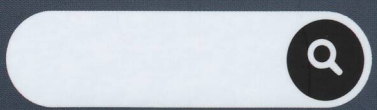
SETH STEPHENS-DAVIDOWITZ

Nguyễn Hạo Nhiên & Nguyễn Hưởng dịch

# MỌI NGƯỜI ĐỀU NÓI DỐI

## DỮ LIỆU LỚN, DỮ LIỆU MỚI

và những điều Internet tiết lộ về chính chúng ta



EVERYBODY LIES

Mọi người đều nói dối.

Người ta nói dối số li đã uống trước khi về nhà. Họ nói dối số lần đi tập gym một tuần, về giá đôi giày mới mua, và cả về chuyện có đọc quyển sách mà họ đã nói hay không. Họ gọi điện báo nghỉ bệnh khi vẫn khỏe như vâm. Họ nói sẽ liên lạc nhưng rồi bặt vô âm tín. Họ nói rằng chuyện không liên quan đến bạn mặc dù có liên quan. Họ nói họ yêu bạn dù rằng họ không hề yêu. Họ nói họ vui dù rằng đang buồn chán. Họ nói họ thích phụ nữ dù thực tế họ thích đàn ông.

Người ta nói dối với bạn bè. Họ nói dối với ông chủ. Họ nói dối với trẻ con. Họ nói dối với cha mẹ. Họ nói dối với bác sĩ. Họ nói dối với chồng. Họ nói dối với vợ. Họ nói dối với chính mình.

Tận dụng lợi thế cực mạnh của Dữ Liệu Lớn cùng những phương pháp khai thác dữ liệu vô cùng độc đáo và thông minh, tác giả Seth Stephens-Davidowitz đã làm lộ diện điều mà mỗi người thực sự muốn nói sâu bên trong tâm hồn mình.

ISBN 978-604-922-753-0

Giá: 150.000 VND

SETH STEPHENS-DAVIDOWITZ

*Nguyễn Hạo Nhiên & Nguyễn Hưởng dịch*

# MỌI NGƯỜI ĐỀU NÓI DỐI

## DỮ LIỆU LỚN, DỮ LIỆU MỚI

và những điều Internet tiết lộ về chính chúng ta

## Everybody Lies

Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are

Nhà xuất bản Kinh tế Tp. Hồ Chí Minh

**EVERYBODY LIES: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are**



**MỌI NGƯỜI ĐỀU NÓI DỐI: Dữ Liệu Lớn, Dữ Liệu Mới Và Những Điều Internet Tiết Lộ Về Chính Chúng Ta**



Dịch từ bản gốc tiếng Anh **EVERYBODY LIES: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are**, tác giả **Seth Stephens-Davidowitz**.

Người dịch: Nguyễn Hạo Nhiên & Nguyễn Hưởng.

Xuất bản theo hợp đồng giữa Công ty TNHH Ecoblader và Fletcher & Company.

Bản tiếng Việt xuất bản lần đầu năm 2019 theo hợp đồng liên kết của Công ty TNHH Ecoblader và Nhà xuất bản Kinh tế Tp. Hồ Chí Minh.

Dành tặng cha mẹ

# Mục Lục

# Lời Nói Đầu

Kể từ khi các triết gia mơ tưởng về một cái "máy đọc ý nghĩ"—một thiết bị tưởng tượng sẽ hiển thị tư tưởng của một người trên màn hình—các nhà khoa học xã hội đã không ngừng tìm kiếm công cụ để phơi bày bản chất con người. Trong sự nghiệp tâm lí học thực nghiệm của tôi, biết bao nhiêu công cụ khác nhau đến rồi lại đi, và tôi đã thử tất cả—thang đánh giá, thời gian phản ứng, độ giãn đồng tử, não ảnh chức năng, thậm chí chúng tôi đã nghiên cứu cả những bệnh nhân động kinh được cấy các điện cực, vui vẻ trải qua hàng giờ liền tại một cuộc thí nghiệm ngôn ngữ trong khi chờ lên cơn.

Tuy nhiên, không một phương pháp nào kể trên cho ta góc nhìn thông suốt vào trí não con người. Đây là một sự đánh đổi ác độc. Tư tưởng con người là tập hợp những mệnh đề phức tạp; không như kiểu Woody Allen đọc nhanh quyển *Chiến tranh và hòa bình*, ta không chỉ nghĩ "Quyển này viết về mấy người Nga." Thế nhưng, các mệnh đề ấy, với tất cả những ánh hào quang đa chiều rối rắm, rất khó để nhà khoa học phân tích. Tất nhiên, khi mọi người trải lòng, ta hiểu được sự phong phú trong dòng chảy ý thức của họ, nhưng các mẩu độc thoại không phải là bộ dữ liệu lí tưởng để kiểm định giả thuyết. Mặt khác, nếu tập trung vào các phương pháp dễ định lượng—như thời gian phản ứng với lời nói, hoặc phản ứng của da với tranh ảnh—thì ta có thể làm thống kê, nhưng đồng thời cũng đã xay nhuyễn cấu trúc phức tạp của nhận thức thành một con số đơn độc. Ngay cả các phương pháp chụp não ảnh phức tạp nhất cũng chỉ có thể cho biết một tư tưởng được tỏa ra trong không gian 3 chiều như thế nào, chứ không cho biết tư tưởng đó *hàm chứa* những gì.

Làm như đánh đổi giữa tính dễ kiểm soát và sự phong phú chưa đủ khó, các nhà nghiên cứu bản chất con người còn bị hành hạ bởi *Luật số nhỏ*—cái tên mà Amos Tversky và Daniel Kahneman đặt cho tư duy ngụy biện rằng các tính trạng của tổng thể sẽ được phản ánh trong bất cứ mẫu nào, dù nhỏ tới đâu đi nữa. Ngay cả các nhà khoa học giỏi toán nhất cũng bị sai lệch về mặt trực giác trong việc xác định số đối tượng nghiên cứu cần có để có thể loại bỏ đi các ngoại lệ ngẫu nhiên và khái quát hóa cho tất cả người Mĩ, chưa nói là cho toàn bộ loài *Homo sapiens*. Mẫu nghiên cứu còn lệch lạc hơn khi được thu thập theo phương pháp lấy mẫu thuận tiện (convenience sampling), như là tặng tiền uống bia cho sinh viên chẳng hạn.

Quyển sách này nói về một phương pháp nghiên cứu trí não hoàn toàn mới. Dữ Liệu Lớn từ các tìm kiếm Internet và các phản hồi trực tuyến không phải là máy đọc ý nghĩ, nhưng Seth Stephens-Davidowitz chỉ ra rằng nó giúp cho ta cơ hội nhìn trộm chưa từng có vào tâm hồn con người. Trong trạng thái riêng tư trước bàn phím, người ta thú nhận những điều kì lạ nhất, đôi lúc (như trong các trang hẹn hò hoặc các tìm kiếm lời khuyên nghề nghiệp) vì họ còn phải chịu những hệ quả trong đời thực, và đôi lúc lại chính vì họ *không* phải gánh chịu hệ quả: Người ta có thể tự bộc lộ một ước muốn hoặc nỗi sợ mà không bị một người thực phản ứng làm mất tinh thần. Dù sao đi nữa, những người này không chỉ đang nhấn nút, mà còn gõ hàng ngàn tỉ chuỗi kí tự để nói rõ vô vàn suy nghĩ mênh mông của họ. Và còn hay ở chỗ, họ đang để các dấu hiệu kĩ thuật số này ở dạng dễ tổng hợp và phân tích. Họ đến từ mọi nẻo đường đời. Họ đang tham gia vào các thí nghiệm kín, với kích thích đa dạng và với các phản hồi thời gian thực được lập bảng rõ ràng. Hơn hết, họ còn vui vẻ cung cấp các dữ liệu này dưới dạng một lượng số liệu khổng lồ.

*Mọi người đều nói dối* không chỉ là một bằng chứng về khái niệm (proof of concept).[1] Nhiều lần các định kiến của tôi về đất nước và giống

[1] [Người dịch - ND] *Proof of concept* là sự hiện thực hóa một ý tưởng hay phương pháp nào đó, chủ yếu để chứng minh ý tưởng/phương pháp ấy là khả thi, có ý nghĩa ứng dụng thực tiễn. Trong trường hợp này, quyển sách *Mọi người đều nói dối* giúp chứng minh phương pháp nghiên cứu bằng dữ liệu là khả thi và có nhiều tiềm năng ứng dụng.

nòi bị đảo ngược bởi những khám phá của Stephens-Davidowitz. Sự ủng hộ không ngờ cho Donald Trump từ đâu đến? Khi Ann Landers hỏi độc giả của mình năm 1976 là họ thấy hối tiếc đã có con hay không và đã rất sốc khi thấy rằng đa số đã hối tiếc, phải chăng bà đã bị lạc lối do mẫu có tính tự chọn (self selected), không mang tính đại diện?[1] Liệu Internet có đáng trách vì cuộc khủng hoảng "bong bóng máy lọc" (filter bubble) cuối những năm 2010?[2] Điều gì châm ngòi cho các vụ phạm tội vì thù hận? Người ta có tìm các câu đùa để phấn chấn lên không? Và dù nghĩ rằng không gì có thể làm mình sốc, tôi đã phải sốc nhiều lần bởi những gì Internet tiết lộ về hoạt động tình dục con người—gồm việc phát hiện rằng hàng tháng một số phụ nữ tìm kiếm cụm từ "humping stuffed animals." Không thí nghiệm sử dụng thời gian phản ứng, sự giãn đồng tử, hay não ảnh chức năng nào có thể moi ra được thông tin đó.

Mọi người sẽ thích *Mọi người đều nói dối*. Với tính tò mò không biết mệt và tài dí dỏm dễ mến, Stephens-Davidowitz chỉ ra con đường mới cho khoa học xã hội Thế kỉ XXI. Khi đã có cánh cửa lôi cuốn bất tận vào những nỗi ám ảnh con người này, ai mà cần đến máy đọc suy nghĩ nữa?

***—Steven Pinker, 2017***

---

[1] [ND] Trong một số nghiên cứu, có một số đặc tính khiến một nhóm người tích cực tham gia trả lời cho các nghiên cứu, trong khi các nhóm không sở hữu đặc tính đó thì không muốn tham gia, khiến cho mẫu thu thập được bị thiên lệch so với tổng thể. Trong trường hợp trên, người viết đặt ra nghi vấn: Có thể những người đang hối tiếc vì đã có con quan tâm đến vấn đề này nên chấp nhận trả lời nghiên cứu của Landers; còn những người không cảm thấy hối tiếc vốn không quan tâm đến việc tham gia trả lời nghiên cứu ngay từ đầu. Điều này có thể khiến cho mẫu nghiên cứu của Landers có rất nhiều người hối tiếc vì đã có con và bỏ qua rất nhiều người không hối tiếc, sai lệch với thực tế. Mẫu bị thiên lệch như thế này gọi là mẫu tự chọn (self-selected sample), và thường không mang tính đại diện cho tổng thể.

[2] [ND] *Filter bubble* là từ được đặt ra bởi Eli Pariser, nhằm chỉ việc các thuật toán trên Internet lọc và chỉ cung cấp các thông tin phù hợp với quan điểm của người dùng (dựa trên các thông tin của người dùng trong quá khứ), tạo ra một bong bóng ngăn không cho các quan điểm đa dạng, đối lập tiếp cận người dùng.

# Dẫn Nhập

## Các đặc điểm chính của một cuộc cách mạng

Chắc chắn ông ta sẽ thua, họ nói.

Trong loạt bầu cử sơ bộ đảng Cộng hòa năm 2016, các chuyên gia thăm dò ý kiến kết luận rằng Donald Trump sẽ không trụ nổi. Dù sao thì Trump cũng đã xúc phạm nhiều nhóm thiểu số. Các cuộc thăm dò và những người diễn giải bảo chúng tôi rằng ít người Mĩ nào chấp nhận bị xúc phạm như thế.

Hầu hết các chuyên gia thăm dò ý kiến lúc đó nghĩ rằng Trump sẽ thua cuộc tổng tuyển cử. Quá nhiều cử tri có-vẻ-sẽ-đi-bầu nói họ không ưa cái bộ dạng và quan điểm của ông ta.

Nhưng thực ra có một số đầu mối cho thấy Trump có thể thắng cả các cuộc bầu cử sơ bộ lẫn cuộc tổng tuyển cử—trên Internet.

***

Tôi là một chuyên gia dữ liệu Internet. Hàng ngày, tôi theo dõi dấu vết kĩ thuật số mà người ta để lại khi lang thang trên mạng. Từ các lần gõ phím và nhấp chuột, tôi cố gắng hiểu ta thực sự muốn gì, ta thực sự sẽ làm gì, và ta thực sự là ai. Xin để tôi giải thích cách tôi bắt đầu trên con đường bất thường này.

Câu chuyện bắt đầu—có vẻ như lâu lắm rồi—với cuộc bầu cử tổng thống 2008 và một câu hỏi tranh luận trong ngành khoa học xã hội từ rất lâu rồi: Thành kiến chủng tộc quan trọng đến thế nào tại Mĩ?

Barack Obama bấy giờ đang chạy đua làm ứng viên tổng thống người Mĩ gốc Phi đầu tiên của một đảng lớn. Ông thắng—khá dễ dàng. Và các cuộc thăm dò cho rằng chủng tộc không phải là một yếu tố ảnh hưởng đến cử tri Mĩ. Chẳng hạn, Gallup đã thực hiện nhiều cuộc thăm dò trước và sau cuộc tuyển cử đầu tiên của Obama. Họ kết luận gì? Cử tri Mĩ phần lớn không quan tâm việc Barack Obama là người da đen. Ngay sau cuộc tuyển cử, 2 giáo sư nổi tiếng tại UC Berkeley nghiên cứu các dữ liệu thu thập từ khảo sát, ứng dụng thêm các kĩ thuật đào bới dữ liệu phức tạp. Họ đi đến kết luận tương tự.

Và vì vậy, suốt nhiệm kì tổng thống của Obama, điều này trở thành lẽ thường tại nhiều bộ phận truyền thông và các học viện lớn. Những nguồn dữ liệu mà giới truyền thông và các nhà khoa học xã hội đã dùng để tìm hiểu thế giới hơn 80 năm nay cho chúng ta biết rằng đại đa số người Mĩ không quan tâm việc Obama là da đen khi quyết định xem ông có nên là tổng thống của họ hay không.

Đất nước này, sau một thời gian dài bị chế độ nô lệ và bộ luật phân biệt chủng tộc Jim Crow vấy bẩn, dường như cuối cùng đã thôi phán xét qua màu da. Điều này có vẻ muốn nói rằng chủ nghĩa phân biệt chủng tộc đã không còn chỗ đứng trên đất Mĩ. Thực vậy, một số chuyên gia còn tuyên bố là chúng ta đã sống trong một xã hội hậu chủng tộc.

Năm 2012, tôi là nghiên cứu sinh kinh tế học, lạc lối trong cuộc đời, kiệt sức trong công việc, và tự tin, thậm chí vênh váo, rằng tôi khá hiểu biết về cách thế giới này vận hành, về những gì người ta nghĩ và quan tâm trong Thế kỉ XXI. Và khi đối mặt với vụ thành kiến chủng tộc này, tôi cho phép mình tin, trên cơ sở mọi thứ tôi đã đọc về mảng tâm lí học và khoa học chính trị, rằng sự phân biệt chủng tộc rõ ràng chỉ giới hạn ở một tỉ lệ nhỏ người Mĩ—đa số trong đó là người đảng Cộng hòa bảo thủ, và hầu hết đều đang sống ở Miền Nam xa xôi.

Sau đó, tôi phát hiện Google Trends.

Google Trends, một công cụ được phát hành một cách không phô trương năm 2009, cho người dùng biết một từ hoặc cụm từ đã được tìm kiếm thường xuyên như thế nào ở các vị trí khác nhau vào các thời điểm

khác nhau. Nó được quảng cáo là một công cụ thú vị—có lẽ dành cho các hội bạn thảo luận tên tuổi nào đang được ưa thích nhất, hay phong cách thời trang nào đang nóng sốt bất ngờ. Các phiên bản đầu tiên có một lời cảnh cáo khôi hài là mọi người "sẽ không muốn viết luận án tiến sĩ" với dữ liệu từ đó—và câu nói này lập tức động viên tôi viết luận án với dữ liệu từ Google Trends.[1]

Hồi đó, dữ liệu tìm kiếm Google dường như không phải là một nguồn thông tin thích hợp cho nghiên cứu học thuật "nghiêm túc." Không như các khảo sát, dữ liệu tìm kiếm Google không được tạo ra để giúp ta hiểu tâm hồn con người. Google được phát minh để người ta tìm hiểu về thế giới, không phải để các nhà nghiên cứu hiểu về con người. Nhưng hóa ra các dấu vết mà ta để lại khi tìm kiếm kiến thức trên Internet lại đang tiết lộ rất nhiều điều.

Nói cách khác, việc tìm kiếm thông tin, tự nó, chính là thông tin. Thời điểm và địa điểm người ta tìm kiếm các sự kiện, trích dẫn, trò đùa, nơi chốn, nhân vật, đồ vật, hoặc sự trợ giúp, hóa ra, có thể cho ta biết họ thực sự nghĩ, thực sự ao ước, thực sự lo sợ, và thực sự làm những gì, nhiều hơn bất cứ ai có thể đoán được. Điều này đặc biệt đúng vì người ta nhiều lúc chẳng hề nghi ngờ Google và thổ lộ trong đó: "Tôi ghét lão chủ." "Tao say." "Bố tôi đánh tôi."

Hoạt động thường ngày—gõ một từ hoặc cụm từ vào cái ô trắng hình chữ nhật bé xíu—để lại một vết sự thật, thứ mà khi nhân lên hàng triệu lần sẽ dần tiết lộ những thực tế sâu sắc. Từ đầu tiên tôi gõ vào Google Trends là "God" (Chúa). Tôi biết được rằng các tiểu bang tìm kiếm Google đề cập từ "God" nhiều nhất là Alabama, Mississippi, và

---

[1] Google Trends là nguồn cung cấp rất nhiều dữ liệu cho tôi. Tuy nhiên, vì nó chỉ cho phép so sánh tần suất tương đối các tìm kiếm khác nhau nhưng không báo cáo con số tuyệt đối cho một tìm kiếm cụ thể, nên tôi phải thường xuyên bổ sung nó bằng Google AdWords—công cụ này báo cáo chính xác tần số của mọi tìm kiếm. Hầu hết trường hợp tôi còn có thể làm sắc nét bức tranh với sự trợ giúp bởi thuật toán dựa trên Trends do chính tôi xây dựng, cái mà tôi đã mô tả trong luận án "Essays Using Google Data" và trong bài "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data" đăng trên tạp chí *Journal of Public Economics*. Luận án, đường dẫn tới bài báo, và lời giải thích đầy đủ về dữ liệu và mật mã dùng trong tất cả nghiên cứu gốc được trình bày trong sách này có sẵn trên website của tôi, sethsd.com.

Arkansas—vùng Vành đai Kinh thánh. Và các tìm kiếm đó rất thường xuyên vào Chủ nhật. Việc đó không có gì đáng ngạc nhiên, nhưng điều thú vị là dữ liệu tìm kiếm tiết lộ một mô thức thật rõ ràng. Tôi thử gõ "Knicks" (một đội bóng rổ), hóa ra nó được Google nhiều nhất ở Thành phố New York. Một điều dễ hiểu khác. Sau đó tôi gõ tên tôi. "Rất tiếc," Google Trends nói. "Không có đủ lượng tìm kiếm" để hiện kết quả. Tôi được biết, Google Trends sẽ cung cấp dữ liệu chỉ khi nhiều người cùng thực hiện một tìm kiếm giống nhau.

Nhưng sức mạnh của các tìm kiếm Google không phải nằm ở chỗ có thể cho ta biết rằng Chúa được kính mến ở Miền Nam, đội Knicks được yêu thích ở Thành phố New York, hoặc tôi không nổi tiếng ở đâu cả. Cuộc khảo sát nào cũng có thể cho bạn biết điều đó. Sức mạnh trong dữ liệu Google nằm ở chỗ, mọi người nói cho cỗ máy tìm kiếm khổng lồ đó biết những thứ mà có thể họ sẽ không nói với bất kì ai khác.

Lấy chuyện tình dục làm ví dụ (một chủ đề mà tôi sẽ khảo sát chi tiết hơn trong phần sau quyển sách này). Rõ là không thể tin các khảo sát về đời sống tình dục được. Tôi phân tích dữ liệu từ General Social Survey, đây được xem là một trong những nguồn cung cấp thông tin về hành vi của người Mĩ có ảnh hưởng và đáng tin cậy nhất. Theo khảo sát đó, với tình dục khác giới, phụ nữ nói họ sinh hoạt tình dục trung bình 50 lần/năm, dùng 1 bao cao su 16% số lần. Thế thì tổng cộng phải có khoảng 1.1 tỉ bao cao su được dùng mỗi năm. Nhưng nam giới dị tính lại nói họ dùng 1.6 tỉ bao cao su/năm. Những con số đó, về bản chất, là phải giống nhau. Vậy thì ai đang nói thật, nam giới hay nữ giới?

Hóa ra không bên nào nói thật cả. Theo Nielsen, công ty đo lường và thông tin toàn cầu, chuyên theo dõi hành vi người tiêu dùng, chưa tới 600 triệu bao cao su được bán ra mỗi năm. Vậy mọi người đều nói dối; cái khác duy nhất là mức độ nói dối mà thôi.

Việc nói dối thực ra rất phổ biến. Nam giới chưa bao giờ kết hôn nói họ dùng trung bình 29 bao cao su mỗi năm. Thế thì tổng cộng nhiều hơn tổng số bao cao su được bán ở Mĩ cho người đã kết hôn và độc thân gộp lại. Người đã kết hôn cũng phóng đại mức độ hoạt động tình dục của họ.

Trung bình, nam giới U65 đã kết hôn nói trong các cuộc khảo sát rằng họ sinh hoạt mỗi tuần 1 lần. Chỉ 1% nói họ đã qua 1 năm không sinh hoạt tình dục. Nữ giới đã kết hôn cho biết họ sinh hoạt ít hơn một chút nhưng không chênh lệch nhiều lắm.

Các tìm kiếm Google cung cấp một bức tranh tình dục trong hôn nhân kém sinh động hơn nhiều—và theo tôi là chính xác hơn nhiều. Trên Google, phàn nàn nhiều nhất về hôn nhân là không sinh hoạt tình dục. Các tìm kiếm "sexless marriage" (hôn nhân không tình dục) thường gấp 3.5 lần "unhappy marriage" (hôn nhân không hạnh phúc) và gấp 8 lần "loveless marriage" (hôn nhân không tình yêu). Ngay cả các cặp chưa kết hôn cũng phàn nàn khá thường xuyên về việc không có hoạt động tình dục. Các tìm kiếm Google cụm từ "sexless relationship" (mối quan hệ không tình dục) đứng thứ nhì chỉ sau các tìm kiếm "abusive relationship" (mối quan hệ lạm dụng). (Dữ liệu này, tôi phải nhấn mạnh, tất cả đều được trình bày nặc danh. Google, dĩ nhiên, không cho biết dữ liệu về các tìm kiếm của bất kì cá nhân cụ thể nào.)

Và các tìm kiếm Google cho thấy một bức tranh nước Mĩ khác xa với cái thiên đường hậu chủng tộc được phác họa bởi các cuộc khảo sát. Tôi còn nhớ lần đầu gõ "nigger" (mọi đen)[1] vào Google Trends. Tôi ngây thơ thật. Từ này rất độc hại, nên tôi nghĩ lượng tìm kiếm sẽ rất thấp. Trời ạ, là tôi sai. Ở Mĩ, từ "nigger(s)" đạt số lượng tìm kiếm gần như tương đương với từ "migraine(s)" (bệnh đau nửa đầu), "economist" (nhà kinh tế), và "Lakers" (tên một đội bóng rổ). Tôi tự hỏi liệu các tìm kiếm lời nhạc rap có đang làm sai lệch các kết quả đó không? Không. Cái từ đó khi dùng trong các bài hát rap hầu như luôn luôn viết là "nigga(s)." Vậy động cơ để người Mĩ tìm kiếm từ "nigger" là gì? Thường là khi họ tìm các câu chuyện cười chế nhạo người Mĩ gốc Phi. Thực vậy, 20% các tìm

[1] [ND] Do một số từ khóa tìm kiếm mang tính đặc trưng tùy vào ngôn ngữ, nên với một số từ, nhóm dịch sẽ để nguyên gốc kèm từ tạm dịch ở bên cạnh. Với một số từ ngữ nhạy cảm nhưng không ảnh hưởng đến ý chung của sách (và bạn đọc có thể dễ dàng tìm hiểu bằng cách tra từ khóa đó trên Google), nhóm dịch sẽ giữ nguyên và không giải nghĩa. Nhóm cũng sẽ xử lí tương tự với các từ ngữ tiếng Anh phổ biến trên các trang mạng xã hội lớn hay các thiết bị điện tử thông dụng, vì nhiều bạn đọc quen với các từ tiếng Anh hơn là từ tiếng Việt (ví dụ, nút "Home" trên iPhone sẽ dễ hiểu hơn khi được giữ nguyên gốc).

kiếm với từ "nigger" cũng bao hàm từ "jokes" (chuyện cười). Các tìm kiếm thông thường khác bao gồm các cụm từ "stupid niggers" (bọn mọi đen ngu ngốc) và "I hate niggers" (tao ghét bọn mọi đen).

Có hàng triệu tìm kiếm như vậy mỗi năm. Rất nhiều người Mĩ ở cõi riêng tư trong nhà mình đang thực hiện những câu hỏi phân biệt chủng tộc gây sốc. Càng nghiên cứu, thông tin thu được càng đáng lo hơn.

Vào đêm tuyển cử đầu tiên của Obama, khi hầu hết các bình luận tập trung vào ca ngợi Obama và công nhận bản chất lịch sử của sự kiện này, cứ khoảng 100 tìm kiếm Google chứa từ "Obama" thì có 1 tìm kiếm có luôn cả "kkk" (Ku-Klux-Klan) hoặc "nigger(s)." Nghe có vẻ không nhiều lắm, nhưng nghĩ xem, có đến hàng ngàn lí do phi chủng tộc để tra Google về con người ngoài cuộc trẻ trung này cùng gia đình đầy quyến rũ sắp nắm giữ vị trí quyền lực nhất thế giới. Vào đêm tuyển cử, các tìm kiếm và lượt đăng kí Stormfr***[1] (một trang mạng dân tộc chủ nghĩa da trắng có mức phổ biến cao khác thường ở Mĩ) tăng gấp 10 lần bình thường. Trong một số bang, có nhiều tìm kiếm "nigger president" (tổng thống mọi đen) hơn "first black president" (tổng thống da đen đầu tiên).

Rõ ràng có tồn tại tình trạng đen tối và thù hận, thứ đã bị che giấu khỏi các nguồn truyền thống nhưng khá rõ ràng trong các tìm kiếm mà người ta thực hiện.

Các tìm kiếm kia khó mà ăn nhập với một xã hội trong đó phân biệt chủng tộc là một yếu tố nhỏ nhoi. Năm 2012, tôi biết đến Donald J. Trump chủ yếu với vai trò doanh nhân kiêm nhà trình diễn trong chương trình truyền hình thực tế. Tôi cũng như bất cứ ai khác không nghĩ rằng 4 năm sau đó ông ấy sẽ là một ứng viên tổng thống thật sự. Nhưng các tìm kiếm khó chịu trên kia thật dễ dàng hòa hợp với thành công của một ứng viên—qua các vụ tấn công người nhập cư, trong những cơn bực mình và nóng giận—thường nhắm vào các khuynh hướng tồi tệ nhất của mọi người.

***

[1] [ND] Với các trang web có nội dung nhạy cảm, kích động, nhóm dịch sẽ ẩn đi một phần tên trang web để tránh vô ý lan truyền các nội dung này.

Các tìm kiếm Google đó còn cho ta biết rằng những gì ta nghĩ về địa điểm có sự phân biệt chủng tộc cũng sai nốt. Các khảo sát và hiểu biết thông thường cho rằng phân biệt chủng tộc hiện đại chủ yếu diễn ra ở Miền Nam và ở những người theo đảng Cộng hòa. Nhưng các nơi có tỉ lệ tìm kiếm phân biệt chủng tộc cao nhất cũng có vùng phía bắc New York, tây Pennsylvania, đông Ohio, vùng công nghiệp Michigan và vùng nông thôn Illinois, cùng với West Virginia, vùng phía nam Louisiana, và Mississippi. Dữ liệu tìm kiếm Google cho biết, ranh giới thực sự không phải là giữa Nam với Bắc, mà là giữa Đông với Tây. Ta không thấy tình trạng này nhiều ở phía tây Mississippi. Và phân biệt chủng tộc không giới hạn trong những người đảng Cộng hòa. Thực ra, các tìm kiếm phân biệt chủng tộc ở những nơi có tỉ lệ người đảng Cộng hòa cao không cao hơn những nơi có tỉ lệ người đảng Dân chủ cao. Nói cách khác, các tìm kiếm Google giúp vẽ một bản đồ mới về phân biệt chủng tộc ở Mĩ—và bản đồ này trông rất khác với những gì ta tưởng. Người đảng Cộng hòa ở Miền Nam có thể sẽ dễ thừa nhận là mình phân biệt chủng tộc hơn. Nhưng nhiều người đảng Dân chủ ở Miền Bắc cũng có thái độ tương tự.

Sau đó 4 năm, bản đồ này đã chứng tỏ giá trị của mình trong việc giải thích thành công chính trị của Trump.

Năm 2012, tôi sử dụng bản đồ phân biệt chủng tộc tôi đã phát triển khi dùng các tìm kiếm Google để đánh giá lại chính xác ảnh hưởng của chủng tộc đối với Obama. Dữ liệu rất rõ ràng. Tại những nơi có lượng tìm kiếm phân biệt chủng tộc cao, Obama kém hơn đáng kể so với John Kerry, ứng viên tổng thống đảng Dân chủ người da trắng 4 năm trước đó. Mối quan hệ đó không được giải thích bởi bất cứ nhân tố nào khác về các vùng này, bao gồm trình độ giáo dục, tuổi tác, việc đi nhà thờ, hoặc việc sở hữu súng. Các tìm kiếm phân biệt chủng tộc không dự báo kết quả kém hơn ở bất cứ ứng viên Dân chủ nào khác. Chỉ với Obama thôi.

Và kết quả ám chỉ một ảnh hưởng khá lớn. Obama mất khoảng 4 điểm phần trăm[1] toàn quốc chỉ từ sự phân biệt chủng tộc rõ ràng. Kết

[1] [ND] "Điểm phần trăm" chỉ mức tăng tuyệt đối về số phần trăm. Ví dụ, từ 10% lên 15% nghĩa là tăng 5 "điểm phần trăm" (15% - 10%), nhưng là tăng 50% so với con số ban đầu (5%/10%). Trong quyển sách này, điểm phần trăm được kí hiệu là %p.

quả này cao hơn con số kì vọng ở bất cứ khảo sát nào. Barack Obama, dĩ nhiên, đã đắc cử và tái đắc cử tổng thống nhờ một số điều kiện rất thuận lợi cho người đảng Dân chủ, nhưng ông phải vượt qua nhiều thử thách hơn sự tưởng tượng của bất cứ ai lệ thuộc vào các nguồn dữ liệu truyền thống—tức là gần như tất cả mọi người. Tồn tại một lượng người phân biệt chủng tộc đủ để giúp thắng cuộc bầu cử sơ bộ hoặc làm thay đổi kết quả tổng tuyển cử trong một năm không thuận lợi đối với người đảng Dân chủ.

Nghiên cứu của tôi ban đầu bị 5 tạp chí học thuật từ chối. Nhiều người trong ban bình duyệt—xin độc giả cho phép tôi bất mãn tí chút ở đây—nói rằng không thể nào có chuyện quá nhiều người Mĩ ấp ủ tư tưởng phân biệt chủng tộc xấu xa như thế. Điều này đơn giản là không khớp với những gì người ta vẫn luôn nói. Hơn nữa, các tìm kiếm Google có vẻ như là một bộ dữ liệu kì dị.

Vì bây giờ chúng ta đã chứng kiến lễ nhậm chức của Tổng thống Donald J. Trump, phát hiện của tôi bắt đầu có vẻ hợp lí hơn.

***

Càng nghiên cứu, tôi càng phát hiện ra rằng Google có nhiều thông tin bị các cuộc thăm dò bỏ qua nhưng có thể hữu ích trong việc hiểu một cuộc bầu cử—giữa rất, rất nhiều chủ đề khác.

Có thông tin về việc ai sẽ thực sự đi bầu. Hơn một nửa công dân không đi bầu nói với các cuộc khảo sát ngay trước bầu cử rằng họ có ý định đi bầu, làm sai lệch ước tính số lượng người. Trong khi đó, các tìm kiếm Google cụm từ "how to vote" (cách bầu cử) hoặc "where to vote" (nơi bầu cử) nhiều tuần trước một cuộc bầu cử có thể dự đoán chính xác vùng nào trong nước sẽ có số lượt đi bầu lớn.

Thậm chí có thể có thông tin là họ sẽ bầu cho ai. Ta có thể dự báo người ta sẽ bầu ứng viên nào chỉ dựa trên những gì họ tìm kiếm không? Rõ ràng, ta không thể chỉ nghiên cứu các ứng viên nào được tìm kiếm thường xuyên nhất. Nhiều người tìm kiếm một ứng viên vì họ yêu ông ta. Một số người lại tìm kiếm một ứng viên vì họ ghét ông ta. Ngay cả vậy, Stuart Gabriel, giáo sư tài chính UC Los Angeles, và tôi đã phát hiện

một đầu mối đáng ngạc nhiên về dự định bầu cử của mọi người. Một tỉ lệ lớn các tìm kiếm liên quan đến bầu cử chứa các câu hỏi với tên của cả hai ứng viên. Trong cuộc tuyển cử năm 2016 giữa Trump và Hillary Clinton, một số người tìm kiếm "thăm dò Trump Clinton". Những người khác tìm các điểm nổi bật nhất từ "Clinton Trump tranh luận." Thực tế, 12% các truy vấn tìm kiếm với "Trump" cũng gồm cả từ "Clinton." Hơn 1/4 lượng truy vấn tìm kiếm với "Clinton" cũng gồm cả từ "Trump."

Chúng tôi đã phát hiện rằng các tìm kiếm dường như trung tính này thực ra có thể là đầu mối cho thấy người ta ủng hộ ứng viên nào.

Bằng cách nào? Thứ tự các ứng viên xuất hiện. Nghiên cứu của chúng tôi chỉ ra rằng người ta chắc chắn sẽ đặt ứng viên họ ủng hộ nằm trước trong các tìm kiếm bao gồm tên của cả hai ứng viên.

Trong 3 cuộc tuyển cử trước, ứng viên xuất hiện trước trong nhiều tìm kiếm hơn nhận nhiều phiếu bầu hơn. Thú vị hơn, thứ tự các ứng viên được tìm kiếm có thể dự báo một bang cụ thể sẽ đi con đường nào.

Thứ tự xuất hiện của ứng viên trong các tìm kiếm dường như còn chứa thông tin mà các cuộc thăm dò có thể bỏ qua. Trong cuộc tuyển cử năm 2012 giữa Obama và đảng viên Cộng hòa Mitt Romney, nhà thống kê kiêm nhà báo bậc thầy Nate Silver dự báo chính xác kết quả của tất cả 50 tiểu bang. Tuy nhiên, chúng tôi thấy rằng tại các tiểu bang để tên Romney trước Obama trong các tìm kiếm nhiều hơn, Romney thực sự làm tốt hơn dự báo của Silver. Tại các tiểu bang thường xuyên để tên Obama trước Romney, Obama làm tốt hơn dự báo của Silver.

Chỉ báo này có thể chứa thông tin mà các cuộc thăm dò bỏ qua vì cử tri đang tự dối mình hoặc không thoải mái tiết lộ ý định thực sự của họ với người thăm dò. Có lẽ nếu họ nói rằng họ không quyết định được trong năm 2012, nhưng thường tìm kiếm "thăm dò Romney Obama," "Romney Obama tranh luận," và "cuộc tuyển cử Romney Obama," thì tức là họ đã có kế hoạch bầu cho Romney ngay từ đầu rồi.

Vậy Google đã dự báo Trump à? Chà, chúng tôi vẫn còn phải làm nhiều việc—và phải có thêm nhiều nhà nghiên cứu tham gia—trước khi biết cách tốt nhất để dùng dữ liệu Google dự báo kết quả bầu cử. Đây là

một môn khoa học mới, mà chúng tôi chỉ mới có vài cuộc bầu cử có tồn tại dữ liệu kiểu này. Dĩ nhiên tôi không định nói ta đang ở thời điểm—hoặc sẽ có thời điểm—khi mà các cuộc thăm dò ý kiến công chúng sẽ hoàn toàn bị loại bỏ khi cần dự báo các cuộc bầu cử.

Nhưng rõ ràng có những điềm báo, ở nhiều thời điểm, trên Internet, cho thấy Trump có thể đã làm tốt hơn dự đoán của các cuộc thăm dò.

Trong suốt cuộc tổng tuyển cử, có các dấu hiệu cho thấy cử tri có thể ủng hộ Trump. Người Mĩ da đen nói với các cuộc thăm dò là họ sẽ đi bầu với số lượng lớn để chống lại Trump. Nhưng các tìm kiếm Google về thông tin bầu cử tại những vùng chủ yếu là người da đen lại rất thấp. Vào ngày bầu cử, Clinton bị thiệt hại bởi số người da đen đi bầu thấp.

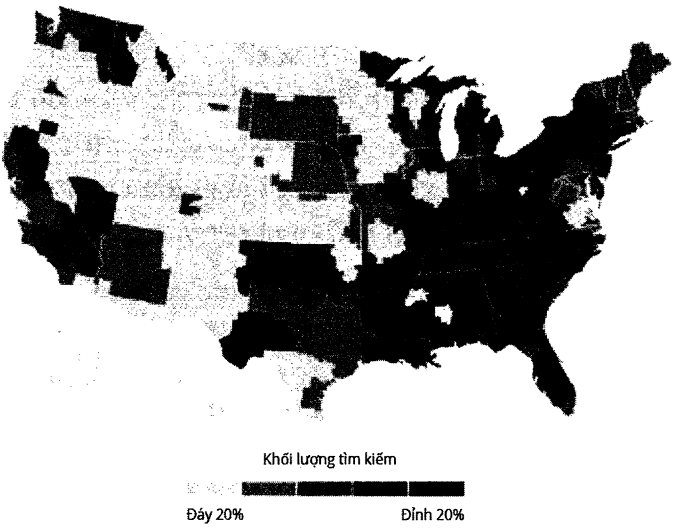
Thậm chí có những dấu hiệu cho thấy các cử tri được cho là chưa quyết định bầu cho ai lại đang theo bước chân Trump. Gabriel và tôi phát hiện rằng có nhiều tìm kiếm "Trump Clinton" hơn "Clinton Trump" tại các tiểu bang chủ chốt ở vùng Trung Tây mà Clinton được kì vọng là sẽ chiến thắng. Thực tế, Trump thắng cử là nhờ ông đã đạt kết quả tốt hơn rất nhiều so với các cuộc thăm dò tại các địa phương này.

Nhưng theo tôi, đầu mối chính cho thấy Trump có thể thắng cử—trước hết là ở các cuộc bầu cử sơ bộ—là do sự phân biệt chủng tộc ngấm ngầm mà nghiên cứu Obama của tôi đã khám phá ra. Các tìm kiếm Google tiết lộ một bức tranh đen tối và thù hận ở một số người Mĩ đáng kể, thứ mà các chuyên gia trong suốt nhiều năm đã bỏ qua. Dữ liệu tìm kiếm tiết lộ rằng chúng ta sống trong một xã hội rất khác với xã hội mà giới hàn lâm và báo chí—dựa vào các cuộc thăm dò—vẫn tưởng. Nó tiết lộ một cơn thịnh nộ xấu xa, đáng sợ, tràn lan; và nó đang chờ đợi một ứng viên lên tiếng.

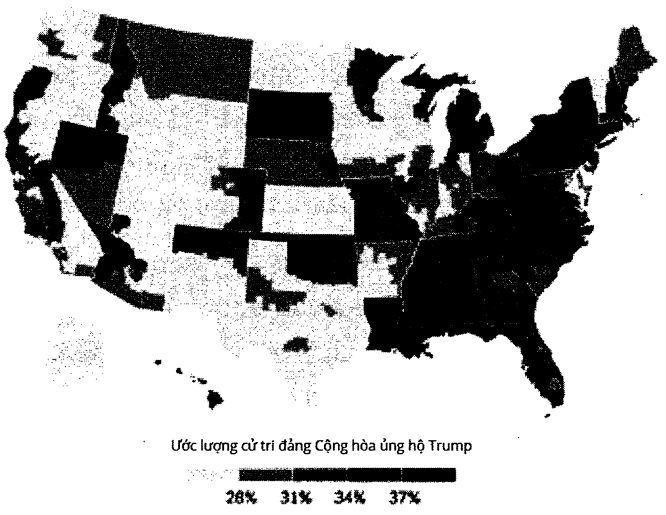
Người ta thường nói dối—với chính mình và với người khác. Năm 2008, người Mĩ nói với các cuộc khảo sát rằng họ không còn để ý về chủng tộc nữa. Sau đó 8 năm, họ chọn Donald J. Trump làm tổng thống, người đã retweet một phát biểu sai lệch trên Twitter rằng người da đen chịu trách nhiệm về đa số vụ ám sát người Mĩ da trắng, người đã bảo vệ những người ủng hộ ông trong vụ bạo hành một người phản đối thuộc

phong trào Black Lives Matters tại một trong các cuộc hội họp của ông, và là người đã do dự trong việc từ chối sự ủng hộ của một cựu thủ lĩnh tổ chức Ku Klux Klan. Chính sự phân biệt chủng tộc ngấm ngầm gây thiệt hại Barack Obama đã giúp Donald Trump.

Tỉ lệ tìm kiếm phân biệt chủng tộc




Tình hình ủng hộ Trump trong cuộc bầu cử sơ bộ đảng Cộng hòa



Từ đầu các cuộc bầu cử sơ bộ, ta đều biết Nate Silver đã tuyên bố rằng hầu như không có cơ hội nào cho Trump thắng cử. Khi các cuộc bầu cử sơ bộ tiến triển và ngày càng thấy rõ rằng Trump đang nhận được sự ủng hộ lan rộng, Silver quyết định nghiên cứu dữ liệu để tìm hiểu điều gì đang diễn ra. Sao mà Trump lại đang làm tốt đến thế?

Silver chú ý thấy rằng các vùng mà Trump thể hiện tốt nhất tạo nên một bản đồ kì lạ. Trump làm tốt ở những vùng Đông Bắc và vùng công nghiệp Trung Tây, cũng như ở Miền Nam. Ông làm tệ hơn thấy rõ ở Miền Tây. Silver tìm các biến để thử giải thích bản đồ này. Phải chăng biến đó là tình trạng thất nghiệp? Là tôn giáo? Sự sở hữu súng? Tỉ lệ người nhập cư? Sự chống đối Obama?

Silver phát hiện rằng nhân tố đơn lẻ tương quan tốt nhất với sự ủng hộ Trump tại các cuộc bầu cử sơ bộ đảng Cộng hòa chính là thứ tôi đã phát hiện 4 năm trước đó. Những vùng ủng hộ Trump với số lượng lớn nhất là những vùng đã tìm kiếm từ "nigger" nhiều nhất trên Google.

***

Suốt 4 năm qua, hầu như ngày nào tôi cũng phân tích dữ liệu Google. Có một giai đoạn tôi làm nhà khoa học dữ liệu tại Google, họ thuê tôi sau khi biết về nghiên cứu phân biệt chủng tộc của tôi. Tôi tiếp tục khám phá dữ liệu này khi viết xã luận và viết các bài báo dữ liệu cho *New York Times*. Tiếp tục có những thông tin mới. Bệnh thần kinh; tình dục; ngược đãi trẻ em; nạo phá thai; quảng cáo; tôn giáo; sức khỏe. Không phải là các chủ đề nhỏ. Bộ dữ liệu này—thứ không hề tồn tại vài chục năm trước—đã cho ta những góc nhìn mới bất ngờ lên tất cả. Giới kinh tế học và khoa học xã hội luôn săn tìm các nguồn dữ liệu mới, vậy tôi xin nói thẳng: Bây giờ tôi đã tin rằng các tìm kiếm Google là bộ dữ liệu quan trọng nhất từng được thu thập về tâm hồn con người.

Tuy nhiên, bộ dữ liệu này không phải là công cụ duy nhất mà Internet đã cung cấp để hiểu thế giới chúng ta. Tôi sớm nhận ra có những mỏ vàng kĩ thuật số khác nữa. Tôi tải về toàn bộ Wikipedia, nghiền ngẫm hết các hồ sơ Facebook, và cóp nhặt Stormfr***. Ngoài ra, P***hub, một trong các trang mạng khiêu dâm lớn nhất trên Internet, đã

cho tôi dữ liệu đầy đủ về các tìm kiếm và lượt xem video của những người nặc danh khắp thế giới. Nói cách khác, tôi đã lặn rất sâu vào cái mà bây giờ được gọi là Big Data—Dữ Liệu Lớn. Tôi đã phỏng vấn hàng chục người khác—các nhà hàn lâm, nhà báo dữ liệu, và doanh nhân—họ cũng đang khám phá các lĩnh vực mới này. Nhiều nghiên cứu của họ sẽ được thảo luận ở đây.

Nhưng trước tiên, tôi xin thú nhận: Tôi không định định nghĩa chính xác Dữ Liệu Lớn là gì. Tại sao? Tại vì đó là một khái niệm vốn rất mơ hồ. Lớn bao nhiêu mới là lớn? 18,462 quan sát là Dữ Liệu Nhỏ còn 18,463 quan sát là Dữ Liệu Lớn chăng? Tôi chỉ muốn nhìn bao quát những thứ đủ điều kiện là Dữ Liệu Lớn: Mặc dù hầu hết dữ liệu tôi dùng là từ Internet, tôi cũng sẽ thảo luận các nguồn khác nữa. Chúng ta đang sống qua một cuộc bùng nổ về số lượng và chất lượng thông tin. Phần nhiều thông tin mới đó lưu chuyển từ Google và các phương tiện truyền thông xã hội. Một phần trong đó là sản phẩm của sự số hóa thông tin, thứ trước đây vốn bị ẩn trong các tủ hồ sơ. Một phần dữ liệu là từ các nguồn lực ngày càng gia tăng dành cho nghiên cứu thị trường. Một số nghiên cứu thảo luận trong sách này không dùng các bộ dữ liệu lớn mà chỉ áp dụng phương pháp tiếp cận mới và sáng tạo với dữ liệu—các phương pháp cốt yếu trong một thời đại tràn đầy thông tin.

Vậy chính xác thì tại sao Dữ Liệu Lớn lại mạnh mẽ như vậy? Thử nghĩ về tất cả lượng thông tin rải rác đầy trên mạng vào một ngày nào đó mà xem. Mà thực tế là ta có một con số chính xác cho lượng thông tin này. Một ngày trung bình trong thời kì đầu Thế kỉ XXI, nhân loại tạo ra 2.5 tỉ tỉ byte dữ liệu.

Và các byte này chính là manh mối.

> Một cô đang buồn chán vào một chiều thứ Năm. Cô Google thêm một vài "chuyện cười vui." Cô mở email. Cô đăng nhập Twitter. Cô google cụm từ "chuyện cười mọi đen."
>
> Một ông đang cảm thấy buồn. Ông Google "các triệu chứng trầm cảm" và "các câu chuyện trầm cảm." Ông chơi một ván solitaire.
>
> Một cô thấy thông báo bạn cô đính hôn trên Facebook. Cô gái—còn độc thân—chặn luôn bạn mình.

Một anh tạm nghỉ Google về NFL và nhạc rap để hỏi máy tìm kiếm: "Mơ mình hôn đàn ông có bình thường không?"

Một cô nhấp vào một câu chuyện trên BuzzFeed về "15 con mèo dễ thương nhất."

Một anh thấy cùng câu chuyện mèo. Nhưng trên màn hình anh nó được gọi là "15 con mèo đáng yêu nhất." Anh không nhấp chuột.

Một bà Google "Con trai tôi có phải thiên tài không?"

Một ông Google "cách để con gái tôi giảm cân."

Một bà đang đi nghỉ với 6 bà bạn thân nhất. Tất cả các bạn bà luôn miệng nói họ đang rất là vui. Bà lẻn đi Google "cô đơn khi xa chồng."

Một ông, chồng người phụ nữ nói trên, đang đi nghỉ với 6 ông bạn thân nhất. Ông lẻn đi để Google "dấu hiệu vợ bạn đang ngoại tình."

Một số dữ liệu trên đây chứa các thông tin mà có lẽ không bao giờ được thú nhận với bất cứ ai. Nếu tổng hợp lại tất cả, giữ nó nặc danh để không làm lộ những nỗi sợ, ước mơ, và hành vi của bất cứ cá nhân cụ thể nào, rồi bổ sung vào đó một ít khoa học dữ liệu, ta sẽ có được một cái nhìn mới mẻ về con người—hành vi, ước muốn, bản chất của họ. Thực vậy, xin nói chuyện vĩ đại một chút: Tôi đã bắt đầu tin rằng dữ liệu mới—thứ ngày càng có nhiều trong thời đại kĩ thuật số—căn bản sẽ mở rộng hiểu biết về nhân loại. Kính hiển vi cho thấy trong một giọt nước ao chứa đựng nhiều thứ hơn ta tưởng. Kính viễn vọng cho thấy có nhiều thứ trên bầu trời đêm hơn ta tưởng. Và dữ liệu kĩ thuật số mới sẽ cho thấy có nhiều thứ trong xã hội loài người hơn ta tưởng. Đó có thể là kính hiển vi hoặc kính viễn vọng của thời đại chúng ta—có thể mang đến những hiểu biết quan trọng, thậm chí mang tính cách mạng nữa.

Có một nguy cơ trong những tuyên bố như thế—các tuyên bố ấy không chỉ nghe có vẻ vĩ đại mà còn hiện đại nữa. Nhiều người đã có những tuyên bố to tát về sức mạnh của Dữ Liệu Lớn. Nhưng họ lại thiếu bằng chứng.

Điều này truyền cảm hứng cho những người hoài nghi Dữ Liệu Lớn—tức là rất nhiều người—gạt bỏ việc tìm kiếm các bộ dữ liệu lớn

hơn. "Ở đây tôi không định nói rằng không có thông tin gì trong Dữ Liệu Lớn," tác gia kiêm nhà thống kê Nassim Taleb đã viết. "Ở đó có nhiều thông tin. Khó khăn nằm ở chỗ, cây kim đang lẫn trong một đống cỏ khô ngày càng lớn."

Thế thì, một trong những mục tiêu ban đầu của sách này là cung cấp bằng chứng còn thiếu về những thứ ta có thể được làm với Dữ Liệu Lớn—về cách tìm các cây kim trong các đống cỏ khô ngày càng lớn đó. Tôi hi vọng là mình cung cấp đủ ví dụ về việc Dữ Liệu Lớn đã mang những hiểu biết mới vào trong ngành tâm lí học và hành vi con người để bạn thấy được các đặc điểm chính của một thứ thực sự mang tính cách mạng.

"Khoan đã, Seth," bạn nói. "Anh đang hứa hẹn một cuộc cách mạng. Anh đang thi vị hóa các bộ dữ liệu to lớn và mới mẻ này rồi. Nhưng cho đến nay, anh đã dùng tất cả dữ liệu đáng kinh ngạc, khác thường, hấp dẫn, đột phá này chỉ để nói tôi biết 2 điều căn bản: Có nhiều người phân biệt chủng tộc ở Mĩ, và nhiều người, đặc biệt là nam giới, phóng đại khả năng tình dục của họ."

Tôi thừa nhận đôi khi dữ liệu mới chỉ xác nhận những điều đã rõ ràng. Nếu bạn nghĩ các phát hiện trên là quá hiển nhiên, hãy đợi cho đến khi bạn tới Chương 4, ở đó tôi sẽ chỉ bạn thấy bằng chứng rõ ràng, đáng tin cậy từ các tìm kiếm Google cho thấy nam giới cực kì lo lắng và bất an về... kích cỡ dương vật của họ.

Tôi khẳng định, việc chứng minh những thứ mà bạn có thể đã nghi ngờ nhưng lại ít có bằng chứng là hoàn toàn có giá trị. Nghi ngờ cái gì đó là một chuyện. Chứng minh nó là chuyện khác. Nhưng nếu Dữ Liệu Lớn chỉ có thể xác nhận các nghi ngờ, nó sẽ không có tính cách mạng đến vậy. May thay, Dữ Liệu Lớn có thể làm hơn thế rất nhiều. Rất nhiều lần, dữ liệu chỉ cho tôi thấy thế giới hoạt động ngược hoàn toàn dự đoán của tôi. Tiếp theo đây là vài ví dụ có thể làm bạn bất ngờ hơn.

Bạn có thể nghĩ rằng một nguyên nhân chính của sự phân biệt chủng tộc là sự bất an và dễ tổn thương về mặt kinh tế. Thế thì, bạn sẽ nghĩ rằng khi người ta mất việc làm, sự phân biệt chủng tộc gia tăng. Nhưng

thực ra, các tìm kiếm phân biệt chủng tộc và số thành viên của trang Stormfr*** đều không tăng khi thất nghiệp tăng.

Bạn có thể nghĩ rằng tình trạng lo lắng sẽ ở mức cao nhất tại các thành phố lớn đề cao giáo dục quá đà. Bệnh thần kinh thành thị là một hình mẫu khá nổi tiếng. Nhưng các tìm kiếm Google phản ánh sự lo lắng—như "các triệu chứng lo lắng" hoặc "giúp khỏi lo lắng"—có khuynh hướng cao hơn ở các nơi có trình độ giáo dục thấp, thu nhập trung bình thấp, và nơi mà phần lớn dân cư sống ở các vùng nông thôn. Tỉ lệ tìm kiếm về sự lo lắng ở vùng nông thôn Bắc New York cao hơn ở Thành phố New York.

Bạn có thể nghĩ rằng một cuộc tấn công khủng bố giết chết hàng chục hoặc hàng trăm người sẽ kéo theo nỗi lo lắng lớn và tràn lan. Chủ nghĩa khủng bố, theo định nghĩa, thường truyền dẫn cảm giác khiếp sợ. Tôi nghiên cứu các tìm kiếm Google phản ánh sự lo lắng. Tôi kiểm tra các tìm kiếm này tăng bao nhiêu tại một nước vào những ngày, tuần, và tháng theo sau mỗi cuộc tấn công khủng bố lớn ở châu Âu hoặc Mĩ từ năm 2004. Vậy, trung bình, các tìm kiếm liên quan đến sự lo lắng đã tăng bao nhiêu? Không tăng chút nào.

Bạn có thể nghĩ rằng người ta tìm kiếm chuyện cười thường xuyên hơn khi họ buồn. Nhiều nhà tư tưởng vĩ đại nhất lịch sử đã nói rằng chúng ta xem sự hài hước là cách thoát khỏi khổ đau. Sự hài hước từ lâu được xem là một cách để đương đầu với những buồn chán, nỗi đau, nỗi thất vọng không thể tránh trong cuộc sống. Như Charlie Chaplin nói, "Tiếng cười là thuốc bổ, thứ làm ta khuây khỏa, thứ chấm dứt nỗi đau."

Tuy nhiên, các tìm kiếm chuyện cười lại rất thấp vào những ngày thứ Hai, ngày mà người ta cho là họ rất khổ sở. Các tìm kiếm đó rất thấp vào những ngày nhiều mây và mưa; và tụt hẳn sau một bi kịch lớn, như khi 2 quả bom giết chết 3 và làm bị thương hàng trăm người trong sự kiện Boston Marathon năm 2013. Thực ra người ta thích tìm kiếm chuyện cười khi mọi thứ đang tốt đẹp hơn là khi gặp trắc trở.

Đôi lúc một bộ dữ liệu mới tiết lộ một hành vi, ham muốn, hoặc lo lắng mà tôi chưa bao giờ nghĩ tới. Có rất nhiều xu hướng tình dục rơi

vào thể loại này. Ví dụ, bạn có biết rằng tại Ấn Độ thì tìm kiếm số một bắt đầu với cụm từ "chồng tôi muốn..." là "chồng tôi muốn tôi cho anh ấy bú" không? Bình luận này ở Ấn Độ thường xuyên hơn rất nhiều so với ở các nước khác. Hơn nữa, các tìm kiếm khiêu dâm về những hình ảnh mô tả phụ nữ cho nam giới bú tại Ấn Độ và Bangladesh cao hơn gấp 4 lần tại bất cứ nước nào khác trên thế giới. Dĩ nhiên tôi sẽ không bao giờ nghĩ là có chuyện đó trước khi xem dữ liệu.

Hơn nữa, trong khi việc nam giới bị ám ảnh bởi kích cỡ dương vật của họ có thể không đáng ngạc nhiên lắm, sự bất an lớn nhất về cơ thể ở phụ nữ, theo Google, lại thực sự bất ngờ. Theo dữ liệu mới này, nếu nam lo lắng về kích cỡ dương vật, thì tương đương với đó, nữ lo lắng về việc liệu âm đạo có bốc mùi hay không. Phụ nữ thực hiện các tìm kiếm thể hiện sự lo lắng về cơ quan sinh dục của họ gần như ngang bằng với nam giới. Và nỗi lo lớn nhất ở phụ nữ là mùi—và muốn tìm hiểu xem có thể cải thiện tình hình như thế nào. Dĩ nhiên tôi không biết chuyện đó trước khi xem dữ liệu.

Đôi khi dữ liệu mới tiết lộ những khác biệt văn hóa mà tôi chưa bao giờ suy nghĩ đến. Một ví dụ: Đàn ông khắp thế giới phản ứng bằng những cách rất khác nhau với các bà vợ đang mang thai. Tại Mexico, các tìm kiếm nhiều nhất về "vợ bầu của tôi" gồm có "frases de amor para mi esposa embarazada" (những lời yêu thương gởi đến bà bầu của tôi) và "poemas para mi esposa embarazada" (những bài thơ dành cho bà bầu của tôi). Tại Mĩ, các tìm kiếm nhiều nhất gồm có "my wife is pregnant now what" (vợ tôi có bầu giờ sao đây) và "my wife is pregnant what do I do" (vợ tôi có bầu tôi làm gì đây).

Nhưng quyển sách này không chỉ là một bộ sưu tập các sự thật kì lạ hoặc các nghiên cứu riêng biệt, dù sẽ có nhiều thứ như thế. Bởi vì đây là các phương pháp quá mới mẻ và chắc chắn sẽ ngày càng mạnh hơn, tôi sẽ trình bày một số ý tưởng về cách vận hành của các phương pháp này, cũng như điều khiến đây trở thành các phương pháp mang tính đột phá. Tôi cũng sẽ thừa nhận những giới hạn của Dữ Liệu Lớn trong các phần sau.

Một phần nhiệt huyết trao cho tiềm năng của cuộc cách mạng dữ liệu đã bị đặt sai chỗ. Hầu hết những người hâm mộ Dữ Liệu Lớn không ngớt nói về chuyện các bộ dữ liệu này có thể trở nên rộng lớn ra sao. Sự ám ảnh với kích cỡ dữ liệu không phải là chuyện gì mới. Trước thời Google, Amazon, và Facebook, trước khi cụm từ "Big Data" tồn tại, đã có một hội nghị tổ chức tại Dallas, Texas, bàn về "Các bộ dữ liệu lớn và phức tạp." Jerry Friedman, giáo sư thống kê Đại học Stanford và là đồng nghiệp của tôi khi làm việc tại Google, còn nhớ hội nghị năm 1977 đó. Một nhà thống kê nổi tiếng đứng lên phát biểu. Ông giải thích rằng ông đã tích lũy được tận 5 gigabyte dữ liệu khủng. Nhà thống kê nổi tiếng tiếp theo đứng lên phát biểu. Ông bắt đầu, "Diễn giả vừa rồi đã có hàng gigabyte. Cái đó không là gì hết. Tôi có hàng terabyte." Điểm nhấn của cuộc hội thảo, nói cách khác, là ở lượng thông tin, chứ không phải việc có thể làm với lượng thông tin ấy, cũng chẳng phải về các câu hỏi mà ta sẽ trả lời được nhờ nó. "Tôi thấy thời đó thật buồn cười," Friedman nói, "cái mà người ta khoe là độ lớn của bộ dữ liệu. Chuyện đó bây giờ vẫn còn xảy ra đấy."

Quá nhiều nhà khoa học dữ liệu ngày nay đang tích lũy các bộ dữ liệu khổng lồ nhưng lại chỉ cho ta biết những điều vặt vãnh—ví dụ như đội Knicks rất nổi tiếng ở New York chẳng hạn. Quá nhiều doanh nghiệp đang chết đuối trong mớ dữ liệu khổng lồ. Họ có rất nhiều terabyte nhưng lại chẳng có bao nhiêu hiểu biết có giá trị. Kích cỡ của một bộ dữ liệu, tôi nghĩ, thường được đánh giá quá cao. Có một lời giải thích tinh tế nhưng quan trọng cho điều này: Ảnh hưởng càng lớn, số các quan sát cần thiết để thấy nó càng ít. Ta chỉ cần đụng cái bếp lò nóng một lần là biết nó nguy hiểm ngay. Ta có thể cần phải uống cà phê hàng ngàn lần để quyết định xem nó có khuynh hướng làm ta đau đầu không. Bài học nào quan trọng hơn? Rõ ràng là cái bếp lò nóng: Do cường độ tác động quá lớn nên xuất hiện rất nhanh, với rất ít dữ liệu.

Thực vậy, các công ty Dữ Liệu Lớn thông minh nhất thường cắt giảm dữ liệu của họ. Tại Google, các quyết định lớn đều chỉ dựa trên một mẫu nhỏ xíu trong toàn bộ dữ liệu. Ta không phải lúc nào cũng cần cả tấn dữ liệu để tìm ra các hiểu biết quan trọng. Ta cần dữ liệu đúng. Một lí do

chính khiến các tìm kiếm Google rất có giá trị không phải là do có rất nhiều tìm kiếm, mà chính là do người ta quá thành thật trong các tìm kiếm ấy. Người ta nói dối với bạn bè, người yêu, bác sĩ, cuộc khảo sát, và với cả chính mình. Nhưng trên Google, họ có thể chia sẻ thông tin khó nói, chẳng hạn, về hôn nhân không tình dục, các vấn đề sức khỏe tâm thần, sự bất an, và sự thù địch với người da đen.

Quan trọng nhất, để trích xuất được hiểu biết từ Dữ Liệu Lớn, ta phải hỏi những câu hỏi đúng. Giống như ta không thể chĩa đại một cái kính viễn vọng lên bầu trời đêm và bảo nó hãy tự khám phá Diêm Vương Tinh, ta cũng không thể tải về cả đống dữ liệu và bảo nó phát hiện các bí mật về bản chất con người giùm mình được. Ta phải nhìn vào những nơi hứa hẹn—kiểu như các tìm kiếm Google bắt đầu với "chồng tôi muốn..." tại Ấn Độ chẳng hạn.

Quyển sách này sẽ chỉ ra cách tốt nhất để tận dụng Dữ Liệu Lớn và giải thích chi tiết tại sao nó lại mạnh mẽ đến vậy. Dọc theo cuộc hành trình, bạn cũng sẽ biết về những điều tôi và những người khác đã phát hiện bằng dữ liệu, bao gồm:

- Bao nhiêu nam giới là đồng tính?
- Quảng cáo có hiệu quả không?
- Tại sao American Pharoah lại là một con ngựa đua tuyệt vời?
- Truyền thông có thiên vị không?
- Những lời nói hớ (Freudian slip) có thành thật không?
- Ai là người gian lận thuế?
- Đi học đại học ở đâu có quan trọng không?
- Bạn có thể thắng thị trường chứng khoán không?
- Nuôi con ở đâu là tốt nhất?
- Điều gì khiến một câu chuyện lan truyền?
- Bạn nên nói về điều gì ngày hẹn hò đầu tiên nếu bạn muốn gặp lần thứ hai?

...và nhiều, nhiều nữa.

---

Nhưng trước khi đến với tất cả các thứ đó, chúng ta cần thảo luận một câu hỏi căn bản hơn: Tại sao chúng ta lại cần dữ liệu? Và với câu hỏi đó, tôi xin được giới thiệu bà ngoại tôi.

# I

# Dữ liệu, lớn và nhỏ

# CHƯƠNG 1

## Trực giác sai lầm

Nếu bạn 33 tuổi và đã dự vài kì lễ Tạ ơn liên tục mà vẫn không có người yêu, chủ đề chọn bạn đời chắc chắn sẽ nổi lên. Và hầu như mọi người đều sẽ có ý kiến.

"Seth cần một cô khùng khùng, giống nó," chị tôi nói.

"Chị khùng thì có! Ảnh cần một cô bình thường, để giữ cân bằng cho ảnh chứ," em trai tôi nói.

"Seth không có khùng," mẹ tôi nói.

"Em khùng rồi! Thằng Seth rõ là khùng chứ còn gì nữa," cha tôi nói.

Bỗng nhiên, bà ngoại rụt rè, nói năng nhẹ nhàng của tôi, yên lặng suốt bữa ăn, bây giờ lại lên tiếng. Mấy giọng nói New York ồn ào im bặt, mọi con mắt dồn về bà lão nhỏ nhắn có mái tóc vàng cắt ngắn và vẫn còn mang chất giọng Đông Âu. "Seth, cháu cần một cô gái dễ thương. Không quá đẹp. Thật thông minh. Tốt với mọi người. Năng hoạt động xã hội một chút, để cháu có cơ hội tham gia các hoạt động. Có óc hài hước, vì cháu rất có óc hài hước."

Tại sao lời khuyên của bà lại được gia đình tôi chú ý và kính trọng như thế? Dễ hiểu thôi, bà ngoại 88 tuổi của tôi đã chứng kiến nhiều thứ hơn tất cả những người khác tại bàn ăn này. Bà đã quan sát biết bao nhiêu cuộc hôn nhân, êm đẹp cũng nhiều, không êm đẹp cũng nhiều. Và suốt mấy mươi năm cuộc đời, bà đã tổng hợp được một danh mục

những phẩm chất tạo nên các mối quan hệ thành công. Tại bàn ăn lễ Tạ ơn đó, với vấn đề đó, bà ngoại tôi có số điểm dữ liệu lớn nhất. Bà ngoại tôi là Dữ Liệu Lớn.

Trong quyển sách này, tôi muốn làm sáng tỏ khoa học dữ liệu. Dù muốn hay không, dữ liệu đang đóng một vai trò ngày càng quan trọng trong suốt cuộc sống của chúng ta—và vai trò của nó sẽ ngày càng lớn. Báo chí bây giờ có những mục chuyên viết về dữ liệu. Các công ty có nhóm chuyên phân tích dữ liệu. Nhà đầu tư cấp cho các start-up hàng chục triệu đô la nếu họ có thể chứa nhiều dữ liệu hơn. Ngay cả nếu bạn chưa bao giờ học cách chạy hồi quy hoặc tính khoảng tin cậy, bạn cũng sẽ gặp nhiều dữ liệu—trong các trang bạn đọc, các cuộc họp kinh doanh bạn dự, hay trong cuộc tán gẫu bạn nghe bên các bình nước văn phòng.

Nhiều người lo lắng về sự phát triển này. Họ bị dữ liệu đe dọa, dễ dàng bị lạc lối và mơ hồ trong thế giới các con số. Họ nghĩ rằng hiểu biết định lượng về thế giới là dành cho một vài người phi thường mạnh não trái, không phải dành cho họ. Ngay khi gặp các con số, họ sẵn sàng lật qua trang, kết thúc cuộc họp, hoặc thay đổi chủ đề nói chuyện.

Nhưng tôi đã trải qua 10 năm trong ngành phân tích dữ liệu và may mắn cùng làm việc với nhiều chuyên gia đầu ngành. Một trong những bài học quan trọng nhất tôi đã học được là đây: Khoa học dữ liệu xịn lại ít phức tạp hơn người ta vẫn tưởng. Thực vậy, khoa học dữ liệu thuộc loại xịn nhất lại bất ngờ chứa đựng tính trực giác, bản năng.

Điều gì khiến khoa học dữ liệu có tính trực giác? Tự cốt lõi, khoa học dữ liệu liên quan đến việc phát hiện các mô hình và dự báo biến này sẽ ảnh hưởng đến biến kia như thế nào. Chúng ta thực hiện quá trình này suốt.

Thử tưởng tượng cách bà ngoại khuyên tôi về cách tìm người phù hợp. Bà sử dụng cơ sở dữ liệu lớn các mối quan hệ mà trí não bà đã lưu giữ suốt gần một thế kỉ cuộc đời—trong các câu chuyện bà đã nghe từ gia đình, bạn bè, người quen. Bà giới hạn phân tích của mình ở một mẫu gồm các mối quan hệ, trong đó người đàn ông có nhiều phẩm chất mà tôi có—tính khí nhạy cảm, có xu hướng tự tách biệt, có óc hài hước. Bà

nhắm vào các phẩm chất chính của người phụ nữ—cô tốt như thế nào, cô thông minh ra sao, cô đẹp cỡ nào. Bà liên hệ các phẩm chất chính này của người nữ với một tính chất chính của mối quan hệ: mối quan hệ đó là *tốt* hay *xấu*. Cuối cùng, bà thông báo kết quả. Nói cách khác, bà phát hiện các mô hình và dự báo biến này sẽ ảnh hưởng đến biến kia như thế nào. Bà ngoại là một nhà khoa học dữ liệu.

Bạn cũng là một nhà khoa học dữ liệu. Khi còn nhỏ, bạn để ý rằng khi bạn khóc, mẹ chú ý tới bạn. Đó là khoa học dữ liệu. Khi đến tuổi trưởng thành, bạn để ý rằng nếu bạn phàn nàn quá nhiều, người ta ít muốn ở gần bạn hơn. Đó cũng là khoa học dữ liệu. Bạn để ý thấy, khi người ta ít ở gần bạn hơn, bạn ít vui hơn. Khi bạn ít vui hơn, bạn ít thân thiện hơn. Khi bạn ít thân thiện hơn, người ta lại ít muốn ở gần bạn hơn nữa. Khoa học dữ liệu. Khoa học dữ liệu. Khoa học dữ liệu.

Bởi vì khoa học dữ liệu rất tự nhiên, các nghiên cứu Dữ Liệu Lớn có giá trị nhất có thể được hiểu bởi hầu như bất cứ một người nhanh trí nào. Nếu bạn không thể hiểu một nghiên cứu, vấn đề có thể là do nghiên cứu đó, chứ không phải do bạn.

Khoa học dữ liệu loại xịn mang tính trực giác à, bằng chứng đâu? Gần đây tôi tình cờ gặp một trong những nghiên cứu có thể nói là quan trọng nhất được thực hiện trong mấy năm gần đây. Đó cũng là một trong những nghiên cứu có tính trực giác, bản năng nhất mà tôi từng thấy. Tôi muốn bạn nghĩ không chỉ về tầm quan trọng của nghiên cứu—mà còn về mức độ tự nhiên và mức độ giống bà ngoại tôi nữa.

Nghiên cứu này là của một nhóm các nhà nghiên cứu từ Đại học Columbia và Microsoft. Nhóm muốn tìm hiểu xem các triệu chứng nào sẽ dự báo ung thư tụy. Căn bệnh này có tỉ lệ sống sau 5 năm rất thấp—chỉ khoảng 3%—nhưng việc tầm soát sớm có thể gấp đôi cơ hội cho bệnh nhân.

Phương pháp của các nhà nghiên cứu thế nào? Họ sử dụng dữ liệu từ hàng chục ngàn người dùng nặc danh từ Bing, máy tìm kiếm của Microsoft. Họ mã hóa một người dùng vừa được chẩn đoán ung thư tụy dựa trên các tìm kiếm không thể nhầm lẫn, ví dụ như "vừa được chẩn

đoán ung thư tụy" hoặc "Tôi được bảo là tôi bị ung thư tụy, chuyện gì sẽ xảy ra."

Kế đến, các nhà nghiên cứu xem các tìm kiếm về triệu chứng sức khỏe. Họ so sánh số nhỏ người dùng sau đó được chẩn đoán ung thư tụy với những người không bị. Nói cách khác, các triệu chứng nào dự báo rằng, trong vài tuần hoặc vài tháng tới, một người dùng sẽ được chẩn đoán là ung thư tụy?

Kết quả thật ấn tượng. Việc tìm kiếm chứng đau lưng và sau đó là triệu chứng vàng da hóa ra chính là dấu hiệu ung thư tụy; việc tìm kiếm chỉ mỗi chứng đau lưng thì không chắc chắn một người có bị ung thư tụy hay không. Tương tự, việc tìm kiếm chứng khó tiêu và sau đó đau bụng là bằng chứng của ung thư tụy, trong khi việc tìm kiếm chỉ mỗi chứng khó tiêu mà không có đau bụng nghĩa là không chắc bị ung thư tụy. Các nhà nghiên cứu có thể xác định 5 đến 15% trường hợp mà hầu như không mắc phải trường hợp dương tính sai (false positive) nào. Tỉ lệ này nghe không có gì to tát cho lắm; nhưng nếu bị ung thư tụy, thì ngay cả 10% khả năng tăng gấp đôi cơ hội sống sót cũng là may mắn lắm rồi.

Bài báo nói rằng chi tiết nghiên cứu này rất khó để những người không phải chuyên gia hiểu đầy đủ. Nó có nhiều biệt ngữ kĩ thuật, chẳng hạn như kiểm định Kolmogorov-Smirnov—ý nghĩa của kiểm định này, tôi phải thú nhận là tôi đã quên. (Kiểm định này là một cách để quyết định xem mô hình có đúng khớp với dữ liệu hay không.)

Tuy nhiên, hãy chú ý tính tự nhiên và trực giác ở mức độ căn bản nhất của nghiên cứu này. Các nhà nghiên cứu nhìn vào một lượng lớn các trường hợp về y khoa và thử kết nối các triệu chứng với một bệnh cụ thể. Bạn có biết còn ai khác dùng hệ thống phương pháp này để tìm hiểu bệnh tình của một người không? Các ông chồng bà vợ, cha mẹ, y tá, và bác sĩ. Dựa trên kinh nghiệm và kiến thức, họ thử kết nối các cơn sốt, đau đầu, chảy mũi, và đau bao tử với những bệnh khác nhau. Nói cách khác, các nhà nghiên cứu ở Columbia và Microsoft đã viết một bài nghiên cứu đột phá bằng cách sử dụng hệ thống phương pháp tự nhiên và rõ ràng mà mọi người hay dùng để chẩn đoán sức khỏe.

Nhưng đợi đã. Hãy đi chậm lại một chút. Nếu hệ thống phương pháp khoa học dữ liệu thường có tính tự nhiên và trực giác như tôi nói, thì điều này sẽ làm phát sinh một câu hỏi căn bản về giá trị của Dữ Liệu Lớn. Nếu con người vốn đã là nhà khoa học dữ liệu tự nhiên, nếu khoa học dữ liệu có tính trực giác, tại sao ta lại cần máy tính và phần mềm thống kê? Tại sao ta cần kiểm định Kolmogorov-Smirnov? Ta không thể chỉ dùng trực giác hay sao? Ta không thể làm như bà ngoại, như y tá và bác sĩ hay sao?

Điều này dẫn đến một cuộc tranh luận dữ dội sau khi *Blink* – quyển sách bán chạy của Malcolm Gladwell—được phát hành, ca tụng sự thần kì của các bản năng trực giác. Gladwell kể về những người chỉ dựa vào trực giác đã có thể nói một bức tượng là giả hay thật; một đấu thủ tennis sẽ đánh trật hay không trước khi chạm banh; một khách hàng sẵn sàng chi trả bao nhiêu. Các anh hùng trong *Blink* không chạy hồi quy; họ không tính khoảng tin cậy; họ không chạy kiểm định Kolmogorov-Smirnov. Thế mà họ đã có những dự báo rất đáng chú ý. Nhiều người đã ủng hộ một cách trực giác quan điểm ủng hộ trực giác của Gladwell: Họ tin trực giác và cảm xúc của họ. Giới hâm mộ *Blink* có thể xem trọng hiểu biết của bà tôi khi khuyên về mối quan hệ mà không có sự trợ giúp của máy tính. Giới hâm mộ *Blink* có thể ít xem trọng các nghiên cứu của tôi hoặc các nghiên cứu khác trong sách này, vì các nghiên cứu ấy lại dùng máy tính. Nếu Dữ Liệu Lớn—loại dùng máy tính, chứ không phải kiểu bà ngoại—là một cuộc cách mạng, nó phải chứng tỏ rằng nó mạnh hơn trực giác đơn thuần—thứ mà Gladwell đã cho thấy là mang lại các kết quả rất đáng chú ý.

Nghiên cứu của Columbia và Microsoft cung cấp một ví dụ rõ ràng cho chuyện khoa học dữ liệu nghiêm ngặt và máy tính có thể dạy ta những thứ mà chỉ mỗi trực giác không thể nào nhìn thấy. Đây cũng là một trường hợp mà kích cỡ bộ dữ liệu là quan trọng. Đôi khi kinh nghiệm không đủ nhiều cho trực giác đơn thuần trích xuất được hiểu biết. Nhiều khả năng là bạn—hoặc bạn thân hay người nhà của bạn—sẽ không nhìn thấy đủ các trường hợp ung thư tụy để rút ra sự khác biệt giữa *không tiêu + đau bụng* với chỉ *không tiêu*. Thực vậy, vì bộ dữ liệu Bing

ngày một lớn hơn, các nhà nghiên cứu sẽ phải chọn ra nhiều mô hình tinh tế hơn trong việc tính thời gian các triệu chứng—cho bệnh này và nhiều bệnh khác—những thứ mà ngay cả bác sĩ có thể cũng không phát hiện ra.

Hơn nữa, mặc dù trực giác cho ta một cảm giác tổng quan rất tốt về cách thế giới vận hành, nó lại thường không chính xác về chi tiết. Ta cần dữ liệu để làm sắc nét bức tranh. Xem xét ảnh hưởng của thời tiết lên tâm trạng, chẳng hạn. Có thể bạn sẽ đoán được rằng người ta có nhiều khả năng cảm thấy u buồn vào một ngày *âm* 10°C hơn một ngày 20°C. Thực tế, điều này đúng. Nhưng có lẽ ta không đoán được sự khác biệt nhiệt độ này có thể tạo ra ảnh hưởng lớn cỡ nào. Tôi khảo sát mối tương quan giữa các tìm kiếm Google về chứng suy nhược của một vùng và một loạt các yếu tố, bao gồm điều kiện kinh tế, trình độ giáo dục, và mức độ đi nhà thờ của vùng đó. Khí hậu mùa đông đánh bật hết tất cả các yếu tố còn lại. Vào các tháng mùa đông, nơi khí hậu ấm áp như ở Honolulu, Hawaii, các tìm kiếm về chứng suy nhược ít hơn 40% so với nơi khí hậu lạnh, như Chicago, Illinois. Chính xác ảnh hưởng này có ý nghĩa như thế nào? Một người lạc quan về hiệu quả của thuốc chống suy nhược sẽ thấy rằng các thuốc hiệu quả nhất làm giảm tỉ lệ suy nhược chỉ khoảng 20%. Theo các con số từ Google, chuyển nhà từ Chicago tới Honolulu sẽ có hiệu quả ít nhất là gấp đôi việc dùng thuốc trị chứng suy nhược mùa đông.[1]

Đôi lúc trực giác mà không được dẫn dắt bởi phân tích máy tính cẩn thận có thể sai bét. Ta có thể bị che mắt bởi kinh nghiệm và thành kiến của chính mình. Thực vậy, ngay cả dù bà tôi có thể dùng hàng chục năm kinh nghiệm để đưa ra lời khuyên về mối quan hệ tốt hơn những người khác trong gia đình, bà vẫn có một số góc nhìn đáng ngờ về các yếu tố tạo nên một mối quan hệ trường tồn. Ví dụ, bà thường xuyên nhấn mạnh với tôi tầm quan trọng của việc có bạn bè chung. Bà tin rằng đây là một trong các yếu tố chủ chốt tạo nên sự thành công trong hôn nhân của bà: Bà đã trải qua hầu hết những buổi chiều ấm áp với chồng, tức là ông

[1] Xin nói rõ: Sau khi hoàn thành nghiên cứu này, tôi lại chuyển từ California (miền Nam) tới New York (miền Bắc). Dùng dữ liệu để biết nên làm gì thì dễ. Thực sự làm theo mới là khó.

ngoại tôi, trong mảnh sân sau nhỏ bé ở Queens, New York, ngồi trên đôi ghế xếp và tán gẫu với một nhóm láng giềng thân thiết.

Tuy nhiên, giờ tôi phải cự lại bà ngoại của tôi thôi: Khoa học dữ liệu cho rằng học thuyết của bà ngoại là sai. Một nhóm các nhà khoa học máy tính gần đây đã phân tích bộ dữ liệu lớn nhất từ trước đến nay về các mối quan hệ con người—Facebook. Họ xem xét rất nhiều cặp đôi mà, tại thời điểm nào đó, đã từng có trạng thái "in a relationship" (ở trong một mối quan hệ). Một số cặp trong số này vẫn còn "in a relationship." Một số cặp khác đã chuyển sang trạng thái "single" (độc thân). Các nhà nghiên cứu phát hiện, việc có chung nhóm bạn thân là một chỉ báo rất mạnh cho thấy rằng mối quan hệ ấy sẽ *không* đi đến đâu. Có lẽ việc đi chơi hàng đêm với người yêu cùng với một nhóm nhỏ bạn bè lặp đi lặp lại không phải là điều tốt; các nhóm xã hội tách biệt có thể giúp làm cho các mối quan hệ mạnh hơn lên.

Như có thể thấy, khi loại bỏ máy tính và chỉ xử lí bằng trực giác, trực giác đơn thuần đôi khi có thể làm ta kinh ngạc. Nhưng nó cũng có thể sai lầm lớn. Bà ngoại có thể đã rơi vào một cái bẫy nhận thức: Chúng ta có khuynh hướng phóng đại sự xác đáng (relevance) trong kinh nghiệm của riêng mình. Theo lối nói của các nhà khoa học dữ liệu, ta *đặt trọng số* cho dữ liệu, và ta lại cho quá nhiều trọng số vào một điểm dữ liệu cụ thể: chính bản thân ta.

Bà ngoại quá tập trung vào các cuộc tán gẫu mỗi chiều với ông ngoại và bạn bè nên bà không nghĩ đầy đủ về các cặp đôi khác. Bà quên quan sát kĩ cặp vợ chồng người em rể, họ tám chuyện mỗi đêm với một nhóm nhỏ bạn bè không đổi, nhưng đã choảng nhau thường xuyên rồi cuối cùng cũng li dị đó thôi. Bà quên quan sát kĩ cha mẹ tôi—con gái và con rể của bà. Cha mẹ tôi nhiều đêm hai người đi hai đường—cha tôi thì tới câu lạc bộ nhạc jazz hoặc chơi bóng với bạn của cha, mẹ tôi thì tới nhà hàng hoặc nhà hát với bạn của mẹ; tuy vậy họ vẫn sống hạnh phúc với nhau.

Khi dựa trên trực giác, ta cũng có thể bị lầm lạc bởi xu hướng say mê những điều gây ấn tượng sâu sắc. Ta có khuynh hướng đánh giá quá cao

sự phổ biến của bất cứ thứ gì có thể làm nên một câu chuyện đáng nhớ. Ví dụ, khi được khảo sát, người ta kiên định xếp lốc xoáy là nguyên nhân gây tử vong thường xuyên hơn bệnh suyễn. Thực ra, bệnh suyễn gây tử vong nhiều hơn khoảng 70 lần. Tử vong do bệnh suyễn không nổi bật—và do đó không được lên bản tin. Tử vong do lốc xoáy thì lại là tin tức rất ấn tượng.

Nói cách khác, ta thường sai lầm về cách thế giới vận hành khi chỉ dựa vào những gì ta nghe hoặc trải nghiệm. Mặc dù phương pháp của khoa học dữ liệu thường có tính trực giác, kết quả lại thường phản trực giác. Khoa học dữ liệu thực hiện một tiến trình tự nhiên và trực giác—xác định các mô hình và hiểu chúng—rồi tiêm huyết thanh sự thật vào đó để chỉ cho ta thấy rằng thế giới vận hành theo cách hoàn toàn khác với cách ta đã hình dung. Đó là những gì đã diễn ra khi tôi nghiên cứu các yếu tố dự báo thành công trong môn bóng rổ.

***

Khi còn bé, tôi có một giấc mơ và chỉ một mà thôi: Tôi muốn lớn lên làm một nhà kinh tế kiêm nhà khoa học dữ liệu. Không. Tôi chỉ đang nói dối thôi. Hồi đó, tôi rất muốn làm một cầu thủ bóng rổ chuyên nghiệp, để theo bước chân người hùng của tôi, Patrick Ewing, tâm điểm toàn sao của đội New York Knicks.

Đôi khi tôi nghĩ rằng bên trong mỗi nhà khoa học dữ liệu là một đứa con nít đang cố tìm hiểu tại sao các giấc mơ tuổi thơ của mình lại không thành sự thật. Vậy thì không có gì bất ngờ khi gần đây tôi lại lao vào nghiên cứu xem làm sao để được chơi ở giải bóng rổ nhà nghề NBA. Kết quả nghiên cứu rất đáng ngạc nhiên. Thực vậy, nó lại một lần nữa chỉ ra rằng khoa học dữ liệu có thể thay đổi cách ta nhìn thế giới, và cho thấy các con số có thể phản trực giác đến thế nào.

Câu hỏi cụ thể tôi xem xét là: Xuất thân từ gia đình nghèo hay trung lưu thì có nhiều khả năng chơi ở giải NBA hơn?

Hầu hết mọi người sẽ đoán là gia đình nghèo. Người ta tin rằng lớn lên trong hoàn cảnh khó khăn, kiểu như trong khu chung cư thuộc sở hữu nhà nước với một người mẹ tuổi "teen" đơn thân, sẽ giúp nuôi

dưỡng động cơ cần thiết để đạt trình độ đỉnh cao trong môn thể thao cạnh tranh dữ dội này.

Quan điểm này được nhắc đến bởi William Ellerbee, một huấn luyện viên bóng rổ trường trung học ở Philadelphia, trong một cuộc phỏng vấn với *Sports Illustrated*. "Trẻ em vùng ngoại ô có khuynh hướng chơi bóng vì niềm vui," Ellerbee nói. "Trẻ em nội thành xem bóng rổ là vấn đề sống chết." Rồi xong. Tôi được cả cha lẫn mẹ nuôi dưỡng ở ngoại ô New Jersey. LeBron James, cầu thủ giỏi nhất thế hệ tôi, sinh ra trong một gia đình nghèo, nuôi lớn bởi một người mẹ đơn thân 16 tuổi ở Akron, Ohio.

Thực vậy, một khảo sát Internet tôi thực hiện chỉ ra rằng đa số người Mĩ suy nghĩ giống huấn luyện viên Ellerbee. Tôi đã nghĩ: Hầu hết các cầu thủ NBA lớn lên trong nghèo khó.

Suy nghĩ này có đúng không?

Ta hãy xem dữ liệu. Không có nguồn dữ liệu đầy đủ về kinh tế xã hội học của các cầu thủ NBA. Tuy nhiên, sau một hồi đóng vai thám tử, tận dụng dữ liệu từ cả đống nguồn—basketball-reference.com, ancestry.com, Cục Thống kê Dân số Mĩ, và nhiều nguồn khác—chúng tôi đã có thể hiểu được gia cảnh nào thực sự có ích cho con đường đến với NBA. Nghiên cứu này dùng đa dạng nguồn dữ liệu, lớn có, nhỏ có, trực tuyến có, và ngoại tuyến có. Tuy rất hứng thú với một số nguồn kĩ thuật số mới, một nhà khoa học dữ liệu giỏi cũng sẽ sẵn sàng tham khảo các nguồn kiểu cũ nếu có ích. Cách tốt nhất để có câu trả lời đúng cho một câu hỏi là phối hợp tất cả dữ liệu có sẵn.

Dữ liệu phù hợp đầu tiên là nơi sinh của mọi cầu thủ. Với mỗi hạt trong nước Mĩ,[1] tôi ghi lại số người da đen và da trắng sinh ra trong thập niên 1980. Sau đó tôi ghi lại số người vào được giải NBA. Tôi so sánh thông tin này với thu nhập bình quân hộ gia đình của hạt. Tôi cũng đặt cơ cấu chủng tộc của hạt làm yếu tố đối chứng (control), vì nam da đen có nhiều khả năng vào NBA hơn nam da trắng khoảng 40 lần (và đây có thể là chủ đề cho cả một quyển sách khác).

[1] [ND] Hạt (county) là đơn vị hành chính dưới tiểu bang. Một số nơi còn gọi là quận.

Dữ liệu cho biết rằng ta sẽ có nhiều cơ hội vào NBA hơn đáng kể nếu sinh ra trong một hạt giàu có. Một trẻ da đen sinh ra tại một trong các hạt giàu có nhất nước Mĩ chẳng hạn, có hơn gấp đôi khả năng vào chơi NBA so với một trẻ da đen sinh ra tại một trong những hạt nghèo nhất. Với trẻ da trắng, lợi thế giữa việc sinh ra tại một trong những hạt giàu nhất so với việc sinh ra tại một trong những hạt nghèo nhất là 60%.

Điều này muốn nói, trái với hiểu biết truyền thống, rằng người nghèo thực ra không có mặt nhiều tại NBA. Tuy nhiên, dữ liệu này không hoàn hảo, vì nhiều hạt giàu có ở Mĩ, chẳng hạn như New York County (Manhattan), cũng bao gồm các khu dân cư nghèo, như Harlem. Vậy thì chuyện tuổi thơ khó khăn giúp ta đến với NBA vẫn có khả năng xảy ra. Chúng ta vẫn cần nhiều đầu mối, nhiều dữ liệu hơn nữa.

Thế là tôi đã nghiên cứu gia cảnh của các cầu thủ NBA. Thông tin này được tìm thấy trong các câu chuyện báo chí và trên các mạng xã hội. Hệ thống phương pháp này khá mất thời gian, vì vậy tôi giới hạn chỉ phân tích nhóm 100 cầu thủ NBA người Mĩ gốc Phi sinh vào thập niên 1980 ghi điểm nhiều nhất. Các siêu sao NBA ít có khả năng được sinh ra bởi một người mẹ tuổi teen hoặc người mẹ ngoài giá thú hơn nam da đen trung bình ở Mĩ 30%. Nói cách khác, gia cảnh của các cầu thủ NBA da đen giỏi nhất cũng cho thấy rằng một gia đình sung túc, ấm cúng là lợi thế lớn để thành công.

Ngay cả vậy, dữ liệu nơi sinh cấp hạt hoặc gia cảnh của một mẫu nghiên cứu nhỏ đều không thể cho thông tin hoàn hảo về tuổi thơ của tất cả cầu thủ NBA. Vì vậy, tôi vẫn không hoàn toàn tin rằng các gia đình trung lưu, có đủ cha mẹ, sẽ sinh ra nhiều ngôi sao NBA hơn các gia đình nghèo có cha hoặc mẹ đơn thân. Càng có thêm nhiều dữ liệu càng tốt.

Sau đó, tôi nhớ thêm một điểm dữ liệu nữa có thể cung cấp các đầu mối biết nói về gia cảnh của một người. Trong một bài báo, 2 nhà kinh tế học Roland Fryer và Steven Levitt cho rằng tên của người da đen là một chỉ báo cho gia cảnh kinh tế xã hội của họ. Fryer và Levitt nghiên cứu giấy khai sinh ở California vào thập niên 1980 và phát hiện rằng, trong những người Mĩ gốc Phi, các bà mẹ nghèo, thiếu giáo dục, đơn thân có

khuynh hướng đặt tên con khác với các cặp cha mẹ trung lưu, có giáo dục.

Trẻ em gia cảnh tốt nhiều khả năng được đặt những cái tên thường gặp, như Kevin, Chris, và John. Trẻ em xuất thân gia đình khó khăn trong các khu nhà tập thể nhiều khả năng được đặt những cái tên khá độc đáo, như Knowshon, Uneek, và Breionshay. Trẻ em Mĩ gốc Phi sinh ra nghèo khổ có gần gấp đôi khả năng được đặt tên không giống một trẻ nào sinh cùng năm.

Vậy còn tên của các cầu thủ NBA da đen thì sao? Nghe giống người da đen trung lưu hay nghèo khổ hơn? Khi xem xét cùng thời kì, các cầu thủ NBA sinh tại California chỉ có phân nửa khả năng được cha mẹ đặt cho những cái tên độc đáo so với trung bình nam da đen, một khác biệt có ý nghĩa thống kê.

Bạn có tình cờ quen ai nghĩ rằng NBA là giải dành cho trẻ nghèo không? Hãy bảo anh ta nghe cho kĩ trận đấu sắp tới trên sóng truyền thanh. Bảo anh ta chú ý có bao nhiêu lần Russell dẫn bóng qua mặt Dwight và sau đó cố gắng thảy bóng lọt qua hai cánh tay dang rộng của Josh vào đôi bàn tay đang chờ đợi của Kevin. Nếu NBA thực sự là một giải có nhiều người da đen nghèo, thì lời tường thuật nghe sẽ hoàn toàn khác. Sẽ có rất nhiều người mang những cái tên độc lạ kiểu như LeBron hơn.

Bây giờ, chúng ta đã gom được 3 mẩu bằng chứng khác nhau—nơi sinh (hạt), thân thế các bà mẹ của những người ghi điểm cao nhất, và tên của các cầu thủ. Không nguồn nào là hoàn hảo. Nhưng cả 3 ủng hộ cho cùng một câu chuyện. Địa vị kinh tế xã hội tốt hơn có nghĩa là cơ hội vào được NBA cao hơn. Nói cách khác, hiểu biết truyền thống là sai.

Trong số tất cả những người Mĩ gốc Phi sinh vào thập niên 1980, khoảng 60% có cha hoặc mẹ ngoài giá thú. Nhưng tôi ước tính rằng trong số những người Mĩ gốc Phi sinh vào thập niên đó vào được NBA, đa số đủ cả cha lẫn mẹ. Nói cách khác, NBA chủ yếu không gồm những người có gia cảnh như LeBron James. Có nhiều người như Chris Bosh hơn ở Texas, được cha mẹ đầy đủ nuôi dưỡng, họ gieo ở anh sự yêu

thích các thiết bị điện tử; hoặc như Chris Paul, con trai thứ hai của cặp vợ chồng trung lưu ở Lewisville, North Carolina. Cả gia đình của Chris Paul đã tham gia với anh trong một tập của chương trình *Family Feud* năm 2011.

Mục đích của một nhà khoa học dữ liệu là hiểu thế giới. Một khi ta tìm thấy kết quả phản trực giác, ta có thể dùng thêm khoa học dữ liệu để giúp ta giải thích tại sao thế giới không như ta thấy. Ví dụ, tại sao dân trung lưu tương đối giỏi bóng rổ hơn dân nghèo? Có ít nhất 2 lời giải thích.

Thứ nhất, vì dân nghèo thường thấp lùn hơn. Các học giả từ lâu biết rằng sự chăm sóc và dinh dưỡng thời thơ ấu đóng vai trò rất lớn cho sức khỏe lúc trưởng thành. Đây là lí do một người trung bình ở thế giới phát triển bây giờ cao hơn 4 inch (10.16 cm) so với 150 năm trước. Dữ liệu cho thấy người Mĩ xuất thân từ gia đình nghèo thì thấp hơn, vì chế độ chăm sóc và dinh dưỡng đầu đời kém hơn.

Dữ liệu còn có thể cho ta biết ảnh hưởng của chiều cao trong việc vào được NBA. Rõ ràng, bằng trực giác, bạn có thể thấy chiều cao rất có ích cho một cầu thủ bóng rổ tài năng. Chỉ việc đối chiếu chiều cao của cầu thủ điển hình trên sân với người hâm mộ điển hình đứng xem là rõ. (Cầu thủ NBA trung bình cao 6'7" [~ 2m]; người Mĩ nam trung bình cao 5'9" [~ 1.75m].)

Chiều cao quan trọng tới mức nào? Cầu thủ NBA đôi khi nói dối một chút về chiều cao của họ, và không có bảng phân phối chiều cao hoàn hảo của nam giới Mĩ. Tuy nhiên, sử dụng ước lượng toán học thô dựa trên dự đoán về phân phối chiều cao của nam giới Mĩ cũng như dựa trên số liệu của NBA, ta dễ dàng xác định rằng ảnh hưởng của chiều cao là rất lớn—có lẽ còn nhiều hơn ta đã hình dung. Tôi ước tính rằng mỗi inch cộng thêm vào chiều cao sẽ tăng gấp đôi cơ hội được chơi ở NBA. Và điều này là đúng đối với toàn bộ dải phân phối chiều cao. Một nam cao 5'11" có gấp đôi cơ hội vào được NBA so với một nam cao 5'10". Một nam cao 6'11" có gấp đôi cơ hội vào được NBA so với một nam cao 6'10". Dường như, trong số các nam cao dưới 6 feet (~ 1.83 m), chỉ khoảng 1

trong 2,000,000 người là vào được NBA. Trong số các nam cao trên 7 feet (~ 2.13 m), tôi và những người khác đã ước tính, đâu khoảng 1 trong 5 người vào được NBA.

Bạn sẽ chú ý thấy, dữ liệu làm sáng tỏ tại sao giấc mơ thành ngôi sao bóng rổ của tôi đã trật đường ray. Không phải vì tôi lớn lên ở vùng ngoại ô. Đó là vì tôi chỉ cao có 5'9" và da trắng (chưa kể còn chậm chạp). Tôi còn lười nữa. Sức chịu đựng thì kém, tư thế ném bóng không ra gì, và thỉnh thoảng lại hoảng hốt khi bóng vào tay.

Lí do thứ hai khiến các chàng trai xuất thân hoàn cảnh khó khăn có thể phải rất vất vả để vào được NBA là vì đôi khi họ thiếu một số kĩ năng xã hội. Dùng dữ liệu của hàng ngàn học sinh, các nhà kinh tế học đã phát hiện rằng các gia đình trung lưu, có đủ cha mẹ, nói chung cơ bản là giỏi hơn trong việc nuôi dạy những đứa trẻ đáng tin, kỉ luật, kiên trì, tập trung, và có óc tổ chức.

Vậy các kĩ năng xã hội kém làm trật đường ray một sự nghiệp bóng rổ đầy hứa hẹn như thế nào?

Ta hãy xem câu chuyện Doug Wrenn, một trong những tài năng bóng rổ có triển vọng nhất thập niên 1990. Huấn luyện viên trường cậu, Jim Calhoun tại Đại học Connecticut, người đã huấn luyện toàn sao NBA tương lai, nói Wrenn nhảy cao nhất trong các cầu thủ ông từng làm việc chung. Nhưng Wrenn có một tuổi thơ đầy khó khăn. Cậu được nuôi dưỡng bởi một bà mẹ đơn thân ở Blood Alley, một trong các khu dân cư khó khăn nhất Seattle. Tại Connecticut, cậu thường xuyên đụng chạm với những người xung quanh. Cậu hay chế nhạo các cầu thủ khác, nghi ngờ huấn luyện viên, và mặc đồ rộng lùng thùng vi phạm nội quy đội bóng. Cậu còn dính tới pháp luật—chôm giày của một cửa hàng và phản ứng với cảnh sát. Hết chịu nổi, Calhoun đá cậu ra khỏi đội.

Wrenn có cơ hội thứ hai tại Đại học Washington. Nhưng ở đó, cậu lại trật đường ray tiếp vì thiếu khả năng hòa hợp với mọi người. Cậu cãi với huấn luyện viên về thời gian thi đấu và cách ném bóng. Kết cục, cậu lại bị loại khỏi đội này. Wrenn thành cầu thủ bị NBA chê, bị đá vòng quanh các giải thấp hơn, phải về ở với mẹ, và cuối cùng phải vô tù nằm vì tội

hành hung. "Sự nghiệp của tôi chấm hết," Wrenn nói với *Seattle Times* năm 2009. "Ước mơ của tôi, khát vọng của tôi chấm hết. Doug Wrenn đã chết. Cầu thủ bóng rổ đó, chàng trai đó đã chết. Chấm hết rồi." Wrenn có đủ tài năng để không chỉ chơi ở NBA, mà còn có thể là một cầu thủ vĩ đại, thậm chí huyền thoại. Nhưng cậu không bao giờ phát triển khí chất phù hợp để trụ lại được bất kì đâu, thậm chí là ngay cả với một đội trường đại học. Giá mà có một tuổi thơ ổn định, Wrenn có thể đã là Michael Jordan tiếp theo rồi.

Michael Jordan, dĩ nhiên, cũng nhảy cực kì cao. Cộng thêm một cái tôi lớn và tính cạnh tranh dữ dội—một cá tính đôi khi không khác gì Wrenn. Jordan đã từng là một đứa trẻ khó dạy. Ở tuổi 12, anh đã bị đuổi học vì đánh lộn. Nhưng ít nhất anh có một thứ mà Wrenn thiếu: sự giáo dục ở một gia đình trung lưu, ổn định. Cha của anh là nhân viên giám sát thiết bị ở General Electric, mẹ làm ở ngân hàng. Và hai người đã giúp anh định hướng nghề nghiệp.

Thực vậy, cuộc đời Jordan đầy những câu chuyện về việc gia đình dẫn dắt anh tránh khỏi những cái bẫy mà một tài năng lớn đầy tính cạnh tranh có thể rơi vào. Sau khi Jordan bị đuổi học, mẹ anh phản ứng bằng cách đưa anh theo bà làm việc. Anh không được phép rời xe mà phải ngồi yên trong chỗ đậu xe để đọc sách. Sau khi anh được đội Chicago Bulls tuyển, cha mẹ và các anh chị em thay phiên đến thăm để bảo đảm là anh tránh xa những cám dỗ thường đến cùng tiền tài và danh vọng.

Sự nghiệp của Jordan không kết thúc bằng một câu than thở không ai để ý đến trên tờ *Seattle Times* như Wrenn. Nó kết thúc với một diễn văn được hàng triệu người theo dõi khi anh được đưa vào danh sách Basketball Hall of Fame. Trong diễn văn của mình, Jordan nói anh cố gắng tiếp tục "tập trung vào những điều tốt đẹp về cuộc sống—biết mọi người nhận thức về mình thế nào, mình tôn trọng họ thế nào [...] mình được công chúng nhận biết ra sao. Dừng lại và nghĩ về những điều mình làm. Và tất cả điều đó đều nhờ cha mẹ tôi."

Dữ liệu cho ta biết Jordan hoàn toàn đúng khi cảm ơn gia đình trung lưu có đầy đủ cha mẹ của mình. Dữ liệu cho ta biết rằng trong các gia

đình ít may mắn hơn, trong các cộng đồng ít may mắn hơn, có những tài năng đẳng cấp NBA nhưng không thể có mặt ở NBA. Những người này có gene, có tham vọng, nhưng không bao giờ phát triển được khí chất phù hợp để trở thành các siêu sao bóng rổ.

Và không, dù ta có nghĩ sao đi nữa, chuyện ở trong môi trường sống tuyệt vọng—khi mà bóng rổ là "vấn đề sống chết"—cũng chả ích gì đâu. Các câu chuyện như của Doug Wrenn có thể minh họa điều này. Và dữ liệu đã chứng minh nó.

Tháng 6/2013, LeBron James được phỏng vấn trên truyền hình sau khi giành chức vô địch NBA lần thứ hai. (Đến nay đã là lần thứ ba.) "Tôi là LeBron James," anh nói. "Đến từ Akron, Ohio. Từ nội ô thành phố. Không ai nghĩ tôi có thể ở được đây." Twitter và các mạng xã hội khác bùng nổ lời phê bình. Làm sao mà một tài năng tuyệt đỉnh, được phát hiện là tương lai của bóng rổ từ lúc còn ở độ tuổi trẻ đến mức vô lí như thế, lại có thể tuyên bố mình là kẻ ngồi chiếu dưới như vậy được? Thực ra, những người có hoàn cảnh khó khăn, dù năng lực thể thao có tốt đến thế nào đi nữa, đều gặp nhiều khó khăn hơn bình thường. Nói cách khác, thành công của James thậm chí còn phi thường hơn ta tưởng. Dữ liệu cũng đã chứng minh điều đó.

# II

# Sức mạnh của Dữ Liệu Lớn

# CHƯƠNG 2

## Freud có đúng không?

Gần đây tôi thấy từ "pedestrian" (khách bộ hành) được viết là "penistrian" (khách "bộ hạ") Bạn phát hiện ra chứ? Là "*penis*trian" thay vì "pedestrian." Tôi thấy từ đó trong một bộ dữ liệu lớn về lỗi đánh máy của mọi người. Một người thấy ai đó đi bộ và viết từ "penis." Phải có ý nghĩa gì đó, đúng không?

Gần đây tôi biết chuyện có một người đàn ông mơ vừa ăn một trái chuối vừa bước đến bàn thờ làm lễ cưới vợ. Tôi thấy nó trong một bộ dữ liệu lớn về các giấc mơ người ta ghi lại trên một ứng dụng. Một nam tưởng tượng cưới một nữ khi đang ăn một thứ có hình dương vật. Cũng phải có ý nghĩa gì đó, đúng không?

Sigmund Freud có đúng không? Từ khi các lí thuyết của ông lần đầu được mọi người chú ý, câu trả lời thành thật nhất cho câu hỏi này là một cái nhún vai.[1] Karl Popper, triết gia người Anh gốc Áo, nói rõ nhất về vụ này. Popper nói rằng các lí thuyết của Freud không thể bị chứng minh là sai. Không có cách để kiểm chứng xem ý tưởng ấy là đúng hay sai.

Freud có thể sẽ nói người viết từ "penistrian" đang tiết lộ một ham muốn tình dục có thể bị đè nén. Người đó có thể phản ứng là mình chả có ý gì cả; rằng cô đơn giản chỉ mắc lỗi đánh máy, và hoàn toàn có thể

[1] [ND] Freud nổi tiếng với thuyết cho rằng, những lần lỡ lời hay buột miệng (Freudian slip) là sự bộc lộ những điều bị đè nén trong tiềm thức. Ở đây, những từ ngữ viết nhầm có liên quan đến tình dục, theo cách giải thích của Freud, thể hiện những ham muốn bị đè nén.

đánh nhầm thành "pedaltrian" chẳng hạn. Đó trở thành một trường hợp ông cãi qua bà cãi lại không hồi kết. Freud có thể nói quý ông mơ ăn một trái chuối vào ngày cưới đang thầm nghĩ đến một cái dương vật, tiết lộ ước muốn thực sự là cưới một người nam chứ không phải một người nữ. Quý ông đó có thể nói ông chỉ tình cờ mơ đến một trái chuối thôi. Ông hoàn toàn có thể tình cờ mơ đang ăn một trái táo khi làm lễ. Trường hợp này cũng sẽ chỉ cãi qua cãi lại thôi. Không có cách nào để kiểm chứng lí thuyết của Freud trong thực tế.

Đến bây giờ, mọi chuyện đã khác.

Khoa học dữ liệu giúp cho nhiều phần trong thuyết của Freud có thể kiểm chứng được—và nó đã đưa nhiều lí thuyết của ông vào bài kiểm tra thực tế. Ta hãy bắt đầu với biểu tượng hình dương vật trong các giấc mơ. Dùng một bộ dữ liệu lớn các giấc mơ được ghi lại, ta có thể thấy tần suất xuất hiện của các thứ có hình giống dương vật. Thực phẩm là nơi thích hợp để tập trung nghiên cứu. Nó xuất hiện trong nhiều giấc mơ, và nhiều thực phẩm có hình dương vật—chuối, dưa leo, xúc xích... Thế thì ta có thể đo lường các yếu tố sẽ khiến ta mơ nhiều về một số thực phẩm nhất định—như mức độ thường xuyên ăn thực phẩm ấy, mức độ ưa thích loại thực phẩm ấy, và vâng, ta cũng đưa vào cả yếu tố nó có giống dương vật trong tự nhiên hay không nữa.

Ta có thể kiểm chứng 2 loại thực phẩm phổ biến tương đương nhau, nhưng một loại giống dương vật, loại kia thì không, và xem tần suất xuất hiện trong mơ của 2 loại này có giống nhau không. Nếu các thực phẩm hình dương vật không được mơ đến nhiều hơn các thực phẩm khác, thế thì hình dạng giống dương vật không phải là yếu tố quan trọng trong các giấc mơ. Nhờ Dữ Liệu Lớn, phần lí thuyết này của Freud hoàn toàn có thể kiểm chứng.

Tôi nhận dữ liệu từ Shadow, một ứng dụng yêu cầu người dùng ghi lại giấc mơ. Tôi mã hóa các thực phẩm được đề cập trong hàng chục ngàn giấc mơ.

Về tổng thể, điều gì khiến ta mơ đến thực phẩm? Nhân tố dự báo mạnh mẽ nhất là tần suất ta tiêu thụ các thực phẩm đó. Chất được mơ

đến nhiều nhất là nước. Các thực phẩm top 20 gồm thịt gà, bánh mì, sandwich, và cơm—tất cả đều là phi-Freud (không mang tính gợi nhắc gì đến tình dục).

Nhân tố dự báo thứ hai là mức độ ưa thích loại thực phẩm ấy. Hai thực phẩm ta mơ đến thường nhất là sô cô la và bánh pizza, rõ ràng là phi-Freud, nhưng quan trọng là đều rất hấp dẫn nữa.

Vậy các thực phẩm hình dương vật thì sao? Chúng có lẻn vào các giấc mơ của ta với tần suất lớn bất ngờ không? Không.

Chuối là trái cây thông thường thứ nhì xuất hiện trong các giấc mơ. Nhưng đó cũng là trái cây thông dụng thứ nhì. Vậy ta không cần nhờ Freud giải thích việc ta thường mơ về chuối. Dưa leo là loại rau củ quả thông thường thứ 7 xuất hiện trong các giấc mơ. Đó cũng là loại rau củ quả thông dụng thứ 7. Vậy một lần nữa hình dáng thực phẩm không phải là thứ giúp giải thích sự hiện diện của nó trong trí não ta khi ta ngủ. Bánh mì xúc xích được mơ với tần suất thấp hơn rất nhiều so với bánh mì hamburger. Điều này là đúng ngay cả khi ta tính luôn việc người ta ăn nhiều hamburger hơn bánh mì xúc xích.

Về tổng thể, khi phân tích hồi quy (một phương pháp cho phép ta phân tách ảnh hưởng của nhiều nhân tố tác động khác nhau) tất cả trái cây và rau củ, tôi phát hiện rằng việc có hình dạng giống dương vật không giúp thực phẩm có thêm khả năng xuất hiện trong mơ. Lý thuyết này của Freud có thể được kiểm chứng—và theo dữ liệu của tôi, nó sai.

Tiếp đến, hãy xem xét các lời nói buột miệng (Freudian slip). Freud đưa ra giả thuyết rằng chúng ta dùng các lỗi—nói nhầm hay viết nhầm—để tiết lộ ham muốn tiềm thức của mình, thường là ham muốn tình dục. Ta có thể dùng Dữ Liệu Lớn để kiểm chứng điều này không? Đây là một cách: Xem thử các lỗi ấy—các lời nói buột miệng—có nghiêng về hướng tục tĩu không. Nếu ham muốn tình dục bị chôn giấu được thể hiện trong các lời nói buột miệng, phải có một số lỗi vượt trội có các từ như "penis," "cock," and "sex."

Vậy số lỗi liên quan đến tình dục có nhiều bất thường không?

Thế là tôi nghiên cứu một bộ dữ liệu gồm hơn 40,000 lỗi đánh máy do các nhà nghiên cứu Microsoft thu thập. Bộ dữ liệu gồm các lỗi mà người ta phạm phải nhưng rồi sửa lại ngay. Trong hàng chục ngàn lỗi này, có nhiều cá nhân đánh máy nhầm các lỗi liên quan đến tình dục. Có trường hợp sai "penistrian" như đã nói. Cũng có người gõ "sexurity" thay vì "security" và "cocks" thay vì "rocks." Nhưng cũng có nhiều lỗi vô hại. Người ta viết "*p*indows" và "*f*egetables," "after*m*oons" và "refri*d*erators."

Để kiểm tra điều này, đầu tiên tôi dùng bộ dữ liệu Microsoft để lập mô hình tần suất người ta viết sai các mẫu tự cụ thể. Tôi tính tần suất người ta nhầm *t* thành *s*, *g* thành *h*. Sau đó tôi tạo ra một chương trình máy tính chuyên đánh máy nhầm tương tự với kiểu nhầm lẫn của con người. Hãy tạm gọi nó là Error Bot. Error Bot thay *t* bằng *s* với cùng tần suất với người dùng trong nghiên cứu của Microsoft. Nó thay *g* bằng *h* cũng theo kiểu như vậy. Và cứ như thế. Tôi chạy chương trình theo giống các từ người ta đã sai trong nghiên cứu Microsoft. Nói cách khác, Error Bot cố gắng đánh vần đúng "pedestrian" và "rocks," "windows" và "refrigerator;" nhưng nó sẽ nhầm *r* thành *t* với cùng tần suất của người bình thường, và viết sai, "tocks" thay vì "rocks" chẳng hạn. Nó cũng sẽ nhầm *r* thành *c* thường xuyên như con người, và viết nhầm thành "cocks."

Vậy chúng ta biết được gì từ việc so sánh Error Bot với những người bất cẩn bình thường? Sau khi thực hiện vài triệu lỗi theo kiểu đặt sai mẫu tự mô phỏng theo cách sai của con người, Error Bot đã phạm vô số lỗi có liên quan đến tình dục. Nó sai chính tả kiểu "seashell" thành "*sex*shell," "lipstick" thành "lips*dick*," và "luckiest" thành "*fuck*iest," cũng như nhiều lỗi tương tự khác. Mấu chốt là Error Bot (dĩ nhiên là không có tiềm thức) cũng có thể phạm các lỗi được xem là liên quan đến tình dục y chang như người thật. Dân khoa học xã hội chúng tôi thích nói rằng vấn đề này "cần có thêm nhiều nghiên cứu," nghĩa là các lỗi đánh máy thường được xem là có liên quan đến tình dục mà con người phạm phải chẳng qua chỉ là tình cờ mà thôi.

Nói cách khác, việc người ta phạm các lỗi như "penistrian," "sexurity," và "cocks," không nhất thiết phải liên quan đến những thứ tục tĩu. Các lỗi gõ nhầm này có thể được giải thích hoàn toàn bằng tần suất lỗi đánh máy. Người ta phạm rất nhiều lỗi. Và nếu phạm lỗi đủ nhiều, thì rồi cũng có lúc ta nói các thứ như "lipsdick," "fuckiest," và "penistrian." Nếu một con khỉ gõ phím đủ lâu, cuối cùng nó cũng sẽ viết được "to be or not to be" như Shakespeare. Nếu một người gõ đủ lâu, cuối cùng cô ta cũng sẽ gõ nhầm thành "penistrian."

Lí thuyết của Freud cho rằng các lỗi tiết lộ các ham muốn tiềm thức thực ra có thể kiểm chứng—và theo phân tích dữ liệu của tôi, nó sai.

Dữ Liệu Lớn cho ta biết một trái chuối luôn chỉ là một trái chuối, và "penistrian" chỉ là "pedestrian" viết sai chính tả.

* * *

Vậy Freud có hoàn toàn sai? Không hẳn. Lần đầu truy cập dữ liệu P***hub, tôi phát hiện ở đó một điều khiến tôi ít nhất cũng hơi nghiêng về phía Freud. Thực vậy, đây là một trong số những điều đáng ngạc nhiên nhất mà tôi đã phát hiện được khi nghiên cứu dữ liệu: Một con số thật kinh hoàng những người vào các trang khiêu dâm lại đang tìm kiếm các nội dung loạn luân.

Trong số 100 tìm kiếm hàng đầu của nam giới trên P***hub, một trong những trang khiêu dâm phổ biến nhất, có 16 là về chủ đề loạn luân. Xin cảnh báo là các cụm từ sau sẽ hơi "nặng đô" một chút *(người dịch xin phép không chuyển ngữ)*: Các tìm kiếm đó bao gồm "brother and sister," "step mom f*cks son," "mom and son," "mom f*cks son," và "real brother and sister." Phần nhiều các tìm kiếm loạn luân của phái nam là các cảnh mẹ và con trai. Còn phái nữ? 9 trong 100 tìm kiếm hàng đầu của nữ trên P***hub là các video chủ đề loạn luân, và các cảnh cũng tương tự—mặc dù giới tính của bố mẹ và con thường ngược lại. Như vậy, phần nhiều các tìm kiếm loạn luân của nữ là các cảnh bố và con gái.

Không khó để xác định trong dữ liệu này phức cảm Oedipus (Oedipal complex) của Freud. Ông đã đưa ra giả thuyết rằng có một ham muốn cận phổ quát thời thơ ấu (mà về sau bị đè nén) liên quan đến tình

dục với bố mẹ khác giới. Giá mà nhà tâm lí học người Vienna này sống đủ lâu để sử dụng các kĩ năng phân tích của ông với dữ liệu P***hub, nơi mà sự quan tâm đến bố mẹ khác giới dường như đã được người trưởng thành xác nhận rất rõ ràng, và là nơi ít có thứ gì bị đè nén.

Dĩ nhiên, dữ liệu P***hub không thể cho biết chắc chắn người ta đang tưởng tượng về ai khi xem các video đó. Họ có thực sự đang tưởng tượng quan hệ với chính bố mẹ họ không? Các tìm kiếm Google có thể cung cấp thêm một số đầu mối cho rằng có nhiều người ham muốn như thế.

Xem xét tất cả tìm kiếm theo mẫu câu "I want to have sex with my..." Gợi ý hàng đầu để hoàn tất tìm kiếm này là "mom." Về tổng thể, hơn 3/4 các tìm kiếm theo mẫu này là loạn luân. Và điều này không phải do ta chọn đúng mẫu câu đặc biệt. Các tìm kiếm của mẫu câu "I am attracted to..." còn cho thấy sự vượt trội hơn của ham muốn loạn luân. Nhưng thôi, tôi xin thừa nhận và làm Ngài Freud thất vọng: Đây không phải là những tìm kiếm phổ biến. Chỉ vài ngàn người mỗi năm ở Mĩ thú nhận là thích mẹ mình. Ai đó nên báo cho Freud biết rằng các tìm kiếm Google, như sẽ được thảo luận sau trong sách này, thỉnh thoảng bị bóp méo nhằm hướng đến những điều cấm kị.

Nhưng khoan. Vẫn còn nhiều sở thích khác mà tôi vốn nghĩ là sẽ được tìm kiếm thường xuyên hơn. Sếp? Nhân viên? Học sinh/sinh viên? Nhà trị liệu? Bệnh nhân? Bạn thân nhất của vợ? Bạn thân nhất của con gái? Em vợ? Vợ bạn thân nhất? Không có ước muốn nào có thể cạnh tranh với mẹ. Có lẽ, kết hợp với dữ liệu P***hub, điều này thực sự có ý nghĩa gì đó.

Khẳng định chung của Freud cho rằng tình dục có thể định hình qua các trải nghiệm tuổi thơ cũng được dữ liệu Google và P***hub ủng hộ. Dữ liệu tiết lộ rằng nam giới ít nhất cũng có một số tưởng tượng bất thường liên quan đến tuổi thơ. Theo các tìm kiếm của các bà vợ về chồng họ, một số điều ưa thích hàng đầu của nam giới trưởng thành là muốn mang tã và được cho bú, đặc biệt là ở Ấn Độ. Hơn nữa, khiêu dâm hoạt hình—với những cảnh sex sống động của các nhân vật trong những bộ

phim mà các cậu thiếu niên ưa thích—cũng có mức độ phổ biến cao. Hoặc xem xét các nghề của nữ được nam tìm kiếm thường xuyên nhất trong khiêu dâm. Nam 18-24 tuổi thường tìm kiếm các cô trông trẻ. Nam 25-64 tuổi cũng vậy. Cả nam 65 tuổi trở lên nữa. Đối với nam giới mọi nhóm tuổi, giáo viên và trưởng nhóm cổ vũ đều nằm trong top 4. Rõ ràng, những năm đầu đời có vẻ đóng một vai trò rất lớn trong những tưởng tượng của nam giới trưởng thành.

Tôi chưa thể dùng tất cả dữ liệu chưa từng có này về tình dục người trưởng thành để hiểu chính xác các sở thích tình dục hình thành ra sao. Qua vài thập niên kế tiếp, các nhà khoa học xã hội khác và tôi sẽ có thể tạo ra các lí thuyết mới có thể kiểm chứng về tình dục người trưởng thành và kiểm tra bằng dữ liệu thực tế.

Tôi đã có thể dự báo vài chủ đề cơ bản chắc chắn sẽ là một bộ phận của lí thuyết tình dục người trưởng thành dựa trên dữ liệu. Rõ ràng sẽ không phải là câu chuyện giống như Freud kể, với các giai đoạn cụ thể, rạch ròi, phổ quát về tuổi thơ và sự kìm nén. Tuy nhiên, dựa trên quan sát dữ liệu P***hub đầu tiên của mình, tôi hoàn toàn chắc chắn rằng phán quyết cuối cùng về tình dục người trưởng thành sẽ nổi bật một số chủ đề chính mà Freud đã nhấn mạnh. Tuổi thơ sẽ đóng vai trò chủ đạo. Các bà mẹ cũng vậy.

---

Chắc chắn không thể nào phân tích Freud bằng cách này vào thời điểm 10 năm trước. Càng chắc chắn là không thể vào 80 năm trước, khi Freud vẫn còn sống. Vậy ta hãy cùng suy nghĩ xem tại sao các nguồn dữ liệu này lại hữu ích. Bài tập này có thể giúp ta hiểu tại sao Dữ Liệu Lớn lại mạnh mẽ đến vậy.

Xin nhớ, chúng ta đã biết rằng bản thân việc có hàng núi dữ liệu không tự động giúp phát sinh hiểu biết. Kích cỡ dữ liệu được đánh giá quá cao so với thực chất. Vậy thì tại sao Dữ Liệu Lớn lại mạnh đến vậy? Tại sao nó sẽ tạo ra một cuộc cách mạng về cách ta nhìn chính mình? Tôi khẳng định có 4 sức mạnh độc đáo trong Dữ Liệu Lớn. Phân tích sau đây về Freud cung cấp một sự minh họa rõ ràng.

Có thể bạn đã thấy, ngay từ đầu, rằng chúng ta đã nghiêm túc đem chuyện khiêu dâm ra bàn vấn đề Freud. Và chúng ta sẽ sử dụng dữ liệu từ các trang khiêu dâm thường xuyên trong sách này. Cũng hơi bất ngờ, dữ liệu khiêu dâm hiếm khi được các nhà xã hội học sử dụng, họ hầu hết thấy dễ chịu khi dựa vào các bộ dữ liệu khảo sát truyền thống đã giúp họ xây dựng sự nghiệp của mình. Nhưng chỉ cần nghĩ kĩ một chút, ta sẽ thấy rằng sự phổ biến của các trang khiêu dâm—kèm dữ liệu tìm kiếm và lượt xem—là bước tiến quan trọng nhất để hiểu tình dục của con người... Thực vậy, có lẽ đây là bước tiến quan trọng nhất từ trước tới giờ. Dữ liệu là thứ Schopenhauer, Nietzsche, Freud, và Foucault đã rất muốn có. Dữ liệu này không tồn tại khi họ còn sống. Cách đây vài thập niên cũng không. Nhưng bây giờ thì nó đã tồn tại. Có nhiều nguồn dữ liệu độc đáo, về một loạt chủ đề, cho ta những cánh cửa đi vào các lĩnh vực mà trước đây ta chỉ có thể đoán. *Cung cấp các loại dữ liệu mới chính là sức mạnh thứ nhất của Dữ Liệu Lớn.*

Dữ liệu khiêu dâm và dữ liệu tìm kiếm Google không chỉ mới mà còn chân thật. Trước thời đại kĩ thuật số, người ta giấu người khác các suy nghĩ "khó nói" của mình. Trong thời đại kĩ thuật số, họ vẫn giấu người khác, nhưng không giấu Internet, đặc biệt là các trang như Google và P***hub, vì sự nặc danh của họ đã được bảo vệ. Các trang này giống như một loại huyết thanh nói thật kĩ thuật số—và nhờ đó ta đã có thể phát hiện sự quyến rũ phổ biến của loạn luân. Dữ Liệu Lớn cho phép ta thấy được người ta thật sự muốn gì và thật sự làm gì, chứ không chỉ những điều họ nói là họ sẽ làm. *Cung cấp dữ liệu chân thật chính là sức mạnh thứ hai của Dữ Liệu Lớn.*

Bởi vì bây giờ có rất nhiều dữ liệu, thông tin có ý nghĩa thậm chí tồn tại trên những lát cắt nhỏ bé của một tổng thể. Chẳng hạn, ta có thể so sánh số người mơ về dưa leo với số người mơ về cà chua. *Cho phép ta phóng to các nhóm nhỏ chính là sức mạnh thứ ba của Dữ Liệu Lớn.*

Dữ Liệu Lớn còn có một sức mạnh ấn tượng nữa—sức mạnh không được dùng trong nghiên cứu nhanh của tôi về Freud nhưng có thể dùng trong tương lai: Nó cho phép ta tiến hành các thí nghiệm nhanh, có kiểm soát. Điều này cho phép ta kiểm định tính nhân quả, chứ không chỉ các

tương quan. Các loại kiểm định này bây giờ chủ yếu được dùng bởi các doanh nghiệp, nhưng sẽ chứng tỏ là một công cụ mạnh cho các nhà khoa học xã hội. *Cho phép ta thực hiện nhiều thí nghiệm nhân quả chính là sức mạnh thứ tư của Dữ Liệu Lớn.*

Bây giờ là lúc "đập hộp" mỗi sức mạnh này và khám phá chính xác tại sao Dữ Liệu Lớn lại quan trọng đến thế.

# CHƯƠNG 3

## Tái hình dung dữ liệu

Lúc 6 giờ sáng một ngày thứ Sáu đặc biệt mỗi tháng, đường phố của hầu khắp Manhattan sẽ rất hoang vắng. Cửa hiệu dọc theo các đường phố này đóng cửa, mặt tiền bị che lại bởi các cổng bảo vệ bằng thép, các căn hộ bên trên thì tối tăm im lìm.

Trái lại, các tầng tòa nhà của ngân hàng đầu tư toàn cầu Goldman Sachs ở khu thương mại Manhattan đèn đóm sáng rực. Thang máy đưa hàng ngàn nhân viên đến nơi làm việc. Trước 7 giờ sáng, hầu hết các bàn làm việc đều sẽ có người ngồi.

Vào các ngày khác, mô tả cả khu phố ở thời điểm này là đang ngái ngủ thì cũng chẳng sai. Tuy nhiên, vào đúng sáng thứ Sáu này, ở đây sẽ bừng bừng năng lượng và sự hứng khởi. Đây chính là ngày một thông tin tác động mạnh thị trường chứng khoán sẽ đến.

Mấy phút sau khi tiết lộ, thông tin này sẽ được các trang tin tức thuật lại. Mấy giây sau khi tiết lộ, thông tin này sẽ được thảo luận, tranh cãi, và mổ xẻ rất ồn ào tại Goldman và hàng trăm công ty tài chính khác. Nhưng phần nhiều hoạt động tài chính thực sự những ngày này diễn ra chỉ trong mấy phần ngàn giây. Goldman và các công ty tài chính khác đã trả hàng chục triệu đô cho cáp quang để giảm thời gian thông tin truyền từ Chicago đến New Jersey chỉ 4 ms (mili-giây) (từ 17 xuống 13). Các công ty tài chính có các thuật toán sẵn sàng để đọc thông tin và dựa vào đó đặt lệnh—tất cả chỉ trong vài ms. Sau khi thông tin quan trọng này

được tiết lộ, thị trường sẽ chuyển động trong thời gian chưa bằng một cái nháy mắt.

Vậy thông tin quan trọng này là gì mà lại giá trị với Goldman và nhiều cơ quan tài chính khác đến thế?

Tỉ lệ thất nghiệp hàng tháng.

Tuy nhiên, tỉ lệ đó—một con số có ảnh hưởng sâu sắc lên thị trường chứng khoán đến mức các định chế tài chính đã làm bất cứ điều gì để tối đa hóa tốc độ nhận, phân tích, và hành động dựa theo nó—lại xuất phát từ một khảo sát qua điện thoại mà Cục Thống kê Lao động (BLS) thực hiện. Tính đến lúc được công bố, thông tin này đã cũ gần 3 tuần—tức là gần 2 tỉ mili-giây.

Khi các công ty đang chi hàng triệu đô la để tranh thủ từng mili-giây của dòng chảy thông tin, việc chính phủ mất thời gian quá lâu để tính toán tỉ lệ thất nghiệp có lẽ khiến bạn cảm thấy kì lạ.

Thực vậy, việc công bố các con số quan trọng này sớm hơn là một trong các chương trình nghị sự đầu tiên của Alan Krueger khi ông đảm nhiệm chức chủ tịch Hội đồng Cố vấn Kinh tế của Tổng thống Obama năm 2011. Ông đã không thành công. Ông kết luận: "Hoặc BLS không có đủ nguồn lực, hoặc họ bị sa lầy trong tư duy Thế kỉ XX."

Với việc chính phủ rõ ràng không thể theo kịp nhịp độ trong thời gian gần, có cách nào để ít nhất nắm sơ bộ số liệu thống kê thất nghiệp nhanh hơn không? Trong thời đại công nghệ cao này—khi gần như mỗi cú nhấp chuột mà một người thực hiện trên Internet đều được ghi lại đâu đó—ta có thực sự phải đợi hàng tuần để biết có bao nhiêu người đang thất nghiệp không?

Một giải pháp tiềm năng được truyền cảm hứng bởi công trình của một cựu kĩ sư Google, Jeremy Ginsberg. Ginsberg chú ý thấy rằng dữ liệu sức khỏe, giống như dữ liệu thất nghiệp, được chính phủ công bố khá chậm trễ. Trung tâm Phòng chống Dịch bệnh mất 1 tuần để công bố dữ liệu bệnh cúm, dù bác sĩ và bệnh viện sẽ có lợi nếu nắm dữ liệu đó sớm hơn.

Ginsberg nghĩ rằng những người bị cúm có thể thực hiện các tìm kiếm liên quan đến bệnh cúm. Chủ yếu, họ sẽ báo các triệu chứng cho Google. Các tìm kiếm này, ông nghĩ, có thể cung cấp một hiểu biết khá chính xác về tỉ lệ bệnh cúm hiện tại. Thực vậy, các tìm kiếm như "các triệu chứng cúm" và "đau cơ" đã chứng tỏ là các chỉ báo quan trọng về tốc độ lây lan dịch cúm.[1]

Trong khi đó, các kĩ sư Google tạo ra một dịch vụ, Google Correlate, cung cấp cho các nhà nghiên cứu bên ngoài phương tiện để thí nghiệm với kiểu phân tích này trên vô vàn lĩnh vực, không chỉ riêng sức khỏe. Các nhà nghiên cứu có thể lấy bất cứ chuỗi dữ liệu nào mà họ đang theo dõi theo thời gian và xem các tìm kiếm Google nào tương quan nhất với bộ dữ liệu đó.

Ví dụ, khi dùng Google Correlate, Hal Varian, trưởng kinh tế gia tại Google, và tôi đã có thể chỉ ra các tìm kiếm nào tương quan sát với giá nhà ở nhất. Khi giá nhà đang lên, người Mĩ có khuynh hướng tìm kiếm các cụm từ như "thế chấp 80/20," "người xây nhà mới," và "tỉ lệ tăng giá." Khi giá nhà đang xuống, người Mĩ có khuynh hướng tìm kiếm các cụm từ như "short sale process," "underwater mortgage," và "mortgage forgiveness debt relief."[2]

Vậy liệu các tìm kiếm Google có thể đóng vai trò như một phép thử tình trạng thất nghiệp tương tự giá nhà và bệnh cúm không? Liệu rằng chỉ đơn giản từ những gì người ta đang Google, ta có thể biết có bao nhiêu người thất nghiệp, và ta có thể làm thật tốt trước khi chính phủ tổng hợp xong kết quả khảo sát không?

Một ngày nọ, tôi đưa tỉ lệ thất nghiệp ở Mĩ từ năm 2004 đến 2011 vào Google Correlate.

---

[1] Mặc dù phiên bản đầu tiên của Google Flu còn nhiều lỗi lớn, các nhà nghiên cứu gần đây đã hiệu chỉnh lại mô hình, với nhiều thành công hơn.

[2] [ND] "Short sale" nhà là bán nhà trước khi bị tịch thu (còn gọi là pre-foreclosure sale). "Underwater mortgage" chỉ khoản vay có giá trị cao hơn giá trị tài sản thế chấp (căn nhà). "Mortgage Forgiveness Debt Relief" là tên một đạo luật năm 2007 liên quan đến phương hướng giải quyết phần thuế khi xử lí nợ vay thế chấp nhà.

Từ hàng ngàn tỉ tìm kiếm Google trong suốt thời gian đó, bạn nghĩ cái gì sẽ liên quan mật thiết nhất với tình trạng thất nghiệp? Bạn có thể nghĩ đến cụm từ "phòng trợ cấp thất nghiệp"—hoặc cái gì đó tương tự. Tìm kiếm đó cao nhưng không phải cao nhất. "Việc làm mới"? Cũng không phải cao nhất.

Tìm kiếm cao nhất trong thời kì đó là "S***load" (và từ này có thay đổi tùy thời). Đúng thế, tìm kiếm thường xuyên nhất là một trang khiêu dâm. Mới nhìn thì có vẻ kì lạ, nhưng những người thất nghiệp thường có nhiều thời gian rảnh. Nhiều người nằm mẹp ở nhà, đơn độc và buồn chán. Một trong các tìm kiếm tương quan cao khác (từ này thì không thuộc nhóm 18+) là "Spider Solitaire." Một lần nữa, không ngạc nhiên đối với nhóm người được gọi là tỉ phú thời gian.

Dĩ nhiên, tôi không định dựa trên mỗi phân tích này mà cho rằng theo dõi từ khóa "S***load" hoặc "Spider Solitaire" là cách tốt nhất để dự báo tỉ lệ thất nghiệp. Các trò giải trí của người thất nghiệp có thể thay đổi theo thời gian (có lúc, "R**tube," một trang khiêu dâm khác, là một trong số các từ khóa tương quan mạnh nhất) và không một từ khóa đơn lẻ nào thu hút được số đông người thất nghiệp. Nhưng nói chung tôi đã phát hiện rằng một hỗn hợp các tìm kiếm liên quan đến giải trí có thể giúp ta theo dõi được tỉ lệ thất nghiệp—và sẽ là một bộ phận của mô hình dự báo tỉ lệ thất nghiệp tốt nhất.

Ví dụ này minh họa sức mạnh thứ nhất của Dữ Liệu Lớn—tái hình dung tiêu chuẩn dữ liệu. Thông thường, giá trị của Dữ Liệu Lớn không phải ở kích cỡ mà chính là ở chỗ nó có thể cung cấp các loại thông tin mới để nghiên cứu—thông tin mà trước đó chưa bao giờ được thu thập.

Trước Google, vẫn có một số thông tin về các hoạt động giải trí—doanh số vé xem phim, chẳng hạn, có thể cho ta vài gợi ý về lượng thời gian rảnh. Nhưng cơ hội biết người ta bỏ bao nhiêu thời gian để chơi bài solitaire hoặc xem đồ khiêu dâm thì chỉ mới đây thôi—và đó là dữ liệu rất mạnh. Trong ví dụ đang bàn, dữ liệu đó có thể giúp ta đo lường sức khỏe nền kinh tế nhanh hơn—ít nhất là tạm dùng được cho đến khi chính phủ tìm ra cách khảo sát nhanh hơn.

***

Cuộc sống ở khuôn viên Google tại Mountain View, California, rất khác với cuộc sống tại trụ sở Manhattan của Goldman Sachs. Lúc 9 giờ sáng, các văn phòng ở Google gần như trống trơn. Nếu có nhân viên nào quanh đó, thì chắc là để ăn sáng miễn phí—bánh nhân chuối việt quất, lòng trắng trứng bác, nước dưa leo lọc. Một số nhân viên có thể đã đi khỏi thành phố: Họ đi họp ở ngoài công ty tại Boulder hay Las Vegas, hoặc có thể đang đến Lake Tahoe trượt tuyết miễn phí. Khoảng thời gian ăn trưa, các sân bóng chuyền bãi biển và sân bóng đá sẽ đầy kín. Món bánh burrito ngon nhất tôi từng ăn là tại nhà hàng Mexico của Google.

Làm thế nào mà một trong các công ty công nghệ cạnh tranh nhất và lớn nhất thế giới lại có vẻ thoải mái và rộng rãi như vậy? Google khai thác Dữ Liệu Lớn theo cách mà chưa có công ty nào khác từng làm để xây dựng một dòng tiền tự động. Google đóng một vai trò quan trọng trong quyển sách này vì hiện nay, các tìm kiếm Google là nguồn Dữ Liệu Lớn vượt trội nhất. Nhưng phải nhớ, thành công của Google tự bản chất được xây dựng dựa trên hoạt động thu thập một loại dữ liệu mới.

Nếu đã có dịp dùng Internet trong Thế kỉ XX, bạn hẳn còn nhớ nhiều máy tìm kiếm khác có mặt lúc đó—MetaCrawler, Lycos, AltaVista, chẳng hạn. Và bạn chắc còn nhớ rằng các máy tìm kiếm này, tốt lắm đi nữa, cũng chỉ hơi đáng tin thôi. Đôi khi, nếu may mắn, các máy đó cũng tìm thấy được điều bạn muốn. Thông thường thì không may mắn thế đâu. Nếu bạn gõ "Bill Clinton" vào các máy tìm kiếm phổ biến nhất cuối thập niên 1990, các kết quả hàng đầu gồm một trang web ngẫu nhiên có mấy dòng tuyên bố "Bill Clinton cùi bắp" hoặc một trang chủ yếu là chuyện tiếu lâm về Clinton. Hiếm khi có thông tin phù hợp nhất về vị tổng thống Mĩ lúc bấy giờ.

Năm 1998, Google xuất hiện. Và các kết quả tìm kiếm của nó rõ ràng là tốt hơn hẳn kết quả của mọi đối thủ. Nếu ta gõ "Bill Clinton" vào Google năm 1998, ta được cung cấp website của ông, địa chỉ email của Nhà Trắng, và những trang tiểu sử tốt nhất của người này trên Internet. Google bấy giờ có vẻ rất thần kì.

Các nhà sáng lập Google, Sergey Brin và Larry Page, đã làm điều gì khác biệt?

Các máy tìm kiếm khác định vị cho người dùng các website thường xuyên có cụm từ mà họ đang tìm kiếm. Nếu bạn đang tìm thông tin về "Bill Clinton," các máy tìm kiếm đó sẽ tìm trên toàn bộ Internet các website nhắc đến Bill Clinton nhiều nhất. Có nhiều lí do khiến hệ thống xếp hạng này không hoàn hảo, một trong các lí do đó là nó rất dễ bị lợi dụng. Một trang chuyện đùa có dòng văn bản "Bill Clinton Bill Clinton Bill Clinton Bill Clinton Bill Clinton" được giấu đâu đó trên trang thường ghi điểm cao hơn website chính thức của Nhà Trắng.[1]

Điều Brin và Page đã làm là tìm cách ghi lại một loại thông tin mới có giá trị hơn hẳn phép đếm từ đơn giản. Các website, khi thảo luận một chủ đề, thường liên kết với các trang họ nghĩ là hữu ích nhất để hiểu chủ đề đó. Ví dụ, báo *New York Times*, nếu đề cập Bill Clinton, có thể sẽ cho phép độc giả nhấp chuột lên tên ông để được đưa đến website chính thức của Nhà Trắng.

Theo góc nhìn này, mỗi khi website tạo đường dẫn, nó đã biểu đạt quan điểm đâu là nguồn tin tốt nhất về Bill Clinton. Brin và Page có thể gom tất cả các ý kiến kiểu này về mọi chủ đề. Họ có thể huy động các quan điểm của *New York Times*, của hàng triệu người dùng Listserv, của hàng trăm blogger, và của mọi người khác trên Internet. Nếu cả một đám người nghĩ rằng đường dẫn quan trọng nhất cho "Bill Clinton" là website chính thức của ông, đây có thể là website mà hầu hết những người tìm kiếm "Bill Clinton" sẽ muốn xem.

Các đường dẫn này là dữ liệu mà các máy tìm kiếm khác không xét đến; chúng bất ngờ dự báo rất chuẩn thông tin hữu ích nhất về một chủ đề được cho. Điểm chính ở đây là Google không thống trị sự tìm kiếm đơn thuần bằng cách thu thập nhiều dữ liệu hơn mọi người. Họ thống trị bằng cách tìm ra một loại dữ liệu *tốt hơn*. Chưa tới 2 năm sau khi trình

[1] Năm 1998, nếu bạn tìm "cars" (xe hơi) trên một máy tìm kiếm phổ biến trước thời Google, bạn sẽ nhận được kết quả đầy các trang khiêu dâm. Các trang khiêu dâm này đã viết từ "cars" thường xuyên bằng chữ trắng trên nền trắng để lừa máy tìm kiếm. Thế là họ nhận thêm vài cú nhấp chuột từ những người muốn mua xe nhưng lại bị nội dung khiêu dâm thu hút.

làng, Google, với sức mạnh đến từ việc phân tích đường dẫn, đã phát triển thành cỗ máy tìm kiếm phổ biến nhất của Internet. Ngày nay, cả Brin và Page đều sở hữu gần 60 tỉ USD.

Giống với Google, mọi người khác cũng đang cố gắng dùng dữ liệu để hiểu thế giới. Trọng tâm Cuộc cách mạng Dữ Liệu Lớn không phải là ngày càng nhiều dữ liệu, mà chính là thu thập đúng dữ liệu.

Nhưng Internet không phải là nơi duy nhất bạn có thể thu thập dữ liệu mới, và việc lấy được dữ liệu đúng có thể mang lại kết quả đột phá. Quyển sách này chủ yếu nói về việc dữ liệu trên web có thể giúp ta hiểu mọi người tốt hơn ra sao. Tuy nhiên, phần tiếp theo không có gì liên quan đến dữ liệu web. Thực vậy, nó thậm chí còn không liên quan đến con người. Thế nhưng, nó thực sự giúp minh họa điểm chính của chương này: giá trị to lớn của dữ liệu mới, trái với niềm tin thông thường. Và các nguyên tắc rút ra từ đó sẽ rất hữu ích để hiểu cuộc cách mạng dữ liệu kĩ thuật số ngày nay.

## Cơ thể là dữ liệu

Mùa hè năm 2013, một con ngựa tía, kích thước trên trung bình, bờm đen, đang yên vị trong một cái chuồng nhỏ vùng Bắc New York. Nó là 1 trong 152 con ngựa tại đợt bán ngựa 1 năm tuổi tuyển chọn của Fasig-Tipton, lứa tháng 8 ở Saratoga Springs, và cũng là 1 trong 10,000 con ngựa 1 năm tuổi được bán đấu giá trong năm đó.

Các ông các bà giàu có, khi bỏ ra nhiều tiền để mua về một con ngựa đua, đều muốn có vinh dự đặt tên cho nó. Như vậy con ngựa tía chưa có tên, và như hầu hết các con ngựa tại cuộc đấu giá, nó được gọi tên theo số chuồng, 85.

Ít có điểm gì khiến Số 85 nổi bật tại cuộc đấu giá. Dòng dõi của nó tốt, nhưng không xuất sắc. Bố nó, Pioneerof the Nile, là ngựa đua hàng đầu, nhưng các con của Pioneerof the Nile lại chưa có nhiều thành tích. Cũng có những nghi ngờ dựa trên bộ dạng của Số 85. Ví dụ, có một vết xước ở cổ chân khiến nhiều người mua lo rằng nó có thể đã từng bị thương.

Người chủ của Số 85 lúc này là một trùm rượu bia người Ai Cập, Ahmed Zayat. Ông đến Bắc New York, hi vọng bán được con ngựa này và mua vài con khác.

Như hầu hết các ông chủ khác, Zayat thuê một nhóm chuyên gia giúp ông chọn ngựa để mua. Nhưng chuyên gia của Zayat hơi khác với chuyên gia của các ông chủ kia. Chuyên gia chọn ngựa bạn thường thấy tại một sự kiện như vậy là những ông tuổi trung niên, chủ yếu đến từ Kentucky hoặc vùng thôn quê Florida, ít học, nhưng gia đình đã hoạt động trong ngành kinh doanh ngựa từ lâu. Tuy nhiên, chuyên gia của Zayat lại đến từ một công ty nhỏ gọi là EQB. Đứng đầu EQB không phải là một chuyên gia ngựa trường phái cũ. Đứng đầu EQB, trái lại, là Jeff Seder, một người lập dị, sinh ra ở Philadelphia với cả một chồng bằng cấp từ Harvard.

Zayat đã làm việc với EQB trước đó, vì vậy quy trình đã khá quen thuộc. Sau vài ngày đánh giá ngựa, nhóm của Seder sẽ quay lại chỗ Zayat với khoảng 5 con ngựa họ đề nghị mua để thay thế Số 85.

Tuy nhiên, lần này thì khác. Nhóm của Seder quay lại chỗ Zayat và bảo là họ không thể thi hành đề nghị của ông. Đơn giản là họ không thể đề nghị ông nên mua con ngựa nào trong 151 con còn lại được chào bán ngày hôm đó. Thay vì vậy, họ đề xuất một yêu cầu thật khẩn thiết và bất ngờ. Zayat tuyệt đối không được bán con ngựa Số 85. Con ngựa này, EGB tuyên bố, không chỉ là con tốt nhất trong cuộc đấu giá; nó là con tốt nhất trong năm, và có thể là trong cả thập niên. "Ông muốn bán nhà cũng được," cả nhóm khuyên ông. "Nhưng đừng bán con ngựa này."

Ngày tiếp theo, một cách êm đềm, con ngựa Số 85 đã được mua với giá 300,000 USD bởi một người tự xưng là Incardo Bloodstock. Bloodstock, sau được tiết lộ, là tên giả của Ahmed Zayat. Nghe theo Seder, Zayat đã mua lại con ngựa của chính mình, một hành động hầu như chưa thấy bao giờ. (Luật đấu giá không cho phép Zayat đơn giản rút con ngựa khỏi cuộc đấu giá, vì thế cần phải giao dịch bằng tên giả.) 62 con ngựa tại cuộc đấu giá được bán giá cao hơn con ngựa Số 85, có 2 con mang về cho chủ hơn 1 triệu USD mỗi con.

Sau đó 3 tháng, cuối cùng Zayat đã chọn được một cái tên cho Số 85: American Pharoah. Và 18 tháng sau, vào một buổi chiều thứ Bảy 24°C ở ngoại ô Thành phố New York, American Pharoah trở thành con ngựa đầu tiên trong hơn 3 thập niên giành được danh hiệu Triple Crown (thắng 3 cuộc đua lớn).

Jeff Seder biết gì về con ngựa Số 85 mà rõ ràng không ai khác biết được? Làm thế nào mà người đàn ông tốt nghiệp Harvard này lại có thể giỏi đánh giá ngựa như thế?

Tôi lần đầu gặp Seder, bấy giờ ông đã 64 tuổi, vào một chiều tháng 6 nóng cháy da tại Ocala, Florida, hơn 1 năm sau khi American Pharoah đoạt danh hiệu Triple Crown. Đó là sự kiện chào hàng kéo dài cả tuần cho các chú ngựa 2 năm tuổi, giao dịch theo kiểu đấu giá, tương tự như sự kiện năm 2013 mà Zayat đã tự mua lại ngựa của mình vậy.

Seder có giọng nói oang oang giống Mel Brooks, một cái đầu đầy tóc, và dáng đi nhún nhảy rất nổi bật. Ông mặc quần khaki có dây đeo, sơ mi đen có logo của công ty ông trên đó, và một thiết bị trợ thính.

3 ngày tiếp theo, ông kể tôi nghe chuyện đời ông—và cả chuyện ông học cách xem ngựa như thế nào. Đó không phải là một đường thẳng. Sau khi tốt nghiệp Harvard hạng xuất sắc và được hội Phi Beta Kappa vinh danh, Seder tiếp tục lấy bằng luật và bằng kinh doanh, cũng tại Harvard. Ở tuổi 26, ông làm nhà phân tích cho Citigroup ở Thành phố New York, nhưng cảm thấy không vui và quá mệt mỏi. Một ngày kia, khi đang ngồi ở tiền sảnh văn phòng mới của công ty trên Đại lộ Lexington, ông vô tình ngắm nhìn bức bích họa một cánh đồng rộng lớn. Bức tranh gợi ông nhớ lại tình yêu đồng quê và các chú ngựa của ông. Ông về nhà và ngắm mình trong gương, người mặc bộ đồ đi làm chỉn chu. Bấy giờ ông đã biết rằng mình sinh ra không phải là để làm một nhân viên ngân hàng và sống ở Thành phố New York. Sáng hôm sau, ông bỏ việc.

Seder chuyển đến vùng quê Pennsylvania và thong thả trải qua nhiều loại công việc trong ngành dệt và y học thể thao trước khi dâng trọn đời mình cho niềm đam mê: tiên đoán thành công của ngựa đua. Tỉ lệ thành công trong ngành đua ngựa khá khó nhai. Trong 1,000 con ngựa

2 năm tuổi chào bán tại cuộc đấu giá Ocala, một trong những nơi uy tín nhất cả nước, có lẽ cuối cùng chỉ 5 con sẽ thắng cuộc đua và mang về một túi tiền đầy. Điều gì sẽ xảy ra với 995 con ngựa kia? Khoảng 1/3 chạy quá chậm. Cũng khoảng 1/3 khác sẽ bị thương—hầu hết vì chân cẳng không chịu nổi áp lực lớn khi phi nước đại hết tốc lực. (Mỗi năm, hàng trăm con ngựa chết trên đường đua Mĩ, hầu hết do gãy chân.) Và 1/3 còn lại sẽ ở trong tình trạng mà ta có thể gọi là hội chứng Bartleby. Bartleby, công chứng viên trong truyện ngắn xuất sắc của Herman Melville, nghỉ làm việc và trả lời mọi yêu cầu của ông chủ với câu "Tôi không muốn làm." Nhiều con ngựa, ngay từ đầu sự nghiệp chạy đua, bắt đầu nhận thấy rằng chúng không cần chạy nếu không thấy thích. Khi mới vào đua, chúng sẽ chạy rất nhanh, nhưng đến một thời điểm nào đó, chúng sẽ chạy chậm lại hoặc không chạy nữa. Tại sao lại phải chạy quanh một cái vòng tròn hết tốc lực, đặc biệt khi móng guốc và chân cẳng ta đau thế chứ? "Tôi không muốn chạy," các chú ngựa quyết định thế. (Tôi rất có cảm tình với các Bartleby, dù là ngựa hay người.)

Khi cơ hội quá nhỏ nhoi, làm thế nào các ông chủ có thể chọn được con ngựa có lợi nhuận? Từ xưa tới nay, người ta tin rằng cách tốt nhất để tiên đoán một con ngựa có thành công hay không là phân tích dòng dõi của nó. Làm chuyên gia về ngựa nghĩa là có thể nói một mạch mọi thứ mà ai cũng muốn biết về bố, mẹ, ông bà nội ngoại, anh chị em của một con ngựa nào đó. Ví dụ, các nhân viên khảo sát ngựa sẽ thông báo rằng một con ngựa lớn "đạt đến kích thước hợp lí" nếu dòng dõi mẹ của nó cũng có nhiều con ngựa lớn.

Tuy nhiên, có một vấn đề. Mặc dù rất quan trọng, dòng dõi vẫn chỉ có thể giải thích một phần nhỏ thành công của ngựa đua. Hãy thử xem xét thông tin đầy đủ về anh chị em ruột của tất cả các con ngựa được tặng danh hiệu *Tuấn mã của năm*, giải thưởng hàng năm danh giá nhất của trò đua ngựa. Các con ngựa này đều có dòng dõi xịn—gia phả của nó không thua gì nhiều con ngựa nổi tiếng trong lịch sử thế giới. Tuy nhiên, hơn 3/4 không thắng cuộc đua lớn nào. Dữ liệu cho ta biết, phương pháp truyền thống tiên đoán thành công của ngựa vẫn còn nhiều chỗ trống cần cải thiện.

Chuyện dòng dõi không dự báo được nhiều thực ra chẳng có gì đáng ngạc nhiên. Nghĩ về con người mà xem. Hãy tưởng tượng một ông chủ đội bóng đang thi đấu ở NBA mua cầu thủ cho đội trẻ—gồm các cậu bé 10 tuổi—dựa trên dòng dõi. Ông sẽ thuê một nhân viên khảo sát Earvin Johnson III, con trai của "Magic" Johnson.[1] "Cho đến nay, cậu bé có số đo tốt," một nhân viên có thể nói. "Số đo hợp lí, thuộc dòng Johnson. Cậu bé chắc sẽ có tầm nhìn, lòng vị tha, số đo, và tốc độ tuyệt vời. Cậu có vẻ thuộc nhân cách hướng ngoại, rất tốt. Bước đi tự tin. Dễ coi. Ca này tuyệt vời." Thật không may, 14 năm sau đó, ông chủ này sẽ có một blogger thời trang ở kênh truyền hình *E!* cao 1m88 (lùn so với cầu thủ chuyên nghiệp). Earvin Johnson III có thể rất giỏi trong thiết kế đồng phục, nhưng có lẽ chẳng giỏi lắm trên sân bóng rổ.

Ngoài blogger thời trang, một ông chủ NBA mà chọn đội bóng theo cách tương tự phương pháp chọn ngựa truyền thống chắc chắn sẽ chộp ngay Jeffrey và Marcus Jordan, hai con trai của Michael Jordan. Cả hai người này đều chỉ là cầu thủ đẳng cấp đội bóng đại học tầm thường. Xin chúc họ may mắn nếu phải gặp Cleveland Cavaliers, đội bóng của LeBron James, một cầu thủ có mẹ chỉ cao 1m65. Hoặc hãy tưởng tượng một quốc gia bầu lãnh đạo dựa trên dòng dõi. Chúng ta sẽ được lãnh đạo bởi những người như George W. Bush. (Xin lỗi, chịu hổng nổi.)

Ngoài dòng dõi, người chọn ngựa cũng dùng thông tin khác nữa. Ví dụ, họ phân tích dáng đi của các con ngựa 2 năm tuổi và khảo sát trực quan. Tại Ocala, tôi đã bỏ hàng giờ tán chuyện với nhiều nhân viên, đủ lâu để biết được rằng mỗi người lại có một tiêu chí, chẳng ai giống ai.

Bên cạnh đầy rẫy mâu thuẫn và sự không chắc chắn tràn lan là việc một số người mua ngựa dường như có kinh phí vô tận. Thế là ta có một thị trường rất không hiệu quả. Cách đây 10 năm, Số 153 là con ngựa 2 năm tuổi chạy nhanh nhất, trông rất đẹp đối với hầu hết những người chọn ngựa, và có dòng dõi tuyệt vời—hậu duệ của Northern Dancer và Secretariat, hai trong số các ngựa đua xịn nhất mọi thời đại. Một tỉ phú Ireland và một lãnh đạo Dubai đều muốn mua nó. Họ sa vào một cuộc

[1] [ND] Earvin "Magic" Johnson Jr. là cầu thủ bóng rổ chuyên nghiệp, cao đến 2.06 m.

đấu giá, và tình hình nhanh chóng biến thành cuộc thi của lòng kiêu hãnh. Hàng trăm người chơi ngựa sững sờ đứng xem giá mỗi lúc một cao hơn, cho đến khi giá chốt ở mức 16 triệu USD, vượt xa giá cao nhất từng được trả cho một con ngựa. Số 153 được đặt tên là The Green Monkey, đua 3 cuộc, kiếm được chỉ 10,000 USD cho chủ, và về hưu.

Seder không bao giờ quan tâm đến các phương pháp đánh giá ngựa truyền thống. Ông chỉ quan tâm đến dữ liệu. Ông có kế hoạch đo lường nhiều thuộc tính của ngựa đua và xem các thuộc tính nào tương quan với thành tích. Phải chú ý rằng Seder đã thực hiện kế hoạch của ông 5 năm trước khi World Wide Web được phát minh ra. Nhưng chiến lược của ông dựa rất nhiều vào khoa học dữ liệu. Và những bài học từ câu chuyện của ông có thể áp dụng cho bất cứ ai dùng Dữ Liệu Lớn.

Suốt nhiều năm, cuộc theo đuổi của Seder chẳng tạo ra được gì ngoài sự thất vọng. Ông lấy số đo lỗ mũi ngựa, tạo ra bộ dữ liệu đầu tiên và lớn nhất thế giới về kích thước lỗ mũi ngựa và tổng thu nhập cuối cùng của từng con. Và ông đã phát hiện ra rằng kích thước lỗ mũi ngựa không dự báo được thành công. Ông đo điện tâm đồ để khám tim ngựa và cắt rời chân ngựa chết để đo lượng cơ co rút nhanh. Ông đã có lần cầm xẻng ở ngoài một chuồng trại để đo kích cỡ phân ngựa, dựa trên một thuyết cho rằng sụt cân quá nhiều trước giải đấu có thể làm ngựa chậm lại. Không yếu tố nào ở đây tương quan với thành công trong cuộc đua cả.

Thế rồi, cách đây 12 năm, ông đã có phát kiến lớn đầu tiên. Seder quyết định đo kích thước nội tạng của ngựa. Vì việc này không thực hiện được với công nghệ bấy giờ, ông đã tự chế máy siêu âm xách tay. Kết quả thật đáng chú ý. Ông phát hiện rằng kích thước của tim, đặc biệt là tâm thất trái, là một chỉ báo mạnh mẽ cho thành công của ngựa, và là biến số quan trọng nhất. Một cơ quan khác cũng quan trọng là lá lách: Các con ngựa lá lách nhỏ hầu như chẳng làm ăn gì được.

Seder có thêm 2 thành công nữa. Ông số hóa hàng ngàn video ngựa phi nước đại và thấy rằng một số dáng phi thực sự tương quan với thành công trên đường đua. Ông còn phát hiện rằng một số ngựa 2 năm tuổi thở khò khè sau khi chạy khoảng 200m. Những con ngựa như thế đôi khi

bán được đến 1 triệu đô la, nhưng dữ liệu của Seder bảo ông rằng các con thở khò khè hầu như không bao giờ thành công cả. Do đó, ông giao cho một trợ lí ngồi gần vạch đích và loại bỏ các con thở khò khè.

Trong khoảng 1,000 con ngựa tại cuộc bán đấu giá Ocala, chừng 10 con sẽ đạt tất cả các tiêu chuẩn của Seder. Ông bỏ qua yếu tố dòng dõi hoàn toàn, trừ khi điều đó ảnh hưởng đến giá con ngựa. "Dòng dõi cho ta biết một con ngựa có thể có một cơ hội rất nhỏ để phát triển thành ngựa xịn," ông nói. "Nhưng nếu tôi đã thấy nó xịn, tôi quan tâm nó đã tốt nhờ đâu làm gì nữa chứ?"

Một đêm nọ, Seder mời tôi tới phòng ông tại khách sạn Hilton ở Ocala. Trong phòng, ông kể tôi nghe về tuổi thơ, gia đình, và sự nghiệp của mình. Ông cho tôi xem hình vợ, con gái, và con trai ông. Ông bảo tôi ông là 1 trong 3 học sinh Do Thái tại trường trung học Philadelphia, và khi vào trường ông cao 1m47. (Ở đại học ông cao lên 1m75.) Ông kể tôi nghe về con ngựa ưa thích của ông: Pinky Pizwaanski. Seder đã mua và đặt tên con ngựa này theo tên một nài ngựa đồng giới. Ông cảm thấy Pink luôn nỗ lực ngay cả khi nó không phải là con ngựa thành công nhất.

Cuối cùng, ông cho tôi xem bộ hồ sơ bao gồm tất cả dữ liệu ông đã ghi lại về Số 85, bộ hồ sơ đã tạo nên dự báo chuẩn nhất trong sự nghiệp của ông. Ông đang cho đi bí quyết của mình sao? Có thể, nhưng ông nói ông không quan tâm. Đối với ông, điều quan trọng hơn cả việc bảo vệ bí quyết của mình là chứng minh nó đúng, cho thế giới thấy rằng 20 năm bẻ cẳng, xúc phân, và lắp máy siêu âm là không hề uổng công.

Đây là một vài dữ liệu về Số 85:

Bách phân vị của Số 85 (American Pharoah) lúc 1 tuổi

| | **Bách phân vị** |
|---|---|
| Chiều cao | 56 |
| Cân nặng | 61 |
| Dòng dõi | 70 |
| Tâm thất trái | 99.61 |

Đó, đơn giản và rõ ràng, lí do khiến Seder và nhóm ông đã rất mê con ngựa Số 85. Tâm thất trái của nó đạt mức bách phân vị 99.61![1]

Không chỉ thế, tất cả các cơ quan quan trọng khác của nó, bao gồm phần còn lại của tim và lá lách, cũng đều lớn khác thường. Nói chung, Seder đã phát hiện, khi chạy đua, tâm thất trái càng lớn càng tốt. Nhưng tâm thất trái lớn như thế này cũng có thể là một dấu hiệu bệnh lí nếu các cơ quan khác lại nhỏ. Ở American Pharoah, tất cả các cơ quan chủ chốt đều lớn hơn trung bình, và tâm thất trái rất lớn. Dữ liệu như hét lên rằng 100,000 hoặc thậm chí 1 triệu con ngựa mới có một con như Số 85.

***

Các nhà khoa học dữ liệu có thể học được gì từ dự án của Seder?

Trước hết, và có lẽ quan trọng nhất, nếu bạn muốn dùng dữ liệu mới để cách mạng hóa một lĩnh vực, tốt nhất là đi vào lĩnh vực mà các phương pháp cũ chẳng ra tích sự gì. Các nhân viên chọn ngựa chằm hăm vào dòng dõi đã để lại nhiều chỗ trống cho sự cải tiến. Các máy tìm kiếm chằm hăm vào đếm từ cũng thế.

Một điểm yếu trong nỗ lực dự báo bệnh cúm bằng dữ liệu tìm kiếm của Google là bạn hoàn toàn có thể dự báo bệnh cúm rất tốt chỉ bằng dữ liệu tuần vừa qua, cộng thêm một thủ thuật điều chỉnh dữ liệu theo mùa đơn giản. Vẫn có tranh luận về việc cần phải có bao nhiêu dữ liệu tìm kiếm để đưa vào mô hình mạnh mẽ và đơn giản đó. Theo tôi, các tìm kiếm Google hứa hẹn sẽ có nhiều giá trị hơn khi dùng để đo lường các bệnh mà dữ liệu hiện có vẫn còn yếu. Vậy là một cái gì đó kiểu như Google STD[2] sẽ có giá trị về lâu về dài hơn Google Flu.

Bài học thứ hai là, khi dự báo, bạn không cần lo lắng quá nhiều về lí do tại sao mô hình lại hiệu quả. Seder không thể giải thích đầy đủ cho tôi tại sao tâm thất trái lại rất quan trọng trong việc dự báo thành công của một con ngựa. Ông cũng không thể giải thích chính xác giá trị của lá

[1] [ND] Đạt mức bách phân vị 99.61 (99.61 percentile) nghĩa là Số 85 có tâm thất trái tốt hơn 99.61% số ngựa được khảo sát.

[2] [ND] STD là bệnh lây qua đường tình dục. Google STD (tên tác giả đặt) có thể là mô hình dự báo bệnh lây qua đường tình dục dựa trên tìm kiếm Google.

lách. Có lẽ một ngày nào đó các bác sĩ chuyên khoa tim và huyết học sẽ giải quyết các bí ẩn này. Nhưng bây giờ việc đó không quan trọng. Seder làm công việc dự báo, chứ không phải giải thích. Và, trong ngành dự báo, bạn chỉ cần biết cái gì hiệu quả, chứ không cần biết tại sao.

Ví dụ, Walmart dùng dữ liệu từ doanh số ở tất cả cửa hàng của họ để biết sản phẩm nào cần cho lên kệ. Trước Siêu bão Frances, trận bão hủy diệt đổ bộ vào vùng Đông Nam năm 2004, Walmart nghi ngờ—một cách chính xác—rằng thói quen mua sắm của mọi người có thể thay đổi khi thành phố sắp bị bão hoành hành. Họ nghiền ngẫm dữ liệu doanh số từ các trận bão trước để xem người ta muốn mua gì. Và đáp án là gì? Bánh Pop-Tarts dâu. Sản phẩm này bán nhanh gấp 7 lần bình thường trong những ngày sắp có bão.

Dựa trên phân tích đó, Walmart chất bánh Pop-Tarts dâu lên các xe tải theo đường Interstate 95 thẳng đến các cửa hàng nằm trên đường đi của bão. Và thực tế, lượng Pop-Tarts này bán rất chạy.

Tại sao lại là Pop-Tarts? Có thể vì bánh này không cần để lạnh hoặc nấu nướng. Tại sao lại là dâu? Không biết. Nhưng khi bão đổ bộ, người ta rõ ràng tìm tới Pop-Tarts dâu. Vì vậy vào những ngày trước bão, Walmart giờ đây thường cho lên kệ hàng chồng hộp Pop-Tarts dâu. Lí do của mối quan hệ này không quan trọng. Nhưng bản thân mối quan hệ là quan trọng. Có thể một ngày nào đó các nhà khoa học thực phẩm sẽ tìm ra sự liên hệ giữa các trận bão và các loại bánh nướng với mứt dâu. Nhưng trong khi chờ họ giải thích, Walmart vẫn cần trữ Pop-Tarts dâu tại các cửa hàng khi bão sắp đến và để dành bánh gạo Rice Krispies lại bán cho những ngày nắng ráo.

Bài học này cũng rất rõ ràng trong câu chuyện của Orley Ashenfelter, một nhà kinh tế học ở Princeton. Ashenfelter trong ngành rượu vang cũng giống như Seder trong ngành ngựa đua vậy.

Cách đây hơn một thập niên, Ashenfelter rất nản chí. Ông đã mua nhiều rượu vang đỏ từ vùng Bordeaux của Pháp. Thỉnh thoảng loại vang này rất ngon, xứng đáng với cái giá cao của nó. Tuy nhiên, có nhiều lần, nó lại chỉ là một nỗi thất vọng tràn trề.

Ashenfelter tự hỏi, tại sao mua rượu vang cùng một giá mà kết quả lại quá khác biệt như thế?

Một ngày kia, Ashenfelter nhận được lời khuyên từ một người bạn nhà báo am hiểu rượu vang. Thực ra, có một cách để biết một loại vang có ngon hay không. Người bạn cho Ashenfelter biết, chìa khóa chính là thời tiết trong mùa trồng nho.

Sự quan tâm của Ashenfelter được khơi gợi. Ông tiếp tục điều tra xem điều này đúng hay không, và để xem liệu có thể luôn mua được rượu vang ngon hơn hay không. Ông tải xuống 30 năm dữ liệu thời tiết vùng Bordeaux. Ông cũng thu thập giá cả các cuộc đấu giá rượu vang. Các cuộc đấu giá, diễn ra nhiều năm sau khi chai vang được bán lần đầu tiên, sẽ cho ta biết loại vang đó có ngon hay không.

Kết quả thật đáng kinh ngạc. Một tỉ lệ rất lớn chất lượng của chai vang có thể được giải thích đơn giản bởi thời tiết suốt mùa trồng nho.

Thực vậy, chất lượng một loại vang có thể được phân tách thành một công thức đơn giản, ta có thể gọi là *Định luật trồng nho thứ nhất*:

> Giá = 12.145 + 0.00117 lượng mưa mùa đông + 0.0614 nhiệt độ mùa trồng nho trung bình – 0.00386 lượng mưa lúc thu hoạch.

Vậy tại sao chất lượng rượu vang ở vùng Bordeaux hoạt động như thế này? Điều gì giải thích *Định luật trồng nho thứ nhất*? Có vài ý tưởng giải thích cho công thức rượu vang của Ashenfelter—nhiệt độ và lượng nước tưới giai đoạn đầu là cần thiết để nho chín đúng.

Nhưng chi tiết chính xác công thức dự báo của ông vượt xa bất cứ lí thuyết nào và chắc chắn sẽ không bao giờ được hiểu đầy đủ, ngay cả với các chuyên gia trong lĩnh vực.

Tại sao 1 cm mưa mùa đông tính trung bình lại giúp thêm 0.1 xu vào giá một chai vang đỏ đủ tuổi? Tại sao không phải 0.2 xu? Tại sao không phải 0.05? Không ai có thể trả lời các câu hỏi này. Nhưng nếu có thêm 1,000 cm mưa vào mùa đông, bạn phải sẵn sàng trả thêm 1 USD cho một chai vang.

Thực vậy, mặc dù không biết chính xác tại sao phép hồi quy của ông lại ra như thế, Ashenfelter vẫn dùng nó để mua rượu vang. Theo ông, "Nó cho kết quả tuyệt vời." Chất lượng của các loại vang ông uống được cải thiện đáng kể.

Nếu mục đích của bạn là dự báo tương lai—loại vang nào ngon, sản phẩm gì sẽ bán chạy, con ngựa nào sẽ chạy nhanh—bạn không cần phải quá lo lắng về việc tại sao mô hình của mình lại ra chính xác như thế. Chỉ cần tính đúng số là được. Đó là bài học thứ hai trong câu chuyện ngựa đua của Jeff Seder.

Bài học cuối cùng rút ra từ nỗ lực dự báo của Seder: Bạn phải tư duy thoáng và linh hoạt trong việc quyết định cái gì nên được coi là dữ liệu. Không có gì chứng tỏ các nhân viên chọn ngựa thời xưa mù tịt về dữ liệu trước khi Seder xuất hiện. Họ nghiên cứu kĩ các lần đua và biểu đồ dòng dõi. Cái hay của Seder là ở chỗ tìm kiếm các dữ liệu mà người khác trước đó chưa thấy, là ở chỗ cân nhắc sử dụng các nguồn dữ liệu phi truyền thống. Với nhà khoa học dữ liệu, góc nhìn tươi mới và độc đáo sẽ có thể mang đến thành công.

## Từ ngữ là dữ liệu

Một ngày năm 2004, hai nhà kinh tế học trẻ có chuyên môn trong mảng truyền thông, bấy giờ là các nghiên cứu sinh tiến sĩ tại Harvard, đang đọc về quyết định hợp pháp hóa hôn nhân đồng giới vừa diễn ra ở Massachusetts.

Hai nhà kinh tế học này, Matt Gentzkow và Jesse Shapiro, chú ý thấy có điều gì đó rất thú vị: 2 tờ báo sử dụng ngôn ngữ rất khác nhau để tường thuật cùng một câu chuyện. *Washington Times*, tờ báo nổi tiếng là bảo thủ, đặt tiêu đề câu chuyện: "Người đồng tính 'cưới nhau' ở Massachusetts" (Homosexuals 'Marry' in Massachusetts). *Washington Post*, tờ báo nổi tiếng là tự do, tường thuật rằng đã có một chiến thắng cho "các cặp đồng giới" (same-sex couples).

Không có gì ngạc nhiên khi các cơ quan báo chí khác nhau nghiêng theo những hướng khác nhau, khi các tờ báo thuật một câu chuyện giống

nhau nhưng tập trung vào những mục tiêu khác nhau. Thực vậy, nhiều năm trước đó, Gentzkow và Shapiro đã cân nhắc xem họ có thể dùng năng lực kinh tế học để giúp tìm hiểu khuynh hướng truyền thông hay không. Tại sao một số tờ báo dường như có góc nhìn tự do, và số khác thì lại bảo thủ hơn?

Nhưng Gentzkow và Shapiro không thực sự có ý tưởng nào về cách xử lí câu hỏi này; họ không biết làm sao đo lường được tính chủ quan truyền thông một cách hệ thống và khách quan.

Điều mà Gentzkow và Shapiro thấy thú vị về câu chuyện hôn nhân đồng giới không phải là việc các cơ quan báo chí đưa tin khác nhau, mà là *cách* họ đưa tin khác nhau: Mọi thứ bộc lộ qua sự khác biệt rõ ràng trong cách lựa chọn từ ngữ. Năm 2004, *Washington Times* dùng từ "homosexuals," một cách nói lỗi thời và miệt thị để mô tả người đồng tính; trái lại, *Washington Post* dùng cụm từ "same-sex couples," nhấn mạnh rằng các mối quan hệ đồng giới cũng chỉ là một hình thức tình cảm lãng mạn bình thường mà thôi.

Các học giả tự hỏi liệu ngôn ngữ có thể là chìa khóa để hiểu xu hướng hay không. Người tự do và người bảo thủ có hay dùng các cụm từ khác nhau không? Từ ngữ mà báo chí dùng trong các câu chuyện có thể biến thành dữ liệu không? Điều này có thể tiết lộ gì về báo chí Mĩ? Ta có thể biết mỗi tờ báo là tự do hay bảo thủ không? Và ta có thể biết tại sao không? Năm 2004, đây không phải là các câu hỏi vô ích. Hàng tỉ từ ngữ trong báo chí Mĩ không còn bị mắc kẹt trên báo in hoặc bản chụp microfilm nữa. Một số website bấy giờ đã ghi lại tất cả các từ chứa trong mỗi câu chuyện ở hầu như mọi tờ báo tại Mĩ. Gentzkow và Shapiro có thể nạo vét các trang này và nhanh chóng kiểm tra khả năng đo lường sự thiên lệch báo chí qua từ ngữ. Và, bằng cách này, ta có thể hiểu sâu sắc hơn về cách vận hành của truyền thông báo chí.

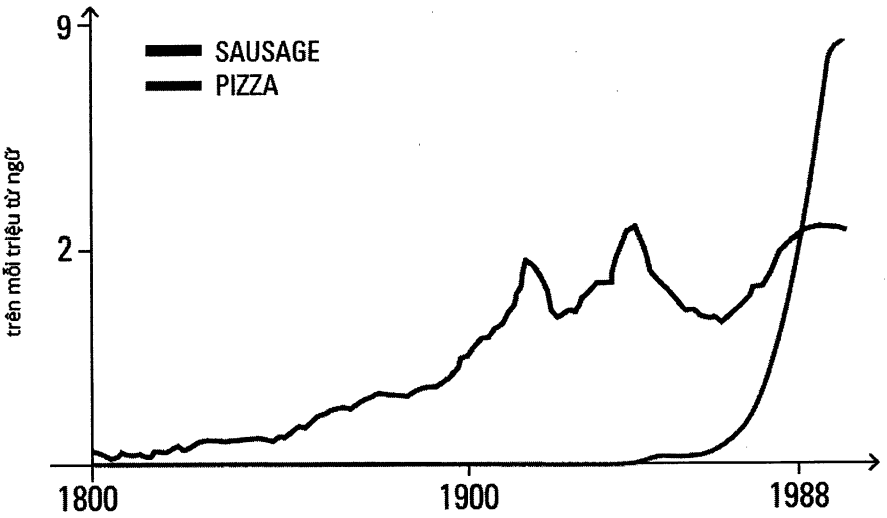
Nhưng trước khi mô tả những gì họ phát hiện, ta hãy tạm gác lại câu chuyện của Gentzkow và Shapiro để thảo luận cách các học giả trong nhiều lĩnh vực rộng lớn đã tận dụng loại dữ liệu mới này (từ ngữ) để hiểu rõ hơn bản chất con người.

Dĩ nhiên, ngôn ngữ xưa nay luôn là một chủ đề đáng quan tâm đối với các nhà khoa học xã hội. Tuy nhiên, nghiên cứu ngôn ngữ nói chung đòi hỏi phải đọc kĩ các văn bản, và việc biến các mảng văn bản đồ sộ thành dữ liệu là bất khả thi. Bây giờ, với máy tính và công nghệ số hóa, việc lập bảng từ ngữ từ các bộ tài liệu khổng lồ là rất dễ dàng. Như thế, ngôn ngữ đã trở thành đối tượng của phân tích Dữ Liệu Lớn. Các đường dẫn mà Google đã sử dụng được cấu thành từ các từ ngữ. Các tìm kiếm Google mà tôi nghiên cứu cũng vậy. Từ ngữ có mặt rất thường xuyên trong sách này. Ngôn ngữ quan trọng đối với cuộc cách mạng Dữ Liệu Lớn đến nỗi nó xứng đáng được dành một phần riêng. Thực vậy, bây giờ nó đang được sử dụng nhiều đến nỗi có cả một lĩnh vực dành cho nó: "văn bản là dữ liệu" (text as data).

Một bước tiến chủ chốt trong lĩnh vực này chính là Google Ngrams. Vài năm trước, 2 nhà sinh học trẻ, Erez Aiden và Jean-Baptiste Michel, cho các trợ lí đếm từng từ một trong các văn bản cũ và bụi bặm để cố phát hiện những hiểu biết mới về cách lan truyền của các cách dùng từ. Một ngày kia, Aiden và Michel nghe về một dự án mới của Google—số hóa một phần lớn sách vở của thế giới. Ngay lập tức, các nhà sinh học nhận ra rằng đây sẽ là một cách để hiểu lịch sử ngôn ngữ dễ dàng hơn nhiều.

"Chúng tôi nhận ra phương pháp của mình đã quá lỗi thời," Aiden nói với tạp chí *Discover*. "Rõ ràng là bạn không thể cạnh tranh với ông thần kĩ thuật số này." Vì vậy, họ quyết định cộng tác với Google. Với sự giúp đỡ của các kĩ sư Google, họ tạo ra một dịch vụ giúp tìm kiếm một từ hoặc cụm từ trong hàng triệu quyển sách đã được số hóa. Nó sẽ cho các nhà nghiên cứu biết tần số xuất hiện của từ hoặc cụm từ đó trong mỗi năm, từ 1800 đến 2010.

Vậy ta có thể biết được gì từ tần số mà các từ hoặc cụm từ xuất hiện trong sách ở những năm khác nhau? Đơn cử, ta đã biết thêm về sự phổ biến chậm chạp của sausage (xúc xích) và sự phổ biến nhanh chóng của pizza.




Nhưng còn có những bài học sâu sắc hơn điều đó rất nhiều. Chẳng hạn, Google Ngrams có thể cho ta biết bản sắc dân tộc hình thành như thế nào. Một ví dụ hấp dẫn được trình bày trong quyển sách *Uncharted* của Aiden và Michel.

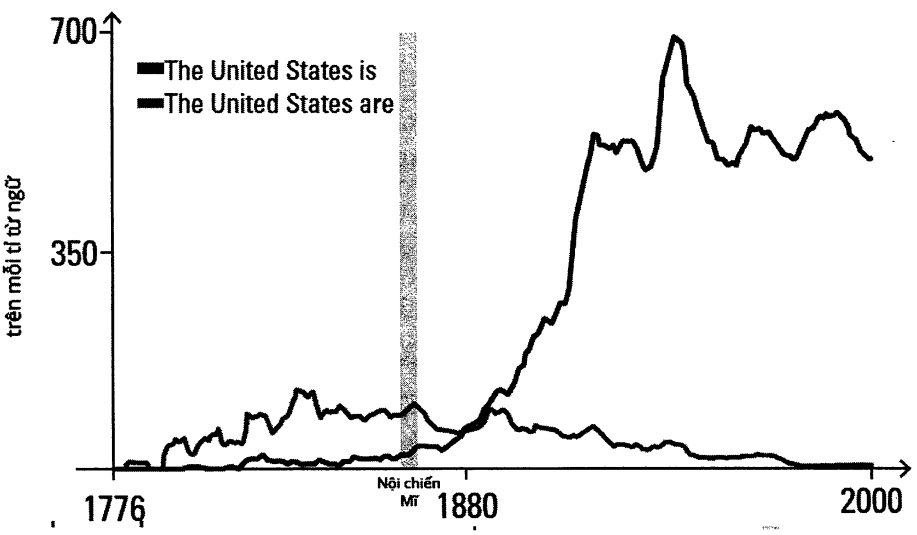
Trước hết, có một câu hỏi nhanh. Bạn nghĩ nước Mĩ đang là một quốc gia hòa hợp hay chia rẽ? Nếu giống hầu hết mọi người, bạn sẽ nói nước Mĩ bị chia rẽ trong những ngày này do mức độ phân cực chính trị cao. Thậm chí bạn có thể nói đất nước này trước giờ vẫn đã chia rẽ như thế rồi. Rõ ràng, nước Mĩ bây giờ được mã hóa bằng màu sắc: Các tiểu bang đỏ là Cộng hòa; các tiểu bang xanh là Dân chủ. Nhưng, trong *Uncharted,* Aiden và Michel lưu ý một điểm dữ liệu hấp dẫn, tiết lộ mức độ chia rẽ chính xác của nước Mĩ trong quá khứ. Điểm dữ liệu đó chính là ngôn ngữ mà người dân dùng để nói về nước Mĩ.

Chú ý từ ngữ tôi dùng khi nói về nước Mĩ. Tôi viết, "The United States *is* divided" (nước Mĩ bị chia rẽ). Tôi nói đến nước Mĩ như là một danh từ số ít. Điều này là tự nhiên, là cách dùng chuẩn mực và đúng ngữ pháp. Tôi dám chắc là các độc giả đọc bản tiếng Anh đã không để ý.

Tuy nhiên, không phải trước nay người Mĩ vẫn nói kiểu này. Trong những ngày đầu của đất nước, người Mĩ dùng hình thức số nhiều cho

United States. Ví dụ, John Adams, trong Thông điệp Liên bang 1799, đã viết "the United States in *their* treaties with his Britanic Majesty" (Hợp chủng quốc Hoa Kì trong các hiệp ước *của họ* với Hoàng đế Anh Quốc). Nếu sách của tôi mà viết trong năm 1800, tôi sẽ nói, "The United States *are* divided." Khác biệt nhỏ về cách dùng này từ lâu là một sự hấp dẫn đối với các sử gia, vì nó cho thấy đã có một điểm chuyển giao, thời điểm mà kể từ đó nước Mĩ thôi nghĩ mình là một tập hợp gồm các tiểu bang và bắt đầu nghĩ mình là một quốc gia thống nhất. Vậy điều này xảy ra khi nào? *Uncharted* viết rằng các sử gia chưa bao giờ chắc chắn, vì chưa có cách nào có hệ thống để kiểm tra điều đó. Nhưng từ lâu nhiều người nghi rằng nguyên nhân chính là Nội chiến Mĩ (1861-1865). Thực vậy, James McPherson, nguyên chủ tịch Hội sử học Mĩ và là người đoạt giải Pulitzer, nói thẳng tuột: "Cuộc chiến đánh dấu sự chuyển dịch cụm từ United States sang danh từ số ít."

Nhưng hóa ra McPherson đã sai. Google Ngrams đã cho Aiden và Miche một cách có hệ thống để kiểm tra điều này. Họ có thể thấy tần số các sách Mĩ dùng cụm từ "The United States are..." so với "The United States is..." cho từng năm trong lịch sử của đất nước. Sự thay đổi rất chậm, và sau khi Nội chiến kết thúc đã lâu thì mới tăng tốc.



Sau Nội chiến 15 năm, người ta vẫn còn dùng "The United States are..." nhiều hơn "The United States is... ," chứng tỏ đất nước vẫn còn bị chia rẽ về mặt ngôn ngữ học. Các chiến thắng quân sự diễn ra nhanh hơn những thay đổi trong đầu óc.

---

Bàn về chuyện hòa hợp quốc gia đến đây là đủ rồi. Vậy còn mức độ hòa hợp của một đôi nam nữ thì sao? Từ ngữ cũng có ích đấy.

Ví dụ, ta có thể dự báo một đôi nam nữ sẽ tiếp tục hẹn hò lần thứ hai dựa trên cách họ nói chuyện ở lần thứ nhất.

Điều này được chỉ ra bởi một nhóm liên ngành các nhà khoa học Stanford và Northwestern: Daniel McFarland, Dan Jurafsky, và Craig Rawlings. Họ đã nghiên cứu hàng trăm người hẹn hò nhanh khác giới và cố gắng quyết định xem điều gì dự báo việc họ sẽ cảm thấy có mối dây liên kết với nhau và muốn hẹn hò lần thứ hai.

Đầu tiên, họ dùng dữ liệu truyền thống. Họ hỏi những người hẹn hò về chiều cao, cân nặng, sở thích, và kiểm tra xem các yếu tố này liệu có tương quan với chuyện có cảm tình với ai hay không. Thông thường thì phụ nữ thích đàn ông cao hơn mình và cùng chung sở thích; nhìn chung thì nam giới thích phụ nữ gầy hơn mình và cùng chung sở thích. Chẳng có gì mới cả.

Nhưng các nhà khoa học này còn thu thập một loại dữ liệu mới. Họ yêu cầu những người hẹn hò mang máy ghi âm theo. Nội dung ghi âm các buổi hò hẹn sau đó được số hóa. Thế là các nhà khoa học có thể mã hóa từ ngữ được dùng, cũng như tiếng cười và tông giọng. Họ có thể kiểm tra các cặp nam nữ phát tín hiệu quan tâm ra sao và cả cách các cặp này quyến rũ đối phương nữa.

Vậy dữ liệu ngôn ngữ cho ta biết điều gì? Đầu tiên là cách nam giới hoặc nữ giới thể hiện mình đang quan tâm. Với nam, có một cách khá rõ ràng để phát tín hiệu mình bị hấp dẫn: Anh ta cười to với các câu nói đùa của bên nữ. Một cách khác ít rõ ràng hơn: Khi nói, anh ta hạn chế thay đổi cao độ giọng nói. Có nghiên cứu cho rằng giọng đều đều

thường được phụ nữ xem là nam tính. Vậy điều này ám chỉ rằng đàn ông, có lẽ trong tiềm thức, sẽ phóng đại vẻ nam tính của họ khi thích một người nữ.

Các nhà khoa học phát hiện rằng bên nữ phát tín hiệu quan tâm bằng cách thay đổi cao độ giọng, nói nhẹ nhàng hơn, và lượt nói ngắn hơn. Cũng có những đầu mối quan trọng cho thấy mức độ quan tâm của bên nữ dựa trên các từ ngữ mà cô ta dùng. Một cô gái có lẽ là không hứng thú lắm khi dùng các từ ngữ rào đón (hedge word) "chắc là" hoặc "em đoán là."

Các chiến hữu à, nếu phụ nữ rào đón các phát biểu của họ về bất cứ chủ đề nào—nếu cô nàng "thì cũng" thích đồ uống của mình hoặc "đại loại cũng" lạnh hoặc "chắc là" sẽ dùng một món khai vị khác—bạn có thể chắc rằng cô nàng "đại loại thì cũng chắc là" không khoái bạn rồi.

Các cô gái *nhiều khả năng* đang hứng thú khi nói về bản thân. Hóa ra, với một chàng trai đang muốn có người yêu, từ đẹp nhất mà anh chàng có thể nghe từ miệng của một cô gái là từ *"Em"*: Đó là một dấu hiệu cho thấy cô nàng đang cảm thấy thoải mái. Phụ nữ cũng có thể đang quan tâm nếu dùng các cụm từ gây chú ý như "Anh biết hông?" (Ya know?) và "Ý em muốn nói" (I mean). Tại sao? Các nhà khoa học nhận xét rằng các cụm từ này mời gọi sự chú ý của người nghe. Nó thân mật, ấm áp, và cho thấy người nói đang nghĩ đến chuyện kết giao, bạn có biết ý tôi muốn nói gì hông?

Vậy thì nam giới và nữ giới có thể giao tiếp như thế nào để đối phương thích mình? Dữ liệu cho biết rằng có nhiều cách nói để bên nam gia tăng cơ hội được nàng thích. Phụ nữ thích các anh đi theo sự dẫn dắt của họ. Không có gì đáng ngạc nhiên, các cô gái nhiều khả năng sẽ đổ nếu chàng trai cười với những câu nói đùa của cô, và tiếp tục chuyện trò về các chủ đề cô đưa ra chứ không thường xuyên thay đổi sang các chủ đề anh ta muốn nói.[1] Phụ nữ cũng thích các anh thể hiện sự ủng hộ và

[1] Một lí thuyết tôi đang nghiên cứu: Dữ Liệu Lớn chỉ xác nhận mọi thứ mà Leonard Cohen quá cố đã từng nói. Ví dụ, Leonard Cohen có lần cho đứa cháu trai mánh tán tỉnh phụ nữ sau đây: "Nghe nhiều vào. Rồi nghe thêm nữa. Và khi cháu nghĩ cháu nghe đã đời rồi, cứ nghe tiếp." Điều này có vẻ gần giống với những gì các nhà khoa học phát hiện.

cảm thông. Nếu anh chàng nói, "Hay quá!" hoặc "Hay ghê," nàng sẽ rất có khả năng đổ trước chàng. Kết quả cũng tương tự nếu anh chàng dùng các cụm từ như "Khó khăn thật" hoặc "Chắc em buồn lắm."

Đối với nữ giới, có một ít tin xấu đây, vì dữ liệu có vẻ như xác nhận một sự thật khó chấp nhận về nam giới. Chuyện trò chỉ đóng một vai trò nhỏ trong cách họ phản ứng với phụ nữ. Dáng vẻ bên ngoài của cô gái quan trọng hơn tất cả các thứ khác. Dù vậy, có một từ mà bên nữ có thể dùng để cải thiện đôi chút cơ hội bên nam thích mình, đó là từ chúng ta đã biết: "Em." Nam giới rất có thể sẽ có cảm tình với một phụ nữ hay nói về bản thân. Và như đã nói ở đoạn trước, phụ nữ cũng có nhiều khả năng sẽ đổ sau một cuộc hẹn hò mà ở đó cô nàng được nói về bản thân. Như vậy, nếu trong cuộc hẹn đầu tiên mà có xuất hiện một cuộc bàn luận về cô gái, thì đó quả là một dấu hiệu tuyệt vời. Người nữ phát tín hiệu là đang thấy thoải mái và trân trọng việc người nam không lấn sân cuộc trò chuyện. Còn người nam thì thích việc người nữ trải lòng. Trong trường hợp này, cuộc hẹn thứ hai là chắc chắn.

Cuối cùng, có một chỉ báo tình trạng rắc rối rõ ràng trong "biên bản" hẹn hò: dấu chấm hỏi. Nếu có nhiều câu hỏi đặt ra tại một cuộc hẹn hò, ít có khả năng cả hai bên nam nữ sẽ đổ nhau. Điều này có vẻ phản trực giác; ta có thể nghĩ rằng các câu hỏi là dấu hiệu của sự quan tâm. Nhưng không phải vậy ở cuộc hẹn hò đầu tiên. Ở cuộc hẹn đầu, hầu hết các câu hỏi là dấu hiệu của sự buồn chán. "Anh có sở thích gì không?" "Em có bao nhiêu anh chị em?" Đây là những thứ người ta hay nói khi không có chuyện gì để nói. Cuộc hẹn hò đầu tiên tuyệt vời có khi chỉ có một câu hỏi ở cuối: "Em có muốn đi chơi với anh không?" Nếu đây là câu hỏi duy nhất trong cuộc hẹn hò, câu trả lời chắc chắn là "Có chứ."

Nam và nữ không chỉ nói chuyện khác nhau khi đang tán tỉnh nhau. Nhìn chung, bình thường họ cũng nói chuyện theo cách khác nhau nữa.

Một nhóm các nhà tâm lí học phân tích từ ngữ được dùng trong hàng trăm ngàn bài đăng tải trên Facebook. Họ đo tần số của mọi từ ngữ được cả nam và nữ sử dụng. Sau đó họ có thể xác định các từ nào là nam tính nhất và các từ nào là nữ tính nhất trong ngôn ngữ Anh.

Đa số các sở thích dùng từ này đều hiển nhiên. Ví dụ, nữ giới nói về "mua sắm" và "tóc của tôi" thường xuyên hơn nam giới rất nhiều. Nam giới nói về "bóng đá" và "Xbox" thường xuyên hơn nữ giới rất nhiều. Có lẽ bạn không cần một nhóm nhà tâm lí học phân tích Dữ Liệu Lớn để nói cho bạn biết điều đó.

Tuy nhiên, có một số những phát hiện thú vị hơn. Nữ giới dùng từ "tomorrow" (ngày mai) thường hơn nam giới rất nhiều, có lẽ vì nam giới không giỏi nghĩ về tương lai cho lắm. Thêm chữ cái "o" vào từ "so" là một trong các đặc điểm ngôn ngữ nữ tính nhất. Trong số các từ được nữ giới dùng nhiều nhất có các từ "soo," "sooo," "soooo," "sooooo," và "soooooo."

Có lẽ tuổi thơ của tôi tiếp xúc nhiều với những phụ nữ mà thỉnh thoảng cũng không ngại chửi thề. Nhưng tôi trước giờ cứ nghĩ chửi rủa là một việc khá bình đẳng. Không phải vậy. Một số từ như "f*ck," "sh*t," "f*cks," "bullsh*t," "f*cking," và "f*ckers" được nam dùng thường xuyên hơn nữ rất nhiều.

Đây là các đám mây từ (word cloud) chỉ các từ chủ yếu được nam giới dùng và các từ chủ yếu được nữ giới dùng. Từ càng lớn ở tranh của phái nào, thì càng được riêng phái đó dùng nhiều.

Nam giới

Nữ giới

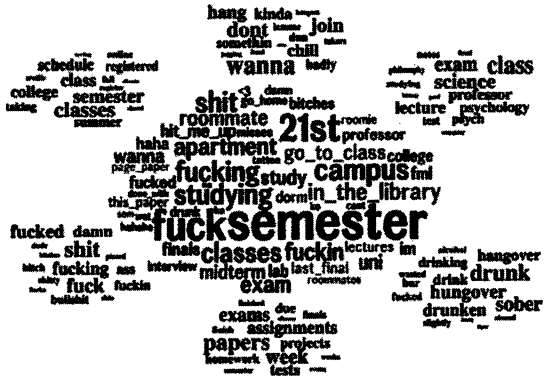

Điều tôi thích ở nghiên cứu này là dữ liệu mới thông tin cho ta biết về các mô hình đã tồn tại từ lâu nhưng ta lại không để ý đến. Nam giới và nữ giới luôn nói theo các cách khác nhau. Nhưng, suốt hàng chục ngàn năm, dữ liệu này biến mất ngay khi sóng âm phai mờ đi trong không gian. Bây giờ dữ liệu này đã được bảo quản trên máy tính và có thể được máy tính phân tích.

Hoặc có lẽ để theo đúng giới tính của tôi, tôi phải nói theo phong cách như sau: "Ngôn ngữ đã từng biến *cmn* mất tiêu. Bây giờ tụi mình có thể tạm nghỉ coi đá bóng và nghỉ chơi Xbox một lát để học cái *qq* này đi. Nói chung là vậy, nếu có thằng *** nào đó quan *cmn* tâm."

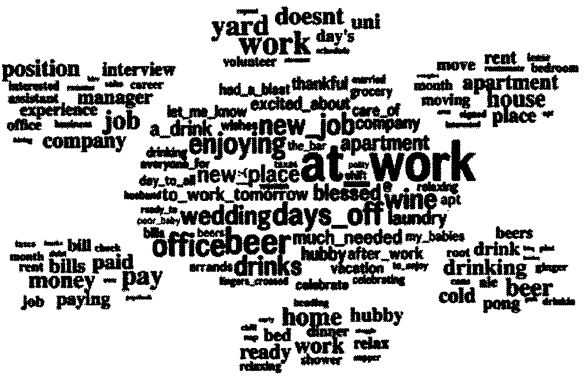
Không chỉ nam giới và nữ giới nói khác nhau. Người ta cũng bắt đầu đổi từ ngữ khi lớn tuổi. Điều này có thể cho ta vài đầu mối về tiến trình lão hóa. Dưới đây, cũng trích từ nghiên cứu trên, là các từ được dùng bất cân xứng nhất bởi các độ tuổi khác nhau trên Facebook. Tôi gọi biểu đồ này là "Nhậu nhẹt. Làm việc. Cầu nguyện." Tuổi teen, người ta nhậu nhẹt. Tuổi hăm, người ta làm việc. Tuổi băm trở lên, người ta cầu nguyện.

## NHẬU NHẸT. LÀM VIỆC. CẦU NGUYỆN.

19-22 tuổi




23-29 tuổi




30-65 tuổi

Có một công cụ mới và mạnh để phân tích văn bản gọi là phân tích cảm xúc (sentiment analysis). Các nhà khoa học bây giờ có thể ước tính một đoạn văn bản cụ thể vui hoặc buồn tới mức nào.

Bằng cách nào? Các nhóm nhà khoa học đã nhờ rất nhiều người mã hóa hàng chục ngàn từ trong ngôn ngữ Anh thành tích cực hoặc tiêu cực. Các từ tích cực nhất, theo phương pháp này, gồm có "happy," "love," và "awesome." Các từ tiêu cực nhất có "sad," "death," và "depression." Theo cách này, họ đã xây dựng một chỉ số tâm trạng của một tập hợp từ khổng lồ.

Khi dùng chỉ số này, họ có thể đo lường tâm trạng trung bình của các từ trong một đoạn văn bản. Nếu ai đó viết "I am happy and in love and feeling awesome," bản phân tích cảm xúc sẽ mã hóa đó là văn bản cực vui. Nếu ai đó viết "I am sad thinking about all the world's death and depression," bản phân tích cảm xúc sẽ mã hóa đó là văn bản cực buồn. Các đoạn văn bản khác sẽ nằm đâu đó ở khoảng giữa.

Vậy bạn có thể biết được gì khi mã hóa tâm trạng của văn bản? Các nhà khoa học dữ liệu Facebook đã chỉ ra một khả năng thú vị. Họ có thể ước tính chỉ số GNH (Gross National Happiness—tổng hạnh phúc quốc gia) của một nước mỗi ngày. Nếu các thông điệp trạng thái của mọi người có khuynh hướng tích cực, đất nước đó được cho là vui vào ngày hôm đó. Nếu các thông điệp có khuynh hướng tiêu cực, đất nước đó được cho là buồn vào ngày hôm đó.

Một trong số phát hiện của các nhà khoa học dữ liệu Facebook: Giáng sinh là một trong những ngày vui nhất của năm. Tôi đã từng nghi ngờ bản phân tích này—và bây giờ tôi hơi nghi ngờ cả dự án này. Nói chung, tôi nghĩ nhiều người âm thầm buồn vào lễ Giáng sinh vì họ cô đơn hoặc đang cự cãi với gia đình. Nói chung hơn nữa, tôi có khuynh hướng không tin các bài đăng Facebook, lí do gì thì tôi sẽ thảo luận rõ hơn trong chương tới—nói sơ là do xu hướng nói dối về cuộc sống của chúng ta trên mạng xã hội.

Nếu đơn lẻ một mình và khốn khổ vào lễ Giáng sinh, bạn có thực sự muốn quấy rầy tất cả bạn bè khi đăng lên là bạn bất hạnh như thế nào

không? Tôi nghĩ có nhiều người phải trải qua một Giáng sinh buồn mà vẫn đăng lên Facebook là họ cảm thấy rất biết ơn "cuộc sống tuyệt vời, cực khủng, ấn tượng, hạnh phúc" của mình. Bấy giờ, bài đăng của họ được mã hóa và nâng GNH của nước Mĩ lên. Nếu muốn mã hóa GNH, ta nên dùng nhiều nguồn hơn là chỉ các bài đăng Facebook.

Nói là vậy, chuyện Giáng sinh, xét toàn cục, là một dịp vui xem ra cũng hợp lí. Các tìm kiếm Google về sự trầm cảm và các khảo sát Gallup cũng cho ta biết rằng Giáng sinh nằm trong số những ngày vui nhất của năm. Và, trái với lời đồn, số vụ tự tử giảm mạnh quanh các ngày lễ. Ngay cả nếu có một số người buồn bã và cô đơn vào lễ Giáng sinh, số người vui vẫn đông hơn nhiều.

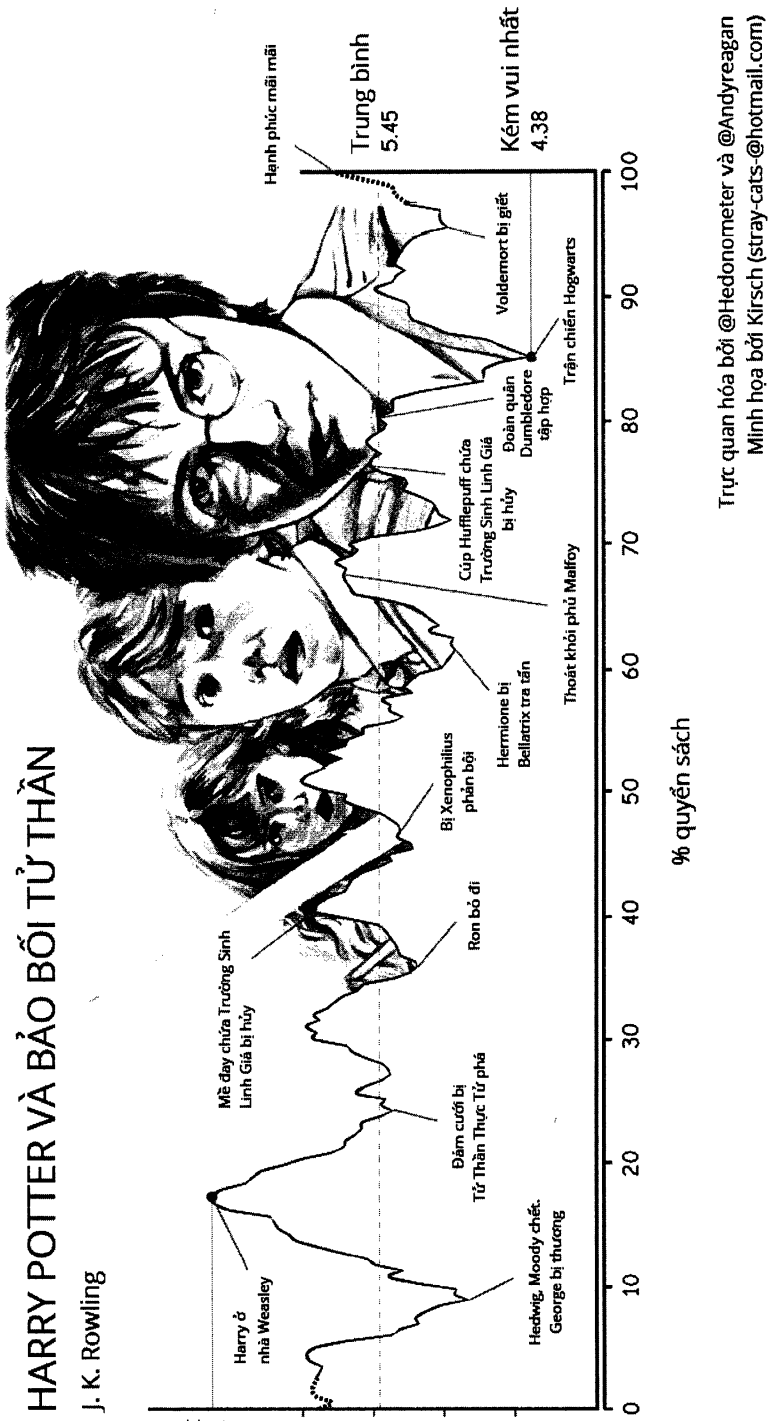
Ngày nảy ngày nay, khi người ta ngồi xuống để đọc, hầu hết thời gian được dành cho việc đọc kĩ các bài đăng Facebook. Nhưng, ngày xửa ngày xưa, cách đây không lâu lắm, người ta đọc truyện, đôi khi là đọc từ trong sách. Ở đây, phân tích cảm xúc cũng có thể cho ta biết nhiều điều.

Một nhóm nhà khoa học do Andy Reagan lãnh đạo—nay thuộc Học viện thông tin UC Berkeley—đã tải về văn bản từ hàng ngàn quyển sách và kịch bản phim. Sau đó họ có thể mã hóa mức độ vui hoặc buồn ở mỗi thời điểm của câu chuyện.

Ví dụ, xem xét quyển *Harry Potter and the Deathly Hallows (Harry Potter và bảo bối tử thần)*. Dưới đây là biểu đồ mô tả sự thay đổi tâm trạng câu chuyện, kèm theo các điểm chính trong diễn biến. Biểu đồ được lập bởi nhóm nhà khoa học trên.

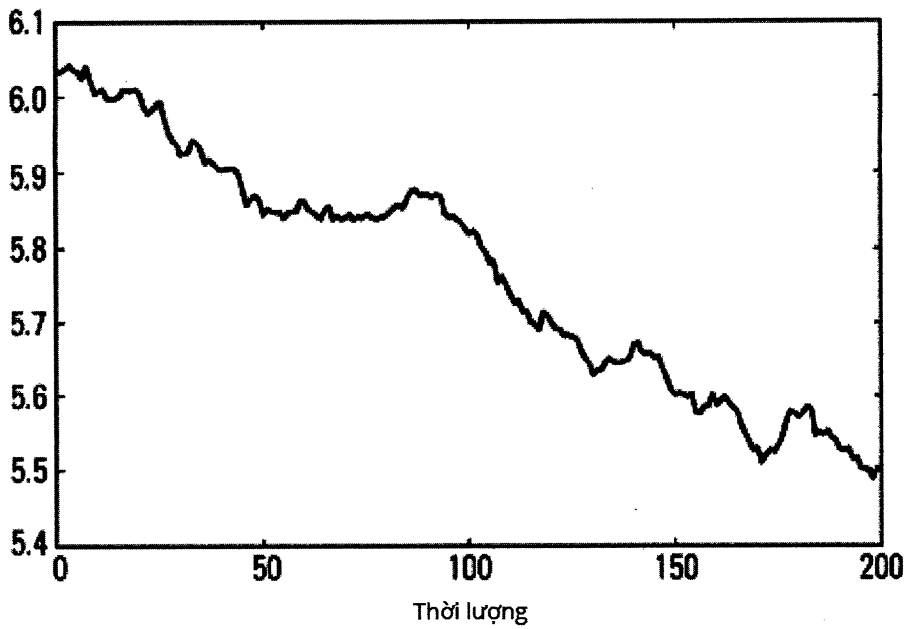
Chú ý rằng có nhiều chỗ lên xuống trong tâm trạng mà bản phân tích cảm xúc đã phát hiện ăn khớp với các sự kiện chính.

Hầu hết các câu chuyện có cấu trúc đơn giản hơn. Bi kịch *King John* của Shakespeare chẳng hạn. Trong vở kịch này, không có gì tốt đẹp cả. Vua John nước Anh bị yêu cầu từ bỏ ngai vàng. Ông bị rút phép thông công vì không vâng lời Giáo hoàng. Chiến tranh nổ ra. Cháu trai ông chết, có lẽ do tự tử. Những người khác chết. Cuối cùng, John bị một tu sĩ bất bình đầu độc.




Trực quan hóa bởi @Hedonometer và @Andyreagan
Minh họa bởi Kirsch (stray-cats-@hotmail.com)

Và đây là biểu đồ phân tích cảm xúc theo tiến triển của vở kịch *King John*.



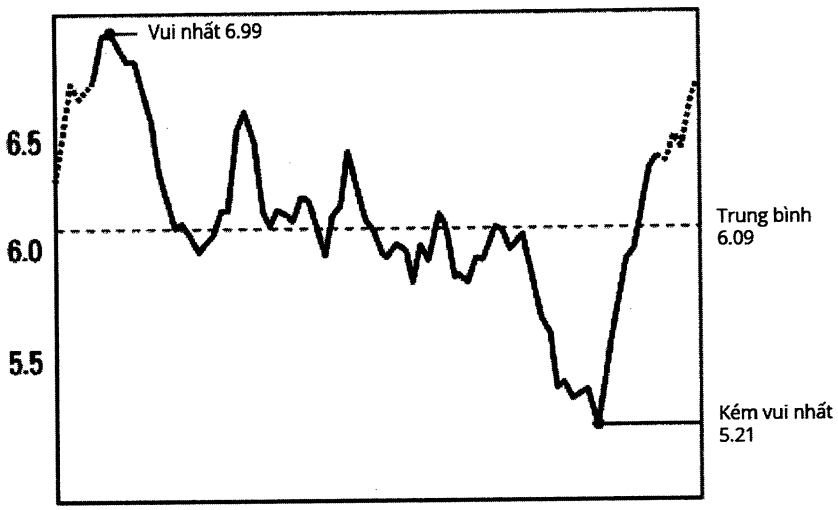

Nói cách khác, chỉ bằng cách phân tích từ ngữ, máy tính đã dò ra được rằng mọi thứ đi từ xấu đến xấu hơn, và rồi là xấu nhất.

Hoặc thử xem xét phim *127 Hours*. Tóm tắt cốt truyện cơ bản của phim này như sau:

> Một vận động viên đến Công viên quốc gia Canyonlands ở Utah để leo núi. Anh kết bạn với những người leo núi khác nhưng rồi chia tay với họ. Bất ngờ, anh trượt chân, bàn tay và cổ tay bị mắc kẹt trong một tảng đá. Anh tìm mọi cách để thoát ra nhưng đều thất bại. Anh hết hi vọng. Cuối cùng, anh cắt bỏ cánh tay và thoát ra. Anh cưới vợ, bắt đầu một gia đình, và tiếp tục leo núi, tuy nhiên bây giờ anh luôn để lại lời nhắn bất cứ khi nào anh đi.

Và đây là bản phân tích cảm xúc theo tiến triển của phim, cũng do nhóm Reagan thực hiện.

**127 Hours** - Danny Boyle




Các nhà khoa học máy tính phát hiện rằng một tỉ lệ lớn các câu chuyện khớp với 1 trong 6 cấu trúc tương đối đơn giản. Theo nhóm Reagan, đó là:

- Rách rưới rồi rủng rỉnh (thăng)
- Rủng rỉnh rồi rách rưới (trầm)
- Người rừng (trầm, rồi thăng)
- Icarus (thăng, rồi trầm)
- Lọ Lem (thăng, rồi trầm, rồi thăng)
- Oedipus (trầm, rồi thăng, rồi trầm)

Có thể có những ngoắt ngoéo nhỏ không được nêu trong biểu đồ đơn giản này. Ví dụ, *127 Hours* được xếp vào nhóm *người rừng*, tuy rằng có những lúc trong đoạn trầm, cảm xúc vẫn tạm thời đi lên. Cấu trúc lớn, bao quát của hầu hết câu chuyện khớp với 1 trong 6 thể loại. *Harry Potter và bảo bối tử thần* là một ngoại lệ.

Có nhiều câu hỏi khác ta có thể trả lời. Ví dụ, cấu trúc các câu chuyện đã thay đổi như thế nào theo thời gian? Các câu chuyện có trở nên phức tạp hơn không? Các nền văn hóa có khác nhau về loại câu chuyện không? Người ta thích loại câu chuyện nào nhất? Cấu trúc câu chuyện

khác nhau có hấp dẫn nam giới và nữ giới theo cách khác nhau không? Người dân ở các nước khác nhau thì thế nào?

Cuối cùng, dữ liệu văn bản cho ta những hiểu biết chưa từng có về điều khán giả thực sự muốn, nó có thể khác với những gì các tác giả hoặc các nhà quản trị vẫn nghĩ. Đã có một số đầu mối theo hướng này.

Ta hãy cùng xem xét một nghiên cứu của 2 giáo sư Wharton, Jonah Berger và Katherine L. Milkman, về loại câu chuyện thường được chia sẻ. Họ kiểm tra xem chuyện tích cực hay tiêu cực có nhiều khả năng vào danh sách được email nhiều nhất của *New York Times* hơn. Họ tải về mọi bài báo *Times* trong một giai đoạn dài 3 tháng. Dùng phân tích cảm xúc, các giáo sư mã hóa tâm trạng các bài báo. Ví dụ về các câu chuyện tích cực có: "Những người mới đến mắt mở to vì phải lòng Thành phố" và "Giải Tony cho lòng từ thiện." Các câu chuyện như "Những lời đồn trên mạng gắn với vụ tự tử của nữ diễn viên Hàn Quốc" và "Đức: Người cho Gấu Bắc cực con ăn đã tử vong" rõ ràng là tiêu cực.

Các giáo sư còn có thông tin về nơi bố trí câu chuyện. Có phải nó nằm ở trang chủ không? Ở phía trên bên phải? Trên bên trái? Và họ có cả thông tin về thời gian câu chuyện được đăng lên. Khuya thứ Ba? Sáng thứ Hai?

Họ có thể so sánh 2 bài báo—một bài tích cực, một bài tiêu cực—xuất hiện trong cùng một vị trí trên trang báo *Times*, được đăng cùng một thời gian, và xem bài nào nhiều khả năng được gởi email hơn.

Vậy cái gì được chia sẻ nhiều hơn, bài báo tích cực hay tiêu cực?

Là các bài báo tích cực. Như các tác giả kết luận, "Nội dung càng tích cực càng có khả năng lan truyền."

Điều này có vẻ trái ngược với suy nghĩ thông thường về báo chí, rằng người ta bị hấp dẫn bởi các câu chuyện bạo lực và thảm họa. Có thể đúng là truyền thông báo chí cung cấp cho người ta nhiều câu chuyện đen tối. Có một bí quyết trong giới làm tin tức, "Tin chảy máu, tin chạy mau" (If it bleeds, it leads). Tuy nhiên, nghiên cứu của các giáo sư Wharton cho rằng người ta thực sự có thể muốn các chuyện vui hơn.

Nghiên cứu ấy cho ta một bí quyết mới: "Tin dễ cười, tin dễ gửi" (If it smiles, it's emailed). Nghe cũng vần đấy chứ?

---

Chuyện văn bản buồn vui thế là đủ rồi. Giờ ta có thể nhận ra từ ngữ thuộc phe tự do hay phe bảo thủ bằng cách nào? Và điều đó cho ta biết gì về truyền thông báo chí hiện đại? Việc này phức tạp hơn một chút, nó đưa ta trở lại với Gentzkow và Shapiro. Chắc bạn còn nhớ, họ là các nhà kinh tế học đã phát hiện rằng hôn nhân đồng giới được mô tả theo cách khác nhau trong 2 tờ báo khác nhau và tự hỏi liệu có thể dùng ngôn ngữ để phát hiện xu hướng chính trị hay không.

Điều đầu tiên 2 học giả trẻ đầy tham vọng này đã làm là khảo sát các biên bản họp Quốc hội. Vì biên bản này đã được số hóa, họ có thể tải về mọi từ ngữ được dùng bởi mọi nghị sĩ đảng Dân chủ năm 2005 và mọi từ ngữ được dùng bởi mọi nghị sĩ đảng Cộng hòa năm 2005. Sau đó, họ có thể xem liệu có chuyện một cụm từ nào đó thường được dùng nhiều hơn hẳn bởi người đảng Dân chủ hay người đảng Cộng hòa hay không.

Quả thực là có. Dưới đây là vài ví dụ.

| CỤM TỪ ĐƯỢC ĐẢNG DÂN CHỦ DÙNG NHIỀU HƠN HẲN | CỤM TỪ ĐƯỢC ĐẢNG CỘNG HÒA DÙNG NHIỀU HƠN HẲN |
|---|---|
| Thuế tài sản thừa kế | Thuế thừa kế tài sản |
| Tư nhân hóa an sinh xã hội | Cải cách an sinh xã hội |
| Rosa Parks | Saddam Hussein |
| Các quyền người lao động | Các quyền tài sản tư nhân |
| Người nghèo | Chi tiêu chính phủ |

Điều gì giải thích những khác biệt trong ngôn ngữ này?

Đôi khi người Dân chủ và người Cộng hòa dùng từ ngữ diễn tả khác nhau cho cùng một khái niệm. Năm 2005, phe Cộng hòa cố gắng cắt thuế thừa kế liên bang. Họ có xu hướng mô tả nó là "death tax" (thuế tử

vong) (nghe như thuế áp lên người mới qua đời). Phe Dân chủ mô tả nó là "estate tax" (thuế tài sản) (nghe như thuế áp lên người giàu). Tương tự, đảng Cộng hòa cố gắng chuyển An sinh xã hội vào nhóm tài khoản hưu trí cá nhân. Với đảng Cộng hòa, đây là "cải cách." Với đảng Dân chủ, đây là "sự tư nhân hóa [tài sản công]"—nghe nguy hiểm hơn nhiều.

Đôi khi khác biệt trong ngôn ngữ là do nhấn mạnh. Người Cộng hòa và người Dân chủ đều rất kính trọng Rosa Parks, vị anh hùng về quyền công dân. Nhưng người Dân chủ nói về bà thường xuyên hơn. Cũng như thế, người Dân chủ và người Cộng hòa đều nghĩ rằng Saddam Hussein, cựu lãnh đạo Iraq, là một kẻ độc tài xấu xa. Nhưng người Cộng hòa thường đề cập ông ta trong nỗ lực thanh minh cho Cuộc chiến Iraq. Tương tự, "quyền người lao động" và quan tâm "người nghèo" là các nguyên tắc cốt lõi của đảng Dân chủ. "Quyền tài sản tư nhân" và cắt "chi tiêu chính phủ" là các nguyên tắc cốt lõi của đảng Cộng hòa.

Và những khác biệt này trong từ ngữ là đáng kể. Ví dụ, năm 2005, các nghị sĩ Cộng hòa dùng cụm từ "death tax" 365 lần và "estate tax" chỉ 46 lần. Với các nghị sĩ Dân chủ thì ngược lại. Họ dùng cụm từ "death tax" chỉ 35 lần và "estate tax" 195 lần.

Và nếu từ ngữ có thể cho ta biết một nghị sĩ là Dân chủ hay Cộng hòa, các học giả đã nhận thấy, chúng cũng có thể cho ta biết một tờ báo là tả khuynh hay hữu khuynh. Giống như các nghị sĩ Cộng hòa, các báo bảo thủ cũng rất thường dùng cụm từ "death tax" để thuyết phục những người phản đối nó. Tờ *Washington Post* tương đối tự do đã dùng cụm từ "estate tax" nhiều gấp 13.7 lần cụm từ "death tax." Tờ *Washington Times* bảo thủ dùng "death tax" và "estate tax" gần như ngang nhau.

Nhờ sự kì diệu của Internet, Gentzkow và Shapiro có thể phân tích ngôn ngữ được dùng trong rất nhiều tờ báo của cả nước. Các học giả sử dụng 2 website, newslibrary.com và proquest.com—các trang này đã cùng nhau số hóa 433 tờ báo. Sau đó, họ đếm tần số 1,000 cụm từ mang nội dung chính trị được dùng trong các tờ báo để đo lường khuynh hướng chính trị của các báo đó. Tờ báo tự do nhất theo thước đo này chính là *Philadelphia Daily News;* bảo thủ nhất: *Billings* (Montana) *Gazette.*

Khi đã có thước đo toàn diện đầu tiên về sự thiên lệch của truyền thông ở một lượng lớn các kênh, ta có thể trả lời câu hỏi quan trọng nhất về báo chí: Tại sao một số xuất bản phẩm tả khuynh còn số khác lại hữu khuynh?

Hai nhà kinh tế học nhanh chóng nhắm vào một yếu tố then chốt: quan điểm chính trị của địa phương. Với một vùng nhìn chung là tự do, như Philadelphia và Detroit, tờ báo lớn ở đó có khuynh hướng tự do. Nếu là vùng bảo thủ hơn, như Billings và Amarillo, Texas, tờ báo lớn ở đó có khuynh hướng bảo thủ. Nói cách khác, bằng chứng chỉ rõ rằng báo chí có khuynh hướng cung cấp những gì độc giả muốn.

Người ta hay nghĩ là chủ tờ báo sẽ có ảnh hưởng đến khuynh hướng đưa tin, nhưng như một quy luật, người sở hữu có ít ảnh hưởng lên khuynh hướng chính trị của tờ báo hơn ta nghĩ. Hãy chú ý điều diễn ra khi cùng một người hoặc một công ty sở hữu các tờ báo ở những thị trường khác nhau. Xem xét Công ty New York Times. Họ sở hữu tờ báo mà Gentzkow và Shapiro thấy là nghiêng về tự do: *New York Times*, một tờ báo có trụ sở tại Thành phố New York, nơi có khoảng 70% dân số là phe Dân chủ. Vào thời điểm nghiên cứu, công ty này cũng sở hữu tờ báo nghiêng về phía bảo thủ: *Spartanburg Herald-Journal*, trụ sở tại Spartanburg, South Carolina, nơi có khoảng 70% dân số là phe Cộng hòa. Dĩ nhiên, có những ngoại lệ: News Corporation của Rupert Murdoch sở hữu tờ báo mà hầu như ai cũng thấy là bảo thủ—*New York Post*. Nhưng, trên tổng thể, các phát hiện cho thấy rằng thị trường—chứ không phải chủ sở hữu—quyết định xu hướng của tờ báo.

Nghiên cứu này tác động sâu sắc đến cách chúng ta nghĩ về truyền thông báo chí. Nhiều người nghĩ báo chí Mĩ bị những con người hoặc tập đoàn giàu có kiểm soát nhằm gây ảnh hưởng đến quần chúng, dẫn dắt người ta đến với quan điểm chính trị của họ. Tuy nhiên, bài báo của Gentzkow và Shapiro cho rằng đây không phải là động cơ chủ chốt của các ông chủ. Trái lại, các ông chủ báo Mĩ chủ yếu cung cấp những gì quần chúng muốn để các ông giàu có hơn nữa mà thôi.

À, còn một câu hỏi nữa—một câu hỏi lớn, nhiều tranh cãi, và có lẽ khiêu khích hơn. Truyền thông báo chí Mĩ, nói chung, tả khuynh hay hữu khuynh? Truyền thông nói chung là tự do hay bảo thủ?

Gentzkow và Shapiro phát hiện rằng báo chí nhìn chung là tả khuynh. Báo chí nói chung, xét theo ngôn từ, giống với nghị sĩ đảng Dân chủ hơn là nghị sĩ đảng Cộng hòa.

"À ha!" các độc giả bảo thủ chắc đang muốn hét lên: "Tôi đã nói rồi mà!" Nhiều người bảo thủ từ lâu đã nghi ngờ báo chí có xu hướng cố gắng điều khiển quần chúng ủng hộ các quan điểm cánh tả. Không phải vậy. Thực ra, sự thiên lệch theo hướng tự do được hiệu chỉnh theo những gì người đọc báo muốn. Nhìn chung thì độc giả báo chí hơi tả khuynh. (Họ có dữ liệu về điều đó.) Và báo chí nói chung cũng hơi tả khuynh để cung cấp các quan điểm mà độc giả đòi hỏi.

Không có âm mưu lớn nào hết. Chỉ đơn thuần là chủ nghĩa tư bản mà thôi.

Kết quả của Gentzkow và Shapiro ám chỉ rằng truyền thông báo chí vẫn thường hoạt động như mọi ngành khác trên hành tinh này. Giống như các siêu thị tìm hiểu người ta muốn loại kem gì và trưng lên kệ loại kem đó, báo chí cũng tìm hiểu người ta muốn loại quan điểm nào và đưa loại quan điểm đó lên các trang báo. "Đó chỉ là một ngành kinh doanh," Shapiro nói với tôi. Đó là những gì ta có thể học được khi phân tách và định lượng các thứ có nhiều thành tố xoắn vào nhau như tin tức, phân tích, và ý kiến ra thành các bộ phận cấu thành: từ ngữ.

## Hình ảnh là dữ liệu

Theo truyền thống, khi giới học thuật và thương gia muốn có dữ liệu, họ tiến hành khảo sát. Dữ liệu đến gọn gàng, được rút ra từ các con số hoặc các ô chọn trong khảo sát. Cách này không còn đúng nữa. Những ngày của dữ liệu có cấu trúc, sạch sẽ, đơn giản, dựa trên khảo sát đã qua rồi. Trong thời đại mới này, các dấu vết lộn xộn chúng ta bỏ lại khi đi qua cuộc sống đang dần trở thành nguồn dữ liệu hàng đầu.

Như ta đã thấy, từ ngữ là dữ liệu. Cú nhấp chuột là dữ liệu. Đường dẫn là dữ liệu. Lỗi đánh máy là dữ liệu. Chuối trong giấc mơ là dữ liệu. Giọng nói là dữ liệu. Thở khò khè là dữ liệu. Nhịp tim là dữ liệu. Kích thước lá lách là dữ liệu. Các tìm kiếm, theo tôi, là dữ liệu tiết lộ nhiều điều nhất.

Và hóa ra, hình ảnh cũng là dữ liệu.

Giống như từ ngữ, một thời bị giới hạn chỉ tồn tại trong các sách và tạp chí trên các kệ sách bụi bặm để rồi giờ đây đã được số hóa, hình ảnh cũng được giải phóng khỏi các tập album và hộp các-tông. Chúng cũng đã được chuyển thành mảnh nhỏ và đưa lên đám mây. Và khi văn bản có thể cho ta các bài học lịch sử—ví dụ như sự thay đổi trong cách nói, thì hình ảnh cũng có thể cho ta các bài học lịch sử—ví dụ như sự thay đổi trong cách người ta đứng tạo dáng chụp hình.

Ta hãy cùng xem xét một nghiên cứu rất tài tình của 4 nhà khoa học máy tính tại Brown và Berkeley. Họ đã lợi dụng một tiện ích của thời đại số: Nhiều trường trung học đã quét hình các quyển niên giám của họ và lưu trực tuyến. Qua Internet, các nhà nghiên cứu phát hiện 949 quyển niên giám đã được quét của các trường trung học Mĩ giai đoạn 1905-2013. Dữ liệu này gồm hàng chục ngàn ảnh chân dung học sinh cuối cấp. Dùng phần mềm máy tính, họ đã tạo ra được một khuôn mặt "chung" từ các hình ảnh mỗi thập niên. Nói cách khác, họ có thể xác định vị trí và hình dạng chung của mũi, mắt, môi, và tóc của mọi người. Đây là các khuôn mặt chung từ suốt hơn thế kỉ qua, được chia theo giới tính:

Bạn có chú ý thấy gì không? Người Mĩ—đặc biệt là phụ nữ—đã bắt đầu mỉm cười. Họ đi từ khuôn mặt gần như không cảm xúc đầu Thế kỉ XX đến tươi cười vào cuối thế kỉ.

Vậy tại sao có sự thay đổi đó? Người Mĩ hạnh phúc hơn sao?

Không. Các học giả khác đã giúp trả lời câu hỏi này. Lí do thật thú vị, ít nhất là đối với tôi. Khi ảnh chụp mới được phát minh, người ta nghĩ nó giống như tranh vẽ. Không có gì khác để so sánh. Thế là, người mẫu ảnh bắt chước cách tạo dáng của người mẫu tranh. Và vì ngồi để vẽ chân dung thì không thể nào chống miệng mỉm cười suốt nhiều giờ cho đến khi xong được, người ta chọn nét mặt nghiêm nghị. Người mẫu ảnh cũng chọn nét mặt như thế.

Cuối cùng điều gì khiến họ thay đổi? Dĩ nhiên là kinh doanh, lợi nhuận, và marketing. Giữa Thế kỉ XX, công ty phim và máy ảnh Kodak rất nản lòng bởi số lượng ảnh người ta chụp quá ít. Thế là họ nghĩ ra một chiến lược khiến mọi người chụp nhiều lên. Quảng cáo của Kodak bắt đầu liên hệ ảnh chụp với sự hạnh phúc. Mục đích là gieo cho mọi người thói quen chụp ảnh bất cứ khi nào họ muốn cho người khác thấy họ đang rất chi là hạnh phúc. Tất cả những ảnh chụp niên giám mỉm cười đó là một trong những kết quả của chiến dịch thành công này (như hầu hết các ảnh chụp ta thấy trên Facebook và Instagram ngày nay vậy).

Nhưng dữ liệu dạng ảnh có thể cho ta biết nhiều điều hơn là chỉ chuyện mấy cô cậu học sinh bắt đầu cười khi nào. Thật ngạc nhiên, hình ảnh cũng có thể cho ta biết nền kinh tế đang tốt xấu ra sao.

Ta sẽ xem một bài báo học thuật có tiêu đề rất khiêu khích: "Đo lường tăng trưởng kinh tế từ ngoại tầng không gian." (Measuring Economic Growth from Outer Space). Khi thấy một bài báo có tiêu đề hấp dẫn như thế, chắc chắn là tôi sẽ lao vào đọc. Các tác giả bài báo này—J. Vernon Henderson, Adam Storeygard, và David N. Weil—bắt đầu bằng nhận xét rằng ở nhiều nước đang phát triển, các thước đo hiện có về tổng sản phẩm quốc nội (GDP) là không hiệu quả. Đó là vì phần lớn các hoạt động kinh tế diễn ra ngoài sổ sách, và các cơ quan chính phủ có nhiệm vụ đo lường đầu ra kinh tế thường có nguồn dữ liệu hạn chế.

Ý tưởng khác thường của các tác giả là gì? Họ có thể giúp đo lường GDP dựa trên lượng ánh sáng ở các nước này vào ban đêm. Họ nhận thông tin đó từ các ảnh chụp của một vệ tinh Không quân Mĩ bay vòng quanh trái đất 14 lần mỗi ngày.

Tại sao ánh sáng vào ban đêm có thể là một thước đo GDP tốt? Ừ thì, ở những nơi rất nghèo trên thế giới, người ta phải rất vất vả để trả tiền điện. Và kết quả là, khi điều kiện kinh tế kém, các hộ gia đình và làng xóm sẽ giảm đáng kể lượng ánh sáng sử dụng vào ban đêm.

Ánh sáng đêm giảm mạnh ở Indonesia suốt cuộc khủng hoảng tài chính châu Á 1998. Ở Hàn Quốc, ánh sáng đêm tăng 72% từ 1992 đến 2008, tương ứng với sự phát triển kinh tế nổi bật suốt thời kì này. Ở Triều Tiên, trong cùng thời kì đó, ánh sáng đêm thực sự giảm mạnh, tương ứng với thành tích kinh tế ảm đạm suốt thời kì này.

Năm 1998, ở Nam Madagascar, một số lượng lớn hồng ngọc và bích ngọc được phát hiện. Thị trấn Ilakaka từ chỗ chỉ là một trạm dừng xe tải bỗng hóa thành một trung tâm thương mại lớn. Hầu như không có ánh sáng đêm ở Ilakaka trước năm 1998. Nhưng 5 năm tiếp theo, có sự bùng nổ ánh sáng ban đêm tại đây.

Các tác giả thừa nhận dữ liệu ánh sáng đêm của họ còn lâu mới là một thước đo hoàn hảo đầu ra kinh tế. Rõ ràng là bạn không thể biết chính xác một nền kinh tế đang như thế nào chỉ bằng lượng ánh sáng mà các vệ tinh thu được ban đêm. Các tác giả không khuyên dùng thước đo này đối với các nước phát triển, như Mĩ chẳng hạn, ở đó dữ liệu kinh tế hiện có là chính xác hơn. Và công bằng mà nói, ngay cả ở các nước đang phát triển, họ thấy rằng ánh sáng đêm cũng chỉ hữu ích tương đương các thước đo chính thức thôi. Nhưng kết hợp dữ liệu khiếm khuyết của chính phủ với dữ liệu ánh sáng đêm không hoàn hảo sẽ cho ra một ước tính tốt hơn là chỉ dùng mỗi một nguồn thông tin đơn lẻ. Nói cách khác, ta có thể cải thiện hiểu biết của mình về các nền kinh tế đang phát triển khi dùng tranh ảnh thu từ ngoại tầng không gian.

Joseph Reisinger, một tiến sĩ khoa học máy tính có giọng nói nhẹ nhàng, cũng có cùng nỗi thất vọng với các bộ dữ liệu hiện có về nền kinh

tế ở những nước đang phát triển. Tháng 4/2014, Reisinger thấy rằng, Nigeria đã cập nhật con số ước tính GDP, ghi nhận thêm các khu vực kinh tế mà lúc trước họ đã bỏ qua. GDP ước tính của họ bấy giờ tăng hơn 90%.

"Họ là nền kinh tế lớn nhất châu Phi," Reisinger nói, giọng từ từ cao lên. "Chúng ta thậm chí còn không biết điều cơ bản nhất mà chúng ta muốn biết về đất nước đó."

Ông muốn có cái nhìn sắc nét hơn về tình hình kinh tế. Giải pháp của ông đúng là một ví dụ về cách tái hình dung những gì hình thành nên dữ liệu và giá trị của việc làm đó.

Reisinger sáng lập Công ty Premise. Ông thuê một nhóm nhân viên ở các nước đang phát triển, được trang bị smartphone. Công việc của nhân viên là gì? Chụp ảnh các hoạt động bất thường thú vị có thể có tầm quan trọng về kinh tế.

Nhân viên có thể chụp ảnh bên ngoài các trạm xăng hoặc những thùng đựng trái cây trong siêu thị. Họ chụp ảnh cùng một địa điểm nhiều lần khác nhau. Hình ảnh được gởi về Premise, và nhóm nhân viên thứ hai ở đây—các nhà khoa học máy tính—biến các ảnh chụp này thành dữ liệu. Các nhà phân tích của công ty có thể mã hóa mọi thứ từ độ dài dòng người ở các trạm xăng đến số táo hiện có trong một siêu thị, độ chín của các trái táo này, giá niêm yết trên thùng táo. Dựa trên ảnh chụp đủ các loại hoạt động, Premise có thể ước tính đầu ra và lạm phát kinh tế. Tại các nước đang phát triển, dòng người dài ở các trạm xăng là một chỉ báo hàng đầu cho thấy khó khăn kinh tế. Không có táo hoặc táo chưa chín cũng vậy. Các hình ảnh thực địa tại Trung Quốc của Premise giúp họ phát hiện lạm phát thực phẩm ở đó năm 2011 cũng như giảm phát thực phẩm năm 2012, rất lâu trước khi có dữ liệu chính thức.

Premise bán thông tin này cho các ngân hàng và các quỹ phòng hộ, cũng như cộng tác với World Bank.

Ý tưởng của Premise vẫn đang tạo ra những thành tựu giá trị. World Bank gần đây quan tâm đến quy mô nền kinh tế thuốc lá ngầm ở Philippines. Đặc biệt, họ muốn biết ảnh hưởng từ các nỗ lực của chính

phủ gần đây, với các cuộc càn quét ngẫu nhiên, nhằm đàn áp thẳng tay những hãng sản xuất thuốc lá trốn thuế. Ý tưởng của Premise là gì? Là chụp hình các hộp thuốc lá nhìn thấy trên đường phố. Xem trong đó có bao nhiêu hộp có dán tem thuế—tức là loại thuốc lá hợp pháp. Họ đã phát hiện rằng bộ phận kinh tế ngầm này, quy mô vốn rất lớn năm 2015, đã thu nhỏ lại đáng kể năm 2016. Các nỗ lực của chính phủ đã có hiệu quả, tuy nhiên, để biết được điều đó cần phải có dữ liệu mới.

***

Như chúng ta đã thấy, dữ liệu đã được tái hình dung một cách mạnh mẽ trong thời đại kĩ thuật số, và nhiều hiểu biết sâu sắc đã được phát hiện từ lượng thông tin mới này. Biết được cái gì điều khiển xu hướng truyền thông, cái gì làm nên một cuộc hẹn hò đầu tiên tốt đẹp, và các nền kinh tế đang thực sự phát triển ra sao, tất cả chỉ là bước khởi đầu.

Không phải ngẫu nhiên, người ta cũng đã kiếm được rất nhiều tiền từ dữ liệu mới đó, đơn cử là hàng chục tỉ của các *me-xừ* Brin và Page. Bản thân Joseph Reisinger cũng làm ăn không tệ. Các nhà quan sát ước tính Premise bây giờ đang thu về hàng chục triệu đô la doanh thu hàng năm. Giới đầu tư gần đây đã rót 50 triệu USD vào công ty này. Điều đó có nghĩa là một số nhà đầu tư xem Premise thuộc nhóm các doanh nghiệp giá trị nhất thế giới—trước hết là trong ngành chụp và bán ảnh, cùng đẳng cấp với tạp chí *Playboy*.

Nói cách khác, đang có một lượng giá trị ngoại hạng đang chờ các học giả và doanh nhân trong việc tận dụng các loại dữ liệu mới, trong việc suy nghĩ rộng hơn về thứ được gọi là dữ liệu. Ngày nay, nhà khoa học dữ liệu không thể giới hạn mình ở một góc nhìn hẹp hoặc truyền thống về dữ liệu. Ngày nay, ảnh chụp các đoàn người tại siêu thị là dữ liệu có giá trị. Độ đầy của các thùng hàng siêu thị là dữ liệu. Độ chín của táo là dữ liệu. Ảnh chụp từ không gian là dữ liệu. Độ cong của môi là dữ liệu. Mọi thứ là dữ liệu!

Và với tất cả dữ liệu mới này, cuối cùng chúng ta đã có thể nhìn thấu những lời nói dối của mọi người.

# CHƯƠNG 4

## Huyết thanh sự thật số

Mọi người đều nói dối.

Người ta nói dối số li đã uống trước khi về nhà. Họ nói dối số lần đi tập gym một tuần, về giá đôi giày mới mua, và cả về chuyện có đọc quyển sách mà họ đã nói hay không. Họ gọi điện báo nghỉ bệnh khi vẫn khỏe như vâm. Họ nói sẽ liên lạc nhưng rồi bặt vô âm tín. Họ nói rằng chuyện không liên quan đến bạn mặc dù có liên quan. Họ nói họ yêu bạn dù rằng họ không hề yêu. Họ nói họ vui dù rằng đang buồn chán. Họ nói họ thích phụ nữ dù thực tế họ thích đàn ông.

Người ta nói dối với bạn bè. Họ nói dối với ông chủ. Họ nói dối với trẻ con. Họ nói dối với cha mẹ. Họ nói dối với bác sĩ. Họ nói dối với chồng. Họ nói dối với vợ. Họ nói dối với chính mình.

Và chắc chắn họ nói dối với các cuộc khảo sát.

Đây là khảo sát ngắn của tôi dành cho bạn:

> Bạn đã bao giờ gian lận trong thi cử chưa? _____
>
> Bạn đã bao giờ tưởng tượng đến việc giết ai đó chưa? _____

Bạn có định nói dối? Nhiều người nói giảm các hành vi và suy nghĩ đáng xấu hổ của mình trên tờ khảo sát. Họ muốn mình trông có vẻ tốt, ngay cả dù hầu hết các khảo sát là nặc danh. Điều này được gọi là *thiên kiến mong đợi xã hội* (social desirability bias).

Một bài báo quan trọng năm 1950 cung cấp bằng chứng mạnh mẽ cho thấy các khảo sát có thể trở thành nạn nhân của thiên kiến đó. Các nhà nghiên cứu thu thập dữ liệu từ các nguồn chính thức về cư dân Denver: bao nhiêu phần trăm đã bầu cử, bao nhiêu phần trăm đóng góp từ thiện, và bao nhiêu phần trăm có thẻ thư viện. Sau đó họ khảo sát cư dân ở đây để xem các con số có khớp không. Kết quả thật sốc. Những gì cư dân ở đó nói với khảo sát rất khác với dữ liệu các nhà nghiên cứu đã thu thập. Dù không cần phải khai tên, rất nhiều người vẫn phóng đại chuyện đăng kí cử tri, hành vi bầu cử, và cả chuyện đóng góp từ thiện.

| | KHẢO SÁT | CHÍNH THỨC |
|---|---|---|
| - Đăng kí bầu cử | 83% | 69% |
| - Đi bầu trong cuộc tuyển cử tổng thống gần đây nhất | 73% | 61% |
| - Đi bầu trong cuộc tuyển cử thị trưởng gần đây nhất | 63% | 36% |
| - Có thẻ thư viện | 20% | 13% |
| - Đã từng đóng góp cho chiến dịch từ thiện Community Chest gần đây | 67% | 33% |

Điều gì đã thay đổi trong 65 năm qua? Trong thời đại Internet, không có thẻ thư viện không còn là chuyện đáng xấu hổ nữa. Nhưng, mặc dù những thứ đáng xấu hổ hoặc đáng mong đợi có thể đã thay đổi, khuynh hướng lừa dối người thăm dò vẫn còn rất mạnh.

Một khảo sát gần đây hỏi các cử nhân Đại học Maryland nhiều câu hỏi về trải nghiệm đại học của họ. Các câu trả lời được đối chiếu với dữ liệu chính thức. Người ta luôn cung cấp thông tin sai, theo hướng làm cho họ trông có vẻ tốt hơn. Chưa tới 2% nói rằng điểm trung bình tốt nghiệp của họ dưới 2.5 (với mức điểm 4.0 là cao nhất). (Trong thực tế, khoảng 11% tốt nghiệp thấp hơn điểm này.) 44% nói họ đã đóng góp cho trường trong năm qua. (Trong thực tế, chỉ khoảng 28% đã đóng góp.)

Chắc chắn là việc nói dối đóng vai trò quan trọng trong việc làm lệch lạc kết quả dự báo của các cuộc thăm dò về chiến thắng năm 2016 của Donald Trump. Các cuộc thăm dò nói chung đánh giá mức độ ủng hộ

Donald Trump thấp hơn thực tế khoảng 2 %p. Một số người có thể đã xấu hổ không dám nói rằng mình ủng hộ Trump. Một số người có thể đã nói họ chưa quyết định được, dù thực ra họ đã ủng hộ Trump từ lâu rồi.

Tại sao người ta lại thông tin sai lạc trong các cuộc khảo sát nặc danh? Tôi hỏi Roger Tourangeau, cựu giáo sư nghiên cứu tại Đại học Michigan và có lẽ là chuyên gia hàng đầu thế giới về thiên kiến mong đợi xã hội. Thói quen "nói dối vô hại" là một phần quan trọng của vấn đề. Tourangeau giải thích: "Khoảng 1/3 thời gian, người ta nói dối trong đời sống thực tế. Các thói quen đó nhiễm vào các cuộc khảo sát."

Thỉnh thoảng chúng ta lại có cái thói quen kì lạ là nói dối chính mình. "Chẳng hạn, người ta thường không muốn thú nhận với bản thân rằng mình đã từng là một sinh viên cà giựt," Tourangeau nói.

Việc nói dối với bản thân có thể giải thích tại sao quá nhiều người nói họ ở mức trên trung bình. Vấn đề này lớn đến đâu? Hơn 40% kĩ sư của một công ty nói họ thuộc top 5%. Hơn 90% giáo sư đại học nói họ làm việc trên mức trung bình. Có 1/4 số học sinh tốt nghiệp trung học nghĩ họ thuộc top 1% về khả năng hòa hợp với người khác. Nếu đang tự dối mình, hẳn là ta cũng không thể thành thật trong khảo sát.

Một yếu tố khác cũng khiến người ta nói dối trong các cuộc khảo sát chính là mong muốn tạo ấn tượng tốt với người phỏng vấn, trong trường hợp có người phỏng vấn. Theo cách nói của Tourangeau, "Một người trông giống như người cô yêu mến của bạn bước vào. [...] Bạn có muốn nói với người cô yêu mến của mình rằng bạn vừa hút cần tháng trước không?"[1] Bạn có muốn thú nhận rằng bạn đã không đóng góp tiền cho ngôi trường cũ yêu dấu của bạn không?

---

[1] Một lí do khác để nói dối đơn giản là nhằm phá rối cuộc khảo sát. Đây là một vấn đề lớn cho các nghiên cứu liên quan đến thanh thiếu niên tuổi teen (từ 13 đến 19), khiến chuyện hiểu nhóm tuổi này càng thêm phức tạp. Các nhà nghiên cứu ban đầu thấy có tương quan giữa việc được nhận làm con nuôi và một loạt các hành vi tiêu cực của các thanh thiếu niên, như dùng ma túy, uống rượu, và trốn học. Trong nghiên cứu kế tiếp, họ phát hiện mối tương quan này hoàn toàn được giải thích bởi 19% số thanh thiếu niên khai khống là được nhận nuôi. Nghiên cứu theo sau phát hiện rằng một tỉ lệ lớn thanh thiếu niên khai trong các khảo sát là mình cao hơn 7 feet (~ 2 mét), nặng hơn 400 pound (~ 180 kg), hoặc có 3 con. Một khảo sát phát hiện rằng 99% số học sinh nói với nhà nghiên cứu rằng mình có 1 tay/chân giả hóa ra chỉ đang nói đùa.

Vì lí do này, điều kiện phỏng vấn càng ít mang tính cá nhân, người ta sẽ càng thật thà. Để gợi được câu trả lời chân thật, khảo sát Internet tốt hơn khảo sát điện thoại, khảo sát điện thoại tốt hơn khảo sát trực tiếp mặt đối mặt. Khi chỉ một mình, người ta sẽ thú nhận nhiều hơn khi có người khác ngồi trong phòng với họ.

Tuy nhiên, đối với các chủ đề nhạy cảm, mọi phương pháp khảo sát đều sẽ khơi gợi hành vi cung cấp thông tin sai lạc. Tourangeau ở đây đã dùng một từ mà thường hay được các nhà kinh tế học sử dụng: "động cơ." Người ta không có động cơ nói sự thật trong các cuộc khảo sát.

Vậy thì làm sao biết được đồng bào ta thực sự đang nghĩ và làm gì?

Trong một số trường hợp, vẫn có các nguồn dữ liệu chính thức ta có thể tham chiếu để tìm sự thật. Chẳng hạn, nếu người ta nói dối về đóng góp từ thiện, ta vẫn có thể có các con số thực tế về hoạt động quyên góp trong một vùng từ chính các hội từ thiện. Nhưng nếu ta đang cố tìm hiểu các hành vi không được liệt kê trong dữ liệu chính thức, hoặc điều người ta đang suy nghĩ—niềm tin, cảm xúc, và ước muốn thực sự của họ—không có nguồn thông tin nào khác ngoại trừ những gì mọi người hạ cố nói ra trong các cuộc khảo sát. Từ trước tới nay, tình hình là vậy.

Đây là sức mạnh thứ hai của Dữ Liệu Lớn: Một số nguồn trực tuyến khiến người ta thú nhận những thứ mà họ sẽ không thú nhận ở đâu khác. Các nguồn đó chính là huyết thanh sự thật số. Hãy nghĩ về các tìm kiếm Google. Hãy duyệt lại các điều kiện khiến người ta chân thật hơn. Trực tuyến? Có. Một mình? Có. Không có người quản lí cuộc khảo sát? Có luôn.

Và các tìm kiếm Google còn một thuận lợi lớn khác để khiến người ta nói sự thật: động cơ. Nếu bạn thích chuyện cười phân biệt chủng tộc, bạn không có động cơ chia sẻ sở thích đó của mình trong các cuộc khảo sát. Tuy nhiên, bạn lại có động cơ tìm kiếm trên mạng chuyện cười phân biệt chủng tộc mới nhất và hay nhất. Nếu nghĩ mình có thể đang bị trầm cảm, ta không có động cơ thú nhận điều này trong cuộc khảo sát. Tuy nhiên, ta thực sự có động cơ hỏi Google về triệu chứng và các cách chữa trị tiềm năng.

Ngay cả nếu ta đang nói dối với chính mình, Google vẫn có thể biết sự thật. Vài ngày trước bầu cử, bạn và một số người trong khu phố có thể thực lòng nghĩ là mình sẽ lái xe đến nơi bầu cử và tiến hành bỏ phiếu. Nhưng, nếu các bạn chưa tìm kiếm thông tin nào về cách bầu hoặc nơi bầu, các nhà khoa học dữ liệu như tôi có thể biết được rằng số cử tri thực sự đi bầu sẽ thấp. Tương tự, có thể bạn chưa thú nhận với chính mình rằng bạn có thể bị trầm cảm, dù rằng bạn đang Google về những cơn khóc lóc liên hồi và tình trạng khó rời khỏi giường. Tuy nhiên, bạn sẽ xuất hiện trong những tìm kiếm liên quan đến trầm cảm mà tôi đã phân tích trong sách này.

Hãy nghĩ về trải nghiệm của chính bạn khi dùng Google. Tôi đoán thỉnh thoảng bạn đã gõ một số thứ vào ô tìm kiếm, tiết lộ một hành vi hoặc suy nghĩ mà bạn thường không muốn thú nhận trong bối cảnh lịch sự. Thực vậy, có bằng chứng rất rõ ràng là đại đa số người Mĩ đang nói cho Google biết một số điều rất riêng tư. Chẳng hạn, người Mĩ tìm kiếm "porn" (khiêu dâm) nhiều hơn "weather" (thời tiết). Nhân tiện nói luôn, điều này khó mà tương thích với dữ liệu khảo sát, vì chỉ khoảng 25% nam và 8% nữ thú nhận là mình có xem phim ảnh khiêu dâm.

Có thể bạn cũng đã chú ý thấy một sự chân thật nhất định trong các tìm kiếm Google khi nhìn cách cỗ máy tìm kiếm này tự động gợi ý hoàn thành các câu hỏi của bạn. Các gợi ý của nó dựa trên các tìm kiếm phổ biến nhất mà những người khác đã thực hiện. Vậy các kết quả hoàn thành tự động cho ta biết mọi người đang Google những gì. Thực ra, kết quả hoàn thành tự động có thể hơi sai lạc một chút. Google sẽ không gợi ý một số chữ mà nó thấy không thích hợp, như "cock," "f*ck," và "porn." Điều này có nghĩa là chức năng hoàn thành tự động giảm bớt sự sỗ sàng những suy nghĩ trên Google của người ta so với thực tế. Dù vậy, một số thứ nhạy cảm vẫn thường xuất hiện.

Nếu gõ "Tại sao..." thì hai phương án hoàn thành tự động đầu tiên của Google ở thời điểm tôi viết dòng này là "Tại sao bầu trời màu xanh?" và "Tại sao có ngày nhuận?"—ám chỉ đây là hai cách phổ biến nhất để hoàn thành tìm kiếm này. Kết quả thứ ba: "Tại sao phân tôi màu xanh lá?" Chức năng hoàn thành tự động có thể khá khó chịu. Nếu gõ "Có

bình thường không khi muốn...," gợi ý đầu tiên là "giết." Nếu gõ "Có bình thường không khi muốn giết...," gợi ý đầu tiên là "gia đình tôi."[1]

Bạn có cần thêm bằng chứng rằng các tìm kiếm Google có thể cho ta một bức tranh khác biệt với thế giới ta thường thấy không? Hãy thử nghiên cứu các tìm kiếm liên quan đến sự hối tiếc xung quanh quyết định nên có con hay không. Trước khi quyết định, một số người sợ sẽ chọn sai. Và hầu như luôn luôn, câu hỏi là liệu họ sẽ hối tiếc *đã không có con* hay không. Khả năng người ta hỏi Google *họ sẽ hối tiếc vì đã không có con hay không* cao hơn gấp 7 lần *họ sẽ hối tiếc vì đã có con hay không*.

Sau khi quyết định—sinh (hoặc nhận nuôi) hay không—người ta đôi khi thú nhận với Google rằng họ hối hận với lựa chọn của mình. Chuyện nghe có thể khá sốc: Sau khi quyết định xong, các con số đảo ngược. Những người có con có gấp 3.6 lần khả năng sẽ bảo với Google rằng họ hối tiếc vì quyết định của mình, so với những người không có con.

Một cảnh báo nên nhớ trong suốt chương này: Google có thể biểu lộ thiên kiến hướng đến những suy nghĩ không đứng đắn, những suy nghĩ mà người ta cảm thấy là không thể bàn với bất cứ ai khác. Nếu ta đang muốn phát hiện các suy nghĩ bị che giấu, khả năng tìm bới các suy nghĩ ấy của Google sẽ rất hữu ích. Và sự khác biệt lớn giữa hối tiếc đã có con và không có con trong trường hợp này có vẻ hàm ý rằng, sự hối tiếc ấy là hối tiếc thực sự.

Chúng ta hãy tạm dừng một lát để nghiên cứu xem, hành vi tìm kiếm các cụm từ kiểu như "Tôi hối tiếc vì đã có con" có ý nghĩa gì. Google tự thể hiện mình là một nguồn để ta tìm kiếm thông tin trực tiếp, về các chủ đề như thời tiết hôm nay, ai thắng trận đấu tối qua, hoặc tượng Nữ thần Tự do dựng lên khi nào. Nhưng đôi khi ta gõ các suy nghĩ không bị kiểm duyệt của mình vào Google, mà không hi vọng lắm là nó sẽ có thể giúp ta. Trong trường hợp này, cửa sổ tìm kiếm đóng vai trò như kiểu phòng xưng tội.

---

[1] [ND] Kết quả hoàn thành tự động (auto-complete) trong tìm kiếm Google thường xuyên thay đổi, và còn phụ thuộc vào lịch sử tìm kiếm của mỗi người (nếu bạn không bật chức năng trình duyệt ẩn danh—incognito). Do vậy, các kết quả trên đây của tác giả có thể không giống với kết quả trong máy của bạn. Tuy nhiên, độ kì cục có lẽ không khác nhau nhiều.

Có hàng ngàn tìm kiếm mỗi năm, ví dụ, "Tôi ghét thời tiết lạnh," "Người ta thật phiền phức," và "Tôi buồn." Dĩ nhiên, hàng ngàn tìm kiếm Google cho mục "Tôi buồn" đó đại diện chỉ một phần rất nhỏ của hàng trăm triệu người cảm thấy buồn trong một năm nào đó. Nghiên cứu của tôi đã phát hiện: Những tìm kiếm thể hiện suy nghĩ (chứ không phải tìm kiếm thông tin) chỉ được thực hiện bởi một mẫu nhỏ trong số những người có suy nghĩ đó. Tương tự, nghiên cứu của tôi cho thấy rằng 7,000 tìm kiếm của người Mĩ hàng năm cho mục "Tôi hối tiếc đã có con" đại diện một mẫu rất nhỏ những người đã từng có suy nghĩ đó.

Trẻ con rõ ràng là một niềm vui lớn cho hầu hết mọi người. Mặc dù mẹ tôi luôn sợ rằng "con và cái phép phân tích dữ liệu ngu xuẩn của con" sẽ hạn chế số cháu nội của bà, nghiên cứu này vẫn chưa thay đổi ước muốn có con của tôi. Nhưng sự hối tiếc đó rất thú vị—và là một khía cạnh khác của con người mà chúng ta thường không nhìn thấy trong các dữ liệu truyền thống. Văn hóa của chúng ta luôn ngập tràn hình ảnh gia đình tuyệt vời, hạnh phúc. Hầu hết mọi người sẽ không bao giờ xem có con là điều mà mình sẽ có thể hối tiếc. Nhưng thực sự có một số người hối tiếc. Có thể họ không thú nhận điều này với ai cả—trừ Google.

## Sự thật về giới tính

Có bao nhiêu nam giới Mĩ đồng tính? Đây là câu hỏi huyền thoại trong mảng nghiên cứu về tính dục. Tuy nhiên, nó thuộc nhóm các câu hỏi khó trả lời nhất. Các nhà tâm lí học không còn tin con số ước tính nổi tiếng của Alfred Kinsey nữa. Dựa trên các khảo sát lấy mẫu chủ yếu là tù nhân và người bán dâm, ông cho rằng 10% nam giới Mĩ là đồng tính. Các khảo sát tiêu biểu bây giờ cho ta biết con số ấy vào khoảng 2% đến 3%. Nhưng sở thích tình dục từ lâu đã thuộc nhóm chủ đề mà người ta có khuynh hướng nói dối. Tôi nghĩ mình có thể dùng Dữ Liệu Lớn để trả lời câu hỏi này tốt hơn.

Trước tiên, hãy nói thêm về dữ liệu khảo sát kể trên. Các khảo sát cho ta biết số nam đồng tính ở các tiểu bang dễ tính (tolerant state) nhiều hơn hẳn các tiểu bang khó tính (intolerant state). Ví dụ, theo một khảo sát Gallup, tỉ lệ dân số đồng tính ở Rhode Island, tiểu bang ủng hộ hôn

nhân đồng tính nhất, hầu như cao gấp đôi Mississippi, tiểu bang ít ủng hộ hôn nhân đồng tính nhất.

Có 2 lời giải thích hợp lí cho vấn đề này. Thứ nhất, nam đồng tính sinh ở các tiểu bang khó tính có thể chuyển đến các tiểu bang dễ tính. Thứ hai, nam đồng tính ở các tiểu bang khó tính có thể không tiết lộ là mình đồng tính; thậm chí rất có thể họ sẽ nói dối.

Một số hiểu biết về lời giải thích số một—tính di động của người đồng tính—có thể được lượm lặt từ một nguồn Dữ Liệu Lớn khác: Facebook. Trang này cho phép người dùng ghi rõ họ thích giới tính nào. Khoảng 2.5% người dùng Facebook nam (có ghi sở thích giới tính) ghi rằng họ thích nam giới; gần như tương đồng với những gì các khảo sát chỉ ra. Và Facebook cũng chỉ ra những khác biệt lớn về dân số đồng tính ở các tiểu bang có độ dễ tính cao thấp khác nhau: Trên Facebook, số người đồng tính ở Rhode Island cao gấp hơn 2 lần ở Mississippi.

Facebook còn có thể cung cấp thông tin về sự dịch chuyển. Tôi đã mã hóa được quê nhà của một nhóm người dùng Facebook đồng tính công khai. Điều này cho phép tôi trực tiếp ước tính có bao nhiêu nam đồng tính chuyển ra khỏi các tiểu bang khó tính đến các nơi dễ tính hơn trong nước. Vậy kết quả ra sao? Rõ ràng có sự dịch chuyển—ví dụ, từ Oklahoma City đến San Francisco. Nhưng tôi ước tính số nam giới ôm các đĩa CD nhạc Judy Garland và chuyển tới những nơi có tư tưởng cởi mở hơn chỉ giúp giải thích chưa đến phân nửa sự khác biệt về số người đồng tính công khai ở các bang dễ tính so với các bang khó tính.[1]

Hơn nữa, Facebook cho phép tập trung vào học sinh trung học. Đây là một nhóm đặc biệt, vì nam sinh trung học hiếm khi có cơ hội tự chọn nơi ở. Nếu sự dịch chuyển là nhân tố tạo ra sự khác biệt ở số người đồng tính công khai trong từng bang, sự khác biệt này sẽ không xuất hiện

---

[1] Một số người có thể thấy khó chịu việc tôi liên hệ chuyện nam thích nhạc Judy Garland với chuyện họ thích quan hệ tình dục với đàn ông, ngay cả khi tôi chỉ nói đùa. Và tôi dĩ nhiên không có ý ám chỉ rằng nam đồng tính đều mê các danh ca nữ. Nhưng dữ liệu tìm kiếm cho thấy rằng định kiến này có phần đúng. Tôi ước tính rằng một người đàn ông tìm kiếm thông tin về Judy Garland có khả năng tìm kiếm phim ảnh khiêu dâm đồng tính gấp 3 lần khiêu dâm khác giới. Dữ Liệu Lớn cho ta biết, một số định kiến là đúng.

trong nhóm người dùng là học sinh trung học. Vậy dữ liệu trung học nói lên điều gì? Nam sinh trung học đồng tính công khai ở các bang khó tính ít hơn rất nhiều. Chỉ 2 phần ngàn nam sinh trung học ở Mississippi là đồng tính công khai. Vậy vấn đề chẳng phải chỉ do sự dịch chuyển.

Nếu số nam đồng tính sinh ra ở các bang và sự dịch chuyển không thể giải thích đầy đủ tại sao một số bang có nhiều nam đồng tính công khai hơn đáng kể các bang khác, thì việc giấu giếm hẳn sẽ đóng một vai trò lớn. Điều đó đưa ta trở lại với Google, vì rất nhiều người sẵn sàng chia sẻ những điều thầm kín với cỗ máy này.

Liệu có thể dùng các tìm kiếm khiêu dâm để kiểm tra *thực tế* có bao nhiêu nam đồng tính ở mỗi bang không? Sự thật là có. Bằng dữ liệu từ các tìm kiếm Google và Google AdWords, tôi ước tính cả nước có khoảng 5% tìm kiếm khiêu dâm nam là muốn tìm khiêu dâm nam đồng tính. (Đây bao gồm các tìm kiếm cho các từ ngữ như "R**ket Tube," một trang khiêu dâm đồng tính phổ biến, cũng như "gay porn.")

Con số này ở các vùng có khác nhau không? Trên tổng thể, có nhiều tìm kiếm khiêu dâm đồng tính ở các bang dễ tính hơn các bang khó tính. Điều này dễ hiểu, là do một số nam đồng tính chuyển khỏi những nơi khó tính và đến ở những nơi dễ tính. Nhưng khác biệt không lớn như trong các khảo sát hoặc trên Facebook. Ở Mississippi, tôi ước tính 4.8% tìm kiếm khiêu dâm nam là tìm khiêu dâm đồng tính, cao hơn nhiều các con số được chỉ ra bởi các khảo sát hoặc Facebook, và rất gần Rhode Island—nơi có 5.2% tìm kiếm khiêu dâm là tìm khiêu dâm đồng tính.

Vậy bao nhiêu nam giới Mĩ là đồng tính? Cách đo lường theo lượng tìm kiếm khiêu dâm của nam giới—với khoảng 5% là đồng tính—có vẻ là một con số ước tính hợp lí số người đồng tính thực tế ở Mĩ. Và có một cách khác, ít trực tiếp hơn để có được con số này. Nó đòi hỏi một chút khoa học dữ liệu. Chúng ta có thể sử dụng mối quan hệ giữa sự dễ tính và số dân đồng tính công khai. Xin hãy kiên nhẫn với tôi một chút ở đây.

Nghiên cứu sơ bộ của tôi chỉ ra rằng, ở một bang, cứ 20%p ủng hộ hôn nhân đồng tính tương đương với việc số nam giới đồng tính công khai trên Facebook tăng khoảng 1.5 lần. Dựa vào đó, ta có thể ước tính số

nam đồng tính công khai sinh ở một nơi tưởng tượng có mức độ dễ tính đạt 100%, tức là tất cả mọi người ủng hộ hôn nhân đồng tính. Theo ước tính của tôi, con số đó là khoảng 5%, rất khớp với dữ liệu từ các tìm kiếm khiêu dâm. Môi trường ủng hộ đồng tính cao nhất là môi trường của các nam sinh trung học tại Bay Area ở California. Khoảng 4% nam sinh ở đó là đồng tính công khai trên Facebook, có vẻ khớp với tính toán của tôi.

Tôi phải lưu ý rằng tôi chưa thể ước tính con số đồng tính nữ. Số tìm kiếm khiêu dâm không hữu dụng lắm, vì phụ nữ xem phim ảnh khiêu dâm ít hơn rất nhiều, làm cho mẫu ít có tính đại diện. Và trong số những người có xem, thậm chí những phụ nữ thích đàn ông trong đời thực dường như cũng thích xem khiêu dâm đồng tính nữ. Ngót 20% video được phụ nữ xem trên P***Hub là đồng tính nữ.

Dĩ nhiên, 5% nam giới Mĩ đồng tính chỉ là một con số ước tính. Một số nam giới là lưỡng tính; một số—đặc biệt khi còn trẻ—không chắc chắn mình thuộc nhóm nào. Rõ ràng, bạn không thể tính con số này một cách chính xác như tính số người đi bỏ phiếu hoặc đi xem một bộ phim.

Nhưng trong ước tính của tôi, có một điều rõ ràng: Rất nhiều nam giới tại Mĩ, đặc biệt ở các bang khó tính, vẫn sống trong sự giấu giếm. Họ không tiết lộ sở thích tình dục của mình trên Facebook. Họ không thừa nhận nó trên các khảo sát. Và trong nhiều trường hợp, thậm chí họ còn kết hôn với phụ nữ.

Hóa ra việc các bà vợ nghi ngờ chồng mình đồng tính là chuyện khá thường xuyên. Họ thể hiện sự nghi ngờ đó trong một tìm kiếm phổ biến đến mức bất ngờ: "Có phải chồng tôi đồng tính?" Từ "đồng tính" có khả năng nằm trong các tìm kiếm bắt đầu với "Có phải chồng tôi..." nhiều hơn 10% từ đứng thứ hai—"ngoại tình." Nó phổ biến hơn "nghiện rượu" 8 lần và hơn "bị trầm cảm" 10 lần.

Đây có lẽ là thứ nói lên nhiều điều nhất: Các tìm kiếm nghi vấn xu hướng tính dục của người chồng phổ biến hơn rất nhiều ở các vùng khó tính nhất. Các bang có tỉ lệ phụ nữ hỏi điều này cao nhất là South Carolina và Louisiana. Thực vậy, ở 21 trong số 25 tiểu bang nơi mà câu hỏi này thường được hỏi nhất, tỉ lệ ủng hộ hôn nhân đồng tính thấp hơn

mức trung bình cả nước.

Google và các trang khiêu dâm không phải là các nguồn dữ liệu hữu ích duy nhất về xu hướng tính dục của nam giới. Dữ Liệu Lớn cho ta nhiều bằng chứng khác về chuyện giấu giếm giới tính. Tôi phân tích các quảng cáo trên Craigslist với chủ đề nam giới tìm kiếm "các cuộc gặp gỡ qua đường" (casual encounter). Tỉ lệ quảng cáo tìm kiếm các cuộc gặp gỡ qua đường này có khuynh hướng lớn hơn ở các tiểu bang khó tính. Trong số các bang có tỉ lệ cao nhất là Kentucky, Louisiana, và Alabama.

Và để hiểu chi tiết hơn, ta hãy trở lại dữ liệu tìm kiếm Google. Một trong các tìm kiếm phổ biến nhất ngay trước hoặc sau "gay porn" là "gay test" (bài kiểm tra đồng tính). (Các bài kiểm tra này cho nam giới biết họ có phải đồng tính không.) Và các tìm kiếm "gay test" là phổ biến gấp khoảng 2 lần ở các tiểu bang khó tính nhất.

Chuyện một người cứ tìm tới tìm lui hết "gay porn" rồi lại "gay test" cho ta biết điều gì? Có lẽ, điều đó cho thấy tâm trí người này đang rối rắm, nếu không muốn nói là đang bị hành hạ. Thật hợp lí khi nghĩ rằng một số người đang hi vọng việc họ quan tâm đến khiêu dâm đồng tính không có nghĩa là họ đồng tính.

Dữ liệu tìm kiếm Google không cho phép ta thấy lịch sử tìm kiếm theo thời gian của một người dùng cụ thể. Tuy nhiên, năm 2006, AOL tiết lộ một mẫu các tìm kiếm của người dùng cho các nhà nghiên cứu học thuật. Đây là một số các tìm kiếm của một người dùng nặc danh trong khoảng thời gian 6 ngày (trích nguyên văn, kể cả các lỗi đánh máy):

Friday 03:49:55 free gay picks

Friday 03:59:37 locker room gay picks

Friday 04:00:14 gay picks

Friday 04:00:35 gay sex picks

Friday 05:08:23 a long gay quiz

Friday 05:10:00 a good gay test

Friday 05:25:07 gay tests for a confused man

Friday 05:26:38 gay tests

Friday 05:27:22 am i gay tests

Friday 05:29:18 gay picks

Friday 05:30:01 naked men picks

Friday 05:32:27 free nude men picks

Friday 05:38:19 hot gay sex picks

Friday 05:41:34 hot man butt sex

Wednesday 13:37:37 am i gay tests

Wednesday 13:41:20 gay tests

Wednesday 13:47:49 hot man butt sex

Wednesday 13:50:31 free gay sex vidio

Đây có vẻ là một người nam không thoải mái với xu hướng tính dục của mình. Và dữ liệu Google cho biết vẫn có nhiều nam giới như anh ta. Thực vậy, hầu hết họ sống ở các bang khó tính về quan hệ đồng giới.

Để hiểu rõ hơn nữa về những con người ở đằng sau các con số này, tôi hỏi một bác sĩ tâm thần ở Mississippi, chuyên gia giúp đỡ nam giới đồng tính không công khai, xem có bệnh nhân nào của ông muốn nói chuyện với tôi không. Một người đã đồng ý. Ông bảo tôi ông là một giáo sư nghỉ hưu, U70, và đã kết hôn với một phụ nữ duy nhất hơn 40 năm.

Cách đây chừng 10 năm, bị stress nặng, ông đi khám bác sĩ tâm thần và cuối cùng thừa nhận xu hướng tính dục của mình. Ông luôn biết mình thích nam giới, ông nói, nhưng nghĩ rằng điều này là phổ biến, nam giới nào cũng vậy và chỉ đang giấu trong lòng mà thôi. Ngay sau khi bắt đầu liệu pháp, ông có cuộc gặp gỡ tình dục đồng giới đầu tiên và duy nhất với một nam sinh viên gần 30 tuổi của ông—một trải nghiệm ông mô tả là "tuyệt vời."

Ông và vợ ông không quan hệ tình dục. Ông nói rằng ông cảm thấy có lỗi nếu chấm dứt cuộc hôn nhân hoặc công khai hẹn hò với đàn ông. Ông hối tiếc hầu như mọi quyết định lớn trong đời mình.

Vị giáo sư hưu trí và vợ sẽ tiếp tục trải qua một đêm không tình yêu lãng mạn, không tình dục. Mặc dù đã tiến bộ rất nhiều, sự thiếu khoan

dung dai dẳng với tình dục đồng giới sẽ tiếp tục khiến hàng triệu người Mĩ khác phải hành xử tương tự.

***

Có thể bạn không sốc khi biết rằng 5% nam giới là đồng tính và nhiều người vẫn sống giấu giếm. Đã có những thời điểm hầu hết mọi người hẳn rất sốc khi biết điều này. Và vẫn có những nơi nhiều người cũng sẽ bị sốc khi nghe tin.

"Ở Iran chúng tôi không có người đồng tính như ở nước bạn," Mahmoud Ahmadinejad, tổng thống Iran bấy giờ, đã khẳng định năm 2007. "Ở Iran chúng tôi không có hiện tượng này." Tương tự, Anatoly Pakhomov, thị trưởng Sochi, Nga, ngay trước khi thành phố ông đăng cai Thế vận hội Mùa đông 2014, đã nói như sau, "Thành phố chúng tôi không có người đồng tính." Tuy nhiên, hành vi Internet tiết lộ rằng có sự quan tâm đáng kể về khiêu dâm đồng tính ở Sochi và Iran.

Điều này gợi lên một câu hỏi: Có những quan tâm tình dục phổ biến nào ở Mĩ ngày nay vẫn bị xem là gây sốc không? Điều đó tùy thuộc vào chuyện bạn xem điều gì là bình thường, và vào mức độ dễ sốc của bạn.

Hầu hết các tìm kiếm nhiều nhất trên P***Hub đều không bất ngờ. Với nam, đó là các cụm từ: "teen," "thr**some," và "bl*wjob"; với nữ thì là "passionate love making," "nipple sucking," và "man eating p*ssy."

Rời dòng chủ lưu, dữ liệu P***Hub cho ta biết về một số sở thích dị thường mà có thể bạn chưa bao giờ nghĩ là có tồn tại. Có những phụ nữ tìm "an*l apples" và "humping stuffed animals." Có những nam giới tìm kiếm "snot fetish" và "nude crucifixion." Nhưng các tìm kiếm này là hiếm—chỉ vào khoảng 10 lần mỗi tháng trên trang khiêu dâm lớn này.

Một điểm liên quan khác trở nên khá rõ ràng khi xem lại dữ liệu P***Hub: Ai cũng có người phù hợp trên đời. Phụ nữ, không có gì đáng ngạc nhiên, thường tìm kiếm các anh "cao," các anh "đen," và các anh "đẹp trai." Nhưng thỉnh thoảng họ cũng tìm kiếm các anh "lùn," các anh "xanh xao," và các anh "xấu xí." Có những phụ nữ tìm kiếm các anh "tàn tật," "anh phúng phính trym nhỏ," và "ông già mập xấu xí." Nam

giới thường tìm kiếm phụ nữ "gầy," phụ nữ "ngực bự," và phụ nữ "tóc vàng." Nhưng đôi khi họ cũng tìm kiếm phụ nữ "mập," phụ nữ "ngực siêu nhỏ," và phụ nữ "tóc xanh lá." Có những nam giới tìm kiếm phụ nữ "trọc," phụ nữ "siêu nhỏ," và phụ nữ "không núm." Dữ liệu này có thể làm những người không cao, không đen, và không đẹp trai hoặc không gầy, ngực không lớn, và tóc không vàng lạc quan hơn một chút.[1]

Còn những tìm kiếm vừa phổ biến vừa đáng ngạc nhiên thì sao? Trong 150 tìm kiếm phổ biến nhất của nam giới, điều ngạc nhiên nhất với tôi là các tìm kiếm loạn luân mà tôi đã thảo luận trong chương nói về Freud. Các đối tượng ít được thảo luận khác là "shemale" (tìm kiếm phổ biến thứ 77) và "granny" (thứ 110). Trên tổng thể, khoảng 1.4% tìm kiếm P***hub của nam giới là phụ nữ có dương vật. Khoảng 0.6% (0.4% đối với nam giới dưới tuổi 34) tìm kiếm người lớn tuổi. Chỉ có 1 trong 24,000 tìm kiếm P***hub của nam giới rõ ràng tìm trẻ dưới 13 tuổi; có thể điều này liên quan đến sự thật là P***hub, vì các lí do hiển nhiên, cấm tất cả mọi hình thức khiêu dâm trẻ em, và việc sở hữu nó là phạm pháp.

Trong các tìm kiếm hàng đầu của nữ giới trên P***Hub là một thể loại khiêu dâm mà, tôi xin cảnh báo, sẽ khiến nhiều bạn đọc khó chịu: tình dục có sử dụng bạo lực với phụ nữ. Ngót 25% tìm kiếm khiêu dâm giới tính bình thường của nữ giới nhấn mạnh sự đau đớn và/hoặc làm nhục phụ nữ—ví dụ như "painful an*l crying," "public disgrace," và "extreme brutal gangb***." Có 5% tìm kiếm tình dục không đồng thuận—"hiếp" hoặc "cưỡng bức"—dù các video loại này bị cấm trên P***Hub. Và tỉ lệ tìm kiếm tất cả các từ ngữ này trong nữ giới phổ biến ít nhất gấp 2 lần trong nam giới. Với thể loại khiêu dâm trong đó bạo lực được thực hiện với phụ nữ, bản phân tích dữ liệu của tôi chỉ ra rằng nó hầu như luôn luôn hấp dẫn khác thường đối với nữ giới.

Dĩ nhiên, khi cố gắng dần chấp nhận điều này, phải nhớ rằng có một sự khác biệt giữa tưởng tượng và đời thực. Vâng, trong thiểu số nữ giới

---

[1] Tôi nghĩ dữ liệu này còn chứa đựng những gợi ý cho chiến lược hẹn hò tối ưu. Rõ ràng, người ta phải tỏ tình, bị từ chối nhiều, và không để bụng khi bị từ chối. Tiến trình này cuối cùng sẽ cho phép bạn tìm thấy một bạn tình thích kiểu người như bạn nhất. Xin nhắc lại, dù bạn có trông thế nào đi nữa, luôn tồn tại những người thích bạn. Tin tôi đi.

vào trang P***Hub, có một nhóm nhỏ tìm phim ảnh hiếp dâm. Rõ ràng ai cũng biết, điều này *không có nghĩa là phụ nữ muốn bị hiếp dâm trong đời thực* và dĩ nhiên cũng không khiến tội hiếp dâm bớt phần kinh khủng. Dữ liệu khiêu dâm chỉ cho ta biết là đôi khi người ta tưởng tượng những điều họ không có và có thể sẽ không bao giờ đề cập nó với người khác.

***

Chốn riêng tư bí mật không chỉ chứa đựng những điều tưởng tượng. Khi nói tới chuyện tình dục, người ta giữ nhiều bí mật, ví dụ như về tần suất hoạt động tình dục.

Trong phần giới thiệu, tôi nhận xét rằng người Mĩ khai là đã dùng bao cao su nhiều vượt xa số lượng được bán hàng năm. Do đó, bạn có thể nghĩ điều này nghĩa là họ nói rằng mình dùng bao cao su trong khi quan hệ nhiều hơn thực tế. Nhưng bằng chứng cho thấy họ còn phóng đại số lần hoạt động tình dục của họ nữa. Khoảng 11% phụ nữ độ tuổi từ 15 đến 44 nói họ tích cực chuyện tình dục, hiện tại không mang thai, và không dùng biện pháp tránh thai. Dựa vào các thông số trên, các nhà khoa học dè dặt nhất cũng sẽ tính ra có khoảng 10% phụ nữ trong số này sẽ có thai mỗi tháng. Tuy vậy, ước tính này đã lớn hơn tổng số trường hợp mang thai ở Mĩ ($^{1}/_{113}$ số phụ nữ ở độ tuổi sinh con). Trong cái văn hóa bị ám ảnh bởi tình dục của chúng ta, có thể rất khó để thú nhận rằng mình không sinh hoạt tình dục nhiều đến thế.

Nhưng nếu muốn tìm hiểu hoặc tìm lời khuyên, một lần nữa, bạn sẽ có động cơ để nói cho Google biết. Trên Google, phàn nàn về chuyện bạn đời không muốn quan hệ tình dục nhiều gấp 16 lần phàn nàn về chuyện bạn đời không muốn nói chuyện. Phàn nàn về chuyện người tình ngoài giá thú không muốn quan hệ tình dục nhiều gấp 5.5 lần chuyện người tình không trả lời tin nhắn.

Và các tìm kiếm Google chỉ ra một thủ phạm đáng ngạc nhiên cho phần nhiều các mối quan hệ không tình dục này. Phàn nàn bạn trai không chịu quan hệ tình dục nhiều gấp đôi bạn gái không chịu quan hệ. Cho đến nay, tìm kiếm có ý phàn nàn về bạn trai hàng đầu chính là "Bạn trai không chịu quan hệ tình dục với tôi." (Các tìm kiếm Google không

được tách ra theo giới, nhưng, vì bản phân tích trước nói rằng 95% nam có giới tính bình thường, ta có thể đoán rằng không quá nhiều tìm kiếm "bạn trai..." đến từ nam giới.)

Ta nên lí giải thế nào? Điều này có thực sự ám chỉ bạn trai ít sinh hoạt tình dục hơn bạn gái không? Không hẳn. Như đã nói, các tìm kiếm Google có xu hướng thiên lệch về những thứ khó nói. Nam giới có thể cảm thấy thoải mái hơn nữ giới khi nói với bạn bè rằng nửa kia của mình ít quan tâm chuyện tình dục. Tuy nhiên, ngay cả nếu dữ liệu Google không ám chỉ rằng nam giới thực sự tránh sinh hoạt gấp đôi nữ giới, nó cũng chứng tỏ rằng trên thực tế, nam giới tránh sinh hoạt nhiều hơn là họ vẫn nói.

Dữ liệu Google còn chỉ ra một lí do khiến người ta tránh chuyện tình dục quá thường xuyên: sự lo lắng quá độ, mà phần nhiều là lo lắng sai chỗ. Hãy bắt đầu với các mối lo của nam giới. Việc nam giới lo lắng về độ lớn của "phụ tùng" không phải chuyện gì mới, nhưng mức độ của sự lo lắng này khá sâu xa.

Nam giới Google các câu hỏi về cơ quan sinh dục của họ nhiều hơn các bộ phận cơ thể khác: nhiều hơn tổng số câu hỏi về phổi, gan, bàn chân, tai, mũi, họng, và não gộp lại. Nam giới thực hiện các tìm kiếm về cách làm dương vật to lên nhiều hơn cách lên dây guitar, làm trứng ốp la, hoặc cách thay bánh xe. Lo lắng được Google nhiều nhất của nam giới về các chất steroid không phải là nó có gây hại sức khỏe hay không, mà là dùng nó có làm teo dương vật hay không. Câu hỏi được Google nhiều nhất của nam giới về chuyện thể xác hoặc tinh thần của họ sẽ thay đổi thế nào khi họ lớn tuổi chính là dương vật có bị nhỏ lại hay không.

Phụ chú: Một trong các câu hỏi phổ biến hơn trên Google liên quan đến cơ quan sinh dục nam là "How big is my p*nis?" ("Hàng" của tôi to bao nhiêu?). Chuyện người ta hỏi Google câu này thay vì dùng thước, theo tôi, chính là biểu hiện của tinh hoa thời đại kĩ thuật số ngày nay.[1]

---

[1] Tôi muốn gọi quyển sách này là *How Big Is My P*nis? What Google Searches Teach Us About Human Nature (Dương vật của tôi lớn cỡ nào? Những gì các tìm kiếm Google dạy chúng ta về bản chất con người)*, nhưng nhà biên tập của tôi cảnh báo rằng như thế sẽ rất khó bán, có thể người ta ngượng không dám mua quyển sách có tựa đề đó tại một hiệu sách ở sân bay. Bạn đồng ý chứ?

Phụ nữ có quan tâm về kích thước dương vật không? Hiếm, theo các tìm kiếm Google. Cứ mỗi tìm kiếm của phụ nữ về súng của đối phương, tương ứng sẽ có khoảng 170 tìm kiếm của đàn ông về súng của chính họ. Thực sự, cũng có những lúc hiếm hoi phụ nữ thể hiện nỗi lo lắng về dương vật của đối phương, thường là về kích thước, nhưng không hẳn là vì nó nhỏ. Hơn 40% phàn nàn về kích thước dương vật của đối phương nói rằng nó quá lớn. "Đau" là từ được Google nhiều nhất trong các tìm kiếm với cụm từ "___ trong khi quan hệ." ("Chảy máu," "tè," "khóc thét," và "xì hơi" làm thành top 5.) Tuy nhiên, chỉ 1% các tìm kiếm của nam giới nhằm thay đổi kích thước dương vật là tìm thông tin về cách làm cho nó nhỏ lại.

Câu hỏi về tình dục phổ biến thứ hai của nam giới là cách kéo dài quá trình giao hợp. Một lần nữa, những bất an của nam giới dường như không khớp với những lo lắng của nữ giới. Con số tìm kiếm hỏi cách làm cho bạn trai lên đỉnh nhanh hơn và lên đỉnh chậm hơn gần như bằng nhau. Thực ra, quan tâm phổ biến nhất của phụ nữ liên quan đến sự cực khoái của bạn trai không phải là nó diễn ra khi nào, mà là tại sao nó không chịu diễn ra.

Chúng ta không thường bàn về các vấn đề hình ảnh cơ thể khi nói đến nam giới. Và mặc dù đúng là chuyện quan tâm đến ngoại hình cá nhân nói chung nghiêng về phía nữ giới, tình hình nay cũng không còn thiên lệch như các định kiến xưa nay. Theo phân tích Google AdWords của tôi—tập trung đo lường các website người ta hay lui tới—trong tổng số người quan tâm đến sắc đẹp và sự săn chắc thì có 42% nam, giảm cân là 33% nam, và giải phẫu thẩm mĩ là 39% nam. Trong số tất cả tìm kiếm theo mô-típ "how to..." (cách để...) liên quan đến ngực, khoảng 20% hỏi cách loại bỏ "ngực xệ" ở đàn ông.

Tuy nhiên, dù số nam giới thiếu tự tin về cơ thể của họ cao hơn ta nghĩ, thì trong vấn đề bất an về ngoại hình, nữ vẫn cứ nhiều hơn nam. Vậy huyết thanh sự thật số có thể tiết lộ điều gì về sự thiếu tự tin của nữ giới? Hàng năm tại Mĩ, có hơn 7,000,000 tìm kiếm nhắm vào việc nâng ngực. Các con số thống kê chính thức cho chúng ta biết rằng khoảng 300,000 phụ nữ đi nâng ngực hàng năm.

Nữ giới còn khá bất an về phần mông. Tuy vậy, thời gian gần đây, nhiều phụ nữ đã đổi ý về điều mà họ không thích ở vòng 3 của mình.

Năm 2004, ở một số vùng tại Mĩ, tìm kiếm phổ biến nhất về việc thay đổi phần bàn tọa là cách làm cho nó nhỏ lại. Mong muốn làm cho mông lớn hơn đại đa số tập trung ở những vùng có nhiều người da đen. Tuy nhiên, bắt đầu năm 2010, ước muốn có mông lớn hơn tăng lên ở những nơi còn lại của nước Mĩ. Sự quan tâm này đã tăng gấp 3 lần trong vòng 4 năm. Năm 2014, các tìm kiếm hỏi cách làm mông lớn lên nhiều hơn là cách làm mông nhỏ lại ở mọi tiểu bang. Ngày nay, cứ mỗi 5 tìm kiếm nhắm vào nâng ngực ở Mĩ thì lại có 1 tìm kiếm nhắm vào nâng mông. (Cảm ơn, Kim Kardashian!)

Sự gia tăng sở thích có một cái mông lớn hơn ở nữ giới liệu có khớp với các sở thích của nam giới hay không? Thật thú vị, khớp. Các tìm kiếm "khiêu dâm mông to," vốn từng tập trung ở các cộng đồng người da đen, gần đây đã tăng vọt trên khắp nước Mĩ.

Nam giới còn muốn gì khác ở cơ thể phụ nữ? Như đã đề cập, và như hầu hết các bạn đều sẽ thấy rất hiển nhiên, nam giới tỏ ra thích ngực lớn. Khoảng 12% tìm kiếm hình ảnh khiêu dâm có chủ đề cụ thể là tìm kiếm ngực lớn. Tìm kiếm này cao hơn gần 20 lần lượng tìm kiếm khiêu dâm ngực nhỏ.

Dù vậy, chẳng rõ điều này có đồng nghĩa với việc nam giới muốn nữ giới nâng ngực hay không. Khoảng 3% tìm kiếm khiêu dâm ngực lớn nói rõ họ muốn nhìn ngực tự nhiên.

Các tìm kiếm Google về vợ và nâng ngực được chia ra hai nửa ngang nhau: Một nửa thì hỏi cách thuyết phục vợ nâng ngực, còn nửa còn lại thì không hiểu tại sao vợ lại muốn nâng ngực.

Thử nhìn tìm kiếm phổ biến nhất về ngực của bạn gái: "Tôi yêu ngực của bạn gái tôi." Không rõ nam giới đang hi vọng tìm thấy gì từ Google khi thực hiện tìm kiếm này.

Tựa như nam giới, nữ giới cũng có những câu hỏi về phụ tùng của họ. Thực vậy, họ thắc mắc về âm đạo nhiều gần bằng nam giới thắc mắc

về dương vật. Những lo lắng của nữ giới về âm đạo thường liên quan đến sức khỏe. Nhưng ít nhất 30% các câu hỏi của họ bàn về những nỗi lo khác. Nữ giới muốn biết cách cạo, làm khít, và cải thiện hương vị của nó. Một quan tâm phổ biến đến ngạc nhiên là cách cải thiện mùi của nó.

Nữ giới thường lo lắng nhất là âm đạo có mùi cá, tiếp theo là giấm, hành, ammonia, tỏi, phô mai, mùi thân thể, nước tiểu, bánh mì, chất tẩy trắng, phân, mồ hôi, kim loại, bàn chân, rác, và thịt thối.

Nói chung, nam giới không thực hiện nhiều tìm kiếm Google liên quan đến cơ quan sinh dục của đối phương. Nam giới thực hiện tìm kiếm về âm đạo của bạn gái khoảng cùng số lượng nữ giới tìm kiếm về dương vật của bạn trai.

Khi nam giới tìm kiếm về âm đạo của đối phương, đó thường là để phàn nàn về điều nữ giới lo lắng nhất: mùi. Chủ yếu, nam giới đang cố tìm cách nói với một phụ nữ về mùi hôi mà không làm tổn thương cô ấy. Tuy nhiên, đôi khi các câu hỏi của nam giới về mùi tiết lộ sự bất an của chính họ. Nam giới thỉnh thoảng hỏi cách dùng mùi để xác định chuyện ngoại tình—ví dụ, họ thắc mắc khi đó có phải nó có mùi bao cao su, hoặc mùi tinh dịch của một người đàn ông khác.

Chúng ta nên làm gì với tất cả sự bất an thầm kín này? Rõ ràng có vài tin tốt lành ở đây. Google cho ta những lí do hợp lí để bớt lo hơn. Phần nhiều các nỗi sợ sâu xa nhất về cảm nhận của bạn tình đối với mình là phi lí. Một mình, bên máy tính, không có động cơ nói dối, những người bạn tình cho thấy mình thường khá sâu sắc và độ lượng. Thực vậy, chúng ta tất cả đều quá bận phán xét cơ thể chính mình nên ít còn năng lượng để phán xét cơ thể của người khác.

Có thể còn một mối quan hệ giữa 2 trong số những nỗi lo lớn được tiết lộ trong các tìm kiếm về tình dục trên Google: thiếu hoạt động tình dục và bất an về sự hấp dẫn cũng như hiệu suất tình dục. Có thể 2 vấn đề này có liên quan. Có thể nếu bớt lo về tình dục, ta sẽ sinh hoạt tình dục nhiều hơn.

Các tìm kiếm Google có thể cho ta biết gì khác về tình dục? Ta có thể so sánh giữa hai giới, để thấy ai rộng lượng hơn. Xét các tìm kiếm nhắm

đến việc cải thiện trình độ "yêu" bằng miệng khác giới. Ai tìm nhiều bí quyết loại này hơn, nam hay nữ? Ai rộng lượng về mặt tình dục hơn, nam giới hay nữ giới? Là nữ giới đó bạn. Gộp hết tất cả các khả năng, tôi ước tính tỉ lệ là 2:1 nghiêng về nữ giới, xét về chuyện tìm kiếm bí quyết "yêu" bằng miệng nhằm thỏa mãn đối phương.

Và khi nam tìm kiếm lời khuyên về cách "yêu" bằng miệng, họ thường không phải đang tìm cách thỏa mãn người khác. Nam giới tìm kiếm cách tự làm cho chính mình nhiều tương đương cách làm cho phụ nữ. (Thông tin này là một trong số những điều tôi tâm đắc nhất ở dữ liệu tìm kiếm Google.)

## Sự thật về thái độ thù ghét và thành kiến

Tình dục và tình cảm lãng mạn không phải là các chủ đề duy nhất bị nỗi xấu hổ bao trùm, và vì vậy, cũng không phải là các chủ đề duy nhất mà người ta giữ bí mật. Nhiều người, vì lí do tốt đẹp, có xu hướng giấu các thành kiến trong tư tưởng mình. Nhiều người cảm thấy họ sẽ bị phán xét nếu thú nhận là họ phán xét người khác dựa trên sắc tộc, xu hướng tính dục, hoặc tôn giáo. Bạn có thể gọi đây là sự tiến bộ. Thế nhưng, nhiều người Mĩ vẫn thú nhận. (Xin cảnh báo bạn đọc, đây lại là một phần khác chứa nội dung gây khó chịu.)

Bạn có thể thấy điều này trên Google. Người dùng thỉnh thoảng đặt các câu hỏi như "Tại sao người da đen thô lỗ?" hoặc "Tại sao người Do Thái xấu xa?" Bên dưới, theo thứ tự, là top 5 từ tiêu cực dùng trong các tìm kiếm về các nhóm người.

Một vài khuôn mẫu nổi bật lên. Ví dụ, người Mĩ gốc Phi là nhóm duy nhất đối mặt thành kiến "thô lỗ." Gần như nhóm nào cũng là nạn nhân của thành kiến "ngu xuẩn"; hai nhóm duy nhất không bị ngu là Do Thái và Hồi giáo. Thành kiến "xấu xa" áp dụng cho người Do Thái, người Hồi giáo, người đồng tính không phải da đen, người Mễ, người châu Á, và người Cơ Đốc.

Người Hồi giáo là nhóm duy nhất bị thành kiến là khủng bố. Khi một người Mĩ gốc Hồi giáo hành xử giống với thành kiến này, phản ứng

có thể tức thời và nguy hiểm. Dữ liệu tìm kiếm Google có thể cho ta một cái nhìn từng phút vào những đợt bộc phát cơn giận nạp đầy lòng thù hận ấy.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **MĨ GỐC PHI** | **rude** thô lỗ | **racist** phân biệt chủng tộc | **stupid** ngu xuẩn | **ugly** xấu xí | **lazy** lười biếng |
| **DO THÁI** | **evil** xấu xa | **racist** phân biệt chủng tộc | **ugly** xấu xí | **cheap** ti tiện | **greedy** tham lam |
| **HỒI** | **evil** xấu xa | **terrorists** khủng bố | **bad** xấu tính | **violent** bạo lực | **dangerous** nguy hiểm |
| **MỄ** | **racist** phân biệt chủng tộc | **stupid** ngu xuẩn | **ugly** xấu xí | **lazy** lười biếng | **dumb** dốt đặc |
| **CHÂU Á** | **ugly** xấu xí | **racist** phân biệt chủng tộc | **annoying** phiền phức | **stupid** ngu xuẩn | **cheap** ti tiện |
| **ĐỒNG TÍNH** | **evil** xấu xa | **wrong** sai trái | **stupid** ngu xuẩn | **annoying** phiền phức | **selfish** ích kỉ |
| **CƠ ĐỐC** | **stupid** ngu xuẩn | **crazy** khùng điên | **dumb** dốt đặc | **delusional** ảo tưởng | **wrong** sai trái |

Hãy xem chuyện đã diễn ra ngay sau cuộc xả súng ở San Bernardino, California, ngày 2/12/2015. Sáng hôm đó, Rizwan Farook và Tashfeen Malik đi vào một cuộc họp với các đồng nghiệp của Farook, mang theo súng lục, súng trường bán tự động và giết chết 14 người. Tối đó, chính xác là mấy phút sau khi truyền thông lần đầu tường thuật cái tên có âm hưởng Hồi giáo của một trong hai kẻ xả súng, một số người California đã quyết định xong điều họ muốn làm với người Hồi giáo: giết.

Tìm kiếm Google nhiều nhất ở California có chứa từ "Hồi giáo" lúc bấy giờ là "giết bọn Hồi giáo." Và nói chung, người Mĩ tìm kiếm cụm từ "giết bọn Hồi giáo" ngang với "công thức rượu martini," "triệu chứng đau nửa đầu," và "danh sách đội Cowboys." Vào những ngày theo sau

vụ San Bernardino, cứ mỗi người Mĩ quan tâm đến "Islamophobia" (chứng sợ người Hồi giáo) thì lại có một người tìm kiếm cụm từ "giết bọn Hồi giáo." Trước vụ xả súng, các tìm kiếm thù ghét chiếm khoảng 20% tất cả tìm kiếm về người Hồi giáo. Những giờ sau vụ xả súng, hơn một nửa tổng số tìm kiếm về người Hồi giáo bỗng đầy thù hận.

Và dữ liệu tìm kiếm theo từng phút cho thấy làm dịu cơn cuồng nộ đó là khó vô cùng. Sau vụ xả súng 4 ngày, tổng thống bấy giờ là Obama đã có một bài diễn văn vào giờ cao điểm gửi đến cả nước. Ông muốn bảo đảm với người Mĩ rằng chính phủ có thể vừa ngăn chặn nạn khủng bố vừa trấn an chứng sợ người Hồi giáo nguy hiểm này.

Obama kêu gọi lòng tốt của mọi người, nói về tầm quan trọng của sự đoàn kết và lòng khoan dung. Bài diễn văn thật mạnh mẽ và cảm động. Tờ *Los Angeles Times* ca ngợi Obama đã "[cảnh báo] không để nỗi sợ hãi che mờ óc phán xét của chúng ta." Tờ *New York Times* gọi bài diễn văn là vừa "cứng rắn" vừa "trấn an." Trang Think Progress ca ngợi đó là "một công cụ cần thiết của sự cai trị tốt, hướng đến việc cứu cuộc sống những người Mĩ Hồi giáo." Bài diễn văn của Obama, nói cách khác, được cho là thành công lớn. Nhưng có phải vậy không?

Dữ liệu tìm kiếm Google cho thấy điều ngược lại. Tôi cùng Evan Soltas (lúc đó còn ở Princeton) khảo sát dữ liệu. Trong bài diễn văn, tổng thống nói, "Trách nhiệm của tất cả người Mĩ—với mọi niềm tin tôn giáo—là loại bỏ sự phân biệt đối xử." Nhưng các tìm kiếm gọi người Hồi giáo là "bọn khủng bố," "xấu xa," "bạo lực," và "ma quỷ" tăng gấp đôi trong và ngay sau bài diễn văn. Tổng thống còn nói, "Trách nhiệm của chúng ta là loại bỏ các bài kiểm tra tôn giáo với những người mà chúng ta chấp nhận được vào đất nước này." Nhưng tìm kiếm tiêu cực về người Syria tị nạn—một nhóm chủ yếu là người Hồi giáo bấy giờ đang tuyệt vọng tìm nơi ẩn náu an toàn—lại tăng 60%, trong khi số tìm kiếm về cách giúp người Syria tị nạn giảm 35%. Obama yêu cầu người Mĩ "đừng quên rằng tự do mạnh hơn nỗi sợ." Tuy nhiên, tìm kiếm "giết bọn Hồi giáo" tăng gấp 3 trong suốt bài diễn văn của ông. Thực vậy, hầu như mọi tìm kiếm tiêu cực về người Hồi giáo tăng vọt trong và sau bài diễn văn của Obama, và hầu như mọi tìm kiếm tích cực đều sụt giảm.

Nói cách khác, Obama có vẻ đã nói toàn những điều đúng. Tất cả các phương tiện truyền thông truyền thống chúc mừng Obama về lời lẽ hàn gắn vết thương của ông. Nhưng dữ liệu mới từ Internet, huyết thanh sự thật số, cho thấy bài diễn văn đó thực tế đã phản lại mục tiêu chính của nó. Thay vì trấn an đám đông giận dữ như mọi người đã nghĩ, dữ liệu Internet cho ta biết rằng Obama thực tế đã kích động đám đông. Thứ ta nghĩ là hiệu quả có thể gây ảnh hưởng hoàn toàn trái ngược với mong đợi. Đôi khi ta cần dữ liệu Internet để bớt tự khen ngợi chính mình.

Vậy lẽ ra Obama nên nói gì để dập tắt hình thức căm thù đang bùng phát khắp nước Mĩ này đây? Chúng ta sẽ quay lại sau. Ngay bây giờ chúng ta sẽ xem xét một thành kiến lâu đời tại Mĩ, hình thức căm ghét thực sự nổi bật hơn các hình thức còn lại, sự căm ghét có tính hủy hoại nhất và là chủ đề của nghiên cứu đã khởi đầu cho quyển sách này. Trong công trình của tôi với dữ liệu tìm kiếm Google, một sự thật nói lên nhiều điều nhất mà tôi phát hiện về lòng căm ghét trên Internet là sự phổ biến của từ "nigger" (mọi đen).

Ở hình thức số ít hoặc số nhiều, từ "nigger(s)" xuất hiện trong 7,000,000 tìm kiếm của người Mĩ hàng năm. (Một lần nữa, từ này dùng trong các bản nhạc rap hầu như luôn là "nigga," không phải "nigger," vậy là ảnh hưởng từ lời nhạc hip-hop là không đáng kể.) Tìm kiếm "nigger jokes" phổ biến gấp 17 lần các tìm kiếm "kike jokes," "gook jokes," "spic jokes," "chink jokes," và "fag jokes" gộp lại.[1]

Các tìm kiếm "nigger(s)"—hoặc "nigger jokes"—phổ biến nhất khi nào? Bất cứ khi nào người Mĩ gốc Phi xuất hiện trên tin tức. Một trong số các thời điểm mà những tìm kiếm trên xuất hiện nhiều nhất là ngay sau Siêu bão Katrina, khi truyền hình và báo chí chiếu hình ảnh những người da đen tuyệt vọng ở New Orleans đang cố gắng để sống sót. Các tìm kiếm đó cũng tăng vọt trong kì bầu cử đầu tiên của Obama. Và các tìm kiếm "nigger jokes" tăng trung bình khoảng 30% vào ngày lễ Martin Luther King Jr.

---

[1] [ND] Đây là các câu chuyện cười với từ ngữ miệt thị từng nhóm người: kike: Do Thái; gook: châu Á; spic: Nam Mĩ; chink: Trung Quốc; fag: đồng tính.

Sự hiện diện khắp nơi đáng sợ của vết nhơ chủng tộc này khiến ta hoài nghi một số hiểu biết hiện tại về sự phân biệt chủng tộc.

Bất cứ học thuyết phân biệt chủng tộc nào cũng phải giải thích một vấn đề nan giải lớn tại Mĩ. Một mặt, đại đa số người Mĩ da đen nghĩ họ bị thành kiến—họ có dư bằng chứng về phân biệt đối xử ở các trạm cảnh sát, các cuộc phỏng vấn việc làm, và các quyết định của tòa án. Mặt khác, rất ít người Mĩ da trắng thú nhận mình là phân biệt chủng tộc.

Lời giải thích vượt trội nhất trong giới khoa học gia chính trị gần đây cho rằng điều này phần lớn là do định kiến tiềm thức lan rộng. Người Mĩ da trắng có thể có tư tưởng tốt đẹp, nhưng họ cũng có một định kiến tiềm ẩn, nó ảnh hưởng cách họ đối xử với người Mĩ da đen. Các nhà học thuật đã phát minh một phép thử hay cho định kiến ấy. Nó được gọi là bài kiểm tra liên tưởng ngầm (implicit association test).

Các thử nghiệm luôn chỉ ra rằng, để liên tưởng các khuôn mặt người da đen với từ tích cực (như "tốt"), hầu hết mọi người mất thêm vài ms so với từ tiêu cực (như "khiếp"). Với khuôn mặt người da trắng, kết quả ngược lại. Thời gian dôi thêm đó là bằng chứng có tồn tại định kiến tiềm thức—thứ thành kiến mà có thể chính bản thân người có thành kiến cũng không ý thức được.

Tuy nhiên, có một lời giải thích khác cho sự phân biệt đối xử mà người Mĩ gốc Phi cảm thấy và người da trắng phủ nhận: sự phân biệt chủng tộc *có ý thức* nhưng đã được che giấu. Giả sử sự phân biệt chủng tộc có ý thức là phổ biến, người ta biết rất rõ điều đó nhưng họ không chịu thú nhận—và dĩ nhiên là không thú nhận trong một cuộc khảo sát. Đó là điều mà dữ liệu tìm kiếm dường như muốn nói. Rõ là không phải tiềm thức khi người ta tìm kiếm "nigger jokes." Và khó mà tin rằng người Mĩ đang Google từ "nigger" với cùng tần số như cụm từ "đau nửa đầu" và "nhà kinh tế" mà trong đầu không chất chứa sự phân biệt chủng tộc *có ý thức* về người Mĩ gốc Phi. Trước thời dữ liệu Google, ta không có một thước đo thuyết phục về sự thù ghét lây lan này. Bây giờ thì ta đã có. Vì vậy, ta đã có thể thấy được nhiều điều từ đó.

Như đã biết, dữ liệu giải thích tại sao số phiếu bầu của Obama năm

2008 và 2012 bị sụt ở nhiều vùng. Nó cũng tương quan với khoảng cách tiền lương da trắng-da đen, theo công bố của một nhóm nhà kinh tế gần đây. Nói cách khác, các vùng mà tôi đã phát hiện có nhiều tìm kiếm phân biệt chủng tộc nhất cũng trả lương thấp cho người da đen. Và sau đó là hiện tượng Donald Trump. Như đã nhận xét ở phần giới thiệu, khi Nate Silver, chuyên gia thăm dò ý kiến, tìm biến số địa lí tương quan mạnh nhất với sự ủng hộ cuộc bầu cử sơ bộ đảng Cộng hòa năm 2016 cho Trump, ông đã phát hiện biến số cần tìm trong bản đồ phân biệt chủng tộc tôi đã phát triển. Biến số đó là các tìm kiếm "nigger(s)."

Các học giả gần đây đã xây dựng được thước đo thành kiến ngầm đối với người da đen từng tiểu bang, cho phép tôi so sánh ảnh hưởng của sự phân biệt chủng tộc *có ý thức*, được đo bởi các tìm kiếm Google, và định kiến tiềm thức. Ví dụ, tôi kiểm tra xem sự phân biệt có ý thức và tiềm thức đã chống lại Obama bao nhiêu trong cả hai lần bầu cử tổng thống của ông. Dùng phép phân tích hồi quy, tôi thấy rằng, để dự báo Obama có thành tích kém ở đâu, các tìm kiếm Google phân biệt chủng tộc của một vùng sẽ rất hữu ích. Thành tích của Obama rất ít tương quan với kết quả bài kiểm tra liên tưởng ngầm.

Để kích thích và khuyến khích thêm nghiên cứu trong lĩnh vực này, tôi xin đưa ra phỏng đoán sau đây, sẵn sàng được kiểm tra bởi các học giả trong nhiều lĩnh vực. Cách giải thích đầu tiên cho sự phân biệt đối xử với người Mĩ gốc Phi ngày nay không phải là vì những người đồng ý tham gia thí nghiệm đã có các liên tưởng tiềm thức giữa các từ tiêu cực và người da đen, mà là vì hàng triệu người Mĩ da trắng vẫn đang làm những việc cho thấy sự phân biệt chủng tộc có ý thức, như tìm kiếm cụm từ "nigger jokes" chẳng hạn.

***

Sự phân biệt đối xử mà người da đen thường gặp tại Mĩ dường như được tiếp nhiên liệu bởi sự thù địch có ý thức bị che giấu. Nhưng, với những nhóm khác, thành kiến trong tiềm thức có thể có tác động căn cơ hơn. Ví dụ, tôi có thể dùng các tìm kiếm Google để tìm bằng chứng về định kiến tiềm thức chống một bộ phận dân số khác: các cô gái trẻ.

Vậy ai đang ấp ủ định kiến về các cô gái trẻ?

Chính là cha mẹ họ.

Hầu như chẳng có gì ngạc nhiên khi các bậc cha mẹ thường hào hứng nghĩ là con họ có năng khiếu. Thực vậy, trong tất cả tìm kiếm Google bắt đầu bằng "Có phải bé 2 tuổi của tôi...," từ tiếp theo phổ biến nhất là "có năng khiếu." Nhưng câu hỏi này không được hỏi ngang bằng nhau giữa các bé trai và bé gái. Khả năng cha mẹ hỏi "Con trai tôi có năng khiếu không?" nhiều gấp 1.5 lần "Con gái tôi có năng khiếu không?" Cha mẹ thể hiện một định kiến tương tự khi dùng các cụm từ khác liên quan đến trí tuệ mà họ ngại không nói ra, như, "Con trai tôi có phải là thiên tài?"

Phải chăng các bậc cha mẹ đang phát hiện ra những khác biệt hợp lí giữa các bé gái và bé trai? Có lẽ các bé trai hay dùng từ đao to búa lớn hơn các bé gái, hoặc có dấu hiệu khách quan về tài năng? Không. Có chăng là ngược lại. Ở các độ tuổi nhỏ, con gái luôn được chứng minh là có vốn từ vựng lớn hơn và dùng nhiều câu phức hơn. Tại các trường ở Mĩ, khả năng góp mặt ở các chương trình năng khiếu của bé gái cao hơn bé trai 9%. Dù vậy, cha mẹ khi nhìn quanh bàn ăn gia đình dường như lại thấy nhiều bé trai tài năng hơn bé gái tài năng.[1] Thực vậy, trên mọi từ tìm kiếm liên quan đến trí tuệ mà tôi kiểm tra, bao gồm cả các tìm kiếm chỉ sự thiếu trí tuệ, cha mẹ có nhiều khả năng hỏi về con trai hơn là con gái. Cũng có nhiều tìm kiếm "có phải con trai tôi chậm" hoặc "ngu" hơn là các tìm kiếm tương tự với con gái. Nhưng các tìm kiếm có từ tiêu cực như "chậm" và "ngu" ít thiên lệch rõ ràng về phía con trai hơn các tìm kiếm có từ tích cực, chẳng hạn như "có tài" hoặc "thiên tài."

Thế thì các nỗi lo lắng lớn nhất của cha mẹ liên quan đến con gái của họ là gì? Trước tiên là bất cứ cái gì liên quan đến ngoại hình. Xem xét các câu hỏi về cân nặng của trẻ. Cha mẹ gõ Google câu "Con gái tôi có thừa cân không?" gấp đôi câu "Con trai tôi có thừa cân không?" Số lần cha mẹ

[1] Để kiểm tra sâu hơn giả thiết cha mẹ đối xử với con trai và con gái khác nhau, tôi đang thực hiện thu thập dữ liệu từ các website nuôi dạy con cái. Ở đây có số cha mẹ lớn hơn nhiều số cha mẹ thực hiện các tìm kiếm cụ thể nêu trên.

hỏi cách giúp con gái họ giảm cân gấp đôi số lần hỏi cách giúp con trai họ giảm cân. Cũng tương tự chuyện tài năng, định kiến giới tính này không dựa trên cơ sở thực tế. Khoảng 28% bé gái thừa cân, trong khi đó 35% cậu trai thừa cân. Ngay cả dù tỉ lệ cho thấy con trai thừa cân nhiều hơn con gái, cha mẹ vẫn lo hoặc nghĩ rằng con gái thường thừa cân hơn nhiều so với con trai.

Phụ huynh cũng thường hỏi xem con gái họ đẹp hay không, gấp rưỡi số lần hỏi con trai họ đẹp hay không. Chuyện xấu hay không cũng vậy: Phụ huynh cũng thường hỏi xem con gái họ xấu hay không, gấp 3 số lần hỏi con trai họ xấu hay không. (Cũng chẳng thể hiểu tại sao người ta lại nghĩ Google biết là con họ đẹp hay xấu nữa.)

Nói chung, cha mẹ thường dùng các từ tích cực hơn trong các câu hỏi về con trai họ. Họ hay hỏi *có phải con trai họ "vui vẻ" không* hơn là hỏi *có phải con trai họ "trầm cảm" không*.

Độc giả có quan điểm tự do có thể nghĩ rằng các định kiến này phổ biến ở các vùng bảo thủ, nhưng tôi không tìm thấy bằng chứng nào cả. Thực vậy, tôi không tìm thấy mối quan hệ có ý nghĩa giữa bất cứ định kiến nào ở đây với tình hình chính trị hoặc văn hóa của bang. Cũng không có bằng chứng là các định kiến này đã giảm kể từ năm 2004, năm mà dữ liệu tìm kiếm Google mới xuất hiện. Dường như định kiến tiêu cực về các bé gái này lan rộng và ăn sâu hơn ta nghĩ.

***

Giới tính không phải là nơi duy nhất định kiến xuất hiện.

Vikingmaiden88 là một cô gái 26 tuổi. Cô thích đọc lịch sử và làm thơ. Câu trích dẫn ở phần chữ kí của cô là một câu của Shakespeare. Tôi lượm lặt tất cả thông tin này từ mục tiểu sử và các đăng tải của cô trên Stormfr***, trang thù địch trực tuyến phổ biến nhất của Mĩ. Tôi cũng biết rằng Vikingmaiden88 thích nội dung trên trang của tờ báo mà tôi làm việc, tờ *New York Times*. Cô có một đăng tải nhiệt tình về một bài ở *Times*.

Gần đây tôi đã phân tích hàng chục ngàn tiểu sử Stormfr*** như thế, trong đó các thành viên có thể nhập địa điểm, ngày sinh, sở thích, và các

thông tin khác.

Stormfr*** được sáng lập năm 1995 bởi Don Black, một cựu thủ lĩnh Ku Klux Klan. Các "nhóm xã hội" phổ biến nhất của trang là "Hiệp hội đảng viên Quốc xã" và "Người hâm mộ và ủng hộ Adolf Hitler." Năm qua, theo Quantcast, khoảng 200,000 đến 400,000 người Mĩ vào trang này mỗi tháng. Một báo cáo gần đây của Southern Poverty Law Center chỉ ra gần 100 vụ án mạng trong 5 năm qua liên quan tới thành viên Stormfr***.

Các thành viên Stormfr*** không phải là những người tôi đã đoán. Họ thường trẻ, ít nhất là theo thông tin ngày sinh tự khai báo. Độ tuổi bắt đầu tham gia trang này phổ biến nhất là 19. Và thành viên 19 tuổi đăng kí nhiều gấp 4 lần thành viên 40 tuổi. Người dùng Internet và mạng xã hội thường trẻ, nhưng không trẻ đến thế.

Phần tiểu sử không có mục giới tính. Nhưng tôi đã xem tất cả các đăng tải và tiểu sử đầy đủ của một mẫu đại diện ngẫu nhiên người dùng Mĩ, và hóa ra ta có thể tìm ra giới tính của hầu hết thành viên: Tôi ước tính rằng khoảng 30% các thành viên Stormfr*** là nữ.

Các bang có tỉ lệ số thành viên Stormfr*** trên tổng số dân cao nhất là Montana, Alaska, và Idaho. Các bang này chủ yếu là da trắng. Phải chăng điều này có nghĩa là chuyện lớn lên trong môi trường ít đa dạng sẽ thúc đẩy sự thù ghét?

Không phải. Trái lại, vì các tiểu bang đó có tỉ lệ người da trắng (không bao gồm người Do Thái) cao hơn, nên khả năng có thành viên trang web chuyên tấn công người Do Thái và người không phải da trắng cũng sẽ cao hơn. Tỉ lệ người dùng mục tiêu của Stormfr*** tham gia trên thực tế lại cao ở những vùng có nhiều nhóm thiểu số. Điều này đặc biệt đúng khi khảo sát các thành viên của Stormfr*** từ tuổi 18 trở xuống, những người không thể tự chọn nơi sinh sống.

Trong nhóm tuổi này, California, bang có số dân thuộc nhóm thiểu số vào loại đông nhất, có tỉ lệ thành viên cao hơn 25% số trung bình cả nước.

Một trong các nhóm phổ biến nhất trên đây là "Ủng hộ chủ nghĩa bài

Do Thái." Tỉ lệ các thành viên tham gia nhóm này tương quan rõ ràng với dân số người Do Thái của bang. New York, bang có đông người Do Thái nhất, có tỉ lệ thành viên trên tổng dân số cao hơn trung bình trong nhóm này.

Năm 2001, Dna88 tham gia Stormfr***, tự mô tả mình là lập trình viên Internet 30 tuổi, "dễ nhìn, ý thức về chủng tộc" sống ở "*Jew* York City." Bốn tháng sau, anh ta viết hơn 200 bài đăng, như "Các tội ác Do Thái chống nhân loại" và "Tiền máu Do Thái," và chỉ người ta đến một website, jew*****.com. Trang này tự xưng là một "thư viện học thuật" về "tính chất tội phạm của phong trào phục quốc Do Thái."

Thành viên Stormfr*** phàn nàn về việc các nhóm thiểu số nói ngôn ngữ khác và thường hay phạm tội. Nhưng điều tôi thấy thú vị nhất là những lời phàn nàn về sự cạnh tranh trong thị trường hẹn hò.

Một ông tự xưng là William Lyon Mackenzie King, ăn theo tên một cựu thủ tướng Canada từng đề nghị rằng "Canada nên là một nước của người da trắng," đã viết vào năm 2003 rằng ông cố gắng để "kìm chế cơn thịnh nộ" của mình sau khi nhìn thấy một phụ nữ da trắng "bế trên tay thằng nhỏ lai mọi đen xấu xí." Trong tiểu sử của mình, Whitepride26, một nữ sinh viên 41 tuổi ở Los Angeles, nói, "Tôi ghét người da đen, người Latin, và đôi khi là người châu Á, đặc biệt khi cánh đàn ông thấy bọn họ hấp dẫn hơn một phụ nữ da trắng."

Một số diễn biến chính trị nhất định cũng góp phần. Ngày chứng kiến sự gia tăng thành viên lớn nhất trong lịch sử Stormfr*** cho đến nay là ngày 5/11/2008, ngay sau khi Barack Obama được bầu làm tổng thống. Tuy nhiên, không có sự gia tăng mức độ quan tâm tại Stormfr*** trong đợt ứng cử của Donald Trump, và chỉ tăng chút ít ngay sau khi ông thắng cử. Trump chỉ cưỡi con sóng chủ nghĩa dân tộc da trắng mà thôi. Không có bằng chứng cho thấy rằng chính ông là người tạo ra con sóng chủ nghĩa dân tộc da trắng.

Cuộc bầu cử của Obama làm dấy lên phong trào chủ nghĩa dân tộc da trắng. Cuộc bầu cử của Trump dường như chỉ là một sự hồi đáp.

Một thứ có vẻ như không ảnh hưởng gì nhiều: kinh tế. Không có mối quan hệ giữa số thành viên đăng kí mới hàng tháng với tỉ lệ thất nghiệp của bang. Các tiểu bang bị ảnh hưởng Đại suy thoái nhiều cũng không có sự gia tăng tương đối trong các tìm kiếm trang Stormfr*** trên Google.

Nhưng có lẽ điều thú vị và đáng ngạc nhiên nhất là một số chủ đề trao đổi của các thành viên Stormfr***. Nó tương tự các chủ đề mà bạn tôi và tôi thường bàn. Có thể là do tôi ngây thơ, nhưng tôi đã hình dung những người theo chủ nghĩa dân tộc da trắng sống ở một thế giới khác với nhóm bạn bè chúng tôi. Không phải thế, họ vẫn có những chuỗi bài trò chuyện dài ca ngợi loạt phim truyền hình *Game of Thrones* hay so sánh các trang hẹn hò trực tuyến, như PlentyOfFish và OkCupid.

Và một điểm mấu chốt cho thấy người dùng Stormfr*** đang sống trong thế giới tương tự những người như tôi: sự phổ biến của *New York Times* trong số người dùng Stormfr***. Không chỉ có Vikingmaiden88 lui tới trang của *Times*. Trang này phổ biến đối với nhiều thành viên Stormfr*** khác. Thực vậy, khi so sánh người dùng Stormfr*** với những người hay lui tới trang Yahoo News, hóa ra đám đông trên Stormfr*** có gấp đôi khả năng sẽ truy cập trang báo nytimes.com.

Các thành viên một trang web thù địch lại đi nghiên cứu kĩ trang nytimes.com theo chủ nghĩa tự do hết cỡ à? Làm sao mà vậy được? Nếu một lượng lớn thành viên Stormfr*** lấy tin tức từ nytimes.com, điều đó có nghĩa rằng hiểu biết truyền thống của chúng ta về những người theo chủ nghĩa dân tộc da trắng là sai. Cũng có nghĩa rằng hiểu biết truyền thống của chúng ta về cách Internet hoạt động cũng sai nốt.

## Sự thật về Internet

Hầu hết mọi người đều đồng ý rằng Internet chia rẽ người Mĩ, khiến hầu hết mọi người chui rúc trong các trang web dành cho những người giống mình. Đây là cách Cass Sunstein của Trường Luật Harvard mô tả tình hình: "Thị trường truyền thông của chúng ta đang di chuyển nhanh [đến một tình trạng mà] người ta giới hạn bản thân trong quan điểm của chính mình—phe tự do thì chỉ xem và đọc nội dung tự do; ôn hòa thì tìm

ôn hòa; bảo thủ thì tìm bảo thủ; phát xít mới thì tìm phát xít mới."

Góc nhìn này cũng dễ hiểu. Xét cho cùng, Internet cho ta hầu như vô hạn lựa chọn để tiêu thụ tin tức. Tôi có thể đọc bất cứ cái gì tôi muốn. Bạn có thể đọc bất cứ cái gì bạn muốn. Vikingmaiden88 có thể đọc bất cứ cái gì cô ta muốn. Và mọi người, nếu ngồi lại với các thiết bị của chính họ, có xu hướng tìm kiếm các quan điểm xác thực những gì họ tin. Như thế, dĩ nhiên, Internet hẳn là đang tạo ra sự phân cực chính trị.

Có một vấn đề với góc nhìn chuẩn mực này: Dữ liệu cho ta biết rằng nó hoàn toàn sai.

Bằng chứng chống lại cái hiểu biết truyền thống này đến từ một nghiên cứu năm 2011 của Matt Gentzkow và Jesse Shapiro, hai nhà kinh tế học mà chúng ta đã gặp ở phần trước.

Gentzkow và Shapiro thu thập dữ liệu về hành vi lướt tìm thông tin của một mẫu đại diện lớn cho người Mĩ. Bộ dữ liệu của họ bao gồm thông tin về hệ tư tưởng—được người tham gia tự khai: Ta biết rằng mỗi người tham gia tự xem mình là tự do hay bảo thủ. Họ dùng dữ liệu này để đo lường sự phân li về chính trị trên Internet.

Bằng cách nào? Họ đã thực hiện một thí nghiệm tâm tưởng rất thú vị. Chọn mẫu đại diện ngẫu nhiên 2 người Mĩ mà tình cờ cả 2 đang vào cùng một website báo chí. Xác suất 1 trong 2 người thuộc phe tự do và người kia phe bảo thủ là bao nhiêu? Nói cách khác, tần số những người tự do và bảo thủ "gặp nhau" trên các trang báo chí là bao nhiêu?

Thử suy nghĩ về điều này sâu hơn. Giả sử phe tự do và phe bảo thủ trên Internet không bao giờ đọc báo trực tuyến cùng một chỗ. Nói cách khác, người tự do thì chỉ lui tới các website tự do, người bảo thủ thì chỉ vào các website bảo thủ. Nếu điều này mà đúng, khả năng 2 người Mĩ trên một trang tin nào đó có quan điểm chính trị đối lập sẽ là 0%. Internet sẽ hoàn toàn *phân li*. Những người tự do và những người bảo thủ sẽ không bao giờ hòa trộn với nhau.

Trái lại, giả sử người tự do và người bảo thủ không khác nhau chút nào trong cách thu thập tin tức. Nói cách khác, khả năng một người tự

do hoặc một người bảo thủ vào một trang tin nào đó là ngang nhau. Nếu điều này đúng, khả năng 2 người Mĩ trên một website báo chí có quan điểm chính trị đối lập sẽ là khoảng 50%. Internet sẽ hoàn toàn *không có sự phân li*. Phe tự do và phe bảo thủ sẽ hoàn toàn hòa trộn với nhau.

Vậy dữ liệu cho ta biết điều gì? Tại Mĩ, theo Gentzkow và Shapiro, khả năng 2 người lui tới cùng một trang báo chí có quan điểm chính trị khác nhau là khoảng 45%. Nói cách khác, Internet gần với mốc hoàn toàn không phân li hơn rất nhiều so với mốc hoàn toàn phân li. Người tự do và người bảo thủ "gặp" nhau trên mạng suốt.

Để hiểu rõ sự không phân li trên Internet, ta phải so sánh nó với mức độ phân li ở những mặt khác trong đời sống. Gentzkow và Shapiro có thể lặp lại phân tích của họ với nhiều tương tác ngoại tuyến khác. Khả năng 2 thành viên gia đình có quan điểm chính trị khác nhau là bao nhiêu? 2 người hàng xóm? 2 đồng nghiệp? 2 người bạn?

Khi dùng dữ liệu từ General Social Survey, Gentzkow và Shapiro phát hiện là tất cả các con số này thấp hơn khả năng 2 người trên cùng một website báo chí có quan điểm chính trị khác nhau.

**XÁC SUẤT NGƯỜI BẠN GẶP CÓ QUAN ĐIỂM CHÍNH TRỊ ĐỐI LẬP**

| | |
|---|---|
| Trên một website báo chí | 45.2% |
| Đồng nghiệp | 41.6% |
| Hàng xóm ngoại tuyến | 40.3% |
| Thành viên gia đình | 37.0% |
| Bạn bè | 34.7% |

Nói cách khác, bạn có nhiều khả năng sẽ gặp một người có quan điểm đối lập trên mạng hơn là ngoài đời.

Tại sao Internet không bị phân li nhiều như ta vẫn tưởng? Có 2 yếu tố hạn chế sự phân li chính trị trên Internet.

Thứ nhất, cũng hơi bất ngờ, ngành báo chí Internet bị thống trị bởi một vài trang khổng lồ. Ta thường nghĩ Internet là nơi thu hút các thành phần cực đoan. Thực ra, có những trang dành cho mọi người, bất kể

quan điểm. Có "chỗ đáp" cho người ủng hộ và chống quyền sở hữu súng, cho nhà hoạt động về các quyền hút xì gà và dùng đồng đô la kim loại, người theo chủ nghĩa vô chính phủ và người theo chủ nghĩa dân tộc da trắng. Nhưng tất cả các trang này chiếm một phần rất nhỏ lượng thông tin của Internet. Thực vậy, năm 2009, 4 trang—Yahoo News, AOL News, msnbc.com, và cnn.com—thu thập hơn một nửa lượt xem tin tức trên toàn Internet. Yahoo News vẫn là trang báo phổ biến nhất của người Mĩ, với gần 90 triệu khách độc lập (unique visitor) mỗi tháng—gấp chừng 600 lần lượng khán giả của Stormfr***. Các trang truyền thông đại chúng như Yahoo News thu hút một lượng khán giả rộng lớn, đa dạng về chính trị.

Lí do thứ hai khiến Internet không bị phân li là vì nhiều người có chính kiến mạnh mẽ thường vào các trang có quan điểm đối lập chỉ để nổi giận và tranh cãi. Những kẻ nghiện chính trị không giới hạn bản thân chỉ vào các trang dành cho mình. Người hay vào thinkprogress.org và moveon.org—hai trang tự do cực đoan—có nhiều khả năng vào foxnews.com (một trang bảo thủ) hơn người dùng Internet bình thường. Người hay vào rushlimbaugh.com hoặc glennbeck.com—hai trang bảo thủ cực đoan—có nhiều khả năng vào nytimes.com (một trang tự do) hơn người dùng Internet bình thường.

Nghiên cứu của Gentzkow và Shapiro dựa trên dữ liệu giai đoạn 2004–2009, tương đối sớm trong lịch sử Internet. Có thể Internet đã phân li hơn kể từ lúc đó chăng? Các phương tiện truyền thông xã hội và đặc biệt là Facebook có làm thay đổi kết luận của họ? Rõ ràng, nếu bạn bè ta mà có chung quan điểm chính trị với ta, sự trỗi dậy của mạng xã hội sẽ đồng nghĩa với sự trỗi dậy của các quan điểm một chiều. Đúng không?

Một lần nữa, chuyện không hề đơn giản như vậy. Mặc dù đúng là bạn bè trên Facebook có nhiều khả năng có chung quan điểm chính trị, nhưng một nhóm nhà khoa học dữ liệu—Eytan Bakshy, Solomon Messing, và Lada Adamic—đã phát hiện rằng lượng lớn thông tin mà người ta lấy trên Facebook lại đến từ những người có quan điểm đối lập.

Sao có thể như vậy được? Bạn bè ta mà thường không có chung quan

điểm chính trị với ta sao? Thực ra là có. Nhưng có một lí do quan trọng là Facebook có thể dẫn đến một cuộc tranh luận chính trị đa dạng hơn những dịp xã giao ngoại tuyến. Người ta thường có bạn bè trên Facebook nhiều hơn ngoài đời. Và các mối quan hệ yếu (weak tie) được Facebook tạo điều kiện cho gặp nhau này rất có thể là những người có quan điểm chính trị đối lập.

Nói cách khác, Facebook cho ta tiếp xúc nhiều với các liên kết xã hội yếu—người quen thời trung học, người anh họ thần kinh xa ba tầm đại bác, bạn của bạn của bạn mà có thể là ta kiểu-như-đại-loại-như có quen biết. Đây là những người có thể ta chưa bao giờ đi vui chơi ăn uống cùng nhau. Có thể ta không bao giờ mời họ đến ăn tối. Nhưng ta lại kết bạn Facebook với họ. Và ta nhìn thấy các đường dẫn của họ đến các bài báo có những quan điểm mà lẽ ra có thể ta sẽ chẳng bao giờ xem đến.

Tóm lại, Internet thực sự mang những người khác quan điểm chính trị lại với nhau. Một người phe tự do bình thường có thể trải qua buổi sáng với ông chồng phe tự do và các con phe tự do; buổi chiều thì gặp đồng nghiệp phe tự do; trên đường đi làm thì thấy toàn thông điệp tự do; buổi chiều thì tập cùng bạn học phe tự do tại lớp yoga. Khi về nhà và đọc kĩ vài bình luận phe bảo thủ trên cnn.com hoặc nhấp vào một đường dẫn Facebook từ người quen thời trung học theo đảng Cộng hòa, đây có thể là thời điểm bà tiếp xúc với phe bảo thủ nhiều nhất trong ngày.

Có thể tôi chưa bao giờ gặp người theo chủ nghĩa dân tộc da trắng tại quán cà phê ruột của tôi ở Brooklyn. Nhưng cả Vikingmaiden88 và tôi đều thường xuyên vào *New York Times*.

## Sự thật về ngược đãi trẻ em và nạo phá thai

Internet có thể cho ta những hiểu biết không chỉ về thái độ quấy rối mà còn hành vi quấy rối. Thực vậy, dữ liệu Google có thể hiệu quả trong việc cảnh báo về các khủng hoảng mà tất cả các nguồn thông thường không thể thấy. Rốt cuộc, người ta đều vào Google khi gặp rắc rối.

Hãy xem xét việc ngược đãi trẻ em trong thời kì Đại suy thoái.

Khi suy thoái kinh tế bắt đầu cuối năm 2007, tất nhiên nhiều chuyên

gia lo lắng về ảnh hưởng của nó đối với trẻ em. Suy cho cùng, nhiều bậc cha mẹ đã bị căng thẳng và suy sụp, và đây là các yếu tố nguy cơ lớn dẫn đến sự ngược đãi. Sự ngược đãi trẻ em có thể tăng vọt.

Sau đó, dữ liệu chính thức xuất hiện, và dường như sự lo lắng ấy là vô căn cứ. Các cơ quan bảo vệ trẻ em thông báo rằng họ ít gặp các trường hợp ngược đãi hơn. Hơn nữa, sự sụt giảm lại lớn nhất ở các bang bị ảnh hưởng cuộc khủng hoảng nặng nề nhất. "Các dự báo âm u tăm tối chưa thành sự thật," Richard Gelles, một chuyên gia phúc lợi trẻ em tại Đại học Pennsylvania, nói với *Associated Press* năm 2011. Vâng, dù có vẻ phản trực giác, tình trạng ngược đãi trẻ em dường như đã giảm mạnh trong thời kì suy thoái.

Nhưng có đúng là tình trạng ngược đãi trẻ em giảm mạnh khi rất nhiều người lớn thất nghiệp và cực kì khốn khổ không? Tôi không thể tin điều này được. Vì vậy tôi nghiên cứu dữ liệu Google.

Hóa ra, một số trẻ thực hiện nhiều tìm kiếm bi thảm và nhói tim trên Google—chẳng hạn như "mẹ đánh tôi" hoặc "cha đánh tôi." Các tìm kiếm này cho thấy một bức tranh khác hẳn—và thật nhói lòng—về những gì thực sự diễn ra trong suốt thời kì này. Số lượng tìm kiếm kiểu này tăng vọt trong thời kì Đại suy thoái, rất sát với tỉ lệ thất nghiệp.

Đây là điều tôi nghĩ đã xảy ra: Số vụ báo cáo các trường hợp ngược đãi trẻ em giảm, chứ không phải bản thân tình trạng ngược đãi trẻ em giảm. Xét cho cùng, người ta ước tính rằng chỉ một tỉ lệ nhỏ các trường hợp ngược đãi trẻ em được báo cáo với chính quyền. Và trong một cuộc suy thoái, phần nhiều những người hay báo cáo các trường hợp ngược đãi trẻ em (giáo viên và cảnh sát, chẳng hạn) và quản lí các trường hợp ấy (nhân viên bảo vệ trẻ em) rất có thể đang quá tải hoặc đã mất việc.

Có nhiều câu chuyện trong thời kì suy thoái kinh tế về những người đã cố gắng báo cáo các trường hợp tiềm ẩn, nhưng rồi vì phải chờ đợi quá lâu nên đã bỏ cuộc.

Thực vậy, có nhiều bằng chứng hơn, lần này không phải từ Google, cho thấy tình trạng ngược đãi trẻ em thực sự tăng lên trong thời kì suy thoái. Khi một trẻ em chết do bị ngược đãi hoặc bỏ bê thì phải được báo

cáo. Những cái chết đó, dù hiếm, cũng đã tăng ở các bang bị tác động nặng nhất bởi cuộc suy thoái.

Và một số bằng chứng từ Google cho thấy nhiều người hơn đang nghi ngờ có sự ngược đãi ở những vùng bị tác động mạnh. Khi đối chiếu với các tỉ lệ thời tiền suy thoái và các xu hướng quốc gia, các bang chịu ảnh hưởng nặng nề nhất đã gia tăng tỉ lệ tìm kiếm về sự ngược đãi và bỏ bê trẻ em. Cứ mỗi %p gia tăng trong tỉ lệ thất nghiệp đồng nghĩa với việc tỉ lệ tìm kiếm "ngược đãi trẻ em" hoặc "bỏ bê trẻ em" tăng 3%. Có lẽ, hầu hết những người này đã chẳng bao giờ báo cáo thành công tình trạng ngược đãi, vì các bang này lại có lượng báo cáo sụt giảm mạnh nhất.

Tìm kiếm của các trẻ trong cuộc tăng. Tỉ lệ tử vong trẻ em tăng mạnh. Tìm kiếm của những người nghi ngờ có sự ngược đãi tăng ở các bang bị tác động mạnh. Nhưng số báo cáo các trường hợp ngược đãi lại giảm. Một cuộc suy thoái dường như khiến thêm nhiều trẻ em nói với Google rằng cha mẹ đang đánh đập chúng, và cũng khiến có thêm nhiều người nghĩ rằng họ thấy có sự ngược đãi. Nhưng các cơ quan bị quá tải lại nắm được ít trường hợp hơn.

Tôi nghĩ là mình hoàn toàn có thể nói rằng cuộc Đại suy thoái đã khiến tình trạng ngược đãi trẻ em thêm tồi tệ, mặc dù các thước đo truyền thống không chỉ ra điều đó.

***

Hễ khi nào nghi ngờ đang có những nạn nhân nằm ngoài sổ sách, tôi lại quay sang dữ liệu Google. Một trong các lợi ích tiềm năng của dữ liệu mới, và của việc biết cách diễn giải dữ liệu, là khả năng giúp đỡ những người dễ bị tổn thương—những người có thể không được người có thẩm quyền để ý tới.

Vì vậy, khi gần đây Tòa án Tối cao xem xét tác động của các luật khiến việc nạo phá thai khó khăn hơn, tôi lại quay sang dữ liệu tìm kiếm. Tôi nghi ngờ phụ nữ bị luật này ảnh hưởng có thể tìm đến những cách phá thai chui. Họ đã làm vậy. Và các tìm kiếm này cao nhất ở các bang đã thông qua luật hạn chế nạo phá thai.

Dữ liệu tìm kiếm này vừa hữu ích vừa đáng ngại.

Năm 2015, ở Mĩ, có hơn 700,000 tìm kiếm Google tìm hiểu về tự nạo phá thai. Để so sánh, có khoảng 3.4 triệu tìm kiếm về cơ sở y tế có nạo phá thai trong năm đó. Điều này chứng tỏ rằng một tỉ lệ không nhỏ phụ nữ nghĩ về nạo phá thai đã dự tính tự làm việc đó.

Nữ giới tìm kiếm khoảng 160,000 lần các cách dùng thuốc phá thai qua các kênh không chính thức—"mua thuốc phá thai trực tuyến" và "thuốc phá thai miễn phí." Họ hỏi Google về việc phá thai bằng cây thuốc như ngò tây hoặc bằng vitamin C. Có khoảng 4,000 lượt tìm kiếm hướng dẫn về nạo phá thai bằng móc áo, bao gồm khoảng 1,300 lượt tìm kiếm cụm từ chính xác "how to do a coat hanger abortion" (cách phá thai bằng móc áo). Cũng có vài trăm lượt tìm kiếm cách nạo phá thai thông qua việc dùng thuốc tẩy tử cung và đấm vào bụng người mang thai.

Điều gì khiến người ta quan tâm tới việc tự nạo phá thai? Thông tin địa lí và thời gian của các tìm kiếm Google chỉ ra một thủ phạm tiềm năng: Khi khó phá thai hợp pháp, phụ nữ tìm vào các phương pháp nằm ngoài giấy tờ sổ sách.

Tỉ lệ tìm kiếm tự nạo phá thai khá ổn định từ 2004 đến hết 2007. Nó bắt đầu tăng vào cuối năm 2008, trùng với cuộc khủng hoảng tài chính và thời kì suy thoái theo sau đó. Nó tăng mạnh năm 2011, vọt lên 40%. Viện Guttmacher, một tổ chức về các quyền sinh sản, chọn năm 2011 là thời kì khởi đầu cuộc trừng phạt nạo phá thai gần đây của cả nước; thời điểm này, 92 điều khoản cấp tiểu bang giới hạn nạo phá thai đã được ban hành. Khi so sánh với Canada, quốc gia chưa phải chứng kiến một cuộc trừng phạt về các quyền sinh sản, không có sự gia tăng đáng kể các lượt tìm kiếm về tự phá thai trong thời gian này.

Bang có tỉ lệ tìm kiếm Google tự nạo phá thai cao nhất là Mississippi, một bang có khoảng 3 triệu người và chỉ có 1 cơ sở nạo phá thai (tính đến thời điểm viết dòng này). 8 trong 10 bang có tỉ lệ tìm kiếm về tự nạo phá thai cao nhất được Viện Guttmacher xem là thù địch hoặc rất thù địch với nạo phá thai. Không bang nào trong 10 bang có tỉ lệ tìm kiếm về tự nạo phá thai thấp nhất bị xếp vào nhóm thù địch hoặc rất thù địch.

Dĩ nhiên, chúng ta không thể biết từ các tìm kiếm Google có bao nhiêu phụ nữ đã tự phá thai thành công, nhưng bằng chứng cho thấy đó có thể là một con số không hề nhỏ. Một cách để làm sáng tỏ điều này là so sánh dữ liệu nạo phá thai và dữ liệu sinh con.

Năm 2011, năm cuối cùng có dữ liệu nạo phá thai cấp tiểu bang đầy đủ, phụ nữ sống ở các tiểu bang có ít cơ sở nạo phá thai có ít lượt phá thai hợp pháp hơn rất nhiều.

Hãy so sánh 10 bang có số cơ sở nạo phá thai tính trên đầu người cao nhất (có New York và California) với 10 bang có số cơ sở nạo phá thai tính trên đầu người thấp nhất (có Mississippi và Oklahoma). Phụ nữ sống ở các bang có ít cơ sở nạo phá thai nhất có số lượt phá thai hợp pháp ít hơn 54%—tức là ít hơn 11 lượt phá thai tính trên 1,000 phụ nữ thuộc nhóm tuổi 15-44. Phụ nữ sống ở các bang có ít cơ sở nạo phá thai nhất cũng có nhiều lượt sinh hơn. Tuy nhiên, số chênh lệch không đủ để bù cho con số nạo phá thai thấp hơn. Chỉ có thêm 6 lượt sinh bình thường cho mỗi 1,000 phụ nữ ở độ tuổi sinh con.

Nói cách khác, có vẻ đã có một số lượt mang thai bị mất ở các vùng khó nạo phá thai nhất. Các nguồn chính thức không cho ta biết điều gì đã xảy ra với 5 lượt sinh bị bỏ qua đó cho mỗi 1,000 phụ nữ ở các bang khó nạo phá thai kia.

Google cung cấp một số đầu mối khá tốt.

Chúng ta không thể mù quáng tin tưởng dữ liệu chính phủ Mĩ. Chính phủ có thể nói cho ta biết rằng sự ngược đãi trẻ em hoặc nạo phá thai đã giảm mạnh và các chính khách sẽ ca tụng thành tựu này. Nhưng kết quả ta đang thấy có thể là sản phẩm lỗi của các phương pháp thu thập dữ liệu. Sự thật có thể khác—và đôi khi cũng đen tối hơn nhiều.

## Sự thật về bạn bè trên Facebook

Quyển sách này nói về Dữ Liệu Lớn nói chung. Nhưng chương này chủ yếu nhấn mạnh các tìm kiếm Google, thứ mà tôi cho là tiết lộ một thế giới ẩn rất khác với thế giới ta nghĩ là ta thấy. Vậy các nguồn Dữ Liệu

Lớn khác có phải là huyết thanh sự thật số? Sự thật là, nhiều nguồn Dữ Liệu Lớn, ví dụ như Facebook, thường là đối nghịch của huyết thanh sự thật số.

Tương tự như trong các khảo sát, ta không có động cơ nói thật trên mạng xã hội. Trên mạng xã hội, ta còn có nhiều động cơ để làm bản thân trông có vẻ ngon lành hơn nữa kìa. Xét cho cùng, ở đó, ta không nặc danh. Ta đang quyến rũ một nhóm khán giả, và đang nói cho bạn bè, các thành viên gia đình, đồng nghiệp, người quen, và người lạ biết ta là ai.

Để xem dữ liệu rút ra từ mạng xã hội có thể thiên lệch đến thế nào, hãy xét tính phổ biến tương đối của *Atlantic,* một nguyệt san hàn lâm, so với *National Enquirer,* một tạp chí phiếm đàm, giải trí. Cả hai xuất bản phẩm có lượng phát hành mỗi số báo trung bình tương tự, bán vài trăm ngàn bản. (*National Enquirer* là tuần san, vì vậy thực tế tờ này bán tổng số bản nhiều hơn.) Hai tạp chí này cũng có số lượt tìm kiếm Google tương đương nhau.

Tuy nhiên, trên Facebook, khoảng 1.5 triệu người thích *Atlantic* hoặc thảo luận các bài báo của *Atlantic* trên trang cá nhân. Chỉ khoảng 50 ngàn người thích hoặc thảo luận nội dung của *Enquirer.*

**TÍNH PHỔ BIẾN SO SÁNH THEO CÁC NGUỒN KHÁC NHAU GIỮA *ATLANTIC* VÀ *NATIONAL ENQUIRER***

| | |
|---|---|
| Phát hành | 1 Atlantic : 1 National Enquirer |
| Tìm kiếm Google | 1 Atlantic : 1 National Enquirer |
| Like trên Facebook | 27 Atlantic : 1 National Enquirer |

Để đánh giá tính phổ biến của tạp chí, dữ liệu phát hành là sự thật nền tảng. Dữ liệu Google gần khớp với nó. Và dữ liệu Facebook thì có xu hướng chống báo lá cải, khiến cho nó trở thành dữ liệu kém nhất để biết người ta thực sự thích cái gì.

Và giống như sở thích đọc báo, cuộc sống cũng vậy. Trên Facebook, ta thể hiện cái tôi văn hóa, chứ không phải cái tôi chân thực. Tôi có dùng

dữ liệu Facebook trong sách này, cụ thể là trong chương này, nhưng luôn cảnh giác tính chất ấy.

Để hiểu biết rõ hơn về những gì mạng xã hội bỏ qua, chúng ta hãy trở lại phim ảnh khiêu dâm một lát. Trước hết, ta cần xử lí niềm tin thông thường rằng Internet bị các thứ dâm ô thống trị. Điều này không đúng. Đa số nội dung trên Internet là phi khiêu dâm. Ví dụ, trong top 10 website được lui tới nhiều nhất, không trang nào là khiêu dâm cả. Vậy tính phổ biến của khiêu dâm, dù lớn, cũng đừng nên nói quá.

Nói là vậy, nhưng khi nhìn vào hoạt động like và share hình ảnh khiêu dâm thì rõ ràng là Facebook, Instagram, và Twitter chỉ cung cấp một cái nhìn hạn chế vào những gì thực sự phổ biến trên Internet. Có những bộ phận lớn trên web rất phổ biến nhưng lại ít xuất hiện trên phương diện xã hội.

Video nổi tiếng nhất mọi thời đại, tính đến lúc viết đoạn này, là "Gangnam Style" của Psy, một video nhạc pop theo phong cách ngốc nghếch châm biếm những người Triều Tiên chạy theo mốt. Video này đã được xem khoảng 2.3 tỉ lượt trên YouTube kể từ khi ra đời năm 2012. Tính phổ biến của nó là rõ ràng, dù bạn có đang vào trang nào. Nó được chia sẻ hàng chục triệu lượt khắp các hệ thống mạng xã hội khác nhau.

Video khiêu dâm phổ biến nhất mọi thời đại có thể là "Great Body, Great ***, Great *******." Nó đã được xem hơn 80 triệu lượt. Nói cách khác, cứ 30 lượt xem "Gangnam Style," thì có khoảng ít nhất 1 lượt xem "Great Body, Great ***, Great *******." Nếu mạng xã hội cho ta một cái nhìn chính xác về các video mà người ta xem, "Great Body, Great ***, Great *******" phải được đăng hàng triệu lần. Nhưng video này được chia sẻ trên mạng xã hội chỉ vài chục lần và đều được đăng bởi các sao khiêu dâm, chứ không phải người dùng bình thường. Rõ ràng người ta không có nhu cầu quảng cáo sự quan tâm của họ với video này.

Facebook là nơi để nổ-với-bạn-bè-về-cuộc-sống-xịn-ngầu-của-tôi. Ở thế giới Facebook, người trưởng thành bình thường có vẻ như đang có cuộc sống hôn nhân hạnh phúc, hay đi nghỉ ở vùng Caribbean, và đọc báo khoa học *Atlantic*. Trong thế giới thực, nhiều người hay cáu giận,

đang phải xếp hàng tính tiền ở siêu thị, toàn đọc báo lá cải *National Enquirer*, phớt lờ các cuộc gọi của bạn đời—và hai người họ đã giường ai nấy ngủ nhiều năm rồi. Trong thế giới Facebook, cuộc sống gia đình có vẻ hoàn hảo. Trong thế giới thực, cuộc sống gia đình rối tung. Thỉnh thoảng có thể rối đến độ một số ít người còn hối tiếc vì đã có con. Trong thế giới Facebook, dường như người trưởng thành trẻ tuổi nào cũng đều đang ăn chơi nhảy múa, tiệc tùng hoành tráng tối thứ Bảy. Trong thế giới thực, hầu hết nằm chèo queo ở nhà, cày phim trên Netflix. Trong thế giới Facebook, một bạn gái đăng lên 26 hình ảnh hạnh phúc khi đang cùng bạn trai "đưa nhau đi trốn." Trong thế giới thực, ngay sau khi đăng hình, cô Google cụm từ "bạn trai không chịu quan hệ với tôi." Và, có lẽ cùng thời điểm đó, người bạn trai đang xem "Great Body, Great ***, Great *******."

| SỰ THẬT SỐ | DỐI TRÁ SỐ |
|---|---|
| - Tìm kiếm | - Post mạng xã hội |
| - Lượt xem | - Like mạng xã hội |
| - Nhấp chuột | - Profile hẹn hò |
| - Swipe | |

## Sự thật về khách hàng

Sáng sớm ngày 5/9/2006, Facebook giới thiệu một cập nhật lớn cho homepage. Các phiên bản đầu tiên của Facebook chỉ cho phép người dùng nhấp chuột lên profile bạn bè để biết họ đang làm gì. Website này, được xem là một thành công lớn, lúc đó có 9.4 triệu người dùng.

Nhưng sau nhiều tháng làm việc tích cực, các kĩ sư đã tạo ra cái gọi là *News Feed*, nó cung cấp cho người dùng các cập nhật về những hoạt động của tất cả bạn bè họ.

Người dùng ngay lập tức báo rằng họ ghét News Feed. Ben Parr, một sinh viên Đại học Northwestern, đã tạo nhóm "Students Against Facebook news feed" (Sinh viên chống news feed của Facebook). Anh nói "news feed quá ớn, quá thọc mạch, và là một tính năng phải bị loại bỏ." Trong vòng vài ngày, nhóm đã có 700,000 thành viên. Một sinh viên

năm 3 tại Đại học Michigan nói với *Michigan Daily*, "Tôi thực sự rất ngán Facebook mới. Nó làm tôi cảm thấy giống một kẻ thọc mạch."

David Kirkpatrick kể câu chuyện này trong quyển *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*—quyển sách được Facebook công nhận viết về lịch sử website này. Ông gọi lần tung ra News Feed là "khủng hoảng lớn nhất mà Facebook từng đối mặt." Nhưng Kirkpatrick tường thuật rằng khi ông phỏng vấn Mark Zuckerberg, đồng sáng lập và là người đứng đầu công ty đang phát triển nhanh chóng đó, vị CEO này không hề bối rối.

Lí do là gì? Zuckerberg đã tiếp cận được với huyết thanh sự thật số: các con số về cú nhấp chuột và đăng nhập Facebook của mọi người. Kirkpatrick viết:

> Thực ra Zuckerberg biết rằng người ta thích News Feed, dù họ có nói gì trong các nhóm đi nữa. Anh có dữ liệu để chứng minh điều đó. Tính trung bình, người dùng đã bỏ nhiều thời gian hơn trên Facebook so với trước khi có News Feed. Và họ cũng làm nhiều thứ ở đó hơn nữa—nhiều hơn đột biến. Tháng 8, người dùng xem 12 tỉ trang trên Facebook. Nhưng đến tháng 10, khi có thêm News Feed, họ xem 22 tỉ trang.

Và đó không phải là tất cả bằng chứng mà Zuckerberg có thể tùy ý sử dụng. Ngay cả sự lan truyền nhanh chóng của nhóm chống News Feed cũng chính là bằng chứng sức mạnh của News Feed. Nhóm phát triển rất nhanh chính xác vì quá nhiều người nghe là bạn bè họ đã tham gia—và họ biết điều này qua News Feed.

Nói cách khác, mặc dù người ta đang tham gia phản đối om sòm rằng họ rất không vui về việc nhìn thấy toàn bộ chi tiết cuộc sống của bạn bè trên Facebook, họ lại đang vào Facebook để xem tất cả chi tiết cuộc sống của bạn bè. News Feed vẫn tồn tại. Facebook bây giờ có hơn 1 tỉ người dùng tích cực hàng ngày.

Trong quyển sách *Zero to One*, Peter Thiel, nhà đầu tư giai đoạn đầu của Facebook, nói rằng các doanh nghiệp lớn được xây trên những bí mật, bí mật về tự nhiên hoặc bí mật về con người. Jeff Seder, người đã

xuất hiện tại Chương 3, đã tìm thấy bí mật tự nhiên rằng kích thước tâm thất trái dự báo thành tích của ngựa. Google đã tìm thấy bí mật tự nhiên của sức mạnh thông tin trong các đường dẫn.

Thiel định nghĩa "bí mật về con người" là "thứ mà mọi người không biết về bản thân hoặc thứ mà họ che giấu vì họ không muốn người khác biết." Các loại doanh nghiệp này, nói cách khác, được xây dựng trên lời nói dối của mọi người.

Bạn có thể nói rằng toàn bộ Facebook được thành lập trên một bí mật không vui về mọi người mà Zuckerberg biết khi còn học ở Harvard. Zuckerberg, đầu năm thứ hai đại học, đã tạo ra một website cho sinh viên gọi là Facemash. Được làm dựa trên một trang gọi là "Am I Hot or Not?" (Tôi có hấp dẫn không?), Facemash đăng hình 2 sinh viên Harvard và sau đó để các sinh viên khác làm trọng tài xem ai đẹp hơn.

Trang web của anh sinh viên năm 2 này được chào đón bởi sự giận dữ. Báo *Harvard Crimson,* trong một bài xã luận, đã buộc tội chàng trai Zuckerberg "phục vụ mặt xấu nhất" của mọi người. Các nhóm người Mĩ gốc Latin và gốc Phi buộc tội anh phân biệt giới tính và chủng tộc. Tuy nhiên, trước khi các nhà quản trị Harvard cắt Internet của Zuckerberg—chỉ vài giờ sau khi trang web được thành lập—450 người đã xem và bầu chọn 22,000 lượt trên các hình ảnh khác nhau. Zuckerberg đã biết một bí mật quan trọng: Người ta có thể nói là họ tức giận, họ có thể chê bai cái gì đó là phản cảm, thế nhưng, họ vẫn sẽ nhấp chuột.

Và anh còn biết thêm một điều: Dù cho nghề nghiệp có cần nghiêm túc, trách nhiệm, và tôn trọng sự riêng tư của người khác đi nữa, người ta, ngay cả sinh viên Harvard, vẫn đều rất thích đánh giá dáng vẻ của mọi người. Các lượt xem và bầu chọn cho anh biết điều đó. Sau đó—vì Facemash đã tạo ra quá nhiều tranh cãi—anh biết chính xác người ta có thể rất quan tâm đến các sự việc hời hợt về người khác—những người mà họ cũng kiểu gọi là có biết biết—và rồi khai thác nó để lập ra công ty thành công nhất của thế hệ mình.

Netflix cũng đã học được bài học tương tự từ rất sớm: Đừng tin những gì người ta nói; hãy tin những gì người ta làm.

Đầu tiên, công ty cho phép người dùng tạo ra một danh sách phim họ muốn xem nhưng hiện tại chưa có thời gian. Theo cách này, khi họ có nhiều thời gian hơn, Netflix có thể nhắc họ về các phim đó.

Tuy nhiên, Netflix chú ý thấy có gì đó kì lạ trong dữ liệu. Người dùng liệt kê rất nhiều phim vào danh sách của mình. Nhưng những ngày sau đó, khi được nhắc lại về các phim trong danh sách, họ hiếm khi nhấp chuột.

Vấn đề là gì? Nếu hỏi người dùng họ định xem những phim gì trong vài ngày tới, họ sẽ liệt kê đầy danh sách các phim danh giá, trí thức, chẳng hạn như các phim tư liệu trắng đen về Thế chiến II hoặc phim nước ngoài nội dung nghiêm túc. Tuy nhiên, vài ngày sau, họ lại muốn xem những phim giống như họ thường muốn xem: các phim hài đơn giản hoặc phim tình cảm lãng mạn. Người ta cứ thường xuyên nói dối với chính mình.

Đối mặt với sự sai lệch này, Netflix thôi yêu cầu người ta liệt kê những thứ họ muốn xem trong tương lai và bắt đầu xây dựng một mô hình dựa trên hàng triệu cú nhấp chuột và lượt xem từ các khách hàng tương tự. Công ty bắt đầu chào đón người dùng với các danh sách phim không dựa trên những gì họ nói thích mà dựa trên những gì dữ liệu nói họ có nhiều khả năng sẽ xem. Kết quả: Khách hàng ghé Netflix thường xuyên hơn và xem nhiều phim hơn.

"Các thuật toán biết bạn rõ hơn bạn biết bản thân mình," theo Xavier Amatriain, cựu khoa học gia máy tính tại Netflix.

## Liệu ta có thể đối mặt với sự thật?

Có thể bạn thấy các phần của chương này thật tiêu cực. Huyết thanh sự thật số đã tiết lộ sở thích phán xét người khác dựa trên dáng vẻ bề ngoài; sự tồn tại của hàng triệu nam đồng tính không công khai; một tỉ lệ đáng kể phụ nữ mơ tưởng về chuyện hiếp dâm; lòng thù hận lan tràn chống người Mĩ gốc Phi; cuộc khủng hoảng về ngược đãi trẻ em và tự nạo phá thai bị che giấu; và sự bùng nổ lòng thù hận đầy bạo lực với người Hồi giáo, thứ ngày càng tồi tệ hơn khi tổng thống kêu gọi lòng

khoan dung. Chẳng phải điều gì vui vẻ. Thông thường, sau khi tôi trình bày nghiên cứu của mình, người ta hay đến gặp tôi và nói, "Seth, tất cả đều rất hay. Nhưng *quá* tiêu cực."

## GIÁ TRỊ LỚN CỦA VIỆC BỎ QUA ĐIỀU NGƯỜI TA NÓI VỚI BẠN

| ĐIỀU NGƯỜI TA NÓI | THỰC TẾ | NHỜ ĐÓ MÀ... |
|---|---|---|
| Họ không muốn lén theo dõi bạn bè. | Ít có thứ gì trên thế gian này họ muốn hơn là theo dõi và phán xét bạn bè. | Mark Zuckerberg, đồng sáng lập Facebook, sở hữu 55.2 tỉ USD. |
| Họ không muốn mua sản phẩm được sản xuất tại các xí nghiệp bóc lột. | Họ sẽ mua sản phẩm tốt, "giá cả phải chăng." | Phil Knight, đồng sáng lập Nike, sở hữu 25.4 tỉ USD. |
| Họ muốn nghe tin tức vào buổi sáng. | Họ muốn nghe về các người lùn quan hệ tình dục với các sao khiêu dâm vào buổi sáng. | Howard Stern sở hữu 500 triệu USD.[1] |
| Họ không thích đọc về các trò trói tay trói chân và bạo dâm. | Họ muốn đọc về các trò trói tay trói chân và bạo dâm giữa một sinh viên trẻ vừa ra trường và một đại gia. | *50 Shades of Gray* đã bán được 125 triệu bản. |
| Họ muốn các chính trị gia đưa ra các lập trường chính sách. | Họ muốn các chính trị gia không đi sâu vào chi tiết nhưng có vẻ cứng rắn và tự tin. | Donald Trump |

Tôi chẳng thể giả vờ không có khoảng tối trong số dữ liệu này. Nếu mọi người thường xuyên nói với ta những gì họ nghĩ là ta muốn nghe, nói chung ta sẽ được nghe những thứ mang tính an ủi hơn là sự thật. Huyết thanh sự thật số sẽ cho ta biết rằng thế giới tệ hại hơn ta nghĩ.

Ta cần biết điều này không? Biết về các tìm kiếm Google, dữ liệu khiêu dâm, và ai nhấp chuột lên cái gì—những chuyện ấy có thể không làm bạn nghĩ, "Cái này tuyệt quá. Mình có thể hiểu mình thực sự là ai."

[1] [ND] Howard Stern là người dẫn chương trình *The Howard Stern Show*, thường liên quan đến các vấn đề tin đồn và tình dục.

Trái lại, bạn sẽ nghĩ, "Cái này kinh khủng quá. Mình có thể hiểu mình thực sự là ai."

Nhưng sự thật này có ích—không chỉ cho Mark Zuckerberg hoặc những người đang mong thu hút các cú nhấp chuột hoặc khách hàng. Ít nhất có 3 cách mà kiến thức này có thể cải thiện cuộc sống của chúng ta.

Thứ nhất, ta có thể cảm thấy được an ủi khi biết rằng không phải chỉ một mình ta lâm vào cảnh bất an và có hành vi đáng xấu hổ. Có thể cũng tốt khi biết người khác cũng bất an về cơ thể họ. Có thể cũng tốt cho nhiều người—đặc biệt những ai kém hoạt động tình dục—khi biết cả thế giới này không phải ai cũng sinh hoạt dày đặc như thỏ. Và có thể cũng tốt cho một nam sinh trung học ở Mississippi đang mê anh tiền vệ trong đội bóng khi biết rằng, dù số nam giới đồng tính công khai rất thấp, vẫn có nhiều người khác cũng có cảm xúc hấp dẫn đồng giới tương tự mình.

Có một lĩnh vực khác—một lĩnh vực mà ta chưa thảo luận—ở đó các tìm kiếm Google có thể cho thấy bạn không đơn độc. Khi bạn còn nhỏ, có thể một giáo viên đã nói với bạn rằng, nếu có thắc mắc gì thì nên giơ tay lên hỏi, vì nếu bạn mơ hồ thì những người khác cũng mơ hồ. Nếu giống tôi, bạn sẽ phớt lờ lời khuyên của giáo viên và cứ ngồi ì ra im lặng, sợ không dám mở miệng. Bạn nghĩ các câu hỏi của mình quá dốt; câu hỏi của những người khác sâu sắc hơn. Dữ liệu tổng hợp nặc danh ở Google có thể cho ta biết các thầy cô nói chẳng sai tí nào. Nhiều câu hỏi cơ bản, kém sâu sắc cũng ẩn núp đâu đó trong đầu óc những người khác nữa.

Hãy xem thử đâu là các câu hỏi hàng đầu của người Mĩ trong buổi phát biểu Thông điệp liên bang 2014 của Obama.

**BẠN KHÔNG PHẢI LÀ NGƯỜI DUY NHẤT MUỐN BIẾT:**
**CÁC CÂU HỎI ĐƯỢC GOOGLE HÀNG ĐẦU TRONG THÔNG ĐIỆP LIÊN BANG**

---

Obama bao nhiêu tuổi?

Ai đang ngồi cạnh Biden?

Tại sao Boehner mang cà vạt xanh lá?

Tại sao Boehner màu da cam?

Bạn có thể đọc các câu hỏi này và nghĩ là nó chả liên quan gì đến nền dân chủ của chúng ta. Để ý về màu cà vạt hoặc màu da của một người thay vì nội dung bài diễn văn của tổng thống chứng tỏ chúng ta chẳng quan tâm gì nhiều. Không biết John Boehner (bấy giờ là người phát ngôn Hạ viện) là ai cũng chứng tỏ ta chẳng quan tâm nhiều đến chính trị.

Tôi thích nghĩ đến các câu hỏi chứng minh cho lời khuyên của các giáo viên hơn. Đây là những loại câu hỏi mà người ta thường không nêu ra, vì nghe có vẻ quá ngu ngốc. Nhưng rất nhiều người thắc mắc—và vào Google tìm hiểu.

Thực vậy, tôi nghĩ Dữ Liệu Lớn có thể cung cấp phiên bản cập nhật Thế kỉ XXI cho câu danh ngôn nổi tiếng sau: "Đừng bao giờ so sánh cái bên trong của ta với cái bên ngoài của mọi người."

Phiên bản danh ngôn thời Dữ Liệu Lớn: "Đừng bao giờ so sánh các tìm kiếm trên Google của ta với các đăng tải trên mạng xã hội của mọi người."

Ví dụ, hãy so sánh cách người ta mô tả các ông chồng trên mạng xã hội công khai và trong các tìm kiếm nặc danh.

**CÁCH NGƯỜI TA MÔ TẢ CHỒNG HỌ PHỔ BIẾN NHẤT**

| TRÊN MẠNG XÃ HỘI | TRONG CÁC TÌM KIẾM |
|---|---|
| the best<br>*tốt nhất* | gay<br>*đồng tính* |
| my best friend<br>*bạn tốt nhất của tôi* | a jerk<br>*gã khốn* |
| amazing<br>*tuyệt vời* | amazing<br>*tuyệt vời* |
| the greatest<br>*tuyệt nhất* | annoying<br>*phiền phức* |
| so cute<br>*dễ thương* | mean<br>*xấu tính* |

Vì chỉ có thể xem các đăng tải trên mạng xã hội của người khác chứ không xem được các tìm kiếm của họ, ta dễ phóng đại việc có bao nhiêu phụ nữ thường nghĩ chồng họ là "tốt nhất," "tuyệt nhất," và "dễ thương." Ta hay giảm thiểu việc có bao nhiêu phụ nữ nghĩ chồng họ là "gã khốn," "phiền phức," và "xấu tính."[1] Bằng phân tích dữ liệu nặc danh tổng hợp, ta có thể hiểu rằng mình không phải là người duy nhất thấy hôn nhân và cuộc sống sao quá đỗi khó khăn. Ta có thể dần biết cách dừng so sánh tìm kiếm của mình với các đăng tải trên mạng xã hội của người khác.

Lợi ích thứ hai của huyết thanh sự thật số là nó cảnh báo cho ta biết về những người đang phải chịu đựng. Human Rights Campaign đã yêu cầu tôi hợp tác để giúp giáo dục nam giới ở một số tiểu bang về khả năng công khai mình là đồng tính. Họ nhắm đến phương pháp sử dụng dữ liệu tìm kiếm Google tổng hợp và nặc danh để quyết định đâu là địa phương đáng tập trung nguồn lực nhất. Tương tự, các cơ quan bảo vệ trẻ em đã liên lạc với tôi để biết những địa phương có thể có tình trạng ngược đãi trẻ em vượt xa con số họ đang ghi nhận được.

Một chủ đề đáng ngạc nhiên mà tôi cũng được liên lạc để nhờ giúp đỡ: mùi âm đạo. Lần đầu tôi viết về điều này là trên *New York Times*, và tôi đã viết bằng giọng điệu khôi hài. Phần đó khiến tôi và những người khác cứ cố nín cười suốt.

Tuy nhiên, sau đó tôi xem bảng tin nhắn (xuất hiện khi có người thực hiện các tìm kiếm này) thì thấy vô số bài đăng từ các cô gái trẻ—họ tin rằng cuộc sống của mình bị hủy hoại là do lo lắng về mùi âm đạo. Đó chẳng phải chuyện đùa. Các chuyên gia giáo dục giới tính đã liên lạc với tôi, hỏi cách tận dụng một số dữ liệu Internet để giảm thiểu nỗi ám ảnh của các cô gái trẻ.

Mặc dù tôi cảm thấy mình hơi thiếu hiểu biết chiều sâu về vấn đề này, họ lại rất nghiêm túc, và tôi tin khoa học dữ liệu có thể giúp được.

---

[1] Tôi đã phân tích dữ liệu Twitter. Xin cảm ơn Emma Pierson đã giúp tải xuống phần này. Tôi không bao gồm các từ mô tả ông chồng đang làm gì, thứ thường thấy trên mạng xã hội nhưng không có ý nghĩa trong tìm kiếm. Ngay cả các mô tả này cũng nghiêng về phía tích cực. Cách phổ biến nhất để mô tả chồng đang làm gì trên mạng xã hội là "làm việc" và "nấu ăn."

Giá trị cuối cùng—và theo tôi là mạnh mẽ nhất—của huyết thanh sự thật số thực ra là khả năng đưa ta đi từ vấn đề đến giải pháp. Với nhiều hiểu biết hơn, ta có thể tìm thấy cách giảm thiểu nguồn tạo ra thái độ tiêu cực.

Hãy trở lại bài diễn văn của Obama về chứng sợ người Hồi giáo. Mỗi lần Obama nói rằng mọi người nên tôn trọng người Hồi giáo hơn, chính những người mà ông đang cố gắng tiếp cận lại trở nên điên cuồng hơn.

Tuy nhiên, tìm kiếm Google tiết lộ rằng có một câu nói thực sự kích nổ loại phản ứng tổng thống đương thời Obama mong muốn. Ông nói, "Người Mĩ gốc Hồi là bạn, là láng giềng, là đồng nghiệp, là anh hùng thể thao của chúng ta và, vâng, họ là những nam nữ quân nhân của chúng ta nữa, những người sẵn sàng hi sinh để bảo vệ đất nước."

Sau câu nói này, lần đầu tiên trong hơn một năm, danh từ được Google kèm với "Muslim" (Người Hồi giáo) nhiều nhất không phải là "khủng bố," "cực đoan," hoặc "tị nạn." Đó là "vận động viên," tiếp đến là "quân nhân." Và, thực tế, "vận động viên" giữ vị trí cao nhất suốt cả ngày hôm sau.

Khi ta lên lớp những người đang giận dữ, dữ liệu tìm kiếm ám chỉ rằng cơn thịnh nộ của họ có thể tăng lên. Nhưng khi tế nhị kích thích sự tò mò, cung cấp thông tin mới, và đưa ra những hình ảnh mới về nhóm người đang chọc khoáy cơn giận của họ, ta có thể chuyển ý nghĩ của họ theo hướng khác tích cực hơn.

Hai tháng sau bài diễn văn đầu tiên, Obama lại phát biểu trên truyền hình về chứng sợ người Hồi giáo, lần này tại một nhà thờ Hồi giáo. Có lẽ ai đó trong văn phòng tổng thống đã đọc mục *Times* của Soltas và tôi—tập trung bàn về cái gì hiệu quả và cái gì không hiệu quả ở bài diễn văn trước. Nội dung bài diễn văn lần này rất khác biệt.

Obama ít bỏ thời gian khẳng định giá trị của lòng khoan dung. Trái lại, ông tập trung chủ yếu vào việc kích thích tính tò mò của mọi người và thay đổi nhận thức của họ về người Mĩ Hồi giáo. Obama nói, phần nhiều các nô lệ đến từ châu Phi là người Hồi giáo; Thomas Jefferson và John Adams có quyển kinh Koran; nhà thờ Hồi giáo đầu tiên trên đất Mĩ

là ở North Dakota; một người Mĩ Hồi giáo đã thiết kế các nhà cao tầng ở Chicago. Obama lại nói về các vận động viên và các quân nhân Hồi giáo; ông cũng nói đến các sĩ quan cảnh sát và lính cứu hỏa, giáo viên và bác sĩ Hồi giáo nữa.

Và phân tích của tôi về các tìm kiếm Google chỉ ra rằng bài diễn văn này thành công hơn bài trước. Phần nhiều các tìm kiếm thù địch, giận dữ chống người Hồi giáo giảm mạnh trong mấy giờ sau cuộc nói chuyện của tổng thống.

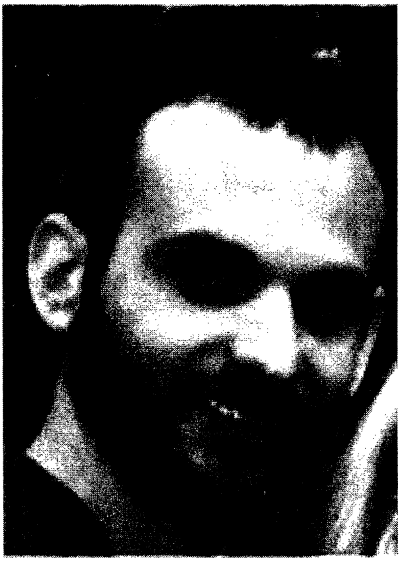
Có các cách tiềm năng khác để dùng dữ liệu tìm kiếm nhằm biết được điều gì gây ra—hoặc giảm thiểu—sự thù ghét. Ví dụ, ta có thể xem tìm kiếm phân biệt chủng tộc thay đổi thế nào sau khi một tiền vệ da đen gia nhập đội bóng của thành phố, hoặc tìm kiếm phân biệt giới tính thay đổi thế nào sau khi một phụ nữ được bầu làm lãnh đạo. Ta có thể thấy sự phân biệt chủng tộc phản ứng thế nào với chính sách cộng đồng, hoặc sự phân biệt giới tính phản ứng thế nào với luật quấy rối tình dục mới.

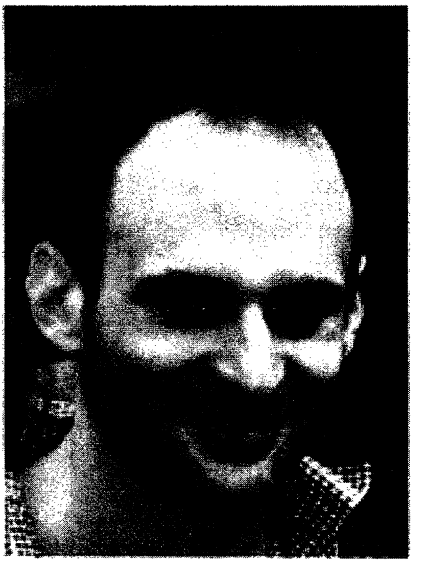
Biết về các thành kiến tiềm ẩn cũng rất có ích. Ví dụ, ta có thể nỗ lực hơn nữa để thấy hứng thú với trí tuệ của các cô gái nhỏ và bớt tỏ ra quá quan tâm về bề ngoài của các bé. Dữ liệu tìm kiếm Google và các suối nguồn sự thật khác trên Internet cho ta một cái nhìn chưa từng có vào các góc tối nhất của tâm hồn con người. Điều này, tôi thừa nhận, đôi khi khó đối mặt. Nhưng nó cũng có thể giúp ta tự chủ. Ta có thể dùng dữ liệu đó để chống lại sự đen tối. Thu thập dữ liệu phong phú các vấn đề của thế giới là bước đầu hướng tới việc sửa chữa thế giới này.

# CHƯƠNG 5

## Phóng to

Em trai tôi, Noah, nhỏ hơn tôi 4 tuổi. Lần đầu tiên gặp chúng tôi, hầu hết mọi người đều thấy hai anh em giống nhau một cách kì lạ. Cả hai đều nói rất to, hói đầu kiểu như nhau, và đều gặp khó khăn lớn trong việc giữ nhà cửa gọn gàng.





Seth Stephens-Davidowitz (trái)—fan bóng chày

Noah Stephens-Davidowitz (phải)—antifan bóng chày

Nhưng vẫn có những khác biệt: Tôi thì tính từng đồng. Noah toàn mua hàng cực đỉnh. Tôi thích nhạc Leonard Cohen và Bob Dylan. Noah thì thích Cake và Beck.

Có lẽ khác biệt đáng chú ý nhất giữa chúng tôi là thái độ đối với môn bóng chày. Tôi mê bóng chày, và đặc biệt, tình yêu dành cho đội New York Mets luôn là phần cốt lõi trong bản tính của tôi. Noah thì thấy bóng chày chán không chịu nổi, và việc ghét môn thể thao này từ lâu là phần cốt lõi trong bản tính của chú ấy.[1]

Làm sao mà 2 anh em với gene tương tự nhau, cùng cha mẹ nuôi dưỡng, ở cùng thành phố, lại có tình cảm đối nghịch về bóng chày như thế? Điều gì quyết định con người anh em tôi? Cơ bản hơn, điều gì *không ổn* với Noah? Có một lĩnh vực đang lên trong tâm lí học phát triển, khai thác các dữ liệu lớn về người trưởng thành và tương quan các dữ liệu ấy với các sự kiện chủ chốt thuở còn thơ ấu. Nó có thể giúp ta xử lí câu hỏi này và các câu hỏi liên quan. Ta có thể gọi việc sử dụng Dữ Liệu Lớn để trả lời các câu hỏi thuộc tâm lí học là Tâm Lí Lớn (Big Psych).

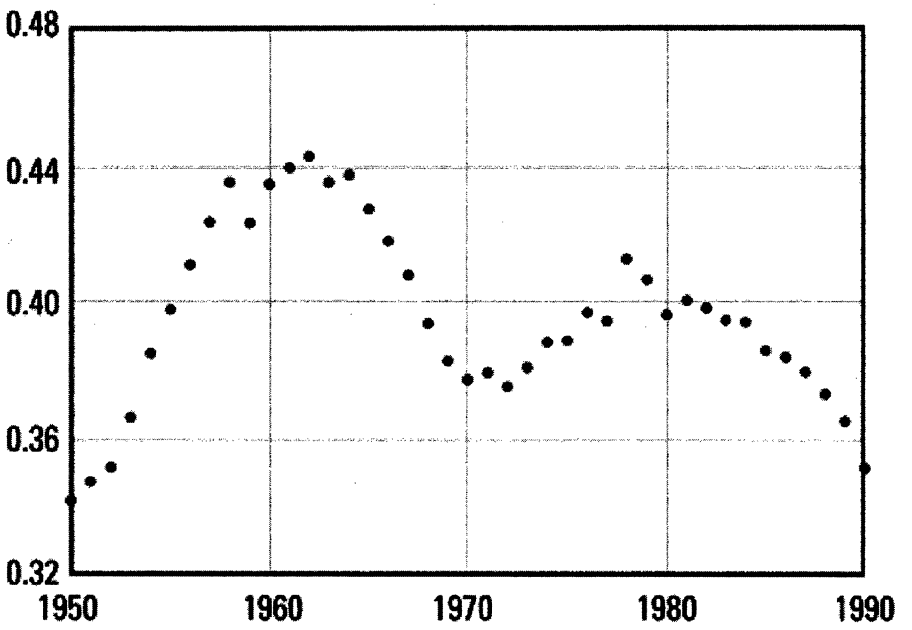
Để thấy cách lĩnh vực này vận hành, hãy xem xét một nghiên cứu mà tôi đã thực hiện về mức độ ảnh hưởng của các trải nghiệm tuổi thơ đến việc bạn ủng hộ đội bóng chày nào—hoặc không ủng hộ đội nào cả. Cho nghiên cứu này, tôi dùng dữ liệu "like" của các đội bóng chày trên Facebook. (Trong chương trước, tôi nhận xét rằng dữ liệu Facebook có thể làm sai lạc lớn trong các chủ đề nhạy cảm. Với nghiên cứu này, tôi giả định rằng không ai, ngay cả fan cuồng đội Phillies, thấy ngại khi thừa nhận là mình hâm mộ một đội bóng nào đó trên Facebook.)

Trước hết, tôi tải về số nam giới ở mọi lứa tuổi "like" 1 trong 2 đội bóng chày của New York. Đây là số phần trăm người hâm mộ đội bóng Mets, theo năm sinh.

---

[1] Xin tiết lộ đầy đủ: Khi tôi kiểm tra lại thông tin quyển sách này, Noah phủ nhận việc ghét bóng chày là phần chính yếu trong nhân cách của chú ấy. Noah thừa nhận ghét môn bóng chày, nhưng chú tin lòng tốt, lòng yêu trẻ, và trí tuệ của chú là các thành phần cốt lõi trong nhân cách của chú—và chuyện chú ấy ghét bóng chày thậm chí còn không nằm trong top 10. Tuy nhiên, tôi kết luận rằng đôi lúc khó mà thấy được bản tính của chính mình một cách khách quan, và với tư cách người quan sát bên ngoài, tôi thấy rằng việc ghét bóng chày thực sự thể hiện rõ Noah là ai, dù chú có nhận ra việc đó hay không. Vì vậy, tôi giữ nguyên không sửa phần này.

**Phần trăm nam hâm mộ bóng chày New York ưa thích đội Mets, theo năm sinh**

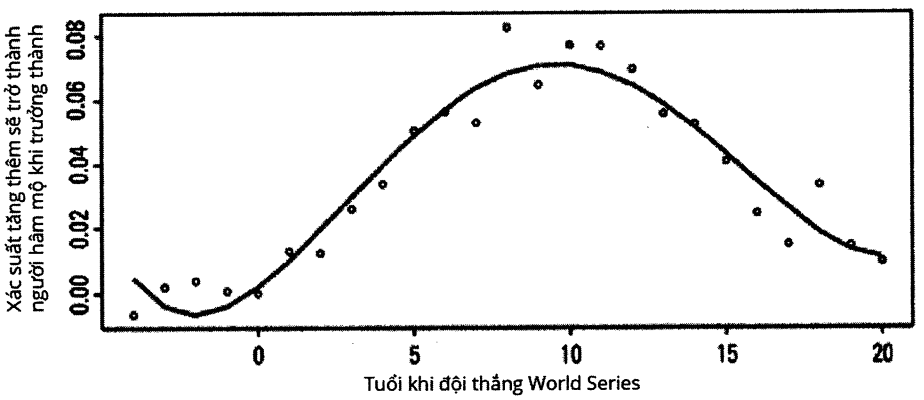



Dấu chấm càng cao thì người hâm mộ Mets càng nhiều. Sự mến mộ đội bóng lên rồi xuống 2 lần, đội Mets rất được những người sinh vào khoảng năm 1962 và 1978 mến mộ. Những người hâm mộ bóng chày có thể hiểu điều gì đang diễn ra ở đây. Đội Mets thắng chỉ 2 giải World Series: năm 1969 và 1986. Những nam hâm mộ này lúc đó khoảng 7 đến 8 tuổi khi Mets thắng giải. Như thế, một nhân tố dự báo rất mạnh về cộng đồng hâm mộ Mets, ít nhất với các cậu trai, là liệu Mets có thắng giải World Series không khi người hâm mộ đang khoảng 7-8 tuổi.

Thực vậy, chúng ta có thể mở rộng bản phân tích này. Tôi tải dữ liệu Facebook cho thấy bao nhiêu người hâm mộ ở mọi lứa tuổi "like" các đội tuyển Major League Baseball.

Tôi phát hiện cũng có một con số cao bất thường các nam hâm mộ đội Baltimore Orioles sinh năm 1962 và các nam hâm mộ đội Pittsburgh Pirates sinh năm 1963. Các nam hâm mộ đó từng là các cậu bé 8 tuổi khi

các đội này vô địch. Thực vậy, khi tính tuổi nhóm hâm mộ đông nhất của các đội tôi nghiên cứu, sau đó tính ra những người hâm mộ này bao nhiêu tuổi [khi đội vô địch], tôi có biểu đồ sau:




Một lần nữa ta thấy rằng năm quan trọng nhất trong đời của một người để củng cố vị thế đội bóng yêu thích khi người này trưởng thành chính là lúc anh ta trên dưới 8 tuổi. Nói chung, từ 5 đến 15 tuổi là thời kì chủ chốt để thu hút một cậu bé. Chiến thắng khi một cổ động viên ở tuổi 19 hoặc 20 thì mức độ thu hút sự hâm mộ của anh ta chỉ bằng khoảng 1/8 nếu thắng lúc anh ta 8 tuổi. Ở tuổi 19-20, hoặc anh ta đã hâm mộ một đội suốt đời rồi, hoặc đã không hâm mộ đội nào cả.

Bạn có thể muốn hỏi: Còn các nữ hâm mộ bóng chày thì sao? Mô hình ít đột ngột hơn nhiều, nhưng độ tuổi tác động nhiều nhất dường như là 22 tuổi.

Đây là nghiên cứu tâm đắc của tôi. Nó liên quan đến 2 trong các chủ đề tôi ưa thích nhất: bóng chày và các nguồn gây bất mãn khi tôi trưởng thành. Tôi bị dính chặt vào đội Mets năm 1986 và đã phải đau khổ dài dài với các thất bại đội này kể từ đó. Noah thì may mắn được sinh ra sau đó 4 năm và tránh được nỗi đau này.

Bóng chày không phải là chủ đề quan trọng nhất trên thế giới, các giáo sư hướng dẫn luận án tiến sĩ của tôi thường nhắc tôi như vậy. Nhưng phương pháp này có thể giúp ta xử lí các câu hỏi tương tự, bao gồm nguyên nhân quyết định khuynh hướng chính trị, tính dục, sở thích

âm nhạc, và các thói quen tài chính của mọi người. (Tôi đặc biệt quan tâm về nguồn gốc các ý kiến lập dị của em trai tôi ở 2 chủ đề cuối.) Dự báo của tôi là chúng ta sẽ thấy rằng nhiều hành vi và sở thích ở tuổi trưởng thành, ngay cả những thứ ta xem là căn cơ nhất, giúp xác định ta là ai, có thể được giải thích bởi các sự kiện võ đoán như ta được sinh ra khi nào và điều gì đang diễn ra vào những năm chủ chốt lúc ta còn nhỏ.

Thực ra, đã có một số nghiên cứu về nguồn gốc các khuynh hướng chính trị. Yair Ghitza, khoa học gia trưởng tại Catalist, một công ty phân tích dữ liệu, và Andrew Gelman, một nhà khoa học chính trị kiêm nhà thống kê tại Đại học Columbia, đã thử kiểm tra ý kiến truyền thống cho rằng hầu hết mọi người lúc nhỏ đều theo khuynh hướng tự do, rồi ngày càng trở nên bảo thủ khi lớn tuổi. Đây là quan điểm thể hiện trong câu nói nổi tiếng thường được gán cho Winston Churchill: "Ai dưới 30 mà không phải người tự do thì không có trái tim; còn ai trên 30 mà không phải người bảo thủ thì không có đầu óc."

Ghitza và Gelman nghiền ngẫm 60 năm dữ liệu khảo sát, sử dụng hơn 300,000 quan sát về khuynh hướng bầu cử. Trái với câu nói của Churchill, hai ông phát hiện rằng thanh thiếu niên khi thì nghiêng về tự do, khi thì nghiêng về bảo thủ. Tuổi trung niên và tuổi lớn hơn cũng thế.

Các nhà nghiên cứu này phát hiện rằng quan điểm chính trị thực ra hình thành theo cách không khác gì việc ta yêu thích một đội thể thao cả. Có một thời kì quan trọng in dấu ấn lên người ta suốt đời. Giai đoạn từ 14 đến 24 tuổi, nhiều người Mĩ sẽ hình thành quan điểm trên cơ sở mức độ được ưa thích của tổng thống đương thời. Một tổng thống Cộng hòa được ưa thích hoặc Dân chủ không được ưa thích sẽ khiến nhiều người trẻ theo phe Cộng hòa. Một tổng thống Cộng hòa không được ưa thích hoặc Dân chủ được ưa thích đưa nhóm người ở độ tuổi nhạy cảm này vào phe Dân chủ.

Và các quan điểm này, vào các năm chủ chốt này, nói chung là sẽ tồn tại suốt cả đời.

Để xem điều này hoạt động ra sao, hãy so sánh những người Mĩ sinh năm 1941 và những người sinh sau đó 1 thập niên.

Những người nhóm 1941 trưởng thành trong thời Dwight D. Eisenhower làm tổng thống, một người Cộng hòa khá được ưa thích. Đầu thập niên 1960, mặc dù dưới 30 tuổi, thế hệ này nghiêng mạnh về phía đảng Cộng hòa. Và các thành viên thế hệ này kiên định nghiêng về đảng Cộng hòa khi họ đã có tuổi.

Những người Mĩ sinh sau đó 10 năm—thế hệ Baby Boomer—trưởng thành trong các nhiệm kì tổng thống của John F. Kennedy, một người Dân chủ cực kì được ưa thích; Lyndon B. Johnson, một người Dân chủ được ưa thích lúc đầu; và Richard M. Nixon, một người Cộng hòa cuối cùng đã về vườn trong sự nhục nhã. Các thành viên thế hệ này nghiêng về phe tự do suốt cả đời.

Với tất cả dữ liệu này, các nhà nghiên cứu đã quyết định được năm quan trọng nhất cho sự phát triển quan điểm chính trị: tuổi 18.

Và họ phát hiện rằng các hiệu ứng dấu ấn này là rất đáng kể. Mô hình của họ ước tính rằng trải nghiệm về Eisenhower làm tăng khoảng 10 %p người theo phe Cộng hòa suốt đời trong số những người Mĩ sinh năm 1941. Trải nghiệm về Kennedy, Johnson, và Nixon cho phe Dân chủ một lợi thế 7 %p trong số những người Mĩ sinh năm 1952.

Tôi đã nói rõ rằng tôi khá hoài nghi các dữ liệu từ khảo sát, nhưng tôi ấn tượng với số lượng lớn phản hồi ở nghiên cứu này. Thực vậy, nghiên cứu này hẳn không thể nào được thực hiện bằng một cuộc khảo sát nhỏ. Các nhà nghiên cứu cần hàng trăm ngàn quan sát, kết hợp từ nhiều cuộc khảo sát, để xem quan điểm thay đổi như thế nào khi người ta có tuổi.

Kích cỡ dữ liệu cũng quan trọng trong nghiên cứu bóng chày của tôi. Tôi cần phóng to không chỉ vào người hâm mộ mỗi đội, mà còn vào số người ở mọi lứa tuổi nữa. Cần phải có hàng triệu quan sát, và Facebook cũng như các nguồn kĩ thuật số khác hàng ngày vẫn cung cấp lượng quan sát lớn đến thế.

Đây là chỗ mà sự to lớn của Dữ Liệu Lớn thực sự đóng vai trò quan trọng. Bạn cần nhiều pixel trong một tấm ảnh để có thể phóng to rõ nét một phần tấm ảnh đó. Tương tự, bạn cần nhiều quan sát trong một bộ

dữ liệu để có thể phóng to rõ ràng một bộ phận nhỏ của dữ liệu đó—ví dụ, để xem đội Mets nổi tiếng cỡ nào đối với nam giới sinh năm 1978. Một khảo sát nhỏ vài ngàn người sẽ không có mẫu đại diện đủ lớn cho nam giới thuộc nhóm này.

Đây là sức mạnh thứ ba của Dữ Liệu Lớn: Dữ Liệu Lớn cho phép ta phóng to một cách có ý nghĩa các phần nhỏ của một bộ dữ liệu để có những hiểu biết mới. Và ta có thể phóng to các chiều kích khác chứ không chỉ độ tuổi. Nếu có đủ dữ liệu, ta có thể thấy mọi người ở từng thành thị cụ thể cư xử như thế nào. Và ta có thể thấy người ta cư xử ra sao từng giờ hoặc thậm chí từng phút nữa.

Trong chương này, hành vi con người sẽ được xem xét cận cảnh.

## Điều gì đang diễn ra tại các hạt, thành phố, và thị trấn?

Giờ nghĩ lại thì thật đáng ngạc nhiên. Thế nhưng, khi Raj Chetty, bấy giờ là giáo sư tại Harvard, và một nhóm nghiên cứu nhỏ mới thu thập được một bộ dữ liệu khá lớn—hồ sơ thuế của tất cả người Mĩ từ năm 1996—họ không chắc sẽ tìm thấy gì ở đó. Cục thuế (IRS) cung cấp dữ liệu này vì họ nghĩ các nhà nghiên cứu có thể dùng để giúp làm sáng tỏ hiệu quả của chính sách thuế.

Những cố gắng ban đầu để sử dụng Dữ Liệu Lớn trong tay của nhóm Chetty thực ra đã dẫn đến nhiều ngõ cụt. Các cuộc kiểm tra kết quả những chính sách thuế tiểu bang và liên bang của họ chủ yếu đi đến chung những kết luận mà mọi người có chỉ bằng việc dùng các khảo sát. Có lẽ các câu trả lời của Chetty, khi dùng hàng trăm triệu điểm dữ liệu IRS, chỉ chính xác hơn một chút mà thôi. Nhưng có câu trả lời tương tự như mọi người, dù là chính xác hơn một chút, không phải là một thành tựu khoa học xã hội lớn. Đó không phải là loại công trình mà các báo chuyên ngành hàng đầu háo hức đăng.

Hơn nữa, việc tổ chức và phân tích tất cả dữ liệu IRS rất mất thời gian. Nhóm Chetty chìm trong dữ liệu, mất nhiều thời gian hơn mọi người chỉ để tìm thấy các câu trả lời tương tự như mọi người.

Bắt đầu có vẻ như những người hoài nghi Dữ Liệu Lớn đã đúng. Bạn không cần dữ liệu của hàng trăm triệu người Mĩ để hiểu chính sách thuế; một khảo sát 10,000 người đã là đủ nhiều. Nhóm Chetty thoái chí cũng là điều dễ hiểu.

Và rồi, cuối cùng thì các nhà nghiên cứu nhận ra sai lầm. "Dữ Liệu Lớn rõ ràng không phải là làm thứ mà bạn vẫn làm với các khảo sát, chỉ khác ở chỗ dùng nhiều dữ liệu hơn," Chetty giải thích. Họ đang đặt quá ít câu hỏi dữ liệu về bộ dữ liệu khổng lồ họ nắm trong tay. "Dữ Liệu Lớn thực sự phải cho phép bạn dùng các mô hình hoàn toàn khác với những gì bạn sẽ có từ các khảo sát," Chetty nói thêm. "Bạn có thể, ví dụ, phóng to các vùng địa lí."

Nói cách khác, với dữ liệu về hàng trăm triệu người, nhóm Chetty có thể xác định các khuôn mẫu trong các thành phố, thị trấn, và khu dân cư, dù lớn hay nhỏ.

Là một nghiên cứu sinh tại Harvard, tôi có mặt trong phòng thảo luận khi Chetty trình bày kết quả ban đầu thu được từ hồ sơ thuế của tất cả mọi người Mĩ. Các nhà khoa học xã hội hay nói về số quan sát trong công trình của mình—cụ thể là họ có bao nhiêu điểm dữ liệu. Nếu một nhà khoa học xã hội đang thực hiện một khảo sát 800 người, ông sẽ nói, "Chúng tôi có 800 quan sát." Nếu ông đang thực hiện nghiên cứu trong phòng thí nghiệm với 70 người, ông sẽ nói, "Chúng tôi có 70 quan sát."

"Chúng tôi có 1.2 tỉ quan sát," Chetty nói, nét mặt thản nhiên. Khán giả khúc khích cười một cách lo lắng.

Tại phòng thảo luận đó và về sau là một loạt các bài báo, Chetty và các đồng tác giả của ông bắt đầu cho chúng ta những hiểu biết mới rất quan trọng về cách nước Mĩ vận hành.

Xét câu hỏi này: Phải chăng nước Mĩ là miền đất của cơ hội? Nếu cha mẹ không giàu, tự thân bạn có cơ hội để phát tài không?

Cách truyền thống để trả lời câu hỏi này là xem xét một mẫu đại diện những người Mĩ và so sánh mẫu này với dữ liệu tương tự từ các nước khác.

Dưới đây là dữ liệu cho nhiều nước về tính bình đẳng cơ hội. Câu hỏi nêu ra: Cơ hội mà một người có cha mẹ thuộc nhóm 20% có thu nhập thấp nhất ngoi lên được nhóm 20% có thu nhập cao nhất là bao nhiêu?

| KHẢ NĂNG MỘT NGƯỜI CÓ CHA MẸ NGHÈO SẼ TRỞ NÊN GIÀU CÓ (Ở MỘT SỐ QUỐC GIA) | |
|---|---|
| Mĩ | 7.5 |
| Anh | 9.0 |
| Đan Mạch | 11.7 |
| Canada | 13.5 |

Như bạn có thể thấy, điểm số nước Mĩ *không tốt lắm*.

Nhưng phân tích đơn giản này bỏ qua câu chuyện thực tế. Nhóm của Chetty đã phóng to theo vùng địa lí. Họ phát hiện cơ hội rất khác biệt tùy vào nơi bạn sinh ra tại Mĩ.

| KHẢ NĂNG MỘT NGƯỜI CÓ CHA MẸ NGHÈO SẼ TRỞ NÊN GIÀU CÓ (Ở MỘT SỐ ĐỊA PHƯƠNG TẠI MĨ) | |
|---|---|
| San Jose, CA | 12.9 |
| Washington, DC | 10.5 |
| *Trung bình nước Mĩ* | *7.5* |
| Chicago, IL | 6.5 |
| Charlotte, NC | 4.4 |

Một số nơi tại Mĩ, cơ hội thành công của trẻ nghèo cao ngang bằng bất cứ nước phát triển nào trên thế giới. Ở một số nơi khác, cơ hội thành công của trẻ nghèo thấp hơn bất cứ nước phát triển nào trên thế giới.

Các khuôn mẫu này sẽ không bao giờ được phát hiện trong một khảo sát nhỏ—có thể chỉ bao gồm một vài người ở Charlotte và San Jose, và vì vậy sẽ không cho phép ta phóng to lên như thế này.

Thực ra, nhóm của Chetty có thể phóng to hơn thế nữa. Bởi họ có rất nhiều giữ liệu—dữ liệu về *từng* người Mĩ—họ có thể phóng to cả các nhóm nhỏ chuyển từ thành phố này đến thành phố khác để xem điều đó có thể ảnh hưởng đến tiền đồ của họ như thế nào: các nhóm người chuyển từ Thành phố New York đến Los Angeles, Milwaukee đến Atlanta, San Jose đến Charlotte. Điều này giúp họ kiểm tra quan hệ nhân quả, chứ không chỉ tương quan (một điểm nổi bật mà tôi sẽ bàn trong chương kế tiếp). Và, vâng, việc chuyển đến đúng thành phố ở độ tuổi định hình cuộc sống sẽ tạo nên khác biệt lớn.

Vậy có phải nước Mĩ là "miền đất của cơ hội" hay không?

Câu trả lời chẳng phải có mà cũng chẳng phải không. Câu trả lời là: Một số vùng thì phải, và một số vùng thì không.

Các tác giả viết, "Nước Mĩ tốt hơn nên được mô tả là một tập hợp các xã hội, một số xã hội trong đó là những 'vùng đất của cơ hội' với tốc độ dịch chuyển[1] (mobility) cao qua các thế hệ, và một số xã hội khác thì ít có trẻ em nào thoát khỏi cái nghèo."

Thế ở những vùng có mức độ dịch chuyển thu nhập cao có gì đặc biệt? Điều gì khiến một số nơi làm khá tốt việc san bằng cơ hội, cho phép trẻ em nghèo có một cuộc sống khá tốt? Những vùng chi nhiều hơn về giáo dục tạo được cơ hội tốt hơn cho trẻ nghèo. Nơi có nhiều người theo đạo và ít tội phạm cũng làm tốt hơn. Địa phương có nhiều người da đen thì kém hơn. Thật thú vị, điều này ảnh hưởng không chỉ trẻ em da đen mà còn trẻ em da trắng sống ở đó nữa. Những nơi có nhiều người mẹ đơn thân cũng kém hơn. Ảnh hưởng này cũng không chỉ đối với trẻ em có mẹ đơn thân mà còn với cả trẻ em có đủ cha mẹ sống ở nơi có nhiều người mẹ đơn thân nữa. Một số các kết quả trên cho thấy, bạn bè của trẻ nghèo cũng đóng vai trò rất quan trọng. Nếu bạn bè có nền tảng khó khăn và ít cơ hội, bản thân trẻ đó cũng sẽ phải vất vả hơn để thoát nghèo.

---

[1] [ND] Sự dịch chuyển (ở đây tác giả nói về sự dịch chuyển xã hội—social mobility) chỉ sự chuyển động của con người từ vị thế xã hội này sang vị thế xã hội khác. Sự dịch chuyển đang được nói tới là về thu nhập qua thế hệ, chủ yếu chỉ sự dịch chuyển từ vị thế thấp của đời trước sang vị thế cao hơn ở đời sau (hay nói đơn giản là đời con giàu có hơn đời cha).

Dữ liệu cho ta biết rằng một số vùng trong nước Mĩ cho trẻ nhiều cơ hội hơn để thoát nghèo. Vậy những nơi nào cho người ta nhiều cơ hội thoát lưỡi hái Tử thần nhất?

***

Chúng ta thường nghĩ cái chết thật công bằng. Xét cho cùng, không ai tránh được cái chết. Dân đen hay vua chúa, kẻ vô gia cư hay Mark Zuckerberg, ai cũng sẽ chết.

Nhưng tuy là người giàu không tránh được cái chết, dữ liệu lại cho ta biết rằng bây giờ họ đã có thể hoãn chết lại. Phụ nữ Mĩ trong top 1% thu nhập cao nhất tính trung bình sống thọ hơn 10 năm so với phụ nữ Mĩ trong top 1% thu nhập thấp nhất. Với nam thì là 15 năm.

Khuôn mẫu này thay đổi thế nào ở những vùng khác nhau tại Mĩ? Tuổi thọ của bạn có khác nhau tùy thuộc nơi bạn sống không? Sự sai lệch này có khác giữa người giàu và người nghèo không? Một lần nữa, bằng cách phóng to về vùng địa lí, nhóm của Raj Chetty đã tìm thấy câu trả lời.

Thật thú vị, đối với những người Mĩ giàu có nhất, tuổi thọ hầu như không bị ảnh hưởng bởi nơi họ sống. Khi bạn đã có ê hề tiền bạc, bạn có thể kì vọng rằng mình sẽ sống đến khoảng 89 tuổi nếu là nữ và khoảng 87 tuổi nếu là nam. Người giàu khắp mọi nơi có khuynh hướng phát triển các thói quen lành mạnh hơn—nói chung, họ tập thể dục nhiều hơn, ăn uống kĩ hơn, hút hít ít hơn, và ít có khả năng bị béo phì hơn. Người giàu thì có tiền mua máy chạy bộ, ăn bơ hữu cơ, đăng kí lớp yoga. Và họ có thể mua các thứ này ở bất cứ xó trời nào trên đất Mĩ.

Đối với người nghèo, câu chuyện sẽ khác. Với những người Mĩ nghèo nhất, tuổi thọ thay đổi lớn tùy thuộc nơi họ sống. Thực vậy, sống ở nơi thích hợp có thể cộng thêm 5 năm vào tuổi thọ của một người nghèo.

Vậy tại sao một số nơi có vẻ như giúp người nghèo khổ sống thọ hơn rất nhiều? Các thành phố nơi mà người nghèo sống thọ nhất có chung những thuộc tính gì?

Đây là 4 thuộc tính của 1 thành phố—trong đó 3 thuộc tính không tương quan với tuổi thọ người nghèo, còn 1 thì có. Xem thử bạn có đoán được thuộc tính nào là thuộc tính cần tìm không nhé.

**ĐIỀU GÌ KHIẾN NGƯỜI NGHÈO TRONG THÀNH PHỐ SỐNG LÂU HƠN?**

Thành phố có mức độ tín ngưỡng cao.

Thành phố có mức độ ô nhiễm thấp.

Thành phố có tỉ lệ cư dân được bảo hiểm y tế cao.

Nhiều người giàu sống trong thành phố.

3 thuộc tính đầu—tôn giáo, môi trường, và bảo hiểm y tế—không tương quan với tuổi thọ người nghèo. Biến số thực sự quan trọng, theo Chetty và những người thực hiện nghiên cứu này: số lượng người giàu sống trong thành phố. Đông người giàu đồng nghĩa với việc người nghèo trong thành phố sẽ sống thọ hơn. Người nghèo ở Thành phố New York chẳng hạn, sống thọ hơn nhiều so với người nghèo ở Detroit.

Tại sao sự hiện diện của người giàu lại là một biến dự báo mạnh tuổi thọ của người nghèo? Một giả thiết—có tính chất suy đoán—đã được đưa ra bởi David Cutler, một trong các tác giả cuộc nghiên cứu và là một trong các cố vấn của tôi. Hành vi lây lan có thể đang điều khiển một phần hiện tượng này.

Một số lượng lớn nghiên cứu chỉ ra rằng các thói quen có tính lây nhiễm. Vậy người nghèo sống gần người giàu có thể học theo nhiều thói quen của họ. Một số các thói quen này—ví dụ như dùng từ ngữ sang chảnh—chắc là không ảnh hưởng đến sức khỏe. Những thói quen khác—như tập thể dục chẳng hạn—rõ ràng sẽ có tác động tích cực. Thực vậy, người nghèo sống gần người giàu tập thể dục nhiều hơn, hút thuốc ít hơn, và ít bị béo phì hơn.

* * *

Nghiên cứu của nhóm Raj Chetty mà cá nhân tôi tâm đắc là phần điều tra lí do tại sao một số người lại gian lận thuế trong khi những người khác thì không. Giải thích nghiên cứu này hơi phức tạp một chút.

Mấu chốt nằm ở chỗ biết rằng có một cách thức dễ dàng để những người làm tư nhân đang nuôi 1 con tối đa hóa số tiền họ nhận được từ chính phủ. Nếu bạn báo cáo rằng bạn có thu nhập phải đóng thuế chính xác là 9,000 USD trong một năm nào đó, chính phủ sẽ viết cho bạn một ngân phiếu 1,377 USD—đây là tiền tín dụng thuế thu nhập (Earned Income Tax Credit—EITC), một khoản trợ cấp giúp giảm thuế tiền lương cho những người lao động nghèo. Báo cáo nhiều hơn số đó thì thuế tiền lương sẽ tăng lên. Báo cáo ít hơn số đó thì trợ cấp EITC giảm. Thu nhập chịu thuế 9,000 USD là tối ưu.

Và bạn có biết không, 9,000 USD là khoản thu nhập chịu thuế thường gặp nhất được báo cáo bởi những người làm tư có 1 con.

Những người Mĩ này có điều chỉnh lịch làm việc để bảo đảm họ kiếm được lượng thu nhập hoàn hảo không? Không. Khi những người lao động này được kiểm toán ngẫu nhiên—một điều rất hiếm khi xảy ra—hầu như người ta đều phát hiện là họ không ở gần mốc 9,000 USD. Thực chất họ kiếm được ít hơn hoặc nhiều hơn mức này rất xa.

Nói cách khác, họ gian lận bằng cách khai khống rằng đang ở mức thu nhập giúp họ kiếm được tấm ngân phiếu béo bở nhất từ chính phủ.

Vậy gian lận thuế loại này có tiêu biểu không, và ai trong số những người làm tư nhân có 1 con có nhiều khả năng vi phạm nhất? Hóa ra, Chetty và các đồng nghiệp báo cáo rằng có sự khác biệt rất lớn về mức độ gian lận thuế loại này. Tại Miami, trong số những người thuộc nhóm này (làm tư, 1 con), thật ngạc nhiên, có đến 30% báo cáo họ kiếm được 9,000 USD. Tại Philadelphia thì chỉ có 2%.

Điều gì dự báo ai sẽ gian lận thuế? Những nơi có nhiều người gian lận và những nơi có ít người gian lận có điểm gì đặc biệt? Chúng ta có thể so sánh mối tương quan giữa tỉ lệ gian lận với các yếu tố thống kê dân số cấp thành phố, và hóa ra có 2 chỉ số dự báo mạnh: (1) mức độ tập trung cao những người trong vùng đủ tiêu chuẩn nhận EITC, và (2) mức độ tập trung cao các chuyên gia ngành thuế trong khu dân cư đó.

Các yếu tố này chỉ ra điều gì? Nhóm Chetty có một hướng giải thích. Động lực chủ chốt để gian lận thuế theo cách này chính là thông tin.

Hầu hết những người đóng thuế đang làm tư và có 1 con đơn giản không biết rằng con số thần kì để nhận tấm ngân phiếu béo bở từ chính phủ là 9,000 USD. Nhưng sống gần những người có hiểu biết—láng giềng hoặc trợ lí thuế—sẽ tăng đáng kể cơ hội họ biết được điều đó.

Thực vậy, nhóm Chetty phát hiện thêm nhiều bằng chứng là kiến thức đóng vai trò quyết định. Khi người Mĩ chuyển từ một vùng mà kiểu gian lận thuế này thấp đến vùng gian lận cao, họ học và áp dụng trò gian đó. Theo thời gian, tình trạng gian lận lan từ vùng này đến vùng khác khắp nước Mĩ. Giống như một loại virus, gian lận thuế có tính lây lan.

Hãy dừng lại một lát và suy nghĩ xem nghiên cứu này cho thấy điều gì. Nó chỉ ra rằng, khi tìm hiểu xem ai sẽ gian lận thuế, câu trả lời không phải là xác định ai lương thiện và ai không lương thiện. Câu trả lời là xác định ai biết cách gian lận và ai không biết.

Vậy khi ai đó nói với bạn họ sẽ không bao giờ gian lận thuế, rất có thể là họ...—bạn cũng đoán ra rồi đấy—đang nói dối. Nghiên cứu của Chetty chỉ ra rằng nhiều người sẽ gian lận thuế nếu biết cách.

Nếu bạn muốn gian lận thuế (tôi *không* có khuyên làm điều này nhé), bạn nên sống gần các chuyên gia ngành thuế hoặc sống gần những người đã gian lận thuế—họ có thể chỉ cách cho bạn. Vậy còn nếu muốn con cái nổi tiếng thế giới, bạn nên sống ở đâu? Khả năng phóng to dữ liệu này và xem xét thật chi tiết cũng có thể giúp trả lời câu hỏi đó.

Tôi tò mò muốn biết những người Mĩ thành công nhất từ đâu đến, vì vậy một ngày nọ tôi quyết định tải Wikipedia về. (Ngày nay bạn cũng có thể làm việc đó.) Với một chút mã hóa, tôi có một bộ dữ liệu hơn 150,000 người Mĩ được các biên tập viên của Wikipedia cho là nổi bật đủ để được có một bài viết riêng. Bộ dữ liệu bao gồm nơi sinh (chi tiết đến cấp độ hạt-county), ngày sinh, nghề nghiệp, và giới tính. Tôi gộp nó lại với dữ liệu nơi sinh cấp hạt được thu thập bởi Trung tâm Thống kê Y tế Quốc gia. Tôi tính toán cơ hội được đưa vào Wikipedia nếu bạn sinh ra ở từng hạt.

Liệu việc được ghi tiểu sử trong Wikipedia có là một dấu ấn quan trọng cho thấy bạn có thành tựu nổi bật không? Chắc chắn vẫn có những

giới hạn. Các biên tập viên Wikipedia thường trẻ và là nam giới, điều đó có thể làm thiên lệch mẫu đại diện. Và một số kiểu nổi tiếng không có giá trị cho lắm. Ted Bundy, chẳng hạn, được xuất hiện trên Wikipedia vì hắn đã giết hàng chục phụ nữ trẻ. Dù vậy, tôi vẫn có thể loại bỏ các tội phạm mà không ảnh hưởng nhiều đến kết quả.

Tôi giới hạn nghiên cứu trong thế hệ Baby Boomer (những người sinh trong khoảng 1946—1964) bởi họ đã có gần cả cuộc đời để nổi tiếng. Trong nhóm người thế hệ ấy sinh ở Mĩ, cứ 2058 người thì có 1 người được xem là nổi tiếng đủ để vào Wikipedia. Khoảng 30% nổi tiếng là nhờ thành tựu trong ngành nghệ thuật hoặc giải trí, 29% nhờ thể thao, 9% qua đường chính trị, và 3% là trong mảng học thuật hoặc khoa học.

Sự thật đáng kinh ngạc đầu tiên tôi chú ý thấy trong dữ liệu là sự khác biệt vùng địa lí rất lớn về khả năng thành công, ít nhất là theo tiêu chí của Wikipedia. Cơ hội bạn được nổi tiếng lệ thuộc rất nhiều vào nơi bạn sinh ra.

Khoảng 1 trong 1209 người thế hệ Baby Boomer sinh ở California có mặt trong Wikipedia. Chỉ 1 trong 4496 người sinh ở West Virginia góp mặt. Khi phóng to theo hạt, kết quả ấn tượng hơn nhiều. Khoảng 1 trong 748 người sinh ở hạt Suffolk, Massachusetts, nơi có thành phố Boston, có trong Wikipedia. Tại một số nước khác, tỉ lệ thành công thấp hơn 20 lần.

Tại sao một số nơi trong nước có vẻ khá "xuất sắc" trong việc sản sinh ra những người nổi tiếng của nước Mĩ? Tôi khảo sát kĩ các hạt đứng đầu. Hóa ra gần như tất cả các hạt đó đều thuộc 1 trong 2 loại sau đây:

Thứ nhất, và điều này khiến tôi ngạc nhiên, phần nhiều các hạt này có một thành phố đại học[1] tầm cỡ. Hầu như mỗi lần tôi thấy tên của một hạt mà tôi chưa từng nghe đến nằm gần đầu danh sách (như Washtenaw, Michigan), tôi lại phát hiện rằng nó chịu tác động bởi một thành phố đại học cổ điển (trong trường hợp này là Ann Arbor). Các hạt

[1] [ND] Thành phố đại học (college town) là một cộng đồng dân cư chịu ảnh hưởng lớn từ một hoặc nhiều đại học đặt tại đó. Các hoạt động kinh tế-xã hội tại đây cũng thường xoay quanh đại học đặt tại địa phương. Mô hình này tương tự các làng đại học tại Việt Nam (nhưng với quy mô lớn hơn).

được vẻ vang nhờ các thành phố đại học—Madison, Wisconsin; Athens, Georgia; Columbia, Missouri; Berkeley, California; Chapel Hill, North Carolina; Gainesville, Florida; Lexington, Kentucky; và Ithaca, New York— tất cả đều ở top 3%.

Tại sao vậy? Một phần rất có thể là nhờ vốn gene: Con cái các giáo sư và nghiên cứu sinh có khuynh hướng thông minh (một đặc điểm mà, trong trò chơi thành công lớn, có thể vô cùng hữu ích). Và, thực tế, việc có nhiều người tốt nghiệp đại học trong vùng là một chỉ báo mạnh sự thành công của những người sinh ra ở đó.

Nhưng rất có thể có một điều khác nữa: cơ hội tiếp xúc sớm với sự đổi mới. Một trong các lĩnh vực mà thành phố đại học sản sinh ra những con chim đầu đàn rất thành công là lĩnh vực âm nhạc. Trẻ em sống ở thành phố đại học sẽ được tiếp xúc các buổi hòa nhạc độc đáo, các đài phát thanh đặc sắc, và cả các cửa hàng băng đĩa nhạc độc lập nữa. Điều này không chỉ xuất hiện trong mảng nghệ thuật. Thành phố đại học còn ươm mầm nhiều doanh nhân nổi bật hơn cả mong đợi. Có lẽ việc tiếp xúc ban đầu với nghệ thuật và ý tưởng hiện đại cũng rất có ích.

Sự thành công của các thành phố đại học không chỉ bất chấp vùng miền. Nó còn bất chấp cả chủng tộc. Người Mĩ gốc Phi rõ ràng rất ít xuất hiện trên Wikipedia trong các lĩnh vực phi thể thao, đặc biệt là kinh doanh và khoa học. Rõ ràng điều này có liên quan nhiều đến phân biệt chủng tộc. Nhưng có một hạt nhỏ, ở đó dân số năm 1950 là 84% da đen, đã tạo ra những người nổi tiếng với tỉ lệ gần với các hạt cao nhất.

Trong gần 13,000 người thế hệ Baby Boomer sinh ở hạt Macon, Alabama, có 15 người xuất hiện trong Wikipedia—tức là 1 trong 852 người. Tất cả đều là người da đen. Trong đó, 14 người từ thị trấn Tuskegee, nơi có Đại học Tuskegee, một đại học da đen lâu đời do Booker T. Washington thành lập. Danh sách gồm có thẩm phán, nhà văn, và nhà khoa học. Thực vậy, một trẻ da đen sinh ở Tuskegee có cùng xác suất trở thành người nổi tiếng trong một lĩnh vực ngoài thể thao ngang bằng một trẻ da trắng sinh ở một số thành phố đại học đa số người da trắng có điểm số cao nhất trong nghiên cứu này.

Thuộc tính thứ hai rất có khả năng khiến người bản địa của một hạt thành công là sự hiện diện trong hạt đó một thành phố lớn. Được sinh ra ở hạt San Francisco, hạt Los Angeles, hoặc Thành phố New York sẽ cho bạn xác suất cao nhất góp mặt trong Wikipedia. (Tôi gom 5 hạt của Thành phố New York chung lại với nhau, bởi vì nhiều mục Wikipedia không xác định cụ thể khu vực nơi sinh.)

Các vùng thành thị thường có nhiều hình mẫu thành công. Để thấy giá trị của việc ở gần người thành công trong một nghề khi trẻ, hãy so sánh Thành phố New York, Boston, và Los Angeles. Trong 3 thành phố, New York có tỉ lệ nhà báo nổi tiếng cao nhất; Boston có tỉ lệ nhà khoa học nổi tiếng cao nhất; và Los Angeles có tỉ lệ diễn viên nổi tiếng cao nhất. Hãy nhớ, chúng ta đang nói về những người sinh ra ở đó, chứ không phải những người nơi khác chuyển đến. Điều này vẫn đúng ngay cả sau khi trừ đi những người có cha mẹ nổi tiếng trong lĩnh vực đó.

Các hạt vùng ngoại vi không chứa thành phố đại học lớn có thành tích kém hơn nhiều so với các hạt tương đương ở khu vực thành thị. Cha mẹ tôi, giống nhiều người khác cùng thế hệ bùng nổ, chuyển từ các vỉa hè đông đúc đến các đường phố có bóng cây—ở đây là từ Manhattan đến hạt Bergen, New Jersey—để nuôi 3 đứa con. Đây là một sai lầm tiềm ẩn, ít nhất nhìn từ góc độ muốn có con nổi tiếng. Một trẻ sinh ở Thành phố New York có hơn 80% khả năng góp mặt trong Wikipedia so với một trẻ sinh ở hạt Bergen. Đây chỉ là các mối tương quan, nhưng nó cũng chỉ rõ rằng lớn lên gần nơi có các ý tưởng lớn thì tốt hơn lớn lên với một cái sân vườn lớn.

Các tác động nổi bật được xác định ở đây có thể còn mạnh hơn nữa nếu tôi có dữ liệu tốt hơn về nơi người ta đã sống qua thời thơ ấu, vì nhiều người lớn lên ở các hạt khác với hạt họ được sinh ra.

Sự thành công của các thành phố đại học và thành phố lớn rất ấn tượng khi bạn chỉ nhìn vào dữ liệu. Nhưng tôi cũng đào sâu thêm để thực hiện một phân tích thực nghiệm phức tạp hơn.

Làm như trên, tôi thấy rằng có một biến số khác cũng là một chỉ báo mạnh cho việc được vào Wikipedia: tỉ lệ người nhập cư ở hạt nơi sinh. Tỉ

lệ phần trăm cư dân sinh ở nước ngoài trong một vùng càng lớn, tỉ lệ trẻ sinh ra ở đó thành công nổi bật càng cao. (Nhớ lấy, ngài Donald Trump nhé!) Nếu hai nơi có dân số đại học và thành thị tương tự nhau, nơi có nhiều dân nhập cư hơn sẽ sản sinh nhiều người Mĩ nổi tiếng hơn. Cái gì giải thích điều này?

Phần lớn điều này có vẻ như trực tiếp là do con cái người nhập cư. Tôi đã tìm kiếm thấu đáo lí lịch của 100 người da trắng nổi tiếng nhất thế hệ Baby Boomer, theo dự án Pantheon của MIT, dự án này cũng đang nghiên cứu dữ liệu Wikipedia. Hầu hết những người này thuộc ngành giải trí. Ít nhất 13 người có mẹ sinh ở nước ngoài, bao gồm Oliver Stone, Sandra Bullock, và Julianne Moore. Tỉ lệ này gấp hơn 3 lần số trung bình cả nước trong thời kì đó. (Nhiều người có cha là dân nhập cư, bao gồm Steve Jobs và John Belushi, nhưng dữ liệu này khó so sánh với các số trung bình cả nước hơn, vì thông tin về những người cha không phải khi nào cũng có trên các giấy khai sinh.)

Còn các biến số không tác động đến thành công thì sao? Một biến số tôi thấy không bất ngờ cho lắm là số tiền mà bang chi cho giáo dục. Ở các tiểu bang có tỉ lệ cư dân sống ở thành thị tương tự nhau, việc chi cho giáo dục không tương quan với tỉ lệ tạo ra các tác gia, nghệ sĩ, hoặc doanh nhân nổi tiếng.

Thật thú vị khi so sánh nghiên cứu Wikipedia của tôi với một trong các nghiên cứu của nhóm Chetty đã thảo luận ở trước. Nhớ lại, nhóm của Chetty đã tìm hiểu những vùng nào có khả năng đưa người dân từ tầng lớp dưới lên tầng lớp thượng trung lưu. Nghiên cứu của tôi cố gắng tìm hiểu những vùng nào có khả năng giúp người dân đạt được danh tiếng. Kết quả khác nhau đến mức đáng kinh ngạc.

Chi nhiều cho giáo dục giúp trẻ với tới tầng lớp thượng trung lưu. Việc đó ít giúp họ trở thành tác gia, nghệ sĩ, hoặc doanh nhân nổi tiếng. Nhiều nhân vật thành công lớn trong số này ghét trường học. Một số thì bỏ học.

Nhóm Chetty phát hiện rằng Thành phố New York không phải là nơi đặc biệt tốt để nuôi trẻ nếu bạn muốn con mình đạt đến tầng lớp

thượng trung lưu. Nhưng tôi phát hiện, đó là một nơi tuyệt vời nếu bạn muốn con có cơ hội nổi tiếng.

Khi xem xét các yếu tố điều khiển sự thành công, sự khác biệt lớn giữa các hạt bắt đầu có ý nghĩa. Nhiều hạt có tất cả các thành phần chủ yếu cho sự thành công. Xin hãy trở lại với Boston. Với rất nhiều trường đại học, thành phố đang nung nấu vô số ý tưởng đổi mới. Đó là một vùng thành thị có nhiều người rất thành công, cho lớp trẻ những tấm gương thành đạt. Và nó thu hút nhiều dân nhập cư, con cái họ có động lực áp dụng những bài học này.

Còn nếu địa phương không có tính chất nào trong đó thì sao? Phải chăng nó được an bài là sẽ sản sinh ra ít siêu sao hơn? Không hẳn. Có một con đường khác: chuyên môn hóa tột bậc. Hạt Roseau, Minnesota, một hạt miền quê nhỏ có ít người nhập cư và không có trường đại học lớn, là một ví dụ rõ ràng. Khoảng 1 trong 740 người sinh ra ở đây góp mặt trong Wikipedia. Bí quyết của họ là gì? Cả 9 người là cầu thủ hockey chuyên nghiệp, rõ ràng được các chương trình hockey trường trung học và thanh niên đẳng cấp thế giới của hạt giúp đỡ.

Giả sử bạn không thích con mình thành ngôi sao hockey, phải chăng điểm chính ở đây là chuyển đến Boston hay Tuskegee nếu muốn con cái trong tương lai có lợi thế tối đa? Việc đó cũng chẳng hại gì. Nhưng ở đây còn có các bài học lớn hơn. Thông thường, các nhà kinh tế và xã hội học tập trung vào những phương thức tránh kết quả xấu, như sự nghèo khổ và tội phạm. Nhưng mục đích của một xã hội tốt đẹp không chỉ là hạn chế số người bị bỏ rơi lại đằng sau; mà còn là giúp càng nhiều người nổi bật càng tốt. Có lẽ nỗ lực phóng to những nơi mà hàng trăm ngàn người Mĩ nổi tiếng nhất được sinh ra có thể cho ta vài chiến lược ban đầu: khuyến khích nhập cư, trợ cấp các trường đại học, và hỗ trợ hoạt động nghệ thuật.

***

Thông thường, tôi nghiên cứu nước Mĩ. Vì vậy khi nghĩ về việc phóng to theo địa lí, tôi phóng to theo các thành thị—tôi xem xét những nơi như hạt Macon, Alabama, và hạt Roseau, Minnesota. Nhưng một lợi

thế lớn (và vẫn đang dần lớn hơn) khác của dữ liệu từ Internet là rất dễ thu thập dữ liệu trên khắp thế giới. Bấy giờ ta có thể thấy các nước khác nhau thế nào. Và các nhà khoa học dữ liệu đã có cơ hội lẻn vào ngành nhân chủng học.

Một chủ đề hơi ngẫu nhiên mà tôi khám phá gần đây: Quá trình mang thai diễn ra thế nào ở các nước khác nhau khắp thế giới? Tôi khảo sát các tìm kiếm Google của thai phụ. Điều đầu tiên tôi phát hiện là một sự tương tự đáng kinh ngạc về các triệu chứng cơ thể mà phụ nữ than phiền.

Tôi kiểm tra tần suất các triệu chứng khác nhau được tìm kiếm kết hợp với từ "mang thai." Ví dụ, bao nhiêu lần từ "mang thai" được tìm kiếm cùng với từ "buồn nôn," "đau lưng," hoặc "táo bón"? Các triệu chứng của Canada rất gần với các triệu chứng ở Mĩ. Các triệu chứng ở các nước như Anh, Úc, và Ấn Độ tất cả cũng gần như tương tự.

Thai phụ trên khắp thế giới rõ ràng cũng thèm các thứ giống nhau. Tại Mĩ, tìm kiếm Google hàng đầu về loại này là "thèm kem thời kì mang thai." 4 thứ tiếp theo là muối, kẹo, trái cây, và thức ăn nhiều gia vị. Tại Úc, những cơn thèm đó đều không khác gì nhiều: Danh sách đặc biệt có muối, kẹo, sô cô la, kem, và trái cây. Còn Ấn Độ thì sao? Câu chuyện cũng tương tự: thức ăn nhiều gia vị, kẹo, sô cô la, muối, và kem. Thực vậy, top 5 là rất giống nhau ở tất cả các nước mà tôi xem xét.

Bằng chứng sơ bộ cho thấy rằng không nơi nào trên thế giới có thực đơn hoặc môi trường làm thay đổi nghiêm trọng trải nghiệm cơ thể khi mang thai.

Nhưng những suy nghĩ xung quanh việc mang thai rõ ràng là rất khác biệt.

Bắt đầu với các câu hỏi về những gì thai phụ có thể làm một cách an toàn. Các câu hỏi hàng đầu ở Mĩ: Thai phụ có thể "ăn tôm," "uống rượu," "uống cà phê," hoặc "dùng thuốc Tylenol" không?

Khi nói về những lo lắng như thế, các nước khác không giống Mĩ hoặc không giống nhau lắm. Thai phụ có thể "uống rượu" hay không lại

nằm ngoài top 10 các câu hỏi ở Canada, Úc, hoặc Anh. Những lo lắng của Úc chủ yếu liên quan đến việc ăn các sản phẩm làm từ sữa trong khi mang thai, đặc biệt là phô mai kem. Tại Nigeria, nơi có 30% dân số dùng Internet, câu hỏi hàng đầu là phụ nữ mang thai có thể uống nước lạnh hay không.

Các lo lắng này có hợp lí không? Còn tùy. Có bằng chứng rõ ràng là thai phụ đang ngày càng có nguy cơ bị nhiễm một số vi khuẩn từ phô mai chưa tiệt trùng. Người ta đã thấy mối liên hệ giữa việc uống quá nhiều rượu và kết quả tiêu cực cho đứa bé. Ở một số nơi trên thế giới, người ta tin rằng uống nước lạnh có thể làm đứa bé viêm phổi; tôi không biết bằng chứng y học nào nói về điều này cả.

Những khác biệt lớn trong các câu hỏi đặt ra trên khắp thế giới rất có thể là do sự tràn ngập thông tin đến từ nhiều nguồn tạp nham tại mỗi nước: các nghiên cứu khoa học hợp chuẩn, các nghiên cứu khoa học tàm tạm, các chuyện mê tín, và chuyện ba láp ba xàm. Rất khó để phụ nữ biết nên tập trung vào nguồn nào—hoặc nên Google nguồn nào.

Ta có thể thấy một khác biệt rõ ràng khác khi xem các tìm kiếm hàng đầu theo cấu trúc "làm thế nào để ___ trong thời kì mang thai?" Tại Mĩ, Úc, và Canada, tìm kiếm hàng đầu là "làm thế nào để ngăn ngừa rạn da trong thời kì mang thai?" Nhưng tại Ghana, Ấn Độ, và Nigeria, việc ngăn ngừa rạn da thậm chí còn không thuộc top 5. Các nước này hay quan tâm nhiều hơn đến cách quan hệ tình dục hoặc cách ngủ.

Rõ ràng có nhiều điều để học hỏi từ việc phóng to các khía cạnh sức khỏe và văn hóa ở những nơi khác nhau trên thế giới. Phân tích ban đầu của tôi cho thấy Dữ Liệu Lớn cho ta biết con người ít mạnh mẽ hơn ta vẫn tưởng khi đối diện với sự thay đổi sinh học. Tuy nhiên, chúng ta thường đưa ra những lời giải thích thật khác biệt về tất cả các điều này.

**TOP 5 TÌM KIẾM "LÀM THẾ NÀO ĐỂ ___ TRONG THỜI KÌ MANG THAI"**

| MĨ | ẤN ĐỘ | ÚC | ANH | NIGERIA | NAM PHI |
|---|---|---|---|---|---|
| ngừa rạn da | ngủ | ngừa rạn da | giảm cân | quan hệ tình dục | quan hệ tình dục |
| giảm cân | quan hệ | giảm cân | ngừa rạn da | giảm cân | giảm cân |
| quan hệ tình dục | quan hệ tình dục | tránh rạn da | tránh rạn da | làm tình | ngừa rạn da |
| tránh rạn da | tình dục | ngủ | ngủ | khỏe mạnh | ngủ |
| giữ dáng | chăm sóc | quan hệ tình dục | quan hệ tình dục | hết nôn | hết nôn |

**TOP 5 TÌM BẮT ĐẦU VỚI "PHỤ NỮ MANG THAI CÓ THỂ ___ KHÔNG?"**

| | | | | | |
|---|---|---|---|---|---|
| **MĨ** | ăn tôm | uống rượu | uống cà phê | dùng thuốc Tylenol | ăn sushi |
| **ANH** | ăn tôm | ăn cá hồi hun khói | ăn bánh phô mai | ăn phô mai mozzaella | ăn mayonnaise |
| **ÚC** | ăn phô mai kem | ăn tôm | ăn thịt heo muối | ăn kem chua | ăn phô mai feta |
| **NIGERIA** | uống nước lạnh | uống rượu | uống cà phê | quan hệ tình dục | dùng cây chùm ngây (ăn được) |
| **SINGAPORE** | uống trà xanh | ăn kem lạnh | ăn sầu riêng | uống cà phê | ăn khóm |
| **TÂY BAN NHA** | ăn pa tê | ăn đùi heo muối | dùng thuốc paracetamol (giảm đau) | ăn cá ngừ | tắm nắng |
| **ĐỨC** | đi máy bay | ăn xúc xích salami | đi sauna | ăn mật ong | ăn phô mai mozzarella |
| **BRAZIL** | nhuộm tóc | dùng thuốc Dipirona (giảm đau) | dùng thuốc paracetamol | đi xe đạp | đi máy bay |

## Cách chúng ta trải qua từng phút từng giờ

"Những cuộc phiêu lưu của một thanh niên có các mối quan tâm chính là hãm hiếp, siêu bạo lực, và Beethoven."

Đó là cách bộ phim gây tranh cãi *A Clockwork Orange* của Stanley Kubrick được quảng cáo. Trong phim, nhân vật chính Alex DeLarge thực hiện các hành động bạo lực kinh hoàng mà mặt không biến sắc. Ở một trong những cảnh phim chấn động nhất, hắn hiếp một phụ nữ trong khi gào rống bản "Singin' in the Rain."

Gần như ngay lập tức, có báo cáo nhiều vụ bắt chước. Thực vậy, một nhóm người hãm hiếp một em gái 17 tuổi và hát chính bài đó. Bộ phim bị cấm chiếu ở nhiều nước châu Âu, và một số các cảnh kinh hoàng hơn được cắt bỏ đối với một phiên bản chiếu tại Mĩ.

Có nhiều ví dụ đời thực bắt chước phim ảnh, khi mà có những người dường như bị thôi miên bởi cái họ vừa xem trên màn hình. Bộ phim băng đảng *Colors* ra rạp và được nối theo sau bởi một vụ xả súng kinh hoàng. Bộ phim băng đảng *New Jack City* ra rạp và được tiếp nối bằng những vụ bạo loạn.

Có lẽ phiền phức nhất là vụ 4 ngày sau khi phát hành phim *The Money Train*, các ông quậy dùng xăng bật lửa đốt trạm thu phí xe điện ngầm, bắt chước y hệt một cảnh trong phim. Khác biệt duy nhất giữa vụ đốt nhà hư cấu và đời thực: Trong phim, người trực trạm chạy thoát. Ngoài đời thực, anh bị chết cháy.

Cũng có vài bằng chứng từ những thí nghiệm tâm lí cho thấy các chủ thể bị nhiễm phim bạo lực sẽ giận dữ và thù địch hơn, ngay cả nếu họ không bắt chước chính xác một cảnh phim.

Nói cách khác, các giai thoại và thí nghiệm cho thấy phim bạo lực có thể kích động hành vi bạo lực. Nhưng nó thực sự có ảnh hưởng lớn cỡ nào? Chúng ta đang nói về 1 hoặc 2 vụ giết người mỗi thập niên hay hàng trăm vụ mỗi năm? Các giai thoại và thí nghiệm không trả lời được câu hỏi này.

Để xem Dữ Liệu Lớn có trả lời được không, 2 nhà kinh tế học, Gordon Dahl và Stefano DellaVigna, đã kết hợp 3 bộ Dữ Liệu Lớn từ năm 1995 đến 2004: dữ liệu tội phạm hàng giờ của FBI, các con số phòng vé, và một thước đo độ bạo lực của mọi bộ phim từ kids-in-mind.com.

Thông tin họ đang dùng là đầy đủ—mọi bộ phim và mọi tội ác tính theo từng giờ tại các thành phố khắp nước Mĩ. Điều này rất quan trọng.

Mấu chốt trong nghiên cứu của họ: Một số ngày cuối tuần, phim đứng hạng nhất là phim bạo lực—*Hannibal* hoặc *Dawn of the Dead,* chẳng hạn. Một số ngày cuối tuần khác, phim số một lại là phim phi bạo lực, như *Runaway Bride* hoặc *Toy Story*.

Các nhà kinh tế học đó có thể thấy chính xác bao nhiêu vụ giết người, hãm hiếp, và hành hung xảy ra vào các ngày cuối tuần khi một phim bạo lực nổi tiếng được phát hành, đem so sánh với số vụ giết người, hãm hiếp, và hành hung vào các ngày cuối tuần khi một phim yên bình nổi tiếng được phát hành.

Vậy họ đã phát hiện điều gì? Khi phim bạo lực được chiếu, tội phạm có tăng như một số thí nghiệm chỉ ra không? Hay vẫn như cũ?

Vào các ngày cuối tuần có phim bạo lực nổi tiếng, hai nhà kinh tế học phát hiện: Tội phạm giảm mạnh.

Bạn không đọc nhầm đâu. Vào các ngày cuối tuần có phim bạo lực nổi tiếng, khi hàng triệu người Mĩ bị phơi nhiễm với các hình ảnh kẻ này giết kẻ kia, tội phạm giảm mạnh thấy rõ.

Khi nhận một kết quả kì lạ và bất ngờ thế này, suy nghĩ đầu tiên của bạn hẳn là có nhầm lẫn chỗ nào đó rồi. Mỗi tác giả đều cẩn thận xem lại công đoạn mã hóa. Không có sai sót. Tiếp theo bạn nghĩ là có một biến số nào khác sẽ giải thích các kết quả này. Họ kiểm tra xem thời gian trong năm có ảnh hưởng các kết quả hay không. Không. Họ thu thập dữ liệu về thời tiết, nghĩ rằng có lẽ yếu tố này đang tác động đến mối quan hệ kia bằng cách nào đó. Không phải luôn.

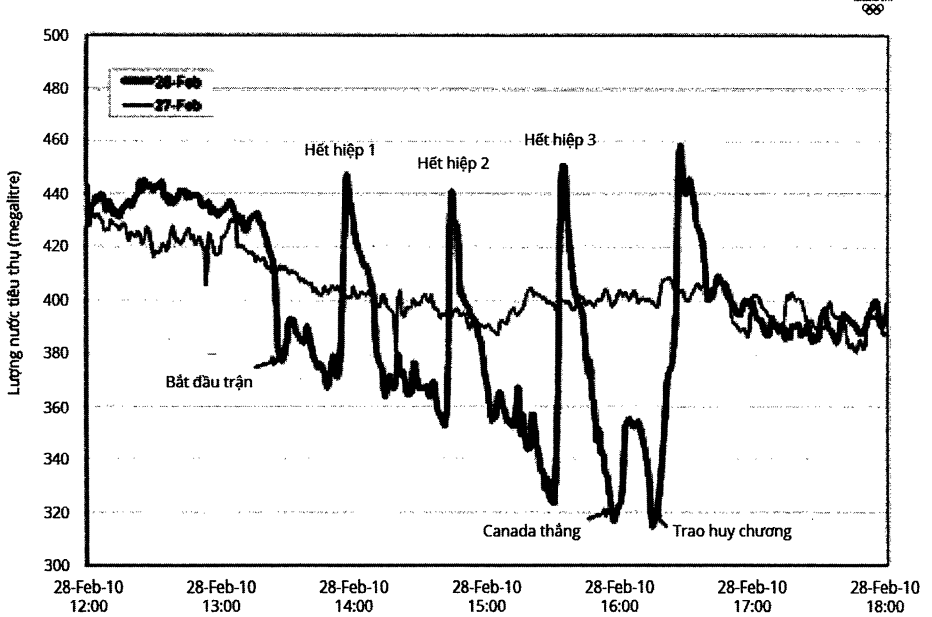
"Chúng tôi kiểm tra tất cả giả thiết, mọi thứ chúng tôi làm," Dahl nói với tôi. "Chúng tôi không phát hiện gì sai cả."

Bất chấp các giai thoại, bất chấp bằng chứng trong phòng thí nghiệm, và dù nghe có vẻ kì lạ, việc chiếu một phim bạo lực chẳng hiểu sao lại khiến tội phạm giảm mạnh. Làm sao có thể vậy được?

Chìa khóa giúp ta hiểu rõ vấn đề chính là sử dụng Dữ Liệu Lớn để phóng to hơn nữa. Dữ liệu khảo sát truyền thống cung cấp thông tin hàng năm, hoặc tốt lắm là hàng tháng. Nếu may mắn, ta có thể có dữ liệu của một kì nghỉ cuối tuần. Bằng cách so sánh—vì chúng ta dùng các bộ dữ liệu đầy đủ, chứ không phải các khảo sát mẫu đại diện nhỏ—chúng ta có thể tập trung vào từng giờ, thậm chí từng phút. Việc này cho phép ta hiểu hơn về hành vi con người.

Đôi khi những thay đổi thất thường theo thời gian thật thú vị, nếu không nói là cực kì bất ngờ. EPCOR, một công ty tiện ích ở Edmonton, Canada, báo cáo dữ liệu tiêu thụ nước từng phút trong trận đấu hockey giành huy chương vàng Olympic 2010 giữa Mĩ và Canada, trận đấu mà ước tính 80% người Canada đã xem. Dữ liệu cho ta biết rằng ngay sau mỗi hiệp đấu kết thúc, mức độ tiêu thụ nước tăng vọt. Các phòng vệ sinh khắp Edmonton rõ ràng đang tích cực dội nước.

**EPCOR** **Lượng nước tiêu thụ ở Edmonton trong trận chung kết Olympic**



Các tìm kiếm Google cũng có thể được phân tích theo từng phút, tiết lộ một số mô thức thú vị. Ví dụ, tìm kiếm "unblocked games" (game được gỡ chặn) tăng vọt lúc 8 giờ sáng các ngày cuối tuần và vẫn cao đến 3 giờ chiều. Điều này rõ ràng liên quan đến các hành vi phản ứng của học sinh khi nhà trường chặn truy cập game di động trong khuôn viên nhà trường nhưng không cấm học sinh dùng điện thoại di động.

Tỉ lệ tìm kiếm "thời tiết," "cầu nguyện," và "tin tức" lên cao nhất trước 5:30 sáng, bằng chứng cho thấy hầu hết mọi người thức giấc sớm hơn tôi nhiều. Tỉ lệ tìm kiếm "tự tử" lên cao nhất lúc 12:36 sáng và ở mức thấp nhất vào khoảng 9 giờ sáng, bằng chứng cho thấy hầu hết mọi người ít đau khổ vào buổi sáng hơn tôi nhiều.

Dữ liệu chỉ ra rằng khoảng 2 đến 4 giờ sáng là thời gian chủ yếu dành cho các câu hỏi vĩ đại: Ý thức nghĩa là gì? Tự do ý chí có tồn tại không? Có sự sống trên các hành tinh khác không? Sự phổ biến của các câu hỏi này vào lúc đêm khuya có thể một phần là kết quả của việc dùng cần sa. Tỉ lệ tìm kiếm "cách quấn một điếu" lên cao nhất vào khoảng 1-2 giờ sáng.

Và trong bộ dữ liệu lớn của mình, Dahl và DellaVigna có thể xem mức độ phạm tội thay đổi theo giờ như thế nào vào những ngày cuối tuần có chiếu phim. Họ phát hiện rằng tội phạm giảm mạnh khi các phim bạo lực nổi tiếng được chiếu—so sánh với những ngày cuối tuần khác—bắt đầu lúc trời vừa tối. Nói cách khác, tội phạm thấp hơn trước khi các cảnh bạo lực bắt đầu, lúc những người đi xem phim có thể vừa mới bước vào rạp.

Bạn có đoán được tại sao không? Đầu tiên, hãy nghĩ xem ai chắc chắn sẽ chọn xem một phim bạo lực. Đó là thanh niên—đặc biệt là các anh hung hăng.

Tiếp đến, thử nghĩ xem tội ác thường xảy ra ở đâu. Hiếm khi là trong rạp. Có những ngoại lệ, nổi bật nhất là vụ xả súng có tính toán năm 2012 ở một rạp tại Colorado. Nhưng, nói chung, nam giới đến rạp thường không có súng và ngồi yên lặng.

Cho các thanh niên hung hăng cơ hội xem phim kinh dị *Hannibal*, và họ sẽ đi xem phim. Cho các thanh niên hung hăng cơ hội xem phim *Runaway Bride* dành cho "bánh bèo," họ sẽ lượn khỏi rạp và quẹo vào quán bar, câu lạc bộ, hoặc đến tiệm bắn bi-da, nơi mà khả năng xảy ra bạo lực cao hơn nhiều.

Phim bạo lực giữ chân những người bạo lực tiềm năng không để họ ra đường.

Câu đố đã được giải đáp. Đúng không? Chưa đúng hẳn. Có một cái lạ nữa trong dữ liệu. Kết quả bắt đầu đúng khi các bộ phim bắt đầu chiếu; tuy nhiên, nó không dừng lại sau khi bộ phim kết thúc và rạp đã đóng cửa. Những tối chiếu phim bạo lực, tội phạm thấp hơn nhiều khi về khuya, từ nửa đêm đến 6 giờ sáng.

Ngay cả nếu mức độ phạm tội thấp hơn khi các thanh niên đó ở trong rạp xem phim, chẳng phải nó sẽ tăng lên sau khi họ ra về và không còn việc để làm nữa sao? Họ vừa xem một phim bạo lực, thứ mà các thí nghiệm nói làm cho người ta giận dữ và hung hăng hơn.

Bạn có thể nghĩ ra lời giải thích nào cho việc tại sao tội phạm vẫn giảm mạnh sau khi phim kết thúc không? Sau nhiều suy nghĩ, các tác giả, những chuyên gia tội phạm học, đã có thêm một phút giây bừng ngộ. Họ biết rằng rượu góp một phần lớn vào các vụ phạm tội. Các tác giả đã ngồi đủ nhiều các rạp chiếu phim để biết rằng hầu như không có rạp nào tại Mĩ bán rượu cả. Thực vậy, các tác giả thấy rằng tội phạm liên quan đến rượu tụt hẳn vào những giờ khuya sau khi chiếu các phim bạo lực.

Dĩ nhiên, kết quả của Dahl và DellaVigna vẫn có giới hạn. Ví dụ, họ không thể kiểm tra các ảnh hưởng kéo dài nhiều tháng sau—để xem sự sụt giảm mạnh hành vi phạm tội có thể kéo dài bao lâu. Sự phơi nhiễm thường xuyên với các phim bạo lực cuối cùng dẫn đến nhiều bạo lực hơn vẫn có thể xảy ra. Tuy nhiên, nghiên cứu của họ cũng đặt ảnh hưởng tức thời của phim bạo lực (đến nay là chủ đề chính của các thí nghiệm này) vào đúng góc nhìn. Có lẽ phim bạo lực thực sự ảnh hưởng một số người, khiến họ nóng giận và hung hăng khác thường. Tuy nhiên, có một thứ rõ

ràng là ảnh hưởng mọi người và khiến họ bạo lực hơn, bạn có biết đó là thứ gì không? Giao du với những người bạo lực tiềm năng khác. Và nhậu nhẹt.[1]

Bây giờ thì điều này cũng hợp lí. Nhưng trước khi Dahl và DellaVigna bắt đầu phân tích hàng đống dữ liệu, mọi thứ trông không hợp lí đến thế.

Một điểm quan trọng nữa trở nên rõ ràng khi ta phóng to dữ liệu: Thế giới rất phức tạp. Hành động ta làm hôm nay có thể rất lâu sau này mới gây ảnh hưởng, hầu hết không định trước được. Ý tưởng lan tỏa—khi thì chậm chạp; khi thì rất nhanh, như virus vậy. Người ta phản ứng với các động cơ theo những cách không thể đoán trước.

Các mối liên kết và quan hệ này, các vụ tăng vọt và tràn ngập này, là không thể truy tìm bằng những khảo sát nhỏ hoặc phương pháp dữ liệu truyền thống được. Nói đơn giản, thế giới quá phức tạp và quá phong phú để dữ liệu nhỏ làm nên chuyện.

## Những kẻ song trùng

Tháng 6/2009, David"Big Papi" Ortiz xem như xong.

Suốt nửa thập niên trước, Boston đã phải lòng danh thủ bóng chày sinh quán Dominican có nụ cười thân thiện với hàm răng hở này.

---

[1] Câu chuyện này chỉ ra những thứ tưởng là xấu lại có thể tốt như thế nào nếu chúng giúp ngăn ngừa cái xấu hơn. Ed McCaffrey, một cựu cầu thủ bóng bầu dục học trường Stanford, dùng lập luận này để biện hộ cho việc để cả 4 con trai của ông chơi bóng: "Bọn giặc này xung lắm. Và, vì vậy, nếu tụi nó mà không chơi bóng, tụi nó sẽ trượt ván, tụi nó sẽ leo trèo, tụi nó sẽ chơi đuổi bắt ngoài sân, tụi nó sẽ chơi bắn súng sơn. Tôi muốn nói là tụi nó sẽ không ngồi yên một đống đâu. Và, vì vậy, cách nhìn của tôi là, kệ đi, ít nhất trong môn thể thao bóng bầu dục cũng có luật lệ. [...] Các con tôi đã từng vào phòng cấp cứu vì ngã trên cao xuống, đụng xe đạp, trượt ván, leo cây ngã. Ý tôi là, đủ kiểu cả. [...] Ừ, đó là môn thể thao va chạm mạnh. Nhưng bọn giặc nhà tôi cũng sẽ có được cái tính là ít nhất không mê nhảy từ trên núi xuống hay làm mấy trò điên khùng như thế. Vậy, tôi nghĩ, đó là bạo lực có tổ chức." Lập luận của McCaffrey, trình bày trong một cuộc phỏng vấn trên *The Herd with Colin Cowherd*, là một lập luận tôi chưa hề nghe trước đó. Sau khi đọc bài báo Dahl/DellaVigna, tôi xem xét nghiêm túc lập luận này. Một lợi thế của các bộ dữ liệu lớn trong đời thực, chứ không phải dữ liệu phòng thí nghiệm, là chúng có thể giúp nhặt ra các ảnh hưởng loại này.

Anh đã đấu 5 trận All-Star[1] liên tục, đoạt một giải cầu thủ xuất sắc nhất, và giúp kết thúc đợt hạn hán chức vô địch kéo dài 86 năm của Boston. Nhưng vào mùa bóng 2008, ở tuổi 32, các chỉ số của slugger này sụt giảm. BA của anh đã rớt 68 điểm, OBP rớt 76 điểm, SLG rớt 114 điểm.[2] Và vào đầu mùa bóng 2009, các chỉ số của Ortiz giảm sâu hơn nữa.

Đây là cách Bill Simmons, một người chuyên viết về thể thao và rất hâm mộ đội Boston Red Sox, mô tả những gì diễn ra các tháng đầu mùa bóng 2009: "Rõ ràng David Ortiz không còn xuất sắc môn bóng chày nữa. [...] Các tay đánh bóng bự con giống như các sao khiêu dâm, các đô vật, các trung phong bóng rổ NBA và các bà vợ trẻ làm cảnh của đại gia: Hết thời là hết." Những người hâm mộ các môn thể thao tin tưởng vào mắt họ, và mắt của Simmons cho ông biết Ortiz đã hết thời. Thực vậy, Simmons dự báo Ortiz sẽ ngồi ghế dự bị hoặc bị thanh lí hợp đồng sớm.

Ortiz có thực sự hết thời không? Nếu bạn là người đứng đầu đội Boston năm 2009, bạn có loại anh ta không? Nói rộng hơn, làm sao dự báo được thành tích của một cầu thủ bóng chày trong tương lai? Nói rộng hơn nữa, làm sao dùng Dữ Liệu Lớn để dự báo mọi người sẽ làm gì trong tương lai?

Đây là một lí thuyết sẽ đưa bạn đi sâu vào khoa học dữ liệu: Hãy xem xét những gì các nhà thống kê bóng chày—sabermetrician (những người dùng dữ liệu để nghiên cứu bóng chày)—đã làm và chờ đợi nó lan tỏa đến các lĩnh vực khác của môn khoa học dữ liệu. Bóng chày nằm trong số các lĩnh vực đầu tiên có bộ dữ liệu đầy đủ chính xác về mọi thứ, và có cả một đội quân thiện chiến sẵn lòng hiến dâng đời mình để tìm hiểu dữ liệu đó. Bây giờ, hầu như mọi lĩnh vực đều có mặt khoa học dữ liệu hoặc đang hướng đến chuyện sử dụng vũ khí này. Bóng chày đến

---

[1] [ND] All-Star là trận đấu mang tính biểu diễn. Các cầu thủ xuất sắc nhất trong giải đấu sẽ được tuyển chọn lập thành 2 đội thi đấu giao hữu với nhau.

[2] [ND] Slugger: người đánh bóng mạnh (giúp cầu thủ bên tấn công chạy vượt nhiều base trong một cú đánh); BA: batting average; OBP: on-base percentage; SLG: slugging percentage. Đây là các chỉ số đánh giá khả năng tấn công của cầu thủ đánh bóng. Tại Việt Nam, bóng chày không phổ biến, nên người dịch sẽ để nguyên các thuật ngữ và chỉ giới thiệu tổng quát về các thuật ngữ này chứ không giải thích sâu.

trước; mọi lĩnh vực khác tiếp theo sau. Ngành thống kê bóng chày nuốt cả thế giới.

Cách đơn giản nhất để dự báo tương lai của một cầu thủ bóng chày là giả sử rằng anh ta sẽ tiếp tục chơi như hiện tại. Nếu một cầu thủ đã gặp nhiều khó khăn 1.5 năm qua, bạn có thể đoán rằng anh ta sẽ lại gặp nhiều khó khăn 1.5 năm tiếp theo.

Bằng phương pháp này, Boston lẽ ra nên loại David Ortiz.

Tuy nhiên, có thể còn nhiều thông tin phù hợp hơn. Thập niên 1980, Bill James, được xem là người sáng lập ngành thống kê bóng chày, nhấn mạnh tầm quan trọng của độ tuổi. James phát hiện rằng cầu thủ bóng chày đạt đỉnh cao sớm—khoảng độ tuổi 27. Các đội bóng thường không để ý các cầu thủ xuống năng lực bao nhiêu khi đã có tuổi. Họ còn trả quá nhiều tiền cho các cầu thủ lớn tuổi.

Bằng phương pháp tiến bộ hơn này, rõ ràng Boston lẽ ra vẫn nên loại bỏ David Ortiz.

Nhưng việc thêm vào biến độ tuổi này có thể vẫn có thiếu sót gì đó. Không phải tất cả cầu thủ đều đi theo một lối mòn suốt cả đời. Một số cầu thủ có thể đạt đỉnh cao lúc 23 tuổi, số khác thì 32 tuổi. Cầu thủ thấp có thể già theo kiểu khác với cầu thủ cao, cầu thủ mập khác với cầu thủ gầy. Các nhà thống kê bóng chày thấy rằng có nhiều loại cầu thủ, mỗi loại già theo một kiểu khác nhau. Câu chuyện này còn tệ hơn nữa với Ortiz: "Các cầu thủ đánh bóng bự con" trong thực tế nói chung thường đạt đỉnh cao sớm và sụp đổ ngay sau tuổi 30. Nếu Boston cân nhắc quá khứ gần đây, tuổi tác, và kích thước của David Ortiz, không cần phải phân vân, lẽ ra họ nên cắt hợp đồng luôn.

Sau đó, năm 2003, nhà thống kê Nate Silver giới thiệu một mô hình mới, ông gọi là PECOTA, để tiên đoán thành tích cầu thủ. Nó chứng tỏ là mô hình tốt nhất—và cũng ngầu nhất nữa. Silver tìm kiếm doppelganger, hay những kẻ song trùng—những người giống các cầu thủ hiện tại. Đây là cách làm: Xây dựng một cơ sở dữ liệu gồm mọi cầu thủ Major League Baseball, hơn 18,000 người. Cho vào đó mọi thứ bạn biết về các cầu thủ: chiều cao, tuổi, và vị trí; số lượt home run, các chỉ số

BA, walk, và strikeout cho mỗi năm của sự nghiệp. Bây giờ, tìm 20 cầu thủ trông giống Ortiz nhất cho đến thời điểm đó của anh—những người chơi như anh khi anh 24, 25, 26, 27, 28, 29, 30, 31, 32, và 33 tuổi. Nói cách khác, tìm những người giống Ortiz. Sau đó xem sự nghiệp những người này tiến triển thế nào.

Tìm các song trùng là một ví dụ khác của phương pháp phóng to. Nó phóng to lên nhóm nhỏ những người giống một đối tượng nào đó nhất. Và, như với mọi lần phóng to khác, ta có càng nhiều dữ liệu càng tốt. Hóa ra, các song trùng của Ortiz cho một dự đoán rất khác về tương lai của Ortiz. Những người giống Ortiz bao gồm Jorge Posada và Jim Thome. Các cầu thủ này bắt đầu sự nghiệp hơi chậm; có những đợt bộc phát đáng ngạc nhiên khi gần 30 tuổi, có sức mạnh đẳng cấp thế giới; và sau đó thì vất vả khi đã ngoài 30 tuổi.

Thế là Silver dự báo về Ortiz dựa trên cuộc đời các cầu thủ kia. Và đây là những gì ông phát hiện: Họ lấy lại sức mạnh. Đối với các bà vợ trẻ của đại gia, Simmons có thể đúng: Hết thời là hết. Nhưng với các cầu thủ giống Ortiz, thời của anh chỉ tạm ra đi để rồi lại trở về.

Phương pháp song trùng—phương pháp tốt nhất để dự báo thành tích cầu thủ bóng chày—nói Boston nên kiên nhẫn với Ortiz. Và Boston thực tế đã kiên nhẫn với anh cầu thủ có tuổi của họ. Năm 2010, điểm trung bình của Ortiz tăng lên .270. Anh đạt 32 lượt home run và lại vào đội All-Star. Đây là khởi đầu của loạt 4 trận All-Star liên tục. Năm 2013, đánh ở vị trí thứ 3 truyền thống trong đội hình, ở tuổi 37, hệ số của Ortiz là .688 khi Boston đánh bại St. Louis 4-2 trong trận World Series. Ortiz được bầu chọn cầu thủ xuất sắc nhất World Series.[1]

Ngay khi tôi đọc xong phương pháp Nate Silver dự báo đường đi của các cầu thủ, tôi bắt đầu nghĩ ngay đến việc có khi mình cũng có một song trùng.

---

[1] Đến phần này của sách, các bạn có thể cho rằng tôi hay có góc nhìn cay độc về các câu chuyện lạc quan. Tôi muốn đây là một câu chuyện lạc quan, vì vậy tôi sẽ để sự cay độc của mình xuống mục cước chú. Tôi nghi PECOTA đã phát hiện rằng Ortiz là một người đã từng dùng steroid và sẽ bắt đầu dùng lại. Từ góc độ của dự báo, thực sự cũng khá ngầu nếu PECOTA dò ra được điều đó—nhưng nó làm cho câu chuyện kém xúc động hơn.

Phương pháp song trùng rất hứa hẹn trong nhiều lĩnh vực, không chỉ thể thao. Tôi có thể tìm thấy người có chung nhiều mối quan tâm nhất với tôi không? Có thể nếu tôi tìm thấy người giống tôi nhất, chúng tôi có thể đi chơi với nhau. Có lẽ anh ta sẽ biết vài nhà hàng mà chúng tôi thích. Có thể anh ta sẽ giới thiệu tôi các thứ mà tôi không nghĩ là mình lại mê.

Phương pháp song trùng phóng to các cá thể và thậm chí cả những đặc điểm của cá thể nữa. Và, như với mọi thứ phóng to khác, càng nhiều dữ liệu thì càng sắc nét. Giả sử tôi tìm kiếm người giống mình trong một bộ dữ liệu khoảng 10 người. Tôi có thể tìm thấy ai đó cùng chung sở thích sách vở với tôi. Giả sử tôi tìm kiếm người giống mình trong một bộ dữ liệu khoảng 1,000 người. Tôi có thể tìm thấy ai đó rất mê sách vật lí. Nhưng giả sử tôi tìm kiếm người giống mình trong một bộ dữ liệu hàng trăm triệu người. Bấy giờ tôi có thể tìm được ai đó thực sự là giống mình.

Một ngày kia, tôi đi săn song trùng của mình trên mạng xã hội. Khi dùng toàn bộ hồ sơ Twitter, tôi tìm kiếm những người có chung các sở thích với tôi nhiều nhất.

Dĩ nhiên bạn có thể biết nhiều điều về các sở thích của tôi từ những người tôi theo dõi trên tài khoản Twitter. Nói chung, tôi theo dõi khoảng 250 người, thể hiện niềm say mê của tôi về thể thao, chính trị, phim hài, khoa học, và các ca sĩ dân ca Do Thái u buồn.

Vậy có ai trong vũ trụ ngoài kia là anh em sinh đôi trên Twitter của tôi, cũng theo dõi hết 250 tài khoản tôi đang theo dõi không? Dĩ nhiên là không. Những người giống ta không giống hệt ta, chỉ tương tự thôi. Cũng không có ai theo dõi trùng 200 tài khoản mà tôi đang theo. Thậm chí 150 cũng không.

Tuy nhiên, cuối cùng tôi cũng tìm thấy một tài khoản đáng kinh ngạc, theo dõi trùng 100 tài khoản với tôi: Country Music Radio Today. Hả? Thì ra, Country Music Radio Today là một con bot (nó không còn tồn tại nữa). Nó tự động theo dõi 750,000 hồ sơ Twitter, hi vọng rằng họ sẽ đáp lễ mà theo dõi lại nó.

Tôi có một cô bạn gái cũ mà tôi nghi là sẽ rất khoái cái kết quả này. Có lần cô ấy bảo tôi rằng tôi giống một con robot hơn là một con người.

Thôi, dẹp hết chuyện đùa sang một bên, phát hiện ban đầu của tôi rằng người giống tôi là một con bot theo dõi 750,000 tài khoản ngẫu nhiên đã chỉ ra một điểm rất quan trọng trong phương pháp song trùng. Để phương pháp này thực sự chính xác, bạn không muốn tìm người chỉ đơn thuần là thích các thứ giống bạn. Bạn còn muốn tìm người ghét những thứ bạn ghét nữa.

Các sở thích của tôi rõ ràng thể hiện không chỉ qua các tài khoản tôi theo dõi, mà còn qua các tài khoản tôi chọn không theo dõi nữa. Tôi thích thể thao, chính trị, phim hài, và khoa học nhưng không thích ẩm thực, thời trang, và ca kịch. Các theo dõi của tôi chỉ ra rằng tôi thích Bernie Sanders nhưng không thích Elizabeth Warren, thích Sarah Silverman nhưng không thích Amy Schumer, thích *New Yorker* nhưng không thích *Atlantic,* thích các bạn tôi Noah Popp, Emily Sands, và Josh Gottlieb nhưng không thích ông bạn Sam Asher. (Xin lỗi Sam nhé. Nhưng tôi tạm chặn ông trên Twitter feed rồi.)

Trong tất cả 200 triệu người trên Twitter, ai có hồ sơ giống tôi nhất? Hóa ra song trùng của tôi là nhà báo Dylan Matthews của trang *Vox*. Đây là một nỗi thất vọng của tôi, vì tôi làm trò này nhằm tìm thêm các nguồn bài viết mới để đọc. Tôi đã theo dõi Matthews trên Twitter và Facebook lâu rồi, và luôn đọc các bài của anh đăng trên *Vox*. Vậy biết rằng anh ấy giống tôi chưa thực sự giúp thay đổi cuộc đời tôi. Thế nhưng, vẫn khá vui khi biết người giống ta nhất trên thế giới, đặc biệt nếu đó là người mà ta ngưỡng mộ. Rồi khi tôi hoàn tất quyển sách này và thôi làm ẩn sĩ, có thể Matthews cùng tôi sẽ giao du và bàn luận về các bài viết của James Surowiecki.

Tìm kiếm người giống Ortiz chỉ dành cho giới hâm mộ bóng chày. Và tìm kiếm người giống tôi thì thật thú vị, ít nhất với tôi. Nhưng các tìm kiếm này có thể tiết lộ điều gì khác nữa? Một ví dụ nhé, phương pháp song trùng đã được nhiều công ty Internet lớn dùng để cải thiện các đề xuất và trải nghiệm người dùng. Amazon dùng một thứ tương tự

phương pháp song trùng để đề xuất sách bạn có thể thích. Họ xem những người giống bạn chọn gì và dựa vào đó mà đề nghị.

Pandora cũng làm giống như vậy trong việc chọn bài hát mà có thể bạn muốn nghe. Và đây cũng là cách Netflix tìm ra các bộ phim mà bạn có thể thích. Tác động của phương pháp này sâu sắc đến nỗi khi kĩ sư Greg Linden của Amazon lần đầu giới thiệu các tìm kiếm song trùng để dự báo loại sách ưa thích của người đọc, các đề xuất được cải tiến tốt đến nỗi người sáng lập Amazon—Jeff Bezos—phải quỳ xuống trước mặt Linden mà la lớn, "Tại hạ xin bái phục!"

Nhưng điều thực sự thú vị về phương pháp song trùng cực khủng không phải là cách nó thường được dùng bây giờ. Điều thú vị nằm ở những chỗ nó không được ứng dụng. Có các lĩnh vực lớn trong cuộc sống có thể được cải thiện nhiều bởi kiểu cá thể hóa dựa trên các tìm kiếm song trùng. Lấy chuyện sức khỏe làm ví dụ.

Isaac Kohane, một nhà khoa học máy tính kiêm nghiên cứu y khoa tại Harvard, đang thử đưa nguyên tắc này vào y học. Ông muốn tổ chức và thu thập tất cả thông tin sức khỏe của chúng ta để, thay vì dùng phương pháp một-kiểu-cho-tất-cả, các bác sĩ có thể tìm các bệnh nhân giống như bạn. Sau đó họ có thể áp dụng các chẩn đoán và cách điều trị tập trung hơn.

Kohane xem đây là một sự mở rộng tự nhiên cho lĩnh vực y học và thậm chí không phải là điều gì căn cơ cả. "Chẩn đoán là gì?" Kohane hỏi. "Chẩn đoán thực ra là một phát biểu nói rằng bạn có chung những đặc tính với các quần thể được nghiên cứu trước đó. Khi tôi chẩn đoán bạn bị nhồi máu cơ tim, xin thứ lỗi, điều đó có nghĩa là tôi nói bạn có tình trạng sinh lí bệnh—mà tôi biết từ những ca bệnh khác—cho thấy là bạn đã bị nhồi máu cơ tim."

Một chẩn đoán, tự bản chất, là một loại tìm kiếm song trùng. Vấn đề là các bộ dữ liệu mà bác sĩ dùng để làm các chẩn đoán thì nhỏ. Ngày nay, chẩn đoán thường dựa trên kinh nghiệm của một bác sĩ với nhóm bệnh nhân mà ông ta đã điều trị, và có lẽ được bổ sung bởi các bài báo hàn lâm từ những nhóm nhỏ mà các nhà nghiên cứu khác đã gặp. Tuy nhiên,

như chúng ta đã thấy, để tìm kiếm bệnh nhân tương tự thực sự hiệu quả, nó sẽ phải bao gồm nhiều trường hợp hơn nữa.

Đây là lĩnh vực mà Dữ Liệu Lớn thực sự có ích. Vậy điều gì khiến nó chậm chạp như thế? Tại sao nó chưa được dùng rộng rãi? Vấn đề nằm ở chỗ thu thập dữ liệu. Hầu hết các báo cáo y khoa vẫn tồn tại trên giấy, vùi trong các chồng hồ sơ, và các báo cáo được vi tính hóa lại thường bị khóa trong các định dạng không tương thích. Kohane nhận xét rằng dữ liệu bóng chày tốt hơn dữ liệu y tế. Nhưng các biện pháp đơn giản vẫn thường được việc. Kohane hay nói đi nói lại chuyện nhắm vào thứ dễ làm và hiệu quả trước. Ví dụ, ông tin rằng đơn thuần tạo ra một bộ dữ liệu đầy đủ biểu đồ chiều cao và cân nặng của trẻ và bất cứ bệnh gì trẻ có thể mắc phải sẽ cách mạng hóa nhi khoa. Đường tăng trưởng mỗi đứa trẻ bấy giờ có thể được so sánh với đường tăng trưởng của mọi đứa trẻ khác. Máy tính có thể tìm thấy những đứa trẻ có quỹ đạo phát triển tương tự và tự động đánh dấu các kiểu hình hay có vấn đề. Nó có thể phát hiện chiều cao một đứa trẻ ngừng tăng trưởng sớm, mà trong một số trường hợp chắc chắn là do 1 trong 2 nguyên nhân: giảm hoạt động tuyến giáp hoặc u não. Chẩn đoán sớm trong cả 2 trường hợp đều rất có ích. "Đây là những ca hiếm hoi," theo Kohane, "1,000 ca mới có 1. Trẻ em, nói chung, đều khỏe mạnh. Tôi nghĩ chúng ta có thể chẩn đoán sớm hơn, ít nhất là sớm hơn 1 năm. 100% là có thể."

Doanh nhân James Heywood có cách tiếp cận khác để xử lí các khó khăn liên quan đến dữ liệu y khoa. Ông tạo website PatientsLikeMe.com để các cá nhân có thể khai báo thông tin của chính mình—tình trạng, cách điều trị, và các tác dụng phụ. Ông đã có nhiều thành công trong việc lập biểu đồ các diễn biến khác nhau của bệnh và so sánh với hiểu biết thông thường của chúng ta về các bệnh này.

Mục tiêu của ông là tuyển đủ người, phủ hết các tình trạng bệnh, để ai cũng có thể tìm thấy người có sức khỏe giống mình. Heywood hi vọng rằng bạn có thể tìm thấy những người cùng độ tuổi và giới tính, cùng lịch sử bệnh lí, có các triệu chứng tương tự như bạn—và thấy cách chữa nào đã hiệu quả với họ. Thực vậy, đó sẽ là một phương pháp y khoa rất khác biệt.

## Các câu chuyện dữ liệu

Trên nhiều phương diện, phương pháp phóng to có giá trị đối với tôi hơn các phát hiện đặc biệt của một nghiên cứu đặc biệt nào đó, vì nó cung cấp một cách mới để nhìn và nói về cuộc sống.

Khi mọi người biết tôi là nhà khoa học dữ liệu và viết sách, đôi khi họ chia sẻ một số thông tin hoặc khảo sát nào đó với tôi. Tôi thường thấy các dữ liệu này chán ngắt—tĩnh và tẻ nhạt. Nó không có chuyện để kể.

Tương tự, bạn bè đã cố gắng lôi kéo tôi tham gia đọc tiểu thuyết và tiểu sử cùng họ. Nhưng các thứ này cũng không làm tôi quan tâm. Tôi luôn luôn tự đặt cho mình câu hỏi, "Điều đó có xảy ra trong các tình huống khác không? Nguyên tắc phổ quát hơn là gì?" Các câu chuyện của họ có vẻ nhỏ và không mang tính đại diện.

Những gì tôi muốn trình bày trong sách này là một điều hoàn toàn khác biệt đối với tôi. Nó dựa trên dữ liệu và các con số; nó mang tính minh họa và ảnh hưởng sâu rộng. Tuy thế, dữ liệu lại phong phú đến nỗi bạn có thể hình dung những người đang ẩn mình dưới đó. Khi phóng to mỗi phút tiêu thụ nước của Edmonton, tôi *thấy* người ta đứng dậy khỏi ghế vào cuối hiệp đấu. Khi phóng to những người chuyển từ Philadelphia đến Miami và bắt đầu gian lận thuế, tôi *thấy* những người này đang nói chuyện với hàng xóm ở khu chung cư và biết về trò gian lận thuế. Khi phóng to người hâm mộ bóng chày mọi lứa tuổi, tôi *thấy* tuổi thơ của chính tôi và của em trai tôi, cùng hàng triệu người trưởng thành vẫn đang gào thét bên một đội được họ ủng hộ từ hồi 8 tuổi.

Một lần nữa, tôi sẽ nói chuyện nghe có vẻ phô trương một chút: Tôi nghĩ các nhà kinh tế học và khoa học dữ liệu có mặt trong quyển sách này đang tạo ra không chỉ một công cụ mới, mà còn một thể loại hoàn toàn mới. Điều tôi đã cố trình bày trong chương này, và trong hầu hết quyển sách này, là dữ liệu rất lớn và rất phong phú, cho phép ta phóng to đến độ, tuy không tập trung vào bất cứ cá nhân cụ thể không mang tính đại diện nào, ta vẫn có thể kể nên các câu chuyện phức tạp và đầy sự liên tưởng.

# CHƯƠNG 6

## Cả thế giới là một phòng thí nghiệm

Ngày 27/2/2000 bắt đầu như một ngày bình thường tại khuôn viên Mountain View của Google. Mặt trời chiếu sáng, một số người đạp xe, kĩ thuật viên massage đang làm việc, các nhân viên thì đang uống nước dưa leo. Và sau đó, vào cái ngày bình thường này, vài kĩ sư Google có một ý tưởng đã khai mở một bí mật mà ngày nay điều khiển đa phần Internet. Các kĩ sư đã phát hiện cách tốt nhất để khiến bạn nhấp chuột, quay lại, và ở lại trên các trang của họ.

Trước khi mô tả những gì họ làm, chúng ta cần bàn về sự tương quan và quan hệ nhân quả, một vấn đề lớn trong phân tích dữ liệu—và là một vấn đề mà chúng ta chưa chú tâm đúng mức.

Hầu như mỗi ngày các phương tiện truyền thông oanh tạc chúng ta bằng những nghiên cứu dựa trên cơ sở tương quan. Ví dụ, ta được bảo rằng ai mà uống một lượng rượu điều độ thì thường có sức khỏe tốt hơn. Đó là một sự tương quan.

Điều này có nghĩa là uống một lượng rượu điều độ sẽ cải thiện sức khỏe—một quan hệ *nhân quả* chăng? Chưa chắc. Cũng có thể là sức khỏe tốt khiến người ta uống một lượng rượu điều độ. Các nhà khoa học xã hội gọi đây là *nhân quả đảo ngược*. Hoặc cũng có thể là tồn tại một nhân tố độc lập dẫn đến cả việc uống rượu điều độ lẫn sức khỏe tốt. Có lẽ bỏ nhiều thời gian với bạn bè dẫn đến cả việc tiêu thụ rượu điều độ lẫn sức

khỏe tốt. Các nhà khoa học xã hội gọi đây là *thiên kiến bỏ sót biến* (omitted-variable bias).

Vậy thì, chúng ta có thể thiết lập quan hệ nhân quả chính xác hơn thế nào? Tiêu chuẩn vàng là phải có một thí nghiệm chọn mẫu ngẫu nhiên, có đối chứng. Đây là cách làm: Bạn phân chia ngẫu nhiên người ta thành 2 nhóm. Nhóm thực nghiệm (treatment group) được yêu cầu làm hoặc nhận cái gì đó. Nhóm đối chứng (control group) thì không làm gì cả. Sau đó bạn xem mỗi nhóm phản hồi thế nào. Sự khác biệt đầu ra giữa 2 nhóm thể hiện tác động nhân quả.[1]

Ví dụ, để kiểm tra liệu uống rượu điều độ có tốt cho sức khỏe không, bạn chọn ngẫu nhiên một số người uống 1 li vang mỗi ngày trong 1 năm, chọn ngẫu nhiên những người khác không uống rượu trong 1 năm, và sau đó so sánh sức khỏe được báo cáo từ cả 2 nhóm. Vì 2 nhóm được chọn ngẫu nhiên, không có lí do gì để nghĩ 1 nhóm có sức khỏe ban đầu tốt hơn hoặc tham gia hoạt động xã hội nhiều hơn. Bạn có thể tin rằng các tác động của rượu vang là nguyên nhân. Các thí nghiệm chọn mẫu ngẫu nhiên có đối chứng là bằng chứng đáng tin cậy nhất trong mọi lĩnh vực. Nếu một viên thuốc qua được một cuộc thí nghiệm chọn mẫu ngẫu nhiên có đối chứng, nó có thể được phân phối cho công chúng. Nếu không qua được cuộc kiểm tra này, nó sẽ không được lên quầy thuốc.

Thí nghiệm chọn mẫu ngẫu nhiên cũng ngày càng được dùng nhiều trong các ngành khoa học xã hội. Esther Duflo, một nhà kinh tế học người Pháp tại MIT, đã lãnh đạo chiến dịch gia tăng mức độ ứng dụng nhiều thí nghiệm trong kinh tế học phát triển, một lĩnh vực cố gắng tìm ra cách tốt nhất để giúp người nghèo trên thế giới. Ta hãy cùng xem xét nghiên cứu của Duflo cùng các đồng nghiệp về cách cải thiện giáo dục vùng thôn quê Ấn Độ, nơi hơn một nửa học sinh cấp 2 không đọc được một câu đơn giản. Một lí do tiềm năng khiến học sinh gặp khó khăn như thế là vì giáo viên không có mặt đều đặn. Tại một số trường ở thôn quê Ấn Độ, có ngày có hơn 40% giáo viên vắng mặt.

---

[1] [ND] Một số nơi còn gọi treatment group là nhóm điều trị (đặc biệt trong y khoa), control group là nhóm kiểm soát.

Thí nghiệm của Duflo là gì? Cô và các đồng nghiệp chia ngẫu nhiên các trường thành 2 nhóm. Trong nhóm thực nghiệm, ngoài tiền lương cơ bản, giáo viên được trả một số tiền nhỏ—50 rupee (khoảng 1.15 USD)—cho mỗi ngày họ có mặt để dạy. Còn nhóm kia (nhóm đối chứng) thì không được trả thêm tiền có mặt. Kết quả thật đáng chú ý. Khi giáo viên được trả tiền, mức độ vắng mặt giảm xuống một nửa. Thành tích kiểm tra của học sinh cũng cơ bản tiến bộ, kết quả tốt nhất là ở các em nữ nhỏ tuổi. Đến cuối cuộc thí nghiệm, các em nữ trong những trường mà giáo viên được trả tiền đến lớp tăng thêm 7 %p khả năng biết viết.

Theo một bài báo trên *New Yorker,* khi Bill Gates biết đến công trình của Duflo, ông rất ấn tượng và đã nói, "Chúng tôi *cần* cấp kinh phí cho cô."

## Kiến thức căn bản về thử nghiệm A/B

Vậy thí nghiệm chọn mẫu ngẫu nhiên có đối chứng là tiêu chuẩn vàng để chứng minh mối quan hệ nhân quả, và việc sử dụng thí nghiệm loại này đã lan khắp các ngành khoa học xã hội. Điều đó đưa chúng ta trở lại các văn phòng của Google ngày 27/2/2000. Google làm gì vào ngày đó mà đã cách mạng hóa Internet vậy?

Hôm ấy, có vài kĩ sư quyết định thực hiện một thí nghiệm trên trang của Google. Họ phân chia ngẫu nhiên người dùng thành 2 nhóm. Nhóm thực nghiệm được cho xem 20 kết quả trên mỗi trang tìm kiếm. Nhóm đối chứng được cho xem 10 kết quả như thông thường. Các kĩ sư bấy giờ so sánh sự hài lòng của 2 nhóm dựa trên số lần họ quay lại Google.

Đây là một cuộc cách mạng à? Không có vẻ gì cách mạng lắm. Tôi đã thấy các thí nghiệm chọn mẫu ngẫu nhiên được các công ty dược và các nhà khoa học xã hội sử dụng lâu rồi. Làm thế nào mà việc sao chép thí nghiệm loại này lại thay đổi thế giới chứ?

Điểm quan trọng—và điều này được các kĩ sư Google nhanh chóng nhận ra—là các thí nghiệm trong thế giới kĩ thuật số có một thuận lợi lớn so với các thí nghiệm trong thế giới ngoại tuyến. Dù có thể đáng tin cậy, các thí nghiệm chọn mẫu ngẫu nhiên ngoài đời thực thường đòi hỏi

nhiều nguồn lực. Với nghiên cứu của Duflo, ta phải tiếp xúc các trường, phải chuẩn bị kinh phí, phải trả tiền một số giáo viên, và phải kiểm tra tất cả học sinh. Thí nghiệm ngoại tuyến có thể tốn hàng ngàn hoặc hàng trăm ngàn đô la và mất hàng tháng hoặc hàng năm để thực hiện.

Trong thế giới số, thí nghiệm chọn mẫu ngẫu nhiên có thể rẻ và nhanh. Bạn không cần tuyển và trả tiền cho người tham gia. Thay vì thế, bạn có thể viết một dòng mã để ngẫu nhiên đưa họ vào một nhóm. Bạn không cần người dùng điền vào bộ câu hỏi khảo sát. Thay vì thế, bạn có thể đo lường các chuyển động và các cú nhấp chuột. Bạn không cần mã hóa thủ công và phân tích các phản hồi. Bạn có thể xây dựng một chương trình tự động làm việc đó cho bạn. Bạn không cần phải tiếp xúc bất cứ ai. Thậm chí bạn không cần phải bảo người dùng là họ đang tham gia thí nghiệm nữa.

Đây là sức mạnh thứ tư của Dữ Liệu Lớn: Nó khiến cho thí nghiệm chọn mẫu ngẫu nhiên—thứ có thể phát hiện các tác động nhân quả thực sự—dễ thực hiện hơn rất, rất nhiều: bất cứ khi nào, gần như bất cứ nơi đâu, miễn là trên mạng. Trong thời đại Dữ Liệu Lớn, cả thế giới là một phòng thí nghiệm.

Hiểu biết này nhanh chóng lan tỏa khắp Google và sau đó là phần còn lại của Thung lũng Silicon, ở đó thí nghiệm chọn mẫu ngẫu nhiên có đối chứng đã được đặt lại một cái tên mới: "thử nghiệm A/B" (A/B test). Năm 2011, các kĩ sư Google chạy 7 ngàn thử nghiệm A/B. Và con số này ngày càng gia tăng.

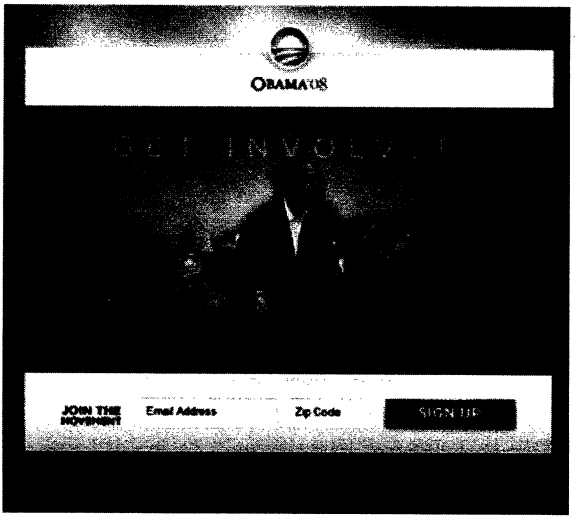
Nếu Google muốn tìm hiểu làm thế nào để tăng số người nhấp chuột lên quảng cáo trên các trang của họ, họ có thể thử 2 sắc độ xanh dương trong quảng cáo—một sắc độ cho Nhóm A, một cho Nhóm B. Google bấy giờ có thể so sánh tỉ lệ nhấp chuột. Dĩ nhiên, thử nghiệm quá dễ dàng có thể khiến ta lạm dụng. Một số nhân viên cảm thấy rằng vì thực hiện quá dễ nên Google đã thử nghiệm quá nhiều. Năm 2009, một nhà thiết kế chán nản bỏ việc sau khi Google khảo sát 41 sắc độ xanh dương chỉ khác nhau chút ít trong các thử nghiệm A/B. Tuy vậy, lập trường ủng hộ nghệ thuật thay vì nghiên cứu thị trường đầy ám ảnh của nhà thiết kế

này hầu như không ngăn chặn được sự lan tỏa của phương pháp thử nghiệm A/B.

Facebook bây giờ chạy 1 ngàn thử nghiệm A/B mỗi ngày, có nghĩa là một nhóm nhỏ kĩ sư tại Facebook thực hiện thí nghiệm chọn mẫu ngẫu nhiên có kiểm soát trong 1 ngày nhiều hơn toàn ngành dược thực hiện trong 1 năm.

Thử nghiệm A/B đã lan ra khỏi phạm vi các công ty công nghệ lớn nhất. Một cựu nhân viên Google, Dan Siroker, đã mang phương pháp này đến cuộc vận động tranh cử tổng thống đầu tiên của Obama—họ thử nghiệm A/B các mẫu trang chủ, lời chào mời qua email, và phiếu quyên góp. Sau đó, Siroker mở một công ty mới, Optimizely, cho phép các tổ chức thực hiện thử nghiệm A/B nhanh. Năm 2012, Optimizely được Obama cũng như đối thủ của ông, Mitt Romney, sử dụng để tối đa hóa số người đăng kí, số tình nguyện viên, và các khoản đóng góp. Nó cũng được dùng bởi đủ loại công ty đa dạng như Netflix, TaskRabbit, và tạp chí *New York*.

Để thấy việc thử nghiệm có giá trị thế nào, hãy xem cách Obama dùng nó để lôi kéo nhiều người hơn vào cuộc vận động của ông. Trang chủ của Obama ban đầu bao gồm một hình ảnh ứng cử viên và một nút bấm dưới hình mời mọi người "Sign Up."



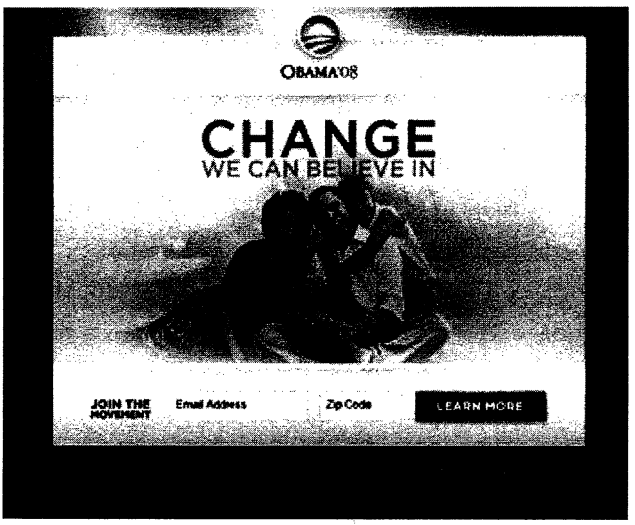
Đây có phải cách tốt nhất để chào mời mọi người không? Với sự giúp đỡ của Siroker, nhóm của Obama thử một hình ảnh và nút bấm khác xem có thể khiến nhiều người đăng kí hơn hay không. Liệu sẽ có nhiều người hơn nhấp chuột nếu trang chủ đổi lại có hình Obama với vẻ mặt nghiêm nghị hơn hay không? Liệu sẽ có nhiều người hơn nhấp chuột nếu nút bấm đổi lại là "Join Now" hay không? Nhóm của Obama cho người dùng xem các phối hợp hình ảnh và nút bấm khác nhau rồi đo lường số người nhấp chuột. Thử xem bạn có dự đoán được hình nào và nút nào thắng không nhé.

THỬ NGHIỆM HÌNH



THỬ NGHIỆM NÚT


Hình thắng cuộc là hình gia đình Obama và nút "Learn More." Và thắng rất đậm. Bằng cách dùng phối hợp đó, nhóm vận động của Obama ước tính nó lôi kéo thêm 40% số người đăng kí, thu được khoảng 60 triệu USD kinh phí bổ sung.




Có một lợi ích lớn khác nữa: các thử nghiệm tiêu chuẩn vàng này có thể được thực hiện rất rẻ và dễ dàng: Nó giải phóng ta khỏi trực giác, thứ mà vẫn tồn tại những hạn chế. Một lí do cơ bản khiến thử nghiệm A/B quan trọng: không thể đoán trước được mọi người. Trực giác thường không dự báo được mọi người sẽ phản ứng ra sao.

Trực giác của bạn có đúng về website tối ưu của Obama không?

Dưới đây là một số thử nghiệm nữa cho trực giác của bạn. Trang *Boston Globe* thử nghiệm A/B các tiêu đề để tìm ra tiêu đề nào khiến nhiều người nhấp chuột lên câu chuyện nhất. Thử đoán các tiêu đề thắng cuộc trong các cặp này nhé:

**Ở MỖI CẶP, CÓ 1 TIÊU ĐỀ TỐT HƠN HẲN TIÊU ĐỀ CÒN LẠI TRONG VIỆC THU HÚT NHẤP CHUỘT:**

| | TIÊU ĐỀ A | TIÊU ĐỀ B |
|---|---|---|
| 1 | Can the SnotBot drone save the whales?<br>Máy bay không người lái SnotBot có cứu được cá voi không? | Can this drone help save the whales?<br>Máy bay không người lái này có cứu được cá voi không? |
| 2 | Of course "deflated balls" is a top search term in Massachusetts<br>Dĩ nhiên "những quả bóng xì hơi" là từ tìm kiếm hàng đầu ở Massachusetts | This top Mass. Google search term is pretty embarrassing<br>Từ tìm kiếm Google nhiều nhất này ở Mass. khá là khó đỡ |
| 3 | Hookup contest at heart of St. Paul rape trial<br>Cuộc thi "yêu" là tâm điểm phiên tòa xử vụ hiếp dâm tại St. Paul | No charges in prep school sex scandal<br>Không có buộc tội trong vụ bê bối tình dục tại trường dự bị đại học |
| 4 | Woman makes bank off rare baseball card<br>Người phụ nữ kiếm cả ngân hàng tiền từ hình cầu thủ bóng chày quý hiếm | Woman makes $179,000 off rare baseball card<br>Người phụ nữ kiếm $179,000 từ hình cầu thủ bóng chày quý hiếm |
| 5 | MBTA projects annual operating deficit will double by 2020<br>Mức thâm hụt hoạt động hàng năm các dự án MBTA sẽ tăng gấp đôi vào năm 2020 | Get ready: the MBTA's deficit is about to double<br>Chuẩn bị đi: Mức thâm hụt của MBTA sắp tăng gấp đôi |
| 6 | How Massachusetts helped win you the right to birth control access<br>Massachusetts đã giúp bạn có được quyền kiểm soát sinh đẻ ra sao | How Boston University helped end "crimes against chastity"<br>Đại học Boston đã giúp chấm dứt "tội ác chống lại trinh tiết" ra sao |
| 7 | When the first subway opened in Boston<br>Khi tuyến tàu điện ngầm đầu tiên được khai trương ở Boston | Cartoons from when the first subway opened in Boston<br>Các phim hoạt hình từ khi tuyến tàu điện ngầm đầu tiên được khai trương ở Boston |
| 8 | Victim and family in prep-school rape trial blame toxic culture | Victim and family in prep-school rape trial releases statement |

| | | |
|---|---|---|
| | Nạn nhân và gia đình trong phiên tòa xử vụ hiếp dâm tại trường dự bị đại học đổ lỗi cho văn hóa độc hại | Nạn nhân và gia đình trong phiên tòa xử vụ hiếp dâm tại trường dự bị đại học đưa ra lời phát biểu |
| 9 | Guy in "Free Brady" hat is only one able to foil Miley Cyrus prank<br><br>Gã đội mũ "Free Brady" là người duy nhất có thể đánh bại trò chơi khăm Miley Cyrus | Pats fan gets an eyeful for recognizing an undercover Miley Cyrus<br><br>Người hâm mộ Pats trố mắt nhìn khi nhận ra Miley Cyrus giả dạng |

Bạn đoán xong chưa? Đáp án đúng được in đậm dưới đây.

| | TIÊU ĐỀ A | TIÊU ĐỀ B | KẾT QUẢ |
|---|---|---|---|
| 1 | **Can the SnotBot drone save the whales?**<br><br>**Máy bay không người lái SnotBot có cứu được cá voi không?** | Can this drone help save the whales?<br><br>Máy bay không người lái này có cứu được cá voi không? | Thêm 53% nhấp chuột cho A |
| 2 | Of course "deflated balls" is a top search term in Massachusetts<br><br>Dĩ nhiên "những quả bóng xì hơi" là từ tìm kiếm hàng đầu ở Massachusetts | **This top Mass. Google search term is pretty embarrassing**<br><br>**Từ tìm kiếm Google nhiều nhất này ở Mass. khá là khó đỡ** | Thêm 986% nhấp chuột cho B |
| 3 | Hookup contest at heart of St. Paul rape trial<br><br>Cuộc thi "yêu" là tâm điểm phiên tòa xử vụ hiếp dâm tại St. Paul | **No charges in prep school sex scandal**<br><br>**Không có buộc tội trong vụ bê bối tình dục tại trường dự bị đại học** | Thêm 108% nhấp chuột cho B |
| 4 | **Woman makes bank off rare baseball card**<br><br>**Người phụ nữ kiếm cả ngân hàng tiền từ hình cầu thủ bóng chày quý hiếm** | Woman makes $179,000 off rare baseball card<br><br>Người phụ nữ kiếm $179,000 từ hình cầu thủ bóng chày quý hiếm | Thêm 38% nhấp chuột cho A |
| 5 | MBTA projects annual operating deficit will double by 2020<br><br>Mức thâm hụt hoạt động hàng năm các dự án MBTA sẽ tăng | **Get ready: the MBTA's deficit is about to double**<br><br>**Chuẩn bị đi: Mức thâm hụt của MBTA sắp tăng gấp đôi** | Thêm 62% nhấp chuột cho B |

| | | | |
|---|---|---|---|
| | gấp đôi vào năm 2020 | | |
| 6 | How Massachusetts helped win you the right to birth control access<br><br>Massachusetts đã giúp bạn có được quyền kiểm soát sinh đẻ ra sao | **How Boston University helped end "crimes against chastity"**<br><br>**Đại học Boston đã giúp chấm dứt "tội ác chống lại trinh tiết" ra sao** | Thêm 188% nhấp chuột cho B |
| 7 | **When the first subway opened in Boston**<br><br>**Khi tuyến tàu điện ngầm đầu tiên được khai trương ở Boston** | Cartoons from when the first subway opened in Boston<br><br>Các phim hoạt hình từ khi tuyến tàu điện ngầm đầu tiên được khai trương ở Boston | Thêm 33% nhấp chuột cho A |
| 8 | Victim and family in prep-school rape trial blame toxic culture<br><br>Nạn nhân và gia đình trong phiên tòa xử vụ hiếp dâm tại trường dự bị đại học đổ lỗi cho văn hóa độc hại | **Victim and family in prep-school rape trial releases statement**<br><br>**Nạn nhân và gia đình trong phiên tòa xử vụ hiếp dâm tại trường dự bị đại học đưa ra lời phát biểu** | Thêm 76% nhấp chuột cho B |
| 9 | Guy in "Free Brady" hat is only one able to foil Miley Cyrus prank<br><br>Gã đội mũ "Free Brady" là người duy nhất có thể đánh bại trò chơi khăm Miley Cyrus | **Pats fan gets an eyeful for recognizing an undercover Miley Cyrus**<br><br>**Người hâm mộ Pats trố mắt nhìn khi nhận ra Miley Cyrus giả dạng** | Thêm 67% nhấp chuột cho B |

Tôi dự đoán là bạn đúng hơn một nửa, có lẽ bằng cách tự cân nhắc xem bản thân bạn sẽ nhấp chuột lên câu nào. Nhưng chắc bạn không đoán đúng tất cả các cặp.

Tại sao? Bạn đã bỏ qua điều gì? Bạn thiếu những hiểu biết nào về hành vi con người? Bạn có thể học được gì từ các sai lầm của bạn?

Chúng ta thường hỏi những câu hỏi như thế này sau khi thực hiện các dự đoán sai.

Nhưng hãy thử xem việc rút ra các kết luận tổng quát từ những tiêu đề *Globe* đó khó đến thế nào. Trong cặp tiêu đề đầu tiên, việc thay đổi

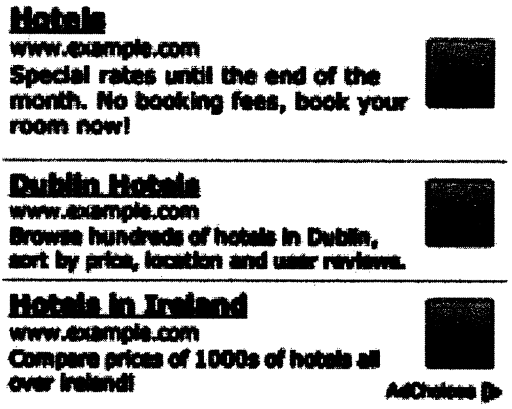
một từ duy nhất ("này" thành "SnotBot") dẫn đến thắng lớn. Trường hợp này cho thấy tiêu đề chi tiết sẽ thắng. Nhưng trong cặp tiêu đề thứ hai, "những quả bóng xì hơi"—từ ngữ chi tiết hơn—lại thua. Cặp thứ tư, "kiếm cả ngân hàng tiền" đánh bại con số $179,000. Ở đây bạn có thể nghĩ rằng dùng từ lóng sẽ hay hơn. Nhưng từ lóng "cuộc thi 'yêu'" lại thua trong cặp thứ 3.

Bài học ở phần thử nghiệm A/B là cẩn thận với những bài học tổng quát. Clark Benson, CEO của ranker.com, một trang tin tức và giải trí chủ yếu dựa vào thử nghiệm A/B để chọn tiêu đề và thiết kế trang. "Suy cho cùng, bạn không quy kết được gì cả," Benson nói. "Đúng là phải thử mọi thứ."

Thử nghiệm lấp chỗ trống trong hiểu biết của ta về bản chất con người. Các chỗ trống này sẽ luôn tồn tại. Nếu chỉ dựa trên kinh nghiệm sống là biết đáp án thì thử nghiệm sẽ không có giá trị. Nhưng ta không biết, vấn đề nằm ở chỗ đó.

Một lí do khác khiến thử nghiệm A/B rất quan trọng là dường như các thay đổi nhỏ có thể gây tác động lớn. Như Benson nói, "Tôi thường xuyên ngạc nhiên với các yếu tố nho nhỏ nhưng có giá trị ngoại cỡ trong thử nghiệm."

Tháng 12/2012, Google thay đổi các quảng cáo của họ. Họ thêm một mũi tên chỉ sang phải, bọc trong một hình vuông.



Mũi tên này thật kì dị. Nó chỉ sang phải, vào chỗ trống không. Thực vậy, khi các mũi tên này xuất hiện lần đầu, nhiều khách hàng Google phê phán. Họ tự hỏi, tại sao lại thêm các mũi tên vô nghĩa vào quảng cáo?

Ừ thì, Google muốn bảo vệ bí quyết kinh doanh, vì vậy họ không nói chính xác các mũi tên có giá trị thế nào. Nhưng họ nhấn mạnh rằng các mũi tên này đã thắng trong cuộc thử nghiệm A/B. Lí do Google thêm các mũi tên là nó khiến nhiều người nhấp chuột. Và thay đổi nhỏ dường như vô nghĩa này đem đến cho Google và các đối tác quảng cáo của họ hàng đống tiền.

Vậy bạn có thể tìm thấy các thay đổi nhỏ tạo ra lợi nhuận ngoại cỡ này như thế nào? Bạn phải thử rất nhiều thứ, thậm chí nhiều cái có vẻ không quan trọng. Thực vậy, người dùng Google đã thấy rất nhiều lần các quảng cáo được thay đổi một chút xíu để rồi lại trở về hình thức trước đó. Họ đã vô tình trở thành thành viên của các nhóm thực nghiệm trong các thử nghiệm A/B—và mỗi người chỉ thấy một số phiên bản điều chỉnh nhỏ mà không hề hay biết.

THỬ NGHIỆM CANH GIỮA (không hiệu quả)

**Best Selling iPad 2 Case**
**The ZAGGmate™ - Tough Aluminum Case**
**with build in Bluetooth Keyboard**
**www.zagg.com**

THỬ NGHIỆM NGÔI SAO MÀU XANH (không hiệu quả)

**Foster's Hollywood Restaurant Reviews, Madrid, Spain ...**
www.tripadvisor.co.uk > ... > Madrid > Madrid Restaurants ▾ TripAdvisor ▾
★★★★★ Rating: 3 - 118 reviews
**Foster's Hollywood, Madrid: See 118 unbiased reviews of Foster's Hollywood, rated 3 of 5 on TripAdvisor and ranked #3647 of 6489 restaurants in Madrid**

THỬ NGHIỆM FONT MỚI (không hiệu quả)

**Live Stock Market News**
**Free Charts, News and Tips from UTVi Experts. Visit us Today!**
UTVi.com/Stocks

Các thay đổi ở hình trên không bao giờ được thực hiện đại trà. Chúng đã thua. Nhưng đó là một phần của quy trình tuyển chọn các thay đổi hiệu quả. Đường đi đến vinh quang của "mũi tên chỉ sang phải" được lót bằng các ngôi sao xấu xí, các cách căn chỉnh chữ sai lạc, và các font chữ phô trương.

Có thể rất vui khi đoán điều gì khiến người ta nhấp chuột. Và nếu bạn là người đảng Dân chủ, có thể rất thú vị khi biết các thử nghiệm đã mang đến cho Obama nhiều tiền hơn. Nhưng có một mặt tối đối với thử nghiệm A/B.

Trong quyển sách *Irresistible* xuất sắc của mình, Adam Alter viết về sự trỗi dậy của các thói nghiện mang tính hành vi trong xã hội đương thời. Nhiều người đang thấy rằng một số phương diện của Internet khiến ta ngày càng khó từ bỏ.

Bộ dữ liệu tâm đắc của tôi—các tìm kiếm Google—có thể cho ta vài đầu mối về những thứ gây nghiện nhất. Theo Google, hầu hết các thói nghiện vẫn là các thứ mà người ta đã đánh vật nhiều thập niên qua—ma túy, tình dục, và rượu, chẳng hạn. Nhưng Internet đang bắt đầu tác động mạnh lên danh sách đó—với "khiêu dâm" và "Facebook" đang nằm trong top 10 các thói nghiện được đề cập.

**Các thói nghiện được tìm kiếm hàng đầu trên Google, 2016**

| | |
|---|---|
| Ma túy | Đường |
| Tình dục | Tình yêu |
| Khiêu dâm | Cờ bạc |
| Rượu bia | Facebook |

Thử nghiệm A/B có thể có góp một chân trong việc khiến Internet gây nghiện khủng khiếp đến thế này.

Quyển *Irresistible* dẫn lời Tristan Harris, một "nhà đạo đức thiết kế," giải thích tại sao người ta khó cưỡng lại một số website trên Internet đến

thế: "Có 1 ngàn người phía bên kia màn hình, công việc của họ chính là phá vỡ sự tự kiểm soát bản thân của bạn."

Và những người này đang dùng thử nghiệm A/B.

Qua thử nghiệm, Facebook có thể biết rằng đổi nút nào đó sang một màu nào đó sẽ khiến người ta trở lại trang của họ thường xuyên hơn. Vậy là họ đổi nút sang màu đó. Tiếp đến, họ có thể biết rằng một font chữ nào đó khiến người ta trở lại trang của họ thường xuyên hơn. Vậy là họ đổi sang font chữ đó. Tiếp đến, họ có thể biết rằng gởi email cho người ta tại một thời điểm nhất định khiến người ta trở lại trang của họ thường xuyên hơn. Vậy là họ gởi email tại đúng thời điểm đó.

Sớm thôi, Facebook sẽ được tối ưu để tối đa hóa thời gian người ta bỏ ra trên Facebook. Nói cách khác, chỉ cần phát hiện đủ các thay đổi hiệu quả trong thử nghiệm A/B, bạn sẽ có một website gây nghiện. Đó là kiểu phản hồi mà các công ty thuốc lá không bao giờ có.

Thử nghiệm A/B đã trở thành công cụ trong ngành sản xuất game. Alter viết, World of Warcraft thử nghiệm A/B các phiên bản game của mình. Một nhiệm vụ có thể yêu cầu bạn khử ai đó. Một nhiệm vụ khác có thể yêu cầu bạn cứu cái gì đó. Nhà thiết kế game có thể giao cho các nhóm người chơi những nhiệm vụ khác nhau, sau đó xem nhiệm vụ nào giữ được nhiều người chơi hơn. Ví dụ, họ có thể thấy rằng nhiệm vụ cứu người khiến ta trở lại thường xuyên hơn 30%. Nếu thử rất, rất nhiều nhiệm vụ, họ sẽ phát hiện ngày càng nhiều nhiệm vụ được ưa thích. Cứ mỗi 30% tăng thêm này cộng lại và tạo ra một game có khả năng giữ chân rất nhiều thanh niên trưởng thành đang sống nhờ trong tầng hầm nhà cha mẹ.

Nếu bạn thấy cái vụ này hơi không ổn, tôi đồng ý với bạn. Chúng ta sẽ bàn thêm một chút về các vấn đề đạo đức của việc này và các mặt khác của Dữ Liệu Lớn ở phần gần cuối sách. Nhưng dù tốt hay xấu, thí nghiệm bây giờ đã là một công cụ quan trọng trong bộ công cụ của nhà khoa học dữ liệu. Và có một hình thức thí nghiệm khác đang nằm trong bộ công cụ đó. Nó được dùng để đặt đủ loại câu hỏi, bao gồm cả chuyện các quảng cáo truyền hình có thật sự hiệu quả hay không.

## Các thí nghiệm tàn bạo nhưng khai sáng của tự nhiên

Ngày 22/1/2012, đội New England Patriots đấu với đội Baltimore Ravens trong trận chung kết giải bóng bầu dục AFC.

Trận đấu còn lại 1 phút. Đội Ravens bị dẫn trước, nhưng họ đang có bóng. 60 giây tiếp theo sẽ quyết định đội nào được chơi trận Siêu cúp Super Bowl.[1] 60 giây tiếp theo sẽ giúp quyết định thành tựu của các cầu thủ. Và phút cuối cùng của trận đấu này sẽ làm một điều khác có ảnh hưởng sâu sắc hơn nhiều, dưới góc độ của một nhà kinh tế học: 60 giây cuối cùng đó sẽ giúp giải đáp một bí mật, một lần cho mãi mãi: Các quảng cáo có hiệu quả hay không?

Ý niệm các quảng cáo cải thiện doanh số rõ ràng là quan trọng đối với nền kinh tế của chúng ta. Nhưng nó khó chứng minh vô cùng. Thực vậy, đây là một ví dụ kinh điển cho thấy sự khó khăn của việc phân biệt giữa mối tương quan và quan hệ nhân quả.

Không nghi ngờ gì nữa, các sản phẩm mà quảng cáo nhiều nhất thì cũng có doanh số cao nhất. 20th Century Fox chi 150 triệu USD marketing bộ phim *Avatar*, nó trở thành phim có doanh thu cao nhất mọi thời đại. Nhưng bao nhiêu trong 2.7 tỉ USD doanh thu bán vé *Avatar* là nhờ vào marketing mạnh tay? Một phần lí do 20th Century Fox chi thật nhiều tiền cho việc quảng bá có lẽ là vì họ biết mình đang nắm trong tay một sản phẩm đáng mơ ước.

Các công ty tin rằng họ biết các quảng cáo của họ hiệu quả tới đâu. Các nhà kinh tế học hoài nghi chuyện đó. Giáo sư kinh tế học Steven Levitt của Đại học Chicago, trong khi cộng tác với một công ty điện tử, không thấy ấn tượng khi công ty này cố thuyết phục ông là họ biết các quảng cáo của họ rất hiệu quả. Levitt tự hỏi, sao mà họ có thể tự tin như vậy được nhỉ?

Công ty giải thích rằng, hàng năm, vào những ngày trước Ngày của cha, họ dốc sức chi cho quảng cáo truyền hình. Gần như chắc chắn, hàng năm trước lễ Ngày của cha, họ có doanh số cao nhất. Ừ, có thể đó chỉ là

[1] [ND] Đội vô địch AFC sẽ đấu với đội vô địch NFC trong trận Siêu cúp Super Bowl.

vì nhiều người hay mua đồ điện tử cho cha, đặc biệt để làm quà, bất kể có quảng cáo hay không.

"Họ vận dụng quan hệ nhân quả ngược hoàn toàn," Levitt nói trong một bài giảng. Ít nhất, có thể là như thế. Chúng ta không biết chắc. "Vấn đề đó khó thực sự," Levitt nói thêm.

Mặc dù vấn đề này quan trọng cần giải quyết, các công ty lại rất ngại tiến hành thí nghiệm nghiêm ngặt. Levitt cố thuyết phục công ty điện tử đó thực hiện một thí nghiệm chọn mẫu ngẫu nhiên có đối chứng để biết chính xác các quảng cáo truyền hình của họ hiệu quả tới đâu. Vì thử nghiệm A/B chưa thực hiện được trên truyền hình, họ sẽ phải thử bằng cách không quảng cáo ở một số vùng.

Đây là cách công ty đó trả lời: "Ông điên à? Chúng ta không thể không quảng cáo ở 20 thị trường. CEO sẽ giết cả bọn đấy." Câu đó kết thúc sự cộng tác của Levitt với công ty.

Điều đó đưa ta trở lại trận bóng Patriots – Ravens. Kết quả của một trận bóng bầu dục có thể giúp ta quyết định các tác động nhân quả của quảng cáo như thế nào? Ừ, nó không thể nói cho ta biết các tác động của một chiến dịch quảng cáo cụ thể từ một công ty cụ thể. Nhưng nó có thể cung cấp bằng chứng về tác động trung bình của quảng cáo từ nhiều chiến dịch lớn.

Thì ra, có một thí nghiệm quảng cáo ẩn trong các trận đấu như thế này. Nó hoạt động như sau: Đến lúc diễn ra các trận chung kết, các công ty đã mua và sản xuất xong các quảng cáo Super Bowl của họ rồi. Vào thời điểm các doanh nghiệp quyết định sẽ chạy quảng cáo nào, họ không biết đội nào sẽ chơi trong trận Super Bowl.

Tuy nhiên, kết quả của các trận chung kết tranh vé Super Bowl sẽ ảnh hưởng rất lớn đối với người thực sự xem trận Siêu cúp Super Bowl. Hai đội cuối cùng giành quyền đấu Super Bowl sẽ kéo theo họ một lượng khán giả khổng lồ. Nếu New England—đội ở gần khu vực Boston—mà thắng, người dân ở Boston sẽ xem trận Super Bowl nhiều vượt xa người dân ở Baltimore. Và ngược lại.

Đối với các công ty, nó giống như một cú tung đồng xu để quyết định họ sẽ cho quảng cáo tiếp cận thêm hàng chục ngàn người ở Baltimore hay ở Boston, một cú tung đồng xu mà sẽ diễn ra sau khi quảng cáo của họ đã được mua và sản xuất xong.

Bây giờ, hãy trở lại sân đấu. Jim Nantz trên đài CBS đang thông báo kết quả cuối cùng của thí nghiệm này.

> Billy Cundiff! Anh đang chuẩn bị san bằng tỉ số! Chắc chắn sẽ có hiệp phụ. 2 năm qua, anh đã ghi 16 bàn kiểu này, không trật phát nào! Sút!!! Sút rồi!!! Coi chừng! Coi chừng! Không được rồi. [...] Đội Patriots đang quỳ xuống chào sân và họ sẽ thực hiện cuộc hành trình đến thành phố Indianapolis. Họ đang hướng đến trận Siêu cúp Super Bowl 46!

Hai tuần sau, Super Bowl 46 có tỉ suất người xem (audience share) là 60.3 ở Boston và 50.2 ở Baltimore. Có thêm 60 ngàn người ở Boston xem các quảng cáo 2012.

Năm tiếp theo, cũng chính 2 đội đó gặp nhau ở trận chung kết AFC. Lần này, Baltimore thắng. Lần này, thời lượng xem quảng cáo Super Bowl 2013 sẽ tăng thêm ở Baltimore.

| | **Tỉ suất người xem của Super Bowl 2012 (Boston tham gia)** | **Tỉ suất người xem của Super Bowl 2013 (Baltimore tham gia)** |
|---|---|---|
| **Boston** | 56.7 | 48.0 |
| **Baltimore** | 47.9 | 59.6 |

Hal Varian (trưởng kinh tế gia tại Google), Michael D. Smith (kinh tế gia tại Carnegie Mellon), và tôi đã dùng 2 trận đấu này và tất cả các trận Super Bowl khác từ 2004 đến 2013 để kiểm tra xem các quảng cáo Super Bowl có hiệu quả hay không, và nếu có thì bao nhiêu. Cụ thể chúng tôi xem thử khi một công ty quảng cáo một bộ phim trong trận Super Bowl, họ có thấy sự gia tăng đột biến trong doanh thu phòng vé ở các thành phố có lượng khán giả xem trận đấu cao hơn hay không.

Thực tế là họ có thấy. Người dân các thành phố có đội bóng thi đấu trận Super Bowl thường xem các phim đã quảng cáo trong trận Super Bowl với tỉ lệ cao hơn đáng kể người dân các thành phố có đội bóng vừa mất quyền thi đấu. Thêm nhiều người tại các thành phố đó xem quảng cáo. Thêm nhiều người tại các thành phố đó quyết định đi xem phim.

Một lời giải thích tiềm năng khác là việc có một đội đấu trận Super Bowl khiến cho ta có nhiều khả năng đi xem phim hơn. Tuy nhiên, chúng tôi đã kiểm thử một nhóm phim có ngân sách tương tự, được phát hành ở các thời điểm tương tự, nhưng không quảng cáo trong trận Super Bowl. Không có lượng người xem tăng lên tại các thành phố có các đội đấu trận Super Bowl.

OK, như bạn có thể đã đoán, quảng cáo có hiệu quả. Điều này không có gì quá bất ngờ.

Nhưng không chỉ là có hiệu quả chung chung, các quảng cáo có hiệu quả đến mức không thể tin được. Thực vậy, khi lần đầu thấy kết quả, chúng tôi kiểm tra lần 2, lần 3, và lần 4 để chắc chắn là đúng—bởi tỉ suất sinh lợi là rất lớn. Bộ phim trung bình trong mẫu của chúng tôi đã chi khoảng 3 triệu USD cho 1 vị trí quảng cáo (ad slot) ở Super Bowl. Họ đã nhận được 8.3 triệu USD doanh thu phòng vé tăng thêm, tỉ suất hoàn vốn đầu tư lên đến 2.8.

Kết quả này được xác nhận bởi 2 nhà kinh tế khác, Wesley R. Hartmann và Daniel Klapper, họ đã nghĩ ra một ý tưởng tương tự và độc lập trước đó. Họ nghiên cứu các quảng cáo bia và nước giải khát chạy trong trận Super Bowl, đồng thời tận dụng thời lượng xem quảng cáo tăng thêm tại các thành phố có đội tham dự. Họ phát hiện tỉ suất hoàn vốn đầu tư là 2.5. Mặc dù rất tốn kém, kết quả của chúng tôi và của họ cho thấy quảng cáo Super Bowl kích cầu hiệu quả đến nỗi, nếu tính đúng thì các công ty này đang mua quảng cáo với cái giá quá hời.

Và tất cả điều này có ý nghĩa gì cho công ty điện tử mà Levitt đã hợp tác khi trước? Có thể các quảng cáo Super Bowl hiệu quả hơn các hình thức quảng cáo khác. Thế nhưng, ít nhất nghiên cứu của chúng tôi cũng cho thấy rằng quảng cáo vào Ngày của cha có thể là một ý tưởng hay.

***

Một ưu điểm của thí nghiệm Super Bowl là không nhất thiết phải cố ý đưa ai vào nhóm thực nghiệm hoặc nhóm đối chứng. Nó diễn ra ngẫu nhiên, tùy vào kết quả trận bóng. Nói cách khác, nó diễn ra một cách tự nhiên. Tại sao đó lại là một thuận lợi? Bởi vì các thí nghiệm phi tự nhiên, ngẫu nhiên và có đối chứng, mặc dù rất mạnh và dễ dàng thực hiện hơn trong thời đại kĩ thuật số, vẫn không phải khi nào cũng thực hiện được.

Đôi khi chúng ta không thể cùng hành động kịp thời. Đôi khi, như trường hợp công ty điện tử không muốn thực hiện thử nghiệm quảng cáo nói trên, kết quả quá quan trọng nên ta không thể mạo hiểm mà thử nghiệm.

Đôi khi thí nghiệm là bất khả thi. Giả sử ta quan tâm chuyện một quốc gia phản ứng với việc mất một lãnh đạo như thế nào. Điều đó có dẫn đến chiến tranh không? Nền kinh tế có ngưng hoạt động không? Hay có khi không thay đổi gì nhiều? Rõ ràng, ta không thể chỉ đơn giản là thủ tiêu một số tổng thống và thủ tướng để xem điều gì xảy ra được. Điều đó không chỉ bất khả thi mà còn trái đạo lí nữa. Qua nhiều thập niên, các trường đại học đã xây dựng các hội đồng thẩm định cơ sở (IRB) để quyết định xem thí nghiệm được đề xuất có đạo đức hay không.

Vậy nếu muốn biết các tác động nhân quả trong một tình huống nhất định, nhưng thí nghiệm lại trái đạo đức hoặc bất khả thi, thì ta có thể làm gì? Ta có thể sử dụng cái mà nhà kinh tế gọi là các thí nghiệm tự nhiên—"tự nhiên" ở đây được định nghĩa đủ rộng để bao gồm cả các trận bóng bầu dục.

Dù tốt hay xấu (OK, rõ ràng là xấu), chắc chắn có nhiều thành phần ngẫu nhiên trong cuộc sống. Không ai biết chắc thứ gì hoặc người nào đang chịu trách nhiệm về vũ trụ này. Nhưng có một điều rõ ràng: Bất cứ ai đang điều khiển thế giới—dù là các quy luật cơ học lượng tử, Chúa, hay một chú nhóc mặt mụn mặc xà lỏn đang mô phỏng vũ trụ chúng ta trên máy tính—người đó sẽ *không* cần được IRB phê duyệt.

Tự nhiên thí nghiệm trên chúng ta suốt. Hai người bị bắn. Một viên đạn dừng ngay sát một cơ quan sinh tử. Viên kia thì không. Các rủi ro

này khiến cuộc sống bất công. Nhưng, có một điều an ủi, chính các rủi ro đó làm các nhà kinh tế học dễ nghiên cứu cuộc sống hơn một chút. Các nhà kinh tế học dùng sự ngẫu nhiên của cuộc sống để kiểm định các tác động nhân quả.

Trong số 43 tổng thống Mĩ, 16 vị đã là nạn nhân của những vụ ám sát nghiêm trọng, và 4 vị đã chết. Lí do tại sao một số tổng thống lại sống sót chủ yếu là ngẫu nhiên.

So sánh John F. Kennedy và Ronald Reagan. Cả 2 ông đều bị những viên đạn đi thẳng đến các bộ phận cơ thể dễ tổn thương nhất. Viên đạn bắn JFK vỡ não, giết chết ông ngay sau đó. Viên bắn Reagan dừng lại cách tim ông chỉ mấy lóng tay, cho phép các bác sĩ cứu sống ông. Reagan sống, JFK thì chết. Không có lí do gì cả—chỉ thuần hên xui thôi.

Những vụ ám sát các lãnh đạo này và sự ngẫu nhiên quyết định họ sống hay chết là điều thường xảy ra khắp thế gian. So sánh Akhmad Kadyrov (Chechnya) và Adolf Hitler (Đức). Cả 2 người chỉ cách một quả bom hoạt động tốt có mấy tấc thôi. Kadyrov chết. Hitler thì thay đổi lịch làm việc, rời khỏi căn phòng bị đặt bom sớm vài phút để đón một chuyến tàu, và thế là sống sót.

Và ta có thể dùng tính ngẫu nhiên lạnh lùng của tự nhiên—giết chết Kennedy nhưng không giết Reagan—để xem điều gì xảy ra khi lãnh đạo một nước bị ám sát. Hai nhà kinh tế học, Benjamin F. Jones và Benjamin A. Olken, đã làm đúng như thế. Nhóm đối chứng ở đây là các nước ở thời điểm những năm ngay sau một vụ ám sát hụt—ví dụ, nước Mĩ giữa thập niên 1980. Nhóm thực nghiệm là các nước ở thời điểm những năm ngay sau một vụ ám sát thành công—ví dụ, nước Mĩ giữa thập niên 1960.

Thế thì, tác động của việc lãnh đạo bị ám sát là gì? Jones và Olken thấy rằng các vụ ám sát thành công thay đổi đột ngột lịch sử thế giới, đưa các nước đến những con đường hoàn toàn khác. Lãnh đạo mới khiến các nước hòa bình trước đó đi đến chiến tranh và các nước chiến tranh trước đó đạt được hòa bình. Lãnh đạo mới khiến các nước bùng nổ kinh tế bắt đầu sụp đổ và các nước sụp đổ kinh tế bắt đầu bùng nổ.

Thực vậy, các kết quả của cuộc thí nghiệm tự nhiên dựa trên các vụ ám sát này đã đánh đổ hàng thập niên thống trị của hiểu biết truyền thống về cách thức các quốc gia vận hành. Nhiều nhà kinh tế học trước đây có khuynh hướng cho rằng các lãnh đạo phần lớn là những con bù nhìn yếu đuối bị các thế lực bên ngoài đưa đẩy. Không phải vậy, theo phân tích thí nghiệm tự nhiên của Jones và Olken.

Nhiều người thường không xem nghiên cứu về các vụ tấn công ám sát nhắm vào các lãnh đạo thế giới này là một ví dụ cho Dữ Liệu Lớn. Con số lãnh đạo chết hoặc suýt chết do ám sát trong nghiên cứu chắc chắn là nhỏ—số cuộc chiến nổ ra hoặc không nổ ra do các sự kiện đó cũng vậy. Các bộ dữ liệu kinh tế cần thiết để mô tả quỹ đạo đặc trưng của một nền kinh tế là lớn, nhưng hầu hết đều có trước thời kì số hóa.

Tuy nhiên, các thí nghiệm tự nhiên như thế—dù bây giờ hầu như chỉ có các nhà kinh tế sử dụng—là rất mạnh và ngày càng quan trọng trong một thời đại của những bộ dữ liệu tốt hơn và lớn hơn. Đây là công cụ mà các nhà khoa học dữ liệu sẽ sớm tìm dùng.

Và vâng, đến bây giờ thì chắc cũng đã rõ, các nhà kinh tế học đang đóng một vai trò quan trọng trong việc phát triển khoa học dữ liệu. Ít nhất là tôi muốn nghĩ vậy, vì đó là ngành học của tôi.

---

Liệu còn nơi nào khác ta có thể tìm thấy các thí nghiệm tự nhiên—hay nói cách khác, những tình huống mà dòng sự kiện ngẫu nhiên đặt mọi người vào các nhóm thực nghiệm và đối chứng?

Ví dụ rõ ràng nhất là một cuộc xổ số, đó là lí do tại sao các nhà kinh tế học thích xổ số. Họ không chơi—vì họ xem đây là một hành vi phi lí trí; họ chỉ nghiên cứu thôi. Nếu một quả bóng bàn mang số 3 lên nằm đầu, ông Jones sẽ giàu. Nếu quả bóng đó mà số 6, ông Johnson sẽ giàu.

Để thử nghiệm các tác động nhân quả của vận may về tiền bạc, các nhà kinh tế học so sánh những người trúng số với những người mua vé số nhưng trật. Các nghiên cứu này nói chung đã phát hiện ra rằng việc

trúng số không làm bạn hạnh phúc trong ngắn hạn, nhưng lại khiến bạn hạnh phúc trong dài hạn.[1]

Các nhà kinh tế học cũng có thể dùng tính ngẫu nhiên của xổ số để xem cuộc sống người ta thay đổi ra sao khi một người láng giềng của họ giàu lên. Dữ liệu cho thấy rằng việc láng giềng của bạn trúng số có thể ảnh hưởng đến cuộc sống của chính bạn. Nếu láng giềng của bạn trúng số, ví dụ, rất có thể bạn sẽ mua một chiếc xe đắt tiền, BMW chẳng hạn. Tại sao? Các nhà kinh tế học khẳng định, gần như chắc chắn nguyên nhân là do ganh tị khi thấy láng giềng mua xe đắt tiền. Bản chất con người mà. Nếu ông Johnson thấy ông Jones lái chiếc BMW mới cáu, ông Johnson cũng muốn có một chiếc.

Thật không may, ông Johnson thường không mua nổi BMW, đó là lí do các nhà kinh tế học thấy rằng láng giềng của những người trúng số có nhiều khả năng bị phá sản hơn hẳn bình thường. Trong trường hợp này, cố gắng đu theo cho kịp gia đình ông Jones là không thể được.

Nhưng thí nghiệm tự nhiên không cần phải ngẫu nhiên rõ ràng như xổ số. Một khi bạn bắt đầu chủ ý tìm sự ngẫu nhiên, bạn sẽ thấy nó khắp nơi—và có thể dùng nó để hiểu cách thế giới của chúng ta vận hành.

Các bác sĩ là một phần của một thí nghiệm tự nhiên. Thỉnh thoảng, chủ yếu là do tùy hứng, chính phủ Mĩ thay đổi công thức bồi hoàn tiền bệnh nhân Medicare cho y bác sĩ. Bác sĩ một số địa phương thấy các khoản phí áp cho một số cách điều trị nhất định tăng lên. Bác sĩ một số nơi khác thì thấy các khoản phí ấy giảm xuống.

Hai nhà kinh tế học—Jeffrey Clemens và Joshua Gottlieb, một bạn học cũ của tôi—đã nghiên cứu các tác động của thay đổi tùy hứng này. Các bác sĩ có luôn luôn chăm sóc bệnh nhân như nhau—điều mà họ xem là rất cần thiết? Hay họ bị chi phối bởi các động cơ tài chính?

Dữ liệu chỉ rõ rằng các bác sĩ có thể được động viên bởi các khoản tiền khuyến khích. Tại các địa phương có tiền bồi hoàn cao hơn, một số

---

[1] Một nghiên cứu nổi tiếng năm 1978 cho rằng trúng số không làm bạn hạnh phúc đã bị lật đổ hoàn toàn.

bác sĩ thiên về sử dụng cách điều trị được bồi hoàn tốt hơn—chẳng hạn như phẫu thuật đục thủy tinh thể, nội soi đại tràng, và chụp MRI.

Tiếp theo là câu hỏi lớn: Bệnh nhân của họ có khỏe hơn sau khi được chăm sóc thêm hay không? Clemens và Gottlieb báo cáo là chỉ có "tác động nhỏ về sức khỏe." Các tác giả không thấy tác động có ý nghĩa thống kê về tỉ lệ tử vong. Thí nghiệm tự nhiên này đã chỉ ra, nếu cung cấp động cơ tài chính để các bác sĩ cho dùng một số cách điều trị nhất định, một số bác sĩ sẽ dùng các cách điều trị này nhiều hơn, nhưng không thay đổi gì nhiều sức khỏe bệnh nhân và dường như cũng không giúp kéo dài cuộc sống của họ.

Các thí nghiệm tự nhiên có thể giúp trả lời các câu hỏi sinh-tử. Chúng cũng có thể giúp trả lời các câu hỏi mà, với một số người trẻ, trông không khác gì sinh-tử.

---

Trường trung học Stuyvesant (thường gọi là "Stuy") nằm trong một tòa nhà gạch, màu nâu tanin, trị giá 150 triệu USD, 10 tầng, nhìn ra dòng sông Hudson, cách World Trade Center vài dãy phố, ở khu thương mại Manhattan. Nói tóm lại, Stuy rất ấn tượng. Trường có 55 lớp AP (nâng cao), 7 ngôn ngữ, và các môn tự chọn về lịch sử Do Thái, khoa học viễn tưởng, và văn học Á-Mĩ. Khoảng 1/4 học sinh tốt nghiệp được nhận vào đại học thuộc nhóm Ivy League hoặc có uy tín tương tự. Stuyvesant đã đào tạo Lisa Randall, giáo sư vật lí tại Harvard; David Axelrod, chiến lược gia của Obama; Tim Robbins, diễn viên đoạt giải Oscar; và tiểu thuyết gia Gary Shteyngart. Những người đã phát biểu tại lễ trao bằng tốt nghiệp ở đó có Bill Clinton, Kofi Annan, và Conan O'Brien.

Điều duy nhất nổi bật hơn các cống hiến và danh sách học sinh tốt nghiệp của Stuyvesant là giá học phí: 0 đô la. Đó là một trường trung học công và có lẽ là trường tốt nhất cả nước. Thực vậy, một nghiên cứu gần đây đã sử dụng 27 triệu đánh giá của 300,000 học sinh và phụ huynh để xếp loại tất cả các trường trung học công tại Mĩ. Stuy xếp số 1. Vậy, không có gì đáng ngạc nhiên, các bậc cha mẹ trung lưu đầy tham vọng ở

New York cùng con cháu cũng tham vọng không kém có thể bị ám ảnh bởi thương hiệu của Stuy.

Đối với Ahmed Yilmaz,[1] con trai của một nhân viên bảo hiểm kiêm giáo viên tại Queens, Stuy là "*the* high school" (ý nói trường trung học duy nhất hoặc hàng đầu).

"Các gia đình tầng lớp lao động và nhập cư xem Stuy như lối thoát," Yilmaz giải thích. "Nếu con bạn học Stuy, nó sẽ vào một đại học chuẩn mực top 20. Gia đình ấy sẽ ổn."

Vậy làm thế nào để vào được trường Stuyvesant? Bạn phải sống ở 1 trong 5 quận của Thành phố New York và có điểm cao hơn điểm chuẩn trong kì thi đầu vào. Thế đấy. Không thư giới thiệu, không tiểu luận, không ưu tiên công trạng, không ưu tiên sắc tộc. 1 ngày, 1 bài kiểm tra, 1 điểm số. Nếu con số của bạn trên ngưỡng, bạn được vào.

Cứ đến tháng 11, khoảng 27,000 thí sinh trẻ ở New York dự kì thi đầu vào. Cạnh tranh thật khốc liệt. Chưa tới 5% số thí sinh vào được Stuy.

Yilmaz giải thích rằng mẹ anh đã "đầu tắt mặt tối" và bỏ ra số tiền ít ỏi bà dành dụm được để con trai chuẩn bị cho kì thi. Sau nhiều tháng ôn luyện mỗi buổi chiều kể cả ngày nghỉ cuối tuần, Yilmaz tự tin là sẽ vào được Stuy. Anh vẫn còn nhớ cái ngày anh nhận được bao thư chứa kết quả. Anh đã trượt chỉ vì 2 câu hỏi.

Tôi hỏi anh lúc đó cảm thấy như thế nào. Anh trả lời, "Bạn cảm thấy thế nào nếu thế giới của bạn sụp đổ khi bạn đang học cấp 2?"

Giải an ủi của Yilmaz không hề tồi tàn chút nào—một tấm vé vào Bronx Science, một trường công đặc biệt có thứ hạng cao khác. Nhưng đó không phải là Stuy. Và Yilmaz cảm thấy Bronx Science là một trường chuyên biệt dành cho dân kĩ thuật hơn. Sau đó 4 năm, anh bị Đại học Princeton từ chối. Anh vào Đại học Tufts và đến giờ đã trải qua một vài nghề. Hiện tại anh là một nhân viên khá thành công tại một công ty kĩ

[1] Tôi đã thay đổi tên và một vài chi tiết của nhân vật.

thuật, mặc dù anh nói công việc của anh "làm mụ mẫm đầu óc" và không được đền bù xứng đáng như anh muốn.

Hơn một thập niên sau, Yilmaz thú nhận rằng đôi lúc anh tự hỏi cuộc đời sẽ ra sao nếu như anh học trường Stuy. "Mọi thứ chắc sẽ khác," anh nói. "Chính xác là mọi người tôi biết đều sẽ khác." Anh tự hỏi liệu Stuyvesant có đưa anh đến với các điểm SAT cao hơn, được vào một trường đại học như Princeton hay Harvard (cả hai đại học này được anh xem là ngon hơn Tufts nhiều), có lẽ là cả một nghề nghiệp hái ra tiền và thành công nữa.

Thật thú vị, nhưng có thể đây cũng là một cách tự hành hạ khi tưởng tượng ra các giả thuyết. Đời mình sẽ ra sao nếu mình tiến tới với người ấy? Nếu mình nhận công việc đó? Nếu mình đi học trường này? Nhưng các câu hỏi sẽ-ra-sao-nếu này dường như không trả lời được. Cuộc đời không phải trò chơi điện tử. Bạn không thể chơi lại các tình huống khác cho đến khi có được kết quả như ý.

Milan Kundera, nhà văn sinh ở Czech, có một câu nói súc tích về điều này trong tiểu thuyết *The Unbearable Lightness of Being* của ông: "Cuộc sống con người xảy ra chỉ một lần, và lí do chúng ta không thể nói quyết định nào của mình là tốt hoặc xấu bởi vì trong một tình huống cụ thể, chúng ta chỉ có thể quyết định một lần duy nhất; chúng ta không được ban cho cuộc đời thứ hai, thứ ba hoặc thứ tư để so sánh các quyết định khác nhau."

Yilmaz sẽ không bao giờ trải qua cái cuộc đời giả tưởng, trong đó anh đã ghi được thêm 2 điểm trong kì thi đó.

Nhưng có lẽ chúng ta cũng biết được vài điều về việc cuộc đời anh ấy có thể khác hay không bằng cách thực hiện một nghiên cứu lượng lớn học sinh trường Stuyvesant.

Một phương pháp lệch lạc và ngây thơ sẽ là so sánh tất cả học sinh học trường Stuyvesant với tất cả học sinh không học trường đó. Ta có thể phân tích họ đã làm các bài thi AP và SAT thế nào—và họ được nhận vào những đại học nào. Nếu làm vậy, ta sẽ thấy rằng những học sinh học trường Stuyvesant có điểm các bài thi tiêu chuẩn cao hơn nhiều và được

nhận vào các đại học tốt hơn đáng kể. Nhưng như ta đã thấy trong chương này, loại bằng chứng đó, tự nó, là không thuyết phục. Có thể lí do học sinh Stuyvesant làm bài tốt hơn nhiều là bởi Stuy thu hút các học sinh giỏi hơn nhiều ngay từ đầu vào. Tương quan ở đây không chứng minh quan hệ nhân quả.

Để thử nghiệm các tác động *nhân quả* của trường Stuyvesant, ta cần so sánh 2 nhóm hầu như giống nhau: nhóm học trường Stuy (nhóm thực nghiệm) và nhóm không học (nhóm đối chứng). Chúng ta cần một thí nghiệm tự nhiên. Nhưng chúng ta có thể tìm nó ở đâu?

Câu trả lời: các học sinh như Yilmaz, những người đạt điểm rất, rất gần điểm chuẩn để được vào Stuyvesant.[1] Các học sinh trượt ngay gần điểm chuẩn là nhóm đối chứng; các học sinh vừa đủ điểm chuẩn là nhóm thực nghiệm.

Hầu như không có lí do để nghĩ rằng học sinh tại hai phía điểm chuẩn chênh nhau nhiều về tài năng hay động lực. Điều gì khiến người này chỉ cao hơn 1 hoặc 2 điểm bài thi so với người kia? Có thể người ghi điểm thấp hơn đã thiếu ngủ 10 phút hoặc đã ăn sáng ít dinh dưỡng hơn. Có thể người ghi điểm cao hơn đã nhớ một từ đặc biệt khó trong bài thi nhờ một cuộc chuyện trò với bà ngoại 3 năm trước đó.

Thực vậy, thể loại thí nghiệm tự nhiên này—sử dụng điểm cắt đột ngột dưới dạng số—mạnh đến nỗi nó được dân kinh tế học đặt cho một cái tên riêng: gián đoạn hồi quy (regression discontinuity). Bất cứ khi nào có một con số chính xác chia người ta thành 2 nhóm khác nhau (một điểm gián đoạn), các nhà kinh tế học có thể so sánh (hoặc hồi quy) kết quả của những người rất, rất gần với điểm cắt.

Hai nhà kinh tế học, M. Keith Chen và Jesse Shapiro, đã lợi dụng một điểm cắt đột ngột được sử dụng bởi các nhà tù liên bang để thử nghiệm

---

[1] Khi tìm kiếm những người gần đạt điểm đỗ như Yilmaz, tôi đã rất ấn tượng bởi số lượng người suýt đậu—những người từ 20 đến dưới 60 tuổi. Họ vẫn nhớ trải nghiệm dự thi này từ khi mới 13, 14 tuổi và nói về việc trượt sát điểm chuẩn bằng ngôn từ rất ấn tượng. Nhóm này có cựu nghị sĩ quốc hội kiêm ứng viên thị trưởng Thành phố New York là Anthony Weiner; ông nói ông trượt Stuy vì thiếu chỉ 1 điểm. "Họ không muốn có tôi," ông nói với tôi, trong một cuộc phỏng vấn điện thoại.

tác động của các điều kiện nhà tù khó khăn đối với tội phạm tương lai. Các tù nhân liên bang tại Mĩ được cung cấp một điểm số, dựa trên bản chất tội phạm và lịch sử phạm tội của họ. Điểm số quyết định các điều kiện ở tù của họ. Những người điểm số đủ cao sẽ đến cơ sở có an ninh cao, nghĩa là ít được tiếp xúc với người khác, ít được tự do di chuyển, và chắc là nhiều bạo lực hơn từ phía cai ngục và các tù nhân khác.

Một lần nữa, sẽ không công bằng khi so sánh toàn bộ người tù bị đưa đến các nhà tù an ninh cao với toàn bộ người tù bị đưa đến các nhà tù an ninh thấp. Nhà tù an ninh cao có nhiều kẻ giết người và hiếp dâm hơn, nhà tù an ninh thấp thì nhiều phạm nhân dùng ma túy và trộm vặt hơn.

Nhưng những người ngay trên hoặc ngay dưới điểm cắt hầu như có lịch sử phạm tội và nền tảng giống nhau. Tuy nhiên, cái điểm mốc đơn lẻ vô giá trị này lại đồng nghĩa với trải nghiệm nhà tù cực kì khác biệt.

Kết quả? Các nhà kinh tế học phát hiện rằng những người tù được đưa đến các điều kiện khắc nghiệt hơn rất có thể phạm thêm tội ác khi ra tù. Điều kiện nhà tù khắc nghiệt, chứ không phải việc ngăn cản họ phạm tội, làm cho họ chai lì đi và bạo lực hơn khi trở lại thế giới bên ngoài.

Vậy điểm gián đoạn hồi quy trong trường hợp trường Stuyvesant cho ta biết điều gì? Một nhóm nhà kinh tế học từ MIT và Duke—Atila Abdulkadirog˘lu, Joshua Angrist, và Parag Pathak—đã thực hiện nghiên cứu. Họ so sánh đầu ra của học sinh New York ở 2 bên điểm chuẩn. Nói cách khác, các nhà kinh tế học này nghiên cứu hàng trăm học sinh đã *trượt* trường Stuyvesant bởi vì thiếu 1 hoặc 2 điểm. Hai ông so sánh họ với hàng trăm học sinh đã có một ngày thi tốt hơn và *đỗ* vào Stuy nhờ 1 hoặc 2 điểm tăng thêm. Thước đo thành công của họ là điểm AP, điểm SAT, và điểm xếp hạng các trường đại học mà cuối cùng họ đã theo học.

Các kết quả gây sửng sốt của họ được làm rõ bởi tiêu đề họ đặt cho bài báo: "Ảo tưởng tinh hoa" (Elite Illusion). Ảnh hưởng của trường Stuyvesant là bao nhiêu? Không, không, không, và không. Học sinh ở cả 2 bên điểm chuẩn sau này đều có điểm AP và điểm SAT ngang ngửa nhau và theo học các trường đại học có mức độ uy tín không chênh lệch gì nhau.

Toàn bộ lí do khiến học sinh Stuy thành đạt hơn trong cuộc sống so với học sinh không học trường Stuy: Học sinh Stuyvesant vốn đã giỏi hơn từ đầu vào. Stuy *không giúp* bạn làm bài thi AP tốt hơn hay làm bài thi SAT tốt hơn, cũng *chẳng giúp* bạn sau này vào một trường đại học tốt hơn.

Các nhà kinh tế học viết, "Chuyện cạnh tranh căng thẳng để có chỗ trong trường dường như không được bù đắp bằng sự cải thiện thành tích học tập đối với một bộ phận rất lớn học sinh."

Tại sao việc học trường nào lại có thể không quan trọng? Một số câu chuyện nữa có thể giúp đi đến câu trả lời. Xem tiếp 2 học sinh, Sarah Kaufmann và Jessica Eng, cả 2 thanh niên New York từ nhỏ đã mơ được học Stuy. Điểm của Kaufmann vừa đủ điểm chuẩn. "Tôi không nghĩ mình có thể gặp điều thú vị như thế một lần nào nữa," Kaufmann hồi tưởng. Điểm của Eng thì ngay dưới điểm chuẩn; cô trượt chỉ vì thiếu 1 câu đúng. Kaufmann vào được trường Stuy mơ ước. Eng thì không.

Vậy cuộc đời của họ cuối cùng thế nào? Cả hai cho đến nay đều có sự nghiệp thành công, xứng đáng—như hầu hết những người New York có điểm thi trong top 5%. Thật mỉa mai, Eng lại thích trải nghiệm thời trung học hơn. Bronx Science, nơi cô học, là trường trung học duy nhất có một bảo tàng về vụ diệt chủng Holocaust. Eng phát hiện là mình yêu thích nghề giám tuyển trưng bày bảo tàng và đã theo học ngành nhân chủng học tại Cornell.

Kaufmann cảm thấy bị lạc trôi một chút tại Stuy—học sinh quá nặng nề chuyện điểm số và trường thì quá chú trọng việc kiểm tra chứ không phải việc dạy học. Cô gọi trải nghiệm ấy "đúng là một cái túi hỗn hợp." Nhưng đó là một trải nghiệm giúp cô học được nhiều điều. Cô nhận ra rằng, với đại học, cô sẽ chỉ quan tâm các trường tổng quan (liberal arts school), chú trọng hơn vào việc giảng dạy. Cô được nhận vào trường mơ ước, Đại học Wesleyan. Ở đó cô tìm thấy niềm đam mê giúp đỡ người khác, và bây giờ cô là một luật sư công ích.

Người ta thích ứng với trải nghiệm của mình, và những người sẽ thành công thường tìm thấy lợi thế trong bất cứ tình huống nào. Các yếu

tố giúp bạn thành công là tài năng và động cơ của bạn, chứ không phải là việc ai đọc bài diễn văn tại lễ tốt nghiệp hoặc các lợi thế khác mà các trường tên tuổi lớn nhất thường sở hữu.

Đây chỉ là một nghiên cứu riêng lẻ, và nó có lẽ bị yếu đi bởi sự thật là hầu hết các học sinh xém đỗ trường Stuyvesant cuối cùng cũng học một trường tốt khác. Nhưng ngày càng có thêm bằng chứng cho thấy, mặc dù học một trường tốt là quan trọng, chuyện học ở trường được xem là xịn nhất cũng chẳng giúp bạn có thêm được gì nhiều.

Chuyện đại học chẳng hạn. Liệu có khác biệt giữa việc học một trong những đại học tốt nhất thế giới như Harvard và một trường có tiếng như Penn State không?

Một lần nữa, có mối tương quan rõ ràng giữa thứ hạng trường học và số tiền người ta kiếm được. Sau 10 năm vào nghề, người tốt nghiệp Harvard trung bình kiếm được 123,000 USD. Người tốt nghiệp Penn State trung bình kiếm được 87,800 USD.

Nhưng mối tương quan này không ám chỉ quan hệ nhân quả.

Hai nhà kinh tế học, Stacy Dale và Alan B. Krueger, đã nghĩ ra một cách khéo léo để kiểm định vai trò của các đại học tinh hoa đối với tiềm năng kiếm sống tương lai của các sinh viên tốt nghiệp. Họ có một bộ dữ liệu lớn theo dõi cả một lượng thông tin lớn về học sinh trung học, gồm cả trường đại học họ ứng tuyển, trường họ được nhận vào, trường họ theo học, nền tảng gia đình họ, và thu nhập của họ khi trưởng thành.

Để có nhóm thực nghiệm và nhóm đối chứng, Dale và Krueger so sánh các học sinh có nền tảng tương tự, được cùng trường nhận vào nhưng chọn theo học các trường khác nhau. Một số học sinh được Harvard nhận đã theo học Penn State—có lẽ để gần người yêu hơn hoặc do trường đó có một giáo sư họ muốn học. Nói cách khác, theo các ban tuyển sinh, những học sinh này tài năng không kém gì những học sinh được vào Harvard. Nhưng họ có những trải nghiệm giáo dục khác nhau.

Vậy khi 2 học sinh có nền tảng tương tự, cùng được Harvard nhận nhưng một người lại chọn học Penn State, điều gì xảy ra? Kết quả của

các nhà nghiên cứu cũng gây sửng sốt như trường hợp Stuyvesant. Các sinh viên đó cuối cùng có thu nhập suýt soát nhau trong sự nghiệp của mình. Nếu tiền lương tương lai là thước đo, thì các học sinh tương tự nhau, được nhận vào các trường uy tín tương tự nhưng chọn học trường khác nhau, cuối cùng cũng sẽ có vị trí tương đương nhau.

Báo chí rải đầy các bài về những người thành công lớn đã theo học các trường nhóm Ivy League: Những người như Bill Gates, nhà sáng lập Microsoft; Mark Zuckerberg và Dustin Moskovitz, các nhà sáng lập Facebook; tất cả đều học Harvard. (Dù vậy, tất cả đều bỏ học, điều đó đặt ra các câu hỏi về giá trị của việc học trường thuộc Ivy League.)

Vẫn còn đó các câu chuyện những người có tài năng đủ để được nhận vào một trường nhóm Ivy League nhưng đã chọn học một trường ít uy tín hơn, và đã có cuộc sống cực kì thành công: Warren Buffett, bắt đầu tại trường Wharton thuộc Đại học Pennsylvania, một trường kinh doanh nhóm Ivy League, nhưng đã chuyển đến Đại học Nebraska–Lincoln vì rẻ hơn. Ông ghét Philadelphia, và ông nghĩ các tiết học ở Wharton thật buồn chán. Dữ liệu chỉ ra rằng, ít nhất là về mặt thu nhập, việc chọn theo học một trường ít uy tín hơn là một quyết định khá ổn, đối với Buffett và những người khác.

***

Quyển sách này tên là *Mọi người đều nói dối*. Theo đây, tôi chủ yếu muốn nói rằng người ta nói dối—với bạn bè, với các khảo sát, và với chính mình—để làm mình trông có vẻ xịn hơn.

Nhưng nhân gian cũng nói dối với ta bằng việc cho ta những dữ liệu lỗi và sai lạc. Thế giới chỉ cho ta rất nhiều người tốt nghiệp Harvard thành công, ít người tốt nghiệp Penn State thành công hơn, và ta cứ quy kết rằng có một lợi thế rất lớn khi học trường Harvard.

Bằng cách lí giải thông minh các thí nghiệm của tự nhiên, ta có thể hiểu đúng dữ liệu của thế giới—để phát hiện cái gì thực sự hữu ích và cái gì vô ích.

***

Các thí nghiệm tự nhiên cũng liên quan đến chương trước nữa. Chúng thường đòi hỏi phóng to—theo nhóm thực nghiệm và nhóm đối chứng: các thành phố trong thí nghiệm Super Bowl, các hạt trong thí nghiệm định giá Medicare, các học sinh gần điểm cắt trong thí nghiệm Stuyvesant. Và việc phóng to, như được thảo luận trong chương trước, thường đòi hỏi các bộ dữ liệu lớn, toàn diện—thứ ngày càng có nhiều khi thế giới được số hóa. Vì không biết tự nhiên sẽ chọn thực hiện thí nghiệm khi nào, ta không thể thiết lập một cuộc khảo sát nhỏ để đo lường kết quả. Ta cần nhiều dữ liệu sẵn có để học hỏi từ những sự can thiệp của tự nhiên này. Ta cần Dữ Liệu Lớn.

Còn một điểm quan trọng nữa cần làm rõ về các thí nghiệm—của chính chúng ta hoặc của tự nhiên—trong chương này. Phần lớn quyển sách này tập trung vào việc hiểu thế giới—Obama đã phải gánh chịu bao nhiêu vì sự phân biệt chủng tộc, bao nhiêu nam giới đồng tính, hay cả nam và nữ bất an về cơ thể của họ ra sao. Nhưng các thí nghiệm có đối chứng hoặc thí nghiệm tự nhiên này có một khuynh hướng ứng dụng thực tế hơn. Chúng cải thiện việc ra quyết định, giúp ta biết các hoạt động can thiệp nào là hiệu quả hoặc không hiệu quả.

Các công ty có thể biết cách kiếm thêm khách hàng. Chính phủ có thể biết cách dùng tiền bồi hoàn để động viên các bác sĩ theo cách tốt nhất. Học sinh có thể biết trường nào có giá trị nhất. Các thí nghiệm này chỉ ra tiềm năng của Dữ Liệu Lớn—để thay thế các phỏng đoán, quan niệm truyền thống, và các mối tương quan giả mạo bằng những thứ thực sự hiệu quả—*theo quan hệ nhân quả.*

# III

# Dữ Liệu Lớn: Xử lí cẩn thận

# CHƯƠNG 7

## Dữ Liệu Lớn hay Phế Liệu Lớn? Điều nó không làm được.

"Seth, Lawrence Summers muốn gặp anh," email viết, hơi có vẻ bí ẩn. Đó là lời nhắn từ một trong các cố vấn luận án tiến sĩ của tôi, Lawrence Katz. Katz không bảo tôi tại sao Summers quan tâm tới công trình của tôi, mặc dù sau đó tôi phát hiện ra là Katz đã biết từ đầu.

Tôi ngồi tại phòng đợi bên ngoài văn phòng của Summers. Sau một chút chậm trễ, nguyên bộ trưởng Ngân khố Hoa Kì, nguyên chủ tịch Harvard, và là người đoạt một số giải thưởng lớn nhất về kinh tế học, cho gọi tôi vào.

Summers bắt đầu cuộc gặp bằng việc đọc bài báo của tôi về tác động của sự phân biệt chủng tộc đối với Obama mà thư kí của ông đã in ra cho ông. Summers là một người đọc tốc độ. Khi đọc, ông thỉnh thoảng thè lưỡi ra bên phải, trong khi mắt đảo nhanh trái phải và xuống dưới trang. Cách Summers đọc một bài báo khoa học xã hội làm tôi liên tưởng đến một nghệ sĩ piano đang trình diễn một bản sonata. Ông tập trung đến độ như quên hết mọi thứ khác. Trong chưa đầy 5 phút, ông đọc xong bài viết 30 trang của tôi.

"Anh nói rằng các tìm kiếm Google từ 'nigger' chứng tỏ sự phân biệt chủng tộc," Summers nói. "Điều đó có vẻ đúng. Nó dự báo những nơi Obama ít được ủng hộ hơn Kerry. Hay lắm. Chúng ta thực sự có thể xem Obama và Kerry là như nhau không?"

"Họ được các nhà khoa học chính trị xếp là có ý thức hệ tương tự," tôi trả lời. "Cũng không có mối tương quan giữa phân biệt chủng tộc và các thay đổi trong việc bỏ phiếu ở Hạ viện. Kết quả vẫn mạnh ngay cả khi chúng ta thêm các nhóm đối chứng cho các tiêu chí nhân khẩu học, đi nhà thờ, và sở hữu súng." Đây là cách các nhà kinh tế học chúng tôi nói chuyện. Tôi đã trở nên sôi nổi.

Summers dừng lại và liếc nhìn tôi. Ông liếc nhanh chiếc ti vi trong văn phòng, nó đang mở kênh CNBC, rồi lại chằm chằm nhìn tôi, rồi nhìn ti vi, rồi trở lại nhìn tôi. "OK, tôi thích bài báo này," Summers nói. "Anh đang nghiên cứu gì khác nữa không?"

60 phút tiếp theo có lẽ là thời gian đầu óc tôi cảm thấy hồ hởi nhất trong đời. Summers và tôi bàn về lãi suất và lạm phát, cảnh sát và tội phạm, kinh doanh và từ thiện. Có một lí do khiến nhiều người gặp Summers là mê ngay. Trong đời mình, tôi đã may mắn được nói chuyện với một số người thông minh khác thường; Summers khiến tôi ấn tượng nhất. Ông luôn bận rộn với các ý tưởng, hơn mọi thứ khác; đó dường như là điều thường khiến ông gặp rắc rối. Ông phải từ chức chủ tịch Harvard sau khi cho rằng việc nữ giới ít xuất hiện trong các ngành khoa học có thể là do nam giới có độ biến thiên về IQ cao hơn. Nếu phát hiện một ý tưởng hay, Summers thường nói ra, ngay cả nếu nó làm một số người nghe khó chịu.

Thế là cuộc họp thâm lạm lịch làm việc của chúng tôi nửa giờ. Cả hai say sưa đàm đạo, nhưng tôi vẫn không hiểu tại sao mình lại có mặt ở đó, không biết khi nào nên đi, và cũng chẳng biết làm sao mình biết được khi nào nên đi. Đến thời điểm này, tôi có cảm giác rằng chính Summers cũng có thể đã quên lí do tại sao ông lại tổ chức cuộc gặp gỡ này.

Và rồi ông nêu câu hỏi triệu đô—hoặc có thể là cả tỉ đô. "Anh nghĩ là anh có thể dự báo thị trường chứng khoán với dữ liệu này không?"

À ha. Thì ra đây là lí do tại sao Summers cho mời tôi tới văn phòng nói chuyện.

Summers không phải là người đầu tiên hỏi tôi câu hỏi đặc biệt này. Cha tôi trước đến giờ nói chung là ủng hộ những sở thích nghiên cứu

phi truyền thống của tôi. Nhưng có một lần chính ông đã mở đầu chủ đề. "Phân biệt chủng tộc, ngược đãi trẻ em, nạo phá thai," ông nói. "Con không thể kiếm được đồng nào bằng chuyên môn này của con sao?" Bạn bè và những người khác trong gia đình cũng nêu chủ đề đó. Đồng nghiệp và người lạ trên Internet cũng vậy. Mọi người dường như muốn biết tôi có thể dùng các tìm kiếm Google—hoặc Dữ Liệu Lớn khác—để chọn mã chứng khoán hay không. Bây giờ là tới nguyên bộ trưởng Ngân khố Hoa Kì. Trường hợp này nghiêm túc hơn.

Vậy liệu các nguồn Dữ Liệu Lớn mới có thể dự báo thành công chứng khoán đi đường nào không? Câu trả lời ngắn gọn là không.

Trong các chương trước chúng ta đã thảo luận 4 sức mạnh của Dữ Liệu Lớn. Chương này chỉ nói về các hạn chế của Dữ Liệu Lớn—cả những gì ta không thể làm và, đôi khi, ta không được làm với nó. Tôi xin bắt đầu bằng cách kể câu chuyện thất bại của Summers và chính tôi trong nỗ lực đánh bại thị trường.

Trong Chương 3, chúng ta đã biết dữ liệu mới rất có thể mang lại những nguồn lợi lớn khi hoạt động nghiên cứu hiện có của một lĩnh vực nào đó còn yếu. Một sự thật đáng buồn của thế giới này là việc thu thập những hiểu biết mới về phân biệt chủng tộc, ngược đãi trẻ em, hoặc nạo phá thai sẽ dễ dàng hơn rất nhiều so với việc thu thập một hiểu biết mới và có ích về hoạt động của một doanh nghiệp. Đó là vì các nguồn lực khổng lồ đã được dồn hết cho việc tìm kiếm ngay cả cái lợi thế mong manh nhất trong việc đo lường hoạt động kinh doanh. Cạnh tranh về tài chính rất dữ dội. Đó đã là một bất lợi đối với chúng tôi rồi.

Summers, một người không quen hào hứng nói về trí thông minh của người khác, đã chắc chắn rằng các quỹ phòng hộ đã đi trước chúng tôi rất xa. Tôi hoàn toàn bị thu hút trong suốt cuộc nói chuyện bởi muốn biết mức độ tôn trọng họ của ông, và cũng để biết sẽ có bao nhiêu đề xuất của tôi bị ông tin là các quỹ phòng hộ đã làm rất tốt rồi. Tôi tự hào chia sẻ với ông một thuật toán tôi đã nghĩ ra cho phép tôi nhận thêm nhiều dữ liệu Google Trends hoàn chỉnh. Ông nói thuật toán đó rất thông minh. Khi tôi hỏi ông liệu Renaissance, một quỹ phòng hộ định

lượng, có nghĩ ra được thuật toán đó không, ông khúc khích cười và nói, "Có, dĩ nhiên họ đã tìm ra cái đó."

Khó khăn trong việc theo kịp các quỹ phòng hộ không phải là vấn đề cơ bản duy nhất mà Summers và tôi đụng phải trong khi dùng các bộ dữ liệu lớn mới để đánh bại thị trường.

## Lời nguyền của tính đa chiều

Giả sử chiến lược dự báo thị trường chứng khoán của bạn là tìm một đồng xu may mắn—được tìm qua các thử nghiệm cẩn thận. Đây là phương pháp của bạn: Bạn đánh số 1,000 đồng xu—1 đến 1,000. Mỗi buổi sáng, trong suốt 2 năm, bạn tung từng đồng xu, ghi lại kết quả ngửa hay sấp, và rồi ghi chú chỉ số Standard & Poor's Index lên hay xuống ngày hôm đó. Bạn nghiên cứu cẩn thận toàn bộ dữ liệu. Eureka! Bạn đã tìm thấy điều gì đó. Hóa ra 70.3% lần Đồng xu 391 ngửa thì S&P Index lên. Mối quan hệ đó có nhiều khả năng có ý nghĩa thống kê. Bạn đã tìm thấy đồng xu may mắn!

Chỉ tung Đồng xu 391 mỗi buổi sáng và mua chứng khoán bất cứ khi nào đồng xu ngửa. Những tháng ngày mì gói đã qua rồi. Đồng xu 391 là chiếc vé đến với cuộc đời tươi đẹp!

Hoặc không.

Bạn đã trở thành nạn nhân của "lời nguyền của tính đa chiều" (the curse of dimensionality). Nó có thể tấn công bất cứ khi nào bạn có nhiều biến số—hay "chiều kích" (trong trường hợp này là 1,000 đồng xu)—nhưng không có đủ lượt quan sát (trong trường hợp này là 504 ngày giao dịch suốt 2 năm đó). Có nhiều khả năng một trong những chiều kích đó (Đồng xu 391) sẽ gặp may [và ngẫu nhiên xuất hiện tương quan]. Giảm số biến xuống—tung chỉ 100 đồng xu thôi—khả năng một trong các đồng xu đó gặp may sẽ giảm đi rất nhiều. Tăng số lượt quan sát lên—thử dự báo hành vi của chỉ số S&P Index suốt 20 năm—và các đồng xu sẽ khó mà "may mắn" nhiều lần như thế.

Lời nguyền đa chiều là một vấn đề quan trọng với Dữ Liệu Lớn, vì các bộ dữ liệu mới thường cho ta nhiều biến hơn gấp vạn lần các nguồn

truyền thống—mọi từ ngữ tìm kiếm, mọi hạng mục tweet... Nhiều người tuyên bố có thể dự báo thị trường khi sử dụng một nguồn Dữ Liệu Lớn nào đó, nhưng cuối cùng hóa ra chỉ đang sập bẫy lời nguyền. Tất cả điều họ thực sự làm là tìm cái tương đương với Đồng xu 391 thôi.

Lấy ví dụ một nhóm nhà khoa học máy tính từ Đại học Indiana và Dại học Manchester, họ tuyên bố có thể dự báo các thị trường sẽ đi theo đường nào dựa vào những gì người ta trao đổi trên tweet. Họ xây dựng một thuật toán để mã hóa các trạng thái từng ngày của thế giới dựa trên các tweet. Họ dùng các kĩ thuật tương tự kĩ thuật phân tích cảm xúc được thảo luận ở Chương 3. Tuy nhiên, họ mã hóa không chỉ một trạng thái mà nhiều trạng thái—vui, giận, tử tế, và nhiều trạng thái khác. Họ phát hiện rằng đa số các tweet thể hiện sự bình tĩnh—ví dụ như "tôi cảm thấy bình tĩnh"—dự báo rằng Dow Jones Industrial Average có khả năng đi lên 6 ngày sau đó. Một quỹ phòng hộ đã được sáng lập để khai thác các phát hiện của họ.

Có vấn đề gì ở đây?

Vấn đề cơ bản là họ đã kiểm tra quá nhiều chỉ báo. Và nếu bạn kiểm tra đủ nhiều chỉ báo, chỉ bằng may rủi ngẫu nhiên, một trong các chỉ báo đó sẽ có ý nghĩa thống kê. Họ đã kiểm tra khá nhiều cảm xúc. Và họ kiểm tra mỗi cảm xúc 1 ngày trước, 2 ngày trước, 3 ngày trước, và lên đến 7 ngày trước hành vi thị trường chứng khoán mà họ đang cố gắng dự báo. Và tất cả các biến số này được dùng để thử giải thích chỉ một vài tháng (tức là vài trăm ngày, hay vài trăm quan sát) Dow Jones lên và xuống mà thôi.

Sự bình tĩnh 6 ngày trước không phải là một chỉ báo chuẩn xác của thị trường chứng khoán. Sự bình tĩnh 6 ngày trước là Dữ Liệu Lớn tương đương với Đồng xu 391 giả tưởng của chúng ta. Quỹ phòng hộ dựa trên tweet đã đóng cửa chỉ sau 1 tháng do kết quả không mấy sáng sủa.

Các quỹ phòng hộ cố gắng canh thị trường dựa vào các tweet không phải là các quỹ duy nhất chiến đấu với lời nguyền của tính đa chiều. Cũng ở tình thế đó là vô số nhà khoa học đã cố gắng tìm kiếm các chìa khóa di truyền cho câu hỏi chúng ta là ai.

Nhờ dự án Bản đồ gene người, bây giờ ta có thể thu thập và phân tích ADN đầy đủ của mọi người. Tiềm năng của dự án này dường như rất lớn.

Có thể chúng ta tìm được gene gây ra bệnh tâm thần phân liệt. Có thể chúng ta phát hiện được gene gây ra các bệnh Alzheimer, Parkinson và ALS. Có thể chúng ta tìm được gene tạo ra... trí thông minh. Liệu có gene giúp bổ sung một mớ điểm IQ không? Có gene tạo ra thiên tài không?

Năm 1998, Robert Plomin, một nhà di truyền học hành vi nổi tiếng, tuyên bố đã tìm ra câu trả lời. Ông thu nhận một bộ dữ liệu bao gồm ADN và IQ của hàng trăm học sinh. Ông so sánh ADN của "thiên tài"—những người có IQ 160 trở lên—với ADN của những người có IQ trung bình.

Ông tìm thấy một khác biệt đầy ấn tượng trong ADN của 2 nhóm này. Nó nằm ở một góc nhỏ của nhiễm sắc thể số 6, một gene mờ nhưng mạnh mẽ có góp phần trong sự chuyển hóa của bộ não. Một phiên bản của gene này, IGF2r, phổ biến gấp 2 lần ở các thiên tài.

"Gene đầu tiên có liên quan đến trí thông minh cao đã được tìm ra," là tiêu đề bài báo trên *New York Times*.

Bạn có thể nghĩ đến nhiều câu hỏi đạo đức mà phát hiện của Plomin đặt ra. Liệu cha mẹ có nên được phép kiểm tra IGF2r của con cái? Liệu họ có nên được phép bỏ một thai nhi có biến thể IQ thấp không? Liệu ta có nên thay đổi mọi người về mặt di truyền để cho họ IQ cao không? IGF2r có tương quan với chủng tộc không? Ta có muốn biết câu trả lời cho câu hỏi đó không? Liệu nghiên cứu về di truyền học IQ có nên được tiếp tục không?

Trước khi các nhà đạo đức sinh học phải xử lí bất cứ câu hỏi gai góc nào ở đây, có một câu hỏi căn bản hơn cho các nhà di truyền học, bao gồm cả Plomin: Liệu kết quả có chính xác không? Có thực sự là IGF2r có thể dự báo IQ không? Có thực sự là các thiên tài có gấp đôi khả năng mang biến thể gene này không?

Không. Một vài năm sau nghiên cứu ban đầu, Plomin tiếp cận một mẫu những người khác—cũng có thông tin ADN và điểm IQ. Lần này, IGF2r không tương quan với IQ. Plomin đã rút lại tuyên bố—dấu hiệu cho thấy ông là một nhà khoa học tốt.

Thực tế, đây là một mô thức chung trong nghiên cứu di truyền học và IQ. Đầu tiên, các nhà khoa học thông báo rằng họ đã phát hiện một biến thể gene dự báo IQ. Sau đó các nhà khoa học thu thập dữ liệu mới và phát hiện khẳng định ban đầu của họ là sai.

Ví dụ, trong một bài báo gần đây, một nhóm nhà khoa học do Christopher Chabris lãnh đạo đã khảo sát 12 tuyên bố nổi tiếng về các biến thể gene liên quan với IQ. Họ khảo sát dữ liệu từ 10 ngàn người. Họ không tìm ra sự tương quan cho bất cứ điều gì trong 12 tuyên bố đó.

Vấn đề với tất cả tuyên bố này là gì? Lời nguyền của tính đa chiều. Bây giờ các nhà khoa học đã biết, bộ gene người khác nhau theo hàng triệu cách. Đơn giản là có quá nhiều gene để kiểm tra.

Nếu kiểm tra đủ nhiều tweet để xem chúng có tương quan với thị trường chứng khoán không, bạn sẽ thấy 1 tweet nào đó tương quan chỉ do tình cờ. Nếu kiểm tra đủ nhiều biến thể gene để xem chúng có tương quan với IQ không, bạn sẽ tìm thấy 1 biến thể tương quan chỉ do tình cờ.

Bạn có thể vượt qua lời nguyền của tính đa chiều như thế nào? Bạn phải có một chút khiêm tốn về công trình của mình và đừng đem lòng yêu kết quả có được. Bạn phải đặt các kết quả này vào các thử nghiệm bổ sung. Ví dụ, trước khi cá cược khoản tiền dành dụm cả đời vào Đồng xu 391, bạn sẽ muốn thấy nó "làm ăn" thế nào trong vài năm tiếp theo. Các nhà khoa học xã hội gọi đây là thử nghiệm "ngoài mẫu" (out-of-sample test). Khảo sát càng nhiều biến, bạn càng phải khiêm tốn. Khảo sát càng nhiều biến, thử nghiệm ngoài mẫu càng phải nghiêm ngặt. Theo dõi mọi thử nghiệm là điều hết sức quan trọng. Sau đó bạn có thể biết chính xác khả năng mình đã trở thành nạn nhân của lời nguyền là bao nhiêu, cũng như mình nên hoài nghi kết quả đến mức độ nào. Điều đó đưa chúng ta trở lại với Summers và tôi. Sau đây là cách chúng tôi cố gắng đánh bại thị trường.

Ý tưởng ban đầu của Summers là dùng các tìm kiếm để dự báo doanh số tương lai của các sản phẩm chủ chốt, ví dụ như iPhone, từ đó soi rọi tương lai giá chứng khoán của công ty, ở đây là Apple. Thực tế là có tồn tại tương quan giữa các tìm kiếm từ "iPhone" và doanh số iPhone. Khi người ta đang Google từ "iPhone" nhiều, bạn có thể cá chắc là nhiều điện thoại đang được bán ra. Tuy nhiên, thông tin này đã được thể hiện qua giá cổ phiếu Apple hiện tại. Rõ ràng, khi có nhiều tìm kiếm từ "iPhone" trên Google, các quỹ phòng hộ cũng đã hiểu rằng iPhone sẽ bán chạy, bất kể họ dùng dữ liệu tìm kiếm hay một nguồn nào khác.

Ý tưởng tiếp theo của Summers là dự báo đầu tư tương lai vào các nước đang phát triển. Nếu một số lượng lớn nhà đầu tư định rót tiền vào các nước như Brazil hoặc Mexico trong tương lai gần, thì giá cổ phiếu các công ty tại các nước này chắc chắn sẽ lên. Có khi ta có thể dự báo tăng vốn đầu tư bằng các tìm kiếm Google chủ chốt—như "đầu tư ở Mexico" hoặc "cơ hội đầu tư ở Brazil." Cách này đưa chúng tôi vào ngõ cụt. Vấn đề là gì? Các tìm kiếm đó quá hiếm hoi. Thay vì tiết lộ các mô thức có ý nghĩa, dữ liệu tìm kiếm này chỉ nhảy tùm lum mà thôi.

Chúng tôi cũng đã thử các tìm kiếm cho từng cổ phiếu riêng lẻ. Có lẽ nếu người ta đang tìm "GOOG" (mã cổ phiếu Google) nghĩa là họ sắp mua Google. Các tìm kiếm này chỉ dự báo là chứng khoán đó sẽ được giao dịch nhiều, nhưng không dự báo được chứng khoán sẽ lên hay xuống. Một hạn chế lớn là các tìm kiếm này không cho ta biết một người có quan tâm đến việc mua hoặc bán chứng khoán đó hay không.

Một ngày kia, tôi hào hứng chỉ cho Summers thấy một ý tưởng mới: Các tìm kiếm quá khứ từ "mua vàng" (buy gold) có vẻ tương quan với các vụ tăng giá vàng tương lai. Summers bảo tôi là tôi nên kiểm tra xem nó còn chính xác hay không. Nó đã không còn hiệu quả, có lẽ vì một quỹ phòng hộ nào đó đã phát hiện mối quan hệ tương tự trước tôi rồi.

Cuối cùng, qua vài tháng, chúng tôi không thấy gì hữu ích. Rõ ràng, nếu tìm dấu hiệu tương quan với tình hình thị trường trong mỗi tỉ từ ngữ tìm kiếm Google, chúng tôi hẳn sẽ thấy một tương quan nào đó, dù là yếu ớt. Nhưng chắc chắn đó sẽ chỉ là Đồng xu 391 may mắn mà thôi.

## Quá đề cao những thứ có thể đo được

Tháng 3/2012, Zoë Chance, giáo sư marketing tại Yale, nhận được một thiết bị đo bước đi nhỏ màu trắng trong hộp thư của cô tại phố New Haven, Connecticut. Cô định nghiên cứu xem thiết bị đo số bước chân đi trong ngày và cho điểm ấy có thể truyền cảm hứng tập thể dục nhiều hơn như thế nào.

Những gì xảy ra tiếp theo, như cô kể lại trong một bài diễn thuyết TEDx, là một cơn ác mộng Dữ Liệu Lớn. Chance quá bị ám ảnh và quá nghiện việc làm tăng các con số đến nỗi bắt đầu đi bộ khắp nơi, từ nhà bếp lên phòng khách, tới phòng ăn, xuống tầng hầm, và cả trong văn phòng nữa. Cô đi bộ lúc sáng sớm, lúc đêm khuya, gần như mọi giờ giấc trong ngày—20 ngàn bước trong một khoảng thời gian 24 giờ bất kì. Cô kiểm tra thiết bị đo bước đi của mình hàng trăm lần mỗi ngày, và phần nhiều thời gian giao tiếp còn lại của cô là với những người dùng thiết bị đo bước đi trên mạng, họ thảo luận các chiến lược để cải thiện thành tích. Cô nhớ đã gắn thiết bị đo bước đi lên đứa con gái 3 tuổi khi nó đang bước, bởi cô đã quá ám ảnh với việc gia tăng điểm số.

Chance bị ám ảnh với việc tối đa hóa con số này đến nỗi quên hết mọi thứ khác. Bà quên lí do tại sao người ta lại muốn có điểm số cao hơn—là để tự động viên tập thể dục, chứ không phải để nhờ con gái bước phụ thêm vài bước. Cô cũng không hoàn thành một nghiên cứu học thuật nào về thiết bị đo bước đi đó cả. Cuối cùng, cô vứt bỏ cái máy sau một đêm kiệt sức vì cố gắng kiếm thêm một số bước đi. Mặc dù nghề của cô là nghiên cứu dựa trên dữ liệu, trải nghiệm đó đã ảnh hưởng sâu sắc đến cô. "Nó khiến tôi hoài nghi liệu việc được tiếp cận thêm dữ liệu bổ sung có luôn là một việc tốt hay không," Chance nói.

Đây là một câu chuyện cực đoan. Nhưng nó chỉ ra một vấn đề tiềm ẩn với những người dùng dữ liệu để ra quyết định. Các con số có thể rất cám dỗ. Ta có thể bị chúng thu hút, và khi đó ta có thể không nhìn thấy những điều quan trọng hơn. Zoë Chance ít nhiều đã không nhìn thấy phần còn lại của cuộc sống.

Ngay cả những say mê nhất thời ít ám ảnh hơn với các con số cũng có thể có những hạn chế. Ta hãy xem xét tình trạng đề cao quá độ việc kiểm tra ở các trường học Mĩ trong Thế kỉ XXI—và phán xét giáo viên dựa trên điểm số của học sinh. Mặc dù ước muốn có được các thước đo khách quan về những gì diễn ra trong lớp học là hoàn toàn hợp lí, có nhiều thứ không dễ gì ghi lại được bằng các con số. Hơn nữa, tất cả những phương pháp kiểm tra đó gây áp lực khiến nhiều giáo viên dạy chỉ để kiểm tra—và tệ hơn nữa. Một số nhỏ, như được chứng minh trong bài báo của Brian Jacob và Steven Levitt, đã gian lận hoàn toàn trong khi thực hiện các bài kiểm tra đó.

Vấn đề là đây: Những thứ ta có thể đo lường thì thường không chính xác là cái mà ta quan tâm. Ta có thể đo lường học sinh làm đúng bao nhiêu câu hỏi đa lựa chọn. Ta không thể dễ dàng đo lường tư duy phản biện, tính tò mò, hoặc sự phát triển cá nhân. Chỉ cố gia tăng một con số riêng lẻ, dễ đo lường—điểm kiểm tra hoặc số bước chân trong một ngày—không phải lúc nào cũng giúp đạt được điều ta thực sự muốn.

Facebook cũng đụng đầu mối nguy hiểm này. Công ty có hàng tấn dữ liệu về cách mọi người dùng website. Dễ dàng để thấy một câu chuyện cụ thể trên News Feed có được like, nhấp chọn, bình luận, hoặc chia sẻ hay không. Nhưng, theo Alex Peysakhovich, một nhà khoa học dữ liệu Facebook mà tôi đã viết chung về các vấn đề này, không có chỉ số nào kể trên có thể đại diện một cách hoàn hảo cho các câu hỏi quan trọng hơn: Trải nghiệm dùng trang này như thế nào? Câu chuyện có liên kết người dùng với bạn bè họ không? Nó có cung cấp thông tin cho họ về thế giới không? Nó có làm họ cười không?

Hoặc như cuộc cách mạng dữ liệu của bóng chày thập niên 1990. Nhiều đội bắt đầu dùng các số thống kê ngày càng khó hiểu chứ không dựa vào những người săn cầu thủ truyền thống để ra quyết định. Thật dễ đo các chỉ số tấn công và ném bóng, nhưng không dễ đo lường các chỉ số phòng thủ, vì vậy cuối cùng một số đội đã đánh giá thấp tầm quan trọng của phòng thủ. Trong quyển *The Signal and the Noise,* Nate Silver ước tính rằng đội Oakland A's, một đội dựa trên dữ liệu được nói đến

trong quyển *Moneyball*, đã mất 8 đến 10 trận thắng mỗi năm giữa thập niên 90 vì phòng thủ cùi bắp.

Giải pháp không phải khi nào cũng là tìm thêm Dữ Liệu Lớn. Ta cần một loại nước sốt đặc biệt để tối đa hiệu quả Dữ Liệu Lớn: sự phán đoán của con người và các khảo sát nhỏ, những thứ mà ta có thể gọi là *dữ liệu nhỏ*. Trong một cuộc phỏng vấn với Silver, Billy Beane, bấy giờ là tổng giám đốc của Oakland A's và là nhân vật chính trong quyển *Moneyball*, nói rằng ông thực sự đã bắt đầu tăng ngân sách cho hoạt động săn cầu thủ truyền thống.

Để lấp các khoảng trống trong vốn dữ liệu khổng lồ của mình, Facebook cũng phải dùng một phương pháp cũ: hỏi xem người ta nghĩ gì. Mỗi ngày khi họ tải lên News Feed, hàng trăm người dùng Facebook được hỏi về những câu chuyện họ xem ở đó. Nói cách khác, các bộ dữ liệu được tự động thu thập của Facebook (lượt thích, nhấp chuột, bình luận) được bổ sung bởi dữ liệu nhỏ hơn ("Bạn muốn xem bài đăng này trong News Feed của bạn không? Tại sao?"). Vâng, ngay cả một tổ chức Dữ Liệu Lớn thành công ngoạn mục như Facebook đôi khi cũng sử dụng nguồn thông tin bị chê nhiều trong sách này: khảo sát nhỏ.

Thực vậy, vì cần có dữ liệu nhỏ bổ sung cho cơ sở dữ liệu chính—những bộ thu thập lớn các cú nhấp chuột, like, và bài đăng—các đội dữ liệu Facebook rất khác so với suy nghĩ của bạn. Facebook thuê các nhà tâm lí học xã hội, các nhà nhân chủng học, và các nhà xã hội học chính xác là để tìm những thứ mà các con số đã bỏ sót.

Một số nhà giáo dục cũng đang cảnh giác hơn với các điểm mù trong Dữ Liệu Lớn. Hiện có một nỗ lực quy mô quốc gia đang dần lan rộng nhằm bổ sung cho các thử nghiệm đại trà bằng dữ liệu nhỏ. Các cuộc khảo sát sinh viên đã tăng lên nhanh chóng. Các cuộc khảo sát phụ huynh và dự giờ giáo viên cũng vậy.

"Các học khu (school district) nhận ra là họ không nên chỉ tập trung vào điểm kiểm tra," Thomas Kane, một giáo sư ngành giáo dục tại Harvard, nói. Một nghiên cứu 3 năm của Bill & Melinda Gates Foundation xác minh giá trị về giáo dục của cả hai loại dữ liệu lớn và

nhỏ. Các tác giả phân tích xem các mô hình dựa trên điểm kiểm tra, các cuộc khảo sát sinh viên, hoặc dự giờ giáo viên có tối ưu hay không trong việc đo lường giáo viên nào cải thiện tốt nhất việc học của học sinh. Khi họ kết hợp 3 thước đo lại với nhau vào thành 1 điểm số hỗn hợp, họ nhận được kết quả tốt nhất. Báo cáo kết luận, "Mỗi thước đo đều bổ sung một điều gì đó có giá trị."

Thực vậy, chính vì biết rằng nhiều hoạt động Dữ Liệu Lớn dùng dữ liệu nhỏ để lấp các lỗ hổng nên tôi đã có mặt tại Ocala, Florida, để gặp Jeff Seder. Chắc bạn còn nhớ, ông ấy là chuyên gia về ngựa học trường Harvard, người đã dùng các bài học từ một bộ dữ liệu lớn để dự báo thành công của chú ngựa American Pharoah.

Sau khi chia sẻ tất cả các file máy tính và công thức toán học với tôi, Seder thú nhận rằng ông còn một vũ khí khác: Patty Murray.

Murray, cũng như Seder, rất thông minh và có năng lực xuất chúng. Bà có bằng đại học Bryn Mawr. Bà cũng bỏ New York để sống ở thôn quê. "Tôi thích ngựa hơn người," Murray thừa nhận. Nhưng Murray hơi truyền thống hơn một chút trong các phương pháp đánh giá ngựa. Cũng như nhiều nhân viên chọn ngựa khác, bà đích thân khảo sát ngựa, xem nó đi ra sao, kiểm tra các vết sẹo và vết thâm, và chất vấn chủ ngựa.

Sau đó Murray cộng tác với Seder để chọn các con ngựa cuối cùng mà họ muốn giới thiệu. Murray phát hiện các vấn đề của ngựa, các vấn đề mà dữ liệu của Seder bỏ sót—mặc dù đó là bộ dữ liệu sáng tạo và quan trọng nhất từng được thu thập về ngựa từ trước đến giờ.

Tôi dự báo một cuộc cách mạng dựa trên những điều mà Dữ Liệu Lớn tiết lộ. Nhưng điều này không có nghĩa là ta có thể chỉ cần ném dữ liệu vào bất cứ câu hỏi nào. Dữ Liệu Lớn không loại trừ hoàn toàn các phương pháp khác, những phương pháp mà con người đã phát triển qua nhiều thiên niên kỉ để hiểu thế giới. Tất cả các phương pháp ấy đều bổ sung cho nhau.

# CHƯƠNG 8

## Thêm dữ liệu, thêm vấn đề? Điều ta không nên làm.

Đôi khi, sức mạnh của Dữ Liệu Lớn ấn tượng đến mức đáng sợ. Nó nêu lên các câu hỏi về đạo đức.

### Sự nguy hiểm của các công ty được trao sức mạnh

Gần đây, 3 nhà kinh tế học—Oded Netzer, Alain Lemaire (Đại học Columbia), và Michal Herzenstein (Đại học Delaware)—tìm cách dự báo khả năng một người đi vay sẽ trả lại tiền. Các học giả này sử dụng dữ liệu từ Prosper, một website cho vay đồng đẳng (peer-to-peer lending). Người vay tiềm năng viết một đoạn ngắn mô tả tại sao họ cần một khoản vay và lí do cho thấy họ sẽ trả khoản vay đó; dựa theo đó, người cho vay tiềm năng quyết định rót tiền cho họ hay không. Nói chung, khoảng 13% người vay xù nợ.

Thì ra ngôn ngữ mà người vay tiềm năng dùng là một chỉ báo mạnh cho khả năng trả nợ vay của họ. Và đó vẫn là một chỉ báo quan trọng ngay cả khi đối chứng chỉ báo ấy với các thông tin liên quan khác mà người cho vay có thể nhận được về những người vay tiềm năng, bao gồm cả điểm tín dụng và thu nhập.

Dưới đây là 10 cụm từ mà các nhà nghiên cứu thấy thường được dùng khi xin vay nợ. 5 cụm từ trong số đó tương quan tích cực với việc trả lại nợ vay. 5 cụm từ kia tương quan tiêu cực với việc trả lại nợ vay. Nói cách khác, 5 từ có khuynh hướng được dùng bởi những người bạn

có thể tin tưởng, 5 từ kia là dấu hiệu không thể tin tưởng. Xem thử bạn có đoán đúng không nhé.

| | | |
|---|---|---|
| **God**<br>Chúa | **lower interest rate**<br>lãi suất thấp hơn | **after-tax**<br>sau thuế |
| **promise**<br>hứa | **will pay**<br>sẽ trả | **hospital**<br>bệnh viện |
| **debt-free**<br>không mắc nợ | **graduate**<br>tốt nghiệp | **minimum payment**<br>khoản trả tối thiểu |
| **thank you**<br>cảm ơn | | |

Bạn có thể nghĩ—hoặc ít nhất là hi vọng—rằng một người lịch sự đã hứa hẹn thì có nhiều khả năng sẽ trả nợ vay nhất. Nhưng thực tế không như vậy. Dữ liệu cho thấy, loại người này có khả năng trả nợ dưới mức trung bình.

Đây là các cụm từ được nhóm theo khả năng trả nợ vay.

### TỪ NGỮ DÙNG TRONG HỒ SƠ XIN VAY CỦA NGƯỜI NHIỀU KHẢ NĂNG SẼ TRẢ NỢ

| | | |
|---|---|---|
| **debt-free**<br>không mắc nợ | **after-tax**<br>sau thuế | **graduate**<br>tốt nghiệp |
| **lower interest rate**<br>lãi suất thấp hơn | **minimum payment**<br>khoản trả tối thiểu | |

### TỪ NGỮ DÙNG TRONG HỒ SƠ XIN VAY CỦA NGƯỜI NHIỀU KHẢ NĂNG SẼ KHÔNG TRẢ NỢ

| | | |
|---|---|---|
| **God**<br>Chúa | **will pay**<br>sẽ trả | **hospital**<br>bệnh viện |
| **promise**<br>hứa | **thank you**<br>cảm ơn | |

Trước khi thảo luận những hàm ý đạo đức của nghiên cứu này, chúng ta hãy suy nghĩ kĩ, với sự trợ giúp của các tác giả nghiên cứu trên, xem kết quả tiết lộ điều gì về mọi người. Chúng ta nên hiểu gì về từ ngữ trong các nhóm đó?

Trước tiên, chúng ta hãy xem xét loại ngôn ngữ báo hiệu một người nào đó nhiều khả năng sẽ trả nợ. Các cụm từ như "lãi suất thấp hơn" hoặc "sau thuế" cho thấy một trình độ hiểu biết tài chính nhất định ở phía người vay, vì vậy có lẽ không ngạc nhiên khi các cụm từ đó tương quan với khả năng trả nợ của họ. Hơn nữa, nếu người đó nói về các thành tựu tích cực như là "tốt nghiệp" đại học và "không mắc nợ," họ cũng có nhiều khả năng trả nợ.

Bây giờ hãy xem xét loại ngôn ngữ báo hiệu một người nào đó nhiều khả năng sẽ xù nợ. Thông thường, nếu người ta nói với bạn họ sẽ trả nợ cho bạn, họ sẽ không trả. Lời hứa càng quả quyết, họ càng có khả năng thất hứa. Nếu người ta viết "tôi hứa tôi sẽ trả, vậy vì Chúa, hãy giúp tôi," tức là họ thuộc nhóm ít có khả năng trả nợ nhất. Khi kêu gọi lòng tốt của bạn—giải thích rằng họ cần tiền bởi vì họ có người thân nằm "bệnh viện"—cũng có nghĩa là họ nhiều khả năng sẽ không trả. Thực vậy, đề cập một thành viên trong gia đình—chồng, vợ, con trai, con gái, mẹ, hoặc cha—là một dấu hiệu người ta sẽ không trả nợ vay. Một từ khác báo hiệu khả năng xù nợ là "giải thích," có nghĩa là nếu người ta đang cố gắng giải thích tại sao họ sẽ có thể trả nợ, họ nhiều khả năng sẽ xù luôn.

Các tác giả không lí luận tại sao cảm ơn lại là bằng chứng cho thấy đối tượng nhiều khả năng sẽ xù nợ.

Tóm lại, theo các nhà nghiên cứu này, việc người vay cung cấp kế hoạch chi tiết lộ trình trả nợ và đề cập các cam kết đã giữ trong quá khứ là bằng chứng họ sẽ trả nợ. Hứa hẹn và kêu gọi lòng tốt của người cho vay là dấu hiệu rõ ràng họ sẽ xù nợ. Không cần quan tâm lí do—bất chấp những điều mà nghiên cứu này cho ta thấy sự thật về bản chất con người, rằng hứa hẹn là dấu hiệu cho thấy người ta sẽ không giữ lời—các học giả đã cho ta một thông tin cực kì giá trị trong việc dự báo xù nợ. Ai mà đề cập đến Chúa thì khả năng xù nợ gấp 2.2 lần. Đây là một trong số

các chỉ báo cao nhất rằng người ta sẽ không trả nợ.

Nhưng các tác giả cũng tin tưởng nghiên cứu của họ nêu lên các câu hỏi đạo đức. Mặc dù đây chỉ là nghiên cứu học thuật, một số công ty cũng công bố rằng họ có sử dụng dữ liệu trực tuyến trong việc duyệt các khoản cho vay. Điều này chấp nhận được không? Bạn có muốn sống trong một thế giới mà các công ty dùng từ ngữ bạn viết để dự báo bạn sẽ trả nợ hay không? Nghe thật sởn gai ốc—và đáng sợ.

Người cần vay trong tương lai gần có thể phải lo lắng không chỉ về lịch sử tài chính mà còn về hoạt động của mình trên mạng nữa. Và họ có thể bị phán quyết dựa trên các yếu tố có vẻ vô lí—ví dụ liệu họ có dùng cụm từ "Cảm ơn" hoặc cầu "Chúa" không chẳng hạn. Hơn nữa, lỡ có trường hợp một phụ nữ có nhu cầu hợp lí muốn giúp cô em gái đang nằm bệnh viện và sẽ chắc chắn trả lại nợ vay sau đó thì sao? Thật tệ khi trừng phạt cô chỉ vì đa phần những người đi vay lấy lí do để trả tiền thuốc hay bị vạch mặt là nói dối. Một thế giới hoạt động theo cách này quả thật là đáng sợ.

Đây là câu hỏi đạo đức: Phải chăng các công ty có quyền phán xét sự phù hợp của chúng ta đối với dịch vụ của họ dựa trên các tiêu chí trừu tượng, nhưng có tính dự báo về mặt thống kê và không liên quan trực tiếp với các dịch vụ đó?

Bỏ thế giới tài chính lại phía sau, ta hãy xét các vấn đề lớn hơn, ví dụ như hoạt động tuyển dụng. Nhà tuyển dụng ngày càng tăng cường sục sạo mạng xã hội khi cân nhắc ứng viên. Điều đó có thể không dấy lên các câu hỏi đạo đức nếu nhà tuyển dụng đang tìm bằng chứng cho thấy ứng viên đã nói xấu sếp cũ, hay làm lộ bí mật của công ty cũ. Thậm chí ta vẫn có thể biện hộ được cho việc từ chối thuê những người mà các bài đăng Facebook hoặc Instagram của họ cho thấy họ dùng rượu quá nhiều. Nhưng chuyện sẽ ra sao nếu nhà tuyển dụng tìm thấy một chỉ báo dường như vô hại nhưng có tương quan với điều gì đó họ quan tâm?

Các nhà nghiên cứu tại Đại học Cambridge và Microsoft cung cấp cho 58 ngàn người dùng Facebook Mĩ một loạt các bài kiểm tra về nhân cách và trí tuệ. Họ phát hiện rằng các like trên Facebook thường tương

quan với IQ, mức độ hướng ngoại, và sự tận tâm. Ví dụ, những người thích Mozart, các cơn giông, và khoai tây chiên xoắn trên Facebook thường có IQ cao hơn. Những người thích mô-tô Harley-Davidson, nhóm nhạc đồng quê Lady Antebellum, hoặc trang "I Love Being a Mom" thường có IQ thấp hơn. Một vài tương quan trong số này có thể là hậu quả lời nguyền của tính đa chiều. Nếu bạn kiểm nghiệm đủ nhiều tiêu chí, một số tiêu chí sẽ tương quan ngẫu nhiên. Nhưng một số sở thích có thể tương quan một cách hợp lí với IQ.

Tuy nhiên, có vẻ không công bằng nếu một người thông minh, chỉ vì thích xe Harley, lại không thể nhận được một công việc tương xứng với kĩ năng—tất cả là bởi anh không biết là mình đang phát các tín hiệu cho thấy anh có trí tuệ thấp.

Công bằng mà nói, đây không phải là vấn đề mới hoàn toàn. Từ lâu người ta đã bị phán xét bởi các yếu tố không trực tiếp liên quan đến khả năng làm việc—độ chặt của cái bắt tay, hay mức độ gọn gàng trong cách ăn mặc. Nhưng cách mạng dữ liệu mang đến một mối nguy: Khi ngày càng nhiều thứ trong cuộc sống được định lượng, các phán xét dựa trên những yếu tố đại diện này cũng ngày càng khó hiểu và phiền nhiễu hơn. Việc dự báo tốt hơn có thể dẫn đến phân biệt đối xử tinh vi hơn và độc ác hơn.

Dữ liệu tốt hơn cũng có thể dẫn đến một hình thức phân biệt đối xử khác, điều mà các nhà kinh tế học gọi là phân biệt giá. Các doanh nghiệp thường cố gắng tính giá chuẩn nhất cho hàng hóa hoặc dịch vụ của mình. Lí tưởng là tính được mức giá tối đa mà khách hàng sẵn sàng chi trả. Theo cách này, họ sẽ trích rút được lợi nhuận tối đa.

Hầu hết các doanh nghiệp thường chọn một cái giá cho tất cả mọi người. Nhưng đôi khi họ ý thức được rằng các thành viên của một nhóm nào đó sẽ chịu chi cao hơn. Đây là lí do các rạp chiếu phim tính giá cao hơn đối với khách hàng trung niên—độ tuổi có thu nhập đạt đỉnh—so với sinh viên hoặc người cao tuổi. Đó cũng là lí do các hãng hàng không thường tính giá cao hơn đối với những người mua vé vào phút chót. Họ phân biệt giá.

Dữ Liệu Lớn giúp các doanh nghiệp giỏi hơn trong việc biết khách hàng nào sẽ sẵn sàng chi tiền—và thế là họ tính giá cắt cổ một số nhóm người nhất định. Optimal Decisions Group là một công ty tiên phong trong việc ứng dụng khoa học dữ liệu để dự báo người tiêu dùng sẵn sàng chi bao nhiêu cho bảo hiểm. Họ đã làm điều đó bằng cách nào? Họ dùng một phương pháp mà chúng ta đã thảo luận trong sách này. Họ tìm các khách hàng cũ giống các khách hàng đang định mua bảo hiểm nhất—và xem tiền bảo hiểm họ sẵn sàng đóng là bao nhiêu. Nói cách khác, họ tìm các song trùng. Phương pháp song trùng rất thú vị nếu nó giúp ta dự báo xem một cầu thủ có trở lại thời kì huy hoàng của anh ta hay không. Phương pháp song trùng rất tuyệt vời nếu nó giúp ta chữa lành bệnh cho một người nào đó. Nhưng nếu phương pháp song trùng lại giúp một công ty trích rút tận đồng xu cuối cùng của bạn thì sao? Thế thì không êm lắm. Chú em xem tiền như rác của tôi sẽ có quyền khiếu nại nếu chú ấy bị tính giá cao hơn ông anh kẹo kéo này.

Đánh bạc là lĩnh vực trong đó khả năng phóng to về khách hàng chất chứa rất nhiều nguy cơ. Các sòng bạc lớn đang dùng một thứ tương tự phương pháp song trùng để hiểu rõ hơn khách hàng của họ. Mục đích của họ? Để trích rút lợi nhuận tối đa—để bảo đảm tiền của bạn chui vào két của họ nhiều nhất có thể.

Đây là cách nó hoạt động: Các sòng bạc tin rằng mỗi con bạc đều có một "điểm nhói." Đó là số tiền thua đủ để khiến con bạc cạch mặt sòng bạc một thời gian dài. Ví dụ, giả sử "điểm nhói" của Helen là 3,000 USD. Điều này có nghĩa là nếu cô ta thua 3,000 USD, bạn sẽ mất một khách hàng, có lẽ là trong vài tuần hoặc vài tháng. Nếu Helen thua 2,999 USD, cô sẽ không vui. Nói cho cùng, có ai mà thích thua tiền đâu, phải không? Nhưng cô sẽ không quá mất tinh thần đến độ sẽ không trở lại vào tối mai.

Hãy tưởng tượng rằng bạn đang quản lí một sòng bạc. Và tưởng tượng là Helen đã đến để chơi máy đánh bạc. Kết quả tối ưu là gì? Rõ ràng, bạn muốn Helen đến càng gần "điểm nhói" càng tốt, nhưng không vượt qua nó. Bạn muốn cô ta thua 2,999 USD, đủ để bạn có lợi nhuận lớn mà không quá đáng đến nỗi cô sẽ không sớm trở lại chơi tiếp.

Bạn có thể làm điều này như thế nào? Vâng, có những cách khiến Helen dừng chơi khi cô ta đã thua một khoản tiền nhất định. Bạn có thể mời cô dùng bữa miễn phí chẳng hạn. Chỉ cần bạn mời đủ hấp dẫn, Helen sẽ rời máy đánh bạc để đi ăn.

Nhưng có một thách thức lớn với cách này. Làm sao bạn biết "điểm nhói" của Helen? Vấn đề là, người ta có "điểm nhói" khác nhau. Với Helen, đó là 3,000 USD. Với John, có thể là 2,000 USD. Với Ben, có thể là 26,000 USD. Nếu bạn thuyết phục Helen dừng đánh bạc khi cô mới thua 2,000 USD, tức là bạn đã bỏ phí một phần lợi nhuận. Nếu bạn chờ quá lâu—sau khi Helen đã thua 3,000 USD—tức là bạn đã làm mất khách trong một khoảng thời gian. Hơn nữa, Helen có thể không muốn nói bạn biết điểm nhói của mình. Thậm chí chính cô có thể cũng không biết.

Vậy bạn nên làm gì? Nếu đọc tới đoạn này trong sách, có thể bạn đã đoán được câu trả lời. Bạn dùng khoa học dữ liệu. Bạn có thể biết mọi thứ về một số khách hàng của bạn—tuổi tác, giới tính, mã vùng, và hành vi đánh bạc. Và, từ hành vi đánh bạc đó—những lần thắng, thua, đến, và đi của họ—bạn ước lượng được "điểm nhói" của họ.

Bạn thu thập tất cả thông tin bạn biết về Helen và tìm kiếm các con bạc tương tự như cô—những người giống cô, không nhiều thì ít. Sau đó bạn tính ra họ có thể chịu đựng được bao nhiêu. Con số đó nhiều khả năng tương đương với điểm nhói của Helen. Thực tế, đây là những gì sòng bạc Harrah's đã làm. Một công ty Dữ Liệu Lớn tên Terabyte trợ giúp họ.

Scott Gnau, tổng giám đốc Terabyte, đã giải thích trong quyển sách xuất sắc *Super Crunchers* những gì các nhà quản lí sòng bạc làm khi họ thấy một khách quen đang tiến sát điểm nhói: "Họ bước ra và nói, 'Tôi thấy anh hôm nay đang xui rồi. Tôi biết anh thích món bít tết của chúng tôi. Đây, tôi muốn mời anh đưa bà xã đi dùng bữa chiều ngay bây giờ, chúng tôi bao.'"

Nghe có vẻ hào phóng hết sức: mời ăn bít tết miễn phí. Nhưng thực sự thì có tính toán cả. Sòng bạc chỉ đang cố kéo khách hàng dừng chơi trước khi họ thua nhiều đến độ sẽ ra đi một khoảng thời gian dài. Nói

cách khác, mục tiêu là dùng phân tích dữ liệu phức tạp để cố rút càng nhiều tiền từ khách hàng càng tốt trong dài hạn.

Chúng ta có quyền lo sợ rằng việc tận dụng ngày càng tốt dữ liệu trên mạng sẽ trao cho các sòng bạc, công ty bảo hiểm, người cho vay, và các công ty khác quá nhiều quyền năng đối với chúng ta.

Mặt khác, Dữ Liệu Lớn cũng đã và đang giúp người tiêu dùng bụp lại các doanh nghiệp chặt chém hoặc bán các sản phẩm kém chất lượng.

Một vũ khí quan trọng là các website đăng tải các đánh giá về nhà hàng cũng như các dịch vụ khác—Yelp là một ví dụ. Một nghiên cứu gần đây của nhà kinh tế học Michael Luca ở Đại học Harvard cho thấy số phận các doanh nghiệp nằm dưới quyền sinh sát của các đánh giá trên Yelp. So sánh các bình luận đó với dữ liệu doanh số ở tiểu bang Washington, ông phát hiện rằng mất 1 sao trên Yelp sẽ khiến doanh thu của nhà hàng rớt từ 5 đến 9%.

Người tiêu dùng cũng nhận được sự trợ giúp trong trận chiến với các doanh nghiệp từ các trang so sánh giá—như Kayak và Booking.com. Như đã được thảo luận trong quyển *Freakonomics*, khi một trang mạng bắt đầu công bố sự chênh lệch giá bảo hiểm nhân thọ có kì hạn giữa các công ty khác nhau, các mức giá này đột ngột giảm. Nếu một công ty bảo hiểm mà chặt chém, khách hàng sẽ biết và mua từ công ty khác. Tổng các khoản tiết kiệm cho tất cả người tiêu dùng là bao nhiêu? 1 tỉ đô la mỗi năm.

Nói cách khác, dữ liệu trên Internet có thể cho các doanh nghiệp biết nên tránh khách hàng nào và nên khai thác khách hàng nào. Nó cũng nói cho khách hàng biết nên tránh doanh nghiệp nào và doanh nghiệp nào đang cố bòn rút họ. Dữ Liệu Lớn đến nay đã giúp cả hai phía trong cuộc chiến giữa người tiêu dùng và các công ty. Chúng ta phải bảo đảm đó là một cuộc chiến công bằng.

## Sự nguy hiểm của các chính quyền được trao sức mạnh

Khi bạn trai cũ xuất hiện tại một buổi tiệc sinh nhật, Adriana Donato biết rằng anh ta đau khổ. Cô biết rằng anh ta điên. Cô biết rằng anh ta đã

vật vã vì trầm cảm. Khi anh ta mời cô đi dạo một vòng xe, có một điều mà cô sinh viên ngành động vật học 20 tuổi Donato đã không biết. Cô không biết bạn trai cũ của cô, James Stoneham 22 tuổi, vừa bỏ ra 3 tuần tìm kiếm thông tin về cách ám sát một người, về luật liên quan đến tội danh giết người, và lẫn trong đó thỉnh thoảng là các tìm kiếm về Donato.

Nếu mà cô biết điều này, có lẽ cô đã không bước lên xe. Có lẽ, tối hôm đó cô đã không bị đâm cho đến chết.

Trong phim *Minority Report*, các nhà tâm linh cộng tác với các sở cảnh sát để ngăn chặn tội ác trước khi chúng xảy ra. Liệu có nên cấp Dữ Liệu Lớn cho các sở cảnh sát để ngăn chặn tội ác trước khi chúng xảy ra không? Lẽ ra ít nhất Donato phải được cảnh báo về các tìm kiếm nguy hiểm kia của bạn trai cũ chứ? Lẽ ra cảnh sát phải thẩm vấn Stoneham trước chứ?

Trước tiên, phải công nhận là ngày càng có nhiều bằng chứng cho thấy các tìm kiếm Google liên quan đến hành vi phạm tội có tương quan với hành vi phạm tội trong thực tế. Christine Ma-Kellams, Flora Or, Ji Hyun Baek, và Ichiro Kawachi đã chỉ ra rằng các tìm kiếm Google liên quan đến tự tử tương quan mạnh với tỉ lệ tự tử cấp tiểu bang. Ngoài ra, Evan Soltas và tôi đã chỉ ra rằng các tìm kiếm tiêu cực về người Hồi giáo hàng tuần—như "Tôi ghét người Hồi giáo" hoặc "diệt người Hồi giáo"—tương quan với các tội ác chống người Hồi giáo tuần đó. Nếu càng nhiều người tìm kiếm những thứ thể hiện rõ họ đang muốn làm một điều gì đó, thì càng nhiều người sẽ thực sự làm điều ấy.

Vậy ta nên làm gì với thông tin này? Một ý tưởng đơn giản, khá ít tranh cãi: Ta có thể dùng dữ liệu cấp vùng để phân bổ nguồn lực. Nếu một thành phố tăng mạnh các tìm kiếm liên quan đến tự tử, ta có thể nâng cao nhận thức về vấn đề tự tử tại thành phố này. Chính quyền thành phố hoặc các tổ chức phi lợi nhuận có thể chạy quảng cáo giải thích người ta có thể nhận sự giúp đỡ ở đâu. Tương tự, nếu một thành phố tăng mạnh các tìm kiếm "diệt người Hồi giáo," các sở cảnh sát có lẽ cần thay đổi cách tuần tra đường phố. Họ có thể phái thêm nhân viên đến bảo vệ nhà thờ Hồi giáo địa phương.

Nhưng có một bước rất khó thực hiện: theo dõi các cá nhân trước khi có tội ác xảy ra. Điều này, trước hết, có vẻ xâm phạm quyền riêng tư. Có một sự khác biệt lớn về đạo đức giữa việc chính quyền có dữ liệu tìm kiếm của hàng ngàn hoặc hàng trăm ngàn người và việc sở cảnh sát có dữ liệu tìm kiếm của một cá nhân. Có một sự khác biệt lớn về đạo đức giữa việc bảo vệ một nhà thờ địa phương và việc lục soát nhà một ai đó. Có một sự khác biệt lớn về đạo đức giữa việc quảng bá việc ngăn ngừa tự tử và việc giam giữ ai đó trong một bệnh viện tâm thần ngược lại ý muốn của họ.

Tuy nhiên, lí do phải thật thận trọng khi dùng dữ liệu cấp cá nhân còn vượt cả phạm vi đạo đức. Cũng có một lí do về mặt dữ liệu nữa. Đó là một sự cách biệt lớn giữa việc thử dự báo các hoạt động của một thành phố đến việc thử dự báo các hoạt động của một cá nhân.

Chúng ta hãy trở lại vụ tự tử. Mỗi tháng, có khoảng 3.5 triệu tìm kiếm Google tại Mĩ liên quan đến tự tử, với đa số trong đó ám chỉ ý định tự tử—các tìm kiếm như là "tự sát," "tự tử," và "cách tự tử." Nói cách khác, mỗi tháng, cứ 100 người Mĩ thì có hơn 1 tìm kiếm liên quan đến tự tử. Điều này làm ta chợt nhớ đến câu nói của triết gia Friedrich Nietzsche: "Ý nghĩ tự tử là một điều an ủi tuyệt vời: Bằng cách đó, người ta trải qua được nhiều đêm tăm tối." Dữ liệu tìm kiếm Google chứng tỏ điều đó đúng, ý nghĩ tự tử thực sự khá phổ biến. Tuy nhiên, mỗi tháng, có chưa tới 4 ngàn vụ tự tử tại Mĩ. Ý định tự tử vô cùng phổ biến. Nhưng tự tử thực sự thì không. Vì vậy, sẽ không hợp lí lắm nếu để cảnh sát thường xuyên xuất hiện tại cửa nhà của tất cả những người đã có lần tìm trên mạng những từ khóa cho thấy họ đang muốn nã một phát tung não mình ra—nếu thế, cảnh sát sẽ không có thời gian cho bất cứ việc gì khác nữa.

Hay thử xét các tìm kiếm căm ghét người Hồi giáo đầy ác ý kia. Năm 2015, có khoảng 12,000 lượt tìm kiếm cụm từ "giết Hồi giáo" tại Mĩ. Có 12 vụ ám sát người Hồi giáo được cho là tội ác do thù ghét. Rõ ràng, tuyệt đại đa số những người thực hiện các tìm kiếm kinh khủng này không thực hiện hành vi tương ứng trong thực tế.

Có một cách giải thích bằng toán học cho sự khác nhau giữa dự báo hành vi của một cá nhân với dự báo hành vi trong một thành phố. Đây là một thí nghiệm tâm tưởng đơn giản. Hãy giả định như sau:

- Có 1 triệu người và 1 nhà thờ Hồi giáo trong thành phố.
- Nếu một người không tìm kiếm "diệt người Hồi giáo," thì chỉ có 1/100,000,000 khả năng người đó sẽ tấn công nhà thờ.
- Nếu một người có tìm kiếm "diệt người Hồi giáo," khả năng này tăng mạnh. Khả năng tấn công của người này là 1/10,000.
- Sự thù ghét người Hồi giáo tăng vút, số tìm kiếm "giết Hồi giáo" tăng từ 100 lên 1,000.

Trong tình huống này, toán học chỉ ra rằng khả năng nhà thờ Hồi giáo bị tấn công tăng lên gấp 5 lần, từ khoảng 2% lên 10%. Nhưng các khả năng một cá nhân đã tìm kiếm từ khóa "giết Hồi giáo" thực sự tấn công nhà thờ vẫn chỉ là 1/10,000.

Phản ứng thích hợp trong tình huống này không phải là tống giam tất cả những người tìm kiếm từ khóa "giết Hồi giáo." Cũng không phải là cho cảnh sát đến viếng nhà họ. Khả năng bất cứ cá nhân nào trong số những người này sẽ phạm tội là rất nhỏ. Tuy nhiên, câu trả lời thích hợp sẽ là bảo vệ nhà thờ, bây giờ nó có 10% khả năng bị tấn công.

Rõ ràng, nhiều tìm kiếm đáng sợ không hề dẫn tới các hành vi khủng khiếp.

Nói là vậy, ít nhất về mặt lí thuyết thì có thể có một số loại tìm kiếm cho xác suất khá cao sẽ có một hành vi khủng khiếp tiếp nối sau đó. Ít nhất về mặt lí thuyết thì, ví dụ, trong tương lai các nhà khoa học dữ liệu có thể xây dựng một mô hình phát hiện được rằng các tìm kiếm của Stoneham liên quan đến Donato là đủ có ý nghĩa để bắt đầu lo lắng quan tâm.

Năm 2014, có khoảng 6,000 tìm kiếm chính xác cụm từ "how to kill your girlfriend" (cách giết bạn gái) và 400 vụ ám sát bạn gái. Nếu tất cả những tên giết người này đã tìm kiếm chính xác cụm từ này, điều đó sẽ có nghĩa là 1/15 số người tìm kiếm "cách giết bạn gái" đã làm thật. Dĩ

nhiên, rất nhiều—có lẽ là hầu hết—những người ám sát bạn gái đã không tìm kiếm chính xác cụm từ này. Điều này sẽ có nghĩa là xác suất tìm kiếm này dẫn đến giết người là thấp hơn, có lẽ thấp hơn rất nhiều.

Nhưng nếu các nhà khoa học dữ liệu có thể xây dựng một mô hình chỉ ra rằng mối đe dọa đối với một cá nhân cụ thể, cứ cho là 1/100 đi, thì có thể chúng ta sẽ muốn làm điều gì đó với thông tin này. Ít nhất, người bị đe dọa có thể có quyền được thông báo rằng có 1/100 khả năng cô ta sẽ bị ám sát bởi một người cụ thể.

Tuy nhiên, nói chung, phải thật thận trọng khi dùng dữ liệu tìm kiếm để dự báo tội phạm ở mức độ cá nhân. Dữ liệu rõ ràng cho ta biết rằng có rất, rất nhiều các tìm kiếm đáng kinh sợ hiếm khi dẫn đến các hành động khủng khiếp. Và cho đến nay, chưa có bằng chứng cho thấy chính quyền có thể dự báo một hành động khủng khiếp cụ thể, với xác suất cao, chỉ từ việc khảo sát các tìm kiếm này. Vậy chúng ta phải thật thận trọng về việc cho phép chính quyền can thiệp vào ở mức độ cá nhân dựa trên dữ liệu tìm kiếm. Đây không chỉ vì lí do đạo đức hoặc pháp lí. Nó còn vì các lí do về khoa học dữ liệu nữa—ít nhất là tính đến thời điểm này.

# Kết Luận

## Bao nhiêu người đọc hết sách?

Sau khi kí hợp đồng sách, tôi đã có một tầm nhìn rõ ràng về cách cấu trúc quyển sách. Chắc bạn cũng nhớ, lúc mới bắt đầu, tôi mô tả một cảnh tại bàn ăn lễ Tạ ơn của gia đình tôi. Các thành viên gia đình tranh luận về sự tỉnh táo của tôi và cố hiểu tại sao đã 33 rồi mà tôi có vẻ vẫn không tìm được cô nàng hợp ý.

Bấy giờ, quyển sách này coi như đã sẵn đường tiến triển đến kết luận rồi. Tôi sẽ gặp và cưới cô gái. Tuy nhiên, tốt hơn hết là tôi sẽ dùng Dữ Liệu Lớn để gặp cô gái hợp ý. Có thể tôi sẽ len lỏi trong các mẩu thông tin từ suốt quá trình tán tỉnh. Rồi câu chuyện sẽ tập trung hết lại ở phần kết luận, phần này sẽ mô tả ngày đám cưới của tôi và, một công đôi việc, tôi sẽ dùng sách này làm bức thư tình gởi cô vợ mới cưới luôn.

Thật chẳng may, cuộc sống đã không khớp với tầm nhìn của tôi. Tự giam mình trong nhà và tránh xa nhân thế để viết sách có lẽ không có ích gì lắm cho đời sống tình cảm lãng mạn của tôi. Và tôi, lạy trời, vẫn cần phải kiếm một cô vợ. Quan trọng hơn, tôi cần một kết luận mới.

Tôi mải mê nghiên cứu nhiều quyển sách tâm đắc để cố tìm kiếm các yếu tố làm nên một kết luận tuyệt vời. Tôi kết luận là các kết luận hay nhất sẽ đưa lên bề mặt một điểm quan trọng mà trước đến giờ vẫn nằm đó, lượn lờ ngay dưới bề mặt. Với quyển sách này, điểm lớn đó là đây: Khoa học xã hội đang trở thành một ngành khoa học thực sự. Và ngành khoa học mới mẻ này đã sẵn sàng để cải thiện cuộc sống của chúng ta.

Ở đầu Phần II, tôi thảo luận về bài phê bình Sigmund Freud của Karl Popper. Tôi thấy, Popper không nghĩ thế giới quan chập mạch của Freud là khoa học. Nhưng tôi không đề cập điều gì về bài phê bình của Popper cả. Nó thực ra rộng hơn nhiều chứ không chỉ là một cuộc tấn công vào Freud. Popper không nghĩ *bất cứ* nhà khoa học xã hội nào thực sự là khoa học cả. Đơn giản là Popper không ấn tượng với tính nghiêm ngặt của những gì mà những người được gọi là nhà khoa học này đang làm.

Điều gì thúc đẩy cuộc thập tự chinh của Popper? Khi ông trao đổi với những trí thức tinh hoa thời đại ông—các nhà vật lí giỏi nhất, các nhà sử học giỏi nhất, các nhà tâm lí học giỏi nhất—Popper nhận thấy một sự khác biệt đáng kinh ngạc. Khi các nhà vật lí nói, Popper tin ở những gì họ đang làm. Dĩ nhiên, đôi khi họ cũng phạm sai lầm. Dĩ nhiên, đôi khi họ cũng bị lừa bởi các thiên kiến tiềm thức của mình. Nhưng các nhà vật lí đang tham gia một tiến trình giúp khám phá những chân lí sâu xa về thế giới, với đỉnh cao là Thuyết tương đối của Einstein. Trái lại, khi các nhà khoa học xã hội nổi tiếng nhất thế giới nói, Popper nghĩ là mình đang phải nghe một mớ từ ngữ vô nghĩa đao to búa lớn.

Popper hầu như không phải là người duy nhất có sự phân biệt này. Gần như mọi người đều đồng ý rằng các nhà vật lí, sinh học, hóa học là các nhà khoa học thực sự. Họ dùng các thí nghiệm khắt khe để tìm hiểu xem thế giới vật chất hoạt động thế nào. Trái lại, nhiều người nghĩ rằng các nhà kinh tế học, xã hội học, và tâm lí học là các nhà khoa học tào lao, quăng ra toàn những biệt ngữ vớ vẩn cho có việc để làm.

Trước đây thì đúng như vậy, nhưng cuộc cách mạng Dữ Liệu Lớn đã thay đổi điều đó. Nếu mà bây giờ Karl Popper còn sống và tham dự buổi trình bày của Raj Chetty, Jesse Shapiro, Esther Duflo, hoặc của chính tôi (xin được nổ chút xíu), tôi chắc chắn là ông sẽ không phản ứng như trước nữa. Thật tình mà nói, khả năng ông sẽ chất vấn xem các nhà lí thuyết chuỗi xuất sắc ngày nay có thực sự khoa học hay chỉ đang chơi trò thể dục trí não chiều theo đam mê của mình còn nhiều hơn là khả năng ông sẽ chất vấn dân khoa học xã hội chúng tôi.

Nếu một bộ phim bạo lực chiếu ở một thành phố, tội phạm tăng hay

giảm? Nếu nhiều người tiếp xúc một quảng cáo nào đó hơn, liệu có nhiều người dùng sản phẩm hơn không? Nếu một đội bóng chày thắng khi một chàng trai ở tuổi 20, liệu anh ta có chắc chắn mê đội bóng đó khi anh ta 40 tuổi không? Đây toàn là những câu hỏi rõ ràng với câu trả lời có-hoặc-không rõ ràng. Và trong các núi dữ liệu chân thật đó, chúng ta có thể tìm thấy các câu trả lời.

Đây đích thị là khoa học, không phải giả khoa học.

Nhưng điều này không có nghĩa là cách mạng khoa học xã hội sẽ hình thành dưới dạng các quy luật giản đơn, bất hủ.

Marvin Minsky, nhà khoa học quá cố của viện MIT—một trong những người đầu tiên nghiên cứu tiềm năng của trí tuệ nhân tạo—cho rằng tâm lí học đã chệch hướng do cố sao chép vật lí học. Vật lí học đã thành công trong việc tìm kiếm các quy luật đơn giản có thể áp dụng mọi lúc mọi nơi.

Bộ não con người, Minsky nói, có thể không bị chi phối bởi các quy luật như thế. Trái lại, bộ não chắc chắn là một hệ thống giải pháp phức tạp—một bộ phận sửa lỗi sai ở các bộ phận khác. Nền kinh tế và hệ thống chính trị có thể cũng phức tạp tương tự.

Vì lí do này, cách mạng khoa học xã hội không thể hình thành dưới dạng các công thức ngắn gọn, như kiểu $E = MC^2$. Thực vậy, nếu ai đó tuyên bố khởi phát một cuộc cách mạng khoa học xã hội dựa trên một công thức ngắn gọn, bạn nên hoài nghi là vừa.

Cuộc cách mạng đó, trái lại, sẽ đến từng chút một, từng nghiên cứu một, từng phát hiện một. Dần dần, ta sẽ hiểu biết hơn về các hệ thống phức tạp của trí óc và xã hội con người.

***

Kết luận hay thì phải tóm lược được nội dung, nhưng nó còn phải chỉ ra con đường đến với nhiều thứ hơn trong tương lai.

Với quyển sách này, việc đó dễ. Các bộ dữ liệu tôi đã thảo luận trong đây mang tính cách mạng, nhưng hầu như chưa được khám phá đầy đủ. Còn quá nhiều thứ để học hỏi. Thành thật mà nói, tuyệt đại đa số giới

học thuật đã phớt lờ sự bùng nổ dữ liệu do thời đại kĩ thuật số gây ra. Các nhà nghiên cứu tình dục nổi tiếng nhất thế giới bám chặt với những điều đã được thử và được xem là đúng. Họ hỏi về ham muốn của vài trăm đối tượng; họ không hỏi các website như P***Hub để lấy dữ liệu. Các nhà ngôn ngữ học nổi tiếng nhất thế giới phân tích các văn bản riêng lẻ; họ phớt lờ phần lớn các mô thức được tiết lộ trong hàng tỉ quyển sách. Các phương pháp được đem dạy cho nghiên cứu sinh ngành tâm lí học, khoa học chính trị, và xã hội học hầu hết chưa được cuộc cách mạng số chạm tới. Lãnh thổ rộng lớn và hầu như chưa được khám phá do sự bùng nổ dữ liệu mở ra đã được bỏ lại cho một số nhỏ các giáo sư cấp tiến, các nghiên cứu sinh nổi loạn, và những người nghiên cứu vì niềm vui.

Điều đó sẽ thay đổi.

Với mỗi ý tưởng tôi đã bàn trong sách này, có hàng trăm ý tưởng quan trọng không kém đang chờ được xử lí. Nghiên cứu được thảo luận ở đây chỉ là chóp của chóp tảng băng trôi, một vết xước trên vết xước của bề mặt mà thôi.

Vậy còn điều gì khác đang đến?

Xin giới thiệu một hướng mở rộng triệt để phương pháp đã được dùng cho một trong các nghiên cứu y tế thành công nhất mọi thời đại.

Giữa Thế kỉ XIX, John Snow, một bác sĩ người Anh, muốn tìm hiểu nguyên nhân của một cuộc bùng nổ dịch tả đang hoành hành ở London.

Ý tưởng của ông thế này: Ông vẽ bản đồ mọi ca dịch tả trong thành phố. Từ đó, ông phát hiện bệnh phần lớn tập trung quanh một cái bơm nước. Điều này ám chỉ bệnh lan truyền qua nước nhiễm trùng, chứng minh niềm tin truyền thống thời bấy giờ—rằng bệnh truyền qua không khí bẩn—là sai.

Dữ Liệu Lớn—và khả năng phóng to—làm cho nghiên cứu kiểu này khá dễ thực hiện. Với bất cứ bệnh gì, ta có thể khám phá dữ liệu tìm kiếm Google hoặc dữ liệu sức khỏe số khác. Ta có thể tìm xem có xó xỉnh nhỏ bé nào trên thế giới mà sự lan truyền bệnh này cao bất thường hoặc

thấp bất thường không. Sau đó ta có thể tìm hiểu xem các nơi này có gì chung. Có gì trong không khí chăng? Trong nước? Trong các tiêu chuẩn xã hội?

Ta có thể làm cách này cho chứng đau nửa đầu. Ta có thể làm cách này cho bệnh sỏi thận. Ta có thể làm cách này cho chứng lo âu, rồi trầm cảm, rồi Alzheimer, rồi ung thư tụy, rồi huyết áp cao, rồi đau lưng, rồi táo bón và chảy máu mũi. Ta có thể làm cách này cho mọi thứ. Bản phân tích mà Snow đã từng làm, ta có thể làm được 400 lần (một việc mà khi đang viết dòng này, tôi đã bắt đầu làm rồi.)

Ta có thể gọi cách này—lấy một phương pháp đơn giản và dùng Dữ Liệu Lớn để thực hiện phương pháp đó hàng trăm lần trong một thời gian ngắn—là nhân bản khoa học (science at scale). Vâng, các ngành khoa học xã hội và hành vi chắc chắn sẽ nhân bản. Phương pháp phóng to theo tình trạng sức khỏe sẽ giúp các ngành khoa học này nhân bản. Còn một thứ khác cũng giúp các ngành khoa học có thể nhân bản: thử nghiệm A/B. Chúng ta đã thảo luận về thử nghiệm A/B trong bối cảnh các doanh nghiệp khiến người dùng nhấp chuột vào các tiêu đề và quảng cáo—và đây là ứng dụng nổi bật nhất của phương pháp đó. Nhưng thử nghiệm A/B có thể được dùng để khám phá những thứ căn bản—và có giá trị về mặt xã hội—hơn là chỉ để nghĩ ra mấy mũi tên khiến người ta nhấp chuột vào quảng cáo.

Benjamin F. Jones, một nhà kinh tế học tại Northwestern, đang thử dùng thử nghiệm A/B để giúp trẻ em học tốt hơn. Ông đã giúp tạo ra nền tảng EDU STAR, cho phép các trường học thử nghiệm ngẫu nhiên các giáo án khác nhau.

Nhiều công ty đang ở trong ngành kinh doanh phần mềm giáo dục. Với EDU STAR, học sinh đăng nhập một máy tính và ngẫu nhiên tiếp xúc với các giáo án khác nhau. Sau đó các em làm những bài kiểm tra ngắn để xem mình đã hiểu bài đến đâu. Nói cách khác, trường học sẽ biết phần mềm nào giúp học sinh nắm bắt kiến thức hiệu quả nhất.

Hiện giờ, cũng như tất cả các hệ thống thử nghiệm A/B xuất sắc khác, EDU STAR đang tạo ra những kết quả đáng kinh ngạc. Một giáo

án mà nhiều nhà giáo dục rất hào hứng có phần mềm sử dụng các trò chơi để giúp dạy học sinh toán phân số. Dĩ nhiên, nếu bạn biến toán học thành một trò chơi, học sinh sẽ vui hơn, học nhiều hơn, và làm bài kiểm tra tốt hơn. Đúng không? Sai. Các học sinh được dạy phân số thông qua trò chơi có kết quả kiểm tra kém hơn các học sinh học phân số theo cách chuẩn mực hơn.

Việc làm cho trẻ học nhiều hơn là một hướng ứng dụng hấp dẫn và hữu ích về mặt xã hội của thử nghiệm A/B—một công cụ mà Thung lũng Silicon đã đi tiên phong, vốn để khiến mọi người nhấp chuột vào nhiều quảng cáo hơn. Việc khiến mọi người ngủ nhiều hơn cũng hấp dẫn và hữu ích không kém.

Người Mĩ trung bình ngủ 6.7 giờ mỗi đêm. Hầu hết người Mĩ muốn ngủ nhiều hơn. Nhưng đến 11 giờ đêm rồi, và chương trình *SportsCenter* thì đang phát sóng, YouTube thì đang vẫy gọi. Vậy chuyện ngủ nghê thì cứ từ từ. Jawbone, một công ty thiết bị đeo (wearable) có hàng trăm ngàn khách hàng, đã thực hiện hàng ngàn thử nghiệm để cố tìm ra những kiểu can thiệp khiến người dùng làm điều họ muốn làm: đi ngủ sớm hơn.

Jawbone đã ghi một bàn thắng lớn khi dùng một mục tiêu có 2 nhánh. Thứ nhất, yêu cầu khách hàng trung thành với một mục tiêu không quá tham vọng. Gởi cho họ một tin nhắn như thế này: "Hình như bạn đã ngủ không được nhiều suốt 3 ngày qua. Tại sao bạn không đặt mục tiêu đi ngủ trước 11:30 tối nay nhỉ? Chúng tôi biết bình thường bạn dậy lúc 8 giờ sáng." Sau đó người dùng sẽ có một lựa chọn: nhấp chuột lên "Tôi tham gia."

Thứ hai, khi đến 10:30, Jawbone sẽ gởi một tin nhắn khác: "Ta đã nhất trí là bạn sẽ ngủ lúc 11:30. Bây giờ đã là 10:30. Tại sao không bắt đầu luôn từ bây giờ nhỉ?"

Jawbone thấy chiến lược này đã giúp tăng thêm 23 phút ngủ. Họ không khiến khách hàng thực sự đi ngủ lúc 10:30, nhưng đã khiến khách hàng đi ngủ sớm hơn.

Dĩ nhiên, mọi bộ phận của chiến lược này phải được tối ưu hóa qua nhiều thí nghiệm. Khởi động mục tiêu ban đầu quá sớm—yêu cầu người

dùng đi ngủ trước 11 giờ đêm—thì ít người sẽ theo. Yêu cầu người dùng đi ngủ lúc nửa đêm thì kết quả nhận được chẳng đáng bao nhiêu.

Jawbone dùng thử nghiệm A/B để tìm thứ tương đương với mũi tên Google chỉ sang phải đối với giấc ngủ. Nhưng thay vì kiếm thêm vài cú nhấp chuột cho các đối tác quảng cáo của Google, họ đã tạo ra thêm vài phút nghỉ ngơi cho những người Mĩ đã mệt lả.

Thực vậy, toàn bộ lĩnh vực tâm lí có thể dùng các công cụ của Thung lũng Silicon để cải tiến mạnh mẽ thí nghiệm của mình. Tôi đang háo hức chờ đón bài báo tâm lí đầu tiên mà, thay vì nêu chi tiết vài cuộc thí nghiệm được thực hiện với vài sinh viên, sẽ chỉ ra các kết quả từ 1 ngàn thử nghiệm A/B nhanh.

Thời đại mà giới nghiên cứu bỏ hàng tháng trời vào việc tuyển mộ một nhóm nhỏ sinh viên để thực hiện một thử nghiệm đơn lẻ sẽ kết thúc. Thay vì thế, giới nghiên cứu sẽ dùng dữ liệu kĩ thuật số để thử nghiệm vài trăm hoặc vài ngàn ý tưởng trong chỉ vài giây. Ta sẽ có thể biết nhiều hơn trong thời gian ngắn hơn rất nhiều.

Dữ liệu văn bản sẽ dạy cho ta rất nhiều thứ. Các ý tưởng lan truyền như thế nào? Các từ mới hình thành ra sao? Các từ biến mất như thế nào? Các câu chuyện đùa hình thành như thế nào? Tại sao một số từ thì mắc cười còn số khác thì không? Các phương ngữ phát triển như thế nào? Tôi cá là, trong vòng 20 năm nữa, chúng ta sẽ có những hiểu biết sâu sắc về tất cả các câu hỏi này.

Tôi nghĩ chúng ta có thể nghĩ đến việc tận dụng hành vi trực tuyến của trẻ em—nặc danh một cách thích hợp—để bổ sung cho các bài kiểm tra truyền thống, nhằm biết các em đang học hỏi và phát triển ra sao. Chính tả của các em có chuẩn không? Các em đang có dấu hiệu mắc chứng khó đọc không? Các em có đang phát triển các sở thích trưởng thành và trí tuệ hơn không? Các em có bạn bè không? Có các đầu mối cho tất cả câu hỏi này trong hàng ngàn cú gõ phím mà mỗi đứa trẻ thực hiện hàng ngày.

Và có một lĩnh vực khác, không kém phần quan trọng, nơi có rất nhiều hiểu biết nội quan đang hình thành.

Trong bài hát "Shattered" của ban nhạc Rolling Stones, Mick Jagger mô tả tất cả những gì làm cho Thành phố New York— Quả Táo Lớn—lại lôi cuốn đến vậy. Tiếng cười. Niềm vui. Cô đơn. Chuột. Rệp. Tự hào. Tham lam. Những con người mặc túi giấy. Nhưng Jagger dành nhiều từ ngữ nhất cho điều khiến thành phố này thực sự đặc biệt: "sex và sex và sex và sex."

Giống như Quả Táo Lớn, Dữ Liệu Lớn cũng vậy. Nhờ cách mạng kĩ thuật số, những hiểu biết đang dần đến với lĩnh vực sức khỏe. Ngủ. Học. Tâm lí. Ngôn ngữ. Và sex và sex và sex và sex.

Một câu hỏi tôi hiện đang khám phá: Có bao nhiêu biến số về xu hướng tính dục? Ta thường nghĩ ai đó là đồng tính hoặc bình thường. Nhưng xu hướng tính dục rõ ràng phức tạp hơn thế. Người đồng tính và người bình thường cũng có nhiều loại—một số nam giới thích "các cô tóc vàng," số khác thì thích "các cô tóc đen." Liệu các sở thích này có mạnh bằng các sở thích giới tính không? Một câu hỏi khác tôi đang nghiên cứu: Các sở thích tính dục đến từ đâu? Giống như ta có thể tìm ra các năm quan trọng quyết định sự hâm mộ bóng chày hoặc các quan điểm chính trị, bây giờ ta có thể tìm các năm quan trọng quyết định sở thích tính dục của người trưởng thành. Để biết các câu trả lời này, chắc bạn sẽ phải mua quyển sách kế tiếp của tôi, dự định đặt tiêu đề là *Mọi người vẫn nói dối—Everybody (Still) Lies.*

Sự ra đời của sản phẩm khiêu dâm—và dữ liệu đi cùng với nó—là một sự phát triển mang tính cách mạng trong khoa học về tính dục con người.

Mất rất nhiều thời gian để các ngành khoa học tự nhiên bắt đầu thay đổi cuộc sống chúng ta—để tạo ra penicillin, vệ tinh, và máy tính. Có thể cũng mất nhiều thời gian để Dữ Liệu Lớn đưa các ngành khoa học xã hội và hành vi đến những tiến bộ quan trọng trong cách chúng ta yêu thương, học hành, và sinh sống. Nhưng tôi tin các tiến bộ như thế đang đến gần. Tôi hi vọng bạn thấy ít nhất là những nét chính của các bước tiến ấy từ quyển sách này. Thực vậy, tôi hi vọng rằng một vài người trong số các bạn đang đọc sách này sẽ giúp tạo ra những tiến bộ ấy.

***

Muốn viết kết luận hay, tác giả phải nghĩ về lí do tại sao mình lại viết quyển sách này. Mục tiêu anh ta đang cố gắng đạt được là gì?

Tôi nghĩ lí do lớn nhất tôi viết quyển sách này là vì một trong những trải nghiệm có tính quyết định nhất trong đời tôi. Bạn biết đấy, cách đây hơn 10 năm, *Freakonomics* ra đời. Quyển sách bán chạy đáng kinh ngạc đó mô tả nghiên cứu của Steven Levitt, một nhà kinh tế học của Đại học Chicago từng đoạt nhiều giải thưởng, được đề cập thường xuyên trong sách này. Levitt là một "nhà kinh tế học kì cục," dường như có thể dùng dữ liệu để trả lời bất cứ câu hỏi nào mà đầu óc khác thường của ông có thể nghĩ ra: Đô vật sumo có gian lận không? Thí sinh các trò chơi truyền hình có phân biệt đối xử không? Người môi giới nhà đất có đưa cho bạn những phi vụ giống như các phi vụ họ tìm cho chính họ không?

Tôi lúc đó vừa mới ra trường, chuyên về triết học, hầu như trong đầu không có bất kì một ý tưởng nào về chuyện tôi muốn làm gì với đời mình cả. Sau khi đọc *Freakonomics,* tôi đã biết mình sẽ làm gì. Tôi muốn làm điều Steven Levitt đã làm. Tôi muốn mải mê nghiên cứu hàng núi dữ liệu để tìm ra cách thế giới *thực sự* vận hành. Tôi quyết định sẽ theo ông và lấy bằng tiến sĩ kinh tế học.

Rất nhiều thứ đã thay đổi trong 12 năm làm việc. 2 nghiên cứu của Levitt bị phát hiện có các lỗi mã hóa. Levitt nói một số điều lệch lạc về mặt chính trị về sự ấm lên toàn cầu. *Freakonomics* không còn được ưa chuộng trong giới trí thức nữa.

Nhưng tôi nghĩ, ngoại trừ một vài lỗi, những năm đó đã rất ưu ái một luận điểm lớn hơn mà Levitt đang cố chỉ ra. Levitt nói với chúng ta rằng sự kết hợp của tính tò mò, tính sáng tạo, và dữ liệu có thể cải thiện mạnh mẽ hiểu biết của chúng ta về thế giới. Có những câu chuyện ẩn trong dữ liệu đã sẵn sàng để được kể, và điều này đã được chứng minh là đúng nhiều lần.

Tôi hi vọng quyển sách này có thể ảnh hưởng đến người khác giống như quyển *Freakonomics* đã ảnh hưởng đến tôi. Tôi hi vọng có một người trẻ tuổi nào đó đang đọc sách này ngay bây giờ và còn hơi mơ hồ về điều

muốn làm với cuộc sống của mình. Nếu bạn có một ít kĩ năng thống kê, dồi dào tính sáng tạo và tính tò mò, hãy gia nhập ngành phân tích dữ liệu.

Quyển sách này (chẳng biết tôi có nói liều lắm không), thực ra, có thể được xem là *Freakonomics* cấp độ tiếp theo. Một khác biệt lớn giữa các nghiên cứu thảo luận trong *Freakonomics* và các nghiên cứu thảo luận trong sách này là ở độ tham vọng. Vào thập niên 1990, khi Levitt làm nên tên tuổi, không có sẵn nhiều dữ liệu như thế. Levitt tự hào khi đi theo các câu hỏi khác thường, nơi mà dữ liệu đã hiện hữu. Hầu như ông phớt lờ các câu hỏi lớn, nơi mà dữ liệu không hiện hữu. Tuy nhiên, ngày nay, với rất nhiều dữ liệu có sẵn về mọi chủ đề, sẽ rất hợp lí nếu bạn muốn lần theo các câu hỏi lớn, sâu sắc, chạm đến cốt lõi của bản chất con người.

Tương lai của ngành phân tích dữ liệu thật sáng sủa. Tôi tin chắc rằng Kinsey kế tiếp sẽ là nhà khoa học dữ liệu. Foucault kế tiếp sẽ là nhà khoa học dữ liệu. Freud kế tiếp sẽ là nhà khoa học dữ liệu. Marx kế tiếp sẽ là nhà khoa học dữ liệu. Salk kế tiếp cũng rất có thể là một nhà khoa học dữ liệu.

***

Thôi, nói chung đó là những cố gắng của tôi để làm một số điều mà một kết luận hay thường có. Nhưng tôi nhận ra rằng các kết luận xuất sắc chứa đựng nhiều điều hơn gấp bội. Hơn nhiều lắm. Một kết luận xuất sắc phải vui mà lạ. Nó phải cảm động. Một kết luận xuất sắc còn phải sâu sắc, hài hước, và buồn. Một kết luận xuất sắc, nói gọn trong một hoặc hai câu, phải có một điểm tóm lược mọi thứ đã xuất hiện trước đó, và cả mọi thứ đang dần xuất hiện nữa. Nó cũng phải có một điểm mới mẻ, độc đáo—một cú ngoặt. Một quyển sách hay phải kết thúc bằng một vụ nổ lớn, thông minh, khôi hài, khiêu khích.

Bây giờ có lẽ là thời điểm tốt để nói một chút về quá trình viết sách của tôi. Tôi không phải là một người thích viết dông dài. Quyển sách này chỉ vào khoảng 75,000 từ, hơi ngắn cho một chủ đề phong phú như thế này.

Nhưng những gì tôi thiếu về độ rộng, tôi đã bù đắp bằng sự say mê. Tôi bỏ ra 5 tháng và viết 47 bản nháp cho bài báo nói về tình dục đầu tiên trên *New York Times*. Bài ấy có 2,000 từ. Một số chương trong sách này tôi viết nháp 60 lần. Tôi có thể bỏ hàng giờ tìm đúng từ cho một câu ở dưới cước chú.

Gần hết năm qua tôi sống như một kẻ ẩn dật. Chỉ tôi và chiếc máy tính. Tôi sống trong khu vực nhộn nhịp nhất Thành phố New York, nhưng hầu như không bao giờ đi ra ngoài. Theo tôi, đây là kiệt tác của mình, ý tưởng tốt nhất mà tôi có trong đời. Và tôi sẵn sàng hi sinh bất cứ điều gì để làm cho đúng. Tôi muốn có khả năng biện hộ cho từng từ trong quyển sách này. Điện thoại của tôi đầy những email tôi quên phản hồi, những thư mời điện tử tôi không bao giờ mở, và những tin nhắn hẹn hò Bumble tôi đã phớt lờ.[1]

Sau 13 tháng làm việc tích cực, cuối cùng tôi có thể gởi đi một bản thảo gần như hoàn chỉnh. Tuy nhiên, có một phần bị thiếu: kết luận.

Tôi giải thích với người biên tập của mình, Denise, rằng phần đó có thể mất thêm ít tháng nữa. Tôi nói với cô ấy rất có thể là 6 tháng. Phần kết luận, theo tôi, là phần quan trọng nhất của quyển sách. Và tôi chỉ mới bắt đầu tìm hiểu điều gì sẽ làm nên một kết luận xuất sắc. Chẳng cần phải nói, Denise không hài lòng.

Rồi một ngày kia, một người bạn gởi email cho tôi đề tài nghiên cứu của Jordan Ellenberg. Ellenberg, một nhà toán học tại Đại học Wisconsin, tò mò về việc bao nhiêu người thực sự đọc hết một quyển sách. Ông nghĩ ra một phương pháp tài tình để kiểm tra ý tưởng đó bằng Dữ Liệu Lớn. Amazon báo cáo có bao nhiêu người trích dẫn các đoạn khác nhau trong

---

[1] Vì mọi người đều nói dối, bạn nên nghi vấn nhiều chỗ trong câu chuyện này. Có thể tôi không phải là người làm việc say mê. Có thể tôi không làm việc cật lực cho quyển sách này. Có thể tôi, cũng như rất nhiều người khác, đã phóng đại công việc của mình. Có thể 13 tháng "làm việc vất vả" của tôi gồm hàng tháng trời tôi không làm gì cả. Có thể tôi không sống ẩn dật. Có thể, nếu bạn kiểm tra trang Facebook của tôi, bạn sẽ thấy hình tôi đi chơi với bạn bè trong suốt thời kì tôi bảo là ẩn dật ấy. Hoặc có thể tôi sống ẩn dật thật, nhưng không phải là tự nguyện. Có thể tôi đã trải qua nhiều đêm một mình, không thể làm việc, hi vọng vô ích rằng ai đó sẽ mời tôi đi đâu đó. Có thể không ai gửi thơ mời tôi cả. Có thể không ai nhắn tin cho tôi trên Bumble. Mọi người đều nói dối. Mọi người kể chuyện đều không đáng tin cậy.

các sách. Ellenberg nhận ra rằng ông có thể so sánh có bao nhiêu trích dẫn nằm ở phần đầu sách so với phần cuối sách. Nghiên cứu này cho biết sơ bộ về khuynh hướng đọc đến cuối sách của người đọc. Theo cách đo lường của ông, hơn 90% người đọc hoàn thành tiểu thuyết *The Goldfinch* của Donna Tartt. Trái lại, chỉ 7% đọc hết kiệt tác *Thinking, Fast and Slow* của nhà kinh tế học đoạt giải Nobel là Daniel Kahneman. Phương pháp sơ bộ này ước tính: Chưa tới 3% đọc đến cuối quyển sách được thảo luận và khen ngợi nhiều—*Capital in the 21st Century*—của nhà kinh tế học Thomas Piketty. Nói cách khác, người ta thường không đọc hết các quyển chuyên luận của các nhà kinh tế học.

Một trong những luận điểm của sách này là chúng ta phải theo dõi Dữ Liệu Lớn *ở bất cứ nơi nào* nó đóng vai trò dẫn dắt, để rồi hành động thích hợp. Tôi hi vọng rằng hầu hết người đọc sẽ bám theo mọi từ ngữ của tôi và thử dò tìm các mô thức liên kết các trang cuối cùng với những trang xảy ra trước đó. Nhưng, dù tôi có hăng hái trau chuốt câu cú của mình đến thế nào đi nữa, hầu hết mọi người cũng sẽ đọc 50 trang đầu, nắm một vài điểm, và tiếp tục sống cuộc sống của riêng mình mà thôi.

Như vậy, tôi sẽ kết luận quyển sách này theo cách thích hợp duy nhất: bằng cách làm theo những điều mà dữ liệu đã cho thấy là người ta thực sự làm, chứ không phải những gì người ta nói là họ sẽ làm. Tôi sẽ đi làm vài ba cốc bia với mấy ông bạn và thôi cạy cục viết cái kết luận chết tiệt này đây. Dữ Liệu Lớn đã cho tôi biết sự thật rằng, chẳng còn mấy ai đọc tới đoạn này.

# Lời Cảm Ơn

Quyển sách này là nỗ lực của một nhóm.

Các ý tưởng này được phát triển trong thời gian tôi còn là sinh viên tại Harvard, rồi nhà khoa học dữ liệu tại Google, và sau đó là tác giả viết cho *New York Times*.

Hal Varian, người mà tôi cùng làm việc tại Google, đã ảnh hưởng rất lớn đến các ý tưởng của sách này. Tôi chỉ có thể nói: Hal luôn đi trước thời đại ông 20 năm. Quyển sách *Information Rules* của ông, viết cùng với Carl Shapiro, căn bản dự báo tương lai. Và bài báo "Predicting the Present" ông viết cùng Hyunyoung Choi cũng đặt nền móng cho cuộc cách mạng Dữ Liệu Lớn trong các ngành khoa học xã hội được mô tả trong quyển sách này. Ông còn là một cố vấn tuyệt vời và tốt bụng, rất nhiều người đã từng được làm việc với ông có thể chứng thực điều này. Một tính cách Hal kinh điển: làm hầu hết việc trong bài báo mà bạn đang là đồng tác giả với ông, và sau đó nhất quyết rằng tên bạn phải đứng trước tên ông. Sự phối hợp giữa thiên tài và tính hào phóng của Hal là điều mà tôi hiếm khi được gặp.

Tác phẩm và ý tưởng của tôi phát triển là nhờ Aaron Retica, ông biên tập tất cả các mục báo của tôi trên *New York Times*. Aaron là người uyên bác. Không hiểu sao mà thứ gì ông cũng biết: âm nhạc, lịch sử, thể thao, chính trị, xã hội học, kinh tế học, và chỉ Chúa mới biết còn gì khác nữa. Ông chịu trách nhiệm hầu hết những điều hay trong các mục báo *Times* có tên tôi trên đó. Những thành viên khác trong nhóm làm các mục báo này gồm có Bill Marsh (các tác phẩm đồ họa của ông tiếp tục làm tôi

kinh ngạc), Kevin Mc-Carthy, và Gita Daneshjoo. Quyển sách này bao gồm nhiều đoạn trích từ các mục báo này, đã được xin phép in lại đầy đủ.

Steven Pinker, người đã sẵn lòng nhận viết lời nói đầu, từ lâu đã là anh hùng của tôi. Ông đã đặt ra chuẩn mực cho một quyển sách hiện đại về khoa học xã hội—một sự khám phá hấp dẫn các nguyên tắc cơ bản của bản chất con người, luận giải được các nghiên cứu tốt nhất từ một loạt các ngành khoa học. Chuẩn mực đó là mục tiêu mà tôi sẽ phấn đấu để đạt được suốt cả đời mình.

Luận án tiến sĩ của tôi—quyển sách này phát triển từ đó—được viết dưới sự chỉ dẫn của các cố vấn tài năng và kiên nhẫn: Alberto Alesina, David Cutler, Ed Glaeser, và Lawrence Katz.

Denise Oswald là một nhà biên tập tuyệt vời. Nếu bạn muốn biết tài biên tập của cô, hãy so sánh bản thảo cuối cùng này với bản thảo đầu tiên của tôi—à mà thực ra, bạn không thể nào so sánh được vì tôi sẽ chẳng bao giờ cho bất cứ ai khác xem cái bản thảo đầu tiên đáng xấu hổ đó đâu. Tôi cũng xin được cảm ơn những thành viên khác của nhóm tại HarperCollins: Michael Barrs, Lynn Grady, Lauren Janiec, Shelby Meizlik, và Amber Oliver.

Eric Lupfer—đại diện của tôi, người nhìn thấy tiềm năng trong dự án này ngay từ đầu—đã có công trong việc đưa ra đề nghị, và giúp hoàn tất dự án.

Còn về khả năng kiểm tra sự kiện siêu đẳng thì tôi xin được cảm ơn Melvis Acosta.

Tôi cũng đã học hỏi được rất nhiều trong cuộc sống chuyên môn và học thuật của mình từ Susan Athey, Shlomo Benartzi, Jason Bordoff, Danielle Bowers, David Broockman, Bo Cowgill, Steven Delpome, John Donohue, Bill Gale, Claudia Goldin, Suzanne Greenberg, Shane Greenstein, Steve Grove, Mike Hoyt, David Laibson, A.J. Magnuson, Dana Maloney, Jeffrey Oldham, Peter Orszag, David Reiley, Jonathan Rosenberg, Michael Schwarz, Steve Scott, Rich Shavelson, Michael D. Smith, Lawrence Summers, Jon Vaver, Michael Wiggins, và Qing Wu.

Xin cảm ơn Tim Requarth và NeuWrite đã giúp đỡ tôi phát triển tác phẩm của mình.

Về việc giúp diễn dịch các nghiên cứu, tôi xin được cảm ơn Christopher Chabris, Raj Chetty, Matt Gentzkow, Solomon Messing, và Jesse Shapiro.

Tôi đã hỏi xin lời khuyên của Emma Pierson và Katia Sobolski về một chương trong sách. Hai vị đã quyết định đọc luôn cả quyển—vì lí do gì thì tôi không hiểu—và cho tôi những lời khuyên uyên thâm về mỗi đoạn văn.

Mẹ tôi, Esther Davidowitz, đã đọc cả quyển sách rất nhiều lần và giúp tôi cải tiến đáng kể. Mẹ còn dạy cho tôi, bằng cách làm gương, rằng tôi nên đi theo trí tò mò của mình, dù nó dẫn đến bất cứ nơi đâu. Khi tôi đang phỏng vấn vào một vị trí làm công việc hàn lâm, một giáo sư quay tôi: "Mẹ anh nghĩ gì về công việc anh muốn làm này?" Ý nói là mẹ tôi có thể xấu hổ chuyện tôi nghiên cứu về tình dục và các đề tài cấm kị khác. Nhưng tôi luôn biết bà tự hào về tôi vì đã đi theo trí tò mò của mình, dù nó dẫn đến bất cứ nơi đâu.

Nhiều người đọc các phần và cho những bình luận hữu ích. Tôi xin được cảm ơn Eduardo Acevedo, Coren Apicella, Sam Asher, David Cutler, Stephen Dubner, Christopher Glazek, Jessica Goldberg, Lauren Goldman, Amanda Gordon, Jacob Leshno, Alex Peysakhovich, Noah Popp, Ramon Roullard, Greg Sobolski, Evan Soltas, Noah Stephens-Davidowitz, Lauren Stephens Davidowitz, và Jean Yang. Thực ra, Jean cơ bản là bạn thân nhất của tôi trong khi tôi viết sách này, vì vậy, tôi cũng cảm ơn Jean về điều đó.

Về việc giúp thu thập dữ liệu, tôi xin được cảm ơn Brett Goldenberg, James Rogers, và Mike Williams tại MindGeek; Rob McQuown và Sam Miller tại Baseball Prospectus. Tôi xin chân thành biết ơn sự giúp đỡ tài chính đến từ Alfred Sloan Foundation. Có một thời điểm, trong khi đang viết sách này, tôi bị sa lầy, mất phương hướng, và suýt bỏ rơi dự án. Sau đó tôi về quê với cha tôi, Mitchell Stephens. Suốt hơn một tuần lễ, Cha đã đưa tôi trở lại. Cha đưa tôi đi dạo và hai cha con tranh luận về tình

yêu, cái chết, thành công, hạnh phúc, và viết lách—sau đó bảo tôi ngồi xuống để hai cha con có thể rà soát từng câu trong sách. Chắc là tôi không thể nào hoàn thành quyển sách này nếu không có Cha.

Tất cả những sai sót còn lại, dĩ nhiên, là của riêng mình tôi.

# Ghi Chú

## Dẫn nhập

5 *Cử tri Mĩ phần lớn không quan tâm việc Barack Obama là người da đen:* Katie Fretland, "Gallup: Race Not Important to Voters," The Swamp, *Chicago Tribune,* June 2008.
5 *giáo sư nổi tiếng tại UC Berkeley nghiên cứu các dữ liệu thu thập từ khảo sát:* Alexandre Mas and Enrico Moretti, "Racial Bias in the 2008 Presidential Election," *American Economic Review* 99, no. 2 (2009).
5 *xã hội hậu chủng tộc:* Ngày 12/11/2009, trong một tập của chương trình truyền hình của mình, Lou Dobbs nói rằng chúng ta sống trong một "xã hội hậu đảng phái, hậu chủng tộc." Ngày 27/1/2010, trong một tập của chương trình truyền hình của mình, Chris Matthews nói rằng Tổng thống Obama "hậu chủng tộc trên mọi phương diện." Để biết thêm các ví dụ khác, xem Michael C. Dawson and Lawrence D. Bobo, "One Year Later and the Myth of a Post-Racial Society," *Du Bois Review: Social Science Research on Race* 6, no. 2 (2009).
7 *Tôi phân tích dữ liệu từ General Social Survey:* Chi tiết những tính toán có ở trang web của tôi, sethsd.com, trong file csv tên "Sex Data." Dữ liệu từ General Social Survey có ở http://gss.norc.org/.
7 *chưa tới 600 triệu bao cao su:* Dữ liệu được cung cấp cho tác giả.
9 *các tìm kiếm và lượt đăng kí Stormfr***:* Phân tích dữ liệu Google Trends của tác giả. Tôi cũng thu thập dữ liệu về toàn bộ thành viên trang Stormfr***, như đã nói trong Seth Stephens-Davidowitz, "The Data of Hate," *New York Times,* July 13, 2014, SR4. Dữ liệu liên quan có thể được tải ở sethsd.com trong mục dữ liệu "Stormfr***."
9 *có nhiều tìm kiếm "nigger president" (tổng thống mọi đen) hơn "first black president" (tổng thống da đen đầu tiên):* Phân tích dữ liệu Google Trends của tác giả. Các bang mà ở đó điều này đúng gồm có Kentucky, Louisiana, Arizona, và North Carolina.
11 *Nghiên cứu của tôi ban đầu bị 5 tạp chí học thuật từ chối:* Bài báo này cuối cùng được đăng với tên Seth Stephens-Davidowitz, "The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data," *Journal of Public Economics* 118 (2014). Chi tiết về bài nghiên cứu này có thể được tìm thấy tại nguồn trích dẫn trên. Thêm vào đó, dữ liệu cũng có ở trang web của tôi, sethsd.com, trong mục dữ liệu "Racism."
15. *nhân tố đơn lẻ tương quan tốt nhất:* "Strongest correlate I've found for Trump support is Google searches for the n-word. Others have reported this too" (February 28, 2016, tweet). Xem thêm Nate Cohn, "Donald Trump's Strongest Supporters: A New Kind of Democrat," *New York Times,* December 31, 2015, A3.
14 Bản đồ này thể hiện phần trăm số tìm kiếm Google có chứa từ "nigger(s)." Chú ý rằng, bởi vì số đo lường là phần trăm số tìm kiếm Google, nên nó sẽ không cao hơn ở những nơi đông dân, hay những nơi mà mọi người thực hiện nhiều tìm kiếm. Cũng xin chú ý rằng một số sự khác

biệt giữa bản đồ này và bản đồ sự ủng hộ của Trump là có nguyên do dễ thấy. Sự ủng hộ Trump sụt giảm ở Texas và Arkansas vì đó là các bang "sân nhà" của 2 đối thủ của ông, Ted Cruz và Mike Huckabee.

14 Đây là dữ liệu khảo sát của Civis Analytics vào tháng 12/2015. Dữ liệu bầu cử thực tế ở đây lại không hữu dụng bằng, bởi vì nó bị ảnh hưởng nhiều bởi nơi diễn ra bầu cư sơ bộ cũng như thể thức bầu cử. Các bản đồ được in lại với sự cho phép từ *New York Times*.

16 *2.5 tỉ tỉ byte dữ liệu:* "Bringing Big Data to the Enterprise," IBM, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

18 *cây kim đang lẫn trong một đống cỏ khô ngày càng lớn:* Nassim M. Taleb, "Beware the Big Errors of 'Big Data,' " *Wired*, February 8, 2013, http://www.wired.com/2013/02/big-data-means-big-errors-people.

19 *các tìm kiếm phân biệt chủng tộc và số thành viên của trang Stormfr***:* Tôi đã nghiên cứu sự phân biệt chủng tộc trên Internet đã thay đổi như thế nào ở những vùng có mức độ ảnh hưởng bởi Đại khủng hoảng khác nhau. Tôi nghiên cứu cả tỉ lệ tìm kiếm Google từ "nigger(s)" lẫn lượng đăng kí thành viên Stormfr***. Dữ liệu liên quan có thể được tải ở sethsd.com, trong mục dữ liệu "Racial Animus" và "Stormfr***."

19 *các tìm kiếm Google phản ánh sự lo lắng:* Seth Stephens-Davidowitz, "Fifty States of Anxiety," *New York Times*, August 7, 2016, SR2. Xin chú ý, tuy các tìm kiếm Google cung cấp mẫu lớn hơn nhiều, nhưng mô thức này vẫn thống nhất với các bằng chứng từ khảo sát. Ví dụ, xem William C. Reeves et al., "Mental Illness Surveillance Among Adults in the United States," *Morbidity and Mortality Weekly Report Supplement* 60, no. 3 (2011).

19 *các tìm kiếm chuyện cười:* Điều này đã được thảo luận ở Seth Stephens-Davidowitz, "Why Are You Laughing?" *New York Times*, May 15, 2016, SR9. Dữ liệu liên quan có thể được tải tại sethsd.com, trong mục dữ liệu "Jokes."

20 *"chồng tôi muốn tôi cho anh ấy bú":* Điều này đã được thảo luận ở Seth Stephens-Davidowitz, "What Do Pregnant Women Want?" *New York Times*, May 17, 2014, SR6.

20 *hình ảnh mô tả phụ nữ cho nam giới bú:* Phân tích dữ liệu P***hub của tác giả.

20 *Phụ nữ thực hiện các tìm kiếm:* Điều này đã được thảo luận ở Seth Stephens-Davidowitz, "Searching for Sex," *New York Times*, January 25, 2015, SR1.

20 *"poemas para mi esposa embarazada":* Stephens-Davidowitz, "What Do Pregnant Women Want?"

21 *Friedman nói:* Tôi đã phỏng vấn Jerry Friedman qua điện thoại ngày 27/10/2015.

21 *các quyết định lớn đều chỉ dựa trên một mẫu nhỏ xíu:* Hal R. Varian, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives* 28, no. 2 (2014).

## Chương 1

26 *khoa học dữ liệu thuộc loại xịn nhất lại bất ngờ chứa đựng tính trực giác:* Tôi đang nói về góc độ phân tích dữ liệu mà tôi biết—khoa học dữ liệu dùng để giải thích và dự báo hành vi con người. Tôi không nói về trí thông minh nhân tạo đang thực hiện những điều, ví dụ như, lái xe hơi. Các phương pháp này, tuy cũng tận dụng các công cụ phát hiện từ bộ não con người, nhưng khó hiểu hơn rất nhiều.

27 *các triệu chứng nào sẽ dự báo ung thư tụy:* John Paparrizos, Ryan W. White, and Eric Horvitz, "Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs: Feasibility Study and Results," *Journal of Oncology Practice* (2016).

30 *Khí hậu mùa đông đánh bật hết tất cả:* Nghiên cứu này đã được thảo luận trong Seth Stephens-Davidowitz, "Dr. Google Will See You Now," *New York Times*, August 11, 2013, SR12.

31 *bộ dữ liệu lớn nhất từ trước đến nay về các mối quan hệ con người:* Lars Backstrom and Jon Kleinberg, "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014).

32 *người ta kiên định xếp lốc xoáy:* Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

32 *bệnh suyễn gây tử vong nhiều hơn khoảng 70 lần:* Giai đoạn 1979-2010, tính trung bình, 55.81 người Mĩ tử vọng vì lốc xoáy và 4216.53 người Mĩ tử vọng vì hen suyễn. Xem Annual U.S. Killer Tornado Statistics, National Weather Service, spc.noaa.gov/climo/torn/fatalmap.php và Trends in Asthma Morbidity and Mortality, American Lung Association, Epidemiology and Statistics Unit.

32 *Patrick Ewing:* Video Ewing ưa thích của tôi là "Patrick Ewing's Top 10 Career Plays," YouTube video, posted September 18, 2015, youtube.com/watch?v=Y29gMuYymv8; và "Patrick Ewing Knicks Tribute," YouTube video, posted May 12, 2006, youtube.com/watch?v=8T2l5Emz u-I.

33 *"bóng rổ là vấn đề sống chết":* S. L. Price, "Whatever Happened to the White Athlete?" *Sports Illustrated,* December 8, 1997.

33 *một khảo sát Internet:* Đây là một khảo sát khách hàng Google mà tôi thực hiện vào ngày 22/10/2013. Tôi hỏi, "Bạn đoán rằng đa số cầu thủ NBA được sinh ra ở đâu?" 2 lựa chọn là "các khu nhà nghèo" (poor neighborhoods) và "các khu nhà trung lưu" (middle-class neighborhoods); 59.7% người chọn "các khu nhà nghèo."

34 *tên của người da đen là một chỉ báo cho gia cảnh kinh tế xã hội:* Roland G. Fryer Jr. and Steven D. Levitt, "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics* 119, no. 3 (2004).

35 *Trong số tất cả những người Mĩ gốc Phi sinh vào thập niên 1980:* Centers for Disease Control and Prevention, "Health, United States, 2009," Table 9, Nonmarital Childbearing, by Detailed Race and Hispanic Origin of Mother, and Maternal Age: United States, Selected Years 1970–2006.

35 *Chris Bosh... Chris Paul:* "Not Just a Typical Jock: Miami Heat Forward Chris Bosh's Interests Go Well Beyond Basketball," Palm BeachPost.com, February 15, 2011, palmbeachpost.com/news/sports/basketball/not-just-a-typical-jock-miami-heat-for ward-chris-b/nLp7Z/; Dave Walker, "Chris Paul's Family to Compete on 'Family Feud,' nola.com, October 31, 2011, http://www.nola.com/tv/index.ssf/2011/10/chris_pauls_family_to_compete.html.

36 *cao hơn 4 inch:* "Why Are We Getting Taller as a Species?" *Scientific American,* http://www.scientificamerican.com/article/why-are-we-getting-taller/. Thú vị thay, người Mĩ đã dừng tăng chiều cao. Amanda Onion, "Why Have Americans Stopped Growing Taller?" ABC News, July 3, 2016, http://abc news.go.com/Technology/story?id=98438&page=1. Tôi đã tranh luận rằng một trong những lí do tạo nên sự gia tăng mạnh mẽ số cầu thủ NBA sinh ở nước ngoài là một số nước đang dần bắt kịp nước Mĩ về chiều cao. Số cầu thủ NBA sinh ở Mĩ cao 7 ft (~ 2.1 m) tăng 16 lần trong giai đoạn 1946-1980, tương ứng với sự phát triển chiều cao của Mĩ. Đến nay, con số đó đã chững lại, vì người Mĩ đã dừng cao lên. Trong khi đó, số cầu thủ 7 ft từ các nước khác lại tăng nhiều. Tôi phát hiện, sự gia tăng mạnh mẽ nhất trong số những cầu thủ quốc tế chủ yếu ở những người cực kì cao từ các quốc gia như Thổ Nhĩ Kì, Tây Ban Nha, và Hi Lạp—những nơi có sự gia tăng đáng kể về sức khỏe trẻ em và chiều cao người trưởng thành trong thời gian gần đây.

36 *người Mĩ xuất thân từ gia đình nghèo:* Carmen R. Isasi et al., "Association of Childhood Economic Hardship with Adult Height and Adult Adiposity among Hispanics/Latinos: The HCHS/SOL Socio-Cultural Ancillary Study," *PloS One* 11, no. 2 (2016); Jane E. Miller and Sanders Korenman, "Poverty and Children's Nutritional Status in the United States," *American Journal of Epidemiology* 140, no. 3 (1994); Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan, and Jens Ludwig, "The Economic Costs of Childhood Poverty in the United States," *Journal of Children and Poverty* 14, no. 1 (2008).

36 *người Mĩ nam trung bình cao 5'9":* Cheryl D. Fryar, Qiuping Gu, and Cynthia L. Ogden, "Anthropometric Reference Data for Children and Adults: United States, 2007–2010," *Vital and Health Statistics Series* 11, no. 252 (2012).

37 *khoảng 1 trong 5 người vào được NBA:* Pablo S. Torre, "Larger Than Real Life," *Sports Illustrated*, July 4, 2011.
37 *gia đình trung lưu, có đủ cha mẹ:* Tim Kautz, James J. Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans, "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success," National Bureau of Economic Research Working Paper 20749, 2014.
37 *Wrenn nhảy cao nhất:* Desmond Conner, "For Wrenn, Sky's the Limit," *Hartford Courant*, October 21, 1999.
37 *Nhưng Wrenn:* Chuyện của Doug Wrenn được kể trong Percy Allen, "Former Washington and O'Dea Star Doug Wrenn Finds Tough Times," *Seattle Times*, March 29, 2009.
38 *"Doug Wrenn đã chết"*: Ibid.
38 *Jordan đã từng là một đứa trẻ khó dạy:* Melissa Isaacson, "Portrait of a Legend," ESPN.com, September 9, 2009, espn.com/chicago/columns/story?id=4457017&columnist=isaacson_me lissa. Một quyển tiểu sử Jordan khá hay là của Roland Lazenby, *Michael Jordan: The Life* (Boston: Back Bay Books, 2015).
38 *Cha của anh là:* Barry Jacobs, "High-Flying Michael Jordan Has North Carolina Cruising Toward Another NCAA Title," *People*, March 19, 1984.
38 *Cuộc đời Jordan đầy những câu chuyện về việc gia đình dẫn dắt anh*: Isaacson, "Portrait of a Legend."
38 *diễn văn được hàng triệu người theo dõi:* Michael Jordan's Basketball Hall of Fame Enshrinement Speech, YouTube video, posted February 21, 2012, youtube.com/watch?v=X LzBMGXfK4c. Phương diện thú vị nhất trong bài diễn văn của Jordan không phải là những cảm xúc của anh về cha mẹ; mà là rằng anh vẫn cảm thấy cần phải chỉ ra những sự coi thường ở đầu sự nghiệp. Có lẽ sự ám ảnh cả đời với những sự coi thường là cần thiết để trở thành cầu thủ bóng rổ xuất sắc nhất mọi thời đại.
39 *LeBron James được phỏng vấn trên truyền hình:* "I'm LeBron James from Akron, Ohio," YouTube video, posted June 20, 2013, https://www.youtube.com/watch?v=XceMbPVAggk.

## Chương 2

43 *việc có hình dạng giống dương vật không giúp thực phẩm có thêm khả năng:* Tôi mã hóa một thực phẩm là có hình dạng dương vật nếu thực phẩm ấy có chiều dài vượt xa chiều rộng và có vẻ tròn. Tôi tính vào đó dưa chuột, bắp, cà rốt, cà tím, bí đao, và chuối. Dữ liệu và mã có ở sethsd.com.
44 *lỗi đánh máy do các nhà nghiên cứu Microsoft thu thập:* Bộ dữ liệu này có thể được tải từ https://www.microsoft.com/en-us/download/details.aspx?id=52418. Các nhà nghiên cứu yêu cầu người dùng Amazon Mechanical Turk mô tả các hình ảnh. Họ phân tích dữ liệu gõ phím và ghi chú lại những lúc người dùng sửa lại từ. Chi tiết có tại Yukino Baba and Hisami Suzuki, "How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs," Proceedings of the Fiftieth Annual Meeting of the Association for Computational Linguistics, 2012. Dữ liệu, mã, và mô tả chi tiết hơn về nghiên cứu này có ở sethsd.com.
46 *Xem xét tất cả tìm kiếm theo mẫu câu "I want to have sex with my..."*: Toàn bộ dữ liệu được thể hiện ở trang bên (xin cảnh báo là không dễ coi):

**"I WANT TO HAVE SEX WITH..."**

| | Số tìm kiếm hàng tháng theo cụm từ chính xác |
|---|---|
| my mom | 720 |
| my son | 590 |
| my sister | 590 |
| my cousin | 480 |
| my dad | 480 |
| my boyfriend | 480 |
| my brother | 320 |
| my daughter | 260 |
| my friend | 170 |
| my girlfriend | 140 |

46 *khiêu dâm hoạt hình:* Ví dụ, *"porn"* là một trong những từ phổ biến nhất xuất hiện trong các tìm kiếm Google về nhiều chương trình hoạt hình nổi tiếng khác nhau, cụ thể như dưới đây:

**PHIM HOẠT HÌNH VÀ KHIÊU DÂM**
**(NHỮNG TÌM KIẾM GOOGLE PHỔ BIẾN NHẤT CHO MỘT SỐ PHIM HOẠT HÌNH)**

| | | | |
|---|---|---|---|
| **family guy porn** | watch the simpsons | **futurama porn** | scooby doo games |
| family guy episodes | **the simpsons porn** | futurama leela | scooby doo movie |
| family guy free | the simpsons online | futurama episodes | **scooby doo porn** |
| watch family guy | the simpsons movie | futurama online | scooby doo velma |

47 *các cô trông trẻ:* Dựa trên tính toán của tác giả, đây là các nghề nghiệp của nữ giới xuất hiện nhiều nhất trong các tìm kiếm của nam về khiêu dâm, chia theo lứa tuổi:

**NGHỀ NGHIỆP CỦA PHỤ NỮ TRONG CÁC TÌM KIẾM KHIÊU DÂM CỦA NAM, PHÂN THEO LỨA TUỔI NAM GIỚI**

| | 18-24 | 25-64 | 65+ |
|---|---|---|---|
| 1. | Giữ trẻ | Giữ trẻ | Giữ trẻ |
| 2. | Giáo viên | Huấn luyện viên yoga | Đội trưởng đội cổ vũ |
| 3. | Huấn luyện viên yoga | Giáo viên | Bác sĩ |
| 4. | Đội trưởng đội cổ vũ | Đội trưởng đội cổ vũ | Giáo viên |
| 5. | Bác sĩ | Môi giới bất động sản | Môi giới bất động sản |
| 6. | Gái điếm | Bác sĩ | Y tá |
| 7. | Môi giới bất động sản | Gái điếm | Huấn luyện viên yoga |
| 8. | Y tá | Thư kí | Thư kí |
| 9. | Thư kí | Y tá | Gái điếm |

## Chương 3

50 *các thuật toán sẵn sàng để:* Matthew Leising, "HFT Treasury Trading Hurts Market When News Is Released," Bloomberg Markets, December 16, 2014; Nathaniel Popper, "The Robots Are Coming for Wall Street," *New York Times Magazine,* February 28, 2016, MM56; Richard Finger, "High Frequency Trading: Is It a Dark Force Against Ordinary Human Traders and Investors?" *Forbes,* September 30, 2013, forbes.com/sites/richardfinger/2013/09/30/high-frequency-trading-is-it-a-dark-force-against-ordinary-human-traders-and-investors/#50875fc751a6.

51 *Alan Krueger:* Tôi đã phỏng vấn Alan Krueger qua điện thoại ngày 8/5/2015.

52 *chỉ báo quan trọng về tốc độ lây lan dịch cúm:* Bài nghiên cứu đầu tiên là Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* 457, no. 7232 (2009). Các sai sót trong mô hình gốc được thảo luận ở David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176 (2014). Mô hình được điều chỉnh được trình bày ở Shihao Yang, Mauricio Santillana, and S. C. Kou, "Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO," *Proceedings of the National Academy of Sciences* 112, no. 47 (2015).

52 *các tìm kiếm nào tương quan sát với giá nhà ở:* Seth Stephens-Davidowitz and Hal Varian, "A Hands-on Guide to Google Data," mimeo, 2015. Xem thêm Marcelle Chauvet, Stuart Gabriel, and Chandler Lutz, "Mortgage Default Risk: New Evidence from Internet Search Queries," *Journal of Urban Economics* 96 (2016).

54 *Bill Clinton:* Sergey Brin and Larry Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Seventh International World-Wide Web Conference, April 14–18, 1998, Brisbane, Australia.

55 *các trang khiêu dâm*: John Battelle, The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture (New York: Penguin, 2005).

55 *huy động các quan điểm:* Một lập luận hay về điều này: Steven Levy, *In the Plex: How Google Thinks, Works, and Shapes Our Lives* (New York: Simon & Schuster, 2011).

57 *"Ông muốn bán nhà cũng được":* Câu trích dẫn này có trong Joe Drape, "Ahmed Zayat's Journey: Bankruptcy and Big Bets," *New York Times,* June 5, 2015, A1. Tuy nhiên, bài báo này đã nhầm câu nói này là của Seder. Thực ra đây là câu của một thành viên trong đội Seder.

58 *Tôi lần đầu gặp Seder:* Tôi đã phỏng vấn Jeff Seder và Patty Murray ở Ocala, Florida, từ 12 đến 14/6/2015.

59 *Khoảng 1/3:* Các nguyên nhân khiến ngựa đua thất bại được ước tính sơ bộ bởi Jeff Seder, dựa trên những năm tháng trong ngành.

59 *hàng trăm con ngựa chết:* Supplemental Tables of Equine Injury Database Statistics for Thoroughbreds, http://jockeyclub.com/pdfs/eid_7_year_tables.pdf.

59 *hầu hết do gãy chân:* "Postmortem Examination Program," California Animal Health and Food Laboratory System, 2013.

59 *Tuy nhiên, hơn 3/4 không thắng cuộc đua lớn nào:* Avalyn Hunter, "A Case for Full Siblings," *Bloodhorse,* April 18, 2014, http://www.bloodhorse.com/horse-racing/articles/115014/a-case-for-full-siblings.

60 *Earvin Johnson III:* Melody Chiu, "E. J. Johnson Loses 50 Lbs. Since Undergoing Gastric Sleeve Surgery," *People,* October 1, 2014.

60 *LeBron James, một cầu thủ có mẹ chỉ cao 1m65:* Eli Saslow, "Lost Stories of LeBron, Part 1," ESPN.com, October 17, 2013, http://www.espn.com/nba/story/_/id/9825052/how-lebron-james-life-changed-fourth-grade-espn-magazine.

61 *The Green Monkey:* See Sherry Ross, "16 Million Dollar Baby," New York *Daily News,* March 12, 2006, and Jay Privman, "The Green Monkey, Who Sold for $16M, Retired," ESPN.com, February 12, 2008, http://www.espn.com/sports/horse/news/story?id =3242341. A video of the

auction is available at "$16 Million Horse," YouTube video, posted November 1, 2008, https://www.youtube.com/watch?v=EyggMC85Zsg.

63 *Một điểm yếu trong nỗ lực dự báo bệnh cúm bằng dữ liệu tìm kiếm của Google:* Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, "Predicting Consumer Behavior with Web Search," *Proceedings of the National Academy of Sciences* 107, no. 41 (2010).

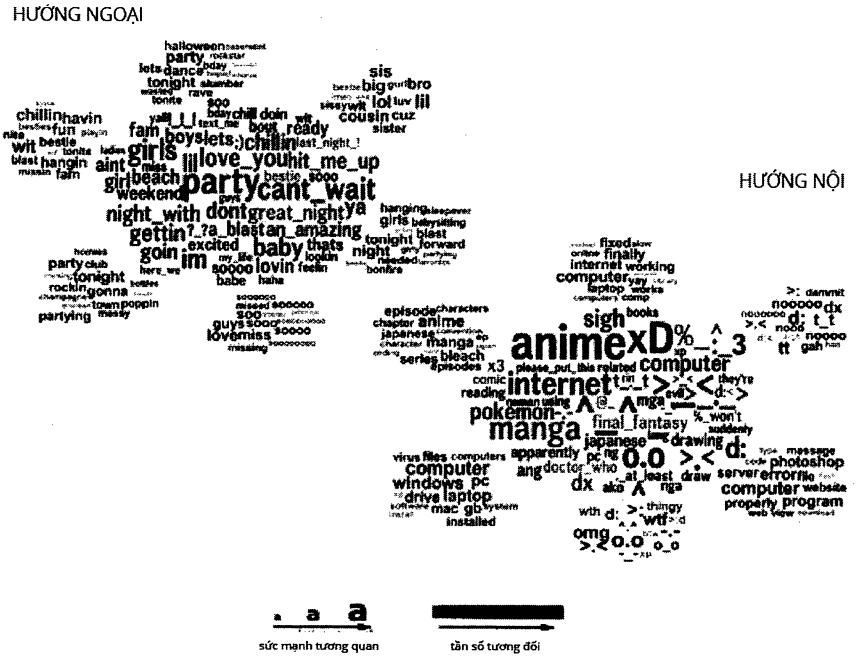
64 *Pop-Tarts dâu:* Constance L. Hays, "What Wal-Mart Knows About Customers' Habits," *New York Times*, November 14, 2004.

66 *"Nó cho kết quả tuyệt vời":* Tôi đã phỏng vấn Orley Ashenfelter qua điện thoại ngày 27/10/2016.

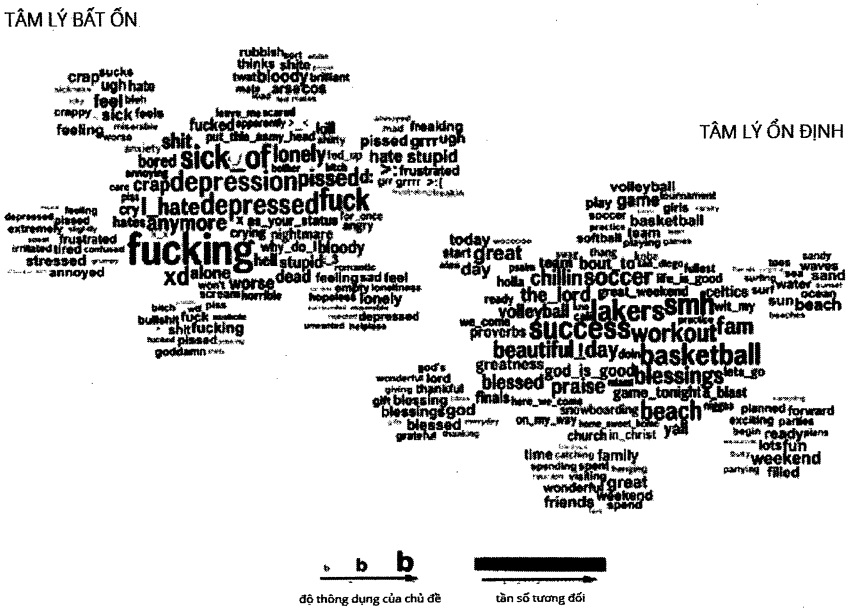
71 *nghiên cứu hàng trăm người hẹn hò nhanh khác giới:* Daniel A. McFarland, Dan Jurafsky, and Craig Rawlings, "Making the Connection: Social Bonding in Courtship Situations," *American Journal of Sociology* 118, no. 6 (2013).

72 *Leonard Cohen có lần cho đứa cháu trai mánh tán tỉnh phụ nữ:* Jonathan Greenberg, "What I Learned From My Wise Uncle Leonard Cohen," *Huffington Post*, November 11, 2016.

73 *từ ngữ được dùng trong hàng trăm ngàn bài đăng tải trên Facebook:* H. Andrew Schwartz et al., "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PloS One* 8, no. 9 (2013). Bài báo này cũng phân tách các cách mà người ta nói, dựa trên kết quả bài kiểm tra tính cách. Và họ phát hiện như sau:




(xem tiếp ở trang bên)




78 *văn bản từ hàng ngàn quyển sách và kịch bản phim:* Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds, "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes," *EPJ Data Science* 5, no. 1 (2016).

82 *loại câu chuyện thường được chia sẻ:* Jonah Berger and Katherine L. Milkman, "What Makes Online Content Viral?" *Journal of Marketing Research* 49, no. 2 (2012).

85 *Tại sao một số xuất bản phẩm tả khuynh:* Nghiên cứu này được phân tích rõ trong Matthew Gentzkow and Jesse M. Shapiro, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica* 78, no. 1 (2010). Khi dự án này bắt đầu, Gentzkow và Shapiro chỉ mới là nghiên cứu sinh tiến sĩ, giờ đây họ đã là các nhà kinh tế học hàng đầu. Gentzkow, hiện là giáo sư Stanford, đã thắng Huy chương John Bates Clark năm 2014—giải thưởng được trao cho nhà kinh tế học hàng đầu dưới 40 tuổi. Shapiro, hiện là giáo sư Đại học Brown, là biên tập viên ở tờ báo danh giá *Journal of Political Economy*. Bài báo về sự thiên lệch truyền thông của họ là một trong những bài báo được trích dẫn nhiều nhất của mỗi người.

85 *Rupert Murdoch:* Việc Murdoch sở hữu tờ báo bảo thủ *New York Post* có thể được giải thích là do New York quá rộng lớn nên có thể có đất cho các tờ báo với nhiều quan điểm khác nhau. Tuy nhiên, rõ ràng là *Post* vẫn liên tục thua lỗ. Ví dụ, xem Joe Pompeo, "How Much Does the 'New York Post' Actually Lose?" *Politico,* August 30, 2013, politico.com/media/story/2013/08/how-much-does-the-new-york-post-actu ally-lose-001176.

86 *Shapiro nói với tôi:* Tôi đã phỏng vấn Matt Gentzkow và Jesse Shapiro ngày 16/8/2015 tại Royal Sonesta Boston.

87 *quét hình các quyển niên giám:* Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A. Efros, "A Century of Portraits: A Visual Historical Record of American High School Yearbooks," paper presented at International Conference on Computer Vision, 2015. Hình ảnh được in lại với sự cho phép của các tác giả.

88 *người mẫu ảnh bắt chước cách tạo dáng của người mẫu tranh:* Ví dụ, xem Christina Kotchemidova, "Why We Say 'Cheese': Producing the Smile in Snapshot Photography," *Critical Studies in Media Communication* 22, no. 1 (2005).

89 *đo lường GDP dựa trên lượng ánh sáng:* J. Vernon Henderson, Adam Storeygard, and David N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review* 102, no. 2 (2012).

90 *GDP ước tính của họ bấy giờ tăng hơn 90%:* Kathleen Caulderwood, "Nigerian GDP Jumps 89% as Economists Add in Telecoms, Nollywood," *IBTimes*, April 7, 2014, ibtimes.com/nigerian-gdp-jumps-89-economists-add-telecoms-nolly wood-1568219.
91 *Reisinger nói:* Tôi đã phỏng vấn Joe Reisinger qua điện thoại ngày 10/6/2015.
103 *50 triệu USD:* Leena Rao, "SpaceX and Tesla Backer Just Invested $50 Million in This Startup," *Fortune*, September 24, 2015.

## Chương 4

93 *Một bài báo quan trọng năm 1950:* Hugh J. Parry and Helen M. Crossley, "Validity of Responses to Survey Questions," *Public Opinion Quarterly* 14, 1 (1950).
93 *Một khảo sát gần đây hỏi các cử nhân Đại học Maryland:* Frauke Kreuter, Stanley Presser, and Roger Tourangeau, "Social Desirability Bias in CATI, IVR, and Web Surveys," *Public Opinion Quarterly* 72(5), 2008.
93 *lệch lạc kết quả dự báo của các cuộc thăm dò:* Một bài báo lập luận rằng nói dối có thể là một vấn đề trong việc cố gắng dự báo sự ủng hộ đối với Trump, xem Thomas B. Edsall, "How Many People Support Trump but Don't Want to Admit It?" *New York Times*, May 15, 2016, SR2. Nhưng một lập luận cho rằng đây không phải là một nhân tố lớn, xem Andrew Gelman, "Explanations for That Shocking 2% Shift," *Statistical Modeling, Causal Inference, and Social Science*, November 9, 2016, http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/.
107 *says Tourangeau:* Tôi đã phỏng vấn Roger Tourangeau qua điện thoại ngày 5/5/2015.
94 *quá nhiều người nói họ ở mức trên trung bình:* Chuyện này được thảo luận trong Adam Grant, *Originals: How Non-Conformists Move the World* (New York: Viking, 2016). Nguồn ban đầu là David Dunning, Chip Heath, and Jerry M. Suls, "Flawed Self-Assessment: Implications for Health, Education, and the Workplace," *Psychological Science in the Public Interest* 5 (2004).
94 *phá rối cuộc khảo sát:* Anya Kamenetz, " 'Mischievous Responders' Confound Research on Teens," *nprED*, May 22, 2014, www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens. The original research this articlediscusses is Joseph P. Robinson-Cimpian, "Inaccurate Estimationof Disparities Due to Mischievous Responders," *Educational Researcher*43, no. 4 (2014).
96 *người Mĩ tìm kiếm "porn" (khiêu dâm) nhiều hơn "weather" (thời tiết):* google.com/trends/explore?date=all&geo=US&q=porn,weather.
96 *thú nhận là mình có xem phim ảnh khiêu dâm:* Amanda Hess, "How Many Women Are Not Admitting to Pew That They Watch Porn?" *Slate*, October 11, 2013, slate.com/blogs/xx_factor/2013/10/11/pew_online_viewing_study_percentage_of_women_who_watch_online_porn_is_growing.html.
96 *"cock," "f*ck," và "porn":* Nicholas Diakopoulus, "Sex, Violence, and Autocomplete Algorithms," *Slate*, August 2, 2013, slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html.
97 *gấp 3.6 lần khả năng sẽ bảo với Google rằng họ hối tiếc:* Tôi ước tính, bao gồm nhiều cách nói khác nhau, có khoảng 1,730 tìm kiếm Google của người Mĩ hàng tháng nói rõ họ hối tiếc đã có con. Có chỉ khoảng 50 tìm kiếm thể hiện sự hối tiếc vì đã không có con. Có khoảng 15.9 triệu người Mĩ trên 45 tuổi không có con. Có khoảng 152 triệu người Mĩ có con. Điều này muốn nói, sau khi đã tính theo tỉ lệ dân số, những người có con có gấp khoảng 3.6 lần khả năng thể hiện sự hối tiếc trên Google so với những người không có con. Rõ ràng, như đã được đề cập trong nội dung sách, nhưng đáng nhấn mạnh lại một lần nữa, các lời thú nhận này trên Google chỉ được thực hiện bởi một số lượng người nhỏ, có chọn lọc—có thể là những người đã cảm thấy một sự hối tiếc đủ mạnh đến nỗi tức thời quên rằng Google không thể giúp gì họ cả.
98 *tiểu bang ủng hộ hôn nhân đồng tính nhất:* Các số ước tính là từ Nate Silver, "How Opinion on Same-Sex Marriage Is Changing, and What It Means," FiveThirtyEight, March 26, 2013,

http://fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/?_r=0.

99 *Khoảng 2.5% người dùng Facebook nam (có ghi sở thích giới tính) ghi rằng họ thích nam giới:* Phân tích của tác giả về dữ liệu quảng cáo Facebook. Tôi không bao gồm người dùng Facebook ghi "nam và nữ." Phân tích của tôi cho rằng con số % người dùng không nhỏ nói họ thích nam và nữ nghĩ câu hỏi này ám chỉ việc thích trong tình bạn chứ không phải thích theo kiểu tình yêu.

100 *khoảng 5% tìm kiếm khiêu dâm nam là muốn tìm khiêu dâm nam đồng tính:* Như đã được thảo luận, Google Trends không phân tách các tìm kiếm theo giới tính. Google AdWords có phân tách xem trang theo các tiêu chí khác nhau theo giới tính. Tuy nhiên, dữ liệu này còn lâu mới chính xác. Để ước tính các tìm kiếm theo giới tính, đầu tiên tôi dùng dữ liệu tìm kiếm đó để có con số ước tính cả tiểu bang số % tìm kiếm khiêu dâm đồng tính theo tiểu bang. Sau đó tôi chuẩn hóa dữ liệu này theo dữ liệu giới tính Google AdWords. Một cách khác để có dữ liệu giới cụ thể là dùng dữ liệu P***hub . Tuy nhiên, P***hub có thể là mẫu quá chọn lọc, vì nhiều người đồng tính có thể dùng các trang chỉ tập trung vào khiêu dâm đồng tính. P***hub chỉ ra rằng khiêu dâm đồng tính trong nam giới ở đây thấp hơn con số dựa trên tìm kiếm Google. Tuy nhiên, nó khẳng định rằng không có mối quan hệ mạnh mẽ giữa sự dễ tính với tình dục đồng giới và việc xem khiêu dâm đồng tính. Tất cả dữ liệu này và các ghi chú sâu hơn có sẵn trên website của tôi, tại sethsd.com, trong mục "Sex."

101 *Khoảng 4% nam sinh ở đó là đồng tính công khai trên Facebook:* Tính toán của tác giả về dữ liệu quảng cáo Facebook: Ngày 8/2/2017, khoảng 300 nam sinh trung học trong thị trường truyền thông San Francisco-Oakland-San Jose trên Facebook nói họ thích nam giới. Khoảng 7,800 nói họ thích phụ nữ.

104 *"Ở Iran chúng tôi không có người đồng tính":* " 'We Don't Have Any Gays in Iran,' Iranian President Tells Ivy League Audience," Daily Mail.com, September 25, 2007, http://www.dailymail.co.uk/news/article-483746/We-dont-gays-Iran-Iranian-president-tells-Ivy-League-audience.html.

104 *"Thành phố chúng tôi không có người đồng tính":* Brett Logiurato, "Sochi Mayor Claims There Are No Gay People in the City," *Sports Illustrated*, January 27, 2014.

104 *hành vi Internet tiết lộ rằng có sự quan tâm đáng kể về khiêu dâm đồng tính ở Sochi và Iran:* Theo Google AdWords, có hàng chục ngàn tìm kiếm mỗi năm cho "геи порно" (khiêu dâm đồng tính). Số phần trăm tìm kiếm khiêu dâm đồng tính ở Sochi tương tự ở Mĩ. Google AdWords không bao gồm dữ liệu cho Iran. P***hub cũng không cho thấy dữ liệu về Iran. Tuy nhiên, PornMD nghiên cứu dữ liệu tìm kiếm của họ và báo cáo rằng 5 trong top 10 từ ngữ tìm kiếm ở Iran là khiêu dâm đồng tính. Mục này bao gồm "daddy love" và "hotel businessman" và được báo cáo trong Joseph Patrick McCormick, "Survey Reveals Searches for Gay Porn Are Top in Countries Banning Homosexuality," *PinkNews*, http://www.pinknews.co.uk/2013/03/13/survey-reveals-searches-for-gay-porn-are-top-in-countries-banning-homosexuality/. Theo Google Trends, khoảng 2% tìm kiếm khiêu dâm ở Iran là khiêu dâm đồng tính, thấp hơn ở Mĩ nhưng vẫn cho thấy là có sự quan tâm rộng rãi.

106 *Khi nói tới chuyện tình dục:* Stephens-Davidowitz, "Searching for Sex." Dữ liệu cho mục này có thể được tìm thấy trên website của tôi, sethsd .com, trong mục "Sex."

106 *Khoảng 11% phụ nữ độ tuổi từ 15 đến 44:* Current Contraceptive Status Among Women Aged 15–44: United States, 2011–2013, Centers for Disease Control and Prevention, http://www.cdc.gov/nchs/data/databriefs/db173_table.pdf#1.

106 *khoảng 10% phụ nữ trong số này sẽ có thai mỗi tháng:* David Spiegelhalter, "Sex: What Are the Chances?" BBC News, March 15, 2012, http://www.bbc.com/future/story/20120313-sex-in-the-city–or-elsewhere.

106 *1/113 số phụ nữ ở độ tuổi sinh con:* Có khoảng 6.6 triệu trường hợp mang thai mỗi năm và 62 triệu phụ nữ ở độ tuổi từ 15 đến 44.

111 *"yêu" bằng miệng khác giới:* Như đã đề cập, tôi không biết giới tính của người tìm kiếm Google. Tôi đang giả sử rằng đa số áp đảo các tìm kiếm xem cách thực hiện *cunnilingus* là bởi

nam giới và đa số áp đảo các tìm kiếm xem cách thực hiện fellatio là nữ giới. Tôi giả sử như vậy bởi vì đại đa số người có xu hướng tính dục khác giới, và bởi vì có thể nhu cầu tìm hiểu cách làm hài lòng bạn đồng tính sẽ ít hơn.

111 *top 5 từ tiêu cực:* Phân tích dữ liệu Google AdWords của tác giả.

112 *giết:* Evan Soltas and Seth Stephens-Davidowitz, "The Rise of Hate Search," *New York Times*, December 13, 2015, SR1. Dữ liệu và các chi tiết có thể được tìm thấy trên website của tôi, sethsd.com, trong mục "Islamophobia."

114 *gấp 17 lần:* Phân tích dữ liệu Google Trends của tác giả.

114 *ngày lễ Martin Luther King Jr.:* Phân tích dữ liệu Google Trends của tác giả.

116 *tương quan với khoảng cách tiền lương da trắng-da đen:* Ashwin Rode and Anand J. Shukla, "Prejudicial Attitudes and Labor Market Outcomes," mimeo, 2013.

117 *Chính là cha mẹ họ:* Seth Stephens-Davidowitz, "Google, Tell Me. Is My Son a Genius?" *New York Times*, January 19, 2014, SR6. Dữ liệu cho các tìm kiếm chính xác có thể được tìm thấy khi dùng using Google AdWords. Các số ước tính có thể được tìm thấy với Google Trends, bằng cách so sánh các tìm kiếm có các từ "gifted" và "son" so với "gifted" và "daughter." Chẳng hạn, so sánh google.com/trends/explore?date=all&geo=US&q=gifted%20son,gifted%20daughter và google.com/trends/explore?date=all&geo=US&q=overweight%20son,overweight%20daughter.
Một ngoại lệ đối với mô thức chung là có nhiều câu hỏi về sons' brains và daughters' bodies hơn là các tìm kiếm "fat son" và "fat daughter." Điều này dường như liên quan đến mức độ phổ biến của khiêu dâm loạn luân được thảo luận trước đây. Khoảng 20% tìm kiếm với các từ "fat" và "son" cũng gồm luôn từ "porn."

117 *khả năng góp mặt ở các chương trình năng khiếu của bé gái cao hơn bé trai 9%:* "Gender Equity in Education: A Data Snapshot," Office for Civil Rights, U.S. Department of Education, June 2012, http://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-edu cation.pdf.

118 *Khoảng 28% bé gái thừa cân, trong khi đó 35% cậu trai thừa cân:* Data Resource Center for Child and Adolescent Health, childhealthdata.org/browse/survey/results?q=2415&g=455&a=3879&r=1.

118 *tiểu sử Stormfr***:* Stephens-Davidowitz, "The Data of Hate." Dữ liệu liên quan có thể được tải xuống tại sethsd.com, trong phần dữ liệu có tiêu đề "Stormfr***."

120 *Stormfr*** trong đợt ứng cử của Donald Trump:* Tìm kiếm Google quan tâm đến Stormfr*** trong tháng 10/2016 tương tự với các mức trong suốt tháng 10/2015. Điều này hoàn toàn trái ngược với tình hình trong suốt cuộc bầu cử đầu tiên của Obama. Tháng 10/2008, tìm kiếm quan tâm đến Stormfr*** đã tăng gần như 60% so với tháng 10 trước. Vào ngày sau khi Obama trúng cử, tìm kiếm Google về Stormfr*** tăng lên khoảng 10 lần. Vào ngày sau khi Trump trúng cử, tìm kiếm Stormfr*** tăng khoảng 2.5 lần. Mức này gần như tương đương với mức tăng ngày sau khi George W. Bush trúng cử năm 2004 và phần lớn có thể phản ánh sự quan tâm tin tức trong giới ghiền chính trị.

122 *sự phân li về chính trị trên Internet:* Matthew Gentzkow and Jesse M. Shapiro, "Ideological Segregation Online and Offline," *Quarterly Journal of Economics* 126, no. 4 (2011).

124 *bạn bè trên Facebook:* Eytan Bakshy, Solomon Messing, and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (2015). Họ phát hiện rằng, trong 9% người dùng Facebook tích cực tuyên bố hệ tư tưởng của họ, khoảng 23% bạn bè họ (trong số những người cũng tuyên bố một hệ tư tưởng) có hệ tư tưởng đối nghịch và 28.5% tin tức họ xem trên Facebook là từ hệ tư tưởng đối nghịch đó. Các con số này không thể so sánh trực tiếp với các con số khác về sự phân li bởi vì chúng chỉ bao gồm mẫu người dùng Facebook nhỏ tuyên bố hệ tư tưởng của họ. Giả sử là những người dùng này rất có khả năng là tích cực về chính trị và kết giao với những người dùng tích cực về chính trị khác có chung hệ tư tưởng. Nếu điều này là đúng, sự đa dạng giữa tất cả người dùng sẽ lớn hơn rất nhiều.

125 *một lí do quan trọng là Facebook:* Một yếu tố làm cho các phương tiện truyền thông xã hội đa dạng đến kinh ngạc là nó thưởng lớn cho những bài báo rất nổi tiếng và được chia sẻ rộng rãi, bất kể khuynh hướng chính trị của chúng là gì. Xem Solomon Messing and Sean Westwood,

"Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online," 2014.

125 *Người ta thường có bạn bè trên Facebook nhiều hơn ngoài đời:* See Ben Quinn, "Social Network Users Have Twice as Many Friends Online as in Real Life," *Guardian*, May 8, 2011. Bài báo này thảo luận một nghiên cứu năm 2011 của Cystic Fibrosis Trust, họ phát hiện rằng người dùng mạng xã hội trung bình có 121 bạn trực tuyến so với 55 bạn ngoài đời thực. Theo một nghiên cứu của Pew Research năm 2014, người dùng Facebook trung bình đã có hơn 300 bạn bè. Xem Aaron Smith, "6 New Facts About Facebook," February 3, 2014,.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

125 *các mối quan hệ yếu:* Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic, "The Role of Social Networks in Information Diffusion," *Proceedings of the 21st International Conference on World Wide Web*, 2012.

126 *"Các dự báo âm u tăm tối chưa thành sự thật":* "Study: Child Abuse on Decline in U.S.," Associated Press, December 12, 2011.

126 *có đúng là tình trạng ngược đãi trẻ em giảm mạnh:* See Seth Stephens-Davidowitz, "How Googling Unmasks Child Abuse," *New York Times*, July 14, 2013, SR5, and Seth Stephens-Davidowitz, "Unreported Victims of an Economic Downturn," mimeo, 2013.

126 *vì phải chờ đợi quá lâu nên đã bỏ cuộc:* "Stopping Child Abuse: It Begins With You," *The Arizona Republic*, March 26, 2016.

127 *những cách phá thai chui:* Seth Stephens-Davidowitz, "The Return of the D.I.Y. Abortion," *New York Times*, March 6, 2016, SR2. Dữ liệu và các chi tiết có thể được tìm thấy trên website của tôi, sethsd.com, trong mục "Self-Induced Abortion."

130 *lượng phát hành mỗi số báo trung bình tương tự:* Alliance for Audited Media, Consumer Magazines, http://abcas3.auditedmedia.com/ecirc/magtitle search.asp.

130 *trên Facebook:* Tính toán của tác giả, ngày 4/10/2016, dùng Ads Manager của Facebook.

131 *top 10 website được lui tới nhiều nhất:* "List of Most Popular Websites," Wikipedia. Theo Alexa, trang web theo dõi hành vi lướt web, ngày 4/9/2016, trang khiêu dâm nổi tiếng nhất là XVid***, và đây là website nổi tiếng thứ 57. Theo SimilarWeb, ngày 4/9/2016, trang khiêu dâm nổi tiếng nhất là XVid***, và đây là website nổi tiếng thứ 17. Top 10, theo Alexa, là Google, YouTube, Facebook, Baidu, Yahoo!, Amazon, Wikipedia, Tencent QQ, Google India, and Twitter.

132 *Sáng sớm ngày 5/9/2006*: Câu chuyện này là từ David Kirkpatrick, The Facebook Effect: The Inside Story of the Company That Is Connecting the World (New York: Simon & Schuster, 2010).

133 *các doanh nghiệp lớn được xây trên những bí mật*: Peter Thiel and Blake Masters, Zero to One: Notes on Startups, or How to Build the Future (New York: The Crown Publishing Group, 2014).

135 *theo Xavier Amatriain:* Tôi đã phỏng vấn Xavier Amatriain qua điện thoại ngày 5/5/2015.

137 *các câu hỏi hàng đầu của người Mĩ trong buổi phát biểu Thông điệp liên bang 2014 của Obama*: Phân tích dữ liệu Google Trends của tác giả.

140 *lần này tại một nhà thờ Hồi giáo:* "The President Speaks at the Islamic Society of Baltimore," YouTube video, posted February 3, 2016, https://www.youtube.com/watch?v=LRRVdVqAjdw.

141 *các tìm kiếm thù địch, giận dữ chống người Hồi giáo giảm mạnh:* Phân tích dữ liệu Google Trends của tác giả. Các tìm kiếm "kill Muslims" thấp hơn thời điểm so sánh trước đó 1 tuần. Hơn nữa, các tìm kiếm có chữ "Muslims" và 1 trong top 5 các từ tiêu cực về nhóm này cũng thấp hơn.

## Chương 5

166 *how childhood experiences influence which baseball team you support:* Seth Stephens-Davidowitz, "They Hook You When You're Young," *New York Times*, April 20, 2014, SR5. Dữ liệu và mã cho nghiên cứu này có thể được tìm thấy trên website của tôi, sethsd.com, trong mục "Baseball."

170 *the single most important year:* Yair Ghitza and Andrew Gelman, "The Great Society, Reagan's Revolution, and Generations of Presidential Voting," bản thảo chưa xuất bản.
173 *Chetty explains:* Tôi đã phỏng vấn Raj Chetty bằng điện thoại ngày 30/7/2015.
176 *escape the grim reaper:* Raj Chetty et al., "The Association Between Income and Life Expectancy in the United States, 2001–2014," *JAMA* 315, no. 16 (2016).
178 *Contagious behavior may be driving some of this:* Julia Belluz, "Income Inequality Is Chipping Away at Americans' Life Expectancy," vox.com, April 11, 2016.
178 *why some people cheat on their taxes:* Raj Chetty, John Friedman, and Emmanuel Saez, "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review* 103, no. 7 (2013).
180 *I decided to download Wikipedia:* Phần này từ Seth Stephens-Davidowitz, "The Geography of Fame," *New York Times,* March 23, 2014, SR6. Dữ liệu có thể được tìm thấy trên website của tôi, sethsd .com, trong mục "Wikipedia Birth Rate, by County." Về việc giúp tải về và mã hóa nơi sinh cấp hạt mọi mục Wikipedia, tôi cảm ơn Noah Stephens-Davidowitz.
183 *a big city:* Để có thêm bằng chứng về giá trị của các thành phố, xem Ed Glaeser, *Triumph of the City* (New York: Penguin, 2011). (Glaeser là cố vấn lớp sau đại học của tôi.)
191 *many examples of real life imitating art*: David Levinson, ed., Encyclopedia of Crime and Punishment (Thousand Oaks, CA: SAGE, 2002).
191 *subjects exposed to a violent film will report more anger and hostility*: Craig Anderson et al., "The Influence of Media Violence on Youth," Psychological Science in the Public Interest 4 (2003).
192 *On weekends with a popular violent movie:* Gordon Dahl and Stefano DellaVigna, "Does Movie Violence Increase Violent Crime?" *Quarterly Journal of Economics* 124, no. 2 (2009).
195 *Google searches can also be broken down by the minute:* Seth Stephens-Davidowitz, "Days of Our Digital Lives," *New York Times,* July 5, 2015, SR4.
196 *alcohol is a major contributor to crime:* Anna Richardson and Tracey Budd, "Young Adults, Alcohol, Crime and Disorder," *Criminal Behaviour and Mental Health* 13, no. 1 (2003); Richard A. Scribner, David P. MacKinnon, and James H. Dwyer, "The Risk of Assaultive Violence and Alcohol Availability in Los Angeles County," *American Journal of Public Health* 85, no. 3 (1995); Dennis M. Gorman, Paul W. Speer, Paul J. Gruenewald, and Erich W. Labouvie, "Spatial Dynamics of Alcohol Availability, Neighborhood Structure and Violent Crime," *Journal of Studies on Alcohol* 62, no. 5 (2001); Tony H. Grubesic, William Alex Pridemore, Dominique A. Williams, and Loni Philip-Tabb, "Alcohol Outlet Density and Violence: The Role of Risky Retailers and Alcohol-Related Expenditures," *Alcohol and Alcoholism* 48, no. 5 (2013).
196 *letting all four of his sons play football:* "Ed McCaffrey Knew Christian McCaffrey Would Be Good from the Start—'The Herd,'" YouTube video, posted December 3, 2015, https://www.youtube.com/watch?v=boHMmp7DpX0.
197 *analyzing piles of data:* Các nhà nghiên cứu đã phát hiện nhiều điều bằng cách tận dụng dữ liệu tội phạm này và phân tách theo những số gia nhỏ về thời gian. Một ví dụ? Đơn kiện bạo lực gia đình tăng ngay sau khi đội bóng đá của một thành phố thua trận đấu mà người ta kì vọng sẽ thắng. Xem David Card and Gordon B. Dahl, "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior," *Quarterly Journal of Economics* 126, no. 1 (2011).
197 *Here's how Bill Simmons:* Bill Simmons, "It's Hard to Say Goodbye to David Ortiz," ESPN.com, June 2, 2009, http://www.espn.com/espnmag/story?id=4223584.
198 *how can we predict how a baseball player will perform in the future*: Điều này được thảo luận trong Nate Silver, The Signal and the Noise: Why So Many Predictions Fail—But Some Don't (New York: Penguin, 2012).
199 *"beefy sluggers" indeed do, on average, peak early:* Ryan Campbell, "How Will Prince Fielder Age?" October 28, 2011, http://www.fan graphs.com/blogs/how-will-prince-fielder-age/.
199 *Ortiz's doppelgangers':* Dữ liệu này tôi được Rob McQuown của Baseball Prospectus cung cấp cho.

204 *Kohane asks:* Tôi đã phỏng vấn Isaac Kohane bằng điện thoại ngày 15/6/2015.
205 *James Heywood is an entrepreneur:* Tôi đã phỏng vấn James Heywood bằng điện thoại ngày 17/8/2015.

## Chương 6

178 *Ngày 27/2/2000:* Câu chuyện này được thảo luận, cũng có ở những nơi khác, trong Brian Christian, "The A/B Test: Inside the Technology That's Changing the Rules of Business," *Wired,* April 25, 2012, http://www.wired.com/2012/04/ff_abtesting/.
180 *Khi giáo viên được trả tiền, mức độ vắng mặt giảm xuống một nửa:* Esther Duflo, Rema Hanna, and Stephen P. Ryan, "Incentives Work: Getting Teachers to Come to School," *American Economic Review* 102, no. 4 (2012).
209 *khi Bill Gates biết đến công trình của Duflo*: Ian Parker, "The Poverty Lab," New Yorker, May 17, 2010.
181 *các kĩ sư Google chạy 7 ngàn thử nghiệm A/B*: Christian, "The A/B Test."
181 *41 sắc độ xanh dương chỉ khác nhau chút ít:* Douglas Bowman, "Goodbye, Google," stopdesign, March 20, 2009, http://stopde sign.com/archive/2009/03/20/goodbye-google.html.
182 *Facebook bây giờ chạy:* Eytan Bakshy, "Big Experiments: Big Data's Friend for Making Decisions," April 3, 2014, https://www.face book.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/. Nguồn thông tin về các nghiên cứu được có ở "How many clinical trials are started each year?" Quora post, https://www.quora.com/How-many-clinical-trials-are-started-each-year.
182 *Optimizely:* Tôi đã phỏng vấn Dan Siroker bằng điện thoại ngày 29/4/2015.
184 *thu được khoảng 60 triệu USD kinh phí:* Dan Siroker, "How Obama Raised $60 Million by Running a Simple Experiment," Optimizely blog, November 29, 2010, blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/.
184 *Trang Boston Globe thử nghiệm A/B các tiêu đề:* Các thử nghiệm A/B và kết quả trang *Boston Globe* được cung cấp cho tác giả. Một số chi tiết về cuộc thử nghiệm của Globe có thể tìm thấy ở "The Boston Globe: Discovering and Optimizing a Value Proposition for Content," Marketing Sherpa Video Archive, https://www.marketingsherpa.com/video/boston-globe-optimization-summit2. Bài này gồm có một cuộc trò chuyện được ghi âm giữa Peter Doucette của *Globe* và Pamela Markey tại MECLABS.
188 *Benson nói:* Tôi đã phỏng vấn Dan Siroker bằng điện thoại ngày 29/4/2015.
188 *Họ thêm một mũi tên chỉ sang phải, bọc trong một hình vuông:* "Enhancing Text Ads on the Google Display Network," Inside Ad-Sense, December 3, 2012, adsense.googleblog.com/2012/12/enhancing-text-ads-on-google-display.html.
189 *khách hàng Google phê phán:* Ví dụ, xem, "Large arrows appearing in google ads—please remove," DoubleClick Publisher Help Forum, productforums.google.com/forum/#!topic/dfp/p_TRMqWUF9s.
190 *sự trỗi dậy của các thói nghiện mang tính hành vi trong xã hội đương thời*: Adam Alter, Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked (New York: Penguin, 2017).
190 *Các thói nghiện được tìm kiếm hàng đầu trên Google, 2016:* Phân tích dữ liệu Google Trends của tác giả.
193 *Levitt nói trong một bài giảng:* Điều này đã được thảo luận trong một video hiện đang ở vị trí nổi bật trên trang Freakonomics của Harry Walker Speakers Bureau, harrywalker.com/speakers/authors-of-freakonomics/.
195 *quảng cáo bia và nước giải khát chạy trong trận Super Bowl:* Wesley R. Hartmann and Daniel Klapper, "Super Bowl Ads," bản thảo chưa xuất bản, 2014.

196 *chú nhóc mặt mụn mặc xà lỏn:* Một bài viết cho thấy chúng ta có nhiều khả năng đang sống trong một chương trình giả lập máy tính, xem Nick Bostrom, "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53, no. 211 (2003).
197 *Trong số 43 tổng thống Mĩ: Los Angeles Times* staff, "U.S. Presidential Assassinations and Attempts," *Los Angeles Times*, January 22, 2012, http://timelines.latimes.com/us-presidential-assassinations-and-attempts/.
197 *So sánh John F. Kennedy và Ronald Reagan:* Benjamin F. Jones and Benjamin A. Olken, "Do Assassins Really Change History?" *New York Times*, April 12, 2015, SR12.
197 *Kadyrov chết:* Một video về cuộc tấn công: "Parade surprise (Chechnya 2004)," YouTube video, posted March 31, 2009, https://www.youtube.com/watch ?v=fHWhs5Q kfuY.
197 *Hitler thì thay đổi lịch làm việc:* Câu chuyện này cũng được thảo luận trong Jones and Olken, "Do Assassins Really Change History?"
197 *tác động của việc lãnh đạo bị ám sát là gì:* Benjamin F. Jones and Benjamin A. Olken, "Hit or Miss? The Effect of Assassinations on Institutions and War," *American Economic Journal: Macroeconomics* 1, no. 2 (2009).
199 *trúng số không làm bạn hạnh phúc trong ngắn hạn:* Đã được nói trong John Tierney, "How to Win the Lottery (Happily)," *New York Times*, May 27, 2014, D5. Tierney's piece discusses the following studies: Bénédicte Apouey and Andrew E. Clark, "Winning Big but Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health," *Health Economics* 24, no. 5 (2015); Jonathan Gardner and Andrew J. Oswald, "Money and Mental Wellbeing: A Longitudinal Study of Medium-Sized Lottery Wins," *Journal of Health Economics* 26, no. 1 (2007); and Anna Hedenus, "At the End of the Rainbow: Post-Winning Life Among Swedish Lottery Winners," unpublished manuscript, 2011. Tierney cũng đã chỉ ra nghiên cứu nổi tiếng năm 1978—Philip Brickman, Dan Coates, and Ronnie Janoff-Bulman, "Lottery Winners and Accident Victims: Is Happiness Relative?" *Journal of Personality and Social Psychology* 36, no. 8 (1978)—trong đó phát hiện rằng thắng xổ số không làm bạn vui hơn được dựa trên mẫu rất nhỏ.
199 *láng giềng của bạn trúng số:* Xem Peter Kuhn, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn, "The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery," *American Economic Review* 101, no. 5 (2011), and Sumit Agarwal, Vyacheslav Mikhed, and Barry Scholnick, "Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies," working paper, 2016.
199 *láng giềng của những người trúng số có nhiều khả năng bị phá sản hơn hẳn bình thường:* Agarwal, Mikhed, and Scholnick, "Does Inequality Cause Financial Distress?"
199 *bác sĩ có thể được động viên bởi các khoản tiền khuyến khích:* Jeffrey Clemens and Joshua D. Gottlieb, "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review* 104, no. 4 (2014). Chú ý rằng các kết quả này không ám chỉ là bác sĩ xấu xa. Thực tế, các kết quả có thể rắc rối hơn nữa nếu các quy trình phụ thêm mà bác sĩ chỉ định khi họ được trả thêm tiền lại thực sự cứu sống bệnh nhân. Nếu mà đúng như vậy thì có nghĩa là bác sĩ cần được trả đủ tiền để chỉ định cách điều trị cứu sống bệnh nhân. Thay vì vậy, các kết quả của Clemens và Gottlieb cho rằng, các bác sĩ sẽ chỉ định các cách điều trị cứu sống bệnh nhân dù cho họ được trả bao nhiêu tiền đi nữa. Đối với các quy trình mà không giúp ích gì nhiều, các bác sĩ phải được trả đủ tiền để ra chỉ định điều trị. Nói cách khác: Các bác sĩ không quá chú ý đến các khoản tiền khuyến khích khi sử dụng những thứ đe dọa mạng sống; họ chỉ rất chú ý đến các khoản tiền khuyến khích khi sử dụng những thứ không quan trọng.
200 *150 triệu USD:* Robert D. McFadden and Eben Shapiro, "Finally, a Face to Fit Stuyvesant: A High School of High Achievers Gets a High-Priced Home," *New York Times*, September 8, 1992.
200 *Trường có 55 lớp:* Thông tin các khóa học có sẵn trên website của Stuy, stuy.enschool.or g/index.jsp.
200 *Khoảng 1/4 học sinh tốt nghiệp được nhận vào đại học thuộc nhóm Ivy League:* Anna Bahr, "When the College Admissions Battle Starts at Age 3," *New York Times*, July 29, 2014, www.nytimes.com/2014/07/30/upshot/when-the-college-admissions-battle-starts-at-age-3.html.

200 *Stuyvesant đã đào tạo:* Sewell Chan, "The Obama Team's New York Ties," *New York Times*, November 25, 2008; Evan T. R. Rosenman, "Class of 1984: Lisa Randall," *Harvard Crimson*, June 2, 2009; "Gary Shteyngart on Stuyvesant High School: My New York," YouTube video, posted August 4, 2010, https://www.youtube.com/watch?v=NQ_phGkC-Tk; Candace Amos, "30 Stars Who Attended NYC Public Schools," New York *Daily News*, May 29, 2015.
200 *Những người đã phát biểu tại lễ trao bằng tốt nghiệp:* Carl Campanile, "Kids Stuy High Over Bubba: He'll Address Ground Zero School's Graduation," *New York Post*, March 22, 2002; United Nations Press Release, "Stuyvesant High School's 'Multicultural Tapestry' Eloquent Response to Hatred, Says Secretary-General in Graduation Address," June 23, 2004; "Conan O'Brien's Speech at Stuyvesant's Class of 2006 Graduation in Lincoln Center," YouTube video, posted May 6, 2012, https://www.youtube.com/watch?v=zAMkUE9Oxnc.
200 *Stuy xếp số 1:* Xem k12.niche.com/rankings/public-high-schools/best-overall/.
201 *Chưa tới 5% số thí sinh vào được Stuy:* Pamela Wheaton, "8th-Graders Get High School Admissions Results," *Insideschools*, March 4, 2016, http://insideschools.org/blog/item/1001064-8th-graders-get-high-school-admissions-results.
204 *người tù bị đưa đến các nhà tù an ninh cao:* M. Keith Chen and Jesse M. Shapiro, "Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach," *American Law and Economics Review* 9, no. 1 (2007).
204 *Ảnh hưởng của trường Stuyvesant là bao nhiêu?* Atila Abdulkadiroğlu, Joshua Angrist, and Parag Pathak, "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica* 82, no. 1 (2014). Kết quả vô hiệu tương tự được phát hiện độc lập bởi Will Dobbie and Roland G. Fryer Jr., "The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools," *American Economic Journal: Applied Economics* 6, no. 3 (2014).
206 *người tốt nghiệp Harvard trung bình kiếm được:* Xem http://www.payscale.com/college-salary-report/bachelors.
207 *các học sinh tương tự nhau, được nhận vào các trường uy tín tương tự nhưng chọn học trường khác nhau, cuối cùng cũng sẽ có vị trí tương đương nhau*: Stacy Berg Dale and Alan B. Krueger, "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," Quarterly Journal of Economics 117, no. 4 (2002).
207 *Warren Buffett*: Alice Schroeder, The Snowball: Warren Buffett and the Business of Life (New York: Bantam, 2008).

## Chương 7

214 *Nhiều người tuyên bố có thể dự báo thị trường:* Johan Bollen, Huina Mao, and Xiaojun Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science* 2, no. 1 (2011).
214 *Quỹ phòng hộ dựa trên tweet đã đóng cửa chỉ sau 1 tháng:* James Mackintosh, "Hedge Fund That Traded Based on Social Media Signals Didn't Work Out," *Financial Times*, May 25, 2012.
216 *không tìm ra sự tương quan:* Christopher F. Chabris et al., "Most Reported Genetic Associations with General Intelligence Are Probably False Positives," *Psychological Science* (2012).
218 *Zoë Chance:* Câu chuyện này được thảo luận trong TEDx Talks, "How to Make a Behavior Addictive: Zoë Chance at TEDx Mill River," YouTube video, posted May 14, 2013, https://www.youtube.com/watch?v=AHfiKav9fcQ. Một số chi tiết của câu chuyện, như màu sắc của thiết bị đo bước đi được thêm vào từ các cuộc phỏng vấn. Tôi phỏng vấn Chance bằng điện thoại ngày 20/4/2015, và bằng email ngày 11/7/2016, và ngày 8/9/2016.
218 *Các con số có thể rất cám dỗ:* Phần này là từ Alex Peysakhovich and Seth Stephens-Davidowitz, "How Not to Drown in Numbers," *New York Times*, May 3, 2015, SR6.

219 *đã gian lận hoàn toàn trong khi thực hiện các bài kiểm tra:* Brian A. Jacob and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118, no. 3 (2003).
220 *Thomas Kane:* Tôi đã phỏng vấn Thomas Kane bằng điện thoại ngày 22/4/2015.
220 *"Mỗi thước đo đều bổ sung một điều gì đó có giá trị":* Bill and Melinda Gates Foundation, "Ensuring Fair and Reliable Measures of Effective Teaching," k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

## Chương 8

222 *Gần đây, 3 nhà kinh tế học:* Oded Netzer, Alain Lemaire, and Michal Herzenstein, "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," 2016.
222 *khoảng 13% người vay xù nợ:* Peter Renton, "Another Analysis of Default Rates at Lending Club and Prosper," October 25, 2012, http://www.lendacademy.com/lending-club-prosper-default-rates/.
225 *các like trên Facebook thường tương quan:* Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," *PNAS* 110, no. 15 (2013).
229 *các doanh nghiệp nằm dưới quyền sinh sát của các đánh giá trên Yelp:* Michael Luca, "Reviews, Reputation, and Revenue: The Case of Yelp," unpublished manuscript, 2011.
230 *các tìm kiếm Google liên quan đến tự tử:* Christine Ma-Kellams, Flora Or, Ji Hyun Baek, and Ichiro Kawachi, "Rethinking Suicide Surveillance: Google Search Data and Self-Reported Suicidality Differentially Estimate Completed Suicide Risk," *Clinical Psychological Science* 4, no. 3 (2016).
231 *khoảng 3.5 triệu tìm kiếm Google tại Mĩ liên quan đến tự tử:* Phần này dùng một phương pháp được thảo luận trên website của tôi ở các ghi chú về việc tự phá thai. Tôi so sánh các tìm kiếm trên Google mục "suicide" với các tìm kiếm "how to tie a tie." Có 6.6 triệu tìm kiếm Google cho "how to tie a tie" năm 2015. Có gấp 6.5 lần số tìm kiếm trong mục "suicide." 6.5*6.6/12 » 3.5.
231 *Có 12 vụ ám sát người Hồi giáo được cho là tội ác do thù ghét:* Bridge Initiative Team, "When Islamophobia Turns Violent: The 2016 U.S. Presidential Election," May 2, 2016, available at http://bridge.georgetown.edu/when-islamophobia-turns-violent-the-2016-u-s-presidential-elections/

## Kết luận

235 *Điều gì thúc đẩy cuộc thập tự chinh của Popper?:* Karl Popper, *Conjectures and Refutations* (London: Routledge & Kegan Paul, 1963).
237 *vẽ bản đồ mọi ca dịch tả trong thành phố:* Simon Rogers, "John Snow's Data Journalism: The Cholera Map That Changed the World," *Guardian*, March 15, 2013.
238 *Benjamin F. Jones:* Tôi đã phỏng vấn Benjamin Jones qua điện thoại ngày 1/6/2015. Nghiên cứu này cũng được thảo luận trong Aaron Chatterji và Benjamin Jones, "Harnessing Technology to Improve K–12 Education," Hamilton Project Discussion Paper, 2012.
245 *người ta thường không đọc hết các quyển chuyên luận của các nhà kinh tế học:* Jordan Ellenberg, "The Summer's Most Unread Book Is... ," *Wall Street Journal*, July 3, 2014.

# Chỉ Mục Từ

**MỌI NGƯỜI ĐỀU NÓI DỐI: DỮ LIỆU LỚN, DỮ LIỆU MỚI**
và những điều Internet tiết lộ về chính chúng ta

Tác giả
**Seth Stephens-Davidowitz**

Biên dịch
**Nguyễn Hạo Nhiên - Nguyễn Hưởng**

Chịu trách nhiệm xuất bản
**PGS.TS. Nguyễn Ngọc Định**

Biên tập
**Nguyễn Ngọc Định**

Trình bày bìa
**Nguyễn Hạo Nhiên**

Sửa bản in
**Trương Thị Thu Nga**

Mã số ISBN
**978-604-922-753-0**

Đơn vị liên kết xuất bản
**Công ty TNHH Ecoblader**
Địa chỉ: 168G Lưu Hữu Phước, Phường 15, Quận 8, TP. HCM
SĐT: 0868612291 - Email: contact@ecoblader.com

**Nhà xuất bản Kinh tế TP. Hồ Chí Minh**
Số 279 Nguyễn Tri Phương, Phường 5, Quận 10, TP. Hồ Chí Minh.
Website: www.nxb.ueh.edu.vn – Email: nxb@ueh.edu.vn
Điện thoại: (028) 38.575.466 – Fax: (028) 38.550.783

In 2000 cuốn, khổ 16x24 cm tại Công ty Cổ Phần In Khuyến Học Phía Nam. Địa chỉ: Lô B5-8 đường D4, KCN Tân Phú Trung, Củ Chi, TP.HCM. Số xác nhận ĐKXB: 2757-2019/CXBIPH/02-25/KTTPHCM. Quyết định số: 61/QĐ-NXBKTTPHCM cấp ngày 1/8/2019. In xong và nộp lưu chiểu Quý III/2019.