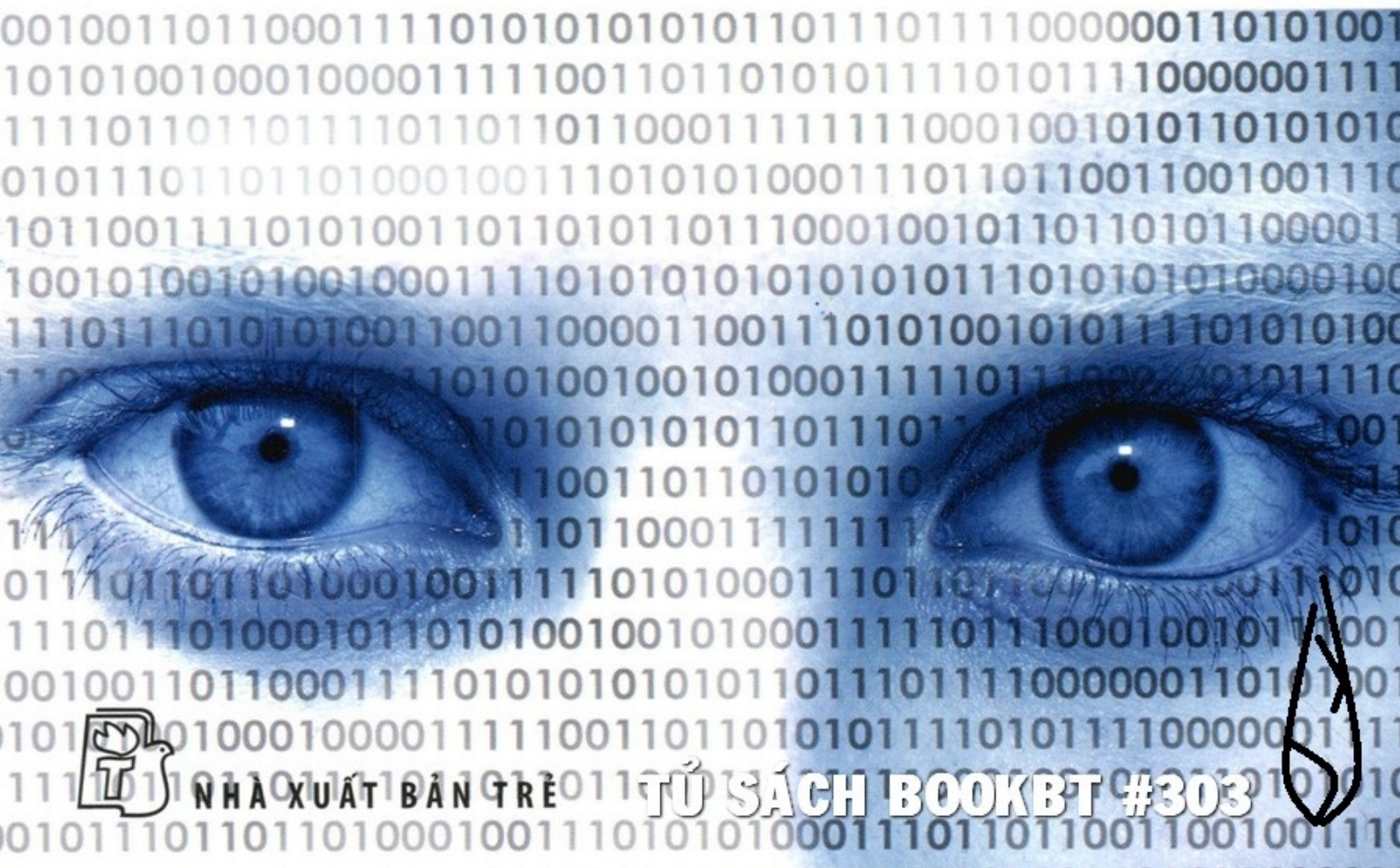
KHOA HỌC KHÁM PHÁ

Viktor Mayer-Schönberger
và Kenneth Cukier

Big Data
A revolution that will transform how we live, work, and think

DỮ LIỆU LỚN

Cuộc cách mạng sẽ làm thay đổi cách chúng ta sống, làm việc và tư duy



NHÀ XUẤT BẢN TRẺ

TỦ SÁCH BOOKBT #303

Viktor Mayer-Schönberger
và Kenneth Cukier
Vũ Duy Mẫn *dịch*

Big Data
A revolution that will transform how we live, work, and think

DỮ LIỆU LỚN

Cuộc cách mạng sẽ làm thay đổi cách chúng ta sống, làm việc và tư duy

NHÀ XUẤT BẢN ...
TỦ SÁCH BOOKBT #303

Thông tin sách

Tên sách: **Dữ liệu lớn (Tủ sách Khoa học Khám phá)**

Nguyên tác: **Big data**

Tác giả: **Viktor Mayer-Schonberger, Kenneth Cukier**

Người dịch: **Vũ Duy Mẫn**

Nhà phát hành: **NXB Trẻ**

Nhà xuất bản: **NXB Trẻ**

Khối lượng: **350g**

Kích thước: **14.5 x 20.5 cm**

Ngày phát hành: **344**

Số trang: **03/2014**

Giá bìa: **120.000đ**

Thể loại: **Khoa học Khám phá**

Thông tin ebook

Thực hiện ebook: **thanhbt**

Ngày hoàn thành: **22/08/2017**

Dự án ebook #303 thuộc Tủ sách BOOKBT 

Giới thiệu

Màu sơn nào có thể cho bạn biết một chiếc xe đã qua sử dụng vẫn còn trong tình trạng tốt? Làm thế nào các công chức ở thành phố New York có thể xác định các hố ga nguy hiểm nhất trước khi chúng phát nổ? Và làm thế nào những cuộc tìm kiếm của Google dự đoán được sự lây lan của dịch cúm H1N1? Chìa khóa để trả lời những câu hỏi này, và nhiều câu hỏi khác, là dữ liệu lớn. “Dữ liệu lớn” đề cập đến khả năng đang phát triển của chúng ta để nắm giữ các bộ sưu tập lớn thông tin, phân tích, và rút ra những kết luận đôi khi sâu sắc đáng ngạc nhiên.

Lĩnh vực khoa học đang nổi lên này có thể chuyển vô số hiện tượng - từ giá vé máy bay đến các văn bản của hàng triệu cuốn sách - thành dạng có thể tìm kiếm được, và sử dụng sức mạnh tính toán ngày càng tăng của chúng ta để khám phá những điều chúng ta chưa bao giờ có thể nhìn thấy trước. Trong một cuộc cách mạng ngang tầm với Internet hoặc thậm chí in ấn, dữ liệu lớn sẽ thay đổi cách chúng ta nghĩ về kinh doanh, y tế, chính trị, giáo dục, và sự đổi mới trong những năm tới. Nó cũng đặt ra những mối đe dọa mới, từ sự kết thúc không thể tránh khỏi của sự riêng tư cho đến khả năng bị trừng phạt vì những thứ chúng ta thậm chí còn chưa làm, dựa trên khả năng của dữ liệu lớn có thể dự đoán được hành vi tương lai của chúng ta.Trong tác phẩm thông tuệ tuyệt vời và gây nhiều ngạc nhiên này, hai chuyên gia hàng đầu giải thích dữ liệu lớn là những gì, nó sẽ làm thay đổi cuộc sống của chúng ta như thế nào, và những gì chúng ta có thể làm để bảo vệ chính mình khỏi các mối nguy hiểm của nó. Dữ liệu lớn là cuốn sách lớn đầu tiên về điều to lớn sắp diễn ra. Bạn

đọc có thể quét các QR Code bên trong sách và trên bìa sách để xem các đoạn phim minh họa.

Tặng B và V V.M.S.

Tặng cha mẹ của tôi

K.N.C.

1. HIỆN TẠI

NĂM 2009 MỘT VI-RÚT CÚM mới được phát hiện. Kết hợp các yếu tố của các vi-rút gây cúm gà, chủng mới này, được gọi là H1N1, đã lây lan nhanh chóng. Trong vài tuần, các cơ sở y tế khắp thế giới lo sợ một đại dịch khủng khiếp đang xảy ra. Một số nhà bình luận đã cảnh báo về một dịch bệnh có quy mô của dịch cúm Tây Ban Nha vào năm 1918, lây nhiễm cho nửa tỷ người và làm chết hàng chục triệu người. Tồi tệ hơn là không hề có vắc-xin nào để chống lại vi-rút mới này. Hy vọng duy nhất của cơ quan y tế là giảm mức lây lan. Nhưng để làm điều đó, họ cần biết bệnh đã lan tới đâu.

Ở Mỹ, Trung tâm Kiểm soát và Phòng chống Bệnh dịch (CDC) đã yêu cầu các bác sĩ thông báo về các ca bệnh cúm mới. Nhưng bức tranh thật về đại dịch vẫn luôn bị chậm trễ một hoặc hai tuần. Nhiều người có thể bị bệnh vài ngày rồi mới đi gặp bác sĩ. Việc chuyển tiếp thông tin về các cơ quan trung ương đòi hỏi thời gian, và CDC chỉ xử lý các con số mỗi tuần một lần. Với một bệnh dịch lây lan nhanh, hai tuần chậm trễ cũng giống như dài vô tận. Sự chậm trễ này đã hoàn toàn vô hiệu hóa các cơ quan y tế tại những thời điểm gay cấn nhất.

Lúc việc đó xảy ra, vài tuần trước khi vi-rút H1N1 xuất hiện rầm rộ trên các phương tiện truyền thông, các kỹ sư của công ty Internet khổng lồ Google đã đăng một bài đáng chú ý trên tạp chí khoa học *Nature*. Nó đã tạo một chuyện giật gân trong giới chức y tế và các nhà khoa học máy tính, nhưng ngoài ra thì ít được quan tâm. Các tác giả lý giải Google có thể "dự đoán" sự lây lan của bệnh cúm mùa đông ở Mỹ như thế nào, không chỉ ở mức độ toàn quốc, mà còn chi tiết tới mức vùng và thậm chí tới mức tiểu bang. Google có thể đạt được điều này bằng cách xem xét những gì người sử dụng đã tìm kiếm trên Internet. Bởi Google nhận được hơn ba tỷ câu hỏi tìm kiếm mỗi ngày và lưu giữ tất cả chúng, nên nó có vô số dữ liệu để phân tích.

Google lấy 50 triệu cụm từ được tìm kiếm phổ biến nhất của người Mỹ và so sánh chúng với dữ liệu của CDC về sự lây lan của bệnh cúm mùa giữa các năm 2003 và 2008. Ý tưởng là để xác định các khu vực bị lây nhiễm vi-rút cúm thông qua những gì người ta tìm kiếm trên Internet, và không ai khác có nhiều dữ liệu, năng lực tính toán và hiểu biết về thống kê như Google.

Dù các chuyên viên của Google phỏng đoán các lệnh tìm kiếm có thể nhằm thu lượm thông tin về cúm - gõ các câu đại loại như "thuốc ho và sốt" - nhưng không phải vậy: họ không biết, và họ đã thiết kế một hệ thống không quan tâm tới điều đó. Tất cả những gì hệ thống của họ làm là phát hiện mối tương quan giữa tần suất của một số câu hỏi tìm kiếm và sự lây lan của bệnh cúm theo thời gian và không gian. Tổng cộng, họ xử lý một lượng đáng kinh ngạc 450 triệu mô hình toán học khác nhau để kiểm tra các điều kiện tìm kiếm, so sánh các dự đoán của họ với các trường hợp bệnh thực tế từ CDC trong năm 2007 và 2008. Và họ đã vớ được vàng: phần mềm của họ tìm thấy một sự kết hợp của 45 điều kiện tìm kiếm mà khi sử dụng cùng với một mô hình toán học, có một mối tương quan mạnh mẽ giữa phỏng đoán của họ và các số liệu chính thức trên toàn quốc. Giống như CDC, họ có thể cho biết cúm đã lây lan tới đâu, nhưng khác với CDC, họ có thể nói điều đó gần như trong thời gian thực, chứ không phải trễ một hoặc hai tuần.

Do vậy, khi dịch bệnh H1N1 tấn công vào năm 2009, hệ thống của Google đã chứng tỏ là một chỉ báo có ích hơn và nhanh hơn so với các số liệu thống kê của chính phủ thường chậm trễ. Các quan chức y tế đã được trang bị những thông tin có giá trị.

Điều gây ấn tượng là phương pháp của Google không liên quan gì đến việc phân phối gạc miệng hoặc liên hệ với các phòng khám. Thay vào đó, nó được xây dựng trên "dữ liệu lớn" - khả năng của xã hội khai thác thông tin theo những cách thức mới để đưa ra

những kiến thức hữu ích hay những sản phẩm và dịch vụ có giá trị đáng kể. Với nó, khi đại dịch kế tiếp xảy ra, thế giới sẽ có sẵn một công cụ tốt hơn để dự đoán và do đó ngăn chặn sự lây lan.

Phim minh họa phương pháp của Google

Y tế công chỉ là một lĩnh vực trong đó dữ liệu lớn đang làm nên một sự khác biệt vĩ đại. Nhiều lĩnh vực khác cũng đang được định hình lại bởi dữ liệu lớn. Dịch vụ mua vé máy bay là một thí dụ.

Năm 2003, Oren Etzioni cần bay từ Seattle tới Los Angeles để dự lễ cưới em trai của ông. Nhiều tháng trước đó, ông lên mạng và mua một vé máy bay, tin rằng càng mua sớm, vé càng rẻ. Trên chuyến bay, do tò mò, Etzioni hỏi người ngồi kế bên xem giá vé của ông ta là bao nhiêu và ông ta mua khi nào. Hóa ra ông ta trả thấp hơn nhiều so với Etzioni, mà thậm chí ông ta mới chỉ mua vé gần đây. Khá tức giận, Etzioni hỏi một hành khách khác và

một hành khách khác nữa. Hầu hết họ đã trả ít tiền hơn.

Với hầu hết chúng ta, ý nghĩa của cảm giác bị lừa có thể đã tiêu tan khi chúng ta gấp khay bàn ăn trước mặt, dựng thẳng ghế và khóa thắt lưng an toàn. Nhưng Etzioni là một trong những nhà khoa học máy tính hàng đầu của Mỹ. Ông nhìn thế giới như một chuỗi các bài toán dữ-liệu-lớn có thể giải được. Và ông đang làm chủ chúng từ khi là người đầu tiên tốt nghiệp Đại học Harvard về chuyên ngành khoa học máy tính vào năm 1986.

Từ căn phòng của mình tại Đại học Washington, ông đã khởi xướng những công ty dữ-liệu-lớn trước khi thuật ngữ "dữ liệu lớn" được biết tới. Ông đã giúp phát triển một trong những công cụ tìm kiếm Web đầu tiên, MetaCrawler, được đưa ra sử dụng vào năm 1994 rồi sau được bán cho InfoSpace, lúc đó là một công ty bất động sản trực tuyến lớn. Ông đã đồng sáng lập Netbot, trang web mua hàng so sánh lớn đầu tiên, sau đó bán nó cho Excite. Ông khởi động công ty làm công cụ trích ý nghĩa từ các văn bản, gọi là ClearForest, sau này được Reuters mua lại.

Trở lại câu chuyện chính, Etzioni quyết tìm ra cách để có thể biết liệu một giá vé ta thấy trên mạng có phải là một giá tốt hay không. Một chỗ ngồi trên máy bay là một thương phẩm: mỗi chỗ về cơ bản là hoàn toàn giống với những chỗ khác trên cùng chuyến bay. Nhưng giá lại rất khác nhau, dựa trên vô số yếu tố mà chủ yếu chỉ chính các hãng bay mới biết.

Etzioni đi đến kết luận ông không cần giải mã ý nghĩa hay nguyên nhân giá cả khác nhau. Thay vào đó, ông đơn giản phải dự đoán liệu giá được báo có khả năng tăng hay giảm trong tương lai. Điều này là khả thi, nếu không nói là dễ. Những gì cần thiết là phân tích tất cả các vé bán cho một tuyến đường và khảo sát các giá phải trả tương quan với số ngày mua trước lúc khởi hành.