XU HƯỚNG HIỆN TẠI VÀ TƯƠNG LAI NHỮNG THÁCH THỨC TRONG THỐNG KÊ HỌC: DỮ LIỆU LỚN

(Trích Báo cáo Khoa học London năm 2014: Hội thảo Tương lai của Thống kê học)

Xu hướng hiện tại trong thống kê được đề cập nhiều tại Hội thảo Tương lai của Thống kê học là Dữ liệu lớn (Big data), điều này không có gì phải nghi ngờ. Dữ liệu lớn hiện diện ở khắp mọi nơi và mỗi đối tượng lại có những suy nghĩ khác nhau khi nghe về chúng. Đối với những người bình thường, Big data mang lại những câu hỏi về sự riêng tư và bảo mật thông tin như: Những thông tin nào của tôi được chia sẻ, và làm thế nào để mọi người có thể truy cập vào đó? Còn đối với các nhà khoa học máy tính, thì câu hỏi đặt ra lại về vấn đề lưu trữ và quản lý dữ liệu, giao tiếp và tính toán. Và đối với các nhà thống kê, Big data đưa ra một tập danh sách các vấn đề hoàn toàn khác nhau: Làm thế nào chúng ta lấy những thông tin có thể sử dụng được ra khỏi cơ sở dữ liệu rất lớn và phức tạp mà phương pháp truyền thống của chúng ta không thể xử lý?

Tại hội thảo này, tất cả các quan điểm đã được trình bày, từ quan điểm "Big data là một cơ hội mà các nhà thống kê không nên bỏ lỡ" đến quan điểm "Big data sẽ thay đổi số liệu thống kê như là chúng ta biết đến nó" hay quan điểm trái ngược như "Big data chỉ là nhất thời và chúng ta không nên đầu tư vào như những gì phóng đại về nó". Nội dung Phần này của báo cáo sẽ không thừa nhận bất cứ quan điểm nhìn nhận nào nhưng sẽ tóm tắt qua tình hình và những thách thức mà Big data đặt ra đối với Khoa học Thống kê.

Ví dụ về Big data

Một số hình thức phổ biến nhất của Big data là:

- Cơ sở dữ liệu thương mại, chẳng hạn như dữ liệu số nhân viên của Google hay Facebook.
- Chính phủ hay gắn với số liệu chính thức.
- Cơ sở dữ liệu Hệ gen người. Ví dụ, Một bộ gen của con người có hơn 3 tỷ cặp gen cơ bản. Hay dữ liệu của 1000 dự án về Hệ gen người, hiện tại đã thu thập được 200 terabytes (200 nghìn tỷ byte) dữ liệu.
- Số liệu về bộ não người. Một lần quét qua não bộ một người bao gồm các dữ liệu về hơn 200.000 điểm mạng lưới thần kinh não bộ, mà có thể được đo nhiều lần với hơn 300 lần đo về thời gian. Dự án Hệ kết nối con người, đó là thu thập hình ảnh

MRI của 1.200 bệnh nhân trong khoảng thời gian 5 năm và đã đưa ra công khai số liệu 7 tỷ byte dữ liệu tính đến đầu năm 2014.

- Cơ sở dữ liệu vật lý và thiên văn học. Ví dụ, thí nghiệm máy va chạm Hadron gia tốc hạt lớn tạo ra hơn 1 petabyte (1.000 nghìn tỷ byte) dữ liệu mỗi năm. Hoặc là The Large Synoptic Survey Telescope (LSST) hay còn gọi là kính thiên văn Synoptic, dự kiến hoạt động vào năm 2020, sẽ tạo ra 1 petabyte dữ liệu mỗi đêm.

Không chỉ là Lớn, mà còn rất phức tạp

Các nhà thống kê cho rằng, Big data thách thức một số mô hình cơ bản. Ví dụ như vấn đề: "p lớn, n nhỏ", (trong đó "p" không phải là một giá trị mà là số lượng các biến, n là số lượng các điểm dữ liệu). Thống kê truyền thống cung cấp phương pháp để phân tích dữ liệu khi số lượng các biến p là nhỏ và số lượng các điểm dữ liệu n là lớn. Ví dụ, một nhà sinh vật học muốn ước tính số lượng của một loài cá cụ thể trên một dòng sông khác nhau, phụ thuộc vào các biến như độ sâu của sông, kích thước của các lưu vực sông, hàm lượng ôxy trong nước, và nhiệt độ nước. Đây là bốn biến (p=4), và các dữ liệu có thể được lấy từ 50 hoặc 100 dòng sông khác nhau (n ≈ 50:100). Nói rằng, kích thước mẫu n vượt xa số lượng các biến p.

Trong một số những ứng dụng Big data (mặc dù không phải là tất cả), tình trạng này lại ngược lại. Nghiên cứu về gen người, các nhà nghiên cứu thu thập dữ liệu trên 100 bệnh nhân ung thư (n=100) để xác định gen gây ra nguy cơ ung thư. Thật không may, có 20.000 gen trong hệ gen người (p > 20.000) và các biến thể của gen có nhiều hơn thế. Nghiên cứu mở rộng bộ gen người nhìn vào nửa triệu "SNPs" (SNPs là chuỗi đơn nucleotide) hoặc vị trí trên hệ gen người, nơi sự thay đổi có thể xảy ra. Số lượng các biến (p = 500.000) lớn hơn nhiều so với cỡ mẫu (n = 100). Tương tự như vậy, trong các nghiên cứu về thần kinh, các biến (p) tương ứng với các điểm mạng lưới thần kinh hoặc khu vực trung tâm mạng lưới thần kinh thường lớn hơn rất nhiều lần số người (n) tham gia trong cuộc khảo sát.

Trong cả hai trường hợp nghiên cứu trên, mục đích là để phát triển một mô hình mô tả rõ mối quan hệ của biến kết quả (ví dụ như quy mô của số lượng cá, hay là sự hiện diện của bệnh ung thư) là liên quan đến các biến p khác và để xác định các biến quan trọng trong mối quan hệ đó. Mô hình này phụ thuộc vào các thông số, đó là xác định số lượng các mối quan hệ cho mỗi biến phù hợp với mô hình dữ liệu liên quan đến việc ước tính các thông số từ dữ liệu và đánh giá các bằng chứng cho thấy chúng khác nhau từ số không (phản ánh một biến quan trọng). Khi p lớn hơn n, số lượng các biến là rất lớn liên quan đến các thông tin về người tham gia. Hàng nghìn các thông số liên quan sẽ xuất hiện với ý nghĩa thống kê nếu sử dụng dữ liệu thống kê

nhỏ. Trong thống kê truyền thống, nếu dữ liệu chứa một cái gì đó mà có cơ hội (1/1.000.000) xảy ra, thì ta có thể chắc chắn rằng nó không phải tình cờ. Nhưng nếu bạn nhìn vào 500.000 điểm dữ liệu, như trong thế giới Big data thì nó không phải là điều bất thường để bạn thực hiện khám phá một điểm trong một triệu điểm dữ liệu đó. Đây là cơ hội không thể bị bỏ lỡ như một lời giải thích về quan điểm "Big data là một cơ hội mà các nhà thống kê không nên bỏ lỡ". Các nhà thống kê gọi đây là hiệu ứng "sự hiện diện ở khắp mọi nơi", và nó là một vấn đề lớn đối với Big data, chứ không phải là giả thuyết, hay phương pháp tiếp cận khoa học và các nhà thống kê đã tìm thấy một số cách tốt để giải quyết với hiệu ứng này.

Hầu hết các bộ dữ liệu chỉ có một vài mối quan hệ rõ ràng giữa các biến, và mọi thứ khác đều không rõ ràng. Vì vậy, ta không nên quan trọng hóa hầu hết các tham số đơn giản. Một cách để làm cho các tham số không quan trọng là việc giả định tất cả những tham số đó bằng không. Một số tiến bộ kỹ thuật trong những năm gần đây đã thực hiện được điều này để trích xuất thông tin cần thiết có ý nghĩa từ một khối dữ liệu. Kỹ thuật này được gọi là L1-minimization, hoặc Lasso, được phát minh bởi Robert Tibshirani vào năm 1996. Thật trùng hợp, L1-minimization được phát hiện gần như đồng thời trong lĩnh vực xử lý hình ảnh, cho phép khai thác một hình ảnh sắc nét từ dữ liệu thiếu thông tin hoặc không rõ ràng.

Một kỹ thuật được áp dụng rộng rãi, đặc biệt là nghiên cứu Hệ gen người và chẩn đoán hình ảnh đó là tỷ lệ phát hiện sai (FDR) của Yoav Benjamini và Yosi Hochberg được đề xuất năm 1995, là một phương pháp thay thế để kiểm tra ý nghĩa về thống kê có tính đến việc xem xét sự hiện diện các biến ở khắp mọi nơi có hiệu quả. Ví dụ, nếu một nghiên cứu tìm thấy 20 SNPs (chuỗi đơn nucleotide) là tác nhân có ý nghĩa gây ra ung thư, và nó có một tỷ lệ phát hiện sai FDR là 10%, thì sau đó bạn mong đợi hai trong số 20 SNPs "phát hiện" là sai, trên mức trung bình. Như vậy, FDR không cho bạn biết đó là những người thân của mình và (nếu có) thì có thể là giả mạo, nhưng đôi khi việc đó có thể được xác định bởi một nghiên cứu tiếp theo.

Những ví dụ này chỉ ra rõ ràng rằng Big data chỉ là một thách thức hay một sự phiền toái. Đây cũng là một cơ hội cho các nhà thống kê đánh giá lại các giả định của họ và mang lại những ý tưởng mới về Big data.

n lớn, p lớn, t (thời gian) nhỏ

Khi cả hai n và p phát triển, các nhà thống kê phải đối mặt với khó khăn khác mà họ chưa bao giờ gặp trước đây đó là: áp lực của thời gian (t).

Thống kê truyền thống luôn được thực hiện trong một chế độ ngoại tuyến, và rất nhiều lý thuyết đã được phát triển trong thời kỳ này (những năm 1900) khi mà "trực tuyến" chưa tồn tại. Các nhà nghiên cứu thu thập dữ liệu, sau đó tiến hành phân tích. Thời gian để phân tích về cơ bản là không giới hạn, và các dữ liệu đã được chia nhỏ. Do đó, các nhà thống kê không bao giờ phải suy nghĩ về việc thực hiện tính toán một cách hiệu quả.

Tuy nhiên, trong thời đại của Big data mọi thứ đã khác. Các công ty web kiếm tiền bằng cách cố gắng hướng những dự đoán của người dùng và khuyến khích các hành vi sử dụng nhất định (như khi nhấp chuột vào một quảng cáo được tài trợ bởi một khách hàng). Để dự báo tỷ lệ phản ứng, cần mô hình thống kê với n lớn (hàng triệu lần nhấp chuột) và p lớn (hàng chục nghìn biến, trong đó các biến như: các quảng cáo để chạy, vị trí để đặt trên trang web, vv...). Trong trường hợp này, n lớn hơn nhiều so với p, do đó kỹ thuật thống kê cổ điển có thể áp dụng được trong lý thuyết.

Thống kê tài chính

Cuộc khủng hoảng tài chính năm 2007 và 2008, sau đó là cuộc suy thoái kinh tế đã làm sáng tỏ một số sai sót cơ bản trong việc đánh giá rủi ro. Chỉ số Basel là để giám sát yêu cầu các tổ chức tài chính báo cáo rủi ro các khoản đầu tư của họ. Tuy nhiên, chỉ số này đã trở thành tiêu chuẩn (đo giá trị rủi ro) trong cộng đồng tài chính, vì lý do đơn giản là sự thuận tiện của một phép đo được phát triển bởi một công ty mà có rất ít lý thuyết thống kê về nó. Ngoài ra, các phương pháp được sử dụng để đánh giá chuyển đổi tín dụng (chẳng hạn như đoạn chuỗi nối Gaussian) đã có những thiếu sót, và tương tự như vậy nó cũng đã trở nên phổ biến đơn giản vì "những người khác cũng đang làm việc đó". Đặc biệt, các phương pháp này cho rằng sự kiện là nhất định, chẳng hạn như mặc định của một chủ sở hữu nhà ở LasVegas và một chủ sở hữu nhà ở Miami là độc lập, ít bị phụ thuộc ít, hoặc chủ yếu phụ thuộc. Trong điều kiện bình thường, đây là một giả định không gây hại. Tuy nhiên, trong cơn khủng hoảng tài chính, sự kiện như vậy trở nên có tương quan. Một số thống kê có vẻ như là đã cảnh báo từ lâu trước năm 2007 về việc vô cùng nguy hiểm này, họ đã được nghe thông qua vắn tắt, nhưng lại bỏ qua nó.

Nó vẫn còn phải chờ xem liệu tổ chức tài chính có thể học được từ chính bản thân cảnh sát, nhưng đó là công bằng để nói rằng nếu tồn tại một giải pháp, thì nó sẽ đòi hỏi một cách tiếp cận ít nợ nhất để tập trung nhiều hơn nghe về những nguyên tắc thống kê.

Tuy nhiên, các thuật toán cần thiết để thực hiện một phân tích hồi quy như vậy là không tốt. Các công ty web chỉ có (1/100 giây) để quyết định làm thế nào để đáp

ứng với lần nhấp chuột của người dùng. Không chỉ vậy các mô hình phải liên tục thay đổi để thích ứng với người dùng mới và sản phẩm mới. Internet là một thử nghiệm giống như vậy, với quy mô lớn, liên tục thay đổi và không bao giờ kết thúc.

Để giải quyết vấn đề về thời gian, các nhà thống kê đã bước đầu chấp nhận và thích ứng với ý tưởng của các nhà khoa học máy tính, những người luôn đặt ra vấn đề về tốc độ xử lý thông tin lên hàng đầu. Mục tiêu trong một số trường hợp có thể không phải là cung cấp được một câu trả lời hoàn hảo, mà là để cung cấp một câu trả lời nhanh và tốt nhất.

Tuy nhiên, các nhà thống kê không đủ khả năng để ngăn chặn lại những dòng suy nghĩ của mình cùng một lúc, vì họ có những chuyên môn riêng mà các công ty của các nhà khoa học máy tính (phần lớn) là không thể hiểu hết và cho rằng: các nhà thống kê hiểu không chắc chắn. Trong trường hợp nhà khoa học máy tính nhìn thấy một khả năng, thì nhà thống kê nhìn thấy một loạt các khả năng. Dự đoán được tốt hơn khi chúng được coi như là dự báo, và vốn dĩ trong đó đã có một sự không chắc chắn, nhưng cũng chỉ có các nhà thống kê là duy nhất đủ điều kiện thực hiện suy luận trừu tượng để kết nối giữa các biến từ dữ liệu và xác định kết nối đó là có thật hay là giả mạo.

Một cách để làm phương pháp thống kê hiệu quả hơn là để họ thực hiện song song giữa việc viết các thuật toán chạy trên nhiều máy tính hoặc nhiều bộ vi xử lý cùng một lúc. Tại hội thảo ở London, Michael Jordan đã trình bày một trường hợp nghiên cứu ứng dụng như vậy được gọi là Bag of Little Bootstraps (BLB). "Bootstrap" là một phương pháp chuẩn, được phát minh bởi Bradley Efron vào năm 1979, cũng giống như nhiều phương pháp thống kê, đó là tính toán chuyên sâu. Năm 2012, Jordan và một nhóm đồng nghiệp thực hiện ứng dụng BLB trên nền tảng điện toán đám mây của Amazon với những bộ dữ liệu khác nhau trong các miền công cộng. Kết quả mà BLB tạo ra thường xuyên được so sánh với Bootstrap, và được tạo ra nhanh hơn. Và những người ủng hộ việc nghiên cứu rất thích điều này như các công ty công nghệ cao: Amazon, Google, SAP, Cisco, Oracle, v.v... Đây là điều công bằng để nói rằng, Silicon Valley đã nhận thấy sự cần thiết phải có các công cụ thống kê mới để đối phó với Big data.

Các loại dữ liệu mới

Hiện nay đã xuất hiện một xu hướng song song với Big data là dữ liệu kiểu mới. Các dữ liệu này là những con số không đơn giản, có thể dưới hình thức của một chức năng, hình ảnh, hình dạng, hoặc mạng. Ví dụ, thế hệ đầu tiên "dữ liệu chức năng" có thể là một chuỗi thời gian, các phép đo về ôxy trong máu được chụp tại một

điểm cụ thể và vào những thời điểm khác nhau. Ngược lại với dữ liệu truyền thống đa biến hoặc không cấu trúc, các chức năng quan sát trong trường hợp này là một mẫu từ một không gian dữ liệu đa chiều (vì nó liên quan đến sự hiểu biết về quá trình ôxy hóa ở nhiều thời điểm). Dữ liệu "đa chiều" như vậy yêu cầu các phương pháp phức tạp hơn.

Nhưng đó chưa phải là tất cả, tại hội thảo ở London, Jane Ling Wang nói về dữ liệu chức năng thế hệ mới, trong đó bao gồm tương quan chức năng ngẫu nhiên của dữ liệu. Các chức năng có giá trị có thể bằng con số, hình ảnh hoặc hình dạng. Ví dụ, quan sát bộ não tại thời điểm t có thể đó là dữ liệu của não bộ về khu vực đang hoạt động tại thời điểm t.

Dữ liệu hình ảnh não bộ và hệ thần kinh người là những ví dụ điển hình của dữ liệu chức năng thế hệ mới, và đó là những trọng tâm của hai sáng kiến nghiên cứu gần đây về lập bản đồ não người, trong đó, một là của chính quyền Obama và một là của Liên minh châu Âu. Các dự án này nhằm mục đích lập bản đồ các hoạt động tế bào thần kinh của não bộ toàn thể nhân loại và tìm hiểu bộ não con người hoạt động như thế nào.

Dữ liệu chức năng thế hệ tiếp theo là Big data, nó không chỉ lớn, mà còn rất phức tạp. Chúng yêu cầu các phát minh, ý tưởng mới dựa trên toán học, những quy ước như thống kê học, chẳng hạn hình học (là để mô tả hình dạng trừu tượng) hoặc cấu trúc liên kết (để mô tả không gian mà các dữ liệu được lấy từ mẫu). Đây là những ví dụ về các cách thức mà Big data không chỉ thách thức, mà còn làm phong phú thêm việc thực hành thống kê.

Riêng tư và Bảo mật

Năm 2013 là năm mà nhiều người Mỹ để ý tới những dữ liệu đang được thu thập về bản thân họ với khối lượng lớn, nhờ vào sự công bố công khai chương trình khai thác dữ liệu do Cơ quan an ninh quốc gia thực hiện được gọi là PRISM. Trong bối cảnh này, vấn đề về sự riêng tư và bảo mật dữ liệu cá nhân là mối quan tâm của các cá nhân và nó trở nên nghiêm trọng hơn. Nó sẽ dễ dàng cho các nhà thống kê khi nói rằng: "Không phải vấn đề của chúng tôi", nhưng, trên thực tế thì họ có thể là một phần của các giải pháp cho những vấn đề trên.

Hai cuộc hội thảo tại London, được đưa ra bởi Stephen Fienberg và Cynthia Dwork, tập trung vào các vấn đề riêng tư và bảo mật. Fienberg khảo sát lịch sử về sự bảo mật và chỉ ra một điều đơn giản, nhưng không thực tế và rõ ràng: Theo như tài liệu có liên quan của Chính phủ, quá khứ là tồi tệ hơn nhiều so với hiện tại. Hồ sơ Cục điều tra dân số Mỹ không được bảo mật cho đến năm 1910. Tính bảo mật pháp

lý được đưa ra trong hai thập kỷ tiếp theo, đầu tiên là để bảo vệ các doanh nghiệp và sau đó là để bảo vệ cho các cá nhân.

Tuy nhiên, Luật Chiến tranh thế giới thứ hai (năm 1942), đã xóa bỏ những bảo đảm pháp lý này. Block-by-block hay còn gọi là khối dữ liệu đã được sử dụng để xác định các khu vực mà người Mỹ gốc Nhật đang sống, và hồ sơ điều tra dân số cá nhân được cung cấp nhiều lần cho cơ quan pháp luật như cơ quan mật vụ hay Cục Điều tra Liên bang. Các hành động đó đã được xóa bỏ vào năm 1947, nhưng thiệt hại của nó là sự tin tưởng của người dân thì không thể lấy lại một cách dễ dàng.

Có nhiều cách để tổ chức bảo quản hồ sơ sau khi được thu thập mà không gây ảnh hưởng đến các thông tin cá nhân, chẳng hạn như cuộc điều tra dân số. Những phương pháp bao gồm việc thêm dữ liệu ngẫu nhiên (báo cáo của người viết Blog A có thu nhập 50.000 đô la mỗi năm và máy tính xử lý cho thêm vào nó một số ngẫu nhiên - 10.000 đô la, nói rằng con số này được rút ra từ bảng phân phối các giá trị ngẫu nhiên); trao đổi dữ liệu (số lượng người phụ thuộc vào người A được trao đổi với người B); hoặc ma trận mặt nạ (toàn bộ dữ liệu p biến của n người được biến đổi bởi hiểu biết bản chất hoạt động trong toán học đã "làm mờ" dữ liệu của tất cả những người xung quanh cùng một lúc). Các nhà thống kê, bao gồm những người ở Cục điều tra dân số Hoa Kỳ đã có nhiều việc làm cụ thể trong việc xây dựng các cơ chế và tính chất của các phương pháp này để tăng tính bảo mật thông tin cá nhân.

Mật mã học là một chuyên ngành khác, áp dụng toán học để biến đổi dữ liệu làm cho dữ liệu không thể bị lộ, hay lộ ra chỉ với một mật khẩu duy nhất, hoặc lộ ra với chi phí lớn mà đối thủ không có đủ khả năng chi trả. Từ những năm 1970, mật mã học đã được thông qua những thay đổi riêng. Khi mật mã học là khoa học về che giấu, chúng có thể được tạo nên bởi Chính phủ, một số ít là gián điệp hay quân đội. Còn hiện nay có nhiều hơn để bảo mật thông tin, và có sẵn cho tất cả mọi người, bất cứ ai sử dụng một thẻ ngân hàng tại máy ATM là đang sử dụng mật mã hiện đại.

Một trong những xu hướng hấp dẫn nhất về Big data trong thập kỷ qua là phát triển sự hợp tác giữa các số liệu thống kê và các cộng đồng mật mã. Dwork, một nhà mật mã học, phát biểu tại Hội thảo về phân biệt bảo mật, một cách tiếp cận mới cung cấp đồng thời đảm bảo phân biệt bảo mật với xác suất chắc chắn trong khi thừa nhận an ninh hoàn hảo là điều không thể. Phân biệt bảo mật cung cấp một cách để đo lường an ninh để nó trở thành một loại hàng hóa: Cũng giống như một người dùng có thể mua càng nhiều để bảo mật cho dữ liệu của mình khi họ cần.

Tuy nhiên, trước nhiều thách thức về sự bảo mật và riêng tư, và các vấn đề đó không có nghĩa là được giải quyết. Hầu hết các phương pháp ẩn danh không làm

được với p hay n có quy mô lớn. Hoặc là họ thêm rất nhiều dữ liệu không rõ ràng mà các phân tích mới trở nên gần như không thể thực hiện được hoặc họ làm giảm đi sự bảo mật riêng tư. Dữ liệu mạng như là một thách thức đặc biệt đối với sự bảo mật riêng tư bởi vì rất nhiều thông tin đã thiết lập với các mối quan hệ giữa các cá nhân. Tóm lại, có vẻ như "không có bữa ăn trưa miễn phí" trong sự cân bằng giữa bảo mật riêng tư và thông tin.

Và đối với các nhà thống kê, Big data giới thiệu một tập danh sách về các vấn đề hoàn toàn khác nhau: Làm thế nào chúng ta lấy những thông tin có thể sử dụng được ra khỏi cơ sở dữ liệu rất lớn và phức tạp mà nhiều phương pháp truyền thống của chúng ta không thể xử lý?

Chất lượng dữ liệu

Một trong những dịch vụ cung cấp được đánh giá thấp trong thế giới Big data mà các nhà thống kê có thể nhìn vào là chất lượng của dữ liệu thống kê với một con mắt hoài nghi. Truyền thống này là ăn sâu trong cộng đồng thống kê, đã được thử nghiệm đầu tiên vào những năm 1940. Dữ liệu đi kèm với một nguồn gốc xuất xứ. Nếu dữ liệu có xuất xứ từ một thử nghiệm ngẫu nhiên với các yếu tố gây sai lệch có thể xác định và kiểm soát, sau đó dữ liệu này có thể được sử dụng để suy luận thống kê. Nếu dữ liệu có xuất xứ từ một thí nghiệm được thiết kế kém hoặc thậm chí tệ hơn, hay nếu họ cứ làm đầy dữ liệu vào một máy chủ web của công ty mà không suy nghĩ theo những thiết kế của thí nghiệm thì các dữ liệu sẽ giống hệt nhau và có thể là vô giá trị. Trong thế giới của Big data, ai đó đã đặt câu hỏi như sau:

- Các dữ liệu thu thập theo cách dựa vào khía cạnh ưa thích? Hầu hết các dữ liệu thu thập trên Internet, trên thực tế, đi kèm với việc lấy mẫu nghiêng theo ý thích của người làm thu thập dữ liệu. Những người điền vào phiếu một cuộc điều tra không nhất thiết phải đại diện cho toàn bộ dân số nói chung. (Xem tại phần 3 của Báo cáo này "Tại sao theo khía cạnh ưa thích").

- Có những dữ liệu bị mất hoặc không đầy đủ? Trong các ứng dụng Web, thường có một số lượng lớn các dữ liệu không rõ ràng. Ví dụ, trang web của bộ phim Netflix muốn giới thiệu phim mới để cho người dùng xem bằng cách sử dụng một mô hình thống kê, nhưng nó chỉ có ít thông tin về các bộ phim mà người dùng đã đánh giá. Họ đã dành một triệu đô la làm giải thưởng cho một cuộc thi để xác định được một cách tốt hơn cho việc đó.

- Có các loại dữ liệu khác nhau? Nếu dữ liệu từ nhiều nguồn khác nhau, trong đó một số dữ liệu có thể là đáng tin cậy hơn những dữ liệu khác. Nếu tất cả những con số được đưa vào cùng một công cụ phân tích, giá trị của dữ liệu chất lượng cao

sẽ giảm bởi các dữ liệu chất lượng thấp hơn. Mặt khác, ngay cả dữ liệu chất lượng thấp có thể chứa một số thông tin có giá trị. Ngoài ra, những dữ liệu ở các định dạng khác nhau như con số, văn bản, mạng lưới các "điểm" hoặc các siêu liên kết. Nó có thể không được rõ ràng đối với các nhà thu thập dữ liệu là làm thế nào để tận dụng lợi thế của các loại thông tin không truyền thống này.

Các nhà thống kê không chỉ biết làm thế nào để hỏi những câu hỏi đúng, nhưng, tùy thuộc vào câu trả lời, họ có thể đã có sẵn các giải pháp thực tế.

Nhà thống kê bởi bất kỳ tên gọi khác...

"Nhà khoa học dữ liệu là người có thể tính toán giỏi hơn so với bất kỳ nhà thống kê nào, và có thể phân tích dữ liệu tốt hơn so với bất kỳ nhà khoa học máy tính nào" - Josh Wills (Giám đốc khoa học dữ liệu, Cloudera)

"Nhà khoa học dữ liệu là nhà thống kê có thể hiểu những bộ phận của khoa học máy tính và toán học có liên quan đến việc thao tác dữ liệu" - David Hand (nhà thống kê, Đại học Imperial College London)

"Nhà khoa học dữ liệu là một nhà thống kê có năng lực" - Hadley Wickham (nhà thống kê, Đại học Rice)

Đối với tất cả những lý do được liệt kê ở phần trước thì các nhà thống kê có thể cung cấp dữ liệu có giá trị lớn cho các tổ chức thu thập Big data. Tuy nhiên, tại Hội thảo tương lai của Thống kê học, có sự lo ngại về việc sinh viên học thống kê có thể bị hạn chế. Chẳng hạn như các tuyển dụng tại thung lũng Silicon mà sinh viên đang nộp đơn xin việc không có chỉ định dành cho "thống kê" mà dành cho "các nhà khoa học dữ liệu". Các sinh viên tốt nghiệp thì muốn những công việc này. Thông tin được truyền tai nhau rằng, có những sinh viên thấy làm một công việc tại Google hấp dẫn như đối với việc học tập tại một trường đại học Top 10. Tuy nhiên, họ không phải luôn luôn có được chúng. Vì người sử dụng lao động muốn tuyển sinh viên có khả năng viết phần mềm cho các công trình và những người có thể giải quyết vấn đề mà họ không phải tìm hiểu trong sách. Một nhận thức là tiến sĩ mới được đào tạo thống kê thường không có những khả năng này.

Có nhiều ý kiến khác nhau về việc các nhà thống kê phải có hành động đáp lại với các nhu cầu mới cho các nhà khoa học dữ liệu. Thống kê ở một số ngành bắt đầu cung cấp các bậc học về khoa học dữ liệu. Một dự án hợp tác của các cơ quan khoa học máy tính và ngành Thống kê, Đại học California tại Berkeley đã bắt đầu chương trình đào tạo thạc sĩ về khoa học dữ liệu, hay là Đại học Warwick, Anh cung cấp

bằng đại học đầu tiên về khoa học dữ liệu, một lần nữa cho thấy sự cộng tác giữa thống kê và khoa học máy tính.

Trong bộ phận thống kê truyền thống, đã có những việc làm để chuẩn bị thêm cho sinh viên về công việc "khoa học dữ liệu"? Hơn nữa đào tạo khoa học máy tính có vẻ là một ý tưởng tốt, và cần phải tiến xa hơn việc đơn thuần là học ngôn ngữ máy tính. Các sinh viên cần phải tìm hiểu làm thế nào để sản xuất phần mềm mạnh hơn và nhanh hơn. "Điều đó cần phải hoàn thành trước khi chúng ta chết", câu châm biếm của Rafael Irizarry.

Nhưng các sinh viên dành thời gian nhiều hơn để học khoa học máy tính, một ít thời gian còn lại sẽ dành cho đào tạo thống kê truyền thống. Các cuộc thảo luận về những gì là thành phần "cốt lõi" có thể được thảo luận, tranh đấu, hoặc nếu đó là nền tảng "cốt lõi" cung cấp cho tất cả sinh viên, thậm chí còn chưa thống nhất giống nhau. Một vài ý kiến tạm gọi là ít nhấn mạnh trên cơ sở toán học trừu tượng của các đối tượng. Tuy nhiên, một số người tham dự cảm thấy rằng sự hợp nhất của chủ đề này là thế mạnh của mình, và làm họ nhớ đến những cuộc hội thảo thống kê và họ hiểu nội dung các bài thuyết trình. Thậm chí họ thừa nhận rằng mọi thứ đang thay đổi, xu hướng hiện nay là hướng tới một lĩnh vực đa dạng và phân nhánh hơn. Xu hướng này nên được chấp nhận hay phản đối? Áp lực của Big data sẽ là "rơm mà phá vỡ sự trở lại của lạc đà" (*là câu thành Tiếng Ả rập ngữ có hàm ý chỉ với thức ăn đơn giản như rơm mà con lạc đà đã được nạp năng lượng vượt quá khả năng của mình để di chuyển*), hoặc là các điều kiện cần thiết thúc đẩy sự thay đổi trong quá trình lâu dài? Các câu hỏi như thế này là không có gì thậm chí còn tiếp cận được sự đồng thuận của các bên tham gia.

Một số người tham gia cuộc họp cho rằng cần phải thay đổi hệ thống khen thưởng ở cấp bậc sau tiến sĩ. Hiện nay, việc này đang xúc tiến và công bố quyền sở hữu trên các tạp chí khoa học truyền thống. Nói chung, các tạp chí lý thuyết (truyền thống) được coi là có uy tín hơn các tạp chí ứng dụng, và các tạp chí thống kê mang nhiều thông tin quan trọng hơn so với các tạp chí các ngành khác, chẳng hạn như thông tin về dữ liệu gen hoặc biến đổi khí hậu. Hầu như không có bộ phận thống kê nào sẽ cung cấp nhiều thông tin quan trọng cho một phần mềm vì nó sẽ dành cho một bài nghiên cứu. Nhưng nếu để họ chuẩn bị công việc về Big data cho sinh viên, thì các bộ phận thống kê đó sẽ phải thực hành những gì họ giảng. Họ sẽ phải trao phần thưởng cho giảng viên có nghiên cứu lý luận, áp dụng thực tế, và các chương trình về thống kê học. Điều này đòi hỏi có một sự thay đổi văn hóa quan trọng trong thống kê.

Một số lo ngại về sự trì trệ và không chắc chắn sẽ dẫn đến sự thiếu xót của một số hoạt động, đó sẽ là khóa học tồi tệ nhất của chúng ta. Marie Davidian, cựu Chủ tịch Hiệp hội Thống kê Mỹ đã viết "Tôi tin rằng các ngành khoa học thống kê là ở ngã ba đường, và những gì chúng ta đang làm... sẽ có ảnh hưởng sâu sắc đối với tình trạng tương lai, quy luật của chúng ta". Sự xuất hiện của Big data, khoa học dữ liệu, phân tích, và như thế đòi hỏi chúng ta có trách nhiệm như là một ngành không thể ngồi yên... nhưng phải chủ động về việc thiết lập cả hai vai trò của chúng ta trong cuộc cách mạng dữ liệu và đưa ra thống nhất bộ các nguyên tắc mà tất cả các đơn vị tham gia nghiên cứu khoa học, đào tạo và hợp tác... Có rất nhiều người trong nghề đang tỏ ra giận giữ về khoa học dữ liệu và tương tự họ cho rằng chúng ta nên tích cực cố gắng làm giảm đi uy tín những điều này và làm những việc hướng tới như đổi tên bất cứ điều gì cần phải làm với dữ liệu là số liệu thống kê. Tại thời điểm này, những tên và khái niệm mới đang dừng lại đây, và điều đó là phản tác dụng khi chúng ta đầu tư quá nhiều nguồn lực để cố gắng thay đổi điều này. Chúng ta nên thúc đẩy thống kê phát triển như là một quy luật và làm rõ vai trò quan trọng của thống kê trong bất kỳ hoạt động có liên quan đến dữ liệu.

Terry Speed, người chiến thắng năm 2013 ở giải thưởng của Thủ tướng Chính phủ dành cho ngành khoa học tại Úc, cung cấp một điểm mới về Big data sau cuộc thi: "Có phải chúng ta đang làm một công việc xấu đến như vậy mà chúng ta cần phải đổi tên mình thành các nhà khoa học dữ liệu để nắm bắt trí tưởng tượng của các sinh viên, cộng tác viên, hoặc khách hàng trong tương lai? Có phải chúng ta quá thiếu tự tin... mà chúng ta run sợ như trong thời điểm xuất hiện một người có khả năng giành mất ngôi vị của chúng ta khi đang ở trên sân khấu? Hoặc, đã thực sự có một sự thay đổi cơ bản xung quanh chúng ta, vì vậy mà cách thức truyền thống để chúng ta thích ứng và phát triển không còn phù hợp?... Tôi nghĩ rằng chúng ta đã có cả hai đó là một truyền thống tuyệt đẹp và một tương lai rộng mở, chúng dài hơn so với khả năng tập trung của các cơ quan tài trợ, các khoa đại học, và các tổ chức... Ngay hiện tại chúng ta có thể bỏ lỡ hàng triệu thứ được coi là khoa học dữ liệu, nhưng điều đó không phải lý do để chúng ta ngừng cố gắng và làm tốt nhất những gì chúng ta có thể làm, về điều gì đó là xa hơn, rộng hơn và có chiều sâu hơn so với khoa học dữ liệu. Như với toán học nói chung, chúng ta đang làm công việc đó cho lâu dài. Đừng để mất tinh thần của chúng ta".