DỮ LIỆU KHÔNG PHẢI LÀ "KẺ THÙ" CỦA NGÀNH NHÂN VĂN

LINH CHI (dịch)

"Dữ liệu lớn (big data) đang tới để thống trị những cuốn sách của bạn". Đã gần một thập kỷ kể từ khi một nhà văn lên tiếng chống lại sự tham gia của khoa học dữ liệu trong các ngành khoa học xã hội và nhân văn trong một bài viết trên tờ Los Angeles Review of Books. Ông lo lắng về dịch vụ Google Books và sự nhen nhóm của các phương pháp tính toán trong lĩnh vực nghiên cứu văn học. Cho đến thời điểm hiện tại, mối đe dọa này được xem là đã bị thổi phồng quá mức, khi mà ngay sau đó Google Books đã chuyển giao lại dự án số hóa của họ cho Hathi Trust, một tổ chức nhân văn phi lợi nhuận cũng sử dụng các phương pháp tính toán và biến dự án này thành một thị trường ngách nhỏ.

Tuy vậy, sự lo lắng lại dấy lên trong những năm gần đây, khi các nhà quản lý giáo dục và nhà tài trợ đã bắt đầu đầu tư rất nhiều tiền vào khoa học dữ liệu (data science), một lĩnh vực nghiên cứu liên ngành.

Khoa học dữ liệu xuất phát chủ yếu từ khoa học máy tính, toán học và thống kê, nhưng tham vọng của nó là bao quát hết các trường đại học. UVA đã khởi động một sáng kiến nhằm tuyển dụng các nhân sự có chuyên môn về khoa học dữ liệu từ các ngành khoa học tự nhiên, khoa học xã hội và nhân văn. Các nhà khoa học dữ liệu giờ đây đã bắt đầu lấn sân sang nghiên cứu văn học và văn hóa.

Trong khi đó, các ngành khoa học xã hội và nhân văn thì đang chịu cảnh bị hạn chế tài trợ, đóng băng hoạt động tuyển dụng, các chương trình đào tạo tiến sĩ bị ngừng vô thời hạn. Dữ liệu lớn thực sự có xu hướng như đang tới để "thống trị những cuốn sách của bạn". Jill Lepore - một nhà sử học công tác tại ĐH Harvard, một trí thức nổi bật - là một đại diện cho tiếng nói chống lại dữ liệu trong ngành nhân văn.

ĐIỂM TỐI CỦA KHOA HỌC DỮ LIỆU

Trong "If Then", tựa sách mới nhất của Lepore, cô đã kể lại câu chuyện về tập đoàn Simulmatics - nơi tiên phong trong việc sử dụng mô hình dự đoán bằng máy tính để dự đoán xem các cử tri nhỏ (microtype) sẽ phản ứng thế nào với các chiến dịch chính trị. Một nhân vật quan trọng trong Simulmatics là Ithiel de Sola Pool, ông được xem là người cha tinh thần của Phòng thí nghiệm truyền thông MIT, đồng thời đại diện cho những gì tồi tệ nhất của khoa học dữ liệu. Pool và Similmatics đã đưa những giả định theo chủ nghĩa nam tính vào trong những mô hình tính toán của họ. Lepore viết nguyên văn như sau:

"Khi nói về 'hành vi con người', ý của họ là hành vi của đàn ông; và khi nói về 'trí tuệ nhân tạo', ý của họ là trí tuệ của chính họ - một huyễn tưởng về trí tuệ của chính họ mà họ đang muốn cấy vào một bộ máy".

Những giả định này đã và vẫn đang tiếp tục

gây ra những thiệt hại không đếm xuể. Lepore tranh luận rằng Simulmatics đã đặt loài người chúng ta vào con đường dẫn đến Amazon, Facebook, Google, những thứ mà giờ đây đang khiến chúng ta bị "đày đọa và mắc kẹt". "Những mô phỏng tự động hành vi của con người trở thành điều kiện/tình trạng của con người". Và trong quá trình đó, các tư duy nhân văn đã bị hạ thấp một cách trầm trọng.

Trong một cuộc đối thoại với Fran Berman trong khuôn khổ Sáng kiến Khoa học Dữ liệu Harvard vào tháng 9 năm ngoái, Lepore đã mở rộng thêm các lập luận của mình. Sự thật (facts) trở nên có quyền lực vào năm 1215 khi bồi thẩm đoàn xuất hiện, thay thế hình thức xét xử bằng cách thử tội (trial by ordeal - một hình thức xét xử cổ xưa, trong đó quyết định về việc một bị cáo là vô tội hay có tội danh được xác định bằng cách buộc họ phải trải qua một trải nghiệm đau đớn, nguy hiểm). Sự kiện trên cùng với cuộc Cải cách Tin lành (Protestant Reformation) về sau đã góp phần phế truất vị thế của Chúa như là người duy nhất nắm rõ mọi sự vô tội và có tội trên thế giới. Bí ẩn đã phải nhường đường cho tri thức.

Các con số lại soán ngôi của sự thật, trở thành hệ hình bằng chứng thống trị (evidentiary paradigm) khi chủ nghĩa tư bản, dân chủ và thống kê trỗi dậy vào cuối thế kỷ 18; gắn liền với nhu cầu của Chính phủ mong muốn thay thế các thảo luận và phân định bằng sự đo lường. Dữ liệu lại bắt đầu chiếm chỗ của các con số khi các máy lập bảng (tabulating machines) là trọng tâm của các cuộc tính toán dân số vào năm 1890, và trở nên thống trị hoàn toàn khi UNIVAC (dòng máy tính kỹ thuật số điện tử đa năng đầu tiên được thương mại hoá) cùng các kỹ thuật tính toán khác xuất hiện vào những năm 1950.

"Sự trỗi dậy của một kỷ nguyên dữ liệu theo một cách nào đó đang khiến chúng ta quay trở lại thời kỳ của sự bí ẩn. Máy móc là Chúa, và các nhà khoa học máy tính là các mục sư. Chúng ta, những người còn lại, chỉ biết hướng về họ và trông cậy rằng họ đang làm đúng". Dữ liệu đã giết chết sự thật và chấm dứt tri thức. Lepore thừa nhận rằng tình trạng này cũng được thúc đẩy bởi các ưu tiên tài chính của các trường đại học.

Lepore đã đặt ra một câu hỏi lớn về vị trí của dữ liệu trong quá trình sản xuất tri thức cũng như những bằng chứng quan trọng về việc khoa học được điều khiển bởi dữ liệu đang làm tổn hại tới các ngành khoa học nhân văn như thế nào. Nếu như thực sự muốn bảo vệ ngành khoa học nhân văn, chúng ta cần phải trả lời câu hỏi trên một cách xác đáng nhất có thể. Và để làm được điều đó, có lẽ ngay từ đầu cần loại bỏ niềm tin nhị phân rằng dữ liệu chỉ thuộc về khoa học, không có liên đới với các ngành nhân văn.

KHOA HỌC NHÂN VĂN BẮT TAY VỚI KHOA HỌC DỮ LIỆU

Thực tế, trong lịch sử của các ngành khoa học nhân văn, việc kết hợp với khoa học dữ

liệu không hề mới. Thậm chí, các học giả nhân văn vốn đã thường xuyên thu thập và dựa vào dữ liệu từ thời kỳ sơ khởi của các trường đại học nghiên cứu hiện đại và có thể là cả trước đó.

Trong cuốn sách "The Teaching Archive", hai tác giả Rachel Sagner Buurma và Laura Heffernan đã dạy cho chúng ta về Caroline Spurgeon, một nhà nghiên cứu về Shakespeare đầu thế kỷ 20, có niềm tin sâu đậm vào các giá trị của nhân văn nhưng cũng ưa thích các phương pháp định lượng, với các công trình phân tích các tác phẩm của Shakespeare bằng phương pháp đếm và phân loại các hình thái tu từ (figures of speech). Họ giới thiệu cho chúng ta về Edith Rickert, người đã tận dụng hết mức "phương pháp phân tích mã" (code analysis) mà bà đã học được khi làm việc trong cơ quan tình báo quân đội thời kỳ Chiến tranh thế giới thứ I. Vào năm 1920, Rickert, cùng một người khác cũng tiến hành nghiên cứu song song thời gian đó nhưng độc lập là I. A. Richards, đã "yêu cầu các sinh viên của mình không chỉ cẩn thận phân tích các văn bản văn học, mà cả tham gia hợp tác vào quá trình phân tích và tổng hợp các đơn vị dữ liệu về người đọc và văn bản". Chính thông qua quá trình thu thập dữ liệu này mà sinh viên và giảng viên "nhận thấy họ đang thảo luận một số lượng đáng kinh ngạc các câu hỏi phức tạp về thể thơ và ngữ cảnh lịch sử."

Buurma và Heffernan còn kể cho chúng ta về Josephine Miles, một giảng viên Văn học Anh ở Đại học Berkeley trong khoảng thời gian từ năm 1939 đến 1978. Miles luôn dạy sinh viên phải "có quan điểm riêng về sự thật (facts), thay vì chỉ tường thuật lại chúng", vì bà tin rằng điều đó là rất quan trọng trong "một xã hội hiện đại và một môi trường đại học nghiên cứu hiện đại coi cá nhân là [một thứ] được xác định bởi dữ liệu". Nhưng bà cũng không phản đối việc dùng dữ liệu trong nghiên cứu văn học. Từ những năm 1950, Miles đã tiến hành những công trình nghiên cứu đột phá sử dụng phương pháp tính toán bằng máy tính. Bà đã xây dựng được những bộ dữ liệu khổng lồ về lịch sử của thơ, cho phép bà lần theo các dấu vết của sự xuất hiện và biến mất của các hình thái tu từ trong suốt chiều dài nhiều thập kỷ và thế kỷ, để tìm ra các điểm đứt gãy và liên tục còn chưa được biết tới.

Truyền thống này vẫn kéo dài tới hiện nay. Lauren F. Klein, phó giáo sư của ngành Văn học Anh và lý thuyết & phương pháp định tính tại Đại học Emory, đã sử dụng các phương pháp tính toán để phục hồi lại những thông tin bị bỏ qua trong các văn bản lưu trữ, đặc biệt là những thông tin về người Mỹ gốc Phi. Cô đã dùng các lá thư của Thomas Jefferson làm dữ liệu để soi xét lại cuộc đời và quá trình lao động của đầu bếp người da đen James Hemings - người bị Jefferson bắt làm nô lệ. Cô cũng phân tích các tờ báo ở thế kỷ 19 để tìm và tiết lộ những hoạt động ẩn giấu

của người phụ nữ da đen đầu tiên thực hiện hoạt động xuất bản ở Bắc Mỹ - Mary Ann Shadd - bởi những sản phẩm của Shadd không được công nhận trong thời kỳ đó.

Các ngành nhân văn vẫn luôn coi trọng và sử dụng dữ liệu, thậm chí là thu được nhiều lợi ích từ các phương pháp tính toán và các mô hình dự đoán. Điều đó dẫn chúng ta tới đâu? Lúc này, "chẩn đoán" của Lepore lại cực kỳ chính xác. Các công ty, tập đoàn công nghệ lớn sẽ thu phục chúng ta cùng sự chú ý của chúng ta vào dữ liệu. Và khẳng định của khoa học dữ liệu rằng "sẽ chiếm lĩnh mọi cách thức hiểu biết khác" đang đe dọa "sự gần như là bỏ rơi tri thức nhân văn".

Nhưng dữ liệu không phải là kẻ thù của ngành nhân văn. Sự hạ thấp giá trị của tri thức xuống thành một thứ tập trung vào hiệu quả kinh tế mới là kẻ thù của ngành nhân văn. Khi mà sự thịnh hành của khoa học dữ liệu trong giới các nhà tài trợ từ thiện các nhà quản lý trường học thể hiện một đặc tính tân tự do mà ở đó nhìn nhận giáo dục dưới góc độ lợi tức đầu tư thì ta nên chống lại nó. Nhiều nhà nghiên cứu nhân văn kỹ thuật số cũng đặt sự kháng cự vào làm trọng tâm trong các công trình của họ. Chẳng hạn, Klein tiên phong cho một "chủ nghĩa nữ quyền dữ liệu", "một cách tư duy về dữ liệu - cả về những công dụng và giới hạn của nó - mà được xác nhận lại bằng những trải nghiệm trực tiếp bằng một cam kết hành động và bằng một tư tưởng giao thoa với chủ nghĩa nữ quyền". Chủ nghĩa nữ quyền dữ liệu phê bình khoa học dữ liệu khi nó củng cố "những bất bình đẳng sẵn có" và sử dụng khoa học dữ liệu "để thách thức và thay đổi sự phân chia quyền lực."

Hãy thừa nhận là dữ liệu đã cung cấp những hiểu biết về văn học và văn hóa mà chúng ta không thể nhận thấy trước đây, cho dù là về lịch sử của thơ, về ẩm thực của James Hemings hay các công trình biên tập của Mary Ann Shadd. Tháng 4 năm nay, tạp chí Cultural Analytics và Post45 đã xuất bản một số đặc biệt những công trình theo định hướng khoa học dữ liệu. Các bài viết trong đó nói về Goodreads đã tái thiết lập lại các thể loại sách như thế nào; các nội dung trên Internet đã góp phần định hình những chương trình truyền hình như The Good Place hay những tiểu thuyết như Lincoln in the Bardo như thế nào; hay làm thế nào mà một hội thảo của các nhà văn tại Bang Iowa đã làm đảo ngược những logic trước đó về chủ nghĩa khu vực Mỹ, tới mức giờ đây nhiều nhà văn trong đó có John Irving và Marilynne Robinson đã chuyển từ các khu đô thị về các miền nông thôn như Iowa và chọn đó làm không gian sáng tác tiểu thuyết - tất cả đều được khám phá nhờ khoa học dữ liệu.

Chính văn học và văn hóa cũng đang ngày càng được định hình bởi dữ liệu. Hãy thử nghĩ tới sự hiện diện của các tác giả trên các không gian văn học kỹ thuật số, hay Instapoets - những nhà thơ nổi tiếng thông qua việc tiếp cận được một lượng lớn độc giả trên Instagram. Hay sự sinh sôi một cách ngoan cố các thể loại sách do sự phân phối của Kindle Direct Publishing. Vào thời điểm năm 2021, việc chối bỏ sự tồn tại của dữ liệu cũng chính là đang liều mình tách biệt bản thân khỏi nền tảng bản thể học và xã hội học của những công trình nghiên cứu. Các nhà nghiên cứu nhân văn cần sử dụng dữ liệu như một trong nhiều đối tượng của nghiên cứu, đồng thời đấu tranh chống lại những bên sử dụng dữ liệu để chống lại chúng ta.

Nguồn

Dan Sinykin (April 29, 2021). Data Is Not the Enemy of the Humanities. The Chronicle of Higher Education.