Câu chuyện về khoa học dữ liệu: Chỉ vài kẻ thắng cuộc?

10/07/2017 16:50 - **Hồ Tú Bảo**

Hồ Tú Bảo – Giáo sư CNTT tại viện JAIST – Nhật Bản

Thành viên viện Toán cao cấp – Ngô Bảo Châu

Bài đăng trên tạp chí Tia sáng

Vừa qua tại hội nghị châu Á - Thái Bình Dương về Khai phá Dữ liệu (PAKDD) ở Hàn Quốc, giáo sư Sang Kyun Cha (Viện trưởng Viện Dữ liệu lớn Hàn Quốc) - người có uy tín trong cả hai giới hàn lâm và công nghiệp nước này, đã nói về cuộc cách mạng công nghiệp lần thứ tư (CMCN4), và một nhận định của ông đã ám ảnh tôi suốt hội nghị: "Mối đe doạ là chỉ một số ít quốc gia thắng cuộc sẽ lấy tất"("Threat: A small number of winners take all").

Việt Nam cần xây dựng được nguồn dữ liệu của các ngành nghề. Nguồn: taichinhdientu.vn

Nếu nhận định này là đúng, thì một số ít quốc gia thắng cuộc trong CMCN4 sẽ là ai? Vì sao họ thắng cuộc? Việt Nam có nằm trong số đông các nước sẽ thua cuộc không? Nếu có thì làm sao vượt ra để thực hiện mục tiêu phát triển đất nước ?

Cốt lõi của CMCN4 là sản xuất thông minh dựa trên các đột phá của công nghệ số. Có thể hiểu công nghệ số gồm hai nội dung chính: số hóa và dùng dữ liệu số hóa. Tiến bộ của khoa học đã cho phép con người dần *số hóa* được hầu hết mọi thực thể trên đời (hệ gien người, cây lúa, chiếc ô-tô, khách sạn, doanh nghiệp, cơ quan công quyền…), và hầu hết mọi thực thể trong thế giới thực của chúng ta có thể được kết nối với nhau qua các phiên bản số của chúng trong không gian internet (internet vạn vật). Việc kết nối này thực chất là kết nối dữ liệu số hóa của các thực thể và do đó tạo ra một không gian số hoá rất lớn và rất phức tạp, gọi là *dữ liệu lớn* (big data), hiện đang vượt quá khả năng xử lý của con người. Những điều trên đang dẫn đến sự *thay đổi phương thức sản xuất* của con người: Hoạt động sản xuất trong thế giới thực được điều hành và quyết định một cách thông minh từ không gian số các kết nối kể trên.

Có thể nói, về bản chất, quốc gia nào thắng cuộc trong CMCN4 là quốc gia làm chủ được các nguồn dữ liệu lớn và phức tạp này, đưa chúng vào mọi lĩnh vực của sản xuất và cuộc sống, làm cho sản xuất và cuộc sống thông minhvà hiệu quả hơn qua các phương pháp của trí tuệ nhân tạo. Mối đe dọa giáo sư Sang Kyun Cha nói đến chính là chỉ một số ít quốc gia nắm được, phát triển và dùng được công nghệ số hiệu quả, và do đó sẽ thắng trong CMCN4.

Ý thức được điều này, trong vòng vài năm qua các nước phát triển đều xây dựng chương trình chiến lược quốc gia của mình cho thời gian tới. Nước Mỹ có "Chiến lược quốc gia về sản xuất tiên tiến" cho ba thập kỷ. Nước Đức có

"Công nghiệp 4.0". Nước Pháp có "Bộ mặt mới của công nghiệp Pháp". Hàn Quốc có "Chương trình tăng trưởng của Hàn Quốc trong tương lai". Trung Quốc có "Sản xuất tại Trung Quốc năm 2025". Nhật Bản có "Xã hội thông minh" (Smart society). Singapore có "Quốc gia thông minh" (Smart nation)… Cốt lõi của các chương trình đó chính là câu chuyện của số hóa, kết nối và phân tích dữ liệu lớn. Và xuyên suốt ba khía cạnh công nghệ này chính là *khoa học dữ liệu* (KHDL).

Quốc gia nào thắng cuộc trong CMCN4 là quốc gia làm chủ được các nguồn dữ liệu lớn và phức tạp này, đưa chúng vào mọi lĩnh vực của sản xuất và cuộc sống, làm cho sản xuất và cuộc sống thông minh và hiệu quả hơn qua các phương pháp của trí tuệ nhân tạo.

Đại thể KHDL là khoa học về việc quản trị và phân tích dữ liệu để tìm ra các hiểu biết, các tri thức hành động, các quyết định dẫn dắt hành động. KHDL gồm ba phần chính: Tạo ra và quản trị dữ liệu, phân tích dữ liệu, và chuyển kết quả phân tích thành giá trị của hành động. Nôm na bước thứ nhất là về số hóa và bước thứ hai là về dùng dữ liệu. Việc phân tích và dùng dữ liệu lại dựa vào ba nguồn tri thức: toán học (thống kê toán học), công nghệ thông tin (học máy) và tri thức của lĩnh vực ứng dụng cụ thể.

Nếu phân tích dữ liệu về nhu cầu thị trường ta có thể quyết định cần nuôi bao nhiêu lợn mỗi nơi mỗi lúc. Nếu có và phân tích được dữ liệu mô phỏng các phương án xả lũ vào mùa mưa ta có thể chọn được cách xả lũ ít thiệt hại nhất. Nếu có và phân tích được các bệnh án điện tử của người bệnh ta có thể tìm ra được phác đồ thích hợp hơn cả cho người bệnh. Amazon đã phân tích các lần mua hàng trước của bạn để dự đoán những món đồ bạn có thể sẽ thích mua và gửi quảng cáo tới, v.v. Khi nghe nói về các thành tựu đột phá gần đây của Trí tuệ nhân tạo người nghe có thể cũng chưa biết rằng phần lớn chúng đều dựa vào các phương pháp và đột phá của KHDL.

Để thực hiện chương trình của mình, một trong các việc Hàn Quốc đã làm là lập ra Viện Dữ liệu Lớn (Big Data Institute) trong Đại học Quốc gia Seoul (SNU) vào tháng 4/2014. Tuy nằm trong SNU, Viện này mang sứ mạng quốc gia, liên kết khoảng 220 giáo sư người Hàn Quốc hoạt động cho lĩnh vực liên ngành này, nhằm dẫn dắt sự dịch chuyển quốc gia về giáo dục và nghiên cứu với KHDL. Viện được đầu tư rất lớn, như phòng thí nghiệm về KHDL đô thị của Viện đã nhận được kinh phí 9 triệu USD của thành phố Seoul vào tháng 4 năm nay cho giai đoạn 3 năm.

Trung Quốc đã sớm đẩy mạnh KHDL trên toàn quốc. Tháng 8/2012 quốc gia này đã khởi động chương trình hoa tiêu về dữ liệu lớn với kinh phí 1,3 tỷ nhân dân tệ. Vào tháng 10/2015, Ban Chấp hành Trung ương Đảng Cộng sản Trung Quốc đã ra "Chiến lược quốc gia về dữ liệu lớn". Các viện về khoa học dữ liệu, về dữ liệu lớn được thành lập ở Đại học Bắc Kinh, Đại học Thanh Hoa, Đại học Thâm Quyến... và nhiều đại học trên cả nước cũng như ở nhiều tỉnh thành. Cơ sở của những quyết sách như vậy là một sinh hoạt thường kỳ về khoa học và công nghệ với tên gọi "học tập theo nhóm" (group study) của Bộ Chính trị của Đảng Cộng sản Trung Quốc. Trong các buổi này, họ mời các nhà khoa học uy tín giới thiệu về các tiến bộ quan trọng của khoa học và công nghệ.

Có thể nói Việt Nam hầu như không thể nằm trong số ít thắng cuộc, theo nghĩa thắng ở việc làm ra các sản phẩm của công nghệ cao cho thiên hạ. Tuy nhiên, với các mục tiêu phát triển của mình, phải chăng ta có những cơ hội để thắng chính mình, để làm được và dùng được KHDL một cách hiệu quả, rộng hơn là công nghệ số, cho những mục tiêu phát triển của đất nước?

Mục tiêu của Singapore là trở thành một quốc gia thông minh dựa trên làm chủ nguồn dữ liệu lớn. Gần đây, Singapore lập ra một tổ hợp về khoa học dữ

liệu (data science consortium) gồm Đại học quốc gia Singapore (NUS), Đại học kỹ thuật Nanyang (NTU), Đại học Quản lý Singapore (SMU) và Trung tâm nghiên cứu A*STAR để thúc đẩy hợp tác giữa công nghiệp và trường viện để phát triển KHDL. Không có thế mạnh chế tạo máy móc như nhiều cường quốc khác, Singapore chọn con đường riêng của mình trong CMCN4.

Con đường phát triển và cách đi của mỗi quốc gia tất nhiên phụ thuộc vào nhiều yếu tố. Đó là thể chế chính trị với các chính sách phát triển, là tình trạng kinh tế, xã hội, văn hoá, giáo dục và đào tạo, khoa học và công nghệ… Tuy nhiên khi CMCN4 bùng nổ, mọi quốc gia muốn phát triển đều phải dựa nhiều hơn vào khoa học và công nghệ, vào các nguồn dữ liệu thay cho các nguồn tài nguyên thiên nhiên. Đương nhiên các quốc gia đang sản xuất với nhiều công nghệ cao cũng như có nền khoa học phát triển sẽ có nhiều cơ hội thắng cuộc. Câu hỏi là có cơ hội nào cho các nước đang phát triển vươn lên trong CMCN4 không? Đâu sẽ là con đường và cách chúng ta đi trong CMCN4?

Nhìn vào bức tranh toàn cầu và các yếu tố kể trên, có thể nói Việt Nam hầu như không thể nằm trong số ít thắng cuộc, theo nghĩa thắng ở việc làm ra các sản phẩm của công nghệ cao cho thiên hạ. Tuy nhiên, với các mục tiêu phát triển của mình, phải chăng ta có những cơ hội để thắng chính mình, để làm được và dùng được KHDL một cách hiệu quả, rộng hơn là công nghệ số, cho những mục tiêu phát triển của đất nước?

Giáo dục toán học của ta có truyền thống. Ta có lực lượng làm về công nghệ thông tin khá đông đảo và có kỹ năng tốt. Quan trọng hơn cả, ta có những thế hệ người trẻ tuổi thông minh, khát vọng vươn lên cho đời mình và cho đất nước. Trong khoá học ngắn hạn về KHDL do Viện nghiên cứu cao cấp về Toán (VIASM) tổ chức giữa tháng 5 vừa qua ở Hà Nội và thành phố Hồ Chí Minh, đã có hơn một nghìn người đăng ký và tham gia, hầu hết còn trẻ. Có

thể cảm nhận được khát khao hiểu biết và mong muốn vươn lên trên từng khuôn mặt và trong từng câu hỏi của người học.

Để phát huy được sức mạnh của KHDL ta phải có chính sách và nỗ lực xây dựng nguồn tài nguyên dữ liệu, về công cuộc số hóa. Ngoài dữ liệu do máy móc đo đạc đưa lại (thí dụ từ các thiết bị đo trong y học, trong giao thông vận tải, khí tượng thời tiết…) hay nguồn dữ liệu từ các mạng xã hội, ta phải xây dựng được nguồn dữ liệu của các ngành nghề (kinh tế, nông nghiệp, du lịch, tài chính, thương mại, giáo dục…), nguồn dữ liệu của xã hội dân sự từ các hệ thống chính phủ điện tử… và các nguồn dữ liệu về thế giới quanh ta.

Theo nhận định chủ quan của tôi, nếu như về sản xuất ô-tô, đồ điện tử… ta có thể sau Hàn Quốc nhiều chục năm, nhưng nếu kết hợp tốt các nhà khoa học trong và ngoài nước ta cũng sẽ không sau Hàn Quốc bao nhiêu về KHDL

Kỹ thuật phân tích dữ liệu phát triển rất nhanh trong những năm vừa qua. Trong quãng thời gian ấy, một số nhà nghiên cứu người Việt làm việc ở đại học và các công ty lớn trên thế giới đã và đang nắm bắt được cũng như tham gia vào giải quyết những bài toán thời sự nhất, khó khăn nhất của KHDL. Phần lớn trong số họ đều sẵn sàng và mong muốn được học hỏi cũng như chia sẻ kinh nghiệm và hiểu biết của mình với đồng nghiệp trong nước.

Theo nhận định chủ quan của tôi, nếu như về sản xuất ô-tô, đồ điện tử… ta có thể sau Hàn Quốc nhiều chục năm, nhưng nếu kết hợp tốt các nhà khoa học trong và ngoài nước, ta cũng sẽ không sau Hàn Quốc bao nhiêu về KHDL.

KHDL mới là một yếu tố cần của sự phát triển trong CMCN4, nhưng là yếu tố cần ta có thể đạt được. Để KHDL có thể là một mũi đột phá trong sự phát triển, những yếu tố cần khác là sự mạnh mẽ của các quyết tâm chính trị, sự sáng suốt của các chính sách, sự cách tân của các doanh nghiệp…

Với nhu cầu phát triển của đất nước khi một cuộc cách mạng khoa học và công nghệ mới đang bắt đầu, với khao khát vươn lên của những người trẻ tuổi, trách nhiệm của những người giữ trọng trách, nỗ lực kết nối lực lượng khoa học và công nghệ trong và ngoài nước… liệu chúng ta có thể thay đổi được cách nghĩ, cách làm, tìm ra được đường đi để “thắng chính mình” trong các mục tiêu phát triển ? Và để không là người thua cuộc?