Tổng luận số 2 /2015

DỮ LIỆU LỚN VÀ XU HƯỚNG ĐỔI MỚI SÁNG TẠO DỰA TRÊN DỮ LIỆU

CỤC THÔNG TIN KHOA HỌC VÀ CÔNG NGHỆ QUỐC GIA
Địa chỉ: 24, Lý Thường Kiệt, Hoàn Kiếm, Hà Nội. Tel: (04)38262718, Fax: (04)39349127
Ban biên tập: TS. **Lê Xuân Định** (*Trưởng ban*), KS. Nguyễn Mạnh Quân,
ThS. Đặng Bảo Hà, ThS. Phùng Anh Tiến.

Mục lục

Lời giới thiệu

Thế giới đang chứng kiến một cuộc cách mạng công nghiệp mới được thúc đẩy bởi các dữ liệu số, tính toán và tự động hóa. Sự giao thoa của một số xu hướng công nghệ và kinh tế xã hội, bao gồm cả việc sử dụng Internet ngày càng tăng và sự suy giảm ở chi phí thu thập, truyền tải, lưu trữ và phân tích dữ liệu, dẫn đến việc tạo ra những khối lượng dữ liệu khổng lồ - gọi chung là "dữ liệu lớn" (Big Data), đây chính là nguồn lực có thể khai thác để thúc đẩy hình thành các ngành công nghiệp mới, các quy trình và sản phẩm mới. Các hoạt động kinh tế và xã hội từ lâu đã dựa vào dữ liệu. Tuy nhiên giờ đây, khối lượng, tốc độ và chủng loại dữ liệu được sử dụng đang gia tăng mạnh mẽ trên phạm vi toàn bộ nền kinh tế, và quan trọng hơn là giá trị kinh tế và xã hội lớn hơn của chúng đang mở ra cơ hội về một sự thay đổi hướng tới mô hình kinh tế xã hội dựa trên dữ liệu. Trong mô hình này, dữ liệu là tài sản cốt lõi có thể tạo ra lợi thế cạnh tranh quan trọng, chi phối đổi mới sáng tạo, tăng trưởng và phát triển bền vững.

Đổi mới sáng tạo dựa vào dữ liệu có giá trị kinh tế to lớn, với doanh thu từ các sản phẩm và dịch vụ Dữ liệu lớn đã vượt quá 18 tỷ USD trong năm 2013, và theo Feff Kelly (2014) thì giá trị này có thể đạt 50 tỷ USD vào năm 2017. Để hiện thực hóa trọn vẹn tiềm năng của dữ liệu lớn, các quốc gia cần có một khuôn khổ chính sách nhất quán, phù hợp về thu thập, truyền tải, lưu trữ, cung cấp và sử dụng dữ liệu, đặc biệt là trong các lĩnh vực như bảo vệ quyền riêng tư, tiếp cận dữ liệu mở, kỹ năng và việc làm, cơ sở hạ tầng và đo lường, v.v... Đây cũng chính là những nội dung thông tin mà cuốn Tổng luận **"Dữ liệu lớn và xu hướng đổi mới sáng tạo dựa trên dữ liệu"** muốn cung cấp với các độc giả. Tài liệu được biên soạn dựa trên các báo cáo của OECD về vai trò tiềm năng của dữ liệu và phân tích dữ liệu trong việc tạo ưu thế cạnh tranh và hình thành vốn tri thức, thúc đẩy đổi mới sáng tạo và tăng trưởng bền vững. Phần đầu của Tài liệu cung cấp những khái niệm và định nghĩa đã được công nhận rộng rãi về Dữ liệu lớn, cũng như việc tạo ra và sử dụng dữ liệu trong các lĩnh vực ứng dụng của nền kinh tế. Tiếp theo tài liệu mô tả các cách thức khai thác dữ liệu như một nguồn lực thúc đẩy tăng trưởng kinh tế và phát triển bền vững, và trong phần cuối, tài liệu đề cập đến các vấn đề chính sách chủ yếu trong hoạch định chính sách công nhằm thúc đẩy đổi mới sáng tạo dựa vào dữ liệu.

Xin trân trọng giới thiệu.

CỤC THÔNG TIN KH&CN QUỐC GIA

Bảng các chữ viết tắt

API	Giao diện lập trình ứng dụng
BI	Trí tuệ doanh nghiệp
CAGR	Tỷ lệ tăng trưởng tổng hợp lũy kế hàng năm
DDI	Đổi mới sáng tạo dựa vào tăng trưởng
HDD	Ổ đĩa cứng
ICT	Công nghệ thông tin - truyền thông
IoT	Internet kết nối vạn vật
KBC	Vốn tri thức
M&A	Mua bán và sáp nhập
M2M	Giao tiếp máy tới máy
NC&PT	Nghiên cứu và phát triển
NoSQL	Cơ sở dữ liệu phân tán không quan hệ
OECD	Tổ chức hợp tác và phát triển kinh tế
PET	Công nghệ bảo vệ quyền riêng tư
PMNM	Ứng dụng phần mềm nguồn mở
PSI	Thông tin khu vực công
SHTT	Sở hữu trí tuệ
SMS	Tin nhắn văn bản
SSD	Ổ đĩa thể rắn

I. ĐỔI MỚI DỰA TRÊN DỮ LIỆU - NGUỒN LỰC TĂNG TRƯỞNG VÀ PHÁT TRIỂN KINH TẾ

1.1. Dữ liệu lớn và các khái niệm liên quan

Trong thời đại hiện nay, dữ liệu đang ngày càng thấm sâu vào cuộc sống của con người hơn bao giờ hết. Chúng ta mong muốn sử dụng dữ liệu để giải quyết các vấn đề, nâng cao phúc lợi và tạo ra thịnh vượng kinh tế. Việc thu thập, lưu trữ, và phân tích dữ liệu đang tuân theo quỹ đạo có xu hướng đi lên và dường như không có ranh giới, hoạt động này được thúc đẩy bằng những gia tăng về năng lực xử lý, chi phí giảm mạnh trong tính toán và lưu trữ, và số lượng ngày càng tăng các công nghệ cảm biến nhúng trong tất cả các loại thiết bị. Vào năm 2011, một số ước tính rằng khối lượng thông tin được tạo ra và sao chép lại sẽ vượt mức 1,8 zettabytes. Trong năm 2013, ước tính có 4 zettabytes dữ liệu được tạo ra trên toàn thế giới.

1 zettabyte (ZB) = 10^{21}bytes. Một byte tương đương với một ký tự trong văn bản. Có thể tưởng tượng rằng, nếu cứ mỗi giây, mỗi một người dân tại Hoa Kỳ chụp một bức ảnh số, cứ thế liên tục trong vòng một tháng. Tất cả số ảnh đó đem tập hợp lại với nhau sẽ bằng khoảng một zettabyte.

Mỗi ngày có hơn 500 triệu bức ảnh được tải lên và chia sẻ trên mạng xã hội, cùng với các đoạn video với độ dài tổng cộng đến 200 giờ được tải lên mỗi phút. Nhưng khối lượng thông tin mà mọi người tự tạo ra, các thông tin liên lạc gồm các cuộc gọi thoại, email và văn bản, các bức ảnh, video và âm nhạc được tải lên vẫn không là gì so với lượng thông tin số được tạo ra về chúng mỗi ngày.

Các xu hướng này vẫn đang tiếp diễn. Hiện nay chúng ta mới ở vào giai đoạn rất sơ khai của cái gọi là "Internet vạn vật" (IoT), khi tất cả các thiết bị, các phương tiện và các công nghệ "mang trên người" có thể giao tiếp được với nhau. Các tiến bộ công nghệ sẽ làm giảm chi phí của việc tạo ra, thu thập, quản lý và lưu trữ thông tin xuống chỉ còn bằng một phần sáu chi phí được tính vào năm 2005. Và kể từ năm 2005, đầu tư doanh nghiệp vào phần cứng, phần mềm, nhân lực và dịch vụ đã tăng 50% đạt 4 nghìn tỷ USD.

"Internet vạn vật" là thuật ngữ dùng để mô tả khả năng các thiết bị có thể giao tiếp được với nhau sử dụng các cảm biến nhúng, liên kết với nhau thông qua các mạng kết nối có dây và không dây. Các thiết bị này có thể bao gồm cả nhiệt kế, xe hơi và thậm chí cả viên thuốc mà bạn nuốt vào để các bác sĩ có thể theo dõi sức khỏe bộ máy tiêu hóa của bạn. Các thiết bị kết nối này sử dụng Internet để truyền, diễn giải và phân tích dữ liệu.

1.1.1. Dữ liệu và các yếu tố thúc đẩy tạo và sử dụng dữ liệu

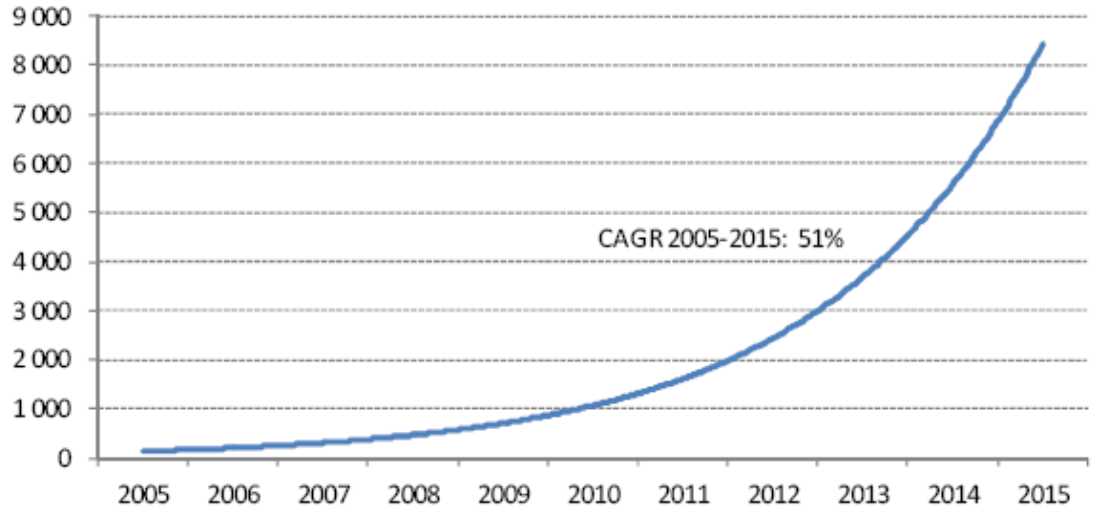
Việc số hóa gần như mọi phương tiện truyền thông và sự chuyển hướng ngày càng tăng

của các hoạt động kinh tế và xã hội sang sử dụng Internet (thông qua các dịch vụ điện tử như các mạng xã hội, thương mại điện tử, y tế điện tử và chính phủ điện tử) đang tạo ra nhiều petabyte (hàng triệu gigabyte) dữ liệu cứ sau mỗi giây. Ví dụ như mạng kết nối xã hội Facebook được biết có đến 900 triệu người tham gia trên toàn thế giới và tạo ra trung bình hơn 1500 trạng thái cập nhật mỗi giây (Hachman, 2012; Bullas, 2011).

Với việc khai thác và kết nối (thế giới thực) ngày càng tăng của các bộ cảm biến thông qua các mạng cố định và di động (mạng cảm biến), ngày càng có nhiều các hoạt động ngoại tuyến cũng được ghi lại bằng kỹ thuật số, dẫn đến một làn sóng bổ sung dữ liệu không ngừng.

Nhiều tài liệu chỉ ra rằng, riêng trong năm 2010, các doanh nghiệp lưu trữ tổng thể hơn 7 exabyte (hàng tỷ gigabyte) dữ liệu mới trên các ổ đĩa, trong khi người tiêu dùng bảo quản hơn 6 exabyte dữ liệu mới (MGI, 2011). Điều đó dẫn đến một lượng dữ liệu tích lũy ước tính hơn 1000 exabyte vào năm 2010; một nhà phân tích ước tính rằng con số này sẽ tăng lên gấp 40 lần vào cuối thập kỷ này (IDC, 2012).

Hình 1: Kho dữ liệu ước tính trên phạm vi toàn thế giới, đơn vị exabyte (tỷ gigabyte)



Nguồn: OECD dựa trên dự báo nghiên cứu của IDC Digital Universe.

Tạo dữ liệu, thu thập và truyền tải

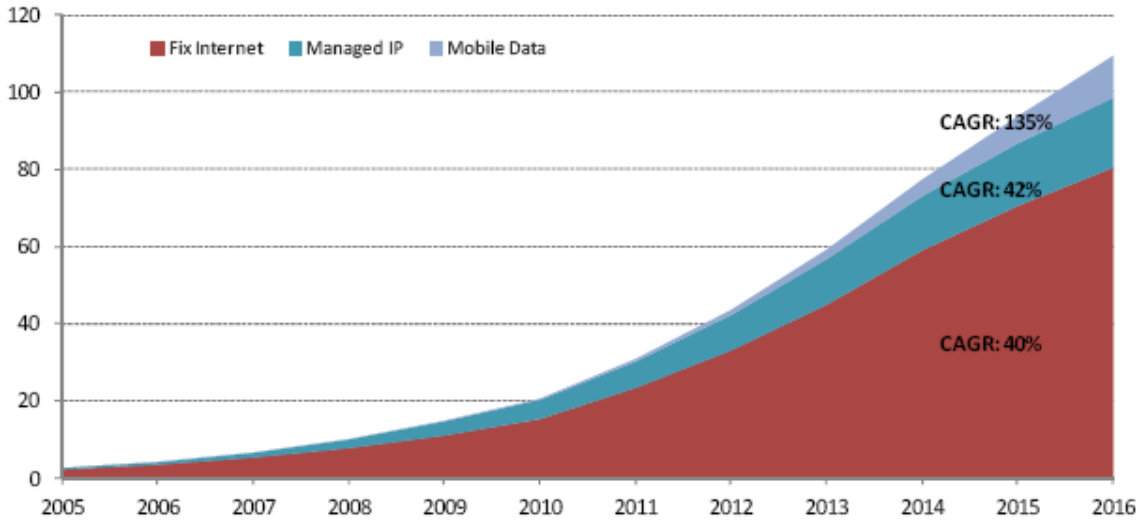
Lượng dữ liệu gia tăng một cách mạnh mẽ chủ yếu bị tác động bởi sự hội tụ của những phát triển công nghệ quan trọng, đáng chú ý là truy cập băng thông rộng ở mọi nơi và sự phổ biến các thiết bị và ứng dụng ICT thông minh, như các dụng cụ đo thông minh, lưới điện và giao thông vận tải thông minh dựa trên các mạng cảm biến và sự giao tiếp máy với máy (M2M). Chi phí truy cập Internet giảm mạnh trong vòng 20 năm qua là một yếu tố chi phối quan trọng. Ví dụ vào năm 2011, người tiêu dùng ở Pháp phải trả khoảng 33 USD một tháng cho một kết nối băng thông rộng tốc độ 51 Mbit/s, trong khi chi phí cho

một kết nối bằng quay số (với tốc độ chậm hơn đến 1000 lần) là 75 USD vào năm 1995. Điện thoại di động đã trở thành một thiết bị thu thập dữ liệu hàng đầu, kết hợp dữ liệu định vị địa lý với kết nối Internet để hỗ trợ các dịch vụ trên phạm vi rộng và ứng dụng mới liên quan đến giao thông, môi trường và y tế. Nhiều dịch vụ và ứng dụng đó dựa (hoặc tham gia vào) việc thu thập và sử dụng dữ liệu cá nhân. Bổ sung cho sự truy cập Internet ngày càng gia tăng và hiệu quả hơn, hầu hết các thiết bị di động được trang bị các mảng giao thức gia tăng để trao đổi dữ liệu cục bộ (như Wifi, Bluetooth, Near Field Communications (NFC) với khả năng truyền dữ liệu ngang hàng (peer-to-peer). Các thiết bị này còn có thể quay video, chụp ảnh và ghi âm thanh (thường gắn với thông tin định vị).

Vào năm 2011, toàn thế giới có gần sáu tỷ thuê bao di động, trong đó khoảng 13% (780 triệu) là điện thoại thông minh có khả năng thu thập và truyền dữ liệu định vị địa lý (ITU, 2012; Cisco, 2012). Cũng vào năm này, các thiết bị điện thoại di động tạo ra khoảng 600 petabyte (triệu gigabyte) dữ liệu mỗi tháng (Cisco, 2012). Với sự phổ cập điện thoại di động (số thuê bao trên 100 dân) vượt quá 100% tại hầu hết các nước OECD và sự phổ biến băng thông rộng không dây đạt gần 50%, thì nguồn dữ liệu này sẽ gia tăng đáng kể khi mà điện thoại thông minh trở thành thiết bị cá nhân phổ biến. Cisco (2012) ước tính rằng lưu lượng dữ liệu sản sinh ra từ điện thoại di động sẽ đạt gần 11 exabyte (hàng tỷ gigabyte) vào năm 2016, có nghĩa là tăng gần gấp đôi mỗi năm (xem hình 2).

Hình 2: Lưu lượng IP toàn cầu hàng tháng, 2005-16.

Đơn vị: exabyte (1 exabyte = 1 tỷ gigabyte)



Nguồn: OECD dựa trên số liệu của Cisco (2012).

Sự gia tăng dữ liệu di động không chỉ do sự gia tăng số điện thoại di động, được dự báo sẽ chiếm đến một nửa tổng lưu lượng di động vào năm 2016 (Cisco, 2012). Các thiết bị

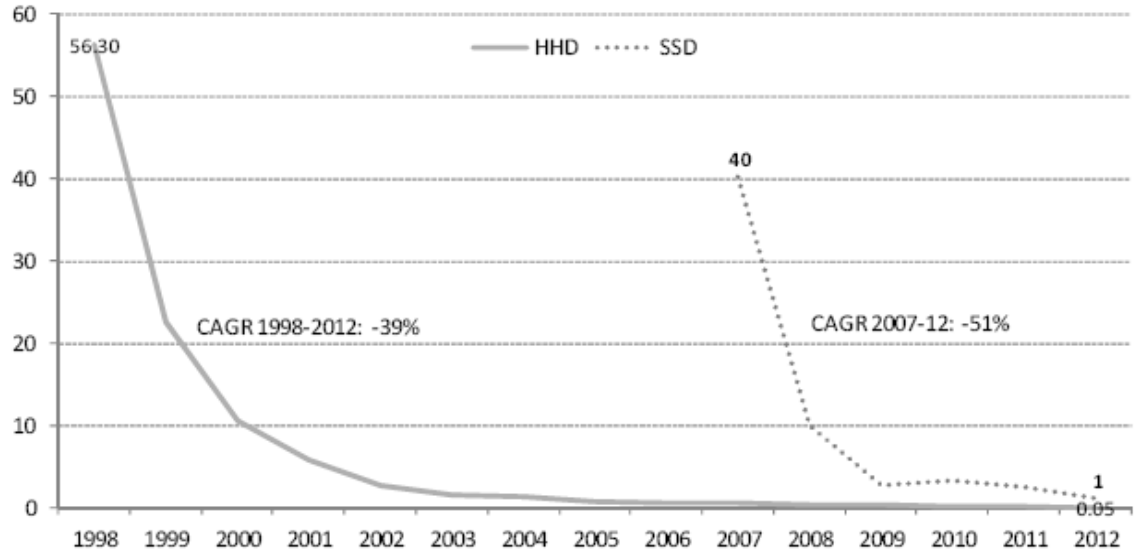
thông minh khác đang phát triển thậm chí còn nhanh hơn. Ví dụ, các dụng cụ đo thông minh thu thập và truyền dữ liệu thời gian thực ngày càng tăng (OECD, 2012), và xe ô tô thông minh giờ đây đã có thể truyền dữ liệu thời gian thực về hiện trạng các linh kiện trong xe và về môi trường (OECD, 2012). Nhiều thiết bị thông minh trong số này được dựa trên cơ sở các mạng kết nối cảm biến và thiết bị đi kèm có thể cảm nhận và tương tác với môi trường thông qua các mạng di động. Các bộ cảm biến và thiết bị đi kèm trao đổi dữ liệu thông qua các kết nối không dây "tạo khả năng tương tác giữa con người hay máy tính với môi trường xung quanh" (Verdone et al., 2008). Hơn 30 triệu bộ cảm biến kết nối tương tác hiện đang được triển khai trên phạm vi toàn thế giới trong các lĩnh vực như an ninh, y tế, môi trường, các hệ thống giao thông vận tải hay hệ thống kiểm soát năng lượng, số lượng của chúng đang tăng lên với tỷ lệ khoảng 30% một năm (MGI, 2011).

1.1.2. Lưu trữ và xử lý dữ liệu

Nếu như những phát triển công nghệ nêu trên chủ yếu thúc đẩy sự sản sinh và truyền tải dữ liệu, thì việc sử dụng dữ liệu đã trở nên dễ dàng hơn nhiều nhờ vào sự giảm mạnh chi phí lưu trữ, xử lý và phân tích dữ liệu. Trước đây, chi phí lưu trữ dữ liệu đã không khuyến khích việc giữ lại dữ liệu đã không còn hoặc có vẻ như không còn cần thiết (OECD, 2011). Nhưng chi phí lưu trữ đã giảm đến mức thấp để có thể lưu trữ dữ liệu trong thời gian dài, thậm chí là vô thời hạn. Điều này có thể được minh họa qua chi phí trung bình cho mỗi gigabyte ổ đĩa cứng (HDD), chi phí này đã giảm từ 56 USD năm 1998 xuống 0,05 USD năm 2012, tốc độ giảm trung bình hàng năm là gần 40% (xem hình 3). Với các công nghệ lưu trữ thế hệ mới như ổ đĩa thể rắn (SSD) chẳng hạn, chi phí trên mỗi gigabyte thậm chí còn giảm nhanh hơn.

Hình 3: Chi phí trung bình lưu trữ dữ liệu cho người tiêu dùng, 1998-2012

Đơn vị: USD/gigabyte

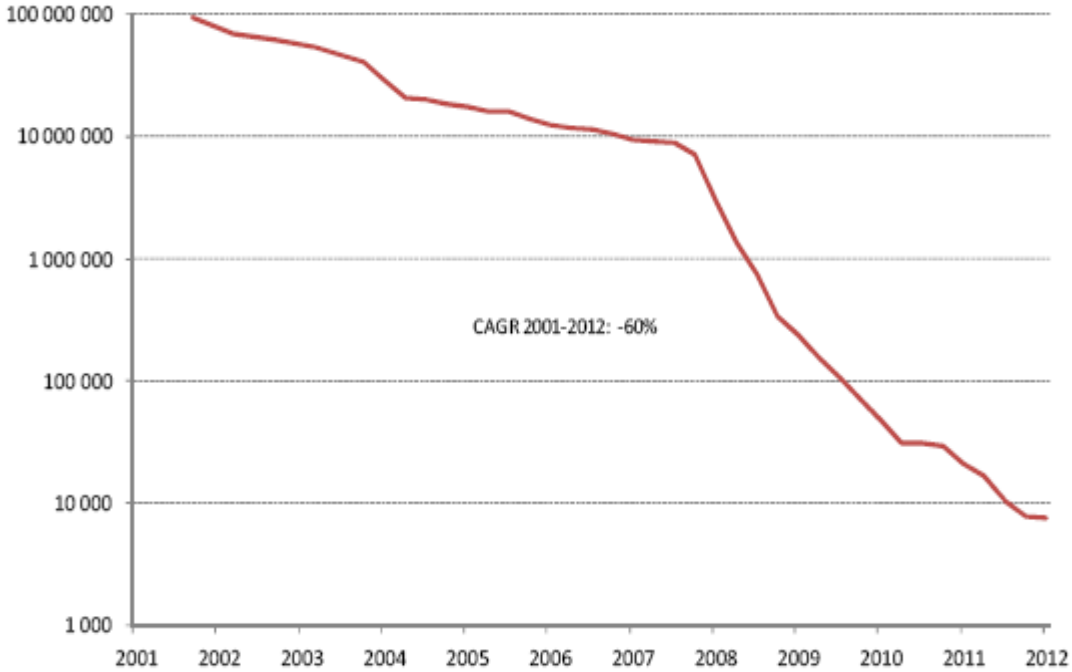


Nguồn: OECD trên cơ sở Pingdom (2011).

Định luật Moore phát biểu rằng tính năng xử lý tăng gấp đôi cứ sau 18 tháng, liên quan đến chi phí hay độ lớn chủ yếu đã được xác minh. Điều này đặc biệt đáng chú ý đối với các công cụ xử lý dữ liệu, chúng ngày càng trở nên có tính năng mạnh, tinh xảo, hiện diện mọi nơi và có giá rẻ, tạo điều kiện dễ dàng tìm kiếm dữ liệu, kết nối và truy xuất nguồn gốc, không chỉ các chính phủ và các tập đoàn lớn mà nhiều người khác đều có thể thực hiện được. Ví dụ như trong lĩnh vực di truyền, các máy lập trình tự gen ADN giờ đây có thể đọc được khoảng 26 triệu ký tự mã di truyền ở người trong chưa đầy một phút, và chi phí lập trình tự mỗi bộ gen đã giảm 60% một năm, trung bình từ 100 triệu USD năm 2001 xuống chưa đến 10.000 USD vào năm 2012 (xem hình 4).

Hình 4: Chi phí lập trình tự bộ gen, 2001-11

Đơn vị USD (theo thang đo logarit)



Nguồn: OECD dựa theo Viện nghiên cứu bộ gen người quốc gia Hoa Kỳ (www.genome.gov/sequencingcosts/)

Điện toán đám mây đóng vai trò quan trọng trong việc gia tăng khả năng lưu trữ và xử lý dữ liệu. Nó được mô tả như một "mô hình dịch vụ tính toán dựa trên một tập hợp tài nguyên máy tính có thể truy cập theo cách thức linh hoạt, mềm dẻo và theo nhu cầu với yêu cầu quản lý thấp" (OECD, 2012). Đặc biệt, đối với các doanh nghiệp vừa và nhỏ (SMEs), và cả các chính phủ không thể hoặc không muốn thực hiện những đầu tư lớn, phải thanh toán trước cho các công nghệ ICT, điện toán đám mây mang lại khả năng cho các tổ chức chi trả cho các nguồn lực siêu tính toán theo phương thức chi tiêu tùy theo khả năng (pay-as-you-go).

Các ứng dụng phần mềm nguồn mở (PMNM) bao gồm đầy đủ các giải pháp cần thiết cho dữ liệu lớn, chẳng hạn như để lưu trữ, xử lý và phân tích (bao gồm cả hiển thị trực quan - visualization), cũng góp phần đáng kể vào việc làm cho phân tích dữ liệu lớn có thể tiếp cận đến dân số rộng lớn hơn. Nhiều công cụ dữ liệu lớn được các công ty Internet phát triển ban đầu giờ đây được phổ biến rộng khắp nền kinh tế tạo ra các hàng hóa và dịch vụ mới dựa vào dữ liệu. Ví dụ, Hadoop, khung lập trình mã nguồn mở để quản trị dữ liệu phân tán, được lấy cảm hứng từ một bài báo của các nhân viên Google, Dean và Ghemawat (2004). Ban đầu nó được Yahoo! tài trợ và được các công ty Internet như Amazon, Facebook 11, 12 và LinkedIn khai thác và tiếp tục phát triển, sau đó được cung cấp bởi các nhà cung cấp cơ sở dữ liệu và máy chủ doanh nghiệp truyền thống như IBM, Oracle, Microsoft, và SAP như là một phần dòng sản phẩm của họ, và hiện đang được sử dụng rộng rãi cho các hoạt động dữ liệu chuyên sâu tại các doanh nghiệp thuộc đủ các loại như Wal-Mart (bán lẻ), Chevron (năng lượng) và Morgan Stanley (dịch vụ tài chính).

Ngày càng có nhiều nhà phân tích dữ liệu chuyên môn hóa và các nhà môi giới dữ liệu chào mời dữ liệu để sử dụng cho các mục đích như quảng cáo, kiểm tra lý lịch tuyển dụng việc làm, cấp tín dụng và thực thi pháp luật. Số các doanh nghiệp chào bán dữ liệu đã tăng mạnh trong những năm gần đây. Tại thời điểm năm 2013, tổ chức privacyrights.org đã liệt kê chỉ riêng ở Hoa Kỳ có đến 180 công ty môi giới dữ liệu trực tuyến đăng ký. Các hãng môi giới dữ liệu rất đa dạng, từ các công ty chuyên môn hóa giữa các doanh nghiệp (business-to-business) đến các dịch vụ nội bộ hóa đơn giản. Có thể kể đến các công ty như LexisNexis đã từng tuyến bố họ tiến hành hơn 12 triệu kiểm tra lý lịch một năm, và BlueKai Exchange tuyên bố là thị trường dữ liệu lớn nhất thế giới cho các nhà quảng cáo, công ty này sở hữu dữ liệu về hơn 300 triệu người tiêu dùng và hơn 30.000 thuộc tính dữ liệu. Theo thông tin công bố trên trang web của mình, BlueKai Exchange cho biết họ xử lý hơn 750 triệu sự kiện dữ liệu và giao dịch, thực hiện hơn 75 triệu cuộc bán đấu giá các thông tin cá nhân mỗi ngày.

1.1.3 Định nghĩa dữ liệu lớn

Có nhiều định nghĩa về "dữ liệu lớn" (Big data), và chúng có thể khác nhau tùy thuộc vào việc bạn là nhà khoa học máy tính, nhà phân tích tài chính hay một doanh nhân đang thuyết minh ý tưởng đầu tư mạo hiểm.

Nhiều tác giả mô tả đơn giản "dữ liệu lớn" như những kho chứa dữ liệu lớn (Large pools of data) (McGuire *et al*., 2012). Loukides (2010) định nghĩa đó là dữ liệu mà trong đó "*chính bản thân độ lớn của dữ liệu đã trở thành một phần của vấn đề*". Viện Nghiên cứu toàn cầu McKinsey (*McKinsey Global Institute* - MGI) cũng đưa ra định nghĩa tương tự "*đó là dữ liệu có độ lớn vượt quá khả năng các công cụ phần mềm cơ sở dữ liệu tiêu biểu có thể nắm bắt, lưu trữ, quản trị và phân tích*".

Hầu hết các định nghĩa phản ánh năng lực công nghệ ngày càng gia tăng để nắm bắt, tổng hợp và xử lý khối lượng dữ liệu với độ lớn, tốc độ và sự đa dạng lớn chưa từng thấy. Nói theo cách khác, "*dữ liệu giờ đây được cung cấp nhanh hơn, độ bao phủ và phạm vi*

lớn hơn, và bao gồm các chủng loại quan trắc và đo lường mới chưa từng có trước đây". Chính xác hơn, các tập hợp dữ liệu lớn là "*những tập hợp dữ liệu lớn, đa dạng, phức hợp, kéo dài (longitudinal), và/hoặc phân tán được tạo ra từ các công cụ, các cảm biến, các giao dịch trên Internet, email, video, các dữ liệu duyệt web, và/hoặc tất cả các nguồn số liệu khác có sẵn hiện có và trong tương lai*".

Theo định nghĩa của IBM, *Dữ liệu lớn* là sự thu thập, quản lý và phân tích dữ liệu, những việc đó đã vượt xa dữ liệu cấu trúc tiêu biểu, nó có thể được truy vấn với hệ thống quản trị dữ liệu quan hệ - thường với những tệp phi cấu trúc, video kỹ thuật số, hình ảnh, dữ liệu cảm biến, tệp lưu nhật ký, bất cứ dữ liệu nào không có trong hồ sơ với các phạm vi tìm kiếm khác.

Tên gọi Dữ liệu lớn không chỉ cho thấy tính chất lớn mà nó còn có tính phức tạp, hai tính chất này ở dữ liệu lớn luôn đi cùng nhau, trong đó tính chất "phức tạp" còn đặc trưng và thách thức hơn vấn đề về độ lớn của dữ liệu. Định nghĩa của IBM về dữ liệu lớn được đặc trưng bằng ba chữ V: Variety, Velocity và Volume. Chữ V đầu tiên chỉ sự đa dạng, sự liên kết chằng chịt của dữ liệu với nhiều kiểu dữ liệu phi cấu trúc, như dòng hình ảnh kỹ thuật số (digital video streams), dữ liệu cảm biến, cũng như các nhật ký tệp xử lý. Chữ V thứ hai chỉ tính chất chuyển động liên tục của dòng dữ liệu rất lớn cần xử lý, khác với cách truyền thống chúng ta thu nhận và xử lý dữ liệu theo từng mẻ (batch). Tốc độ dữ liệu gia tăng bởi vì băng thông mạng - điển hình như tốc độ gigabit ngày nay (gigE, 10G, 40G, 100G) được so sánh với tốc độ megabit. Chữ V thứ ba chỉ độ lớn của dữ liệu ở mức terabytes (10^{12}), rồi petabytes (10^{15} bytes), và cả exabytes (10^{18} bytes). IBM ước lượng, có 2,5 x 10^{18} bytes dữ liệu được tạo ra mỗi ngày.

Trong một số trường hợp, dữ liệu lớn được xác định bằng khả năng phân tích các tập hợp dữ liệu phi cấu trúc, chủ yếu từ các nguồn khác nhau như các web log, truyền thông xã hội, thông tin di động, các bộ cảm biến và các giao dịch tài chính. Điều này đòi hỏi khả năng liên kết các tập hợp dữ liệu; đó là điều cần thiết do thông tin mang tính phụ thuộc nhiều vào bối cảnh và có thể không có giá trị nếu không đúng với bối cảnh. Điều này cũng yêu cầu khả năng trích xuất thông tin từ các dữ liệu phi cấu trúc, có nghĩa là các dữ liệu còn thiếu một mô hình được xác định trước (rõ ràng hay tiềm ẩn). Các ước tính tỷ trọng dữ liệu phi cấu trúc ở các doanh nghiệp có thể chiếm từ 80% đến 85% và phần lớn chưa được khai thác hoặc khai thác quá ít. Trước đây, việc trích xuất giá trị từ các dữ liệu phi cấu trúc là công việc tốn nhiều công sức. Bằng phân tích dữ liệu lớn, các kho dữ liệu phi cấu trúc có thể liên kết và phân tích để trích xuất được những thông tin có giá trị tiềm tàng theo một cách thức tự động và hiệu quả.

Tiềm năng để tự động liên kết các tập hợp dữ liệu phi cấu trúc có thể minh họa qua sự tiến hóa của các công cụ tìm kiếm. Các nhà cung cấp dịch vụ tìm kiếm trên mạng như Yahoo! đã bắt đầu bằng các thư mục web có tính cấu trúc cao do con người biên tập. Các dịch vụ này đã không thể mở rộng phạm vi do nội dung online gia tăng. Các nhà cung cấp dịch vụ tìm kiếm đã phải áp dụng các chương trình tự động duyệt các nội dung web

(crawle) "*phi cấu trúc*". Yahoo! đã áp dụng duyệt tự động trang web là một nguồn chủ yếu của các kết quả tìm kiếm vào năm 2002. Khi đó Google đã sử dụng công cụ tìm kiếm của mình (dựa trên cơ sở thuật toán PageRank) đã được 5 năm, và thị phần của hãng này trong dịch vụ tìm kiếm đã chiếm hơn 80% vào năm 2012.

Ba đặc tính - số lượng, tốc độ và đa dạng, được coi là những đặc trưng chính của dữ liệu lớn và thường được viết tắt là 3V. Tuy nhiên, đây là các đặc tính kỹ thuật, chúng phụ thuộc vào sự phát triển của các công nghệ lưu trữ và xữ lý dữ liệu. Đến năm 2012, công ty nghiên cứu Gartner (hãng META Group) bổ sung thêm rằng Big Data, ngoài ba tính chất trên thì còn phải "cần đến các dạng xử lí mới để trợ giúp việc ra quyết định, khám phá sâu vào sự vật/sự việc và tối ưu hóa các quy trình làm việc". Khái niệm mới về Big Data 2014 của Gartner đưa ra mô hình **"5V"** bổ sung thêm hai tính chất quan trọng của Big Data, đó là *Veracity* (Độ chính xác): Một trong những tính chất phức tạp nhất của BigData là độ chính xác của dữ liệu. Với xu hướng kết nối mạng xã hội và truyền thông xã hội ngày nay và sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng di động làm cho bức tranh xác định về độ tin cậy và chính xác của dữ liệu ngày một khó khăn hơn. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là đặc tính quan trọng của BigData.

Value (Giá trị thông tin): Giá trị thông tin cũng là đặc tính quan trọng của xu hướng công nghệ Big Data. Đặc tính này liên quan đến giá trị kinh tế xã hội ngày càng gia tăng có thể thu được từ việc sử dụng dữ liệu lớn. Đây chính là giá trị kinh tế và xã hội tiềm năng cuối cùng sẽ thúc đẩy việc tích lũy, xử lý và sử dụng dữ liệu. Vì vậy, sẽ là thích hợp khi vượt xa hơn các khía cạnh kỹ thuật thuần túy về độ lớn, tốc độ và sự đa dạng để xem xét đến khía cạnh kinh tế xã hội của dữ liệu lớn như một "nhân tố sản xuất mới" (Gentile, 2011; Jones, 2012).

Điều thực sự quan trọng về Dữ liệu lớn là những gì nó thực hiện. Ngoài việc chúng ta định nghĩa Dữ liệu lớn như một hiện tượng công nghệ, tiềm năng sử dụng đa dạng đối với phân tích dữ liệu đặt ra những câu hỏi quan trọng về việc liệu các chuẩn mực luật pháp, đạo đức và xã hội của chúng ta đã đủ để bảo vệ sự riêng tư và các giá trị khác trong một thế giới dữ liệu lớn hay chưa. Khả năng tính toán và mức độ tinh vi chưa từng có tiền lệ đã làm cho những khám phá, những sáng tạo và tiến bộ bất ngờ trở nên khả dụng phục vụ chất lượng cuộc sống của chúng ta. Nhưng những năng lực đó, hầu hết đều không thể nhìn thấy hay có sẵn đối với những người tiêu dùng bình thường, nó cũng tạo ra một sự bất cân xứng về quyền lực giữa những ai nắm giữ dữ liệu và những người cung cấp chúng một cách cố ý hoặc không cố ý.

Một phần của thách thức nằm ở việc hiểu được nhiều ngữ cảnh khác nhau trong đó dữ liệu bắt đầu có hiệu lực. Dữ liệu lớn có thể được coi là tài sản, một nguồn lực công, hay một biểu hiện đặc trưng cá nhân. Các ứng dụng dữ liệu lớn có thể là động lực thúc đẩy kinh tế tương lai hoặc cũng là mối đe dọa đối với quyền tự do được ưu chuộng. Dữ liệu lớn có thể là tất cả những điều đó. Cả công nghệ dữ liệu lớn và các lĩnh vực công nghiệp

hỗ trợ nó đều đang không ngừng đổi mới và thay đổi.

1.2. Giá trị của dữ liệu ngày càng gia tăng trong nền kinh tế

Doviệc lưu trữ và xử lý dữ liệu ngày càng trở nên tinh xảo, phổ biến và có chi phí rẻ, nên các tổ chức trong nền kinh tế đang sử dụng những lưu lượng dữ liệu lớn cho các hoạt động hàng ngày của mình. Brynjolfsson et al. (2011) ước tính rằng sản lượng đầu ra và năng suất của các công ty áp dụng ra quyết định dựa trên dữ liệu cao hơn từ 5-6% so với ước tính các khoản đầu tư khác của họ vào sử dụng công nghệ thông tin. Các doanh nghiệp này cũng hoạt động tốt hơn theo các khía cạnh sử dụng tài sản, thu nhập trên vốn cổ phần và giá trị thị trường. Đầu tư gia tăng vào quản trị và phân tích dữ liệu phản ánh một phần vai trò kinh tế ngày càng tăng của dữ liệu. Ví dụ, riêng giá thị trường của các hệ thống quản trị cơ sở dữ liệu quan hệ đã có giá hơn 21 tỷ USD trong năm 2011, tăng trung bình 8%/năm kể từ năm 2002. Có lẽ điều đáng quan tâm hơn đối với dữ liệu lớn đó là nhu cầu về các hệ thống cơ sở dữ liệu không quan hệ (NoSQL), trí tuệ doanh nghiệp (BI) và phần mềm phân tích đã gia tăng mạnh trong những năm gần đây khi phân tích dữ liệu tiếp tục phát triển, đặc biệt là đối với việc ra quyết định dựa trên dữ liệu.

Khối lượng dữ liệu liên quan có thể khác biệt đáng kể giữa các ngành, một số lĩnh vực có thể có cường độ dữ liệu chuyên sâu hơn so với các lĩnh vực khác. Theo MGI (2011), cường độ dữ liệu (được tính theo khối lượng dữ liệu bình quân mỗi tổ chức) thuộc loại cao nhất trong lĩnh vực dịch vụ tài chính (bao gồm các dịch vụ chứng khoán, đầu tư và ngân hàng), truyền thông và các phương tiện thông tin đại chúng, các tổ chức tiện ích (cung cấp hàng hóa cơ bản như điện, nước), chính phủ, và chế tạo linh kiện. Trong các lĩnh vực này, mỗi một tổ chức lưu trữ trung bình hơn 1000 terabytes (hay một petabyte - một triệu tỷ) dữ liệu vào thời điểm năm 2009. Một xếp hạng tương tự có thể rút ra từ con số ước tính về số các nhà chuyên gia quản trị và phân tích dữ liệu (các nhà khoa học dữ liệu) bình quân trên mỗi 1000 nhân viên trong từng lĩnh vực. Giả định ngầm có thể rút ra là các ngành này càng sử dụng nhiều nhân lực khoa học dữ liệu hơn khi các hoạt động càng có cường độ chuyên sâu dữ liệu hơn.

Theo các cuộc điều tra dân số tại Hoa Kỳ, số các ngành sử dụng bình quân một nhà quản trị cơ sở dữ liệu hoặc nhiều hơn bình quân 10.000 nhân viên đã tăng lên trong vòng chín năm gần đây. Vào năm 2012, có năm ngành công nghiệp có tỷ lệ sử dụng các nhà quản trị cơ sở dữ liệu lớn nhất là các lĩnh vực: hoạt động tài chính (22 nhà quản trị cơ sở dữ liệu trên 10.000 nhân viên); dịch vụ chuyên môn và kinh doanh (12); bán buôn và bán lẻ (6); chế tạo (6); thông tin, hành chính công và các dịch vụ khác (5). Tỷ lệ bình quân các quản trị viên cơ sở dữ liệu trong các lĩnh vực này cũng đã tăng lên đáng kể trong những năm gần đây, với đỉnh điểm có đến hơn 160 nhà quản trị cơ sở dữ liệu trên 10.000 nhân viên tại Hoa Kỳ vào năm 2011. Hầu hết các lĩnh vực thâm dụng dữ liệu cũng có xu hướng có cường độ sử dụng ICT cao (chi tiêu ICT tính theo tỷ trọng sản lượng đầu ra); tuy nhiên, lĩnh vực khai thác khoáng sản lại chỉ sử dụng một số lượng nhỏ các nhà quản trị cơ sở dữ liệu.

Sự khác biệt về cường độ dữ liệu cho thấy giá trị của dữ liệu có thể khác nhau đáng kể giữa các ngành (OECD, 2012d). Các nghiên cứu thực nghiệm chỉ ra sự phụ thuộc ngữ cảnh không chỉ ở cấp doanh nghiệp, mà còn cả ở cấp nhân viên (Acquisti et al., 2011). Điều này làm cho đánh giá tác động kinh tế vĩ mô khó khăn hơn, và cho thấy sự cần thiết phải nghiên cứu cụ thể để hiểu được tác động trong từng lĩnh vực hoặc từng phần trong chuỗi giá trị dữ liệu. Các nghiên cứu cụ thể đã chỉ ra giá trị tiềm năng của dữ liệu trong năm lĩnh vực. Các lĩnh vực này đã được xác định trong các tài liệu và các nghiên cứu trước đây của OECD là những lĩnh vực có khả năng sử dụng dữ liệu cao, coi đó như một nguồn lực của đổi mới sáng tạo và tăng năng suất (OECD 2009b; 2012a; 2012b; 2012c). Năm lĩnh vực đó bao gồm: quảng cáo (trực tuyến), hành chính công, chăm sóc sức khỏe, tiện ích, dịch vụ hậu cần và giao thông vận tải. Trong các lĩnh vực này, một số được lựa chọn bởi họ khai thác dữ liệu dưới mức, mặc dù đó là các lĩnh vực thâm dụng dữ liệu (hành chính công, tiện ích trong một chừng mực nào đó). Các lĩnh vực khác hiện nay còn có cường độ dữ liệu thấp nhưng sẽ phải đối mặt với khối lượng dữ liệu mới ngày càng gia tăng, chẳng hạn như dòng dữ liệu nhấp chuột (click-stream data) trong quảng cáo trực tuyến, dữ liệu định vị địa lý (vận tải), dữ liệu đo lường thông minh (tiện ích), và hồ sơ y tế (chăm sóc sức khỏe), trong đó nếu khai thác đầy đủ, có thể tạo ra những lợi ích tăng thêm. Tính gộp lại với nhau, các lĩnh vực này chiếm trung bình khoảng một phần tư tổng giá trị gia tăng tại mười quốc gia thuộc OECD có số liệu đầy đủ. Tổng thể, triển vọng của dữ liệu lớn nằm ở một hoặc nhiều lĩnh vực liên quan đến đổi mới sáng tạo sau đây:

- Sử dụng dữ liệu để tạo ra các sản phẩm mới (hàng hóa và dịch vụ). Điều này bao gồm việc sử dụng dữ liệu như một sản phẩm (sản phẩm dữ liệu) hay như một thành phần chủ yếu của sản phẩm (sản phẩm thâm dụng dữ liệu);
- Sử dụng dữ liệu để tối ưu hóa hoặc tự động hóa các quy trình sản xuất hoặc cung ứng (các quy trình dựa vào dữ liệu). Điều này bao gồm việc sử dụng dữ liệu để nâng cao hiệu quả phân phối các nguồn năng lượng (lưới điện thông minh), hậu cần và giao thông vận tải (hậu cần và giao thông vận tải thông minh).
- Sử dụng dữ liệu để cải tiến marketing, ví dụ bằng cách cung cấp quảng cáo và tư vấn cá nhân hóa hay các loại hình phân biệt đối xử liên quan đến marketing (marketing dựa vào dữ liệu) cũng như sử dụng dữ liệu để thiết kế sản phẩm thử nghiệm (thiết kế sản phẩm dựa vào dữ liệu) (Brian, 2012);
- Sử dụng dữ liệu để phục vụ cho các phương thức tổ chức và quản lý mới hoặc để cải tiến các thực hành hiện tại (tổ chức dựa trên dữ liệu và ra quyết định dựa vào dữ liệu) (Brynjolfsson et al., 2011).
- Sử dụng dữ liệu để tăng cường nghiên cứu và phát triển (NC&PT dựa vào dữ liệu). Điều này bao gồm các phương pháp mới thâm dụng dữ liệu phục vụ khám phá khoa học bằng cách tăng thêm "một lĩnh vực nghiên cứu mới dựa trên việc khai thác những hiểu biết mới từ các tập hợp dữ liệu rộng lớn và đa dạng" (EC, 2010).

1.2.1. Quảng cáo trực tuyến

Dữ liệu được tạo ra khi người dùng sử dụng Internet có thể tạo ra giá trị và mang lại cho các công ty các cơ hội để cải tiến các hoạt động và tiếp thị các sản phẩm của mình theo cách có hiệu quả hơn. Việc tiến hành marketing dựa vào dữ liệu hoàn toàn có thể thực hiện, ví dụ, dòng dữ liệu nhấp chuột được thu thập sử dụng sự kết hợp giữa mã phần mềm như web-bugs và cookies cho phép các nhà quảng cáo theo dõi các thói quen duyệt web của khách hàng. Đối với các doanh nghiệp, việc khai thác dòng dữ liệu nhấp chuột (click-stream data) cung cấp các phương tiện mới để cải tiến việc quản lý quan hệ khách hàng. Trước đây, khi một khách hàng tương tác ngoại tuyến với một công ty, dấu vết thông tin thường phân tán và hạn chế. Một doanh nghiệp chỉ có thể thu thập các dữ liệu quét khi khách hàng thanh toán sử dụng thẻ khách hàng thường xuyên để suy đoán về mối quan tâm của khách hàng đối với một phạm vi rộng hơn các sản phẩm. Bằng dòng dữ liệu nhấp chuột, các doanh nghiệp giờ đây có được nhiều thông tin hơn. Ví dụ, các công ty giờ đây có các thông tin về các trang web để giới thiệu công ty với người sử dụng, cho dù sử dụng một công cụ tìm kiếm hay sử dụng các cụm từ đều có thể tiếp cận trang web công ty. Điều này cho phép các doanh nghiệp có thể phân bổ ngân sách marketing của mình hiệu quả hơn và nhằm mục tiêu vào các trang web có thể tiếp cận với những khách hàng có giá trị nhất của họ. Ngoài ra, các doanh nghiệp có thể phát hiện chính xác những gì người sử dụng muốn tìm kiếm trên một trang web. Điều này cho phép họ nâng cao kinh nghiệm trực tuyến của người sử dụng dựa trên bằng chứng thực nghiệm và các phương pháp thống kê như thử nghiệm kiểm tra phân tách (A/B testing), không phải chỉ cải thiện kinh nghiệm của các nhà phát triển web.

Việc thu thập dữ liệu không giới hạn ở trang web của công ty. Bằng cách sử dụng các nhà cung cấp dịch vụ, như các trang web mạng xã hội và các mạng lưới quảng cáo, các doanh nghiệp cũng có thể thu thập dữ liệu được tạo ra ở các nơi khác. Những dữ liệu như vậy hiện diện ngày càng tăng thông qua các thị trường dữ liệu và có thể kết hợp với dữ liệu từ các nguồn như: dữ liệu điều tra dân số, hồ sơ bất động sản, đăng ký xe, v.v... Những dữ liệu đó bổ sung thêm hồ sơ người dùng sau đó được bán cho các nhà quảng cáo đang tìm kiếm những người tiêu dùng với các hồ sơ cụ thể để nhằm cải thiện việc nhằm mục tiêu hành vi. Ví dụ, comScore, một nhà môi giới dữ liệu có trụ sở tại Hoa Kỳ, thu thập dữ liệu trên các trang web được hơn 2 triệu người tham gia trên toàn thế giới truy cập, bao gồm các thuật ngữ tìm kiếm mà họ sử dụng trên các công cụ tìm kiếm và cả lịch sử mua sắm trực tuyến của họ. Hãng comScore sau đó bao gói lại các thông tin này và bán các báo cáo và dịch vụ dữ liệu cho thấy các xu hướng doanh thu thương mại điện tử, lưu lượng truy cập trang web và các chiến dịch quảng cáo trực tuyến. Báo cáo như vậy được chào bán cho các công ty Fortune 500 (*Fortune 500 là* bảng xếp hạng danh sách 500 công ty lớn nhất Hoa Kỳ theo tổng thu nhập mỗi công ty*)* và các công ty truyền thông.

Nhìn chung, đặc biệt là trong 5 năm gần đây, doanh thu từ quảng cáo trực tuyến đã tăng nhanh hơn rất nhiều so với những gì mà các kênh quảng cáo truyền thống đã làm

được trong 15 năm đầu tiên. Ví dụ như trong quý một của năm 2012, doanh thu từ quảng cáo trực tuyến của 500 nhà quảng cáo hàng đầu tại Hoa Kỳ đã đạt 8,4 tỷ USD, theo Báo cáo Quảng cáo Internet IAB gần đây nhất (BusinessWire, 2012). Con số này cao hơn 1,1 tỷ USD (15%) so với quý đầu của năm 2011. Trong năm 2011, AdWords đã tạo ra trung bình hơn 20 triệu USD một tháng từ 20 trang web hàng đầu. Kết quả này phần lớn nhờ vào khả năng gia tăng nhằm vào khách hàng tiềm năng và các kết quả đánh giá. Tuy nhiên, giá trị gia tăng không chỉ giới hạn ở doanh thu quảng cáo. Ở đây còn có nhiều lợi ích cho người tiêu dùng. Theo McKinsey (2010), người tiêu dùng tại Hoa Kỳ và Châu Âu được hưởng lợi ích trị giá 100 tỷ euro năm 2010 từ các dịch vụ web hỗ trợ quảng cáo. Giá trị này còn cao hơn gấp ba lần doanh thu từ quảng cáo và cho thấy giá trị tạo ra cho người dùng còn lớn hơn thu nhập từ quảng cáo.

1.2.2. Các cơ quan chính phủ và khu vực công

Khu vực công là người sử dụng và cũng là nguồn dữ liệu quan trọng. Trên thực tế đây là một trong số các khu vực sử dụng dữ liệu với cường độ lớn nhất của nền kinh tế. Ví dụ như tại Hoa Kỳ, các cơ quan thuộc khu vực công lưu trữ trung bình 1,3 petabytes dữ liệu vào thời điểm năm 2011, là khu vực thâm dụng dữ liệu lớn thứ năm đất nước. Tuy nhiên, bằng chứng cho thấy rằng khu vực công không khai thác được đầy đủ tiềm năng của dữ liệu do khu vực này tạo ra và thu thập được, và cũng không khai thác được tiềm năng của dữ liệu do các nơi khác tạo ra (MGI, 2011; OECD, 2012). Tuy nhiên, khả năng truy cập được cải thiện và việc dùng lại dữ liệu khu vực công (PSI) mang lại nhiều lợi ích tiềm năng, chẳng hạn như cải thiện tính minh bạch trong khu vực công, việc cung cấp các dịch vụ công cộng trở nên hiệu quả và sáng tạo hơn hoặc được cá nhân hoá hơn, và việc hoạch định chính sách công và ra quyết định cũng kịp thời hơn.

Các ước tính chỉ ra rằng việc khai thác dữ liệu tốt hơn có thể đẩy mạnh hiệu quả, và có thể giúp tiết kiệm hàng tỷ đôla cho khu vực công. Theo MGI (2011), việc sử dụng đầy đủ dữ liệu lớn tại 23 chính phủ lớn nhất châu Âu có thể giảm các chi phí hành chính từ 15% đến 20%, tạo nên các giá trị mới tương đương từ 150 tỷ euro đến 300 tỷ euro, và thúc đẩy năng suất tăng trưởng 0,5% mỗi năm trong vòng 10 tới. Những lợi ích chủ yếu sẽ là hiệu quả lớn hơn (do tính minh bạch lớn hơn), thu thuế gia tăng (do các dịch vụ phù hợp với yêu cầu của khách hàng), và ít gian lận và sai sót hơn (nhờ phân tích dữ liệu tự động). Các nghiên cứu tương tự của Vương quốc Anh cho thấy, khu vực công có thể tiết kiệm 2 tỷ Bảng trong phát hiện gian lận và tạo ra 4 tỷ Bảng nhờ vào quản lý hiệu suất tốt hơn do sử dụng phân tích dữ liệu lớn (CEBR, 2012).

Các ước tính trên còn chưa bao gồm những lợi ích đầy đủ đối với việc hoạch định chính sách có thể thu được nhờ vào dữ liệu thời gian thực và thống kê. Một lĩnh vực có mối quan tâm ngày càng tăng trong bối cảnh này đó là an ninh nội bộ và thực thi pháp luật. Ví dụ như CitiVox là một công ty mới khởi sự giúp các chính phủ khai thác các nguồn dữ liệu phi truyền thống như SMS (tin nhắn văn bản) và truyền thông xã hội để bổ sung cho số liệu thống kê tội phạm chính thức. Khách hàng hiện tại là các Chính phủ ở

Trung và Nam Mỹ, những nơi có tỷ lệ tội phạm khá lớn không bị tố cáo. Bằng cách cung cấp cho các công dân các phương tiện kỹ thuật số để tố cáo tội phạm, hệ thống của CitiVox cho phép các cá nhân có thể giữ kín danh tính. Đồng thời, các nhà hoạch định chính sách và các cơ quan thực thi pháp luật có thể khai thác các dữ liệu gọi đến về các mẫu hình tội phạm mà sẽ không bị phát hiện (hoặc không đủ nhanh) thông qua các số liệu thống kê chính thức.

Hơn nữa, các ước tính trên không bao gồm lợi ích có thể đạt được thông qua việc cung cấp thông tin của khu vực công, theo khuyến cáo của Hội đồng OECD về *Tăng cường truy cập và sử dụng thông tin khu vực công hiệu quả hơn* (OECD, 2008) được định nghĩa là một phạm vi rộng các thông tin có thể sử dụng thương mại "bao gồm các sản phẩm và dịch vụ thông tin phát sinh, được tạo ra, thu thập, xử lý, bảo quản, lưu trữ, phổ biến, hoặc được tài trợ bởi Chính phủ hoặc cho chính phủ hay các tổ chức công". Các kết quả có lợi đối với đời sống kinh tế và xã hội có thể liệt kê như thời tiết đối với ùn tắc giao thông, thống kê tội phạm địa phương, các chức năng chính phủ minh bạch hơn, chẳng hạn như mua sắm hay kiến thức giáo dục và văn hóa phục vụ dân số rộng lớn hơn qua các tạp chí và kho dữ liệu mở cũng như các thư viện điện tử.

Do tiềm năng của dữ liệu khu vực công (PSI) đang trở nên được công nhận rộng rãi hơn, một số chính phủ đã tiến hành các xúc tiến "dữ liệu mở" có thể làm tăng nhanh tác động và vai trò của PSI. Các xúc tiến này đang trở thành một phương tiện có giá trị để phát triển hàng hóa và dịch vụ bổ sung và khuyến khích sự nổi lên của các "doanh nghiệp cộng đồng" cung cấp các dịch vụ xã hội dựa trên dữ liệu khu vực công. Bằng cách cung cấp truy cập và dùng lại dữ liệu chính phủ mở, các chính phủ đẩy mạnh việc thiết kế và cung cấp dịch vụ đổi mới sáng tạo, mà không cần phải xây dựng các giải pháp từ nguồn đến đích (end-to-end) mới. Ví dụ, người dân ngày càng sử dụng PSI có sẵn để phát triển các ứng dụng điện thoại di động tạo điều kiện dễ dàng tiếp cận các dịch vụ hiện có và cung cấp các dịch vụ mới (m-government). Ngoài ra, thông qua hợp tác với các cộng đồng trực tuyến, chất lượng dữ liệu có thể được cải thiện và tính toàn vẹn của dữ liệu chính phủ được kiểm tra cẩn thận.

Đầu tư vào PSI tại Hoa Kỳ đã được ước tính có trị giá hàng chục tỷ USD (Uhlir, 2009). Việc lập mô hình ban đầu chỉ ra rằng trong hơn ba thập kỷ qua, lợi ích của truy cập mở tới tài liệu lưu trữ có thể cao hơn chi phí gần tám lần (Houghton et al., 2010). Một nghiên cứu khác, đánh giá các nguồn thông tin khu vực công tại châu Âu (MEPSIR) (EC, 2006) đã kết luận rằng thị trường PSI dùng lại trực tiếp trong năm 2006 đối với các quốc gia EU25 cộng thêm Na Uy có trị giá 27 tỷ euro.

1.2.3. Y tế

Lượng dữ liệu sử dụng trong ngành y tế ngày càng gia tăng, liên quan đến việc quản lý hệ thống y tế và sử dụng phổ biến các hồ sơ y tế điện tử. Các xét nghiệm chẩn đoán, hình ảnh trong y tế và ngân hàng các mẫu phẩm sinh học cũng đang tạo ra những dữ liệu mới. Hiện nay, có những bộ sưu tập ảnh chụp y tế rất lớn, riêng ảnh chụp quang tuyến vú ở

Hoa Kỳ đã lên đến 2,5 petabytes được lưu trữ hàng năm (EC, 2010).

Có thể nói là những lợi ích từ dữ liệu mang lại cho khu vực công cũng tương đương như đối với lĩnh vực y tế, việc sử dụng dữ liệu tốt hơn có thể có những tác động quan trọng đối với ngành này cũng như đối với toàn bộ nền kinh tế. Trong lĩnh vực y tế, dữ liệu có thể giúp hệ thống chăm sóc sức khỏe nâng cao được hiệu quả, độ an toàn, đặt tâm điểm vào bệnh nhân và còn giúp các nhà nghiên cứu và các bác sĩ đánh giá các kết quả, xác định các mối tương quan không được quan sát trước đây, và thậm chí có thể dự đoán được những thay đổi trong quá trình lâm sàng thiết yếu và đưa ra các biện pháp can thiệp (Bollier, 2010). Khi dữ liệu dân số từ các nguồn khác nhau được liên kết với dữ liệu của ngành y tế, một số nguyên nhân gây ra bệnh tật có thể được hiểu rõ hơn. Một ví dụ là việc phân tích các yếu tố môi trường của các bệnh liên quan đến dinh dưỡng, áp lực và sức khỏe tâm thần (OECD-NSF, 2011).

Việc chia sẻ dữ liệu y tế thông qua các hồ sơ y tế điện tử có thể tạo cơ hội tiếp cận với dịch vụ y tế và có thể mang đến những hiểu biết sâu phục vụ đổi mới sản phẩm và dịch vụ, kể cả nghiên cứu về các loại thuốc và phương pháp điều trị mới. Các nguồn dữ liệu sức khỏe cá nhân khác có thể bao gồm các ứng dụng giám sát từ xa, thu thập số liệu về các điều kiện lâm sàng cụ thể hoặc các điều kiện sinh hoạt hàng ngày, ví dụ như để biết được khi nào thì một người sức khỏe yếu cần được giúp đỡ. Dữ liệu sức khỏe cá nhân cũng ngày càng được nhiều cá nhân cung cấp, được lưu trữ và trao đổi trực tuyến thông qua các mạng xã hội chú trọng y tế. Mạng xã hội PatientsLikeMe không chỉ cho phép những người có vấn đề sức khỏe có thể tương tác, tìm kiếm sự an ủi và học hỏi từ những người khác có cùng hoàn cảnh, nó còn có vai trò như cơ sở bằng chứng về dữ liệu cá nhân để phân tích và là nền tảng cho việc kết nối bệnh nhân với các thử nghiệm lâm sàng. Mô hình kinh doanh này phụ thuộc vào việc làm hài hòa giữa lợi ích của bệnh nhân với lợi ích của ngành; PatientsLikeMe bán các dữ liệu đã được xử lý, tổng hợp, mã hóa danh tính (de-identified) cho các đối tác, bao gồm các công ty dược phẩm và các nhà sản xuất thiết bị y tế, để giúp họ hiểu rõ hơn về các trải nghiệm thực tế của bệnh nhân và quá trình tác động của một căn bệnh. PatientsLikeMe còn chia sẻ dữ liệu bệnh nhân với các cộng sự nghiên cứu trên toàn thế giới.

Các nhà cung cấp dịch vụ y tế lớn như Kaiser Permanente (một tập đoàn y tế quản lý tại Mỹ) sử dụng các tập hợp dữ liệu để phát hiện ra những tác dụng bất lợi không được lường trước của thuốc, như Vioxx tuy không bị phát hiện trong các thử nghiệm lâm sàng nhưng đã được phát hiện thông qua khai thác các dữ liệu tạo ra khi loại thuốc này được kê đơn và sử dụng (MGI, 2011). Viện Y học và kinh nghiệm lâm sàng Vương quốc Anh cũng đã sử dụng các bộ dữ liệu lâm sàng lớn để đánh giá hiệu quả chi phí của các loại thuốc và phương pháp trị liệu mới, dẫn đến các kết quả được cải thiện với chi phí thấp hơn. Nhìn rộng hơn, dữ liệu liên kết có thể làm giảm các chi phí liên quan đến điều trị không đúng mức hoặc quá mức, nó còn có thể giúp phòng chống các căn bệnh mãn tính bằng cách xác định các nguyên nhân hành vi và qua đó hướng dẫn các can thiệp trước khi

phát bệnh (Bollier, 2010). MGI (2011) ước tính rằng dữ liệu lớn nếu được sử dụng trên toàn bộ hệ thống chăm sóc sức khỏe của Hoa Kỳ, như các hoạt động lâm sàng, thanh toán và định giá dịch vụ, NC&PT, có thể tiết kiệm được hơn 300 tỷ USD, hai phần ba số này xuất phát từ việc giảm được 8% chi phí chăm sóc sức khỏe. Những ước tính này vẫn chưa bao gồm các lợi ích từ phân tích dữ liệu, tạo cơ hội cho hoạch định các chính sách y tế công cộng kịp thời thông qua các số liệu thống kê thời gian thực giống như những dữ liệu tìm kiếm trên mạng để đánh giá xu hướng phát triển bệnh cúm ngay trong thời gian thực (Polgreen et al, 2009).

1.2.4. Dịch vụ tiện ích

Tiện ích "thông minh" được triển khai để phục vụ sản xuất, phân phối và tiêu thụ năng lượng hiệu quả hơn, nhưng ngày càng được sử dụng cho các nguồn tài nguyên thiên nhiên khác như nước. Ví dụ, lưới điện "thông minh" là các mạng điện có khả năng thông tin và truyền thông nâng cao, có thể giải quyết được những thách thức lớn của ngành điện lực trong chuỗi giá trị từ phát điện đến tiêu thụ. Những thách thức này bao gồm quản lý mức tiêu thụ đỉnh, mà thường dẫn đến chi phí phát thải CO_2 cao, và tích hợp các nguồn năng lượng tái tạo dễ bay hơi trong quá trình sản xuất năng lượng và giảm thất thoát trong truyền tải và phân phối năng lượng.

Tiện ích "thông minh" chủ yếu dựa trên dữ liệu thu thập được thông qua "công-tơ thông minh" tại các hộ gia đình và người tiêu dùng và với các nguồn năng lượng khác. Các thiết bị thông minh này tạo ra khả năng liên lạc hai chiều trên chuỗi giá trị, cho phép không chỉ thu thập dữ liệu tiêu thụ trong thời gian thực, mà còn có thể trao đổi dữ liệu về giá cả trong thời gian thực và (thực hiện) các tín hiệu điều khiển bật hoặc tắt các thiết bị trong gia đình và doanh nghiệp. Các ước tính chỉ ra rằng việc kết nối một triệu ngôi nhà vào lưới điện thông minh có thể tạo ra 11 gigabyte dữ liệu một ngày; điều này có thể làm nảy sinh những thách thức to lớn đối với quản trị và phân tích dữ liệu (OECD, 2009). Trong khi vòng phản hồi thông tin cho phép người tiêu dùng có thể điều chỉnh sự tiêu thụ của họ theo năng lực sản xuất, các nhà cung cấp dịch vụ tiện ích giờ đây có thể tiến hành phân tích dữ liệu để xác định các mẫu hình tiêu thụ tổng thể và dự báo nhu cầu. Điều đó có thể giúp họ điều chỉnh năng lực sản xuất và cơ chế định giá phù hợp với nhu cầu tương lai. Nói chung, việc sử dụng các ứng dụng lưới điện thông minh dựa trên dữ liệu có thể làm giảm lượng phát thải CO2 hơn 2 gigatonnes (tương đương 79 tỷ euro) .

Ngoài ra, dữ liệu thu thập được từ các mạng phân phối cho phép các nhà cung cấp dịch vụ tiện ích có thể xác định những thiệt hại và rò rỉ trong quá trình phân phối năng lượng và các nguồn lực khác. Bằng cách triển khai các bộ đo cảm biến nước thông minh kết hợp với phân tích dữ liệu, hãng Aguas Antofagasta, một công ty tiện ích cung cấp nước của Chile đã có thể xác định các sự cố rò rỉ nước trên toàn bộ mạng lưới phân phối và giảm được thất thoát nước từ 30% xuống 23% trong vòng 5 năm qua, do đó tiết kiệm được 800 triệu lít nước một năm.

Cũng giống như trong trường hợp dữ liệu khu vực công, việc mở ra dữ liệu đồng hồ đo

thông minh đến với thị trường đã dẫn đến một lĩnh vực công nghiệp mới cung cấp hàng hóa và dịch vụ đổi mới sáng tạo dựa trên những dữ liệu này, đã góp phần vào tăng trưởng xanh và tạo ra số lượng lớn việc làm xanh. Ví dụ như Opower, một doanh nghiệp mới khởi sự có trụ sở tại Hoa Kỳ đã liên kết với các nhà cung cấp dịch vụ tiện ích để thúc đẩy hiệu quả năng lượng dựa trên phân tích dữ liệu đồng hồ đo thông minh. Công ty này đã huy động được 14 triệu USD đầu tư mạo hiểm (VC) tài trợ trong năm 2008 và 50 triệu USD trong hai năm sau đó. Ba năm sau khi thành lập, Opower đã có hơn 230 nhân viên.

1.2.5. Hậu cần và giao thông vận tải

Ngành hậu cần và giao thông vận tải tuy sử dụng dữ liệu với cường độ thấp nhưng đang đối mặt với lưu lượng dữ liệu ngày càng tăng. Đây có thể là cơ hội để ngành này tăng hiệu quả vận chuyển hàng hoá và hành khách thông qua định tuyến đường thông minh và các dịch vụ mới dựa trên các ứng dụng thông minh.

Định tuyến thông minh dựa trên dữ liệu giao thông thời gian thực được sử dụng cũng như thu thập nhờ vào các hệ thống định vị. Một số hệ thống là các thiết bị phần cứng chuyên dụng, nhưng đa số là các hệ thống định vị cá nhân hoạt động như phần mềm chạy trên điện thoại thông minh hoặc tích hợp trong xe ô tô. Các ứng dụng này sử dụng dữ liệu với cường độ cao. Ví dụ, TomTom, hãng dẫn đầu về phần cứng và phần mềm định vị, vào năm 2012 trong các cơ sở dữ liệu của mình đã có hơn 5000 nghìn tỷ điểm dữ liệu từ thiết bị định vị của hãng và từ các nguồn khác, mô tả thời gian, vị trí, hướng và tốc độ của người dùng cá nhân ẩn danh, và họ bổ sung thêm 5 tỷ điểm dữ liệu mỗi ngày. Tổng thể theo ước tính của MGI (2011) cho thấy, các kho dữ liệu định vị địa lý cá nhân toàn cầu đạt ít nhất 1 petabyte vào năm 2009, và đang tăng khoảng 20% một năm. Đến năm 2020, kho dữ liệu này được dự báo sẽ cung cấp 500 tỷ USD trị giá trên toàn thế giới dưới hình thức tiết kiệm thời gian và nhiên liệu hay giảm được 380 triệu tấn phát thải CO_2. Con số này chưa bao gồm giá trị mang lại thông qua các dịch vụ định vị khác.

Cũng như các nhà cung cấp hệ thống định vị, các nhà vận hành khác cũng cung cấp những khối lượng dữ liệu lớn. Ví dụ, các nhà vận hành mạng di động sử dụng các tín hiệu di động tháp điện thoại để kiểm tra chéo vị trí của người sử dụng điện thoại di động và xác định các mẫu hình liên quan đến sự cố và ùn tắc dựa trên phân tích dữ liệu. Các dữ liệu và thông tin này được gợi ý bán cho các nhà cung cấp hệ thống định vị, và cho cả bên thứ ba như các chính phủ. Ví dụ, Orange - công ty dịch vụ viễn thông di động Pháp sử dụng công nghệ Floating Mobile Data (FMD) thu thập dữ liệu lưu thông điện thoại di động để xác định tốc độ và mật độ lưu lượng tại một điểm nhất định của mạng lưới đường bộ và suy ra thời gian đi lại hay sự hình thành ùn tắc giao thông. Các dữ liệu lưu lượng điện thoại di động ẩn danh được bán cho các bên thứ ba, bao gồm cả các cơ quan chính phủ, để xác định các điểm nóng cần can thiệp công cộng, và cho các công ty tư nhân như Mediamobile, nhà cung cấp hàng đầu các dịch vụ thông tin giao thông ở châu Âu.

Một lĩnh vực khác trong đó việc sử dụng dữ liệu có triển vọng mang lại lợi ích đáng kể cho ngành hậu cần và vận chuyển đó là việc sử dụng các ứng dụng thông minh dựa trên

giao tiếp máy-máy (M2M). Ví dụ ô tô thông minh có xu hướng được trang bị các bộ cảm biến để giám sát và truyền về hiện trạng các bộ phận của xe, cũng như môi trường xe đang chuyển động. Điều này cho phép các dịch vụ như OnStar và Sync, do các nhà sản xuất xe hơi cung cấp cho các chủ xe và bao gồm bảo vệ chống trộm, định vị và dịch vụ khẩn cấp. Các mô hình kinh doanh mới và các hình thức lệ phí và thuế mới, chẳng hạn như định giá đường đi động dựa trên dữ liệu GPS và M2M cũng đang cung cấp giá trị gia tăng đáng kể. MGI (2011) ước tính đến năm 2020, việc sử dụng dịch vụ thu phí tự động dựa trên vị trí của điện thoại di động sẽ tạo ra từ 4 đến 10 tỷ USD giá trị cho người tiêu dùng cuối cùng và 2 tỷ USD doanh thu cho các nhà cung cấp dịch vụ.

1.3. Đổi mới sáng tạo dựa trên dữ liệu - nguồn lực tăng trưởng và phát triển mới

1.3.1. Sự phát triển hệ sinh thái dữ liệu lớn

Công nghệ thông tin và truyền thông (ICT) trông cậy nhiều vào các khoản đầu tư vốn tri thức (knowledge-based capital - KBC). Điều này đặc biệt rõ nét trong cơ cấu tài sản của các công ty Internet như Google và Facebook, là nơi có tài sản vật chất chỉ chiếm khoảng 15% giá trị của các công ty tính vào thời điểm 31/12/2013. Các công ty Internet cũng đạt được năng suất rất cao nhờ vào các khoản đầu tư cho nguồn vốn KBC liên quan đến phần mềm và đặc biệt là dữ liệu. Tuy nhiên, so với các công ty ICT khác, cũng phụ thuộc nhiều vào các khoản đầu tư phần mềm và dữ liệu, các công ty Internet có năng suất cao hơn nhiều. Trong số 250 công ty ICT hàng đầu của OECD, các công ty Internet tạo ra trung bình gần 1 triệu USD doanh thu bình quân mỗi nhân viên trong năm 2011, trong khi các công ty ICT hàng đầu khác tạo ra trung bình từ 500.000 USD (doanh nghiệp phần mềm) đến 200.000 USD (doanh nghiệp dịch vụ IT).

Một phân tích về các mô hình kinh doanh cho thấy, các công ty Internet cùng có một điểm chung chủ yếu ngoài việc dựa trên Internet như là xương sống của hoạt động kinh doanh, đó là việc sử dụng "dữ liệu lớn" (OECD, 2012). Bằng cách thu thập và phân tích dữ liệu lớn, chủ yếu được cung cấp bởi người dùng Internet (tức là người tiêu dùng), các công ty Internet có thể tự động hóa các quy trình của mình và họ tiến hành thử nghiệm và thúc đẩy các sản phẩm và các mô hình kinh doanh mới với tốc độ nhanh hơn nhiều so với phần còn lại của ngành công nghiệp. Đặc biệt, việc sử dụng hiệu quả dữ liệu và phân tích cho phép các công ty Internet có thể mở rộng quy mô kinh doanh của mình với chi phí thấp hơn nhiều so với các công ty ICT khác, một hiện tượng đang tiến xa hơn so với điều mà Brynjolfsson et al. (2008) mô tả như là việc mở rộng quy mô không có khối lượng.

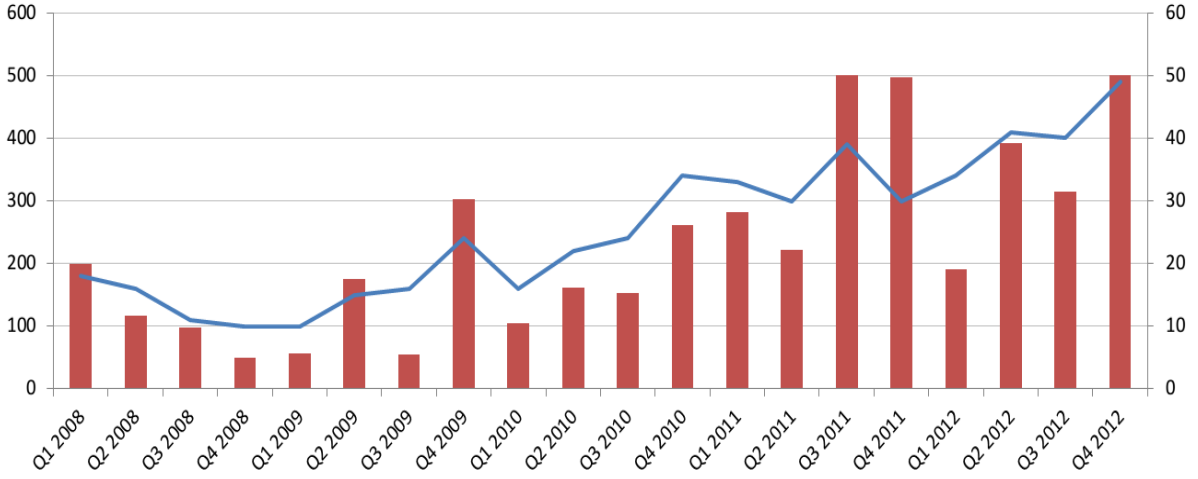
Các doanh nghiệp ICT khác đã bắt đầu nhận thức được "*dữ liệu lớn*" như một cơ hội kinh doanh mới và đang thực hiện những khoản đầu tư quan trọng để theo kịp và tham gia vào các hoạt động khai thác "*dữ liệu lớn*". Ước tính của IDC (2012) chỉ ra rằng, "*công nghệ và dịch vụ dữ liệu lớn*" sẽ tăng từ mức 3 tỷ USD năm 2010 lên 17 tỷ USD vào năm 2015, nghĩa là đạt tỷ lệ tăng trưởng tổng hợp lũy kế hàng năm (CAGR) gần 40%. Các công nghệ và dịch vụ liên quan đến lưu trữ được dự báo sẽ là phân khúc phát triển nhanh nhất, tiếp theo là kết nối mạng và dịch vụ, điều này giải thích vai trò ngày càng tăng của

công ty thiết bị IT trên thị trường tương đối mới này. Nhiều công ty ICT hàng đầu đang cố gắng củng cố vị trí trên thị trường của mình thông qua việc phát triển các sản phẩm "*dữ liệu lớn*" mới, chủ yếu dựa trên các giải pháp mã nguồn mở được phát triển ban đầu bởi các công ty Internet như trong trường hợp một công nghệ dữ liệu lớn quan trọng là Hadoop.

Nhưng các công ty ICT hàng đầu cũng ngày càng củng cố vị trí của mình thông qua việc mua lại các công ty mới khởi nghiệp chuyên môn hóa về công nghệ và dịch vụ dữ liệu lớn và/hoặc thông qua hợp tác với các đối thủ cạnh tranh tiềm năng trong các dự án mã nguồn mở như Hadoop. Dữ liệu do Orrick (2012) cung cấp về các giao dịch sáp nhập và mua lại (M&A) ở Hoa Kỳ cho thấy từ năm 2008 các hoạt động M&A đã tăng lên đáng kể về khối lượng và số lượng giao dịch (Hình 5). Theo Orrick (2012), IBM là hãng thâu tóm các công ty dữ liệu lớn mạnh nhất trong năm 2012, tiếp theo là Oracle.

Hình 5: Các hoạt động tài chính liên quan đến dữ liệu lớn, Q1/2008 - Q4/2012 (Đơn vị: triệu USD)

Khối lượng đầu tư (trái) và số giao dịch (phải)



Nguồn: OECD dựa trên Orrick (2012)

Kết quả là ngày càng có thêm nhiều doanh nghiệp bước vào thị trường dữ liệu lớn, cung cấp nhiều chủng loại công nghệ và dịch vụ để thu thập, tích hợp, lưu trữ, phân tích và trực quan hóa dữ liệu. Hiệu ứng tổng hợp của các hoạt động này đó là sự nổi lên một hệ sinh thái "dữ liệu lớn", trong đó hàng hóa và dịch vụ được phát triển phục vụ cho các ứng dụng dựa vào dữ liệu trong toàn bộ xã hội. Một phân tích về hệ sinh thái này cho thấy các loại hình đối tượng tham gia chủ yếu sau:

(1) Các nhà cung cấp dịch vụ Internet cung cấp mạng trục (backbone) của hệ sinh thái dữ liệu này;

(2) Các nhà cung cấp cơ sở hạ tầng IT mang đến các công cụ quản trị dữ liệu và các tài

nguyên tính toán quan trọng như các máy chủ lưu trữ dữ liệu, phần mềm quản trị cơ sở dữ liệu, và điện toán đám mây;

(3) Các nhà cung cấp phân tích dữ liệu, mang đến các giải pháp phần mềm cho phân tích dữ liệu, bao gồm cả trực quan hóa dữ liệu;

(4) Các nhà cung cấp dữ liệu, chủ yếu là người tiêu dùng, các chính phủ thông qua các xúc tiến dữ liệu mở, các doanh nghiệp như các nhà môi giới dữ liệu và thị trường dữ liệu và cả các chủ sở hữu các thiết bị và hệ thống kết nối (Internet vạn vật);

(5) Các doanh nghiệp dựa vào dữ liệu, những người tạo ra các hoạt động đổi mới sáng tạo dựa trên nguồn tài nguyên được cung cấp từ hệ sinh thái dữ liệu trong các lĩnh vực như bán lẻ, tài chính, quảng cáo, khoa học và y tế.

Mối tương tác giữa các thành phần tham gia này thông qua các lớp như được minh họa ở hình 6, trong đó các lớp phía dưới cung cấp hàng hóa và dịch vụ cho các lớp trên. Ví dụ, các nhà kinh doanh dựa trên dữ liệu dựa vào khả năng truy cập vào dữ liệu và các công cụ phân tích cũng như các cơ sở hạ tầng IT như điện toán đám mây để cung cấp các dịch vụ đổi mới sáng tạo của mình.

Hình 6: Hệ sinh thái dữ liệu lớn gồm các lớp người tham gia chính

<table>
<tr><td colspan="2">Các nhà kinh doanh dựa trên dữ liệu và các nhà đổi mới sáng tạo trong xã hội (các công ty mới khởi sự, doanh nhân cộng đồng)</td></tr>
<tr><td>Các nhà cung cấp phân tích dữ liệu (các giải pháp phần mềm phân tích)</td><td>Các nhà cung cấp dữ liệu (người tiêu dùng, chính phủ, môi giới dữ liệu và IoT)</td></tr>
<tr><td colspan="2">Các nhà cung cấp cơ sở hạ tầng IT (công cụ quản trị cơ sở dữ liệu, điện toán đám mây)</td></tr>
<tr><td colspan="2">Các nhà cung cấp dịch vụ Internet (băng thông rộng cố định và di động)</td></tr>
</table>

Hệ sinh thái dữ liệu còn có một thuộc tính quan trọng đó là bản chất toàn cầu vốn có của nó. Hệ sinh thái dữ liệu lớn liên quan đến các luồng dữ liệu xuyên biên giới do bản chất toàn cầu của các thành phần tham gia trong đó và do sự phân bố toàn cầu của các công nghệ và các nguồn lực được sử dụng để tạo ra giá trị. Ví dụ, dữ liệu có thể thu thập từ người tiêu dùng hay các thiết bị đặt ở một nước với các thiết bị và ứng dụng được phát triển ở một nước khác. Sau đó dữ liệu có thể được xử lý ở một nước thứ ba và dùng để cải tiến hoạt động marketing cho người tiêu dùng ở nước đầu tiên và/hoặc người tiêu dùng khác trên toàn cầu. Ngoài ra, các cơ sở hạ tầng ICT thường được sử dụng để thực hiện phân tích dữ liệu bao gồm các trung tâm dữ liệu và phần mềm hiếm khi chỉ được cung cấp trong vòng ranh giới một nước, thực tế chúng được phân phối trên toàn cầu để tận dụng các biến thể của nhiều yếu tố bao gồm, khối lượng công việc địa phương, môi trường (nhiệt độ và ánh nắng mặt trời), và cung ứng kỹ năng và lao động (và chi phí). Ví dụ như công ty Kaggle chuyên cung cấp nền tảng nguồn lực đám đông (crowd-sourcing) để dựa vào đó các chính phủ, doanh nghiệp và cá nhân trên toàn thế giới gửi (post) dữ liệu của họ

và để cho những người khác cạnh tranh tạo ra các kết quả phân tích tốt nhất (Rao, 2011). Bên cạnh đó, nhiều dịch vụ dựa vào dữ liệu được phát triển bởi các nhà kinh doanh có khả năng tận dụng được những tài nguyên có sẵn của các doanh nghiệp lớn, họ làm cho các dịch vụ đổi mới của mình (bao gồm cả dữ liệu của họ) trở nên khả dụng thông qua các giao diện lập trình ứng dụng (API), nhiều trong số đó được đặt tại nước ngoài. Ví dụ, Ushahidi, một công ty phần mềm phi lợi nhuận có trụ sở tại Nairobi, Kenya, cung cấp các dịch vụ thu thập dữ liệu, trực quan hóa và đồ họa tương tác dựa trên các API sẵn có của các công ty Internet như Google và Twitter.

1.3.2. Xu hướng đổi mới sáng tạo dựa trên dữ liệu

Việc sử dụng dữ liệu để tạo ra giá trị không chỉ giới hạn ở các công ty ICT, mặc dù có bằng chứng mạnh mẽ cho thấy rằng các công ty ICT vẫn đang dẫn đầu trong sử dụng phân tích dữ liệu tiên tiến. Theo Tambe (2014), chỉ có 30% số đầu tư vào công nghệ của Hadoop có nguồn gốc từ khu vực ngoài ICT, trong đó đặc biệt phải kể đến các doanh nghiệp thuộc lĩnh vực tài chính, giao thông vận tải, tiện ích, bán lẻ, y tế, dược phẩm và các công ty công nghệ sinh học. Tuy nhiên, mối quan tâm đến dữ liệu lớn từ các doanh nghiệp nằm ngoài lĩnh vực ICT trên toàn bộ nền kinh tế đang gia tăng nhanh, các công nghệ và dịch vụ khai thác dữ liệu được coi như một nguồn lực quan trọng để tạo ra giá trị và thúc đẩy đổi mới sáng tạo hay cải tiến các sản phẩm, quy trình, và thị trường hiện tại (tức là đổi mới sáng tạo dựa trên dữ liệu).

Nhiều tổ chức nói chung đã được hưởng lợi từ đầu tư vào dữ liệu dưới dạng các cơ sở dữ liệu truyền thống phục vụ cho đổi mới sáng tạo. Riêng thị trường của các hệ thống quản trị cơ sở dữ liệu quan hệ đã có trị giá hơn 21 tỷ USD trong năm 2011, tăng trưởng trung bình 8% một năm kể từ năm 2002 (OECD, 2013). Theo số liệu thống kê, các khoản đầu tư vào phần mềm và dữ liệu (của toàn bộ nền kinh tế) có tỷ trọng trung bình gần bằng 2% giá trị gia tăng của khu vực doanh nghiệp tại các nước OECD, các doanh nghiệp tại các nước như Đan Mạch (4%), Thụy Điển (3%), Vương quốc Anh (2%) và Hoa Kỳ (2%) dẫn đầu về tỷ trọng đầu tư so với giá trị gia tăng khu vực doanh nghiệp. Các quốc gia này (trừ Thụy Điển) cũng cho thấy một sự gia tăng mạnh về đầu tư phần mềm và dữ liệu trong thời kỳ khủng hoảng. Mặc dù số liệu thống kê chính thức cung cấp một bằng chứng mạnh mẽ về vai trò ngày càng tăng của phần mềm và dữ liệu, tuy nhiên chúng không phản ánh đầy đủ sự đóng góp ngày càng tăng của dữ liệu đối với tăng trưởng kinh tế. Điều đó không chỉ do các số liệu thống kê chính thức về dữ liệu vẫn còn quá ít, mà còn do hầu hết các lợi ích liên quan đến việc sử dụng dữ liệu vẫn chưa được các giao dịch thị trường nắm bắt (Mandel 2012; 2013).

Đổi mới sáng tạo dựa trên dữ liệu phục vụ tăng trưởng

Việc khai thác dữ liệu và phân tích có thể tạo ra giá trị gia tăng quan trọng thông qua đổi mới dựa trên dữ liệu liên quan đến một loạt các hoạt động, từ tối ưu hóa chuỗi giá trị và dây chuyền sản xuất đến sử dụng hiệu quả hơn các nguồn lực, các mối quan hệ khách hàng tốt hơn, và phát triển các thị trường mới. Trong nhiều lĩnh vực, đổi mới dựa trên dữ

liệu có khả năng phá vỡ và thách thức vị trí thống trị hiện tại của các thị trường truyền thống. Trong giao thông vận tải, sự gia tăng khả năng định vị của các thiết bị di động đã tạo khả năng cho một loạt các dịch vụ định vị mới, trong đó có các dịch vụ hậu cần và định vị cá nhân. TomTom - nhà cung cấp phần cứng và phần mềm định vị hàng đầu, hiện nay có hơn 9 nghìn tỷ điểm thu thập dữ liệu từ các thiết bị định vị và các nguồn khác, mô tả thời gian, vị trí, hướng đi và tốc độ di chuyển của người dùng cá nhân ẩn danh, và hãng này giờ đây đang bổ sung thêm 6 tỷ điểm đo lường mỗi ngày. Các kết quả phân tích dữ liệu được phản hồi trở lại các thiết bị định vị để thông báo cho lái xe về tình hình hiện tại và dự đoán về giao thông. Điều này có thể giúp tiết kiệm thời gian và giảm ùn tắc giao thông, đặc biệt là ở các thành phố. Về tổng thể, các ước tính chỉ ra rằng vùng lưu dữ liệu định vị địa lý cá nhân toàn cầu tăng trưởng với tỷ lệ 20% một năm kể từ năm 2009. Đến năm 2020, vùng dữ liệu này có thể mang lại 500 tỷ USD giá trị trên toàn thế giới dưới các hình thức tiết kiệm thời gian và nhiên liệu, hay làm giảm được 380 megatonnes (triệu tấn) khí thải CO_2 theo như ước tính của MGI (2011).

Ngay cả các lĩnh vực truyền thống như bán lẻ, thể thao, giày dép, chế tạo cũng đang bị phân đoạn thông qua việc sử dụng dữ liệu và phân tích, và một số trường hợp ngày càng phát triển theo hướng dịch vụ hơn, một xu hướng đã được nhiều tài liệu mô tả bằng từ "*servicification*" (dịch vụ hóa) (Lodefalk, 2010). Các công ty như Tesco, chuỗi siêu thị của Anh, khai thác những luồng dữ liệu lớn được tạo ra từ các chương trình thẻ khách hàng thân thiết của họ. Các chương trình này của Tesco đến nay, đếm được hơn 100 rổ thị trường trong một giây và 6 triệu giao dịch mỗi ngày, đã giúp Tesco phát triển từ một hãng bán lẻ hàng giá rẻ địa phương thành công ty thương mại quốc gia, định hướng khách hàng và dịch vụ có sức hấp dẫn rộng rãi trong các nhóm xã hội. Các công ty bán lẻ như Walmart thậm chí còn tiến bộ hơn trong việc sử dụng dữ liệu và phân tích. Công ty này đã phát triển các dịch vụ dữ liệu phân tích riêng của mình thông qua công ty con Walmart Labs, một tổ chức đang tích cực đóng góp cho sự (đồng) phát triển phân tích mã nguồn mở. Ví dụ như giải pháp (nội bộ) của Walmart Labs mang tên Social Genome đã cho phép Walmart có thể tiếp cận tới các khách hàng tiềm năng, bao gồm cả bạn bè của khách hàng trực tiếp, người đã từng đề cập đến các sản phẩm cụ thể trên mạng, và hãng này đã cung cấp giảm giá các sản phẩm đó. Social Genome được xây dựng dựa trên dữ liệu công cộng từ web (bao gồm cả dữ liệu truyền thông xã hội) cũng như từ dữ liệu độc quyền của Walmart như mua thông tin liên lạc và thông tin mua sắm của khách hàng. "*Điều này đã dẫn đến một cơ sở kiến thức rộng lớn, luôn thay đổi, liên tục cập nhật với hàng trăm triệu thực thể và các mối quan hệ*" (Big Data Startups, 2013).

Trong lĩnh vực chế tạo, các công ty ngày càng sử dụng nhiều các bộ cảm biến gắn trên máy móc sản xuất và các sản phẩm phân phối để thu thập và xử lý dữ liệu về hoạt động của máy móc và sản phẩm. Xu hướng này được tạo khả năng nhờ vào giao tiếp máy nối máy (M2M) và phân tích các dữ liệu cảm biến, đã từng được mô tả như "*Internet công*

nghiệp" (Bruner, 2013) hay "*mạng chế tạo*" (network manufacturing). Dữ liệu cảm biến được sử dụng để theo dõi và phân tích hiệu quả của sản phẩm, để tối ưu hóa hoạt động ở mức độ toàn bộ hệ thống, và để phục vụ cho dịch vụ sau bán hàng bao gồm các hoạt động bảo trì phòng ngừa. Các dữ liệu tiếp theo còn được sử dụng để phân tích và dự báo về các thành phần linh kiện có khả năng dễ bị ảnh hưởng và các kết quả có thể sử dụng để tối ưu hóa thiết kế sản phẩm và kiểm soát sản xuất. Điều này cũng có thể bao gồm cả thiết kế sản phẩm và điều khiển sản xuất của các nhà cung ứng, trong trường hợp đó những hiểu biết về phân tích dữ liệu được hợp tác chia sẻ với các nhà cung ứng và trong một số trường hợp thậm chí có thể thương mại hóa như một phần của dịch vụ mới cho các nhà cung ứng và khách hàng tiềm năng. Ví dụ, hãng Schmitz Cargobull của Đức, nhà sản xuất thùng xe tải và toa móc lớn nhất thế giới đã sử dụng M2M và dữ liệu cảm biến để giám sát việc bảo trì, điều kiện di chuyển và các tuyến di chuyển của các xe kéo do hãng sản xuất (Chick et al 2014). Các kiến thức được tạo ra từ phân tích dữ liệu được sử dụng để giúp khách hàng của Schmitz Cargobull giảm thiểu được những sự cố hỏng hóc khi sử dụng. Các dịch vụ tương tự được quan sát thấy trong lĩnh vực thiết bị sản xuất năng lượng, nơi M2M và dữ liệu cảm biến được sử dụng để tối ưu hóa các điều kiện ngẫu nhiên trong các hoạt động lập kế hoạch dự án phức tạp. Bằng chứng định lượng về tác động kinh tế tổng thể của dữ liệu và phân tích trong lĩnh vực chế tạo vẫn còn hiếm, nhưng những ước tính có thể thực hiện được ví dụ như đối với Nhật Bản cho thấy việc sử dụng và phân tích các dạng dữ liệu của các công ty chế tạo Nhật Bản có thể mang lại tiết kiệm chi phí bảo trì trị giá gần 5 nghìn tỷ yên (tương đương với việc vận chuyển được nhiều hơn 15% lô hàng) vào năm 2010 và hơn 50 tỷ yên nhờ tiết kiệm điện (MIC, 2013).

Việc sử dụng và phân tích dữ liệu còn có thể tạo khả năng dịch vụ hóa trong các lĩnh vực công nghệ thấp như dệt may (bao gồm cả thể thao và công nghiệp da giày) và nông nghiệp. Nike - hãng sản xuất giày và dụng cụ thể thao của Hoa Kỳ đã thiết kế lại nhiều sản phẩm của mình thành các dịch vụ dựa trên dữ liệu, chúng được tích hợp thông qua nền tảng trực tuyến Nike+. Dữ liệu được thu thập thông qua các cảm biến Nike+ được gắn trên giày điền kinh, hoặc gần đây hơn họ thông qua FuelBand, miếng da bao cổ tay theo dõi các hoạt động và lượng calo tiêu hao trong ngày. Mặc dù nhiệm vụ cốt lõi của hãng là hỗ trợ mọi người hoạt động thể chất và khỏe mạnh không thay đổi, nhưng Nike giờ đây ngày càng cung cấp nhiệm vụ này như một dịch vụ thông qua sử dụng dữ liệu, cho phép người dùng thiết lập các mục tiêu của họ, theo dõi sự tiến bộ của mình và cung cấp cả các yếu tố xã hội có thể làm gián đoạn thị trường đối với các huấn luyện viên cá nhân. Một chiến lược đổi mới dựa trên dữ liệu tương tự có thể quan sát được giữa các đối thủ cạnh tranh như công ty đồ dùng và giày dép thể tao Adidas của Đức, đã đưa ra dịch vụ dựa trên dữ liệu của mình mang tên miCoach cũng để xâm nhập và chi phối thị trường huấn luyện viên cá nhân.

Ngành nông nghiệp hiện tại cũng đang trở nên hiện đại hóa hơn nhờ vào đổi mới dựa

trên dữ liệu, đôi khi được thể hiện qua thuật ngữ "nông nghiệp chính xác", lĩnh vực này đang ngày càng được khai thác để nâng cao năng suất và giảm tác động môi trường, xây dựng các bản đồ mã hóa địa lý các vùng nông nghiệp và giám sát trong thời gian thực mọi hoạt động từ gieo hạt, tưới nước, phân bón và thu hoạch. Kết quả là, nông dân ngày nay có một khối lượng phong phú các dữ liệu thông tin nông nghiệp, trong đó các công ty như Monsanto, John Deere và DuPont Pioneer đang cố gắng khai thác thông qua hàng hóa và dịch vụ mới dựa trên dữ liệu. Ví dụ hãng John Deere đã tận dụng "Internet công nghiệp" như các cảm biến tích hợp vào các thiết bị mới nhất của mình "để giúp nông dân quản lý xe cộ của họ và giảm thời gian chết của máy kéo cũng như tiết kiệm nhiên liệu" (Big Data Startup, 2013). Các dữ liệu cảm biến tương tự sau đó có thể được liên kết với dữ liệu lịch sử và theo thời gian thực, ví dụ như về dự báo thời tiết, điều kiện đất đai, sử dụng phân bón, và đặc điểm cây trồng để tối ưu hóa và dự báo sản xuất nông nghiệp. Một số kết quả phân tích dữ liệu được cung cấp cho nông dân thông qua nền tảng MyJohnDeere.com (và các ứng dụng liên quan của nó) cho phép nông dân có thể tối ưu hóa sự lựa chọn cây trồng, nơi trồng và thời gian cày xới (Big Data Startup, 2013). Nhìn chung, việc sử dụng dữ liệu và phân tích được một số chuyên gia ước tính có thể nâng cao được sản lượng từ 5-10 giạ bình quân mỗi mẫu hay tăng được lợi nhuận khoảng 100 USD/mẫu (Noyes, 2014). Sự gia tăng năng suất này diễn ra vào đúng thời điểm khi OECD và FAO (Tổ chức nông lương Liên hợp quốc) (OECD và FAO, 2012) kêu gọi tăng 60% sản lượng lương thực thế giới để có thể nuôi sống dân số ngày càng gia tăng, được dự kiến sẽ đạt 9 tỷ người vào năm 2050.

Trong khi các bằng chứng được nêu ở trên cho thấy một mối liên kết tích cực giữa đổi mới dựa trên dữ liệu và tăng trưởng năng suất trên toàn bộ nền kinh tế, một vài nghiên cứu thực nghiệm đã chỉ ra những ước tính định lượng thuyết phục. Ở cấp độ doanh nghiệp, nghiên cứu của Brynjolfsson et al. (2011) về 330 công ty ở Hoa Kỳ cho thấy rằng, sản lượng và năng suất của các công ty áp dụng việc ra quyết định dựa trên dữ liệu tăng 5% đến 6% cao hơn so với các khoản đầu tư khác và so với sử dụng ICT. Các doanh nghiệp này cũng hoạt động tốt về khía cạnh sử dụng tài sản, thu lợi về vốn cổ phần và giá trị thị trường. Một nghiên cứu tương tự của Bakhshi et al. (2014) dựa trên 500 công ty tại Vương quốc Anh phát hiện ra rằng, những doanh nghiệp sử dụng nhiều dữ liệu trực tuyến về khách hàng và người tiêu dùng có năng suất cao hơn từ 8% đến 13%. Việc "sử dụng phân tích dữ liệu" và "báo cáo về những hiểu biết dựa vào dữ liệu" có mối liên kết mạnh với tăng năng suất. Một nghiên cứu gần đây của Tambe (2014) dựa trên phân tích hồ sơ của 175 triệu người sử dụng LinkedIn, bao gồm cả những người có kỹ năng về công nghệ "*dữ liệu lớn*" chỉ ra rằng, đầu tư của doanh nghiệp vào các công nghệ "*dữ liệu lớn*" có liên quan đến gia tăng năng suất 3%, nhưng điều này chỉ đúng với các doanh nghiệp: (i) đã tiếp cận được đến các tập hợp dữ liệu quan trọng; và (ii) đã kết nối với các mạng lưới lao động có trình độ chuyên môn về các công nghệ "*dữ liệu lớn*" cụ thể.

Việc sử dụng phân tích dữ liệu của doanh nghiệp phụ thuộc chủ yếu vào chủng loại tập hợp dữ liệu. Dữ liệu hoạt động kinh doanh và dữ liệu điểm bán hàng thường phụ thuộc vào phân tích dữ liệu, trong khi dữ liệu trực tuyến bao gồm cả dữ liệu truyền thông xã hội và dữ liệu nhấp chuột ít khi được sử dụng trong các doanh nghiệp. Theo một khảo sát của Economist Intelligence Unit (2012) lấy ý kiến của hơn 600 giám đốc điều hành kinh doanh trên thế giới, hai phần ba người trả lời nói rằng, việc thu thập và phân tích các dữ liệu là nền tảng cho chiến lược kinh doanh công ty của họ và cho cả việc ra quyết định hàng ngày. Người được hỏi đặc biệt đánh giá cao "*dữ liệu hoạt động kinh doanh*", coi đó là các tập hợp dữ liệu có giá trị nhất, trong khi lĩnh vực bán lẻ lại cho rằng dữ liệu về điểm bán hàng có ý nghĩa quan trọng. Trong số các công ty được khảo sát ở Vương quốc Anh, chỉ có 18% được xác định là người sử dụng dữ liệu trực tuyến tinh xảo. Một nghiên cứu do IRISGROUP thực hiện ở Đan Mạch (2013) xác nhận số lượng "*người tiên phong*" còn hạn chế. Một nghiên cứu Gartner năm 2013 cho thấy *"dữ liệu lớn*" là một động lực đang tăng ở các doanh nghiệp tại Hoa Kỳ: vào năm 2012 có 58% doanh nghiệp được hỏi cho biết họ đã triển khai hoặc lên kế hoạch về các dự án "*dữ liệu lớn*", đến năm 2013 con số này đã lên đến 64% doanh nghiệp (Gartner, năm 2013).

Một phân tích các hoạt động làm việc theo nghề và ngành cho thấy các dịch vụ hành chính công cũng như giáo dục và y tế có khả năng là hai lĩnh vực trong đó việc áp dụng phân tích dữ liệu có thể gây tác động lớn nhất trong giai đoạn tương đối ngắn. Các lĩnh vực này sử dụng tỷ lệ lớn nhất các công việc thực hiện nhiều nhiệm vụ liên quan đến việc thu thập và phân tích thông tin và đang ngày càng trở nên thâm dụng dữ liệu. Tuy nhiên, các công việc vẫn còn được thực hiện ở mức tin học hóa tương đối chậm. Việc triển khai phân tích dữ liệu nhằm mục tiêu do đó có thể làm tăng thêm được hiệu quả trong các lĩnh vực này.

Trong trường hợp khu vực công (không kể tình báo và an ninh), một số bằng chứng cho thấy việc sử dụng dưới mức các dữ liệu được tạo ra và thu thập (MGI, 2011; OECD, 2012). Theo MGI (2011), việc sử dụng đầy đủ phân tích dữ liệu tại 23 chính phủ lớn nhất thuộc châu Âu có thể giảm được từ 15% đến 20% chi phí hành chính, tương đương với việc tạo ra 150 tỷ đến 300 tỷ euro giá trị mới, và thúc đẩy năng suất tăng nhanh hơn 0,5% mỗi năm trong vòng mười năm tới.

Đổi mới dựa trên dữ liệu cải thiện phúc lợi xã hội

Các nghiên cứu đã chỉ ra bản chất có tính phá hủy của đổi mới dựa trên dữ liệu và những tác động tích cực của nó đối với tăng năng suất. Tuy nhiên, như đã nhấn mạnh, các nghiên cứu vẫn chưa phản ánh sự đóng góp đầy đủ của đổi mới dựa trên dữ liệu đối với phúc lợi, do có nhiều tác động xã hội liên quan đến việc sử dụng dữ liệu và phân tích rất khó hoặc không thể đo đạc được. Việc tạo điều kiện cho các công dân sử dụng dữ liệu mở do các chính phủ cung cấp thông qua các xúc tiến dữ liệu mở có thể làm tăng sự mở cửa, tính minh bạch và trách nhiệm giải trình của các hoạt động chính phủ và do đó làm tăng

niềm tin của công chúng vào chính phủ. Đồng thời, nó có thể tạo khả năng cho một phạm vi không giới hạn các dịch vụ thương mại và xã hội trong toàn xã hội. Ví dụ, doanh nghiệp cộng đồng ngày càng sử dụng dữ liệu mở có sẵn theo như khuyến cáo của OECD (2008) về tăng cường tiếp cận và sử dụng hiệu quả hơn thông tin khu vực công, kết hợp với các nguồn dữ liệu công bố công khai khác để phát triển các ứng dụng tạo điều kiện tiếp cận đến các dịch vụ công hiện tại và cung cấp các dịch vụ bổ sung mới trên toàn xã hội. Ví dụ, như CitiVox. Ước tính về tác động kinh tế của PSI (509 tỷ euro năm 2008 đối với sử dụng lại PSI trong OECD) tập trung vào việc tái sử dụng thương mại của PSI vì thế không bao quát được đầy đủ các lợi ích xã hội.

Một ví dụ khác là có nhiều nguồn dữ liệu bao gồm điện thoại di động và các trang web trong đó truyền thông xã hội đang được khai thác và sử dụng để cải thiện phúc lợi của các cá nhân ở các nước đang phát triển. Thông qua các sáng kiến quốc tế, như Paris21, Hợp tác Thống kê phục vụ phát triển trong thế kỷ 21, và *Global Pulse* của Liên Hợp Quốc (UN) một sáng kiến được Văn phòng điều hành của Tổng thư ký Liên Hợp Quốc khởi xướng nhằm đáp ứng yêu cầu cần có dữ liệu kịp thời hơn để theo dõi và giám sát tác động của các cuộc khủng hoảng kinh tế-xã hội toàn cầu và địa phương (United Nations, 2012). Một ví dụ khác là tổ chức phi lợi nhuận Ushaidi có trụ sở tại Kenya đã tạo ra phần mềm để thu thập và hiển thị dữ liệu dùng để phân tích và trực quan hóa ví dụ như báo cáo của các nhân chứng về bạo lực thông qua email và tin nhắn văn bản hoặc các thông báo về sự sẵn có của các loại thuốc quan trọng được cung cấp thông qua viện trợ nhân đạo ở các nước đang phát triển trên toàn thế giới, những lợi ích này đã không được phản ánh trong các thống kê kinh tế.

Trong lĩnh vực khoa học, sự ra đời của các công cụ và phương pháp nghiên cứu mới sử dụng dữ liệu cường độ cao đã dẫn đến một lĩnh vực mới được gọi là "*khám phá khoa học thâm dụng dữ liệu*" (“data-intensive scientific discovery), được xây dựng dựa trên sử dụng các phương pháp mô tả thực nghiệm, các mô hình lý thuyết và các mô phỏng hiện tượng phức tạp (BIAC, 2011). Các công cụ mới như máy siêu gia tốc hay các kính viễn vọng, và cả Internet như một công cụ thu thập dữ liệu, đã được sử dụng làm công cụ trong những phát triển mới về khoa học, chúng làm thay đổi cả quy mô và độ chi tiết của các dữ liệu đang được thu thập. Ví dụ, dự án Trạm quan sát bầu trời bằng kỹ thuật số (Digital Sky Survey) bắt đầu vào năm 2000, trong tuần lễ đầu tiên đã thu thập được một lượng dữ liệu còn lớn hơn số dữ liệu đã tích lũy được trong suốt lịch sử thiên văn (The Economist, 2010), và kính viễn vọng vô tuyến mới lớn nhất thế giới (SKA) này có thể tạo ra được 1 petabyte dữ liệu cứ mỗi 20 giây (EC, 2010). Hơn nữa, khả năng ngày càng tăng của phân tích dữ liệu đã làm cho nó có thể rút ra được sự hiểu biết sâu từ những tập hợp dữ liệu rất lớn một cách nhanh chóng. Ví dụ như trong lĩnh vực di truyền, máy lập trình tự gen ADN dựa trên phân tích dữ liệu lớn giờ đây có thể đọc được khoảng 26 tỷ ký tự về mã di truyền của người chỉ trong vài giây. Điều này diễn ra kèm với sự giảm chi phí đáng kể trong lập

trình tự ADN trong vòng 5 năm gần đây.

Những phát triển gần đây về khoa học rõ ràng đã có tác động đáng kể đến nghiên cứu trong lĩnh vực chăm sóc sức khỏe, là nơi có những thay đổi về dân số theo hướng xã hội già hóa và chi phí y tế tăng cao, đang đòi hỏi các dịch vụ nhằm vào bệnh nhân với hiệu quả cao hơn và khả năng đáp ứng nhanh hơn. Trọng tâm của đổi mới dựa trên dữ liệu trong lĩnh vực y tế là dữ liệu y học quốc gia bao gồm, nhưng không giới hạn, các hồ sơ y tế điện tử, các dữ liệu chụp hình hệ thần kinh và dữ liệu dịch tễ học. Việc sử dụng lại một cách có hiệu quả các tập hợp dữ liệu này có triển vọng nâng cao được hiệu quả và chất lượng chăm sóc sức khỏe. Ví dụ như ở Phần Lan, nội dung, chất lượng và hiệu quả chi phí điều trị đối với một số căn bệnh chọn lọc đang được phân tích bằng cách liên kết dữ liệu bệnh nhân trên toàn bộ trình tự chăm sóc từ khi nhập viện, đến sự chăm sóc của bác sĩ cộng đồng, đến các loại thuốc được kê đơn và tử vong (OECD, 2013d). Các kết quả của sự phân tích này được công bố rộng rãi và được trao cho bệnh nhân, dẫn đến nâng cao chất lượng tại các bệnh viện ở Phần Lan.

Trong khi dữ liệu nghiên cứu và y học truyền thống đóng một vai trò quan trọng đối với đổi mới dựa trên dữ liệu trong khoa học và nghiên cứu y học, thì các nguồn dữ liệu mới cũng được các nhà nghiên cứu cũng như các cá nhân tìm kiếm, nhằm cải tiến nghiên cứu và điều trị bệnh tật cũng như để tận dụng lợi thế của đổi mới dựa trên dữ liệu để phòng bệnh và chăm sóc sức khỏe tốt hơn cho chính bản thân. Ví dụ, mạng xã hội PatientsLikeMe không chỉ cho phép những người có vấn đề về sức khỏe tương tác với nhau, để có được sự an ủi và học hỏi từ những người khác mắc cùng chứng bệnh, mạng này còn cung cấp một cơ sở bằng chứng về dữ liệu cá nhân cho phân tích và tạo nền tảng kết nối bệnh nhân với các thử nghiệm lâm sàng. Một ví dụ khác, được gọi là phong trào tự phát Quantified Self-movement lấy cảm hứng từ những người tham gia để sử dụng tất cả các loại công cụ, như Fitbit, nhằm theo dõi từng động thái và nhịp đập của tim, và tạo khả năng cho các cá nhân có thể nâng cao sức khoẻ và sự an toàn nói chung.

II. CÁC CÔNG NGHỆ VÀ CHÍNH SÁCH THÚC ĐẨY ĐỔI MỚI SÁNG TẠO DỰA TRÊN DỮ LIỆU

2.1. Các kênh khai thác đổi mới sáng tạo dựa trên dữ liệu để phục vụ tăng trưởng kinh tế

Đổi mới sáng tạo dựa trên dữ liệu đóng góp cho tăng trưởng kinh tế thông qua hai "kênh" khác biệt trong suốt chu trình giá trị dữ liệu, đó là: đặc tính kinh tế của dữ liệu như một nguồn lực cơ sở hạ tầng; và các cơ chế tạo ra giá trị của việc sử dụng và phân tích dữ liệu. Dữ liệu được xem xét như một cơ sở hạ tầng, một khái niệm khá phổ biến trong số các nhà kinh tế và nhà hoạch định chính sách, đặc biệt là đối với những người hoạt động trong lĩnh vực cơ sở hạ tầng viễn thông và mạng Internet. Quan niệm như vậy sẽ giúp hiểu

được đặc điểm kinh tế của dữ liệu, làm rõ thêm đặc trưng của đổi mới dựa trên dữ liệu, và đặc biệt làm tăng lợi nhuận theo quy mô và phạm vi có thể đến cùng với việc sử dụng dữ liệu.

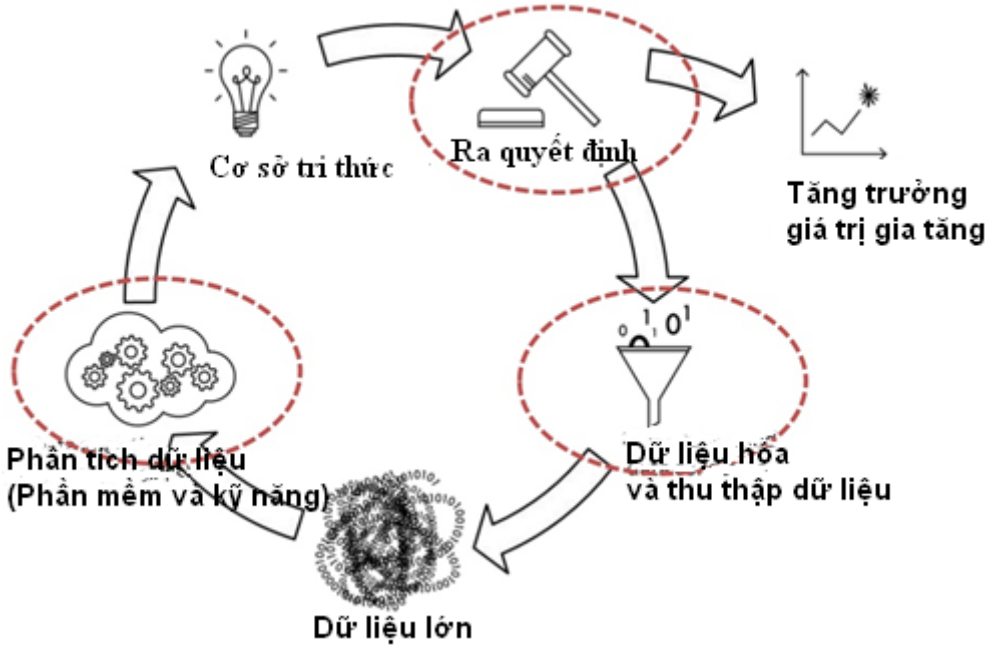
2.1.1. Vòng đời giá trị dữ liệu*: từ dữ liệu hóa (datafication) đến phân tích dữ liệu và ra quyết định*

Đổi mới sáng tạo dựa vào dữ liệu được mô tả rõ nhất bằng một quá trình có tính đến các giai đoạn khác nhau mà qua đó dữ liệu được chuyển hóa để cuối cùng dẫn đến sự đổi mới sáng tạo. Hình 7 minh họa một chu trình giá trị dữ liệu dựa trên nhận thức rằng, sự đổi mới dựa trên dữ liệu không phải là một quá trình tuyến tính, do đó không thể mô tả đầy đủ bằng chuỗi giá trị đơn giản. Ngược lại, đổi mới dựa vào dữ liệu có liên quan đến các vòng lặp phản hồi tại một số giai đoạn trong quá trình tạo ra giá trị. Các giai đoạn được xác định bao gồm:

- Dữ liệu hóa và thu thập dữ liệu là hoạt động tạo ra dữ liệu thông qua việc số hóa truyền thông, giám sát các hoạt động bao gồm cả các hoạt động thế giới thực (ngoại tuyến) và các hiện tượng thông qua các bộ cảm biến.
- Dữ liệu lớn là kết quả của quá trình dữ liệu hóa và thu thập dữ liệu dẫn đến một vùng dữ liệu lớn có thể khai thác thông qua phân tích dữ liệu. Dữ liệu trong trạng thái này thường không có ý nghĩa cố hữu, không có bất kỳ cấu trúc nào hay mối quan hệ trong bản thân nó.
- *Phân tích dữ liệu*: Cho đến khi được xử lý và diễn giải thông qua phân tích dữ liệu, dữ liệu lớn có thể không có giá trị, bởi vì ban đầu không có thông tin rõ ràng. Phân tích dữ liệu là một tập hợp các kỹ thuật và công cụ phần mềm được sử dụng để trích xuất thông tin từ dữ liệu. Theo OECD (2012), giá trị của dữ liệu có tính phụ thuộc cao vào bối cảnh và dựa vào cách dữ liệu liên kết với các bộ dữ liệu khác như thế nào, đó chính là mục đích mà phân tích dữ liệu hướng đến. Cuối cùng, phân tích dữ liệu ngày càng được thực hiện thông qua điện toán đám mây.
- *Cơ sở tri thức*: là những kiến thức mà các cá nhân hoặc các hệ thống (bao gồm cả các tổ chức) tích lũy được thông qua phân tích dữ liệu theo thời gian. Nó có đặc trưng nội hàm trong con người khi đạt được những hiểu biết sâu (học tập). Tuy nhiên, nó cũng có thể hàm chứa trong các sản phẩm hữu hình và vô hình, bao gồm sách, thủ tục chuẩn, và vốn tri thức như bằng sáng chế, thiết kế và phần mềm. Nơi có học máy tham gia, cơ sở tri thức phản ánh hiện trạng của hệ thống học tập. Cơ sở tri thức là những "viên ngọc quý" của tổ chức dựa trên dữ liệu, và do đó được hưởng sự bảo vệ đặc biệt thông qua pháp luật (ví dụ như bí mật thương mại) và các phương tiện kỹ thuật.
- *Ra quyết định dựa trên dữ liệu*: Giá trị kinh tế và xã hội của dữ liệu chủ yếu gặt hái được tại hai thời điểm: thứ nhất khi dữ liệu được chuyển hóa thành tri thức (có

được hiểu biết sâu) và sau đó khi nó được sử dụng cho việc ra quyết định (hành động). Giai đoạn ra quyết định là quan trọng nhất đối với các doanh nghiệp. Theo một khảo sát của Economist Intelligence Unit (2012), gần như 60% các nhà lãnh đạo doanh nghiệp sử dụng "*dữ liệu lớn*" để hỗ trợ ra quyết định và gần 30% sử dụng để tự động ra quyết định.

Hình 7: Vòng đời giá trị dữ liệu



2.1.2. Dữ liệu là nguồn lực cơ sở hạ tầng

Đặc tính kinh tế của dữ liệu chỉ ra rằng dữ liệu là nguồn lực cơ sở hạ tầng, theo lý thuyết có thể phục vụ số người dùng không hạn chế với các mục đích khác nhau, như một đầu vào để sản xuất hàng hóa và dịch vụ. Sự gia tăng lợi nhuận theo quy mô và phạm vi do việc sử dụng dữ liệu tạo ra là nguồn gốc của tăng năng suất dựa trên dữ liệu được các doanh nghiệp hiện thực hóa, khi dữ liệu được sử dụng, chẳng hạn như để phát triển các thị trường hỗn hợp, trong đó việc thu thập dữ liệu tại một đầu của thị trường lại tạo khả năng cho sản xuất hàng hóa và dịch vụ mới ở (các) đầu kia của thị trường (ví dụ như sử dụng dữ liệu được tạo ra từ các dịch vụ mạng xã hội cho mục đích quảng cáo).

Các đặc tính kinh tế của dữ liệu dẫn đến việc xem xét dữ liệu như một cơ sở hạ tầng hoặc nguồn lực cơ sở hạ tầng. Ban đầu điều này nghe có vẻ phản trực giác vì theo truyền thống, cơ sở hạ tầng thường được coi là những phương tiện vật chất quy mô lớn phục vụ tiêu dùng công cộng như các hệ thống giao thông bao gồm các hệ thống đường cao tốc hay đường sắt, các hệ thống thông tin liên lạc như các mạng điện thoại hoặc băng thông rộng, và các phương tiện và dịch vụ cơ bản như các tòa nhà, hệ thống thoát nước và cung

cấp nước (Frischmann, 2012). Tuy nhiên, như đã được Hội đồng nghiên cứu quốc gia Hoa Kỳ công nhận (NRC, 1987), khái niệm về cơ sở hạ tầng còn đề cập đến các cơ sở phi vật chất như các hệ thống giáo dục cũng như các hệ thống quản lý, bao gồm cả hệ thống tòa án. Theo Frischmann (2012), khái niệm rộng hơn về cơ sở hạ tầng này cho thấy nên nhìn cơ sở hạ tầng theo góc độ chức năng chứ không phải theo quan điểm ngữ nghĩa thuần túy.

Dữ liệu là hàng hóa phi cạnh tranh

Một loại hàng hóa được coi là có tính cạnh tranh thuần túy như dầu mỏ chỉ có thể tiêu thụ một lần. Ngược lại hàng hóa không có tính cạnh tranh như dữ liệu về nguyên tắc có thể tiêu thụ với số lần không giới hạn. Đặc tính này là nguồn gốc của những hiệu ứng lan tỏa quan trọng, nó cung cấp mối liên kết lý thuyết chủ yếu với tăng trưởng năng suất yếu tố tổng (Corrado et al. 2009). Nhưng nó cũng làm nảy sinh những vấn đề liên quan đến việc làm thế nào để phân bổ một cách tốt nhất nguồn tài nguyên dữ liệu này.

Trong khi đối với hàng hóa cạnh tranh thuần túy, điều được chấp nhận rộng rãi là phúc lợi xã hội sẽ đạt được tối đa khi một loại hàng hóa cạnh tranh được tiêu thụ bởi những người đánh giá nó cao nhất, và cơ chế thị trường là phương tiện hiệu quả nhất để phân phối hàng hoá đó và cũng là để phân bổ các nguồn lực cần thiết cho sản xuất các hàng hóa như vậy, điều này không phải lúc nào cũng đúng đối với hàng hóa phi cạnh tranh (Frischmann, 2012). Đối với hàng hoá không có tính cạnh tranh, tình hình phức tạp hơn do các loại hàng hóa này luôn đi kèm với một mức độ tự do bổ sung liên quan đến quản lý nguồn lực. Như Frischmann (2012) nhấn mạnh, phúc lợi xã hội sẽ không thể tối đa hóa khi hàng hóa chỉ được tiêu thụ bởi những người đánh giá chúng cao nhất, mà phải là tất cả những ai coi trọng nó. Việc tối đa hoá khả năng tiếp cận đến hàng hóa phi cạnh tranh về lý thuyết sẽ tối đa hóa phúc lợi xã hội như mọi lợi ích tự bổ sung mà không mất thêm chi phí.

Dữ liệu là tư liệu sản xuất

Dữ liệu thường được mô tả như một loại "*dầu mỏ mới*". Tuy nhiên, ngoài bản chất phi cạnh tranh của dữ liệu, ở đây còn có một tính chất khác nữa tương tự như vậy: đó là dữ liệu không phải là một loại hàng hóa tiêu dùng như một quả táo hay một mặt hàng trung gian như dầu mỏ. Trong hầu hết các trường hợp, dữ liệu có thể được phân loại như tư liệu sản xuất.

Tư liệu sản xuất theo OECD là "*hàng hóa, khác với nguyên liệu đầu vào và nhiên liệu, được sử dụng để sản xuất các loại hàng hóa và/hoặc dịch vụ khác*". Trái ngược với hàng tư liệu sản xuất, hàng hóa trung gian như: nguyên liệu thô (dầu mỏ) được sử dụng hết, làm cạn kiệt, hoặc nếu không thì được chuyển hóa để sử dụng như đầu vào cho sản xuất hàng hoá khác (Frischmann, 2012). Ngoài ra, tư liệu sản xuất "*phải được sản xuất ra như những sản phẩm đầu ra từ các quy trình sản xuất*", điều này giải thích tại sao "*tài nguyên thiên nhiên như đất đai, khoáng sản, than đá, dầu mỏ, hay khí tự nhiên, hoặc các hợp đồng, hợp đồng cho thuê và cấp giấy phép"* không được coi là tư liệu sản xuất (UN, 2008).

Dữ liệu đôi khi có thể được tiêu thụ để đáp ứng nhu cầu tiêu dùng trực tiếp. Ví dụ như khi tìm kiếm các số liệu thống kê của OECD chẳng hạn, chúng sẽ thông báo cho người đọc về tình hình kinh tế-xã hội. Tuy nhiên, trong hầu hết các trường hợp, dữ liệu thường được sử dụng như một đầu vào đối với hàng hóa hoặc dịch vụ, và điều này đặc biệt đúng đối với lượng dữ liệu lớn, chúng là phương tiện và không bị cạn kiệt. Nói cách khác, nhu cầu đối với "*dữ liệu lớn*" không bị chi phối bởi chính bản thân "*dữ liệu lớn*", mà là do những ích lợi từ việc sử dụng nó có thể mang lại. Theo nghĩa đó, thậm chí các sản phẩm dữ liệu thuần túy như thiết kế đồ họa thông tin (infographics) (là dạng thức thể hiện các thông tin, dữ liệu hay tri thức) là kết quả của minh họa thuật toán được áp dụng đối với dữ liệu.

Dữ liệu cũng không phải là một hàng hóa trung gian bởi vì nó không bị cạn kiệt khi sử dụng do có tính không cạnh tranh: trái ngược dầu mỏ, việc sử dụng dữ liệu về nguyên tắc không làm ảnh hưởng đến tiềm năng nó còn có thể đáp ứng nhu cầu của những người khác. Điều này không có nghĩa là dữ liệu không thể bị loại bỏ sau khi đã sử dụng. Trong nhiều trường hợp, dữ liệu chỉ được sử dụng một lần. Tuy nhiên, trong khi chi phí lưu trữ dữ liệu trong quá khứ không khuyến khích việc lưu giữ các dữ liệu được cho là không cần thiết nữa, giờ đây chi phí lưu trữ đã giảm đến mức cho phép có thể lưu trữ dữ liệu trong thời gian dài, thậm chí là vô hạn định. Điều này đã làm tăng khả năng dữ liệu được sử dụng như nguồn tư liệu sản xuất và yếu tố sản xuất.

Đặc tính tư liệu sản xuất của dữ liệu có ý nghĩa quan trọng đối với tăng trưởng kinh tế: do dữ liệu là một nguồn vốn phi cạnh tranh, về lý thuyết nó có thể được sử dụng đồng thời bởi nhiều người dùng cho nhiều mục đích, như một yếu tố đầu vào để sản xuất một số lượng không hạn chế các hàng hóa và dịch vụ. Đây chính là mối liên kết lý thuyết chủ yếu với tăng trưởng năng suất yếu tố tổng, mà trên thực tế được áp dụng trên các thị trường hỗn hợp được tạo năng lực nhờ dữ liệu, có nghĩa là nền tảng kinh tế trong đó các nhóm người dùng khác nhau tạo ra lợi ích (tác động ngoại lai hay hiệu ứng lan tỏa) đến các lĩnh vực khác.

Dữ liệu là đầu vào đa mục đích

Như Frischmann (2012) đã giải thích, "*nguồn lực cơ sở hạ tầng cho phép nhiều hệ thống (thị trường và phi thị trường) có thể hoạt động và đáp ứng nhu cầu của nhiều loại người dùng khác nhau*". Chúng không phải là yếu tố đầu vào được tối ưu hóa cho một mục đích hữu hạn cụ thể, "*chúng cung cấp chức năng cơ bản, đa mục đích*". Đặc biệt, cơ sở hạ tầng tạo khả năng sản xuất một loạt các loại hàng hóa tư nhân, công cộng, và xã hội, mà người dùng có thể tự do sản xuất theo khả năng của họ.

Việc sử dụng dữ liệu, thường phụ thuộc vào các nguồn dữ liệu. Ví dụ, dữ liệu nông nghiệp chủ yếu được sử dụng cho hàng hóa và dịch vụ nông nghiệp. Tuy nhiên, về mặt lý thuyết không có giới hạn về mục đích sử dụng dữ liệu và việc sử dụng lại dữ liệu có thể mang lại nhiều lợi ích với giả định rằng dữ liệu được tạo ra trong một lĩnh vực có thể cung

cấp những hiểu biết khi được áp dụng trong các lĩnh vực khác. Điều này rất rõ trong trường hợp dữ liệu mở thuộc khu vực công, trong đó một bộ dữ liệu sử dụng ban đầu vì mục đích hành chính được các doanh nghiệp sử dụng lại để tạo ra các dịch vụ mới mà ban đầu không hề được dự tính trước. Hoặc trong trường hợp chăm sóc sức khỏe, đặc biệt là nghiên cứu bệnh Alzheimer, là nơi có dữ liệu bán lẻ và mạng xã hội được các nhà nghiên cứu cân nhắc khi nghiên cứu về tác động của các mẫu hình hành vi và dinh dưỡng đến tiến triển bệnh tật.

Dữ liệu làm tăng lợi nhuận theo quy mô và phạm vi

Việc sử dụng dữ liệu có thể tạo ra lợi nhuận lớn nhờ quy mô và phạm vi do dữ liệu là nguồn vốn phi cạnh tranh, có thể được tái sử dụng bằng các vòng phản hồi tích cực, giúp tăng cường tác dụng ở phía cung và phía cầu. Tuy nhiên, điều này chỉ đúng ở một mức độ nhất định bởi tích lũy dữ liệu còn đi kèm với các chi phí (như lưu trữ) và rủi ro (vi phạm quyền riêng tư và rủi ro an ninh kỹ thuật số):

Về phía cung:

(1) *Tăng lợi nhuận nhờ quy mô*: Sự tích lũy dữ liệu có thể dẫn tới những cải thiện đáng kể ở dịch vụ dựa vào dữ liệu, điều đó có thể thu hút được nhiều người dùng hơn, dẫn tới dữ liệu có thể được thu thập nhiều hơn. Sự "*phản hồi tích cực này làm cho người đã mạnh trở nên mạnh hơn và kẻ yếu trở nên yếu hơn, dẫn đến các kết quả cực đoan*" (Shapiro và Varian, 1999). Ví dụ, nhiều người sử dụng các dịch vụ như Google Search, hoặc các công cụ tiến cử như của Amazon cung cấp, hay hệ thống định vị như của TomTom, các dịch vụ càng tốt hơn do cung cấp chính xác hơn các địa chỉ trang web và sản phẩm theo yêu cầu, hay cung cấp thông tin giao thông tốt hơn, chúng càng thu hút được nhiều người sử dụng hơn.

(2) *Tăng lợi nhuận nhờ phạm vi*: Việc đa dạng hóa dịch vụ có thể dẫn đến những hiểu biết tốt hơn nếu có thể liên kết dữ liệu. Điều này đạt được là do liên kết dữ liệu cho phép tăng thêm những hiểu biết sâu, dẫn đến tăng lợi nhuận nhờ vào phạm vi. Dữ liệu liên kết là một phương tiện để ngữ cảnh hóa dữ liệu và do đó là nguồn gốc cho những hiểu biết và giá trị lớn hơn so với từng phần riêng biệt (hay kho dữ liệu). Như Newman (2013) đã nhấn mạnh trong trường hợp Google: "*Google không chỉ thu thập dữ liệu từ chỗ mọi người sử dụng công cụ tìm kiếm của mình. Nó còn thu thập dữ liệu về những gì mà mọi người quan tâm khi viết trong các tài khoản Gmail của mình, những gì mọi người xem trên YouTube, những nơi họ được định vị khi sử dụng dữ liệu từ Google Maps, toàn bộ các mảng dữ liệu khác từ việc sử dụng điện thoại Android của Google, và thông tin người dùng được cung cấp từ mạng lưới các dịch vụ trực tuyến của Google*". Tập hợp dữ liệu đa dạng này đã cho phép công ty tạo ra những hồ sơ còn chi tiết hơn về người sử dụng của mình, điều không thể làm được nếu chỉ bằng các dịch vụ đơn lẻ.

Về phía cầu:

(1) *Tác động mạng lưới* (lợi thế kinh tế nhờ quy mô trọng cầu): nhiều dịch vụ và nền tảng dựa vào dữ liệu như cộng đồng mạng xã hội được đặc trưng bằng những hiệu ứng mạng lưới rộng, là nơi mà việc sử dụng các dịch vụ làm tăng quá mức số lượng người dùng. Điều này tăng cường lợi nhuận nhờ quy mô và phạm vi về phía cung.

(2) *Thị trường hỗn hợp*: dữ liệu có thể tạo khả năng cho các thị trường hỗn hợp. Việc sử dụng lại dữ liệu tạo ra lợi nhuận lớn nhờ quy mô và phạm vi, điều đó dẫn đến vòng phản hồi tích cực có lợi cho các doanh nghiệp nằm ở một đầu thị trường, và có thể làm tăng khả năng thành công tại một đầu khác của thị trường.

2.1.3. Các cơ chế tạo ra giá trị từ việc sử dụng và phân tích dữ liệu:

Có nhiều cơ chế thông qua đó giá trị có thể được tạo ra từ dữ liệu. Mặc dù có nhiều hình thức tạo giá trị, có thể phân biệt một số cơ chế chung như sau:

- Đạt được hiểu biết sâu (sáng tạo tri thức): phân tích dữ liệu là phương tiện kỹ thuật để rút ra những hiểu biết sâu và những công cụ nâng cao khả năng hiểu biết, tác động hoặc kiểm soát tốt hơn các đối tượng dữ liệu về những hiểu biết này (ví dụ như các hiện tượng tự nhiên, các hệ thống xã hội, các cá nhân). Ví dụ, các tổ chức ngày càng dựa vào các mô phỏng và thực nghiệm không chỉ để hiểu biết rõ hơn về hành vi của các cá nhân, mà còn để tìm hiểu, đánh giá và tối ưu hóa tác động có thể từ các hành động của họ đến những cá nhân đó.
- Ra quyết định của con người: hướng tới một nền văn hóa kinh doanh thử nghiệm dựa vào dữ liệu và sử dụng nguồn lực cộng đồng (*crowd sourcing).*

Việc dữ liệu được tạo ra và thu thập ở mọi nơi đã tạo khả năng cho các tổ chức có thể căn cứ quá trình ra quyết định của mình dựa vào dữ liệu nhiều hơn so với trước đây. Nổi bật có hai xu hướng chính gồm: (i) việc ra quyết định của con người ngày càng dựa trên các thử nghiệm nhanh dựa vào dữ liệu. (ii) việc sử dụng nguồn lực cộng đồng: "là việc có được các dịch vụ, các ý tưởng, hoặc nội dung cần thiết bằng cách thu hút sự đóng góp của một nhóm lớn dân chúng và đặc biệt là từ các cộng đồng trực tuyến" (Merriam-Webster, 2014) đã có thể dễ dàng đạt được nhờ vào khả năng đúc kết thông tin từ dữ liệu phi cấu trúc trên Internet và chia sẻ dữ liệu với các nhà phân tích khác.

Việc sử dụng phân tích dữ liệu trong quá trình ra quyết định cho thấy một sự thay đổi trong cách thức các tổ chức ra quyết định dựa vào dữ liệu. Người ra quyết định không nhất thiết phải hiểu được hiện tượng, trước khi hành động. Nói theo cách khác: thực tế phân tích đến trước, sau đó là hành động, và cuối cùng là sự hiểu biết. Ví dụ, một công ty như Wal-Mart Stores có thể thay đổi nơi đặt sản phẩm tại các cửa hàng của mình dựa trên các mối tương quan mà không nhất thiết phải biết tại sao sự thay đổi đó có tác động có lợi đến doanh thu của họ.

- Tự động ra quyết định (tự động hóa quyết định): phân tích dữ liệu (thông qua các

thuật toán học máy) nâng cao năng lực của máy móc và các hệ thống tự điều khiển, khiến cho chúng có thể học hỏi từ dữ liệu về các tình huống trước đó và có thể tự đưa ra các quyết định dựa trên phân tích các dữ liệu này. Các máy móc và hệ thống tự điều khiển này đang ngày càng trở nên có tính năng cao hơn do chúng có thể thực hiện được các nhiệm vụ mà trước đây thường đòi hỏi sự can thiệp của con người. Xe không người lái của Google là một ví dụ minh họa điển hình, nó được dựa trên các thuật toán học máy được làm giàu bằng dữ liệu thu thập từ các bộ cảm biến có kết nối với xe và từ các dịch vụ như Google Maps và Google Street View.

Máy móc tự động điều khiển được dự báo là có một tiềm năng lớn trong lĩnh vực hậu cần, chế tạo và nông nghiệp. Trong chế tạo, các robot có truyền thống được sử dụng chủ yếu ở những nơi cần đến tốc độ, độ chính xác, sự khéo léo và khả năng làm việc trong điều kiện nguy hiểm. Tuy nhiên, robot truyền thống chỉ nhanh trong các môi trường được xác định rất chính xác và việc thành lập một nhà máy robot sẽ phải mất hàng tháng, thậm chí hàng năm, để lập kế hoạch chính xác đến từng milimet về mọi di chuyển của robot. Tương tự như vậy, robot hậu cần vận chuyển các cấu kiện thành phẩm tuyến đường được dàn dựng chính xác. Các robot có thể được gắn các cảm biến, nhưng hầu hết các di chuyển cần được lên kế hoạch trước và lập trình, điều đó không cho phép có nhiều linh hoạt trong việc sản xuất sản phẩm. (*Vì lý do này, việc sản xuất các thiết bị điện tử tiêu dùng vẫn thường được thực hiện bằng tay, bởi vì vòng đời của các thiết bị điện tử tiêu dùng và thời gian để tiếp thị (đưa ra thị trường) quá ngắn*).

2.1.4. Các thách thức chính đặt ra đối với đổi mới sáng tạo-dựa trên-dữ liệu (DDI)

Vai trò kinh tế và xã hội của các dữ liệu không phải là mới. Các hoạt động kinh tế và xã hội từ lâu đã xoay quanh việc phân tích và sử dụng dữ liệu. Thậm chí trước cuộc cách mạng kỹ thuật số, dữ liệu đã được sử dụng, ví dụ như để khám phá khoa học và để giám sát các hoạt động kinh doanh như trong kế toán. Hơn nữa, trong kinh doanh, những khái niệm như "*trí tuệ doanh nghiệp*" (Luhn, 1958) và "*kho dữ liệu*" (Keen, 1978; Sol, 1987) đã xuất hiện trong những năm 1960 và trở nên phổ biến vào cuối thập niên 1980 khi máy tính ngày càng được sử dụng như những hệ thống hỗ trợ quyết định (DSS). Ngành tài chính là một ví dụ điển hình cho việc sử dụng lâu dài các hệ thống DSS tinh vi, ví dụ như để phát hiện gian lận và đánh giá rủi ro tín dụng.

Tuy nhiên, sự hợp lưu của ba xu hướng kinh tế-xã hội và công nghệ đã làm cho đổi mới sáng tạo dựa trên dữ liệu trở thành một hiện tượng mới hiện nay. Ba xu hướng đó bao gồm: (i) sự tăng trưởng theo cấp số nhân trong tạo ra và thu thập dữ liệu, (ii) sử dụng phân tích dữ liệu rộng rãi, bao gồm cả các doanh nghiệp mới khởi sự và doanh nghiệp vừa và nhỏ (SME), và (iii) xuất hiện sự thay đổi mô hình trong sáng tạo tri thức và ra quyết định. Tất cả những xu hướng này diễn ra trong suốt vòng đời giá trị dữ liệu. Sự hợp lưu của các xu hướng này theo mỗi giai đoạn của vòng đời giá trị dữ liệu đã tạo khả năng khai thác dữ liệu cho các dịch vụ theo cách thức mà trước đây không thể thực hiện được.

Mức độ phổ biến của các xu hướng này ở cấp quốc gia có thể ảnh hưởng đến sự sẵn

sàng của các nước để tận dụng lợi thế của đổi mới sáng tạo dựa vào dữ liệu. Điều này không có nghĩa là phải có đủ tất cả các yếu tố mới có thể hiện thực hóa được những lợi ích của đổi mới dựa vào dữ liệu. Tính chất toàn cầu của hệ sinh thái dữ liệu cho phép các nước có thể khai thác được những lợi ích của đổi mới dựa vào dữ liệu thông qua dữ liệu, phân tích, hàng hóa và dịch vụ dựa trên dữ liệu được sản sinh ở các nơi khác. Tuy nhiên, có thể giả định rằng các quốc gia phát triển mạnh theo những xu hướng này có nhiều khả năng hơn để tận dụng lợi thế của đổi mới dựa vào dữ liệu, do họ phát triển cung cấp dữ liệu, sử dụng và phân tích dữ liệu mạnh hơn. Những yếu tố này là những thách thức then chốt và không phải quốc gia nào cũng thực hiện đủ tốt để đạt được tiềm năng đầy đủ của đổi mới sáng tạo dựa trên dữ liệu.

Để đánh giá tốt hơn mức độ sẵn sàng của các nước trong tận dụng lợi thế của đổi mới dựa trên dữ liệu, cần phân biệt các thách thức (i) từ phía cung và (ii) từ phía cầu mà các nước phải đối mặt. Ngoài ra, có một số (iii) thách thức xã hội có liên quan đến những tác động có thể có của đổi mới dựa trên dữ liệu mà các nhà hoạch định chính sách cần phải giải quyết để bảo tồn các giá trị chung của nền dân chủ thị trường, đồng thời thúc đẩy tăng trưởng toàn diện và đẩy mạnh phúc lợi xã hội.

Các thách thức từ phía cung liên quan đến việc cung cấp và phân tích dữ liệu bao gồm:

- Đầu tư vào băng thông rộng di động và các rào cản luồng lưu chuyển dữ liệu tự do: Băng thông rộng di động có tiềm năng cho phép thực hiện DDI, đặc biệt là ở các vùng xa và kém phát triển (ví dụ như nông nghiệp). Tuy nhiên, tỷ lệ thâm nhập trong năm 2013 vẫn còn ở mức thấp, như ở các nước Chile, Thổ Nhĩ Kỳ, Hungary và Mexico. Tương tự, việc bảo vệ riêng tư, an ninh hoặc các thông tin kinh doanh bí mật là những lý do chính đáng để hạn chế luồng lưu chuyển dữ liệu tự do xuyên biên giới, giữa các ngành, các tổ chức, người tiêu dùng và các công dân, nhưng có thể ảnh hưởng xấu đến đổi mới dựa trên dữ liệu, ví dụ như hạn chế thương mại và cạnh tranh.
- Các vấn đề truy cập, sở hữu, và khuyến khích dữ liệu: DDI có thể đòi hỏi những khoản đầu tư đáng kể để phát triển và bảo trì các cơ sở dữ liệu, siêu dữ liệu và các thuật toán liên quan. Một số tổ chức và cá nhân có thể thiếu các động cơ khuyến khích chia sẻ các dữ liệu mà họ sở hữu và kiểm soát. Quyền sở hữu trí tuệ thường được đề xuất như một giải pháp để khắc phục các vấn đề khuyến khích. Tuy nhiên, trái ngược với các tài sản vô hình khác, dữ liệu thường liên quan đến việc chuyển nhượng các quyền hạn khác nhau giữa các bên nắm giữ dữ liệu, điều đó thách thức khả năng áp dụng khái niệm "*quyền sở hữu*". Trong trường hợp dữ liệu được coi là "*dữ liệu cá nhân*", khái niệm sở hữu thậm chí còn kém thực tế hơn, do các chế độ bảo mật riêng tư trao quyền kiểm soát rõ ràng cho các chủ thể dữ liệu.
- *Khả năng tiếp cận tới phân tích và điện toán đám mây*: Việc áp dụng phân tích dữ liệu được quyết định bởi một số yếu tố trong đó có quyền SHTT. Các phương thức cấp phép nguồn mở đang ngày càng được sử dụng để bảo vệ lợi ích của nhà đầu tư

trong khi cho phép hợp tác phát triển và sử dụng phân tích mở. Điện toán đám mây, thường được mô tả như một mô hình dịch vụ tính toán linh hoạt, mềm dẻo và đáp ứng theo yêu cầu, có tác dụng làm tăng khả năng lưu trữ và phân tích trên phạm vi toàn bộ nền kinh tế. Tuy nhiên, việc thiếu khả năng tương tác và nguy cơ bị phụ thuộc vào một nhà cung có thể gây cản trở việc áp dụng nó. Việc thiếu các tiêu chuẩn mở là một vấn đề trong lĩnh vực nền tảng cụ thể, ví dụ như dịch vụ (PaaS), nơi tài nguyên tính toán được cung cấp thông qua một nền tảng.

Các thách thức phía cầu liên quan đến khả năng tận dụng lợi thế của DDI:

- Kỹ năng và năng lực trong quản lý và phân tích dữ liệu: các khảo sát gần đây khẳng định rằng việc thiếu kỹ năng quản lý và phân tích dữ liệu là một rào cản quan trọng đối với việc áp dụng DDI, trong các lĩnh vực như khoa học, chăm sóc sức khỏe và khu vực công. Các chuyên gia dữ liệu chiếm khoảng 0,5% tổng số việc làm ở các nước như Phần Lan, Thụy Điển, Estonia, và Hoa Kỳ, trong khi Luxembourg và Hà Lan có nhiều hơn 1% tổng số lao động của họ là các chuyên gia dữ liệu. Tuy nhiên các kỹ năng như vậy cần phải được bổ sung năng lực theo lĩnh vực cụ thể để có thể diễn giải và đưa ra những quyết định tốt nhất dựa trên phân tích dữ liệu. Đây cũng là những lĩnh vực có tiềm năng tạo việc làm quan trọng theo một số ước tính cho thấy.
- Thay đổi tổ chức: sự bổ sung cho nhau giữa thay đổi về tổ chức và sử dụng ICT là rất quan trọng đối với sự tăng trưởng năng suất của doanh nghiệp. Các nghiên cứu hiện tại cho thấy rằng sự bổ sung cho nhau giữa thay đổi tổ chức và phân tích dữ liệu cũng có ý nghĩa quan trọng. Sự thay đổi về tổ chức có thể bị gián đoạn và do đó rất khó để thực hiện. Điều này có thể dẫn đến tình thế lưỡng nan của nhà cải cách, nơi mà các công ty thành công đặt quá nhiều trọng tâm vào sự thành công hiện tại, và do đó không chú trọng đổi mới dài hạn.
- Tinh thần kinh doanh: các doanh nghiệp mới khởi sự đang gia tăng với sự tập trung vào cung cấp hàng hóa và dịch vụ dữ liệu liên quan (bao gồm cả phân tích dữ liệu và các công cụ trực quan). Các doanh nghiệp mới khởi sự này thường nhanh nhạy hơn và có thể đáp ứng nhu cầu đặc biệt của khách hàng, là nơi mà các công ty lớn với các sản phẩm dữ liệu chung của họ thường không thể đáp ứng. Tuy nhiên, việc đạt được thành công đối với các doanh nghiệp dựa vào dữ liệu phụ thuộc vào các điều kiện kinh tế thuận lợi cho tinh thần khởi nghiệp nói chung, bao gồm cả khung pháp lý ảnh hưởng đến khả năng tiếp cận thị trường bán hàng, tiếp cận tài chính và thị trường lao động.

Các thách thức xã hội ảnh hưởng đến cả hai phía cung và cầu với các tác động bất lợi tiềm tàng đến các giá trị cốt lõi của các nền kinh tế thị trường dân chủ và phúc lợi của tất cả các công dân gồm:

- Mất tự chủ và tự do: Tiến bộ về phân tích dữ liệu làm cho nó thậm chí có thể suy luận ra các thông tin nhạy cảm hàm chứa trong những dữ liệu tầm thường. Việc sử

dụng sai những hiểu biết này có thể ảnh hưởng đến các giá trị và nguyên tắc cốt lõi, chẳng hạn như quyền tự chủ cá nhân, bình đẳng và tự do ngôn luận, và có thể có ảnh hưởng rộng lớn hơn đến toàn thể xã hội. Khả năng lọc (biết phân biệt) là có thể nhờ phân tích dữ liệu, có thể dẫn đến hiệu quả cao hơn, nhưng cũng làm hạn chế khả năng của cá nhân để thoát ra khỏi tác động của các chỉ số kinh tế-xã hội tồn tại từ trước. Các hành động phản ứng để giải quyết những thách thức này bao gồm cải thiện tính minh bạch, tiếp cận và nâng cao khả năng cho các cá nhân, thúc đẩy các tổ chức sử dụng dữ liệu cá nhân theo cách có trách nhiệm và sử dụng công nghệ trong các dịch vụ bảo vệ riêng tư.

- Tập trung và thống trị thị trường: kinh tế học dữ liệu thuận lợi cho tập trung và thống trị thị trường. Theo tài liệu của OECD về cạnh tranh trong nền kinh tế kỹ thuật số, các thị trường dựa trên dữ liệu có thể dẫn đến kết quả "người chiến thắng có tất cả", khi mà sự tập trung là kết quả có thể của sự thành công trên thị trường. Có một số yếu tố đặc trưng của đổi mới dựa trên dữ liệu có thể thách thức cách tiếp cận truyền thống được các cơ quan quản lý cạnh tranh sử dụng để đánh giá những lạm dụng và tác hại tiềm năng của sự thống trị thị trường và các vụ sáp nhập. Các yếu tố đó bao gồm: (i) thách thức trong việc xác định thị trường liên quan, và trong việc đánh giá (ii) mức độ tập trung thị trường, và (iii) tác hại tiêu dùng tiềm năng do vi phạm quyền riêng tư.
- Sự thay đổi về quyền lực làm tăng thêm bất bình đẳng hiện tại: những hiểu biết dựa trên dữ liệu tốt hơn đi kèm với sự hiểu biết tốt hơn về các đối tượng dữ liệu và về cách tốt nhất để gây ảnh hưởng hoặc kiểm soát chúng. Nơi tích tụ dữ liệu dẫn đến sự tập trung và bất cân xứng thông tin lớn hơn, những thay đổi quan trọng về quyền lực có thể chuyển từ: i) cá nhân đến tổ chức (bao gồm người tiêu dùng đến các doanh nghiệp, và công dân đến chính phủ);. ii) các doanh nghiệp truyền thống đến các doanh nghiệp dựa vào dữ liệu, xuất phát từ gia tăng lợi nhuận theo quy mô và rủi ro tiềm ẩn của tập trung và thống trị thị trường; iii) chính phủ đến các doanh nghiệp dựa vào dữ liệu, trong đó các doanh nghiệp có thể có được nhiều kiến thức về các công dân hơn so với chính phủ; và iv) từ các nền kinh tế kém phát triển đến các nền kinh tế dựa vào dữ liệu.
- Thay đổi cơ cấu trên các thị trường lao động: tự động ra quyết định nhờ vào các ứng dụng "thông minh" có thể là ứng dụng của đổi mới dựa trên dữ liệu với những tác động lớn nhất đến năng suất (lao động). Các ứng dụng này đang trở nên ngày càng mạnh hơn và có thể thực hiện ngày càng nhiều các nhiệm vụ thâm dụng tri thức và lao động, và sẽ yêu cầu sự can thiệp của con người ít hơn so với trước đây. Điều này có thể có tác động quan trọng đến việc làm, đặc biệt là những công việc mang tính chất "giao dịch", dẫn đến sự thay đổi cấu trúc hơn nữa trên các thị trường lao động với những tác động tiềm năng đến bất bình đẳng trong thu nhập.
- Hạn chế phương pháp tiếp cận bảo mật truyền thống: Để tạo thuận lợi cho đổi mới

sáng tạo, đổi mới dựa trên dữ liệu đòi hỏi một môi trường kỹ thuật số mở và nối kết, cũng như linh hoạt, cho phép lưu trữ, truy cập và chia sẻ những khối lượng dữ liệu khổng lồ, đa dạng trên khắp hệ sinh thái dữ liệu. Những đặc điểm liên quan đến nhau này làm tăng tính phức tạp của quản trị an ninh kỹ thuật số, kêu gọi một cách tiếp cận hiện đại hơn dựa trên rủi ro thu hút sự tham gia của tất cả các bên liên quan.

2.2. Các công nghệ thúc đẩy đổi mới sáng tạo dựa trên dữ liệu

Khối lượng dữ liệu sẵn có ngày càng gia tăng đã thúc đẩy sự phát triển của các kỹ thuật và công nghệ mới trong mọi giai đoạn của vòng đời dữ liệu, từ thu thập, lưu trữ, thao tác, phân tích, sử dụng đến phổ biến dữ liệu. Ngược lại, những công nghệ này làm gia tăng giá trị của dữ liệu thô, đưa đến việc thu thập nhiều hơn và thậm chí tính khả dụng của dữ liệu cũng tăng lên.

2.2.1. Quản trị dữ liệu

Thu thập

Thu thập dữ liệu là bước đầu tiên trong chu trình đổi mới dựa vào dữ liệu. Tính đến năm 2012, khoảng 2,5 tỷ gigabyte dữ liệu đã được thu thập mỗi ngày trên toàn cầu, một phần đáng kể trong số đó là video. Trong khi đó, toàn bộ bộ sưu tập in trong Thư viện Quốc hội Hoa Kỳ mới chỉ chiếm khoảng 10.000 gigabyte.

Hai nguồn chính của dữ liệu số mới là các thiết bị cảm biến vật lý và các biểu ghi điện tử. Hầu hết các thiết bị điện tử, kích thước và giá thành của nhiều thiết bị cảm biến đã giảm đáng kể trong thập kỷ qua trong khi chức năng của chúng tăng đáng kể. Công nghệ cảm biến dẫn đến một loạt những thiết bị đo các biến số vật lý như nhiệt độ, áp suất, định vị, thành phần hóa học, dòng điện, chuyển động, hàm lượng ánh sáng và nhiều biến số khác. Các thiết bị cảm biến là một phần không thể thiếu của Internet vạn vật - IoT, một khái niệm được sử dụng để mô tả một thế giới nơi hàng ngày, các đối tượng, từ máy bay tới tủ lạnh và giày chạy, có thể giao tiếp với nhau và với người sử dụng chúng. Ví dụ, máy bay Boeing 787 tạo ra hơn một nửa terabyte dữ liệu trong mỗi chuyến bay từ các động cơ, thiết bị hạ cánh và các thiết bị khác. Các thiết bị cảm biến có tính chuyên dụng cao và nhiều biến số của thiết bị thường được sử dụng để đo một biến môi trường nhất định trong các phạm vi ứng dụng khác nhau. Các nhà khoa học dữ liệu thường sử dụng các kỹ thuật xử lý tín hiệu và lập mô hình thống kê để thu được những hiểu biết từ dữ liệu cảm biến, ví dụ như Trung tâm Khí tượng quốc gia sử dụng việc lập mô hình khí hậu trong các dự báo của mình. Lượng dữ liệu cảm biến sẽ tiếp tục tăng khi các thiết bị cảm biến hiệu quả hơn và rẻ hơn, và các công ty đã nhúng chúng vào các thiết bị ngày càng nhiều. Sự ra đời của các bộ xử lý giá rẻ, tiêu thụ điện năng thấp cũng sẽ hỗ trợ cho sự gia tăng dữ liệu cảm biến, cho phép các công ty có thể nhúng năng lực xử lý thông minh vào bất kỳ thiết bị nào.

Biểu ghi điện tử bao gồm các dữ liệu có cấu trúc và phi cấu trúc. Dữ liệu có cấu trúc là

dữ liệu được tổ chức chặt chẽ và dễ dàng truy vấn, chẳng hạn như dữ liệu bảng về các giao dịch, chi tiết tài khoản và các hoạt động trực tuyến khác. Theo thiết kế, việc phân tích dữ liệu có cấu trúc thường đơn giản hơn; các ứng dụng nhất định, chẳng hạn như phân tích mạng lưới và lập mô hình dự báo, cần đến dữ liệu có cấu trúc. Dữ liệu phi cấu trúc là dữ liệu được tổ chức kém hơn và không thích hợp để truy vấn, chẳng hạn như hình ảnh, video và âm thanh. Ví dụ, một biểu ghi điện tử của phòng thí nghiệm của một bệnh viện hay một bảng kê khai hàng hóa vận chuyển được số hóa của một công ty vận tải thường được lưu trữ theo các định dạng có cấu trúc; tin tức, video trực tuyến và các đánh giá sản phẩm bằng văn bản thường là các dữ liệu phi cấu trúc.

Dữ liệu có cấu trúc được các tổ chức, cả công và tư, thu thập với số lượng lớn. Ví dụ, Công ty Dịch vụ bưu phẩm hợp nhất (Hoa Kỳ) nhận trung bình 39,5 triệu yêu cầu theo dõi đường đi bưu phẩm mỗi ngày và Công ty Visa xử lý hơn 172 triệu giao dịch thẻ mỗi ngày. Tuy nhiên, phần lớn dữ liệu được thu thập hiện nay là phi cấu trúc và nhiều trong số đó dưới dạng video. Tính đến tháng 6 năm 2012, cứ mỗi phút người dùng đã tải lên YouTube 48 giờ video.

Những tiến bộ đạt được trong các mạng cố định và không dây cũng ảnh hưởng đến lượng dữ liệu được thu thập và hàng loạt các cơ hội cho đổi mới dựa vào dữ liệu. Một phân tích của Cisco năm 2013 cho rằng lưu lượng sử dụng internet trên toàn cầu thông qua các mạng viễn thông sẽ tăng lên gần ba lần từ năm 2012 đến năm 2017, với tổng số 3,1 exabyte mỗi ngày.

Lưu trữ

Dữ liệu phải được lưu trữ ngay sau khi thu thập. Vệc lưu trữ dữ liệu hiệu quả và linh hoạt có thể làm đơn giản hóa phân tích dữ liệu và tiết kiệm đáng kể chi phí. Trong hai thập kỷ qua, lưu trữ dữ liệu đã được hưởng lợi từ những thành tựu đạt được trong đổi mới sáng tạo phần mềm và phần cứng.

Phần cứng được cải tiến cho phép chi phí lưu trữ giảm mạnh; năm 1980, chi phí cho lưu trữ một gigabyte dữ liệu vào khoảng 440.000 USD, thì năm 2013, chi phí này chỉ khoảng 0,05 USD. Những tiến bộ đạt được tại các trung tâm dữ liệu cũng đã làm cho việc lưu trữ dữ liệu với số lượng lớn của các tổ chức dễ dàng hơn và với chi phí thấp hơn do sử dụng các phương pháp lưu trữ điện toán đám mây từ xa. Ngoài những cải tiến đáng kể về phần cứng, các nhà phát triển đã tạo ra một loạt các phần mềm cơ sở dữ liệu được thiết kế để lưu trữ dữ liệu phi cấu trúc và có khả năng mở rộng "dữ liệu lớn". Các cơ sở dữ liệu với ngôn ngữ truy vấn có cấu trúc (SQL) truyền thống dựa vào các cấu trúc được tổ chức chặt chẽ, đôi khi không phù hợp với dữ liệu đầu vào không đồng nhất và thay đổi. Những hệ thống này, đã được sử dụng trong nhiều thập kỷ để lưu trữ các tập tin của nhân viên, dữ liệu doanh số bán hàng và các thông tin được tổ chức chặt chẽ khác, không dễ dàng mở rộng cho nhiều ứng dụng khoa học dữ liệu hiện đại, chẳng hạn như lưu trữ tài liệu.

Các công ty của Hoa Kỳ, cùng với cộng đồng mã nguồn mở toàn cầu, là những người tiên phong trong việc phát triển các công nghệ khắc phục một số những hạn chế này. Nói

chung, các công nghệ mới được gọi là công nghệ không phải SQL hay NoSQL (not only SQL), để biểu thị sự loại bỏ các tính chất SQL khác nhau, bao gồm cả những hạn chế về lưu trữ tập trung và sửa đổi dữ liệu. Ví dụ về các công nghệ NoSQL độc quyền bao gồm BigTable của Google, Dynamo của Amazon và Facebook của Cassandra, tất cả các công nghệ này đã thúc đẩy sự phát triển các công nghệ mã nguồn mở cho phép lưu trữ và phân tích dữ liệu lớn. Ví dụ, Công ty Apache Software Foundation phát triển HBase, cơ sở dữ liệu phổ biến với dữ liệu lớn, dựa trên công trình nghiên cứu ban đầu do Google thực hiện.

2.2.2. Phân tích dữ liệu

Phân tích dữ liệu là những gì làm cho dữ liệu lớn trở nên sống động. Phân tích dữ liệu chính là rút ra ý nghĩa từ dữ liệu, một phần được thực hiện bằng cách xác định mối tương quan giữa các biến số và đưa ra các dự đoán về các sự kiện trong tương lai. Nếu không có phân tích, các tập hợp dữ liệu lớn có thể được lưu trữ và được truy xuất, toàn bộ hoặc có chọn lọc, nhưng những dữ liệu truy xuất sẽ chính là những dữ liệu đầu vào.

Tốc độ tăng trưởng mạnh mẽ của dữ liệu phi cấu trúc đã thúc đẩy sự phát triển của các kỹ thuật như khai phá văn bản (text mining), xử lý ngôn ngữ tự nhiên và hình ảnh máy tính (computer vision), tất cả đều có thể giúp dữ liệu phi cấu trúc có ý nghĩa. Các nhà phát triển công nghệ cũng đã rất nỗ lực để tạo ra phần mềm thao tác và phân tích dữ liệu, bao gồm cả ngôn ngữ lập trình số, phần mềm thống kê, các công cụ phân tích kinh doanh chuyên dụng và các tiện ích "dữ liệu lớn". Đầu tư vào các ứng dụng phân tích này có thể mang lại lợi nhuận cao; một nghiên cứu năm 2011 cho thấy các công ty kiếm được trung bình 10,66 USD cho mỗi đôla đầu tư cho các ứng dụng phân tích.

Do tính linh hoạt, các ngôn ngữ lập trình, chẳng hạn như ngôn ngữ điện toán thống kê R và các ngôn ngữ điện toán số Matlab và Julia, được sử dụng để phân tích và thao tác dữ liệu trong nhiều lĩnh vực. Ngôn ngữ lập trình cho phép người sử dụng tạo ra và phân bổ các chức năng riêng của chúng; ví dụ, ngôn ngữ điện toán thống kê R có các chức năng chuyên biệt cho rất nhiều lĩnh vực, bao gồm phân tích hình ảnh y tế, toán kinh tế và phân tích sinh thái. Ngôn ngữ lập trình Python đa năng đã được mở rộng để bao gồm cả khả năng thống kê. Ngoài ra còn có một loạt các phần mềm thống kê chuyên dụng, chẳng hạn như SAS, SPSS và Stata.

Một tập hợp con quan trọng của phần mềm thống kê là phần mềm phân tích kinh doanh được các công ty sử dụng để đưa ra các quyết định kinh doanh dựa vào dữ liệu. Phần mềm phân tích kinh doanh rất đa dạng và bao gồm các gói có sẵn từ các nhà cung cấp như Adobe, IBM, Microsoft, Oracle, SAP và SAS. Những công cụ thân thiện người dùng này cho phép các nhà phân tích thăm dò và thao tác dữ liệu bằng cách sử dụng các lệnh được cài đặt trước trong phần mềm. Mặc dù ít linh hoạt hơn các ngôn ngữ lập trình, phần mềm phân tích kinh doanh có thể đặc biệt hữu ích cho các ngành công nghiệp sử dụng các số liệu thống kê được xác định rõ ràng, chẳng hạn như ngành bảo hiểm.

Ngoài ra, các công cụ đã được phát triển đặc biệt cho dữ liệu lớn, chẳng hạn như

Hadoop, một nền tảng mã nguồn mở cho các ứng dụng liên quan đến phân tích các bộ dữ liệu lớn. Các tổ chức của nhiều ngành công nghiệp, bao gồm chăm sóc y tế, nông nghiệp và ngành tiện ích, sử dụng chức năng cốt lõi của Hadoop để xử lý những lượng dữ liệu lớn. Nhiều nhà phát triển đã tạo ra các phần mở rộng và các tiện ích phụ cho các trường hợp sử dụng cụ thể, chẳng hạn như phân tích thời gian thực. Ví dụ, một phòng thí nghiệm tại Viện Y khoa Howard Hughes ở Maryland sử dụng một nền tảng phân tích thời gian thực dựa vào Hadoop để phân tích và hiển thị các mô hình hoạt động của não trong thời gian thực.

Mặc dù việc phân tích được tiến hành bằng cách sử dụng phần mềm, những cải tiến trong phần cứng máy tính, đặc biệt là các bộ xử lý cho phép xử lý dữ liệu nhanh hơn, rẻ hơn và tiêu thụ ít năng lượng hơn trong những năm qua. Phần cứng được chế tạo để phân tích dữ liệu quy mô lớn bao gồm bộ xử lý đa lõi được Intel và AMD liên tục tinh chỉnh; phần cứng máy chủ hiệu năng cao của IBM (dựa trên các công nghệ được phát triển cho dự án Watson của IBM); máy chủ "*bộ nhớ lớn*" với dung lượng lưu trữ cao của Oracle; và các thiết bị được tối ưu hóa cho "*dữ liệu lớn*" từ HP và EMC. Sự gia tăng của điện toán song song và xử lý đám mây đã làm cho tốc độ của bộ xử lý đạt đến mức ít nút thắt cổ chai hơn để phân tích dữ liệu so với các thập kỷ trước, nhưng những tiến bộ gia tăng của các nhà cung cấp phần cứng vẫn là một động lực quan trọng cho các ứng dụng hiệu suất cao.

Phân tích dữ liệu, bao gồm việc sử dụng một số công nghệ tính toán khác nhau đang nuôi dưỡng cho cuộc cách mạng dữ liệu lớn. Việc phân tích để tạo ra giá trị mới trong các tập dữ liệu lớn, lớn hơn rất nhiều tổng giá trị của các dữ liệu thành phần.

Khai phá dữ liệu (data mining)

Khai phá dữ liệu, đôi khi được đánh đồng với phân tích, nhưng thực chất khai phá dữ liệu chỉ là một tập hợp con của phân tích dữ liệu, dùng để chỉ một quá trình tính toán để phát hiện ra các mẫu trong các tập dữ liệu lớn. Phân tích là sự hội tụ của nhiều lĩnh vực nghiên cứu khoa học, bao gồm cả toán học ứng dụng, khoa học máy tính, thống kê, cơ sở dữ liệu, trí tuệ nhân tạo và học máy (machine learning). Giống như các công nghệ khác, những tiến bộ trong khai phá dữ liệu có giai đoạn nghiên cứu và phát triển, trong đó các thuật toán và các chương trình máy tính mới được phát triển và các giai đoạn tiếp theo là thương mại hóa và ứng dụng.

Các đầu ra mong muốn của khai phá dữ liệu có thể có nhiều dạng, mỗi dạng có các thuật toán chuyên dụng riêng, cụ thể:

- Thuật toán phân loại: cố gắng gán các đối tượng hoặc sự kiện với các thể loại đã được biết đến. Ví dụ, một bệnh viện có thể muốn phân loại bệnh nhân xuất viện theo nguy cơ phải nhập viện trở lại ở các mức cao, trung bình hoặc thấp.
- Thuật toán phân cụm: nhóm các đối tượng hoặc sự kiện thành các thể loại tương tự, ví dụ như "mèo".

- Thuật toán hồi quy: (còn gọi là thuật toán dự đoán số) cố gắng dự đoán số lượng số. Ví dụ, một ngân hàng có thể muốn dự đoán, từ những chi tiết trong đơn xin vay tiền, xác suất của một mặc định.
- Kỹ thuật liên kết: cố gắng để tìm ra mối quan hệ giữa các mục trong tập dữ liệu. Sản phẩm gợi ý của Amazon và các bộ phim đề xuất của Netflix là những ví dụ cho kỹ thuật này.
- Thuật toán phát hiện dị thường: tìm kiếm các ví dụ không điển hình trong một tập hợp dữ liệu, ví dụ, phát hiện các giao dịch gian lận trên tài khoản thẻ tín dụng.
- Kỹ thuật tổng hợp: để tìm và đưa ra các tính chất nổi bật trong dữ liệu. Các ví dụ bao gồm các bản tóm tắt thống kê đơn giản (ví dụ, điểm thi trung bình của học sinh theo trường và giáo viên) và phân tích cấp cao hơn (ví dụ, một danh sách các sự kiện quan trọng về một cá nhân được thu thập từ tất cả các thông tin đăng trên web có liên quan đến cá nhân đó).

Khai phá dữ liệu đôi khi bị nhầm lẫn với học máy. Học máy là trường con rộng của khoa học máy tính trong nghiên cứu khoa học và công nghiệp. Khai phá dữ liệu sử dụng máy học cũng như các ngành khác, trong khi máy học có ứng dụng cho các lĩnh vực khác chứ không phải là khai phá dữ liệu, ví dụ như ngành khoa học người máy.

Khai phá dữ liệu về khả năng có những hạn chế, cả về thực tiễn và lý luận, thực hiện, cũng như giới hạn về độ chính xác có thể đạt được. Việc khái phá dữ liệu có thể tìm ra các mẫu và các mối quan hệ, nhưng nó thường không cho người sử dụng biết giá trị hay ý nghĩa của những mô hình này. Ví dụ, học có giám sát dựa vào các đặc điểm của những kẻ khủng bố đã được biết có thể tìm ra những người tương tự, nhưng họ có thể là hoặc có thể không phải là kẻ khủng bố; và nó sẽ bỏ qua các loại khủng bố khác, những người không phù hợp với hồ sơ.

Khai phá dữ liệu có thể xác định những mối quan hệ giữa các hành vi và/hoặc các biến, nhưng những mối quan hệ này không phải lúc nào cũng biểu thị quan hệ nhân quả. Nếu người dân sống dưới đường dây điện cao thế có tỷ lệ mắc bệnh cao hơn, điều này có thể có nghĩa là đường dây điện là một mối nguy hiểm cho sức khỏe cộng đồng; hoặc nó có thể có nghĩa là những người sống dưới đường dây điện có xu hướng nghèo và có quyền được chăm sóc sức khỏe đầy đủ. Các tác động chính sách là khá khác nhau. Trong khi các biến gây nhiễu (trong ví dụ này là thu nhập) có thể được sửa chữa khi chúng được biết và hiểu rõ, không có cách nào chắc chắn để biết liệu tất cả các biến đã được xác định hay không. Việc quy mối quan hệ nhân quả là đúng trong dữ liệu lớn là một lĩnh vực nghiên cứu vẫn còn ở giai đoạn sơ khai.

Nhiều phép phân tích dữ liệu khai thác mối tương quan có thể hoặc không thể phản ánh quan hệ nhân quả. Một số phép phân tích dữ liệu phát triển thông tin không hoàn hảo, hoặc là do những hạn chế của các thuật toán, hoặc do sự lấy mẫu chệch. Sử dụng hỗn tạp những phân tích này có thể gây ra sự kỳ thị đối với các cá nhân hoặc sự thiếu công bằng

vì sự liên kết không chính xác với một nhóm cụ thể. Khi sử dụng các phân tích dữ liệu, phải đặc biệt bảo vệ sự riêng tư của trẻ em và các nhóm được bảo vệ khác.

Dữ liệu thực tế có thể chưa đầy đủ và có nhiễu. Những vấn đề về chất lượng của dữ liệu làm giảm hiệu suất của các thuật toán khai phá dữ liệu và các kết quả đầu ra tối nghĩa. Khi điều kiện kinh tế cho phép, việc sàng lọc cẩn thận và chuẩn bị dữ liệu đầu vào có thể cải thiện chất lượng của các kết quả, nhưng sự chuẩn bị dữ liệu này thường phải sử dụng nhiều lao động và tốn kém. Người sử dụng, đặc biệt là trong lĩnh vực thương mại, phải đánh đổi chi phí lấy tính chính xác, đôi khi với những hậu quả tiêu cực đối với các cá nhân có thông tin trong dữ liệu. Ngoài ra, dữ liệu thực tế có thể chứa các sự kiện cực đoan hay các giá trị ngoại lệ. Các giá trị ngoại lệ có thể là các sự kiện thực sự, ngẫu nhiên, tồn tại rất nhiều trong dữ liệu; hoặc chúng có thể là kết quả của các lỗi nhập dữ liệu hoặc truyền dữ liệu. Trong cả hai trường hợp, chúng có thể làm lệch mô hình và làm giảm hiệu suất. Nghiên cứu về các giá trị ngoại lệ là một lĩnh vực nghiên cứu thống kê quan trọng.

Trộn dữ liệu và tích hợp thông tin

Trộn dữ liệu là sự kết hợp của nhiều bộ dữ liệu không đồng nhất thành một dạng đồng nhất để chúng có thể được xử lý tốt hơn cho khai phá và quản lý dữ liệu. Trộn dữ liệu được sử dụng trong một số lĩnh vực kỹ thuật như mạng cảm biến, xử lý video/hình ảnh, robot và các hệ thống thông minh, v.v...

Tích hợp dữ liệu khác với trộn dữ liệu, trong đó tích hợp là sự kết hợp rộng hơn các tập dữ liệu và giữ lại tập thông tin lớn hơn. Kỹ thuật cắt giảm hay thay thế thường được sử dụng trong trộn dữ liệu. Trộn dữ liệu được hỗ trợ bởi khả năng tương tác dữ liệu, khả năng để hai hệ thống giao tiếp và trao đổi dữ liệu. Trộn dữ liệu và tích hợp dữ liệu là các kỹ thuật quan trọng cho quản trị kinh doanh thông minh. Các nhà bán lẻ đang tích hợp các cơ sở dữ liệu trực tuyến, tại cửa hàng và danh mục bán hàng để tạo ra các bức tranh hoàn chỉnh hơn về khách hàng của họ. Ví dụ Williams-Sonoma đã tích hợp các cơ sở dữ liệu khách hàng với thông tin về 60 triệu hộ gia đình. Các biến bao gồm thu nhập của hộ gia đình, giá trị nhà ở và số trẻ em được theo dõi. Công ty này tuyên bố rằng thư điện tử nhằm mục tiêu dựa trên các thông tin này có tỷ lệ phản hồi nhiều hơn từ 10 đến 18 lần so với thư không nhằm mục tiêu. Đây là một minh họa đơn giản về cách nhiều thông tin hơn có thể dẫn đến các suy luận tốt hơn. Các kỹ thuật có thể giúp bảo vệ sự riêng tư đang được quan tâm.

Hiện nay, các kỹ thuật trộn dữ liệu đa cảm biến rất được quan tâm. Những thách thức kỹ thuật lớn nhất được giải quyết hiện nay, nói chung thông qua phát triển các thuật toán mới và tốt hơn, liên quan đến độ chính xác/độ phân giải dữ liệu, giá trị ngoại lệ và dữ liệu không xác thực, dữ liệu mâu thuẫn, phương thức (cả dữ liệu không đồng nhất và đồng nhất) và chiều, tương quan dữ liệu, sắp xếp dữ liệu, liên kết trong dữ liệu, xử lý tập trung hóa so với xử lý phi tập trung, thời gian hoạt động và khả năng xử lý hiện tượng động so với tĩnh. Những quan ngại về tính riêng tư có thể phát sinh từ độ trung thực và độ chính xác của cảm biến cũng như mối tương quan của dữ liệu từ nhiều cảm biến. Một đầu ra của

một cảm biến có thể không phải là vấn đề nhạy cảm, nhưng sự kết hợp từ hai hay nhiều đầu ra có thể làm tăng mỗi lo ngại về vấn đề riêng tư.

Nhận diện hình ảnh và giọng nói

Các công nghệ nhận diện hình ảnh và giọng nói có thể trích xuất thông tin từ các kho dữ liệu lớn hình ảnh, video và bài phát biểu được ghi âm hoặc phát sóng.

Khai thác quang cảnh đô thị có thể được thực hiện bằng cách sử dụng nhiều nguồn dữ liệu từ các bức ảnh và video đến LiDAR mặt đất (Light Detecting And Ranging - kỹ thuật viễn thám sử dụng laser). Trong khu vực của chính phủ, các mô hình thành phố đang trở nên quan trọng đối với việc lập kế hoạch và trực quan hóa đô thị. Chúng cũng rất quan trọng đối với một loạt các môn học bao gồm cả lịch sử, khảo cổ học, địa lý và nghiên cứu đồ họa máy tính. Các mô hình thành phố dưới dạng số cũng là trung tâm của các ứng dụng trực quan hóa và lập bản đồ khách hàng phổ thông như Google Earth và Bing Maps, cũng như các hệ thống định vị GPS. Khai thác quang cảnh là một ví dụ của việc nắm giữ không chủ ý thông tin cá nhân và có thể được sử dụng để trộn dữ liệu làm tiết lộ thông tin cá nhân.

Các công nghệ nhận diện khuôn mặt, đang bắt đầu trở thành hiện thực trong các ứng dụng thương mại và thực thi pháp luật. Các công nghệ này có thể thu được, bình thường hóa và nhận dang các khuôn mặt di chuyển trong cảnh động. Việc giám sát bằng video thời gian thực được thực hiện bằng các hệ thống một camera (và một số bằng các hệ thống nhiều camera, cả hai hệ thống này đều có thể nhận ra các đối tượng và phân tích hoạt động) có hàng loạt ứng dụng trong cả khu vực công và tư nhân, chẳng hạn như an ninh quốc gia, phòng chống tội phạm, điều khiển giao thông, dự báo, phát hiện tai nạn và theo dõi bệnh nhân, người già và trẻ em ở nhà. Tùy thuộc vào ứng dụng, việc sử dụng giám sát bằng video ở các cấp được triển khai khác nhau.

Các khả năng khác của nhận diện hình ảnh bao gồm:

- Tổng hợp video và phát hiện thay đổi hiện trường (có nghĩa là, chọn một số hình ảnh tóm tắt một khoảng thời gian)
- Định vị địa lý chính xác hình ảnh từ vệ tinh hoặc máy bay
- Sinh trắc học dựa vào hình ảnh
- Hệ thống giám sát đòi hỏi có sự tương tác của con người
- Tái nhận diện người và phương tiện, tức là theo dõi một người hay một phương tiện khi người hay phương tiện đó di chuyển từ cảm biến này đến cảm biến khác.
- Nhận diện các loại khác hoạt động khác nhau của con người
- Tổng kết ngữ nghĩa (có nghĩa là chuyển đổi hình ảnh thành các bảng tóm tắt bằng văn bản)

Mặc dù các hệ thống này được dự kiến có thể theo dõi các đối tượng qua camera và phát hiện các hoạt động bất thường trong một khu vực rộng lớn bằng cách kết hợp thông

tin từ nhiều nguồn, việc tái nhận diện các đối tượng vẫn còn khó khăn (một thách thức đối với theo dõi liên camera), ví dụ như giám sát bằng video trong môi trường đông đúc.

Nhận diện giọng nói tự động đã xuất hiện ít nhất từ những năm 1950, nhưng những phát triển gần đây trong 10 năm qua đưa đến các khả năng mới. Các văn bản nói (ví dụ, các phát thanh viên của một bản tin đọc một phần của một tài liệu) hiện nay có thể được nhận diện với độ chính xác cao hơn 95% do sử dụng các thuật kỹ thuật hiện đại. Để nhận diện chính xác giọng nói tự nhiên thì khó hơn nhiều. Trong những năm gần đây, đã có sự gia tăng đáng kể các kho ngữ liệu các dữ liệu giọng nói tự nhiên do đó các nhà nghiên cứu có thể cải thiện được độ chính xác.

Trong vài năm tới, các giao diện nhận diện giọng nói sẽ được áp dụng nhiều hơn. Ví dụ, nhiều công ty đang nghiên cứu kỹ thuật nhận diện giọng nói để điều khiển TV và xe hơi, tìm một chương trình trên truyền hình, hoặc lên lịch ghi hình DVR. Các nhà nghiên cứu tại Nuance nói rằng họ đang tích cực lên kế hoạch thiết kế công nghệ giọng nói như thế nào để tích hợp vào các máy tính mặc được. Google đã thực hiện một số chức năng cơ bản này trong sản phẩm Google Glass của mình và hệ thống Xbox one của Microsoft đã tích hợp thị giác máy và đầu vào âm thanh của đa micrô để điều khiển các chức năng hệ thống.

Phân tích mạng xã hội

Phân tích mạng xã hội đề cập đến việc trích xuất thông tin từ một loạt các đơn vị liên kết theo giả định rằng các mối quan hệ của chúng rất quan trọng và các đơn vị này không hành xử một cách độc lập. Các mạng xã hội thường ở trong một ngữ cảnh trực tuyến. Các ví dụ rõ ràng nhất là các nền tảng trực tuyến dành riêng cho phương tiện truyền thông xã hội như Facebook, LinkedIn và Twitter, trong đó cung cấp sự truy cập mới vào tương tác xã hội bằng cách cho phép người dùng kết nối trực tiếp với nhau qua Internet để giao tiếp và chia sẻ thông tin. Các mạng xã hội ngoại tuyến cũng có thể để lại dấu vết kỹ thuật số có thể phân tích được, chẳng hạn như trong các bản ghi siêu dữ liệu điện thoại ghi lại các cuộc điện thoại hay tin nhắn đã trao đổi và trong thời gian bao lâu.

Phân tích các mạng xã hội ngày càng có thể thực hiện được bởi các bộ sưu tập dữ liệu số ngày càng tăng để kết nối mọi người lại với nhau, đặc biệt là khi nó có tương quan với dữ liệu hoặc siêu dữ liệu ngoài cá nhân này. Các công cụ để thực hiện các phân tích như vậy đang được phát triển và có sẵn, được thúc đẩy một phần bởi số lượng ngày càng tăng của nội dung truyền thông xã hội có thể truy cập thông qua các giao diện lập trình ứng dụng mở đến các nền tảng mạng xã hội trực tuyến. Loại phân tích này là một vũ đài tích cực cho nghiên cứu. Phân tích mạng xã hội bổ sung cho phân tích cơ sở dữ liệu thông thường và một số các kỹ thuật được sử dụng (ví dụ, phân cụm trong các mạng liên kết) có thể được sử dụng trong cả hai bối cảnh.

Phân tích mạng xã hội có thể sẽ mạnh hơn vì sự liên kết dễ dàng của các loại đa dạng của thông tin (ví dụ, trộn dữ liệu đáng kể là có thể). Nó phù hợp với trực quan hóa các kết quả, trong đó hỗ trợ trong việc giải thích các kết quả phân tích. Nó có thể được sử dụng để

tìm hiểu về con người thông sự liên kết của họ với những người khác, trong một bối cảnh xu hướng của người liên kết với những người khác có một số điểm tương đồng với mình.

Phân tích mạng xã hội đang mang lại các kết quả có thể làm mọi người ngạc nhiên. Đặc biệt, nhận dạng đặc thù của một cá nhân là dễ dàng hơn so với chỉ từ phân tích cơ sở dữ liệu. Hơn nữa, nó có thể đạt được thông qua các loại dữ liệu đa dạng hơn so với nhiều người có thể hiểu được, góp phần vào sự xói mòn của tình trạng nặc danh. Cấu trúc mạng của một cá nhân là duy nhất và bản thân nó phục vụ như là một kỹ hiệu nhận dạng; việc xảy ra đồng thời trong thời gian và không gian là một phương tiện nhận dạng quan trọng; và các loại dữ liệu khác nhau có thể được kết hợp để thúc đẩy sự nhận dạng. Phân tích mạng xã hội được sử dụng trong điều tra pháp y hình sự để hiểu được các liên kết, phương tiện và động cơ của những người có thể đã phạm tội. Đặc biệt, phân tích mạng xã hội đã được sử dụng để hiểu rõ hơn về các mạng lưới khủng bố bí mật mà động cơ của chúng có thể khác với động cơ của các mạng công khai.

Trong lĩnh vực thương mại, một điều đã được biết từ lâu là những gì bạn bè của một người thích hay mua có thể ảnh hưởng đến những gì anh ta hoặc cô ta có thể mua. Ví dụ, trong năm 2010, một báo cáo cho thấy việc có một người bạn sở hữu iPhone sẽ làm cho khả năng sở hữu iphone của người đó tăng lên ba lần so với các loại điện thoại khác. Một người có hai người bạn sở hữu iPhone thì khả năng có thể có một chiếc iphone tăng lên 5 lần. Những tương quan như vậy xuất hiện trong phân tích mạng xã hội và có thể được sử dụng để giúp dự đoán xu hướng sản phẩm, các chiến dịch tiếp thị hướng tới các sản phẩm may mặc mà một cá nhân có thể có nhiều khả năng muốn có và nhằm vào các khách hàng (được cho là có "*giá trị mạng*" cao hơn) với một vai trò trung tâm (và một số lượng lớn ảnh hưởng) trong một mạng xã hội.

Có những căn bệnh thường lây lan qua tiếp xúc trực tiếp giữa các cá nhân (con người hoặc động vật), việc hiểu các mạng xã hội thông qua bất kì vật trung gian nào, có thể đưa ra khả năng tiếp xúc trực tiếp và do đó có thể hỗ trợ việc theo dõi và giám sát bùng nổ của bệnh. Một nghiên cứu mới đây của các nhà nghiên cứu trên Facebook đã phân tích mối quan hệ giữa vị trí địa lý của người dùng cá nhân và của bạn bè của họ. Từ phân tích này, họ đã có thể tạo ra một thuật toán để dự đoán vị trí của một người dùng cá nhân dựa trên vị trí của một số ít các bạn bè trong mạng của họ, chỉ đơn giản là nhìn vào địa chỉ IP của người dùng.

Có rất nhiều dịch vụ thương mại "*lắng nghe xã hội*", như Radian6/Salesforce Cloud, Collective Intellect, Lithium và những dịch vụ khác, khai phá dữ liệu từ mạng xã hội để sử dụng trong tình báo kinh doanh. Cùng với mạng xã hội, thông tin này có thể được sử dụng để đánh giá các thay đổi ảnh hưởng và sự lây lan của các xu hướng giữa các cá nhân và cộng đồng để thông báo các chiến lược tiếp thị.

2.2.3. Sử dụng dữ liệu

Mục đích cuối cùng của phân tích dữ liệu là để hỗ trợ việc ra quyết định tốt hơn, cho dù những quyết định này được thực hiện bởi một người điều hành trong một văn phòng, một robot trong nhà máy, hoặc một người nào đó ở nhà. Tự động hóa dựa vào dữ liệu có thể

đơn giản hóa các quyết định được thực hiện bởi các robot, trong khi thông tin được tổ chức sử dụng các hệ thống hỗ trợ ra quyết định, trực quan hóa dữ liệu và các công nghệ ánh xạ có thể hỗ trợ con người.

Hệ thống hỗ trợ ra quyết định

Hệ thống hỗ trợ ra quyết định là các công cụ tương tác giúp người sử dụng đưa ra các quyết định tốt hơn và nhanh hơn trong các môi trường phức tạp, đa biến. Hệ thống hỗ trợ ra quyết định sử dụng các mô hình và các mô phỏng để dự đoán các kết quả và sau đó đưa ra các khuyến nghị cho người ra quyết định. Ví dụ, một nhà quản lý xây dựng có thể sử dụng hệ thống hỗ trợ ra quyết định giúp chọn nhà thầu phụ có sự kết hợp tốt nhất giữa rủi ro và doanh thu cho một dự án nhất định.

Những hệ thống như vậy đặc biệt phổ biến ở các bệnh viện, nơi các hệ thống hỗ trợ ra quyết định lâm sàng có thể sử dụng thông tin của bệnh nhân để cảnh báo cho bác sĩ nếu một đơn thuốc ảnh hưởng đến các loại thuốc khác hay các bệnh khác. Các hệ thống hỗ trợ ra quyết định cũng có thể được sử dụng trong nhiều lĩnh vực khác, bao gồm cả giám sát môi trường. Ví dụ, hệ thống hỗ trợ ra quyết định cho an toàn hàng hải ở Địa Trung Hải đã được thiết kế cho các chính phủ thành viên của EU để giúp giảm thiểu những rủi ro tràn dầu ở Địa Trung Hải. Do các kỹ thuật phân tích dữ liệu như lập mô hình dự báo và xử lý ngôn ngữ tự nhiên tiếp tục phát triển, khả năng của các hệ thống hỗ trợ ra quyết định cũng phát triển theo.

Tự động hóa

Trong khi nhiều phân tích dữ liệu được triển khai để giúp con người đưa ra các quyết định chính xác hơn, dữ liệu cũng có thể được sử dụng để kích hoạt các hoạt động tự động trong hệ thống máy tính và robot. Ví dụ, Nest, máy điều nhiệt thông minh, có thể sử dụng dữ liệu cảm biến để xác định khi ngôi nhà có người và điều chỉnh hệ thống sưởi và làm mát của ngôi nhà một cách thích hợp. Xe ô tô tự lái của Google có thể nhận dữ liệu về các điều kiện đường sá và luồng giao thông để điều hướng hiệu quả và tránh va chạm. Một báo cáo năm 2013 của công ty nghiên cứu thị trường Markets and Markets dự đoán rằng thị trường giao tiếp máy-máy sẽ đạt 290 tỷ USD năm 2017, tăng 650% so với năm 2011.

Máy học, một ngành của khoa học máy tính liên quan đến các hệ thống có hiệu suất được cải thiện bằng việc bổ sung dữ liệu mới, cung cấp các phương pháp ra quyết định tự động trong một loạt các ứng dụng. Máy học đã được sử dụng rộng rãi trong khoa học người máy, chẳng hạn như thị giác máy tính và hoạt động tự động trong các môi trường nhà máy, cũng như trong các hệ thống khuyến nghị trực tuyến, chẳng hạn như những hệ thống được sử dụng bởi dịch vụ nhạc trực tuyến Spotify và trang web hẹn hò trực tuyến OKCupid.

Trực quan hóa

Một cách để các nhà khoa học dữ liệu có thể truyền tải phân tích của họ đến người ra quyết định là thông qua trực quan hóa. Trực quan hóa được sử dụng trong một loạt các lĩnh vực và có thể từ các đồ thị đường đơn giản giá cổ phiếu đến các sơ đồ mạng xã hội phức tạp cho thấy sự lây lan của bệnh dịch. Trong các trường hợp nơi các mẫu trong dữ

liệu có thể được xác định dễ dàng hơn khi dữ liệu được hiển thị, trực quan hóa cũng có thể được sử dụng để tiến hành phân tích dữ liệu. Trực quan hóa dữ liệu được đưa vào nhiều công cụ phần mềm phân tích kinh doanh, chẳng hạn như Tableau. Các nền tảng và ngôn ngữ chuyên dụng dành cho các ứng dụng cụ thể, chẳng hạn như Gephi cho mạng và hiển thị đồ thị và xử lý hiển thị tương tác. Ngôn ngữ lập trình Javascript rất phổ biến để các ứng dụng tùy chỉnh hiển thị dữ liệu, cung cấp các thư viện mã nguồn mở, được sử dụng rộng rãi như D3.

Các ứng dụng ánh xạ đã thúc đẩy sự phát triển rộng rãi phần mềm các hệ thống thông tin địa lý (GIS), cho phép các đặc trưng không gian được tích hợp vào phân tích dữ liệu. Có các công nghệ chuyên dụng cho tất cả các khía cạnh của đổi mới dựa vào dữ liệu không gian địa lý, bao gồm các cơ sở dữ liệu, máy chủ và các công cụ trực quan hóa. Các nhà cung cấp phần mềm độc quyền chính bao gồm ESRI (nhà cung cấp ArcGIS), Google (nhà cung cấp Google Maps, Earth và Street View) và Oracle (nhà cung cấp Spatial and Graph). Các dịch vụ GIS mã nguồn mở, chẳng hạn như những dịch vụ được công ty công nghệ không gian địa lý MapBox tạo ra, cũng đang phát triển ngày càng phổ biến. Các công cụ từ những nhà cung cấp trên đang được sử dụng rộng rãi trong ngành công nghiệp và chính phủ. Ví dụ, chính quyền Obama đã sử dụng phần mềm GIS để bổ sung thêm các lớp dữ liệu và tính tương tác vào các bản đồ trên trang web Recovery.gov của mình.

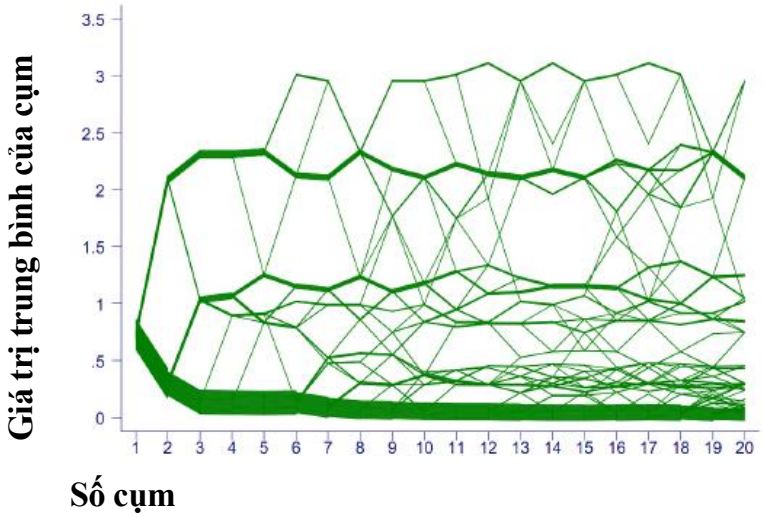
Trình bày thông tin theo cách mà mọi người có thể tiếp thu nó một cách hiệu quả là một thách thức quan trọng cần phải được đáp ứng nếu phân tích dữ liệu là để dẫn đến hành động cụ thể. Loài người đã tiến hóa để đạt hiệu quả cao trong nhận thức một số loại mô hình với các giác quan của mình nhưng vẫn tiếp tục phải đối mặt với những hạn chế đáng kể trong khả năng của bản thân để xử lý các loại dữ liệu khác như số lượng lớn các dữ liệu số hoặc văn bản. Vì lý do này, hiện nay có một lượng lớn nghiên cứu và đổi mới trong lĩnh vực trực quan hóa, ví dụ, các kỹ thuật và công nghệ được sử dụng để tạo ra các hình ảnh, sơ đồ, hoặc hình ảnh động để giao tiếp, hiểu và cải thiện kết quả của phân tích dữ liệu lớn. Dưới đây là một số ví dụ về lĩnh vực quan trọng và đang phát triển hỗ trợ dữ liệu lớn.

a) Đám mây từ khóa (Tag cloud)

Văn bản của một báo cáo hiển thị dưới hình thức một đám mây thẻ (từ khóa), có thể là một danh sách các từ được đánh giá mức độ quan trọng, trong đó các từ xuất hiện thường xuyên nhất được hiển thị lớn hơn và các từ ít xuất hiện thường xuyên hơn sẽ được hiển thị nhỏ hơn. Đây là cách trực quan giúp người đọc lĩnh hội nhanh chóng các khái niệm nổi bật nhất trong một văn bản dài.

b) Clustergram

Clustergram là một kỹ thuật trực quan hóa được sử dụng cho phân tích cụm, hiển thị các thành phần riêng của một tập dữ liệu được gán thành các cụm khi số lượng các cụm tăng lên. Sự lựa chọn số cụm là một tham số quan trọng trong phân tích cụm. Kỹ thuật này cho phép các nhà phân tích có được sự hiểu biết tốt hơn về cách các kết quả của cụm khác với số khác của các cụm.

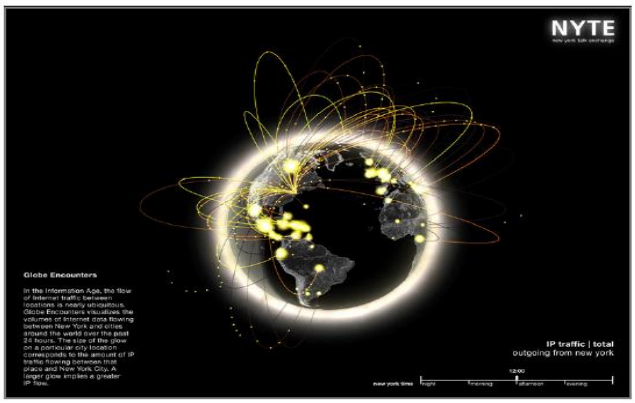


c) Dòng lịch sử

Dòng lịch sử là một kỹ thuật trực quan hóa lập các biểu đồ tiến hóa của một tài liệu khi nó được biên tập bởi nhiều tác giả. Thời gian nằm trên trên trục hoành, trong khi những đóng góp vào văn bản nằm trên trục tung; mỗi tác giả có một mã màu khác nhau và chiều dài của trục tung biểu thị số lượng văn bản được viết bởi mỗi tác giả. Bằng cách trực quan hóa lịch sử của một tài liệu theo cách này, những hiểu biết khác nhau dễ dàng xuất hiện.

d) Dòng thông tin không gian

Một kỹ thuật trực quan hóa khác là kỹ thuật mô tả các dòng thông tin không gian. Ví dụ chúng tôi chỉ ra ở đây có tên gọi New York Talk Exchange. Nó cho thấy lượng dòng dữ liệu của giao thức Internet giữa New York và các thành phố trên khắp thế giới. Kích thước của ánh sáng trên một vị trí thành phố cụ thể tương ứng với tổng lưu lượng IP lưu thông giữa các nơi đó và TP. New York; ánh sáng càng sáng hơn, dòng lưu thông càng lớn. Sự trực quang hóa này cho phép chúng ta xác định một cách nhanh chóng thành phố nào được kết nối chặt chẽ nhất với New York về khối lượng thông tin liên lạc của chúng.



2.2.4. Phổ biến dữ liệu

Các tổ chức, bao gồm các cơ quan chính phủ, thường muốn chia sẻ dữ liệu của họ với những tổ chức khác. Trước đây, các bộ dữ liệu thường được phổ biến thông qua các phương tiện số, chẳng hạn như đĩa CD, nhưng sử dụng các đối tượng vật lý để phổ biến có những hạn chế nhất định như khối lượng dữ liệu hạn chế, sự phân bổ chậm và tốn kém. Hiện nay, dữ liệu có trên các trang web, thường miễn phí trực tiếp cho người sử dụng. Một số tổ chức chỉ cung cấp quyền truy cập vào các tập dữ liệu thô; những tổ chức khác phát triển các giao diện lập trình ứng dụng để các nhà phát triển khác tái sử dụng dữ liệu của họ dễ dàng hơn.

Gần đây hơn, phần mềm chuyên dụng quản lý số lượng lớn các bộ dữ liệu mở của các tổ chức được xây dựng, chủ yếu là từ doanh nghiệp mới khởi nghiệp Socrata. Phần mềm này tương đối mới và các phần mềm khác cũng đã bắt đầu xuất hiện gần đây. Trong một số trường hợp, các tổ chức đã phát triển các nền tảng phổ biến dữ liệu mở của họ trong nội bộ tổ chức; một ví dụ là Data.gov của Hoa Kỳ. Những nhà sáng tạo ra nền tảng này sau đó phổ biến phần mềm của họ cho cộng đồng nguồn mở.

2.2.5. Cơ sở hạ tầng của dữ liệu lớn

Phân tích dữ liệu lớn đòi hỏi không chỉ các thuật toán và dữ liệu, mà còn cả các cơ sở vật chất, nơi lưu trữ và phân tích dữ liệu. Các dịch vụ an ninh liên quan được sử dụng đối với dữ liệu cá nhân cũng là một thành phần thiết yếu trong cơ sở hạ tầng. Trước đây loại cơ sở hạ tầng này thường chỉ thuộc về các tổ chức lớn, giờ đây nó có thể phổ biến đến các doanh nghiệp nhỏ và các cá nhân thông qua "*đám mây*". Khi mà phạm vi chia sẻ cơ sở hạ tầng phần mềm được mở rộng, thì các dịch vụ cơ sở hạ tầng bảo mật thông tin cá nhân cũng có thể được sử dụng dễ dàng hơn.

Các trung tâm dữ liệu

Một cách để nghĩ tới nền tảng dữ liệu lớn đó là cơ sở vật chất của các "trung tâm dữ liệu". Trong những năm gần đây, các trung tâm dữ liệu đã trở thành một loại hàng hóa gần như đạt chuẩn. Một trung tâm dữ liệu điển hình là một tòa nhà lớn, giống như kho chứa trên một nền bê tông kích thước bằng vài sân bóng đá. Nó được đặt ở vị trí có thể tiếp cận nguồn điện giá rẻ với kết nối cáp quang và kết nối trực tiếp với mạng xương sống Internet, thường là ở một vùng nông thôn hoặc biệt lập. Các trung tâm dữ liệu điển hình tiêu thụ 20-40 megawatt điện (tương đương với một thành phố 20.000-40.000 dân) và chứa đến hàng chục ngàn máy chủ và ổ đĩa cứng, với tổng số lên đến hàng chục petabytes. Trên thế giới, có khoảng 6000 trung tâm dữ liệu đạt quy mô này, Hoa Kỳ chiếm khoảng một nửa số này. Các trung tâm dữ liệu là vị trí cụ thể của dữ liệu lớn với mọi hình thức của nó. Các tập hợp dữ liệu lớn thường được sao chép tại nhiều trung tâm dữ liệu để nâng cao tính cả hiệu suất và độ chắc chắn. Hiện nay thị trường dịch vụ trung tâm dữ liệu đang phát triển nhanh.

Công nghệ phần mềm chuyên dụng cho phép các dữ liệu tại nhiều trung tâm dữ liệu

(và phân tán qua hàng chục ngàn bộ vi xử lý và ổ đĩa cứng) có thể tác hợp để thực hiện các nhiệm vụ phân tích dữ liệu, qua đó cho phép mở rộng quy mô và hiệu suất tốt hơn. Ví dụ, MapReduce (vốn là một công nghệ độc quyền của Google, nhưng giờ đây là một thuật ngữ được sử dụng tổng quát) là một mô hình lập trình về các hoạt động thực thi song song trên các bộ vi xử lý với số lượng gần như không giới hạn; Hadoop là một nền tảng lập trình mã nguồn mở phổ biến và là thư viện lập trình dựa trên những ý tưởng tương tự; NoSQL (Not Structured Query Language) là một tập hợp các công nghệ cơ sở dữ liệu, tháo gỡ nhiều giới hạn của các cơ sở dữ liệu truyền thống và "*quan hệ*", cho phép mở rộng tốt hơn trên nhiều bộ xử lý trong một hoặc nhiều trung tâm dữ liệu.

Nghiên cứu đương đại đang được nhằm vào thế hệ tiếp theo của Hadoop. Đại diện một nhánh là Accumulo, do Cơ quan An ninh Quốc gia Hoa Kỳ khởi xướng và chuyển tiếp thành cộng đồng mã nguồn mở Apache . Một ví dụ khác là Berkeley Data Analytics Stack, một nền tảng mã nguồn mở vượt trội Hadoop về phân tích dữ liệu từ nhiều bộ nhớ (memory-intensive) và được sử dụng bởi các công ty như Foursquare, Conviva, Klout, Quantifind, Yahoo, và Amazon Web Services. Đôi khi được gọi là "NoHadoop" (dịch chuyển từ SQL sang NoSQL), các công nghệ phù hợp với xu hướng này bao gồm Dremel của Google, MPI (thường được sử dụng trong siêu máy tính), Pregel (sử dụng cho đồ họa), và Cloudscale (phân tích thời gian thực).

Đám mây

Có thể hiểu "đám mây" như là một tập hợp các nền tảng và dịch vụ có thể thực hiện được nhờ vào việc thông dụng hóa vật chất các trung tâm dữ liệu. Khi nói rằng dữ liệu nằm "trong đám mây", không chỉ đề cập đến các ổ đĩa cứng cụ thể tồn tại (ở một nơi nào đó) với các dữ liệu, mà đó là cả một cơ sở hạ tầng phức tạp gồm các chương trình ứng dụng, phần mềm lớp trung gian (middleware), các giao thức mạng, và các mô hình kinh doanh cho phép dữ liệu được đăng nhập, truy cập, và sử dụng, tất cả với chi phí phân phối cạnh tranh. Các tổ chức thương mại cung cấp đám mây tồn tại trong một hệ sinh thái có nhiều cấp thứ bậc và nhiều mô hình giá trị gia tăng khác nhau cùng tồn tại. Ở đây có nhiều cách chuyển giao trách nhiệm giữa người dùng cuối và các trung tâm dữ liệu cụ thể.

Các nhà cung cấp đám mây hiện nay mang lại một số lợi ích an ninh (và thông qua đó, lợi ích bảo mật) so với các trung tâm dữ liệu thông thường của các doanh nghiệp trước đây hay các máy tính của các doanh nghiệp nhỏ. Các dịch vụ có thể bao gồm bảo vệ và giám sát tốt hơn, cũng như hỗ trợ tập trung hóa nhân lực, đào tạo, và giám sát. Các dịch vụ đám mây cũng đặt ra nhiều thách thức mới về an ninh, một đối tượng nghiên cứu hiện nay. Cả lợi ích và rủi ro đều xuất phát từ sự tập trung hóa các nguồn lực: Thêm nhiều dữ liệu được một tổ chức cụ thể nắm giữ (mặc dù phân bố trên nhiều máy chủ hoặc các trang web), và một nhà cung cấp đám mây có thể thực hiện tốt hơn so với các trung tâm dữ liệu được tổ chức riêng biệt bằng cách áp dụng các tiêu chuẩn cao về tuyển dụng và quản lý con người và hệ thống.

Việc sử dụng đám mây và các tương tác cá nhân cùng với nó (bất kể cố ý hay không)

được dự báo sẽ tăng mạnh trong những năm tới. Sự gia tăng của cả hai ứng dụng di động, tăng cường sử dụng điện thoại di động và máy tính bảng như là nền tảng, và các bộ cảm biến phân bổ rộng có liên quan với việc sử dụng ngày càng tăng của các hệ thống đám mây để lưu trữ, xử lý, và các tác nghiệp dựa trên thông tin khác đóng góp bởi các thiết bị phân tán. Mặc dù sự tiến bộ về môi trường di động cải thiện khả năng sử dụng các ứng dụng đám mây di động, tuy nhiên nó có thể gây phương hại đến tính riêng tư đến mức nó có thể che giấu hiệu quả hơn sự trao đổi thông tin từ người sử dụng. Khi có thêm tính năng di động lõi được chuyển sang đám mây, một lượng lớn thông tin sẽ được trao đổi, và người dùng có thể ngạc nhiên bởi bản chất của thông tin không còn cục bộ hóa trong điện thoại di động của mình. Ví dụ, màn hình hiển thị (screen rendering) dựa trên đám mây (hoặc "màn hình ảo hóa") cho điện thoại di động sẽ có nghĩa là hình ảnh hiển thị trên màn hình điện thoại di động trên thực tế sẽ được tính toán trên đám mây và truyền đến thiết bị di động. Điều đó có nghĩa là tất cả các hình ảnh trên màn hình của thiết bị di động đều có thể truy cập và thao tác từ đám mây.

Kiến trúc đám mây cũng đang được sử dụng ngày càng tăng để hỗ trợ phân tích dữ liệu lớn, cả các doanh nghiệp lớn (như Google, Amazon, eBay) và các doanh nghiệp nhỏ hay cá nhân, những người sử dụng đột xuất hay thường xuyên các nền tảng đám mây công cộng (như Amazon Web Services, Google Cloud Platform, Microsoft Azure) thay cho việc mua sắm cơ sở hạ tầng riêng. Các dịch vụ truyền thông xã hội như Facebook và Twitter đang được triển khai và phân tích bởi các nhà cung cấp thông qua sử dụng các hệ thống đám mây. Các dịch vụ này đại diện cho một dạng dân chủ hóa phân tích, có tiềm năng tạo điều kiện thuận lợi cho các doanh nghiệp mới và nhiều hơn. Triển vọng tương lai bao gồm khám phá các phương án hợp nhất hoặc kết nối các ứng dụng đám mây và làm giảm một số không đồng nhất trong các giao diện lập trình ứng dụng cho các ứng dụng đám mây.

3.3. Các vấn đề chính sách để khai thác đổi mới dựa sáng tạo trên dữ liệu như một nguồn lực tăng trưởng mới

3.3.1. Các thách thức chính sách đặt ra đối với đổi mới sáng tạo dựa trên dữ liệu

Chính phủ có một vai trò quan trọng trong việc thúc đẩy các điều kiện thuận lợi cho đổi mới sáng tạo dựa trên dữ liệu diễn ra trong một môi trường đáng tin cậy. Sau đây là các thách thức chính sách đã được xác định cho đến nay:

Xem xét toàn bộ vòng đời giá trị của dữ liệu

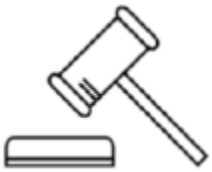
Việc thiết kế các chính sách hiệu quả để thúc đẩy đổi mới dựa trên dữ liệu, trong khi giảm thiểu rủi ro, đòi hỏi một sự hiểu biết cơ bản về quá trình tạo ra giá trị. Một số chính sách (như truy cập mở đến dữ liệu) sẽ ảnh hưởng đến các giai đoạn cụ thể của vòng đời giá trị của dữ liệu trong khi những chính sách khác (ví dụ như bảo mật riêng tư) sẽ có tác động đến toàn bộ vòng đời giá trị. Việc xem xét toàn bộ vòng đời giá trị của dữ liệu là rất quan trọng vì nhiều lĩnh vực chính sách bổ sung cho nhau. Nói cách khác, việc tập trung chỉ vào một lĩnh vực chính sách sẽ có tác động rất ít nếu không được hỗ trợ bởi các biện

pháp chính sách bổ sung. Ví dụ, việc thúc đẩy truy cập mở trong một nền kinh tế mà không thúc đẩy các kỹ năng phân tích dữ liệu và tinh thần kinh doanh liên quan đến dữ liệu sẽ không đưa đến những lợi ích đầy đủ của đổi mới sáng tạo dựa trên dữ liệu trong phạm vi quốc gia.

Hình 8: Các giai đoạn chính của vòng đời giá trị của dữ liệu và các vấn đề chính sách

Dữ liệu hóa và thu thập dữ liệu

Phân tích dữ liệu



Quyết định dựa trên dữ liệu

Mở rộng lượng dữ liệu có thể thu thập được	Tăng cường năng lực phân tích dữ liệu	Thúc đẩy việc ra quyết định có trách nhiệm cho tăng trưởng và sự phồn thịnh
- Dữ liệu mở - Tính di chuyển của dữ liệu	- Hạ tầng điện toán đám mây - Nhà khoa học về dữ liệu	- Năng lực trong các lĩnh vực cụ thể - Tinh thần kinh doanh

Các vấn đề chính sách liên quan: Tính riêng tư, quyền sở hữu trí tuệ, cạnh tranh, thuế, thương mại...

Bảo vệ hiệu quả sự riêng tư và quyền tự do cá nhân

Việc sợ mất quyền tự chủ và tự do có thể tạo ra một phản ứng mạnh đối với đổi mới dựa trên dữ liệu, dẫn đến sự ít tham gia hơn của các cá nhân và sự miễn cưỡng đóng góp dữ liệu cá nhân, nguồn dữ liệu cần thiết cho đổi mới dựa trên dữ liệu. Do đó, việc bảo vệ hiệu quả sự riêng tư là một điều kiện quan trọng để duy trì lòng tin vào đổi mới dựa trên dữ liệu. Chính phủ nên khuyến khích việc bảo vệ hiệu quả sự riêng tư khi xem xét toàn bộ vòng đời giá trị của dữ liệu, từ sưu tập dữ liệu, đến phân tích dữ liệu, đến ra quyết định dựa trên dữ liệu. Các biện pháp sau đây có thể được áp dụng: (i) tăng cường thực tiễn phân tích dữ liệu minh bạch, (ii) tiếp cận tốt hơn và trao quyền cho các chủ thể dữ liệu (chủ thể dữ liệu là cá nhân mà dữ liệu có liên quan đến họ), (iii) thúc đẩy việc sử dụng dữ

liệu có trách nhiệm của những người kiểm soát dữ liệu (người kiểm soát dữ liệu là người hoặc một mình hoặc cùng với những người khác xác định mục tiêu và cách thức tổ chức hoặc xử lý dữ liệu của cá nhân) và (iv) thúc đẩy quản lý rủi ro về quyền riêng tư bao gồm tất cả các bên liên quan.

Thúc đẩy văn hóa quản lý rủi ro kỹ thuật số trên toàn hệ sinh thái dữ liệu

Phương pháp bảo đảm an ninh truyền thống có thể hạn chế việc hiện thực hóa các lợi ích của đổi mới dựa trên dữ liệu. Chính phủ cần thúc đẩy văn hóa quản lý rủi ro an ninh kỹ thuật số trong đó yêu cầu những người kiểm soát dữ liệu và các nhà ra quyết định hiểu được cách thức để tiếp cận an ninh trong một bối cảnh kỹ thuật số để phục vụ tốt nhất các mục tiêu kinh tế và xã hội của họ.

Việc đẩy mạnh văn hóa quản lý rủi ro thường gắn liền với sự hiểu biết về chu trình quản lý rủi ro an ninh kỹ thuật số bao gồm các bước sau: đánh giá rủi ro (bước 1) và xử lý rủi ro (bước 2), tức là xác định xem cần phải xử lý rủi ro như thế nào (bước 3), giảm thiểu rủi ro (bước 4), chuyển nó cho người khác (ví dụ như thông qua hợp đồng, bảo hiểm hay thoả thuận hợp pháp khác) (bước 5) hoặc tránh rủi ro bằng cách không thực hiện hoạt động này (bước 6). Nếu một người quyết định giảm thiểu rủi ro, việc đánh giá rủi ro giúp xác định các biện pháp an ninh cần được lựa chọn và áp dụng ở đâu và khi nào, đứng trên góc độ của các hậu quả của các sự kiện không chắc chắn đối với các mục tiêu kinh tế và xã hội (bước 7). Cuối cùng, rủi ro còn lại không thể bỏ qua. Một kế hoạch được chuẩn bị (bước 8) cũng cần được thiết lập để hạn chế và quản lý các hậu quả của các sự cố khi chúng xảy ra và làm giảm khả năng leo thang.

Cung cấp các ưu đãi cho Internet tốc độ cao và mở

Sự phổ biến nhanh chóng băng thông rộng ở các quốc gia OECD và các nền kinh tế đối tác của nó là một trong những hỗ trợ cơ bản nhất cho đổi mới dựa trên dữ liệu. Băng thông rộng tốc độ cao, và đặc biệt băng thông rộng di động, là cơ sở hạ tầng cơ bản cho dòng dữ liệu tự do và trao đổi được thu thập từ xa thông qua các ứng dụng Internet và hiện nay thông qua các thiết bị thông minh ngày càng nhiều và kết nối với nhau tạo thành Internet vạn vật. Hơn nữa, tính chất toàn cầu và phân bổ của hệ sinh thái dữ liệu làm cho Internet mở là một điều kiện quan trọng cho đổi mới dựa trên dữ liệu.

Chính phủ cần tiếp tục thúc đẩy băng thông rộng di động và hỗ trợ mối quan tâm chung để tìm sự đồng thuận về cách duy trì Internet mở và sôi động. Hội nghị Cấp cao của OECD về Nền kinh tế Internet diễn ra ngày 28-29/6/2011 đã thảo luận về tính mở của Internet và cách tốt nhất để đảm bảo sự tăng trưởng liên tục và đổi mới nền kinh tế Internet. Thông cáo kết quả dự thảo, đưa đến Khuyến nghị về các nguyên tắc cho hoạch định chính sách Internet, bao gồm một số nguyên tắc cơ bản cho hoạch định chính sách Internet với mục tiêu để đảm bảo cho Internet duy trì mở và năng động, "*cho phép mọi người nói lên khát vọng dân chủ của mình và bất kỳ hoạch định chính sách nào liên quan đến nó cũng phải thúc đẩy tính mở và được đặt nền tảng trên sự tôn trọng nhân quyền và các quy định của pháp luật*". Bốn nguyên tắc đầu tiên sau đây rất phù hợp cho việc sử

dụng dữ liệu. Điều này không có nghĩa là các nguyên tắc khác không quan trọng đối với đổi mới sáng tạo dựa trên dữ liệu:

(1) Thúc đẩy và bảo vệ luồng thông tin tự do toàn cầu: Nền kinh tế Internet, cũng như khả năng học tập của mỗi cá nhân, chia sẻ thông tin và kiến thức, thể hiện bản thân, tập hợp và lập hội, phụ thuộc vào luồng thông tin tự do toàn cầu. Để khuyến khích các luồng thông tin tự do trực tuyến, làm việc cùng nhau để thúc đẩy khả năng tương thích toàn cầu tốt hơn trên một tập hợp đa dạng các luật và quy định là rất quan trọng. Trong khi thúc đẩy luồng thông tin tự do, các chính phủ cũng cần hướng tới việc bảo vệ tốt hơn các dữ liệu của các cá nhân, trẻ em, người tiêu dùng, các quyền sở hữu trí tuệ và giải quyết các vấn đề an ninh mạng. Để thúc đẩy luồng thông tin tự do, chính phủ cũng nên tôn trọng các quyền cơ bản.

(2) Thúc đẩy tính mở, bản chất phân tán và liên kết của Internet: Là một mạng phi tập trung của các mạng máy tính, Internet đã đạt được sự kết nối toàn cầu mà không thuộc sự phát triển của bất cứ cơ chế quản lý quốc tế nào. Sự phát triển của một cơ chế quản lý chính thức như vậy có thể hủy hoại sự phát triển của nó. Tính mở của Internet đối với các thiết bị, các ứng dụng và dịch vụ mới đóng một vai trò quan trọng trong sự thành công của nó trong việc thúc đẩy đổi mới, sáng tạo và tăng trưởng kinh tế. Tính mở này bắt nguồn từ sự tương tác liên tục phát triển và sự độc lập giữa các thành phần kỹ thuật khác nhau của Internet, cho phép hợp tác và đổi mới trong khi tiếp tục hoạt động độc lập với nhau. Sự độc lập này cho phép những thay đổi chính sách và quy định trong một số thành phần mà không cần những thay đổi ở những thành phần khác hoặc có tác động đối với đổi mới và hợp tác. Tính mở của Internet cũng bắt nguồn từ sự chấp nhận trên toàn cầu các tiêu chuẩn kỹ thuật hỗ trợ các thị trường sản phẩm và truyền thông toàn cầu. Việc duy trì tính trung lập của công nghệ và chất lượng phù hợp cho tất cả các dịch vụ Internet cũng rất quan trọng để đảm bảo một môi trường Internet mở và năng động. Cung cấp dịch vụ truy cập Internet mở là rất quan trọng cho nền kinh tế Internet.

(3) Thúc đẩy đầu tư và cạnh tranh trong các dịch vụ và mạng tốc độ cao: Dịch vụ và mạng tốc độ cao cần thiết cho sự tăng trưởng kinh tế trong tương lai, tạo việc làm, năng lực cạnh tranh cao hơn và để mọi người được hưởng một cuộc sống tốt hơn. Các chính sách công cần thúc đẩy cạnh tranh mạnh mẽ trong việc cung cấp Internet băng thông rộng tốc độ cao cho người dùng với giá cả phải chăng và thúc đẩy đầu tư để đạt được độ bao phủ địa lý lớn nhất của Internet băng thông rộng. Các chính sách công cũng cần thúc đẩy mức đầu tư tốt nhất bằng cách tạo ra nhu cầu đối với các mạng và dịch vụ băng thông rộng tốc độ cao, đặc biệt là trong các lĩnh vực nơi chính phủ đóng vai trò quan trọng như trong giáo dục, y tế, phân phối năng lượng và giao thông vận tải. Chính sách công sẽ giúp thúc đẩy sự đa dạng của nội dung, các nền tảng, các ứng dụng, các dịch vụ trực tuyến và các công cụ truyền thông của người dùng khác sẽ tạo ra nhu cầu cho các mạng và dịch vụ, cũng như cho phép người dùng được hưởng lợi đầy đủ từ các mạng và dịch vụ này và truy cập vào sự đa dạng của nội dung mà không có phân biệt đối xử, bao gồm các nội dung

văn hóa và ngôn ngữ theo lựa chọn.

(4) Đẩy mạnh và cho phép chuyển giao dịch vụ xuyên biên giới: Các nhà cung cấp cần có khả năng cung cấp các dịch vụ xuyên Internet qua biên giới và trung lập về mặt công nghệ theo cách thúc đẩy khả năng tương tác của các dịch vụ và công nghệ, ở nơi thích hợp. Người sử dụng cần có khả năng truy cập và tạo ra nội dung hợp pháp và chạy các ứng dụng theo sự lựa chọn của họ. Để đảm bảo hiệu quả về chi phí và các hiệu quả khác, các rào cản đối với vị trí, sự tiếp cận và việc sử dụng các công cụ dữ liệu và các chức năng xuyên biên giới cần được giảm thiểu, việc cung cấp các biện pháp bảo vệ dữ liệu và an ninh dữ liệu thích hợp được thực hiện một cách phù hợp và phản ánh sự cân bằng cần thiết giữa tất cả các quyền, quyền tự do và các nguyên tắc cơ bản.

Khuyến khích việc tiếp cận đến dữ liệu và luồng dữ liệu tự do qua biên giới của quốc gia và tổ chức

Luồng dữ liệu tự do qua biên giới của quốc gia và tổ chức là một nhân tố hỗ trợ quan trọng cho đổi mới dựa trên dữ liệu. Chính phủ nên khuyến khích sự tiếp cận tốt hơn với luồng dữ liệu tự do trên toàn bộ nền kinh tế. Điều này không chỉ bao gồm việc tăng cường tiếp cận và tái sử dụng dữ liệu của khu vực công, những lợi ích đáng kể được dự kiến có thể thu được từ việc chia sẻ dữ liệu xuyên khu vực. Điều này có thể thực hiện được thông qua việc thúc đẩy các dữ liệu mở và dữ liệu dùng chung một cách phổ thông hơn. Theo Frischmann (2012), dữ liệu dùng chung có thể: (i) tạo điều kiện cho việc sản xuất liên doanh hoặc hợp tác với các nhà cung cấp, các khách hàng hay thậm chí các đối thủ cạnh tranh, (ii) hỗ trợ và khuyến khích đổi mới dựa vào người sử dụng bao gồm các hoạt động tạo ra giá trị của người sử dụng (bao gồm cả người tiêu dùng và công dân), (iii) tối đa hóa giá trị tùy chọn của dữ liệu khi các đầu tư vào dữ liệu là không thể đảo ngược và có sự không chắc chắn cao về các nguồn lực của giá trị thị trường trong tương lai. và cuối cùng nhưng không kém phần quan trọng là (iv) trợ cấp (chéo) một cách hiệu quả cho việc sản xuất hàng hóa xã hội và công cộng mà không cần phải dựa vào thị trường hay các chính phủ để "*chọn người chiến thắng*".

Dữ liệu mở là chế độ chia sẻ dữ liệu mạnh mẽ nhất. Các chế độ khác tồn tại giữa dữ liệu mở và dữ liệu đóng, với các yếu tố chính ảnh hưởng đến mức độ mở của gồm: (i) thiết kế công nghệ (bao gồm dữ liệu trên web, có thể đọc được bằng máy và khả năng liên kết), (ii) quyền sở hữu trí tuệ (bao gồm các chế độ pháp lý như bản quyền, các hình thức sở hữu trí tuệ đối với cơ sở dữ liệu và các bí mật thương mại) và (iii) sự định giá.

Việc trao quyền cho các cá nhân (người tiêu dùng) thông qua khả năng mang theo dữ liệu (data portability) có thể tiếp tục thúc đẩy luồng dữ liệu tự do qua biên giới quốc gia và tổ chức. Dữ liệu được phân loại theo (i) dữ liệu đóng góp (contributed data), (ii) dữ liệu quan sát (observed data) và (iii) dữ liệu ngoại suy (inferred data) có thể giúp các nhà hoạch định chính sách thiết kế các cơ chế thích hợp để cân bằng các quyền cá nhân với lợi ích hợp pháp của doanh nghiệp.

Thiết lập các khuôn khổ quản trị dữ liệu cho truy cập, chia sẻ và khả năng liên tác của dữ liệu

Các chế độ quản trị dữ liệu có thể có một tác động đối với việc truy cập, chia sẻ và tính liên tác (interoperability) của dữ liệu. Chúng bao gồm những thách thức mà các cá nhân, doanh nghiệp và các nhà hoạch định chính sách phải đối mặt trong mọi lĩnh vực, trong đó dữ liệu được sử dụng mà không phân biệt các loại dữ liệu. Các chế độ quản trị dữ liệu có thể có tác động đối với các khuyến khích chia sẻ và tiềm năng của dữ liệu được sử dụng theo cách thức liên tác. Các yếu tố được xem xét cho một chế độ quản trị dữ liệu hiệu quả bao gồm:

- Giá trị và định giá dữ liệu
- Liên kết và tích hợp dữ liệu
- Chất lượng và xử lý dữ liệu
- Quyền sở hữu và kiểm soát dữ liệu

Thúc đẩy nghiên cứu và phát triển các công nghệ phân tích dữ liệu và tăng cường bảo vệ quyền riêng tư

Chất lượng của những hiểu biết dựa vào dữ liệu phụ thuộc vào chất lượng của các thuật toán được sử dụng để phân tích dữ liệu (bên cạnh việc lựa chọn thuật toán phù hợp và chất lượng của dữ liệu). Đồng thời, kiến thức về các cơ chế được sử dụng để trích xuất thông tin làm phong phú cho nghiên cứu về các cơ chế bảo vệ và kiểm soát tốt hơn việc khai thác thông tin. Vì vậy, NC&PT trong phân tích dữ liệu có thể được tiến hành đồng thời với NC&PT các công nghệ bảo vệ quyền riêng tư (privacy enhancing technologies-PET). Tuy nhiên, bằng chứng cho thấy rằng các động cơ khuyến khích khu vực tư nhân tiến hành NC&PT về phân tích dữ liệu là nhiều hơn so với PET. Ví dụ, số lượng đơn xin cấp bằng sáng chế về các công nghệ PET liên quan đến bảo vệ sự riêng tư vẫn còn rất thấp và thậm chí đã giảm trong năm 2011, trong khi đơn xin cấp bằng sáng chế liên quan đến phân tích dữ liệu liên tục tăng. Vì vậy chính phủ cần thúc đẩy NC&PT không chỉ tập trung vào phân tích dữ liệu mà còn tập trung vào các công nghệ PET.

Đảm bảo việc cung cấp và phát triển các kỹ năng và năng lực phân tích dữ liệu

Việc gặt hái những lợi ích đầy đủ của dữ liệu đòi hỏi một mức độ đủ cao năng lực phân tích dữ liệu trong nền kinh tế và xã hội. Bên cạnh việc cung cấp các công cụ điện toán đám mây và phân tích dữ liệu, cần thiết phải nâng cao các kỹ năng phân tích dữ liệu (nhà khoa học dữ liệu). Các kỹ năng và năng lực cụ thể về cách giải thích và tận dụng tối đa các kết quả phân tích dữ liệu cũng quan trọng. Chính phủ cần đảm bảo việc cung cấp và phát triển các kỹ năng và năng lực phù hợp thông qua (i) các tổ chức giáo dục chính thức và (ii) đào tạo tại chỗ và đào tạo nghề công nghệ thông tin và truyền thông.

Khuyến khích tinh thần khởi nghiệp doanh nghiệp dựa vào dữ liệu và thay đổi tổ chức trên toàn bộ nền kinh tế

Đổi mới dựa trên dữ liệu muốn đạt được một mức độ lớn phải được thực hiện bởi các

nhà doanh nhân, họ nhận thức được tiềm năng của phân tích dữ liệu trong các tổ chức của mình cũng như trong các thị trường khác.

Đối với các doanh nhân trong một tổ chức, những thách thức chính sẽ là thay đổi tổ chức: Chuyển đổi từ một tổ chức truyền thống sang tổ chức dựa trên dữ liệu có thể đòi hỏi sự thay đổi văn hóa có thể rất khó để thực hiện. Như Bakhshi et al. (2014) nhấn mạnh: Thực hiện những thay đổi bổ sung để gặt hái lợi nhuận đầy đủ từ phân tích dữ liệu có thể "bao gồm những thay đổi gây phá vỡ, do đó có thể gây tranh cãi trong các cơ cấu tổ chức và quy trình kinh doanh".

Chính phủ có thể đóng một vai trò quan trọng trong việc khuyến khích các doanh nghiệp dựa vào dữ liệu và thay đổi tổ chức thông qua việc cung cấp các thực tiễn tốt nhất và khuyến khích cung cấp vốn mạo hiểm.

Kết luận

Khuyến nghị các lĩnh vực chính sách công hỗ trợ đổi mới sáng tạo dựa vào dữ liệu

Cơ hội kinh tế của đổi mới sáng tạo dựa vào dữ liệu là rất lớn. Như OECD đã kết luận, "sự gia tăng độ lớn, tốc độ và đa dạng dữ liệu được sử dụng trên toàn bộ nền kinh tế, và quan trọng hơn là giá trị kinh tế và xã hội lớn hơn của nó, báo hiệu một sự thay đổi hướng tới một mô hình kinh tế xã hội định hướng dữ liệu. Trong mô hình này, dữ liệu là tài sản cốt lõi có thể tạo ra lợi thế cạnh tranh và chi phối đổi mới, tăng trưởng và phát triển bền vững". Sự tăng trưởng về số lượng dữ liệu được tạo ra trên cơ sở hàng ngày đến nay đã vượt quá bất kỳ một sự hiểu biết tiềm năng nào về độ lớn của nó. Một ước tính gần đây đã đưa ra con số 161 exabytes một năm - hay tương đương với khối lượng thông tin được lưu trữ tại 37.000 thư viện có độ lớn tương đương Thư viện Quốc hội Hoa Kỳ. Với độ lớn như vậy, tiềm năng kinh tế và xã hội là vô cùng to lớn.

Giá trị từ phân tích dữ liệu có thể tính toán trong điều kiện kinh tế thực. Chi tiêu cho cơ sở hạ tầng CNTT để phân tích dữ liệu theo ước tính của Gartner đạt 37 tỉ USD vào năm 2013. Cũng báo cáo này chỉ ra rằng vào năm 2015, đổi mới sáng tạo dựa vào dữ liệu sẽ tạo ra được 4,4 triệu việc làm IT trên toàn cầu.

Việc hiểu được giá trị có thể nắm bắt được từ sự đổi mới sáng tạo dựa vào dữ liệu là điều quan trọng bởi chính bản thân dữ liệu không có giá trị sẵn có. Khối lượng dữ liệu được tạo ra thường gây nhầm lẫn hoặc đặt không đúng chỗ và làm chệch hướng các cuộc tranh luận chú trọng vào các vấn đề về độ lớn hơn là phân tích. Như Hilbert đã lập luận, "không phụ thuộc vào tầm cỡ độ lớn ở mức Peta, Exa, hoặc zettabyte, đặc điểm then chốt của sự thay đổi mô hình này chính là việc xử lý phân tích dữ liệu được đặt ra ở vị trí hàng đầu của việc ra quyết định trí tuệ". Các số liệu thống kê kinh tế chỉ là những đại diện cho giá trị mà đổi mới dựa vào dữ liệu tạo ra. Nhiều hiệu quả của thông tin số không thể nắm bắt bằng các phép đo kinh tế truyền thống như GDP hay GVA. Chỉ có thể thông qua phân tích, kết hợp các sản phẩm hoặc dịch vụ mới làm cho núi dữ liệu khổng lồ tạo ra giá trị hoặc hiệu quả cho xã hội.

Giá trị từ đổi mới dựa vào dữ liệu không dành riêng cho khu vực nhà nước hay tư nhân. Eric Byrnjolfsson phát hiện rằng các doanh nghiệp áp dụng việc ra quyết định dựa trên dữ liệu thì nâng cao được sản lượng và năng suất lên từ 5-6%. Tương tự, các chính phủ có thể cải thiện được các dịch vụ mà họ cung cấp cho công dân bằng cách mang đến các kỹ năng và kỹ thuật để xử lý những dữ liệu riêng của mình. Ngoài ra còn có một áp lực ngày càng tăng đối với các chính phủ để thực hiện các chính sách dựa trên bằng chứng; để tuân theo quy luật rằng "những gì đo đếm được thì được cải tiến". Điều này đòi hỏi không chỉ thu thập dữ liệu bổ sung mà còn phải xử lý nó. Đó không phải là chỉ chính phủ có thể có ý tưởng về cách sử dụng các dữ liệu thu thập được như thế nào. Dữ liệu còn giúp tiết kiệm tiền: các chính phủ thuộc EU có thể giảm chi phí hành chính 15-20%, giá trị tương đương 150-300 tỷ euroo.

Hiện nay, ngày càng có nhiều chính phủ công bố các bộ dữ liệu mở để thúc đẩy đổi mới sáng tạo trong công chúng. Cho dù đó là việc công khai các lịch trình giao thông công cộng để cho các nhà phát triển ứng dụng sáng tạo các sản phẩm tiêu dùng mới hay sự gia tăng tính minh bạch trong các dịch vụ công bằng cách mở cửa dữ liệu cho các tổ chức phi chính phủ, thì các cơ hội cho các tổ chức thuộc khu vực công có ý nghĩa rất quan trọng.

Các cơ hội mang lại là cả về kinh tế lẫn xã hội. Các bệnh viện và hệ thống y tế có thể chữa bệnh và khắc phục các rủi ro hệ thống thông qua đổi mới dựa trên dữ liệu; các trường học có thể phân tích xem học sinh tương tác như thế nào với tài liệu giảng dạy để nâng cao kết quả giáo dục; việc bố trí các nguồn lực được phân bổ hiệu quả hơn thông qua sử dụng phân tích dữ liệu. Thật sự khả năng là vô tận, chỉ cần chúng ta có nền tảng và kỹ năng để phân tích các kho dữ liệu được sản sinh và thu thập.

Tất cả các cơ hội kinh tế và xã hội đó cũng tạo ra những nguy hiểm và rủi ro, vì vậy chúng cần được phân tích và phản ứng thận trọng. Thách thức đầu tiên đó là đảm bảo rằng thông tin cá nhân không bị tiết lộ dù vô tình hay bất đắc dĩ thông qua việc chia sẻ các tập hợp dữ liệu. Những mối quan tâm đó cần được giải quyết và các rủi ro cần được giảm thiểu trước nhằm duy trì niềm tin của công chúng trong sử dụng các dịch vụ kỹ thuật số và để xã hội có thể tận dụng được những lợi thế mà đổi mới sáng tạo dựa trên dữ liệu có thể mang lại. Điều này có thể mang lại lợi ích cho các cá nhân cũng như cho xã hội nói chung và vì thế cách tiếp cận của các nhà hoạch định chính sách phải là một tập hợp các quy định hỗ trợ chứ không phải là những cấm đoán.

Do khu vực tư nhân sẽ thực hiện nhiều nỗ lực tiên phong trong sử dụng và phân tích dữ liệu, các chính phủ có thể và nên hỗ trợ cho những nỗ lực đó. Đặc biệt, đổi mới dựa vào dữ liệu đòi hỏi một lực lượng lao động có kỹ năng, công nghệ tiên tiến và sự tiếp cận dữ liệu. Các nhà hoạch định chính sách có thể hỗ trợ những nỗ lực đó bằng cách xem xét các cơ hội chính sách công trong bối cảnh khu vực công là một trong những nơi có cường độ sử dụng dữ liệu cao nhất trong nền kinh tế. Các lĩnh vực chính sách công cần chú trọng để hỗ trợ cho đổi mới sáng tạo dựa trên dữ liệu gồm:

Nhân lực

Hiện tại, thế giới còn thiếu nhân lực có kiến thức, kỹ năng và năng lực để hỗ trợ đổi mới

dựa vào dữ liệu. Nguồn nhân lực này không chỉ bao gồm các nhà lập trình có kỹ năng về học máy và Hadoop, mà còn bao gồm các nhà quản lý, các nhà thiết kế và các chuyên gia truyền thông. Ví dụ, năm 2012, công ty phân tích thị trường Gartner dự tính đến năm 2015, chỉ có một phần ba trong số 4,4 triệu việc làm trong lĩnh vực dữ liệu lớn sẽ được tuyển dụng. Trong khi một số trường đại học gần đây đã bắt đầu đưa các chương trình khoa học dữ liệu, phân tích kinh doanh và học máy vào chương trình đào tạo, những nỗ lực này có thể không đáp ứng nhanh chóng được các nhu cầu trước mắt.

Các quốc gia có thể cung cấp nhân tài làm việc trong các lĩnh vực liên quan đến dữ liệu sẽ có lợi thế trong nền kinh tế toàn cầu. Các nhà hoạch định chính sách có cơ hội để giúp thúc đẩy sự tăng trưởng số nhân lực có kiến thức về dữ liệu bằng cách tài trợ cho các khóa học mở, trực tuyến về các môn học liên quan đến dữ liệu và mở rộng tuyển sinh các lớp thống kê và khoa học máy tính. Các trường trung học cũng có thể hỗ trợ bằng cách tạo ra các yêu cầu về toán linh hoạt hơn, do đó học sinh có thể tham dự các khóa học khoa học máy tính hay thống kê. Mặc dù những nỗ lực như vậy chắc chắn phải mất một thời gian để đem lại kết quả nhưng chúng có thể giúp mở ra những cơ hội mới cho người lao động và mở rộng sự sẵn có của nhân lực đa ngành có kiến thức về dữ liệu cho các công ty về dài hạn.

Chính phủ cũng có thể giúp thúc đẩy sự phát triển vốn nhân lực cần thiết bằng cách trở thành người đi đầu, chứ không phải là người tụt hậu, trong việc thực hiện đổi mới dựa vào dữ liệu. Các cơ quan chính phủ có thể sử dụng dữ liệu để tiết kiệm tiền bạc và cung cấp dịch vụ tốt hơn cho người dân. Một báo cáo năm 2012 của Viện Toàn cầu McKinsey ước tính rằng bằng cách làm như vậy, các quốc gia phát triển của châu Âu có thể tiết kiệm 100 tỷ euroo (149 tỷ USD) mỗi năm chỉ riêng trong việc cải thiện hiệu quả hoạt động.

Bằng cách trở thành quốc gia sớm áp dụng đổi mới dựa vào dữ liệu, các cơ quan chính phủ có thể giúp xây dựng các cộng đồng am hiểu dữ liệu (data-savvy communities) địa phương, chứng minh tính khả thi của các công nghệ khác nhau và thúc đẩy mối quan tâm đến đổi mới dựa vào dữ liệu trong công chúng. Cuối cùng, các cơ quan chính phủ cấp quốc gia và địa phương cần tham gia trực tiếp vào cộng đồng khoa học dữ liệu và tham gia vào các cuộc thi lập trình, thi mã hóa dành cho mọi công dân và các sự kiện khác được cộng đồng khoa học dữ liệu tổ chức.

Công nghệ

Chính phủ cũng có thể giúp thúc đẩy sự phát triển các công nghệ tạo năng lực sử dụng dữ liệu. Năm 2012 tại Hoa Kỳ, chính quyền Obama đã công bố sáng kiến NC&PT dữ liệu lớn với khoản tài trợ 200 triệu USD. Các nỗ lực tài trợ như vậy cần được tiếp tục và mở rộng do các lợi ích của những công nghệ này có thể có các hiệu ứng lan tỏa tích cực đối với toàn bộ nền kinh tế. Như một số nhà kinh tế lưu ý, đầu tư cho tín dụng thuế NC&PT tạo ra hơn một đôla cho nghiên cứu từ mỗi đôla thuế nộp trước. Hơn nữa, khi các cơ quan chính phủ phát triển phần mềm riêng của họ, họ nên phổ biến cho các cộng đồng mã nguồn mở để những người khác có thể tái sử dụng nó và dựa vào nó. Làm như vậy sẽ giúp đảm bảo rằng các công dân phát huy tối đa những lợi ích của tiền thuế được dùng cho nghiên cứu và phát

triển.

Để đảm bảo rằng tiền đầu tư cho nghiên cứu của chính phủ đang hướng vào những thách thức cấp bách nhất trong khu vực công và tư nhân, một cơ quan chính phủ, với ngân sách công lớn, nên phát triển một lộ trình NC&PT về các chủ đề liên quan như phân tích dữ liệu, lưu trữ dữ liệu và điện toán phân tán cũng như các chủ đề riêng tư và bảo mật. Điều này có thể đặc biệt thành công trong các lĩnh vực nơi các tiến bộ công nghệ có thể làm giảm các rào cản để thích ứng. Ví dụ, những quan ngại về tính riêng tư có thể được giải quyết thông qua các công nghệ và phương pháp mới trong các lĩnh vực như xóa vết định dạng dữ liệu, đảm bảo an toàn thông tin trong quá trình khai thác dữ liệu, bảo mật, xác thực đa bên và khả năng liên tác số. Các hợp tác công tư, chẳng hạn như Liên hiệp Quốc gia về khoa học dữ liệu của Hoa Kỳ (NCDS), cũng có thể giúp mang lại kiến thức chuyên môn sâu rộng để thiết lập các ưu tiên nghiên cứu và ban hành các chuẩn.

Cuối cùng, chính phủ có thể khuyến khích việc sử dụng và tái sử dụng dữ liệu bằng cách khuyến khích chuẩn hóa. Do các chuẩn dữ liệu có thiên hướng mang lại lợi ích cho phạm vi rộng các bên liên quan trong một khu vực nhất định, sự đồng thuận rộng rãi thường có thể đạt được; tuy nhiên trong một số trường hợp, sự hỗ trợ của chính phủ có thể giúp đẩy nhanh quá trình này. Tại Hoa Kỳ, sự lãnh đạo của Ủy ban Chứng khoán và giao dịch (SEC) trong xây dựng chuẩn XBRL về hồ sơ doanh nghiệp là một ví dụ điển hình về vai trò tạo điều kiện thuận lợi của chính phủ trong ban hành các chuẩn dữ liệu. Hoa Kỳ cũng sẽ tiếp tục hỗ trợ Liên minh Dữ liệu nghiên cứu quốc tế để làm cho dữ liệu khoa học và các cụ phân tích tương thích trên toàn thế giới.

Dữ liệu

Nếu không có dữ liệu, đổi mới sáng tạo dựa vào dữ liệu là không thể. Kết quả là, chính phủ có một vai trò quan trọng không chỉ trong việc thu thập và cung cấp dữ liệu, mà còn trong việc tạo ra các khuôn khổ pháp lý phù hợp để thúc đẩy việc chia sẻ dữ liệu và nâng cao nhận thức của công chúng về tầm quan trọng của chia sẻ dữ liệu.

Các cơ quan chính phủ nên để người dùng tiếp cận dữ liệu riêng của họ một cách kịp thời và ở định dạng hữu ích. Việc làm cho dữ liệu được nhận dạng đầy đủ và duy nhất, công khai trực tuyến ở định dạng có thể đọc được bằng máy và kịp thời sẽ cho phép các doanh nghiệp, các nhà nghiên cứu, các tổ chức phi lợi nhuận và người dân có thể tái sử dụng. Một cách để đạt được điều này là thông qua các chính sách dữ liệu mở rõ ràng ở tất cả các cấp của chính phủ, chẳng hạn như Điều lệ Dữ liệu mở 2013 của G8, Chương trình nghị sự Dữ liệu mở của Hoa Kỳ, hoặc chính sách dữ liệu mở của thành phố Toronto.

Tương tự như vậy, các nhà hoạch định chính sách cần tiếp tục theo đuổi các nỗ lực để cho phép các cá nhân truy cập vào dữ liệu cá nhân của chính họ. Hai ví dụ của nỗ lực này ở Hoa Kỳ là Sáng kiến Nút bấm xanh (Green Button) khuyến khích các công ty tiện ích tạo điều kiện thuận lợi để người tiêu dùng có thể truy cập vào dữ liệu sử dụng năng lượng tại nhà của họ và các Sáng kiến Nút bấm lam (Blue Button) để các cựu chiến binh có thể truy cập hồ sơ y tế của họ. Bằng cách theo đuổi quy tắc "*mở mặc định*", các cơ quan chính quyền ở tất cả các cấp có thể khuyến khích các nghiên cứu và thử nghiệm mở rộng rất quan

trọng để khởi phát đổi mới dựa vào dữ liệu. Khi các công ty không tự nguyện cung cấp cho khách hàng của mình cơ hội truy cập vào dữ liệu riêng ở định dạng điện tử, có thể tái sử dụng, các nhà hoạch định chính sách có thể cần can thiệp. Đây không phải là việc bắt buộc các công ty phải từ bỏ quyền sở hữu dữ liệu, mà là yêu cầu họ cố gắng cung cấp cho khách hàng những bản sao dữ liệu riêng của họ.

Các nhà hoạch định chính sách cũng cần đảm bảo rằng họ tạo ra các khuôn khổ pháp lý và luật pháp để khuyến khích chia sẻ dữ liệu và tái sử dụng trong các ngành công nghiệp khác nhau. Đổi mới sáng tạo dựa vào dữ liệu diễn ra khi các tổ chức, cá nhân có thể thu thập, sử dụng và tái sử dụng dữ liệu cho các mục đích mà họ có thể không hình dung ban đầu. Ví dụ, cuộc điều tra dân số đầu tiên của Hoa Kỳ ban đầu được tiến hành cho mục đích duy nhất là xác định đại biểu Quốc hội, nhưng dữ liệu của nó đã được áp dụng cho một loạt các ứng dụng trong khu vực công và tư nhân, từ tăng trưởng kinh tế đến phân tích y tế công cộng. Để hỗ trợ cho các ứng dụng không được lường trước như vậy, các nhà hoạch định chính sách cần tạo không gian cho sự đổi mới ngẫu nhiên. Điều này có nghĩa là các khung pháp lý nên hỗ trợ sự di chuyển của dữ liệu giữa các cá nhân, trong và giữa các quốc gia và các tổ chức. Những nỗ lực của một số quốc gia áp đặt các luật "*khu trú dữ liệu*" hạn chế luồng thông tin tự do toàn cầu chứ không phải là khuyến khích lưu thông dữ liệu xuyên biên giới.

Các nhà hoạch định chính sách cũng nên tránh các quy định hạn chế không cần thiết về thu thập và chia sẻ dữ liệu. Khi những hạn chế sử dụng là cần thiết chúng cần được thực hiện với sự kiềm chế. Các quy định của pháp luật ngăn chặn việc sử dụng dữ liệu có thể dẫn đến một tình huống gọi là "*bi kịch chống lại những cái chung*".

Điều này xảy ra khi sự tồn tại của quá nhiều rào cản pháp lý và quan liêu tạo ra chi phí giao dịch cao hạn chế việc sử dụng và trao đổi dữ liệu. Ví dụ, sự không chắc chắn về quyền sở hữu dữ liệu có thể ngăn chặn một công ty tạo ra một ứng dụng dựa vào dữ liệu hữu ích. Để không làm giảm tính năng của các ứng dụng dữ liệu có lợi, các cuộc thảo luận chính sách cần tập trung giải quyết việc dữ liệu có thể được sử dụng như thế nào, chứ không phải là việc quyết định liệu nó có nên được thu thập và trao đổi hay không. Những sử dụng đưa đến tác hại cụ thể nên bị cấm, nhưng các nhà hoạch định chính sách cần tạo ra chính sách mở thừa nhận phạm vi rộng không thể dự báo trước của các ứng dụng dựa vào dữ liệu trong tương lai, đặc biệt là trong các lĩnh vực y tế và giáo dục.

Ở đây tồn tại những cơ hội tuyệt vời tận dụng dữ liệu để giải quyết các vấn đề xã hội quan trọng và khuyến khích tăng trưởng kinh tế, tuy nhiên, để đạt được đầy đủ tiềm năng của đổi mới dựa vào dữ liệu, các nhà hoạch định chính sách phải tạo ra cơ sở hạ tầng và khung chính sách cần thiết. Bước đầu tiên để làm điều đó là phải hiểu và đánh giá cao tầm quan trọng của đổi mới dựa vào dữ liệu trong khu vực công và tư nhân.

Biên soạn: ***Đặng Bảo Hà***
Nguyễn Lê Hằng

Tài liệu tham khảo

1. OECD: DATA-DRIVEN INNOVATION FOR GROWTH AND WELL-BEING: INTERIM SYNTHESIS REPORT. 10/2014.

2. OECD: EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH: MAPPING THE POLICY ISSUES RAISED BY "BIG DATA". 6/2013.

3. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 6/2011.

4. Market Analysis: Worldwide Big Data technology and services 2012-2015 Forecast. www.idc.com

5. White Paper: Data-Driven Innovation in South-East Europe**.** Economics Institute, Serbia; Inženjerski biro, Croatia; Economics Institute, Bosnia and Herzegovina; Economic Program Center for the Study of Democracy, Bulgaria, 12/2014**.**

6. Jeff Kelly, "Big Data Vendor Revenue and Market Forecast," *Wikibon,* 12 Feb. 2014.

7. Daniel Castro & Travis Korte: Data Innovation 101: An Introduction to the Technologies and Policies Supporting Data-Driven Innovation. Center for Data Innovation, 11/2013.

8. Report to the President: BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE. The President's Council of Advisors on Science and Technology (PCAST), 5/2014.

9. The Future of Data-driven Innovation. U. S. Chamber of Commerce Foundation, 10/2014.

10. BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES. Executive Office of the President , 5/2014.

11. Big Data for Development: Challenges & Opportunities. Global Pulse, 5/2012.

12. David Abecassis, Nico Flores, Sara Montakhab: Data-driven innovation in Japan - supporting economic transformation . Analysys Mason Limited, 10/2014.