Data Innovation 101: An Introduction to the Technologies and Policies Supporting Data-Driven Innovation

By Daniel Castro & Travis Korte | November 4, 2013

Summary: New technologies have made it easier and cheaper to collect, store, analyze, use, and disseminate data. But while the potential for vastly more data-driven innovation exists, many organizations have been slow to adopt these technologies. Policymakers around the world should do more to spur data-driven innovation in both the public and private sectors.

If oil was the fuel of the twentieth-century economy, data will be the fuel of the twenty-first century. Indeed, by enabling people to better understand the complex world around us and to use that understanding to make better decisions, large and small, data has the potential to drive innovation in a broad range of areas, improving both economic productivity and quality of life. A growing number of organizations, from Netflix to the National Oceanic and Atmospheric Administration (NOAA), have already begun to harness large-scale data analysis technologies to great effect. For example, 75 percent of what users watch on Netflix is found through the company's advanced recommendation algorithm; and NOAA has refined its predictive weather models to anticipate storms ten times more accurately than it did 20 years ago.[1]

Making sense of the vast amounts of data collected about people and the world around them is necessary to address major social challenges, including improving health care, education, public safety, transportation, energy, and the environment. Some countries have already begun moving in this direction. For example, the Australian government has identified data as key to driving innovation in the public sector, including through initiatives to maintain infrastructure, improve health care, and reduce response times for emergency services.[2] But data-driven innovation is not merely a tool for developed countries; developing countries, too, are

harnessing the power of data, including for humanitarian aid and development.[3]

Researchers use data to open doors for future applications and to spawn entirely new areas of inquiry. Scientists ushered in medical genomics with the Human Genome Project, advanced particle physics at the European Center for Nuclear Research (CERN), and mapped the stars with the Sloan Digital Sky Survey, all relying on enormous quantities of data. Some landmark research datasets have continued to pay dividends decades after their collection began. For example, data from the Framingham Heart Study, a long-term cardiovascular study launched in 1948, has recently been applied to analyses of obesity and divorce. In the future, researchers will derive insights into neuroscience and computing using massive datasets from the EU's Human Brain Project and the U.S. BRAIN Initiative. Cancer research will be supported through data-sharing fostered by the U.S. National Cancer Informatics Program.

Data-driven innovation can also strengthen economies by reducing waste and improving services. For example, in the United States, the value from data in health care exceeds $300 billion annually.[4] And globally, the use of smart grid data could reduce carbon dioxide emissions by more than two billion tons by 2020, according to a 2013 OECD estimate.[5]

A conversation about data-driven innovation is possible now because new technologies have made it easier and cheaper to collect, store, analyze, use, and disseminate data. But while the potential for vastly more data-driven innovation exists, many organizations have been slow to adopt these technologies. Policymakers around the world should do more to spur data-driven innovation in both the public and private sectors, including by supporting the development of human capital, encouraging the advancement of innovative technology, and promoting the availability of data itself for use and reuse.

THE BENEFITS OF DATA

Data leads to better understanding and decision making among individuals, businesses, and government.

Individuals use data to make better decisions about everything from what they buy to how they plan for the future. These decisions can be minor, such as deciding whether to carry an umbrella based on weather forecasts, or major, such as deciding where to go to college based on school evaluations or predictions of future career earnings. Traffic data helps individuals find the most efficient route from point A to point B, saving time and gas in the process. Data from the electricity grid can help homeowners save on utility bills. User reviews on sites like Amazon help consumers discover the products that they like best.[6] Yelp's restaurant reviews help people decide where to enjoy their next meal, and (since the site has recently begun to integrate additional data from city health inspections) to factor food safety into these decisions.[7]

Businesses use data to find new customers, automate processes, and inform business decisions. For example, Visa's data-driven Advanced Authorization service alerts banks to potential fraudulent transactions in real-time, identifying as much as $1.5 billion in fraud around the world annually.[8] Coca Cola uses complex models to ensure that every batch of orange juice it blends tastes consistently fresh.[9] Intel uses predictive modeling on data from its chip manufacturing plants to anticipate failures, prioritize inspections and cut monitoring costs.[10] GlaxoSmithKline conducts text analytics on data collected from online forums so that it can better understand and respond to the concerns of parents who delay vaccinating their children.[11] Wind energy companies, such as Vestas, use complex weather models to determine the optimal locations for their turbines.[12]

Government agencies use data to cut costs, prioritize social services, and keep citizens safe. The U.S. Securities and Exchange Commission analyzes data reported by publicly-traded companies to identify suspicious filings and inform fraud investigations.[13] The European Space Agency deploys satellites equipped with remote sensing technologies to track and analyze changes in the global environment and help forecast weather events, such as hurricanes and droughts.[14] The U.S. Centers for Disease Control and Prevention uses social network analysis to better understand and stem the spread of communicable diseases.[15] The UK's Royal Mail uses analytic software to determine the most efficient delivery routes and make sure parcels get to their destinations as quickly as possible.[16] The U.S. Institute of Educational Sciences conducts randomized trials, inspired by clinical research, to collect data and measure the impact on learning outcomes of certain educational variables, such as choice of instructional materials.[17] New York City's Fire Department prioritizes inspections based on risk assessments derived from building data which has resulted in the city reducing the number of annual fire deaths to the lowest since recordkeeping began in 1916.[18]

Public Health	The U.S. Centers for Disease Control is using social network analysis to better understand and stem the spread of communicable diseases.
Education	The U.S. Institute of Educational Sciences conducts randomized trials, inspired by clinical research, to collect data and measure the impact of certain educational variables, such as choice of instructional materials.
Developing Countries	The non-profit company Ushahidi creates software to easily collect and visualize important information for those providing humanitarian aid in developing countries, such as the availability of critical drugs in southeast Africa.[19]

Transportation	The City of Dublin, Ireland combines real-time data streaming with traffic information collected from a number of sources to map city bus locations and combat traffic jams.[20]
Environment	The European Space Agency deploys satellites equipped with remote sensing technologies to track and analyze changes in the global environment and help forecast weather events, such as hurricanes and droughts.
Public Safety	New York City's Fire Department prioritizes inspections based on risk assessments derived from building data which has resulted in the city reducing the number of annual fire deaths to the lowest since recordkeeping began in 1916.
Retail	User reviews on sites like Amazon help consumers discover the products that they like best.
Government	The U.S. Securities and Exchange Commission analyzes reporting data from publicly-traded companies to identify suspicious filings and inform fraud investigations.
Energy	Wind energy companies use complex weather models to determine the optimal locations for their turbines.
Manufacturing	Intel uses predictive modeling on data from its chip manufacturing plants to anticipate failures, prioritize inspections and cut monitoring costs.

Table 1: Ten examples of benefits from data-driven innovation

UNDERSTANDING THE JARGON

The topic of data-driven innovation is so new that many terms are poorly understood. These include "big data," "open data," "data science," and "cloud computing."

"Big data" refers to data that cannot be processed using traditional database systems, either due to the relative size and heterogeneity of the data set, or the speed at which it is updated. Big data has existed for decades in fields such as astronomy and atmospheric science, but the growth of digital data collection has rapidly brought the topic into many other fields.[21] In some areas, big data technologies have allowed researchers to analyze entire populations, without having to rely on samples; in addition to enabling faster analysis, this has also resulted in increased model accuracy.[22]

"Open data" refers to data that is made freely available without restrictions.[23] Such data is most useful when it is released in non-proprietary, machine-readable formats. Open data can be used to drive innovation within and beyond the organization that created it because it

allows other organizations to make use of it. A 2013 McKinsey Global Institute report estimated that open data could add over $3 trillion in total value annually to the education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance sectors worldwide.[24]

"Data science" refers to the range of technologies, as well as statistical and computational techniques, that are used to analyze and derive insights from data. The term "data science" does not exclusively refer to these techniques and technologies as they are applied to big data; it also encompasses data analysis that is conducted using smaller data sets and traditional database systems.[25]

"Cloud computing" is the practice of "renting" remotely-located IT services, including processing capabilities, information storage, and software applications, on an as-needed basis. Cloud computing turns fixed costs into variable costs, and it allows organizations to scale their computing resources to meet demand.[26]

WHAT TECHNOLOGIES ARE BEHIND DATA-DRIVEN INNOVATION?

The increasing availability of data has spurred the development of new techniques and technologies at every stage of the data lifecycle, including data collection, storage, manipulation, analysis, use, and dissemination. In turn, these technologies have increased the value of raw data, leading to more collection and even greater availability of data.

COLLECTION

Collecting data is the first step in the data innovation lifecycle. As of 2012, about 2.5 billion gigabytes of data were collected each day globally, a significant portion of which was video.[27] In comparison, the entire print collection in the U.S. Library of Congress amounts to about 10,000 gigabytes.[28]

The two major sources of new digital data are physical sensors and electronic records.[29] As with most electronics, the size and cost of many sensors has decreased significantly over the last decade, while their capabilities have increased significantly.[30] Sensing technologies encompass an extremely broad array of devices, which measure physical variables such as temperature, pressure, geolocation, chemical composition, electric current, movement, light content, and many others. Sensors are an integral part of the "Internet of Things," a concept used to describe a world where everyday objects, from aircraft to refrigerators and running shoes, can communicate with each other and their users.[31] For example, Boeing 787 aircraft generate over half a terabyte of data per flight from the engines, landing gear, and other equipment.[32] Sensors are highly specialized and multiple varieties of sensors often are used to measure a given environmental variable across different application domains. Data scientists frequently use signal processing and statistical modeling techniques to derive insights from sensor data; for example the

National Weather Service uses climate modeling in its forecasts. The amount of sensor data will continue to grow as sensors become even more effective and cheaper, and companies embed them in more and more devices. The availability of cheap, low-power processors will also support this growth, enabling companies to embed intelligence in any device.

Electronic records include structured and unstructured data. Structured data refers to data that is highly organized and easily queried, such as tabular data on transactions, account details, and other online activity. By design, structured data is often simpler to analyze; certain applications, such as network analysis and predictive modeling, require structured data. Unstructured data refers to data that is less organized and that does not lend itself to being queried, such as images, video and audio. An electronic lab report from a hospital, for example, or a digitized shipping manifest from a trucking company would typically be stored in structured formats; news stories, online videos, and text-based product reviews usually are unstructured.

Structured data is collected in large volumes by a variety of public and private-sector organizations. The United Parcel Service, for example, receives an average of 39.5 million tracking requests each day, and Visa processes over 172 million card transactions daily.[33] However, the vast majority of data being collected today is unstructured, and much of it is video. As of June 2012, YouTube users uploaded forty-eight hours of new video every minute.[34]

Improvements to physical and wireless networks also influence the amount of data collected and the range of opportunities available for data-driven innovation. A 2013 Cisco analysis predicted that the amount of internet traffic moving globally through telecommunications networks will increase nearly threefold between 2012 and 2017, to a total of 3.1 exabytes per day.[35]

STORAGE

Once data is collected, it must also be stored. Efficient and flexible data storage can simplify data analysis and provide significant cost savings. In the past two decades, data storage has benefitted from software as well as hardware innovation.

Improved hardware has allowed storage costs to drop precipitously; in 1980, it cost around $440,000 to store one gigabyte of data, while in 2013, it cost about $0.05.[36] Advances in data centers have also made it easier and cheaper for organizations to store vast quantities of data remotely using cloud-based storage options. In addition to these dramatic hardware improvements, developers have created a range of database software designed to store unstructured data and achieve “big data” scalability. Traditional structured query language (SQL) databases rely on restrictive organizational structures, and do not always lend themselves to heterogeneous and changing data inputs. These systems, which have been used for decades to store employee files, sales data, and other well-

organized information, are not easily extensible to many modern data science applications, such as document storage.

U.S. companies, along with the global open-source community, have been pioneers in developing technologies that overcome some of these shortcomings. Collectively, these are referred to as NoSQL technologies, to indicate their rejection of various SQL characteristics, including constraints on centralized storage and data modification. Examples of proprietary NoSQL technologies include Google's BigTable, Amazon's Dynamo, and Facebook's Cassandra, all of which have fostered the development of open-source technologies that enable big data storage and analytics. For example, the Apache Software Foundation developed HBase, a popular database used with big data, based on initial work done by Google.

ANALYSIS

Data analysis extracts meaning from data, in part by identifying correlations between variables and making predictions about future events.

The vast growth of unstructured data has spurred the development of techniques such as text mining, natural language processing, and computer vision, all of which can help make sense of this data. Developers also have devoted considerable effort to creating software for data manipulation and analysis, including numerical programming languages, statistical software, dedicated business analytics tools, and "big data" utilities. The use of analytics applications can provide high returns on investment; one 2011 study found that companies earn an average of $10.66 for every dollar spent on analytics applications.[37]

Due to their versatility, programming languages, such as the statistical computing language R and the numerical computing languages Matlab and Julia, are used for data analysis and manipulation in a broad range of fields.[38] Programming languages, while relatively difficult to learn, allow users to create and distribute their own functions; R, for example, offers specialized functions for an extremely broad range of fields, including medical image analysis, econometrics, and ecological analysis.[39] In addition, the general purpose Python programming language has been extended to include statistical capabilities. There also exist a variety of dedicated statistical software, such as SAS, SPSS, and Stata.[40]

An important subset of statistical software is business analytics software, which companies use to make data-driven business decisions. Business analytics software offerings are highly diverse and include off-the-shelf packages from providers such as Adobe, IBM, Microsoft, Oracle, SAP, and SAS. These user-friendly tools allow analysts to probe and manipulate data using pre-set commands built into the software. While less flexible than programming languages, business analytics software can be particularly useful for industries that use well-defined metrics, such as the insurance industry.[41]

In addition, tools have been developed specifically for big data, such as Hadoop, an open-source platform for applications that involve analyzing large data sets. Organizations in a wide range of industries, including health care, agriculture, and utilities, use Hadoop's core functionality to process large amounts of data.[42] Many developers have created extensions and add-ons that serve specific use cases, such as real-time analytics.[43] For example, one lab in Maryland's Howard Hughes Medical Institute uses a Hadoop-based real-time analytics platform to analyze and visualize patterns of brain activity in real time.[44]

Although analysis is conducted using software, improvements in computer hardware, particularly processing units, have enabled faster, cheaper and more energy-efficient data processing capabilities over the years. Hardware built for large-scale data analysis includes the multi-core processor offerings being continually refined by Intel and AMD; IBM's high-performance server hardware (based on technologies developed for its Watson project); Oracle's storage-intensive "big memory" server; and devices optimized for "big data" from HP and EMC. The rise of parallel computing and cloud-based processing has made processor speed somewhat less of a bottleneck to data analysis than it was in previous decades, but hardware vendors' incremental advances remain an important driver of high-performance applications.[45]

USE

The ultimate purpose of data analysis is to support better decision-making, whether those decisions are made by an executive in an office, a robot on the factory floor, or someone at home. Data-driven automation can simplify decisions made by robots, while information organized using decision support systems, data visualization, and mapping technologies can help humans.

Decision Support System

Decision support systems are interactive tools that help users make better and faster decisions in complex, multivariate environments. Decision support systems use models and simulations to predict outcomes and then present recommendations to decision-makers. For example, a construction manager may use a decision support system to help choose which subcontractor offers the best combination of risk and revenue for a given project.

Such systems are especially popular in hospitals, where clinical decision support systems can use patient information to alert a doctor if a prescription will interfere with other medications or conditions.[46] Decision support systems can also be found in many other domains, including environmental monitoring. For example, the Mediterranean Decision Support System for Maritime Safety was designed for EU member governments to help mitigate the risks of Mediterranean oil spills.[47] As data analysis techniques such as predictive modeling and natural language processing continue to develop, so will the capabilities of decision support systems.

Automation

While much of data analysis is deployed to help humans make better-informed decisions, data can also be used to trigger automated actions in computer systems and robots. For example, Nest, a smart thermostat, can use sensor data to determine if a house is occupied and adjust the house's heating and cooling appropriately. Google's self-driving car can take data about roadway conditions and incoming traffic to navigate efficiently and avoid collisions. A 2013 report from research firm Markets and Markets projected that the market for machine-to-machine communications would reach $290 billion in 2017, a 650% increase from 2011.[48]

Machine learning, a branch of computer science concerned with systems whose performance improves with the addition of new data, offers methods for automated decision-making in a range of applications. Machine learning has seen widespread use in robotics, such as computer vision and automated movement in factory environments, as well as in online recommendation systems, such as those used by music streaming service Spotify and online dating site OKCupid.[49]

Visualization

One way data scientists can communicate their analysis to decision makers is through visualizations. Visualizations are used in a broad array of fields, and can range from simple line graphs of stock prices to complex social network diagrams showing the spread of disease. In cases where patterns in the data may be more easily identified when the data is displayed, visualizations can be used for conducting analysis as well.[50] Data visualization is built into many business analytics software tools, such as Tableau. Specialized platforms and languages exist for particular applications, such as Gephi for network and graph visualization, and Processing for interactive visualization. The Javascript programming language is popular for custom data visualization applications, offering widely used open-source libraries such as D3.

Mapping applications have driven the development of a wide array of geographical information systems (GIS) software, which allows spatial features to be integrated into data analysis. There are specialized technologies for all aspects of geospatial data-driven innovation, including databases, servers and visualization tools. Major proprietary software providers include Esri (provider of ArcGIS), Google (provider of Google Maps, Earth and Street View), and Oracle (provider of Spatial and Graph). Open-source GIS offerings, such as those created by geospatial technology firm MapBox, are growing in popularity as well. Tools from these providers are widely used in industry and government. For example, the Obama administration used GIS software to add data layers and interactivity to the maps on its Recovery.gov website.[51]

DISSEMINATION

Organizations, including government agencies, often want to share their data with others. In the past, data sets often were disseminated through digital media, such as CDs, but using physical objects for dissemination

meant data volume was limited and distribution slow and costly. Today, data is made available on websites, often at no direct cost to the user. Some organizations simply provide access to raw data files; others develop application programming interfaces (APIs) to make it easier for other developers to reuse their data.

More recently, dedicated software to manage the large number of open data sets released by organizations has emerged, chiefly from start-up Socrata. This software niche is relatively new, and other players have begun to emerge only recently. In some cases, organizations have developed their open data dissemination platforms internally; an example is the U.S. hub Data.gov. The creators of this platform subsequently released their software to the open-source community.[52]

HOW SHOULD POLICYMAKERS SUPPORT DATA-DRIVEN INNOVATION?

Data-driven innovation is critical to economic growth and improvements in quality of life. While the private sector will drive much of this progress, governments can and should aid this effort. In particular, data-driven innovation requires a skilled workforce, innovative technology, and access to data. Policymakers can support efforts to ensure all of these needs are met.

HUMAN CAPITAL

There is a global shortage of workers with the knowledge, skills, and abilities to support data-driven innovation. These workers include not only programmers skilled with machine learning and Hadoop, but also data-literate managers, designers and communications specialists. For example, research firm Gartner projected in 2012 that, by 2015, only one-third of the 4.4 million available big data jobs will be filled.[53] While some universities recently have begun offering programs devoted to data science, business analytics and machine learning, these efforts may not gain momentum quickly enough to meet the imminent demand.

Countries that can provide the talent necessary to work in data-related fields will have an advantage in the global economy.[54] Policymakers have an opportunity to help accelerate the growth of a data-literate workforce by funding efforts to develop open, online courses in data-related subjects and expand enrollment in statistics and computer science classes. Secondary schools can also help by creating more flexible math requirements, so that students can take computer science or statistics courses. Although such efforts undoubtedly will take some time to pay off, they could help open new opportunities for workers and expand the availability of interdisciplinary, data-literate workers for companies in the long term.

Government can also help spur the development of the necessary human capital by becoming a leader, rather than a laggard, in the adoption of data-driven innovation. Government agencies can use data to save money and deliver better services to citizens. A 2012 report from the McKinsey

Global Institute estimated that by doing so the developed countries of Europe could save €100 billion ($149 billion) annually in operational efficiency improvement alone.[55]

By becoming early adopters, government agencies can help build local data-savvy communities, demonstrate the feasibility of different technologies, and foster public enthusiasm for data-driven innovation. To this end, government agencies at the federal, state and local levels should continue to engage directly with the data science community and participate in civic hackathons, public coding challenges, and other events hosted by the data science community.

TECHNOLOGY

The government can also help accelerate the development of technologies that facilitate use of data. In 2012 in the United States, the Obama Administration announced a one-time big data research and development (R&D) initiative with $200 million in funding.[56] Funding efforts such as these should be continued and expanded since the benefits of these technologies can have strong positive spillover effects and benefits throughout the economy. As some economists have noted, investment in R&D tax credits produces more than one dollar of research for every foregone tax dollar.[57] Moreover, when government agencies develop their own software, they should make it available to the open-source community so that others can reuse it and build on it. Doing so will help ensure that citizens maximize the benefits of tax dollars spent on development.

To ensure that government research dollars are directed at the most pressing challenges in the public and private sectors, a government agency, with broad public input, should develop an R&D roadmap on relevant topics such as data analytics, data storage, and distributed computing, as well as privacy and security topics. This may be especially fruitful in areas where technological advances could reduce barriers to adoption. For example, some privacy concerns could be addressed through new technologies and methods in areas such as data de-identification, privacy-preserving data mining, secure, multi-party authentication, and interoperable digital credentials.[58] Public-private partnerships, such as the United States' National Consortium for Data Science (NCDS), can also help bring in wide-ranging expertise for setting research priorities and promulgating standards.[59]

Finally, government can help encourage data use and reuse by encouraging standardization. Since data standards tend to benefit a wide range of stakeholders across a given sector, broad consensus often can be reached; in some cases, however, a strong first mover in government can help accelerate the process. In the United States, the Securities and Exchange Commission's (SEC) leadership in instituting the XBRL standard for corporate filings serves as a prime example of the government's facilitative role in promulgating data standards.[60] The United States should also continue to support the international Research Data Alliance in its

efforts to make scientific data and analysis tools interoperable throughout the world.[61]

DATA

Without data, data-driven innovation is impossible. As a result, the government has an important role to play not only in collecting and supplying data, but also in creating the appropriate legal frameworks to foster data sharing, and in raising public awareness about the importance of sharing data.

Government agencies should make their own data available to users in a timely way and useful format. Making complete and uniquely identified data available publicly online in a machine-readable format and in a timely manner will allow for reuse by businesses, researchers, non-profit organizations, and citizens. In addition, government agencies should ensure that data produced directly in their behalf by contractors fall under open data policies. One way to achieve this is through explicit open data policies at all levels of governments, such as the G8's 2013 Open Data Charter, the U.S. Open Data Agenda, or the City of Toronto's open data policy.[62]

Similarly, policymakers should continue to pursue efforts to allow individuals access to their own personal data. Two examples of this in the United States are the Green Button Initiative, which encourages utility companies to give consumers access to data on their home energy use and the Blue Button initiative, which gives veterans access to their health records. By pursuing an ethic of "open by default," government agencies at all levels can encourage the sorts of open-ended exploration and experimentation that are crucial for sparking data-driven innovation. When companies do not voluntarily provide their customers access to their own data in a reusable, electronic format, policymakers may need to intervene. This is not to say that companies must give up ownership of that data, only that they should strive to provide customers with copies of their own data.

Policymakers should also ensure that they create the legal and regulatory frameworks to encourage data sharing and reuse in different industries. Data-driven innovation occurs when organizations and individuals can collect, use, and reuse data for purposes that they might not have originally envisioned. The first U.S. Census, for example, was initially conducted for the sole purpose of determining congressional representation, but its data has since been applied to a host of public and private-sector uses, from economic growth to public health analyses. To support such unforeseen applications, policymakers should make space for serendipitous innovation. This means that regulatory frameworks should support the movement of data among individuals and within and between nations and organizations. Attempts by some countries to impose "data residency" laws restrict the global free flow of information rather than encourage cross-border data flow.[63]

Policymakers should also avoid unnecessarily restrictive regulations on the collection and sharing of data. When restrictions on use are necessary they should be implemented with restraint. Legal rules preventing the use of data can lead to a situation known as the "tragedy of the anticommons." This occurs when the existence of too many legal and bureaucratic barriers create high transaction costs that restrict the use and exchange of data. For example, uncertainty over data ownership may prevent a company from creating a useful data-driven application. In order not to undermine beneficial applications of data, policy discussions should focus on resolving how data can be used, rather than on deciding whether it can be collected and exchanged. Uses that result in specific harm should of course be prohibited, but policymakers must craft open-ended policies acknowledging the unpredictable breadth of future data-driven applications, particularly in the health and education sectors.

CONCLUSION

There are incredible opportunities to leverage data to address important social issues and encourage economic growth. However, to achieve the full potential of data-driven innovation, policymakers must create the necessary infrastructure and policy framework. The first step to doing that is to understand and appreciate the critical importance of data innovation in the public and private sector.

REFERENCES

1. "Xavier Amatriain and Justin Basilico, "Netflix Recommendations: Beyond the 5 Stars (Part 1)," The Netflix Tech Blog, 2012, http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html and Frank Konkel, "Sandy Shows Storm-Prediction Progress," Federal Computer Week, 2012, http://fcw.com/articles/2012/11/05/sandy-hurricane-center.aspx.

2. Australian Government Information Management Office, "The Australian Public Service Big Data Strategy," 2013, http://agict.gov.au/sites/default/files/Big%20Data%20Strategy.pdf.

3. Global Pulse, "Harnessing Innovation to Protect the Vulnerable," 2013, United Nations, http://www.unglobalpulse.org.

4. Peter Groves et al, "The 'Big Data' Revolution in U.S. Health Care," McKinsey & Company, 2013, http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care and Global e-Sustainability Initiative and Boston Consulting Group, "SMART 2020: Enabling the Low Carbon Economy in the Information Age," 2008, http://www.smart2020.org/_assets/files/Smart2020UnitedStatesReportAddendum.pdf.

5. OECD, "Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by 'Big Data,'" OECD Publishing, 2013, http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en.

6. Derrick Harris, "You Might Also Like … To Know How Online Recommendations Work," GigaOM, 2013, http://gigaom.com/2013/01/29/you-might-also-like-to-know-how-online-recommendations-work/.

7. Jeremy Stoppelman, "Now on Yelp: Restaurant Inspection Scores," Yelp Web Log, 2013, http://officialblog.yelp.com/2013/01/introducing-lives.html.

8. World Economic Forum and Boston Consulting Group, "Unlocking the Value of Personal Data: From Collection to Usage," 2013, http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf.

9. Stanford, Duane, "Coke Engineers Its Orange Juice—With an Algorithm," BloombergBusinessweek, January 31, 2013, http://www.businessweek.com/articles/2013-01-31/coke-engineers-its-orange-juice-with-an-algorithm.

10. Ian King, "Karl Kempf, Intel's Money-Saving Mathematician," Bloomberg Businessweek, 2012, http://www.businessweek.com/articles/2012-05-31/karl-kempf-intels-money-saving-mathematician.

11. Stephanie Baum, "Why it's Good News that GSK is Monitoring What Parents are Saying About Vaccinations," Medcity News, 2013,

http://medcitynews.com/2013/05/gsk-is-monitoring-public-forums-on-vaccinations-and-why-thats-good-news/.

12. Nino Marchetti, "Determining Optimal Placement of Wind Farms," 2012, theenergycollective.com/namarchetti/80862/put-wind-power-its-place-weather-key and IBM, "Vestas: Turning Climate into Capital with Big Data," 2011, http://www-01.ibm.com/software/success/cssdb.nsf/CS/JHUN-8MY8YK?OpenDocument&Site=default&cty=en_us.

13. Dave Michaels, "SEC Data to Transform High-Speed Trading Debate, White Says," Bloomberg, 2013, http://www.bloomberg.com/news/2013-10-02/sec-to-transform-high-speed-trading-debate-with-data-white-says.html.

14. European Space Agency, "Sentinel-3," 2013, http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-3.

15. Centers for Disease Control, "Use of Social Networks to Identify Persons with Undiagnosed HIV Infection," 2005, http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5424a3.htm.

16. Giro, "GeoRoute—Routing and Scheduling," 2013, http://www.giro.ca/en/products/georoute/.

17. Gina Kolata, "Guesses and Hype Give Way to Data in Study of Education," The New York Times, 2013, http://www.nytimes.com/2013/09/03/science/applying-new-rigor-in-studying-education.html?_r=0New.

18. Marc La Vorgna, John McCarthy and Frank Gribbon,"Mayor Bloomberg And Fire Commissioner Cassano Announce New Risk-based Fire Inspections Citywide Based On Data Mined From City Records," NYC.gov, 2013, http://www1.nyc.gov/office-of-the-mayor/news/163-13/mayor-bloomberg-fire-commissioner-cassano-new-risk-based-fire-inspections-citywide#/12.

19. Monica Hesse, "Crisis Mapping Brings Online Tool to Haitian Disaster Relief Effort," Washington Post, 2010, http://www.washingtonpost.com/wp-dyn/content/article/2010/01/15/AR2010011502650.html.

20. Jeff Bertolucci, "Dublin Points Big Data Tech at Traffic Jams," InformationWeek, 2013, http://www.informationweek.com/big-data/news/big-data-analytics/dublin-points-big-data-tech-at-traffic-j/240155213.

21. Gartner, "Big Data," 2013, http://www.gartner.com/technology/topics/big-data.jsp.

22. Kenneth Cukier and Viktor Mayer-Schoenberger, *Big Data: A Revolution that Will Transform How We Live, Work and Think*, Eamon Dolan/Houghton Mifflin Harcourt, 2013.

23. Open Data Handbook, "What is Open Data?", Open Knowledge Foundation, 2012, http://opendatahandbook.org/en/what-is-open-data/.

24. James Manyika et al, "Open Data: Unlocking Innovation and Performance with Liquid Information," McKinsey Global Institute, 2013, http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.

25. DJ Patil, "Building Data Science Teams," O'Reilly Radar, 2011, http://radar.oreilly.com/2011/09/building-data-science-teams.html#what-makes-data-scientist.

26. Daniel Castro, "Cloud Computing: An Overview of the Technology and the Issues Facing American Innovators," testimony before the U.S. House of Representatives Subcommittee on Intellectual Property, Competition and the Internet http://judiciary.house.gov/hearings/Hearings%202012/Castro%2007252012.pdf.

27. IBM, "What is Big Data?" 2013, http://www.ibm.com/big-data/us/en/.

28. Leslie Johnston, "How Many Libraries of Congress Does it Take?" The Library of Congress Digital Preservation Blog, 2012, http://blogs.loc.gov/digitalpreservation/2012/03/how-many-libraries-of-congress-does-it-take/.

29. Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution," Harvard Business Review, 2012, http://hbr.org/2012/10/big-data-the-management-revolution.

30. Daniel Cooley, "Wireless Sensor Networks Evolve to Meet Mainstream Needs," RTC Magazine, 2012, http://rtcmagazine.com/articles/view/102871.

31. Daniel Castro, “The Internet of Things Requires New Thinking on Data,” The Information Technology and Innovation Foundation, 2013, http://www.innovationfiles.org/the-internet-of-things-requires-new-thinking-on-data/.

32. Matthew Finnegan, "Boeing 787s to Create Half a Terabyte of Data per Flight, Says Virgin Atlantic," ComputerworldUK, 2013, http://www.computerworlduk.com/news/infrastructure/3433595/boeing-787s-create-half-terabyte-of-data-per-flight-says-virgin-atlantic/.

33. Bennett J. Loudon, "UR Announces $50 Million 'Big Data' Plan," Democrat & Chronicle, 2013, www.democratandchronicle.com/story/money/business/2013/10/18/ur-announces-50-million-big-data-plan/3007857/.

34. Zach Whittaker, “How Much Data is Consumed Every Minute?”, ZDNet, 2012, http://www.zdnet.com/blog/btl/how-much-data-is-consumed-every-minute/80666.

35. “VNI Forecase Highlights,” Cisco Visual Networking Index, n.d., http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html (accessed October 22, 2013).

36. Statistic Brain, "Average Cost of Hard Drive Storage," 2013, http://www.statisticbrain.com/average-cost-of-hard-drive-storage/.

37. Nucleus Research: "Analytics Pays Back $10.66 for Every Dollar Spent," 2011, http://www.ironsidegroup.com/wp-content/uploads/2012/06/l122-Analytics-pays-back-10.66-for-every-dollar-spent.pdf.

38. Avi Bryant, "MATLAB, R, and Julia: Languages for Data Analysis," O'Reilly Strata, 2012, http://strata.oreilly.com/2012/10/matlab-r-julia-languages-for-data-analysis.html. R, Python and Julia are open source; SAS, SPSS, Stata and Matlab are proprietary.

39. Revolution Analytics, "R Applications and Extensions," 2013, http://www.revolutionanalytics.com/what-is-open-source-r/r-language-features/applications.php.

40. Robert Muenchen, "The Popularity of Data Analysis Software," r4stats.com, 2013, http://r4stats.com/articles/popularity/.

41. Tableau Software, "Tableau Generates Action from Market Research Data," 2013, http://www.tableausoftware.com/learning/applications/studies/tableau-makes-money-chase.

42. Alina Popescu, "Hadoop Advances Healthcare Research–Children's Hospital Use Case | #hadoopsummit," Silicon Angle, 2013, http://siliconangle.com/blog/2013/06/27/leveraging-hadoop-to-advance-healthcare-research-childrens-hospital-use-case-hadoopsummit/, Bala Venkatrao et al, "Taming the Elephant – Learn How Monsanto Manages Their Hadoop Cluster to Enable Genome/Sequence Processing," presentation at Strata Conference New York, 2012, http://strataconf.com/stratany2012/public/schedule/detail/25695, and Christophe Bisciglia, "The Smart Grid: Hadoop at the Tennessee Valley Authority (TVA)," Cloudera Blog, 2009, http://blog.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/.

43. Apache Software Foundation, “Spark: Lightning-Fast Cluster Computing,” 2013, http://spark.incubator.apache.org.

44. Apache Software Foundation, "Powered by Spark," Apache Wiki, 2013, https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark.

45. Statistic Brain, "Average Cost of Hard Drive Storage," 2013, http://www.statisticbrain.com/average-cost-of-hard-drive-storage/.

46. Daniel Castro, "Explaining International IT Application Leadership: Health IT," The Information Technology and Innovation Foundation, 2009, http://www.itif.org/files/2009-leadership-healthit.pdf.

47. Mediterranean Decision Support System for Maritime Safety, "Project Outline," 2013, http://www.medess4ms.eu/project-outline.

48. MarketsandMarkets, "Internet of Things (IoT) & Machine-To-Machine (M2M) Communications Market Research Report," 2013, http://www.marketsandmarkets.com/Market-Reports/internet-of-things-market-573.html.

49. Travis Korte, “Data Scientists Should be the New Factory Workers,” The Center for Data Innovation, 2013, http://www.datainnovation.org/2013/09/data-scientists-should-be-the-new-factory-workers/, "Music recommendations at Spotify - Erik Bernhardsson," Vimeo video, 1:27:37, 2013, http://vimeo.com/57900625, and David Gelles, "Inside Match.com," FT Magazine, 2011, http://www.ft.com/intl/cms/s/2/f31cae04-b8ca-11e0-8206-00144feabdc0.html.

50. Travis Korte, “Mapping the Scientific Research Landscape,” The Center for Data Innovation, 2013, http://www.datainnovation.org/2013/10/mapping-the-scientific-research-landscape/.

51. The Recovery Accountability and Transparency Board, "Map Gallery," 2009, http://www.recovery.gov/Transparency/Pages/Maps.aspx#z.

52. Data.gov, “Data.gov Releases Open Source Software,” 2012, http://www.data.gov/welcome-open-government-platform.

53. "Uploads from Gartner Gartner," YouTube video, 2:21, posted by "Gartner Gartner," 2012, http://www.youtube.com/watch?v=mXLy3nkXQVM.

54. Claire Cain Miller, "Data Science: The Numbers of Our Lives," The New York Times, 2012, http://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html.

55. James Manyika et al, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

56. The Executive Office of the President of the United States, "Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments," 2012, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

57. Robert Atkinson, "The Research and Experimentation Tax Credit: A Critical Policy Tool for Boosting Research and Enhancing U.S. Economic Competitiveness," The Information Technology and Innovation Foundation, 2006, http://www.itif.org/files/R&DTaxCredit.pdf.

58. Daniel Castro, “The Need for an R&D Roadmap for Privacy,” Information Technology and Innovation Foundation, 2012, http://www2.itif.org/2012-privacy-roadmap.pdf.

59. Stanley Ahalt, PhD, “Establishing a National Consortium for Data Science,” The National Consortium for Data Science, 2012,

http://data2discovery.org/dev/wp-content/uploads/2012/09/NCDS-Consortium-2-pager_0709121.pdf.

60. Securities and Exchange Commission, “Interactive Data to Improve Financial Reporting,” 2009, http://www.sec.gov/rules/final/2009/33-9002.pdf.

61. Travis Korte, “Research Data Alliance Tackles the Rising Tide of Scientific Data,” The Center for Data Innovation, 2013, http://www.datainnovation.org/2013/07/research-data-alliance-tackles-the-rising-tide-of-scientific-data/.

62. Daniel Castro and Travis Korte, “G8 Charter Puts Open Data on International Agenda,” The Center for Data Innovation, 2013, http://www.datainnovation.org/2013/07/g8-charter-puts-open-data-on-international-agenda/.

63. Stephen Ezell, Robert Atkinson and Michelle Wein, "Localization Barriers to Trade: Threat to the Global Innovation Economy," The Information Technology and Innovation Foundation, 2013, http://www.itif.org/publications/localization-barriers-trade-threat-global-innovation-economy.

ABOUT THE AUTHORS

Daniel Castro is the director of the Center for Data Innovation. Mr. Castro writes and speaks on a variety of issues related to information technology and internet policy, including privacy, security, intellectual property, internet governance, e-government, and accessibility for people with disabilities. He has a B.S. in Foreign Service from Georgetown University and an M.S. in Information Security Technology and Management from Carnegie Mellon University.

Travis Korte is a research analyst at the Center for Data Innovation specializing in data science applications and open data. He has a background in journalism, computer science and statistics. He graduated with highest honors from the University of California, Berkeley, having studied critical theory and completed coursework in computer science and economics.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation conducts high-quality, independent research and educational activities on the impact of the increased use of data on the economy and society. In addition, the Center for Data Innovation formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities, and improve quality of life. The Center for Data Innovation also sponsors the annual Data Innovation Day.

contact: info@datainnovation.org

datainnovation.org