CHAPTER 1.8

From Big Data to Big Social and Economic Opportunities: Which Policies Will Lead to Leveraging Data-Driven Innovation's Potential?

PEDRO LESS ANDRADE
JESS HEMERLY
GABRIEL RECALDE
PATRICK RYAN
Public Policy Division, Google, Inc.

Over the last few years, myriad examples of innovation in data analysis have emerged, creating new business models for data-driven innovation. For example, businesses are developing ways for real-time weather information to be communicated to devices in the field that can advise farmers on pest activity, water supply, and inclement weather.[1] The Royal Netherlands Meteorological Institute has found a way to generate extremely accurate rainfall information using nothing more than existing data from cell-tower installations.[2] The next phase of the Internet's evolution has us on a clear path toward a "revolution of data."[3] Every year, the costs associated with the production, collection, storage, and dissemination of data come down, making those data more readily available. This process is fomented by the increasing migration of many social and economic activities to the web.[4] More data are generated today than ever before; this is a positive trend that will inevitably continue: 90 percent of the world's information generated through the history of mankind has been generated over the last two years,[5] while data generated per year is growing at a rate of 40 percent.[6]

In this chapter we will focus on the social and economic value of data, but from the point of view of use and purpose rather than volume. We will therefore talk about data driven-innovation instead of "big data," and will provide case studies from different areas, with a special consideration of how data-driven innovation in the public sector could improve policymaking. We will finish the chapter by describing the main issues that should be addressed by policymakers, who can leverage the potential of data-driven innovation in their communities through forward looking policies.

WHY SPEAK OF DATA-DRIVEN INNOVATION INSTEAD OF BIG DATA?

It has become axiomatic that more data are produced every year, and somehow this phenomenon has driven commentators to call this revolution "the age of big data." However, what is commonly known as big data is not a new concept, as the use of data to build successful products and services, optimize business processes, or make more efficient data-based decisions already has an established history. Innovative uses of data have been key to developing new products and making more efficient decisions for quite a long time, and these activities have become more common and more efficient with the availability of modern computing. Crunching data, statistics, and trends in new ways has always helped change the way that entire sectors operate. Agriculture is one of the first major sectors to have benefitted from the aggregation and analysis of data: in 1793, the *Farmer's Almanac* found a

The opinions here are the views of the authors and do not represent the views and positions of their employer.

niche when it published its first report more to provide landowners with guidance for what they might expect in the coming year and to plan their crops accordingly.[7] It took 183 years for paper-based, manpower-intensive analysis to become automated, but in 1950, a team of meteorologists used the Electronic Numerical Integrator and Computer (ENIAC) computer to make the first successful numerical weather prediction.[8]

Moreover, the term *big data* is ambiguous, and it sets up data as a negative because of the implication that "big" is "bad." Indeed, many common definitions of big data usually focus not on size but instead on various characteristics, including the frequency of production, speed, volume, variety, and capacities needed to manage and process information.[9] McKinsey, for example, describes *big data* as "datasets whose size is beyond the ability of typical database tools to capture, store, manage and analyze."[10] The implications of this definition are that the main features of big data (quantity, speed, variety) are technical properties that depend not on the data itself, but instead on the evolution of computing, storage, and processing technologies.[11] What may look like big data today will not likely be as "big" in the near future.

Thus, what is important about data is not their volume, but how they may contribute to innovation and therefore be used to create value. Data alone do not possess inherent value; instead it is the processing of data in innovative ways that brings new economic and social benefits, and this value creates a virtuous circle to feed into more use of data-based decision-making and analysis.[12] In other words, it is the use of data that really matters.[13] One way to measure this value is to measure the socioeconomic metrics (or to estimate the future potential) obtained from the use of data. The excitement that we are seeing with new deployments of data to fuel innovation is not just because of the volume of data, nor is it about the data themselves. As pointed out by the Software and Information Industry Association, "transformative data can be big or small or even the 'needle' of data found in a giant haystack."[14]

The truth is that data are data, and that has not changed for centuries. When "big data" is no longer a trendy concept, data will continue to drive innovation, and solutions for new problems will come from new ways of analyzing and interpreting data, regardless of volume or our technological capacities to manage it. In the next section, we will address what we see in the future for data-driven innovation.

THE BENEFITS OF DATA-DRIVEN INNOVATION

Many sectors benefit from data-driven innovation: healthcare (e.g., diagnosis and treatment), financial services (e.g., analyzing market trends and economic conditions), and transportation and public administration (e.g., metrics on what citizens want and where economic development is headed), to name a few. In one example, a philanthropic research center stores and analyzes the cancer genome and the sequences and mutations of more than 10,000 cancer cases to understand the complexity of the disease.[15] In another recent project, a university-based group of academics mined data from 60 years of historical weather records to identify the factors that are most predictive of hurricane activity.[16] In the private sector, PayPal has developed a system that screens payments and combines them with IP addresses, browser information, and other technical data to identify and prevent fraudulent activity in online payments, bolstering trust for commercial exchanges on the Internet.[17] A startup firm has developed a no-cost platform for users that helps travelers predict flight delays using an algorithm that scours data on every domestic flight for the past 10 years and matches it to real-time conditions.[18] Finally, the United Nations is working with governments around the world to understand global trends related to hunger, poverty, disease, and job loss.[19]

However, because data-driven innovation takes place across various sectors of the economy and society, it is sometimes difficult to quantify its full economic impact. For example, using any traditional measure of Gross Domestic Product (GDP), Wikipedia does not contribute any economic value.[20] This makes no sense in today's world, as economists are now demonstrating that a fundamental problem exists in our ability to quantify the value of data, and this gap misleads policymakers in their drive to maximize economic surplus.[21] As Michael Mandel has observed, "economists have been systematically trained to think of the economy as divided into two big categories: 'goods' and 'services.'"[22] Data are neither a good nor a service and so they escape traditional economic analysis. This highlights the complication of discussing data: although the value often creates an economic reward, such measurements are not easy to make. The Internet itself has been a strong contributor to economic growth for more than two decades, but only in the past couple of years have economists undertaken serious attempts to quantify the Internet's impact on the world's economies.[23]

One example of innovative data use that has a difficult-to-quantify economic value proposition is Google's Flu Trends, which provides near real-time estimates of flu activity for a number of countries around the world. Flu Trends provides its analysis based on aggregated search queries.[24] Some of these estimates have been compared with official historic influenza data from relevant countries with surprisingly high levels of accuracy, and in some cases Flu Trends provided information weeks ahead of official records. In the case of H1N1, the world's citizens were searching online long before official statistics were available.[25] Additionally, the data from Flu Trends are open, available for everybody

to download and use. A group of researchers from the Johns Hopkins University, for example, used these data to develop a practical influenza forecast model designed to provide medical centers with advance warning of the expected number of flu cases, thus allowing sufficient time to implement interventions before outbreaks.[26] This example illustrates how the openness and accessibility of data are crucial to keeping the wheel of innovation rolling by allowing others to access and manipulate the data in transformative ways.

Similarly, the rapid collection and processing of information has helped in recent natural disasters. After a devastating earthquake hit Haiti in 2010, a group of researchers from the Karolinska Institute and Columbia University analyzed calling data of over 2 million mobile phones to detect the pattern of population movements across the country. This information was then handed to humanitarian agencies to allocate relief resources in a more efficient way.[27] The findings of the project, called Flow Minder, suggest that population movements during disasters may be more predictable than had previously been understood.[28]

These examples show that there are ethical and responsible ways of analyzing big sets of data and equally ethical and responsible ways of using them to provide high-value solutions for citizens, whether or not they have a clear quantifiable economic value at the outset. More efficient preparation for outbreaks and better understanding of post-disaster movement ultimately mean more cost-effective deployment of public services.

High-value products and services and more efficient deployment of resources are not the only outcomes of data-driven innovation. Studies suggest that there is a direct connection between data-driven decision-making in business and improved firm performance. Firms that adopt data-driven decision-making have an output and productivity that is 5 percent to 6 percent higher than would be expected, given their other investments and their information technology (IT) usage.[29] Another study has shown that the use of Internet computing tools can also help firms reach decisions more efficiently, across a broad range of industries, as they allow firms of all sizes to leverage data-driven analysis without needing to make huge investments in their IT infrastructure.[30]

As is the case for businesses, policymakers are entrusted to make decisions for the citizenry with very little information. Politicians recognize the need to base their recommendations on objective information, and they are expected to move quickly, just as business managers are.

In fact, the public sector is one the most data-intensive sectors of all. According to McKinsey, the US government had over 848 petabytes of data stored in 2009—second only to the manufacturing sector.[31] What is usually known as "data-driven policymaking" involves the collection of information related to how roads are traveled, to determine trends in utility consumption and the provision of government services (Box 1), and to promote creativity and new ideas within government agencies.[32] Statistical agencies inside of governments, such as census departments, have long been established to maintain data about the nation. Thus data-driven policymaking is not new, but the opportunities brought by the advances on information and communication technologies make data-driven policymaking increasingly accessible to government officials. Further, open government initiatives put these data into the hands of the public, facilitating a new kind of transparency and civic engagement for curious and interested citizens. Data can benefit society when they are open.[33]

By providing a way to check assumptions, detect problems, clarify choices, prioritize resources, and identify solutions, data-driven policymaking injects data-based rationality into the policymaking process, all of which could also create economic benefits.[34] According to the Organisation for Economic Co-operation and Development (OECD), by fully exploiting public data, governments in the European Union could reduce administrative costs by 15 percent to 20 percent, creating the equivalent of €150 billion to €300 billion.[35] In other words, data-driven policymaking moves policymaking out of the realm of intuition and dogma by creating a sound evidentiary basis for decisions.

However, studies suggest that the public sector still does not fully exploit the potential of the data it generates and collects, nor does it exploit the potential of data generated elsewhere. The "revolution of data" still needs to make its way within government agencies. Although the government is one of the sectors with the greatest potential to capture value from data-driven innovation, it also has one of the lowest productivity growth rates because it lags behind business and industry in fully embracing data.

Box 1: Hong Kong Efficiency Unit

The Hong Kong Efficiency Unit acts as a single point of contact for handling public inquiries and complaints on behalf of many government departments. After collecting thousands of complaints each year, its staff recognized the social messages hidden in the complaints data, which in fact provided important feedback on public service. Using a platform called the "Complaints Intelligence System," they now use the complaints information collected to gain a better understanding of daily issues by uncovering trends, patterns, and relationships inherent in the complaints.

Source: Government of the Hong Kong Special Administrative Region, Hong Kong Efficiency Unit, 2013.

Box 2: Harvard Transparency Project

The Transparency Policy Project at Harvard's Kennedy School studied the relationship between transit data format and accessibility and the number of applications for that system. Of the five transit agencies they studied, the TriMet in Portland, Oregon, and the Massachusetts Bay Transit Authority (MBTA) in Boston have generated the highest ratio of applications per transit rider (1 to 7,000 and 1 to 27,000, respectively). Meanwhile, the most reluctant agency to adopt open data, Washington DC's Metro, had only 10 applications serving its customers in 2012 (1 to 121,400).

Source: Rojas, 2012.

SETTING THE STAGE FOR A DATA-DRIVEN ECONOMY

Apart from producing and using data for better policymaking processes, the public sector can also play its part by promoting and fostering data-driven innovation and growth throughout economies. To realize the potential of data-driven innovation, policymakers need to develop coherent policies for the use of data. This could be achieved by: (1) making public data accessible through open data formats, (2) promoting balanced legislation, and (3) supporting education that focuses on data science skills.

Open data initiatives

The use of data across sectors can drive innovation and economic growth. However, many generators of data—including governments—do not share their data. As we have seen, the public sector is one of the main producers and collectors of data. Open data initiatives that make data in the public sector accessible to everyone contribute to data-driven innovation and create value for governments. For example, aggregate public transport data may be used by developers to create useful applications for passengers (see Box 2). This access to real-time information could result in a greater number of passengers and, subsequently, to more income for the transport authorities. In addition, accessible public data usually lead to better data because data users can test structure and help to fix mistakes (see Box 3). Improvements in the quality of data mean better data-based solutions and, ultimately, better policy.

It is important to note that opening up public data does not necessarily lead to the disclosure of personal data. Public data that may contain personal information of citizens should be shared in an aggregate or fully de-identified way to protect citizens' privacy. We will go into more detail around the discussions on privacy and personal data in the following section.

How to get the best of data-driven innovation

The increasing ease of linking and analyzing information usually raises concerns about individual privacy protection. Personal data are the type that has drawn the most attention, from a regulatory point of view, in relation to data-driven innovation. The challenge is to achieve a reasonable balance between individuals' right to privacy and the emerging opportunities in data-driven innovation.

For this reason, in order to capitalize on opportunities for economic growth via innovation, flexible and adaptable policies are needed. We need to focus on using datasets responsibly and ensuring that personally identifiable information is accessible only by those who are authorized to do so, without limiting innovation. In other words, privacy protection frameworks should support secure and reliable data flows while enhancing responsible, risk-reducing behavior regarding the use of personal data.

Legislation should take into account the tension between data-driven innovation and the principle of data minimization. This principle essentially states that the collection of personal data should be limited to what is relevant and necessary to accomplish a specific purpose, and for only as long as necessary. This tension usually materializes in two regulatory discussions: first, the definition of *personal data;* and second, the model of consent by users. These considerations are both critical, but framing things in this way leads to the inevitable conclusion that fewer data are better.

A key dividend of data-driven innovation is the possibility of finding new insights by analyzing existing data and combining them with other data. This can sometimes blur the lines between personal and non-personal data, as well as the uses for which consent may have been given.[36] A practical definition of *personal data* should be based on the real possibility of identifying an individual during the treatment of data.[37] This is why applying existing approaches to personal data may result in overly broad definitions that can have unintended negative consequences for data-driven innovation.

For the same reason that combining and correlating datasets is a key feature of data-driven innovation, the full potential of data collected may not be clear at the time of collection. A consent model that is appropriate to the data-driven economy should provide a path for individuals to participate in research through informed consent. In this model, they would become aware of the benefits of their participation as well as potential privacy risks. For this reason, the legislative considerations for data collection should not assume that less is always more and should take into consideration the data-intensive direction of some of the economy's growing sectors.

Building skills for the future

An economy where both the public and private actors who base their decisions on data analysis will demand highly skilled workers with backgrounds in

Box 3: Can open data lead to better data?

Moscow's city government published about 170 datasets with geo coordinates at the Moscow opendata portal. After examining the data, Russian members of the OpenStreetMap community found many errors and mistakes, including wrong geo coordinates. After publishing their research, most of the issues were solved by Moscow state officials.

Clinicians from the Imperial College London, while reviewing open statistical data from the United Kingdom's National Health Service, found that records said that 20,000 male patients required midwifery services between 2009 and 2010. After this research was published, data systems were improved.

Source: Open Knowledge Foundation, 2013.

data analysis, information science, metadata and data visualization. The demand for engineers who specialize in technologies such as machine learning and natural language processing will also increase, and a gap between the supply and demand for these types of skills may hinder data-driven innovation's full potential. The United States itself will need up to 190,000 more workers with deep analytical expertise by 2018.[38] This clear demand for skilled workers is further evidence of data-driven innovation's potential benefits for economies.

CONCLUSION

We have already begun to see the impact technology has had on the volume and speed at which data may be generated, analyzed, and put to use. Thirty years ago we needed an army of data-entry clerks to feed an information into a system; today, the information is already available in a machine-readable format. We carry devices with sensors that can provide incredible amounts of information in real time. Every day, the world adds petabytes of information into social networks and other Internet platforms.

Talking about this phenomenon as "big data," however, misses the true potential of data. Instead, we should focus our discussion on *data-driven innovation,* as this relates to the results and outcomes of data use—from generating innovative products and service to improving business and government efficiency. Many other examples provided earlier have shown that data-driven solutions have transformative social impact as well.

However, achieving the full potential of data-driven innovation demands challenging the outdated paradigms established in a significantly less data-intensive world. To achieve the maximum benefits from data-driven innovation, policymakers must take into account the possibility that regulation could preclude economic and societal benefits. Decisions that affect data-driven innovation are usually focused on the problems of privacy and data protection, but fail to consider economic and social benefits that regulation could preclude. It is by looking at the big picture surrounding big data that we can create the right environment for data-driven innovation, and that the individuals, organizations, and economies that may benefit from it can thrive.

NOTES

1 Gray 2013.

2 The Economist 2013a.

3 Mayer-Schonberger and Cukier 2013.

4 OECD 2013.

5 IBM 2013.

6 Manyika et al. 2011.

7 The Old Farmer's Almanac, no date, "History of the Farmer's Almanac."

8 Platzman 1979.

9 Hemerly 2013.

10 Manyika et al. 2011.

11 OECD 2013.

12 According to Hilbert (2013, p. 4), "the crux of the 'Big Data' paradigm is actually not the increasingly large amount of data itself, but its analysis for intelligent decision-making."

13 Hemerly 2013.

14 SIIA 2013.

15 Burke 2012.

16 McCormick University 2012. See also Chen 2013.

17 Sims 2011.

18 See www.flightcaster.com.

19 United Nations 2012.

20 Tapscott and Williams 2007.

21 Pélissié du Rausas et al. 2011.

22 Mandel 2012, p. 1.

23 A collection of studies that quantify the contribution of the Internet to GDP is available at www.valueoftheweb.com.

24 The Economist 2013b.

25 The Economist 2011.

26 Dugas et al. 2013.

27 Lu et al. 2012.

28 Talbot 2013.

29 Brynjolfsson et al. 2011.

30 Cacciola and Gibbons 2012.

31 Manyika et al. 2011.

32 Esty and Rushing 2007.

33 Rojas 2012.

34 Esty and Rushing 2007.

35 Manyika et al. 2011.

36 Hemerly 2013.

37 For example, an IP address, by itself, cannot be linked to nor identify an individual, because it identifies only a device connected to a network.

38 Manyika et al. 2011.

REFERENCES

Brynjolfsson, E., L. M. Hitt, and H. H. Kim. 2011. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" April 22. http://dx.doi.org/10.2139/ssrn.1819486.

Burke, A. 2012. "Haussler Says Cancer Genomics Needs Tech Geeks." *New York Genome Center Blog,* November 13. Available at http://nygenome.org/blog/haussler-says-cancer-genomics-needs-tech-geeks.

Cacciola, S. and R. Gibbons. 2012. "Coase Meets the Cloud: How and When Can Outsourcing IT Improve Organizational Performance?" White Paper, November 28. Available at http://web.mit.edu/rgibbons/www/Coase%20Meets%20the%20Cloud_Final.pdf.

Chen, Z., W. Hendrix, H. Guan, I. K. Tetteh, A. Choudhary, F. Semazzi, and N. F. Samatova. 2013. "Discovery of Extreme Events-Related Communities in Contrasting Groups of Physical System Networks." *Data Mining and Knowledge Discovery* 27 (2): 225–58. Available at http://link.springer.com/article/10.1007/s10618-012-0289-3?null.

Dugas, A.F., M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman. 2013. "Influenza Forecasting with Google Flu Trends." *PLoS ONE* 8 (2). Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3572967/pdf/pone.0056176.pdf.

The Economist. 2011. "Attention, Citizens!" *The Economist,* November 16. Available at http://www.economist.com/node/21538656.

——. 2013a. "Counting Raindrops." *The Economist,* February 9. Available at http://www.economist.com/news/science-and-technology/21571384-how-use-mobile-phone-networks-weather-forecasting-counting-raindrops.

——. 2013b. "Have Germs, Will Travel." *The Economist,* January 25. Available at http://www.economist.com/blogs/gulliver/2013/01/tracking-flu-0.

Esty, D. C. and R. Rushing. 2007. *Governing by the Numbers: The Promise of Data-Driven Policymaking in the Information Age.* April. Center for American Progress. Available at http://www.americanprogress.org/wp-content/uploads/issues/2007/04/pdf/data_driven_policy_report.pdf.

Government of the Hong Kong Special Administrative Region. No date. Hong Kong Efficiency Unit Homepage. Available at http://www.eu.gov.hk/eindex.html.

Gray, J. 2013. "Semios: Big Data and Digitizing the Farm." *Canadian Startup News,* August 7. Available at http://www.betakit.com/semios-big-data-and-digitizing-the-farm/.

Hemerly, J. 2013. "Public Policy Considerations for Data-driven Innovation." *Computer* (IEEE Computer Society) 46 (6): 25–31.

Hilbert, M. 2013. "Big Data for Development: From Information- to Knowledge Societies." January 15. Available at http://dx.doi.org/10.2139/ssrn.2205145.

IBM. 2013. "The IBM Big Data Platform." New York: IBM Corporation. Available at http://public.dhe.ibm.com/common/ssi/ecm/en/imb14135usen/IMB14135USEN.PDF.

Lu, X., L. Bengtsson, and P. Holme. 2012. "Predictability of Population Displacement after the 2010 Haiti Earthquake." *PNAS* 109 (29), July 17. Available at http://www.pnas.org/content/early/2012/06/11/1203882109.full.pdf+html.

Mandel, M. 2012. "Beyond Goods and Services: The (Unmeasured) Rise of the Data-Driven Economy." Progressive Policy Institute, Policy Memo. October. Available at http://www.progressivepolicy.org/wp-content/uploads/2012/10/10.2012-Mandel_Beyond-Goods-and-Services_The-Unmeasured-Rise-of-the-Data-Driven-Economy.pdf.

Manyika, J., M. Chui, B. Brown, J. Bughin, R, Dobbs, C. Roxburgh, and A. H. Byers. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." *McKinsey Global Institute Report,* May. Available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Mayer-Schonberger, V. and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* New York: Houghton Mifflin Harcourt.

McCormick University, 2012. "Big-Data Approach Leads to More Accurate Hurricane Forecasting." *News from McCormick,* September 25. Available at http://www.mccormick.northwestern.edu/news/articles/2012/09/more-accurate-hurricane-forecasting-using-big-data.html.

OECD (Organisation for Economic Co-operation and Development). 2013. "Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by 'Big Data'." *OECD Digital Economy Papers* 222, June 18. Available at http://dx.doi.org/10.1787/5k47zw3fcp43-en.

The Old Farmer's Almanac. No date. "History of the Farmer's Almanac." Available at http://www.almanac.com/content/history-old-farmers-almanac.

Open Knowledge Foundation. 2013. "How Can Open Data Lead to Better Data Quality?" September 3. Available at http://blog.okfn.org/2013/09/03/how-can-open-data-lead-to-better-data-quality/.

Platzman, G. W. 1979. "The ENIAC Computations of 1950: Gateway to Numerical Weather Prediction." *Bulletin of the American Meteorological Society* 60 (4): 302–12.

Pélissié du Rausas, P., J. Manyika, E. Hazan, J. Bughin, M. Chui, and R. Said. 2011. *Internet Matters: The Net's Sweeping Impact on Growth, Jobs, and Prosperity.* McKinsey Global Institute, McKinsey & Company, May. Available at http://www.mckinsey.com/insights/high_tech_telecoms_internet/internet_matters.

Rojas, F. M. 2012. "Transit Transparency: Effective Disclosure through Open Data." Transparency Policy Project, Harvard Kennedy School, June. Available at http://www.transparencypolicy.net/assets/FINAL_UTC_TransitTransparency_8%2028%202012.pdf.

SIIA (Software and Information Industry Association). 2013. "Data-Driven Innovation: A Guide for Policymakers: Understanding and Enabling the Economic and Social Value of Data." *SIIA White Paper.* Available at http://goo.gl/QWjGhY.

Sims, D. 2011. "Big Data Thwarts Fraud." *O'Reilly Strata,* February 8. Available at http://strata.oreilly.com/2011/02/big-data-fraud-protection-payment.html.

Talbot, D. 2013. "Big Data from Cheap Phones." *MIT Technology Review,* April 23. Available at http://www.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/.

Tapscott, D. and A. Williams. 2007. *Wikinomics: How Mass Collaboration Changes Everything.* New York: Portfolio Trade Books.

United Nations. 2012. "Big Data for Development: Challenges & Opportunities." Global Pulse, May. Available at http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.