CHAPTER 1.4

Big Data: Balancing the Risks and Rewards of Data-Driven Public Policy

ALEX PENTLAND
MIT

In June 2013, massive US surveillance of phone records and Internet data was revealed by former National Security Agency (NSA) contractor Edward Snowden, who called these activities the "architecture of oppression." His disclosures ignited an overdue public debate on the balance between personal privacy and our growing digital capabilities regarding the collection and use of personal data. Finding this balance is an issue of vital and urgent interest to corporations and governments as well as to ordinary citizens around the world. This chapter will outline both the risks and the rewards of this new age of big data, address policy issues in this area, and provide practical recommendations for a way forward.

Data about human behavior, such as census data, have always been essential for both government and industry to function. In recent years, however, a new methodology for collecting data about human behavior has emerged. By analyzing patterns within the "digital breadcrumbs" that we all leave behind us as we move through the world (call records, credit card transactions, and global positioning system, or GPS, location fixes, for example), scientists are discovering that we can begin to explain many things—such as financial crashes, revolutions, panics—that previously appeared to be random events. These new tools, with the perspective they provide on life in all its complexity, shape the future of social science and public policy. Just as the microscope and telescope revolutionized the study of biology and astronomy, "socioscopes" have the potential to revolutionize regulation and public policy.

The risk of deploying this sort of data-driven policy and regulation comes from the danger of putting so much personal data into the hands of either companies or governments. Fortunately, new approaches to regulation and technology that can help protect personal privacy from exploitation have been developed. These approaches can mitigate the problem of government overreach as well. Both regulation and technology must continue to evolve in order to provide more scientific, real-time public policy while protecting citizens from the dangers of exploitative companies or an all-knowing authoritarian government. This chapter will provide practical recommendations to achieve these goals.

A BIG DATA TAXONOMY

It is probably hopeless to try to provide a detailed taxonomy of data types and uses because the technology is progressing so quickly. But it is possible to provide a broad taxonomy framed in terms of control. The three main divisions within the spectrum of data control are: (1) *data commons,* which are available to all, with at most minor limitations on use; (2) *personal or proprietary data,* which are typically controlled by individuals or companies, and for which legal and technology infrastructure must provide strict control and auditing of use; and (3) the *secret data of governments,* which typically has less direct public oversight and more stringent controls. The issues of data commons will be addressed first, followed by concerns about personal and proprietary data, and, finally, issues of secret government data.

The preferred lens for examining these issues is experimentation in the real world rather than arguments from theory or first principles, because using massive, live data to design institutions and policies is outside of our traditional way of managing things. In this new digital era we cannot rely only on existing policy, tradition, or even laboratory science, because the strengths and weaknesses of big data analysis are very different from those obtained through standard information sources. To begin to manage our society in a data-driven manner requires us to move beyond academic debate and laboratory question-and-answer processes. Instead, we need to try out new policy ideas within living laboratories—real, diverse communities that are willing to try a new way of doing things—in order to test and prove our ideas. This is new territory and so it is important for us to constantly try out new ideas in the real world in order to see what works and what does not (see Box 1).

Data commons

The first entry in the data taxonomy is the *data commons.* A key insight is that our data are worth more when shared because they can inform improvements in systems such as public health, transportation, and government. Using a "digital data commons" can potentially give us unprecedented ability to measure how our policies are performing so we can know when to act quickly and effectively to address a situation.

We already have many data commons available: maps, census data, and financial indices, for example. With the advent of big data, we can potentially develop many more types of data commons; these commons can be both accessible in real time and far more detailed than previous, hand-built data commons (e.g., census data, etc.). This is because the new digital commons depend mostly on data that are already produced as a side effect of ongoing daily life (e.g., digital transaction records, cell phone location fixes, road toll records, etc.), and because they can be produced automatically by computers without human intervention.

One major concern with these new data commons is that they can endanger personal privacy. Another, secondary, concern involves the tension between proprietary interests, both commercial and personal, and the goal of putting data in the commons. Acceding to these proprietary interests might tend to reduce the richness of such a commons, which would diminish the ability of such a data commons to enable significant public goods.

To explore the viability of a big data commons, what is perhaps the world's first true big data commons was unveiled on May 1, 2013. In this Data for Development (D4D) initiative, 90 research organizations from around the world reported hundreds of results from their analysis of data describing the mobility and call patterns of the citizens of the entire African country Côte d'Ivoire.[1] The data were donated by the mobile carrier Orange, with help from the University of Louvain (Belgium) and the MIT Human Dynamics Laboratory (United States), along with collaboration from Bouake University (Côte d'Ivoire), the United Nation's Global Pulse, the World Economic Forum, and the GSMA (the mobile carriers' international trade association). The D4D program was led by Nicolas De Cordes (Orange), Vincent Blondel (Louvain), Alex Pentland (MIT), Robert Kirkpatrick (UN Global Pulse), and Bill Hoffman (World Economic Forum).

The research projects conducted by the 90 participating organizations explored the use of this data commons, covering many different aspects of better governance. An example of using the D4D data to improve social equality was highlighted by work done by researchers at the University College of London, who developed a method for mapping poverty from the diversity of cell phone usage. As people have more disposable income, they explore or sample their environment more, and their patterns of movement and patterns of phone calls become increasingly diverse. Measurement of this additional exploration allows us to make a surprisingly accurate estimate of their disposable income. Another example of using the D4D data to enhance social equality is the mapping of ethnic boundaries by researchers from the University of California, San Diego. This method relies on the fact that ethnic and language groups communicate far more within their own group than they communicate with other groups. This project is significant because, while we know that ethic violence often erupts along such boundaries, the government and aid agencies are usually uncertain about the geography of these social fault zones.

The D4D data were also utilized to understand and promote operational efficiency through an analysis of Côte d'Ivoire's public transportation system by IBM's Dublin laboratory. This analysis showed that, for very little cost, the average commute time in Abidjan—Côte d'Ivoire's biggest city—could be cut by 10 percent. Other research groups demonstrated similar potential for operational improvements in the areas of government, commerce, agriculture, and finance.

Finally, examples of using D4D data to improve social resiliency include analysis of disease spread by groups from Novi Sad University (Serbia), École Polytechnique Fédérale de Lausanne (EPFL, Switzerland), and Birmingham (United Kingdom). These research groups showed that small changes in the public health system could potentially cut the spread of flu by 20 percent as well as significantly reduce the spread of HIV and malaria.

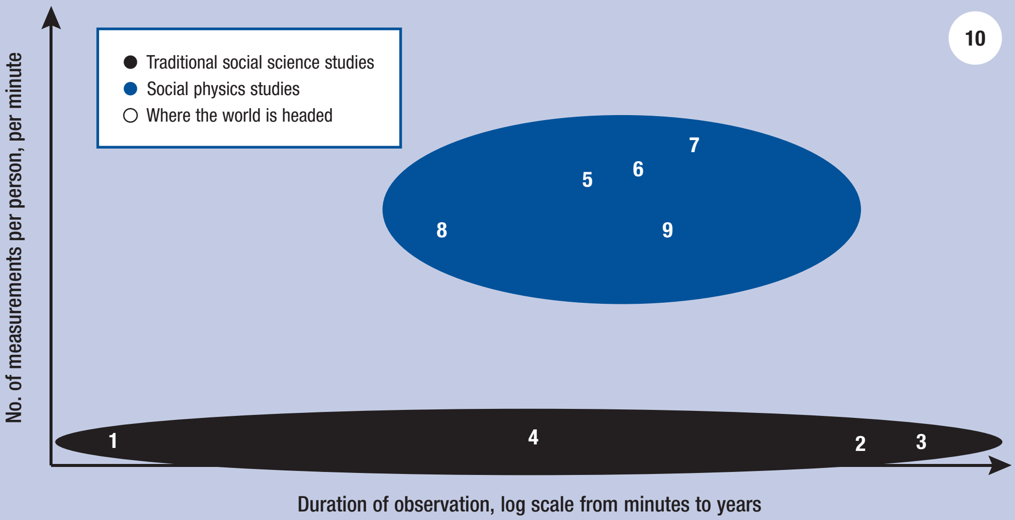
Box 1: The future of big data and governance

The Data for Development (D4D) data commons is only a small first step toward improving governance by using big data. Much more can be accomplished because our current understanding of policy and human society is based on very limited data resources. Currently, most social science is based either on analysis of laboratory experiments or on survey data. These approaches miss the critical fact that it is the details of which people you interact with, and how you interact with them, that truly matter. Social phenomena are made up of billions of small transactions between individuals—people trading not only goods and money but also information, ideas, or just gossip. There are patterns in those individual transactions that drive phenomena such as financial crashes and Arab Springs. We need to understand these micro-patterns because they do not just average out to the classical way of understanding society. Big data gives us—for the first time—a chance to view society in all its complexity, composed of millions of networks of person-to-person exchanges.

Figure A compares social science living labs with traditional experiments. The horizontal axis presents the duration of the data collection; the vertical axis shows the richness of the information collected.

If we had an all-seeing view, we could potentially arrive at a true understanding of how society works and develop scientifically proven methods to fix our problems. Unfortunately, as illustrated in Figure A, almost all data from traditional social science (labeled “1” in the figure) are near the (0,0) coordinate, meaning that these datasets represent information gathered from under a hundred people and for only for a few hours. The studies labeled “2” and “3” are some of the largest social science studies to date.[1] In the last decade, computational social scientists have begun to discover how to leverage big data and have been using datasets from companies such as cell phone carriers and social media firms. Typical examples of these studies are labeled “4.”

Figure A: Qualitative overview of social science living labs and traditional experiments



Note: Datasets identified in the figure are derived from the following: 1 = most social science experiments, 2 = the Midwest Field Station Study, 3 = the Framingham heart study, 4 = large call record datasets, 5 = reality mining, 6 = social evolution, 7 = friends and family, 8 = sociometric badge studies, 9 = the D4D dataset, and 10 = where the world is headed (see text for explanation).

Unfortunately, even these large datasets are impoverished because they measure only a few variables at a time, thus providing only a very limited view of human nature. Recently data scientists have developed living lab technologies for harvesting digital breadcrumbs, and are now obtaining much richer descriptions of human behavior. The studies labeled “5,” “6,” “7,” and “8” are living lab studies that use smart phones or electronic name badges (sociometers) to collect data.[2] The point labeled “9” is the D4D dataset that covers the entire country of Côte d’Ivoire.[3]

Just a brief examination of Figure A makes it easy to see that these living lab datasets are many orders of magnitude richer than previous social science datasets. These large, digital datasets contain extraordinary amounts of objective, continuous, dense data that allow us to build quantitative, predictive models of human behavior in complex, everyday situations.

Importantly, the point labeled “10” shows where the world is headed. In just a few short years we are likely to have available incredibly rich data about the behavior of virtually all of humanity on a continuous basis. The data mostly already exist in cell phone networks, credit card databases, and elsewhere, but currently only technical gurus have access to them. As these digital data become more widely available for scientific inquiry, we will be able to understand and manage ourselves in ways better suited to our complex, interconnected, and networked society.

Notes

[1] See Barker 1968; Dawber 1980.

[2] For details about these living lab studies, see Pentland 2014; Mobile Territorial Lab (MTL), available at http://www.mobileterritoriallab.eu/.

[3] See the D4D challenge, available at http://www.d4d.orange.com/home.

These selected results are just a small sample of the impressive work that is made possible by this rich and unique data commons. These results and others like them are available at http://www.d4d.orange.com/home. Each of these D4D research projects has demonstrated the great potential of a big data commons for improving people's living conditions. From the point of view of Orange, it also demonstrates the potential for new lines of business that combine this data commons with customers' personal data: imagine phone applications that advise commuters about which bus will get them to work quickest, or that help citizens reduce their risk of catching the flu.

The work of these 90 research groups also suggests that many of the privacy fears associated with the release of data about human behavior may be generally misunderstood. In this data commons, the data were processed by advanced computer algorithms (e.g., sophisticated sampling and the use of aggregated indicators) so that it was unlikely that any individual could be re-identified. In fact, no path to re-identification was discovered even though several of the research groups studied this specific question.

In addition, although the data were freely available for any legitimate research in which a group was interested, the data were distributed under a legal contract that specified that they could be used only for the purpose proposed and only by the specific people making the proposal. A similar technology-legal framework is used in trust networks described in the next section. The use of both advanced computer algorithms and contract law to specify and audit how personal data may be used and shared is the goal of new privacy regulations in the European Union, the United States, and elsewhere.

Personal and proprietary data

The second category in the data taxonomy is *personal and proprietary data,* which are typically controlled by individuals or companies, and for which legal and technology infrastructure that provides strict control and auditing of use is needed. The current best practice is a system of data sharing called *trust networks.*[2] Trust networks are a combination of a computer network that keeps track of user permissions for each piece of personal data and a legal contract that specifies both what can and cannot be done with the data and what happens if there is a violation of the permissions. This is the model of personal data management that is most frequently proposed within the World Economic Forum Personal Data Initiative.

In such a system, all personal data have attached labels specifying what the data can, and cannot, be used for. These labels are exactly matched by terms in a legal contract between all the participants stating penalties for not obeying the permission labels and giving the right to audit the use of the data. Having permissions, including the provenance of the data, allows automatic auditing of data use and allows individuals to change their permissions and withdraw their individual data.

Today, long-standing versions of trust networks have proven to be both secure and robust. The best known example is the SWIFT network for inter-bank money transfer; its most distinguishing feature is that it has never been hacked. When asked why he robbed banks, bank robber Willie Sutton famously said, "Because that's where the money is." In today's world, the SWIFT network is where the money is—trillions of dollars are moved through the network each day. This trust network has not only kept the robbers away, but it also makes sure the money reliably goes where it is supposed to go. Until recently, such systems were available only to the "big guys." To give individuals a similarly safe method of managing personal data, the MIT Human Dynamics Laboratory (http://hd.media.mit.edu), in partnership with the Institute for Data Driven Design (http://idcubed.org), have helped build openPDS (open Personal Data Store)—a consumer version of this type of system. We are now testing it with a variety of industry and government partners.[3]

A major concern about trust networks is the cost associated with keeping track of permissions and supporting the capability for automated auditing. Since many companies already maintain such data structures in order to support internal compliance and auditing functions, the cost concern does not appear to be a major barrier. Another more serious concern, however, is the extent to which incidental data about human behavior must be included in the permissions and auditing framework. Such data are typically collected in the course of normal operations in order to support those operations (e.g., the location of a cell phone is required to complete phone calls), but without specific informed consent. A final concern is that a trust network system may be too complex for average people to use, or that it will not inspire (or deserve) the sort of user trust that the name suggests.

In order to investigate these concerns, a living lab has been launched with the city of Trento in Italy, supported by Telecom Italia, Telefonica, the MIT Human Dynamics Laboratory, the Fondazione Bruno Kessler, the Institute for Data Driven Design, and local companies within Trento. Importantly, this living lab has the approval and informed consent of all its participants—they know that they are part of a real-world experiment whose goal is to invent a better way of living.[4]

The objective of this living lab is to develop new ways of sharing data to promote greater civic engagement and information diffusion. One specific goal is to build upon and test trust-network software such as the openPDS system by deploying a set of "personal data services" designed to enable users to collect, store, manage, disclose, share, and use data about themselves. For example, the openPDS system lets the community of young families learn from each other without the work of entering data by hand or the risks associated with sharing through current social media. These data can then be used for the personal self-empowerment of each member, or (when aggregated) for the creation of a data commons that supports improvement of the community—for example, a map that shows disposable income for each neighborhood can stimulate better distribution of community services. The ability to share data safely should enable better idea flow among individuals, companies, and government; we want to see if these tools can in fact increase productivity and creative output at the scale of an entire city.

The Trento living lab will also investigate how to deal with the sensitivities of collecting and using deeply personal data in real-world situations. For example, it will explore different techniques and methodologies to protect the users' privacy while at the same time being able to use personal data—typically mobility, financial, and medical records—to generate a useful data commons. It will also explore different user interfaces for privacy settings, for configuring the data collected, for the data disclosed to applications, and for those data shared with other users, all in the context of a trust framework. Although the Trento experiment is still in its early days, the initial reaction from participating families is that these sorts of data-sharing capabilities are valuable, and they feel safe sharing their data using the openPDS system.

Government data

The third category in the taxonomy is *secret government data.* A major risk of deploying data-driven policies and regulations comes from the danger of putting so much personal data into the hands of governments. But how can it happen that governments, especially authoritarian governments, choose to limit their reach? The answer is that unlimited access to data about the citizen behavior is a great danger to the government as well as to its citizenry. Consider the NSA's response to the recent Snowden leaks:

> "This failure originated from two practices that we need to reverse," Ashton B. Carter, the deputy secretary of defense, said recently. "There was an enormous amount of information concentrated in one place," he said. "That's a mistake." And second, no individual should be given the kind of access Mr. Snowden had, Mr. Carter said.[5]

That is, the government must organize big data resources in a distributed manner, with each different type of data separated and dispersed among many locations, using many different types of computer systems and encryption. Similarly, human resources should be organized into cells of access and permission that are localized both spatially and by data type. Both computer and human resources should always be redundant and fragmented in order to avoid overly powerful central actors.

The logic behind this observation is that databases that have different types of data that are physically and logically distributed, and that also have heterogeneous computer and encryption systems, are hard to attack, both physically as well as through cyberattack. This is because any single exploit is likely to gain access to only a limited part of the whole database. Similarly, the resilience of organizations with a heterogeneous cell-like human and permissions structure is familiar from intelligence and terrorist organizations. Importantly, resistance to attack by adopting a distributed organization is a particularly pressing issue for authoritarian governments, because unfettered access to data about citizen behavior can be a major aid to organizing a successful coup to overthrow the government.

What does all this have to do with the danger that a big data government will trample individual freedoms? The key insight is that for these types of data systems, each type of data analysis operation has a characteristic pattern of communication between different databases and human operators. As a consequence, it is possible to monitor the functioning of the data analysis process without gaining access to, or endangering, the analysis content. In short, one can use "metadata about metadata" in order to monitor the *use* of metadata, and with some reasonable confidence one can ensure that only normal and usual analysis operations are being conducted without reference to specific content. Governments that structure their data resources in this manner can more easily monitor attacks and misuse of all sorts.

As a concrete example, let us assume a system in which different types of databases are physically distributed. In this case one can observe the amount and pattern of traffic between the different databases. These patterns are characteristic of the analysis being performed, and so deviations from the normal patterns of communication between databases are cause for concern. In this manner, an open civil authority can perform substantial, fairly effective monitoring of the functioning of a classified agency. In most cases it is sufficient that each element of the system monitor only local traffic.

A familiar example of this type of monitoring is the "many eyes" security strategy. When patterns of communication among different departments are visible (as with physical mail), then the patterns of normal operations are also visible to many employees, even though the content of the operations (the content of the requested records) remains hidden. For example, a health official responsible for maintaining health records will be able to see if those records are suddenly being accessed by the finance records office with unusual frequency, and may inquire if that is proper. In contrast, when copies of all the data types are all in one place (as when all the records are located in one filing cabinet), it is easy for people to conduct unauthorized analyses.

The computer architecture for the type of system that relies on multiple, distributed types of oversight is very similar to that of the trust networks described in the previous section: distributed data stores with permissions, provenance, and auditing for sharing among data stores. In this case, however, the data stores are segmented by their referent—for example, tax records for individuals, tax records for companies, import records from country X to port Y, and so on—rather than having one data store per person. Because the architecture is so similar to the citizen-centric personal data stores, it enables easier and safer sharing of data between citizens and government. For this reason, several states within the United States are beginning to test this architecture for both internal and external data analysis services.

Finally, it should not escape the reader's attention that all of these lessons also apply to companies with large, complex databases. Misbehavior by employees, industrial espionage, and cyberattack are among the greatest dangers that companies face in the big data era. A distributed architecture of databases joined with a network that supports permissions, provenance, and auditing can reduce risk and increase resilience of companies' internal data analysis functions.

SUMMARY

We are entering a big data world, where governance is far more driven by data than it has been in the past. Basic to the success of a data-driven society is the protection of personal privacy and freedom. Discussions at the World Economic Forum have made substantial contributions to altering the privacy and data ownership standards around the world in order to give individuals unprecedented control over data that are about them, while at the same time providing for increased transparency and engagement in both the public and private spheres.

We still face the challenge that large organizations, in particular governments and corporations, may be tempted to abuse the power of the data that they hold. To address this concern, we need to establish best practices that are in the interest of both large organizations and individuals. This chapter has suggested one path that can limit potential abuses of power while at the same time providing greater security for organizations that use big data. The key policy recommendations for all large organizations, commercial or government, are that:

1. Large data systems should store data in a distributed manner, separated by type (e.g., financial vs. health) and real-world categories (e.g., individual vs. corporate). These systems should be managed by a department whose function is focused on those data, with sharing permissions set and monitored by personnel from that department. Best practice would have the custodians of data be regional and use heterogeneous computer systems. With such safeguards in place, it is difficult to attack many different types of data at once, and it is more difficult to combine data types without authentic authorization.
2. Data sharing should always maintain provenance and permissions associated with data, and should support automatic, tamper-proof auditing. Best practice would share answers only to questions about the data (e.g., by using the pre-programmed structured query language, or SQL, queries known as "Database Views") rather than sharing the data themselves, whenever possible. This allows improved internal compliance and auditing and helps to minimize the risk of unauthorized information leakage by providing the minimum amount of information required.
3. Systems controlled by partner organizations, and not just one's own systems, should be secure. External data sharing should take place only between data systems that have similar local control, permissions, provenance, and auditing, and should include the use of standardized legal agreements such as those employed in trust networks, as described earlier. Without such safeguards, data can be siphoned off at either the data source or at the end consumer, without even attacking central system directly.
4. The need for a secure data ecosystem extends to the private data of individuals and the proprietary data of partner companies. As a consequence, best practice for data flows to and from individual citizens and businesses is to require them to have secure personal data stores and be enrolled in a trust network data sharing agreement.[6]
5. All entities should employ secure identity credentials at all times. Best practice is to base these credentials on biometric signatures.[7]
6. Create an "open" data commons that is available to partners under a lightweight legal agreement, such as the trust network agreements. Open data can generate great value by allowing third parties to improve services.

Although these recommendations might seem cumbersome at first glance, they are for the most part easily implemented with the standard protocols already found within modern computer databases and networks. In many cases, the use of distributed data stores and management are *already* part of current practice, and so the entire system will be simpler and cheaper to implement than a centralized solution: all that is really new is the careful use of provenance, permissions, and auditing within a legal or regulatory framework such as a trust network. Most importantly, these recommendations will result in a data ecosystem that is more secure and resilient, allowing us to safely reap the advantages of using big data to help set and monitor public policy.

NOTES

[1] See the D4D challenge, available at http://www.d4d.orange.com/home.

[2] For examples of trust networks, see Pentland 2009; World Economic Forum 2011; and the Institute for Data Driven Design, available at http://idcubed.org.

[3] For details about openPDS, see http://idcubed.org/open-platform/openpds-project/.

[4] For information about the Mobile Territorial Lab (MTL), see http://www.mobileterritoriallab.eu/.

[5] Sanger 2013.

[6] Pentland 2009; World Economic Forum 2011; http://idcubed.org.

[7] See http://openid.net/connect/.

REFERENCES

Barker, R. 1968. *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior.* Palo Alto, CA: Stanford University Press.

Dawber, T. 1980. *The Framingham Study: The Epidemiology of Atherosclerotic Disease.* Cambridge, MA: Harvard University Press.

ID3 (Institute for Data Driven Design, or idcubed). Available at http://idcubed.org.

MTL (Mobile Territorial Lab). Available at http://www.mobileterritoriallab.eu/.

OpenID Connect. Available at http://openid.net/connect/.

Orange. *D4D Challenge.* Available at http://www.d4d.orange.com/home.

Pentland, A. 2009. "Reality Mining of Mobile Communications: Toward a New Deal on Data." In *The Global Information Technology Report 2008–2009: Mobility in a Networked World.* Geneva: World Economic Forum. 75–80. Available at www.insead.edu/v1/gitr/wef/main/fullreport/files/Chap1/1.6.pdf.

———. 2014. *Social Physics: How Good Ideas Spread—The Lessons from a New Science.* New York: Penguin Press.

Sanger, D. E. 2013. "A Washington Riddle: What Is 'Top Secret '?" *The New York Times Sunday Review,* August 3. Available at http://www.nytimes.com/2013/08/04/sunday-review/a-washington-riddle-what-is-top-secret.html?_r=0.

World Economic Forum. 2011. *Personal Data: The Emergence of a New Asset Class.* Geneva: World Economic Forum. Available at http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.