Data, Consent, Privacy, and Insight

Daniel A. Reed, University of Iowa

From Toffler's iconic 1970 book, *Future Shock* and its meditations on "much change in too short a period of time" through Friedman's 2005 book, *The World is Flat,* and its commentary on globalization and its far reaching effects, to contemporary assessments of global information flows, much has been written about the accelerating pace of technical change and the associated socioeconomic consequences. Some assessments have been simplistic and motivated by specific social agendas; others have been nuanced and deep. All agree the changes are deep, profound, and substantive. Consider just a few, illustrative examples:

- **Urbanization** is creating new megacities while, even as the world's population grows, rural areas are being depopulated. United Nations predictions suggest that over 66 percent of the population will live in megacities by the year 2050 [1], up from 30 percent in 1950.
- **Stratification** is concentrating a larger fraction of the world's wealth in an increasingly small fraction of the population. Today, the top one percent control the overwhelming majority of global wealth, and socioeconomic mobility (i.e., the ability to change the economic stratum of one's birth) continues to decline.
- **Disintermediation** of existing supply and distribution models and chains is creating new businesses while eliminating others. From e-commerce (e.g., Amazon) through professional services (e.g., E-Trade and Zillow) to consumer services (e.g., Uber and AirBnB), all focus on direct consumer engagement.
- **Polarization** of social perspectives and political opinions is one consequence of such rapid shifts, as people respond to change with anger and fear, fed by a continuous stream of targeted news and information, based on data analytics and machine learning.

Against this backdrop of social issues, technological change continues apace, contributing to and accelerating the social change. As the 21st century industrial revolution, the digitization of our world surely ranks at the top of that technological change. Consider just a few examples:

- **Big data** arises from a combination of e-commerce transactions, the explosive growth of smartphones, the nascent, but rapidly growing Internet of Things (IoT), and the automation of government and business services.
- **Deep learning,** enabled by both big data and massive computing capabilities, is increasingly enabling computing systems to equal or exceed human capabilities on a wide range of speech and vision tasks, as well as routine and (often) non-routine cognitive tasks.
- **Automation** is a direct consequence of deep learning, with machines now managing many manufacturing tasks

and increasingly supplanting humans in areas such as medicine and finance.
- **Biomedicine** advances, triggered by inexpensive DNA sequencing and new insights, have created tailored treatments for heretofore untreatable illnesses, albeit sometimes at exorbitant costs. It has also brought a new world of consumer health sensors and the quantified self movement.
- **Environmental** change and global warming create severe weather, adversely affect agriculture, and may render some portions of the planet uninhabitable. Concurrently, smart sensors and data analytics have enabled precision agriculture and detailed environmental monitoring.

The importance of data in both enabling these technical changes and in potentially remediating the more pernicious effects of others cannot be overemphasized. With that backdrop, the remainder of this paper discusses the scale and scope of big data, the privacy and legal challenges created by digital data flows, and the emerging issues surrounding sensors and passive data. It concludes with some thoughts on a new model of digital privacy, one that combines bounded lifetimes, limited sharing transitivity, and claims-based access.

Data Gets Big

The phrase "big data" has been widely adopted to describe the explosive growth of digital data from a wide variety of sources. Big is, of course, a relative term, depending on both context and use. Just as New Orleans, LA is a "big city" if one lives in rural Iowa, New Orleans is a small city by comparison with Chicago or Beijing. More practically, "big data" means the data volume exceeds the utility and efficacy of the traditional tools used in the relevant context. For a small office, that might mean exceeding the expertise of local users with standard desktop tools. For an academic, government, or business, it might mean exceeding the capabilities of the organization's enterprise storage systems.

The growth of "big data" has been both driven by and aided by the rise of e-commerce, smartphones, and commercial cloud services. Anyone who has browsed the web, updated their social network, purchased an item online, or used a business service, has relied on a rich array of cloud services. In turn, these cloud services operate atop a network of massive data centers. Built by Amazon, Google, Microsoft, Facebook, and others, each data center exceeds the scale of the entire Internet just a few years ago. As such, each costs hundreds of millions of dollars to construct and contains tens of thousands of servers. Each of the major cloud vendors operates a worldwide network of such data centers, serving both consumers and enterprise customers.

Privacy, Ethics, and Law in the Digital Age

As data has become digital and migrated from local devices to the global cloud of data centers, law and policy have struggled to keep pace. Our legal notions of privacy and security are all rooted in the concepts of person and place. In the United States, these derive from the castle doctrine and English common law. We expect that our homes are legally secure, protected from search and seizure without due process. Likewise, we expect our physical selves to be also be legally inviolate.

In contrast to these physical concepts of person and place, where location defines the governing laws and policies, the data associated with our digital personas can be and often are geographically distributed across a worldwide set cloud data centers. Equally importantly, the decisions about where that data resides rest with the cloud service operators, not the consumers with whom that data is associated nor is it dependent on where the consumer may physically reside.

Not only can the data be in multiple jurisdictions, sometimes those jurisdictions can be in legal conflict. This can be true not only across state boundaries but also across international ones. As a hypothetical example, consider a Kenyan national whom a German company employs. As he or she travels from Nairobi to Berlin, then on to the United States and then China, each time he or she uses her smartphone, he or she leaves a global trail of digital data and an equally complex set of conflicting legal jurisdictions.

As this example illustrates, there deep, profound, and unresolved issues about global jurisdiction and legal applicability. For instance, can a country insist that a cloud service provider produce data regarding one of its citizens, even if that data is stored outside the borders of the country? What about citizens of other countries? Alternatively, must the entity seek legal access in the country where the data is stored? What rights do citizens have to protect their digital personas in each jurisdiction? How are conflicting legal expectations managed?

These are not merely hypothetical questions; they are issues currently being litigated. The U.S. Supreme Court recently agreed to hear arguments in a case involving Microsoft [2]. The U.S. sought access to data on a U.S. citizen who was a suspected drug dealer. Microsoft turned over data stored domestically, but it refused to supply the data stored in a data center in Dublin, Ireland, citing the potential precedent that would require it to turn over similar data to other countries.

The high court will also consider whether a warrant is required to access smartphone location data (i.e., the location history of the smartphone based on cellular tower connections). This case illustrates the power of not just data, but metadata. It is not what was discussed on the smartphone that is of value, but where the calls were made.

Although a judicial resolution of such issues would benefit all, the ultimate and appropriate disposition of such questions should be via an update to the *Stored Communications Act of 1986.* Thirty years of technological change have made many of its provisions no longer relevant. To put this in perspective, remember that the first popular web browsers did not appear until the 1990s, smartphones where unknown, and floppy disks were the common medium of data exchange.

Passive and Active Data

The use of email or a smartphone is an active event, requiring the user to engage in an explicit action (i.e., sending an email or making a call). The same is true when making an electronic purchase or using a social network service. In each case, the user has agreed to the terms of service as a condition of use. To be sure, these end user license agreements (EULAs) are often arcane and difficult for a layperson to understand, but they do enumerate the rights of use by both the provider and the consumer. Thus, one may well question whether consent was truly informed, but there is an explicit consent and any use requires explicit action by the user.

The rapid growth of wireless sensors and the Internet of Things (IoT) often removes the element of explicit consent, as data capture is an implicit artifact of other activities. Interconnected smart appliances, networked security systems and cameras, and smart cars all raise questions regarding appropriate and acceptable use of captured data. Who owns the data? Who controls it use?

Wearable health or exercise monitors make these questions intensely personal. Although a user may have accepted the terms of use when the device was purchased, no explicit action is required to generate data; the device captures data continuously during wear and often stores it in a cloud service. More perniciously, this data is highly personal and in a medical context would be protected by the Health Insurance Portability and Accountability Act (HIPAA).

In addition to personal health monitoring, intelligent home assistants can and do capture behavior and context from daily life. Amazon's Echo and Alexa assistant, Google Home and Google Assistant, and Apple's Home Pod and Siri have already raised privacy concerns, with cloud-based voice recognition systems listening for user commands. How the balance of privacy, ease of use, and consumer value will be resolved is yet determined.

Toward a New Model of Data Sharing

As noted earlier, our models of privacy are all rooted in concepts of person and place. Moreover, data sharing is largely a binary choice – to share or not to share. Even when more nuanced policies exist, managing the configuration details is often confusing and difficult.

As we adapt to a brave new world of cyberespionage, corporate and government data breaches, and global data flows, perhaps it is time to reconsider some of our approaches from first principles. These might include more nuanced notions of data ownership, privacy, and security, resting on three principles:

- **Bounded lifetimes.** Today, there are no constraints on how long digital data may persist across the Internet. Once released, it remains as long as any search engine or cloud service algorithm asserts that its retention may have economic value. Employment recruiters and government agencies both exploit this to conduct background checks. Instead, one might attach a lifetime to data at the time of its release, requiring the data's destruction at the end of that period.
- **Controlled Transitivity.** Similarly, data released today is most often available to all entities. Instead, consider a model where data released to an individual or organization has limited transitivity (i.e., it cannot be passed on to others without explicit consent). Thus, one might share a photograph with one person but that individual would be unable to share the photograph with anyone else.
- **Claims-based Access**. Finally, once data is released, there are few limitations on how that data is used. Claims-based access would specify the purpose for which data could be used. Hence, one might make data

available for personal, non-commercial use, but forbid any other uses.

Combining these three ideas creates a more nuanced model for data sharing. As an integrated example, one might share a digital document only with four team members for one week, allowing them to read it but not create copies or share it with others.

Concluding Thoughts

In the early days of the Internet, the *New Yorker* published a cartoon [3] that sparked an Internet meme: "On the Internet, nobody knows you are a dog." The notion of an uncharted frontier, where anonymity ruled, disappeared long ago. Today, ubiquitous sensors and consumer-connected devices; big data from e-commerce, social networks, intelligent assistants, and smart devices; and deep learning, mean the Internet not only knows you are a dog, your smart dog collar shares where you walk and your dog-food preferences are cross-referenced and matched with purchasable chew toys.

Within the broader context of social and technological change, we must ask wise and thoughtful questions about how this data is used and by whom. Only by concurrently considering social implications and technological capabilities can be create sustainable approaches.

References

1. United Nations, *World Urbanization Prospects*, https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf, 2014
2. United States v. Microsoft, http://www.scotusblog.com/case-files/cases/united-states-v-microsoft-corp, 2017
3. Wikipedia, "On the Internet, Nobody Knows You're a Dog," https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog, 2017