CHAPTER 1
Big Data, Big Impact

A people that values its privileges above its principles soon loses both.

—Dwight D. Eisenhower

I had chosen to use my work as a reflection of my values.

—Sidney Poitier

Target knows. Apple Computer knows, too. So do LinkedIn, Netflix, Facebook, Twitter, Expedia, national and local political campaigns, and dozens of other organizations that all generate enormous economic, social, and political value. They know that that the age of Big Data is here and it's here to stay. The swelling ranks of organizations that increasingly depend on big-data technologies include dozens of familiar names and a growing number you've never heard of.

On February 16, 2012, the *New York Times* published an article about Target's ability to identify when a customer is pregnant. Target declined to comment or participate in the story, but it was written and published anyway. The onslaught of commentary and subsequent news raised numerous questions ranging from the legality of Target's actions to the broader public concern about private, personal information being made more public.

On April 20, 2011, two security researchers announced that iPhones were regularly recording the position of each device to a hidden file. While Apple readily acknowledged that the claim was true, the resulting hubbub made clear that it was the method by which that file was generated and stored that caused security concerns. The decision to use that technological method had clear and direct ethical consequences in the real world.

Who was involved in making that decision? A lone engineer in a back room making the technology perform in the way that made the most sense? Was there a broader business discussion of whether that function should be available at all? To what level of detail were the security and other risks discussed?

In August of 2011, Facebook faced criticism when it was thought to be exposing the names and phone numbers of everyone in the contacts on mobile devices that used the "Contacts" feature of the Facebook mobile application. It responded and clarified how the feature worked and provided people with a method to remove that information from their Facebook account. Why wasn't that clarification and method provided in conjunction with releasing the feature in the first place?

In 2011, when the CEO of GoDaddy published a tweet about killing elephants in Africa and publicly supported the controversial Stop Online Piracy Act (SOPA), the negative customer response resulted in the domain registrar reportedly losing tens of thousands of customers. The Kenneth Cole brand was damaged when they were perceived to be using the political uprising in Cairo in the spring of 2011 to promote their products. Apologies and a damaged brand reputation followed. In 2010, Wal-Mart was alleged to be using a fake online community to build support for new stores in areas where the idea was not popular. One of the public relations firms was allegedly responsible.

As you are likely considering how your organization would respond in similar situations, consider the fact that all these examples share one common factor: big-data technology. As these examples show, one impact of big data is that actions have far greater consequences, at a more accelerated pace, and direct repercussions for a company's brand quality, customer relationships, and revenue. As a result, big data is *forcing* new considerations about our values and behavioral actions—especially as it gives more people more ways to engage, communicate, and interact. One outcome of this growing presence of big-data technology is that business operations are changing and increasing the sheer amount of information they generate so fast that the big data phenomenon is starting to raise ethical questions.

As Brad Peters recently wrote in *Forbes*, it literally "changes the social contract" (*http://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/*). The nature of that change is complex. One primary motivation for this work is to address both individuals and organizations and suggest that more explicit and transparent discussion is needed —a discussion that inherently contains ethical components.

And although those ethical topics are centered on individual people, the implications span a variety of areas. In the same way that big data raises personal privacy concerns, it generates new questions about personal identity, notably who owns our personal data and how the increased presence and availability of more data influence our reputations.

For both individuals and organizations, four common elements define what can be considered a framework for big data ethics:

Identity
: What is the relationship between our offline identity and our online identity?

Privacy
: Who should control access to data?

Ownership
: Who owns data, can rights to it be transferred, and what are the obligations of people who generate and use that data?

Reputation
: How can we determine what data is trustworthy? Whether about ourselves, others, or anything else, big data exponentially increases the amount of information and ways we can interact with it. This phenomenon increases the complexity of managing how we are perceived and judged.

Both individuals and organizations have legitimate interests in understanding how data is being handled. Regardless of your role in an organization, or if you even work in technology, nearly everyone's life is touched by big-data technology today. Which means this framework has the potential to inform both the benefits big data provides and the potential risks from unintended consequences for a truly staggering number of people.

As an example, New York Judge Gary Brown recently found that an IP address is not sufficient evidence to identify copyright infringers (*http://torrentfreak.com/judge-an-ip-address-doesnt-identify-a-person-120503/*). Although this legal finding was focused on copyright issues, it could have far-reaching implications for questions about all four elements of big-data ethics. If a person is not an IP address (and who, really, ever thought they were identical?), then can any data generated via a specific IP address be legitimately associated with a single, unique individual?

Digital marketers have struggled with this for years. But the risk of unintended consequences as big data evolves becomes more widespread—well beyond targeted marketing. Consider how Google filters its understanding of your content preferences if you share equal time on the same computer with one or more people in your household. My interest in beach vacation spots is much less relevant to someone with whom I might share my Internet connection who is afraid of the ocean and can't swim. Improving the relevancy of targeted marketing is a major challenge, but the challenges and potential risks don't end with online advertising.

A realistic scenario illustrates some of the challenges people and organizations face. Imagine that an elderly relative's glucose and heart monitoring device shares the same IP address as the rest of your household. As a matter of course, all data from those medical devices is captured and stored by a healthcare provider. Now imagine that through an internal data leak, the hospital inadvertently mixes up their medical condition with your own. After all, you both live at the same address, could be the same gender, and might have the same last name.

This is not an economic risk, although it's easy to imagine bills for healthcare services being assigned to the wrong person as a result of the mix-up. But the legal decoupling of an IP address from a specific, individual person points to the presence of risks that exist right now, with technology that is already in widespread usage. The risk is that although there is value and benefit to healthcare innovations using technology, the real-world relationship between the Internet technologies used and the people who benefit from them is not sufficiently understood.

"Spoofing" (pretending to be someone you're not) has a long and storied history—both on and off the Internet. But in this scenario, the unintentional confusion between a relative's medical condition and your own, which is based on the assumption that a single person generates data originating via a single IP address, could have disastrous consequences if you're ever rushed to the emergency room.

Judge Brown's legal decision encourages a must-needed exploration of the nuances of privacy, identity, reputation, and data ownership. The direct impact of failing to understand the complexities and nuance of the relationships between big-data technologies and the people who use them can, in this example, literally be a matter of life and death.

Why Big Data?

At this point you might be asking, "Why not just *any* data?" After all, many organizations have been struggling to figure out how to manage their data for some time now, right? Common definitions of the popular phrase for the phenomenon "big data" are based on distinctions between the capabilities of legacy database technologies and new data storage and processing techniques and tools such as Hadoop clusters, Bloom filters, and R data analysis tools. Big data is data too big to be handled and analyzed by traditional database protocols such as SQL (which makes *big data* a term that may evolve over time; what is now big data may quite rapidly become small). In this sense, size is just one aspect of these new technologies. The risks and ethical considerations also come from a few related factors.

The *volume*, *variety*, and *velocity* of available information exponentially increase the complexity of information that companies need to manage, and these factors generate questions they haven't previously encountered in the course of doing business.

The volume at which new data is being generated is staggering. We live in an age when the amount of data we expect to be generated in the world is measured in exabytes and zettabytes. By 2025, the forecast is that the Internet will exceed the brain capacity of everyone living on the entire planet.

Additionally, the variety of sources and data types being generated expands as fast as new technology can be created. Performance metrics from in-car monitors, manufacturing floor yield measurements, all manner of healthcare devices, and the growing number of Smart Grid energy appliances all generate data.

More importantly, they generate data at a rapid pace. The velocity of data generation, acquisition, processing, and output increases exponentially as the number of sources and increasingly wider variety of formats grows over time. It is widely reported that some 90% of the world's data has been created in the last two years (*http://www.economist.com/node/21537967*). The big data revolution has driven massive changes in the ability to process complex events, capture online transactional data, develop products and services for mobile computing, and process many large data events in near real time.

In the last few years of working with organizations who use big data technologies, it became clear to us that there were divided opinions on just what were the ethical issues and constraints in a dizzying variety of big-data situations. Without a formal and explicit framework for having ethical discussions in business environments, people often revert to their own moral code. Which, although it's a great place to start, can quickly devolve into a "But, that's creepy…"/"No, it's not" debate that goes nowhere fast. What frequently happens in those cases is that the discussion becomes mired by frustration, the meeting ends, and the question doesn't get answered. The potential for harm due to unintended consequences can quickly outweigh the value the big-data innovation is intended to provide.

So, while business innovators are excited about the potential benefits they can create from the design and development of a wide range of new products and services based on big-data technologies, the size, variety, and velocity of information available raises new questions. Some of those questions are about the implications of the acquisition, storage, and use of large quantities of data about people's attributes, behavior, preferences, relationships, and locations.

Fundamentally, these questions are *ethical*. They relate to your values and how we apply them while creating products and services. And your values are at the heart of how you balance the promise of useful innovation against the risk of harm. Whether you are aware of them or not, your values inform how you conceive of and execute on designs for products and services based largely on information gleaned from massive amounts of data. They are critical inputs to the calculus you perform when weighing the promise of those benefits against the risks of unintended consequences.

This implies that there is a balance to be achieved between those risks and the benefits of the innovations that big data can provide. This book is intended, in part, to help organizations develop a framework for having explicit ethical discussions to help maintain that balance.

What Is Big Data Forcing?

Society, government, and the legal system have not yet adapted to the coming age of big-data impacts such as transparency, correlation, and aggregation. New legislation is being drafted, debated, and ratified by governments all over the world at a rapid pace.

Only a generation or two ago, one could fairly easily drop "off the grid" and disappear within the continental United States. Today, it would be nearly impossible for a person to do much of anything without generating a data trail that a reasonably knowledgeable and modestly equipped investigator could follow to its end (*http://www.wired.com/vanish/2009/11/ff_vanish2/*).

Big data is persistent. And it is persistent in a way that business and society have never experienced before. The Library of Congress is archiving all tweets since 2006. And when the Library of Congress archives something, they intend for it to *stay* archived. Facebook has tacitly acknowledged that deleting your account does not delete all the data associated with your account (*http://arstechnica.com/gadgets/2012/05/on-facebook-deleting-an-app-doesnt-delete-your-data-from-their-system/*).

Eric Freeman and David Gelernter coined the phrase "lifestream" to describe:[1]

> "…a time-ordered stream of documents that functions as a diary of your electronic life; every document you create and every document other people send you is stored in your lifestream. The tail of your stream contains documents from the past (starting with your electronic birth certificate). Moving away from the tail and toward the present, your stream contains more recent documents—papers in progress or new electronic mail; other documents (pictures, correspondence, bills, movies, voice mail, software) are stored in between. Moving beyond the present and into the future, the stream contains documents you will need: reminders, calendar items, to-do lists."

Freeman and Gelernter intended lifestream to inform software architectures and structures for managing personal electronic information, but the concept is useful in understanding how the persistence of big data influences critical, essential characteristics of individual lives. Big data often includes "metadata," which can add another layer (or several layers) of information about each of us as individuals onto the physical facts of our existence. For example, the architecture and technology of big data allows the location of where you physically were when you made a tweet to be associated with each message.

And those additional layers are explicit. They can contain a vast array of ancillary information only tangentially related to the essence of any given financial or social transaction. Big data can reconstruct your entire travel history anywhere on the planet. It supplies the information necessary to tie together intentionally disparate facets of your personality in ways we sometimes cannot fully control. Pictures of you on spring break are presumably not intended to be considered as relevant material when applying for a job, and big data has significantly changed how reputation is managed in such situations.

This data trail is just one example of how big-data technologies allow broader and deeper insight into human behavior and activity than ever before. Innovators of all types have

1. *http://cs-www.cs.yale.edu/homes/freeman/lifestreams.html*

realized the potential for turning those insights into new and valuable products and services. This wealth of data promises to improve marketing, management, education, research and development, healthcare, government, services, and a host of other aspects of our lives. Big data is already being used to improve insights into effective education policies and to improve our ability to predict dangerous weather conditions in microclimate-sized geographies.

But the forcing function big data creates raises questions about data handling with a new urgency. These challenges are potentially troubling because they often extend beyond the management controls of a single organization. Big-data technologies influence the very meaning of important concepts such as privacy, reputation, ownership, and identity for both individuals and corporations. As information is aggregated and correlated by not only the originating entity, but also by those who may seek to further innovate products and services using the original information, we frequently don't (or can't, even) control how that information is used once it is out of our hands.

Big data also allows us to congregate in online communities whose populations sometimes exceed those of entire countries. Facebook is the most well known example, but there are literally thousands of online communities across the Internet, each of which contains specific, unique snippets or facets of information about each of its members. We are just now realizing the impact of this phenomenon on our identities, the concept of ownership, how we view ourselves and our relationships, trust, reputation, and a host of other, more traditionally self-managed aspects of our lives.

Because the data is frequently data about people and their characteristics and behavior, the potential use and abuse of this acquired data extends in a great many directions. Direct benefits are now being realized, but concerns about the consequences of having personal data captured, aggregated, sold, mined, re-sold, and linked to other data (correlated) are just now beginning to see the light of day.

And these risks are not just limited to individual people. They apply equally, if not more, to organizations. Corporations are not in the business of harming their customers. Hospitals are not in the business of violating their patients' confidentiality. Nonprofit research facilities are not in the business of sharing their test subjects' personally identifiable information. Yet, through the normal course of everyday business operations, which increasingly utilize big-data technologies, the risk of various harms increases.

And the type, size, and impact of those risks are difficult to determine in advance. We have, as a society, only just begun to understand the implications of the age of big data.

Consider the following:

- The social and economic impact of setting insurance rates based on browser or location history, e.g., visits to sites with information about chest pain or a detailed record of your vehicle's GPS history (*http://www.wired.com/threatlevel/2011/09/onstar-tracks-you/*).

OnStar quickly reversed its decision in response to privacy concerns. See *http://www.computerworld.com/s/article/9220337/OnStar_reverses_course_on_controversial_GPS_tracking_plans*.

- The use of genetic information to influence hiring practices.
- "Predicting" criminal behavior through extrapolation from location, social network, and browsing data. *Minority Report*–style "predictive policing" is already in place in some major urban areas (see *http://www.cbsnews.com/8301-18563_162-57412725/lapd-computer-program-prevents-crime-by-predicting-it/*).
- Retrieval of metadata about a person based on a picture snapped with a mobile phone in a "dating" app that gave access to criminal records, browsing history, or a site of dating reviews of individual people.

At risk are the very benefits of big data innovation itself. In late 2011 and early 2012, the Stop Online Piracy Act (SOPA) put before Congress was met with fierce resistance from a wide variety of industries, organizations, and individuals. The primary reason was the belief that the provisions of the proposed law would severely constrain innovation in the future using technical tools such as big data (*http://en.wikipedia.org/wiki/Stop_Online_Piracy_Act*).

Part of the debate centered around the belief that the members of Congress supporting the bill were either misinformed by interested parties about how the technology worked and how innovation was made possible, or they were just simply unaware of the realities of how Internet and big data technologies worked in the first place. In either case, SOPA represents a classic example of how a lack of transparent and explicit discourse about how a critical piece of our economy and society works had the potential to significantly limit our collective ability to benefit from those tools.

As big data's forcing function drives data further into our organizations and individual lives, balancing risk and innovation will continue to be an urgent need that must be met in order to maintain the ability of big data to generate benefit rather than harm.

Big Data Is Ethically Neutral

While big-data technology offers the ability to connect information and innovate new products and services for both profit and the greater social good, it is, like all technology, ethically neutral. That means it does not come with a built-in perspective on what is right or wrong or what is good or bad in using it. Big-data technology has no value framework. Individuals and corporations, however, do have value systems, and it is only by asking and seeking answers to ethical questions that we can ensure big data is used in a way that aligns with those values.

Such discussions require explicitly exploring those values and developing ethical perspectives, which can be difficult. Ethics is a highly personal topic and comes loaded with lots of polarizing vocabulary, such as *good*, *bad*, *right*, and *wrong*. We all have personal moral codes, which naturally vary from individual to individual. The lack of a common vocabulary for expressing the relationship between what we personally believe in and what we, as members of a common enterprise, plan to do with big data can create constraints on productive discussion and obstacles to finding consensus.

That said, this isn't a book about dictating operational policies or changes to case or statute law. Business executives, managers, judges, and elected officials must see to that. This also isn't a book about business ethics—at least as traditionally conceived. Business is concerned primarily with profit and innovation. Ethical inquiries, as a formal practice, are of interest only as far as they impact profitable operations and the ongoing development of products and services that meet the needs of a dynamic market.

There is, however, an inherently social component to business, and in fact, big data and social media have only exaggerated this reality in recent years. The mere act of conducting commerce, exchanging goods and services for items of value (often in the form of currency), is an activity that typically involves people. And people have values. The purpose of this book is to build a framework for facilitating ethical discussions in business environments designed to expose those values and help organizations take actions that align with them.

The big-data forcing function is bringing business functions and individual values into greater contact with each other. Big data is pushing corporate action further and more fully into individual lives through the sheer volume, variety, and velocity of the data being generated. Big-data product design, development, sales, and management actions expand their influence and impact over individuals' lives in ways that may be changing the common meaning of words like *privacy*, *reputation*, *ownership*, and *identity*.

Its sheer size and omnipresence is essentially forcing new questions into play about our identities, the evolution of personal privacy, what it means to own data, and how our online data trails influence our reputations—both on- and offline. Organizations from business to education and from research to manufacturing and professional services have tremendous amounts of information available about their customers, their operations, and nearly every other measurable aspect of their existence. Before the rapid growth of big-data technology in the last five years, changes in organizational processes or policies had a delayed effect on customer's lives, if any. Whether a customer's personal data was accessible or not was typically a matter of how many individuals or organizations had access to customer records.

Big data operates at such a scale and pace now that such changes in policies and practices extend further and faster and touch more people. Thus, changes in business functions have a much greater impact on people's lives. The expansion of traditional operations

touches our lives every day in ways we can hardly keep track of, let alone manage. The reality is that the ways in which legislation, social norms, economics, or reasonable expectations of normal interaction will change as a result of the growing presence of big data is simply unknown.

And it is precisely because these things are unknown that ethical dialog should be encouraged. Open and explicit dialog about aligning values with actions to balance the risks with the benefits of big-data innovations is one method you can use to ensure that you negotiate the trade-off well—and in your favor. Identifying those moments when decisions turn into actions, or *ethical decision points*, is the first step to developing a capacity to have those discussions both "in-the-room" on the fly and more formally in the development of transparent perspectives and policies.

Don't Tell Me What to Do

It is also not the aim of this book to be prescriptive, in the sense of laying down some hard-and-fast list of rules for the ethical handling of data. Indeed, these issues are often too specialized to a given business model, sector, or industry to allow for that. The aim, rather, is to illustrate the benefits of directly addressing these questions, to discuss key factors that go into developing a coherent and consistent approach for ethical inquiry, and to set out a framework for and encourage discussion. This discussion can take place not just in boardrooms, executive meetings, courtrooms, and legislatures, but also in working meetings, hallways, and lunchrooms—a discussion that is explicit, collaborative, and transparent.

The goal of addressing these questions directly through explicit and transparent dialog is to better understand and mitigate risks to relationships with customers and partners, and to better express the benefits of big-data innovations. Unfavorable perceptions and bad press affect the bottom line. Even the *perception* of unethical data handling creates a risk of negative consequences, diminishing internal support for business goals and external relationships with customers. This is not merely a question of transparency or good management; it is a broader ethical question about maintaining the consistent alignment of actions and values as big data evolves and becomes even more embedded and influential in people's lives.

In short, this book won't tell you what to do with your data. The intent is to help you engage in productive ethical discussions raised by today's big-data-driven enterprises, propose a framework for thinking and talking about these issues, and introduce a methodology for aligning actions with values within an organization. That framework will provide a set of tools that any enterprise can adopt to become an organization in which customers, partners, and other stakeholders can trust to act in accordance with explicit values coherently and consistently.

Important Concepts and Terms

Identifying ethical decision points helps to develop perspectives and policies that drive values alignment in business operations, products, and services involving personal data. To do that, you have to know what values you have and where they might not be aligned. And this can be a complex activity with a specialized vocabulary. The following are some useful terms in developing that vocabulary:

Rights and interests

It is common for people to speak of privacy *rights*, but talk of rights brings with it the suggestion that such rights are absolute, which presumes to prejudge some of the issues at hand. In order to avoid prejudgment, we will speak of privacy *interests* and other sorts of interests, with the explicit understanding that a right is a kind of interest, the strongest and most absolute kind.

For example, an absolute privacy right with respect to the usage of your medical data includes the right to stipulate that no privacy risk at all is to be taken with this data. But suppose that you are brought unconscious to the emergency room and treated—with data being generated in the process. This data might be useful in the development of better treatments for you and others in your situation. Do we really want to hold that the use or sharing of this data without your consent is absolutely forbidden? Even with the next nurse or doctor on staff? Perhaps we do want to hold that there is such a right, but to think that there is one should be an outcome, not a presupposition of the sort of discussion that we advocate.

This is all complicated by the fact that to have such a right is itself an ethical view. Supporting an absolute right inherently contains an ethical position and diminishes an ability to be objective about whether or not that position aligns with our values. Thinking in terms of privacy interests (as opposed to rights) allows for more objective freedom in assessing the strength of ethical claims.

Personal data

The commonly assumed distinction between *personally identifying information* and other data is largely an artifact of technological limitations that often can be overcome. In order to move forward, we need a very broad term for the sort of data that is at issue when people are concerned about privacy. In usage here, *personal data* will simply be any data generated in the course of a person's activities.

A responsible organization

The difference between doing right and doing what various people *think* is right is a significant one for the present topic. A *responsible organization* is an organization that is concerned both with handling data in a way that aligns with its values and with being perceived by others to handle data in such a manner. Balancing these two nonequivalent concerns is something a responsible organization must work to achieve.

So, big data is big, fast, and can contain a wide variety of information. It's here to stay, and it offers huge promise of economic gain, social benefit, and cultural evolution. And it's forcing ethical questions into places and environments where previously they haven't been critical to answer. How are people and organizations supposed to respond? This book advocates learning how to engage in explicit, transparent, and productive ethical inquiry.

The next chapters discuss how that kind of ethical inquiry can help align your values with your actions to both enhance innovation and to reduce risks. The discussion begins with a demonstration that ethical misalignment is present in even the most successful and well-run organizations, and then offers a vocabulary and a framework for engaging in the ethical inquiry needed to gain better alignment.