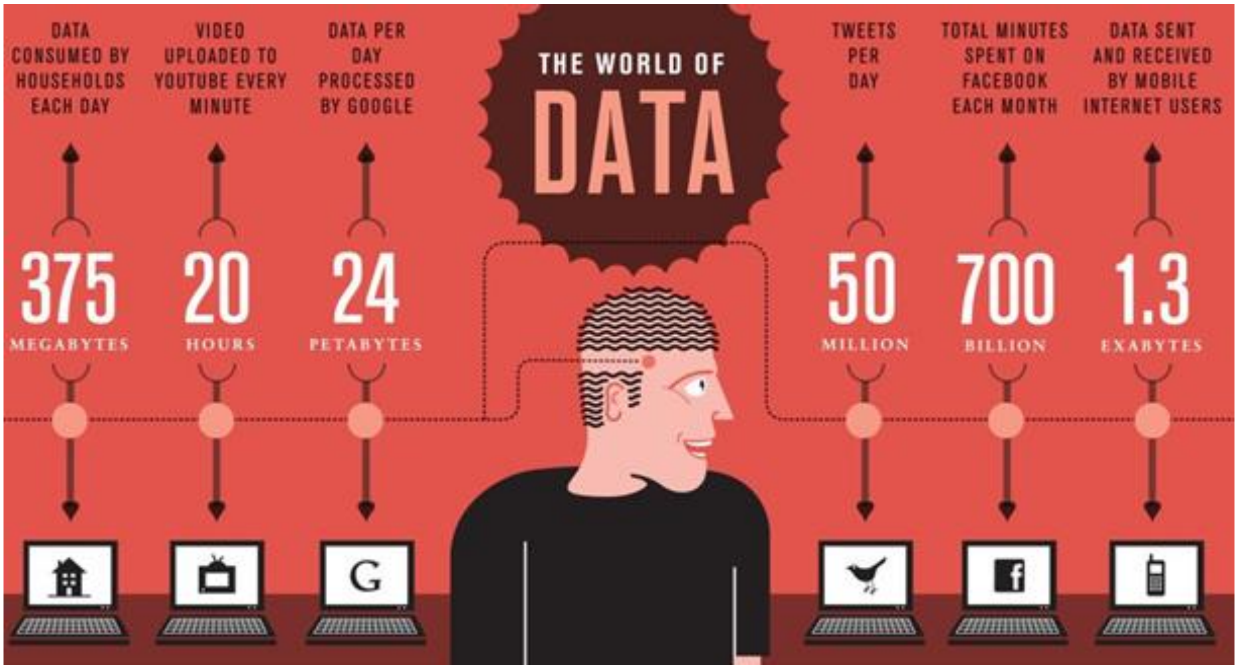
Chapter 20 – Ethics of Big Data

Matthew Rothmeyer

Summary

When considering the advance of technology and the prevalent and pervasive nature of electronic data, many questions of both an ethical and practical nature arise. While many of these relate specifically to individuals (What information should one share and how does one protect that information?) many are more applicable to the corporations and entities with the capital and knowledge to make use of this information on a large scale. These are questions relating to the ownership of such data, the responsibility of protecting data, and obligations an organization might have to both the owners of the data and the interests of those invested in said organization. All of these questions, and many others, can be captured under the concept of *The Ethics of Big Data*. This chapter will explore this important domain, providing an introduction and examination of some of the most pressing questions, as well as examples of what considerations one must make to remain ethically sound when using Big Data.

Introduction – What is Big Data?

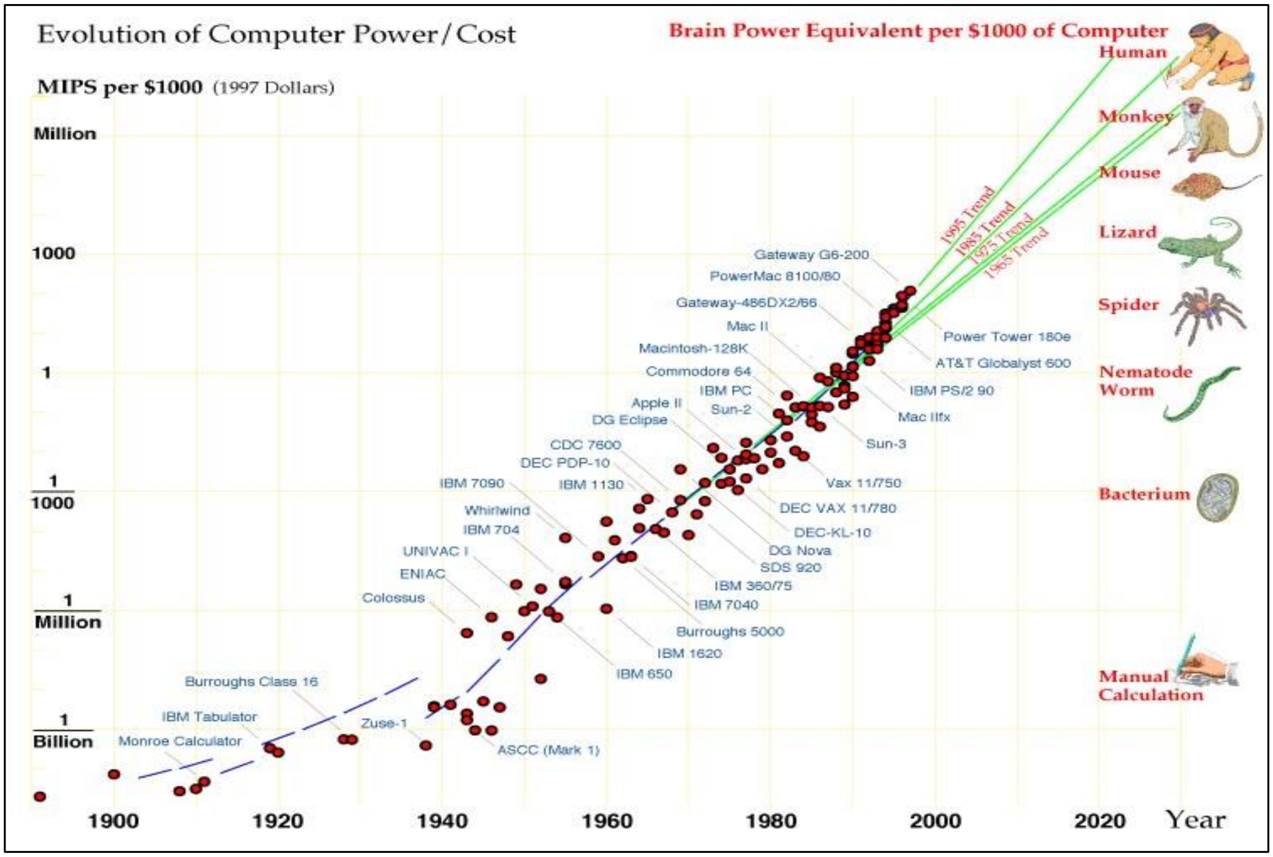

DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
THE WORLD OF DATA
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
G

1 Illustrating the growth of data

In order to understand the ethics of Big Data and why such ethics are meaningful, it is useful to have some grasp on what Big Data actually is. As such several definitions or ways to consider big data are presented below.

Big Data is often a catch all term referring to an incredibly expansive data set (or collection of data) that is beyond the technological capabilities of traditional data management software. In practice this usually equates to the need for special tools that aid in the process of capturing, searching, analyzing, and visualizing this data to be used. Big data is often encountered in practice in many scientific fields such as meteorology (complex weather patterns), astrophysics (cosmological interaction), and biological simulations and computations (genome sequencing). In fact, any field of research that involves processing many different input variables or “data points” could be considered to use Big Data. In such cases the amount of useful data captured is often restricted by storage capacity and processing power (which, as illustrated in part by figure 1 and 2, are rapidly increasing) as opposed to the actual presence of available, measurable data.

Big data is not limited to use in scientific studies however, but also finds use in business applications such as advertising, finance, internet search, and business administration. In these areas data sets that were once small have been expanded due to improving technology and the many avenues that technology creates for measuring and processing data (remote sensing, logs, wireless networks, grid computing, etc.). In fact, according to Hilbert and Lopez, the ability to store information has doubled approximately every 40 months and the ability to compute information has doubled every 14 months over the past few decades [1]. As a result of this development, many organizations with significant capital are able to acquire technology that enables processing and deriving conclusions from data where such capabilities were previously impossible to instrument.



2: Growth of computing power

Because what constitutes as big data is often considered to be pinned on the technology used to process and store it, said technology can also provide another way to define this domain. When making such considerations Big Data can be thought as the situation arising from the vastly increased speed and quality of gathering and analyzing personal data based on the growth of computing power [2]. Under this definition what might be referred to as big data today could quickly be overshadowed by even larger quantities of fine grain data in the future. Consider for example that, at one point in time, the ability to search national phone and mail directories electronically may have constituted as Big Data. Today such a thing might seem trivial while, at the same time, the idea that one might track the locations and habits of individuals over their lifespan does not seem that farfetched.

A final consideration of Big Data, one that unlike technology rarely changes, is the source and use of the data collected. This is often considered the most important attribute and the source of much of the ethical quandaries relating to Big Data. This is because data sets of the described magnitude can often be combined in ways that provide information not germane to the initial measurements. To put it another way, big data (especially that used for business purposes) is often composed of sets that can raise privacy concerns when used to draw certain conclusions. As such, one might conclude that another appropriate definition of big data is simply “data big enough to raise practical rather than merely theoretical concerns about the effectiveness of anonymization” [3]

Why is Big Data Important?

At this point one might question why big data is important to an individual. Why should you, as the reader of this book, spend time considering Big Data and its ethical nuances? These kinds of questions, while common, often point to a lack of understanding in how big data is used.

Davis and Patterson, in their work, The Ethics of Big Data, discuss several reasons why Big Data is so important when compared to normal data, and why businesses and professionals alike need to be prepared. When considering everything these boil down to what they call the “volume, variety, and velocity of the data.” [4]

3: A visualization of an electronic footprint

The volume of the data, or the amount of data both being generated and recorded, is massive and is continuing to grow. As the ability to generate data through technology becomes increasingly cheap the number of devices generating data will grow exponentially, thus filling the increasing capacity for data processing.

The types of items reporting are also tied to the variety of the data, or specifically the “variety and sources of data types” [4] that are coming into being at such a rapid pace. The future will be, for better or for worse, a world of smart, location aware objects existing in an “Internet of Things.” Almost every item a person can acquire can and will become at the very least a constantly updating data point in a massive database, and at the most, a database in its own right, communicating with other entities to share data and draw conclusions. A refrigerator will record its contents, a vacuum cleaner will note the amount of captured dirt, and typical cleaning times, and your car will be able to note which gas stations you frequent most often and what radio stations are your favorite.

The velocity of data, or the rate that it can be output, is also increasing exponentially. Several sources report that a vast majority of the world’s data has been generated in the past several years [5], as the ability to actually use this massive amount of data has grown. This increase in capacity has allowed the process of tracking, cataloging, and categorizing information about an individual to become relatively simple with the right resources, as opposed to in the past when such a thing was neigh impossible for anyone outside the largest corporations and medium to large governments.

This information often finds its beginnings, in many cases, as the result of interactions between an individual and some sort of electronic service. These interactions, more often than not, leave remains. These remnants might be inputs into a web form, an email address entered when signing up for an online account, or a list of past purchases at your favorite grocery store. Even communication that does not exist in electronic form often has a record of that kind associated with it (bills, bank statements, tax forms). These artifacts, when combined, are often referred to as an electronic footprint. A clearer definition is simply data that exists as the result of some interaction with an electronic system, either direct or indirect. This electronic footprint comprises a large portion of what many people consider to be Big Data, partly because it permeates the day to day lives of an individual and also because it is persistent, often lasting far longer than many people would guess (sometimes forever). Many times this information is also not of a mundane nature, instead of web forms or email addresses the data might be composed of travel records or monetary transactions, pictures from vacations and social events.

The unfortunate reality is that most individuals are at best only partially aware of this information and are at worst completely unaware. In many cases this information is separate and disparate, belonging to different companies and existing in many incompatible formats. In some instances however groups either exchange or control a significant portion of this information and have the tools or capabilities to compare it. In this scenario, an electronic footprint can often be used to gather and infer information that was not present in the initial data set. In some cases these inferences can be entirely harmless and expected. However, when taken to the logical extreme, these capabilities can be used to compile a history not only of one’s actions but also of their personality traits and habits. They allow an organization to, in a real sense, map an individual in a very personal way.

These uses are important because the affects they have are not limited to just one company or group. Take for example, an organization interviewing a prospective employee. It could be possible that the aforementioned company simply looks at the resume, schedules an interview, and makes decisions based off of human interaction and qualifications. It could also be possible that said

ARRIVALS
BIG DATA
TG'11
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

4 A hypothetical use or abuse of big data

company makes use of Big Data to determine that the interviewee has some undesirable genetic traits, is somewhat of an introvert, and once, while at university, made some poor decisions over spring break. It is possible that this data could end up costing the interviewee a job, in some cases before he or she had an opportunity to defend themselves. Now consider the scenario in which such a thing becomes popular among hiring organizations and after a time, might become commonplace in society. This would have far reaching consequences and would affect every organization from a small business to a government. Instead of just one group, the thoughts and feelings of everyone have been changed, in the opinion of many, for the worse.

This environment of changing opinions and social norms is one that has been, in part, forced by Big Data and the significant changes it has introduced to the capabilities of large organizations. On one hand Big Data promises to improve many aspects of our lives ranging from predicting dangerous storms to improving consumer shopping experiences, yet at the same time Big Data is changing important concepts such as privacy, and personal or organizational reputation in subtle ways that are often difficult to predict. These changes can be dangerous as there is almost no precedent for what is appropriate and a universal set of guidelines to what is ethical when concerning Big Data has yet to be written. As a result there have been many cases where Big Data has been used in ways that would be considered questionable or would not be considered to lie within the ethical boundaries of an entity. A few examples of these instances are listed below:

Target target's Pregnant Mothers

In early 2002 Target approached one of its newly hired statisticians, Andrew Pole, about a new application for big data, pregnancy prediction. Target, as well as many other large retailers, thrives off of determining the spending habits of their customers and providing them with products that fit their needs and desires, even if they weren't fully aware of those needs and desires. There was however one problem,

> *"Most shoppers don't buy everything they need at one store. Instead, they buy groceries at the grocery store and toys at the toy store, and they visit Target only when they need certain items they associate with Target — cleaning supplies, say, or new socks or a six-month supply of toilet paper." [6]*

Target sells a wide variety of items and would wish to appear to their customers as the one and only stop they need to make when purchasing goods for their home. In most cases however one's shopping habits are determined by brand loyalty or some other such concept. Targets analysts noticed that in most cases these habits rarely changed except for a few very specific circumstances.

> *"One of those moments —* the *moment, really — is right around the birth of a child, when parents are exhausted and overwhelmed and their shopping patterns and brand loyalties are up for grabs. But as Target's marketers explained to Pole, timing is everything. Because birth records are usually public, the moment a couple have a new baby, they are almost instantaneously barraged with offers and incentives and advertisements from all sorts of companies." [6]*

After some research what Pole and many other mathematicians discovered was that, given enough data about a particular topic (be it purchasing habits, spending frequency, the days a person gets groceries), and the ability to process that data, one can determine almost anything about an individual. What was unique about this discovery was that, more often than not, the “particular topic” said data was centered around often did not need to be related to the object of investigation. Pole eventually found that certain purchases, large quantities of lotion, vitamin supplements, hand sanitizers, and scent free soaps were almost always associated with an upcoming due date. Not only was this data correct, but in many cases it was very accurate. Pole and others at Target were able to assign what they called a “pregnancy prediction score” to shoppers that showed how likely a shopper was to be pregnant. In some cases Pole was even able to pin the actual due date of a pregnancy to a small window of time. There was even a case in Minneapolis where a father angrily complained about his daughter receiving coupons for baby items only to later apologize after questioning his daughter and finding out that she was, in fact, pregnant and that he had not been aware.

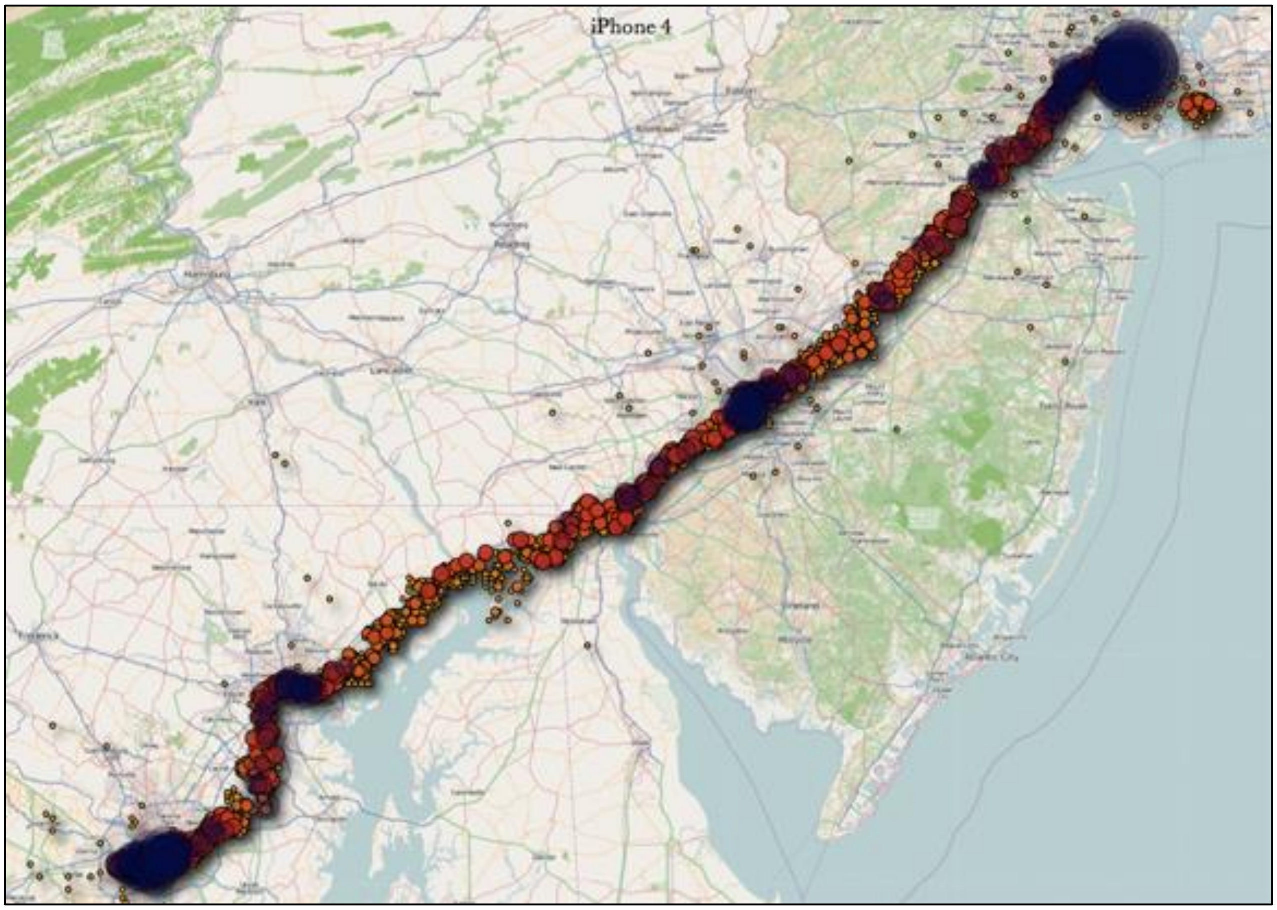
This case is important when examining the problems caused by big data for two reasons. First, few if any of the customers had explicitly given Target (or in some cases anyone) information about their pregnancy and yet Target was able to determine this fact with a high degree of accuracy. This is important because it is an excellent example of how a large amount of seemingly unrelated data points can be used to determine very specific pieces of information about an individual, information that in many cases would be considered private. Second, an organization using information in this way can end up damaging its own reputation. In this case Target ended up being the center of an “onslaught of commentary and subsequent news” which “raised numerous questions ranging from the legality of Target’s actions to the broader public concern about private, personal information being made more public.” [4]

Apple Records More Than Music

In April of 2011 security researches Alasdair Allan and Pete Warden announced at the Where 2.0 conference that several apple products, specifically the iphone and 3g iPad, had been recording an individual’s location data to a secret and hidden file [13]. Not only was this data being recorded but it was specifically being preserved through backups, restores, and even in some cases device wipes.

According to Apple the purpose of this data was for aiding your phones in locating itself in relation to the rest of the world. Apple stated in a Q & A that :

“The iPhone is not logging your location. Rather, it’s maintaining a database of Wi-Fi hotspots and cell towers around your current location, some of which may be located more than one hundred miles away from your iPhone, to help your iPhone rapidly and accurately calculate its location when requested. Calculating a phone’s location using just GPS satellite data can take up to several minutes. iPhone can reduce this time to just a few seconds by using Wi-Fi hotspot and cell tower data“ [7]



6: A visualization of the data recorded by Apple devices

However this did little to assay the concerns of the general public as there was no guarantee that Apple would not, at some future date, use this information for less than benign purposes. Also at issue in this case was the fact that the data stored on the phone was under no form of encryption, meaning that if a security flaw was ever uncovered that would allow a user’s to access those files on your phone, there would be nothing stopping them from learning your travel history. Even a stored backup of the phone did not, by default, encrypt or protect this data by default.

This case, just as the target case, raised very important issues. Specifically, what kinds and quantities of data should an organization be allowed to track? In the past this kind of question wasn’t particularly an issue because technological limits on processing and recording capabilities minimized the utility that data could provide. Now however, the amount of data that can be recorded is limited by the capital capabilities of the organization in question, and it seems the trend is for said organizations to grab as much data as possible with the hopes that it will be useful or profitable later. Apple, like target, also received a heavy amount of criticism from the public and faced the prospect of a tarnished image

It’s Not All Bad

Though the above examples illustrate the dangers of big data there have been several noted cases of when using Big Data and tracking user experiences has both been profitable and well received by the Community. Netflix, a company that provides on demand streaming of digital media founded in ’97 has done well with its use of Big Data. Boasting 194 million users [8] the company has a wealth of personal data to work with. It uses this data to create systems to recommend movies based off of both past viewing habits and user interaction with surveys. This data also allows Netflix to view user interest trends in both movies and television and adjust their offerings accordingly. As an example Netflix noted that a large number of people watched the show Arrested Development (one which had been prematurely cancelled), and in reaction backed a brand new season of the show.   Reactions by users were incredibly positive to both events, in part because Netflix gave them what they wanted, but also because Netflix stayed firmly within accepted and expected business practices.

7: Netflix Max uses Big Data to improve user experience

Google is another example of a company that takes steps to use Big Data in ways that most people find appropriate and beneficial. A prime example is the spell checking utility found in Google’s web browser, known as Google Chrome. The utility takes misspellings and corrections and records them in a database allowing the service to improve its overall functionality with use over time. Google’s text to speech works in a similar way, recording a user’s speech to improve its ability to recognize specific words and phrases. Though much of this information seems private, Google takes two significant yet simple (almost to the point of common sense) steps that help ease concerns. First, Google always asks before it records data in the above circumstances. This seems like something small but it ends up having a large impact. Many large organizations fail to take into account the fact that there is a significant difference between giving up privacy voluntarily to help others, and having it taken from you without your consent or knowledge. Second, Google informs you of the nature of the data its

8: Crome takes steps to make Big Data Ethical

taking and that it will take steps to make your data anonymous, further easing concerns of every day individuals. Last, Google informs users of the benefits their data will give to themselves or others, allowing users to make informed and conscious decisions and giving them perspective on the usefulness of their contribution.

Why do we need Ethics for Big Data?

The cases with target and apple both illustrate a lack of consensus on what is appropriate for organizations to record and use for the purposes of monetary gain. Part of this problem originates from the fact that Big Data is such a vast domain with a large variety of situations in which the capabilities it provides could be abused. This lack of standard means that, in most cases, individuals have to rely on their own personal code of ethics to make decisions regarding what constitutes and acceptable use of Big Data. Unfortunately this often degrades to the "but that's creepy … / No, it's not" [4] argument, which usually ends up helping no one.

The other part of the problem is how exciting and influential big data can be. That is not to say that using Big Data is inherently wrong, but that the vast and lucrative applications of this domain can often encourage a sort of recklessness in business decisions that can be unsafe. That fact, coupled with the general wide reach associated with big data, creates a situation where a single mistake or poor use of data can affect a very large number of people in a very short period of time.

These questions involving how to use data about real people, and the atmosphere currently permeating the field both lead to a single conclusion. That these questions are ethical in nature, and that a code or system of ethics that would give system designers and architects a frame of reference when deciding how to balance the risk of harm with the reward of innovation is entirely necessary if we are to avoid the kinds of blunders made by apple and target. This system would need to take into account the many different applications of ethics (personal, professional, business, and societal). Before delving into what ethics for big data should look like it is important to discuss some related key terms and concepts. In their work Davis and Patterson describe several of these in detail, of which a brief summary is provided below.

Rights and Interests

In their book Davis and Patterson make the distinction between rights and interests when discussing Big Data ethics. They point out the important distinction that the word right often brings with it the context and presumption of an absolute right that is inviolable. Unfortunately the use of data is so wide that the idea of an absolute right (absolute privacy for example) often hinders the process of development. The idea that a right is absolute should be as they put it "an outcome, not a presupposition." The word right in itself is complicated because in many cases it makes presuppositions about ethical views that shouldn't exist in this context as there really are no views to presuppose. They conclude that, in many cases, considering the interests of the client or the providers of the data allows for a more objective viewpoint to be taken as opposed to considering the "rights" of the client.

Personal Data

How one defines personal data is also important to nail down. This is largely due to the fact that personal information or what is can be tagged to an individual, often has a lot to do with available technology and can change rapidly. In the past only specific data (like a phone number) might be considered personal. In order for the ethics of Big Data to be sound it is important that this term is wide reaching. As such it is suggested to consider any data that is generated by an individual’s activities to be personal (because with enough effort that information could be used to identify someone).

A Responsible Organization

Davis and Patterson note that there is a significant “difference between doing right and doing what various people think is right” especially when relating to what is right for Big Data. As mentioned earlier it is often the case that anyone from a software architect to a manager can get caught up in all the “potential” of Big Data to the degree to which he or she might bend the rules slightly or simply do what is accepted, rather than what moral or ethical obligations would suggest. A responsible organization is not just concerned with how they are viewed in the eyes of others but is also concerned with dealing with data in such a way that actions align with the values of the company, and how those two concerns should interact.

What Does Big Data Ethics Look Like?

After defining Big Data, considering its importance, and also addressing why Big Data needs a code of ethics, we can come to a few conclusions. First, Big Data is not going anywhere anytime soon. It is too useful and lucrative of a tool to be thrown out because of the challenge of giving it ethical guidelines. Second, Big Data is both massive and diverse, and as such needs a set of guidelines that take those things into account. Finally Big Data is forcing questions that need to be answered should we all wish to avoid disaster. As Neil Richards and Jonathan King point out “The problem is that our ability to reveal patterns and new knowledge from previously unexamined troves of data is moving faster than our current legal and ethical guidelines can manage.” [9] Given what we have learned from those before us we can make considerations of our own in relation to the Ethics of Big Data and come up with a set of useful principles for remaining ethically sound and for facilitating ethical discussion.

Be Clear and Concise

First, any set of ethical principles and their implementations should be clear and concise as much as possible. This is an idea referred to by many sources as “Radical Transparency” [10]. This means letting the users know exactly what you or the system you architect does with their data while making assumptions for the level of technical expertise for each user. “Users do understand that nothing is for free; they just want to be told. Otherwise it would be like receiving a free book from the local bookstore and finding out later that the store still charged your credit card for it.” [10] There is almost nothing worse than being unable to explain to your users, in context, the reasons why you are taking and using their personal data. This scenario almost always plays out when a developer or security firm pours through one’s carefully architected software and finds something suspicious or ominous that the users weren’t told about. In many cases said finding is an artifact or a result of some entirely benign process or an unintended fluke, but it is very hard to make that argument when you were not forward with your users to begin with and are trying to play the damage control game. In order to avoid this scenario simply tell the user everything that your software is doing and, in the best case, everything the company has recorded or determined about them. This will not only build trust but will allow you to explain and spin to some degree the reasons why you are collecting data, instead of responding to an angry mob of customers who have already made up their minds as to who is in the wrong.

Give Users Power Over their Data

After telling the users everything that one's organization keeps records of, give the users a chance to decide what they wish to share and make that tool or decision simple. In an article on Big Data Ethics Jeffrey F. Rayport suggests that "One way to avoid an Orwellian nightmare is to give users a chance to figure out for themselves what level of privacy they really want." [11] This ties into the first point in that a simple and concise explanation and set of tools prevents users from being surprised and offended. Take for example the gaming store who, in 2010, added a clause to their Terms of Service that granted the company ownership of a shopper's eternal soul [12]. As humorous as that example might be, it highlights a common problem, the trend of lengthy and complex privacy agreements. Even if an individual does give up his or her rights through some sort of documentation, if the document deferring those rights is complex or vague such that it is hard for a common person to understand, the owner of that document will undoubtedly come under intense scrutiny for their actions. However, a simple and uncomplicated agreement will, at the very least, pass the blame of ignorance from the organization in question to the user who failed to read the 3 line description in the privacy settings page. In many cases a simple agreement entirely avoids inciting the anger of users and the public alike.

9: TOS agreements are often incredibly complex and difficult to understand. This characteristic often draws criticism from the public.

Communicate Value

Paired with a user's understanding of privacy is their understanding of the inherent value of their information In most cases, the more a company understands its clientele or user group, the better its service and products are likely to be at serving that group. Sometimes this understanding necessitates keeping user information that might be considered private. When an organization doesn't tell its users about the benefits of sharing this data, a user is likely to not want to do so. Most organizations realize this and, in order to remain competitive, take this user information without telling said users, a practice that has no ethical foundation. As such it is important to inform a user about the value of their data. Users realize, or can be made to realize, that everything comes at some cost. They are also often willing to pay that price as long as they know what they are getting in return. Netflix and Google (see above), are prime examples of companies that inform their users and have had great success. This form of transparency also works to promote Big Data Ethics as it encourages accountability and good business practices.

This partly because users will no longer feel that their trust was or might be violated (and will continue to do business with said organization) but is also due to the fact that, as organizations become more transparent with the use of big data, the ease of keeping them accountable will increase.

The Importance of Security

Security, while not directly related to Big Data, is an important aspect of the related ethics. An architect needs to very carefully define what personal or private data is necessary (instead of desired), and how that might conflict with the interests of the owners of that data. Once an architect has determined what data an application needs, it is important that he or she build in security around that data. Often the data collected is valuable, and it is the organization who lost that data that will take the blame, not those that took it. As such it is the ethical responsibility of an organization to protect not only the input data, but also the inferences that can be made with Big Data, from those who would obtain it illegally or without permission.

Building In Privacy

Another lesson to be learned is that the architect of an application should attempt to include privacy within the design. Just as security is often difficult to build into a piece of software after its completion so follows privacy. If an organization and designer considers the privacy and interests of a user from the beginning they are far less likely to do something ominous. Building privacy into applications also not only allows for one to differentiate their application from others (and thus gain business), but promotes a society that values these principles, instead of one that is consistently encouraged or tricked into giving them up.

Final Questions

After taking into account the above considerations one might find that there is some choice or decision in implementation that is not captured by said principles. In those cases it is important to fall back on pre-existing ethical perspectives as they provide questions that may rule out whatever option one is considering. The questions to consider are the following:

1. How does this architectural choice affect my organization as a whole? Will this use of Big Data hurt my organization if discovered either in the form of lost clients, public backlash, or tarnished reputation?
2. How does this architectural choice fit into the view of personal ethics? Does this choice violate a user’s privacy interests without any acceptable reason or benefit? Does this action negatively impact the life of an individual? Does this action actively go against the architect’s personal code of ethics?
3. How does this architectural choice fit into the view of Legal Ethics? Is this use of Big Data and private information illegal in one’s country or location of residence? Would this use inspire lawsuits or legal action that would be negative to the organization if discovered?
4. How does this architectural choice fit into the view of Professional Ethics? Will this decision or use of personal data affect how the public views Software Architects or engineers? Will that view change be detrimental and hurt the opinion of the profession as a whole.

5. How does this architectural choice fit into the ethical views of society? Is this decision socially acceptable? Will this use of data promote a change or changes in what society views as acceptable that are harmful, especially if those affected by these changes are not fully aware of the downsides when making their decision?

If the answers to any of these questions are negative and due to the nature of one's use of Big Data then it is important to reconsider the action to be taken or make changes such that there are no violations of the above principles.

Big Data Ethics in Practice

Now that we've seen guiding principles for Big Data Ethics, it is useful to look at real world examples of its implementation. Big Data Ethics is, more often than not, the result of communication and discussion within an organization about how best to implement the above concepts. As such it is beneficial for the reader to see where the industry is at today. In their work Davis and Patterson present their findings from several of the top fortune 500 companies relating to several aspects of how big data is used.

What Companies Have Clear Policies

Obvious to any observer is the variation in policy statements between organizations. This in many cases generates distrust in organizations as a lack of clear or consistent statements lead users to believe that a company is hiding practices. Making policies clear and uniform makes it simple for an Architect to align his work with company ethics and makes businesses accountable for their actions (as users can see a clear picture of what they are or are not giving away). Unfortunately this is still an area that needs significant work. Davis and Paterson found that, almost "all of the policies surveyed made some type of distinction between 'personally identifying' and 'anonymized' data. Nearly half of those, however, did not explain how they defined the distinction—or exactly what protections were in place." [4]

Defining these aspects of use is incredibly important because of how rapidly changes in the capabilities of Big Data are changing. Something that might not be personally identifying today may be that way tomorrow, and how a company has defined that information might allow it to be used for such purposes. How a company defines data that is anonymized is important for the same reason. Often this kind of data is open to use by the company because, at the current time, it can't be used to identify an individual. If that data is only anonymous because of capabilities, and not because there is no way to correlate that data with an individual, problems will arise.

What Companies Give Users Power Over their Data

Davis and Paterson also discovered that one of the most used methods for giving users power over their data was allowing a user to "opt-out" of data being used in specific ways or shared between organizations for business reasons. Unfortunately opting out meant not using a product or not agreeing to a terms of service instead of providing the user a means to still benefit from the product and not have their data taken. Also, while some organizations gave users the opportunity to still use their products, the methods for opting out or restricting data were in many cases difficult and/or complex, requiring signing and mailing several forms in some cases.

It is easy to see why an organization might take these kinds of actions. In many cases requiring a user to opt-in nets almost no benefit as it requires extra input from a user often without any clear benefit. Fear of the unknown as Davis and Paterson put it, is also a problem, as many customers will simply not opt-in (or would opt-out if the methods were easy) because they did not understand and would not take the time to understand the benefits. The problem with this kind of practice is that it is impossible to inform a user of what can be done with their data, as the capabilities of Big Data are often changing. A person who didn't opt out today might very well have chosen to do so a year from now because of what information that data, when combined with other data sets, might reveal.

Who Owns What?

It might be argued that, other than for scientific and educational purposes, Big Data exists mainly to generate revenue. In many cases one must either own or license something to generate revenue from that thing. This leads too many questions about the data that customers provide to organizations and how control of those assets should be distributed between the user and said organization. While many organizations, as will be discussed below, state explicitly that they will not sell their users data, they make no attempts to assign ownership to any one entity. This lack of exposition means that, as far as an organization is concerned, they can use the data in any way that benefits their business, which is a frightening conclusion to be sure (though less so if the companies inform users and allow them to opt out). Unfortunately there is no consensus across organizations that have been found in regards to this topic, and as such this remains an area that would benefit from further scrutiny. This scrutiny would hopefully result in agreements between corporations and individuals that were explicit in stating what can and cannot be done with data, instead of ones that make a few rules and leave everything not mentioned up to the group that controls the data.

How is data bought and sold

In their research Davis and Patterson found that over 75% of interviewed companies said explicitly they would not sell personal data. There were however, no companies that would make concrete statements about their decision to or not to buy personal data. This leads to the observation that this area, the decision to buy data, is something that needs to be challenged by members of organizations as well as their customers. This is especially important because, more often than not, those who have provided personal data have no control over who buys it, and the existence of buyers encourages companies to sell data, with or without their user's knowledge.

Overall, if one draws anything from the above investigation, it is that while many companies are on the right track towards ethical use of Big Data, there is still much work to be done. In many cases companies and organizations will only do what takes them out of public scrutiny instead of what would be best for everyone (not selling, but buying data), or

9: how user data is bought and sold

only what is required by the rule of law. It will only be by applying the previously mentioned principals to the current business climate that software architects will be able to change the ethical practices of business involving Big Data for the better.

Privacy Erosion

This chapter concludes with a short discussion on privacy erosion, a topic related to Big Data and one to keep in mind when discussing its related ethics. As discussed earlier, Information Technology changes how we as a society access, search, and make decisions regarding data. As the rate of data generation and capture increases rapidly (from added sensing capabilities and cheapening data storage), so do the inferences that can be made from said captured data. Many times, these actions and transformations can reveal, intentionally or unintentionally, data that would violate a person's civil liberties (especially when considering governments or large organizations). There might one day be a point when Big Data calls into question the right to privacy that many governments give to their people (the 4th amendment in the U.S. for example). Online surveillance is becoming the norm: ISP's (internet service providers) track and sell data about consumers, websites download cookies that can be used to track information, and cellular companies can track the locations of users through cellular towers.

The growing concern is that this erosion of privacy, or the difficulty in keeping one's information within one's own control, is becoming the standard for the future. Society will, over time, become more comfortable with the erosion of privacy we see today simply by the fact that, generations from now, the expectation of privacy one grows up with will be entirely different then what we have today, especially when considering that the erosion of privacy can be beneficial when used in the proper way. It is important then, as a closing note, to consider how the decisions of the reader as an Architect will affect future generations, as that type of foresight is often absent from planning meetings or presentations in front of superiors.

References

[1] Hilbert, Martin, and Priscila López. "The world's technological capacity to store, communicate, and compute information." Science 332.6025 (2011): 60-65.

[2] Mateosian, Richard. "Ethics of Big Data." IEEE Micro 33.2 (2013): 0060-61.

[3] Wen, Howard. "Big Ethics for Big Data." Data. O'Reilly, 11 June 2012. Web. 29 Apr. 2014. <http://strata.oreilly.com/2012/06/ethics-big-data-business-decisions.html>.

[4] Davis, Kord. Ethics of big data. O'Reilly Media, Inc., 2012.

[5] Conway, Rob. "Where angels will tread." The Economist. 17 Nov. 2011. The Economist Newspaper. 29 Apr. 2014 <http://www.economist.com/node/21537967>.

[6] Duhigg, Charles. "How Companies Learn Your Secrets." The New York Times. 18 Feb. 2012. The New York Times. 28 Apr. 2014 <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0>.

[7] "Apple - Press Info - Apple Q&A on Location Data." Apple - Press Info - Apple Q&A on Location Data. 27 Apr. 2011. Apple. 29 Apr. 2014 <http://www.apple.com/pr/library/2011/04/27Apple-Q-A-on-Location-Data.html>.

[8] "Netflix." Wikipedia. 28 Apr. 2014. Wikimedia Foundation. 28 Apr. 2014 <http://en.wikipedia.org/wiki/Netflix>.

[9] King, Jonathan H., and Neil M. Richards. "What's Up With Big Data Ethics?" Data. 21 Mar. 2014. O'Reilly. 29 Apr. 2014 <http://strata.oreilly.com/2014/03/whats-up-with-big-data-ethics.html>.

[10] Rijmenam, Mark. "Big Data Ethics: 4 principles to follow by organisations." BigDataStartups. 11 Mar. 2013. Big Data Startups. 29 Apr. 2014 <http://www.bigdata-startups.com/big-data-ethics-4-principles-follow-organisations/>.

[11] Rayport, Jeffrey F. "What Big Data Needs: A Code of Ethical Practices | MIT Technology Review." MIT Technology Review. 26 May 2011. MIT Technology Review. 29 Apr. 2014 <http://www.technologyreview.com/news/424104/what-big-data-needs-a-code-of-ethical-practices/>.

[12] Bosker, Bianca. "7,500 Online Shoppers Accidentally Sold Their Souls To Gamestation." The Huffington Post. 17 Apr. 2010. TheHuffingtonPost.com. 29 Apr. 2014 <http://www.huffingtonpost.com/2010/04/17/gamestation-grabs-souls-o_n_541549.html>.

[13] Allan, Alasdair, and Pete Warden. "Got an iPhone or 3G iPad? Apple is recording your moves." OReilly Radar. 27 Apr. 2011. O'Reilly. 29 Apr. 2014 <http://radar.oreilly.com/2011/04/apple-location-tracking.html>.

Figures

[1] O'Keefee, Anthony. "Blog." Big Data. 29 Apr. 2014 <http://www.lucidity.ie/blog/173-big-data>

[2] "Moravec Robot book figure." Moravec Robot book figure. 29 Apr. 2014 <http://www.frc.ri.cmu.edu/~hpm/book98/fig.ch3/p060.html>.

[3] Melissa. "Safety: Protecting your digital footprint." Digital Family Summit. Digital Family Summit. 29 Apr. 2014 <http%3A%2F%2Fwww.digitalfamilysummit.com%2F2012%2Fsafety-protecting-your-digital-footprint%2F>.

[4] Gregorious, Thierry. Wikipedia, http://commons.wikimedia.org/wiki/File:Big_data_cartoon_t_gregorius.jpg

[5] "Big Data: How Target Knows You Are Pregnant - Yu-kai Chou & Gamification." Yukai Chou Gamification. 29 Apr. 2014 <http://www.yukaichou.com/loyalty/big-data-how-target-knows-you-are-pregnant/>.

[6] Allan, Alasdair, and Pete Warden. "Got an iPhone or 3G iPad? Apple is recording your moves." OReilly Radar. 27 Apr. 2011. O'Reilly. 29 Apr. 2014 <http://radar.oreilly.com/2011/04/apple-location-tracking.html>.

[7] http://www6.pcmag.com/media/images/391044-netflix-max.jpg?thumb=y

[8] http://www.logobird.com/wp-content/uploads/2011/03/new-google-chrome-logo.jpg

[9] Downey, Sarah A. "9 easy ways to beat identity thieves." Online Privacy Abine. 22 Jan. 2013. Online Privacy Blog. 29 Apr. 2014 <http://www.abine.com/blog/2013/beat-identity-thieves/>.

[10] Tacma.net