# Data Is the New Currency

Carrie Gates
Dell Research
Round Rock, TX
carrie_gates@dell.com

Peter Matthews
CA Labs
London, UK
peter.matthews@ca.com

## ABSTRACT

Data is growing. We are all aware of this in the IT industry, it is a common mantra. The elephant in the room is the ownership of that data and the use of that data. As with many new technologies, its legal and personal implications are not well understood until the technology has matured. Data ownership has by default resided with organisations that hold the data; utility companies, websites, retailers and data aggregators and brokers. If data could be owned by the people it identifies, the data handlers would have to pay to use that data for sales and marketing purposes. We are not suggesting payment would be the mostly illusory free services and hidden discounts that are the current answer but real money or an asset that can be bartered or purchased. If even the poorest people can gain an income this would revolutionise personal finance for those who register on the web either for purchases or to make their name available for use. We propose such a revolution in data ownership and urge all data generators to establish their rights to a business asset that they frequently gift to already wealthy organisations. Don't be a mere generator of data, become a personal data aggregator, collecting and controlling all your data. Time to man the barricades against entities who are using your data for their profit and pleasure, not yours.

## 1. INTRODUCTION

Data has been collected from the earliest times. The change from hunter-gatherer to a more static agrarian culture generated the need to record assets on tokens. According to historians, between 4100 to 3800 BCE written language began to develop. A concerted effort by William the Bastard to register all the assets in his new kingdom of England resulted in the Domesday Book [1]. This document compiled in 1085 to 1086 CE is still a functioning legal document regarding land claims and is one of the earliest moves to create the dossier society [2]. Many cultures before and since have gathered data about their citizens, not always with beneficial effect. The records of the Stasi come to mind.

Data gathering has passed out of the government based records keeping and archival into the public domain. In the early days of computing, utility companies compiled customer data and later sold access to other organizations for marketing and sales initiatives. This data gathering has accelerated on an exponential scale resulting in whole new areas of business. Data aggregators or brokers range from the obvious, Google, Amazon and your favorite government departments, to the more hidden MEDBase200[1]. All of these have one thing in common: they claim ownership and asset rights on your data.

You may be offered a small discount, real or illusory, for the use of the data. "Free" emailers are one of the best known examples. They make money from advertisers targeting the customers of the email service and also make money for providing demographic and other data to marketeers and sales. The cost of running the email service and the profits from advertising and marketing are hidden from you. All you get is an email service that is managed by an organisation who makes money from selling data about its users. We are not suggesting that the reputable free email suppliers are taking intimate details from your emails but that there is a vulnerability of individuals in this circumstance.

One issue with data aggregators we highlight in this paper is the lack of legal oversight. Lawyers have been debating data privacy for many years and working on potential solutions for some time but The General Audit office of the US Government has concluded that: "No overarching federal privacy law governs the collection and sale of personal information among private-sector companies, including information resellers." Individuals are being constantly warned about data theft and how they can prevent it, however there are no such constraints on data aggregators and list owners who sell lists for many purposes for less than 100 dollars. Neither are there any constraints on accuracy. It is hoped that you can check the validity of personal data held by a data aggregator, but you can't check that if you don't know who is aggregating data. The lack of regulation and oversight pervades data management and the twin issues of privacy and personal data are discussed in later sections.

This paper explores the legal status of your data, the value of your data, as well as some commonly understood issues with data. The paper explores the massive potential of your data, current ownership of your data and why that ownership should revert to you. The potential personal and

[1] `http://www.medbase200.com`


*NSPW'14,* September 15–18, 2014, Victoria, BC, Canada.

http://dx.doi.org/10.1145/2683467.2683477.

societal benefits of owning your own data can be significant but only realised if we all agitate for a revolution in data ownership.

The rest of this paper is organized as follows: in Section 2 we provide definitions for some of the terms we use, including currency, privacy and data. We then go on to talk about data ownership in Section 3. From there, we attempt to determine the value of data (and privacy) in Section 4, where we discuss the value of individual data for both direct marketing purposes and on the blackmarket. In Section 5 we identify some of the problems associated with data collection and storage. From there, we describe three possible solutions — ranging from legal to technological to economic — followed by drawing a set of conclusions in Section 7.

# 2. DEFINITIONS

## 2.1 Currency

Currency is defined in the Oxford English Dictionary as a system of money in general use in a particular country. This definition of currency depends on a definition of money. Money is defined as the assets, property and resources owned by someone or something.

From these definitions it can be assumed that personal data can be classed as an asset, property or resource of an individual and therefore currency. This creates the need to debate what is personal data. As an example, if a utility company holds your personal name and address details as a result of your consumption of a service the data can be considered as money or at least an asset owned by the utility company. There is however a viable claim that this data should be owned by the individual, as we explore in Section 6.3.

## 2.2 Privacy

### *2.2.1 Legal*

In the United States privacy legislation is founded [2] on a paper in 1890 by Samual Warren and Louis Brandeis Warren in the Harvard Law Review, who extolled "a right to be left alone" [32]. From this basis, both constitutional and statutory rights have been defined, with three distinct privacy rights being recognized: [5]

1. The right to privacy with regards to access to personal data: Several acts exist to limit the access of organizations and government to what has been defined as personal data in the context of the act. Some examples of acts that limit access to data include:

   - Privacy Act of 1971 — this act defines how the United States federal government can use data
   - Gramm-Leach Bliley Act of 1999 — this act defines protections for access to financial data
   - Health Insurance Portability and Accountability Act of 1996 — this act defines protections for access to individual health data

2. The right to privacy in regards to personal autonomy: This right is a "liberty" recognized as part of the due process clause of section 1 of the 14th Amendment. This right is narrowly defined with the constraints of protecting the "privacy of family, marriage, motherhood, procreation, and child rearing" [4]. It is actually this right that was used in the arguments in Roe v. Wade, 410 U.S. 113 (1973), and argues for the right of personal autonomy of the mother, and that the state has no compelling interest until viability. Similarly, this right also protects parents, allowing them the right to personal autonomy in their choice of parenting methods and the number of children to raise.

3. The right of publicity: This right protects an individual's personal likeness, stating that it can not be used commercially without permission. This right is not protected at a federal level, but rather at the state level, with many states not recognizing this right explicitly. For example, some states protect this right via laws on unfair competition.

Civil law also recognizes some right to privacy, such as that identified in the Restatement (Second) of Torts §652C. This tort law (civil law that is not criminal in nature and that carries a lower burden of proof) recognizes three types of invasions of privacy [15]: (1) intrusion into seclusion, (2) appropriation of name or likeness, and (3) unreasonable publicity. (Note that some add a fourth invasion — false light — to the list, e.g., [2], [28].) Intrusion into seclusion refers to someone having their solitude intruded upon (physically or otherwise), such as by having someone break into their home, or wiretap their phone, or use binoculars to spy on someone. Appropriation of name of likeness refers to a person using someone else's identity for commercial reasons without permission (e.g., as a product endorsement). Unreasonable publicity refers to the right that a person has to not have articles about their private life publicized (e.g., as in made public via publication), if such publicity is not of legitimate concern to the public and would be considered to be highly offensive to a reasonable person. False light refers to presenting someone in a manner that is both false and negative.

At a federal level, balanced against this right of privacy is the need for a government to protect its citizens from harm. Thus the government has recognized a need for surveillance and the collection of data on specific subjects that they suspect may be involved in activities such as espionage or terrorism. This need has been enshrined in the Foreign Intelligence Surveillance Act of 1978. In this paper we note this as a competing need against privacy, but do not investigate the balance between government protections and personal privacy further.

### *2.2.2 Personally Identifiable Information and Re-identification*

Personally Identifiable Information (PII) has been identified by the Government Accountability Office (GAO) in the US as "any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information." [13] The key point to note about this definition is that it assumes that there is some single key that is unique to an individual (e.g., name, mother's maiden name, social

security number) and that can be used to link data about an individual. From here, the definition focuses on the types of data that have traditionally been considered to be private (e.g., health, finances).

However, it has been found that individuals can be identified reasonably reliably from sets of generally non-unique data, much to the chagrin of organizations that have released data publicly for research purposes. For example, Narayanan and Schmatikov found that they were able to de-anonymize users in the Netflix Prize dataset by cross-correlating the data with the Internet Movie Database (IMDb[2]), including determining other data about individuals such as their apparent political preferences [25].

In general, when we discuss PII, we are actually concerned with the act of re-identification. Narayanan and Schmatikov describe re-identification as having two key properties: "(1) they are reasonably stable across time and contexts, and (2) the corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability." [26] Indeed, it is these two properties that allowed the Netflix movie preference data to be linked back to individuals. As it turns out, people are actually surprisingly unique in their preferences, and thus re-identification becomes possible from surprising sets of data. This makes the issue of what data can be released publicly, or used even by organizations in non-public settings, more complicated as it becomes more difficult to guarantee the anonymity of the subjects in the data set.

### 2.2.3 Individual

In addition to legal definitions, individuals also have a definition for privacy, and this definition tends to vary by both individual and context. (See, for example, Kwasny *et al.* [19], for preliminary results examining individual privacy definitions and perspectives across generations.) For example, an individual's definition of what they consider to be private data will be different based on their employer (e.g., locations requiring high-level clearances versus a university), their location and culture, and their context.

As an example of how the privacy of certain data varies by context, consider the simple example of going out on a date on Friday night. Is this private data? In the case that you have just started seeing someone, then this might be data that is only shared with close friends. But if the relationship is serious, then it might be data that gets posted on Facebook. On the other hand, if you are married and your date was someone other than your spouse, then you might consider this to be extremely private data that should not be shared with anyone!

Now the question becomes does this piece of data — the fact that you went out on a date on Friday night — have any value? Excluding any possible interest by the government or law enforcement (depending on who you are or who you are dating), this data could be valuable in the aggregate form for research (e.g., how many single women are dating?). The data is perhaps of higher value to the restaurant you visited, as they might desire to attract you back to the restaurant (maybe through some special offer of free roses with dinner and wine), and want to provide personalized service (e.g., given that you had this wine last week, might I recommend this other wine this week?). And perhaps of higher value still is this data to competitors who might wish to lure you away. Thus this data has value for targeted marketing.

Individuals also have different notions of what they therefore consider to be a breach of privacy. Continuing with this example, for some people, having the restaurant send provide incentives to return (such as the free roses) might be considered a nice gesture, while for others this might be considered an intrusion into their private life (why should the restaurant keep track of my bringing dates there?).

The perception of a breach of privacy sometimes has more impact than a real breach of privacy. A good example would be the controversy related to the relative merits of the use of cookies or deep packet inspection (DPI) for behavioural advertising. A study by Khuen and Mueller [18] shows that the perceptions of technologists related to the forging of cookies and redirection needed to implement DPI gave the impression of a flawed technology and potential breach of privacy. Comparing DPI to cookies as a way of activity tracking showed no more privacy problems for DPI but it became a failed technology because of the perception of privacy breaches.

## 2.3 Data

We focus here on definitions of data taken from the field of information management. We note that even this field, which one would expect to have developed crisp definitions, suffers from a multitude of assumed definitions. Zins published a series of definitions in 2007 [33], as provided by a number of experts in the field. The commonality noted was that the majority of experts provided definitions that used nonmetaphyscial, human-exclusive, cognitive-based approaches. In this paper, Hanne Albrechtsen provides a computationally-focused definition of data, information and knowledge:

> **Data.** In computational systems data are the coded invariances. In human discourse data are that which is stated, for instance, by informants in an empirical study. **Information** is related to meaning of human intention. In computational systems information is the contents of databases, the web, etc. In human discourse systems information is the meaning of statements as they are intended by the speaker/writer and understood/misunderstood by the listener/reader. **Knowledge** is embodied in humans as the capacity to understand, explain and negotiate concepts, actions and intentions.

Personal data is defined as data which relate to a living individual who can be identified. This definition comes from the UK Government's guidance on the meaning of the Data Protection Act of 1998. There are other definitions, however this definition is used as it represents a Europe wide definition. This definition excludes data that is created by an individual, covered under copyright, and data such as purchases and online activity as long as there are no personally identifying data. Privacy has already been discussed in Section 2.2. It is interesting to note that again lawyers have been reviewing the differences for many years. Karen McCullagh argues in her 2008 paper [22] that the provisions of the UK Data Protection act of 1998 are flawed in that they refer to personal data as defined above and do not make any provision for private data. Even with such a clear, if concise

[2] `http://www.imdb.com`

definition of personal data the international legal domain is confused. In the EU this definition is used widely and is deemed adequate. The US has no federal data protection provision but uses Personal Identifiable Information (PII) as an accepted term in the various state and federal laws that fragment and clash in wild abandon.

The scope of this paper goes beyond personal data in EU terms and PII in US terms. Here, we define data as the invariants about a person or activity at a particular point in time (e.g., age, address, income, items purchased). We define information as being the meaning derived from that data — also referred to as inferred data — such as inferring education level based on a person's address. We also make the distinction between active and passive data creation.

Active creation of data, writing content, developing source code, writing emails is frequently covered by copyright law. If you wonder who owns the emails you write, take a close look at the terms and conditions of companies like Google or your own employment contract. Emails and writing notes and documents are covered by copyright laws in your jurisdiction, assuming that this data is not stored offshore in a jurisdiction that does not recognise copyright. Copyright laws are complex and have provided and will provide sizeable incomes for the legal profession.

Passive data creation is achieved by buying something. The data that you create is a function of an activity, buying from a web site, renting a broadband or using power from a utility company. This data is not only classed as private but it is also personal data due to the personal identification that it contains. Other data that is passively created concerns physical or mental features such as eye colour, nervous habits and DNA. It can be argued that this is again personal data because of the increasing likelihood of being able to identify an individual based on that data.

The increasing proliferation of data that comes from devices and sensors may result in passive data creation and we only consider this data when we are considering data that is stored for identification such as medical scans. The evolution of the Internet of Things may alter these definitions going forward however and may need to be considered in the future.

In this paper, we refer to data as including: personal data (in EU terms), passive personal data creation (as described above) and inferred data. In general, we consider data that can be tied back to an individual, and that is *about* the individual, whether provided directly by the individual or not.

### 2.3.1 Data Usage

Data that has been collected or inferred can have a variety of uses, the most obvious and best known of which is for marketing purposes. Loyalty cards were first instigated in order to track user purchases and target advertising based on the user's puchasing history. There are, however, other reasons that an entity might want to collect personal data. For example, governments might want to use personal data in order to determine what it should employ for ecomonic policies, or what effect specific government programs might be having. A good example of this is the census. Starting with asset lists like the Domesday book governments have gathered personal information for taxation and identification purposes and later for forward planning. Use of lager data sets can identify trends in changing populations and

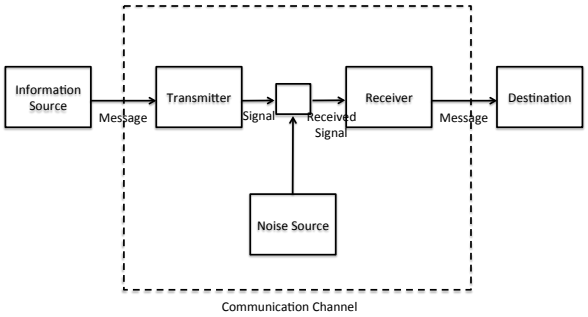


**Figure 1: Schematic Diagram of a General Communication System [30]**

behaviours. This is also one of those increasingly common demands from government that you cannot ignore. In many jurisdictions in the UK and USA, for example, it is a legal requirement to complete the census report.

### 2.3.2 Shannon's Model of Communication

In 1948, Shannon published a mathemathical model of data transmission [30]. While the mathematical modeling of data and information is beyond the scope of this paper, we do note the schematic diagram of a general communication system that he provided, as reproduced in Figure 1. This model can actually be modified to represent data as it gets stored and transformed. In the simplest case, the information source is the individual, while the message is the data being provided (e.g., address, age). This situation becomes more complex in the case of data, since data can be collected simultaneously from multiple sources and combined, with inferences made based on the values of the combined data. Such a situation does not exist when examining strictly communication in the telecommunications sense. We therefore expand Shannon's model, as represented in Figure 2.

In this new model, we have demonstrated that there are multiple information sources, all feeding into one location (transmitter). These information sources might include the individual providing the data, but can also be other organizations that have collected data from (or about) the user and who are now selling it. This information is combined and used to then further *infer* data about a user (e.g., likely income based on address, if the property is being rented or is owned, and employer). Inferred data is considered to be a noise source as it is less likely to be correct than collected data. (Although we note that collected data might also be incorrect due to items such as the age of the data itself.) The noise source then combines with the collected data and is sent to the receiver (e.g., the receiving company) who then store it at some final destination (e.g., in some database). The final information that is stored can then act as an input to later data collection for the same receiver (which we put as a noise source, since it may already contain noise from

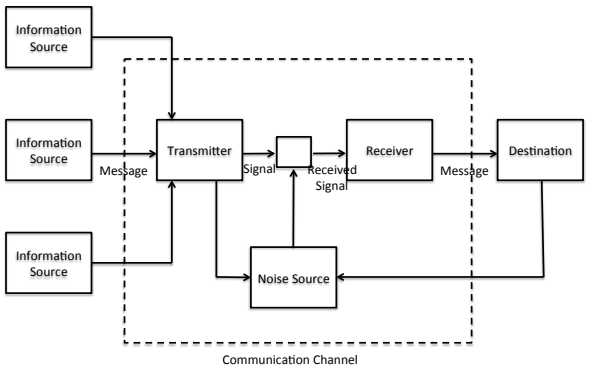


**Figure 2: Schematic Diagram of a General Data Aggregation System**

how the data was previously obtained), generating a feedback loop that continues to degrade the quality of the data collected by that particular receiver.

The key takeaway for this model in the context of this paper is that we define data to not only include direct information willingly provided by an individual under some circumstance, but also the combination of that data from multiple sources, and the *inferred data* that comes from the combination of that data. Thus in our context, (personal) data is data *about* an individual. The second takeaway is that the resulting data about a given individual (e.g., salary, purchasing habits) might not be accurate. The third takeaway is that the individual, while potentially being one of the data inputs to a given receiver, has no insight into the other data sources being tapped, nor the final data stored about him.

# 3. DATA OWNERSHIP

The concept of data as currency hinges on the question "who owns data about an individual?" At first sight it seems logical that an individual owns data that refers to them. A review of section 2.2.1 will show that there are several laws that protect such assets as your image. The Restatement (Second) of Torts noted in that section recognises the appropriation of name or likeness as an invasion of privacy. Some US states have a "right of publicity" protecting individuals from commercial use of their image without permission. Most privacy and identity protection laws appear to have a common point; an individual has to give permission for their data to be used commercially. Whatever the legal position it is evident that in registering on a website for retail purposes an individual is most likely to be giving permission for the use of their data by that website. This permission is often noted in the various agreements that are signed with a click and not read by all but the most diligent. It is possible that millions of people are sleepwalking into giving permission to use their data to both reputable sites and to some dubious sites. Further, as identified in Section 2.3.2, data can be combined from multiple sources and used to generate entirely new data about an individual. What ownership rights does an individual have to data *about* him, that was not at any point provided *by* him?

## 3.1 Who Owns Your Identity?

Ownership of data is inextricably linked with ownership of identity. Identity theft is frequently facilitated by access to PII data. While there may be a common understanding that there is personal ownership of personal data the reality is that this data is "owned" by the organisation that has that data in its domain. Most of these organisations have acquired data as a result of the individual's retail or other online activities. These organizations, through their web sites, demand data in exchange for the use of online facilities. Enter any website and unless your activities are superficial you will most likely be asked for your name and email address. These have become almost mandatory identifiers and the requests for data are often justified because they "will make your next visit easier/faster/cheaper". The entity behind the website is almost certainly expecting to use this personal data to expand their business by direct marketing, sales or just selling access to some of the data. Individuals are presented with no other option. Most websites will expect registration prior to use of their facilities. There is an escalation of online proliferation of personal data generated by the use of the web. Traditional businesses will also take data but this is secondary to their providing a service that is well understood. The ownership of data is more difficult to establish if the business holding the data sees that data as a major asset and a major influence on their stock price. Although the data may be personal to an individual the analysis of that data is frequently claimed to have almost magical properties. Anyone who has read press releases about the marvels of Big Data Analytics will understand this.

## 3.2 New Challenges

There are a few classes of personal data that are hot debating topics and likely to stay hot for some time to come. Google glass has started a whole new debate about recording personal conversations and activities. The record feature of a unit that is worn full time by the user has caused problems in bars and other social locations. There is a big difference between your actions being recorded by a security camera and the same actions being recorded with Glass or other devices and ending up on social media. There is a school of thought that maintains activities in a public place are in the public domain however people are becoming increasingly concerned by the unregulated surveillance that can be carried out by people wearing things like Glass. This has led to useful technology being banned in bars and meetings. If the data being recorded was owned by the person being recorded then the person recording would have to ask permission before uploading the data or face legal action. Ownership of an individuals image even if they are in a big party should still reside with the individual. This may remove some TV programs based on "funny" personal film clips from the air... but would that necessarily be a bad thing?

Another class of personal data that will stretch the data ownership argument is DNA. The Myriad Genetics litigation has been running since 2009 and shows no sign of stopping soon. This litigation concerns patents awarded on BRCA1 and BRCA2 genetic markers for increased risk of breast can-

cer. DNA is considered sufficiently unique to be used as unquestioned proof of identity in criminal cases. Genes are assigned to universities and companies. An individual's genetic makeup is a unique identifier and their rights to their personal genome are potentially in conflict with the growing market in genetics based medicine. Ownership of personal DNA may be established by the courts in the long distant future but until then most people are in the dark that part of their genetic makeup may be the property of a company on the other side of the world.[3]

These examples indicate that ownership of data and thus identity is not clear. There are many views of ownership ranging from the individual who fails to understand that in putting data onto a website they may be gifting a valuable asset to a company, to the genetics company that has invested millions isolating a particular gene and who wants protected exploitation rights. As new technology or new uses for old technology continue to affect data, ownership will continue to be debated.

# 4. VALUE OF DATA

## 4.1 Value of Your Identity

While there is no definitive price list available to determine what your identity costs, we can infer values based on cases where your identity is being sold, such as on the black market. SecureWorks recently (December 2013) released a report that provided the prices they were observing on sites that focus on selling identity-related data. Highlights from this report include: [6]

- $25 in the US and $40 in Europe for a fullz (a fullz includes a complete identity, such as name, address, phone, email, passwords, birthdate, ID number (e.g., SSN), plus a bank account or credit card number)
- $4 for US Visa or Mastercard up to $8 for UK or European credit cards (the price difference is because it is cheaper and easier to launder money from Europe than it is from the US)
- $25-$100 doxing (which is when a hacker is hired to get all the data she can about a target victim, via social engineering and/or infecting the victim with an data-stealing trojan)
- $300 for a bank account with $70K - $150K in it
- $11 for a birthdate

In addition to the value of identity data on the black market, we can infer a value for that data from the cost (in both time and money) to a victim of identity theft. For example, Javelin Strategy & Research found that: [17]

- the average resolution time for victims of all types of identity theft was 12 hours (and so multiply this by the victim's hourly rate to obtain a final value),
- the average consumer cost of identity theft resolution was $354,
- the average resolution time in new account fraud was 26 hours (this is the case where someone opens a new account in a victim's name, as opposed to using a victim's existing accounts), and
- the average consumer cost of identity theft resolution in new account fraud was $1,205.

Finally, there is the issue of identity insurance, which is starting to gain traction within the United States. This directly measures the value of my identity to *me*, in terms of what I am willing to pay in order to protect it. IdentityGuard is one company that sells such insurance — for $20/month a person can have their identity insured for up to $1 million.[4]

We note that comparing each of these sources does not provide a consistent value for a person's identity; however, it does provide some indication of the value for this type of data in different contexts.

## 4.2 Value for Targeted Marketing

As noted in Section 2.2.3, data has value for targeted marketing purposes. Take, for example, Target's foray into targeted marketing as described in a New York Time Magazine article by Charles Duhigg [10]. In particular, Target hired a statistician and charged him with determining if it is possible to determine if a woman is pregnant early in her pregnancy (e.g., first trimester) based on her buying habits. The reason that Target was interested in this data was because it is one of the few times in a person's life where they are likely to change their buying habits, and Target wanted to take advantage of this shift in order to attract new customers and convince existing customers to expand their shopping to include more items sold by Target (e.g., groceries in addition to house wares). Pole, the statistician in question, was successful in determining this data, to the point where Target sent targeted advertising to a teenager living at home with coupons for new baby items — and in so doing demonstrated that Target knew the girl was pregnant before her father did. (Her father was originally incensed at the advertising until his daughter broke the news to him.) Target found that "... even if you're following the law, you can do things where people get queasy" (Pole), and so changed their model from obvious targeted advertising (how did Target know I was pregnant when I never told them?) to embedding the coupons for baby items amongst random other items (e.g., BBQs) so that the recipient did not realize they were being targeted. "As long as we don't spook her, it works."

A second example comes from advertising that OfficeMax sent to Mike Seay, which inadvertently listed the recipient as "Mike Seay Daughter Killed in Car Crash or Current Business" [27]. Mike Seay's daughter had died in a car crash one year earlier. In this instance OfficeMax has blamed a third party data broker, leaving one to wonder if the data broker is collecting data linking people to deceased relatives (in order to market coffins?) or if this was a sensitivity notation that was put in the wrong field (e.g., added by someone to insure that either Mike Seay would not be contacted, or that the contacter would be sympathetic from knowing about his situation).

These examples are not unusual and, in fact, there are many lists available for purchase using criteria that might be

[3]We note that the US court ruled in 2013 that genes cannot be patented; however, this is a US-centric ruling and does not necessarily represent the ultimate views of other nations.

[4]`https://www.identityguard.com/how-identity-guard-works/id-insurance/`

considered objectionable. Pam Dixon, the CEO for World Privacy Forum, testified before a senate committee in December 2013 regarding data brokers and some of the lists available [9]. She notes:

> Data brokers sell lists of people suffering from mental health diseases, cancer, HIV/AIDS, and hundreds of other illnesses. Data brokers sell lists of people who live in or near trailer parks so that these undesirable consumers can be targeted for suppression. Data brokers sell lists of people who are late on payments, often to those who make predatory offers to those in financial trouble. Data brokers sell lists of people who are impulse buyers or "eager senior buyers".

Later in her testimony, she shows web sites where one can purchase lists of police officers including family data and home address, rape sufferers (at 7.9¢ per name), a list of seniors suffering from dementia, and a list of people with "addictive behaviors, alcohol and drugs". The type of data provided includes home addresses, the presence of any children and their ages, income, ethnicity, occupation, and even if they are a pet owner. Acxiom, one data brokerage, clusters individuals based on the data collected, into one of 70 categories, allowing companies to purchase very specific lists. Further, Acxiom provides enough data on each cluster to help guide marketing efforts. She notes that few data brokers allow opt-out, and that many that do often make it difficult or even expensive (> $1000) to do so.

Another example comes from Strategiclists[5] who, while they do not publish their list prices online, note that they will work with clients to refine their requirements to a much more targeted list, such as "Women, who own a dog or dogs, who live within a 5 mile radius of a certain zip code, who have a household income of at least $70,000, who have an interest in organic cooking, and who have responded to a mail offer in the past." Charles Duhigg notes in his article that: "Target can buy data about your ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own." [10].

## 4.3 So ...

What does this data cost? LeadsPlease publishes their mailing list prices online[6]. Prices vary by the size of the database purchased and the type of data (e.g., new home owners versus consumers), ranging from 6¢ per record (for 50,000 consumer records) to 20¢ per record (for 250 new homeowner records). The prices provided by sites listed by Pam Dixon [9] were 7.9¢ per person. But what is the value to the individual who is being targeted in a marketing list because he has HIV/AIDS? Is his privacy worth more than 8¢?

In the case of DNA, Myriad Genetics spent years in court protecting patents on BRCA1 and BRCA2 (genetic markers for increased risk of breast cancer) and cDNA (a synthetic product for detecting breast cancer). They claimed that they had spent years and millions of dollars performing research to identify the genetic markers, and therefore were entitled to being able to patent them. In June 2013 the Supreme Court struck down claim one (regarding the patents on the genetic markers) but upheld claim two (regarding the synthetic product that was developed). Thus, essentially, the Supreme Court ruled that a company cannot patent genes [20]. Prior to this ruling, Myriad Genetics had a $1 billion annual market from breast cancer tests. If this data (the genetic markers) were actually owned by the women who have them, then a back-of-the-envelope calculation shows that some women would be worth approximately $50 annually.[7]

Data about a person also has different values at different points in time. For example, if a person is going through a major life event (e.g., she is pregnant), then this piece of data is particularly valuable because this represents a stage in life when you are likely to change your purchasing habits. It's difficult to quantify what this particular piece of data is worth (and, of course, it's worth more to the first company to learn of it, versus the second, since presumably the first company gets more opportunity to sway you towards shopping with them); however, Charles Duhigg notes in his article that "The company [Target] doesn't break out figures for specific divisions, but between 2002 — when Pole was hired — and 2010, Target's revenues grew from $44 billion to $67 billion." [10].

# 5. PROBLEMS WITH DATA

There are several issues wrapped up in the collection of data, particularly when it comes to individuals, where data is subject to change and is easily incorrectly recorded. Some of these issues include:

**Data Correctness**

1. Data accuracy: The data collected on individuals (or inferred about individuals) is often incorrect, and without knowing that this data is even being collected, let alone who is maintaining the database, users are not able to correct data about themselves. Some data aggregators are providing mechanisms for users to investigate and correct their own data (see, for example, Acxiom's About the Data[8]; however, most consumers do not even know about data aggregators in general, let alone about Acxiom specifically, and are even more unlikely to know that they can view and correct their data with Acxiom. Further, Acxiom is only one of thousands of data aggregators (current estimate being approximately 3500 [9]). It is highly unlikely that all 3500 data aggregators maintain accurate data on each individual.[9]

[5] `http://strategiclists.com`

[6] `http://www.leadsplease.com/mailing-list-prices`

[7] Since 1 in 8 women get breast cancer, and there are 316 million people in the US, 50.8% of whom are women. Thus approximately 20 million women have the marker, and given a market of $1 billion annually, this becomes $50 annually / woman.

[8] `http://www.aboutthedata.com`

[9] As an aside, one of the authors tried using aboutthedata.com, and found that what data was there was wildly incorrect, such as education and income. The author's occupation was listed as "white collar / clerical".

2. Timeliness of data: Even in cases where the data collected about an individual is accurate at the time of collection, there is often no tie between the date collected and the data, and no time-out on this data. As a result, data that is no longer correct may continue to exist and propagate about an individual, with potentially negative repercussions.

3. Data inference: The data being collected and stored is not necessarily directly provided by the individual or gathered from databases where an individual has provided data. Rather, data is now being inferred about individuals based on other data (e.g., income can be inferred based on on your address and if you are renting or owning). There is, of course, varying levels of accuracy achieved by inferring data, and as such, this data inference can result in wildly incorrect data being propagated about individuals.

**Use of the Data**

4. Ownership of data: We argue in this paper that individuals should own, and therefore control access to, their own data. The current environment, however, does not define data ownership sufficiently, and the data owner tends to be the organization who "holds" (collects, stores) the data, rather than who the data is about. This can lead to contention when individuals learn how data about them is being used, often without their consent or knowledge.

5. Control of data: Tied closely to data ownership is the control of data. Data aggregators (again, because data is their income) understand that they (legally) own the data they collect, and put in appropriate controls on this data (both technical where possible and legal where necessary) to insure that it is not distributed without appropriate authorization (and likely remuneration).

6. Use of data: There are laws governing that data can only be used for the purpose for which it was collected; however, these laws do not apply to all organizations, all types of data, or in all circumstances. As a result, some organizations have developed approaches to collecting public data (and purchasing other data), which they then combine in order to infer new data. Given that this new data was not collected from an individual, it is not subject to the same laws regarding its use needing to be approved by the individual that the data describes.

7. Ethical issues: Several ethical issues arise surrounding who owns what data, and how much control they should have over that data, in addition to what uses that data can by put. For example, is it ethical to sell lists of people who suffer from HIV/AIDS? Without a solid ethical underpinning, our legal framework for protecting individuals will continue to weaken.

8. Confirmation bias: Given a large collection of data, it is possible to prove anything you want by selectively ignorning any data that does not confirm your hypothesis. That this issue requires consideration can be underscored by experiences such as that of Brandon Mayfield, who was assumed to be the Spanish train bomber by the FBI, who used circumstantial data (a web search on vacations in Spain was performed — let's ignore that it was by his daughter for a high school class) while ignoring other data (Brandon Mayfield had not had a valid passport for ten years — which obviously means that he is using a fake passport) in order to support their (over-zealous) hypothesis [16]. Note that this does not address the sister issue of false positives.

**User Concerns**

9. Protection of data: Data that has legal protections (e.g., such as medical/health data governed by HIPAA) often also has appropriate protections in place; however, data that is outside of such regulations does not need to adhere to the same level of protection and audit controls. As such, there exists the potential for data to be insufficiently protected. (Having said this, given that data aggregators earn income via their data, they likely have sufficient protection on their investment; however, other organizations that maintain data for some secondary purpose are less likely to invest in appropriate controls.)

10. User acceptance: The user acceptance regarding how and what data gets collected, by whom, and how it is used, is complex. For example, users seem to accept that Facebook collects data and uses it to sell targeted advertising space (e.g., how many "single" or "undefined" men in the room receive ads for dating sites?), and do not seem to consider beyond this who might have access to the data they post and for what purpose. Users are, however, suspicious of new technologies with more obvious data collection capabilities. For example, Google Glass, with its ability to record video, is not widely accepted, even though someone with a cell phone can record the same video (albeit more obviously). This is a complex issue that can vary dramatically between individuals (even those within the same demographic clusters), and which has not been sufficiently studied nor addressed.

11. Analytics / scoring: Related to user acceptance and data inference is the analytics applied across data for clustering purposes and to generate consumer scores for people across various categories. "The quality and relevance of the data used, the transparency of the methodology, and the reasonableness of the application are the major factors that determine the fairness of any scoring activity." [9] Much of the "fairness" of such systems has not been addressed, either within the legal context or publicly.

# 6. SO YOU SAY YOU WANT A REVOLUTION?[8]

Much about the current system is entrenched, such as the working (legal) definitions behind privacy (which are based on a notion that predates the current data proliferation problem) and how companies and governments can

[8]With apologies to the Beatles

currently access and use data for their own purposes. While the government has tried to put some restrictions in place to control data access, it has been relatively little and very late. The result is that there now exist massive databases containing large amounts of personal data that can be used for marketing purposes, and organizations such as Facebook, Google and Acxiom can continue to collect data on people and sell it for advertising. To combat this at this stage requires nothing short of a revolution in three different areas: the very definition of privacy itself, technical solutions and capabilities, and the economic model underpinning data usage.

## 6.1 New Definition of Privacy

The current legal system developed around privacy is based on a definition from 1890, long before data aggregation became common and easy, and long before data analytics were employed to determine what products should be marketed to you. Thus the "right to be left alone" does not include the right to not have data collected about you! We need a new definition of privacy that recognizes the current environment, and yet is still general enough to adapt to changes in that environment. Towards that end, we propose a definition of privacy based on the right to control data about you.

This idea is not new. For example, McFarland stated that privacy means "shielding one's personal life from unwanted scrutiny" [23]. And in fact, of the four types of invasions of privacy recognized by the Restatement Second of Torts (intrusion into seclusion, appropriation of name or likeness, unreasonable publicity, false light), three are about data *about* you, and how this data can be used. Thus there is some basis in American law for making such a change.

The first reference to changing the definition of privacy in this manner that the authors can find comes from James Moor, who in 1989 coined the term "informational privacy", which he defined as "the right to control of access to personal information" [24].

Such a change in definition is necessary, as it will inform the legal context within which data aggregators will need to operate. For example, consumer scores are currently created and used to determine to what deals a consumer might gain access [9]. These scores do not use financial data that is regulated, and so are not subject to the Fair Credit Reporting Act. Further, the data aggregators argue that the data they use (in addition to the weighting and underlying algorithms) are proprietary, and so consumers cannot even find out what data is being used about them to generate these scores. But, by expressly stating that an individual has a right to control access to their personal data, it forces data companies to divulge what data they are collecting, and it also provides consumers with the right to either allow this collection or "opt out". The end result is a change in the economic model where data aggregators can provide some micropayment back to the consumer whenever their data is accessed (e.g., if it currently costs 6¢ per record, then perhaps the consumer is entitled to 1¢ each time his record is sold). This allows data aggregators a model in which they can still operate and make a profit, while also allowing individuals to make a choice about sharing their data.

We do provide a word of caution, however, when it comes to employing regulatory and legal frameworks in order to protect individual privacy. It is possible for laws to be written that have not been thoroughly considered in terms of possible adverse consequences, or where the technical solutions are difficult or infeasible. The unintended consequences of the Digital Millennium Copyright Act [21] have been discussed at length. For a more recent example, the European Court of Justice has decreed that individuals have a "right to be forgotten", and that they can request that Google and other search engines not link to specific information. The result has been that Google has had to respond to over 100,000 requests within the first few months of the law passing — putting a company, Google, in the position of determining what information should effectively be censored. (See [31] for a good discussion on this law and its consequences.)

## 6.2 Technical Solutions

Technical solutions need to be developed that can help individuals manage their privacy. Unfortunately, there is no one uber-app that can do everything needed in order to protect individual privacy. Instead, there are several interrelated, promising areas, each of which help to provide a piece of the privacy puzzle. We focus on four key areas here.

First, users often have a need for anonymity. While it is possible to go to a physical store and pay for an item with cash, thus generating no long term linkage of the sale to you (unless, of course, the store has cameras and maintains the video, or does any additional processing of the video). But this has the disadvantage of limiting the goods to which an individual has access. But to purchase an item online means that there is a long-term record of that purchase, both with the merchant, and with the account used to pay the merchant. This does not include any other data that might be collected (such as IP linkages at the ISP level, or any tracking done by companies such as Google — assuming that the merchant was originally discovered via a search — using cookies). Some of this data collection can be mitigated through the use of technologies such as TOR[11], which implements Onion Routing [29], but this does not mitigate data collection at the end points. Additional mechanisms for anonymity need to be developed. The authors recognize that this will raise concerns from law enforcement and certain government agencies, and acknowledge that a discussion will need to be had to determine the appropriate balance between privacy and anonymity of the populace versus the needs for safety and security as achieved through surveillance. The authors posit that much of the backlash against the current government is because policies were put into place without first having the discussion.

Second, technology is needed that allows a user greater control over their data. Specifically, much data has a time limit, after which it should no longer be available, and yet it stays in unknown databases potentially forever. A mechanism that allows data to self-destruct after some period of time is required. A start on research in this area has been performed, for example, by Geambasu *et al.*, who developed a system called Vanish [12].

Third, technology needs to be developed that allows end users to "jam" signals when data is being collected about them without their permission. An example of such a system would be an approach to preventing Google Glass from recording your image (e.g., by your mobile phone giving off

[11] `http:\\www.tor.com`

some signal that results in your image being blurred and indistinguishable in any recorded video). This technology is reminiscent of technology used in "A Scanner Darkly" [8], where the police wear uniforms that show a constantly shifting facade so that they cannot be identified outside of their uniform. This is a form of Private Information Retrieval (PIR), and the authors acknowledge that the cryptography community is doing research into how images (e.g., from a security camera) might be encrypted, and yet a search could be performed on the encrypted data given some image (e.g., the picture of a face that one suspects was captured by the security camera).

Lastly, and perhaps most importantly, technology needs to be developed that allows consumers and other end users to track their data. It is of little use to have a legal framework where users can control access to their data if users do not know where their data actually resides. Users need to know what data about them is being stored where, and when it is accessed, and by whom, and for what purpose. Additionally they need the ability to approve or deny access, or perhaps to set some price for access to their data. At this point, it seems too late for technological interventions, but should policies change then technology will be needed to implement the new policies and legal frameworks.

## 6.3 New Economic (Data) Model

Through all the discussions in this paper it is clear that data has a value. Currently the value of data tends to reside with the holder or container of the data. An electricity company has some personal data, some consumption data and some general power generation/usage data. The utility companies have been selling name and address details as mailing lists for years. The amount of data that is now being sold is enormous, growing and clearly has a value to the purchaser, with the holder of the data getting the economic value. Whole IT ecosystems are supported in this way e.g. Google, Amazon, Facebook, etc.

In the wider world more and more tasks are being automated by increasingly intelligent machines. This is leading to a change in the value of employees, many of whom are seen to be surplus to requirements. In a September 2013 report, Frey and Osbourne examined how susceptible jobs are to computerisation [11]. They claim that 47% of total US employment is at high risk of being computerised within the next decade or two. There are other supporting arguments and anecdotal evidence of automation-related job losses. Who pays to keep the unemployed fed and housed? At the moment there are large numbers of employees paying small to large amounts of tax to governments who disburse benefits to the unemployed. If the figures in the paper are correct there may be as much as 50% of working age people unemployed. The Socioeconomic imperatives will demand some solution. Even today it is difficult to get proportionate tax from large corporations who use international boundaries to move their tax liability to low taxation areas. Unless there is a solution we may be at the start of a dystopian future that is much loved by the SCiFi fans (e.g., watch Bladerunner [7]).

A partial solution is to become a personal data aggregator. Gathering data that is held at various locations (GMail, Facebook posts, tweets, etc.) gives a more complete picture of you as an entity of interest. While you may own your data on Facebook you have given Facebook unlimited permission to exploit that data. Third parties would love to use the Facebook data but the Facebook terms and conditions are such that third parties have no access to your Facebook data. As an owner of the data, copying it from Facebook to your local machine is allowed. Once the data is on the local machine you can make that data available to third parties, for a price. Indeed you can add value in making available personal but anonymised data from twitter and medical records. The totality of data that you have access to as the owner, once you have the data under your own control, is mouth watering to any data analysis team. Being a data aggregator who can aggregate from more than one source you have a very valuable asset, even if your commercial activity on the web is negligible.

There is one source of revenue for individuals that may provide income for the unemployed and relieve government and taxpayers of some of the headaches to come. Jaren Lanier has coined the phrase "Siren Server" to describe organisations that are data aggregators. These siren servers have made massive amounts of money by selling individuals data to third parties (such as Acxiom, mentioned in Section 4.2). He suggests that individuals should be paid proportionately to the value of their data to the siren server. Micro payments from many sales may provide an income based on your value as a consumer. Even the unemployed need to buy power, food and clothes. Employment will change for some. Today new paid roles generated by social media, such as reputation protection and restoration are being created. In the future new roles like avatar stylist for virtual worlds, or gaming consultant or, more interestingly, online value discovery may emerge but only when data is more the property of the individual than the mere holder of the data. It would be great to see a new role of data value agent, a person who promotes your data to organisations who have a need to market to people like you. Selling your data might be the only way of generating an income to supplement the minuscule unemployment benefit that the government may pay. Of course the more net worth and the more data the more wealth that can be transferred to the owner. We are unsure if this will polarise into high class data sellers and low class data sellers, but it will be interesting to watch developments.

One concern with generating income from data is that the worth of data (as discussed in Section 4) would not be sufficient for someone to use as an income. We note that we would expect that the data economy will change as user control over data changes. That is, while it might cost only 6¢ currently for a user record, this is because a data aggregator can sell records in bulk without user permission. If the underlying model changes such that a user can state if they are willing to allow a company to have specific data about them (such as their purchasing habits, or if they are a woman who owns a dog, lives near a certain zip code, has a household income greater than $70,000 and who has an interest in organic cooking), then the user has control over the price for their data. This is likely to increase the cost for a single record, where supply and demand will end up ultimately determining the going rate for specific data.

An additional concern about the commercial use of personal data is the joint ownership of data. Joint copyright of an article or a book is generally handled in the publishing contract or the transfer of rights forms as part of publication but the joint ownership of a patent has potential for disaster. [14] Joint ownership of data could fall into the

same common law issues. If a utility bill is in joint names can the data be defined as personal property owned jointly? If jointly owned data is regarded as personal property the rights to sell or use that data can be exercised by any of the property owners without the permission of the other owners or any accounting to those other owners. We have established that data may fulfil the requirements of a currency, but we have yet to decide if the data is personal property and falls into the joint ownership trap.

Although this is a somewhat depressing vision of the future it is not all bad. Some roles will remain with humans for the foreseeable future. Anything that requires manual dexterity will be relatively safe. Brynjolfssen and McAfee, in their book Second Machine Age, have described a robot that can fold a towel, but takes 24 minutes [3]. So towel folding will be safe employment, but maybe not the career choice of every graduate.

# 7. CONCLUSIONS

Data ownership and data privacy are two interlinked aspects of an individual's identity. People used to laugh at the naive view that having your photograph taken stole a person's soul. It may not be too long before the majority of the world wakes up to the very real possibility that an individual has few if any rights to online data.

The legal position is fragmented and partial. The Australian legal community is debating a new proposal that would make it an offence to record private conversations and activities without a person's consent. The EU has data protection and privacy directives that are generally implemented across member states and the US has a mix of complementary or conflicting state and federal legislation that will affect this debate.

A major concern about the status quo is the lack of visibility of the use of personal data. The dinner time phone call from cold calling salesmen and marketeers who are "not trying to sell you anything" is the result of an unknown harvesting of your data. Individuals should not be relaxed that a purchase of a Statin for medical reasons can result in their data being sold by a medical company to a holiday company selling stress relieving vacations for people with heart problems. Greater transparency of the use of data is important. As we have established earlier data has a value and one of the recipients of the value should be the individual that the data references. Realising that value may be difficult but there are considerations that can help. We would recommend that individuals observe the **Three Laws of Data Value**:

1. The value of your data decreases with the amount you share it. Even Facebook admits this.
2. The value of your data increases with it's uniqueness.
3. The value of your data decreases over time.

Observing these three laws will help an individual maximise the value of their data but that is of little use unless the ownership of the data is well established. In addition to the three laws above there are another three laws that would ensure personal ownership of personal data. Legislators can take these as a guideline to establish more coherent legislation. We define the **Three Laws of Data Ownership**:

1. An individual's data is an asset and they have automatic rights of ownership
2. Other entities can use an individual's data assets with permission, however permission may be withheld if the data is not accurate and timely
3. Governments can use an individual's data assets after notification, however a user can prevent the use of data that is inaccurate or out of date.

Governments and individuals need to become more active to ensure that data ownership is retained by the individual, giving them the rights to manage and sell their data wherever it is stored. The current situation leads to apathy on the part of the individual who has few tools to hand to ensure that data that is stored about them is accurate and not being used without explicit permission. There is complacence on the part of governments and data storage and collection organisations that there is no real challenge to their data supremacy. All parties need to have a better understanding of what data is being stored and for what reasons it is being used. Transparency that enables a user to discover and validate the data that is being stored should be matched by transparency to the individual regarding the use of that data.

The real revolution will come when a new economic model is developed that enables individuals to generate income from the data they own and their activities with the data, reducing the dependence on paid work. This would revolutionise the world and go some way to developing the utopian dream where people do not worry about money and no-one is truly poor.

## Bibliography

# 8. REFERENCES


[1] bbc.co.uk. The domesday book. `http://www.bbc.co.uk/history/british/normans/doomsday_01.shtml`. Last Visited: April 16, 2014.

[2] Karl D. Belgum. Who leads at half-time? three conflicting visions of internet privacy policy. `http://www.richmond.edu/jolt/v6i1/belgum.html`, 1999. Last Visited: April 9, 2014.

[3] Brynjolfssen and McAfee. *Second Machine Age*. IEEE Computer Society Press, 1996.

[4] Cornell University Law School Legal Information Institute (LII). Personal autonomy. `http://www.law.cornell.edu/wex/personal_Autonomy`. Last Visited: March 16, 2014.

[5] Cornell University Law School Legal Information Institute (LII). Right of privacy: An overview. `http://www.law.cornell.edu/wex/privacy`. Last Visited: March 16, 2014.

[6] Dell SecureWorks. The underground hacking economy is alive and well. `http://www.secureworks.com/resources/blog/the-underground-hacking-economy-is-alive-and-well/`, December 2013. Last Visited: March 16, 2014.

[7] Philip K Dick. *Do Androids Dream of Electric Sheep*. Doubleday, 1968.

[8] Philip K Dick. *A Scanner Darkly*. Doubleday, 1977.

[9] Pam Dixon. What information do data brokers have on consumers, and how do they use it?

http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=e290bd4e-66e4-42ad-94c5-fcd4f9987781, December 2013. Last Visited: April 2, 2014.

[10] Charles Duhigg. How companies learn your secrets. *New York Times Magazine*, February 2012.

[11] Carl Benedikt Frey and Michael A. Osbourne. The future of employment: How susceptible are jobs to computerisation? http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf, September 2013. Last Visited: April 16, 2014.

[12] Roxana Geambasu, Tadayoshi Kohno, Amit A. Levy, and Henry M. Levy. Vanish: increasing data privacy with self-destructing data. In *Proceedings of The 18th Usenix Security Symposium*, Berkeley, CA, 2009. Usenix Association.

[13] Government Accountability Office. Privacy: Alternatives exist for enhancing protection of personally identifiable information. http://www.gao.gov/new.items/d08536.pdf, May 2008. GAO Report 08-536.

[14] Wendell Ray Guffey. Joint ownership of patents: A trap for the unwary. http://www.intelproplaw.com/Articles/cgi/download.cgi?v=1091393695, 2004.

[15] Harvard Law School. Restatement of the law, second, torts, s 652. http://cyber.law.harvard.edu/privacy/Privacy_R2d_Torts_Sections.htm, 1977. Last Visited: March 16, 2014.

[16] Matthew Harwood. The terrifying surveillance case of Brandon Mayfield. http://america.aljazeera.com/opinions/2014/2/the-terrifying-surveillancecaseofbrandonmayfield.html, February 2014. Last Visited: April 13, 2014.

[17] Javelin Strategy & Research. 2012 identity fraud survey report, February 2012.

[18] Andreas Kuehn and Milton Mueller. Profiling the profilers: Deep packet inspection and behavioral advertising in Europe and the United States, September 2012. http://dx.doi.org/10.2139/ssrn.2014181.

[19] Michelle Kwasny, Kelly Caine, Wendy A. Rogers, and Arthur D. Fisk. Privacy and technology: folk definitions and perspectives. In *Proceedings of the CHI '08 Extended Abstracts on Human Factors in Computing Systems*, pages 3291–3296. ACM, 2008.

[20] Adam Liptak. Justices, 9–0, bar patenting human genes. http://www.nytimes.com/2013/06/14/us/supreme-court-rules-human-genes-may-not-be-patented.html?_r=0, June 2013. Last Visited: April 13, 2014.

[21] Jacqueline Lipton. The law of unintended consequences: The digital millennium copyright act and interoperability. *Washington and Lew Law Review*, 62(2), 2005. http://scholarlycommons.law.wlu.edu/wlulr/vol62/iss2/3.

[22] Karen McCullagh. Protecting 'privacy' through control of "personal" data collection: a flawed approach. *International Review of Law, Computers and Technology*, pages 47–58, January 2008.

[23] Michael McFarland. What is privacy? http://www.scu.edu/ethics/practicing/focusareas/technology/internet/privacy/what-is-privacy.html. Last Visited: April 13, 2014.

[24] James H. Moor. *How to Invade and Protect Privacy with Computers*, pages 57–70. Westview Press, Boulder, CO, 1989.

[25] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, May 2008.

[26] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 53(6), June 2010.

[27] Matt Pearce. Dad gets officemax mail addressed 'daughter killed in car crash'. http://articles.latimes.com/2014/jan/19/nation/la-na-nn-officemax-mail-20140119, January 2014. Last Visited: March 31, 2014.

[28] William L. Prosser. Privacy. *California Law Review*, 48:338–423, 1960.

[29] Michael Reed, Paul Syverson, and David Goldschlag. Hiding routing information. In *Proceedings of The First International Workshop in Information Hiding*, pages 137–150. Springer-Verlag, 1996. LLNCS 1174.

[30] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

[31] Jeffrey Toobin. The solace of oblivion. *The New Yorker*, September 2014. http://www.newyorker.com/magazine/2014/09/29/solace-oblivion.

[32] Samuel Warren and Louis Brandeis. The right to privacy. *Harvard Law Review*, 4(5), 1890.

[33] Chaim Zins. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4):479–493, 2007.