Data Science at the Intersection of Security and Privacy

University of California-Berkeley

MIDS Program

W231 Final Project

Spring 2016

Alfred Arsenault

Tarun Chopra

Executive Summary

Balancing security and privacy requirements in complex systems collecting and processing massive amounts of citizen data can be very difficult for Government organizations. The problem is exacerbated by the availability of powerful Data Science techniques such statistical analysis, predictive analytics, and machine learning.

The “Open Government” initiative to make datasets public by default heightens the risks to personal information. Sharing of Government data has many advantages: it promotes Government transparency; it puts data paid for by the taxpayers into the hands of those taxpayers; and it opens the possibility of new uses for data that were not conceived of when the data were collected. However, it changes the calculus of determining risk when releasing data. The release of Government data does not happen in a vacuum. A released dataset lives alongside all other datasets that have been created or released publicly, and it will live alongside all of those that are yet to be released. It is possible that release of a dataset does not harm privacy today, but will in two or three years when other datasets are also available that can be correlated with the one released today.

In this project, the authors studied the Government’s processes for determining and balancing cyber security and privacy requirements. We built on previous work to show where security and privacy are synergistic, and where they conflict.

We then show how the ability to collect, store, and quickly process massive amounts of data impact the situation. Data science capabilities enhance some facets of security, such as providing accountability and identifying malware; while harming others, such as the ability to protect data from direct or indirect disclosure. Similarly, data science capabilities support some facets of privacy, while harming others.

The authors close by providing a set of recommendations to Government organizations about releasing datasets, and a set of recommendations to Data Scientists involved in the process.

Table of Contents

1. Introduction

1.1 Problem statement

The improved data analysis capabilities brought on by the maturing of the data science field have wreaked havoc on the ability of an organization to properly balance its security and privacy requirements. Security and privacy strategies that were adequate to protect an organization, its users and its data subjects only a few years ago are no longer sufficient. The distribution of open data sets, the wide sharing of data with third parties, and the increasingly distributed nature of datasets makes it difficult to determine how much security and privacy is sufficient. [RENCI]

Federal and State Government agencies invest enormous amounts of time and money in the collection of data, and face pressures to share that data widely while maintaining the security of that data and the privacy of the subjects of that data. This is particularly true when security and privacy must be maintained for an indefinite time in the future, when new tools and capabilities will undoubtedly exist. If a dataset were to be released today with all users anonymized, the de-anonymization of subjects and disclosure of personal data a year from now is just as unacceptable and potentially disastrous as the direct disclosure of the information today.

1.2 Purpose

The purpose of this research project was to study the state of the art in privacy, security and data science, and develop a set of recommendations that will help Government organizations and data scientists provide appropriate security and privacy protection to sensitive data.

1.3 Background and Context

The study focused on US Government organizations, primarily in the civilian sector. These Agencies collect enormous amounts of data about nearly everything on the planet. Under the Open Government Initiative[1], the default is for all that data to be shared with the public. While much of the data can be freely shared without harm, some data requires protection. Some of the

[1] See https://www.whitehouse.gov/open

data collected contains Personally Identifiable Information (PII) about citizens and other residents of the United States, and must be protected by law. Other data can be disclosed without harm, but must not be changed without detection, so that data users know they are dealing with the correct dataset.

Since Government organizations have statutory requirements to protect the privacy and security of their data, and since they are broadly encouraged to share all data by default, they were selected as the primary subject of the study.

1.5 Structure of the paper

Section 1 of this paper provides the Introduction, the problem statement, the problem context, and some background information on the problem. Section 2 discusses Security, Privacy and Data Science, as used in this research project. Section 3 illustrates how security and privacy requirements are established by Government organizations, and identifies some of the synergies and conflicts between security and privacy requirements. Section 4 Identifies synergies and conflicts between Data Science and Security and Privacy. Section 4 then goes on to provide 5 detailed examples of the synergies and conflicts. Section 5 provides our conclusions; the authors' recommendations to the Government organization and to the Data Scientist responsible for Security and Privacy.

There are three Appendices to this report. The first identifies the references cited by the authors during this research. The second goes into some detail about the privacy controls identified by the National Institute of Standards and Technology in Appendix J of Special Publication 800-53. The final appendix describes how our recommendations would be impacted by adopting a different approach to privacy; e.g., Nissenbaum's *Contextual Integrity* and Solove's *Privacy Taxonomy*.

2. Security, Privacy and Data Science

In this section we define the terms “security,” “privacy” and “data science” and explain how we use them in this project. Each of the terms has many different definitions, and different researchers choose different definitions to suit their purposes. We selected definitions that we believe are appropriate for the Government context in which our analysis is done.

2.1 Security

Security, or more specifically ***cybersecurity***, is the protection from damage or harm of systems and their resources, e.g., their datasets. Security has four key facets:

> ***Confidentiality*** means that data are disclosed to anyone except for those authorized to see the data
>
> ***Integrity*** means that data are not modified or deleted in an unauthorized way without such modification or deletion being detected.
>
> ***Availability*** means that the system, software, and data can be accessed when needed and used for their intended purposes
>
> ***Accountability*** means that all events that occur in the system - e.g., accesses to or modifications of data - can be traced back to the unique entity that caused the event to occur.

2.2 Privacy

Privacy is an ambiguous term that means different things in different contexts. One of the earliest widely-used definitions was from Westin:

> “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.” [WESTIN]

In recent times other definitions have been proposed, and in fact there are those who assert that it is pointless to establish a definition for the term ***privacy***, because there are no core characteristics that represent privacy in all contexts. [NISSEN], [SOLOVE]

In our research, we decided to focus on the major elements of privacy that would apply to users or Government systems, or to subjects whose data would be contained in Government systems. [2] These elements include: :

> ***Anonymity*** - the ability of a subject to be unknown to those who view the dataset. Not only is it impossible to identify the real subject, but actions of the same real subject cannot be correlated through analysis of the data.
>
> ***Pseudonymity*** - the ability of a subject to have her true identity hidden from those who view the subject, while also having her actions correlated with one another. This allows the subject to establish a reputation in the system while also keeping the "real name" secret. Examples of pseudonyms would include "Publius", the author of the Federalist Papers; and "Satoshi Nakamoto," the creator of Bitcoin and the Blockchain.
>
> ***Fair Use*** - personal data are only collected when needed; collected with the consent of the subject; protected while in custody; and used only for the stated purposes.
>
> ***Access*** - the subject of the data can get access to the data to determine its accuracy and correctness. Further, no one can get access to data which reveals personal information about a subject unless that person is explicitly authorized for such access, and/or the subject has approved the access.
>
> ***Control of Lifecycle*** - the data are controlled throughout the lifecycle from collection to destruction/disposal, so that no inadvertent privacy violations occur after the data have left established controls
>
> ***Control of Aggregation*** - the data are only shared with third parties or other organizations in accordance with the established policies, and/or with the knowledge and consent of the subject. This limits privacy violations due to de-anonymization of users because of the aggregation of information from different datasets.

[2] See Appendix B of this report for a discussion of changes to our model that would occur if we used a different approach; e.g., Solove's taxononmy.

2.3 Data Science

Data Science is a new and evolving interdisciplinary field that addresses the collection and analysis of very large set of heterogeneous data sets. Data Science vastly different from statistics where the later is focused on validating predefined hypothesis through systematic study of the organization, properties, and analysis of data. Data Science as a discipline is different from statistics and other existing fields in key areas of unstructured data, leveraging unlimited compute power for sophisticated data analysis, knowledge discovery through pattern matching and finally providing future predictions through machine learning. Data Science implies multitude of various skills in order to be proficient in quick meaningful analysis of the large amounts of data that is available to man-kind. Key subfields that are important to the research project are described in more details.

Statistical analysis - the use of statistical analysis, and statistical software packages, to draw conclusions about numerical properties related to a dataset; e.g., the mean and median of a dataset; the coefficients in a linear regression; etc.

Data Analytics: deals with examining raw data both structured and unstructured in order to draw meaningful conclusions. One key difference between data and statistical analysis is that in later, data analysis is usually done through statistical method in order to either accept or reject null hypothesis. In data analytics, raw compute power along with software tools are used to identify meaningful patterns without any preconceived assumptions.

Predictive Analytics: is the science of analyzing historical and current facts in order to make predictions about future and unseen situations. More precisely, predictive analytics is finding patterns in the sea of data and predict value in future using the identified pattern.

Data Visualization: is critical in visually grasping difficult concepts or identify new patterns. Data visualization leveraged statistical graphics, plots and information graphics in order to communicate key points to the audience. Key to data visualization is to make complex data both structured/unstructured including images. video easily accessible, understandable and usable.

Machine Learning: is what separate data science from other existing branches like statistics. In the broadest of terms machine learning refers to system learning from previous data and then predicting to unseen examples. The emphasis on prediction is especially strong in machine learning even to the extent learned model with predictive models are more valuable than vice versa.

3. Establishing Security and Privacy Requirements

3.1 Requirement Process

For Federal Government organizations, laws such as the Federal Information Security Management Act (FISMA) specify the process by which cyber security requirements are to be defined. The process is based on a number of documents published by the National Institute of Standards and Technology (NIST). Agencies must look at FIPS Pub 200 [F200] to understand impact levels for information systems, and the security controls that must be applied. They then use FIPS Pub 199 [F199] to determine whether their data and system have a low, medium or high impact upon the United States and its citizens. Once that determination is made, Federal organizations then consult NIST Special Publication 800-53[SP853] to determine what cyber security controls should apply to their system and data. The security controls contained in [SP853] range from personnel and physical security requirements to processes that must be followed to technical capabilities that must be built into the system. These controls can be tailored as appropriate for the given system and data being processed - e.g., a stronger cryptographic algorithm or authentication mechanism can be used to provided additional security; or a control may not be applicable to a specific system because it will not be connected to the Internet.

The process for specifying privacy controls is much less formal. There are laws requiring the protection of Personally-Identifiable Information (PII), and guides such as NIST’s Special Publication 800-122 [SP8122]. However, the actual process for specifying privacy controls centers around Appendix J of Special Publication 800-53[3]. Each Federal organization is required to have either a Chief Privacy Officer (CPO) or a Senior Agency Official for Privacy (SAOP). The audience for Appendix J is the CPO/SAOP. Whenever a system is being designed, implemented, or given approval to operate, the CPO/SAOP must take the list of controls and work with the program manager responsible for the system and data. Together, they will develop a set of controls to be implemented, along with the implementation details. The CPO/SAOP can then ensure that the privacy of US citizens is adequately protected.

[3] See Appendix B for a discussion of the privacy controls contained in Appendix J of [SP853]

3.2 The intersection of security and privacy

When a system is being built, or data are being added to the system, both security and privacy requirements have to be considered. But there is an issue, which is that security and privacy don't always work together. Sometimes security and privacy facets supplement each other, but sometimes they conflict. As an example, consider the privacy facet "anonymity" and the security facet "accountability."

Anonymity means that a subject can interact with the system and there will be no way of knowing who this subject is. This is often recommended as a way of protecting PII; the relevant data such as incidence of a disease is recorded without the association of that data with a person.

Accountability means that every action on a system can be uniquely traced to the subject who caused that action to occur, even if the cause was indirect.

Obviously, no system can provide both true anonymity and true accountability. One of these facets must be sacrificed if the other is to be achieved. Thus, Federal Government system and data owners have to make design tradeoffs to identify which requirements are more important and what other controls may be needed.

In [CONG}, Conger and Lowry discussed "the intersection of security and privacy" and provided a set of tables showing which facets of security and privacy (as defined by them) were complementary, which conflicted, and which had no significant impact. In our study we added additional facets for both security and privacy, but we have taken the same approach. Thus, Table 1 summarizes how the security and privacy facets we discussed in Section 2 of this report interact with one another. The ratings given in each cell represent the opinions of the authors of this study based on our research.

Privacy -> Security	**Anonymity**	**Pseudonymity**	**Controlled lifecycle**	**Fair Use**	**Access**	**Control of Aggregation**
Confidentiality	Support	Support	Support	Support	Support	Harm
Integrity	Neutral	Neutral	Neutral	Support	Support	Harm
Availability	Neutral	Neutral	Harm	Harm	Harm	Support
Accountability	Prevent	Harm	Support	Support	Support	Harm

Table 1: The Relationship Between Security and Privacy

Note that in Table 1, a rating of “neutral” means that the data science field both supports and harms the implementation of the privacy area, but overall the impact is reasonably balanced.

Anonymity and Accountability “Prevent” each other from occurring. As explained above, it is not possible for a single system to provide both true anonymity for users/subjects and true accountability for all actions.

Anonymity and Pseudonymity “Harm” each other. They are not truly mutually exclusive. It is possible to have accountability for all actions while supporting pseudonymity for users/subjects. There will have to records kept off the system mapping the system’s pseudonyms to the actual names of subjects. Pseudonymity is provided because it is possible for users to operate without exposing their real identities on the system; accountability is provided by taking the additional step of manually mapping a pseudonym to a real name when required. The two facets harm each other because the only way to achieve both is to include this off-system manual step, which is time-consuming and expensive, and also puts the real name-pseudonym mapping at risk of disclosure.

By contrast, Anonymity and Confidentiality “Support” each other, because ensuring that data are never disclosed to those not authorized to see it helps prevent disclosure of the real identity of a subject via a de-anonymization attack.

Integrity and Control of Aggregation are shown as having a “Harm” relationship, which may seem counter-intuitive. It helps to recall that Integrity means that no changes or deletions are made to the information without authorization, without such changes or deletions being detected. If the information cannot be changed or deleted by the subject, then the subject has very limited control over the later use of that information in aggregation. E.g., if I am the victim of a crime reported in a law enforcement database, then Integrity means that I cannot later delete the record of the crime of my status, and thus when the database is later shared - even if it has been anonymized - I am at risk of having my victim status discovered when that database is aggregated with other public databases.

4. What’s Data Science Got to do With It?

In Section 3 of this report we described the synergies and contradictions of security and privacy facets. In this section we address data science. In Section 4.1 we provide two more tables, showing the relationship between data science and privacy, and between data science and security, respectively. In Section 4.2 we provide a set of examples that illustrates the results provided in the tables.

4.1 The Intersections of Data Science with Privacy and Security

Table 2 summarizes the relationship between the facets of data science and the facets of privacy. As with Table 1, a rating of “neutral” means that the data science field both supports and harms the implementation of the privacy facet, but overall the impact is reasonably balanced.

	Anonymity	**Pseudonymity**	**Controlled lifecycle**	**Fair Use**	**Access**	**Control of Aggregation**
Statistical Analysis	Harm	Harm	Neutral	Neutral	Neutral	Harm
Machine Learning	Neutral	Neutral	Neutral	Neutral	Neutral	Harm
Data Analytics	Harm	Harm	Neutral	Neutral	Neutral	Harm

Predictive Analytics	Harm	Harm	Neutral	Neutral	Neutral	Harm
Data Visualization	Harm	Harm	Neutral	Neutral	Neutral	Harm

Table 2: The Interaction Between Data Science and Privacy

Statistical Analysis and Anonymity have a “Harm” relationship. Statistical analysis can be used to de-anonymize published datasets with anonymized subjects. For example, suppose that no direct test scores of a single subject will be released, but it is possible to find the mean of any arbitrary set of subjects as long as that set is of sufficient size. Then repeatedly finding the mean of many different sets of subjects can eventually reveal the scores of individuals or of small groups of individuals, effectively de-anonymizing the data set.

On the other hand, if true anonymity can be provided, it is likely that no meaningful statistical analysis can be performed on the dataset. Recall from our definition that anonymity means that transactions involving the same subject cannot be correlated. Thus a truly anonymous dataset would look like a large set of transactions, each involving a separate subject that was only involved in that one transaction. It would be very difficult to draw many meaningful conclusions from analysis of such a dataset.

Machine Learning can either Harm or Support Anonymity; thus it is given a rating of “Neutral.” Machine Learning can be used to help identify malware far better than traditional signature-matching approaches. It can also be used to isolate the true causes of anomalous network traffic. (See Examples 2 and 4, below.) These are synergistic properties. On the other hand, Machine Learning can also be used to help de-anonymize released databases. If subjects that are similar can be identified in the dataset, then learning information about one of those subjects is likely to reveal information about the similar subjects. This is a conflicting relationship. So we assign an overall rating of “Neutral.”

Table 3 summarizes the relationship between the facets of data science and the facets of security. As with Table 1 and Table 2, a rating of “neutral” means that the data science field both supports and harms the implementation of the security facet, but overall the impact is reasonably balanced.

	Confidentiality	Integrity	Availability	Accountability
Statistical Analysis	Harm	Neutral	Neutral	Support
Machine Learning	Neutral	Neutral	Neutral	Neutral
Data Analytics	Harm	Neutral	Neutral	Support
Predictive Analytics	Harm	Neutral	Neutral	Support
Data Visualizatio n	Neutral	Neutral	Neutral	Neutral

Table 3: The Interaction Between Data Science and Security

The relationship between Statistical Analysis and Confidentiality is similar to that between Statistical Analysis and Anonymity in Table 2, although not quite to the same extent. Statistical analysis can still be carried out to perform inference attacks, even if access to individual data items is denied. Meaningful statistical analysis could possibly be done, but only by subjects authorized to view the entire dataset. Thus, the ability of independent researchers to do statistical analysis after the fact would be harmed.

Just as with Machine Learning and Anonymity, the relationship between Machine Learning and Confidentiality is both supportive and conflicting, resulting in an overall rating of “Neutral.”

4.2 Examples

Tables 2 and 3 in Section 4.1 showed the authors' conclusions about the relationships between facets of Data Science with those of Privacy and Security, respectively. The text following the tables provided brief explanations of selected conclusions. In this section we will provide 5 examples that provide a more detailed explanation and justification of our conclusions.

Example 1: Data Analytics harms confidentiality and anonymity

As with many other jurisdictions, the City and County of Denver, CO, makes crime records available for download[4]. Crime reports include date, time, category, and location. While location is sometimes given in terms of a block or an intersection, crimes that occur at a residence or business are listed as occurring at that specific address.

Denver does a good job of ensuring that this database by itself reveals no PII; e.g., it does not list the victim of a domestic assault. However, that information is often available elsewhere. The Denver Property Tax Database[5] makes publicly available the owner of each property. It also provides a "Chain of Title" dating back to at least 1990, so it is easy to determine the property owner at a given time.

This gives significant insight into the perpetrator and victim if it is an owner-occupied residence, but it will not reveal significant information about rental property. Fortunately, there are additional public databases. For example, the Colorado Voter Registration database, including names, addresses, and records of whether or not the person voted, is available from the Colorado Secretary of State's office.[6]

The bottom line is that there are many publicly available datasets, and there is generally limited coordination done to ensure that release of a new dataset will not cause the de-anonymization of records in previously-released datasets.

[4] See http://data.denvergov.org/dataset/city-and-county-of-denver-crime

[5] See https://www.denvergov.org/Property/

[6] See https://sos.state.co.us/Voter. Note that while the information is publicly available, there is a charge for the data as of the time of this writing.

This provides significant opportunities for data scientists with analytics tools and the skills to use them. The ability to glean information from a single large dataset, and to merge it with information from other, seemingly-unrelated datasets, means that many confidentiality and anonymity policies can be bypassed with relative ease.

Example 2: Machine Learning Helps Identify Root Causes of Anomalous Network Traffic

Large Internet Protocol(IP)-based networks are designed to deal with varying patterns and volumes of traffic flowing. Different amounts of traffic flow between different sites depending on the day of the week and the time of the day. E.g., Monday through Friday, 8 am to 6 pm local time, sees heavy traffic among the sites controlled by a business, while evenings and weekends tend to see heavier traffic to recreational or entertainment sites.

An outage or a misconfiguration can cause traffic to vary from what is normally seen, because IP networks are designed to detect outages or heavy traffic and simply route around it.

Network attacks can also cause anomalous network traffic flow. A Distributed Denial of Service (DDOS) flooding attack can cause heavy traffic flow from many different sites simultaneously. A successful DNS attack, or a BGP rerouting attack can cause a site or network to be starved of traffic.

Determining that a traffic pattern is anomalous, and then identifying the root cause of the anomaly, is an area that is greatly helped by Machine Learning. For example, Microsoft Azure relies on Principal Component Analysis to help classify traffic anomalies and identify security threats[7].

- “PCA converts a set of cases, containing possibly correlated variables, into a set of values called principal components. In the case of anomaly detection, for each new input, the anomaly detector first computes its projection on the eigenvectors, and then computes the normalized reconstruction error. This normalized error is the anomaly score. The higher the error, the more anomalous the instance is.”

[7] See https://msdn.microsoft.com/en-us/library/azure/dn913102.aspx

Example 3: Statistical Analysis Makes Inference Attacks Easier; Harms Anonymity & Confidentiality

An ***inference attack*** is one in which an attacker learns something prohibited by policy using only indirect methods. That is, the attacker is not able to directly access a private value, but instead is able to *infer* that value from related information that is available.

Inference attacks have been known and addressed in the database world for decades. As a simple case, suppose that there is a database containing the test scores of 100 people. Individual test scores are considered private information and will not be released. However, the system will provide statistical measures of groups of people, such as the mean test score and the median test score. A policy is established that Means and Medians will not be provided for groups of less than 10 people, because that should provide sufficient protection against an attacker inferring an individual score.

However, even with this control, it may still be possible to infer scores. Suppose that an attacker is able to successively request the means of different groups of people. First, the attacker gets the mean of subjects 1 through 20. Next, she gets the mean of subjects 21 through 40. Then the mean of subjects 1, 3, 5, 7, ...39. Then 2, 4, 6, 8, ...40.

Each mean represents a linear equation involving the variables that make up the mean; e.g., the mean of S1 through S20 can represent:

S1 + S2 + … S20 = 20*M1

While the mean of 1, 3, 5, 7, ...39 is

S1 + S3 + … + S39 = 20 * M2

Linear algebra can now come into play. If by repeated querying we can build a set of 100 linearly-independent equations in 100 unknowns (the 100 test scores), then we have a solvable problem. We can determine each test score!

Even if we can’t directly apply linear algebra as above, careful repeated querying can provide information about sensitive values. When this inference attack involved a single database contained within a single database server, the attack could be limited by, e.g., restricting the number of related queries that could be asked.

In the Data Science world, however, that is likely not possible. First, the databases are often distributed, making access control more difficult. Second, there are many databases from which partial information can be extracted, allowing the build-up of an inference. And third, the datasets are often released to the public. Once the attacker has the database in her hands, there is no way to control what analyses she performs nor what resources she brings to bear.

The result is that the potential of inference attacks has been made significantly higher by the advent of widely-available, powerful tools and released datasets.

Example 4: Data Science has Entirely Changed Malware Detection; Helps Accountability, Integrity, Confidentiality

Malware is software or firmware that can cause harm to a computer and/or its data. Malware includes attack categories such as viruses, Trojan horses, trap doors, worms, ransomware, etc. Most organizations and people are very concerned about keeping malware off their systems to help protect their data.

From the mid-1980s until fairly recently, most solutions that detected and prevented malware were based on a signature-matching approach. Malware would be detected in the wild or in the lab, and a signature - a unique pattern of code or data carried only by this malware - would be identified. Then the malware detection code would search for that pattern, and block any software containing it.

There were two major drawbacks with that approach. The first was that a race developed between the malware writers and the malware detectors. The malware writers developed increasingly sophisticated techniques to modify their code on the fly or between each attack so that it wouldn’t be detected. Defenders were forced to try to keep up with the new variants, which appeared rapidly.

The second and more significant drawback is that the defenders are always behind. The malware detection side is always *reacting* to the latest threat. Every time a new class of vulnerability was detected, the attackers had a free shot to exploit this *zero-day* vulnerability for whatever value they could, before the defenders caught up.

In the last few years, a number of researchers and companies have abandoned the old signature-matching approach and moved to Machine Learning as a mechanism to try to detect previously unseen malware and defend against zero-day attacks. The general approach is to amass and analyze significant amounts of data, showing both malware and harmless software. Classification techniques are used to identify the properties of malware, and distinguish them from the properties of benign software. These classification techniques can then be applied to new software before it is deployed, to help decide whether the new software shares the characteristics of malware, or the characteristics of benign software.

As an example of this, Grace et alia described RiskRanker[8], a two-level analysis system to determine the risk associated with code. The first stage analysis involves traditional signature detection and static code analysis. The second stage analysis brings in behavioral modeling and dynamic analysis to look for unusual characteristics, such as encrypted native code and unsafe byte code loading.

The availability of tools to analyze massive amounts of data is very useful in assessing the risk of mobile apps. Mobile apps from Google Play and the Apple App Store are software programs that have the potential to be executed on millions of systems. They are inviting targets for attackers and sources of concern for security personnel. While both Google and Apple do some vetting of apps prior to making them available for download, that does not provide a guarantee that the apps are not malicious. A number of companies[9] provide services that collect data in real time from a global population of devices (e.g., iPhones, Android phones), then sift through the data to identify complex correlations that would otherwise not be identified as risks.

[8] See http://www.facweb.iitkgp.ernet.in/~niloy/COURSE/Autumn2014/SmartPhone/Books_Paper/RiskRanker.pdf

[9] See e.g., https://www.lookout.com/docs/Lookout%20Security%20Platform_%20Technology%20Whitepaper.pd

Example 5: Data Analytics Facilitate Aggregation Attacks, Harm Anonymity & Confidentiality

An ***aggregation attack*** occurs when an attacker is able to pull together information from many different sources quickly and efficiently to gain information to a larger set of information than would otherwise be permitted.

In many situations the release of a single data item is considered harmless, while the release of a large dataset containing that item is unacceptable. Consider a company's employee information. Release of a single employee's contact information is harmless. Many people put much their entire employment information in their e-mail signature block - employer name, department, physical address, work and mobile phone numbers, e-mail address, etc.

There are many different datasets widely available today, and a variety of tools to support the efficient scraping of data from the different datasets for later correlation and aggregation. This has exacerbated the problem that existed previously. In our employee contact information, suppose that the attacker begins by getting a list of all employees participating in standards bodies. Large technology companies generally have significant numbers of key people involved in standards bodies such as the Internet Engineering Task Force (IETF), the Institute of Electrical and Electronics Engineers (IEEE), the World-Wide Web Consortium (W3C), etc. It is generally not difficult to scrape data from the standards body's mailing lists/website/attendee lists. This will reveal a number of employee names, and possibly also something about the company structure.

The next phase can involve scraping data from sites like LinkedIn, where people generally list their current employer; and Facebook, where some people do. This will reveal another list of employees, and possibly more information about the structure of the company. Lists of former employees may also be found on these sites, and this can also be helpful in understanding the company.

Public companies, and certain large private companies, as well, have required filings with the Securities and Exchange Commission. They may also have required filings with other regulatory bodies; e.g., environmental impact statements with the EPA. These filings are often made public, and can be scraped for information.

The bottom line is that in many cases, once these datasets and others are scraped for information, the attacker will have a respectable part of the employee database, along with a decent understanding of the company structure and how that has changed over time. The sensitivity of this information will depend on the company and on other circumstances, but it is something that should be protected.

5. Recommendations

5.1 Recommendations for the Organization

Government organizations should thoroughly review all considerations while reviewing and implementing policies around data. For example, release of government data sets through open government raises challenging questions about the extent to which individuals may be identifiable when this data is combined with other available data. Hence, great focus needs to be given on anonymizing data before releasing it to the open world. Some of the necessary actions may include developing guidance on privacy issues being risen through particular datasets and also guidance on when those privacy issues can be overridden by the need for transparency and accountability. Greater focus should be driven towards improved approaches to meta data protection and aggregation of data instead of releasing individual records. Many observers have suggested that deliberately injecting errors into the data in a manner which keeps averages the same but changes individual records might increase privacy. This idea could be combined with digital watermarking technology to simultaneously improve privacy and provide a means to track the source of data, thus helping enforce the code of conduct.

In today's political environment, rhetoric is dominated by calls for transparency and accountability through open government initiatives and pressure is on government to proactively release data. It is very important to not lose sight of privacy, and government institutions before releasing any form of data should comply with privacy laws like Fair Information Practice Privacy Principles (FIPPs) and leverage it as identifying, assessing and mitigating privacy risks related to open data sets. focus should be driven towards improved approaches to meta data protection and aggregation of data instead of releasing individual records. Also, before the data is released, a testing procedure should be in place in order to test technology solutions used to anonymize and de-identify the data and the final results should be documented.

5.2 Recommendation for the Data Scientist

Data Scientist have a key role to play in managing privacy and security while dealing with large amount of sensitive data. “Data” being the 21st century natural resource will put a lot of pressure on data scientists to drive quicker insight, competitive advantage and bring meaningful societal changes to the world. Initiatives like “Open Gov” demonstrates the power of big data where both state and federal government are providing transparency and accountability to its citizen. It is very important for data scientists to not lose trust of the citizens and in turn give up the momentum around Big Data and its power to change the world. Given below are some simple steps that data scientists can proactively take in order to enhance their craft and the overall value to society in large.

a) Understand the distinction between privacy and security and be an advocate for data sharing where it doesn’t harm the privacy and security for the data subjects..

b) Take accountability and responsibility for understanding the data, how it is being protected for privacy and security threats and the impact of releasing data in public. For example, before releasing the data for broader analysis, implement a testing procedure for technical solutions that are being used to anonymize and de-identify the data and a process to store the documented results.

c) As discussed throughout the paper, data science is a multi layered discipline and requires a very unique skill set in order to be successful. Data scientist will deal with all kinds of data structured/unstructured, personal and sensitive in nature in order to perform required analysis. Hence, it is critical for data scientist to keep up with the latest trends in the ever evolving field of data science and tightly align with subject matters experts in the field of privacy and security.

d) In the ever volatile global socio-political system, regulatory and compliance requirements are constantly changing in order to keep up with modern threats. For example, European Union recently approved new data protection laws getting potentially enacted by 2017. The laws are the new step in bolstering European's privacy rights and has huge ramifications on the handling of sensitive data. Data scientist will need to continuously monitor and keep pace with ever changing laws and leverage technology in order to keep ahead.

e) Finally and most importantly, the key trait that will be most useful is the willingness to adapt and reach to changing policies, laws, regulations and security threats.

Appendix A: References

[BRILL} Brill, Julie, “Privacy and Data Security in the Age of Big Data and the Internet of Things,” Statement by U.S. Federal Trade Commissioner Julie Brill Delivered at Washington Governor Jay Inslee’s Cyber Security and Privacy Summit, January 5, 2016

[CONG} Conger, Sue, and Brett J. L. Landry, “The Intersection of Security and Privacy,” The Association for Information Systems, 28 January 2009

[GRACE] Grace, Michael, Yayin Zhou, Qiang Zhang, Shihong Zhou, and Xuxian Jiang, “RiskRanker: Scalable and Accurate Zero-day Android Malware Detection,” in Proc. Mobisys ‘12, Lake District, UK, June 2012. Available at http://www.facweb.iitkgp.ernet.in/~niloy/COURSE/Autumn2014/SmartPhone/Books_Paper/RiskRanker.pdf

[LKOT] Lookout, Inc., “The Lookout Security Platform - Advanced Mobile Threat Protection Through Predictive Cybersecurity,” San Francisco, CA. Available at https://www.lookout.com/docs/Lookout%20Security%20Platform_%20Technology%20Whitepaper.pdf

[SP8122] - National Institute of Standards and Technology, “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),” NIST Special Publication Number 800-122, Gaithersburg, MD, April 2010

[SP853] National Institute of Standards and Technology, “Security and Privacy Controls for Federal Information Systems and Organizations,” NIST Special Publication Number 800-53, Revision 4, Gaithersburg, MD, April 2013

[F199] -National Institute of Standards and Technology, “Standards for Security Categorization of Federal Information and Information Systems,” Federal Information Processing Standards Publication (FIPS Pub) 199, Gaithersburg, MD, February 2004

[F200] -National Institute of Standards and Technology, “Minimum Security Requirements for Federal Information and Information Systems,” Federal Information Processing Standards Publication (FIPS Pub) 200, Gaithersburg, MD 2006

[NISS] Nissenbaum, Helen, “Privacy in Context - Technology, Policy and the Integrity of Social Life”, Stanford Law Books, Stanford, CA, 2010

[SCHMT] Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., & Fecho, K., “Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage,”. RENCI, University of North Carolina at Chapel Hill, 2013. Available at http://dx.doi.org/10.7921/G0WD3XHT

[SOL02] Solove, D. J. (2002). Conceptualizing privacy. *California Law Review, 90*, 1087–1155.

[SOL06] Solove, D. J. (2006). Taxonomy of privacy. *University of Pennsylvania Law Review, 154*(3), 477–564.

[WEST] Westin, Alan F., “Privacy and Freedom,” The Bodley Head Limited, 16 April 1970

Appendix B: NIST SP 800-53 Privacy Controls

This research project addressed the use of the privacy controls listed in Appendix J of NIST's Special Publication 800-53. In this Appendix we provide more detail about those controls. Table B-1 lists the controls as they appear in [SP853].

AP-1 AUTHORITY TO COLLECT	DM-2 DATA RETENTION AND DISPOSAL
AP-2 PURPOSE SPECIFICATION	DM-3 MINIMIZATION OF PII USED IN TESTING, TRAINING, AND RESEARCH
AR-1 GOVERNANCE AND PRIVACY PROGRAM	IP-1 CONSENT
AR-2 PRIVACY IMPACT AND RISK ASSESSMENT	IP-2 INDIVIDUAL ACCESS
AR-3 PRIVACY REQUIREMENTS FOR CONTRACTORS AND SERVICE PROVIDERS	IP-3 REDRESS
AR-4 PRIVACY MONITORING AND AUDITING	IP-4 COMPLAINT MANAGEMENT
AR-5 PRIVACY AWARENESS AND TRAINING	SE-1 INVENTORY OF PERSONALLY IDENTIFIABLE INFORMATION
AR-6 PRIVACY REPORTING	SE-2 PRIVACY INCIDENT RESPONSE
AR-7 PRIVACY-ENHANCED SYSTEM DESIGN AND DEVELOPMENT	TR-1 PRIVACY NOTICE
AR-8 ACCOUNTING OF DISCLOSURES	TR-2 SYSTEM OF RECORDS NOTICES AND PRIVACY ACT STATEMENTS
DI-1 DATA QUALITY	TR-3 DISSEMINATION OF PRIVACY PROGRAM INFORMATION

DI-2 DATA INTEGRITY AND DATA INTEGRITY BOARD	UL-1 INTERNAL USE (Limitation)
DM-1 MINIMIZATION OF PERSONALLY IDENTIFIABLE INFORMATION	UL-2 INFORMATION SHARING WITH THIRD PARTIES

Table B-1: NIST Special Publication 800-53 Privacy Controls

The AP requirements address the Government Agency’s legal authority to collect the information (AP-1), and the statement of purpose that the Agency has written regarding its collection and use of PII (AP-2). Government Agencies are not permitted to collect PII without a legal authority to do so. They must also have declared what use they will make of the PII, and then follow that statement.

The AR family of requirements deal with accountability, audit and risk management. AR-1 requires that the Agency designate a CPO or SAOP as the accountable official for PII, and to have a governance and privacy plan. AR-2 requires that the Agency conduct a privacy impact assessment and risk assessment for each system with access to PII. AR-3 requires that the Agency establish privacy roles, responsibilities, and access requirements for contractors and service providers; and include privacy requirements in contracts and other acquisition-related documents.. AR-4 covers the Agency monitoring and auditing its privacy policies and processes to ensure that they are effective.. AR-5 addresses privacy training. The Government Agency has to keep a log of each employee’s status, identifying which employees are authorized access to PII, which have the relevant training and which need training. AR-6 addresses the Agency’s privacy reporting to OMB, Congress and other authorities. AR-7 requires that systems be designed and developed to maximize the protection of PII. AR-8 covers the Agency disclosing PII..

DI-1 and DI-2 cover data quality and integrity. DI-1 requires that the system operator confirm to the greatest extent practicable upon collection or creation of personally identifiable information (PII), the accuracy, relevance, timeliness, and completeness of that information. PII should be collected directly from the individual to the greatest extent practicable. The Agency must check for, and correct as necessary, any inaccurate or outdated PII. The Agency must also issue guidelines ensuring and maximizing the quality, utility, objectivity, and integrity of disseminated information. DI-2 covers the Agency’s Data Integrity Board.

The DM family of requirements are centered around minimizing the amount of PII collected and retained, and minimizing the unnecessary use of such PII. Per DM-1, the Agency must have a plan in place for protecting PII that complies with OMB Memorandum M-07-16, “Safeguarding Against and Responding to the Breach of Personally Identifiable Information.” The plan will follow the Guidance in [SP8122]” The plan must show how the Agency will identify the PII elements needed to accomplish the mission; collect ONLY those PII elements identified and ensure that consent for collection has been obtained where required; and evaluate PII holdings to ensure that they are still required.

DM-1 further says that the Agency should minimize the risk to PII by redacting or anonymizing it where that is feasible. DM-2 requires that PII be retained and later destroyed in accordance with the plan from DM-1. DM-3 says that PII should not be used for training and testing unless such use is absolutely necessary, and even then such use should be kept to a minimum.

The IP requirements deal with individual participation in data collection, and with redress. IP-1 covers the Government’s process for obtaining consent from the subject to collect PII, and IP-2 covers the Government’s processes for allowing a citizen to have access to his/her own data. IP-3 covers the Government’s redress process, and IP-4 covers complaint management.

The SE family of requirements covers the security of PII. For SE-1, as part of its overall PII plan, the Agency must maintain an inventory of all programs and information systems that collect, use, maintain or share PII. This inventory must be updated periodically, and all updates must be provided to the Agency CIO or designated information security official. The CIO or designated official will use this list to help set information security requirements for new systems that handle PII. SE-2 covers the Agency’s Privacy Incident Response plan.

The TR family of requirements addresses transparency of the process. TR-1 requires the Agency to develop and promulgate a Privacy Notice, and TR-2 requires a System of Records Notice, both of which are required by other Federal laws. TR-3 ensures that the public has access to the privacy-related notices and other information.

My goal in the next part of this appendix is to map these controls to the privacy facets that we defined in Section 2.2.

Recall from Section 2.2 of this report that we divided “Privacy” into the following facets:

Anonymity -

Pseudonymity -

Fair Use -

Access -

Control of Lifecycle -

Control of Aggregation

Appendix C: The Impact of Using a DIfferent Definition of Privacy

The facets of privacy that were described in Section 2.2 and used throughout the rest of this paper were those considered most relevant by the authors. But there are different approaches to defining and using “privacy” and it is possible to develop a different set of privacy facets.

There has long been controversy over the “proper” definition of privacy to use. Westin formulated one of the first widely-cited definitions in 1967:

> “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.” [WESTIN]

That has been modified numerous times, and wide agreement has seldom if ever been reached. Recently, the trend has been to not bother with a definition of privacy. The thought is that there is no one single definition that will fit for all purposes, because there are no core characteristics of privacy that fit across all situation.

Solove [SOL02, SOL06] is one of the strongest voices asserting that there is no single essence or core characteristic of “privacy.” Instead, he developed a taxonomy of acts that impinge on privacy. Figure C-1 shows the elements of Solove’s taxonomy. Since this taxonomy contains attacks that violate privacy, our intent is to show that we have addressed each element with our facets, or explicitly ruled an attack out of scope.

Surveillance, Interrogation: These attacks are covered by the security “Confidentiality” facet. Where they are directly relevant to privacy, they would fit in the “Anonymity” facet.

Aggregation: Covered in the “Control of Aggregation” facet. See also Example 5 in Section 4.2 for a description of aggregation attacks in general.

Identification: Covered in the “Anonymity” and “Pseudonymity” facets.

Insecurity: Covered by the Security facets

Secondary Use: Addressed in the “Control of Lifecycle”, “Control of Aggregation,” “Fair Use” and “Access” facets.

Exclusion: Covered in the “Access” and “Fair Use” facets.

Breach of Confidentiality, Disclosure, Exposure, Blackmail: Again, these attacks are covered by the security “Confidentiality” facet. Where they are directly relevant to privacy, they would fit in the “Anonymity” facet.

Appropriation: Addressed in the “Access” and “Fair Use” facets, with some parts covered in “Control of LifeCycle” and “Control of Aggregation.”

Distortion: Addressed in the “Access” and “Fair Use” facets.

Thus, the authors believe that our privacy facets adequately cover all parts of Solove’s Taxonomy that we believe are in scope.

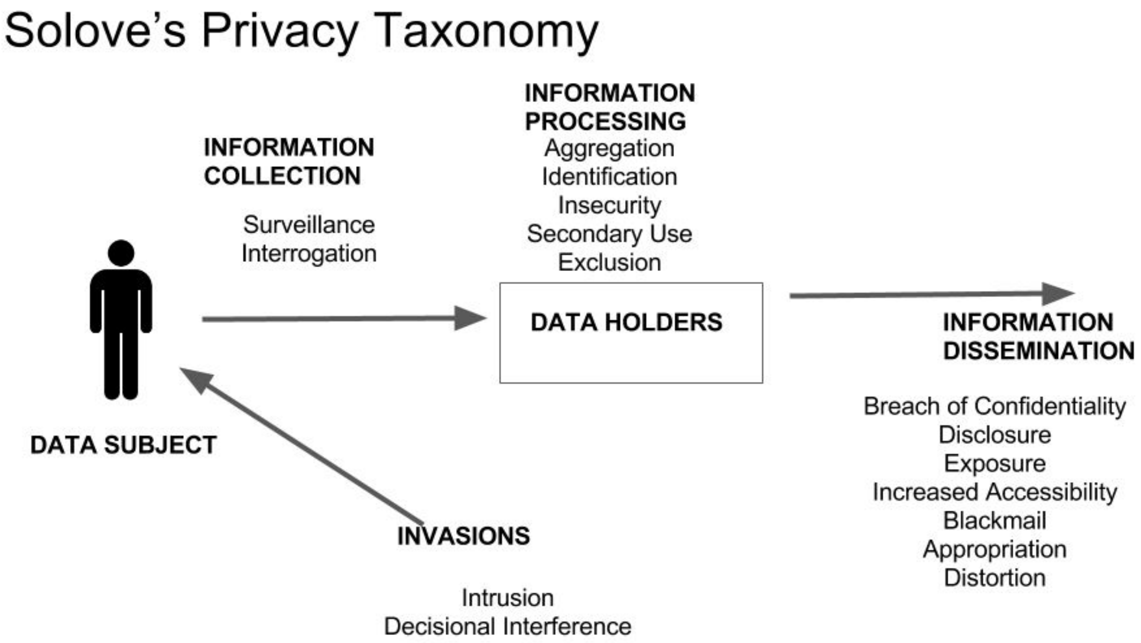

Figure C-1:Solove’s Privacy Taxonomy

Nissenbaum[NISS] agrees with the thought that it is a waste of time attempting to develop a single, formal definition of privacy. Rather, researchers and organizations should focus on contextual integrity - ensuring that data flows appropriately through a system, where "appropriately" depends on the system context.

The authors agree with Nissenbaum in this view, and thus we have limited the context of our analysis to Federal Government environments. In these environments the process for determining security requirements is well-established, and the process for determining privacy requirements is somewhat less formal but still recognized and structured. We believe that the privacy facets and process described in our report are consistent with Nissenbaum's call for a focus on contextual integrity.