FREEDOM OF INFORMATION IN THE AGE OF TERRORISM: THE MOSAIC THEORY IN PRACTICE

by

Harry R Cooper

A Dissertation Presented in Partial Fulfillment

of the Requirements for the Degree

Doctor of Science in Cybersecurity

CAPITOL TECHNOLOGY UNIVERSITY

April 10, 2018

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

FREEDOM OF INFORMATION IN THE AGE OF TERRORISM: THE MOSAIC THEORY IN PRACTICE

by

Harry R Cooper

April 10, 2018

Approved:

Dr. William V. Maconachy, PhD, Chair

Dr. Michael Fain, PhD, Committee

Dr. William Butler, D.Sc., Committee

Accepted and Signed:

_____

William V. Maconachy, PhD Date

_____

Michael Fain, PhD Date

_____

William Butler, D.Sc. Date

_____

Helen G. Barker, DM Date
Chief Academic Officer
Capitol Technology University

ABSTRACT

The premise for this qualitative grounded theory study involved identifying how it is becoming easily more mineable for critical pieces of data. The purpose of this grounded theory study was to identify what specific tools or tool sets make an experiment testing the mosaic theory of intelligence successful. Material analyzed in this study not only identified the specific tools or tool sets, but also produced a wide variety of themes that allowed for the data to be dissected further to answer many questions. The study identified themes affecting such topics as individuals best suited for intelligence work, complexity of finding certain pieces of data, and other factors. Additionally, the study generated some additional avenues for future research into a topic that has until recently remained in the purview of the government.

DEDICATION

I dedicate this dissertation to my family. For Matthew, who has suffered just as much for this dissertation as I did, thanks are not nearly enough. For my mother, sister, cousins, nieces, nephews, etc. I thank you for allowing me to be an absent part of the family while doing this dissertation. To my sister, Christine, thanks for handling our crazy lives while I was writing.

ACKNOWLEDGEMENTS

I wish to thank everyone who has helped me in getting this dissertation done. Without the help of these individuals I would still be sitting at my desk typing away. Thanks again for all the assistance you have given me.

I would first like to thank my husband above everyone else. Not only did he maintain a calm home life for the past few years, but he has acted as my soundboard for some of my crazier ideas and helped ground me in reality.

A deep thank you to Jeffrey Bardin who helped me identify this particular field of study in cybersecurity is something I enjoy doing. I also want to thank Joe Giordano who has served as a mentor to me for many years and helped me in getting involved in teaching.

Finally, I would like to thank Dr. William Maconachy, my chair, Dr. Helen Barker, who helped me, and Dr. Michael Fain, and Dr. William Butler, my committee, as well as Dr. Jeffery Hayman. They are the reason why this paper is readable at the end of the day.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The United States Court of Appeals decision in the case of *United States v. Maynard* (2010) stated the Mosaic Theory allows for one who has a broad view of a situation to realize the importance of small bits of data which may seem trivial to the uninformed. The decision focused on the legality of a 30-day period of GPS tracking of an individual. The case showcased in the court's opinion once an individual knows all of another person's travels one can deduct from this data whether an individual is an alcoholic, a cheating spouse, political associations… (United States of America v. Lawrence Maynard, 2010). The court agreed the collection of significant amounts of seemingly benign facts in part allow someone who can collect these facts to know all such facts about an individual.

There is a theory, the Mosaic Theory of Intelligence, which has been around for over seven decades even having been recognized by the courts as far back as *Halkin v. Helms* (1978). In this case the United States Court of Appeals, District of Columbia Circuit, noted the evolving age of computer technology has played a key role in evolving the business of foreign intelligence (Halkin v. Helms, 1978). The Court specifically notes no longer is foreign intelligence a cloak and dagger affair, but it is instead a construction of a mosaic of thousands of bits and pieces of seemingly innocuous information (Halkin v. Helms, 1978). Once the mosaic has been constructed the Court argues that analysis of the mosaic will provide the builder of the mosaic with a startling clarity of how the whole operates (Halkin v. Helms, 1978).

The Mosaic Theory of Intelligence, sometimes known simply as the Mosaic Theory, has been defined by many individuals, courts, and organizations, but the one definition which provided the most complete insight is from Richards J. Heuer, Jr. in his book the Psychology of Intelligence Analysis. In his definition of the Mosaic Theory of Intelligence Heuer noted small

bits of information and data are collected and put together like a jigsaw puzzle or a mosaic (Heuer, Jr., 1999). The combined nature of these mosaics gives intelligence analysts a much clearer picture into a situation than any single bit of information may have provided by itself (Heuer, Jr., 1999). Heuer noted thusly the collection and retention of all of the small bits of information must be maintained at all times, so intelligence analysts can either today or weeks or months later identify the value of said data in their overall mosaic of information (Heuer, Jr., 1999). This was just one of the rationales for the large technical intelligence collection systems implemented by the various national security agencies (Heuer, Jr., 1999).

Since the *Halkin v. Helms* case in 1978, the world and the technology running the world has advanced greatly progressing from the earliest supercomputer to today's smart watches. Since the Cray-1 supercomputer was first installed in Los Alamos National Laboratory in 1976, there has been an ever-increasing race to produce a faster, smaller, and cheaper alternative to the previous generation of computing. The Cray-1 set a world record in late 1976 for computing power with the speed of 160 million floating-point operations per second, or 160 megaflops (Cray, Inc., 2016). Whereas the Apple Watch 1, a wristwatch released in April of 2015 by Apple, operated at approximately 7 billion floating-point operations per second (Zolfagharifard, 2015). This is a 4275% increase in computing power with an associated 99.9985% price decrease from $38 million for a Cray-1 in 1976 to $549 for an Apple Watch 1 in 2015 (Cray, Inc., 2016).

The consistent progression of significantly increasing computing power and the associated decreasing cost of said computing power has progressed the Mosaic Theory of Intelligence from the purview of countries to the prevue of any individual with a small bit of knowledge, skills and the proper set of tools. In 2013, two new Open Source Intelligence (OSINT) tools became available online for anyone to use, FBStalker and GeoStalker

(Ruslanovich & Alekseevna, 2016). FBStalker made use of Facebook to determine when a user was active online, what their interests were, and even where they commonly visited (Ruslanovich & Alekseevna, 2016). GeoStalker made use of geolocation-related resources to take an address and build out a profile of a target located at an address (Ruslanovich & Alekseevna, 2016). Individuals however are not restricted to just these two tools, there are thousands of tools available readily to the public to gather open source intelligence from both public and semipublic resources. In most cases, it takes only a credit card for an individual to have a data center available to them.

Unfortunately, tools and tool sets are only one part of the overall problem. The data these tools collect or extrapolate play just as crucial a role in this problem. A key example of this data extrapolation can be found in one of today's biggest problems the re-identification of de-identified data. As the researchers behind the study titled, A Systematic Review of Re-Identification Attacks on Health Data, noted the overall success rate for all re-identification attacks on health data was between 26% and 34% (El Emam, Jonker, Arbuckle, & Malin, 2011).

This re-identification problem exists across the whole spectrum of the data available online. The ability to link an individual to data without their name is straightforward and used every day. Linking housing data, deeds, liens, mortgages, etc. are all very simple if someone has one piece of information, a home address. If an individual has an address, they can easily within minutes make use of local, county, state, and federal databases to get access to the details of the individual(s) who own the home. Thusly, it is key one understand the tools are just one part of the overall problem which exists in today's data rich society.

The Mosaic Theory of Intelligence does not discriminate in who can make use of the theory. The theory can be used by anyone from a stalker looking up information about the

individual they are stalking to a nation state attempting to piece together information for an upcoming operation against a country. As such, governmental bodies are working to establish a set of rules and regulations to limit more critical data. One such case being Presidential Executive Orders 12356, 12958 and 13292. Executive Order 13292, Further Amendment to Executive Order 12958, as Amended, Classified National Security Information, states, compilations of unclassified pieces of information may be classified if it meets certain standards (Executive Order No. 13292, 2003). The standards are that the compiled information reveals an additional association or relationship classified in the Executive Order, or it reveals an additional association or relationship which is not revealed in the individual items of information (Executive Order No. 13292, 2003). Unfortunately, these regulations and rules are there to protect the government only, leaving the rest of society and their information at the whims of the personal users, the companies they entrust their data to, and many others.

It should be noted the Mosaic Theory of Intelligence and the act known as doxing are very closely related but differ in their intentions and even sometimes their methods. Doxing is defined as an intentional release onto the internet by a third party of the personal information of an individual for the purposes of humiliation, threatening, intimidation, or punishment of said individual (Douglas, 2016, p. 199). Whereas, the Mosaic Theory of Intelligence argues small bits of information and data are collected and put together like a jigsaw puzzle or a mosaic (Heuer, Jr., 1999). The combined nature of these mosaics gives intelligence analysts a much clearer picture into a situation than any single bit of information may have provided by itself (Heuer, Jr., 1999). While an individual may use identical tools, tactics, techniques, or procedures to carry out their underlying research of an individual, it is the end goal which separates the two acts from one another. Finally, of note, a study of the validity of the data compiled using the Mosaic

Theory of Intelligence, which is collectively called sources and methods, is not available to be included as the intelligence community does not put out these figures due to national security reasons and showcasing perceived vulnerabilities and shortcomings of the system and/or user in analysis.

This chapter first covered the background of this study and why it is of important social, scientific and theoretical interest. This was followed by a review of the specifics regarding the study put forth by the researcher. A brief overview of the study regarding the observation of what tool set or sets are successful in building the most detailed mosaic on a given topic or target has also been provided. The researcher has further observed whether there are additional demographical details which improved the success of a given tool set.

The study itself made use of a single individual as the target of the study and asked the study participants to provide as detailed of a dossier of the individual as possible. This specific target has been covered extensively by the press. The data collected on this topic by the researcher has been used to generate a base for what information is available. Additionally, the researcher served as a validator of data gathered in the study and verified the accurateness of said data against the researcher developed database.

Background

In addition to the tools available to the average savvy internet user, there is a completely different subset of tools and organizations making use of the Mosaic Theory. In February of 2009, Google published a paper in conjunction with the Centers for Disease Control and Prevention, CDC, which presented a system which could predict possible flu outbreaks (Ginsberg, et al., 2009). By making use of the billions and billions of queries hitting Google each day, Google could refine a small subset of search queries. Google could then extrapolate based

off factors such as location, rate of search word appearance, and many other factors, an estimate of the spread of influenza in a specific region. Using these methods of taking tons of disparate information and combining the data into an actionable prediction, Google could estimate levels of weekly influenza activity with a lag of about one day, compared to the CDC's typical lag of about two weeks.

Target Corporation's use of business analytics to dig through their sales data presents a unique peek into the success of analytics in identifying potential sales to a customer but with unintended side effects. Bucking the trend in the retail sector of sending coupons and advertisements to the parents of newborns based on public birth records, Target decided to approach the sales opportunity in a more proactive fashion (Sprague, 2015). Using their business analytics system to delve into a shopper's buying habits, Target could discern from purchases made by individual customers, those customers who were most likely pregnant. This usage of the Mosaic Theory allowed for Target to begin the direct marketing of products to a customer months in advance of other retailers. There may have been an unfortunate side effect to this marketing push, as Target informed a father or his daughter's pregnancy before she had informed him of her good news (Sprague, 2015).

The anticipated impact of this study as seen by the researcher is that the intelligence community, the data privacy community, and other interested parties will gain an understanding of the tools and individuals who may prove to be of the most concern to them. The progression of technology often makes life easier, but as technology progresses it is the responsibility of the communities involved to understand the role said technology has had in eroding former barriers of protection. The Mosaic Theory of Intelligence was seen in the 1970s as available only to the largest players, who had access to significant computing resources, but in the proceeding years

this limitation has disappeared with the advent of services such as cloud computing and on-demand utility computing (Rosenzweig, 2015). The public and the security communities need to understand anyone with enough talent, skills, and drive can bring to bear a great amount of resources to find anything they want (Rosenzweig, 2015).

Theoretical Standpoint

From a theoretical standpoint, this study addressed issues concerning the necessity of identifying the best tools and tool sets to carry out a successful mosaic of a target or a topic. The data these tools and tool sets uncover can prove intrusive and worrisome when used by someone for purposes outside of the expected usage, such as deed searches, web searches, etc. for said technology. Justice Ginsburg in the United States of America v. Lawrence Maynard, reviewed the application of the Mosaic Theory of Intelligence in this case, and noted a whole month of collection of a person's movements cannot be construed to be public as the likelihood of a single stranger observing all of these movements is nil (United States of America v. Lawrence Maynard, 2010). Justice Ginsberg argues the public nature of the acts is waived when one collects all of the acts for a sustained length of time as it has the effect of unmasking private information and private routines of the individual which would not otherwise have been exposed in a single observance (United States of America v. Lawrence Maynard, 2010). The Maynard ruling showed that tools, in this case a GPS tracker, are prone to abuse by individuals and organizations alike. The ruling also showcased the usage of technological tools to gather information in conveyance of producing a mosaic on a target can have many different implications such as legal matters or national matters.

From an international perspective, the Mosaic Theory can be abused by nations for their own purposes. One very distinct instance of this abuse is the extraordinary rendition of foreign

nationals by the United States to Guantanamo Bay. In the case of Alla Ali bin Ali Ahmed, et al. v. Barack H. Obama, et al., the United States put forth the assertion that a mosaic of disparate pieces of information regarding the allegations against the detainee should only be examined in the whole (Alla Ali Bin Ali Ahmed, et al. v Barack H. Obama, et al., 2009). The argument was since no single piece of evidence or testimony could prove the guilt of the detainee then the totality of the evidence and testimony must be accepted in whole as proof of guilt (Alla Ali Bin Ali Ahmed, et al. v Barack H. Obama, et al., 2009). The United States shifted the Mosaic Theory of Intelligence away from a method of intelligence gathering to a method of prosecution.

From a national perspective, the American public was given a sneak peek into the ways the Mosaic Theory was being used by the government for mass surveillance (Mornin, 2014). In June 2013, The Guardian newspaper published the first of many articles covering the mass surveillance of Americans by the National Security Agency (NSA) (Mornin, 2014). In this article, the newspaper noted the blanket order issued by the Foreign Intelligence Surveillance Court (FISA) required Verizon Communications to provide the NSA with the call records of millions of Americans, regardless if they were suspected of any wrongdoing. The data included the numbers of both parties, location data, call duration, unique identifiers, telephone calling card numbers, trunk identifiers, International Mobile Subscriber Identity (IMSI) numbers, and the time of all calls (Mornin, 2014). As the article notes, the primary conclusion to be drawn from this widespread collection of data is the government wanted to build out a comprehensive picture of who any individual contacted, how, and when and even from where (Mornin, 2014).

From a social perspective, one can simply look again at the National Security Agency and its collection policies. The NSA argued under the “hop” or “chain” analysis method they use, the NSA can collect and review not just a suspect’s phone records, but also the phone

records of everyone he calls, everyone who calls those people and yet again everyone who calls those people (Mornin, 2014). This chaining analysis means if a suspect called 40 unique individuals, the three hops analysis would allow the NSA to review the records of 2.5 million individuals (Mornin, 2014). Without knowing or inferring the intent of the NSA, one can simply look at these numbers and the three hops analysis and see the combination of data done under the guise of national security has aimed the Mosaic Theory against one's social interactions and activities.

From a financial perspective, implementation of systems which follow the basics of the Mosaic Theory of Intelligence can be a moneymaker for companies. These companies have earned the nickname "Privacy Merchants" by Amitai Etzioni in his aptly named journal article, "The Privacy Merchants: What is to be Done? (Etzioni, 2012)". As of 2005, one of these companies, Choicepoint (now LexisNexis Risk Solutions), had records on over 220 million people. The data was culled by corporate data miners to build out a profile of each person and included information such as, demographics, income, net worth, property holdings, social security numbers, known addresses, known phone numbers, driving records, employment, criminal records, credit card transactions, and much more (Etzioni, 2012). This type of information is then sold onto companies, law enforcement, insurers, individuals, and others who can pay the fee to gain access to the database.

From a military perspective, the usage of disparate information across news, social media, and other points of data can be used by one's enemies to identify and track down members of the military involved in specific incidents. A case study performed by Dr. Hayman on the use of reidentification to recover redacted United States (U.S.) Army Crewmembers showcases this is a real threat (Hayman, 2015). The participants in the study were given a 15-

minute online research task after being provided only the date and type of a helicopter crash. The study found of the ten participants involved, all ten participants could successfully identify by name and rank all service members onboard the accident aircraft even though the Army had redacted said information from the official accident report. The case study in reidentification showcases these tools used for reidentification, which are many times the same tools used in mosaic work, can be used by anyone to identify service members involved in an incident and thusly expose them to possible retribution from third parties, such as the enemy or from a casualty's family (Hayman, 2015).

Problem Statement

This dissertation explores the use of, the Mosaic Theory of Intelligence. This theory has been recognized by the courts since Halkin v. Helms (1978). Since this case, the state of computer technology in the late 1970s has been surpassed rapidly in the three decades since. A general problem exists where the technology to mine data is available to everyone, and mining can have far reaching criminal ramifications by building a mosaic of information on any given topic. In 2014, C.E. Walsh identified a new problem affecting data which had not existed until recently. Walsh identified a problem exists in the increasing amounts of information which is gathered each day and how it is becoming easily more mineable for critical pieces of data (Walsh, 2014).

A general problem statement is that data mining which once was a tool only usable by large corporations or governments, is now available to the public at large at a minimal cost, and mining can have far reaching criminal ramifications by building a mosaic of information on any given topic (Rosenzweig, 2015). The specific problem is what specific tools or tool sets make an experiment testing the mosaic theory of intelligence successful. Secondly, what specific subset of

individuals can make use of those tools or tool sets to build a successful profile of a targeted individual when given a basic set of information on said individual.

The problems posed by the expansion of the usage of the Mosaic Theory of Intelligence to the masses is capable of growing into an enormous undertaking. One such problem is the crime of stalking. While there are thousands of cases out there which involve ex-boyfriends, ex-girlfriends, or ex-spouses cyber stalking their former love, that does not rise to a level associated with the Mosaic Theory as the individual is known intimately beforehand. A non-intimate stalking incident though does fit the Mosaic Theory very well, as the individual being stalked does not know the stalker and likewise the stalker does not know details about their victim ahead of time. In recent years, there has been an increase in celebrities, wealthy individuals, sports figures, and more being stalked by overly zealous fans (James & MacKenzie, 2018).

One case of cyberstalking making use of the Mosaic Theory of Intelligence is the case of Pradeep Manukonda. Manukonda decided to target the ultimate target on Facebook, Facebook Founder and CEO, Mark Zuckerberg (Hill, 2011). In January of 2011, Manukonda began a campaign of stalking Zuckerberg, going to Facebook headquarters, Zuckerberg's home, and sent messages through Facebook to Zuckerberg and his loved ones (Hill, 2011). Manukonda was able to gather the information needed to carry out his stalking activities from multiple websites, including Facebook as well as Gawker. On January 13th of 2011, Gawker posted pictures of Zuckerberg's house along with a statement noting it was just down the street (Hill, 2011). From this information, Manukonda put together the location and subsequently arrived at his home on January 24th but was intercepted by security at the last second as he approached the front steps (Hill, 2011).

This is just one instance of how the Mosaic Theory of Intelligence can be misused. There are many individuals out there who can abuse the use of Mosaic Theory techniques for their own gain. One such usage was the attempted use of the Mosaic Theory as a defense for insider trading charges researcher (The Economist, 2011). Raj Rajaratnam, former boss of the hedge fun Galleon Group, was charged in October 2009 for insider trading. During his trial in Federal Court, his lawyers claimed Rajaratnam's success in identifying opportune times to invest was not associated to improperly gathered information, but rather to legitimate analysis done by a meticulous researcher (The Economist, 2011). This qualitative study will focus on the tools and tool sets which are used by individuals, rather than corporations or governments, so it can be discerned what tools are best at accessing information for potentially nefarious purposes.

Purpose of the Study

The purpose of this qualitative study was to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target. The experiment conducted for this study focused on the tool sets used by the participants in building out a detailed mosaic profile. The preselected target of this detailed mosaic profile was, Aaron Hernandez, a former tight end for the New England Patriots.

The research study made use of an online mosaic profile with participants of varying education levels making up the targeted research population. Various demographics of each participant was collected but no personally identifying information was included in any part of the study. No bias on behalf of the researcher in the participant population was anticipated, as the participants were not chosen ahead of time for this study. Rather, they were gathered from a pool of individuals in the cybersecurity field. The participants was note involved in anything other than the completion of an open-ended profile on a specified target. Finally, the collection

methodology chosen for this study was an open-ended questionnaire. This method was chosen as the data requested from the study participants was extremely varied and depending upon their source may or may not have been accurate (Creswell, 2012).

Significance of the Study

The most significant impact anticipated from this study was that the intelligence community, the data privacy community, and other interested parties could gain an understanding of the tools and individuals which may prove to be of the most concern to them. The progression of technology makes life easier, but as technology progresses it is the responsibility of the communities involved to understand the role said technology has had in eroding former barriers of protection. The Mosaic Theory of Intelligence which was seen in the 1970s as available only to the largest players is no longer the case as more average individuals are making use of this theory (Rosenzweig, 2015). The public and the security communities need to understand anyone with enough talent, skills, and drive can bring to bear a great amount of resources to find anything they want.

The secondary impact as seen by the researcher was that the public would obtain a better understanding of what happens to the information they release every single day. The desire of the researcher was showing average citizens posting a picture from one's night out with the girls or a picture of oneself and fraternity brothers partying during Spring Break can have lifelong consequences. In addition to informing the public of the dangers of their own informational releases, the study hoped to make the public aware the information we give up is no longer restricted to the organization we hand said data over to but to many others. Whether this disclosure is done under open government laws, reselling customer information, or even just

accepting cookies when browsing, the data is never under one's own control but is available to many others.

Nature of the Study

The research design chosen for this study was a qualitative analysis using the grounded theory research design. A single individual was targeted rather than a disparate group of individuals. While a large target set could serve to increase the success of any given tool set, they were excluded due to the inability of validating the information given or by allowing the participants to use existing knowledge of a targeted individual. A wide variety of data was collected from participants and stored in a database along with the associated tool or tool set the data was garnered from allowing for multiple evolving interpretations to be derived from the data (Nieswiadomy, 2008).

The grounded theory research design was chosen as the stated goal of identifying the tools, tool sets, and key demographics remained an unknown or emergent piece of information. The grounded theory allowed interpretations to be continually derived from the raw data, or more succinctly, the story emerged from the data (Nieswiadomy, 2008). This flexibility of allowing interpretations to be continually derived extends to all parts of the study and allows the theory to be self-correcting (Nieswiadomy, 2008). As data is collected adjustments can be made to allow for interpretation of the newly collected data (Nieswiadomy, 2008). This grounded theory approach, made use of an open-ended questionnaire, also referred to as a profile form, to guide the participants to what information was being looked for as part of the study.

The study planned to involve many participants, with estimations of 15 individuals with an undergraduate degree, 15 individuals with a graduate degree, and 15 individuals with a doctoral degree or currently are doctoral students. Participants were given 25 minutes, controlled

by the JotForm software being used, to perform the study's task and submit the results of their work. The participants were instructed to make use of any tools they had access to such as search engines, publicly accessible government databases, Maltego, and other software packages. Students were restricted from contacting the subject's family or from purchasing any data online. The participants were asked to identify as much information as possible on the targeted individual while being monitored programmatically by the researcher via the JotForm program for multiple submissions from the same IP to ensure no participant interactions.

The data which was provided to the participants is limited to the full name of the target, his year of birth, his state of birth, and the fact he had an interesting life history before dying in jail. This basic data was chosen as these pieces of information are the most commonly used facts when giving a brief introduction of a person to a large group. The study's intent was to identify that open source information is more easily mineable with the proper tools, so an individual who is unknown to the participants but has a decent exposure of data across open sources was necessary to ensure the best test results.

The subsequent analysis of the results was used to identify the best tool set as well as any key demographics which enhanced one's ability to carry out the task. Outside of the profile form's mosaic profile information a small section of personal demographics was requested from each study participant. It was anticipated the extent of the demographics requested wwould be age, gender, education level, current GPA, and a self-rated tech skill scale. Finally, at the conclusion of the exercise, a short set of post-study open-ended questions were asked to identify any potentially relevant data which was not anticipated and accounted for by either the profile form or the demographic questions.

Overview of the Research Method

A grounded theory approach to the research was appropriate as the focus of the study, identifying tool sets and key demographics, are an unknown or emergent piece of information (Nieswiadomy, 2008). The grounded theory allowed interpretations to be continually derived from the raw data, or more succinctly, the story emerged from the data (Nieswiadomy, 2008). A qualitative method was also the appropriate method to use as the data being sought was not known by the researcher but was rather being discovered by the researcher in the study (Creswell, 2014). The researcher followed the standard steps in the research, starting first with the literature review, then the study, and finally the recording of the results and subsequent processing of the results.

The data which was collected during the study were the participant demographics, mosaic profile form data, and the post-study follow up questions. The quantitative research design was rejected as it focused on collecting measurements, analyzing said data, and finally determining whether the original hypotheses is true (Creswell, 2014). The qualitative research design was accepted as it is focused on understanding the world and its effects on the study's subjects. Through the collection of a subject's recollection of an event, the researcher works to determine the views of not only the individual but possible outside forces acting on the subject such as historical or cultural norms (Creswell, 2014). Finally, the mixed methods research design was also rejected as it is focused on the understanding one must first collect diverse types of data to provide the most complete understanding of a problem, and then the researcher can narrow their inquiry down to a smaller population with which they can conduct interviews (Creswell, 2014). The goal is to have the interviewed subjects' views help explain the results of the initial survey (Creswell, 2014).

Overview of Design Appropriateness

This study was seeking to determine what tool set or sets are successful in building the most detailed mosaic on a given topic or target as well as what specific subset of individuals use those tools or tool sets. A grounded theory approach to the research was appropriate as the focus of the study was an unknown or emergent piece of information. The grounded theory allows interpretations to be continually derived from the raw data, or more succinctly, the story emerged from the data (Nieswiadomy, 2008).

The other qualitative research designs were each compared and contrasted below to explain why they were not chosen. The first design rejected was a biographical study. Biographical studies are the collection of archival information and media to produce an exhaustive account of a life experience with either a wide or narrow focus (Nieswiadomy, 2008). Since, the information the study was looking for was not yet known and has not been previously identified by anyone; the biographical approach was not proper for this study (Nieswiadomy, 2008).

The second design rejected was a phenomenology study. Phenomenological studies are aimed at understanding an occurrence or experience while taking into account the point of view of an individual, such as their reactions, perceptions, and feelings (Nieswiadomy, 2008). The information the study was looking for does not mesh with this design as the data being collected was the tool set chosen by an individual to carry out their task, and the emotions of the individual are disregarded. The third design rejected was an ethnography study. Ethnography studies are aimed mainly at sociological or anthropological studies where one makes directed field observations of a group of individuals or a culture (Nieswiadomy, 2008). Since the study is not

carrying out any type of sociological experiments this design can be safely excluded from consideration.

The fourth and final design rejected was a case study. Case studies are best described as an in-depth analysis of people, events, and relationships, bounded by a unifying factor (Nieswiadomy, 2008). This design was the most likely competitor to the chosen design of grounded theory. The main problem with making use of the case study method was the unknown factors which existed in the study, principally, which tool sets the users would make use of in the exercise, and the ability of this design to allow the data to drive modifications to the study's theory as the results emerge (Corbin & Strauss, 2015).

This study was designed for the express purpose of observing what tools were being used without any guidance from the researcher. Had the focus of the study instead been to determine the demographics of the users of a specific set of predefined tools, then a case study would have been used. In conclusion, while many types of study design are available to a researcher to choose from, the methods used and the data being sought will not necessarily fall under each design type. Thusly, a responsible researcher must make a decision based off of their methods and data being sought and pick the most accurate study design (Creswell, 2014).

Hypotheses/Research Questions

The study addressed a set of research questions to hopefully be fully answered by the study research, the experiment, and finally the results of the experiment. The identification of the tool sets that demonstrated the most detail in completing a successful mosaic profile was at the core of the study and needed to be answered. Next, the identification of the demographics of individuals who demonstrate the highest rate of success in completing a successful mosaic profile opened up the data for additional interpretation. The identification of any demographics

of individuals who made use of the particular tool sets would allow for organizations to extrapolate new policies and procedures.

An additional secondary set of questions has also been put together as the researcher felt the answers to these secondary questions would provide additional insight into the results of the study. Was there a significant correlation of one age group to Google or is Google age agnostic. What percentage of data found was incorrect, false flag. Did any subjects make use of a paid service, if so, did it increase accuracy or completeness. Would the location of the experiment play a role in the project. Does the combination of age, gender, and location of subjects pose a risk of cross correlation of data to expose subjects.

The primary research question, which tool sets demonstrated the most detail in completing a successful mosaic profile, was set up to identify the best tools which made use of the Mosaic Theory of Intelligence. The term tools in this study covers the programs, websites, databases and other electronic sources used to collect or access data. The study was seeking to not only identify the tool or tool sets most widely used, but in addition identify those tools with the greatest success in locating pieces of information. The identification of the best tool or tool sets could provide all interested parties with a better understanding of the spread of information across the internet.

Due to the chosen usage of the qualitative research method and the grounded theory research design a hypothesis was not required for this study, as these methods aim to generate a hypothesis rather than testing a hypothesis (Corbin & Strauss, 2015). The primary objective of this study was to identify the best tools or tool sets in creating a mosaic profile. In addition, it also aimed to identify via secondary objectives and research questions whether there existed a specific subset of demographics of the participants who were better suited to make use of the

identified tools to get the most concise mosaic profile. Additionally, the research was also structured to allow for the conclusions to be reviewed by future researchers to see if there were additional factors not identified in the study which may have increased the level of completeness presented via a tool set or demographic (Corbin & Strauss, 2015).

To assess the success or failure of a given tool or tool set, the researcher needed to first carry out the study and collect the data generated by the study participants. The source of the data to be compiled and reviewed came from the questionnaires, or profile forms, and contained not only data on the target but also demographic data on the study participant and the post study questionnaire data. The profile forms were then evaluated using a criterion-referenced test which defined a level of performance where the only thing of importance was said participant's performance (Salkind, 2016). Once all data was compiled, the researcher analyzed the data looking to extract from the data the answers to the research questions. The most complete profile forms were reviewed to determine the tools used in the production of those profile forms to gain an initial understanding of the most successful tool. Additional data analysis then focused on various correlations, such as identifying the average usage of a tool individually and overall, the validity of the results returned by each tool, the demographics of the users of each tool, and other possible correlations.

The research questions of the study were built upon the grounded theory based off the unique features present in this theory as identified by Corbin and Strauss; the theory is constructed upon various concepts which are derived from the data collected only during the research process and not prior to beginning and research analysis and data collection are interrelated (Corbin & Strauss, 2015). The primary research question focused on identifying tools or tools set which produce the most complete mosaic profile. The tools or tool sets were not

identified ahead of time and then analyzed based off a previous hypothesis, rather the tools or tool sets would emerge from the data collected from the study participants.

The second component of grounded theory states that the research analysis and data collection are interrelated (Corbin & Strauss, 2015). A cursory review of the other two primary research questions and subsequent secondary questions showcased this component of the grounded theory in the setup of this study. Once, a tool or tool set was identified from the data collected, the tool or tool set was then reviewed with regards to its usage by certain demographics, success rates, or vulnerabilities to inaccurate data. The data collection and subsequent analysis did not stop there, it continued throughout the entire study process. These analyses included questions such as whether location of where the study took place will have an effect on the results or whether the collection of the demographic data combined with the location of the study would present itself as a target for the tools or tool sets identified to produce a successful mosaic on the study participants.

Theoretical Framework

The theoretical framework explored by this study was what tools or tool sets have the greatest success in presenting as complete of a mosaic profile as possible. A qualitative study using the grounded theory was used as the base for the research that was conducted. The study focused primarily on the tools or tool sets which produced the most complete mosaic profile of a given target, in this case Aaron Hernandez, a former tight end for the New England Patriots. The more complete a profile produced by a tool or tool set can present a greater risk to the individual being focused upon.

There was very little information publicly available on this topic, hence why this topic was chosen by the researcher. There exists a wide collection of literature on the mosaic theory's

effect on legal and criminal matters, as well as, the overall discussion of the pervasive nature of the internet into one's life and the associated loss of control of the personal data of said individuals. This lack of literature resulted in a struggle in finding other studies or experiments which correlated to the study laid out herein. Thusly, the literature review section of this dissertation was expanded beyond the mosaic theory of intelligence and looked at closely related fields such as big data, business analytics, doxing, and others.

There was one study, which was the closest match possible to the proposed study, which did serve as a key component during the research and development phase of this study. The study by Jeffery W. Hayman is titled, "Case Study: Suggested Best Practices for Redacting U.S. Army Aviation Accident Reports to Reduce Opportunities for Doxing of Re-identified U.S. Army Aircrew" (Hayman, 2015). This study by Hayman provided support to the underlying construction of this proposed study and its implementation. Hayman's study and dissertation proved randomly selected participants in a study could in fact find information on individuals with the most basic of information about their target (Hayman, 2015).

While Hayman's study did focus on the act of reidentification, it should be noted the differences between the basic mechanics of reidentification and a mosaic profile are negligible at their core. In fact, Hayman's study used a roughly equivalent arrangement for the measuring instruments which this study also provided (Hayman, 2015). Finally, Hayman's results showed the success of participants in producing a reidentification of redacted information was so successful the researcher made modifications to the original study to account for the success of Hayman's study (Hayman, 2015). These modifications primarily affected the study with the addition of some additional secondary questions to hopefully identify to a deeper degree what increased a participant's success.

Definitions

To ensure the best understanding of the terms utilized in this study, a set of definitions have been provided to provide a clear understanding of the terms and their usage in this study. These definitions include:

Demographics is defined as, a subset of the general population sharing some form of a common characteristic, such as age, gender, class, etc… (William Collins Sons & Co. Ltd., 2016).

Data Mining is defined as, the extrapolation of information from existing data in a database which allows for the discovery of one's habits (William Collins Sons & Co. Ltd., 2016).

Doxing is defined as, the practice of publishing on the internet personal information on an individual or group without their consent (Schneier, 2015).

Mosaic is defined as, a high quality and detailed profile produced by technology which yields significant amounts of information such as where one goes, one's associations, such as political, religious, amorous as well as the pattern of our professional and vocational pursuits (Selva, Shulman, & Rumsey, 2016).

Mosaic Theory is defined as, the ability to the see in a broad view of numerous pieces of data the important details which would otherwise seem trivial to the uninformed (United States of America v. Lawrence Maynard, 2010).

Open Source Information is defined as, information available to everyone which is stored on publicly accessible media without any expectation of privacy (Hayman, 2015, p. 29).

Open Source Intelligence (OSINT) is defined as, intelligence produced from publicly available sources of information, such as news broadcasts, government documents, or

information available online, which is collected and distributed to an audience for the purpose of addressing a specific intelligence query (Bazzell, 2016, pp. III-IV).

Tool sets is defined as, a set of tools which are predefined to work with specific data and programs (William Collins Sons & Co. Ltd., 2016).

Assumptions

When producing any type of research, a researcher must make certain assumptions. The researcher has identified the following assumptions which if proven false could be devastating to the overall study. While the goal of removing these assumptions is lofty, Anselm Strauss, the father of Grounded Theory, stated individuals are the products of their culture it is important to recognize when either your own or the participants' biases, assumptions, or beliefs are injecting themselves into the analysis. (Corbin & Strauss, 2015).

The first assumption which was critical to the success of the study was that every individual in the experiment would be able to produce some quantity of information when building the mosaic profile on the supplied target. This assumption relied upon the mosaic cyber skills of the study participant, which was difficult for a researcher to quickly assess (Palmer, 2015). The researcher in this study felt confident as the mosaic cyber skills were being assumed due to the advanced educational level of the participants.

The second assumption which was critical to the success of the study was the usage of a variety of tools. While it was expected Google Search would be identified as a key tool for locating mosaic information, it was assumed many of these individuals would have access to additional tools they are familiar with using. This assumption relied upon the basis that participants will possess differing levels of experience (Corbin & Strauss, 2015). While these

differing levels exist, we must understand them and consider these differing level as properties of our participants, and track their effects.

The final assumption which was critical to the success of this study was that the researcher could have access to enough participants at each of the varying levels of educational levels which served as a key demographic of the participants partaking in the study. The study planned to involve many participants, with estimations of 15 individuals with an undergraduate degree, 15 individuals with a graduate degree, and 15 individuals with a doctoral degree or currently are doctoral students. All attempts by the researcher were made to ensure the number of participants was going to be as close to the estimates as possible. These attempts included agreements with various industry groups to garner access to established individuals of the cybersecurity industry.

Scope

The scope of this study involved the use of a mosaic profile on former tight end for the New England Patriots, Aaron Hernandez. The participants were given a small amount of basic data limited to the full name of the target, his year of birth, his state of birth, and the fact he has had an interesting life before dying in prison. The study participants were chosen solely based on the availability of the prechosen group of individuals from the Cyber Security Forum Initiative (CSFI) LinkedIn group and the High Tech Crime Consortium (HTCC) mailing list spread across the educational spectrum. The study presented the participants with a mosaic profile form asking for certain pieces of specific data as well as open ended data. The participants were given a 25-minute period, controlled by the JotForm software which was used, to do their task of producing a mosaic profile on the target. At the end of the 25 minutes, the participants were asked to complete a short set of post study questions.

Upon completion of all study exercises, the data from the participants was collected, collated, recorded, and analyzed by the researcher. The data was first used to identify the most successful tools or tool sets in completing a mosaic profile. Once this had been determined, participant demographics were then compared to discern any trends which may emerge from the data as the increased success by a participant based on any of the collected demographics. Finally, a secondary set of questions were also examined to see what additional correlations or key data could be extracted from the data.

Upon completion of the study, the researcher presented what tools and tool sets are most successful as well as any other key pieces of data which emerged from the study and its results. This presentation of the results allowed the study's audience to make any decisions or conclusions regarding this field and what effect it may have on them. This study did not reveal at its conclusion any pertinent details on either the study's target nor on the study's participants to ensure their identity remains as anonymous as possible.

Limitations

The researcher chose to limit the target of this study of the mosaic theory of intelligence and the success of mosaic profiles to a single individual. This decision was made to account for multiple potential pitfalls. The first potential pitfall was the privacy of any individuals who might have been chosen by a participant as their study target. The second potential pitfall was the ability to validate any of the data and its accuracy. The third potential pitfall was the ability to correlate the information collected to the tools listed by the study participant versus possible prior knowledge of the study target.

A second potential limitation faced by the researcher was the ability to access enough individuals to provide a statistically relevant pool of participants across the educational

spectrum. The researcher first validated the minimum number of participants needed in each educational level against a comparable study done by Jeffery Hayman, which used a single level of participants garnered using the snowball method (Hayman, 2015). The researcher was in contact with industry groups who had agreed to provide access to individuals at each of the educational levels under review. The researcher prepared to work with other groups if necessary and had held preliminary talks with at least two other institutions prior to acceptance by the industry groups.

Delimitations

The first delimitation imposed by the researcher was to only allow the investigation of a single targeted individual. The researcher made this decision to counteract previously discussed pitfalls which could have affected the results of the study as well as overall validity. The second delimitation imposed was the specific choice of the target with privacy being of concern, the decision was made to use a deceased individual. The chosen target was a well-known and recently deceased football player. Additionally, the individual has in the past had these types of activities carried out by the press and others due to his celebrity status.

The final delimitation imposed was the choice of the size of the pools of study participants. As previously mentioned, a comparative study utilizing snowball sampling made use of just 10 participants (Hayman, 2015), but it was felt this was too low of a number to provide statistical relevance to the demographics to be drawn from the pool of participants. As this qualitative study utilized the grounded theory, it was determined too small of a pool would have negative impact on the inferences which could be drawn on demographics so the pool was purposely expanded to provide much higher numbers of individuals (Creswell, 2014).

Chapter Summary

This first chapter was aimed at establishing the core aspects of the researcher's study including the problem and purpose statements. The specific problems laid out were what specific tool sets made an experiment testing the mosaic theory of intelligence successful. Secondly, what specific subset of individuals used those tool sets to build a successful profile of a targeted individual when given a basic set of information on said individual. In addition to identifying these key pieces of information, the chapter also described a base for why this study has an inherent interest for the cybersecurity, intelligence, and privacy fields. Finally, the chapter also describes some of the current activities that are being completed by groups such as Google, Target, the CDC, and more that are unknowingly making use of the same techniques of the Mosaic Theory of Intelligence with great success (Sprague, 2015) (Ginsberg, et al., 2009).

The purpose of this qualitative study was to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target. These tools and tool sets can prove very dangerous when used by someone for purposes outside of their expected usage. The Maynard ruling showed that tools, in this case a GPS tracker, are prone to abuse by individuals and organizations alike (United States of America v. Lawrence Maynard, 2010). The ruling also showcased the usage of technological tools to gather information in conveyance of producing a mosaic on a target can have many different implications such as legal matters or national matters.

In Chapter 2, a literature review focusing on current and historical references was explored across a wide selection of topics such as, big data, business analytics, judicial rulings at the state, national, and international levels, the intelligence community, social networks, surveillance, and many other associated topics. Discussion of previous studies which correlate to

this topic such as the Hayman study (Hayman, 2015) were also reviewed. Finally, Chapter 2 further investigated the themes presented here in Chapter 1 as background for the need and purpose of the study.

CHAPTER 2: REVIEW OF THE LITERATURE

Intelligence gathering and analysis has developed greatly over the past seventy-five years. The second Director of the Central Intelligence Agency (CIA), Lieutenant General Hoyt S. Vandenberg, said intelligence work was the equivalent of building a picture piece by piece (Hilsman, Jr., 1952). However, the third Director of the CIA, Rear Admiral Roscoe H. Hillenkoetter, identified in 1948 the basis for the Mosaic Theory of Intelligence when he said the job of an intelligence operator is to identify vital facts from the all the extraneous data and put it together like a gigantic jigsaw puzzle which presents the picture the decision makers need (Hilsman, Jr., 1952). In the decades following this description of the Mosaic Theory of Intelligence, technology has advanced significantly. In 2014, C.E. Walsh stated a general problem exists in increasing amounts of information are becoming mineable for critical data (Walsh, 2014). Thusly, the purpose of this qualitative study is to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target.

Salkind (2009) noted a successful research proposal and study are built upon a logical and systematical review of the literature available on the key topics of one's proposal and study. An extensive and thorough review of the literature on one's research topic is the only way for a researcher to validate not only the uniqueness of the research question they are looking to answer but also to provide them with a solid groundwork upon which they can base their work as well (Salkind, 2016). This extensive review though must not just be a presentation of literature alone, this research should also guide the study from the larger problem to a more narrowed and nuanced issue the study plans to answer (Creswell, 2014). The research question of this qualitative grounded theory study was what tools or tool sets demonstrate the most success in completing a successful mosaic profile.

Title Searchers, Articles, Research Documents, and Journals

A search of accessible resources focusing on the identification of tools and tool sets used to perform searches capable of building an intelligence mosaic was included in this study. Primary sources of applicable literature were the online library at Capitol Technology University, the online library of Utica College, Google Scholar, ProQuest dissertation database, Worldcat, the Central Intelligence Agency Library, the Federal Register, and Justia Law. Additional sources of primary literature included Academic Search Premier, LexisNexis, Taylor & Francis, IEEE Computer Society, Homeland Security Digital Library, ACM Digital Library, Defense Technical Information Center, Google Books, and Cambridge University Press.

Key words and key phrases used to locate the relevant content for the literature review included, Mosaic Theory of Intelligence, Mosaic Theory, big data, business analytics, Freedom of Information Act, FOIA, intelligence, social network, surveillance, data mining, open government, doxing, profiling, Mosaic Profile, regulatory filings, corporate knowledge, Google Dorking, Open Source Intelligence (OSINT), artificial intelligence, machine learning, knowledge, Central Intelligence Agency (CIA), and the National Security Agency (NSA). To gather the widest possible collection of articles and papers, the usage of search variants, word variants, and search restrictions was used. Key words and phrases were chosen either due to their use in the research question or for their proximity to the underlying Mosaic Theory of Intelligence. The literature review was narrowed also to a window of the last five years for most of the content which was collected. There are some documents in this review which go back further than the five-year filter as they presented a rich context and history to the overall topic of this paper. Finally, the literature review was presented in a topical order rather chronological

order so a narrative was constructed which allowed the reader to understand each key topic in both its own context as well as the context of how the topic fits into the overall subject matter.

The Mosaic Theory of Intelligence

The third Director of the CIA, Rear Admiral Roscoe H. Hillenkoetter, identified in 1948 the basis for the Mosaic Theory of Intelligence when he said the job of an intelligence operator is to identify vital facts from the all the extraneous data and put it together like a gigantic jigsaw puzzle which presents the picture the decision makers need (Hilsman, Jr., 1952). However, over the seven decades which have passed since Hillenkoetter's definition, the Mosaic Theory of Intelligence has moved from being just a description of intelligence collection, rather it is now more often used as an argument for the justification of withholding information (Jaffer, 2010). The definition of the Mosaic Theory of Intelligence becomes very muddled when looked at historically and contemporarily.

Jaffer defined the Mosaic Theory of Intelligence in his article, The Mosaic Theory, as the justification for the government to withhold information from the public, the reason for silencing its citizens, or for the government's right to collect information which should otherwise be kept confidential (Jaffer, 2010). Whereas David Pozen defined the Mosaic Theory of Intelligence as a basic precept of intelligence gathering, collecting disparate pieces of data which have little value individually, but when joined together provide a more significant picture of a topic which is greater than the sum of its parts (Pozen, 2005). However, Pozen continued his definition to include the resulting mosaic, when undertaken by one's adversary, can prove just as dangerous in identifying one's weaknesses and vulnerabilities (Pozen, 2005). It is this two-part definition by Pozen which serves as the underlying definition for the entirety of this paper.

Historically, the Mosaic Theory has been utilized by the government mostly for the blocking of information from the American public rather than describing the collection activities of the relevant intelligence agencies (Pozen, 2005). The earliest usages of the Mosaic Theory were made by the Executive branch of the government to block the release of information requested by the public under the Administrative Procedure Act of 1946 and its successor the Freedom of Information Act (FOIA) (Pozen, 2005). FOIA was passed by United States Congress and subsequently upheld by the Supreme Court of the United States of America as serving the crystal-clear objective of piercing the secrecy veil of the Executive branch and opening its actions to the scrutiny of its citizens (Pozen, 2005). Unfortunately, the Executive branch took umbrage with this perceived invasion of their work and sought wide ranging exclusions to FOIA under numerous claims of executive privilege and national security, which Congress and the courts kept narrowing (Pozen, 2005).

The battle between the Executive branch and the Legislative and Judicial branches lasted for many years until the Executive branch won a case in 1972, United States v. Marchetti (Pozen, 2005). Chief Judge Clement Haynsworth in his opinion in United States v. Marchetti opined the significance of one piece of information while insignificant in its own right, might prove to be highly significant when viewed in the larger context, and the courts are ill-equipped to judicially review such data claims made by the Executive branch (Pozen, 2005). This case led then Professor Antonin Scalia to state the ruling is responsible for rendering FOIA as “a relatively toothless beast” (Scalia, 1982). It was also this case Pozen argued serves as the first practical argument of the Mosaic Theory of Intelligence as an acceptable defense against public disclosure as well as a limiter of judicial review (Pozen, 2005).

The first direct mention of the basic tenant of the Mosaic Theory of Intelligence came in the case Halkin v. Helms, which involved the collection of international communications by the National Security Agency (Pozen, 2005). In the ruling, the judges noted the age of computer technology had shifted intelligence gathering from cloak and daggers to a mosaic construction (Halkin v. Helms, 1978). The next case which solidified the presence of the Mosaic Theory in the realm of government openness and national security was the case of Halperin v. Central Intelligence Agency (CIA) (Pozen, 2005). In Halperin, the court solidified the stance the courts should defer to the Executive branch when reviewing cases the Executive branch claims executive privilege or national security grounds via the use of the mosaic rationale (Pozen, 2005). The final case which settled the usage of the Mosaic Theory was the 1985 case of CIA v. Sims, which concerned the request of information under FOIA on MKULTRA, which was CIA funded research into brainwashing and interrogation (Pozen, 2005). In this case the Supreme Court issued its ruling, which forms the guidance for all other courts in the United States, stating great deference must be given to the Executive branch, in this case the CIA Director, on the basis they are more familiar with the whole picture and are better judges on whether specific information presents a mosaic opportunity to our adversaries (Pozen, 2005).

This extended discussion and usage of the Mosaic Theory in judicial matters does not mean the theory is only useful in winning court cases or in successfully withholding information from the public, it has with the extraordinary advances in technology become a tool for everyone to use (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). In a 2012 competition called, The Nokia Mobile Data Challenge, researchers were challenged to use machine learning to identify certain characteristics of users based only on their Global Positioning System (GPS) and cell phone tower data (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). The researchers using only the data

provided were able to estimate a user's gender, marital status, occupation, and age (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). Additional algorithms were produced to also predict the likely location of a user in the future, but they announced when the data from a user's friends was included the reliability of the future prediction increased (Bellovin, Hutchins, Jebara, & Zimmeck, 2013).

The technological advances, as evidenced in the aforementioned Nokia Mobile Data Challenge, have expanded the mosaic theory out of the hands of nation state actors and into the hands of any individual with a little money and a little knowledge. The technological advancement which has had the most impact on the transition of Mosaic Theory usage is machine learning (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). When a computer using machine learning has been properly setup using the train and test method, where an individual trains the machine with a known data set, the machine can subsequently run largely automated and with a high level of reliability (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). Machine learning, at its core, is trained to take all of the pieces of a mosaic and put it together by identifying key dependencies, correlations, and clusters within the data (Bellovin, Hutchins, Jebara, & Zimmeck, 2013).

This expanded usage of machine learning has continued into other aspects of society and government. In Santa Cruz, California, the police are making use of machine learning and algorithms originally used to predict aftershocks from earthquakes, to now predict specific areas and specific time frames which are at the highest risk for future crimes (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). Machine learning alone does not bring the Mosaic Theory down to the level of the individual, rather it requires the additional resource of ever expanding data sets which would contain possible pieces of the overall mosaic (Bellovin, Hutchins, Jebara, &

Zimmeck, 2013). It is the combination of this proliferation of expanding data sets and the ability to analyze and cross-reference instantly the data using machine learning which has led to a new term in cybersecurity community called databuse (Wittes, 2011).

Databuse was first opined by Benjamin Wittes, in his report titled, Databuse: Digital Privacy and the Mosaic (Wittes, 2011). Wittes argues data is being used and abused by organizations far beyond what individuals believe they have agreed to (Wittes, 2011). This lack of understanding by the individuals does not stop or even slow the increasing buildup of data collection rather it contributes to the increase as users continue to share data under their misunderstood beliefs of the situation (Wittes, 2011). This shared data can be either publicly available or private, as the organizations storing the private data have permitted access to specific organizations or even have gone as far as selling said data to data merchants (Wittes, 2011).

The scale of the data we generate every day which feeds back into this idea of databuse is staggering (Wittes, 2011). On any given day, an individual can generate data from their public activities, such as using an automated toll pass system, swiping their credit card at a gas station, even driving through an intersection or down a highway which is covered by CCTVs. Unfortunately, the data is not just being generated unknowingly, rather individuals will knowingly hand out their information in exchange for a perceived reward or benefit (Wittes, 2011). Inducements for which individuals are willing to hand their data over for can range from a discount from a grocery store, to a free email account, or even for just access to a news website (Wittes, 2011). These examples of data sharing are just a small portion of the information individuals share and just a small amount of the data which data merchants will make use of to build their mosaic of a given individual (Wittes, 2011).

Wittes points out the individual has knowingly or unknowingly engaged with the Mosaic Theory and its mosaic of the individual exuberantly with the usage of social media (Wittes, 2011). When an individual wonders the web, they “like” or “recommend” things which appeal to them and may even go as far as leaving feedback or comments on a page, and by performing these actions they feed the ever-growing set of data (Wittes, 2011). Unfortunately, the individual does not stop there, they join online communities such as MySpace, Facebook, or Twitter, where they put out to the internet massive amounts of personal information (Wittes, 2011). The individual shares their work history, family members, relationships, likes, photographs, videos, even their random thoughts, and they share all of this because they want to be interesting enough to garner friends and enhance themselves (Wittes, 2011).

The reality is individuals are willing to accept the underlying idea of having mosaics created about them, such as credit reports, loyalty reward programs, and even targeted advertising, if they are receiving a perceived benefit or reward (Wittes, 2011). However, when the individual does not receive a benefit, such as a bad credit report, the individual will complain about the collection and assert claims of violation of their privacy (Wittes, 2011). In the end, the individual will accept a mosaic being created of them only if they are benefiting from said mosaic (Wittes, 2011). When the individual is not benefiting from the mosaic though, they will view the mosaic of themselves as a detriment to their personal security (Wittes, 2011).

As the tools and data to build mosaics have become more readily available, the intersection of personal security and physical security has emerged in all aspects of an individual’s life. One such case, is the Google built system called Virtual Alabama which is used for homeland security (Citron & Gray, 2013). Virtual Alabama makes use of Google’s three-dimensional satellite and aerial imagery combined with geospatial analytics and numerous data

sets to build out a real-time mosaic which will reveal relationships, trends, and patterns in the data which may be of concern to its users (Citron & Gray, 2013). Virtual Alabama is not restricted to days or weeks old data sets, but rather can track moving objects, monitor live sensors, and integrate near-real time data sets it is provided (Citron & Gray, 2013). The system pulls data in real-time from traffic cameras, public and even private video streams, GPS location of all law enforcement vehicles, schematics of many buildings, the sex offender database containing names and addresses of all registered offenders, and land deeds (Citron & Gray, 2013). Finally, all the state's 1500 public schools are linking in their live streaming cameras (Citron & Gray, 2013).

Virtual Alabama is touted by governmental officials as a great success and a critical tool in protecting the public, but some question the motives of the government and whether the United States is becoming a surveillance state driven by the Mosaic Theory (Citron & Gray, 2013). This push toward a greater use of the Mosaic Theory can be seen in the building out of fusion centers by federal, state, and local governments (Citron & Gray, 2013). Fusion centers are built to operate on the Mosaic Theory as evidenced by the services they consume such as, criminal records, social security numbers, property records, car rentals, credit reports, postal and shipping records, utility bills, gaming records, insurance claims, social network activity, drug store records, grocery store records, biometric data, fingerprints, facial recognition profiles, law enforcement surveillance records, law enforcement cameras, and so much more (Citron & Gray, 2013). Finally, some even believe fusion centers have access to broadband providers' records which show each subscriber's online activities and communications (Citron & Gray, 2013).

This overwhelming large collection of data on individuals raises many concerns over privacy and the constitutional rights of citizens. This privacy concern raises the underlying

question of the ethics of the individuals making use of the collected data or guiding the collection of data (Citron & Gray, 2013). In 2012, the U.S. Senate's Permanent Subcommittee on Investigations reported on worryingly frequent internal Department of Homeland Security warnings about fusion centers being used to carryout surveillance on individuals specifically aimed at their activities protected under the first amendment (Citron & Gray, 2013). These warnings covered such prohibited activities as using law enforcement to keep track of political bumper stickers and the owner of the vehicles the bumper stickers were on to the reporting of individuals who attended a talk on marriage and Islam at a mosque (United States Senate, 2012).

This perceived violation of one's privacy though does not end with the individual making use of the mosaic data, rather it has permeated itself throughout the government and law enforcement as well. In the United States v. Jones, five of the Supreme Court justices wrote citizens should be able to reasonably expect privacy even in the collection of data, accepting limited surveillance may be reasonable if one is suspected of a crime, but not acceptable if the surveillance goes on for an extended period of time (United States v. Jones, 2012). United States v. Jones involved the usage of a GPS tracker on a suspect's car for over a one month period (United States v. Jones, 2012). Justice Sotomayor specifically noted the extended collection of GPS data was the equivalent of the Mosaic Theory, since by looking at the collected data one could discern personal and private information (United States v. Jones, 2012). She noted certain trips such as those to a psychiatrist, an abortion clinic, an AIDS treatment center, a strip club, a by-the-hour motel, a gay bar, etc. can all be used to identify personal aspects of the individual which are protected under said individual's right to privacy (United States v. Jones, 2012). The problem thusly presented is to what extent does an individual have to give up certain liberties and rights to remain safe. Additionally, does an individual have to accept the seeming

inevitability of the loss of privacy to the machinations of internet in order to access a website or an online tool, do they have to accept the creation of a digital mosaic of oneself (Citron & Gray, 2013).

Doxing

Doxing is the act of intentionally and publicly releasing personal information on the internet for the express purpose of intimidating or punishing the victim of the doxing (Douglas, 2016). Doxing is a term developed by hackers as part of their leetspeak, or cultural lexicon, which represents the dropping of documents on an individual as a form of revenge (Douglas, 2016). Doxing, at its core intent, is illegal, but there are instances in which private information is readily available through public databases and thusly possibly legal (Coleman, 2014). Additionally, the act of doxing and in specific targeting doxing, is not always 100% reliable or accurate as witnessed in the 2011 doxing debacle by HBGary Federal and it's CEO Aaron Barr (Olson, 2012). Mr. Barr chose to perform an intelligence gathering exercise in retribution for attacks Anonymous had carried out, and further he did it for the express purpose of gaining a reputation and financial reward for the company as HBGary Federal needed money (Olson, 2012). Anonymous, upon learning of the claims of Mr. Barr, decided to socially engineer Mr. Barr into linking a key member of Anonymous to a completely innocent individual (Olson, 2012). This false identification proved to be ruinous to HBGary Federal and Mr. Barr's reputation (Coleman, 2014).

The form of targeting doxing is of particular interest in this paper as it focuses on disclosing personal information an individual expects to remain private, obscure, or obfuscated (Douglas, 2016). Targeting doxing seems to have a close correlation to the Mosaic Theory of Intelligence in the underlying methods of both activities is the discovery of personal information

to build out a profile of a given individual. It is seems to be the end goal of the two activities which diverge in the actions taken by the profile builder.

Carrie Gates and Peter Matthew in their paper, Data is the New Currency: Becoming a Data Whore, noted doxing has become a cornerstone of the online black market with the average cost for a doxing ranging between $25 and $100 depending on the amount of personal information included (Gates & Matthews, 2014). Gates argues the idea one's identity, not just personally identifiable information (PII), has a value in the online black market (Gates & Matthews, 2014). This is evident in her argument that the more information collected and identified on an individual the higher the cost for the information (Gates & Matthews, 2014). The basic precept which can be drawn from this is the more data one can collect on an individual to build out the most complete profile possible will allow one to reap a bigger reward.

Doxing is not restricted to its use in the dark web economy, it has been used numerous times by hackers and hacker groups seeking revenge for perceived wrongs one may have committed against them (Mathews, Aghili, & Lindskog, 2013). One such example is the August 2, 2011 Federal Bureau of Investigations (FBI) Intelligence Bulletin on Cyber Intelligence which was sent to all law enforcement agencies alerting them to the doxing of law enforcement officers by the hacking groups, Anonymous and LulzSec, in retaliation for the increased activities by law enforcement aimed at the groups (Mathews, Aghili, & Lindskog, 2013). When a hacktivist has chosen to dox an individual, they make use of numerous tools which have been previously mentioned, such as public records, social media networks, governmental agencies, browsing history, GPS data from geo-location services, etc. (Mathews, Aghili, & Lindskog, 2013).

Sony Pictures Entertainment in late 2014 was the victim of a wide ranging cyber-attack which made use of numerous attack vectors, such as distributed denial of service attacks, website

defacement, network intrusion, threats of real world attacks, but the most damaging of these vectors proved to be the doxing (Haggard & Lindsay, 2015). The Guardians of Peace (GoP), a North Korean based group of hacktivists, attacked Sony Pictures Entertainment in response to a film which was due to be released within weeks of the attack, The Interview (Haggard & Lindsay, 2015). The Interview was a satirical comedy which posited the assassination of the North Korean leader Kim Jong Un by a fictional talk show host Dave Skylark (Haggard & Lindsay, 2015). The GoP's doxing attack involved a significant collection of information such as internal emails, financial records, film contracts, PII on actors and actresses, employee health records, and much more (Haggard & Lindsay, 2015). The doxing was damaging not only from the stand point of all of the trade secrets, contract negotiations, and scathing emails being released, but in the fact the GoP was able to release the information over a period of a few weeks with each week's release building on top of the releases from the previous week (Haggard & Lindsay, 2015). This form of structured doxing allowed for the attacker to land multiple punches against Sony instead of one big doxing (Haggard & Lindsay, 2015).

Big Data

Big data encompasses a wide range of technologies, schemas, and approaches to data, but at its most basic is the storage and analysis of large complex data sets making use of various data techniques such as NoSQL, Map Reduce, or machine learning (Ward & Barker, 2013). In its most general form, big data is focused on two distinct topics, data storage and data analysis (Ward & Barker, 2013). The term big data has a rich history in trying to identify its etymological origins as evidenced in a newspaper article by Steve Lohr of the New York Times (Lohr, 2013). The problem lies in the fact the term and the definition are separate from each other, the technology and software are encompassed by the term big data existed well before the term itself

(Lohr, 2013). It is with this understanding the earliest attribution to the coining of the term big data lies with John Mashey, of Silicon Graphics, one of the largest high power computing firms at the time. Based on this etymological history it is arguable the definition of big data was purposely designed to encompass many subtopics within it, such as retail analytics, search tracking, ad tracking, data mining, and data merchants, all of which are covered below.

Retail analytics is the progression forward of retail groups, credit card companies, and data miners to mine the transactional data present in their historical databases to identify trends, forecast stock levels, and even predict an individual's purchases or needs (Duhigg, 2012). Neil Ashe, Walmart's CEO of global e-commerce stated in a 2013 speech, "We want to know what every product in the world is. We want to know who every person in the world is. And we want to have the ability to connect them together in a transaction." (Neef, 2014, p. 146). Nordstrom and Home Depot are two other retailers who wish to know where their customers are at all times when they are in their stores (Goodman, 2015). These two companies are just one of many retailers using technology developed by Euclid which tracks all customers in a retailer's store by accessing the WiFi in their cellular phones and getting the mac address (Goodman, 2015). Euclid's software can discern and track individuals both in real time and historically (Goodman, 2015). This allows for the retailer to understand buying patterns, customer movements, and much more (Goodman, 2015).

Walmart is, as evidenced by the words of the CEO of global e-commerce, making significant strides in their collection and use of big data. Walmart processes more than one million transactions every hour and had a database estimated at 2.5 petabytes as of 2012 (Marr, 2015). Historically, Walmart has been ahead of the curve in retail analytics, as evidenced by their 2004 study of sales and stock data after Hurricane Charley to determine what would be needed

before the next storm, Hurricane Frances (Hays, 2004). Walmart used their predictive technology to determine what items would be purchased and were surprised to find strawberry Pop-Tarts sales increased by a factor of seven and beer was the top-selling item store wide (Hays, 2004).

Walmart has gone even further than most retailers with the creation of a subsidiary called @WalmartLabs which is tasked with using its Big Fast Data Team to find new ways to mine the data the retailer has on customers to increase sales (SAS Institute, 2016). One such innovation from this group was the new search engine system which runs on Walmart.com called Polaris (Walmart Corporation, 2012). Polaris was developed using data from Walmart's Social Genome project which the retailer built to use data from social networking sites to better understand the relationships between customers and products based on semantics and syntax analysis (Walmart Corporation, 2012). By using this data, programmers were able to better code the search engine to produce more relevant results to a customer's search query (Walmart Corporation, 2012). Initial testing of the Polaris search engine showed a 10-15% increase in shoppers completing purchases which is a significant reduction in cart abandonment (Walmart Corporation, 2012).

The search tracking and ad tracking topics are combined due to the close correlation between the two when looking at the world's most popular search engine, Google (Goodman, 2015). Google started as a simple website with a basic premise to allow individuals to search the internet and get the most relevant results to one's query (Goodman, 2015). Unfortunately, Google found running a search engine was not cheap and needed to monetize their traffic, which resulted in the creation of Google Adwords (Goodman, 2015). What most people did not realize then or even now is Google's search engine has shifted from being a simple search website into something more akin to a data hog, as it tracks every single search done by an individual to build

and develop a marketing profile of said individual with its own unique identifier (Goodman, 2015).

This unique identifier within Google is where the tracking takes place. As an individual moves from one Google product to another, it passes the unique identifier along so an individual's activities in other Google product can be tracked as well (Goodman, 2015). One such product is the free email system Google provides all its users called Gmail (Goodman, 2015). By providing individuals with a free email account with a vast amount of storage, Google convinced millions of users to move to their mail system, which just happened to allow Google to scan email content at said time to further build out one's marketing profile (Goodman, 2015). The use of this unique identifier continued to all aspects of Google's ecosystem of tools, Google Contacts and Google+ allow an individual to connect with one's friends but also lets Google identify one's social network, Google Maps gives users free GPS and step by step driving directions but also provides Google with the knowledge of where everyone goes every day, even the Android operating system which now powers a significant portion of the mobile phones in use today, and many more of the Google tools used each day (Goodman, 2015). It is this collection of highly refined and highly detailed data Google has on each individual which allows Google to operate as one of the world's largest companies, because it has the holy grail of marketing and advertising a detailed and accurate profile built by every individual personally (Goodman, 2015).

But Google did not stop at just using Google's own sites and tools to their advantage, they moved out into the world of advertising and ad tracking on the web with their AdSense network (Liu, Sheth, Weinsberg, Chandrashekar, & Govindan, 2013). With AdSense, Google could get website owners to join into their ad network and insert code to serve ads to the visitors

of those websites, while at the same time using the previously mentioned unique identifier to track their movements across these websites (Liu, Sheth, Weinsberg, Chandrashekar, & Govindan, 2013). This ad tracking is what allows Google to provide the depth of detail as the software is in essence tracking a significant portion of the individual's web browsing history (Liu, Sheth, Weinsberg, Chandrashekar, & Govindan, 2013). The end result is as a user goes between websites, the ads which are served on the website are not generic and correlated to the content on the website but rather the ads are based on the user's most recent web searches, emails, and browsing history (Liu, Sheth, Weinsberg, Chandrashekar, & Govindan, 2013).

Google and Walmart are not alone in using the data they have collected from their customers or users, virtually all businesses of a decent size make use of the art of data mining. One such business to make use of data mining to enhance their "product offering" was the Major League Baseball club the Oakland Athletics (A's) (Marr, 2015). Bill James, a baseball talent advisor, developed a hypothesis that with enough data and data points, undervalued talent could be identified (Marr, 2015). James took this idea to the general manager of the Oakland A's, Billie Beane, and together the two tweaked and experimented with various data mining techniques to choose a team of undervalued talent (Marr, 2015). The result of this experiment was the Oakland A's was able to field a team of undervalued players which made the play offs in both 2002 and 2003 even though they had the third lowest payroll in the entire league (Marr, 2015).

Data mining can go even further than identifying undervalued baseball talent, it can be used to identify highly sensitive personal attributes (Marr, 2015). In 2013, Cambridge University and Microsoft Research Labs carried out a study focused on identifying personal attributes of individuals solely based on their Facebook Likes (Marr, 2015). The study made use of Like data

from 58,000 volunteers and was able to identify attributes such as race, religion, gender, political affiliation, sexual orientation, relationship status, and illegal substance use as well as traits such as intelligence, emotional state, outgoing, openness, and conscientiousness (Kosinski, Stillwell, & Graepel, 2013). Some examples of this identification showed a Like for swimming, Jesus, Pride and Prejudice and Indiana Jones produced a prediction the individual was satisfied with life, versus a Like for So So Happy, Dot Dot Curve, Girl Interrupted, The Adams Family and Kurt Donald Cobain produced a prediction the individual was emotionally unstable or neurotic (Marr, 2015).

The key piece of data from this study though is what level of accurateness was this form of data mining able to achieve. According to the study summary, the model was able to correctly identify a man's sexual orientation with an 88% accuracy, identify the race of an individual with a 95% accuracy, or identify the political affiliation of an individual with 85% accuracy (Kosinski, Stillwell, & Graepel, 2013). The study concluded Like data is not the only data set which is susceptible to this reverse identification of sensitive personal attributes, other data sets susceptible to this method could be browsing histories, search queries, or purchase histories (Kosinski, Stillwell, & Graepel, 2013). These data sets are data sets which have been covered in previous discussions on big data in this paper. It is this fact which brings the literature review to its next topic, data merchants.

Data merchants are a group of virtually unknown corporations whom act as data aggregators shadowing every aspect of an individual both online and in the real world (Etzioni, 2012). These corporations have few laws to restrict what they can gather and sell to any paying customer, which normally only cover medical and financial records (Etzioni, 2012). However, these corporations have figured out ways around these restrictions by using a method of using

innocent facts to extrapolate medical and financial details (Etzioni, 2012). One example showcases this is parsing one's web browsing history to identify key words, such as searching the word depression on a website, or browsing medical websites or support chat groups for individuals with depression (Etzioni, 2012). These corporations would be able to thusly tag one's record in their database as being depressed and not be in violation of federal or state laws (Etzioni, 2012).

In 2014, there were an estimated 3,500 data aggregators also called data merchants in business (Gates & Matthews, 2014). Acxiom Corporation is one such data merchant in the United States where they are based in Conway, Arkansas (Singer, 2012). In 2012, the New York Times reported Acxiom had the largest database on consumers in the world and was integrating more than 50 trillion data transactions per year (Singer, 2012). The company further reported they had data profiles of more than 500 million active consumers, with each data profile for each consumer containing approximately 1,500 unique data points (Singer, 2012). In a throwback to the study performed by Cambridge University and Microsoft Research Labs, Acxiom also assigns every consumer in their database to a set of socioeconomic clusters using their classification system PersonicX, which looks at an individual's thousands of data points and discerns for its customers what type of promotion or marketing approaches will work best to convert said individual into a customer (Singer, 2012).

Acxiom and other data merchants would not be viable businesses if they were selling something no one wants, but it is the reverse which is true, clients are clamoring for their product (Singer, 2012). In 2012, 47 of the Fortune 100 companies as well as the United States Federal Government had contracts with Acxiom (Singer, 2012). Acxiom's customers such as Wells Fargo, Toyota, Ford, Macy's, etc. are able to access a Consumer Data Products Catalog on their

client website and choose what data group they are looking for (Singer, 2012). This catalog offers demographics, physical traits, behaviors, medical, financial and many other searchable factors with categories like, Christian families, smokers, or something more esoteric like Latinos with an age between 25-34 who live within a two-hour drive of a mountain and are avid sports fans (Singer, 2012).

The questions big data and its subtopics raise are the same questions which are raised by mosaic profiles produced using the Mosaic Theory of Intelligence. These questions revolve around the loss of privacy of an individual (Etzioni, 2012). No longer can an individual protect their privacy by curbing the government, as the government has become clients of the data merchants (Etzioni, 2012). When one considers the data which is collected by these data merchants, one should be chilled by the possible uses someone can make of the data they purchase, such as one's political or social views (Etzioni, 2012). Unfortunately, that someone who becomes a client of one of these data merchants could be any individual, organization, or governmental body with the money to pay for access to these services (Etzioni, 2012). Therefore, one should assume the data they believe to be private and personal is no longer private or personal, it is now just another piece of data available to the public (Etzioni, 2012).

Law Enforcement and Surveillance

One of the customers of these data merchants are the government and by extension law enforcement (Etzioni, 2012). ChoicePoint, one of the larger data merchants, had as of 2015 at least thirty-five contracts with government agencies such as the Department of Justice, the Federal Bureau of Investigation, the Drug Enforcement Agency, the Internal Revenue Service, and the Bureau of Citizenship and Immigration Services (Etzioni & Rice, 2015). A 2006 government study found fifty-two federal agencies had launched at least 199 data-mining

projects which relied on commercial data merchants (Etzioni & Rice, 2015). The fact is law enforcement surveilling an individual is not a new tactic as surveillance has been used for centuries (Sagar, 2015). In 1653, the English Commonwealth established a Secret Office in their Council of State which was charged with looking for suspicious letters being posted and copying them (Sagar, 2015). Many other countries followed this model and created identical institutions in their own countries, such as the United States (Sagar, 2015).

In the United States, the United States Court of Appeals for the Sixth Circuit ruled in 2010, law enforcement could not perform these same type of activities when law enforcement had carried out a warrantless search of about 27,000 emails of a suspect as it was a violation of one's reasonable expectation of privacy even though it was stored in a third party service provider (Carter, 2015). The court's decision basically affirmed email is afforded the same protections under the Fourth Amendment as other traditional forms of communication (Carter, 2015). The reality though is this case and United States v. Jones are just two cases against the over extending of law enforcement when it comes to data mining.

One of these instances where the government and law enforcement has extended their reach into data mining to carry our surveillance is Virtual Alabama. Virtual Alabama, which was mentioned earlier, was produced by Google for homeland security (Citron & Gray, 2013). The system pulls data in real-time from traffic cameras, public and even private video streams, GPS location of all law enforcement vehicles, schematics of many buildings, the sex offender database containing names and addresses of all registered offenders, and land deeds (Citron & Gray, 2013). Finally, all the state's 1500 public schools are linking in their live streaming cameras (Citron & Gray, 2013). The end result of this combination of data, is law enforcement

individuals can actively surveil individuals from anywhere by accessing this real-time system (Citron & Gray, 2013).

Another instance is the Santa Cruz Police Department which has implemented a predictive policing system to assist them with a 30% increase in service calls and a 20% decline in staffing (Greengard, 2012). Every day the predictive policing system produces 10 hot spot maps for the patrol officers (Greengard, 2012). These maps are produced based on historical police records which have been geocoded (Greengard, 2012). In the first six months of the program, Santa Cruz reported a 15% decrease in burglary and property theft crimes over the previous year (Greengard, 2012).

But Santa Cruz is just one of many cities who have moved toward this predictive policing. The Los Angeles Police Department in conjunction with the University of California, Los Angeles (UCLA), has created a separate division called the Real Time Analysis and Critical Response Division (Ferguson, 2012). This division makes use of three years' worth of criminal activity, specifically burglary, automobile theft, and theft from automobiles (Ferguson, 2012). The system weights the data based on newest to oldest crimes and creates 500 feet by 500 feet grids where the highest rate of suspected or predicted crime will take place (Ferguson, 2012). The reality is with the consistently shrinking municipal and state budgets, the government needs to find ways to be more cost effective with their policing resources (Ferguson, 2012). Due to this reality, law enforcement is embracing predictive policing, data merchants, and increasingly high tech methods of surveillance (Ferguson, 2012). Trends can already be seen progressing toward the usage of data mining, mosaic profiles, and predictive policing to track not just hot spots of crime, but individuals as well, as evidenced by the increasing use of predictive recidivism

evaluation systems to determine whether an individual gets early release or parole (Ferguson, 2012).

Chicago's police department has implemented one such system, which they call their "heat list", this list tracks individuals whom the computers determine via risk analysis and other methods as being the most likely to commit an act of violence (Joh, 2014). However, New York City puts all of these systems to shame with their Domain Awareness System (DAS) which was co-developed with Microsoft (Joh, 2014). DAS collects and analyzes data in real time across the whole of New York City using over 3,000 surveillance cameras, over 200 automatic license plate readers, over 2,000 radiation sensors, and information from databases such as the license plate tracking system, the police database and other databases (Joh, 2014). The goal of DAS is similar to the goals of Walmart with their databases, the police want to know in real-time who or what poses a threat and any connections between persons, items and places (Joh, 2014). The system is built to carry out tasks, such as, automatically detecting unattended bags, tracking where a suspect's car has been over the past few months, or in real time identify the presence of any vehicle which is linked to an individual on a watch list and feed information to an officer from where the car is currently located, where it has been, and even a full profile of the individual including their criminal history (Joh, 2014).

It is the intersection of these technologies and the usage of such technologies by law enforcement which have caused law enforcement, surveillance, and predictive policing to run afoul of Fourth Amendment privacy rights and the Mosaic Theory of Intelligence (Joh, 2014). This intersection came to a head in the 2012 Supreme Court case of United States v. Jones (Joh, 2014). In United States v. Jones, police had used a GPS tracking device on the defendant's car for a period of twenty-eight days tracking all the defendant's movements, public and private

(Joh, 2014). This extended period of data collection caused concern for at least five justices that the Mosaic Theory, normally used by the federal government to restrict access to government data, has moved into the realm of daily life and was available with publicly available technology (Joh, 2014).

Justice Sotomayor specifically noted the extended collection of GPS data was the equivalent of the Mosaic Theory, since by looking at the collected data one could discern personal and private information (United States v. Jones, 2012). She noted certain trips such as those to a psychiatrist, an abortion clinic, an AIDS treatment center, a strip club, a by-the-hour motel, a gay bar, etc. can all be used to identify personal aspects of the individual which are protected under said individual's right to privacy (United States v. Jones, 2012). A 2010 paper by Chaoming Song and other researchers noted using a set of cell phone tower data points, a rough equivalent of GPS, researchers were able to predict future movements with an accuracy of 93% (Song, Qu, Blumm, & Barabasi, 2010). In a 2013 paper by Yves-Alexandre and other researchers found with as little as four cellular phone tower data points, the researchers could uniquely identify 95% of the individuals (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013). This shows small amounts of data as well as small pieces of data can be joined together programmatically, algorithmically, or mentally to create a rich mosaic profile of an individual (Bellovin, Hutchins, Jebara, & Zimmeck, 2013). It is with this knowledge the next topic is discussed, Social Networks.

Social Networks

Social media is the local five and dime of the 21st century, a gathering place for individuals of all ages, races, genders, politics, etc.. These gathering places are growing every day in leaps and bounds, one service, Twitter, generates nearly 400 million tweets every single

day (Bello-Orgaz, Jung, & Camacho, 2016). One journal article even recommends the creation of a new term, social big data, as data from social network is exploding in its daily generation (Bello-Orgaz, Jung, & Camacho, 2016). It is this explosion of data which has made the mining of social media data a top priority for all types of organizations (Bello-Orgaz, Jung, & Camacho, 2016).

A study of Twitter utilization for marketing metrics was carried out and they identified 19% of Twitter accounts mention a brand name product and of said group nearly 20% expressed an opinion or sentiment on said product (Bello-Orgaz, Jung, & Camacho, 2016). It is from this data analysis, which gathers up millions of individual's personal tweets, that marketing groups then advise their clients on their approaches to specific targeted consumer groups (Bello-Orgaz, Jung, & Camacho, 2016). A study done in 2010 showed over 40% of social network users had posted private information on themselves to a social network (Hajili & Lin, 2016). An additional study in 2011 quantified the disparity between an individual's desired privacy settings in Facebook versus the actual privacy settings in Facebook, and the study found only 37% of the time did the desired privacy settings match the actual privacy settings (Hajili & Lin, 2016).

In the book, Social Media as Surveillance: Rethinking Visibility in a Converging World, the author argues an individual's social media profile is a digital representation of the individual's human body (Trottier, 2016). The digital body is generated by first filling up the repository, in this case an individual's social media profile, with personal and private information and secondly by stepping outside of the digital home and creating social linkages or friends (Trottier, 2016). This act of friending results in the production of even more personal data as it identifies, one's friends, one's family, one's significant other, one's children, or even celebrities one admires (Trottier, 2016). Unfortunately, the digital body needs to express an

individual's thoughts, actions, activities, likes, dislikes, etc. into the social group one built, so one's friends and family now how the individual is doing, where the individual is going or even who the individual is dating (Trottier, 2016).

These activities are essential to the social networking site as it tries to build a profile of an individual for marketing purposes, so it encourages more and more interaction from the digital self (Trottier, 2016). Each one of these interactions are seemingly benign when viewed individually, but it is the combination of these digital puzzle pieces which when put together present a chilling mosaic of the real human being and one's personal and private life (Trottier, 2016). However, as these social networks become larger and more integrated into society, the act of stepping away from a social network becomes harder as individuals expect one to interact with them via these social networks rather than via a phone call or a letter (Trottier, 2016). This reinforcing behavior is designed by the social networks for the purpose of keeping an individual involved in social media and to keep sharing more pieces of data with the company (Trottier, 2016). It is not just the peer pressure which is aimed at keeping an individual involved on social media but even the advertising is aimed at keeping one on the site and sharing by encouraging one to make new friends, reunite with old school buddies, or share news and details with one's geographically distant family and more (Trottier, 2016).

The social media networks, such as Facebook, see all of this data individuals have provided as a veritable gold mine, and this belief is not just a turn of phrase as they will mine an individual and an individual's information to make every cent possible off of the individual (Trottier, 2016). Facebook offers a set of purchasable "business solutions" to any company wanting to utilize their social network (Trottier, 2016). These solutions offer companies access to an extremely detailed and in-depth group of individuals who meet their marketing profile, but it

is not just the data an individual shares which Facebook is selling but knowledge their systems are able to identify or mine by looking at the whole of an individual's digital body (Trottier, 2016). However, it is not just businesses using the data provided on social media to their advantage, law enforcement has also been getting involved more heavily in recent years. In 2011, there was a riot in Vancouver and virtually everyone at the riot posted videos and photos from the riot, so the police took advantage of social media and used those videos and photos to run facial recognition and crowdsourcing to identify quite a large number of the rioters (Trottier, 2016).

The reality of social media is anyone and everyone has access to one's digital body and the information it contains. A group of tools have been created to allow for individuals to dig deeply into an individual's online life, a few of these tools are fbStalker, GeoStalker, and Cree.py (Ruslanovich & Alekseevna, 2016). fbStalker collects information on an individual from their Facebook page, gathering data such as videos, photos, posts, comments, notes, date and time stamps, and geolocation of all data (Ruslanovich & Alekseevna, 2016). GeoStalker collects information on an individual from multiple sites such as Foursquare, Instagram, Flickr, etc. gathering data such as network data, geolocation of data, photos, profiles, and notes (Ruslanovich & Alekseevna, 2016). The last of these tools is Cree.py, an open source tool, which is built to collect geolocation data from data on an individual's Tweets, Instagram posts, Google+ posts, and Flickr posts and present them on a Google map so an individual could learn an individual's home address, work address, or where the individual tends to go (Ruslanovich & Alekseevna, 2016).

The fact is when one applies the Mosaic Theory to social media, the result is a mosaic which is much more detailed and more invasive of one's privacy then even Justice Sotomayor

envisioned in her opinion against long term surveillance (United States v. Jones, 2012). Unfortunately, this enhanced invasion of privacy is perfectly legal under the third-party doctrine as data is stored for varying lengths of time by internet service providers (ISPs) as it is transmitted between the sender and the receiver (Tokson, 2011). The same loss of privacy applies when the individual agrees to let Facebook retain the data provided on their servers as again it involves a third-party, Facebook (Bedi, 2014). This loss of privacy to one's social media data is proving to be a boon for the next topic of the literature review, the United States Intelligence Agencies (Bedi, 2014).

United States Intelligence Agencies

The various intelligence agencies of the United States government are many, with the current number standing at sixteen agencies and the Office of the Director of National Intelligence (Richelson, 2015). The role and responsibilities of each of these agencies varies but the underlying goal of each is to provide the necessary intelligence needed for the stakeholder/decision maker to make a decision based on a thorough understanding and context of a situation (Richelson, 2015). There are four activities every intelligence agency carries out, collection, analysis, counterintelligence, and covert action (Richelson, 2015). The first two of these activities are the activities covered primarily in this literature review.

The first of these activities is collection which can itself cover many areas such as open source collection, human source collection, interrogation, and technical collection (Richelson, 2015). Technical collection is the particular activity on which the Mosaic Theory thrives as it involves the collection from all different types of electronic information (Richelson, 2015). The intelligence community when it comes to technical collection has taken the approach of "Collect it all, process it all, exploit it all, sniff it all, know it all" (Hu, 2015). It is this mantra toward

collection which pulls the intelligence communities into the discussion of big data and the Mosaic Theory (Crampton, 2015).

What goes into this Intelligence Community (IC) big data system via collection activities? The reality is there is more collected then could be covered in this paper, as one defense contractor calls it limitless intelligence, so we will just peek at a few items (Crampton, 2015). The National Security Agency (NSA) captures and analyzes the full content of all phone calls in at least two countries, the NSA also analyzes metadata of numerous communications and one such processing system, SHELLTRUMPET, processed its one trillionth metadata record on December 31, 2012 (Hu, 2015). Another NSA project PRISM upstream collected in the first six months of 2011 approximately 13.25 million internet transactions (Donohue, 2015). Finally, an unnamed program collected data flowing between Google and Yahoo datacenters in the United States and the United Kingdom (U.K.) which was estimated to have collected hundreds of millions of records, but this program was carried out by the U.K.'s Government Communications Headquarters (GCHQ) to allow the United States intelligence agencies to be protected.

The reality is the collection of data has grown by leaps and bounds as the amount of data produced has grown. In 2012, it was estimated 2.7 zettabytes of worth of data was being stored across the world (Joh, 2014). IBM stated in 2011 the ninety percent of the world's data had been generated in the last two years (Joh, 2014). Google stated we create more data in two days than all data from the beginning of human civilization to 2003 (Joh, 2014). Even the Library of Congress started a program to archive all public tweets in 2010 and by January 2013 had collected 170 billion tweets (Joh, 2014).

It is the idea everything must be collected which has begun guiding our intelligence agencies. In 1982, one intelligence official noted everybody with digital communications are a

target, and the agencies have been hard at work collecting their data (Hu, 2015). The Chief Technology Officer of the Central Intelligence Agency (CIA) pushed this same approach toward data and the Mosaic Theory when he stated in an interview in 2013 the value of a piece of data is not known until you can connect it with another piece of data you collect in the future (Hu, 2015).

As Pozen noted a basic precept of intelligence gathering is the idea of disparate items of information taking on added importance when combined (Hu, 2015). The NSA had gotten a blanket order issued by the Foreign Intelligence Surveillance Court (FISA) which required Verizon Communications to provide the NSA with the call records of millions of Americans (Mornin, 2014). The NSA argues under the "hop" or "chain" analysis method they use, the NSA can collect and review not just a suspect's phone records, but also the phone records of everyone he calls, everyone who calls those people and everyone who calls those people (Mornin, 2014). This means if a suspect called 40 unique individuals, the three hops analysis would allow the NSA to review the records of 2.5 million individuals (Mornin, 2014). Without knowing or inferring the intent of the NSA, one can simply look at these numbers and this three hops analysis and see the combination of data done under the guise of national security has turned the Mosaic Theory against one's social interactions and activities.

Some additional tools which perform functions making use the Mosaic Theory are listed here. The first is from the NSA and is called Social Network Analysis Collaboration Knowledge Services (SNACKS) which uses texts to construct the hierarchy of organizations and their personnel (Hu, 2015). Another tool is a GCHQ program called TEMPORA, which buffers and retains about 25% of the internet traffic in the U.K. for up to thirty days in a database which allows an analyst to search the communications, search terms, browsing habits, and more on any

individual (Hu, 2015). Another tool from the NSA is XKEYSCORE, which allows near real-time querying of all data the NSA has on the internet traffic of a given individual including emails, websites, searcher, and more (Hu, 2015). DISHFIRE is a NSA database which stores many years' worth of text messages from across the world, collecting on average almost 200 million text messages a day (Hu, 2015). TRACFIN is a NSA database which stores credit card purchase histories (Hu, 2015).

MARINA is a NSA database which stores the metadata on millions of internet users for as long as a year at a time (Hu, 2015). This metadata allows for the NSA to gather an individual's personal information and develop a mosaic profiles of the individual (Hu, 2015). The MARINA system will even begin the building of a patter-of-life analysis (Hu, 2015). A pattern-of-life analysis is yet another form of intelligence collection which is in the same vein of a mosaic profile, as it uses collected surveillance data document and discern an individual's habits, which can then be used to predict actions they may carry out in the future (Franz, 2017). One could draw the inference from looking at these tools which the various programs mentioned are indicative of the Mosaic Theory concerns which Justice Sotomayor referenced in United States v Jones (Gentithes, 2015). This concern is concerning enough one should review the current laws which are affected by this increased collection methodology.

United States Constitution and Federal Laws

The Constitution of the United States of America has served as the bedrock of the country for over 220 years. However, the times have changed and technology has progressed and we are faced with a modern world being guided by a 220-year-old document (Baggett, Foster, & Simpkins, 2017). Thomas Jefferson, the writer of the Constitution, addressed this when he argued laws must change to keep pace with the progression of mankind (Baggett, Foster, &

Simpkins, 2017). It is these laws and interpretations which must be reviewed to understand where the government stands on the Mosaic Theory of Intelligence.

The First Amendment provides United States citizens with the freedom of religion, the freedom of speech, the freedom to assemble and the freedom to petition the government (Bedi, 2014). However, some judicial scholars argue the era of big data and surveillance are in violation of those basic rights, as the sustained surveillance and mass collection of data can be used to interfere with one's intellectual freedoms and thus the first amendment violation (Citron & Gray, 2013). The belief of these scholars is an unchecked usage and collection of data such as internet search histories, email, web traffic, and telephone calls should rise to the level of a first amendment rights violation and these types of data requests must be managed and controlled by the court (Citron & Gray, 2013). This belief has a base of data upon which it is built, in 2012, the U.S. Senate's Permanent Subcommittee on Investigations reported on worryingly frequent internal Department of Homeland Security warnings about fusion centers being used to carryout surveillance on individuals specifically aimed at their activities protected under the first amendment (Citron & Gray, 2013). These warnings covered such prohibited activities as using law enforcement to keep track of political bumper stickers and the owner of the vehicles the bumper stickers were on to the reporting of individuals who attended a talk on marriage and Islam at a mosque (United States Senate, 2012).

The Fourth Amendment provides United States citizens with the right to be secure in their persons, houses, papers, and effects against unreasonable search and seizures and further no warrant should be issued without a basis for probable cause (Bedi, 2014). Unfortunately, the Fourth Amendment does not address the distributed nature of technology, so the courts rely currently upon what has been termed the third-party doctrine, which states information freely

given by an individual to a third party may be disclosed to the government without a resulting violation of the Fourth Amendment (Schlabach, 2015). This third-party doctrine is what has allowed the continuing erosion of one's Fourth Amendment rights where technology is involved in today's world as virtually all communications rely upon a third party (Schlabach, 2015). A large collection of legal scholars as well as even a Supreme Court Justice have discussed the removal of the third-party doctrine from any discussions involving technology (Schlabach, 2015).

Supreme Court Justice Sotomayor in United States v. Jones opined a reconsideration may be warranted of the premise in which an individual relinquishes their rights to privacy when voluntarily disclosing information to third parties (Schlabach, 2015). Justice Sotomayor further argued the current interpretation of the third-party doctrine would allow any governmental body to access without warrant an individual's cache of phone numbers, e-mail addresses, web browsing history, and even shopping lists stored online (Schlabach, 2015). It is this overly broad interpretation of the third-party doctrine which has lead Justice Sotomayor to infer in her opinion in United States v. Jones that the Mosaic Theory of Intelligence could apply to many other technologies than just GPS (Schlabach, 2015). The concurrence opinion by Justice Sotomayor in United States v. Jones has already had some effect on subsequent rulings by the United States Supreme Court in their ruling in Riley v. California (Schlabach, 2015). In this case, Chief Justice Roberts ruled cellular phone searches require a warrant as more than 90% of US citizens carry their phone on themselves as a container of their digital records and the quantity of data makes the device qualitatively different from physical records in one's possession (Schlabach, 2015).

These rulings are based in part on the case United States v. Warshak, which held a user enjoys an expectation of privacy in their emails, regardless of the fact the commercial internet

service provider may transmit or store said email (Kugler & Strahilevitz, 2015). This case argued and was concurred that as forms of communication change so should the Fourth Amendment (Schlabach, 2015). These rulings are in direct conflict with the Stored Communications Act, which made it legal for government officials to get the contents of an individual's online communications without a warrant, and thusly in United States v. Warshak the Stored Communications Act was deemed unconstitutional (Schlabach, 2015). Further complicating the issues present in the privacy of email is what length of time in a warrant goes beyond legal policework and stretches into the timeframe of a violation of one's rights under the courts' United States v. Jones ruling and the Mosaic Theory of Intelligence (Schlabach, 2015).

In the United States v. Jones, five of the Supreme Court justices wrote citizens should be able to reasonably expect privacy even in the collection of data, accepting limited surveillance may be reasonable if one is suspected of a crime, but not acceptable if the surveillance goes on for an extended period (United States v. Jones, 2012). United States v. Jones involved the usage of a GPS tracker on a suspect's car for over a one month period (United States v. Jones, 2012). Justice Sotomayor specifically noted the extended collection of GPS data was the equivalent of the Mosaic Theory, since by looking at the collected data one could discern personal and private information (United States v. Jones, 2012). She noted certain trips such as those to a psychiatrist, an abortion clinic, an AIDS treatment center, a strip club, a by-the-hour motel, a gay bar, etc. can all be used to identify personal aspects of the individual which are protected under an individual's right to privacy (United States v. Jones, 2012).

Aaron Hernandez

Aaron Hernandez was a former tight end for the New England Patriots (SI Wire, 2017). While playing at the University of Florida Aaron and his team won the National Championship

and as such he was named an All-American. Mr. Hernandez was drafted in 2010 by the New England Patriots in the 2010 NFL draft (SI Wire, 2017). Mr. Hernandez was released from the New England Patriots in June of 2013 due to his involvement in the murder of Odin Lloyd, who was dating the sister of his fiancé (SI Wire, 2017). On April 15, 2015, Mr. Hernandez was found guilty of first degree murder (SI Wire, 2017).

Unfortunately, Mr. Hernandez's legal troubles were just starting as he was subsequently charged in the murder of two Daniel de Abreu and Safiro Furtado. On April 14, 2017, Mr. Hernandez was acquitted of these charges (SI Wire, 2017). Unfortunately, Mr. Hernandez subsequently took his life while in prison for the murder of Odin Lloyd on April 19, 2017 (SI Wire, 2017). Mr. Hernandez was chosen due to his high profile within the professional football community and the subsequent coverage of his murder trials. The target of this study should be an individual who has a basic public profile which the participants can associate to, as well as be recent enough to have data which was mineable by the tools and tool sets we were working to identify.

The Mosaic Theory of Intelligence and Gaps in Literature

Jaffer defined the Mosaic Theory of Intelligence in his article, The Mosaic Theory, as the justification for the government to withhold information from the public, the reason for silencing its citizens, or for the government's right to collect information which should otherwise be kept confidential (Jaffer, 2010). Whereas David Pozen defined the Mosaic Theory of Intelligence as a basic precept of intelligence gathering, collecting disparate pieces of data which have little value individually, but when joined together provide a more significant picture of a topic which is greater than the sum of its parts (Pozen, 2005). However, Pozen continued his definition to

include the resulting mosaic, when undertaken by one's adversary, can prove just as dangerous in identifying one's weaknesses and vulnerabilities (Pozen, 2005).

The Mosaic Theory has been around for over seven decades but has primarily been used only by the government and the courts (Halkin v. Helms, 1978). The government would make use of the Mosaic Theory to build their intelligence collection and analysis infrastructure on, then use the Mosaic Theory to then block access to government information from the public due to the fact one could possibly do the same type of work against the government (Graziano, 2016). The courts on the other hand have until the past decade only accepted the usage of the Mosaic Theory for purposes of blocking access to information, but this has changed (Schlabach, 2015).

There is a study by Jeffery W. Hayman, "Case Study: Suggested Best Practices for Redacting U.S. Army Aviation Accident Reports to Reduce Opportunities for Doxing of Re-identified U.S. Army Aircrew", which closely matched this study and served as a key component during the research and development phase of this study (Hayman, 2015). One such issue of the mosaic theory being addressed currently in the Supreme Court is the case of Carpenter v. United States (Carpenter v. United States, 2017). Nathan Wessler, a lawyer for the American Civil Liberties Union who is representing Timothy Carpenter, stated:

The court could not have imagined the technological landscape today, highly sensitive digital records like search queries entered into Google, a person's complete Web browsing history showing everything we read online, medical information or fertility tracking data from a smartphone . . . would be vulnerable. (Carpenter v. United States, 2017)

Yet these items are vulnerable and are being used by everyone including the government and law enforcement.

Whether the person carrying out the activity is using Google to search across numerous public databases the search engine has indexed or another tool such as Intelius, a website that provides public data on people and their connections to other individuals, the end goal remains the same, to gather tiny bits and pieces of information on an individual. The motives behind this are bound to vary with each individual, but the goal at the end remains the same to find out information on an individual that if asked for of that individual they could refuse. So, whether it is a stalker wanting to get detailed information on their victim or a prospective employer researching a potential employee's past for any potential problems, the Mosaic Theory of Intelligence is being used every day (Pozen, 2005).

Now, we see the slow movement of the courts toward the possibility the Mosaic Theory can be used for something other than blocking access to government information, it is now usable by the government intelligence agencies and law enforcement to invade an individual's constitutional rights and their expectation of privacy (Schlabach, 2015). It is unfortunately for this very reason, courts are just now seeing the effects this theory can have on an individual's rights, and there exists a literature gap when it comes to the Mosaic Theory of Intelligence (Schlabach, 2015).

The goal of this literature review was to account for the other forms of technology, policing, business intelligence, big data, and more which all make an unconscious use of the Mosaic Theory. The belief was these other topics allow us to explore a topic, the Mosaic Theory of Intelligence, which has not had much research done on it outside of the legal community as far as can be identified in academic searches.

Chapter Summary

Chapter 2 has underscored the viability of tools and data being available for individuals of all characteristics to be able to carry out the exercise laid forth in this paper. It has showcased as well these activities are taking place every single day on every single person. The literature review also discussed gaps in the literature with regards to the topic of the Mosaic Theory of Intelligence.

The purpose of this qualitative study was to observe what tools or tool sets were successful in building the most detailed mosaic on a given topic or target. Chapter 2 supported the reasoning for this study to be carried out by showing the Mosaic Theory and the tools to carry out activities associated with the Mosaic Theory have shifted away from governments and corporations and are now in the hand of a motivated individual. In addition, it showed that while a gap exists in the literature, a review of such topics as big data, social networks, and law enforcement and surveillance can provide a base of knowledge upon which the study can stand.

The next chapter will cover the research methodology of this study. Chapter 3 will cover research methodology, appropriateness of the method, the population to be studied, and the reliability and validity of the study. Further chapters will cover the findings of this study and identify and discuss any possible recommendations for further research.

CHAPTER 3: RESEARCH METHODS

The purpose of this qualitative study was to observe what tools or tool sets were successful in building the most detailed mosaic on a given topic or target. This chapter reviewed the chosen research method and design as well as the validity of those choices. In addition, the chapter reviewed the study's instrumentation, validity, reliability, population, sampling, and location parameters of the study. Finally, the processes of both the data collection and data analysis procedures was discussed.

Research Method

The research design for the study was a qualitative analysis making use of a grounded theory mosaic profile to identify the tools or tool sets which were the most successful in building the most detailed mosaic on a given topic or target. The first step in developing the research method for this study was identifying which analysis method would be the most appropriate (Corbin & Strauss, 2015). Looking at the topic of research, the identification of the tools which were most successful in building the most detailed mosaic on a given topic or target, one can conclude the variables in this study, the tools, was unknown and must be identified (Corbin & Strauss, 2015). As Corbin and Strauss identified in their book, Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, two of the most common reasons given for the use of qualitative research over quantitative research is the basis one is trying to research in areas not yet fully explored or they are trying to identify key variables which can be tested later via quantitative research (Corbin & Strauss, 2015). Thusly, since the variables of this study are unknown and must be identified, both of these reasons are applicable to this research topic and thusly guided us to a qualitative study.

Once the qualitative research analysis method was decided upon an extensive review was undertaken to ensure the study will make use of the appropriate research design. Five separate research designs were considered during the design phase of the study, biographical studies, phenomenological studies, ethnography studies, case studies, and grounded theory studies. The research design chosen for this study was the grounded theory. The grounded theory study method is built around the idea that the research study's theory is developed from and grounded in the data which has been collected and analyzed (Nieswiadomy, 2008). As Nieswiadomy notes, grounded theory uses both approaches to theory development, inductive and deductive (Nieswiadomy, 2008).

This approach of allowing the data to drive the theory was essential as the researcher was proceeding this study with a minimal set of expectations on what tools are going to be used by the study participants. The collection of the data took place online over a period of a few weeks, this allowed for the researcher to implement the process called constant comparison. Constant comparison is a process in which newly collected data is constantly compared to data which has already been gathered previously (Nieswiadomy, 2008). This allowed for the identification of pertinent concepts and possible assignment of data coding. And while the researcher must maintain an open mind one can make use of intuitive processes in interpreting the data coming in to identify specific trends and key data points (Nieswiadomy, 2008).

When identifying and testing of hypotheses, the grounded theory flips the standard approaches of having a hypothesis and then running a study to prove or disprove it on its head. Rather, the grounded theory focuses on going into the study with no hypotheses identified ahead of time, but instead to allow the hypothesis to be generated from the data coming in (Nieswiadomy, 2008). The grounded theory derived hypothesis therefore is self-correcting,

meaning as the study data is collected, adjustments are made to the theory to allow for newly obtained data to be interpreted (Nieswiadomy, 2008). Figure 1 depicts the interrelationship between data collection and analysis in grounded theory.

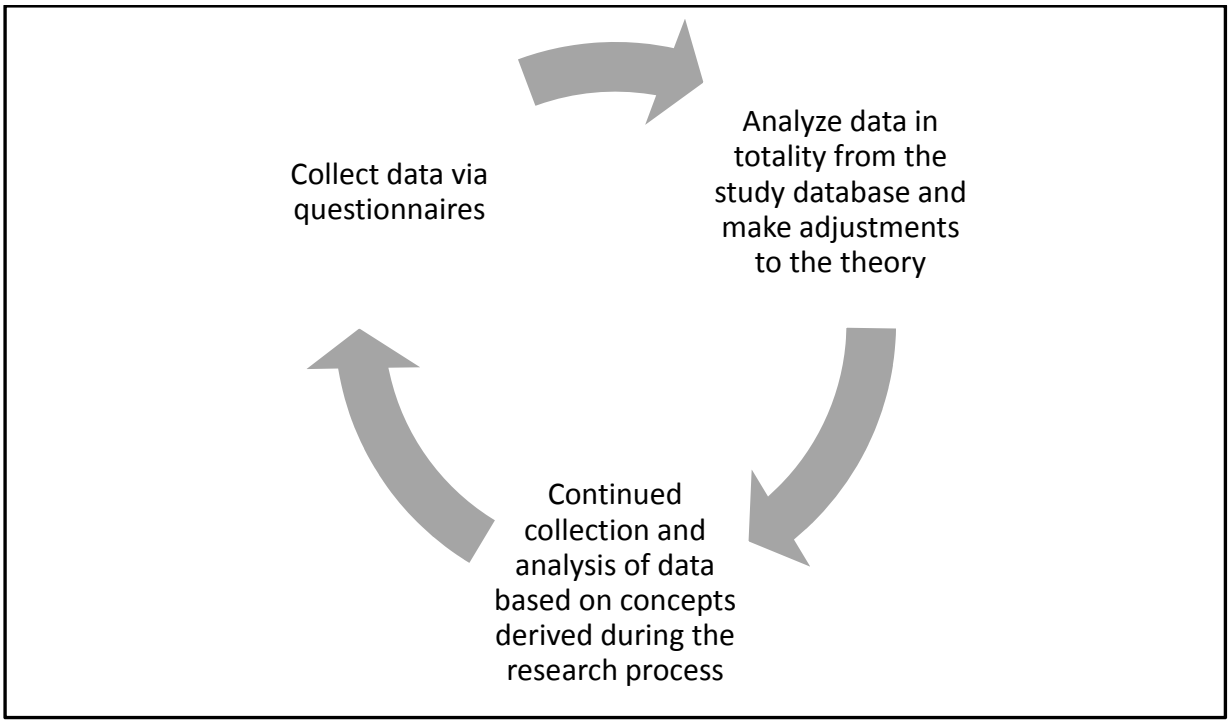


Figure 1. Interrelationship Between Data Collection and Analysis

The types of data collected by a grounded theory study are numerous. The most common types of data for a grounded theory study are interviews and observations (Corbin & Strauss, 2015). However, virtually any source of data, written, observed, or recorded can be used to collect the data from the study's subjects. This research study made use of questionnaires, or profile forms, and contained not only data on the target generated by the study participants, but also contained demographic data on the study participant and the post study questionnaire data. Figure 2 below identifies the types of data collected and the contribution expected from each particular type of data.

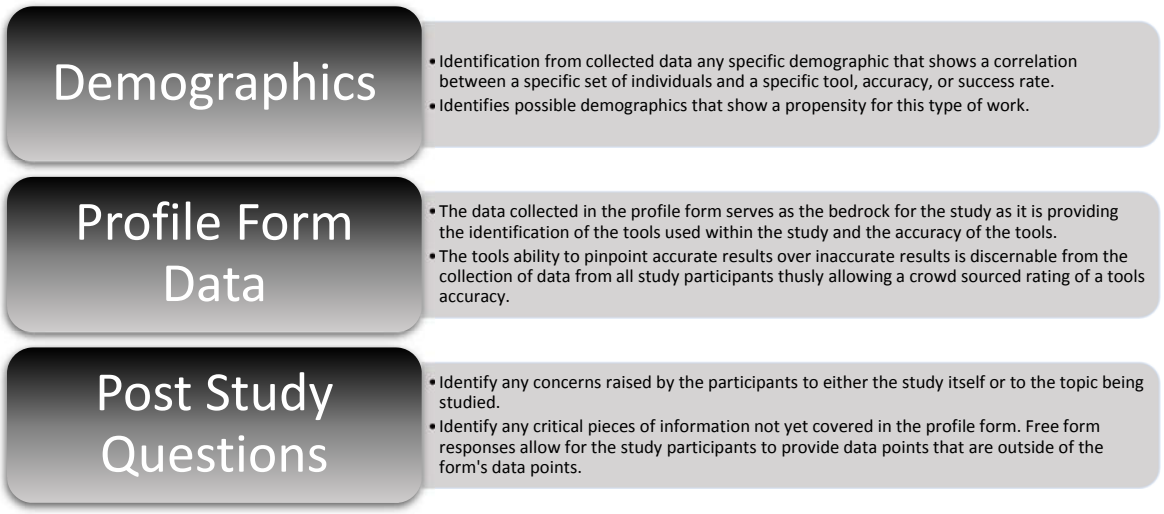


Figure 2. Data Collection Types and Contribution for This Study

The progression of research followed for this study are a literature review, collection and coding of data, review of collected data, and finally observations. The collection and coding of data as well as the review of collected data was ran three separate times as this research study was making use of three separate groups of study participants. This repetitive collection and review was also one of the reasons behind the choice of the grounded theory research design choice (Corbin & Strauss, 2015). The use of this research design allowed for any problems in the study to be identified in between each iteration and allow for correction of said problems (Corbin & Strauss, 2015). The data collected from each participant was the demographics form, the profile form, and finally the post-study questionnaire form.

Once the data was collected from this study it was analyzed by the researcher. The data which can vary widely based on the success level of each individual was validated for accuracy during the coding phase to allow for the researcher to draw accurate conclusions at the end of the study. This validation was carried out to ensure not only are the tools identified which provide data, but validation of said data can be used to judge the susceptibility of the tool itself to be

deceived (Corbin & Strauss, 2015). This study was guided by the research question, which tool sets demonstrated the most detail in completing a successful mosaic profile.

Design Appropriateness

This qualitative study utilized the grounded theory method and design. The grounded theory design is the most appropriate form of research when the research study's theory is developed from and grounded in the data which has been collected and analyzed (Nieswiadomy, 2008). This grounded theory design collected data from the participants and then studied said data to identify the most successful tool sets which demonstrated the most detail in completing a successful mosaic profile (Corbin & Strauss, 2015). The study made use of various data gathering techniques such as the collection of participant demographics, collection of mosaic profiles, and the post-study questionnaire from each participant. The tools and tool sets the participants made use of for this exercise were not restricted but the topic/person they produced the mosaic profile on was restricted to a single individual.

The individual who was the target of the mosaic profile was chosen based on his celebrity status and a history of previous attempts at digging up information on the target. This study presented the participants with a brief five-minute introduction video to the target of the mosaic profile exercise as well as a brief overview of what a tool and tool set are with respect to this exercise. The data collected wasa compared against the known correct answers, identified and validated ahead of time by the researcher, and each use of a tool was collected and identified as to whether it provided real or fake information (Corbin & Strauss, 2015). Figure 3 below shows a breakdown of a profile form's review and coding methodology.

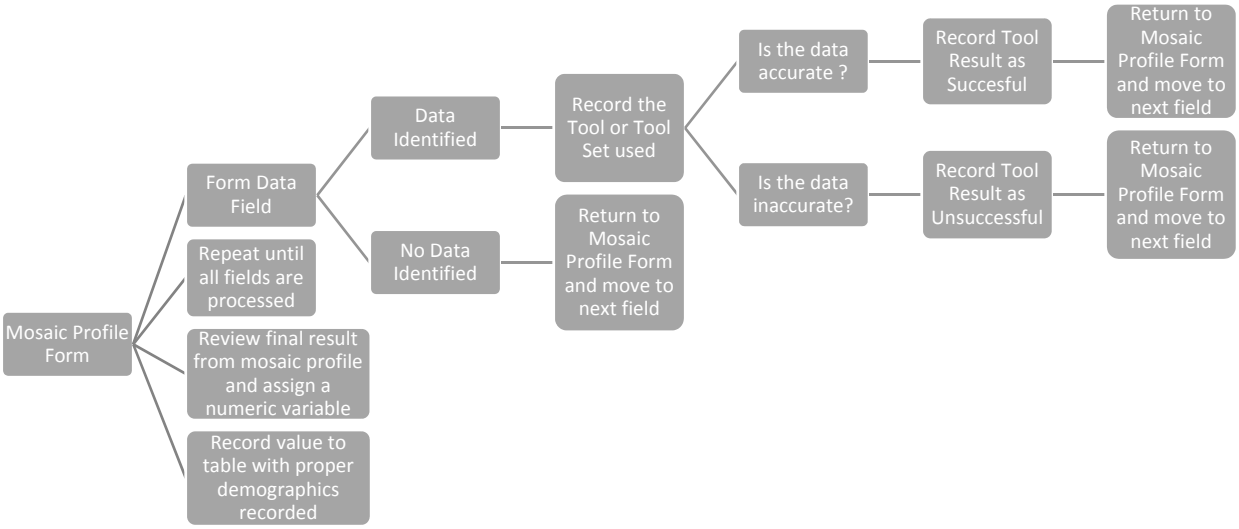


Figure 3. Mosaic Profile Review and Coding Methodology

As the process shown above in Figure 3 shows, the data collected by the participants was tracked based on a multitude of factors from tools used, accuracy of data at the field level, final overall mosaic completeness score, and the demographics associated with said participant. The primary focus of the results was to provide a reliable list of tools and tool sets which produced the most accurate mosaic profiles in completeness, accuracy, and usage level (Corbin & Strauss, 2015). After these tools were identified, the roles these tools play in our society can be reviewed more in depth and considered for possibly blocking the usage of said tools on certain pools or sources of information.

The grounded theory method was appropriate for this study as the data which was coming from each of the study participants was feeding back into the study to identify the tools which are most successful, rather than going into the study with a preconceived set of tools which would either restrict the participants or not identify unknown or emergent tools (Corbin & Strauss, 2015). Corbin and Strauss note a researcher will have to live with some ambiguity about the meaning of the data and be willing to follow the leads in the data (Corbin & Strauss, 2015). The focus of the research question was *what* tools or tool sets demonstrated the most *success* in

completing a successful mosaic profile, as opposed to which of the following tools do participants use. This shows the tools or tool sets are unknowns which are emerging from the data rather than being a pre-defined list of tools chosen by the researcher, therefore the grounded theory research method and design was the only logical choice (Corbin & Strauss, 2015).

The literature review performed as part of this dissertation has provided some information relative to this study. The literature review showed the Mosaic Theory of Intelligence is not as widely known or covered in detail by academics, reporters or even technology experts. Thusly, the researcher had to make use of additional related fields of study such as business analytics, doxing, big data, social networking, surveillance, and various court systems. The literature review when expanded by these related fields of study assisted in allowing for the more precise formulation of the research question as opposed to a hypothesis which would be required for a quantitative study (Corbin & Strauss, 2015).

A pilot study was conducted to determine the viability of the research question and its applicability to the overall research topic (Creswell, 2014). The pilot study involved the testing of the survey tools and the ability for a pre-selected individual to produce a valid mosaic profile of the target Aaron Hernandez. This pre-selected individual, an industry individual finishing their bachelor's degree in cybersecurity, was aware of the Mosaic Theory and has performed exercises close to this exercise in the past. The goal was to discern what level of information participants had gathered on the individual and to ascertain the validity of the information with the target (Creswell, 2014). Once a baseline of information was established as to what the researcher had found in the past, the researcher could better prepare a mosaic profile form which would ask questions which were known to be available (Creswell, 2014).

Qualitative research is focused on identifying a set of theories and testing those theories through careful study. And the grounded theory research design is designed to take that one step farther, by allowing the study itself to test the theory and modify it through the collection of data. Corbin and Strauss emphasize a researcher should not begin their research with a pre-identified list of concepts (Corbin & Strauss, 2015). Thusly, this research and its associated research study needed the maximum amount of flexibility to allow the participants to identify the tools and tool sets rather than allowing the researcher to impugn the study with preconceived notions.

Research Question

The research question, which tool sets demonstrated the most detail in completing a successful mosaic profile, served as the core aspect of this study. An online mosaic profile website was created to allow for the study participants to partake in the mosaic profile experiment and provide via their activities the data the study needed to answer the research question. Once, the data had been collected from each of these submissions, the researcher collated the experiment data and updated the study as needed to ensure the most accurate results (Corbin & Strauss, 2015). In addition, a specific post-study set of questions was added to the study to allow for corrections to the experiment and its environment to ensure the strongest set of results (Benn et al., 2015). The end result hopefully provided the researcher and the public with insight into what is considered the best tools for implementing the Mosaic Theory of Intelligence on any given topic (Benn et al., 2015).

This qualitative grounded theory study made use of an open-ended questionnaire. Open-ended questionnaires are appropriate for this study as the end result of the form was to identify the tools or tool sets used by each individual to identify the most widely used and the most widely effective tools in creating a mosaic profile (Züll, 2016). Creswell specifically notes

closed ended questions are primarily used when one is attempting to support a theory or concept which has already been stated, whereas open-ended questions are aimed at exploring the responses given by the study participants (Creswell, 2012).

The study made use of the open-ended questionnaire to extract from each participant the tool or tool set they used for each piece of collected data on the target subject. One drawback noted by Corbin and Strauss with this type of study instrument is the abundance of data which will be given by the study participants which can vary from short to long strings of data which must be analyzed one by one by the researcher (Corbin & Strauss, 2015). This concern was addressed to ensure short answers and minimal interpretation by the researcher. The usage of a short example list of tools was provided, brevity in response was stressed in the introduction to the exercise and a physically smaller box on the questionnaire form were all implemented to reduce the possibility of unnecessarily long responses.

Creswell recommends these types of activities be used to ensure the focus of the results is aimed at answering the research question (Creswell, 2012). Meanwhile Corbin and Strauss also note, one cannot let the richness and detail of the data gathered to be lost (Corbin & Strauss, 2015). The research questionnaire thusly was modified to account for this possible loss with a set of post-study questions asked of each participant upon completion. The belief was these questions would allow the study participants to provide a context for their work on the open-ended questionnaire. By providing context to their answers, the researcher can ensure the richness of their responses are retained.

The research question developed for this study followed the form of research questions typically associated with qualitative research, in there is a central question and associated sub-questions (Creswell, 2014). The central question as previously mentioned was, which tool sets

demonstrate the most detail in completing a successful mosaic profile? But this was followed by a series of sub-questions which allowed for the central question to be explored in more depth. Which demographics of individuals demonstrated the highest rate of success in completing a successful mosaic profile? Which demographics of individuals made use of the particular tool sets?

Finally, a secondary set of additional questions were also proposed to see if they could possibly provide additional insight to the final results of the study. Was there a significant correlation of one age group to Google? Or was Google age agnostic? What percentage of data found was incorrect, false flag? Did any subjects make use of a paid service? If so, did it increase accuracy or completeness? Did the location of the experiment play a role in the project? Did the combination of age, gender, and location of subjects pose a risk of cross correlation of data to expose subjects?

Study Participant Open-ended Questionnaire and Post-study Questionnaire

This grounded theory study was designed to enable participants in the study to guide the study and its conclusions by focusing on the data collected from the participant's responses (Corbin & Strauss, 2015). This study had two documents which made up the collection of this data. The first document which was provided was an open-ended questionnaire, referred to as an intelligence profile in the documents for easier understanding by the participants (Züll, 2016). The second document which was provided was a post-study questionnaire (Benn et al., 2015). The open-ended questionnaire allowed the participants to identify as many pieces of information as possible on the research study target while at the same time noting the tools they were making use of when identifying said data (Benn et al., 2015). This open-ended document is illustrated in Figure 4.

Mosaic Profile

Target: Aaron Josef Hernandez

Aaron Hernandez, a former Tight End for the New England Patriots, has recently passed away in Shirley, MA. Outside the basics of his being born in 1989 in Connecticut, and having an interesting college, pro, and post-pro football career, little is being shared. You are an intelligence analyst hired to create an intelligence profile of the target. Your tasking is to craft an intelligence brief which provides clear, chronological, credible and detailed information on the target.

A structured form has been provided on the next few pages for the most commonly gathered information. Anything not fitting into one of the provided categories, should be entered in the section titled, Other Data. Finally, at the end of the profile, you will find a small section asking for you to answer a few basic demographical questions, please make sure you complete this section before submitting the document.

Tool Examples:

Google Search Engine, Google Scholar, Google Maps, Facebook, Facebook Graph, Classmates.com, Ancestry.com, Realtor.com, City/County/State governmental databases, Spokeo.com, Intelius, Twitter, Instagram, IMDB, PeopleFinder, USA People Search, InstantCheckmate.com, and many other websites and tools.

Just make sure you do not get stuck in a never ending chain.

Age:		Date of Birth:	
Tool		*Tool*	
Gender:		Marital Status:	
Tool		*Tool*	
Political Affiliation:		Religion:	
Tool		*Tool*	

Spouse(s):	
Tool	
Mother:	
Tool	

Father:	
Tool	
Sibling(s):	
Tool	
Children:	
Tool	
Other Relatives:	
Tool	
Current Address(s):	
Tool	
Previous Addresses:	
Tool	
Employment History: (include military/government work)	
Tool	
Language(s):	
Tool	
Education (All Levels):	
Tool	
Criminal/Legal History:	
Tool	
Photos:	
Tool	
Favorite(s): (color, movie, etc)	
Tool	
Medical History:	
Tool	
Financials:	
Tool	
Email Address(s):	
Tool	
Phone Number(s):	
Tool	
Conferences, Symposia, or other public speaking events:	
Tool	
Certifications:	
Tool	

Published Works:	
Tool	
Social Media Profiles:	
Tool	
Any news references:	
Tool	
Any scandals:	
Tool	
Member of Organization(s):	
Tool	
Security Clearances:	
Tool	
Other 1:	
Tool	
Other 2:	
Tool	
Other 3:	
Tool	
Other 4:	
Tool	
Other 5:	
Tool	

Demographical Information of Study Participant

Gender:	
Age:	
Current Education Level:	
Current GPA:	
Self Rated Tech Skills: *Scale of 1 to 10* *1 being no skills* *10 being In-Depth Skills*	

Figure 4. Open-ended Questionnaire

The open-ended questionnaire seen in Figure 4 gave a wide array of fields for a study participant to try and discover details on the research study target. Each field on the questionnaire has been validated by the researcher to be a piece of information which has been previously found by the researcher performing this exercise in the past. The study participants

were asked to record not only the information they discovered but the tool or tool set they used to collect said data. Study participants were informed should they fail to identify any information or a minimal amount of information their results would still be of interest to the study and its conclusions. A short explanation of the key pieces of data are described below.

1. Introduction. This portion of the form was designed to familiarize the study participants with the basic information of the study target, Aaron Hernandez. A few high-level pieces of biographical data, which are publicly known, were given to the participants to begin the study. The preselected target was chosen as he has lived recently enough to have a wide exposure online of his life and was a former public figure.
2. Exercise description. This portion was laying out for the study participant the activity they were being requested to complete.
3. Tool examples. In order to ensure easier processing of the data entered into the form by the participants, a list of some well-known tools were given so the participant could see how the tools should be identified.
4. Questionnaire fields. This portion of the questionnaire was made up of fields which had previously been identified as available on this individual via the use of various tools and tool sets. Each field requesting a piece of information on the target had an associated field located immediately below it which requested the participant to identify the tool they used to garner said piece of information.
5. Gender and age. This portion of the form was focused on collecting key demographical pieces of information on the study participants. These first two fields

were collected to determine whether one's age or gender played any role in the success of an individual in this study.

6. Current educational level and current GPA. This portion of the form was focused on answering one of the key sub-questions of the study, as to whether education played a role in the success of an individual in this study.
7. Self rated tech skills. This portion of the form was a subjective question posed to the study participant to get an understanding of the individual's confidence in their tech skills. This would also hopefully help in identifying whether a participant's success was dependent on their perceived level of technical skills.

This grounded theory study was attempting to identify which tool sets demonstrate the most detail in completing a successful mosaic profile. The mosaic profile form presented in the open-ended questionnaire form in Figure 4 above was used to not identify any key piece or pieces of information on the study target, rather it was focused on seeing which tool or tool set were able to garner the information being requested (Züll, 2016). The study participants were given a target to give them a focus on an activity which would have them think critically, identify a tool or tool set, test their identification of a tool, and if successful present the result along with the tool they identified (Züll, 2016).

The results from the participant's activity were used by the researcher and the study to illuminate or identify what tools or tool sets were the most successful in identifying these critical pieces of information. It was crucial before identifying any tool as successful to remember both true and false data have been seeded over the years by individuals (Creswell, 2014). A verification of validity of the data collected must not be accepted as a successful identification of data, if said data is incorrect, so the data was validated before identifying a tool as successful.

Once these tools and tool sets were identified the research progressed into the next phase (Creswell, 2014). In the next phase the focus was on identifying the answers to the sub-questions of did a demographic of individuals demonstrate higher rates of success as well as did certain demographics of individuals make wider use of specific tools or tool sets.

The second document to be used by the researcher in this study was the post-study questionnaire. This questionnaire was filled out immediately following the completion of the study exercise. The questions were a mixture of closed-ended and open-ended questions, a few closed-ended questions were asked to allow for metrics to be drawn from the data, while the open-ended questions were in place to allow for the participant to provide context for their work. As Creswell notes, one cannot let the richness and detail of the data gathered to be lost (Creswell, 2012). The belief was these open-ended questions would allow the study participants to provide the requested context for their work. By providing context to their answers, the researcher could ensure the richness of their responses is retained.

Post-study Questionnaire

1. Have you ever performed this type of activity in the past? This question will be asked to discern whether an individual may have prior experience with this activity and whether that experience may affect the results collected. None of the study participants are expected to have done a mosaic profile before due to its currently obscure nature.
2. If you answered yes, please explain: This question is to better define the perceived knowledge an individual may believe they have with regards to this activity type. This data will provide insight into what correlation of activities individuals believe mimic a mosaic profile.

3. On a scale of 1 to 10, with 1 being very difficult and 10 being very easy, where would you classify this activity? This question will be used to identify the perceived difficulty of the exercise by the study participants. This particular metric is expected to be useful when looking into the demographical aspects of the participant pool and the success of the participants. Question three will hopefully also present a piece of interesting data in the correlation of a study participant's self-identified technical skills level with the perceived difficulty of the exercise.
4. Which tool or tool set did you find to be the most useful? Please explain: This question will be used to develop an understanding as to whether a tool is perceived correctly or incorrectly to be of the most use. Does a tool that is easier to use but garners less information get identified as most useful or is the tool that provided the greatest amount of information identified as the most useful.
5. Did you feel that the time given to do this exercise was enough? This question is of the biggest concern to the researcher. Due to time constraints imposed by class schedules, the maximum amount of time that could be used for the profile portion of this study was 25 minutes. Prior experience in this type of exercise by the researcher has led the researcher to know the longer spent on this exercise the more detail that can be found. But the belief is the 25 minute window should be enough to identify the best and most widely used tools or tool sets.
6. If given additional time what percentage more data do you think you could discover? This question is asked of the participant to see if the exercise is perceived to have gotten easier or harder by the end of the exercise. This perception could then be used

to possibly extrapolate additional metrics comparing time spent to the completeness of the profile.

7. What information was easiest to find? This question will hopefully make the participant reflect on the exercise and their successes. In addition, this could make the participant question the amount of data that is out on the internet and publicly available.
8. How would you rate the validity of your results? This question encourages the student to reflect on the study and identify the validity of their work. This question presents a unique opportunity to compare the perceived validity of one's work with the actual validity of their work and whether any factors or demographics play a role in this difference of perception.
9. Does this study make you reconsider how much information may be available on you on the internet? In preparation for the final question, the researcher wanted to ensure the participant would internalize the exercise that the participant just completed. The researcher hopes this question will encourage an internal reflection of the topic by the participant.
10. Is there anything else you would like to share such as feelings or concerns? This final question is where all the other questions have been leading. This is the question that will allow the participant to inject the context of their observations, feelings on the exercise, and concerns regarding the topic. By presenting this open-ended question to the participant, the goal is to allow them to inject into the study the richness of data that is otherwise lost among hard figures.

In this final stage of the study, the post-study questionnaire, the participants were allowed to look back over their open-ended questionnaire to allow for an accurate reflection to be drawn as to the work they completed. Upon the completion of the thirty-minute window, the website collected all documents previously presented to the participants. Participants were notified they should not mention this exercise to any of their fellow colleagues nor share any of the information collected during this study with any others. Upon completion of each group of individuals, the data collected was coded and reviewed and any mention of the subject's name was redacted. Upon completion of the study, all documents were scanned to a pdf file, stored on a USB drive, and placed into a lockbox at the researcher's office. Documents will be destroyed after three years. The open-ended questionnaire as well as the post-study questionnaire are located in Appendix A of this dissertation.

Instrumentation

This grounded theory study made use of multiple artifacts which make up one of the necessary aspects of instrumentation in any type of study. Instrumentation can come in many forms, such as interviews, observations, videos, documents, drawings, diaries, group meetings, memoirs, newspapers, historical documents, biographies, etc. (Corbin & Strauss, 2015). For this study, the study participants were given two documents which covered both the study exercise as well as a post-study questionnaire.

The research question, which tool sets demonstrated the most detail in completing a successful mosaic profile, guided this grounded theory study. The first artifact to collect was the mosaic profile, also referenced as the open-ended questionnaire (Corbin & Strauss, 2015). This document was given to study participants at the beginning of the exercise and served as the primary data collection mechanism for the overall study. The participants also made use of

electronic records in the form of websites, databases, software tools, and other methods of data collection, but they were only asked to notate on the mosaic profile those tools which produced pertinent data (Corbin & Strauss, 2015).

Upon completion of the mosaic profile, the study participants were presented with a post-study questionnaire which was designed to clarify the work the study participants performed during the exercise. The data to be collected on this post-study questionnaire was being collected to not only clarify their work but to also to ensure, as Creswell notes, and not let the richness and detail of the data gathered to be lost (Creswell, 2012). All data collection was being performed with the express purpose of allowing the researcher to identify the tools or tool sets which were best at producing successful mosaic profiles. As Corbin and Strauss emphasize there is no research without data (Corbin & Strauss, 2015). To do research one must collect, code and analyze said data as part of their overall research.

The data collected from this study exercise as well as other associated data used to verify the validity of the study participant's results were stored into a central study database (Creswell, 2014). The data once collected and coded into the database allowed the researcher to begin the analysis of the results and work toward the identification of the most successful tools or tool sets in regard to mosaic profiles (Creswell, 2014). The goal of identifying these tools or tool sets was just the beginning of the analysis process, as the database was then be used to answer the additional questions of identifying demographic advantages, demographics of tool usage, validity of data collected, etc. (Creswell, 2014). Figure 5 below addresses the study database and what information, or artifacts made up the data stored and analyzed in the database.

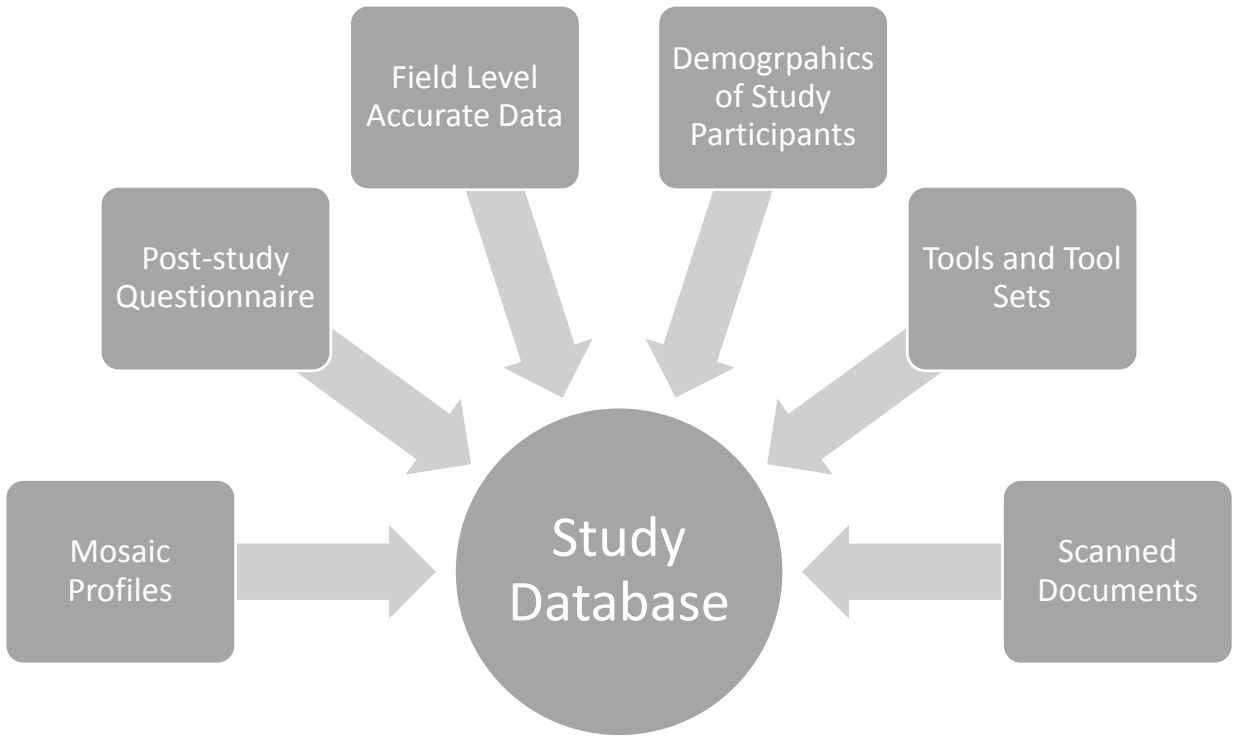


Figure 5. Study Database

The study database diagram above lays out the specific categories of artifacts which were collected during the research study. Per Salkind, the data collection process has four distinct steps which must be followed (Salkind, 2016). First, the researcher must construct a data collection form which can be used to organize the data you collect. Second, a coding strategy must be identified and implemented which can be used to ensure accurate translation of the data into the database. Third, the researcher must carry out the data collection portion of the exercise. And finally, the collected data from the exercise must be entered in to the database using the identified collection form and proper coding methodology (Salkind, 2016).

A study protocol was developed and followed during this grounded theory study to ensure a consistent and reliable set of rules and procedures were followed. The study protocol which was developed is shown in Appendix D (Hayman, 2015). Creswell stated an essential

process in qualitative research data collection is designing a protocol which guides the recording of the data which is collected (Creswell, 2012).

The data stored into the study database was coded per the coding strategy identified by the researcher. Flick noted the key rule to follow in data coding is to always record study data in coding which is both explicit and as discrete as possible (Flick, 2014). The overall goal of this coding was to reduce the clutter which may be present in the data collected from the participant, without losing the true meaning of the data (Flick, 2014). The coding process used for the data collected in this study involved the use of two different coding methods.

The first form of coding method used in the study was in vivo coding. In vivo coding was used in the initial data entry process of copying the data from the collected documents produced by the study participants, in particular the open-ended questionnaire as well as the post-study questionnaire. Saldana recommends in vivo be used as the first coding method in grounded theory studies (Saldana, 2015). Saldana identified a common reason for the usage of this coding method was to prioritize and honor the study participant's voice in recording the data (Saldana, 2015). This coding method was best utilized in the post-study questionnaire.

The second form of coding method used in the study was pattern coding. According to Miles, Huberman, and Saldana, pattern codes are precise codes which can identify an emergent theme or explanation (Miles, Huberman, & Saldana, 2013). These codes combine the whole of the material into more precise units of analysis, a sort of meta-code (Miles, Huberman, & Saldana, 2013). This second form of coding was used to allow for the researcher to take advantage of not only emergent themes in the data but also in the overall research which made use of the grounded theory study method. The combination of these two coding methods allowed

the researcher to retain the participant's voice, but at the same time, narrow down the data provided by the study participant into a more clearly defined set of categories.

The study protocol depicted in Appendix D was followed at all times during data collection and during both of the data coding cycles, in vivo coding and pattern coding. All data collected were entered into the study database first utilizing the in vivo coding. Subsequently the researcher reviewed the data collected and recorded additional details of the data utilizing the pattern coding method. At the completion of this study, a set of qualitative results were observed, which detailed the analysis of the data to be collected. The instruments used for this grounded theory study were a grounded theory study protocol, a coding strategy, and the overall study database.

Validity and Reliability

This study employed a qualitative grounded theory research study design and method. When it comes to qualitative research validity and reliability are debated as to their usefulness. Corbin and Strauss stress qualitative research has both scientific as well as creative or artistic components, and thusly the quality of the final product should reflect both aspects (Corbin & Strauss, 2015). While Whittemore, Chase, and Mandle argued innovative thinking can mesh with reasonable claims, evidence presented, and applicable methods (Corbin & Strauss, 2015). This disagreement within the research community though does not mean validity and reliability should be ignored, rather the opposite it should be embraced and adapted to work with qualitative research.

Validity has numerous definitions throughout the scientific community and while a whole paper can and has probably been written on the topic, a definition must be identified for use in this study. Two different definitions were reviewed in depth to discern which definition would be

used in defining validity for this study. Creswell argues validity in qualitative research is more of a checking for accuracy of the findings from the standpoint of the researcher, the participant, or the readers of an account (Creswell, 2014). Martyn Hammersley, of Open University in London, stated a research account can be considered valid if "it represents accurately those features of the phenomena that it is intended to describe, explain, or theorize" (Guest & Namey, 2015). For the purpose of this study, the Hammersley definition seemed to be the most appropriate definition of validity and was used going forward.

Reliability was a bit easier to classify in terms of research work. Reliability in its most generic form means the results from an instrument are stable and consistent (Creswell, 2012). Creswell describes reliability as results which should be nearly identical when researchers administer an instrument multiple times at different times (Creswell, 2012). In short, Creswell is stating good research will have measures or observations which are reliable and repeatable (Creswell, 2012). The structure to be used by the researcher in this grounded theory study for validity and reliability was defined by three tests, internal validity, external validity, and reliability.

Validity

This dissertation relied on the collection and analysis of data to ensure the validity of the grounded theory study. The first test, internal validity in qualitative research, focuses on the basic concept of how congruent are the findings with reality (Shenton, 2004). One of the methods which can be used to ensure internal validity is to employ tactics to help ensure honesty in informants (Shenton, 2004). The tactic employed by the researcher to ensure honesty in the participants was to allow them to choose to be excluded from the study, as well as, encouraging frank responses with no notice or feedback as to the correct or incorrect answer on the mosaic

profile (Shenton, 2004). The second method employed by the researcher to ensure internal validity was to reach out to a third party and provide the third party with this study and its raw data for peer review and scrutiny to ensure further validity (Shenton, 2004).

The second test, external validity in qualitative research, is defined by Merriam and Tisdell as "concerned with the extent to which the findings of one study can be applied to other situations". To meet this definition, external validity was controlled by providing a specific subset of information to the readers of the study. The subset of information which was presented at the outset was the number of organizations taking part in the study and where they are based, any restrictions in the type of people who contributed data, and the number of participants involved in the fieldwork, as well as the data collection methods which were employed, the number and length of the data collection sessions, and the time period over which the data was collected (Shenton, 2004). Finally, this study does not generalize as most qualitative studies do not generalize, but the contents of the study may be transferable to provide explanations for comparable situations and studies (Yin, 2016).

Reliability

The third test, reliability, addresses the issue by employing techniques to show, when work is repeated, in the same context, and makes use of the same methods and the same participants, similar results should be produced. This study addressed the issue of reliability by ensuring as much detail as possible is given to other researchers on the processes used within the study (Shenton, 2004). The three core pieces of information Shenton recommends be provided to allow for a deep understanding of the study are, the research design and its implementation, the operational detail of data gathering, and reflective appraisal of the project (Shenton, 2004).

Population and Sampling

The general population for this grounded theory study consisted of members of various industry, business, academic, and governmental cybersecurity communities and organizations. The population was included to ensure individuals at all levels can impart their knowledge on the topic. The population was selected due to the research problem being focused on the production of mosaic profiles. As the topic of mosaic profiles and cyber intelligence are unique skills, the researcher believed the best results could be gained from individuals who are part of the industry groups, CSFI and HTCC, and are aware of these tools and skills.

In the case of qualitative studies, Creswell states because researchers are trying to understand a central phenomenon, the random sampling a quantitative study might use to generalize against a population is not necessary (Creswell, 2012). Instead qualitative study researchers can proceed directly to people and places which can best help in understanding the central phenomenon via purposeful sampling (Creswell, 2012). Purposeful sampling is aimed toward providing the strongest set of data by focusing the selection of a study's population on four key aspects, the setting (where research will take place), the actors (who will participate), the events (the study exercise), and the process (the activities carried out by the participants in regard to the exercise) (Creswell, 2014). The belief was the participants in this study would be able to collect enough information via their usage of various tools and tool sets to provide a rich data set to be analyzed.

Creswell states purposeful sampling is the optimal choice as it presents the researcher with a population which is information rich in the phenomenon being studied (Creswell, 2012). The study's targeted population was members of various industry communities. The population was selected due to the research problem being focused on the production of mosaic profiles. As

the topic of mosaic profiles and cyber intelligence are unique skills, the researcher believed the best results could be gained from individuals who makeup these communities, CSFI and HTCC, and are aware of these tools and skills. For the study to proceed a sampling method was chosen. Based on the predefined characteristics of the overall population, the sampling method which best fits the study is homogenous sampling. Homogenous sampling allowed for the selection of participants based on a similar trait or characteristic, which in the case of this study was a deep interest in cybersecurity (Creswell, 2012).

Sampling sizes vary across the various research designs as well as the phenomenon being studied. Creswell though does specify in his book, Research Design: Qualitative, Quantitative and Mixed Methods Approaches, studies which use the grounded theory research design should have a minimum of twenty to thirty participants (Creswell, 2014). The sampling size though can be larger if there is belief additional data may be garnered from a larger group. Creswell offers for these types of grounded theory studies that a different sampling size can be used which is called saturation (Creswell, 2014). Saturation sampling size states you stop collecting data when the categories or themes being studied are saturated, when the data collected is no longer identifying new data or results (Creswell, 2014).

The population size available to the researcher at the various study locations was larger than needed under the guidelines laid out by Creswell, so the population was narrowed down to a smaller group. The problem presented to the researcher was identifying which characteristics were the most crucial in the population and narrowing the population accordingly. One demographic of study participants, educational level, is believed by the researcher to be the most defining demographic in succeeding in the study exercise. The educational level of the

participants was reviewed against the sampling guidelines laid out by Creswell to determine the most accurate size of the demographic specific groups (Creswell, 2014).

Ultimately, the overall size of the entire population necessary to produce the best results was identified as being approximately forty total study participants. The population was further reduced based off the defined demographic of educational level. As Creswell mentioned, grounded theory studies tend to have sample sizes of approximately twenty to thirty, so this was used as the starting number for each educational level (Creswell, 2014). The idea of saturation posed by Creswell was then used to discern what variance in the population would be acceptable to produce the most accurate data (Creswell, 2014).

The final number of participants for the bachelor educational level was set at approximately fifteen participants based on the premise individuals will skew toward younger individuals who have significant online experience. The final number of participants for the master educational level was set at approximately fifteen participants based on the premise this group may contain a wider age range of individuals who may have less online experience. Finally, the final number of participants for the doctoral educational level was set at approximately fifteen participants based on the premise this group will contain participants with a more defined set of skills due to their advanced degree.

This grounded theory study was conducted online. Additional demographical information was gathered by the researcher during the study, but none have been identified to be a critical demographic which could affect the success of an individual in producing a mosaic profile (Axinn, Link, & Groves, 2009). The tools and tool sets which were identified by the study participants can be accessed online via any computer and as such all participants will be required to have a computer of their own which they will use to carry out the study (Axinn, Link, &

Groves, 2009). No other factors affected the overall research question. The research question was, which tool sets demonstrate the most detail in completing a successful mosaic profile?

Confidentiality

Qualitative research at its core focuses on creating an in-depth exploration of a central phenomenon, and the most generally accepted way of exploring the phenomenon is to involve the population affected by the phenomenon (Creswell, 2012). The involvement of participants presented the researcher with the responsibility of protecting their identity. This grounded theory study did not gather any personally identifiable information, commonly referred to as PII, as defined by the National Institute of Standards and Technology (NIST) (McCallister, Grance, & Scarfone, 2010). The PII defined by NIST is presented in Figure 6. The researcher did not make use of any interviews or other interactions with the participants, so the only data used to track and identify the documents provided to the participants and given back to the participants was a unique number identifier for each participant packet.

Name	• full name, maiden name, mother's maiden name, or alias
Personal Identification Number	• social security number (SSN), passport number, driver's license number, taxpayer identification number, patient identification number, and financial account • or credit card number
Address Information	• street address or email address
Asset Information	• Internet Protocol (IP) or Media Access Control (MAC) address or other host-specific persistent static identifier that consistently links to a particular person or small, well defined group of people
Telephone Numbers	• mobile, business, and personal numbers
Personal Characteristics	• photographic image (especially of face or other distinguishing characteristic), x-rays, fingerprints, or other biometric image or template data (e.g., retina scan, voice signature, facial geometry)
Information Identifying Personally Owned Property	• vehicle registration number or title number and related information
Information Linkable to an Above Item	• date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information

Figure 6. PII as defined by NIST (McCallister, Grance, & Scarfone, 2010)

The researcher further made use of an informed consent notification at the beginning of the participant packet (Flick, 2014). No formal letters of acceptance to participate were issued so collection and security of those letters was not necessary. In addition, since no PII was ever requested by the researcher, no PII was stored in the study database. First and last contact with the participants took place via an online website (Flick, 2014). Acceptance of participation in the study by participants was gathered by the student continuing to take part in the exercise after reading the informed consent notification (Flick, 2014).

The population this study drew its participants from was industry communities. The sample chosen via homogenous sampling was a subset of the population which had a deep interest in cybersecurity (Flick, 2014). This sample was further broken down into three separate groups based on their educational level. Punch in his book, Introduction to Social Research: Quantitative and Qualitative Approaches, noted a key strength of the qualitative research method is the study is carried out in a naturalistic setting for the participants (Punch, 2014), so the setting of the participant's home was deemed the most appropriate setting for the study to take place.

Coding of the participant responses was done to facilitate the confidentiality of the study participants even though no PII was being collected on the subjects. The method used was a simple combination of two pieces of data, the educational level of the participant and a randomly assigned participant id number. The code was expected to follow the structure seen in these examples, B-2, M-46, D-5. The abbreviations are identified as B for Bachelor, M for Master, and D for Doctorate. This method thusly allowed for a quick identification of results as well as a count of participants involved in the study at each educational level.

Participants were asked to devote 25 minutes, controlled by the JotForm software being used, to the core aspect of the study, the identification of data on a mosaic profile. The data compiled by the participants on the subject of the mosaic profile, Aaron Hernandez, were collected without any confirmation being given to the participants as to the accuracy of their results. In addition, the master document containing all the accurate results for each field in the mosaic profile was only seen by the researcher and was never provided to the participants. These steps were taken to not only ensure the confidentiality of the participants but also the confidentiality of the subject, Aaron Hernandez.

Procedures for Data Collection

The beginnings of any quality qualitative study's data collection lie in following a set of steps set forth by Creswell in his book Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. These steps include setting the boundaries of the study, collecting information through observations, interviews, documents or visual materials, and establishing the protocol for recording the information gathered (Creswell, 2014). The first step of setting out the boundaries of the study has been previously discussed under the section describing the population. The focus here was on the collection of the information produced and the protocol for recording the information produced. The collection of information in this study included the mosaic profiles which were produced by the participants using the open-ended questionnaire, the post-study questionnaire, the coded responses of the study participants, and finally the analysis which was completed at the end of the data collection phase.

Upon completion of the study, all documents were scanned to a pdf file, stored on a USB drive, and placed into a lockbox at the researcher's office. The lockbox is accessible only by the researcher. Physical documents collected during the study will be kept for approximately three

months to ensure there were no coding or scanning issues during the data recording steps. At the end of these three months, all physical documents will be destroyed by the researcher. Finally, all data on the USB drive as well as the USB drive itself will be destroyed three years after completion of this study.

This grounded theory study made use of a two-cycle coding method for data recording. A thorough review of coding methods as laid out by Saldana in his book, The Coding Manual for Qualitative Research, identified the two-cycle coding method was the appropriate choice for this study (Saldana, 2015). The first coding cycle involved the usage of the in vivo coding method. This method was chosen as the most appropriate method which matches with the underlying research design, the grounded theory design.

Grounded theory is built around the idea the research study's theory is developed from and grounded in the data which has been collected and analyzed (Nieswiadomy, 2008). As such, in vivo coding retains and prioritizes the participant's voice by recording their exact words. By recording their exact words, the study retains the rich context which exists in the participant's words, including key phrases or words which can provide additional context to the study outside of just looking at the words. The documents which were collected during the exercise, the mosaic profile and the post-study questionnaire were thusly recorded verbatim into the study database.

The second coding cycle involves the usage of the pattern coding method as the researcher wanted to be able to narrow down distinct themes within the data. This method was also chosen as the most appropriate method which correlates with the grounded theory design. Pattern coding allowed for codes to be created which identify emergent themes in the data which tracks with the research design nicely. The second coding cycle allowed for the identification of

common themes identified by participants in their responses as well as identifying the level of accuracy participants were able to attain in their research. The expectation was the two coding cycles would allow for not only the identification of which tool sets demonstrated the most detail in completing a successful mosaic profile, but also identify any emerging trends from the data along the lines of demographics, success rates, etc.

Procedures for Data Analysis

This grounded theory study followed the process developed by Creswell for data analysis in qualitative design as found in his book, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (Creswell, 2014). The process and its seven steps are identified in Figure 7 below. The first step was to gather together all the raw data produced by the study, which consists of the mosaic profiles, demographics, and post-study questionnaires (Creswell, 2014). The second step was to organize and prepare all data for analysis which consisted of keying the raw data into the study database using the in vivo coding method, making a scan of the documents, and then validating the data entry for accuracy (Creswell, 2014).

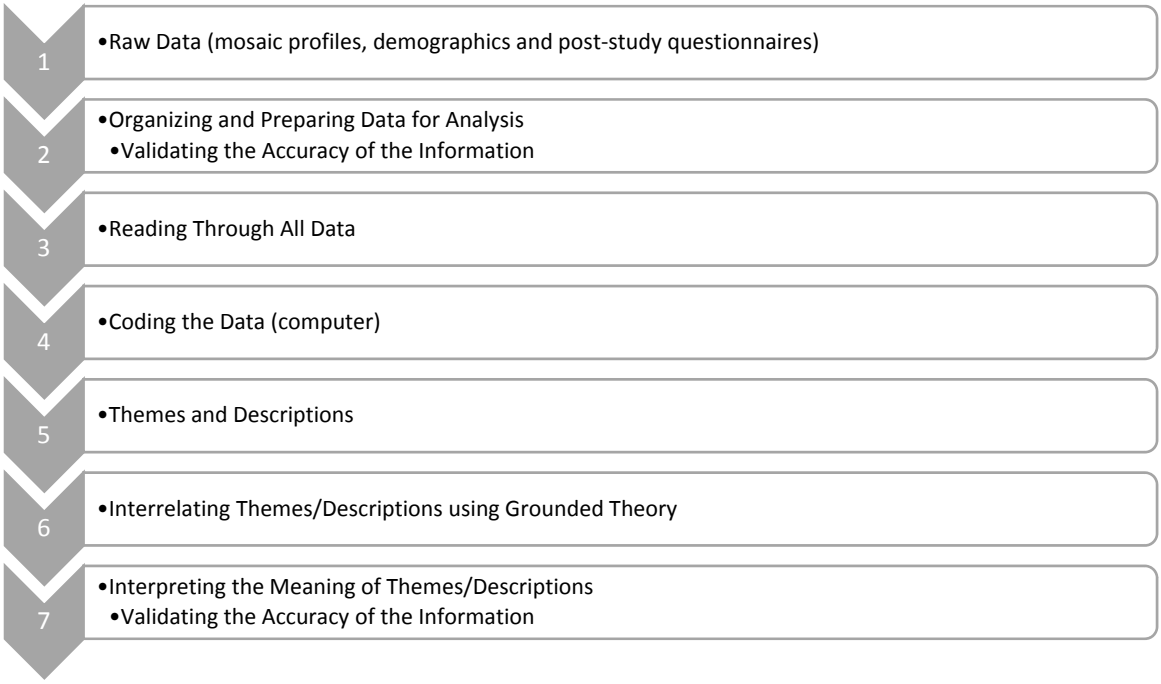


Figure 7. Data Analysis in Qualitative Research (Creswell, 2014)

The third step was to read through all data collected (Creswell, 2014). The goal of this step was to generate a general sense of information and what it may mean. The fourth step was to code all of the data (Creswell, 2014). This step involved the second cycle of coding, pattern coding, to identify emergent themes in the data. The fifth step was to make use of the themes identified by the pattern coding to generate more detailed descriptions of the themes by developing them into a description which could be recognized as a major finding in the study (Creswell, 2014).

The sixth step was to interrelate the themes and descriptions identified in step five by advancing how the themes and descriptions were represented in the qualitative narrative (Creswell, 2014). This step was done by creating a narrative passage to convey the findings of the analysis as well as a process diagram to show the process used to arrive at the findings of the analysis (Creswell, 2014). The seventh and final step was to interpret the meanings of the themes and descriptions developed in step five. This step was where the researcher gets to answer one of

the most common questions in research, what were the lessons learned (Creswell, 2014)? The answer to this question is the interpretation of the data, what did the researcher conclude as the final answer.

This grounded theory study implemented these seven steps described by Creswell as it progressed. Data was analyzed by comparing data garnered by the participants on their mosaic file against the master profile to validate the accuracy of the collected data (Creswell, 2014). The minimal amount of information was provided to the participants to do a basic identification of the target which served as the beginning point of their research using the tools or tool sets of their choice (Creswell, 2014). The participants then produced a mosaic profile of their own to gather additional data outside of the basics provided.

The researcher validated the accuracy of the data collected by the individuals during the second step of the process, organizing and preparing data for analysis. The study database was configured so during data entry the value inputted would be directly compared to the master accurate value stored in the database. This validation method was chosen to decrease the number of times the data would be processed, as well as lessening the number of tools, such as comparison tables, etc. which would be needed (Flick, 2014). It should be noted this method also allowed for the accurate data to be stored in one spot only, thusly reducing the risk of exposure. This combined with the coding used for participants as mentioned earlier ensured both the participants and the target of the study are both protected from possible data exposures (Flick, 2014).

The data was next reviewed by the researcher in step three, and then the researcher coded the data the second time using the pattern coding method during step four to identify emergent themes (Creswell, 2014). During step four, the researcher also entered coding which identified

the validity of the data collected, the completeness of the data collected, and accuracy of the data collected. The coding identifying validity, completeness, and accuracy then allowed the themes to emerge during step five of the data analysis process (Creswell, 2014).

Step six of the data analysis allowed the researcher to start to discover the emerging themes in the data and how those can develop into the overall findings of the study (Creswell, 2014). During this step, the researcher anticipated the primary research question would be answered, which tools or tool sets demonstrated the most detail in completing a successful mosaic profile? By correlating the accuracy and the completeness of the data to the tools used by the participants, the tool which produced the most pieces of accurate data across all of the mosaic profiles should emerge (Creswell, 2014). Answers for the other research questions and secondary questions were expected to also emerge from the data as well, but those answers were most accurately answered during step seven of the data analysis. In step seven of the data analysis, the data was be reviewed in depth to see what conclusions could be drawn from the resulting data (Creswell, 2014). It was expected during this step the researcher would progress the other primary questions and secondary questions beyond just basic answers, but rather provide a rich context to the answers to these questions.

Chapter Summary

The purpose of this qualitative grounded theory study was to identify which tool sets demonstrated the most detail in completing a successful mosaic profile. This chapter discussed the chosen research method and the appropriateness of method and the chosen design. By exploring these matters in depth, the researcher could identify the grounded theory research was the best choice due to its ability to be altered as the data is collected ensuring that the theory is grounded in the data (Nieswiadomy, 2008). This chapter also identified the use of an open-ended

questionnaire was the best tool to use for the study exercise as it allowed the richness and detail gathered to come through in the data (Creswell, 2012).

This chapter also discussed the coding and raw data collection which would take place by the researcher. In vivo and pattern coding were selected as the coding methods used during the data processing based on Saldana's explanation of the available methods (Saldana, 2015). The design and structure of the study database were developed and presented in this chapter as well. The open-ended questionnaire as well as the post-study questionnaire were both presented, and the sampling of the population was finalized. This chapter concluded with a discussion of Creswell's seven step process of data analysis and a discussion of how the analysis was utilized in this study to produce the results which were being sought (Creswell, 2014).

CHAPTER 4: RESULTS

This grounded theory study of the most successful tool set or sets which are the most successful in building a detailed mosaic profile was conducted with 43 individuals who are in the cybersecurity field. The research compiled by the researcher in Chapters 1 and 2 showed the explosion of data being collected and collated online is exposing more and more individuals to a potential exposure of their personal data online. In addition to the exposure of personal data online, Chapters 1 and 2 also explored the extent of which companies have gone in collecting data on individuals in the expectation of higher spending, more focused advertising, and many other reasons.

The reasoning behind why individuals may want to gather a mosaic profile of an individual were discussed in Chapters 1 and 2, with the Target pregnancy fiasco being a prime example of the power of a mosaic profile. Chapter 3 laid out the research method and design of the study. This grounded theory study contained a literature review to examine the existing data and knowledge which exists to validate the overall study's premise. In addition to this literature review, the study made use of an online mosaic profile questionnaire form, demographics, and a post study questionnaire of individuals performing a 25-minute research activity on the internet. The data which was collected from the study was reviewed and coded according to the previously specified two-cycle coding process, in vivo and pattern coding as identified by Saldana (2015).

Pilot Study

A pilot study was carried out to ensure the viability of the research question and its applicability to the overall research topic (Creswell, 2014). The pilot study was also carried out to ensure the readability of the consent form and the overall instructions, the usability of the

mosaic profile form, and the effectiveness of the post-study questionnaire. The pilot study also established a rough baseline of how much data an individual could collect inside of the aforementioned 25-minute time frame. The results of this pilot study not only provided a rough baseline of information but assured the researcher the underlying focus of performing a mosaic profile was a task the targeted group would be capable of.

The pre-selected individual for the pilot study is an industry individual finishing their bachelor's degree in cybersecurity and is aware of the Mosaic Theory and has performed exercises close to this exercise in the past. The individual completed the task within the allotted 25-minute time frame and provided some basic feedback in his post-study questionnaire. There was only one adjustment made to the study based on the feedback from the pilot study participant. The wording of the introduction and instruction paragraph at the top of the form was altered to add the following information, "but if the data comes from a site you clicked to from Google such as PeopleFinder note it as PeopleFinder".

Findings

The results of this grounded theory study have provided insight into the tool set or sets which are the most successful in building a detailed mosaic profile. The findings of this study were extracted from the results data mathematically as well as through the two coding methods, in vivo and pattern coding. The first coding cycle, in vivo coding, was used by the researcher during the data input and organization stage. The data inserted into the study database was specifically left in the original words provided by the individuals who participated. Saldana recommends in vivo be used as the first coding method in grounded theory studies (Saldana, 2015). Saldana also identified a common reason for the usage of this coding method was to prioritize and honor the study participant's voice in recording the data (Saldana, 2015).

The researcher upon completion of entering all data into the study database moved onto the second coding method, pattern coding. This second form of coding was used to allow for the researcher to take advantage of emergent themes in the data. These codes combine the whole of the material into more precise units of analysis, a sort of meta-code (Miles, Huberman, & Saldana, 2013). The data which emerged from this second coding combined with the results to the underlying research questions allowed the researcher to develop these emergent themes. It is a combination of these themes and the answers to the research questions which allow for the researcher to determine the most successful tools, as well as possible areas of concern.

Participant Observations

A total of 43 individuals completed the mosaic profile search and submitted the forms for review. All but two of the participants, submitted a mosaic profile form which contained at least one piece of data. The data submitted by the participants was entered into the study database and then processed using mathematical analysis to derive the answers to the research questions.

The majority of participants were successful in producing a mosaic profile on the study subject, Aaron Hernandez. The form contained a possible 31 pieces of unique data to be collected on Mr. Hernandez and the study participants identified an average of 13.2 fields of information during the study. Virtually every participant made use of the search engine, Google, with 27 out of the 43 participants getting more than 50% of their results from Google. The majority of the data collected was found to be accurate, with only 8 pieces of incorrect data present in a total of 568 pieces of data collected, an error rate of 1.4%. This data and the other data collected during the study served to develop not only a tool set or tool sets of concern, but to also develop a set of conclusions to the overall exercise of mosaic profiles.

Method of Selection

The goal of this grounded theory study was to get as many cybersecurity individuals as possible to participate in an exercise focused on mosaic profiles and the tool sets which are of concern. The approach to this goal made use of purposeful sampling and in particular homogenous sampling. Homogenous sampling allows for the selection of participants based on a similar trait or characteristic, which in the case of this study was a deep interest in cybersecurity (Creswell, 2012). While the overall homogenous sampling method was effective, the researcher did note a lower than expected turn out of participants at the Doctoral level of educational experience.

General Participant Information

Table 1 shows the demographical information gathered from each participant in the study. All forty-three participants in the study met the requirement of working on or holding a college degree and having a deep interest in cybersecurity. A ratio of roughly 2 males for every one female participant emerged from the population of participants. The average age of the participants who took part in the study was approximately 41 years old, with the youngest participant being 25 years old and the oldest participant being 65 years old. No personally identifiable information such as name or email address were collected so the participant paper id is the principal identifier.

Table 1

Demographics of Participants

Paper ID #	Age	Gender	Education	GPA	Tech Skills
1	42	Female	Bachelor	4	8
2	30	Male	Bachelor	3	10
3	52	Male	Bachelor	3.3	7
4	34	Male	Bachelor	3.2	9
5	25	Male	Bachelor	2.7	4
6	56	Female	Bachelor	2.8	7
7	28	Male	Bachelor	2.9	7
8	27	Male	Bachelor	3.2	9
9	38	Female	Bachelor	3.6	6
10	35	Female	Bachelor	3.5	6
11	28	Female	Bachelor	3	8
12	41	Male	Bachelor	3.7	10
13	54	Male	Bachelor	3.8	7
14	32	Male	Bachelor	2.7	10
15	35	Male	Bachelor	3.5	8
16	28	Female	Bachelor	3.25	8
17	37	Female	Bachelor	3	5
18	34	Female	Bachelor	3.7	8
19	34	Male	Bachelor	3.7	10
20	37	Male	Bachelor	2.9	5
21	49	Male	Bachelor	4	8
22	46	Male	Bachelor	6	7
23	55	Female	Bachelor	3.5	1
24	46	Male	Bachelor	3.89	5
25	29	Male	Masters	3.98	9
26	39	Female	Masters	3.8	7
27	28	Male	Masters	5	7
28	27	Male	Bachelor	3.5	1
29	45	Male	Masters	4	9
30	37	Male	Masters	3.6	7
31	35	Male	Masters	4	8
32	40	Male	Masters	3.93	7
33	63	Male	Masters	3.93	9
34	51	Male	Masters	3.8	8
35	45	Male	Masters	3.95	8
36	65	Female	Masters	3.5	7
37	40	Female	Masters	3.83	2
38	59	Female	Masters	3.9	5
39	28	Female	Doctoral	3.4	8
40	28	Female	Doctoral	3.4	8
41	54	Male	Doctoral	4	8
42	64	Male	Doctoral	3.9	9
43	49	Male	Doctoral	4	10

Direct Observations

This study was carried out online by participants not under direct observation by the researcher. So direct observations which can be identified are based off data collected which the participant did not have any control over. The following are just a few brief summaries drawn about the participants from the data.

None of the participants made use of any paid service or services to collect data on the target of the mosaic profile. Most participants honored the 25-minute time frame they were given, with a few minor exceptions. A few participants went over the allotted time and a few participants ended their study a few minutes early.

Table 2 shows the results of the participants work in relation to the amount of information they were able to identify. Each participant is represented in this table and the basic collection information is included. The fields attempted column represents any field in the mosaic field where an individual provided a piece of data. Extraneous data such as none, n/a, and other non-responsive answers were not counted toward the fields attempted value. The validity of the results does include one minor adjustment made by the researcher, if a participant correctly identified the date of birth but accidentally miscalculated the age by a year, the age field was counted as correct.

Table 2 was presented in this style by the researcher to ensure no information on the study's target, Aaron Hernandez was exposed via this study. The columns presented were chosen to present the data identified by the participants and to quantify the validity of the data they collected. The researcher thusly extrapolated from the provided results the total number of fields attempted by each participant and the validity of the results of each field of data. The total number of available fields of 31 is included to represent the number the participant was attempting to reach during the study. Finally, the results of Table 2 data show only 8 out of the 568 pieces of data collected were found to be incorrect for an error rate of 1.4%.

Table 2

Participant's work and validity

Paper ID #	Total Fields	Fields Attempted	Fields Correct	Fields Correct %	Fields Incorrect	Fields Incorrect %
1	31	14	14	100.00%	0	0.00%
2	31	16	16	100.00%	0	0.00%
3	31	17	17	100.00%	0	0.00%
4	31	13	13	100.00%	0	0.00%
5	31	16	16	100.00%	0	0.00%
6	31	13	13	100.00%	0	0.00%
7	31	13	13	100.00%	0	0.00%
8	31	16	15	93.75%	1	6.25%
9	31	17	17	100.00%	0	0.00%
10	31	17	17	100.00%	0	0.00%
11	31	14	14	100.00%	0	0.00%
12	31	10	10	100.00%	0	0.00%
13	31	9	9	100.00%	0	0.00%
14	31	15	15	100.00%	0	0.00%
15	31	15	15	100.00%	0	0.00%
16	31	11	11	100.00%	0	0.00%
17	31	16	15	93.75%	1	6.25%
18	31	20	20	100.00%	0	0.00%
19	31	16	15	93.75%	1	6.25%
20	31	12	11	91.67%	1	8.33%
21	31	17	17	100.00%	0	0.00%
22	31	12	12	100.00%	0	0.00%
23	31	4	4	100.00%	0	0.00%
24	31	8	7	87.50%	1	12.50%
25	31	11	11	100.00%	0	0.00%
26	31	14	14	100.00%	0	0.00%
27	31	1	0	0.00%	0	0.00%
28	31	1	0	0.00%	0	0.00%
29	31	15	15	100.00%	0	0.00%
30	31	21	20	95.24%	1	4.76%
31	31	21	20	95.24%	1	4.76%
32	31	9	9	100.00%	0	0.00%
33	31	10	10	100.00%	0	0.00%
34	31	13	13	100.00%	0	0.00%
35	31	9	9	100.00%	0	0.00%
36	31	11	10	90.91%	1	9.09%
37	31	24	24	100.00%	0	0.00%
38	31	11	11	100.00%	0	0.00%
39	31	13	13	100.00%	0	0.00%
40	31	13	13	100.00%	0	0.00%
41	31	11	11	100.00%	0	0.00%
42	31	13	13	100.00%	0	0.00%
43	31	16	16	100.00%	0	0.00%

Table 3 provide the data to the primary research question of the study; the most successful tool set or sets which are the most successful in building a detailed mosaic profile. Every tool identified by the participants for each piece of data collected was counted and entered into the study database. There is a small deviation of approximately 13 uses between the fields collected and the tools used to collect the field where some participants forgot to specify the tool used to collect the data. The researcher did not associate these 13 uses to any tool set or tool sets

as this could skew the results. Finally, Table 3 displays the tool sets and shows no paid services were used by any of the participants.

Table 3

Most Successful Tools Used to Build a Mosaic Profile

Tool Set	Total Uses of Tool (555 Total)	Percent of overall usage
Google	371	66.85%
Wikipedia	111	20%
Facebook	16	2.88%
Twitter	2	0.36%
Biography.com	14	2.52%
CNN	8	1.44%
USA Today	3	0.54%
Zillow	2	0.36%
Spokeo	4	0.72%
Government Databases	3	0.54%
Others (2 uses or less)	21	3.78%

Table 4 provides data for the research question as to whether the usage of Google was correlated to one age group or was Google age agnostic. The study data shows while virtually all the participants used Google to some extent, those who made use of it for more than 50% of their overall results were identified as being a heavy user of Google. These 27 individuals listed in Table 4 who used Google heavily allows for the researcher to identify any possible extensive use by a particular age group.

Table 4

Heavy Google Use with Age

Paper ID #	Age	Total # of Uses	Google	Google %
1	42	14	12	85.71%
2	30	16	16	100.00%
4	34	13	12	92.31%
8	27	16	16	100.00%
9	38	17	9	52.94%
10	35	17	9	52.94%
11	28	14	13	92.86%
15	35	15	8	53.33%
16	28	11	9	81.82%
19	34	16	9	56.25%
20	37	12	11	91.67%
21	49	17	14	82.35%
22	46	12	12	100.00%
23	55	4	4	100.00%
24	46	8	8	100.00%
25	29	11	11	100.00%
29	45	15	12	80.00%
30	37	21	21	100.00%
31	35	21	18	85.71%
32	40	9	9	100.00%
33	63	10	10	100.00%
34	51	13	11	84.62%
36	65	11	11	100.00%
37	40	24	15	62.50%
38	59	11	11	100.00%
39	28	13	11	84.62%
40	28	13	11	84.62%
43	49	16	16	100.00%

Table 5 shows all the demographics of the participants in the study along with the success rate of their producing the most complete and accurate mosaic profile. This data will allow the researcher to answer the question as to whether there exists a key demographic of participants who display a higher rate of success versus the rest of the participants. The success rate which is assigned to each participant is calculated as follows:

(((Fields Attempted/Total Fields)+((Fields Attempted-Average Fields)/100))-Fields Incorrect %)

The Fields Attempted value divided by Total Fields value provides a baseline score percentage for each participant's level of completeness of the mosaic profile. This value is then increased by adding a success bonus to any participant who exceeded the average number of fields identified by all participants which is calculated at 13.2 fields. The success bonus is calculated as (Fields Attempted -Average Fields)/100. The final factor in the equation is the

removal of percentage points for any participant who encountered false or incorrect data, which is calculated by subtracting from the success rate the Fields Incorrect Percentage.

Table 5

Success Rate of Participants with Demographics

Paper ID #	Submission Rating	Your Age	Your Gender	Education	Recent GPA
1	45.96%	42	Female	Bachelor	4
2	54.41%	30	Male	Bachelor	3
3	58.64%	52	Male	Bachelor	3.3
4	41.74%	34	Male	Bachelor	3.2
5	54.41%	25	Male	Bachelor	2.7
6	41.74%	56	Female	Bachelor	2.8
7	41.74%	28	Male	Bachelor	2.9
8	48.16%	27	Male	Bachelor	3.2
9	58.64%	38	Female	Bachelor	3.6
10	58.64%	35	Female	Bachelor	3.5
11	45.96%	28	Female	Bachelor	3
12	29.06%	41	Male	Bachelor	3.7
13	24.83%	54	Male	Bachelor	3.8
14	50.19%	32	Male	Bachelor	2.7
15	50.19%	35	Male	Bachelor	3.5
16	33.28%	28	Female	Bachelor	3.25
17	48.16%	37	Female	Bachelor	3
18	71.32%	34	Female	Bachelor	3.7
19	48.16%	34	Male	Bachelor	3.7
20	29.18%	37	Male	Bachelor	2.9
21	58.64%	49	Male	Bachelor	4
22	37.51%	46	Male	Bachelor	6
23	3.70%	55	Female	Bachelor	3.5
24	8.11%	46	Male	Bachelor	3.89
25	33.28%	29	Male	Masters	3.98
26	45.96%	39	Female	Masters	3.8
27	0.00%	28	Male	Masters	5
28	0.00%	27	Male	Bachelor	3.5
29	50.19%	45	Male	Masters	4
30	70.78%	37	Male	Masters	3.6
31	70.78%	35	Male	Masters	4
32	24.83%	40	Male	Masters	3.93
33	29.06%	63	Male	Masters	3.93
34	41.74%	51	Male	Masters	3.8
35	24.83%	45	Male	Masters	3.95
36	24.19%	65	Female	Masters	3.5
37	88.22%	40	Female	Masters	3.83
38	33.28%	59	Female	Masters	3.9
39	41.74%	28	Female	Doctoral	3.4
40	41.74%	28	Female	Doctoral	3.4
41	33.28%	54	Male	Doctoral	4
42	41.74%	64	Male	Doctoral	3.9
43	54.41%	49	Male	Doctoral	4

The data in Table 5 provides not only the success rate of each participant but their demographic markers as well. The overall average of the success rate across all participants was 41.68%. The overall average of the success rate for each respective educational level was Bachelor at 41.69%, Master's at 41.32%, and Doctoral at 42.58%. The overall average of the

success rate for each respective age level was 25-35 at 43.65%, 36-45 at 46.46%, 46-55 at 35.65%, and 56-99 at 34.00%.

Gender was the final demographic to be correlated to see if there was an increase in the success rate based on the gender of the participant. The data was collected and reviewed via two different comparison methods to ensure accuracy. The overall average of the success rate for the respective genders are females at 45.50% and males at 39.64%. An additional analysis was run against the success rates as a graded level. When all the success rates were given a rating of high, medium, average, or low the correlation showed 27% of the females rated medium or better compared to 36% of males who rated medium or better.

The direct observations of the data provided us with the data needed for the researcher to answer the research questions which were discussed in Chapters 1 and 3, but the observations also led to a collection of additional information or themes which developed out of the data. These themes were discovered in the data both during the in vivo coding process of the post-study questions, as well as during the pattern coding process where distinct patterns were noted and assigned. A thorough review of these themes must be undertaken to better understand the data.

Emerging Themes from Data Collection

The data collected from the mosaic profile forms as well as the post-study questionnaires guided the researcher to identify the tool set or tool sets which provide the most detail in completing a successful mosaic profile as well as a subset of questions such as demographics, bad data, and more. From the analysis carried out during this process a set of nine themes also emerged. The data was run through two separate coding cycles to ensure to ensure these themes

were accurate developed out of the data and not out of any bias. The following are the nine themes which developed from the data.

1. News sites are favored outside of Google and Wikipedia.
2. The results are believed valid by default.
3. Prior experience with these types of activities does not affect success.
4. Time is a factor.
5. Privacy concerns only play a role for some participants.
6. Finding the exercise fun seemed to increase the success rate.
7. Date of birth and basic demographics classified easiest to find.
8. Most participants found the exercise easy.
9. Perceived tech skills have no effect on success rate.

Research Question

The research question was designed to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target. To answer this question and a subset of additional research questions, the researcher enlisted 43 individuals to participate in an online exercise producing a mosaic profile on a famous individual, Aaron Hernandez. Table 6 displays the nine themes which emerged out of the data beyond the research questions. These themes and the research questions will all be addressed in Chapter 5.

Table 6

Theme Matrix

Theme ID	Title	Basis
1	News sites are favored outside of Google and Wikipedia.	Tool Set calculations
2	The results are believed valid by default.	Post Study Question on validity
3	Prior experience with these types of activities does not affect success.	Post Study Questions on whether they had done this activity before
4	Time is a factor.	Post Study Questions on additional time and collection
5	Privacy concerns only play a role for some participants.	Post Study Questions on privacy concerns and feedback
6	Finding the exercise fun seemed to increase the success rate.	Post Study Question on feedback
7	Date of birth and basic demographics classified easiest to find.	Post Study Question on data searches
8	Most participants found the exercise easy.	Post Study Question on rating the difficulty of the activity
9	Perceived tech skills have no effect on success rate.	Derived from data correlation of success rate and tech skills

Theme 1: News sites are favored outside of Google and Wikipedia.

The participants in the study were asked to identify each tool set they used during the process of building the mosaic profile. The data shows 14 out of the 43 participants in the study made use of a news website to gather data on the target. These websites included sites such as CNN, USAToday, and Biography.com.

Theme 2: The results are believed valid by default.

The participants in the study were asked to collect as much data as possible and as accurate as possible on an individual. Upon completion of the mosaic profile activity, the participants were asked to answer a post study questionnaire. One question asked of the participants was, “How would you rate the validity of your results? (use 1-100%)”. Out of the 43 participants 38 participants self-rated their results as at a minimum of 75%, while the remaining

5 participants placed their results between 50% and 74%. The results show no participant placed the validity of their results below 50%.

Theme 3: Prior experience with these types of activities does not affect success.

The post study questionnaire answered by each of the participants asked them many questions to better quantify their results. One set of questions asked the participants if “Have you ever performed this type of activity in the past?” followed up by “If you answered yes, please explain:”. The data showed 4 participants had answered yes to performing this activity in the past with answers ranging from having done research on individuals in the past through hunting criminals. The success rate of the 4 individuals who claimed previous experience were rated as one of medium quality, one of low quality, and two of average quality.

Theme 4: Time is a factor.

The post study questionnaire answered by each of the participants asked them, “Did you feel that the time given to do this exercise was enough?” and “If given additional time what percentage more data do you think you could discover? (use 1-100%)”. The results show 19 out of the 43 participants felt the time they were given was not sufficient enough for the given activity. When the participants were then asked to quantify how much additional data they could discover if additional was given, 21 out of 43 participants felt they could increase their results by at least 50%.

Theme 5: Privacy concerns only play a role for some participants.

The post study questionnaire answered by each of the participants asked them, “Does this study make you reconsider how much information may be available on you on the internet?”. The results show 24 out of 43 participants answered yes to this question and are questioning

what information is out there on themselves. The post study questionnaire also asks student for additional feedback and two participants left comments addressing privacy. One participant, paper id # 2, expressed unease stating, “Some of the information requested is private and I do not feel comfortable trying to find it.”. Whereas another participant, paper id #19, expressed very little concern stating, “I don’t feel very concerned since the target I was given was a high-profile murderer/football player. I’m a nobody compared to him.”.

Theme 6: Finding the exercise fun seemed to increase the success rate.

The post study questionnaire answered by each of the participants asked them, “Is there anything else you would like to share such as feelings or concerns?”. This open-ended question provided additional insight into each participant’s thinking and in at least 6 cases, participants noted some aspect of fun or enjoyment in the study. One participant, paper id # 4, stated simply, “Fun study”. The success rate of each of the participants who found the exercise fun was calculated as being above average.

Theme 7: Date of Birth and basic demographics classified easiest to find.

The post study questionnaire answered by each of the participants asked them, “What information was the easiest to find?”. The results show 29 out of 43 of the participants found the date of birth and demographic fields to be the easiest pieces of data to locate online. The results also show 38 of the 43 participants correctly identified the date of birth of the subject individual.

Theme 8: Most participants found the exercise easy.

The post study questionnaire answered by each of the participants asked them, “On a scale of 1 to 10, with 1 being very difficult and 10 being very easy, where would you classify this activity?”. The results show 29 out of the 43 participants rated the exercise as a 5 or greater.

The results also show 16 out of the 29 participants who rated the exercise as easy gave a rating of 7 or greater.

Theme 9: Perceived tech skills have no effect on success rate.

The participants were asked at the beginning of the study for their demographical information, specifically, age, gender, educational level, GPA, and self-rated technical skills. The results show the highest success rate of all 43 participants, 88.22%, was from a participant who rated their technical skills at a 2. The average technical skills rating for all participants was 7.2. The results also show the average quality and low-quality results come from participants who have an average technical skill of 7.1.

Chapter Summary

Chapter 4 presented the findings from this grounded theory study. The study made use of a wide variety of instruments and tools to achieve the best results. Both direct and informal observations were used to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target. Once data collection was complete all data was processed through two coding cycles to in vivo and pattern coding to allow for identification of additional themes.

These themes developed from the analysis carried out during the coding process were run through two separate coding cycles to ensure to ensure these themes were accurately developed out of the data and not out of any bias. The nine themes which developed from the data provide additional context to data beyond the primary research question or other research questions. Some of the data collected may not have played a role in the development of these themes, but most of the results underpin most of these themes. One example is the individual who collected

the most data and had the highest success rate was the third lowest in technical skills of all 43 participants.

The theoretical and practical implications of these findings are presented in Chapter 5. Associated recommendations and potential areas of further research will also be discussed. The research questions and themes will be addressed individually. Chapter 5 will also outline the conclusions and recommendations based on the findings of this grounded theory study.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

The United States Court of Appeals decision in the case of United States v. Maynard (2010) stated the Mosaic Theory allows for one who has a broad view of a situation to realize the importance of small bits of data which may seem trivial to the uninformed. The Court specifically notes no longer is foreign intelligence a cloak and dagger affair, but it is instead a construction of a mosaic of thousands of bits and pieces of seemingly innocuous information (Halkin v. Helms, 1978). The Mosaic Theory of Intelligence does not discriminate in who can make use of the theory. The theory can be used by anyone from a stalker looking up information about the individual they are stalking to a nation state attempting to piece together information for an upcoming operation against a country. The purpose of this grounded theory study was to observe what tool set or sets are successful in building the most detailed mosaic on a given topic or target.

The significance of this study to the cybersecurity field was to allow interested parties to gain an understanding of the tools and individuals which may prove to be of the most concern to them when it comes to the Mosaic Theory of Intelligence. A general problem exists where the technology to mine data is available to everyone, and mining can have far reaching criminal ramifications by building a mosaic of information on any given topic. The specific problem is what specific tools or tool sets make an experiment testing the mosaic theory of intelligence successful. This qualitative grounded theory study attempted to identify the most successful tools or tool sets to build a complete mosaic profile, as well as whether a subset of individuals has an innate talent for these profiles.

This study identified instances where the government has used the Mosaic Theory of Intelligence to block access to information under the Freedom of Information Act, through to the

tossing out of evidence in cases which made it to the Supreme Court of the United States where the actions of law enforcement were equated to mosaic profiles. The increased focus on producing more and more detailed profiles of individuals by the government, corporations, websites, and many more were discussed in the literature review. The legality of these actions and their usage was discussed as well to provide a context of the current legal environment. Chapter 4 presented the results of this qualitative grounded theory study and demonstrated any individual can do a mosaic profile to some degree with the right combination of tools. Chapter 5 will discuss the results of the study focusing on answering not only the underlying research questions but also addressing the nine additional themes which emerged from the data.

Limitations

The underlying purpose of any research is to continue improving and to add to the body of knowledge in one's field (Creswell, 2012). The research problem studied was to identify the tool sets which demonstrated the most detail in completing a successful mosaic profile. The researcher chose to limit the target of this study of the mosaic theory of intelligence and the success of mosaic profiles to a single individual, Aaron Hernandez, to address two potential pitfalls privacy of the targeted individual and accuracy of the data.

A second limitation faced by the researcher was the ability to access enough individuals to provide a statistically relevant pool of participants across the educational spectrum. As the topic of mosaic profiles and cyber intelligence are unique skills, the researcher believed the best results could be gained from individuals who have a deep interest in cybersecurity. Therefore, the researcher employed the homogenous sampling method which allows for the selection of participants based on a similar trait or characteristic.

Recommendations

The study's findings, combined with the literature review, yielded the answers to the research questions as well as an emerging set of themes which were presented in Chapter 4. The primary research question of which tool set demonstrated the most detail in completing a successful mosaic profile is discussed in detail, including the overwhelming usage of the Google search engine as a primary tool set for information. This section will analyze each of the research questions and themes to identify information which allows for the adding of information to the body of knowledge.

Primary Research Question

The data collected during the study aimed to answer the underlying primary research question, what is the most successful tool set or sets which are the most successful in building a detailed mosaic profile. The data showed overwhelmingly that Google was the tool set used to collect the most information with 66.85% of all information collected by the study participants coming from Google. The data showed the second tool to be used the most was Wikipedia with 20% of the data collected coming from Wikipedia. The study did identify additional tools which were used but no tool reached a level of significance by themselves. These additional tools can be found in Table 3 of Chapter 4. The researcher recommends individuals investigate the right to be forgotten laws in their jurisdiction to see if these laws can provide an individual with a possible resolution.

Secondary Research Question

The secondary research question of the study was the identification of any demographics of individuals who made use of a particular tool set or individuals who demonstrate the highest

rate of success in completing a successful mosaic profile. The overall average of the success rate across all participants was 41.68%.

The overall average of the success rate for each respective educational level was Bachelor at 41.69%, Master's at 41.32%, and Doctoral at 42.58%. The overall average of the success rate for each respective age level was 25-35 at 43.65%, 36-45 at 46.46%, 46-55 at 35.65%, and 56-99 at 34.00%. The overall average of the success rate for the respective genders are females at 45.50% and males at 39.64%. An additional analysis was run against the success rates as a graded level. When all the success rates were given a rating of high, medium, average, or low the correlation showed 27% of the females rated medium or better compared to 36% of males who rated medium or better.

The data showed regardless of gender, age, or educational level there seems to be no single unique demographical attribute which increases the success rate of a participant. Since the data has not produced a demographic attribute which increases the success of an individual the recommendation is a focus be put on other attributes which may have been identified elsewhere in the study for addressing this issue. The data in Table 5 of Chapter 4 provided the information for these observations.

Additional Secondary Subset of Questions

The data collected also answered a subset of secondary questions which were posed at the beginning of the study. The data provided in Table 4 in Chapter 4 shows while virtually all participants used Google to some extent, only those participants who made use of it for over 50% of their results were used to answer the question as to whether the usage of Google was age agnostic. The 27 participants which met this criterion were found to be across all age groups,

therefore allowing us to conclude Google is in fact age agnostic. No recommendations are warranted for this result.

The data provided in Table 2 in Chapter 4 allows one to address the research question as to what percentage of data was found to be incorrect. The participants in the study collected 568 pieces of data on the target across all the tool sets used. The incident rate of false or inaccurate data was found to be 1.4% or 8 pieces of data. The recommendation which can be drawn from this data is while there are many false pieces of data on the internet, when it comes to the tools listed in this study, the accuracy of the data seems to be very high.

The next answer derived from the data in Table 3 of Chapter 4 showed no participants made use of a paid service during the study exercise. The data in the study database also was used to answer the final two research questions as to whether the location of the experiment played a role in the project and whether a combination of the location and demographics pose a risk of exposing the identities of the participants. The system used to collect the data only collected IP addresses of the participants, so the location was not able to be discerned with fine enough detail to pose a risk. Unfortunately, as the data collected was insufficient or not present for these three research questions, no recommendations can be drawn from the data.

Theme 1: News sites are favored outside of Google and Wikipedia

The participants were asked during the study to identify each tool set used during the exercise. The data collected showed after excluding the top two tool sets, Google and Wikipedia, 14 out of the 43 participants in the study made use of at least one news site. The websites included such sites as CNN, USAToday, and Biography.com. The recommendation based on this theme is as an individual becomes more famous and their information is exposed more across

media websites, they should make use of media management personnel to manage their digital exposure.

Theme 2: The results are believed valid by default

The study exercise asked each participant to collect as much and as accurate as possible data on an individual, and as the data in Table 2 in Chapter 4 shows the rate of erroneous data was only 1.4% of all data collected. Upon completion of the mosaic profile activity, the participants were asked to answer a post study questionnaire. One question asked of the participants was, “How would you rate the validity of your results? (use 1-100%)”.

Out of the 43 participants, 38 participants self-rated their results as at a minimum of 75%, while the remaining 5 participants placed their results between 50% and 74%. The results show no participant placed the validity of their results below 50%. The recommendation based on this theme is while the incident rate of errors was very low at 1.4%, it is the perceived validity of the results identified by the individual which should be of concern. The cybersecurity industry needs to be more skeptical about the data that exists online and should work on a concerted effort to address the perceived validity of online data.

Theme 3: Prior experience with these types of activities does not affect success

The participants were asked at the completion of the study exercise to answer a short set of post study questions. One of the questions asked the participants if “Have you ever performed this type of activity in the past?”. The data collected from these two questions showed 4 out of the 43 participants had performed this type of activity in the past for a variety of reasons. The success rate of these 4 individuals were rated as one of medium quality, one of low quality, and two of average quality. The recommendation based on this theme is prior experience at this type of data collection activity does not guarantee a higher level of success in mosaic data collection.

This means organizations who want to make use of this type of work should test the skills of the individuals they interview rather than trust a stated history of experience.

Theme 4: Time is a factor

The study exercise was limited to a 25-minute time frame for each participant to carry out the mosaic profile collection. The researcher was concerned as to whether the limited time frame may have affected the results, so a set of questions were asked after the study including, “Did you feel that the time given to do this exercise was enough?” and “If given additional time what percentage more data do you think you could discover? (use 1-100%)”. The data showed 19 out of the 43 participants felt the time limit did affect their results.

The data also showed 21 out of the 43 participants felt more time would allowed them to have increased their results by at least 50%. The recommendations based on this theme are first, a mosaic profile or any other form of mosaic work is not an exercise which can be rushed, rather it takes time to develop as an individual works through the stream of data they must sift through. The second recommendation is mosaic work should never be considered completed, as the data behind the work is constantly changing and being indexed on a consistent basis.

Theme 5: Privacy concerns only play a role for some participants

The issue of privacy played a key role throughout the entirety of the study and its associated exercises. Whether the focus was on ensuring the privacy of the individual being researched, Aaron Hernandez, or ensuring the anonymity of the participants who partook in the study, the underlying motive has been to ensure and protect everyone’s privacy. The literature review covered some of the negative effects one can expect when privacy is ignored in the pursuit of more data and open access to data. The researcher felt it was necessary to carry this concern of privacy into the exercise and created a question about privacy on the post study

questionnaire. The post study questionnaire answered by the participants asked, “Does this study make you reconsider how much information may be available on you on the internet?”. The data showed 24 out of 43 participants answered yes to this question.

The post study questionnaire also asks student for additional feedback and two participants left comments addressing privacy. One participant, paper id # 2, expressed unease stating, “Some of the information requested is private and I do not feel comfortable trying to find it.”. Whereas another participant, paper id #19, expressed very little concern stating, “I don’t feel very concerned since the target I was given was a high-profile murderer/football player. I’m a nobody compared to him.”. The recommendation based on this theme is the cybersecurity industry must do more to educate internet users to the prevalence of their data existing online without their knowledge. The prevalence of this data will not be altered until the majority of internet users demand its removal.

Theme 6: Finding the exercise fun seemed to increase the success rate

The research study was designed and developed to identify a key demographic or tool set which would increase the success of a mosaic profile exercise, and while those questions were answered by the data, an unknown variable was also identified. The post study questions included an open-ended question requesting additional insight on each participant’s thinking. Participants were asked, “Is there anything else you would like to share such as feelings or concerns?”. During the coding process, in particular the pattern coding step, at least 6 participants noted some aspect of fun or enjoyment in the study.

One participant, paper id # 4, expressed their feelings by simply stating, “Fun study”. The data in each of the 6 cases showed the success rate of each of the participants who found the exercise fun was calculated as being above average. This data showed a possible existence of a

correlation between individuals who find the activity to be interesting or fun to potentially be better at the overall mosaic exercise. The recommendation based on this theme is, in addition to additional research being needed to identify the underlying reason behind an individual finding enjoyment in this type of exercise, organizations should focus on identifying individuals who present a sense of joy or amusement when presented with this task and encourage and develop these individuals further.

Theme 7: Date of Birth and basic demographics classified easiest to find

The participants were asked in a post study question to identify "What information was the easiest to find?". The data shows 29 out of 43 of the participants found the date of birth and demographic fields to be the easiest pieces of data to locate online. The results also show 38 of the 43 participants correctly identified the date of birth of the subject individual. The recommendation based on this theme is the usage of date of birth as a security question should be reconsidered as the prevalence of such data seems to be widespread.

Theme 8: Most participants found the exercise easy

The post study questionnaire answered by each of the participants asked them, "On a scale of 1 to 10, with 1 being very difficult and 10 being very easy, where would you classify this activity?". The results show 29 out of the 43 participants rated the exercise as a 5 or greater and of those participants 16 out of the 29 participants gave a rating of 7 or greater. The recommendation based on this theme is with the creation of millions of websites and millions of data services now being prevalent on the web the privacy issue needs to be revisited at a fundamental level. No longer can an individual anticipate or expect their key demographics to be private, the result is either society works on correcting these leakages of data or society must accept that with the internet comes an associated loss of privacy.

Theme 9: Perceived tech skills have no effect on success rate

The participants were asked at the beginning of the study for their demographical information, specifically, age, gender, educational level, GPA, and self-rated technical skills. The results showed the highest success rate of all 43 participants, 88.22%, was from a participant who rated their technical skills at a 2. The average technical skills rating for all participants was 7.2. The results also show the average quality and low-quality results come from participants who claimed an average technical skill of 7.1. The recommendation based on this theme correlates back to Theme 6 with the primary driver of what seems to identify an individual to produce a more successful mosaic profile is the fact the individual finds fun or enjoyment in the task not, their technical skills, gender, age group or any other demographic reviewed in this study.

Recommendations for Further Research

The literature review covered the fact a literature gap exists when looking at the Mosaic Theory of Intelligence outside of the federal court system. While this grounded theory study was carried out to extend the body of knowledge outside of the courts, it is just the beginning of the work which must be undertaken to fully understand how technology and the Mosaic Theory of Intelligence are going to progress into the future. The rest of this chapter will examine this need and provide some recommendations for further research.

A recommended area of further study should be a mixed methods analysis of exactly how much of an effect the emerging variable of finding mosaic work fun has on the success of an individual to do mosaic work. While this variable emerged from the qualitative results of the post study questionnaire, the possible identification of a subset of individuals who naturally succeed at this work could provide organizations with an advantage when recruiting individuals.

Another area of recommended further study should be a qualitative analysis of identifying the tools or tool sets successful in building a mosaic profile with the explicit removal of Google and Wikipedia from the study. While these two tool sets emerged as the most reliable tools used by the average individual with an interest in cybersecurity, it would be beneficial to run a study removing these particular tool sets to continue to develop the list of the most effective tool sets. The possible study would force the participants to be more creative in their tool choices which would hopefully expand the pool of tools to be analyzed. Finally, as some of the tools used were different than what was anticipated, a possible last area of research would be to identify from both study participants and subject matter experts a list of tools and discern what tools are tools and what tools are simply an avenue to the tools.

Chapter Summary

The problem statement motivating this study expressed a concern about the technology to mine data being available to everyone, and mining can have far reaching criminal ramifications by building a mosaic of information on any given topic. The purpose of this study was to identify the most successful tool sets which are the most successful in building a detailed mosaic profile. The study revealed Google and Wikipedia are by far the most successful tool sets for mosaic work, so the study has succeeded in adding to the body of knowledge with regards to the Mosaic Theory of Intelligence.

The study through its usage of a post study questionnaire, the collection of demographics, and the use of a two-cycle coding method was able to extend the data results beyond the research questions. The researcher was able to develop from the data an additional nine themes which extended the results of the study to additional conclusions and recommendations. The most

surprising of these themes was the emergence of the variable of enjoyment or fun as a demographic of the individual and their success.

This study approached a topic, the Mosaic Theory of Intelligence, which has not gotten much exposure outside the court system. The researcher chose this topic after identifying some concerning patterns in previous research and went forward with this study to start developing an interest in the topic in the cybersecurity field as well as providing a small part of the groundwork of knowledge future research can be built upon. As noted previously in this study, study results cannot be generalized to a larger population, and these results pertain only to this particular study.

References

Alla Ali Bin Ali Ahmed, et al. v Barack H. Obama, et al., No. 05-1678 (United States District Court for the District of Columbia June 11, 2009).

Axinn, W. G., Link, C. F., & Groves, R. M. (2009). *Responsive survey design, demographic data collection, and models of demographic behavior*. (NSFG Survey Methodology Working Papers, No. 09-005). Retrieved University of Michigan, Institute for Social Research website: https://www.psc.isr.umich.edu/pubs/pdf/ng09-005.pdf

Baggett, R. K., Foster, C. S., & Simpkins, B. K. (2017). *Homeland security technologies for the 21st century.* Santa Barbara, CA: Praeger.

Bazzell, M. (2016). *Open source intelligence techniques: Resources for searching and analyzing online information.* Seattle, WA: CreateSpace Independent Publishing Platform.

Bedi, M. (2014). Social networks, government surveillance, and the Fourth Amendment Mosaic Theory. *Boston University Law Review, 94*, 1809-1880. Retrieved from http://www.bu.edu/bulawreview/

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-49. doi:10.1016/j.inffus.2015.08.005

Bellovin, S. M., Hutchins, R. M., Jebara, T., & Zimmeck, S. (2013). When enough is enough: Location tracking, Mosaic Theory, and machine learning. *New York University Journal of Law and Liberty, 8*, 555-628. Retrieved from http://lawandlibertyblog.com/

Benn J, Arnold G, D'Lima D, et al. (2015). Evaluation of a continuous monitoring and feedback initiative to improve quality of anaesthetic care: a mixed-methods quasi-experimental

study. *Health Services and Delivery Research, 3*(32). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK305883/pdf/Bookshelf_NBK305883.pdf

Carpenter v United States, No. 16-402 (United States Supreme Court November 29, 2017).

Carter, A. (2015). *The future of the Fourth Amendment: Guardian of the First Amendment.* Minneapolis, MN: University of Minnesota.

Citron, D., & Gray, D. (2013). Addressing the harm of total surveillance: A reply to Professor Neil Richards. *Harvard Law Review Forum, 126,* 262-274. Retrieved from http://harvardlawreview.org/topics/forum/

Coleman, G. (2014). *Hacker, Hoaxer, Whistleblower, Spy: The many faces of Anonymous.* [Kindle version]. Retrieved from http://www.amazon.com/

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Thousand Oaks, CA: Sage Publications.

Crampton, J. W. (2015). Collect it all: national security, big data and governance. *GeoJournal, 80*(4), 519-531. Retrieved from https://link.springer.com/journal/10708

Cray, Inc. (2016, November 5). *History: Seymour Cray & Cray Research to Cray Inc. | Cray*. Retrieved from Cray: http://www.cray.com/company/history

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Creswell, J. W. (2014). *Research design* (4th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

de Montjoye, Y.A., Hidalgo, D. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports, 3*, 1-5. doi:10.1038/srep01376

Donohue, L. (2015). Section 702 and the collection of international telephone and internet content. *Harvard Journal of Law and Public Policy, 38*(1), 117-267. doi:10.2139/ssrn.2436418

Douglas, D. M. (2016). Doxing: A conceptual analysis. *Ethics and Information Technology, 18*(3), 199-210. doi:10.1007/s10676-016-9406-0

Duhigg, C. (2012, February 19). Psst, you in aisle 5. *The New York Times Magazine*, p. MM30.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS One, 6*(12). doi:10.1371/journal.pone.0028071

Etzioni, A. (2012). The privacy merchants: What is to be done? *University of Pennsylvania Journal of Constitutional Law, 14*(4), 929-951. Retrieved from https://www.law.upenn.edu/journals/conlaw/

Etzioni, A., & Rice, C. J. (2015). *Privacy in a cyber age: Policy and practice.* New York, NY: Palgrave Macmillan.

Exec. Order No. 13,292, 68 FR 15315 (2003)

Ferguson, A. G. (2012). Predictive policing and reasonable suspicion. *Emory Law Journal, 62*(2), 259-325. Retrieved from http://law.emory.edu/elj/

Flick, U. (2014). *An introduction to qualitative research.* Thousand Oaks, CA: SAGE Publications.

Franz, N. (2017). Targeted killing and pattern-of-life analysis: Weaponised media. *Media, Culture & Society, 39*(1), 111-121. doi:10.1177/0163443716673896

Gates, C., & Matthews, P. (2014). Data is the new currency: Becoming a data whore. *NSPW '14 Proceedings of the 2014 New Security Paradigms Workshop* (pp. 105-116). doi:10.1145/2683467.2683477

Gentithes, M. (2015). Tranquility & mosaics in the Fourth Amendment: How our collective interest in constitutional tranquility renders data dragnets like the NSA's telephony metadata program a search. *Tennessee Law Review, 82*, 1-38. Retrieved from https://tennesseelawreview.org/

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012-1014. doi:10.1038/nature07634

Goodman, M. (2015). *Future crimes.* New York, NY: Doubleday.

Graziano, S. (2016). An unconstitutional work of art: Discussing where the Federal Government's discrete intrusions into one's privacy become an unconstitutional search through Mosaic Theory. *Minnesota Journal of Law, Science and Technology, 17*(2), 977-1011. Retrieved from http://scholarship.law.umn.edu/mjlst/

Greengard, S. (2012). Policing the future. *Communications of the ACM, 55*(3), 19-21. doi:10.1145/2093548.2093555

Guest, G., & Namey, E. E. (2015). *Public health research methods.* Los Angeles, CA: Sage.

Haggard, S., & Lindsay, J. R. (2015). North Korea and the Sony hack: Exporting instability through cyberspace. *AsiaPacific Issues, 117*, 1-8. Retrieved from http://www.asia-studies.com/aps.html

Hajili, N., & Lin, X. (2016). Exploring the security of information sharing on social networking sites: The role of perceived control of information. *Journal of Business Ethics, 133*(1), 111-123. doi:10.1007/s10551-014-2346-x

Halkin v. Helms, 598 F.2d 1, 3 (United States Court of Appeals for the District of Columbia Circuit 1978).

Hayman, J. W. (2015). *Case study: Suggested best practices for redacting U.S. Army aviation accident reports to reduce opportunities for doxing of re-identified U.S. Army aircrew* (Unpublished doctoral dissertation). Capitol Technology University, Laurel.

Hays, C. L. (2004, November 14). What Wal-Mart knows about customers' habits. *The New York Times*. Retrieved from http://www.nytimes.com

Heuer, Jr., R. J. (1999). *Psychology of intelligence analysis.* Fairfax, VA: Central Intelligence Agency.

Hill, K. (2011, February 8). Did Mark Zuckerberg's stalker get his intel from Gawker?. Forbes. Retrieved from Forbes: http://www.forbes.com

Hilsman, Jr., R. (1952). Intelligence and policy-making in foreign affairs. *World Politics, 5*(1), 1-45. doi:10.2307/2009086

Hu, M. (2015). Taxonomy of the Snowden disclosures. *Washington and Lee Law Review, 72*(4), 1679-1767. Retrieved from http://lawreview.journals.wlu.io/

Jaffer, J. (2010). The Mosaic Theory. *Social Research: An International Quarterly, 77*(3), 873-882. Retrieved from http://www.socres.org/

James, D. V., & MacKenzie, R. D. (2018). Stalking and harrassment. In J. L. Ireland, P. Birch, & C. A. Ireland (Eds.), *The Routledge International handbook of human aggresion: Current issues and perspectives*. New York, NY: Routledge.

Joh, E. E. (2014). Policing by numbers: Big data and the Fourth Amendment. *Washington Law Review, 89*(1), 35-68. Retrieved from https://www.law.uw.edu/wlr

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*(15), 5802-5805. doi:10.1073/pnas.1218772110

Kugler, M. B., & Strahilevitz, L. (2015). Surveillance duration doesn't affect privacy expectations: An empirical test of the Mosaic Theory. *University of Chicago Public Law Working Paper*, 1-55. Retrieved from http://chicagounbound.uchicago.edu/public_law_and_legal_theory/

Liu, B., Sheth, A., Weinsberg, U., Chandrashekar, J., & Govindan, R. (2013). AdReveal: Improving transparency into online targeted advertising. *HotNets-XII Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks* (pp. 1-7). doi:10.1145/2535771.2535783

Lohr, S. (2013, February 4). Searching for origins of the term ‘big data’. *The New York Times*, p. B4.

Marr, B. (2015). *Big data: Using SMART big data, analytics and metrics to make better decisions and improve performance.* West Sussex, England: Wiley.

Mathews, R. S., Aghili, S., & Lindskog, D. (2013). *A study of doxing, its security implications and mitigation strategies for organizations.* Edmonton, AB: Concordia University College of Alberta.

Miles, M. B., Huberman, A., & Saldana, J. (2013). *Qualitative data analysis: A methods sourcebook* (3rd ed.). London, England: Sage Publications.

Mornin, J. D. (2014). NSA metadata collection and the Fourth Amendment. *Berkeley Technology Law Journal, 29*(4), 985-1006. Retrieved from http://btlj.org/

Neef, D. (2014). *Digital exhaust: What everyone should know about big data, digitization and digitally driven innovation.* Upper Saddle River, NJ: Pearson Education.

Nieswiadomy, R. M. (2008). *Foundations of nursing research* (5th ed.). Upper Saddle River, NJ: Pearson Education.

Olson, P. (2012). *We are Anonymous*. [Kindle version] Retrieved from http://www.amazon.com

Palmer, B. W. (2015). Study participants and informed consent. *Monitor on Psychology*, 46(8), 62. Retrieved from http://www.apa.org/monitor/

Pozen, D. E. (2005). The Mosaic Theory, national security, and the Freedom of Information Act. *The Yale Law Journal, 115*(3), 628-679. Retrieved from http://www.yalelawjournal.org/

Punch, K. F. (2014). *Introduction to social research: Quantitative and qualitative approaches* (3rd ed.). London, England: Sage Publications.

Richelson, J. T. (2015). *The U.S. intelligence community.* Boulder, CO: Westview Press.

Rosenzweig, P. (2017, November 29). In defense of the Mosaic Theory. *LawFare.* Retrieved from: https://www.lawfareblog.com/defense-mosaic-theory

Ruslanovich, G. R., & Alekseevna, S. I. (2016). Social networks as a source of information about the applicants. *Symbol Science, 1,* 38-40. Retrieved from http://cyberleninka.ru/article/

Sagar, R. (2015). Against moral absolutism: Surveillance and disclosure after Snowden. *Ethics and International Affairs, 29*(2), 145-159. doi:10.1017/S0892679415000040

Saldana, J. (2015). *The coding manual for qualitative researchers* (3rd ed.). London, England: Sage Publications.

Salkind, N. J. (2016). *Exploring research* (9th ed.). Upper Saddle River, NJ: Pearson Education.

SAS Institute. (2016). *How Walmart makes data work for its customers.* Cary: SAS Institute.

Scalia, A. (1982). The Freedom of Information Act has no clothes. *Regulation, 6*(2), 14-19. Retrieved from https://www.cato.org/regulation/

Schlabach, G. R. (2015). Privacy in the cloud: The Mosaic Theory and the Stored Communications Act. *Stanford Law Review, 67,* 677-721. Retrieved from https://www.stanfordlawreview.org/

Schneier, B. (2015, January 2). Doxing as an attack [Web log comment]. Retrieved from https://www.schneier.com/blog/archives/2015/01/doxing_as_an_at.html

Selva, L., Shulman, W., & Rumsey, R. (2016). Rise of the Mosaic Theory: Implications for cell site location tracking by law enforcement. *The John Marshall Journal of Information Technology & Privacy Law, 32*(4), 235-256. Retrieved from http://repository.jmls.edu/jitpl/

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information,* (22), 63-75. Retrieved from http://www.iospress.nl/journal/education-for-information/

SI Wire. (2017, April 19). Key moments in the life of Aaron Hernandez. Retrieved from Sports Illustrated: https://www.si.com/nfl/2017/04/19/aaron-hernandez-timeline-career-life

Singer, N. (2012, June 17). You for sale. *The New York Times,* p. BU1.

Song, C., Qu, Z., Blumm, N., & Barabasi, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018-1021. doi:10.1126/science.1177170

Sprague, R. (2015). Welcome to the machine: privacy and workplace implications of predictive analytics. *Richmond Journal of Law & Technology, XXI*(4), 1-46. Retrieved from http://jolt.richmond.edu/

The Economist. (2011, April 14). The mosaic defence. *The Economist*. Retrieved from http://www.economist.com.

Tokson, M. (2011). Automation and the Fourth Amendment. *Iowa Law Review, 96*, 581-647. Retrieved from https://ilr.law.uiowa.edu/

Trottier, D. (2016). *Social media as surveillance: Rethinking visibility in a converging world.* New York, NY: Routledge.

U.S. Department of Commerce, National Institute of Standards and Technology. (2010). *Guide to protecting the confidentiality of personally identifiable information (PII).* (NIST Publication No. 800-122). Retrieved from http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

United States of America v. Lawrence Maynard, 615 F.3d 544 (2010) (United States Court of Appeals, District of Columbia Circuit August 6, 2010).

United States Senate. (2012). *Federal support for and involvement in state and local fusion centers.* Permanent Subcommittee on Investigations. Washington DC: United States Senate.

United States v. Jones, No. 10-1259 (United States Supreme Court January 23, 2012).

Walmart Corporation. (2012, August 30). *Walmart announces new search engine to power Walmart.com* [Press release]. Retrieved from http://corporate.walmart.com/_news_/news-archive/2012/08/30/walmart-announces-new-search-engine-to-power-walmartcom

Walsh, C. (2014). Surveillance technology and the loss of something a lot like privacy: an examination of the "Mosaic Theory" and the limits of the Fourth Amendment. *St. Thomas Law Review, 24*(2), 169-251. Retrieved from http://stthomaslawreview.org/

Ward, J. S., & Barker, A. (2013). *Undefined by data: A survey of big data definitions.* Ithaca: Cornell University Library. Retrieved from https://arxiv.org/pdf/1309.5821v1.pdf

William Collins Sons & Co. Ltd. (n.d.). *Collins English dictionary - complete & unabridged 2012 digital edition*. Retrieved from https://www.collinsdictionary.com.

Wittes, B. (2011). *Databuse: Digital privacy and the Mosaic.* Retrieved from The Brookings Institution, Governance Studies website: https://www.brookings.edu/wp-content/uploads/2016/06/0401_databuse_wittes.pdf

Yin, R. K. (2016). *Qualitative research from start to finish* (Second ed.). New York, NY: The Guilford Press.

Zolfagharifard, E. (2015, May 26). How the Apple Watch is as powerful as two Cray supercomputers: Graphic reveals the incredible advances in computing power. *Daily Mail Online.* Retrieved from: http://www.dailymail.co.uk

Züll, C. (2016). *Open-Ended Questions. GESIS Survey Guidelines.* Mannheim, Germany: GESIS – LeibnizInstitute for the Social Sciences. doi: 10.15465/gesis-sg_en_002

Song, C., Qu, Z., Blumm, N., & Barabasi, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018-1021. doi:10.1126/science.1177170

APPENDIX A: KEY LITERATURE REVIEW SEARCH TERMS

artificial intelligence

big data

business analytics

Central Intelligence Agency (CIA)

corporate knowledge

data mining

doxing

Freedom of Information Act (FOIA)

Google Dorking

intelligence

knowledge

machine learning

Mosaic Profile

Mosaic Theory of Intelligence

Mosaic Theory

National Security Agency (NSA)

open government

Open Source Intelligence (OSINT)

profiling

regulatory filings

social network

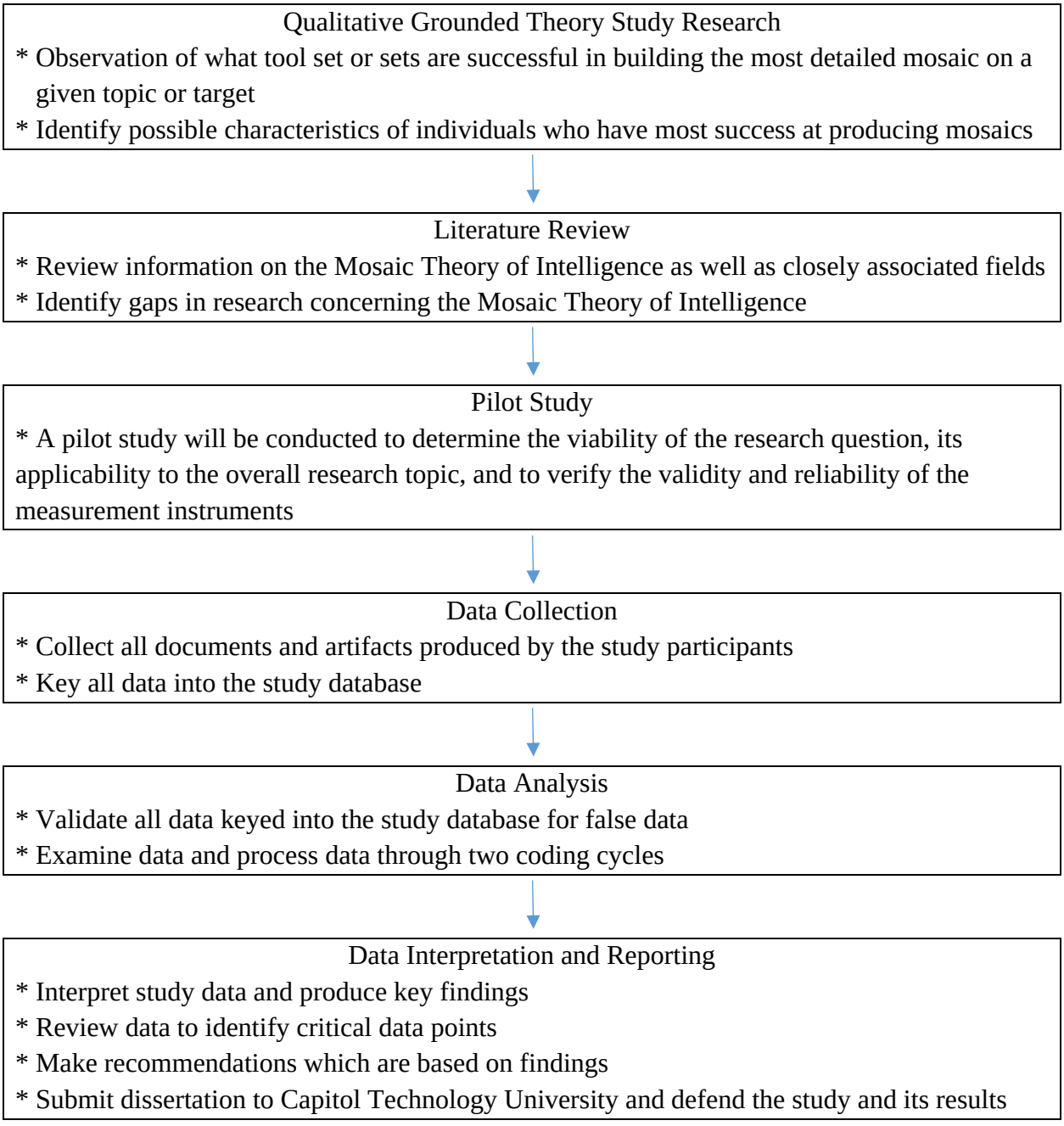
surveillance

APPENDIX B: LITERATURE SEARCH

Key Word Search	Peer Reviewed Works Reviewed	Germinal Works Reviewed	Books Reviewed	Studies Reviewed	Totals
Mosaic Theory-Big Data					
Artificial Intelligence and Machine Learning	12	4	2	2	**20**
CIA and NSA	46	37	3	15	**101**
Profiling, Google Dorking, Doxing, and Mosaic Profile	51	27	3	5	**86**
Intelligence, Knowledge, Corporate Knowledge, and Regulatory Filings	62	26	2	1	**91**
Big Data, Data Mining, and Business Analytics	34	21	2	1	**58**
Freedom of Information Act & Open Government	35	28	1	0	**64**
Social Network	43	24	2	8	**77**
Surveillance	73	42	1	9	**125**
Open Source Intelligence	4	4	0	2	**10**
Mosaic Theory of Intelligence	62	29	0	5	**96**
Research Methodologies					
Qualitative	25	11	0	25	**61**
Quantitative	23	8	0	23	**54**
Total Documents Reviewed	**470**	**245**	**16**	**96**	**843**

APPENDIX C: METHODOLOGY MAP

Research Methodology



APPENDIX D: STUDY METHODOLOGY

Outline – Protocol for Mosaic Profile Study

1. Introduction to the study and purpose of the protocol
    a. Study questions
    b. Study objectives
    c. Theoretical framework for the study
    d. Role of protocol in guiding the study investigator
2. Data Collection Procedures
    a. Name of the site for the study and participants
    b. Data collection plan, calendar for the study
    c. Study preparation
    d. Chain of custody of the data
3. Study Database
    a. Database plan
    b. Coding framework
4. Study Coding Procedures
    a. Coding procedures after the artifacts and documents are collected
        i. Primary coding
        ii. Secondary coding
5. Outline of the study report
    a. Introduction to the study
    b. Phases of the study
    c. Outcome of the study

d. Analysis of the study artifacts

6. Study Questions

    a. Which tool sets demonstrated the most detail in completing a successful mosaic profile?

    b. Which demographics of individuals demonstrated the highest rate of success in completing a successful mosaic profile?

    c. Which demographics of individuals made use of the particular tool sets?

    d. Was there a significant correlation of one age group to Google?

    e. Is Google age agnostic?

    f. What percentage of data found was incorrect, false flag?

    g. Did any subjects make use of a paid service?

    h. If so, did it increase accuracy or completeness?

    i. Will the location of the experiment play a role in the project?

    j. Does the combination of age, gender, and location of subjects pose a risk of cross correlation of data to expose subjects?

APPENDIX E: STUDY CONSENT FORM

Informed Consent Statement

Mosaic Profiles: The Mosaic Theory in Practice

INTRODUCTION

You are invited to join a research study to look at the best tool sets that can be used to build a successful mosaic profile. Please take whatever time you need to think about your role in this study. The decision to join, or not to join, is up to you.

In this research study, we are evaluating what tools you make use of to collect random pieces of data regarding a targeted individual. Additionally, we are looking to see if there is any unique demographic factor that plays a role in one's success in finding this information.

WHAT IS INVOLVED IN THE STUDY?

If you decide to participate you will be asked to perform a one hour exercise where you will be able to make use of whatever tools of your choice to carry out the act of intelligence collection. You will be asked to answer questions on the provided profile form and notate what tools you used to gather that data. We think this will take you 25 minutes.

The investigator may stop the study or take you out of the study at any time they judge it is in your best interest. They may also remove you from the study for various other reasons. They can do this without your consent.

You can stop participating at any time. If you stop you will not be affected in any manner.

RISKS

This study involves the following risks, no risks are anticipated. While there may be other risks that we cannot predict, we have strived to protect you from any extraneous risks. Remember to note make use of any tool that is questionably illegal.

BENEFITS TO TAKING PART IN THE STUDY?

It is reasonable to expect the following benefits from this research: none. However, we can't guarantee that you will personally experience benefits from participating in this study. Others may benefit in the future from the information we find in this study.

CONFIDENTIALITY

We will take the following steps to keep information about you confidential, and to protect it from unauthorized disclosure, tampering, or damage. All notes, questionnaires/profile forms, records and data will be kept strictly confidential. Neither your participation in this study or your response will ever be disclosed to anyone other than the researcher. All responses are aggregated together and study results will be drawn from this aggregated data, so that no individual response can be determined. Upon completion of this study, all data involving participant information will be kept secure for a three year period before being destroyed. Confidentiality can only be violated if someone references a criminal activity or poses a threat to an individual.

INCENTIVES

None.

YOUR RIGHTS AS A RESEARCH PARTICIPANT?

Participation in this study is voluntary. You have the right not to participate at all or to leave the study at any time. Deciding not to participate or choosing to leave the study will not result in any penalty or loss of benefits to which you are entitled, and it will not harm your relationship with your educational institute. To withdraw, simply notify the researcher and hand in your paperwork.

CONTACTS FOR QUESTIONS OR PROBLEMS?

Call Harry Cooper at 412-362-5907 or email harry@twcsec.com if you have questions about the study, any problems, unexpected physical or psychological discomforts, any injuries, or think that something unusual or unexpected is happening.

Contact Helen Barker, Associate Dean of Academics at (240) 965-2485 if you have any questions or concerns about your rights as a research participant.

Consent of Subject (or Legally Authorized Representative)

By continuing onto the next page, I acknowledge that I have read the informed consent statement and agree to it.

Privacy Act Statement

Authority: 5 U.S.C. § 552a authorizes the collection of this information.

Purpose: The purpose of this study is to collect information that will be used to assess the best tool sets that can be used by an individual to perform a mosaic profile. The information from this study may be used to further the knowledge of those interested in the effects that the Mosaic Theory of Intelligence may have on our national security. Findings may be shared with governmental and non-governmental institutions. Additionally findings may be published in a professional journal or presented at conferences, symposia, and scientific meetings. In none of the above cases will the data be used or reported that may identify a given individual in the study.

Participation: Participation in this study is voluntary. No compensation will be given for participating in this study. If you do not wish to participate there will be no malice. Additionally, should a participant find that they must withdraw during the study, you may do so without prejudice.

Confidentiality: All notes, questionnaires/profile forms, records and data will be kept strictly confidential. Neither your participation in this study or your response will ever be disclosed to anyone other than the researcher. All responses are aggregated together and study results will be drawn from this aggregated data, so that no individual response can be determined. Upon completion of this study, all data involving participant information will be kept secure for a three year period before being destroyed. Confidentiality can only be violated if someone references a criminal activity or poses a threat to an individual.

Routine Uses: The information provided in this study will be analyzed as part of a formal research study to be submitted to Capitol Technology University as part of their Doctorate in Cybersecurity program. The data files will be maintained solely by the principal researcher.

By continuing onto the next page, I acknowledge that I have read the privacy and confidentiality statements and agree to them.

APPENDIX F: MOSAIC PROFILE FORM FOR STUDY

Mosaic Profile (http://mosaicprofile.com/)

Target: Aaron Josef Hernandez

Aaron Hernandez, a former Tight End for the New England Patriots, has recently passed away in Shirley, MA. Outside the basics of his being born in 1989 in Connecticut, and having an interesting college, pro, and post-pro football career, little is being shared. You are an intelligence analyst hired to create an intelligence profile of the target. Your tasking is to craft an intelligence brief that provides clear, chronological, credible and detailed information on the target.

A structured form has been provided on the next few pages for the most commonly gathered information. Anything not fitting into one of the provided categories, should be entered in the section titled, Other Data. Finally, at the end of the profile, you will find a small section asking for you to answer a few basic demographical questions, please make sure you complete this section before submitting the document.

Tool Examples:

Google Search Engine, Google Scholar, Google Maps, Facebook, Facebook Graph, Classmates.com, Ancestry.com, Realtor.com, City/County/State governmental databases, Spokeo.com, Intelius, Twitter, Instagram, IMDB, PeopleFinder, USA People Search, InstantCheckmate.com, and many other websites and tools.

Just make sure you do not get stuck in a never ending chain.

Age:		Date of Birth:	
Tool		*Tool*	
Gender:		Marital Status:	
Tool		*Tool*	
Political Affiliation:		Religion:	
Tool		*Tool*	

Spouse(s):	

Tool	
Mother:	
Tool	
Father:	
Tool	
Sibling(s):	
Tool	
Children:	
Tool	
Other Relatives:	
Tool	
Current Address(s):	
Tool	
Previous Addresses:	
Tool	
Employment History: (include military/government work)	
Tool	
Language(s):	
Tool	
Education (All Levels):	
Tool	
Criminal/Legal History:	
Tool	
Photos:	
Tool	

Favorite(s): (color, movie, etc)	
Tool	
Medical History:	
Tool	
Financials:	
Tool	
Email Address(s):	
Tool	
Phone Number(s):	
Tool	
Conferences, Symposia, or other public speaking events:	
Tool	
Certifications:	
Tool	
Published Works:	
Tool	
Social Media Profiles:	
Tool	
Any news references:	
Tool	
Any scandals:	
Tool	
Member of Organization(s):	
Tool	

Security Clearances:	
Tool	
Other 1:	
Tool	
Other 2:	
Tool	
Other 3:	
Tool	
Other 4:	
Tool	
Other 5:	
Tool	

Demographical Information of Study Participant

Gender:	
Age:	
Current Education Level:	
Current GPA:	
Self Rated Tech Skills: *Scale of 1 to 10* *1 being no skills* *10 being In-Depth Skills*	

APPENDIX G: POST-STUDY QUESTIONNAIRE

Post-study Questionnaire

1) Have you ever performed this type of activity in the past? Yes or No

2) If you answered yes, please explain:

_____

_____

3) On a scale of 1 to 10, with 1 being very difficult and 10 being very easy, where would you classify this activity? _____

4) Which tool or tool set did you find to be the most useful? Please explain:

_____

_____

5) Did you feel that the time given to do this exercise was enough? Yes or No

6) If given additional time what percentage more data do you think you could discover? _____ (please use: 0-100%)

7) What piece of information was easiest to find? _____

8) How would you rate the validity of your results? _____ (please use: 0-100%)

9) Does this study make you reconsider how much information may be available on you on the internet? Yes or No

10) Is there anything else you would like to share such as feelings or concerns?

_____

_____

_____

APPENDIX H: PATTERN CODING THEMATIC ANALYSIS

Theme 1 – News sites are favored outside of Google and Wikipedia	
Representative comment/detail	Top profile was 1/3 from new sites and tool set calculations
Percentage of participants making similar comments/detail	33%
Pattern identified	News sites provide lots of detail
Emerging theme	News sites are favored outside of Google and Wikipedia

Theme 2 – The results are believed valid by default	
Representative comment/detail	No participant placed their validity below 50%
Percentage of participants making similar comments/detail	100% above 50% 88% above 75%
Pattern identified	Data assumed valid
Emerging theme	The results are believe valid by default

Theme 3 – Prior experience with these types of activities does not affect success	
Representative comment/detail	“I have done research on people in the past.” Average quality result
Percentage of participants making similar comments/detail	10%
Pattern identified	Average or low success
Emerging theme	Prior experience with these types of activities does not affect success

Theme 4 – Time is a factor	
Representative comment/detail	"exercise like this in real life could take days or weeks to be accurate" About half of the people wanted more time.
Percentage of participants making similar comments/detail	44%
Pattern identified	Participants wanted more time and felt they could increase data by 50%with more time.
Emerging theme	Time is a factor

Theme 5 – Privacy concerns only play a role for some participants	
Representative comment/detail	"I do not feel comfortable trying to find it"
Percentage of participants making similar comments/detail	56%
Pattern identified	Many expressed concern about their own data
Emerging theme	Privacy concerns only play a role for some participants

Theme 6 – Finding the exercise fun seemed to increase the success rate	
Representative comment/detail	"Fun study" and all calculated above average
Percentage of participants making similar comments/detail	14%
Pattern identified	Fun seems to increase success
Emerging theme	Finding the exercise fun seemed to increase the success rate

Theme 7 – Date of Birth and basic demographics classified easiest to find	
Representative comment/detail	89% accurately identified Date of Birth
Percentage of participants making similar comments/detail	68%
Pattern identified	Demographics identified in post study as easiest to find
Emerging theme	Date of Birth and basic demographics classified easiest to find

Theme 8 – Most participants found the exercise easy	
Representative comment/detail	“this was really engaging” ease of task average 6
Percentage of participants making similar comments/detail	67%
Pattern identified	Majority seemed to find activity easy
Emerging theme	Most participants found the exercise easy

Theme 9 – Perceived tech skills have no effect on success rate	
Representative comment/detail	Average tech skills for respondents was 7.2 out of 10 but the higher the number the more average the success rate
Percentage of participants making similar comments/detail	100%
Pattern identified	Higher tech skills do not match higher results
Emerging theme	Perceived tech skills have no effect on success rate