NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

APPLICATION OF BIG DATA ANALYTICS TO SUPPORT HOMELAND SECURITY INVESTIGATIONS TARGETING HUMAN SMUGGLING NETWORKS

by

Thomas A. Hodge

March 2018

Thesis Co-Advisors: Doug MacKinnon
Lauren Fernandez

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE APPLICATION OF BIG DATA ANALYTICS TO SUPPORT HOMELAND SECURITY INVESTIGATIONS TARGETING HUMAN SMUGGLING NETWORKS			**5. FUNDING NUMBERS**
6. AUTHOR(S) Thomas A. Hodge			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Human smuggling organizations facilitating the smuggling of aliens into the United States have an unlawful network supporting their illicit transnational activities. Identifying those networks and the key facilitators is challenging due to high volumes of disparate data.

This research focuses on how big data analytics can improve the effectiveness and efficiency of Homeland Security Investigations (HSI) targeting human smuggling networks. The purpose of this thesis is to determine whether applying big data analytics to data associated with human smuggling will make network identification of illegal aliens more efficient while producing the necessary articulable facts to substantiate enough probable cause for subsequent investigative actions. An experimental data analytics application called Citrus is used to examine the efficiency and effectiveness of data analytics supporting criminal investigations.

Citrus revealed that big data analytics can effectively produce knowledge, including probable cause, more efficiently for HSI in targeting criminal networks. The implications are significant, as the application of data analytics may reshape analytical tradecraft, and compel HSI to revamp data systems. Increases in efficiencies through data analytics may be limited without changes in judicial processes. Upgrading processing capacities for obtaining warrants will become vital as analytics becomes more prevalent.

14. SUBJECT TERMS big data analytics, human smuggling			**15. NUMBER OF PAGES** 97
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

APPLICATION OF BIG DATA ANALYTICS TO SUPPORT HOMELAND SECURITY INVESTIGATIONS TARGETING HUMAN SMUGGLING NETWORKS

Thomas A. Hodge
Supervisory Intelligence Research Specialist,
Homeland Security Investigations, Phoenix, Arizona
B.S., Southern Illinois University, 2001
M.A., Arizona State University, 2014

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF ARTS IN SECURITY STUDIES
(HOMELAND SECURITY AND DEFENSE)**

from the

**NAVAL POSTGRADUATE SCHOOL
March 2018**

Approved by: Doug MacKinnon
Thesis Co-Advisor

Lauren Fernandez
Thesis Co-Advisor

Erik Dahl
Associate Chair for Instruction
Department of National Security Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Human smuggling organizations facilitating the smuggling of aliens into the United States have an unlawful network supporting their illicit transnational activities. Identifying those networks and the key facilitators is challenging due to high volumes of disparate data.

This research focuses on how big data analytics can improve the effectiveness and efficiency of Homeland Security Investigations (HSI) targeting human smuggling networks. The purpose of this thesis is to determine whether applying big data analytics to data associated with human smuggling will make network identification of illegal aliens more efficient while producing the necessary articulable facts to substantiate enough probable cause for subsequent investigative actions. An experimental data analytics application called Citrus is used to examine the efficiency and effectiveness of data analytics supporting criminal investigations.

Citrus revealed that big data analytics can effectively produce knowledge, including probable cause, more efficiently for HSI in targeting criminal networks. The implications are significant, as the application of data analytics may reshape analytical tradecraft, and compel HSI to revamp data systems. Increases in efficiencies through data analytics may be limited without changes in judicial processes. Upgrading processing capacities for obtaining warrants will become vital as analytics becomes more prevalent.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

API	application programming interface
AS	advanced search
ATISMART	accelerated-time simulation for traffic flow
BIC	Business Integrity Commission
BRIS	border risk identification system
CBP	Customs and Border Protection
CI	collective intelligence
CIO	chief information officer
CO	co-occurrence
DHS	Department of Homeland Security
DSI	document set investigation
HFT	high frequency trading
HSI	Homeland Security Investigations
HSO	human smuggling organizations
ICE	Immigration and Customs Enforcement
OBP	Office of Border Patrol
QE	quality of evidence
SWBA	Southwest Border Anti-Money Laundering Alliance

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The continuous pressure from large volumes of aliens attempting to enter the country illegally creates a persistent challenge for the 20,000 office of border patrol (OBP) agents attempting to apprehend hundreds of thousands of aliens annually.[1] Human smuggling organizations (HSO) facilitating the smuggling of aliens into the United States have an unlawful network supporting their illicit transnational activities. Identifying those networks and the key facilitators is challenging due to high volumes of disparate data.

The research question for this thesis is how can big data analytics improve the effectiveness and efficiency of Homeland Security Investigations (HSI) targeting human smuggling networks? The purpose of this thesis is to determine whether applying big data analytics to data associated with human smuggling will make network identification of illegal aliens more efficient while producing the necessary articulable facts to substantiate enough probable cause for subsequent investigative actions.

An experimental data analytics application called Citrus was used to examine the efficiency and effectiveness of data analytics supporting criminal investigations. Citrus is a free-for-government use software tool designed by Sandia National Laboratories. The Department of Homeland Security Science and Technology provided this application to HSI for testing and evaluation. Citrus was built to discover, trend, and link disparate data and is being used as an analytics application developed and implemented by HSI Phoenix.

The research compares the results of queries being conducted manually to a same set of problems with Citrus, which measures the total number of targets, network identification, whether enough documentation is obtained to justify problem cause, and the timelessness of the results. To test the efficiency and the effectiveness of data analytics with Citrus, two types of tests were conducted, search and discovery.

- The search query consists of extrapolating the highest volume of phone numbers from 45, 90, and 180 days of Arizona border patrol station reports

[1] "Stats and Summaries," U.S. Customs and Border Protection, 1, accessed August 21, 2017, https://www.cbp.gov/newsroom/media-resources/stats?title=Border+Patrol.

and correlating those results to other indexes relating to financial and communication data.

- The discovery query tests two reports created with Citrus. The quality of evidence report measures levels of probable cause against known and unknown entities from within a large dataset. The co-occurrence report is designed to determine the total number of phone number pairings from thousands of OBP reports, which has led to potential human smuggling network identification.

Efficiency is measured by comparing the time it takes to complete a search and a discovery query manually versus using the Citrus application, in identifying subjects and human smuggling networks. Effectiveness is measured by comparing the number of subjects and networks identified manually versus using the Citrus application for the search and discovery queries.

Determining whether *probable cause* can be obtained simultaneously while increasing the efficacy of human smuggling analysis is also a measure of effectiveness. Understanding the level of probable cause required to substantiate a warrant is important in targeting HSO networks, as it becomes the basis for proving human smuggling criminal violations. Probable cause exists when reasonable suspicion related to human smuggling activities can be articulated in a legal sense. For this study, probable cause exists when specific phone numbers are recorded frequently in OBP reports linked to communications records with other phone numbers also associated with OBP reports, and combined with financial transactions used by the same phone numbers.

Big data analytics is generally defined as large volumes of data that require large computer capacities and applications to discover meaningful insights.[2] The value of big data is realized when applied to understanding large datasets that eventually lead to making

[2] Emerging Technologies Big Data Community of Interest, *HM Government Horizon Scanning Programme Emerging Technologies: Big Data* (London: HM Government, 2014), 2, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389095/Horizon_Scanning_-_Emerging_Technologies_Big_Data_report_1.pdf.

better decisions.[3] Big data analytics as a research topic is relatively new and has the interest of several industries, including the government.[4]

Data is growing at exponential rates.[5] Scores of authors promote the use of big data analytics and the potential value. From garnering greater insights to potentially altering the standard scientific method, big data analytics is a growing technology that has great benefits. The public sector, however, is developing the technology at slower rates compared to the private sector.[6] The application of data analytics may allow the conversion of large datasets into insights that result in better decisions for the public sectors.

Although big data analytics presents value and opportunity for the public sector, academic literature is scarce in supporting big data analytics in practice for public entities, according to Gamage.[7] Theoretical frameworks for big data analytics are also lacking. Additionally, literature for managers is scant that describes how best to develop and integrate big data analytics.[8] Moreover, a framework specifically for law enforcement or for federal investigations is nonexistent. HSI has a broad mission and validating how analytics is applied against different criminal programmatic areas is necessary before applying analytics nationally.[9]

The results of this research demonstrate that Citrus works well for triaging large amounts of data. The efficiency of Citrus to sift through voluminous amounts of reporting and communication and financial data was exponentially better. The effectiveness also

---

[3] Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management* 35, no. 2 (April 2015): 140, https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

[4] Jonathan Seddon and Wendy L. Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," *Journal of Business Research* 70, no. C (2017): 301.

[5] Ase Dragland, "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years," *ScienceDaily*, 1, May 22, 2013, https://www.sciencedaily.com/releases/2013/05/130522085217.htm.

[6] Pandula Gamage, "New Development: Leveraging 'Big Data' Analytics in the Public Sector," *Public Money & Management* 36, no. 5 (2016): 385.

[7] Gamage, 385.

[8] Gerard George, Martine R. Haas, and Alex Pentland, "Big Data and Management," *Academy of Management Journal* 57, no. 2 (2014): 321.

[9] "Homeland Security Investigations," U.S. Immigration and Customs Enforcement, 1, accessed October 15, 2017, https://www.ice.gov/hsi.

proved to be substantially better with Citrus when compared to the same analysis process conducted manually. The amount of additional reports and the capability to calculate probable cause was decisively more effective with Citrus, even though dataset was limited.

Investigative discoveries may be made more efficient and effective with data analytics. The implications for HSI are significant, particularly relating to changing analytical tradecraft, revamping data systems, and increasing investigative process capacities as summarized as follows.

(1) Analytical Tradecraft

The application of data analytics may reshape analytical tradecraft. Citrus demonstrates that analysts are able to create and answer hypotheses on a deeper level that leads to greater network identifications. With data analytics, new forms of analytical tradecraft can be produced, as data analytics potentially creates an unlimited means of reviewing and analyzing data in bigger ways.

(2) Merging Data

Advancing data analytics requires HSI to remove barriers between data systems, which are imperative to maximizing the value of data analytics. HSI should move beyond systems designed to work well against one particular dataset to aggregated data from across the breadth of systems. Revamping the current HSI systems architecture may be necessary in evolving to a more data-driven organization through analytics.

(3) Investigative Processes

With increases in efficiencies through data analytics, the analysis process and production may outpace investigation processes. If analytics can immediately identify which entities or persons within the data already possess enough probable cause exist, theoretically the HSI investigation process can be accelerated. This acceleration will have an impact on the judicial process, particularly relating to processing capacities of the courts. Upgrading the processing capacities for obtaining warrants will become vital as analytics becomes more prevalent.

In summary, HSI can be more effective and efficient in investigating and targeting criminal networks with data analytics. Exploring and investing further in the technology should be a high priority, as data analytics offers HSI enormous potential.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to express my deepest appreciation for my wife's support throughout my time attending the Naval Postgraduate School, Center for Homeland Security and Defense. She was incredibly encouraging and provided me ample time to complete my studies. I would not have been able to complete my thesis without her backing. She made the challenging experience much more enjoyable. For her, I am very grateful.

Dr. Doug Mackinnon and Dr. Lauren Fernandez were fantastic thesis advisors. They were engaging, responsive, and supportive. I am most appreciative of their dedication throughout the process. Their feedback was challenging and productive. I am thankful to have been under their advisement and owe them a debt of gratitude.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Organizations facilitating the smuggling of aliens into the United States have an unlawful network supporting their illicit transnational activities. Identifying those networks and the key facilitators is challenging due to high volumes of disparate data. The purpose of this thesis is to determine whether applying big data analytics to data associated with human smuggling will make network identification of illegal aliens more efficient while producing the necessary articulable facts to substantiate enough probable cause for subsequent investigative actions.

Valuable evidentiary information relating to human smuggling networks resides in disparate data principally through reports gathered after personnel from the Office of Border Patrol (OBP) apprehend migrant aliens. The OBP reports outline the details of the aliens' migration patterns, as well as the subjects and associated networks responsible for the smuggling activities; they are disseminated to Homeland Security Investigations (HSI), which is the largest investigative agency for the Department of Homeland Security (DHS). HSI gathers the reports and manually sifts through the volumes of information in an effort to identify the human smuggling networks. Pertinent information, such as names of the smugglers, phone numbers, and financial accounts that assist in the identification of the human smuggling networks embedded in the OBP reporting, is extracted for further analysis. Subpoenas are issued to collect more information stemming from the data collected in the original OBP reports. All this information is compiled for HSI analysts to analyze, identify, and prioritize the most prolific human smuggling networks. This process is largely done manually.

Manually filtering through and analyzing high volumes of reports is time consuming and inefficient. Moreover, reliably being able to identify subjects and networks is virtually impossible primarily because of human limitations. With large volumes of data, the ability to process so much information exceeds human capacities. Analysts are unable to keep pace with increasing volumes of data, potentially leaving valuable insights undiscovered. Under a manual analysis process, HSI is unable to measure the accuracy of data analysis if accessible data is unprocessed. Meaning, if large portions of data,

particularly evidentiary data, is unable to be evaluated, confidence in analysis is reduced. As a result, HSI's confidence in network analysis is immeasurable.

This research is important because of the potential big data analytics holds in improving HSI investigations. Big data analytics has the potential to identify threat networks early, which allows agencies to disrupt transnational criminal organizations more effectively.

Leveraging big data analytics has been directed at the national level. In 2012, the Obama administration launched a $200 million dollar initiative to advance big data research efforts across the breadth of several federal agencies, including such agencies as the Department of Defense and Department of Energy.[1] Furthermore, relative to the national-level big data strategy, the DHS provides more explicit guidance for the use of big data for investigators. According to the 2014 *Quadrennial Homeland Security Review*, homeland security agencies "must continually improve [their] ability to make sense of vast amounts of intelligence and other information—the so-called 'big data' challenge—while rigorously protecting the privacy and civil liberties of Americans."[2]

Higher volumes of data increase the difficulty for investigators to "connect the dots." Projections in both data growth and the need for more analytics are growing exponentially. According to *Forbes*, mankind produced more data in the past few years compared to the total amount of data created prior to 2013, and the volumes are accelerating.[3] The mass production of smart phones is certainly contributing to the explosions growth in big data. By 2020, over six billion smart phones will be in use

[1] Rick Weiss and Lisa-Joy Zgorski, *Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments* (Washington, DC: Office of Science and Technology Policy, Executive Office of the President, 2012).

[2] Catherine Dale, *The 2014 Quadrennial Defense Review (QDR) and Defense Strategy: Issues for Congress* (Washington, DC: Federation of American Scientists, 2014), 38, http://search.proquest.com/docview/1641843659/51044E2789BD4713PQ/1.

[3] Bernard Marr, "Big Data: 20 Mind-Boggling Facts Everyone Must Read," Forbes, September 30, 2015, http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/.

worldwide.[4] This number of phones will have a direct impact on Homeland Security Investigators, as investigations often revolve around criminal communications.

Big data analytics as a research topic is relatively new and several industries, including and government agencies are interested in it.[5] Although big data analytics presents value and opportunity for the public sector, academic literature is lacking in researching big data analytics in practice for public entities.[6] This research should contribute to academic research by demonstrating how big data analytics can enable investigators to target criminal networks more effectively and efficiently. Additionally, the research results may provide a model of measuring the performance of analytics when used to support federal law enforcement investigations, as well as the limits of the technology. Understanding these results should provide useful considerations for developing, integrating, and applying future analytics against other criminal programmatic areas.

This research examines how applying big data analytics to disparate data related to human smuggling activities can result in greater insights, as well as the discovery of unknown trends, subjects, and the extensiveness of human smuggling networks. The analytics may also enable HSI to prioritize key nodes within the human smuggling networks efficiently, while concurrently producing the necessary *probable cause* necessary to support federal investigations.

A. RESEARCH QUESTION

How can big data analytics improve the effectiveness and efficiency of HSI targeting human smuggling networks?

---

[4] Marr.

[5] Jonathan Seddon and Wendy L. Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," *Journal of Business Research* 70, no. C (2017): 301.

[6] Pandula Gamage, "New Development: Leveraging 'Big Data' Analytics in the Public Sector," *Public Money & Management* 36, no. 5 (2016): 385.

B. RESEARCH DESIGN

The purpose of this research is to test an experimental data analytics application to examine the efficiency and effectiveness of data analytics supporting criminal investigations. Big data analytics currently is not being widely studied or tested within HSI.

The data analytics application examined is called Citrus. The application is a free-for-government use software tool designed by Sandia National Laboratories. The Department of Homeland Security Science and Technology provided this application to HSI for testing and evaluation. Citrus was built to discover, trend, and link disparate data and is being used as an analytics application developed and implemented by HSI Phoenix.

1. Application Background

Citrus is a large library of algorithms with a flexible means of ingesting and conditioning data that can be tested against different problems and datasets. Conceptually, the search capability is designed to enhance collection, targeting, and trending processes by greatly reducing man-hours spent on data management and refocusing resources on analysis. Citrus also compliments other analytical systems being employed by HSI, such as Falcon developed by Palantir, which allows HSI to analyze investigation information related to criminal activities manually.[7] The outputs of Citrus can be transferred to other HSI systems. For HSI systems lacking analytics, discoveries made with Citrus can then be migrated over to those existing systems for further analysis with datasets not resident with Citrus.

Citrus is an indexing software with a suite of applications. Indexing is a data structure that increases efficiencies in searching and retrieving data.[8] Data is extracted to create an index that is then searchable using various functions. The typical challenge for HSI analysts is that they are required to correlate and analyze information from different

[7] "Privacy Impact Assessment for the FALCON Search & Analysis System," Google, 1, accessed October 12, 2017, https://www.google.com/search.

[8] Omar El Gabry, "Database—Indexing, Transactions & Stored Procedures (Part 9)," *Medium* (blog), September 15, 2016, 1, https://medium.com/omarelgabrys-blog/database-indexing-and-transactions-part-9-a24781d429f8.

datasets. Citrus was developed to enable analysts to correlate and analyze large, disparate datasets, both structured and unstructured, quickly and efficiently.

Citrus has distinct capabilities. For example, advanced search (AS) allows users to search through a diverse set of documents and document types and summarize documents in various ways while discovering relationships among terms. Secondly, document set investigation (DSI) splits documents into categories through search and machine-learning algorithms, which allow users to review documents, identify documents of interest, and have potentially relevant documents automatically suggested. These algorithms may be able to help HSI identify criminal networks, relationships, and patterns of behavior. If proven successful, this discovery of new information will dramatically increase analysts' abilities to increase network analysis, while identifying new targets with enough evidence to articulate probable cause.

Citrus streamlines processes through automation. Citrus' existing scripting capability, which is a programming language that automates tasks, combined with customizations specific to HSI problem sets, provides flexible and rapid automation of data ingestion, link analysis, and relationship discovery.[9] The reduction in front end data management should allow analysts to spend more time on actual analysis, prioritize more efficiently, and process more casework effectively.

2. Application Requirements

Prior to implementing and testing Citrus, HSI Phoenix spent several months with the data scientists from Sandia National Laboratories. The purpose was to define the analytics requirements. In 2016, the University of Hawaii published the results of a big data architecture study. The purpose of the study was to determine how best to design big data systems supporting analytics.[10] The study focused on an outsourcing company that

[9] Margaret Rouse, "What Is Script?" Tech Target Network, 1, accessed October 25, 2017, http://whatis.techtarget.com/definition/script.

[10] Hong-Mei Chen, Rick Kazman, and Serge Haziyev, "Agile Big Data Analytics Development: An Architecture-Centric Approach," in *System Sciences (HICSS), 2016 49th Hawaii International Conference On* (Piscataway, NJ: IEEE, 2016), 5378, http://ieeexplore.ieee.org/abstract/document/7427853/.

had successfully provided big data analytic solutions to thousands of companies.[11] The primary lesson learned from the study revealed that software engineers and the end-users should be teamed early in the developmental process.[12] HSI applied this lesson in developing Citrus. The analytics supporting this study have been tailored specifically to the inputs from HSI Phoenix. This format is important when considering the results of the tests, as the time to develop the analytics should be taken into account when judging the efficiencies of the application.

Human smuggling is the criminal programmatic area selected for this research because it is a top priority for the DHS. Additionally, the information used to analyze human smuggling activities generally stems from three datasets: OBP reports, communication, and financial data. Scoping the research to a particular type of criminal activity with a manageable set of data should result in findings that are fundamental to understanding efficiencies and effectiveness of data analytics supporting criminal investigations.

A summative evaluation was conducted to measure two key objectives:

- Does applying Citrus increase the efficiencies in identifying subjects and networks associated with human smuggling?

- Does applying Citrus increase the effectiveness in identifying subjects and networks associated with human smuggling?

Given a set of search criteria, the research compares the results of queries being conducted manually to a same set of problems with Citrus that measures the total number of targets, network identification, whether enough documentation is obtained to justify problem cause, and the timelessness of the results.

Data results relating to efficiencies and the effectiveness of research being applied with the same process with Citrus and without is collected. Two types of tests were conducted, search and discovery.

---

[11] Chen, Kazman, and Haziyev, "Agile Big Data Analytics Development," 5381.

[12] Chen, Kazman, and Haziyev, 5382.

- The search query consists of extrapolating the highest volume of phone numbers from 45, 90, and 180 days of Arizona border patrol station reports and correlating those results to other indexes relating to financial and communication data.

- The discovery query tests two reports created with Citrus. The quality of evidence (QE) report measures levels of probable cause against known and unknown entities from within a large dataset. The CO report is designed to determine the total number of phone number pairings from thousands of OBP reports, which leads to potential human smuggling network identification.

Efficiency is measured by comparing the time it takes to complete a search query and a discovery query manually and with the application in identifying subjects and human smuggling networks. Efficiencies in obtaining pertinent information related to human smuggling are also measured by the number of steps taken to obtain the same results with and without the application.

Effectiveness is measured by comparing the number of subjects and networks identified manually and with the application for the search and discovery queries. Determining whether probable cause can be obtained simultaneously while increasing the efficacy of human smuggling analysis is also a measure of effectiveness. Probable cause exists when reasonable suspicion related to human smuggling activities can be articulated. For this study, probable cause exists when specific phone numbers are recorded frequently in OBP reports linked to communications records with other phone numbers also associated with OBP reports, and combined with financial transactions used by the same phone numbers.

3. Limitations

The study is limited to one HSI office, against one type of transnational network, human smuggling, with limited datasets related to human smuggling. The measured data is further divided between 45-, 90-, and 180-day periods. Testing more broadly in terms of

data and analytics with more personnel likely may yield different results. Following a defined methodology for analyzing human smuggling networks, including ingesting the same type of border patrol reports, can control for research consistencies. If other personnel follow the same process, results can then be compared.

C. POSSIBLE BENEFITS OF RESEARCH FOR HOMELAND SECURITY

The intended outcome of the research is to examine the efficacy of big data analytics when applied to datasets specific to a particular criminal network. More broadly, this research may help HSI determine how best to develop data analytics supporting investigations and produce a model that can be replicated to assist in the investigations of different types of criminal networks. To that end, any collateral understandings relating to the development process is captured.

II. BIG DATA ANALYTICS OVERVIEW

Big data is becoming a popular term.[13] According Gandomi and Haider, the term big data probably emerged originally in the mid-1990s but it gained popularity in 2011.[14] Industries ranging from health care to national security are exploring big data application. Big data continues to present incremental solutions to numerous industries.[15] In that respect, the definition, value, associated characteristics, and data growth are important to understanding the application of big data analytics.

A. BIG DATA ANALYTICS DEFINED

Big data and big data analytics can be viewed separately, although the terms are used interchangeably and synonymously. Sivarajah et al. arrived at a similar conclusion after reviewing 227 articles related to big data analytics methods and tools associated with the technology.[16] Sivarajah et al. nuanced the terms by stating big data has the potential to assist in the decision-making process and big data analytics provides the means for extracting insights or value from the big data.[17] A study from the UK government has a similar view and fused the definition, "big data refers to both large volumes of data with high level of complexity and the analytical methods applied to them which require more advanced techniques and technologies in order to derive meaningful information and insights in real time."[18]

---

[13] Zhihan Lv et al., "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics," *IEEE Transactions on Industrial Informatics* 13, no. 4 (2017): 1891.

[14] Amir Gandomi and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management* 35, no. 2 (April 2015): 138, https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

[15] Lv et al., "Next-Generation Big Data Analytics," 1891.

[16] Uthayasankar Sivarajah et al., "Critical Analysis of Big Data Challenges and Analytical Methods," *Journal of Business Research* 70 (2017): 275.

[17] Uthayasankar Sivarajah et al., 265.

[18] Emerging Technologies Big Data Community of Interest, *HM Government Horizon Scanning Programme Emerging Technologies: Big Data* (London: HM Government, 2014), 2, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389095/Horizon_Scanning_-_Emerging_Technologies_Big_Data_report_1.pdf.

George, Haas, and Pentland and Gartner have consistent descriptions in that big data is the "massive amount of data collected" from many sources.[19] While George, Haas, and Pentland further define it by adding that big data stems from an "increasing plurality of sources," Snijders, Matzat, and Reips provides a consistent, more succinct definition and proposes that big data exists when data sizes exceed the processing capacities of common software tools.[20] Although authors consistently reference big, large, and increasing volumes when defining, practitioners of big data generally do not apply a particular file size when referencing big data. Big data is very often thought of in terms of how insights can be gained from larger volumes of data.[21] Gartner elaborates and states that big data also is "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation."[22] Henry and Venkatraman support that definition and generally define analytics as allowing organizations to study large datasets and respond according to requirements and operations.[23] Desouza merges the large dataset definition with the idea of meeting business requirements by stating that "big data is an evolving concept that refers to the growth of data and how it is used to optimize business processes, create customer value, and mitigate risks."[24]

Palem's definition parallels most others, "big data analytics enables an organizations to reliably collect and analyze large volumes of data," but further cataloged

---

[19] Gerard George, Martine R. Haas, and Alex Pentland, "Big Data and Management," *Academy of Management Journal* 57, no. 2 (2014): 321; "Big Data?" Gartner IT Glossary, May 25, 2012, http://www.gartner.com/it-glossary/big-data/.

[20] George, Haas, and Pentland, 321; Chris Snijders, Uwe Matzat, and Ulf-Dietrich Reips, "'Big Data': Big Gaps of Knowledge in the Field of Internet Science," *International Journal of Internet Science* 7, no. 1 (2012): 1.

[21] George, Haas, and Pentland, 321.

[22] Gartner IT Glossary, "Big Data."

[23] Regina Henry and Santosh Venkatraman, "Big Data Analytics the Next Big Learning Opportunity," *Journal of Management Information and Decision Sciences; Weaverville* 18, no. 2 (2015): 18.

[24] Kevin Desouza, *Realizing the Promise of Big Data, Implementing Big Data Projects* (Washington, DC: IBM Center for the Business of Government, 2014), 10, http://observgo.uquebec.ca/observgo/fichiers/26986_Realizing%20the%20Promise%20of%20Big%20Data.pdf.

big data as a "solution enabler" and "storage platform."[25] As a solution enabler, big data reduces the time it takes to solve a problem or question.[26] As a storage platform, big data provides the means accessing large swaths of data from a variety of sources.[27] In regards to a variety of sources, Yaqoob dissects the types of data within big data when defining the set of structured, unstructured, and semi-structured data accumulated from heterogeneous data sources.[28] The variety of big data sources is an important context for further defining big data analytics with respect to understanding the characteristics of big data.

B. BIG DATA VALUE

The value of big data is released when it is applied to make better decisions.[29] Big data has implications from the micro to macro levels. At the smallest scale, human behaviors of team interactions can be tracked and evaluated with personal sensors.[30] Big data analytics has strategic value, as it offers the capability to monitor disease outbreaks, trafficking patterns, or community attitudes of large populations from social media.[31] In healthcare, big data analytics has the opportunity to increase the quality of care while reducing costs.[32] For example, Columbia University used analytics to reduce the amount of time it takes to identify complications in brain injuries.[33] The opportunity for big data for researchers is to determine how big data can produce new value.[34]

---

[25] Gopalakrishna Palem, "Formulating an Executive Strategy for Big Data Analytics," *Technology Innovation Management Review* 4, no. 3 (March 2014): 25.

[26] Palem, 25.

[27] Palem, 25.

[28] Ibrar Yaqoob, "Information Fusion in Social Big Data: Foundations, State-of-the-Art, Applications, Challenges, and Future Research Directions," *International Journal of Information Management*, April 19, 2016, 2–4.

[29] Gandomi and Haider, "Beyond the Hype," 140.

[30] George, Haas, and Pentland, "Big Data and Management," 325.

[31] George, Haas, and Pentland, 325.

[32] Kent Bottles, Edmon Begoli, and Brian Worley, "Understanding the Pros and Cons of Big Data Analytics," *Physician Executive* 40, no. 4 (2014): 12.

[33] IBM, *Data-Driven Healthcare Organizations Use Big Data Analytics for Big Gains* (Somers, NY: IBM Corporation, 2013), 2–8, http://www-03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf.

[34] George, Haas, and Pentland, "Big Data and Management," 324.

The strongest evidence of value from big data analytics is in the demand for data analytics skillsets. The Department of Labor predicts that four million jobs in data analytics will exist by 2018, including an estimated gap between 140,000 and 190,000 skilled personnel.[35] The collection of data from people, as well as the exponential increase in data generated from machines connected to the internet, is resulting in an inconceivable repository of data.[36] The value of big data, however, will only be realized if the data is analyzed to produce greater insights into problems, reveal informative patterns, and allow better decisions.[37] Henry and Venkatraman suggest that with the rapid growth in data, organizations need to have a commensurate increase in analytics technology and skillsets to meet the big data growth.[38] Although this article was based on a business construct, Henry and Venkatraman suggest that managers today have opportunities to make better, more objective decisions by applying analytics to their respective sets of big data.[39]

C. BIG DATA CHARACTERISTICS

One of the primary characteristics of big datasets relates to *structured* versus *unstructured* data. Traditional data is structured.[40] Meaning, the data collected is recorded consistently, generally in rows or columns in a common spreadsheet. Conversely, big data is normally unstructured. Henry and Venkatraman state that IBM estimates that 80% of data produced is unstructured, largely in the form of emails, texts, and video.[41] Gandomi and Haider claim that the percentage is higher and asserts that 95% of big data is unstructured but criticized that most studies revolve around structured data.[42]

Data originates from so many various sources that generate numerous forms of data. The inconsistent data forms, whether created by humans or machines, produces the

---

[35] Henry and Venkatraman, "Big Data Analytics the Next Big Learning Opportunity," 17–23.

[36] Henry and Venkatraman, 17.

[37] Henry and Venkatraman, 17, 20.

[38] Henry and Venkatraman, 20.

[39] Henry and Venkatraman, 21.

[40] Henry and Venkatraman, 19.

[41] George, Haas, and Pentland, "Big Data and Management," 25.

[42] Gandomi and Haider, "Beyond the Hype," 137.

unstructured formats, such as word documents, PDF files, text messages, etc., as noted by Gandomi and Haider. Finding and understanding relationships between large datasets is key for homeland security, as much of the data collected is unstructured. This relationship is central to this research because unstructured data is growing at twice the rate of conventional structured data, and unstructured data lacks the structure to be analyzed by computers often in the form text, images, audio, and video.[43]

The "Vs" of big data are commonly referenced as the primary characteristics. According to Seddon and Currier, in 2001, Laney introduced volume, velocity, and variety as the original three V's in describing big data.[44] In 2013, research by Van Rijmenam introduced veracity, variability, visualization, and value, bringing seven V's into the descriptive lexicon of big data.[45] Desouza explains the importance as the characteristics showcase the big data challenges with developing a supportable analytics system.[46] The Vs, however, should not be considered equal, as emphasized by Desouza.[47]

Volume is the means of transferring and storing data efficiently.[48] Kitchin, however, uses a specific measure of volume and references as petabytes and terabytes.[49] Although Kitchin believes volume can be measured, Gandomi and Haider referenced volume as the "magnitude of data" but injected that big data is relative and defining as a particular file size is impractical.[50] As storage and processing capacities increase, the actual size of what is considered *big* varies.[51] Further, different file types may require different types of data

[43] George, Haas, and Pentland, "Big Data and Management," 25; Gandomi and Haider, 138.

[44] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302.

[45] Mark Van Rijmenam, "Why the 3v's Are Not Sufficient to Describe Big Data," Big Data Startup, 1, 2013, http://www. bigdata-startups.com/3vs-sufficient-describe-big-data.

[46] Desouza, *Realizing the Promise of Big Data*, 11.

[47] Desouza, 11.

[48] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302.

[49] Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* 1, no. 1 (2014): 1.

[50] Gandomi and Haider, "Beyond the Hype," 138.

[51] Gandomi and Haider," 138.

management applications, e.g., Excel or video. Therefore, the type and size of the data matters when defining "big."[52]

Velocity describes now fast data is collected and processed in real-time, as inferred by Kitchen.[53] In addition to speed and process, Gandomi and Haider added the rate of analysis, which they underscore as particularly important for businesses requiring analysis of real-time data where information is considered *perishable*, which supports Kitchin's description.[54]

Variety refers to the different types of structured datasets, according to Seddon and Currie, but that description is in the minority.[55] Most reviews reference variety as both structured and unstructured forms.[56] Gandomi and Haider argue succinctly that variety refers to the "structural heterogeneity" of a dataset.[57] Al Nuaimi et al. also support this view and refer to variety as the different types of data generated, with the understanding that most of the data is unstructured.[58] Henry and Venkatraman came to the same conclusion that most of the data is unstructured. Desouza contended that variety can be the most challenging *V*, as organizations struggle with integrating heterogeneous datasets with new types of data formats, particularly with legacy systems.[59]

Veracity refers to the accuracy of how the data is analyzed.[60] Gandomi and Haider took the opposite view and described veracity as the unreliability of data, but the meaning

---

[52] Gandomi and Haider," 138.

[53] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302; Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 1.

[54] Gandomi and Haider, "Beyond the Hype," 138.

[55] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302.

[56] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 1; Desouza, *Realizing the Promise of Big Data*, 11.

[57] Gandomi and Haider, "Beyond the Hype," 138.

[58] Eiman Al Nuaimi et al., "Applications of Big Data to Smart Cities," *Journal of Internet Services and Applications* 6, no. 1 (August 2015): 4, https://doi.org/10.1186/s13174-015-0041-5.

[59] Desouza, *Realizing the Promise of Big Data*, 12.

[60] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302.

is consistent with other authors.[61] This meaning is important in the context of social media as verifying the truthfulness, or intent, of postings is difficult to discern.[62]

Variability is the means of measuring the change in data flows.[63] In other words, variability refers to the flow rates of data from the origins and multitudes of data sources.[64] This flow rate creates challenges in processing data, as the data must be cleaned, structured, and merged from multiple different sources in applying analytics.[65]

Visualization is important in understanding big data, as it is descriptive for displaying models, trends, and patterns, which is common practice in data analytics.[66]

Value describes how an organization profits from a data analytics strategy.[67] Value also refers to the advantages big data offers an organization, and becomes more important when increased and merged with larger datasets.[68]

D. DATA GROWTH

Projections in big data growth and the need for more analytics are flourishing. According to *Forbes,* mankind produced more data in the past few years compared to the total amount of data created prior to 2013, and the volumes are accelerating.[69] The mass production of smart phones is certainly contributing to the explosions growth in big data. By 2020, over six billion smart phones will be in use worldwide.[70] This large number of smart phones will have a direct impact on Homeland Security Investigators, as investigations often revolve around criminal communications.

---

[61] Gandomi and Haider, "Beyond the Hype," 139.

[62] Gandomi and Haider, 139.

[63] Desouza, *Realizing the Promise of Big Data*, 11.

[64] Gandomi and Haider, "Beyond the Hype," 139.

[65] Gandomi and Haider, 139.

[66] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 302.

[67] Seddon and Currie, 302.

[68] Gandomi and Haider, "Beyond the Hype," 139; Al Nuaimi et al., "Applications of Big Data to Smart Cities," 4.

[69] Marr, "Big Data," 1.

[70] Marr, 1.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LITERATURE REVIEW

This literature review underpins the thesis topic of evaluating big data analytics supporting HSI) targeting criminal networks. The volume of data within HSI is growing, which creates bigger challenges in processing large datasets related to investigations. Big data analytics offers investigators tremendous benefits but implementation challenges incorporating it nationwide are likely. To that end, this literature review is categorized according to the theoretical frameworks for development, public versus private sector application, risks, and challenges related to implementing big data analytics.

A. BIG DATA ANALYTICS THEORETICAL FRAMEWORKS

Big data analytics as a research topic is new and has the interest of several industries, including the government.[71] Although big data analytics presents *value* and opportunity for the public sector, academic literature supporting big data analytics in practice for public entities is lacking, according to Gammage.[72]

George, Haas, and Pentland suggest that big data will change social and economic policies, as well as research methods over the next decade.[73] Rubin et al. also assert that big data analytics is relatively new and the convergence of data from fields, such as science and economics, has the potential to transform societies.[74] Kitchin acknowledges that big data analytics is popular within business but the sciences are also beginning to recognize the potential.[75] Where business is leveraging data analytics to increase revenue and market share, science can apply to increase worldly understandings.[76] Desouza noted that big data is abundant within the physical sciences but data within the social sciences is limited.[77]

---

[71] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 301.

[72] Gamage, "New Development," 385.

[73] George, Haas, and Pentland, "Big Data and Management," 324.

[74] David Rubin et al., "Harnessing Data for National Security," *The SAIS Review of International Affairs* 34, no. 1 (2014): 122.

[75] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 3.

[76] Kitchin, 3.

[77] Desouza, *Realizing the Promise of Big Data,* 12.

Desouza offered, for example, the case for human trafficking. The issue is global but the associated data is incomplete, disparate, and unstructured.[78] This inconsistency makes a complex problem more difficult in terms of assembling big data where larger scale analytics can be applied.[79]

The potential in leveraging big data analytics exists but supporting the application is lacking theory. Chen, Li, and Wang found that big data analytics is progressing technically, but a theoretical framework is required for design and integration.[80] Seddon and Currie concurred in that big data analytics is "under-theorized" and "empirically underrepresented in business research."[81] Furthermore, Sivarajah et al. contend that the research domain is nascent despite the recent increase in research.[82] Sivarajah et al. also found the vast majority of related research articles were analytical and more research using case studies in different organizations is needed.[83]

Kitchin argues that big data analytics in revolutionizing research principally because of the unprecedented amount of data available.[84] With the exponential and rapid growth in data, Kitchin promotes the idea that big data has vast implications in how knowledge is produced.[85] Kitchin asserts that hypotheses become less relevant because so many different patterns and relationships can be identified from within the massive datasets.[86] As such, big data presents the potential for establishing a new research model supporting multiple science disciplines.[87] Kitchin endorses the idea of data-driven science,

---

[78] Desouza, 12.

[79] Desouza, 12.

[80] Kun Chen, Xin Li, and Huaiqing Wang, "On the Model Design of Integrated Intelligent Big Data Analytics Systems," *Industrial Management & Data Systems* 115, no. 9 (2015): 1678.

[81] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 301.

[82] Sivarajah et al., "Critical Analysis of Big Data Challenges and Analytical Methods," 277.

[83] Sivarajah et al., 278.

[84] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 1.

[85] Kitchin, 2.

[86] Kitchin, 4.

[87] Kitchin, 3.

which creates a newfound extension of the scientific method.[88] Kitchin supports his assertion by quoting Chris Anderson, former chief editor of *Wired* magazine, by claiming big data analytics can produce knowledge so vast that it can be "the end of theory." Kitchin argues that with so much data and the means to make unexplored discoveries, insights can be produced "free of theory."[89] To produce a new, data-driven research method, a new theoretical framework must be created.[90]

George, Haas, and Pentland also contend that big data changes research.[91] The normal means of measuring correlation is less effective due to large volumes of data. In other words, more data will inevitably lead to some correlations that potentially can result in false correlations.[92] A move from focusing on means and averages to outliers is another consideration. With such large datasets, outliers become more significant as they may represent larger populations even though the data is relatively smaller.[93] The positive side of research as it relates to big data is the vast amounts of "multifaceted richness."[94] With so much data available of deep analysis, George, Haas, and Pentland argue that big data will allow researchers to move beyond correlations and causality to *consilience*. Big data introduces opportunities to merge a voluminous amount of data from a plethora of independent sources that can allow research to make greater conclusions.[95]

Although the research and theory for big data analytics is under development, several frameworks have been explored. Seddon and Currie used a model based on the seven V's characterizing big data. The authors conducted interviews of traders to determine not only how big data analytics was being applied competitively, but also how each "V" was being leveraged to maximize the technology. For example, billions of transactions are

---

[88] Kitchin, 3.

[89] Kitchin, 3.

[90] Kitchin, 10.

[91] George, Haas, and Pentland, "Big Data and Management," 323.

[92] George, Haas, and Pentland, 323.

[93] George, Haas, and Pentland, 323.

[94] George, Haas, and Pentland, 323.

[95] George, Haas, and Pentland, 324.

conducted by the minute in high frequency trading (HFT). Traders used speed (volume) to maintain a competitive edge. Additionally, searching through streaming social media feeds of unstructured data (variability) was conducted to assist with real-time forecasting. The study concluded that big data analytics technology has greatly changed financial trading. By segmenting the review by the seven V's of big data, the study provides insights into precisely how the technology is employed and the advantages that can be obtained within a particular industry.[96]

Chen, Li, and Wang argue that big data analytics, although proving to be beneficial in business, requires a design model to integrate analytics capable of meeting organizational demand.[97] In testing for a design model, Chen, Li, and Wang applied the collective intelligence (CI) model proposed by Schut. According to Schut, CI systems are complex, adaptive, can organize autonomously, and have emergent behaviors.[98] Schut argues the CI systems are widespread in multiple sciences, including computer science, primarily because designing the application requirements phase of a complex system are so important.[99] Schut proposes that the CI framework has two models, general and specific. The general model is used for validation and the specific model is used for architecture implementation.[100]

Chen, Li, and Wang applied Schut's CI model to test the integration of big data analytics in an ecommerce company. Chen, Li, and Wang concluded that the CI framework was suitable for designing and implementing big data analytics applications.[101] More specifically, the authors contend that both a generic and specific models are necessary for

---

[96] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 305–306.

[97] Chen, Li, and Wang, "On the Model Design of Integrated Intelligent Big Data Analytics Systems," 1666.

[98] Martijn C. Schut, "On Model Design for Simulation of Collective Intelligence," *Information Sciences* 180, no. 1 (2010): 132.

[99] Schut, 132.

[100] Schut, 137–39.

[101] Chen, Li, and Wang, "On the Model Design of Integrated Intelligent Big Data Analytics Systems," 1678.

systems design.[102] From a managerial perspective, the authors suggest that big data analytics applications should be packaged as components supporting a specific process that are integrated incrementally and tested in different situations.[103]

Al Nuaimi et al. provided an example framework for integrating big data analytics for governments in developing a "smart city." Al Nuaimi et al. defined a smart city "as an integrated living solution that links many life aspects such as power, transportation, and buildings in a smart and efficient manner to improve the quality of life for the citizens of such city."[104] The overarching goal of the smart city concept is to improve resource efficiencies, and big data analytics presents opportunity to support that main objective.[105] The smart city concept is similar to other findings in that public sectors are looking to improvement efficiencies areas, such as transportation and utilities management.[106] For example, Al Nuaimi et al. references a system called accelerated-time simulation for traffic flow (ATISMART). The purpose of the system is to manage traffic flows efficiently through a network of sensors to provide real-time data of traffic conditions.[107] Al Nuaimi et al. also found that big data characteristics were important in evaluating big data analytics, especially in regards to key characteristics of velocity, volume, and variety.[108]

George, Haas, and Pentland suggest that with the abundance of data flooding our society, academics will have opportunities to study and understand processes and behaviors at multiple levels.[109] Organizations can move beyond quarterly updates to daily or even live updates and trends affecting operations. This philosophy was presented in the context of business applications and not through the lens of law enforcement but the principle has

---

[102] Chen, Li, and Wang, 1678.

[103] Chen, Kun, Li, Xin, and Wang, Huaiqing, "On the Model Design of Integrated Intelligent Big Data Analytics Systems," *Emerald Insight* 115, no. 9 (2015): 1678.

[104] Al Nuaimi et al., "Applications of Big Data to Smart Cities," 2.

[105] Al Nuaimi et al., 14.

[106] Al Nuaimi et al., 5.

[107] Al Nuaimi et al., 7.

[108] Al Nuaimi et al., 7.

[109] George, Haas, and Pentland, "Big Data and Management," 325.

value.[110] Although big data is a common term now with business, the authors assert that more testing and fielding should be done due to a shortage of "published management scholarship" that explores how best to apply big data analytics.[111]

B. PRIVATE VERSUS PUBLIC SECTOR USAGE

The literature surrounding big data analytics centers on business applications in the private sector. Seddon and Currie support that the most popular companies known for leveraging big data are Amazon, Google, and Facebook where analytics are applied to analyze high volumes of data for their respective clients.[112] These aforesaid companies are leaders for big data analytics, but the financial and banking industries also have decades of experiences in managing large volumes of data.[113]

Numerous examples of the private sector using big data analytics to improve business operations can be found. For example, General Motors invested in telematics, which is an analytics system design to assist stranded drivers and to provide vehicle security and maintenance diagnostic services.[114] Additionally, Henry and Venkatraman provided an example regarding a vehicle sales website, Edmunds.com. According to the article, Edmunds.com receives "50,000 events per minute and produces 60 to 70 gigabytes of data per day." Analyzing so much real-time data is humanly impossible. Edmunds.com was able to apply analytics software to monitor security breaches, as well as consumer behaviors hourly and subsequently provide decision-makers with real-time, prescribed notifications.[115]

---

[110] George, Haas, and Pentland, 325.

[111] George, Haas, and Pentland, 321.

[112] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 301.

[113] Seddon and Currie, 301.

[114] "OnStar: GM's Not-So-Secret Weapon," Bloomberg, 2, May 30, 2013, https://www.bloomberg.com/news/articles/2013-05-30/onstar-gms-not-so-secret-weapon.

[115] Henry and Venkatraman, "Big Data Analytics the Next Big Learning Opportunity," 22.

One of the primary differences between private and public sector big data is access.[116] Users of third-party applications, for example, voluntarily submit their personal identifiable information to private companies.[117] This submission of data allows mass collection, which increases the developer's ability to have streaming, live data readily available.[118] Desouza presented the example of an application created by Nike that tracks a user's lifestyle. Nike then in turn collaborated with Sprout, which allows users to provide their lifestyle data for employer-sponsored fitness programs. Moreover, private companies that collect vast amount of data on their customers also sell data. Amazon.com, for instance, sells buying habits of customers to other companies to promote products and drive sales.[119] This type of ubiquitous data voluntarily offered and sold distinguishes private sector from public sector data.[120]

For the public sector, accessing such large datasets from external organizations is challenging.[121] Governments that obtain large amounts of public data are negatively perceived.[122] Publicly collected data is often viewed as intrusive and the perceptions are that the data will benefit a distrustful government.[123] A case in point was when a group of researchers obtained and published the addresses of gun owners living in New York counties through a Freedom of Information Act request that occurred immediately following the tragic school shooting at the Sandy Hook Elementary school. The intent was to inform the public of the pervasiveness of gun ownership but the fallout was putting those legal gun owners at risk by exposing their locations to criminals wanting to target the owners.[124]

---

[116] Desouza, *Realizing the Promise of Big Data*, 13.

[117] Desouza, 13; Rubin et al., "Harnessing Data for National Security," 123.

[118] Desouza, 13.

[119] Desouza, 14.

[120] Desouza, 13.

[121] Desouza, 14.

[122] Desouza, 14.

[123] Desouza, 14.

[124] Desouza, 15.

According to Desouza, the private sector also has made more investments in big data compared to the public sector, although some infamous attempts made by the federal government failed: the troubled Affordable Care Act site, Healthcare.Gov, and the Federal Bureau of Investigation case management systems, Virtual Case File, both of which fell flat initially and cost the federal government millions.[125]

Gamage asserts that although the private sector is gaining competitive advantages with big data analytics, opportunities do exist for the public sector in leveraging big data analytics.[126] For the public sector, implementing big data analytics is a new concept but presents similar opportunities as the private sector to streamline processes in the management of operations, personnel, and other vital resources.[127] Rubin et al. argue that the government has made improvements in identifying the silos of data but more needs to be done to allow progress by federal agencies in the development of big data analytics.[128] This argument is consistent with Desouza's study after canvassing chief information officers (CIOs) from multiple levels of government. Desouza found that the CIOs recognize that data reservoirs are invaluable to effective decision-making to make operations more efficient.[129]

Furthermore, Gamage reviewed the usage of big data analytics in the public sector from 16 different global governments.[130] The findings demonstrated that big data analytics applications are being used in the public sector to improve efficiencies, particularly relating to transportation, human resources, and agricultural. Examples of applications in government healthcare programs are also available. For example, Qatar is using data analytics to help predict the spread of diseases by aggregating and analyzing health records. In the United Kingdom, infection control data was exchanged between hospitals and big data analytics was applied. The results substantially reduced infection rates that saved the

---

[125] Desouza, 15.

[126] Gamage, "New Development," 385.

[127] Desouza, *Realizing the Promise of Big Data*, 6, 10.

[128] Rubin et al., "Harnessing Data for National Security," 122.

[129] Desouza, *Realizing the Promise of Big Data*, 7.

[130] Gamage, "New Development," 387–88.

hospitals millions.[131] Of the 16 countries reviewed by Gamage, only one country was noted to be using big data analytics in a law enforcement capacity, the United States. Gamage noted that the Internal Revenue Service was using data analytics to detect fraudulent tax returns.[132]

Although the examples of big data analytics in law enforcement are limited, the technology has been proven to help "isolate the needle in the haystack," particularly when applied to fraud and cybercrimes.[133] Desouza also provided examples of big data analytics being leveraged in a federal law enforcement capacity. In 2006, the United States Postal Service began leveraging big data analytics to save money and improve efficiencies by detecting fraud in suspicious packages. The analytics involved are able to scan and review billions of records in milliseconds.[134] Moreover, the Department of State is using analytics to detect potential areas of instability and violence through large datasets collected from social media and other sensors.[135]

From a city law-enforcement perspective, New York City is employing the Business Integrity Commission (BIC), which was designed to detect and identify suspects illegally collecting biodegradable waste from local restaurants. Using analytics, BIC crossed referenced grease production from the Department of Health and Mental Hygiene permit data with the Department of Environmental Protection sewer backup data. The results produced hotspots where unlicensed activity was occurring.[136] Although law

---

131 Gamage, 388.

132 Gamage, 388.

133 Rubin et al., "Harnessing Data for National Security," 123.

134 Desouza, *Realizing the Promise of Big Data*, 18.

135 Desouza, 18–19.

136 NYC Business Integrity Commission and NYC Environmental Protection, *New York City Business Integrity Commission, Department of Environmental Protection, and Mayor's Office of Policy and Strategic Planning Launch Comprehensive Strategy to Help Businesses Comply within Grease Disposal Regulations*, no. 71 (New York City, NYC Business Integrity Commission and NYC Environmental Protection, 2012), 1–3, http://www.nyc.gov/html/bic/downloads/pdf/pr/nyc_bic_dep_mayoroff_policy_10_18_12.pdf.

enforcement is just beginning to employ data analytics, Rubin et al. suggests mission success is in jeopardy if big data analytics is not adopted.[137]

C. BIG DATA ANALYTICS RISKS

Although big data analytics offers numerous benefits and value, it is fraught with risks, according to one big data expert, Alexei Efros.[138] Big data analytics should be thought of in terms of what is occurring and not why something is occurring.[139] Take for example this hypothetical scenario. With big data analytics, an analyst could sift through hundreds of border patrol reports and identify common human smuggling networks responsible for moving hundreds of aliens through a particular corridor along the southwest border. The analytics, however, would not necessarily be able to reveal why those hundreds of aliens decided to select that particular corridor to transit. Moving beyond the "what" to the "why" can lead to more effective solutions.[140] If an analyst determined that corrupt officials were complicit in the smuggling of the aliens, a different enforcement action could be applied instead of simply adding more border patrol agents to a specific area with an increase in apprehensions. The analytics saves time and allows efficient correlations, yet the human interaction is required to interpret and answer deeper questions.[141]

1. Ethical Risks

Ethical risks further underscore the importance of answering the "why." Btihaj Ajana studied big data analytics regarding the ethical concerns of immigration. Btihaj reveals that governments are increasingly applying big data analytics technologies to regulate the flows of immigration.[142] Btihaj's worked centered on the Australian Department of Immigration and Citizenship's big data analytics system developed by

---

[137] Rubin et al., "Harnessing Data for National Security," 127.

[138] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 9.

[139] Bottles, Begoli, and Worley, 9.

[140] Bottles, Begoli, and Worley, 9.

[141] Bottles, Begoli, and Worley, 10.

[142] Btihaj Ajana, "Augmented Borders: Big Data and the Ethics of Immigration Control," *Journal of Information, Communication & Ethics in Society* 13, no. 1 (2015): 58.

IBM.[143] The purpose of the border risk identification system (BRIS) is to identify the "risky travelers" and improve border security.[144] Under BRIS, data was collected from various sources from the airlines, historical crossings, and passenger relationships. The data was then complied and algorithms were built to provide profiles on the "riskiest" travelers.[145] Btihaj argues that his type of algorithmic big data analytics is discriminatory and exclusionary to immigrants.[146] Big data analytics is dangerous to immigration because governments can unfairly target undocumented immigrants and prevent migrant population flow through an unchallenged algorithm.[147] He further states that big data analytics for "border management acquires its legitimacy by constructing a divide between the 'belonging citizens' and 'risky others,' attaching itself to things that are valued by the public, such as security and the welfare system, and staging a need for their protection and securitization."[148] Essentially, Btihaj presents that big data analytics can be used as means of digital discrimination and prevents movements of people based on a preconceived risky pattern.

Another risk of big data analytics presented by Btihaj was the concept of apophenia, which essentially means that with big data analytics, people will find patterns even if they do not exist.[149] This concept is tantamount to looking at clouds in the sky and seeing images. This type of pattern analysis can be hazardous, and this concept of pattern analysis through big data analytics in the identification high-risk persons based on analytic profiles is supported by homeland security. Alan Bersin, Assistant Secretary for International Affairs and Chief Diplomatic Officer for the DHS Office of Policy, supported the concept of leveraging analytics to identifying the highest priority threats as the "needles in haystack." By leveraging big data analytics, according to Bersin, the "haystack" becomes

[143] Ajana, 59, 62.

[144] Ajana, 59.

[145] Ajana, 62.

[146] Ajana, 66.

[147] Ajana, 67.

[148] Ajana, 67.

[149] Ajana, 60.

smaller, separating the low-risk from the high-risk people or goods.[150] The caution is what patterns are correctly identified, and the understanding of how those patterns were identified.

2. Data Integrity Risks

Foundational to the value of analytics is the data. Big data does result in more information, which includes false information.[151] Assuming the data itself is without biased and objective is inherently wrong. Data is not randomly assembled and is structured according to human designs.[152] Those hidden biases during collection and analysis present risks that must be recognized to avoid consequences.[153] The point is big data analytics is not flawless. The technology will produce new discoveries and offer greater insights, but the human expertise must be applied when interpreting the results to avoid drawbacks.[154]

Waterman and Bruening assert that uncovering new discoveries is one of the primary values of big data analytics, but compiled data with underlining data integrity problems creates inherent risks.[155] Flawed data can lead to faulty discoveries, which can result in consequential decisions. This problem is common when applying the new big data analytics technologies to old government data systems. The way in which the data is stored differs by each system or application that has been introduced over time.[156] The problem is multiplied when data is merged from different government agencies or private sector organizations. For example, dates of birth are structured in different formats. Merged data that have dates of birth formats that begin with a month integrated with a different dataset where the dates of birth start with the day results in inaccurate identifications. Waterman and Bruening support that big data analytics is able to examine multiple different types of

---

[150] Ajana, 66.

[151] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 10.

[152] Bottles, Begoli, and Worley, 10.

[153] Bottles, Begoli, and Worley, 9.

[154] Bottles, Begoli, and Worley, 10.

[155] K. Krasnow Waterman and Paula J. Bruening, "Big Data Analytics: Risks and Responsibilities," *International Data Privacy Law* 4, no. 2 (2014): 90, https://doi.org/10.1093/idpl/ipu002.

[156] Waterman and Bruening, 90.

data from numerous sources, but inconsistencies in the structure of merged data can produce misleading results.[157]

3. Data Processing

Data processing also creates inherent risks. An important aspect of mitigating risks in data integrity is the process of "cleansing" data. Data that has inaccuracies, incompleteness, or possesses duplicates is considered "dirty" and must be scrubbed to correct.[158] Tools are available that assist in cleaning data but catching mistakes is complicated. One processing error example presented by Waterman and Bruening in their study addressing risks pertaining to big data analytics related to the Enron financial scandal. The initial emails released by the Federal Energy Regulatory Commission totaled over one million. After the emails were scrubbed, removing duplicates and blanks, the total number of emails was reduced to fewer than 600,000 and the total number of users was reduced by 7.5%.[159] These types of discrepancies could skew the larger set of analyses and possibly result in misguided decisions.[160] The integrity of the data and how the data is processed create risks, as the analytical models depend on uncompromised data.[161]

Big data analytics has the potential to enhance Homeland Security Investigations, but the technology does present risks. Merging high volumes of data from disparate datasets has potential risks. To maximize the potential of big data analytics and avoid pitfalls, oversight of data integrity and processing is necessary. Most importantly, verifying the results of findings produced through big data analytics is imperative to protecting individuals, particularly related to potential algorithmic discrimination, and the reputation of the supported organization.[162]

---

[157] Waterman and Bruening, 92.

[158] Margaret Rouse, "What Is Data Scrubbing (Data Cleansing)?" SearchDataManagement, 1, accessed October 28, 2016, http://searchdatamanagement.techtarget.com/definition/data-scrubbing.

[159] Waterman and Bruening, "Big Data Analytics," 92.

[160] Waterman and Bruening, 92.

[161] Waterman and Bruening, 94.

[162] Waterman and Bruening, 95.

D. CHALLENGES

Integrating and applying big data analytics relating to privacy, skillset shortages, data compatibility, and costs present consistent challenges that differ from risks in that these relate to implementation versus potential inherent risks with the data analytics. This delineation is consistent with Schut's CI framework of general and specific models for developing analytics, as different considerations must be made with respect to validating analytics and integrating the technology.[163] These issues are explored further as potential implementation challenges.

1. Privacy

Privacy is a prevalent topic in big data analytics studies. Privacy is referenced repeatedly in literature related to big data analytics, particularly involving personal identifiable information relating to health records, location patterns, and buying habits.[164] According to McNeely and Hahm, "the surveillance potential of big data has raised concerns about invasions of privacy and stark limitations on personal freedom, such that privacy is one of the most prominent problems identified on the big data policy agenda."[165] As the volume of data grows, the government's ability to know more about citizens is alarming, which can potentially lead to discrimination.[166] Aggregated dataset from multiple sources allows large-scale profiling against selected demographics.[167]

George, Haas, and Pentland promote data sharing agreements as a means to mitigate privacy concerns.[168] The challenge is considering the balance between maximizing the value of big data analytics, which is, by definition, an aggregation of data, while protecting privacy.[169] The smart city concept supports the notion that information

---

[163] Schut, "On Model Design for Simulation of Collective Intelligence," 137–39.

[164] Gamage, "New Development," 388.

[165] Connie L. McNeely and Jong-on Hahm, "The Big (Data) Bang: Policy, Prospects, and Challenges," *Review of Policy Research* 31, no. 4 (2014): 308.

[166] Desouza, *Realizing the Promise of Big Data*, 15.

[167] Desouza, 15.

[168] George, Haas, and Pentland, "Big Data and Management," 323.

[169] George, Haas, and Pentland, 323.

sharing can cause friction. For example, investigative casework must be protected; however, sharing the related case data may result in additional connections in further a particular case. Big data analytics works best when the data is widely accessible; however, government agencies have rules that often prevent data merges. With data aggregation, more information is exposed, potentially creating risks to privacy.[170] The core issue is big data presents problems if a means to contest the origin of the data is not available.[171] Even if the data is obfuscated or masked, identifiable information can be obtained. For example, Bottles, Begoli, and Worley noted that AOL released the searched queries of over 600,000 people. Even though the identities of the people were removed, *The New York Times* was able to comb through the queries and identified specific users by the profiled searches.[172]

2. Skillset Shortage

Desouza found that CIOs working in governmental big data analytics projects are struggling to recruit personnel with the requisite skillsets.[173] Large corporations, such as Taco Bell, General Electric, Boeing, and Walt Disney, are seeking employees with data analytics skills to garner greater insights from big data.[174] According to the McKinsey Global Institute report, 1.5 million managers in big data will be required to advance the technology.[175] Additionally, Henry and Venkatraman contend that undergrad business programs are lacking courses that focus on data analytics.[176] This deficiency also has a potential impact on further research. According to Kitchin, a big drawback for potentially developing a new paradigm shift in scientific research is a shortage data analytic skillsets.[177] Moreover, Gammage suggests that personnel shortages will impact the public sector more severely, as the private sector will attract more analytics professionals with

---

[170] Al Nuaimi et al., "Applications of Big Data to Smart Cities," 7.

[171] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 10.

[172] Bottles, Begoli, and Worley, 10.

[173] Desouza, *Realizing the Promise of Big Data*, 7.

[174] Henry and Venkatraman, "Big Data Analytics the Next Big Learning Opportunity," 17.

[175] Desouza, *Realizing the Promise of Big Data*, 7.

[176] Henry and Venkatraman, "Big Data Analytics the Next Big Learning Opportunity," 17.

[177] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 10.

higher compensations.[178] This impact is further exacerbated in that technology associated with and advancing analytics is outpacing the talent being produced, including leadership, necessary to reaching the potential of big data.[179]

3. Compatibility

Another significant challenge in implementing data analytics is data governance. Data governance is not appealing to managers but critical for applying data analytics effectively.[180] Schultz reported on a survey of 20 large companies in how big data analytics was being incorporated.[181] The findings suggested the businesses are still exploring how best to apply big data analytics and integrating with legacy systems and traditional datasets presented steep challenges.[182] Seddon and Currie noted in their research that a senior technologist said, "more accurate analytics is simply blocked by older technology."[183] The smart city concept also found that the collection of datasets, especially the unstructured textual data (variety) created challenges for data governance.[184]

Another pitfall of big data is more data means more false data.[185] Bottles, Begoli, and Worley describe big data as a "disruptive technology." This disruption has positives and negatives. The solution, as proposed by Bottles, Begoli, and Worley, results from big data that must be scrutinized and not accepted at face value. Experts asking the right questions and interpreting the results can reduce the risks of false returns.[186]

---

[178] Gamage, "New Development," 389.

[179] Sivarajah et al., "Critical Analysis of Big Data Challenges and Analytical Methods," 265.

[180] Rubin et al., "Harnessing Data for National Security," 126.

[181] Beth Schultz, "View/Review: Big Data in Big Companies," *Baylor Business Review* 32, no. 1 (2013): 20–21.

[182] Schultz, 20–21.

[183] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 303.

[184] Al Nuaimi et al., "Applications of Big Data to Smart Cities," 7.

[185] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 10.

[186] Bottles, Begoli, and Worley, 12.

4. Cost

Developing big data analytics capabilities is costly.[187] Big data analytics is a nascent technology and leaders tend to be weary of investing in big data analytics with so much uncertainty.[188] According to the Wynyard Group, 88% of law enforcement leaders believe that big data analytics will improve law enforcement effectiveness, but only 34% of those surveyed are actually making investments in big data analytics.[189] This contrast is steep compared to the 73% of private sector companies that have, or plan to make, investments in big data analytics.[190] The technology is not cheap but the value, if implemented appropriately, likely will prove to be more effective.

Based on a series of experiments and literature review, a Gartner survey concluded that developing a successful big data environment depends on orchestrating a series of technologies.[191] Adding big data analytics to existing structures can be unsupportable and require more investments.[192] Investments in big data infrastructures are necessary for federal agencies to maximize the value of the technology.[193] For example, in a recent report by MeriTalk, federal IT experts were surveyed with respect to big data and cyber security. The study found that the federal agencies recognize the value in big data but lack the infrastructure and policy to grow and maximize the effectiveness of big data analytics technology.[194] The flood of data in today's world is unavoidable. Recognizing the need

---

[187] Big Data Senior Steering Group, "The Federal Big Data Research and Development Strategic Plan," Digital Commons, University of Nebraska—Lincoln, 19, 2016, http://digitalcommons.unl.edu/scholcom/20/.

[188] Derek Brown, "Analytics in Government: Picking Up Speed or Caught in an Eddy? (Industry Perspective)," Government Technology, 2, August 8, 2016, http://www.govtech.com/opinion/Analytics-in-Government-Picking-Up-Speed-Caught-in-Eddy-Industry-Perspective.html.

[189] Brown, 2.

[190] Janessa Rivera and Rob van der Meulen, "Gartner Survey Reveals that 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years," September 17, 2014, https://www.gartner.com/newsroom/id/2848718.

[191] Chen, Kazman, and Haziyev, "Agile Big Data Analytics Development," 5381.

[192] Brown, "Analytics in Government," 2.

[193] Big Data Senior Steering Group, "The Federal Big Data Research and Development Strategic Plan," 16.

[194] Frank Konkel, "Report: Big Data, Cybersecurity Intrinsically Linked," 1, February 25, 2014, https://fcw.com/articles/2014/02/25/crit-read-big-data-cyber.aspx.

should be a strong impetus for creating the best cyber infrastructure to support the technology advancements.

The continuous growth in data has caused a greater demand for more robust systems.[195] According to Sivarajah et al., prior to investing in big data analytics, organizations should study their respective data environments.[196] Al Nuaimi et al. suggest that improperly planning for implementing data analytics is costly because it is expensive to implement, as the associated hardware is expensive.[197]

E. SUMMARY

Data is growing at exponential rates.[198] Scores of authors promote the use of big data analytics and the potential value. From garnering greater insights to potentially altering the standard scientific method, big data analytics is a growing technology that has great benefits. The public sector, however, is not exercising and exploring the technology at the speed that the private sector is in converting large datasets into insights to make better decisions.[199]

Theoretical frameworks for big data analytics are lacking and literature is scarce for managers that describes how best to develop and integrate big data analytics.[200] Moreover, a framework specifically for law enforcement or for federal investigations is nonexistent. Chen, Li, and Wang demonstrated the promise of leveraging a CI framework for designing and implementing big data analytics applications that potentially could support HSI.[201] Investigations have a set process, and the CI framework offers a format for developing analytics incrementally that can be tested against different types of criminal

---

[195] Sivarajah et al., "Critical Analysis of Big Data Challenges and Analytical Methods," 275.

[196] Sivarajah et al., 263.

[197] Al Nuaimi et al., "Applications of Big Data to Smart Cities," 11; Sivarajah et al., "Critical Analysis of Big Data Challenges and Analytical Methods," 265.

[198] Dragland, "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years."

[199] Gandomi and Haider, "Beyond the Hype," 35.

[200] George, Haas, and Pentland, "Big Data and Management," 321.

[201] Chen, Li, and Wang, "On the Model Design of Integrated Intelligent Big Data Analytics Systems," 1678.

investigations, particularly with respect to the framework's generic and specific models. HSI has a broad mission and validating how analytics is applied against different criminal programmatic areas is necessary before applying analytics nationally.[202]

Seddon and Currie proved that a model based on the seven V's characterizing big data could also be useful. This model is particularly interesting considering the amount of literature referencing the importance of reviewing big data characteristics when implementing analytics.[203] This implementation has particular implications for HSI, as the respective data variety has a "structural heterogeneity," which creates challenges.[204] The potential for homeland security to leverage big data analytics is appealing but how effective and efficient the technology can be is undetermined.

[202] U.S. Immigration and Customs Enforcement, "Homeland Security Investigations," 1.

[203] Al Nuaimi et al., "Applications of Big Data to Smart Cities," 12.

[204] Gandomi and Haider, "Beyond the Hype," 126.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. HUMAN SMUGGLING OPERATIONS

Immigration and Customs Enforcement (ICE), HSI, is the lead agency responsible for investigating human smuggling.[205] ICE defines human smuggling as "the importation of people into a country via the deliberate evasion of immigration laws," which includes smuggling, transporting, and harboring aliens residing illegally.[206]

The three primary tenants of ICE's strategy to dismantle human smuggling networks are the following:

- Conduct intelligence-driven investigations that target the most significant threatening networks.
- Work in conjunction with Customs and Border Protection (CBP)
- Target the entire transnational network, including the logistical, financial, and employment nodes associated with the illicit activities.[207]

A. OFFICE OF BORDER PATROL APPREHENSIONS

Although the Office of Border Patrol (OBP) apprehensions dropped significantly between 2000 and 2016, compared to levels in the 1980s and 1990s, the total numbers are still in the hundreds of thousands as shown in Figure 1.[208]

[205] "Human Smuggling," U.S. Immigration and Customs Enforcement, 1, accessed August 18, 2017, https://www.ice.gov/human-smuggling.

[206] U.S. Immigration and Customs Enforcement, 1.

[207] U.S. Immigration and Customs Enforcement, 1.

[208] "U.S.-Mexico Border Requires Evidence-Based Humanitarian Solutions, Not Border Walls," WOLA, 1, October 27, 2016, https://www.wola.org/2016/10/wola-report-u-s-mexico-border-requires-evidence-based-humanitarian-solutions-not-border-walls/.

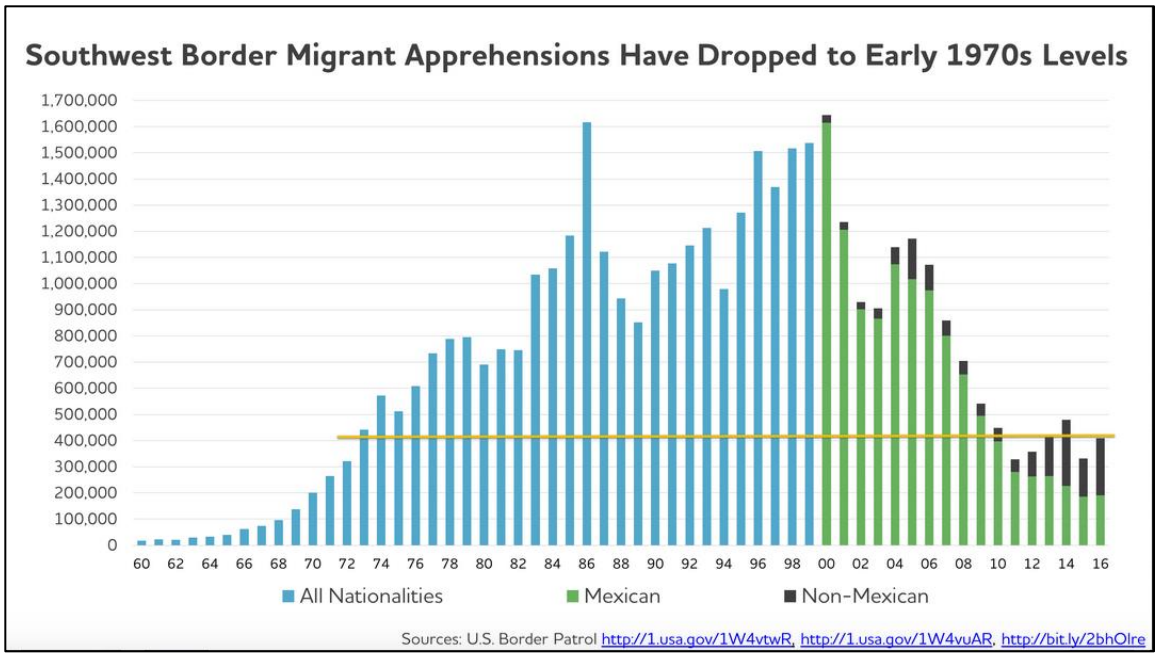

Figure 1. Annual Alien Apprehensions[209]

Furthermore, these figures only include the number of aliens actually caught. According to the Council on Foreign Relations, U.S. immigration enforcement has been largely ineffective despite the billions of dollars the federal government spends annually to enforce immigration U.S. laws.[210] Moreover, this study concluded that the OBP does not apprehend 45–60% of aliens crossing the border illegally.[211] The continuous pressure from large volumes of aliens attempting to enter the country illegally creates a persistent challenge for the 20,000 OBP agents attempting to apprehend hundreds of thousands of aliens annually.[212] Reports related to measurable effects on human smuggling enforcement

[209] Source: "WOLA's Response to February CBP Southwest Border Migration Numbers," WOLA, March 9, 2017, https://www.wola.org/2017/03/wolas-response-february-cbp-southwest-border-migration-numbers/.

[210] Bryan Roberts, Edward Alden, and John Whitley, *Managing Illegal Immigration to the United States* (Washington, DC: Council on Foreign Relations, 2013), 1, https://www.cfr.org/content/publications/attachments/Managing_Illegal_Immigration_report.pdf.

[211] Roberts, Alden, and Whitley, 3.

[212] "Stats and Summaries," U.S. Customs and Border Protection, 1, accessed August 21, 2017, https://www.cbp.gov/newsroom/media-resources/stats?title=Border+Patrol.

actions are unclear.[213] The government publicly acknowledges the resources devoted to human smuggling enforcements but the government fails to report whether those resources applied to enforcement immigration have been effective.[214]

B. HUMAN SMUGGLING OPERATIONS

Increasing the security of the U.S. border began in 1993. The security improvements made it more difficult to enter the country illegally, which led to the creation of the human smuggling industry. David Spencer found during the course of his research in interviewing law enforcement and legal professionals involved in human smuggling enforcement that the immigrant community runs the human smuggling industry.[215] According to Spencer, the increases in border security created "a sophisticated and highly profitable industry dominated by large-scale criminal syndicates that prey on migrants desperate to enter the US without official authorization."[216] Theses smugglers are commonly referred to as *coyotes*.[217] Due to the risks in entering the country unlawfully, professional smugglers are required to increase the likelihood of success.[218] Human smuggling organizations (HSO) capitalize on the desires of thousands of aliens wanting to immigrate into the United States. The ultimate objective of the HSO is to profit from transporting aliens safely to their final destinations within the United States.[219] The demand for the profit-driven smugglers has historically been such that "customer service" is not a priority, as migrant abuses are commonly reported.[220] The smuggling practices

[213] Roberts, Alden, and Whitley, *Managing Illegal Immigration to the United States*, 41–42.

[214] Roberts, Alden, and Whitley, 2.

[215] David Spencer, "Mexican Migrant Smuggling: A Cross-Border Cottage Industry," *Journal of International Migration and Integration* 5, no. 3 (2004): 306.

[216] Spencer, 295.

[217] Spencer, 298.

[218] Spencer, 299.

[219] David Kyle and Rey Koslowski, *Global Human Smuggling: Comparative Perspectives* (Baltimore, MD: JHU Press, 2001), 139.

[220] Spencer, "Mexican Migrant Smuggling," 300.

established in the 1990s continue today, as the processes in which human smuggling operations are conducted have largely remained unchanged.[221]

The sizes of the HSOs along the southwest border range from one or two individuals to larger networks consisting of dozens of personnel.[222] Spencer quoted an attorney who stated that an "organization was too grand of a term" in characterizing the networks involved in human smuggling operations.[223] The human smuggling operations are typically "mom and pop" and are largely unstructured. The structures of the organizations vary from highly organized to loose affiliates.[224] The networks can be assembled based on opportunity to well-organized groups working commercially in the transport of aliens. Kyle and Koslowski characterized the professional human smugglers as "a variety of organizational arrangements that range from a continuum of size and sophistications."[225]

Although the networks are often loose, they can be prolific. The smallest of the human smuggling groups are capable of transporting hundreds of aliens.[226] For example, Spencer found that two human smugglers moved nearly 600 aliens and generated approximately $750K annually by driving loads of people on the weekends.[227] Moreover, becoming involved in human smuggling is easy due to high demand and a lack of obstacles, which enables HSOs to reconstitute quickly against law enforcement successes.[228]

In general, the process works as follows. An alien contacts a Mexico-based facilitator, the facilitator summons a foot guide for navigation across the border, a load driver brings the aliens to a safe house in a metropolitan area along the southwest border,

---

[221] Spencer, 315.

[222] Kyle and Koslowski, *Global Human Smuggling*, 134.

[223] Spencer, "Mexican Migrant Smuggling," 306.

[224] Kyle and Koslowski, *Global Human Smuggling*, 138.

[225] Kyle and Koslowski, 138.

[226] Spencer, "Mexican Migrant Smuggling," 306.

[227] Spencer, 306.

[228] Spencer, 317.

and an additional driver is coordinated to deliver the aliens to their final destination.[229] The payment for the process is negotiated upfront by the alien and family sponsors in the United States who often pay for the transportation to the final destination once the alien is successfully across the border.[230] As noted previously, undocumented aliens run this process, which makes network identification challenging for HSI.

C. CRIMINAL INVESTIGATIONS

Human smuggling has been a persistent problem for the DHS, as evident but for the aforementioned annual OBP apprehensions. Title 8, U.S.C., 1324 is the principal law defining criminality relating to alien smuggling. According to the U.S. Attorney's Office, U.S.C. 1324 "prohibits alien smuggling, domestic transportation of unauthorized aliens, concealing or harboring unauthorized aliens, encouraging or inducing unauthorized aliens to enter the United States, and engaging in a conspiracy or aiding and abetting any of the preceding acts."[231] In short, U.S.C. 1324 prohibits smuggling or even the attempt of smuggling an alien into the United States, which includes encouraging someone to enter the country illegally.[232]

As mentioned previously, a large, unorganized network of people is often involved in human smuggling. Proving criminal violations related to human smuggling, however, is challenging primarily because the prosecutor must demonstrate beyond a reasonable doubt that the offenders involved knew the persons being transported, or harbored, were aliens.[233] A common tactic of the smugglers is portraying themselves to be one of the aliens being smuggled.[234] This tactic makes it difficult for agents to identify and criminally prosecute

---

[229] Kyle and Koslowski, *Global Human Smuggling*, 139.

[230] Kyle and Koslowski, 139.

[231] "1907. Title 8, U.S.C. 1324(a) Offenses," Office of the United States Attorneys, 1, accessed August 21, 2017, https://www.justice.gov/usam/criminal-resource-manual-1907-title-8-usc-1324a-offenses.

[232] Office of the United States Attorneys, 1.

[233] "Alien Smuggling & Harboring Illegal Aliens," Leonardo Law Offices, 1, accessed August 21, 2017, http://www.arizonadefenseattorney.net/alien-smuggling-harboring-illegal-aliens/.

[234] Leonardo Law Offices, 1.

subjects responsible for human smuggling because a large part of the network is undocumented and transnational.[235]

Increasing the probability of apprehensions and enforcing consequences, which requires good evidence, is important toward making immigration enforcement more effective.[236] This enforcement, however, requires better data analyses so that leadership can make more informed decisions in combating illegal immigration.[237]

D. INTELLIGENCE METHODOLOGY

Conducting intelligence-driven investigations and targeting the entire human smuggling network in conjunction with CBP are the strategy principles for ICE. In meeting these principles, HSI Phoenix developed a specific methodology consistent with the strategy principles. HSI Phoenix collects reports from OBP stations conducting interviews of apprehended aliens. One of the primary evidentiary collections methods in demonstrating criminality relates to communications. Contact information related to *coyotes* reported by the apprehended aliens during OBP interviews is subpoenaed and analyzed. Those interview reports are compiled and reviewed by HSI analysts with the purpose of identifying the networks and principal facilitators responsible for human smuggling. In doing so, enough evidence to develop "probable cause" must be collected that can be used to further investigations and prosecute the individuals involved in human smuggling.

E. PROBABLE CAUSE

According to the Fourth Amendment, probable cause is required before a search can be conducted, a warrant can be issued, or an arrest can be made.[238] Probable cause exists when reasonable suspicion is presented to a court that a crime may have been

---

[235] Kyle and Koslowski, *Global Human Smuggling*, 134.

[236] Roberts, Alden, and Whitley, *Managing Illegal Immigration to the United States*, 20.

[237] Roberts, Alden, and Whitley, 52.

[238] "Annotation 1—Fourth Amendment," Findlaw, 1, accessed August 26, 2017, http://constitution.findlaw.com/amendment4/annotation01.html.

committed.[239] Probable cause, however, is subject to interpretation. The U.S. Supreme Court has made several definitional rulings but considers probable cause to be determined predominantly on a case-by-case basis.[240]

The intelligence process for generating evidence related to HSO networks starts with OBP interview reports. OBP agents conduct interviews of undocumented aliens who have been apprehended. In the course of these interviews, OBP agents report that HSO smuggle undocumented aliens into the United States and coordinate the locations of the aliens, drivers, and safe houses via cellular telephone. OBP interviews capture those suspected phone numbers of individuals involved in the alien smuggling organizations and subsequently shares the information with HSI. HSI in turn, issues subpoenas for the call detail records of certain *coyotes* that the apprehended aliens identify. A "call detail record" contains "communication event" data, including cellular telephone numbers, internal electronic numbers for billing and switching purposes, the general geographic location of cell sites accessed by the telephone, and cellular tower utilization information including time and date.

A review of five adjudicated affidavits used in the course of the aforementioned process reveals the common basis for probable cause necessary to issue warrants against phone numbers reportedly involved in human smuggling.[241] The central lines of evidence stem from total number of OBP reports tied to a suspected phone number, volume of communications with the suspected phone number in contact with other phone numbers from OBP interview reports, and the financial transactions connected to the phone number suspected. Based on the review of five previous warrants, enough probable cause was presented when a minimum of two OBP reports, 5,000 call records in contact with at least

[239] John C. Busby, "Probable Cause," Cornell Law School, Legal Information Institute, 1, September 17, 2009, https://www.law.cornell.edu/wex/probable_cause.

[240] Busby, 2.

[241] "Affidavit in Support of Application for Search Warrant, Case No. 16-8143MB," Homeland Security Investigations, April 28, 2017; "Affidavit in Support of Application for Search Warrant, Case No. 16-8144MB," Homeland Security Investigations, April 28, 2017; "Affidavit in Support of Application for Search Warrant, Case No. 16-9434MB," Homeland Security Investigations, May 20, 2017; "Affidavit in Support of Application for Search Warrant, Case No. 16-6352MB," Homeland Security Investigations, July 16, 2017; "Affidavit in Support of Application for Search Warrant, Case No. 16-8070MB," Homeland Security Investigations, September 12, 2017.

30 other suspects' phones, and five financial transactions connected to the suspected phone number was articulated.[242] In other words, if the phone number was listed at least once in two separate OBP interview reports, a high volume of communication with other phones was also listed in other OBP reports, coupled with financial transactions with the same phone number, sufficient probable cause occurred to substantiate a warrant.

Understanding the level of probable cause required to substantiate a warrant is important in targeting HSO networks, as this understanding becomes the basis for proving human smuggling criminal violations. Measuring the level of evidence, however, while concurrently analyzing a human smuggling network is challenging. Put differently, numerous phone numbers may be linked through several OBP reports, but determining whether enough probable cause evidence already is present must be measured individually. Knowing the level of probable cause determines investigative actions or highlights where additional evidence collection is necessary in the furtherance of an investigation. Having the means of efficiently identifying subjects involved and immediately knowing the level of evidence associated is of great benefit. The process for reviewing, prioritizing, and analyzing the communications originating from the OBP interview reports is laborious, particularly when evaluating the level of probable cause, but important for identifying the HSO networks.

[242] Review of the warrants was conducted specifically for this study. Approval of warrants varies according to the discretion of attorneys, magistrates, and judges, etc.

V. TESTING AND ANALYSIS

The associated systems, data, and two types of queries (search and discovery) structure this chapter. Comparing the efficiency and effectiveness of the manual analysis process compared to Citrus are divided by three datasets: OBP reports, communication records, and financial transactions. These same data variables are used to produce probable cause when pursuing human smuggling actors. The research design follows the HSI strategy and intelligence methodology for determining human smuggling targets noted in Chapter IV. The date range for the entire dataset used for this study is from April 2015–September 2017. The OBP reports comprise the pivotal dataset, as those reports provide the most substantial evidence. The communication records and financial transactions supplement the evidence gained from the OBP reports. The prioritized results are based upon the volume of phone numbers that occur from the OBP reports within 45, 90, and 180 days, which range from April 2017–September 2017. The communication records and financial transaction queried for the tests, supplementing the highest volume phone numbers originating in the OBP reports, include the OBP reports date ranges of 45, 90, and 180 days, as well as the historical data back to 2015. The most recent phone numbers are valuable, particularly if a phone numbers is associated with historical reporting. Through this process, targets are prioritized according to the combination of datasets relating to OBP reports, communication records, and financial transactions over set time periods (see Figure 2). At the end, the phone numbers with the highest volume of datasets/evidence are considered the highest priority human smuggling targets.

A. SYSTEMS PROCESS

Data for targeting human smuggling networks in analyzing the OBP reporting, communications, and financial information is pulled from independent systems. The OBP reports are emailed and stored on a local shared drive. The communications records are uploaded and analyzed in Pen-Link, which is a system HSI uses to analyze subpoenaed

telephone records.[243] The financial data is downloaded from the Southwest Border Anti-Money Laundering Alliance (SWBA) system and analyzed separately. The primary concern with the systems process is that data resides in independent and unconnected sources, which requires separate queries and separate coalescing for the decision maker.

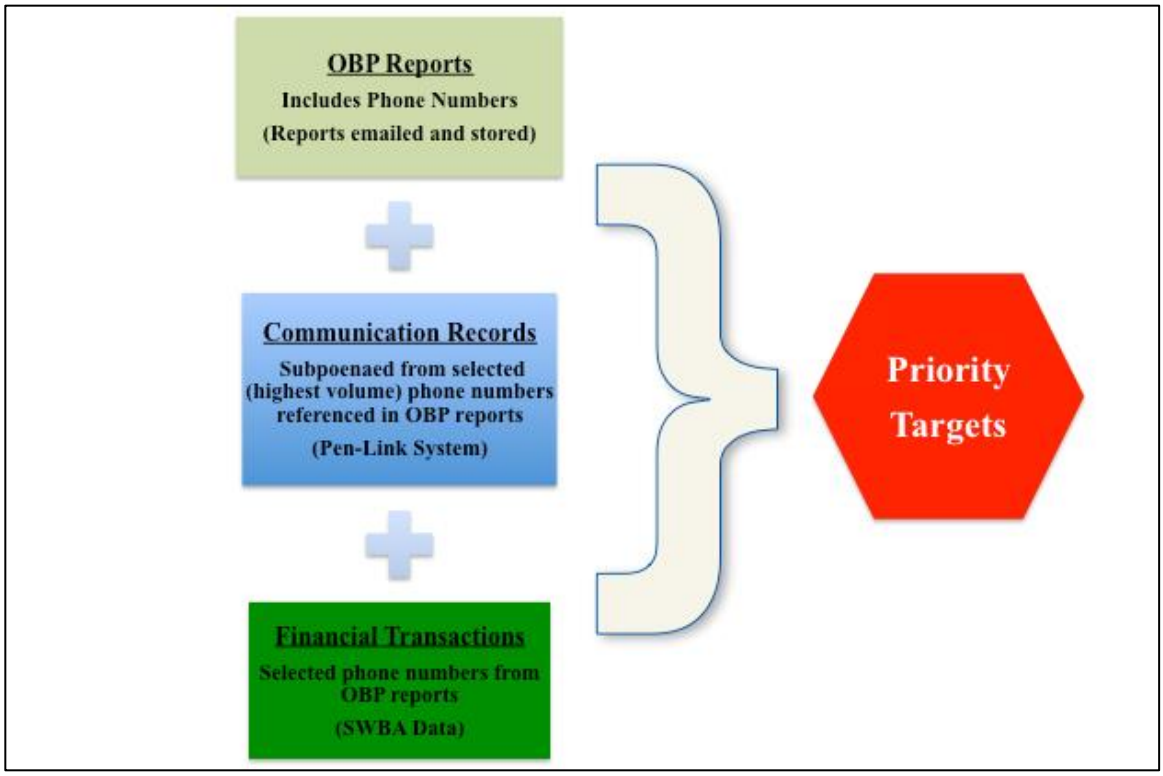

Figure 2. Analysis Process

The *variety* of datasets available to support this research consists of the following:

- Subpoena Returns: 2,100
- Unique phone numbers: 225K
- Number of phone records: 2M

[243] "About Pen-Link," PenLink, Ltd., accessed October 5, 2017, https://www.penlink.com/About/tabid/123/Default.aspx.

- Number of OBP reports: 4,500
    - Average number of discovered phone numbers per OBP report: 9
- Number of financial transactions: 100K

B. DATA CLEANSING

In general, it takes 3–4 hours to clean the phone records from one subpoena return. Phone carriers provide subpoena returns in different formats. Moreover, the data in each return are captured inconsistently. A typical return for phone number 123-456-7890, hypothetically, may be provided in the following formats, as show in Table 1.

Table 1. Unconditioned Phone Record

123-456-7890
(123)-456-7890
11234567890
1234567890
1-521234567890
1234567890
1123-456-7890
11234567890
1-521234567890
(123)-456-7890

Each phone number needs to be cleaned and structured to ensure consistent analysis. If the data were ingested without cleaning, each phone number would be counted as different entities, which affects the data *veracity* that thus skews analysis. Manually cleaning each phones number from hundreds of phone records takes an exorbitant amount of processing time. This step is an important consideration when comparing efficiencies, as the total time in determining human smuggling priority targets does not include the time dedicated to cleaning phone records.

To test whether Citrus could *clean* all phone records, a specific set of heuristics was created. For example, country codes, characters, and erroneous phone number prefixes

were removed. The goal was to determine if Citrus could read each subpoena and clean the associated phone numbers. A script was created to measure the results. Citrus was able to read 270 (88%) of the subpoena returns and adjusted over 410K (20%) phone records. This result builds confidence that Citrus was successful in reading and cleaning phone numbers. To what extent Citrus cleaned the records better or worse compared to a manual process was immeasurable, as the manual process could not be calculated. Assuming the human process was more accurate, the time in cleaning the data outweighs a marginal improvement in cleaning manually.

C. SEARCH QUERY

The purpose of the search query is to determine which phone numbers were recorded the most over 45-, 90-, and 180-day periods. Those prioritized phone numbers are subsequently queried against *unstructured,* historical OBP reporting (beyond 180 days), communications, and financial data to determine the volume of associated data linked to the highest priority phone numbers.

For the purposes of the research, the following methodology was established in testing the efficiency and effectiveness of Citrus compared to a manual process conducting the same search criteria:

- Determine the 10 most occurring phone numbers recorded from 45, 90, and 180 days of reporting.
- Search these top phone numbers through the entire database to determine the number of additional historical OBP reports associated with these phone numbers.
- Determine the total number of financial transactions associated with the top phone numbers.
- Determine the total number of communication contacts of each phone number and the volume of calls between each contact.

1. Manual Results (Efficiency)

Following the analysis process (Figure 1), each dataset (OBP reports, communication records, and financial transactions) must be reviewed independently under the manual process. Sifting through hundreds of OBP reports manually to determine which phone numbers were recorded the most projected to take hundreds of hours, as reading and cataloging phone numbers from four–five OBP reports takes approximately one hour. Once the top 10 phone numbers from each time period are identified, recycling the prioritized phone number list to determine if additional phone numbers have historical OBP reporting is virtually impossible to calculate because of the total number of hours such a process takes to complete. Moreover, manually processing the top 10 phone numbers through Pen-Link in evaluating the communication records is manageable, with results being returned within an hour. The financial connections were timelier, as returns were produced in minutes derived from the SWBA database. See Table 2.

Table 2. Manual Efficiency Results

Search (Efficiency)						
Days of Reporting	**Total Reports**	**OBP Initial Review**	**OBP Recycle**	**Communication Review**	**Financial Review**	**Total**
45	225	50 hours	Incalculable	1 hour	10 minutes	51.1 hours
90	450	100 hours	Incalculable	1 hour	10 minutes	101.1 hours
180	900	200 hours	Incalculable	1 hour	10 minutes	201.1 hours

2. Citrus Results (Efficiency)

Conducting the same queries with Citrus required just seconds, which included recycling the top phone numbers in determining if additional historical OBP reports existed. In other words, Citrus was able to prioritize phone numbers from 45-, 90-, and 180-day increments almost instantly. Moreover, Citrus was able to read and clean hundreds of thousands of phone records from 2,100 subpoena returns, which further saved hundreds of hours processing data; however, it also did not include the time it took to toggle between multiple systems in comparing results. The timesaving for Citrus conducting the same

analysis process manually in triaging phone numbers from thousands of OBP reports, communication and financial records proved to be exponentially more efficient, as seen in Table 3.

Table 3. Citrus Efficiency Results

Search (Efficiency)						
Days of Reporting	**Total Reports**	**OBP Initial Review**	**OBP Recycle**	**Communication Review**	**Financial Review**	**Total**
45	225	10 seconds	10 seconds	10 seconds	10 seconds	10 seconds
90	450	10 seconds	10 seconds	10 seconds	10 seconds	10 seconds
180	900	10 seconds	10 seconds	10 seconds	10 seconds	10 seconds

3. Manual Results (OBP Reports Effectiveness)

Manually evaluating the 10 most occurring phones from the OBP dataset of each time period as shown in Table 4 resulted in 15 individual phone numbers that were listed in two or more time periods. The actual phone numbers were anonymized and referenced as A-O. Five phone numbers (A, G, I, C, D) were recorded in all three reporting periods: three phone numbers (B, E, H) recorded in the 90- and 180-day periods, two phone numbers (L, K) in the 45- and 180-day periods, three phone numbers (N, O, M) recorded only during the 45-day period; and two phone numbers (F, J) only recorded in the 180-day period. The manual results produced a total of 181 OBP reports associated with the top 15 phone numbers. The manual results concluded with phone number D being the highest priority phone number, with a total of 29 OBP reports associated.

Table 4. Manual Effectiveness Results

Search Effectiveness			
	45 Days	**90 Days**	**180 Days**
Phone Number	**OBP Reports**	**OBP Reports**	**OBP Reports**
A	9	11	22
B	0	10	14
C	8	12	18
D	12	19	29

Search Effectiveness			
	45 Days	**90 Days**	**180 Days**
Phone Number	**OBP Reports**	**OBP Reports**	**OBP Reports**
E	0	12	18
F	0	0	13
G	8	12	18
H	0	12	19
I	6	11	19
J	0	0	11
K	9	9	0
L	8	9	0
M	6	0	0
N	5	0	0
O	7	0	0
Grand Total	**78**	**117**	**181**

4. Citrus Results (OBP Effectiveness)

In addition to the same total number of OBP reports from the manual results, the top 15 phone numbers were recycled through the historical OBP database with Citrus, which increased the total number of OBP reports outside of 180 days to 237 that resulted in a 33% increase compared to the overall manual results. The increases in effectiveness, however, varied during each time period. The Citrus average increased to 159% within the 45-day period. Considering the exorbitant amount of time it would take to triage 90 and 180 days with OBP reporting, the effectiveness is arguably more accurate at 159% compared to the 30% overall average. With the exception of phone number D, every phone number had additional associated reports outside of the defined reporting periods that results in more evidence, captured more efficiently, which potentially can be undiscovered without a data analytics application like Citrus.

Citrus instantly prioritized phone numbers at different time periods. This prioritization is significant, as it lends to a greater understanding of uptrends or downtrends relative to the prioritized phone numbers. For example, phone number D was ranked as the top phone number for the manual and Citrus results. With Citrus, however, Citrus instantly displays that phone number D also had the highest number of reports within the 45-day

period. Therefore, D is not only high in volume; the target number is likely surging in activity. Comparably, the majority of the reporting relative to phone number E occurred prior to 90 days and was unlisted within the 45-day period. Phone number E could be obsolete, which would assist in validating that phone number D should be the highest priority target. Citrus was not only more effective in identifying more evidence, it also enabled an increased capability of prioritizing targets more effectively, as shown in Tables 5–7.

Table 5. Citrus 45-Day Effectiveness Results

Phone Number	Days	Manual OBP Reports	Citrus Count	Increase Percentage	Increase Average
A	**45**	9	23	155.56%	
C	**45**	8	27	237.50%	
D	**45**	12	29	141.67%	
G	**45**	8	25	212.50%	
I	**45**	6	27	350.00%	
K	**45**	9	10	11.11%	
L	**45**	8	14	75.00%	
M	**45**	6	9	50.00%	
N	**45**	5	15	200.00%	
O	**45**	7	13	85.71%	159%

Table 6. Citrus 90-Day Effectiveness Results

Phone Number	Days	Manual OBP Reports	Citrus Count	Increase Percentage	Increase Average
A	**90**	11	24	118.18%	
B	**90**	10	22	120.00%	
C	**90**	12	27	125.00%	
D	**90**	19	29	52.63%	
E	**90**	12	27	125.00%	
G	**90**	12	25	108.33%	
H	**90**	12	25	108.33%	
I	**90**	11	27	145.45%	
K	**90**	9	10	11.11%	
L	**90**	9	14	55.56%	**96%**

Table 7. Citrus 180-Day Effectiveness Results

Phone Number	Days	Manual OBP Reports	Citrus Count	Increase Percentage	Increase Average
A	**180**	22	25	13.64%	
B	**180**	14	22	57.14%	
C	**180**	18	27	50.00%	
D	**180**	29	29	0.00%	
E	**180**	18	27	50.00%	
F	**180**	13	16	23.08%	
G	**180**	18	25	38.89%	
H	**180**	19	26	36.84%	
I	**180**	19	27	42.11%	
J	**180**	11	13	18.18%	**33%**

5. Financial Effectiveness

Manually querying each of the highest priority phone numbers directly through the SWBA database produced far greater results. The results were aggregated from all three time periods, considering only one transaction was identified with Citrus within the 45-day timeframe. This result can be explained due to the limited amount of data accessible for Citrus. In other words, the financial dataset for this research was limited in scope to Arizona. The SWBA database is much larger, and therefore, produced more transactions. With application programming interface (API) access, the Citrus results likely would have been compared with the manual results. Phone number O was the sole phone number with a Citrus-identified transaction. The manual query connected with three transactions that used the same phone. Without an equal dataset, the results are inconclusive regarding whether Citrus is more or less effective with regard to financial transactions, but for the purposes of the research, manual querying was more effective, which is demonstrated in Table 8.

Table 8. Financial Effectiveness Results

Search Effectiveness	Manual Financial		Date Ranges		Citrus Financial	Date Ranges
A	3		>180 days		0	
B	0				0	
C	0				0	
D	0				0	
E	15		>180 days		0	
F	3		>180 days		0	
G	0				0	
H	0				0	
I	0				0	
J	0				0	
K	0				0	
L	1		>180 days		0	
M	1	3	90 days	>180 days	0	
N	0				0	
O	3		>180 days		1	45 days
Grand Total	**29**				**1**	

6. Communication Effectiveness

Although the overall communication results were relatively scarce compared to the OBP reports and financial transactions, effectiveness in identifying contacts and communication transactions were greater with Citrus. Phone number J had the highest number of contacts and volume of phone records from the Pen-Link search and Citrus; however, the number of contacts and volume of transactions identified by Citrus increased by 900% and 200%, respectively, albeit from only one phone number that likely can be explained by a manual data conditioning process, as the *variability* of the data is subject to human errors. Citrus was able to condition 88% of the subpoena returns. The increase from Citrus could mean the cleaning process was more accurate. The automated data conditioning process conducted with Citrus likely reduced the human errors, which increased the *veracity* of the data. This result, as shown in Table 9, is further supported in that four phone numbers (B, E, H, and O) were identified by Citrus with contacts and transactions that had remained unnoticed and unrecognized through the manual process recorded in Pen-Link.

Table 9. Communication Effectiveness Results

Search Effectiveness						
Phone Number	**Manual Pen-Link Contacts**	**Manual Pen-link Transactions**	**Manual Pen-Link Date Range**	**Citrus Contacts**	**Citrus Transactions**	**Citrus Date Range**
A	0	0		0	0	
B	0	0		4	20	90 days
C	0	0		0	0	
D	0	0		0	0	
E	0	0		4	18	90 days
F	0	0		0	0	
G	0	0		0	0	
H	0	0		2	12	180 days
I	0	0		0	0	
J	3	1206	45 days	31	3628	45 days
K	0	0		0	0	
L	0	0		0	0	
M	0	0		0	0	
N	0	0		0	0	
O	0	0		1	1	
Grand Total	**3**	**1206**		**42**	**3679**	

7. Overall Effectiveness

With the exception of the financial transactions, Table 10 shows that effectiveness drastically increased with Citrus overall. Conversely, gaps in evidentiary coverage are more effectively highlighted with Citrus. In other words, knowing which targets require more evidence gathering presents a new type of value for investigations. Phone number A, for example, has a high number of OBP reports and financial transactions but no communication results, which indicates that more evidence should be collected against that phone number to determine significance of the potential target. This increase in effectiveness coupled with the exponential increase in efficiency equates to a much more productive means of triaging and identifying subjects actively involved in human smuggling activities.

Table 10. Overall Efficacies

Search Effectiveness								
Phone Number	**Manual OBP Reports**	**Citrus Totals**	**Manual Financial**	**Citrus Financial**	**Manual Pen-Link Contacts**	**Citrus Contacts**	**Manual Pen-Link Transactions**	**Citrus Transactions**
A	22	3	12	0	0	0	0	0
B	14	8	0	0	0	4	0	20
C	18	9	0	0	0	0	0	0
D	29	0	0	0	0	0	0	0
E	18	9	30	0	0	4	0	18
F	13	3	3	0	0	0	0	0
G	18	7	0	0	0	0	0	0
H	19	7	0	0	0	2	0	12
I	19	8	0	0	0	0	0	0
J	11	2	0	0	3	31	1206	3628
K	9	1	0	0	0	0	0	0
L	9	5	2	0	0	0	0	0
M	6	3	4	0	0	0	0	0
N	5	10	3	0	0	0	0	0
O	7	6	3	1	0	1	0	1
Grand Total	**217**	**81**	**57**	**1**	**3**	**42**	**1206**	**3679**

D. DISCOVERY QUERY

The purpose of the discovery queries is to determine the usefulness of Citrus in determining level of evidence (probable cause) when applied against aggregated data, and the effectiveness of identifying potential human smuggling networks when applied against thousands of OBP reports. Applying data analytics to aggregated datasets has the potential to make new discoveries. To test this hypothesis with Citrus, two types of discovery reports were designed and created, the QE report and the CO report.

1. Quality of Evidence Report

The purpose of the QE report is to prioritize phone numbers based upon the level of evidence resident within merged data. The QE report essentially fuses the aforementioned search query process of evaluating a target phone based on OBP reports, financial transactions, and communication records. More critically, the QE report

potentially identifies phone numbers with enough probable cause for HSI to pursue investigative action. A weighted scale was created based on previous affidavits with a maximum value of 2.0.

QE is the sum of the following values:

- 0.8 x number of OBP reports the phone number
- 0.6 x number of contacts of the phone number
- 0.4 x volume of call records between the contacts of the phone number in the same date range
- 0.2 x financial transaction with the phone number

The QE formula was applied to the entire database to produce the prioritized list presented in Table 11.

Table 11. Quality of Evidence Report Results

Phone	**QE Score**	**Action**
Phone Number 1.	1.426	Unpursued
Phone Number 2.	1.418	Arrested
Phone Number 3.	1.416	Arrested
Phone Number 4.	1.415	Arrested
Phone Number 5.	1.414	Arrested

The top five phones numbers with the highest score were reviewed. The top phone number with the highest score of 1.426 was a phone number not currently being actioned by HSI. The target phone was undiscovered within the voluminous amount of data. This substantiates the need for analytics, as important targets will be missed if analysis is conducted manually. With so much data, for this study, with over two million phone records, overlooking an important target is likely. The remaining phone numbers were associated with previous human smuggling targets apprehended and prosecuted for human smuggling activities. This result demonstrated the value in creating analytics that calculate

probable cause. The data associated with the previous apprehended targets became benchmarks for validating the QE calculations. The formula proved to be accurate, as four of the top five phone numbers were associated with prosecuted criminals. This finding creates *value* for HSI in that an automated methodology may be used to identify entities where ample evidence unknowingly exists within a large database. This result arguably can accelerate investigations knowing that enough probable exists necessary, by entity, to purse prosecution resides within a large dataset.

2. Co-Occurrence Report

The purpose of the CO report was to determine potential networks. The premise was that if two or more phone numbers were mentioned collectively in three of more reports, then those phone numbers would be considered networked and become a search priority.

The CO was applied to 900 OBP reports with over 6,400 individual phones, equating to approximately 36 million pairs. Excluding the time to develop the necessary code to perform the calculations, the processing time with Citrus was completed in 10 seconds. The results generated a list of over 100 pairs of phone numbers listed in three or more reports. Three of the top U.S.-based phone numbers were researched further. The results identified were human smuggling *coyotes* with ties to a significant human smuggling organization with open investigations. The undiscovered relationships were that foot guides remained unlisted in currently "open" investigations. This finding indicates suspects listed in multiple OBP reports collectively were not linked. The calculations would have been virtually impossible manually.

Citrus increased the efficiency and effectively greatly. The increases in efficiencies are definitive. With Citrus, investigators would save hundreds of hours in prioritizing targets. Effectiveness also increased, particularly with more comprehensive reports like QE and CO. This increase likely would have significant consequences for investigations, as it would shift priorities, as well as produce more evidence against human smuggling networks.

E. CONTEXT

While Citrus proved to be more efficient and effective, analyzing the individual reports directly could produce different results. Although this analysis would be an impossible feat, understanding the context surrounding each phone number potentially could change judgements. This understanding supports the research by Efros in that analytics explains what is occurring but not necessarily the cause.[244] To explain, if an OBP report presented articulable facts of a human smuggler, including the phone number, who violated aliens, unless the associated phone number was captured several times, or an agent ensured leadership was made aware, a significant incident could be missed. While searching for phone numbers principally based on volume and frequency is valuable, single reports referencing significant events likely would be overlooked.

Furthermore, as asserted by Bottles, Begoli, and Worley, unhidden biases in data create risk.[245] The phone numbers for this research were not subpoenaed randomly. The communication data was assembled without the use of analytics.[246] The phone numbers originally selected for subpoenas were based on undefined assumptions, as no defined measures were established for deciding which numbers would be subpoenaed. In other words, an inherent bias exists within the data. Moreover, OBP interview reports likely were biased. The phone numbers were collected from human interactions. Without knowing exactly how the agents concluded as to why certain phone numbers were recorded lends to biases from the reporting. For example, if a phone number for a bus station was mentioned more than a phone number for a named *coyote*, a key target could be overlooked. OBP agents may assume that all phone numbers, whether mentioned during an interview or discovered written on a receipt from an alien's belongings, are connected to illicit activities. Given the volume of reporting, frequency-based analytics is more effective. Adding context, however, could skew analysis. Experts scrutinizing results from analytics is important in mitigating risks from data biases, as emphasized by Bottles, Begoli, and

[244] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 9.

[245] Bottles, Begoli, and Worley, 10.

[246] The communication data was generated from subpoenaed phone numbers recorded in OBP reports.

Worley.[247] Citrus is exceedingly more efficient and effective but human examination is still important to the final analysis in mitigating inherent data risks.

[247] Bottles, Begoli, and Worley, "Understanding the Pros and Cons of Big Data Analytics," 12.

VI. FINDING, IMPLICATIONS, AND CONCLUSION

Citrus works well for triaging large amounts of data. The efficiency of Citrus to sift through voluminous amounts of reporting and communication and financial data is convincing. The effectiveness also is substantially better with Citrus. The number of additional reports and the capability to calculate probable cause was decisively more effective with Citrus compared to a manual process. The one caveat, however, is including context would arguably add more value. According to Sokol and Chan, "organizations must learn how to apply context to big data, or the conclusions and mission-critical decisions that are made from the analysis might be in error."[248] Context analytics, as proposed by Sokol and Chan, can understand relationships from entities within disparate datasets.[249] This type of solution would enhance analysis, as HSI could create more value with analytics if the entity relationships could be understood in addition to the frequency or volume.

Applying data analytics to prioritize information is ideal—focusing results on a particular set of targets certainly saves time—but having a machine automatically suggest subjects to apprehend would be risky considering the inherent biases and lack of context. Having evidence generated by data analytics, like QE, and subsequently reviewed by legal authorities, also reduces risks, as long as the biases in the production of such results are also understood. Analytics is better overall for triaging large volumes of reporting. The next step for HSI would be to introduce context analytics in the analysis process. Volume plus contextual analytics would enhance the value of analysis, which then subsequently enhances the overall effectiveness of big data.

A. IMPLICATIONS

Investigative discoveries could be made more efficient and effective with data analytics. The implications for Citrus are significant, particularly relating to changing

[248] Lisa Sokol and Steve Chan, "Context-Based Analytics in a Big Data World: Better Decisions," *IBM RedBooks Point-of-View Publication*, 2013, 1.

[249] Sokol and Chan, 2.

analytical tradecraft, prioritizing information, revamping data systems, collecting evidentiary data, increasing investigative process capacities, and developing future data analytics applications.

1. Analytical Tradecraft

The application of data analytics could reshape analytical tradecraft. Hare and Coghill support this reasoning and asserted that analytics will allow analysts to create and answer deeper hypotheses.[250] The CO report is a prime example of a more complex hypothesis that could be tested and answered with data analytics to lead to greater criminal network identifications. In the case of the CO report, a hunch of whether multiple phone numbers were repeated in numerous reports was questioned. The question was conceived because a data analytics tool was available to answer a more complex question. Without analytics, the question would have been impossible to answer. This type of deeper analysis could create new forms of analytical tradecraft, as data analytics potentially creates an unlimited means of reviewing and analyzing data in bigger ways.

2. Prioritizing Information

Citrus demonstrated extraordinary value in triaging information. Instantly prioritizing criminal networks alleviates doubt as to which networks should be targeted first. In addition to instant prioritization, Citrus demonstrated how data analytics can provide measurable reasons as to why certain subjects or networks should be the highest priority. For example, if a tip were received that a certain subject was involved in criminal activity, determining the importance of the reported, suspicious information could be made relevant instantly compared to what data currently exists. HSI is thus able to determine confidently which investigations should receive the most appropriate amount of resources in targeting a particular criminal network.

[250] Nick Hare and Peter Coghill, "The Future of the Intelligence Analysis Task," *Intelligence and National Security* 31, no. 6 (2016): 865.

3. Merging Data

Advancing data analytics would require HSI to remove barriers between data systems, which is imperative to maximizing the value of data analytics. Where currently investigative research is piecemealed from different systems, e.g., Pen-Link, SWBA etc., data analytics presents the opportunity to merge datasets together to allow correlated discoveries much more efficiently that Eggers refers to as "horizontal government" in which government organizations can leverage analytics more efficiently by punching through silos of data.[251] HSI will have to collapse data silos to maximize the potential of data analytics.

HSI should move beyond systems designed to work well against one particular dataset to aggregated data from across the breadth of systems. Data systems should be recognized as more valuable when merged with other datasets; however, caution must be given when merging new data analytics technologies with old government legacy systems. As demonstrated with Citrus, aggregated data can immediately calculate if probable cause exists. Flaws in data integrity however can emerge when data is stored in different systems over time. Seddon and Currie put this issue succinctly by claiming legacy government systems can impede the accuracies of data analytics.[252] HSI fits this mold, as the agency is structured to pull data manually from different systems created and updated from different periods of time. This structure supports Babuta's research regarding big data supporting law enforcement. Babuta concluded that law enforcement data is fragmented in multiple different data systems, which creates inefficiencies in policing.[253] Revamping the current HSI systems architecture would be necessary in evolving to a more data-driven organization through analytics.

---

[251] William D. Eggers, *Delivering on Digital: The Innovators and Technologies that Are Transforming Government* (New York: RosettaBooks, 2016), 2322.

[252] Seddon and Currie, "A Model for Unpacking Big Data Analytics in High-Frequency Trading," 303.

[253] Alexander Babuta, *Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations, and Priorities* (Westminster, London: Royal United Services Institute, 2017), vii, https://rusi.org/publication/occasional-papers/big-data-and-policing-assessment-law-enforcement-requirements.

4. Evidentiary Data Collection

An unanticipated finding from Citrus was the identification of evidentiary gaps. Citrus was designed to make new discoveries, but Citrus also immediately identified where collections should be focused that potentially results in a more efficient evidentiary data collection process. Moreover, if the collection process could be automated, the surge in data for HSI would be tremendous. If a machine could not only identify the gap but also nominate collection requirements on all unknown, needed evidence within a large database, the pieces to the proverbial puzzle could be placed faster. Automating evidentiary data collection should be evaluated further, as long as proper oversight is emplaced to mitigate data integrity risks. Considering the challenges with respect to privacy and government overreach, a machine collecting prescribed evidence can help substantiate as to why the government is collecting larger datasets.[254] If implemented, the government would not be aimlessly collecting data in bulk. Any large data collected could be justified and explained.

5. Investigative Processes

With increases in efficiencies through data analytics, the analysis process and production could outpace investigation processes. The QE report results are compelling but present a possible dilemma. Probable cause drives investigations, which is critical for establishing court orders, warrants, vehicle trackers etc. The debate as to whether enough probable cause exists is key to furthering investigations. If analytics could immediately identifying which entities or persons within the data already possess enough probable cause exist, the HSI investigation process would theoretically be accelerated.

The speed in which investigations could be accelerated with analytics is valuable; however, the judicial process is limited by capacities. With a finite number of attorneys, magistrates, and judges, etc., an uptick in investigative efficiencies may be undercut if the judicial process is unable to support. The unanswered question is if a machine could instantly produce probable cause on hundreds of subjects, would the judicial system be able to maintain pace? Further questions related to court capacities would have to be

[254] Desouza, *Realizing the Promise of Big Data*, 13; McNeely and Hahm, "The Big (Data) Bang," 308.

explored. For example, how many warrants could a single magistrate review and approve daily? This limitation would cause a renewed focus in prioritizations but the outputs and production could be measured, projected, and budgeted.

Bean argues that legislation is overdue to increase the efficiencies of issuing warrants while protecting Fourth Amendment rights.[255] With improvements in technology, as Bean asserts, law enforcement can rapidly produce and transmit affidavits to the courts electronically, which greatly decreases the time to necessary to obtain a warrant. According to Bean, law enforcement officers have a practice of using the exigent circumstances exception of the Fourth Amendment to circumvent the slow judicial process in obtaining warrants.[256] With advancements in technologies, affirming warrants can be streamlined, which arguably will become more prevalent with the advent and expansion of data analytics. As presented by Bean, this trend is supported by an Eighth Circuit decision in that the court judged that streamlining the issuance of warrants protects the Fourth Amendment.[257] By making warrants conveniently obtainable, law enforcement officers are more apt to pursue them. This convenience relates directly to applying for warrants, however. Upgrading the processing capacities for obtaining a warrant will become vital as analytics becomes more prevalent. The basis for a warrant is probable cause. Analytics can be an accelerant for this future process, as probable cause can be generated rapidly, which accelerates warrant applications. Without modernization, the current slow warrant application process will be worsened.

6. Data Analytics Development

Time to develop, craft, and update analytics by criminal programmatic areas potentially would be time consuming initially. HSI Phoenix spent approximately three months defining requirements and developing the Citrus analytics to support one particular criminal programmatic area, that of human smuggling.

---

[255] Andrew H. Bean, "Swearing by New Technology: Strengthening the Fourth Amendment by Utilizing Modern Warrant Technology While Satisfying the Oath or Affirmation Clause," *Brigham Young University Law Review* 2014, no. 4, art. 5 (2014): 949.

[256] Bean, 932.

[257] Bean, 932.

HSI operates worldwide. The agency has a broad mission with authority to investigate over 400 criminal statutes.[258] The way in which HSI Phoenix targets human smuggling networks likely would be different compared to HSI offices from U.S. cities in the Midwest. Different datasets, methodologies would be required; thereby, the analytics would have to be reshaped. Cost in development, hiring data engineers, upgrading cyber infrastructure etc. would be considerations when developing and proliferating analytics. How this process would work for HSI is uncertain. In short, data analytics is a disruptive technology that could conceivably have a ripple effect on HSI and require changes from cyber-infrastructure to personnel with more data qualifications. The data boom is present and analytics is arguably critical for HSI future success, but how HSI develops, integrates, and propagates the technology requires further exploration.[259]

B. CONCLUSION

The research question for this thesis is "How can big data analytics improve the effectiveness and efficiency of HSI targeting human smuggling networks?" With analytics, HSI would not be limited to human capacities since machine calculations would be much more efficient and effective, as demonstrated with Citrus. Put succinctly, data analytics can produce and streamline knowledge for HSI to lead to greater conclusions.[260] Exploring and investing further in the technology should be a high priority, as data analytics offers HSI enormous potential.

[258] "Written Testimony of U.S. Immigration and Customs Enforcement Homeland Security Investigations-Phoenix Special Agent in Charge Matthew Allen for a House Committee on Homeland Security, Subcommittee on Border and Maritime Security Field Hearing Titled "Stopping the Flow of Illicit Drugs in Arizona by Leveraging State, Local and Federal Information Sharing," Department of Homeland Security, 2, May 21, 2012, https://www.dhs.gov/news/2012/05/21/written-testimony-us-immigration-and-customs-enforcement-house-homeland-security.

[259] Marr, "Big Data," 1; Rubin et al., "Harnessing Data for National Security," 21.

[260] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," 1; George, Haas, and Pentland, "Big Data and Management," 324.

LIST OF REFERENCES

Al Nuaimi, Eiman, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. "Applications of Big Data to Smart Cities." *Journal of Internet Services and Applications* 6, no. 1 (August 2015): 1–15.

Babuta, Alexander. *Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations, and Priorities*. Westminster, London: Royal United Services Institute, 2017. https://rusi.org/publication/occasional-papers/big-data-and-policing-assessment-law-enforcement-requirements.

Bean, Andrew H. "Swearing by New Technology: Strengthening the Fourth Amendment by Utilizing Modern Warrant Technology While Satisfying the Oath or Affirmation Clause." *Brigham Young University Law Review* 2014, no. 4, art. 5 (2014): 927–50.

Big Data Senior Steering Group. "The Federal Big Data Research and Development Strategic Plan." Digital Commons, University of Nebraska—Lincoln, 2016. http://digitalcommons.unl.edu/scholcom/20/.

Bloomberg. "OnStar: GM's Not-So-Secret Weapon." May 30, 2013. https://www.bloomberg.com/news/articles/2013-05-30/onstar-gms-not-so-secret-weapon.

Bottles, Kent, Edmon Begoli, and Brian Worley. "Understanding the Pros and Cons of Big Data Analytics." *Physician Executive* 40, no. 4 (August 2014): 6–12.

Brown, Derek. "Analytics in Government: Picking Up Speed or Caught in an Eddy? (Industry Perspective)." Government Technology, August 8, 2016. http://www.govtech.com/opinion/Analytics-in-Government-Picking-Up-Speed-Caught-in-Eddy-Industry-Perspective.html.

Btihaj, Ajana, "Augmented Borders: Big Data and the Ethics of Immigration Control." *Journal of Information, Communication & Ethics in Society* 13, no. 1 (2015): 58–78.

Busby, John C. "Probable Cause." Cornell Law School, Legal Information Institute, September 17, 2009. https://www.law.cornell.edu/wex/probable_cause.

Chen, Hong-Mei, Rick Kazman, and Serge Haziyev. "Agile Big Data Analytics Development: An Architecture-Centric Approach." In *System Sciences (HICSS), 2016 49th Hawaii International Conference On*, 5378–5387. Piscataway, NJ: IEEE, 2016. http://ieeexplore.ieee.org/abstract/document/7427853/.

Chen, Kun, Xin Li, and Huaiqing Wang. "On the Model Design of Integrated Intelligent Big Data Analytics Systems." *Industrial Management & Data Systems* 115, no. 9 (2015): 1666–1682.

Dale, Catherine. *The 2014 Quadrennial Defense Review (QDR) and Defense Strategy: Issues for Congress*. Washington, DC: Federation of American Scientists, 2014. http://search.proquest.com/docview/1641843659/51044E2789BD4713PQ/1.

Department of Homeland Security. "Written Testimony of U.S. Immigration and Customs Enforcement Homeland Security Investigations-Phoenix Special Agent in Charge Matthew Allen for a House Committee on Homeland Security, Subcommittee on Border and Maritime Security Field Hearing Titled "Stopping the Flow of Illicit Drugs in Arizona by Leveraging State, Local and Federal Information Sharing." May 21, 2012. https://www.dhs.gov/news/2012/05/21/written-testimony-us-immigration-and-customs-enforcement-house-homeland-security.

Desouza, Kevin. *Realizing the Promise of Big Data*. Washington, DC: IBM Center for the Business of Government, 2014. http://observgo.uquebec.ca/observgo/fichiers/26986_Realizing%20the%20Promise%20of%20Big%20Data.pdf.

Dragland, Ase. "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years." *ScienceDaily*, May 22, 2013. https://www.sciencedaily.com/releases/2013/05/130522085217.htm.

Eggers, William D. *Delivering on Digital: The Innovators and Technologies that Are Transforming Government.* New York: RosettaBooks, 2016.

El Gabry, Omar. "Database—Indexing, Transactions & Stored Procedures (Part 9)." *Medium* (blog), September 15, 2016. https://medium.com/omarelgabrys-blog/database-indexing-and-transactions-part-9-a24781d429f8.

Emerging Technologies Big Data Community of Interest. *HM Government Horizon Scanning Programme Emerging Technologies: Big Data*. London: HM Government, 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389095/Horizon_Scanning_-_Emerging_Technologies_Big_Data_report_1.pdf.

Findlaw. "Annotation 1—Fourth Amendment." Accessed August 26, 2017. http://constitution.findlaw.com/amendment4/annotation01.html.

Gamage, Pandula. "New Development: Leveraging 'Big Data' Analytics in the Public Sector." *Public Money & Management* 36, no. 5 (2016): 385–390.

Gandomi, Amir, and Murtaza Haider. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35, no. 2 (April 2015): 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gartner IT Glossary. “Big Data.” May 25, 2012. http://www.gartner.com/it-glossary/big-data/.

George, Gerard, Martine R. Haas, and Alex Pentland. “Big Data and Management.” *Academy of Management Journal* 57, no. 2 (2014): 321–326.

Google. “Privacy Impact Assessment for the FALCON Search & Analysis System.” Accessed October 12, 2017. https://www.google.com/search.

Hare, Nick, and Peter Coghill. “The Future of the Intelligence Analysis Task.” *Intelligence and National Security* 31, no. 6 (2016): 858–870.

Henry, Regina, and Santosh Venkatraman. “Big Data Analytics the Next Big Learning Opportunity.” *Journal of Management Information and Decision Sciences; Weaverville* 18, no. 2 (2015): 17–29.

Homeland Security Investigations. “Affidavit in Support of Application for Search Warrant, Case No. 16-6352MB.” July 16, 2017.

———. “Affidavit in Support of Application for Search Warrant, Case No. 16-8070MB.” September 12, 2017.

———. “Affidavit in Support of Application for Search Warrant, Case No. 16-8143MB.” April 28, 2017.

———. Affidavit in Support of Application for Search Warrant, Case No. 16-8144MB.” April 28, 2017.

———. “Affidavit in Support of Application for Search Warrant, Case No. 16-9434MB.” May 20, 2017.

IBM. *Data-Driven Healthcare Organizations Use Big Data Analytics for Big Gains*. Somers, NY: IBM Corporation, 2013. http://www-03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf.

Kitchin, Rob. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society* 1, no. 1 (2014): 1–12.

Konkel, Frank. “Report: Big Data, Cybersecurity Intrinsically Linked.” February 25, 2014. https://fcw.com/articles/2014/02/25/crit-read-big-data-cyber.aspx.

Kyle, David, and Rey Koslowski. *Global Human Smuggling: Comparative Perspectives*. Baltimore, MD: JHU Press, 2001.

Leonardo Law Offices. “Alien Smuggling & Harboring Illegal Aliens.” Accessed August 21, 2017. http://www.arizonadefenseattorney.net/alien-smuggling-harboring-illegal-aliens/.

Lv, Zhihan, Houbing Song, Pablo Basanta-Val, Anthony Steed, and Minho Jo. “Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics.” *IEEE Transactions on Industrial Informatics* 13, no. 4 (2017): 1891–1899.

Marr, Bernard. “Big Data: 20 Mind-Boggling Facts Everyone Must Read.” Forbes, September 30, 2015. http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/.

McNeely, Connie L., and Jong-on Hahm. “The Big (Data) Bang: Policy, Prospects, and Challenges.” *Review of Policy Research* 31, no. 4 (2014): 304–310.

NYC Business Integrity Commission and NYC Environmental Protection. *New York City Business Integrity Commission, Department of Environmental Protection, and Mayor’s Office of Policy and Strategic Planning Launch Comprehensive Strategy to Help Businesses Comply within Grease Disposal Regulations*. no. 71. New York City, NYC Business Integrity Commission and NYC Environmental Protection, 2012. http://www.nyc.gov/html/bic/downloads/pdf/pr/nyc_bic_dep_mayoroff_policy_10_18_12.pdf.

Office of the United States Attorneys. “1907. Title 8, U.S.C. 1324(a) Offenses.” Accessed August 21, 2017. https://www.justice.gov/usam/criminal-resource-manual-1907-title-8-usc-1324a-offenses.

Palem, Gopalakrishna. “Formulating an Executive Strategy for Big Data Analytics.” *Technology Innovation Management Review* 4, no. 3 (March 2014): 25–34.

PenLink, Ltd. “About Pen-Link.” Accessed October 5, 2017. https://www.penlink.com/About/tabid/123/Default.aspx.

Rivera, Janessa, and Rob van der Meulen. “Gartner Survey Reveals that 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years.” September 17, 2014. https://www.gartner.com/newsroom/id/2848718.

Roberts, Bryan, Edward Alden, and John Whitley. *Managing Illegal Immigration to the United States*. Washington, DC: Council on Foreign Relations, 2013. https://www.cfr.org/content/publications/attachments/Managing_Illegal_Immigration_report.pdf.

Rouse, Margaret. “What Is Data Scrubbing (Data Cleansing)?” SearchDataManagement. Accessed October 28, 2016. http://searchdatamanagement.techtarget.com/definition/data-scrubbing.

———. “What Is Script?” Tech Target Network. Accessed October 25, 2017. http://whatis.techtarget.com/definition/script.

Rubin, David, Kim Lynch, Jason Escaravage, and Hillary Lerner. “Harnessing Data for National Security.” *The SAIS Review of International Affairs* 34, no. 1 (2014): 121–28.

Schultz, Beth. “View/Review: Big Data in Big Companies.” *Baylor Business Review* 32, no. 1 (Fall 2013): 20–21.

Schut, Martijn C. “On Model Design for Simulation of Collective Intelligence.” *Information Sciences* 180, no. 1 (2010): 132–155.

Seddon, Jonathan, and Wendy L. Currie. “A Model for Unpacking Big Data Analytics in High-Frequency Trading.” *Journal of Business Research* 70, no. C (2017): 300–307.

Sivarajah, Uthayasankar, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. “Critical Analysis of Big Data Challenges and Analytical Methods.” *Journal of Business Research* 70 (2017): 263–286.

Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. “‘Big Data’: Big Gaps of Knowledge in the Field of Internet Science.” *International Journal of Internet Science* 7, no. 1 (2012): 1–5.

Sokol, Lisa, and Steve Chan. “Context-Based Analytics in a Big Data World: Better Decisions.” *IBM RedBooks Point-of-View Publication*, 2013.

Spencer, David. “Mexican Migrant Smuggling: A Cross-Border Cottage Industry.” *Journal of International Migration and Integration* 5, no. 3 (2004): 295–320.

U.S. Customs and Border Protection. “Stats and Summaries.” Accessed August 21, 2017. https://www.cbp.gov/newsroom/media-resources/stats?title=Border+Patrol.

U.S. Immigration and Customs Enforcement. “Homeland Security Investigations.” Accessed October 15, 2017. https://www.ice.gov/hsi.

———. “Human Smuggling.” Accessed August 18, 2017. https://www.ice.gov/human-smuggling.

Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. “Critical Analysis of Big Data Challenges and Analytical Methods.” *Journal of Business Research* 70 (August 10, 2016): 263–86.

Van Rijmenam, Mark. “Why the 3v’s Are Not Sufficient to Describe Big Data.” Big Data Startup, 2013. http://www. bigdata-startups.com/3vs-sufficient-describe-big-data.

Waterman, K. Krasnow, and Paula J. Bruening. “Big Data Analytics: Risks and Responsibilities.” *International Data Privacy Law* 4, no. 2 (May 2014): 89–95. http://dx.doi.org/10.1093/idpl/ipu002.

Weiss, Rick, and Lisa-Joy Zgorski. *Obama Administration Unveils “Big Data” Initiative: Announces $200 Million in New R&D Investments*. Washington, DC: Office of Science and Technology Policy, Executive Office of the President, 2012.

WOLA. “U.S.-Mexico Border Requires Evidence-Based Humanitarian Solutions, Not Border Walls.” October 27, 2016. https://www.wola.org/2016/10/wola-report-u-s-mexico-border-requires-evidence-based-humanitarian-solutions-not-border-walls/.

———. “WOLA’s Response to February CBP Southwest Border Migration Numbers.” March 9, 2017. https://www.wola.org/2017/03/wolas-response-february-cbp-southwest-border-migration-numbers/.

Yaqoob, Ibrar. “Information Fusion in Social Big Data: Foundations, State-of-the-Art, Applications, Challenges, and Future Research Directions.” *International Journal of Information Management*, April 19, 2016.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California