Small Data Surveillance v. Big Data Cybersurveillance

Margaret Hu*

* Assistant Professor of Law, Washington and Lee School of Law. I would like to extend my deep gratitude to those who graciously offered comments on this draft, or who offered perspectives and expertise on this research through our thoughtful discussions: Sahar Aziz, John Bagby, Judge James Baker, Jack Balkin, Kate Bartlett, Lawrence Baxter, Jody Blanke, Joseph Blocher, danah boyd, Andy Briggs, Dru Brenner-Beck, Jamie Boyle, Rachel Brewster, Guy Charles, Bobby Chesney, Andrew Christensen, Danielle Citron, Geoff Corn, Jennifer Daskal, Nora Demleitner, Charlie Dunlap, Josh Fairfield, Nita Farahany, Mark Graber, David Gray, Woody Hartzog, Janine Hiller, Trina Jones, Orin Kerr, J.J. Kidder, Corinna Lain, Sandy Levinson, Rachel Levinson-Waldman, Maggie Lemos, Erik Luna, Peter Margulies, Russ Miller, Steve Miskinis, Brian Murchison, Richard Myers, Jeff Powell, Jed Purdy, Angie Raymond, David Robinson, Mark Rush, Mark Seidenfeld, Andrew Selbst, Victoria Shannon, Ben Spencer, Dan Tichenor, Steve Vladeck, Russ Weaver, John Weistart, Ben Wittes, Ernie Young, and apologies to those whom I may have inadvertently failed to acknowledge. In addition, this research benefited greatly from the discussions generated from the 2014 Pepperdine Law Review Symposium, "The Future of National Security Law"; 2015 AALS National Conference, National Security Law Section, Call for Papers, National Security Surveillance Panel; Washington and Lee University School of Law, 2015 Law Review Symposium, "Cybersurveillance in the Post-Snowden Age"; Constitutional Law Schmooze, King Carey University of Maryland; 2015 Sorbonne International Symposium, "Freedom of Information and Governmental Transparency in the Open Government Era" in Paris, France; University of Indiana-Bloomington, 2015 "Law and Big Data Ethics" Research Colloquium; 2014 Southeastern Association of Law Schools' Annual Conference, Post-Snowden Group Discussion; 2014 Association of American Law Schools' Annual Meeting, Administrative Law Section, "New Voices in Administrative Law" Workshop; American University, Washington College of Law, 2014 Faculty Workshop; University of Richmond School of Law 2014–15 Faculty Colloquium Series; mass surveillance panel hosted by Texas A&M School of Law, 2014 Law Review Symposium, "New Technologies, Old Law"; Charleston School of Law, Federal Courts Law Journal, "Technology in the Criminal Justice System" Symposium, mass surveillance panel; University of Freiburg, KORSE Centre for Security and Society, "Privacy and Power: Transatlantic Dialogue in the Shadow of the NSA-Affair" Symposium in Freiburg, Germany; "Transnational Dialogue on Surveillance Methods," hosted by Max Planck Institute in Freiburg, Germany; 2013 Duke Law School Summer Faculty Workshop; 2013 "Politics of Surveillance" Symposium, hosted by the Wayne Morse Center for Law and Politics at the University of Oregon School of Law; and 2011 Duke-UNC Junior Faculty Workshop. Many thanks to the research assistance of Lauren Bugg, Russell Caleb Chaplain, Jessica Chi, Katherine Dickinson, Joshua Hockman, Andrew House, Cadman Kiker, Kirby Kreider, Oscar Molina, Markus Murden, Madeline Morcelle, Kelsey Peregoy, Joe Silver, and Cole Wilson. All errors and omissions are my own.

Abstract

This Article highlights some of the critical distinctions between small data surveillance and big data cybersurveillance as methods of intelligence gathering. Specifically, in the intelligence context, it appears that "collect-it-all" tools in a big data world can now potentially facilitate the construction, by the intelligence community, of other individuals' digital avatars. The digital avatar can be understood as a virtual representation of our digital selves and may serve as a potential proxy for an actual person. This construction may be enabled through processes such as the data fusion of biometric and biographic data, or the digital data fusion of the 24/7 surveillance of the body and the 360° surveillance of the biography. Further, data science logic and reasoning, and big data policy rationales, appear to be driving the expansion of these emerging methods. Consequently, I suggest that an inquiry into the scientific validity of the data science that informs big data cybersurveillance and mass dataveillance is appropriate.

As a topic of academic inquiry, thus, I argue in favor of a science-driven approach to the interrogation of rapidly evolving bulk metadata and mass data surveillance methods that increasingly rely upon data science and big data's algorithmic, analytic, and integrative tools. In Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), the Supreme Court required scientific validity determinations prior to the introduction of scientific expert testimony or evidence at trial. I conclude that to the extent that covert intelligence gathering relies upon data science, a* Daubert*-type inquiry is helpful in conceptualizing the proper analytical structure necessary for the assessment and oversight of these emerging mass surveillance methods.*

I. INTRODUCTION

The disclosures of former National Security Agency (NSA) contractor Edward Snowden[1] underscore why, as a matter of statutory and

1. *See, e.g.*, GLENN GREENWALD, NO PLACE TO HIDE: EDWARD SNOWDEN, THE NSA, AND THE U.S. SURVEILLANCE STATE (2014) (discussing in detail the history of the Snowden disclosures). Scholars and experts have focused careful attention on the legal implications of the mass surveillance activities of the NSA and the intelligence community in work both preceding and following the disclosures of former NSA contractor Edward Snowden. *See, e.g.*, Laura K. Donohue, *Section 702 and the Collection of International Telephone and Internet Content,* 38 HARV. J.L. & PUB. POL'Y 117 (2015); Margo Schlanger, *Intelligence Realism and the National Security Agency's Civil Liberties Gap*, 6 HARV. NAT'L SEC. J. 112 (2015); Laura K. Donohue, *Bulk Metadata Collection: Statutory and Constitutional Considerations,* 37 HARV. J.L. & PUB. POL'Y 757 (2014);

constitutional inquiry, it is important to focus attention on the critical distinctions between small data[2] surveillance[3] and big data[4]

Orin S. Kerr, *A Rule of Lenity for National Security Surveillance Law*, 100 VA. L. REV. 1513 (2014); Peter Margulies, *Dynamic Surveillance: Evolving Procedures in Metadata and Content Collection After Snowden*, 66 HASTINGS L.J. 1 (2014); Stephen I. Vladeck, *Big Data Before and After Snowden*, 7 J. NAT'L SEC. L. & POL'Y 333 (2014); Stephen I. Vladeck, *Standing and Secret Surveillance*, 10 J.L. & POL'Y INFO. SOC'Y 551 (2014); Omer Tene, *A New Harm Matrix for Cybersecurity Surveillance*, 12 COLO. TECH. L.J. 391 (2014); Christopher Slobogin, *Panvasive Surveillance, Political Process Theory, and the Nondelegation Doctrine*, 102 GEO. L.J. 1721 (2014); Nathan A. Sales, *Domesticating Programmatic Surveillance: Some Thoughts on the NSA Controversy*, 10 I/S: J.L. & POL'Y INFO. SOC'Y 523 (2014); John Yoo, *The Legality of the National Security Agency's Bulk Data Surveillance Programs*, 37 HARV. J.L. & PUB. POL'Y 901 (2014); Margot E. Kaminski & Shane Witnov, *The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech*, 49 U. RICH. L. REV. 465 (2015); Christopher Slobogin, *Cause to Believe What? The Importance of Defining A Search's Object-or, How the ABA Would Analyze the NSA Metadata Surveillance Program*, 66 OKLA. L. REV. 725 (2014); Anjali S. Dalal, *Shadow Administrative Constitutionalism and the Creation of Surveillance Culture*, 2014 MICH. ST. L. REV. 61 (2014); Patrick Toomey & Brett Max Kaufman, *The Notice Paradox: Secret Surveillance, Criminal Defendants, & the Right to Notice*, 54 SANTA CLARA L. REV. 843 (2014); Paul Ohm, *Electronic Surveillance Law and the Intra-Agency Separation of Powers*, 47 U.S.F. L. REV. 269 (2012). Several important works have been published in recent years, shedding light on mass surveillance technologies, and the policy and programmatic framework of cybersurveillance and covert intelligence gathering. *See, e.g.*, JULIA ANGWIN, DRAGNET NATION: A QUEST FOR PRIVACY, SECURITY, AND FREEDOM IN A WORLD OF RELENTLESS SURVEILLANCE 17–18 (2014); SHANE HARRIS, @WAR: THE RISE OF THE MILITARY-INTERNET COMPLEX (2014); DANA PRIEST & WILLIAM M. ARKIN, TOP SECRET AMERICA: THE RISE OF THE NEW AMERICAN SECURITY STATE (2011); SHANE HARRIS, THE WATCHERS (2010); ROBERT O'HARROW, JR., NO PLACE TO HIDE (2006); JEFFREY ROSEN, THE NAKED CROWD: RECLAIMING SECURITY AND FREEDOM IN AN ANXIOUS AGE (2004).

2. "'Small data,' like 'big data,' has no set definition." Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 329 n.6 (2015). "Small data" has been described in the following way: "Generally, small data is thought of as solving discrete questions with limited and structured data, and the data are generally controlled by one institution." *Id.* (citing JULES J. BERMAN, PRINCIPLES OF BIG DATA: PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION 1–2 (2013)). In many important recent works, scholars and experts have observed the transformational nature of emerging technologies of the Information Society—such as the Internet, digital culture, technological innovations in surveillance capacities—and the legal and privacy implications of such transformative technological developments. *See, e.g.*, JOHN GILLIOM & TORIN MONAHAN, SUPERVISION (2013); SIMON CHESTERMAN, ONE NATION UNDER SURVEILLANCE (2011); CONSTITUTION 3.0: FREEDOM AND TECHNOLOGICAL CHANGE (Jeffrey Rosen & Benjamin Wittes eds., 2011); SUSAN LANDAU, SURVEILLANCE OR SECURITY: THE RISKS POSED BY NEW WIRETAPPING TECHNOLOGIES (2010); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET—AND HOW TO STOP IT (2008); DAVID LYON, SURVEILLANCE STUDIES: AN OVERVIEW (2007); LAWRENCE LESSIG, CODE VERSION 2.0 (2006); JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? (2006) MARK POSTER, INFORMATION PLEASE: CULTURE AND POLITICS IN THE AGE OF DIGITAL MACHINES (2006); A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461 (2000).

3. The term "small data surveillance" is neither widely used nor, to the best of my knowledge,

cybersurveillance.[5] The Snowden disclosures reveal that, in addition to the traditional communications that the intelligence community once sought in a small data world, organizations such as the NSA are increasingly exploiting newly available mass data surveillance, or dataveillance,[6] and cybersurveillance tools[7] in a big data world.[8] Specifically, from the

officially defined. In this Article, the term is used as a way to mark a contrast between traditional intelligence gathering methods (i.e., "small data surveillance") and newly emerging intelligence methods that are digital data-driven, dependent upon supercomputing capacities, and capitalize on big data phenomena and tools (i.e., "big data cybersurveillance"). As technology has transformed the Information Society, surveillance methods have transformed as well. David Lyon, *Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique*, BIG DATA & SOC. 2 (2014) ("[A]s political-economic and socio-technological circumstances change, so surveillance also undergoes alteration, sometimes transformation."). Historically, it appears that in a small data world, intelligence gathering methods have relied upon human intelligence, including human sensory perception analysis, and other communication gathering and analytic methods that have depended upon human judgment and human decisionmaking; traditional evidence based upon analog data and paper-based files; traditional intelligence collection methods, such as traditional signals intelligence and other traditional communications interception; and other data analytic tools that have centered upon traditional research approaches, such as hypothesis-driven methods. *See, e.g.*, ROBERT M. CLARK, INTELLIGENCE COLLECTION (2014); ROBERT WALLACE & H. KEITH MELTON, WITH HENRY R. SCHLESINGER, AND FOREWARD BY GEORGE TENET, SPYCRAFT: THE SECRET HISTORY OF THE CIA'S SPYTECHS FROM COMMUNISM TO AL-QAEDA (2008).

4. "Big data" is difficult to define, as it is a newly evolving field and the technologies that it encompasses are evolving rapidly as well. *See* discussion *infra* Part II.A.1 ("What is Big Data?"). *See generally infra* Parts II–III. Multiple authors have addressed the characteristics of "big data" and the challenges posed by big data technologies. *See, e.g.*, VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013); BERMAN, *supra* note 2; PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT (Julia Lane, Victoria Stodden, Stefan Bender & Helen Nissenbaum eds., 2014); ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES (2014).

5. *See, e.g.*, LESSIG, *supra* note 2, at 209 (describing cybersurveillance or "digital surveillance" as "the process by which some form of human activity is analyzed by a computer according to some specified rule. . . . [T]he critical feature in each [case of surveillance] is that a computer is sorting data for some follow-up review by some human.").

6. Roger Clarke is attributed with introducing the term "dataveillance." *See* Roger A. Clarke, *Information Technology and Dataveillance*, 31 COMM. ACM 498 (1988). Clarke describes dataveillance as the systematic monitoring or investigation of people's actions, activities, or communications through the application of information technology. *Id.*; *see also* LYON, *supra* note 2, at 16 ("Being much cheaper than direct physical or electronic surveillance [dataveillance] enables the watching of more people or populations, because economic constraints to surveillance are reduced. Dataveillance also automates surveillance. Classically, government bureaucracies have been most interested in gathering such data . . . ."); MARTIN KUHN, FEDERAL DATAVEILLANCE: IMPLICATIONS FOR CONSTITUTIONAL PRIVACY PROTECTIONS (2007) (examining constitutional implications of "knowledge discovery in databases" (KDD applications) through dataveillance).

7. The Snowden disclosures have included multiple high-profile revelations on newly emerging

disclosures and other publicly available information, it appears that in the intelligence context, "collect-it-all"[9] tools in a big data world[10] can now

data surveillance, or dataveillance tools, and cybersurveillance methods, and information specific to their implementation. *See, e.g.*, Glenn Greenwald, *NSA Collecting Phone Records of Millions of Verizon Customers Daily*, GUARDIAN (June 5, 2013), http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order; Barton Gellman & Laura Poitras, *U.S., British Intelligence Mining Data from Nine U.S. Internet Companies in Broad Secret Program,* WASH. POST (June 6, 2013), http://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html*;* T.C. Sottek & Josh Kopstein, *Everything You Need to Know About PRISM,* VERGE, (July 13, 2013), http://www.theverge.com/2013/7/17/4517480/nsa-spying-prism-surveillance-cheat-sheet; Scott Shane, *No Morsel Too Minuscule For All-Consuming N.S.A., From Spying on Leader of U.N. to Tracking Drug Deals, an Ethos of 'Why Not'?*, N.Y. TIMES (Nov. 3, 2013) A1, A10, *available at* http://www.nytimes.com/2013/11/03/world/no-morsel-too-minuscule-for-all-consuming-nsa.html?_r=0; Ellen Nakashima & Barton Gellman, *Court Gave NSA Broad Leeway in Surveillance, Documents Show*, WASH. POST (June 30, 2014), http://www.washingtonpost.com/world/national-security/court-gave-nsa-broad-leeway-in-surveillance-documents-show/2014/06/30/32b872ec-fae4-11e3-8176-f2c941cf35f1_story.html.

8. Several scholars have begun to use the term "big data surveillance" to describe how surveillance methods are evolving in light of the emerging pervasiveness of big data technologies. *See, e.g.*, Lyon, *Surveillance, Snowden, and Big Data*, *supra* note 3, at 4 ("The Big Data/surveillance link was recognized by US President Obama on 17 January 2014, when he called for a 'comprehensive review of Big Data and privacy' following the Snowden leaks." (citation omitted)); Mark Andrejevic, *Surveillance in the Big Data Era*, *in* EMERGING PERVASIVE INFORMATION AND COMMUNICATION TECHNOLOGIES (PICT): ETHICAL CHALLENGES, OPPORTUNITIES, AND SAFEGUARDS 56 (Kenneth D. Pimple ed., 2014) ("[I]n the era of 'big data' surveillance, the imperative is to monitor the population as a whole: otherwise it is harder to consistently and reliably discern useful patterns."). Other scholars and experts have documented how the NSA, CIA, and other intelligence organizations capitalize on technological innovation in the evolution and expansion of intelligence gathering tools and methods. *See, e.g.*, JAMES BAMFORD, THE SHADOW FACTORY: THE ULTRA-SECRET NSA FROM 9/11 TO THE EAVESDROPPING ON AMERICA (2008); JAMES BAMFORD, THE PUZZLE PALACE: INSIDE THE NATIONAL SECURITY AGENCY AMERICA'S MOST SECRET INTELLIGENCE ORGANIZATION (1982); William C. Banks, *Programmatic Surveillance and FISA: Of Needles in Haystacks*, 88 TEX. L. REV. 1633 (2010); Peter P. Swire, *Privacy and Information Sharing in the War on Terrorism*, 51 VILL. L. REV. 951 (2006).

9. GREENWALD, *supra* note 1, at 97 (citing NSA slide from Snowden disclosures titled, "New Collection Posture," quoting NSA data collection procedure as "Collect it All"), *available at* http://glenngreenwald.net/pdf/NoPlaceToHide-Documents-Compressed.pdf; *see also* David Cole, *'No Place to Hide' by Glenn Greenwald, on the NSA's Sweeping Efforts to 'Know it All'*, WASH. POST (May 12, 2014) ("In one remarkable [NSA] slide presented at a 2011 meeting of five nations' intelligence agencies and revealed here for the first time, the NSA described its "collection posture" as "Collect it All," "Process it All," "Exploit it All," "Partner it All," "Sniff it All" and, ultimately, "Know it All.").

10. The legal, science, social, and other consequences of what has been termed the "big data revolution" have been an topic of intense academic inquiry. *See, e.g.*, Neil M. Richards & Jonathan H. King; *Three Paradoxes of Big Data*; 66 STAN. L. REV. ONLINE 41 (2013); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. (forthcoming 2016); Kate

potentially facilitate the construction of digital avatars,[11] or the virtual representation[12] of our digital selves.[13] Significant legal and constitutional

---

Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63 (2012). Some scholars have focused particularly on the algorithmic-driven decisionmaking consequences of emerging big data technologies. *See, e.g.*, Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); FRANK PASQUALE, THE BLACK BOX SOCIETY (2015). Other experts have focused on the data mining and predictive analytic capacities of big data tools. *See, e.g.*, STEVEN FINLAY, PREDICTIVE ANALYTICS, DATA MINING AND BIG DATA: MYTHS, MISCONCEPTIONS, AND METHODS (2014); ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE (2013); NATE SILVER, THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL—BUT SOME DON'T (2012); Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435 (2008); Christopher Slobogin, *Government Data Mining and the Fourth Amendment*, 75 U. CHI. L. REV. 317 (2008); Daniel J. Solove, *Data Mining and the Security-Liberty Debate*, 75 U. CHI. L. REV. 343 (2008). At the dawn of the big data revolution, scholars are now actively interrogating the implications of government-led big data uses by the government and law enforcement. *See, e.g.*, Joshua A.T. Fairfield & Erik Luna, *Digital Innocence*, 99 CORNELL L. REV. 981 (2014); David Gray & Danielle Citron, *The Right to Quantitative Privacy*, 98 MINN. L. REV. 62 (2013); Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934 (2013).

11. *See infra* Part IV.A. (discussing why the term "digital avatar" may best describe this phenomenon). The term "digital avatar" is used often in the video gaming context, and most commonly refers to a digitally constructed representation of the computing user or, in some instance, the representation of the user's alter ego or character. *See, e.g.*, Hart v. Electronic Arts, Inc., 717 F.3d 141 (3d Cir. 2012). In *Hart v. Electronic Arts, Inc.*, for example, a class action suit of college athletes alleged that their digital avatars and likeness had been unlawfully appropriated for profit by the video game developer, Electronic Arts, Inc. *See id.*

12. The introduction of virtual reality and virtual worlds has raised increasingly complicated legal questions. For example, in *Brown v. Entertainment Merchants Ass'n*, 131 S.Ct. 2729 (2011), the Supreme Court considered the First Amendment implications of expressive speech of video games. It explained,

> Like the protected books, plays, and movies that preceded them, video games communicate ideas—and even social messages—through many familiar literary devices (such as characters, dialogue, plot, and music) and through features distinctive to the medium (such as the player's interaction with the virtual world). That suffices to confer First Amendment protection.

Id. at 2733; *see also* Joshua A.T. Fairfield, *Mixed Reality: How the Laws of Virtual Worlds Govern Everyday Life*, 27 BERKELEY TECH. L.J. 55, 71 (2012); Joshua A.T. Fairfield, *Virtual Parentalism*, 66 WASH. & LEE L. REV. 1215 (2009); Joshua A.T. Fairfield, *Escape Into the Panopticon: Virtual Worlds and the Surveillance Society*, 118 YALE L.J. POCKET PART 131 (2009); Marc Jonathan Blitz, *The Freedom of 3D Thought: The First Amendment in Virtual Reality*, 30 CARDOZO L. REV. 1141 (2008). Increasingly, scholars are interrogating the legal implications of self-representations and digital avatar representations in virtual worlds. *See, e.g.*, Llewellyn Joseph Gibbons, *Law and the Emotive Avatar*, 11 VAND. J. ENT. & TECH. L. 899 (2009).

13. *See infra* Part IV.A; *see also* David Cole, *Is Privacy Obsolete? Thanks to the Revolution in Digital Technology, Privacy is About to Go the Way of the Eight-Track Player*, THE NATION (Apr. 6,

consequences attach to these rapidly evolving technologies of mass surveillance and bulk data collection, and their application.[14]

To help illustrate this historically significant transition from small data surveillance methods to big data cybersurveillance methods,[15] it is instructive to focus on one sentence extracted from one alleged document released from the Snowden archives. In a *New York Times* article by James Risen and Laura Poitras published on June 1, 2014, titled *NSA Collecting*

2015), http://www.thenation.com/article/198505/privacy-20-surveillance-digital-age# ("Digital technology has exponentially expanded the government's ability to construct intimate portraits of any particular individual by collecting all sorts of disparate data and combining and analyzing them for revealing patterns."); Frank Gillett, *How Will You Manage Your Digital Self?*, INFORMATIONWEEK.COM, (Oct. 30, 2013), http://www.informationweek.com/software/social/how-will-you-manage-your-digital-self/d/d-id/1112130 ("The digital self is not just your work and personal computer files. It includes all of the complex and varied digital information that you and the organizations you deal with generate.").

14. Several scholars have noted how transformative technological shifts have also transformed methods of governance and surveillance as a tool of governance. *See, e.g.*, Jack M. Balkin, *Old-School/New-School Speech* Regulation, 127 HARV. L. REV. 2296, 2297 (2014) ("The digital era is different. Governments can target for control or surveillance many different aspects of the digital infrastructure that people use to communicate: telecommunications and broadband companies, web hosting services, domain name registrars, search engines, social media platforms, payment systems, and advertisers."); Jack M. Balkin, *The Constitution in the National Surveillance State*, 93 MINN. L. REV. 1 (2008) [hereinafter Balkin, *National Surveillance State*]; Jack M. Balkin & Sanford Levinson, *The Processes of Constitutional Change: From Partisan Entrenchment to the National Surveillance State*, 75 FORDHAM L. REV. 489 (2006); Lior Jacob Strahilevitz, *Signaling Exhaustion and Perfect Exclusion*, 10 J. ON TELECOMM. & HIGH TECH. L. 321 (2012); David Lyon, *Biometrics, Identification and Surveillance*, 22 BIOETHICS 499 (2008); Erin Murphy, *Paradigms of Restraint*, 57 DUKE L.J. 1321 (2008).

15. Following the Snowden disclosures, at least one expert has asserted that the NSA is attempting to merge big data tools with small data tools. *See, e.g.*, Kate Crawford, *The Anxieties of Big Data*, THE NEW INQUIRY (May 30, 2014), *available at* http://thenewinquiry.com/essays/the-anxieties-of-big-data/ ("[A Squeaky Dolphin PowerPoint slide from the Snowden disclosures] outlines an expansionist program to bring big data together with the more traditional approaches of the social and humanistic sciences: the worlds of small data. . . . [A]nd it is all about supplementing [big] data analysis with broader sociocultural tools from anthropology, sociology, political science, biology, history, psychology, and economics."). Scholars and experts have also juxtaposed small data policing and surveillance practices with big data policing and surveillance practices as a way to deepen the legal and constitutional discourse. *See, e.g.*, Ferguson, *supra* note 2, at 329 (discussing in the context of predictive policing practices: "[W]hat happens if this small data suspicion [suspicion that is 'individualized to a particular person at a particular place'] is replaced by 'big data' suspicion?"); Toomey & Kaufman, *supra* note 1, at 847 (describing a "notice paradox" whereby in a small data surveillance world "notice was all but automatic in most cases . . . [because] searches were confined to a physical world"; however, with big data surveillance methods and "the rise of electronic surveillance conducted remotely and surreptitiously[,]" the authors observe that "the government has achieved an unprecedented measure of control over when, and to whom, notice [of the surveillance] is given.").

Millions of Faces From Web Images,[16] the authors discuss NSA documents from the Snowden disclosures that focus on biometric data collection[17] (e.g., scanned fingerprints, irises, digital photos and facial recognition technology, and DNA).[18] The authors note that an alleged 2010 NSA document explains: "'It's not just the traditional communications we're after: It's taking a full-arsenal approach that digitally exploits the clues a target leaves behind in their regular activities on the net to compile biographic and biometric information' that can help 'implement precision targeting.'"[19]

16. James Risen & Laura Poitras, *N.S.A. Collecting Millions of Faces from Web Images*, N.Y. TIMES, June 1, 2014, at A1, *available at* http://www.nytimes.com/2014/06/01/us/nsa-collecting-millions-of-faces-from-web-images.html.

17. Whether biometric data is defined in the disclosures is not mentioned; however, the authors specifically reference NSA's interest in the specific types of biometric data: "facial recognition technology" and "facial images [from digital photos, videoconferences, etc.], fingerprints and other identifiers." *Id*. ("While once focused on written and oral communications, the N.S.A. now considers facial images, fingerprints and other identifiers just as important to its mission of tracking suspected terrorists and other intelligence targets, the documents show.").

18. *See, e.g.*, Margaret Hu, *Biometric ID Cybersurveillance*, 88 IND. L.J. 1475 (2013). Biometrics is "[t]he science of automatic identification or identity verification of individuals using physiological or behavioral characteristics." JOHN R. VACCA, BIOMETRIC TECHNOLOGIES AND VERIFICATION SYSTEMS 589 (2007). Numerous scholars and experts have explored the science and application of biometrics and the consequences of this emerging technology. *See, e.g.*, Laura K. Donohue, *Technological Leap, Statutory Gap, and Constitutional Abyss: Remote Biometric Identification Comes of Age*, 97 MINN. L. REV. 407 (2012); JENNIFER LYNCH, FROM FINGERPRINTS TO DNA: BIOMETRIC DATA COLLECTION IN U.S. IMMIGRANT COMMUNITIES AND BEYOND (2012); A. MICHAEL FROOMKIN & JONATHAN WEINBERG, CHIEF JUSTICE EARL WARREN INST. ON LAW & SOC. POLICY, HARD TO BELIEVE: THE HIGH COST OF A BIOMETRIC IDENTITY CARD (2012), *available at* http://www.law.berkeley.edu/files/Believe_Report_Final.pdf; KELLY A. GATES, OUR BIOMETRIC FUTURE: FACIAL RECOGNITION TECHNOLOGY AND THE CULTURE OF SURVEILLANCE (2011); ANIL K. JAIN, ARUN A. ROSS, KARTHIK NANDAKUMAR, INTRODUCTION TO BIOMETRICS (2011); SHOSHANA AMIELLE MAGNET, WHEN BIOMETRICS FAIL: GENDER, RACE, AND THE TECHNOLOGY OF IDENTITY (2011); BIOMETRIC RECOGNITION: CHALLENGES AND OPPORTUNITIES (Joseph N. Pato & Lynette I. Millett eds., 2010) [hereinafter BIOMETRIC RECOGNITION]; DAVID LYON, SURVEILLANCE STUDIES: AN OVERVIEW 118–36 (2007); VACCA, *supra*; ROBERT O'HARROW, JR., NO PLACE TO HIDE 157–89 (2005); Robin Feldman, *Considerations on the Emerging Implementation of Biometric Technology*, 25 HASTINGS COMM. & ENT. L.J. 653 (2003); U.S. GEN. ACCOUNTING OFFICE, GAO-03-174, TECHNOLOGY ASSESSMENT: USING BIOMETRICS FOR BORDER SECURITY (2002) [hereinafter GAO TECHNOLOGY ASSESSMENT], *available at* http://www.gao.gov/assets/160/157313.pdf; SIMSON GARFINKEL, DATABASE NATION: THE DEATH OF PRIVACY IN THE 21ST CENTURY 37–67 (2000).

19. Risen & Poitras, *supra* note 16. The use of the term "targeting" in this alleged 2010 NSA document from the Snowden disclosures does not appear to be defined. However, the term "targeting" in the defense and intelligence context has been defined as "[t]he process of selecting and prioritizing targets and matching the appropriate response to them, considering commander's objectives, operational requirements, capabilities, and limitations." *See* U.S. DEPT. OF DEFENSE,

As this Article will attempt to explain, in the intelligence context, it appears that big data cybersurveillance and mass dataveillance tools may now risk the conflation of the digitally constructed virtual representation of a "target" with an actual person. Viewed through the lens of this risk, it appears that the reference to the "target" in the alleged 2010 NSA document above may be more appropriately and descriptively characterized as a digital avatar in that it appears that the "target" may be a product of data fusion,[20] or an amalgamation of data,[21] (e.g., "digitally exploit[ing] the clues a target leaves behind in their regular activities on the [Inter]net to compile biographic and biometric information"),[22] and may not represent an actual person (e.g., "signature strike" where the identity of the target of a drone strike may be unknown).[23]

OFFICE OF COUNTERINTELLIGENCE, DEFENSE CI & HUMINT CENTER, DEFENSE INTELLIGENCE AGENCY, GLOSSARY (UNCLASSIFIED), TERMS & DEFINITIONS OF INTEREST FOR DOD COUNTERINTELLIGENCE PROFESSIONALS, at GL-167 (2011), *available at* http://fas.org/irp/eprint/ci-glossary.pdf; *see also id*. (defining the counterintelligence community's use of the word "target" as "1) An entity or object considered for possible engagement or other action; 2) in intelligence usage, a country, area, installation, agency, or person against which intelligence operations are directed; 3) an area designated and numbered for future firing; and 4) in gunfire support usage, an impact burst that hits the target.").

20. In the intelligence context, "fusion" or "data fusion" has been described as "the collection of information from myriad sources to be organized and analyzed for a fuller picture of terrorist or other threats." PRIEST & ARKIN, *supra* note 1, at 92. In the consumer context, "data fusion" has been defined in the following way: "Data fusion occurs when data from different sources are brought into contact and new facts emerge[.]" PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY ("PCAST"), BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE (May 2014) [hereinafter PCAST REPORT], *available at* https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf; *see infra* Parts III–IV. Several scholars and experts have explored the legal and surveillance implications of data fusion centers that have been created by the government, particularly after the terrorist attacks of September 11, 2001. *See, e.g.*, PRIEST & ARKIN, *supra* note 1, at 92–93; Slobogin, *Panvasive Surveillance*, *supra* note 1; Danielle Keats Citron & Frank Pasquale, *Network Accountability for the Domestic Intelligence Apparatus*, 62 HASTINGS L.J. 1441 (2011).

21. *See infra* Parts III–IV.

22. Risen & Poitras, *supra* note 16.

23. "Signature strikes" are "a controversial [targeted killing] practice" where the "defining characteristics associated with terrorist activity [of the targets are identified], but whose identities aren't necessarily known." DANIEL KLAIDMAN, KILL OR CAPTURE: THE WAR ON TERROR AND THE SOUL OF THE OBAMA PRESIDENCY 41 (2012). After the Snowden disclosures, several media reports have indicated that the revelations establish "collaboration between the CIA and NSA" in the targeted killing program. Greg Miller, Julie Tate & Barton Gellman, *Documents Reveal NSA's Massive Involvement in Targeted Killing Program*, WASH. POST (Oct. 16, 2013), *available at* http://www.washingtonpost.com/world/national-security/documents-reveal-nsas-extensive-involvement-in-targeted-killing-program/2013/10/16/29775278-3674-11e3-8a0e-4e2cf80831fc_print.html ("[T]he

Nonetheless, it appears that the legal or other consequences that may flow from the big data cybersurveillance or mass dataveillance methods are suffered by the person associated with the suspicious digital data and, potentially, conflated with the guilty digital avatar or the digital avatar's technological surrogate (e.g., a smartphone).[24] In other words, in a big data world, the intelligence community may view the digital avatar or technological surrogate as a proxy for the actual person targeted.[25]

From the disclosures, it appears that the "full-arsenal approach" to newly emerging mass surveillance methods employs data science[26] to

---

[Snowden] documents provide the most detailed account of the intricate collaboration between the CIA and the NSA in the drone campaign."); Jeremy Scahill & Glenn Greenwald, *The NSA's Secret Role in the U.S. Assassination Program*, INTERCEPT (Feb. 9, 2014), *available at* https://firstlook.org/theintercept/2014/02/10/the-nsas-secret-role/ ("According to a former drone operator for the military's Joint Special Operations Command (JSOC) who also worked with the NSA, the agency often identifies targets based on controversial metadata analysis and cell-phone tracking technologies."). Prior to the Snowden disclosures, media reports indicated that drone strikes could be authorized based upon "patterns of suspicious behavior." Greg Miller, *Broader Drone Tactics Sought*, WASH. POST, Apr. 19, 2012, at A1 ("The CIA is seeking authority to expand its covert drone campaign in Yemen by launching strikes against terrorism suspects even when it does not know the identities of those who could be killed, U.S. officials said. Securing permission to use these 'signature strikes' would allow the agency to hit targets based solely on intelligence indicating patterns of suspicious behavior[.]"). Due to the covert nature of the targeted killing program, however, limited information is available on precisely what intelligence may inform drone strikes and signature strikes. *See generally* DAVID E. SANGER, CONFRONT AND CONCEAL: OBAMA'S SECRET WARS AND SURPRISING USE OF AMERICAN POWER 241–70 (2012) (describing use of drones and targeted killing strategy in the "war on terror"); JEREMY SCAHILL, DIRTY WARS (2013); Kevin Jon Heller, *'One Hell of a Killing Machine': Signature Strikes and International Law*, 11 J. INT'L CRIM. JUST. 89, 89 (2013); Kenneth Anderson, *The Secret "Kill List" and the President*, 3 J.L.: PERIODICAL LABORATORY OF LEG. SCHOLARSHIP 93 (2013).

24. For example, according to one media report, a drone operator explained that drone strikes target not a suspicious person necessarily, but, rather may target a digital avatar proxy—suspicious phones. Scahill & Greenwald, *supra* note 23 ("We're not going after people—we're going after their phones, in the hopes that the person on the other end of that missile is the bad guy."); *see also* Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. (forthcoming 2015).

25. Scahill & Greenwald, *supra* note 23 ("According to a former drone operator for the military's Joint Special Operations Command (JSOC) who also worked with the NSA, the agency often identifies targets based on controversial metadata analysis and cell-phone tracking technologies.").

26. Like the terms "small data" and "big data" that are not yet clearly defined as of yet, the terms "data science," "data-driven science" and "big data science" have no set, agreed-upon definition. Generally, however, "[i]n contrast to new forms of empiricism, data-driven science seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon." Rob Kitchin, *Big Data, New Epistemologies and Paradigm Shifts*, BIG DATA & SOC'Y J., Apr.–June 2014, at 1. Data science, thus, has been used to describe a new field of study and academic research that is

determine the data-driven suspicion and guilt of digital avatars.[27] Properly assessing the scientific validity of this approach, therefore, becomes central to the legal inquiry. Consequently, in this Article, I contend that the perceived capacities or presumed intelligence value of big data

---

dependent upon big data technological developments and tools. According to a National Science Foundation Solicitation, the term "big data science & engineering" appears to include the study of the:

> core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure[.]

NAT'L SCI. FOUND., SOLICITATION 12-499, CORE TECHNIQUIES AND TECHNOLOGIES FOR ADVANCING BIG DATA SCIENCE & ENGINEERING (2012), *available at* http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm.

27. Multiple scholars are carefully examining the legal implications of targeted killing policies and drone strikes as a linchpin of U.S. counterterrorism policy. *See, e.g.*, Martin S. Flaherty, *The Constitution Follows the Drone: Targeted Killings, Legal Constraints, and Judicial Safeguards*, 38 HARV. J.L. & PUB. POL'Y 21 (2015); Oren Gross, *The New Way of War: Is There A Duty to Use Drones?*, 67 FLA. L. REV. 1 (2015); Gregory S. McNeal, *Targeted Killing and Accountability*, 102 GEO. L.J. 681 (2014); Douglas Cox & Ramzi Kassem, *Off the Record: The National Security Council, Drone Killings, and Historical Accountability*, 31 YALE J. ON REG. 363 (2014); David W. Opderbeck, *Drone Courts*, 44 RUTGERS L.J. 413 (2014); Matthew Craig, *Targeted Killing, Procedure, and False Legitimation*, 35 CARDOZO L. REV. 2349 (2014); Jenny-Brooke Condon, *Illegal Secrets*, 91 WASH. U.L. REV. 1099 (2014); Jennifer Daskal, *The Geography of the Battleflield: A Framework for Detention and Targeting Outside the 'Hot' Conflict Zone*, 171 U. PA. L. REV. 1165 (2013); Amos Guiora, LEGITIMATE TARGET: A CRITERIA-BASED APPROACH TO TARGETED KILLING (2013), Deborah Pearlstein, *Enhancing Due Process in Targeted Killing*, AMERICAN CONSTITUTION SOCIETY ISSUE BRIEF (Oct. 2013); Richard Murphy & Afsheen John Radsan, *Notice and an Opportunity to Be Heard Before the President Kills You*, 48 WAKE FOREST L. REV. 829 (2013); Alberto R. Gonzales, *Drones: The Power to Kill*, 82 GEO. WASH. L. REV. 1 (2013); Leila Nadya Sadat, *America's Drone Wars*, 45 CASE W. RES. J. INT'L L. 215 (2012); Carla Crandall, *Ready . . . Fire . . . Aim! A Case for Applying American Due Process Principles Before Engaging in Drone Strikes*, 24 FLA. J. INT'L L. 55 (2012); Pardiss Kebriaei, *The Distance Between Principle and Practice in the Obama Administration's Targeted Killing Program: A Response to Jeh Johnson*, 31 YALE L. & POL'Y REV. 151 (2012); Mark V. Vlasic, *Assassination & Targeted Killing-A Historical and Post-Bin Laden Legal Analysis*, 43 GEO. J. INT'L L. 259 (2012); Robert Chesney, *Who May Be Killed? Anwar al-Awlaki as a Case Study in the International Legal Regulation of Lethal Force*, *in* Y.B. INT'L HUMANITARIAN L. (M.N. Schmitt et al. eds., 2011); Lesley Wexler, *Litigating the Long War on Terror: The Role of* al-Aulaqi v. Obama, 9 LOY. U. CHI. INT'L L. REV. 159 (2011); Philip Alston, *The CIA and Targeted Killings Beyond Borders*, 2 HARV. NAT'L SEC. J. 283 (2011); Kenneth Anderson, *Targeted Killing in U.S. Counterterrorism Strategy and Law*, *in* LEGISLATING THE WAR ON TERROR: AN AGENDA FOR REFORM (Ben Wittes ed., 2009); Richard Murphy & Afsheen John Radsan, *Due Process and Targeted Killing of Terrorists*, 31 CARDOZO L. REV. 405 (2009); Daphne Barak-Erez & Matthew C. Waxman, *Secret Evidence and the Due Process of Terrorist Detentions*, 48 COLUM. J. TRANS. L. 3 (2009).

cybersurveillance and bulk metadata collection programs can be more fully interrogated through a rigorous scientific critique. I argue that this scientific validity determination is particularly justified if, as General Michael Hayden, former Director of the NSA and CIA, explains: "We kill people based on metadata."[28]

This Article, therefore, uses this single sentence from a single document released by the Snowden disclosures as a vehicle to illustrate that further dialogue is needed on whether this "full-arsenal approach" to surveillance increasingly relies upon data science, data fusion processes, and the "full-arsenal" of algorithmic, analytic, and integrative big data tools for "precision targeting." As part of the Symposium, *The Future of National Security Law*, the Article aims to accomplish two goals. First, it extends the important conversation on the future of mass surveillance programs "in the Post-Snowden age"[29] that was raised at the Symposium and builds upon my comments at this event. And, second, this Article, specifically, helps to explain why a scientific-driven inquiry might be useful to inform the impending challenges of big data-driven national security policymaking and the role of big data cybersurveillance in national security law.

At the outset, it is important to explain that this research relies exclusively upon publicly available sources. At this time, this academic endeavor is greatly enhanced as the public has been granted access to more classified documents relating to covert intelligence activities than ever before by virtue of the Snowden disclosures, media and investigative reports, and national security revelations through other intelligence sources. Yet, like the work of other scholars engaged in similar research, this work is necessarily constrained in its conclusions and restricted to the information available, which is, of course, incomplete.

Although informed heavily by credible sources and reports on intelligence activities, public statements by intelligence officials, and actual government documents, such as the Snowden disclosures, this symposium piece might be considered best as a thought experiment. As a result of this

28. David Cole, *'We Kill People Based on Metadata'*, N.Y. REV. BOOKS (May 10, 2014), http://www.nybooks.com/blogs/nyrblog/2014/may/10/we-kill-people-based-metadata/; Lee Ferran, *Ex-NSA Chief: 'We Kill People Based on Metadata,'* ABC NEWS (May 12, 2014, 12:59 PM), http://abcnews.go.com/blogs/headlines/2014/05/ex-nsa-chief-we-kill-people-based-on-metadata/.

29. Jane Harman, *Security Policies for a Post-Snowden Age,* WASH. POST OPINIONS (Nov. 7, 2013), http://www.washingtonpost.com/opinions/security-policies-for-a-post-snowden-age/2013/11/07/be307c90-464c-11e3-a196-3544a03c2351_story.html.

thought experiment, I conclude that a scientific critique, such as the one required by the Supreme Court case *Daubert v. Merrell Dow Pharmaceuticals*,[30] may aid in assessing the efficacy of big data-driven national security policymaking and the scientific validity of covert big data cybersurveillance methods.

In *Daubert,* a landmark case, the Court determined that a trial judge must engage in a preliminary assessment of whether scientific testimony is reliable.[31] For example, a trial court must assess whether the scientific testimony promulgated by a scientific expert is based on a methodology that is scientifically valid.[32] A trial court must further determine whether the scientific reasoning is generally accepted, and whether this scientific method or scientific reasoning can be properly and consistently applied to the facts at issue.[33]

This Article simply explains why *Daubert* is relevant to newly emerging big data cybersurveillance and mass dataveillance methods. I reserve for

30. 509 U.S. 579 (1993). Under the *Daubert* standard, the evidence must be not only relevant, but also reliable. *Id*. at 589. Several factors often considered in determining whether the methodology is valid are: (1) whether the theory or technique in question can be and has been tested; (2) whether it has been subjected to peer review and publication; (3) its known or potential error rate; (4) the existence and maintenance of standards controlling its operation; and (5) whether it has attracted widespread acceptance within a relevant scientific community. *Id*. at 593–95. Daubert was the first in line of a trilogy of case exploring the relationship between Federal Rules of Evidence (FRE) 702 and scientific expert testimony admissibility. The trilogy consists of *Daubert*, 509 U.S. at 579 (1993), *General Electric Co. v. Joiner*, 522 U.S. 136 (1997), and *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999).

31. *Daubert*, 509 U.S. at 598. Multiple scholars have carefully explored the significance of the *Daubert* decision. *See, e.g.*, David E. Bernstein, *The Misbegotten Judicial Resistance to the* Daubert *Revolution*, 89 NOTRE DAME L. REV. 27 (2013); David L. Faigman, *The* Daubert *Revolution and the Birth of Modernity: Managing Scientific Evidence in the Age of Science*, 46 U.C. DAVIS L. REV. 101 (2013); Eric Lasker, *Manning the* Daubert *Gate: A Defense Primer in Response to* Milward v. Acuity Specialty Products, 79 DEF. COUNS. J. 128, 128 (2012); Jennifer L. Mnookin, *Expert Evidence, Partisanship, and Epistemic Competence*, 73 BROOK. L. REV. 1009, 1016 (2008); Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CAL. L. REV. 721 (2007); David E. Bernstein & Jeffrey D. Jackson, *The* Daubert *Trilogy in the States*, 44 JURIMETRICS J. 351, 355–56 (2004); Michael J. Saks, *The Aftermath of* Daubert*: An Evolving Jurisprudence of Expert Evidence*, 40 JURIMETRICS J. 229, 233-347 (2000).

32. *Id*.; *see, e.g.*, Kumho Tire Co. v. Carmichael, 526 U.S. 137, 147 (1999) (recognizing and affirming *Daubert*'s evidentiary rationale by stating that "[i]n *Daubert*, this Court held that Federal Rule of Evidence 702 imposes a special obligation upon a trial judge to 'ensure that any and all scientific testimony . . . is not only relevant, but reliable.'" (alteration in original) (quoting *Daubert*, 509 U.S. at 589)).

33. *Daubert*, 509 U.S. at 589.

future scholarship the inquiry of whether the Foreign Intelligence Surveillance Court and other courts should be bound by *Daubert*—or a similar method of scientific inquiry—in evaluating the validity of big data cybersurveillance, mass surveillance, or bulk data collection programs.[34] Rather, this Article claims that the Supreme Court initiated with *Daubert* a tradition of closely interrogating the scientific reasoning and scientific method underlying a proposed piece of evidence as a way to assess whether that evidence should have a legal consequence. Such a consequence might include, for example, an admissibility determination that could result in the introduction of evidence to a jury in a trial. Just as forensic evidence has come to dominate much of the evidence that is debated in the criminal law context, the data science evidence that informs intelligence and law enforcement activities should be increasingly and openly debated. This is especially the case when the scientific method may be needed to assess the efficacy of big data tools used in investigation and prosecution.[35]

In this Article, I address presumptively sanctioned intelligence gathering[36] by governmental entities conducted by both the domestic intelligence and foreign intelligence components.[37] I do not address surveillance of a purely private or corporate enterprise matter.[38] I further

34. Specifically, in future research, I will explore how a *Daubert*-type inquiry could be integrated as method for weighing the evidentiary value of conclusions derived from mass cybersurvellance and big data cybersurveillance programs. In addition, I will examine ways in which *Daubert*-type analyses could be a part of what courts may consider in assessing Fourth Amendment challenges and other privacy-related constitutional challenges to mass surveillance methods. Whether efficacy and scientific validity can be required as a matter of statutory compliance with the Foreign Intelligence Surveillance Act, or any other statutory framework operating to guide the oversight of foreign and domestic intelligence gathering, is an interesting academic inquiry that I also reserve for future scholarship.

35. *See, e.g.*, Fairfield & Luna, *supra* note 10 (discussing how big data tools, including secret intelligence, are increasingly used to facilitate the investigation and prosecution of criminal defendants).

36. The government has defended the legality of its intelligence activities. *See, e.g.*, Kenneth T. Walsh, *Obama Defends NSA Surveillance*, U.S. NEWS (June 18, 2003, 10:04 AM), http://www.usnews.com/news/blogs/ken-walshs-washington/2013/06/18/obama-defends-nsa-surveillance. The constitutionality of the NSA's bulk telephony metadata collection program is the subject of ongoing litigation and is not yet resolved. *See infra* Part II.D.

37. I clarify that this Article focuses on *governmental* information and intelligence gathering and the collection and analysis of cybersurveillance intelligence by the government or official governmental delegates.

38. It has been well acknowledged, of course, the extent to which purely private uses of technologies and corporate enterprise technologies are quickly expanding governmental cybersurveillance capacities. *See, e.g.*, Christian Fuchs, *Societal and Ideological Impacts of Deep*

clarify that this Article is not a blanket rejection of big data tools. There are legitimate uses for big data tools in many contexts, and scholars are actively exploring the consequences and ethics of these tools in private and corporate settings.[39] Without understanding the architecture of mass surveillance and its proponents' aspirations, however, such a legal analysis will not be adequate to its purpose.

An informed dialogue requires taking stock of what we understand to be the current surveillance methods in a big data world and offering a *Daubert*-type lens of scientific validity to these methods. In fact, it is significant to note that a criminal defendant has already attempted to use *Daubert* as a method to critique the scientific validity of a mass cybersurveillance system that had been deployed to collect evidence against the defendant.[40] Though constrained in their public discourse due to the covert nature of their actions, I argue that the intelligence community can and should engage in a

Packet Inspection Internet Surveillance, 16 INFO., COMM. & SOC'Y 1328, 1329 (2013) ("[S]urveillance does not only have a state dimension (police and secret services monitoring citizens in order to catch criminals, terrorists, and repressing political opponents), but also has a corporate dimension: surveillance technology is a very lucrative business. State surveillance is fuelled by private businesses that produce and sell monitoring technologies that allow the surveillance of mobile phone communication, fixed line phones, email, and Internet communication and thereby achieve profit."). Many have also noted that governmental intelligence gathering responsibilities are increasingly delegated to purely private, corporate enterprises in ways that are both official and unofficial. *See, e.g.*, David Talbot, *Bruce Schneier: NSA Spying Is Making Us Less Safe*, MIT TECH. REV. (Sept. 23, 2013), http://www.technologyreview.com/news/519336/bruce-schneier-nsa-spying-is-making-us-less-safe/ (describing how the NSA might be attempting to secure cooperation with the private sector for the implementation of unofficial "backdoor" surveillance programs).

39. *See, e.g.*, Janine Hiller et al., *Privacy and Security in the Implementation of Health Information Technology (Electronic Health Records): U.S. and EU Compared,* 17 B.U. J. SCI. & TECH. L. 1, 15–16 (2011) (exploring the ethical aspects of electronic information collection and sharing in the healthcare industry); Anjanette Raymond, *The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics*, NW. J. INT'L L. & BUS. (forthcoming); John W. Bagby, *Using an Industrial Organization (I/O) Lens to Enhance Predictive Analytics: Disentangling Emerging Relationships in the Electronic Surveillance Supply Chain* (forthcoming); Philip Nichols, *The Biggest Data of All: Preparing for and Preventing Corruption in Algorithmic Healthcare* (forthcoming).

40. For example, a defendant in a recent case decided in the U.S. Court of Appeals for the Ninth Circuit raised efficacy and *Daubert* concerns over secret mass cybersurveillance programs operated by naval intelligence in a criminal prosecution. *See, e.g.*, United States v. Dreyer, 767 F.3d 826, 828 n.1 (9th Cir. 2014) ("Dreyer challenges the admission of evidence related to RoundUp [mass cybersurveillance program], arguing it did not meet the requirements for the admission of expert testimony established by Federal Rule of Evidence 702 and *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S.Ct. 2786, 125 L.Ed.2d 469 (1993)."). The Court did not reach the *Daubert* issue because it found in favor of Dreyer on other grounds. *Id.*

discussion of their scientific methods and the empirical evaluation or scientific testing, if any, of such methods.

Further, it is useful to note that while other fields may enjoy more certainty in their conclusions, a deep interrogation of the scientific underpinnings of covert cybersurveillance methods requires speculation. This analysis, as I have previously stated, is meant to engage the entire concerned community and those who are certainly more informed. The analysis is not intended to invite concrete conclusions. From an academic research perspective, it is practically impossible to interrogate these secret methods without a degree of speculation.

Despite the speculative aspects of this research, the potential legal and other consequences of this topic cannot be overemphasized. The scientific inquiry in the homeland security and national security context, like the evidentiary and criminal procedure contexts, is integral to understanding whether specific rights are protected or not.[41] As science has played an ever-expanding role in determining liability or guilt in both our civil and criminal justice system, courts increasingly recognized the need to anchor the introduction of such evidence upon sound scientific principles.[42] This Article proposes that, similarly, if data science and big data tools are increasingly used in the intelligence and national security programs and policies, scientific validity determinations should be sought prior to the implementation of these emerging cybersurveillance and dataveillance methods.

This Article proceeds in five parts. Following Part I, this introduction, in Parts II through IV, I discuss important background information, intended to frame the analysis and set the factual and legal predicate necessary for future scholarship. Specifically, in Part II, I first offer a brief definition of "small data" and "big data." Next, I will briefly summarize small data surveillance methods and then contrast small data surveillance with big data cybersurveillance—a "collect-it-all" approach to intelligence gathering that

41. *See, e.g.*, Elizabeth E. Joh, *Policing by the Numbers: Big Data and the Fourth Amendment*, 89 WASH. L. REV. 35, 56 (2014) (questioning, for example, "whether predictive software based on historical crime data is similar to other uses of third party information that have already been held to support a reasonable suspicion determination."); Shayana Kadidal, *NSA Surveillance: The Implications for Civil Liberties*, 10 I/S: J.L. & POL'Y FOR INFO. SOC'Y 433, 469–70 (2014) (recognizing the role of effectiveness and ineffectiveness in the arguments of the government regarding what extent Fourth Amendment and Fourth Amendment-like analysis should be considered in surveillance).

42. *See, e.g.*, Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).

is facilitated by bulk data collection, and mass cybersurveillance and dataveillance programs. Finally, I will provide a brief overview of the landmark case of *Daubert v. Merrell Dow Pharmaceuticals*,[43] and its relationship to a tradition of scientific interrogation in the evidentiary context. I recognize that *Daubert* currently plays no role in the Fourth Amendment[44] jurisprudence in evaluating the constitutionality of surveillance tools. Yet, to the extent that newly emerging big data cybersurveillance and mass dataveillance tools are dependent upon data science, I suggest that a *Daubert*-type analysis may be helpful to an analysis of these new surveillance methods.

Next, Part III specifically focuses on private and public datafication—the process of transitioning all human-generated data into digital forms that can be indexed and stored indefinitely. This, as Part III will discuss, facilitates the datafication of the body and biometric surveillance (i.e., 24/7 surveillance of the body), and the datafication or comprehensive surveillance of behavioral, biographical, and other personally identifiable information (i.e., 360° surveillance of the biography). Although technical, this discussion is critical in that it shows why data science reasoning and big data policy rationales appear to be both operative and persuasive in a "collect-it-all" approach to intelligence gathering.

Part IV will strive to explain how big data cybersurveillance tools appear to function to fuse biometric data (e.g., surveillance of the body) with biographic and behavioral data (e.g., surveillance of the biography) to construct digital avatars from our digital selves. Additionally, Part IV explores the virtual reality dimension of the construction of digital avatars and potential scientific limits of a "collect-it-all" approach to intelligence gathering that is big data dependent, given the inherent limitations of big data tools.

Part V concludes that the *Daubert* analyses, now embedded within the judicial oversight function, initiated a close interrogation of the scientific reasoning and scientific method underlying a proposed piece of evidence as a way to assess whether that evidence should have a legal consequence in a

43. *Id.*

44. The Fourth Amendment of the U.S. Constitution provides: "The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized." U.S. CONST. amend. IV.

civil or criminal trial context. If the intelligence community is currently presuming the efficacy and the scientific validity of "collect-it-all" methods, and is allowed to implement these tools without the benefit of a careful scientific-driven inquiry, then the imposition of a *Daubert*-type evidentiary burden is appropriate. In other words, the discussion below will attempt to illustrate why a *Daubert*-type inquiry may be helpful in conceptualizing the proper analytical structure necessary for the assessment and oversight of big data cybersurveillance and mass dataveillance methods.

II. BACKGROUND ON BIG DATA AND BIG DATA CYBERSURVEILLANCE: WHY EXAMINING *DAUBERT* AND DATA SCIENCE MATTERS

Parts II, III, and IV of this Article, in more technical and specific detail, attempt to address a subject that is gaining importance both as a matter of law and a matter of democratic governance[45]: exactly how and why small data surveillance is significantly distinct from big data cybersurveillance in the intelligence gathering context. Today, at the earliest dawn of big data, it is difficult to ascertain the efficacy of government-driven big data cybersurveillance tools in the national security context.[46] It is now openly

45. Journalist and attorney Glenn Greenwald, and journalist and documentary filmmaker Laura Poitras—who reportedly exercise sole possession over the full Snowden files—and other surveillance experts have shared the view that the Snowden disclosures profoundly implicate questions of democratic governance. *See, e.g.*, George Packer, *The Holder of Secrets: Laura Poitras's Closeup View of Edward Snowden*, NEW YORKER (Oct. 20, 2014), *available at* http://www.newyorker.com/magazine/2014/10/20/holder-secrets; GREENWALD, *supra* note 1, at 6 ("[Snowden] has made it clear, with these disclosures, that we stand at a historic crossroads. Will the digital age usher in the individual liberation and political freedoms that the Internet is uniquely capable of unleashing? Or will it bring about a system of omnipresent monitoring and control . . . ?"); LAURA POITRAS, CITIZENFOUR (2014); Peter Maass, *The Intercept's Laura Poitras Wins Academy Award for 'Citizenfour'*, INTERCEPT (Feb. 22, 2014), *available at* https://firstlook.org/theintercept/2015/02/22/poitras-wins-oscar-for-citizenfour/ ("'The disclosures that Edward Snowden revealed don't only expose a threat to our privacy but to our democracy itself,' Poitras said in her acceptance speech [at the 87th Academy Awards, immediately after Poitras received the Oscar for Best Documentary Feature for directing CITIZENFOUR]."); RACHEL LEVINSON-WALDMAN, BRENNAN CTR. FOR JUSTICE, WHAT THE GOVERNMENT DOES WITH AMERICANS' DATA 9 (2013) ("The collection and retention of non-criminal information about Americans for law enforcement and national security purposes poses profound challenges to our democracy and our liberties.").

46. *See* PETER BERGEN, DAVID STERMAN, EMILY SCHNEIDER & BAILEY CAHALL, NEW AMERICA FOUNDATION, DO NSA'S BULK SURVEILLANCE PROGRAMS STOP TERRORISTS? (2014) *available at* http://www.newamerica.net/sites/newamerica.net/files/policydocs/Bergen_NAF_NSA%20Surveillance_1_0.pdf) (arguing that "traditional" investigative tools like the use of informants

debated whether and to what extent emerging bulk collection capacities, as well as big data's mass integrative and predictive technologies, can effectively advance important national security objectives.[47] Regardless of whether bulk data collection, mass surveillance programs, and big data cybersurveillance tools have been adequately evaluated, it appears that these emerging surveillance methods are being rapidly tested and deployed amidst what has been termed the "big data revolution."[48]

As will be discussed more fully below, the NSA cybersurveillance programs revealed by the Snowden disclosures and other recent public reports shed light on a "collect-it-all" approach to intelligence gathering. I argue that more fully examining this "collect-it-all" approach and its implications in a big data world context reinforces the critical need for a *Daubert*-type inquiry of these emerging surveillance technologies. In order to comprehend why such an inquiry of these technologies under a *Daubert* analysis might be necessary, some of the critical distinctions between small data surveillance and big data cybersurveillance must be more clearly understood.

---

have been the primary method used by the NSA in counterterrorism operations in the past).

47. Several recent reports, conducted by both the public and nonprofit sectors, have investigated the efficacy of several of the programs revealed by the Snowden disclosures. *See, e.g.*, RICHARD A. CLARKE, MICHAEL J. MORRELL, GEOFFREY R. STONE, CASS R. SUNSTEIN & PETER SWIRE, REPORT AND RECOMMENDATIONS OF THE PRESIDENT'S REVIEW GROUP ON INTELLIGENCE AND COMMUNICATIONS TECHNOLOGIES LIBERTY AND SECURITY IN A CHANGING WORLD (2013), *available at* https://www.whitehouse.gov/sites/default/files/docs/2013-12-12_rg_final_report.pdf; PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD, REPORT ON THE TELEPHONE RECORDS PROGRAM CONDUCTED UNDER SECTION 215 OF THE USA PATRIOT ACT AND ON THE OPERATIONS OF THE FOREIGN INTELLIGENCE SURVEILLANCE COURT (2014), *available at* https://www.pclob.gov/library/215-Report_on_the_Telephone_Records_Program.pdf; PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD, REPORT ON THE SURVEILLANCE PROGRAM OPERATED PURSUANT TO SECTION 702 OF THE FOREIGN INTELLIGENCE SURVEILLANCE ACT (2014), *available at* https://www.pclob.gov/library/702-Report.pdf; PETER BERGEN ET AL., *supra* note 46; LEVINSON-WALDMAN, *supra* note 45. In addition, several media outlets have provided for an open debate on this issue and other issues related to the Snowden disclosures. *See, e.g.*, Jennifer Stisa Granick & Christoper Jon Sprigman, *The Criminal N.S.A.*, N.Y. TIMES (June 27, 2013), http://www.nytimes.com/2013/06/28/opinion/the-criminal-nsa.html ("If all data is 'relevant,' it makes a mockery of the already shaky concept of relevance."); Bruce Schneier, *NSA Surveillance: A Guide to Staying Secure*, GUARDIAN (Sept. 6, 2013), http://www.theguardian.com/world/2013/sep/05/nsa-how-to-remain-secure-surveillance ("The NSA has turned the fabric of the internet into a vast surveillance platform, but they are not magical."); Glyn Moody, *The Repeated Failure of the US and UK Governments' "Add More Hay" Approach to Surveillance*, TECHDIRT (Dec. 3, 2014), https://www.techdirt.com/articles/20141201/09320729286/repeated-failure-us-uk-governments-add-more-hay-approach-to-surveillance.shtml.

48. Several scholars and experts have referred to the big data phenomenon as a "revolution." *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 4.

A. Big Data v. Small Data

The big data revolution is presenting new challenges to a variety of disciplines, and understanding the highly technical nature of the topic has become essential to properly understanding the implications of big data's impact on the law and constitutional analyses.[49] Similarly, to understand a legal challenge to the scientific validity of data science evidence and other evidence dependent upon applications of big data, first, the meaning of the term and phenomenon of big data itself must be explored. Only through exploration of this highly technical topic can a fuller and more informed statutory and constitutional inquiry be realized. The discussion below will serve several purposes. First, for this Article, it sets a definitional and descriptive baseline necessary for understanding the applicability of a *Daubert*-type evaluative framework to these new technologies. Second, it will also endeavor to build the foundation for future scholarship, and a more thorough statutory and constitutional discussion.

To help understand the significance of how big data is transforming intelligence gathering, Viktor Mayer-Schönberger and Kenneth Cukier anchor the paradigmatic nature of big data by contrasting the conception of a "small data world" with a newly emerging conception of a "big data world."[50] Likewise, for this Article, I have selected a small data world versus big data world framework of analysis to compare and contrast the significant differences between how surveillance methods operate in a small data world versus how surveillance methods now appear to operate in a big data world.

49. Roughly speaking, big data, as an evolving field of research and academic study, appears to involve, for example, the interrogation of a new science (i.e., what has been termed "data science" or "big data science" and "big data engineering"); newly emerging big data tools and methods (e.g., capturing, storing, and analyzing the data generated by the Internet and Social-Mobile-Cloud technologies); big data products (e.g., interoperability among databases, big data mass integration, big data visualization and data pattern mapping, results of predictive analytics, etc.); frameworks for guiding or managing big data (e.g., big data ethics, big data protocols to maintain data integrity); big data end results (e.g., benefits to other knowledge and science pursuits like epidemiology, delivery of consumer services, improvement of decisionmaking in the public or private sectors, etc.); and the unintended consequences of big data (e.g., discriminatory inferences and disparate impact), to identify just a few sub-categories of big data inquiry.

50. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 13.

1. What is Big Data?

What is big data? According to some, "big data is revolutionizing 21st century business without anybody knowing what it actually means."[51] Jonathan Stuart Ward and Adam Barker, thus, recognize that there is a "big data conundrum": Scholars and experts agree there is presently no working definition of the term "big data."[52] Ward and Barker have attempted to craft a definition that "they hope everyone can agree on."[53] That definition is as follows: "Big data is a term describing the storage and analysis of large or complex data sets using a series of techniques[.]"[54] Julie Cohen extends the big data definition further:

> "Big Data" is shorthand for the combination of a technology and a process. The technology is a configuration of information-processing hardware capable of sifting, sorting, and interrogating vast quantities of data in very short times. The process involves mining the data for patterns, distilling the patterns into predictive analytics, and applying the analytics to new data.[55]

Other scholars and experts explain that, "'Big Data' is a generalized, imprecise term that refers to the use of large data sets in data science and predictive analytics.'"[56]

The most widely-recognized definition of big data is commonly anchored by several data-specific characteristics, often referred to as the "3-Vs" of big data: volume, velocity, and variety.[57] "Technologists often use the technical '3-V' definition of big data as 'high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.'"[58]

---

51. *The Big Data Conundrum: How to Define It?*, MIT TECH. REV. (Oct. 3, 2013), http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/ [hereinafter *Big Data Conundrum*].

52. *Id*.

53. *Id*.

54. *Id*.

55. Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1920–21 (2013).

56. Crawford & Schultz, *supra* note 10, at 96 (2014).

57. *Id*.

58. Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 394 n.3 (2014) (quoting *IT Glossary: Big Data*, GARTNER, http://www.gartner.com/it-glossary/big-data/ (last visited May 11, 2015)); *see id*. (citing the original "3-V" big data report, DOUG LANEY, 3D

Increasingly, experts note that a fourth "V" of big data involves the veracity or reliability of the underlying data.[59] Still other experts, such as Rob Kitchin, have identified additional data-specific characteristics of big data, including: "exhaustive in scope, striving to capture entire populations of systems; fine-grained resolution, aiming at maximum detail, while being indexical in identification; relational, with common fields that enable the conjoining of different data-sets; flexible, with traits of extensionality (easily adding new fields) and scalability (the potential to expand rapidly)."[60]

Additionally, and highly relevant to the inquiry of this Article, the federal government has recently adopted several definitions of big data. For example, the White House has recently described big data and its implications in its report *Big Data: Seizing Opportunities, Preserving Values*,[61] often referred to as the "Podesta Report," as the report's inquiry was led by John Podesta, Counselor to the President.[62] The Podesta Report quotes a National Science Foundation document, titled *Core Techniques and Technologies for Advancing Big Data Science & Engineering*, stating: "Big datasets are 'large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.'"[63] The National Institute of Standards and Technology explains further that big data "'exceed(s) the capacity or capability of current or

---

DATA MANAGEMENT: CONTROLLING DATA VOLUME, VELOCITY, AND VARIETY (2001), *available at* http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf).

59. *Big Data Conundrum*, *supra* note 51 ("In 2001, a Meta (now Gartner) report noted the increasing size of data, the increasing rate at which it is produced and the increasing range of formats and representations employed. This report predated the term 'big data' but proposed a three-fold definition encompassing the 'three Vs': Volume, Velocity and Variety. This idea has since become popular and sometimes includes a fourth V: veracity, to cover questions of trust and uncertainty.").

60. Lyon, *Snowden, supra* note 3, at 5 (citing the work of Rob Kitchin); *see also* KITCHIN, *supra* note 4.

61. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014) [hereinafter PODESTA REPORT], *available at* https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

62. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES, INTERIM PROGRESS REPORT (2015), *available at* https://www.whitehouse.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf.

63. PODESTA REPORT, *supra* note 61, at 3 (quoting NATIONAL SCIENCE FOUNDATION, SOLICITATION 12-499, CORE TECHNIQUES AND TECHNOLOGIES FOR ADVANCING BIG DATA SCIENCE & ENGINEERING (2012), *available at* http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm).

conventional methods and systems.'"[64] "In other words, the notion of 'big' is relative to the current standard of computation."[65]

The word "big" in the term "big data," however, is misleading.[66] In another recent White House report, *Big Data and Privacy: A Technological Perspective*, submitted by the President's Council of Advisors on Science and Technology, the Council explains that big data is not only about size, but also about new forms of knowledge creation, data-driven decisionmaking, and the inferences that can be supported by analytics.[67] "Big data is big in two different senses. It is big in the quantity and variety of data that are available to be processed. And, it is big in the scale of analysis (termed 'analytics') that can be applied to those data, ultimately to make inferences and draw conclusions."[68] Moreover, most experts agree that big data relies upon supercomputing and machine learning or artificial intelligence tools and, therefore, by its very definition, big data exceeds the ability of human capacities to make sense of the "big data" without the assistance of algorithmic tools and other computer-enabled devices.[69]

Who are the predominant drivers and users of big data today? Dave Farber, the "Grandfather of the Internet," claims that there are currently two prevalent users of big data: corporations and government agencies.[70] Both of these users appear to exploit big data, but for different ends. "First,

64. *Big Data Conundrum*, *supra* note 51.

65. *Id.* (emphasis omitted).

66. "The MIKE [Method for an Integrated Knowledge Environment] project argues that big data is not a function of the size of a data set but its complexity. Consequently, it is the high degree of permutations and interactions within a data set that defines big data." *Id.*

67. PCAST REPORT, *supra* note 20.

68. *Id.* at ix.

69. *See, e.g.*, MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 11–12 ("Though it is described as part of the branch of computer science called artificial intelligence, and more specifically, an area called machine learning, this characterization is misleading. Big data is not trying to 'teach' a computer to 'think' like humans. Instead, it's about applying math to huge quantities of data in order to infer probabilities[.]").

70. John Horgan, *U.S. Never Really Ended Creepy "Total Information Awareness" Program*, SCI. AM. (June 7, 2013), http://blogs.scientificamerican.com/cross-check/2013/06/07/u-s-never-really-ended-creepy-total-information-awareness-program/ ("Farber recalled that shortly after 9/11, the Defense Advanced Research Projects Agency initiated 'Total Information Awareness,' a surveillance program that called for recording and analyzing all digital information generated by all U.S. citizens. . . . After news reports provoked criticism of the Darpa program, it was officially discontinued. But Farber suspected that [Snowden disclosures] new surveillance programs represent a continuation of Total Information Awareness. 'I can't get anyone to deny that there's a common thread there,' he said.").

corporations can analyze the data for commercially beneficial insights. Second, government agencies can examine the data for evidence that you are engaged in suspicious activities."[71]

This Article contends, based on publicly available information, that it appears that in intelligence gathering, the government believes that it can and should exploit big data tools in the same manner as the private sector.[72] Unlike the private sector, the government must provide a legal basis for mass data collection. According to the government, the statutory basis for bulk telephony metadata collection, for example, derives from Section 215 of the USA PATRIOT Act, which authorizes the following collection: "any tangible things (including books, records, papers, documents, and other items)[.]"[73] These "tangible things", however, must be "relevant to an authorized investigation . . . [to] protect against international terrorism or clandestine intelligence activities."[74] The government has successfully argued in the Foreign Intelligence Surveillance Court that bulk collection of data is necessary *ex ante* under Section 215 of the USA PATRIOT Act. The bulk telephony metadata program provides the government with an aggregate of data (e.g., metadata on all phone calls collected from Verizon on a daily basis, thus, allowing the NSA to collect the "phone records of millions of Verizon customers daily"),[75] and then, once the bulk data is amassed, allows the intelligence community to query a specific identifier within the aggregated database once the relevance of data to an ongoing investigation is established.[76]

As will be discussed in more detail below, whether bulk telephony metadata collection under Section 215 of the USA PATRIOT Act is

71. *Id*.

72. *See, e.g*., Ira "Gus" Hunt, Presentation at Gigaom Structure Data Conference: The CIA's "Grand Challenges" with Big Data (Mar. 20, 2013) [hereinafter Hunt CIA Presentation] (video and transcript available at https://gigaom.com/2013/03/20/even-the-cia-is-struggling-to-deal-with-the-volume-of-real-time-social-data).

73. USA PATRIOT Act § 215, 50 U.S.C. § 1861(a)(1) (2012).

74. *Id*. § 1861(b)(2)(A) (2012).

75. Glenn Greenwald, *NSA Collecting Phone Records of Millions of Verizon Customers Daily*, GUARDIAN (June 5, 2013), http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order.

76. *See, e.g*., Slobogin, *Cause To Believe What*, *supra* note 1 (citing *In re Application of the F.B.I. for an Order Requiring the Prod. of Tangible Things* BR 13-109, 22 (FISA Ct. July 29, 2013) and the USA PATRIOT Act § 215, 50 U.S.C. §§ 1861 (a)(1), (b)(2)(A) (2012)).

constitutional is unresolved.[77] In challenges filed immediately after the Snowden disclosures, federal courts are now attempting to resolve whether the NSA's bulk telephony metadata collection program is consistent with constitutional protections such as the Fourth Amendment's proscription against unreasonable searches and seizures.

2. What is Small Data?

What is small data? Prior to the onset of big data, such a definition was never necessary, as all data at that time would now be considered small by today's comparison. Consequently, at the earliest dawn of the big data revolution, there is no agreed-upon definition on what is "small data." However, "[g]enerally, small data is thought of as solving discrete questions with limited and structured data, and the data are generally controlled by one institution."[78] Explained another way, small data is the world that we think we know[79]: a universe of knowledge that humans can see, touch, analyze, and perceive without the assistance of supercomputing capabilities.

Often, common definitions of a big data set as a necessary prerequisite that a data set have advanced computing storage and processing capacity in order for the data to be sufficiently big enough to qualify as "big data." Almost invariably, big data expressly or implicitly precludes human storage and processing capacity—if a human can comprehend the data without computing and algorithmic assistance, it is not big data. As a result, a small data world involves things that humans can create and grasp using human judgment alone. A big data world is a world filled with big data-driven knowledge and big data products that a human could not perceive using human judgment alone.

Because of its transformative potential, Mayer-Schönberger and Cukier and others have explained that the movement away from a small data world and towards a big data world is properly characterized as a "revolution."[80] Like the industrial revolution, big data signals a historically significant methodological and philosophical shift in how we approach and perceive information, and what we accept as efficiencies of decisionmaking and

77. *See infra* Part II.D.

78. Ferguson, *supra* note 2, at 329 n.6 (citing JULES J. BERMAN, PRINCIPLES OF BIG DATA: PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION 1–2 (2013)).

79. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 17–18.

80. *Id.*

production.[81] Unlike the industrial revolution, however, the big data revolution is taking place in the midst of what has been termed the "Information Society"[82] or the digital age. Therefore, as danah boyd and Kate Crawford explain, big data creates new forms of knowledge and the processes by which we produce knowledge and perception.[83] Cohen builds on this concept of big data as a device of knowledge creation: "Together, the technology and the process [of big data] comprise a technique for converting data flows into a particular, highly data-intensive type of knowledge."[84]

Grappling with why and how small data knowledge is distinct from big data knowledge, therefore, is essential to understanding how small data surveillance and intelligence gathering is fundamentally different from big data surveillance and intelligence gathering.

B. Small Data Surveillance Methods v. Big Data Cybersurveillance Methods

The distinction between small data surveillance methods and big data methods is critically important legally and scientifically. It is important legally because human intelligence is the foundational bulwark for investigative inquiries for law enforcement or intelligence-gathering organizations that ask: who is a suspect, what is reasonable suspicion, etc.[85] Because the data that can be gathered has changed (e.g., bulk telephony metadata or Internet term-selector queries) and methods for gathering that

81. *See* Lyon, *supra* note 3, at 6.

82. One definition of "global information society" offers the following description: "[Global Information Society"] recognizes that science and technology co-exist in a world where technology diminishes geographic, temporal, social, and national barriers to discovery, access, and use of data." REPORT OF THE INTERAGENCY WORKING GROUP ON DIGITAL DATA TO THE COMMITTEE ON SCIENCE OF THE NATIONAL SCIENCE AND TECHNOLOGY COUNCIL 17 (2009) *available at* https://www.whitehouse.gov/files/documents/ostp/opengov_inbox/harnessing_power_web.pdf.

83. danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO., COMM. & SOC'Y 662, 662–79 (2012).

84. Cohen, *supra* note 55.

85. *See, e.g.*, OFFICE OF THE INSPECTOR GEN., THE FEDERAL BUREAU OF INVESTIGATION'S COMPLIANCE WITH THE ATTORNEY GENERAL'S INVESTIGATIVE GUIDELINES (REDACTED) (2005), *available at* http://www.justice.gov/oig/special/0509/chapter3.htm ("Human sources are vitally important to our success against terrorists and criminals. They often give us critical intelligence and information we could not obtain in other ways, opening a window into our adversaries' plans and capabilities. Human sources can mean the difference between the FBI preventing an act of terrorism or crime, or reacting to an incident after the fact." (quoting Dir. Robert Mueller, Fed. Bureau of Investigation)).

data have changed (e.g., seeking all metadata located on telecommunications servers or directly tapping cables), traditional legal concepts guiding the constitutionality of new intelligence technologies must be reexamined. Human intelligence is an essential part of the Fourth Amendment analysis, but in a big data world, human intelligence and judgment is at risk of becoming marginalized.[86]

1. Small Data Surveillance Methods

Small data policing[87] and small data surveillance[88] traditionally relied upon human perception and analysis, and the sensory-based tools and physical-based evidence of a non-digitalized world. In a small data world, as a matter of technological limitation, methods of law enforcement investigation and intelligence gathering historically relied upon human intelligence, including human sensory perception analysis, and other communication gathering and analytic methods that depended upon human judgment and human decisionmaking; traditional evidence based upon analog data and paper-based files; traditional intelligence collection methods, such as traditional signals intelligence and other traditional communications interception; and other data analytic tools that centered upon traditional research approaches, such as hypothesis-driven methods.

The term "small data surveillance" has not been formally defined. However, in this Article, the term is used as a way to mark a contrast between traditional intelligence gathering methods (i.e., "small data surveillance") and newly emerging intelligence methods that are digitally data-driven, dependent upon supercomputing capacities, and capitalize on

86. *See* Donohue, *Bulk Metadata Collection, supra* note 1. Professor Laura Donohue recognized the need for a Fourth Amendment analysis, as well as the tension that exists when collecting programs are either seemingly performing the analysis themselves or are not fully understood such that human analysts can properly dispel Fourth Amendment concerns. *See id.* ("[I]t appears that neither the NSA nor FISC had an adequate understanding of how the algorithms operate. Nor did they understand the type of information that had been incorporated into different databases, and whether they had been subjected to the appropriate legal analysis before data mining."); *see also* Kyllo v. United States, 533 U.S. 27, 33–34 (2001) ("It would be foolish to contend that the degree of privacy secured to citizens by the Fourth Amendment has been entirely unaffected by the advance of technology.").

87. *See, e.g.*, Ferguson, *supra* note 2, at 340 (recognizing that a small data world included investigation through physical means, such as an officer's notice of a suspect's "observable actions," i.e., presence in a high crime neighborhood, standing on a street corner, etc.).

88. *See, e.g.*, CLARK, *supra* note 3; WALLACE & MELTON, *supra* note 3.

big data phenomena and tools (i.e., "big data cybersurveillance"). As technology has transformed the Information Society, surveillance methods have necessarily transformed as well. As David Lyon has explained, "[A]s political-economic and socio-technological circumstances change, so surveillance also undergoes alteration, sometimes transformation."[89]

The characteristic that helps to define the transformational nature of big data technologies is the predictive aspects of the data-driven knowledge. Therefore, a core distinction that separates small data surveillance from big data cybersurveillance is the fact that, historically, information gathering was of an *ex post* nature, not an *ex ante* nature.[90] Jack Balkin elaborates this point: "Older models of law enforcement have focused on apprehension and prosecution of wrongdoers after the fact and the threat of criminal or civil sanctions to deter future bad behavior. The National Surveillance State supplements this model of prosecution and deterrence with technologies of prediction and prevention."[91]

Concurrent with that shift in focus from *ex post* to *ex ante* is an exponential growth in the need for data to be analyzed.[92] *Ex post* information can be limited and focused on specific suspects and events that have triggered the need for the surveillance. *Ex ante* surveillance seeks to discover suspects before they become suspicious, so to speak, and seeks to identify future events before they occur in order to intervene beforehand.[93] Doing this requires—and is facilitated by—the datification of reality in a big data world. In other words, big data cybersurveillance tools appear to be radically changing what the government considers to be the full body of evidence that allows for the careful examination of security- and defense-driven inquiries.[94]

89. Lyon, *supra* note 3, at 2; *see also* Balkin, *National Surveillance State*, *supra* note 14; Balkin & Levinson, *supra* note 14; Murphy, *supra* note 14; CLARK, *supra* note 3; WALLACE & MELTON, *supra* note 3.

90. Balkin, *National Surveillance State*, *supra* note 14, at 10–11 ("Governance in the National Surveillance State is increasingly statistically oriented, and preventative, rather than focused on deterrence and *ex post* prosecution of individual wrongdoing.").

91. *Id.* at 10 (footnote omitted) (citing Scott Charney, *The Internet, Law Enforcement, and Security*, *in* 2 PRACTICING L. INST., FIFTH ANNUAL LAW INSTITUTE 943–44 (Ian C. Ballon et al. eds., 2001)).

92. *See* Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy*, 66 STAN. L. REV. ONLINE 65, 66 (2013).

93. *See* Balkin, *National Surveillance State*, *supra* note 14, at 15–16.

94. *See* Hu, *supra* note 18, at 1479–80 (describing recently introduced forms of "biometric ID

A primary significance of these distinctions is that, in a big data world, the investigative method has been flipped on its head. Ira "Gus" Hunt, Chief Technology Officer of the CIA, explains why the investigative process has been flipped upside down:

> When I started as analyst years ago inside the CIA, the world was pretty simple. It was the world of the few to the many in terms of information flows . . . . The Social Mobile Cloud world has completely inverted that model and has gone to this complex many-to-many model.[95]

He further suggests that, because of the inversion of information flows resulting from a big data world, the nature of the investigatory inquiry has flipped upside down as well. He specifically elaborates that in a small data world, you "move data to the question"[96] (e.g., start with the question or hypothesis and then assess the small data evidence that may be available to assist in the inquiry through human judgment and human evaluative processes). In direct contrast, in a big data world, you "move the question to the data"[97] (e.g., start with the big data evidence that has been amassed and that is available for technologically-derived insights, and then assess the question or hypothesis that might be illuminated by the data through big data tools—data mining and pattern-based analysis, database screening, statistical modeling and algorithms, predictive analytics, or other supercomputing capacities and artificial intelligence tools).

Put another way, it appears that in a small data world, investigators start with a thesis or a suspect and build evidence that allows for the gathering of small data evidence that is capable of supporting a conclusion: whether the arrest and prosecution of an individual is warranted. The question (e.g., who is a suspect and is there evidence that he or she committed the crime) leads to the gathering of evidence that can support the conviction. The government asks whether the evidence can responsibly support government

surveillance").

95. *See* Hunt CIA Presentation, *supra* note 72.

96. Within Hunt's PowerPoint slides, he includes one titled, "Tectonic Technology Shifts." The slide juxtaposes "Traditional Processing" and "Mass Analytics/Big Data." Under "Traditional Processing" of data, Hunt identifies "Move Data to Question" as a characteristic of small data. *Id*.

97. Within Hunt's PowerPoint slides, he includes one titled, "Tectonic Technology Shifts." The slide juxtaposes "Traditional Processing" and "Mass Analytics/Big Data." Under "Mass Analytics/Big Data," Hunt identifies "Move Question to Data" as a characteristic of big data. *Id*.

action (e.g., a warrant, an arrest, a prosecution) premised upon the evidence (e.g., fingerprints, witnesses, etc.).

By contrast, in a big data world, it appears that investigators and analysts start with the data. Presumably, in the case of the intelligence communities within the government, in a big data world these programs function to gather the data in order to begin formulating questions. Instead of forming a thesis about who committed a crime that has already occurred, big data can be accumulated and analyzed to allow the formulation of theses about who is likely to commit a criminal or terrorist act before any event.[98] The thesis can be probabilistic in nature, with the presumption that the broader the array of data analyzed or scope of data integrated, the more accurate the data analytic or algorithmic results will become. Statistically speaking, the predictive "thesis" appears to be true. In such a situation, preemptive action may appear to be justified in the eyes of the government.[99] Perhaps problematically, it also appears that such a thesis would not be subject to rigorous scientific inquiry. In a small data world, statistics are often taken for granted as being scientifically valid.

2. Big Data Cybersurveillance and Mass Dataveillance Methods

As we transition from a small data world to a big data world, it appears that the government may be at the earliest stages of attempting to merge small data evidence and big data evidence for prosecutorial purposes.[100] The inquiry starts with the collection of all available digitalized data in the hope that the data will lead the government to its aspiration—the discovery of the categories or sub-categories of individuals considered suspect—and, consequently, may facilitate the digital construction of the data patterns and

98. *See* Elizabeth E. Joh, *Policing by Numbers: Big Data and the Fourth Amendment*, 89 WASH. L. REV. 35, 42–48 (2014).

99. *See, e.g.*, Jennifer C. Daskal, *Pre-Crime Restraints: The Explosion of Targeted, Noncustodial Prevention*, 99 CORNELL L. REV. 327 (2014); JENNIFER BACHNER, PREDICTIVE POLICING: PREVENTING CRIME WITH DATA AND ANALYTICS 14 (2013) ("The fundamental notion underlying the theory and practice of predictive policing is that we can make probabilistic inferences about future criminal activity based on existing data.").

100. *See, e.g.*, Balkin, *National Surveillance State*, *supra* note 14; Daskal, *supra* note 99; Ferguson, *supra* note 2, at 330 ("At some point inference from this personal data (independent of the observation) may become sufficiently individualized and predictive to justify the seizure of a suspect[,]" and "[t]he next phase will use existing predictive analytics to target suspects without any firsthand observation of criminal activity, relying instead on the accumulation of data points."); *see also, e.g.*, BACHNER, *supra* note 99, at 6.

data analyses that can justify the thesis.[101] Put differently, in a small data world, the accumulation of data in response to a thesis enables the government to drill down vertically on a particular suspect or terrorist target.[102] The vertical collection of data is accumulated to isolate the key pieces of fact that can prove or disprove the thesis.[103] The vertical nature of small data investigations allows the investigator to drill down on one suspect at a time to extract the relevant data from a mass of seemingly irrelevant data.

In a big data world, by contrast, data science logic and big data policymaking rationales demand a panoramic vision of all the data in order to see the patterns and tendencies which can make visible and corroborate theories about who is predisposed to criminal or terrorist behaviors.[104] To explain further, big data analytics and bulk data collection techniques—which have led to the government's impulse to "collect everything and hang on to it 'forever'"[105]—allow the intelligence community to use inferential knowledge to digitally construct potential threats "virtually" in order to flip "virtual" suspects from a horizontal data position (e.g., "collect-it-all" and "everybody is a target") into a vertical data position (e.g., drilling down on any potential suspect or specific target at any given moment that the government deems necessary to preempt actual or algorithmically understood threats).[106]

In a small data world, resource and technological constraints restricted the government to the investigation of individual suspects.[107] Suspects were

101. *See* sources cited *supra* note 100.

102. *See* BACHNER, *supra* note 99, at 24 ("The social network analysis allowed the detective on the case 'to efficiently and effectively move his personnel resources to strategically navigate the suspect into the hands of the police.'").

103. *See, e.g.*, Slobogin, *Panvasive Surveillance*, *supra* note 1; Ferguson, *supra* note 2; WALTER L. PERRY ET AL., RAND CORP., PREDICTIVE POLICING: THE ROLE OF CRIME FORECASTING IN LAW ENFORCEMENT OPERATIONS 11–13 (2013), *available at* http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf (discussing the use of data in predictive policing).

104. *See* Joh, *supra* note 98, at 42–46.

105. *See* Hunt CIA Presentation, *supra* note 72.

106. *See, e.g.*, Lyon, *Snowden*, *supra* note 3; Slobogin, *Panvasive Surveillance*, *supra* note 1; Ferguson, *supra* note 2; PERRY ET AL., *supra* note 103, at 67. This method has allowed the Memphis Police Department to use big data and "respond to predicted threats before a criminal act is committed." PERRY ET AL., *supra* note 103, at 67.

107. Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. CAL. L. REV. 1083, 1149–51 (2002); *see also* Kevin S. Bankston & Ashkan Soltani, *Tiny Constables*

identified through small data methods, for example, the utilization of sensory-based tools and analog-based investigatory methods. As has been well-documented, human misconceptions of a perceived threat, such as racial profiling, and human fallibility, such as faulty eyewitness reporting or inaccurate human intelligence, have led to false targeting.[108]

In a big data world, however, the resource and technological limitations on the government no longer impose inherent restraints on surveillance as in the past.[109] Resource and technological innovation facilitates mass, dragnet surveillance of millions and potentially billions of individuals.[110] This, in turn, enables the potential digital investigation of anyone who engages in electronic communications and, consequently, allows for the construction of the digital avatars of potentially millions and billions of individuals. Big data precrime or preterrorism initiatives are seemingly justified by the statistically-driven evidence.[111] Human biases can be embedded in algorithms, and human misconceptions of a perceived threat can be translated into technological methods for intelligence gathering and automated or semi-automated decisionmaking.[112] The contrast in scale also distinguishes the scale of human misconceptions and human fallibility. The creation of artificial intelligence targeting systems that embed misconceptions and fallibilities can lead to potentially thousands and millions of false suspects.[113]

Thus, both human small data-driven intelligence and big data-driven intelligence are subject to error. Although most understand human frailty and human fallibility in intelligence gathering, at the earliest stages of the big data revolution perhaps it is more difficult to concede big data is susceptible to frailties and fallibilities of its own kind.

---

and the Cost of Surveillance: Making Cents Out of United States v. Jones, 123 YALE L.J. ONLINE 335 (2014).

108. *See, e.g.*, TOM R. TYLER, WHY PEOPLE OBEY THE LAW 117 (1990).

109. Bankston & Soltani, *supra* note 107, at 335.

110. *See* GREENWALD, *supra* note 1; ANGWIN, *supra* note 1.

111. *See, e.g.*, PERRY ET AL., *supra* note 103, at 2–3.

112. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1260–61 (2008); Citron & Pasquale, *supra* note 10, at 1 (discussing the use of big data to create algorithms to assess people).

113. Bruce Schneier, *Why Data Mining Won't Stop Terror*, WIRED (Mar. 9, 2006), http://archive.wired.com/politics/security/commentary/securitymatters/2006/03/70357?currentPage=all (noting that a 99% accurate system would "generate 1 billion false alarms for every real terrorist plot it uncovers").

C. Daubert *and Data Science*

Beyond differences in language and terminology in big data discussions, there does not appear to be widespread recognition of a necessary scientific interrogation of big data-driven programs, though such a tradition of scientific critique has a long-standing existence in other forms in the legal realm. In *Daubert v. Merrell Dow Pharmaceuticals*,[114] the Supreme Court handed down a landmark ruling concerning the standard for admitting expert scientific testimony in a federal trial.[115] In a broader sense, however, the Court reinforced the idea that where scientific evidence is concerned—and prior to its admission into a trial where it can have legal consequences—it should be "not only relevant, but reliable."[116] In such a circumstance where reliability is in question, the judge, through application of *Daubert*, provides a gatekeeping function intended to protect the trial and ultimately the rights of individuals in the trial from inaccurate factual judgments derived from unreliable influences.[117]

In the context of admission of such evidence, the *Daubert* Court interpreted the term "scientific," to "impl[y] a grounding in the methods and procedures of science," and the term "knowledge" to "[connote] more than subjective belief or unsupported speculation," but instead to apply to "any body of known facts or any body of ideas inferred from such facts or accepted as truths on good grounds."[118] Indeed, *Daubert* recognized, reliability of this type of evidence must be supported by "appropriate validation."[119] Thus, the gatekeeping function of the judge that *Daubert* prescribed is necessarily, in the case of scientific evidence, one grounded on careful interrogation of evidence and scientific methodology.[120] *Daubert* itself notes that there are several questions that the judge should use to guide their determination—such as whether the method producing the evidence

114. 509 U.S. 579 (1993).

115. Specifically, the *Daubert* Court agreed that Federal Rule of Evidence 702 provided the correct standard for scientific testimony admissible in trial. *Id.* at 587.

116. *Id.* at 589.

117. *Id.* at 597.

118. *Id.* at 590 (internal quotation omitted).

119. *Id.* at 590.

120. *See, e.g.*, Jennifer Mnookin & David Kaye, *Confronting Science: Expert Evidence and the Confrontation Clause*, 2012 SUP. CT. REV. 99, 99–100 (2012) (noting that *Daubert* and following case law established guidelines to assess scientific validity, which "generated a sense that scientific evidence required special attention and careful scrutiny").

"can be (and has been) tested,"[121] whether it has "been subjected to peer review and publication,"[122] whether it has a "known or potential rate of error,"[123] and if the methodology is generally accepted in the relevant scientific community.[124]

In a criminal trial, where stakes can be high, the very design of trials and the applicable rules that limit considerable evidence is "intended to protect against unreliable evidence," and thus, "inaccurate factual judgments."[125] These factual judgments need such protection as *Daubert* provides, undoubtedly, because these judgments are the foundation of verdicts that can carry serious legal consequence for individuals.

Daubert instituted the close interrogation of both the scientific evidence and the methods that underlie it before potential legal consequences can be attached to an individual on the basis of the science and the scientific evidence.[126] *Daubert* interrogations center upon the assurance of "appropriate validation,"[127] or, in other words, the efficacy of the particular scientific method used pursuant to appropriate scientific standards. While in application the assurance of scientific validity and efficacy is not always perfect, the call for scrutiny of such evidence always remains.[128]

---

121. *Daubert*, 509 U.S. at 593.

122. *Id*.

123. *Id*. at 594.

124. *Id*. (noting that although the *Frye* standard of general acceptance is not determinative of scientific evidence admissibility, it "can yet have a bearing on the inquiry").

125. Keith A. Findley, *Judicial Gatekeeping of Suspect Evidence: Due Process and Evidentiary Rules in the Age of Innocence*, 47 GA. L. REV. 723, 729 (2013) (specifically discussing the Confrontation Clause and its limits on admissible evidence based on constitutional principles).

126. *See, e.g.*, Mnookin & Kaye, *supra* note 120.

127. *Daubert*, 509 U.S. at 590.

128. *See, e.g.*, Rachel Dioso-Villa, *Scientific and Legal Developments in Fire and Arson Investigation Expertise in* Texas v. Willingham, 14 MINN. J.L. SCI. & TECH. 817 (2013) (calling for increased scrutiny on testimony and scientific evidence relating to arson); Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identification*, 95 J. CRIM. L. & CRIMINOLOGY 985 (2005) (calling for increased scrutiny relating to fingerprint identification, specifically articulating a need for error rate reform under *Daubert*); Eric Nielson, *The Admission of Scientific Evidence in a Post-*Crawford *World*, 14 MINN. J.L. SCI. & TECH. 951, 981 (2013) (calling for an increase in scrutiny of lab reports under *Daubert* and concluding that the *Crawford* line of cases does not adequately protect defendants from "shoddy work and practices [impersonating] dependable science in our courts"); Margaret A. Berger, *Expert Testimony in Criminal Proceedings: Questions* Daubert *Does Not Answer*, 33 SETON HALL L. REV. 1125 (2003) (noting that the onset of *Daubert* heightened judges' sensitivity to the need to scrutinize and likely led to the current climate in which handwriting testimony is no longer universally admitted).

Individual criminal defendants, however, are not the only ones that should be concerned with the efficacy and reliability of science used against them. An explicit dialogue concerning the efficacy of these emerging mass surveillance programs, and whether efficacy should be a part of a legal analysis in deciding the constitutionality of these programs, should be encouraged.[129] Surveillance programs operate on a clandestine level by nature, and thus cannot be interrogated for scientific validity as the law currently stands. FISA Courts, which are intended to operate as a check on the usage of surveillance programs for particularized needs, are not publicly accessible.[130] A meaningful discussion on scientific validity may be constrained due to the secret nature of these programs. Thus, the need for such an interrogation itself may not be obvious. The close, *Daubert*-type interrogation of the effectiveness of these programs and the methods behind them is missing, allowing a sentiment that these programs' efficacy can be assumed by both those that support such programs and those that oppose them.[131] Arguably, however, concerns over the efficacy of these programs influence the legal analysis of some jurists faced with the question of the constitutionality of these secret programs.[132] Put differently, a court prohibiting or permitting a mass surveillance program, with an attendant mass collection of what has been traditionally considered private information, will on some level want to know whether the program is efficacious or scientifically sound.

D. Daubert*, the Fourth Amendment, and Post-Snowden Litigation on Bulk Telephony Metadata Collection*

Whether a *Daubert*-type inquiry can be integrated into the Fourth Amendment analytical framework and, if so, exactly how it could be integrated, are questions that exceed the scope of this Article. As I have stated above, I reserve these questions for future scholarship. However, the

129. *See e.g.*, Shayana Kadidal, *NSA Surveillance: Issues of Security, Privacy, and Civil Liberty*, 10 J.L. & POL'Y FOR INFO. SOC'Y 433, 469–70 (2014) (recognizing the role of effectiveness/ineffectiveness in the arguments of the government regarding what extent Fourth Amendment/Fourth Amendment-like analysis should be considered in surveillance).

130. *See, e.g.*, Banks, *supra* note 8.

131. Speculative aspects of this research remain unavoidable due to the covert nature of the surveillance methods and because of the highly technical aspect of these programs themselves.

132. *See, e.g.*, *infra* Part II.D (discussing Klayman v. Obama, 957 F. Supp. 2d 1 (D.D.C. 2013) and ACLU v. Clapper, 959 F. Supp. 2d 724 (S.D.N.Y 2013)).

discussion below attempts to establish a foundation for why an integration of a *Daubert*-type analysis is useful in evaluating the legality and constitutionality of big data cybersurveillance programs generally, and the programs of the Snowden disclosures, in particular.

To help more clearly illustrate how big data cybersurveillance and mass dataveillance appears to be forcing an evolution of the Fourth Amendment doctrine in light of efficacy concerns, it is useful to turn to the litigation that immediately followed the Snowden disclosures. The most mature litigation challenging the legality of what was revealed by the Snowden disclosures thus far is represented by two cases concerning bulk telephony metadata collection. Both of these cases challenged the Section 215 USA PATRIOT Act bulk metadata program in federal court days after the Snowden disclosures first came to light on June 5-6, 2013.[133] In perhaps the most poignant example on the relevancy of a *Daubert*-type analysis, the U.S. District Court Judge for the Southern District of New York, William H. Pauley III, in *ACLU v. Clapper*, and U.S. District Court Judge for the District of Columbia, Richard Leon, in *Klayman v. Obama*, considered the same program—bulk telephony metadata collection—and reached entirely different interpretations of the efficacy of the program.[134]

In both *ACLU v. Clapper* and *Klayman v. Obama*, asserting a combination of statutory and constitutional claims, the plaintiffs challenged the bulk telephony metadata program that pertained to a May 24, 2006, Foreign Intelligence Surveillance Court (FISC) order requiring Verizon to turn over all telephony metadata to the NSA pursuant to Section 215 of the USA PATRIOT Act.[135] It is important to note, however, that plaintiffs

133. For a detailed history of the Snowden disclosures, see generally GREENWALD, *supra* note 1.

134. *See, e.g.*, *Klayman*, 957 F. Supp. 2d at 40–42; *Clapper*, 959 F. Supp. 2d at 729–30.

135. *See* Donohue, *Bulk Metadata Collection*, *supra* note 1, at 759 n.1 (discussing *In re* Application of the Fed. Bureau of Investigation for an Order Requiring the Prod. of Tangible Things from [Telecommunications Providers] Relating to [REDACTED], Order, No. BR 0605 (FISA Ct. May 24, 2006), *available at* https://www.eff.org/sites/default/files/filenode/docket_06-05_1dec201_redacted.ex_-_ocr_0.pdf (released by court order as part of the Electronic Frontier Foundation's Freedom of Information Act (FOIA) litigation)); *see also* OFFICE OF THE INSPECTOR GEN., NAT'L SEC. AGENCY, ST-06-0018, ASSESSMENT OF MANAGEMENT CONTROLS FOR IMPLEMENTING THE FOREIGN INTELLIGENCE SURVEILLANCE COURT ORDER: TELEPHONY BUSINESS RECORDS, *available at* http://www.dni.gov/files/documents/section/pub_Feb%2012%202009%20Memorandum%20of%20US.pdf (see page 94 of 1846 and 1862 Production). For purposes of a more precise citation, I draw from both sources. *See also* Slobogin, *Panvasive Surveillance, supra* note 1, at 1757 ("The FISC has agreed, authorizing such bulk metadata collection for the first time in May 2006, and reauthorizing this collection (from, at a minimum, the three largest service providers) every ninety

litigating government mass surveillance programs in recent years have faced several jurisdictional and doctrinal hurdles. These hurdles have included, for example, overcoming the government's standing challenges[136] and the government's challenges under the State's Secrets Doctrine.[137] In addition, for those challenging mass surveillance and mass dataveillance under the Fourth Amendment's proscription against unreasonable searches and seizures, overcoming the "third-party doctrine" or "third-party records doctrine" of the Fourth Amendment has posed a particularly difficult hurdle.

The third-party doctrine is enshrined within the current Fourth Amendment jurisprudence through *Smith v. Maryland*.[138] In *Smith*, the Court held that there is no reasonable expectation of privacy in telephone numbers that individuals dial on the reasoning that the customer knowingly shares information and records with the telephone provider. Because a telecommunications customer knowingly shares data with a third-party provider, the Court determined in *Smith* that there was neither an actual nor subjective expectation of privacy. Justice Harry A. Blackmun explained that Smith voluntarily waived his privacy right because he "conveyed numerical information to the phone company [third party] and . . . assumed the risk that the company would reveal the information to the police."[139]

Thus, in extending the logic of the third-party doctrine to the present day and in the case of the Snowden disclosures, it could be argued that when data is shared with a third party (e.g., Verizon, Google, Apple, etc.) the Court's Fourth Amendment reasonable expectation of privacy test established in *Katz v. United States*[140] does not hold. Under *Katz*, the two-prong "reasonable expectation of privacy" test requires "first that a person

---

days since then, including in the wake of the Snowden affair.").

136. *See, e.g.*, Clapper v. Amnesty International, 133 S. Ct. 1138 (2013) (holding that plaintiffs lacked standing because of a lack of certainty and traceability of the purported future injury); Jewel v. National Sec. Agency, 673 F.3d 902 (9th Cir. 2011) (holding that residential telephone customers had standing to challenge warrantless eavesdropping). The litigation in *Jewel* is ongoing under Case 4:08-cv-04373-JSW, available at *Order on Motions for Summary Judgment*, ELECTRONIC FRONTIER FOUNDATION, https://www.eff.org/document/order-motions-summary-judgment-1 (last visited May 11, 2015).

137. *See, e.g.*, U.S. v. Reynolds, 345 U.S. 1 (1953) (holding that the government has a evidentiary privilege preventing a court ordered disclosure of intelligence or military secrets). *But see* In re National Security Agency Telecommunications Records Litigation, 564 F. Supp. 2d 1109 (N.D. Cal. 2008) (holding that FISA preempted the state secrets privilege).

138. 442 U.S. 735 (1979).

139. *Id.* at 744.

140. 389 U.S. 347 (1967).

have exhibited an actual (subjective) expectation of privacy."[141] And, as a follow-on inquiry, the test requires an objective expectation of privacy as well: "second, that the expectation [of privacy] be one that society is prepared to recognize as 'reasonable.'"[142]

Citing to *Smith* and in relying upon the third-party doctrine of the Fourth Amendment, Judge Pauley in *ACLU v. Clapper* concluded that the Fourth Amendment was not violated because the bulk telephony metadata was shared by the telecommunications consumer (plaintiff ACLU) with a third party (Verizon).[143] Therefore, no reasonable expectation of privacy under *Katz* could be established.[144]

In contrast, in the case of *Klayman v. Obama*, Judge Leon determined that the third-party doctrine could not be extended to the facts at hand for the following reason: the NSA's mass collection of U.S. telephone data is "almost Orwellian," and a likely violation of the U.S. Constitution.[145] Judge Leon explained, "I cannot imagine a more 'indiscriminate' and 'abitrary invasion' than this systematic and high-tech collection and retention of personal data on virtually every single citizen for purposes of querying and analyzing it without prior judicial approval."[146]

In both cases, however, the judges appeared to first reach for a way to determine the reasonableness of bulk telephony metadata collection to test the Fourth Amendment's outer boundary of what is an "unreasonable" search or seizure. In *ACLU v. Clapper*, Judge Pauley in the Southern District of New York dismissed the ACLU's constitutional claim against the program, stating that, "[t]he effectiveness of bulk telephony metadata collection cannot seriously be disputed,"[147] quickly listing several examples offered by the government itself to support this claim.[148] By contrast, in *Klayman v. Obama*, Judge Leon in the District of D.C., following his agreement with the potential constitutional claim of the petitioner Klayman

141. *Id.* at 361 (Harlan, J., concurring).

142. *Id.*

143. *Clapper*, 959 F. Supp. 2d at 749–52.

144. *Id.* at 752 ("Because *Smith* controls, the NSA's bulk telephony metadata collection program does not violate the Fourth Amendment.").

145. *Klayman*, 957 F. Supp. 2d at 33–37.

146. *Id.* at 42.

147. *Clapper*, 959 F. Supp. 2d at 755.

148. *Id.* Judge Pauley here notes that the examples that he presents in his opinion are "several successes" elucidated from "Congressional testimony and in declarations that are part of the record in this case." *Id.*

against the same program, states that, "the Government does not cite a single instance in which analysis of the NSA's bulk metadata collection actually stopped an imminent attack," while dismissing the examples offered by the government as no better than what traditional methods would have illuminated.[149]

Much like *Daubert* in the evidentiary context, the federal judges in *Klayman v. Obama* and in *ACLU v. Clapper* demanded some level of validity from the bulk telephony metadata program before deciding whether a legal consequence is warranted. While it is impossible to know to what extent these efficacy concerns subtly influenced the legal analyses of Judge Pauley and Judge Leon, or whether either judge would find it to be relevant to constitutionality determinations, the role of efficacy and assurance of some level of scientific validity seemed to inform their legal analysis. The subtle call for a scrutiny of the program's effectiveness found in both opinions perhaps should have a place in a new Fourth Amendment analysis that has been called for elsewhere, where modern technology and dated precedent collide.[150]

Analogous to the criminal context, government electronic surveillance potentially implicates important rights of the individuals to whom the surveillance is applied without the benefits of many of the procedural safeguards in place for, by way of example, ordinary criminal defendants.[151] In a small data world, the rights implicated by, for instance, an illegal search of physical property, at least do not go unnoticed, and have the opportunity through specific procedures to be vindicated.[152] However, now "[w]ith the rise of electronic surveillance conducted remotely and surreptitiously . . . the government has achieved an unprecedented amount of control"[153] without the traditional protections of a small data world. Indeed, as is the case with

149. *Klayman*, 957 F. Supp. 2d at 40.

150. *See, e.g.*, United States v. Jones, 132 S. Ct. 945, 957 (2012) (Sotomayor, J., concurring) (questioning the applicability of the third party doctrine to modern technology and Fourth Amendment analysis); *see also* Kevin Miller, *Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm*, 19 J. TECH. L. & POL'Y 105, 110 (2014).

151. *See, e.g.*, Patrick Toomey & Brett Max Kaufman, *The Notice Paradox: Secret Surveillance, Criminal Defendants, & The Right to Notice*, 54 SANTA CLARA L. REV. 843, 847–48 (2014) (noting Fourth Amendment implications and potentially others to individuals subject to electronic surveillance).

152. *Id.*

153. *Id.* at 847 (specifically referring to the government's control over notice – or the decision not to notify – the person victimized by government electronic surveillance).

big data, which this Article and others following will endeavor to articulate, the government's "collect-it-all" approach is unprecedented, and carries with it the potential for ubiquitous and pervasive surveillance.[154] While the intelligence community admits to this "collect-it-all" approach to digital information collection,[155] all of the legal and constitutional consequences to this collection for individuals are still undetermined.[156] There is little doubt, however, that many of these programs can have consequences for some individuals.[157]

A *Daubert*-type interrogation focusing on scientific validity and reliability prior to legal consequences could, perhaps, illuminate a new understanding of the reasonableness of Fourth Amendment intrusions,[158] provide a legal defense to criminal defendants who are subjected to unwarranted surveillance,[159] or spark a legislative restructuring of the surveillance architecture in existence to better assure scientific reliability prior to the initiation of surveillance programs. This, in essence, is the

154. *See, e.g.*, Nakashima & Gellman, *supra* note 7 (discussing collect it all approach and the unprecedented nature of the NSA to do so with the use of some programs).

155. *See, e.g.*, Miller, *supra* note 150, at 110 (discussing a profile of NSA Director General Keith Alexander, which stated that he "wants as much data as he can get . . . [a]nd he wants to hang on to it for as long as he can").

156. While the author recognizes a substantial need for a thorough review of the Fourth Amendment implications of wide scale, electronic surveillance, this is beyond the scope of this Article's particular purpose.

157. *See, e.g.*, Barton Gellman, *NSA Broke Privacy Rules Thousands of Times Per Year Audit Finds*, WASH. POST (Aug. 15, 2013), http://www.washingtonpost.com/world/national-security/nsa-broke-privacy-rules-thousands-of-times-per-year-audit-finds/2013/08/15/3310e554-05ca-11e3-a07f-49ddc7417125_story.html (reporting that the NSA "counted 2,776 incidents in the preceding 12 months of unauthorized collection, storage, access to or distribution of legally protected communications. Most were unintended. Many involved failures of due diligence or violations of standard operating procedure. The most serious incidents included a violation of a court order and unauthorized use of data about more than 3,000 Americans and green-card holders."); *see also* Memorandum from Chief of Signals Intelligence Division (SID) to SIGINT (May 3, 2012), *available at* http://apps.washingtonpost.com/g/page/national/nsa-report-on-privacy-violations-in-the-first-quarter-of-2012/395.

158. "Reasonableness" here is intended to mean within the context of seminal Fourth Amendment case *Katz v. United States*, 389 U.S. 347 (1967), and those cases interpreting it thereafter.

159. *See, e.g.*, United States v. Dreyer, 767 F.3d 826, 828 n.1 (9th Cir. 2014) (challenging the introduction of evidence derived from a Naval secret agency program based on *Daubert*, while the court did not rule on the merits of this argument, the notion of a *Daubert* challenge to covert surveillance program is notable); *see also* United States v. Chiaradio, 684 F.3d 265, 277 (1st Cir. 2012) (noting that the "defendant argue[d] that he has a right to the source code in order to determine whether he could credibly challenge the reliability of the technology, and thus, block the expert testimony proffered by the government on the E2P2 program and how it implicated the defendant").

underlying argument put forth in this Article. While an exact articulation of how *Daubert* could or should apply to electronic government surveillance, big data programs, and the Fourth Amendment is beyond the scope of this Article, *Daubert*'s close scrutiny of scientific evidence is instructive as to why the technical aspects of this Article are not only legally relevant, but critical to understand. A closer understanding of the technical or scientific aspects of big data cybersurveillance methods can illuminate not only potential program flaws or avenues for legal claims, but can also lead to a better understanding of routinely debated particulars, such as whether or not mass collection is itself a "search" or a "seizure" for Fourth Amendment purposes.[160] Ultimately, discerning between small data and big data, and grasping both the capabilities and flaws of big data, mass collection, and predictive analytics is the foundation for competent legal scrutiny of such programs, and, hopefully, the effectuation of the constitutional rights of individuals subject to mass surveillance and bulk data collection.

In summary, as of yet, the Supreme Court has not determined whether bulk telephony metadata collection is statutorily or constitutionally permitted. In *Klayman v. Obama,* Judge Leon granted a preliminary injunction on the grounds that the plaintiffs are likely to succeed on the merits of their Fourth Amendment claim against the NSA for the telephony metadata collection program.[161] As of the date of this publication, no federal court, in fact, has reached a binding determination on the constitutionality of the bulk telephony metatdata program on the merits. Judge Leon's determination of the likely merits of the plaintiff's constitutional claim in *Klayman v. Obama* in the District of D.C. was made pursuant to a preliminary injunction order and, thus, was not a final determination on the merits. Further, Judge Leon stayed his order of preliminary injunction pending an appeal, warning that the government should "take whatever steps necessary to prepare itself to comply with this order when, and if, it is upheld. Suffice it to say, requesting further time to comply with this order months from now will not be well received and could result in collateral

160. Members of the government, on several occasions, have claimed that mass collection in and of itself is not surveillance, especially where big data programs gather metadata. *See, e.g.*, Dianne Feinstein, *Sen. Dianne Feinstein: Continue NSA Call-Records Program*, USA TODAY, Oct. 20, 2013, http://www.usatoday.com/story/opinion/2013/10/20/nsa-call-records-program-sen-dianne-feinstein-editorials-debates/3112715/ ("The call records program is not surveillance.").

161. 957 F. Supp. 2d 1, 30 (D.D.C. 2013).

sanctions."[162]

Although Judge Pauley of the Southern District of New York found the program constitutional under the Fourth Amendment, and further concluded that §215 of the USA PATRIOT Act impliedly precludes judicial review, and that plaintiffs' claim regarding the scope of §215 would fail on the merits—his decision in *ACLU v. Clapper* was vacated and remanded on May 7, 2015.[163] In the latter opinion, the U.S. Court of Appeals for the Second Circuit did not reach the issue of constitutionality, reversing based on its finding that §215 does not preclude judicial review and that the bulk telephony metadata collection program exceeds the scope of authorization under §215.[164]

Nevertheless, and interestingly, the U.S. Court of Appeals for the Second Circuit opinion concludes with a discussion of the constitutional issues raised by the bulk telephony metadata program, noting that, on this issue, the Supreme Court's "jurisprudence is in some turmoil."[165] Instead of trying to resolve that turmoil, the court called on the legislative branch to "pass judgment on the value of the telephone metadata program as a counterterrorism tool" as a way to help courts assess the reasonableness of the program in the face of constitutional challenges.[166] In other words, the U.S. Court of Appeals for the Second Circuit has advised that it desires information on the validity and efficacy of the surveillance program in question in order to assist the court in deciding its constitutional propriety.

In summary, from the post-Snowden litigation, it appears that whether or not big data cybersurveillance programs or mass dataveillance systems, such as the bulk telephony metadata collection program, meet a test of efficacy or scientific validity is an important inquiry in order to preserve the integrity of the judicial function, and to preserve the Fourth Amendment's proscription against unreasonable searches and seizures. The gatekeeping function of the judiciary is negated in the Fourth Amendment analysis—where mass surveillance and big data cybersurveillance tools may be driven by efficacy presumptions and a scientific justification—if there is no

162. *Id.* at 44.

163. ACLU v. Clapper, No. 14- 42-cv, 2015 WL 2097814 (2d Cir. May 7, 2015).

164. *Id.*

165. *Id.* at *29 (referring to Fourth Amendment jurisprudence leading up to, and including, *United States v. Jones*, 132 S. Ct. 945, 565 U.S. ___ (2012), in the opinion's subsequent discussion at *29-30).

166. *Id.* at *31.

meaningful way to interrogate the reliability of scientific evidence prior to the implementation of the bulk metadata collection.

III. BACKGROUND ON DATAFICATION AND DATA FUSION: WHY UNDERSTANDING BIOMETRIC AND BIOGRAPHIC DATAFICATION AND COLLECTION MATTERS

Also critical to a scientific inquiry, for any purpose, is an understanding of not only the tools of the science of big data, but also the underlying logic, scientific reasoning, policy rationales, processes of testing, evaluation of the vehicles of promulgation, etc. Datafication, thus, is important in that it can help illuminate the logic, rationales, and the processes of big data tools. "Datafication" means "transforming [all information] into a data format to make it quantified."[167] Mass datafication, as will be examined in Part III below, is one process by which information is translated into data for interpretation by algorithmic-driven programs. The following description and discussion of datafication is in service to this Article's larger goal: calling for the increased scientific inquiry into data-driven programs, big data tools, and the surveillance architecture that relies on these programs. Because of the highly technical nature of datafication, it may not be readily obvious why a call for increased scientific validity is crucial.

Datafication can be understood as the process by which all human generated activity and knowledge is converted into datafied information, and then is quantified or assigned status or new meaning. Mass datafication illuminates not only the process by which large amounts of information undergo this transformation, but also, as this Article argues, the rationale behind many preexisting policies that call for the mass collection of datafied information. In addition, datafication, as a process that facilitates new knowledge discovery and production, is a relatively new concept from which statistically-driven assessments and algorithmic-derived inferences can be enabled. As explained below, these assessments and inferences can appear to be statistically significant. Consequently, because datafication and big data tools can form the basis of data-driven suspicion or data-driven guilt that may lead to legal consequences, a *Daubert*-type inquiry should be initiated into this type of process.

To roughly analogize, datafication is to data science as the collection

167. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 15.

and storage of millions of DNA samples[168] are to forensic science.[169] The mass datafication of DNA has allowed for the rapid growth of DNA databases. With the growth and prevalence of forensic DNA evidence, evolving standards for scientific reliability have involved the *Daubert* inquiry.[170] The proliferation of DNA databases is, in fact, a form of datafication. Thus, just as the growth of forensic DNA evidence and DNA databases has forced evolving standards for assessing the scientific reliability of this new technological development, the growth of datafication and the prevalence of data science should now be subjected to similar scientific reliability assessments. A *Daubert*-type inquiry into big data cybersurveillance methods that may be dependent upon datafication and data science is, therefore, appropriate.

Because datafication is enabled by the proliferation of big data, big data and datafication go hand-in-hand. Datafication can also be understood as the underlying drive to force the issue and reinforce the underlying values of big data: a policy impetus currently underway that mandates or delegates, often under law or administrative regulation, the collection or sharing of more and more data to feed the preexisting databases and database-driven policy protocols.[171]

Datafication mandates the acquisition and collection of more and more

168. For example, the FBI currently stores the DNA of over 11.6 million offender profiles and over 612,477 forensic profiles. *CODIS—NDIS Statistics*, FED. BUREAU OF INVESTIGATION, http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics (last visited May 11, 2015). According to the FBI, this information is current as of February 2015. *Id.* "Offender" profiles include those of convicted offenders, detainees, and arrestees. *Id.*; *see also infra* Table 2 (labeled "Examples of Biometric Datafication").

169. Important research has been conducted by scholars in recent years investigating the implications of the mass datafication of forensic evidence through, for example, biometric databases, such as DNA databases. *See, e.g.*, David H. Kaye, *A Fourth Amendment Theory for Arrestee DNA and Other Biometric Databases*, 15 U. PA. J. CONST. L. 1095 (2013); Erin Murphy, *License, Registration, Cheek Swab: DNA Testing and the Divided Court*, 127 HARV. L. REV. 1 (2013); Erin Murphy, *Databases, Doctrine and Constitutional Procedure*, 37 FORDHAM URB. L.J. 803 (2010); Andrea Roth, *Safety in Numbers: Deciding When DNA Alone is Enough to Convict*, 85 N.Y.U. L. REV. 101 (2010); Jennifer L. Mnookin, *Fingerprint Evidence in an Age of DNA Profiling*, 67 BROOK. L. REV. 13 (2001).

170. *See, e.g.*, Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CAL. L. REV. 721 (2007) (discussing the efficacy of the *Daubert* inquiry in relation to the history of forensic DNA evidence: "[E]ven the short history of DNA evidence is specked with examples of both questionable methodological assertions and erroneous applications of techniques.").

171. *See id.*

digital data to feed the preexisting cybersurveillance structures or the construction of new structures, and empower government actions that are determined by digital data collection and processing protocols, and mass data analyses.[172] Datafication can also be characterized as the government's policy interest in actively developing new forms of stored data and transforming analog data[173] (e.g., paper-based files) into digital data (e.g., centralized databases that are digitally stored and indexed, and electronically searchable).

A. Surveillance of the Body: Geolocational Data and Biometric Data

In order to encourage a *Daubert*-type scientific inquiry into big data cybersurveillance tools and newly emerging surveillance methods, it is important to understand precisely and technically how 24/7 tracking of the body is accomplished. This datafication of the body is conducted both through geolocational and biometric data collection, tracking aggregation, storage, and analysis. As will be explained further below, the surveillance of the body can be fused with the surveillance of the biography through big data tools. This is a transformative technology previously unavailable to the intelligence community, and it appears that this new surveillance method has not yet been subjected to scientific interrogation. Further, scholars such as Laura Donohue have concluded that neither preexisting statutory frameworks (e.g., surveillance and privacy statutes) nor constitutional frameworks (e.g., current Fourth Amendment privacy jurisprudence), are likely to operate to protect against the new types of surveillance harms implicated by emerging biometric data tracking technologies.[174]

To provide an overview of the breadth and depth of the surveillance of the body enabled by big data and datafication, the Tables below provide examples of both geolocational datafication and biometric datafication. Geolocational datafication is the process by which the movements of people are tracked and recorded as data.[175] Devices like cellphones and EZ Passes,

172. *See id*.

173. PCAST REPORT, *supra* note 20, at 22 (explaining information that is "born analog" as coming "[f]rom the characteristics of the physical world).

174. *See, e.g.*, Donohue, *Technological Leap*, *supra* note 18.

175. *See* ARTICLE 29 DATA PROTECTION WORKING PARTY, EUR. COMM'N, OPINION 13/2011 ON GEOLOCATION SERVICES ON SMART MOBILE DEVICES 1, 3 (2011), *available at* http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp185_en.pdf.

which facilitate electronic payment of highway tolls, produce data relating to the locations and movements of people.[176] Table 1 provides examples of geolocational datafication. Biometric datafication is the process of transforming individually distinguishing bodily and behavioral characteristics into data—a means of datafying the biological body.[177] Table 2 provides examples of biometric datafication.

Table 1. Examples of Geolocational Datafication

Program	Agency	Volume
Demand for Subscriber Information from Cellphone Carriers	Law enforcement agencies	1.3 million requests in 2011; AT&T, by itself, received over 700 requests per day in 2011.[178] Approximately 1 million requests in 2012;[179] AT&T received over 815 requests per day in 2012.[180]

176. *Id*.

177. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 11.

178. Eric Lichtblau, *More Demands on Cell Carriers in Surveillance*, N.Y. TIMES, July 9, 2012, at A1, *available at* http://www.nytimes.com/2012/07/09/us/cell-carriers-see-uptick-in-requests-to-aid-surveillance.html.

179. This calculation is based upon responses of telecommunications companies to an inquiry by Senator Edward J. Markey. The number does not include data from Sprint Nextel because this company did not provide exact numbers in response to the Senator's inquiry, but instead offered to meet with the Senator to discuss in detail the number of varying types of requests they have received from law enforcement. *See For Second Year in a Row, Market Investigation Reveals More Than One Million Requests by Law Enforcement for Americans Mobile Phone Data*, U.S. SENATOR ED MARKEY MASS. (2013), http://www.markey.senate.gov/news/press-releases/for-second-year-in-a-row-markey-investigation-reveals-more-than-one-million-requests-by-law-enforcement-for-americans-mobile-phone-data.

180. Letter from Timothy P. McKone, Executive Vice President, AT&T, to Edward J. Markey, U.S. Senator, at Attachment A (Oct. 3, 2013), *available at* http://www.markey.senate.gov/documents/2013-10-03_ATT_re_Carrier.pdf.

Automated License Plate Readers	U.S. Department of Homeland Security (DHS), U.S. Department of Justice (DOJ), and other law enforcement agencies nationwide	"[H]undreds of millions of data points reveal[] the travel histories of millions of motorists."[181] The state of Maryland collected more than 85 million license plate records in 2012.[182]
Automated License Plate Readers	Private companies such as MVTrac, which compiles databases for repossession agents[183]	One company, the Digital Recognition Network (DRN), has a database with "over 700 million data points on where American drivers have been."[184]
SunPass RFID Card for Tollbooths	Florida Department of Transportation	More than 8 million transponders sold total as of August 2013.[185]
E-ZPass RFID card for Tollbooths	E-ZPass Group	As of 2012, there are more than 24.3 million E-ZPass tags.[186]
Smartphones	Various companies including Android and Apple	As of January 2014, 58% of American adults own a smartphone.[187]

181. ACLU, YOU ARE BEING TRACKED: HOW LICENSE PLATE READERS ARE BEING USED TO RECORD AMERICANS' MOVEMENTS 7 (2013), *available at* https://www.aclu.org/files/assets/071613-aclu-alprreport-opt-v05.pdf.

182. *Id*. at 13.

183. *Id*. at 28.

184. *Id*.

185. Michael Turnbell, *SunPass to Replace Oldest Transponders*, SUN SENTINEL (Aug. 1, 2013), *available at* http://articles.sun-sentinel.com/2013-08-01/news/fl-sunpass-glitches-20130731_1_transponders-toll-roads-turnpike-enterprise.

186. *See About Us*, E-ZPASS GROUP, http://www.e-zpassiag.com/about-us (last visited May 11, 2015).

187. *Mobile Technology Fact Sheet*, PEW RES. CTR., http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet (last visited May 11, 2015).

Table 2. Examples of Biometric Datafication

Program	Entity	Volume
US-VISIT (United States Visitor and Immigration Status Indicator Technology)	DHS	Approximately 300,000 fingerprint data collected per day from non-citizens crossing U.S. borders.[188]
IAFIS (Integrated Automated Fingerprint Identification System) Biometric Database	Federal Bureau of Investigation (FBI)	Over 75.9 million fingerprints in the criminal master file and over 39.6 million civil fingerprints.[189]
IDENT (Automated Biometric Identification System)	DHS	Processes over 200,000 transactions daily and has over 146 million individual fingerprint records on file. "The monthly growth rate of [approximately] 1 million fingerprint records is expected to continue . . . ."[190]
National DNA Index (NDIS)	FBI	Over 11.6 million offender profiles and over 612,477 forensic profiles.[191]

188. JENNIFER LYNCH, IMMIGR. POL'Y CTR., FROM FINGER PRINTS TO DNA 4 (2012), *available at* http://mygreencard.com/downloads/FingerprintsDNA_May2012.pdf.

189. *Integrated Automated Fingerprint Identification System*: *Fact Sheet*, FED. BUREAU OF INVESTIGATION (Dec. 5, 2013), http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis_facts.

190. *CBP—US-VISIT—Automated Biometric Identification System (IDENT)*, IT DASHBOARD (Aug. 30, 2013), https://myit-2014.itdashboard.gov/.

191. *CODIS—NDIS Statistics*, FED. BUREAU OF INVESTIGATION, http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics (last visited May 11, 2015). According to the FBI, this information is current as of February 2015. *Id*. "Offender" profiles include those of convicted offenders, detainees, and arrestees. *Id*.

FBI Facial Recognition Project	FBI	FBI project to add facial recognition-ready photographs of suspects by 2014.[192] The FBI expects its facial recognition repository to be approximately 70 million photos.[193]
Consular Consolidated Database (CCD)	U.S. Department of State (DoS)	Over 100 million visa cases, 90 million photographs.[194] Currently grows at a rate of 35,000 visa cases every day.[195]
DoD's Next Generation ABIS	U.S. Departemnt of Defense (DoD)	Over 6 million total records, including 1.6 million submissions for fiscal year 2011.[196]
Biometric Records Collected in Afghanistan	DoD	The U.S. military and Afghan government have collected more than 2.5 million biometric records of fingerprints and iris scans.[197]

192. LYNCH, *supra* note 188, at 3; Aliya Sternstein, *FBI to Launch Nationwide Facial Recognition Service*, NEXTGOV (Oct. 7, 2011), http://www.nextgov.com/nextgov/ng_20111007_6100.php.

193. Sternstein, *supra* note 192 ("The bureau expects its collection of shots to rival its repository of 70 million fingerprints once more officers are aware of the facial search's capabilities.").

194. NAT'L SCI. & TECH. COUNCIL SUBCOMM. ON BIOMETRICS & IDENTITY MGMT., THE NATIONAL BIOMETRICS CHALLENGE 6 (2011), *available at* http://www.biometrics.gov/Documents/BiometricsChallenge2011_protected.pdf.

195. U.S. DEP'T OF STATE, CONSULAR CONSOLIDATED DATABASE (CCD) PRIVACY IMPACT ASSESSMENT (PIA) 1 (2010), *available at* http://foia.state.gov/_docs/PIA/ConsularConsolidatedDatabase_CCD.pdf.

196. BIOMETRICS IDENTITY MGMT. AGENCY, ANNUAL REPORT FY11, at 9 (2012).

197. *Biometrics in Afghanistan: The Eyes Have It*, ECONOMIST (July 5, 2012, 3:28 PM), http://www.economist.com/node/21558263 ("Yet America's army and the Afghan government have collected digital records of more than 2.5m of [Afghans].").

B. Surveillance of the Biography: Personally Identifiable Data, Behavioral Data, and Other Biographical Data

Our daily habits and behaviors leave data traces that help profile and make comprehensible to third parties our consuming habits, interests, and social involvements.[198] The public and private sectors increasingly collect, store, and analyze biographical and behavioral data in whatever form it is datafied.[199] According to Balkin, more than the war on terror, it is the development of a technologically-driven and Internet-supported Information Society that has led to the National Surveillance State. "The war on terror may be the most familiar justification for the rise of the National Surveillance State, but it is hardly the sole or even the most important cause."[200] As Balkin further explains, "Government's increasing use of surveillance and data mining is a predictable result of accelerating developments in information technology. As technologies that let us discover and analyze what is happening in the world become ever more powerful, both governments and private parties will seek to use them."[201]

Table 3 provides examples of the kinds of data traces that we leave that enables the construction of behavioral profiles. So far, as discussed above, courts are unresolved on the constitutionality of the NSA's bulk metadata collection program and the legal processes that justify queries of that data,[202] but have not addressed the scientific validity of the queries themselves.[203]

198. Balkin, *National Surveillance State*, *supra* note 14, at 12.

199. *Id*. at 14.

200. *Id*. at 3 (footnote omitted) (citing Andrew Cohen, *The Legal War on Terror: White House Describing Surveillance in Military Terms*, CBS NEWS (Jan. 22, 2006), http://www.cbsnews.com/news/the-legal-war-on-terror-22-01-2006/).

201. *Id*. (footnote omitted) (citing James X. Dempsey & Lara M. Flint, *Commercial Data and National Security*, 72 GEO. WASH. L. REV. 1459, 1464–69 (2004); U.S. GEN. ACCOUNTING OFFICE, GAO-04-548, DATA MINING: FEDERAL EFFORTS COVER A WIDE RANGE OF USES (2004), *available at* http://www.gao.gov/new.items/d04548.pdf).

202. *See supra* Part II.D.

203. *See* ACLU v. Clapper, 959 F. Supp. 2d 724, 755 (S.D.N.Y. 2013) (stating that "[t]he effectiveness of bulk telephony metadata collection cannot be seriously disputed. . . . [T]he Government has acknowledged several successes . . . . [T]hey offer ample justification" and then providing three instances of the NSA's metadata collection program's successes). Yet, the court relied on the government's unofficial testimony of success rather than expert testimony regarding the efficacy of the scientific inquiries used. *Id*.

Table 3. Examples of Behavioral Datafication

Entity	Type of Data	Volume of Data
Google	Web Browsing	Google is more than 100 petabytes in size.[204] Google has more than one trillion indexed URLs and more than 3 million servers.[205] Google experiences more than 7.2 billion page views per day.[206] "Google processes more than 24 petabytes of data per day, a volume that is thousands of times the quantity of all printed material in the U.S. Library of Congress."[207]
Facebook	Uploading Photos; "Like" Clicks; and Comments	Facebook is more than 300 petabytes in size.[208] Facebook has more than 1 billion users as of August 2012.[209] An estimated 35% of all of the world's digital photos are currently stored on Facebook.[210] "[M]ore than 10 million new photos [are] uploaded [on Facebook] every hour."[211] "Facebook members click a 'like' button or leave a comment nearly three billion times per day, creating a digital trail that the company can mine to learn about users' preferences."[212]

204. *See* Hunt CIA Presentation, *supra* note 72.
205. *Id.*
206. *Id.*
207. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 8.
208. Hunt CIA Presentation, *supra* note 72.
209. *Id.*
210. *Id.*
211. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 8.
212. *Id.*

YouTube[213]	Uploading Videos	YouTube is more than 1,000 petabytes in size.[214] Over 72 hours of video is uploaded on YouTube per minute.[215] YouTube has more than 4 billion views per day.[216] "800 million monthly users of Google's YouTube service upload over an hour of video every second."[217]
Twitter	Tweets	Twitter generates more than 124 billion tweets per year.[218] "The number of messages on Twitter grows at around 200 percent a year and by 2012 had exceeded 400 million tweets a day."[219] It is estimated that there are more than 4,500 tweets per second.[220]
Global Texting	Text Messages	There are more than 6.1 trillion texts per year, and there are more than 193,000 texts per second.[221]
Cell Phone and Smartphone	Mobile Calls	There are more than 2.2 trillion cell phone calls per year; roughly more than 19 minutes of cell phone usage per person per day.[222]

213. YouTube was acquired by Google in 2006. Paul R. La Monica, *Google to Buy YouTube for $1.65 Billion*, CNNMONEY (Oct. 9, 2006, 5:43 PM), http://money.cnn.com/2006/10/09/technology/googleyoutube_deal/.

214. *See* Hunt CIA Presentation, *supra* note 72.

215. *Id.*

216. *Id.*

217. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 8.

218. *See* Hunt CIA Presentation, *supra* note 72.

219. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 8.

220. *See* Hunt CIA Presentation, *supra* note 72.

221. *Id.*

222. *Id.*

C. Fusion of 24/7 Surveillance of the Body and 360° Surveillance of the Biography

The various kinds of datafication and new surveillance methods appear to enable the government to engage in a fusion of locational-body surveillance and biographical-behavioral surveillance to infer a suspect status. The fusion process facilitates the government's protocols for identity verification and identity management purposes to enable tracking and data analytics (e.g., identifying a potential suspect or terrorist). To better understand the fusion process under big data tools, it is instructive to refer to the 2014 White House report to the President from the President's Council of Advisors on Science and Technology (PCAST), titled *Big Data and Privacy: A Technological Perspective*.[223] This report described the fusion process in the private sector consumer context in the following manner[224]:

> Data fusion occurs when data from different sources are brought into contact and new facts emerge (see Section 3.2.2). Individually, each data source may have a specific, limited purpose. Their combination, however, may uncover new meanings. In particular, data fusion can result in the identification of individual people, the creation of profiles of an individual, and the tracking of an individual's activities. More broadly, data analytics discovers patterns and correlations in large corpuses of data, using increasingly powerful statistical algorithms. If those data include personal data, the inferences flowing from data analytics may then be mapped back to inferences, both certain and uncertain, about individuals.

This 2014 White House PCAST report recognizes that fusion in data analytics can be used by the government.[225] Specifically, the President's Council Advisors on Science and Technology noted:

> After data are collected, data analytics come into play and may generate an increasing fraction of privacy issues . . . it is the *use* of a product of [big data] analysis, whether in commerce, by

223. PCAST REPORT, *supra* note 20.
224. *Id*. at x.
225. *Id* at xii.

government, by the press, or by individuals, that can cause adverse consequences to individuals.[226]

IV. BACKGROUND ON DIGITAL AVATAR CONSTRUCTION: WHY INTERROGATING THE VIRTUAL REALITY RISKS OF "COLLECT-IT-ALL" INTELLIGENCE GATHERING AND DATA FUSION IN A BIG DATA WORLD MATTERS

In Parts II and III above, this Article describes how big data technologies and the phenomenon of datafication facilitate "collect-it-all" tools that are markedly distinct from the "collect-it-all" tools that were once available to the intelligence community in a small data surveillance context. In Part IV below, I discuss how, in a big data world, "collect-it-all" intelligence gathering can now potentially facilitate the construction of digital avatars. The digital avatar perhaps can best be understood as a virtual representation of our digital selves. This construction may be enabled through processes such as the data fusion of biometric and biographic data, or the digital data fusion of the 24/7 surveillance of the body and the 360° surveillance of the biography. Further, data science rationales and big data tools appear to be driving the expansion of these emerging methods. Consequently, in Part V, I suggest that an inquiry into the scientific validity of the data science that informs big data cybersurveillance programs may be appropriate.

A. Fusion of Biometric and Biographical Data to Construct Digital Avatars

From the Snowden disclosures, it appears that the legal or other consequences that may flow from the big data cybersurveillance or mass dataveillance methods are suffered by the person associated with the suspicious digital data and, potentially, conflated with the guilty digital avatar or the digital avatar's technological surrogate (e.g., a smartphone).[227] In other words, in a big data world, the intelligence community may view the digital avatar or technological surrogate as a proxy for the actual person

226. *Id.*

227. *See* discussion *supra* Part I (citing Scahill & Greenwald, *supra* note 23 ("'We're not going after people—we're going after their phones, in the hopes that the person on the other end of that missile is the bad guy.'" (quoting drone strike operator))).

targeted.[228] Consequently, the discussion below sets forth a description of the data science and technology that I contend facilitates the construction of the digital avatar—data science and underlying scientific presumptions that I assert should be tested against a *Daubert*-type inquiry to check the scientific validity of the methods.

For at least two decades, since the rise of the Information Society, experts and scholars have been searching for the proper vocabulary to describe data surveillance, or "dataveillance,"[229] and the new capacities and consequences of this "'new surveillance.'"[230] This new form of mass dataveillance and cybersurveillance is enabled by the advent of technologies that "datafies"[231] all aspects of information (e.g., all aspects of social life, and human-generated activity and knowledge can be quantified, digitalized, stored, accessed, and analyzed).[232] The terms "data self"[233] and "cyber self"[234] are used in a variety of contexts to describe self-manipulation of an online reputation. The concept of "digital personhood,"[235] however, is different. In contrast, it describes how "digital dossiers"[236] can be created by

228. *See id*. ("According to a former drone operator for the military's Joint Special Operations Command (JSOC) who also worked with the NSA, the agency often identifies targets based on controversial metadata analysis and cell-phone tracking technologies.").

229. Roger Clarke is attributed with first introducing the term "dataveillance" into academic discourse. *See* Roger A. Clarke, *Information Technology and Dataveillance*, 31 COMM. ACM 498 (1988). Clarke describes dataveillance as the systematic monitoring or investigation of people's actions, activities, or communications through the application of information technology. *Id*.; *see also* LYON, *supra* note 2, at 16 ("Being much cheaper than direct physical or electronic surveillance [dataveillance] enables the watching of more people or populations, because economic constraints to surveillance are reduced. Dataveillance also automates surveillance. Classically, government bureaucracies have been most interested in gathering such data . . . ."); MARTIN KUHN, FEDERAL DATAVEILLANCE: IMPLICATIONS FOR CONSTITUTIONAL PRIVACY PROTECTIONS (2007) (examining constitutional implications of "knowledge discovery in databases" (KDD applications) through dataveillance).

230. LYON, *supra* note 2, at 87–88.

231. MAYER-SCHÖNBERGER & CUKIER, *supra* note 4.

232. Hunt CIA Presentation, *supra* note 72; *see, e.g.*, DOUGLAS RUSHKOFF, PRESENT SHOCK (2013); JAMES GLEICK, THE INFORMATION: A HISTORY, A THEORY, A FLOOD (2011).

233. *See, e.g.*, Robert Gordon, *The Electronic Personality and Digital Self*, 56 DISP. RESOL. J. 8 (2001).

234. Chassity N. Whitman & William H. Gottdiener, *The Cyber Self: Facebook as a Predictor of Well-being*, INT'L J. APPLIED PSYCHOANALYTIC STUD. (2015).

235. DANIEL SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE (2004).

236. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, *supra* note 107.

others to construct our "data-double,"[237] "data image,"[238] "digital persona,"[239] "electronic personality and digital self,"[240] etc. The common goal of these multiple terms is an attempt to describe that in an Information Society, "the interest of surveillance [is] not in complete bodies . . . but in fragments of data[.]"[241] Relatedly, the concept of the "proliferation of networked identities and selves[,]"[242] concerns the preservation of the autonomous self within the infrastructure of the Information Society.

For this Article, however, despite other preexisting terminology, there are many reasons why "digital avatar" is a more appropriate term than "digital person," "digital self," etc. One reason is that the intelligence community appears to use similar terminology. Chief Technology Officer of the CIA, Ira "Gus" Hunt, for example, evokes the image of the transporter[243]

237. Kevin D. Haggerty & Richard V. Ericson, *The Surveillant Assemblage*, 51 BRIT. J. SOC. 605 (2000).

238. LYON, *supra* note 2, at 87 (citing David Lyon, THE ELECTRONIC EYE: THE RISE OF SURVEILLANCE SOCIETY 19 (1994)).

239. *Id*. at 87–88 (citing Roger Clarke, *The Digital Persona and Its Application to Data Surveillance*, 10 THE INFORMATION SOCIETY 2, 77–92 (1994)).

240. Gordon, *supra* note 233.

241. LYON, *supra* note 2, at 88 (citing Haggerty & Ericson, *supra* note 237, at 612).

242. *See, e.g.*, Frank Pasquale & Danielle Keats Citron, *Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society*, 89 WASH. L. REV. 1413, 1413–14 (2014) (referring to the work of Professor Tal Z. Zarsky); *see also* JULIE COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE (2012); Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, 89 WASH. L. REV. 1375 (2014); Tal Z. Zarsky, *Mining the Networked Self*, 6 JERUSALEM REV. LEGAL STUD. 120 (2012), *available at* http://jrls.oxfordjournals.org/content/6/1/120.full.pdf?keytype=ref&ijkey=b1gi1dlZvf3iBX4.

243. *See* Hunt CIA Presentation, *supra* note 72 ("[Y]ou're already a walking sensor platform. You guys know this I hope, right? Which is that your mobile device, your smartphone, your iPad, whatever it's going to be, has got a, just, any number of these things [sensors] . . . . What's happened is that if you're a *Star Trek* fan, like I was when I was a kid, what's current now is that this mobile platform, your smartphones, have turned into your communicator, they're becoming your tricorder, and actually they're becoming your transporter, right?"). It is instructive to examine the definitions of both "tricorder" and "transporter," as Hunt uses both terms from *Star Trek* to more descriptively convey a perception of the big data potential of cloud-mobile-smart technologies. In *Star Trek*, a tricorder is a multifunction hand-held device used for sensor scanning, data analysis, and recording data. The transporter device converts a person into a pattern of materials that turns a person into a data signal that can then be transmitted and reconstructed as a person at another location. The popularized catchphrase, "Beam me up, Scotty," from *Star Trek* is commonly associated with the transporter and the imagery of the hologram of the *Star Trek* character being transported as data from one location to another. In a big data world, big data cybersurveillance may be used to facilitate the construction of a multi-dimensional-like virtual representation of our digital selves, thus resulting in the datafication of the person in a similar way to the transporter. Although Hunt does not use the term "digital avatar," his references to the smartphone (and Social-Mobile-

from *Star Trek* to explain the phenomenon of the digital person.[244] Although Hunt did not use the term "digital avatar," the transporter reference evoked an image of a multi-dimensional virtual representation of the digital selves of others. During a talk at Gigaom's Structure Data conference in New York City on March 20, 2013, titled *The CIA's "Grand Challenge" with Big Data*,[245] Hunt appears to describe how the Internet, the Social-Mobile-Cloud phenomenon, and smart technologies combined, facilitate the replacement of the self with the "transporter"[246] through, for example, the multi-dimensional virtual representation of our digital selves. Put another way, the "transporter" metaphor appears to support how our digital avatars may be constructed from the comprehensive aggregation and amalgamation of our digital footprints (e.g., a "full-arsenal approach that digitally exploits the clues a target leaves behind in their regular activities on the net to compile biographic and biometric information that can help implement precision targeting.")[247]

As a consequence of an unprecedented historical phenomenon of datafication—the digitalization of all aspects of knowledge and social activity—the "data-double" is increasingly conflated with the person who has been datafied. Profound social and legal consequences result when private and public entities conflate the data-double with the individual. In the private context, the White House recognizes that "[s]mall bits of data can be brought together to create a clear picture of a person to predict preferences or behaviors."[248] In other words, consumer data-doubles can lead corporations to seek what the White House refers to as "perfect personalization."[249]

Cloud technologies) as possessing a similar functionality as a tricorder and transporter appear to parallel the intelligence community's data collection ambitions by fusing an individual's biometric and biographic data to create a multi-dimensional data likeness of the smartphone user or user of other social-cloud-mobile-smart technologies.

244. STAR TREK, created by Gene Roddenberry, currently owned by Paramount, is a Registered Trademark of Paramount Pictures Corporation. *See, e.g.*, JUSTIN EVERETT, THE INFLUENCE OF STAR TREK ON TELEVISION, FILM, AND CULTURE 186 (2008). Legal scholars have also noted *Star Trek*'s relevancy to the study of the law. *See, e.g.*, Paul Joseph & Sharon Carton, *The Law of the Federation: Images of Law, Lawyers, and the Legal System in "Star Trek: The Next Generation,"* 24 U. TOL. L. REV. 43 (1992).

245. *See* Hunt CIA Presentation, *supra* note 72.

246. *Id.*

247. Risen & Poitras, *supra* note 16.

248. PODESTA REPORT, *supra* note 61, at 7.

249. *Id.*

Similarly, in the intelligence context and in a big data world, it appears that the intelligence community also seeks to understand how "[s]mall bits of data can be brought together to create a clear picture of a person to predict preferences or behaviors."[250] Consequently, the term "digital avatar," although currently used in the virtual gaming context, also appears to be appropriately used in the intelligence gathering context. Just as Hunt's use of the term "transporter" is not intended to be literal, the term "digital avatar" is not intended to be literal, either. Rather, this Article's usage of the term "digital avatar" attempts to identify appropriately descriptive vocabulary that might more accurately capture the capacities and ambitions of big data tools now at the intelligence community's disposal.[251]

In other words, the digital avatar analogy is very apt in the following ways: it appears to capture the intelligence community's ambition to create hologram-like representations of the digital selves of others. The analogy is incomplete in that it does not appear to adequately capture the actual technological capacities of the intelligence community. More data about what the intelligence community is doing would be necessary in order to understand how accurate this analogy is. In the meantime, it provides a useful frame of reference to conceptualize the importance of *Daubert* in assessing the scientific validity of the intelligence community's ambitions and methods.

The continuing integration of big data tools and datafication into our Information Society currently underway marks a moment of historical transformation. As the big data revolution transforms how we capture and analyze data generally—in other words, as we move from a small data world to a big data world—the intelligence community will necessarily adapt. This adaptation means moving from small data intelligence tools to big data intelligence tools. Small data tools and technology represent a reality as we once knew it. Big data tools and technologies facilitate a virtually

250. *Id.*; *see also* LYON, *supra* note 2, at 88 ("[T]he data-double emerges consequent on the interest of surveillance not in complete bodies to be controlled, but in fragments of data emanating from the body." (citing Kevin D. Haggerty & Richard V. Ericson, *The Surveillant Assemblage*, 51 BRIT. J. SOC. 605 (2000))).

251. Although some might object to the use of this term in this context because of the emerging and covert nature of the topic, identifying more consistent and precise vocabulary is challenging. In other words, because the technologies of big data cybersurviellance are new and secretive, further dialogue and transparency of the big data science that underscores the rationales behind these new surveillance methods are needed to develop an agreed-upon terminology for big data cybersurveillance tools and phenomena.

understood reality. The potential consequences of this new virtual reality in the intelligence context—big data cybersurveillance systems built upon datafication and big data knowledge—necessitate scientific validation before full, or further, implementation.

To assist in the interrogation of this fusion of biometric and biographic data, to construct the digital avatar within the surveillance architecture,[252] it is helpful to anchor this discussion around a single NSA document revealed through the Snowden disclosures. In a particularly illuminating disclosure, it was revealed that the intelligence community, such as the NSA, is moving away from "traditional communications."[253] Historically, in a small data world, intelligence gathering and investigatory methods focused on the vertical use of data—for example, drilling down on a particular crime or suspect.[254] In an Information Society and big data world, intelligence gathering and investigatory methods appear to focus now on the horizontal use of data, which is necessary in a world where digital data is gathered indiscriminately and stored indefinitely, and no particular crime or suspect necessarily exists. "Vertical scaling of data"[255] in a strictly technical capacity sense involves the improvement of computer processing power within a machine.[256] "Horizontal scaling of data"[257] involves utilizing the

252. This Article's usage of the term "architecture" attempts to follow the vocabulary of the intelligence community. *See, e.g.*, Hunt CIA Presentation, *supra* note 72. During the 2013 Gigaom's Structure Data presentation, Hunt used the term "architecture" in the following way to help illuminate the topic of his talk, "The CIA's Grand Challenge with Big Data":

> We actually want a push into what we call peta scale memory architectures to do distributed analytics and things like that. Okay, and this is what's driving all these technology shifts that you read about all the time. Alright, and what we think is doing is this is going to drive new competing architectures that will radically shift how things happen in the world.

Id.

253. Risen & Poitras, *supra* note 16 (internal quotation marks omitted).

254. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 4, at 157.

255. *See* Hunt CIA Presentation, *supra* note 72. Within Hunt's PowerPoint slides, he includes one titled, "Tectonic Technology Shifts." *Id.* The slide juxtaposes "Traditional Processing" and "Mass Analytics/Big Data." Under "Traditional Processing," Hunt identifies "Vertical Scaling" of data. *Id.*

256. *See* B. Arputhamary & L. Arockiam, *Data Integration in Big Data Environment,* BONFRING INT'L J. DATA MINING, Feb. 2015, *available at* http://www.academia.edu/10662652/Data_Integration_in_Big_Data_Environment.

257. *See* Hunt CIA Presentation, *supra* note 72. Within Hunt's PowerPoint slides, he includes one titled, "Tectonic Technology Shifts." *Id.* The slide juxtaposes "Traditional Processing" and "Mass Analytics/Big Data." *Id.* Under "Mass Analytics/Big Data," Hunt identifies "Horizontal Scaling" of data. *Id.*; *see also* Richards & King, *Big Data Ethics*, *supra* note 58, at 394 ("Peter Mell, a computer

maximum amount of processing power possible among multiple machines.[258] Because of the ever-increasing data flows that society produces—and the national security infrastructure that has announced its ambition to "collect it all"—more and more supercomputing technologies are necessary to process these vast data sets. Thus, the government has created immense data processing centers such as the NSA's Utah Data Center[259] and data fusion centers in nearly all 50 states.[260] In short, due to the ease of data generation and collection in the digital age—and the shift to horizontal scaling that allows for near limitless computing power—the assumption now leads with the proposition that everyone is a potential suspect.[261]

The fusion process functions not only to forecast the perceived threat of individuals—for example, those perceived to be suspected criminals or terrorists—but increasingly, the fusion process appears to forecast the perceived threats of social and political movements; the perceived threats of mass populations, subpopulations, and classifications of individuals; protest movements; and what the government terms as other "social contagions."[262]

Yet, some in the intelligence community may contend that data fusion and the "collect-it-all" approach is a small data "mosaic theory" approach to surveillance.[263] "The 'mosaic theory' describes a basic precept of

scientist with the National Institute of Standards and Technology, similarly constrains big data to '[w]here the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing.'" (internal citation omitted)).

258. *Id*.

259. James Bamford, *The NSA Is Building the Country's Biggest Spy Center (Watch What You Say),* WIRED (Mar. 15, 2012, 7:24 PM), http://www.wired.com/2012/03/ff_nsadatacenter/ ("At a million square feet, this $2 billion digital storage facility outside Salt Lake City will be the centerpiece of the NSA's cloud-based data strategy and essential in its plans for decrypting previously uncrackable documents.").

260. *See Fusion Center Locations and Contact Information*, U.S. DEPT. OF HOMELAND SECURITY, http://www.dhs.gov/fusion-center-locations-and-contact-information (last visited May 11, 2015).

261. "[A]ccording to [one intelligence] official: 'Everybody's a target; everybody with communication is a target.'" Bamford, *supra* note 259; *see also* James Bamford, *Big Brother is Listening,* ATLANTIC (Apr. 2006), *available at* http://www.theatlantic.com/magazine/archive/2006/04/big-brother-is-listening/304711/ (providing an early analysis of NSA collection techniques pre-Snowden disclosures).

262. *See, e.g.*, *Pentagon Spending Millions to Prepare for Mass Civil Unrest*, RT (June 13, 2014, 8:22 PM), http://rt.com/usa/165844-pentagon-minerva-research-initiative/.

263. In recent Fourth Amendment cases, the "mosaic theory" has not emerged as a theory of investigation or surveillance, but rather as a method to preserve a meaningful way to assess

intelligence gathering: disparate items of information, though individually of limited or no utility to their possessor, can take on added significance when combined with other items of information."[264] Under big data cybersurveillance and mass dataveillance tools, however, the "mosaic theory" has been transformed into a "connect-the-dots" theory where, as one intelligence official explained, "[e]verybody's a target; everybody with communication is a target."[265]

Rachel Levinson-Waldman explains the "connect-the-dots" theory of big data cybersurveillance and mass dataveillance this way: "One chief argument in favor of retaining all information gathered, regardless of its apparent law enforcement value, is that seemingly innocuous information may prove meaningful today or in the future when connected with other 'dots' of information."[266] This theory has been used by multiple leaders in the intelligence community, including Gus Hunt, Chief Technology Officer of the CIA:

> The value of any piece of information is only known when you can connect it with something else that arrives at a future point in time . . . . Since you can't connect dots you don't have, it drives us into a mode of, we fundamentally try to collect everything and hang on to it forever.[267]

Former NSA Director, General Keith Alexander, similarly used the "connect the dots" theory to justify NSA cybersurveillance programs after the

---

reasonable privacy expectations under the Fourth Amendment, as Orin Kerr, Ben Wittes, Danielle Citron, David Gray and others have noted. *See, e.g.*, David C. Gray & Danielle Keats Citron, *A Shattered Looking Glass: The Pitfalls and Potential of the Mosaic Theory of Fourth Amendment Privacy,* 14 N.C. J.L. & TECH. 381, 390 (2013); Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment,* 111 MICH. L. REV. 311, 313 (2012); Christopher Slobogin, *Making the Most* of United States v. Jones *in a Surveillance Society: A Statutory Implementation of Mosaic Theory,* 8 DUKE J. CONST. L. & PUB. POL'Y. 1, 3–4 (2012); Benjamin Wittes, *Databuse: Digital Privacy and the Mosaic*, GOVERNANCE STUDIES AT BROOKINGS INSTITUTION (Apr. 1, 2011), http://www.brookings.edu/research/papers/2011/04/01-databuse-wittes. I reserve for future scholarship a more careful study of the mosaic theory, the Fourth Amendment, and big data cybersurveillance.

264. David E. Pozen, *The Mosaic Theory, National Security, and the Freedom of Information Act,* 115 YALE L.J. 628, 630 (2005).

265. Bamford, *supra* note 259.

266. *See* LEVINSON-WALDMAN, *supra* note 45, at 17 (citing Pozen, *supra* note 264, at 630–31).

267. Matt Sledge, *CIA's Gus Hunt on Big Data: We 'Try to Collect Everything and Hang onto It Forever,'* HUFFINGTON POST (Mar. 20, 2013, 4:52 PM), http://www.huffingtonpost.com/2013/03/20/cia-gus-hunt-big-data_n_2917842.html.

Snowden disclosures.[268] The process of combining these dots into a pattern that suggests terrorist activity is generally called data mining, or 'pattern prediction': analyzing a store of data to tease out patterns connected to certain behaviors, and then looking for matching patterns in other datasets in order to predict other instances in which those behaviors are likely to occur.[269]

Superficially, therefore, it appears that the mosaic theory approach to law enforcement investigations and traditional intelligence gathering in a small data world (e.g., the collection of disparate pieces of intelligence that can be pieced together to form a fuller picture of the potential suspect or crime) parallels the "collect-it-all" approach to data collection or the "connect-the-dots" theory of mass surveillance policymaking in a big data world, as explained above. However, the mosaic theory presupposes an *ex post* investigation of an offense, generally involving a suspect or group of suspects. In contrast, in a big data world, the 'investigation' (e.g., mass data collection policy) takes place *ex ante*, where no crime has occurred and no suspect exists. As explained by the representatives of the intelligence community above, the goal of big data collection, integration, and analytics is to indiscriminately collect data for two primary purposes: first, to apply that data to future security needs (e.g., investigation of an unforeseeable criminal or terrorist investigation that may occur in the future), and, second, purportedly to predict threats and to preempt future national security risks (e.g., construct digital avatars and forecast suspects based upon suspicious digital data from data-mining or database screening, or use pattern-based analysis or algorithmic intelligence to implement statistically-driven threat risk assessments).

Consequently, the manner in which the mosaic theory operates in a small data world context is not easily transferrable to the big data world context. Further, the underlying scientific method and scientific reasoning of the "collect-it-all" approach and "connect-the-dots" theory, and the potential data fusion processes that attach, are not known due to the covert nature of secret intelligence. Yet, the "collect-it-all" approach and "connect-the-dots" theory, as operative in a big data world, are distinctly

268. *Collect It All: America's Surveillance State,* ALJAZEERA (Nov. 7, 2013, 7:15 PM), http://www.aljazeera.com/programmes/faultlines/2013/11/collect-it-all-america-surveillance-state-20131158358543439.html.

269. *See* LEVINSON-WALDMAN, *supra* note 45, at 17.

technologically dependent and appear to be data science driven. The construction of digital avatars and "precision targeting" of digital avatars (e.g., "full-arsenal approach" that fuses the digital data of "biographic and biometric information that can help implement precision targeting")[270] and "precision targeting" of the digital avatar's technological surrogate (e.g., a smartphone), similarly, appear to be animated with data science reasoning and big data policymaking. As a result, the data science and underlying policy rationales deserve close inquiry.

B. Limits of the "Collect-it-All" Approach and Virtual Reality Implications of Big Data Cybersurveillance

As discussed above, big data cybersurveillance and mass dataveillance depend upon a "collect-it-all" approach or a "connect-the-dots" theory of mass surveillance.[271] This new approach to intelligence gathering is highly controversial.[272] Levinson-Waldman has explained that it is a put-the-"haystack-before-the-needle approach to information gathering."[273] Stephen Vladeck framed the controversy in this way: there is a presumption that there is, in fact, a needle in the haystack.[274] Vladeck's point appears to be that presuming there is a needle in the haystack creates a justification for the view that all persons are suspects.

Also worthy of caution is the fact that this presumption presents the potential for multiple challenges,[275] including integrating biases into data-driven systems (e.g., confirmation bias, implicit bias, cognitive bias); path

270. Risen & Poitras, *supra* note 16.

271. *See, e.g.*, GREENWALD, *supra* note 1.

272. *See, e.g.*, Banks, *supra* note 8.

273. Vladeck, *Big Data Before and After Snowden*, *supra* note 1 (citing Rachel Levinson-Waldman, *The Double Danger of the NSA's "Collect It All" Policy on Surveillance*, GUARDIAN, Oct. 10, 2013, *available at* http:// www.theguardian.com/commentisfree/2013/oct/10/double-danger-nsa-surveillance).

274. *Id.* at 334 n.11.

275. Scholars have recently examined the various concerns arising from big data, algorithmic-decisionmaking, and predictive analytics in the private context. *See* PODESTA REPORT, *supra* note 61; Barocas & Selbst, *supra* note 10; Scott R. Peppett, *Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security & Consent,* 93 TEX. L. REV. 85 (2014); Ryan Calo, *Digital Market Manipulation,* 82 GEO. WASH. L. REV. 995 (2014); Crawford & Schultz, *supra* note 10; Pasquale & Citron, *supra* note 242, at 1413–14 (referring to the work of Professor Tal Z. Zarsky); *see also* Tal Z. Zarsky, *Understanding Discrimination in the Scored Society*, *supra* note 242.

dependency (e.g., building systems to guarantee a correlative "hit" or "miss" that is intended to indicate data is suspicious; and assuming statistical certainty that suspicious data proves guilt of terroristic or criminal threat); overreliance on automation and risk of undertrained analysts; and exacerbation of perverse incentives (e.g., metrics of success designed to track number of suspects identified rather than assess whether intelligence can independently verify suspect classification). In other words, presuming that there is a digitally constructed needle (e.g., suspect or terrorist target or precrime-preterrorist threat that can be digitally identified through big data tools) in the government's digitally constructed haystack[276] (e.g., government's attempt to store and analyze all digitally produced data in order to, purportedly, preempt crime and terrorism)[277] can create incentives to construct imaginary needles.

Reality as we understand it is changing in light of big data, which underscores the need for a *Daubert*-type inquiry to assess the accuracy and reliability of big data cybersurveillance programs. As boyd and Crawford have explained, "Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality."[278] But, what is the impact of the new "categorization of reality" or new "nature" of reality in the national security context? For clarification, it is helpful to turn to Jaron Lanier, referred to as "the father of virtual reality."[279]

Following the Snowden revelations, Lanier asserted that the potential scientific theory underlying NSA programs should be subjected to greater scientific scrutiny. Specifically, he offered his observations on the cybersurveillance capacities of the NSA, and explained why big data systems could not capture "the underlying structure of reality."[280] In an interview with *Scientific American*, Lanier explains why big data predictive

276. *See* Vladeck, *Big Data Before and After Snowden*, *supra* note 1, at 334 n.11.

277. *See* Hunt CIA Presentation, *supra* note 72.

278. boyd & Crawford, *supra* note 83.

279. Janet Maslin, *Fighting Words Against Big Data*, N.Y. TIMES, May 6, 2013, at C1, *available at* http://www.nytimes.com/2013/05/06/books/who-owns-the-future-by-jaron-lanier.html?r=0 (reviewing Jaron Lanier's book, *Who Owns the Future?*).

280. Telephone Interview by Seth Fletcher with Jaron Lanier (Oct. 15, 2013) [hereinafter Lanier Interview], *available at* http://www.scientificamerican.com/article/lanier-interview-how-to-think-about-privacy/.

analysis is digitally generated and statistically driven by supercomputing.[281] Yet, these big data methods are not grounded in reality in the scientific sense.[282]

Lanier points to the seductiveness of big data—that it marks a departure from the frailties of small data, which is clearly tied to the frailties of human intelligence.[283] Lanier points out that big data does have a limited predictive certainty[284]:

> If it simply didn't work at all, then that would mean that everyone who tried to do it would fail, and they would stop trying to do it, right? However, you know the problem here is that it is a seductive illusion, and here's how this illusion works: Statistics are correct. The mathematics behind statistics is valid. So what that means is that if you are gathering data about the world and you're trying to predict events that have certain characteristics, which is that they change gradually and that covers most events in the world, then a lot of data—a big data approach—statistical projections will, by definition, work for awhile. You'll be able to project how things are changing.

The "seductive illusion" of big data's predictions springs from the sense that the limited and local predictive certainty can be amplified and expanded into a permanent and overarching predictive mechanism whose accuracy has comparable certitude.[285] But the real world cannot actually be "datafied": there will always be a gap between the real world and virtual world that shows the virtual reality to be limited and instrumental within a narrow context.[286] Explains Lanier, "[A] statistical view of the world like that is very short term. . . . The world has an underlying structure that statistics can never address by its nature . . . ."[287] In short, the efficacy of big data is unquestioned, but only in limited circumstances; its overarching pretensions

281. *Id.*
282. *Id.*
283. *Id.*
284. *Id.*
285. *Id.*
286. *Id.*
287. *Id.*

are simply illusions.[288] "And that's why we have scientists and theories . . . to talk about the structure of reality, not just trend lines. So anything relying on big data and trend lines will hit a wall at some point because [those statistical] models don't actually fit the structure of reality."[289]

---

288. *Id.*

289. *Id.* Lanier's discussion in its entirety:

> I think what we can talk about that is concrete is the economic effect [of big data privacy violations] . . . because I think that is really what unifies the issues of the NSA with the issues of Silicon Valley with the issues of the financial industries. The fundamental driver for the people who own the biggest computers is confused in my opinion. And what the driver is the sense that if you can gather the data of a lot of other people you can use statistics to analyze that data to your own benefit and gain what we can call an automatic benefit. If you can just analyze everything going on in the world very carefully to calculate your move in a more informed way than anyone else, you can create the perfect investment that will always yield a profit, you can create the perfect business that will always grow and yield a profit, or in the case of the NSA and other security agencies around the world, you can just press a button and get automatic security because you have information superiority. And there are a lot of problems with this . . . and one could talk about whether this is fair, one could talk about whether this is sustainable, but the most important problem with it is that it is self-limiting. And it is self-limiting in a somewhat tricky way. If it simply didn't work at all, then that would mean that everyone who tried to do it would fail, and they would stop trying to do it, right? However, you know the problem here is that it is a seductive illusion, and here's how this illusion works: statistics are correct. . . . The mathematics behind statistics is valid. So what that means is that if you are gathering data about the world and you're trying to predict events that have certain characteristics, which is that they change gradually and that covers most events in the world, then a lot of data—a big data approach—statistical projections will, by definition, work for awhile. You'll be able to project how things are changing. So with enough data you should be able to project the future of, oh, I don't know, the stock price, or somebody's purchasing behavior, or somebody's health, or somebody's political leanings, or somebody's likelihood to participate in a crime, all sorts of things like that. But, the problem is that a statistical view of the world like that is very short term. . . . The world has an underlying structure that statistics can never address by its nature, you know. And that's why we have scientists and theories, you know, to talk about the structure of reality, not just trend lines. So anything relying on big data and trend lines will hit a wall at some point because [those statistical] models don't actually reflect the structure of reality. So every financial scheme that seems to be perfect and an automatic generator of money will at some point hit the wall of underlying structure and then crash, demanding a giant public bailout. Exactly the same thing will happen with intelligence agencies that might for a moment think they have this automatic engine of security but it will hit the underlying structure of reality and will suddenly fail. Exactly the same thing will happen with Silicon Valley companies, or so I predict . . . . So what I think really has to happen for us to address privacy is first to understand the underlying mistaken understanding of how statistics can be used to represent reality that's falsely or improperly motivating the people who run the biggest computers. And as soon as their understanding of the advantages of big data can be more mature and they can take on a longer-term

Lanier finally predicts that the big data cybersurveillance methodologies are likely to collapse due to their failure to reflect an underlying "structure of reality"[290] He concludes that a rational discussion on the efficacy of big data cybersurveillance and mass dataveillance methods by the intelligence community is currently difficult for several reasons, including the overestimation of the assumed benefits of big data tools and an overreliance on supercomputing capacities.[291] There is, according to Lanier, a "[m]istaken understanding of advantages of big data. Until the conversation can be made more mature, people with the biggest computers think that they have a magic lamp, so it is hard to have a rational conversation."[292]

According to Lanier, therefore, the conclusions drawn from big data cybersurveillance are not necessarily drawn from a representation of reality or fact in the scientific method sense.[293] Consequently, it is unclear whether targets and threats identified by big data cybersurveillance can be reconciled as real or factually-grounded if analysts abide by definitions of reality or fact that were forged in a small data world. This is because it is, as yet, unresolved whether the artificial intelligence tools and statistical-algorithmic methods of big data cybersurveillance are capable of supporting an "underlying structure of reality."[294] In other words, big data may discover "threats" that do not exist in the real world.[295] Lanier's observation on the immaturity of the discourse surrounding the presumed accuracy and efficacy of big data cybersurveillance is a function of the fact that we are at the earliest dawn of generating and deploying these tools. The infancy of the discussion combined with the illusion of efficacy of big data tools appears to highlight the need for a *Daubert*-type inquiry.

> perspective, then I think we'll have the basis for talking about privacy that is more rational. But as long as the people with the biggest computers feel that they have Aladdin's magic lamp and they can automatically get a benefit from it, it's very hard to have a rational discussion.

Id.

290. *Id.*

291. *Id.*

292. *Id.*; *see also* EVGENY MOROZOV, TO SAVE EVERYTHING, CLICK HERE: THE FOLLY OF TECHNOLOGICAL SOLUTIONISM (2013).

293. *See, e.g.*, Kitchin, *supra* note 26 (suggesting that application of the scientific method differs when utilized in the data-driven science context).

294. Lanier Interview, *supra* note 280.

295. *See, e.g.*, BHAVANI THURAISINGHAM, WEB DATA MINING AND APPLICATIONS IN BUSINESS INTELLIGENCE AND COUNTER-TERRORISM 203 (2013), *available at* https://www.utdallas.edu/~jxr06 1100/paper-for-website/%5B18%5DMining-Terrorism-NGDM04.pdf.

V. CONCLUSION

At the earliest dawn of big data, it is difficult to ascertain the accuracy and efficacy of a big data approach to intelligence gathering and security decisionmaking. Thus, it is difficult to ascertain whether, and to what extent, big data tools can be appropriately applied to manage the risks of perceived security threats. Big data cybersurveillance, unlike small data surveillance, relies upon data science, datafication and dataveillance, artificial intelligence, and algorithmic-driven processes. These big data tools may be used to facilitate the data fusion and construction of our digital avatars which can potentially, in turn, form the basis for precrime targets and security threat forecasting. This predictive analysis is digitally generated and statistically driven by supercomputing; however, it is not grounded in reality in the scientific sense.

Moving away from a traditional intelligence gathering model that had previously engaged small data surveillance methods, it appears that, in a big data world, the intelligence community now employs a "full-arsenal approach that digitally exploits the clues a target leaves behind in their regular activities on the net to compile biographic and biometric information that can help implement precision targeting."[296] A *Daubert*-type inquiry can assist in evaluating whether this "full-arsenal approach" is scientifically sound, and whether and to what extent rapidly evolving bulk metadata and mass data surveillance methods increasingly rely upon data science and big data's algorithmic, analytic, and integrative tools. Further, a *Daubert*-type approach to assessing big data cybersurveillance methods initiates an important conversation: how best to include established scientific validation questions and testing principles within a framework to evaluate the legality and constitutionality of these newly emerging methods.

By necessity—given the opacity and complexity of big data cybersurveillance methods—this Article is highly definitional and descriptive in its approach. This effort requires the investment of significant attention to the technologies revealed by the Snowden disclosures and other recent disclosures on emerging bulk metadata collection, mass data surveillance efforts, and cybersurveillance policymaking developments. In this Article, as a topic of academic inquiry, I have argued that a science-driven approach to the interrogation of rapidly evolving big data-driven

296. Risen & Poitras, *supra* note 16.

mass data surveillance methods deserves to be treated on its own.

Therefore, this Article simply endeavors to explain why *Daubert* is relevant to newly emerging big data cybersurveillance and mass cybersurveillance methods. I conclude that to the extent that covert intelligence gathering relies upon data science, a *Daubert*-type inquiry is helpful in conceptualizing the proper analytical structure necessary for the assessment and oversight of these emerging big data cybersurveillance methods. Establishing the underlying "why," as this Article has attempted to accomplish, now sets the foundation for establishing the underlying "how": the legal analytical structure for integrating a *Daubert*-type inquiry into the Fourth Amendment. In future scholarship, I will address specifically how a *Daubert*-type inquiry, or other scientific-driven analyses, could be included within the Fourth Amendment's analytical framework to evaluate the reasonableness and efficacy of big data cybersurveillance methods. Thus, I reserve for future research the question of whether and how the Foreign Intelligence Surveillance Court and other courts could be informed by *Daubert* in evaluating the validity of big data cybersurveillance, mass surveillance, or bulk data collection programs. I also reserve for future scholarship an analysis of whether the current legal framework suffices to protect constitutional values in the face of big data cybersurveillance and mass dataveillance capacities.

In summary, this Article claims that the Supreme Court initiated with *Daubert* a tradition of carefully understanding and then interrogating the scientific reasoning and scientific method that underpins any proposed evidence that purports to be scientific in nature. *Daubert* is indicative of a trend that illustrates the way in which the law attempts to handle science. Before evidence is deemed worthy of inclusion in trial, when evidence is scientific-based evidence, the validity of the science that informs the evidence must meet a minimum evidentiary threshold.

Daubert currently plays no role in the Fourth Amendment jurisprudence in evaluating the constitutionality of surveillance tools. Further, it appears that to the best of public knowledge, the political branches also do not utilize a *Daubert*-type inquiry in the oversight of mass surveillance and big data cybersurveillance methods. As mentioned in the discussion above, however, a criminal defendant has already attempted to use *Daubert* as a method to critique the scientific validity of a mass cybersurveillance system that had

been deployed to collect evidence against the defendant.[297]

Daubert embedded within the judicial oversight function a close interrogation of the scientific reasoning and scientific method underlying a proposed piece of evidence as a way to assess whether that evidence should have a legal consequence against a defendant, civil or criminal. If the intelligence community is currently presuming the efficacy and the scientific validity of "collect-it-all" methods, the scientific aspects of intelligence gathering deserves further examination and greater transparency. Further, if the intelligence community is currently allowed to implement newly emerging big data cybersurveillance tools, and if the expansion and deployment of these tools are driven by data science reasoning without the benefit of a careful scientific-driven inquiry, then the imposition of a *Daubert*-type evidentiary burden is appropriate. By comparing and contrasting small data surveillance and big data cybersurveillance methods, this Article demonstrates why the governing law on surveillance and data gathering, and Fourth Amendment jurisprudence, should now evolve to assess the efficacy and science of new surveillance methods tested and deployed in a new big data surveillance world.

297. *See supra* note 40 and accompanying text; *see also* United States v. Dreyer, 767 F.3d 826, 828 n.1 (9th Cir. 2014) (discussing the defendant's *Daubert* challenge of a mass cybersurveillance program).

***