Oklahoma Law Review

Volume 66 | Number 4

2014

Big Data Distortions: Exploring the Limits of the ABA LEATPR Standards

Andrew G. Ferguson
University of the District of Columbia, aferguson@udc.edu

BIG DATA DISTORTIONS: EXPLORING THE LIMITS OF THE ABA LEATPR STANDARDS

ANDREW GUTHRIE FERGUSON*

Before moving on to my contribution about how the growing reliance on big data analytics may necessitate a slight modification to the ABA Standards on Law Enforcement Access to Third Party Records (LEATPR Standards),[1] I would like first to pay a few compliments to the drafters of the LEATPR Standards for producing such a systematic, thoughtful, and elegant framework for considering Fourth Amendment freedoms. As anyone who writes about or teaches the Fourth Amendment knows, the doctrine remains a theoretical muddle.[2] Yet, despite a minefield of conflicting precedent, the drafters of the LEATPR Standards have managed to construct a defensible and coherent structure on which to build third party protections. I hope legislatures take note of the logic, scholarship, and wisdom of the committee in providing such a considered analysis of a complex problem.

Introduction

The value in "third party records" is information—masses of revealing information.[3] This data can expose clues about individuals, groups, or patterns of criminal activity.[4] This data can identify, link, and prove

* Associate Professor of Law, David A. Clarke School of Law at the University of the District of Columbia. Thank you to Professor Stephen Henderson for inviting me to participate in the *Oklahoma Law Review* Symposium.

1. *See* ABA STANDARDS FOR CRIMINAL JUSTICE: LAW ENFORCEMENT ACCESS TO THIRD PARTY RECORDS (3d ed. 2013) [hereinafter LEATPR STANDARDS]. Individual standards will be referred to using the format 'STANDARD x-x.'

2. *See* Morgan Cloud, *Rube Goldberg Meets the Constitution: The Supreme Court, Technology and the Fourth Amendment*, 72 MISS. L.J. 5, 6-7 (2002); Orin S. Kerr, *An Equilibrium-Adjustment Theory of the Fourth Amendment*, 125 HARV. L. REV. 476, 479 (2011).

3. VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 2 (2013).

4. *See, e.g.*, JAMES MANYIKA ET AL., MCKINSEY GLOBAL INST., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 87 (2011), *available at* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation; Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240 (2013).

involvement in crime.[5] Much of this data is personal: involving information individuals may hope to keep private from law enforcement officials.[6] As a result, individuals' desire to keep this information private is often in conflict with law enforcement's obligation to aggressively pursue investigations, which may include accessing personal data. This tension between privacy and police investigation has been left unsatisfactorily resolved by the current Fourth Amendment doctrine.[7] Thus, the American Bar Associations' Standards for Criminal Justice proposed Law Enforcement Access to Third Party Records (LEATPR Standards) provide an alternative approach to balance the competing needs in this new world of available data.

The question this article poses is how the LEATPR Standards can survive the impact of big data policing.[8] Big data policing, as described here, involves utilizing vast, networked, commercial databases to investigate and also predict criminal activity.[9] Big data policing involves the use of not just third party, but "fourth party" commercial aggregators[10]

---

5. *See, e.g.*, Chris Jay Hoofnagle, *Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement*, 29 N.C. J. INT'L L. & COM. REG. 595, 595-96 (2004); Jon D. Michaels, *All the President's Spies: Private-Public Intelligence Partnerships in the War on Terror*, 96 CALIF. L. REV. 901, 902 (2008) ("[P]rivate organizations can at times obtain and share information more easily and under fewer legal restrictions than the government can when it collects similar information on its own.").

6. *See, e.g.*, Christopher Slobogin, *Government Data Mining and the Fourth Amendment*, 75 U. CHI. L. REV. 317, 317 (2008); Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1138 (2002) [hereinafter *Access and Aggregation*].

7. *See* Marc Jonathan Blitz, *Video Surveillance and the Constitution of Public Space: Fitting the Fourth Amendment to a World That Tracks Image and Identity*, 82 TEX. L. REV. 1349, 1383 (2004); Orin S. Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 313-14 (2012); Kerr, *supra* note 2, at 480; Raymond Shih Ray Ku, *The Founders' Privacy: The Fourth Amendment and the Power of Technological Surveillance*, 86 MINN. L. REV. 1325, 1326 (2002); Ric Simmons, *From* Katz *to* Kyllo*, A Blueprint for Adapting the Fourth Amendment to Twenty-First Century Technologies*, 53 HASTINGS L.J. 1303, 1321-22 (2002); Christopher Slobogin, *Public Privacy: Camera Surveillance of Public Places and the Right to Anonymity*, 72 MISS. L.J. 213, 217 (2002); James J. Tomkovicz, *Technology and the Threshold of the Fourth Amendment: A Tale of Two Futures*, 72 MISS. L.J. 317, 438 (2002).

8. *See infra* Part I.

9. *Id.*

10. Joshua L. Simmons, Note, *Buying You: The Government's Use of Fourth-Parties to Launder Data About "The People"*, 2009 COLUM. BUS. L. REV. 950, 951-52.

as well as de-identified datasets, which eventually can be re-identified.[11] Without doubt, the LEATPR Standards acknowledge these issues, and arguably cover them.[12] But as set forth in this article, big data distorts the traditional analysis and, thus, the LEATPR Standards may require a few modifications to be useful in the future.[13]

This article begins with a contestable (but defensible) premise: big data will revolutionize policing by offering new avenues to augment current investigation strategies. These new tactics, while having a real cost to privacy, liberty, and autonomy, will also result in more targeted and efficient investigations, and thus will become incredibly attractive to police.[14]

This article focuses on the distorting effects of big data policing. By distortion, I mean that traditional understandings, language, and categories may become blurred by the rise of big data.[15] Fortunately, the LEATPR Standards offer a mechanism to address some of these distorting effects, and with slight modification, can provide a clarifying lens to the problems arising from big data policing.

Part I of this article sets out the promise and problems of big data. Big data is revolutionizing policing, and this section explains the level and amount of data now available for law enforcement use. In simple terms, information about individuals is being catalogued in unprecedented ways, including government and corporate tracking of public records, consumer purchases, financial data, and even health data.[16] The aggregation of these different databases has the potential to allow personal dossiers to be created about each individual. Such commercial dossiers are valuable investigatory tools, and once created, generate real privacy concerns.[17] The promise of big data policing also means that in addition to continuing traditional investigatory database searches for known suspects, law enforcement will be able to use predictive analytics to discover unusual patterns that might signal criminal activity from currently unknown individuals. This section seeks to demonstrate that the potential for this type of investigatory

11. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1724 (2010); Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1877-78 (2011).

12. *See* LEATPR STANDARDS, *supra* note 1, at 2-5 (discussing the need for standards).

13. *See infra* Part II.

14. *See infra* Part I.

15. *See infra* Part II.

16. *See infra* Part I.

17. *Id.*

technique is too promising to resist and that the LEATPR Standards must address mass, anonymous surveillance in a more sophisticated manner.

Part II explores how the LEATPR Standards currently address access to these types of third party records. While the LEATPR Standards adequately govern the traditional law enforcement practice of searching established third party databases for identified records, they present some problems in the era of big data. This Part discusses three direct distortions of big data. First, I argue that the amount and interconnectedness of the available data weakens legal standards like "relevance," "reasonable suspicion," and "probable cause," which are predicated on a traditional model of limited, small data policing.[18] This critique is an analysis of how big data affects the Fourth Amendment, but as the terminology of LEATPR Standards derives from Fourth Amendment doctrine, this critique also implicates the Standards.[19] Second, I argue that the conception of "records" as envisioned in the LEATPR Standards becomes distorted in an era of blended, aggregated databases. Information is no longer siloed in particular identifiable third party institutions, but regularly sold, merged, and incorporated into even larger datasets.[20] In a merged dataset that includes highly private and nonprivate information, how does one know what level of justification is required to search? While the LEATPR Standards suggest defaulting to the highest level of protection based on the highest level of privacy,[21] this would prevent access to a significant amount of valuable data. Third, I consider the problems arising from mass surveillance searches for suspicious activity in de-identified records. These pattern matching searches raise concerns about how one can adequately protect de-identified data as well as larger issues of generalized mass surveillance.

---

18. This is a subject I address in detail in Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. (forthcoming 2015), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2394683.

19. *See* LEATPR STANDARDS, *supra* note 1, at 6-9 (discussing influence of the Fourth Amendment on the Standards).

20. Sarah Ludington, *Reining in the Data Traders: A Tort for the Misuse of Personal Information*, 66 MD. L. REV. 140, 142 (2006) ("[M]ost types of personal information—including names, birthdates, addresses, telephone numbers, clickstream data, travel details (flights, car rentals, hotels, train tickets) and transactional data (who bought what from whom, when, where, and how)—are unregulated, unless the data trader violates its own privacy policy, in which case the Federal Trade Commission (FTC) can hold the company accountable for unfair trade practices.").

21. *See* STANDARD 25-4.2 ("If a record contains different types of information, it should be afforded the level of protection appropriate for the most private type it contains.").

Part III attempts to identify possible solutions to these gaps, with a specific focus on smoothing the distortions of big data. The LEATPR Standards offer a valuable framework for analysis, and this section merely attempts to suggest some modifications to prepare the Standards for the future of big data. The solutions focus on modifications to language in the Standards, addressing each area of weakness discussed in Part II.

I. The Development of Big Data Policing

Like many evolving industries, law enforcement has recognized the promise of big data.[22] Police work involves gathering information about crimes and criminals, and big data offers a new tool to collect and analyze that information.[23] The ability to sort through vast datasets, identify particular people or suspicious patterns, and catalogue the information for future use offers new ways to track and prevent crime.[24] Much of our lives are being recorded through digital trails of information.[25] What we buy, what we read, where we go, and where we live, work, and play are being recorded by private companies.[26] Our interactions with government and public resources are being collected by public institutions.[27] The innovation of big data is the recognition that those disparate pieces of information can be aggregated and studied in mega databases. Powerful new computers, sophisticated algorithms, and cheap storage space have allowed massive

22. *See* Hoofnagle, *supra* note 5, at 595; Robert Block, *Requests for Corporate Data Multiply: Businesses Juggle Law-Enforcement Demands for Information About Customers, Suppliers*, WALL ST. J., May 20, 2006, at A4; Bob Sullivan, *Who's Buying Cell Phone Records Online? Cops*, MSNBC (June 20, 2006), http://www.msnbc.msn.com/id/12534959/.

23. Candice L. Kline, Comment, *Security Theater and Database-Driven Information Markets: A Case for an Omnibus U.S. Data Privacy Statute*, 39 U. TOL. L. REV. 443, 447 (2008); Andrea Peterson, *Your Location History Is Like a Fingerprint. And Cops Can Get it Without a Warrant*, WASH. POST, July 31, 2013, http://www.washingtonpost.com/blogs/the-switch/wp/2013/07/31/your-location-history-is-like-a-fingerprint-and-cops-can-get-it-without-a-warrant/; Glenn R. Simpson, *Big Brother-in-Law: If the FBI Hopes to Get the Goods on You, It May Ask ChoicePoint*, WALL ST. J., Apr. 13, 2001, at A1.

24. Steve Lohr, *Sizing Up Big Data*, N.Y. TIMES, June 20, 2013, at F1, *available at* http://bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/. *See generally* MANYIKA ET AL., *supra* note 4.

25. Hayley Tsukayama, *Alarm on Hill over iPhone Location Tracking*, WASH. POST, Apr. 22, 2011, at A13; Troy Wolverton, *iSpy: Apple's iPhones Can Track Users' Movements*, SAN JOSE MERCURY NEWS, Apr. 20, 2011, http://www.mercurynews.com/ci_17893676.

26. Lohr, *supra* note 24.

27. Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435, 442-43 (2008).

volumes of data to be useful for ordinary criminal investigations.[28] This section briefly sets out how big data will change policing, focusing on two particular aspects of the change: (1) aggregation and personalization of data collection, and (2) predictive analytics.

A. Aggregation and Personalization of Data Collection

Databases and data mining have been around for years.[29] Almost as soon as computers developed the capacity to store information, analysts have been seeking to use that information for their investigations.

> Data mining is the process of looking for new knowledge in existing data. The basic problem addressed by data mining is turning low-level data, usually too voluminous to understand, into higher forms (information or knowledge) that might be more compact (for example, a summary), more abstract (for example, a descriptive model), or more useful (for example, a predictive model).[30]

The move to big data is, thus, a change of degree, not kind, for investigators. But it is a significant change.[31]

In part, this change arises because the amount of data has continued to increase.[32] Every public record, criminal record, and financial record is collected by third party institutions. Direct marketers know things about

---

28. Tene & Polonetsky, *supra* note 4, at 240.

29. Cate, *supra* note 27, at 438 ("'Data mining' is defined in many different ways but is perhaps best understood as encompassing a wide spectrum of data-based activities ranging from 'subject-based' searches for information on specified individuals to 'pattern-based' searches for unusual or predetermined patterns of activities or relationships."); Slobogin, *supra* note 6, at 317; *see also* Christopher Slobogin, *Transactional Surveillance by the Government*, 75 MISS. L.J. 139, 144 (2005).

30. K. A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 1, 22 (2003).

31. *See* Anita L. Allen, *Privacy Law: Positive Theory and Normative Practice*, 126 HARV. L. REV. F. 241, 246 (2013) ("'Big Data' is a nickname for enterprises that collect, analyze, package, and sell data, even uninteresting-looking data, to reveal tastes, habits, personality, and market behavior. Big Data is challenging traditional privacies.").

32. Larry Port, *Disconnect from Tech*, LEGAL MGMT., Nov.-Dec. 2010, at 46, 49-50 ("Google records every click of every search result, your Linkedln and Facebook profiles, including who you associate with, what products you like, and what entertainment you enjoy . . . . If you read books on a Kindle, Amazon knows what books you're reading, what page you're on in those books, and what you've deemed important via your highlights and bookmarks. Any online retailer where you have an account knows what you've browsed and bought.").

our personal lives before our friends and families do.[33] This growing amount of data includes government, consumer, financial, health, and internet records created by us (as users) and recorded by others about us.[34]

The change is not just the volume, but also the interconnectedness of the information available in third party institutions.[35] Linking disparate data sources into one easily searchable source has profound implications for the ease of studying human activity (and criminality). New companies are creating new industries to buy, sell, and study our data.[36] These data aggregators purchase information from private third party institutions and public record holders to create sophisticated consumer datasets for marketing, insurance, and other purposes.[37]

> A byproduct of enhanced technological capabilities is the ease with which data can be populated, aggregated, and exchanged across an increasingly diverse set of corporate interests. These corporate interests span the economy and include retailers (Sears, Hallmark), pharmaceutical companies (Pfizer), technology firms (Microsoft, IBM), banks and financial services firms (Bank One, Bank of America), and automakers (GM, Toyota). Data brokerage companies, such as Acxiom and

33. Charles Duhigg, *Psst, You in Aisle 5*, N.Y. TIMES MAG., Feb. 19, 2012, at MM30, *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

34. Slobogin, *supra* note 29, at 145 ("[A]dvances in data warehousing and data exchange technology in the financial sector allow very easy access to a virtual cornucopia of transaction-related information that can reveal, among other things, 'what products or services you buy; what charities, political causes, or religious organizations you contribute to; . . . where, with whom, and when you travel; how you spend your leisure time; . . . whether you have unusual or dangerous hobbies; and even whether you participate in certain felonious activities.'" (internal citations omitted)).

35. Solove, *Access and Aggregation*, *supra* note 6, at 1185; Daniel J. Solove, *Data Mining and the Security-Liberty Debate*, 75 U. CHI. L. REV. 343, 343 (2008) [hereinafter *Data Mining*].

36. Nicolas P. Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UMKC L. REV. 385, 389 (2012) ("Big data is closely linked both literally and by its scale to the massive datasets compiled by well know [sic] data aggregators such as ChoicePoint or Acxiom. Those datasets often start by aggregating large (but not 'big') structured sets created by state, federal, and local governments, law enforcement, and financial institutions amongst others. Acxiom is reported to hold data on five-hundred million consumers with an average of 1500 data points per data subject.").

37. Allen, *supra* note 31, at 246; Lior Jacob Strahilevitz, *Reputation Nation: Law in an Era of Ubiquitous Personal Information*, 102 NW. U. L. REV. 1667, 1699, 1733 (2008).

> LexisNexis repackage, augment, and sell personal data on individuals to corporate and public sector clients.[38]

These private companies work with and sell information to law enforcement.[39] In fact, law enforcement is an avid user of these commercial data collections.[40] Adding to these private sources of information, the government's own organically developed data mining projects supplement this privately collected data.[41] The result is a valuable source of investigatory information which federal and state police have begun to use on a regular basis.[42]

The aggregation of information allows for the targeting of particular identified individuals and groups. For companies, the goal is an individualized dossier of information about particular persons, groups, and links among different persons.[43] Yet, this same dossier also offers clues to law enforcement seeking information about particular suspects. Since September 11, 2001, "[t]he DOJ, through the FBI, has been collecting telephone logs, banking records, and other personal information regarding thousands of Americans not only in connection with counterterrorism

---

38. Kline, *supra* note 23, at 447.

39. *See* Stephen Rushin, *The Judicial Response to Mass Police Surveillance*, 2011 U. ILL. J.L. TECH. & POL'Y 281, 288.

40. *See* Gerry Smith, *ATF Seeks 'Massive' Database for Faster Investigations*, HUFFINGTON POST, Apr. 8, 2013, http://www.huffingtonpost.com/2013/04/08/atf-database_n_3038271.html ("The federal agency tasked with regulating firearms wants a new weapon in its investigative arsenal: Big Data. The Bureau of Alcohol, Tobacco, Firearms and Explosives is seeking proposals for 'a massive online data repository system' that could allow agents to make faster connections between suspects' names, social security numbers, telephone numbers and utility bills . . . .").

41. Cate, *supra* note 27, at 457 ("There are information aggregation businesses in the private sector that already combine personal data from thousands of private-sector sources and public records. . . . These records are updated daily by a steady stream of incoming data. They provide a one-stop-shop for the government when it wants access to personal data, and most of the government's data mining initiatives depend on access to those data.").

42. *Id.* at 444 ("The FBI aggregates data from multiple databases into its Investigative Data Warehouse ('IDW'). According to press briefings given by the FBI in 2006, the IDW contains more than 659 million records, which come from 50 FBI and outside government agency sources. The system's data mining tools are so sophisticated that they can handle many variations in names and other data, including up to twenty-nine variants of birth dates. The 13,000 agents and analysts who use the system average one million queries a month."); Simpson, *supra* note 23, at A1.

43. Elspeth A. Brotherton, Comment, *Big Brother Gets a Makeover: Behavioral Targeting and the Third-Party Doctrine*, 61 EMORY L.J. 555, 562-63 (2012); Simmons, *supra* note 10, at 991.

efforts, but also in furtherance of ordinary law enforcement."[44] Public records have been digitized so our addresses, employment, criminal activities, and the like are available by a quick search using only a name or identifying number.

> Companies like Acxiom, Docusearch, ChoicePoint, and Oracle can provide the inquirer with a wide array of data about any of us, including basic demographic information, income, net worth, real property holdings, social security number, current and previous addresses, phone numbers and fax numbers, names of neighbors, driver records, license plate and VIN numbers, bankruptcy and debtor filings, employment, business and criminal records, bank account balances and activity, stock purchases, and credit card activity.[45]

Law enforcement can, thus, quite quickly pull up information on individuals from computers in the police station.[46] Creating a mosaic of public, consumer, and health information about criminal suspects is simply too useful for investigators not to take advantage of this new tool. As one investigator stated, "Imagine the ability to instantly take a security camera photograph from a bank robbery and match it using a facial recognition algorithm to a photograph in an out-of-state motor vehicle database, and then to link that person's name to a mobile phone from a private-sector marking database."[47] In the future, once these different sets of data are linked up, the result will be a very valuable integrated, individualized investigative dossier that raises obvious privacy concerns.[48]

---

44. Slobogin, *supra* note 6, at 319-20.

45. *Id.* at 320.

46. *See* Cate, *supra* note 27, at 442-43 ("The Federal Bureau of Investigation ('FBI') maintains extensive databases in its Criminal Justice Information Services Division ('CJISD') that collect data from, and supply data to, a wide array of public- and private-sector entities."); *see also The CJIS Division Turns 20*, CJIS LINK, Mar. 2012, at 2, *available at* http://www.fbi.gov/about-us/cjis/cjis-link/march-2012/the-cjis-division-turns-20 (noting that transactions with the FBI's CJISD National Crime Information Center totaled 2.7 billion searches in 2011).

47. Douglas Page, *Crime Fighting's Next Big Deal*, OFFICER.COM (Sept. 9, 2012), http://www.officer.com/article/10773317/crime-fightings-next-big-deal (quoting Philip Becnel, managing partner of Dinolt, Becnel & Wells Investigative Group).

48. Tene & Polonetsky, *supra* note 4, at 251 ("Big data poses big privacy risks. The harvesting of large sets of personal data and the use of state of the art analytics implicate growing privacy concerns.").

These public-private databases pale in comparison to what information technology companies are learning about us from the internet and mobile devices.[49] "Increasingly and of considerable importance going forward, big data comes from less structured sources including '[w]eb-browsing data trails, social network communications, sensor data and surveillance data.' Much of it is 'exhaust data,' or data created unintentionally as a byproduct of social networks, web searches, smartphones, and other online behaviors."[50] Google not only knows what you have bought, searched for, and viewed online, but also has the ability to figure out where you have been.[51] Of course, should third party institutions like Google partner with credit card companies to know where you shop,[52] police license plate readers to know where you drive,[53] social media to know your habits,[54] or commercial aggregators to know your consumer history,[55] a rather complete personal dossier with all of your personal preferences and patterns could be created. This information would then be potentially available to police investigating a particular person.[56]

---

49. Julia Angwin, *The Web's New Gold Mine: Your Secrets*, WALL ST. J., July 30, 2010, at W1.

50. Terry, *supra* note 36, at 389-90 (internal citations omitted).

51. Andrew William Bagley, *Don't Be Evil: The Fourth Amendment in the Age of Google, National Security and Digital Papers and Effects*, 21 ALB. L.J. SCI. & TECH. 153, 163-64 (2011).

52. MANYIKA ET AL., *supra* note 4, at 85 ("Globally in 2008, there were 90 billion to 100 billion such transactions off line linkable to [point of sale] devices. Law enforcement investigations regularly use such data to establish physical location.").

53. Rushin, *supra* note 39, at 285-86 ("[Automatic License Plate Recognition] systems not only flag passing cars that match a criminal database, but they also record the exact time and location of *all passing cars* into a searchable database, whether or not there is any evidence of wrongdoing. This data can be kept on file indefinitely. In communities with extensive, integrated networks of ALPR cameras, this could potentially amount to mass surveillance of an entire community.").

54. *See* MANYIKA ET AL., *supra* note 4, at 89-90 (recognizing the data available when we willingly join social networking programs, share geo-tagged photos, travel, use neighborhood guides, or a multitude of other everyday activities).

55. Kline, *supra* note 23, at 447-48.

56. *See, e.g.*, Editorial, *The End of Privacy?*, N.Y. TIMES, July 14, 2012, at SR10; *see also* Solove, *Access and Aggregation*, *supra* note 6, at 1138; Solove, *Data Mining*, *supra* note 35, at 343-44; Omer Tene, *What Google Knows: Privacy and Internet Search Engines*, 2008 UTAH L. REV. 1433, 1454; Andy Greenberg, *U.S. Government Requests for Google Users' Private Data Jump 37% in One Year*, FORBES (June 17, 2012), http://www.forbes.com/sites/andygreenberg/2012/06/17/u-s-government-requests-for-google-users-private-data -spike-37-in-one-year/.

The LEATPR Standards recognize the new world of data collection. But, the next stage—the future of big data policing—will be when each of those datasets are linked together into large commercial databases that upend any ability to categorize the content of the data. Traditional categories of "health records" or "financial records" will not be limited to a single dataset of that type of record, but blended with other types of information.

For purposes of this article, the aggregation of disparate databases collecting public and private information on individuals offers two issues to study. First, the types and sources of information included in these aggregated databases include all sorts of private and not-so-private information combined together. Isolating the level of privacy in a particular database might be possible in single use databases (phone records, financial records, health records), but becomes more difficult in an aggregated dataset that includes portions of all of these types of records. Second, commercial aggregators add a level of distance between the parties. As Joshua Simmons has written, some third party institutions are best thought of as "Fourth Parties" who have no relationship to the data except that they purchased it.[57] Unlike third parties who have some contractual relationship with the provider (i.e., phone company to phone consumer), these commercial purchasers of data possess the data simply as a commodity.[58] Finally, as a related concern, there is the technical reality that these Fourth Parties will soon be storing their data on cloud-computing systems hosted by yet another party (Fifth Parties?), which raises the question of whether this location also weakens the protections against law enforcement access.[59] After all, if police can ask the host for access to the information stored, why do they need to ask permission from the owner or custodian of the information?

B. Aggregation and Prediction

Aggregation of information also facilitates the creation of new prediction-based techniques to investigate crimes. Pattern matching algorithms that flag suspicious consumer purchases (fertilizer to build bombs, pseudoephedrine to make methamphetamine, etc.) allow police to spot (or prevent) crimes without any previous knowledge that the crime is

57. Simmons, *supra* note 10, at 990.

58. *Id.*

59. Joshua Gruenspecht, Note, *"Reasonable" Grand Jury Subpoenas: Asking for Information in the Age of Big Data*, 24 HARV. J.L. & TECH. 543, 548-49 (2011).

occurring.[60] Predictive analytics have already been used to determine areas where crime may occur,[61] but new predictive technologies will soon look to predict who will be committing those crimes.

Third party records also allow law enforcement to recognize new patterns of criminal activities in the data. Police can determine social connections or links, visualizing who a particular criminal's associates might be.[62] Police can create profiles of suspects by matching behavior patterns to repeated crimes[63] or offender networks.[64] Police can determine travel patterns and activities of known criminals.[65] Police can determine the location of particular types of crimes, narrowed to particular geographic areas.[66] This pattern matching thus shifts the focus of investigation from a reactive approach[67] to a more forward-thinking, predictive approach that is all based on the accumulated data.[68]

This type of predictive searching presents difficult issues that the LEATPR Standards will need to address. The first involves the practice of generalized pattern matching searches conducted without any particularized suspicion. As the LEATPR Standards suggest, such searches should only be allowed if the identifying information in the data is removed (or hidden).[69] This process of making the identifying data anonymous offers one level of protection. However, it is not a very fulsome protection, as de-

60. Erin Murphy, *Databases, Doctrine & Constitutional Criminal Procedure*, 37 FORDHAM URB. L.J. 803, 830 (2010).

61. Andrew Guthrie Ferguson, *Predictive Policing and Reasonable Suspicion*, 62 EMORY L. J. 259, 267 (2012).

62. Gareth Cook, *Software Helps Police Draw Crime Links*, BOS. GLOBE, July 17, 2003, at A1.

63. Vikas Grover, Richard Adderley, & Max Bramer, *Review of Current Crime Prediction Techniques*, *in* APPLICATIONS AND INNOVATIONS IN INTELLIGENT SYSTEMS XIV 233 (Richard Ellis et al. eds., 2007).

64. *Id.* ("Data is not just a record of crimes, it also contains valuable information that could be used to link crime scenes based on the modus operandi (MO) of the offender(s), suggest which offenders may be responsible for the crime and also identify those offenders who work in teams (offender networks) . . . .").

65. Murphy, *supra* note 60, at 830 ("But the use of databases to generate suspects represents a new kind of investigation altogether—whether based on particular information (e.g., 'who called this number') or upon predefined algorithms (e.g., 'who has traveled to these three countries and bought these two items within a one month period').").

66. Bernhard Warner, *Google Turns to Big Data to Unmask Human Traffickers*, BLOOMBERGBUSINESSWEEK (Apr. 10, 2013), http://www.businessweek.com/articles/2013-04-10/google-turns-to-big-data-to-unmask-human-traffickers.

67. Cook, *supra* note 62, at A1; Sullivan, *supra* note 22.

68. Block, *supra* note 22, at A4.

69. *See* STANDARD 25-5.6.

identified data can easily be re-identified (either directly by requesting the identity of the target or through a process of connecting the dots with other data).[70] In addition, as will be discussed in the next few sections, the traditional legal categories of protection (relevance, reasonable suspicion, probable cause) that might prevent police from re-identifying the individuals behind the suspicious patterns offer little protection in the era of big data.

C. The Big Data Lure

There is much more that could be said about how big data may affect law enforcement's relationship with third party records. The preceding brief summary seeks only to tease out some of the concerns that regulation of access to third party records must confront in a changing technological landscape.

Before moving on with my analysis of how the LEATPR Standards may be affected by big data policing, it is important to acknowledge the incredible promise that big data offers law enforcement. Big data is an important innovation because it offers novel solutions to age old problems. The move toward "smart-policing" or "data-driven policing" is not mere hype,[71] but also recognition that many traditional police techniques lacked empirical support.[72] In trusting the numbers, police departments have seen dramatic improvement in crime suppression.[73] Whether this improvement is a direct result of the use of data is still an open question, but the correlation certainly exists.[74]

---

70. *See infra* Part II.C.1.

71. Although, in truth, there may be some measure of hype with these new technologies. *See, e.g.*, Guy Adams, *LAPD's Sci-Fi Solution to Real Crime*, INDEPENDENT, Jan. 11, 2012, http://www.independent.co.uk/news/world/americas/lapds-scifi-solution-to-real-crime-6287800.html; Joel Rubin, *Stopping Crime Before It Starts*, L.A. TIMES, Aug. 21, 2010, http://articles.latimes.com/2010/aug/21/local/la-me-predictcrime-20100427-1.

72. Nina Cope, *'Intelligence Led Policing or Policing Led Intelligence?': Integrating Volume Crime Analysis into Policing*, 44 BRIT. J. CRIMINOLOGY 188, 191 (2004); DAVID ALAN SKLANSKY, THE PERSISTENT PULL OF POLICE PROFESSIONALISM 4 (Mar. 20011) (from a series of papers titled "New Perspectives in Policing," published on behalf of the John F. Kennedy School of Government Executive Session on Policing & Public Safety, and the National Institute of Justice.)

73. James J. Willis, Stephen D. Mastofsky & David Weisburd, *Making Sense of COMPSTAT: A Theory-Based Analysis of Organizational Change in Three Police Departments*, 41 LAW & SOC'Y REV. 147, 172 (2007).

74. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 3, at 70-72 (discussing how correlation may replace hypothesis in an era of big data).

In part, because this lure of big data is so great, those who study it must be vigilant in asking difficult questions. Police will benefit from access to the data, and thus, they will seek ways to access it. The LEATPR Standards acknowledge this reality and create a rather permissive process that generally allows law enforcement access. The next Part assesses whether the reality of big data, by distorting some of the traditional legal protections, weakens the Standards too much. Part III of this article will then seek to offer some suggestions in response to these concerns.

II. The LEATPR Standards and Big Data

The LEATPR Standards, of course, directly address the rise of centralized sources of personal information included in third party records. Necessarily, the "institutional third party" defined in Standard 25-1.1(e) contemplates private and corporate entities compiling personal data about individuals.[75] The question, however, is whether the existing Standards accurately speak to the new world of commercial big data, aggregated data, and the valuable information in de-identified data searches. This Part proceeds in three steps. First, it examines how the chosen terminology in the Standards may become distorted by the availability of big data. Second, it looks at the phenomenon of blended records that arise when corporate and other entities collect, aggregate, and merge various third party records into one large megadatabase. Third, it discusses how the LEATPR Standards might apply to large scale predictive pattern searches used to identify unknown suspects and even unknown crimes from large de-identified datasets. This Part attempts to show that big data has a distorting effect, altering both the strength of the categories of protection and the ability to access the records available. The point is not to criticize, but to refine the Standards in the face of these larger societal and technological changes brought on by big data.

A. The LEATPR Standards and Language

Central to the logic of the LEATPR Standards is the interrelation between the level of privacy associated with the categories of information[76] and the level of authorization needed to access that information.[77] The chosen levels of authorization mirror well-established Fourth Amendment and criminal procedure terms of art—relevancy, reasonable suspicion, and

75. *See* STANDARD 25-1.1(e).
76. *See* STANDARD 25-4.1.
77. *See* STANDARD 25-4.2.

probable cause. Yet, one question that must be asked is whether these traditional categories of suspicion become distorted in an era of big data. If so, the protection they seemingly (or traditionally) provide may not be sufficiently robust to protect individuals' private information.

In many cases, the LEATPR Standards provide a workable model to address basic law enforcement needs. For traditional investigations, the categories of information (Standard 25-4.1) and the categories of protection (Standard 25-4.2) will be fairly easy to analyze.[78] As set out in the examples section of the LEATPR Standards' Introduction,[79] if police are aware of a particular crime (a shooting in a park), their ability to search particular records of identified suspects (phone records, financial records) will turn on the level of privacy protection granted to those third party records.[80] In the park-shooting example, obtaining the phone records of the 9-1-1 caller, because of the "minimally private" nature of the call, would be permissible if supported by a statement that the evidence is relevant to an investigation.[81]

The key is the term "relevant," which brings up the first point of caution. The legal categories of protection—namely, the standards of relevancy, reasonable suspicion, and probable cause—become weakened in a world of big data. What I seek to point out is that the LEATPR Standards' chosen terminology, borrowed from Fourth Amendment doctrine, is less protective in application because the amount of aggregated, networked information now available distorts the analysis. The next three subparts explain how more information makes it easier to meet these legal thresholds, and thus, easier to justify access to the information that is sought.

1. Big Data Distortions of Relevancy

Relevance is understood as perhaps the lowest threshold to obtain information.[82] Under the Federal Rules of Evidence, evidence is relevant if "it has any tendency to make a fact more or less probable than it would be

78. *See* STANDARD 25-4.1, 25-4.2.

79. *See* LEATPR STANDARDS, *supra* note 1, at 11.

80. *See* STANDARD 25-5.2, 25-5.3.

81. *See* LEATPR STANDARDS, *supra* note 1, at 12-13 (discussing the privacy level of phone records in the 9-1-1 shooting hypothetical).

82. *See* New Jersey v. T.L.O., 469 U.S. 325, 345 (1985) ("[I]t is universally recognized that evidence, to be relevant to an inquiry, need not conclusively prove the ultimate fact in issue, but only have 'any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.'" (quoting FED. R. EVID. 401)).

without the evidence."[83] In practice, there is little required to obtain information under such a low threshold.[84] Courts routinely find information relevant. Most recently, in the high profile case involving the relevance of the National Security Agency's (NSA) access to volumes of telephone metadata, the reviewing court essentially held that the records were relevant because the government argued they were relevant.[85]

Relevance makes its appearance in Standard 25-5.2(a)(iii), authorizing access to a record based on "a judicial determination that the record is *relevant* to an investigation,"[86] or, in Standard 25-5.3(a)(iv), based on "a prosecutorial certification that the record is *relevant* to an investigation."[87] In both situations, the category of justification (relevance) covers "moderately protected information" or "minimally protected information," depending on whether the adopting jurisdiction chooses to require Standard 25-5.3(a)(ii), 25-5.3(a)(iii), or 25-5.3(a)(iv) as its guide.[88]

How does big data help expand the reach of relevancy? First, the sheer amount of personal information available to search in third party records presents new opportunities for police to expand their searches about suspects. There are simply more possible sources for which to search under a relevancy standard, because more information is available in big databases. If police believe someone is selling drugs, all sorts of things might be "relevant" to that suspicion: financial records, travel patterns, associates, substance abuse issues or treatment, consumer purchases, phone calls, etc. Many of these data points were simply not easily available to search before the advent of big data because they were not collected in widely accessible computer networks. In addition, there is a qualitative

---

83. FED. R. EVID. 401 ("Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action.").

84. *See, e.g.*, United States v. Sumner, 522 Fed. App'x 806, 810 (11th Cir. 2013) ("Relevance under the Federal Rules of Evidence is a low standard . . . .").

85. *See In re* Application of F.B.I. for an Order Requiring Production of Tangible Things From [Redacted], No. BR 13-109, 2013 WL 5307991 (Foreign Intelligence Surveillance Ct. Sept. 13, 2013) ("This Court recognizes that the concept of relevance here is in fact broad and amounts to a relatively low standard. Where there is no requirement for specific and articulable facts or materiality, the government may meet the standard under Section 215 if it can demonstrate reasonable grounds to believe that the information sought to be produced has some bearing on its investigations of the identified international terrorist organizations.").

86. STANDARD 25-5.3(a)(iii).

87. STANDARD 25-5.3(a)(iv).

88. *See* STANDARD 25-5.3(a).

change, as inference builds upon inference toward suspicion. Information creates links, clues, and suspicions that, even if focused on innocent correlations, might suggest criminal activity. If the target texted a known drug dealer, it could be an incriminating clue linking him to a conspiracy, or it could just be a contact with an old friend. If the target parked outside a known drug house, it could be an incriminating clue proving his involvement in delivering drugs, or just a coincidence of geography. But, under a relevance standard for investigation, these facts would be relevant to building a case.

For example, assume that in the park-shooting example, the 9-1-1 caller was identified by a telephone number through call records. Once identified, police could run the telephone number through national databases to find out a name, addresses, prior criminal records, prior arrests, and, most interestingly, if the phone number had previously been associated with gun violence.[89] These datasets are not private and are under the government's control. Further, with a name and the connection to the shooting, police might be able to request more detailed information about the 9-1-1 caller on relevancy grounds. A judge might sign off on a relevancy request to search databases that include minimally or moderately private information to see if the individual had any connection with the shooting (beyond being a witness). A judge might also allow police to request twenty-four hours of geolocational data tracking the witness' whereabouts.[90] Or, simply because of the type of crime at issue, certain searches of consumer purchases might be considered relevant to the investigation. Police might wish to access local companies' sales receipts or the witness' credit card receipts to see if the witness purchased the type of ammunition used in the shooting. Again, this type of search for minimally or moderately private information would be relevant to investigate the shooting, even if it only happened to reveal innocent information.

The questions get even harder in the context of mass searches. Continuing with the shooting example, assume police know that thirty people were present at the shooting. Presumably, if every one of those thirty potential witnesses has a cell phone or a smartphone, it might be possible to identify all of the phones that were in the park when the

89. *See National Crime Information Center (NCIC)*, FED'N OF AM. SCIENTISTS (June 2, 2008), http://www.fas.org/irp/agency/doj/fbi/is/ncic.htm.

90. Stephen E. Henderson, *Real-Time and Historic Location Surveillance After* United States v. Jones*: An Administrable, Mildly Mosaic Approach*, 103 J. CRIM. L. & CRIMINOLOGY, 803, 820 (2013) (applying the Standards to location data).

shooting occurred.[91] Under the Standard 25-4.1 categories, the identifying numbers and names would only be minimally private and thus available with a Standard 25-4.2 relevancy showing.[92] Therefore, without any evidence of criminal wrongdoing, the identity and whereabouts of thirty individuals will become known to investigating officers. After all, what these witnesses know (or do not know) is relevant to the investigation. This alone presents a slightly more concerning situation than the 9-1-1 caller who self-identified his whereabouts to the police. People who have no association with the crime will be tracked to a particular place at a particular time without any individualized suspicion of criminal wrongdoing.

Mass surveillance without evidence of particularized criminal wrongdoing certainly raises Fourth Amendment questions.[93] But, it also raises concerns under the LEATPR Standards. Does a person's mere presence in a park along with thirty other people at the time of a shooting make her personal information relevant to an investigation? Could police search law enforcement databases to find out if the telephone numbers correspond with a person in their police databases (or if the number is listed publicly)? Could law enforcement search to see if any of these witnesses have a criminal record? Could police search large data aggregators to find out if the names overlapped at all with the name of the victim (addresses/jobs/associations)? Could they search other phone contacts to see if these numbers provide any connection or motive to the shooting? The answer to each of these questions is likely yes under the LEATPR Standards (since none of these are highly protected data sets). Notice that by merely being proximate to a crime, the idea of relevancy has expanded to greater and greater access to personal information. Add to that the phenomenon of "confirmation bias," whereby individuals see what they expect to see, and police (unintentionally or intentionally) may create an

91. Tsukayama, *supra* note 25, at A13; Sullivan, *supra* note 22; Wolverton, *supra* note 25.

92. *See* STANDARD 25-4.1, 25.4-2.

93. *See* United States v. Garcia, 474 F.3d 994, 998 (7th Cir. 2007) ("Should government someday decide to institute programs of mass surveillance of vehicular movements, it will be time enough to decide whether the Fourth Amendment should be interpreted to treat such surveillance as a search."), *abrogated by* United States v. Jones, 132 S. Ct. 945 (2012).

ever-widening web of suspicion.[94] In the hunt for suspicious links, many more things will appear suspicious.[95]

One larger question for the LEATPR Standards might be whether the term "relevant" provides any real limitation at all. If relevance is the only limitation, it seems likely that in practice the minimally-moderately protected and not protected categories will merge into one category of readily accessible information. Simply stated, if police can conduct mass searches of location and identity based on any reported crime, then the relevancy standard is revealed to offer very little protection.

2. Big Data Distortions of Reasonable Suspicion

Reasonable suspicion is a well-established Fourth Amendment term of art, used in thousands of federal and state cases.[96] In the LEATPR Standards, it has been adopted as one of the types of authorization for moderately private records. Standard 25-5.2(a)(ii) requires "a judicial determination that there is reasonable suspicion to believe the information in the record contains or will lead to evidence of crime,"[97] and Standard 25-5.3(a)(ii) connects that requirement to moderately private records.[98]

The effect of big data is apparent in any analysis of reasonable suspicion. As I have explored in a separate article, reasonable suspicion is essentially a "small data doctrine."[99] From *Terry v. Ohio* onwards, reasonable suspicion has derived from cases involving police officers observing unknown suspects with little information about the suspect.[100] These are small data observations. And, as a doctrine built on cases involving unknown suspects with small data points about observable actions, the reasonable suspicion threshold makes sense. But, as more information about the suspect is provided to the officer, the easier the reasonable suspicion threshold is to

94. Barbara O'Brien, *Prime Suspect: An Examination of Factors that Aggravate and Counteract Confirmation Bias in Criminal Investigations*, 15 PSYCHOL. PUB. POL'Y & L. 315, 315 (2009).

95. United States v. Montero-Camargo, 208 F.3d 1122, 1143 (9th Cir. 2000) (Kozinski, J., concurring) ("Just as a man with a hammer sees every problem as a nail, so a man with a badge may see every corner of his beat as a high crime area.").

96. A Westlaw search of "reasonable suspicion" in the same sentence as "Fourth Amendment" returns over 10,000 results.

97. *See* STANDARD 25-5.2(a)(ii).

98. *See* STANDARD 25-5.3(a)(ii).

99. Ferguson, supra note 18.

100. *See* Terry v. Ohio, 392 U.S. 1, 24-25 (1968).

meet.[101] This becomes apparent by looking at how the Supreme Court has evaluated particularized information in the totality of circumstances analysis.[102] It can also be seen in how courts routinely find reasonable suspicion when more information about an identified suspect is added to the analysis.[103] And, almost universally, in cases with "known suspects" there is a finding of reasonable suspicion.[104] Simply put, the more information known about a suspect, the easier it is to justify a finding of reasonable suspicion.

The move from small data to big data can be significant. Big data—and the ability to know all sorts of personal details about the suspect—weakens the protections of reasonable suspicion. An otherwise innocent observation of a parked car in a motel lot with out-of-state license plates can turn into reasonable suspicion if the license plate identifies the car's owner as a suspected drug dealer who is listed in a database of known drug suspects.[105] Or, a man with a bag lurking outside a darkened home can give rise to reasonable suspicion if the suspect is in an area predicted to be burglarized.[106] The suspect has done nothing different, but the suspicion changes because the officer has been able to obtain more contextualizing information.[107]

---

101. As will be discussed, this is both because the information is particularized and individualized, mirroring the language in *Terry*, and because there is simply more information available.

102. *Compare* Florida v. J.L., 529 U.S. 266, 272 (2000) (holding that the information in an anonymous tip was not enough to create reasonable suspicion), *with* Alabama v. White, 496 U.S. 325, 330 (1990) (finding an anonymous tip, coupled with further police investigation, provided enough to create reasonable suspicion).

103. Ornelas v. United States, 517 U.S. 690, 695 (1996); United States v. Sokolow, 490 U.S. 1, 7 (1989).

104. *See, e.g.*, Commonwealth v. Calderon, 681 N.E.2d 1246, 1248 (Mass. App. Ct. 1997); State v. Gilchrist, 299 N.W.2d 913, 916 (Minn. 1980); State v. Valentine, 636 A.2d 505, 510-511 (N.J. 1994).

105. *See Ornelas*, 517 U.S. at 695.

106. Will Frampton, *With New Software, Norcross Police Practice Predictive Policing*, CBS ATLANTA (Aug. 19, 2013), http://www.cbsatlanta.com/story/23178208/with-new-software-norcross-police-utilize-predictive-policing.

107. Again the phenomenon of confirmation bias plays a role here. Keith A. Findley, *Innocents at Risk: Adversary Imbalance, Forensic Science, and the Search for Truth*, 38 SETON HALL L. REV. 893, 899 (2008) ("Confirmation bias means that police and prosecutors—as human beings—are likely, once they have identified a suspect or formed a theory of guilt, to seek confirming evidence and not seek disconfirming evidence. Accordingly, any ambiguous evidence is likely to be construed as incriminating, any incriminating evidence is likely to be viewed with heightened significance, and any inconsistent evidence is likely to be ignored or marginalized as insignificant or unreliable.").

In addition, the number of data points can affect reasonable suspicion. Similar to the quantification argument with relevancy, sometimes the sheer quantity of facts can satisfy the totality of circumstances test even if those data points are otherwise innocent.[108] Once an officer identifies the suspect, big data gives the officer access to a wealth of information with which to justify reasonable suspicion. The point here is simply that the choice of "reasonable suspicion" terminology may not be as protective as it seems.

Again, going back to our shooting example, to figure out if the 9-1-1 caller was involved in the shooting, police may want to consider possible motives for the crime. Police may wish to figure out if the 9-1-1 caller and the victim have any personal or business connections. To establish (or rule out) a financial motive to the shooting (robbery, bad debts, etc.), police may wish to access the moderately protected financial records of the 9-1-1 caller (or any of the other witnesses) to see if a sum of money had been transferred. To develop reasonable suspicion in a big data world, certain database searches could be conducted on mere relevancy grounds.[109] Law enforcement searches could identify past addresses, employment, or other available records. If any data showed a match between the victim and the suspect, this might establish a personal or business connection. Data searches into stored automobile license plate readers might reveal the overlapping movements of the 9-1-1 caller and the victim. If there was any geographical link, this might signify a personal connection. If police suspected that the 9-1-1 caller was the shooter, consumer searches into past ammunition or gun purchases would certainly satisfy the relevancy standard, though possibly implicate only innocent conduct. Aggregating these data points may well create reasonable suspicion to believe that other records may lead to evidence of a crime, which would allow even more invasive searches in other databases containing moderately private information.

Again, notice that the 9-1-1 caller has not done anything more than report a crime. Yet, police can develop (rightly or wrongly) reasonable suspicion based on connecting information (that may well be innocent). This is the reality of big data. The more information gathered, the easier it is to generate suspicion. As will be discussed, this weakness of the

108. The Supreme Court has never adopted a numbers approach to reasonable suspicion, although in cases like *United States v. Arvizu*, the sheer number of factors, even if each factor was itself innocent, was found to be sufficient to find reasonable suspicion. 534 U.S. 266, 277 (2002).

109. *See supra* Part II.A.1.

reasonable suspicion standard reveals a weakness in the LEATPR Standards that have adopted the same terminology.

3. Big Data Distortions for Probable Cause

Probable cause remains a constitutionally rooted threshold requirement that "exists where 'the facts and circumstances within their [the officers'] knowledge and of which they had reasonably trustworthy information [are] sufficient in themselves to warrant a man of reasonable caution in the belief that' an offense has been or is being committed."[110] Under Standard 25-5.2, "a judicial determination that there is probable cause to believe the information in the record contains or will lead to evidence of crime" is required for "highly protected information."[111]

Probable cause, as a term of art, may be the least affected by big data, because probable cause has always been a standard requiring significant information. While big data provides more tools to collect this information, and likely makes the standard easier to meet, it does not fundamentally change the analysis.

Take, for example, the seminal probable cause case *Illinois v. Gates*.[112] In *Gates*, police received an anonymous letter stating that Sue and Lance Gates were involved in distributing drugs.[113] Further, the letter provided details about the Gates' impending trip to Florida to retrieve the drugs and drive them back to Illinois.[114] Police officers followed Mr. Gates and discovered that many of the plans detailed in the anonymous tip had in fact occurred.[115] With this corroborated information, police officers requested a search warrant.[116] In finding that police observations corroborated the anonymous tip, the Supreme Court upheld the finding of probable cause under a totality of circumstances test.[117]

---

110. Brinegar v. United States, 338 U.S. 160, 175-76 (1949) (quoting Carroll v. United States, 267 U.S. 132, 162 (1925)); *see also* Florida v. Harris, 133 S. Ct. 1050, 1055 (2013) ("The test for probable cause is not reducible to 'precise definition or quantification.' 'Finely tuned standards such as proof beyond a reasonable doubt or by a preponderance of the evidence . . . have no place in the [probable-cause] decision.' All we have required is the kind of 'fair probability' on which 'reasonable and prudent [people,] not legal technicians, act.'" (internal citations omitted)).

111. *See* STANDARD 25-5.2(a)(i), 25-5.3(a)(i).

112. 462 U.S. 213 (1983).

113. *Id.* at 225.

114. *Id.*

115. *Id.* at 226.

116. *Id.*

117. *Id.* at 238.

Now, imagine *Gates* in an era of big data searches. With the reasonable suspicion provided by the tip, police could have obtained financial information that might show a discrepancy in the amount of money earned and the family's purchases.[118] Police could obtain the past flight history of Mr. Gates who was apparently repeating this travel pattern with some frequency.[119] Police could also obtain limited geolocational details of the Gates' car, identify the visitors to the Gates' Florida home (and determine whether they were known to be involved in the drug trade), or compare the phone numbers Gates called to known drug distributers, etc. These data searches could well replace the corroboration provided by following Mr. Gates to Florida and could be done rather efficiently with a computer at the police station. These facts would also likely support a claim of probable cause.

This is not to say that big data changes the traditional probable cause analysis, except in that it makes it easier to reach the probable cause threshold. New interconnected resources will be able to fill in the gaps of information and provide a seemingly stronger set of facts to base a finding of probable cause.[120] New information sources will alter the probabilities that criminal activity is occurring. Probable cause simply becomes more attainable with more information.

The expansion of information sources may well be a positive innovation, as it also likely means that the probable cause established in many cases will be stronger. The risk is merely that, just as in the reasonable suspicion analysis, quantity replaces quality under a totality of circumstances test.

These distortions in the terminology suggest a possible corrective solution—namely, alter the terminology in the Standards to reflect the accurate protective scope envisioned by the drafters. Probable cause, in fact, may be a more appropriate standard in a big data world that has eroded the justifications of lesser protections. As will be discussed in Part III, these changes need not be major, but may be necessary.

---

118. This financial information is likely moderately private information accessible under Standard 25-5.3(a)(ii).

119. This travel information may not be private at all. *See* Susan Stellin, *Security Check Now Starts Long Before You Fly*, N.Y. TIMES, Oct. 21, 2013, at A1 ("The Transportation Security Administration is expanding its screening of passengers before they arrive at the airport by searching a wide array of government and private databases that can include records like car registrations and employment information.").

120. *See, e.g.*, Max Minzner, *Putting Probability Back into Probable Cause*, 87 TEX. L. REV. 913, 913 (2009); Lawrence Rosenthal, *Probability, Probable Cause, and the Law of Unintended Consequences*, 87 TEX. L. REV. SEE ALSO 63, 63 (2009), http://www.texaslrev.com/wp-content/uploads/Rosenthal-87-TLRSA-63.pdf.

B. The LEATPR Standards and Aggregation of Records

Big data means big money because the information collected on consumers is valuable to companies.[121] One of the realities of big data is that unique datasets are being aggregated into massive commercial and public databases. This aggregation means that numerous types of highly private and nonprivate information are mingled in the same blended database.[122] This change has the potential to fundamentally alter how we think of third party records. In the future, centralized third party records filled with personalized information may exist as a commodity sold by data brokers making the records easily accessible to law enforcement and others who are willing to pay for the information.

In many ways, the future is now, as commercial enterprises are collecting everything from private health information and financial credit reports, to more public interactions with law enforcement and government institutions. Data brokers represent merely the beginning of companies seeking to make a profit from merging all sorts of private and public information.[123] Facebook, Amazon, and other social media and commercial sites are already capitalizing on the highly personal information they know about users by selling it to marketers.[124] This vast amount of information—easily searchable for all sorts of reasons—will only grow in sophistication.[125]

---

121. John Furrier, *Big Data Is Big Market and Big Business - $50 Billion Market by 2017*, FORBES (Feb. 17, 2012), http://www.forbes.com/sites/siliconangle/2012/02/17/big-data-is-big-market-big-business/ (announcing the Wikibon prediction that big data will be a $50 billion industry by the year 2017).

122. *See supra* Part I.A.

123. Lois Beckett, *Everything We Know About What Data Brokers Know About You*, PROPUBLICA (Sept. 13, 2013), http://www.propublica.org/article/everything-we-know-about-what-data-brokers-know-about-you; *see also* Robert Epstein, *Google's Gotcha: Fifteen Ways Google Monitors You*, U.S. NEWS & WORLD REPORT (May 10, 2013), http://www.usnews.com/opinion/articles/2013/05/10/15-ways-google-monitors-you ("Google uses your search history to send you personalized ads. That's how it survives, after all. About 97 percent of the company's revenues are from advertising.").

124. Kashmir Hill, *Facebook Joins Forces with Data Brokers to Gain More Intel About Users for Ads*, FORBES (Feb. 27, 2013), http://www.forbes.com/sites/kashmirhill/2013/02/27/facebook-joins-forces-with-data-brokers-to-gather-more-intel-about-users-for-ads/; Marcus Wohlsen, *Amazon's Next Big Business Is Selling You*, WIRED MAG. (Oct. 16, 2012), http://www.wired.com/business/2012/10/amazon-next-advertising-giant/.

125. Leslie Cauley, *Google's G1 Phone Makes It Easy to Track Surfing Habits*, USA TODAY, Feb. 10, 2009, http://usatoday30.usatoday.com/tech/wireless/phones/2009-02-08-google-g1-web-tracking-privacy_N.htm?csp=15; Epstein, *supra* note 123 ("Every search you conduct using Google's ubiquitous search engine – for medical or mental health information, an update on your favorite mayoral candidate, the schedule of your church's

Law enforcement can now purchase access to some commercial databases and aggregated datasets, and the resulting datasets which they can then search with a single query are becoming incredibly valuable for investigations.

This aggregation of many third party institutions' data, into what has been called "fourth party" aggregators,[126] distorts the Standards in two ways. First, it challenges the level of *privacy* that should be given to this aggregated megadatabase (which includes private and nonprivate information). Second, it challenges the level of *justification* needed for access to these aggregated, blended records. The LEATPR Standards recognize these problems.[127] They suggest: "If different types of content are commingled in a single record, as will often be the case, then the protection afforded to that record should be dictated by the most private type of information contained therein."[128] But, taken seriously, that means that records of many of these fourth party institutions might be off limits to law enforcement without the highest level of justification. Such a result is likely not the intended result of the drafters, but does arise because of the distortions of aggregated big data.

1. Blended Privacy in Records

The first issue is whether an aggregated, blended record (including highly private and nonprivate information) changes the level of privacy for the entire records dataset. In other words, does the fact that a highly private record is included in a larger database of records distort the level of privacy under the Standards? For purposes of this article, an aggregated, blended record is defined as the type of massive dataset collected by commercial companies whose central goal is to collect as much information about a person as possible.

There are two reasons why aggregation might change the privacy analysis. The first is that it might distort the meaning of "record." Remember, the key to the LEATPR Standards is that one must first determine the appropriate level of privacy for the records at issue.[129] But,

---

potluck dinner, how to handle kids' tantrums, the cure for halitosis or the latest sex toys – allows the company to track your interests and, over time, build a detailed dossier that describes virtually every aspect of your character, food preferences, religious beliefs, medical problems, sexual inclinations, parenting challenges, political leanings and so on.").

126. *See* Simmons, *supra* note 10.

127. *See* STANDARD 25-4.2(a) commentary.

128. *Id.*

129. STANDARD 25-4.2(a).

in an aggregated, blended dataset, what is the "record" for which you are determining privacy?

For example, if a specific highly private health checklist necessary to get life insurance is combined with other less private information relevant to the insurance company, is the "record" the health checklist or the entire personal file with all of the information (health, financial, personal) owned by the insurance company? If it is the former, then the privacy levels in the Standards work without much difficulty (the health checklist might be designated highly private, but other parts of the file would not be). But, if the "record" is the entire dataset on the individual (the insurance records), then the record might not be clearly highly private and may include nonprivate information. Said another way, if the individual's insurance records are combined with financial, personal, or consumer information in one massive, aggregated database, is a query into that database (a) a single search of a single record (i.e., the entire composite file on the person), or (b) several different searches of different types of records?

In terms of establishing what a "record" means, the LEATPR Standards offer the following definition: "A 'record' contains information, whether maintained in paper, electronic, or other form, that is linked, or is linkable through reasonable efforts, to an identifiable person."[130] This broad definition does not resolve the question of whether an aggregated, blended dataset is a single record or a series of records. The specific health checklist, the complete life insurance file, and the other personal information are all linkable to an identified person, and thus all would be considered records under Standard 25-1.1(g).

Of course, the term "records" is used hundreds of times throughout the LEATPR Standards and is understood to include both narrow and broad conceptions depending on the context. If this is correct, then blended, aggregated datasets held by big data companies can be considered the type of "record" covered by the Standards. But, the level of privacy in that single record that contains lots of different types of information is quite difficult to fathom. As will be discussed in the next section, without a clear idea of the level of privacy in the aggregated dataset, the level of justification police need to access that information is quite difficult to determine.

Second, there is the issue of commodification. If I give my personal information to my insurance company in order to get health insurance, does the private nature of this information change when it is sold to a series of

130. STANDARD 25-1.1(g).

larger data aggregators? Or, in other words, does the commodification of my personal data change the level of privacy in that data?[131] If dozens of drug companies not only know, but have purchased these highly private facts about my health, can I still claim that they are highly private?

The LEATPR Standards provide a four factor test to determine the level of privacy in records. This privacy level is determined by looking at whether

> (a) the initial transfer of such information to an institutional third party is reasonably necessary to participate meaningfully in society or commerce, or is socially beneficial, including to freedom of speech and association; (b) such information is personal, including the extent to which it is intimate and likely to cause embarrassment or stigma if disclosed, and whether outside of the initial transfer to an institutional third party it is typically disclosed only within one's close social network, if at all; (c) such information is accessible to and accessed by non-government person outside the institutional third party; and (d) existing law, including the law of privilege, restricts or allows access to and dissemination of such information or of comparable information.[132]

The first two factors are not affected by the commodification of data. The first factor focuses on the *initial transfer* of such information. By definition, subsequent transfers (to fourth parties) should not alter the analysis. The second factor seems to be referring to intimate information that is not usually disclosed outside a close social network by the individual person. The transfer of information to additional third or fourth party aggregators would not affect this factor.

The third and fourth factors, however, may be affected by the blending and selling of private data to commercial aggregators. If information is sold or merged into large databases, the information will be accessible and accessed by additional institutional third parties.[133] Thus, looking at the

131. Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 HARV. L. REV. 2055, 2057 (2004) ("[A] strong conception of personal data as a commodity is emerging in the United States, and individual Americans are already participating in the commodification of their personal data.").

132. *See* STANDARD 25-4.1(a)-(d).

133. Angwin, *supra* note 49, at W1 ("Hidden inside Ashley Hayes-Beaty's computer, a tiny file helps gather personal details about her, all to be put up for sale for a tenth of a penny. The file consists of a single code—4c812db292272995e5416a323e79bd37—that

third factor, the act of selling the information to a fourth party aggregator (or fifth party, etc.) will cut against the claim of privacy under the definition, even though the individual identified in or linked to the private information has not done anything differently. Perhaps this makes some sense. If for example, the information about a person's ill health (through insurance rates or premiums) is available in half a dozen commercial databases, then why should there not be a lessened sense of privacy?

The fourth factor may also be affected because the laws that prevent institutional third parties from revealing information do not always reach fourth parties.[134] Law enforcement agents may be restricted from obtaining consumer information directly from a drug store (without lawful authority), but they may be able to obtain the same information through a commercial aggregator (or direct marketer) that purchased the same information from the drug store. Currently, the only regulation on law enforcement is on the direct access of the information, and not the commercial purchase of the same information. Thus, the lack of law or regulation covering these big data collections may undermine a claim of privacy under the Standards.

Commodification, thus, seems to affect two of the factors for evaluating the level of privacy. As third party records become a more valuable commodity, this trading of information could affect the level of privacy in those records. Finally, the financial reality that data is a commodity seems to undercut the sense that this information has a strong claim to privacy. Perhaps this is an insight best saved for another forum, but underlying the debate over privacy is the deeper question: Whose data is it? If I buy goods from Apple and Amazon, is the data about those purchases mine or the companies'?[135] If Amazon sells this information, why can I claim any privacy in it? Perhaps, for business or reputation reasons, the company should not sell the information, but the question is *could* they? If I regularly telephone a depression hotline number, is that information private such that my phone company cannot sell my name to a marketer of antidepressants? Under the LEATPR Standards' factors, both types of information would have some type of privacy protection, but the analysis seems to ignore that the particular third parties are commercial businesses

---

secretly identifies her as a 26-year-old female in Nashville, Tenn. The code knows that her favorite movies include 'The Princess Bride,' '50 First Dates' and '10 Things I Hate About You.' It knows she enjoys the 'Sex and the City' series. It knows she browses entertainment news and likes to take quizzes.").

134. Simmons, *supra* note 10, at 951.

135. Larry Downes, *Privacy Panic Debate: Whose Data Is It?*, CNET (Apr. 27, 2011), http://news.cnet.com/8301-13578_3-20057682-38.html.

who have an interest in maximizing their profits. If it is the companies' data as well as mine, why can they not sell it to the highest bidder including law enforcement agencies that want to use the information for criminal investigations?

To be clear, these are not omissions in the Standards, but merely distortions that arise when personal information becomes a commodity and is blended by large big data companies owning multiple, integrated databases.

2. Blending Justifications

The second inquiry involves how to evaluate the level of justification needed to access a large, aggregated database. Police are not always looking for a discrete fact in a database, but many times would like the entire composite picture. If health insurance information (fairly private) is mixed with home purchases (not private), and financial records (some private, some not) are mixed with consumer purchases (some private, some not), how can a law enforcement agent calibrate the appropriate level of justification to search?

The Standards suggest that the default is to require the level of justification for the most private information in the database.[136] Under the LEATPR Standard 25-4.2(a), "[i]f a record contains different types of information, it should be afforded the level of protection appropriate for the most private type it contains."[137] Thus, if there was any highly private information in the dataset, the commercial aggregator could not be queried without a finding of probable cause. This solution is simple in theory, but difficult to apply, and quite restrictive to law enforcement.

First, the LEATPR Standards require law enforcement to know exactly what is in these large aggregated databases. This presents a few practical problems. The databases are proprietary, with companies unwilling to open their collections to public scrutiny.[138] In addition, these datasets are constantly evolving, with companies being purchased, new datasets being collected, and new features integrated into the product line.[139] There exists

136. STANDARD 25-4.2(a).

137. *Id.*

138. *See* Daniel J. Solove & Chris Jay Hoofnagle, *A Model Regime of Privacy Protection*, 2006 U. Ill. L. Rev. 357, 386-87 (2006) (describing the different types of private companies that sell commercial data, including those that maintain proprietary databases).

139. *See, e.g.*, Toby Anderson, *LexisNexis Owner Reed Elseveir Buys ChoicePoint*, USA TODAY, Feb. 21, 2008, http://usatoday30.usatoday.com/money/industries/2008-02-21-reed-choicepoint_N.htm.

a real question of how a law enforcement agent (or legislature) can go about determining the requisite level of justification for these blended datasets without knowing precisely what information exists in the datasets. The puzzle of these aggregated, blended datasets is that information is being collected so rapidly and being linked so easily that aggregators themselves could be offering highly protected information without even knowing it.

Second, by requiring the highest level of justification to access records (if any highly private information exists in the blended record), the Standards create a disincentive for companies to build these datasets. Big data businesses view accumulating and aggregating data as value added. The benefit of big data is mining these unexpected correlations that reveal patterns about particular people.[140] Requiring data to be siloed into specific records and identified by privacy labels would thwart the development of these companies (or at least prevent law enforcement from using them). Perversely, the greater the aggregation of data into larger and more helpful datasets, the harder it would be for police to access the information under Standard 25-4.2.

Finally, the practical effect of the default rule results in three suboptimal options for police. If a state adopted the Standards' rule for blended records, police might: (a) not search these blended records; (b) wait for probable cause to develop; or (c) claim some form of ignorance about what is in the datasets and address claims of privacy violations at a later time. None of these options cleanly allows police to gain the benefits of big data searches in the first instance. Further, the analysis of what is highly protected is so complex (even in an isolated and defined record), that police may be reluctant to use these otherwise helpful datasets.

Again, these problems are more a function of the nature of the big data environment than a fault of the LEATPR Standards, but they do make for a difficult application of the Standards to these large aggregated, blended datasets.

---

140. For example, knowing someone bought cigarettes might show they are a smoker. Knowing someone bought cigarettes at a local bar at 2:00 am, might show they smoke and drink alcohol. Knowing that they got into an accident at 3:00 am, might show that they crashed as a result of their evening activities. Knowing that the accident occurred in an area known for drug dealing and prostitution might suggest involvement in other illicit activities. Knowing they were treated at a health clinic on Monday morning might further the suspicion of their activities. Each fact alone is unrevealing, but together they demonstrate a pattern of dangerous activities that might be relevant to insurance companies, consumer marketers, and law enforcement.

C. The LEATPR Standards and Prediction

Law enforcement has never been simply reactive to crime.[141] Preventing crime, discovering ongoing crime, and deterring crime have always been a concern of police.[142] The rise of big data policing offers new tools to discover criminal patterns and thus solve crimes. And, critical to this change is access to third party information sources.

One such law enforcement innovation allows police to use pattern matching techniques to identify suspicious criminal activity.[143] For example, in order to manufacture illegal methamphetamine, dealers must purchase over the counter amphetamine products found in common cold medications. Tracking the sales of those cold medications from particular stores reveals who is buying the raw ingredients for a deadly drug.[144] This information about cold medicines is held by the third party institution (the drug store), and is both potentially relevant to a criminal investigation and yet also reveals private information about a person's health.

The LEATPR Standards directly address this concern in Standard 25-5.6, which involves de-identified records. Under that provision, police with an official certification can obtain de-identified records from third party institutions.[145] Thus, with the appropriate certification, police could obtain the sales of common cold medications at stores in a particular jurisdiction.

---

141. Craig S. Lerner, *Reasonable Suspicion and Mere Hunches*, 59 VAND. L. REV. 407, 437-39 (2006); Christopher Slobogin, *A World Without a Fourth Amendment*, 39 UCLA L. REV. 1, 39-41 (1991); Andrew E. Taslitz, *Fortune-Telling and the Fourth Amendment: Of Terrorism, Slippery Slopes, and Predicting the Future*, 58 RUTGERS L. REV. 195, 201 (2005).

142. *A Nat'l Interoperable Broadband Network for Pub. Safety: Recent Devs. Before H. Subcomm. on Commc'ns, Tech., & the Internet, Comm. on Energy & Commerce*, 111th Cong. 15 (2009) (statement of William Bratton, Chief of Police, L.A. Police Dep't) ("Very soon, we will be moving to a Predictive Policing model where, by studying real time crime patterns, we can anticipate where a crime is likely to occur.").

143. Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1, 4 (2005) ("Data mining's computerized sifting of personal characteristics and behaviors (sometimes called 'pattern matching') is a more thorough, regular, and extensive version of criminal profiling, which has become both more widespread and more controversial in recent years. Profiling varies in how it is conducted, but often focuses on features such as age, gender, and race or ethnicity, sometimes coupled with behavior.").

144. Jon Bardin, *Kentucky Study Links Pseudophedrine Sales, Meth Busts*, L.A. TIMES, Oct. 16, 2012, http://articles.latimes.com/2012/oct/16/news/la-heb-kentucky-counties-pseudophedrine-meth-busts-20121016 ("Using that data, researchers were able to determine how much of the drug was sold in each Kentucky county and compare it with the number of meth busts in local police logs. . . . In any given county, an increase in pseudophedrine sales of 13 grams per 100 people translated to an additional meth lab busted.").

145. STANDARD 25-5.6.

The Standards further require that finding the identity of a de-identified record requires additional authorization under Standard 25-5.3.[146] Thus, to discover the name of the person who purchased the cold medicine, the police would be required to meet the appropriate category of justification laid out in the Standards based on the level of privacy for the type of information sought.

The problem with Standard 25-5.6 is that while protective of privacy in theory, it really offers little protection in practice. Much has already been written about the straight-forward technological problems of keeping de-identified records anonymous.[147] Data scholars and curious individuals have taken to re-identifying previously de-identified data to demonstrate the lack of protections.[148] Well-known companies such as Netflix and AOL have seen de-identified user information re-identified, generating unflattering news coverage and lawsuits.[149] Health records have received

146. STANDARD 25-5.3.

147. *E.g.*, Schwartz & Solove, *supra* note 11, at 1854-55 ("In behavioral marketing, companies generally do not track individuals by name. Instead, they use software to build personal profiles that exclude this item but that contain a wealth of details about each individual. In lieu of a name, these personal profiles are associated with a single alphanumerical code that is placed on an individual's computer to track their activity. In one reported case, for example, the tracking file consisted of this string: '4c812db292272995e5416a323e79bd37.' These codes are used to decide which advertisements people see, as well as the kinds of products that are offered to them." (citing Angwin, *supra* note 49, at W1)).

148. Tene & Polonetsky, *supra* note 4, at 257 ("[O]ver the past few years, computer scientists have repeatedly shown that even anonymized data can typically be re-identified and associated with specific individuals."); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1141 (2013) ("The concept of k-anonymity originated with the work of Latanya Sweeney, who demonstrated, rather vividly, that birth date, zip code, and sex are enough to uniquely identify much of the U.S. population." (citing Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYS. 557, 558 (2002))).

149. *See, e.g.*, Wu, *supra* note 148, at 1118-20 (describing the Netflix re-identification problem and discussing the research of those who did the re-identification, including Arvind Narayanan and Vitaly Shmatikov); Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, at A1 ("[S]earch by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for 'landscapers in Lilburn, Ga,' several people with the last name Arnold and 'homes sold in shadow lake subdivision gwinnett county georgia.' It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. 'Those are my searches,' she said, after a reporter read part of the list to her.").

the most attention, as the social utility of studying disease has been compromised by the ease with which the health data can be re-identified.[150]

These stories and problems are acknowledged if not satisfactorily resolved in the LEATPR Standards. The Commentary on Standard 25-5.6 explicitly recognizes the dangers of de-identified data.[151] The Standards reference Professor Paul Ohm's resistance to the idea that de-identified data can ever be protected in a world of expanding, aggregated information sources.[152] While other scholars have countered this pessimism by proposing methods to protect de-identified data,[153] importantly the academic debate has not centered on law enforcement access to these records.

In this way, the Standards may not fully address the dangers of police access to de-identified data. Unlike academic researchers attempting to test the vulnerabilities of de-identified data, law enforcement analysts seeking to re-identify information have a particular and urgent goal in mind: solving a crime. If a pattern emerges suggesting criminal activity in a de-identified

---

150. *See* Brief of Amicus Curiae Electronic Frontier Foundation in Support of Petitioners at 12, Sorrell v. IMS Health Inc., 131 S. Ct. 2653 (2011) (No. 10-779) ("The PI data at issue in this case presents grave re-identification issues."); Brief of Amici Curiae Electronic Privacy Information Center (EPIC) and Legal Scholars and Technical Experts in Support of the Petitioners at 24, *Sorrell*, 131 S. Ct. 2653 (No. 10-779) ("Patient [r]ecords are [a]t [r]isk of [b]eing [r]eidentified."); Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 33, 37 (2010) ("Personal information that no longer contains overt identifiers (name, identification number, e-mail address, telephone number) can still be linked with known individuals. Identity can be ascertained from simple, basic, widely available non-unique identifiers (sometimes called *quasi-identifiers*).").

151. STANDARD 25-5.6 commentary.

152. STANDARD 25-1.1(g) commentary; *see also* Ohm, *supra* note 11, at 1746 ("The accretion problem is this: Once an adversary has linked two anonymized databases together, he can add the newly linked data to his collection of outside information and use it to help unlock other anonymized databases. Success breeds further success."); Schwartz & Solove, *supra* note 11, at 1847 ("In sum, whether information can be re-identified depends on technology and corporate practices that permit the linking of de-identified data with already-identified data. Moreover, as additional pieces of identified data become available, it becomes easier to link them to de-identified data because there are likely to be more data elements in common.").

153. *E.g.*, Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1427-28 (2012) (proposing a differential privacy theory); Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, PLOS ONE, Dec. 2011, *available at* http://www.plosone.org/article/info%3Adoi%C2F10.1371%2Fjournal.pone.0028071; Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

source, then police have the ability (if not a duty) to investigate by trying to match other patterns from other data sources to identify the individual. All of the methods of re-identification and all of the government resources can be brought to the case. With enough effort, police should be able to match or surpass the abilities of academic researchers to re-identify data.

In addition, in many cases, because of the other available third party data, police will not need to take the second step under the LEATPR Standards for direct re-identification. Simply stated, once a suspicious pattern has been identified (be it unusual purchases, suspicious comments, or criminal associates), police can use *indirect* re-identification; police have enough other investigatory resources to re-identify the data without relying on the data itself. For example, in our methamphetamine purchases example, police know from de-identified data that a particular person bought cold medicine from a particular store, at a particular time, and on a particular day. While police could go through the direct re-identification process under the Standards to get the identity of the purchaser, police could also review store security tapes, speak with witnesses, track license plate reader records from the stores, or obtain other surveillance tapes for each of the times in question.[154] Matching the images of who was at the stores at the correct time would also identify the suspect. In addition, police could request the cell phone records of all the people in that store location at the time. If a particular phone was in the store at the time of each of the purchases, police could use phone data to identify the purchaser.[155] Then, police could either search police databases for the phone number, or if necessary, ask for a court order now that reasonable suspicion has been generated (since only one phone number matched the time and place of each of the suspicious purchases).

The de-identified data offers other clues as well. With a phone number (even de-identified), police can see who else called (or was called by) that number, thus generating a network of connections for individuals involved

154. Don Babwin, *Chicago Video Surveillance Gets Smarter*, USA TODAY, Sept. 27, 2007, http://usatoday30.usatoday.com/news/nation/2007-09-27-4171345706_x.htm; Cara Buckley, *Police Plan Web of Surveillance for Downtown*, N.Y. TIMES, July 9, 2007, at A0; John Del Signore, *NYPD Tightens Surveillance in Subway's "Ring of Steel"*, GOTHAMIST (Sept. 21, 2010), http://gothamist.com/2010/09/21/nypd_tightens_surveillance_in_subwa.php; David Gambacorta & Morgan Zalot, *Surveillance Cameras Prove Helpful in Crime Probes*, PHILA. INQUIRER, Feb. 1, 2013, http://articles.philly.com/2013-02-01/news/36661895_1_surveillance-cameras-surveillance-network-high-crime-areas; Somini Sengupta, *Privacy Fears Grow as Cities Increase Surveillance*, N.Y. TIMES, Oct. 13, 2013, at A1.

155. *See* Henderson, *supra* note 90, at 805-06 (discussing how the "high country bandits" were apprehended using cell phone surveillance techniques).

in possible illegal activity. Those phone numbers (still de-identified) can then be run through law enforcement databases and re-identified to generate names and addresses. Again, the identifying number allows police to search for all connections to that number, creating a matrix of the suspect's associates. The combination of direct re-identification technology and indirect re-identification through other law enforcement techniques means that much de-identified data can be re-identified by law enforcement. While this is not a negative result, it may call into question the protections built within the Standards, which seem to assume that de-identification of data is a substantial protection.

III. Solutions: Smoothing the Distortions

This Part offers possible solutions to the problems raised in this article. It focuses on three major questions discussed in Part II: (1) the problem of terminology in the Standards, (2) the problem of aggregation of blended records, and (3) the problem of de-identification.

A. Problem of Terminology

As discussed in Part II.A., the Standard's chosen terminology offers little protection in a world of big data. Specifically, relevance and reasonable suspicion, two already weak standards, are easily surmounted by the availability of personalized information that provides a vast quantity (if not quality) of data points and particularized information about a suspect.

The solution, while perhaps frustrating to the drafters of the Standards (who no doubt thought carefully about the chosen language), seeks to ratchet up the protection by changing the terminology. Essentially, I propose a one level shift, increasing the level of protection by one order. Thus, I would propose the following standards to replace the existing terminology: For nonprivate information held by third parties, I would require a relevance standard. For minimally private information, I would require reasonable suspicion. For moderately private information, I would require probable cause to believe the information in the record *will lead to evidence of a crime*. For highly private information, I would require probable cause to believe that the information in the record *will reveal criminal activity*.

These suggestions leave much of the Standards untouched. The determination of the level of privacy of the information remains the same. The equivalence—central to the Standards (in terms of a relation between level of privacy and level of justification)—remains intact (even though the justifications themselves have been ratcheted up one degree). And the

rationale for the levels of justification remains largely unchanged. The more specific arguments for the changed standard are addressed in turn.

Not Private Information. Under the Standards, no justification is needed to access nonprivate information. Under the proposed changes, a relevance justification would be required similar to Standard 25-5.3(b). This would mean that police could not simply vacuum up all personal data, even if other private companies could do so. This change requires police to justify why the information was relevant to an investigation. Of course, as the relevance standard is quite low, this burden should not be difficult to overcome. If the information is not relevant to a police investigation, police should not be collecting it for investigatory purposes.[156] However, similar to Standard 25-5.3(b), this relevance justification could be met by a prosecutorial or agency subpoena without a court order.

Minimally Private Information. Under the current Standards, a relevance subpoena under Standard 25-5.2(b) is all that is required to obtain minimally private information.[157] Under the proposed change, reasonable suspicion via a judicial order would be required. Again, as demonstrated, because reasonable suspicion in a big data world is such an easily surmountable standard, this is not a significant burden. Minimally private information is still quite revealing and to justify access to phone contact records or the like, police should have to determine that there is some particularized and individualized reason to obtain the information. Otherwise, fishing expeditions for data will result on a mass scale.

Moderately Private Information. Under the current Standards, moderately private information can be obtained in one of three ways: (1) a judicial determination that there is reasonable suspicion to believe the information in the record contains or will lead to evidence of a crime[158]; (2)

156. Such a suggestion runs counter to the well-established tradition that law enforcement should have access to the same records that ordinary citizens have access to on a regular basis. For example, if a citizen can search my utility records in a jurisdiction that allows public access to such records, then the argument goes, police should have similar access. *See* STANDARD 25-4.1(c) commentary. The argument presented above would require police to meet a relevance requirement before accessing those same records, thus imposing an additional barrier to access. While recognizing the imposition, the proposal above acknowledges that governmental access to personal data is different than individual access to that same data. The government simply has more power than an individual, and thus checks should be built to inhibit governmental conduct as distinguished from individual access.

157. STANDARD 25-5.3(a)(iii).

158. STANDARD 25-5.2(a)(ii).

a judicial determination that the record is relevant to an investigation[159]; or (3) a prosecutorial certification that the record is relevant to an investigation.[160] These are rather low standards of protection. Under my proposed change, these options would be replaced with a modified probable cause standard—"probable cause that the record *will lead to evidence of a crime*." This language comes directly from the current Standard 25-5.2(a)(ii). Notice that the language of the Standards does not require probable cause that "an offense has been or is being committed,"[161] but only that the record "will lead to evidence of crime."[162] The Standards, thus, adopt a rather indirect requirement,[163] which would allow judges to sign judicial orders several steps removed from actually uncovering the crime. This modified probable cause standard could be adopted to replace 25-5.2(a)(ii)-(iv).

Justifications for this change include that (1) the current standards are quite weak for moderately private information, and (2) the proposed language still provides a great deal of flexibility. Many things will lead to evidence of a crime that may not be criminal themselves. A review of bank transactions may suggest criminal activity without being criminal themselves. It is illegal to steal money, but not to deposit stolen money. Similarly, long term GPS surveillance may lead to a connection with drug dealing (as in *United States v. Jones*),[164] but may not itself demonstrate overt criminal activity. By emphasizing the "will lead to" language, this proposed change ensures a measure of protection while still allowing police to obtain necessary information.

Highly Private Information. Under the current Standard 25-5.2(a)(i), a judicial determination that there is probable cause to believe the information in the record contains or will lead to evidence of a crime is required to obtain highly private information.[165] Under the proposed

159. STANDARD 25-5.2(a)(iii).

160. STANDARD 25-5.2(a)(iv).

161. Brinegar v. United States, 338 U.S. 160, 175-76 (1949).

162. STANDARD 25-5.2(a)(i).

163. The "will lead to evidence" language may however be interpreted to be a tougher standard than this author believes. As Professor Slobogin mentioned at the *Oklahoma Law Review* Symposium, Nov. 15, 2013, the language could be interpreted to require a high standard of proof. "Will" does not equate with "may" and thus "will lead to" could be understood to require a high level of justification. *See also* 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, 741 (2014).

164. 132 S. Ct. 945, 947-48 (2012).

165. STANDARD 25-5.2(a)(i).

change, the probable cause language would really be a probable cause standard—not that there is probable cause that the evidence will lead to evidence of a crime, but probable cause that the records will *reveal criminal activity*. Specifically, the records must provide evidence of past or ongoing criminal activity or be usable in a prosecution for an identifiable criminal activity. For highly private information, the standards should be equally high.

The above suggestions, again, only seek a slight change in the LEATPR Standards to counteract the effect of big data. The chosen terminology seeks to raise the level of protection without disturbing the underlying logic and proportionality reasoning of the existing Standards.

B. Problem of Aggregation

As discussed in Part II.B., the problem of blended records means that there are differing levels of privacy in these large aggregated datasets, which can no longer be identified by a single category of information. This reality distorts the level of privacy and confuses the level of justification needed to access the records.

The solution proposed runs counter to the default rule suggested in the LEAPTR Standards, but is more consistent with the big data environment. It acknowledges the reality that police do not necessarily know what information is included in the blended datasets, and thus, police cannot determine what level of privacy is required. This confusion will result in officers either waiting until they generate the highest level of suspicion (on the chance that there might be highly private information), or forgoing these queries into aggregated, blended datasets.[166]

The solution proposed allows broad access into these mixed datasets, but then requires police to establish minimization processes to guard against revelation of highly private information. Minimization is a well-established concept in surveillance law.[167] Essentially, investigators are required to

166. *See supra* Part II.B.

167. *E.g.*, Stephanie K. Pell & Christopher Soghoian, *Can You See Me Now?: Toward Reasonable Standards for Law Enforcement Access to Location Data that Congress Could Enact*, 27 BERKELEY TECH. L.J. 117, 184 (2012) ("Minimization requirements are not a new idea. They already play a privacy protective role in several other surveillance statutes, including the Wiretap Act, the USA PATRIOT Improvement and Reauthorization Act of 2005 ('PATRIOT Act'), and the Foreign Intelligence Surveillance Act ('FISA')."); Tene & Polonetsky, *supra* note 4, at 259 ("Through various iterations and formulations, data minimization has remained a fundamental principle of privacy law. Organizations are required to limit the collection of personal data to the minimum extent necessary to obtain their legitimate goals. Moreover, they are required to delete data that is no longer used for

ignore or shield information that they acquire that does not fit the categories of information they are justified in collecting.[168] In addition, investigators are generally prohibited from sharing or using that information.[169] Minimization, thus, protects from unintentional disclosures and limits the use of the information for investigative purposes. In wiretap surveillance, police must minimize content from individuals not identified in the wiretap warrant.[170] In national security surveillance of overseas telephone calls, investigators must minimize content from U.S. citizens.[171] If information is uncovered that is unrelated to the targeted justification, investigators may not use the information (subject to some exceptions, including a law enforcement exception).[172] Minimization thus allows an overbroad search, with carefully designated protections to limit the information revealed.

It must be acknowledged that traditional minimization analysis does not neatly map onto the blended records problem. In a traditional minimization

---

the purposes for which they were collected and to implement restrictive policies with respect to the retention of personal data in identifiable form.").

168. Wu, *supra* note 148, at 1173 ("Data minimization provides that 'organizations should only collect PII ('Personally Identifiable Information') that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s).'" (quoting *National Strategy for Trusted Identities in Cyberspace*, WHITE HOUSE, Apr. 2011, at 45)).

169. Peter Swire, *Social Networks, Privacy, and Freedom of Association: Data Protection vs. Data Empowerment*, 90 N.C. L. REV. 1371, 1413 (2012) ("Data minimization posits that holders of personal information should minimize the collection and use of personal information to protect privacy rights.").

170. *Id.* ("Data minimization is an important principle in wiretap law, where the state gains lawful access to the relevant conversations, but should not use the existence of the wiretap to trawl through the rest of the conversations on a phone line."); *see also* 18 U.S.C. § 2518(5) (2012) (requiring that wiretaps "be conducted in such a way as to minimize the interception of communications not otherwise subject to interception").

171. The Foreign Intelligence Surveillance Act, Pub. L. No. 95-511, 92 Stat. 1783 (1978) (codified at 50 U.S.C. §§ 1801-1811 (1982)), requires "minimization" protocols to limit the collection, retention, and dissemination of information relating to United States citizens. *See* 50 U.S.C. §§ 1801(h), 1804(a)(5), 1805(a)(4).

172. *See* 50 U.S.C. § 1805(a)(3) (requiring minimization procedures); *id.* § 1801(h) (defining FISA minimization procedures); *id.* § 1801(h)(3) (detailing law enforcement exception); *see also* Stephanie Cooper Blum, *What Really Is at Stake with the FISA Amendments Act of 2008 and Ideas for Future Surveillance Reform*, 18 B.U. PUB. INT. L.J. 269, 302 (2009) ("While minimization procedures are supposed to prevent the retention and dissemination of information that is not related to foreign intelligence, there are notable exceptions. Under the minimization procedures, 'information that is evidence of a crime which has been, is being, or is about to be committed' can 'be retained or disseminated for law enforcement purposes.'").

situation, the targeted individual guides the limitation.[173] A warrant might allow for all phone calls of a particular person to be recorded, with the understanding that other callers from that phone line would be protected by minimization standards. Similarly, foreign nationals may be targeted for telephone surveillance with the understanding that U.S. citizens in communication with those individuals will be protected by minimization. Both of these hypotheticals share the commonality that the level of privacy is irrelevant when it comes to the targeted individual. Police can obtain both highly private and nonprivate information about the target. Only information about others (not the target) must be minimized.

In the blended records context, one may also have this problem of revealing information about nontargeted individuals, but it is not the main concern. The main problem is that with broad access to blended records police will see highly private information about the target without the appropriate justification. In addition, police may be tempted to use that information to support their investigation.

My minimization proposal seeks to address these concerns, but candidly only addresses the second issue of *use*. I would suggest a minimization process that mirrors the categories of protection in the LEATPR Standard 25-4.2.[174] For aggregated, blended datasets, police would be allowed to access the dataset without the highest level of justification, but only able to retrieve and use information for the level of justification they had authority to access. All other information would be minimized and not available for use as the basis of a justification for further investigation.

For example, assume police suspect an individual of drug dealing. Assume they only have reasonable suspicion (not probable cause), preventing them from obtaining highly private information. Police have access to a large commercial "big data" database that has aggregated and blended a host of available personal, financial, public, consumer, and health data. Some of this data is highly private; some is not private. Under the current LEATPR Standards, because there is some highly private information, police would need probable cause to access the records. Under my proposed modification, police would be able to search this data without probable cause, but only be able to use information that fell within the level of justification they possessed (reasonable suspicion). Thus, in the

173. United States v. Hoffman, 832 F.2d 1299, 1307 (1st Cir. 1987) ("This minimization requirement spotlights the interest in confining intrusions as narrowly as possible so as not to trench impermissibly upon the personal lives and privacy of wiretap targets and those who, often innocently, come into contact with such suspects.").

174. *See* STANDARD 25-4-2(a).

search through the database, should highly private information about substance abuse, mental illness, or the like be revealed, police would be precluded from using that information in their investigation. Of course, the police officers who conducted the search would be aware of the private facts, but protocols could be created to keep this information confidential.

Such a minimization process would obviously turn on drafting appropriate protocols for protection. These protocols would need to specify with clarity the types of information that would be allowed under each justification. Such a categorization would be immensely difficult as a legislature would have to determine *ex ante* what type of content would fit in each category. In addition, the police officer would also have to be able to determine with clarity what level of privacy the information contained. While perhaps the database companies themselves could use technology to sort, categorize, and code the level of privacy (making access to highly private information more difficult), the categorization process will be contested, contingent, and likely confused.

The protocols would, however, precisely determine the limitations on use of this minimized data. As with other minimization protocols, limiting the use and dissemination of the inappropriately obtained information provides a significant protection. Usually, knowledge about a target (even highly private knowledge) has less of an impact than using that knowledge to further an investigation. Protocols designed to restrict law enforcement use of the data could thus be effective protections within aggregated, blended datasets.

C. Problem of De-Identification

As discussed in Part II.C., the problem with de-identification involves the ease with which such data can be re-identified through direct technological or indirect third party surveillance means. One solution is to allow access to de-identified records under Standard 25-5.6 only when there is no chance that the information can be re-identified.

Currently, such a guarantee that de-identified data will remain anonymous is technologically impossible, meaning that in practice de-identified records are not really de-identified and should be recognized as such. Certainly, some companies have taken steps to reduce the possibility that de-identified information will be re-identified. As one example, companies like StreetLight Data, which use de-identified GPS information to track human activity, have developed privacy policies which require that

all identifiable information be removed.[175] The policies further require encrypted identifiers with no access to the decryption algorithms as well as aggregation into groups of fifteen or more so no individualized information is used.[176] But, many times the information police wish to access in the de-identified data will not be useful in that protected format.

Thus, accepting the current technological reality, the de-identified records mentioned in Standard 25-5.6 are really misnamed. Police requesting these records have access to re-identification procedures; therefore the records are merely disguised, not protected. This reality should caution legislatures from adopting Standard 25-5.6 without additional protections.

One protection would be to allow access to de-identified data only when police can provide the necessary level of justification for the type of privacy in the records. Whereas before police had unlimited access to the de-identified records, but had to justify (based on the appropriate level of privacy) particular access within those records, one proposal would be to require the same level of justification to do the initial search. Thus, to access the de-identified drug store sales records, police would need to demonstrate the appropriate justification (i.e., relevance, reasonable suspicion) associated with the drug store records. Then, to get additional access to re-identify a particular record, police would again need to demonstrate that same level of justification, but as to a particular record. This proposal, of course, would be restrictive to law enforcement, precluding many of the pattern-matching searches that look for anomalies in the data without any suspicion at all.[177]

Another solution would be to adopt the current Standards, but simply ban police from using other indirect methods to de-identify the data. If police wish to re-identify the person, the only recourse would be to use the protocols in the Standards. Such a solution, preventing law enforcement from using available (and traditional) techniques to re-identify suspects has

175. *See Privacy Policy*, STREETLIGHT DATA, http://streetlightdata.com/privacy/ (last visited Mar. 18, 2014).

176. *Id.*

177. As a practical matter, this protection might allow police to conduct de-identified searches on the sale of pseudoephedrine because of its connection to the manufacture of illegal methamphetamine, but not to searches on the sale of Advil or other drugs not connected to a suspicion of illegal use. In this way, some de-identified records would be allowed (if connected to a particularized criminal investigation), but large scale mass surveillance of third party records would be prohibited.

little to offer for its support, except that it does protect identities in de-identified data.

Finally, one could simply hope that the technology of de-identification solves the problems of re-identification. Technology may be created that allows truly de-identified data to stay anonymous. This would of course allow the Standards to be used as originally designed. Whether a technological fix can be envisioned, however, is beyond the scope of this article.

IV. Conclusion

Any prediction of how the LEATPR Standards will withstand the distortions of big data must begin with the fundamental question of the purpose of the Standards themselves. My reading of the Standards is that they seek to *regulate*[178] law enforcement access to personal information, but not necessarily affirmatively protect that information. This distinction between regulation of police and protection of individuals has significant implications. As discussed in the section entitled "Need for the Standards," the drafters begin with the recognition that "[g]overnment access to third party records . . . is surely among the most important and common investigatory activities."[179] This is not to say that the Standards do not acknowledge the important interests of "privacy, freedom of expression, and social participation,"[180] but only that the starting point seems to favor law enforcement access rather than personal privacy protections.[181]

As demonstrated, big data policing compounds an already limited protection of personal information in the LEAPTR Standards. The terminology chosen, the default rules, and the technological fixes may

178. LEATPR STANDARDS, *supra* note 1, at 4-5 ("American norms of limited government and principles of freedom of speech and association thus require that law enforcement records access be regulated.").

179. *Id.* at 2.

180. *Id.* at 5.

181. According to Professor Andy Taslitz, who was a member of the drafting committee:

> Law enforcement members were vehemently opposed to any justification requirement whatsoever, predicting that criminal investigations in serious cases would be rendered virtually impossible. The judge, defense lawyers, and law professors on the drafting committee, however, saw some level of justification as essential to prevent governmental overreaching—to regulate, without prohibiting, legitimate law enforcement work.

Andrew E. Taslitz, *Cybersurveillance Without Restraint? The Meaning and Social Value of the Probable Cause and Reasonable Suspicion Standards in Governmental Access to Third-Party Electronic Records*, 103 J. CRIM. L. & CRIMINOLOGY 839, 841-42 (2013).

not—upon analysis—be very protective in a big data era. Thus, a deferential framework in a big data world may ultimately result in very little protection for personal data.

In many ways, the emerging big data world may necessitate a stronger emphasis on protecting private information. To create a balanced approach, consistent with the goal of law enforcement access but cognizant of privacy concerns of highly private information, the LEAPTR Standards may need to evolve with the technology. The solutions proposed are neither complete nor comprehensive, but provide a starting point for discussion about this complex issue. They offer some guidance about ways to improve the LEAPTR Standards in the face of big data's distorting effects.