30

Algorithmic Stages in Privacy of Data Analytics

Process and Probabilities

Ronald P. Loui, Arno R. Lodder, and Stephanie A. Quick

INTRODUCTION

Technological advances continue to produce massive amounts of information from a variety of sources about our everyday lives. The simple use of a smartphone, for example, can generate data on individuals through telephone records (including location data), social media activity, Internet browsing, e-commerce transactions, and email communications. Much attention has been given to expectations of privacy in light of this data collection, especially consumer privacy. Much attention has also been given to how and when government agencies collect and use this data to monitor the activities of individuals.

In previous work,[1] we discussed three aspects of data analytics in the context of intelligence agencies that matter, but are often overlooked: stages, numbers, and the human factor. We made the following observations. First, stages of the algorithmic process should be included in what legal constraints apply to these activities. Second, the law should take into account the difference between people and machines. Third, the law should think more about numbers. Although each of these topics deserves further discussion, we focus in this chapter in particular on the first issue: stages.

The background of our discussion of algorithms is the activities of intelligence agencies, in particular the National Security Agency (NSA), so the General Data Protection Regulation (GDPR) does not apply. However, the general framework for data protection the GDPR has become, even surpassing its application to non-EU countries, an interesting point of reference for issues of data protection. It does lay down some rules for using algorithms, in particular in the context of automated decision-making and profiling. Notably, the GDPR does not take into account the stages of the algorithmic processing. What we discuss in this chapter is equally relevant for the GDPR.

When there are stages of algorithmic processing, privacy interests depend on the risk of harm at each stage, especially in relation to the numbers of persons under consideration at that stage. It is especially important not to confuse the numbers at one stage (for example, an early stage of algorithmic filtering) with the risks at another stage (for example, a later stage of human involvement). We did discuss this previously, but further justify the relevance of this approach.

Also, analysis by a human, with normal human comprehension, is different from analysis by a program, algorithm, or artificial intelligence (AI), even with human supervision of the

[1] A. R. Lodder and R. Loui, Data Algorithms and Privacy in Surveillance: On Stages, Numbers and the Human Factor, in W. Barfield and U. Pagallo (eds.), *Research Handbook of Law and Artificial Intelligence* (Edward Elgar, 2018).

meta-analysis. Privacy concerns are different depending on the sentience of the comprehension; neither is less concerning in toto, as each can raise different issues. This is especially relevant to actual privacy harms, although large-scale bulk collection can have society-wide effects even if late-stage individual harm is minimized.

The legality of a programmatically warranted surveillance regime, such as the NSA's domestic telephony metadata bulk-collection and seeded-search based on social network analysis under Section 215 of the Patriot Act, depends crucially on scale, which may require reference to numbers or ranges of numbers, not just concepts that differentiate kinds of search, suspicion, and exfiltration, and stages in particular algorithmic processing.[2] Sometimes numbers actually matter, so that different scales of quantities can be different qualitatively in nature. This is one way of balancing the competing interests of national security and US Fourth Amendment Constitutional protections. A warrant-granting stage can have the effect of imposing practical limits on numbers, as can retrospective audits. Justice Sotomayor's concurring opinion on *United States* v. *Jones* is important support for this point of view, although that opinion concerned a variety of concepts (for example, precision and persistence of GPS information, not just quantity, and mainly in the context of Third Party Disclosure).[3]

The chapter is divided into three sections. The first section covers algorithmic stages, in particular the privacy harm of someone being part of a dataset is connected to the particular state of algorithmic processing. The second section elaborates on logging, and the third on legality.

ALGORITHMIC STAGES

The press reports during and after the Snowden revelations about NSA activities reported numbers that caught the attention of the public. See, for instance, the following overview:[4]

> NSA Collecting Phone Records of Millions of Verizon Customers Daily[5]
>
> NSA Collects Millions of E-mail Address Books Globally[6]
>
> In one month, March 2013, the NSA collected 97 billion pieces of intelligence from computer networks worldwide, including 3 billion pieces of intelligence from US computer networks.[7]
>
> NSA Can Legally Access Metadata of 25,000 Callers Based on a Single Suspect's Phone[8]

[2] Consider the practical limit at the warrant-granting stage: "[T]he Administration's decision to bring the Terrorist Surveillance Program under FISA resulted in a paper jam at the Foreign Intelligence Surveillance Court as the number of warrant applications soared." R. A. Posner, Privacy, Surveillance, and Law (2008) 75 *Univ. Chi. Law Rev*. 245, 259.

[3] *United States* v. *Jones*, 565 US 400 (2012).

[4] See, e.g., S. Landau, Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations, *IEEE Security & Privacy* (August 2, 2013); D. Lyon, Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique, Big Data & Society (2014); M. V. Hayden, Beyond Snowden: An NSA Reality Check, *World Affairs Journal* (January/February 2014); M. Hu, Taxonomy of the Snowden Disclosures (2015) 72 *Wash. Lee Law Rev*. 1679; S. I. Vladeck, Big Data before and after Snowden (2014) 7 *J. Nat'l Sec. Law Policy* 333; D. Cole, After Snowden: Regulating Technology-Aided Surveillance in the Digital Age (2016) 44 *Cap. Univ. Law Rev*. 677.

[5] G. Greenwald, NSA Collecting Phone Records of Millions of Verizon Customers Daily, *The Guardian* (June 6, 2013).

[6] B. Gellman and A. Soltani, NSA Collects Millions of E-mail Address Books Globally, *Washington Post* (October 14, 2013).

[7] K. Rodriguez and D. Kayyali, On 6/5, 65 Things We Know about NSA Surveillance that We Didn't Know a Year Ago, *Electronic Frontier Foundation* (June 5, 2014).

[8] A. Nordrum, NSA Can Legally Access Metadata of 25,000 Callers Based on a Single Suspect's Phone, *IEEE Spectrum* (May 16, 2016).

> Reined-In N.S.A. Still Collected 151 Million Phone Records in ’16[9]
>
> NSA Collected 151 Million Phone Records in 2016, Despite Surveillance Law Changes[10]
>
> The U.S. National Security Agency collected 534 million records of phone calls and text messages of Americans last year, more than triple gathered in 2016, a U.S. intelligence agency report released on Friday said.[11]

Meanwhile, the actual audits of numbers of queries and US persons at (*ex post*) risk were on a completely different scale:

> In 2005, for example, although more than 2,000 applications were filed . . .[12]
>
> In 2012, the NSA queried 288 primary phone numbers, and through contact chain analysis touched 6,000 numbers.[13]
>
> Estimated number of targets of [Section 702 orders] (recall that only non-USPs are targeted): CY2013=89,138, CY204=92,707, CY2015=94,368, CY2016=106,469, CY2017=129,080.[14]

How can the discrepancy of numerical scales be reconciled? On the one hand, there are reports of millions of records, and millions of persons having records collected. On the other hand, the number of persons being used as seeds is several orders of magnitude lower (log10); even a full order of magnitude lower for those persons targeted as a result of processing those seeds.

Part of the explanation is that there are, or were, many different programs being confused by press reports and privacy advocates: most notably those regarding domestic persons under the Section 215 Patriot Act II authorization of 2001,[15] and those regarding foreign persons under Section 702 of the FISA Amendment Acts of 2008.[16]

Section 215 allows the government to request a court order from the FISA Court that requires third parties to produce tangible things (such as books, records, and documents) if deemed relevant to a terrorism investigation.[17] The Privacy and Civil Liberties Oversight Board (PCLOB) described the process of storing and accessing telephone records collected under the NSA’s program in its 2014 report:

> Once the calling records are properly formatted, NSA houses them within its data repositories. At this point, technical personnel may take additional measures to make the calling records usable for intelligence analysis, including removing “high volume” telephone identifiers and other unwanted data. The NSA is required to limit who has access to the calling records it obtains . . . Calling records must be deleted from the NSA’s repositories no later than five years after the agency receives them. If a calling record shows up in a “query”

[9] C. Savage, Reined-in N.S.A. Still Collected 151 Million Phone Records in ’16, *New York Times* (May 2, 2017).

[10] J. Vincent, NSA Collected 151 Million Phone Records in 2016, Despite Surveillance Law Changes, *The Verge* (May 3, 2017).

[11] D. Volz, Spy Agency NSA Triples Collection of U.S. Phone Records: Official Report, *Reuters* (May 4, 2018).

[12] R. A. Posner, Privacy, Surveillance, and Law (2008) 75 *Univ. Chi. Law Rev.* 245, 260.

[13] Center for Strategic & International Studies, Fact Sheet: Section 215 of the USA PATRIOT Act (February 27, 2014).

[14] Office of the Director of National Intelligence, STATISTICAL TRANSPARENCY REPORT Regarding Use of National Security Authorities ~ Calendar Year 2017 (April 2018).

[15] United and Strengthening America-Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT) Act of 2001, Pub. L. No. 107–56, § 215, 115 Stat. 272, 287–8 (codified at 50 USC §§ 1861–2 (Supp. II 2002)).

[16] FISA Amendments Act of 2008, Pub. L. No. 110–261, § 702, 122 Stat. 2436, 2438–48 (codified at 50 USC § 1881a (2008)).

[17] See 50 USC § 1861.

> performed by an analyst, however – a process described below – the information about that call need not be destroyed after five years.
>
> Initially, NSA analysts are permitted to access the Section 215 calling records only through "queries" of the database. A query is a software-enabled search for a specific number or other selection term within the database. When an analyst performs a query of a telephone number, for instance, the software interfaces with the database and provides results to the analyst that include a record of calls in which that number participated. Analysts perform these queries to facilitate what is called "contact chaining" – the process of identifying the connections among individuals through their calls with each other. The goals of contact chaining are to identify unknown terrorist operatives through their contacts with known suspects, discover links between known suspects, and monitor the pattern of communications among suspects. Presently, the only purpose for which NSA analysts are permitted to search the Section 215 calling records housed in the agency's database is to conduct queries as described above, which are designed to build contact chains leading outward from a target to other telephone numbers. The NSA has stated that it does not conduct pattern-based searches. Instead, every search begins with a specific telephone number or other specific selection term.[18]

Section 702, on the other hand, provides that the government may collect communications from non-Americans reasonably believed to be located outside the United States in order to obtain foreign intelligence information.[19] The PCLOB describes the process of Section 702 surveillance as follows:

> Once a Section 702 certification has been approved, non-U.S. persons reasonably believed to be located outside the United States may be targeted to acquire foreign intelligence information within the scope of that certification . . .
>
> The Section 702 certifications permit non-U.S. persons to be targeted only through the "tasking" of what are called "selectors." A selector must be a specific communications facility that is assessed to be used by the target, such as the target's email address or telephone number. Thus, in the terminology of Section 702, people (non-U.S. persons reasonably believed to be located outside the United States) are targeted; selectors (e.g., email addresses, telephone numbers) are tasked. The users of any tasked selector are considered targets – and therefore only selectors used by non-U.S. persons reasonably believed to be located abroad may be tasked. The targeting procedures govern both the targeting and tasking process . . .
>
> Although targeting decisions must be individualized, this does not mean that a substantial number of persons are not targeted under the Section 702 program. The government estimates that 89,138 persons were targeted under Section 702 during 2013. Once a selector has been tasked under the targeting procedures, it is sent to an electronic communications service provider to begin acquisition.[20]

FISA, while considerably scrutinized and revised since its inception, does not directly engage Fourth Amendment protections, so Section 702 activity is very different from bulk collection under Section 215.[21]

[18] Privacy and Civil Liberties Oversight Board, Report on the Telephone Records Program Conducted under Section 215 of the USA Patriot Act and on the Operations of the Foreign Intelligence Surveillance Court, January 23, 2014, pp. 25, 26–7, www.pclob.gov/library/215-Report_on_the_Telephone_Records_Program.pdf (hereinafter, PCLOB Section 215 Report).

[19] See 50 USC § 1881a.

[20] Privacy and Civil Liberties Oversight Board, Report on the Surveillance Program Operated Pursuant to Section 702 of the Foreign Intelligence Surveillance Act, Jul 2, 2014, pp. 32–3, www.pclob.gov/library/702-Report.pdf (hereinafter, PCLOB Section 702 Report).

[21] In fact, the PCLOB recommended curtailing Section 215 activity while retaining Section 702 activity with few modifications. See PCLOB Section 215 Report, above note 18, and PCLOB Section 702 Report, *ibid.*

Even so, the numbers reported for Section 702 may be surprisingly low in the context of the press reports. It is worth noting that the foreign press may have had little incentive to keep straight the difference between collections for foreign persons and for US persons.

The main reason the numbers are so different is that there are different stages in the processing. From a seed query, with individualized warrant, Section 215 used social network analysis to determine who was 1-hop from the seed (directly in contact), then 2-hops (directly in contact with someone directly in contact), then 3-hops (directly in contact with someone 2-hops from the seed). Anyone within 3-hops would have records added to the "corporate store" that could be searched under other authorities.[22]

From the point of view of numbers of records involved, one starts with a single number, then assuming 100 to 1,000 contacts per hop, scales to 100 after one hop, 10,000 after two hops, and 1 million after three hops. This is the lower estimate, with a fan-out of 100. For an upper estimate, 1,000 after one hop is 1 million after two hops, and 1 billion after three hops. This is from a single seed. No doubt the reduction from three hops to two hops was a significant reduction in scale.[23]

Looking at the process from the point of view of collection and filtering, perhaps ten records of 100 million persons are collected in bulk each day, or 1 billion records. If records are retained for 100 days for social network analysis, that makes 100 million persons' 100 billion records at risk for connection to a seed on any given query. Using the Section 702 Office of the Director of National Intelligence transparency report numbers from 2017, 129,000 persons had records targeted. This is a different program, targeting non-US persons, and post-curtailment by the Obama administration. But it is a useful estimate for illustration. Suppose the result of winnowing in Section 215 social network analysis yielded 100,000 persons' ten records for 100 days. This means that at the first stage of processing, 2-hop connection, 100 million was reduced to 100,000, or a 1,000:1 reduction. In this first stage, 99.9 percent of the records, 100 million – 100,000 or 99.9 million persons' records were discarded. While 100,000 persons 1,000 records were retained, an alarming 100 million results of a single query, an equally spectacular 99.9 million persons' 1,000 records were algorithmically declared uninteresting, or 99.9 billion records. This is a massive discarding or excluding, at least for the purposes of that query.[24]

This is not to take either side of the dispute over whether too much was collected, or whether a sufficient amount of the collection was never "used," hence large numbers of persons' privacy rights barely implicated. It is intended to make clear the point that the numbers need to be understood in terms of the stage of the processing.[25]

[22] This process is described in authoritative detail in the PCLOB Section 215 Report, above note 18.

[23] Others have noted this over-productivity. See, e.g.: "This might be okay – except for the fact that each one of those 300 or so queries can in turn let the NSA suck in 1 million other phone records" (J. Roberts, How Feds Use One "Seed" and 3 "Hops" to Spy on Nearly Everyone, *Gigaom* (December 17, 2013). See also, Nordrum, above note 8.

[24] For similar analysis, see, e.g., D. Storm, NSA Collected 1 Trillion Metadata Records, Harvested 1 Billion Mobile Calls Daily, *Computerworld* (June 30, 2013). Also, "approximately 500 billion communication records [were] intercepted and analyzed in 2012, via a program called One-End Foreign, which relies on the FISA Amendments Act for its legality." D. Ombres, NSA Domestic Surveillance from the Patriot Act to the Freedom Act: The Underlying History, Constitutional Basis, and the Efforts at Reform (2015) 39 *Seton Hall Legis. J.* 27, 32. Both of these, however, are referring to Section 702 activity.

[25] It may be worth noting that the rationale for 3-hop connectivity in multi-modal social network analysis is that productivity at each hop may be more limited. For a shared bank account, or a shared plane flight, 10× fan out per day, or 1,000× per hop per query is a massive overestimate (e.g. sharing an apartment produces at most ten new connections over the entire 100-day period). Not all hops are necessarily telephonic connections.

What are the risks at each stage? For an individual, the risk of being included at the first bulk collection stage is that the individual might be included in the next stage, being warehoused in the "corporate store" and susceptible to production under analysts' more general search. Note that the processing between the first and second stages is entirely algorithmic. The probability of being included at the first stage might be 50 percent (0.50), depending on whether we are starting with a population of US adults, US adult phone users, US adult phone users with business records at a major carrier, etc. The probability of being included at the second stage might be 1/1,000 of that, or 0.0005, depending on what one imagines the novelty of contacts per day might be during the period of temporary social network analytics access. One hundred million persons initially at risk became 100,000 persons with records in the store. The third stage of processing starts when a record included in the second stage is actually produced during a human-initiated query and examined. This probability might be another 1/10 or 1/100 winnowing, resulting in the 1,000 to 10,000 persons targeted for human inspection that would be in line with the later statistical transparency reports.

The fact is that US intelligence agencies do not have the manpower to inspect much larger numbers of records and the point of the algorithmic "front-end" is to produce a smaller number of highly probable, if not reasonably articulable, suspects.

A similar point was made in the PCLOB Section 215 Report:[26]

> [C]hanging program rules to limit contact chaining to two hops … would not unduly diminish the value of the telephony metadata program … Each additional hop from the original "selector" makes the connection more remote and adds exponentially greater numbers of "false positives" to the query results. The value of connections becomes more limited as the contact chain is extended and it becomes more difficult to sift through the results.

FROM BENIGN LOGGING TO AUTHORITARIAN AI

The question of whether algorithmic looking is different from human looking, with regard to privacy protections, is where much of the debate remains open. And if it is different, in what ways can it be less invasive, while in what other ways it can be more invasive.

On the one hand, we see converging respect for what analytics and the fusion of disparate data, even metadata, can yield. Whether this is knowing, or merely probabilistic hypothesizing, is worth further discussion. It may be that claiming and attributing with low probability raise significantly reduced privacy concerns. The same would be true for probably non-persistent conclusions, based on data that is out of date. Privacy interests might even be lessened when justified true beliefs are wrongly known (improper causal chains that are "lucky guesses"). Many of these epistemological concerns are taken seriously in the GDPR recommendation that data be anonymized or pseudonymized.[27]

[26] PCLOB Section 215 Report, above note 18, pp. 170–1.

[27] For example, blurring: "Data blurring uses an approximation of data values to render their meaning obsolete and/or render the identification of individuals impossible" and generalizing, which reduce precision, in addition to unlinking forms of anonymization that make precise attribution improbable. GDPR Report, Data Masking: Anonymisation or Pseudonymisation? (November 7, 2017). K. Lehrer, The Gettier Problem and the Analysis of Knowledge, in G. Pappas (ed.), *Justification and Knowledge* (Springer, 1979). Lehrer's work is an example of the philosophical problem of wrongly knowing something that is nevertheless justified true belief.

There is no question that in an era of big data, concepts and dimensions of privacy are shifting with new appreciation for what can be known with algorithmic inference and AI. On the other side is Richard Posner's claim that the privacy concerns began with the human comprehension:[28] "So the search sequence is interception, data mining, and finally a human search of those intercepted messages that data mining or other information sources have flagged as suspicious. Computer searches do not invade privacy because search programs are not sentient beings. Only the human search should raise constitutional or other legal issues." Posner continues by distinguishing the harm of algorithmic scrutiny and the harm of *ex post* human interest.

In the initial computer sifting designed to pick out data meriting scrutiny by an intelligence officer, only facts bearing on national security will trigger scrutiny. But once an individual is identified as a possible terrorist or foreign agent, the government's interest in him will explode. Besides obtaining contact information, it will want to learn about his ethnicity and national origin; education and skills; previous addresses and travel (especially overseas); family, friends, and acquaintances; political and religious beliefs and activities; finances; any arrest or other criminal record; military service (if any); mental health and other psychological attributes; and a range of consumption activities, the whole adding up to a comprehensive personal profile.

No doubt collection for the purposes of potential downstream human comprehension do raise individual privacy concerns, precisely because sentient beings have a probability of using the data and deriving conclusions.[29] Posner is right, however, that there are many benign examples of computers looking at data that do not usually raise privacy concerns. For example, the Internet is based on packet-forwarding nodes that are presumed to look only at the headers. A fair analogy is the postcard being processed at the Post Office. Even if those packet-forwarding nodes processed the packet "payload" contents, for example, to calculate a "checksum" and verify that the data has not been corrupted, no one seems to care. Perhaps the only action that would result from such a calculation is a request for retransmit at the prior node, completely devoid of human intervention. But if the error is logged, and the error logs are routinely read by a human system administrator, the risk of human comprehension as a result of algorithmic filtering and storing does raise privacy issues. Internet service providers log requests routinely, but only the prospect of those logs being attributed and sold brought attention to the logging.[30]

Another benign example of algorithmic processing is the calculation of baseline normality for abnormality testing. When machine learning is applied to a set of cases that are normal, and a set of cases labeled as abnormal, usually no re-identifiable data can be attributed to an individual in one of the training sets. This is especially true of the individuals in the set labeled as normal: usually it is a large set, and usually no decision is made, no alert sounded, when inputs are classified as normal. In the simplest example of this, training data may simply cause statistical summaries such as means and moments to be retained. Long before there was machine learning, there was information retrieval that required "inverse document frequencies" of words, which were statistical summaries that aggregated many individuals' data.

[28] R. A. Posner, Privacy, Surveillance, and Law (2008) 75 *Univ. Chi. Law Rev.* 245, 253–4.

[29] See, e.g., O. Tene and J. Polonetsky, Privacy in the Age of Big Data: A Time for Big Decisions (2012) 64 *Stan. Law Rev. Online*; I. S. Rubinstein, Big Data: The End of Privacy or a New Beginning? (2013) 3 *Int. Data Privacy Law* 74; or I. Kerr and J. Earle, Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy (2013) 66 *Stan. Law Rev. Online* 65.

[30] See, e.g., B. Fung, What to Expect Now that Internet Providers Can Collect and Sell Your Web Browser history, *Washington Post* (March 29, 2017).

There are few privacy interests for data used in this way, precisely because no human ever looks at the individual's information apart from the aggregate.[31]

Computers, of course, are constantly logging transactions for billing purposes, where summarization, notification, and payment are all done without human comprehension, unless an error condition or audit is triggered.

There are autonomous systems that are not benign processors. Some may even be regarded as malign actors, or even authoritarian AI. The same machine-learning program that did not raise privacy concerns during training could make autonomous decisions that impact persons in ways that anyone would recognize as classic violations of an individual's rights. In most cases, those rights will not be privacy rights, but non-discrimination, speech, employment, equal protection, property, or other rights. It will be because someone or some group delegated too much authority, or because of an error in the program, or a malfunction of the device: something where the control is the problem, not just the access to information and its comprehension.

It is easily conceivable that an autonomous system could access student test records, or statistical correlates of student test records, for an employment decision, in a way that ignores Family Educational Rights and Privacy Act protections. Even here, it is not the machine comprehension of the data that stings, but the use of the data, as an act, that produces a harm or violates a protocol; and that sequence is probably attributable to a programmer. When self-organizing, evolving, self-programmed automata begin to re-identify Health Insurance Portability and Accountability Act anonymized data in patently disallowed ways, without having been programmed explicitly to do so, then Posner's counsel, "[o]nly the human search should raise constitutional or other legal issues," might well be disregarded.

The bottom line is that sometimes computer processing is benign. Collection qua logging is just collection, unless there is potential for further processing of the logs, leading eventually to comprehension by a human, or leading to a decision, action, or inaction. Many algorithms are intended to improve privacy of data, by de-anonymizing or de-linking, reducing precision, masking, blurring, generalizing, hashing, encrypting, or otherwise pseudonymizing.[32]

No one minds that a computer "sees" the data that it is encrypting so that others cannot see it. Other kinds of computer processing can transgress, especially if the programming permits impropriety. Most of the time the automation is discarding data as uninteresting, as a pre-processor to reduce the human workload at a later stage. There is therefore no *a priori* reason to assume that computer processing of data, even large quantities of data, crosses privacy sensibilities. And there is no reason to believe that computer processing of data, of any quantity, is incapable of raising concerns. It depends on the algorithm and how it is embedded in social decision-making and social controls.

LEGALITY AND ALGORITHMIC STAGES

Justice Sotomayor's concurring opinion in *United States* v. *Jones*, regarding police use of GPS tracking, is the clarion call for regarding big data as different from property and persons that can be searched and seized. In *Jones*, the Supreme Court held that attaching a GPS device to a vehicle and using the device to monitor the vehicle's movements constitutes a search under

[31] A classic work in this field is G. Salton, *Automated Text Processing* (Prentice Hall, 1983).

[32] For discussion of various privacy techniques, see, e.g., S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets (2019) 22 *Stan. Technol. Law Rev.* 1; or K. Nissim, A. Bembenek, A. Wood, *et al.*, Bridging the Gap between Computer Science and Legal Approaches to Privacy (2018) 31 *Harv. J. Law Technol.* 2.

the Fourth Amendment. The Court reached a unanimous result, but several justices differed in their reasoning. In particular, Justice Sotomayor explained that technological advances have shaped the evolution of privacy expectations, particularly where physical intrusion is unnecessary for many forms of surveillance. She stated that, at the very least, "longer term GPS monitoring in investigations of most offenses impinges on expectations of privacy." Furthermore:

> In cases involving even short-term monitoring, some unique attributes of GPS surveillance relevant to the *Katz* analysis will require particular attention. GPS monitoring generates a precise, comprehensive record of a person's public movements that reflects a wealth of detail about her familial, political, professional, religious, and sexual associations . . . The Government can store such records and efficiently mine them for information years into the future . . . And because GPS monitoring is cheap in comparison to conventional surveillance techniques and, by design, proceeds surreptitiously, it evades the ordinary checks that constrain abusive law enforcement practices: limited police resources and community hostility.

There are several issues regarding data here. First, there is the collection over time. Second, the precision or alleged precision. Third, the comprehensiveness, or alleged comprehensiveness. Fourth, there is the inferential productivity of the data. Fifth, there is the persistence of the data, which raises questions of its staleness and the European concern over "being forgotten." Finally, there is the issue of quantity, where a practical limit that had been related to cost is no longer a constraint, with cheap data acquisition at scale.

These issues are beyond the scope of this chapter except the last, and to some extent the fourth. We have argued that the algorithm matters, and the numbers at each stage matter, in expressing judicial will and guidance. When there is an issue of quantity, perhaps legislators and courts should find a way to express what quantities are permissible. The quantities, in relation to the stages of the algorithm, may have as rationale the productivity of the inferences at each stage (the fourth issue above), not just the risk of practical harm. Quantities can be expressed as proportions, with suggested norms if not bright lines, and with orders of magnitudes for ranges. Algorithms can be referenced by their type, such as "social network analysis" or "automatic classification," and their high-level features, such as *ex ante* and *ex post*, relative to some massive filtering, querying, or storing.

In some ways, the issue of reasonable quantities was delegated to the FISA Court, which failed to constrain sequential retention as a concomitant of individual warrant granting. Gray and Citron propose a different idea of quantity in relation to privacy.[33] Until now, most proposals for defending Fourth Amendment interests in quantitative privacy have focused on a case-by-case method called the "mosaic theory." Under this approach, the Fourth Amendment is implicated whenever law enforcement officers gather "too much" information during the course of a specific investigation. Critics of the mosaic theory have rightly wondered how courts will determine whether investigators have gathered too much information in any given case and how officers in the midst of ongoing investigations will know whether the aggregate fruits of their efforts are approaching a Fourth Amendment boundary. The best solution that mosaic advocates have so far been able to muster is to draw bright, if arbitrary, lines based on how long officers use an investigative method or technology. These kinds of solutions fail to satisfy because they are under-inclusive and over-inclusive, and also sidestep important conceptual and doctrinal questions. Rather than asking how much

[33] See D. Gray and D. Citron, The Right to Quantitative Privacy (2013) 98 *Minn. Law Rev*. 62.

information is gathered in a particular case, we argue here that Fourth Amendment interests in quantitative privacy demand that we focus on how information is gathered.[34]

While we agree that a mosaic theory of the Fourth Amendment is an important idea, and agree that there is such a thing as gathering too much information under a particular warrant, their conclusion is different from ours (and, apparently, Posner's). They focus on the permissibility or impermissibility of first-stage bulk collection, which has the potential, the mere possibility, of yielding human scrutiny of a particular individual's data at some later stage. This was in fact where much of the press focused its attention during the Snowden disclosures, and led to the disconnect: one side was discussing large numbers at the first stage of algorithmic processing; the other side was discussing small numbers at the final stages of algorithmic processing.

It may be that constraining collection is a lost cause, practically speaking. Like disallowed discrimination in the insurance industry, the data are there whether one tries to disallow their collection or not. The rights of individuals must therefore be secured by severing the particular inferences that cause actual harm.

CONCLUSION

Technological innovations have dramatically expanded the ability to use personal data: the ability to collect, edit, compare, distribute, store, process, and mine data, which were once extraordinarily time-consuming processes, can now be accomplished instantaneously. Unsurprisingly, national intelligence agencies around the world are taking advantage of new information communication technologies in their investigations, particularly those against terrorists. Consequently, there has been an increase in state control over individuals' movements, financial standing, behavior, sexual habits, and other personal characteristics important to self-definition. States justify the high levels of data collection and exchange as part of the fight against crime and terrorism. The central question concerns how to warrant this unbridled gathering, analysis, and exchange of electronic information by intelligence agencies and law enforcement to combat terroristic threats and investigate criminal activities while safeguarding the privacy of ordinary citizens.

We claim that the concept of stages should be an integral part of any discussion about the privacy infringements by intelligence agencies, and should be taken into account in future legislation. We elaborated on the characteristics of stages in algorithms and what the particular consequences of stages might be. This will help to clarify how to guard the values concerning algorithmic processing of data.

[34] Another way of explaining mosaic theory is given here: "The fundamental insight behind the mosaic theory is that we can maintain reasonable expectations of Fourth Amendment privacy in certain quantities of information and data even if we lack reasonable expectations of privacy in the constituent parts of that whole." D. Gray and D. K. Citron, A Shattered Looking Glass: The Pitfalls and Potential of the Mosaic Theory of Fourth Amendment Privacy (2012) 14 *NC J. Law Technol.* 381, 390.