University of the District of Columbia School of Law
Digital Commons @ UDC Law

Journal Articles | Publications

2014

# When Enough is Enough: Location Tracking, Machine Learning and the Mosaic Theory

Renee McDonald Hutchins

Steve Bellovin

Tony Jebara

Sebastian Zimmeck

Follow this and additional works at: https://digitalcommons.law.udc.edu/fac_journal_articles

Part of the Privacy Law Commons, and the Supreme Court of the United States Commons

# When Enough is Enough: Location Tracking, Mosaic Theory, and Machine Learning

**Steven M. Bellovin**
**Columbia University**

**Renée M. Hutchins**
**University of Maryland Francis King Carey School of Law**

**Tony Jebara**
**Columbia University**

**Sebastian Zimmeck**
**Columbia University**

**No. 2013 - 51**

UNIVERSITY *of* MARYLAND
FRANCIS KING CAREY
SCHOOL OF LAW

This paper can be downloaded free of charge at:
The Social Science Research Network Electronic Paper Collection
http://ssrn.com/abstract=2320019

NEW YORK UNIVERSITY
JOURNAL OF LAW & LIBERTY

# WHEN ENOUGH IS ENOUGH: LOCATION TRACKING, MOSAIC THEORY, AND MACHINE LEARNING

**Steven M. Bellovin** [I]
**Renée M. Hutchins** [II]
**Tony Jebara** [III]
**Sebastian Zimmeck** [IV]

**ABSTRACT**: Since 1967, when it decided *Katz v. United States*, the Supreme Court has tied the right to be free of unwanted government scrutiny to the concept of reasonable expectations of privacy.[1] An evaluation of reasonable expectations depends, among other factors, upon an assessment of the intrusiveness of government action. When making such assessment historically the Court considered police conduct with clear temporal, geographic, or substantive limits. However, in an era where new technologies permit the stor-

[I] Professor, Columbia University, Department of Computer Science.

[II] Associate Professor, University of Maryland Francis King Carey School of Law.

[III] Associate Professor, Columbia University, Department of Computer Science.

[IV] Ph.D. candidate, Columbia University, Department of Computer Science.

[1] Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring).

age and compilation of vast amounts of personal data, things are becoming more complicated. A school of thought known as "mosaic theory" has stepped into the void, ringing the alarm that our old tools for assessing the intrusiveness of government conduct potentially undervalue privacy rights.

Mosaic theorists advocate a cumulative approach to the evaluation of data collection. Under the theory, searches are "analyzed as a collective sequence of steps rather than as individual steps."[2] The approach is based on the observation that comprehensive aggregation of even seemingly innocuous data reveals greater insight than consideration of each piece of information in isolation. Over time, discrete units of surveillance data can be processed to create a mosaic of habits, relationships, and much more. Consequently, a Fourth Amendment analysis that focuses only on the government's collection of discrete units of data fails to appreciate the true harm of long-term surveillance—the composite.

In the context of location tracking, the Court has previously suggested that the Fourth Amendment may (at some theoretical threshold) be concerned with the accumulated information revealed by surveillance.[3] Similarly, in the Court's recent decision in *United States v. Jones*, a majority of concurring justices indicated willingness to explore such an approach.[4] However, in general, the Court has rejected any notion that technological enhancement matters to the

[2] Orin Kerr, *The Mosaic Theory of the Fourth Amendment*, 111 MICH. L. REV. 311, 312 (2012).

[3] United States v. Knotts, 460 U.S. 276, 284 (1983).

[4] Justice Scalia writing for the majority left the question open. United States v. Jones, 132 S. Ct. 945, 954 (2012) ("It may be that achieving the same result [as in traditional surveillance] through electronic means, without an accompanying trespass, is an unconstitutional invasion of privacy, but the present case does not require us to answer that question.").

constitutional treatment of location tracking.[5] Rather, it has decided that such surveillance in public spaces, which does not require physical trespass, is equivalent to a human tail and thus not regulated by the Fourth Amendment. In this way, the Court has avoided a quantitative analysis of the amendment's protections.

The Court's reticence is built on the enticingly direct assertion that objectivity under the mosaic theory is impossible. This is true in large part because there has been no rationale yet offered to objectively distinguish relatively short-term monitoring from its counterpart of greater duration.[6] This article suggests that by combining the lessons of machine learning with the mosaic theory and applying the pairing to the Fourth Amendment we can see the contours of a response. Machine learning makes clear that mosaics can be created. Moreover, there are important lessons to be learned on *when* this is the case.

Machine learning is the branch of computer science that studies systems that can draw inferences from collections of data, generally by means of mathematical algorithms. In a recent competition, "The Nokia Mobile Data Challenge,"[7] researchers evaluated machine learning's applicability to GPS and cell phone tower data. From a user's location history alone, the researchers were able to estimate

[5] *Compare Knotts*, 460 U.S. at 276 (rejecting the contention that an electronic beeper should be treated differently than a human tail) *and* Smith v. Maryland, 442 U.S. 735, 744 (1979) (approving the warrantless use of a pen register in part because the justices were "not inclined to hold that a different constitutional result is required because the telephone company has decided to automate") *with* Kyllo v. United States, 533 U.S. 27, 33 (2001) (recognizing that advances in technology affect the degree of privacy secured by the Fourth Amendment).

[6] United States v. Jones, 132 S. Ct. 945 (2012); *see also* Kerr, *supra* note 2, at 329-330.

[7] *See Mobile Data Challenge 2012 Workshop*, NOKIA RESEARCH CENTER, http://research.nokia.com/page/12340.

the user's gender, marital status, occupation and age.[8] Algorithms developed for the competition were also able to predict a user's likely *future* location by observing past location history. The prediction of a user's future location could be even further improved by using the location data of friends and social contacts.[9]

Machine learning of the sort on display during the Nokia competition seeks to harness the data deluge of today's information society by efficiently organizing data, finding statistical regularities and other patterns in it, and making predictions therefrom. Machine learning algorithms are able to deduce information—including information that has no obvious linkage to the input data—that may otherwise have remained private due to the natural limitations of manual and human-driven investigation. Analysts can train machine learning programs using one dataset to find similar characteristics in new datasets. When applied to the digital "bread crumbs" of data generated by people, machine learning algorithms can make targeted personal predictions. The greater the number of data points evaluated, the greater the accuracy of the algorithm's results.

In five parts, this article advances the conclusion that the duration of investigations is relevant to their substantive Fourth Amendment treatment because duration affects the accuracy of the predictions. Though it was previously difficult to explain, for example, why an investigation of four weeks was substantively different from an investigation of four hours, we now have a better

[8] Sanja Brdar, Dubravko Culibrk & Vladimir Crnojevic, *Demographic Attributes Prediction on the Real-World Mobile Data*, MOBILE DATA CHALLENGE WORKSHOP 2012, https://research.nokia.com/files/public/mdc-final202-brdar.pdf

[9] Manlio de Domenico, Antonio Lima & Mirco Musolesi, *Interdependence and Predictability of Human Mobility and Social Interactions*, MOBILE DATA CHALLENGE WORKSHOP 2012, https://research.nokia.com/files/public/mdc-final306_dedomenico.pdf.

understanding of the value of aggregated data when viewed through a machine learning lens. In some situations, predictions of startling accuracy can be generated with remarkably few data points. Furthermore, in other situations accuracy can increase dramatically above certain thresholds. For example, a 2012 study found the ability to deduce ethnicity moved sideways through five weeks of phone data monitoring, jumped sharply to a new plateau at that point, and then increased sharply again after twenty-eight weeks.[10] Similarly, the accuracy of identification of a target's significant other improved dramatically after five days' worth of data inputs.[11] Experiments like these support the notion of a threshold, a point at which it makes sense to draw a Fourth Amendment line.

In order to provide an objective basis for distinguishing between law enforcement activities of differing duration, the results of machine learning algorithms can be combined with notions of privacy metrics, such as *k*-anonymity or *l*-diversity. While reasonable minds may dispute the most suitable minimum accuracy threshold, this article makes the case that the collection of data points allowing predictions that exceed selected thresholds should be generally deemed unreasonable searches in the absence of a warrant.[12] Moreover, any new rules should take into account not only the data being collected but also the foreseeable improvements in the machine learning technology that will ultimately be brought to bear on it; this includes using future algorithms on older data.

[10] *See* Yaniv Altshuler, Nadav Aharony, Michael Fire, Yuval Elovici & Alex Pentland, *Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data*, WS3P, IEEE SOCIAL COMPUTING, Figure 10, (2012), http://arxiv.org/ftp/arxiv/papers/1111/1111.4645.pdf.

[11] *See id*. Figure 9.

[12] Admittedly, there are differing views on sources of authority beyond the Constitution that might justify location tracking. *See, 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 (2012).

In 2001, the Supreme Court asked "what limits there are upon the power of technology to shrink the realm of guaranteed privacy."[13] In this study, we explore an answer and investigate what lessons there are in the power of technology to protect the realm of guaranteed privacy. After all, as technology takes away, it also gives. The objective understanding of data compilation and analysis that is revealed by machine learning provides important Fourth Amendment insights. We should begin to consider these insights more closely.

## TABLE OF CONTENTS

Introduction ....... 561
I. Legal Analysis ....... 564
A. The Development of the "Reasonable Expectation of Privacy" ....... 565
B. Location Tracking as Privacy Violation ....... 567
C. The Emergence of the Mosaic Theory ....... 570
D. The Privacy of the Home as a Basis for location tracking. ....... 576
E. Quantifying the Mosaic ....... 583
II. Machine Learning ....... 589
A. Unsupervised Machine Learning ....... 590
B. Supervised Machine Learning ....... 592
C. Semi-supervised and Network Learning ....... 594
III. Privacy Metrics ....... 596
A. *k*-Anonymity ....... 597
B. *l*-diversity ....... 599
IV. A Machine Learning Approach to the Mosaic Theory ....... 601
A. The Existence of Predictable Location Patterns ....... 603

[13] Kyllo v. United States, 533 U.S. 27, 34 (2001).

B. Determining the Formation of a Mosaic ................................. 606
C. Applying Privacy Metrics......................................................... 615
D. Determining a Privacy Violation ............................................ 620
E. A Notional Experiment ............................................................ 623
V. Conclusions ...................................................................................... 626

## INTRODUCTION

In *Olmstead v. United States*,[14] the first wiretap case considered by the Supreme Court, the Fourth Amendment was interpreted very narrowly. The Court asserted that only physical searches of "material things—the person, the house, his papers or his effects" were relevant under the Fourth Amendment.[15] In *Katz v. United States*,[16] though, the Court reversed this interpretation, saying "the reach of the Amendment cannot turn upon the presence or absence of a physical intrusion into any given enclosure."[17] Since the Court decided *Katz* in 1967 technology has moved further, and the scope of the Fourth Amendment is again being challenged by invention.

One particularly thorny issue of Fourth Amendment analysis is location tracking: is a warrant required to track someone with the aid of a technological device? At first glance, the answer would seem to be "no." Following someone was hardly a new concept in 1789, when the amendment was introduced into the first Congressional session. It is not obvious why technology would change this. The question, then, is this: can newer and perhaps more invasive location tracking technology constitute a difference sufficient to bring location tracking under the ambit of the Fourth Amendment?

[14] Olmstead v. United States, 277 U.S. 438 (1928).

[15] *Id.* at 464.

[16] Katz v. United States, 389 U.S. 347, 353 (1967).

[17] *Id.* at 353.

In the only modern location tracking case to reach the Supreme Court thus far, *United States v. Jones*,[18] many expected that the question would need to be answered. However, as it turned out, technology did not play a role for the holding of the *Jones* Court. Rather, since a tracking device had been attached to Jones's car, the police actions were held to squarely fall within classic Fourth Amendment doctrine: there had been an unauthorized physical intrusion, so a warrant was required independent of the location tracking. As indicated in the concurring opinions, though,[19] five of the justices seemed prepared to move further than the majority opinion did. However, these concurring opinions failed to conclusively identify what test should be used to analyze the relevance of modern location tracking technology under the Fourth Amendment.

One proposed test has been labeled the "mosaic theory." Under the mosaic theory, identifying searches that trigger Fourth Amendment protection requires the analysis of police actions, each of which may not qualify as a search when viewed in isolation but which over time reveal a collective "mosaic" of behavior and characteristics.[20] That is, it is the totality of information gathered that makes a search unreasonable. Such a collection of information is *more* than the sum of its parts; the inferences that can be drawn go far beyond the individual observations.[21] It need not be stressed that data mining and other modern technologies allow even more detailed mosaics to be developed.

[18] United States v. Jones, 132 S. Ct. 945 (2012).

[19] *See id*. at 954 (Sotomayor, J., concurring); *id*. at 957 (Alito, J., concurring).

[20] Kerr, *supra* note 2.

[21] *See, e.g.*, Renée Hutchins, *Tied up in* Knots*? GPS Technology and the Fourth Amendment*, 55 UCLA L. REV. 409, 458 (2007) (stating that the "[police] could generate and compare such records for weeks or months at a time to develop a comprehensive digest of [a person's] friends, associates, preferences, and desires").

In particular, one branch of computer science, machine learning, can cause concern when it comes to privacy and large datasets. Machine learning is just what it sounds like: it is a way for computers to "deduce" patterns in datasets and use those patterns to do further analysis. Specifically, in supervised machine learning, an analyst can "train" a machine learning program using one dataset. The patterns derived can then be used to find the same characteristics in new datasets.[22]

One recent use of machine learning technology is location prediction.[23] Given a training dataset of location data, such as GPS tracking logs, a suitable program can look at a new dataset and make predictions with some degree of accuracy, including where someone is likely to be in the future. In other words, such programs are in a strong sense a technological exemplar of the mosaic theory: based on prior knowledge, they can predict behavioral patterns and characteristics of a subject, accumulating information into a picture of increasing completeness.

Such technological advances have largely been viewed as a source of concern for privacy activists. However, as technology takes away, it also gives. We posit that viewing the Fourth Amendment protection through the lens of machine learning offers important legal guidance. The main idea is this: If there are enough data points that allow for predictions above a certain threshold of accuracy, a mosaic exists. Thus, for the grouping problem—the problem of identifying which data points a set must contain to transform it into a mosaic[24]—we claim that the set must be composed of points that enable predictions above a certain threshold of accuracy. Collection of data in excess of the threshold established by

[22] Machine learning is explained in more detail. *See infra* Section II.
[23] *See, e.g.*, de Domenico et al., *supra* note 9.
[24] Kerr, *supra* note 2, at 333-36.

experiments involving machine learning is *a priori* an unreasonable search.[25]

The remainder of the article is organized as follows: We start with a review of the relevant legal and technological background. Specifically, we give an overview of Fourth Amendment law in the context of location tracking (Section I.); we then provide an introduction to machine learning (Section II.) and discuss privacy metrics, which are mathematical and statistical models aimed at quantifying "privacy" (Section III.). Then we present our major contributions: a demonstration of how sufficient data lets us build a functional mosaic. That is, by using machine learning techniques on a *given amount of data* it is possible to make useful predictions, predictions that go beyond what is actually known, and that are relevant to the Fourth Amendment's analysis of location tracking (Section IV.). Finally, we summarize our contributions (Section V.).

## I. LEGAL ANALYSIS

The Fourth Amendment guarantees the "right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures."[26] In the early years, this language was understood quite literally.[27] Routinely interpreting the amendment narrowly, the Supreme Court stated that it protected

[25] The issue of what sort of authorization should be needed for location tracking can be looked at from a legislative dimension as well. *See* Pell & Soghoian, *supra* note 12.

[26] U.S. CONST. amend. IV.

[27] *See, e.g.*, Goldstein v. United States, 316 U.S. 114, 120 (1942) ("[T]he unlawful interception of a telephone communication does not amount to a search or seizure prohibited by the Fourth Amendment.") (citing Olmstead v. United States, 277 U.S. 438 (1928)); Goldman v. United States, 316 U.S. 129 (1942).

against little more than physical intrusions by law enforcement.[28] By the late 1960s, however, law enforcement was increasingly able to gain access to information about private affairs without actual incursion into protected spaces. The Court (or at least a majority of its members) became more and more concerned about a world of unregulated government surveillance. This concern led the Court to the realization that a more robust interpretation of the Fourth Amendment was needed.

### A. THE DEVELOPMENT OF THE "REASONABLE EXPECTATION OF PRIVACY"

In 1967, in *Katz v. United States*, the Court settled upon an understanding of the amendment that used the concept of a person's "reasonable expectation of privacy" as the boundary line of protection.[29] Justice Harlan explained in his concurring opinion that this boundary imposes "a twofold requirement, first that a person have exhibited an actual (subjective) expectation of privacy and, second, that the expectation be one that society is prepared to recognize as 'reasonable[,]'" that is, it must be objectively reasonable.[30] Rejecting its past fealty to the singular notion of trespass, the Court in *Katz* further explained that "the reach of [the Fourth] Amendment cannot turn upon the presence or absence of a physical intrusion into any given enclosure."[31] Stating plainly its seeming analytical shift,

[28] Olmstead v. United States, 277 U.S. 438 (1928).

[29] *See* Katz v. United States, 389 U.S. 347, 360 (1967) (Harlan, J., concurring).

[30] *Id.* at 361. Shortly after the decision in *Katz* was handed down the full Court adopted, in various majority opinions, the test articulated by Justice Harlan in his *Katz* concurrence. *See, e.g.*, Terry v. Ohio, 392 U.S. 1, 9 (1968) (citing Katz v. United States, 389 U.S. 347, 361 (1967)).

[31] Katz v. United States, 389 U.S. 347, 353 (1967).

the Court announced that "the Fourth Amendment protects people, not places."[32]

Under the Court's evolved understanding of the amendment the heart of the inquiry shifted from explicit consideration of specific police tactics to a broader discussion about what society should reasonably be able to expect the police not to do. Rather than examining, for example, whether a police officer's microphone had physically breached the bedroom threshold, the Court now considered whether society was bound to respect a personal desire that the police not listen in on pillow talk.[33] In the post-*Katz* world, if the answer to the latter inquiry was "no," it mattered little how the police accomplished their eavesdropping.

The beauty of the evolved construction was that its flexibility provided, at least theoretically, broader protection than the unyielding physical invasion test. By focusing on society's expectations of privacy and not on the narrow means that permitted official intrusion, the Court infused elasticity into the analysis that could be responsive to advances in technology. As Justice Harlan explained in *Katz*, the Court's earlier trespass-based interpretation of the Fourth Amendment "is, in the present day, bad physics as well as bad law, for reasonable expectations of privacy may be defeated by electronic as well as physical invasion."[34]

Allowing the Court to define realms of protection based upon societal norms and not physical boundaries provided a flexibility that could erect zones of privacy independent of geography.[35] However, the flexibility of the Court's post-*Katz* analysis—flexibility that was once lauded as its greatest attribute—has become the focus of its greatest criticism. Following *Katz*, the mallea-

[32] *Id.* at 351.
[33] *Id.* at 361 (Harlan, J., concurring).
[34] *Id.* at 362 (Harlan, J., concurring).
[35] *Id.* at 351.

bility of the standard was decried in both conservative and liberal circles as ruinous.[36] With the increased flexibility, legitimate questions arose about how to draw clear lines around what was being protected. This struggle was seen most recently in the area of location tracking and the Court's decision in the *Jones* case.[37]

### B. LOCATION TRACKING AS PRIVACY VIOLATION

Jones was a nightclub owner in the District of Columbia. He was also suspected by law enforcement of dealing drugs. Investigating their suspicions, the local police, working in concert with federal agents, attached a GPS tracking device to Jones's car. Based, in part, upon thousands of pages of location information gathered from the device over a four-week period, Jones was convicted of a drug trafficking conspiracy and other narcotics offenses. He was sentenced to life in prison. Because the police did not adhere to the limitations of the warrant they obtained, however, it became disputed whether Jones's reasonable expectation of privacy was controlled by the Court's earlier decision in *United States v. Knotts*,[38] which approved warrantless location tracking by means of an electronic beeper. While the prosecution argued for the application of *Knotts*, the defense sought to distinguish it.

[36] *See, e.g.*, Richard A. Posner, *The Uncertain Protection of Privacy by the Supreme Court*, 1979 Sup .Ct. Rev. 173; Minnesota v. Carter, 525 U.S. 83, 97 (1998) (Scalia, J., concurring); Akhil Reed Amar, *Fourth Amendment First Principles*, 107 Harv. L. Rev. 757, 759 (1994) ("Fourth Amendment case law is a sinking ocean liner—rudderless and badly off course—yet most scholarship contents itself with rearranging the deck chairs."); Daniel J. Solove, *Fourth Amendment Pragmatism*, 51 B.C. L. Rev. 1511, 1514 (2010) ("We should sidestep the contentious debate about expectations of privacy. . . .")

[37] United States v. Jones, 132 S. Ct. 945 (2012).

[38] United States v. Knotts, 460 U.S. 276, 281 (1983).

In the view of the defense, the GPS unit's ability to collect and store massive amounts of detailed location tracking data for extended periods justified a different constitutional treatment. The defense argued that the enhanced technology represented a change in the substance of the investigation, not simply a change in the form of surveillance. The government in turn argued that it mattered little whether it tracked Jones for two days or two months, whether it used an electronic beeper or a GPS unit. In the government's view, Jones had no reasonable expectation of privacy in his movements on public streets. The *Jones* case, thus, presented what many saw as a difficult but unavoidable choice between two competing understandings of what it means to have a reasonable expectation of privacy under *Katz*. To the delight of some and the dismay of others, however, the Court resolved the case without answering the question.[39]

Rather than deciding whether the extended warrantless tracking violated Jones's reasonable expectation of privacy, the Court instead found that the attachment of the tracking device to Jones's car (coupled with the monitoring of that device) constituted a search within the meaning of the Fourth Amendment. In a unanimous decision, the Court announced that the reasonable expectation of privacy test adopted in *Katz* supplemented (rather than replaced) traditional trespass-based understandings of Fourth Amendment protection exemplified by the Court in *Olmstead*.[40] Thus, after *Jones*, a violation of the Fourth Amendment can be estab-

[39] *See* United States v. Jones, 132 S. Ct. 945, 954 (2012).

[40] *Id.* at 950 (holding that "for most of our history the Fourth Amendment was understood to embody a particular concern for government trespass upon the areas ('persons, houses, papers, and effects') it enumerates. Katz did not repudiate that understanding."); *id*. at 952 (finding that "the *Katz* reasonable-expectation-of-privacy test has been *added to,* not *substituted for,* the common-law trespassory test.") (emphasis in original).

lished with a showing that law enforcement attempted to gather information *either* by an unauthorized physical intrusion of a protected space (the *Olmstead* test) *or* by invading reasonable expectations of privacy (the *Katz* test).[41] In Jones's case, where the monitoring of his movements was accomplished by an unauthorized physical intrusion—attaching the device to the car—the Court held the conduct was unconstitutional on that ground alone.[42]

The Court's refusal to go further and resolve whether the government's conduct in the case would have been unconstitutional under a straightforward application of *Katz*'s reasonable expectation of privacy test reflected the difficulty of translating the concept of objective reasonableness through a quantitative lens. As application of that test presented tricky (and, in his view, unnecessary) questions of line drawing, Justice Scalia, writing for the majority, stated:

> [I]t remains unexplained why a 4-week investigation is "surely" too long . . . . What of a 2-day monitoring of a suspected purveyor of stolen electronics? Or of a 6-month monitoring of a suspected terrorist? We may have to grapple with these "vexing problems" in some future case where a classic trespassory search is not involved and re-

[41] *Id.; see also id.* at 951 n.5 (finding that "[a] trespass on 'houses' or 'effects' or a *Katz* invasion of privacy, is not alone a search unless it is done to obtain information; and the obtaining of information is not alone a search unless it is achieved by such a trespass or invasion of privacy.").

[42] *Jones*, 132 S. Ct. at 949. In *United States v. Katzin*, 732 F.3d 187 (3d Cir. 2013) (vacated by, rehearing, en banc, granted by *United States v. Katzin*, LEXIS 24722 (3d Cir. 2013)), the court discussed and rejected applicability of the automobile exception for warrantless searches. It found that attaching and monitoring a GPS tracker does not serve the purpose of the exception, which consists of permitting law enforcement to preserve existing evidence in an automobile that otherwise might be lost due to automobiles' mobility.

> sort must be had to the *Katz* analysis; but there is no reason for rushing forward to resolve them here.[43]

Finding that the issue in *Jones* could be decided on physical intrusion grounds alone, the majority chose to avoid the "thornier" questions required to assess reasonable expectations of privacy.

### C. THE EMERGENCE OF THE MOSAIC THEORY

The "thornier" questions identified by the *Jones* majority are addressed by what has come to be known as the "mosaic theory."[44] This theory submits that a Fourth Amendment search can be understood either as an individual act by the police or as a sequence of acts in a longer investigation. In the latter case, individual acts by the police are simply tiles in the mosaic; the full picture is what is analyzed under the Fourth Amendment. The mosaic theory is seen as more protective of privacy because obtaining and analyzing the full mosaic may constitute a Fourth Amendment search even if none of the individual tiles trigger constitutional scrutiny.

[43] *Jones*, 132 S. Ct. at 954 (citation omitted).

[44] The term "mosaic theory" was used by the Court of Appeals for the District of Columbia Circuit in *United States v. Maynard*, 615 F.3d 544, 562 (D.C. Cir. 2010) *aff'd sub nom. United States v. Jones*, 132 S. Ct. 945 (2012) ("As with the 'mosaic theory' often invoked by the Government in cases involving national security information, 'What may seem trivial to the uninformed, may appear of great moment to one who has a broad view of the scene.'") (citation omitted). As the court explained, the mosaic theory originated in national security law, particularly, the Freedom of Information Act (FOIA), and is defined in 32 C.F.R. § 701.31 (2005) as "[t]he concept that apparently harmless pieces of information when assembled together could reveal a damaging picture." The term was then referenced by law professor Orin Kerr in a blog post that he published the day the decision in *Maynard* was handed down. *See* Orin Kerr, *D.C. Circuit Introduces 'Mosaic Theory' of Fourth Amendment, Holds GPS Monitoring a Fourth Amendment Search*, VOLOKH CONSPIRACY (Aug. 6, 2010), *available at* http://volokh.com/2010/08/06/d-c-circuit-introduces-mosaic-theory-of-fourth-amendment-holds-gps-monitoring-a-fourth-amendment-search. It has since been embraced by many scholars writing in the field.

Before the *Jones* case reached the Supreme Court, it had been analyzed by the Court of Appeals for the District of Columbia Circuit using the mosaic theory.[45] The court found that the extended surveillance of a target vehicle over the course of some twenty-eight days constituted a warrantless search that was prohibited by the Fourth Amendment. When the case reached the Supreme Court, the majority declined to adopt the mosaic theory articulated by the *Maynard* court. However, while the Court as a whole declined to wade into the fray, five justices (though divided on the precise details) did not share such reticence. As Justice Alito announced, "I would analyze the question presented in this case by asking whether respondent's reasonable expectations of privacy were violated by the long-term monitoring of the movements of the vehicle he drove."[46] In keeping with this sentiment, the concurring justices in *Jones* in two separate opinions took the *Katz* inquiry head on and appear ready to overlay *Katz*'s objective reasonableness prong, in one form or another, with considerations of the mosaic theory.[47]

[45] United States v. Maynard, 615 F.3d 544 (D.C Cir. 2010), *aff'd sub nom.* United States v. Jones, 132 S. Ct. 945 (2012).

[46] Jones, 132 S. Ct. at 958.

[47] Justice Alito's concurring opinion endorsing a mosaic theory of privacy was joined by Justices Ginsburg, Breyer, and Kagan. *See* Jones, 132 S. Ct at 957 (Alito, J., concurring). Justice Sotomayor also wrote separately. *See id.* at 954 (Sotomayor, J., concurring). In her concurrence, Justice Sotomayor similarly expressed a willingness to infuse *Katz* with a quantitative understanding of objective reasonableness. Echoing the late Justice Marshall, Justice Sotomayor then went a step further, and urged reconsideration of the third-party doctrine—a doctrine cited by earlier Courts to defeat Fourth Amendment protection in a host of cases, including *Smith v. Maryland*, 442 U.S. 735, 744-45 (1979), where information was already disclosed to a third party. *Id.* at 957 ("This approach is ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks."). Indeed, some state constitutions do not adhere to the third party doctrine. For example, in *New Jersey v. Earls*, 214 N.J. 564 (2013), the Supreme Court of New Jersey concluded that the privacy protections in N.J. CONST. art. I, para. 7, which are similar to the Fourth Amendment, generally require law enforcement

For example, in her concurrence, Justice Sotomayor announced that, in assessing objective reasonableness under *Katz*, it is relevant that "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."[48] Expressing more plainly her belief that the accumulation of even seemingly innocuous data points might be relevant to constitutional protection, Justice Sotomayor wrote:

> I would take these attributes of GPS monitoring into account when considering the existence of a reasonable societal expectation of privacy in the sum of one's public movements. I would ask whether people reasonably expect that their movements will be recorded and aggregated in a manner that enables the Government to ascertain, more or less at will, their political and religious beliefs, sexual habits, and so on.[49]

Writing for three fellow justices, Justice Alito, too, expressed support for the contention that the government's accumulation of discrete location data points over a period of four weeks was determinative of *Katz*'s objective reasonableness inquiry. In the view of these four justices:

> [T]he use of longer term GPS monitoring in investigations of most offenses impinges on expectations of privacy. For such offenses, society's expectation has been that law enforcement agents and others would not—and indeed, in the main, simply could not—secretly monitor and catalogue

officers to obtain a warrant when requesting cell phone location tracking data from third party phone service providers.

[48] *Jones*, 132 S. Ct. at 955 (Sotomayor, J., concurring).

[49] *Id.* at 956 (Sotomayor, J., concurring).

> every single movement of an individual's car for a very long period.[50]

Significantly, the concurrences in *Jones* were built upon the Court's decades-old observation in *Knotts* that a resource-intensive, round-the-clock, dragnet-type surveillance might justify different constitutional treatment than a low-cost surveillance by a single officer following a suspect in a car for a limited time period.[51]

The concurring justices' willingness to operationalize the observation in *Knotts* marked something of a departure from the Court's prior cases. By and large, the Court's past consideration of technologically enhanced surveillance has treated new forms of surveillance as changes in investigative form, not substance.[52] In the

[50] *Id.* at 964 (Alito, J., concurring).

[51] United States v. Knotts, 460 U.S. 276, 283-84 (1983).

[52] *See* United States v. White, 401 U.S. at 745, 785 (1970) (Harlan, J., dissenting) (observing that "[t]he contention is, in essence, an argument that the distinction between third-party monitoring and other undercover techniques is one of form and not substance. The force of the contention depends on the evaluation of two separable but intertwined assumptions: first, that there is no greater invasion of privacy in the third-party situation, and, second, that uncontrolled consensual surveillance in an electronic age is a tolerable technique of law enforcement, given the values and goals of our political system."). The Court's decision in *Kyllo v. United States* is one clear exception to its general approach to enhanced surveillance—form, not substance. 533 U.S. 27 (2001). Rejecting the observation that equivalent information might have been obtained through unenhanced surveillance, the Court in *Kyllo* determined that the technologically enhanced search was substantively different and thus warranted different constitutional treatment. *Id.* at 35 n.2 ("The fact that equivalent information could sometimes be obtained by other means does not make lawful the use of means that violate the Fourth Amendment."). Four justices, however, rejected this conclusion. Comparing the information revealed by a thermal imager to information apparent to any passerby, the dissenters found the use of the imager a change in investigative form only—and thus not entitled to novel constitutional treatment. *Id.* at 43 (Stevens, J., dissenting) ("Indeed, the ordinary use of the senses might enable a neighbor or passerby to notice the heat emanating from a building, particularly if it is vented, as was the case here. Additionally, any member of the public might notice that one part of a house is warmer than another part or a nearby

Court's view, mere changes in the form of surveillance did not justify novel constitutional treatments. Indeed, the Court has oft repeated the refrain that the Fourth Amendment is not an impediment to improved police efficiency.[53] Particularly, the Court approved the warrantless use of beeper location tracking devices because, in the Court's view, a human tail could obtain similar information. Approving the use of such a device in *Knotts*, the Court commented:

> The fact that the officers in this case relied not only on visual surveillance, but on the use of the beeper to signal the presence of [co-defendant] Petschen's automobile to the police receiver, does not alter the situation. Nothing in the Fourth Amendment prohibited the police from augmenting the sensory faculties bestowed upon them at birth with such enhancement as science and technology afforded them in this case.[54]

However, as mosaic theorists have pointed out (and as the Court has at times acknowledged), the above approach is too simplistic; it depends, in part, on the false assumption that no greater invasion of privacy is occasioned by technologically enhanced surveillance. But as technology increases our ability to store, compare, and continuously obtain new data streams from multiple targets, there is growing recognition of the fact that, in some instances, technological advances do more than simply make police work more efficient; sometimes those advances radically change the substance of the investigation. Such a difference in kind (not just degree), the argument goes, warrants different constitutional treat-

building if, for example, rainwater evaporates or snow melts at different rates across its surfaces.").

[53] *Knotts*, 460 U.S. at 284 ("We have never equated police efficiency with unconstitutionality, and we decline to do so now.").

[54] *Id.* at 281.

ment. The challenges, though, are twofold. First, we must objectively confirm that it is possible, as an absolute matter, for a difference in kind to come to pass. Second, we must identify the point at which that change occurs. As is discussed in greater detail in the sections below, the science provides a clear answer to the first query–machine learning can demonstrate objectively that the collection of numerous data points will eventually tell the observer more than the sum of the data collected. Moreover, while a clear answer to the second question depends on the details of the investigation, we are in principle able to provide such.

As a legal matter, critics of the mosaic theory have identified the above as the two most persuasive challenges facing the theory. Justice Scalia noted this in his majority opinion in *Jones*, commenting, "it remains unexplained why a 4-week investigation is 'surely' too long."[55] Legal academics have echoed a similar grievance in their writing.[56] Thus, in addition to the lessons that can be learned on the scientific front, we should also begin thinking how we might anchor those lessons in the existing legal landscape. In this regard, it should be noted that the scientific advances, which we describe in the sections to follow,[57] are still in development. Accordingly, until we are able to answer with greater objectivity the precise moment at which a change occurs, the applicable legal rules will necessarily be something less than fully developed. In this sense we now turn to consider whether what machine learning currently makes possible can be squared with aspects of existing Fourth Amendment protection. We suggest that the minimal constitutional protection historically afforded to particular types of information and the Court's past willingness to adopt mathematically bright lines in connection

[55] United States v. Jones, 132 S. Ct. 945, 954 (2012).
[56] *See*, *e.g.*, Kerr, *supra* note 2, at 311.
[57] *See infra* Sections II, III.

with other legal concepts are anchoring points around which future courts can begin to structure their thinking as they seek to identify the threshold at which enough is enough–the point at which long-term government surveillance becomes objectively unreasonable.

### D. The Privacy of the Home as a Basis for location tracking.

Interestingly, with the current state of the science, the most relevant strand of precedent comes not from the Court's past adjudication of tracking devices; but rather from the Court's treatment of information about the home. Without question, while no place is afforded unqualified "status" protection under the Fourth Amendment,[58] the Court has consistently said that the home will be afforded the greatest protection possible.[59] Thus, in *New York v. Payton*, the Court acknowledged that "physical entry of the home is the chief evil against which the wording of the Fourth Amendment is directed."[60] In contrast, in *Oliver v. United States*, the Court declined to protect an "open field" behind Oliver's home because "open fields do not provide the setting for those intimate activities that the Amendment is intended to shelter from government interference or surveillance."[61]

Moving one rung up the ladder of abstraction, the Court in protecting the home has articulated a standard that encompasses not only the physical space, but also details about the activities occur-

[58] Oliver v. United States, 466 U.S. 170, 177-78 (observing that "[n]o single factor determines whether an individual legitimately may claim under the Fourth Amendment that a place should be free of government intrusion not authorized by warrant.") (citing Rakas v. Illinois, 439 U.S. 128, 152-153 (1978) (Powell, J., concurring)).

[59] *See*, *e.g*., United States v. Kyllo, 533 U.S. 27 (2001).

[60] 445 U.S. 573, 585-86 (1980) (citing United States v. United States District Court, 407 U.S. 297); *see also id.* at 601 (noting "the sanctity of the home that has been embedded in our traditions since the origins of the Republic.").

[61] *Oliver*, 466 U.S. at 179.

ring therein. Notably, in *Kyllo v. United States*,[62] federal agents chose to investigate a suspected marijuana grower by scanning his home one evening with a thermal imaging device that revealed areas of relative heat. There was no physical intrusion into the suspect's house. Instead, officers were able to observe remotely an area of extreme heat over the garage, which they believed to be consistent with use of the high intensity halide lamps needed to grow marijuana indoors. Following conviction, Kyllo challenged the warrantless use of the imager. Starting from the premise that the warrantless search of a home is, with few exceptions, unconstitutional, the Court found that use of the imager was unlawful because the information it obtained could not otherwise have been gathered without physical trespass into the home's interior.[63]

In the now often-repeated quote from the *Kyllo* decision, the Court declared warrantless use of the thermal imager unconstitutional where the device might "disclose, for example, at what hour each night the lady of the house takes her daily sauna and bath."[64] Declaring details that are traditionally associated with the intimacies of home life protected, the Court held that the warrantless scan of Kyllo's home violated the Fourth Amendment. Without question, the *Kyllo* Court was unwilling to present a laundry list of "intimate details" that it considered worthy of protection. Rather, the Court noted, in the context of the home *all* details are intimate whether those details be the color of the rug in the front hallway or the tim-

[62] United States v. Kyllo, 533 U.S. 27 (2001).

[63] *Id.* at 34-35; *see also* United States v. Karo, 468 U.S. 705 (1984) (striking down the warrantless use of an electronic device tracking the location of a can of ether in a private residence because "had a [Drug Enforcement Administration] agent thought it useful to enter the . . . residence to verify that the ether was actually in the house and had he done so surreptitiously and without a warrant, there is little doubt that he would have engaged in an unreasonable search within the meaning of the Fourth Amendment.").

[64] *Kyllo*, 533 U.S. at 38.

ing of the resident's evening soak. In the Court's view, "obtaining by sense-enhancing technology any information regarding the interior of the home that could not otherwise have been obtained without physical 'intrusion into a constitutionally protected area,' constitutes a search."[65]

In *Florida v. Jardines*, the Court again affirmed that for purposes of Fourth Amendment protection the "home is first among equals."[66] In that case, police suspected that Jardines was growing marijuana in his home. Officers set up surveillance at the residence and determined that Jardines was not home. The officers then sent a drug-sniffing dog and his trainer onto the porch of the house to see if the dog would alert. After several minutes the dog did in fact alert by sitting down at the front door to indicate that it was the source of the strongest odor. The officers left and obtained a warrant based, in part, upon the drug dog's alert at the home's front door. A subsequent search of the house revealed a marijuana growing operation. Jardines challenged the validity of the warrant. He argued that the dog sniff on the front porch constituted a warrantless search within the meaning of the Fourth Amendment. The Court agreed. Of particular relevance to the discussion here, the Court noted that at the very core of Fourth Amendment protection is the right of persons to retreat into their homes free of unwanted government scrutiny. "This right would be of little practical value if the State's agents could stand in a home's porch or side garden and trawl for evidence with impunity."[67]

For purposes of the present conversation, there would be little gained from the Court's historic protection of the home if that protection were motivated solely by concern for the physical space.

[65] *Id.* at 34 (citing Silverman v. United States, 365 U.S. 505, 512 (1961)).
[66] 133 S. Ct. 1409, 1414 (2013).
[67] *Id.*

However, as the Court's decisions make clear, the Fourth Amendment sanctity of the home is about something much broader. Describing the rationale underlying constitutional protection of the intimate activities of the home, the Court has explained that some refuge from public scrutiny is necessary to the concept of ordered liberty:

> A man can still control a small part of his environment, his house; he can retreat thence from outsiders, secure in the knowledge that they cannot get at him without disobeying the Constitution. That is still a sizable hunk of liberty—worth protecting from encroachment. A sane, decent, civilized society must provide some such oasis, some shelter from public scrutiny, some insulated enclosure, some enclave, some inviolate place which is a man's castle.[68]

Echoing a similar understanding of the principles underlying the Fourth Amendment's protection of the home, Justice Kagan in her concurrence in *Jardines* described the police conduct there as objectionable not simply because of the intrusion into a private physical space, but because that intrusion was used to "nos[e] into intimacies you sensibly thought protected from disclosure."[69] Stating plainly the broader principles inspiring the home's protection, Justice Kagan wrote, "And so the sentiment 'my home is my own,' while originating in property law, now also denotes a common understanding—extending even beyond that law's formal protections—about an especially private sphere. Jardines's home was his property; it was also his most intimate and familiar space."[70]

[68] Silverman v. United States, 365 U.S. 505, 512 n.4 (1961) (emphasis added) (citing United States v. On Lee, 193 F.2d 306, 315-16 (C.A.2) (Frank, J., dissenting)).

[69] Florida v. Jardines, 133 S. Ct. at 1418 (Kagan, J., concurring).

[70] *Id.* at 1419.

Without question, these cases do not provide a completely satisfactory answer. To be meaningful, the protection offered by the mosaic theory will need to do more than offer the protection already provided. However, we contend simply that the principles undergirding the home's constitutional protection are a starting point. They provide some guidance in thinking about when, *at a bare minimum*, discrete units of location data will combine to form a mosaic worthy of constitutional protection. In other words, machine learning provides a useful anchor by telling us objectively that aggregation of location tracking data will at point *x* begin to reveal information akin to that which has already received the protection just discussed.

In thinking about how existing legal standards might inform a mosaic theory of Fourth Amendment protection, another piece of the puzzle is provided by the Court's refusal to protect information about the home where, in the Court's view, that information was held out to public scrutiny. The Court's treatment of information held out for public scrutiny helps inform our thinking about where we might defensibly place an outer limit. For example, in *California v. Greenwood* the Court determined that the police could, without a warrant, search sealed trash bags left at the curb for collection.[71] Certainly the information in Greenwood's trash told the police something about what was going on in Greenwood's home. But, explaining the holding, Justice White, stated that, "respondents exposed their garbage to the public sufficiently to defeat their claim to Fourth Amendment protection." [72] In the Court's view, while Greenwood may not have wanted the police to go through his garbage, that expectation was not one that society recognized as reasonable. Because wild animals and mischievous children might rifle

[71] California v. Greenwood, 486 U.S. 35 (1988).
[72] *Id.* at 40.

through trash bags at the curb, the Court reasoned a homeowner should not expect the police to abstain from similar conduct.

Just one year later, in *Florida v. Riley*, the Court authorized warrantless police efforts to obtain information by flying low over Riley's five-acre property in a helicopter.[73] Riley had a mobile home and a greenhouse on the property. Two walls of the greenhouse were enclosed. The other two sides of the greenhouse were completely obscured from ground views by the mobile home, bushes, and a surrounding forest. The greenhouse and home were enclosed by a fence, which was posted with a "Do Not Enter" sign. The top of the greenhouse was almost entirely covered with translucent roofing panels. However, from the low altitude used to fly over the property, the police were able to observe, through a space left open by two missing roof panels, the marijuana plants that Riley grew inside. Asked to rule on the constitutionality of the fly-over, the Court held that any expectation of privacy that Riley may have had was unreasonable—"[b]ecause the sides and roof of his greenhouse were left partially open . . . what was growing in the greenhouse was subject to viewing from the air."[74]

The Court's decisions in *Riley*, *Kyllo*, *Jardines*, and *Greenwood* cannot, with the existing state of the science, provide "the" answer. But, they are "data points" in the Fourth Amendment landscape that provide several interesting insights. First, we can say that the privacy protection afforded to "home life" cannot be said to rise or fall with physical boundaries. In *Jardines*, the Court found a privacy violation upon physical entry onto the suspect's porch, while *Kyllo* found a similar violation with no such physical intrusion. In *Greenwood*, trash bags left outside the home were not protected; and in *Riley*, a similar conclusion was reached, even though the govern-

[73] Florida v. Riley, 488 U.S. 445 (1989).
[74] *Id.* at 450.

ment peeked into a home's backyard. Put simply, the Court's decisions reflect its move from a notion of home privacy that is dependent on physical space to a much more flexible interpretation of what constitutes "home life." Further, the manner in which the Court has drawn a line between protected "intimate details" and unprotected "public information" can help us think about where a line of minimal constitutional protection in the realm of location tracking might lie. When considering whether the mosaic theory is viable as an abstraction, one obvious question is why location tracking data should be compared with the intimate details of the home that were protected in *Jardines* and *Kyllo*, and not with the information held out to public scrutiny in *Greenwood* and *Riley*. Machine learning provides the beginning contours of an answer.

As described in greater detail below,[75] one thing we know for certain is, when aided by machine learning, discrete points of location data reveal far more about a target in the aggregate than simply a chronicle of where the target has been. Viewing that technological reality through the lens of precedent provides one possible answer to the criticism of the mosaic theory as being impossibly imprecise. If the science tells us that the collection of *x* data points enables disclosure of information "that could not otherwise have been obtained without physical 'intrusion into a constitutionally protected area,'" then a plausible argument exists that the law should, *at a bare minimum*, recognize a constitutionally significant search under the mosaic theory at the moment at which *x* data points are collected.[76] Put somewhat more plainly, it could be said as a starting point, we can understand a mosaic worthy of constitutional protection as being established when the collection of location tracking data ena-

[75] *Infra* Section IV.

[76] Certainly, if the collection of any individual data point constitutes a discrete search under existing case law, it could be as such analyzed without resort to the mosaic theory.

bles the police to learn intimate details about a target's home life that could not otherwise be learned without intrusion into the target's private realm. To be certain, this is just a minimal starting point in thinking about where the appropriate layer of constitutional protection must lie, for it goes without saying that the Constitution protects reasonable expectations of privacy well beyond the four walls of the abode.

### E. QUANTIFYING THE MOSAIC

As the above demonstrates, the abstract notion of the mosaic theory can be preliminarily aligned with privacy notions that have previously been articulated in the case law. However, without further development of the science it will be difficult to objectively articulate the precise contours of the theory. Thus, even if it can be said that the minimum level of constitutional protection is tripped when location data enables the discovery of the type of information that already enjoys constitutional protection, the question of line drawing remains. In this section, we explore this line drawing and whether there is any support in the precedent for precise quantification of legal concepts. Efforts to imbue inexact legal concepts with some aspects of numeric measurement are not unique to privacy. In other areas of the law, similar suggestions have been made to translate relatively amorphous notions into more certain mathematical models.[77] The Court, though, has most often declined to endorse a precise mathematical formulation.[78]

[77] *See e.g.*, Ronald J. Bacigal, *Making the Right Gamble: The Odds on Probable Cause*, 74 MISS. L.J. 279 (2004); Edward K. Cheng, *Reconceptualizing the Burden of Proof*, 122 YALE L.J. 1254 (2013); Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970); Erica Goldberg, *Getting Beyond Intuition in the Probable Cause Inquiry*, 17 LEWIS & CLARK L. REV. 1065; John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968); C.M.A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitu-*

In *Maryland v. Pringle*, for example, the Court noted that "the probable-cause standard is a 'practical, nontechnical conception' that deals with 'the factual and practical considerations of everyday life on which reasonable and prudent men, not legal technicians, act.'"[79] Under the facts of that case, a police officer stopped a car for speeding, and, after searching it, found cocaine, of which all three passengers in the car denied ownership.[80] Absent any other facts, each passenger, as a mathematical proposition, was likely to have committed a narcotics offense with a probability of only one-third. Nonetheless, the Court found that the officer had probable cause to arrest the respondent Pringle, one of the three passengers.[81] Though some argued that the Court's decision signaled a new mathematical understanding of probable cause, *i.e.*, 33⅓%, the Court made clear it was not adopting a precise quantitative definition of the term: "[t]he probable cause standard is incapable of precise definition or quantification into percentages because it deals with probabilities and depends on the totality of the circumstances." On the totality

*tional Guarantees?*, 35 VAND. L. REV. 1293 (1982); Michael J. Saks & Robert F. Kidd, *Human Information Processing and Adjudication: Trial by Heuristics*, 15 LAW & SOC'Y REV. 123 (1980-1981); Barbara D. Underwood, *Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment*, 88 YALE L.J. 1408 (1979). *But see, e.g.*, J. D. Jackson, *Probability and Mathematics in Court Fact-Finding*, 31 N. IR. LEGAL Q. 239 (1980); Orin Kerr, *Why Courts Should not Quantify Probable Cause,* THE POLITICAL HEART OF CRIMINAL PROCEDURE 131 (Michael Klarman et al. eds., 2012); Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971).

[78] *See*, *e.g.*, Illinois v. Gates, 462 U.S. 213, 235 (1983) (stating that "an effort to fix some general, numerically precise degree of certainty corresponding to 'probable cause' may not be helpful"). *But see*, *e.g.*, United States v. Schipani, 289 F. Supp. 43, 55-56 (E.D.N.Y. 1968) (stating that the proof of a fact by preponderance of evidence requires a probability of at least 50%).

[79] Maryland v. Pringle, 540 U.S. 366, 370 (2003) (quoting Illinois v. Gates, 462 U.S. 213, 231 (1983)); Illinois v. Gates, 462 U.S. 213, 231 (1983) (quoting Brinegar v. United States, 338 U.S. 160, 175 (1949)); Brinegar v. United States, 338 U.S. 160, 175 (1949).

[80] *Pringle*, 540 U.S. at 368.

[81] *See id.* at 374.

before it, the *Pringle* Court found that "[t]he quantity of drugs and cash in the car indicated the *likelihood* of drug dealing, [to be] an enterprise [among the three passengers] to which a dealer would be *unlikely* to admit an innocent person with the potential to furnish evidence against him."[82]

The Court's reluctance to quantify other legal standards can be seen in its treatment of proof beyond a reasonable doubt and preponderance of the evidence. Noting the usefulness of these standards despite their inability to be quantified, Justice Harlan stated in his concurring opinion in *In re Winship* that "[a]lthough the phrases 'preponderance of the evidence' and 'proof beyond a reasonable doubt' are quantitatively imprecise, they do communicate to the finder of fact different notions concerning the degree of confidence he is expected to have in the correctness of his factual conclusions."[83]

Notwithstanding the above, however, it would be inaccurate to suggest that the Court always eschews the comparative certainty that comes with mathematically precise bright lines. Accordingly, the Court's refusal to adopt a quantitative understanding of a term like probable cause does not stand in the way of the instant suggestion that a more precise understanding of the mosaic theory can (and should) be informed by the developing objective scientific notions. Though the Court's present reluctance to embrace the mosaic theory appears to be driven in part by reluctance to draw an arbitrary constitutional line in the field of location tracking, it has not

[82] *Id*. at 373 (emphasis added); Florida v. Harris, 133 S. Ct. 1050, 1055 (2013) (noting that the "test for probable cause is not reducible to 'precise definition or quantification.'"); Ornelas v. United States, 517 U.S. 690, 695 ("Articulating precisely what 'reasonable suspicion' and 'probable cause' mean is not possible.").

[83] 397 U.S. 358, 370 (1970) (Harlan, J., concurring).

been so reserved universally.[84] In certain contexts, the Court has embraced numerical approaches to rule-making, even while admitting that the precise point selected was somewhat arbitrary.

For example, the Court has determined that a custodial suspect's request for counsel will not bar further uncounseled questioning, so long as the suspect has experienced a fourteen-day "break" in custody.[85] In that case, *Maryland v. Shatzer*, Shatzer, an inmate at a Maryland prison, was questioned about the sexual abuse of his son. After being given *Miranda* warnings, Shatzer indicated that he wanted to speak with an attorney. The questioning detective left, and Shatzer was returned to the general prison population. Three years later, a second detective visited Shatzer, gave him *Miranda* warnings, and questioned him again about the abuse. During this second round of questioning, Shatzer made incriminating statements. Finding that the statements were not obtained in violation of Shatzer's rights, the Supreme Court held that "once the suspect has been out of custody long enough (14 days) to eliminate the coercive effect, there will be nothing to gain" by continuing to recognize a prohibition on future questioning.[86] Writing in concurrence, Justice Stevens noted, "Today's decision . . . offers no reason for its 14-day time period. To be sure, it may be difficult to marshal conclusive evidence when setting an arbitrary time period."[87]

Another instance in which the Court has been willing to quantify constitutional protection to advance a simple rule is with regard

[84] With regard to the Fifth Amendment, the Court has readily acknowledged that lines it has drawn are arbitrary but bright. The Court's refusal to quantify legal concepts like reasonable doubt and probable cause is driven less by a concern for arbitrariness, and more by an appreciation for the complex mental processes underlying such evaluations. In the Court's view quantification in such instances would do more harm than good.

[85] *See* Maryland v. Shatzer, 559 U.S. 98, 130 S. Ct. 1213 (2010).

[86] *Id*. at 1223.

[87] *Id.* at 1231 n.7 (Stevens, J., concurring).

to binary searches (searches that can produce only two results). The Court has clearly said that binary searches at both the high and low ends of the technological scale are permissible without a warrant because of the limited quantity of data they reveal, assuming they do not tread upon other constitutional protections.[88]

Though the Court in some cases has been unwilling to quantify legal concepts, in others it has found that hard numbers aid in the articulation of legal standards. Most importantly, with regard to location tracking, the Court has previously found that the quantity of information collected may be relevant to the intrusiveness of the government's conduct, and thus would be relevant to the appropriate level of constitutional protection afforded.[89] Thus, the suggestion that practical implementation of the mosaic theory will benefit from a more quantitative understanding of objective reasonableness is not contrary to existing doctrine.

The urge to quantify the Fourth Amendment's protection in the context of location tracking is, in part, a call for greater objectivity and, in part, a call for greater protection. As the unanimous decision in *Jones* reflects, however, the Court is not quite ready to make the leap. The Court's reluctance to fully embrace the mosaic theory in this context is not unwarranted. While the concerns of the concurring justices in *Jones* are readily understood at a visceral level, they are more difficult to defend objectively. And while the Court has, at times, been willing to embrace arbitrary numerical standards, a stronger case for change is made if one can explain why it would improve the status quo. As Justice Scalia wryly explained in the majority opinion language quoted above,[90] the quantification of objec-

[88] *Compare* Illinois v. Caballes, 543 U.S. 405, 408-09 (2005); United States v. Place, 462 U.S. 696, 707 (1983); United States v. Jacobsen, 466 U.S. 109, 123-24 (1984) *with* Florida v. Jardines, 133 S. Ct. 1409, 1414 (2013).

[89] United States v. Knotts, 460 U.S. 276, 284 (1983).

[90] *Supra* Section I.0

tive reasonableness advanced by the concurrences is hardly more clear-cut than the generic objective reasonableness standard under *Katz* that it seeks to enhance. Indeed, even the concurring justices conceded that they could not identify the precise point at which monitoring moved from permissible to unconstitutional: "We need not identify with precision the point at which the tracking of this vehicle became a search, for the line was surely crossed before the 4-week mark."[91]

However, what the concurring justices in *Jones* recognized, and what the Court's prior guidance tells us is that it is generally possible to identify a minimum point at which constitutional protection must attach. Put somewhat differently, there is an upper bound for a period of time at which technologically-aided location tracking stops being simply more efficient surveillance and becomes something altogether different substantively. The lessons of machine learning help us to understand where that upper temporal bound lies for they help us to understand exactly what can be learned from the aggregation of various types of data. Moreover, those same lessons will help us more clearly identify to what extent the upper bound can be lowered. If these lessons are taken seriously, the imprecision decried in *Jones* will not be a barrier to quantification much longer.

Before turning to a discussion of the power of machine learning, it should be noted that in constructing the jurisprudence of the Fourth Amendment, the Court has expressed concern for both the current state of scientific knowledge and its likely future ability.[92] A scientific understanding of location tracking that will help to make future abilities clear would thus do much to advance the discussion.

[91] United States v. Jones, 132 S. Ct. 945, 964 (2012) (Alito, J., concurring).

[92] Kyllo v. United States, 533 U.S. 27, 36 (2001) ("While the technology used in the present case was relatively crude, the rule we adopt must take account of more sophisticated systems that are already in use or in development.").

The machine learning principles, described in Section II., combined with the privacy metrics described in Section III., do just that. They provide a rationale for according differential legal treatment to technologically enhanced location tracking of different durations.[93] They help explain why location tracking data gathered for $x$ data points can be substantively different than location tracking data gathered for $x + 2$ data points or $x + 2$ time units. In this sense, machine learning and privacy metrics provide a dispassionate explanation for the *Jones* concurrences' intuitive belief that GPS monitoring of a suspect for twenty-eight days is different than the only hours-long beeper monitoring at issue in *Knotts*.

## II. MACHINE LEARNING

As discussed earlier,[94] under the mosaic theory, a sequence of acts may constitute a Fourth Amendment search even if none of the individual acts trigger constitutional scrutiny. This insight is the core element of the mosaic theory. It acknowledges that the aggregation of observations about a person can lead to a picture that is more revealing than the sum of the individual observations. However, how is this possible? That is the question to which machine learning provides an answer.

Machine learning is a field that seeks to harness today's exponential data deluge by finding patterns in it, making predictions from it, and efficiently organizing it. Machine learning leverages large-scale efficient algorithms from computer science and principled inference methods from statistics. However, machine learning can also be potentially invasive if applied to location data or other data: it can deduce information that may otherwise have been pro-

[93] Related prediction programs are already being used by law enforcement. For example, Mosaic 20 is a domestic violence prediction program currently in use.

[94] *See supra* Section I.0

tected by the natural limitations of manual and human-driven investigation.

Machine learning works best when given a large training set of observations (ideally drawn in some independent manner) with which it estimates models. These models are then used to make predictions on future data outputting a probability measure for the occurrence of an event or existence of a fact. The train/test paradigm can largely be automated and also reliably evaluated. Three natural regimes can be distinguished: unsupervised machine learning, supervised machine learning, and semi-supervised machine learning. Each will be discussed in turn. We caution that this is a very brief overview of a highly mathematical branch of computer science.

## A. UNSUPERVISED MACHINE LEARNING

In unsupervised machine learning, a dataset describing $n$ people is measured and stored as $\{x_1, \dots, x_n\}$. Here, each $x_i$ refers to all the data collected about user $i$ (the profile or location history or some other collection of personal information).[95] A machine learning system automatically finds dependencies, correlations, and clusters in the data without requiring any significant human intervention. More specifically, it could perform the following operations:

[95] Unsupervised machine learning is an umbrella term that covers many aspects of density estimation, Bayesian inference, and maximum likelihood. Bayesian inference dates back to Reverend Thomas Bayes, FRS (1702-61) with a general overview by GEORGE E.P. BOX & GEORGE C. TIAO, BAYESIAN INFERENCE IN STATISTICAL ANALYSIS VOL. 40 (John Wiley & Sons 2011). More recent Bayesian inference approaches involve large sets of interdependent random variables as described by DAVID HECKERMAN, A TUTORIAL ON LEARNING WITH BAYESIAN NETWORKS (Springer 2008). Maximum likelihood was formalized by R.A. Fisher at the start of the 20th century as discussed by John Aldrich, *R.A. Fisher and the making of maximum likelihood 1912-1922,* STATISTICAL SCIENCE 12.3, 162-76 (1997).

- **Clustering:** In *clustering*, a system automatically finds groups of users in the dataset that appear statistically similar. For instance, certain individuals may show a pattern of visiting churches on Sundays while others stay home during that time. After application of a clustering algorithm, it becomes relatively easy for a human investigator to observe prototypes from each cluster and figure out which group it represents (for instance, followers of a particular faith, e.g., Christians). The number of groups to be extracted can be fixed (i.e., find the 5 most important groups) or can be automatically estimated. The groupings could be disjoint, overlapping, hierarchical, or nested in various ways. For instance, sub-groups of religious activity (Baptists, Roman-Catholics, Lutherans, etc.) could emerge under a larger umbrella group (Christians).

- **Detection:** Given data about individuals as an unbiased sample of the population, a *detection system* recovers a probability distribution, $p(x)$, which says how an individual likely behaves under this sample. This permits an investigator to flag anomalous users in the training data (and in future data) as individuals with a $p(x)$ score that is lower than some reasonable threshold. Alternatively, it is possible to identify the handful of users who had the lowest $p(x)$ scores as *outliers*, for example, in a location dataset those who do not exhibit regular location movement. One natural example of an outlier is the mail carrier who spends the workday going door-to-door delivering mail. This is an unusual commute pattern relative to the rest of the population.

- **Visualization and Summarization:** Another application of machine learning is *visualizing trends* in "big data" and highlighting important aspects in it. While each person's record, $x_i$, may contain thousands or millions of bytes of information, a human investigator can only visualize projections of the data in two or three dimensions. Machine learning, however, finds low-dimensional embeddings, which summarize the original data with minimal distortion. For example, the similarities or distances between pairs of visualized low-dimensional embedding-points could be almost equal to the similarities or distances that were measured between pairs of original data points. Alternatively, only the key measurements in the original data points are preserved. For example, from the thousands of latitude and longitude coordinates a user visited that are stored in $x_i$, it is possible to extract one or two important locations such as the user's home or place of work.

- **Inference:** One of the most powerful unsupervised machine learning techniques is arguably *probabilistic inference*. In particular, machine learning is able to find dependencies in parts of a collection of data gathered about users. For instance, if we have observed two types of information for many users, say, their location history and web-browsing history, a machine learning system can learn the dependence and correlations between locations and browsing. This allows the system, for example, to fill-in likely browsing patterns for a new user even though only location history for this user was available. Put another way, we can predict a user will probably visit the website espn.com frequently if that user has frequently attended sports events at stadiums.

### B. Supervised Machine Learning

In supervised machine learning, a dataset of $n$ input and target output pairs, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, is measured.[96] Here, each $x_i$ could, for example, refer to a profile, aggregation of location infor-

[96] Currently popular methods that embody supervised machine learning are described in CARL EDWARD RASMUSSEN & CHRISTOPHER K.I. WILLIAMS, GAUSSIAN PROCESSES FOR MACHINE LEARNING (MIT Press 2006).

mation, or other collection of data about a user while $y_i$ is a label with which this data has been manually annotated. For instance, $y_i$ could refer to the fact that the individual is on a suspicious list. This type of data is more laborious to create since it requires human annotation effort while unsupervised learning is more of a pure data collection exercise. With supervised learning, we can perform the following operations with varying degrees of accuracy:

- **Classification:** One of the most basic supervised machine learning operations is *classification*, that is, the identification of a category for a new observation. In addition to collecting data, $x_i$ , about an individual, classification also requires that we annotate individuals with a discrete label, $y_i$. Collecting such a categorical variable, $y_i$, about an individual often requires some effort, expense, or a need for the subject to volunteer information about themselves. For example, in addition to collecting location data, one may survey a small portion of the population and ask them to report their occupation (student, construction worker, taxi driver, etc). Then, having obtained such labels from the survey, it is possible for a machine learning system to automatically label other individuals using *only* their location data, $x_i$.

- **Regression:** While classification involves obtaining a discrete label, $y_i$, for an individual, *regression* assumes that the discrete label is a scalar. For instance, instead of a category (such as occupation), we may collect the income that the individual received last year as a numerical value. Machine learning then learns a good prediction function from training examples to accurately estimate the salary, $y_i$ of other individuals directly from their location data. For instance, by getting location data from someone who lives in an expensive neighborhood and works in the financial district, it would be possible to estimate a high income level, $y_i$.

- **Prediction:** In *prediction*, the output, $y_i$, is either discrete (as in classification) or continuous (as in regression), but is also specifically a quantity that is only available in the future after the input raw data, $x_i$, is observed from a user. For example, $y_i$ may be the location (latitude and longitude) that the user will visit tomorrow for lunch. Alternatively, $y_i$ may be the party (Republican or Democrat) that a person will vote for in the next election. By observing a population of users for some time, it may be possible to predict that user $i$ will likely go for pizza at the mall in his or her next lunch break. Prediction may help an advertising company determine what ad to target on a mobile device by delivering a relevant message (for instance, to lure the user to a new pizza establishment in the vicinity of his or her next lunch location).

While some of these supervised learning problems are difficult, with increasing amounts of data, the accuracy of the classification, regression, or prediction improves and eventually achieves surprisingly strong performance. Unfortunately, collecting labels in addition to raw data may be an expensive proposition. This leads to a third regime which attempts to leverage large amounts of cheap unsupervised raw data with small amounts of expensive labels to obtain the best of both worlds.

### C. SEMI-SUPERVISED AND NETWORK LEARNING

Semi-supervised learning has recently emerged.[97] It can be thought of as the natural blend of both supervised and unsupervised methods. As in supervised learning, on some individuals, we

[97] Xiaojin Zhu, Zoubin Ghahramani & John Lafferty, *Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions*, THE INT'L CONF. ON MACHINE LEARNING, 912-19 (2003). *See generally* OLIVIER CHAPELLE, BERNHARD SCHÖLKOPF & ALEXANDER ZIEN, SEMI-SUPERVISED LEARNING (MIT Press 2006).

have raw input data as well as a target variable. However, on the vast majority of other individuals, we only have raw input data (say, just location data) without any human label annotation. A major component of semi-supervised learning is learning with network data, which has potentially the largest implications for private and location data.

As social networks and social media proliferate, network data is quickly becoming another important alternative to the training datasets mentioned earlier. Rather than having profile information about $n$ individuals in the form of $\{x_1, \dots, x_n\}$, it is increasingly popular to gather information about interactions between pairs of $n$ individuals represented by potentially $n(n-1)/2$ edges between them in the form $\{e_{1,2}, e_{1,3}, \dots, e_{1,n}, \dots, e_{n-1,n}\}$. Each edge, $e_{i,j}$, between two individuals, $i$ and $j$, represents a relationship, such as a friendship or work relationship.

Such networks can be inferred from mobile communication and location data. For instance, people who call each other can be assumed to be friends and this leads to the formation of a friendship edge between a pair of users. Alternatively, people who spend much time together in similar locations (i.e., co-locate), would also allow an algorithm to infer the presence of an edge or relationship between those two individuals. Moreover, network datasets are natural targets for semi-supervised learning. By knowing some labels on a few individuals in a network (such as their shopping preferences), it is possible to propagate or diffuse this label information to predict labels for others nearby in their network (such as their friends and the friends of their friends).[98]

[98] The theoretical framework behind such network labeling is explicated in Xiaojin Zhu, *Semi-Supervised Learning with Graphs* (May 2005) (Ph.D. thesis, Carnegie Mellon University).

## III. PRIVACY METRICS

Machine learning provides a basis for the mosaic theory's rationale that aggregate information can reveal more than the sum of individual observations. It does not, however, provide a measure for privacy. Furthermore, the legal guidance tells us what sort of information and which realms of life have been traditionally protected, but affords little help for deciding when collected data has tripped that threshold.

However, the quantification of privacy is the subject of various privacy metrics proposed in the computer science literature. While most of these metrics are developed for measuring privacy in databases, they are also used for anonymization in location-based web services and for other location privacy purposes. This section will discuss two of those metrics: *k*-anonymity and *l*-diversity.[99] Their underlying notions can be applied to the output of a machine learn-

[99] *k*-anonymity was the starting point for a whole family of privacy metrics that built upon and extended it: for *l*-diversity *see* Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke & Muthuramakrishnan Venkitasubramaniam, *l-diversity: Privacy Beyond k-anonymity*, 1 ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA 1 (2007); for *t*-closeness *see* Ninghui Li, Tiancheng Li & Suresh Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, INT'L CONF. ON DATA ENG'G, 106 (2007) [hereinafter ICDE]; for *m*-invariance *see* Xiaokui Xiao and Yufei Tao, *m-invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets*, SPECIAL INT. GRP. ON MGMT. OF DATA, 689 (2007) [hereinafter SIGMOD]; for *δ*-presence *see* M. Ercan Nergiz, Maurizio Atzori & Christopher W. Clifton, *Hiding the Presence of Individuals From Shared Databases*, PROC. OF THE 2007 ACM SIGMOD ICDE, 665 (2007). Beyond *k*-anonymity and its progeny, one of the most influential recent privacy metrics is differential privacy. *See* Cynthia Dwork, *Differential Privacy*, 33 INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING, 4052:1 (2006). However, differential privacy is rarely used for purposes of location privacy. For one of the few exceptions *see* Rinku Dewri, *Location Privacy and Attacker Knowledge: Who Are We Fighting against?*, 7 PROC. INT'L. ICST CONF. ON SEC. AND PRIVACY IN COMMC'N NETWORKS (2011).

ing algorithm, thereby allowing for integration into Fourth Amendment doctrine.

### A. *K*-ANONYMITY

Most approaches for quantifying location privacy are based on *k*-anonymity.[100] Under the *k*-anonymity metric, which originated in the context of database privacy,[101] a release of information from a database is *k*-anonymous "if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appears in the release."[102] Applying this metric to location privacy, a person is *k*-anonymous if his or her location is indistinguishable from the location of at least $k-1$ other persons.[103] Such anonymity is achieved by spatial cloaking, that is, a trusted third party or peer-to-peer process transforms the precise location of the person to be anonymized into a larger area, known as the anonymity spatial region. This area must be large enough to contain the location of all $k$ individuals. For an area of

[100] Aris Gkoulalas-Divanis, Panos Kalnis & Vassilios S. Verykios, *Providing k-anonymity in Location Based Services*, 12 SPECIAL INT. GRP. IN KNOWLEDGE DISCOVERY & DATABASES EXPLOR. NEWSL., 3, 5 (2010). *See generally* Marco Gruteser & Dirk Grunwald, *Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking*, 1 PROC. INT'L CONF. ON MOBILE SYS., APPLICATIONS, AND SERVICES, 31 (2003), who developed an early model of *k*-anonymous location information. For further model proposals *see, e.g.*, Roberto J. Bayardo & Rakesh Agrawal, *Data Privacy through Optimal k-anonymity*, 21 PROC. INT'L CONF. ON DATA ENG'G (2005); Bugra Gedik & Ling Liu, *A Customizable k-Anonymity Model for Protecting Location Privacy*, INT'L CONF. ON DISTR. COMPUTER SYS. 1 (2005) [hereinafter ICDCS].

[101] Pierangela Samarati & Latanya Sweeney, *Protecting Privacy when Disclosing Information: k-anonymity and Its Enforcement through Generalization and Suppression*, Tech. Report SRI-CSL-98-04, SRI INT'L COMPUTER SCIENCE LAB. (1998).

[102] Latanya Sweeney, *k-anonymity: A Model for Protecting Privacy*, 10(5) INT. J. OF UNCERTAINTY, FUZZINESS AND KNOWLEDGE-BASED SYS. 557, 557 (2002).

[103] Bhuvan Bamba, Ling Liu, Peter Pesti & Ting Wang, *Supporting Anonymous Location Queries in Mobile Environments with Privacygrid*, PROC. INT'L WWW CONF. 237, 237 (2008).

such size it is guaranteed that the identity of the person to be anonymized cannot be disclosed with a probability larger than $1/k$.[104]

There are many different approaches for selecting the $k-1$ persons for populating the anonymity spatial region. Those approaches can be categorized into location *k*-anonymity, historical *k*-anonymity, and trajectory *k*-anonymity.[105] Location *k*-anonymity protects a person's privacy in a network by building the anonymity spatial region from the current location of all people in the network.[106] This approach is different from historical *k*-anonymity, which uses the location history as a basis for anonymization.[107] Historical *k*-anonymity can be analogized with using people's footprints instead of their current location.[108] Finally, trajectory *k*-anonymity makes use of the location paths of individuals and is therefore particularly useful for preserving privacy in location-based services that cannot be offered in a single communication, such as car navigation.[109]

Given that the degree of anonymity depends on the choice of *k*, which value should *k* have? In order to provide some flexibility, many *k*-anonymity approaches do not provide a fixed value, but

[104] Aris Gkoulalas-Divanis, Panos Kalnis & Vassilios S. Verykios, *Providing k-anonymity in Location Based Services*, 12 SIGKDD EXPLOR. NEWSL. 3, 5 (2010).

[105] *Id.*

[106] *Id. See, e.g.*, Marco Gruteser & Dirk Grunwald, *Anonymous Usage of Location-based Services Through Spatial and Temporal Cloaking*, 1 PROC. INT'L CONF. ON MOBILE SYS., APPLICATIONS, AND SERVICES, 31 (2003).

[107] Aris Gkoulalas-Divanis, Panos Kalnis & Vassilios S. Verykios, *Providing k-anonymity in Location Based Services*, 12 SIGKDD EXPLOR. NEWSL., 3, 7 (2010). *See, e.g.*, Claudio Bettini, X. Sean Wang & Sushil Jajodia, *Protecting Privacy Against Location-based Personal Identification*, 2 PROC. VLDB WORKSHOP ON SECURE DATA MGMT., 185 (2005).

[108] Aris Gkoulalas-Divanis, Panos Kalnis & Vassilios S. Verykios, *Providing k-anonymity in Location Based Services*, 12 SIGKDD EXPLOR. NEWSL. 3, 7 (2010).

[109] *Id.* at 8. *See, e.g.*, Chi-Yin Chow & Mohamed F. Mokbel, *Enabling Private Continuous Queries for Revealed User Locations*, 10 PROC. INT'L SYMPOSIUM ON ADVANCES IN SPATIAL AND TEMPORAL DATABASES 258 (2007).

rather allow for an adaptive solution, which is useful because not everybody has the same privacy expectations.[110] Furthermore, a person may have different privacy expectations at different locations. Therefore, the same value of *k* for every person or for one person at every place is not a good fit.[111] However, an individual's ability to choose the value of *k* requires sufficient knowledge about the number of people in a particular area at a given time.[112] Otherwise, a system may be unable to accumulate *k* persons at the time of requesting a service, which could render time-critical services inoperable.[113] For example, a GPS car navigation system that uses *k*-anonymity to protect the driver's privacy will not work in remote areas when there are not enough other cars.

### B. *L*-DIVERSITY

Another privacy metric employed in the context of location privacy is *l-diversity*. Similar to *k*-anonymity, *l*-diversity was originally proposed to protect the identity of individuals in databases.[114] It is founded on the observation that while *k*-anonymity prevents the disclosure of identities, it does not prevent the disclosure of sensitive attributes, such as height, eye color, ethnicity, or other quasi-identifiers of a person.[115] Against this background, *l*-diversity requires that there are at least *l* values for each sensitive attribute.

[110] *See*, *e.g.*, Chin-Yin Chow, Mohamed F. Mokbel & Xuan Liu, *A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-based Services*, ACM INT'L SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFO. SYST. 171, 172 (2006).

[111] *See* Sheikh Iqbal Ahamed, Md., Munirul Haque & Chowdhury Sharif Hasan, *A Novel Location Privacy Framework without Trusted Third Party Based on Location Anonymity Prediction*, 12 ACM SIGAPP APPLIED COMPUTING REVIEW 24, 25 (2012).

[112] *Id.*

[113] *Id.*

[114] *See generally* Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke & Muthuramakrishnan Venkitasubramaniam, *l-diversity: Privacy Beyond k-anonymity*, 1 ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA 1 (2007).

[115] *See id.* at 2.

More specifically, *l*-diversity means that "[a] *q**-block [that is, a block from a database table that contains a generalized quasi-identifier *q**] is *l*-diverse if it contains at least *l* well-represented values for the sensitive attribute *S*. A table is *l*-diverse if every *q**-block is *l*-diverse."[116]

*l*-diversity can be utilized as a standalone privacy metric. However, it can also be seen as a companion measure to be used in tandem with *k*-anonymity.[117] In the context of location privacy, *l*-diversity allows individuals, for example, to control their state of being unidentifiable from a set of *l* different physical locations, such as churches, clinics, or offices.[118] However, attributes do not necessarily need to be a type of location. They can also be the driving speed, religion, or ethnicity of a person. Comparable to the formation of an anonymity spatial region by selecting $k - 1$ individuals, *l*-diversity achieves privacy protection by extending an anonymity spatial region until $l - 1$ different values of a sensitive attribute are included.[119] For example, if religion is the sensitive attribute, the region is extended until it includes persons with $l - 1$ different religions. That way it could be hidden that a person attends a particular church service.

[116] *Id.* at 16.

[117] *See id.* at 5.

[118] Bhuvan Bamba, Ling Liu, Peter Pesti & Ting Wang, *Supporting Anonymous Location Queries in Mobile Environments with Privacygrid*, PROC. INT'L WWW CONFERENCE 237, 239 (2008).

[119] *See* Byoungyoung Lee, Jinoh Oh, Hwanjo Yu & Jong Kim, *Protecting Location Privacy Using Location Semantics*, 14 PROC. ACM SIGKDD INT.L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING , 1289, 1289 (2011).

## IV. A MACHINE LEARNING APPROACH TO THE MOSAIC THEORY

Having set up our tools—machine learning techniques and privacy metrics—we are now ready to consider how pervasive location tracking impacts the Fourth Amendment in light of the mosaic theory. At its essence, the mosaic theory claims that in surveillance, the whole is greater than the sum of its parts. This means both that law enforcement can learn more than a simple tally of the collected data *and* that, at a certain point, law enforcement can learn disproportionately more relative to the effort they have expended. With regard to this latter point, the practical concern is that the relative ease of data accumulation removes the economic check on abusive police activity that might otherwise exist. These insights of the mosaic theory raise troubling Fourth Amendment concerns. Machine learning demonstrates the truth of these propositions.

Let us begin with the observation that accumulation of too much location information is itself troubling, for it can reveal intimate facts about the target of the surveillance. As Justice Sotomayor expressed in her concurring opinion in *Jones*:

> Disclosed in [GPS] data . . . will be trips the indisputably private nature of which takes little imagination to conjure: trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar and on and on.[120]

By making high-accuracy predictions based on limited data, this problem is exacerbated. Depending upon the predictions being made, the collection of data can become more intrusive substantively. Furthermore, law

[120] United States v. Jones, 132 S. Ct. 945, 955 (Sotomayor, J., concurring) (quoting People v. Weaver, 12 N.Y.3d 433, 441-42 (2009))).

enforcement is able to know more with considerably less effort.[121] As Justice Alito stated in *Jones*, the economic aspect of automatic accumulation of data becomes increasingly troubling:

> In the pre-computer age, the greatest protections of privacy were neither constitutional nor statutory, but practical. Traditional surveillance for any extended period of time was difficult and costly and therefore rarely undertaken. The surveillance at issue in this case—constant monitoring of the location of a vehicle for four weeks—would have required a large team of agents, multiple vehicles, and perhaps aerial assistance. Only an investigation of unusual importance could have justified such an expenditure of law enforcement resources. Devices like the one used in the present case, however, make long-term monitoring relatively easy and cheap.[122]

The fact that location tracking is cheap (and even made cheaper by prediction) is seen as eroding a vital bulwark: "[B]ecause 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."[123] Again, machine learning techniques lower the cost still more, and produce more data. Furthermore, the mechanisms are even more hidden from public scrutiny.

The central question then is this: can the tracking, aggregation, and processing of data by machine learning algorithms constitute a search in violation of the Fourth Amendment? For the answer to this question to be

[121] There is, of course, a considerably higher error rate in data generated by machine learning algorithms, as opposed to items directly observed. This raises the fascinating question of whether it requires more or fewer questions for law enforcement to believe something that is not correct.

[122] *Jones*, 132 S. Ct. at 963 (Alito, J., concurring).

[123] *Id.* at 956 (Sotomayor, J., concurring) (internal quotations omitted).

"yes," two things must be established. First, it must be true that more can be learned from the location tracking data than the sum of the information individually gathered (Subsections A. and B.). Second, it must be demonstrated that the information learned is protected under the privacy test set forth in *Katz* by Justice Harlan[124] (Subsections C. and D.). We must establish more than that, though. We must also show that the mosaic theory is an operationally useful approach to the Fourth Amendment (Subsection E.).

### A. THE EXISTENCE OF PREDICTABLE LOCATION PATTERNS

Is it possible to learn more from location tracking data than the discrete units of data? The answer is a resounding "yes!" There are predictable patterns to people's movements that can be derived from their past locations. A 2010 paper by Chaoming Song and other researchers demonstrates this proposition. Using a set of cell phone tower data points, they showed that human movement was 93% predictable.[125] Song and his co-authors note that the high degree of "regularity is . . . potentially . . . intrinsic to human activities."[126] Moreover, "it is not the 93% predictability that [is] most surprising. Rather, it is the lack of variability in predictability across the population."[127] While the Song paper did not attempt to make actual predictions based upon the datasets it was using, the authors did conclude that the high degree of regularity that is found in hu-

[124] Katz v. United States, 389 U.S. 347, 361 (Harlan, J., concurring).

[125] *See* Chaoming Song, Zehui Qu, Nicholas Blumm & Albert-László Barabási, *Limits of Predictability in Human Mobility*, 327 SCIENCE 1018 (2010), http://www.sciencemag.org/content/suppl/2010/02/18/327.5968.1018.DC1/Song.SOM.pdf. Mobile phone records provide location information only when a person uses his or her phone. *Id.* at 1019. The result is therefore based on the analysis of data from 45,000 users whose location was recorded for more than 20% of hourly intervals and whose location recordings were reliably extrapolated to 100% of hourly intervals. *Id.* at 1019-20.

[126] *Id.* at 1021.

[127] *Id.*

man movement makes it likely that efforts at prediction would succeed.[128]

In a more recent paper Yves-Alexandre de Montjoye and three co-authors present stronger results. They found that as few as four data points derived from coarse cell phone tower data could uniquely identify 95% of individuals.[129] Their conclusions are unambiguous:

> All together, the ubiquity of mobility datasets, the uniqueness of human traces, and the information that can be inferred from them highlight the importance of understanding the privacy bounds of human mobility. We show that the uniqueness of human mobility traces is high and that mobility datasets are likely to be re-identifiable using information only on a few outside locations . . . . This implies that even coarse datasets provide little anonymity.[130]

In order to find how many mobility data points are needed to uniquely identify an individual from a mobility trace, de Montjoye and his co-authors define $I_p$ as the set composed of $p$ mobility data points and $S(I_p)$ as the set of all traces that match the $p$ points.[131] Thus, for example, in the case of tracking three individuals in New York City from Union Square to Washington Square, given that these are the only two mobility data points, that is, $p = 2$, there are

[128] *Id.*

[129] *See* Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, *Unique in the Crowd: The privacy bounds of human mobility*, SCIENTIFIC REPORTS 1376 (March) http://www.nature.com/srep/2013/130325/srep01376/pdf/srep01376.pdf.

[130] de Montjoye et al., *supra* note 129, at 2.

[131] *Id.*

three traces that match these points, $|S(I_{p=2})| = 3$.[132] From this information it is not possible to identify any of the three individuals. However, if one more mobility data point is obtained, that is, $p = 3$, and if it turns out that one individual moves further to the East Village, one to the West Village, and the third to Little Italy, three unique traces will emerge, $|S(I_{p=3})| = 1$. Therefore, this reduction in the cardinality of $S(I_p)$ from 3 to 1 leads to unique identification of all individuals. In this regard, de Montjoye and his co-authors note:

> [T]he information added by a point is highly dependent from the points already known. The amount of information gained by knowing one more point can be defined as the reduction of the cardinality of $S(I_p)$ associated with this extra point. The larger the decrease, the more useful the piece of information is. Intuitively, a point on the MIT campus at 3AM is more likely to make a trace unique than a point in downtown Boston on a Friday evening.[133]

In other words, adding a data point—another observation of someone's location at a given time—can at times dramatically cut the size of $S(I_p)$, i.e., reduce the number of people whose behavior can be matched. Having more data points allows for a better identification. This fits well with the concept of *k*-anonymity: generally speaking, only a few points are necessary to reduce $k$ to 1.[134]

These results, as striking as they are, were obtained with random data. However, as de Montjoye and his co-authors explain, not all data points are equally meaningful. In particular, they note that their random sampling tended to pick out "home" and "office"

[132] The notation "$|X| = 3$" means "set $X$ has cardinality 3", i.e., there are 3 elements in that set.

[133] de Montjoye et al., *supra* note 129, at 3.

[134] *See supra* Section III.A (for an explanation of *k*-anonymity).

points, simply because people are there for longer time than they are on the road.[135] They envision, however, a far more discriminate collection of data points:

> For the purpose of re-identification, more sophisticated approaches could collect points that are more likely to reduce the uncertainty, exploit irregularities in an individual's behavior, or implicitly take into account information such as home and workplace or travels abroad. Such approaches are likely to reduce the number of locations required to identify an individual, vis-à-vis the average uniqueness of traces.[136]

It is important to understand what these two papers do and do not say. Neither gives results that are likely to be of direct benefit to law enforcement. After all, if a comprehensive set of cell phone tower location records is available, there is no need to predict someone's next location; law enforcement can simply demand access to the database. However, the papers do indeed support the notion that there are patterns to people's locations, patterns that are often unique, and which can, in principle, be used to learn more, and more easily, than is present in the records themselves.

### B. Determining the Formation of a Mosaic

To demonstrate the correctness of the mosaic theory, we need to show that location information can answer prosecutors' questions without the aspect in question being directly observable. This is the strongest theoretical contribution of machine learning to the mosaic theory. Experimental results do indeed validate our hypothesis that

[135] de Montjoye et al., *supra* note 129, at 3.
[136] *Id.* (citation omitted).

a point can be objectively identified at which the collection of data becomes greater than the sum of its parts, in that it reveals information not previously known. Consider, for example, a study performed by Yaniv Altshuler and others.[137] It can be observed that some (though not all) of their graphs show a sharp uptick in accuracy when monitoring has been done for a certain amount of time. Figure 1 is the most dramatic: after about 5 weeks of monitoring, and again after about 27 weeks, accuracy in identifying a subject's ethnicity jumps quite sharply.

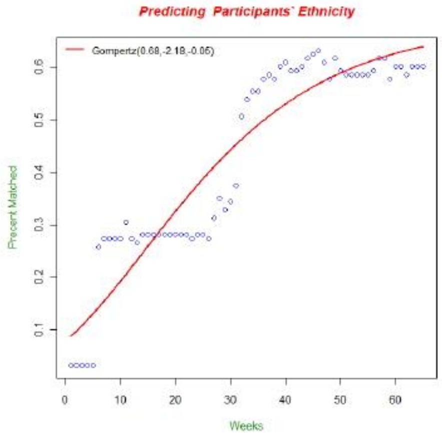

*Figure 1: This is Figure 10 from Altshuler et al.*

Figure 2 is almost as striking; after the initial increase and a plateau, the accuracy in determining whether or not someone is American-born climbs substantially again around the 20 day mark. Such sharp changes in a graph provide an objective basis for defining the existence of a mosaic. Not only is the dataset producing

[137] *See* Yaniv Altshuler et al. "Incremental learning with accuracy prediction of social and individual properties from mobile-phone data," *WS3P, IEEE Social Computing,* 2012.

more accurate predictions at these points of sharp change (i.e., previously unknown information is more likely to be revealed), but too, the output knowledge at these points is growing much faster than the input effort (i.e., law enforcement is learning more with considerably less effort). Effort is, of course, economic cost, per Justice Alito's concurrence in *Jones*.[138] These sharp upward bends in the curves are, therefore, crucial. To the extent that resistance to the mosaic theory is driven by concerns about incomprehensible line drawing, the upticks described above reflect that an objective basis for such line drawing does, in fact, exist.

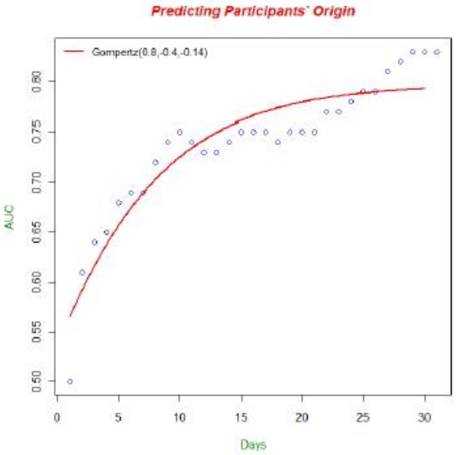

*Figure 2: This is Figure 7 from Altshuler et al.*

It is difficult to provide a formal mathematical definition of such an uptick. However, we can define it descriptively. Suppose we have a graph, similar to those in the Altshuler study, which re-

[138] *See supra* note 122.

lates the amount of monitoring (in the figures on the *x*-axes) to the accuracy of a prediction (in the figures on the *y*-axes). Using straightforward, well-known techniques, one can fit a curve to those points. At any point on this curve, one can visualize its *slope* (i.e., how fast it is rising or falling).[139] However, the slope at a certain point does not tell us what we need for the mosaic theory to hold because it tells us nothing about how the collection of data at one particular point compares with the collection at other points. Rather, all that a steep slope tells us is that a small amount of observation yields a large increase in accuracy.

The *change* in the slope, however, is significant because it provides an objective measure for comparing the slope at different points in time. If the slope is increasing as more data points are considered, and especially if it is increasing rapidly, the change in slope tells us that we have a better chance of learning more proportionally from later than from earlier observations.[140] Once this transformation in the accuracy of factual predictions occurs, a mosaic has been formed. This is true because no longer is the government merely gathering information more efficiently. Rather, at these points on the curve, the government is more precisely generating previously unknown information. It is easily possible to visualize such a curve. Where a sharp bend upwards can be observed, a mosaic has been created.

[139] The notion of the slope of a curve at a given point is well defined mathematically; in calculus, it is known as the *first derivative* of the equation of the curve.

[140] The rate of change of the slope—the first derivative—is known, not surprisingly, as the *second derivative.*

We have illustrated this in Figure 3 using a made-up, but realistically shaped graph.

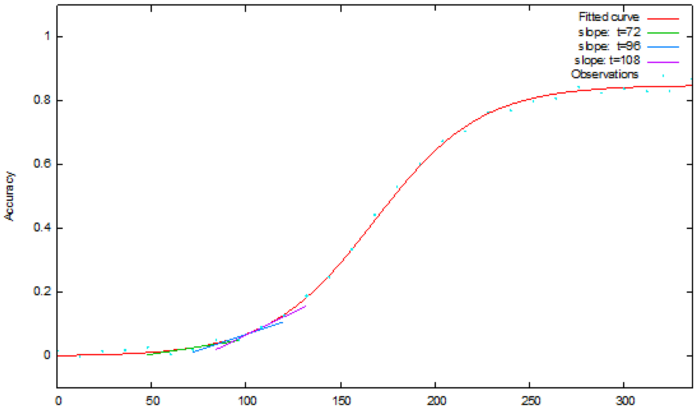

*Figure 3: This is a synthetic (i.e., utterly made up, and not corresponding to any real experiment) graph showing the accuracy of predictions after some number of hours. Note the lines showing the slope at several points. The curve is assumed to have been fitted.*

Figure 4 is a close-up of the crucial section of the graph.

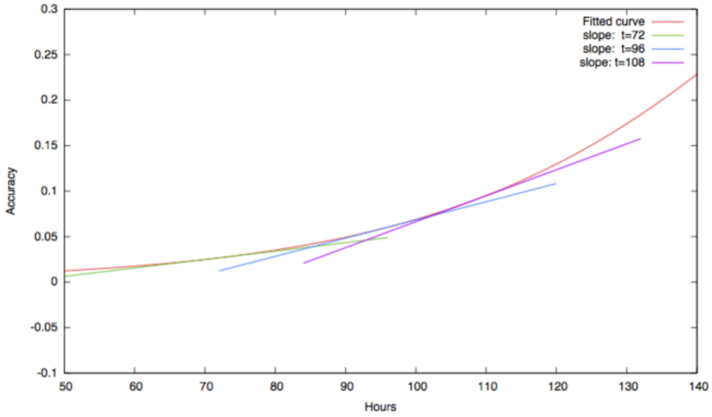

*Figure 4: A close-up of the previous graph showing the slope at several points.*

Each data point gives the accuracy of predictions after that number of hours of observation of the target, based on a (presumed) training dataset. There are lines showing the slope at several different points. Table 1 shows the second derivative of the curve, i.e., the rate of change of the slope. Note the relatively small change after just a few hours of observation, compared with the very large change from 96 to 108 hours. Also, as can be noted, the change becomes smaller after 108 hours. This increase, and later decrease, is crucial. It indicates that a mosaic has been formed, probably around the 108 hours point. If another similar increase and decrease were to be observed at a later point, it could be disregarded as the mosaic was already established at an earlier point, a lower bound.

| Hours | Second derivative x10,000 |
|---|---|
| 0 | 0.02661 |
| 12 | 0.04063 |
| 24 | 0.06187 |
| 36 | 0.09379 |
| 48 | 0.14126 |
| 60 | 0.21060 |
| 72 | 0.30916 |
| 84 | 0.44334 |
| 96 | 0.61357 |
| 108 | 0.80490 |
| 120 | 0.97413 |

*Table 1: The second derivative of the slope at certain points.*

To make the determination if, and at what point, a mosaic has been formed, that is, when enough is enough, the analyst would have to take an appropriate set of data, train the models, see what correlations form, and draw the accuracy curves just discussed. Where the mosaic forms is dependent on the training dataset used, the predictive algorithm employed, and the precise question being asked. Each of these three criteria raises questions.

Of the three criteria, the possession of large amounts of data by law enforcement is the most studied, though not in the context of training a machine learning algorithm.[141] In general, there are many types of data that the government cannot legally collect, or can collect only subject to stringent limitations. These same datasets, however, may be readily available to the private sector. In such situations, government agencies, including law enforcement agencies, have simply purchased data from large-scale data brokers.[142] Thus, for now, we assume that suitable datasets exist and can be obtained, perhaps in anonymized form, and perhaps accessible to law enforcement only to answer particular questions, rather than for general use.[143]

Given their availability, the choice of the training data raises troubling questions. In general, the better the training data match a target, the more accurate the predictions will be. Consider, for example, the location patterns of a stay-at-home mother and a deliv-

[141] *See, e.g.*, Fabio Arcila, *GPS Tracking Out of Fourth Amendment Dead Ends:* United States v. Jones *and the* Katz *Conundrum*, 91 N.C. L. REV. 1 (2012); Stephanie Pell et al., *supra* note 12; *see also Recommendations for Fusion Centers*, THE CONST. PROJECT, http://constitutionproject.org/pdf/fusioncenterreport.pdf.

[142] One case in point is non-content information about subscribers to electronic communications and remote computing services. Carriers are explicitly prohibited from providing this type of information to the government unless a suitable court order is presented or other exceptions are applicable; *see* 18 U.S.C. §§ 2702(c)(6), 2703(c).

[143] Data anonymization is remarkably hard. *See, e.g.*, Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L.REV. 1701 (2010). The computer science literature also gives many examples showing that simply being able to ask questions about the behavior of aggregates in an otherwise-inaccessible database or using outside information on an anonymized dataset can still leak information. *See ,e.g.*, Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, PROC. OF THE 2008 IEEE SYMPOSIUM ON SEC. AND PRIVACY (SP) 111 (2008).

ery-truck driver. They are clearly quite different. Using patterns that are similar to the target's behavior will result in better predictions. While that itself raises issues, such as the compilation of training datasets along ethnic or racial lines, those concerns are beyond the scope of this paper.

Obtaining and selecting training data is not the only point to consider. Formation of a mosaic also depends on the selected algorithm. Algorithms are not static. As in many fields of computer science, there has been rapid progress in recent years. An algorithm that represents a breathtaking advance one year may be commonplace the next and obsolescent the year after that. This in turn means that determinations of when a mosaic has formed, and, hence, when a warrant should be procured, are also not static. Rather, the question must be reexamined at reasonable intervals, certainly no less frequently than every few years. That said, police are increasingly relying on sophisticated predictive software.[144] In Santa Cruz, California, for example, an experimental trial used such software to affect police deployment patterns:

> [. . . ] Santa Cruz's method is more sophisticated than most. Based on models for predicting aftershocks from earthquakes, it generates projections about which areas and windows of time are at highest risk for future crimes by analyzing and detecting patterns in years of past crime data.

[144] Obviously, skilled investigators are also adept at making deductions from patterns of data simply based on their experience. For example, one former police officer made the following comment to us: "It is no secret that Friday and Saturday nights are big with the drug trade. Sometimes money changes hands on Mondays early. That pays for last week's product and [serves as a] down payment for next week's product. A Monday mid-day visit is a tell. If that follows with a Thursday visit and is consistent, we know we have a pick-up, drop-off location. If the same people go to different places, but it follows a pattern, we know when shipments are being made."

> The projections are recalibrated daily, as new crimes occur and updated data is fed into the program.[145]

In one case use of the program turned out to be crucial for arresting two female suspects; one with an outstanding warrant and the other one carrying illegal drugs. "On the day the women were arrested, [. . . ], the program identified the approximately one-square-block area where the parking garage [in which the women were arrested] is situated as one of the highest-risk locations for car burglaries."[146] This success of technologically enhanced police work illustrates that we can expect increasingly more algorithmic use of location data by the police.

In addition to training data and algorithms, the set of questions that can be asked is also relevant for the formation of a mosaic. As we saw from the Altshuler study, in order to achieve a certain predictive accuracy, different questions demand different amounts of data.[147] The set of questions that might be asked, however, is quite large, and generally both fact-specific and dependent on the stage of the investigation:

> If law enforcement had a known target, but was otherwise unaware or only had minimal information about the target's "criminal associates," law enforcement would want to identify those potential criminal associates (which may occur initially through analysis of location data, other methods, or a combination of methods which can include location data) and then track those potential associates to see where they go and who they meet with. It may be that, in this circumstance, law enforcement has very little infor-

[145] *See, e.g.*, Erica Goode, *Sending the Police Before There's a Crime*, N.Y. TIMES, August 16, 2011.

[146] *Id*.

[147] *See* Altshuler et al., *supra* note 10.

> mation about the newly identified "associates," but the monitoring of their movements can reveal information about the *modus operandi* of the organization (to include roles and "criminal knowledge" of various individuals in the organization)—additional important insights beyond just "where did they go."[148]

Although it is probably feasible to come up with a canonical, albeit somewhat large, set of fairly standard questions, it is less clear that this would be entirely satisfactory. For one thing, the set of facts in a given case can be wildly different than anything encountered before.[149] Perhaps even more seriously, the set of questions an investigator *should* ask may differ from what is *actually* asked or intuited. In this sense, machine learning and mosaic theory may raise Equal Protection or Due Process questions, which we flag but not explore further.

### C. APPLYING PRIVACY METRICS

So far, we have discussed the existence of mosaics and how their formation can be detected. Now, we must show how mosaics can be integrated into the reasonable expectation of privacy test. To that end, we use notions of the different privacy metrics that are proposed in the computer science literature and that we discussed earlier.[150] We focus on two metrics that are most promising for accomplishing our task: *k*-anonymity and *l*-diversity. However, before we discuss how *k*-anonymity and *l*-diversity motivate our ap-

[148] Private communication with Stephanie Pell. Pell is a former federal prosecutor who worked on national security cases.

[149] *See, e.g.*, the scenario described in Stephanie Pell & Christopher Soghoian. *supra* note 12, at n.150, which is based on a real investigation.

[150] *See supra* Section III.

proach, we briefly describe some of the major obstacles preventing their direct application.

In order to see why *k*-anonymity and *l*-diversity cannot be applied directly, it first should be noted that both metrics in their original form are inherently limited. This limitation is a consequence of their development for the purpose of preventing identification of individuals in databases. They were not meant to provide a general and comprehensive privacy metric. Rather, they are tools for creating degrees of anonymity within databases that prevent particular entries from being conclusively linked to known identities. It would be peculiar to describe the protections of the Fourth Amendment as primarily concerned with encouraging this sort of randomization.[151]

Further, in the location tracking context, *k*-anonymity and *l*-diversity are mainly used for providing anonymity to users of location-based web services; an application that is very different from the location tracking of suspects in police investigations. Both privacy metrics generally assume a three-party scenario in which a trusted third party, for example, a cell phone network provider, knows the exact location of a user and then forwards only an imprecise anonymity spatial region to the requesting location-based service. Thus, the exact location of the user is only hidden from the location-based service, but not from the network provider. However, our scenario will often not involve a network provider or other trusted third party. Rather, the location information is transmitted directly from the tracking device to the police. Even if a third party is involved, the police may be able to obtain the location information from that party.

[151] *But see* Bernard E. Harcourt & Tracey L. Meares, *Randomization and the Fourth Amendment*, 78 U. CHI. L. REV. 809 (2010), http://www.law.uchicago.edu/files/file/530-317-bh-fourth-amendment_0.pdf.

A final—and central—definitional limitation concerns what is to be protected. Privacy violations can only occur for protected information. The challenge, therefore, is defining the class of information worthy of protection. While *k*-anonymity prohibits the disclosure of identifiers and *l*-diversity extends this prohibition to quasi-identifiers, the class of protected information for our purposes is characterized by the reasonable expectation of privacy, which has some overlap with *k*-anonymity and *l*-diversity but is not completely congruent with those metrics. Given this and the other previously described limitations, the usefulness of *k*-anonymity and *l*-diversity may seem doubtful. However, the situation is not entirely bleak. Despite their constraints, it is possible to leverage their general ideas.

Considering *k*-anonymity first, one attribute of a person is protected: identity. By definition, *k*-anonymity is concerned with size $k$ of the group that satisfies certain criteria; when $k = 1$, the subject is perfectly identified. We see this described by de Montjoye and coauthors: very few people's location traces correspond to the same set of four observations.[152] That is, with four observations, $k = 1$ with high probability. There are certainly scenarios where this might be of interest. For example, suppose that a crime takes place in a certain locale. Given training data from a certain population and a set of *after the fact* location data for one particular suspect, is this person "*predicted*" to have been in that locale at the time the crime was committed?[153] This prediction is possible because correlations are not restricted to predicting future acts. We can use those correlations to ask "whose earlier location is predicted to be most

[152] *See supra* Section IV.0

[153] The tenses admittedly are odd in that sentence. Nevertheless, they are correct when applying location data to a prior act.

consistent with the known *later* locations of a group of suspects?" In other words, we are running the algorithm backwards in time.

In contrast to $k$-anonymity, $l$-diversity deals with a larger set of protected attributes: quasi-identifiers. In general, any attribute can be specified as a quasi-identifier and for each there must be at least $l$ possible values. However, it is an open research question how $l$-diversity—or $k$-anonymity, for that matter[154]—can be reconciled with and mapped to the output of machine learning algorithms. Such mapping is necessary because the algorithms yield an accuracy rate in terms of probability, rather than supplying a set of $l$-diverse "well represented" answers.[155] Therefore, in order to overcome this disconnect, we either need experiments that give answers in terms of $l$-diversity or a different privacy metric in terms of guess accuracy. We propose the former and provide a simple rule for converting probabilities into an $l$-diverse answer: Given that a machine learning algorithm returns a probability, $p$, for the existence of an attribute, it holds that $l = \lfloor 1/p \rfloor$.[156]

Let us illustrate our rule by an example. If investigators believe that a suspected drug dealer driving in his car picked up a bag containing drugs in San Francisco, the machine learning algorithm may return a 40% probability for a pick-up stop in San Francisco.[157] This result can be translated into 2-diversity. Now, why is that the case? In general, the probabilities for selecting the correct answer from two possibilities at random would be 50%, from three possibilities

[154] It is possible to view $k$-anonymity as a special case of $l$-diversity. If identity is the only attribute of interest, saying that there must be $l$ possible values of that attribute is equivalent to saying that $k = l$.

[155] An answer is "well represented" in terms of $l$-diversity if an attribute is hidden among a total of $l$ attributes. For a discussion of $l$-diversity *see supra* Section III.0

[156] The floor notation $\lfloor x \rfloor = y$ means that $y$ equals the largest integer not greater than $x$. Thus, for example, $\lfloor 5.3 \rfloor = 5$.

[157] It is possible to predict a drop-off or pick-up trip with relatively high accuracy. *See infra* Section IV.0

33.1/3%, from four 25%, and so on. Thus, if the probability returned from the machine learning algorithm is greater than 50%, there is a higher chance of being correct when selecting this answer compared to any other answer. This can be interpreted as 1-diversity. However, if the probability returned is not greater than 50%, but greater than 33.1/3%, we have 2-diversity. If it is not greater than 33.1/3%, but greater than 25%, 3-diversity, and so on. Because in our example the probability that the suspect picked up something in San Francisco is 40%, it holds that $l = \lfloor 1/0.4 \rfloor = \lfloor 2.5 \rfloor = 2$, that is, our mapping creates 2-diversity.

The demonstrated conversion rule leads to another observation. The rule in fact provides a rationale based on *k*-anonymity and *l*-diversity for quantifying a reasonable expectation of privacy violation at a 50% probability threshold. Whatever question the investigators ask, it must be checked if the probability of the answer is greater than 50%. If that is the case, the corresponding answer is more likely to be correct than all others. Consequently, the prediction of an attribute (in case of *l*-diversity) or the identification of the suspect (in case of *k*-anonymity) is more likely to be successful than not and we have 1-diversity and 1-anonymity, respectively. Given such result and given that the type of information asked for is protected as well, a point we will address in the next subsection,[158] the 50% probability threshold is crossed and a privacy violation exists.

If either *k*-anonymity or *l*-diversity are used in the manner described, they import a probabilistic understanding of privacy into the reasonable expectation of privacy analysis. In this regard, as noted earlier,[159] the attempts at a purely quantitative definition of, for example, "probable cause" have failed to garner support from a majority of the Court. Of course, one reason they have not been

[158] *See infra* Section IV.0

[159] *See supra* Section II.0

adopted is because judging is not quantitative. We do not, for example, have juries saying, "the probability that this person is guilty is 83%" and then comparing that against the "reasonable doubt" threshold. However, the Court's reluctance to quantify legal concepts like probable cause does not stand as an impediment to the proposal here—quantification of objective reasonableness. Our application of *k*-anonymity or *l*-diversity provides an objective rationale for the probabilistic quantification of reasonableness and, after all, the Court has indicated a willingness to adopt quantitative understandings of legal concepts on far more tenuous grounds than the instant proposal.[160]

Moreover, while the Court has repeatedly instructed that the "probable cause standard is incapable of precise definition or quantification into percentages because it deals with probabilities and depends on the totality of the circumstances,"[161] it has also simultaneously suggested that the protections of the Fourth Amendment may rise or fall based upon the *quantity* and quality of information sought by law enforcement.[162] Consequently, there has to date been no suggestion that science might not provide an objective basis for quantifying privacy. Quite the opposite should hold true, even more so as the mosaic theory can be seamlessly integrated into the traditional *Katz* test for determining violations of reasonable expectations of privacy.

### D. Determining a Privacy Violation

In order to establish a case for the mosaic theory, the final necessary step is to show that machine learning techniques can indeed violate the reasonable expectation of privacy. We assert that it will

[160] *Id.*

[161] Maryland v. Pringle, 540 U.S. 366, 371 (2003).

[162] Dow Chem. Co. v. United States, 476 U.S. 227 (1986).

be a Fourth Amendment violation if machine learning techniques are used to deduce facts that are not otherwise ascertainable without violating clearly established principles, most fundamentally the privacy protections originating from the privacy of the home.[163] Without question, this is just a starting point; as the science develops, so too will our objective understanding of the applicable legal rules.

Suppose that it were possible to learn—with high probability and solely by looking at location data—that a couple was estranged and were sleeping in separate rooms. This is undeniably private information, perhaps even more so than "at what hour each night the lady of the house takes her daily sauna and bath."[164] This sounds like an improbable thing to learn; nevertheless, one reason machine learning is so valuable is that it can discover such correlations, even if no one can explain the causality.[165]

*Kyllo* makes this observation very clear: "We think that obtaining by sense-enhancing technology any information regarding the interior of the home that could not otherwise have been obtained without physical intrusion into a constitutionally protected area constitutes a search."[166] In the language of *Kyllo*, machine learning *is* a "sense-enhancing technology." It allows the detection of information that otherwise would be hidden from human observation.

[163] *See supra* Section I.0

[164] Kyllo v. United States, 533 U.S. 27, 38 (2001).

[165] It is important to remember that machine learning works by finding correlations, rather than by identifying causal relationships. We can imagine a scenario, e.g., that a man who regularly spends Saturday nights at a strip club does so because he's estranged from his wife, but machine learning does not make that leap; it simply finds the pattern. The prediction can be wrong, perhaps because he is an employee rather than a guest, or because he is a plumber who is regularly called out to repair balky pipes, or because this is how a happy couple has chosen to spend their Saturday nights together. That does not invalidate the correlation, which simply says that *most* men with such a location pattern are unhappy in their marriages.

[166] *Kyllo*, 533 U.S. at 34 (internal quotation marks omitted).

It should be noted, however, that machine learning *per se* is not the issue. Sufficient datasets are also required. Given their availability, the aggregation of publicly observable movements can be transformed from a chronicle of where the target has been into something different and new, something much more meaningful and invasive. With such application of machine learning algorithms to location tracking data, a substantive change in the police investigation occurs, not simply a change in the investigation's form.

While it is true that most of the location tracking data is likely obtained from the tracked individual's movements in public, the information deduced from the analysis of the aggregated public data does not need to be. Rather, it can be of a very intimate nature. The deduced information can be of a type and nature that is protected under the evolved interpretation of what constitutes the privacy of the home and its reduced dependency on physical boundaries.[167] If that is the case, it must be protected. This way, even nominally public behavior can be protected.

In the end, which information is awarded Fourth Amendment protection depends on societal expectations. The "reasonable expectation of privacy" of today's Fourth Amendment doctrine accommodates this notion and is explicitly couched in terms of societal expectations, *i.e.*, what people as a whole believe is "reasonable." Consider again Justice Harlan's concurrence in *Katz*: "there is a twofold requirement, first that a person have exhibited an actual (subjective) expectation of privacy and, second, that the expectation be one *that society is prepared to recognize as 'reasonable.'*"[168]

Societal expectations, though, are based on what is customary, and customary behavior by law enforcement is based in part on

[167] *See supra* Section I.0

[168] Katz v. United States, 389 U.S. 347, 361 (Harlan, J., concurring) (emphasis added).

economic factors and is limited by what people will put up with. Thus, visits to "the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the criminal defense attorney, the by-the-hour motel, the union meeting, the mosque, synagogue or church, the gay bar"[169] are protected information under the Fourth Amendment if contemporary societal expectations consider them private.

### E. A NOTIONAL EXPERIMENT

We now propose an experiment to determine, in advance, where the mosaic boundaries are, that is, at what point location tracking requires a warrant in order to not violate the tracked individual's reasonable expectation of privacy. This experiment has not been performed and it is not clear that it actually can be, in particular, given the uncertainty about the set of questions the police may want to ask, the different algorithms that could be employed, and the different types of location data that can be collected. However, assuming that it is possible to perform the experiment, ideally, determination of a mosaic in any given situation, or perhaps for a given time and place—say, New York City one week from now—would be done ahead of time.

The experiment would begin by selecting training data similar to the type of data to be analyzed. Then, the general procedure would be to compile a standard set of questions, based on questions investigators intend to ask during an investigation and facts that are believed to be learnable. From this set of questions, those felt to be innocuous or permissible are discarded. The remainder—questions whose answers are intrusive enough to potentially violate a person's reasonable expectation of privacy (as described previously)[170]

[169] United States v. Jones, 132 S. Ct. 945, 955 (Sotomayor, J., concurring).
[170] *Supra* Section IV.0

or which are impermissible for law enforcement use (for example, as a matter of Due Process)—can be used to create the training dataset and to query a test dataset. From the resulting test dataset curves, one for each question (and perhaps for each question/algorithm pair), the analyst can see if and where a mosaic forms, and obtain a warrant, if necessary.

To our knowledge this procedure has not yet been carried out. Therefore, the absence of such research prevents us, at this time, from giving candidate values for certain standard sets of data, algorithms, and questions: a day, a week, a month. The kinds of questions a law enforcement officer might ask are not those that have typically been examined in the computer science literature. However, despite the lack of specific research in this area, a general trend is already emerging, that is, location patterns generally form according to the regular organization of human life.[171] This regularity may serve as a basis for approximating mosaic formation. In particular, human activities repeat a high degree of regularity from one week to another.[172]

In this regard, a human mobility study by Adam Sadilek and John Krumm shows that, while the location of someone in the distant future is in general highly independent of the recent location, "it is likely to be a good predictor of [the person's] location exactly one week from now."[173] This result is not surprising and intuitively the case in many realms of life as discussed, for example, in the often week-based regularity and organization of the drug trade.[174] Looking at even smaller time increments, Song and his co-authors

[171] Chaoming Song et al., *supra* note 125, at 1018.

[172] Tao Jia & Bin Jiang, *Exploring Human Activity Patterns Using Taxicab Static Points*, 1 ISPRS INT. J. GEO-INF. 89 (2012).

[173] Adam Sadilek & John Krumm, *Far Out: Predicting Long-Term Human Mobility*, PROC. OF THE TWENTY-SIXTH AAAI CONF. ON ARTIFICIAL INTELLIGENCE (2012).

[174] *Supra* Section IV.0

have observed a high degree of potential predictability from daily mobility patterns.[175] Based on these findings, and in absence of any more specific experimental results, the location tracking of someone for more than a week without a warrant appears to be an upper bound in the average case.[176] However, as noted,[177] algorithms change. Therefore, this upper bound may become smaller over time. Tracking the location of a person for even just a few days may be enough to reveal a lot of protected information.

For example, various studies aim to deduce the trip purpose from location data collected for less than a week. In an early study, Jean Wolf and co-authors equipped survey participants with GPS devices for three-day periods and found that it was possible to derive whether a person was going home, to work, or had a different trip purpose with an accuracy of 93.38%.[178] In a similar experiment Zhongwei Deng and Minhe Ji classified trip purposes into seven categories: going to work, going to school, going home, picking-up or dropping-off, shopping or recreation, business visit, and other activities.[179] They were able to achieve an overall accuracy of 87.6%.[180] Obviously, this type of information gives law enforcement

[175] Chaoming Song et al., *supra* note 125, at 1020.

[176] In his proposal for a statutory implementation of the mosaic theory, Christopher Slobogin suggests that searches lasting longer than 48 hours should require a warrant unless exigent circumstances exist. Christopher Slobogin, *Making the Most of* United States v. Jones *in a Surveillance Society: A Statutory Implementation of Mosaic Theory*, 8 DUKE J. CONST. L. & PUB. POL'Y 1, 24 (2012). Drawing the line at 48 hours is informed by the length of time the government may hold an arrestee before a judge must be consulted. *Id.* at 25. However, as Slobogin states, this line drawing is not related to the intrusiveness of the search and, in this sense, arbitrary. *Id.* at 26.

[177] *See supra* Section IV.0

[178] Jean Wolf et al., *Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from Global Positioning System Travel Data*, 1768 TRANSP. RES. REC. 125 (2001).

[179] Zhongwei Deng & Minhe Ji, *Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach,* TRAFFIC AND TRANSPORTATION STUDIES 768 (2010).

[180] *Id.*

a good gauge, for example, to determine where a suspected drug dealer went for picking-up or dropping-off drugs or money, however; it can also reveal an abundance of protected information about the targeted person.

## V. CONCLUSIONS

At least in principle, machine learning lets us answer one of the key challenges posed by the mosaic theory: how to tell if a mosaic exists. In his piece on the mosaic theory, Fourth Amendment scholar Orin Kerr notes[181] the three different expectation of privacy theories for prolonged location tracking in the opinions by Justice Alito ("a degree of intrusion that a reasonable person would not have anticipated"),[182] Justice Sotomayor ("a manner that enables the Government to ascertain, more or less at will, their political and religious beliefs, sexual habits, and so on"),[183] and Judge Ginsburg ("the likelihood a stranger would observe all . . . movements [of a person over the course of a month] is not just remote, it is essentially nil").[184] Machine learning provides clear objective support for the first two theories advanced: it can find surprising correlations, and it permits retrospective inquiries into many different facets of private behavior.[185]

In principle, machine learning also lets us draw lines beyond which a mosaic definitely exists; the process described in this article lets us measure the degree of intrusiveness (i.e., the loss of privacy) of any given set of location observations. Unfortunately, the necessary experiments have not been carried out and the current tech-

[181] *See* Kerr, *supra* note 2, at 330.

[182] United States v. Jones, 132 S. Ct. 945, 964 (2012) (Alito, J., concurring).

[183] *Id*. at 956 (Sotomayor, J., concurring).

[184] *See* United States v. Maynard, 61 F.3d. 544, 560 (D.C. Cir. 2010).

[185] Judge Ginsburg's theory can be already met with the location tracking as such and does not require any machine learning analysis of the recorded data.

nical privacy metrics cannot be integrated into the mosaic theory without modification. The latter point also has to do with the lack of a generally applicable privacy metric. It will be an important task for the future to come up with a metric that is mathematically sound, technically useful, and legally relevant.

The development of the legal doctrine for location tracking is in its infancy. While we provide a basic framework and general rules, there are many details that can have an impact on the legal analysis as the doctrine further develops. For example, it may be that different types of location tracking mandate different legal treatment. Particularly, the fine granularity of GPS tracking data may create a mosaic much faster than cell phone tower data would.[186] Additionally, it could play a role how close the location tracker is to the tracked person. For example, because a cell phone is usually carried around when people leave their homes,[187] its GPS can provide, in many cases, more accurate location tracking data than a GPS device attached to a car. The analysis may also be further complicated by the aggregation of different types of information, for example, when location tracking information is aggregated with other information contained in government databases. The legal and computer

[186] Some courts already applied such distinction. *See, e.g.*, United States v. Rigmaiden, 2013 LEXIS 65633, *35-36 ("The Court cannot conclude that . . . use of cell-site information, obtained from a third party under the [Stored Communications Act], is tantamount to attaching a GPS device to a person's vehicle . . . . The calculations [made from the historical cellsite information] merely identified a general area . . . ."); United States v. Graham, 2012 WL 691531, *6 (noting that "the GPS location data at issue in *Maynard* was far more precise than the historical cell site location data at issue here"). *But see* United States v. Powell, 943 F. Supp. 2d 759, 767-68 (E.D. Mich. 2013) (evaluating both cell phone tower data and GPS phone data under the same Fourth Amendment standard).

[187] A study found that keys, cash, and phone are the objects that most people consider essential when leaving home. *See* Jan Chipchase, Per Persson, Petri Piippo. Mikko Aarras & Tetsuya Yamamoto, *Mobile Essentials: Field Study and Concepting*, PROC. OF THE 2005 CONF. ON DESIGNING FOR USER EXPERIENCE 57 (2005).

science communities should work collaboratively to answer these, and many more questions, in the time to come. Moreover, as such advances are made, the law on location tracking should continue to keep step with the current state of scientific discovery.