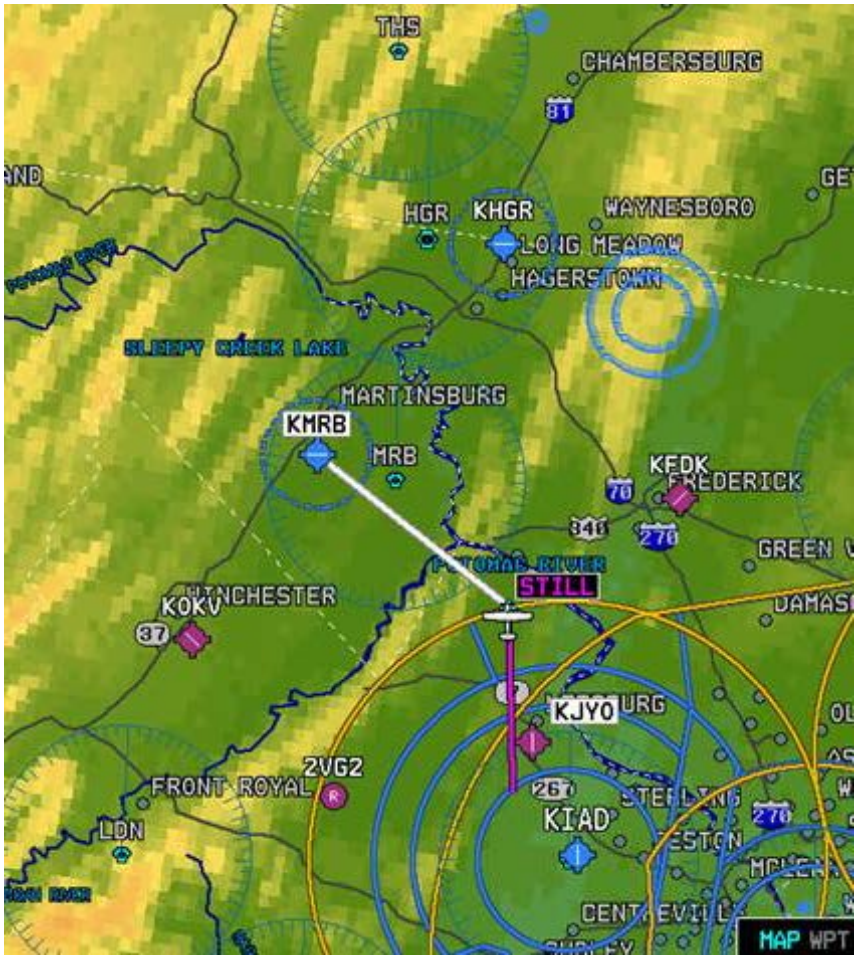
When Enough is Enough: Location Tracking, the Fourth Amendment, and Machine Learning

Steven M. Bellovin
Renée Hutchins
Tony Jebara
Sebastian Zimmeck

Where Are You?

- Who can know?
- Under what conditions?
- In particular, when can law enforcement know?
- *Is a warrant needed?*



(Image from faa.gov)

The Case for No Warrant Needed

- You're in public; anything you do can be observed by anyone
- The Supreme Court held in *Knotts* that using a beeper attached to a drum of chemicals was a legitimate way to follow someone for three days
- How is this different?

On the Other Hand…

- In one recent case (*Jones*), a GPS tracker was affixed for 28 days—does that (should that?) make a difference?
- The economics of GPS-tracking compared with physically following someone have shifted the balance between privacy and law enforcement
- The Court noted in *Kyllo* that technology did make a difference in the legal view of privacy

Legal Background

- Fundamental protection provided by the Fourth Amendment:

  > The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

- The question here: is warrantless GPS tracking *unreasonable?*

Mosaic Theory

(Public domain image from Wikimedia)

- From a collection of individual data points, a very complete picture of an individual can be drawn
- Short observation is fine—but the *total* of many observations is, in fact, unreasonable

(Even One Datapoint May Matter)

"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."

Justice Sotomayor's concurrence in *Jones*

The Legal Status

- Attaching a physical tracker without a warrant is definitely illegal (*Jones*)—it's a physical intrusion
- Legal standard beyond that: "reasonable expectation of privacy" (*Katz*)
- The target must have a subjective expectation of privacy; it must be "one that society is prepared to recognize as 'reasonable'"

What is a “Mosaic”?

- Is there a line between “a few data points” and a mosaic?
- How do you draw this line?
- How does a police officer draw the line?

When is Enough Enough?

"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."

Justice Alito's concurrence in *Jones*

"[I]t remains unexplained why a 4-week investigation is "surely" too long"

Opinion of the Court (by Justice Scalia) in *Jones*

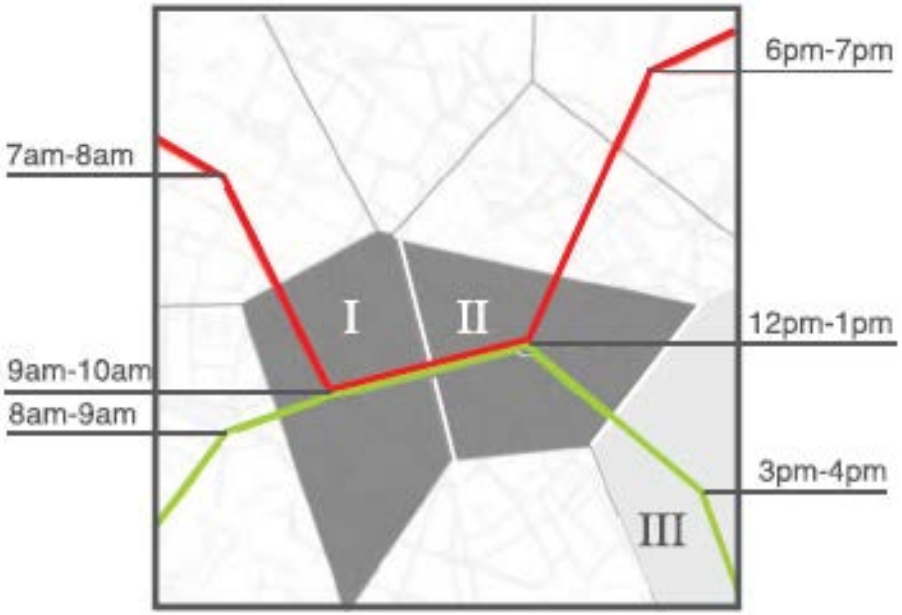
Enter Big Data…

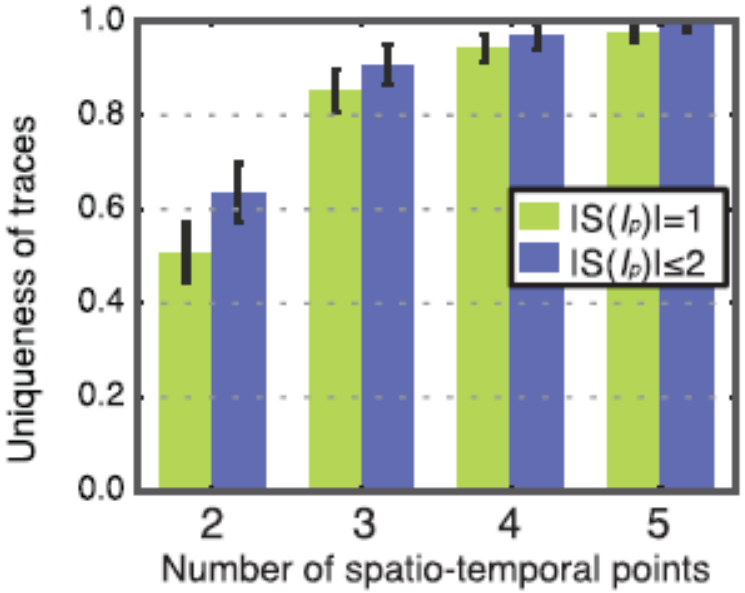
Assertion

- *We have a mosaic when a suitably-trained algorithm can make accurate enough predictions about a person, based on their location history*

- Computer science questions
  - Do mosaics exist?
  - Can machine learning let us draw the line?

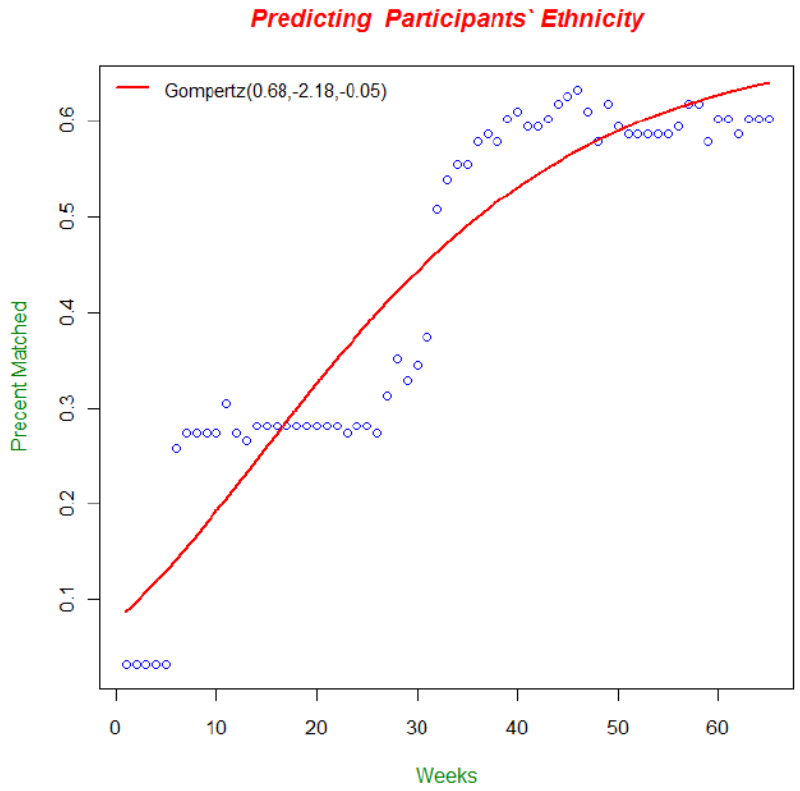
Human Mobility Patterns





de Montjoye et al, Unique in the Crowd: The privacy bounds of human mobility. Nature srep. 3 (2013)

Predictions



Altshuler et al., Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data, *WS3P, IEEE Social Computing* (2012), Figure 10

Creation of a Mosaic

- Intuitively, where the slope is increasing we can learn proportionally more from later observations than from earlier ones, that is, our prediction accuracy increases steeply

- **Where the slope has the highest increase, a transformation in the accuracy of factual predictions occurs and a mosaic is created**

The Second Derivative

- The **Second Derivative** indicates the Rate of Change in the Slope

- At a certain point, law enforcement can learn disproportionately more relative to the effort they have expended

Privacy Metrics

- When do we have a violation of the reasonable expectation of privacy?
- Privacy Metrics
  - k-anonymity
  - l-diversity
  - Others (t-closeness, m-invariance, δ-presence, differential privacy, information theory-based Privacy Metrics)

Mapping Example

- Example: A suspected drug dealer driving in his car picked up a bag somewhere in the Bay Area. After analysis of the location tracking data the machine learning algorithm returns a 40% probability for a pick-up stop in San Francisco.
- Because in our example the probability that the suspect picked up something in San Francisco is 40%, it holds that $l = \lfloor 1/0.4 \rfloor = 2$, that is, our mapping creates 2-diversity. However, because $l = 2 > 1$, we have no violation of the suspect's reasonable expectation of privacy.

Mapping Rules

- Conversion: Probability → l-diversity → reasonable expectation of privacy
- Probability → l-diversity: Given that a machine learning algorithm returns a probability, p, for the existence of an attribute, it holds that $l = \lfloor 1/p \rfloor$.
- l-diversity → reasonable expectation of privacy: If $l = 1$, violation. If $l > 1$, no violation.

Mapping Rationale

- Probabilities for selecting the correct answer from two possibilities at random would be 50%, from three possibilities 33.1/3%, from four 25%, and so on.

- If the probability, *P*, returned from the algorithm is $P > 50\%$, there is a higher chance of being correct when making this selection compared to any other selection. We have 1-diversity. If we have $33.1/3\% > P \leq 50\%$, we obtain 2-diversity. If we have $25\% > P \leq 33.1/3\%$, 3-diversity, and so on.

Probabilities and the Law

- If either *k*-anonymity or *l*-diversity are used for defining the reasonable expectation of privacy, they import a probabilistic understanding of privacy into the law

A Notional Experiment

1. Select training data similar to the type of data to be analyzed.
2. Compile standard set of questions based on questions investigators intend to ask during an investigation and facts that are believed to be learnable.
3. Discard permissible question. The remainder can be used to annotate the training data set and to query a test data set.
4. Analyze test data. From the resulting test data set curves, one for each question (and perhaps for each question/algorithm pair), it can be observed if and where a mosaic forms and the conversion rules can be applied.

Why “Notional”?

- How do we make this approach operational?
- (1) Datasets, (2) Algorithms, (3) Questions
- *Problem: Experiments have not been conducted and experimental results for the type of questions we are interested in are not yet available*

What we can say …

- Location patterns generally form according to the regular organization of human life

  Song et al., Limits of predictability in human mobility, Science 2010: 1018-1021

- The current location of a person "is likely to be a good predictor of [the person's] location exactly one week from now."

  Sadilek & Krumm, Far Out: Predicting Long-Term Human Mobility, Proceeding of AAAI-12, 2012

Yes, There is a Mosaic

- The technical literature supports the idea that mosaics exist
- Although the precise necessary experiments have not been done, we are comfortable in asserting that the limit is about one week
- This limit may decrease as technology improves
- Future Work: Different types of location tracking, Location of the Tracking Device, Aggregation of different types of information