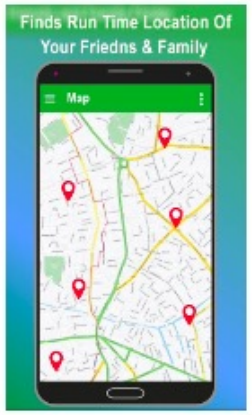
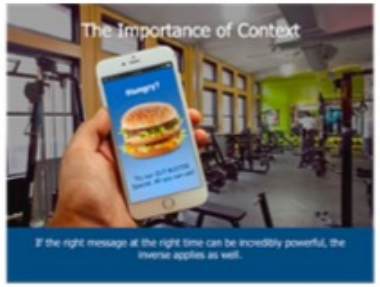
Location Privacy

Asst. Prof. Sinem Sav

With gratitude and special thanks to Carmela Troncoso for some of the slides

Location data is useful…



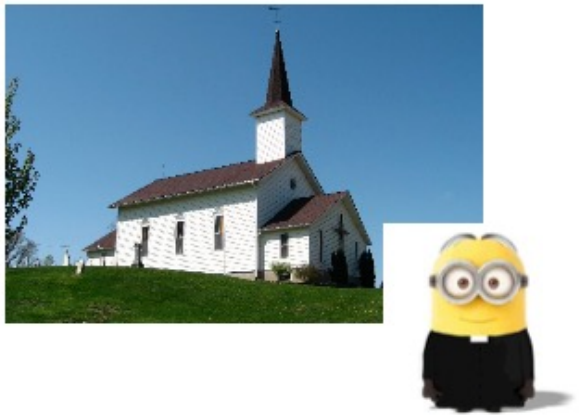


- Google Maps has over 1 billion monthly active users.

- Google Maps is the world's most used mobile app with 54% of global smartphone users accessing it

But location data is sensitive...

Bob



But location data is sensitive...

Bob

Home

Work place

Bob's identity!

Inference: Points of interest (POIs)

- Specific location that someone may find useful or interesting

Why are POIs important?

- **Movements are unique** [De Montjoye et al 2013] [De Montjoye et al 2015]

  4 spatio-temporal points are enough to uniquely identify 95% of people in a mobile phone database of 1.5M people and to identify 90% of people in a credit card database of 1M people

Inference: Points of interest (POIs)

- Specific location that someone may find useful or interesting

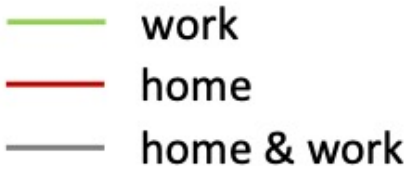
Why are POIs important?

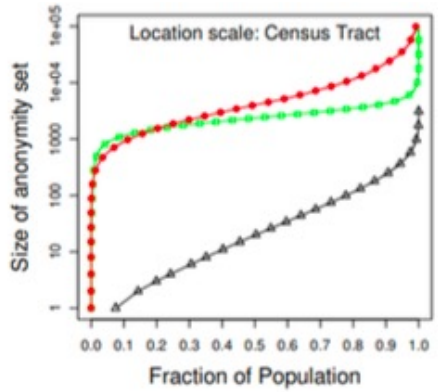
-**Movements are unique** [De Montjoye et al 2013] [De Montjoye et al 2015]

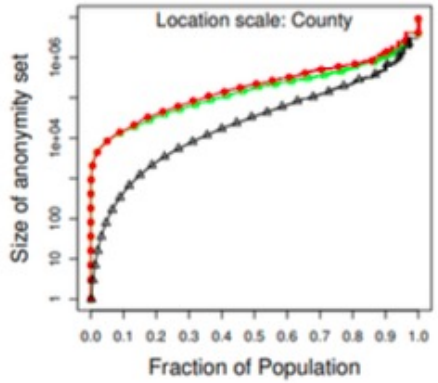
4 spatio-temporal points are enough to uniquely identify 95% of people in a mobile phone database of 1.5M people and to identify 90% of people in a credit card database of 1M people

- **Home and Work: unique identifier** [Golle & Partridge 2009]

given home & work, the median individual's anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census tract and county respectively







A census block is a smallest geogeaphic area and census tract is a geographic area that contains approx. 1200 people in US

Inference: Points of interest (POIs)

- Specific location that someone may find useful or interesting

Why are POIs important?

- **N-top locations: unique identifiers** [Zhang & Bolot 2011]

  [call records] *“top 2" locations likely correspond to home and work locations, the "top 3" to home, work, and shopping/school/commute path locations*

- **Where a user will move next** [Gambs et al 2012]

  Accuracy for the prediction of the next location in the range of 70% to 95%

Hidden Markov Model movement patterns

Inference: Points of interest (POIs)

- Specific location that someone may find useful or interesting

Why are POIs important?

- **Learning about users' motivation** [Bilogrevic et al 2015]

43 % correct classification. Interesting utility impact study [complementary to this lecture]

Inference: Points of interest (POIs)

- Specific location that someone may find useful or interesting

Why are POIs important?

- **Learning about users' motivation** [Bilogrevic et al 2015]

  43 % correct classification. Interesting utility impact study [complementary to this lecture]

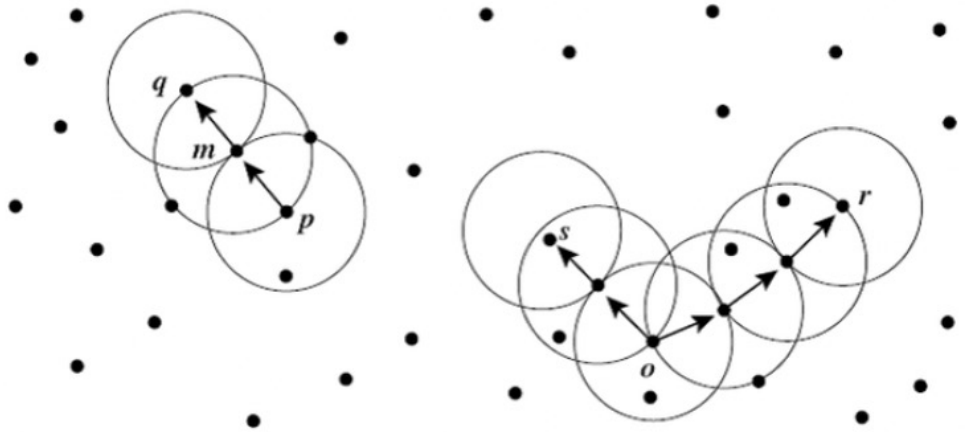
- **Learning Demographics and other Patterns**

  [Pang and Zhang 2017] [Felbo et al 2017][Cho et al 2010] [Liao et al 2005] [Liao et al 2007]
  Machine-learning based frameworks

Inference: Points of interest (POIs)

POI extraction: clustering techniques [Ester et al 1996][Ashbrook & Starner 2003][Krumm 2007]

- DBSCAN [Ester et al 1996]

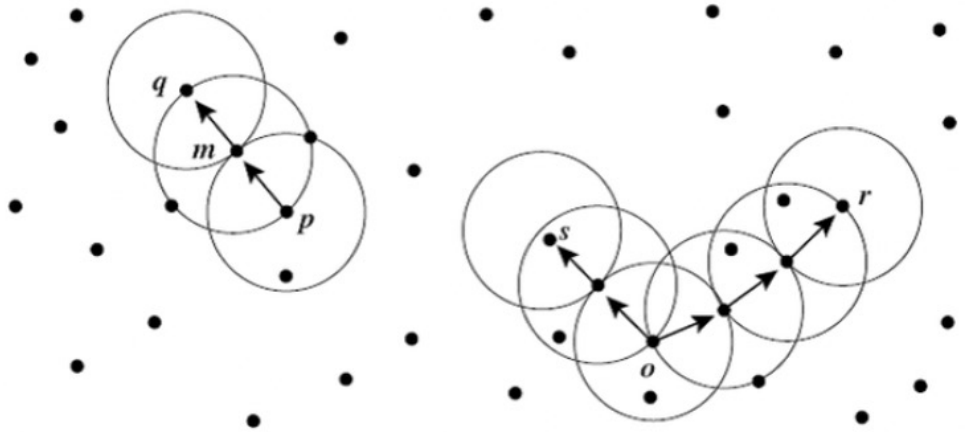


https://www.quora.com/How-does-DBSCAN-algorithm-work

Inference: Points of interest (POIs)

POI extraction: clustering techniques [Ester et al 1996][Ashbrook & Starner 2003][Krumm 2007]

- DBSCAN [Ester et al 1996]



https://www.quora.com/How-does-DBSCAN-algorithm-work

And after finding the clusters?

Home and work: identified by time

One can split clusters even more (e.g., using X-Means [Pelleg & Moore 2000])

Once cluster is found, inferences can be automatized using reverse geo-coding on the centroids

"We stop clustering when the nearest two clusters are over 100 meters apart. The home location is taken as the centroid of the cluster with the most points. " [Krumm 2006]

Inferenc

- Specific location

Why are POIs

- POI ext ner 2003][Krumm 2007]

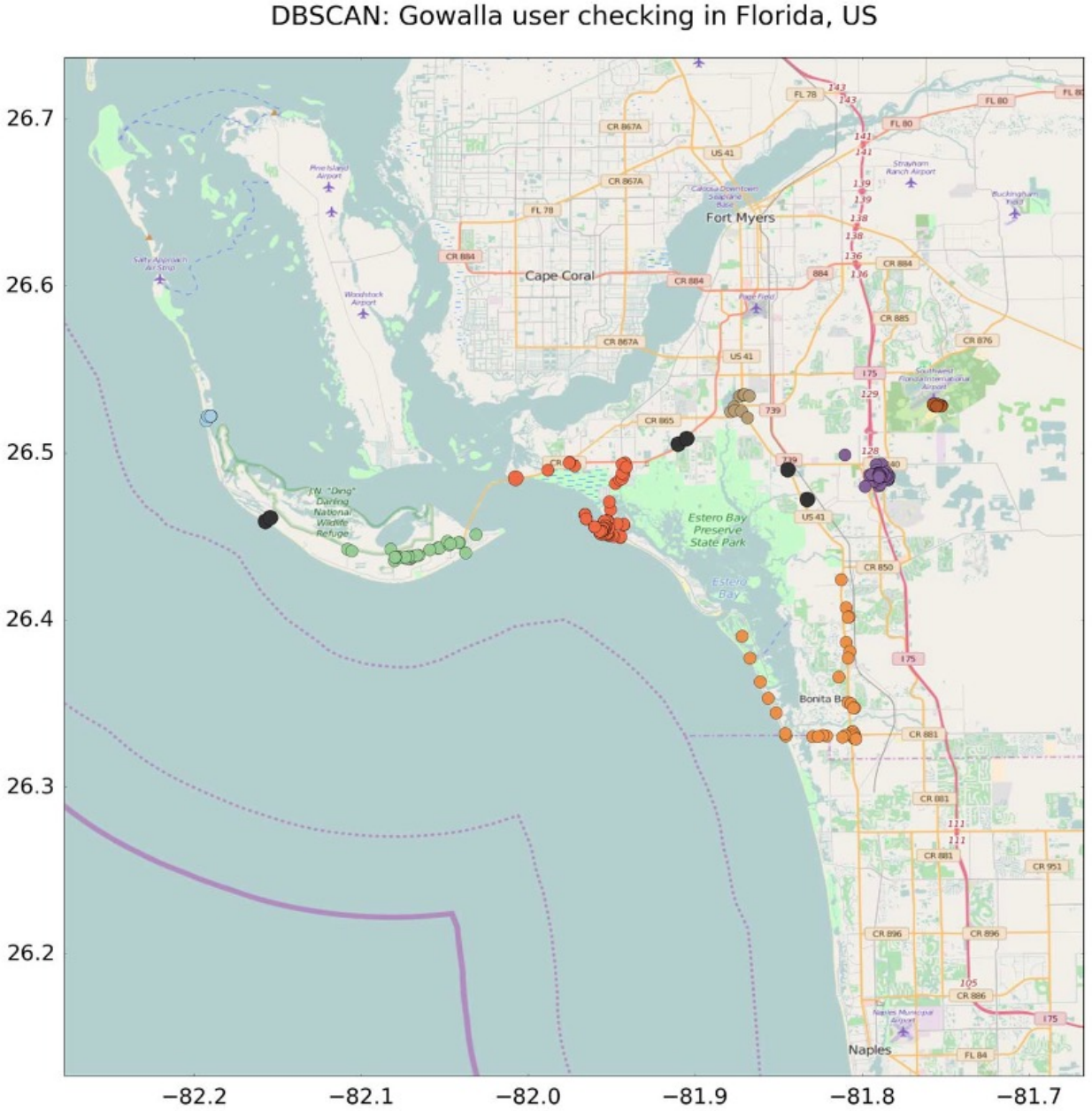


Figure 2. An example of clusters extracted from a user using the Gowalla data set in the area of Cape Coral, Florida. Note the denser collection of points correctly mapped to clusters and the isolated black dots marked as outliers. Figure courtesy of Natalino Busa. Map overlay: OpenStreet Map.

Inference: Points of interest (POIs)

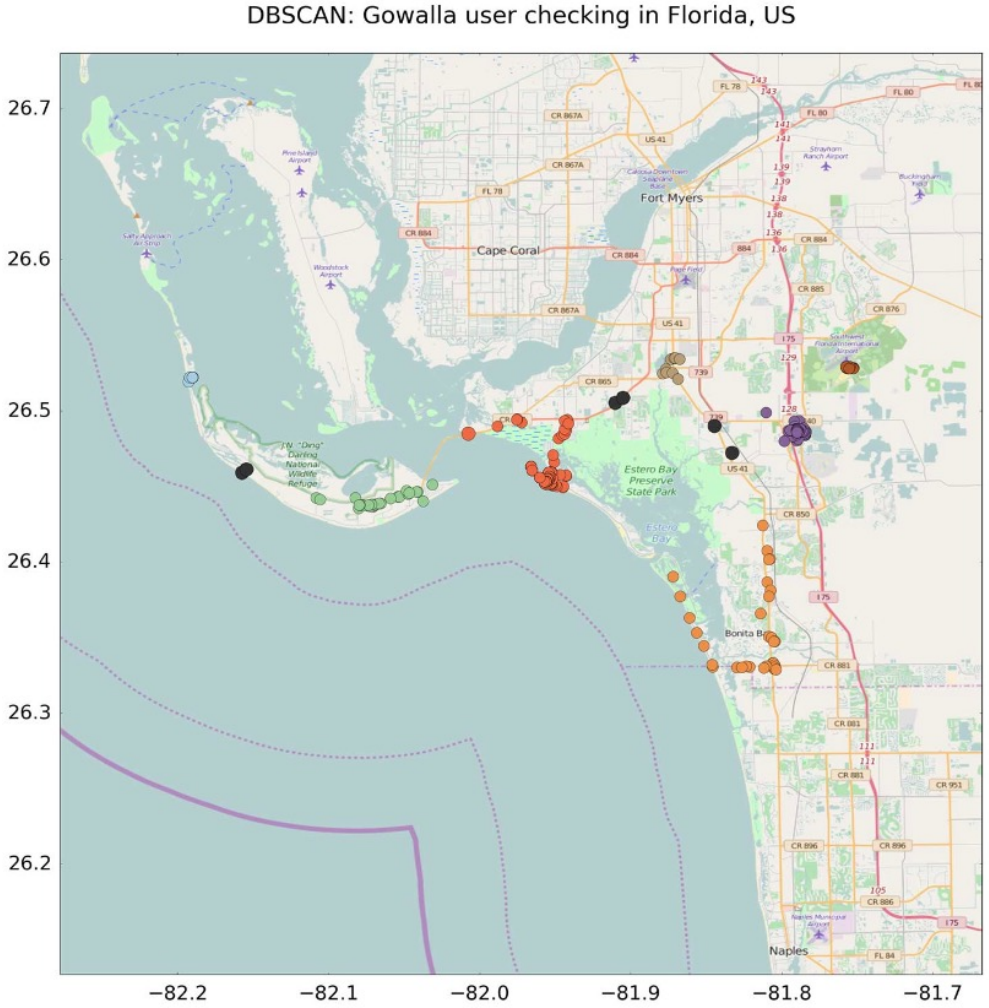


Figure 2. An example of clusters extracted from a user using the Gowalla data set in the area of Cape Coral, Florida. Note the denser collection of points correctly mapped to clusters and the isolated black dots marked as outliers. Figure courtesy of Natalino Busa. Map overlay: OpenStreet Map.

"Figure 2 shows an example of clusters extracted from an anonymous user using Gowalla, a social networking site where users share their locations by checking in at specific places…
…Events have been clustered according to their geographical position. So, for instance, the hike at Estero Bay (dark orange dots), the venues at the airport (brown dots), and the venues at Sanibel Island (green dots) belong to separate clusters (*epsilon* set to 3 km, and *minPoints* set to 3)."

https://www.oreilly.com/content/clustering-geolocated-data-using-spark-and-dbscan/

What is location privacy?

What is location privacy?

- ‘‘is the ability of an individual to move in public space with the expectation that under normal circumstances their location will not be systematically and secretly recorded for later use’’. [Blumberg and Eckersley, 2009]

What is location privacy?

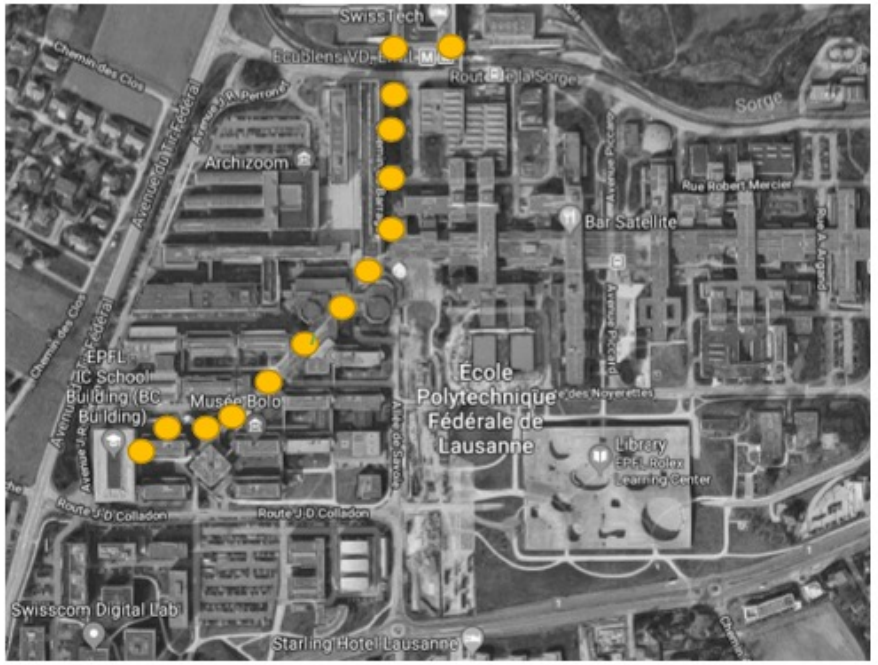
- ‘‘is the ability of an individual to move in public space with the expectation that under normal circumstances their location will not be systematically and secretly recorded for later use’’. [Blumberg and Eckersley, 2009]
- They also argue that there is no absolute location privacy, because: ‘‘...when you leave your home you sacrifice some privacy. Someone might see you enter the clinic on Market Street, or notice that you and your secretary left the Hilton Gardens Inn together.’’ [Blumberg and Eckersley, 2009]

What is location privacy?

- “Location privacy is concerned with ensuring that individuals have control over **who** can access their location information and **how** that information is used.”
- ??

Protecting location privacy

What can we do??



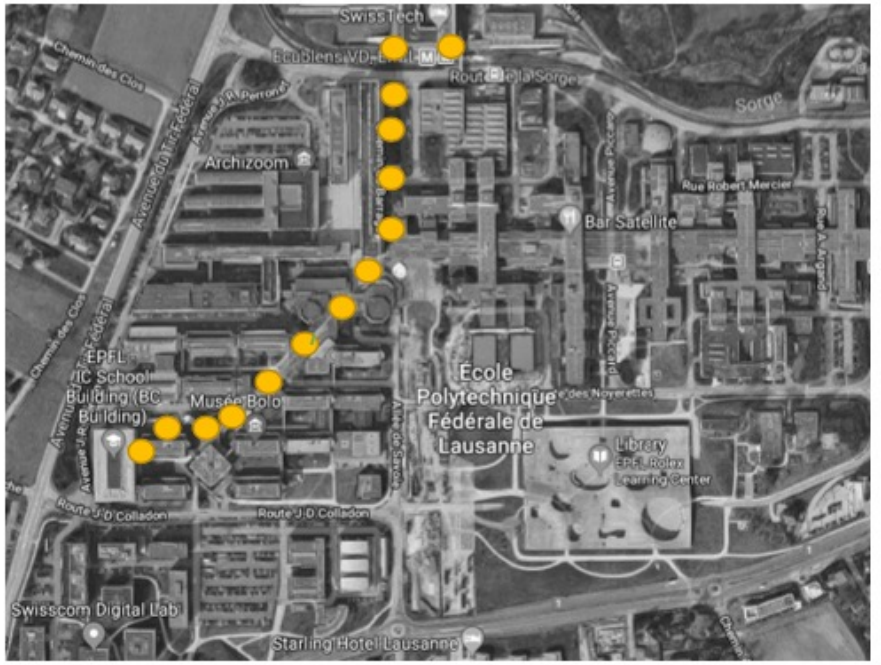
Protecting location privacy

What can we do??

4 main techniques:

Perturbation
Hiding
Generalization
Adding dummies

(+ Synthetic data)



Protecting location privacy

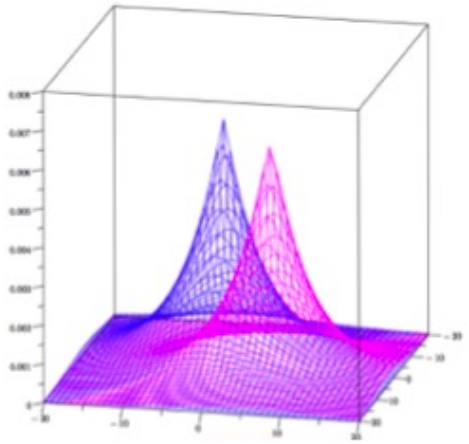
Spatial Obfuscation: Perturbation of locations using noise [Duckham & Kulik 2005]

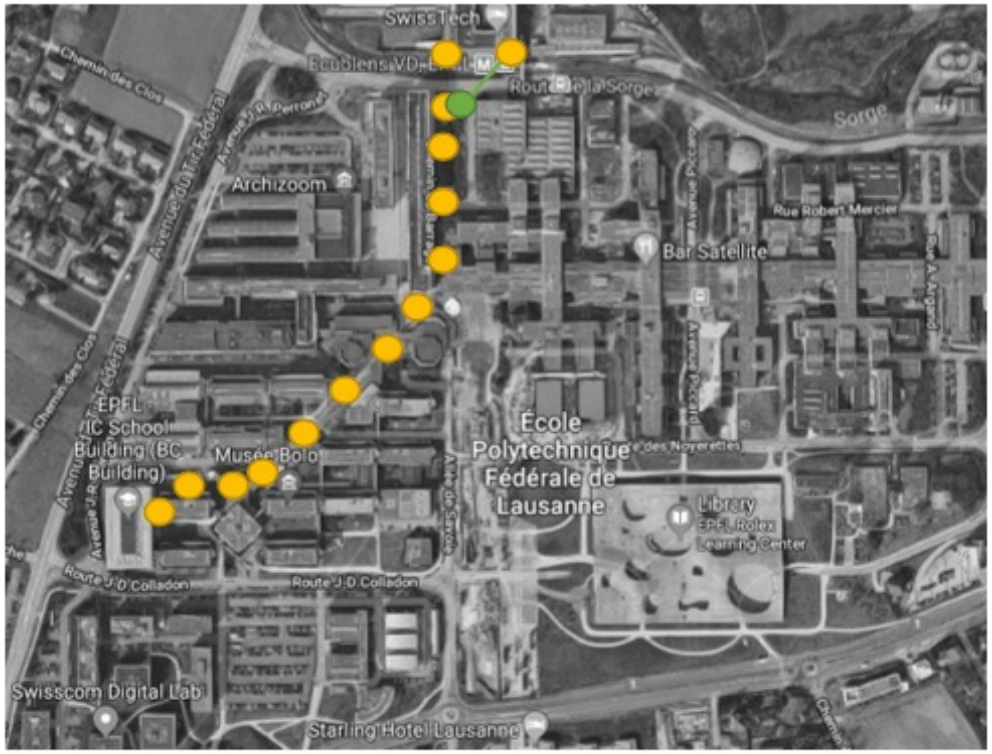
Geo-indistinguishability [Andres et al 2013]

Add 2-dimensional
ϵ - differential privacy noise

$Pr(● | ●) = Pr(● | ●') * \epsilon^{r}$

radius





Protecting location privacy

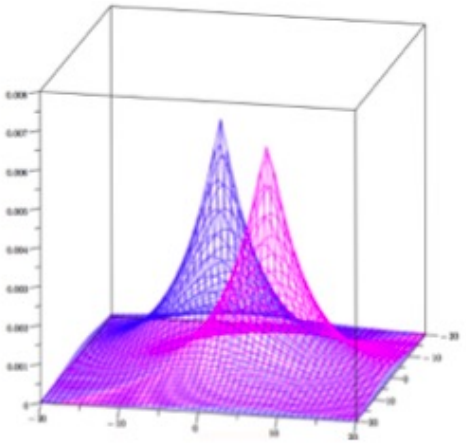
Spatial Obfuscation: Perturbation of locations using noise [Duckham & Kulik 2005]

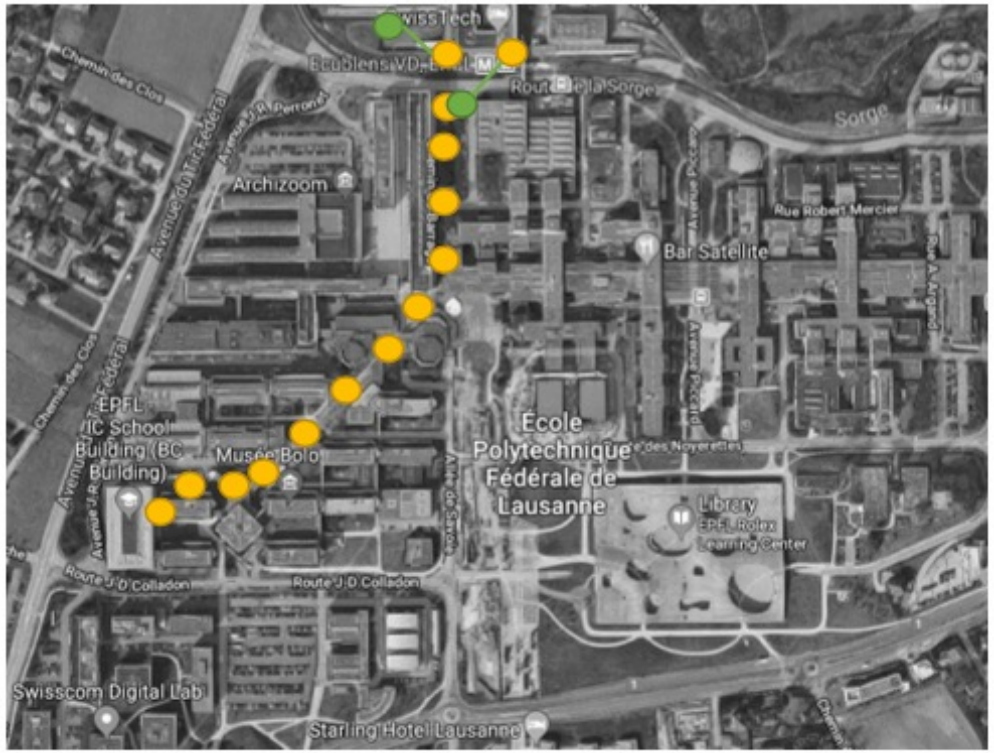
Geo-indistinguishability [Andres et al 2013]

Add 2-dimensional
ϵ - differential privacy noise

$Pr(● | ●) = Pr(● | ●') * \epsilon^{r}$

radius





Protecting location privacy

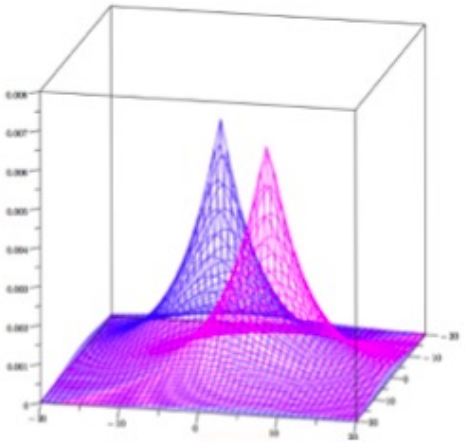
Spatial Obfuscation: Perturbation of locations using noise [Duckham & Kulik 2005]

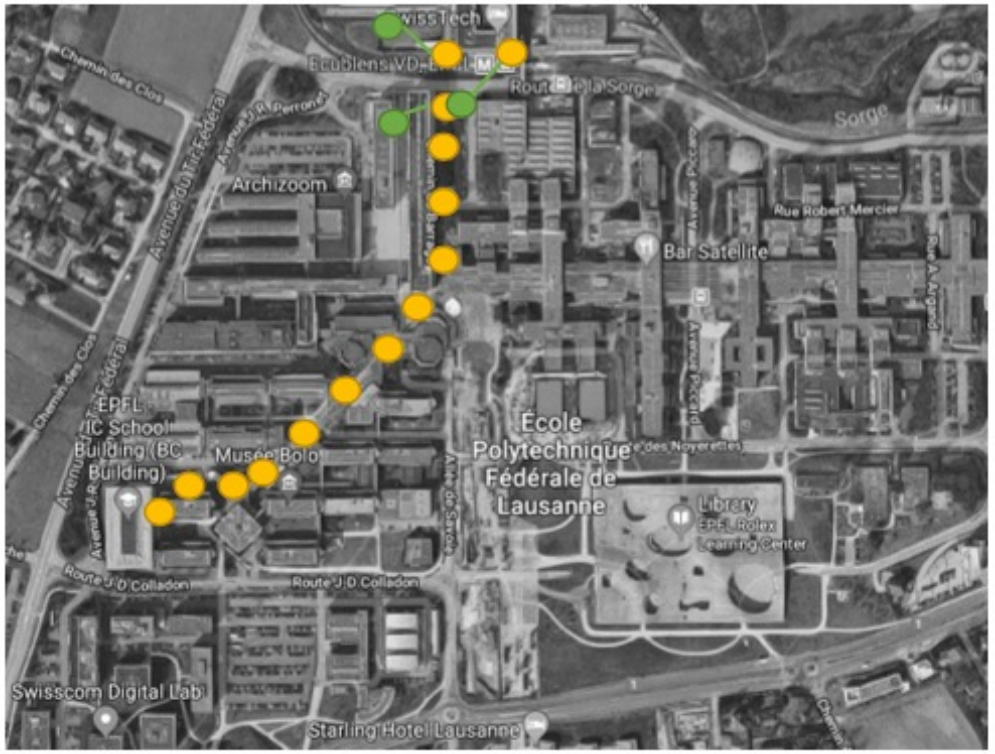
Geo-indistinguishability [Andres et al 2013]

Add 2-dimensional
ϵ - differential privacy noise

Pr(● | ●) = Pr(● | ●') * ϵr

radius





Protecting location privacy

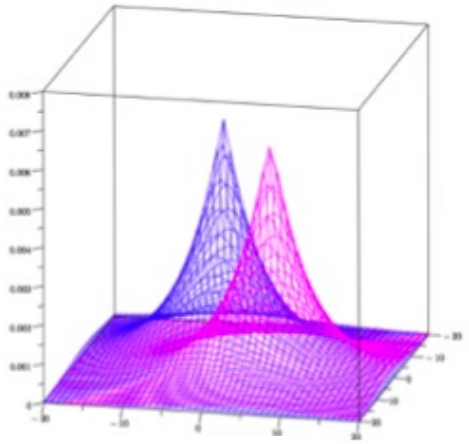
Spatial Obfuscation: Perturbation of locations using noise [Duckham & Kulik 2005]

Geo-indistinguishability [Andres et al 2013]

Add 2-dimensional
ϵ - differential privacy noise

$Pr(z|x) = Pr(z|x') * \epsilon^{r}$

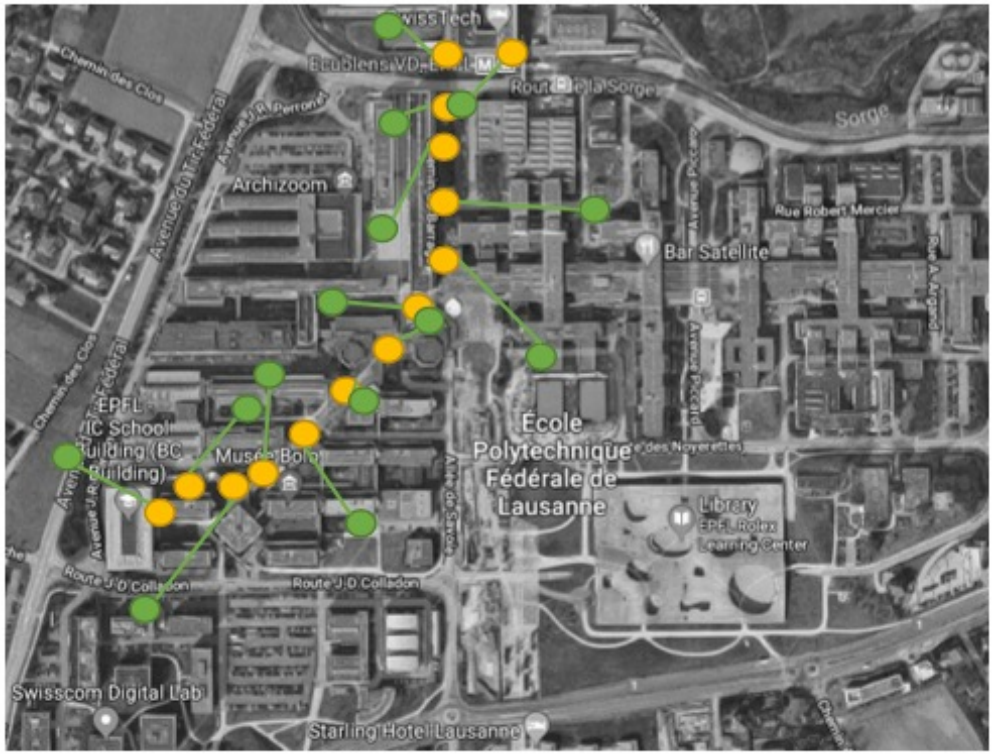
radius



ϵ ↑ —

ϵ ↓ ——————

Significant Privacy vs. utility trade-off
[Oya et al 2017b]



Protecting location privacy

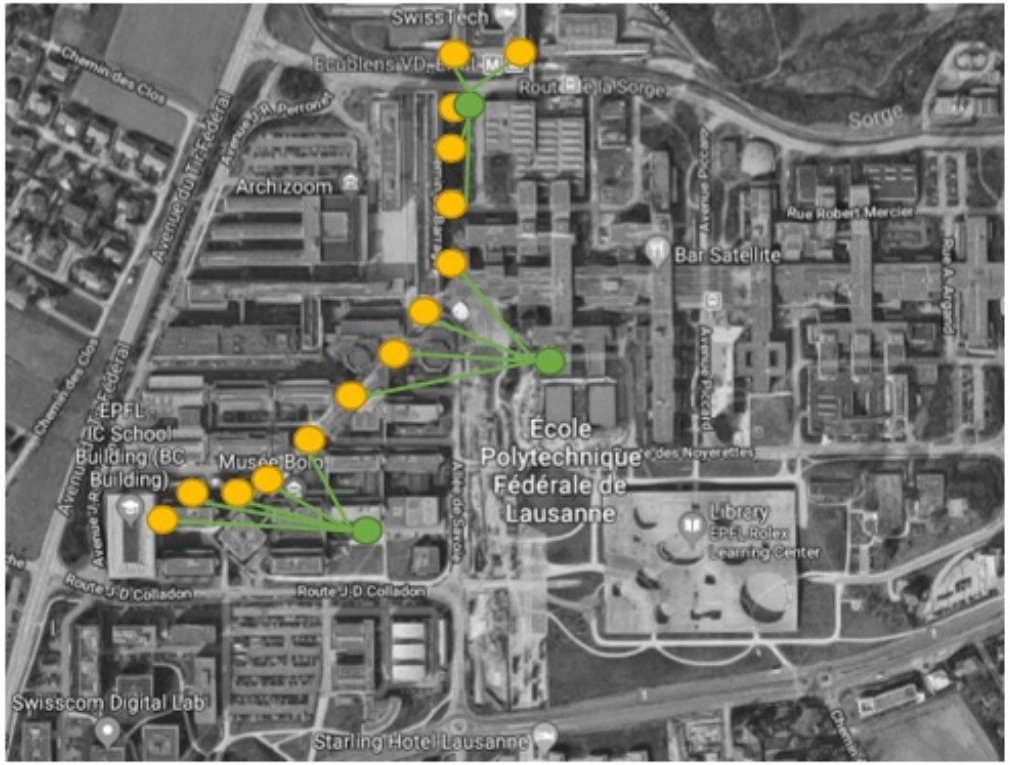
Spatial Obfuscation: Perturbation of locations using noise [Duckham & Kulik 2005]

Geo-indistinguishability [Andres et al 2013]

As with differential privacy, protection decreases linearly with every sample → fast degradation

Release Geo-indinstinguishability: only draw noise when needed to keep utility (i.e., when moving far from previous sample)
[Chatzikokolakis et al 2014]

Significant Privacy vs. utility trade-off
[Oya et al 2017b]

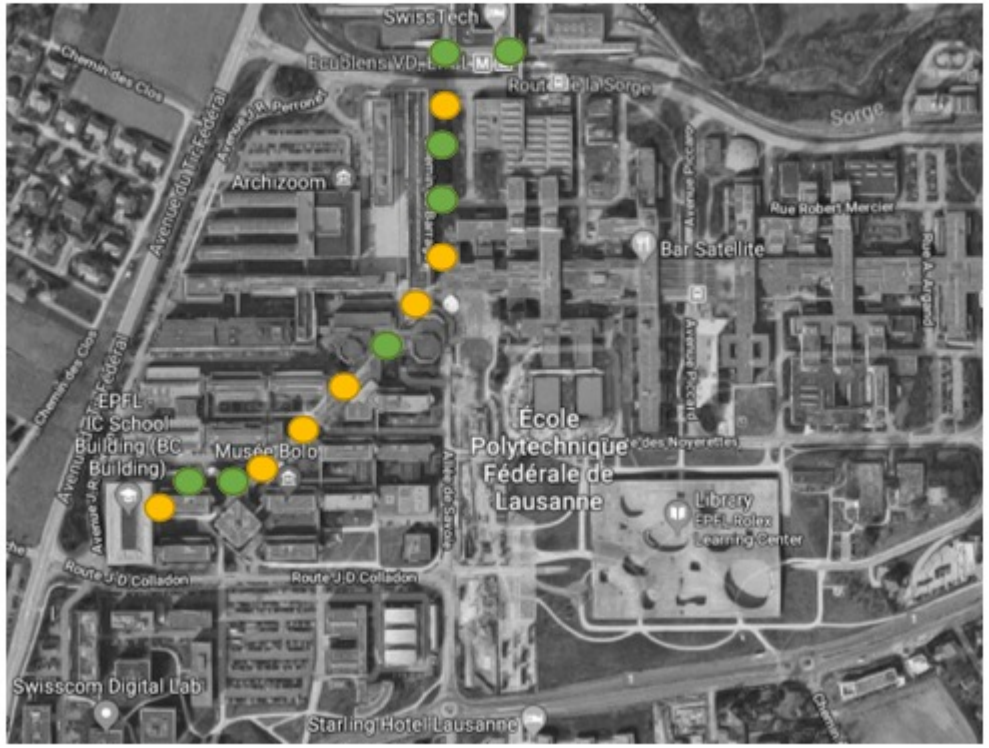


Protecting location privacy

Hiding: Not reporting some of the locations [Huang 2006][Hoh 2007]

Random Hiding: Reveal a percentage of the points chosen at random (e.g, 50% on average)

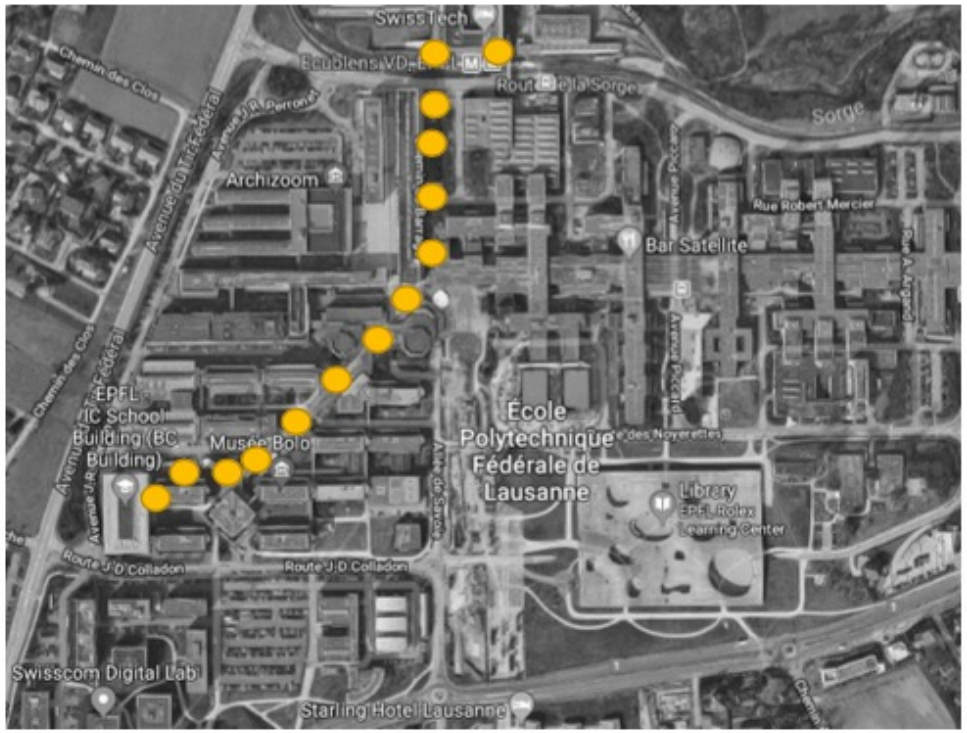
Release: Reveal points only when needed



Protecting location privacy

Generalization: reduce the precision of the reported locations

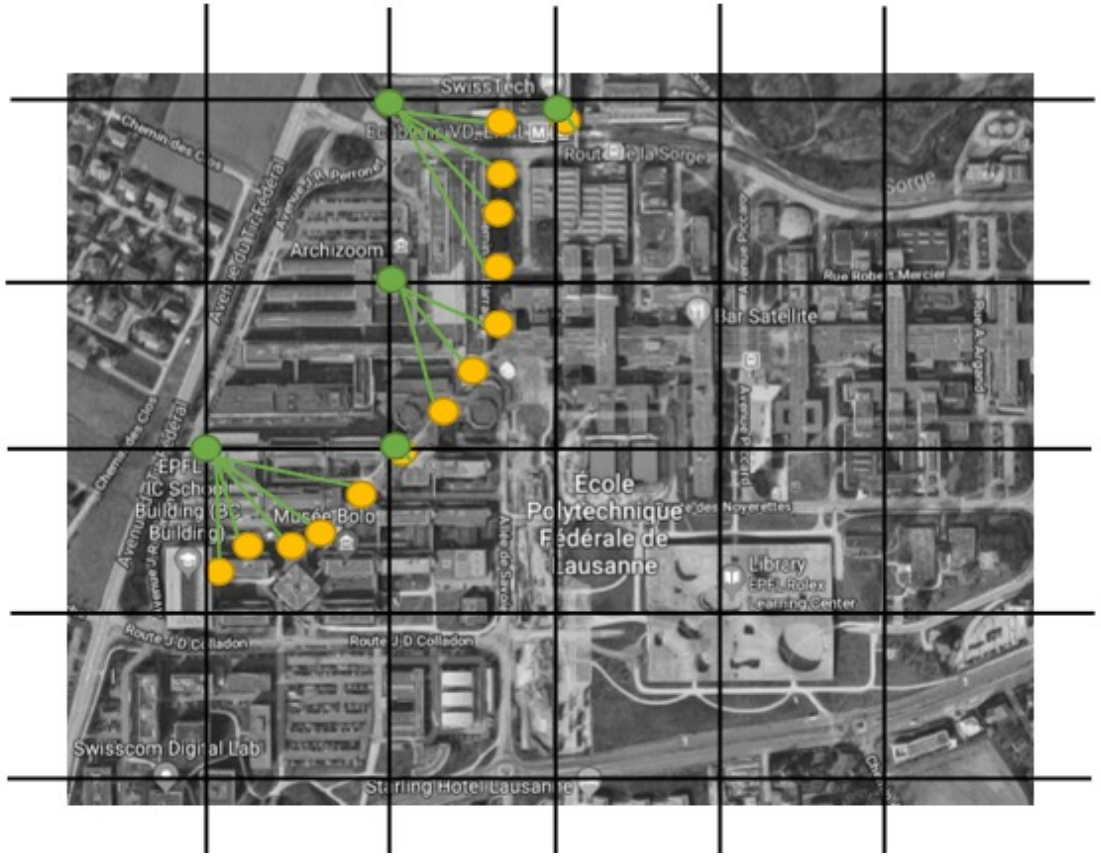
[Bamba et al 2008]



Protecting location privacy

Generalization: reduce the precision of the reported locations
[Bamba et al 2008]

Discretization: Map to grid points (Rounding- Floor)
[Krumm 2009]

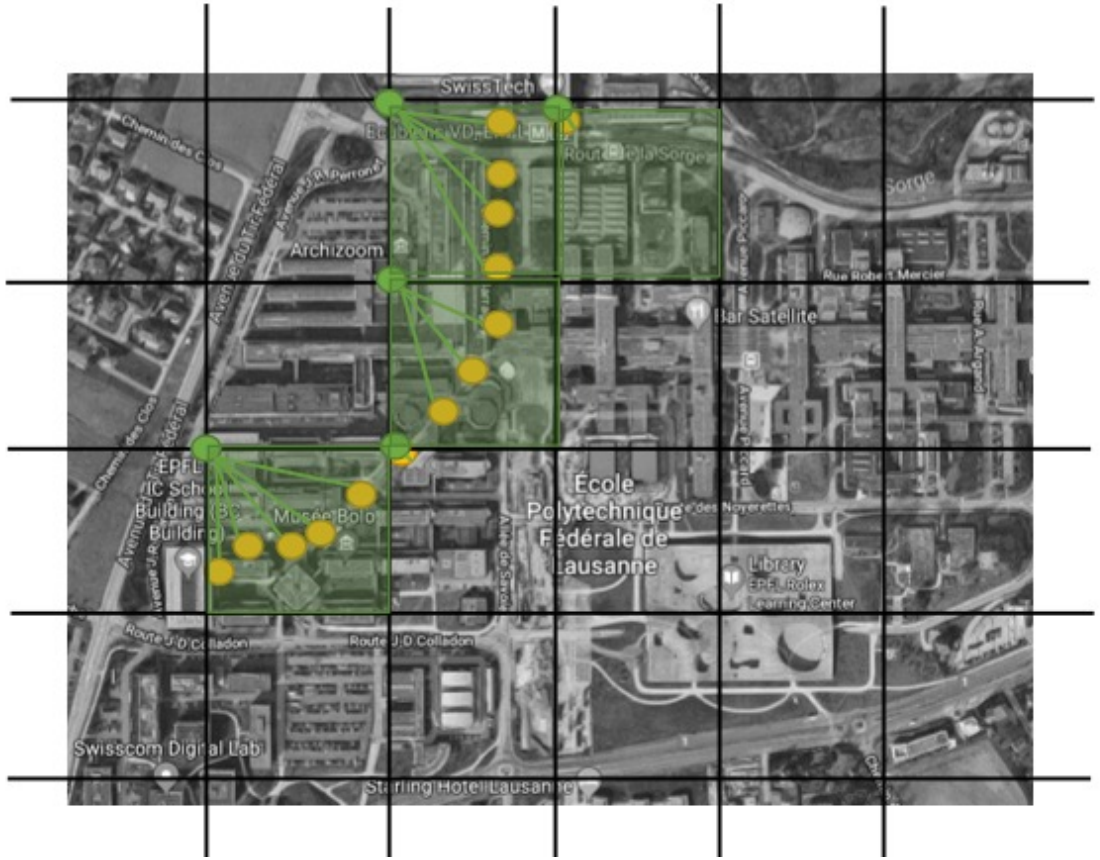


Protecting location privacy

Generalization: reduce the precision of the reported locations
[Bamba et al 2008]

Cloaking: Reveal a region
- Fixed cloaks: always map to the same cloak

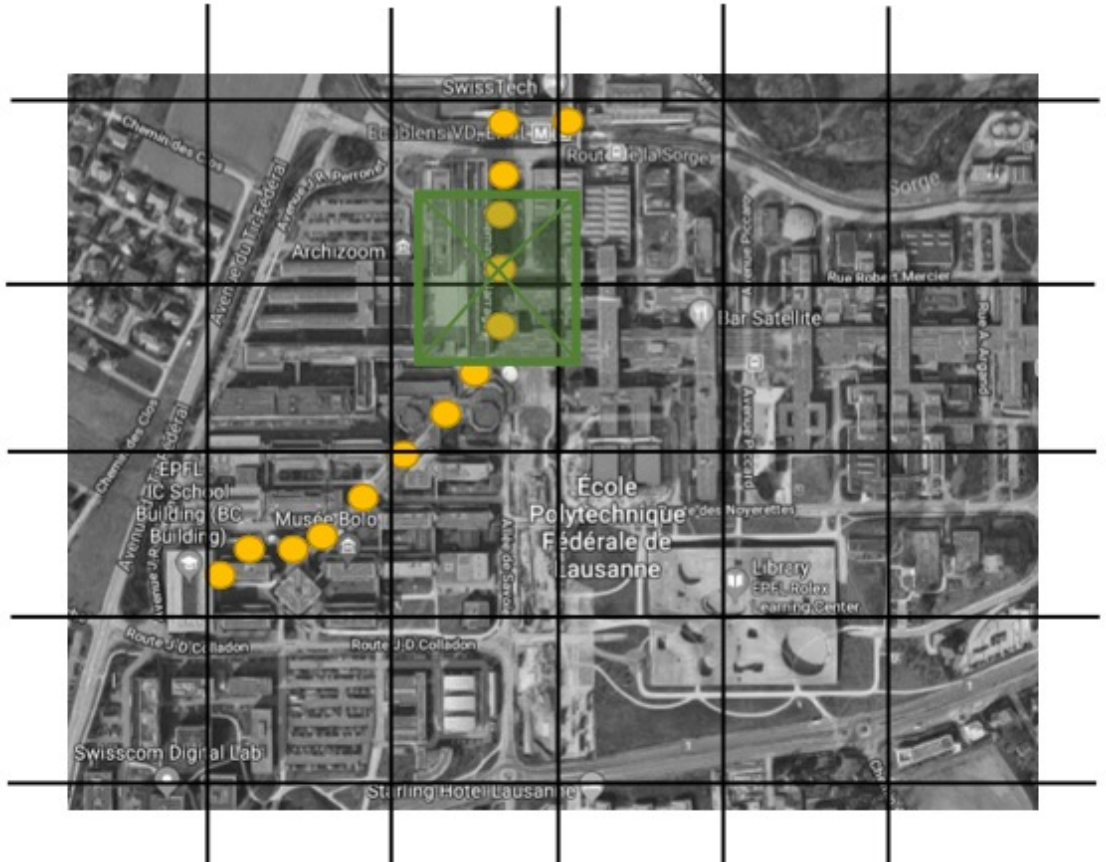
Aggregate location on a cloak can be calculated via different functions, e.g., grid corners, centroid of points etc.



Protecting location privacy

Generalization: reduce the precision of the reported locations
[Bamba et al 2008]

Cloaking: Reveal a region
- Fixed cloaks: always map to the same cloak
- Location-dependent cloaks (centered on location)



Protecting location privacy

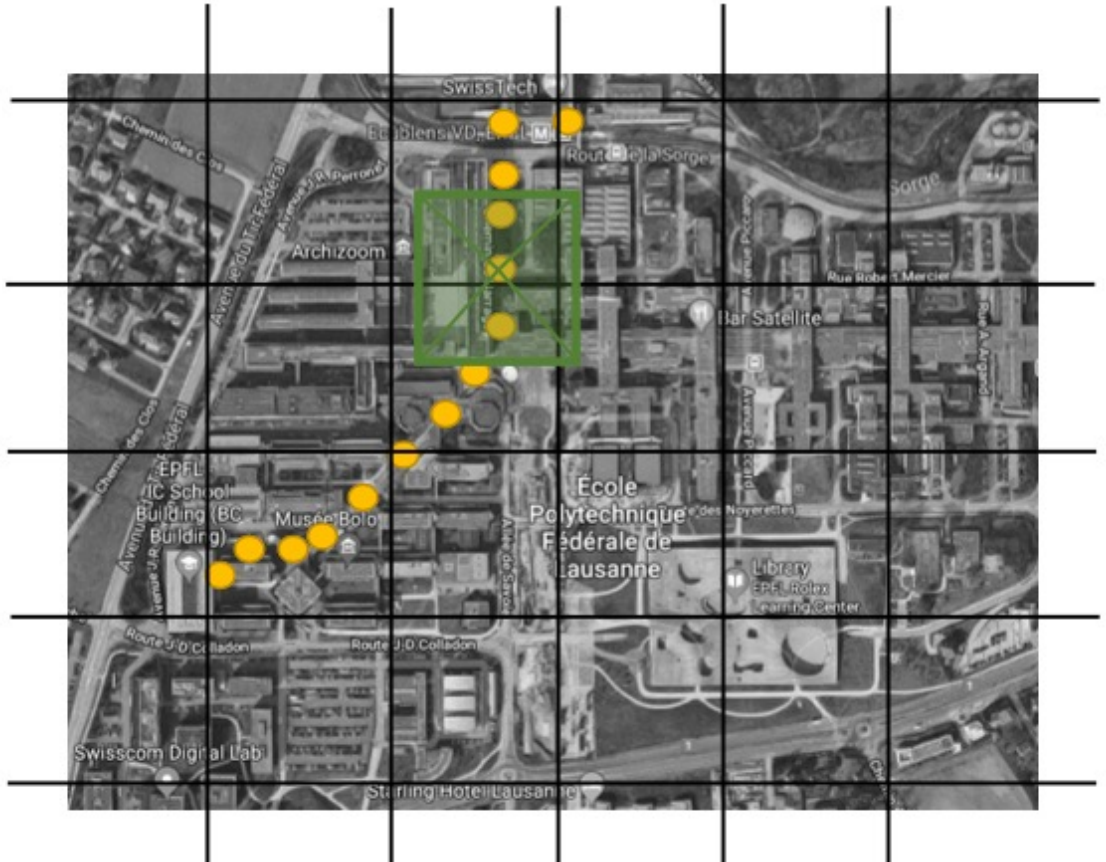
Generalization: reduce the precision of the reported locations

[Bamba et al 2008]

Cloaking: Reveal a region

- Fixed cloaks: always map to the same cloak
- Location-dependent cloaks (centered on location)
- **k-anonymity based**

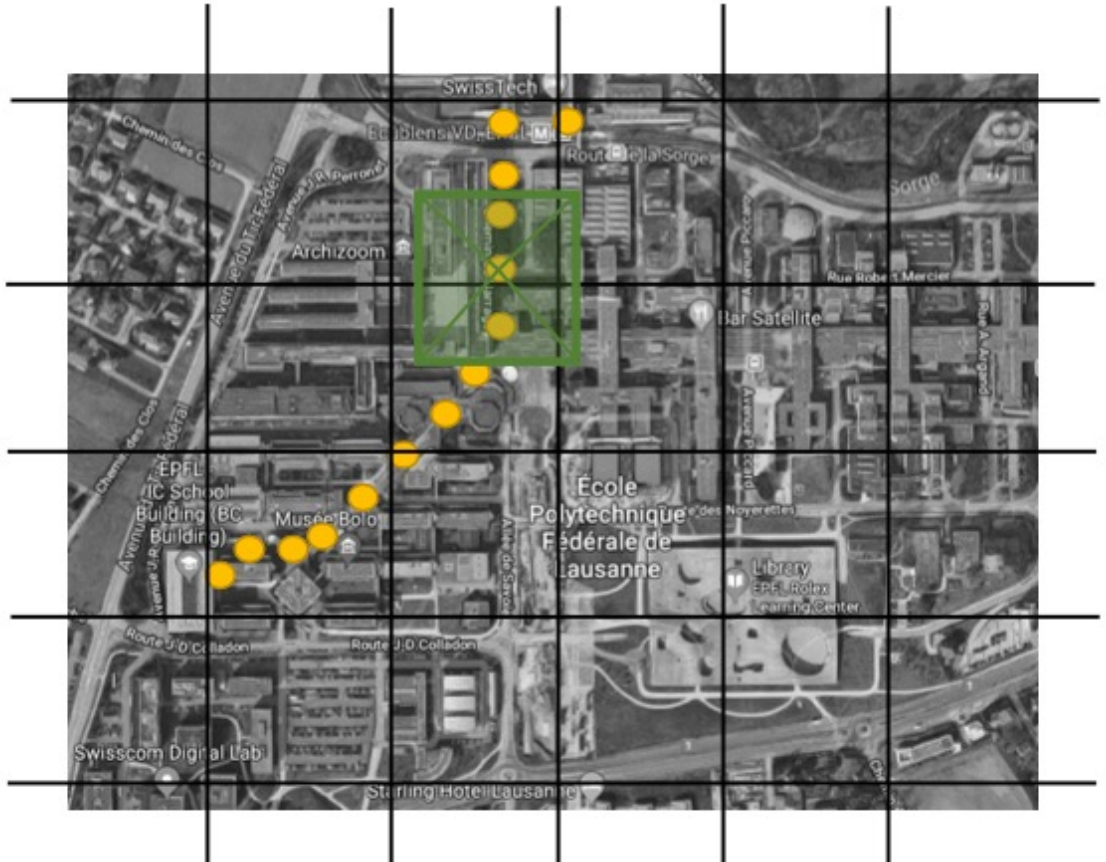
cloak the exact location of an individual by grouping them with at least k-1 other individuals who share similar location attributes (to make it difficult to distinguish the target individual from the other k-1 members



Protecting location privacy

Generalization: reduce the precision of the reported locations
[Bamba et al 2008]

Cloaking: Reveal a region
- Fixed cloaks: always map to the same cloak
- Location-dependent cloaks (centered on location)
- **k-anonymity based**

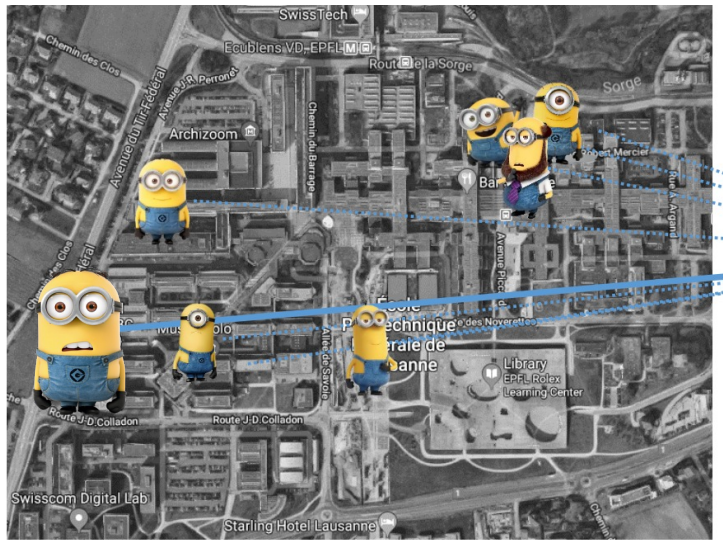


Any problems?

Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works



(x,y,Q) where (x,y) is the location and Q the query

(x,y,Q)

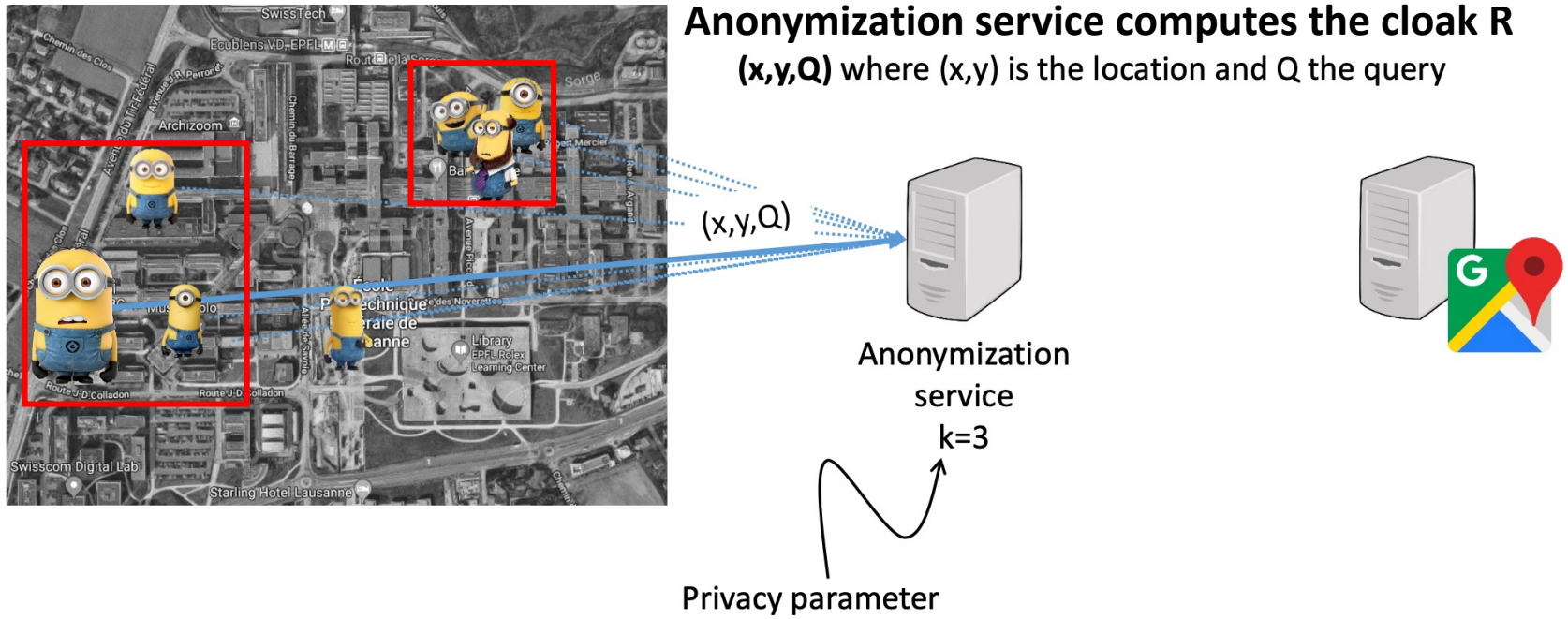
Anonymization
service
k=3

Privacy parameter

Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works



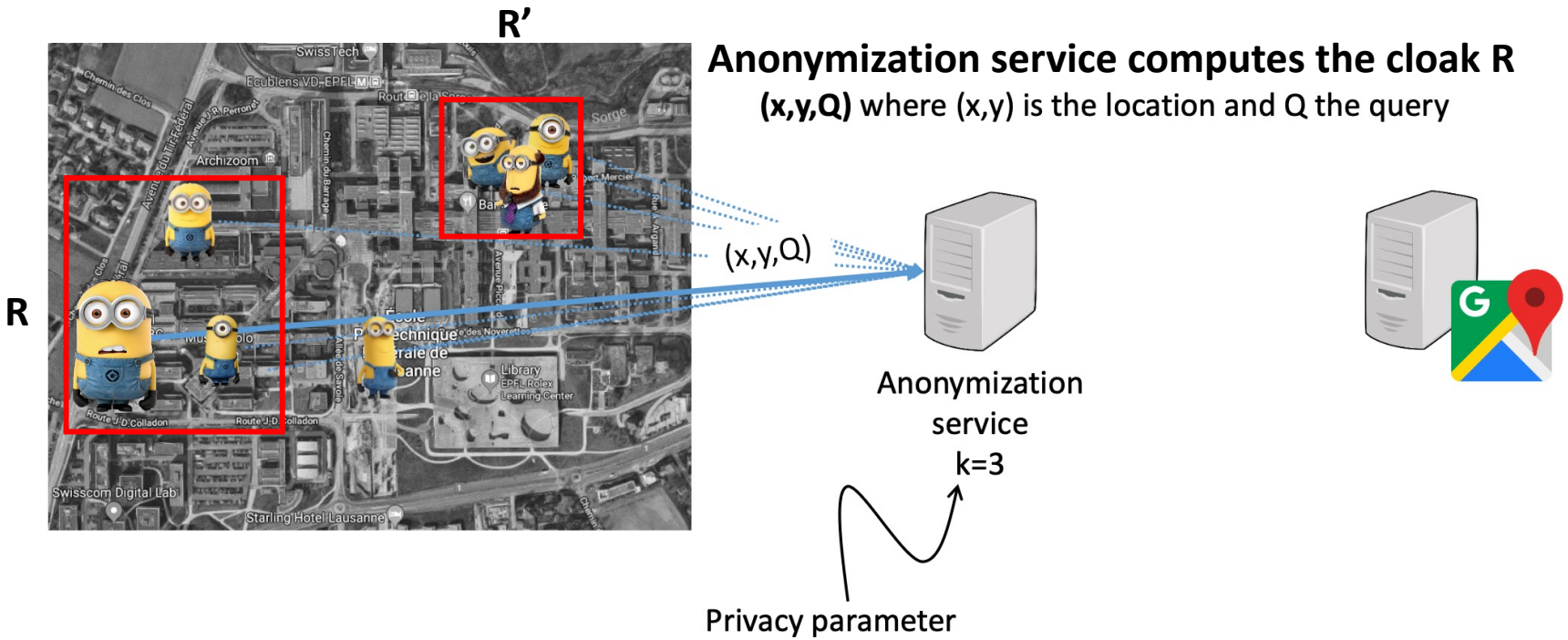
Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works

Problem 1: k!=location privacy

k = 3 , R vs R' location?



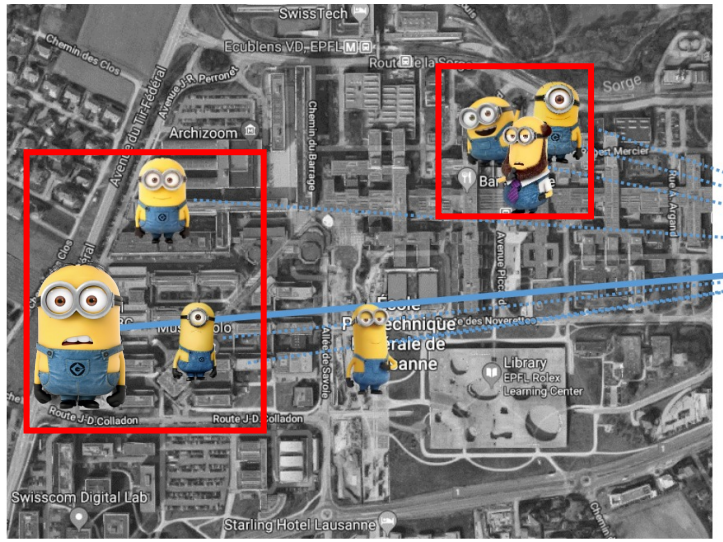
Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works

Problem 1: k!=location privacy

Problem 2: How to choose k?



Anonymization service computes the cloak R

(x,y,Q) where (x,y) is the location and Q the query

(x,y,Q)

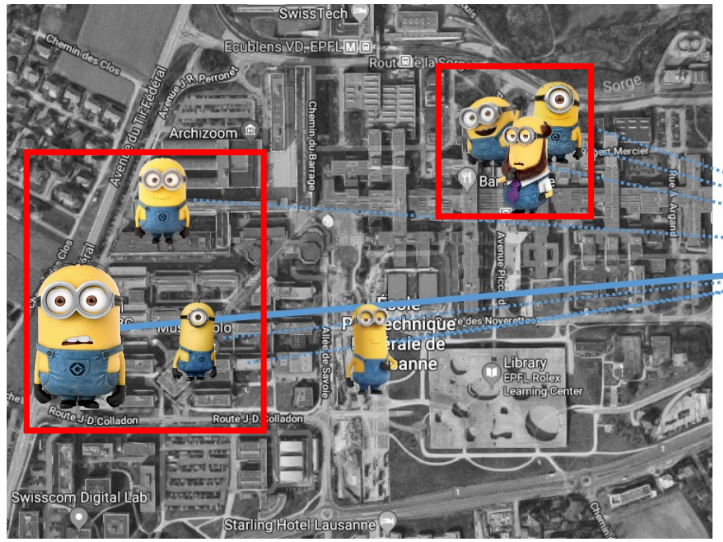
Anonymization service k=3

Privacy parameter

Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works



Anonymization service computes the cloak R

(x,y,Q) where (x,y) is the location and Q the query

(x,y,Q)

Anonymization service
k=3

Privacy parameter

Problem 1: k!=location privacy

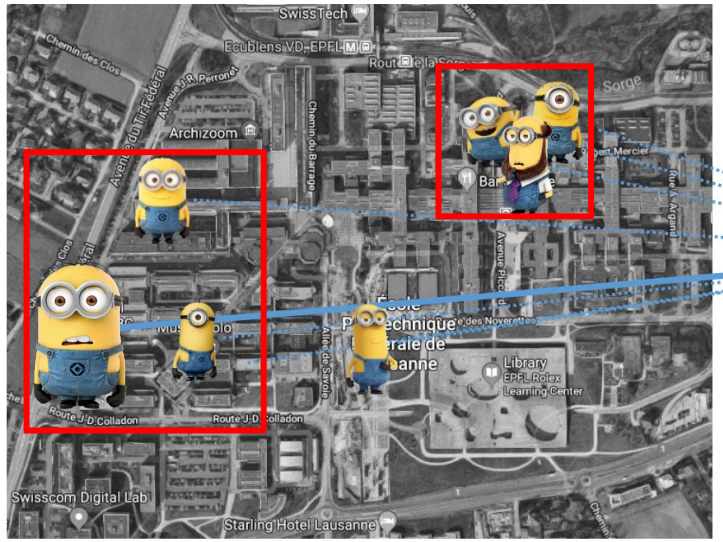
Problem 2: How to choose k?

Problem 3: Auxilliary information, query-based attacks, dynamic environments

Protecting location privacy

Any problems?

[Gruteser & Grunwald 2003] and a long, long, long list of follow-up works



Anonymization service computes the cloak R

(x,y,Q) where (x,y) is the location and Q the query

(x,y,Q)

Anonymization service
k=3

Privacy parameter

Problem 1: k!=location privacy

Problem 2: How to choose k?

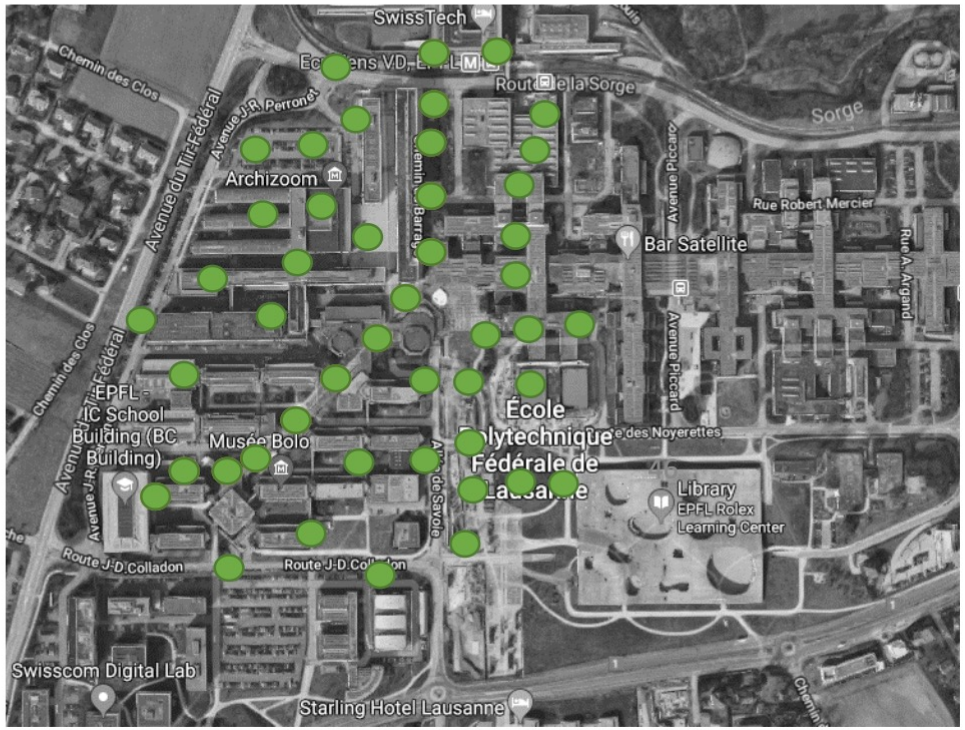
Problem 3: Auxilliary information, query-based attacks, dynamic environments

If server knows location (e.g., query metadata)
Anonymity but not location privacy!!

Protecting location privacy

Dummy Locations: add decoy locations [Meyerovitz & Choudhury 2009]

What is the problem??



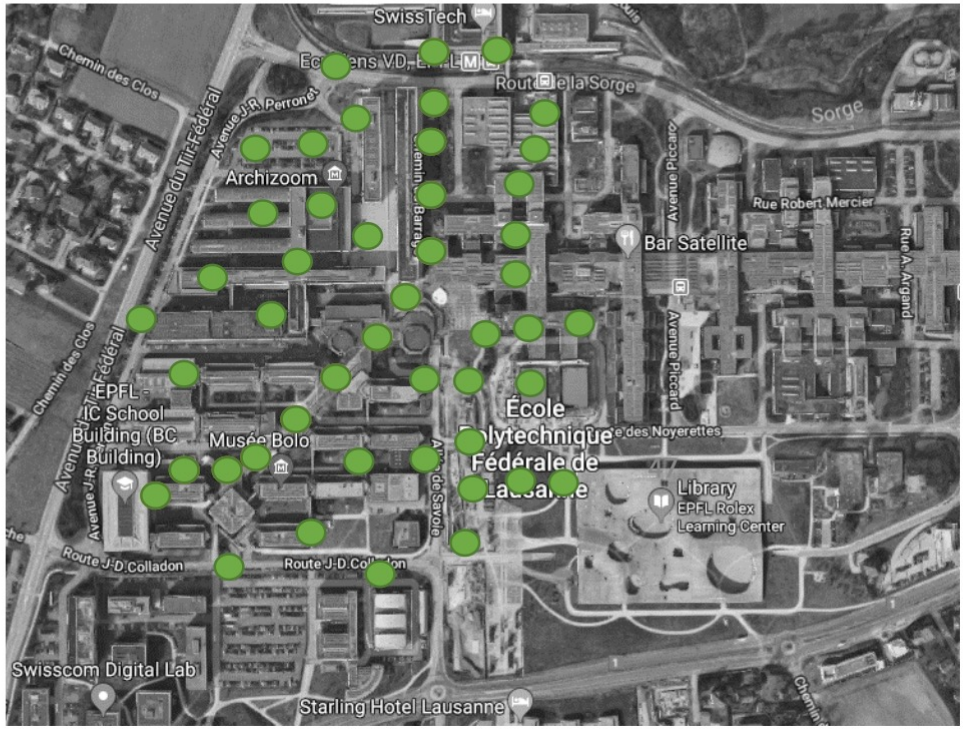
Protecting location privacy

Dummy Locations: add decoy locations [Meyerovitz & Choudhury 2009]

What is the problem??

Difficult to create plausible dummies
[Chow & Golle 2009]

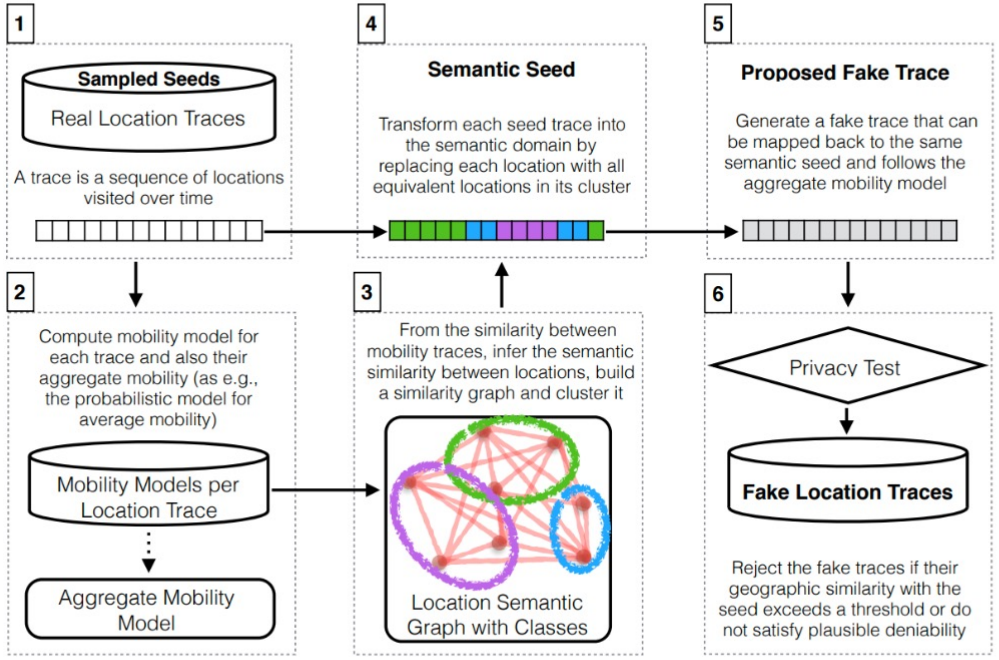
For some use cases:
How to assign values to these points??

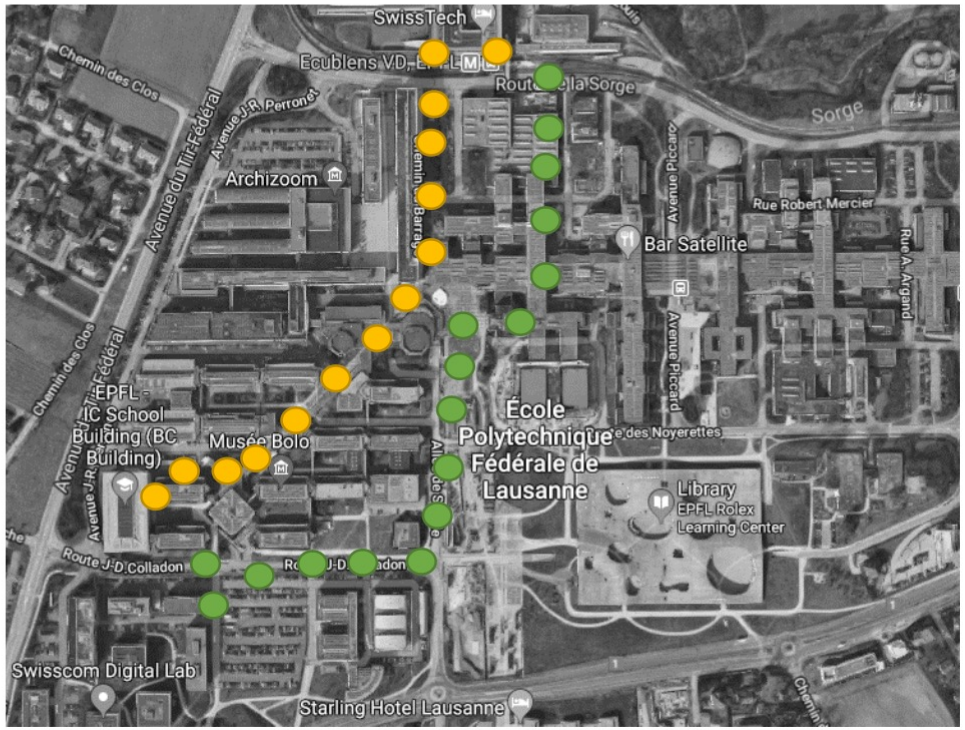


Protecting location privacy

Synthetic Data: fully new locations from the same distribution

[Bindschaedler & Shokri 2016]



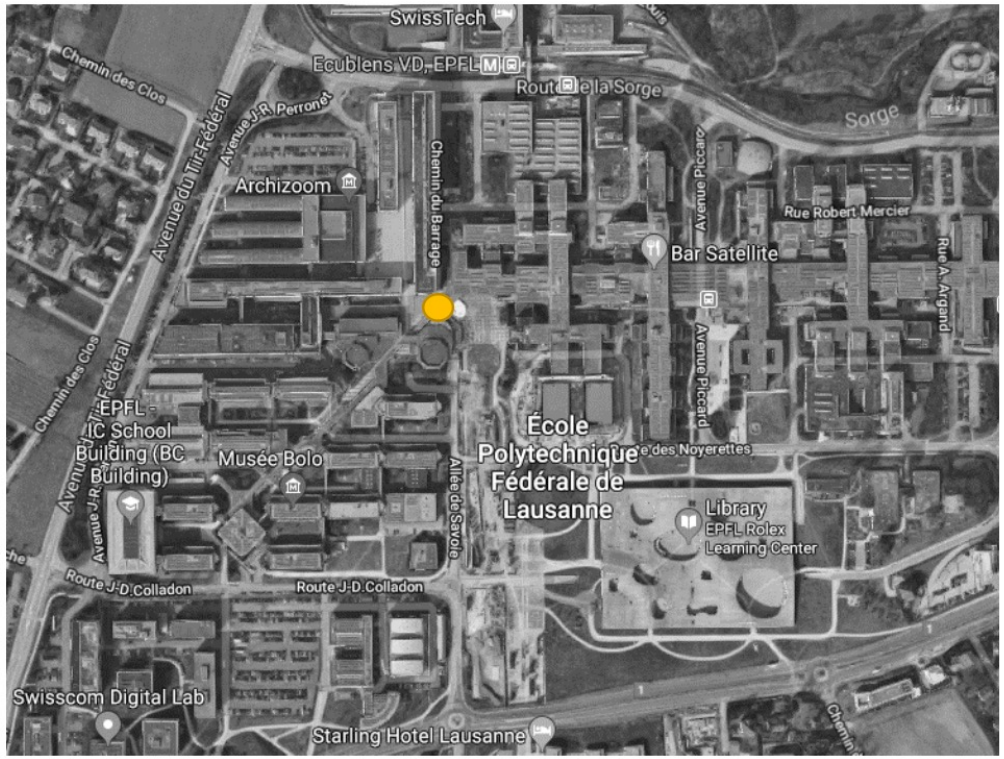


Why not this method??

How to measure privacy?

1) **Strategic adversary (knows defense!)**
   Estimates locations that could have originated the observation: Pr[● | ●]

2) **Privacy error**
   - **Accuracy:** how much variance in estimation
     - Confidence interval
   - **Correctness:** how close to reality
     - Adversary's error [Shokri et al 2011]
   - **Certainty:** how sure of the guess
     - Entropy [Oya et al 2017]



How to measure privacy?

Real location

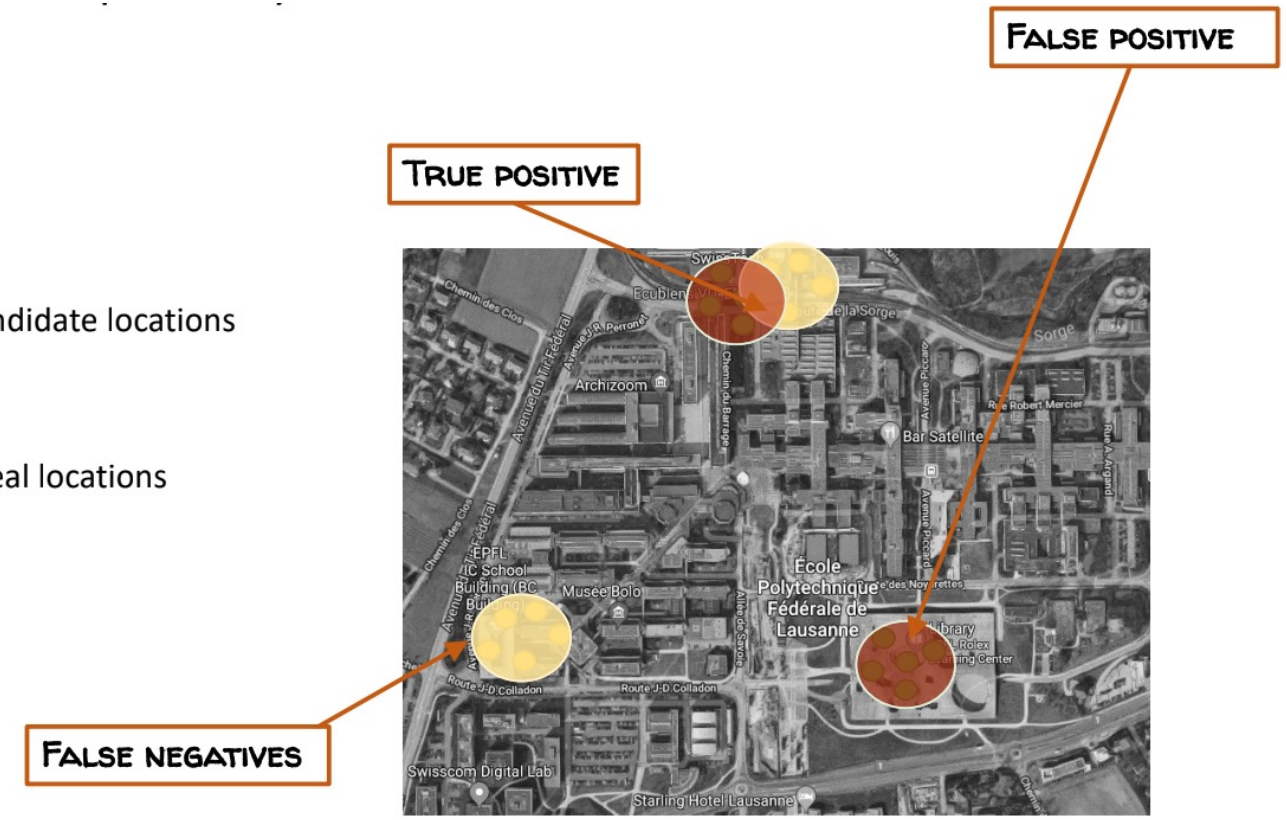
Inferred location

Towards tangible metrics

Privacy is achieved if:

Low precision: many false candidate locations
TP/(TP+FP)

Low recall: not finding many real locations
TP/(TP+FN)



On (The Lack Of) Location Privacy in Crowdsourcing Applications

Authors:
Spyros Boukoros, *TU-Darmstadt;* Mathias Humbert, *Swiss Data Science Center (ETH Zurich, EPFL);* Stefan Katzenbeisser, *TU-Darmstadt, University of Passau;* Carmela Troncoso, *EPFL*

Abstract:
Crowdsourcing enables application developers to benefit from large and diverse datasets at a low cost. Specifically, mobile crowdsourcing (MCS) leverages users' devices as sensors to perform geo-located data collection. The collection of geo-located data though, raises serious privacy concerns for users. Yet, despite the large research body on location privacy-preserving mechanisms (LPPMs), MCS developers implement little to no protection for data collection or publication. To understand this mismatch we study the performance of existing LPPMs on publicly available data from two mobile crowdsourcing projects. Our results show that well-established defenses are either not applicable or offer little protection in the MCS setting. Additionally, they have a much stronger impact on applications' utility than foreseen in the literature. This is because existing LPPMs, designed with location-based services (LBSs) in mind, are optimized for utility functions based on users' locations, while MCS utility functions depend on the values (e.g., measurements) associated with those locations. We finally outline possible research avenues to facilitate the development of new location privacy solutions that fit the needs of MCS so that the increasing number of such applications do not jeopardize their users' privacy.

Mobile Crowdsourcing (MCS)

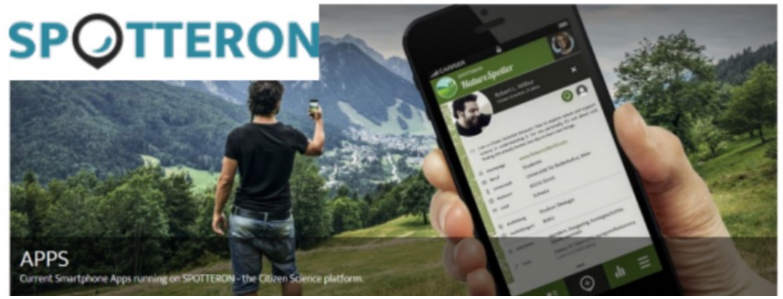
- Rely on collecting data from mobile devices
  - Geo-located measurements

- Democratizes large scale data access: cheap & effective
  - Open-source alternatives to existing services

- Growing ecosystem
  - Millions of data contributors

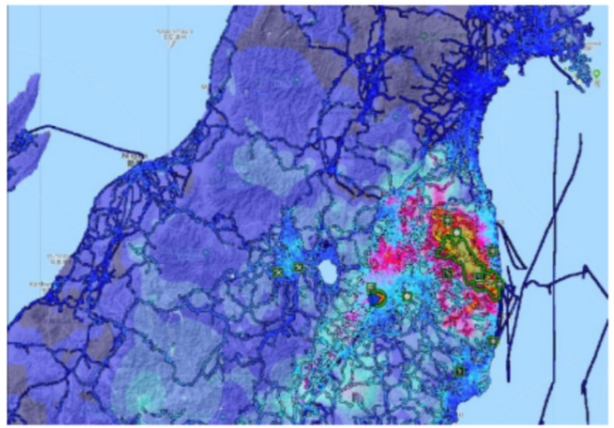
E.g., “Google, Microsoft and Mozilla use crowdsourcing to build WiFi location databases.”

E.g., “OpenStreetMaps, a map generation project from contributed GPS points, reports 4.3 million users in 2018, with 1 million active map editors contributing over 4 billion GPS points.”



Publicly available data from two MCS Projects

Safecast

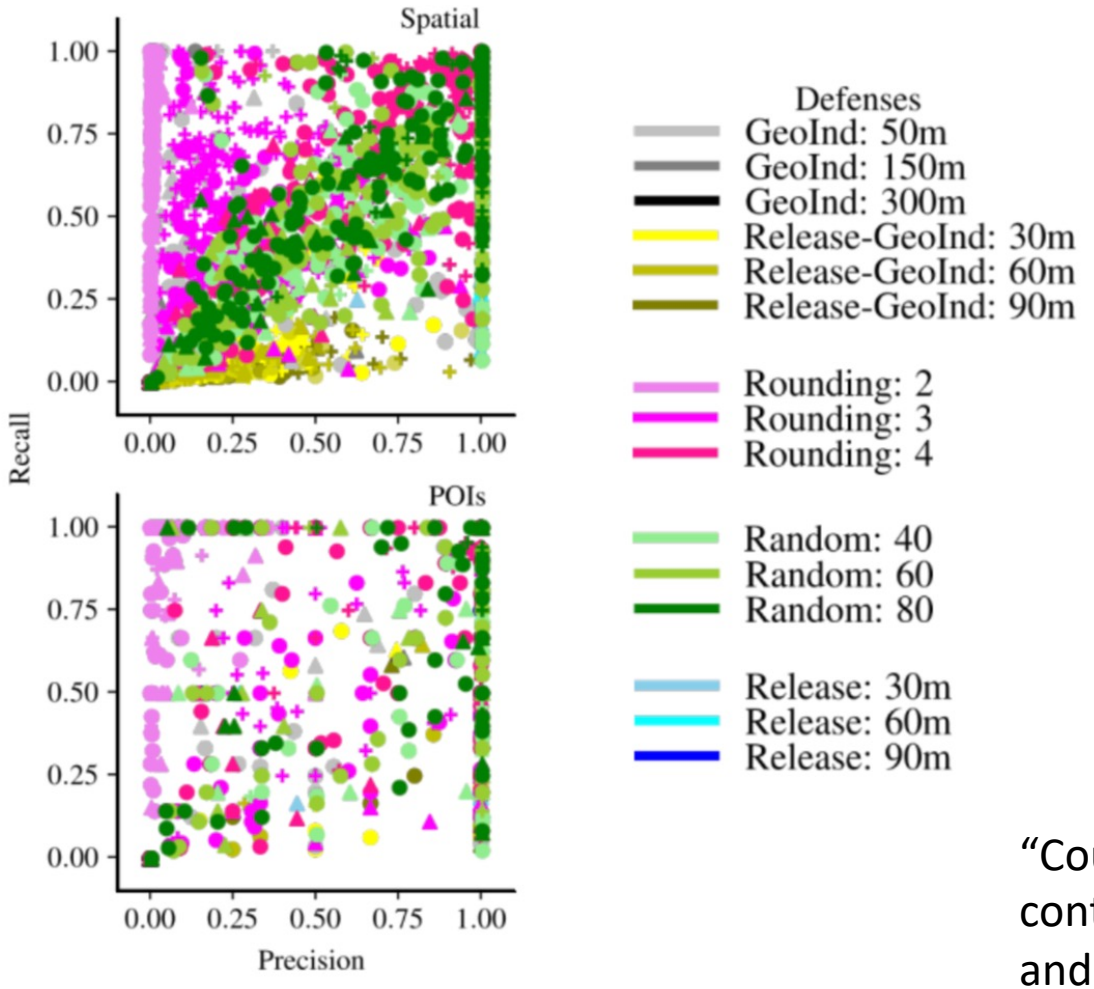


- Users report
  - Location, timestamp, and radiation measurement
- **Goal**: Map radiation levels, finding hotspots
- Identifiable / pseudonymous datasets
  - 56.7M measurements / 540 users

Radiocells

- Users report
  - Location, timestamp, cell information, and device properties
- **Goal**: Map telecomm infrastructure
- “Anonymized” dataset
  - 4 million measurements / 568 users

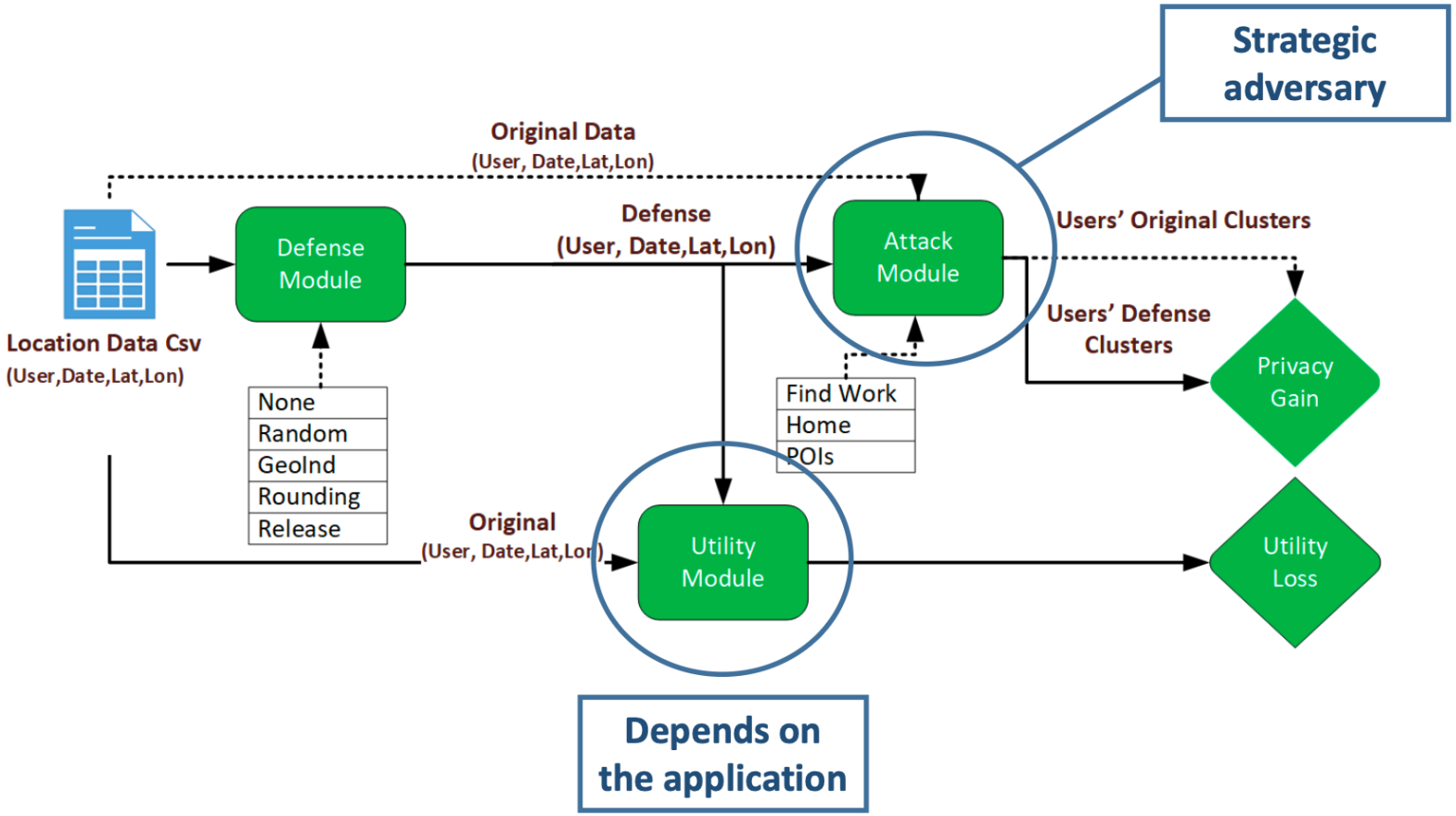
Privacy Gain - Safecast



- Noise addition works for high privacy
  - **Geo-Ind** variants
- Generalization (**Rounding**) ok for medium privacy
- Hiding locations at **Random** does not work, but clever **Release** **(reveal when needed)** helps
- Users with more data were protected better!
  More data => more spread

"Counterintuitively, the LPPMs perform worse for users who contribute fewer points. This is because the attack constructs more, and larger (on average 10 times bigger), clusters for people who share many points than for those sharing fewer points"

How to measure privacy?



Takeaways

Location data contains a lot of information

Simple attacks can reveal many kinds of personal information

More complex attacks or more man power can be destructive

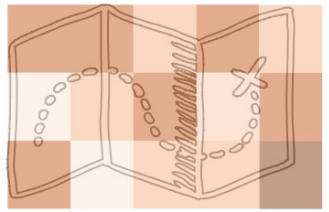
Evaluating privacy risks requires thinking about the defenses: strategic adversaries

State of the art defense mechanisms

have no good privacy/utility frameworks

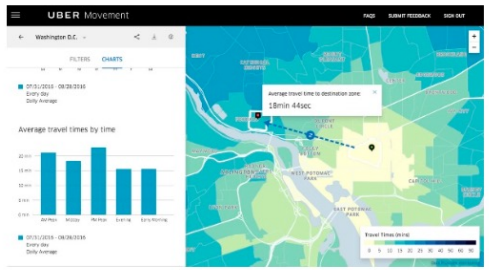
are NOT suitable when utility depends on a function of the location

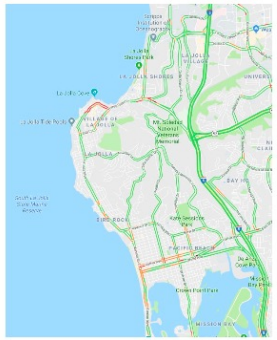
Can we hide in the crowd then?

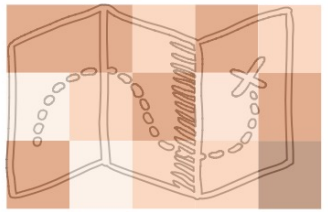


Aggregate Location Data

Can we hide in the crowd then?





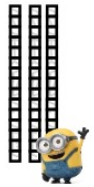


Aggregate Location Data

Even aggregate location is sensitive…

Can we infer if  is in  ?

MEMBERSHIP ATTACKS



ϵ

?

Aggregates might reflect sensitive locations / time

Aggregates might relate to a group of users that share a sensitive characteristic

Regulators can verify possible misuse of the data

Knock Knock, Who's There? Membership Inference on Aggregate Location Data*

Apostolos Pyrgelis
University College London
apostolos.pyrgelis.14@ucl.ac.uk

Carmela Troncoso
IMDEA Software Institute
carmela.troncoso@imdea.org

Emiliano De Cristofaro
University College London
e.decristofaro@ucl.ac.uk

***Abstract*—Aggregate location data is often used to support smart services and applications, e.g., generating live traffic maps or predicting visits to businesses. In this paper, we present the first study on the feasibility of membership inference attacks on aggregate location time-series. We introduce a game-based definition of the adversarial task, and cast it as a classification problem where machine learning can be used to distinguish whether or not a target user is part of the aggregates.**

privacy of the individuals that are part of the aggregates [22, 36]. In this paper, we focus on *membership inference attacks*, whereby an adversary attempts to determine whether or not location data of a target user is part of the aggregates.

Motivation. The ability of an adversary to ascertain the presence of an individual in aggregate location time-series constitutes an obvious privacy threat if the aggregates relate to a group of users that share a sensitive characteristic. For