



# AtlantTIC

Research Center for  
Information & Communication Technologies

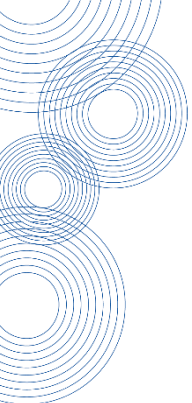
# Location Privacy: Threats and Opportunities

*Fernando Pérez-González*

*Simon Oya*

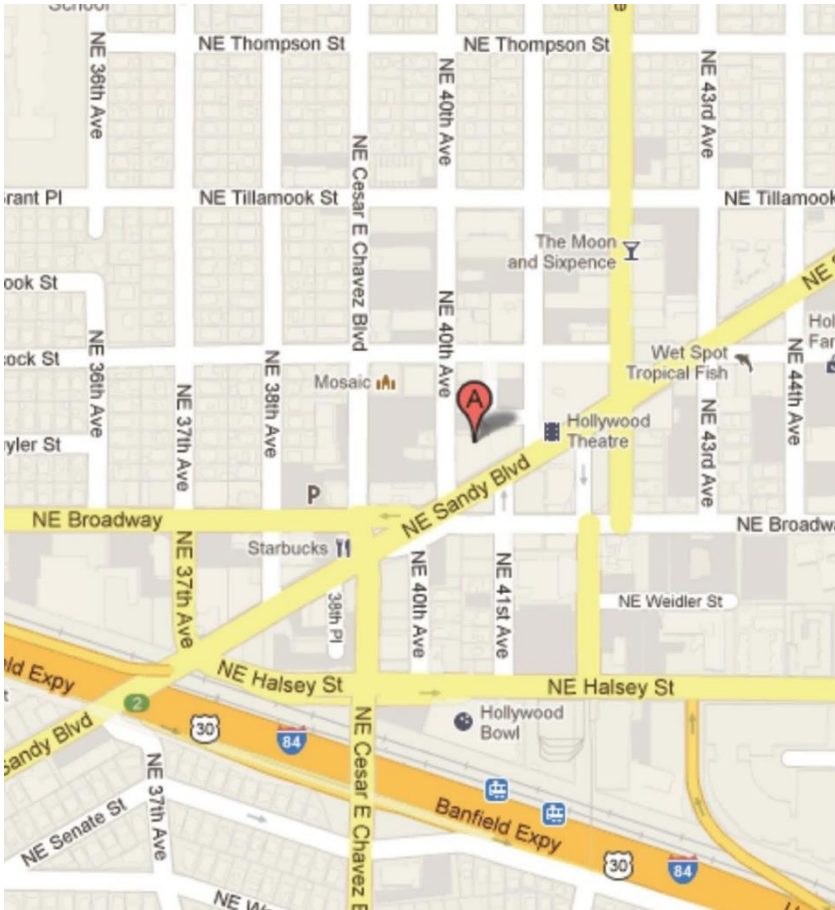
*Signal Theory and  
Communications Department*

*Universidad de Vigo - SPAIN*

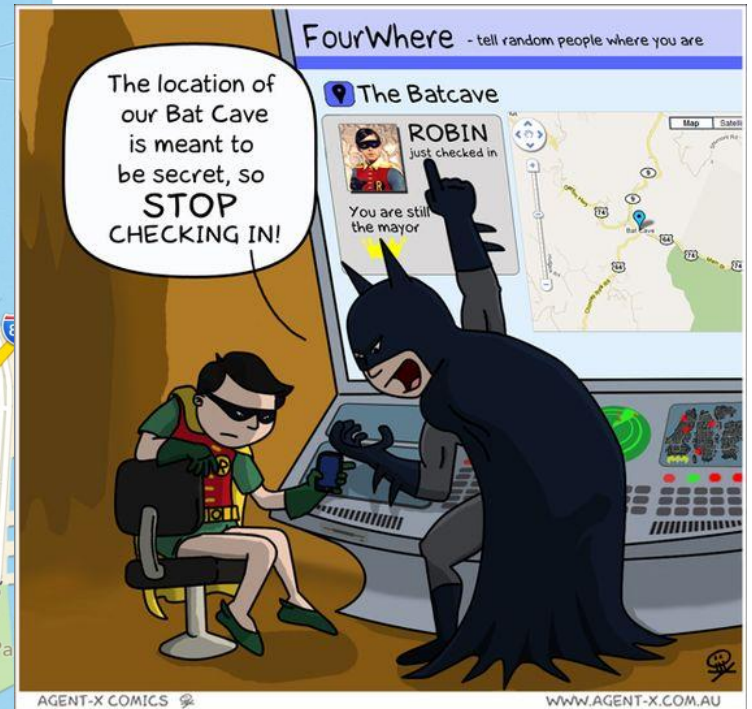
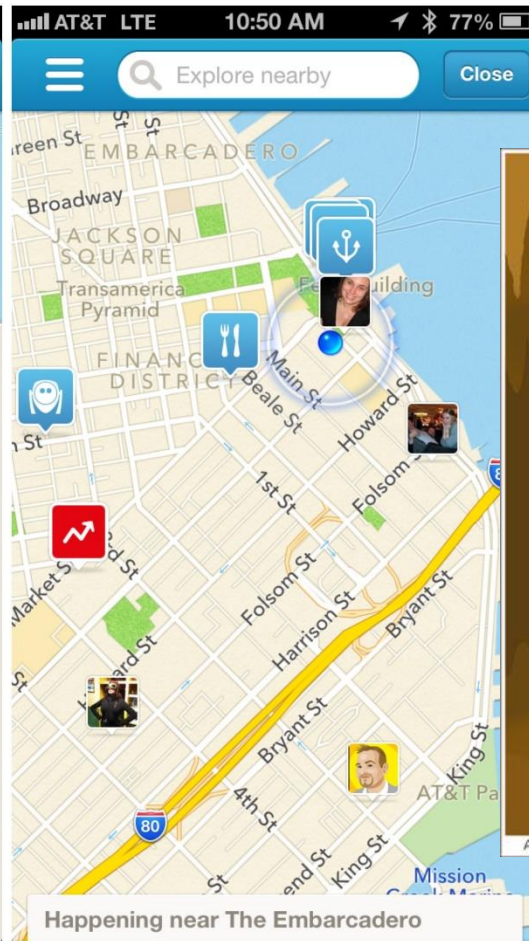
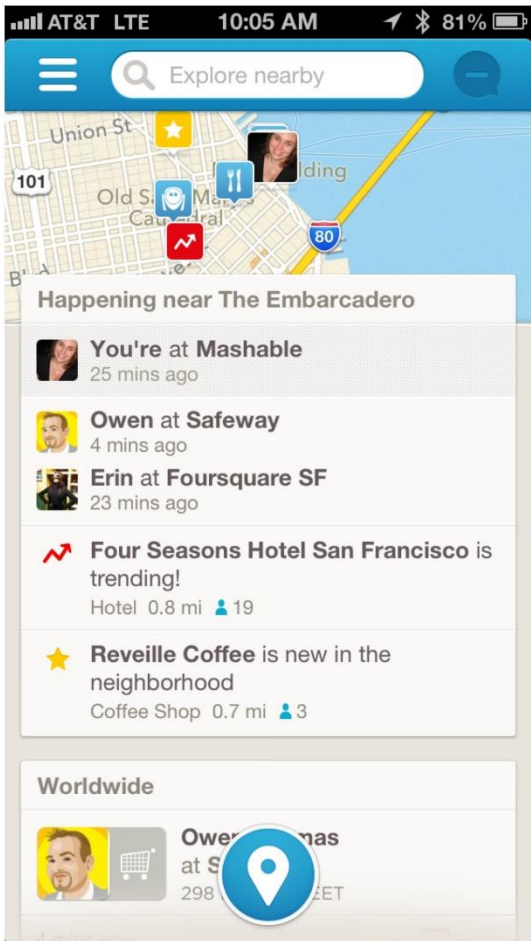


# Why do we like location based apps?

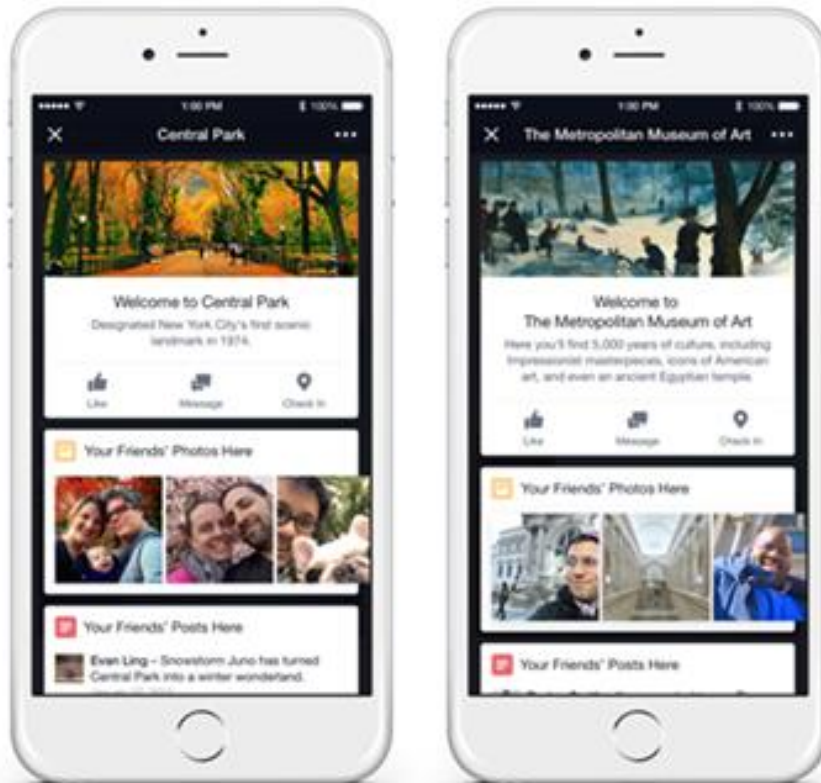
# Google Maps



# Foursquare



# Facebook place tips

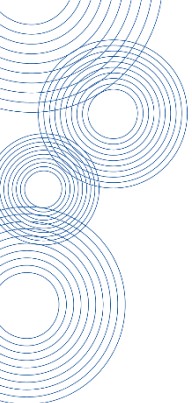


# Waze



And, of course...





# How can you be geolocated? (without you fully knowing)



# IP-based Geolocation

The screenshot displays the GeoIP Tool interface. On the left, a search bar labeled 'Host/IP' contains the IP address 109.73.65.211. Below the search bar, the following geolocation data is listed:

- Nombre Host: 109.73.65.211
- Dirección de IP: 109.73.65.211
- País: United Kingdom
- Código de país: GB (GBR)
- Region: London 1068,GB,I1,"Luton
- Ciudad: London
- Hora local: 31 Oct 10:23 (GMT+0000)
- Código Postal: EC4N
- Latitud: 51.5144
- Longitud: -0.0941

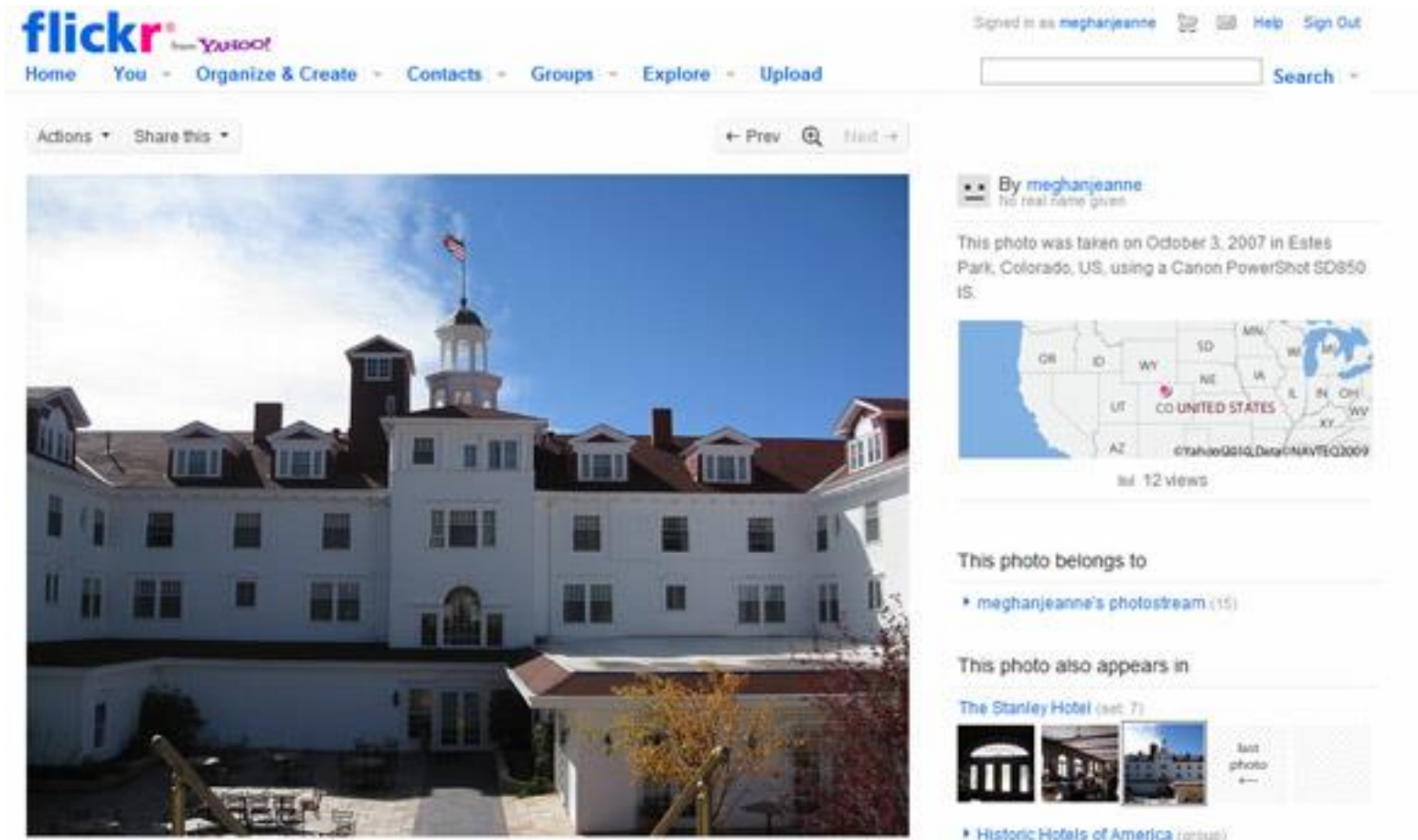
On the right, a Google Map of Europe shows a red pin in London. A tooltip above the pin displays the following information:

- País: United Kingdom
- Ciudad: London
- Dirección de IP: 109.73.65.211

The map also shows various European countries and cities, including Ireland, France, Germany, Poland, and Italy. The Google logo and map data are visible at the bottom of the map area.

Source: GeoIPTool

# Meta-data based Geolocation



The screenshot shows a Flickr page for a photo of a large, white, multi-story building with a central tower and a cupola. The building is identified as The Stanley Hotel. The page includes navigation links, a search bar, and metadata for the photo.

**flickr** by Yahoo!

Home You Organize & Create Contacts Groups Explore Upload

Signed in as meghanjeanne Help Sign Out


Search

Actions Share this

← Prev Next →

**By meghanjeanne**  
no real name given

This photo was taken on October 3, 2007 in Estes Park, Colorado, US, using a Canon PowerShot SD850 IS.

  
©Yahoo!G014,Data©NAVTEQ2009




12 views

This photo belongs to

- meghanjeanne's photostream (15)

This photo also appears in

**The Stanley Hotel** (set: 7)

   [last photo](#)

- Historic Hotels of America (group)

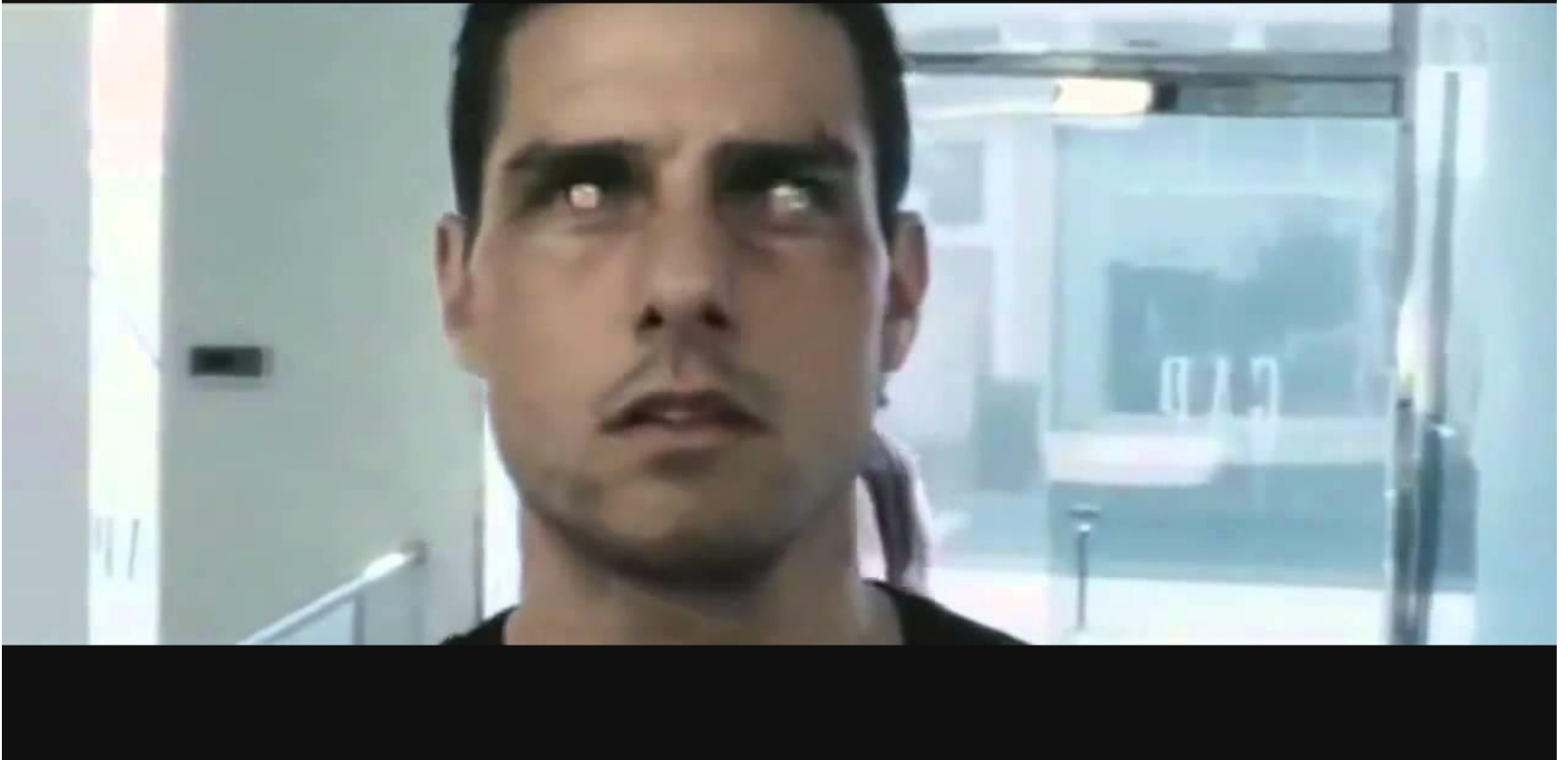
# Landmark recognition Geolocation

A screenshot of a Google search interface. The address bar shows the URL: [https://www.google.co.uk/search?tbs=sbi:AMhZZisxKkIsZBRX0vH7f2NXW3i7CJafjBPl9p\\_1cmwBd0Qu2iKy6L](https://www.google.co.uk/search?tbs=sbi:AMhZZisxKkIsZBRX0vH7f2NXW3i7CJafjBPl9p_1cmwBd0Qu2iKy6L). The search bar contains the text 'statue of five goats'. Below the search bar, the 'Images' tab is selected. The search results show 'About 2 results (0.56 seconds)'. A single image result is displayed, showing a smaller version of the 'Statue of Five Goats' sculpture. To the right of the image, the text reads: 'Image size: 4320 x 3240' and 'No other sizes of this image found.' Below the image, it says 'Best guess for this image: **statue of five goats**'. At the bottom, there is a link: 'Legend of 5 goats | - Guangzhou.chn.info' with the URL [www.guangzhou.chn.info/overview/legend-of-5goats.html](http://www.guangzhou.chn.info/overview/legend-of-5goats.html). Below the link, it says: 'There are many goat statues in Guangzhou and the **Statue of the Five Goats** is the most impressive, and now the one which were built in Yuexiu Park in 1959 ...'.

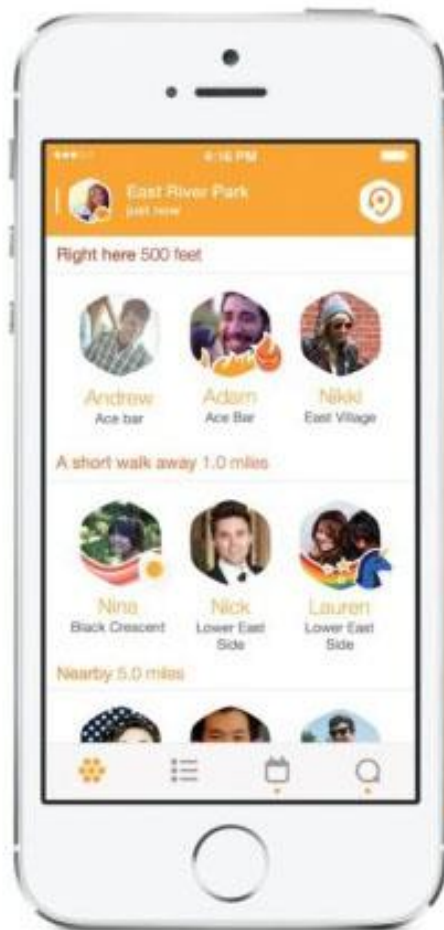
A screenshot of a tweet from the user 'ellouis' (@ellouis). The tweet text reads: 'Just discovered that even though geotagging is disabled in my camera, Google Photos automatically adds location when it recognizes a landmark. Very creepy.' The tweet has 313 likes and was posted at 5:24 PM on August 6, 2018. At the bottom of the tweet, it says '190 people are talking about this'. The tweet is highlighted with a blue glow.



# Biometric geolocation



# Apps-based geolocation



# Credit card usage Geolocation



MailOnline



FREE - On the Microsoft Store

## Mastercard under fire for tracking customer credit card purchases to sell to advertisers

- Credit card firm refuses to reveal 'proprietary' technique that allows it to anonymously track customers and target them with online ads
- Privacy campaigners accuse firm of 'treating details of our personal behaviour like their own property'
- System tracks information about the date, time, amount and merchant
- Credit card firm says system is only operational in US

By [MARK PRIGG](#)

**PUBLISHED:** 15:52, 17 October 2012 | **UPDATED:** 17:36, 17 October 2012



View comments

Mastercard has come under fire for tracking its US customer's purchases and selling the data to advertisers.

The credit card company's MasterCard Advisors Media Solutions Group boasts it can target the most affluent customers and tell advertisers who is most likely to buy their products.

The firm does this by tracking a consumer's credit card details - although it says their identity

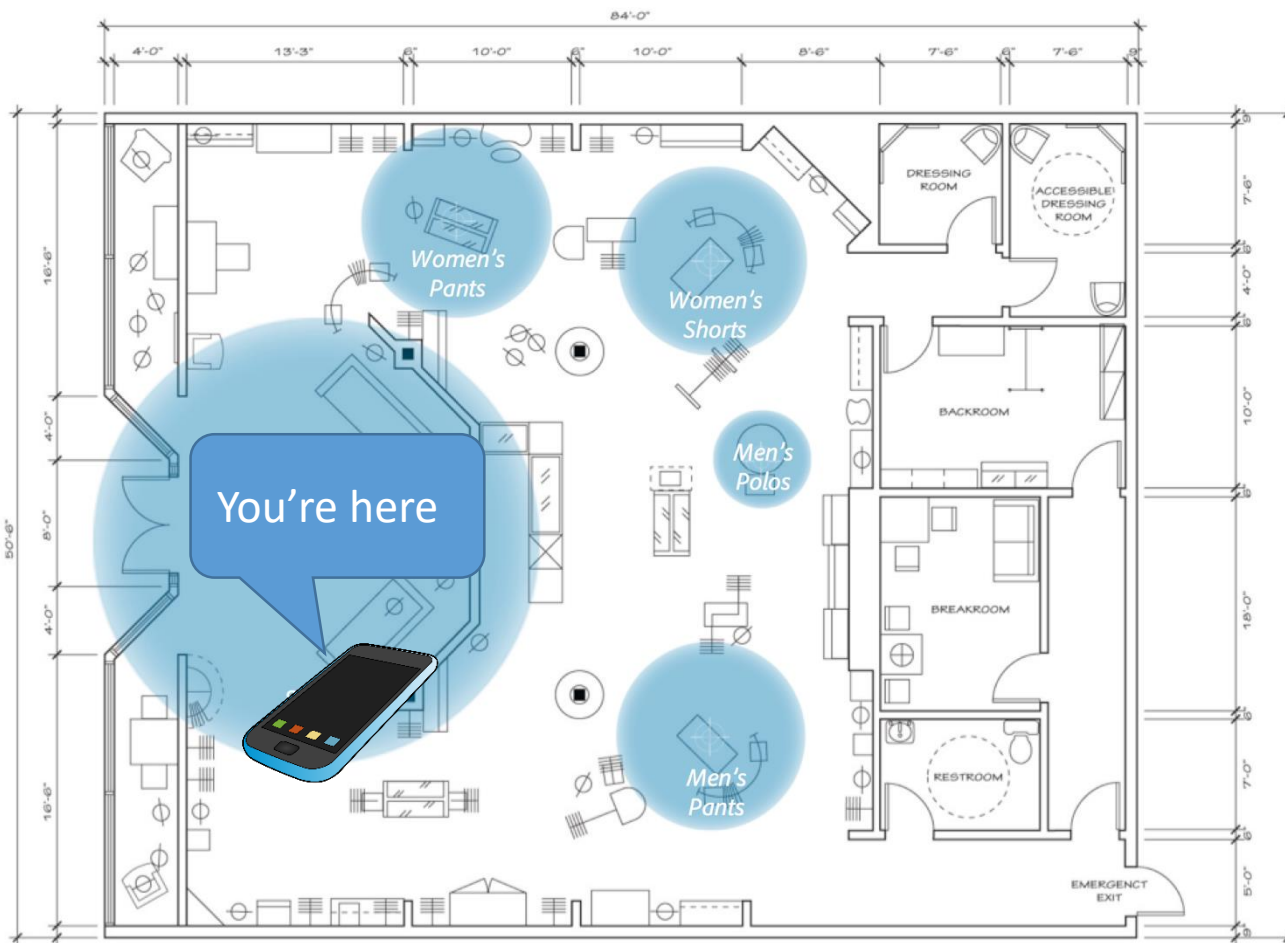


# Radio-based localization



# Proximity-based localization

Source: theblog.adobe.com



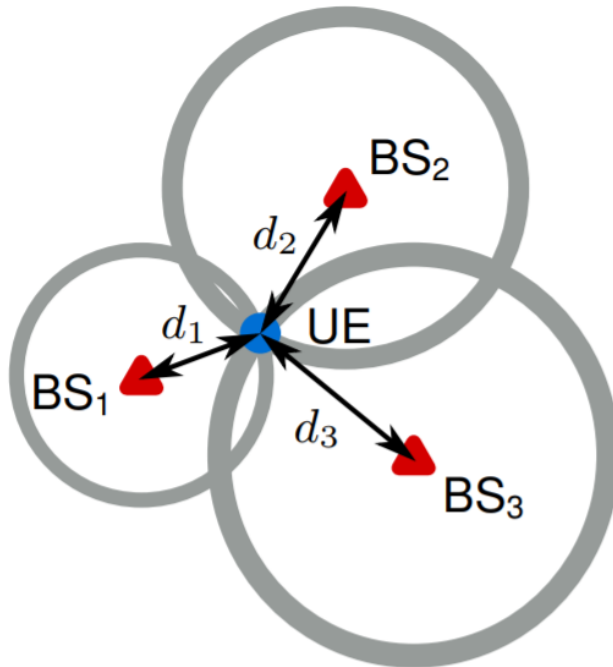
## Store #45

- Store Entry (beacon #1)
  - Women's Pants (beacon #2)
  - Women's Shorts (beacon #3)
  - Men's Polos (beacon #4)
  - Men's Pants (beacon #5)
- Total Beacons: 5

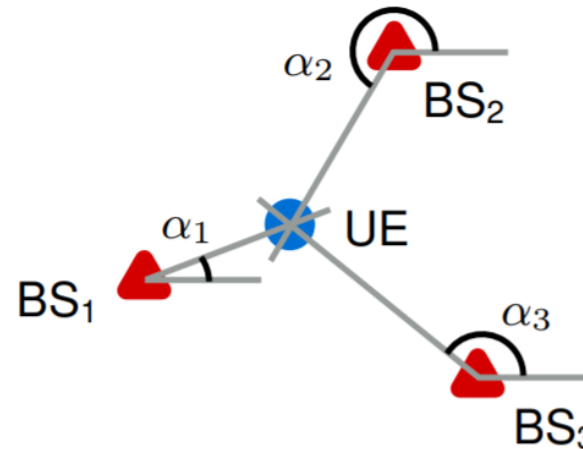


# Trilateration vs. triangulation

## Trilateration

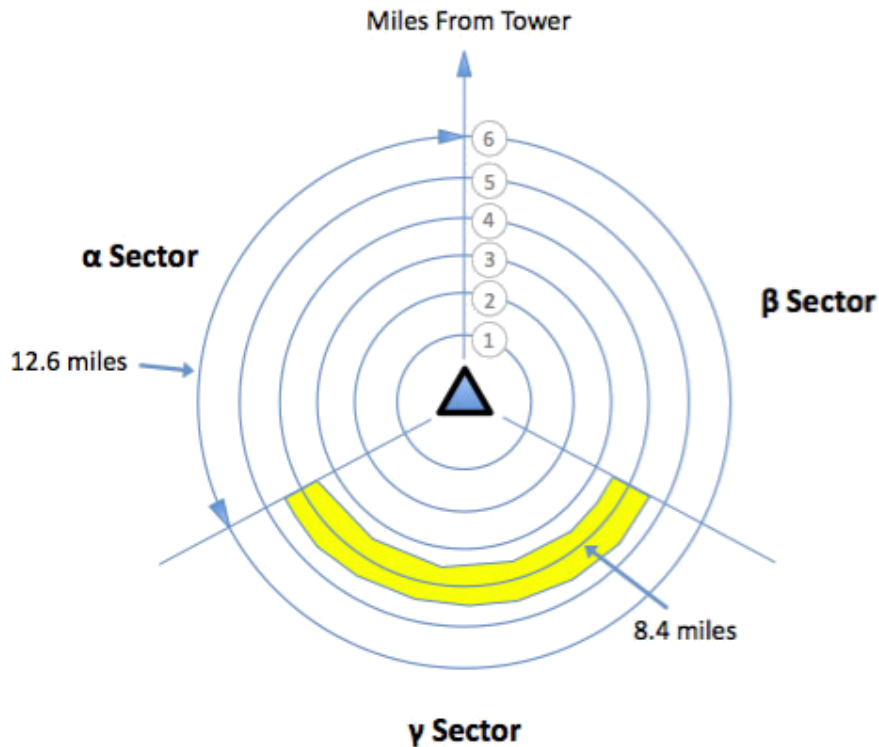


## Triangulation



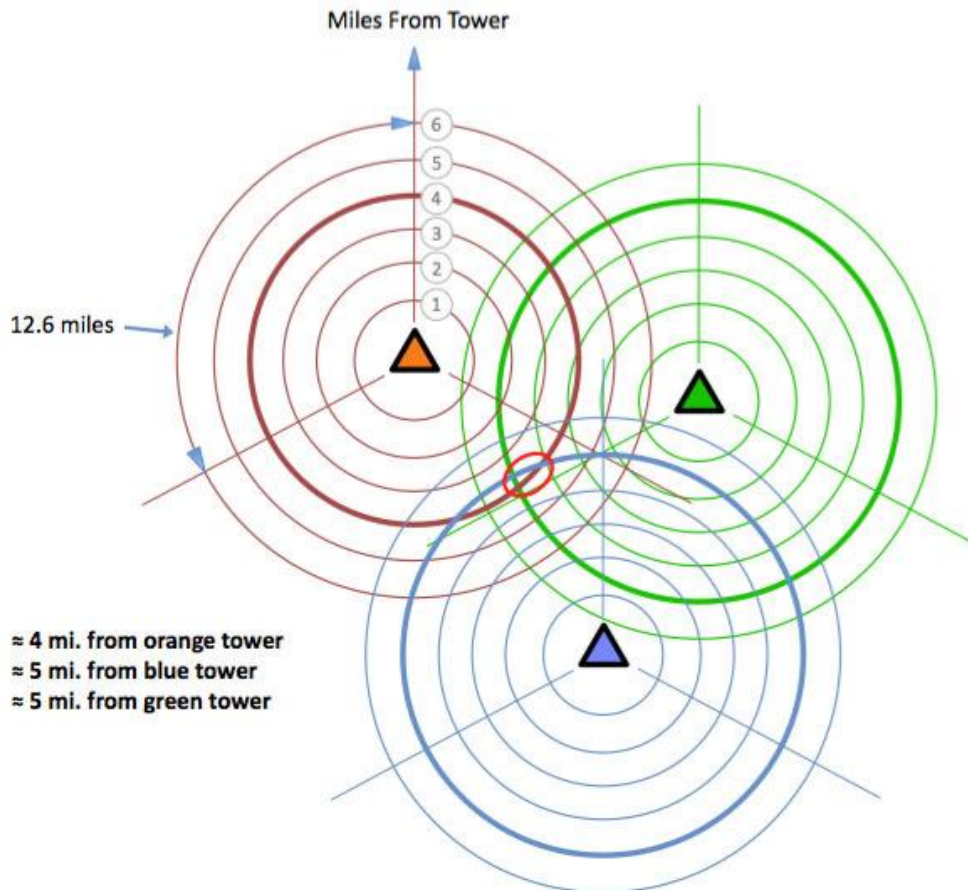
Source: J.A. del Peral et al. "Survey of Cellular Mobile Radio Localization Methods: from 1G to 5G", IEEE Communications Surveys & Tutorials, 2017.

# Signal strength-based triangulation/trilateration



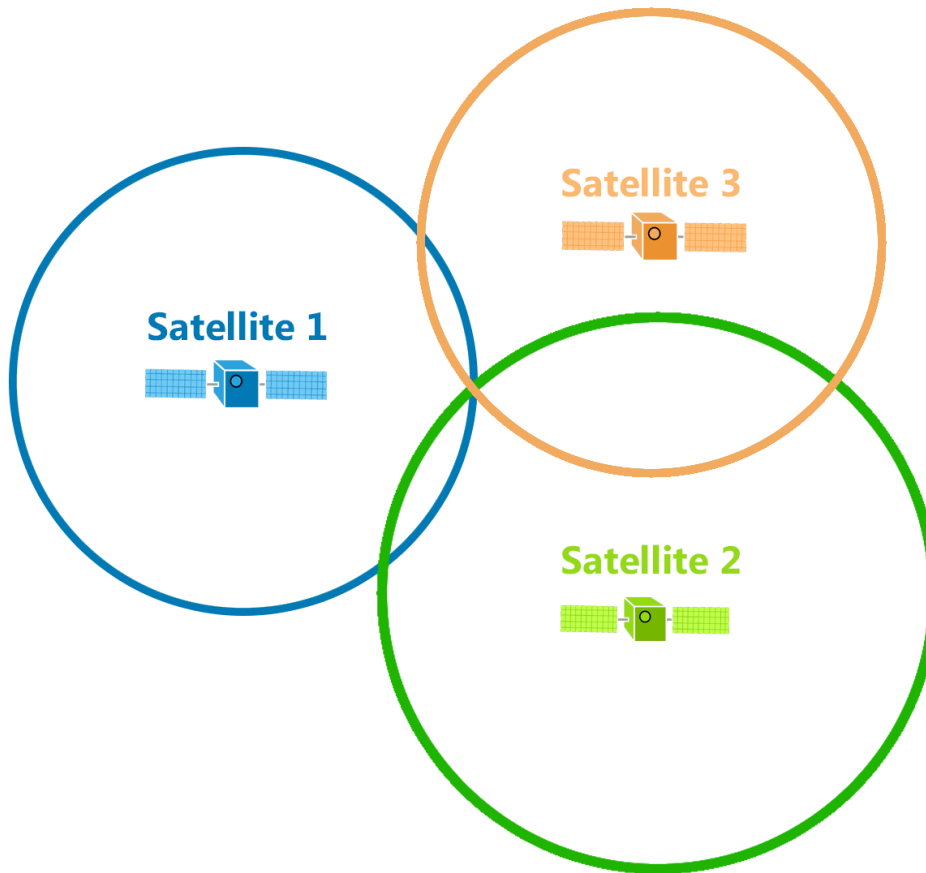
Source: The Wrongful Convictions Blog

# Signal strength-based triangulation/trilateration



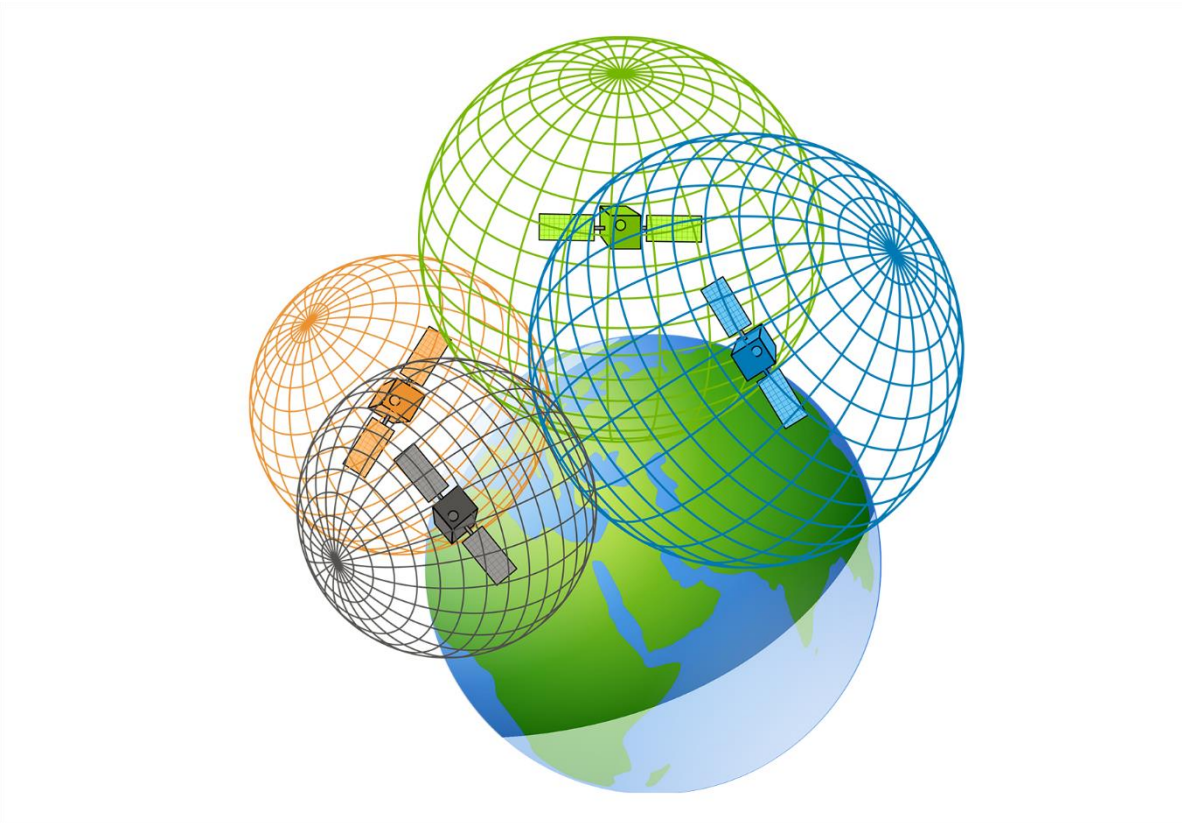
Source: The Wrongful Convictions Blog

# Time of Arrival (ToA) based trilateration



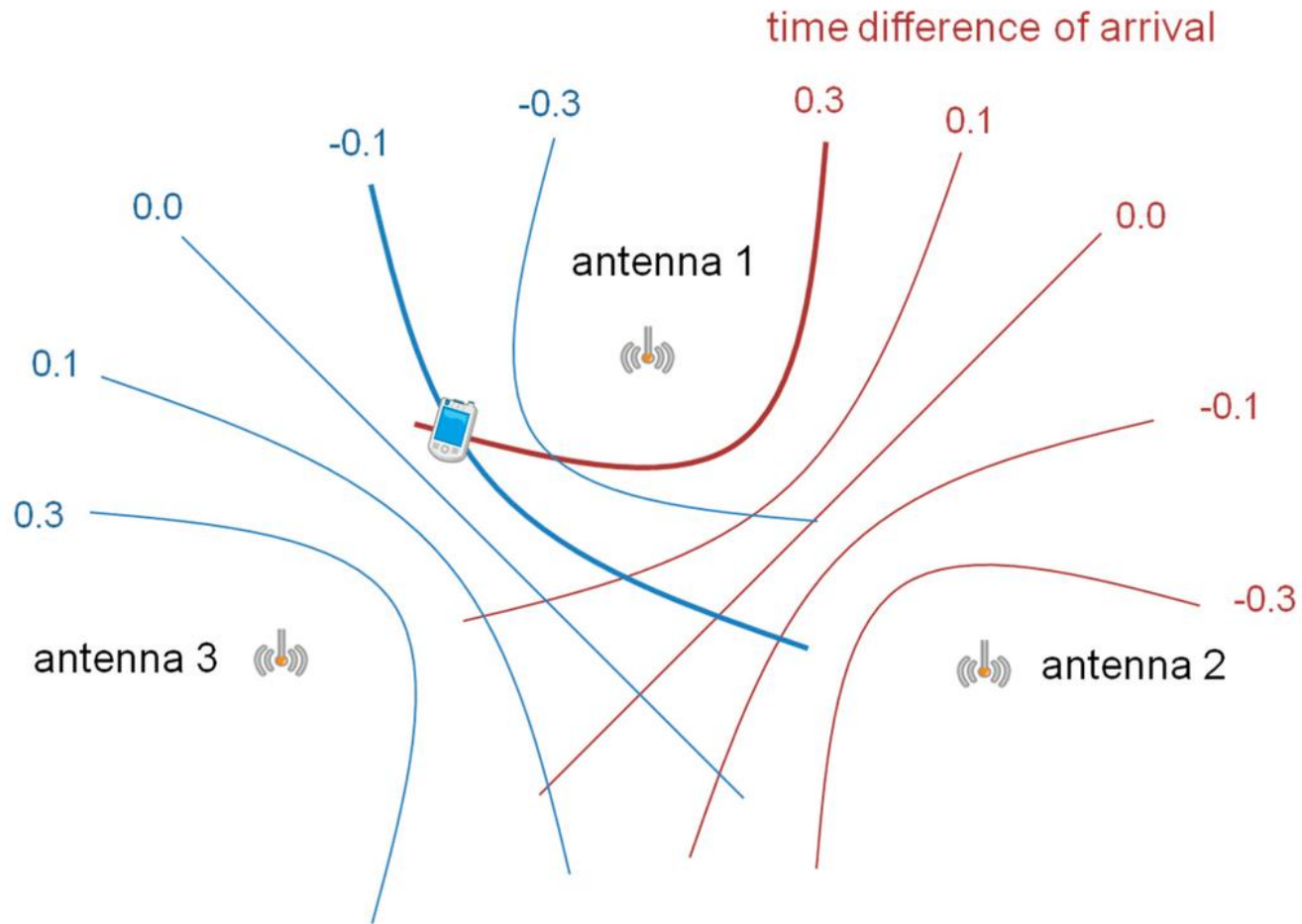
Source:GISGeography.com

# GPS-based trilateration



Source:GISGeography.com

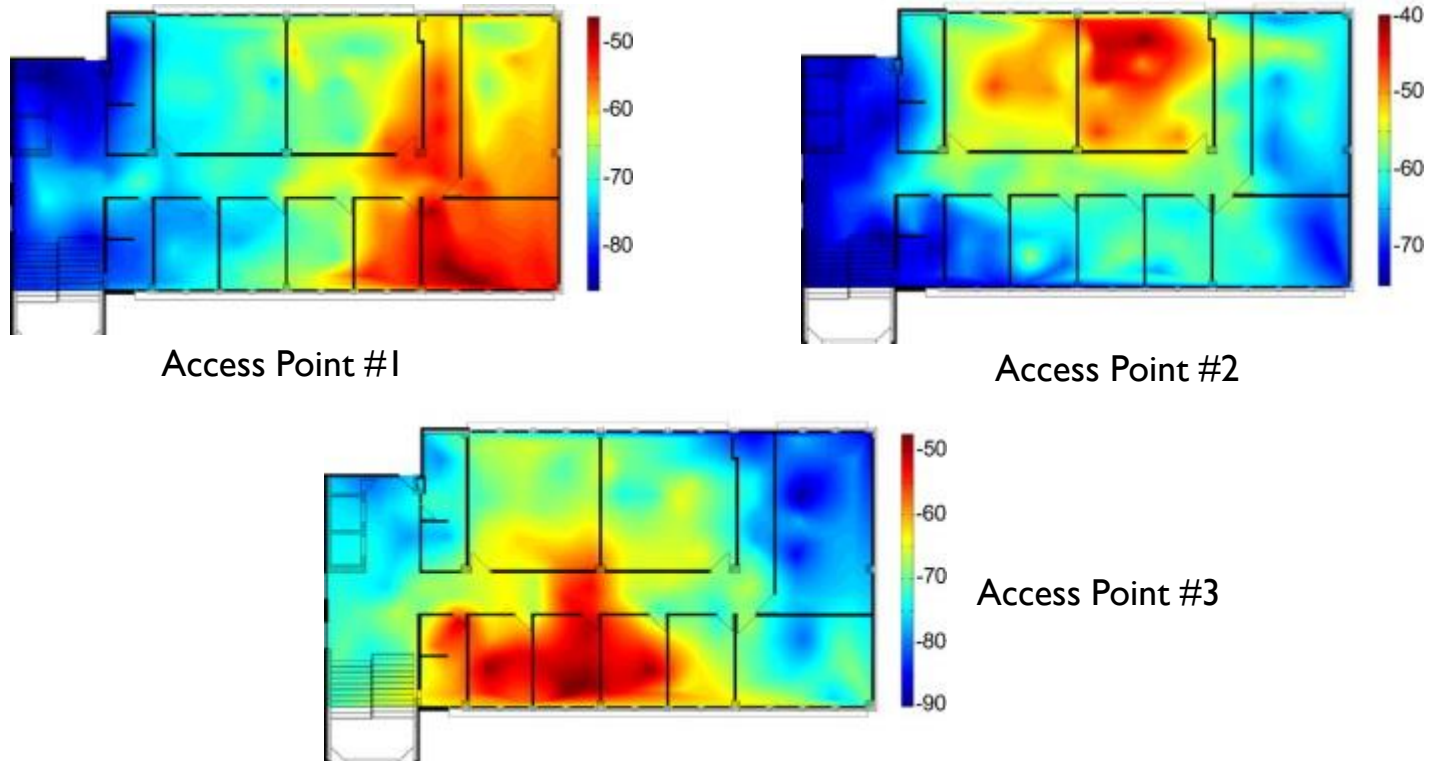
# Multilateration: Time Difference of Arrival (TDOA)



Source:[Fujii et al. 2015]

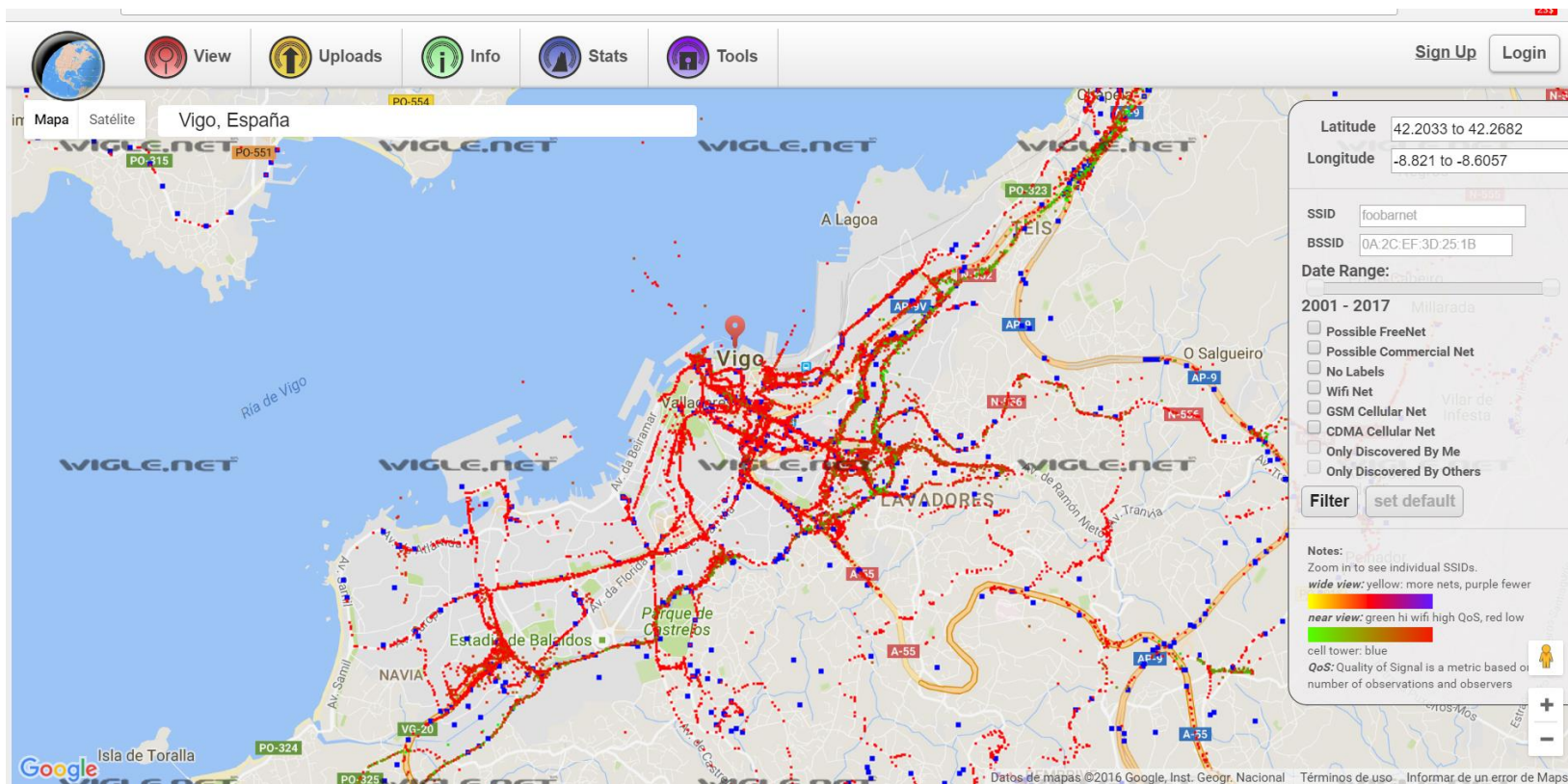
# Fingerprint-based localization

## Signal Strengths



Source: M. Stella, M. Russo, D. Begusic, "Fingerprinting based localization in heterogeneous wireless networks", , Expert Systems with Applications Journal, 2014

# Wardriving geolocation (Wigle)



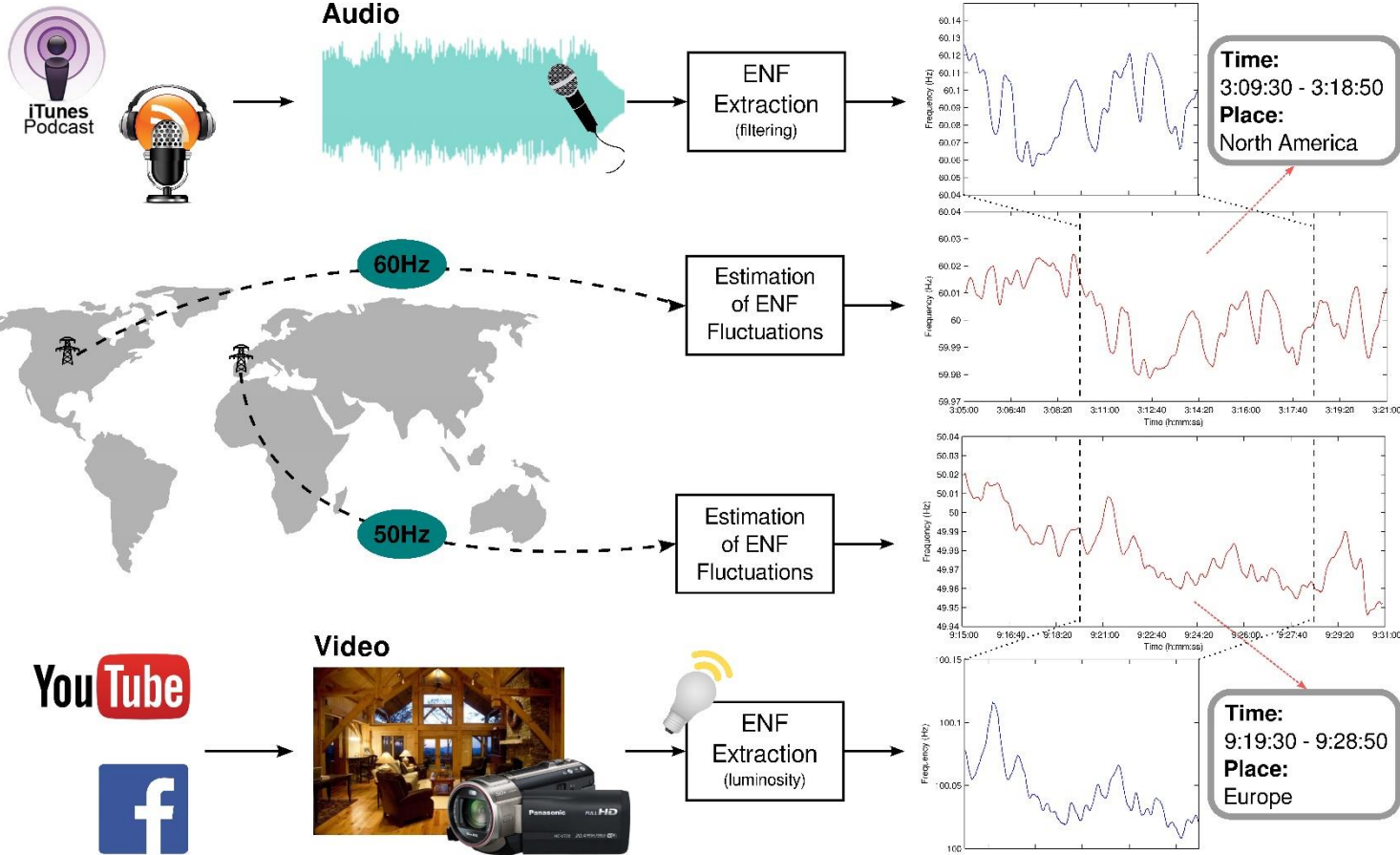
Source:Wigle.net



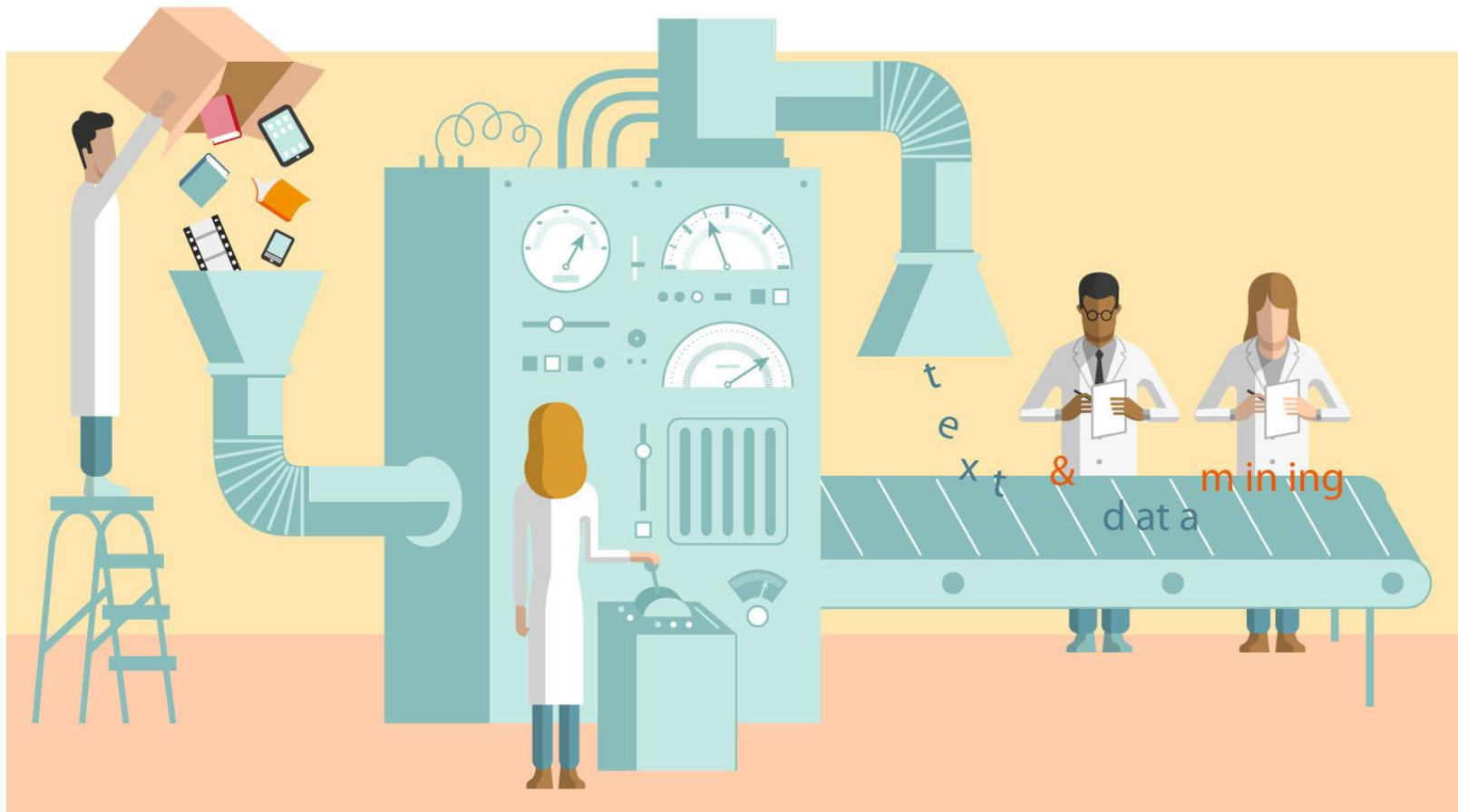
# Geolocation malware

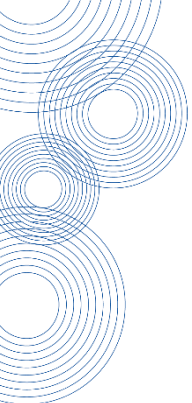


# Electrical Network Frequency Geolocation

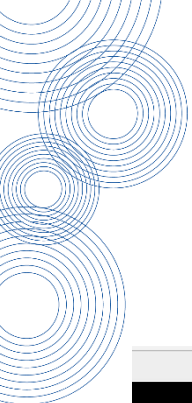


And, of course, combinations of all the above...





# Why is it dangerous?



DOW JONES, A NEWS CORP COMPANY ▾

DJIA Futures ▲ 18116 0.10%   Stoxx 600 ▼ 339.49 -0.38%   U.S. 10 Yr ▲ 2/32 Yield 1.842%   Crude Oil ▼ 48.49 -0.43%   Euro ▼ 1.0956 -0.28%

# THE WALL STREET JOURNAL.

Subscribe Now | Sign In

**SPECIAL OFFER: JOIN NOW**

Home World U.S. Politics Economy Business **Tech** Markets Opinion Arts Life Real Estate



**Dyn Says Cyberattack Has Ended, Investigation Continues**



**Visa Taps Blockchain for Cross-Border Payment Plan**



**Airbnb Revises New York Rules Amid Possible Legislation**



WHAT THEY KNOW

## Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI  
December 24, 2012

It was the same Swingline stapler, on the same [Staples.com](#) website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

Staples seemed to think they were located

### Most Popular Videos

1. **KLM to Make Final Dramatic Landing With 747**



2. **Bottle Flipping Hits a Wall**



# Buster busted!

twitpic

Click here to login or create an account | Sign in with Twitter

Posted on February 9, 2010 by dontrythis

DRIVERS WHO SWITCHED SAVED \$348 A YEAR ON AVERAGE

Allstate QUOTE NOW

More photos by dontrythis

Now it's off to work in my beast. Wait... How'd that DOG get in there?

Login to leave a comment

47 Comments 1 2 3 Next

mableitsml 150 days ago  
This car is missing some explosive devices :)

fuzzy10498 158 days ago  
the mystery of the pooling dog

sewdotcoe 164 days ago  
now THAT is a truly awesome machine.

finkdawg5 173 days ago

Share this photo  
Put this photo on your website

Views 21,833  
Events  
Tags





# PLEASE ROB ME



## Raising awareness about over-sharing

Check out our [guest blog post](#) on the CDT website.

Like Share 32K people like this. Be the first of your friends.

Check your own **Twitter timeline** for checkins

Are you curious if people can see your checkins?  
Enter your Twitter username and find out.

Find!



love ↗

### More Info

[Home](#)

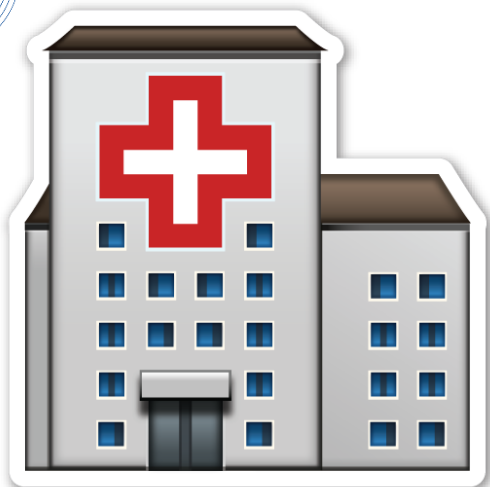
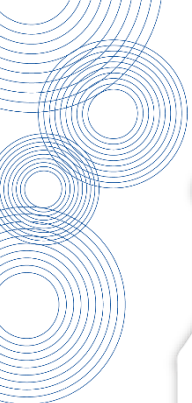
[Why](#)

### Made Possible By

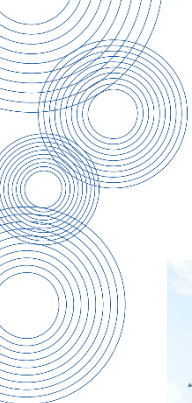
[Foursquare](#)

[Twitter](#)

[@boyvanamstel](#)







# Rogue employees

Facebook Engineer Accused of Stalking Women Online Using ...

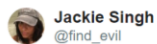
TECH • FACEBOOK

## Facebook Engineer Accused of Stalking Women Online Using Company Data

Facebook's data security troubles certainly don't end with the Cambridge Analytica scandal. The company is now investigating a claim that an engineer used access to Facebook's data to stalk women online.

The social media giant told *Motherboard* that there are "strict technical controls and policies to restrict employee access to user data," emphasizing that Facebook employees can only access the data they need to conduct their jobs.

The allegations against the Facebook engineer surfaced Sunday night in a tweet from Jackie Stokes, the founder of *Spyglass Security*.



Jackie Singh  
@find\_evil

I've been made aware that a security engineer currently employed at Facebook is likely using privileged access to stalk women online.

I have Tinder logs. What should I do with this information?

593 4:50 AM - Apr 30, 2018

SFGATE LOCAL NEWS SPORTS REAL ESTATE BUSINESS A&E FOOD LIVING TRAVEL OBITUARIES

## Google worker arrested for cyberstalking

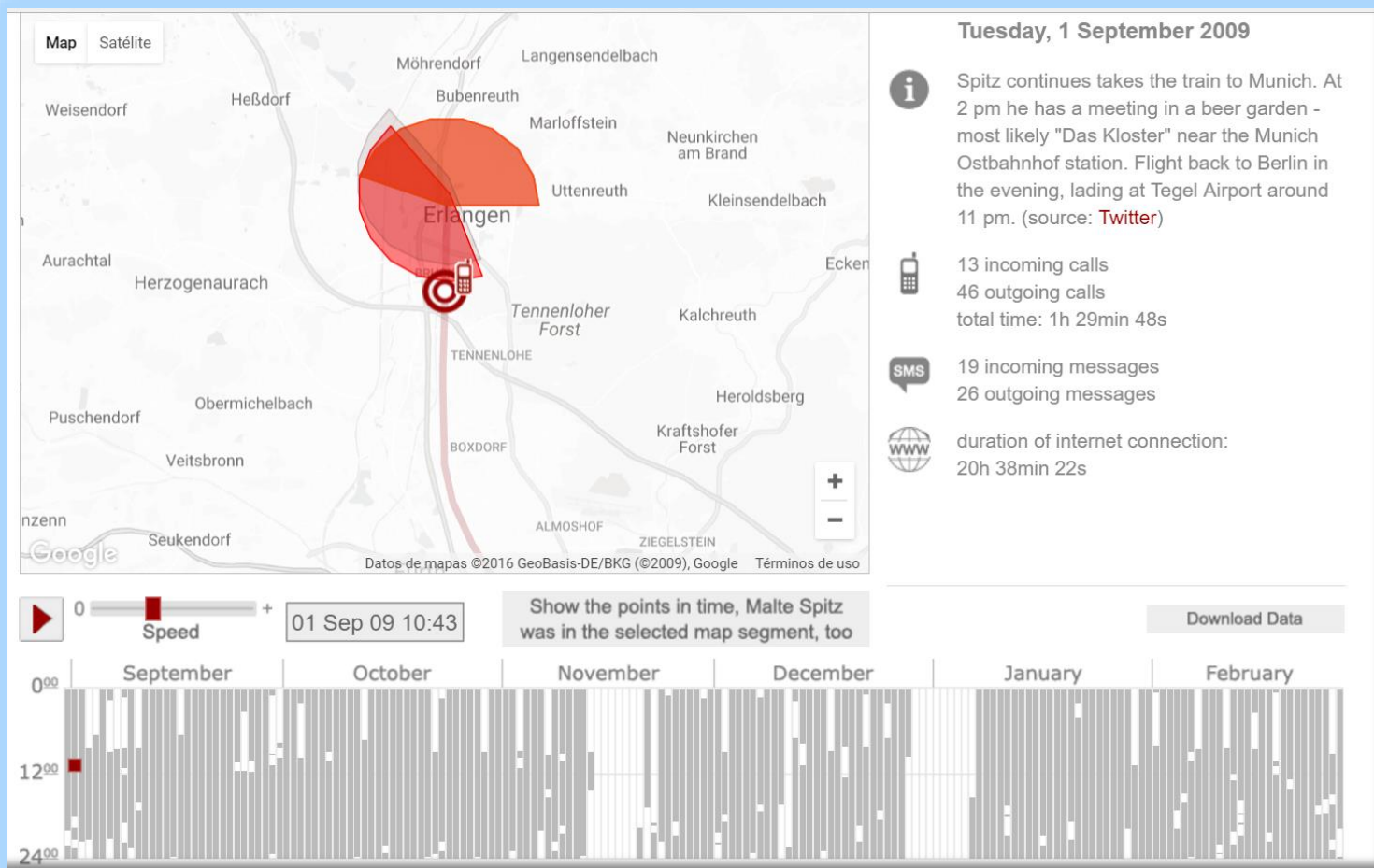
By Erin Allday Updated 7:14 pm PDT, Saturday, October 25, 2014

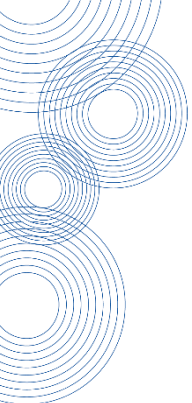
A Google employee from San Jose is facing federal charges in connection with the alleged cyberstalking of a former college classmate and a threat to reveal naked pictures of her if she didn't send him more explicit photos and videos.

According to documents posted Friday on **The Smoking Gun** website, Nicholas Rotundo, 23, was arrested Oct. 4 after an investigation by the FBI and the University of Texas at Dallas. According to the **FBI documents**, Rotundo was an employee of Google in Mountain View and living in San Jose during the 15 months when the online harassment allegedly took place. It's not known whether he is still employed by Google.

The alleged stalking began in June 2013 when a woman, identified in the FBI documents as a University of Texas at Dallas student, received an e-mail inviting her to join a research study on "the public's perception of different

# 6 months in the life of Malte Spitz (2009-2010)



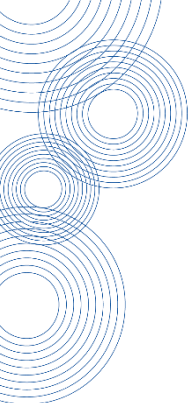


# Are we concerned about it?

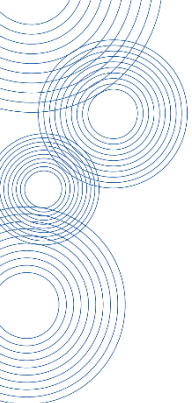
# Are people really concerned about location privacy?

- Survey by Skyhook Wireless (July 2015) of 1,000 Smartphone app users.
- 40% hesitate or don't share location with apps.
- 20% turned off location for all their apps.
- Why people don't share location?
  - 50% privacy concerns.
  - 23% don't see value in location data.
  - 19% say it drains their battery.
- Why people turn off location?
  - 63% battery draining.
  - 45% privacy.
  - 20% avoid advertising.





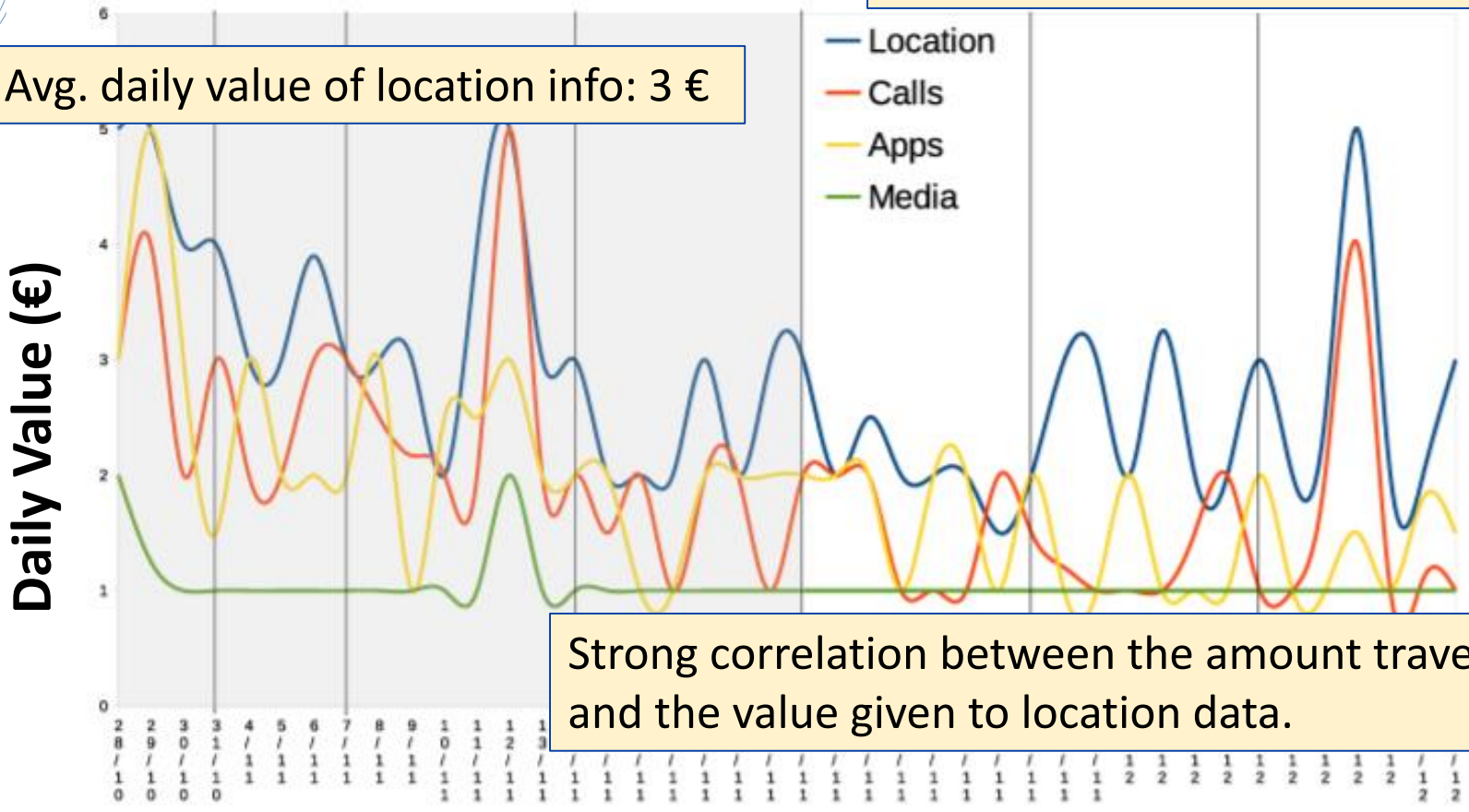
# How much is geolocation data worth?



# How much value do we give to location data? [Staiano et al. 2014\*]

Many participants opted-out of revealing geolocation information.

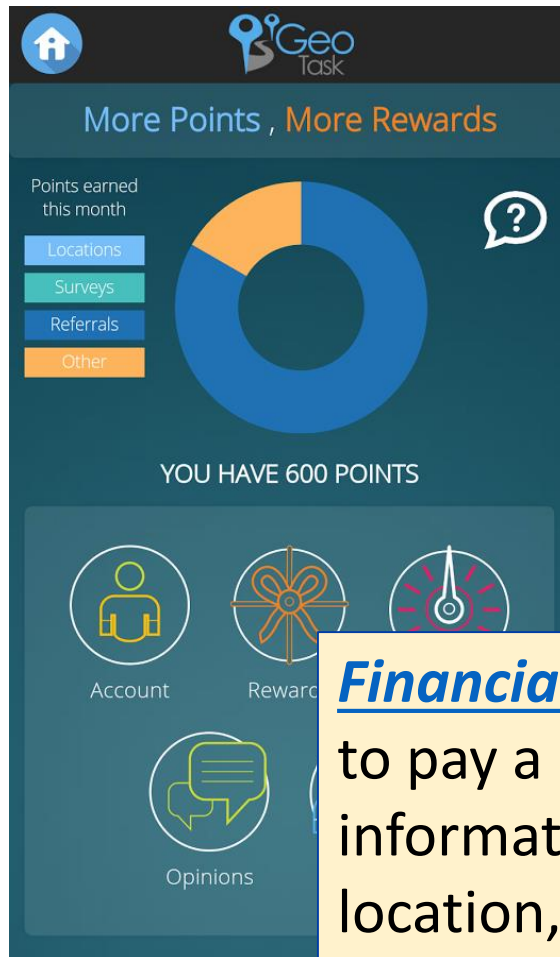
Avg. daily value of location info: 3 €



\* J. Staiano et al. "Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data". ArXiv 2014

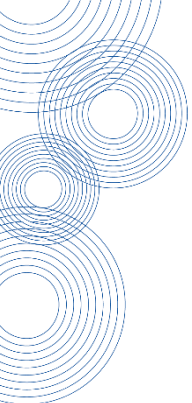


# Earn money as you share data



- GeoTask
- £1 PayPal cash voucher per 100 days of location data sharing (£0.01/day)

**Financial Times** in 2013: advertisers are willing to pay a mere \$0.0005 per person for general information such as their age, gender and location, or \$0.50 per 1,000 people.



Securing Your Journey  
to the Cloud

[Buy Online](#) | [Downloads](#) | [Partners](#) | [United States](#) | [About Us](#) | [Log In](#)

[For Home](#) | [For Business](#) | [Security Intelligence](#) | [Why Trend Micro](#) | [Support](#) |

[Home](#) > [Security Intelligence](#) > [Security News](#) > [Internet of Things](#) > [How Much is Your Personal Data Worth? Survey Says...](#)

Physical location information is sixth at US\$16.10. US citizens priced it at US\$38.40 while consumers in Japan and Europe priced it a paltry US\$4.80 and US\$5.10 respectively.

Glossary  
Research & Analysis

In this day and age where privacy, security and the lack of both (which

Home address is seventh at US\$12.90, with US consumers once more pricing it at US\$17.90. Japanese respondents pegged this information at US\$16.30 while those in Europe priced it at US\$5.00.

# Pay as you drive



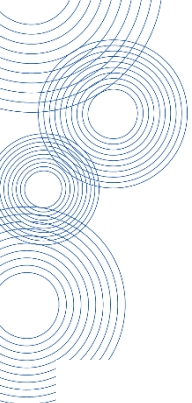
Home / Car Insurance / Pay As You Drive Insurance



## Pay As You Drive Insurance

If you want the security of Comprehensive car insurance but you only drive a little, then

- Formula can be a function of the amount of miles driven, or the type of driving, age of the driver, type of roads used...
- Up to 40% reduction in the cost of insurance.



## Location-Targeted Mobile Ad Spend to Reach \$29.5B in the U.S. in 2020

🕒 June 16, 2016 👤 📁 Press Releases

**That's \$90 per person year!!!**

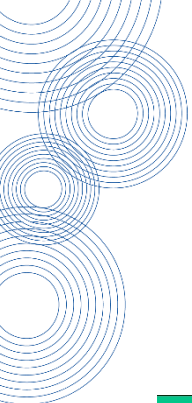
*Native social advertising will represent more than one-quarter (28.1%) of U.S. local-targeted ad spend by 2020, pulling market share from search and display.*

CHANTILLY, Va. (June 16, 2016) – In the spring update to its U.S. Local Advertising Forecast 2016, BIA/Kelsey projects location-targeted mobile ad spending to grow from \$9.8 billion in 2015 to \$29.5 billion in 2020, a 24.6 percent compound annual growth rate.

The forecast offers breakouts of ad spend for search, traditional display, native social, traditional video, and messaging. Search will continue to eclipse all ad formats, holding the largest share of location-targeted ad spend through the forecast period. However, that share will decrease from 57 percent in 2016 to 42 percent in 2020.

**BIA/Kelsey projects U.S. location-targeted mobile ad spending to grow from \$9.8 billion in 2015 to \$29.5 billion in 2020.**

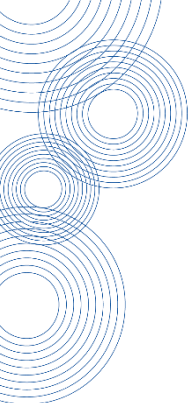




# Carriers Sell Users' Tracking Data in \$5.5 Billion Market

SAP, Germany, estimates wireless carrier revenue from selling mobile-user behavior data in \$5.5 billion in 2015 and \$9.6 billion for 2016. Other estimates for 2020 put it at \$79 billion.

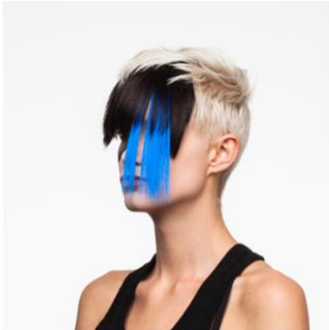




# Countermeasures

# CV Dazzle

- A project by Brooklyn artist Adam Harvey.
- Makeup tips to fool facial recognition software.



*Look N° 5 (a)*

For New York Times Op-Art  
Model: Bre Bitz  
Hair: Pia Vivas  
Makeup: [Giana DeYoung](#)  
Assistant Creative Direction: Tiam Taheri



*Look N° 5 (b)*

For New York Times Op-Art  
Model: Bre Bitz  
Hair: Pia Vivas  
Makeup: [Giana DeYoung](#)  
Assistant Creative Direction: Tiam Taheri



*Look N° 5 (c)*

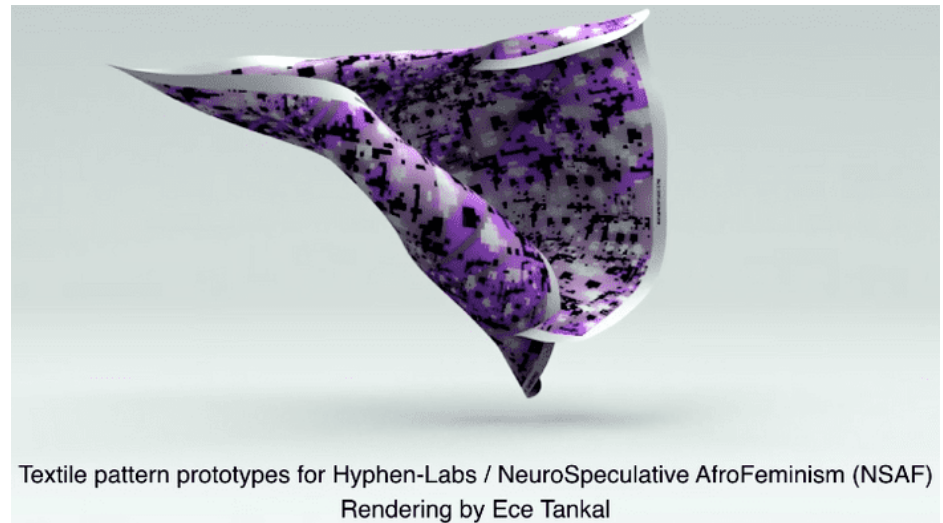
For New York Times Op-Art  
Model: Bre Bitz  
Hair: Pia Vivas  
Makeup: [Giana DeYoung](#)  
Assistant Creative Direction: Tiam Taheri

# Hyperface

- By the same artist, tries to confound the face detection software by creating textile fabrics with lots of ‘faces’.

## Anti-surveillance clothing aims to hide wearers from facial recognition

Hyperface project involves printing patterns on to clothing or textiles that computers interpret as a face, in fightback against intrusive technology



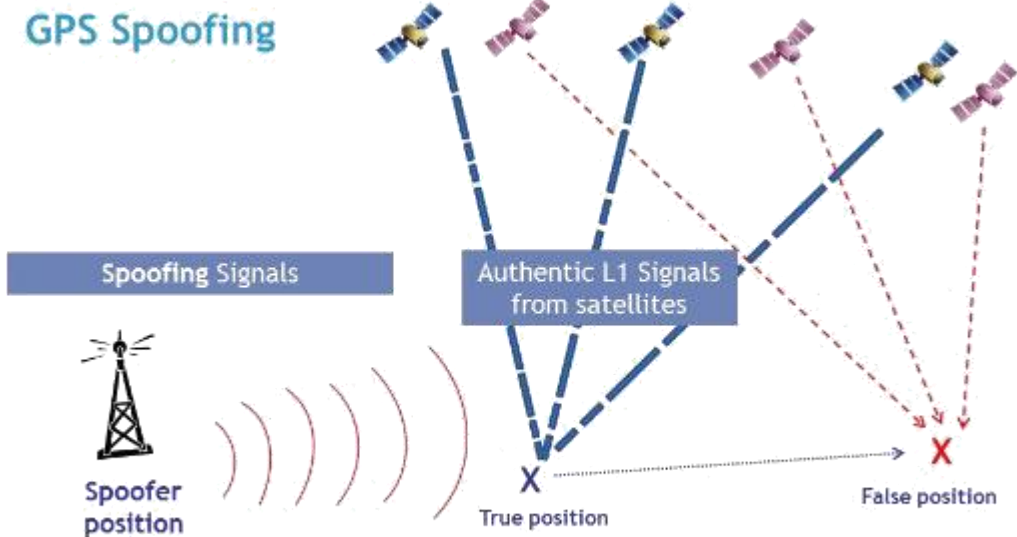


# 'Reflectacles'



# GPS Spoofing

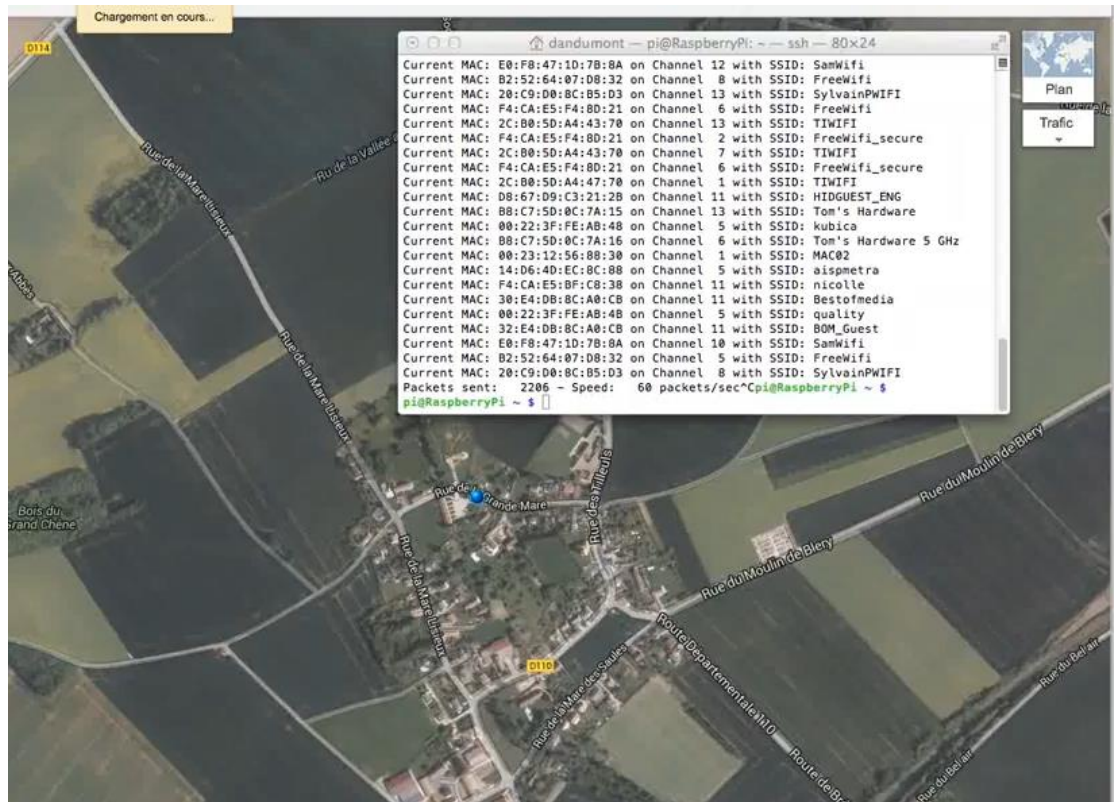
- Deceive a GPS receiver by transmitting fake (but legitimate looking) GPS signals.
- Its becoming more of a threat since the advent of cheap SDR platforms.
- Works with  
Pokemon-Go too!



Source: spirent.com

# Wi-Fi-based location spoofing

- Create fake WiFi networks with extremely cheap hardware.
- List of APs MACs is available at e.g. Wigle. Need to create more fake networks than correct ones at a given point.



Source: <https://www.journaldulapin.com/2013/08/26/dont-trust-geolocation/>

# IP-based location spoofing

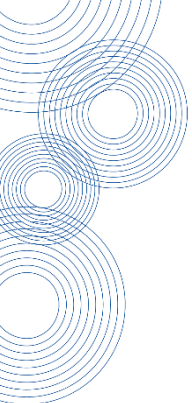
## Geo-spoofing: How to Fake Your Location Using a VPN

By Top10.com Staff | Nov.08, 2018



Advertising Disclosure





# How about anonymization/pseudonymization?

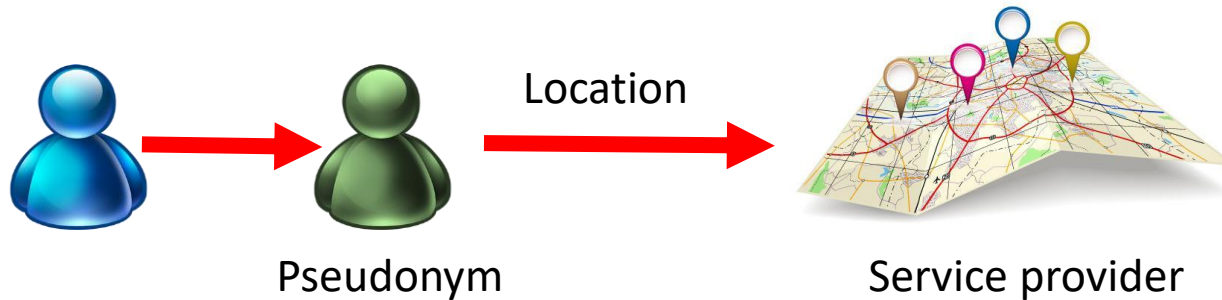
# Anonymization



## Problems:

- Difficult authentication and personalization.
- Operating system or apps may access location before anonymization.

# Pseudonymization

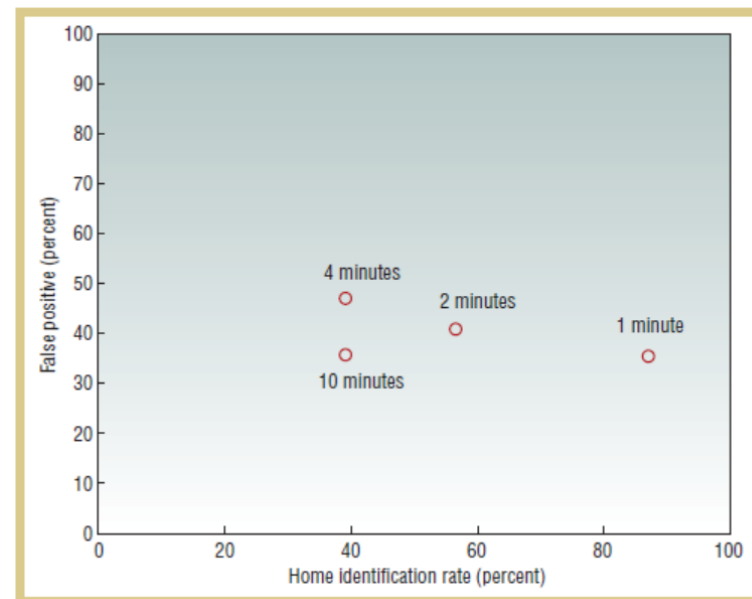
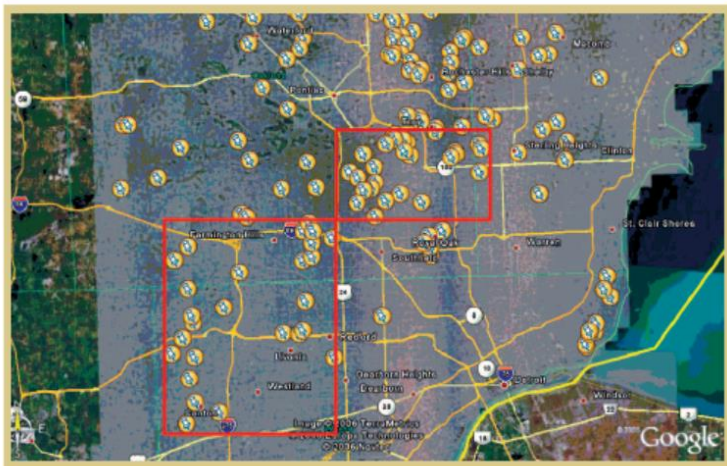


## Problems:

- Operating system or apps may access location data before pseudonymization.
- Deanonimization.

# Deanonymization based on home location [Hoh, Gruteser et al 2006\*]

- Data from GPS traces of larger Detroit area (1 min resolution).
- No data when vehicle parked.
- K-means algorithm for clustering locations + 2 heuristics:
  - Eliminate centroids that don't have evening visits.
  - Eliminate centroids outside residential areas (manually).



\* B. Hoh, M. Gruteser, H. Xiong and A. Alrabady, "Enhancing Security and Privacy in Traffic-Monitoring Systems," in *IEEE Pervasive Computing*, 2006]



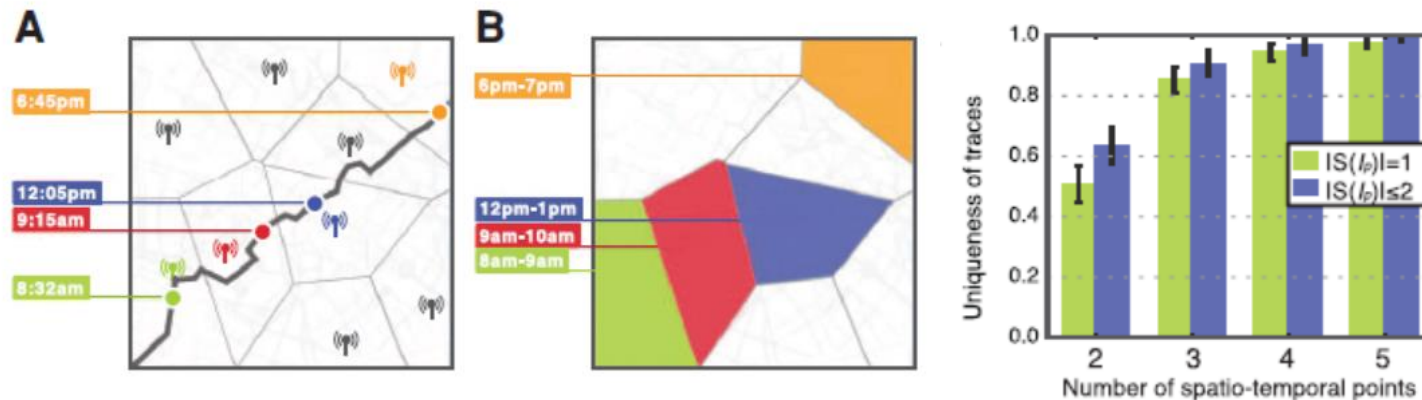
# Deanonymization based on home location [Krumm 2008\*]

- 2- week GPS data from 172 subjects (avg. 6 sec resolution).
- Use heuristic to single out trips by car.
- Then use several heuristics: destination closest to 3 a.m. is home; place where individual spends most time is home; center of cluster with most points is home.
- Use reverse geocoding and white pages to deanonymize. Success measured by finding out name of individual.
- Positive identification rates around 5%.
- Even noise addition with  $\text{std}=500$  m gives around 5% success when measured by finding out correct address.

\* J. Krumm, A Survey of Computational Location Privacy, Personal and Ubiquitous Computing, 2008

# Mobile trace uniqueness [de Montjoye et al 2013\*]

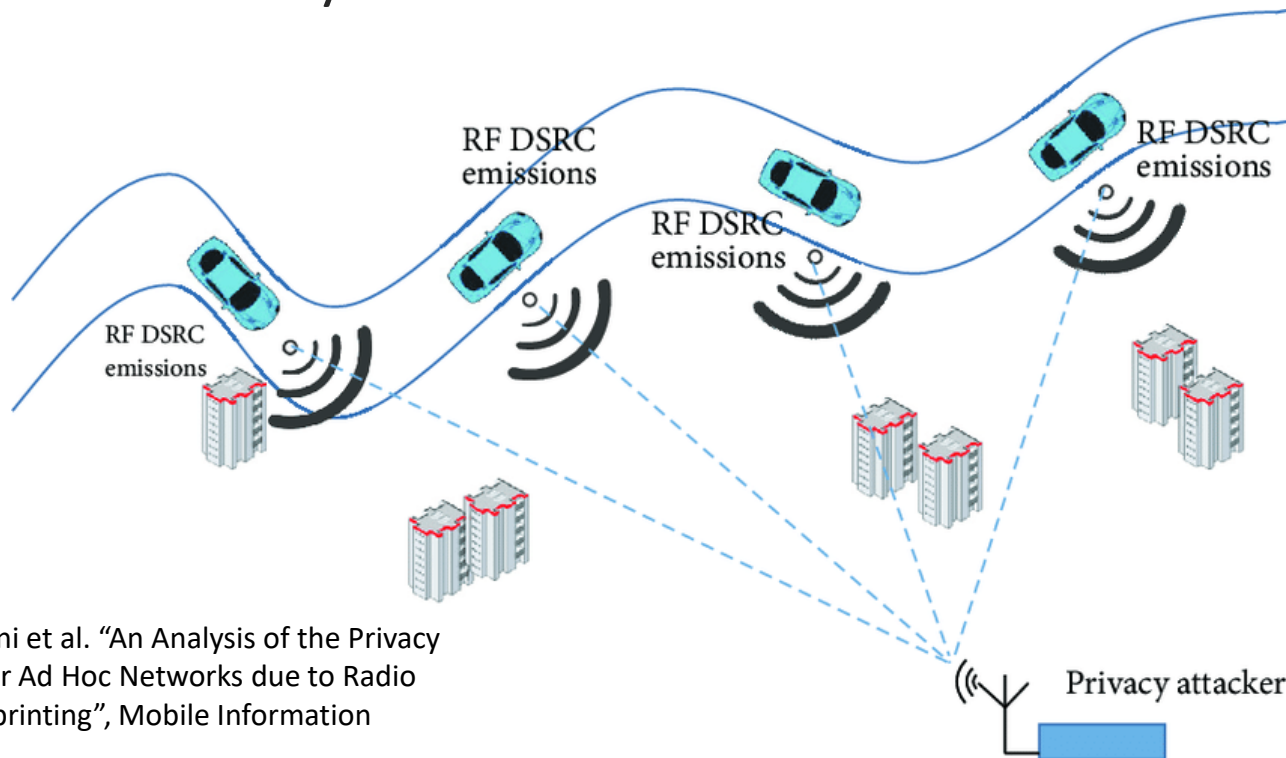
- Study on 15 months of mobility data; 0.5M individuals.
- Dataset with hourly updates and resolution given by cell carrier antennas, only 4 points suffice to identify 95% of individuals.
- Uniqueness of mobility traces decays as 1/10th power of their resolution.



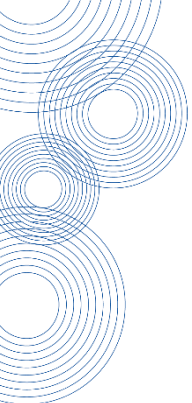
\*Source: Y. de Montjoye et al. Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports, 2013

# Deanonymization in VANETs [Baldini et al 2017\*]

- Even if pseudonyms are used, the RF fingerprint from the Dedicated Short Range Communications transceiver can be used to deanonymize the vehicle.



\* Source: G. Baldini et al. "An Analysis of the Privacy Threat in Vehicular Ad Hoc Networks due to Radio Frequency Fingerprinting", Mobile Information Systems, 2017.



# Location Privacy Protection Mechanisms (LPPMs)

# Writing location in incomprehensible language



Source: CNN

# Privacy-preserving queries

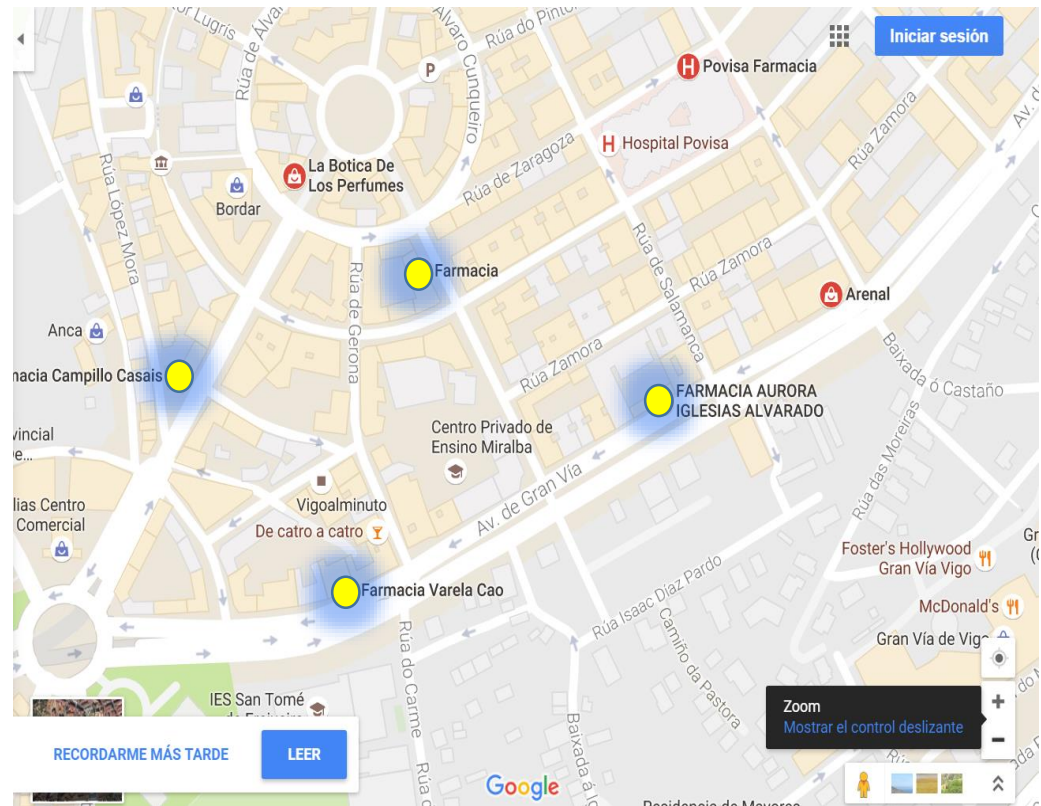
Retrieval in Encrypted Domain



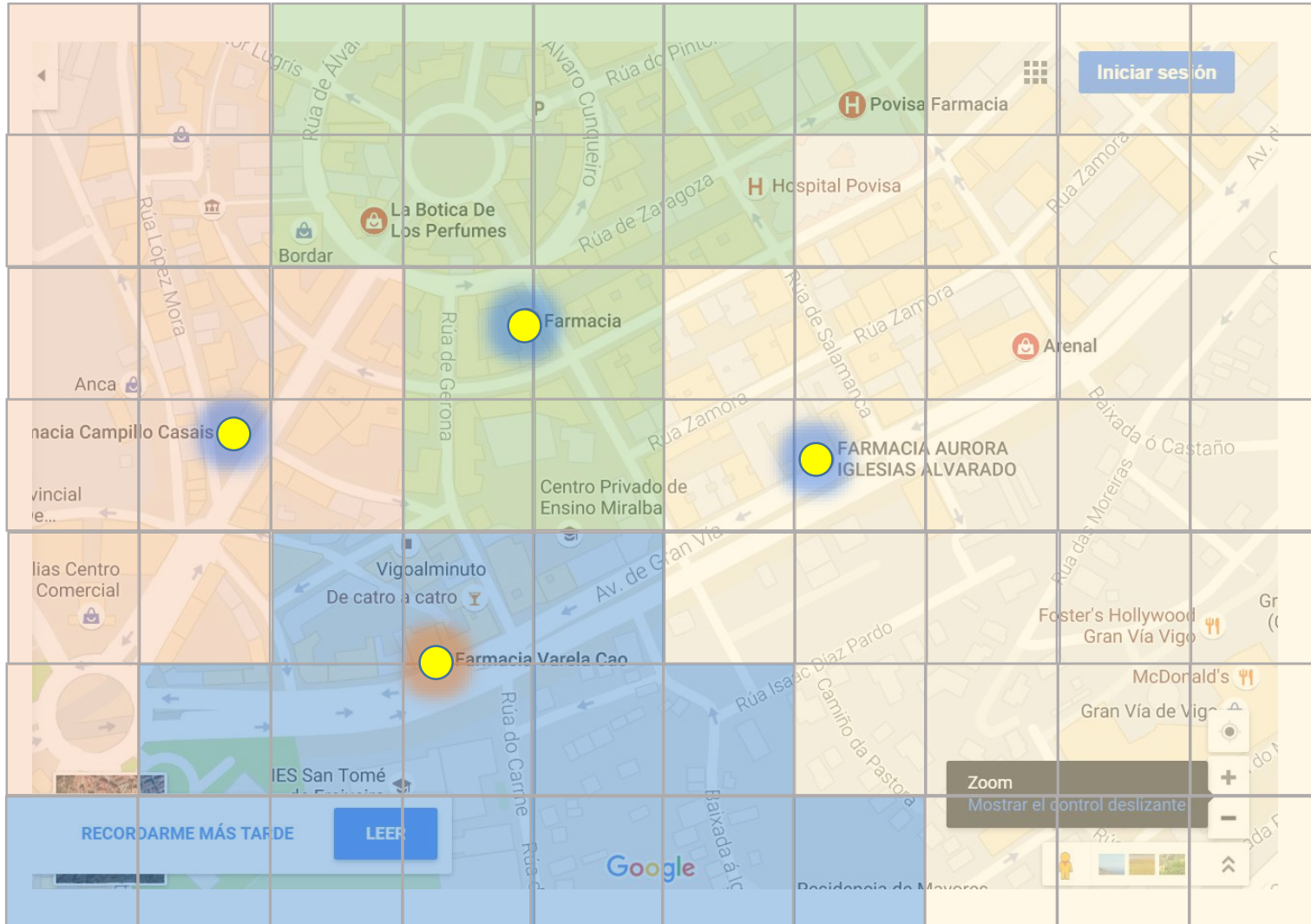
# Private Information Retrieval of Location [Ghinita et al., 2008\*]

- Query the server for the closest pharmacy without it knowing where we are.
- Example:
  - 4 pharmacies
  - Map gridding.
  - Distance rule: square is closer to the pharmacy for which more points are closer.

\*G. Ghinita et al. "Private queries in location based services: Anonymizers are not necessary," in Proc. ACM SIGMOD, Vancouver, BC, Canada, 2008







# Private Information Retrieval of Location (2)





## Private Information Retrieval of Location (3)

- Assign a number to every square in the “cloaking region” (CR).
- Example: CR has  $7 \times 10 = 70$  cells.
- Server constructs a  $7 \times 10$  matrix with 2 bits to indicate color, e.g.
  - 00: 
  - 01: 
  - 10: 
  - 11: 
- Protocol intends to retrieve the two bits for a certain position without the server learning which position that is.

## Private Information Retrieval of Location (4)

- Recall: if  $a, p$  are integers, then  $a \bmod p$  is the remainder of the division of  $a$  by  $p$ .
- Given integer  $N$ , integer  $a$  is a quadratic residue (QR) modulo  $N$  iff there exists integer  $y$  such that

$$a \equiv y^2 \pmod{N}$$

For instance, 5 is QR mod 11 because  
 $5 \equiv 4^2 \pmod{11}$   
but 6 is a Quadratic Non-Residue  
(QNR) mod 11.

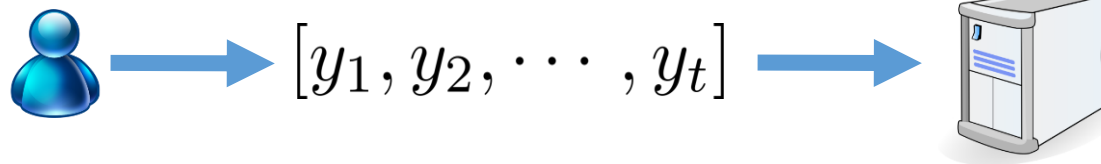
- Modulo a prime number  $p$ , there are  $(p - 1)/2$  QRs and  $(p - 1)/2$  QNRs in  $\{1, \dots, p - 1\}$ .
- Modulo a composite number  $N = p \cdot q$ , integer  $a$  is QR iff it is QR modulo both  $p$  and  $q$ .

## Private Information Retrieval of Location (5)

- Let  $\mathcal{S}$  be the set of integers that are QR modulo both  $p, q$  or QNR modulo both  $p, q$ .
- Quadratic residuosity assumption:
  - Given  $a$ , it is feasible to know whether  $a \in \mathcal{S}$ , but
  - Given  $a \in \mathcal{S}$  it is computationally unfeasible to know whether  $a$  is QR mod  $N$  or QNR mod  $N$ , if the factorization of  $N$  is not known.

Those are QR mod  $N$

- Then, if the user sends a vector of integers in  $\mathcal{S}$

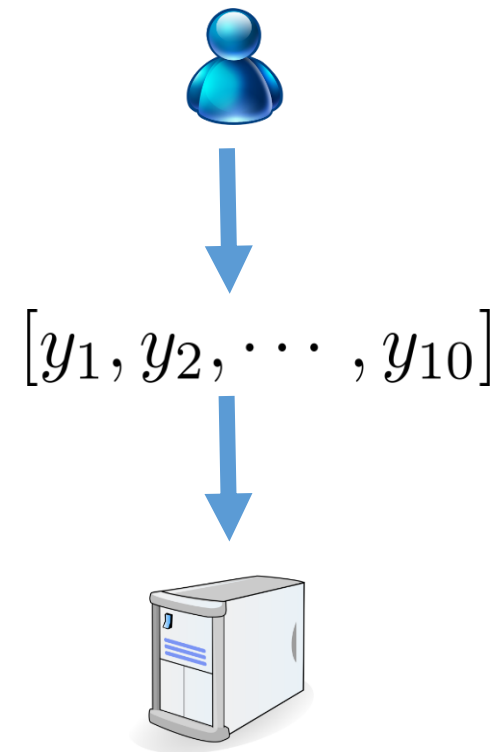
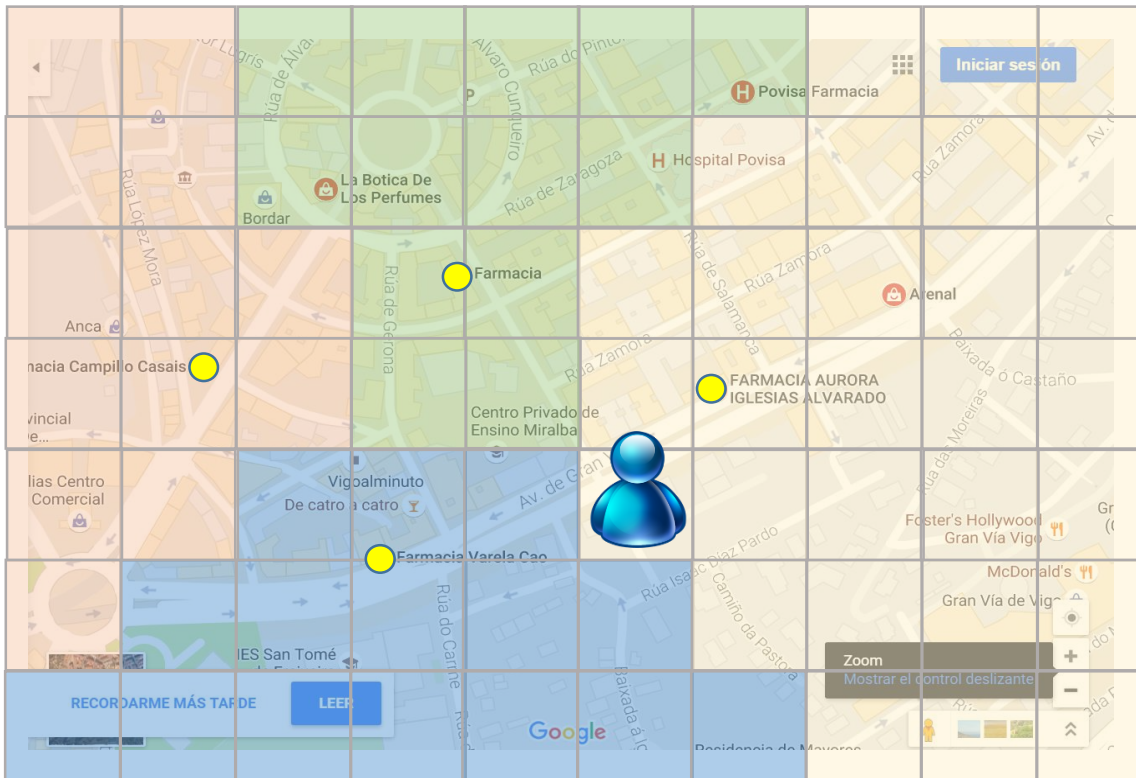


If it's not known, there is 50% chance for QR/QNR.

of which one (say  $y_m$ ) is QNR mod  $N$  and all the others QR mod  $N$ , the server cannot know where is the distinct one!

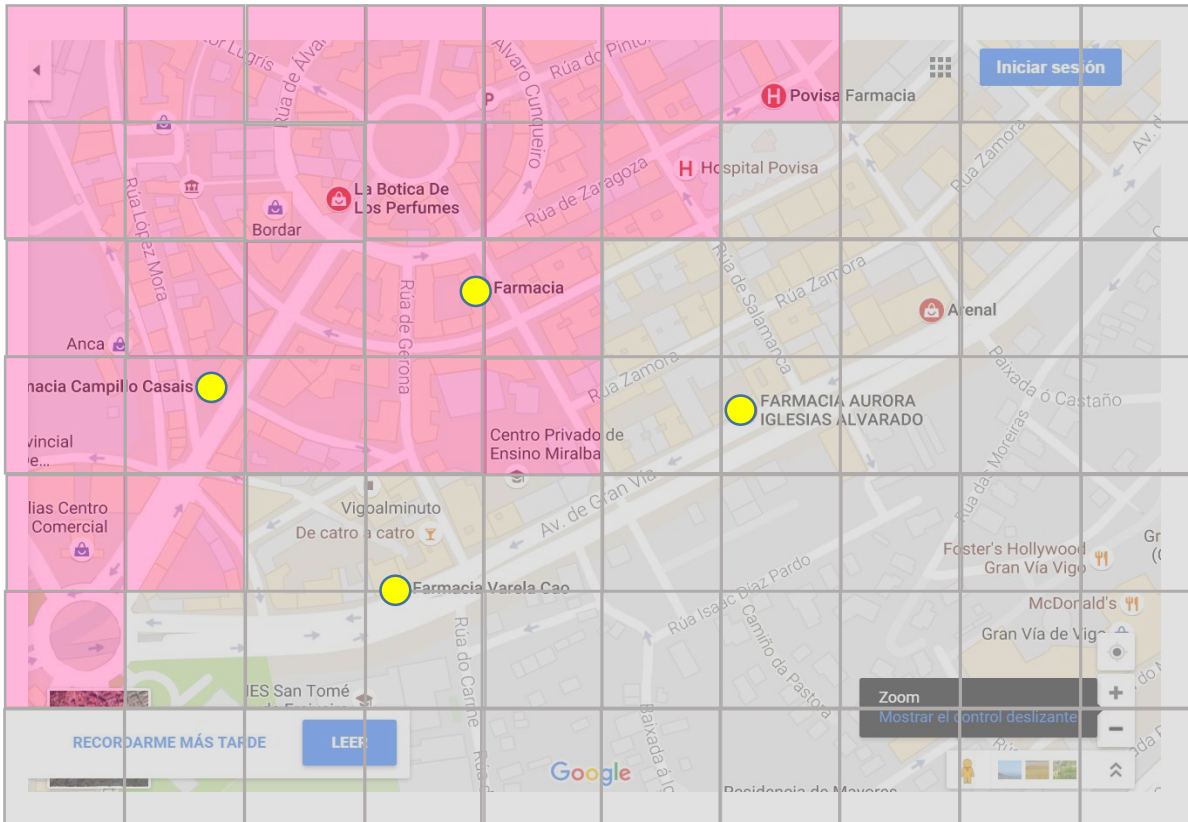
# Private Information Retrieval of Location (6)


- The user sends  $[y_1, y_2, \dots, y_{10}]$ . Only  $y_6$  (corresponding to the column where he is) is QNR. All the rest, QR.




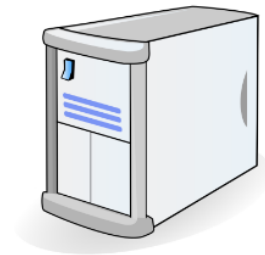
# Private Information Retrieval of Location (6)

- The server has a matrix for each output bit-plane. In our example, for the first bit:



  $M_{i,j} = 0$

  $M_{i,j} = 1$



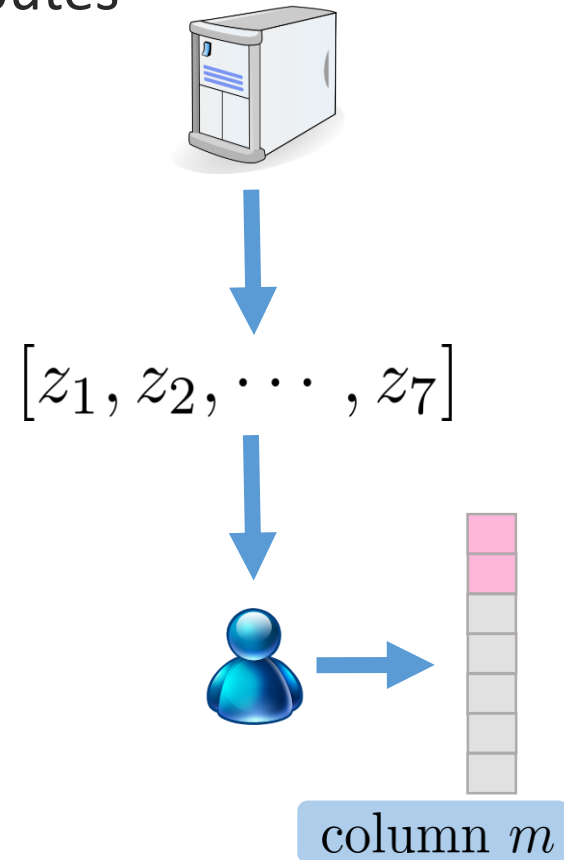
# Private Information Retrieval of Location (7)

- For every row of matrix  $\mathbf{M}$ , the server computes

$$z_r = w_{r,1} \cdot w_{r,2} \cdots w_{r,t}$$

where  $w_{r,j} = \begin{cases} y_j^2, & \text{if } M_{r,j} = 0 \\ y_j, & \text{if } M_{r,j} = 1 \end{cases}$

- Note that:
  - All factors of the form  $y_j^2$  are QR.
  - All factors of the form  $y_j$  are QR except for  $y_m$ .
- Then, the result  $z_r$  is
  - QR if  $M_{r,m} = 0$
  - QNR if  $M_{r,m} = 1$



## Private Information Retrieval of Location (8)

- There is a certain “dummification” of the queries: the user will get the answer for ALL cells in the same column. This increases the bandwidth cost.
- Complexity increases linearly with the number of bits in the answer (2, in our example, because there are 4 pharmacies).
- By using a 2-D reordering method, all points can be put in a 1-D vector and then, instead of sending  $t$  values and getting  $u$  answers, it is possible to send  $t \cdot u$  values and get one answer.
- There is an inherent granularity in the cells, reminiscent of quantization methods, with a corresponding loss of utility. Smaller cells increase accuracy, but also communication and computation costs.

# Homomorphic schemes

- An homomorphism is a mapping between structures that preserves operations.
- For instance, given two sets  $A, B$  and two respective operations  $\diamond, \circ$ , there is a map  $f : A \rightarrow B$  such that

$$f(x \diamond y) = f(x) \circ f(y)$$

- When  $f$  is an encryption function, the existence of an homomorphism allows to do operations over encrypted data without a prior decryption.



## Paillier scheme

- Client generates two large primes  $p, q$  which are secret. From them,  $N = p, q$  is made public.
- Client generates  $g < N^2$  with some additional properties.\*
- Given message  $m < N$  encryption is as follows:

$$c = g^m r^N \pmod{N^2}$$

where  $r < N$  is 'randomness' coprime with  $N$ .

- The randomness can be eliminated if  $p, q$  are known, because then  $\lambda = \text{lcm}\{(p - 1), (q - 1)\}$  is computable, and

$$c^\lambda \equiv g^{m\lambda} \pmod{N^2}$$

\*  $g$  must be coprime with  $N$  and such that its order is a multiple of  $N$ .

## Paillier scheme

- To recover  $m$  from  $c^\lambda$ , write  $g \equiv (1 + N)^{g'}$  mod  $N^2$  and notice that (binomial expansion)

$$(1 + N)^{g'} = 1 + Ng' + \text{terms with powers of } N \text{ higher than } 2$$

- So  $g \equiv 1 + Ng' \pmod{N^2}$  and  $g^{m\lambda} \equiv 1 + m\lambda Ng' \pmod{N^2}$
- Define the extraction function:  $L(u) = (u - 1)/N$
- Then,  $L(g^{m\lambda} \pmod{N^2}) \equiv m\lambda g' \pmod{N}$ , so if we multiply by the modular inverse of  $\lambda g'$  we recover the message.
- Note that  $\lambda g' \equiv L(g^\lambda \pmod{N^2}) \pmod{N}$



## Paillier scheme

- So, given  $c$ , decryption:
- 1) eliminates randomness  $r$  by computing  $c^\lambda$
- 2) extracts  $m\lambda g' \bmod N$  by doing  $L(c^\lambda \bmod N^2)$ , and
- 3) recovers  $m$  by multiplying by  $(\lambda g')^{-1} \bmod N$

# Paillier homomorphisms

- Given two ciphers

$$c_1 = g^{m_1} r_1^N \bmod N^2, \quad c_2 = g^{m_2} r_2^N \bmod N^2$$

- If we multiply them

$$c_1 \cdot c_2 \equiv g^{m_1 + m_2} (r_1 \cdot r_2)^N \bmod N^2$$

- When we decrypt, we get  $m_1 + m_2$ . So the sum of clear messages is equivalent to the product of their ciphers.

## Paillier homomorphisms (2)

- Given one cipher

$$c = g^m r^N \pmod{N^2}$$

- If we raise it to  $t$

$$c^t \equiv g^{mt} (r^t)^N \pmod{N^2}$$

- When we decrypt, we get  $m \cdot t$ . So the product of the message by a constant is equivalent to exponentiation of the cipher.

# Computing Euclidean distances with Paillier

- Client wants the server to compute the distance to a given point without revealing his location.

- It's easier to produce the squared distance.

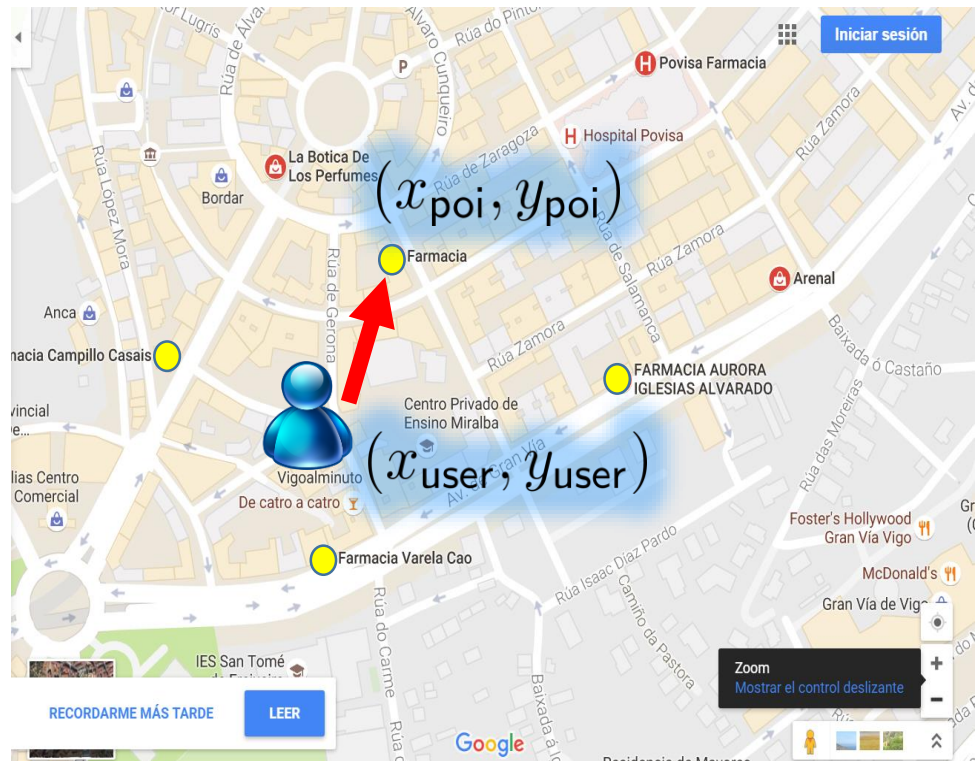
- Using Paillier, client

computes:



- $Enc(1)$
- $Enc(x_{user})$
- $Enc(y_{user})$
- $Enc(x_{user}^2 + y_{user}^2)$

- And sends all to the server.



# Computing Euclidean distances with Paillier

- Knowing the coordinates of the desired point  $(x_{\text{poi}}, y_{\text{poi}})$  server does:



$$(\text{Enc}(1))^{x_{\text{poi}}^2 + y_{\text{poi}}^2} \cdot (\text{Enc}(x_{\text{user}}))^{-2x_{\text{poi}}} \cdot (\text{Enc}(y_{\text{user}}))^{-2y_{\text{poi}}} \cdot \text{Enc}(x_{\text{user}}^2 + y_{\text{user}}^2)$$

- And sends it back to the client. When decrypting, thanks to Paillier homomorphisms, the client gets



$$\begin{aligned} & x_{\text{poi}}^2 + y_{\text{poi}}^2 - 2x_{\text{poi}} \cdot x_{\text{user}} - 2y_{\text{poi}} \cdot y_{\text{user}} + x_{\text{user}}^2 + y_{\text{user}}^2 \\ = & (x_{\text{poi}} - x_{\text{user}})^2 + (y_{\text{poi}} - y_{\text{user}})^2 \end{aligned}$$

## Location white lies



Source: Caro Spark (CC BY-NC-ND)

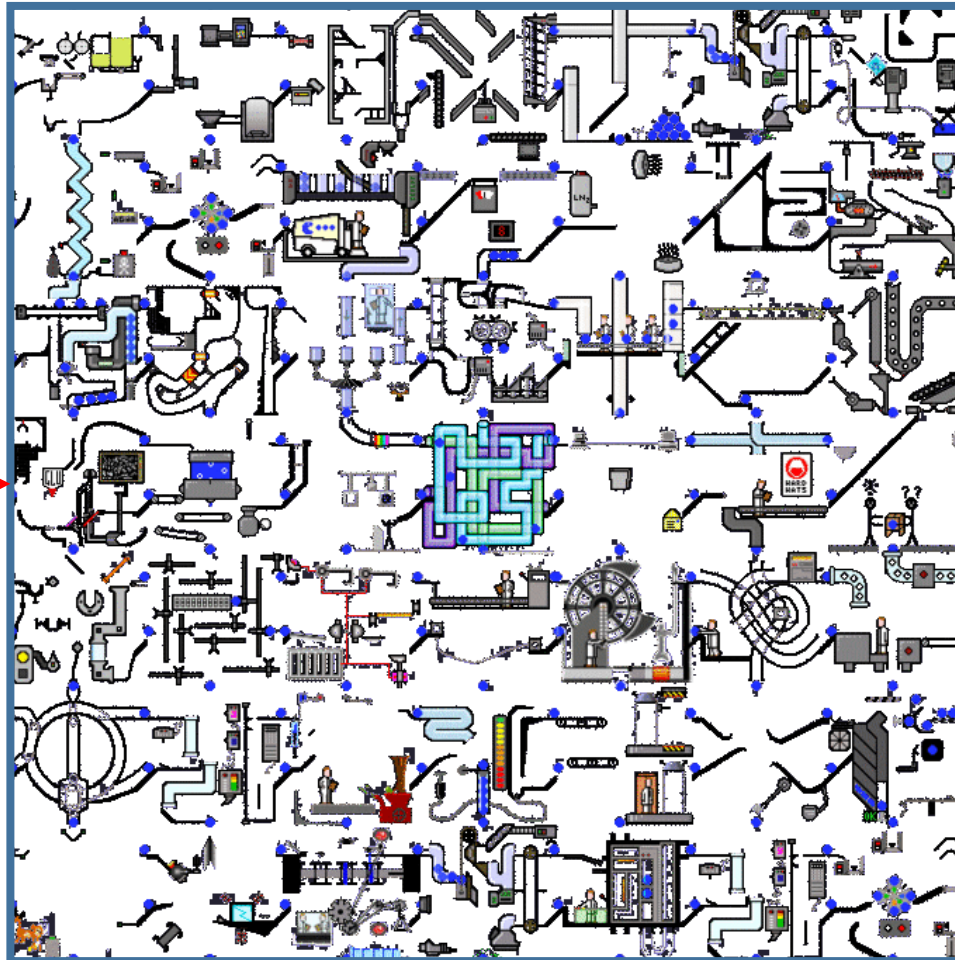


# Perturbation-based LPPMs



Input location

$X$



Output pseudolocation

$Z$

Source: Motherboards.org

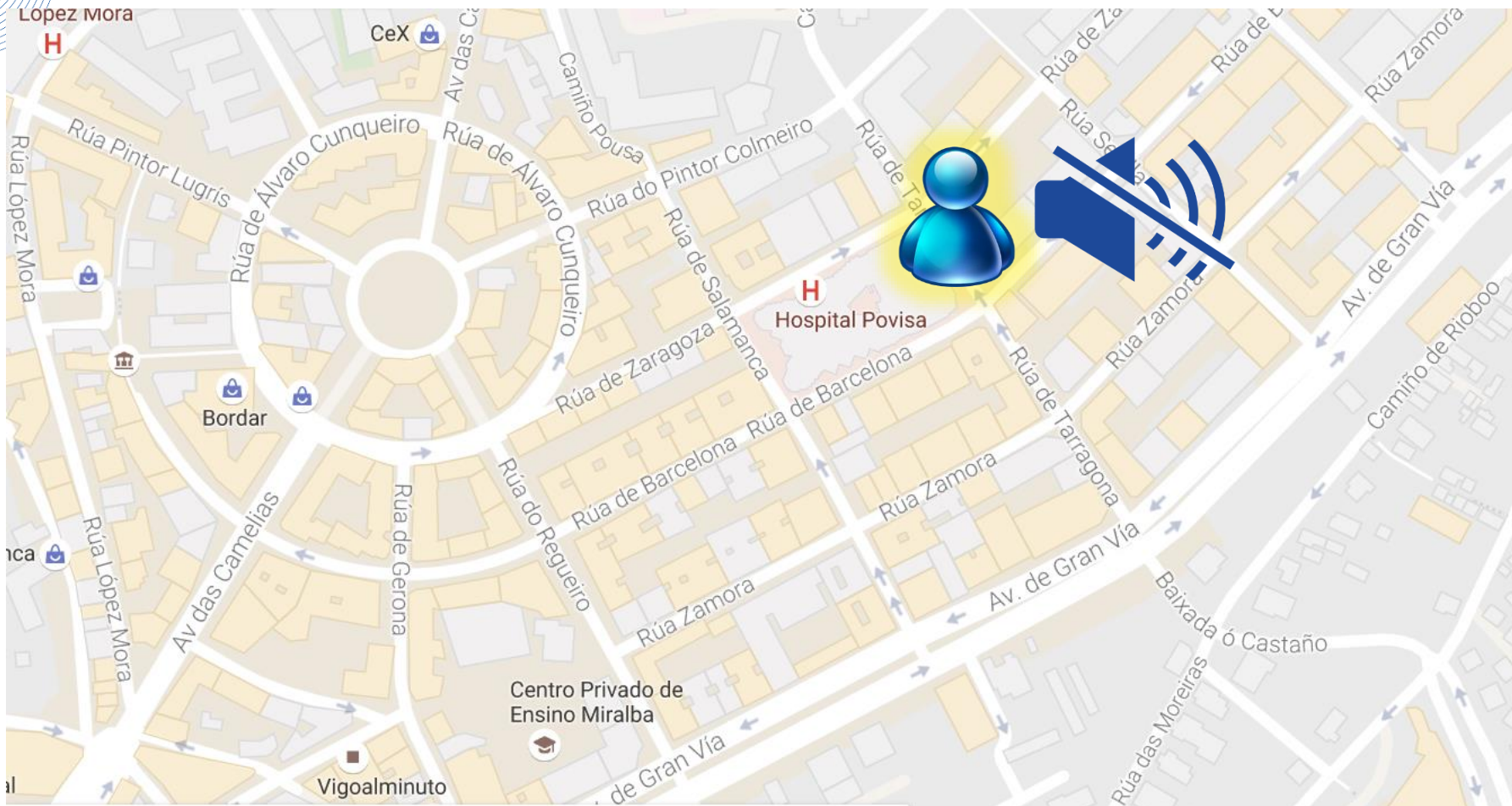


## Perturbation-based LPPMs

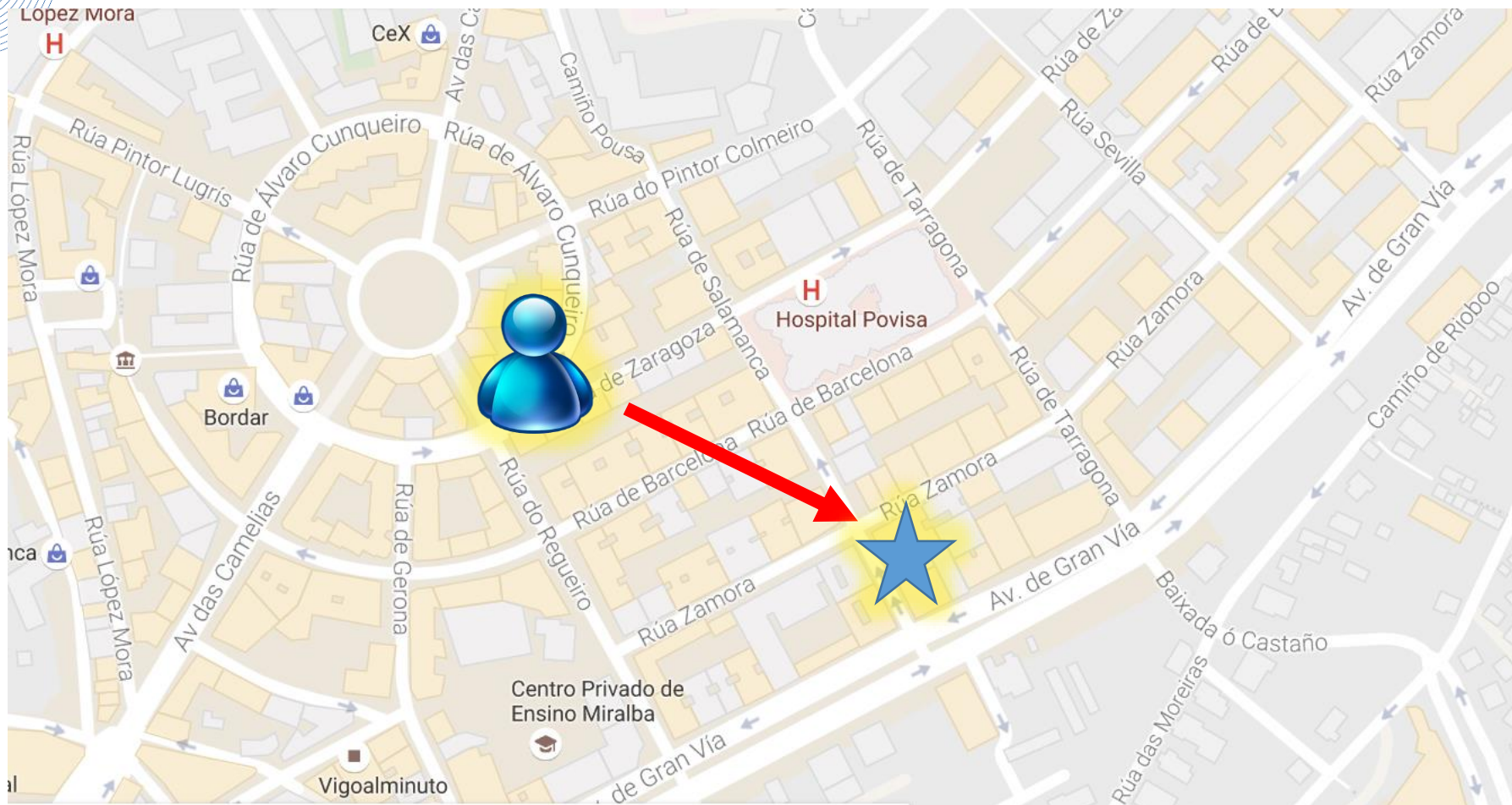
- $Z = \varphi(X)$
- The **mechanism** may be deterministic (e.g., quantization) or stochastic (e.g., noise addition).
- Function  $\varphi(\cdot)$  may depend on other contextual (e.g., time) or user-tunable (e.g., privacy level) parameters.
- When the **mechanism is stochastic**, there is an underlying probability density function, i.e.,

$$f(Z | X)$$

# Hiding

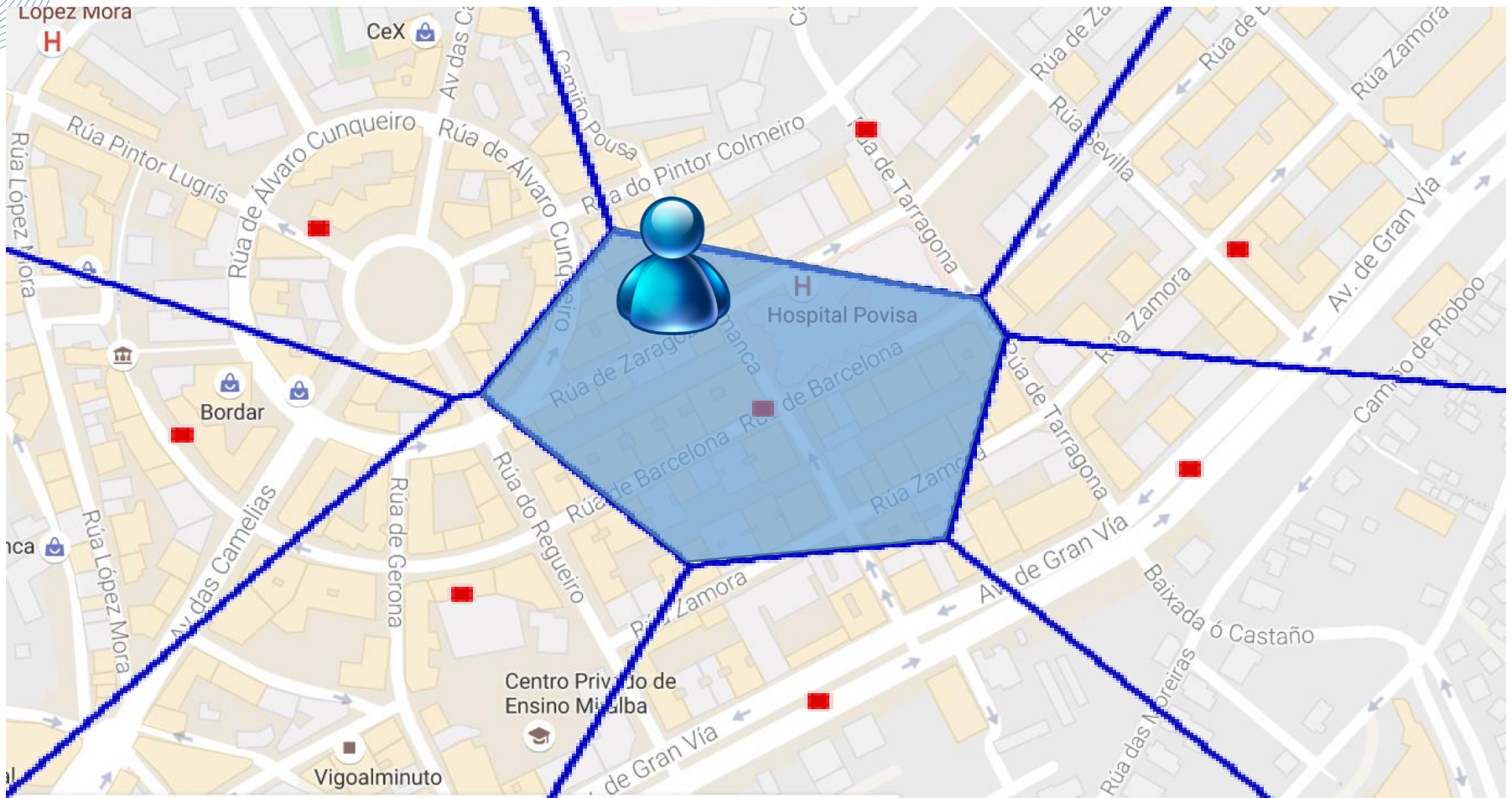


# Perturbation: (independed) noise addition

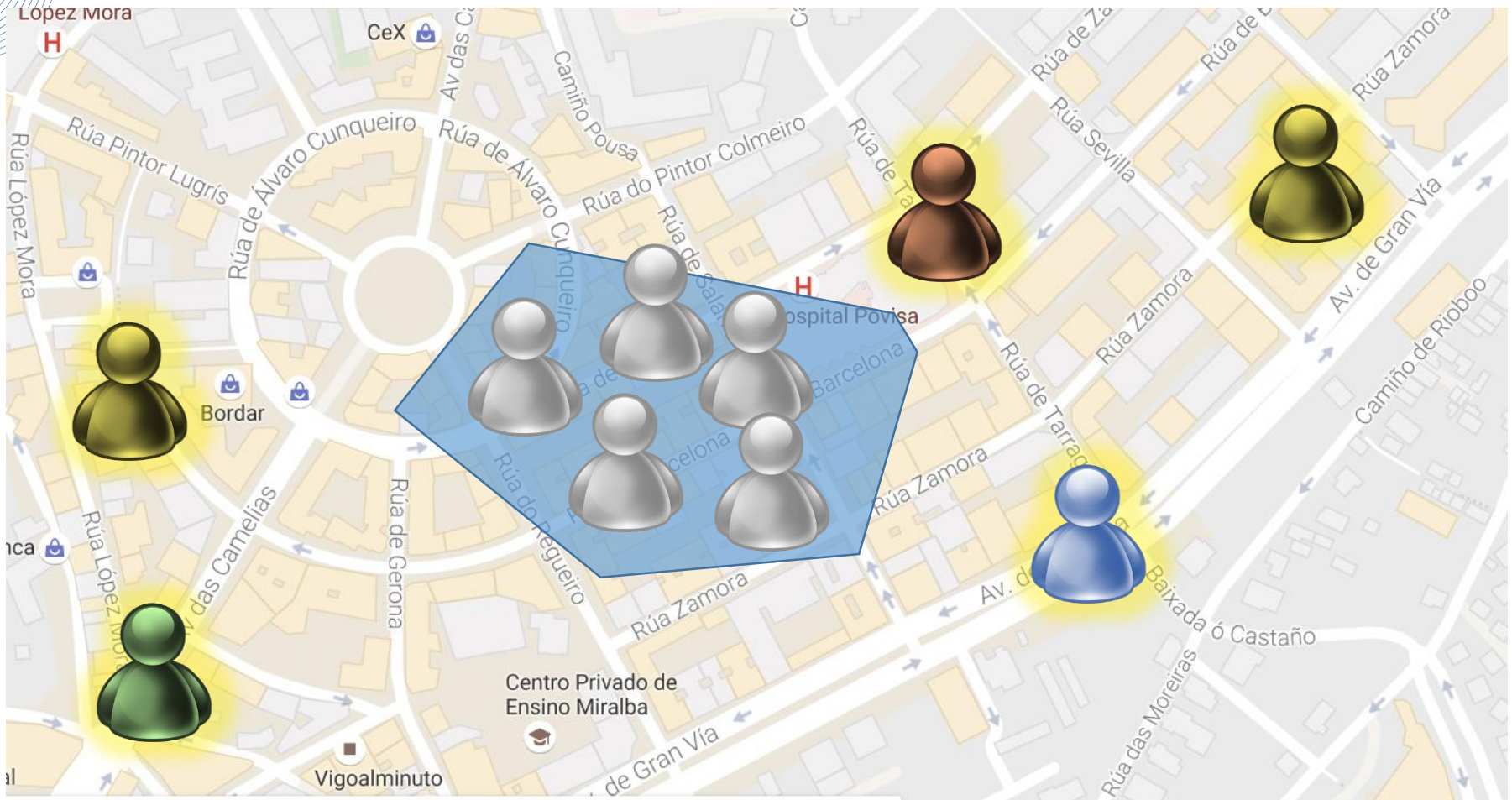




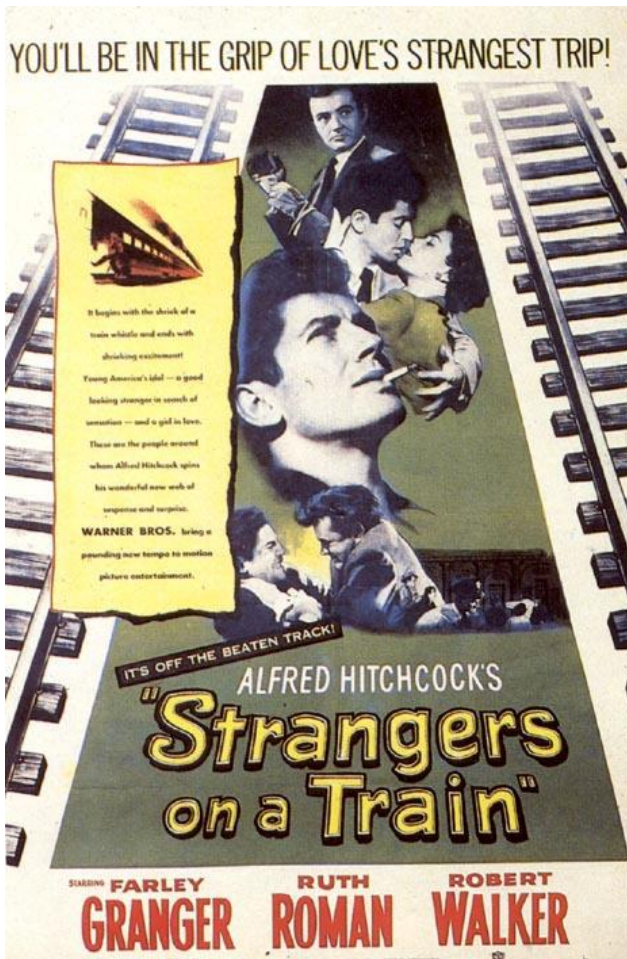
# Obfuscation



# Spatial Cloaking

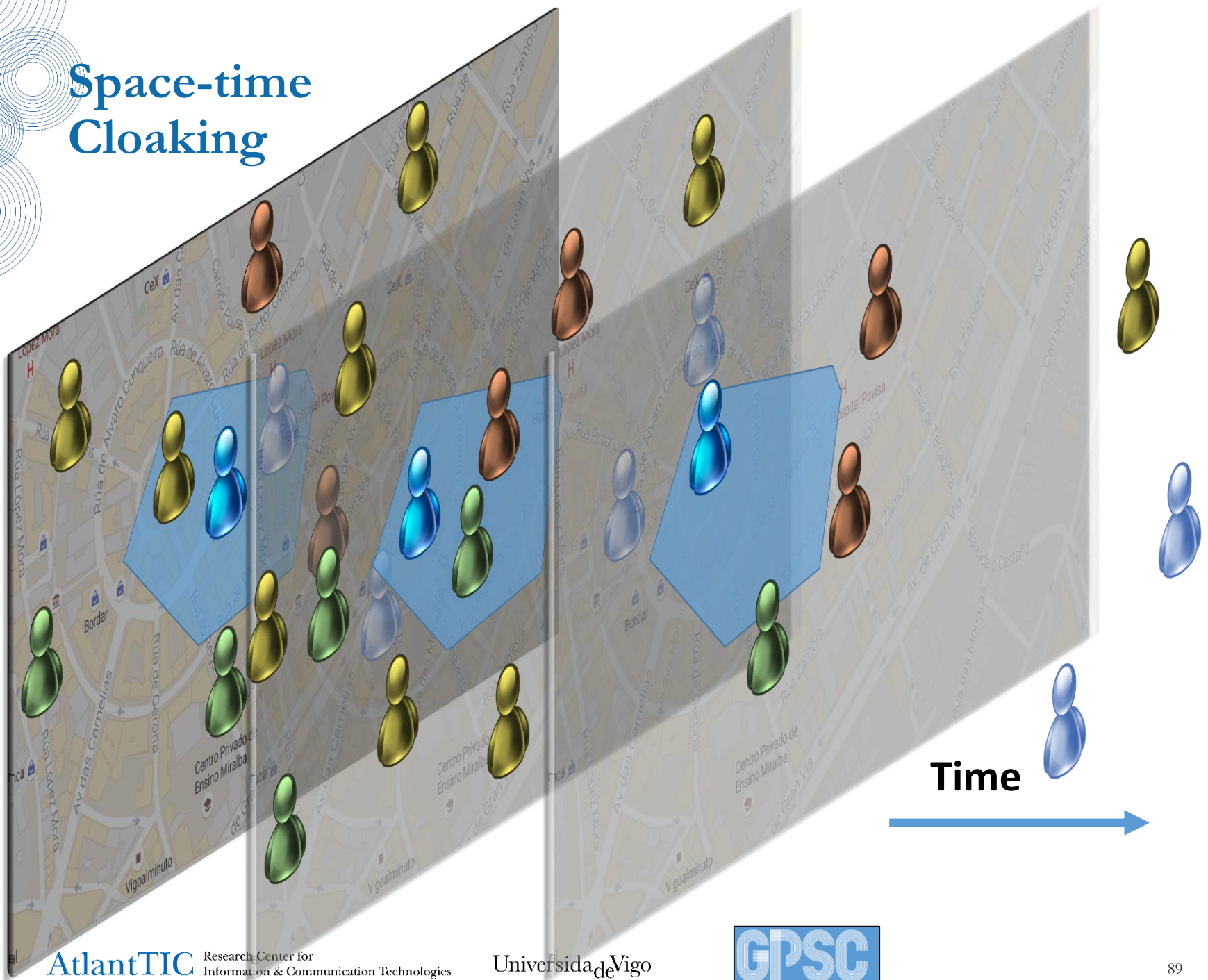


# How to commit the perfect murder



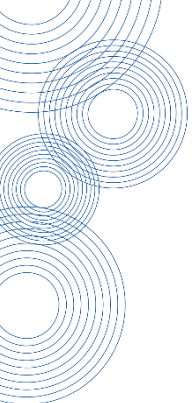


# Space-time Cloaking



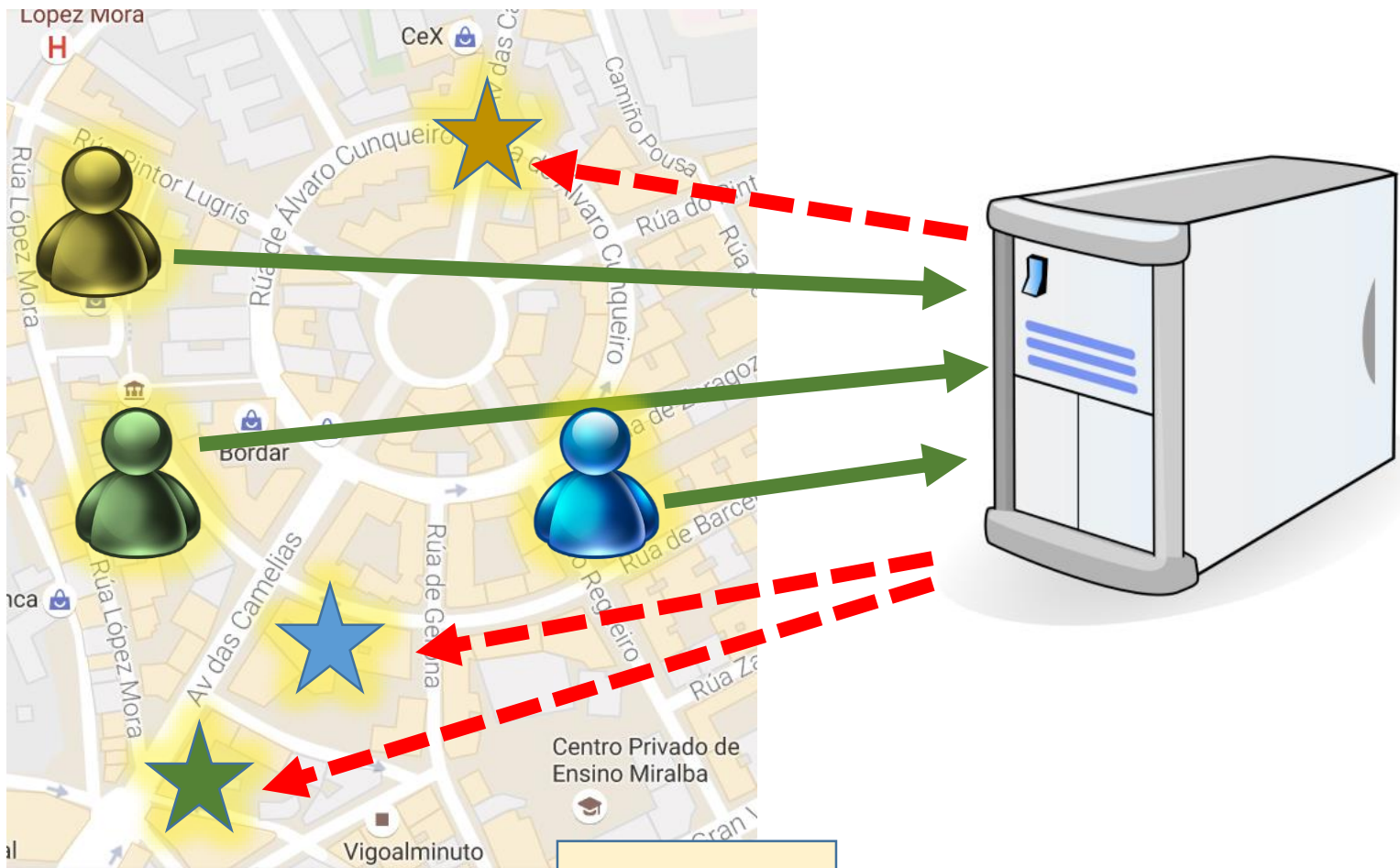
# Dummies





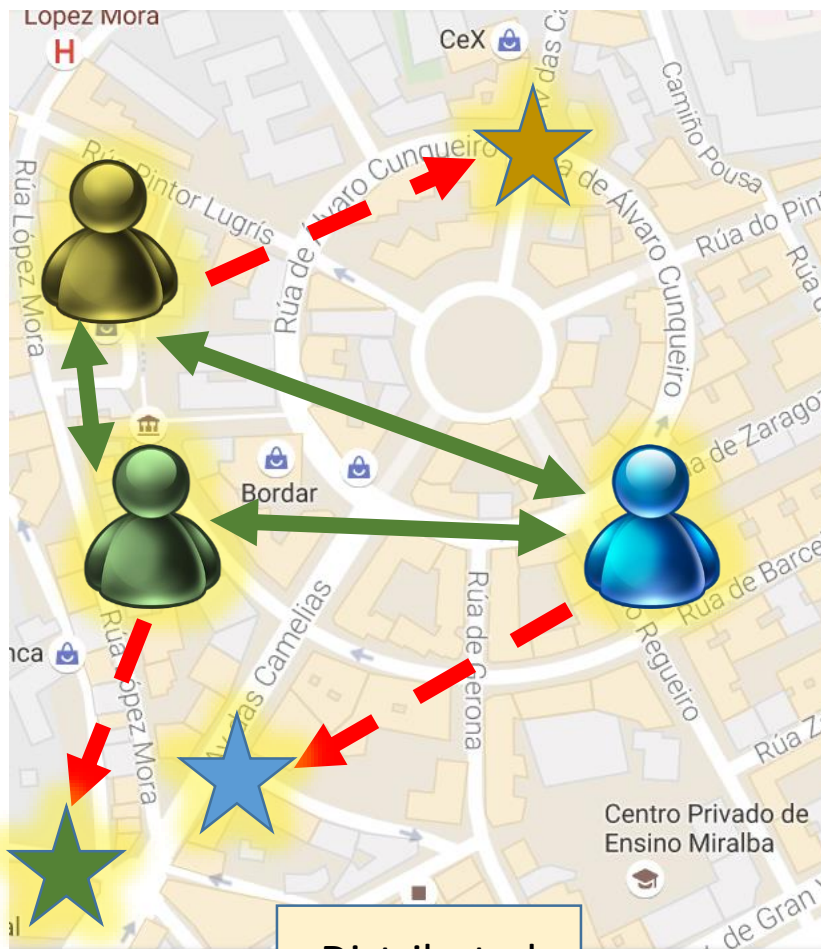
# LPPM Topologies

# Centralized LPPMs



Centralized

# Distributed LPPMs



Distributed

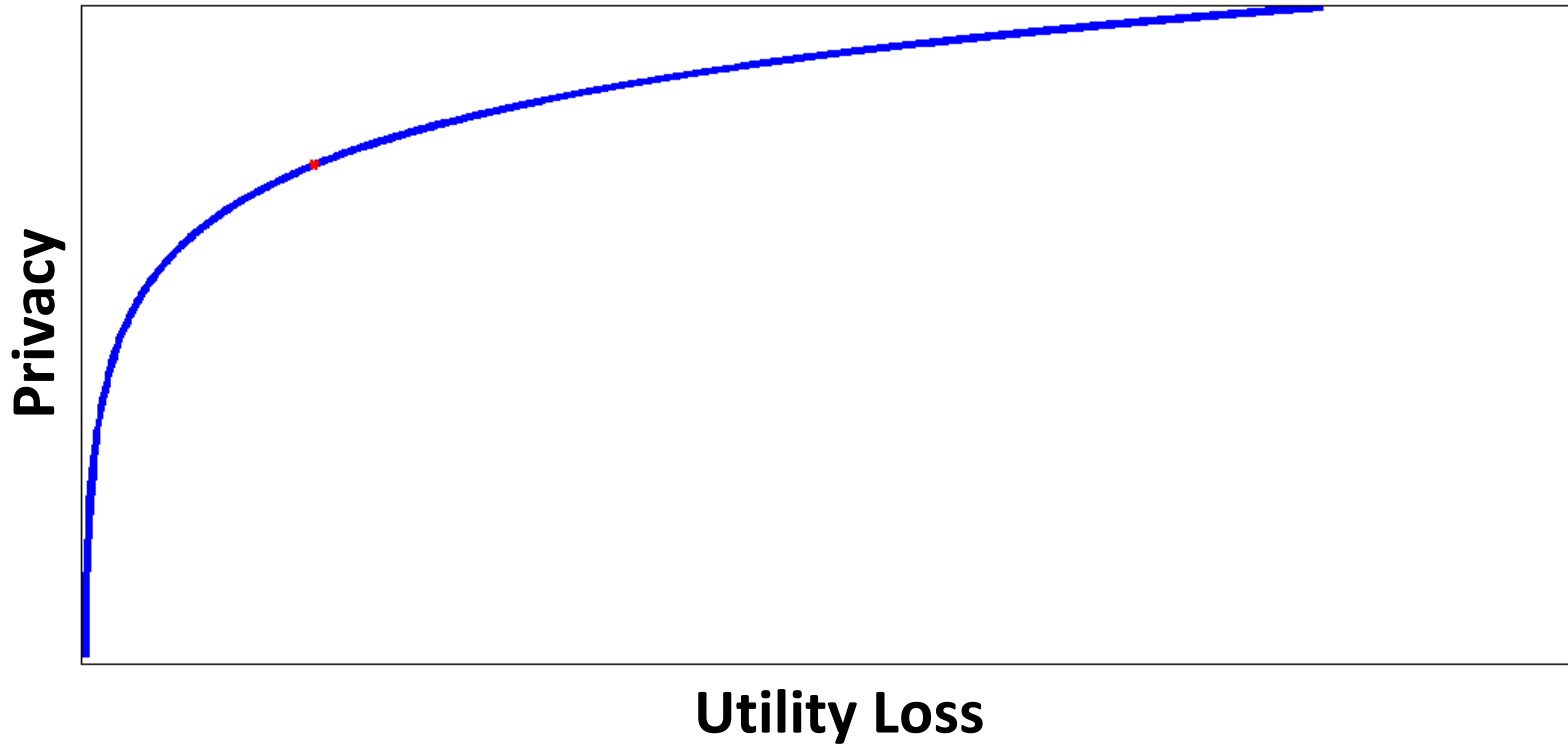
# User-centric LPPMs



User-centric

# Utility vs. Privacy

- In broad terms:

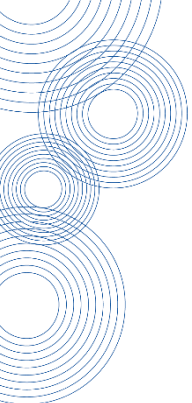


## Very nice, but...

- There are two main problems:  
How do we measure utility?  
How do we measure privacy?







# Quantifying LPPM Performance

# A (little) bit of notation

- Real locations:

$$x^r$$

- Obfuscated locations:

$$z^r$$

- Location Privacy-Preserving Mechanism (LPPM):

$$f(z^r | \mathbf{x}^r, \mathbf{z}^{r-1})$$

- Sometimes, just:

$$f(z^r | x^r)$$



# Mobility Models

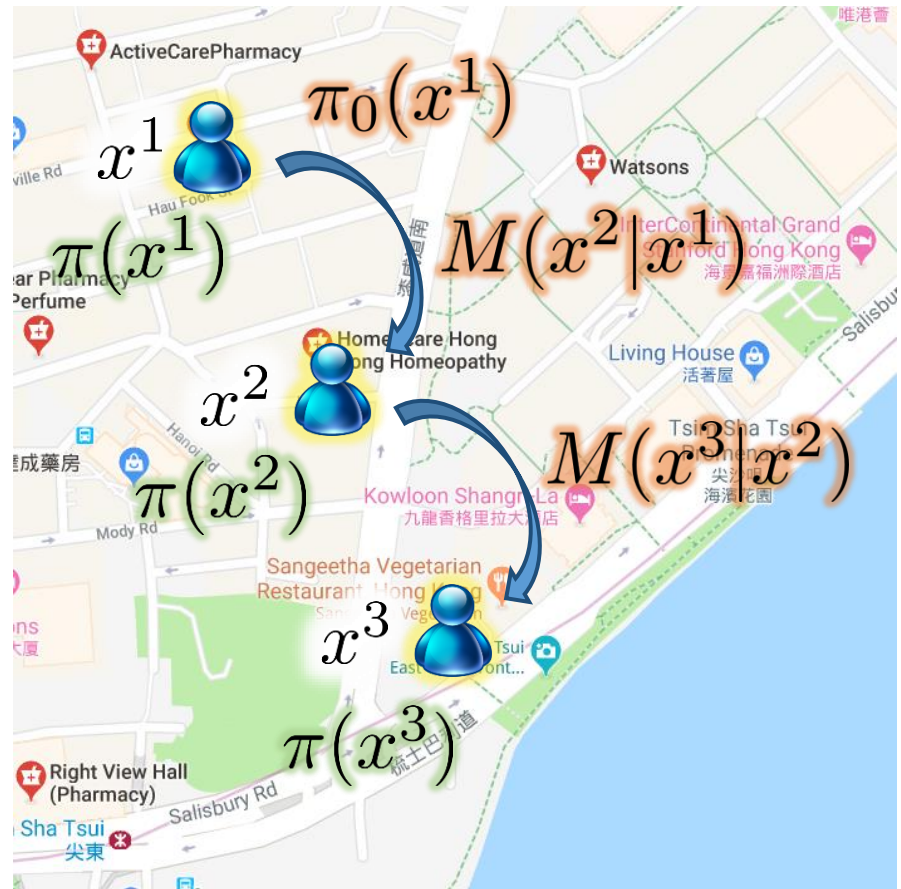
- **Sporadic:** location releases are not temporally-correlated

$$\pi(x^1, x^2, x^3) = \pi(x^1)\pi(x^2)\pi(x^3)$$

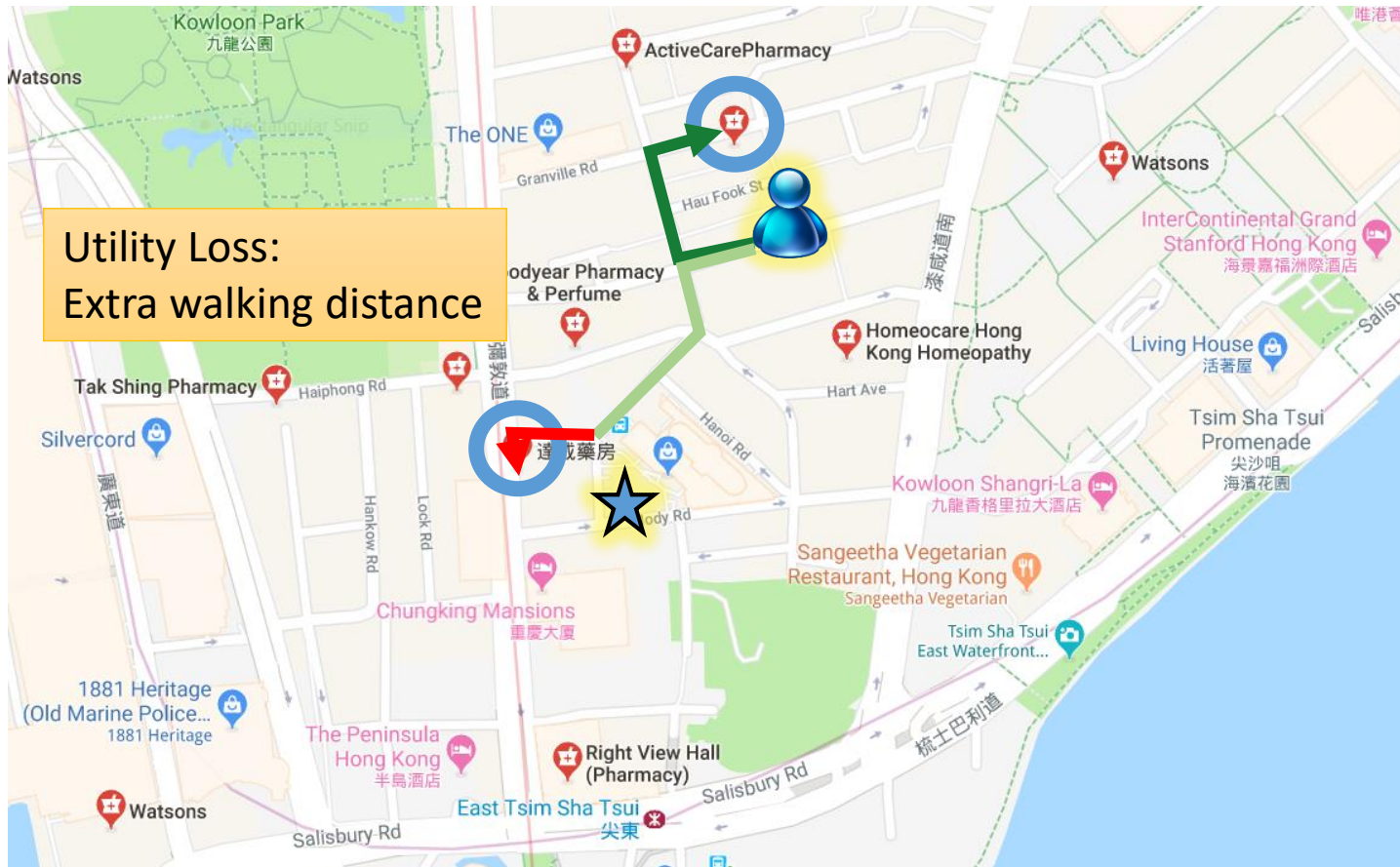
- **Non-sporadic:** temporal correlations

- Markov:  $M(x^{r+1}|x^r)$   
 $\pi_0(x^1)$

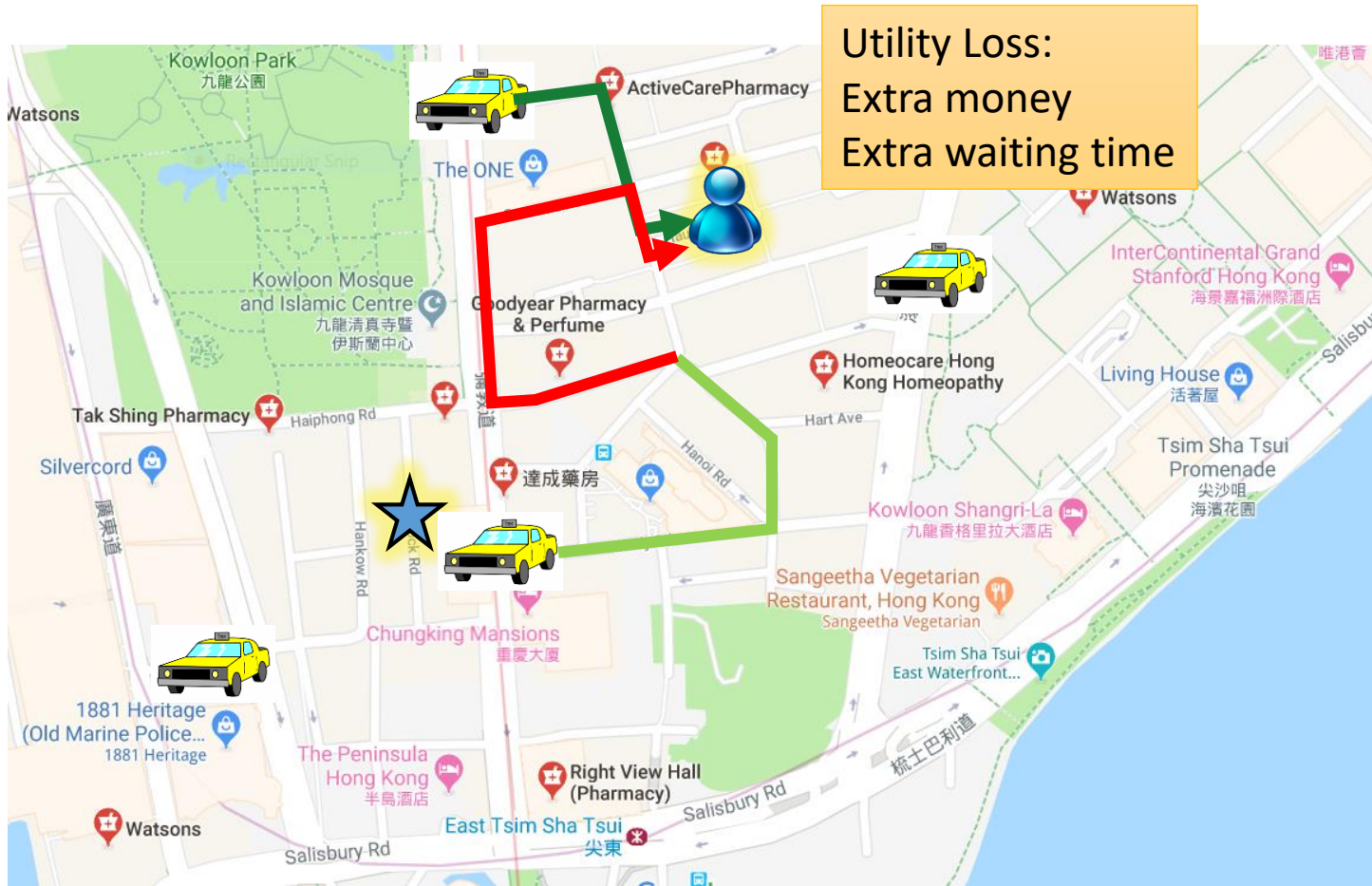
$$\pi(x^1, x^2, x^3) = \pi_0(x^1)M(x^2|x^1)M(x^3|x^2)$$



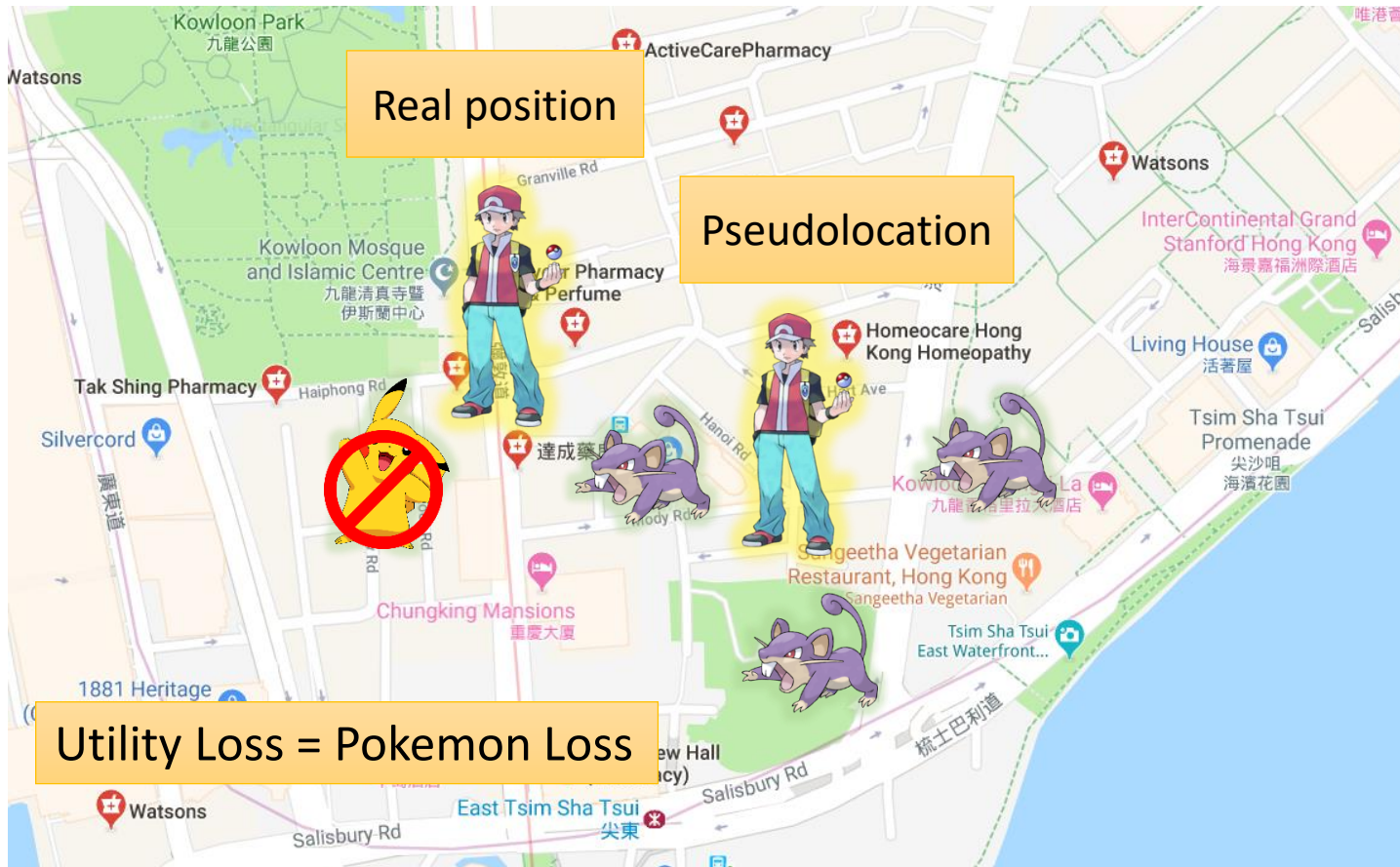
# Measuring Utility



# Measuring Utility





# Measuring Utility



## Average Quality Loss

- Formally, we define a generic point-to-point distance function:  $d_Q(x, z)$
- The most used metric is the **Average Quality Loss**:

$$\bar{Q} \doteq \mathbb{E}\{d_Q(x, z)\}$$

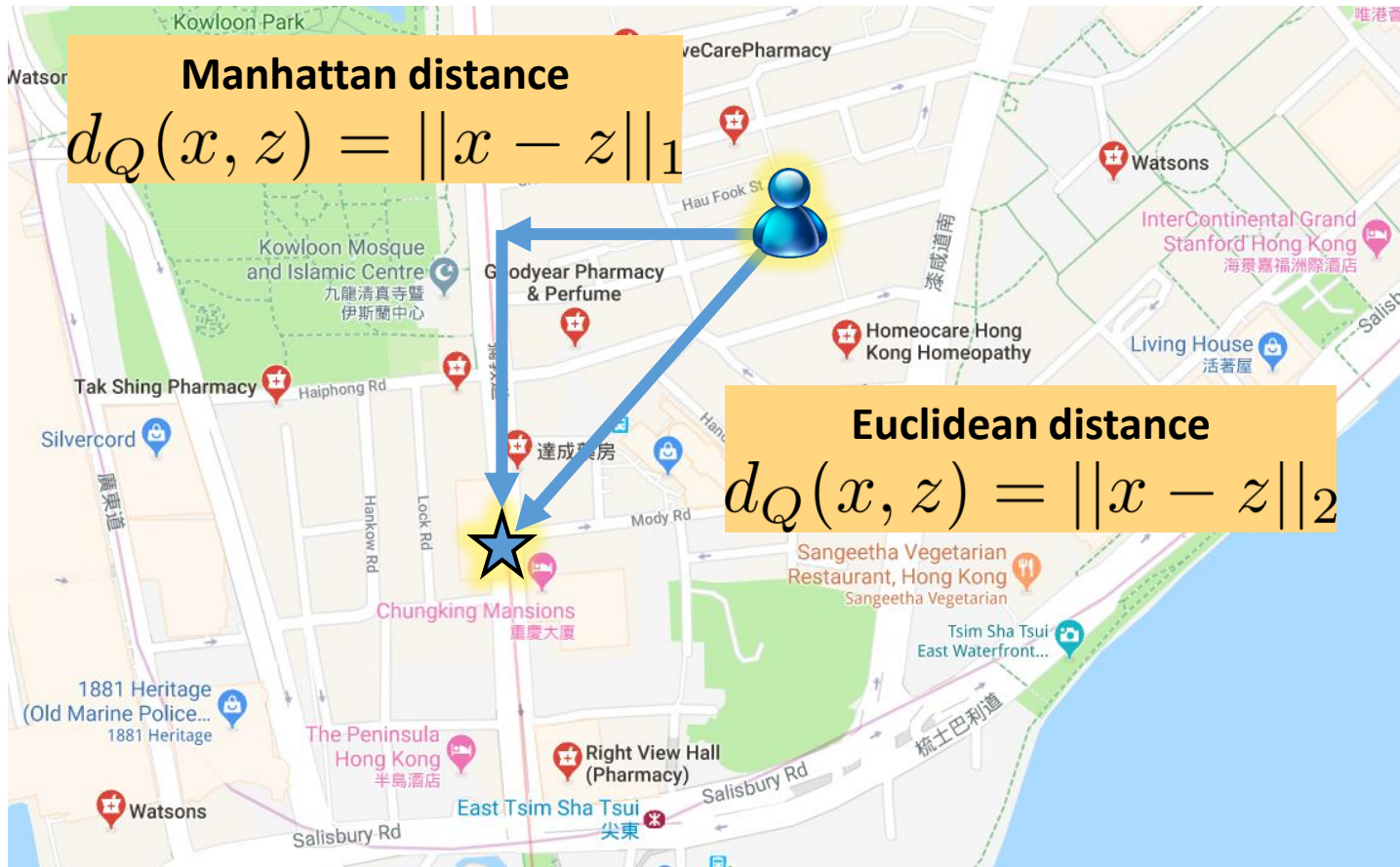
 

$$\bar{Q} = \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x) \cdot f(z|x) \cdot d_Q(x, z).$$

User mobility profile                      LPPM

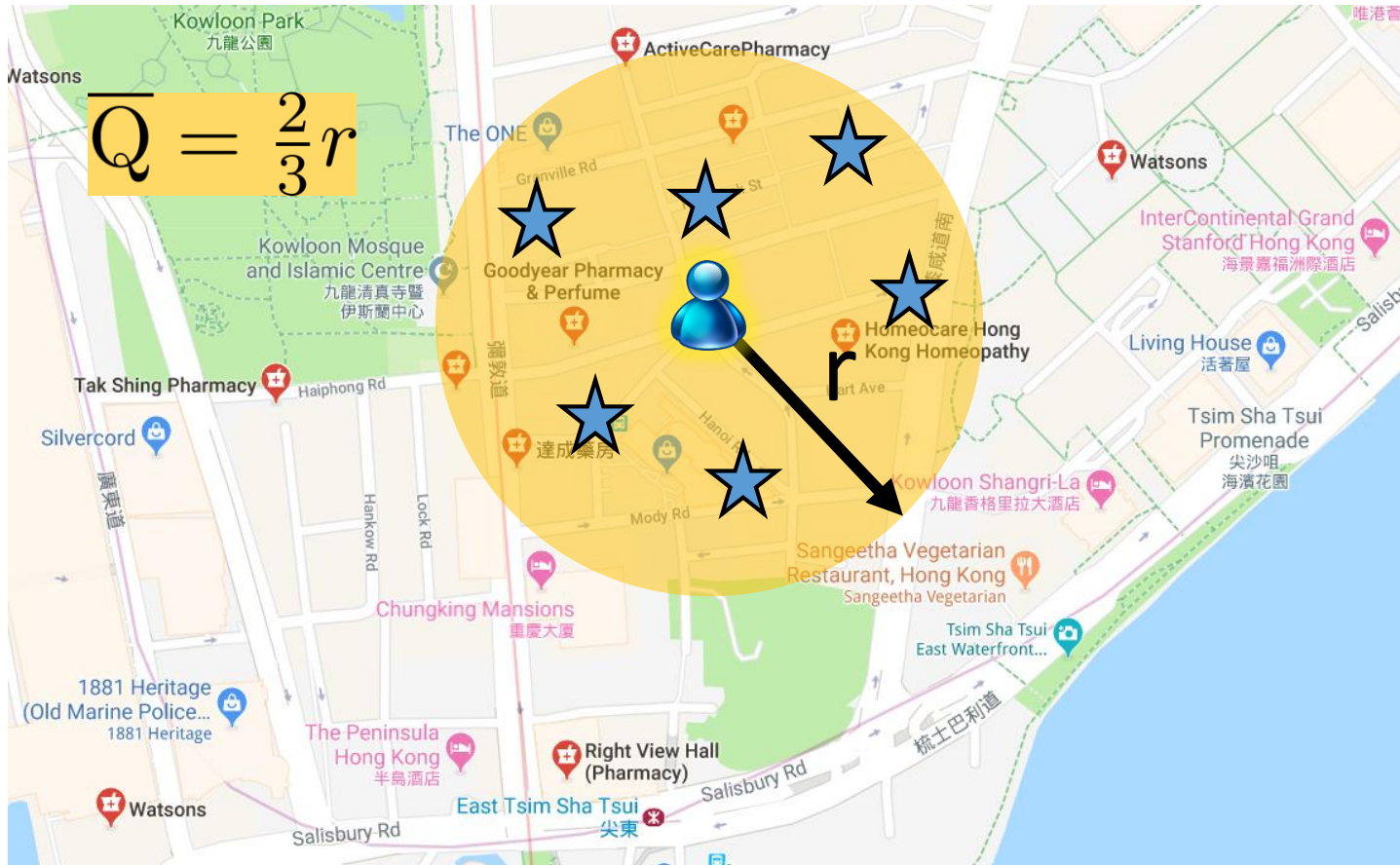
↓    ↓

# Measuring Utility: Typical Choices



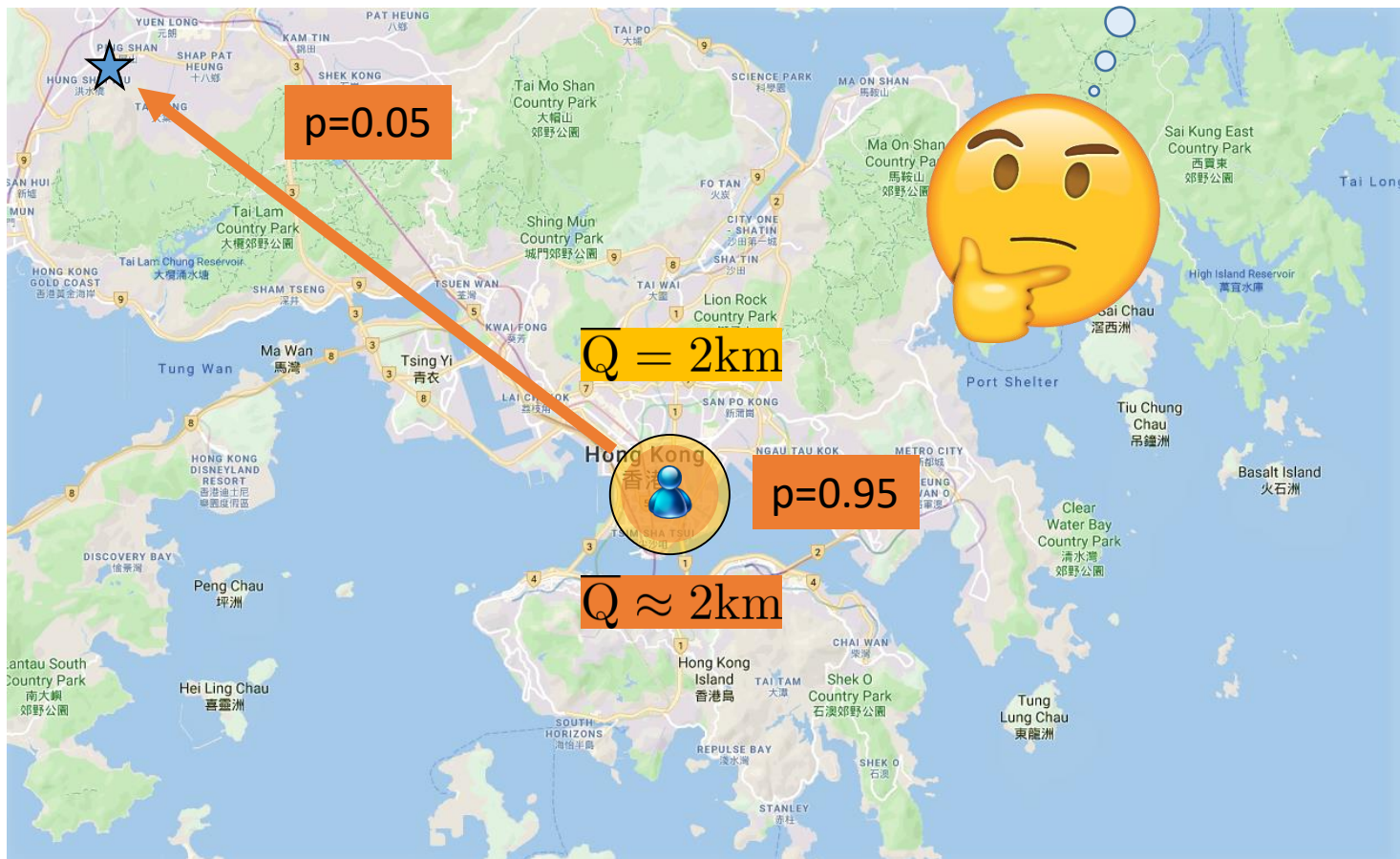


# Average Quality Loss (example)



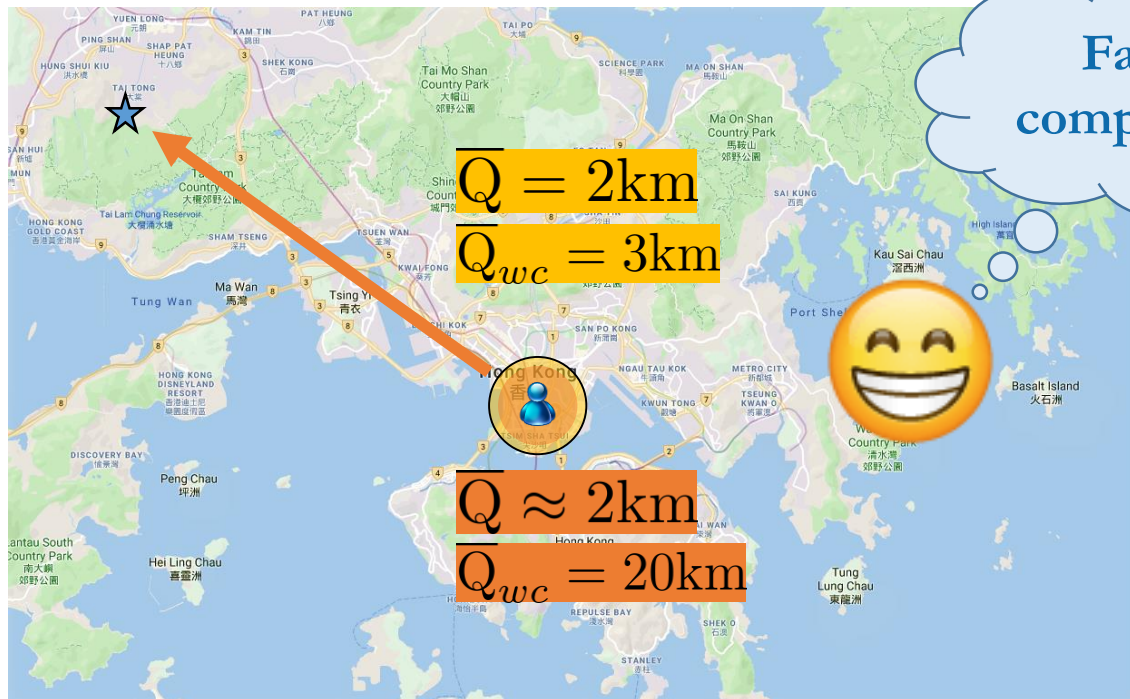
The average loss is great!

or maybe it's not...



# Worst-Case Quality Loss

- Another utility metric:  $\bar{Q}_{wc} \doteq \max_{x,z} d_Q(x,z)$



# Quantifying Privacy

- Privacy... against what/who?
- Shannon's maxim:  
"the enemy knows the system"

$$f(z^r | \mathbf{x}^r, \mathbf{z}^{r-1})$$

$$z^r \quad \mathbf{z}^{r-1}$$

Mobility profile:  $\pi(\mathbf{x}^r)$

- How do we quantify privacy?

An adversary




Wants to learn:  
 $x^r$  or  $\mathbf{x}^r$

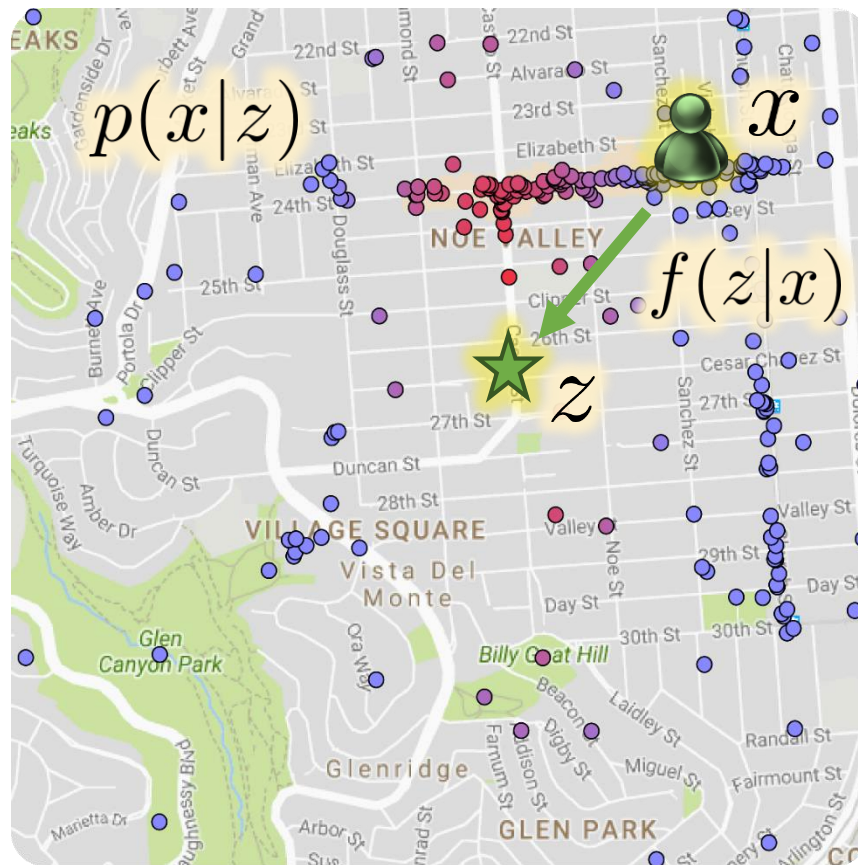
# Optimal Adversary's Attack. Computing the Posterior.



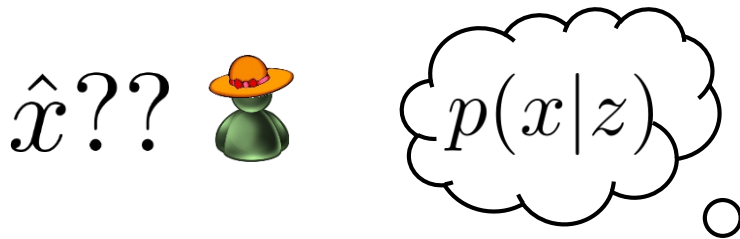
$z$  ★  
 $\pi(x)$   
 $f(z|x)$


 $\hat{x}??$

$$p(x|z) = \frac{\pi(x) \cdot f(z|x)}{\sum_{x \in \mathcal{X}} \pi(x) \cdot f(z|x)}$$



# Optimal Adversary's Attack: $h(\hat{x}|z)$

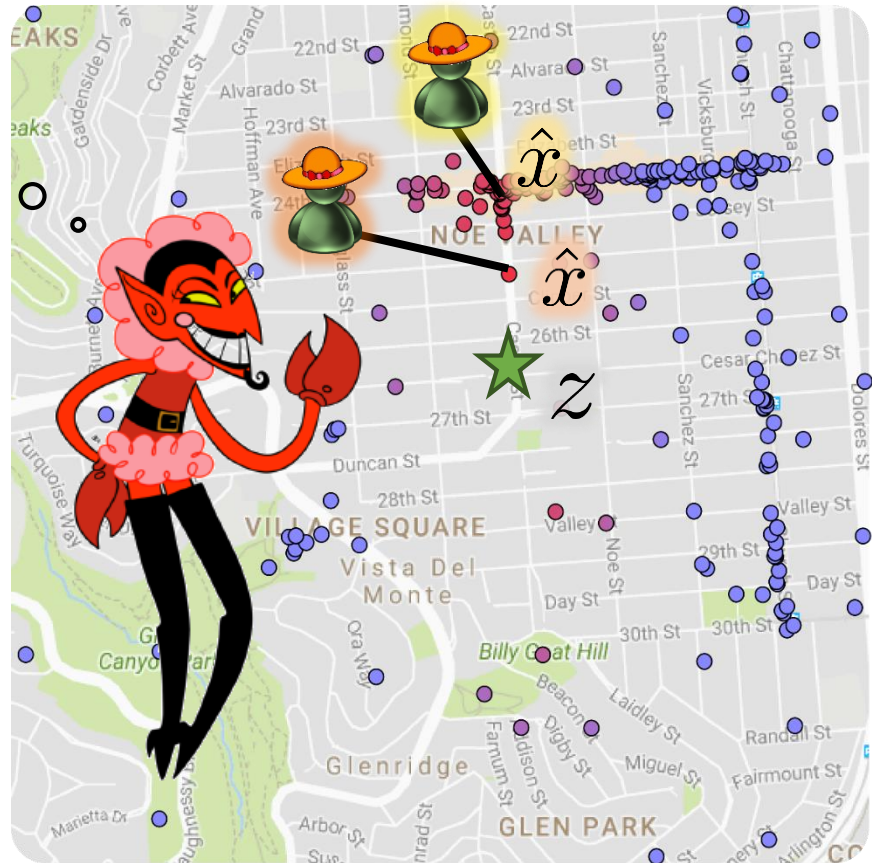


If the adversary just cares about getting the real location right:

$$\hat{x} = \arg \max_x p(x|z)$$

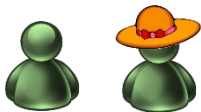
If the adversary wants to get as close as possible to the user on average:

$$\hat{x} = \arg \min_x p(x|z) \|x - z\|_2$$



... more general:

$$d_P(x, \hat{x})$$

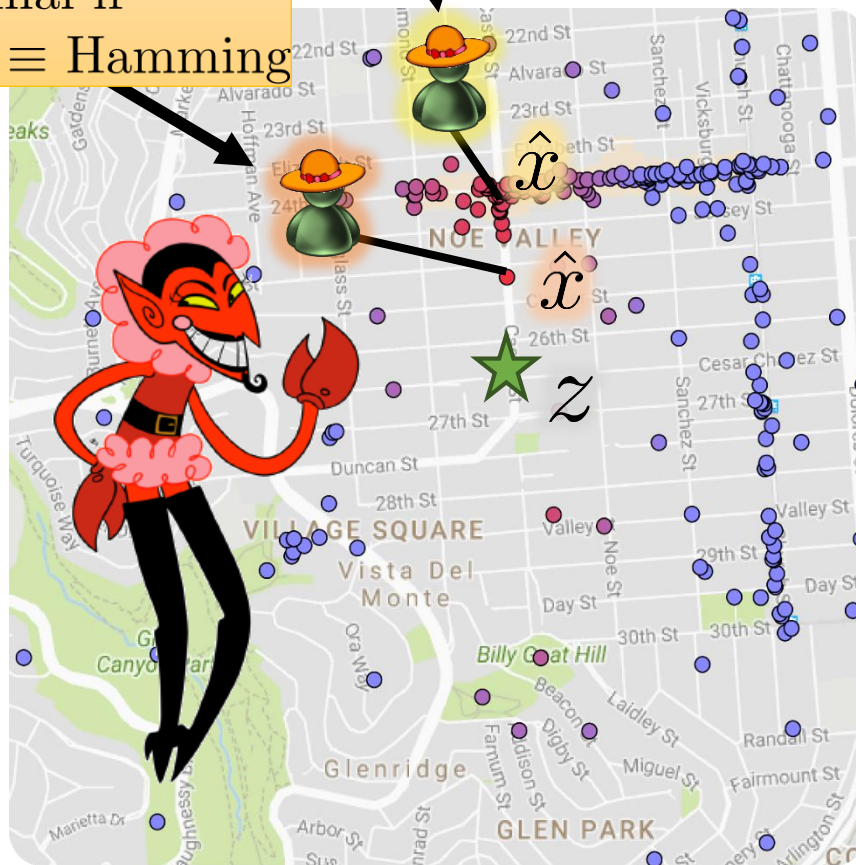


$$\hat{x} = \arg \min p(x|z) \cdot d_P(x, \hat{x})$$

$d_P(x, \hat{x}) \equiv$  Manhattan distance  
Semantic distance  
...

Optimal if  
 $d_P(x, \hat{x}) \equiv$  Euclidean

Optimal if  
 $d_P(x, \hat{x}) \equiv$  Hamming

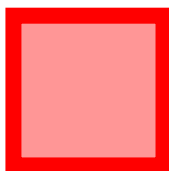


## Example of Semantic Distance:

$$d_P(\text{house}, \text{tree}) = 1 \quad d_P(\text{house}, \text{house}) = 0$$

$$\hat{x} = \arg \min p(x|z) \cdot d_P(x, \hat{x})$$

If  $p(x|z) =$



It's as if the adversary chose her estimation in the "tag domain", instead of the location domain.



	$d_P = 0$			
	★		$d_P = 1$	

- Home
- Park
- Shop
- Café



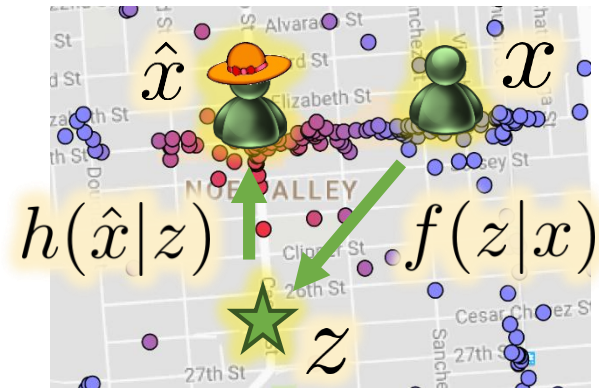
# Ok, but How do We Measure Privacy?



- Privacy is related to how good the adversary's estimation is.
- Average Adversary Error (correctness)

$$P_{\text{AE}} \doteq \mathbb{E}\{d_P(x, \hat{x})\}$$

$$P_{\text{AE}} = \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \sum_{\hat{x} \in \hat{\mathcal{X}}} \pi(x) \cdot f(z|x) \cdot h(\hat{x}|z) \cdot d_P(x, \hat{x}).$$



Typically,  
against the  
optimal attack

$$P_{\text{AE}} = \sum_{z \in \mathcal{Z}} \min_{\hat{x}} \left\{ \sum_{x \in \mathcal{X}} \pi(x) \cdot f(z|x) \cdot d_P(x, \hat{x}) \right\}.$$

Shokri, Reza, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. "Quantifying location privacy." IEEE S&P, 2011.

# Average Adversary Error (example)



$$d_Q(x, z) \equiv \text{Euclidean}$$

$$d_P(x, \hat{x}) \equiv \text{Euclidean}$$

$$d_Q(x, z_1) = 640\text{m}$$

$$d_Q(x, z_2) = 720\text{m}$$

$$d_Q(x, z_3) = 80\text{m}$$

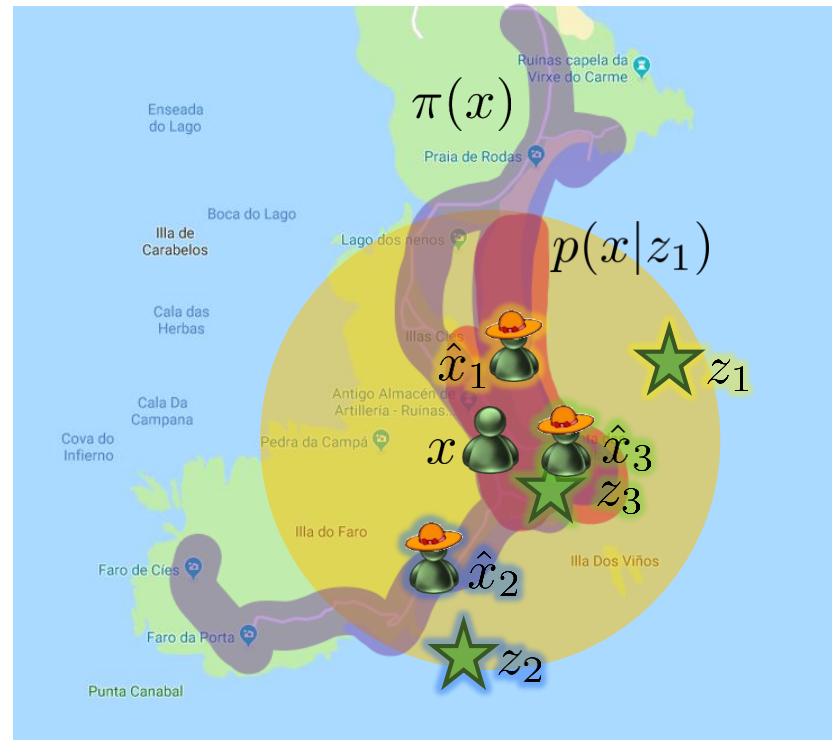
$$d_P(x, \hat{x}_1) = 210\text{m}$$

$$d_P(x, \hat{x}_2) = 350\text{m}$$

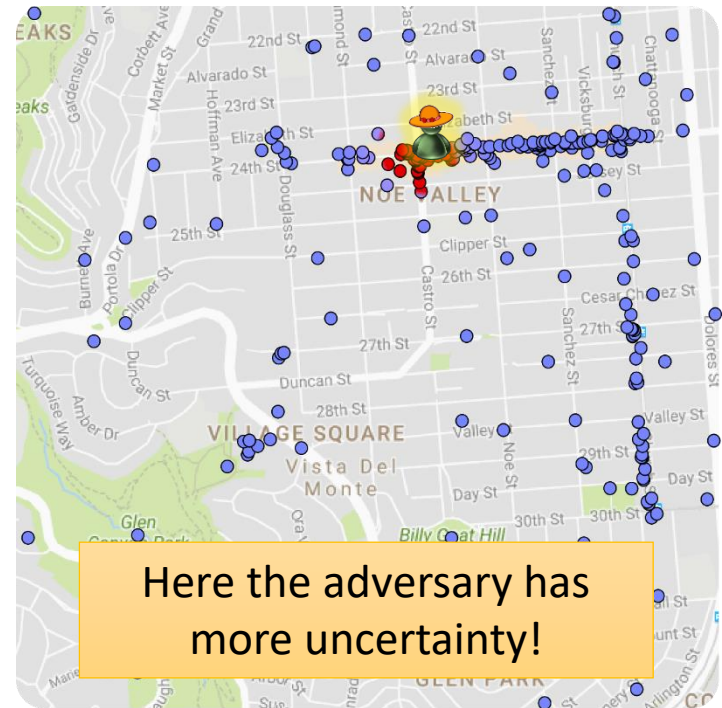
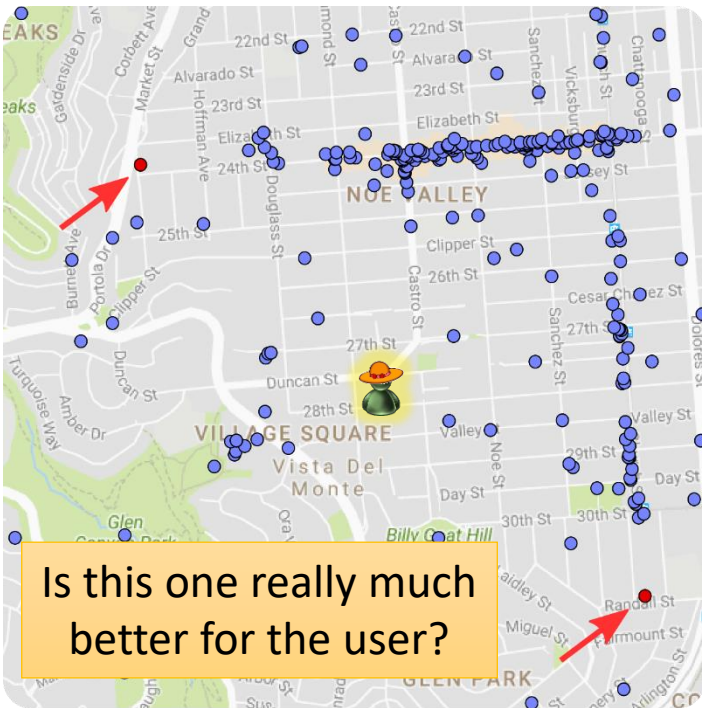
$$d_P(x, \hat{x}_3) = 90\text{m}$$

$$\bar{Q} = \mathbb{E}\{d_Q(x, z)\} = 500\text{m}$$

$$P_{\text{AE}} = \mathbb{E}\{d_P(x, \hat{x})\} = 200\text{m}$$



# The Average Adversary Error is great (?)



$P_{AE}$  ↑ ↑

The avg. error does not capture this “adversary uncertainty”

$P_{AE}$  ↓ ↓

# Conditional Entropy

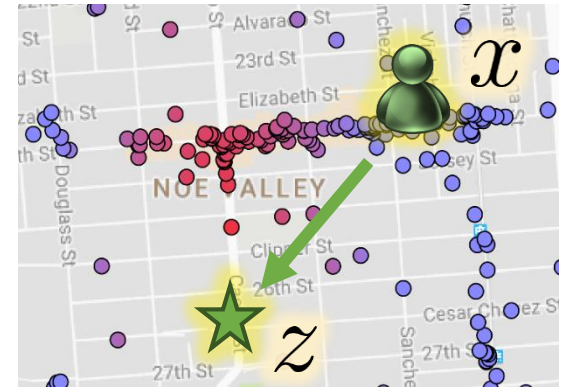


- Given the posterior:  $p(x|z)$
- Uncertainty:  $H(x|z = \star)$

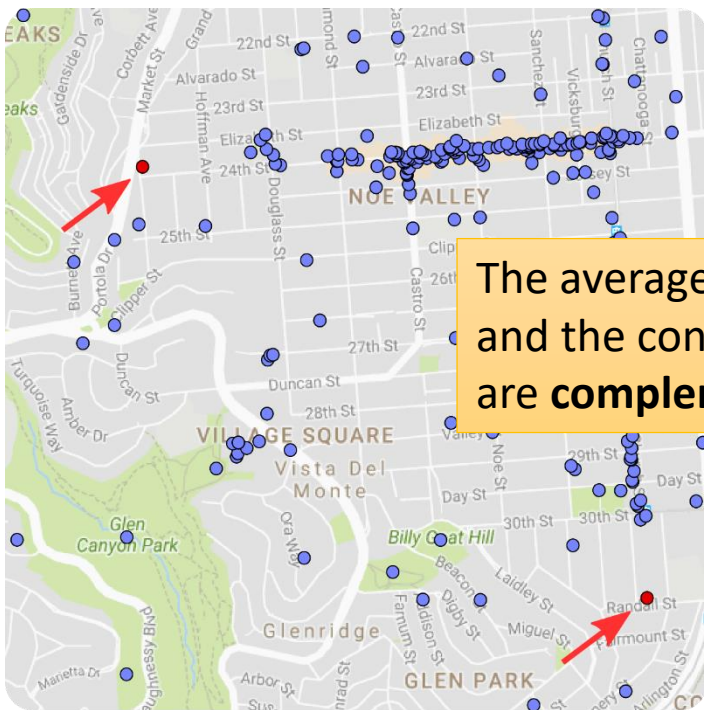
$$H(x|z) = - \sum_{x \in \mathcal{X}} p(x|z) \log p(x|z)$$

- Conditional Entropy (Average Uncertainty):

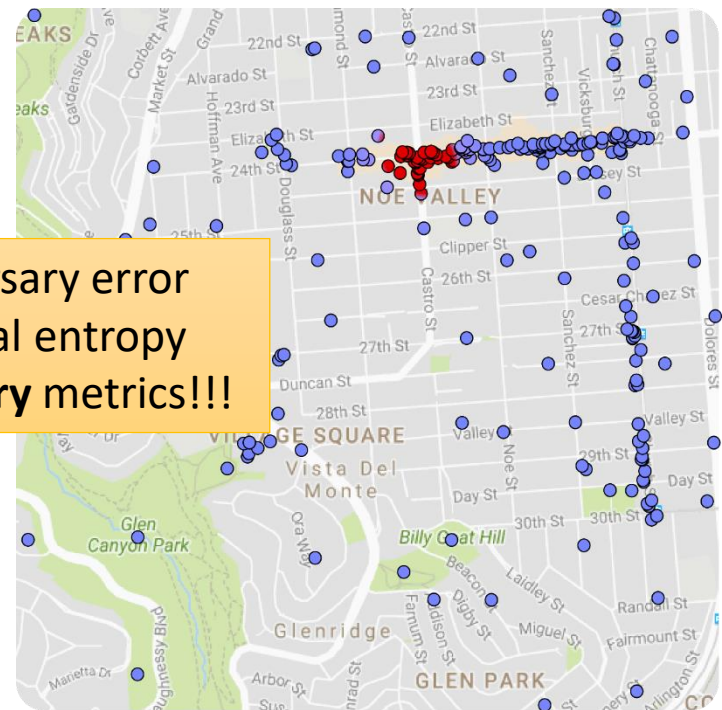
$$P_{\text{CE}} \doteq H(x|z) = \mathbb{E}\{H(x|z = \star)\}$$



# Conditional Entropy (example)



The average adversary error and the conditional entropy are **complementary metrics!!!**



$$P_{AE} \uparrow\uparrow$$

$$P_{CE} = 1 \text{ bit} \downarrow\downarrow$$

$$P_{AE} \downarrow\downarrow$$

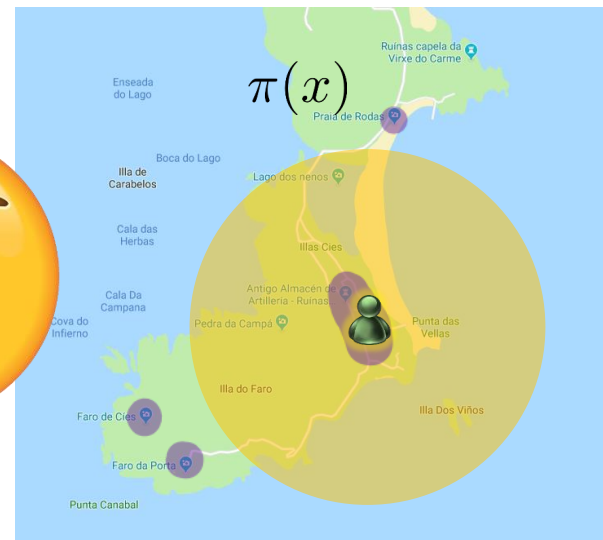
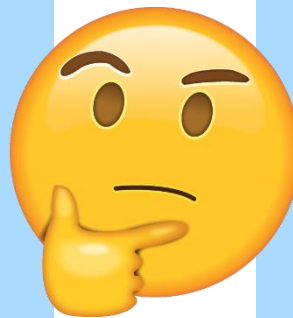
$$P_{CE} = 5 \text{ bits} \uparrow\uparrow$$

# Issues of Adversary-Tailored Metrics

- The **average error** and the **conditional entropy** assume an adversary with a certain knowledge:  $\pi(x^1, x^2, x^3, \dots)$



$$P_{AE} = E\{d_P(x, \hat{x})\} = 200m$$

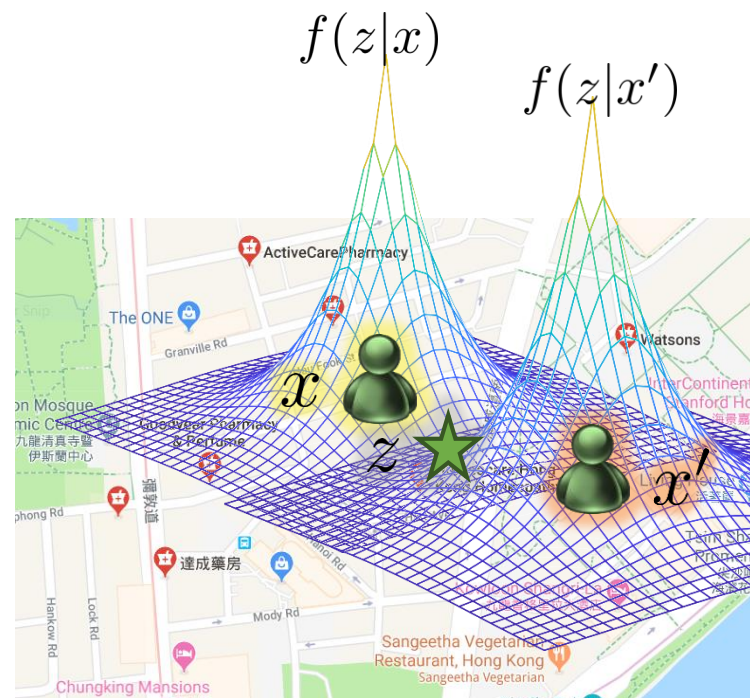
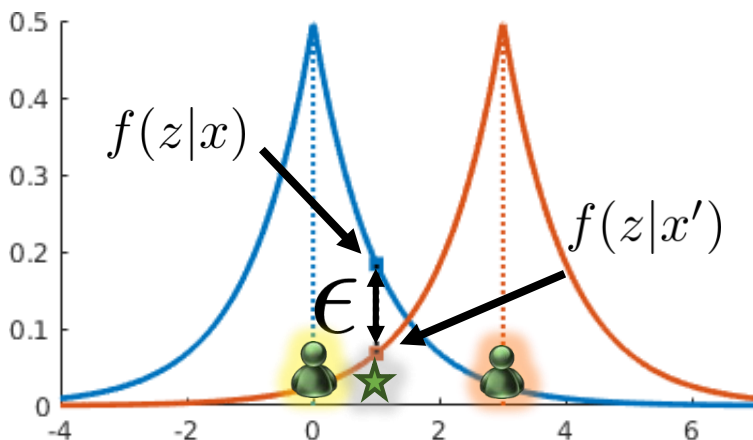


$$P_{AE} = E\{d_P(x, \hat{x})\} = 50m$$

# A Possible Solution... Differential Privacy

- Adversary-agnostic guarantee.
- Used in database privacy and other fields.
- An LPPM “ $f$ ” guarantees  $\epsilon$ -DP if the following holds:

$$f(z|x) \leq e^\epsilon \cdot f(z|x')$$



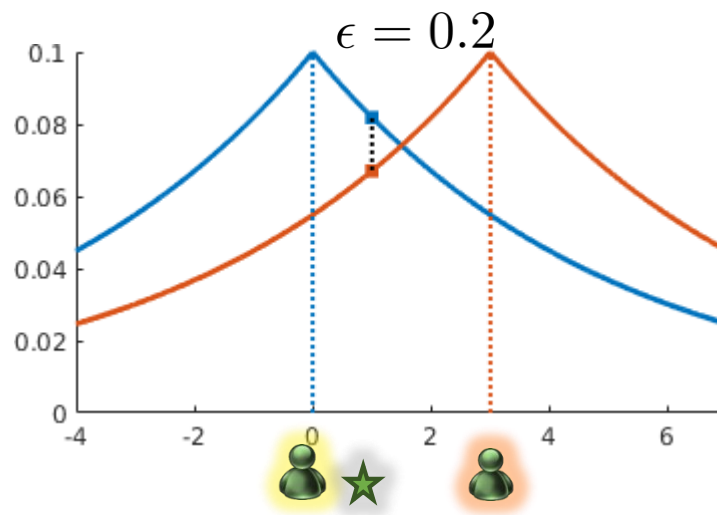
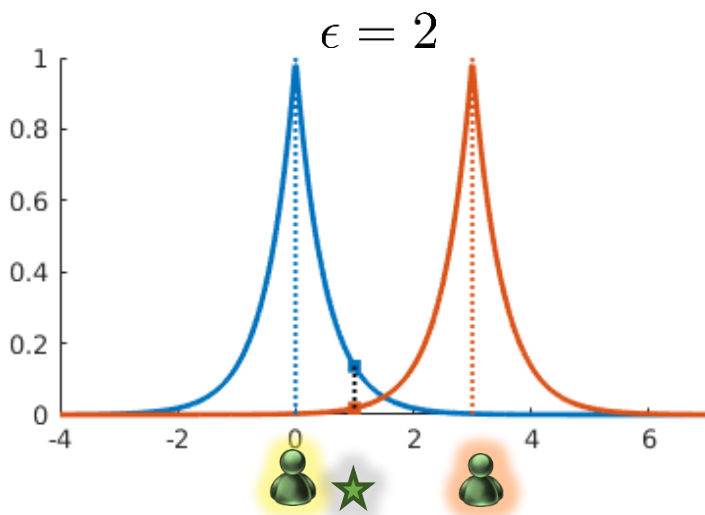
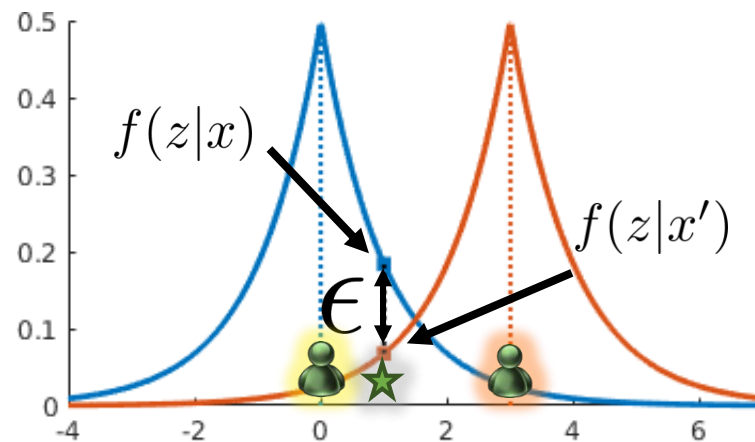
# Differential Privacy

$$f(z|x) \leq e^{\epsilon} \cdot f(z|x')$$

Privacy parameter:

$\epsilon \uparrow \uparrow$  Looser bound  $\rightarrow$  Less privacy

$\epsilon \downarrow \downarrow$  Tighter bound  $\rightarrow$  More privacy



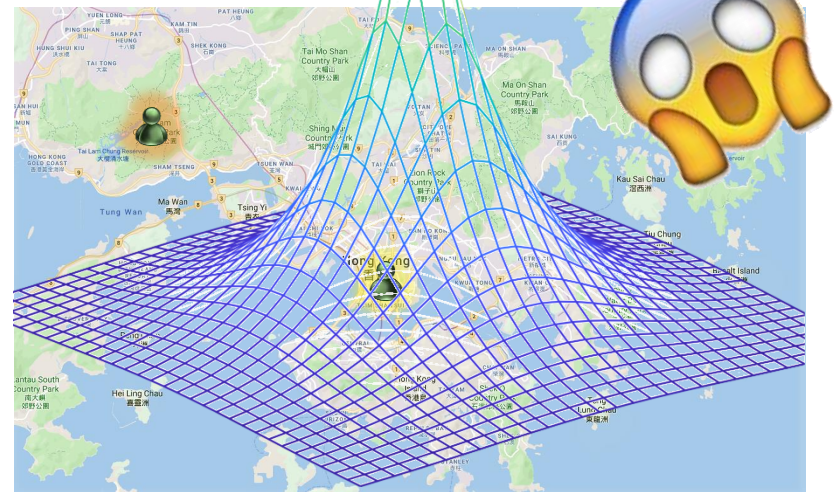
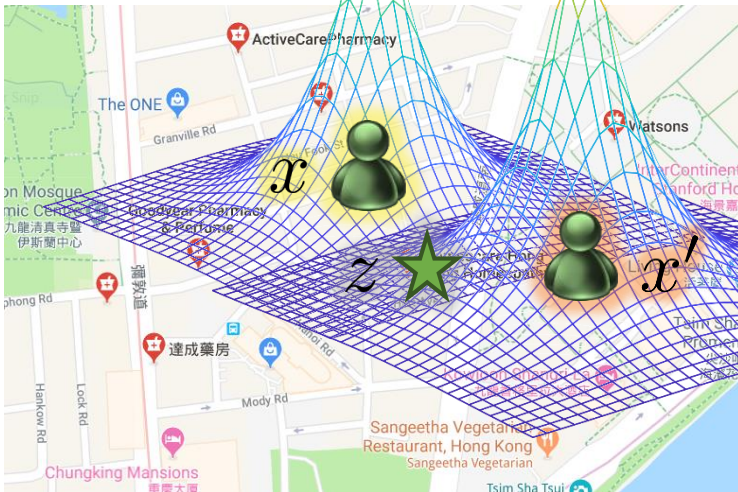


# Differential Privacy

- An LPPM “ $f$ ” guarantees  $\epsilon$ -DP if the following holds:

$$f(z|x) \leq e^\epsilon \cdot f(z|x')$$

$$\forall x, x' \in \mathcal{X} \quad \forall z \in \mathcal{Z}$$



# The Solution is... Geo-Indistinguishability!

- Extension of DP to Location Privacy:

$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x')$$

$$f(\star | \text{person}) \leq e^{\frac{\epsilon \cdot d_2(\text{person}, \text{person}')}{\epsilon'}} \cdot f(\star | \text{person}')$$

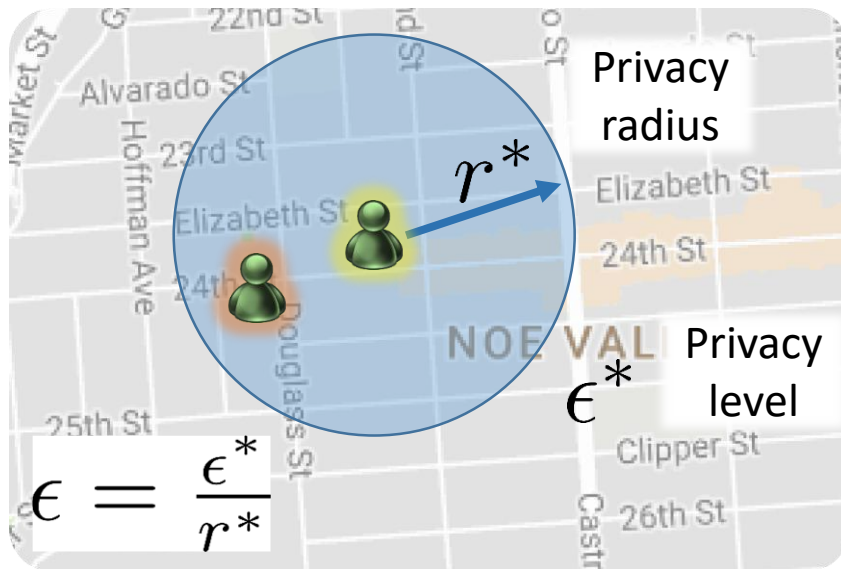
Intuition:

- If the two locations are close:  $d_2(\text{person}, \text{person}') \Downarrow\Downarrow$
- The adversary will find it **hard** to distinguish them:  $\epsilon' \Downarrow\Downarrow$
- If the two locations are far:  $d_2(\text{person}, \text{person}') \Uparrow\Uparrow$
- The adversary will find it **easy** to distinguish them:  $\epsilon' \Uparrow\Uparrow$

Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., & Palamidessi, CCS'13. Geo-indistinguishability: Differential privacy for location-based systems.

# Choosing the Privacy Level

- How do we pick  $\epsilon$ ?
- Typical approach:



- How do we choose  $\epsilon^*$ ?
  - From  $\log(1.4)$  to  $\log(10)$ .
  - Normally,  $\log(2)$ .
- Example:

$$\left. \begin{array}{l} r^* = 0.5\text{km} \\ \epsilon^* = \log(2) \end{array} \right\} \epsilon \approx 0.60\text{km}^{-1}$$

- Inside the region, we get:

$$f(\star | \text{user}) \leq 2 \cdot f(\star | \text{user})$$

Hard to interpret

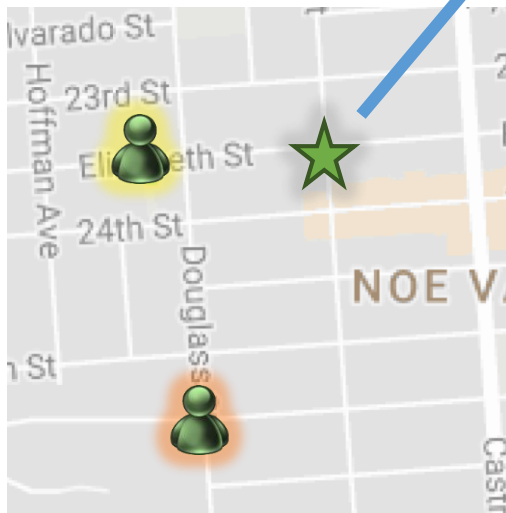
$$d_2(\text{user}, \text{user}) \leq r^* \Rightarrow f(\star | \text{user}) \leq e^{\epsilon^*} \cdot f(\star | \text{user})$$

# Geo-Indistinguishability as an Adversary Error

- Decision adversary:

$$\pi(\text{👤}) = \pi(\text{👤}) = 0.5$$

$f$



- If  $f(\star|\text{👤}) \leq f(\star|\text{👤})$  the adversary decides: 👤

- Prob. of error:

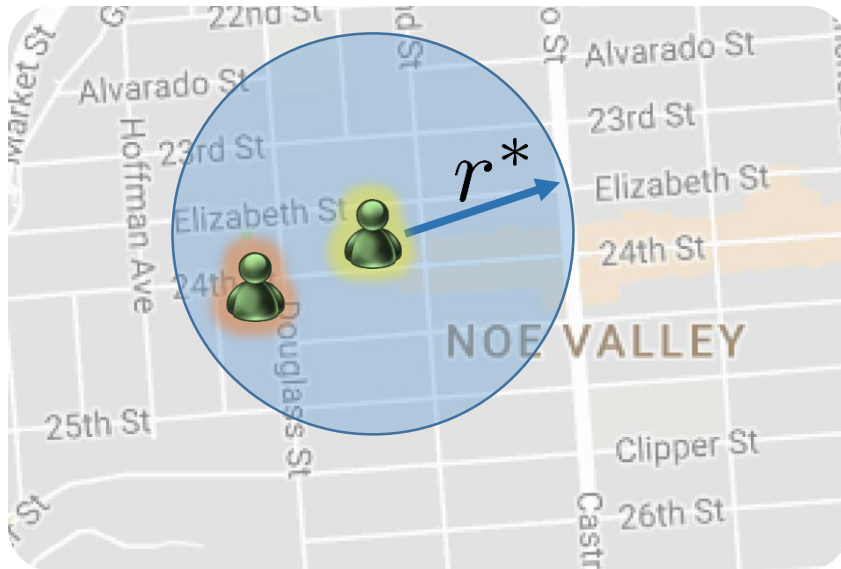
$$p_e(\text{👤}, \text{👤}, \star) = \frac{f(\star|\text{👤})}{f(\star|\text{👤}) + f(\star|\text{👤})}$$

$f$  gives geo-indistinguishability if and only if,

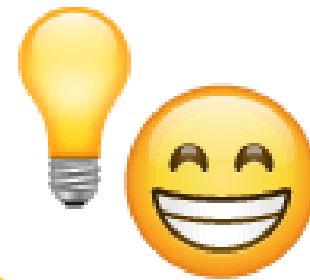
$$p_e(\text{👤}, \text{👤}, \star) \geq p_e^* = \frac{1}{1 + e^{\epsilon \cdot d_2(\text{👤}, \text{👤})}}$$

Simon Oya, Carmela Troncoso, and Fernando Pérez-González. "Is Geo-Indistinguishability What You Are Looking for?." WPES'17

# Geo-Indistinguishability as an Adversary Error



$$p_e(\text{person}, \text{person}, \text{star}) \geq 0.33$$



Easier to interpret

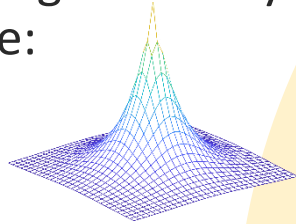
$$\epsilon \approx 0.60\text{km}^{-1}$$

$$f(\text{star} | \text{person}) \leq 2 \cdot f(\text{star} | \text{person})$$



# Geo-Indistinguishability in Numbers

- Most used geo-indistinguishability LPPM: Laplacian noise:



- Example: we want  $p_e \geq 0.4$  for locations inside a region  $r^*$ .

$$\bar{r} \approx 5r^* \quad r_{95} \approx 12r^*$$

The price we pay is too high  
for the privacy we get!!  
Bad privacy-utility trade-off



Reported location  
here on average

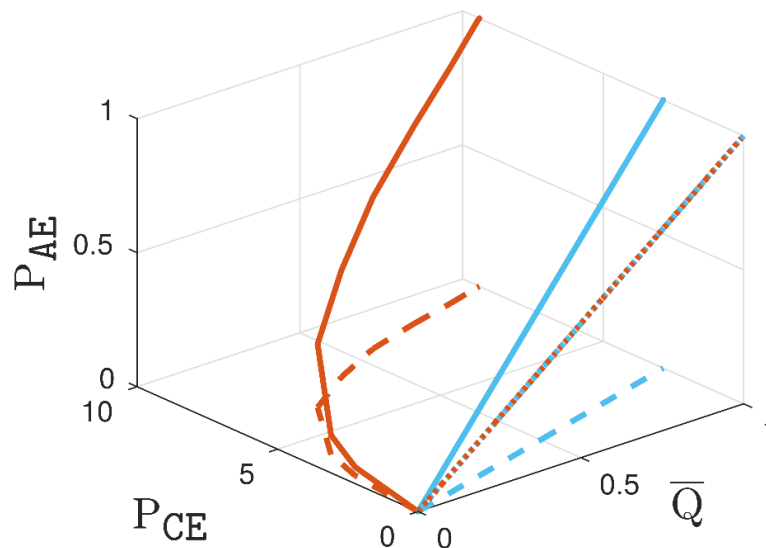
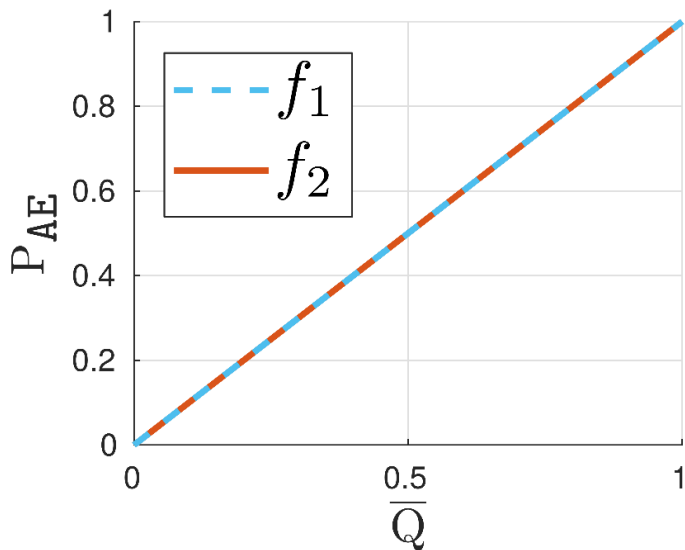
Reported location 95%  
of the time is here

# Quantifying LPPM Performance (Summary)

	Pros	Cons
Average Quality Loss	<ul style="list-style-type: none"> <li>• Intuitive</li> <li>• Versatile</li> </ul> $d_Q$	<ul style="list-style-type: none"> <li>• Average only</li> </ul>
Average Adv. Error	<ul style="list-style-type: none"> <li>• Intuitive</li> <li>• Versatile</li> </ul> $d_P$	<ul style="list-style-type: none"> <li>• Average only</li> <li>• Adversary-dependent</li> </ul>
Conditional Entropy	<ul style="list-style-type: none"> <li>• Intuitive</li> <li>• Probabilistic (non-geographic)</li> </ul>	<ul style="list-style-type: none"> <li>• Average notion</li> <li>• Adversary-dependent</li> </ul>
Geo-indistinguishability	<ul style="list-style-type: none"> <li>• Adversary-agnostic</li> </ul>	<ul style="list-style-type: none"> <li>• Not intuitive</li> <li>• Numerical issues in the user-centric approach</li> <li>• Degrades with further location reports</li> </ul>

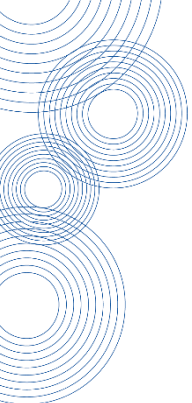
# Conclusions

- There is no universal notion of privacy.
- Privacy is a multi-dimensional notion.



- Privacy and utility are subjective and application-dependent.





# LPPM Design and Evaluation

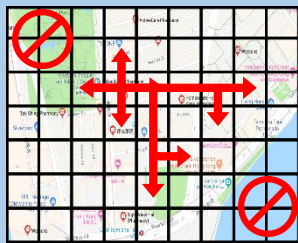
# Location Privacy-Preserving Mechanism (LPPM) Design

Privacy and quality loss requirements

maximize  $P$   
 subject to  $Q \leq Q_{\max}$

Mobility model

$$\pi(x^1, x^2, x^3, \dots)$$

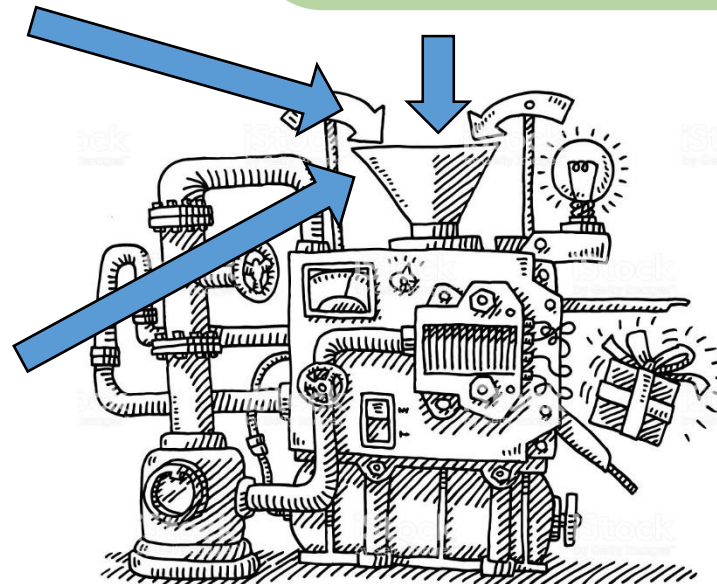


Application requirements



$$x \in \mathcal{X}$$

$$z \in \mathcal{Z}$$



LPPM  
 $f$

# Traditional Approach: Average Adv. Error vs Average Loss

Privacy and quality  
loss requirements

$$\begin{aligned} &\text{maximize} && P_{\text{AE}} \\ &\text{subject to} && \overline{Q} \leq \overline{Q}_{\text{max}} \end{aligned}$$

Mobility model & Application reqs.

$$\pi(x^1, x^2, \dots) = \pi(x^1)\pi(x^2) \dots$$

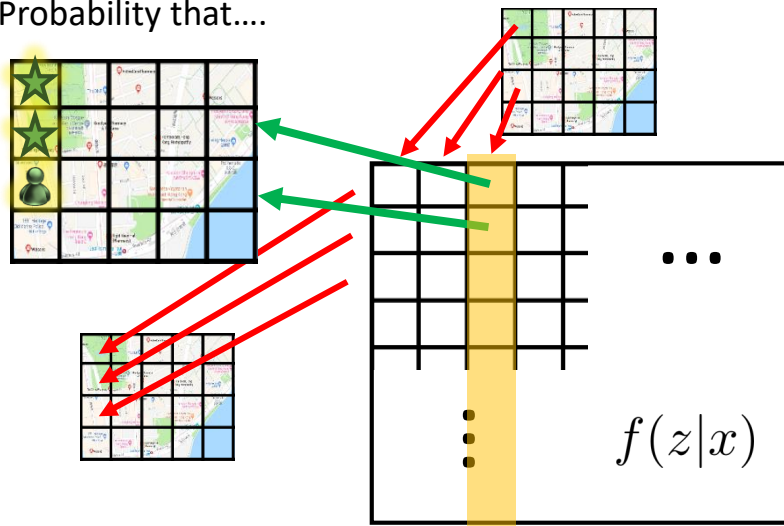


$$\pi(x)$$



**Theorem:** if the mobility model is sporadic, we can design  $f(z^r | z^{r-1}, \mathbf{x}^r)$  as  $f(z^r | x^r)$  and we do not lose privacy.

Probability that....

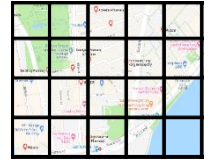


Adds up to 1

# Traditional Approach: Average Adv. Error vs Average Loss

$$\begin{aligned} & \text{maximize} && P_{\text{AE}} \\ & \text{subject to} && \bar{Q} \leq \bar{Q}_{\text{max}} \end{aligned}$$

$$P_{\text{AE}} = \sum_{z \in \mathcal{Z}} \min_{\hat{x}} \left\{ \sum_{x \in \mathcal{X}} \pi(x) \cdot f(z|x) \cdot d_P(x, \hat{x}) \right\}.$$



$N$

$$\begin{aligned} & \text{maximize}_{f(z|x), p_z} && \sum_{z \in \mathcal{Z}} p_z \\ & \text{s.t.} && p_z \leq \sum_{x \in \mathcal{X}} \pi(x) \cdot f(z|x) \cdot d_P(x, \hat{x}), \quad \forall z, \hat{x} \end{aligned}$$

$$\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x) \cdot f(z|x) \cdot d_Q(x, z) \leq \bar{Q}_{\text{max}},$$

$$\sum_{z \in \mathcal{Z}} f(z|x) = 1, \quad \forall x$$

$$f(z|x) \geq 0, \quad \forall x, z$$

$p_z$

$N^2 + N$  variables

$N^2$  constraints

1 constraint

$N$  constraints

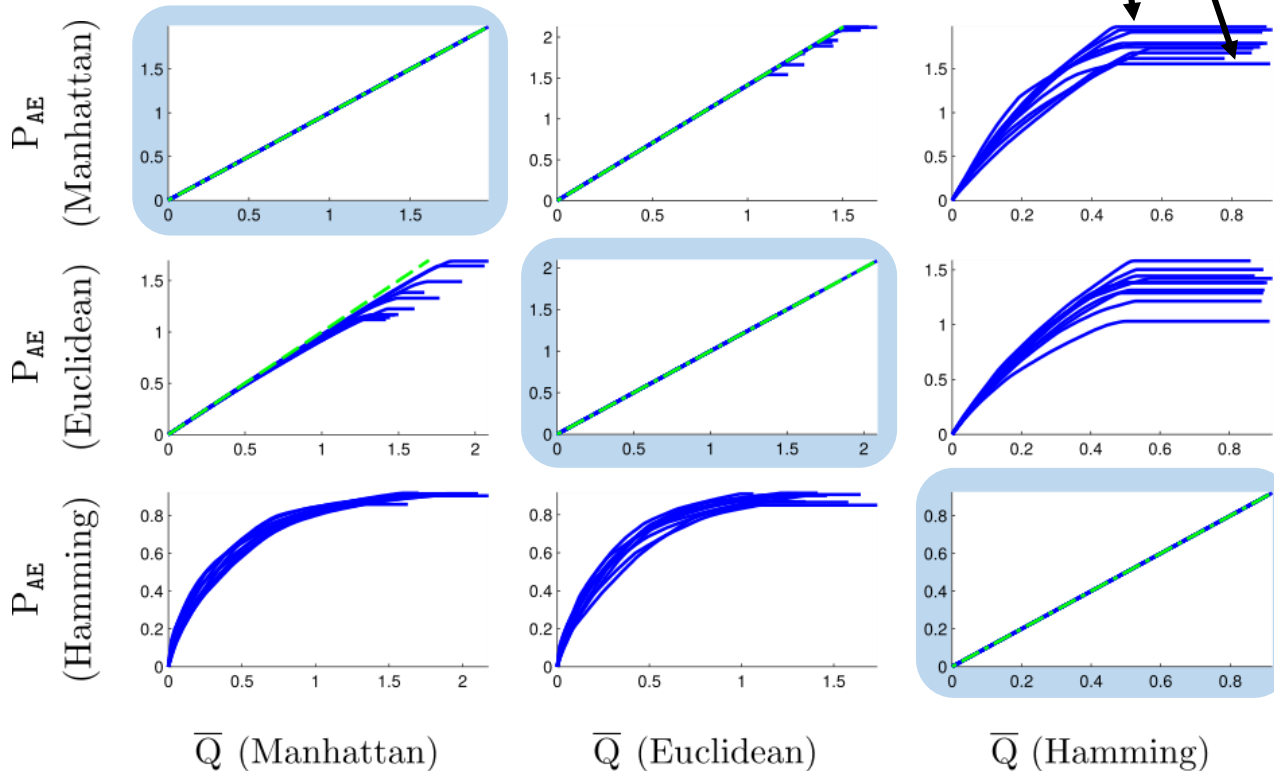
$N^2$  bounds



$$\begin{aligned} & \text{maximize} && P_{\text{AE}} \\ & \text{subject to} && \bar{Q} \leq \bar{Q}_{\text{max}} \end{aligned}$$

# Optimal Performance

- What does privacy vs. utility look like?
- Toy example:



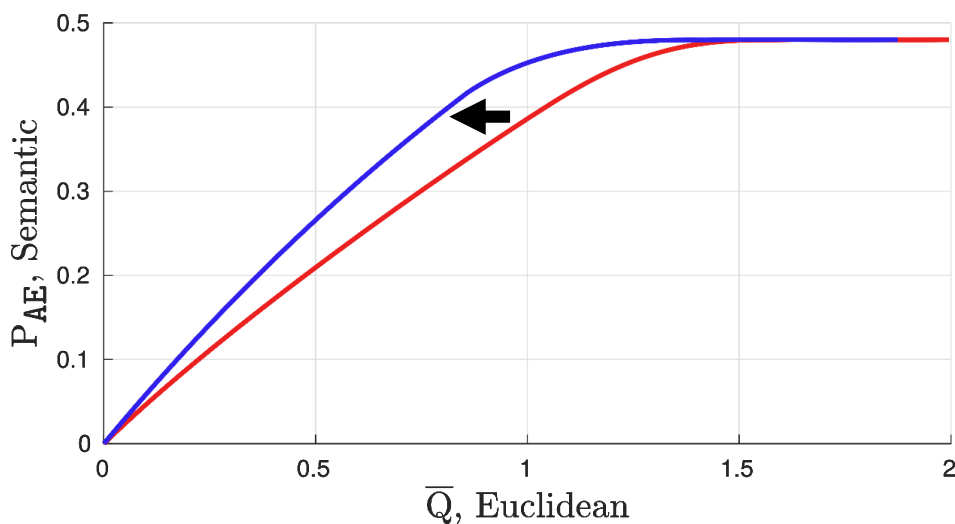
- It depends on...
- User mobility
  - Terrain
  - Application
  - Privacy and utility metrics
  - ...

If  $d_Q \equiv d_P$ , then

$$P_{\text{AE}} = \bar{Q}$$

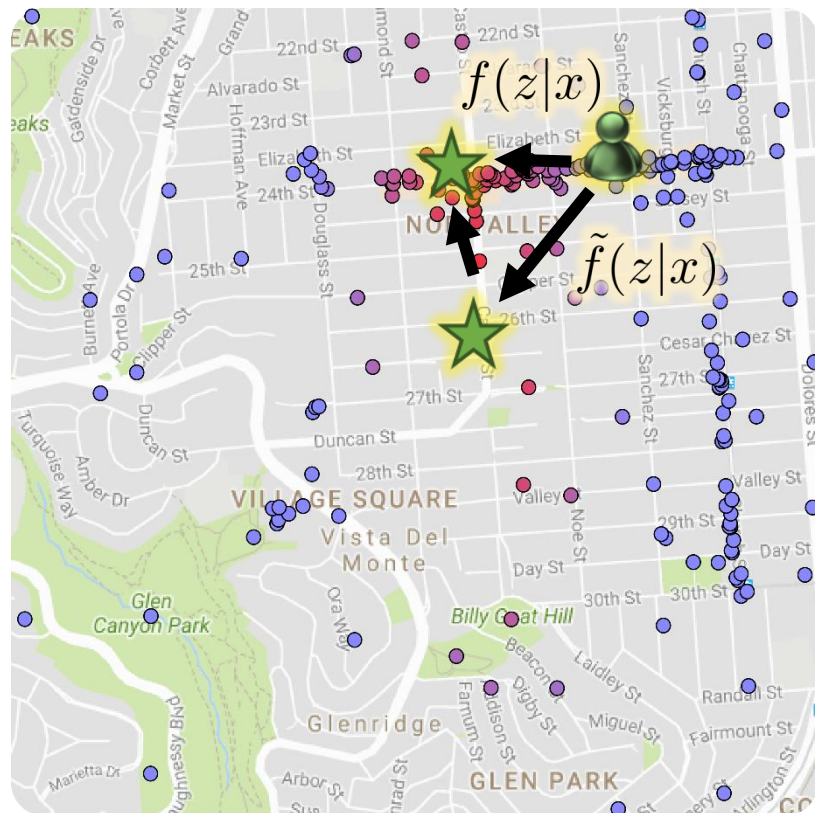
# Optimal Remapping

$$\bar{Q} = \mathbb{E}\{d_Q(x, z)\} \quad P_{AE} = \mathbb{E}\{d_P(x, \hat{x})\}$$



**Theorem:** optimal remappings do not reduce (any) privacy metric.

- Average Error, Conditional Entropy, Geo-Ind.



Chatzikokolakis, Konstantinos, Ehab Elsalamouny, and Catuscia Palamidessi. "Efficient utility improvement for location privacy." *PoPETS'17*.

# Optimal Remapping LPPMs are Optimal!! (if $d_Q \equiv d_P$ )

Proof:

$$\bar{Q} = \mathbb{E}\{d_Q(x, z)\} \quad P_{\text{AE}} = \mathbb{E}\{d_P(x, \hat{x})\}$$

- The attack that “does nothing”,

$$h^*(\hat{x}|z) = 1 \text{ if } \hat{x} = z$$

gives  $P_{\text{AE}}(h^*) = \bar{Q}$ .

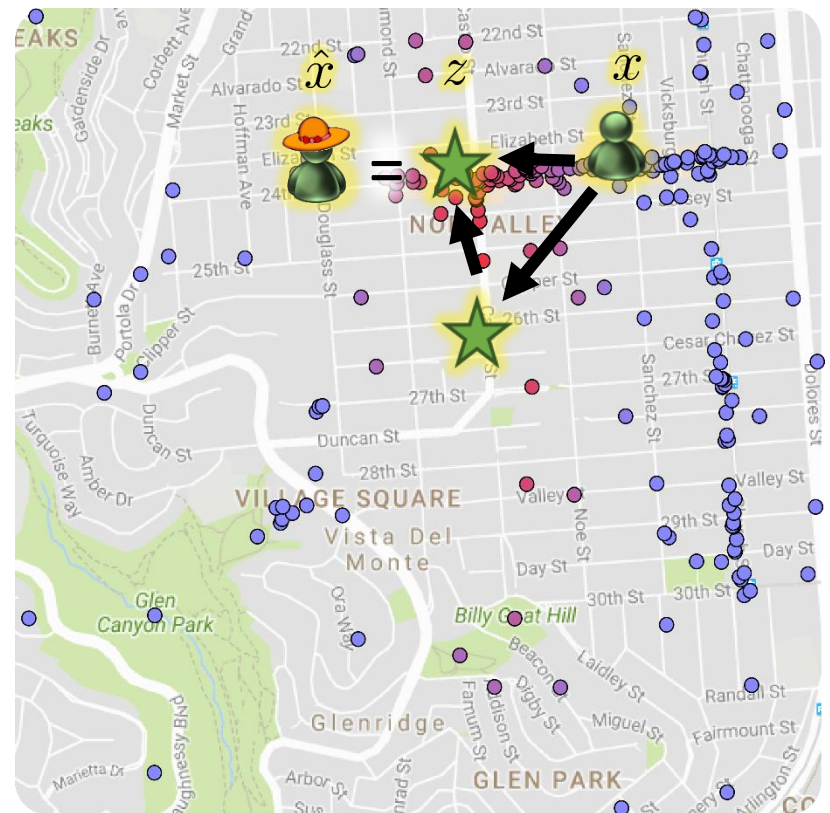
- Therefore, against an optimal attack,

$$P_{\text{AE}} \leq \bar{Q}$$

- What is the optimal attack against these LPPMs?

“do-nothing”

- We have reached the upper bound, and thus optimal remapping LPPMs are optimal in terms of privacy:  $P_{\text{AE}} = \bar{Q}$



# Traditional Approach with the Markov Model

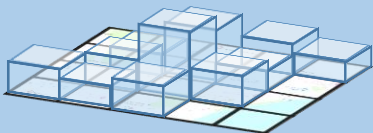
Privacy and quality  
loss requirements

maximize  $P_{AE}$   
subject to  $\bar{Q} \leq \bar{Q}_{\max}$

Mobility model & Application reqs.

$$\pi(x^1, x^2, \dots) = \pi_0(x^1)M(x^2|x^1) \dots$$

$\pi_0(x)$



$M(x^r|x^{r-1})$



- We have to take all the previous releases into account:

$$f(z^r | \mathbf{x}^r, \mathbf{z}^{r-1})$$

- We can find an optimal mechanism by solving a linear program

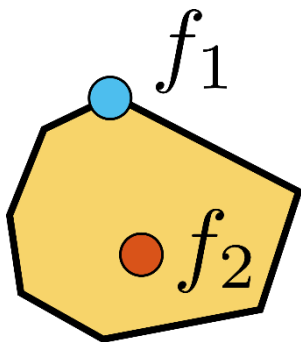
THE COMPUTATIONAL  
COST IS PROHIBITIVE

- We can use remapping techniques to find optimal mechanisms in the Markov model if  $d_Q \equiv d_P$ .



## There are Infinite Optimal Mechanisms

- Applying the optimal remapping to any  $\tilde{f}(z|x)$  gives an optimal mechanism.
- Solving the linear program with different algorithms gives us different LPPMs.
- Optimal LPPMs are in a polytope:

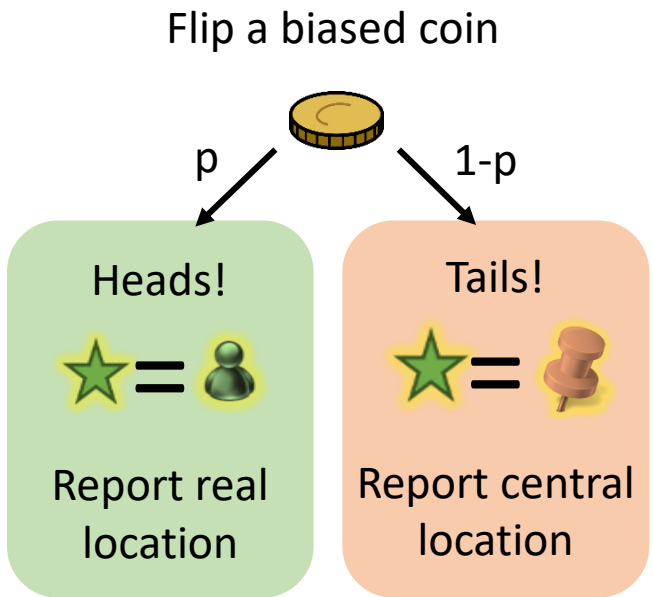


- Are all optimal LPPMs as “good”?
- Let’s study one:



Simon Oya, Carmela Troncoso, and Fernando Pérez-González. "Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms." *CCS'17*.

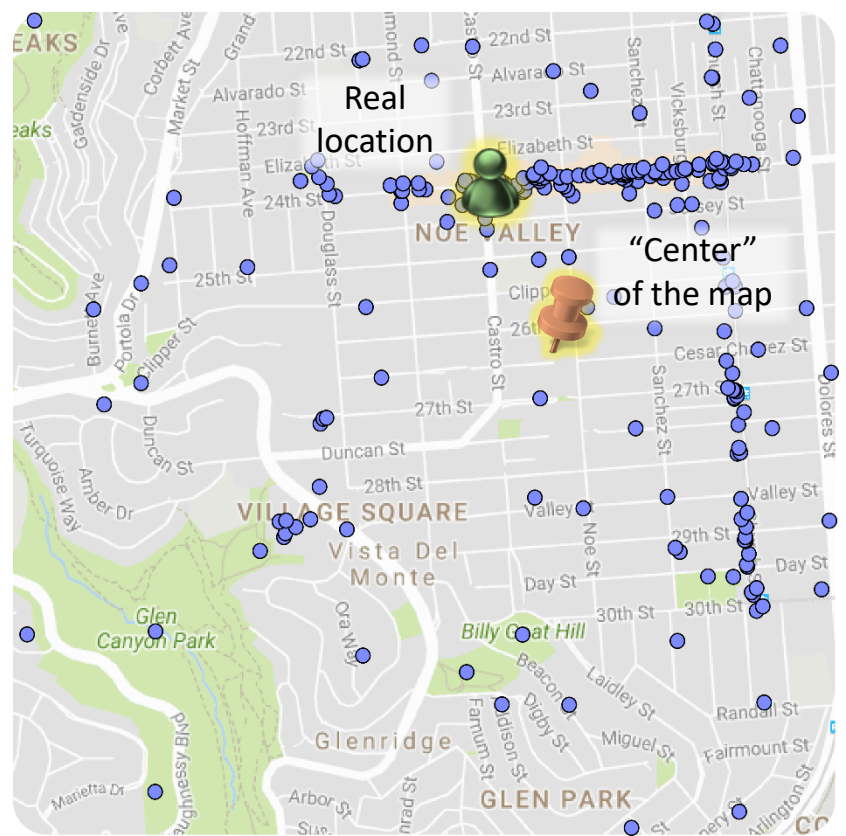
# The Coin Mechanism (also called Location Hiding)



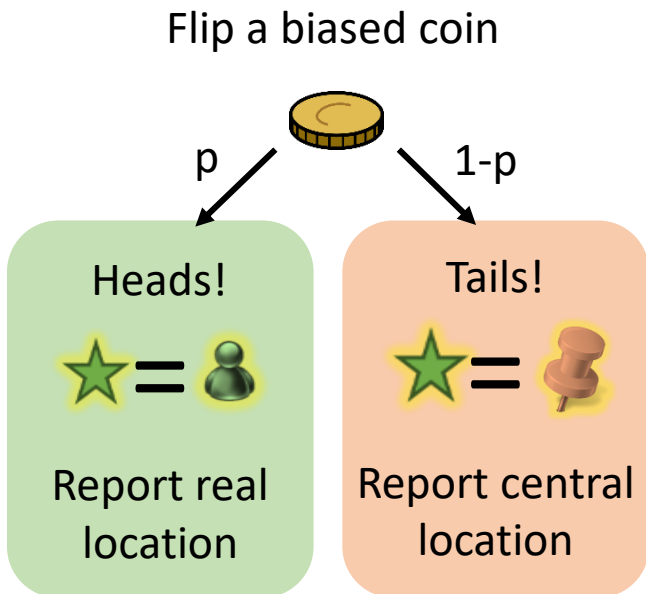
How “good” is this mechanism?

No privacy!

Seems OK...



# The Coin Mechanism (also called Location Hiding)



How "good" is this mechanism?

No privacy!

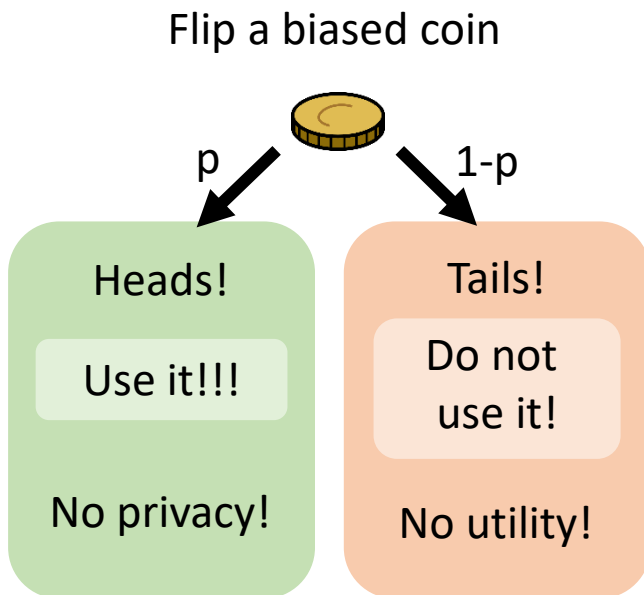
~~Seems OK...~~

No utility!

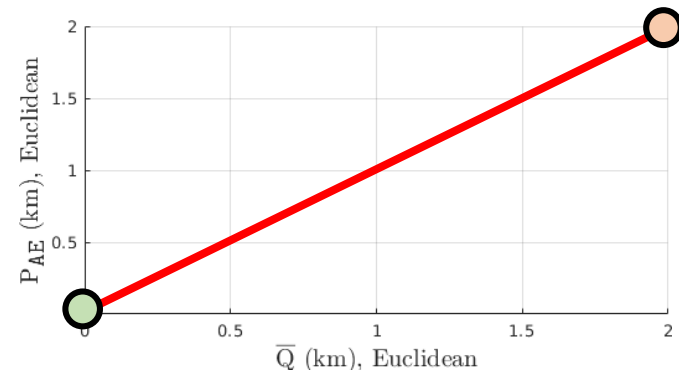


# The Coin Mechanism

- You can use this right now on your phone!!
- Whenever you want to use a location-based service...



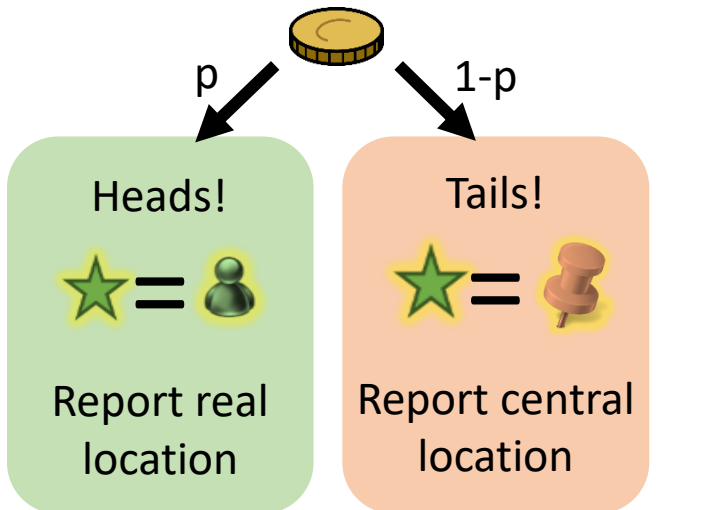
- This mechanism is optimal in terms of  $P_{AE}$  vs.  $\bar{Q}$ .



- Yet it does not seem very “desirable”.
- Where’s the problem?

# The Coin Mechanism and its Conditional Entropy

- The Coin is very “binary”. The Conditional Entropy reveals this issue.



$$H(x|z = \star) = 0$$

$$H(x|z = \star) = H(x)$$

If  $p=1$ :

$$P_{CE} = 0$$

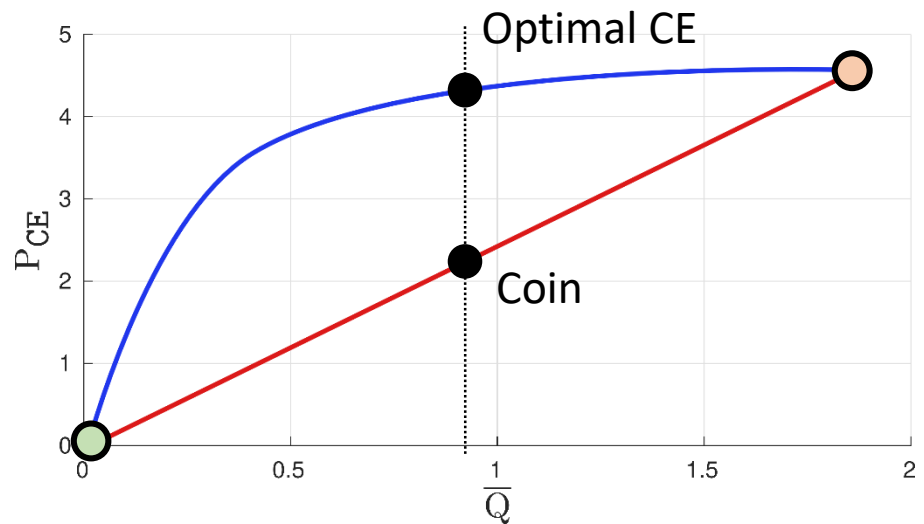
$$\bar{Q} = 0$$

If  $p=0$ :

$$P_{CE} = H(x)$$

$$\bar{Q} = \mathbb{E}\{d_Q(x, \star)\}$$

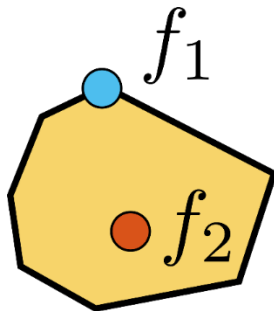
$$P_{CE} \doteq \mathbb{E}\{H(x|z = \star)\}$$




- The Coin performs poorly in terms of Conditional Entropy!

# Recap

- Optimal LPPMs in terms of  $P_{AE}$  vs.  $\bar{Q}$ :
  - Solve a linear program (expensive)
  - Optimal remapping (only if  $d_Q \equiv d_P$ )
- There are infinite optimal LPPMs:



- Careful: they might be “undesirable” 
- Use other metrics for this:
  - Conditional Entropy
  - Worst-case Quality Loss
  - ...
- **Next:**
  - LPPM design to maximize the **Conditional Entropy**.
  - LPPM design to maximize **Geo-indistinguishability**

# Maximizing the Conditional Entropy

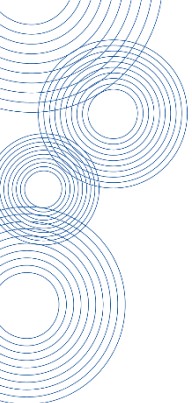
$$\begin{aligned} & \underset{f}{\text{maximize}} && P_{\text{CE}} \\ & \text{s.t.} && \bar{Q} \leq \bar{Q}_{\text{max}} \\ & && f \in \mathcal{P} \end{aligned}$$



$$\begin{aligned} & \underset{f}{\text{minimize}} && I(x; z) \\ & \text{s.t.} && \bar{Q} \leq \bar{Q}_{\text{max}} \\ & && f \in \mathcal{P} \end{aligned}$$

$$P_{\text{CE}} \doteq H(x|z) = H(x) - I(x; z)$$

↑↑      Indep. of LPPM      ↓↓

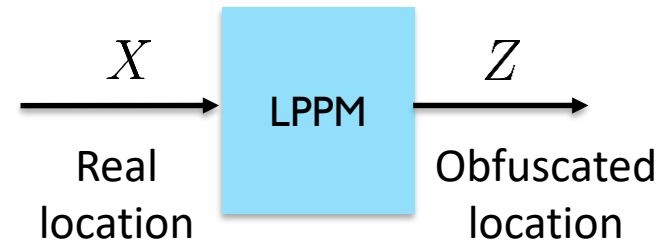


**WAIT!!!**  
This is the missing link with the strategies I covered in the first part !





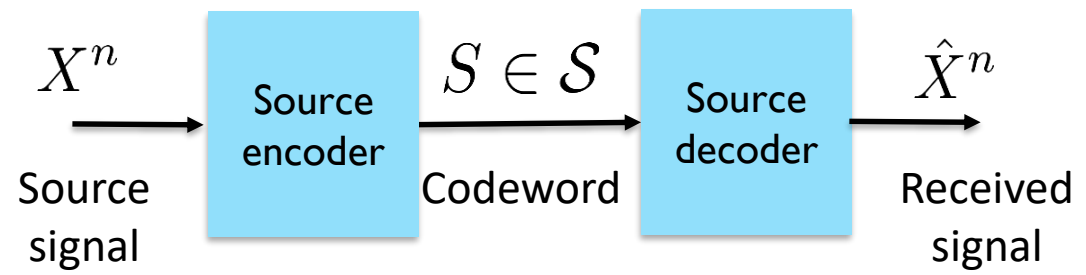
# Sounds known to me...



minimize  $I(x; z)$   
 s.t.  $\bar{Q} \leq \bar{Q}_{\max}$   
 $f \in \mathcal{P}$

**Goal:** minimize  $I(X; Z)$   
 subject to a quality loss constraint  $\bar{Q}(X, Z)$ .

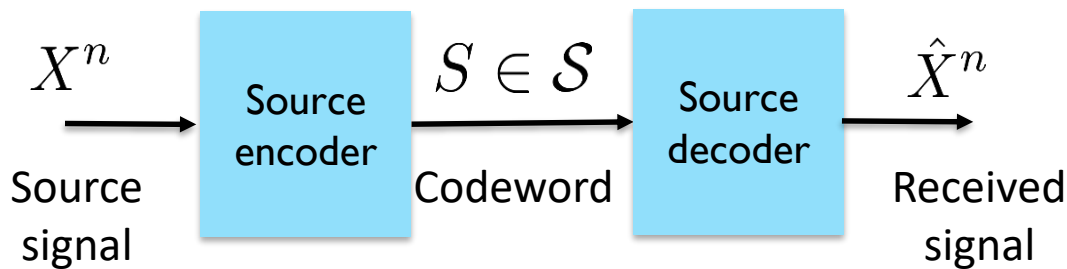
This is source coding!



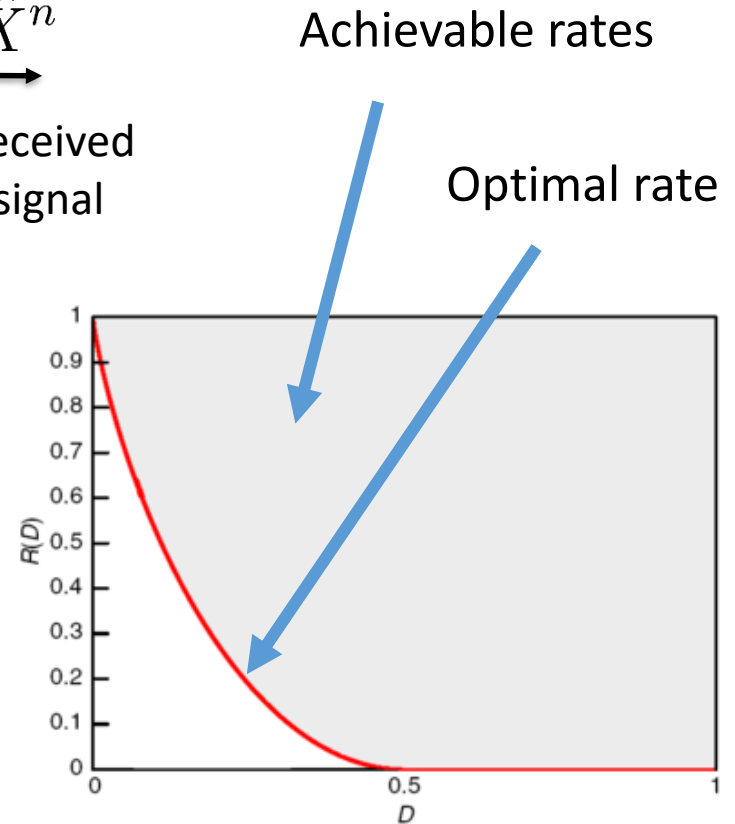
**Goal:** design codebook  $\mathcal{S}$  and mapping  $X^n \rightarrow S$  to minimize  $I(X^n; S)$  with a distortion (reconstruction) constraint  $d(X^n, \hat{X}^n) \leq nD$ .

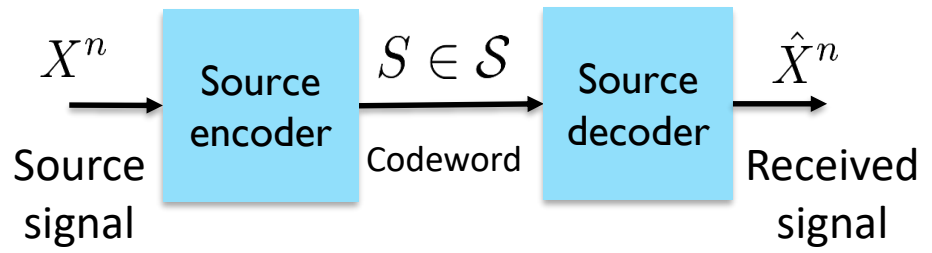


# Rate-Distortion Function



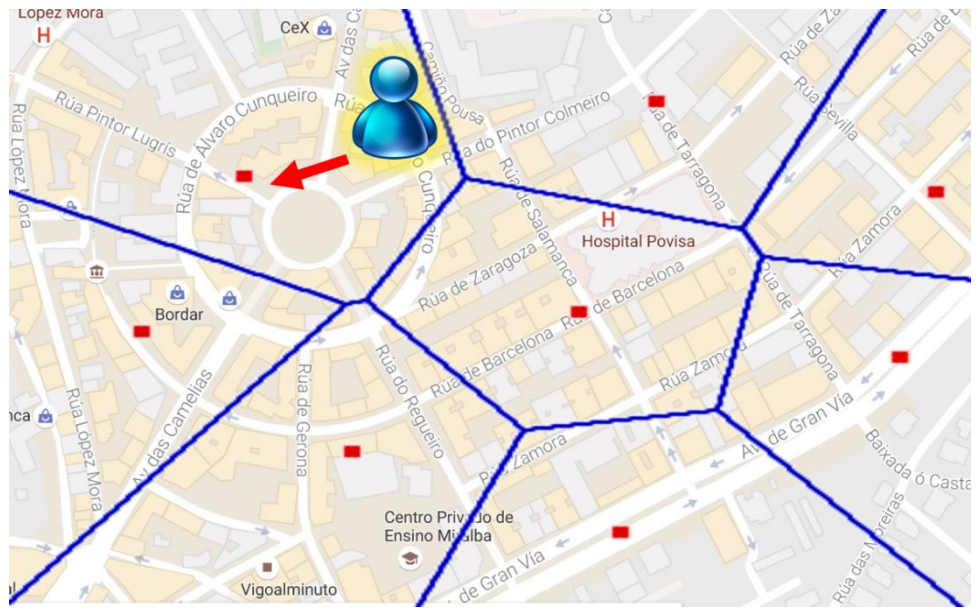
- Quantifies the **rate**, i.e., how many bits are needed (on average) to transmit a symbol, so that the source signal can be reconstructed at reception without exceeding a distortion  $D$ .
- It can be computed analytically in some cases.
- It can be computed empirically using Blahut-Arimoto algorithm.



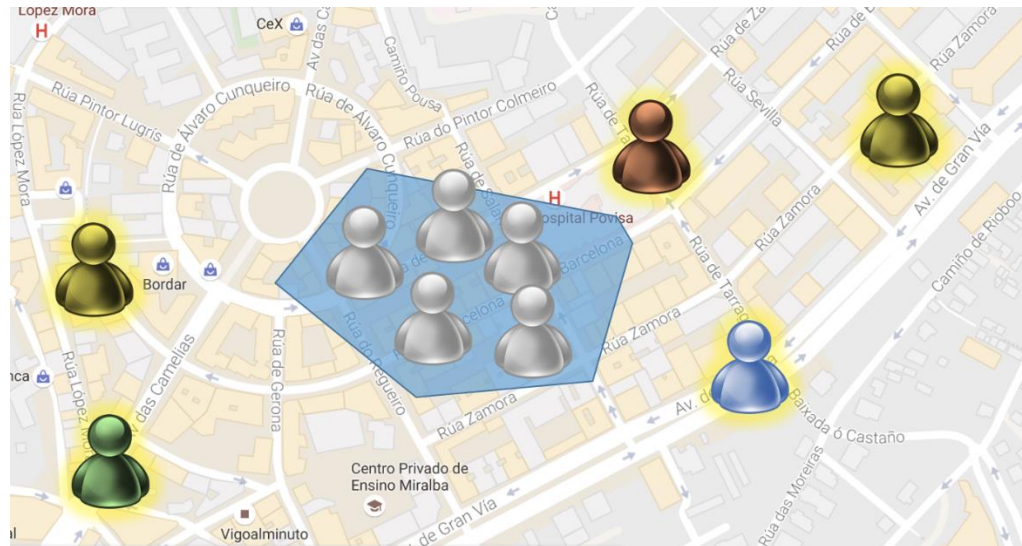


# Practical Source Coding

- Standard approach: **vector quantization** of  $X^n$ . The  $\hat{X}^n$  become the centroids. Target is to reduce  $|\mathcal{S}|$  (e.g. sphere covering) to minimize bandwidth.
- In **location privacy**, this corresponds to exactly this perturbation scheme:

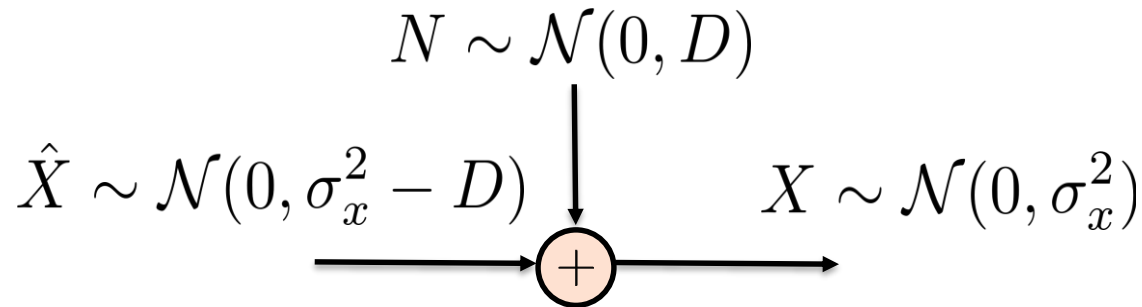


- But vector quantization approaches the Rate-Distortion function for large dimensionality, **and we're in 2-D!!**
- We can improve a bit by adding extra dimensions, (e.g., time slicing, for extra delay) or jointly quantizing several users (e.g. space cloaking), but still...

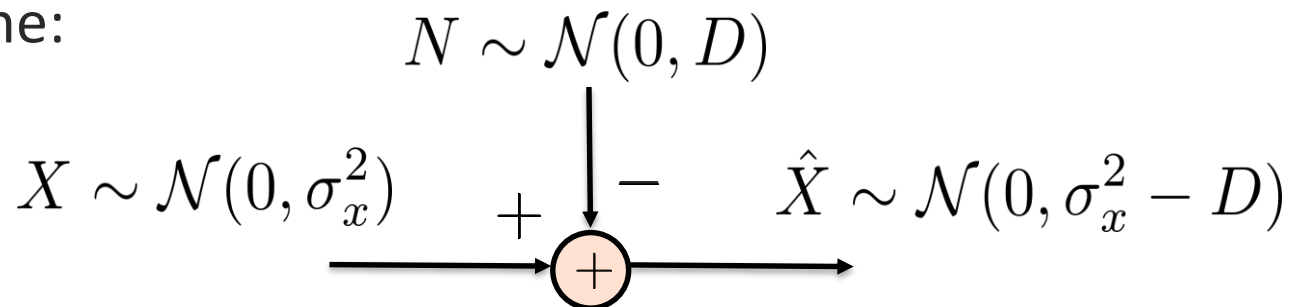


## Rate-Distortion (RD) of an i.i.d. Gaussian Source

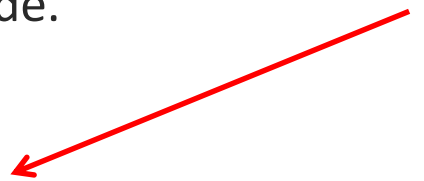
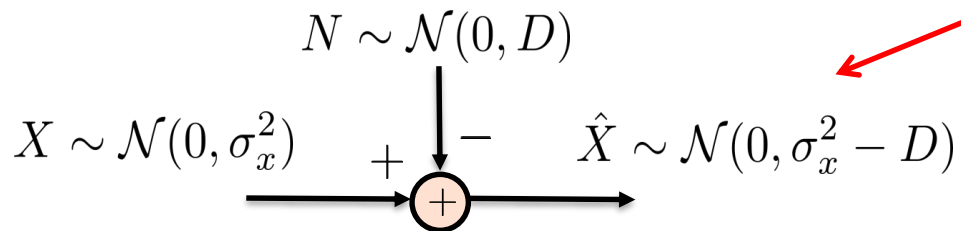
- The “test-channel” is used classically to find the RD for an i.i.d. Gaussian source.



- This is used to show that, for a Gaussian source, it is optimal to use i.i.d. Gaussian codewords and this yields i.i.d. Gaussian “quantization” noise. We can always revert the scheme:



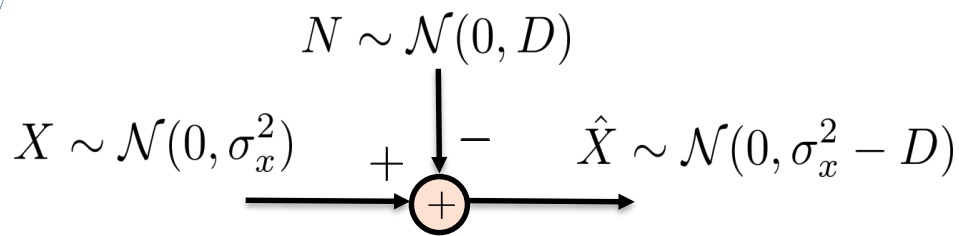
- The “test channel” is a theoretical construction; we see that adding noise would work, but this would produce i.i.d. Gaussian codewords, which are not **practical** as a source code.



- But in location privacy, we do not care about **rate**, but about **privacy**.
- So for us adding noise is OK!  
(recall we already proposed this)



- The test channel would suggest adding independent noise in the Gaussian case, right?



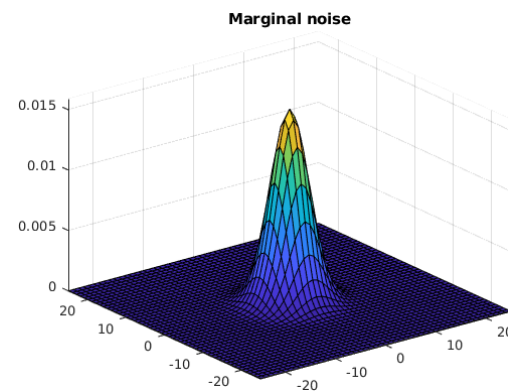
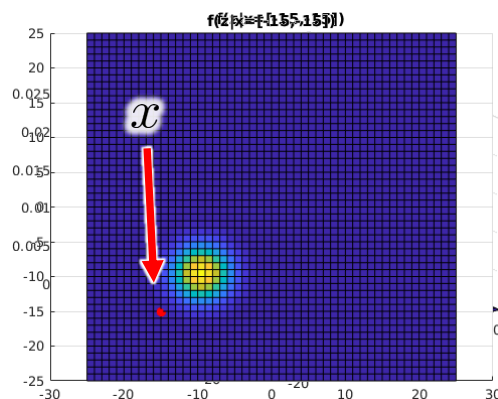
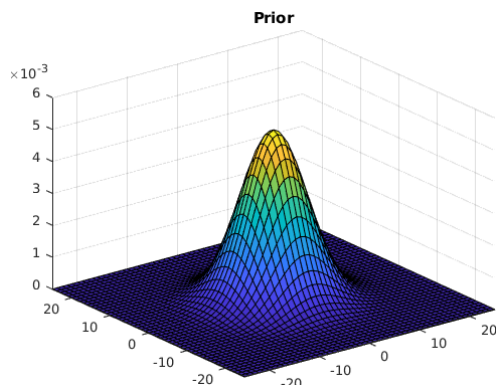
- Not so fast! In the test channel, the noise  $N$  is independent of  $\hat{X}$  but not of  $X$ !!!
- In fact, the RD is achieved when the noise  $N$  has the form:

$$N = \alpha X + N_0$$

with  $\alpha = D/\sigma_x^2$  and  $N_0$  i.i.d. noise independent of  $X$  and with variance

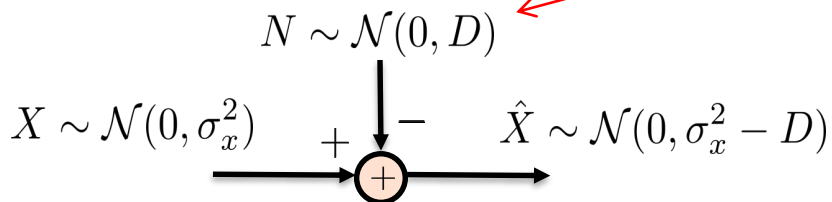
$$\sigma_{N_0}^2 = D - D^2/\sigma^2$$

- This matches what happens with optimal conditional entropy LPPMs if the prior is Gaussian:



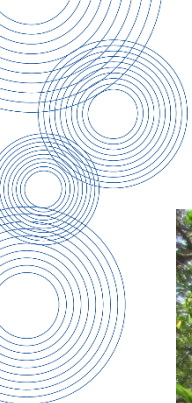
- Gaussian prior.
- We compute the optimal  $P_{CE}$  LPPM.
- This is  $f(z|x = (-15, -15))$
- It's not  $\mathcal{N}(0, \Sigma)$ !!
- The noise is **dependent** on  $x$ !!
- The **marginal noise**:  $n = z - x$

$$f(n) = \sum_{x \in \mathcal{X}} f(n|x) \cdot \pi(x)$$



- It's actually **Gaussian**!!
- This matches the “test channel”.





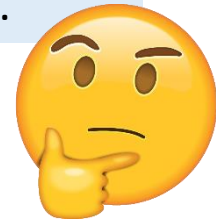
Sorry! Where were we?



# Maximizing the Conditional Entropy

$$\begin{array}{ll} \underset{f}{\text{maximize}} & P_{\text{CE}} \\ \text{s.t.} & \bar{Q} \leq \bar{Q}_{\text{max}} \\ & f \in \mathcal{P} \end{array} \quad \rightarrow \quad \begin{array}{ll} \underset{f}{\text{minimize}} & I(x; z) \\ \text{s.t.} & \bar{Q} \leq \bar{Q}_{\text{max}} \\ & f \in \mathcal{P} \end{array} \quad \rightarrow$$

We could gather  $r \rightarrow \infty$  locations and then perform a quantization.



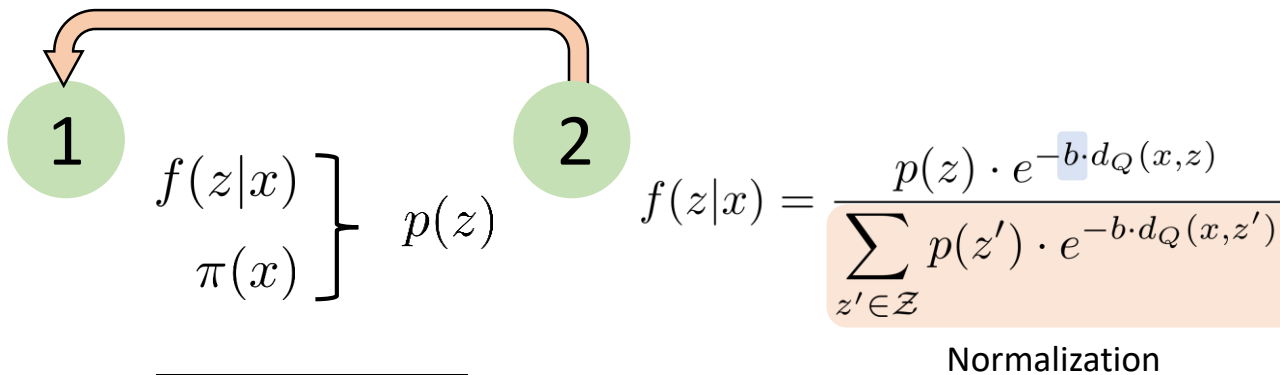
We could also add (dependent) noise to every location... which is more convenient!!

Blahut-Arimoto algorithm computes the encoding  $X \rightarrow \hat{X}$  in source coding. We can use to compute the LPPM  $f(z|x)$  in location privacy!

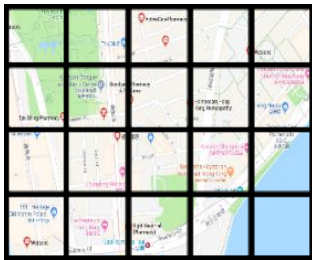
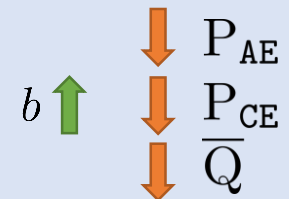
$$\begin{aligned} & \underset{f}{\text{minimize}} && I(x; z) \\ & \text{s.t.} && \bar{Q} \leq \bar{Q}_{\max} \\ & && f \in \mathcal{P} \end{aligned}$$

# Blahut-Arimoto Algorithm

- The exponential distribution maximizes the entropy for a given distortion constraint.
- Blahut-Arimoto: iterative algorithm that tries to make an exponential posterior  $p(x|z)$ .



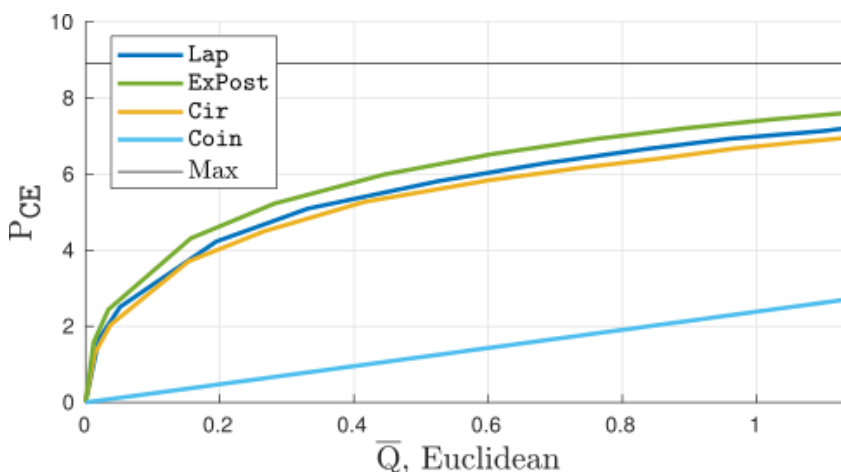
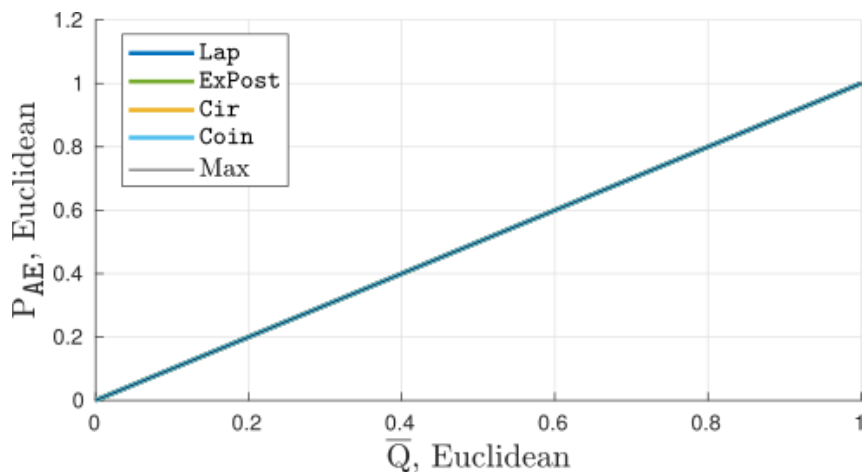
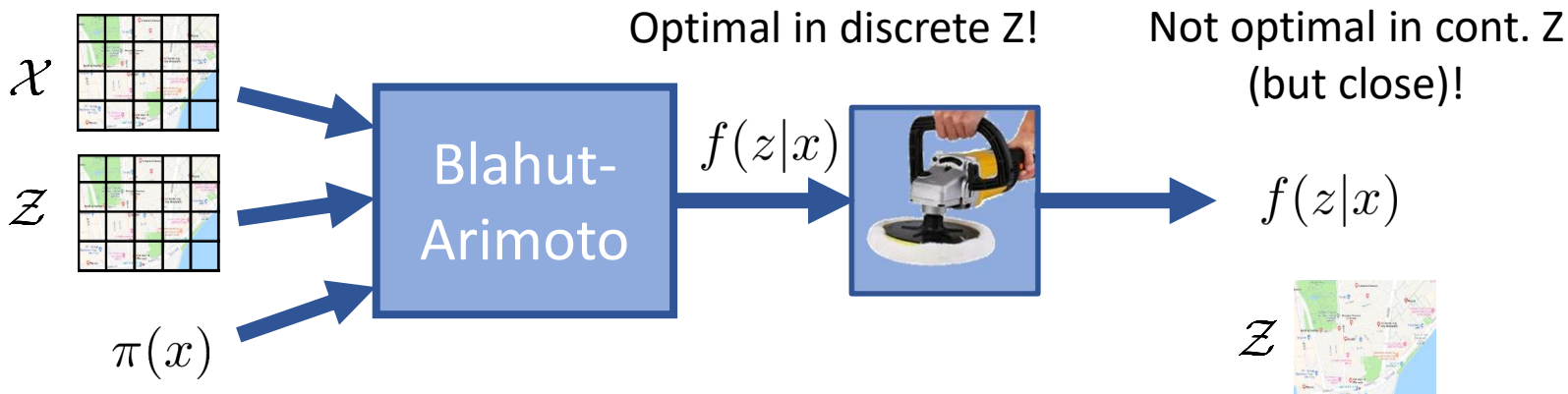
$b$  affects privacy and quality loss



Intuition:

$$p(x|z) = \frac{f(z|x)}{p(z)} \cdot \pi(x) \approx \text{Exponential}$$

# Exponential Posterior LPPM



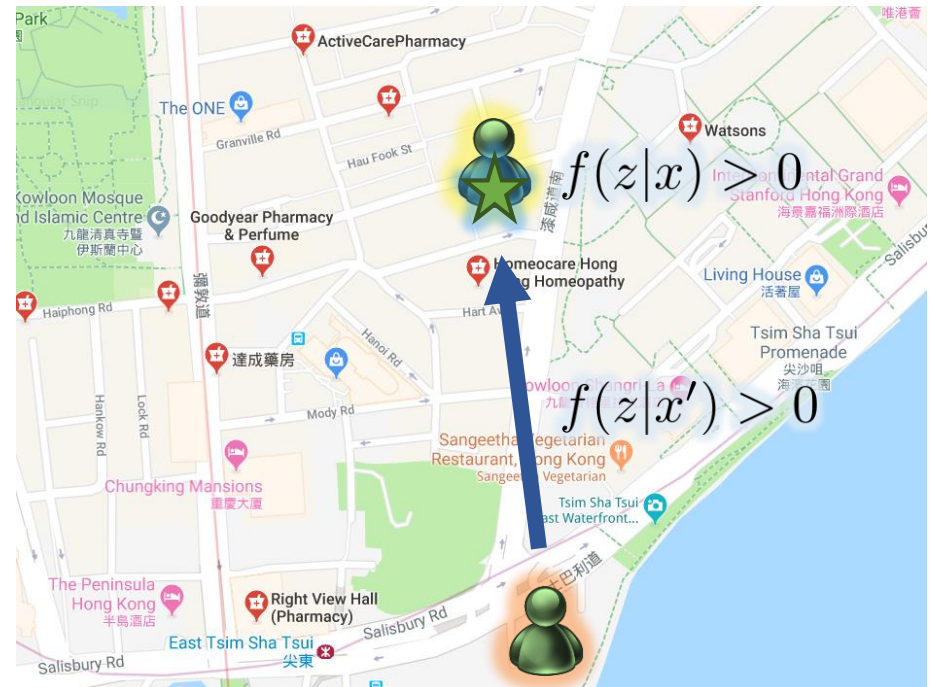
Simon Oya, Carmela Troncoso, and Fernando Pérez-González. "Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms." CCS'17.

# Designing Geo-Indistinguishability LPPMs

- Recap:

$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

- Most LPPMs do not guarantee any level of geo-ind (i.e.,  $\epsilon \rightarrow \infty$  ).
  - E.g., finite mechanisms.

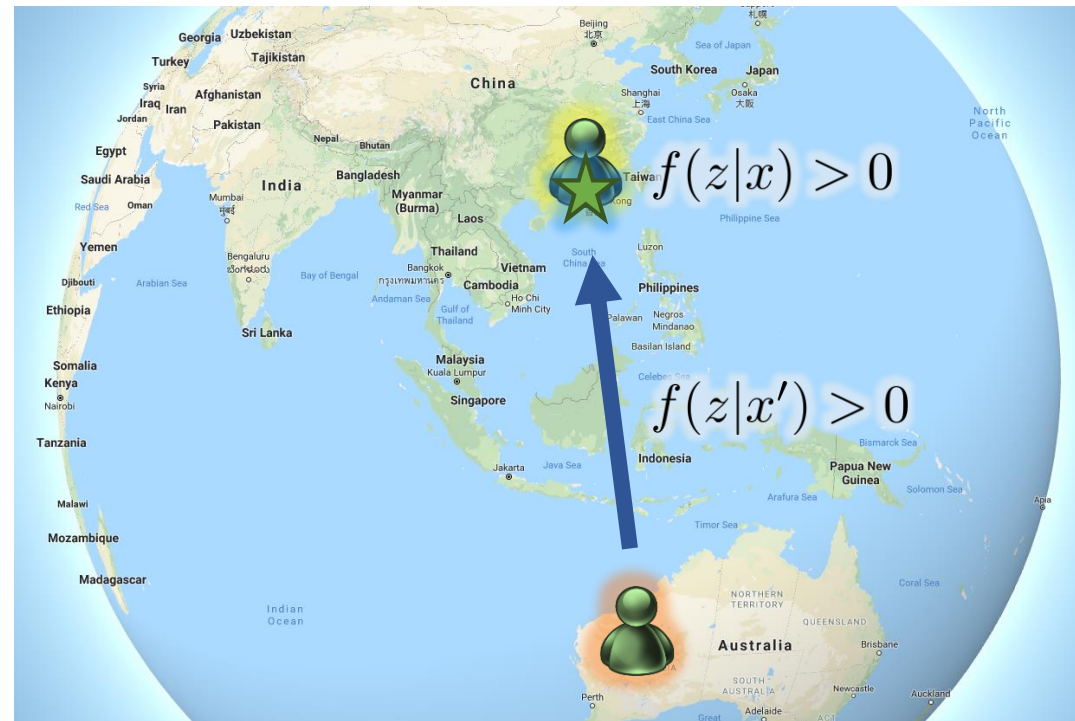


# Designing Geo-Indistinguishability LPPMs

- Recap:

$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

- Most LPPMs do not guarantee any level of geo-ind (i.e.,  $\epsilon \rightarrow \infty$  ).
  - E.g., finite mechanisms.
- We are going to see some geo-ind LPPMs



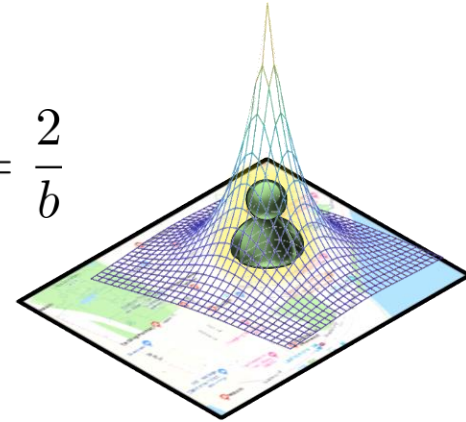
$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

## Laplacian LPPM

- Continuous map:

$$f(z|x) = \frac{b^2}{2\pi} e^{-b \cdot d_2(x, z)}$$

$$\bar{Q} = \frac{2}{b}$$



- The Laplacian LPPM provides  $b$ -geo-indistinguishability.

$$f(z|x) = \frac{b^2}{2\pi} e^{-b \cdot d_2(x, z)} \leq \underbrace{\frac{b^2}{2\pi} e^{-b \cdot d_2(x', z)}}_{f(z|x')} \cdot e^{b \cdot d_2(x, x')}$$

$$d_2(x', z) \leq d_2(x, z) + d_2(x, x') \quad f(z|x')$$

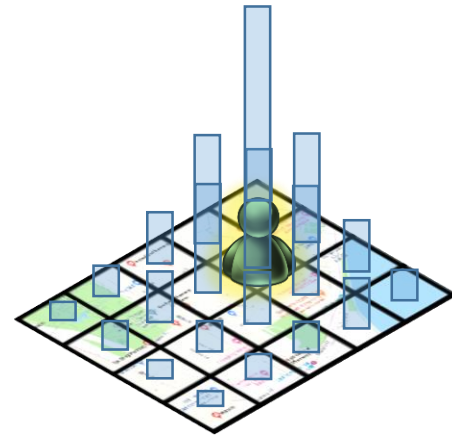
$$-d_2(x, z) \leq -d_2(x', z) + d_2(x, x')$$

$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

## Exponential LPPM

- Discrete set of locations:

$$f(z|x) = \frac{e^{-b \cdot d_2(x, z)}}{\sum_{z' \in \mathcal{Z}} e^{-b \cdot d_2(x, z')}} \quad \text{with } z' \in \mathcal{Z}$$



- This mechanism guarantees  $2b$ -geo-indistinguishability.

$$\begin{aligned} f(z|x) &= \frac{e^{-b \cdot d_2(x, z)}}{\sum_{z' \in \mathcal{Z}} e^{-b \cdot d_2(x, z')}} \leq \frac{e^{-b \cdot d_2(x', z)}}{\sum_{z' \in \mathcal{Z}} e^{-b \cdot d_2(x, z')}} \cdot e^{b \cdot d_2(x, x')} \\ &\leq \frac{e^{-b \cdot d_2(x', z)}}{\sum_{z' \in \mathcal{Z}} e^{-b \cdot d_2(x', z')}} \cdot \frac{e^{b \cdot d_2(x, x')}}{e^{-b \cdot d_2(x, x')}} = e^{2b \cdot d_2(x, x')} \cdot f(z|x') \end{aligned}$$

$$d_2(x, z') \leq d_2(x', z') + d_2(x, x') \longrightarrow -d_2(x, z') \geq -d_2(x', z') - d_2(x, x')$$



$$f(z|x) \leq e^{\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

## What about ExPost?

- Blahut-Arimoto iteration was:

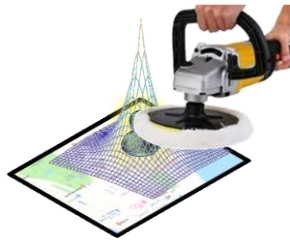
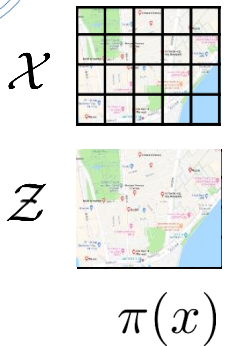
$$\begin{array}{c}
 \textcircled{1} \\
 f(z|x) \\
 \pi(x)
 \end{array}
 \left. \vphantom{\begin{array}{c} f(z|x) \\ \pi(x) \end{array}} \right\} p(z)
 \quad
 \textcircled{2}
 \quad
 f(z|x) = \frac{p(z) \cdot e^{-b \cdot d_Q(x, z)}}{\sum_{z' \in \mathcal{Z}} p(z') \cdot e^{-b \cdot d_Q(x, z')}}$$

- Exponential LPPM:

$$f(z|x) = \frac{e^{-b \cdot d_2(x, z)}}{\sum_{z' \in \mathcal{Z}} e^{-b \cdot d_2(x, z')}}$$

- If  $d_Q(\cdot)$  satisfies the triangle inequality, ExPost provides  $2b$ -geo-indistinguishability.

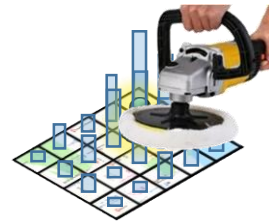
# Laplacian vs. Exponential vs. ExPost



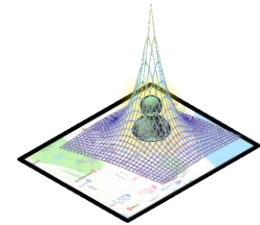
Laplace



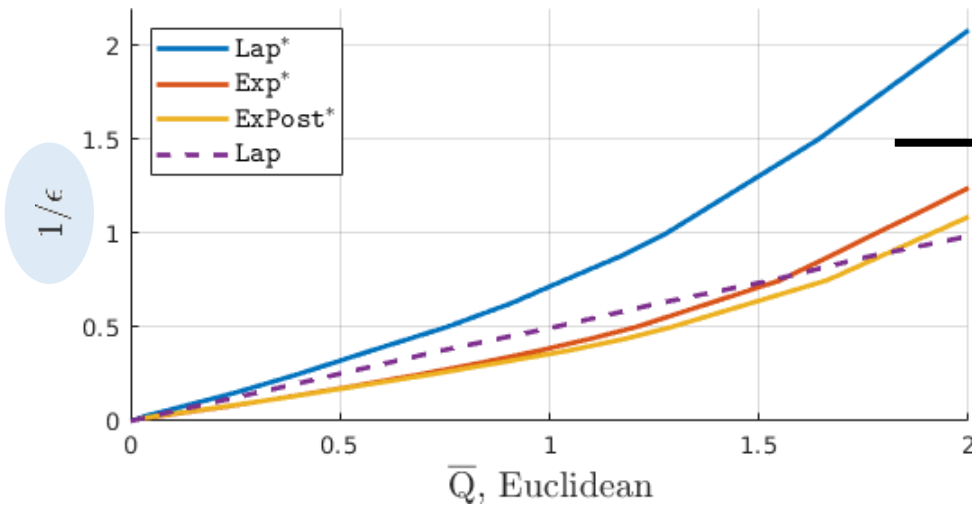
Exponential



ExPost



Laplace



State-of-the-art  
in geo-ind.

Gowalla dataset

# Optimal Geo-Indistinguishability

- We can minimize  $\epsilon$  subject to a quality loss constraint.
- In this case, it is easier to minimize the quality loss subject to an  $\epsilon$  constraint.

$$\underset{f(z|x)}{\text{minimize}} \quad \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x) \cdot f(z|x) \cdot d_Q(x, z)$$

$$\text{s.t.} \quad f(z|x) \leq e^{-\epsilon \cdot d_2(x, x')} \cdot f(z|x') \quad \forall x, x', z$$

$$\sum_{z \in \mathcal{Z}} f(z|x) = 1, \quad \forall x$$

$$f(z|x) \geq 0, \quad \forall x, z$$



$N^2$  variables

$N^3$  constraints

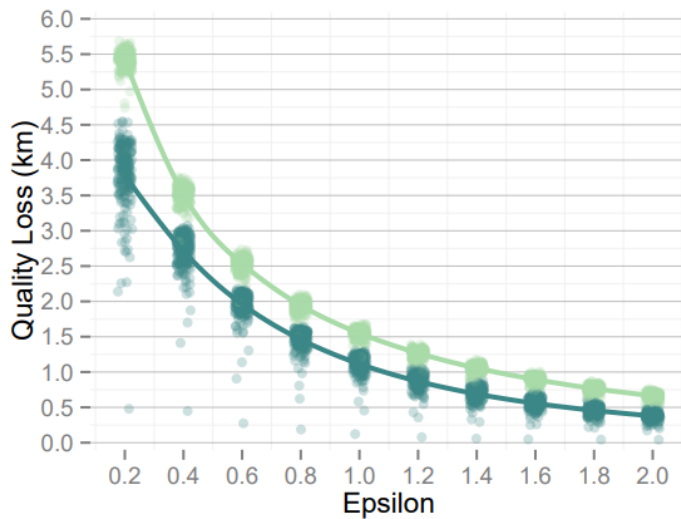
$N$  constraints

$N^2$  bounds

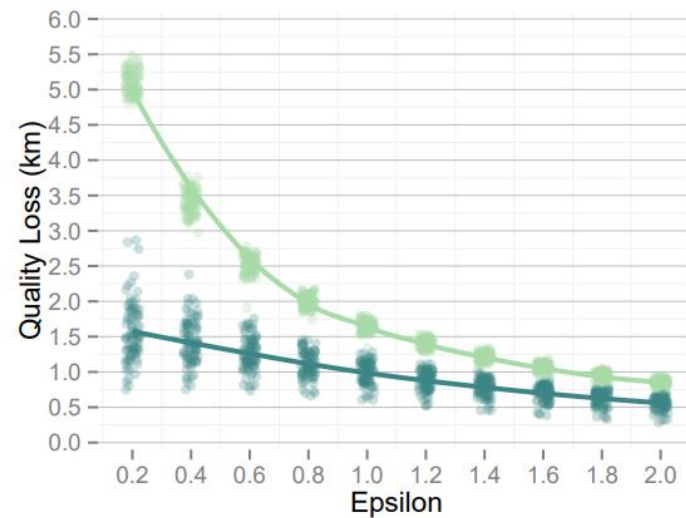
- There are more efficient methods, but still expensive...

# How Good is Optimal Geo-Ind?

T-drive dataset



Geolife dataset





- Optimal Geo-Indistinguishability
- Laplace LPPM (no remapping)

N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. "Optimal geo-indistinguishable mechanisms for location privacy." CCS'14.



# LPPM Design: Summary

## Average Error ( $P_{AE}$ )

- **Linear Program:**



- Performance: 
- Scaling: 

- **Any LPPM + Remapping (RM):**



- Performance:
- If  $d_P = d_Q$ : 
- If  $d_P \neq d_Q$ : ?
- Scaling: 

## Conditional Entropy ( $P_{CE}$ )



- **ExPost (Blahut-Arimoto):**

- Performance: 
- Scaling: 

- **Laplace/Gaussian/Circular + RM:**



- Performance: 
- Scaling: 

- **“Binary mechanisms”:**



- Performance: 
- Scaling: 

## Geo-Indistinguishability

- **Laplace + RM:**

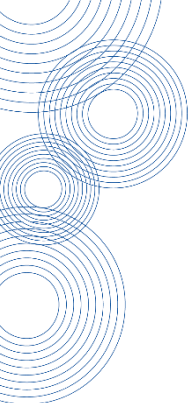
- Performance: 
- Scaling: 

- **Exponential/ExPost + RM:**

- Performance: 
- Scaling: 

- **Optimal Geo-Ind:**

- Performance: 
- Scaling: 



# Practical Considerations for LPPM Design

# Everything was a lie!

- Well, not everything... just our first assumption.



Do we know the real mobility model?

Can we look into the future?

Yes

No



We do not know the real mobility model

I know it!

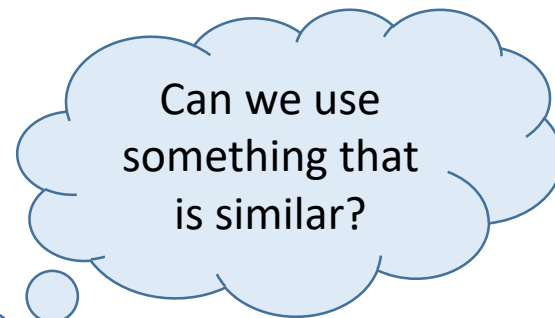


# What can we do?

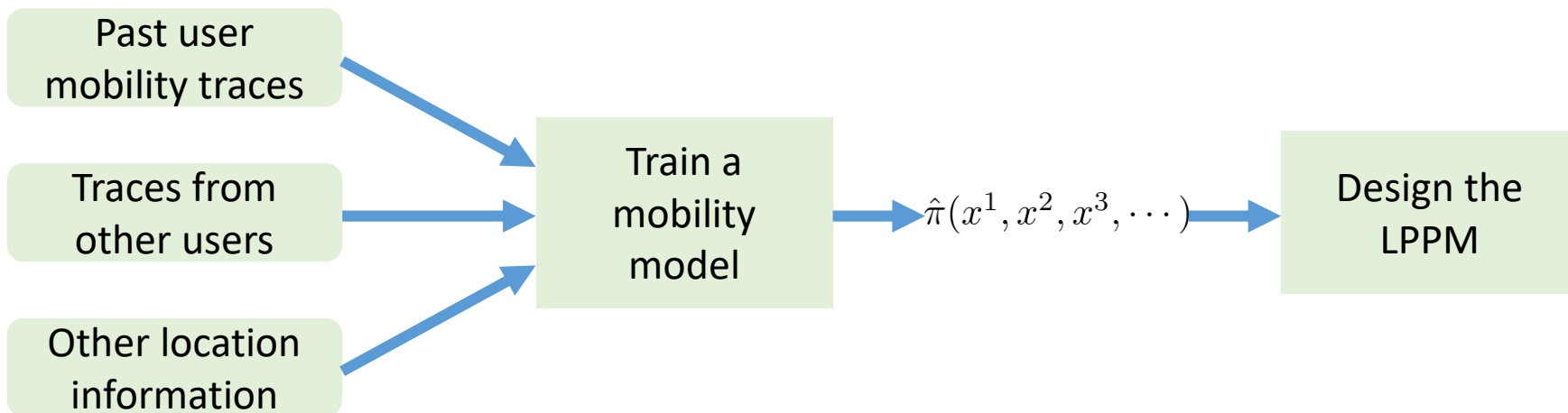


We do not know  
the real mobility  
model

$$\pi(x^1, x^2, x^3)$$



$$\hat{\pi}(x^1, x^2, x^3)?$$





# Wait, but the user knows her “mobility” on the fly!



Some algorithm



$$\hat{\pi}(\dots, x^{r-2}, x^{r-1}, x^r)$$

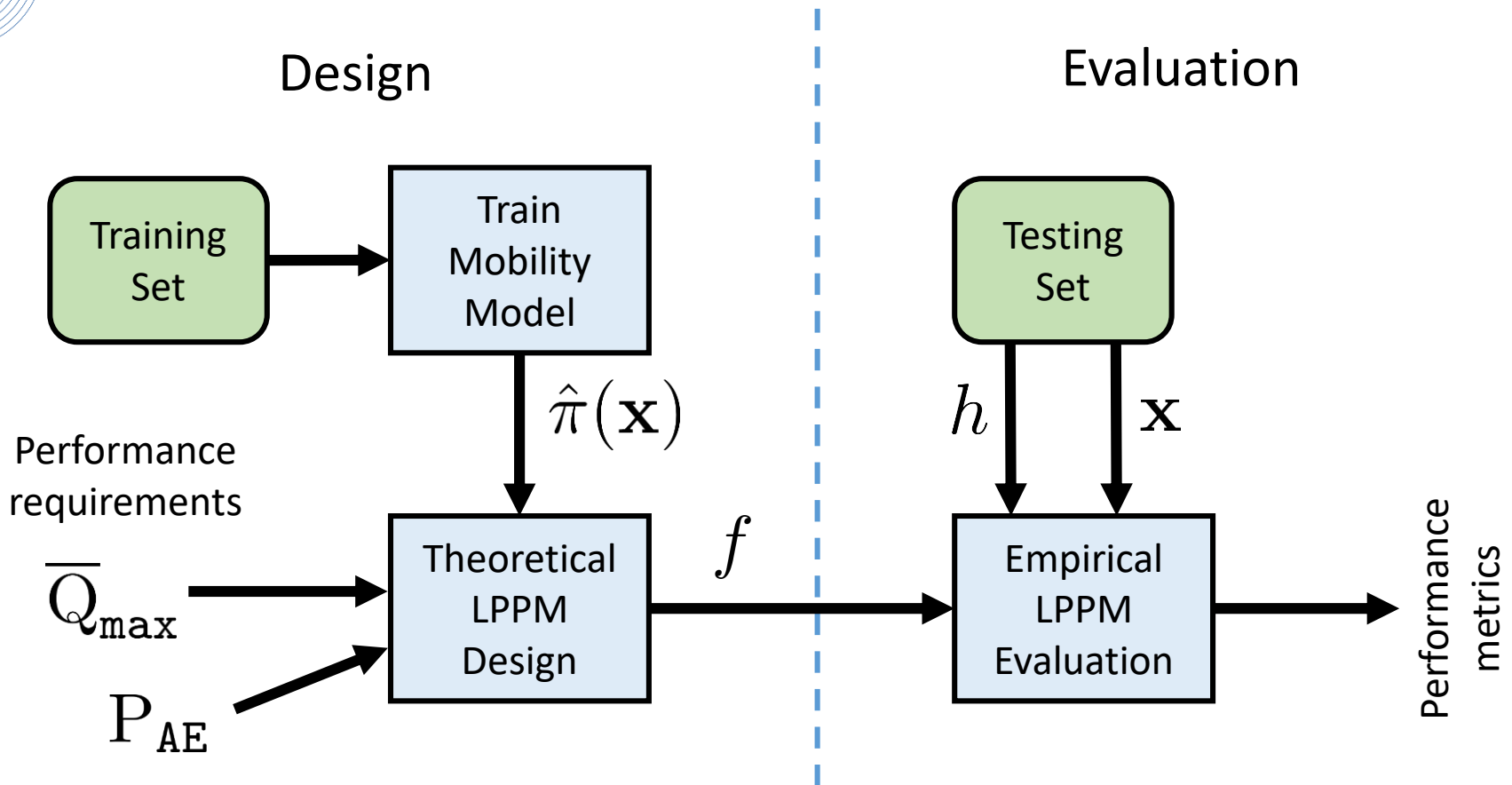
Problem:  $f(z^r | x^1, \dots, x^r, z^1, \dots, z^{r-1})$

Optimal attack:  $p(x^1, \dots, x^r | z^1, \dots, z^r)$



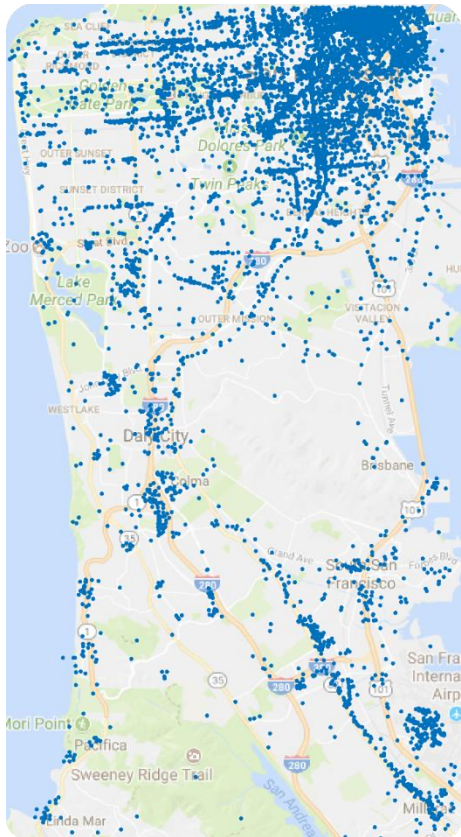
We need to handle  $|\mathcal{X}|^r$  values to compute the optimal attack and measure privacy...

# LPPM Design and Evaluation Framework

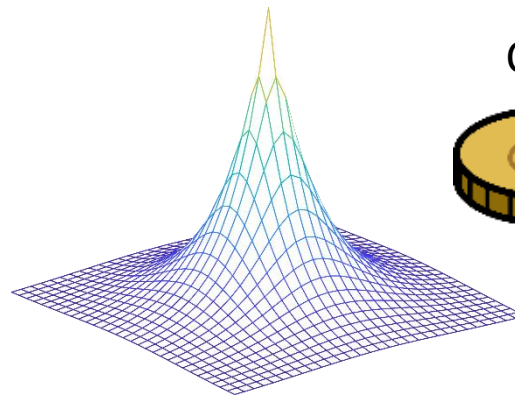


# How Does This Affect LPPM Performance?

- San Francisco Region:



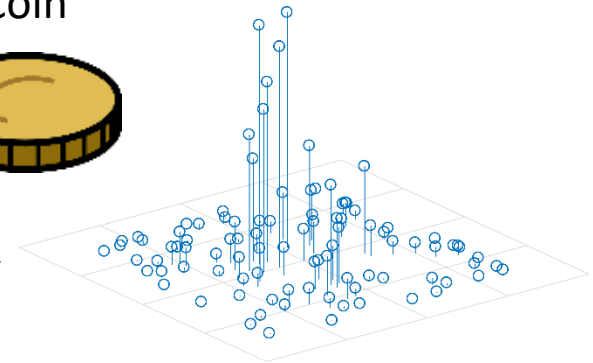
- Gowalla dataset.
  - 16 users for evaluation
  - All the others as training data
- Sporadic mobility assumption.



Laplacian

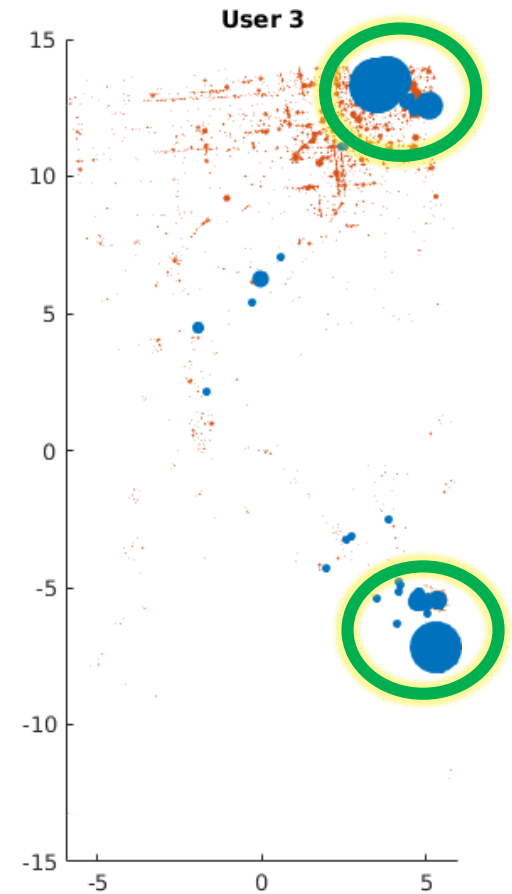
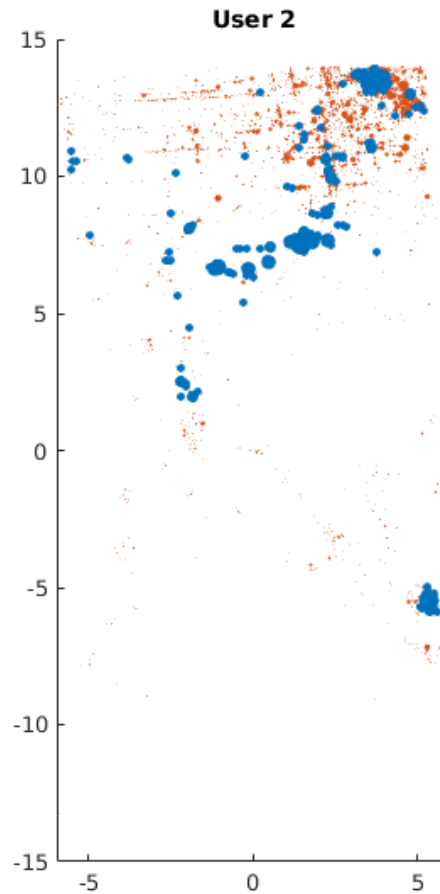
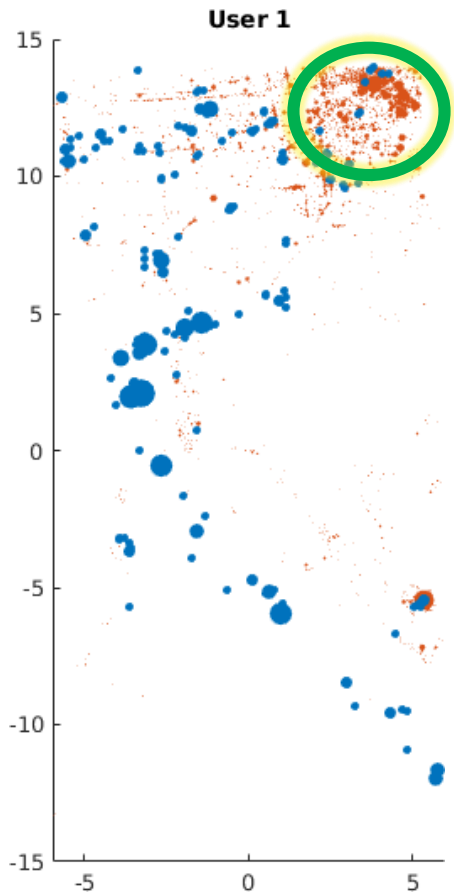


Coin



ExPost

# User's Mobility Profile

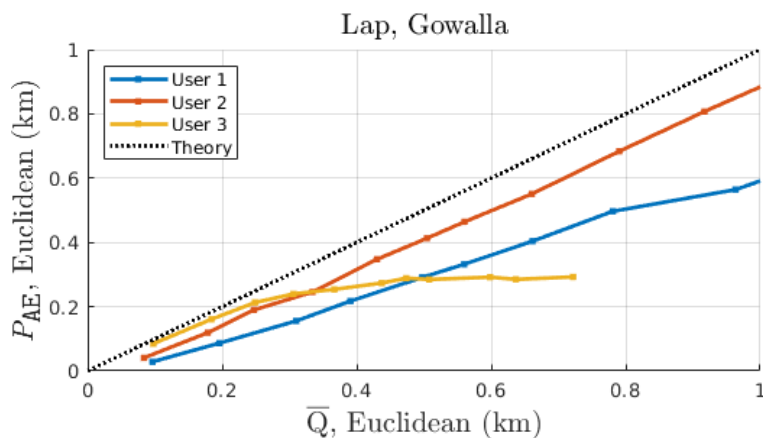
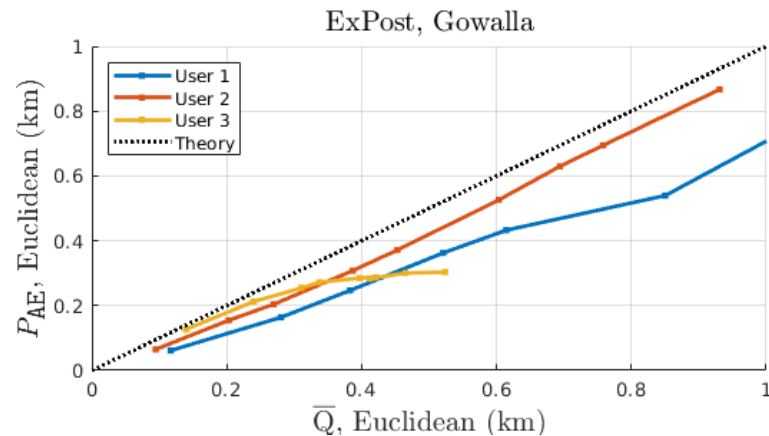
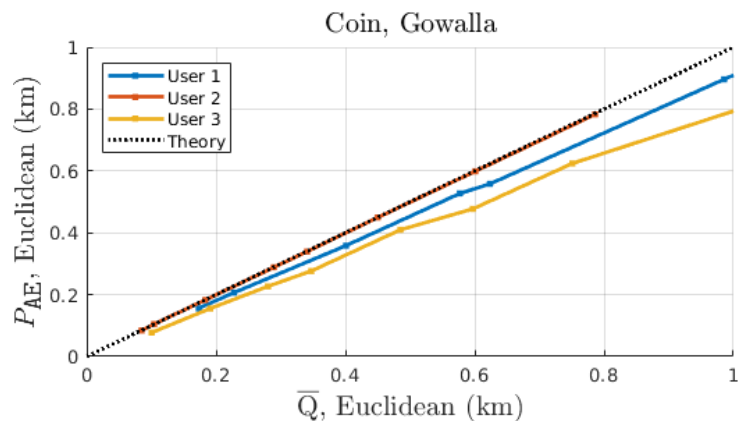


Very spread, not close to avg.

Closer to avg.

Concentrated in two regions.

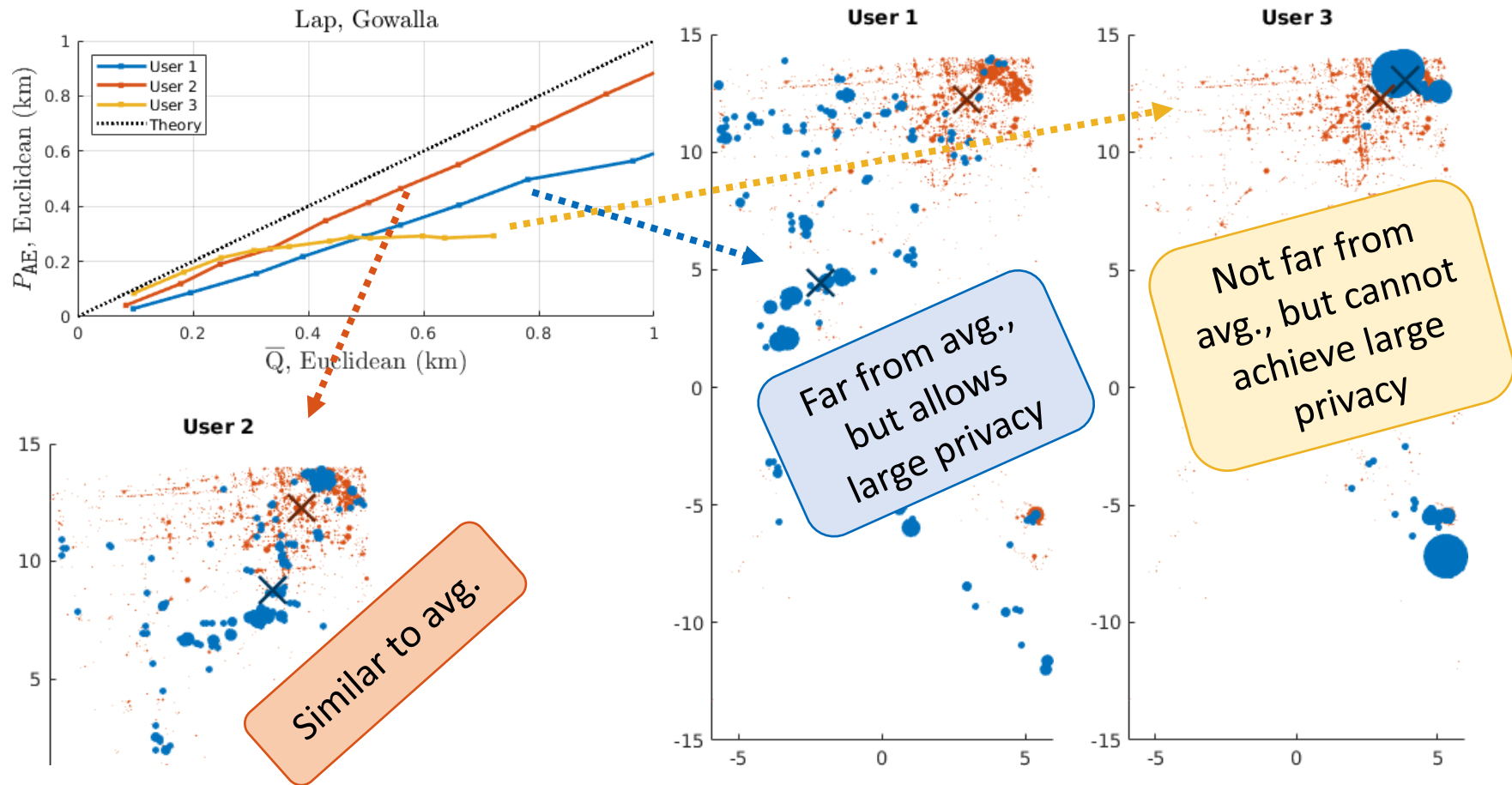
# Average Error Privacy (Gowalla)



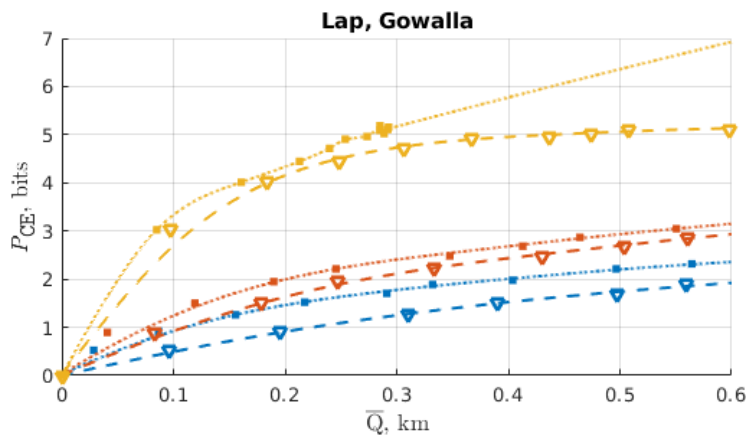
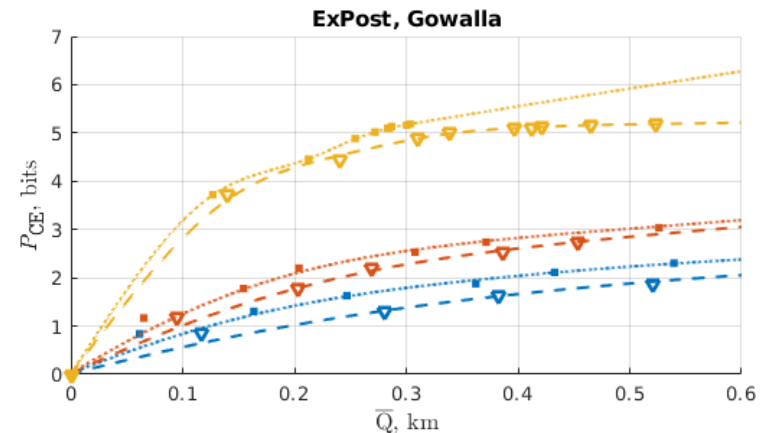
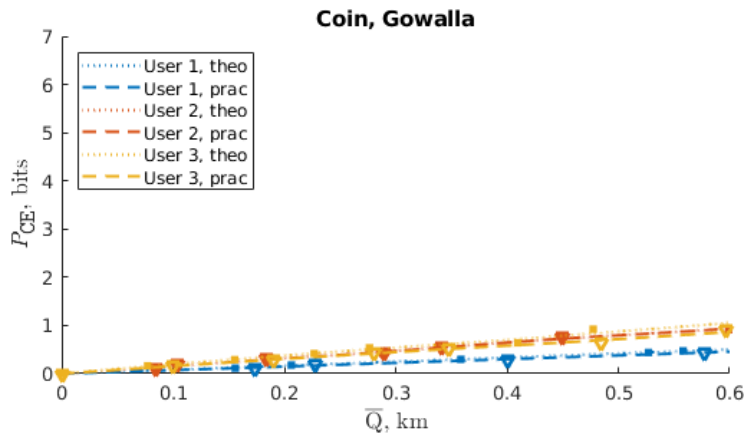
- Performance worsens in evaluation.
- Users: different performance in the evaluation!!
- LPPMs: different performance in the evaluation!!



# Different Performance Among Users (Average Error Privacy)

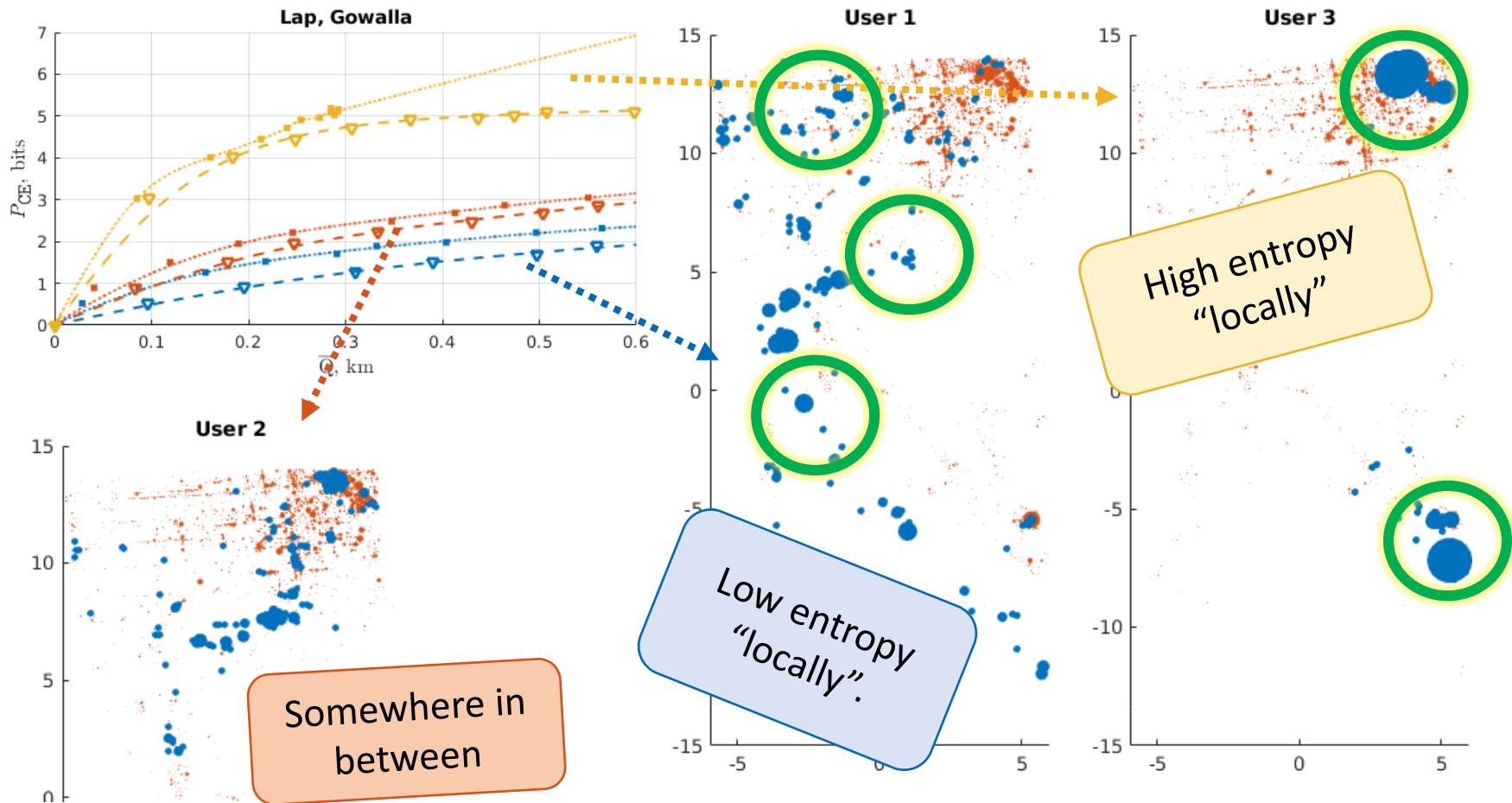


# Conditional Entropy (Gowalla)



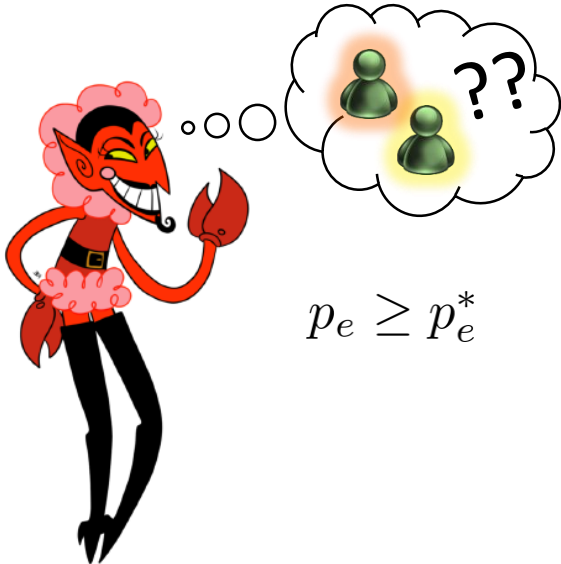
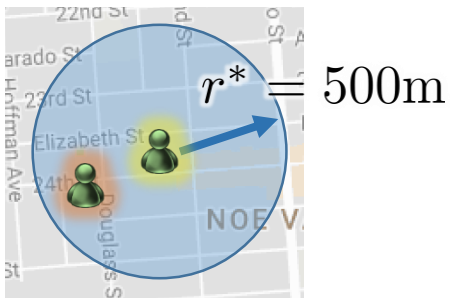
- Performance worsens in evaluation (but not much, due to concavity of entropy).
- Users: different performance in the evaluation!!
- LPPMs: different performance in the evaluation!!

# Different Performance Among Users (Conditional Entropy)

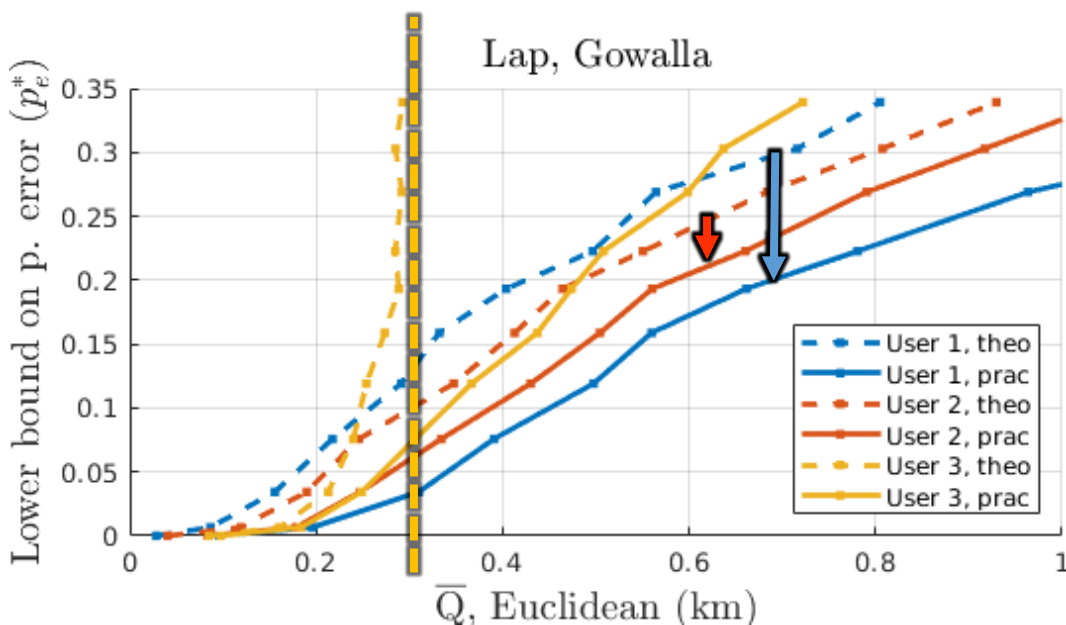
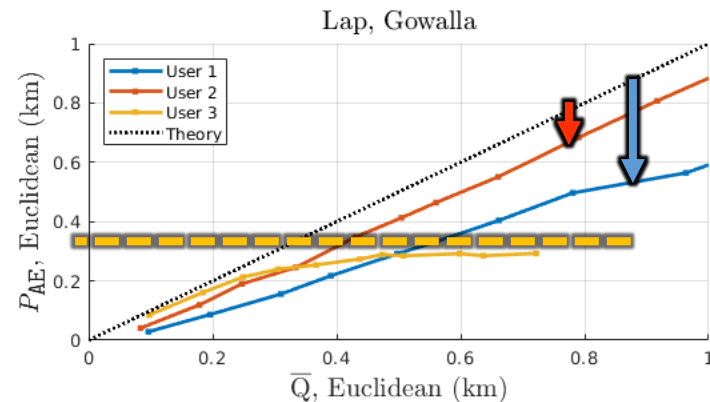




# Geo-Indistinguishability (as an Adversary Error)



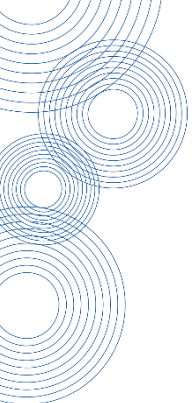
$$p_e \geq p_e^*$$





## Practical Considerations (Conclusions)

- In practice, we do not perfectly know the user's mobility model.
- LPPMs designed with training data are tailored to that data, and thus perform worse if the evaluation data is different.
- Designing LPPMs to protect users in practice is very challenging.



# Challenges Ahead

- 
- We have explained the basics of user-centric perturbation-based location privacy.

How to measure  
privacy and utility

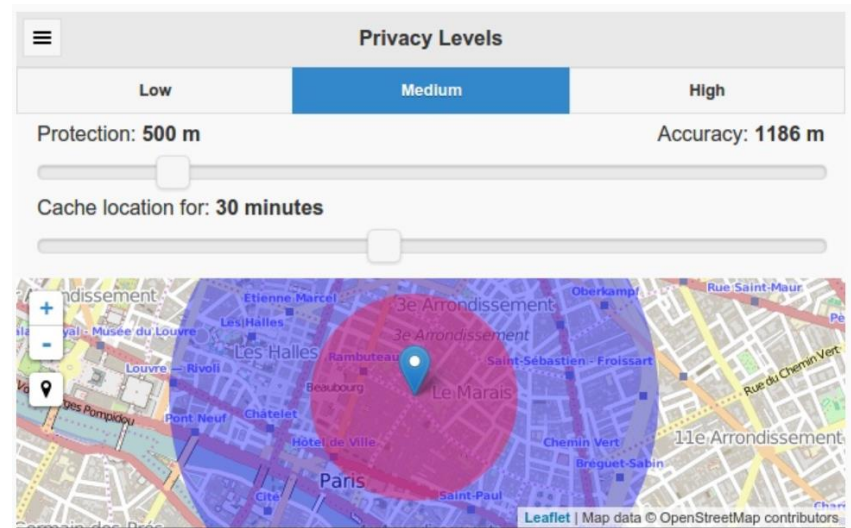
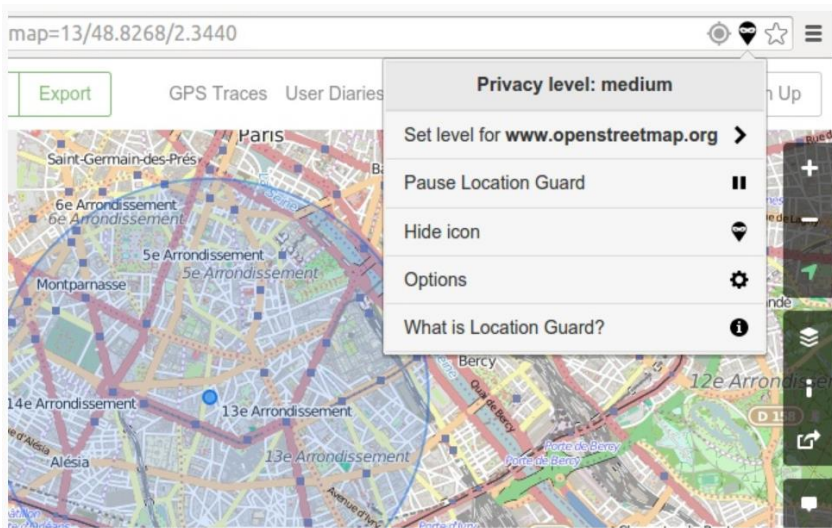
How to design  
LPPMs according to  
those metrics

Practical  
considerations  
when designing  
those LPPMs

- In practice, guaranteeing location privacy is (even) a more complex issue.

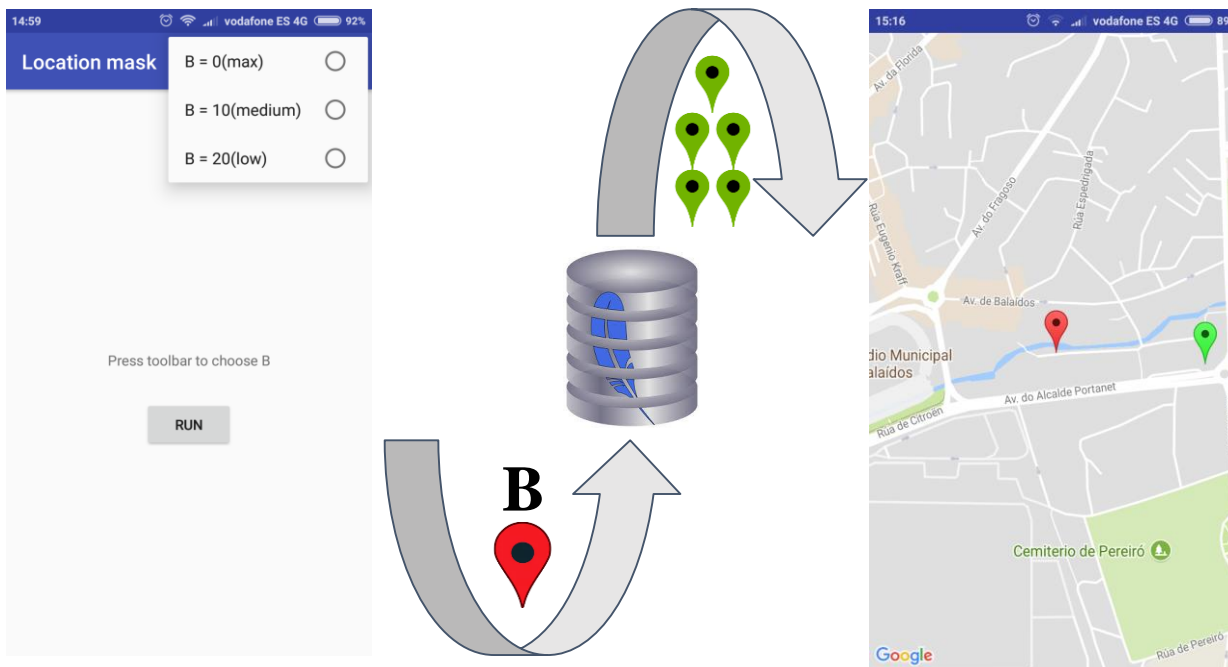
# Location Guard

- Extension for Google Chrome/Firefox.
- You can choose from privacy levels and add Laplacian noise.
- You can also choose a fixed location (this is actually what most people use it for).



# Academic Implementations

- Location Mask, developed by Miguel Gallego Martín (University of Vigo).
- Implementation of the ExPost LPPM.



## The “Privacy Paradox”

- Consumer’s choice to use mobile technologies is primarily driven by considerations of popularity, usability and the price of a given technology despite the potential risk of data misuse.
- But research shows that users are concerned about privacy and misuses of their data.

Eurostat: 71% of Europeans agree that “providing personal information is an increasing part of modern life”.

57% disagree with “providing personal information is not a big issue for them”.



EUROBAROMETER

[http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs\\_431\\_en.pdf](http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf)

# 35 Theories for the “Privacy Paradox” [S. Barth et al 2017\*]

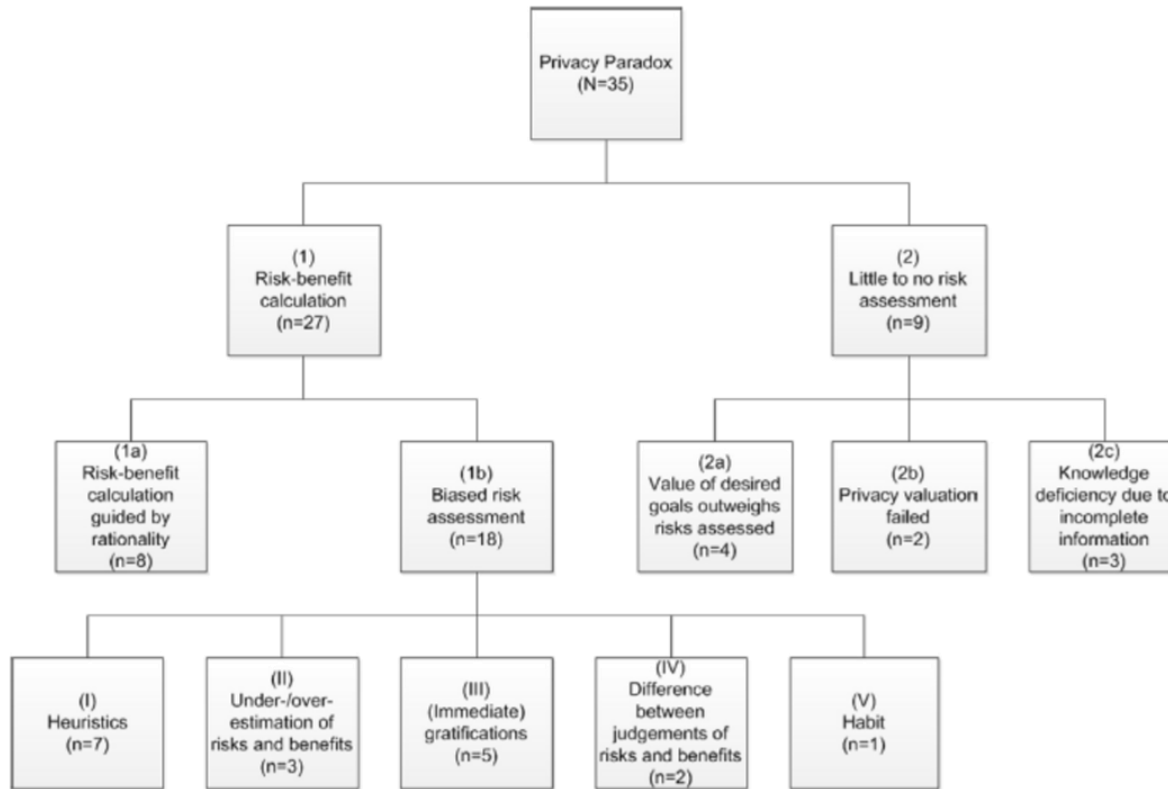


Fig. 1. Overview of categorization theories according to nature of decision making.

S. Barth et al. “The privacy paradox : Investigating discrepancies between expressed privacy concerns and actual online behavior - A systematic literature review”, Telematics and Informatics, 2017.





## Good remarks (in our view)

- People are biased when making choices about privacy vs. utility.
- There are no privacy assessment measures that can be used to make a rational decision.
- People underestimate their own risk but not others' (optimistic bias stance).
- People seek immediate gratification (including habit) and are concerned about being excluded from the group.
- Resignation (users perceive they have little power).
- Asymmetric and incomplete information.

# Asymmetric Information

- In the LBS market, users cannot properly evaluate the amount of privacy they lose.
- This is an instance of “asymmetric information”. Users cannot select the best product because there are hidden costs in terms of privacy.
- Example: Google doesn’t tell you the whole truth with “Location History”.
- Customers may not even consider using other privacy preserving alternatives because of such asymmetry.
- This “adverse selection” hampers innovation.
- This is why a solution is privacy enforcement by law.



## Incomplete Information

- The user does not know what is the utility from the LBS provider.
- Example: Facebook tells you it's for your own good.



“To create personalized products that are unique and relevant to you, we use your connections, preferences, interests and activities based on the data we collect and learn from you and others”.

### ***Facebook Secretly Shared User Data After Saying It Stopped***

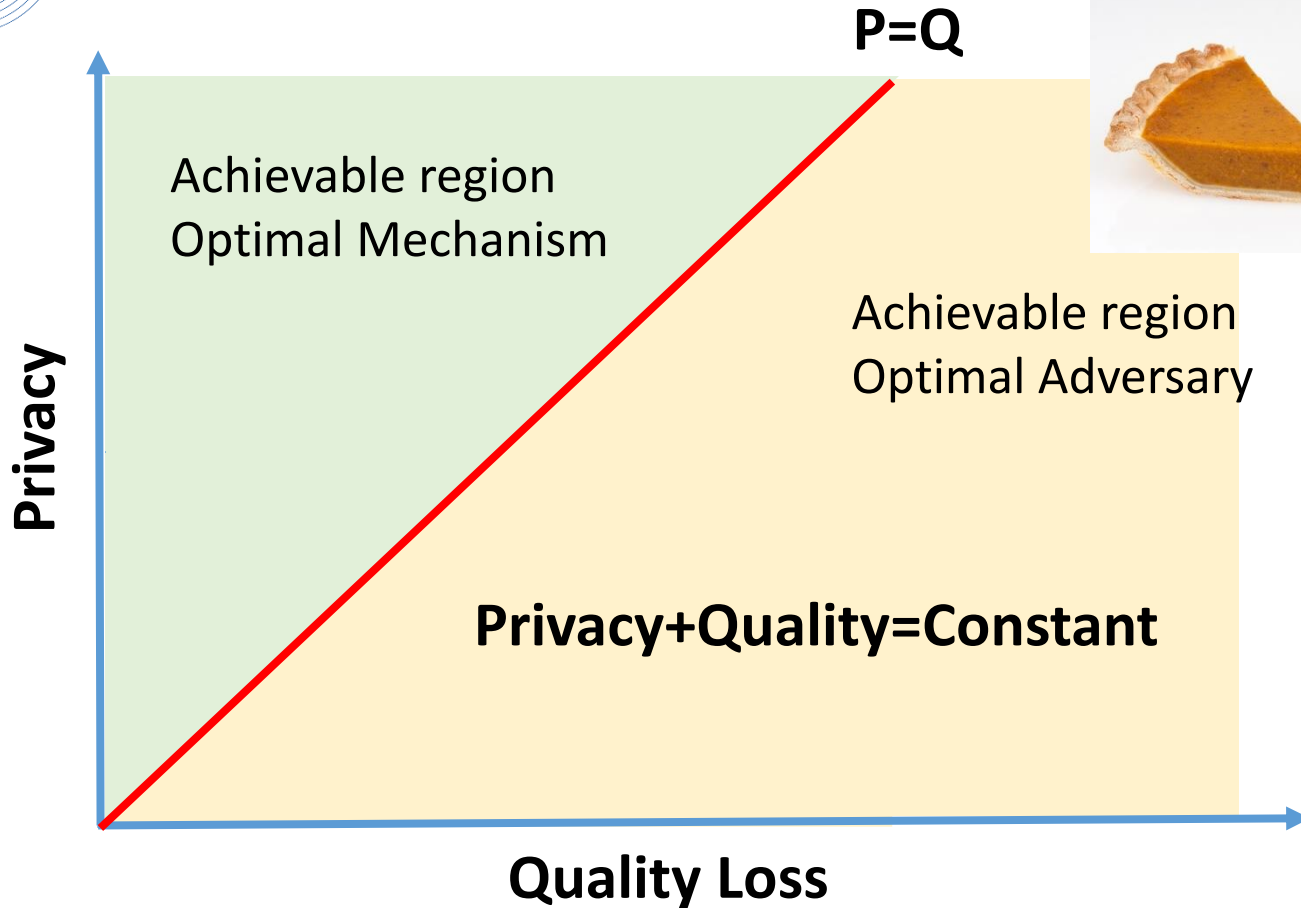
Netflix, AirBnb, Lyft and other companies got special access to info on people's friends without their knowledge, new documents published by Britain's Parliament reveal.



Kelly Weill 12.05.18 12:35 PM ET



# Privacy as a Zero-Sum Game

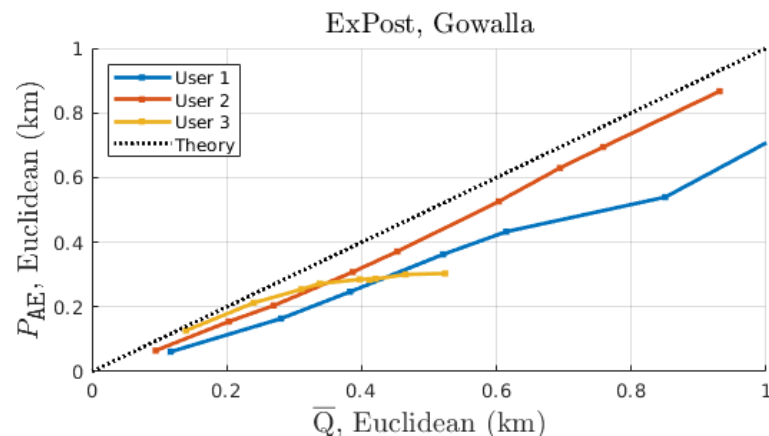


Utility for the adversary is proportional to Quality of user



## Can the user do anything else to increase privacy?

- So far, the quality of the response that we get from the Location Based Service is the resource that we trade in for privacy.
- However, there are other resources we could trade in for privacy:
  - Computational Complexity
  - Bandwidth
  - Delay



Average Quality Loss (compared to the quality of a unprotected location release)

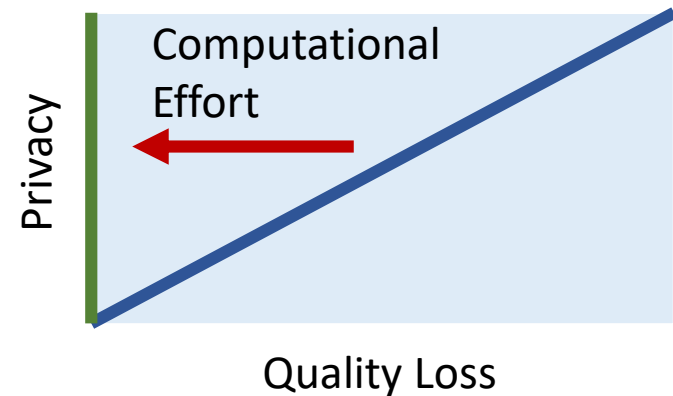
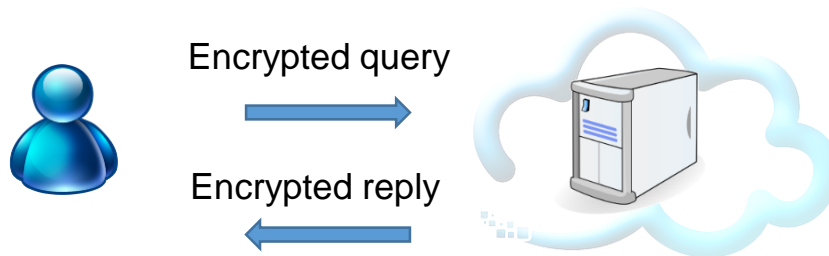
# Computational Complexity for Privacy

- We saw examples of this at the beginning.



I gave it for free because I wanted your data!

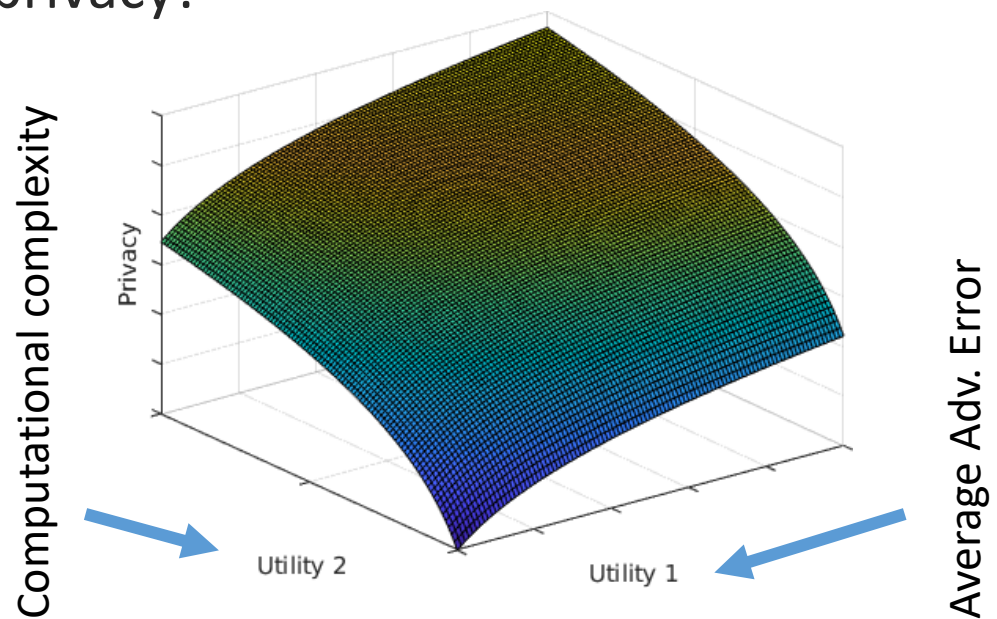
Retrieval in the Encrypted Domain



- But the service provider gets nothing for her collaboration
- There's a notion of "provider utility" behind this, that we have not taken into account!

## Can we find a midpoint?

- Maybe we can let the server get some information (some “server utility”), but also hide some.
- Also, we can rely on both computational complexity and perturbation to achieve privacy!
- Example:
- This is an interesting future line of work.

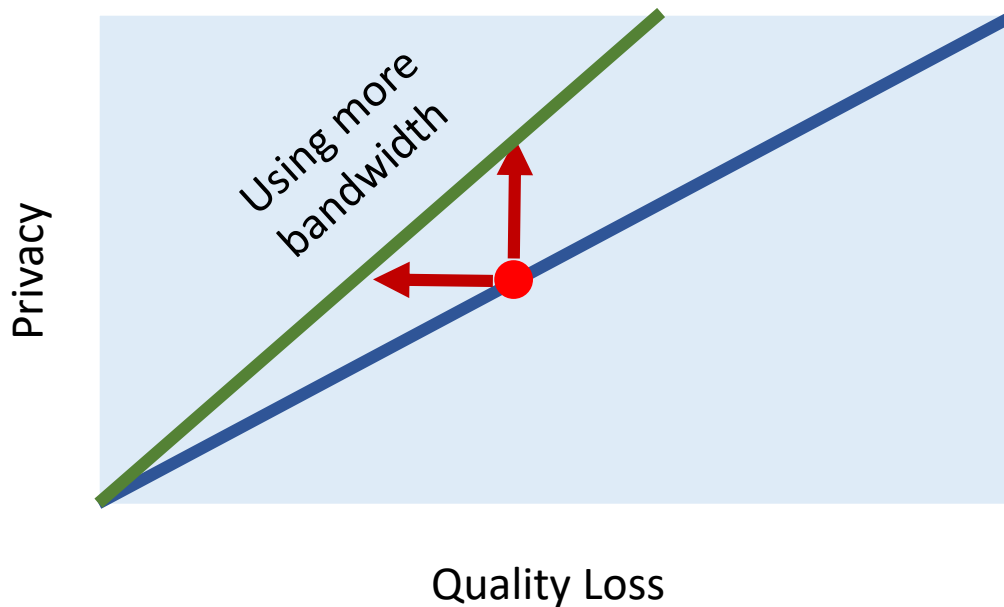






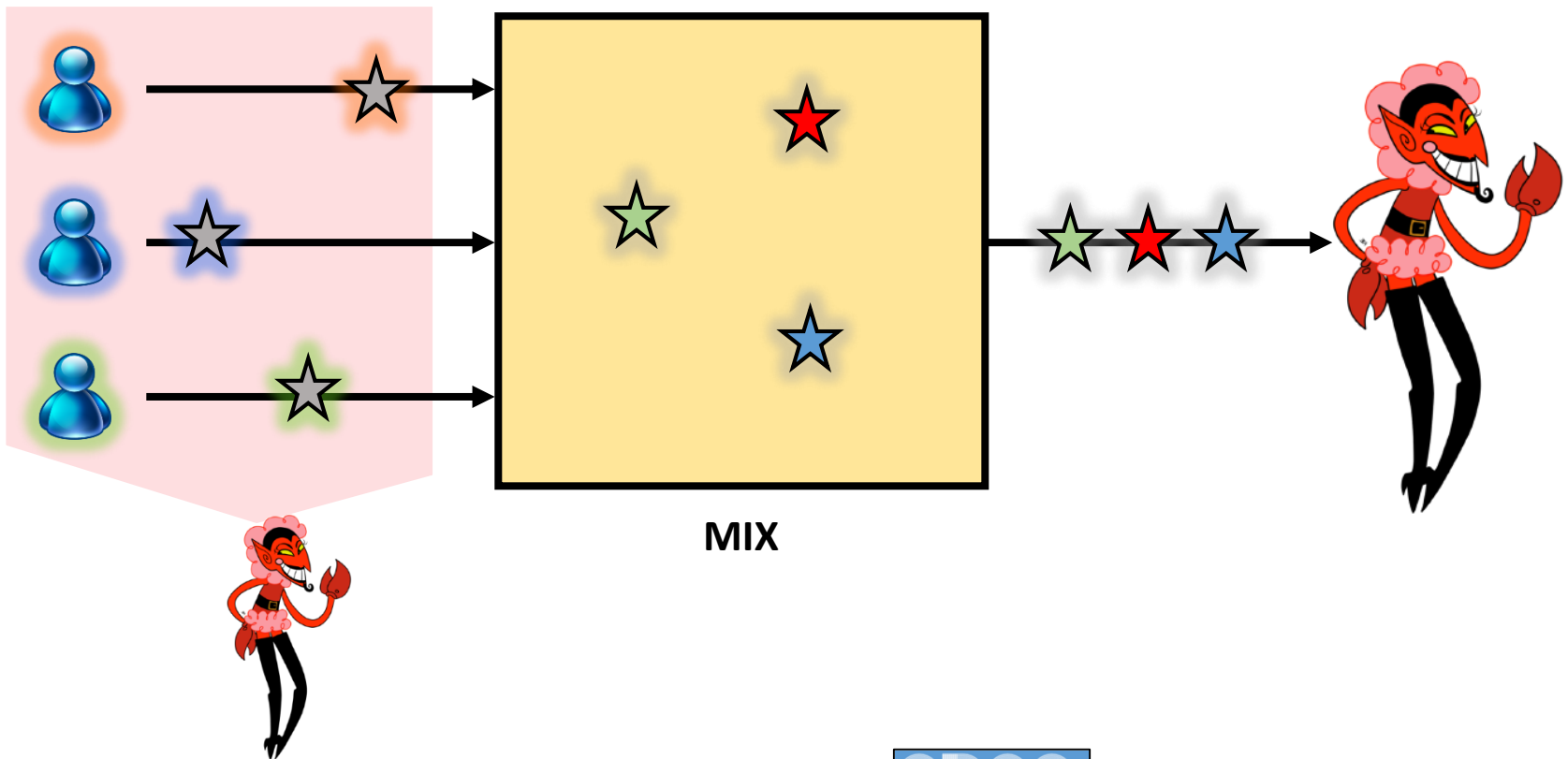
## Bandwidth as a Source of Privacy

- Using more bandwidth (dummy locations) decreases quality of service (or increases privacy).



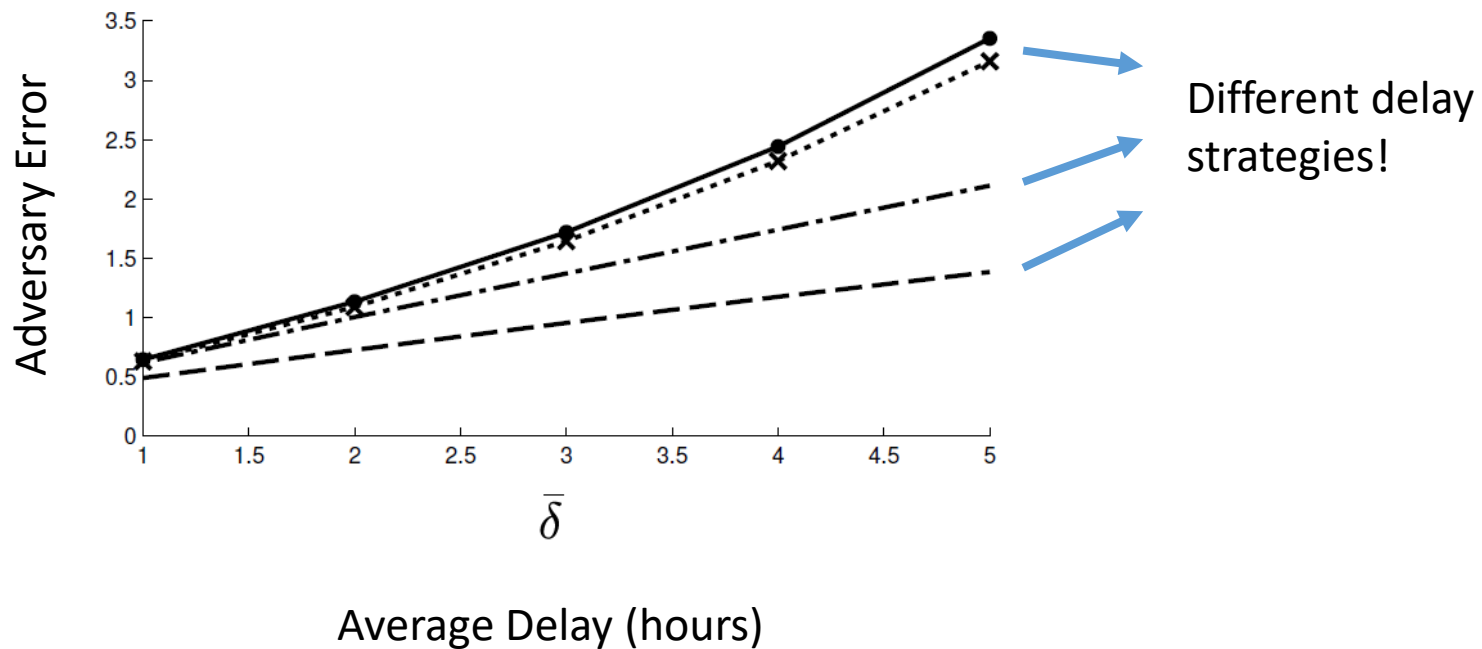
# Delay as a Source of Privacy

- If several users cooperate and “mix” their location reports:



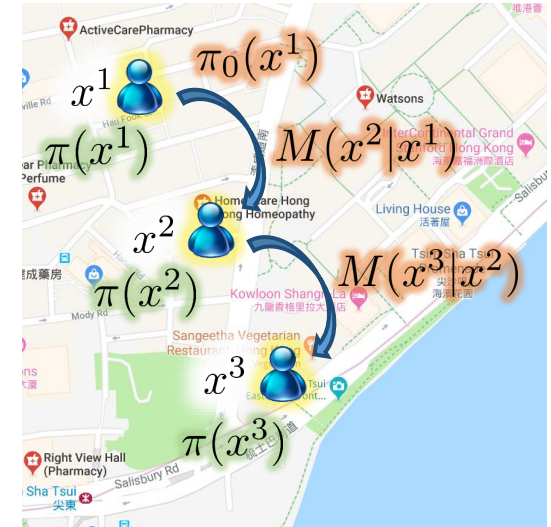
## Example of Performance (from Mix literature)

- Average Adversary Error (MSE) of estimating the mobility profile of a user (not an individual location).

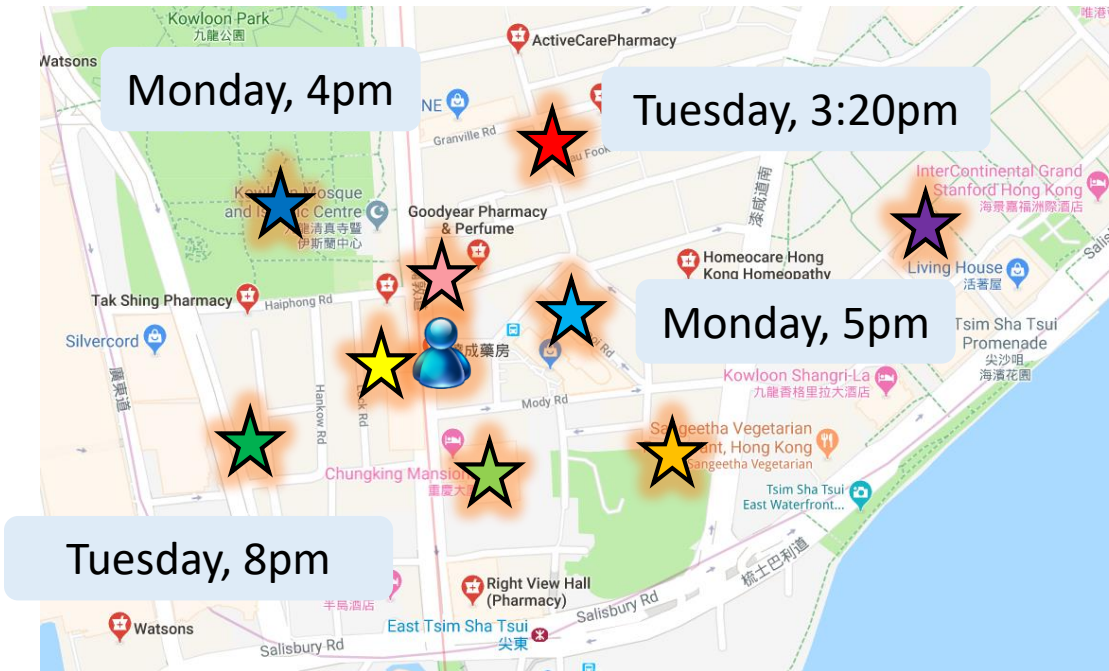


# Other Issues of LPPMs in Practice: More Realistic Mobility Models

- We have seen sporadic mobility models.
- Also, a bit of Markov mobility models.
- In practice, users normally have routines.
  - Leave home for work at the same time.
  - Stay the same time at work.
  - Leave work at the same time.
  - Go to the gym at the same time.
  - ...
- This induces correlations between the user's locations, that can be exploited by an adversary.



# More Realistic Mobility Models



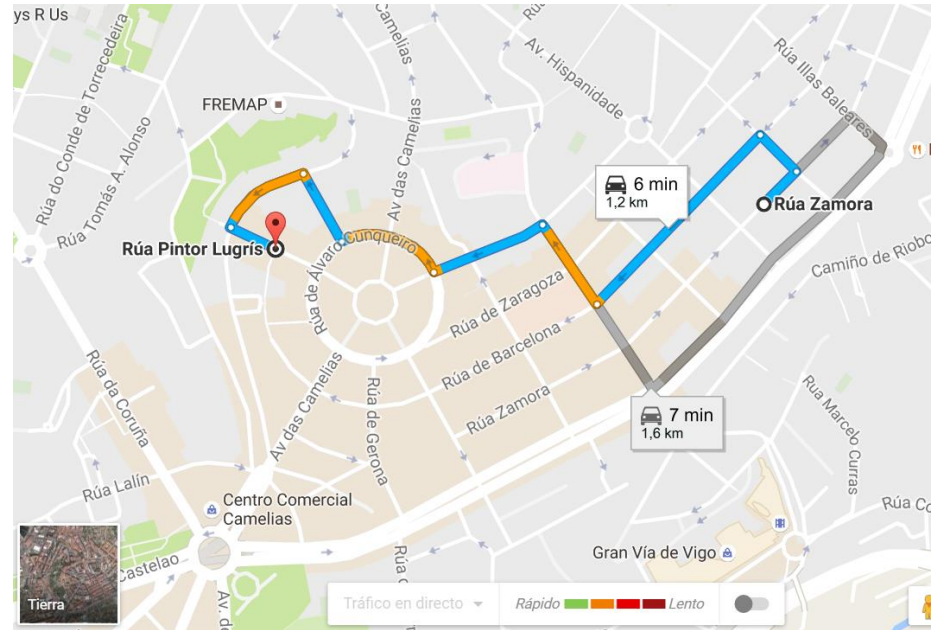
I average them,  
and get the real  
location!!



Defending against this is a difficult challenge. User mobility behaviors are very complex.

# How to Generate Dummy Traces? [Chow, Golle 2009]

- Take polyline from the route offered by Google.
- Generate additional points between existing.
- Points are meant to be equally spaced in time.
- Add random stops.
- Add noise to each vertex to simulate GPS.
- Sample the available vertices and report them.



So...we need more research!

This new spatial cloaking mechanism is a game changer!

