DSC 291 Privacy-sensitive Data Systems (week 2a)

Haojian Jin

Logistics

1. Discussion leader signup
2. Find your partners (2-3 people each team).
3. Reviews due Thursday morning.
4. Arrange a meeting with me in the next two weeks.

Recap: Why is Privacy Hard?

- #1 Privacy is a broad and fuzzy term
- #2 Technological Capabilities Rapidly Growing
- #3 Strong Incentives for Companies to Collect Data
- #4 Same Device/Data, Different Perspectives
- #5 Wide Range of Privacy Risks
- #6 Burden on End-Users Too High
- #7 Low Knowledge, Awareness, Motivation by Devs
- #8 Companies Get Little Pushback on Privacy
- #9 Unclear What the Right Thing To Do Is
- #10 Probabilistic and Emergent Behaviors

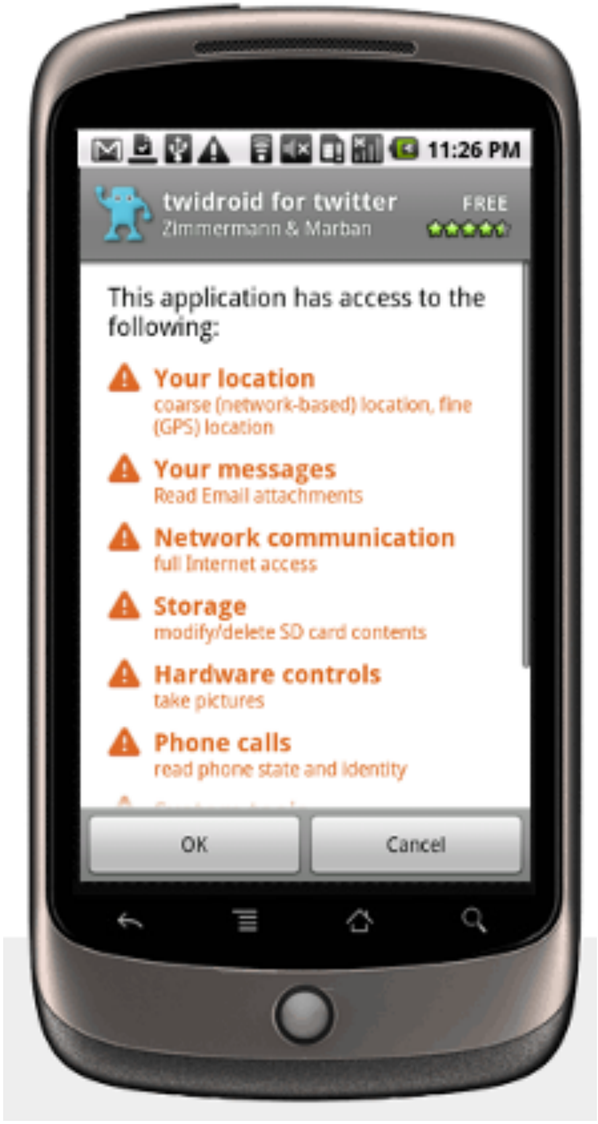
Today's topic:

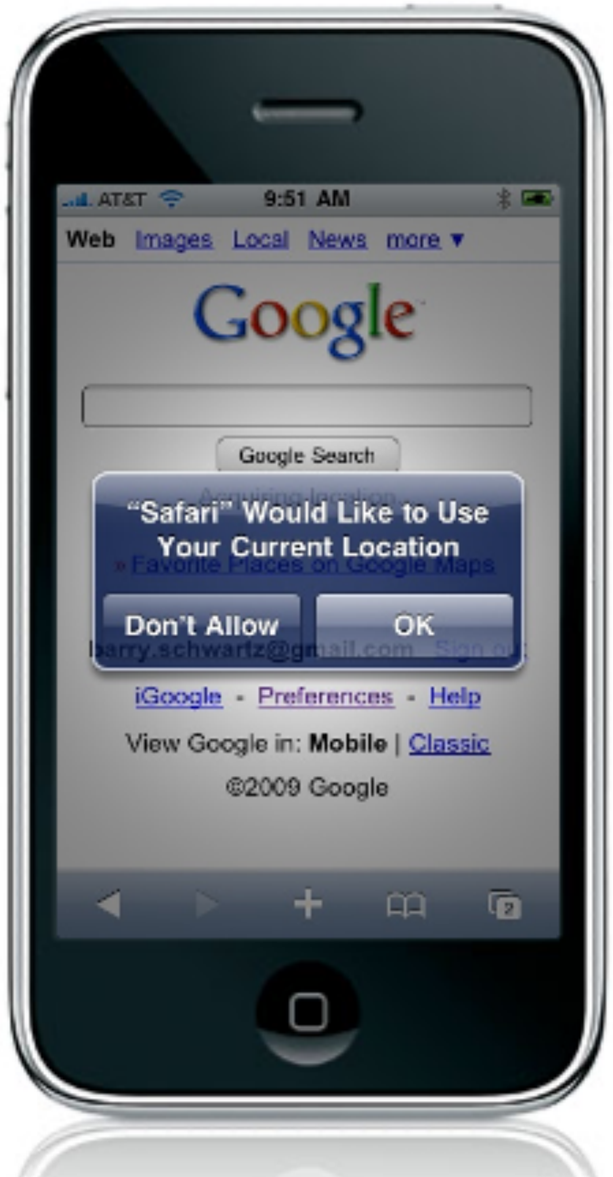
Location Privacy

Schedule - Context of Location Privacy

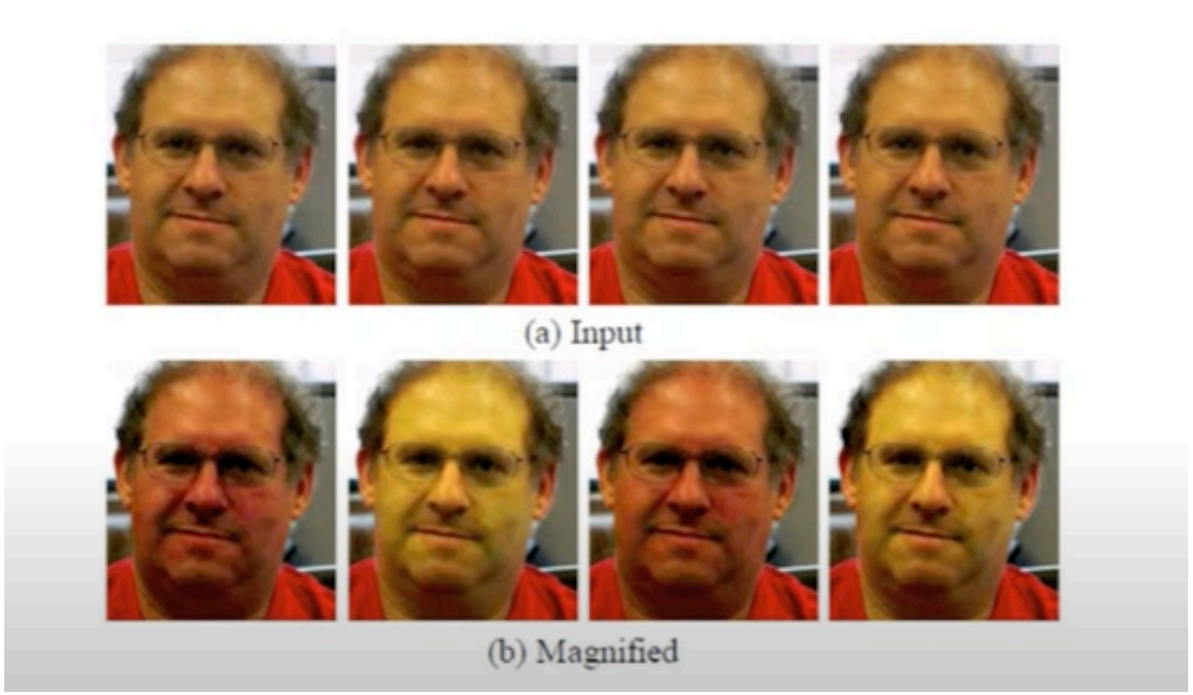
- Why location privacy?
- Location-based applications
- Locating technologies
- Protecting location privacy
- Beyond location privacy
- Purpose framework
  - Data collection, data processing, data usages

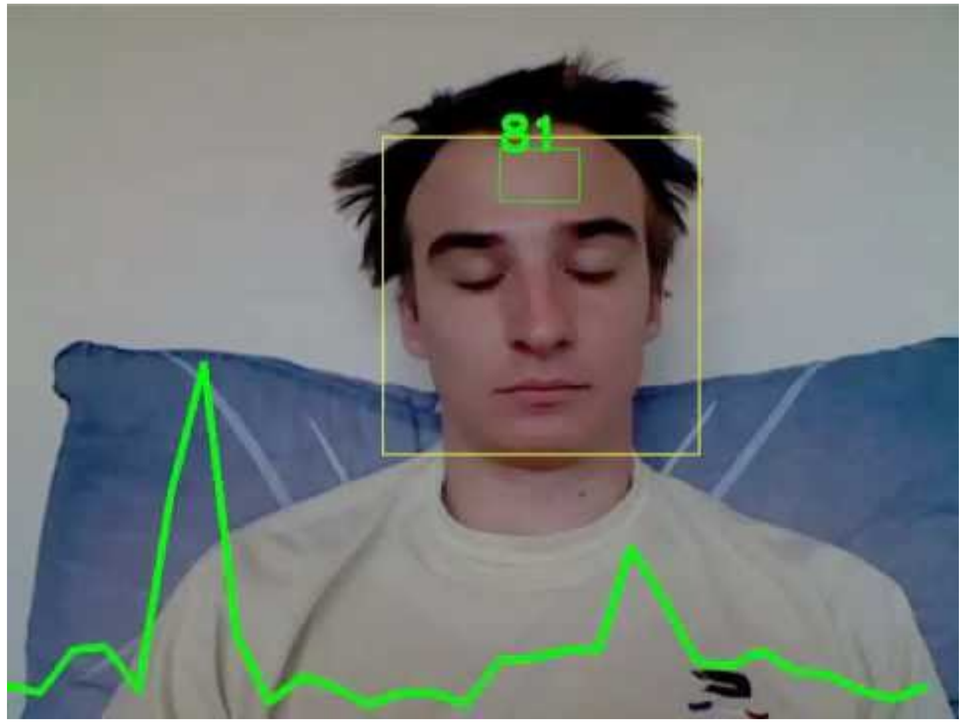
Why location privacy?





Why not face privacy?



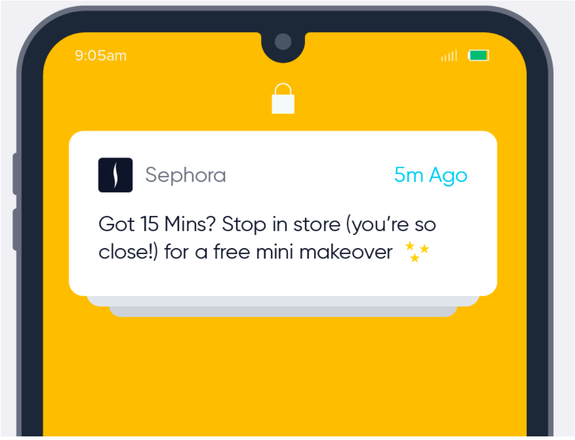


http://people.csail.mit.edu/mrub/evm/

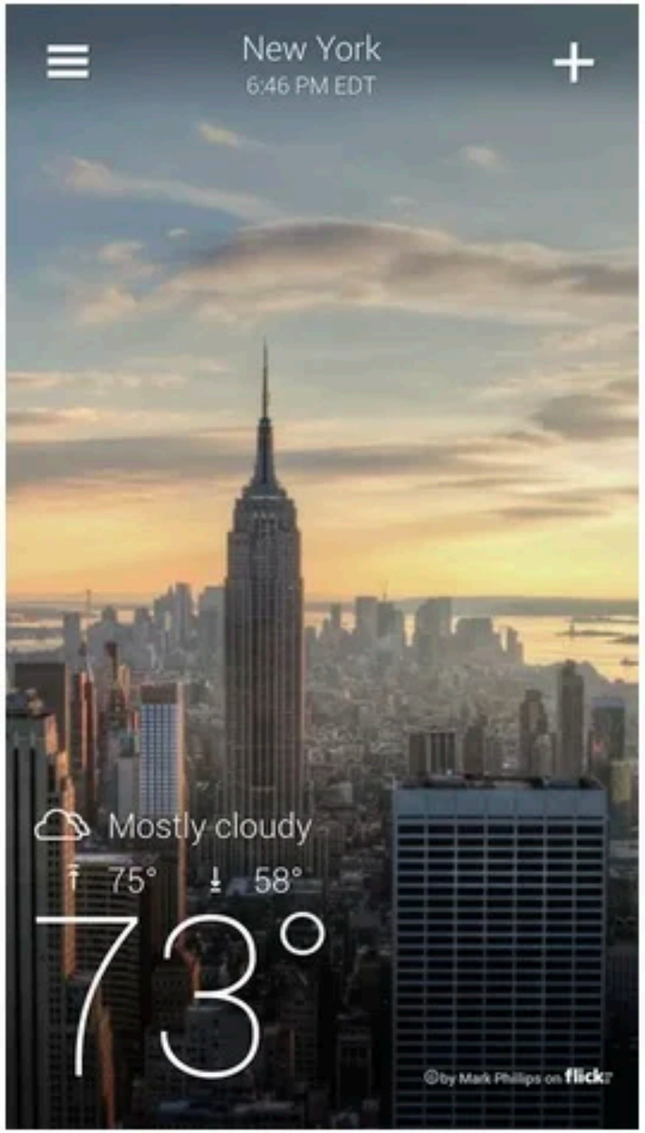
Why location privacy?

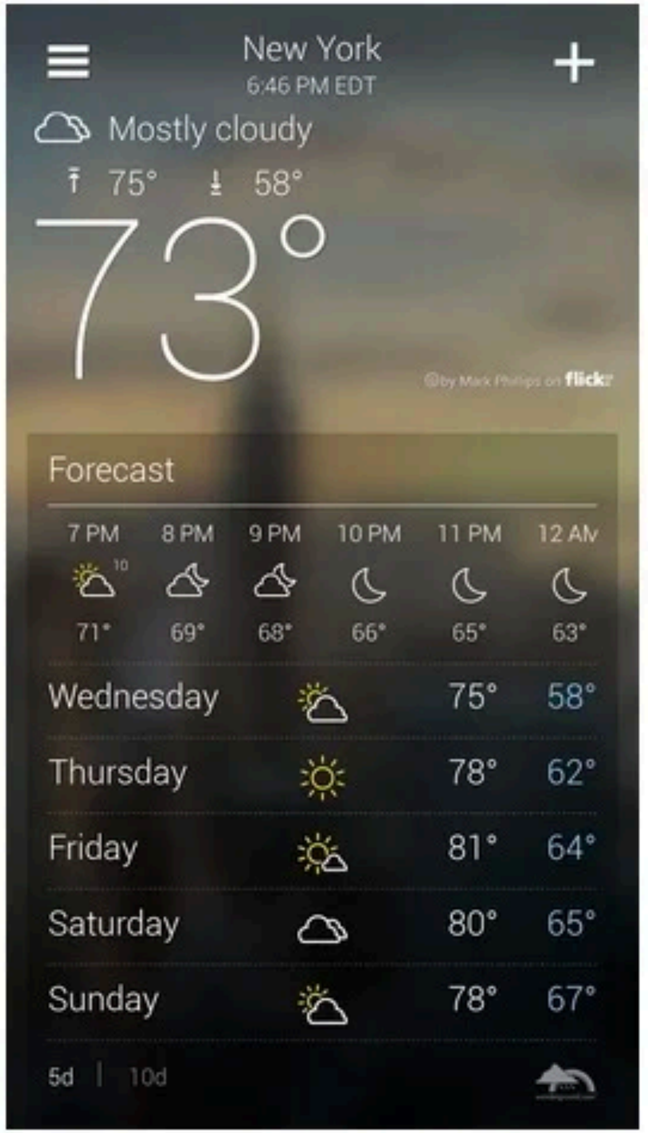
- Location privacy is one of the obvious privacy problems.
- It enables numerous applications.
- It has clear “dangerous” outcome.
- It is a great lens for us to study privacy.

Applications: Advertising/Marketing

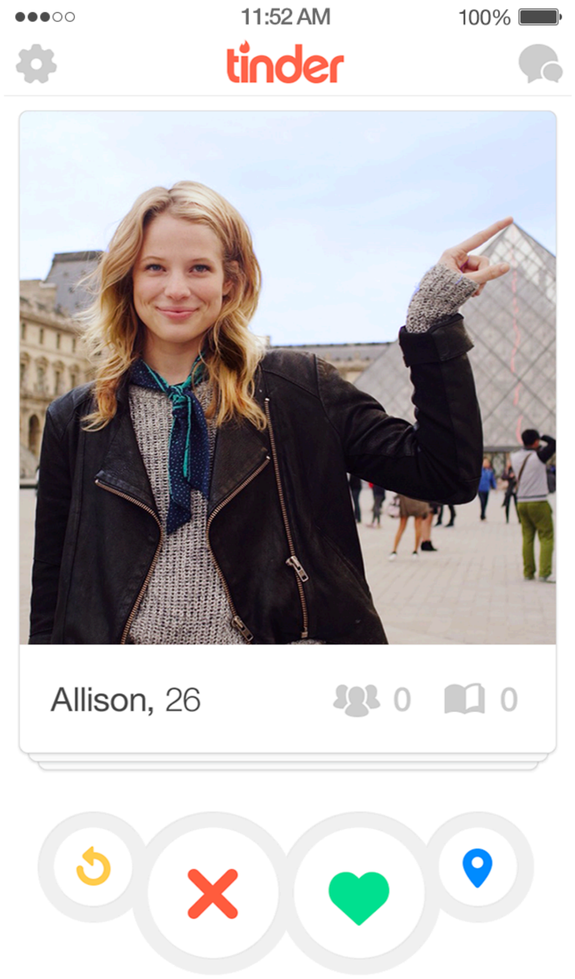


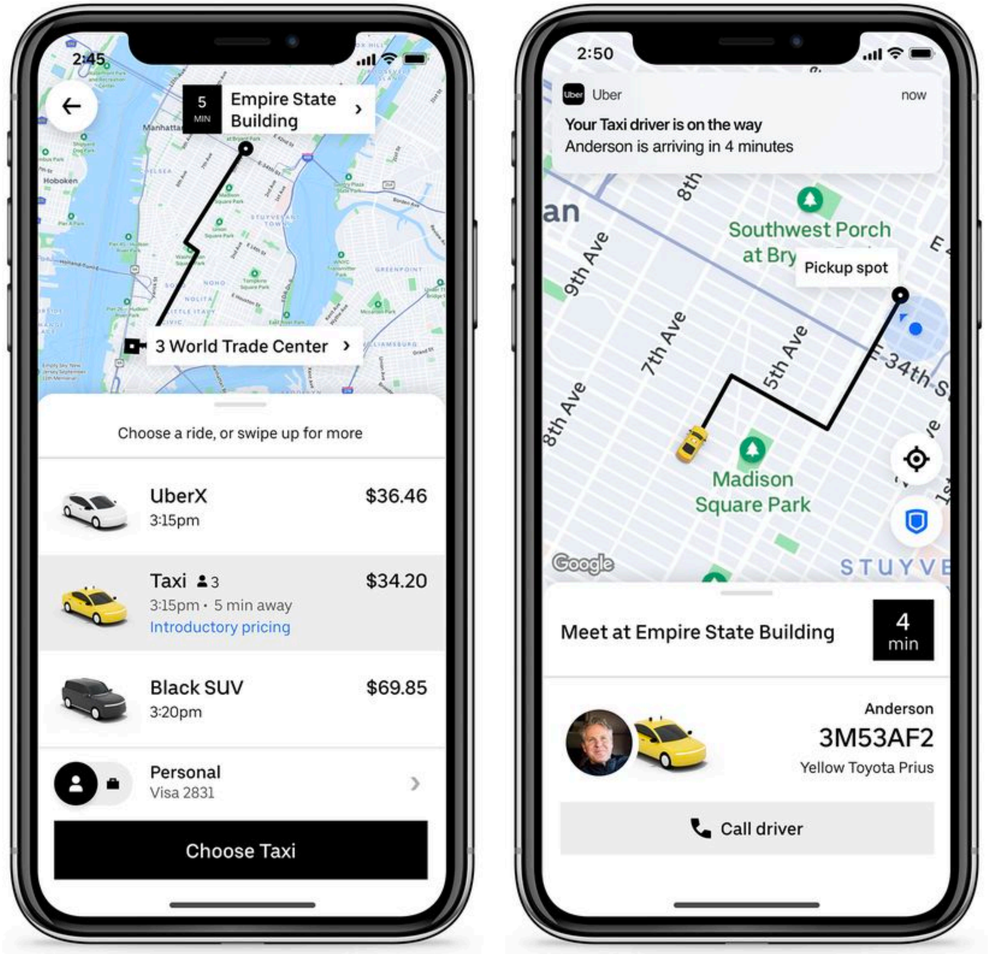
Applications: Information services



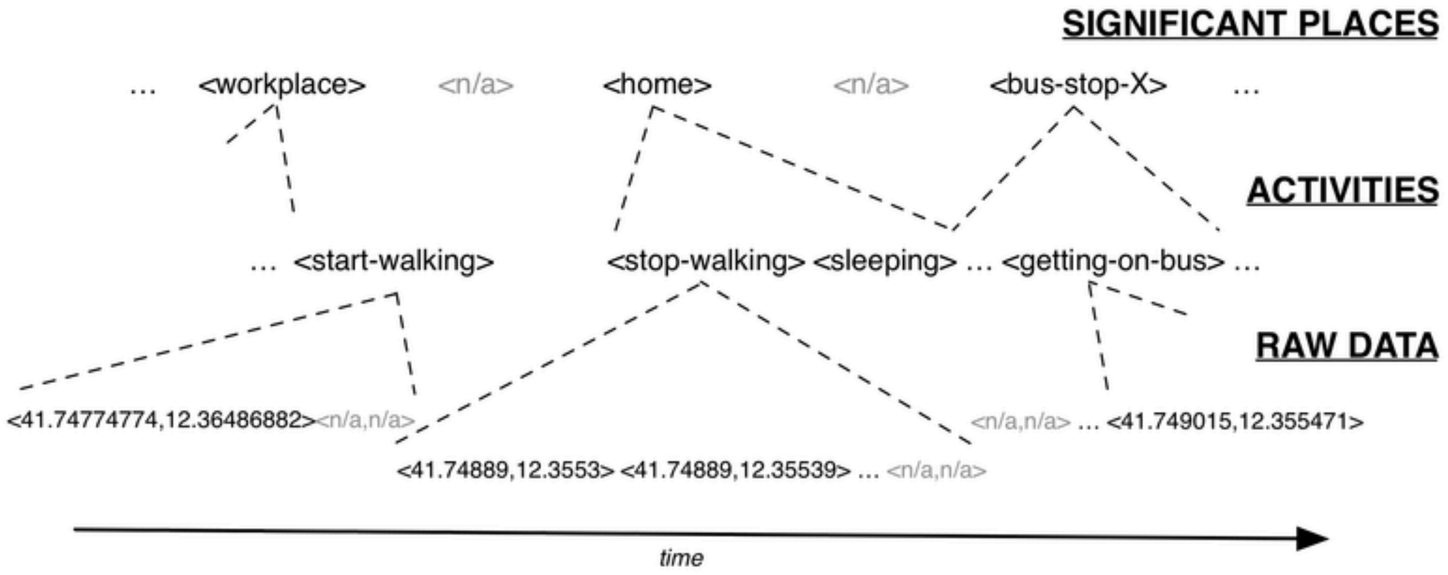


Applications: Friend-finding | Micro-coordination

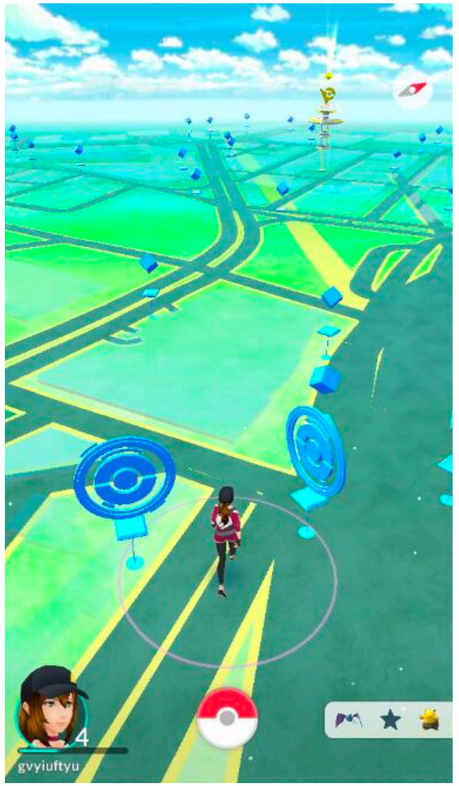




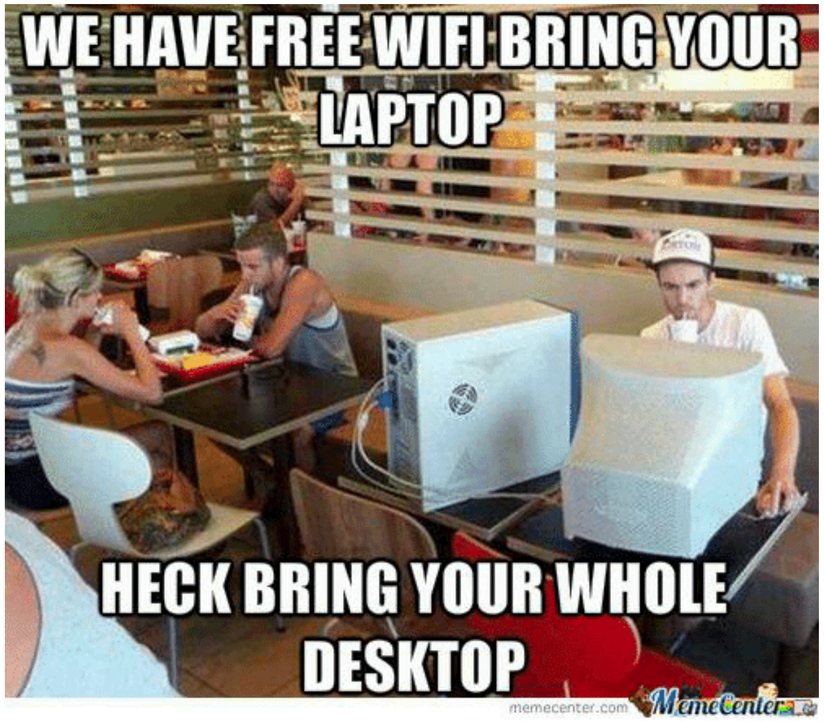
Applications: Activity recognition



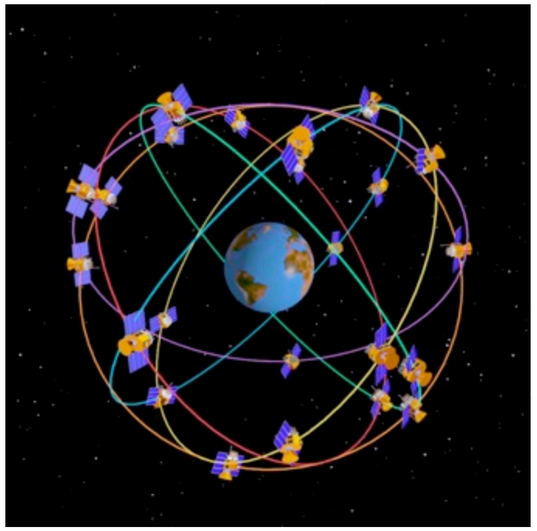
Applications: Games

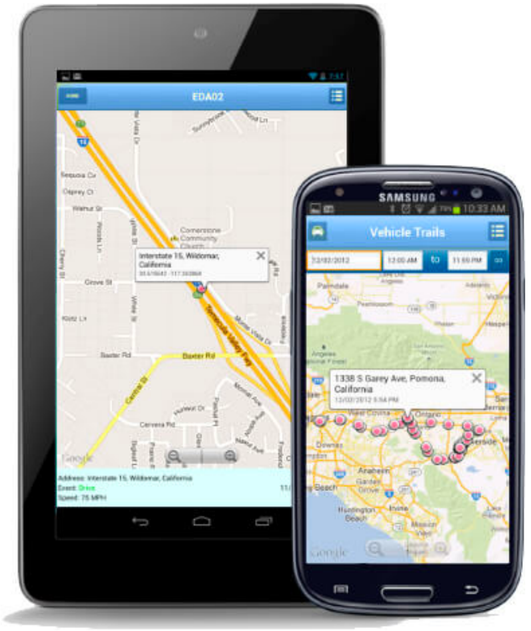


Locating technologies: mobile devices



Locating technologies: GPS



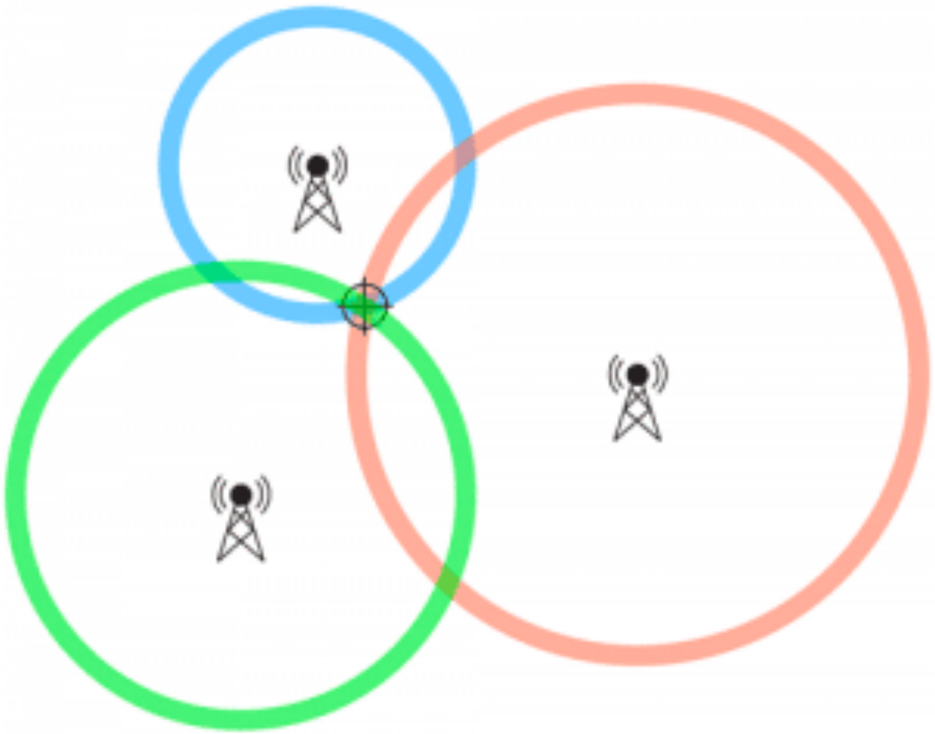


2009: 150 million GPS-equipped phones shipped

2014: 770 million GPS-equipped phones shipped (~5x)

Locating technologies: WiFi fingerprints

Locating technologies: Cellular/Wifi Triangulation



RSSI signal strength

Meter-level accuracy

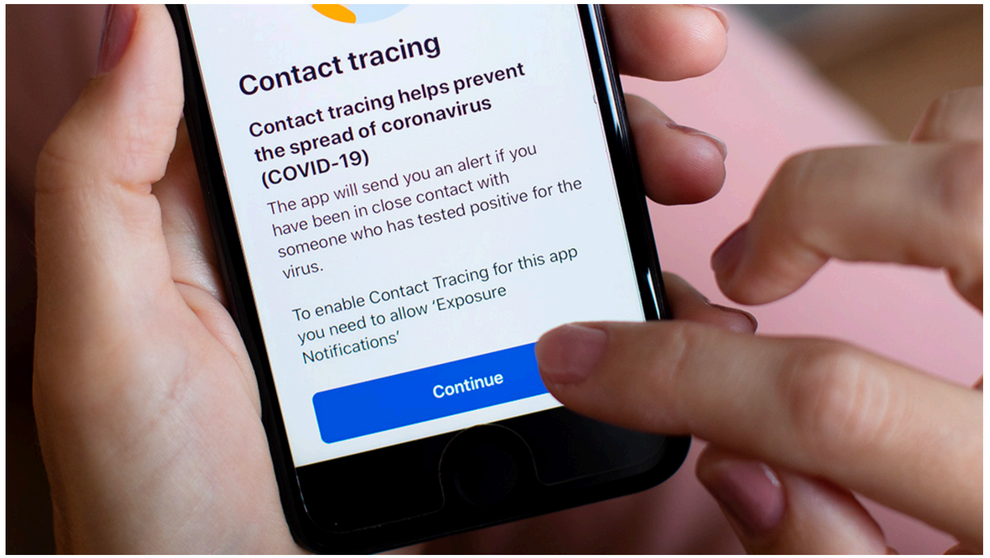
Demo


Corona
Positioning Adjacent Devices
with Asymmetric Bluetooth Low Energy RSSI Distributions

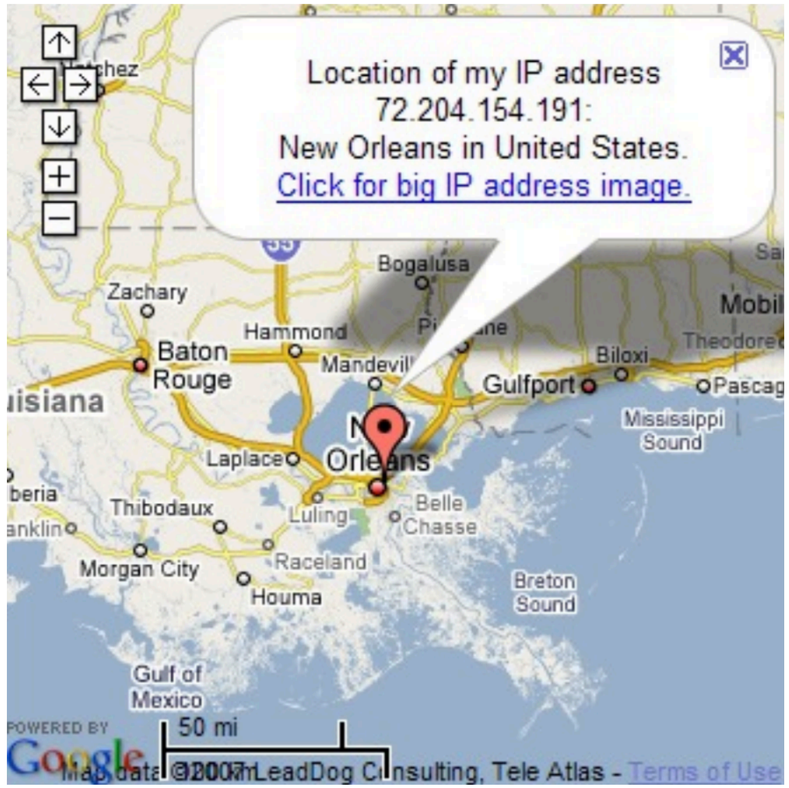
Locating technologies: Acoustic/Radio/UWB/RFID

AirTag features:

- Precision finding
- Bluetooth LE
- Siri Support
- Over a year of battery life
- IP67 rating



Locating technologies: IP location



Default Activity Report

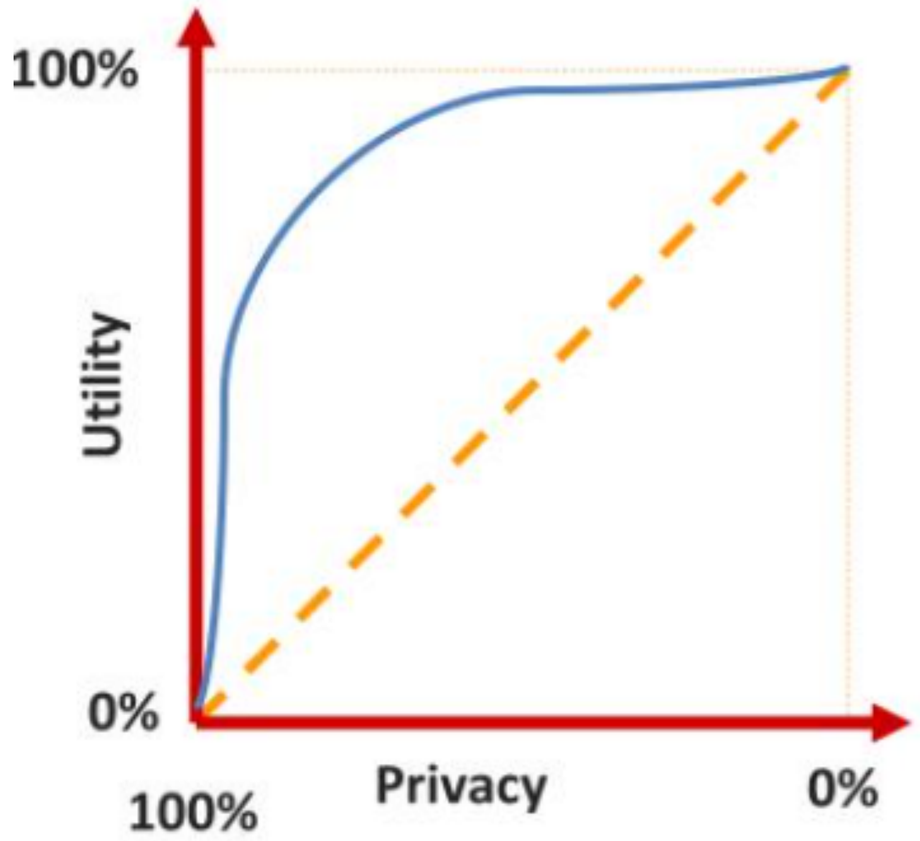
csv

Timestamp	IP Address	City	State/Region	Domain	First Referrer Type	Browse
2023-01-15T00:02:07-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-15T00:02:03-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-15T00:01:34-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-15T00:01:32-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-14T23:59:44-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-14T23:58:26-0800	150.66.46.238	Suita	Ōsaka	shift-3.com	direct	Chrom
2023-01-14T22:33:14-0800	112.254.213.100	Qingdao	Shandong	haojianj.in	direct	Safari Mo
2023-01-14T22:19:01-0800	128.12.122.139	Stanford	California	haojianj.in	social	Chrom
2023-01-14T20:53:38-0800	2001:250:500f:3ea:51cb:e055:266c:8462	Nanjing	Jiangsu	shift-3.com	search	Chrom
2023-01-14T19:37:37-0800	2600:4040:5786:3a00:8a3:d37d:9b79:6df9			haojianj.in	direct	Microsoft E
2023-01-14T18:41:52-0800	67.21.64.115	Los Angeles	California	haojianj.in	direct	Chrom
2023-01-14T18:39:30-0800	67.21.64.115	Los Angeles	California	haojianj.in	direct	Chrom
2023-01-14T18:38:44-0800	67.21.64.115	Los Angeles	California	haojianj.in	direct	Chrom
2023-01-14T18:38:33-0800	67.21.64.115	Los Angeles	California	haojianj.in	direct	Chrom
2023-01-14T18:38:24-0800	67.21.64.115	Los Angeles	California	shift-3.com	direct	Chrom
2023-01-14T17:52:43-0800	72.76.184.152	Leonia	New Jersey	shift-3.com	direct	Chrom
2023-01-14T17:38:55-0800	69.181.77.191	Cupertino	California	shift-3.com	direct	Safari 1
2023-01-14T17:28:39-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom
2023-01-14T17:28:35-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom
2023-01-14T17:28:29-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom
2023-01-14T17:28:28-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom
2023-01-14T17:28:18-0800	67.171.74.117	Pittsburgh	Pennsylvania	blog.haojianj.in	internal	Chrom
2023-01-14T17:26:37-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom
2023-01-14T17:26:25-0800	67.171.74.117	Pittsburgh	Pennsylvania	shift-3.com	direct	Chrom

Protecting location privacy

- System architecture (pre-fetching, pre-processing, decentralized)
  - How you get location
  - Where and how data stored and used
  - User interface and policies
    - When is it shared
    - How is it displayed
  - How do people manage in practice

System architecture: Privacy-utility tradeoff



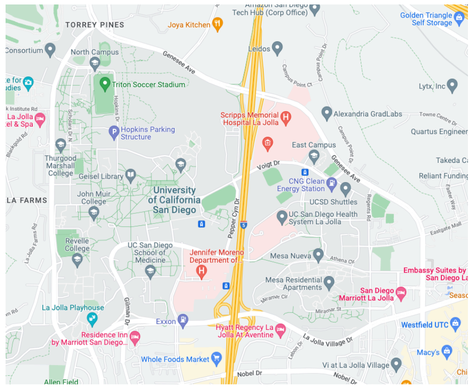
System architecture: How you get location?

Different time-to-live

Real-time	Traffic, Parking spots, Friend Finder
Daily	Weather, Social events, Coupons
Weekly	Movie schedules, Ads, Yelp!
Monthly	Geocaches, Bus schedules
Yearly	Maps, Store locations, Restaurants

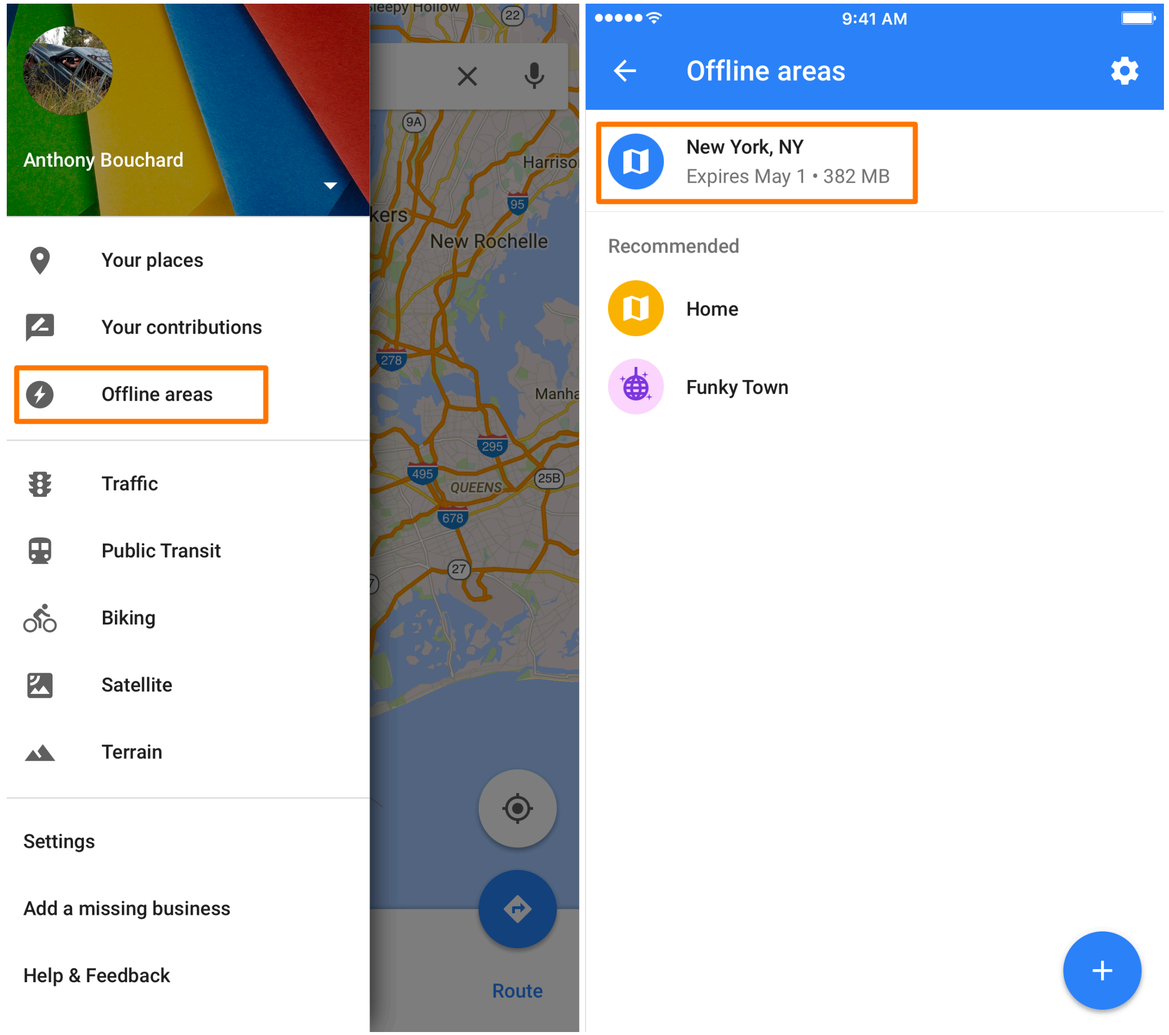
System architecture: How you get location?

- **Pre-fetch** all the content you might need for a geographic area in advance
  - *SELECT * from DB where City='San Diego'*
- Then, use it locally on your device only
  - Assume that you determine your location locally using WiFi or GPS
  - So a content provider would only know you are in San Diego

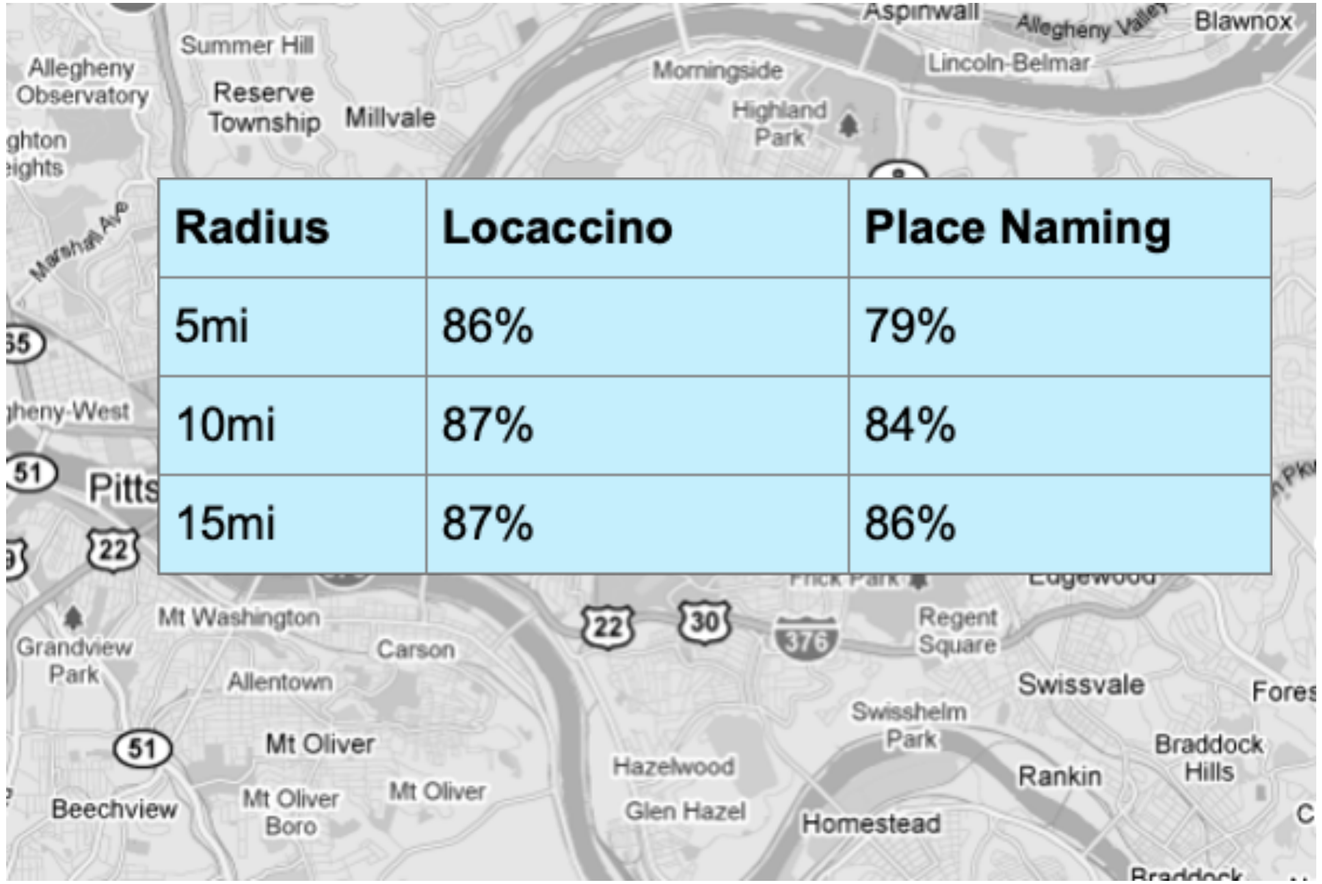


Feasibility of Pre-Fetching

- Are people's mobility patterns regular?
- **Pre-fetching** useful only if we can predict where people will be
  - Locaccino: 2000 users, 460k traces
  - Place naming: 26 people, 118k traces
- For each person, 5mi radius around two most common places (home + work) accounts for what % of mobility data?



Feasibility of Pre-Fetching



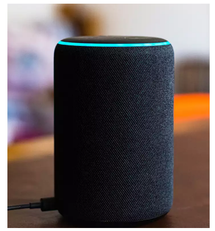
Radius	Locaccino	Place Naming
5mi	86%	79%
10mi	87%	84%
15mi	87%	86%

Feasibility of Pre-Fetching

Data Type	Added %	Removed %	Modified %
Weather	25.00	25.00	67.26
Events	5.28	5.35	11.75
Yelp POI	0.15	0.06	0.04
MSN POI	6.69	6.80	1.43
Bus Schedule	0.00	0.00	0.15
Map Tiles	0.00	0.00	0.00

- Content doesn't change that often
  - Average amount of change per day (over 5 months)

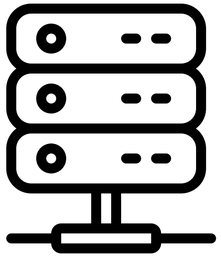
Preprocessing: 77% Smart home apps do not need raw data.

	Sensor	*Raw*	*Needed data*
Hello visitor			
Noise level			55 db

Peekaboo

Smart home app →

```
@purpose: To measure device engagement.
WeeklyUsageHours{
  // operator [properties]
  inject [weekly] ->
  pull [smart TV driver] ->
  aggregate [sum duration] ->
  post [duration]
}
```



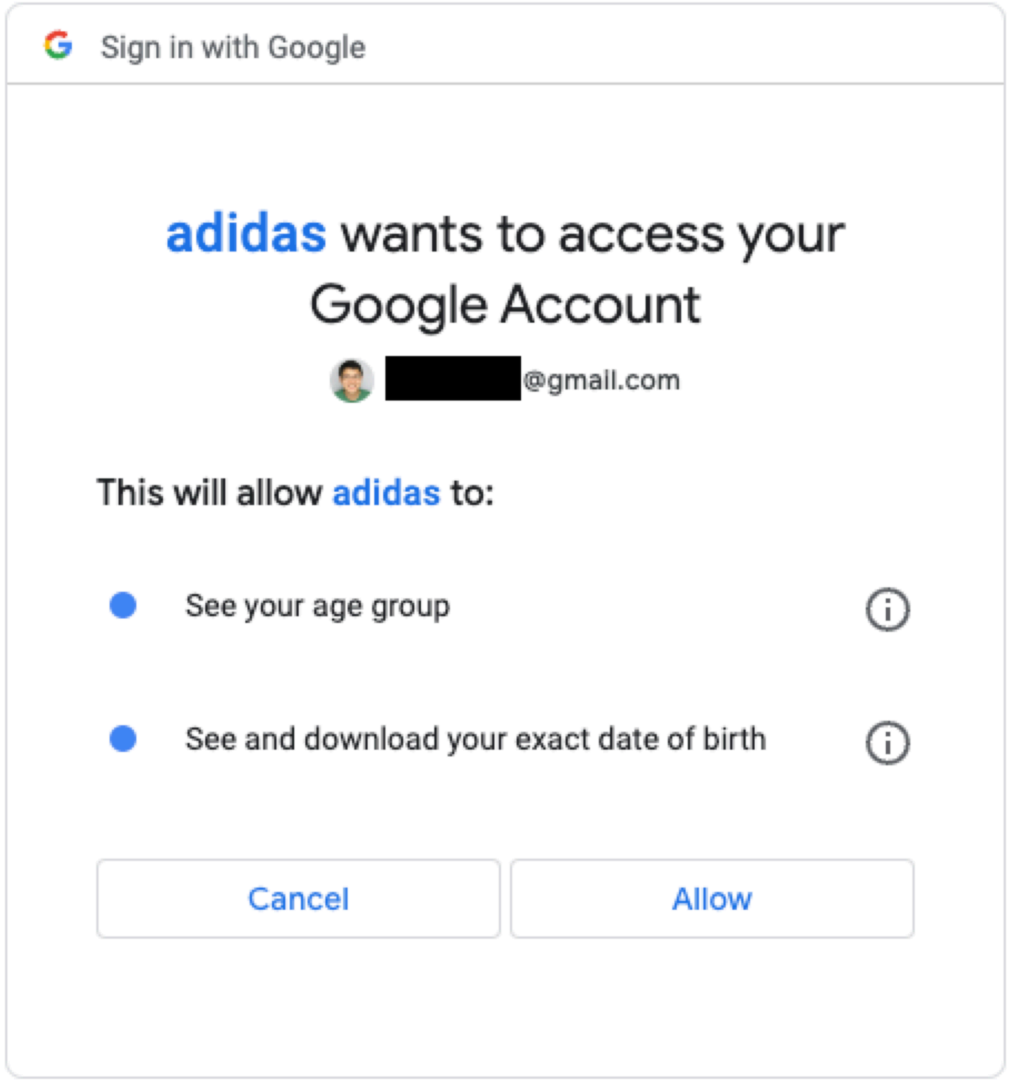
Edge devices

A local hub

Cloud

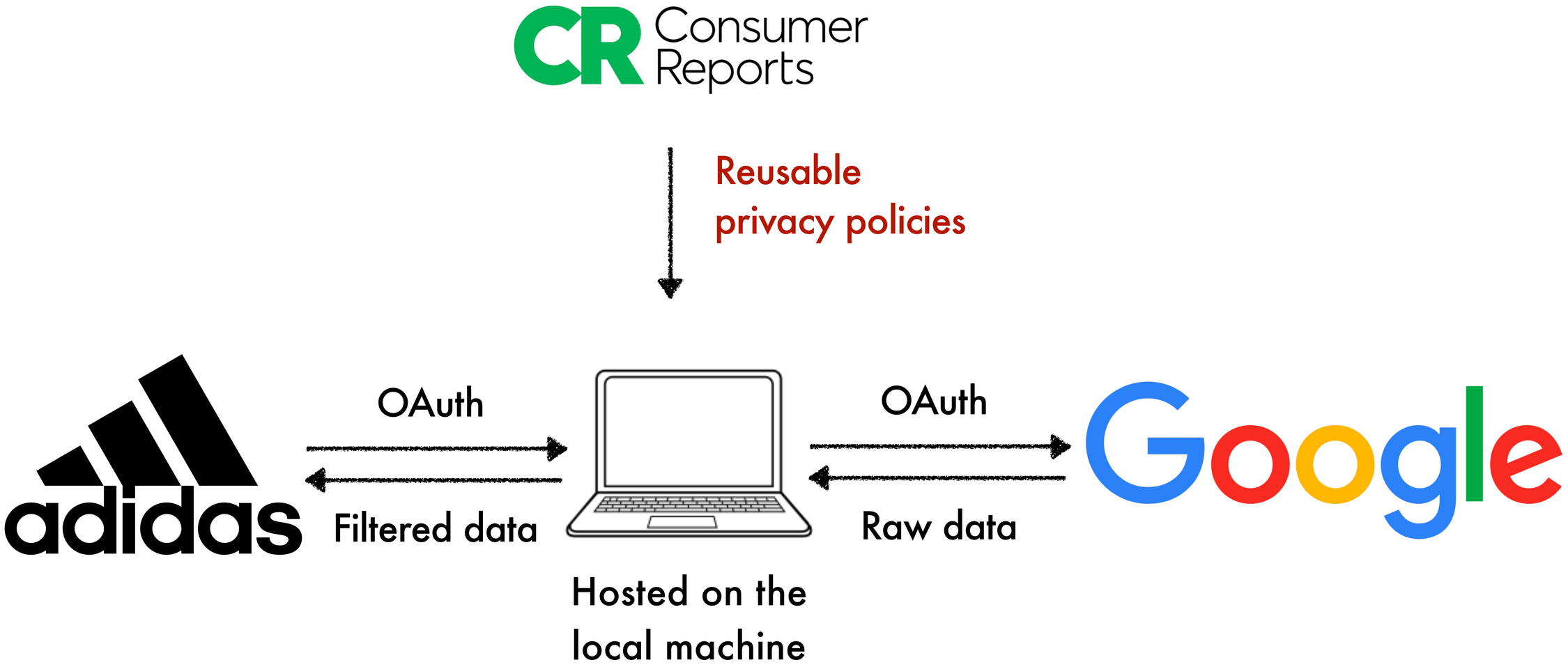
"Privacy firewall"

Developers often access more data than needed.



- Coarse granularity
- All-or-nothing access
- Limited awareness
- Limited control

Privacy Firewall for Personal Data.



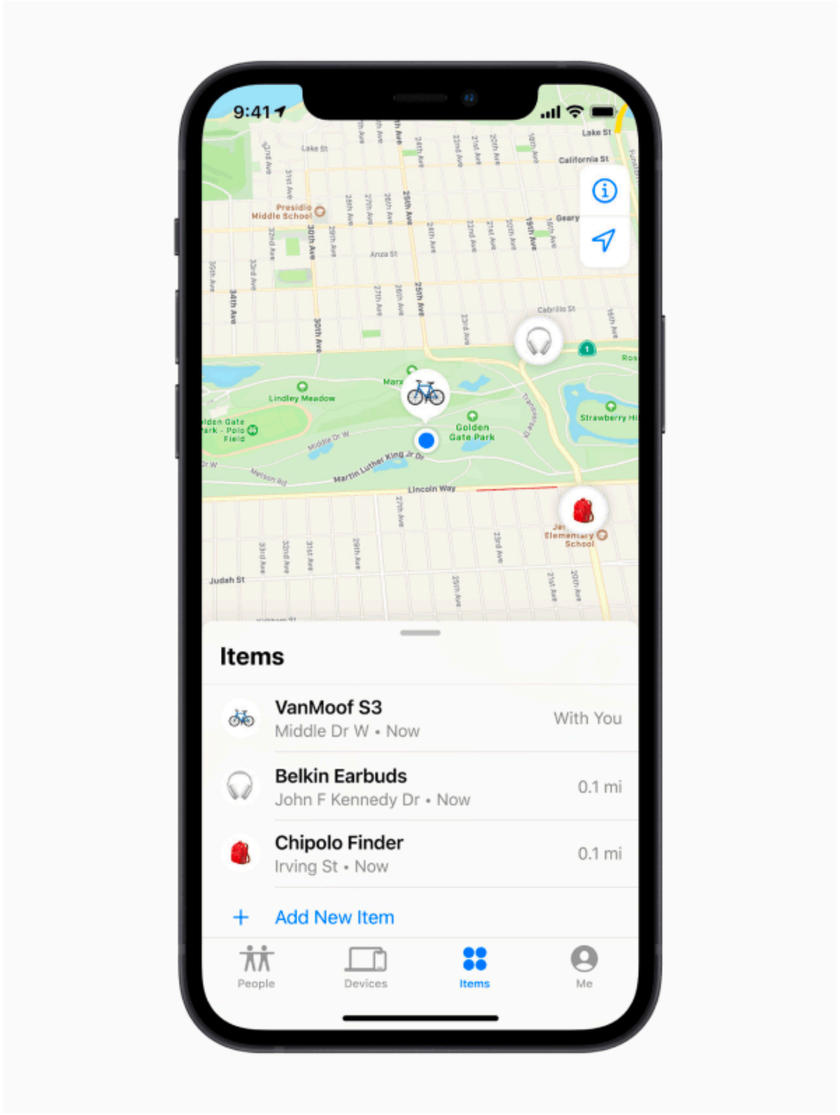
Apple AirTag

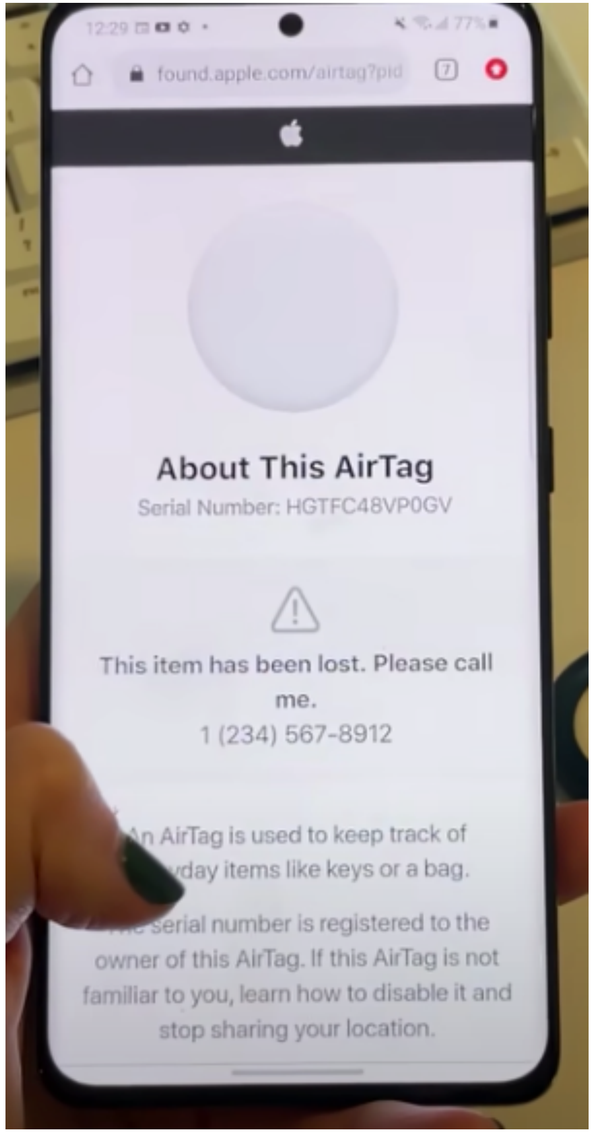
AirTag features:

- Precision finding
- Bluetooth LE
- Siri Support
- Over a year of battery life
- IP67 rating

$29

New applications: find lost tagged item & lost tags





How it works?

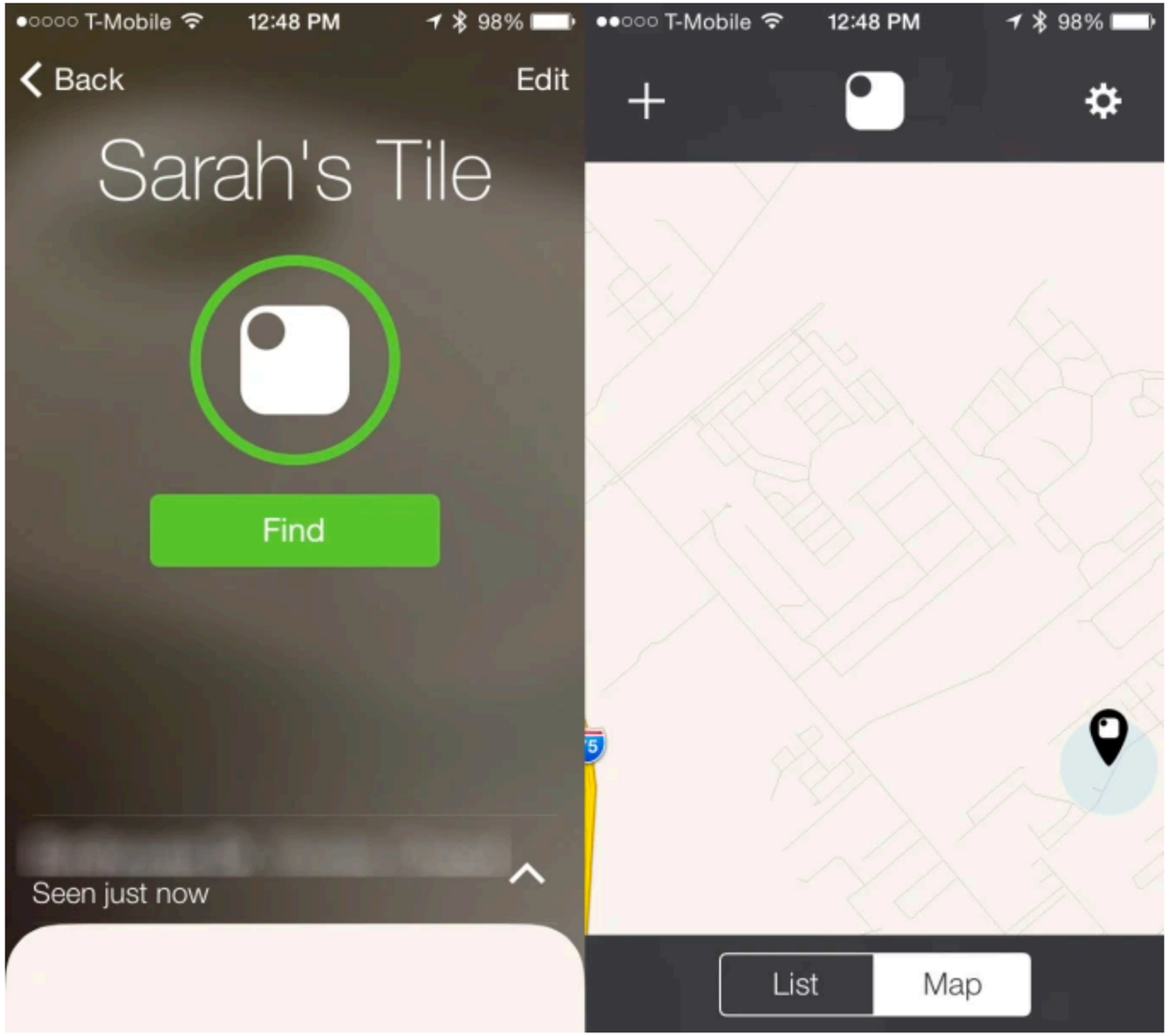
- No internet connection.
- Ad-hoc connections to a billion iOS/Macs.
- Each AirTag sends a unique encrypted bluetooth identifier.
- Other app devices can detect it and relay the location to an owners' account.

My story

Airtag v.s. Tile

Ultra-accurate
Ultra-fast
Apple device network

They work too great!

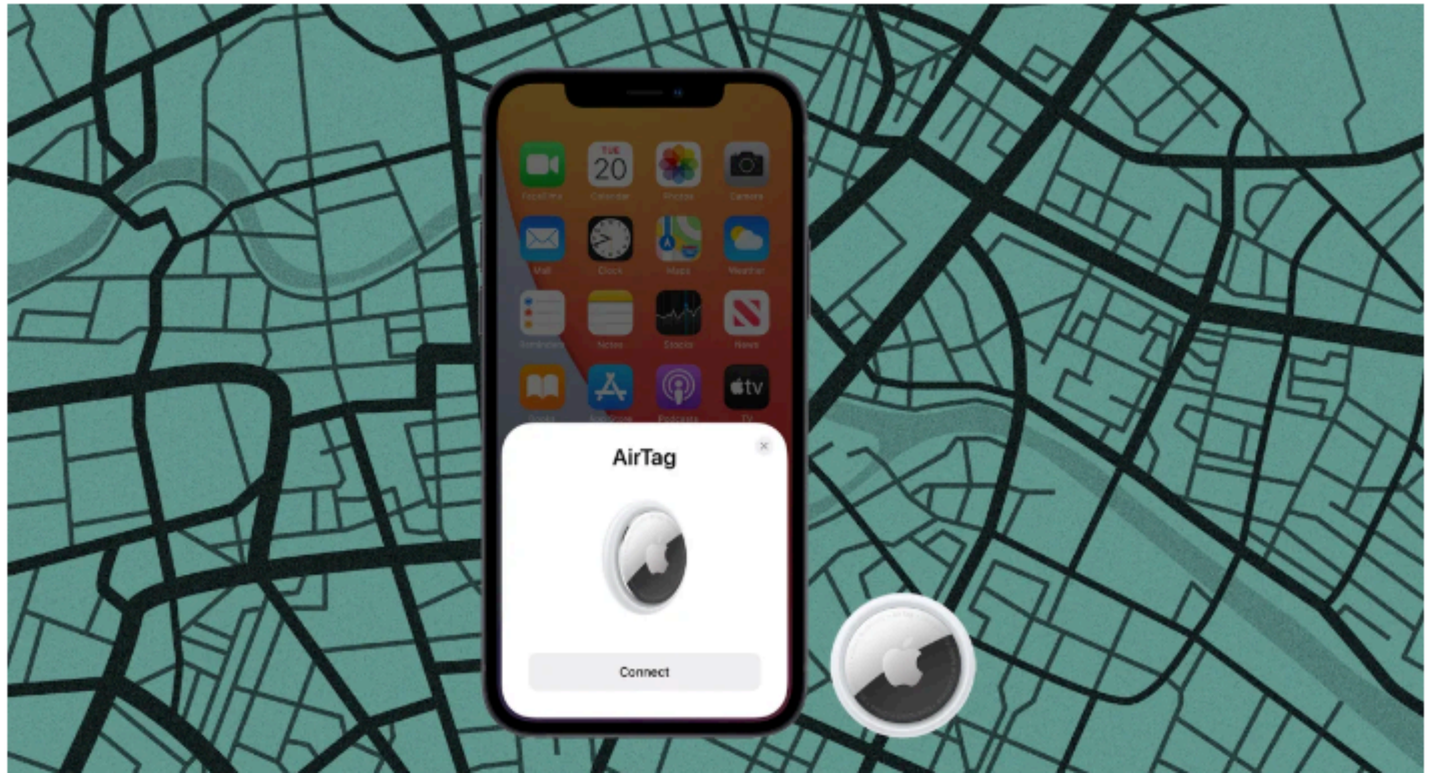


Privacy-first and Stalker-proof

04-22-21

How Apple designed AirTags to be privacy-first and stalker-proof

AirTag users, owners of other Apple devices, and even people who don't own a single Apple product—the company wants its new tracker to respect everybody's privacy.



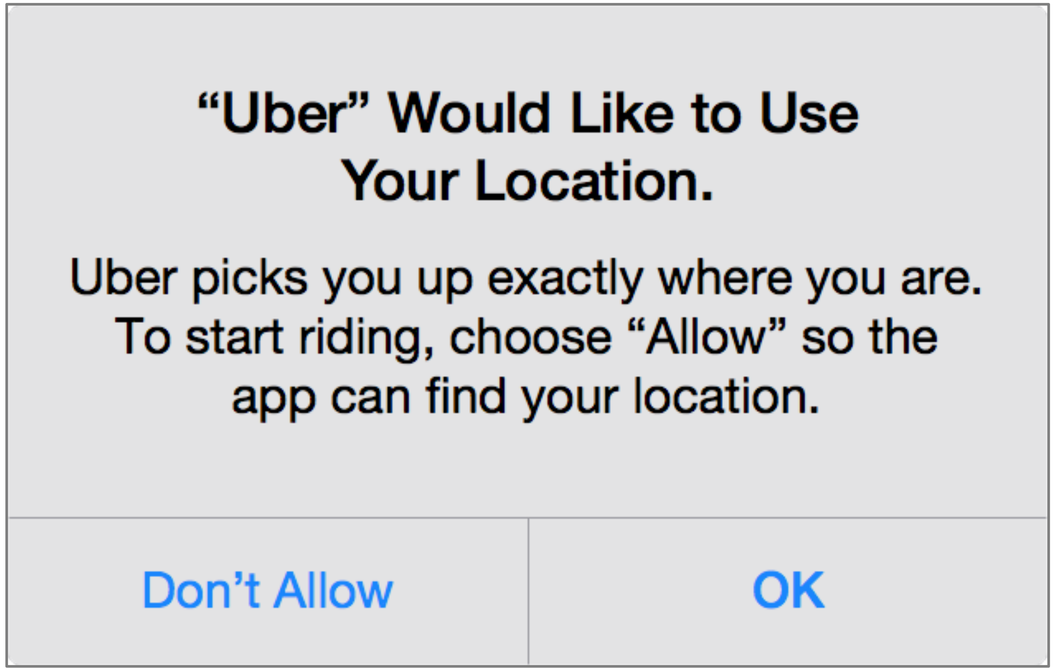
Question?

Procedural v.s. outcome approach?

Who’s privacy?

Who are the stalkers?

Does apple ask for your permissions to scan the tags?



Who's privacy? Who are the attackers?

- AirTag owners' privacy
- Relay proxies' privacy
- Non-consent human's privacy

- Hackers
- Stalkers (owners)
- Apple

End-to-end encryption

- AirTag sends a secure Bluetooth signal to nearby Apple devices.
  - Enrolled in Find My network by default.
- The devices send the location of the AirTag to iCloud.
- Location data and history are never stored on the AirTag.

What apple knows?

- Apple does not know the location of your AirTag and the identity of the devices that help find it.
- Apple knows the encrypted string of your location data and the fact that someone finds your AirTag.
- Apple wants to do it. Because it reduces the amount of received legal process form law enforcement agencies.

What hackers know?

- Disassemble the hardware?
  - No local storage
- Pretend to be the relay devices.
  - The bluetooth msg is encrypted.
- Hack Apple’s server
  - Apple does not know the location of your AirTag and the identity of the devices that help find it.

Stalkers and non-consent human’s privacy

What about Android?

Non-iOS users

- AirTag will play a sound if separated from its owner is currently three days.
  - Three days is not baked into the Tags. It's a server-side setting.
  - They changed it to 8 hours recently.
- If someone finds an AirTag after hearing it make a sound, they can use any NFC device to see if its owner marked it as lost and help return it.

Contact Tracing

Why does privacy matter?

Exclusive: Government scientist Neil Ferguson resigns after breaking lockdown rules to meet his married lover

Prof Ferguson allowed the woman to visit him at home during the lockdown while lecturing the public on the need for strict social distancing

By Anna Mikhailova, DEPUTY POLITICAL EDITOR, Christopher Hope, CHIEF POLITICAL CORRESPONDENT, Michael Gillard *and* Louisa Wells

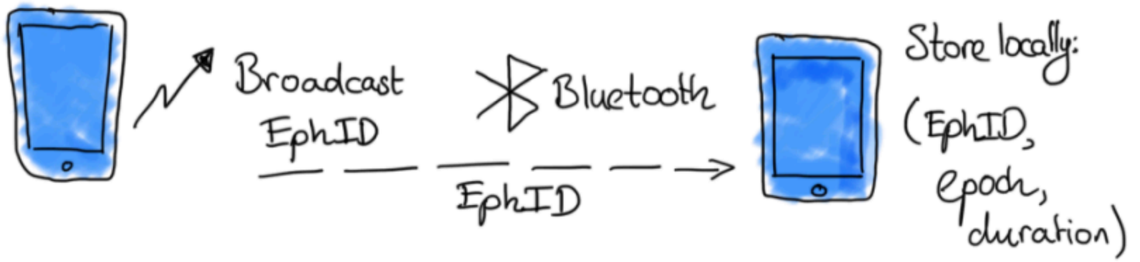
5 May 2020 • 7:7pm

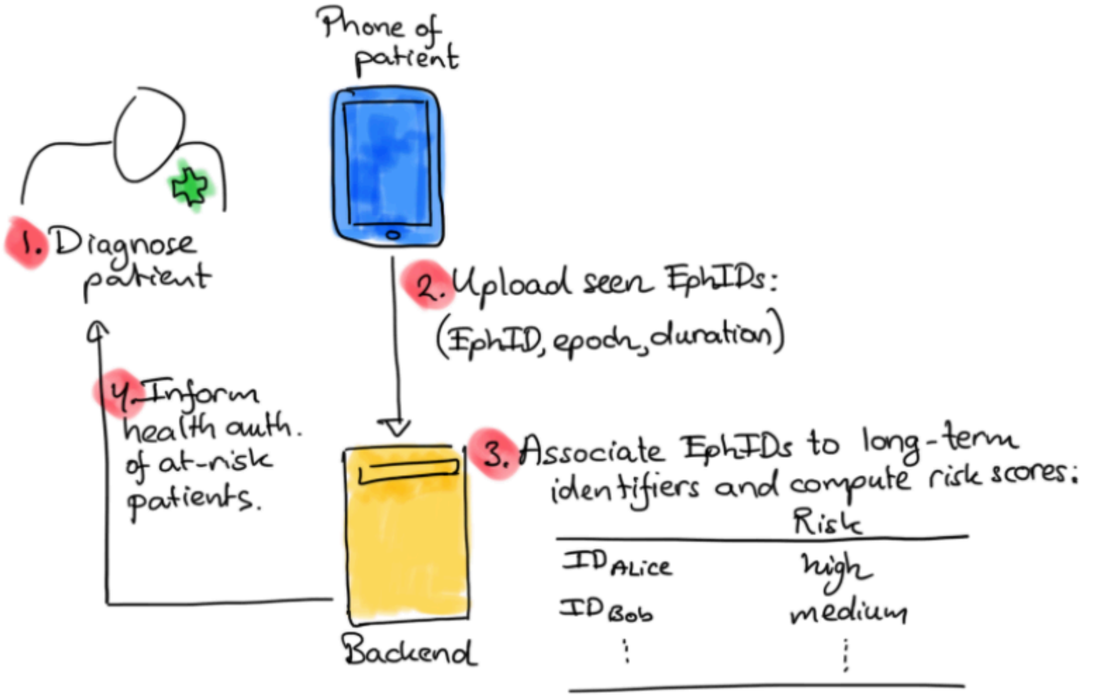
Neil Ferguson and Antonia Staats

The scientist whose advice prompted Boris Johnson to lock down Britain resigned from his Government advisory position on Tuesday night as The Telegraph can reveal he broke social distancing rules to meet his married lover....

1. Social graph
2. Interaction graph
3. Location traceability
4. At-risk individuals
5. Covid-19 positive status
6. Highly exposed locations

A centralized system





Why not centralized?

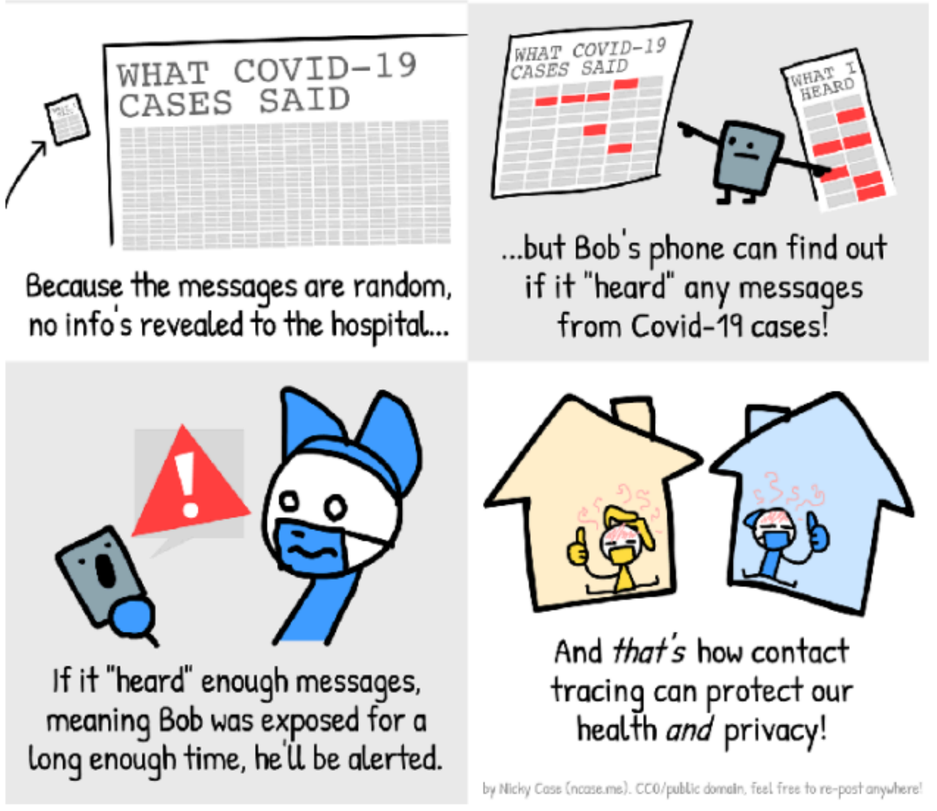
Resecurity
August 29, 2022

Share

COVID-19 data put for sale on the Dark Web

Resecurity, a California-based cybersecurity company protecting Fortune 500, has identified leaked PII stolen from Thailand's Department of Medical Sciences containing information about citizens with COVID-19 symptoms. The incident was uncovered and shared with Thai CERT.

How decentralized systems work?



https://www.youtube.com/watch?v=D__UaR5MQao

Crypto 101: Pseudorandom generator/function

`random.` **`seed`**`(a=None, version=2)`

Initialize the random number generator.

If *a* is omitted or `None`, the current system time is used. If randomness sources are provided by the operating system, they are used instead of the system time (see the `os.urandom()` function for details on availability).

If *a* is an int, it is used directly.

With version 2 (the default), a `str`, `bytes`, or `bytearray` object gets converted to an `int` and all of its bits are used.

With version 1 (provided for reproducing random sequences from older versions of Python), the algorithm for `str` and `bytes` generates a narrower range of seeds.

Changed in version 3.2: Moved to the version 2 scheme which uses all of the bits in a string seed.

Deprecated since version 3.9: In the future, the *seed* must be one of the following types: *NoneType*, `int`, `float`, `str`, `bytes`, or `bytearray`.

Deterministic

Random seed → Long string

(So hard to reverse engineering the seed.)

https://docs.python.org/3/library/random.html

Ephemeral ID generation

How to generate the random messages?

1. Each day, $SK_t = H(SK_{t-1})$

$$EphID_1 \,||\, \ldots \,||\, EphID_n = PRG(PRF(SK_t, \text{“broadcast key”}))$$

Pick a random order to transmit them, the phone stores SK_t it generated during the past 14 days.

Linkable

2. Each epoch draws a random 32-byte per-epoch seed

$$EphID_i = LEFTMOST128(H(seed_i)),$$

Non-linkable

3. At each time window, draws a random 16-byte per-epoch seed

$$EphID_{w,1} \,||\, \ldots \,||\, EphID_{w,n} = PRG(PRF(seed_w, \text{“DP3T-HYBRID”}))$$

Pick a random order to transmit them, the phone stores SK_t it generated during the past 14 days.

Hybrid
temporarily-linkable

Communication

If Alice gets Covid-19, she sends *her* messages to a hospital.

What data does Alice send to the server?

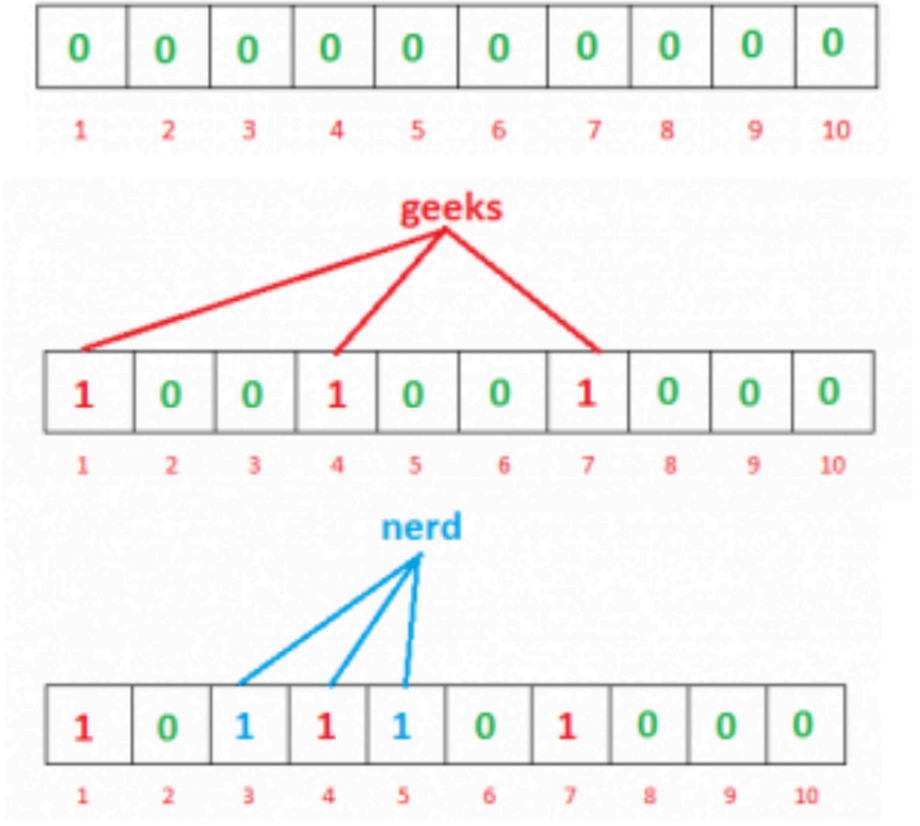
1. The backend collects the pairs (SK_t, t) of COVID-19 positive users.

2. Users have more fined control over their data. The users upload selected {i, Seed_i} to the server.

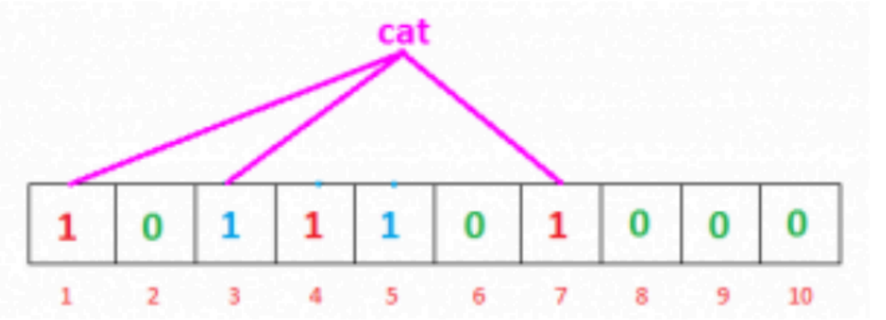
3. Users have more fined control over their data. The users upload selected {w, Seed_i} to the server.

Bloomfilter 101:

Insert



Lookup



Lookup

What data does Bob receive from the server?

1. Phones periodically download these pairs (SK_t, t) and use these pairs to reconstruct the list ephIDs.

2. The server maintains a public Cuckoo filter. Each smartphone uses the filter $\mathtt{F}$ to check.

Non-zero false positive!

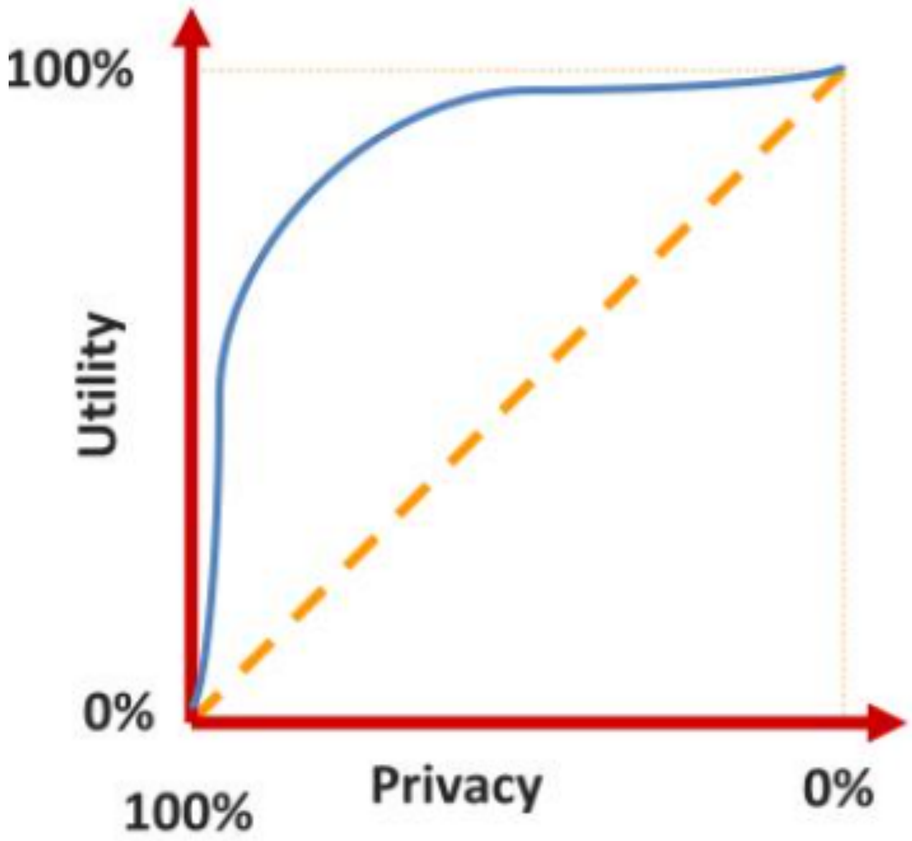
3. Phones periodically download these pairs (w, seed_w) and use these pairs to reconstruct the list of EphIDs for each time window.

How do people manage in practice?

- Risk/benefit analysis
- User perceptions
- Different stakeholders
- Control/Notice => next week

Contexts

System architecture: Privacy-utility tradeoff



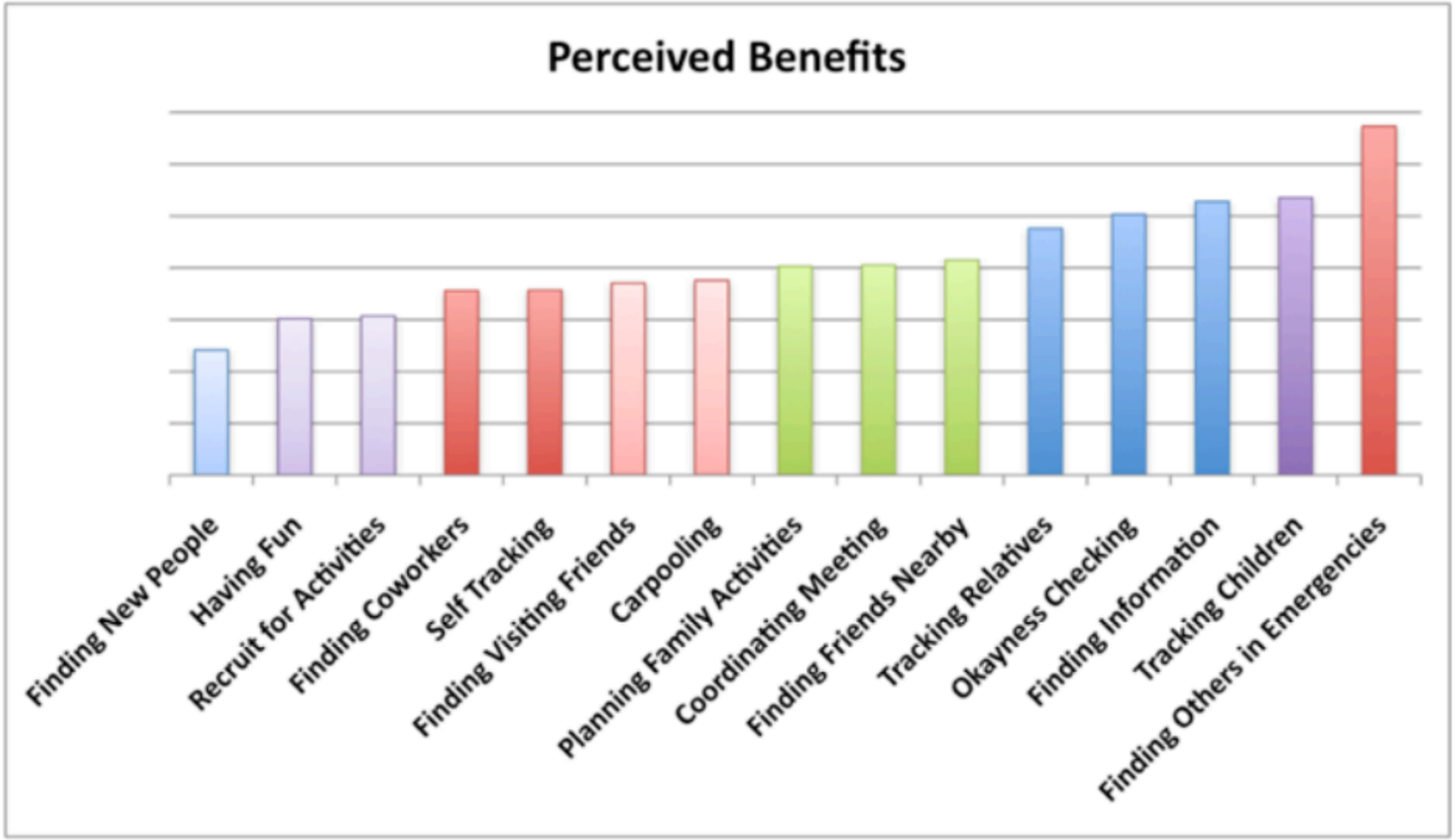
Utility for who?

Individual?

Local community?

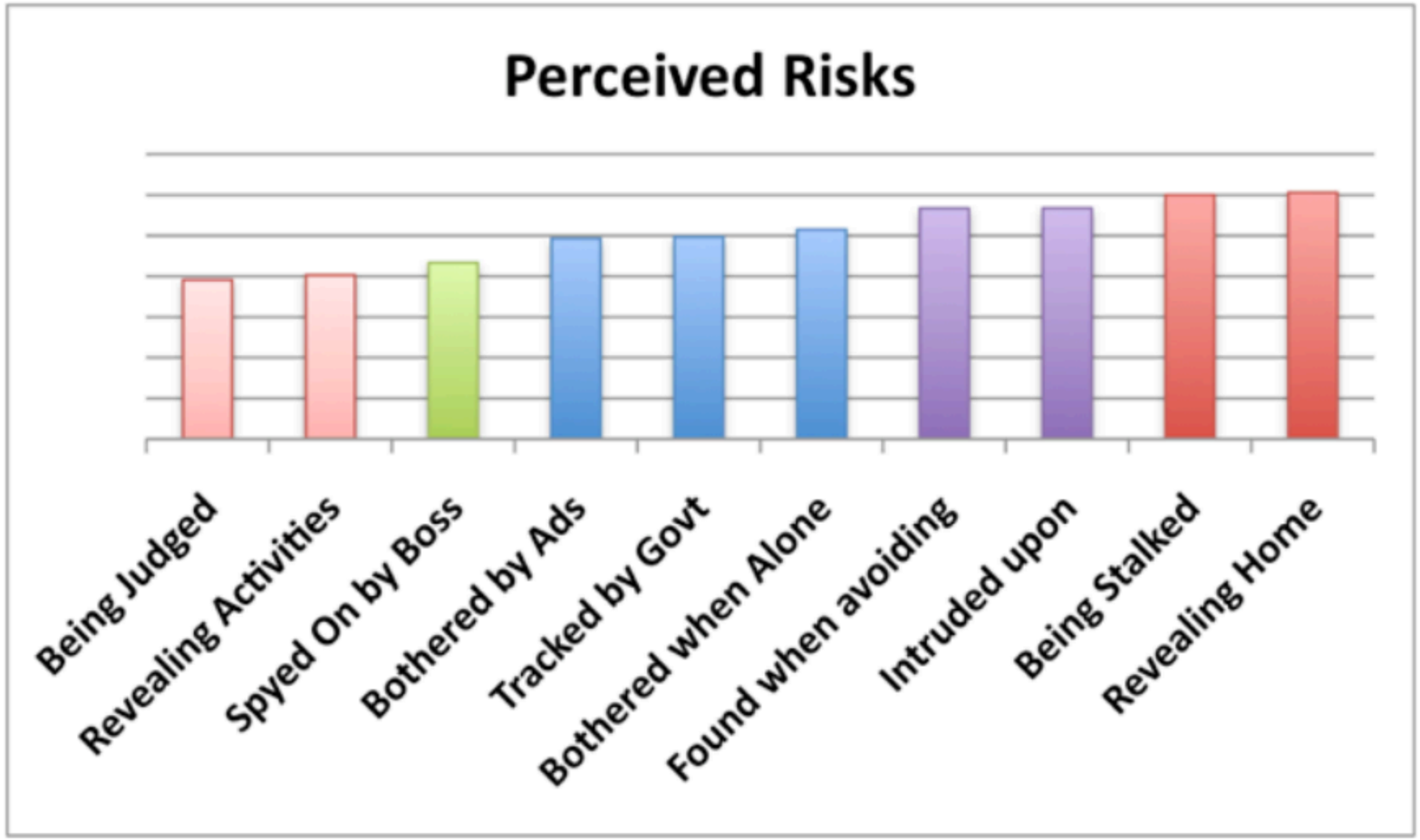
How do people manage in practice?

Benefit Scenarios

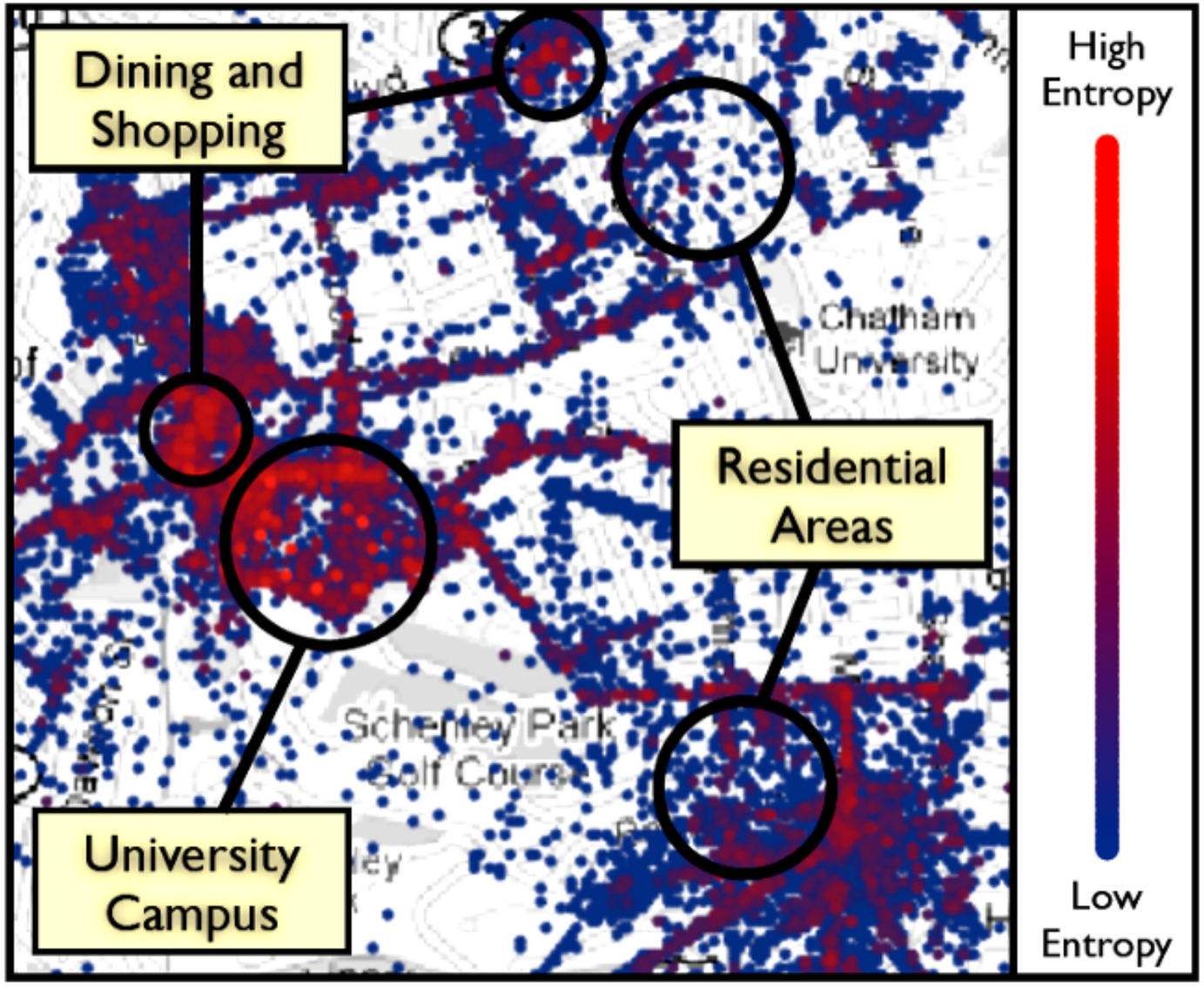


How do people manage in practice?

Risk Scenarios



Information entropy

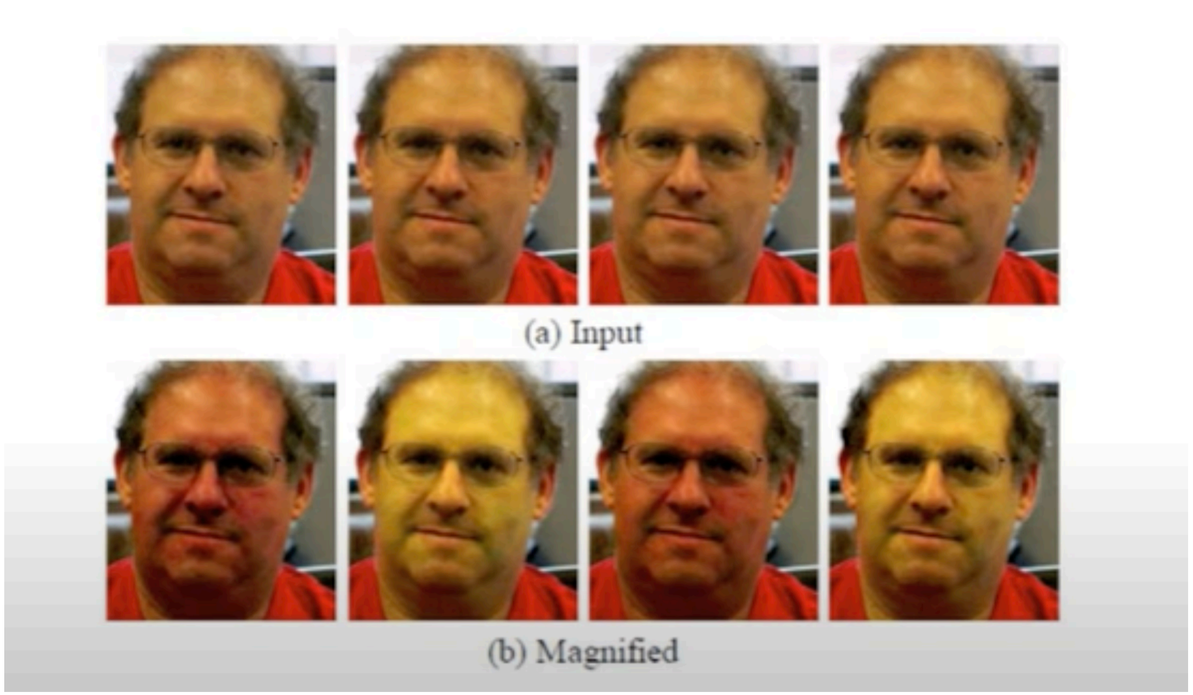


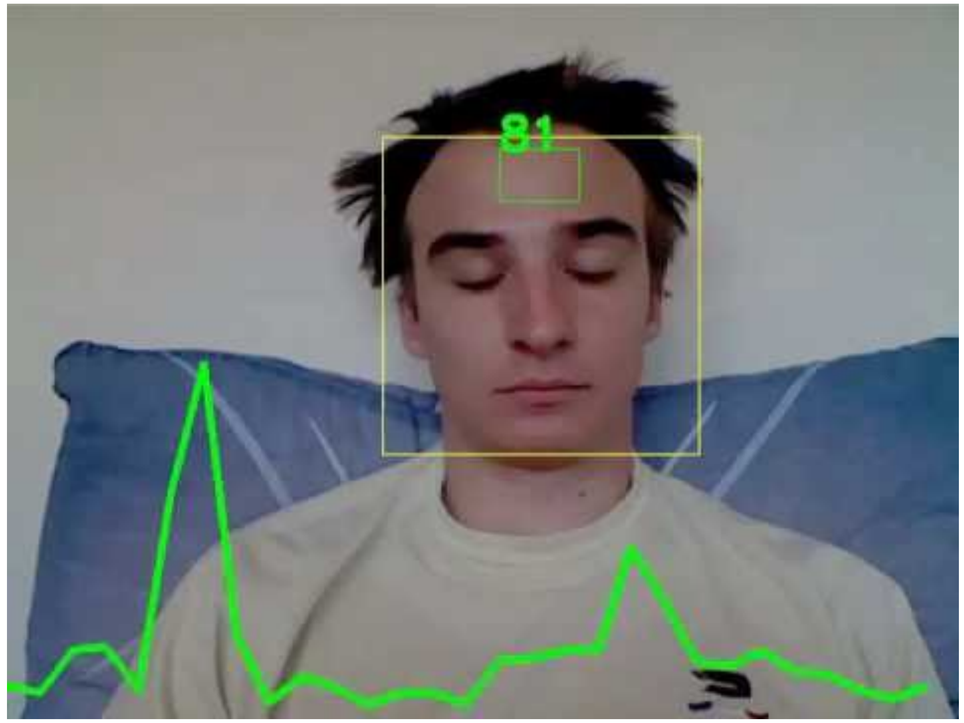
Place naming & Context

- Hey mom, I am at 55.66N 12.59E.
- Hey mom, I am at home.
- Hey mom, I am at UTC.

- Which one is more privacy-sensitive?

Why not face privacy?





http://people.csail.mit.edu/mrub/evm/

Fair Information Practices

- Notice / Awareness
- Choice / Consent
- Access / Participation
- Integrity / Security
- Enforcement / Redress

A CONSUMER INTERNET PRIVACY

BILL of RIGHTS

The Obama Administration believes America must apply our timeless privacy values to the new technologies and circumstances of our times. Citizens are entitled to have their personal data handled according to these principles.

Individual Control
Consumers have a right to exercise control over what personal data companies collect from them and how they use it.

Access and Accuracy
Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity and risk associated with the data.

Transparency
Consumers have a right to easily understandable and accessible information about privacy and security practices.

Focused Collection
Consumers have a right to reasonable limits on the personal data that companies collect and retain.

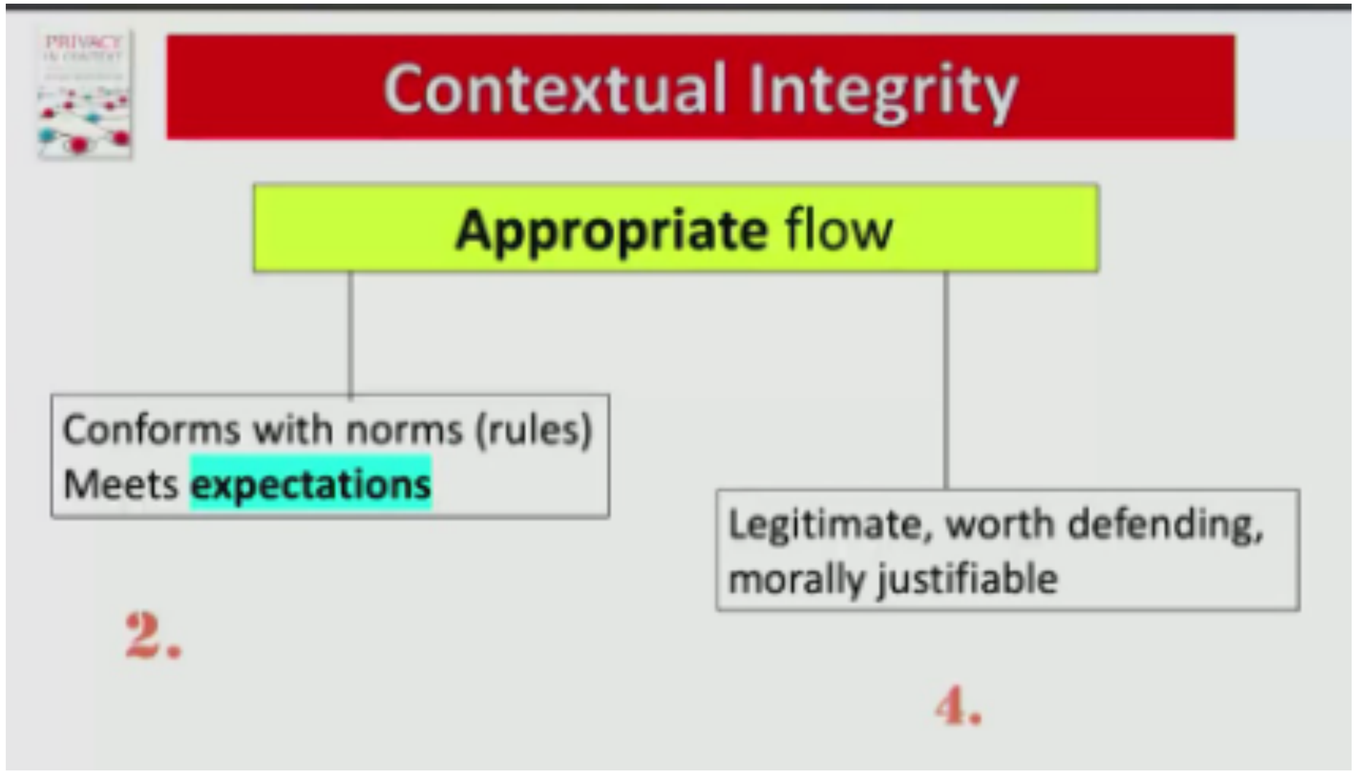
Respect for Context
Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent.

Accountability
Companies should be accountable to enforcement authorities and consumers for adhering to these principles.

Security
Consumers have a right to secure and responsible handling of personal data.

LEARN MORE AT WHITEHOUSE.GOV

Contextual integrity



Purposes

Approximate information flow

THE MARKET FOR "LEMONS":
QUALITY UNCERTAINTY AND THE
MARKET MECHANISM *

GEORGE A. AKERLOF

I. INTRODUCTION

This paper relates quality and uncertainty. The existence of goods of many grades poses interesting and important problems for the theory of markets. On the one hand, the interaction of quality

The Market for Lemons, Nobel Prize, 2001

Schedule - Context of Location Privacy

- Why location privacy?
- Location-based applications
- Locating technologies
- Protecting location privacy
- Beyond location privacy
- Purpose framework
  - Data collection, data processing, data usages

Credits

- Duke. CompSci 590.03 https://www.slideserve.com/glenys/location-privacy-powerpoint-ppt-presentation
- Location-Sharing Technologies: Privacy Risks and Controls
- An overview of location privacy for mobile computing: https://slideplayer.com/slide/4915254/