Thesis Proposal

Chris Riederer
Columbia Computer Science
April 20th, 2017

In 2016,
$12 billion was spent on Location Based Advertising

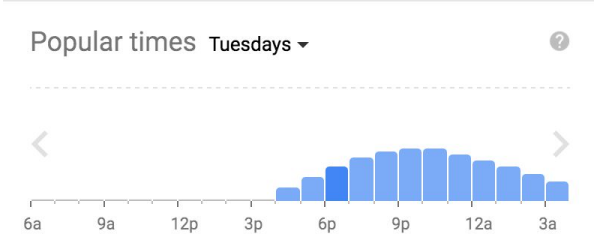
(38% of all smartphone ads)

http://www.biakelsey.com/location-targeted-mobile-ad-spend-reach-32-billion-2021/

Location Data: Useful

Why are firms hungry for location data?

- New products
- Better understanding of users
- Higher revenues!



Location Data: Highly Sensitive

Justice Sotomayor:

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

Goals

In my Ph.D., **I am making the use of location data safer for both users while retaining service provider utility** , by answering many questions:

- When can location data be anonymizeable?
- Are there attacks on location data outside of anonymization?
- How can we make the risks of location data use understandable to users?
- What tools can we build for both users and services to make location data more private?
- Can we prevent algorithmic bias during the use of location data?
- What are ways to grant more user control to location data?
- What beliefs do users have about the collection of their location data?

Talk Outline

1. Anonymity
2. Privacy and Economics
3. Demographics and Bias
4. Proposal for Next Steps

1. Anonymity

k-Anonymity

[Sweeney ‘02]

- A quasi-identifier (QI) is a set of columns in a database.
- A database is k-anonymous for QI if for any record, there are k-1 other records with identical QI.

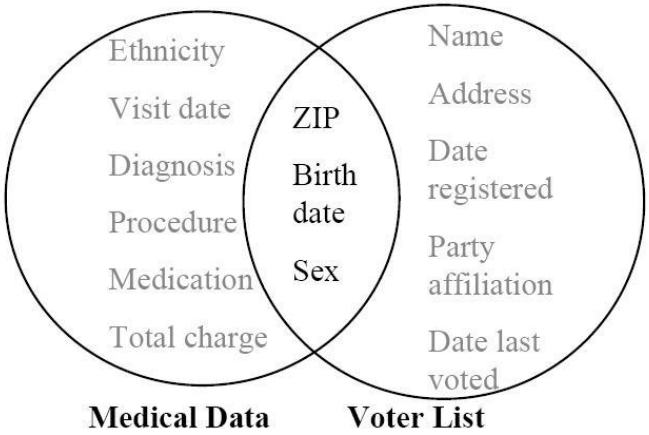
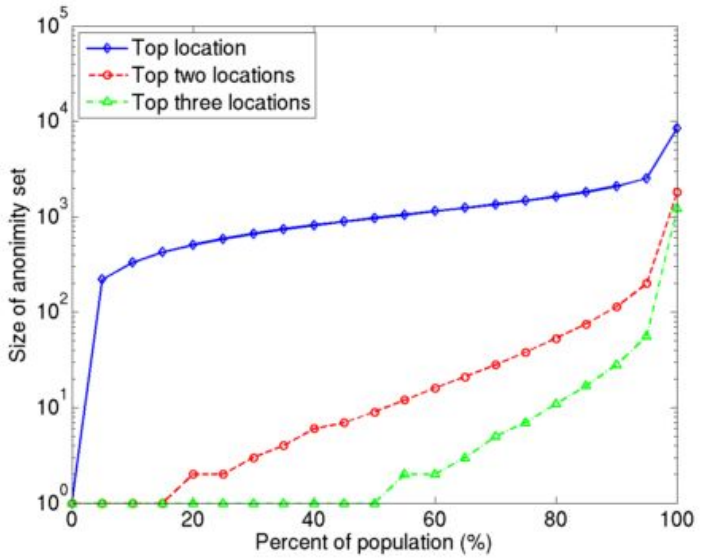


Figure 1 Linking to re-identify data

k-Anonymity of Location Data

[Zang & Bolot ‘11]

- Data: CDR, 25M users, USA
- QI for each user: top *n* most visited locations
- What is their size of anonymity set *k*?
- For 3 points (i.e. n = 3),
  - median of *k*=1 at cell granularity
  - median of *k*=2 at ZIP code level

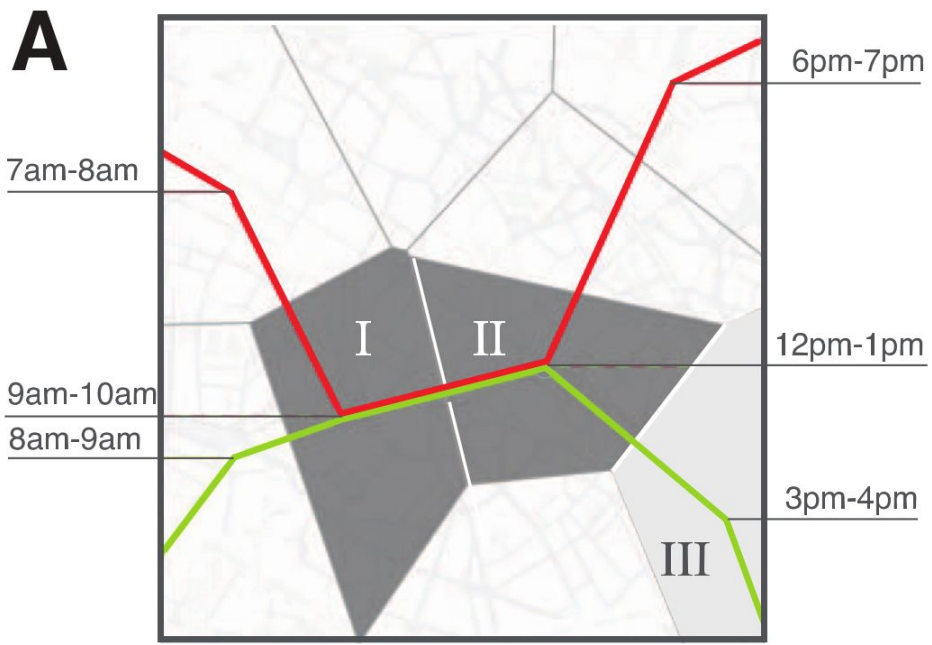


(b) Cell

k-Anonymity of Location Data

[de Montjoye et al ‘13]

- Data: CDR, 1.5M users, European country
- Instead of top-*n* locations, use *m* random points and **include time.**
- QI is now *m* spatiotemporal points
- For *m*=4: 95% of users are 1-anonymous

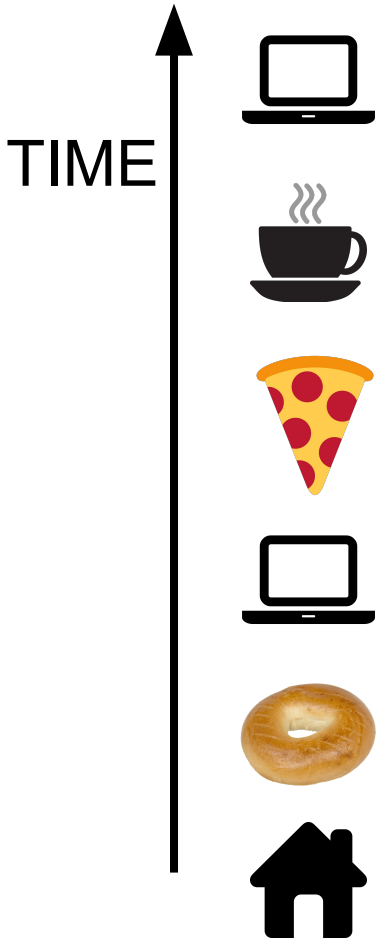


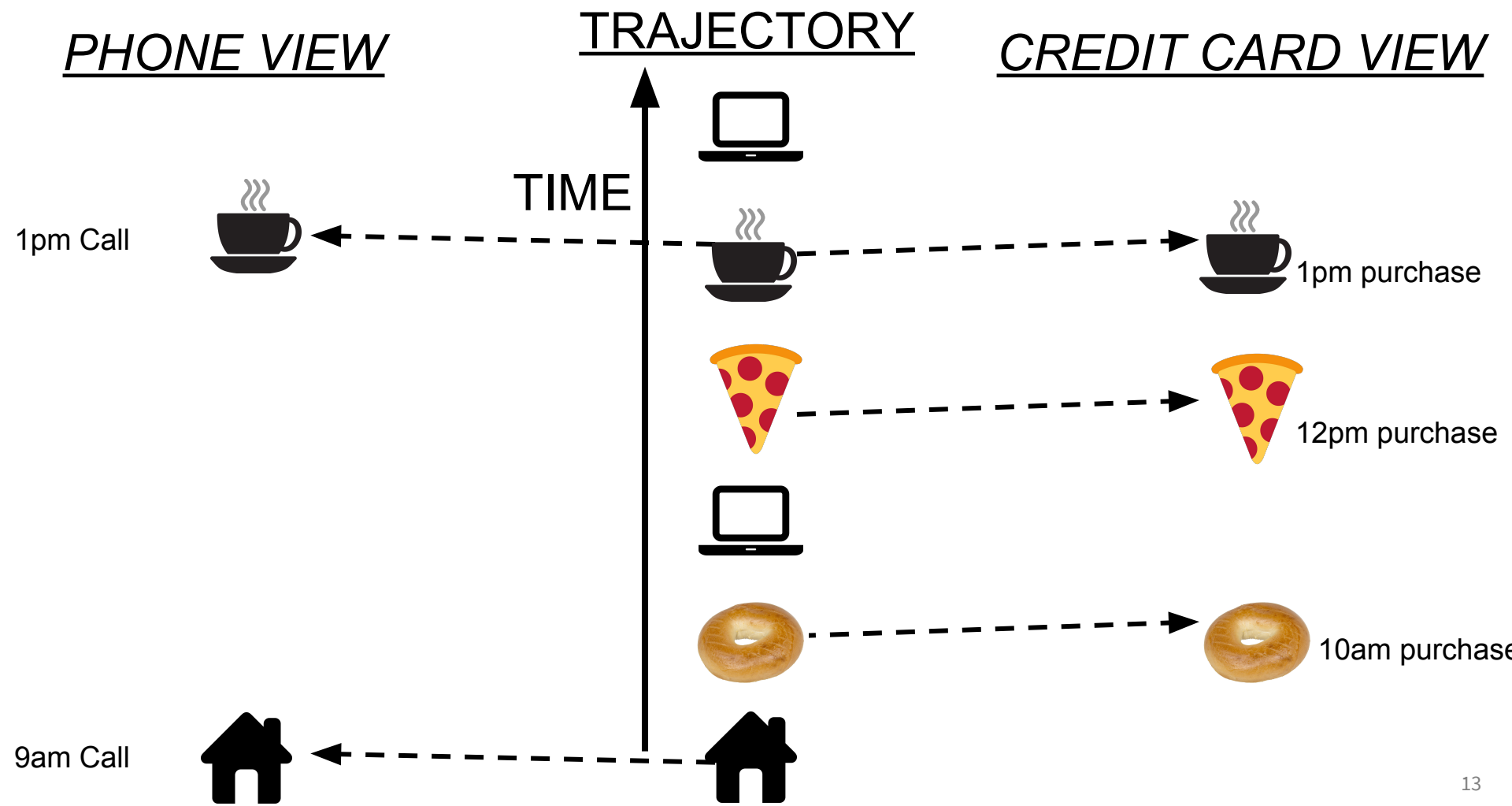
Limits of k-anonymity attack

These attacks only work when the attacker has auxiliary data that intersects the original data.

Going beyond the basic k-anonymity attack: is user mobility unique enough to be identified when sampled or in different ways or from different time periods?

TRAJECTORY




PHONE VIEW
TRAJECTORY
CREDIT CARD VIEW
TIME
1pm Call
1pm purchase
12pm purchase
10am purchase
9am Call

PHONE VIEW

1pm Call

TIME

9am Call

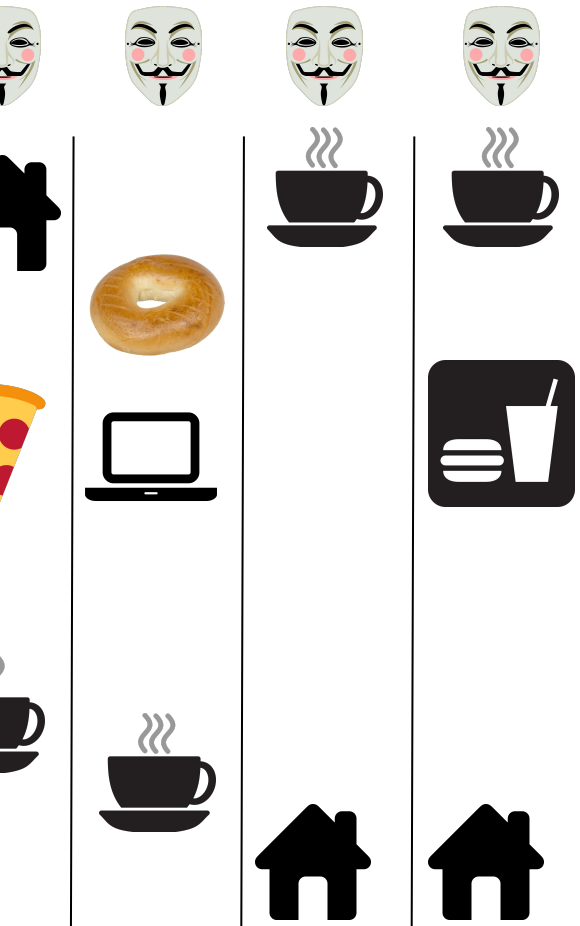
CREDIT CARD VIEW

1pm purchase

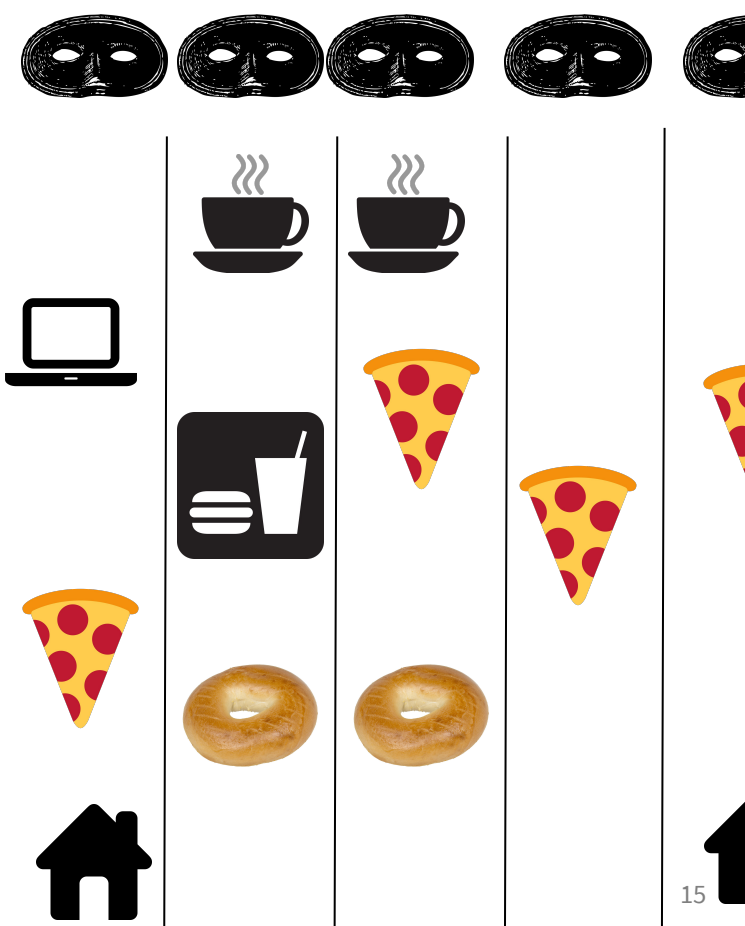
12pm purchase

10am purchase

PHONE VIEW



CREDIT CARD VIEW



Our Approach: Maximum Likelihood

There are many possible permutations!

Theorem: It is possible to efficiently find the most likely σ.

How? Using score function ϕ as edge weights, matching weight is the log likelihood σ.

$$\phi = \frac{e^{\lambda} \sum_{k \geq \max\{a_1, a_2\}} \frac{\lambda^k}{k!} \binom{k}{a_1} (1-p_1)^{k-a_1} \binom{k}{a_2} (1-p_2)^{k-a_2}}{\sum_{k \geq a_1} \frac{\lambda^k}{k!} \binom{k}{a_1} (1-p_1)^{k-a_1} \cdot \sum_{k \geq a_2} \frac{\lambda^k}{k!} \binom{k}{a_2} (1-p_2)^{k-a_2}}$$

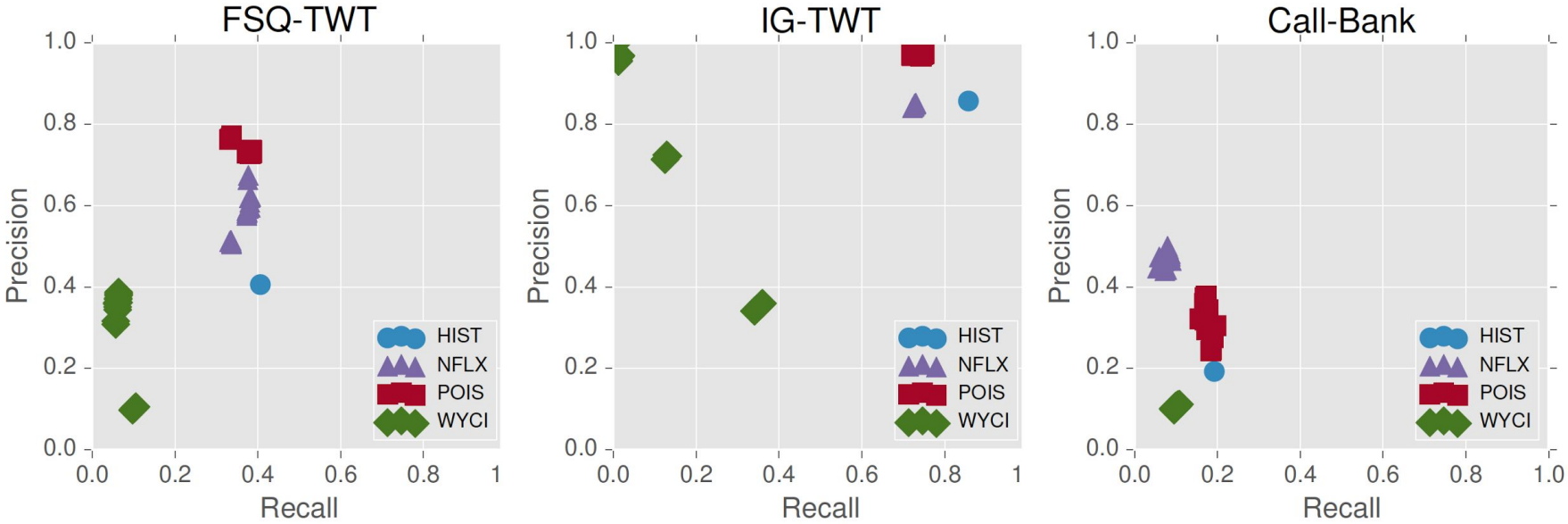


Figure 2.3: Precision and Recall plots for each dataset.

2. Privacy + Economics

Brief Overview of “Ad Economy”

“If you are not paying for it, you're not the customer; you're the product being sold.”

-- Andrew Lewis (and probably others)

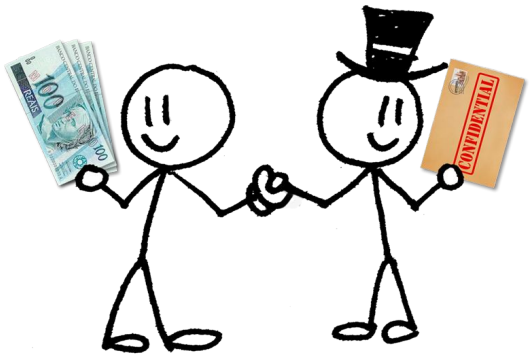
Goals:

- Informed consent
- User control

Transactional Privacy

If system designers can obtain user anonymity, there is a good way to reveal user information in a way that can be accurately priced to benefit users.

We proposed a system [Riederer HotNets ‘11] in which anonymous users auction off the right to be shown ads. Through the use of a differentially private auction (from [McSherry FOCS ‘07]) users obtain a fair price from aggregators.



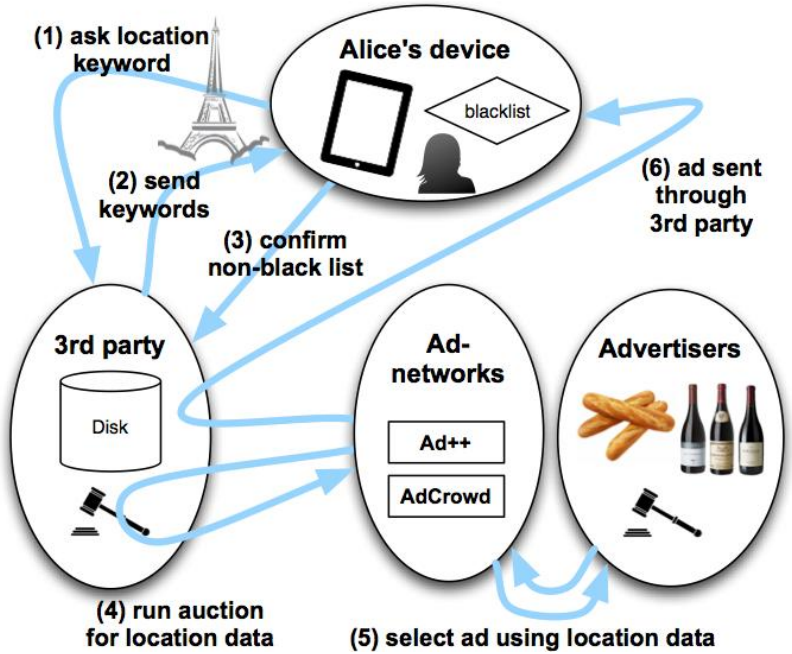
Txn'l Privacy for Location Data

Locations represented as keywords

Users control keywords

Easier for advertisers and users to understand and control.

[Riederer WPES '13]



Papers and follow-ups

Transactional privacy:

[Riederer HotNets ‘11]

[Riederer WPES ‘13]

Understanding user beliefs:

[Carrascal WWW ‘13]

Citing papers w/ mobile angle:

[Yang MobiCom ‘12]

[Xia SIGCOMM ‘13]

[Staiano UbiComp ‘14]

[Hosio CHI ‘ 16]

…

3. Demographics and Bias

Beyond anonymity...

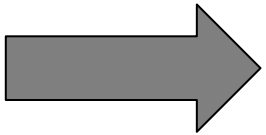
"AI can exacerbate problems of bias."

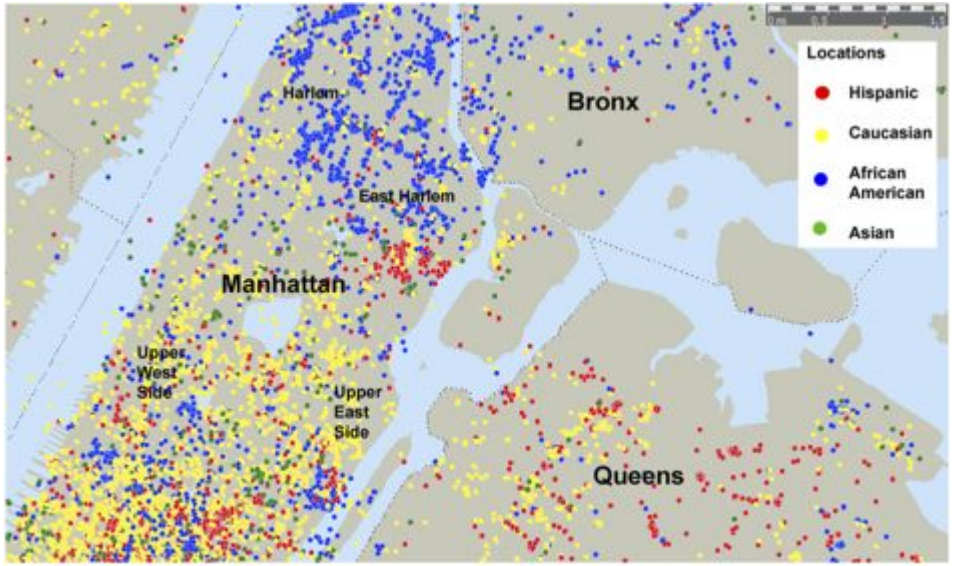
-- Obama White House Report on Preparing for the Future of Artificial Intelligence

Are there still concerns even when users are anonymous?

- Can we predict race?
- Can we study demographics and location data at scale?
- Can we explain the risks to the general public?

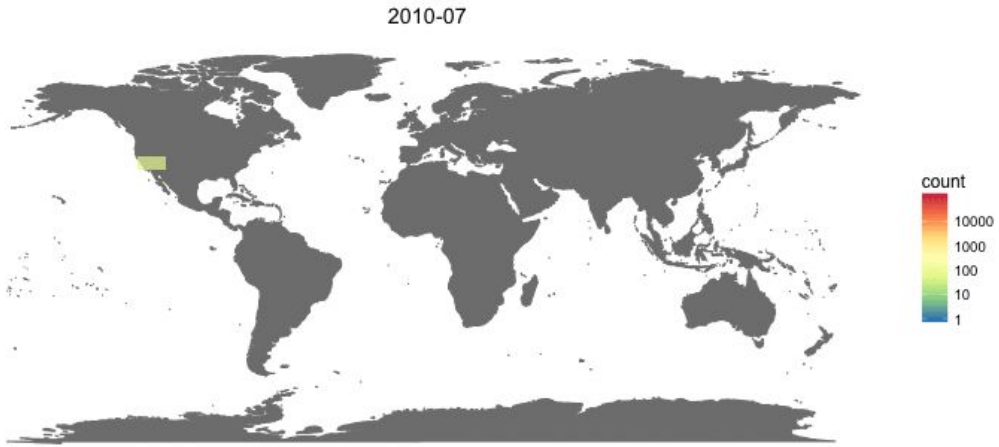
Geotagged Photos to Demographics



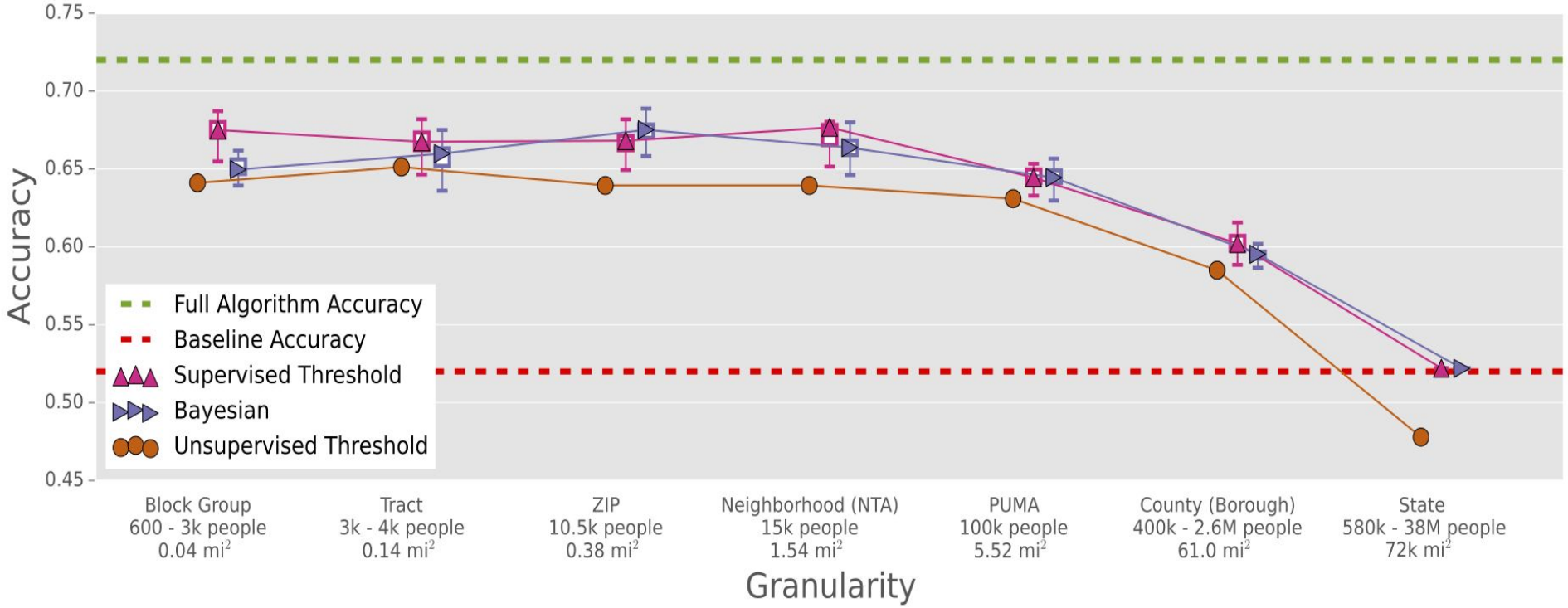


Instagram Data

- 115M Instagram photo’s metadata
  - Name
  - Date
  - Location
  - Comments
  - Likes
  - Tags
  - And more...
- 260k users
- 16M geotags
- 180 countries
- Small subset labeled with Mechanical Turk



Inferring Demographics



Scaling up the Census

Mechanical Turk is accurate but expensive. Scale up labeling by using face recognition.

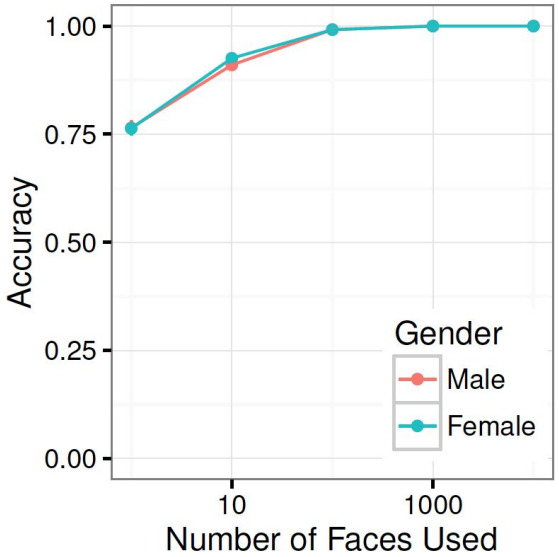
Response JSON

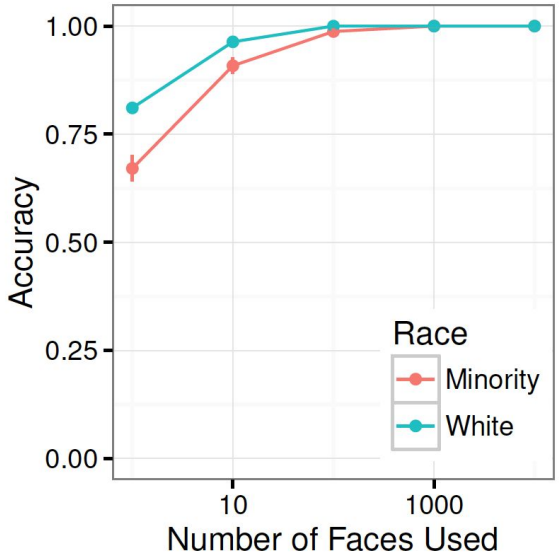

```
      "x": 181
    },
    "right_eye_upper_right_quarter": {
      "y": 147,
      "x": 247
    },
    "mouth_upper_lip_bottom": {
      "y": 234,
      "x": 218
    }
  },
  "attributes": {
    "gender": {
      "value": "Female"
    },
    "age": {
      "value": 51
    },
    "eyestatus": {
      "left_eye_status": {
        "normal_glass_eye_open": 0.187,
        "no_glass_eye_close": 0,
```

Results

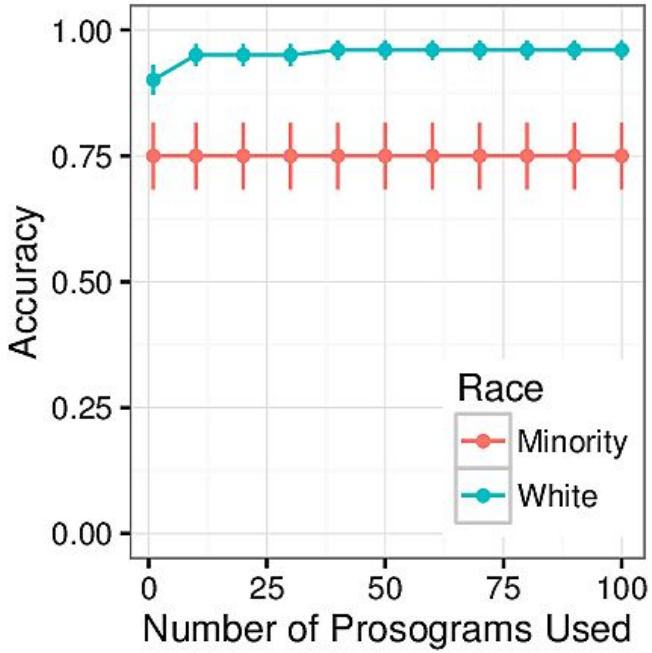
87.5% accuracy for gender

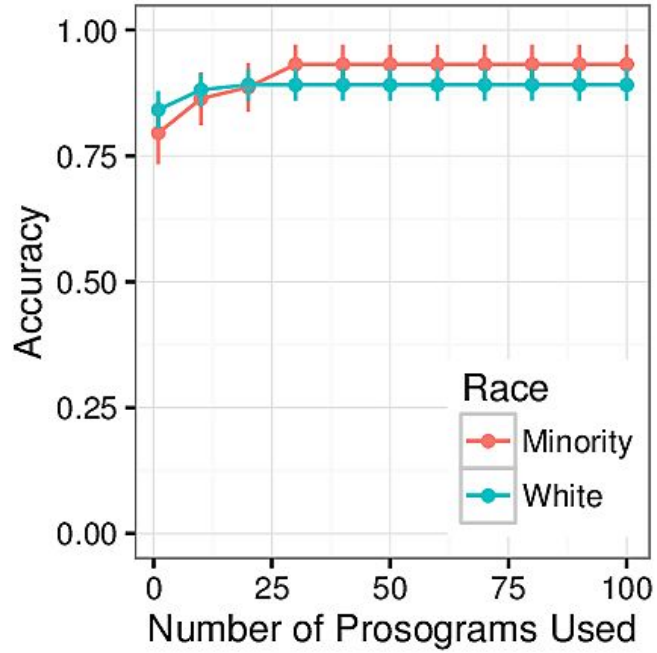
89.7% accuracy for race





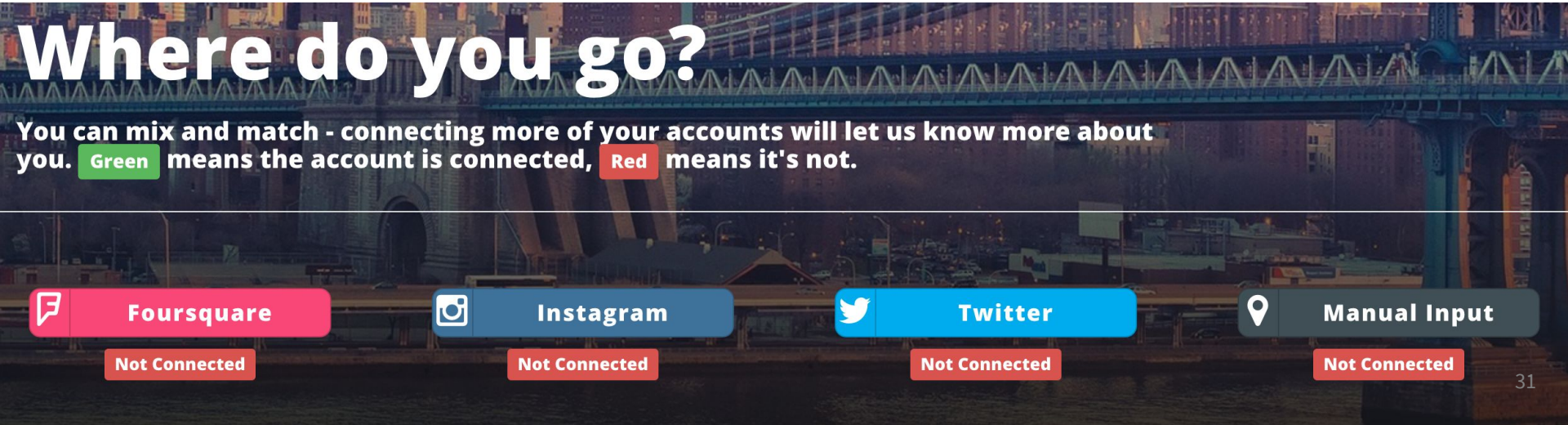
Balanced Error Rate





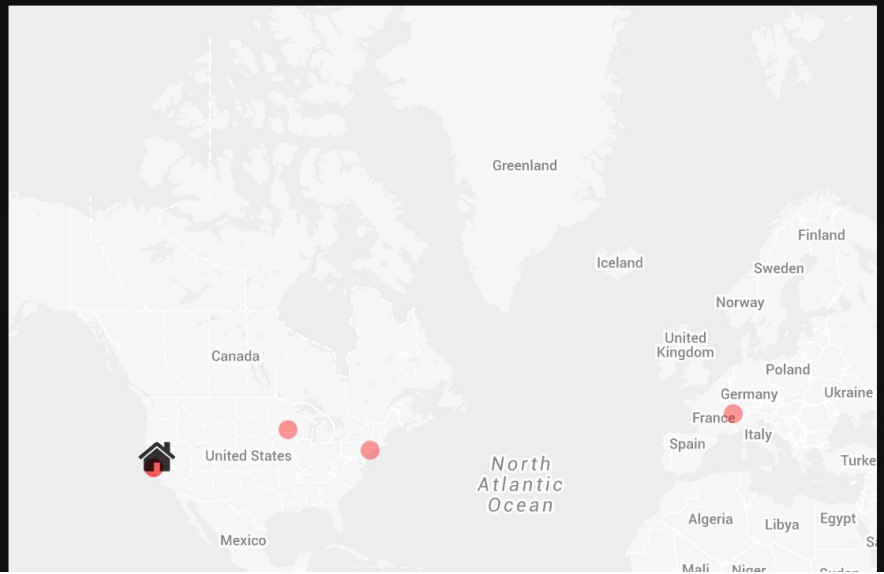
Tools to Inform Users

Learning ways of labeling demographics of users allows us to communicate risk to users. Built online tool, called “FindYou”, which we demoed at WWW in 2016.



Hi, Chris

Scroll down to find out about what we think we know about you



You have **5** geotagged pictures at **5** unique locations in the U.S. You most actively geotag in the **night**. You average 1 geotagged picture every **35.4** days.

Hi, Chris

Scroll down to find out about what we think we know about you

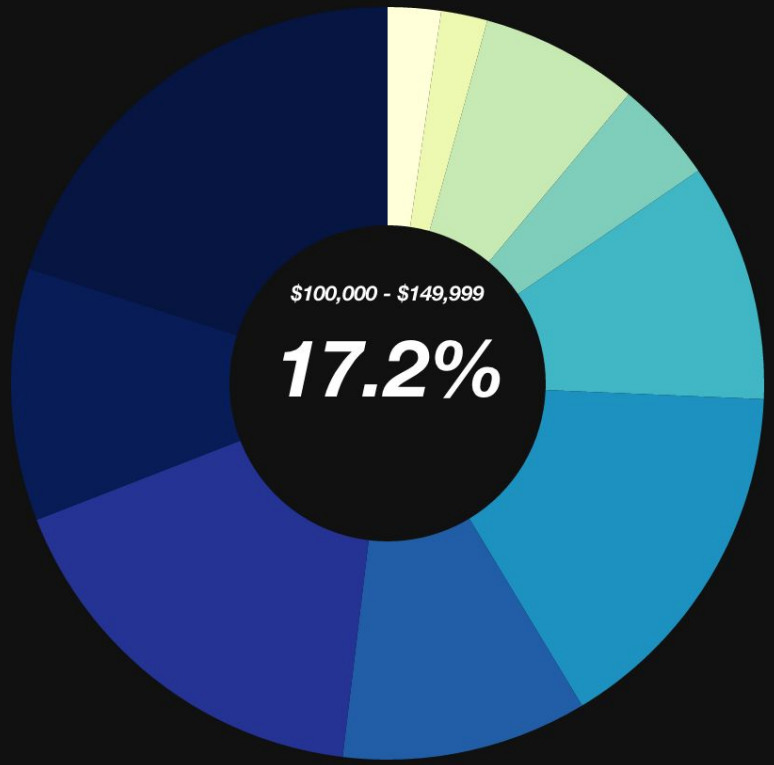
You have **5** geotagged pictures at **5** unique locations in the U.S. You most actively geotag in the **night**. You average 1 geotagged picture every **35.4** days.

Race

We predict you are: White

Are we correct? Yes | No

Average Income Distribution Of All Tracts You Have Visited



Under $10,000: *2.3%* | $10,000 - $14,999: *2.0%* | $15,000 - $24,999: *6.8%*
$25,000 - $34,999: *4.4%* | $35,000 - $49,999: *10.2%* | $50,000 - $74,999: *15.7%*
$75,000 - $99,999: *10.5%* | $100,000 - $149,999: *17.2%* | $150,000 - $199,999: *10.8%*
Over $200,000: *20.1%*

Sc

t you

gged pictures at 5
U.S. You most
night. You average
very 35.4 days.

We

Are w

hite

Average Income Distribution Of All Tracts You Have Visited

We predicted this primarily because:

We predict your home is in **Census Tract 81** in **New York County**, **New York** because this is the tract in which you have the most geotagged locations.

You have **4** geotagged locations here.

This comprises **25.0%** of your total **16** geotagged locations.

In all the maps, it is indicated with the 🏠 icon.

Total population & gender split:

The total population of your home tract is **8,047**.

Of the population over 18, there are **3,849** men and **3,494** women. This means that women comprise **47.6%** of the population over 18 whereas men comprise **52.4%** of the population over 18.

Renters & owners, household size, family size:

Of the **4,992** housing units in your census tract, **3,550** are rented and **1,442** are owned. This means this tract is made up of **71.1%** renters and **28.9%** owners.

The average number of individuals living in a single household in your home tract is **1.58**.

The average family's size in your home tract is **2.64**.

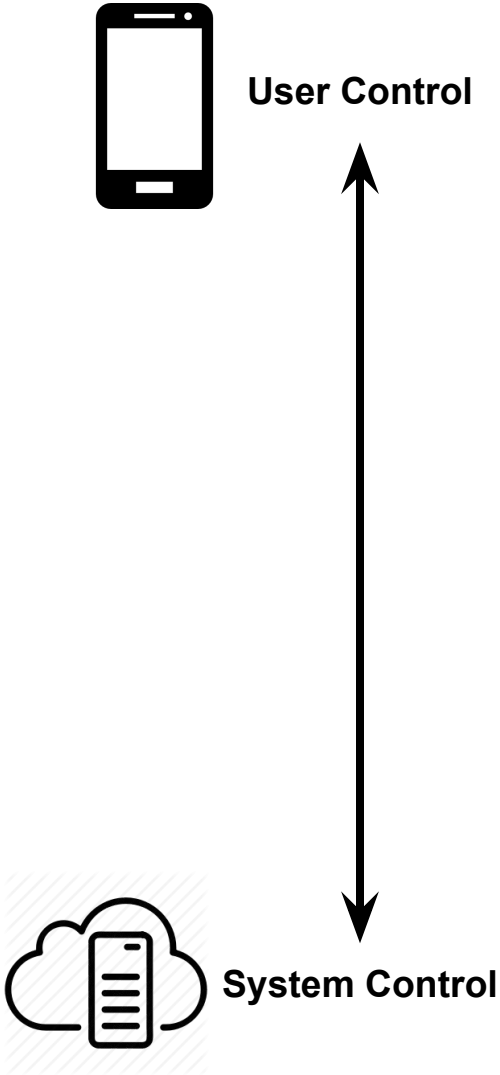
Why ^

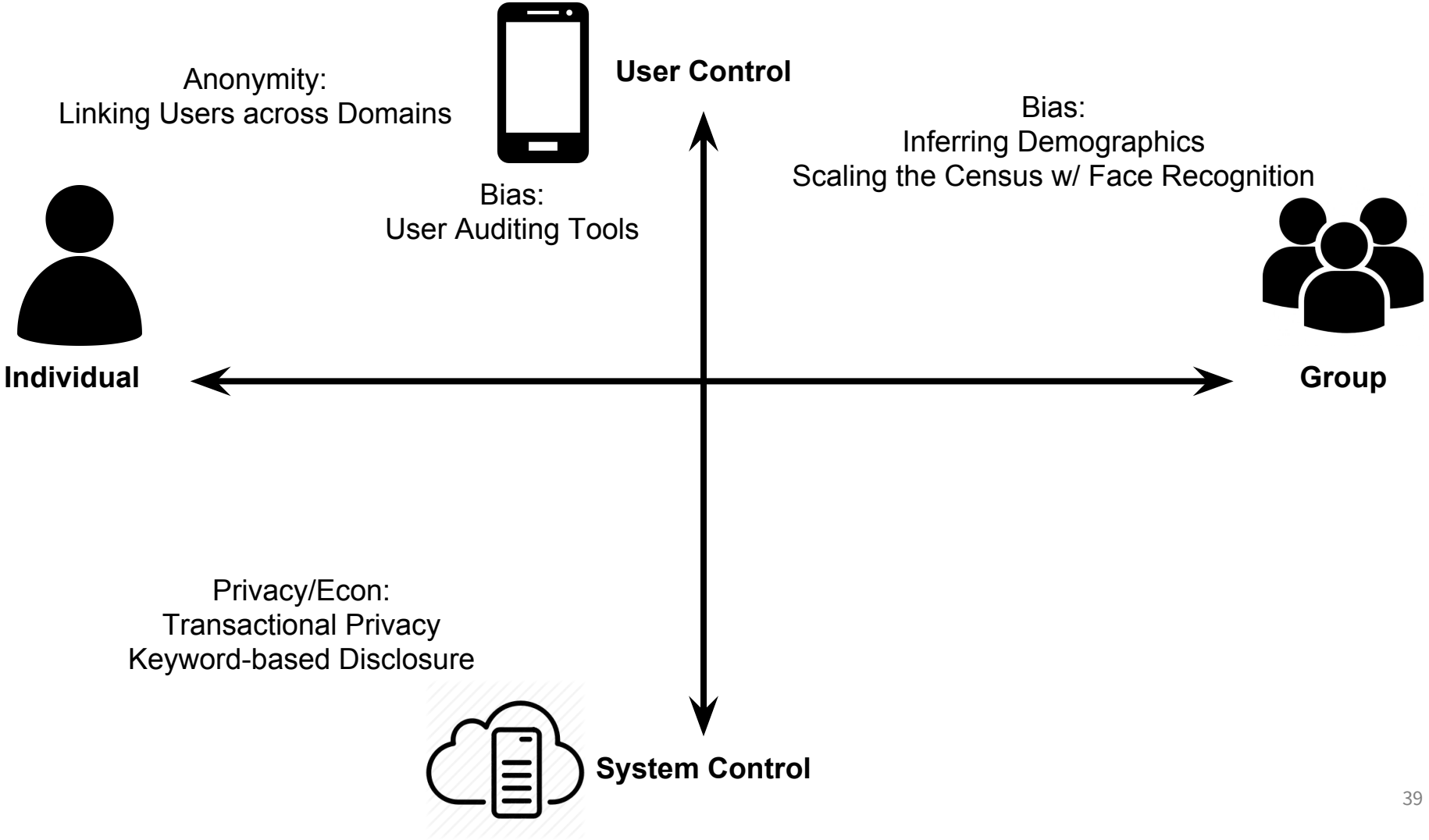
Are w

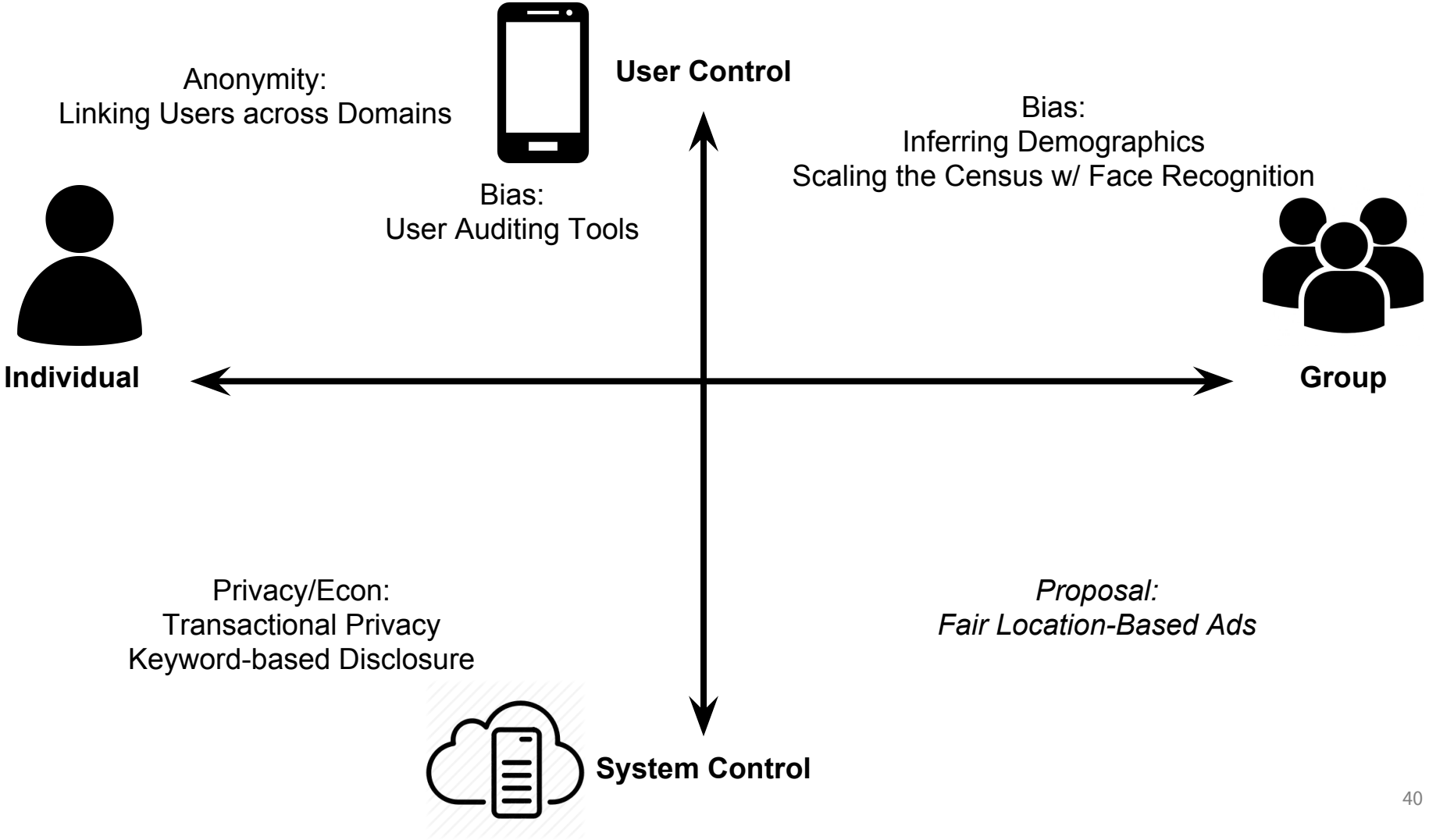
$25,000 - $34,999: *4.4%* $35,000 - $49,999: *10.2%* $50,000 - $74,999: *15.7%*
$75,000 - $99,999: *10.5%* $100,000 - $149,999: *17.2%* $150,000 - $199,999: *10.8%*
Over $200,000: *20.1%*

4. Proposal


Individual
Group


User Control
System Control


Anonymity:
Linking Users across Domains
User Control
Bias:
Inferring Demographics
Scaling the Census w/ Face Recognition
Bias:
User Auditing Tools
Individual
Group
Privacy/Econ:
Transactional Privacy
Keyword-based Disclosure
System Control


Anonymity:
Linking Users across Domains
User Control
Bias:
Inferring Demographics
Scaling the Census w/ Face Recognition
Bias:
User Auditing Tools
Individual
Group
Privacy/Econ:
Transactional Privacy
Keyword-based Disclosure
Proposal:
Fair Location-Based Ads
System Control

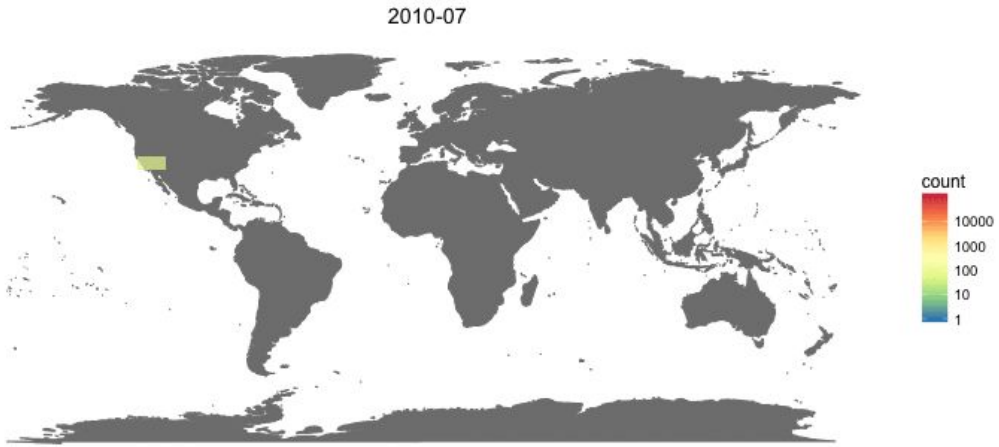
Fair Location-Based Advertising

Plan

1. Simulate location-based advertising scenario
2. Learn where to show ads
3. Debias where or when to display ad
4. Investigate the loss of utility

Instagram Data

- 115M Instagram photo's metadata
  - Name
  - Date
  - Location
  - Comments
  - Likes
  - Tags
  - And more...
- 260k users
- 16M geotags
- 180 countries
- **2M photos, 844k faces for gender, race, "smiling"**



Advertising Scenarios

1. **Context**: show ad at right time and place.

   Simulate based on “#tag”. Predict when user will use a tag based on location.

2. **Behavioral**: show ad based on previous locations.

   Simulate based on travel. Look at locations of users who have traveled a lot. Make prediction for all users. Send offline ad (email?) to users.

“Fairness Through Awareness”

Solve this LP. Outcomes should be similar for similar users, while minimizing loss.

x, y are individuals. **a** is an ad. $\boldsymbol{\mu}_x$ is x’s distribution over possible ads.

d is a distance function for users, **D** is a distance function for distributions

$$\begin{aligned} \min \quad & \mathbf{E}_{x\sim V}\mathbf{E}_{a\sim\mu_x} L(x,a) \\ \text{subject to} \quad & \forall x,y \in V : D(\mu_x,\mu_y) \leq d(x,y) \\ & \forall x \in V : \mu_x \in \Delta(A) \end{aligned}$$

“Fairness Through Awareness”

When using a particular **d**, the Earthmover Distance, then statistical bias between two different groups will be equal to the Earthmover Distance of the distribution of users in expectation.

In other words, we can understand bias between two groups by using a particular distance function.

Theorem 3.3. *Let d be a metric. Then,*

$$\text{bias}_{D_{\text{tv}},d}(S,T) \leq d_{\text{EM}}(S,T)\,.$$

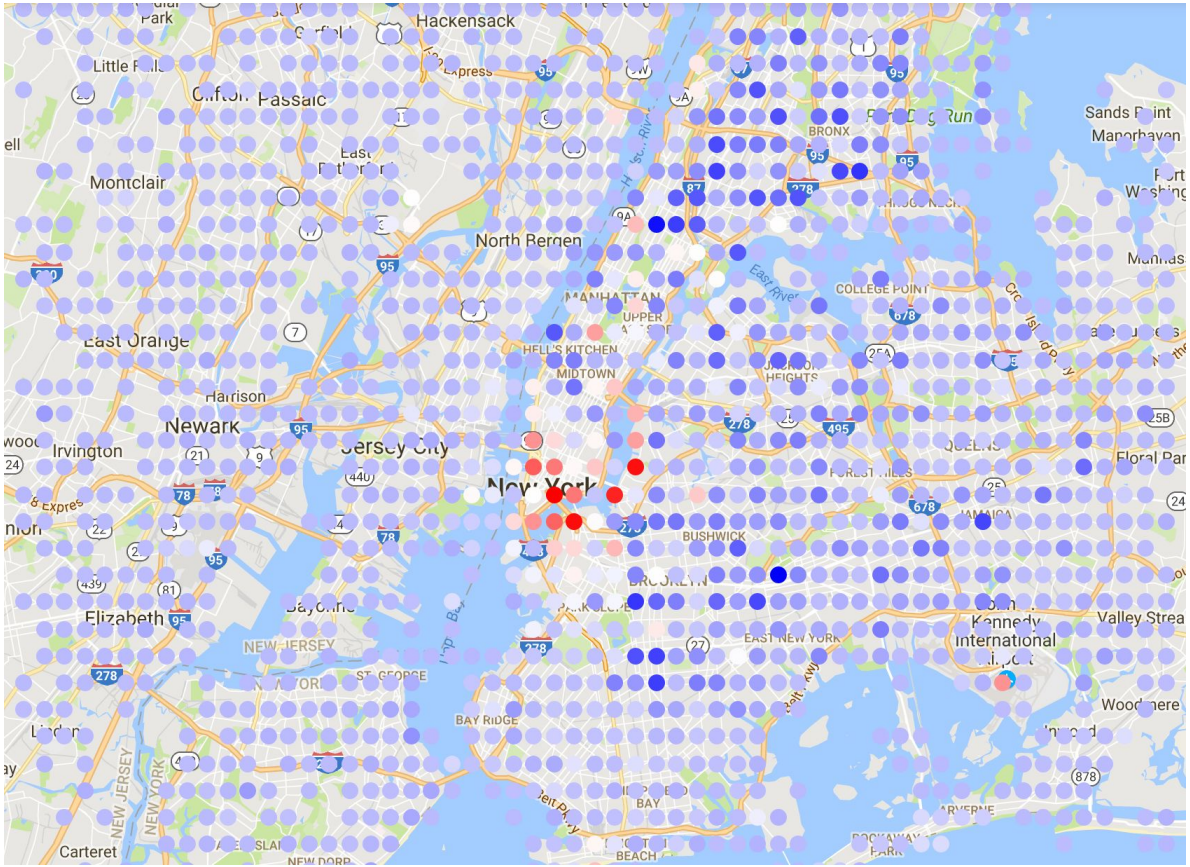
If furthermore $d(x,y) \leq 1$ for all x, y, then we have

$$\text{bias}_{D_{\text{tv}},d}(S,T) \geq d_{\text{EM}}(S,T)\,.$$

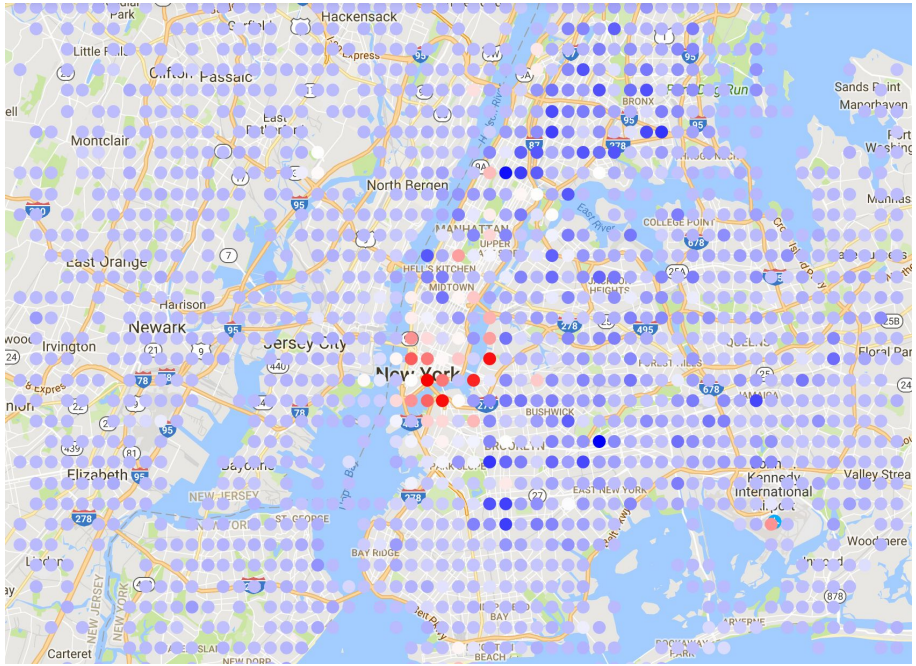
Associated Tags

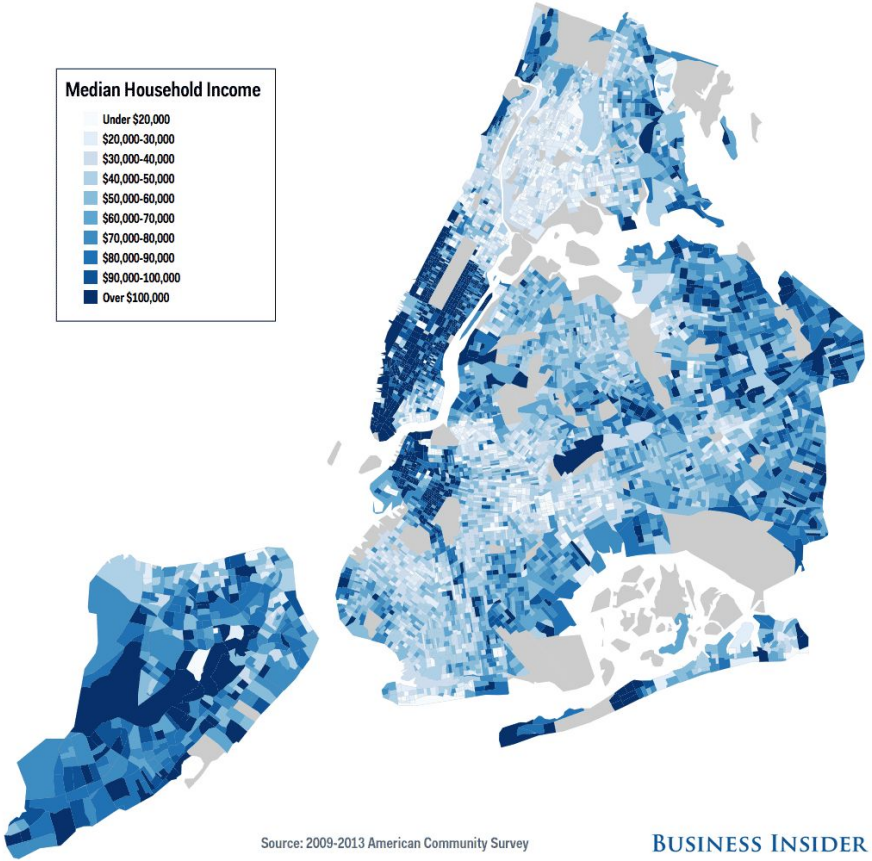
Using χ^2 test...

Female	*Male*	*Asian*	*Black*	*White*
makeup	beard	indonesia	wcw	summer
nails	mensfashion	japan	tbt	winter
dress	jordan	singapore	truth	sun
lipstick	hiphop	korea	hiphop	spring
sisters	gay	thailand	dope	autumn



Color: ~likelihood of traveling internationally.


Hackensack
Passaic
Montclair
North Bergen
Manhattan
East Orange
Harrison
Newark
Irvington
Jersey City
New York
Elizabeth
Bayonne
Carteret
Bronx
Sands Point
Manhasset
College Point
Hell's Kitchen
Midtown
Upper East Side
Bushwick
East New York
Kennedy International
Valley Stream
Woodmere
Floral Park
Bay Ridge
Staten Island
New Jersey
Long Island


Median Household Income
Under $20,000
$20,000-30,000
$30,000-40,000
$40,000-50,000
$50,000-60,000
$60,000-70,000
$70,000-80,000
$80,000-90,000
$90,000-100,000
Over $100,000
Source: 2009-2013 American Community Survey
BUSINESS INSIDER

Timeline

Task	***Target Date***
Thesis Proposal	April 20th (Today!)
Create advertising model	Late April / Early May
Compute bias in ad outcome (correlation of demographics with “click” likelihood)	Early May
Apply debiasing and analyze results	Mid May
FATML Submission	May 26th
Work ready for publication	Early June
Defense	Mid June

Thank you!

Extra Slides

Balance

It will be necessary to find ways to obtain the benefits of big location data while mitigating the risks.

Targeting

Firms gather data in order to target ads.

Infer (potentially sensitive) demographics from location data.

Attacks on location data

Other attacks:

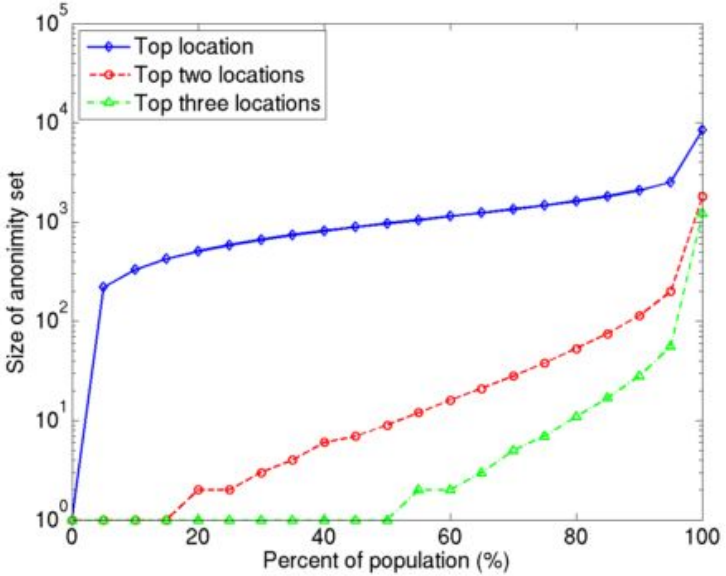
- [Narayan ‘08] “Netflix attack”
  - Uses sparsity to link databases to auxiliary data
  - Robust to some noise
  - Can be applied to discrete or continuous location data
- [Naini ‘14] Histogram Matching
  - Consider percent of time spent in locations
  - Discrete location representation

Attacks on location data

[Rossi ‘14]

- Partition a user’s LBSN checkins into a small set of 10 and a larger set, and attempt to link the two.
- Uses a hybrid solution that combines spatiotemporal similarity with discrete, frequency based similarity.
- Spatiotemporal
- Frequency

k-Anonymity of Location Data



(b) Cell

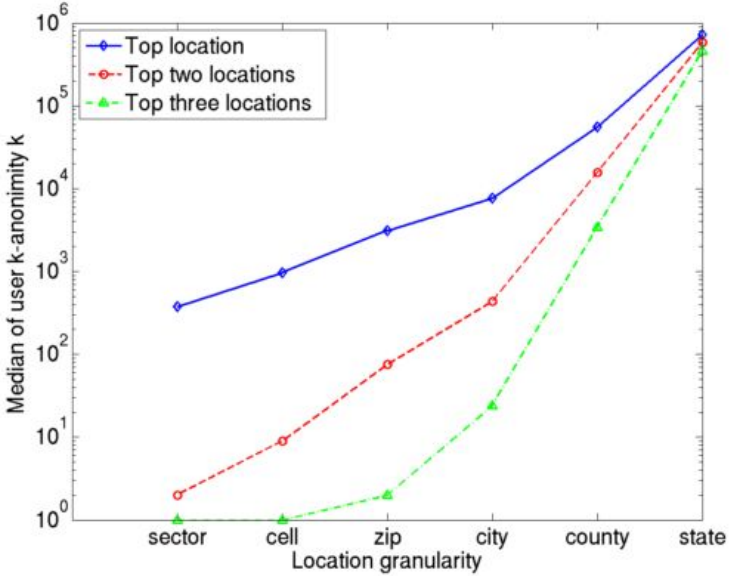
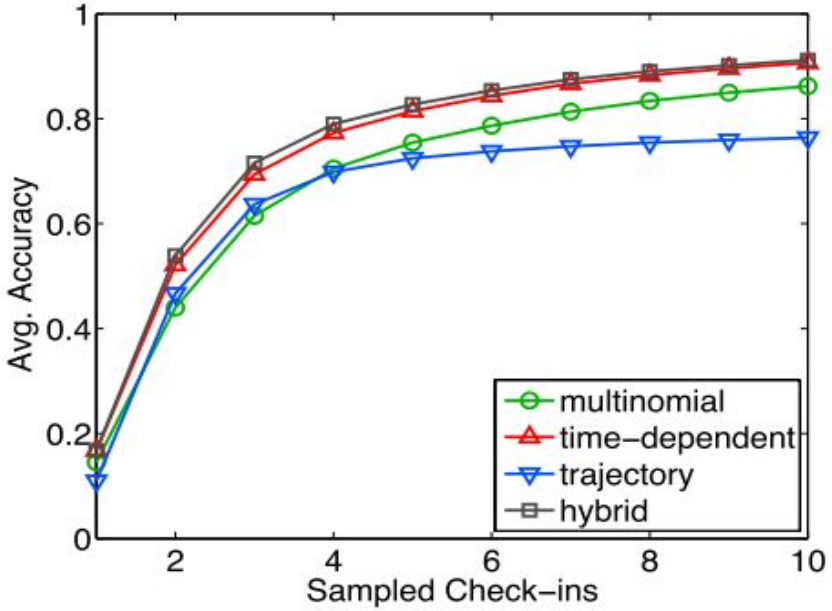


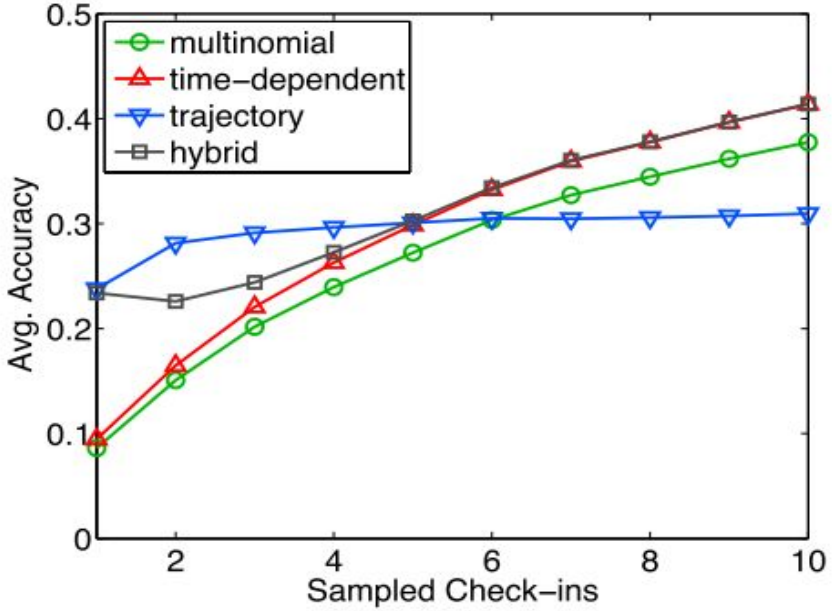
Figure 2: Median size of anonymity set at various granularity levels

Attacks on location data

[Rossi ‘14]



(b) Gowalla



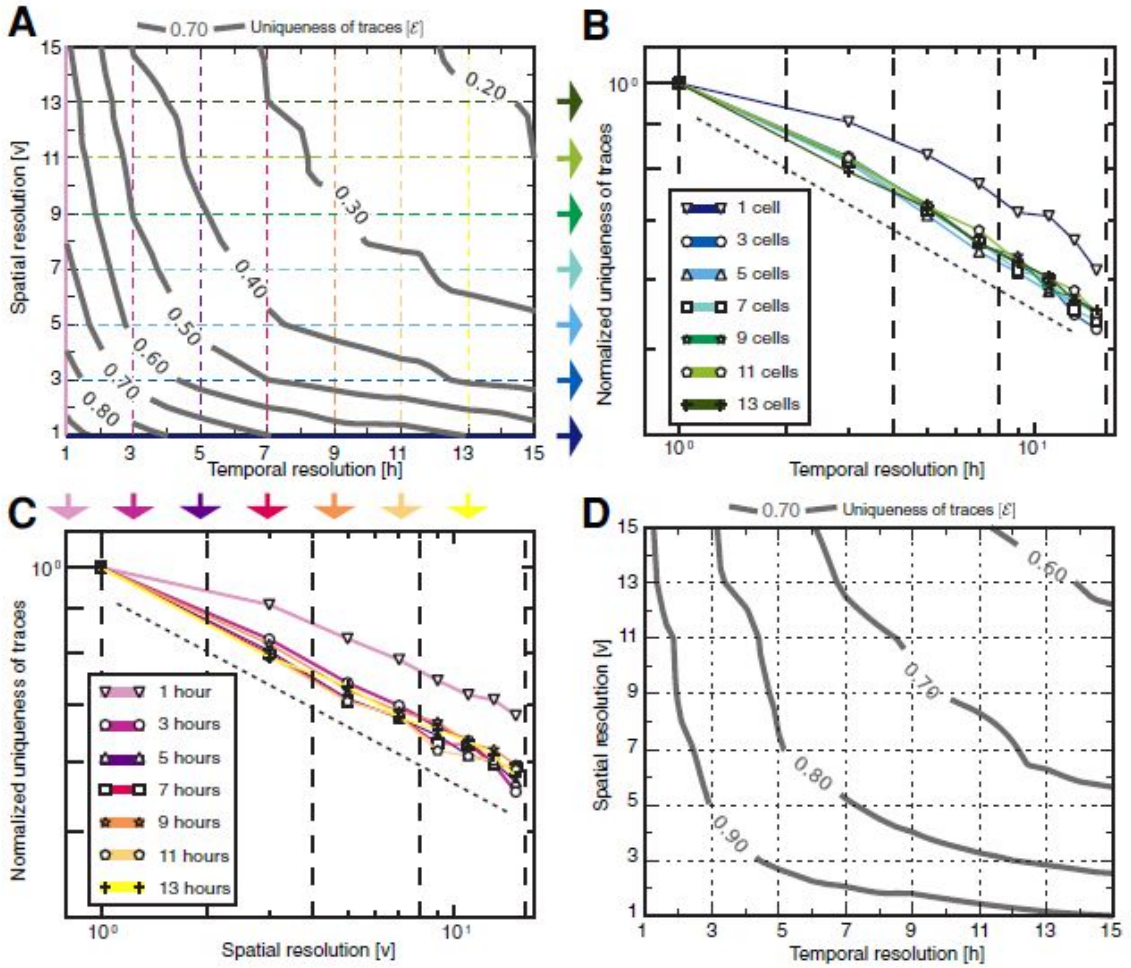
(c) Foursquare

k-Anonymity of Location Data

[de Montjoye et al ‘13]

- Introduced “unicity”, ε, fraction of users with k = 1
- Fit $\varepsilon = \alpha - (vh)^{\beta}$
  - v is spatial granularity,
  - h is time granularity
- $\beta = 0.157 - 0.007p$
  - p = number of points

4 points ->



<- 10 points

k-Anonymity of Location Data

[de Montjoye et al ‘14]

- Data: Credit Card transactions, 1.1M users
- Use *p* random points, including time **and transaction amount.**
- QI as p <space, time> tuples,
  - p=4 -> ε = 0.90
- QI as <time, space, amount> tuples,
  - p=4 -> ε= 0.98

Our Approach: Maximum Likelihood

Is this permutation any good?

Theorem 2: There is a gap between the likelihood of the most likely σ and any other permutation.

Achieving Privacy

Anonymize

- Difficult to anonymize user profiles
- Some hope
- Anonymizing each transaction might go too far

Obfuscate

- Possible to get guarantees with differential privacy
- Pair obfuscation with offloading of computation to local device

Informed Consent and User Choice

- Ask customers if they're okay with it.
- **Put control into users' hands.**

[Riederer HotNets '11]

Face++ API

Response JSON


```
    "x": 181
   },
   "right_eye_upper_right_quarter": {
    "y": 147,
    "x": 247
   },
   "mouth_upper_lip_bottom": {
    "y": 234,
    "x": 218
   }
  },
  "attributes": {
   "gender": {
    "value": "Female"
   },
   "age": {
    "value": 51
   },
   "eyestatus": {
    "left_eye_status": {
     "normal_glass_eye_open": 0.187,
     "no_glass_eye_close": 0,
     "occlusion": 0.001
```

Attacks on Privacy

These are attacks that ***identify*** users.

What about attacks that instead target ***sensitive attributes*** of users?