Mobile Social Media and Demographics Opportunities and Risks

A. Chaintreau, Computer Science Department

Columbia University
Data Science Institute

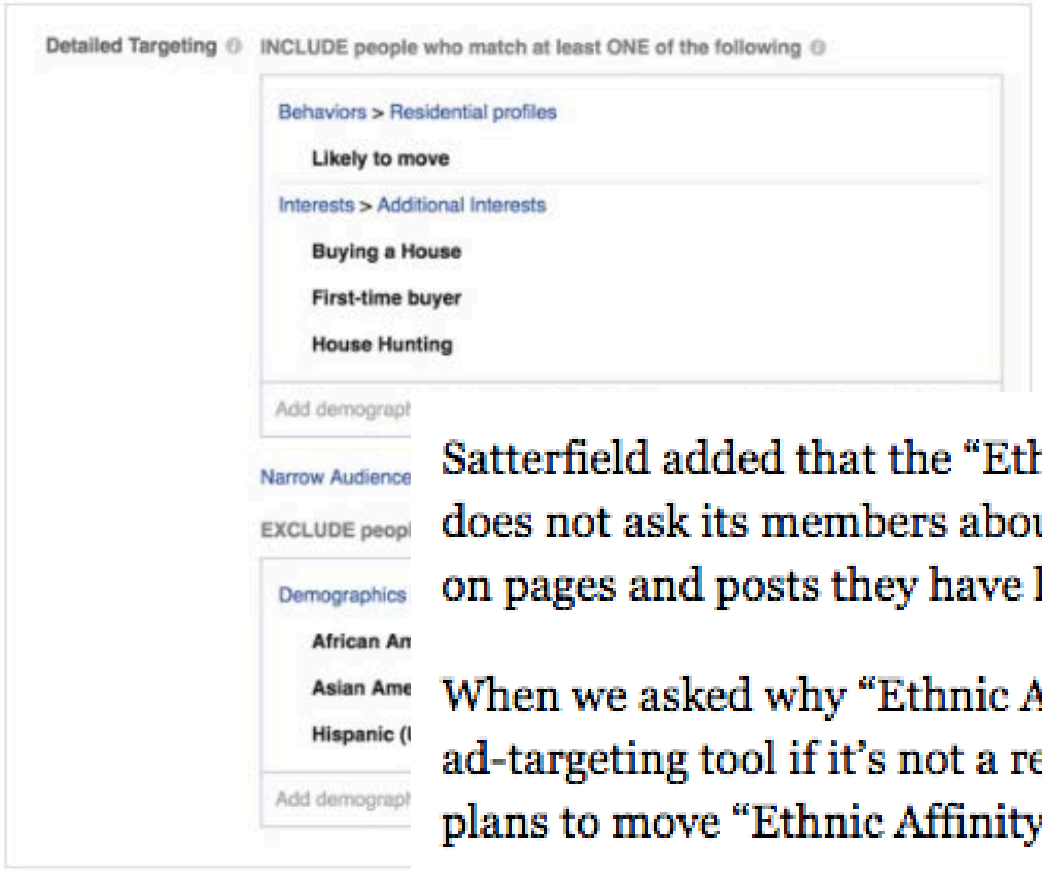
Outline

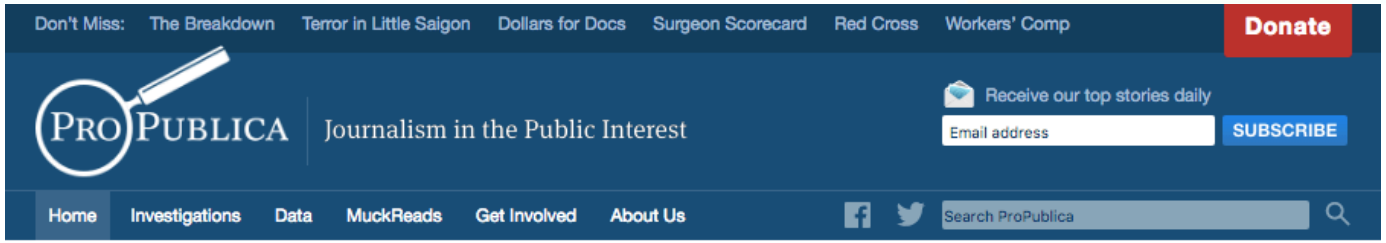
- How I came to care about demographics in big data
  - And why we need new tools for mobile social media

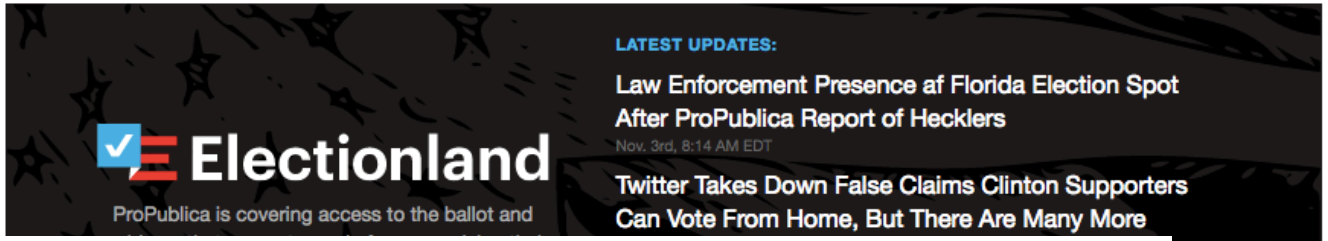
- Risks: Identity Reconciliation using location data
- Opportunities: Demographic study of Urban mobility
- How to scale?

- Conclusion

Big Data to Social Science: We need you!

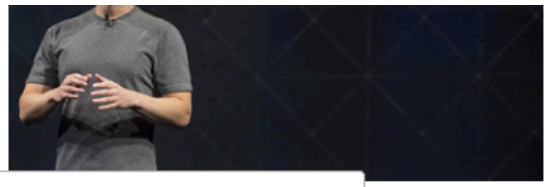




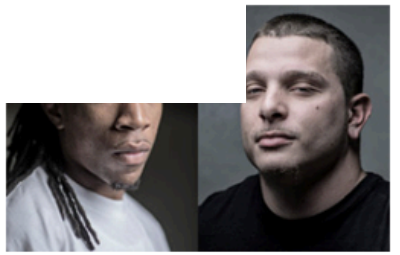


Satterfield added that the "Ethnic Affinity" is not the same as race — which Facebook does not ask its members about. Facebook assigns members an "Ethnic Affinity" based on pages and posts they have liked or engaged with on Facebook.

When we asked why "Ethnic Affinity" was included in the "Demographics" category of its ad-targeting tool if it's not a representation of demographics, Facebook responded that it plans to move "Ethnic Affinity" to another section.



Four members of the Congressional Black Caucus wrote Facebook's Mark Zuckerberg telling him that the company should stop allowing advertisers to exclude people by race.

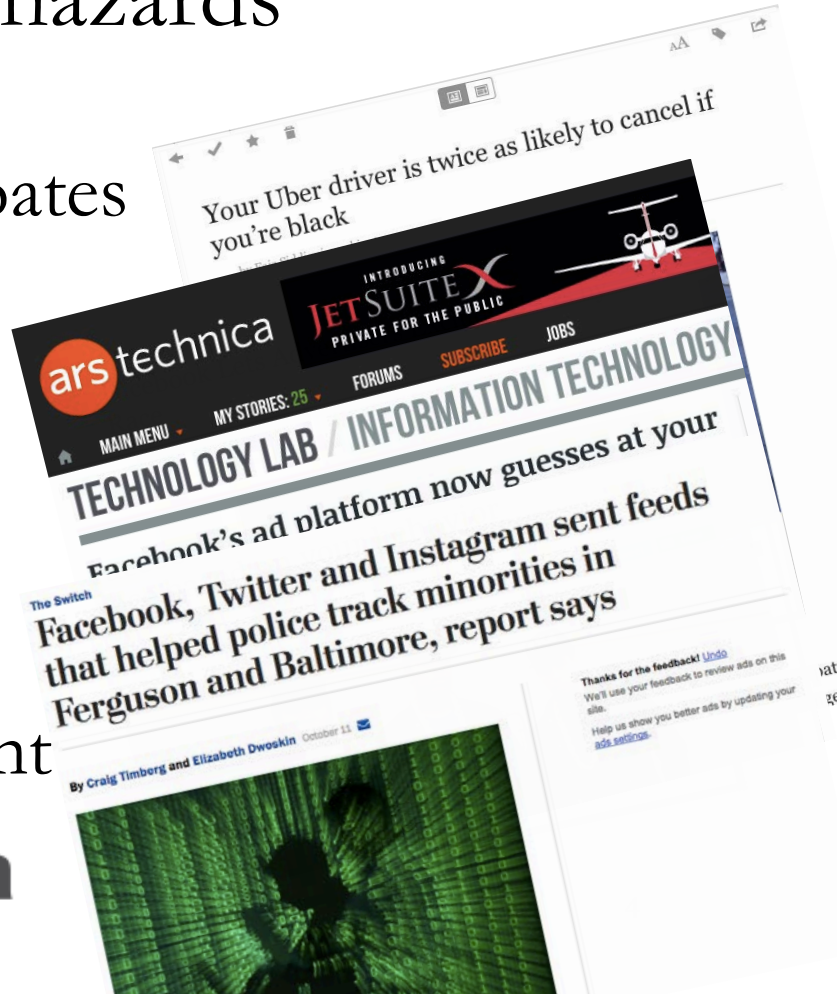


We're investigating algorithmic injustice and

see all

How I came to care about demographics

- Personalization runs the gamut of moral hazards
  - Commerce (Wash. Post, Oct. 31st):
    Data exchanged in "sharing economy" exacerbates discrimination
  - Information (Propublica, Oct. 28th):
    Race induced personalization on Ads
  - Politics (Wash. Post, Oct. 11th):
    FB, Twitter, Instagram used by law enforcement



Geofeedia

My research in a nutshell

- Reconciling Big Data's Disparate Impact with our values

1. Building an infrastructure for transparency
2. Reconciling data to empower users and social scientists
3. When does social sharing reinforce disparate impact

- Promoting scalable reusable tools, incrementally deployable, validated with reproducible data in open access

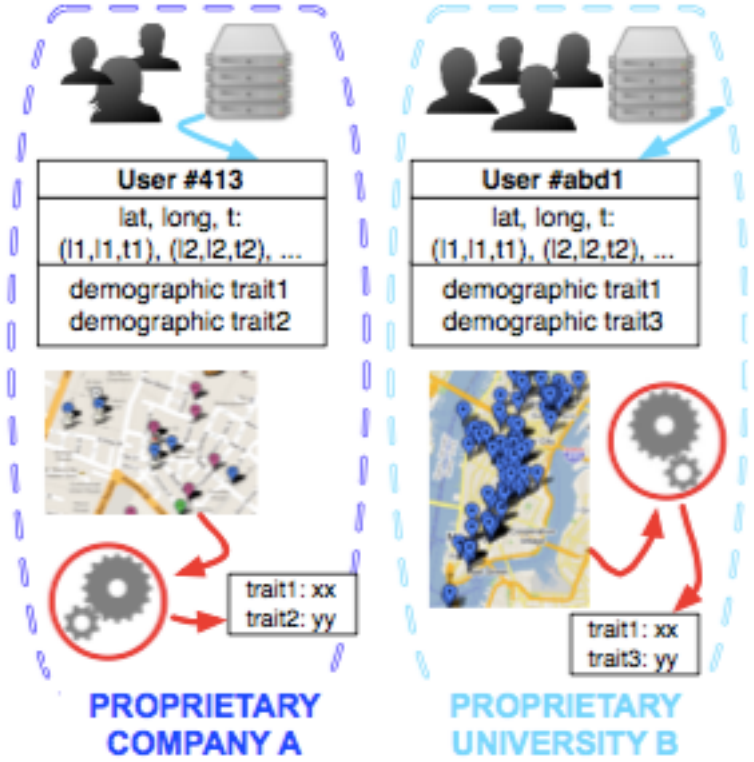
Outline

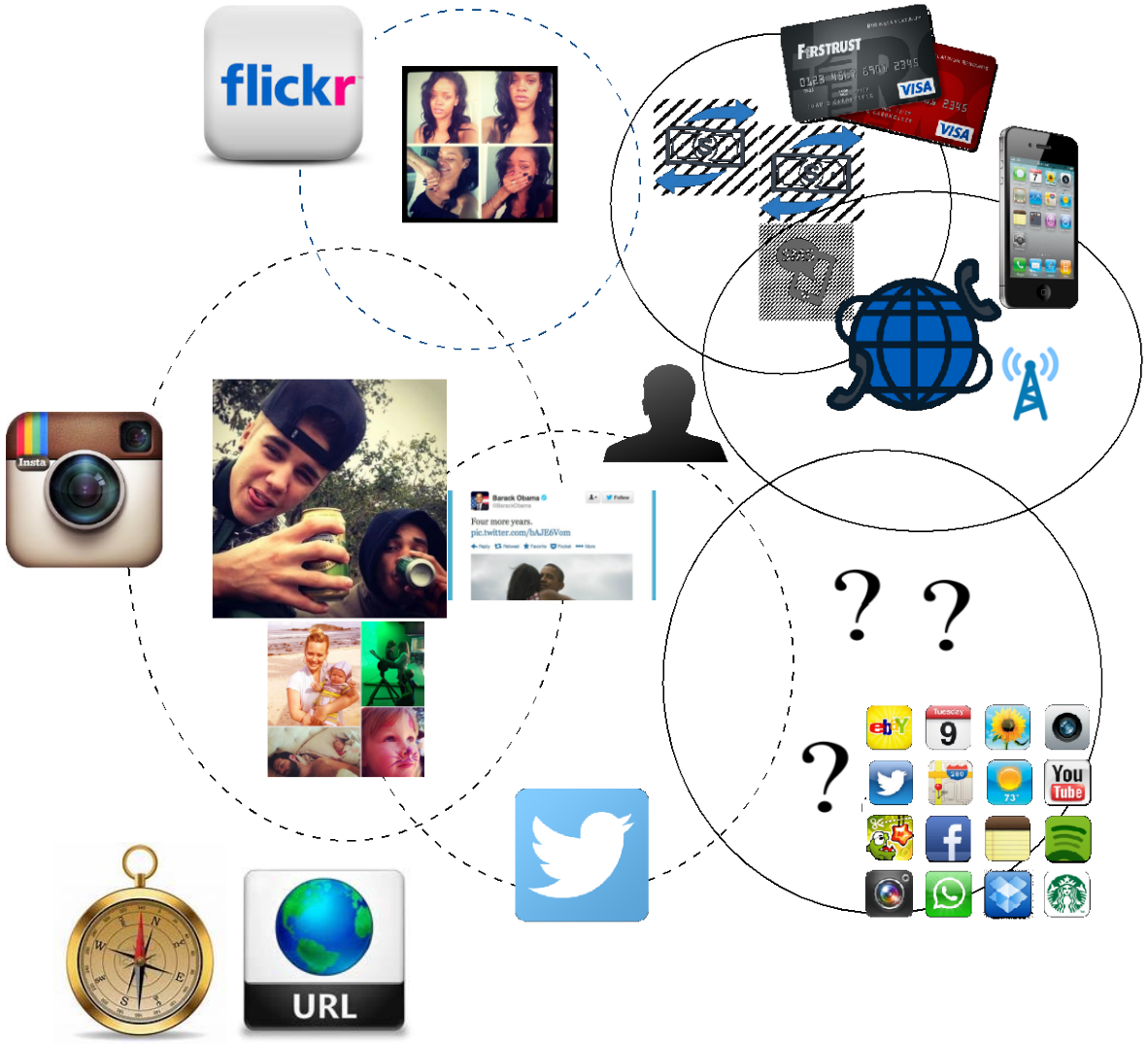
- How I came to care about demographics in big data
  - And why we need new tools for mobile social media

- Risks: Identity Reconciliation using location data
- Opportunities: Demographic study of Urban mobility
- How to scale?

- Conclusion

Mobile Social Media: Bridging the Chasm





Why we need reconciliation tools?

- Privacy typically centers on anonymity in a single-domain
- Big Data has fragmented view (e.g., tweets, calls, transactions) with overlaps in various domains
- "Can two profiles I maintained be linked?", "Can my behavior in be used to discriminate?"

Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study

Hui Zang
Sprint
1 Adrian Ct
Burlingame, CA 94010, USA

Jean Bolot*
Technicolor
735 Emerson St
Palo Alto, CA 94301, USA

SCIENTIFIC REPORTS

OPEN

Unique in the Crowd: The privacy bounds of human mobility

IDENTITY AND PRIVACY

Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,[1*] Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Ubiquitous technologies

scale behavioral data sets to the invention of the microscope (*1*). New fields such as computational social science (*2–4*) rely on metadata to address crucial questions such as fighting malaria, study-

Downloaded from www.sciencemag.org on J

Outline

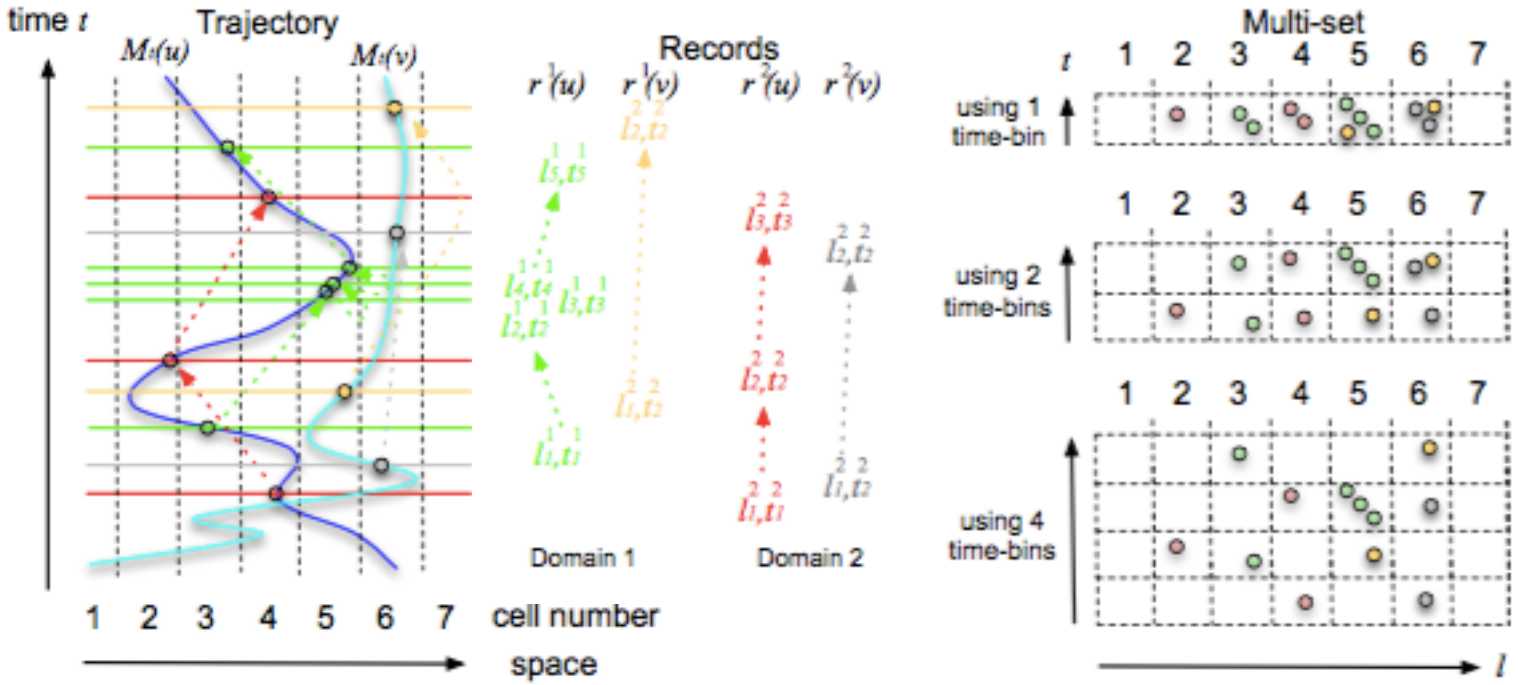
- How I came to care about demographics in big data
  - And why we need new tools for mobile social media

- Risks: Identity Reconciliation using location data
- Opportunities: Demographic study of Urban mobility
- How to scale?

- Conclusion

Identity Reconciliation using location

- Using (space, time)
  - No username, friends, tastes, etc.
- ≠ indiv. anonymity
  - Non-overlapping set
  - cross domain
- Difficult & lack data



How to combine positive and negative signals

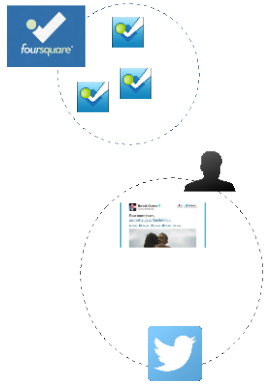
- 2 Poisson Processes sampled
  - From the same original pattern
  - 3 parameters: λ, p_1, p_2
- Our analysis:
  - Computer a max matching.
  - Thm: In expectation, there is only one score maximizer, which is the correct identity reconciliation
  - Parameters? can be estimated from imperfect matching

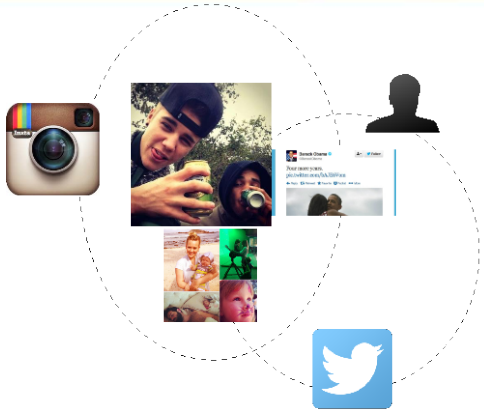
Algorithm 1: Our reconciliation algorithm

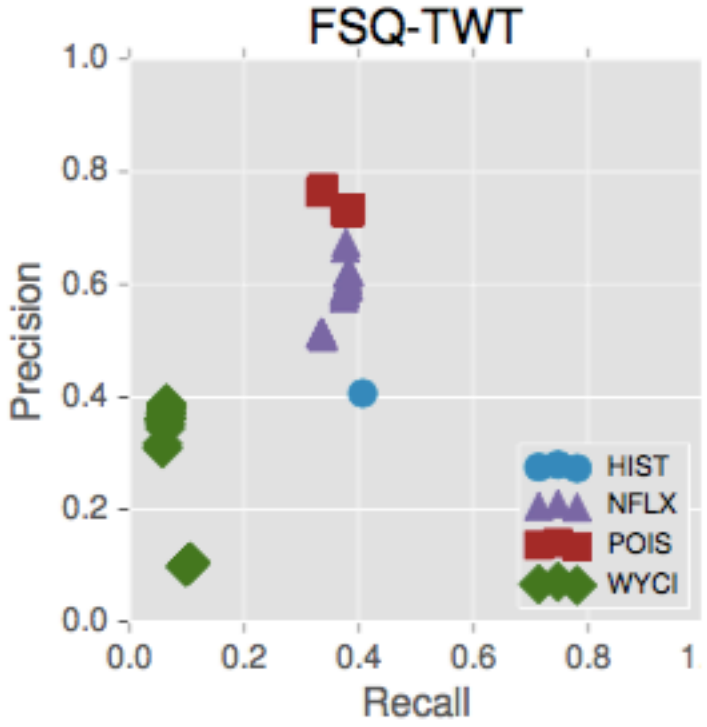
Require: $\forall u \in U : r^1(u), \forall v \in V : r^2(v), \{\lambda_{\ell,t}\}$
  for $(u, v) \in (U \times V)$ **do**
    $w(u, v) = \sum_{t \in T} \sum_{\ell \in L} \ln \phi_{\ell,t}\left(a_1(u, \ell, t), a_2(v, \ell, t)\right)$
  end for
  Let $E = \{w(u, v) : (u, v) \in (U \times V)\}$
  Compute the maximum weighted matching on the bipartite graph $B(U, V, E)$
  return the function that maps matched vertices.

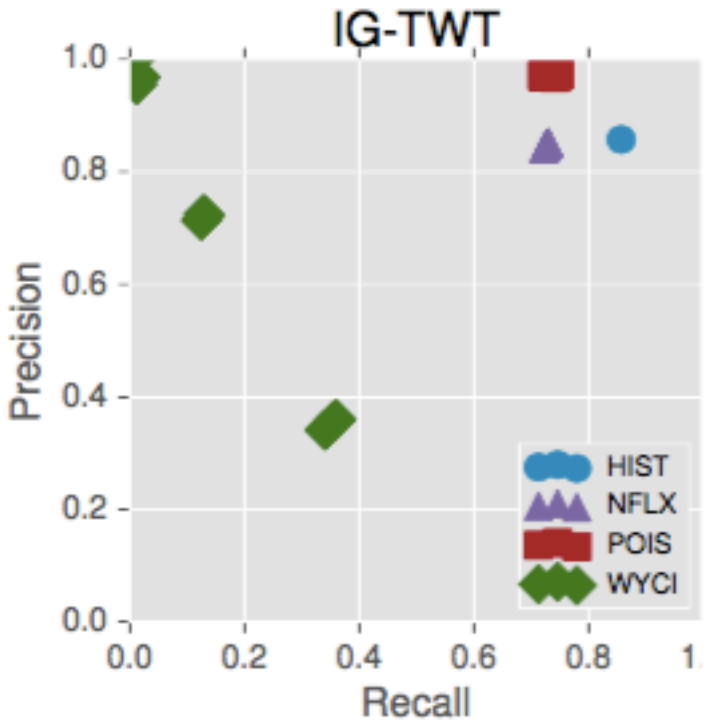
$$\frac{e^{-\lambda(1-p_1-p_2)}}{(\lambda(1-p_1))^{a_1}(\lambda(1-p_2))^{a_2}} \sum_{k \geq \max(a_1,a_2)} \frac{(\lambda(1-p_1)(1-p_2))^k k!}{(k-a_1)!(k-a_2)!}$$

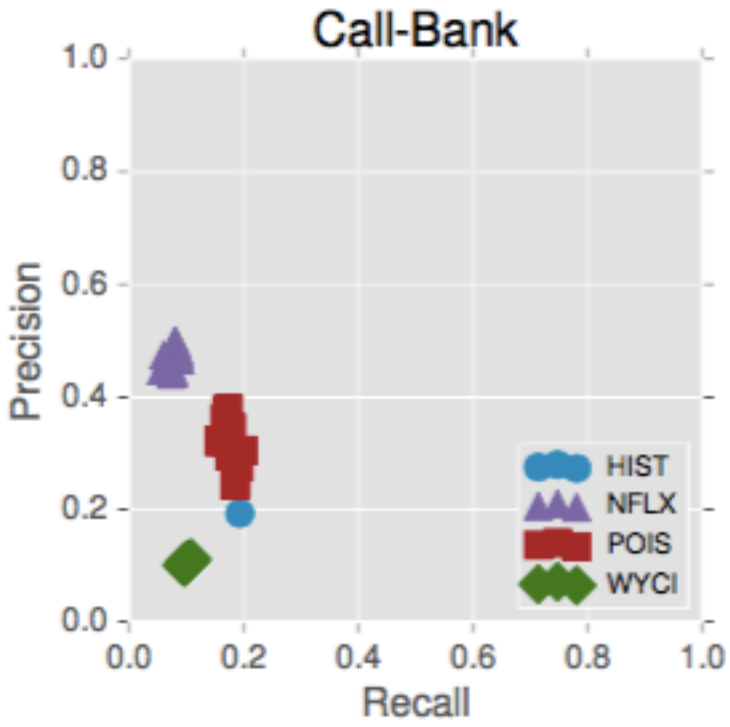
Reality Test in 3 increasingly hard examples











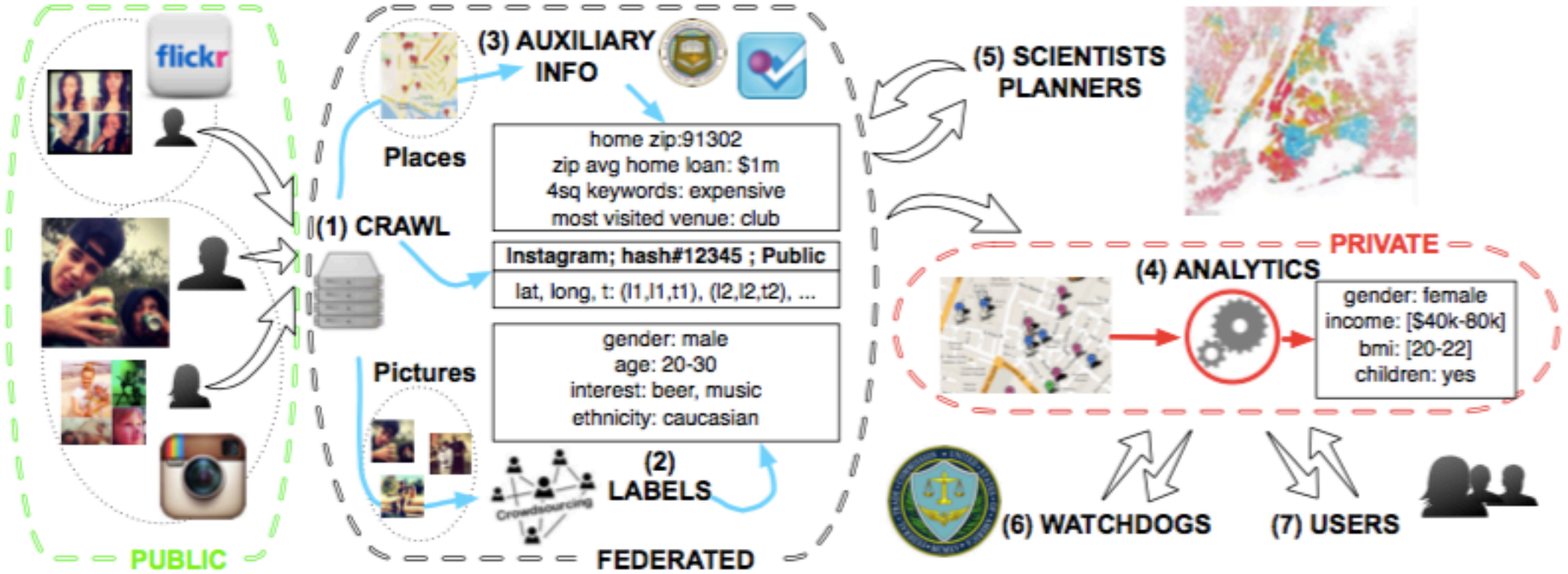
Outline

- How I came to care about demographics in big data
  - And why we need new tools for mobile social media

- Risks: Identity Reconciliation using location data
- Opportunities: Demographic study of Urban mobility
- How to scale?

- Conclusion

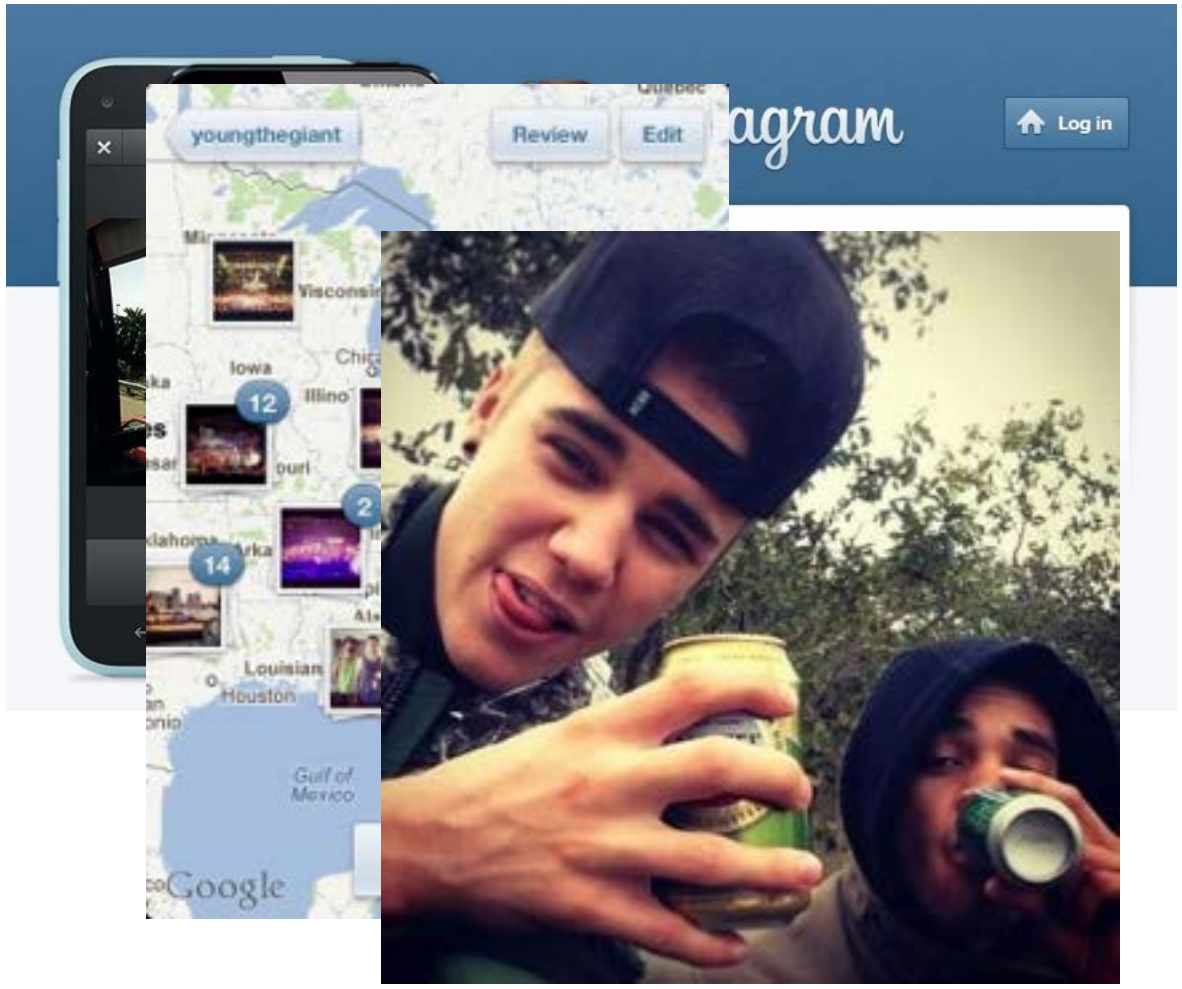
Our proposition: Mobility Data Unchained



But where do you start?

- Geotagged photos
  - 118,000 users
  - 35M geotags
  - Did **not** download or save any photos
  - Only public profiles

- Labeled in 2 ways:
  - Benchmark group
  - Crowdsourcing for scale

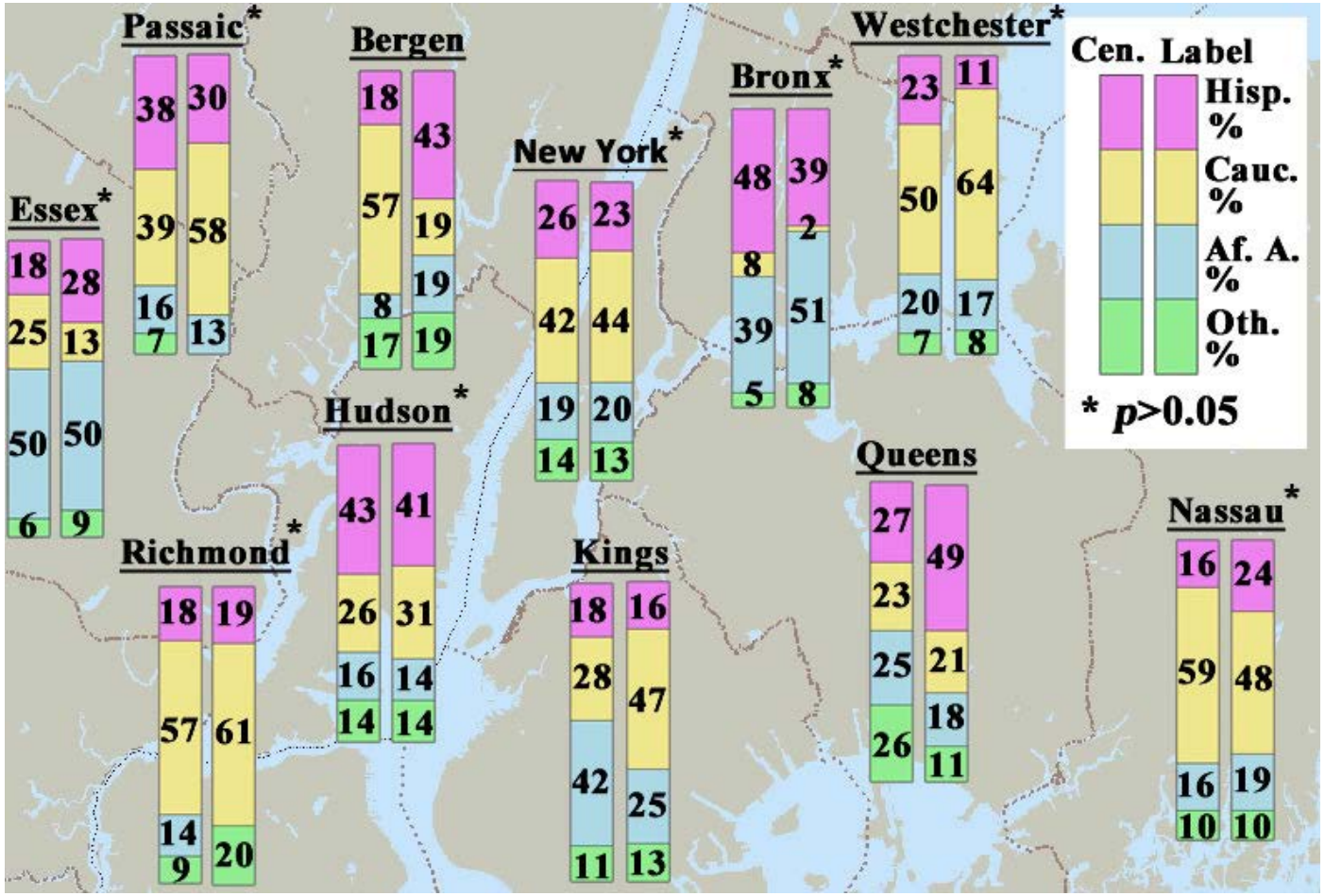


Male, Caucasian, Likes beer

Promise #1: Reduce cost of surveys

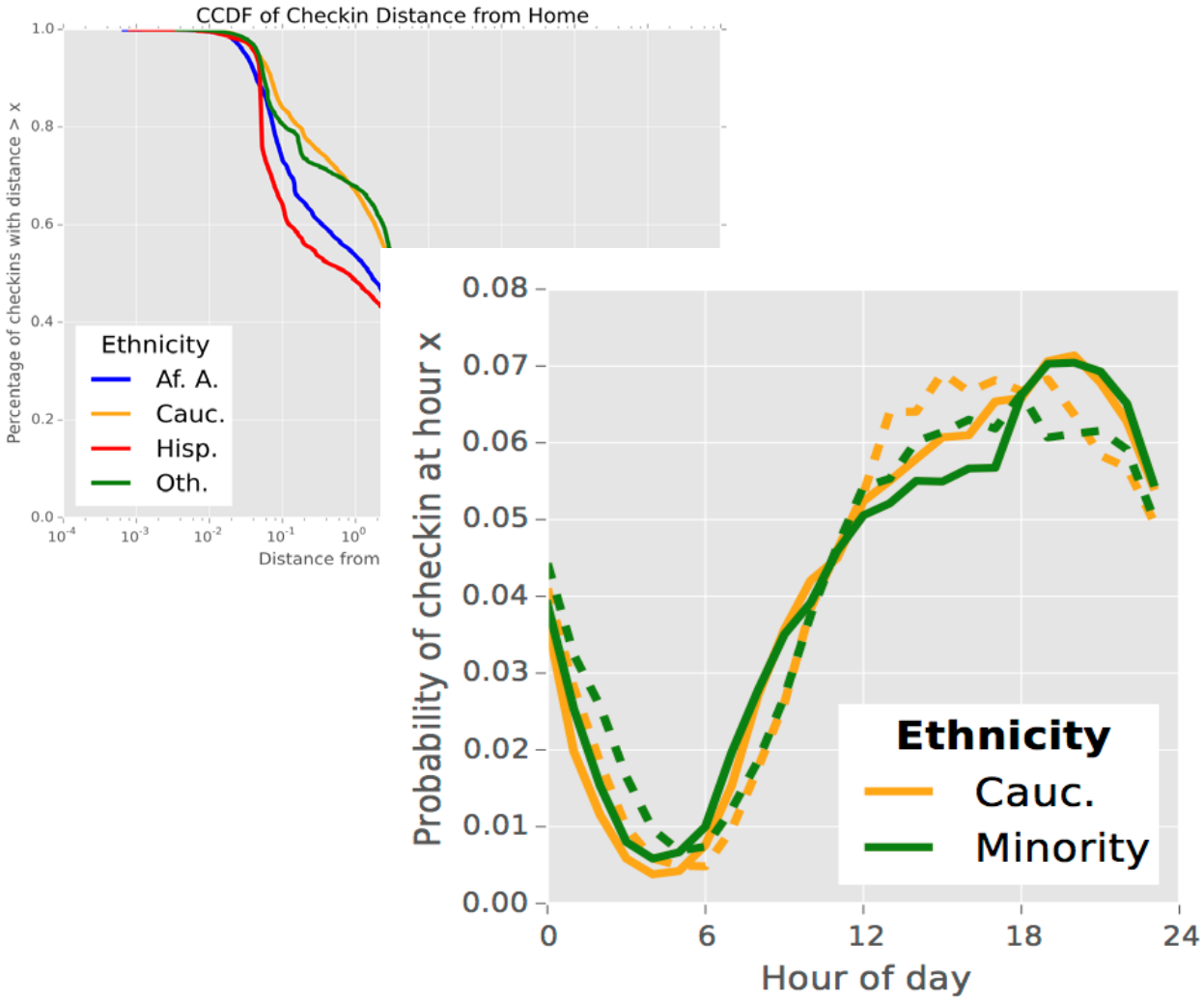
- Survey are (very) costly and slow, Call Details Records too.

- Reproduce CDR trends
  - Weekly maximum distances
  - Median distances traveled
  - Average distance traveled from home



Promise #2: Enrich our understanding

- Distance traveled vary greatly across races
- Activity at night vary greatly across gender
- Even weekly temporal activity pattersns vary across races

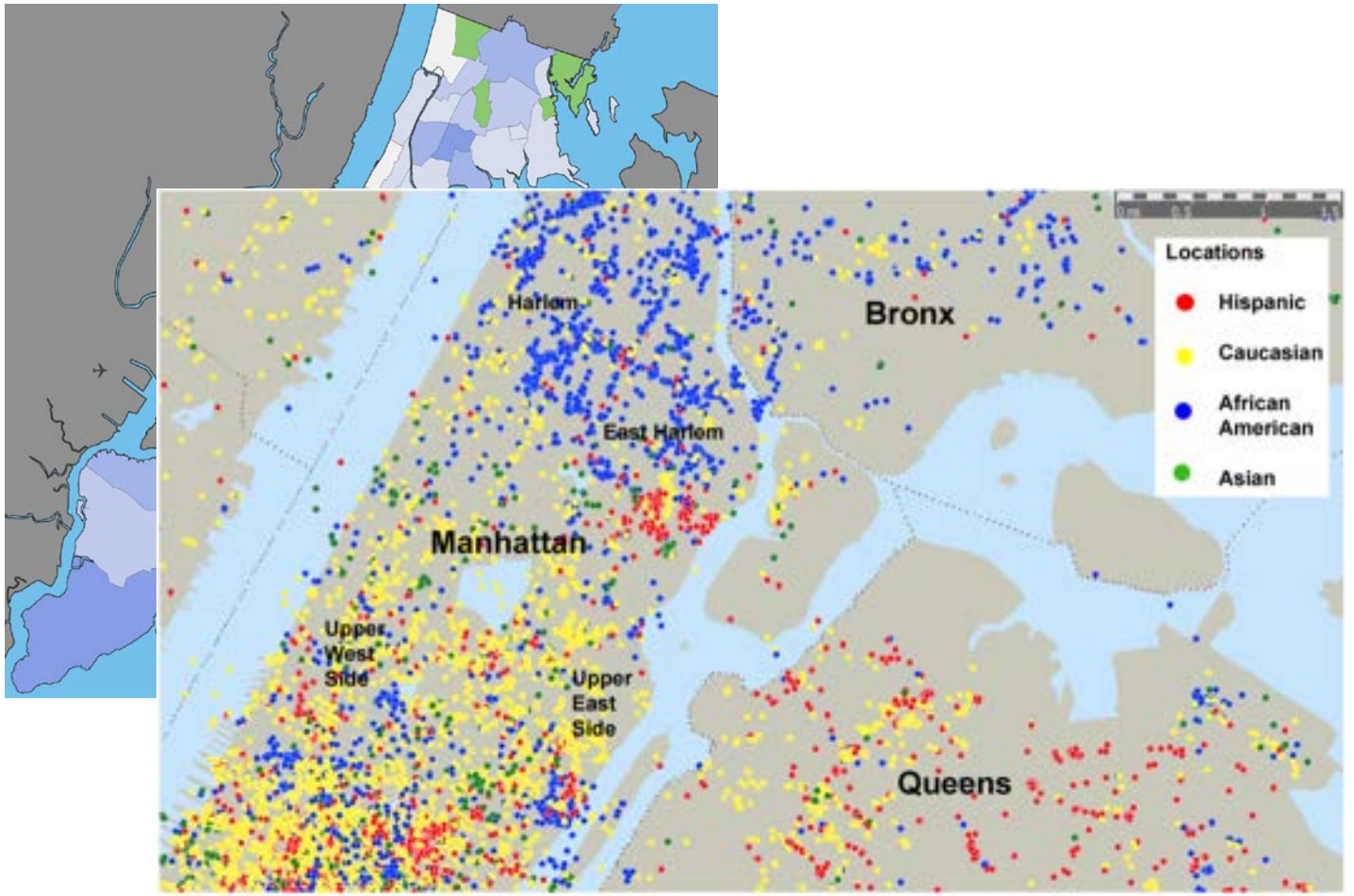


Promise #2: Enrich our understanding

New Study of Segregation

- **Interaction Segregation:** how much people of different groups meet.
- **Typically studied with home location zipcode**

Using public social media we found more **interaction**

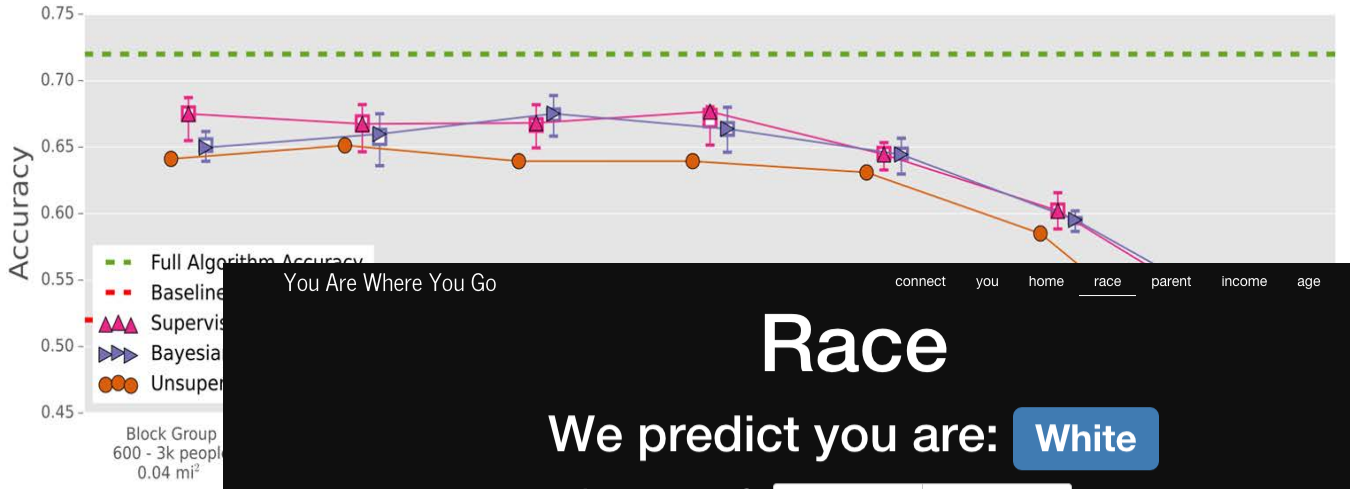


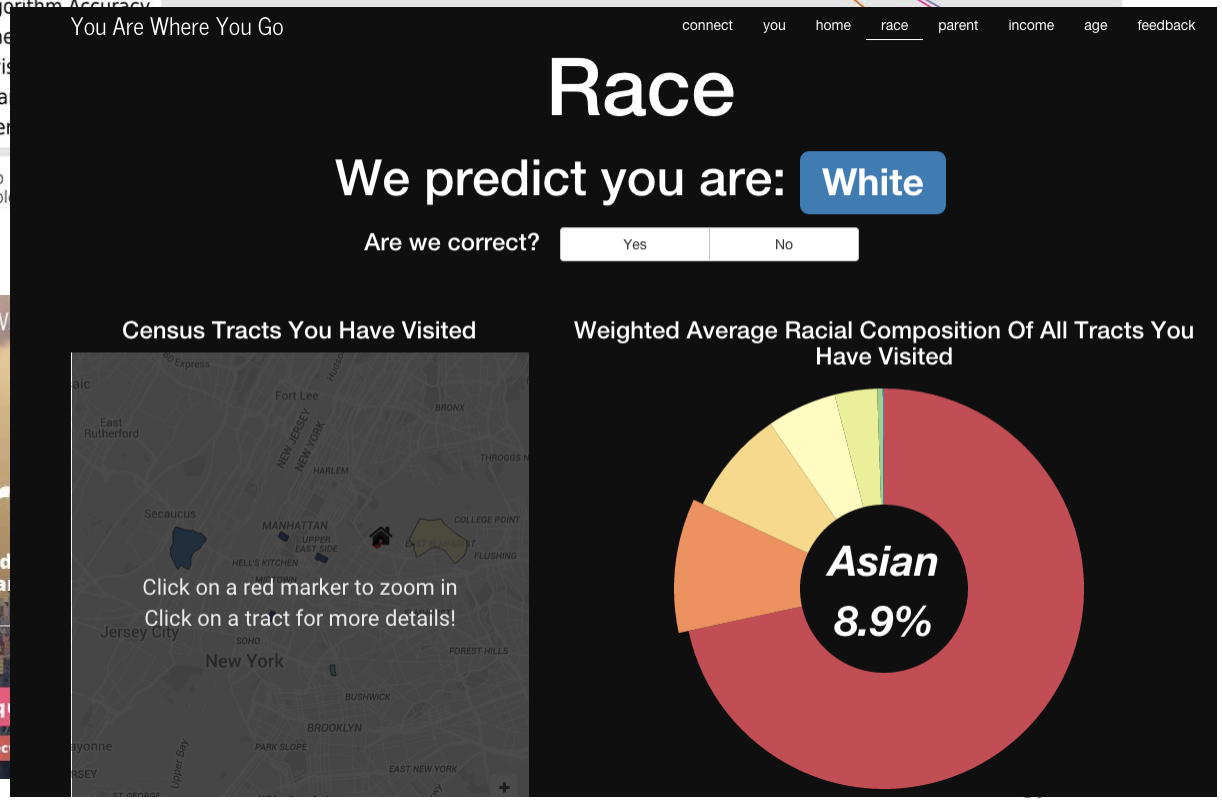
Promise #3: Document Discriminative Power

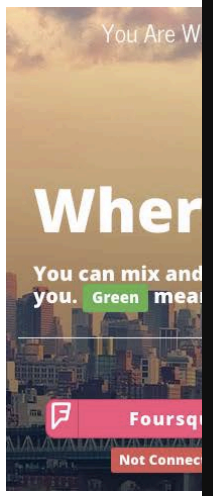
Document the discriminative power of location data

- Even anonymously
- A few coarse-grain location data suffice

Build into an auditing tool
https://find-you.herokuapp.com
(demo at ACM WWW '16)







Outline

- How I came to care about demographics in big data
  - And why we need new tools for mobile social media

- Risks: Identity Reconciliation using location data
- Opportunities: Demographic study of Urban mobility
- How to scale?

- Conclusion

Why we need a Second Generation?

- Collecting accurate labels from Crowdsourcing is bottleneck
  - Does not leverage the scale of information widely available
  - Remains slow and/or the most significant cost
  - Crowd not reproducible, and is it neutral?
- Can we leverage decade of Computer Vision progress?
  - Face++ API : Asian/Black/White, Female/Male, confidence
  - What if multiple faces? In multiple pictures?
    - Majority Rules/Confidence Majority Rules/Normalized Picture/Profile Picture

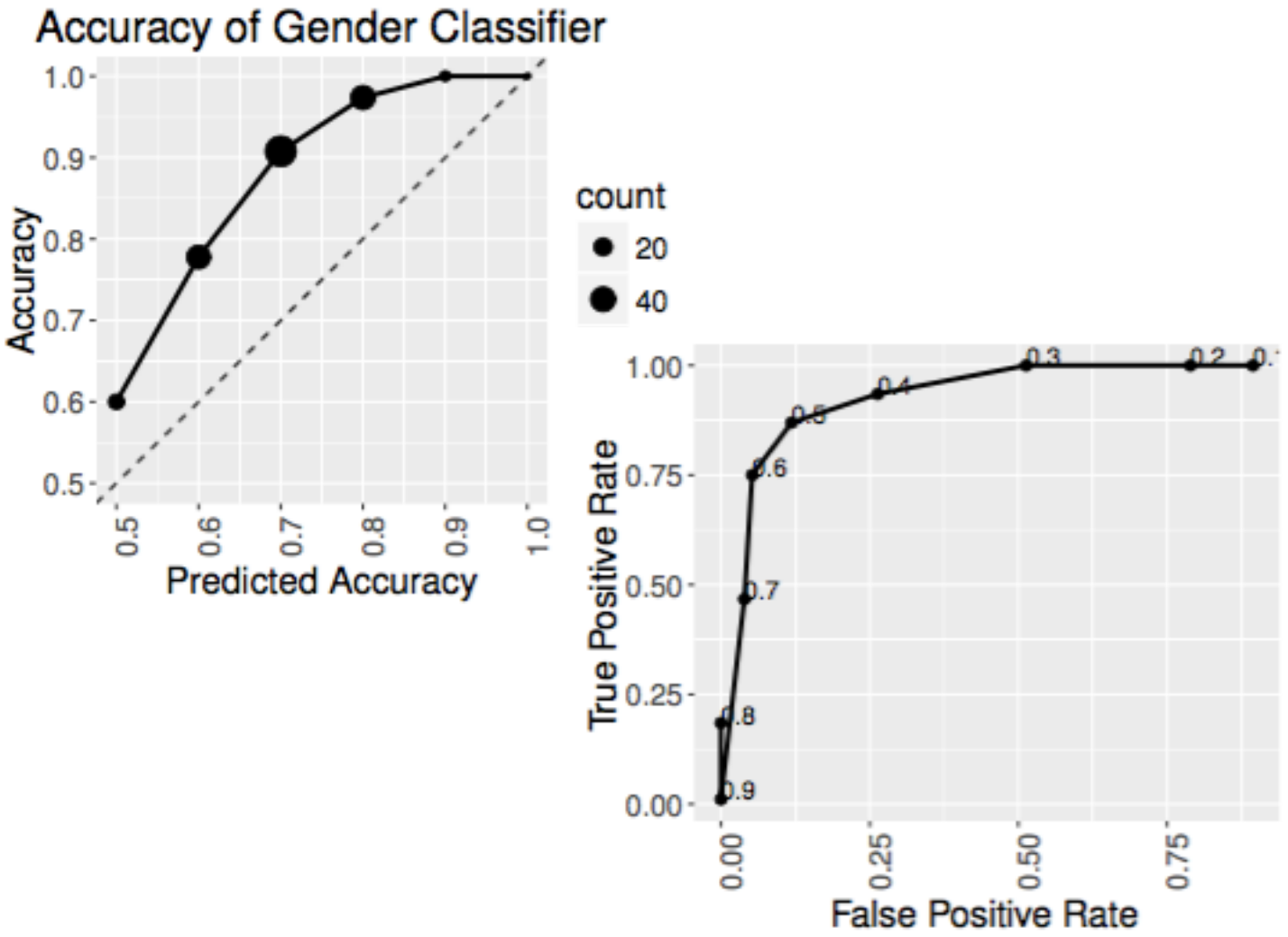
Name	*Cost*	*Speed*	*Accuracy*
Manual	Expensive	Moderate	High
Face Recognition	Free / Cheap	Slow	Moderate
Census	Free	Fast	Low

Table 2: Demographic Labeling Techniques

Lessons learned so far: The good

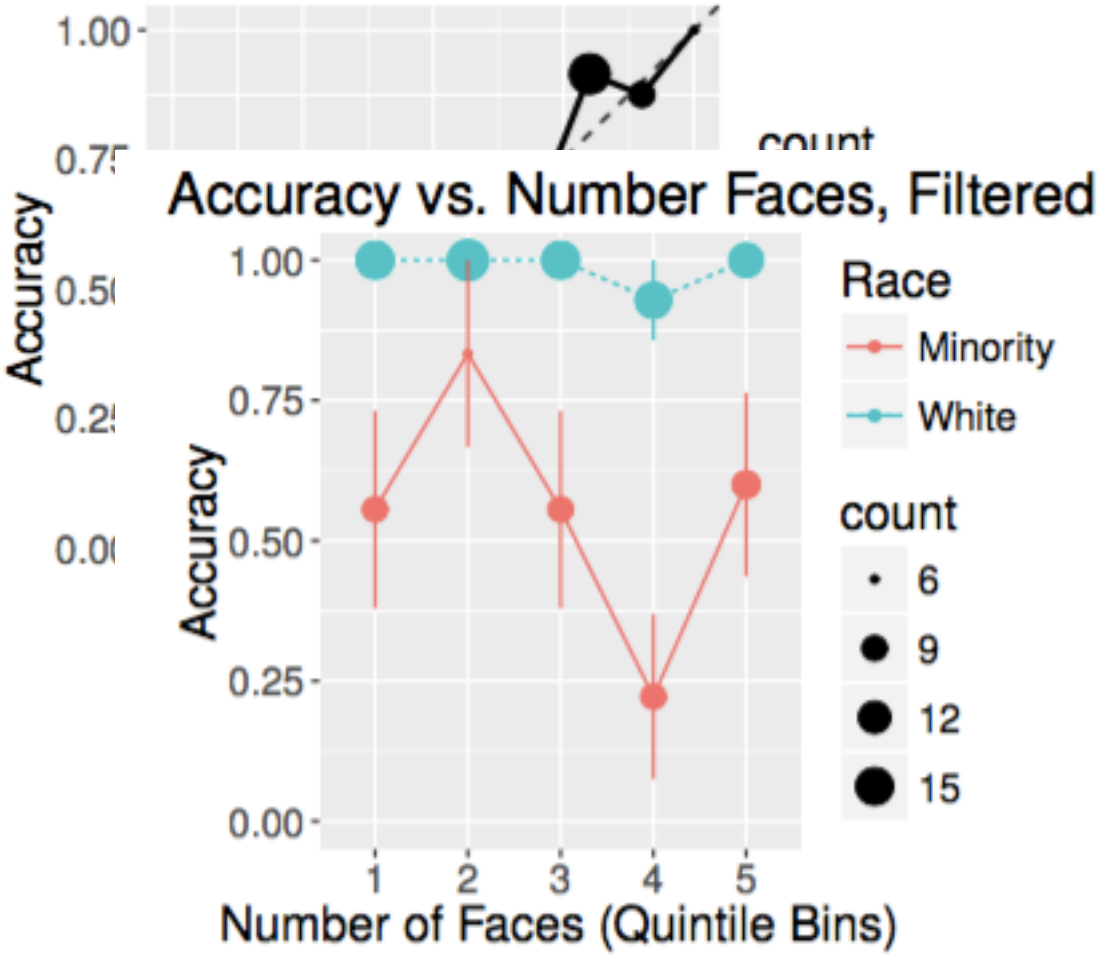
- Promising: gender
  - Accuracy: 87.5%
- Conservative Pred.
- ROC curve

  0.6-threshold finds 75% female with only 7% misclassified male



Lessons learned so far: The ugly

- Race indicates challenges
  - Accuracy ~82%
  - Confidence may be overstated!
- But, wait there is worse!
  - Minority: more data means less accuracy! (tyranny of majority)
  - Simple "neutral" algorithms may under-represent specific groups!



Lessons learned so far: The less ugly

- Balanced Error Rate (BER)
  - Average of errors made across classes (over-represents small classes in effective error rates)
  - Optimizing for this metric can remove some of the bias

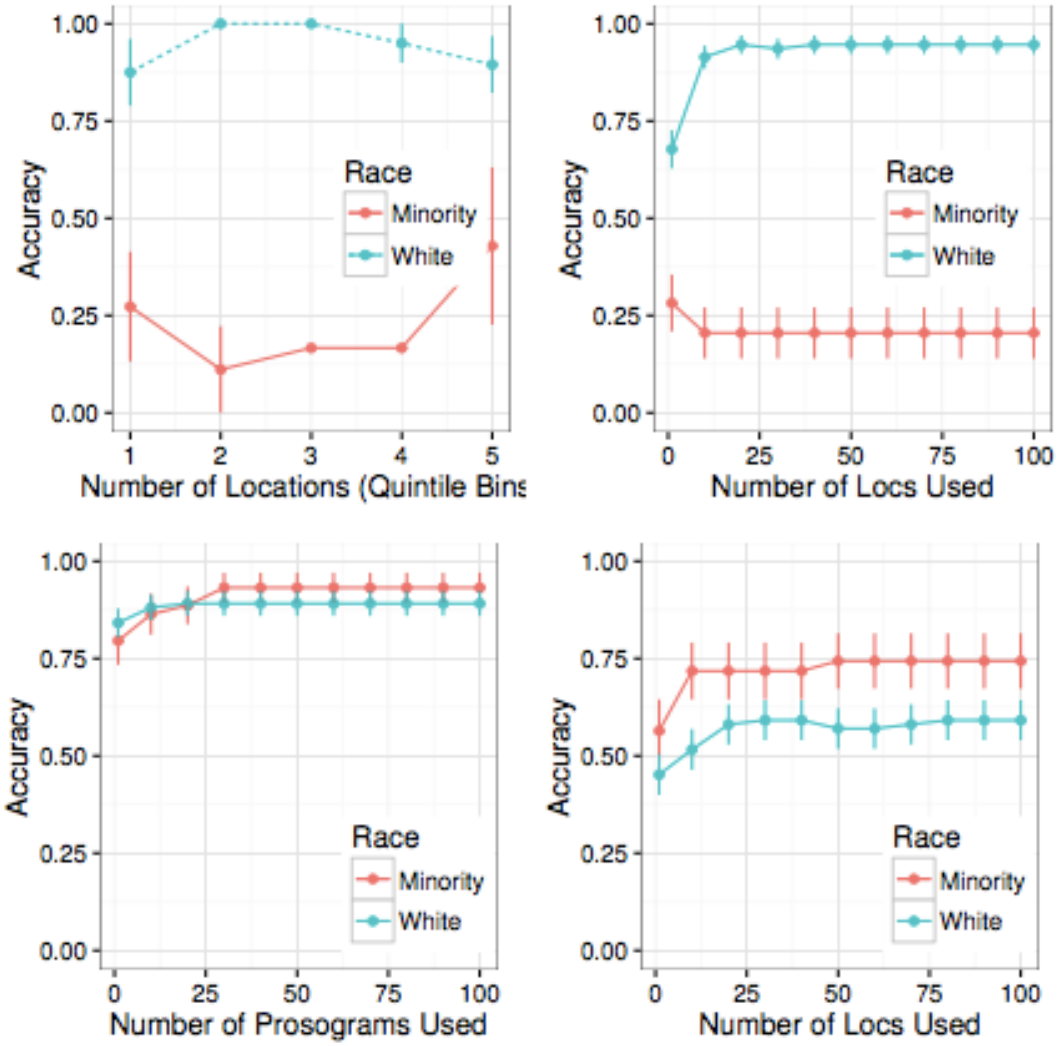


Figure 20: Location or faces vs. accuracy, with a choice of threshold determine by Balanced Error Rate.

Points for discussion

- Should we use Social Media for demographics research?
  - Yes, because it is feasible
  - Yes, if we want to protect users against new discrimination form
  - Yes, if we want to have equal representation
- But experience brings additional caution
  - Crowdsourcing raises cost/neutrality issues?
  - Even innocuous neutral algorithms can under-represent minorities
  - Critical to understand those effects when informing policies

Thank you for your time!

- [1] C. J. Riederer, S. Zimmeck, C. Phanord, A. Chaintreau, and S. M. Bellovin, “I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data,” presented at the COSN '15: Proceedings of the third ACM conference on Online social networks, 2015, pp. 185–195.
- [2] C. Riederer, D. Echikson, S. Huang, and A. Chaintreau, “FindYou: A Personal Location Privacy Auditing Tool (demo),” WWW '16: Proceedings of the 25th international conference on World Wide Web, Apr. 2016.
- [3] C. J. Riederer and A. Chaintreau, “Scaling Up the Census with Social Media,” upcoming (presented at Proceedings of AAAI ICWSM Workshop on The Social Media and Demographic Research Workshop), May 2016.