UNIVERSITY of WASHINGTON

FAT Databases

Bill Howe

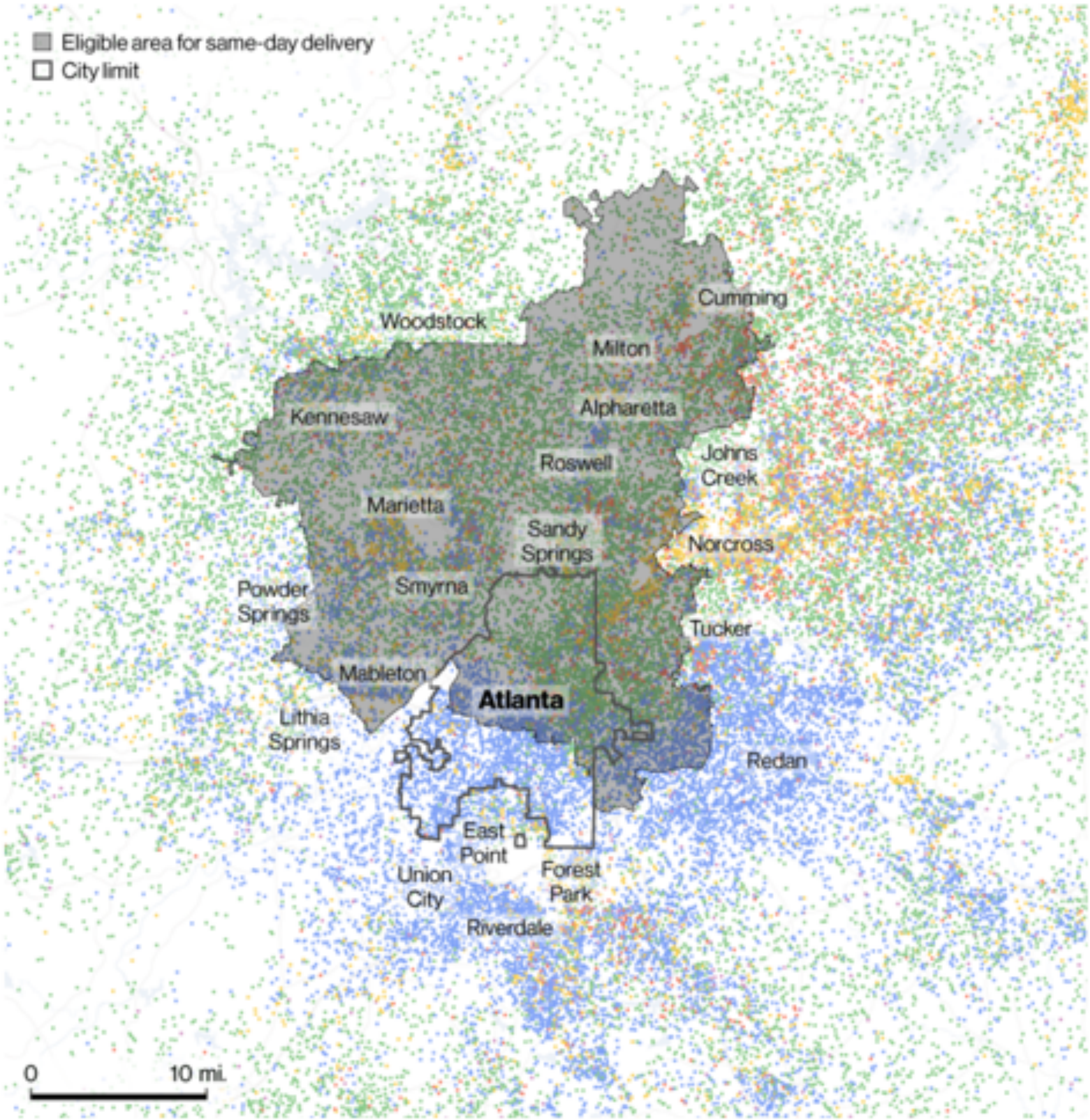
University of Washington

The next ~10 minutes…

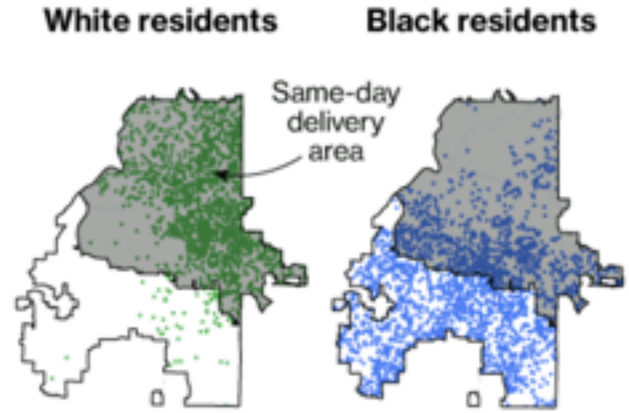
- Some more examples of the problems
- Some topics for DB research

Amazon Prime Now Delivery Area: Atlanta

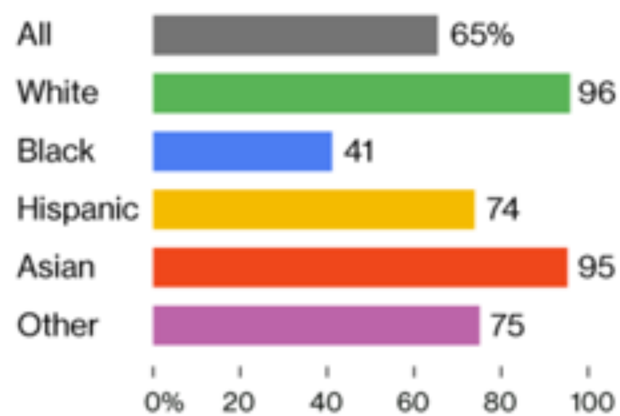
Bloomberg, 2016



The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.



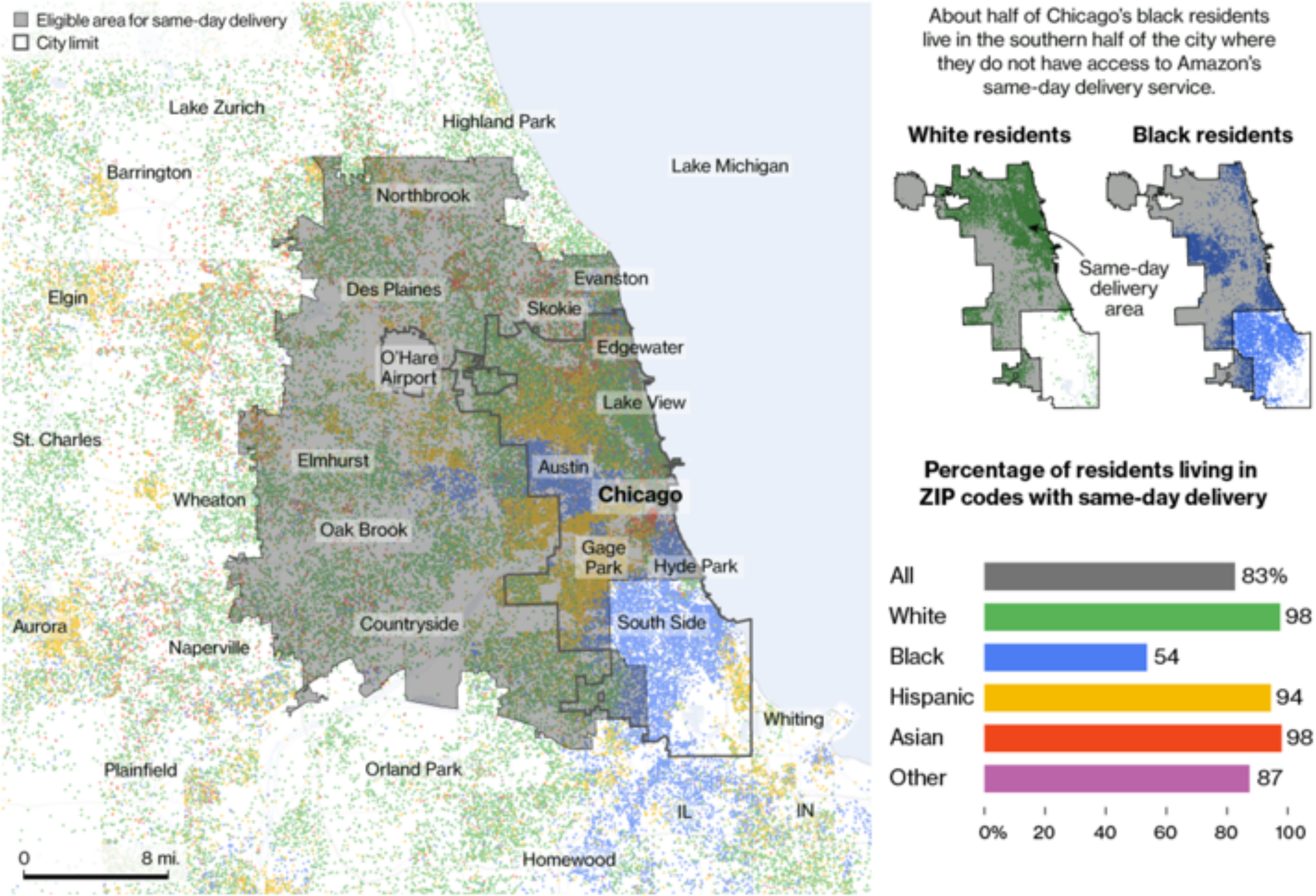
Percentage of residents living in ZIP codes with same-day delivery

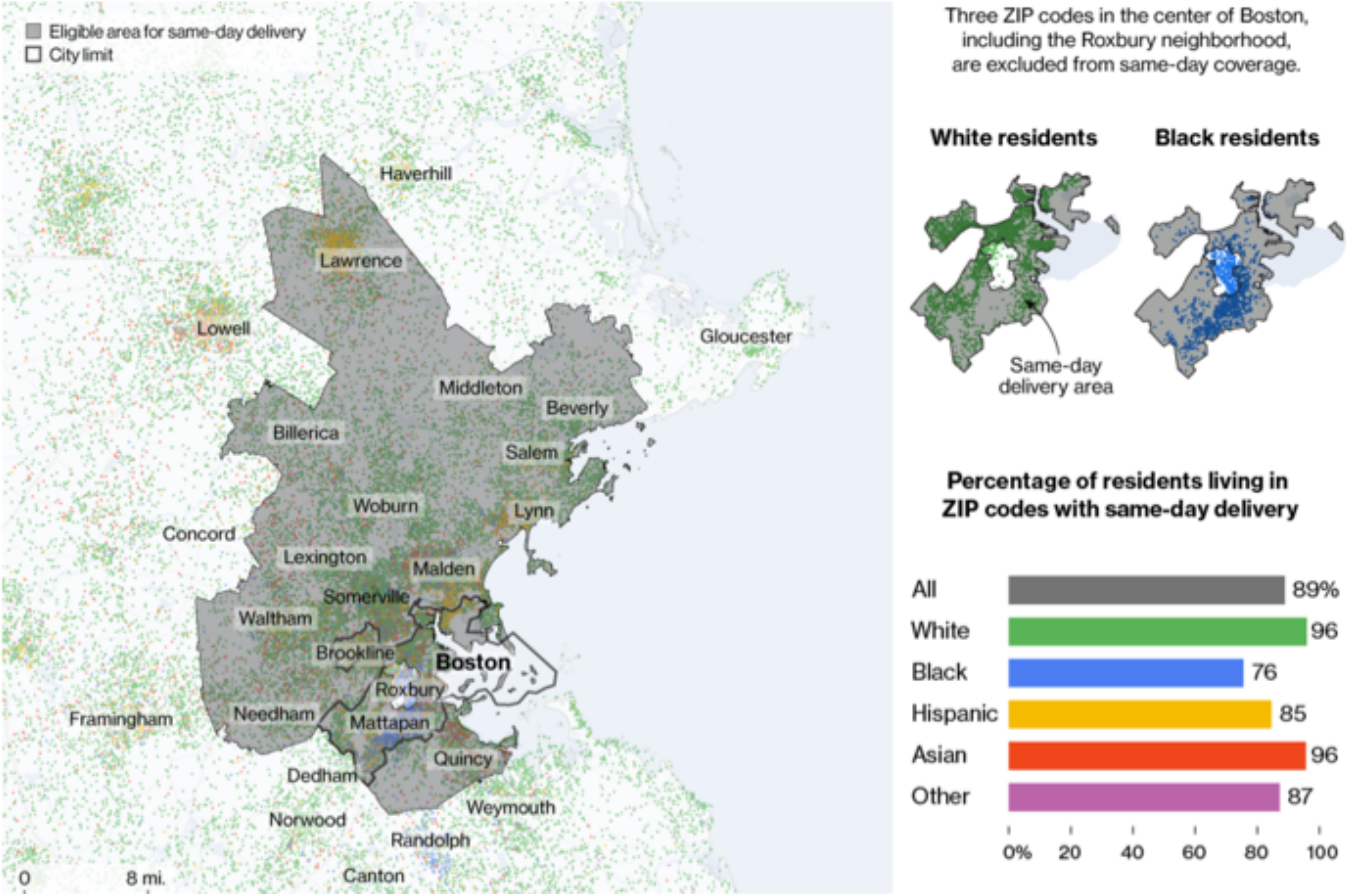


Population percentages are based on American Community Survey estimates and have a 90% confidence interval.

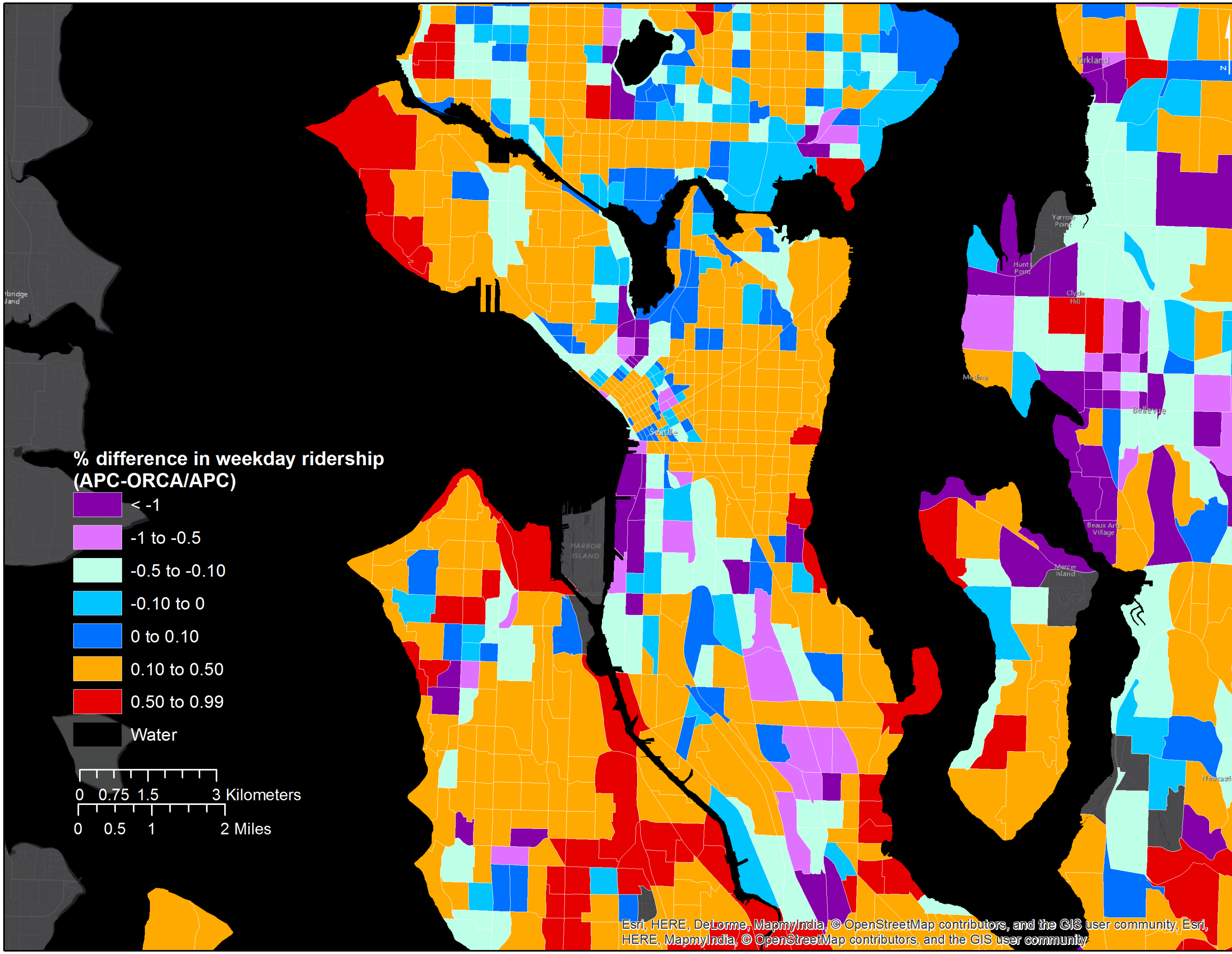
Amazon Prime Now Delivery Area: Chicago

Bloomberg, 2016



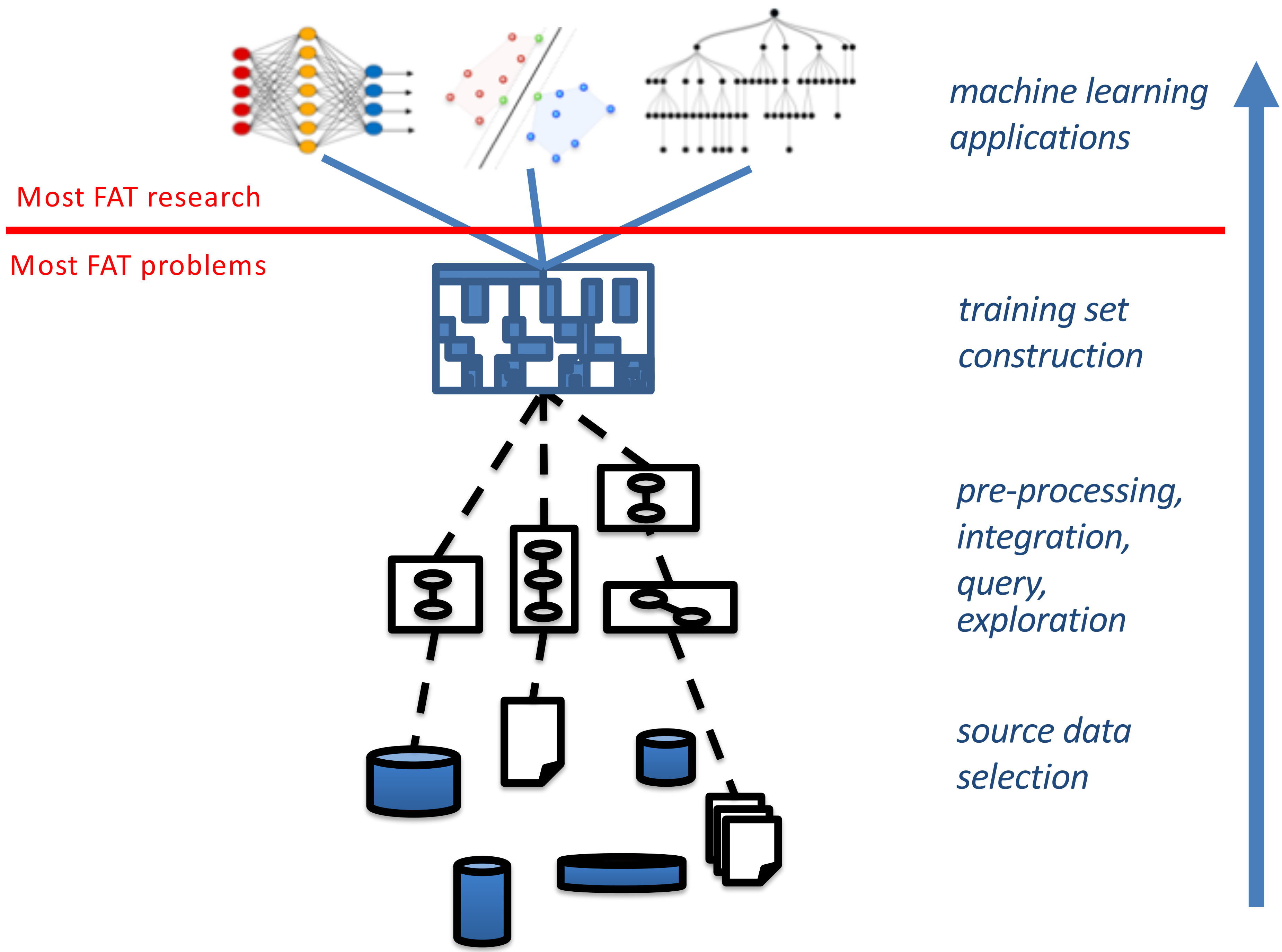

Eligible area for same-day delivery
City limit
Haverhill
Lawrence
Lowell
Gloucester
Middleton
Beverly
Billerica
Salem
Woburn
Lynn
Concord
Lexington
Malden
Somerville
Waltham
Brookline
Boston
Roxbury
Framingham
Needham
Mattapan
Quincy
Dedham
Weymouth
Norwood
Randolph
Canton
0
8 mi.
Three ZIP codes in the center of Boston, including the Roxbury neighborhood, are excluded from same-day coverage.
White residents
Black residents
Same-day delivery area
Percentage of residents living in ZIP codes with same-day delivery
All 89%
White 96
Black 76
Hispanic 85
Asian 96
Other 87
0% 20 40 60 80 100

Data source selection:
Bias in transportation measurements



Mark
Hallenbeck

TRAC



machine learning applications
Most FAT research
Most FAT problems
training set construction
pre-processing, integration, query, exploration
source data selection

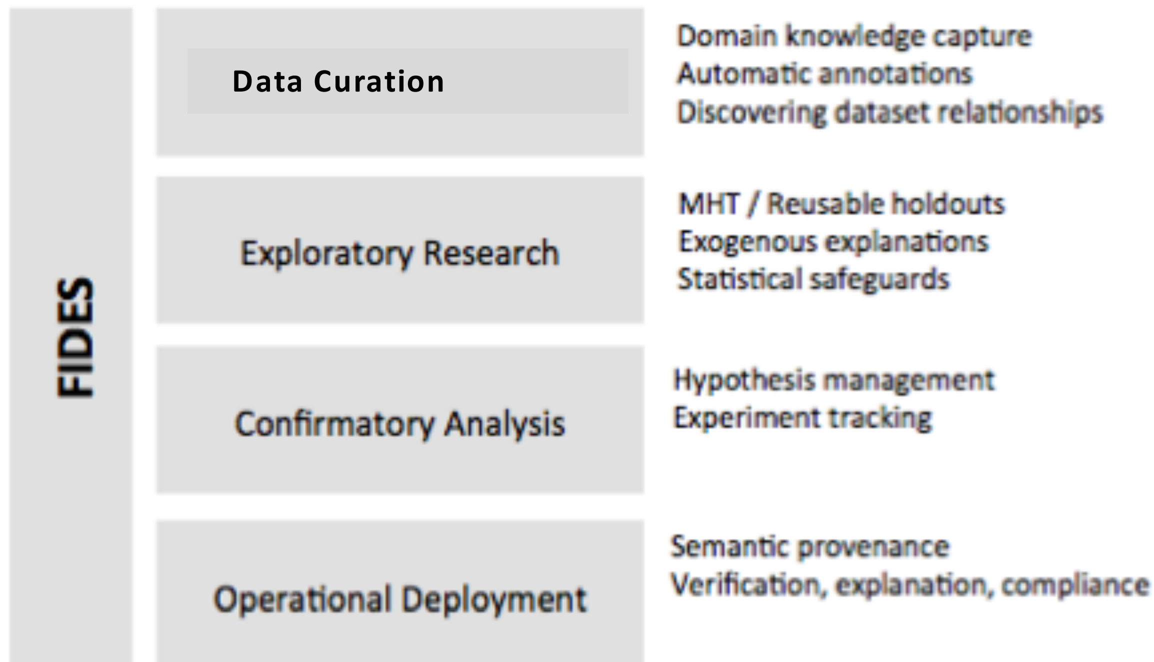
Databases are becoming “training set management systems”

- Claim: Query results are increasingly being used to train models
- So the exact answer to the query result is not that important. It’s important that the trained model gets the right answer on unseen data
- We need declarative specification (i.e., SQL) and management of **high-quality training sets**
- What’s a high-quality training set?
  - It’s a bad training set if the resulting classifier doesn’t work, or overfits
  - It’s a bad training set if it deviates too far from the specification (fidelity)
  - It’s a bad training set if it’s too small (significance), and it *might* be a bad training set if it’s too big (scale)
  - It’s a bad training set if it leaks private information (privacy-preserving)
  - It’s a bad training set if you can’t tell where it came from (provenance)
  - It’s a bad training set if it reinforces discrimination (bias-correcting)

So what do we do about it?

Fides: Responsible Data Management

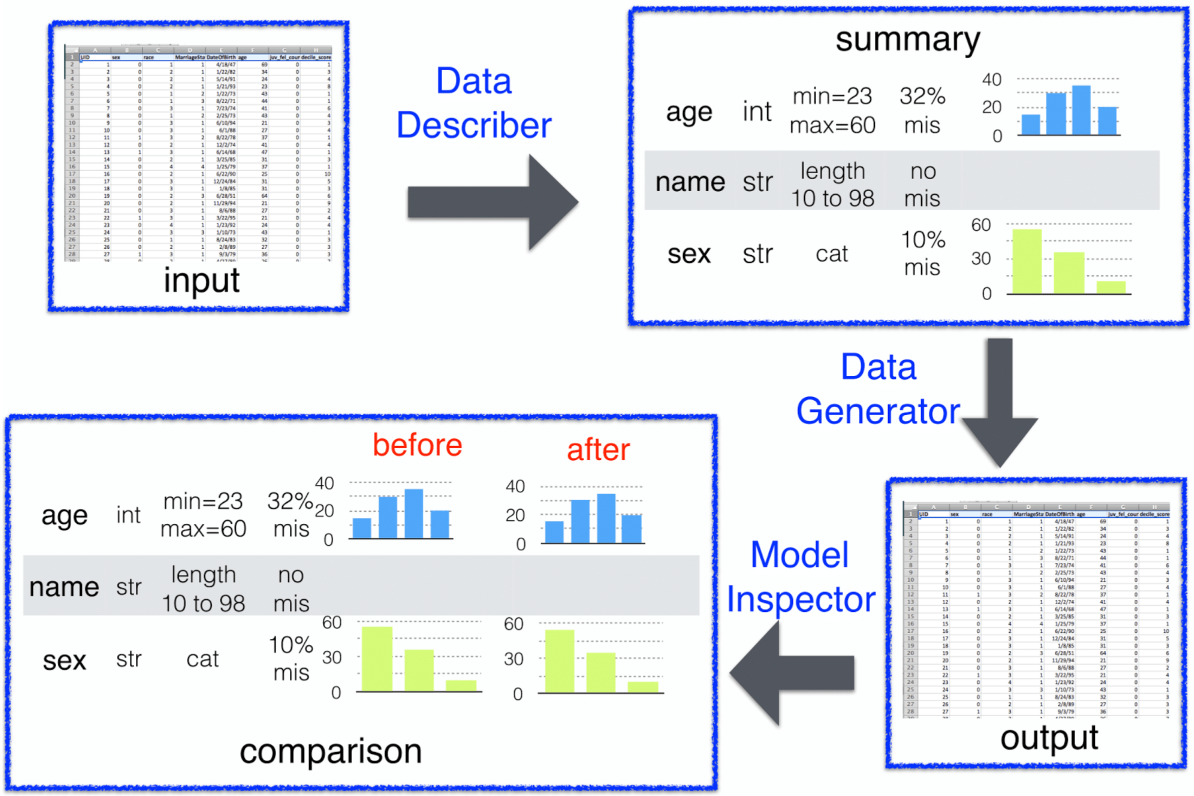
Fairness
Accountability
Transparency
Privacy
Reproducibility



joint with Stoyanovich [US], Abiteboul [FR], Miklau [US], Sahuguet [US], Weikum [DE]

UNIVERSITY of WASHINGTON

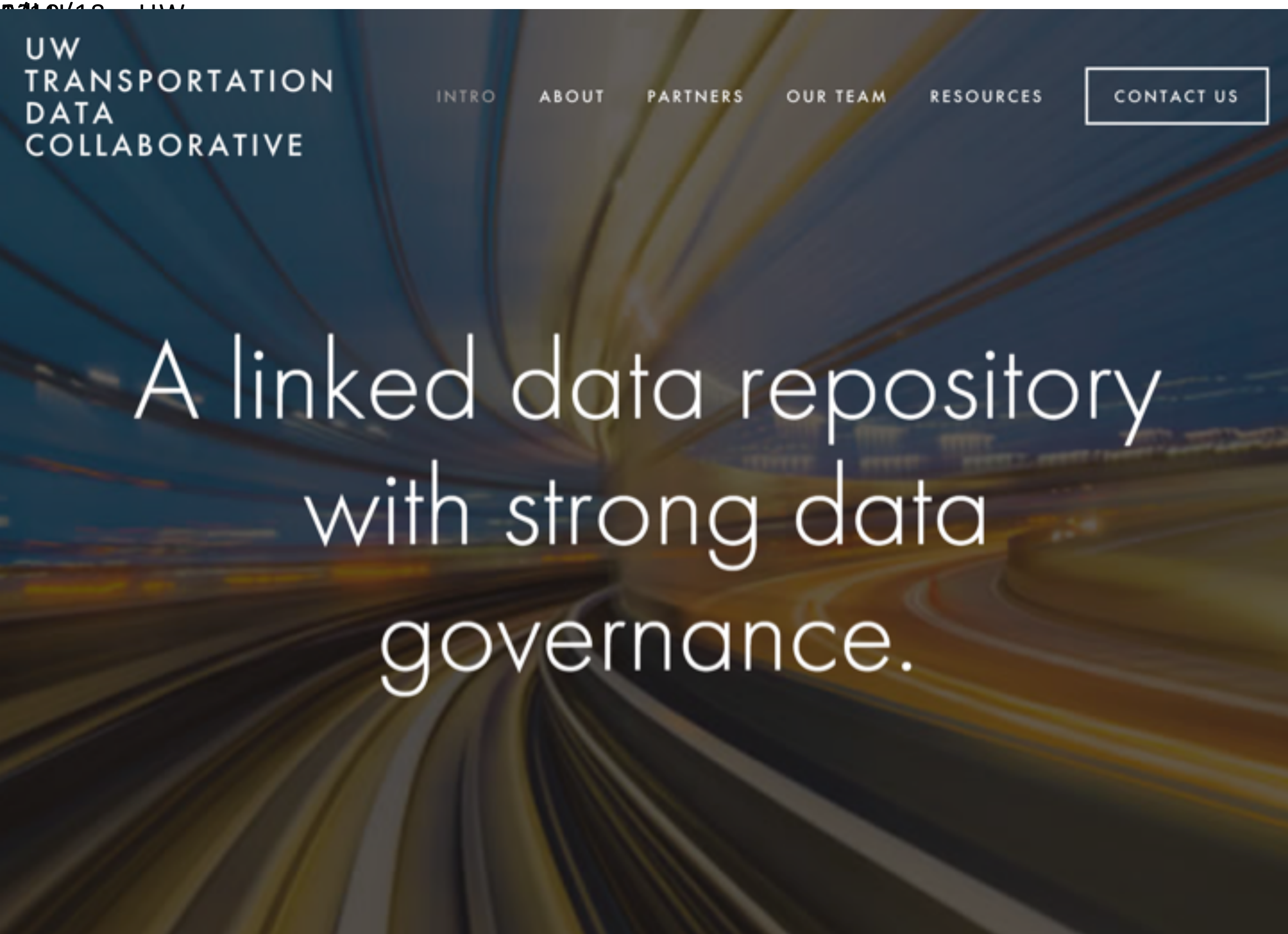
Data Synthesizer: Privacy-preserving synthetic data



With Stoyanovich (Drexel), Gee (Chicago), Ping (Drexel), Herman (UW)

A Nutritional Label for Rankings

[Yang, Stoyanovich, Asudeh, Howe, Jagadish, Miklau SIGMOD 2018 demo]


UW
TRANSPORTATION
DATA
COLLABORATIVE
INTRO
ABOUT
PARTNERS
OUR TEAM
RESOURCES
CONTACT US
A linked data repository with strong data governance.

BACKCHANNEL 06.14.18 07:00 AM

THE BIKE SHARE WAR IS SHAKING UP SEATTLE LIKE NOWHERE ELSE

Residents are divided over whether the city's dockless bike share program is revolutionizing transit—or creating an unwieldy, dangerous mess.

BY MARK HARRIS

Example: Bike Share data

- Companies need to release trip data to comply with Seattle permits and civic transparency
- But there are concerns about privacy, misuse, and competitive advantage
- Setup:

Trip(time, userid, company, orig, dest, gender, helmet)

aggregate

Domain info:
company in {Lime, Spin, Ofo}
origin, dest one of 94 neighborhoods in Seattle}
gender in {M, F, other, null}
Helmetuser in {true, false}

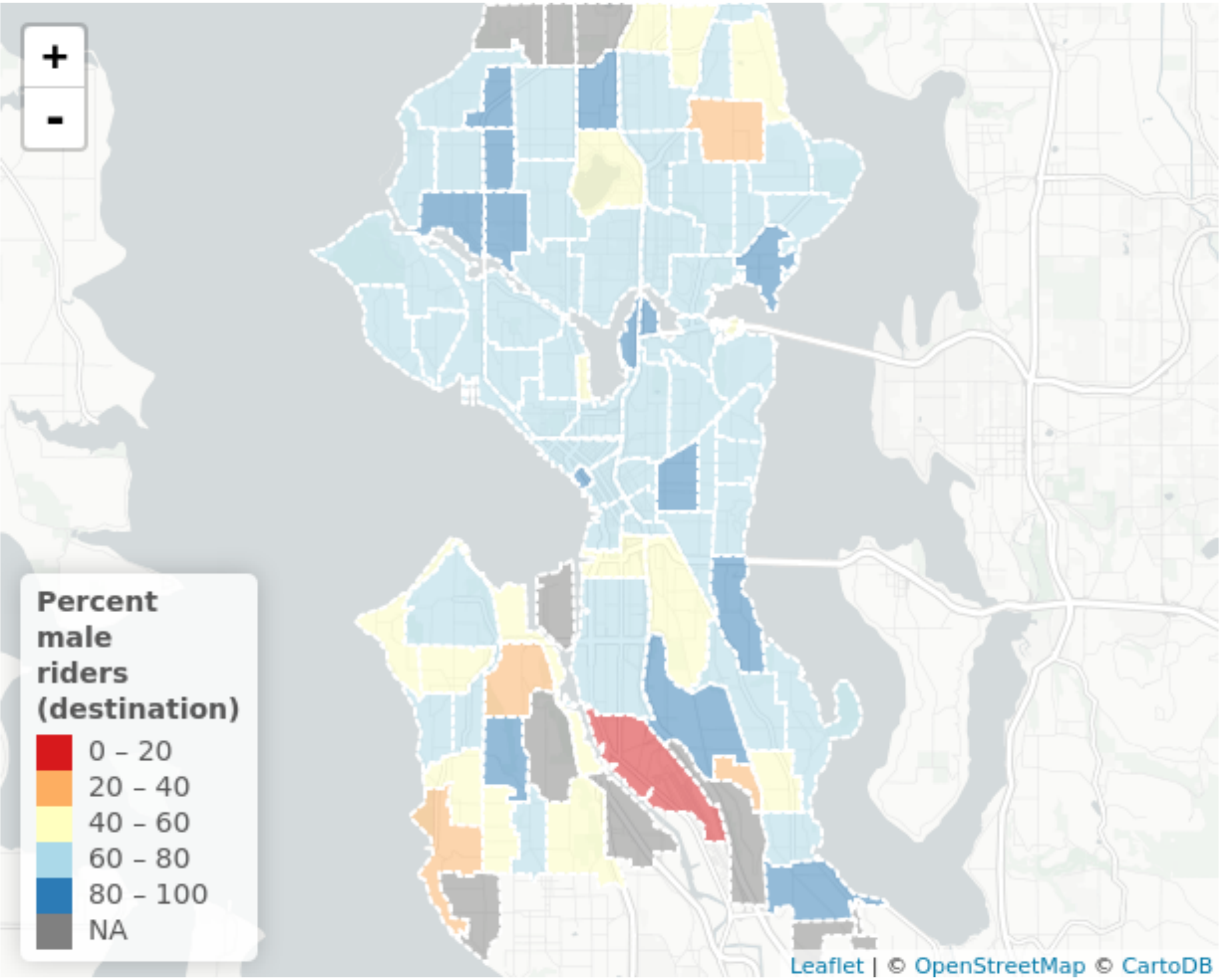
OD(company, origin, dest, gender, helmet, count)

We can release the joint distribution of company, origin, dest, gender, helmet

To release plots like this:

For privacy we can add noise to the counts.

But we also want to remove bias…



Bias-Corrected Data Sharing by Breaking Causal Relationships

Babak Salimi

Luke Rodriguez

Ex: We don’t want race to influence hiring, so set the mutual information between, say, race and GPA to zero before releasing the dataset.

Different strategies:

- You could remove and insert tuples
- You could directly edit the GPA
- You could change the “weight” of each tuple

UNIVERSITY of WASHINGTON

Bias-Corrected Data Sharing by Breaking Causal Relationships

Babak Salimi

Luke Rodriguez

Back to bike share: Hiding competitive advantage

Company doesn’t mind releasing data, but doesn’t want to reveal that they have a marketing campaign targeting women.

Set the mutual information between company (X) and gender (Y) to zero, conditioned on the other attributes (Z).

Compute a new joint distribution of trips, asserting independence between X and Y conditioned on Z

$$P_{R'}(\mathbf{A}) = P_{R'}(X\mathbf{Z})P_{R'}(Y|\mathbf{Z})P_{R'}(\mathbf{U}|XY\mathbf{Z})$$

FAT* 2019 Call for Papers

Important Dates and Links

Submission site	TBD
Abstract Deadline	pre-registration at 11:59PM August 16, 2018 AoE
Full paper submission	11:59PM August 23, 2018 AoE
Notification Date	October 12, 2018
Conference Date	late January/early February 2019

FAT* is an international and interdisciplinary peer-reviewed conference that seeks to publish and present work examining the **fairness, accountability, and transparency of algorithmic systems.**

Topics of Interest

The FAT* conference solicits work from a wide variety of disciplines, including computer science, statistics, the humanities, and law. FAT* welcomes submissions that touch on any of the following topics (broadly construed):

- Fairness
  - Techniques and models for fairness-aware data mining, information retrieval, recommendation, etc.
  - Formalizations of fairness, bias, discrimination; trade-offs and relationships between them
  - Defining, measuring and mitigating biases in data sets; improving data collection processes; combining different sources of information
  - Translation of legal, social, and philosophical models of fairness into mathematical objectives
  - Qualitative, quantitative, and experimental studies on perceptions of algorithmic bias and unfairness
  - Design interventions to mitigate biases in systems, or discourage biased behavior from users
  - Measurement and data collection regarding potential unfairness in systems
  - Understanding how tools from causal inference can help us to better reason about fairness and the interplay between prediction and intervention
  - Analyses of the impact of algorithmic experimentation and exploration