Privacy and Synthetic Data

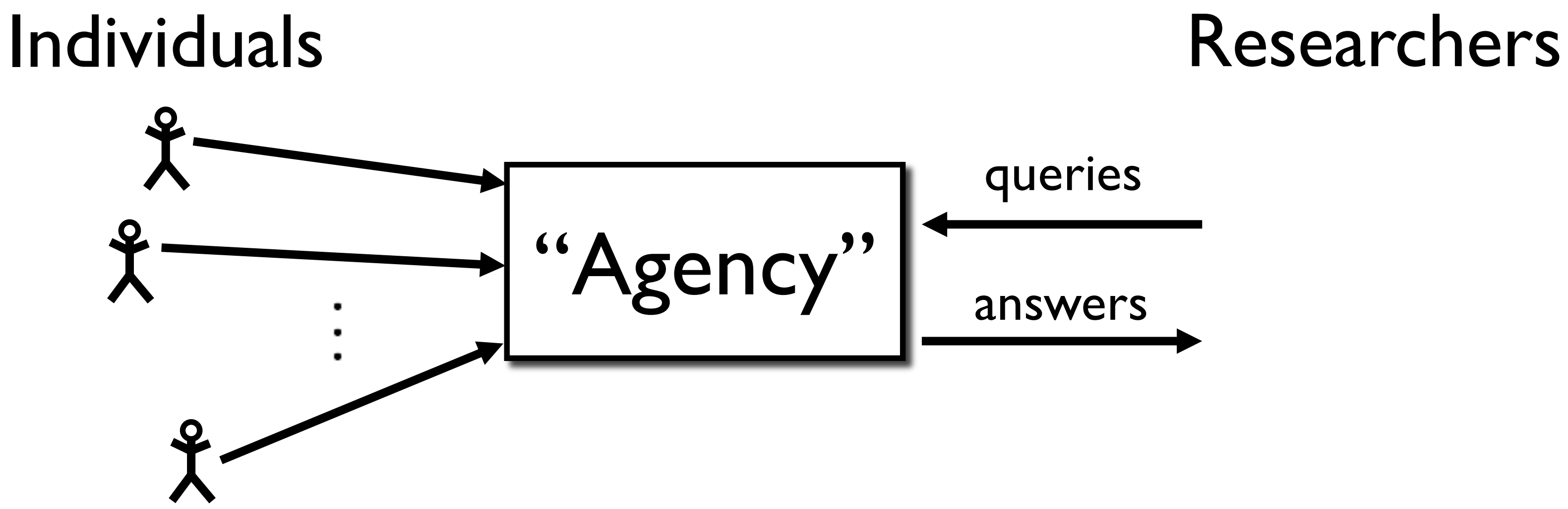
BOSTON UNIVERSITY

Adam Smith

BU Computer Science

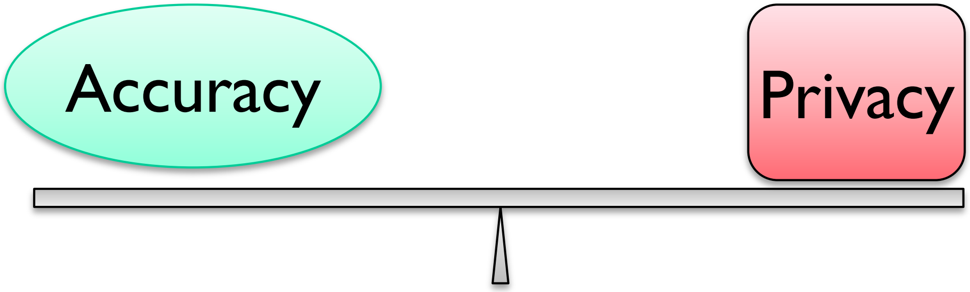
January 18, 2023

Privacy in Statistical Databases

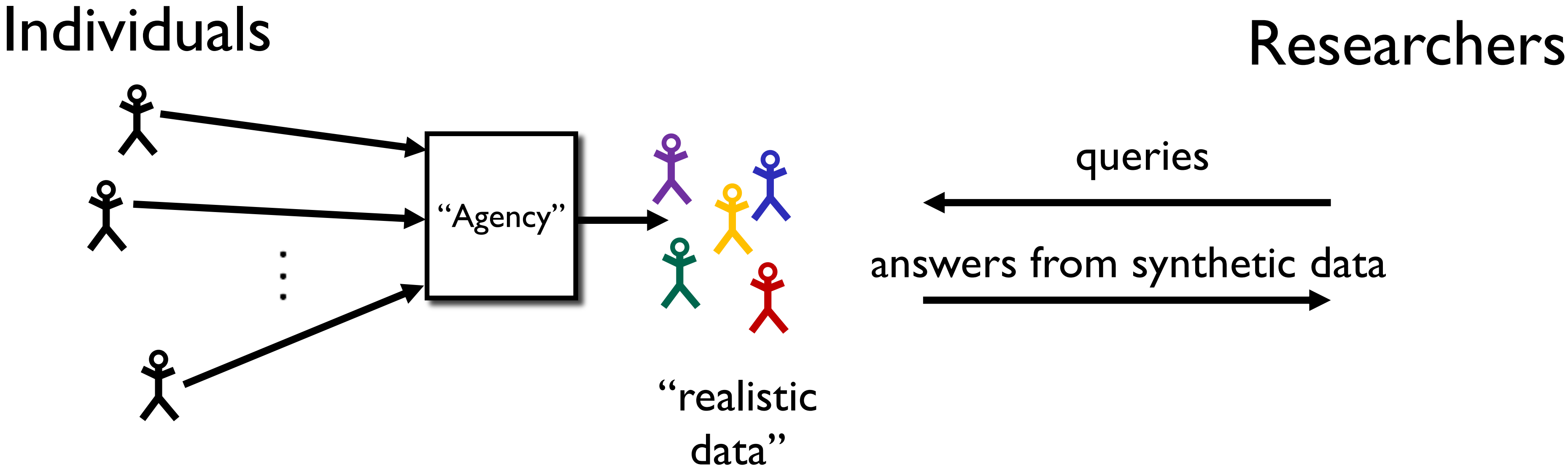


Large collections of personal information

- census data
- medical/public health
- social networks
- education
- system usage



Privacy in Synthetic Data



Individuals

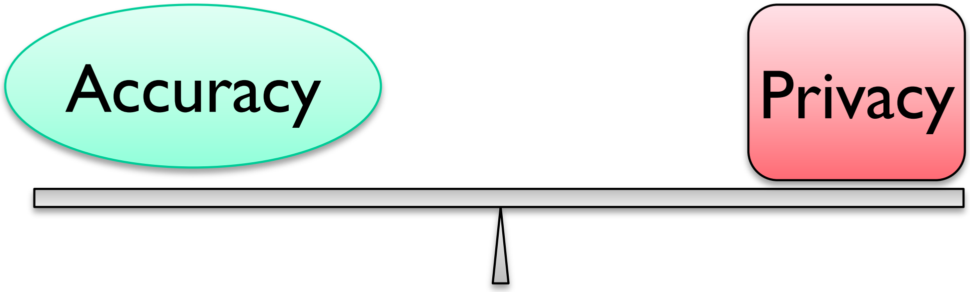
Researchers

"Agency"

queries

answers from synthetic data

"realistic data"



Accuracy

Privacy

- Problem: What is "accuracy" here?
- Ideally, we want data that works for all queries.

First attempt: Remove obvious identifiers

Everything is an identifier

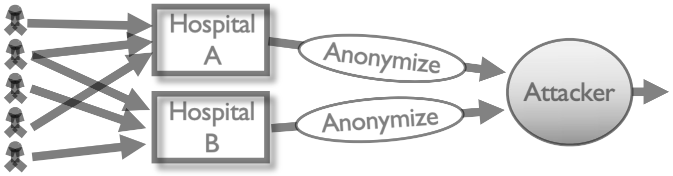
"AI recognizes blurred faces"
[McPherson Shokri Shmatikov '16]

[Gymrek McGuire Golan Halperin Erlich '13]

[Pandurangan '14]

On Taxis and Rainbows

Lessons from NYC's improperly anonymized taxi logs



[Ganta Kasiviswanathan S '08]

Images: whitehouse.gov, genesandhealth.org, medium.com

Is the problem granularity?

What if we only release aggregate information?

Statistics together may encode data

- Example: Average salary before/after resignation
- More generally:

Too many, "too accurate" statistics
reveal individual information

- ➢ Reconstruction attacks [Dinur Nissim 2003, …]
- ➢ Membership attacks [Homer et al, 2008, …]
- ➢ Memorization [Carlini et al. '20, Brown et al. '21, …]

Cannot release everything
everyone would want to know

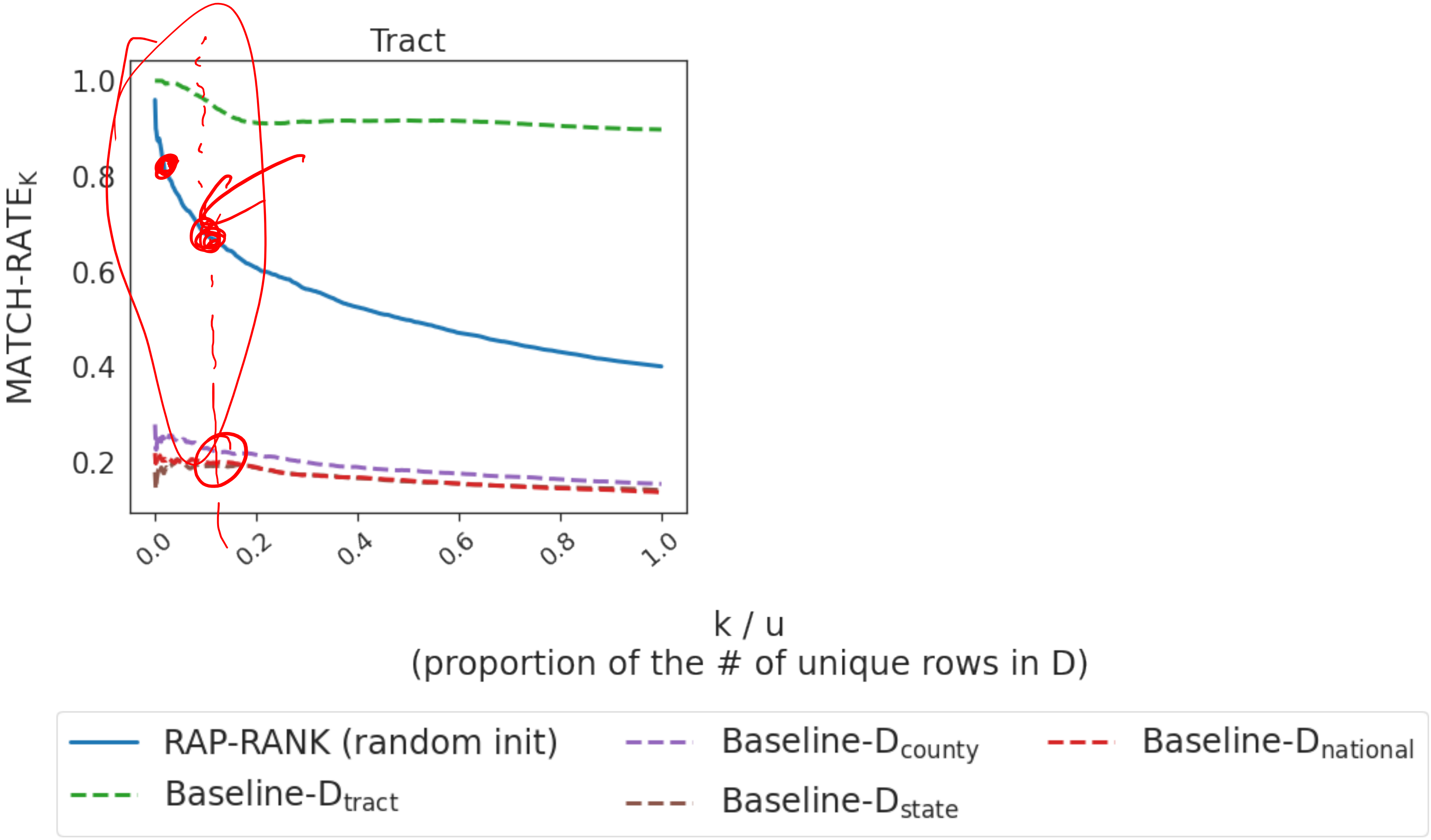
Reconstruction from Census Data

[Dick, Dwork, Kearns, Liu, Roth, Vietri, Wu. arxiv:221103128, 2022]

- Raw data: 2020-05-27 Privacy Protected Microdata File
  - ➢ (Generated to imitate 2010 Census microdata)
- Compute basic demographics (“Census SF 1”)
- Methodology
  - ➢ Find data sets consistent with demographics
  - ➢ Rank records according to how often they are reconstructed
  - ➢ Compare match rate to baseline sample from same data set

Reconstruction from Census Data

[Dick, Dwork, Kearns, Liu, Roth, Vietri, Wu. arxiv:221103128, 2022]

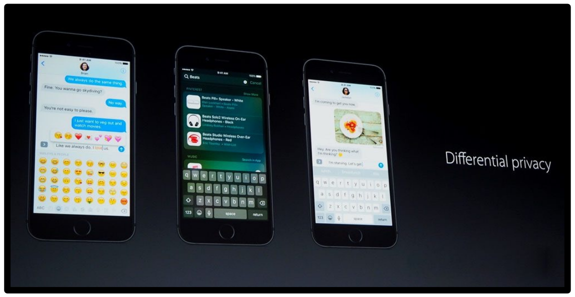


This talk

- Synthetic data
- Why is privacy challenging?
- Differential privacy recap
- How DP synthetic data algorithms (generally) work
- What types of statistics need to be preserved?

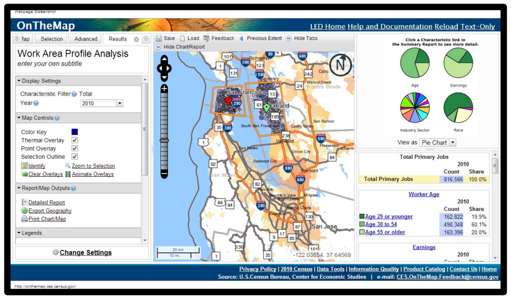
Differential Privacy [Dwork, McSherry, Nissim, S., 2006]

- Many current deployments



Apple

Google


Webpage Screenshot
OnTheMap
LED Home Help and Documentation Reload Text-Only
Map Selection Advanced Results
Save Load Feedback Previous Extent Hide Tabs
Hide Chart/Report
Click a Characteristic link in the Summary Report to see more detail.
Age
Earnings
Industry Sector
Race
View as Pie Chart
San Francisco
Oakland
San Jose
-122.03854, 37.64569

Work Area Profile Analysis
enter your own subtitle

Display Settings

Characteristic Filter: Total
Year: 2010

Map Controls

Color Key ■
Thermal Overlay ☑
Point Overlay ☑
Selection Outline ☑

Identify, Zoom to Selection, Clear Overlays, Animate Overlays

Report/Map Outputs

Detailed Report
Export Geography
Print Chart/Map

Legends

Change Settings

Total Primary Jobs

	2010 Count	2010 Share
Total Primary Jobs	816,566	100.0%

Worker Age

	2010 Count	2010 Share
Age 29 or younger	162,822	19.9%
Age 30 to 54	490,348	60.1%
Age 55 or older	163,396	20.0%

Earnings

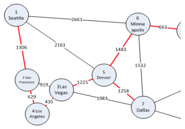
2010

Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home
Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

http://onthemap.ces.census.gov/

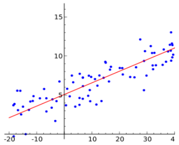
US Census

- Burgeoning field of research



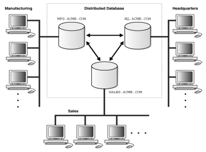
Algorithms

Crypto, security



Statistics, learning

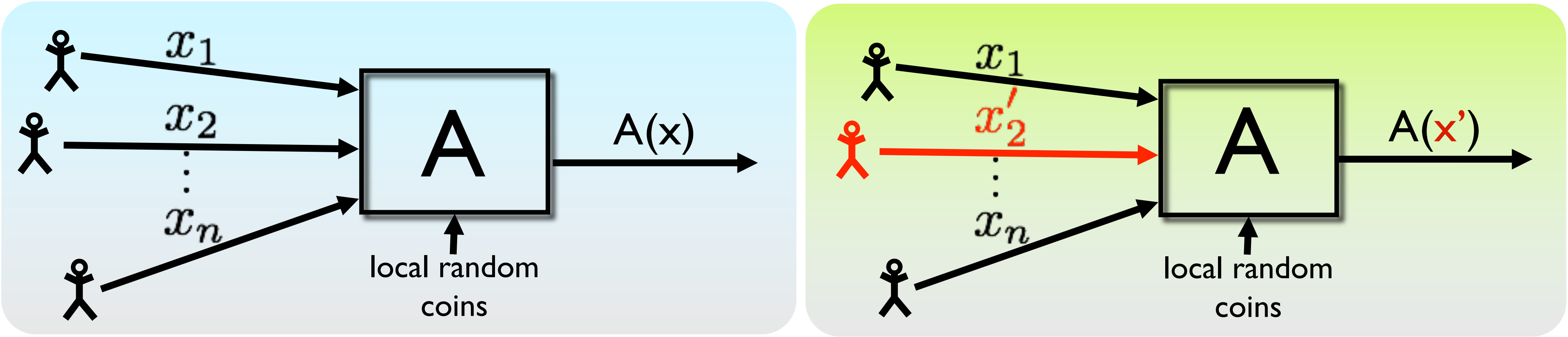
Game theory, economics



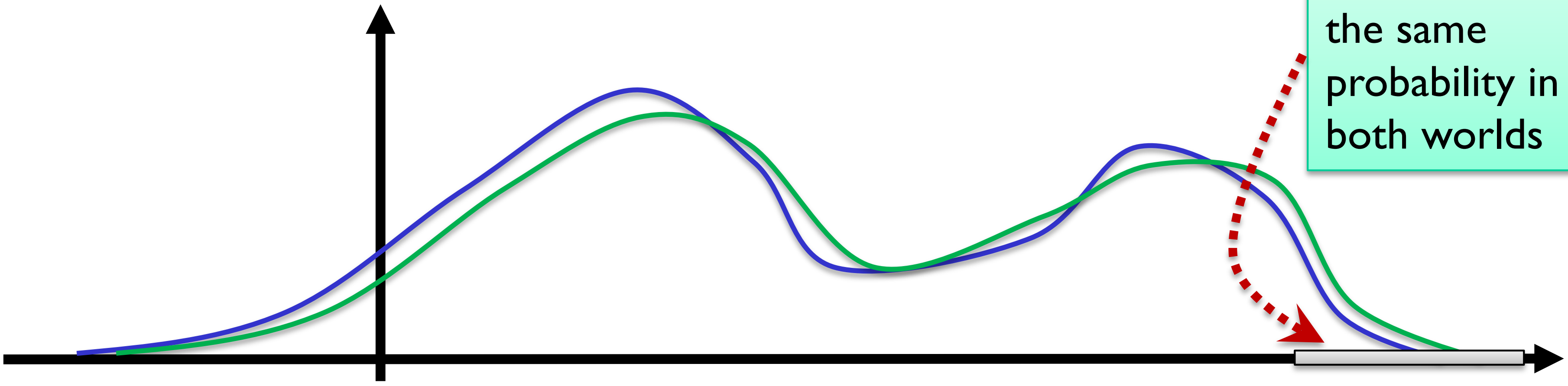
Databases, programming languages

Law, policy

Differential Privacy [Dwork, McSherry, Nissim, S., 2006]

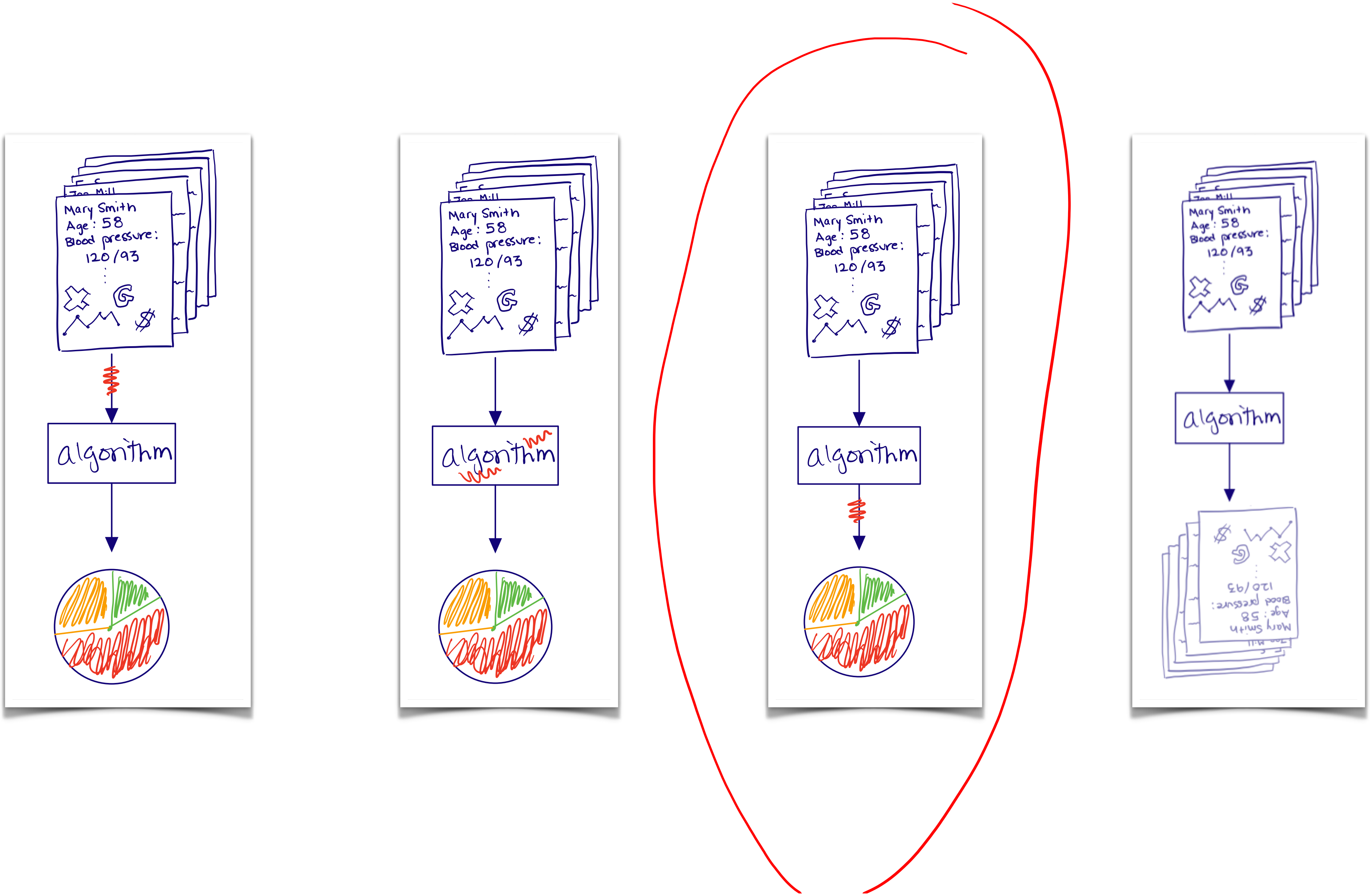


- A thought experiment
  - Change one person's data (or add or remove them)
  - Will the **distribution of outputs** change much?



For any set of outcomes, about the same probability in both worlds

How to achieve DP?



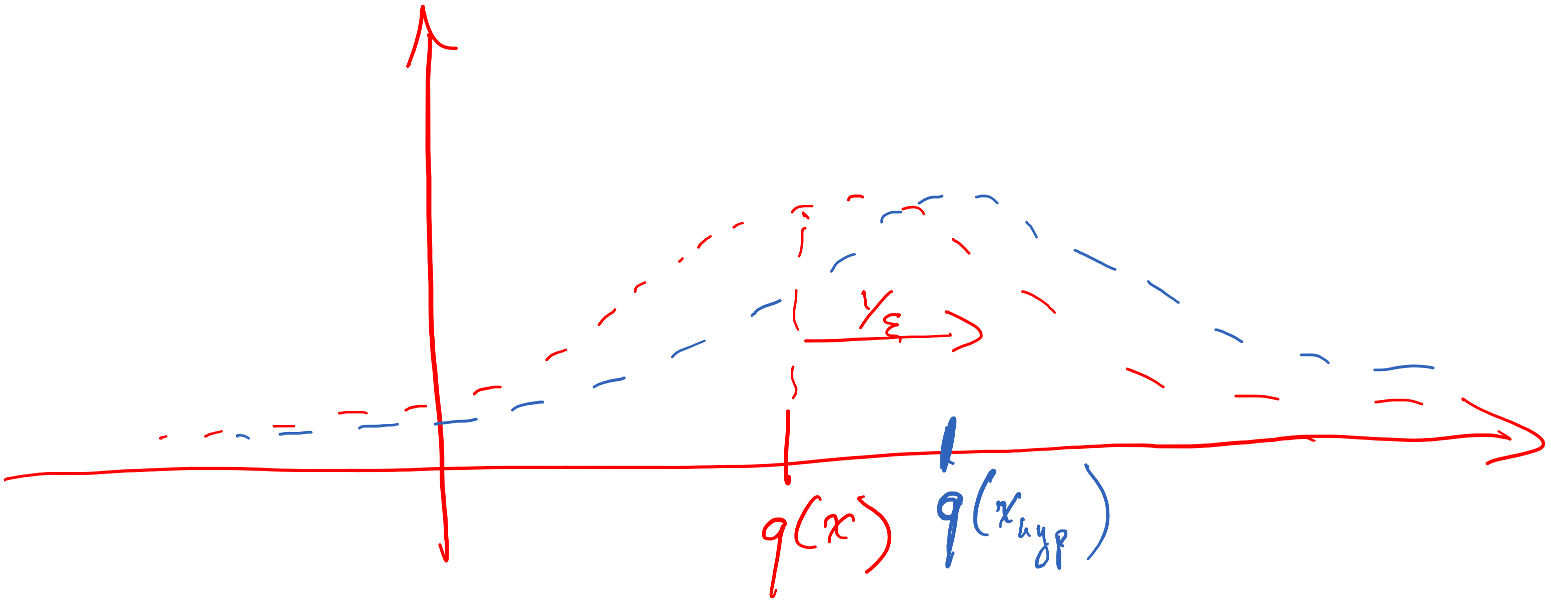
Adding noise to 1 count

- Suppose we want to release

$$q(x) = \text{\# diabetics in data set } x$$

- Parameter $\boxed{\varepsilon}$ measures how much is leaked
- One approach: add Gaussian noise* roughly $\frac{1}{\varepsilon}$

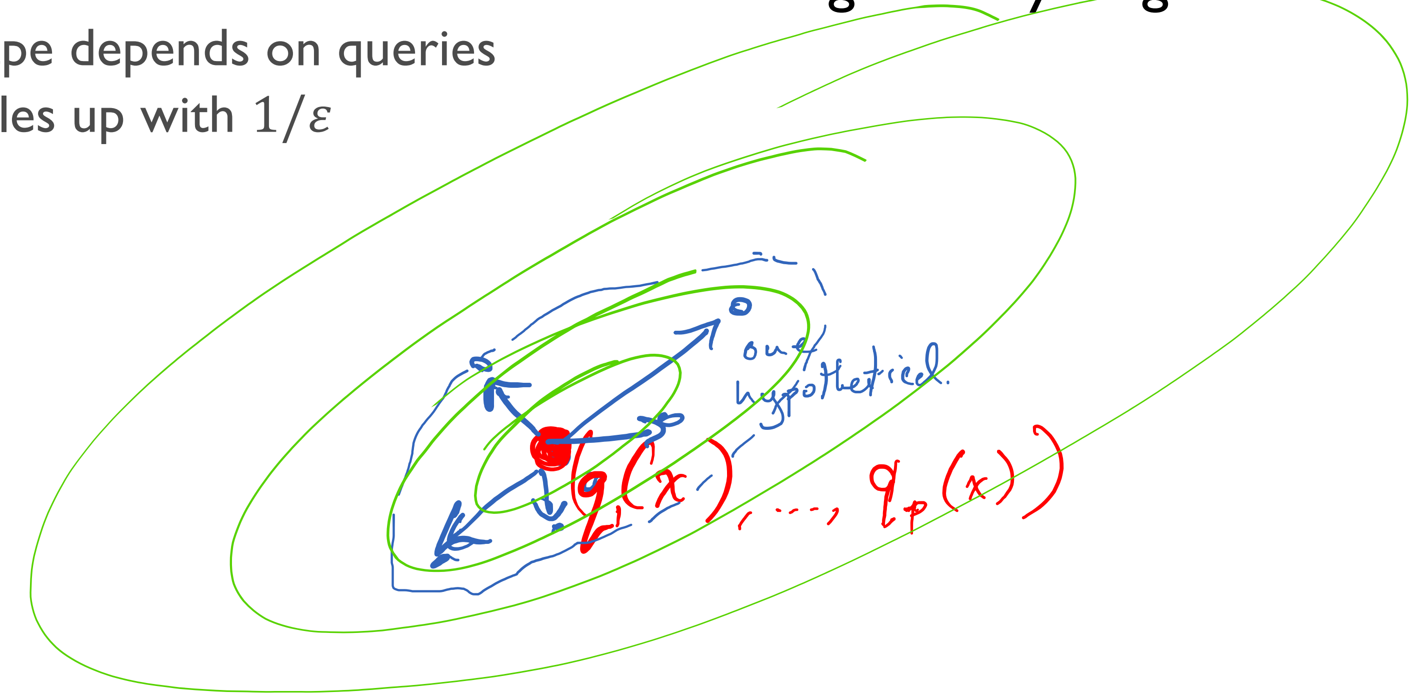
$$A(x) = \boxed{q(x)} + \underline{N(0,\sigma^2)} \quad for \quad \boxed{\sigma \approx 1/\varepsilon}$$



* Provides "concentrated differential privacy" [Dwork-Rothblum, Bun-Steinke]

Adding noise to many counts

- Now suppose we have a **vector** of statistics
  $q_1(x) = diabetics$
  $q_2(x) = \#people\ over\ 80$
  $\vdots$
- Add correlated noise to mask the change of any single record
  - Shape depends on queries
  - Scales up with $1/\varepsilon$



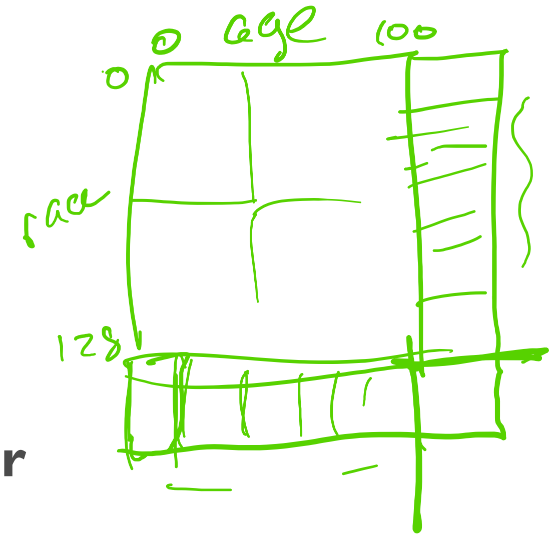
- Tradeoff: complexity vs accuracy

This talk

- Synthetic data
- Why is privacy challenging?
- Differential privacy recap
- How DP synthetic data algorithms (generally) work
- What types of statistics need to be preserved?

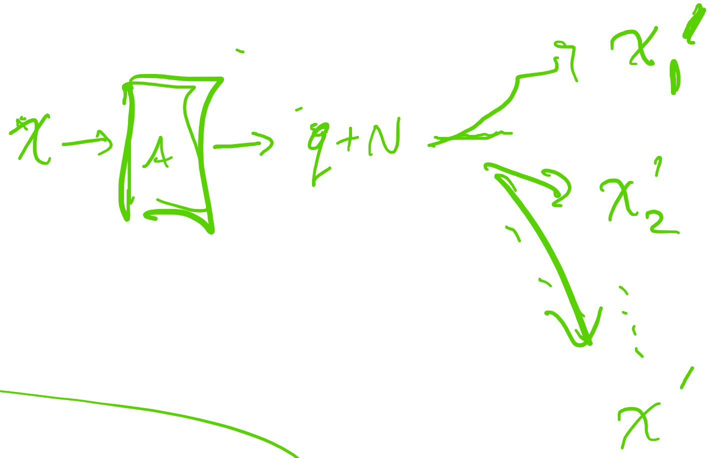
Generic Template: Measure and Fit

- Measure:
  - ➢ Calculate set of predetermined statistics
- Add noise:
  - ➢ Release noisy measurements
- Generate data: **either**
  - ➢ Find data set consistent with noisy measurements, **or**
  - ➢ Sample from distribution fit to noisy measurements



Generic Template: Measure and Fit

- Measure:
  - Calculate set of predetermined statistics
- Add noise:
  - Release noisy measurements
- Generate data: **either**
  - Find data set consistent with noisy measurements, **or**
  - Sample from distribution fit to noisy measurements



Challenges for current research

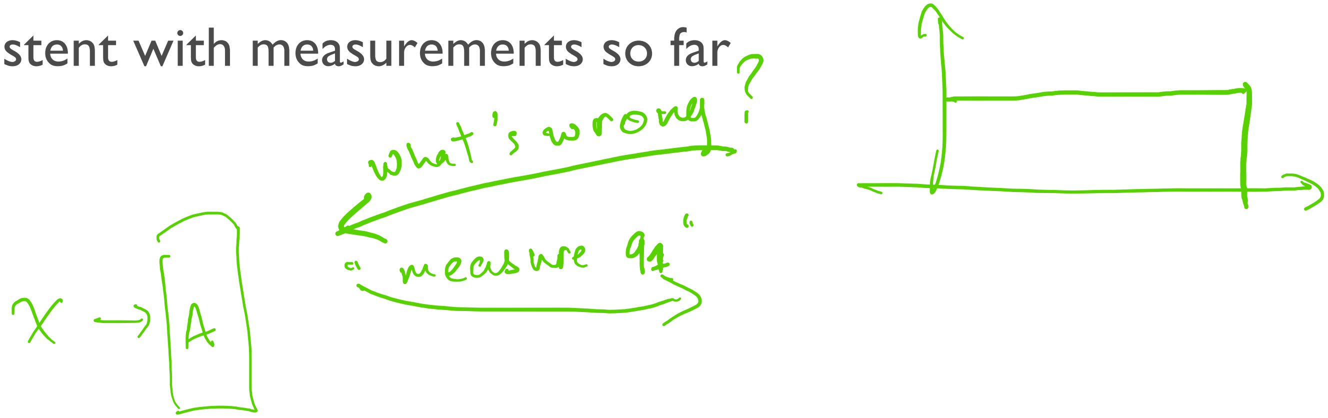
- Which statistics to measure?
- Computation: finding consistent data set
- Inference: Principled uncertainty estimates

Discriminative template

Fix a set of statistics to preserve

Start with an initial random proposed data set and repeat:

- Search:
  - Find a statistic distinguishes the proposed and real data sets
- Add noise:
  - Release noisy measurements of that statistic
- Update proposed data set
  - To make consistent with measurements so far



This talk

- Synthetic data
- Why is privacy challenging?
- Differential privacy recap
- How DP synthetic data algorithms (generally) work
- What types of statistics need to be preserved?

Private Synthetic Data Requires Choices

- What (minimal) set of analyses should be supported?

- Validation on real data
  - ➢ Synthetic data are problematic statistically
  - ➢ Ensuring some available validation is crucial

This talk

- Synthetic data
- Why is privacy challenging?
- Differential privacy recap
- How DP synthetic data algorithms (generally) work
- What types of statistics need to be preserved?