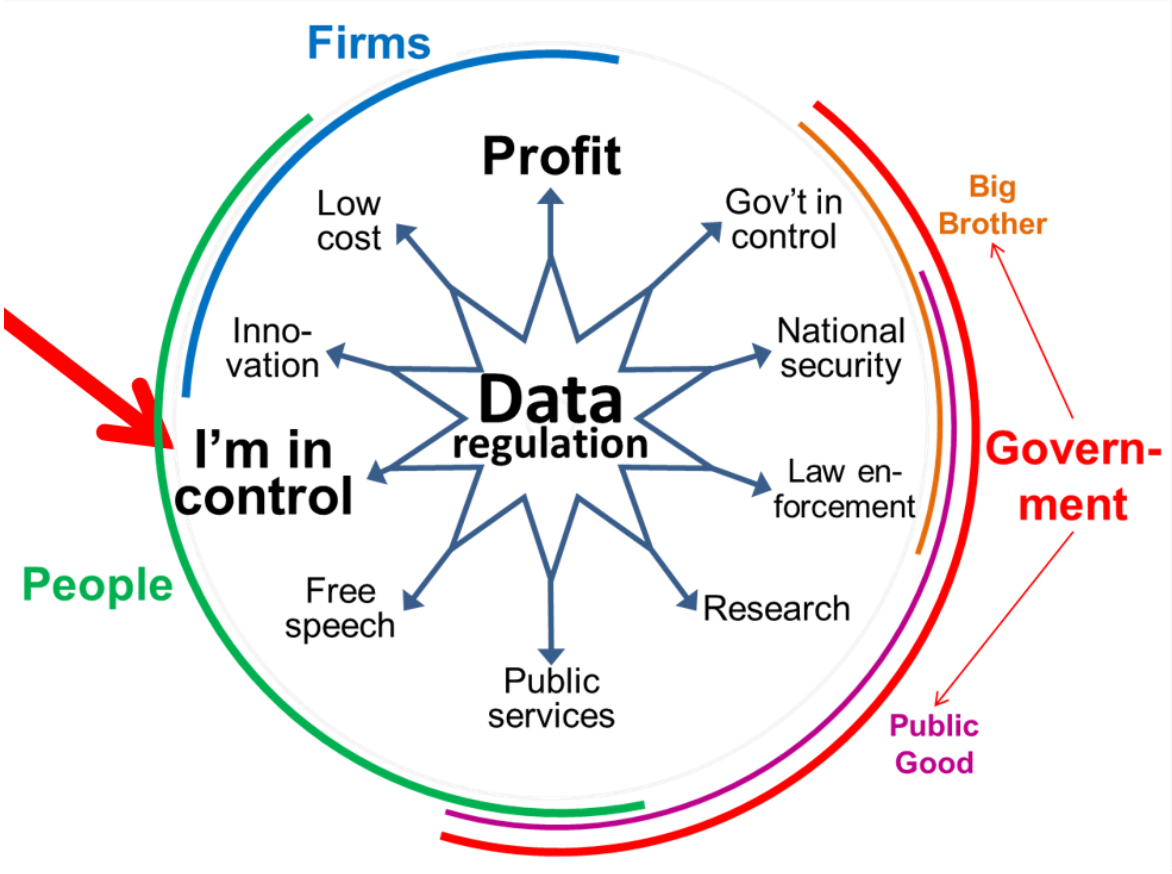
Differential Privacy and the Right to be Forgotten

Cynthia Dwork, Microsoft Research

Limiting Prospective Use

- Lampson's approach empowers me to limit the use of my data, prospectively



Limiting Future Use: Raw Data

Get me off Your [redacted]ing Mailing List

David Mazières and Eddie Kohler
New York University
University of California, Los Angeles
http://www.mailavenger.org/

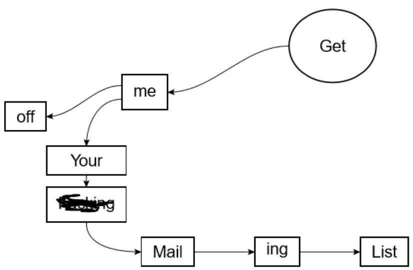


Figure 1: Get me off your [redacted]g mailing list.

Abstract

Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]g mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ng mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]g mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]ing mailing list. Get me off your [redacted]g mailing list. Get me off your [redacted]ing mailing list. Get me off your

Limiting Future Use: Raw Data

Use of blood sample data

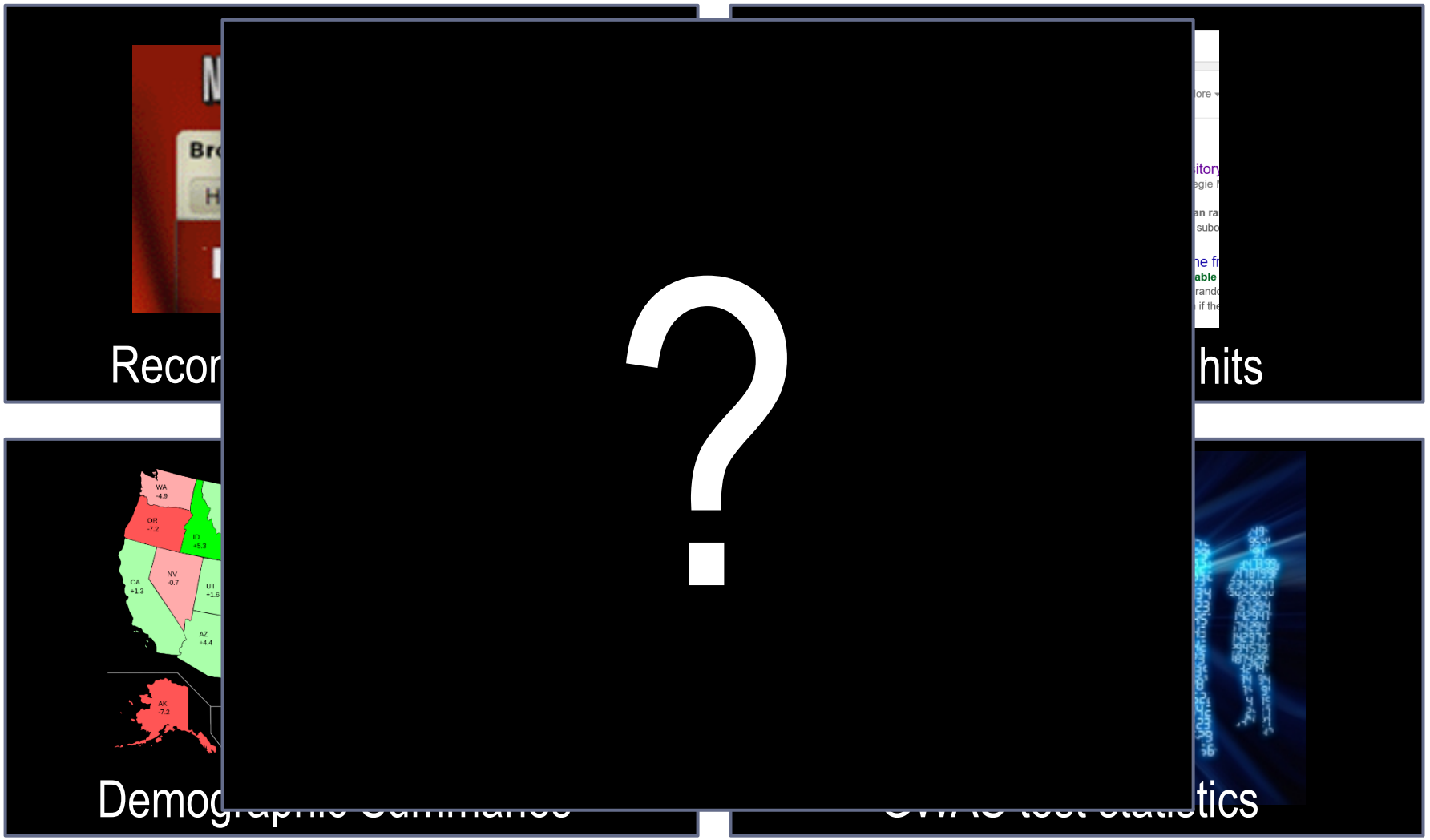
LinkedIn

Showing my data to subscribers

Google

Reporting my past

Limiting Future Use: Entangled Data



Re-Compute Without Me?

- Expensive; Great vector for denial of service attack
- Privacy compromise

Sickle cell trait: 33

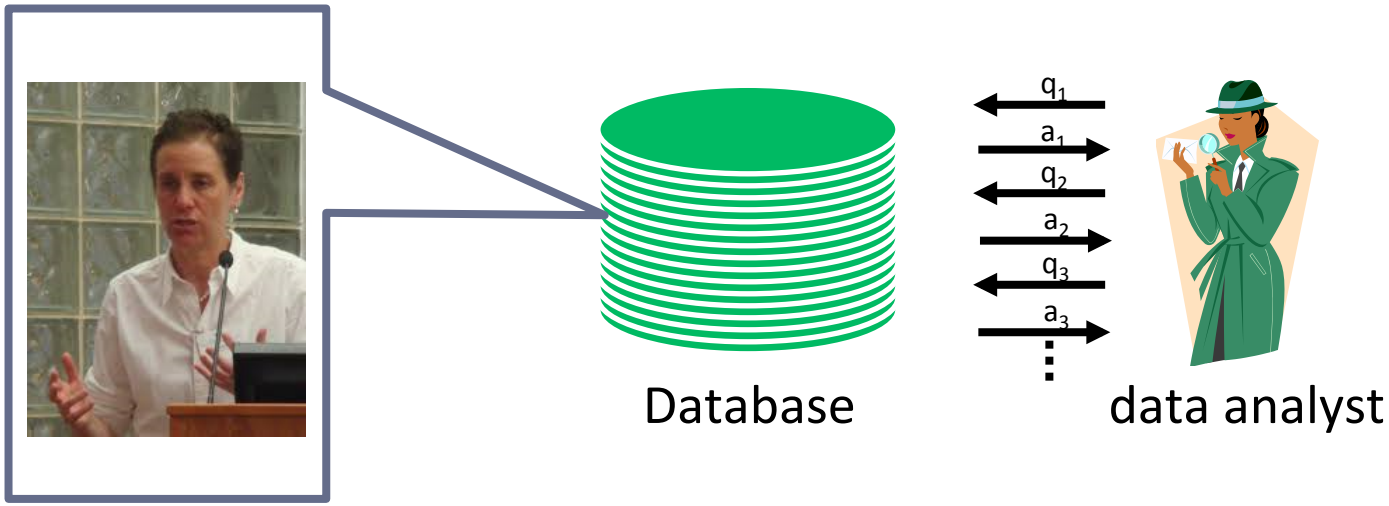
Sickle cell trait: 32

Differential Privacy as a Solution Concept

- Definition of privacy tailored to statistical analysis of big data
- “Nearly equivalent” to not having had one’s data used at all
- Safeguards privacy even under re-computation

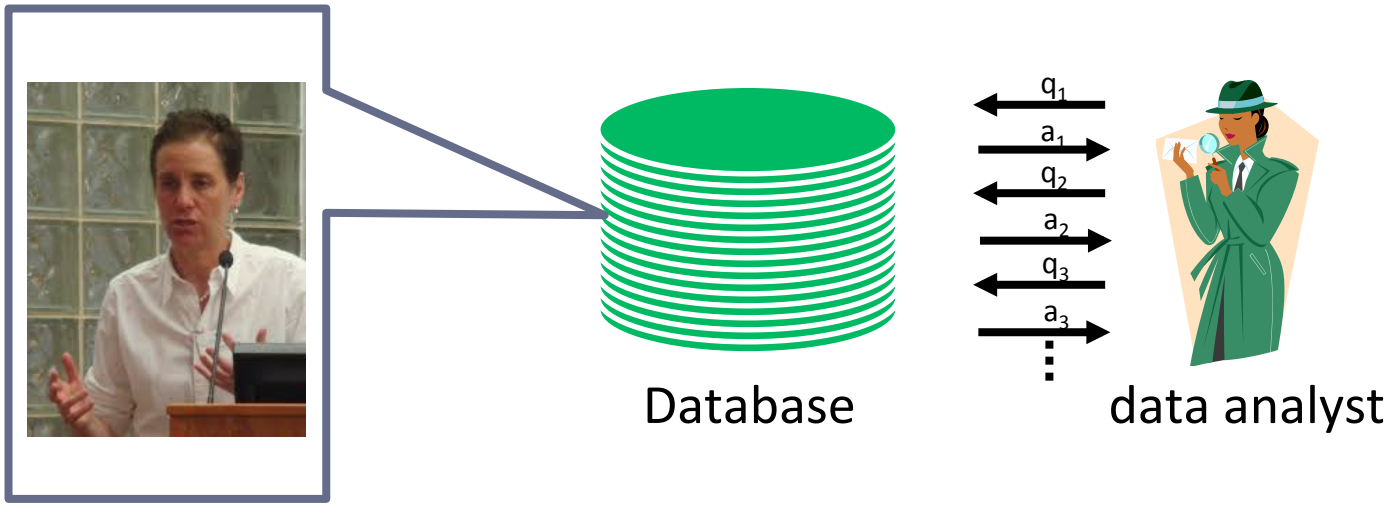
Dwork, McSherry, Nissim, and Smith 2006

Privacy-Preserving Data Analysis?



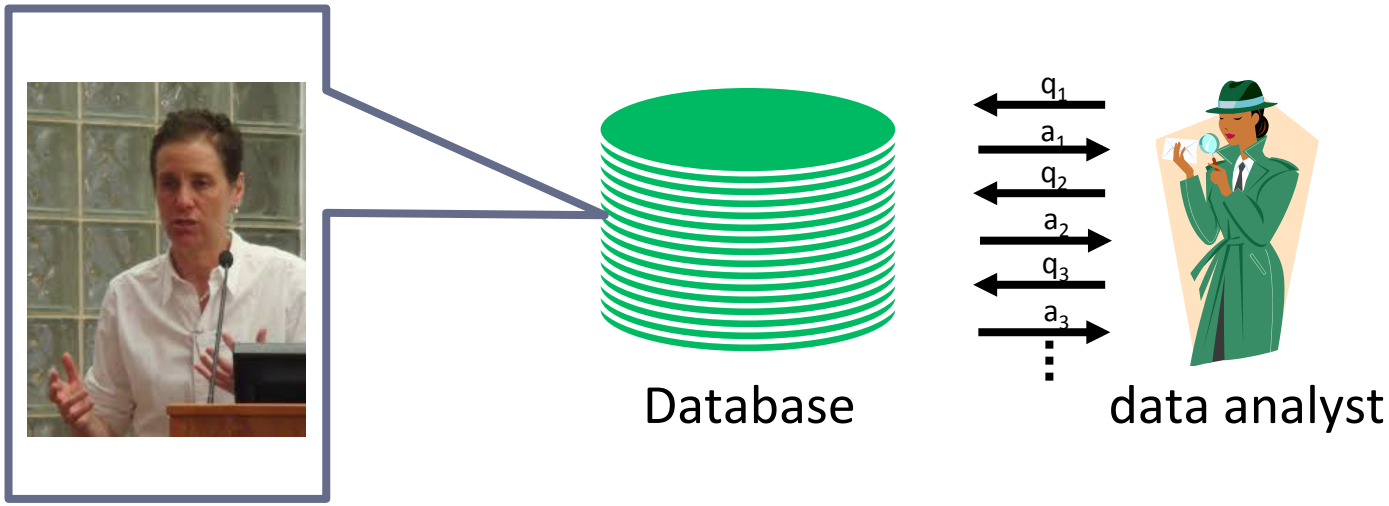
- "Can't learn anything new about Nissenbaum"?

Privacy-Preserving Data Analysis?



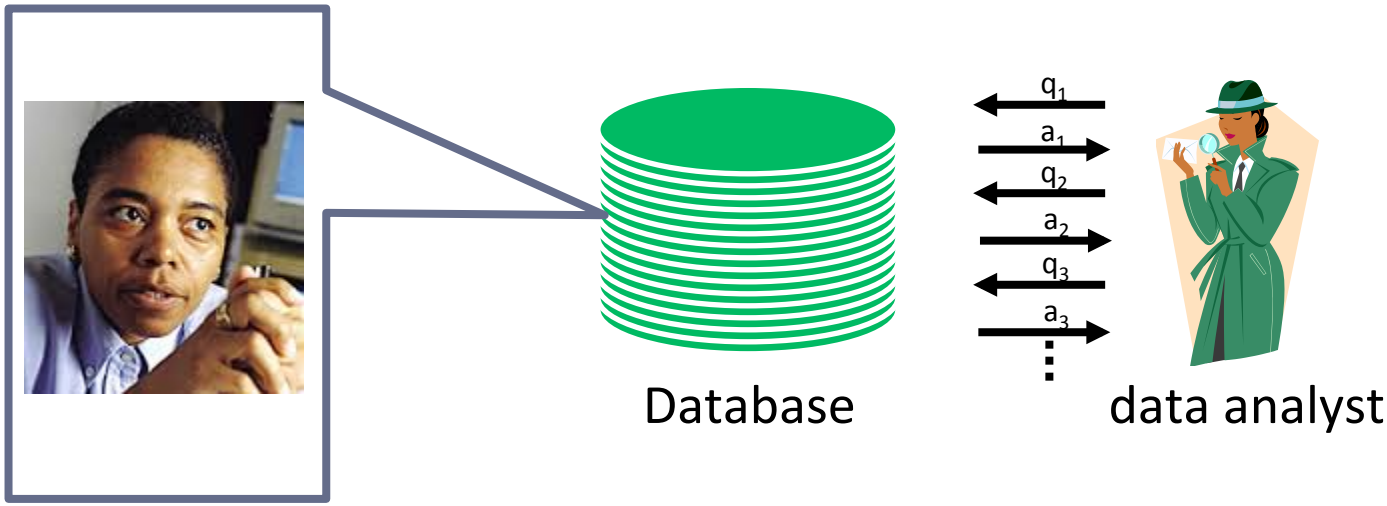
- "Can't learn anything new about Nissenbaum"?
- Then what is the point?

Privacy-Preserving Data Analysis?



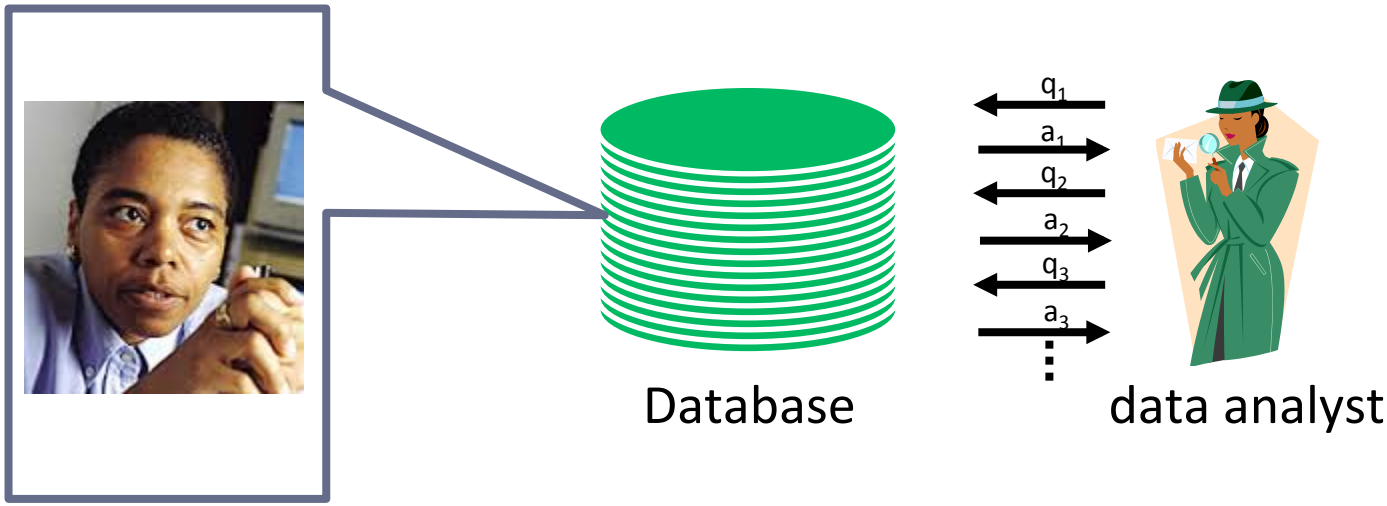
- "Can't learn anything new about Nissenbaum"?
- Then what is the point?

Privacy-Preserving Data Analysis?



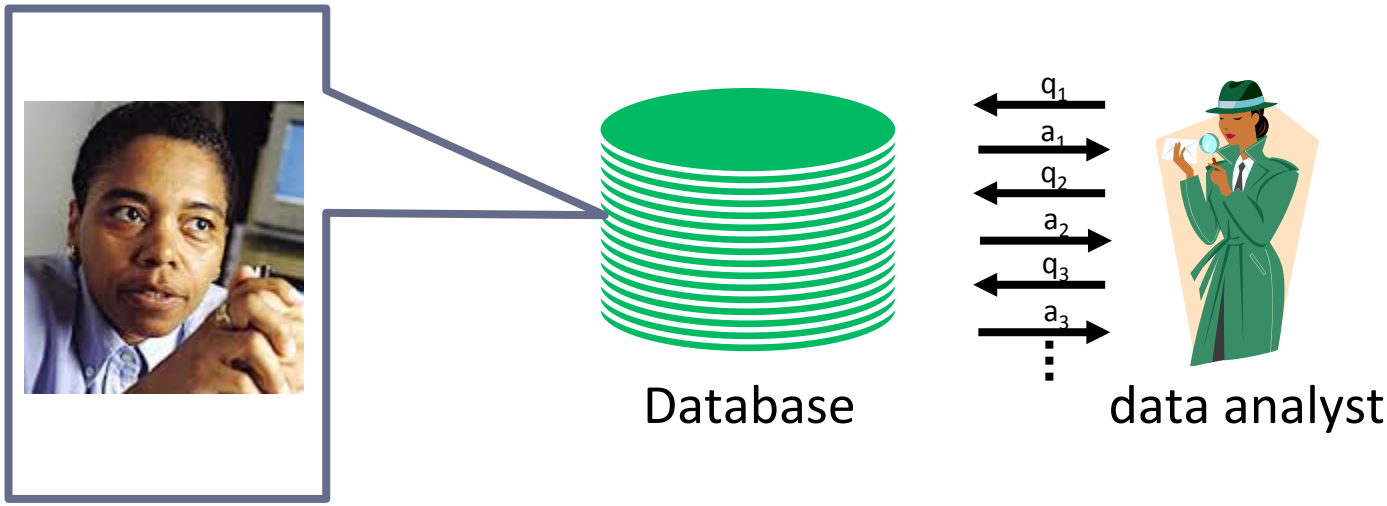
- Ideally: learn same things if Nissenbaum is replaced by another random member of the population

Privacy-Preserving Data Analysis?



- Ideally: learn same things if Nissenbaum is replaced by another random member of the population ("stability")

Privacy-Preserving Data Analysis?



- Stability preserves Nissenbaum's privacy AND prevents over-fitting
- Privacy and Generalization are aligned!

Differential Privacy

- The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset.
  - Nissenbaum's data are deleted, Sweeney's data are added, Nissenbaum's data are replaced by Sweeney's data, etc.
  - "Nearly equivalent" to not having data used in the first place

Formally

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all subsets S of possible outputs

$$\Pr[M(x) \in S] \leq (1+\epsilon)\Pr[M(y) \in S]$$

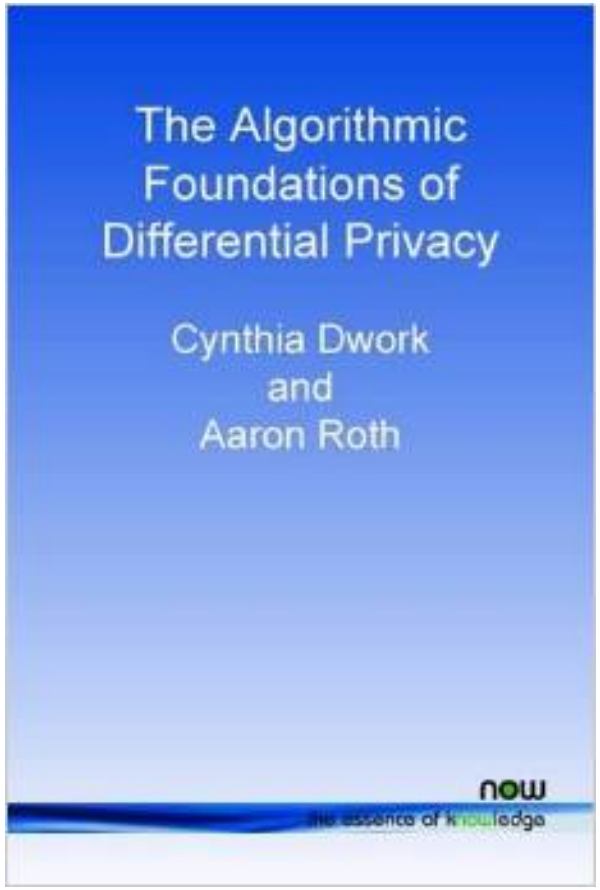
Randomness introduced by M

Properties

- Immune to current and future(!) side information
- Automatically yields group privacy
- Understand behavior under composition
  - Can bound cumulative privacy loss over multiple analyses
    - Permits “re-computation” when data are withdrawn
- Programmable
  - Complicated private analyses from simple private building blocks

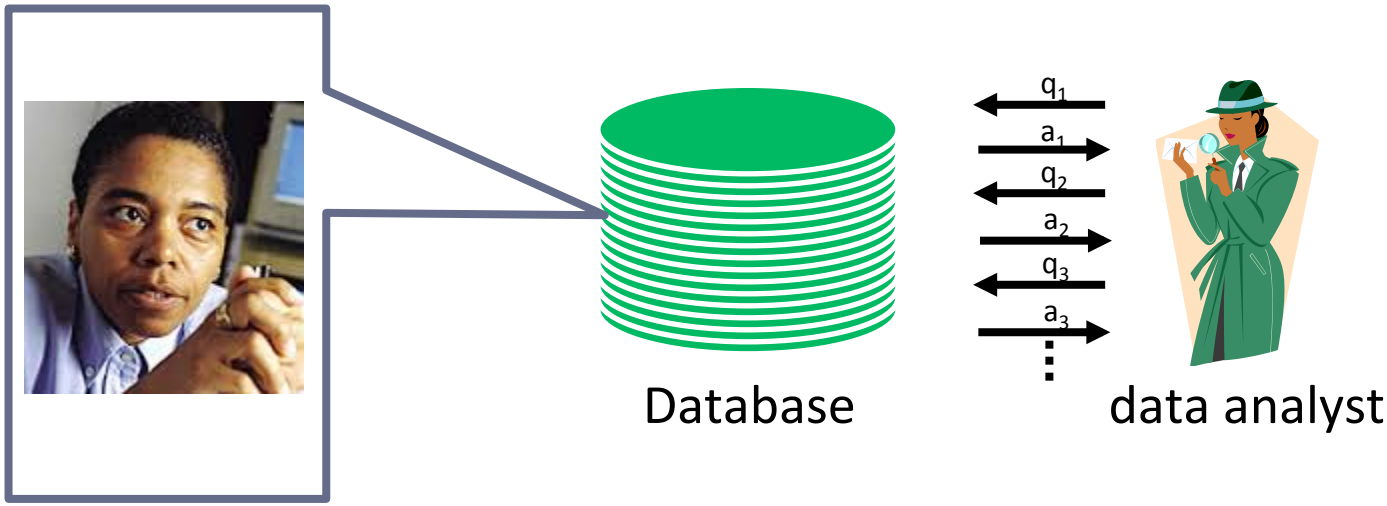
Rich Algorithmic Literature

- Counts, linear queries, histograms, contingency tables (marginals)
- Location and spread (eg, median, interquartile range)
- Dimension reduction (PCA, SVD), clustering
- Support Vector Machines
- Sparse regression/LASSO, logistic and linear regression
- Gradient descent
- Boosting, Multiplicative Weights
- Combinatorial optimization, mechanism design
- Privacy Under Continual Observation, Pan-Privacy
- Kalman filtering
- Statistical Queries learning model, PAC learning
- False Discovery Rate control
- Pan-Privacy, privacy under continual observation …



Which is “Right”?

Which is “Right”?



- Stability preserves Nissenbaum’s privacy AND prevents over-fitting
- Differential privacy protects against false discovery / overfitting due to adaptivity (aka exploratory data analysis)

Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth 2014

Not a Panacea

Fundamental law of information recovery

ϵ: a nexus of policy and technology

[DN03,DMT07,HSR+08,DY08,SOJH09,MN12,BUV14,SU15,DSSUV16]

[Dwork and Mulligan 2013]

Thank you!

Washington, DC, May 10, 2016