

# CS208: Applied Privacy for Data Science Course Overview

Salil Vadhan, James Honaker, Wanrong Zhang,  
Jayshree Sarathy, Mike Shoemate

School of Engineering & Applied Sciences  
Harvard University

January 18, 2022

# Important Announcement

- If you might attend the class *without formally registering this week* (e.g. you want to audit, or might add it later, or will be cross-registering), then please fill out the following Google form by end of today, 1/19:  
<https://forms.gle/7aG5fvcCQTnH6zNj9>
- Why?
  - To calibrate size of teaching staff & room
  - To inform you of Harvard COVID policies (esp. for non-Harvard auditors)

# Plan for today

- First hour: course overview
  - Salil: motivation & definition of differential privacy
  - Wanrong: the theory of DP
  - James: from theory to practice
  - Jayshree: ethics, law, and society
  - Mike: programming and implementation
  - Salil: class structure
- Second hour: Q & A
  - Group discussion
  - Individual questions in breakout rooms

# Data Privacy: The Problem

Given a dataset with sensitive information, such as:

- Census data
- Health records
- Social network activity
- Telecommunications data

Academic research

- Informing policy
- Identifying subjects for drug trial
- Searching for terrorists
- Market analysis
- ...

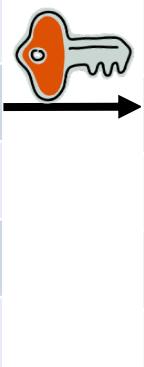
How can we:

- enable “desirable uses” of the data
- while protecting the “privacy” of the data subjects?

????

# Approach 1: Encrypt the Data

Name	Sex	Blood		HIV?
Chen	F	B		Y
Jones	M	A		N
Smith	M	O		N
Ross	M	O		Y
Lu	F	A		N
Shah	M	B		Y

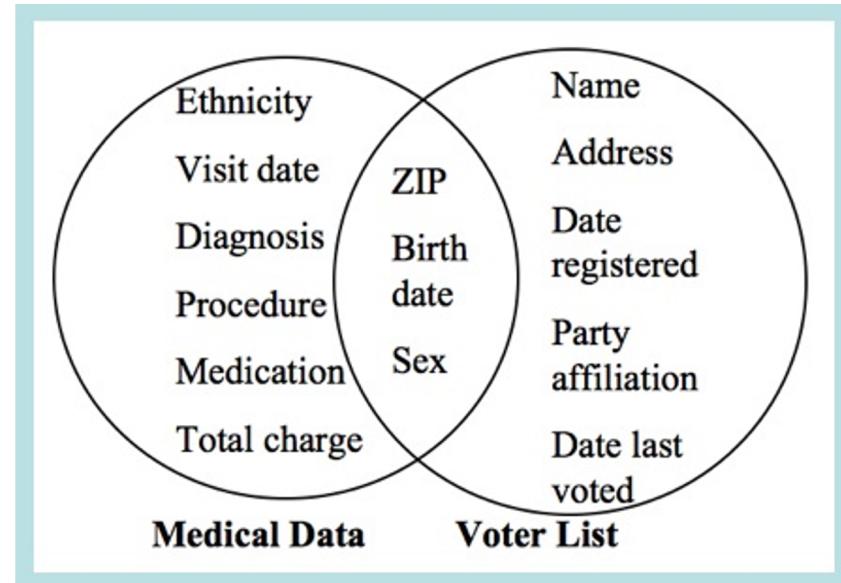


Name	Sex	Blood		HIV?
100101	001001	110101		110111
101010	111010	111111		001001
001010	100100	011001		110101
001110	010010	110101		100001
110101	000000	111001		010010
111110	110010	000101		110101

## Problems?

# Approach 2: Anonymize the Data

Name	Sex	Blood	HIV?
Chen	F	B	Y
Jones	M	A	N
Smith	M	O	N
Ross	M	O	Y
Lu	F	A	N
Shah	M	B	Y



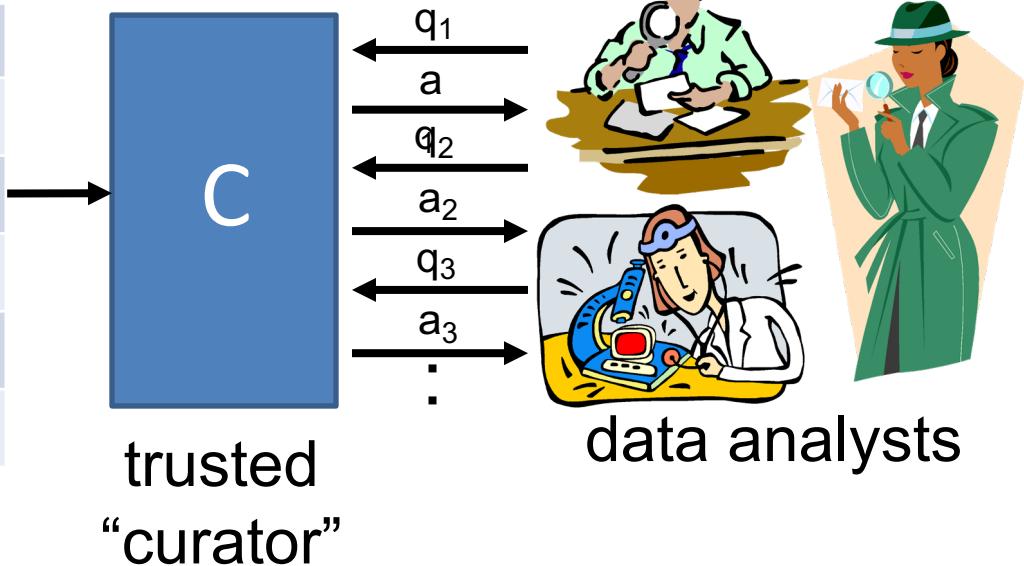
[Sweeney '97]

“re-identification” often easy

## Problems?

# Approach 3: Mediate Access

Name	Sex	Blood	HIV?
Chen	F	B	Y
Jones	M	A	N
Smith	M	O	N
Ross	M	O	Y
Lu	F	A	N
Shah	M	B	Y



# Existing Query Interfaces

**AMERICAN FactFinder** KANSAS MISSOURI VIRGINIA Feedback FAQs Glossary Help

United States Census Bureau

MAN COMMUNITY FACTS GUIDED SEARCH ADVANCED SEARCH DOWNLOAD CENTER

Advanced Search - Search all data in American FactFinder

1 Advanced Search 2 Table Viewer

S0101 AGE AND SEX 2012-2016 American Community Survey 5-Year Estimates

Result 1 of 1 VIEW ALL AS PDF

Table View BACK TO ADVANCED SEARCH

Actions: Modify Table Add/Remove Geographies Bookmark/Save Print Download Create a Map

This table is displayed with default geographies. Click Back to Search to select other geographies using the search options on the left.

Tell us what you think. Provide feedback to help make American Community Survey data more useful for you.

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Versions of this table are available for the following years:

Subject	United States				
	Total	Male	Female		
Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
318,558,162	*****	156,765,322	+/-6,427	161,792,840	+/-6,432

IES NCES National Center for Education Statistics MENU

Search Go

IAP International Data Explorer

IDE IAP PISA PIRLS TIMSS PIAAC TALIS Contact Us

PISA IDE 1. Select Criteria 2. Select Variables 3. Edit Reports 4. Build Reports Help

STEP 4: View each report table by selecting the report name from the drop-down menu. Create report types to edit and preview, each tab created represents one report type to export.

Subject: Age: Mathematics, Reading, and Science, 15 years  
Jurisdiction: International Average (OECD Countries)  
Measure: PISA Mathematics Scale: Overall Mathematics  
Variable: All students  
Year: 2015

Select Report: Report 1 Link to this Page Export Reports

Table Chart Significance Test Gap Analysis Regression Analysis

Averages for PISA mathematics scale: overall mathematics, age 15 years by All students [TOTAL], year and jurisdiction: 2015

Year	Jurisdiction	Average	All students	Standard Error
2015	International Average (OECD Countries)	490	(0.4)	

NOTE: The PISA mathematics scale: overall mathematics ranges from 0 to 1000. Some apparent differences between estimates may not be statistically significant.  
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2015 Mathematics, Reading, and Science Assessment.

NCBI Resources How To Sign in to NCBI

Phenotype-Genotype Integrator

All Databases Search

Search Summary

Search Criteria Phenotype Selection

Trait: Abdominal Fat; Peanut Hypersensitivity  
P-Value: < 1 x 10<sup>-1</sup>

Genotype Selection - Location

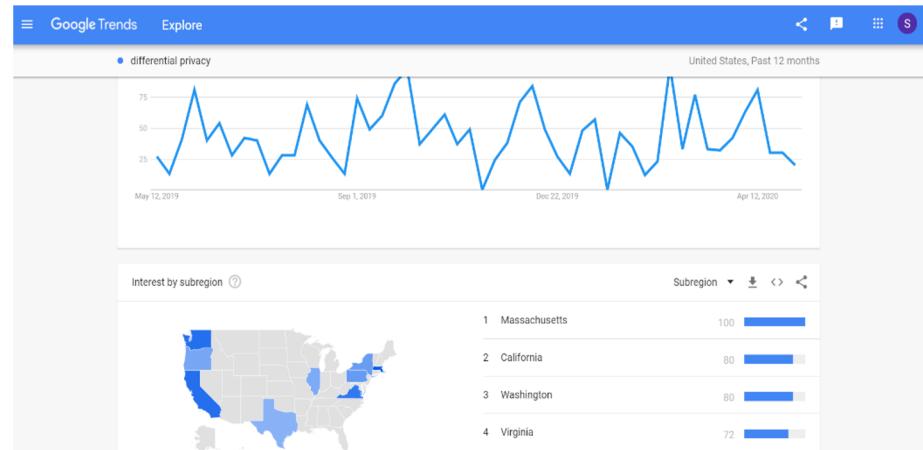
Chromosome: 13

Modify Search

Search Results

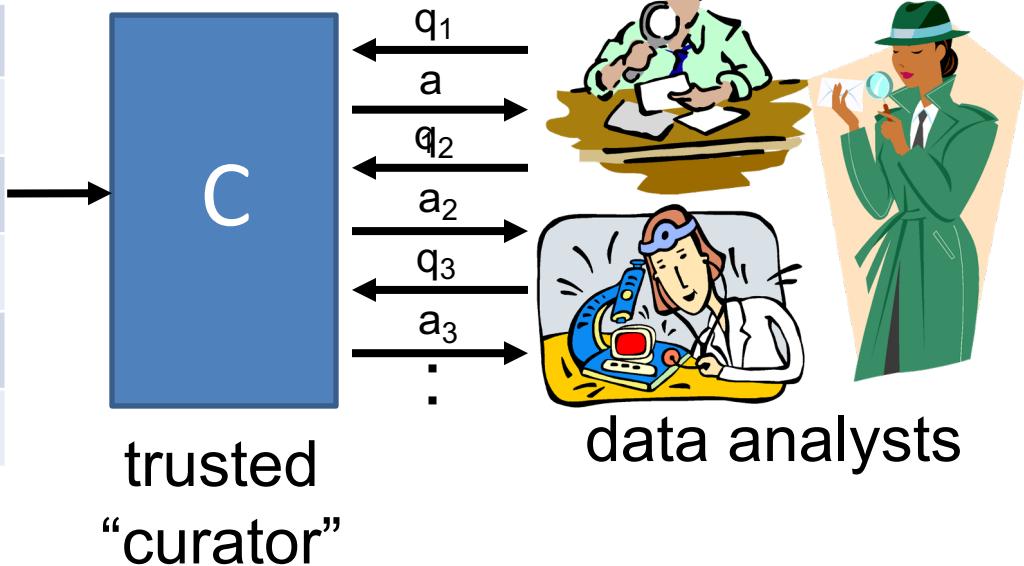
Association Results 1 - 3 of 3 Searched by phenotype trait, SNP chromosome, and P-Value.  
Genes 1 - 4 of 4 Searched by gene IDs retrieved from association results.  
SNPs 1 - 2 of 2 Searched by SNP rs numbers retrieved from association results.  
eQTL Data No data found Searched by SNP rs numbers retrieved from association results and P-Value.  
dbGaP Studies No data found Searched by traits retrieved from association results.  
Genome View 2 SNPs and 4 genes over 1 chromosome.

Modify Search Show All Hide All



# Approach 3: Mediate Access

Name	Sex	Blood	HIV?
Chen	F	B	Y
Jones	M	A	N
Smith	M	O	N
Ross	M	O	Y
Lu	F	A	N
Shah	M	B	Y



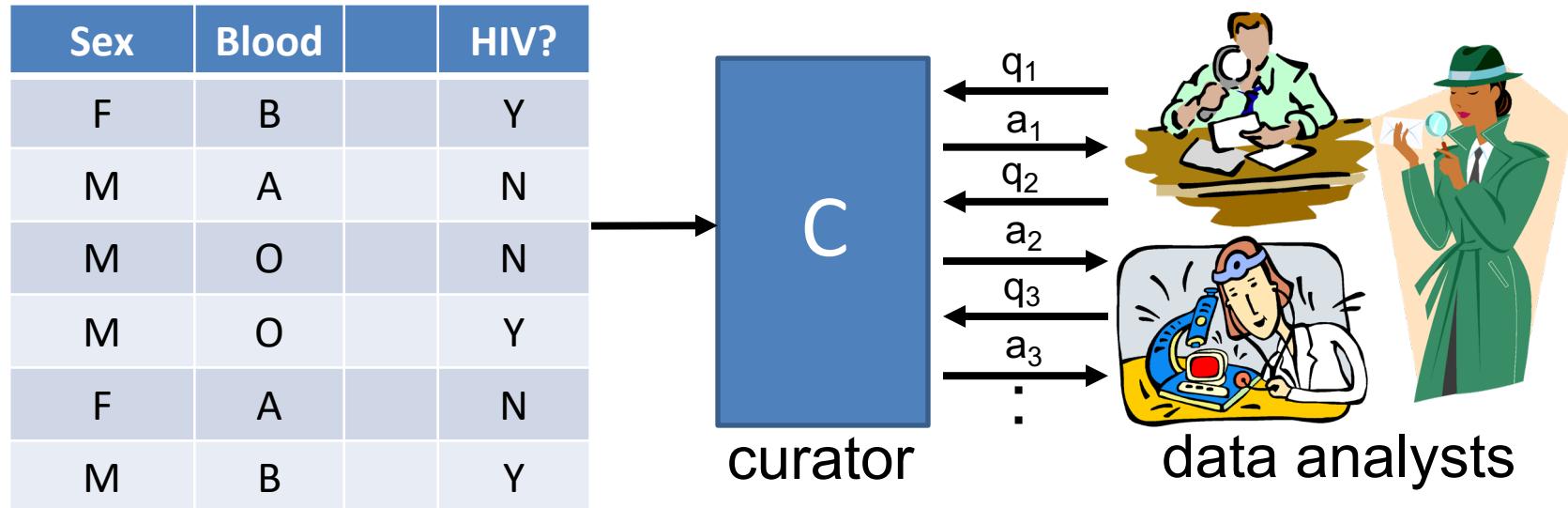
## Problems?

# Privacy Models from Theoretical CS

Model	Utility	Privacy	Who Holds Data?
Differential Privacy	statistical analysis of dataset	individual-specific info	trusted curator
Secure Multiparty Computation	any query desired	everything other than result of query	original users (or semi-trusted delegates)
Fully Homomorphic (or Functional) Encryption	any query desired	everything (except possibly result of query)	untrusted server

# Differential privacy

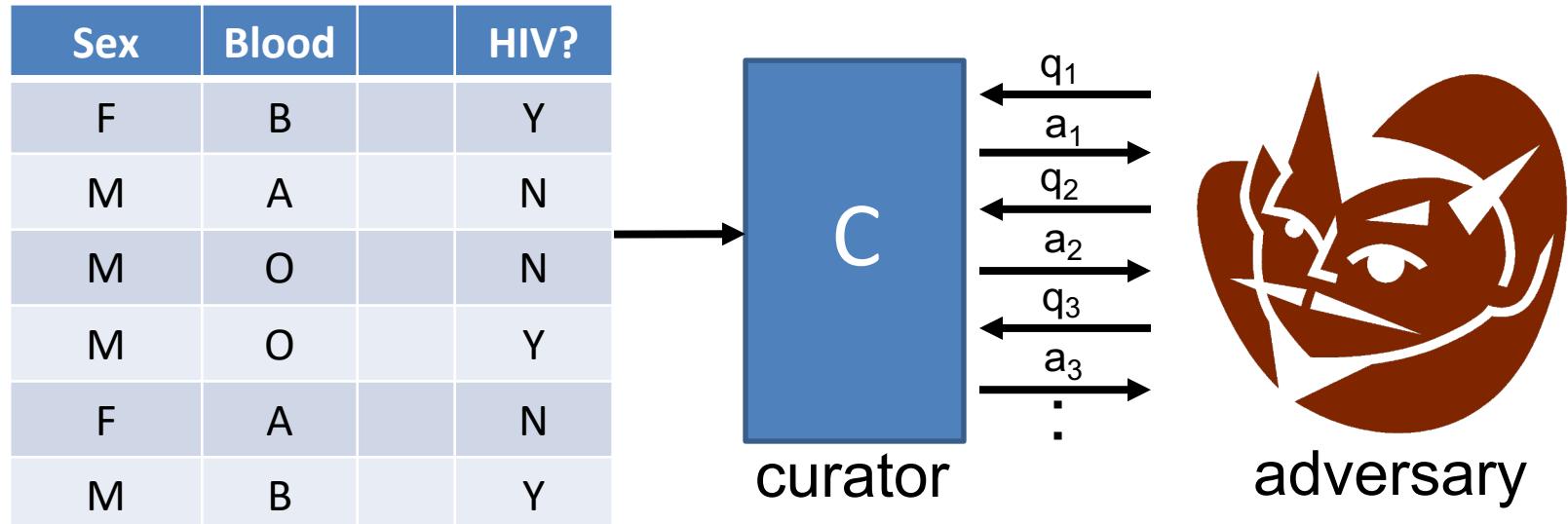
[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



**Requirement:** effect of each individual should be “hidden”

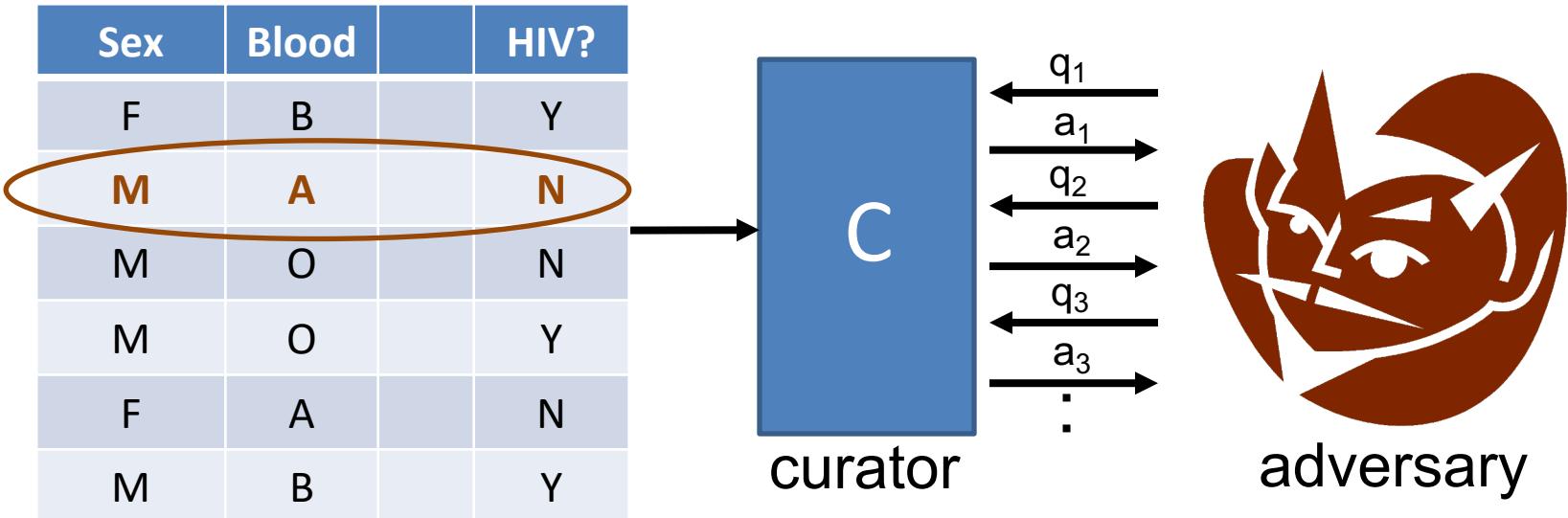
# Differential privacy

[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



# Differential privacy

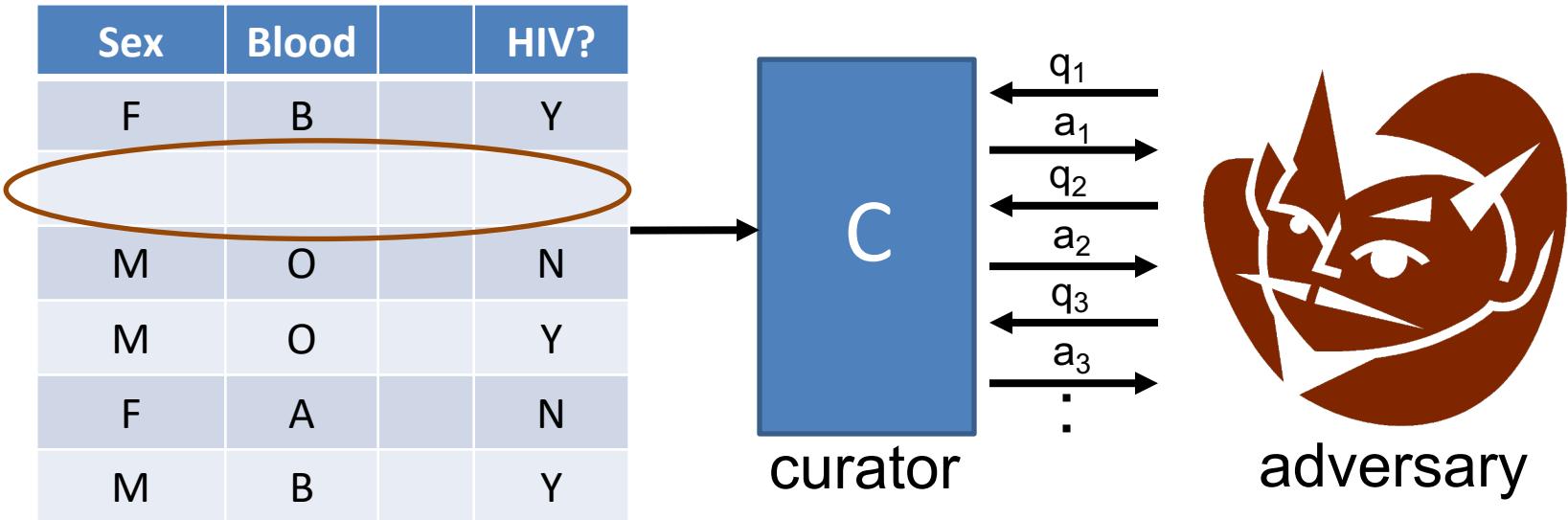
[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



**Requirement:** an adversary shouldn't be able to tell if any one person's data were changed arbitrarily

# Differential privacy

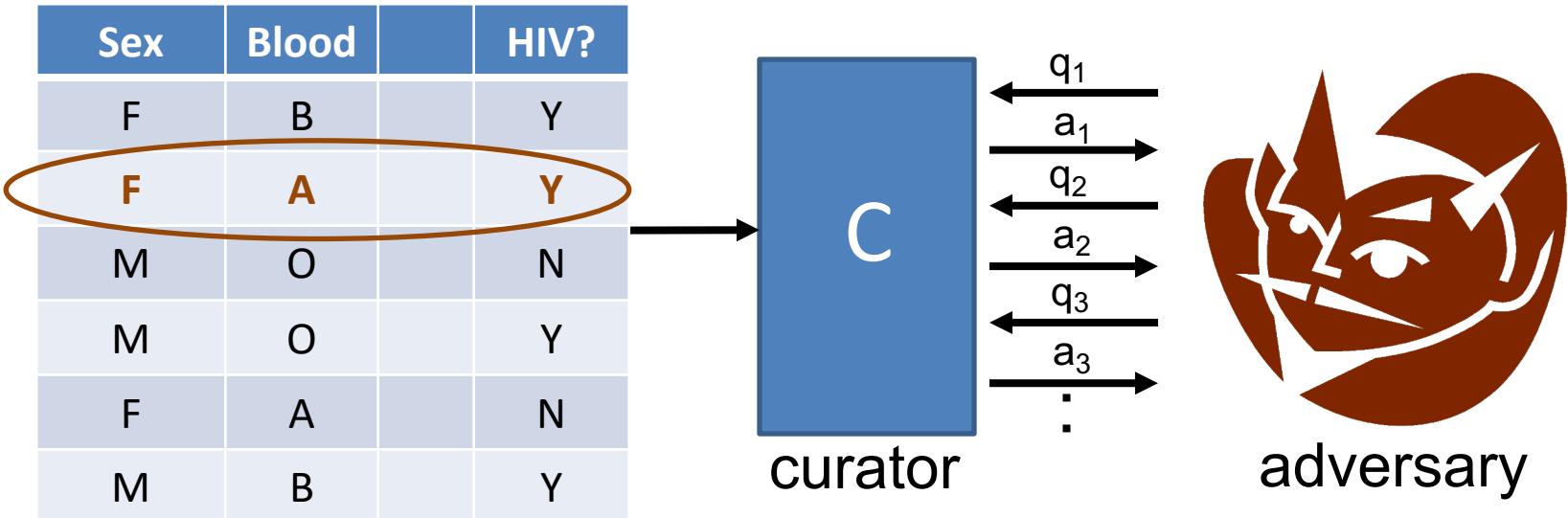
[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



**Requirement:** an adversary shouldn't be able to tell if any one person's data were changed arbitrarily

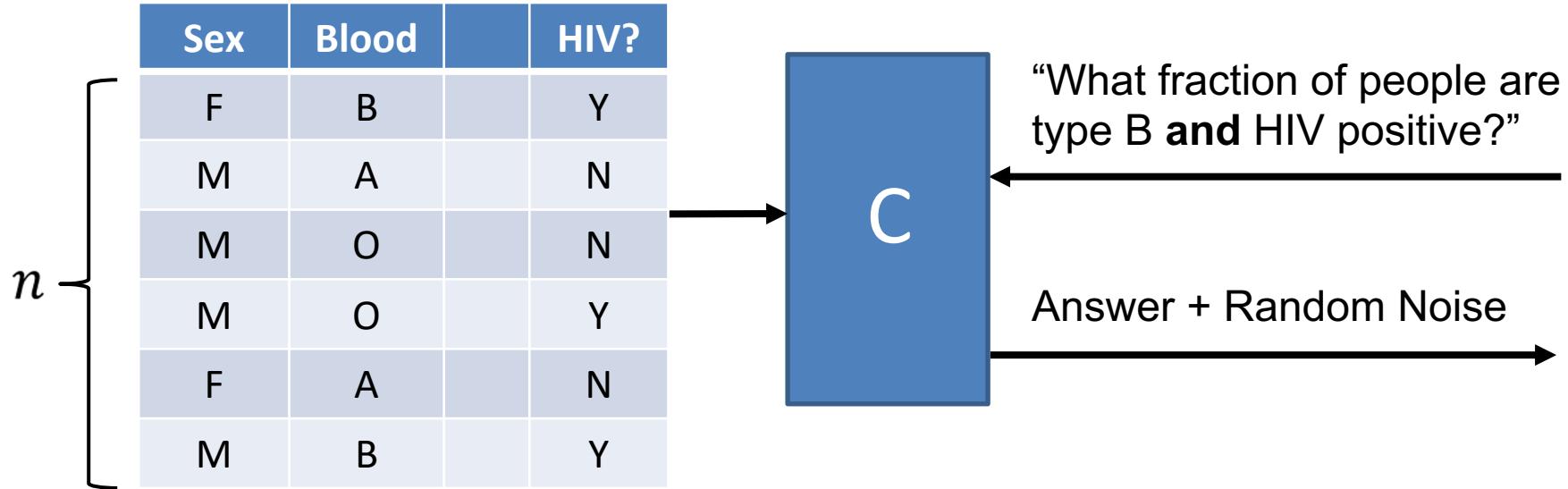
# Differential privacy

[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



**Requirement:** an adversary shouldn't be able to tell if any one person's data were changed arbitrarily

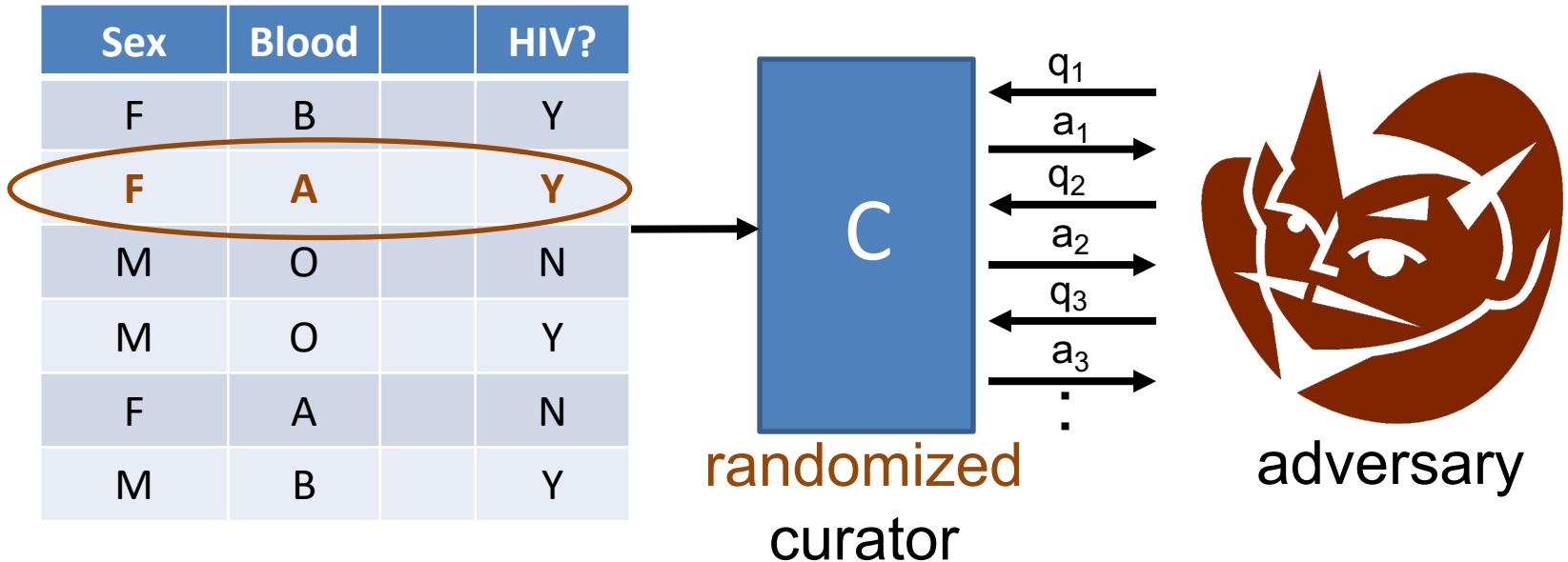
# Simple approach: random noise



- Very little noise needed to hide each person as  $n \rightarrow \infty$ .

# Differential privacy

[Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]



**Requirement:** for all  $D, D'$  differing on one row, and all  $q_1, \dots, q_t$

$$\text{Distribution of } C(D, q_1, \dots, q_t) \approx_{\epsilon} \text{Distribution of } C(D', q_1, \dots, q_t)$$

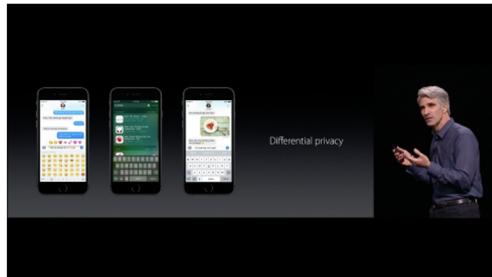
# Algorithm Implementation

- OpenDP Library  
<https://docs.opendp.org>
- Common themes:
  - Many algorithms for the same general class of estimator
  - Side-channel-safe sampling algorithms
  - Accounting for finite data representations
- High demand for robust DP algorithm implementations in industry

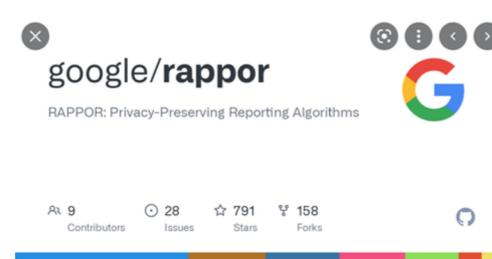
Apple will not  
see your data



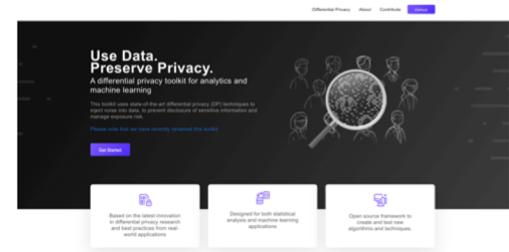
# Differential Privacy Deployed



Apple



Google



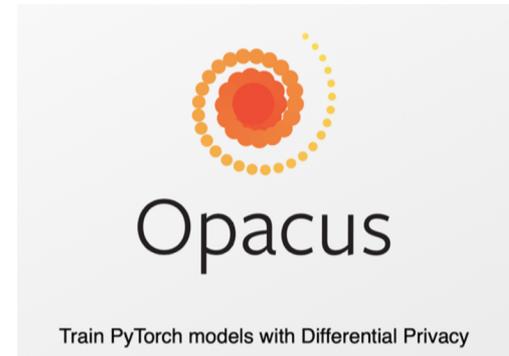
Microsoft



Census Bureau



Uber

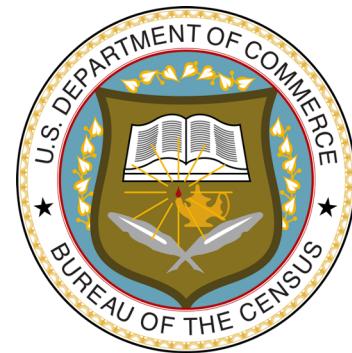


Meta

# Major Deployments of DP

## U.S. Census Bureau

- “OnTheMap” commuter data (2006)
- Planned: all public-use products from 2020 decennial census



## Google

- “RAPPOR” for Chrome Statistics (2014)



## Microsoft

- SmartNoise (2020)



## Apple

- iOS10 and Safari(2016) and PCM (2022)



# Harvard Privacy Tools Project

<http://privacymethods.seas.harvard.edu/>



Google

Alfred P. Sloan  
FOUNDATION

Computer Science, Law, Social Science, Statistics



The Institute for Quantitative Social Science  
HARVARD UNIVERSITY



A **community effort** to build a **trustworthy** and **open-source** suite of differential privacy tools that can be **easily adopted** by custodians of sensitive data to make it available for **research and exploration** in the public interest.

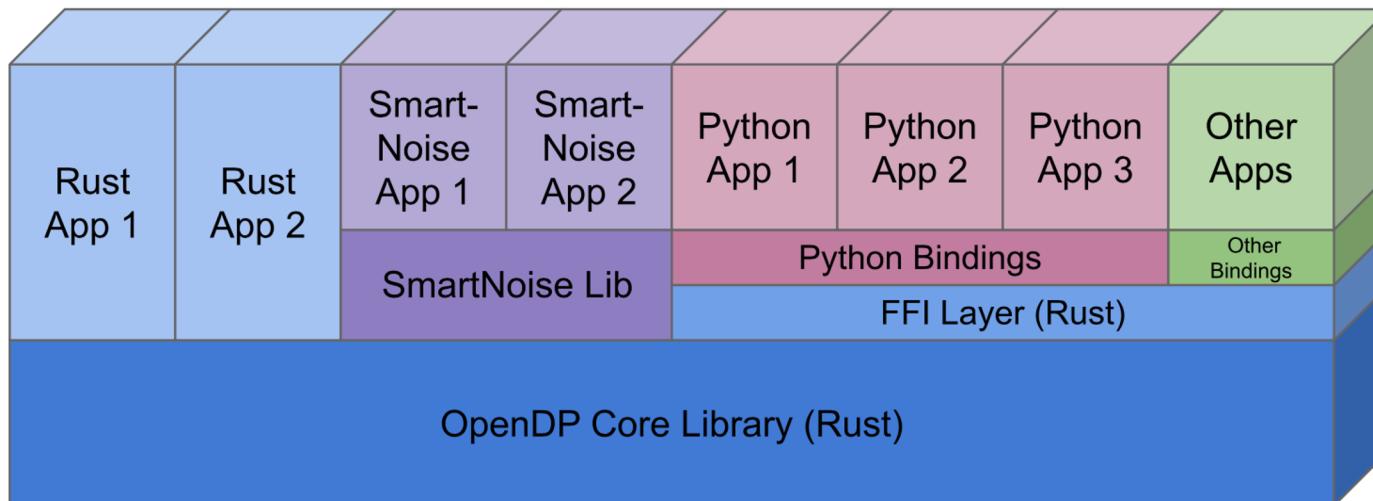
## Why?

- Channel our collective advances on science & practice of DP
- Enable wider adoption of DP
- Address high-demand, compelling use cases
- Provide a starting point for custom DP solutions
- Identify important research directions for the field

Project site: <http://opendp.io>



```
>>> from opendp.meas import make_base_geometric
...
>>> # call the constructor to produce a measurement
>>> base_geometric = make_base_geometric(scale=1.0)
...
>>> # investigate the privacy relation
>>> absolute_distance = 1
>>> epsilon = 1.0
>>> assert base_geometric.check(d_in=absolute_distance, d_out=epsilon)
...
>>> # feed some data/invoke the measurement as a function
>>> aggregated = 5
>>> release = base_geometric(aggregated)
```



# Target: Data Repositories



HARVARD  
LIBRARY

*Share, Cite, Reuse, Archive Research Data*  
Scientific data for reproducible research

## Harvard Dataverse Network

POWERED BY THE  
**Dataverse Network™ PROJECT** v. 3.6.2



Create Account

Log In

Search this Dataverse Network

Search

[Advanced Search](#) [Tips](#)

We're redesigning Dataverse and want your feedback! Please check out our [Beta Site](#)

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. Learn more about the [Dataverse Network](#).

## Dataverses

[Create Dataverse](#)

**706** Dataverses

A **Dataverse** is a container for research data studies, customized and managed by its owner.

### RECENTLY RELEASED DATERVERSES

Eben N. Broadbent

Jun 2, 2014

USoc: Quantitative Methods over the Undergraduate Life Course

May 30, 2014

## Studies

**53,896** Studies, **739,606** Files, **1,015,093** Downloads

A **study** is a container for a research data set. It includes cataloging information, data files and complementary files.

### RECENTLY RELEASED STUDIES

Replication data for: Neoliberal Reform and Protest in Latin American Democracies: A Replication and Correction by Solt, Frederick; Kim, Dongkyu; Lee, Kyu Young; Willardson, Spencer; Kim, Seokdang

Jun 3, 2014



# Target: Data Repositories





Create Account

Log In

## Murray Research Archive Original Collection Dataverse

## INTERGENERATIONAL STUDIES, 1932-1982

hdl:1902.1/00627UNF:3:jYQzhUZ5MxpaKGMylojITA==

Version: 5 – Released: Tue Jun 19 13:50:23 EDT 2012

Cataloging Information

DATA &amp; ANALYSIS

Comments (0)

Versions

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Access to some files is restricted, and those files are not available for downloading. Check the [Terms of Use](#) for more information.

 Select all files[Download All Selected Files](#)**1. Documentation** 006271HD-Intergenerational-Cat

Adobe PDF - 41 MB -

[Download](#)

Description of coded data variables

 006271HD-Intergenerational-Cat

Adobe PDF - 7 MB -

[Download](#)

Blank measures for study

 006271HD-Intergenerational-Cat

Adobe PDF - 173 KB -

[Download](#)

Overview: abstract, research methodology, publications, and other info.

**2. Berkeley Data** 006271HD-Intergenerational-BerkSpou-Data.por

SPSS Portable - 29 KB - 0 downloads

[Restricted](#)

Data on Spouses in Berkeley Sample in SPSS Portable Format

 006271HD-Intergenerational-BerkSpou-Data.tab

Tab Delim

[Restricted](#)

Data on Spouses of Berkeley Sample in Tab

TABUL

Goal: enable wider sharing while protecting privacy

 006271HD-Intergenerational-BerkSubj-Data.por

SPSS Portable - 217 KB - 0 downloads

[Restricted](#)

Data on Subjects in Berkeley Sample in SPSS Portable Format

# Privacy Preserving Interfaces

[My Data](#)[My Profile](#)[Logout](#)

⚠ The DPcreator takes the first 20 variables of the dataset. The default type has been inferred from the dataset. Incorrect type labeling can result in privacy violation.

💡 Any changes will be applied for the purpose of creating the differential privacy release only, and will **not affect** the original data file. dataset. Incorrect type labeling can result in privacy violation.

[Continue](#)

Last saved: 9/20/2021, 10:22:49 PM

⌚ Remaining: 2d 14h 16min

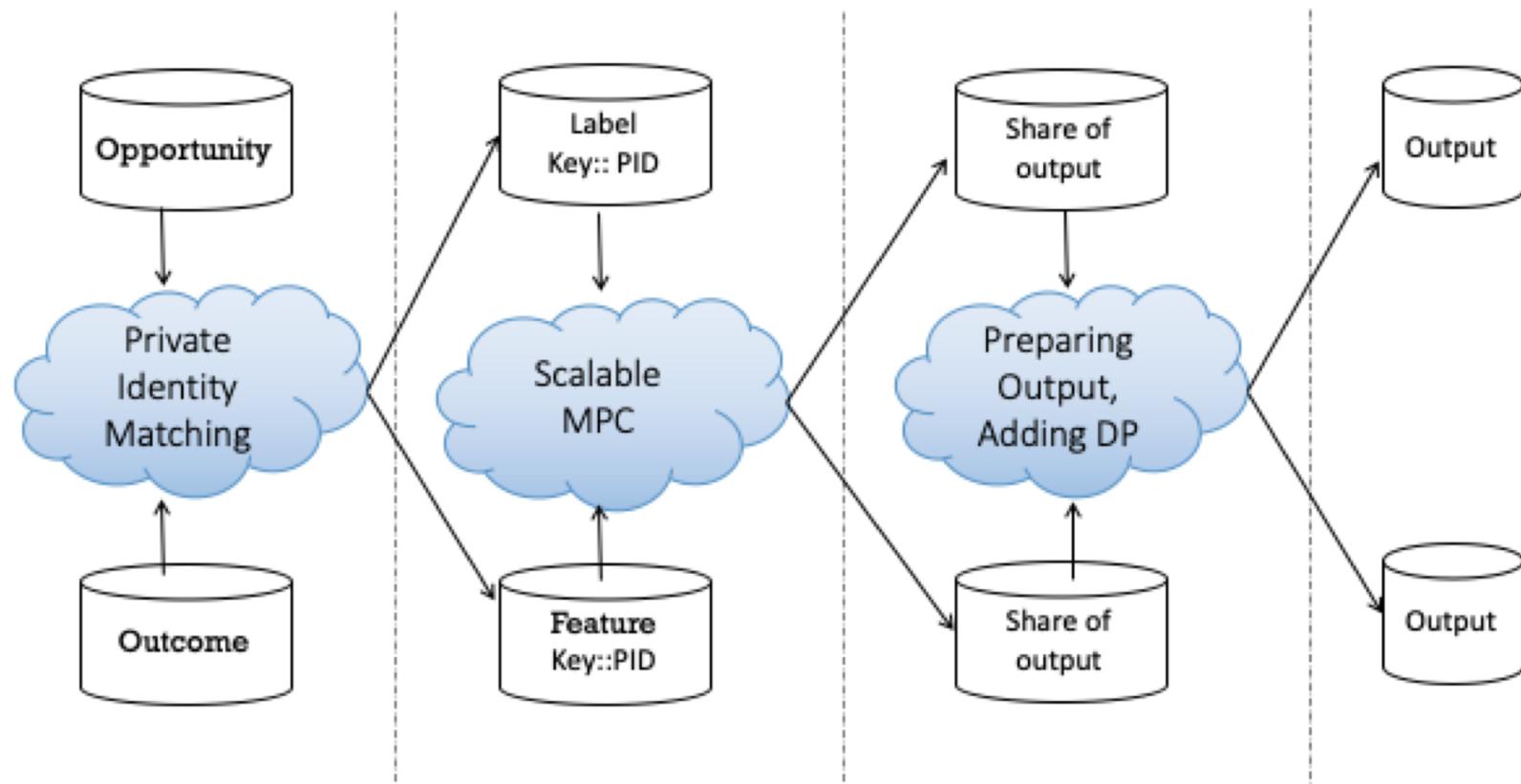
	Variable name	Variable label	Type <small>?</small>	Additional variable information <small>?</small>
1	Trial	Trial	Categorical	<a href="#">Add categories</a>
2	Session	Session	Boolean	
3	Subject	Subject	Categorical	<a href="#">Add categories</a>
4	Language	Language	Boolean	
5	EyeHeight	EyeHeight	Numerical	-8  5
6	TypingSpeed	TypingSpeed	Numerical	<a href="#">Add min</a> <a href="#">Add max</a>

# Challenges for DP in Practice

- Accuracy for “small data” (moderate values of  $n$ )
- Modelling & managing privacy loss over time
  - Especially over many different analysts & datasets
- Analysts used to working with raw data
  - One approach: “Tiered access”
  - DP for wide access, raw data only by approval with strict terms of use (cf. Census PUMS vs. RDCs)
- Cases where privacy concerns are not “local” (e.g. privacy for large groups) or utility is not “global” (e.g. targeting)
- Matching guarantees with privacy law & regulation
- ...

# Challenge for DP in Practice

When to rely on DP and how to combine DP with other privacy enhancing techniques?



# Differential Privacy: Interpretations

$$\text{Distribution of } C(D, q_1, \dots, q_t) \approx_{\epsilon} \text{Distribution of } C(D', q_1, \dots, q_t)$$

- Whatever an adversary learns about me, it could have learned from everyone else's data.
- Mechanism cannot leak "individual-specific" information.
- Above interpretations hold regardless of adversary's auxiliary information.

But

- No protection for information that is not localized to a few rows.
- Differential privacy does not prevent statistical analysis and machine learning.

# Nice Properties of Differential Privacy

## Privacy loss measure ( $\epsilon$ )

“Information structures no longer have to be the binary extremes”

## Post-processing:

“No adversary can break the privacy guarantee”

## Composition:

“The epsilons add up”

# Basic Definitions and Techniques

- Definitions
  - Approximate DP/ local DP/ alternative definitions
- Introducing randomness (**How to add noise and how much noise**)
  - Randomized Response/Laplace/Exponential/Gaussian mechanisms
- Composition theory

# Some Differentially Private Algorithms

## Private Machine Learning

- Private ERM [RBHT12, CMS11, KST12, BST14]
- Modern ML: DPSGD [ACG+16, WBK19, MTZ19] and beyond;  
PATE [PAE+17, PSM+18]

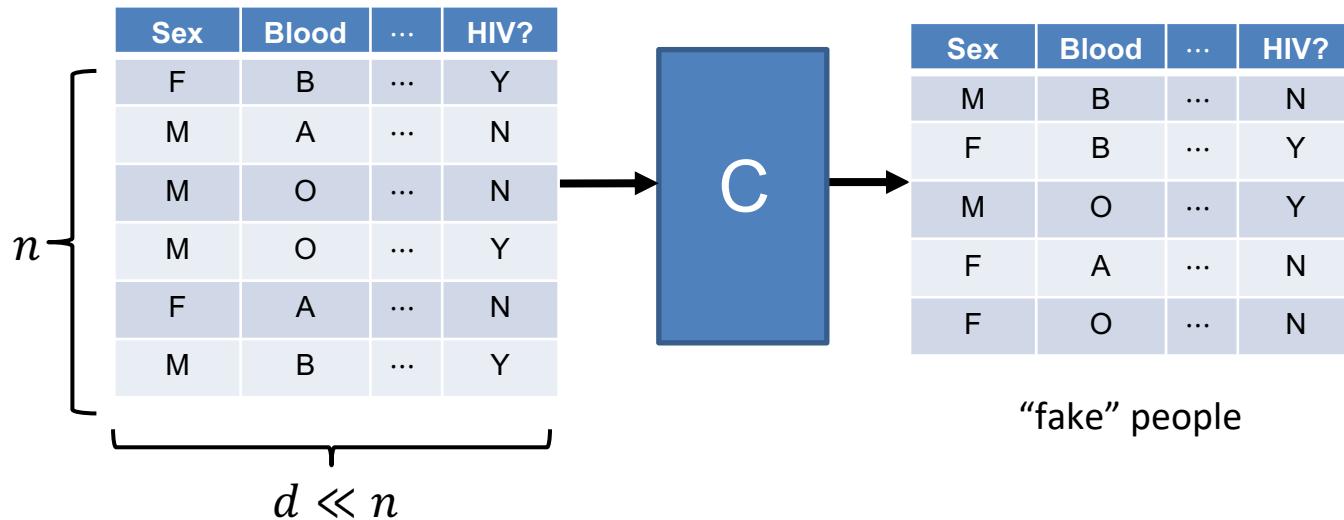
## Private Stats

- Mean estimation & Confidence interval [BDK+2020, KSU20, KV18]
- Hypothesis testing [GLRV16, GR18, CKS+19, CKM+19]
- singular value decomposition [HR12, HR13, KT13, DTTZ14]

## Private Synthetic Data [HLM12, GAH+14, VTB+20, NWD21, LBW21]

...

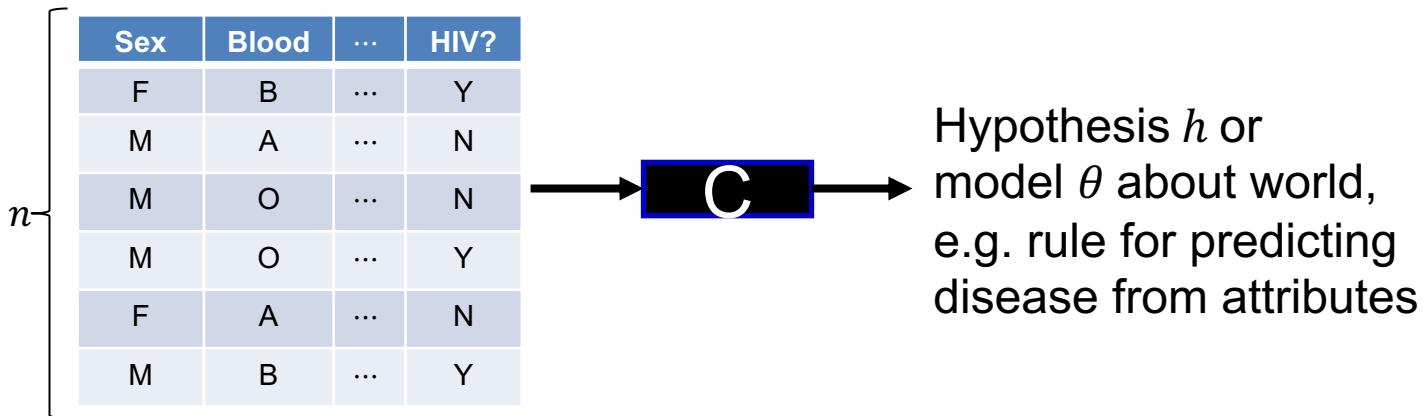
# Amazing possibility: synthetic data



**Utility:** preserves fraction of people with *every* set of attributes!

**Problem:** uses computation time exponential in  $d$ .

# Amazing Possibility II: Statistical Inference & Machine Learning



**Theorem [KLNRS08,S11]:** Differential privacy for vast array of machine learning and statistical estimation problems with little loss in convergence rate as  $n \rightarrow \infty$ .

# DP Theory

Differential privacy research has

- many intriguing theoretical challenges
- rich connections w/other parts of CS theory & mathematics

e.g. cryptography, learning theory, game theory & mechanism design, convex geometry, pseudorandomness, optimization, approximability, communication complexity, statistics, ...

# Ethics, Law, and Society

- Analyze differential privacy deployments from various perspectives
  - **Ethics:** How does differential privacy alter ethical considerations around collecting sensitive data for public interest purposes?
  - **Law and policy:** What is the relationship between differential privacy and existing regulatory standards for privacy protection?
  - **Science & Technology Studies:** How does differential privacy reflect and shape power dynamics among data subjects, data holders, and researchers?
- Identify critiques, gaps, points of tension, and possible solutions