

# Responsible Data Science

## Anonymity and privacy

*March 20 - 29, 2023*

---

**Prof. Elisha Cohen**

Center for Data Science  
New York University

# Reading for weeks 7 & 8

## Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

### Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

### 1 Introduction

Datasets containing *micro-data*, that is, information about specific individuals, are increasingly becoming public in response to "open government" laws and to support data mining research. Some datasets include legally protected information such as health histories; others contain individual preferences and transactions, which many people may view as private or sensitive.

Privacy risks of publishing micro-data are well-known. Even if identifiers such as names and Social Security numbers have been removed, the adversary can use background knowledge and cross-correlation with other databases to re-identify individual data records. Famous attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with a public voter database [25] and privacy breaches caused by (ostensibly anonymized) AOL search data [16].

Micro-data are characterized by high dimensionality

and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no "similar" records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the "fat tail" phenomenon: individual transaction and preference records tend to include statistically rare attributes.

**Our contributions.** Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the probability of successful de-anonymization, the other on the amount of information recovered about the target. Unlike previous work [25], we do not assume *a priori* that the adversary's knowledge is limited to a fixed set of "quasi-identifier" attributes. Our model thus encompasses a much broader class of de-anonymization attacks than simple cross-database correlation.

Our second contribution is a very general class of de-anonymization algorithms, demonstrating the fundamental limits of privacy in public micro-data (section 4). Under very mild assumptions about the distribution from which the records are drawn, the adversary with a small amount of background knowledge about an individual can use it to identify, with high probability, this individual's record in the anonymized dataset and to learn all anonymously released information about him or her, including sensitive attributes. For *sparse* datasets, such as most real-world datasets of individual transactions, preferences, and recommendations, very little background knowledge is needed (as few as 5-10 attributes in our case study). Our de-anonymization algorithm is *robust* to the imprecision of the adversary's background knowledge and to perturbation that may have been applied to the data prior to release. It works even if only a *subset* of the original dataset has been published.

Our third contribution is a practical analysis of the Netflix Prize dataset, containing anonymized movie ratings of 500,000 Netflix subscribers (section 5). Netflix—the world's largest online DVD rental

DOI:10.1145/1866739.1866758

## What does it mean to preserve privacy?

BY CYNTHIA DWORK

# A Firm Foundation for Private Data Analysis

This long history is a testament to the importance of the problem. Statistical databases can be of enormous social value; they are used for apportioning resources, evaluating medical therapies, understanding the spread of disease, improving economic utility, and informing us about ourselves as a species.

The data may be obtained in diverse ways. Some data, such as census, tax, and other sorts of official data, is compelled; other data is collected opportunistically, for example, from traffic on the Internet, transactions on Amazon, and search engine query logs; other data is provided altruistically, by respondents who hope that sharing their information will help others to avoid a specific misfortune, or more generally, to increase the public good. Altruistic data donors are typically promised their individual data will be kept confidential—in short, they are promised "privacy." Similarly, medical data and legally compelled data, such as census data and tax return data, have legal privacy

### » key insights

- In analyzing private data, only by focusing on rigorous privacy guarantees can we convert the cycle of "propose-break-propose again" into a path of progress.
- A natural approach to defining privacy is to require that accessing the database teaches the analyst nothing about any individual. But this is problematic: the whole point of a statistical database is to teach general truths, for example, that smoking causes cancer. Learning this fact teaches the data analyst something about the likelihood with which certain individuals, not necessarily in the database, will develop cancer. We therefore need a definition that separates the utility of the database (learning that smoking causes cancer) from the increased risk of harm due to joining the database. This is the intuition behind differential privacy.
- This can be achieved, often with low distortion. The key idea is to randomize responses so as to effectively hide the presence or absence of the data of any individual over the course of the lifetime of the database.

IN THE INFORMATION realm, loss of privacy is usually associated with failure to control access to information, to control the flow of information, or to control the purposes for which information is employed. Differential privacy arose in a context in which ensuring privacy is a challenge even if all these control problems are solved: privacy-preserving statistical analysis of data.

The problem of *statistical disclosure control*—revealing accurate statistics about a set of respondents while preserving the privacy of individuals—has a venerable history, with an extensive literature spanning statistics, theoretical computer science, security, databases, and cryptography (see, for example, the excellent survey of Adam and Wortmann,<sup>1</sup> the discussion of related work in Blum et al.,<sup>2</sup> and the *Journal of Official Statistics* dedicated to confidentiality and disclosure control).

# Reading for weeks 7 & 8

data security

1 of 26

TEXT  
ONLY

## Understanding **DATABASE RECONSTRUCTION ATTACKS** on Public Data

THESE ATTACKS  
ON STATISTICAL  
DATABASES ARE  
NO LONGER A  
THEORETICAL  
DANGER.

SIMSON GARFINKEL  
JOHN M. ABOWD, AND  
CHRISTIAN MARTINDALE  
U.S. CENSUS BUREAU

In 2020 the U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to DRAs (*database reconstruction attacks*), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so that the reconstruction no longer results in the original data. This has implications for the 2020 Census.

The goal of the census is to count every person once,



WILLIAM DUKE/PHOTO BY DAXIAO PRODUCTIONS/SHUTTERSTOCK

### Can a set of equations keep U.S. census data private?

By Jeffrey Mervis | Jan. 4, 2019, 2:50 PM

The U.S. Census Bureau is making waves among social scientists with what it calls a "sea change" in how it plans to safeguard the confidentiality of data it releases from the decennial census.

The agency announced in September 2018 that it will apply a mathematical concept called differential privacy to its release of 2020 census data after conducting experiments that suggest current approaches can't assure confidentiality. But critics of the new policy believe the Census Bureau is moving too quickly to fix a system that isn't broken. They also fear the changes will degrade the quality of the information used by thousands of researchers, businesses, and government agencies.

The move has implications that extend far beyond the research community. Proponents of differential privacy say a fierce, ongoing legal battle over plans to add a citizenship question to the 2020 census has only underscored the need to assure people that the government will protect their privacy.

#### A noisy conflict

The Census Bureau's job is to collect, analyze, and disseminate useful information about the U.S. population. And there's a lot of it: The agency generated some 7.8 billion statistics about the 308 million people counted in the 2010 census, for example.

At the same time, the bureau is prohibited by law from releasing any information for which "the data furnished by any particular establishment or individual ... can be identified."

Once upon a time, meeting that requirement meant simply removing the names and addresses of respondents. Over the past several decades, however, census officials have developed a bag of statistical tricks aimed at providing additional protection without undermining the quality of the data.

# Reading for weeks 7 & 8

## DataSynthesizer: Privacy-Preserving Synthetic Datasets

Haoyue Ping  
Drexel University, USA  
hp354@drexel.edu

Julia Stoyanovich\*  
Drexel University, USA  
stoyanovich@drexel.edu

Bill Howe†  
University of Washington, USA  
billhowe@cs.washington.edu

### ABSTRACT

To facilitate collaboration over sensitive data, we present DataSynthesizer, a tool that takes a sensitive dataset as input and generates a structurally and statistically similar synthetic dataset with strong privacy guarantees. The data owners need not release their data, while potential collaborators can begin developing models and methods with some confidence that their results will work similarly on the real dataset. The distinguishing feature of DataSynthesizer is its usability — the data owner does not have to specify any parameters to start generating and sharing data safely and effectively.

DataSynthesizer consists of three high-level modules — DataDescriber, DataGenerator and ModelInspector. The first, DataDescriber, investigates the data types, correlations and distributions of the attributes in the private dataset, and produces a data summary, adding noise to the distributions to preserve privacy. DataGenerator samples from the summary computed by DataDescriber and outputs synthetic data. ModelInspector shows an intuitive description of the data summary that was computed by DataDescriber, allowing the data owner to evaluate the accuracy of the summarization process and adjust any parameters, if desired.

We describe DataSynthesizer and illustrate its use in an urban science context, where sharing sensitive, legally encumbered data between agencies and with outside collaborators is reported as the primary obstacle to data-driven governance.

The code implementing all parts of this work is publicly available at <https://github.com/DataResponsibly/DataSynthesizer>.

### CCS CONCEPTS

• Security and privacy → Data anonymization and sanitization; Privacy protections; Usability in security and privacy;

### KEYWORDS

Data Sharing; Synthetic Data; Differential Privacy

\*This work was supported in part by NSF Grants No. 1464327 and 1539856, and BSF Grant No. 201439.

†This work was supported by the University of Washington Information School, Microsoft, the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #8888) through the Data Science Environments program.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

SSDBM '17, Chicago, IL, USA

© 2017 ACM. 978 1 4503 5282 6/17/06. \$15.00

DOI: <http://dx.doi.org/10.1145/3085504.3091117>

### ACM Reference format:

Haoyue Ping, Julia Stoyanovich, and Bill Howe. 2017. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In *Proceedings of SSDBM '17, Chicago, IL, USA, June 27-29, 2017*, 5 pages.  
DOI: <http://dx.doi.org/10.1145/3085504.3091117>

### 1 INTRODUCTION

Collaborative projects in the social and health sciences increasingly require sharing sensitive, privacy-encumbered data. Social scientists, government agencies, health workers, and non-profits are eager to collaborate with data scientists, but formal data sharing agreements are too slow and expensive to create in ad hoc situations — our colleagues report that 18 months is a typical timeframe to establish such agreements! As a result, many promising collaborations can fail before they even begin. Data scientists require access to the data before they can understand the problem or even determine whether they can help. But data owners cannot share data without significant legal protections in place. Beyond legal concerns, there is a general reluctance to share sensitive data with non-experts before they have “proven themselves,” since they do not understand the context in which the data was collected and may be distracted by spurious results.

To bootstrap these collaborations without incurring the cost of formal data sharing agreements, we saw a need to generate datasets that are *structurally and statistically similar* to the real data but that are 1) obviously synthetic to put the data owners at ease, and 2) offer strong privacy guarantees to prevent adversaries from extracting any sensitive information. These two requirements are not redundant: strong privacy guarantees are not always sufficient to convince data owners to release data, and even seemingly random datasets may not prevent subtle privacy attacks. With this approach, data scientists can begin to develop models and methods with synthetic data, but maintain some degree of confidence that their work will remain relevant when applied to the real data once proper data sharing agreements are in place.

We propose a tool named DataSynthesizer to address this problem. Assume that the private dataset contains one table with  $m$  attributes and  $n$  tuples, and that the values in each attribute are homogeneous, that is, they are all of the same data type. We are interested in producing a synthetic dataset such that summary statistics of all numerical, categorical, string, and datetime attributes are similar to the private dataset. What statistics we preserve depends on the data type, as we will discuss in Section 3.

DataSynthesizer infers the domain of each attribute and derives a description of the distribution of attribute values in the private dataset. This information is saved in a dataset description file, to which we refer as data summary. Then DataSynthesizer is able to generate synthetic datasets of arbitrary size by sampling from the probabilistic model in the dataset description file.

## WINNING THE NIST CONTEST: A SCALABLE AND GENERAL APPROACH TO DIFFERENTIALLY PRIVATE SYNTHETIC DATA

RYAN MCKENNA, GEROME MIKLAU, AND DANIEL SHELDON

College of Information & Computer Sciences, The University of Massachusetts, Amherst, MA 10002  
e-mail address: rmckenna@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusetts, Amherst, MA 10002  
e-mail address: miklau@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusetts, Amherst, MA 10002  
e-mail address: sheldon@cs.umass.edu

**ABSTRACT.** We propose a general approach for differentially private synthetic data generation, that consists of three steps: (1) select a collection of low-dimensional marginals, (2) measure those marginals with a noise addition mechanism, and (3) generate synthetic data that preserves the measured marginals well. Central to this approach is **Private-PGM** [42], a post-processing method that is used to estimate a high-dimensional data distribution from noisy measurements of its marginals. We present two mechanisms, **NIST-MST** and **MST**, that are instances of this general approach. **NIST-MST** was the winning mechanism in the 2018 NIST differential privacy synthetic data competition, and **MST** is a new mechanism that can work in more general settings, while still performing comparably to **NIST-MST**. We believe our general approach should be of broad interest, and can be adopted in future mechanisms for synthetic data generation.

### 1. INTRODUCTION

Data sharing within the modern enterprise is extremely constrained by privacy concerns. Privacy-preserving synthetic data is an appealing solution: it allows existing analytics workflows and machine learning methods to be used while the original data remains protected. But recent research has shown that unless a formal privacy standard is adopted, synthetic data can violate privacy in subtle ways [18, 25]. Differential privacy offers such a formalism, and the problem of differentially private synthetic data generation has therefore received considerable research attention in recent years [3, 6, 9, 13, 14, 26, 31, 32, 39, 40, 52–55, 59, 60, 66, 68, 70, 71].

In 2018, the National Institute of Standards and Technology (NIST) highlighted the importance of this problem by organizing the *Differential Privacy Synthetic Data Competition* [56]. This competition was the first of its kind for the privacy research community, and it encouraged privacy researchers and practitioners to develop novel practical mechanisms for this task. The competition consisted of three rounds of increasing complexity. In this paper we describe **NIST-MST**, the winning entry in the third and final round of the competition. Our algorithm is an instance of a general template for differentially private synthetic data generation that we believe will simplify design of future mechanisms for synthetic data.

Our approach to differentially private synthetic data generation consists of three high-level steps, as show in Figure 1: (1) query selection, (2) query measurement and (3) synthetic

*Key words and phrases:* differential privacy, synthetic data, graphical models.





**motivation**

# Truth or dare?

**Did you go out drinking over the weekend?**

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the class for whom **P** holds

1. flip a coin **C1**

1. if **C1** is tails, then **respond truthfully**

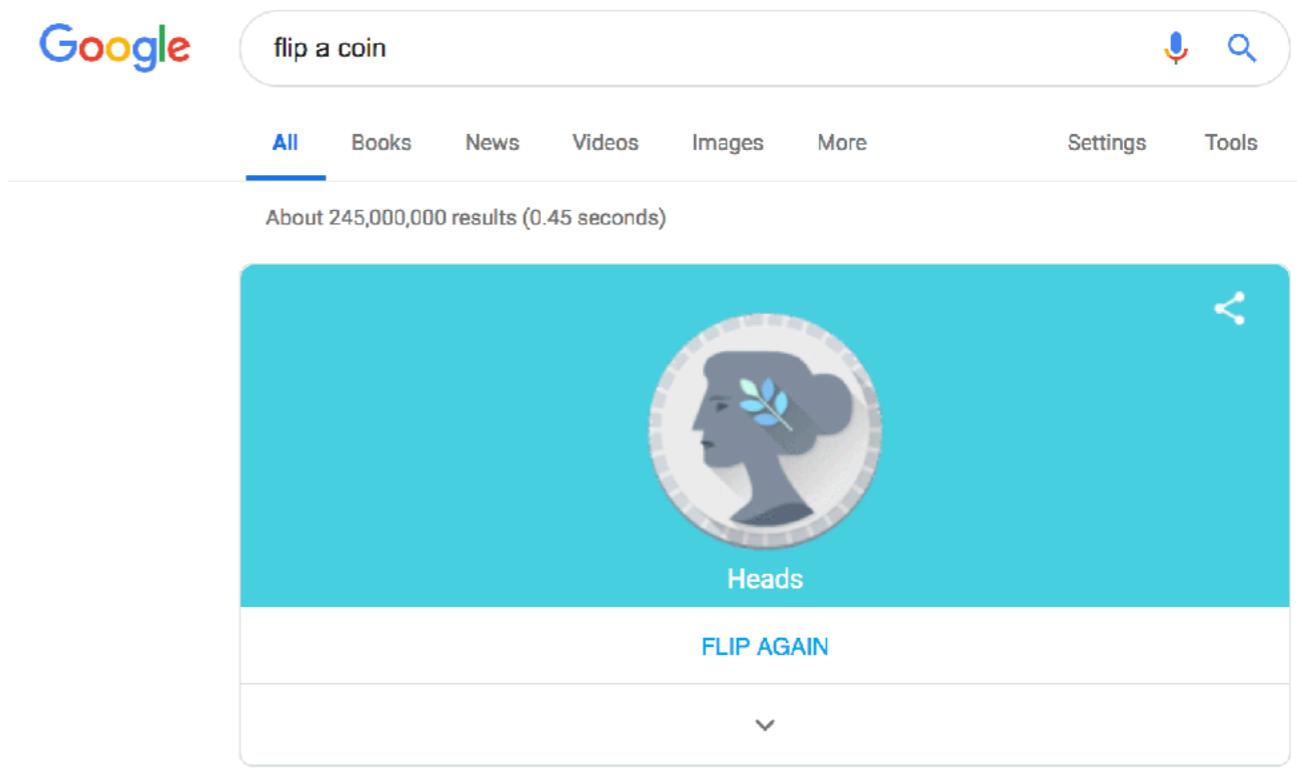
2. if **C1** is heads, then flip another coin **C2**

1. if **C2** is heads then **Yes**

2. else **C2** is tails then respond **No**

the expected number of **Yes** answers is:

$$A = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{p}{2}$$



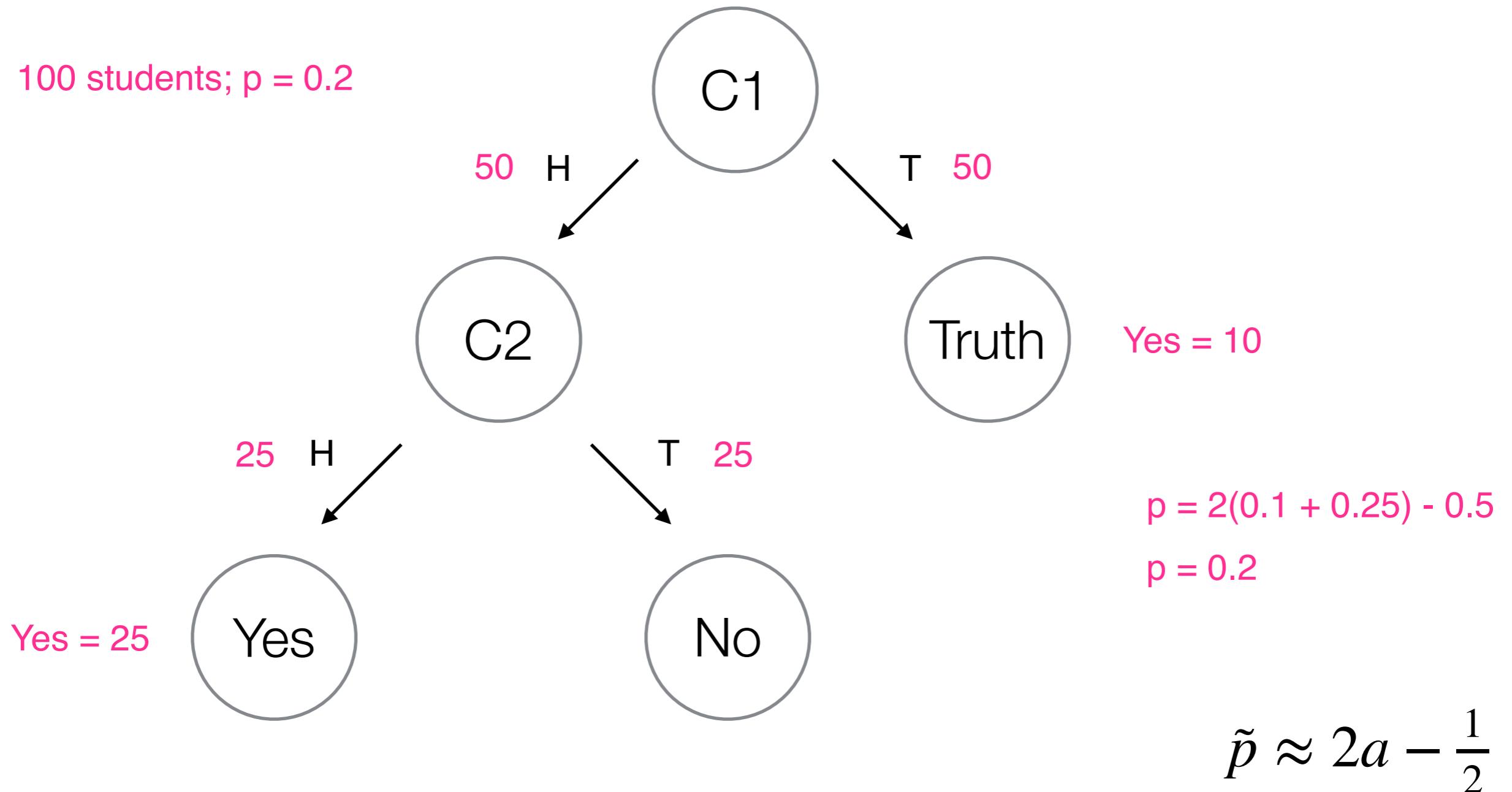
thus, we estimate **p** as:

$$\tilde{p} = 2A - \frac{1}{2}$$

# Truth or dare?

**Did you go out drinking over the weekend?**

100 students;  $p = 0.2$



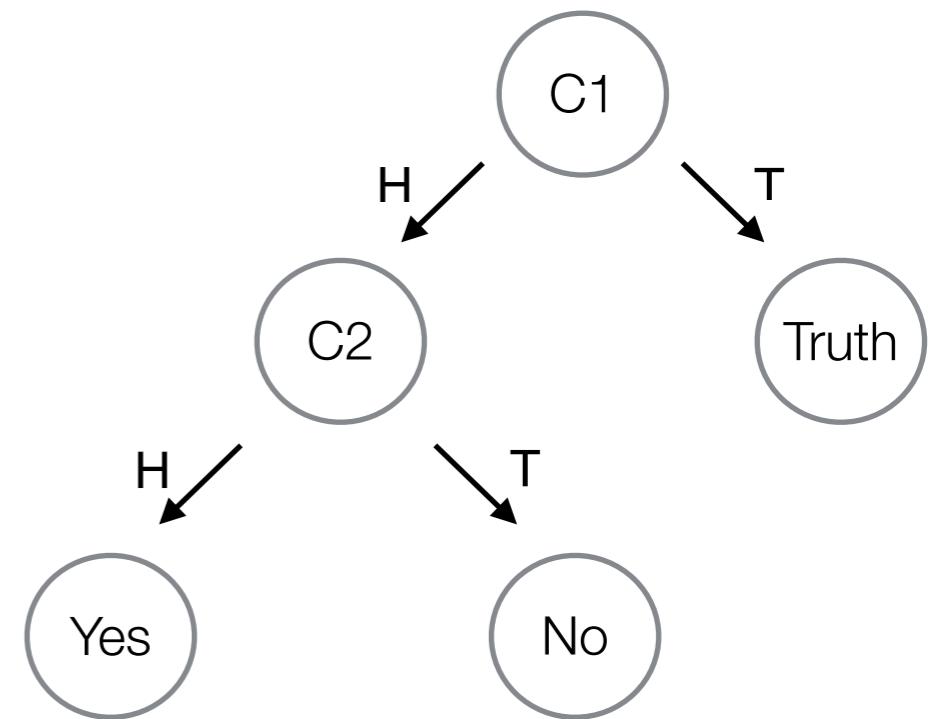
# Truth or dare?

**Did you go out drinking over the weekend?**

Student	Response
A	no
B	yes
C	no
D	no
E	no
F	yes
G	no
H	no
I	no
...	...

T or HH?

T or HH?



# Randomized response

**Did you go out drinking over the weekend?**

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the class for whom **P** holds

1. flip a coin **C1**

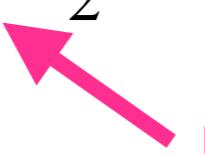
1. if **C1** is tails, then **respond truthfully**
2. if **C1** is heads, then flip another coin **C2**
  1. if **C2** is heads then **Yes**
  2. else **C2** is tails then respond **No**



randomization - adding noise - is what gives plausible deniability a process privacy method

the expected number of **Yes** answers is:

$$A = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{p}{2}$$



privacy comes from plausible deniability

# Privacy: two sides of the coin

protecting an individual  
plausible deniability

---



learning about the population  
noisy estimates

---



do we really  
need  
randomization?

# Some other options

- Data release approaches that fail to protect privacy (these are prominent classes of methods, there are others):
  - **sampling** (“just a few”) - release a small subset of the database
  - **aggregation** (e.g., **k-anonymity** - each record in the release is indistinguishable from at least  $k-1$  other records)
  - **de-identification** - mask or drop personal identifiers
  - **query auditing** - stop answering queries when they become unsafe

# Sampling (“just a few”)

- Suppose that we take a random small sample  $\mathbf{D}'$  of  $\mathbf{D}$  and release it without any modification
- If  $\mathbf{D}'$  is much smaller than  $\mathbf{D}$ , then every respondent is unlikely to appear in  $\mathbf{D}'$
- This technique provides protection for “the typical” (or for “most”) members of the dataset
- But it may be argued that **atypical** individuals are the ones needing stronger protection!
- In any case, this method is problematic because a respondent who does appear has **no plausible deniability!**
- Suppose next that appearing in the sample  $\mathbf{D}'$  has terrible consequences. Then, every time subsampling occurs - some individual suffers horribly!

# Aggregation without randomization

- Alice and Bob are professors at State University.
- In March, Alice publishes an article: “.... the current freshman class at State U is **3,005** students, **202** of whom are from families earning over \$1M per year.”
- In April, Bob publishes an article: “... **201** families in State U’s freshman class of **3,004** have household incomes exceeding \$1M per year.”
- Neither statement discloses the income of the family of any one student. But, taken together, they state that **John, a student who dropped out at the end of March**, comes from a family that earns \$1M. Anyone who has this **auxiliary information** — that John dropped out at the end of March — will be able to learn about the income of John’s family.

this is known as a problem of **composition**, and can be seen as a kind of a **differencing attack**

# A basic differencing attack

- **X**: count the number of HIV-positive people in  $D$
- **Y**: count the number of HIV-positive people in  $D$  not named *Freddie*;
- **X - Y** tells you whether *Freddie* is HIV-positive

what if  $X-Y > 1$ , do we still have a problem?

# Reconstruction: death by a 1000 cuts

- Another serious issue for aggregation without randomization, or with an insufficient amount of randomization: **reconstruction attacks**
- **The Fundamental Law of Information Recovery** (starting with the seminal results by Irit Dinur & Kobbi Nissim, PODS 2003): overly accurate estimates of too many statistics can completely destroy privacy
- Under what conditions can an adversary reconstruct a candidate database  $\mathbf{D}'$  that agrees with the real database  $\mathbf{D}$  in **99%** of the entries?
- Suppose that  $\mathbf{D}$  has  $n$  tuples, and that noise is bounded by some quantity  $E$ . Then there exists an adversary that can reconstruct  $\mathbf{D}$  to within  $4E$  positions, issuing all possible  $2^n$  queries

$$4E = \frac{4n}{401} < \frac{n}{100}$$

- Put another way: if the magnitude of the noise is less than  $n/401$ , then 99% of  $\mathbf{D}$  can be reconstructed by the adversary. Really, any number higher than 401 will work
- **There are also reconstruction results under a limited number of queries**

# Reconstruction: death by a 1000 cuts

## Privacy-Preserving Data Analysis for the Federal Statistical Agencies

January 2017



*John Abowd, Lorenzo Alvisi, Cynthia Dwork, Sampath Kannan, Ashwin Machanavajjhala, and Jerome Reiter*

**we'll discuss the use  
of differential privacy  
by the 2020 US  
Census later today**

The Fundamental Law of Information Recovery has troubling implications for the publication of large numbers of statistics by a statistical agency: it says that the confidential data may be vulnerable to database reconstruction attacks based entirely on the data published by the agency itself. **Left unattended, such risks threaten to undermine, or even eliminate, the societal benefits inherent in the rich data collected by the nation's statistical agencies.** The most pressing immediate problem for any statistical agency is how to modernize its disclosure limitation methods in light of the Fundamental Law.

# De-identification

- Also known as **anonymization**
- Mask or drop identifying attribute or attributes, such as social security number (SSN), name, mailing address
- Turns out that this also doesn't work because **auxiliary information** is available
- Fundamentally, this is due to **the curse of dimensionality**: high-dimensional data is sparse, the more you know about individuals, the less likely it is that two individuals will look alike

**de-identified data can be re-identified with a linkage attack**

# A linkage attack: Governor Weld

In 1997, Massachusetts Group Insurance Commission released "anonymized" data on state employees that showed every single hospital visit!

She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes.

Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code.

Latanya Sweeney, a grad student, sought to show the ineffectiveness of this "anonymization."

For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter.

*Follow up: ZIP code, birthdate, and sex sufficient to identify 87% of Americans!*

<https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/>

# The Netflix prize linkage attack

- In 2006, Netflix released a dataset containing ~100M **movie ratings** by ~500K users (about 1/8 of the Netflix user base at the time)
- **FAQ:** “Is there any customer information in the dataset that should be kept private?”

*“No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy, which you can review here. Even if, for example, you knew all your own ratings and their dates you probably couldn’t identify them reliably in the data because only **a small sample** was included (less than one-tenth of our complete dataset) and that **data was subject to perturbation**. Of course, since you know all your own ratings that really isn’t a privacy problem is it?”*

**The real question:** How much does the adversary need to know about a Netflix subscriber to identify her record in the dataset, and thus learn her complete movie viewing history?

# The Netflix prize linkage attack

- Very little auxiliary information is needed to de-anonymize an average subscriber record from the Netflix Prize dataset
- **Perturbation, you say?** With 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records be uniquely identified in the dataset
- For 68%, two ratings and dates (with a 3-day error) are sufficient
- **Even without any dates, a substantial privacy breach occurs, especially when the auxiliary information consists of movies that are not blockbusters:** Two movies are no longer sufficient, but 84% of subscribers can be uniquely identified if the adversary knows 6 out of 8 moves outside the top 500

**We cannot assume a priori that any data is harmless!**

# The Netflix prize linkage attack

WIRED

An in-the-closet lesbian mother is suing Netflix for privacy invasion, alleging the movie rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its \$1 million contest to improve its recommendation system.

The suit known as [Doe v. Netflix \(.pdf\)](#) was filed in federal court in California on Thursday, alleging that Netflix violated fair-trade laws and a federal privacy law protecting video rental records, when it launched its popular contest in September 2006.

The suit seeks more than \$2,500 in damages for each of more than 2 million Netflix customers.

RYAN SINGEL SECURITY 12.17.09 04:29 PM

## NETFLIX SPILLED YOUR BROKEBACK MOUNTAIN SECRET, LAWSUIT CLAIMS



r/ai

# The Netflix prize linkage attack

WIRED

RYAN SINGEL SECURITY 03.12.10 02:48 PM

## NETFLIX CANCELS RECOMMENDATION CONTEST AFTER PRIVACY LAWSUIT



Netflix is canceling its second \$1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

r/ai

# Query auditing

- Monitor queries: each query is granted or denied depending on what other queries were answered in the past
- If this method were to work, it could be used to detect that a differencing attack is about to take place
- Unfortunately, it doesn't work:
  - **Query auditing is computationally infeasible**
  - Refusal to respond to a query may itself be disclosive
  - We refuse to execute a query, then what? No information access at all?

# Query auditing

- We have a set of (secret) Boolean variables  $\mathbf{X}$  and the result of some *statistical queries* over this set
- A *statistical query*  $\mathbf{Q}$  specifies a subset  $\mathbf{S}$  of the variables in  $\mathbf{X}$ , and returns the sum of the values of all variables in  $\mathbf{S}$

Example:

Relation Employees (name, age, salary)

Query `select sum(salary) from Employees where age > 35`

Suppose that Employees (name, age) is public, but salary is confidential

# Query auditing

- We have a set of (secret) Boolean variables  $\mathbf{X}$  and the result of some *statistical queries* over this set
- A *statistical query*  $\mathbf{Q}$  specifies a subset  $\mathbf{S}$  of the variables in  $\mathbf{X}$ , and returns the sum of the values of all variables in  $\mathbf{S}$
- **The auditing problem:** Decide whether the value of any Boolean variable is determined by the results of the queries
- **Main result:** The Boolean auditing problem is **coNP-complete**
  - coNP-complete is the hardest class of problems in coNP: all coNP problems can be formulated as a special case of any coNP-complete problem
  - if P does not equal NP, then there does not exist a polynomial time algorithm that solves this problem



privacy-  
preserving data  
analysis

# Privacy: two sides of the coin

protecting an individual  
plausible deniability

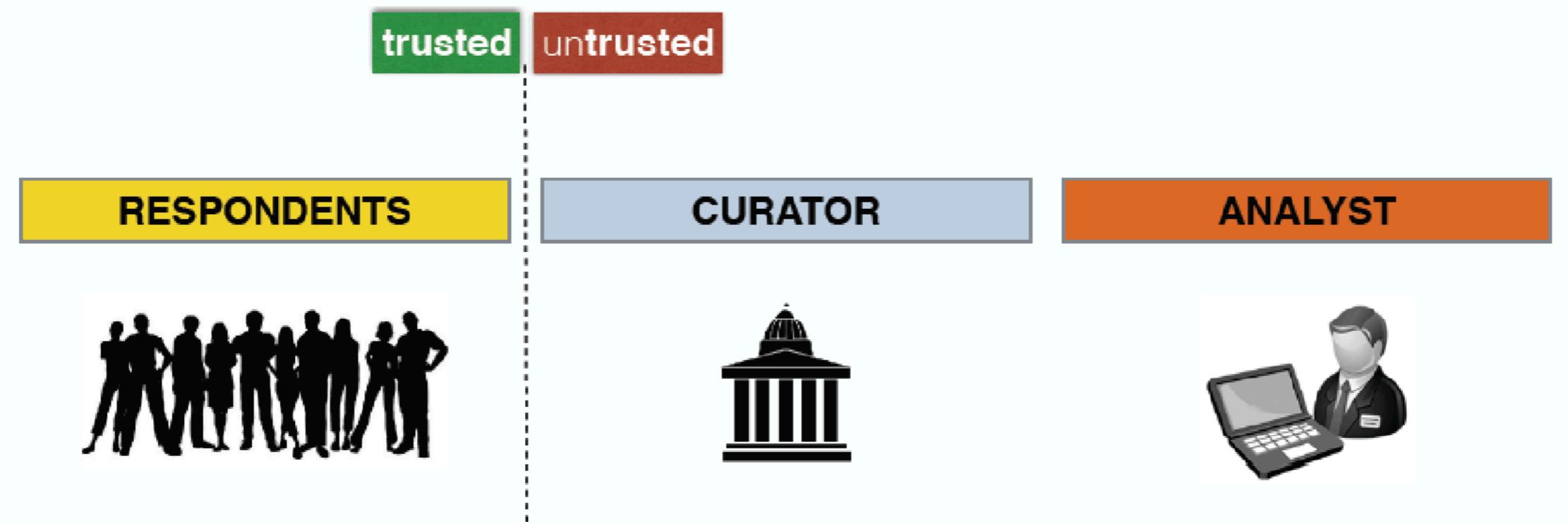
---



learning about the population  
noisy estimates

---

# Privacy-preserving data analysis

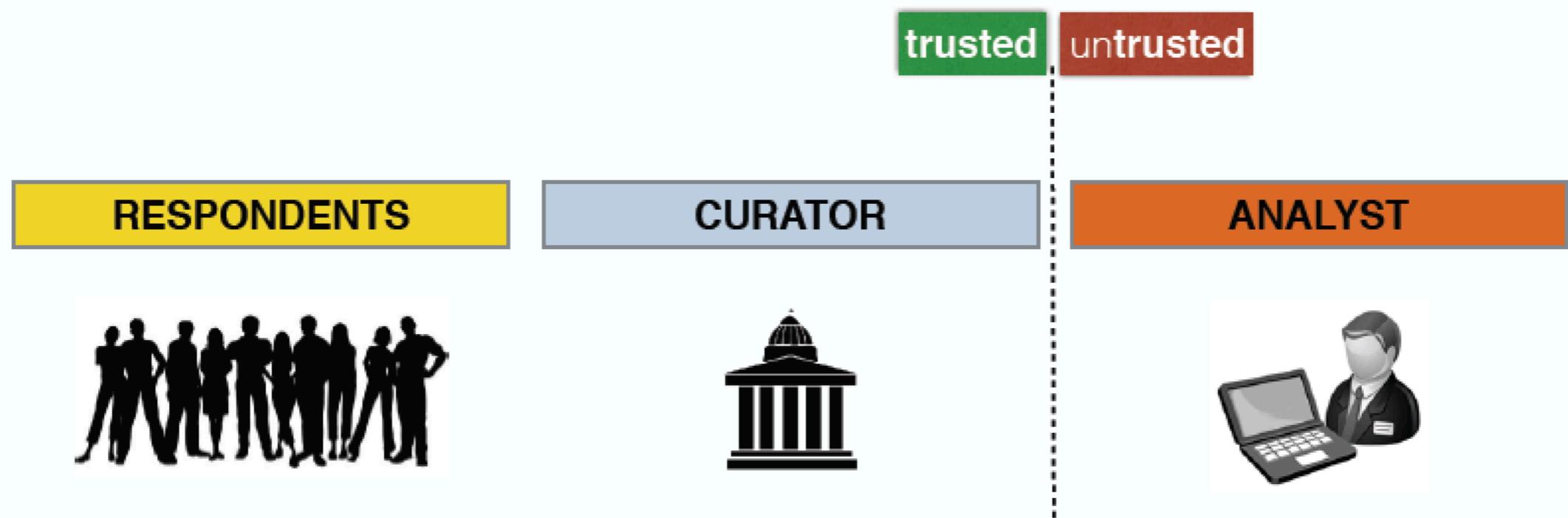


**respondents** contribute their personal data

the **curator** is **untrusted**, collects data, releases it to analysts

the **analyst** is **untrusted**, extracts value from data

# Privacy-preserving data analysis

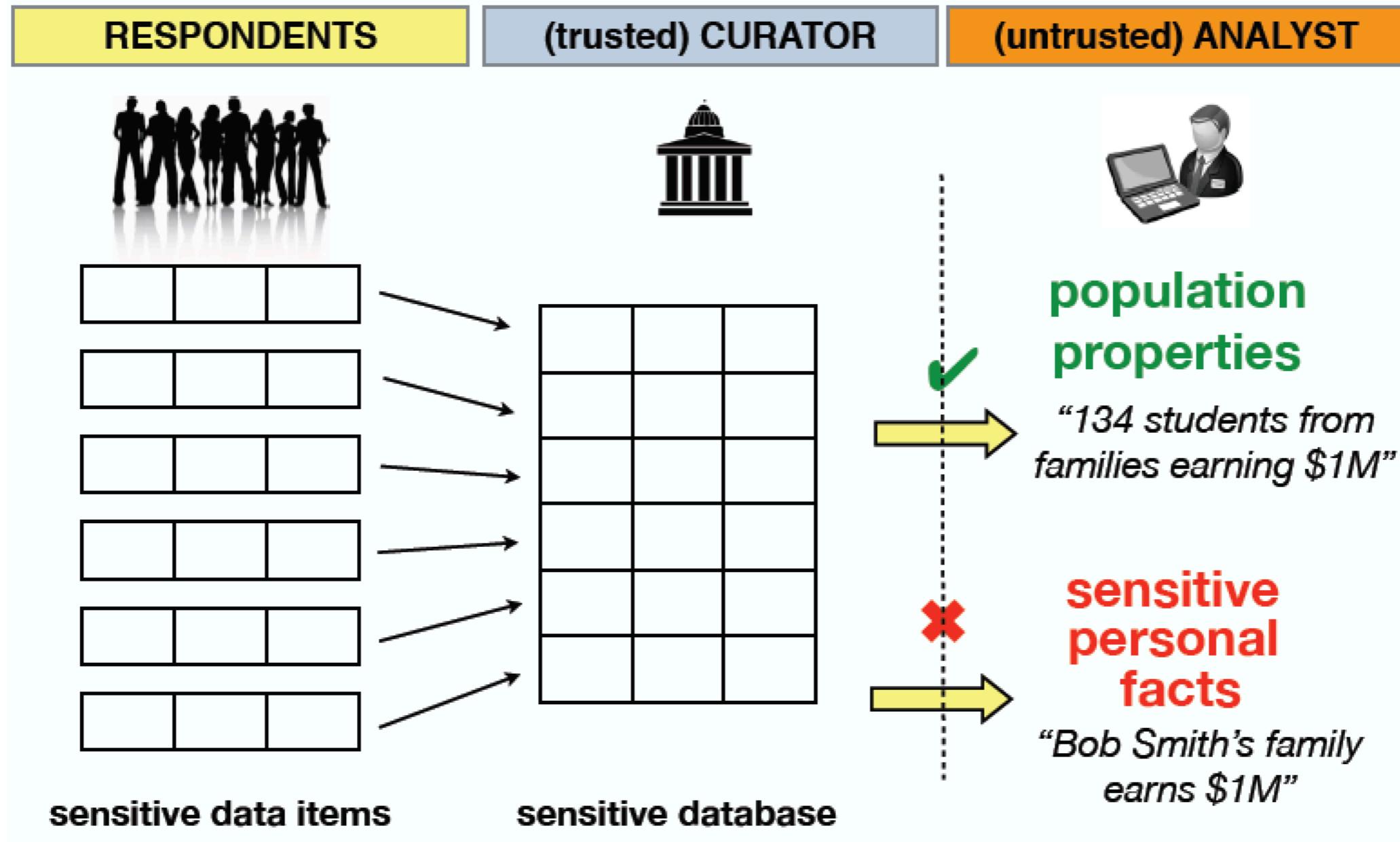


**respondents** in the population seek protection of their personal data

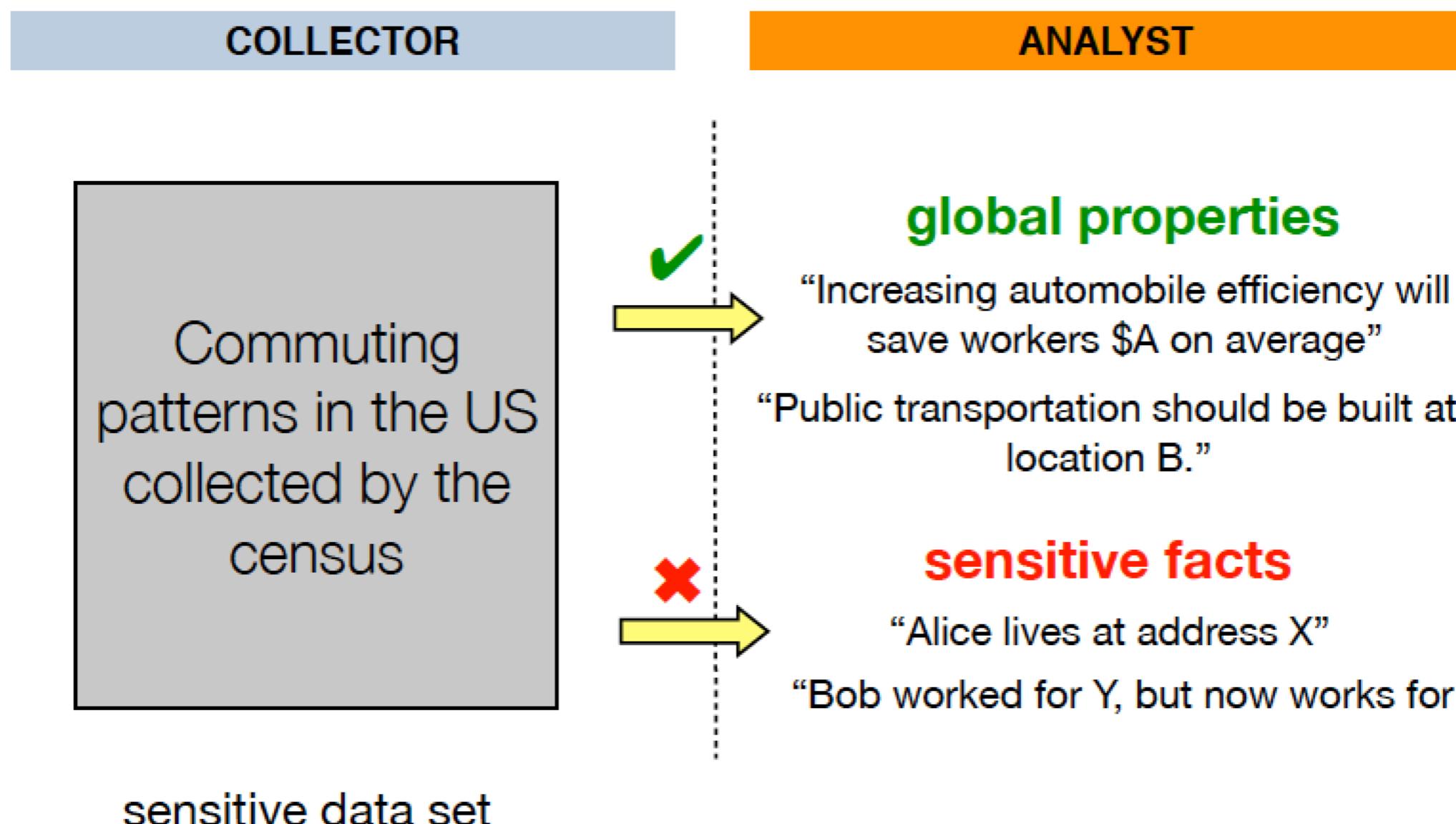
the **curator** is **trusted** to collect data and is responsible for safely releasing it

the **analyst** is **untrusted** and wants to gain the most accurate insights into the population

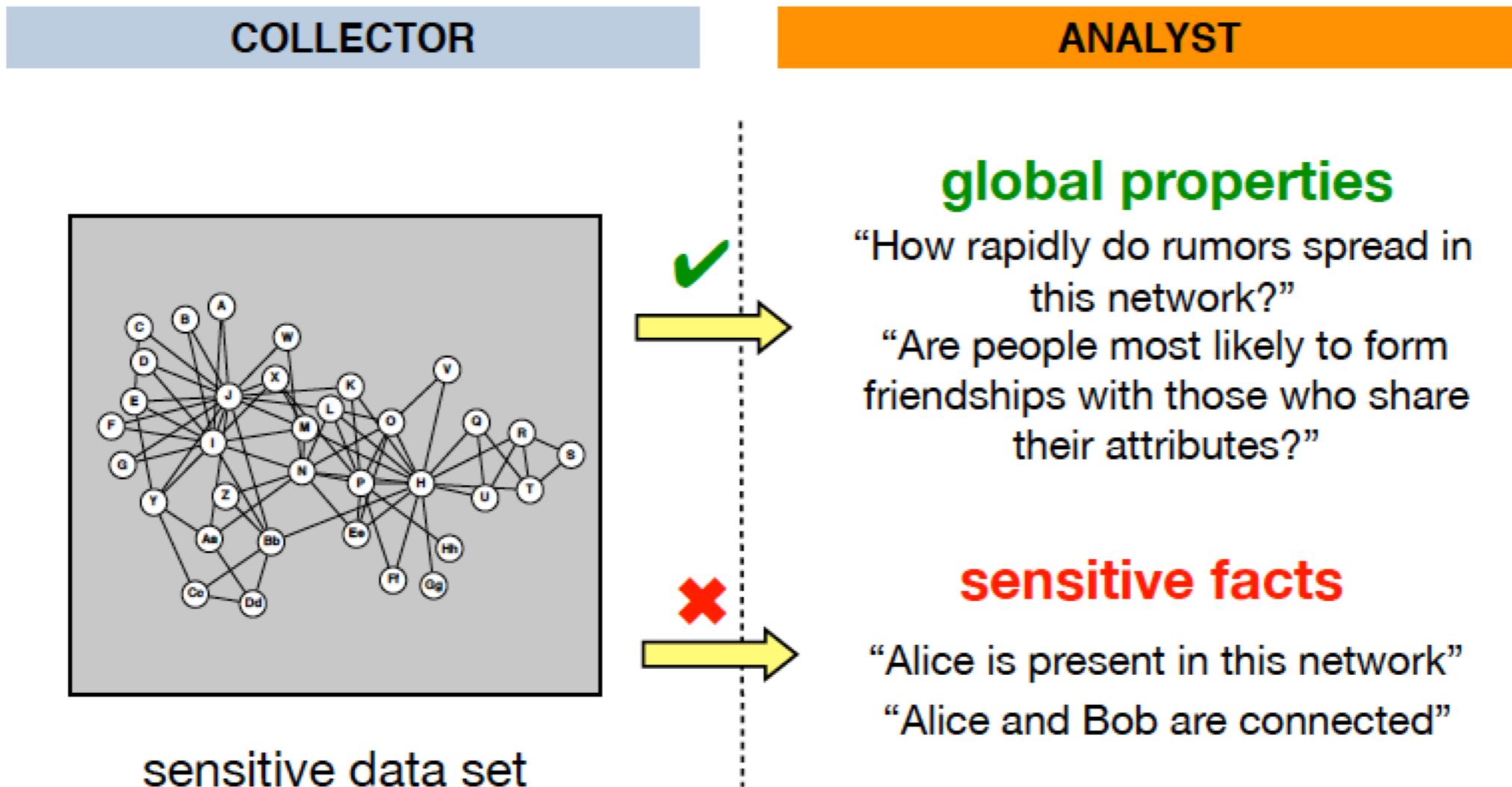
# Privacy-preserving data analysis



# Example: US Census



# Example: Social networks



# Defining private data analysis

- Take 1: If **nothing is learned** about any individual in the dataset, then no individual can be harmed by analysis.
  - **Dalenius' Desideratum:** an *ad omnia* (Latin: “for all”) privacy goal for statistical databases, as opposed to *ad hoc* (Latin: “for this”). Anything that can be learned about a respondent from the statistical database should be learnable without access to the database.
  - Put another way, the adversary’s prior and posterior views about an individual should not be different.
  - This objective is **unachievable** because of auxiliary information.
  - **Example:** Alice knows that John smokes. She read a medical research study that found a causal relationship between smoking and lung cancer. Alice concludes, based on study results and her prior knowledge about John, that he has a heightened risk of developing lung cancer.
  - Further, the risk is to everyone in a particular group (smokers, in this example), **irrespective of whether they participated in the study**.

# Defining private data analysis

- Take 1: If **nothing is learned** about any individual in the dataset, then no individual can be harmed by analysis.
  - **Dalenius' Desideratum:** an “*ad omnia*” (opposed to *ad hoc*) privacy goal for statistical databases: Anything that can be learned about a respondent from the statistical database should be learnable without access to the database.
  - Put another way, the adversary’s prior and posterior views about an individual should not be different.
- Take 2: The information released about the sensitive dataset is virtually indistinguishable **whether or not a respondent’s data is in the dataset**. This is an informal statement of **differential privacy**: that no information **specific to an individual** is revealed.

# Defining private data analysis

DOI:10.1145/1866739.1866758

## What does it mean to preserve privacy?

BY CYNTHIA DWORK

# A Firm Foundation for Private Data Analysis

IN THE INFORMATION realm, loss of privacy is usually associated with failure to control access to information, to control the flow of information, or to control the purposes for which information is employed. Differential privacy arose in a context in which ensuring privacy is a challenge even if all these control problems are solved: privacy-preserving statistical analysis of data.

The problem of *statistical disclosure control*—revealing accurate statistics about a set of respondents while preserving the privacy of individuals—has a venerable history, with an extensive literature spanning statistics, theoretical computer science, security, databases, and cryptography (see, for example, the excellent survey of Adam and Wortmann,<sup>1</sup> the discussion of related work in Blum et al.,<sup>2</sup> and the *Journal of Official Statistics* dedicated to confidentiality and disclosure control).

This long history is a testament to the importance of the problem. Statistical databases can be of enormous social value; they are used for apportioning resources, evaluating medical therapies, understanding the spread of disease, improving economic utility, and informing us about ourselves as a species.

The data may be obtained in diverse ways. Some data, such as census, tax, and other sorts of official data, is compelled; other data is collected opportunistically, for example, from traffic on the Internet, transactions on Amazon, and search engine query logs; other data is provided altruistically, by respondents who hope that sharing their information will help others to avoid a specific misfortune, or more generally, to increase the public good. Altruistic data donors are typically promised their individual data will be kept confidential—in short, they are promised “privacy.” Similarly, medical data and legally compelled data, such as census data and tax return data, have legal privacy.

### » key insights

- In analyzing private data, only by focusing on rigorous privacy guarantees can we convert the cycle of “propose-break-propose again” into a path of progress.
- A natural approach to defining privacy is to require that accessing the database teaches the analyst nothing about any individual. But this is problematic: the whole point of a statistical database is to teach general truths, for example, that smoking causes cancer. Learning this fact teaches the data analyst something about the likelihood with which certain individuals, not necessarily in the database, will develop cancer. We therefore need a definition that separates the utility of the database (learning that smoking causes cancer) from the increased risk of harm due to joining the database. This is the intuition behind differential privacy.
- This can be achieved, often with low distortion. The key idea is to randomize responses so as to effectively hide the presence or absence of the data of any individual over the course of the lifetime of the database.

“A natural approach to defining privacy is to require that accessing the database teaches the analyst nothing about any individual. But this is problematic: **the whole point of a statistical database is to teach general truths**, for example, that smoking causes cancer. Learning this fact teaches the data analyst something about the likelihood with which certain individuals, not necessarily in the database, will develop cancer. We therefore **need a definition that separates the utility of the database** (learning that smoking causes cancer) **from the increased risk of harm due to joining the database. This is the intuition behind differential privacy.**”



differential  
privacy (DP)

# Differential privacy: the formalism

We will define privacy with respect to a database  $\mathbf{D}$  that is made up of rows (equivalently, tuples) representing individuals. Tuples come from some universe of datatypes (the set of all possible tuples).

The  $\ell_1$  norm of a database  $\mathbf{D}$ , denoted  $\|\mathbf{D}\|_1$  is the number of tuples in  $\mathbf{D}$ .

The  $\ell_1$  distance between databases  $\mathbf{D}_1$  and  $\mathbf{D}_2$  represents the number of tuples on which they differ.  $\|\mathbf{D}_1 - \mathbf{D}_2\|_1$

We refer to a pair of databases that differ in at most 1 tuple as **neighboring databases**

$$\|\mathbf{D}_1 - \mathbf{D}_2\|_1 \leq 1$$

Of these  $\mathbf{D}_1$  and  $\mathbf{D}_2$ , one, say  $\mathbf{D}_1$ , is a subset of the other, and, when a proper subset, the larger database  $\mathbf{D}_2$  contains 1 extra tuple.

# Differential privacy: the formalism

The information released about the sensitive dataset is virtually indistinguishable **whether or not a respondent's data is in the dataset**. This is an informal statement of **differential privacy**. That is, no information **specific to an individual** is revealed.

A randomized algorithm  $M$  provides  **$\epsilon$ -differential privacy** if, for all neighboring databases  $D_1$  and  $D_2$ , and for any set of outputs  $S$ :

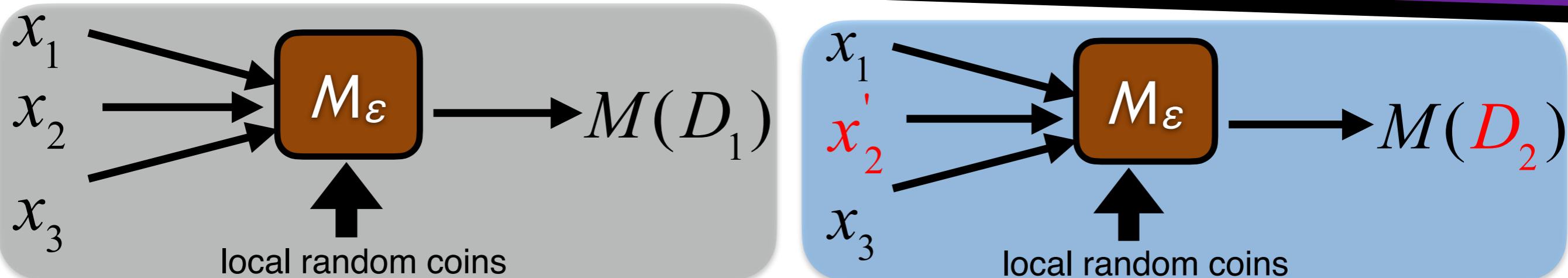
$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S]$$

**$\epsilon$  (epsilon) is a privacy parameter**

 **lower  $\epsilon$  = stronger privacy** 

The notion of **neighboring databases** is integral to plausible deniability:  $D_1$  can represent a database with a particular respondent's data,  $D_2$  can represent a neighboring database but without that respondent's data

# Differential privacy: the formalism



A randomized algorithm  $\mathbf{M}$  provides  **$\epsilon$ -differential privacy** if, for all neighboring databases  $\mathbf{D}_1$  and  $\mathbf{D}_2$ , and for any set of outputs  $\mathbf{S}$ :

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S]$$

Think of database of respondents  $\mathbf{D}=(\mathbf{x}_1, \dots, \mathbf{x}_n)$  as **fixed** (not random),  $\mathbf{M}(\mathbf{D})$  is a random variable distributed over possible outputs

**Neighboring databases** induce **close distributions** on outputs

# Back to randomized response

## Did you go out drinking over the weekend?

1. flip a coin **C1**

1. if **C1** is tails, then **respond truthfully**

2. if **C1** is heads, then flip another coin **C2**

1. if **C2** is heads then **Yes**

2. else **C2** is tails then respond **No**

Denote:

- Truth=Yes by **P**
- Response=Yes by **A**
- **C1**=tails by **T**
- **C1**=heads and **C2**=tails by **HT**
- **C1**=heads and **C2**=heads by **HH**

A randomized algorithm **M** provides  **$\epsilon$ -differential privacy** if, for all neighboring databases **D<sub>1</sub>** and **D<sub>2</sub>**, and for any set of outputs **S**:

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S]$$

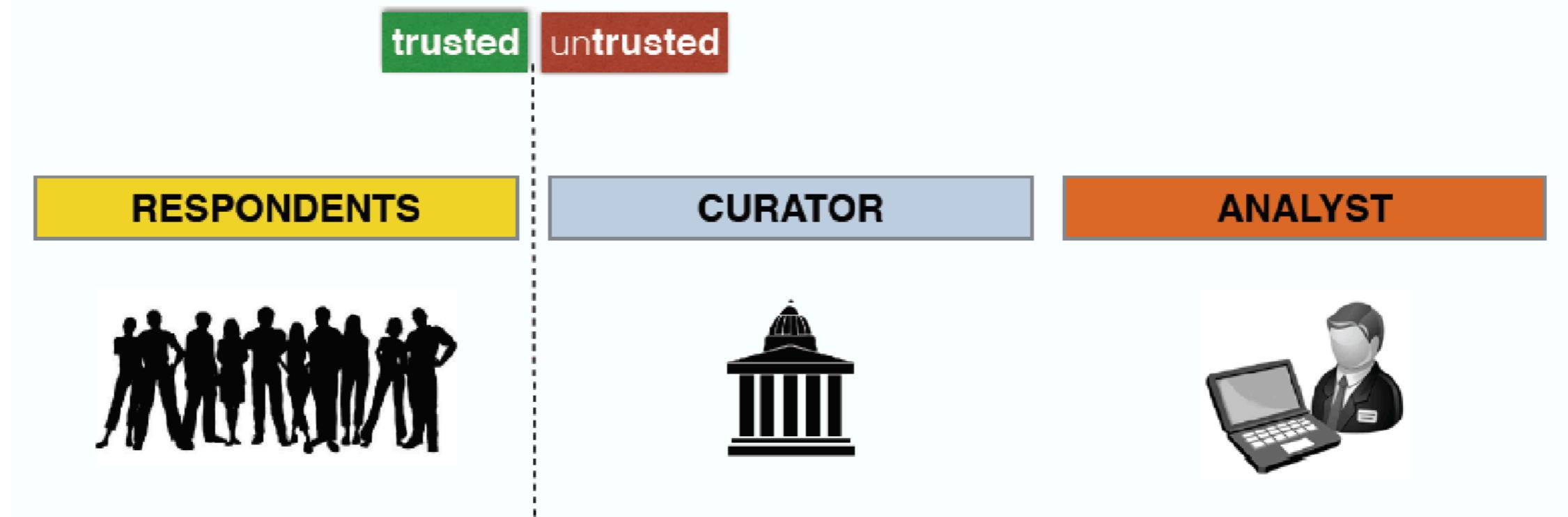
$$\Pr[A | P] = \Pr[T] + \Pr[HH] = \frac{3}{4}$$

$$\Pr[A | \neg P] = \Pr[HH] = \frac{1}{4}$$

$$\begin{aligned}\Pr[A | P] &= 3 \Pr[A | \neg P] \\ \Rightarrow \epsilon &= \ln 3\end{aligned}$$

our version of randomized response is  
( $\ln 3$ )-differentially private

# Local differential privacy

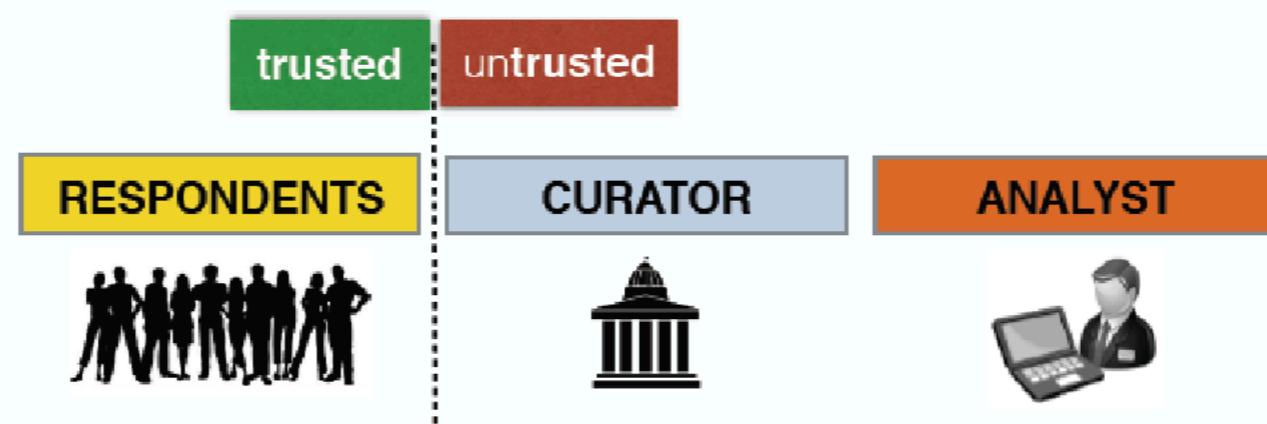
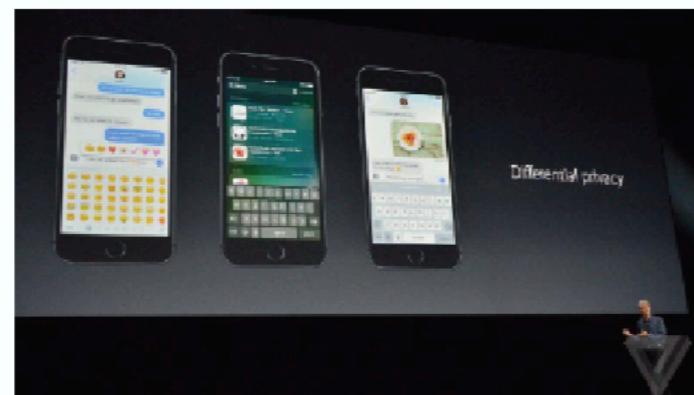


**respondents** contribute their personal data

the **curator** is **untrusted**, collects data, releases it to analysts

the **analyst** is **untrusted**, extracts value from data

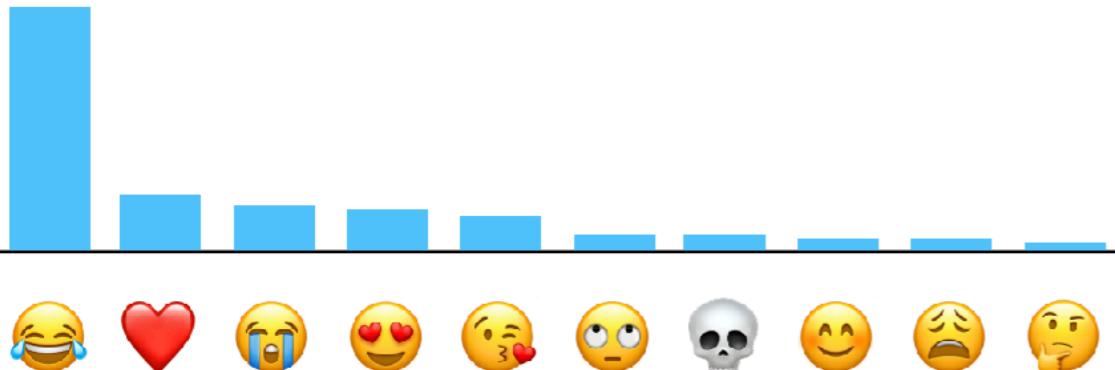
# Differential privacy in the field



# Example: What's your favorite emoji?

## A privacy-preserving system

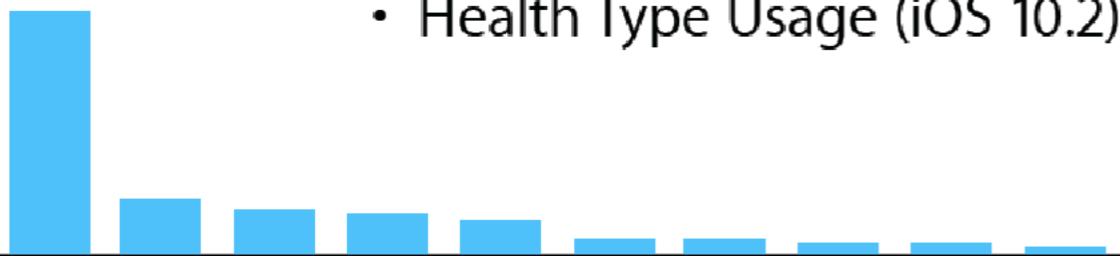
Apple has adopted and further developed a technique known in the academic world as *local differential privacy* to do something really exciting: gain insight into what many Apple users are doing, while helping to preserve the privacy of individual users. It is a technique that enables Apple to learn about the user community without learning about individuals in the community. Differential privacy transforms the information shared with Apple before it ever leaves the user's device such that Apple can never reproduce the true data.



# Example: What's your favorite emoji?

Apple uses local differential privacy to help protect the privacy of user activity in a given time period, while still gaining insight that improves the intelligence and usability of such features as:

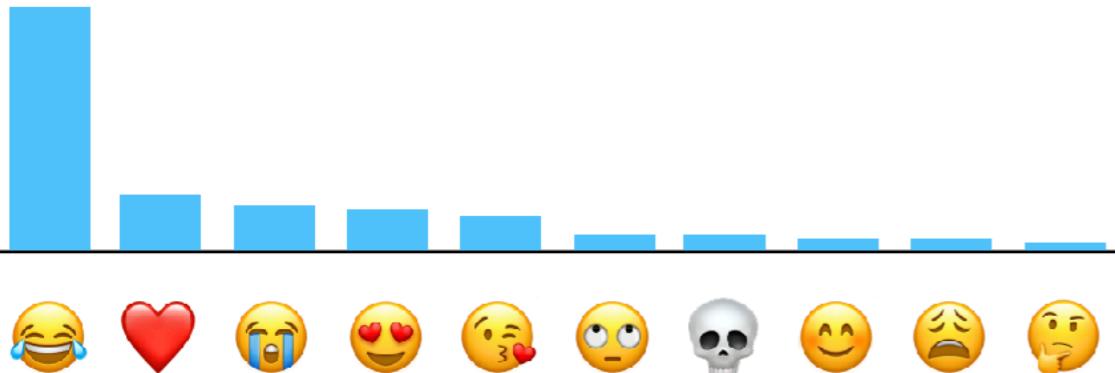
- QuickType suggestions
- Emoji suggestions
- Lookup Hints
- Safari Energy Draining Domains
- Safari Autoplay Intent Detection (macOS High Sierra)
- Safari Crashing Domains (iOS 11)
- Health Type Usage (iOS 10.2)



# Example: What's your favorite emoji?

## Privacy budget

The Apple differential privacy implementation incorporates the concept of a per-donation *privacy budget* (quantified by the parameter epsilon), and sets a strict limit on the number of contributions from a user in order to preserve their privacy. The reason is that the slightly-biased noise used in differential privacy tends to average out over a large numbers of contributions, making it theoretically possible to determine information about a user's activity over a large number of observations from a single user (though it's important to note that Apple doesn't associate any identifiers with information collected using differential privacy).



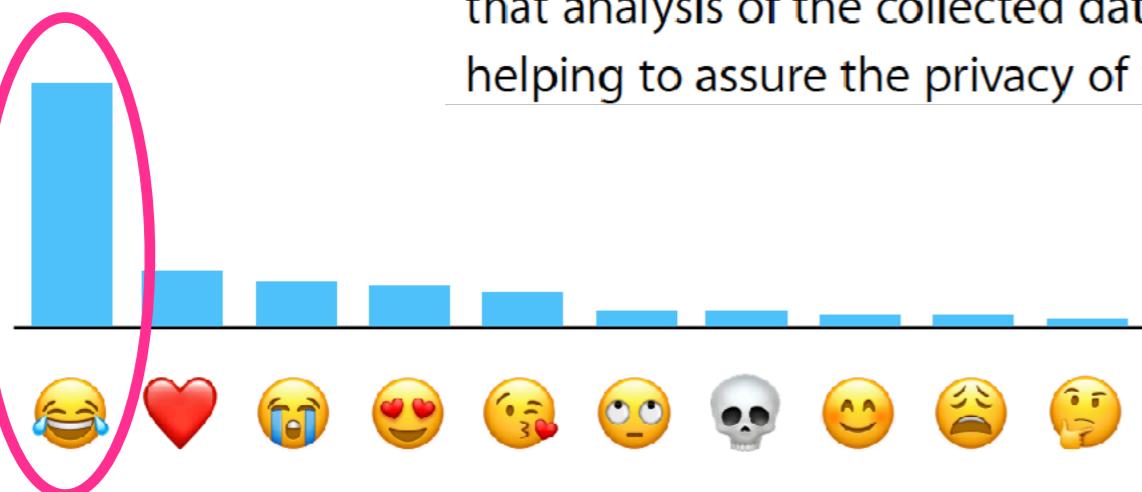
# Example: What's your favorite emoji?

## Count Mean Sketch

In our use of the Count Mean Sketch technique for differential privacy, the original information being processed for sharing with Apple is encoded using a series of mathematical functions known as *hash functions*, making it easy to represent data of varying sizes in a matrix of fixed size.

The data is encoded using variations of a SHA-256 hash followed by a privatization step and then written into the sketch matrix with its values initialized to zero.

The noise injection step works as follows: After encoding the input as a vector using a hash function, each coordinate of the vector is then flipped (written as an incorrect value) with a probability of  $1/(1 + e^{\epsilon/2})$ , where  $\epsilon$  is the privacy parameter. This assures that analysis of the collected data cannot distinguish actual values from flipped values, helping to assure the privacy of the shared information.



# Transparency is important!

ANDY GREENBERG

SECURITY 09.15.2017 09:28 AM

= WIRED

## How One of Apple's Key Privacy Safeguards Falls Short

Apple has boasted of its use of a cutting-edge data science known as "differential privacy." Researchers say they're doing it wrong.

"...[Researchers] examined how Apple's software injects random noise into personal information—ranging from emoji usage to your browsing history to HealthKit data to search queries—before your iPhone or MacBook upload that data to Apple's servers.

Ideally, that obfuscation helps protect your private data from any hacker or government agency that accesses Apple's databases, advertisers Apple might someday sell it to, or even Apple's own staff. But **differential privacy's effectiveness depends on a variable known as the "privacy loss parameter," or "epsilon,"** which determines just how much specificity a data collector is willing to sacrifice for the sake of protecting its users' secrets. By taking apart Apple's software to determine the epsilon the company chose, the researchers found that **MacOS uploads significantly more specific data than the typical differential privacy researcher might consider private.** iOS 10 uploads even more. And perhaps most troubling, according to the study's authors, is that **Apple keeps both its code and epsilon values secret**, allowing the company to potentially change those critical variables and erode their privacy protections with little oversight...."

Epsilon, Epsilon

# A closer look at differential privacy

A randomized algorithm  $M$  provides  **$\epsilon$ -differential privacy** if, for all neighboring databases  $D_1$  and  $D_2$ , and for any set of outputs  $S$ :

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S]$$

 **lower  $\epsilon$  = stronger privacy** 

- The state-of-the-art in privacy technology, first proposed in 2006
- Has precise mathematical properties, captures cumulative privacy loss over multiple uses with the concept of a **privacy budget**
- Privacy guarantee encourages participation by respondents
- Robust against strong adversaries, with auxiliary information, including also **future auxiliary information!**
- Precise error bounds that can be made public

# A closer look at differential privacy

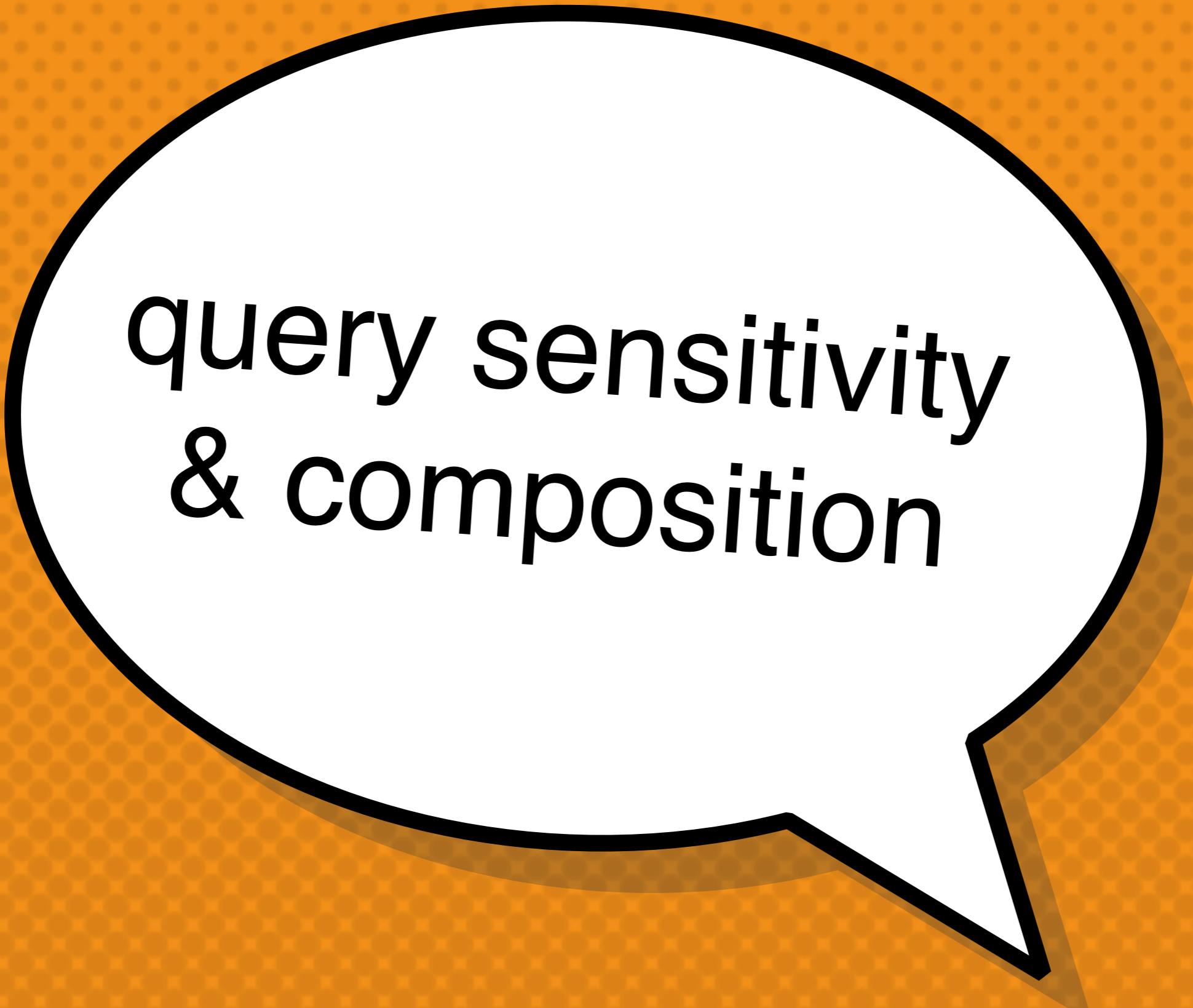
A randomized algorithm  $M$  provides  **$\epsilon$ -differential privacy** if, for all neighboring databases  $D_1$  and  $D_2$ , and for any set of outputs  $S$ :

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S]$$

 **lower  $\epsilon$  = stronger privacy** 

**$\epsilon$  (epsilon)** cannot be too small: think 1/10, not 1/2<sup>50</sup>

Differential privacy is a condition on the **algorithm M** (process privacy). Saying simply that “the output is safe” does not take into account how it was computed, and is insufficient.



query sensitivity  
& composition

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

 **lower  $\epsilon$  = stronger privacy** 

- Example 1: counting queries
  - “How many elements in  $D$  satisfy property  $P$ ?” **What’s  $\Delta q$  ?**
  - “What fraction of the elements in  $D$  satisfy property  $P$ ?”
- Example 2: max / min
  - “What is the maximum employee salary in  $D$ ?” **What’s  $\Delta q$  ?**

**Intuition: for a given  $\epsilon$ , the higher the sensitivity, the more noise we need to add to meet the privacy guarantee**

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query  $q$

query sensitivity  $\Delta q$

select count(\*) from D

1

select count(\*) from D  
where sex = Male and age > 30

?

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query $q$	query sensitivity $\Delta q$
select count(*) from D	1
select count(*) from D where sex = Male and age > 30	1
select MAX(salary) from D	?

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query $q$	query sensitivity $\Delta q$
select count(*) from D	1
select count(*) from D where sex = Male and age > 30	1
select MAX(salary) from D	$MAX(salary)-MIN(salary)$
select gender, count(*) from D group by gender	?

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query $q$	query sensitivity $\Delta q$
select count(*) from D	1
select count(*) from D where sex = Male and age > 30	1
select MAX(salary) from D	$MAX(salary)-MIN(salary)$
select gender, count(*) from D group by gender	1 (disjoint groups, presence or absence of one tuple impacts only one of the counts)

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query  $q$

select gender, count(\*)  
from D group by gender

query sensitivity  $\Delta q$

**1 (disjoint groups)**, presence or  
absence of one tuple impacts only one  
of the counts)

an arbitrary list of  $m$  counting  
queries

?

# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query  $q$

select gender, count(\*)  
from D group by gender

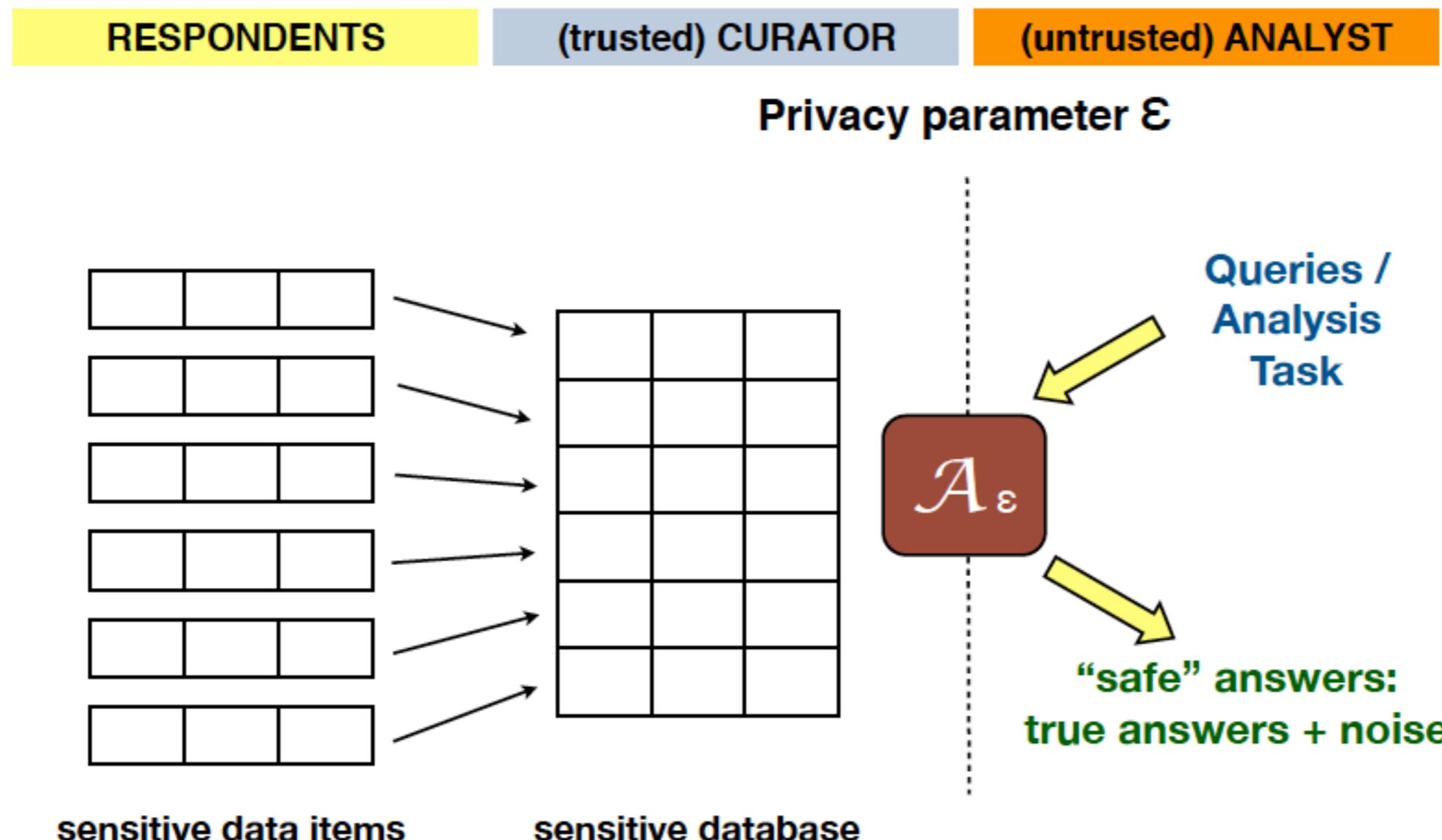
query sensitivity  $\Delta q$

**1 (disjoint groups)**, presence or absence of one tuple impacts only one of the counts)

an arbitrary list of  $m$  counting queries

**$m$**  (no assumptions about the queries, and so a single individual may change the answer of **every query** by 1)

# Adding noise

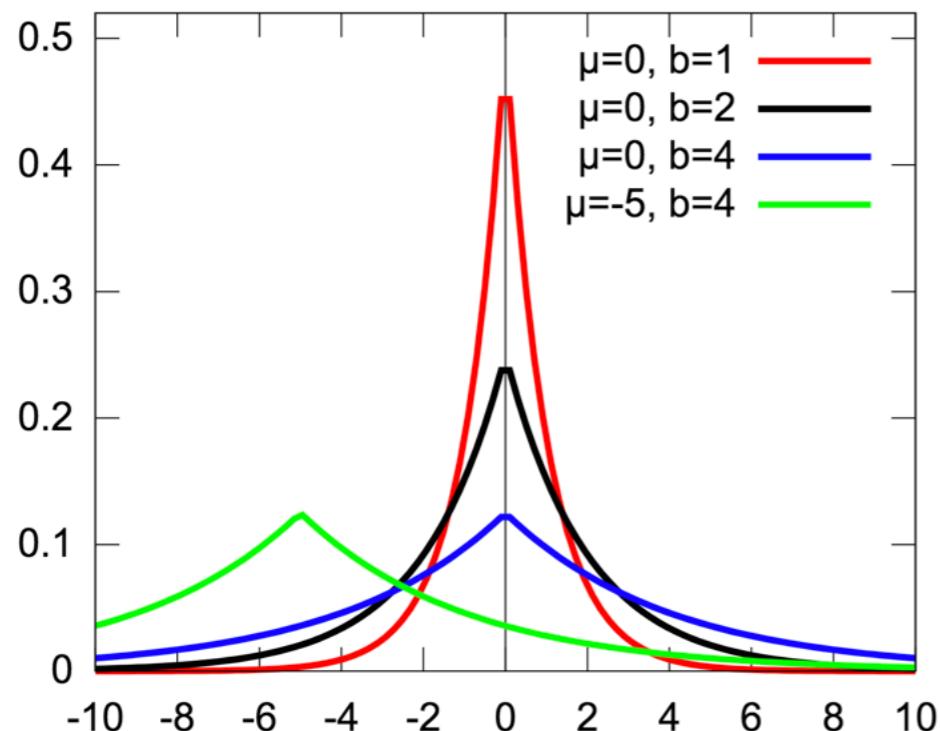


# Adding noise

Use the **Laplace mechanism** to answer  $\mathbf{q}$  in a way that's  $\epsilon$ -differentially private

$$M(\epsilon) : q(D) + \text{Lap}\left(\frac{\Delta q}{\epsilon}\right)$$

The Laplace distribution, centered at 0 with scale  $\mathbf{b}$ , denoted **Lap( $\mathbf{b}$ )**, is the distribution with probability density function:

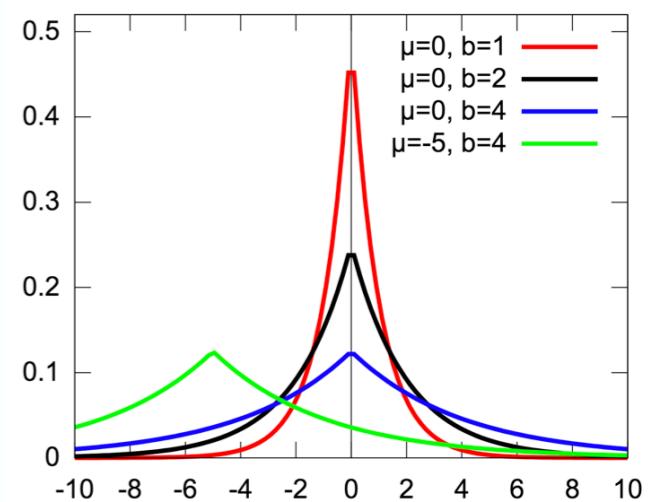
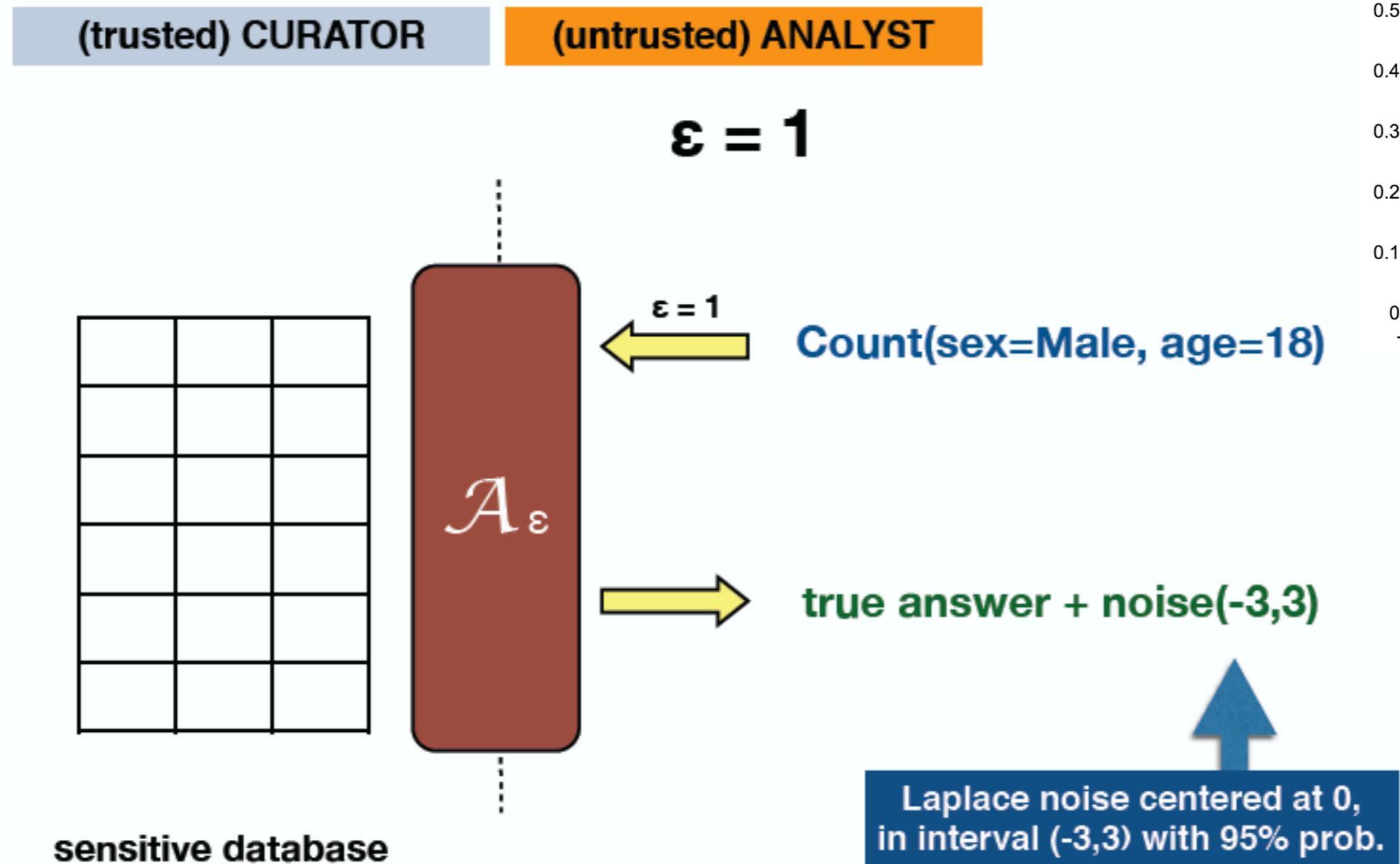


[https://en.wikipedia.org/wiki/Laplace\\_distribution](https://en.wikipedia.org/wiki/Laplace_distribution)

fix sensitivity  $\Delta q$ , verify that more noise is added for lower  $\epsilon$

lower  $\epsilon$  = stronger privacy

# Adding noise



# Query sensitivity

The  $\ell_1$  sensitivity of a query  $q$ , denoted  $\Delta q$ , is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} |q(D) - q(D')|$$

query  $q$

select gender, count(\*)  
from D group by gender

query sensitivity  $\Delta q$

parallel composition

1 (**disjoint groups**, presence or absence of one tuple impacts only one of the counts)

sequential composition

an arbitrary list of  $m$  counting queries

$m$  (no assumptions about the queries, and so a single individual may change the answer of **every query** by 1)

# Sequential composition

- Consider 4 queries executed in sequence
  - Q1: select count(\*) from D under  $\epsilon_1 = 0.5$
  - Q2: select count(\*) from D where sex = Male under  $\epsilon_2 = 0.2$
  - Q3: select count(\*) from D where sex = Female under  $\epsilon_3 = 0.25$
  - Q4: select count(\*) from D where age > 20 under  $\epsilon_4 = 0.25$
- $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 = 1.2$  That is: all queries together are  $\epsilon$ -differentially private for  $\epsilon = 1.2$ . **Can we make a stronger guarantee?**
- This works because **Laplace noise is additive**

More generally: set a **cumulative privacy budget**, and split it between all queries, pre-processing, other data manipulation steps of the pipeline

# Parallel composition

- If the inputs are disjoint, then the result is  $\epsilon$ -differentially private for  $\epsilon = \max(\epsilon_1, \dots, \epsilon_k)$ 
  - Q1: select count(\*) from D under  $\epsilon_1 = 0.5$
  - Q2: select count(\*) from D where sex = Male under  $\epsilon_2 = 0.2$
  - Q3: select count(\*) from D where sex = Female under  $\epsilon_3 = 0.25$
- Q4: select count(\*) from D where age > 20 under  $\epsilon_4 = 0.25$
- $\epsilon = \epsilon_1 + \max(\epsilon_2, \epsilon_3) + \epsilon_4 = 1$  That is: all queries together are  $\epsilon$ -differentially private for  $\epsilon = 1$ .

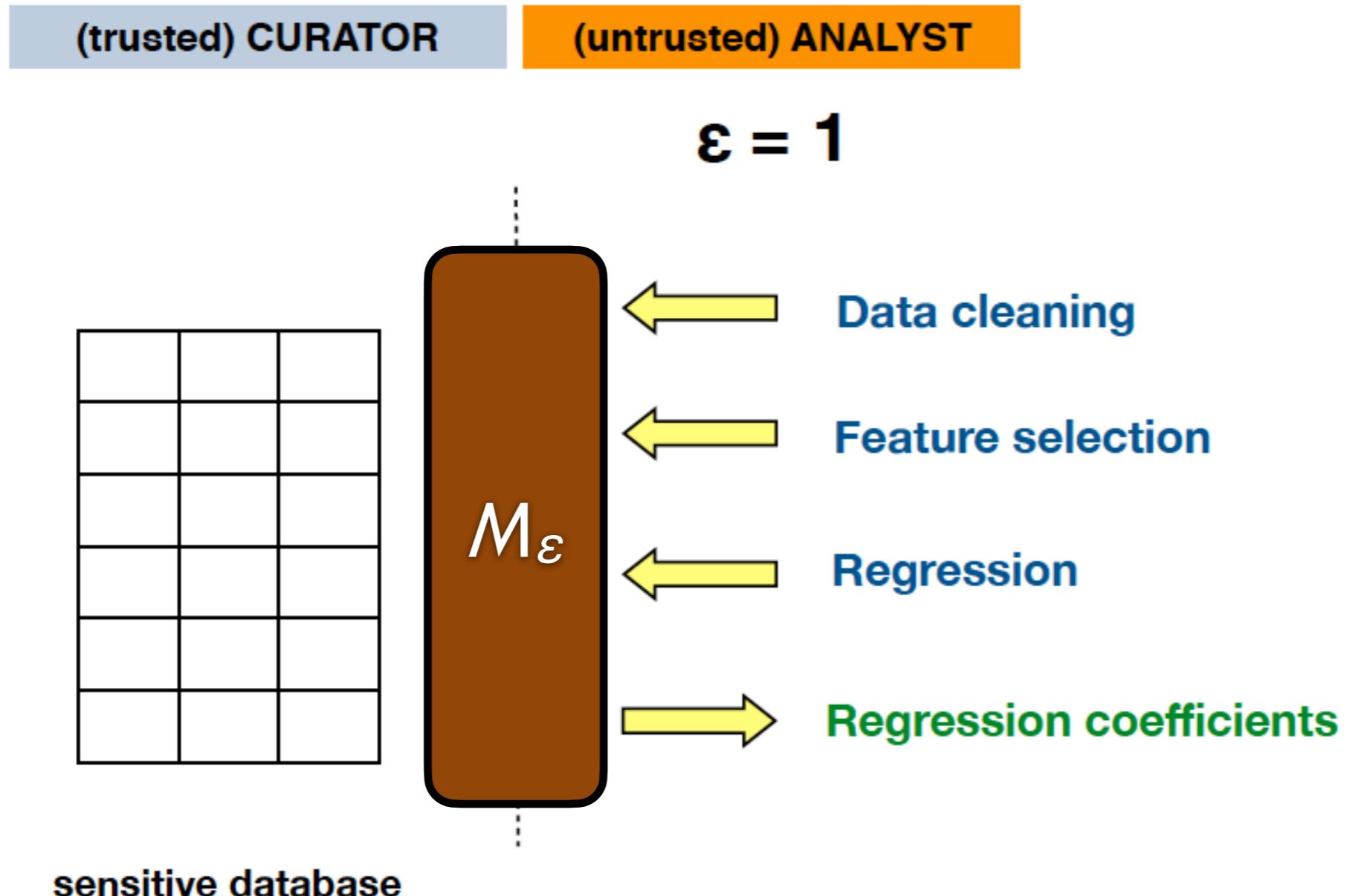
# Composition and consistency

- Consider again 4 queries executed in sequence
  - Q1: select count(\*) from D under  $\varepsilon_1 = 0.5$  returns **2005**
  - Q2: select count(\*) from D where sex = Male under  $\varepsilon_2 = 0.2$  returns **1001**
  - Q3: select count(\*) from D where sex = Female under  $\varepsilon_3 = 0.25$  returns **995**
  - Q4: select count(\*) from D where age > 20 under  $\varepsilon_4 = 0.25$  returns **1789**

Assuming that there are 2 genders in D, Male and Female, there is **no database consistent with these statistics!**

Also don't want any negative counts + may want to impose datatype checks, e.g., no working adults with age = 5 etc.

# Entire workflow must be DP



# Privacy-preserving synthetic data

