Responsible Data Science

Anonymity and privacy

March 20 & 27, 2023

Prof. Julia Stoyanovich

Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING
NYU Center for Data Science
r/ai

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

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,[1] the discussion of related work in Blum et al.,[2] 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.

Reading for weeks 7 & 8

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.

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.

CCS CONCEPTS

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

KEYWORDS

Data Sharing; Synthetic Data; Differential Privacy

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.

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 01002
e-mail address: rmckenna@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusetts, Amherst, MA 01002
e-mail address: miklau@cs.umass.edu

College of Information & Computer Sciences, The University of Massachusetts, Amherst, MA 01002
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.

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

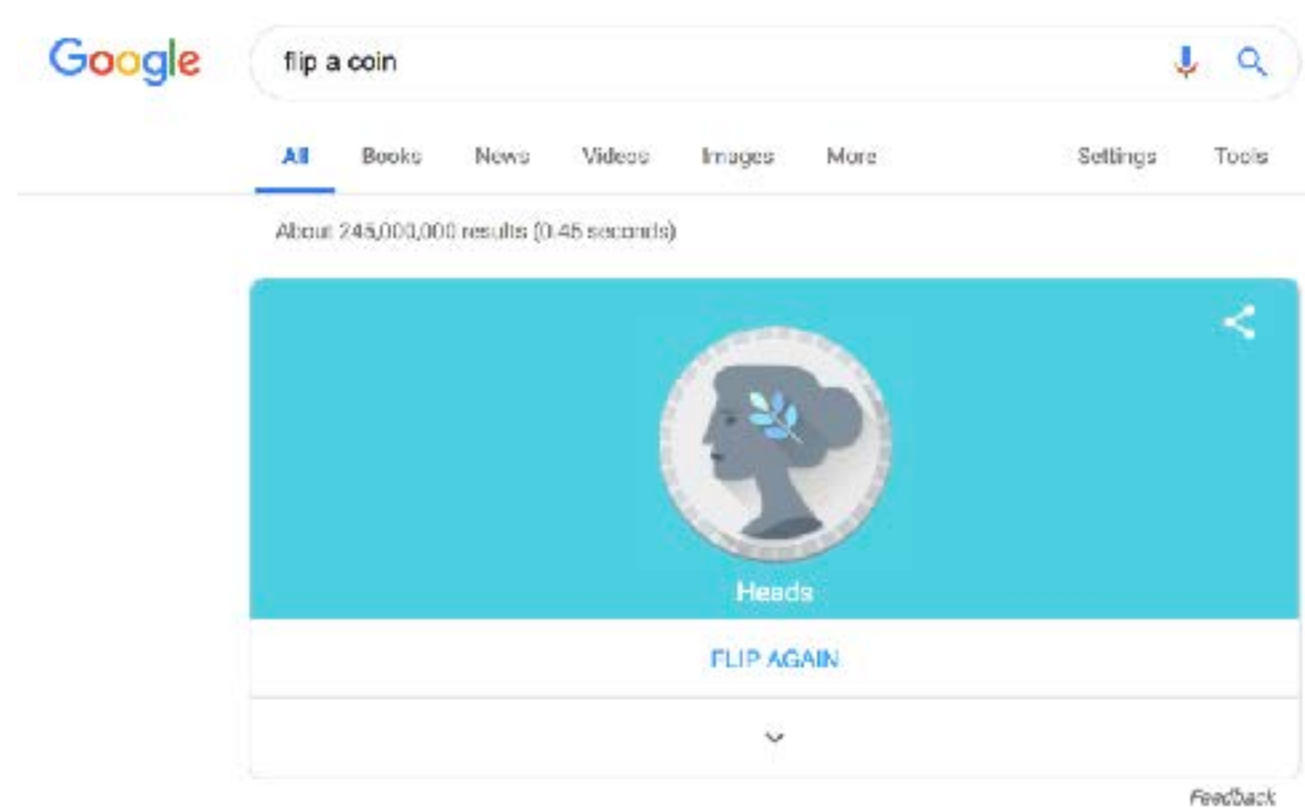
motivation

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:

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

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

r/ai

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 ***D’*** of ***D*** and release it without any modification
- If ***D’*** is much smaller than ***D***, then every respondent is unlikely to appear in ***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 ***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**

r/ai

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 ***D'*** that agrees with the real database ***D*** in **99%** of the entries?

- Suppose that ***D*** has ***n*** tuples, and that noise is bounded by some quantity ***E***. Then there exists an adversary that can reconstruct ***D*** to within ***4E*** positions, issuing all possible $\mathbf{2^n}$ 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 ***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

CCC
Computing Community Consortium
Catalyst

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

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.

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

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!

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

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

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.

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

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 Nexflix 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?

[Narayanan and Shmatikov, *IEEE S&P 2008*]

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!

[Narayanan and Shmatikov, *IEEE S&P 2008*]

r/ai

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.

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 ***X*** and the result of some *statistical queries* over this set

- A *statistical query* ***Q*** specifies a subset ***S*** of the variables in ***X***, and returns the sum of the values of all variables in ***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 ***X*** and the result of some *statistical queries* over this set
- A *statistical query* ***Q*** specifies a subset ***S*** of the variables in ***X***, and returns the sum of the values of all variables in ***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

[Kleinberg, Papadimitriou, Raghavan, *PODS 2000*]

r/ai

privacy-preserving data analysis

Privacy: two sides of the coin

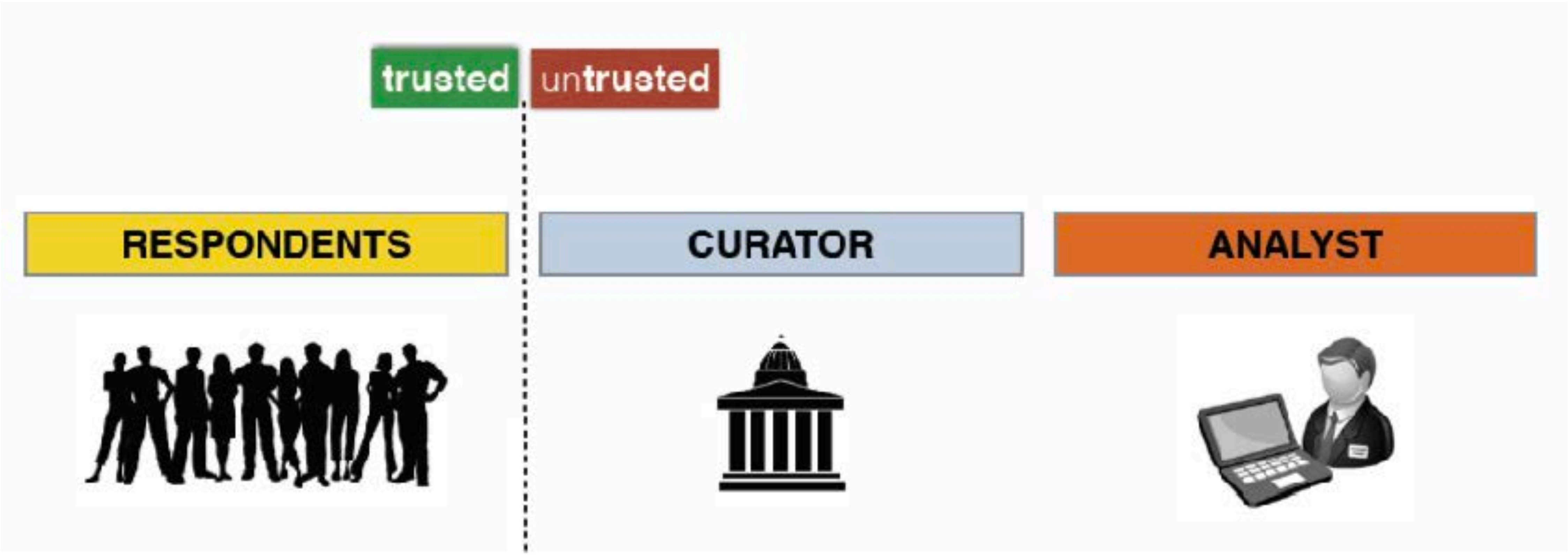
protecting an individual

plausible deniability

learning about the population

noisy estimates

Privacy-preserving data analysis

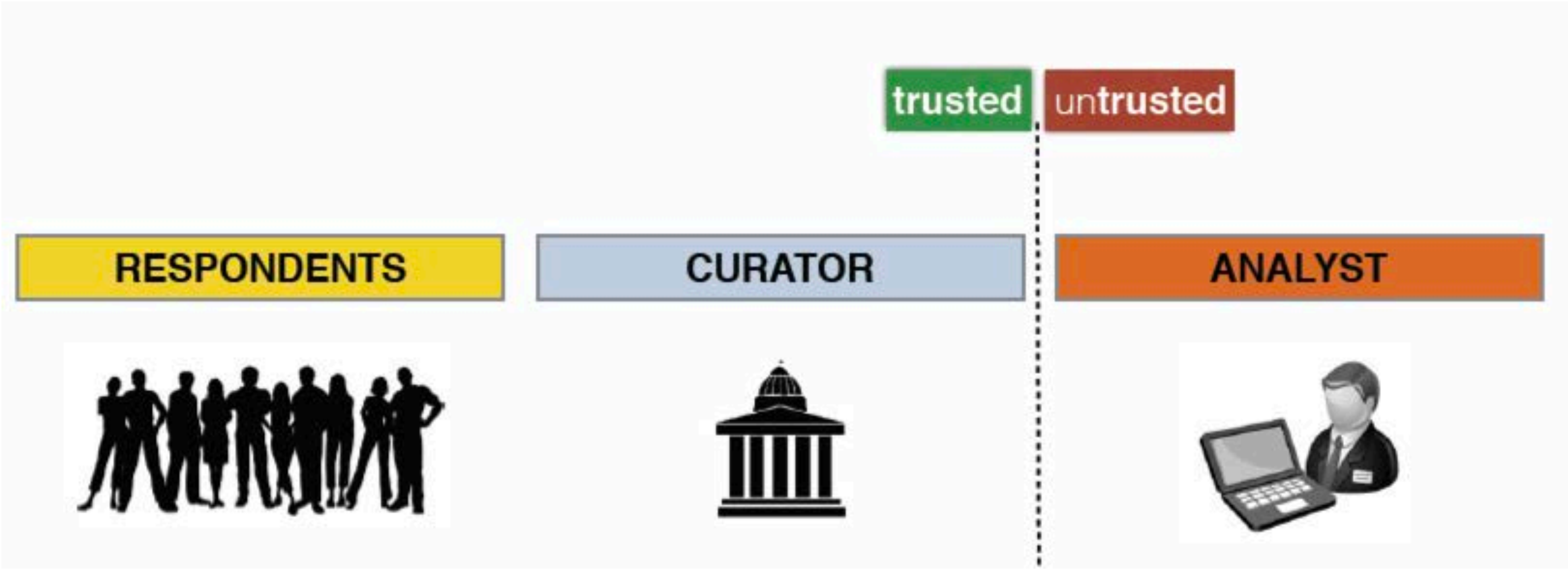


respondents contribute their personal data

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

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

Privacy-preserving data analysis

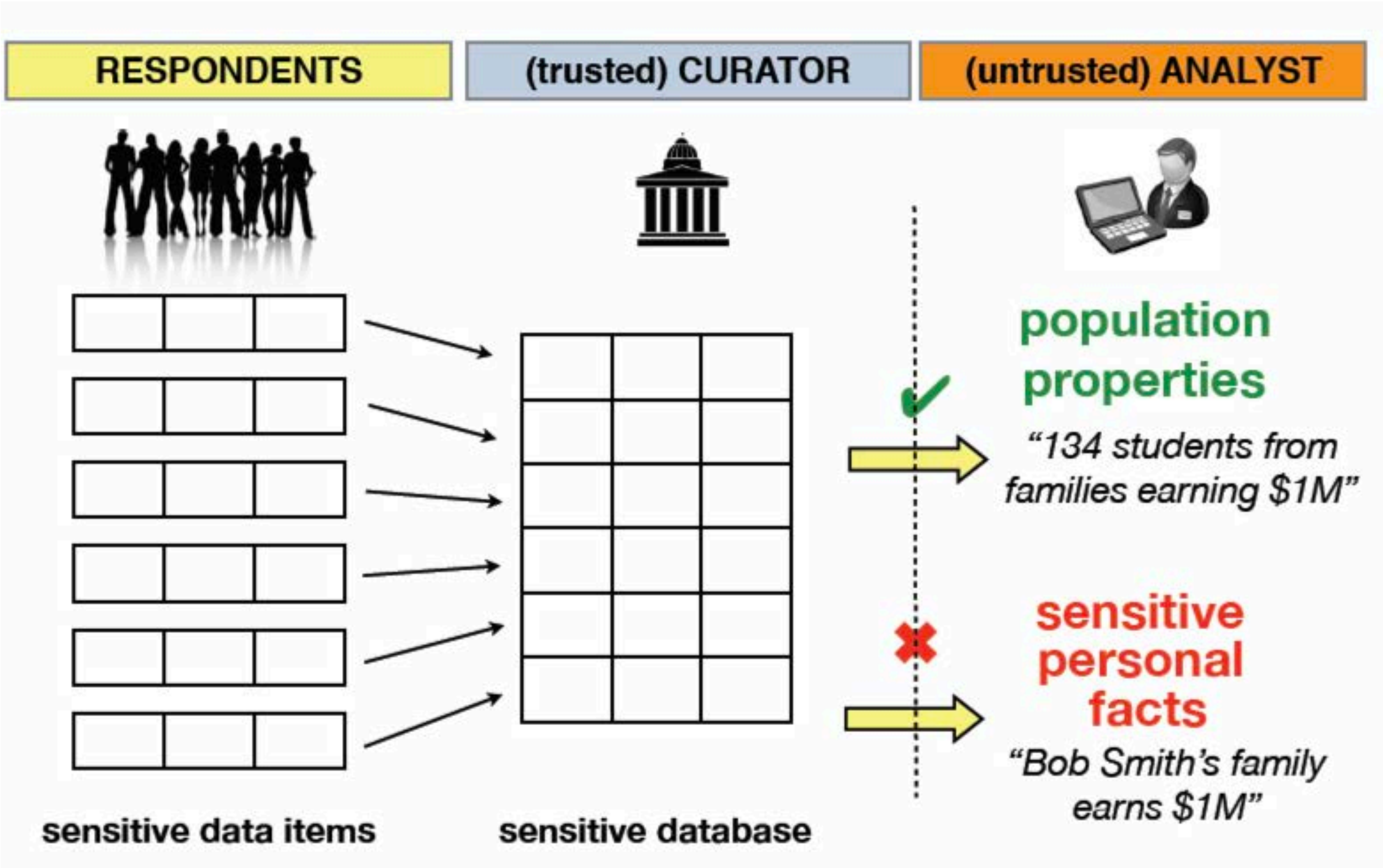


respondents in the population seek protection of their personal data

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



slide by Gerome Miklau

Example: US Census

COLLECTOR

ANALYST

Commuting patterns in the US collected by the census

sensitive data set

✔ **global properties**

"Increasing automobile efficiency will save workers $A on average"

"Public transportation should be built at location B."

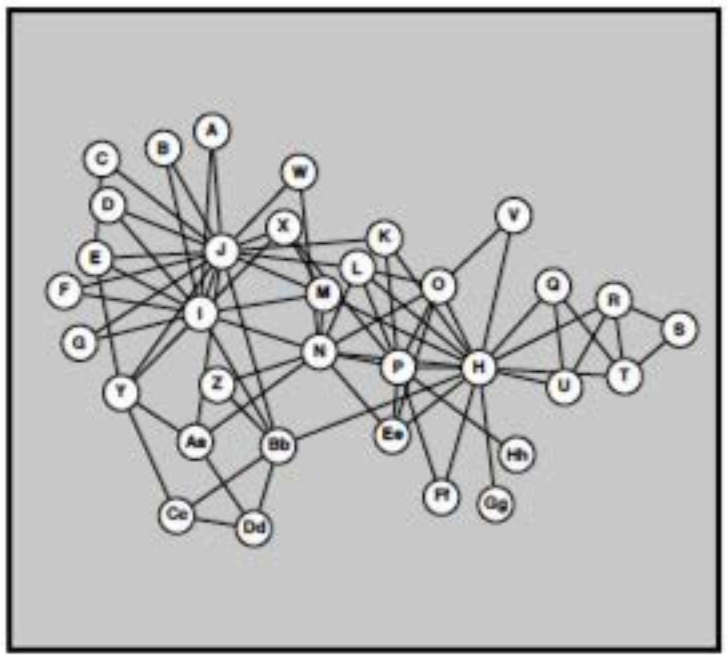
✖ **sensitive facts**

"Alice lives at address X"

"Bob worked for Y, but now works for Z"

Example: Social networks

COLLECTOR	ANALYST



sensitive data set

global properties

"How rapidly do rumors spread in this network?"

"Are people most likely to form friendships with those who share their attributes?"

sensitive facts

"Alice is present in this network"

"Alice and Bob are connected"

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**.

r/ai

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.

r/ai

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,[1] the discussion of related work in Blum et al.,[2] 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.** "

r/ai

differential privacy (DP)

Differential privacy: the formalism

We will define privacy with respect to a database ***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 ℓ_1 norm of a database ***D***, denoted $\|D\|_1$ is the number of tuples in ***D***.

The ℓ_1 distance between databases ***D₁*** *and* ***D₂*** represents the number of tuples on which they differ. $\|D_1 - D_2\|_1$

We refer to a pair of databases that differ in at most 1 tuple as **neighboring databases** $\|D_1 - D_2\|_1 \leq 1$

Of these D_1 and D_2, one, say D_2, is a subset of the other, and, when a proper subset, the larger database 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 **ε-differential privacy** if, for all neighboring databases $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, and for any set of outputs ***S***:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

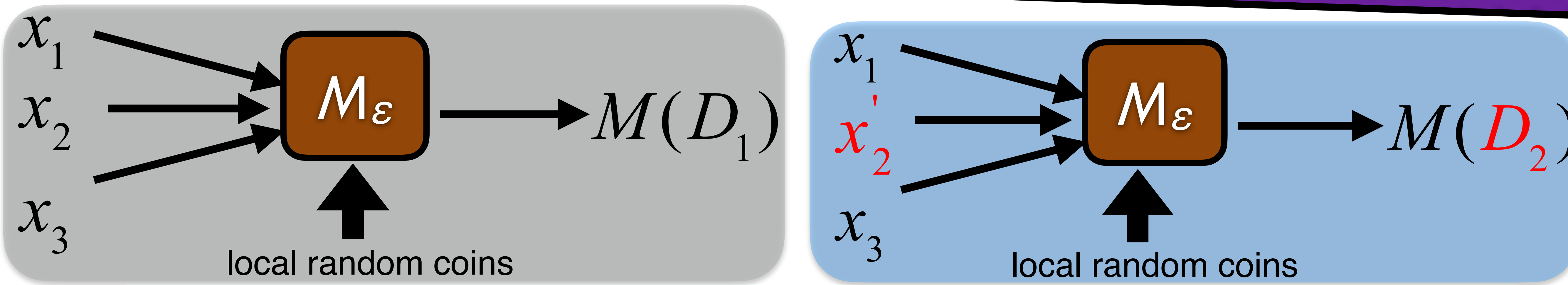
ε (epsilon) is a privacy parameter

lower ε = stronger privacy

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

r/ai

Differential privacy: the formalism



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

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

Think of database of respondents ***$D=(x_1, .., x_n)$*** as **fixed** (not random), ***M(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 **ε-differential privacy** if, for all neighboring databases ***D₁*** and ***D₂***, and for any set of outputs ***S***:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

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

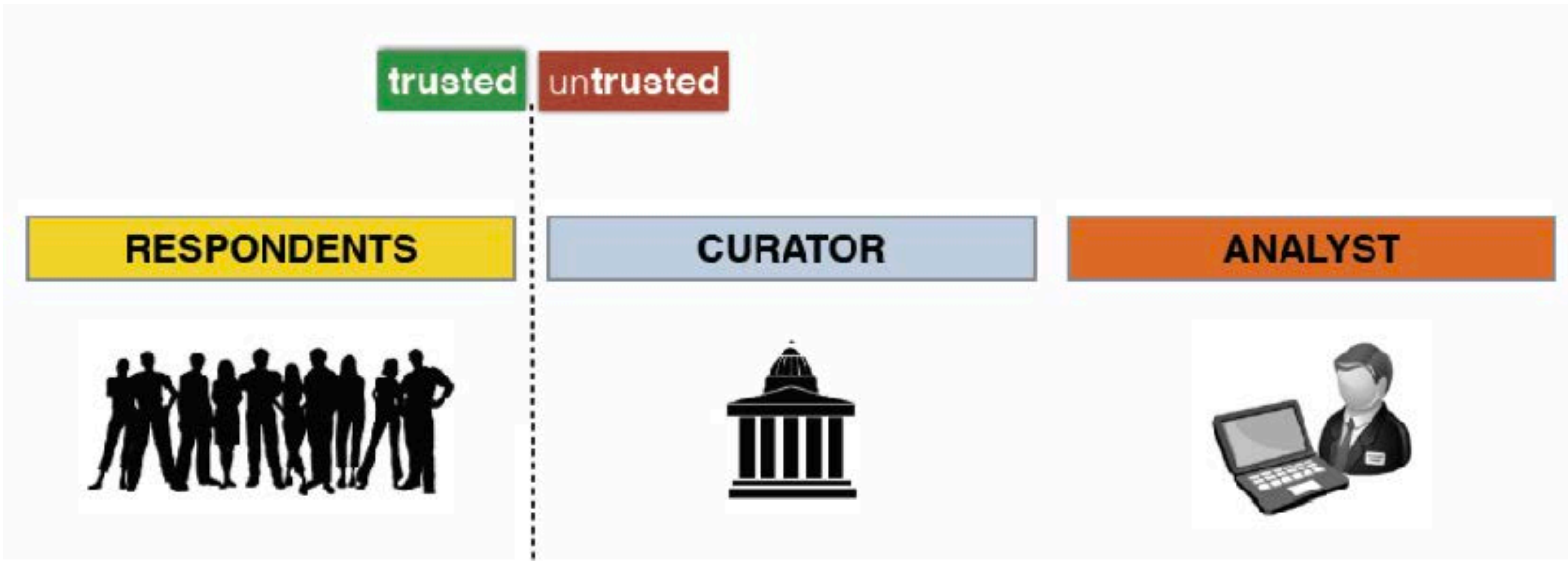
$$\Pr[A \mid \neg P] = \Pr[HH] = \frac{1}{4}$$

$$\Pr[A \mid P] = 3\Pr[A \mid \neg P]$$

$$\Rightarrow \varepsilon = \ln 3$$

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

Local differential privacy

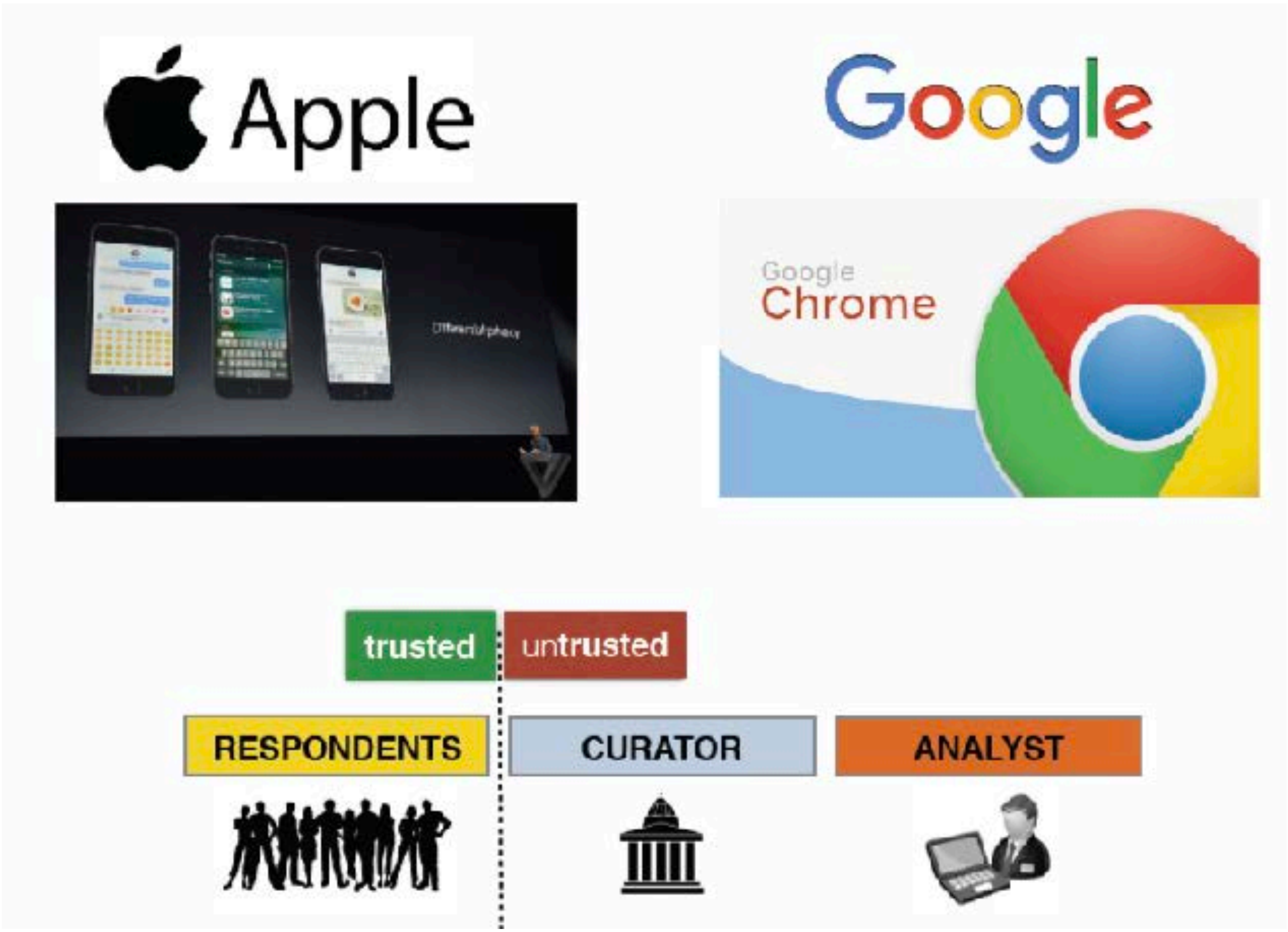


respondents contribute their personal data

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

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

Differential privacy in the field

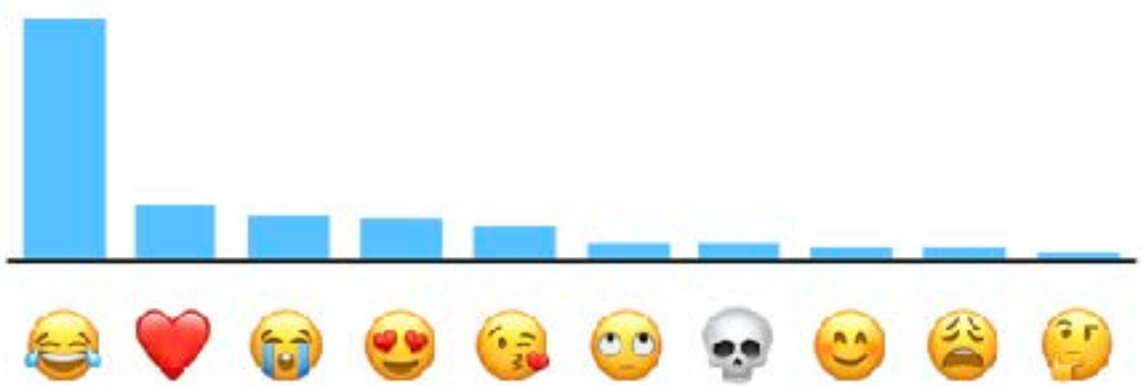


slide by Gerome Miklau

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.



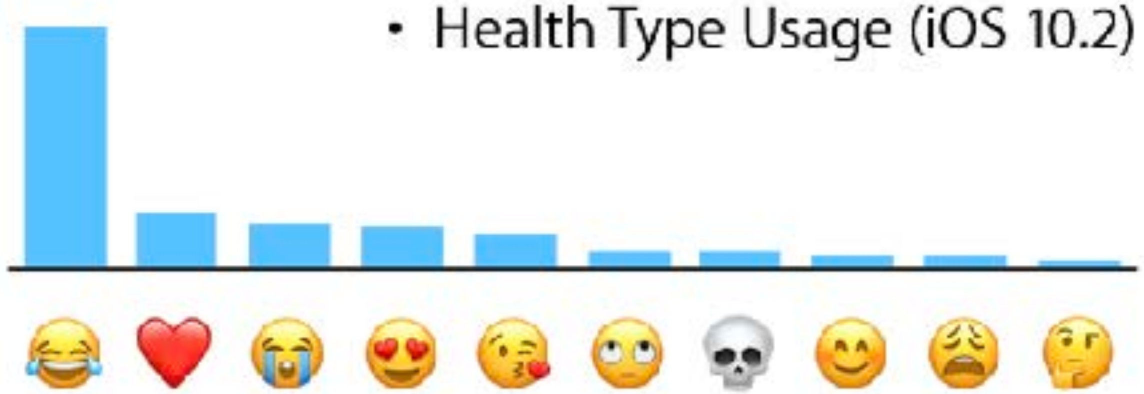
https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

r/ai

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:

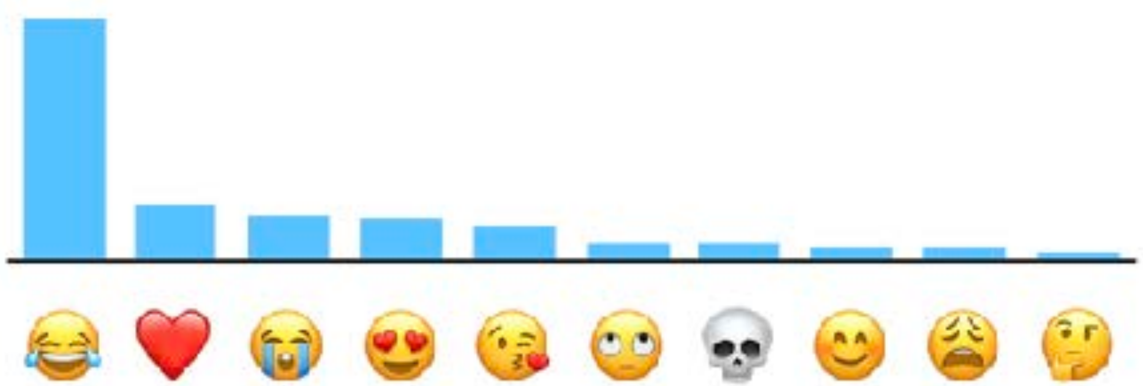
- QuickType suggestions
- Emoji suggestions
- Lookup Hints
- Safari Energy Draining Domains
- Safari Autoplay Intent Detection (macOS High Sierra)
- Safari Crashing Domains (iOS 11)
- Health Type Usage (iOS 10.2)



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).



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

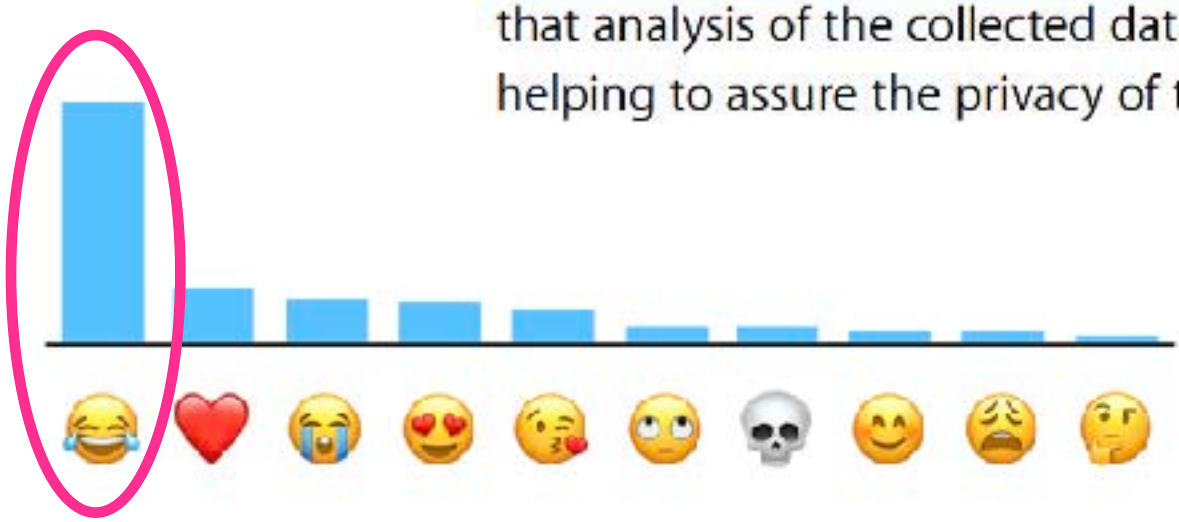
Example: What's your favorite emoji?

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 $1/(1 + e^{\varepsilon/2})$, where ε 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.



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Transparency is important!

ANDY GREENBERG SECURITY 09.15.2017 09:20 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.

Epsilon, Epsilon

"…[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…."

https://www.wired.com/story/apple-differential-privacy-shortcomings/

r/ai

A closer look at differential privacy

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

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

lower ε = 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 **ε-differential privacy** if, for all neighboring databases ***D₁*** and ***D₂***, and for any set of outputs ***S***:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

lower ε = stronger privacy

ε (epsilon) cannot be too small: think 1/10, not $1/2^{50}$

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

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

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

lower ε = stronger privacy

- Example 1: counting queries
  - "How many elements in ***D*** satisfy property ***P*** ?" **What's *Δq* ?**
  - "What fraction of the elements in ***D*** satisfy property ***P*** ?"
- Example 2: max / min
  - "What is the maximum employee salary in ***D*** ?" **What's *Δq* ?**

Intuition: for a given ε, the higher the sensitivity, the more noise we need to add to meet the privacy guarantee

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

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

query *q*	query sensitivity *Δq*
select count(*) from D	1
select count(*) from D where sex = Male and age > 30	?

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

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

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

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, 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 *Δq***
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	?

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

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

query *q*	**query sensitivity *Δq***
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)

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

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

query *q*	**query sensitivity *Δq***
select gender, count(*) from D group by gender	**1** (**disjoint groups**, presence or absence of one tuple impacts only one of the counts)
an arbitrary list of *m* counting queries	?

r/ai

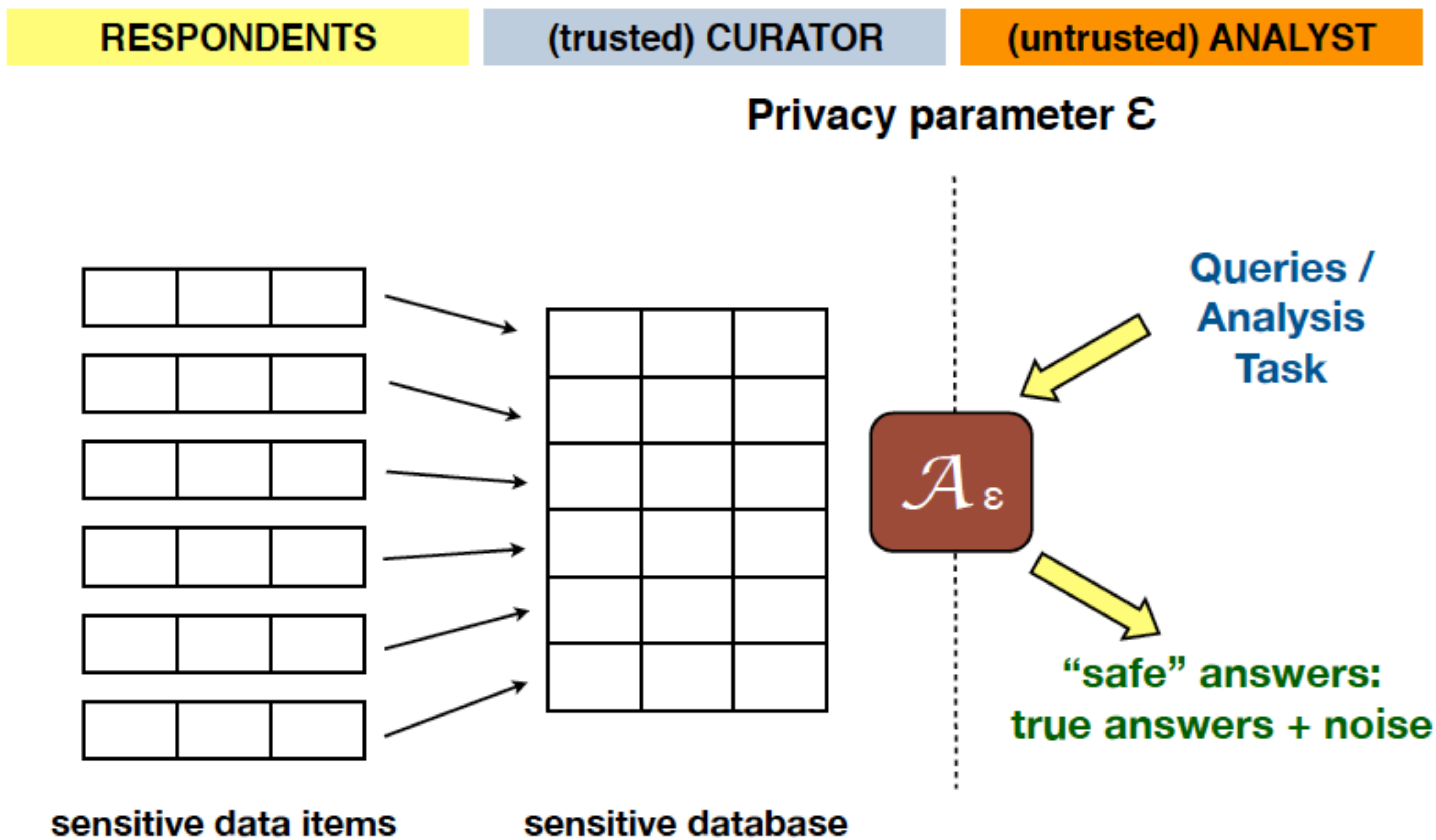
Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, 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 *Δq***
select gender, count(*) from D group by gender	**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



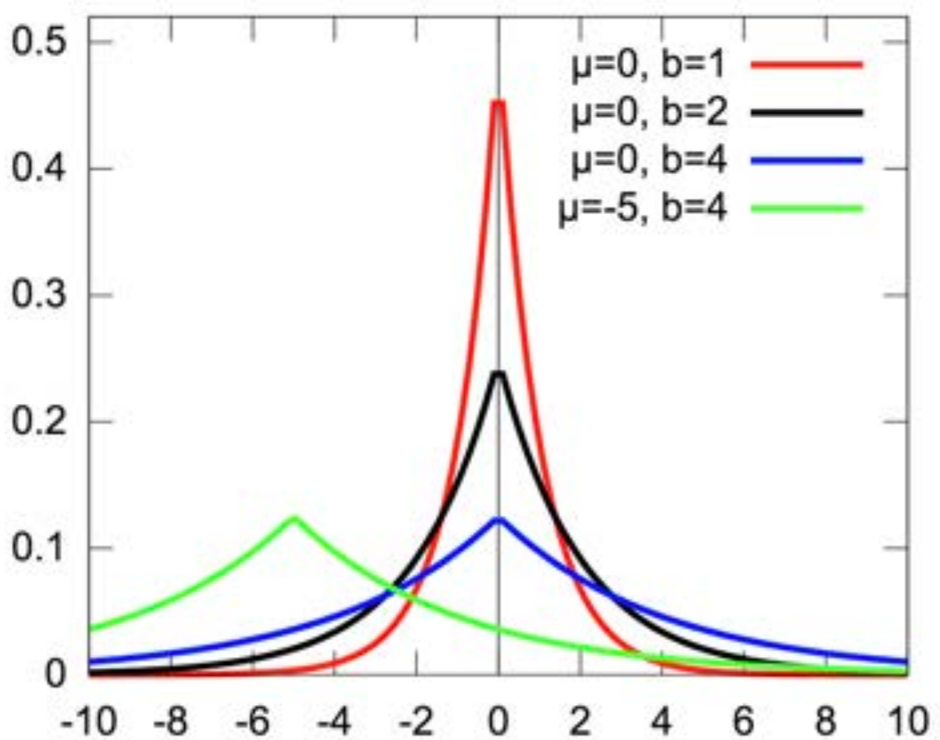
slide by Gerome Miklau

Adding noise

Use the **Laplace mechanism** to answer ***q*** in a way that's ***ε***-differentially private

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

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

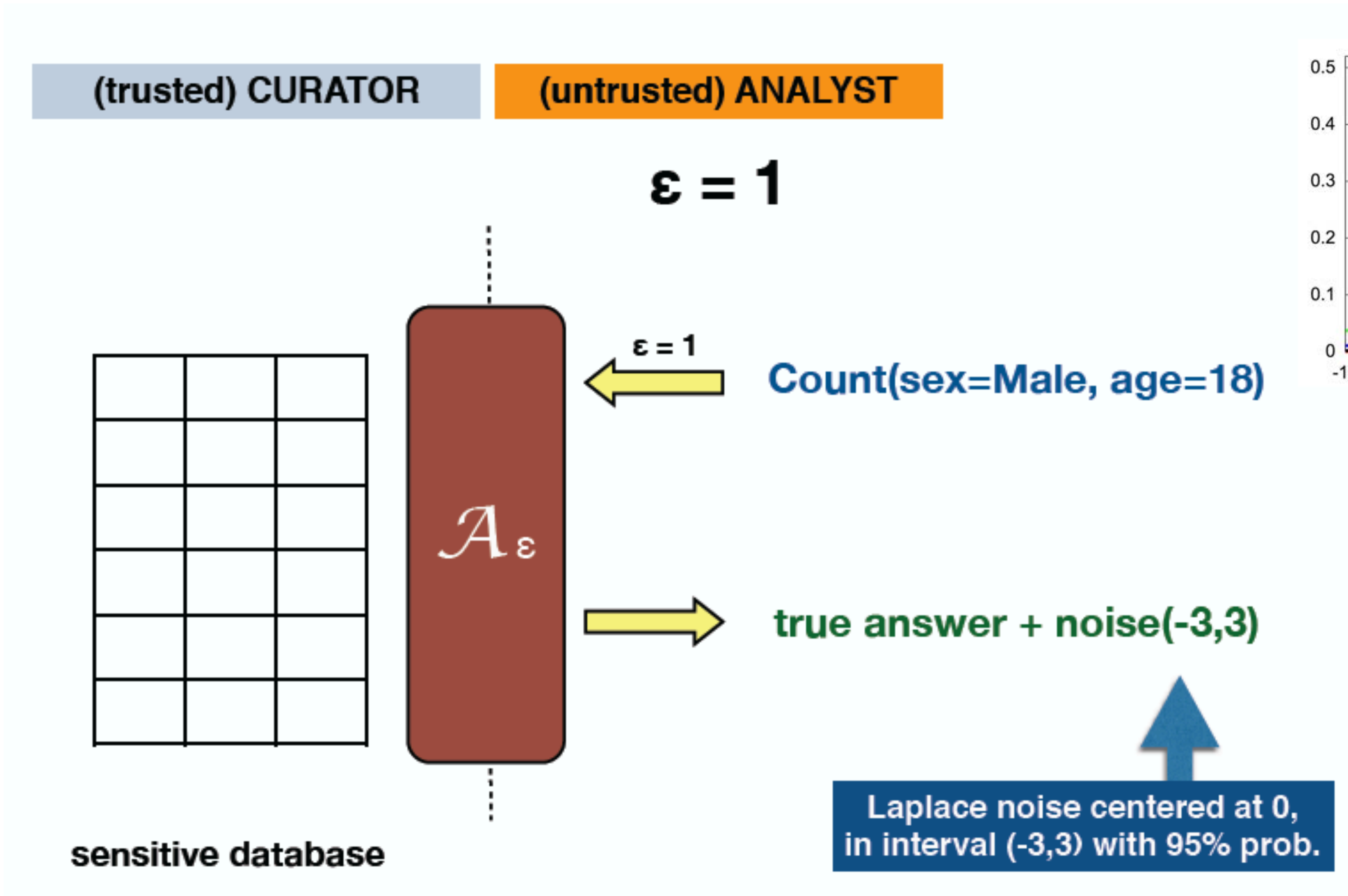


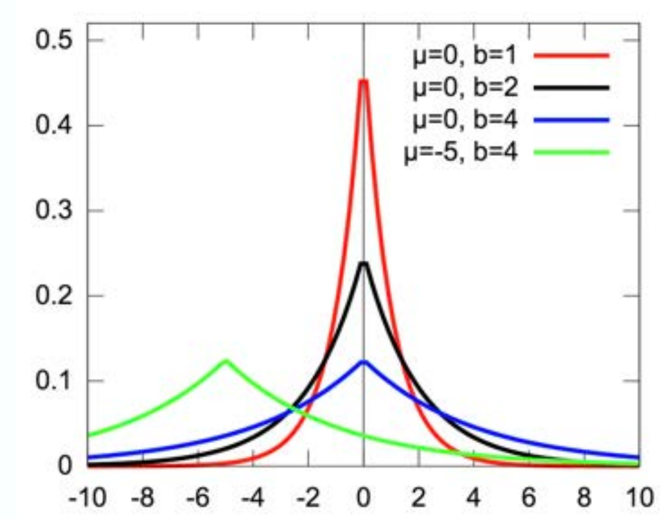
https://en.wikipedia.org/wiki/Laplace_distribution

fix sensitivity ***Δq***, verify that more noise is added for lower ***ε***

lower ε = stronger privacy

Adding noise





Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, 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 *Δq*
parallel composition	
select gender, count(*) from D group by gender	**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)

r/ai

Sequential composition

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

- $\boldsymbol{\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = 1.2}$ That is: all queries together are ε-differentially private for ε =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 ε-differentially private for $\varepsilon = \max(\boldsymbol{\varepsilon_1, \ldots, \varepsilon_k})$
  - Q1: select count(*) from D under $\boldsymbol{\varepsilon_1 = 0.5}$
  - Q2: select count(*) from D where sex = Male under $\boldsymbol{\varepsilon_2 = 0.2}$
  - Q3: select count(*) from D where sex = Female under $\boldsymbol{\varepsilon_3 = 0.25}$
  - Q4: select count(*) from D where age > 20 under $\boldsymbol{\varepsilon_4 = 0.25}$
- $\boldsymbol{\varepsilon = \varepsilon_1 + \max(\varepsilon_2, \varepsilon_3) + \varepsilon_4 = 1}$ That is: all queries together are ε-differentially private for ε =1.

r/ai

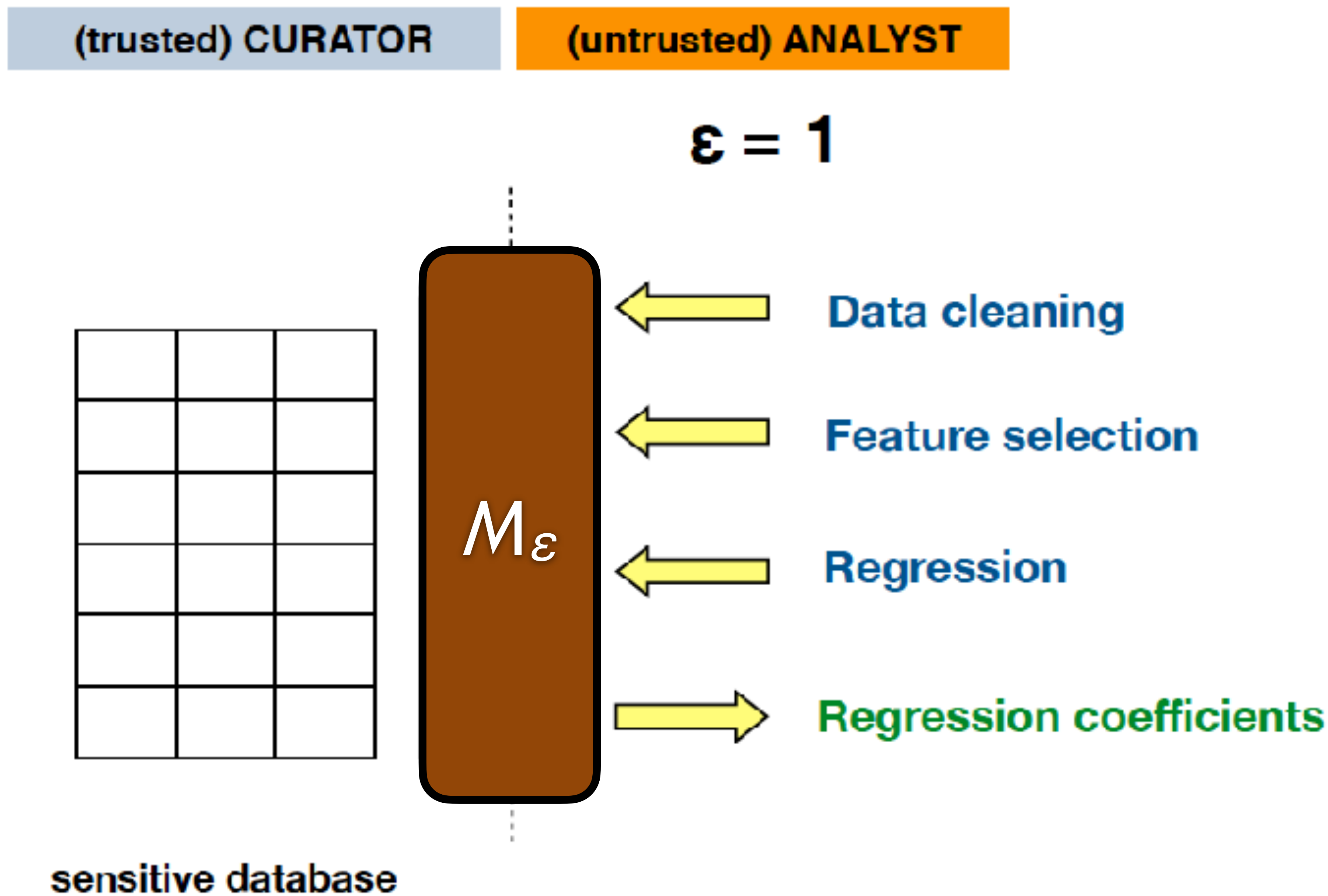
Composition and consistency

- 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**

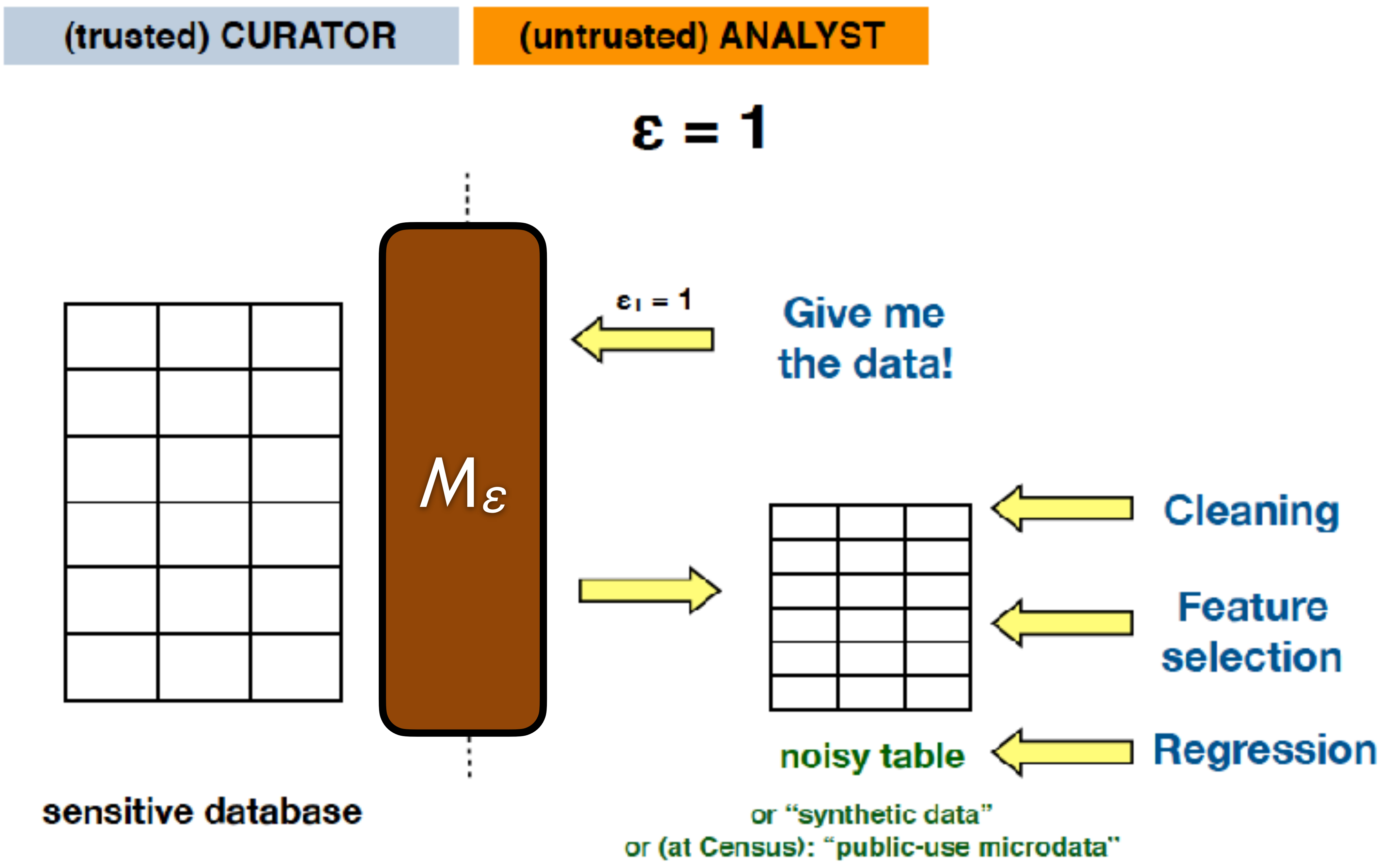
Assuming that there are 2 genders in D, Male and Female, there is **no database consistent with these statistics**!

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



slide by Gerome Miklau

DP synthetic data generation

DP synthetic data

Lots of advantages

- Consistency is not an issue
- Analysts can treat synthetic data as a regular dataset, run existing tools
- No need to worry about the privacy budget
- Can answer as many queries as they want, and any kind of a query they want, including record-level queries

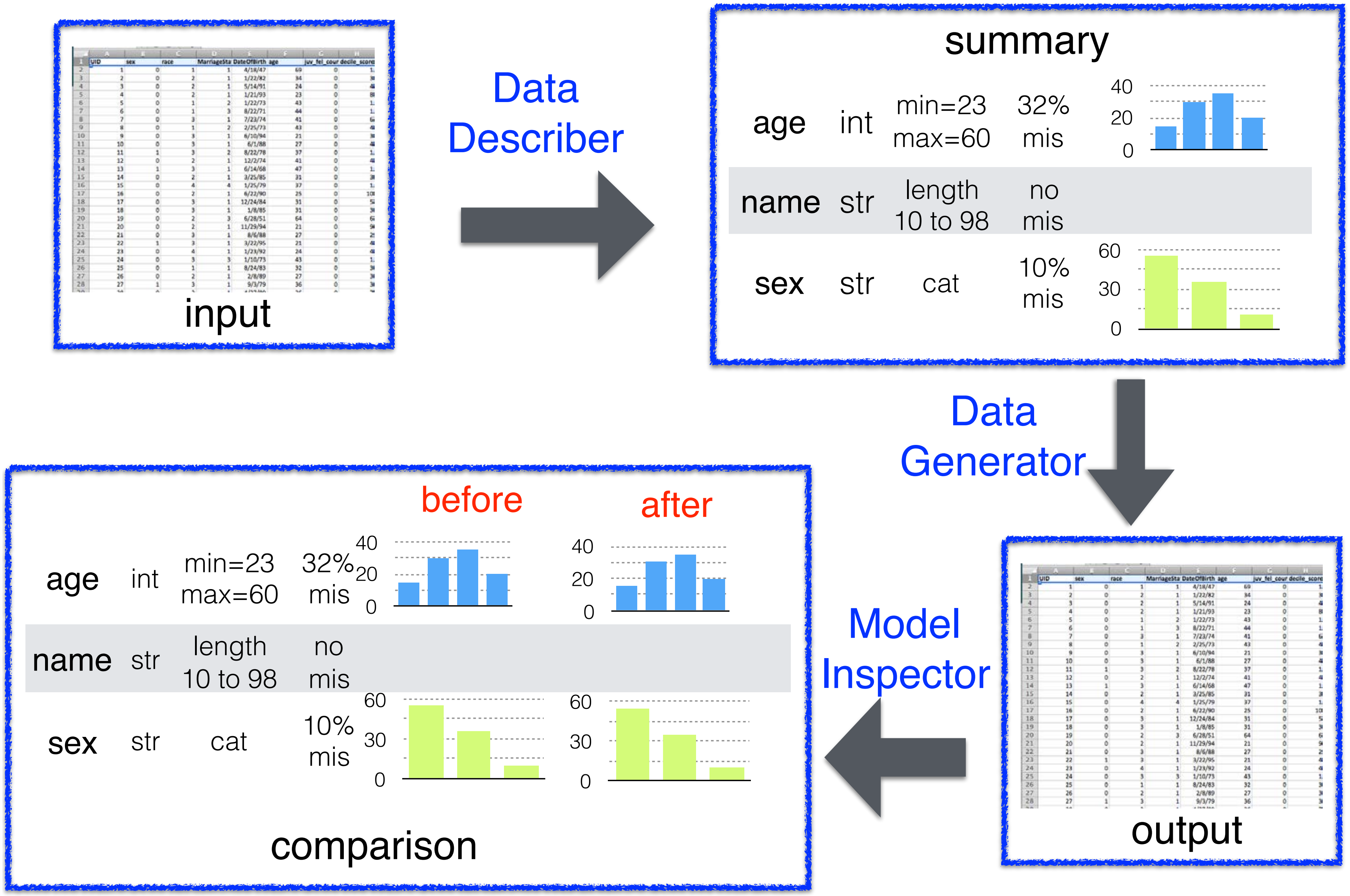
What's the catch?

Recall the Fundamental Law of Information Recovery. It tells us that we cannot answer all these queries accurately and still preserve privacy!

Therefore, when releasing synthetic data, we need to document it with which queries it supports well

r/ai

Data Synthesizer



Data Synthesizer

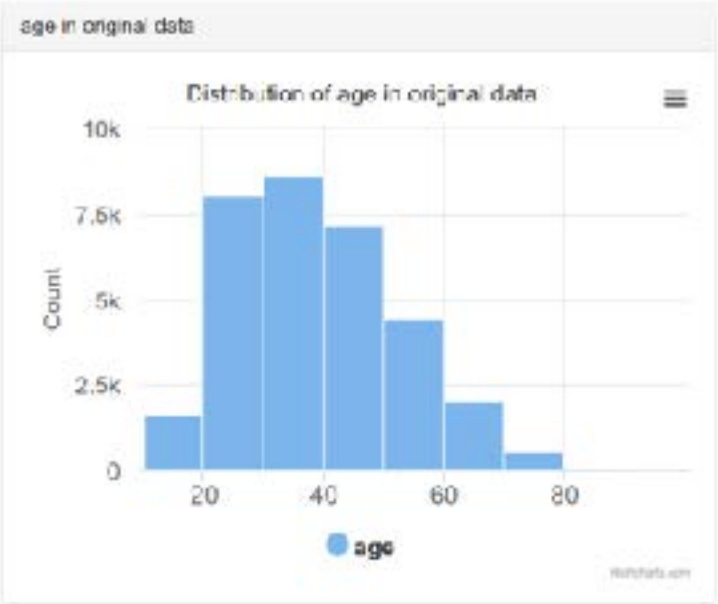
- Main goal: **usability first**
  - user is the data owner
  - the tool picks up data types from the input file: categorical / string / numerical (integer, float) / date-time
  - the tool computes the frequency of missing values per attribute
  - user can then inspect the result, over-ride what was learned about an attribute, e.g., whether it's categorical, or what its datatype is

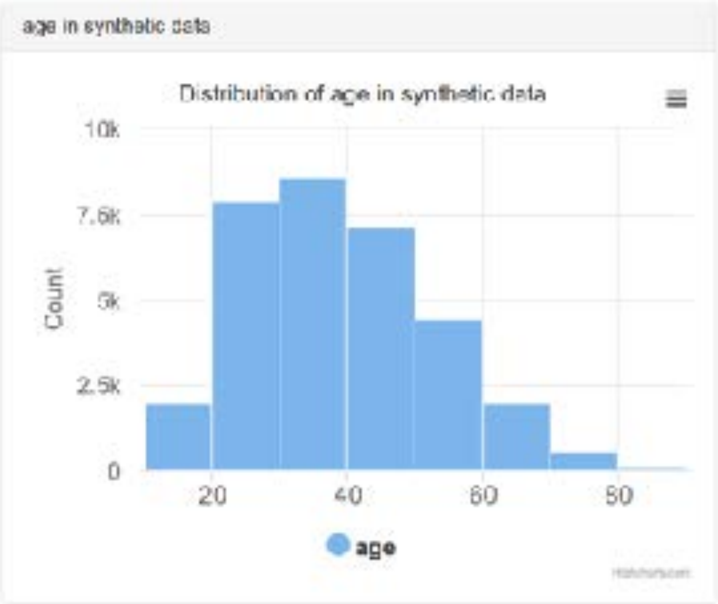
- The tool generates an output dataset of a specified size, in one of three modes
  - **random** - type-consistent random output
  - **independent attribute** - learn a noisy histogram for each attribute
  - **correlated attribute** - learn a noisy Bayesian network (BN)

Data Synthesizer: Independent attributes

Given the over-all privacy budget ε, and an input dataset of size n. Allocate ε/d of the budget to each attribute A_i in $\{A_1, .., A_d\}$. Then for each attribute:

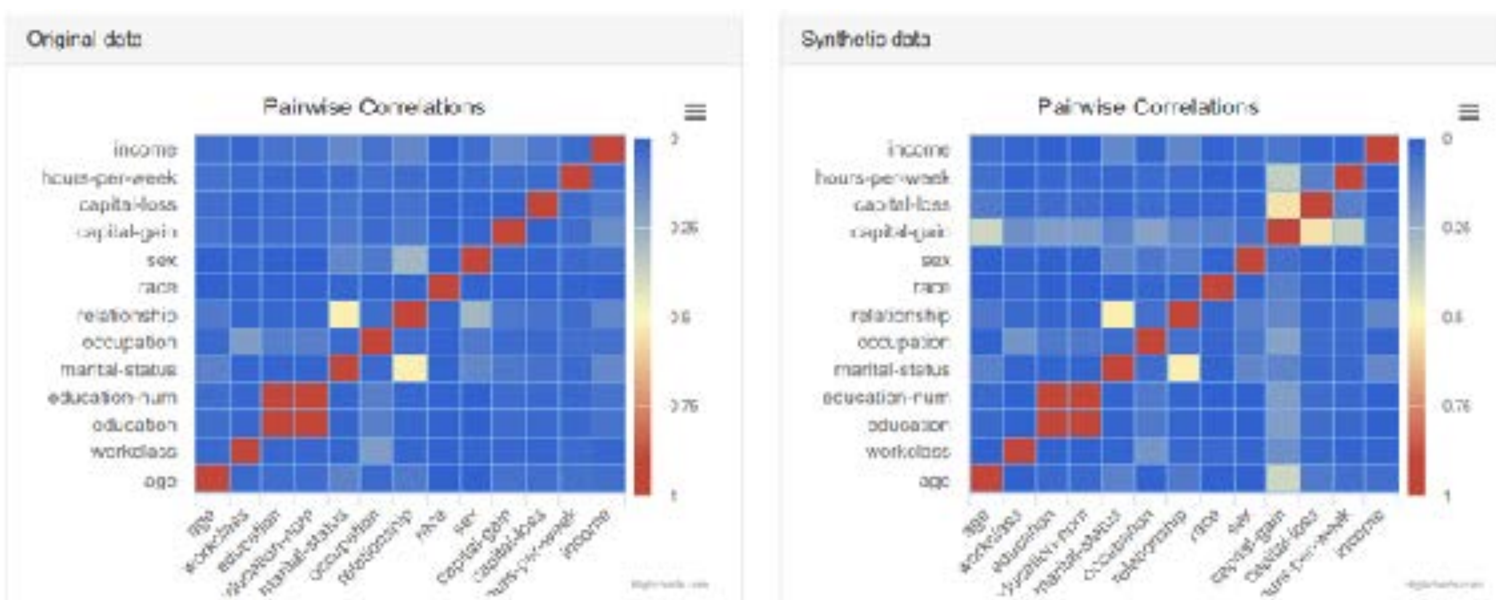
- Compute the i^{th} histogram with t bins (t=20 by default), with query q_i
- The sensitivity Δq_i of this (or any other) histogram query is **2/n** **Why?**
- So, each bin's noisy probability is computed by adding $Lap\left(\frac{2d}{\varepsilon n}\right)$





Data Synthesizer: Correlated attributes

- Learn a differentially private Bayesian network (BN)
- Use the method called **PrivBayes** [Zhang, Cormode, Procopiuc, Srivastava, Xiao, *2016*]
- Privacy budget is split equally between (a) network structure computation and (b) populating the conditional probability tables of each BN node
- User inputs privacy budget **ε** and the maximum number of parents for a BN node ***k*** - you'll play with these settings as part of HW2
- The tool treats a missing attribute value as one of the values in the attribute's domain (not shown in the examples in the next two slides)

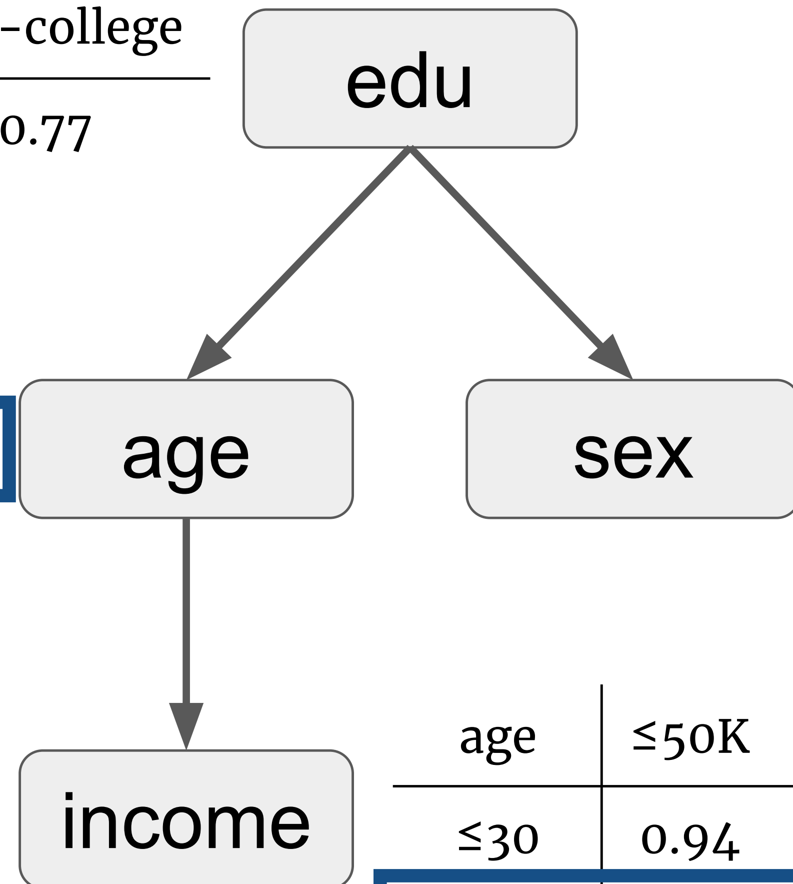


[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Data Synthesizer: Correlated attributes

K=1 not a causal DAG, a regular Bayesian network!

college	non-college
0.23	0.77



edu	≤30	30~50	>50
$\boxed{\text{college}}$	0.24	0.56	0.20
non-college	0.35	0.45	0.20

edu	female	male
college	0.30	0.70
non-college	0.34	0.66

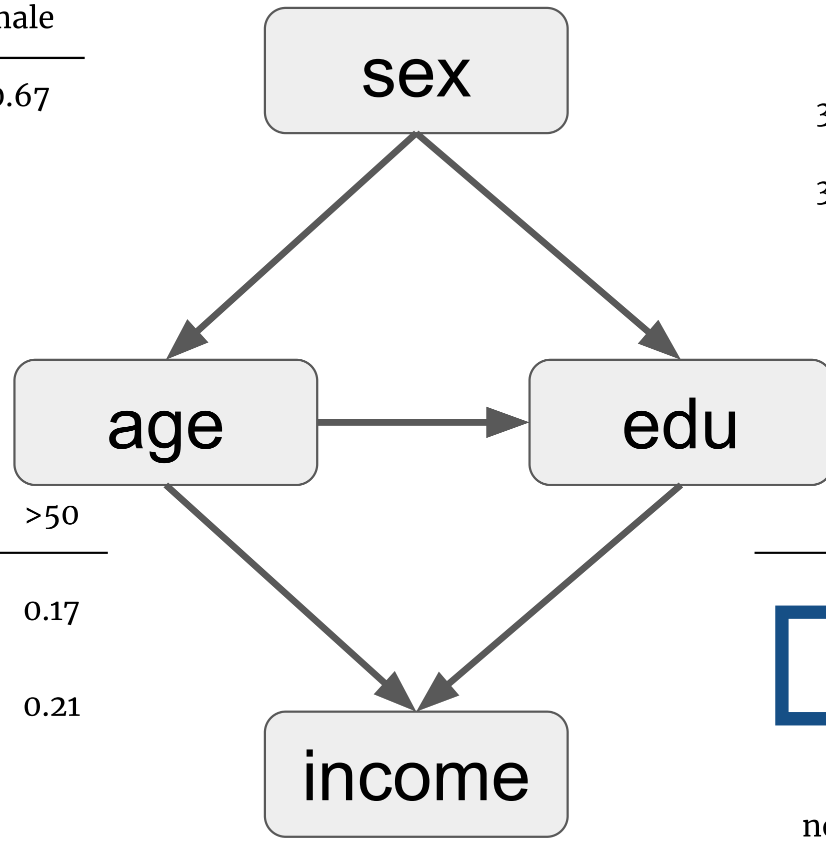
age	≤50K	>50K
≤30	0.94	0.07
$\boxed{30\sim50}$	0.67	0.33
>50	0.68	0.32

Data Synthesizer: Correlated attributes

not a causal DAG, a regular Bayesian network!

K=2

female	male
0.33	0.67



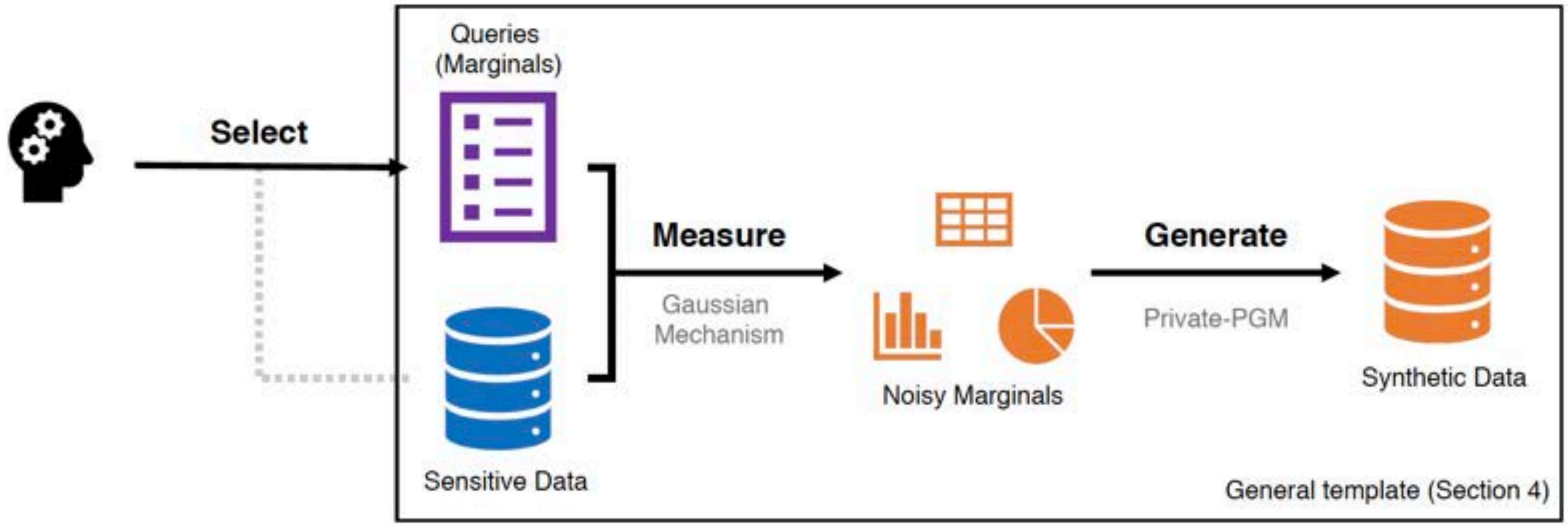
sex	≤30	30~50	>50
female	0.40	0.43	0.17
male	0.29	0.59	0.21

age	sex	college	non-college
≤30	female	0.18	0.82
≤30	male	0.16	0.84
30~50	female	0.25	0.75
30~50	male	0.28	0.72
>50	female	0.17	0.83
>50	male	0.25	0.75

edu	age	≤50K	>50K
college	≤30	0.83	0.17
college	30~50	0.45	0.55
college	>50	0.41	0.59
non-college	≤30	0.96	0.04
non-college	30~50	0.76	0.24
non-college	>50	0.75	0.25

[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

NIST-MST: Tuning synthetic datasets



- **Select** a collection or marginal queries - manually or automatically
- Use the Gaussian mechanism to **measure** those marginals while preserving differential privacy
- Post-process noisy marginals and **generate** synthetic data that respects them

[McKenna, Miklau, Sheldon *2021*] https://github.com/ryan112358/private-pgm/

NIST-MST: Marginals

- For a set of attributes *C in **C***, a **marginal**, is a table that counts the number of occurrences of each combination of possible values of these attributes. Marginals can be **selected** manually by a domain expert or automatically.

- Marginals are **measured** in a DP manner; how epsilon is used can incorporate information about their relative importance (specified as a weight w_C).

SEX	*LABFORCE*	count
M		156
M	*N*	65
M	*Y*	316
F	—	158
F	*N*	282
F	*Y*	23
LABFORCE	***SCHOOL***	**count**
—	*N*	159
	Y	155
N	*N*	288
N	*Y*	59
Y	*N*	336
Y	*Y*	3

(a) True marginals

SEX	*LABFORCE*	count
M		132.428
M	*N*	124.549
M	*Y*	244.365
F	—	173.633
F	*N*	318.029
F	*Y*	−21.358
LABFORCE	***SCHOOL***	**count**
—	*N*	116.021
—	*Y*	186.826
N	*N*	287.215
N	*Y*	171.134
Y	*N*	278.498
Y	*Y*	−46.497

(b) Noisy marginals

SEX	*LABFORCE*	count
M		124.829
M	*N*	121.696
M	*Y*	254.636
F	—	166.034
F	*N*	315.177
F	*Y*	0
LABFORCE	***SCHOOL***	**count**
—	*N*	110.029
—	*Y*	180.834
N	*N*	276.477
N	*Y*	160.396
Y	*N*	254.636
Y	*Y*	0

(c) Private-PGM marginals

Count (LABFORCE=N) is inconsistent in (b)

124.549 + 318.029 = 442.578

287.215 + 171.134 = 458.349

Count (LABFORCE=N) is consistent in (c)

121.696 + 315.177 = 436.873

276.477 + 160.396 = 436.873

Better accuracy in (c) than in (b): L1(a, c) < L1(a, b)

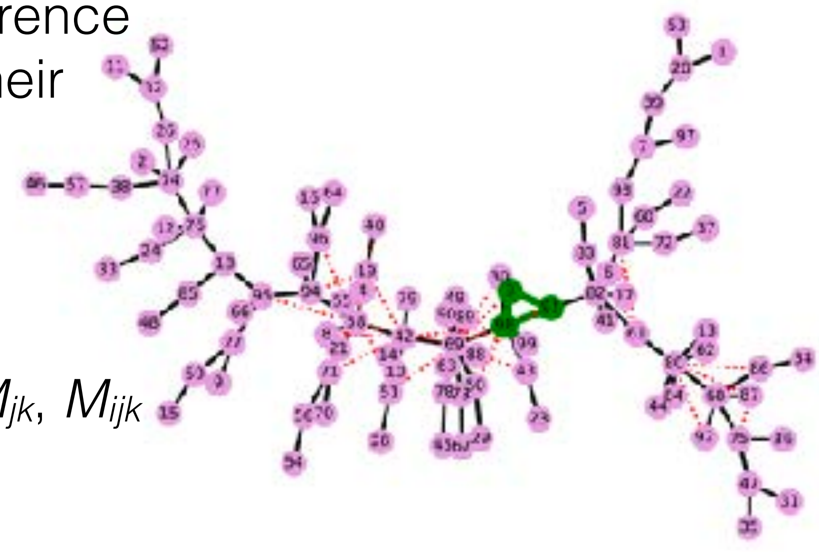
[McKenna, Miklau, Sheldon *2021*] https://github.com/ryan112358/private-pgm/

NIST-MST: Selecting marginals

- Which marginals should we **select** to be measured? Important, because which marginals we focus on will determine which marginals will be preserved well in the synthetic data.
- Marginal selection algorithm takes epsilon as input to determine weights w_C to assign to the selected marginals. It does not consume epsilon.

Summary of Algorithm 5

1. Construct complete graph *G*, where vertices *i* and *j* correspond to attributes, and the weight of edge (*i*, *j*) is the **mutual information** (MI) between *i* and *j*. MI measures a lack of independence between i and j by quantifying the difference between the joint distribution of a pair of variables and to the product of their marginals.
2. Find the maximum spanning tree (MST) of *G*
3. For each pair of adjacent edges (*i*, *j*) and (*i*, *k*), compute marginals M_{ij}, M_{jk}, M_{ijk}
4. Heuristically prune the MST to measure only highly correlated attributes



NIST-MST: Selecting marginals

- Which marginals should we **select** to be measured? Important, because which marginals we focus on will determine which marginals will be preserved well in the synthetic data.
- Marginal selection algorithm takes epsilon as input to determine weights w_C to assign to the selected marginals. It does not consume epsilon.

Algorithm 6: Differentially private measurement selection

Input: D (sensitive dataset), log (measurements of 1-way marginals), ρ (privacy parameter), $\mathcal{C}$ (initial set of (i, j) pairs to measure; empty by default)

Output: $\mathcal{C}$ (final set of (i, j) pairs to measure)

(1) Use `Private-PGM` to estimate all 2-way marginals $\bar{M}_{ij}$ from log

(2) Compute L_1 error between estimated 2-way marginal and actual 2-way marginal for all i, j: $q_{ij}(D) = \left\| M_{ij}(D) - \bar{M}_{ij} \right\|_1$ (this is a sensitivity 1 quantity)

(3) Let $G = (\mathcal{A}, \mathcal{C})$ be the graph where attributes are vertices and edges are pairs of attributes

(4) Let r be the number of connected components in G [6]

(5) Let $\epsilon = \sqrt{\frac{8\rho}{r-1}}$

(6) Repeat $r - 1$ times

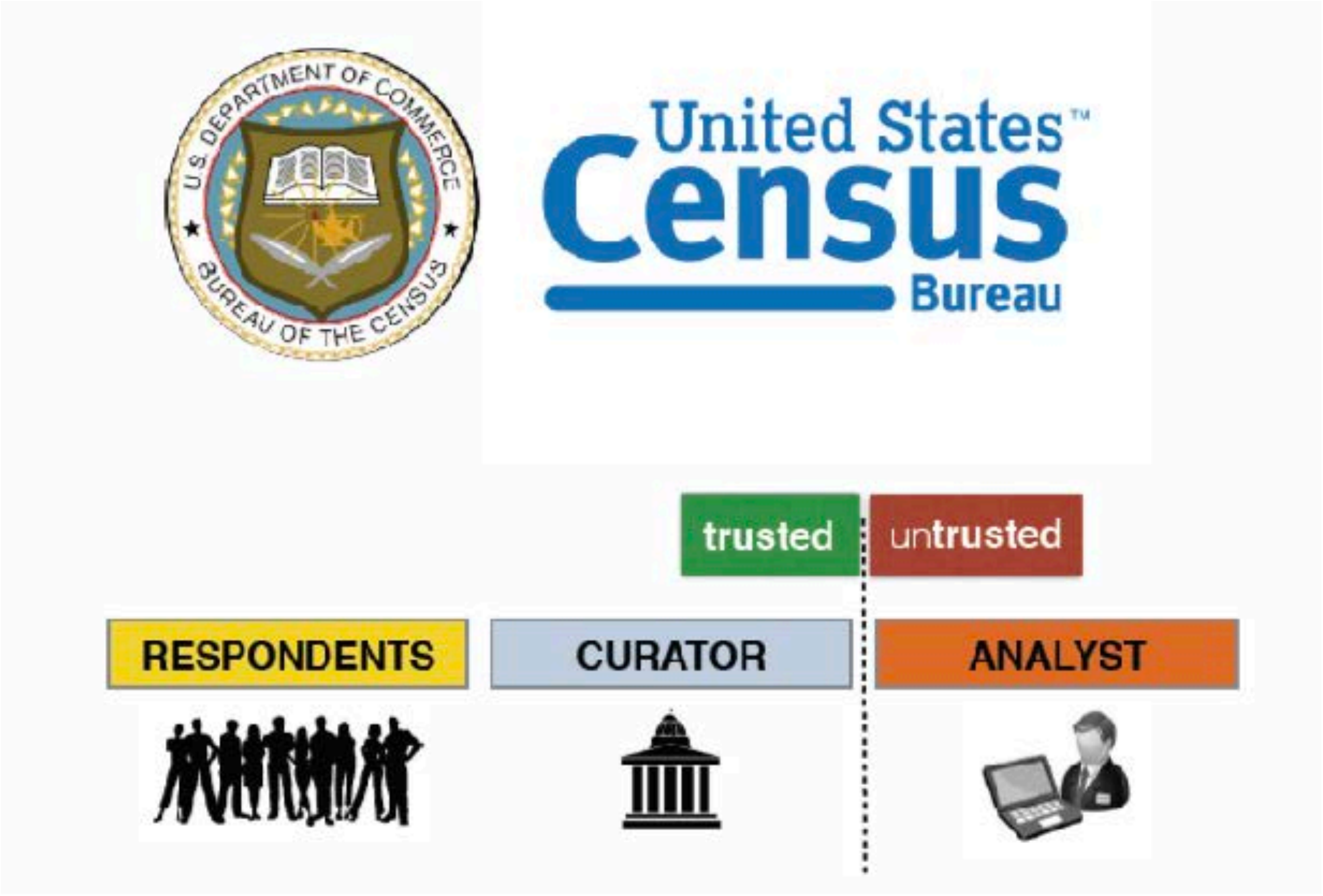
(7) Let S be the set of all attribute pairs (i, j), where i and j are in different connected components of G

(8) Select attribute pair (i, j) by running the `exponential mechanism` with quality score function q_{ij} on set S and privacy parameter ϵ.

(9) Add attribute pair (i, j) to $\mathcal{C}$

DP & the US Census

Differential privacy in the field

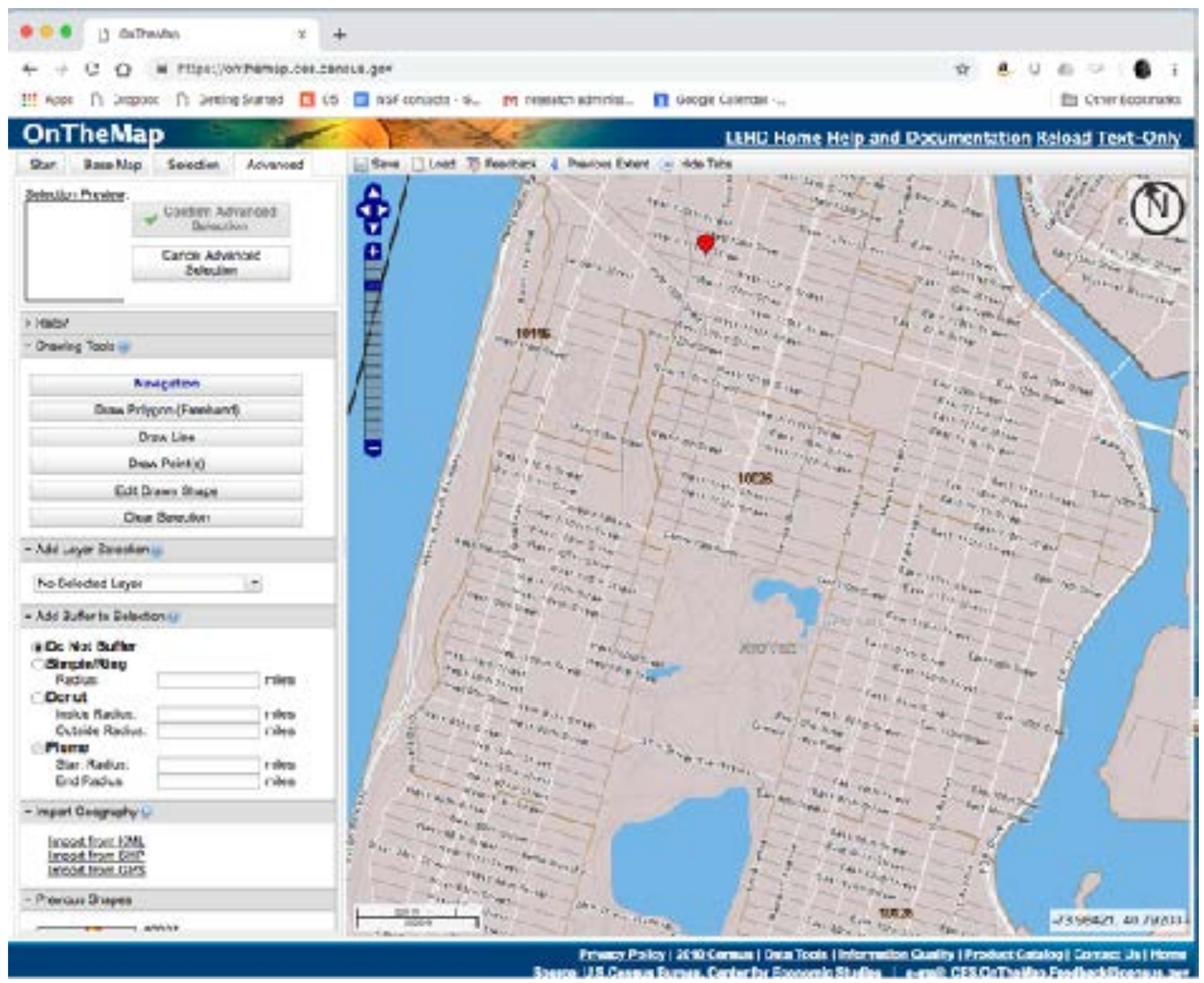


Decennial Census 2020

Differential privacy in the field

First adoption by the US Census Bureau:

OnTheMap (2008), synthetic data about where people in the US live and work



Differential privacy in the field

TheUpshot

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.

The New York Times

By Mark Hansen

Dec. 5, 2018

A 2018 census test letter mailed to a resident in Providence, R.I. The nation's test run of the 2020 Census is in Rhode Island. Michelle R. Smith/Associated Press

At the root of the problem are the tables of aggregate statistics that the bureau publishes. There are hundreds of tables — sex by age, say, or ethnicity by race — summarizing the population at several levels of geography, from areas the size of a city block all the way up to the level of a state or the nation. In 2010, the bureau released tables with nearly eight billion numbers in all. That was about 25 numbers for each person living in the United States, even though Americans were asked only 10 questions about themselves. In other words, the tables were generated in so many ways that the Census Bureau ended up releasing more data in aggregate then it had collected in the first place.

Reconstruction attack: an example

TABLE 1: FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK

STATISTIC	GROUP	AGE		
		COUNT	MEDIAN	MEAN
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American male	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

Note: Married persons must be 15 or over

[Garfinkel, Abowd and Martindale, *ACM Queue 2018*]

Reconstruction attack: an example

Let's assume that the oldest person is 125 years old, and that everyone's age is different. How many possible age combinations are there?

$$\binom{125}{3} = 317{,}750$$

But only 40 combinations have median = 30 and mean = 44

Idea: extract all such constraints, represent them as a mathematical model, have an automated solver find a solution.

TABLE 2: **POSSIBLE AGES FOR A MEDIAN OF 30 AND MEAN OF 44**

A	B	C	A	B	C	A	B	C
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

[Garfinkel, Abowd and Martindale, *ACM Queue 2018*]

r/ai

What does the law say?

Title 13 of U.S. Code authorizes data collection and publication of statistics by the Census Bureau.

Section 9 of Title 13 requires privacy protections: "Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, ... may ... make **any publication whereby the data furnished by any particular establishment or individual under this title can be identified**" (Title 13 U.S.C. § 9(a)(2), Public Law 87-813).

> In 2002, Congress further clarified the concept of identifiable data: it is prohibited to publish "**any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means**" (Pub. L. 107–347, Title V, §502 (4), Dec. 17, 2002, 116 Stat. 2969).

Section 214 of Title 13 outlines penalties: fines up to $5,000 or imprisonment up to 5 years or both per incident (data item), up to $250,000 in total.

DP in the 2020 Census: pushback

MPC
MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Implications of Differential Privacy for Census Bureau Data and Research

Task Force on Differential Privacy for Census Data †
Institute for Social Research and Data Innovation (ISRDI)
University of Minnesota

November 2018
Version 2
Working Paper No. 2018-6

- noisy data - **impact on critical decisions**
- difficult to explain differential privacy / privacy budget to the public - **how do we set epsilon**?
- disagreement about whether using differential privacy is legally required
- messaging is difficult to get right "**the result doesn't change whether or not you participate**" - might discourage participation!

Revealing **characteristics** of individuals vs. their **identity**, is there a distinction?

But the Census collects "generic" **harmless data**, is this really a big deal?

What sorts of trade-offs should we be aware of? Who should decide?

beyond
differential privacy

The Strava Heat Map

WIRED

JEREMY HSU SECURITY JAN 29, 2018 7:14 PM

The Strava Heat Map and the End of Secrets

The US military is reexamining security policies after fitness tracker data shared on social media revealed bases and patrol routes

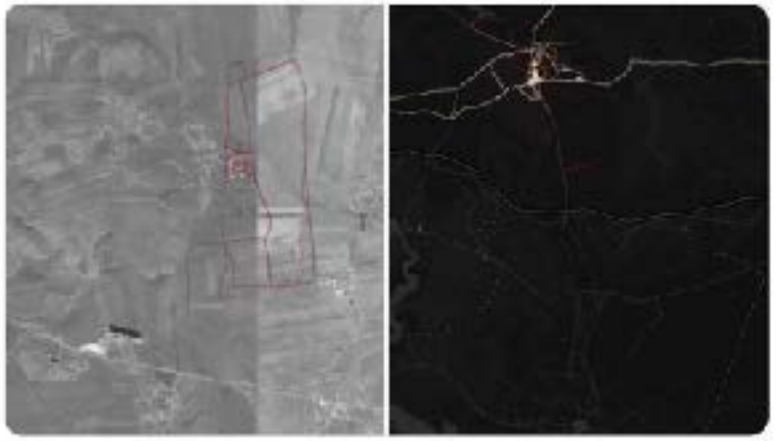
"Over the weekend, researchers and journalists raised the alarm about how **anyone can identify secretive military bases and patrol routes** based on public data shared by a "social network for athletes" called Strava.

This past November, the San Francisco-based Strava announced a huge update to its global heat map of user activity that displays 1 billion activities—including running and cycling routes—undertaken by exercise enthusiasts wearing Fitbits or other wearable fitness trackers. […]

But the biggest danger may come from potential adversaries figuring out "patterns of life," by tracking and even identifying military or intelligence agency personnel as they go about their duties or head home after deployment. These digital footprints that echo the real-life steps of individuals underscore a greater challenge to governments and ordinary citizens alike: each person's connection to online services and personal devices makes it increasingly difficult to keep secrets."

Nathan Ruser
@Nrg8000

Strava released their global heatmap. 13 trillion GPS points from their users (turning off data sharing is an option). medium.com/strava-enginee… … It looks very pretty, but not amazing for Op-Sec. US Bases are clearly identifiable and mappable



1:24 PM · Jan 27, 2018

Read the full conversation on Twitter

2.5K Reply Copy link to Tweet

Read 132 replies

r/ai

Is genetic data your own?

NEWS | SCIENCE AND POLICY

We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans

Researchers call for limiting how ancestry databases can be used to protect privacy

Science

If you're white, live in the United States, and a distant relative has uploaded their DNA to a public ancestry database, there's a good chance **an internet sleuth can identify you from a DNA sample you left somewhere**. That's the conclusion of a new study, which finds that by combining an anonymous DNA sample with some basic information such as someone's rough age, researchers **could narrow that person's identity to fewer than 20 people by starting with a DNA database of 1.3 million individuals**. […]

The study was sparked by **the April arrest of the alleged "Golden State Killer," a California man accused of a series of decades-old rapes and murders**. To find him—and more than a dozen other criminal suspects since then—law enforcement agencies first test a crime scene DNA sample, which could be old blood, hair, or semen, for hundreds of thousands of DNA markers—signposts along the genome that vary among people, but whose identity in many cases are shared with blood relatives. They then upload the DNA data to **GEDmatch, a free online database where anyone can share their data from consumer DNA testing companies such as 23andMe and Ancestry.com to search for relatives who have submitted their DNA**. Searching GEDMatch's nearly 1 million profiles revealed several relatives who were the equivalent to third cousins to the crime scene DNA linked to the Golden State Killer. Other information such as genealogical records, approximate age, and crime locations then allowed the sleuths to home in on a single person.

the origins of data protection

Barrow, Alaska, 1979

Native leaders and city officials, worried about drinking and associated violence in their community, **invited a group of sociology researchers** to assess the problem and work with them to devise solutions.

Methodology:

- 10% representative sample (N=88) of everyone over the age of 15 using a 1972 demographic survey
- Interviewed on attitudes and values about use of alcohol
- Obtained psychological histories & drinking behavior
- Given the Michigan Alcoholism Screening Test
- Asked to draw a picture of a person (to determine cultural identity)

based on a slide by Bill Howe

Study “results”

Alcohol Plagues Eskimos; Alcoholism Plagues Eskimo Village

DAVA SOBEL ();
January 22, 1980,
, Section Science Times, Page C1, Column , words

PERMISSIONS

[DISPLAYING ABSTRACT]

THE Inupiat Eskimos of Alaska's North Slope, whose culture has been overwhelmed by energy development activities, are "practically committing suicide" by mass alcoholism, University of Pennsylvania researchers said here yesterday. The alcoholism rate is 72 percent among the 2,000 Eskimo men and women in the village of Barrow, where violence is becoming the ...

At the conclusion of the study researchers formulated a report entitled **“The Inupiat, Economics and Alcohol on the Alaskan North Slope”**, released **simultaneously** at a press release and to the Barrow community.

The press release was picked up by the New York Times, who ran a front page story entitled **“Alcohol Plagues Eskimos”**

Harms and backlash

Article Preview

Eskimos Irate Over Alcoholism Study

[DISPLAYING ABSTRACT]

PERMISSIONS

BARROW, ALASKA HOT tempers and tension arising from a scientific report that found a high rate of alcoholism in this predominantly Eskimo community have abated somewhat after two days of meetings here at the northernmost point of Alaska.

Study **results were revealed** in the context of a press conference that was held far from the Native village, and **without the presence, much less the knowledge or consent**, of any community member who might have been able to present any context concerning the socioeconomic conditions of the village.

Study results suggested that nearly all adults in the community were alcoholics. In addition to the shame felt by community members, the town's Standard and Poor bond rating suffered as a result, which in turn decreased the tribe's ability to secure funding for much needed projects.

Problems

Edward F. Foulks, M.D., "Misalliances In The Barrow Alcohol Study"

Methodological

- "The authors once again met with the Barrow Technical Advisory Group, who stated their concern that only Natives were studied, and that outsiders in town had not been included."

- "The **estimates of the frequency of intoxication based on association with the probability of being detained** were termed "ludicrous, both logically and statistically."

Ethical

- Participants not in control of how their data is used
- Significant harm: social (stigmatization) and financial (bond rating)

can differential privacy help with this?

need an ethical framework!

Responsible Data Science

Anonymity and privacy

Thank you!