Responsible Data Science

Differential privacy

Prof. Julia Stoyanovich

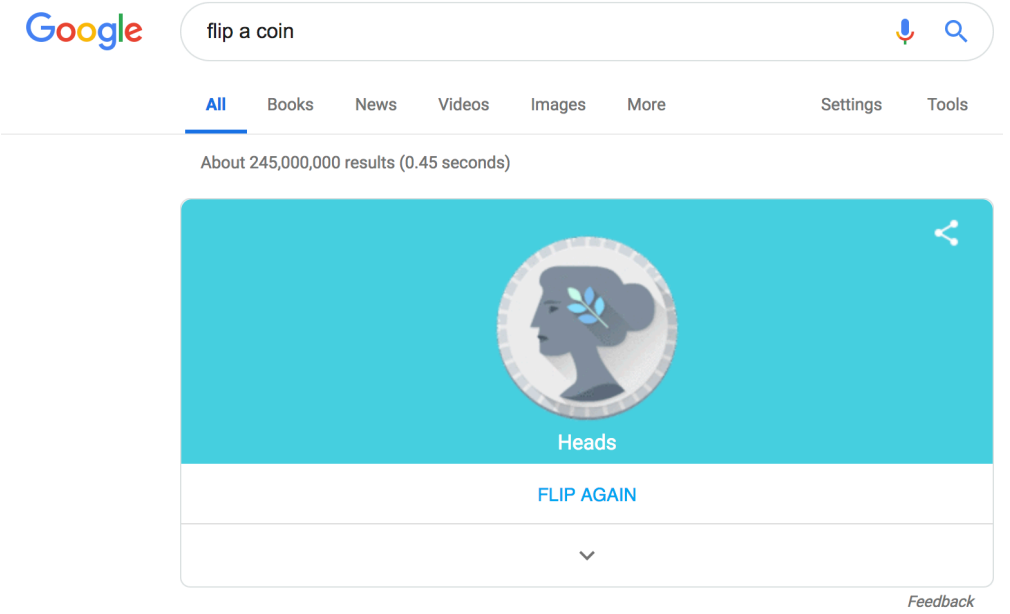
Center for Data Science &
Computer Science and Engineering
New York University

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

Privacy: two sides of the coin

protecting an individual

plausible deniability

learning about the population

noisy estimates

Do we really need randomization?

- 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
- 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**

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/

slide by Bill Howe

The Netflix prize linkage attack

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

- 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?

The Netflix prize linkage attack

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

- 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

RYAN SINGEL SECURITY 12.17.09 04:29 PM

NETFLIX SPILLED YOUR BROKEBACK MOUNTAIN SECRET, LAWSUIT CLAIMS

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.

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
- But:
  - **Query auditing is computationally infeasible**

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

  - 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 is infeasible

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

- 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

Privacy: two sides of the coin

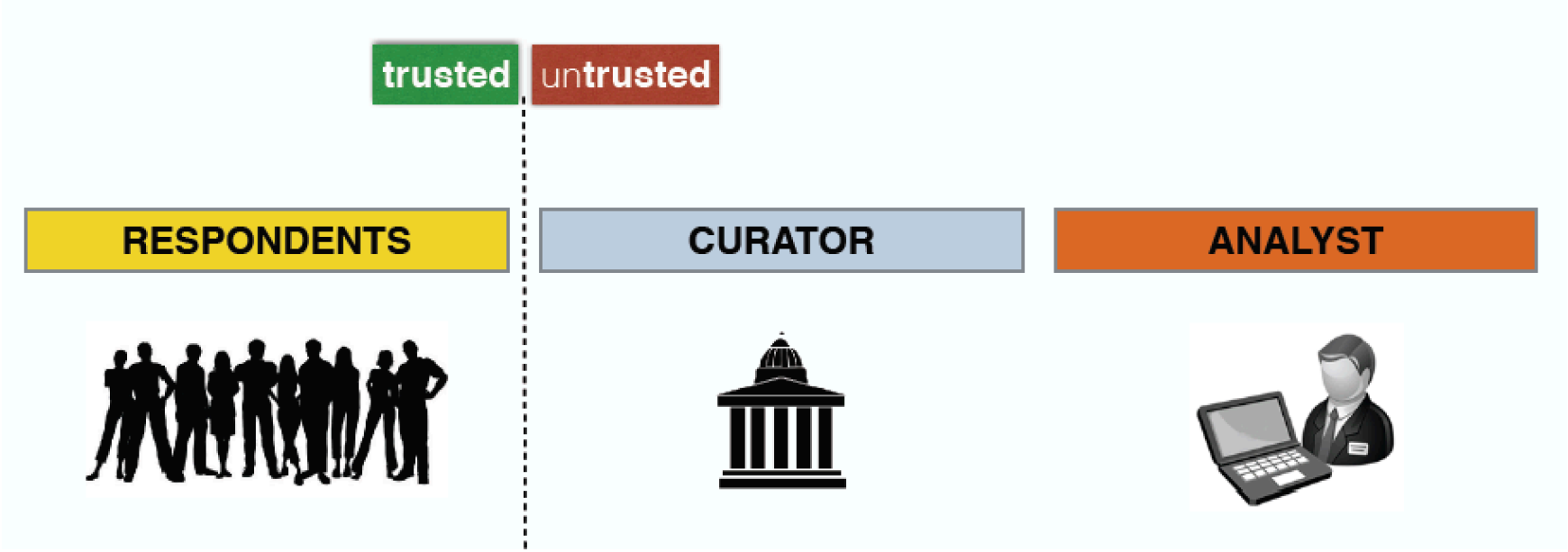
protecting an individual

plausible deniability

learning about the population

noisy estimates

Privacy-preserving data analysis



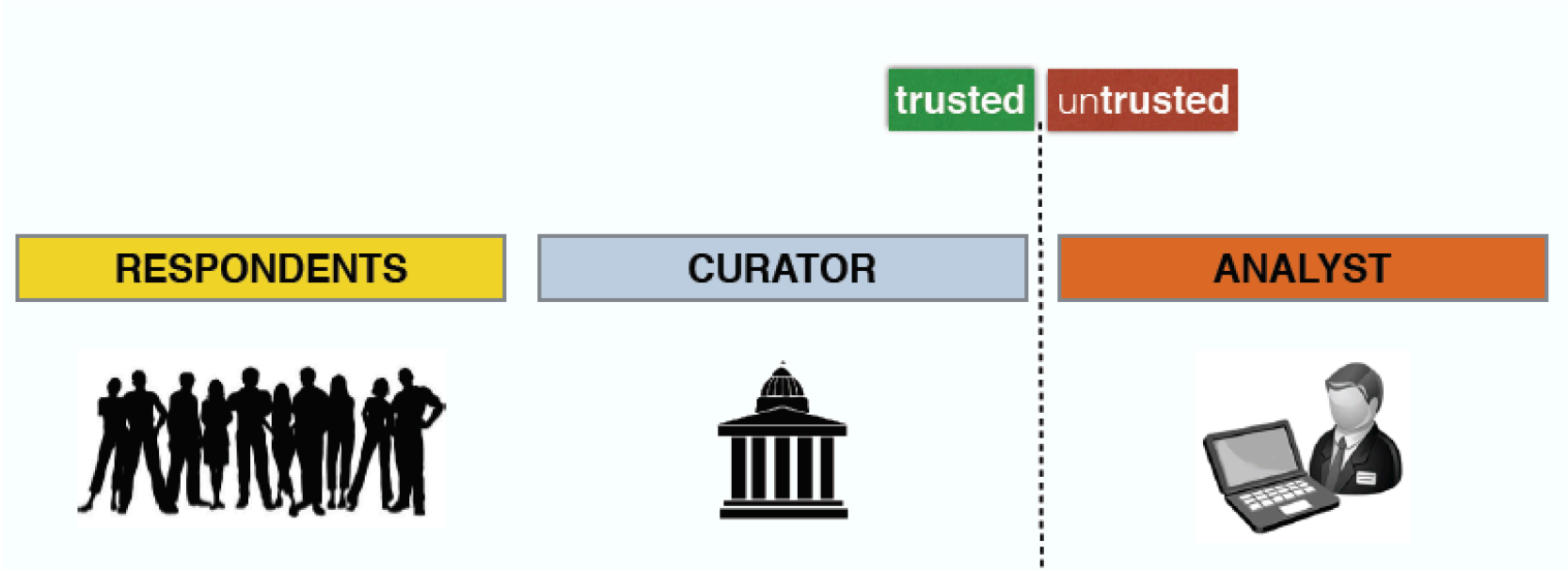
respondents contribute their personal data

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

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

slide by Gerome Miklau

Privacy-preserving data analysis



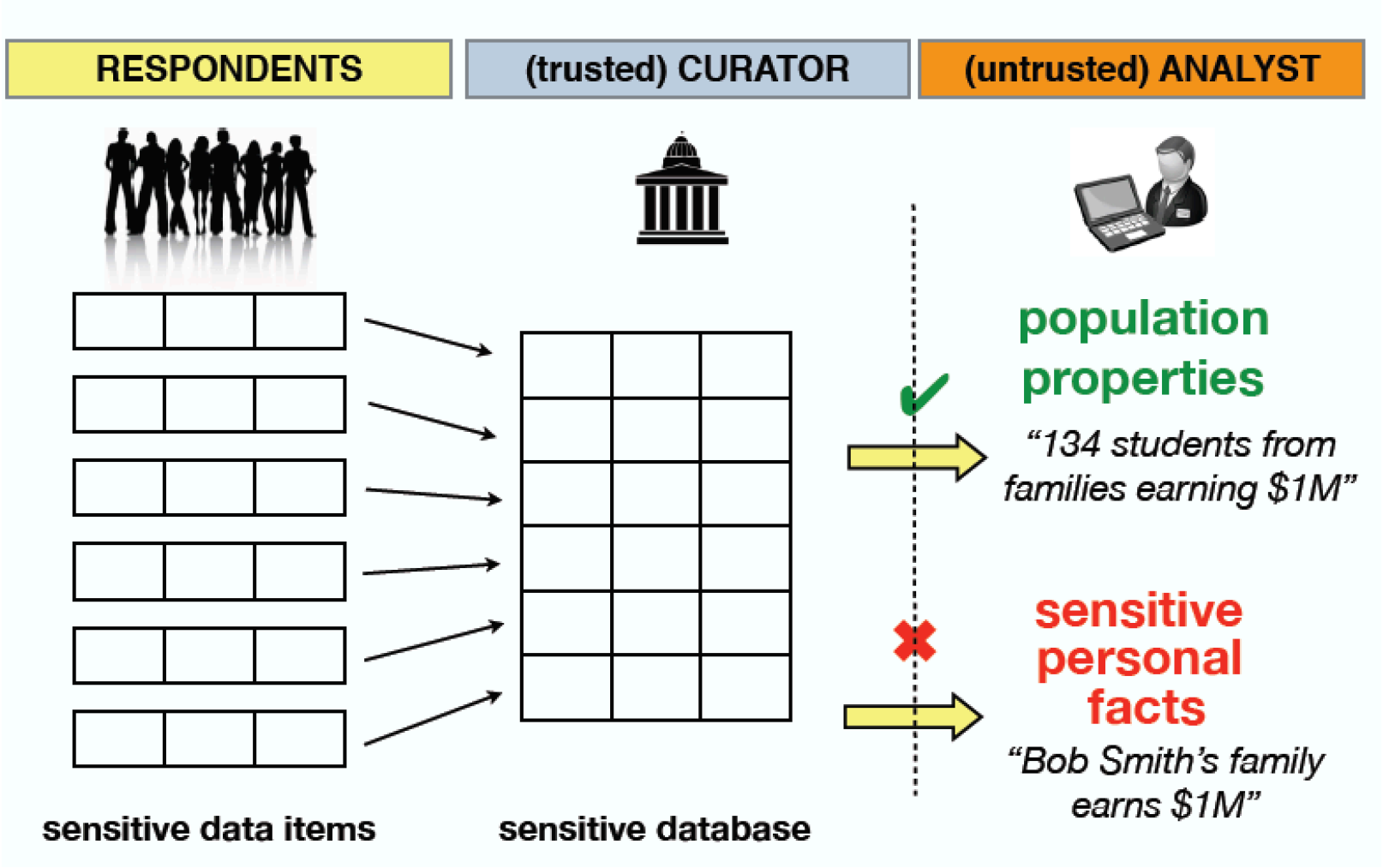
respondents in the population seek protection of their personal data

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

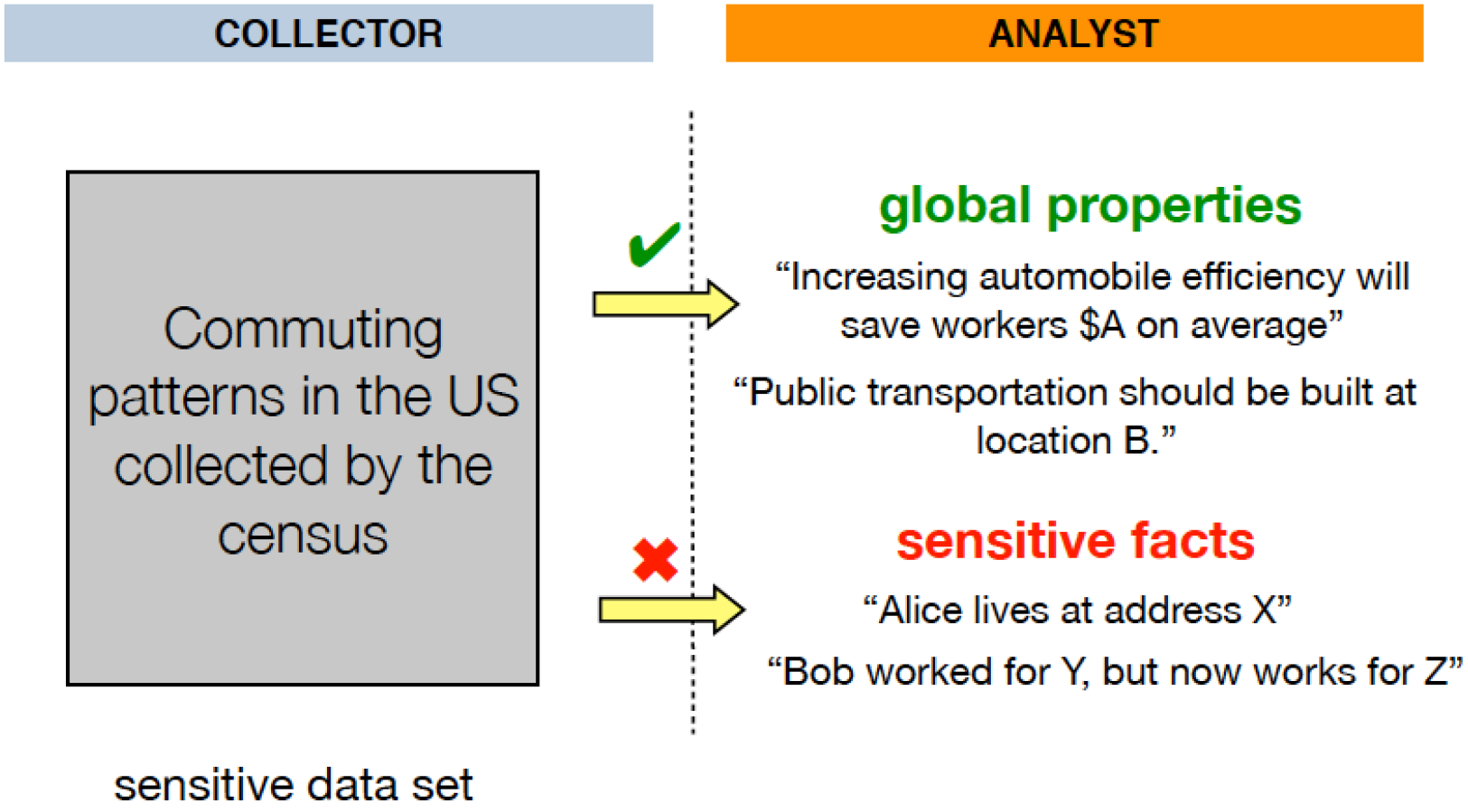
slide by Gerome Miklau

Privacy-preserving data analysis



slide by Gerome Miklau

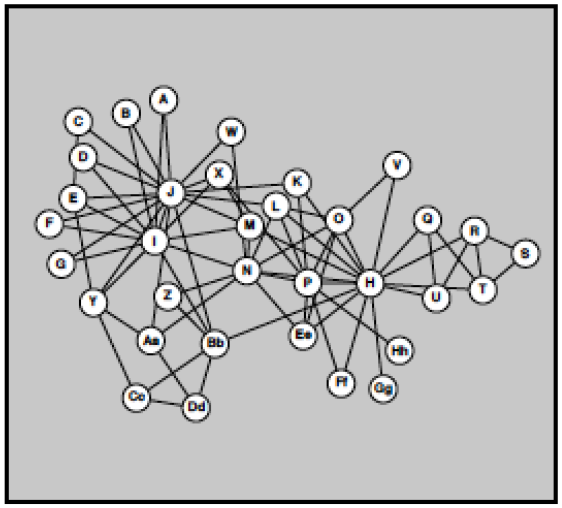
Example: Census data



slide by Gerome Miklau

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"

slide by Gerome Miklau

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

review articles

DOI:10.1145/1866739.1866758

What does it mean to preserve privacy?

BY CYNTHIA DWORK

A Firm Foundation for Private Data Analysis

Communications of the ACM CACM
Homepage archive
Volume 54 Issue 1, January 2011
Pages 86-95

"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: 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 ***D₁*** and ***D₂***, one, say ***D₂,*** is a subset of the other, and, when a proper subset, the larger database ***D₂*** 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 ***D*$_1$** and ***D*$_2$**, and for any set of outputs ***S***:

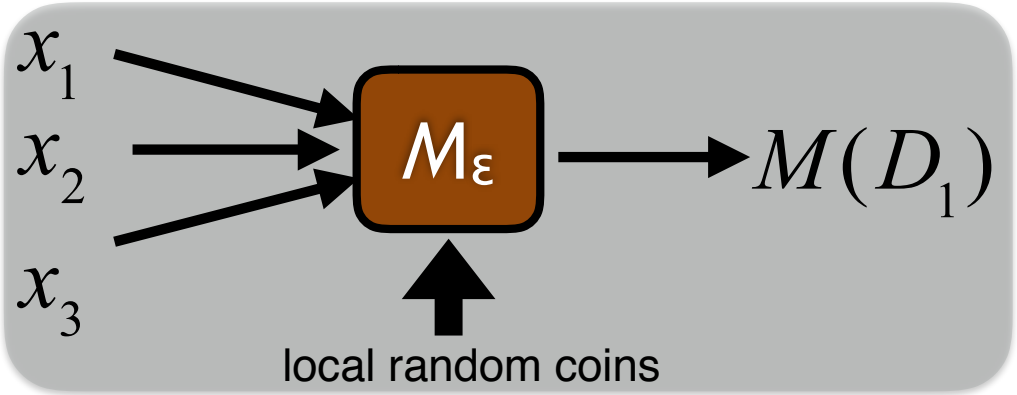
$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

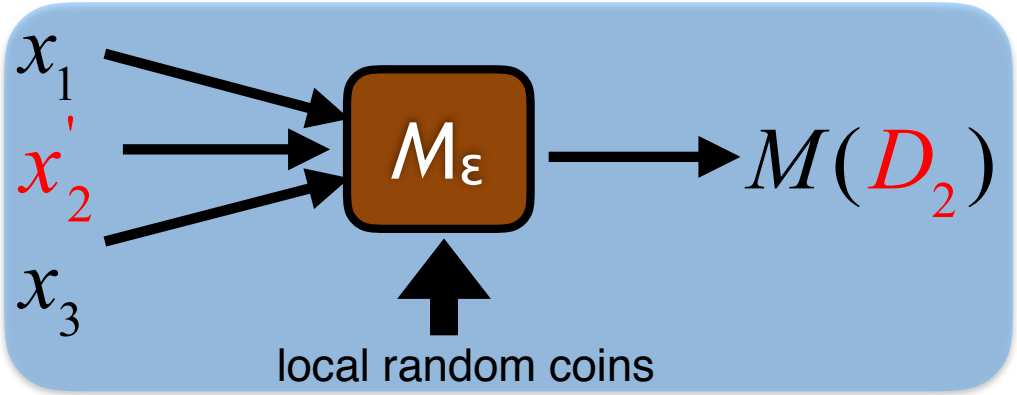
ε (epsilon) is a privacy parameter

lower ε = 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: neighboring databases





A randomized algorithm ***M*** provides **ε-differential privacy** if, for all neighboring databases ***D₁*** and ***D₂***, and for any set of outputs ***S***:

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

Think of database of respondents ***D=(x₁, .., xₙ)*** as **fixed** (not random), ***M(D)*** is a random variable distributed over possible outputs

Neighboring databases induce **close distributions** on outputs

based on a slide by Adam Smith

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***:

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

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

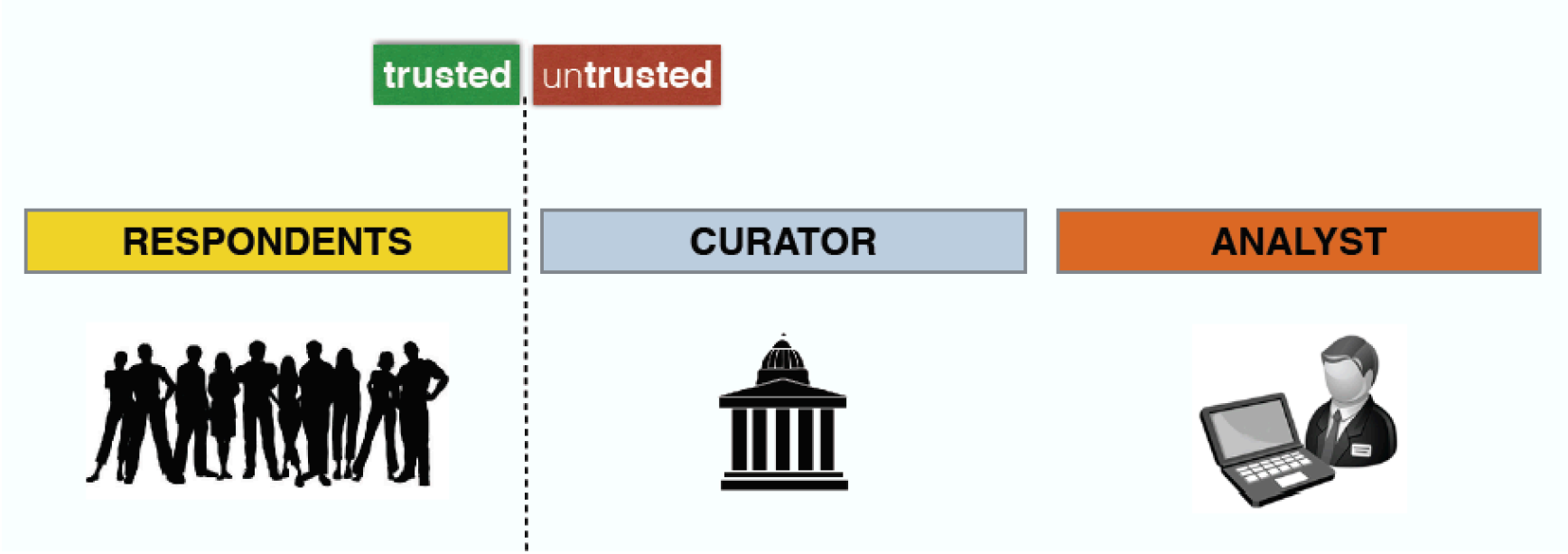
$$\Pr[A \mid \neg P] = \Pr[HH] = \frac{1}{4}$$

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

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

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

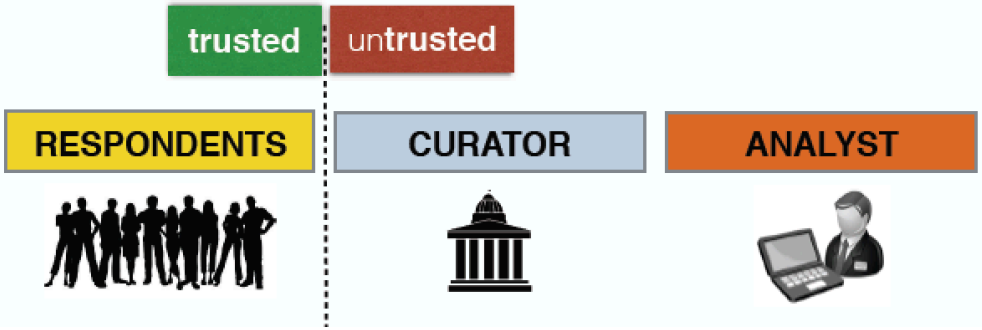
slide by Gerome Miklau

Differential privacy in the field

Apple

Google

Google Chrome



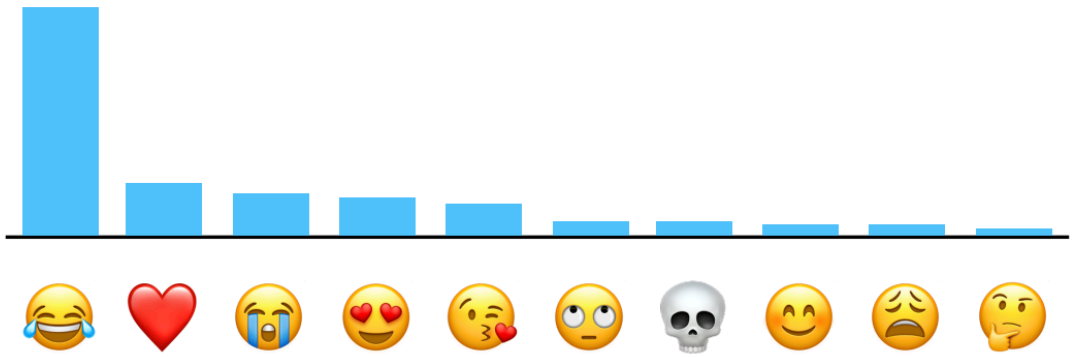
slide by Gerome Miklau

Apple uses local differential privacy

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.

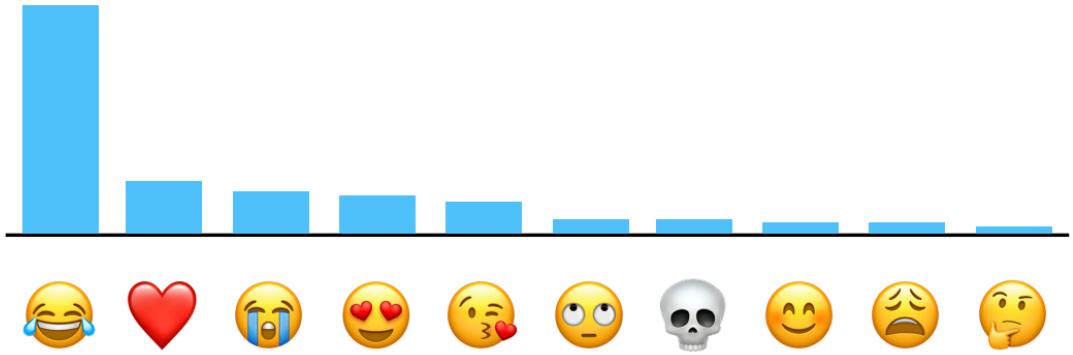


https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Apple uses local differential privacy

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)

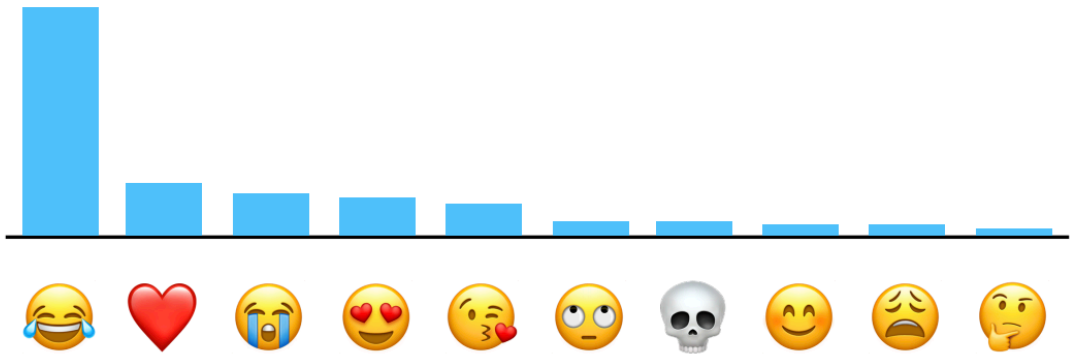


https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Apple uses local differential privacy

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

Apple uses local differential privacy

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

Apple uses local differential privacy

WIRED

ANDY GREENBERG SECURITY 09.15.2017 09:28 AM

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/

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***:

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

ε (epsilon) is a privacy parameter

lower ε means 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***:

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

ε (epsilon) is a privacy parameter

lower ε means 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

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

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

- 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 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	?

Query sensitivity

The 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 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	?

Query sensitivity

The 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)

Query sensitivity

The 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	?

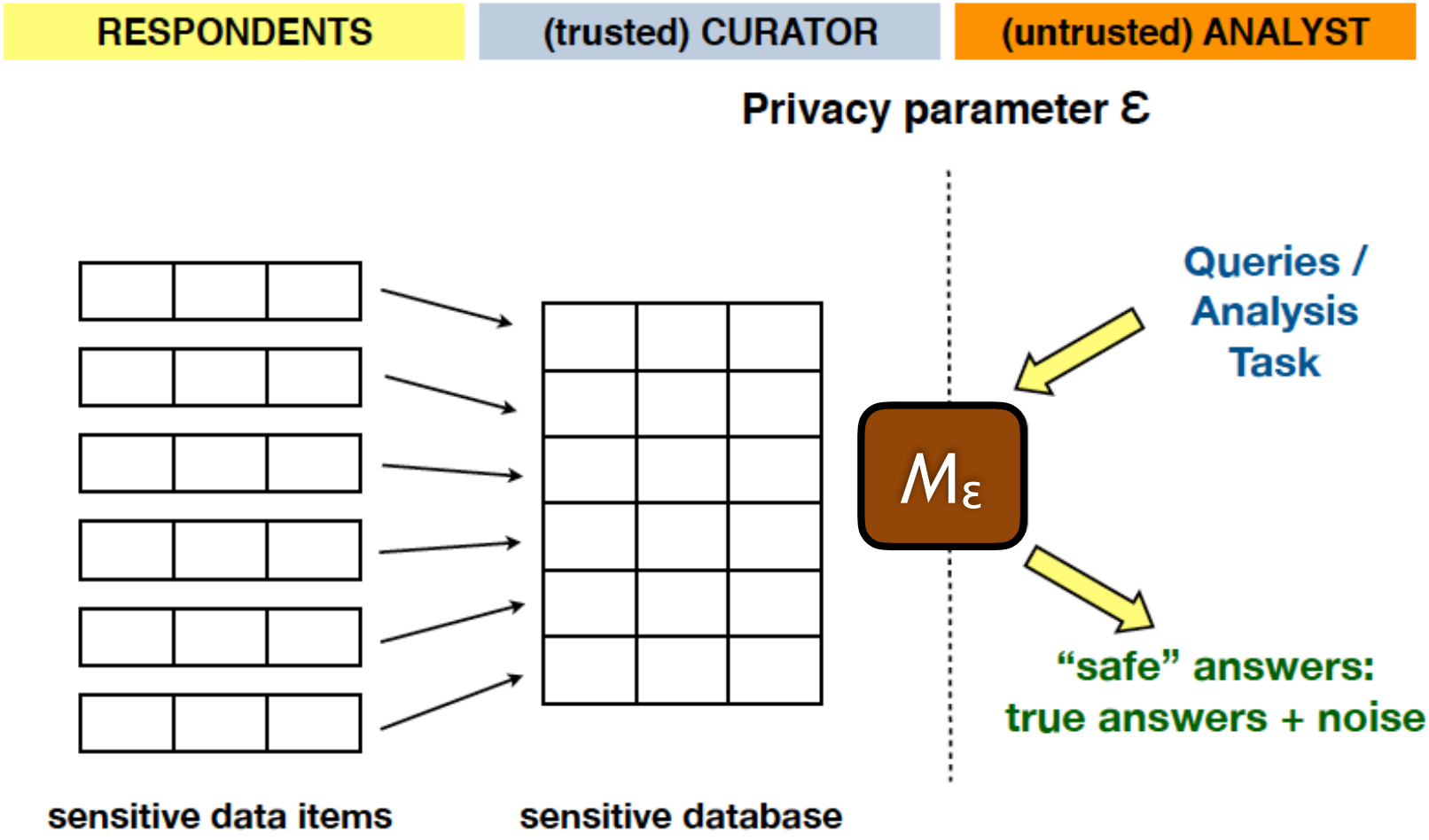
Query sensitivity

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

Adding noise



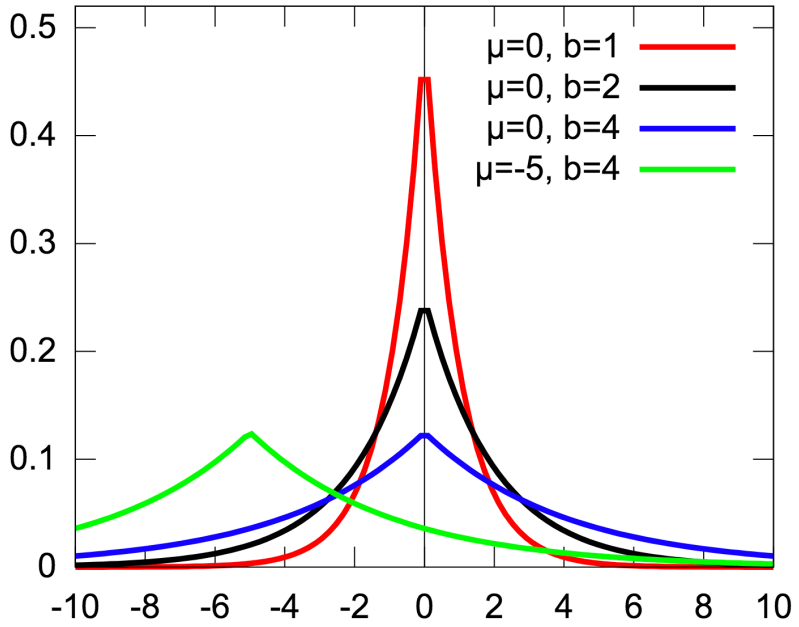
slide by Gerome Miklau

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:

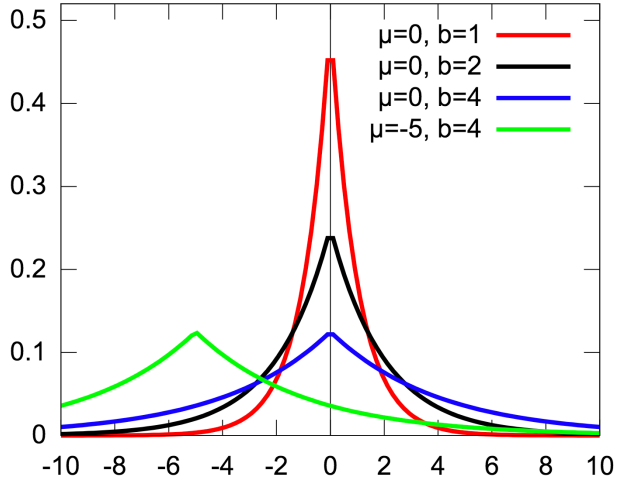


https://en.wikipedia.org/wiki/Laplace_distribution

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

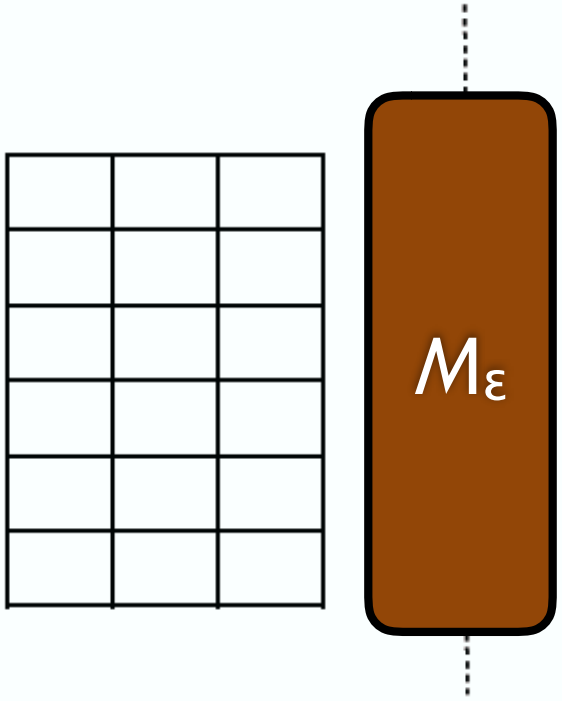
lower ε = stronger privacy

Adding noise



(trusted) CURATOR | (untrusted) ANALYST

$\varepsilon = 1$



sensitive database

$\varepsilon = 1$

Count(sex=Male, age=18)

true answer + noise(-3,3)

Laplace noise centered at 0, in interval (-3,3) with 95% prob.

slide by Gerome Miklau

Query sensitivity

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

Composition

If algorithms M1 and M2 are $\boldsymbol{\varepsilon}$-differentially private, then outputting results of both algorithms is $\mathbf{2\varepsilon}$-differentially private

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)

Sequential composition

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

- $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = 1.2$ That is: all queries together are ε-differentially private for ε =1.25. **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 ε =max($\varepsilon_1, \ldots, \varepsilon_k$)
  - Q1: select count(*) from D under $\varepsilon_1 = 0.5$
  - Q2: select count(*) from D where sex = Male under $\varepsilon_2 = 0.2$
  - Q3: select count(*) from D where sex = Female under $\varepsilon_3 = 0.25$
  - Q4: select count(*) from D where age > 20 under $\varepsilon_4 = 0.25$
- $\varepsilon = \varepsilon_1 + max(\varepsilon_2, \varepsilon_3) + \varepsilon_4 = 1$ That is: all queries together are ε-differentially private for ε =1.

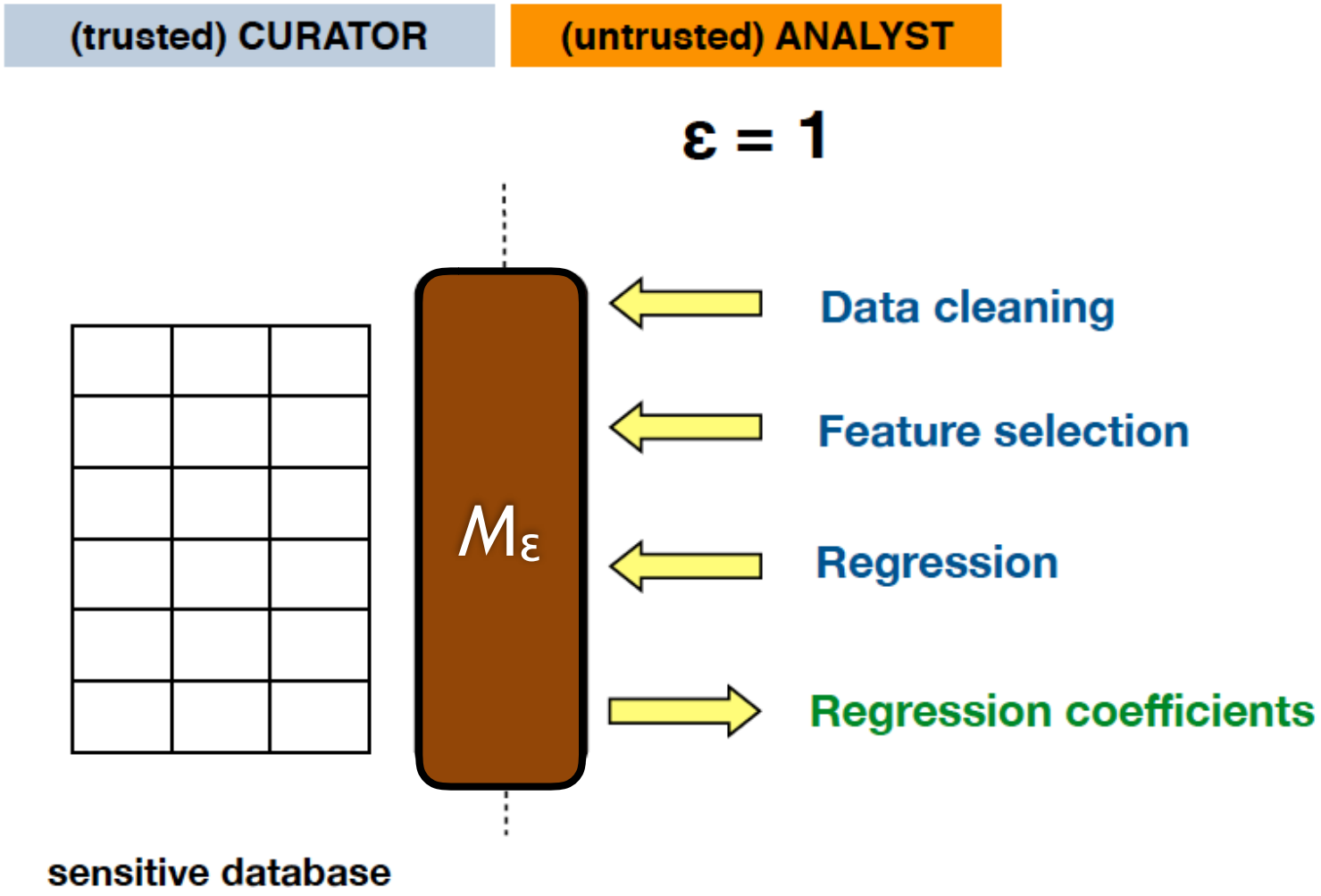
Composition and consistency

- 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**!

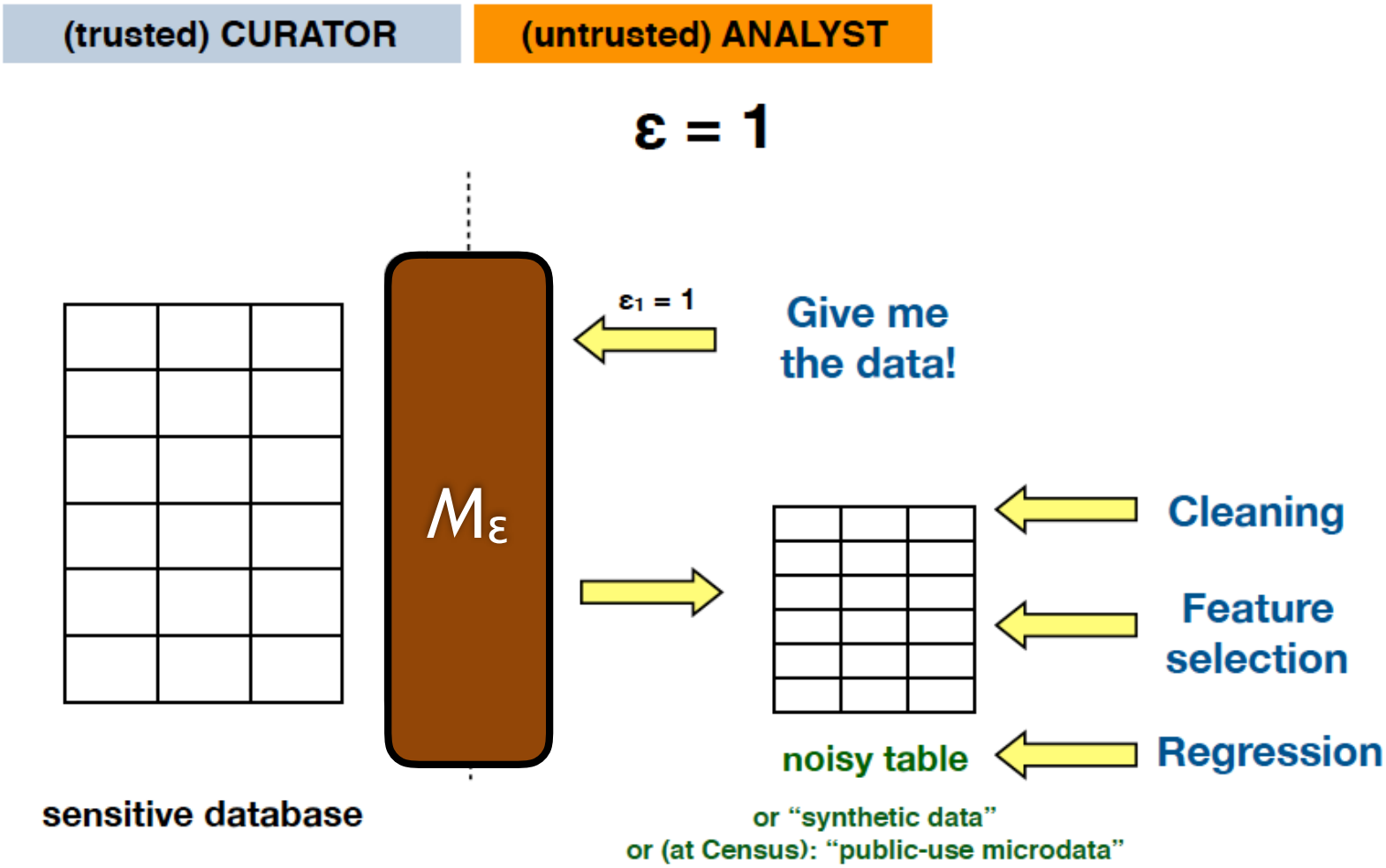
Also don't want any negative counts + may want to impose datatype checks, e.g., no working adults with age = 5 etc.

Entire workflow must be DP



slide by Gerome Miklau

Privacy-preserving synthetic data

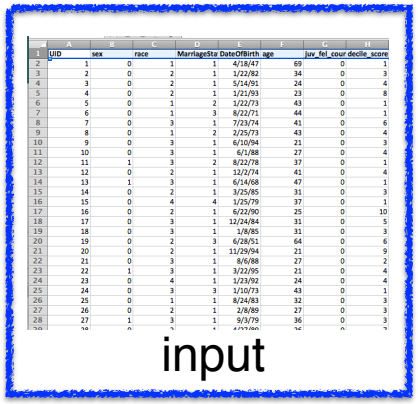


slide by Gerome Miklau

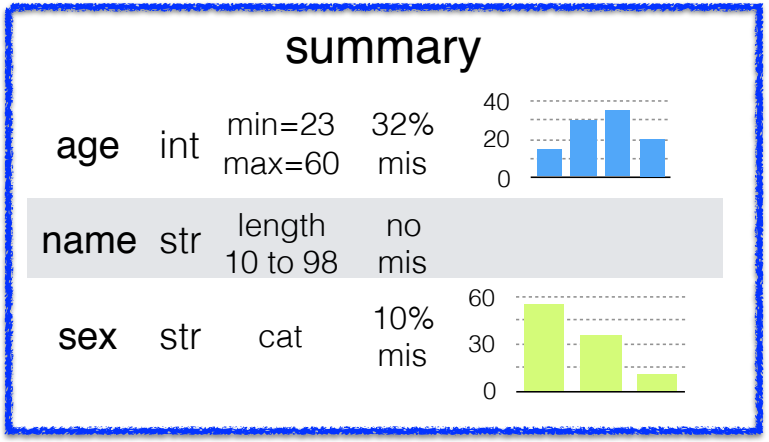
Data Synthesizer

[Ping, Stoyanovich, Howe *2017*]

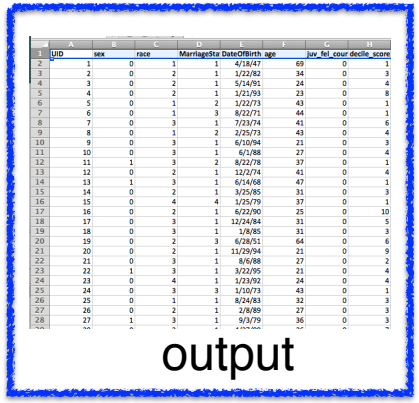
http://demo.dataresponsibly.com/synthesizer/



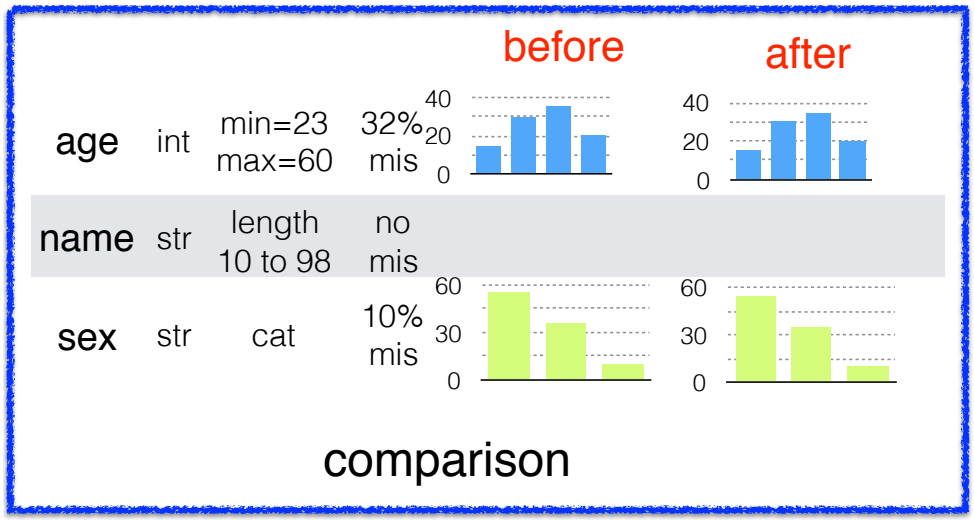
Data Describer



Data Generator



Model Inspector



Privacy-preserving synthetic data, generally

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

Data Synthesizer

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

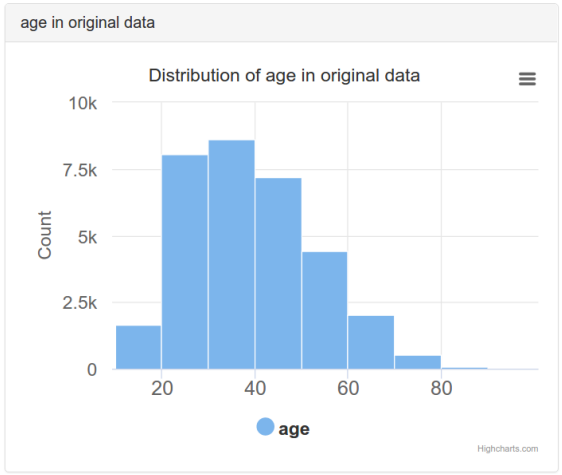
- Main goal: **usability first**
  - user is the data owner
  - the tool picks up data types from the input file: categorical / string / numerical (integer, float) / date-time
  - the tool computes the frequency of missing values per attribute
  - user can then inspect the result, over-ride what was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

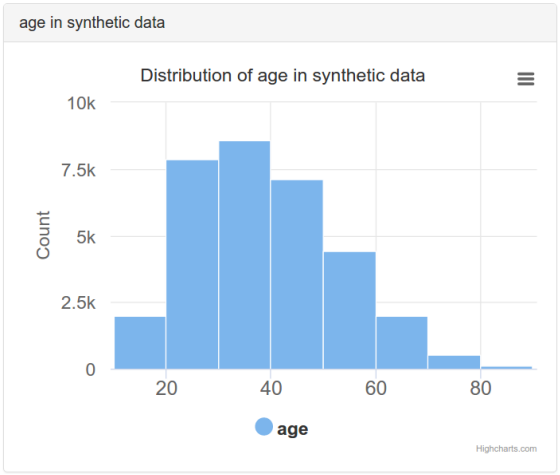
- The tool generates an output dataset of a specified size, in one of three modes
  - **random** - type-consistent random output
  - **independent attribute** - learn a noisy histogram for each attribute
  - **correlated attribute** - learn a noisy Bayesian network (BN)

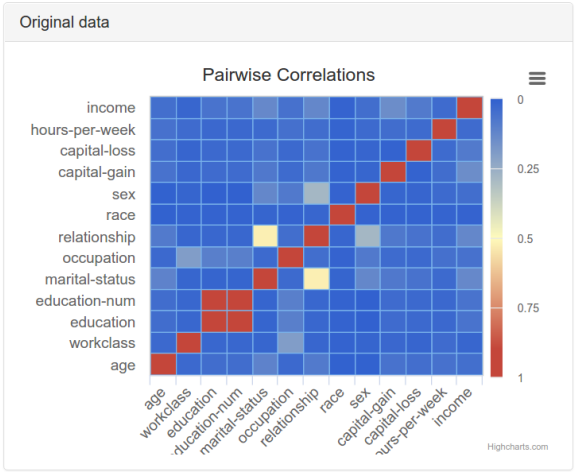
Data Synthesizer

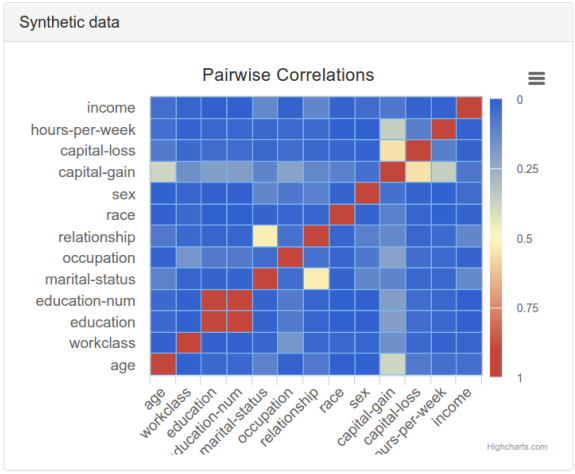
[Ping, Stoyanovich, Howe *2017*]

http://demo.dataresponsibly.com/synthesizer/







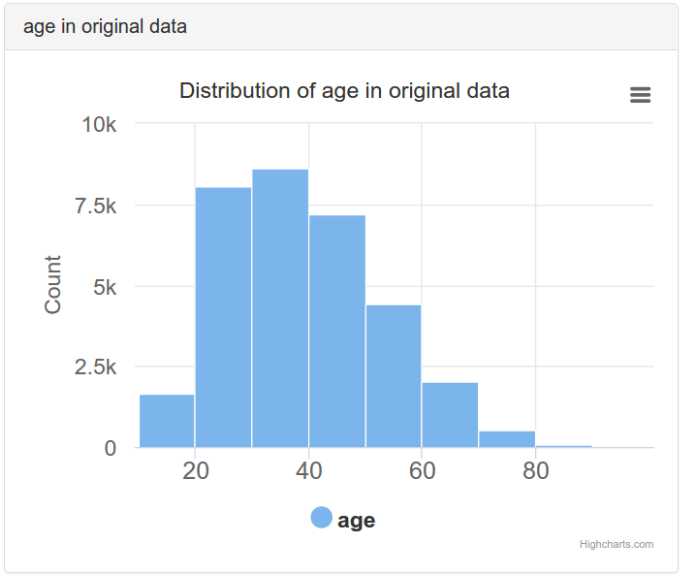


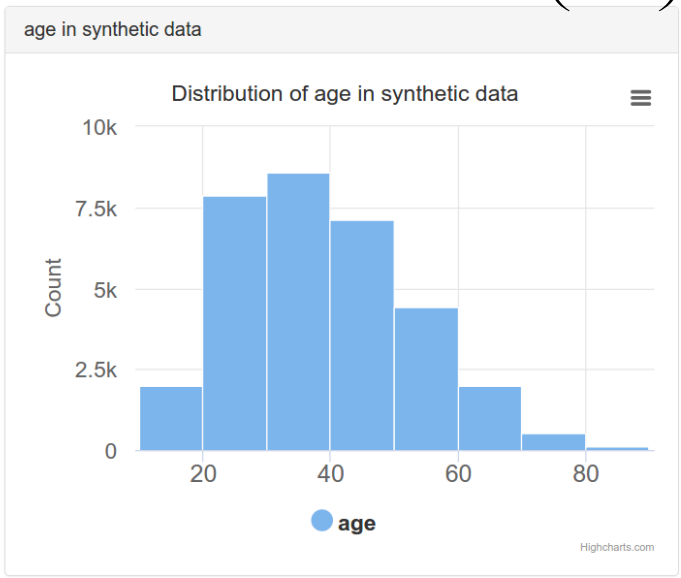
Data Synthesizer: independent attributes

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

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

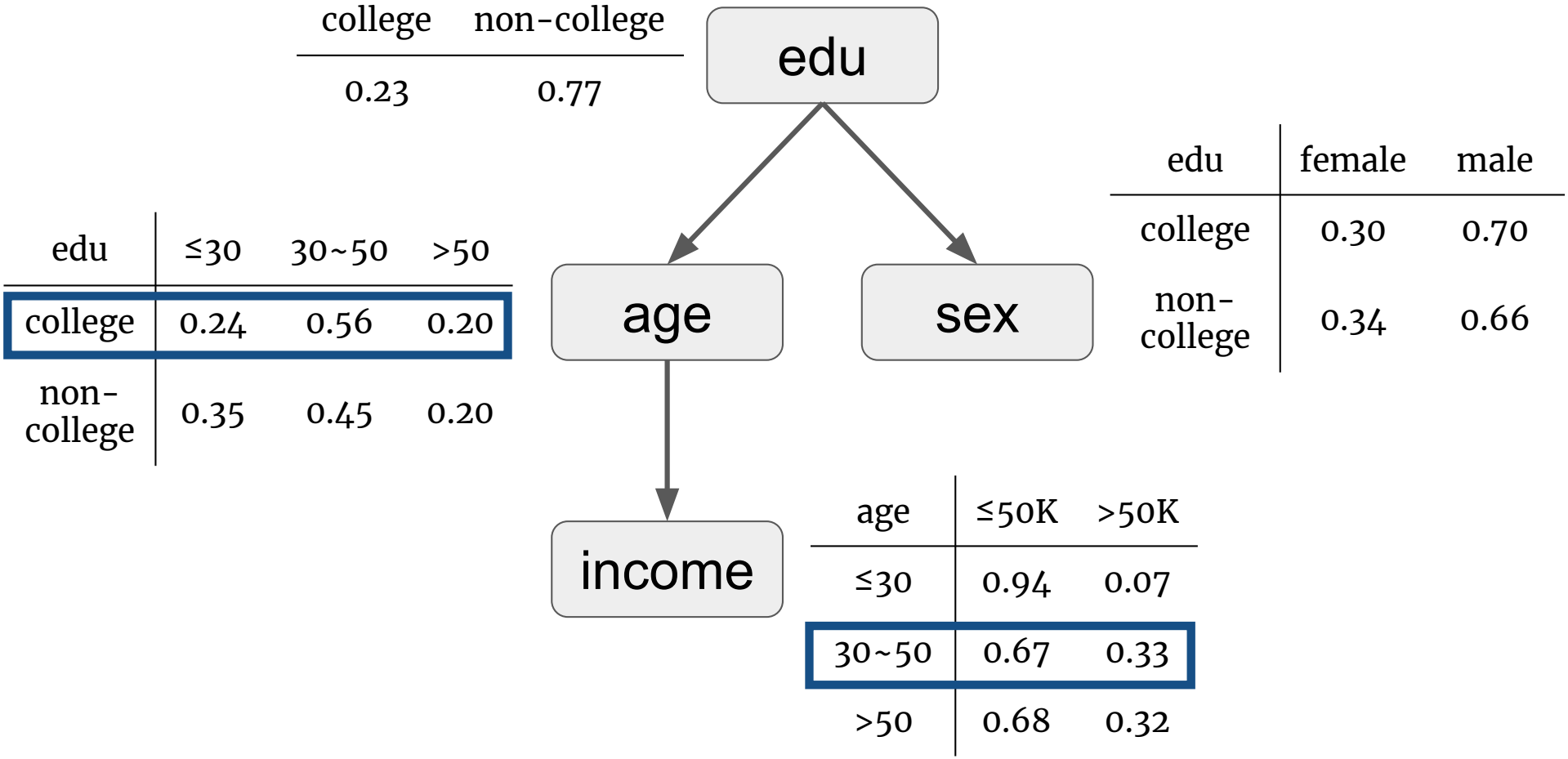
[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

- 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)

Data Synthesizer: correlated attributes

K=1

note that this is not a causal, BN!

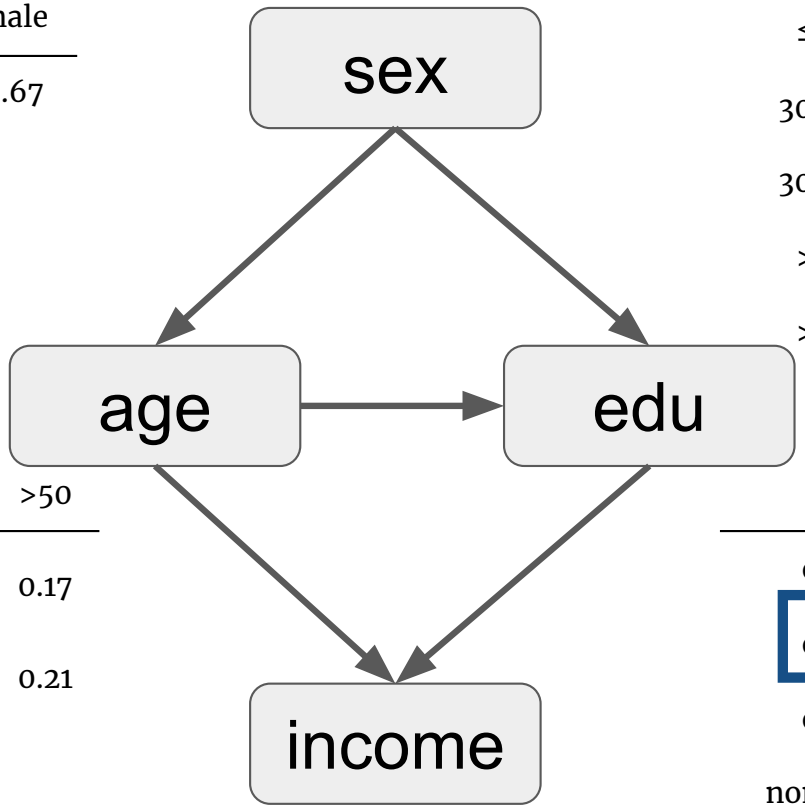


Data Synthesizer: correlated attributes

K=2

note that this is not a causal, BN!

female	male
0.33	0.67

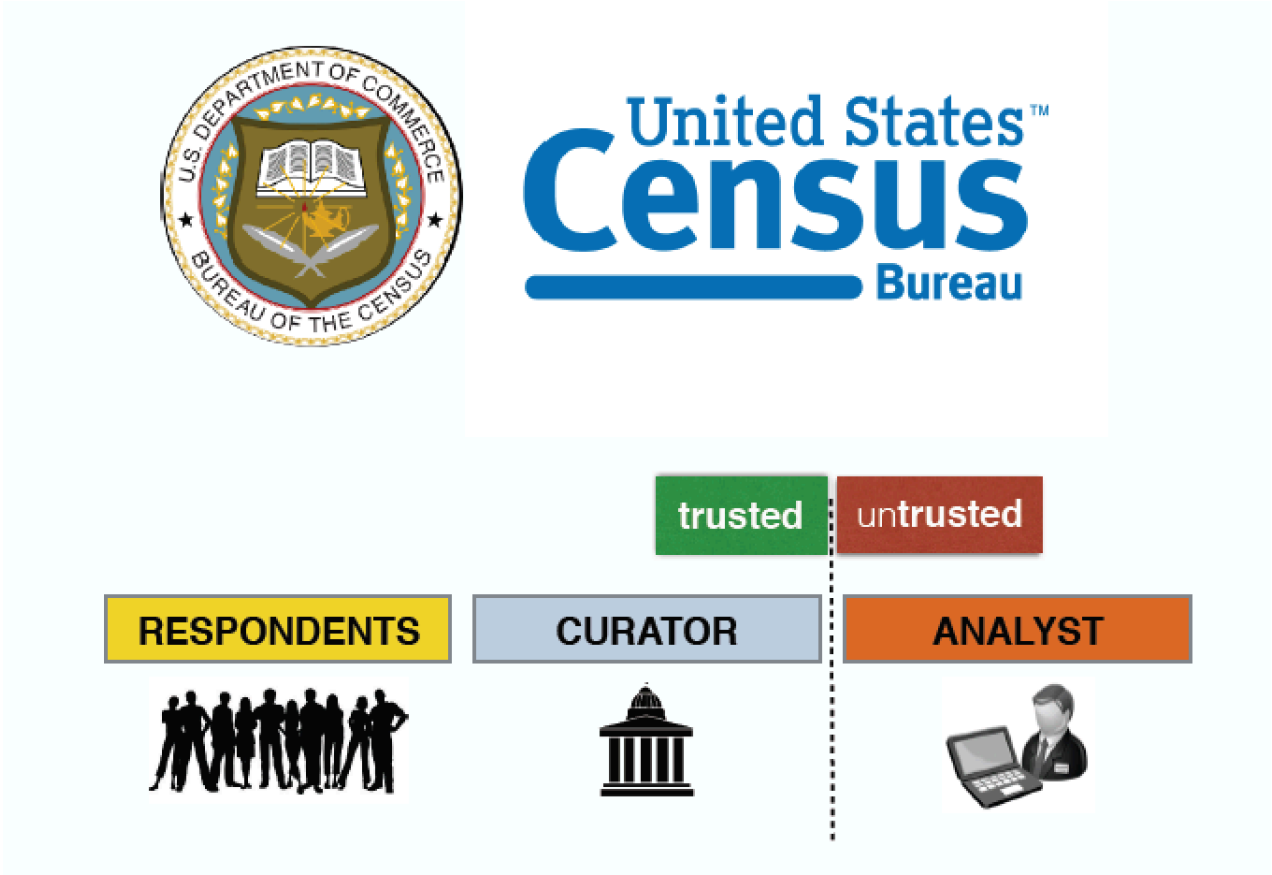


age	sex	college	non-college
≤30	female	0.18	0.82
≤30	male	0.16	0.84
30~50	female	0.25	0.75
30~50	male	0.28	0.72
>50	female	0.17	0.83
>50	male	0.25	0.75

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

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

Differential privacy in the field



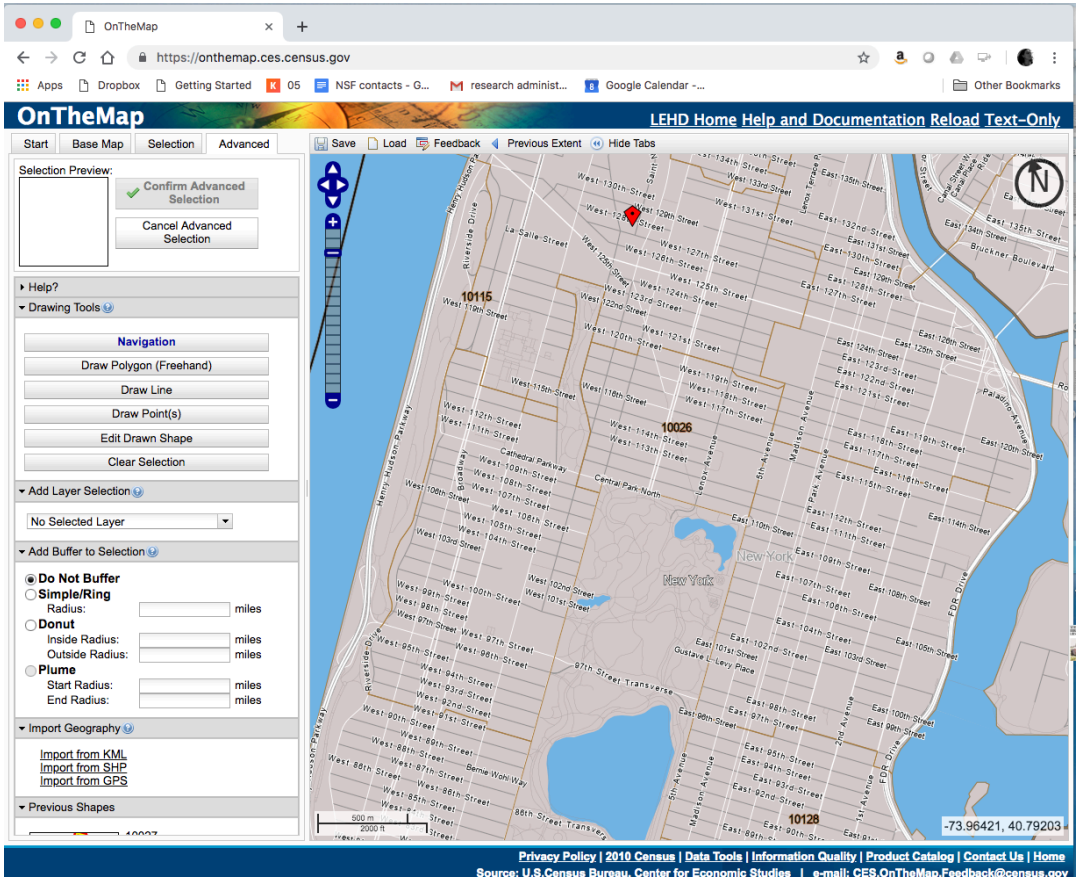
Decennial Census 2020

slide by Gerome Miklau

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



TheUpshot

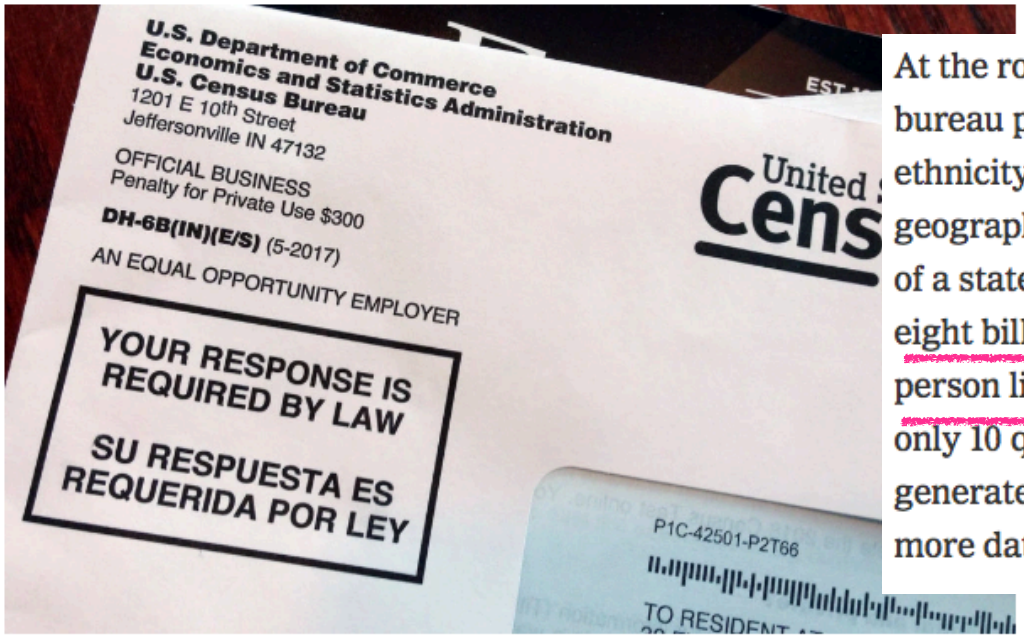
To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

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**

		AGE		
STATISTIC	**GROUP**	**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

https://queue.acm.org/detail.cfm?id=3295691

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

https://queue.acm.org/detail.cfm?id=3295691

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.

Differential privacy in 2020 Census: pushback

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?