*DS-GA 3001.009: Responsible Data Science*

# Privacy and Data Protection

Prof. Julia Stoyanovich
Center for Data Science
Computer Science and Engineering at Tandon

@stoyanoj

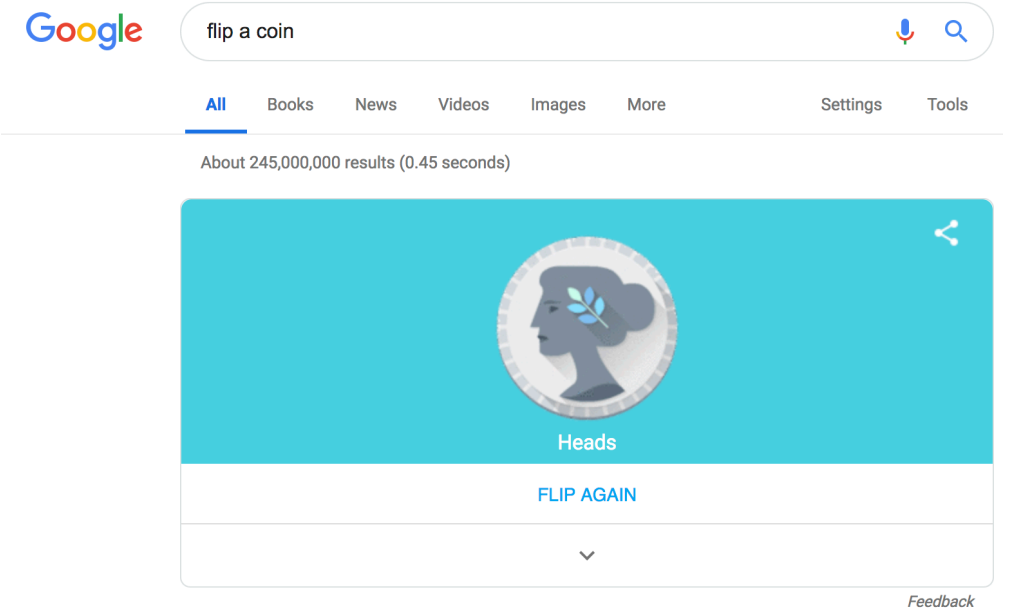
http://stoyanovich.org/
https://dataresponsibly.github.io/

# 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

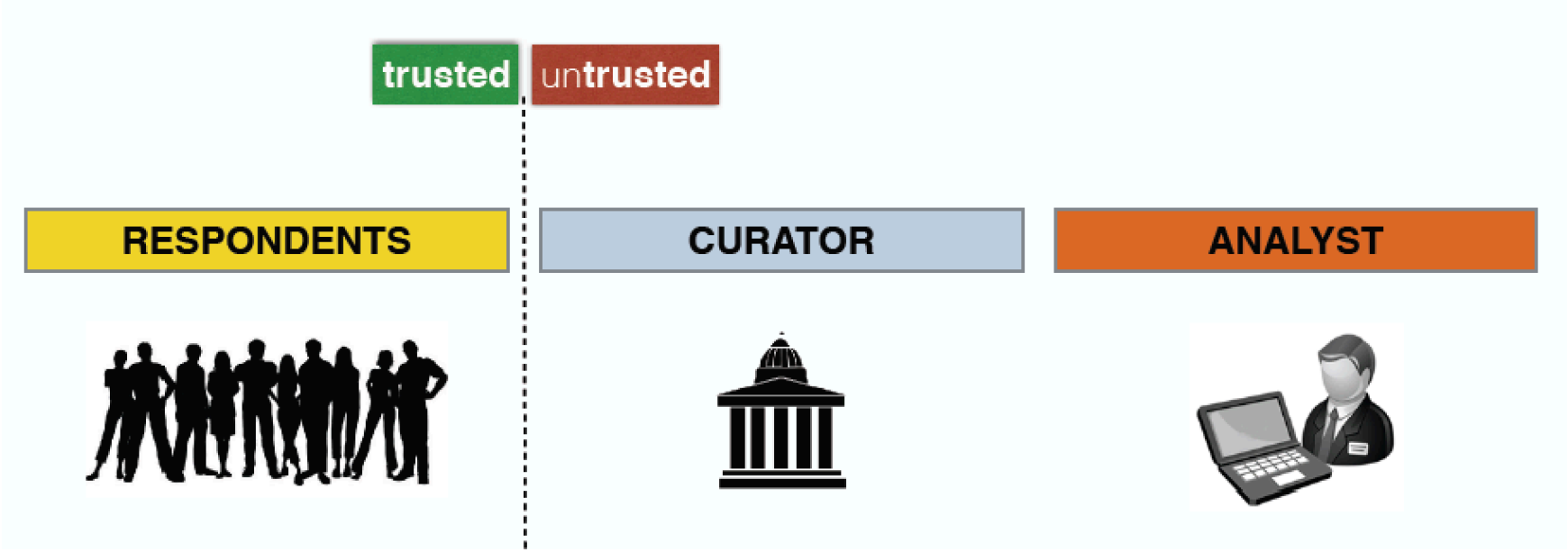
protecting an individual

plausible deniability


learning about the population

noisy estimates

# Privacy-preserving data analysis



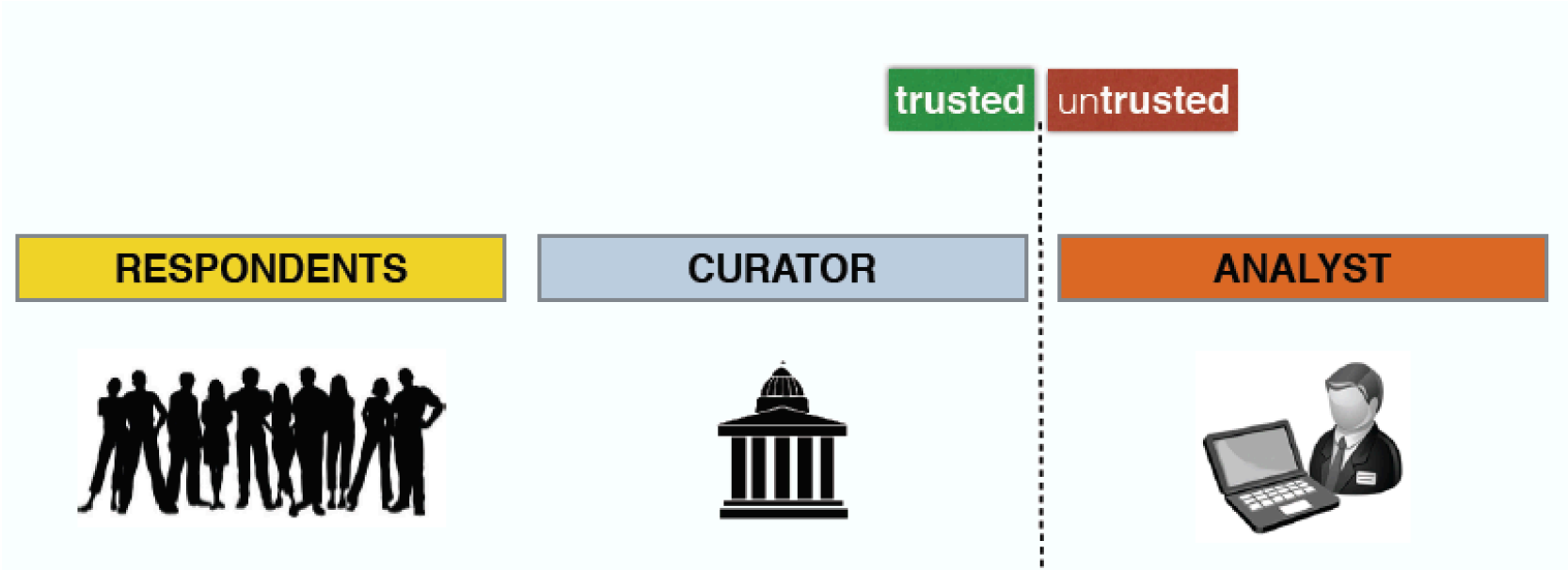

**respondents** contribute their personal data

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



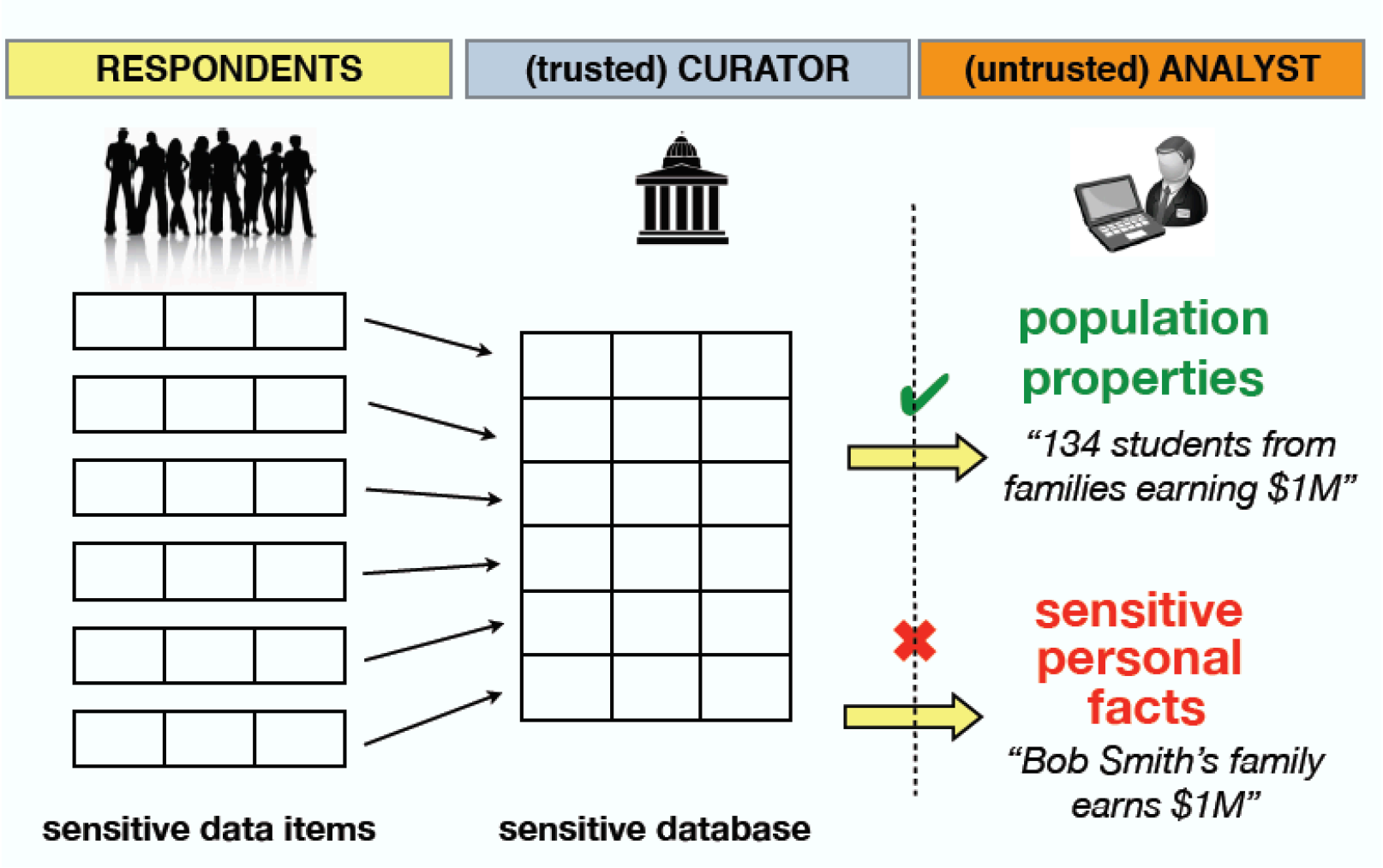

**respondents** in the population seek protection of their personal data

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

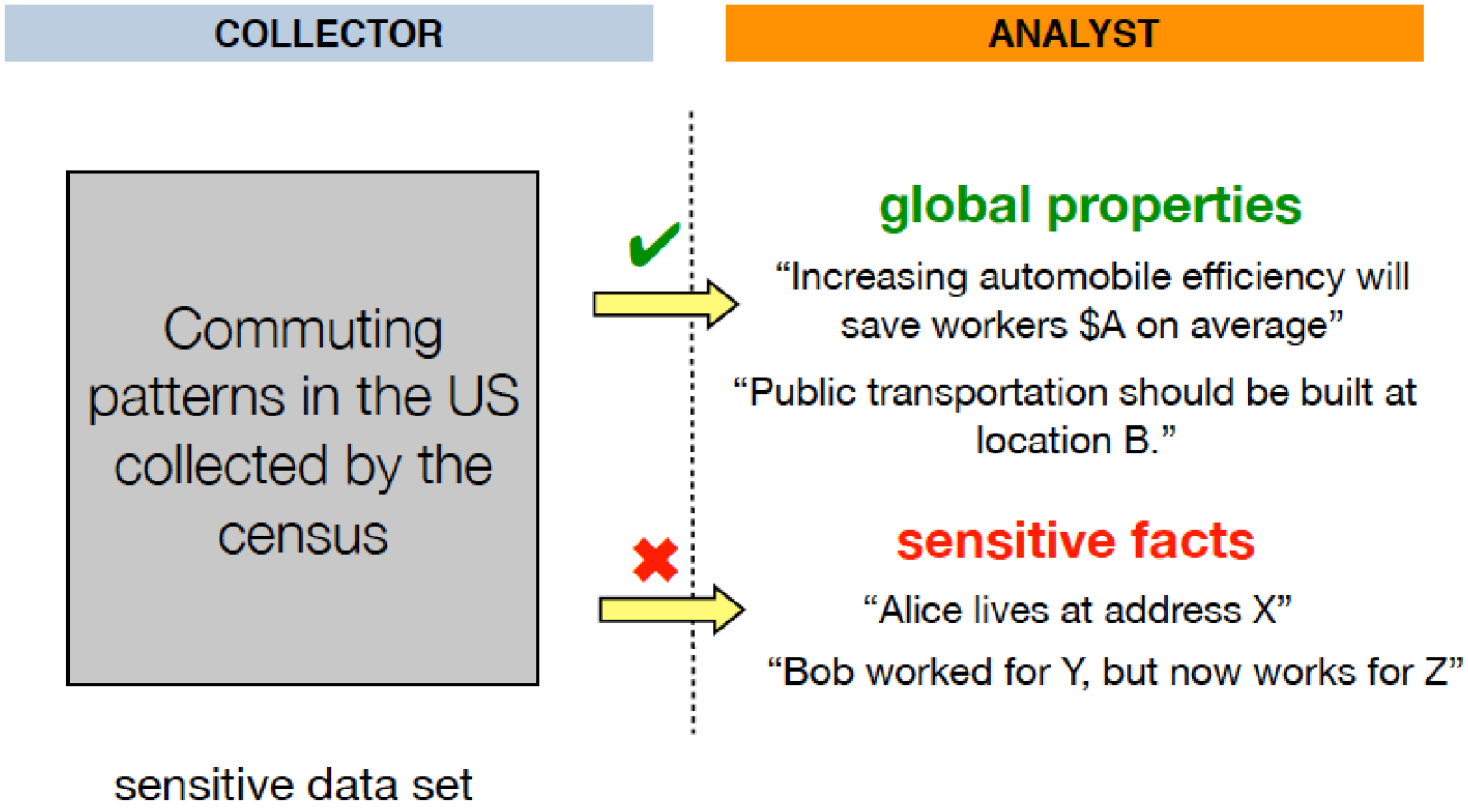
slide by Gerome Miklau

# Privacy-preserving data analysis




slide by Gerome Miklau

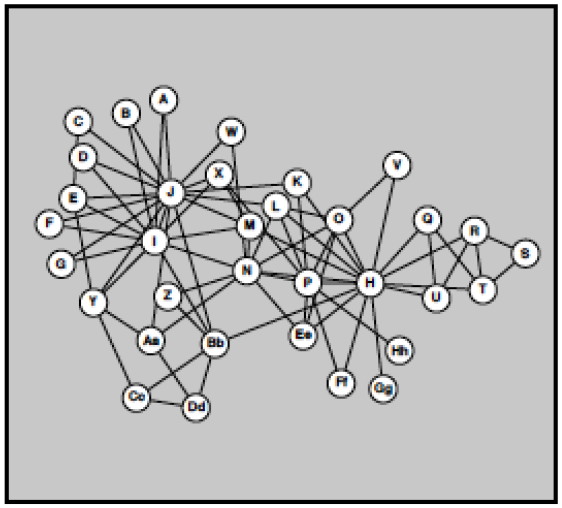
# Example: Census data




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**. We'll return to this when discussing the **Barrow, Alaska alcohol 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 data set 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 $\ell_1$ norm of a database ***D***, denoted $\|D\|_1$ is the number of tuples in ***D***.

The $\ell_1$ 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 data set 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

# 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***:

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

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

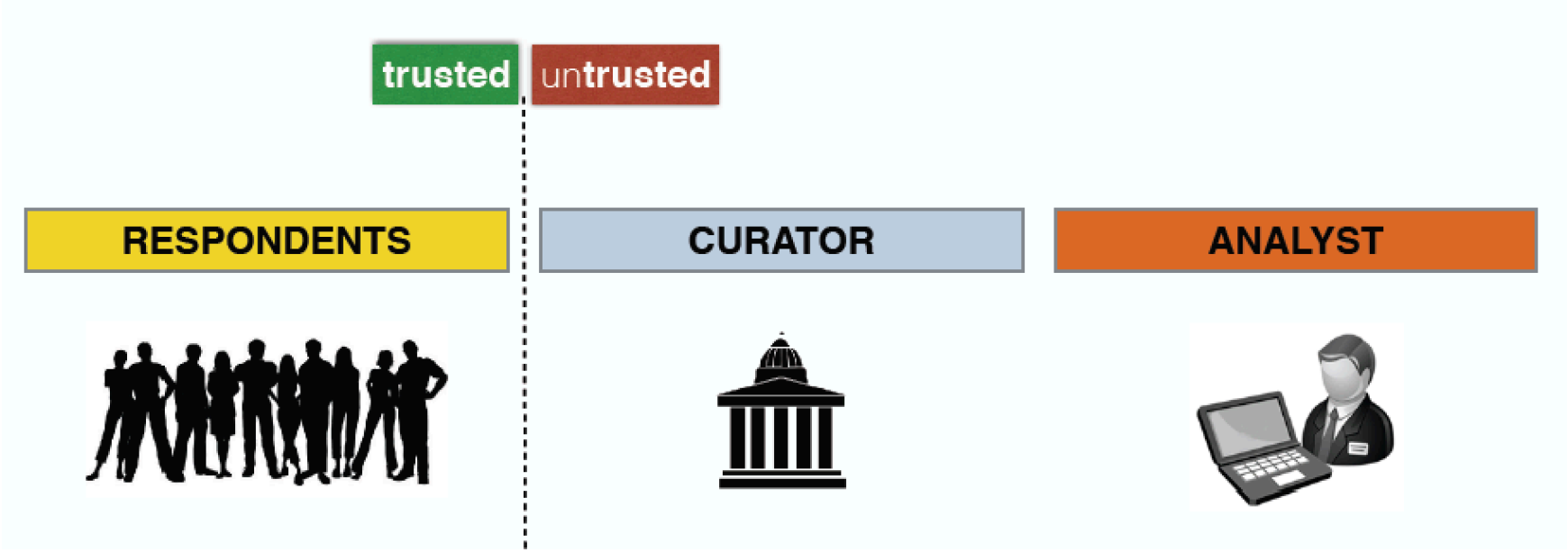
$$\Pr[A \mid \neg P] = \Pr[HT] = \frac{1}{4}$$

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

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

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

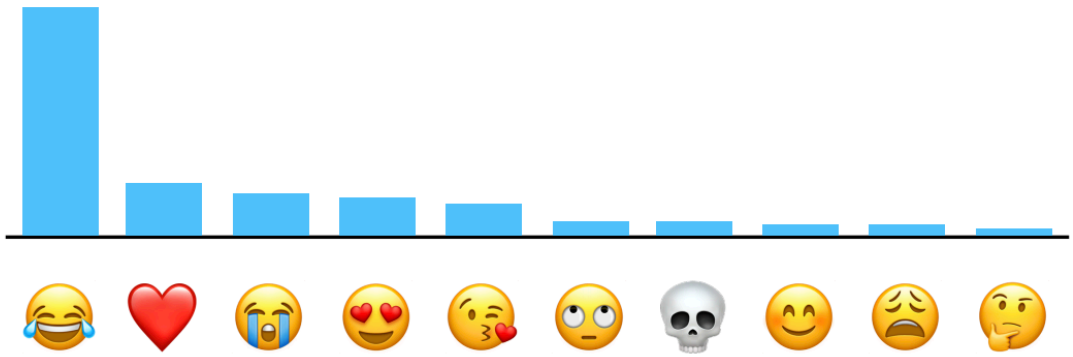
slide by Gerome Miklau

# 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.

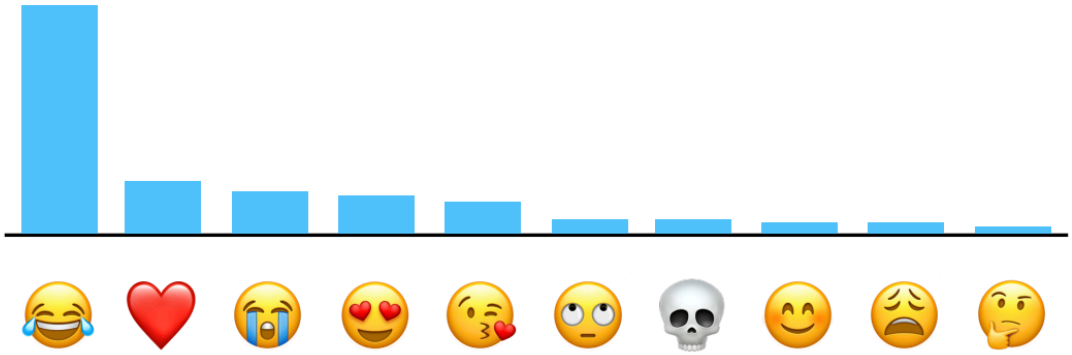



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

# 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)

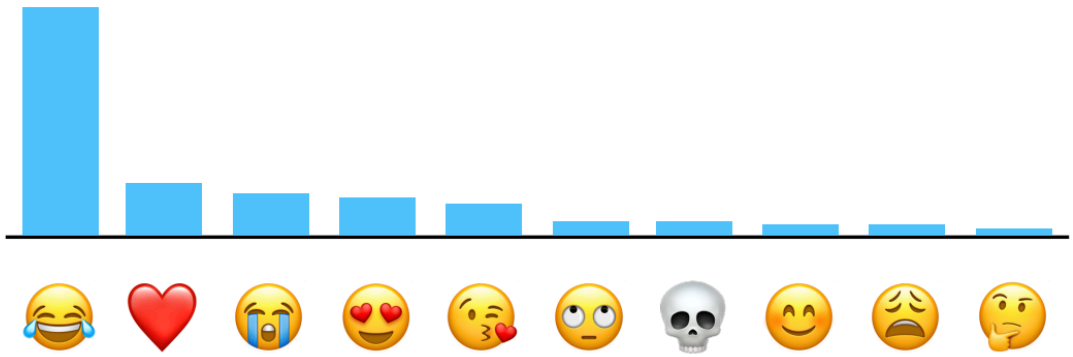



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

# 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 $\varepsilon$ 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

# Do we really need randomization?

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

# Aggregation without randomization

- Alice and Bob are professors at State University. Both underwent "data safety" training.
- In March, Alice publishes an article: "…. the current freshman class at State University is made up of **3,005** students, **202** of whom are from families earning over $1M per year."
- In April, Bob publishes an article: "… **201** families in State University'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** 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**: (1) **X**: count the number of HIV-positive people in ***D***; (2) **Y**: count the number of HIV-positive people in ***D*** not named *Freddie*; (3) **X - Y** tells you whether *Freddie* is HIV-positive

# 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**

CCC
Computing Community Consortium
Catalyst

January 2017

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

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

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

# 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!

# 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

# 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 for 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.

# 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

# Query sensitivity

The $\ell_1$ 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**

***Δq* gives an upper bound on how much we must perturb *D* to preserve privacy**

# 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<br>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<br>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<br>where sex = Male and age > 30 | 1 |
| select MAX(salary) from D | *MAX(salary)-MIN(salary)* |
| select gender, count(*)<br>from D group by gender | ? |

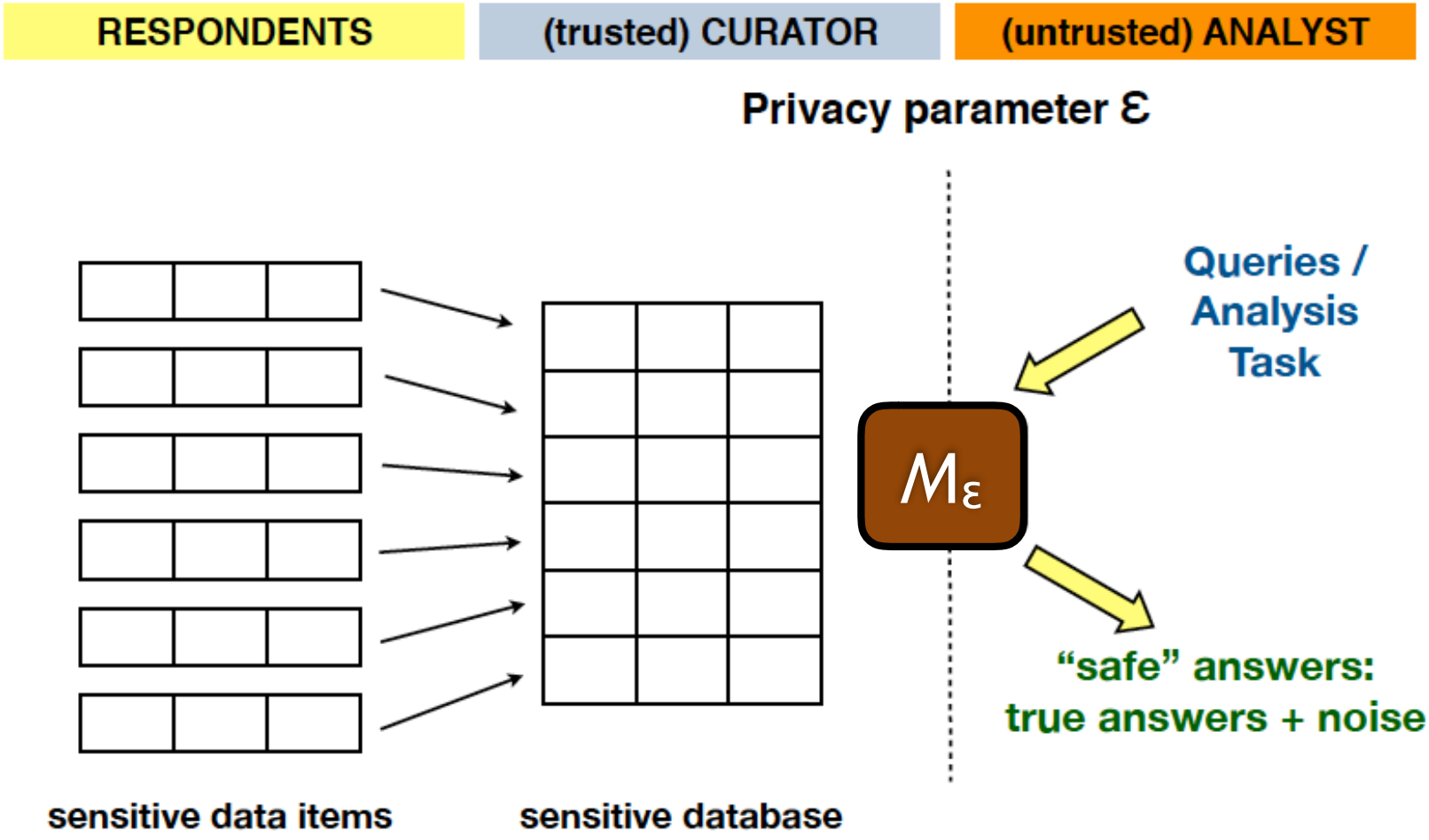
# Query sensitivity

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 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 | ? |

# Adding noise



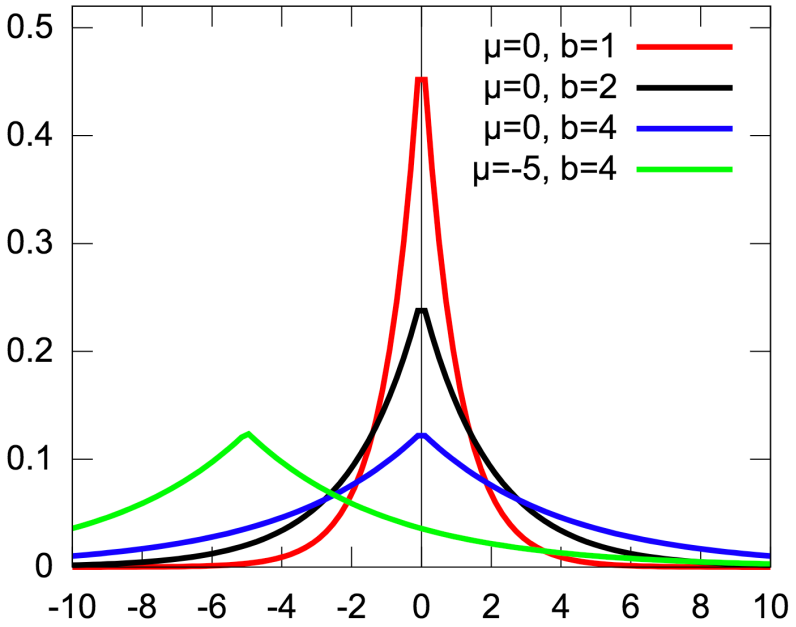

slide by Gerome Miklau

# 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:

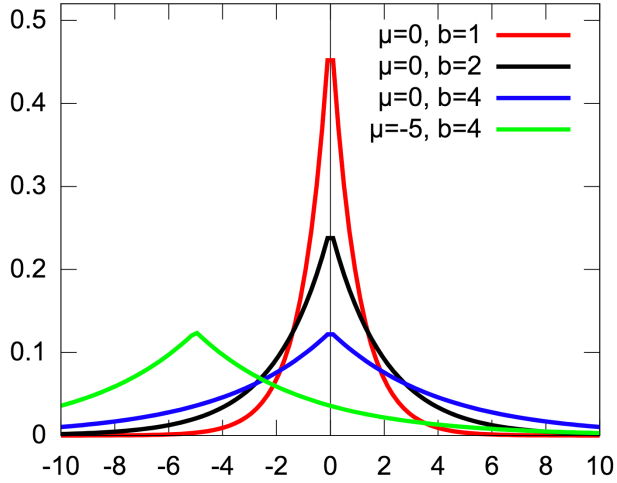



https://en.wikipedia.org/wiki/Laplace_distribution

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

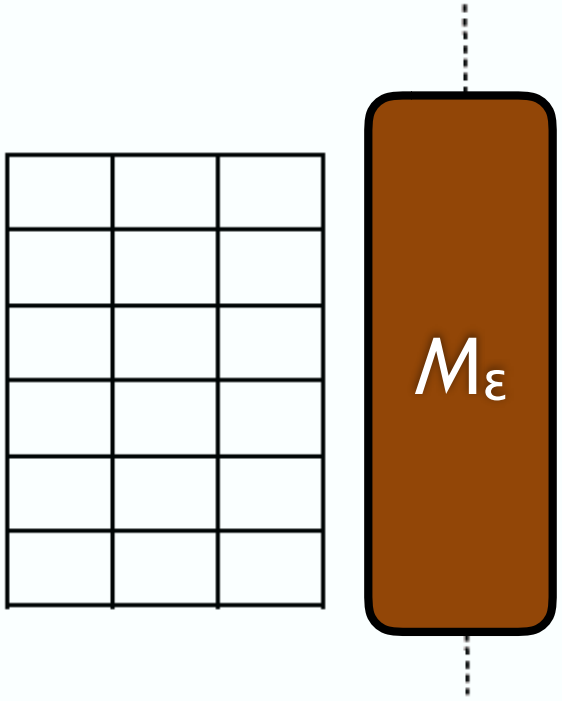
**lower ε = stronger privacy**

# 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) |
| **sequential composition** | |
| an arbitrary list of *m* counting queries | ? |

# 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($\boldsymbol{\varepsilon_1, \ldots, \varepsilon_k}$)
  - 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.

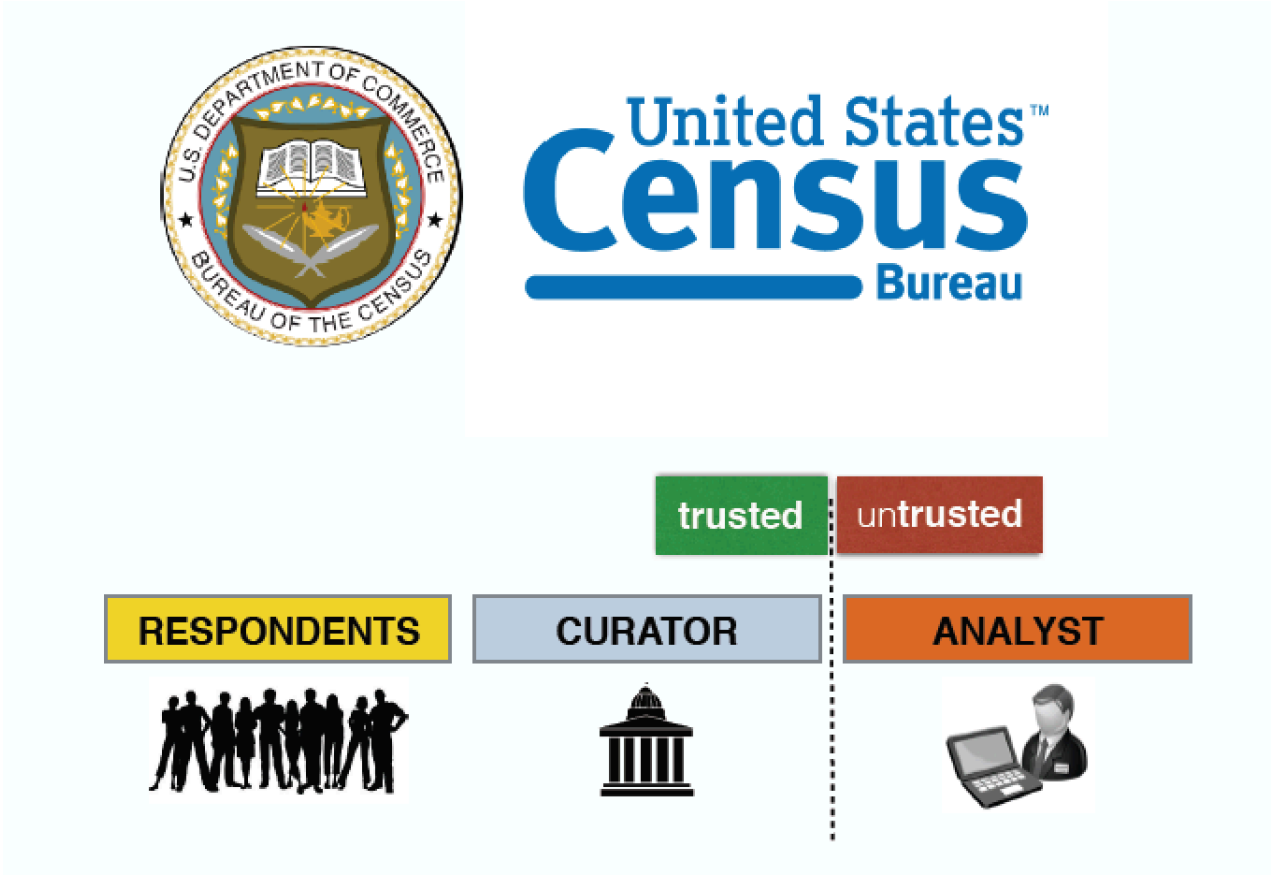
# Composition and consistency

- 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.

# Differential privacy in the field



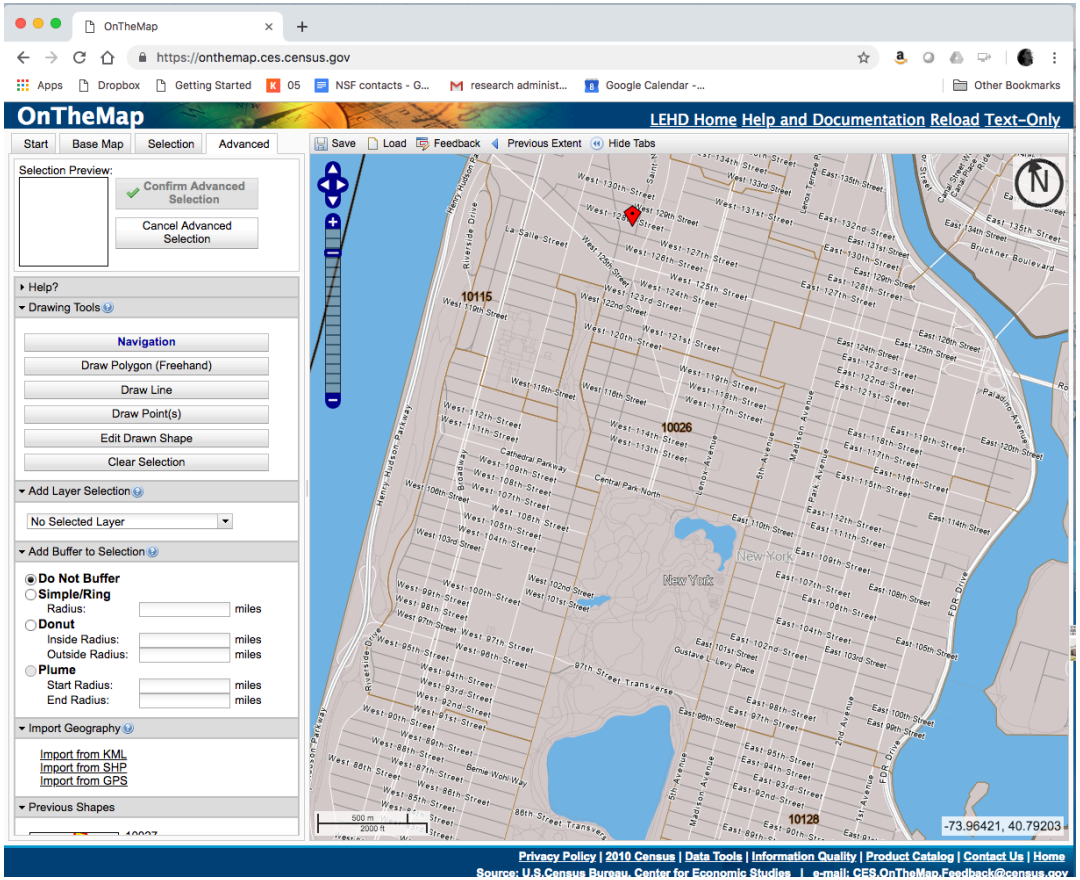

**current goals: Decennial Census 2020**

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

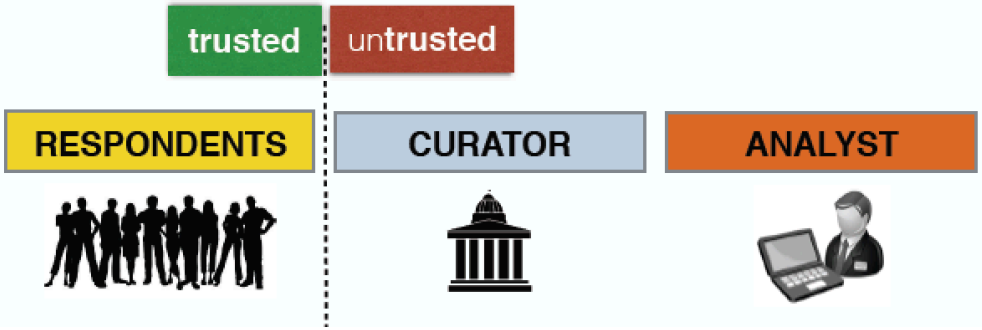


# Differential privacy in the field

Apple


Google

Google Chrome




slide by Gerome Miklau

# Detour: 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 (used 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”**

based on a slide by Bill Howe

# Harms and backlash

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.

**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.

based on a slide by Bill Howe

# 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."

**any chance of selection bias?**

- "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)
- No laws were broken, and harms are not about individual privacy!
- **Who benefits? Who is harmed?**

**data protection …. responsibility …. trust**

based on a slide by Bill Howe

# General Data Protection Regulation
# GDPR

Welcome to gdpr-info.eu. Here you can find the official PDF of the Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 as a neatly arranged website. All Articles of the GDPR are linked with suitable recitals. The European Data Protection Regulation is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe. If you find the page useful, feel free to support us by sharing the project.

## Quick Access