PrivateClean: Data Cleaning and Differential Privacy

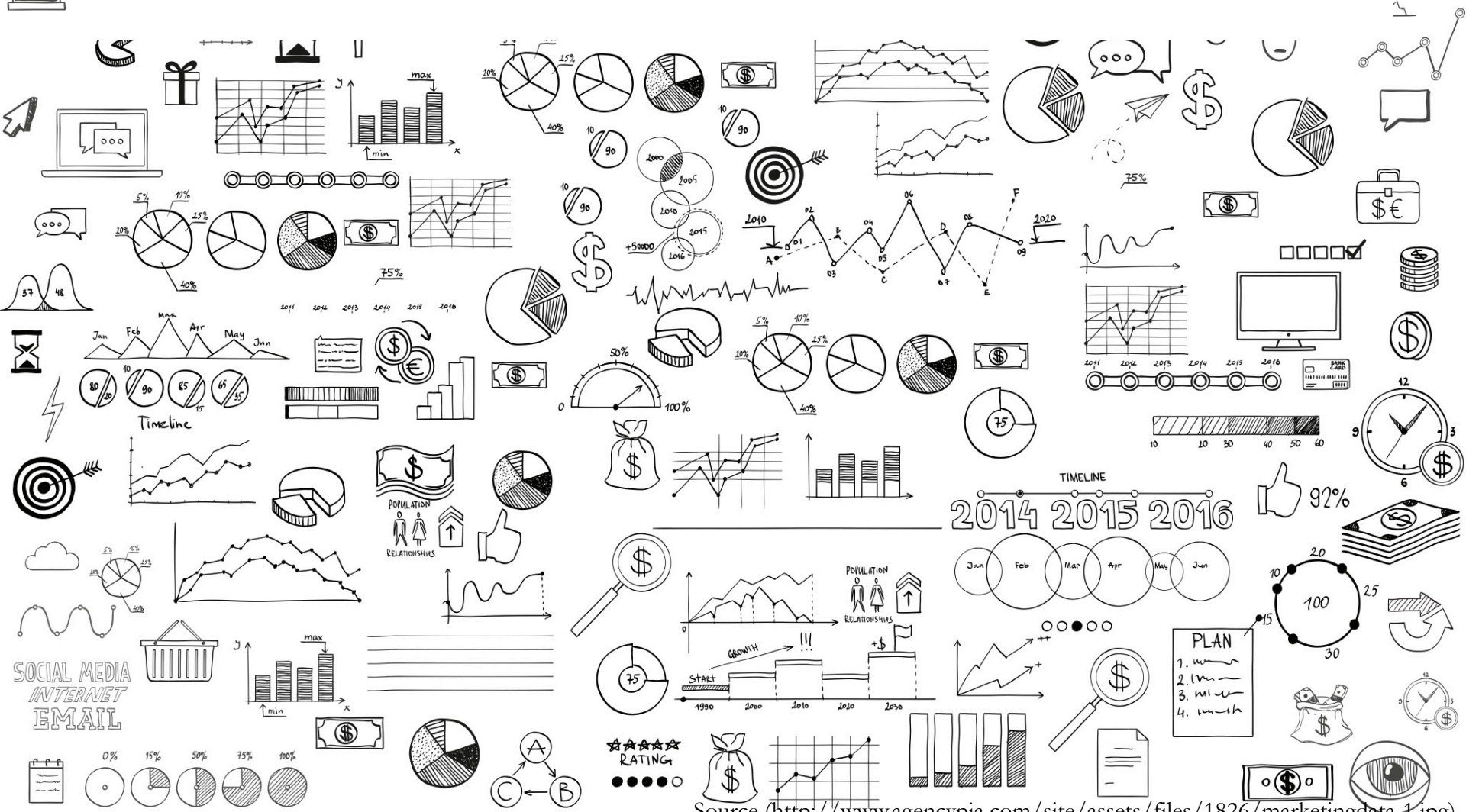
Data Cleaning & Integration

CompSci 590.01 Spring 2017

Presented by Xi He

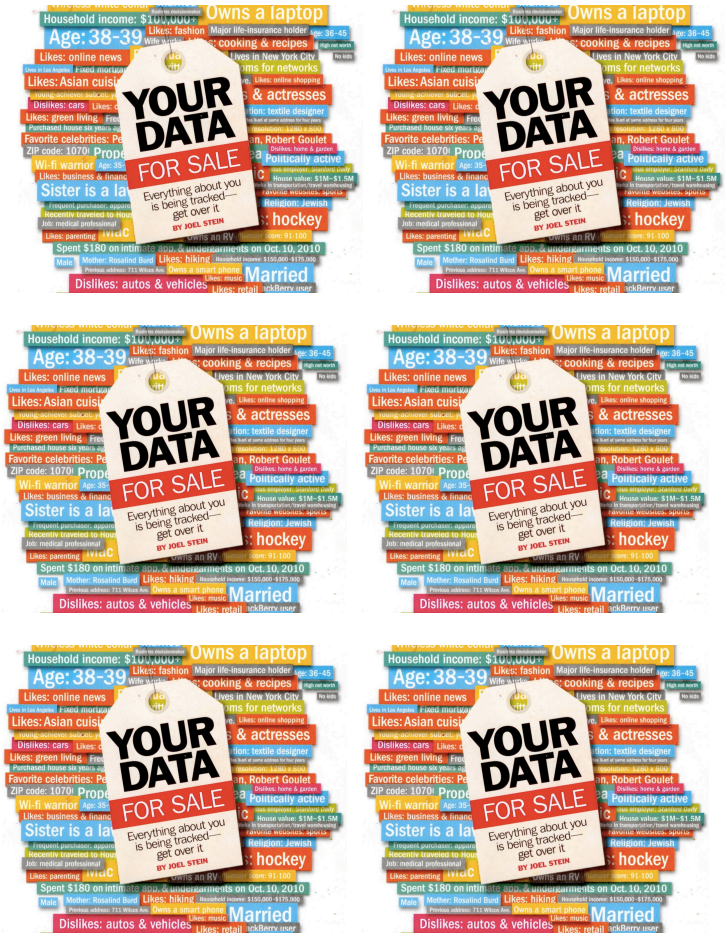
Some contents were based on Sanjay Krishnan et al.'s SIGMOD 2016 slides & Machanavajjhala et al.'s VLDB 2016 DP Tutorial slides

Our world is increasingly data driven



Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)

Aggregated Personal Data is **invaluable**



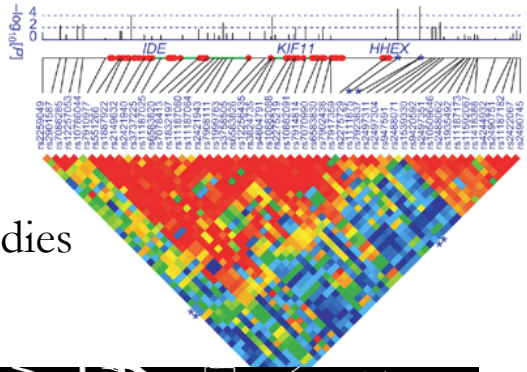
Advertising

Source (esri.com)

Dirty

Privacy

ome Wide
ociation Studies

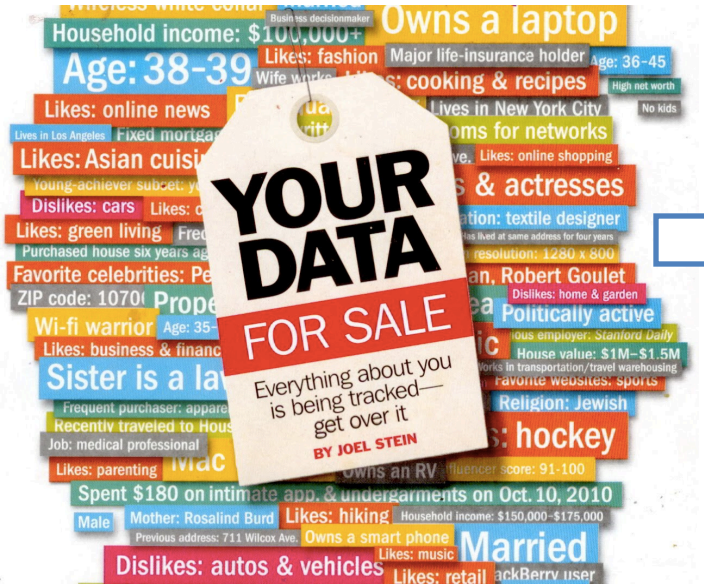


Human Mobility analysis

Outline

- A brief tour to differential privacy
- PrivateClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing
    - Correction for randomized response
    - Cleaning for dirty data
  - Evaluation
- Discussion

Personal data is … well … **personal**!



Source (time.com)

Age
Income
Address
Likes/Dislikes
Sexual Orientation
Medical History

Redlining

Discrimination

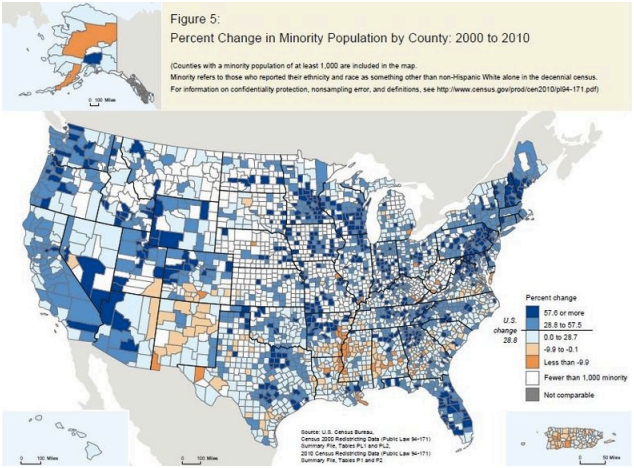
Physical/Financial Harm

Aggregated Personal Data …

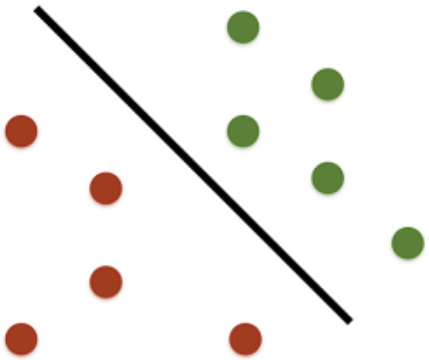
… is made publicly available in many forms.

De-identified records
(e.g., medical)

Statistics
(e.g., demographic)



Predictive models
(e.g., advertising)



That's fine … I am anonymous!

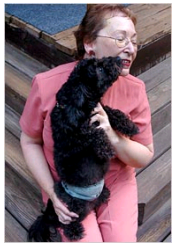
Source (http://xkcd.org/834/)

Anonymity is not enough …

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS



NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The Scientist » The Nutshell

"Anonymous" Genomes Identified

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

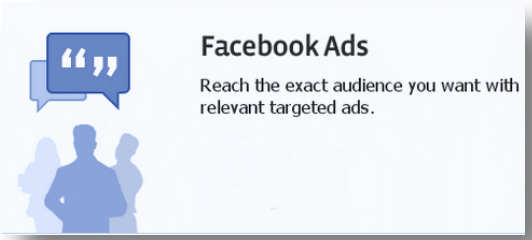
By Dan Cossins | May 3, 2013

… and predictive models can breach privacy too

The New York Times — Business Day — Technolo

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HE

Marketers Can Glean Private Data on Facebook

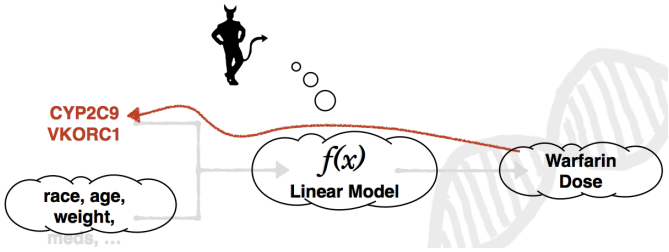


TARGET

TECH | 2/16/2012 @ 11:02AM | 837,678 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing



Need data analysis algorithms that can mine aggregated personal data with provable guarantees of privacy for individuals.

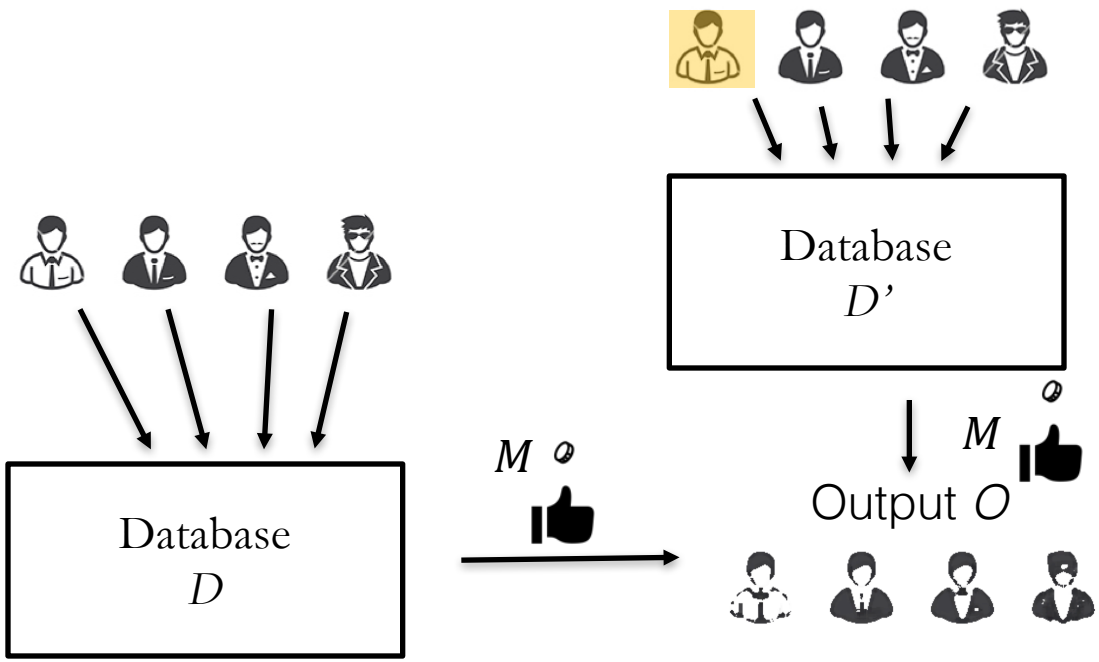
This is the goal of **Differential Privacy**.

Differential Privacy (DP)

Dwork et al. 2006

- Randomized mechanism M satisfies ϵ-DP if for all D, D' that differ in one row, and for every output O

$$\Pr[M(D) = O] \leq \exp(\epsilon) \Pr[M(D') = O]$$



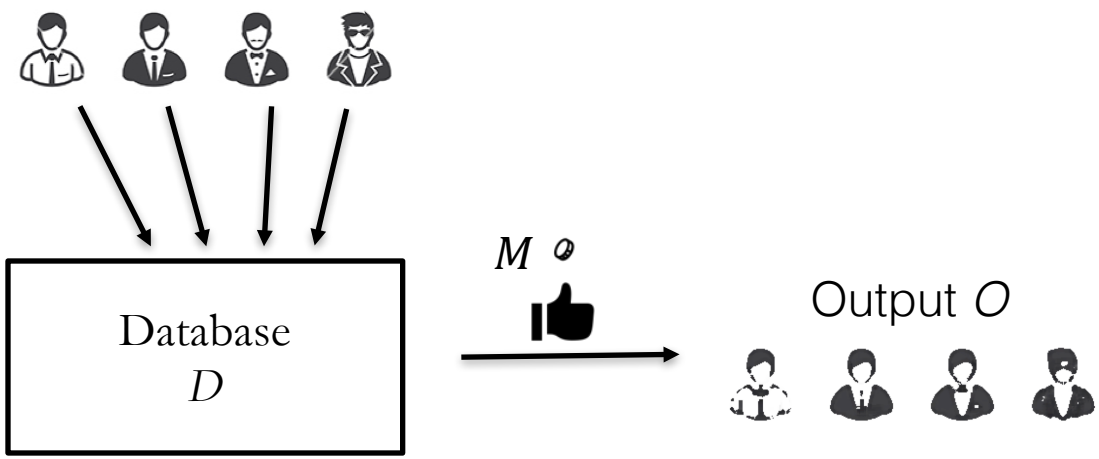
Differential Privacy (DP)

Dwork et al. 2006

- Randomized mechanism M satisfies ϵ-DP if for all D, D' that differ in one row, and for every output O

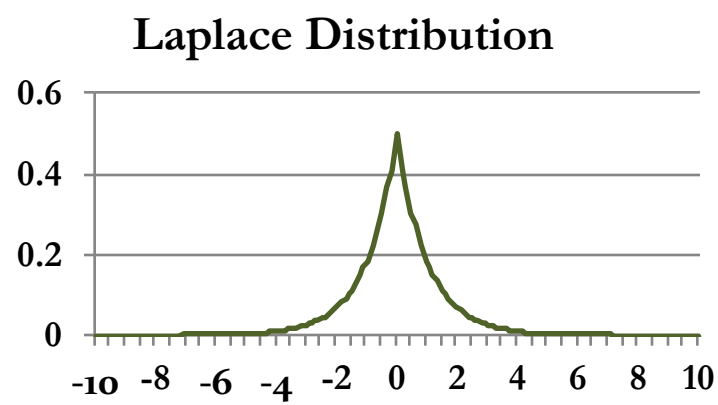
$$\Pr[M(D) = O] \leq \exp(\epsilon) \Pr[M(D') = O]$$

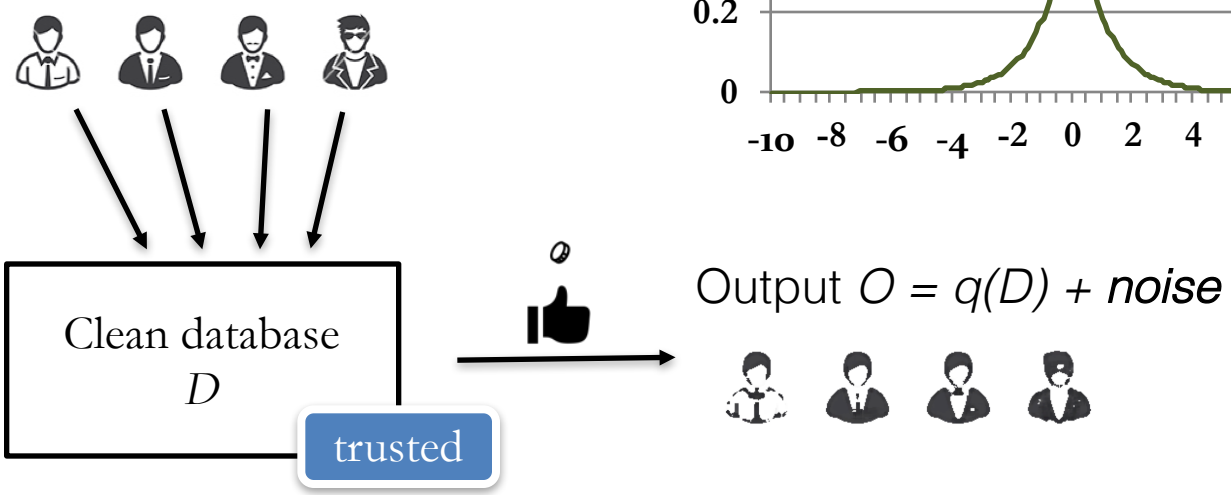
  - D, D' : Simulate the different values of a single record
  - Smaller the ϵ more the privacy (and better the utility)



Basic Algorithms

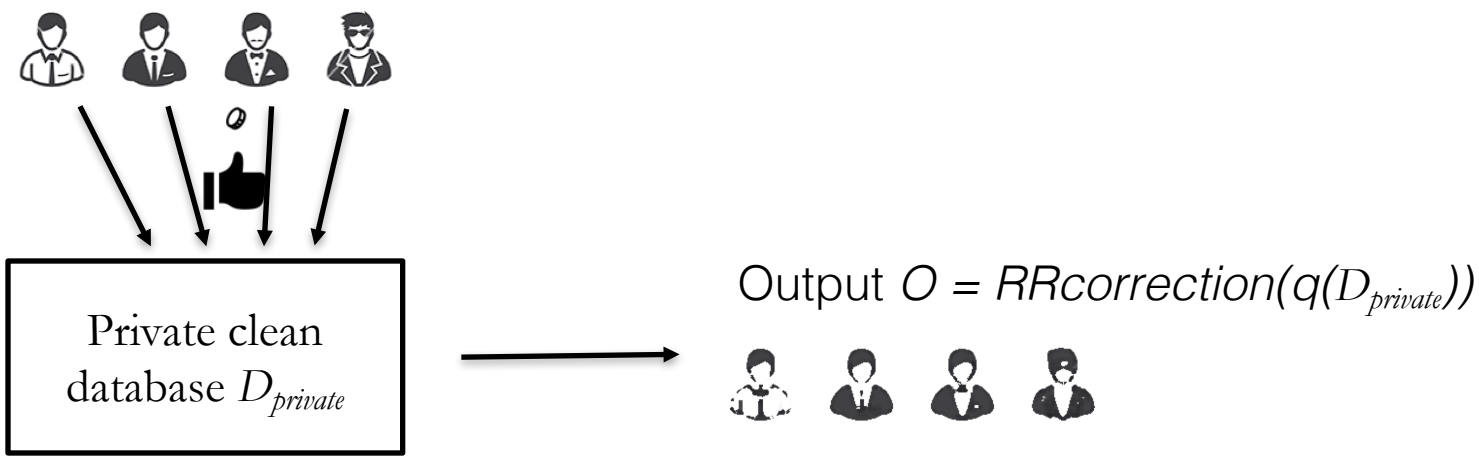
- Output Perturbation to query *q(D)*
  - Laplace Mechanism: add noise ~ $Lap(\frac{\Delta q}{\epsilon})$
  - Assume data collector is trusted





Basic Algorithms

- Input Perturbation
  - Randomized response (RR): each record is perturbed
    - By participant if data collector is not trusted
    - By data collector o.w.
  - RRCorrection is required for $q(D_{private})$



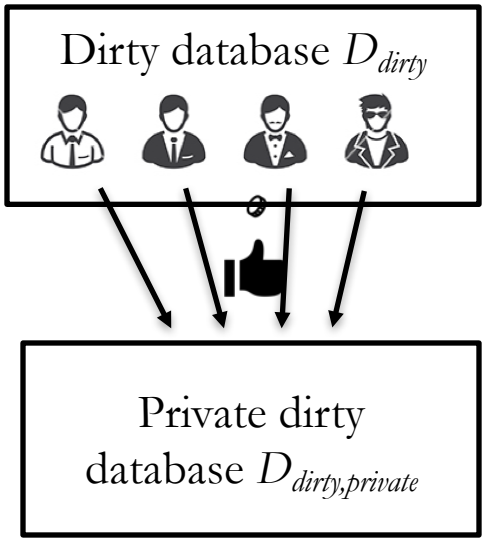
Outline

- A brief tour to differential privacy
- PrivacyClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing
    - Correction for randomized response
    - Cleaning for dirty data
  - Evaluation
- Discussion

Overview of PrivateClean

Sanjay et al. 2016

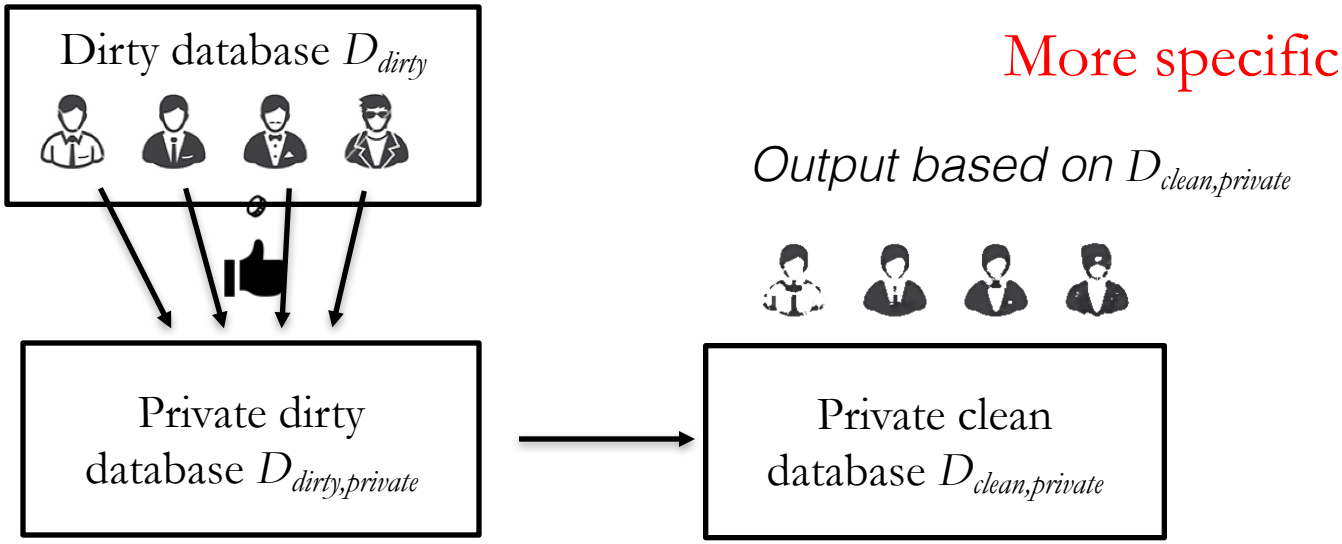
- Part I: Private relation creation:
  - Dirty database D_{dirty} (Categorical/ Numerical attributes)
  - Generate private dirty database $D_{dirty,private}$ with
    - RR for Categorical attributes (Cat)
    - Local version of Laplace Mechanism for Numerical attributes (Num)



Overview of PrivateClean

Sanjay et al. 2016

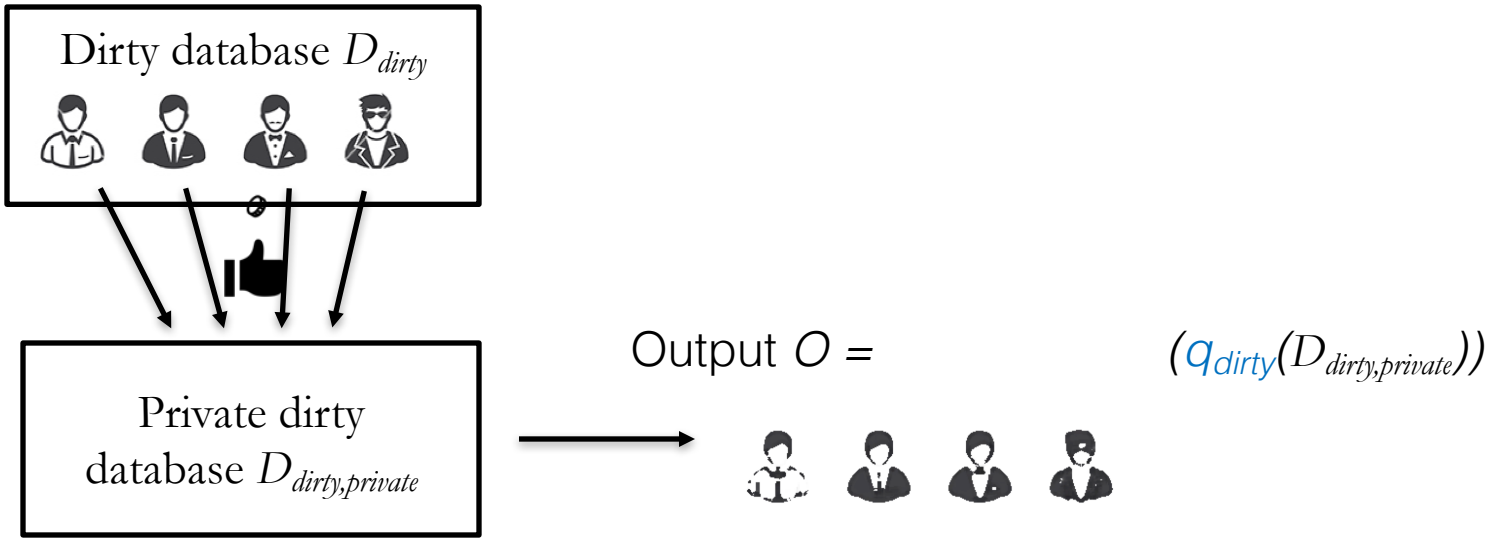
- Party II: Query processing:
  - q : SELECT **SUM/COUNT/AVG**(Num) Where **Cond**(Cat)
  - local cleaner with deterministic UDF over Cat attributes in each row (**Merge/Transform/Extract**)



Overview of PrivateClean

Sanjay et al. 2016

- Party II: Query processing:
  - q : SELECT **SUM/COUNT/AVG**(Num) Where **Cond**(Cat)
  - q_{dirty} : local cleaner with deterministic UDF over Num attributes in each row (**Merge/Transform/Extract**)
  - *RRcorrection*: correction for randomized response



Private Relation Creation

- Example dataset D_{dirty}
  - $q(D_{clean})$: SELECT **COUNT(*)** Where **Major = EECS**

StudentId	Major	Eval
1	IEOR	1.2
2	EECS	1.6
3	Electrical Engineering and Computer Sciences	3
4	Nuclear Engineering	2.1

Merge(…)

EECS

Categorical Attributes

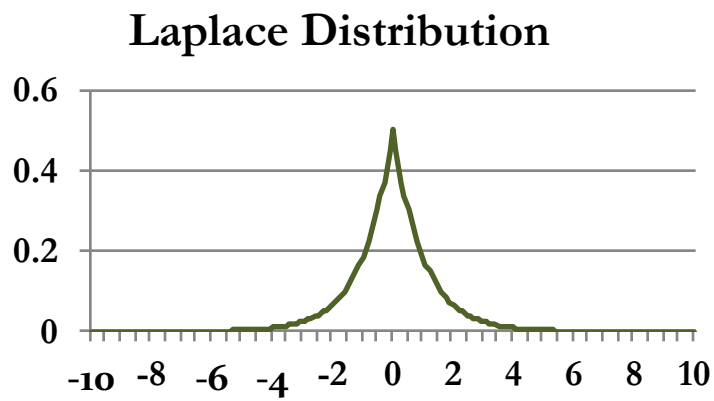
Numerical Attributes

RR for Categorical Attribute A_i

- *dom(A_i)*: distinct values for A_i in D_{dirty}
- For record *r*, let *r[i]* be the value of A_i, RR outputs
  - the true value: *r[i]* with prob. $1 - p_i$
  - a value in *dom(A_i)* uniformly, i.e. *U(dom(A_i))* with prob. p_i
- Given p_i, RR satisfies ϵ_i-DP, where $\epsilon_i = \ln(\frac{2}{p_i} - 1)$
  - Recall ϵ-DP: $\Pr[M(D) = O] \leq \exp(\epsilon) \Pr[M(D') = O]$
  - $\exp(-\epsilon_i) \leq \frac{\Pr[RR(a)=a]}{\Pr[RR(a')=a]} = \frac{1-p_i+p_i/|dom(A_i)|}{p_i/|dom(A_i)|} \leq \exp(\epsilon_i)$
  - Tightest $\epsilon_i = \ln\left(\frac{|dom(A_i)|}{p_i} - |dom(A_i)| + 1\right) \leq \ln(\frac{2}{p_i} - 1)$

Lap Noise for Numerical Attribute A_i

- For record r, let $r[i]$ be the value of A_i, Lap outputs $r[i]+\eta$, where $\eta \sim Lap(\frac{\Delta_i}{\epsilon_i})$,
  - Δ_i is the domain size of A_i in D_{dirty}
  - Mean: 0, Variance: $2\left(\frac{\Delta_i}{\epsilon_i}\right)^2$



- Satisfy ϵ_i-DP
- However, this add a lot of noises for queries such as SELECT **SUM()** Where **Cond**(Cat), unlike output perturbation method

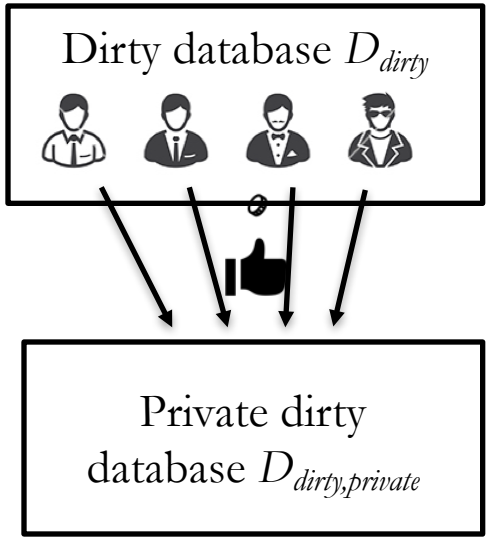
DP Dirty Database

- Generates $D_{dirty,private}$

StudentId	Major	Eval
1	IEOR	1.82321
2	Nuclear Engineering	2.12212
3	Electrical Engineering and Computer Sciences	3.2412
4	EECS	2.88721

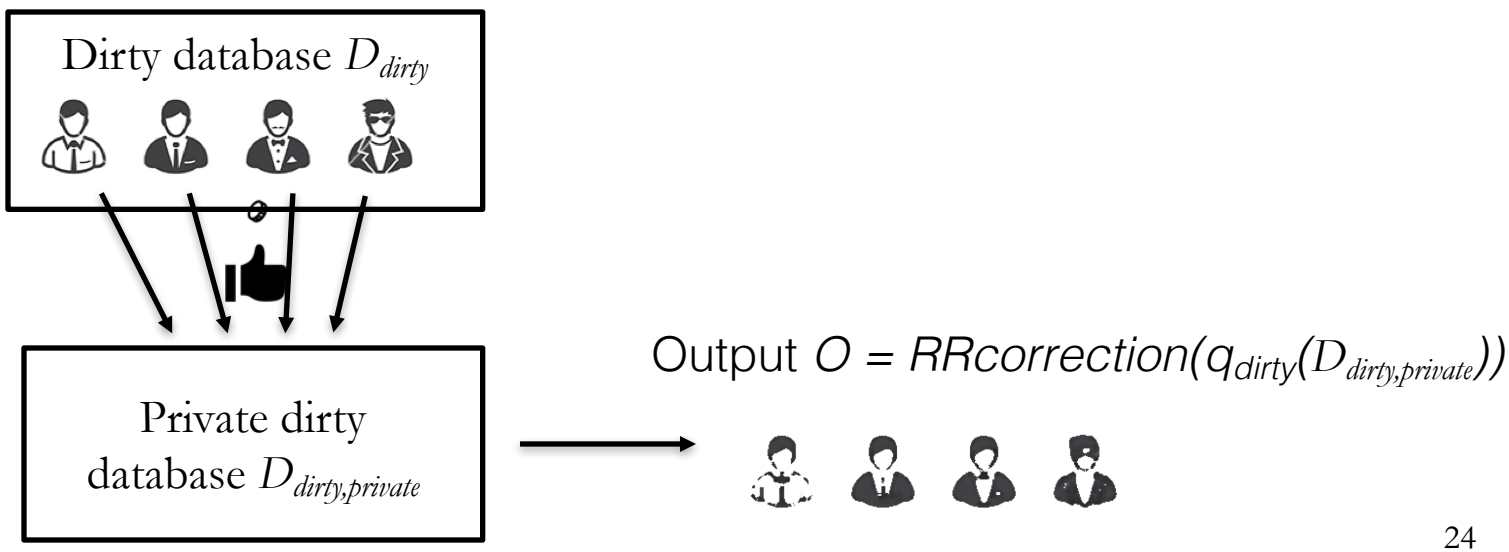
Overall Privacy Guarantee

- Composition: if each A_i is assigned with ϵ_i, the released $D_{dirty,private}$ satisfies $\sum_{A_i} \epsilon_i$ -DP.



Overall Privacy Guarantee

- Composition: if each A_i is assigned with ϵ_i-DP, the released $D_{dirty,private}$ satisfies $\sum_{A_i} \epsilon_i$ -DP.
- Post-processing: queries/cleaning over $D_{dirty,private}$ without looking at of D_{dirty} still satisfy $\sum_{A_i} \epsilon_i$ -DP.



Outline

- A brief tour to differential privacy
- PrivateClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing
    - Correction for randomized response
    - Cleaning for dirty data
  - Evaluation
- Discussion

Correction of RR

- First, we consider private clean database $D_{private}$
- q: SELECT **COUNT(*)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - c : true output for q, S: size of $D_{private}$
  - $E[q(D_{private})] = c\, p_i^+ + (S-c)\, p_i^-$ where
    - p_i^+ : Pr[including a record r that satisfies the predicate] = $(1 - pi) + pi * l/|dom(A_i)|$
    - p_i^- : Pr[The prob. of including a record r that does not satisfy the predicate]= $p_i * l/|dom(A_i)|$
  - estimate c by $(q(D_{private}) - S\, p_i^-)/(p_i^+ - p_i^-)$

Correction of RR

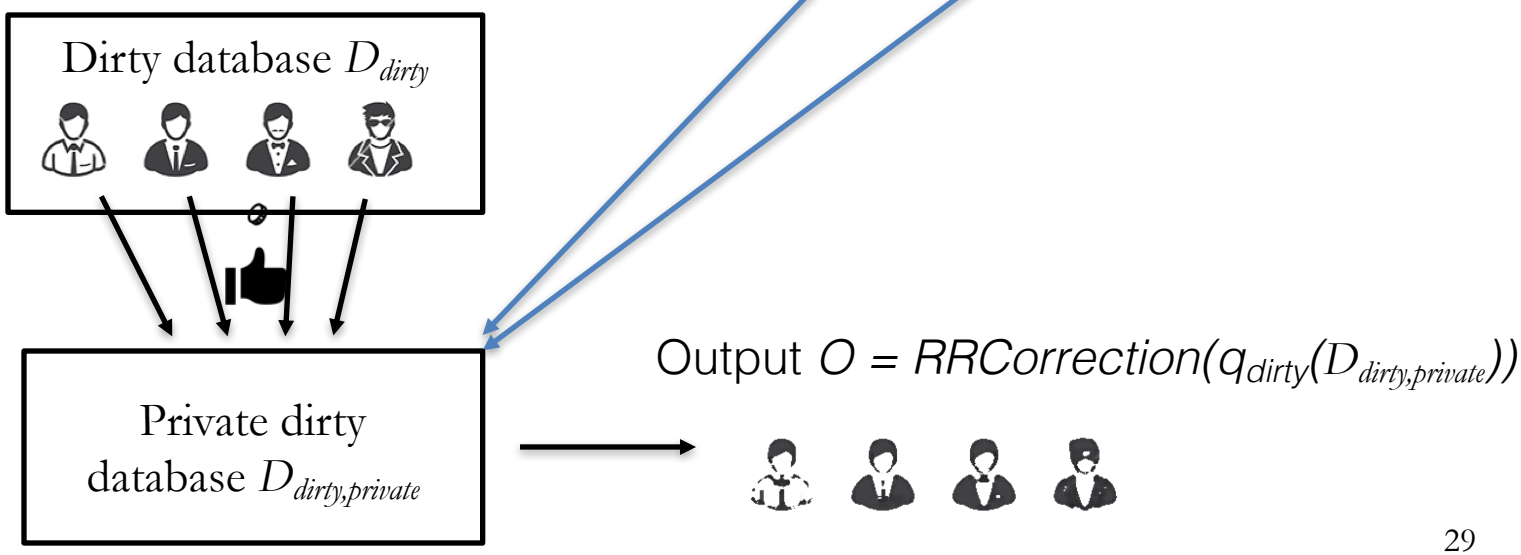
- First, we consider private clean database $D_{private}$
- q: SELECT **COUNT(*)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - estimate c by $(q(D_{private})-S\, p_i^-)/(p_i^+ - p_i^-)$
- q: SELECT **SUM(Num)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - estimate by $(1- p_i^-)q(D_{private})-p_i^- q^c(D_{private}))/(p_i^+ - p_i^-)$
    - where q^c: SELECT **SUM(Num)** Where $\boldsymbol{A_i \notin \{a_1,..,a_l\}}$
- Error: O(randomization * Sqrt(S))

Correction of RR

- First, we consider private clean database $D_{private}$
- q: SELECT **COUNT(*)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - estimate c by $(q(D_{private})-S\, p_i^-)/(p_i^+ - p_i^-)$
- q: SELECT **SUM(Num)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - estimate by $(1- p_i^-)q(D_{private})-p_i^- q^c(D_{private}))/(p_i^+ - p_i^-)$
    - where q^c: SELECT **SUM(Num)** Where $\boldsymbol{A_i \notin \{a_1,..,a_l\}}$
- q: SELECT AVG**(Num)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - estimate by SUM/COUNT

What Happens in Data Cleaning?

- q: SELECT **COUNT(*)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
  - p_i^+ : Pr[including a record r that satisfies the predicate]$(1 - p_i) + p_i * l/|dom(A_i)|$
  - p_i^- : Pr[The prob. of including a record r that does not satisfy the predicate]: $p_i * l/|dom(A_i)|$



What Happens in Data Cleaning?

q: SELECT **COUNT(*)** Where $\boldsymbol{A_i \in \{a_1,..,a_l\}}$
e.g. SELECT **COUNT(*)** Where **Major is EECS or IEOR**

$D_{dirty,private}$

Id	Major	Eval
1	IEOR	1.82321
2	Nuclear Engineering	2.12212
3	Electrical Engineering and Computer Sciences	3.2412
4	EECS	2.88721

***l* = 3**

$D_{clean,private}$

Id	Major	Eval
1	IEOR	1.82321
2	Nuclear Engineering	2.12212
3	EECS	3.2412
4	EECS	2.88721

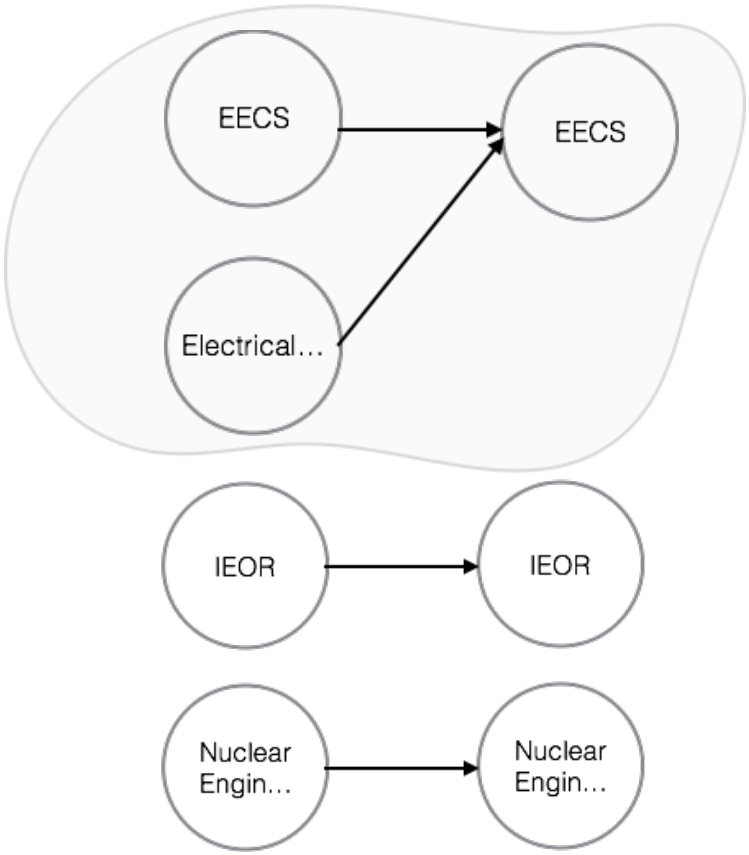
***l* = 2**

$p_i^+ = (1 - p_i) + p_i * l/|dom(A_i)|, p_i^- = p_i * l/|dom(A_i)|$
what is the right choice of l, $|dom(A_i)|$?

Value Provenance Graph per Attribute

Domain(A_i) in

D_{dirty} / $D_{private,dirty}$ D_{clean} / $D_{private,clean}$



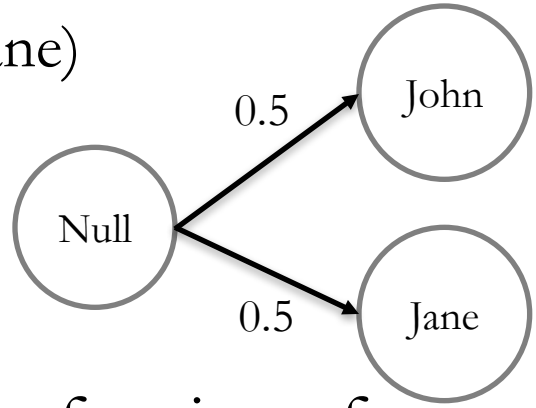
q: SELECT **COUNT(*)**
Where **Major is EECS or IEOR**
From $\boldsymbol{D_{private,clean}}$

q_{dirty}: SELECT **COUNT(*)**
Where **Major is EECS, Electrical .. or IEOR**
From $\boldsymbol{D_{dirty,private}}$

→ $RRcorrection(q_{dirty}(D_{dirty,private}))$

Weighted Provenance Graph for Multiple-Attribute Cleaning

- For example: <section, instructor>
  - (1,Null) → (1, John), (2,Null) → (2, Jane)
  - Graph for attribute instructor



- Edge with weight that represents the fraction of rows with the value in dirty database a mapped to the value in the clean database a'
  - Use this weight as the probability to include records in $D_{dirty,private}$ that satisfy the predicate in q_{dirty}

Outline

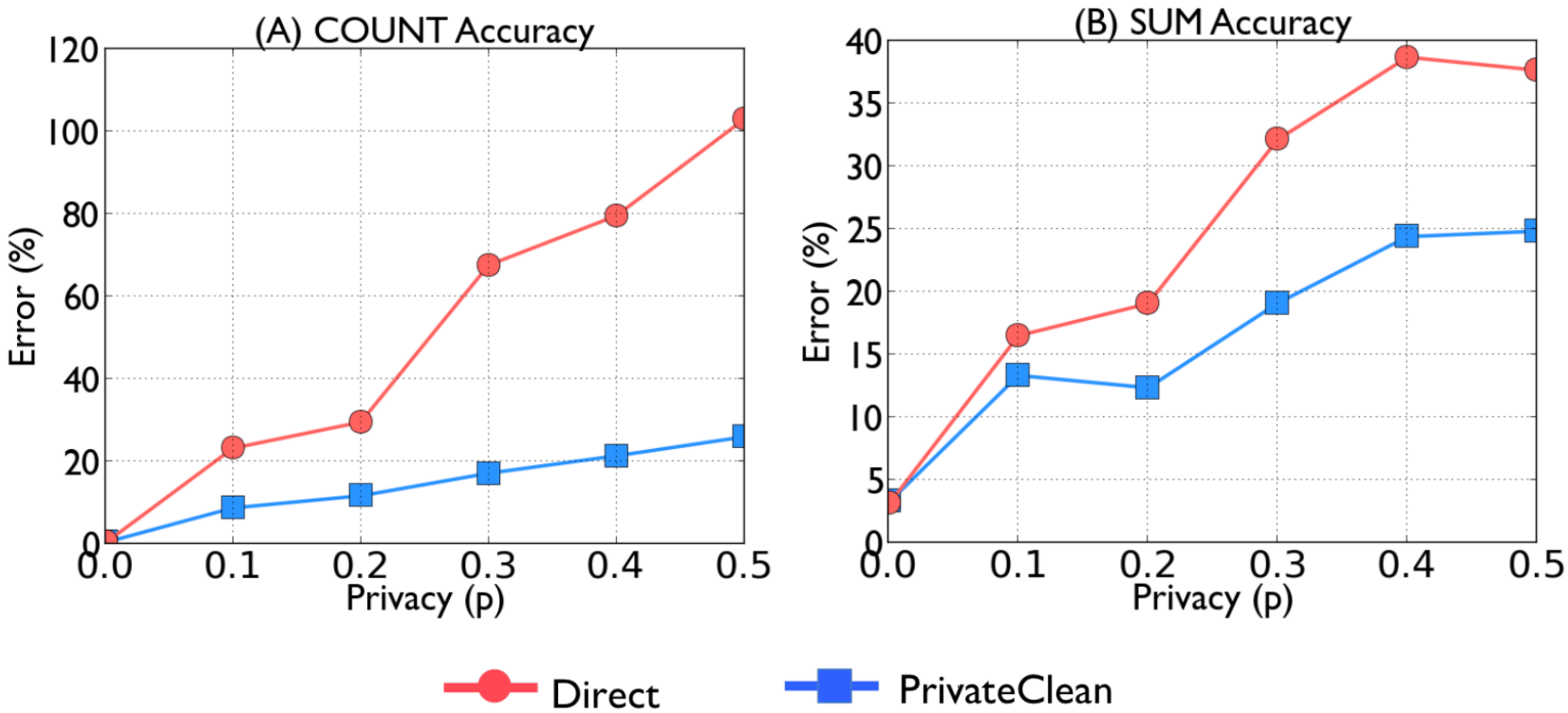
- A brief tour to differential privacy
- PrivateClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing
    - Correction for randomized response
    - Cleaning for dirty data
  - Evaluation
- Discussion

Evaluation of PrivateClean

- Baseline: **Direct** f($D_{clean,private}$) without CorrectionRR.
- Datasets:
  - Synthetic:
    - a single Num attribute [0,100], a single Cat attribute {1,…,N} drawn from a Zipfian distribution
  - TPC-DS:
    - Customer_address (ca_city, ca_county, ca_state, ca_country)
    - [ca_city, ca_county] → [ca_state], MD([ca_country]~[ca_country])
  - IntelWireless: 2.3 million
    - 68 sensor id (private), sensor measurements
  - MCAFE: 406 records
    - Cat: contry code (private), Num: "ethusaism" with value 1-10

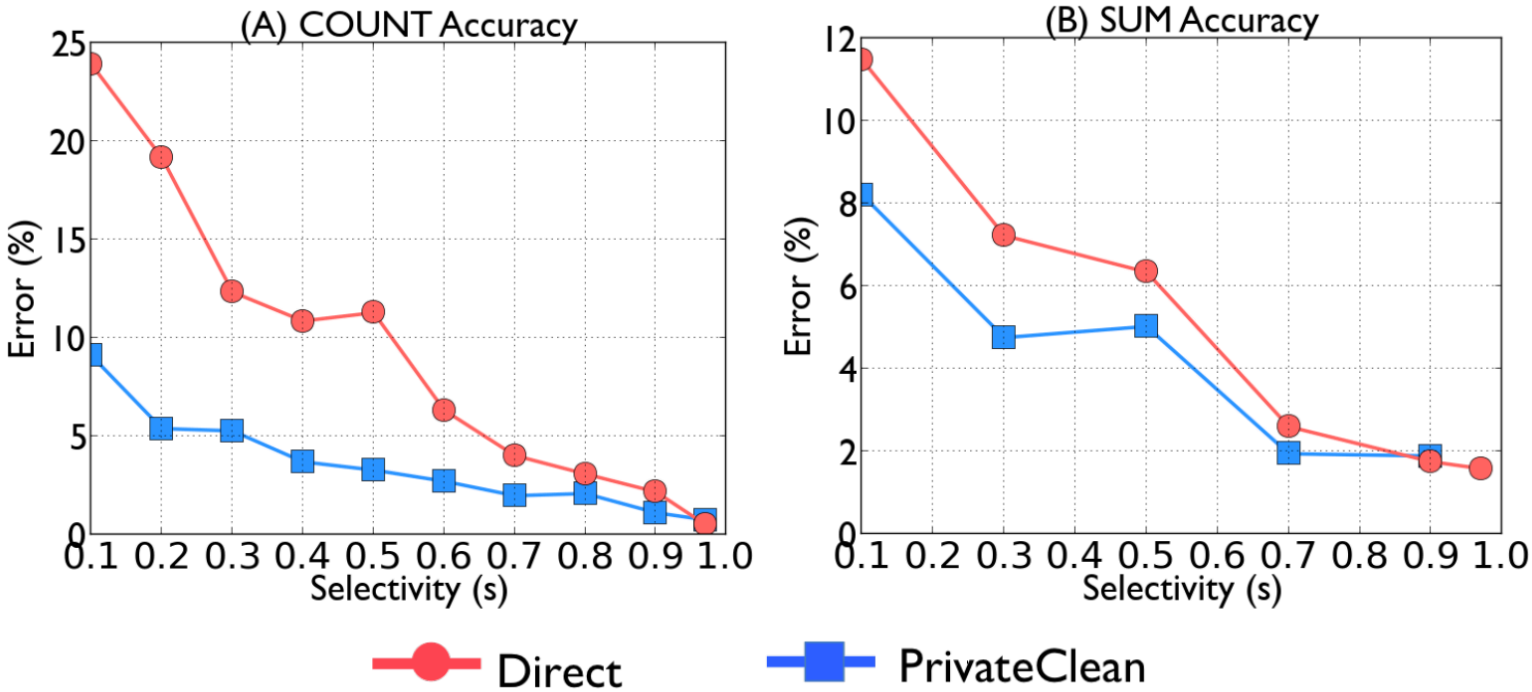
Varying privacy parameters p

- $p_i^+ = (1 - p_i) + p_i * l/|dom(A_i)|$, $p_i^- = p_i * l/|dom(A_i)|$

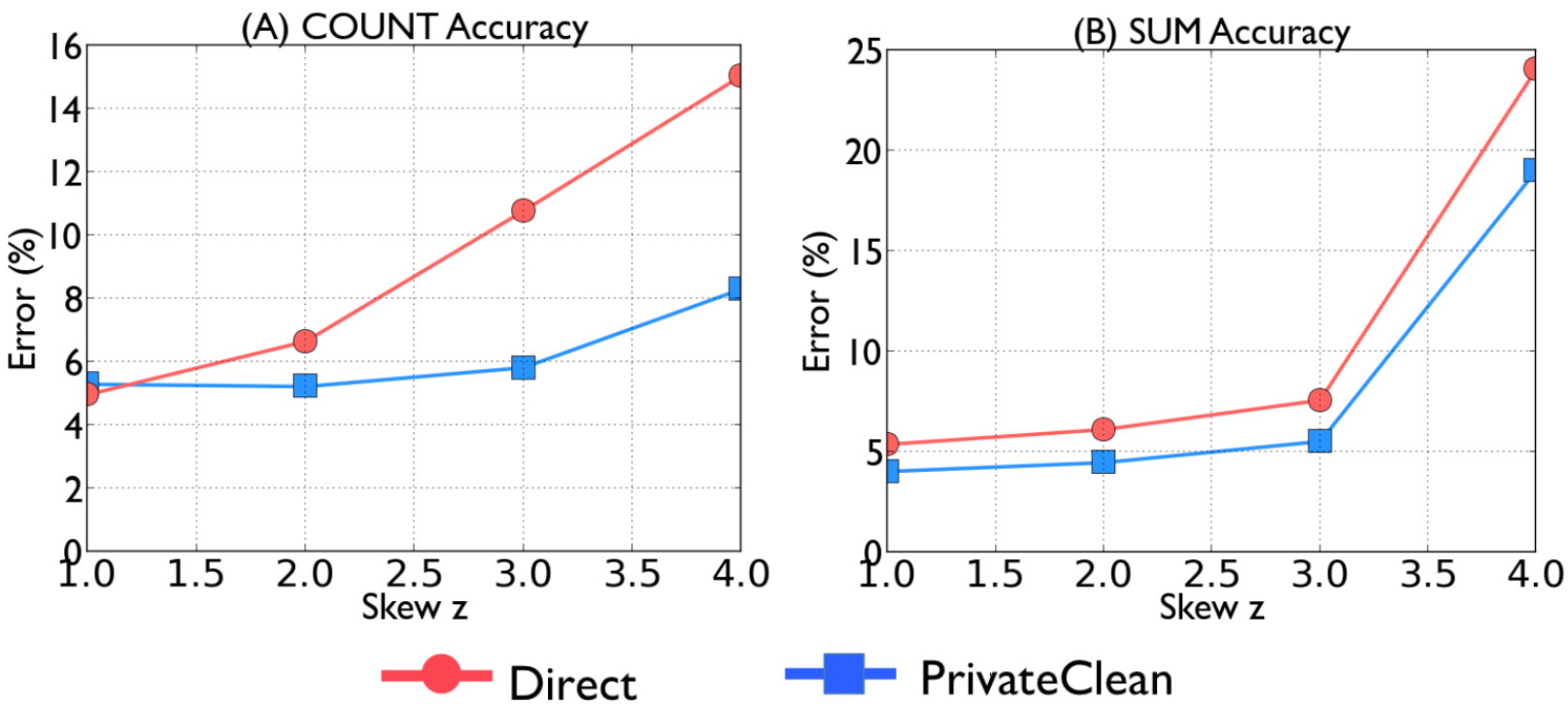


Varying selectivity $l/|dom(A_i)|$ of query

- $p_i^+ = (1 - pi) + pi * l/|dom(A_i)|$, $p_i^- = pi * l/|dom(A_i)|$



Varying skewness of data



Outline

- A brief tour to differential privacy
- PrivateClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing
    - Correction for randomized response
    - Cleaning for dirty data
  - Evaluation
- Discussion

Limitations

- Poor utility for perturbation per row
  - Error has a lower bound Sqrt(|D|) for both RR and Lap
  - Requires # of distinct values << |D| at small privacy budget
- Poor privacy guarantee for high dimension data
  - $\sum_{A_i} \epsilon_i$ -DP.

- Limited model:
  - Consider queries with only 1 single Cat attribute in predicate
  - Consider local cleaner that cleans a single row
  - Transformations for data cleaning are given
  - Mainly a post-processing step for DP database

In real world

- Optimal strategies to clean the data are unknown

- Data owner outsources a 3rd party (who is expert for data cleaning) to clean the data, but requires
  - the 3rd party learns little about individual records in the data (provable privacy), but
  - with the optimal data cleaning strategy (utility)
  - with a small number of interactions (efficiency)
    → 3-way trade-offs

Summary

- Differential privacy: mine aggregated personal data with provable guarantees of privacy for individuals.

- PrivateClean: data cleaning + differential privacy
  - Private relation generation
  - Query processing

- Discussion:
  - A good start for private data cleaning
  - Limited model, lack of optimality in privacy and utility