# Utility Cost of Provable Privacy
## A case study on US Census Bureau Data

Ashwin Machanavajjhala

*ashwin@cs.duke.edu*

# Our world is increasingly data driven

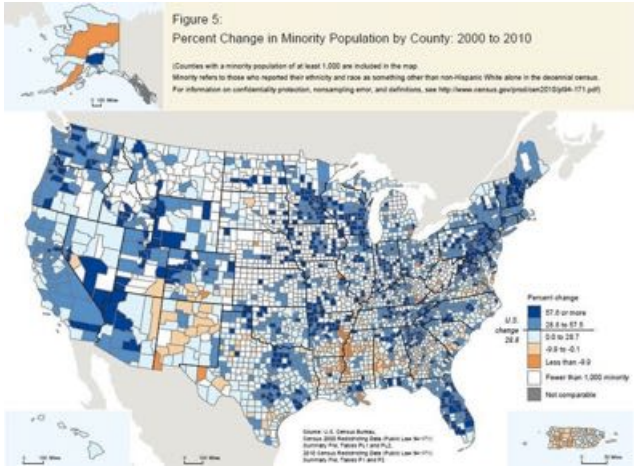

Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)

# Aggregated Personal Data …

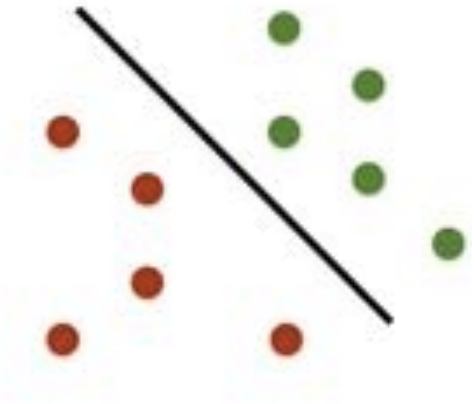
… is made publicly available in many forms.

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

Statistics
(e.g., demographic)




Predictive models
(e.g., advertising)



# … but privacy breaches abound

A Face Is Exposed for AOL Searcher No. 4417749

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

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

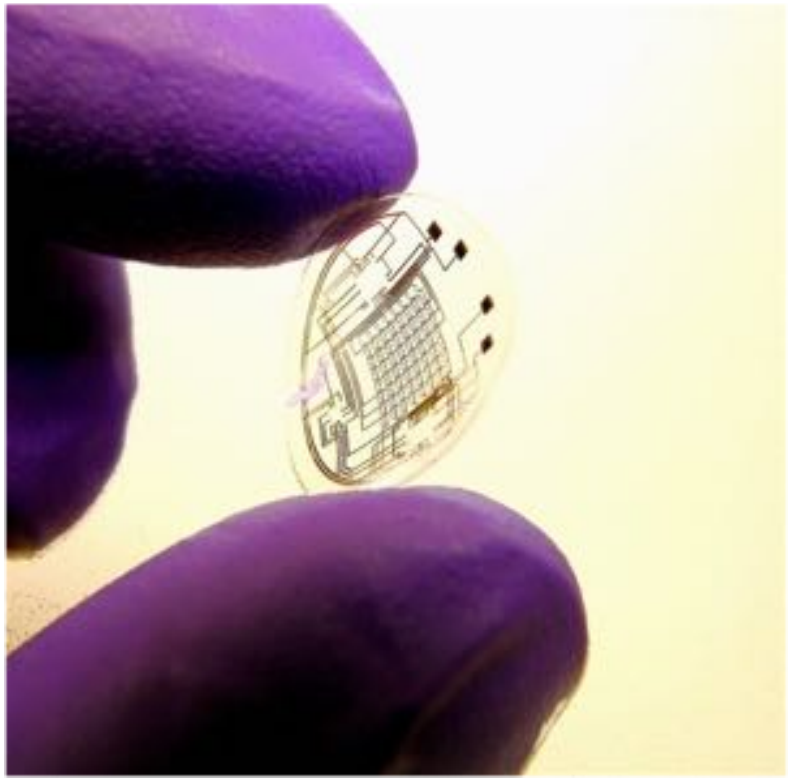
# >50 year old problem … what's new?

… and why are computer scientists excited about this problem?

- Internet
- Rampant data collection
- Sophistication in off-the-shelf analysis techniques
- *and* ...

# >50 year old problem … what's new?

*Privacy through the lens of Computation**



**Acknowledgements: Christos Papadimitriou.*

# >50 year old problem … what's new?

*Privacy through the lens of Computation*

- Privacy is a property of the computation generating the aggregated data and not of the output dataset itself

# >50 year old problem … what's new?

*Privacy through the lens of Computation*

- Ability to *formulate and prove* whether an *adversarially* chosen sequence of computations on sensitive data satisfies privacy.

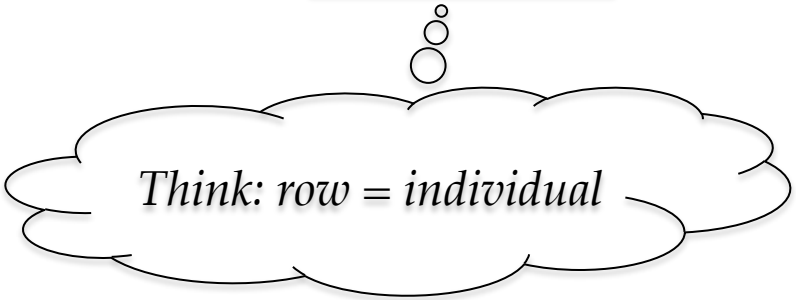
- Ability to *build algorithms* with provable guarantees of privacy.

# Provable guarantees of privacy

## A Firm Foundation for Private Data Analysis

By Cynthia Dwork
Communications of the ACM, Vol. 54 No. 1, Pages 86-95
10.1145/1866739.1866758

***Differential Privacy***: *"The output of an algorithm should not be significantly affected by the addition or deletion of* ***one row*** *in the dataset"*



# Provable Privacy Challenge

- Can traditional algorithms for data release and analysis be **replaced with provably private** algorithms while ensuring **little loss in utility**?

# This talk

- *Yes we can* … a case study on US Census Bureau Data

  - Current algorithm for data release with *no provable guarantees* and parameters used have to be kept *secret*
  - Our new algorithms provably satisfy a strong privacy guarantee (like differential privacy)
  - Can release tabular summaries with *comparable or better utility* than current techniques!

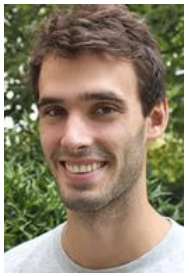
# A case study on US Census Bureau Data

Current interpretation of the law ≈ Mathematical Privacy Requirements on Release Mechanism ← Privacy Definition ← Release Mechanism

# A case study on US Census Bureau Data

Current interpretation of the law ≈ Mathematical Privacy Requirements on Release Mechanism ← Privacy Definition ← Release Mechanism

**Appears in SIGMOD 2017**


Sam Haney


John Abowd


Matthew Graham


Mark Kutzbach

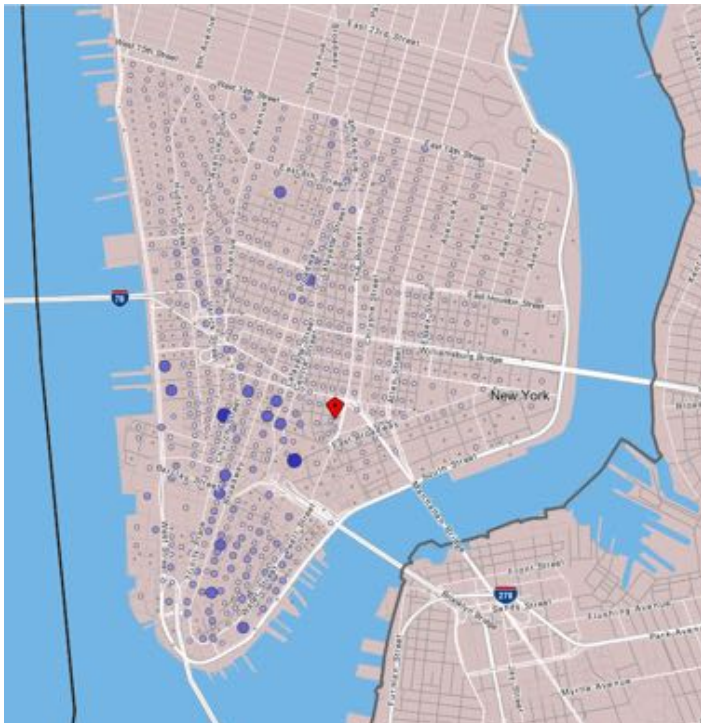

Lars Vilhuber

Duke UNIVERSITY

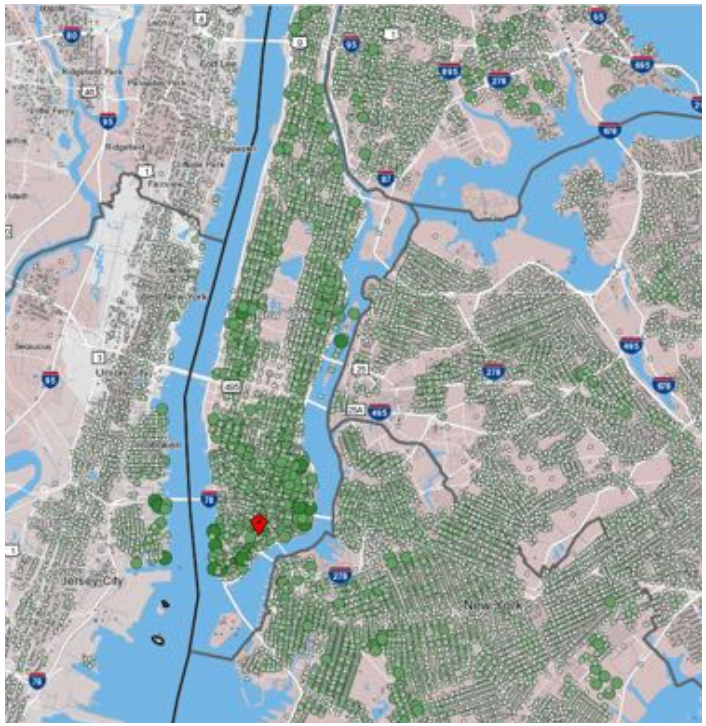
United States Census Bureau

Cornell University

# US Census Bureau's OnTheMap



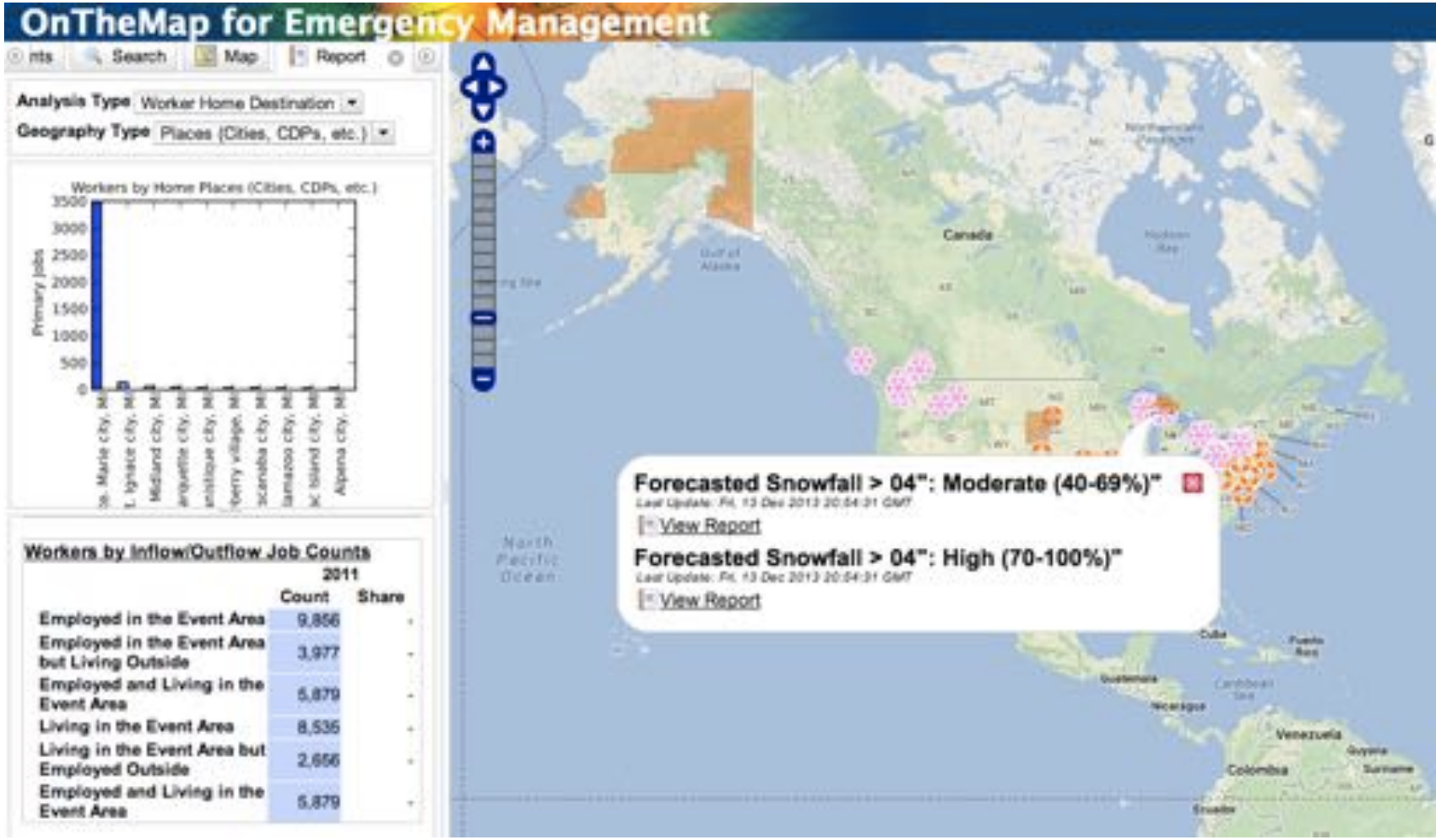

Employment in Lower Manhattan




Residences of Workers Employed in Lower Manhattan

Available at http://onthemap.ces.census.gov/.

# OnTheMap


OnTheMap for Emergency Management
Search
Map
Report
Analysis Type
Worker Home Destination
Geography Type
Places (Cities, CDPs, etc.)
Workers by Home Places (Cities, CDPs, etc.)
Primary Jobs
3500
3000
2500
2000
1500
1000
500
0
Workers by Inflow/Outflow Job Counts
2011
Count
Share
Employed in the Event Area
9,856
Employed in the Event Area but Living Outside
3,977
Employed and Living in the Event Area
5,879
Living in the Event Area
8,535
Living in the Event Area but Employed Outside
2,656
Employed and Living in the Event Area
5,879
Canada
Forecasted Snowfall > 04": Moderate (40-69%)"
View Report
Forecasted Snowfall > 04": High (70-100%)"
View Report
Venezuela
Colombia

# Underlying data: LODES

- Employee
  - Age
  - Sex
  - Race & Ethnicity
  - Education
  - Home location (Census block)

- Job
  - Start date
  - End date
  - Worker & Workplace IDs
  - Earnings

- Employer
  - Geography (Census blocks)
  - Industry
  - Ownership (Public vs Private)

# Underlying data: LODES

- Employee
  - Age
  - Sex
  - Race & Ethnicity
  - Education
  - Home location (Census block)

Released using an algorithm that satisfies a variant of differential privacy [Machanavajjhala et al 2008]

First real world application of differential privacy.

# Goal: Release Tabular Summaries

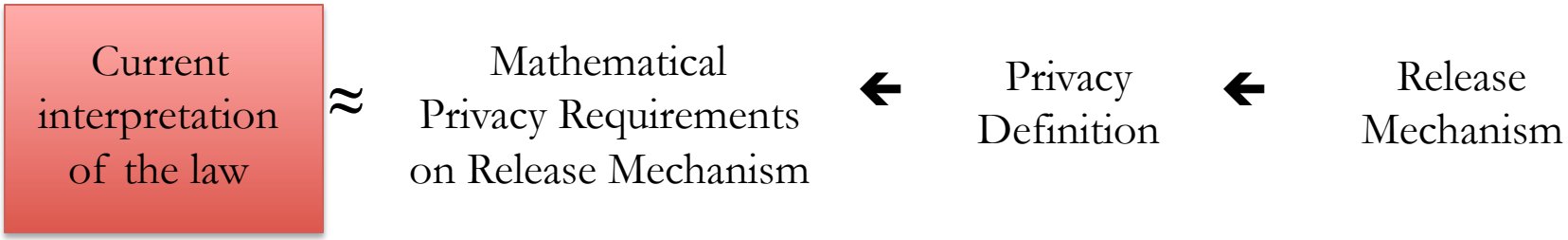
Counting Queries

- Count of jobs in NYC
- Count of jobs held by workers age 30 who work in Boston.

Marginal Queries

- Count of jobs held by workers age 30 by work location (aggregated to county)

# A case study on US Census Bureau Data



# Release of data about employers and employees is regulated by …

- Title 13 Section 9

*Neither the secretary nor any officer or employee …*

*… make any publication whereby the data furnished by any particular establishment or individual under this title can be identified …*

# Disclosure Review Board

- Approves the release of *data* that, in its view, satisfy these statutory confidentiality protection requirements.

# Current Interpretation

- The existence of a job held by a particular individual *must* not be disclosed.
- The existence of an employer business as well as its type (or sector) and location is not confidential.
- The data on the operations of a particular business must be protected.
  - Total employment
  - Number of males/females, etc.

# Current Interpretation

- The existence of a job held by a particular individual *must* not be disclosed.

  **No exact re-identification of employee records … by an informed attacker.**

- The existence of an employer business as well as its type (or sector) and location is not confidential.

  **Can release exact numbers of employers**

- The data on the operations of a particular business must be protected.

  **Informed attackers must have an uncertainty of up to a multiplicative factor $(1+\alpha)$ about the workforce of an employer**

# Current Protection: Input Noise Infusion

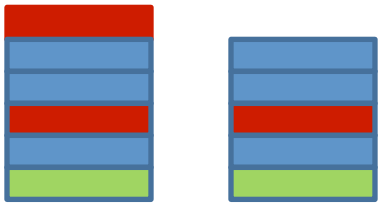
- Each employer assigned a *permanent* multiplicative noise distortion factor
- True employer workforce counts multiplied by noise factor before aggregation
- Additional techniques to hide small counts in tabulations.

- No provable guarantees against inference
  - Method to generate noise factors public, but the parameters used are a *secret*.

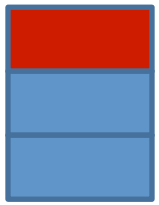
# Can we use differential privacy (DP)?

For every pair of
**Neighboring Tables**

For every output …



$D_1$ $D_2$



$O$

Should not be able to distinguish whether O was generated by $D_1$ or $D_2$

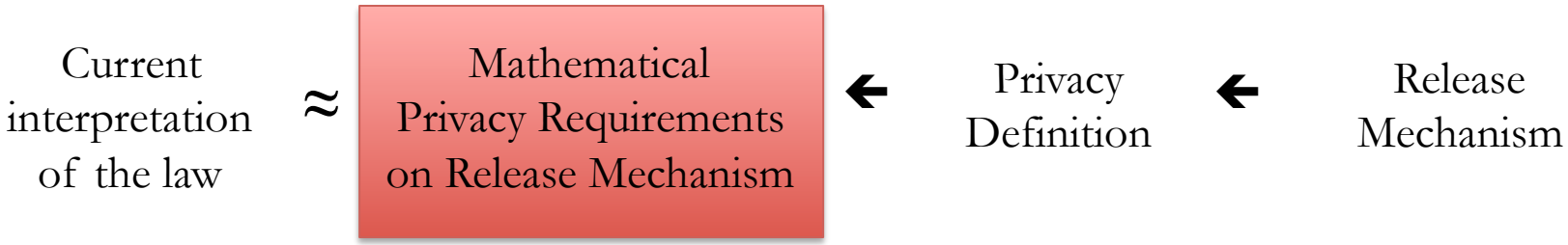
$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \qquad (\varepsilon>0)$$

# Neighboring tables for LODES?

- Tables that differ in …
  - one employee?
  - one employer?
  - something else?

- And how does DP (and its variants) compare to the current interpretation of the law?
  - Who is the attacker? Is he/she informed?
  - What is secret and what is not?

# A case study on US Census Bureau Data



# The Pufferfish Framework

[Kifer and Machanavajjhala PODS '12]

- What is being kept secret?
- Who are the adversaries?
- How is information disclosure bounded?
  - (similar to epsilon in differential privacy)

# Sensitive Information

- **Secrets**: S be a set of potentially sensitive statements
  - "individual j's record is in the data, and j has Cancer"
  - "individual j's record is not in the data"

- **Discriminative Pairs**: $S_{pairs} \subseteq S \times S$
  Mutually exclusive pairs of secrets.
  - ("Bob is in the table", "Bob is not in the table")
  - ("Bob has cancer", "Bob has diabetes")

# Adversaries

- We assume a Bayesian adversary who is can be completely characterized by his/her prior information about the data
  - We do not assume computational limits

- **Data Evolution Scenarios**: set of all probability distributions that could have generated the data ( … think adversary's prior).

  - *No assumptions*: All probability distributions over data instances are possible.

  - *I.I.D.*: Set of all *f* such that: $P(data = \{r_1, r_2, \ldots, r_k\}) = f(r_1) \times f(r_2) \times \ldots \times f(r_k)$

# Pufferfish Privacy Guarantee

$$\forall\, w \in range(M)$$
$$\forall\, (s, s') \in S_{pairs}$$
$$\forall \theta \in D, s.t. \quad P(s|D), P(s'|D) \neq 0$$

$$e^{-\varepsilon} \leq \frac{P(s|M(\mathfrak{D}) = w, \theta)}{P(s'|M(\mathfrak{D}) = w, \theta)} \Big/ \frac{P(s|\theta)}{P(s'|\theta)} \leq e^{\varepsilon}$$

**Posterior odds of s vs s'**

**Prior odds of s vs s'**

# No Free Lunch in privacy

It is not possible to guarantee *any* utility in addition to privacy, *without making assumptions about*

- *the data generating distribution* [Kifer-**Machanavajjhala** SIGMOD '11]
- *the background knowledge available to an adversary* [Dwork-Naor JPC '10]

# DP vs Pufferfish

**Discriminative Pairs**:

- (x in table with value r, x not in table)

- Attackers can't tell whether a record is in the data with value x, or is not in the data.

# DP vs Pufferfish

**Discriminative Pairs**:

- (x in table with value r, x not in table)

**Data Evolution Scenarios**:

Set of all *priors* such that records are **independent**

$$P(data = \{r_1, r_2, \ldots, r_k\}) = f_1(r_1) \times f_2(r_2) \times \ldots \times f_k(r_k)$$

# DP vs Pufferfish

[Kifer-Machanavajjhala PODS'12]

**Discriminative Pairs**:

- (x in table with value r, x not in table)

**Data Evolution Scenarios**:

$$P(data = \{r_1, r_2, \ldots, r_k\}) = f_1(r_1) \propto f_2(r_2) \propto \ldots \propto f_k(r_k)$$

A mechanism ensures the Pufferfish privacy guarantee with above setting of secrets and adversaries *if and only if* it satisfies DP.

# Back to LODES

- **Discriminative Secrets**:
  - (*w* works at *E*, or *w* works at *E'*)
  - (*w* works at *E*, *w* does not work)
  - ($|E| = x$, $|E| = y$), for all $x < y < (1+\alpha)x$
  - …

- **Data evolution scenarios**:
  - All priors where employee records are independent of each other.

# Example of a formal privacy requirement

DEFINITION 4.2 (EMPLOYER SIZE REQUIREMENT). *Let $e$ be any establishment in $\mathcal{E}$. A randomized algorithm $A$ protects establishment size against an informed attacker at privacy level $(\epsilon, \alpha)$ if, for every informed attacker $\theta \in \Theta$, for every pair of numbers $x, y$, and for every output of the algorithm $\omega \in range(A)$,*

$$\left| \log \left( \frac{Pr_{\theta,A}[|e| = x | A(D) = \omega]}{Pr_{\theta,A}[|e| = y | A(D) = \omega]} \Big/ \frac{Pr_{\theta}[|e| = x]}{Pr_{\theta}[|e| = y]} \right) \right| \leq \epsilon \quad (4)$$

*whenever $x \leq y \leq \lceil (1+\alpha)x \rceil$ and $Pr_\theta[w = x], Pr_\theta[w = y] > 0$.*

# A case study on US Census Bureau Data

# Satisfying Privacy Requirements

| | Privacy for Employees | Privacy for Employers |
|---|---|---|
| Input Noise Infusion | No | No |
| DP (employees) | Yes | No |
| DP (employers) | Yes | Yes |

But algorithms that satisfy DP (employers) result in a high loss in utility.

# Employer-Employee Privacy

- A new privacy definition that is *customized to* the privacy requirements, and yet permits useful algorithms.

- Neighboring tables:
  - Differ in the set of employees working at a single establishment
  - $E \subseteq E'$, *and* $|E| \leq |E'| \leq \max((1+\alpha)|E|, |E|+1)$

# Employer-Employee Privacy

DEFINITION 7.2 ($(\alpha, \epsilon)$-ER-EE PRIVACY). *A randomized algorithm $\mathcal{M}$ is said to satisfy $(\alpha, \epsilon)$-ER-EE Privacy, if for every set of outputs $S \subseteq range(M)$, and every pair of strong $\alpha$-Neighbors $D$ and $D'$, we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S]$$

# Employer-Employee Privacy

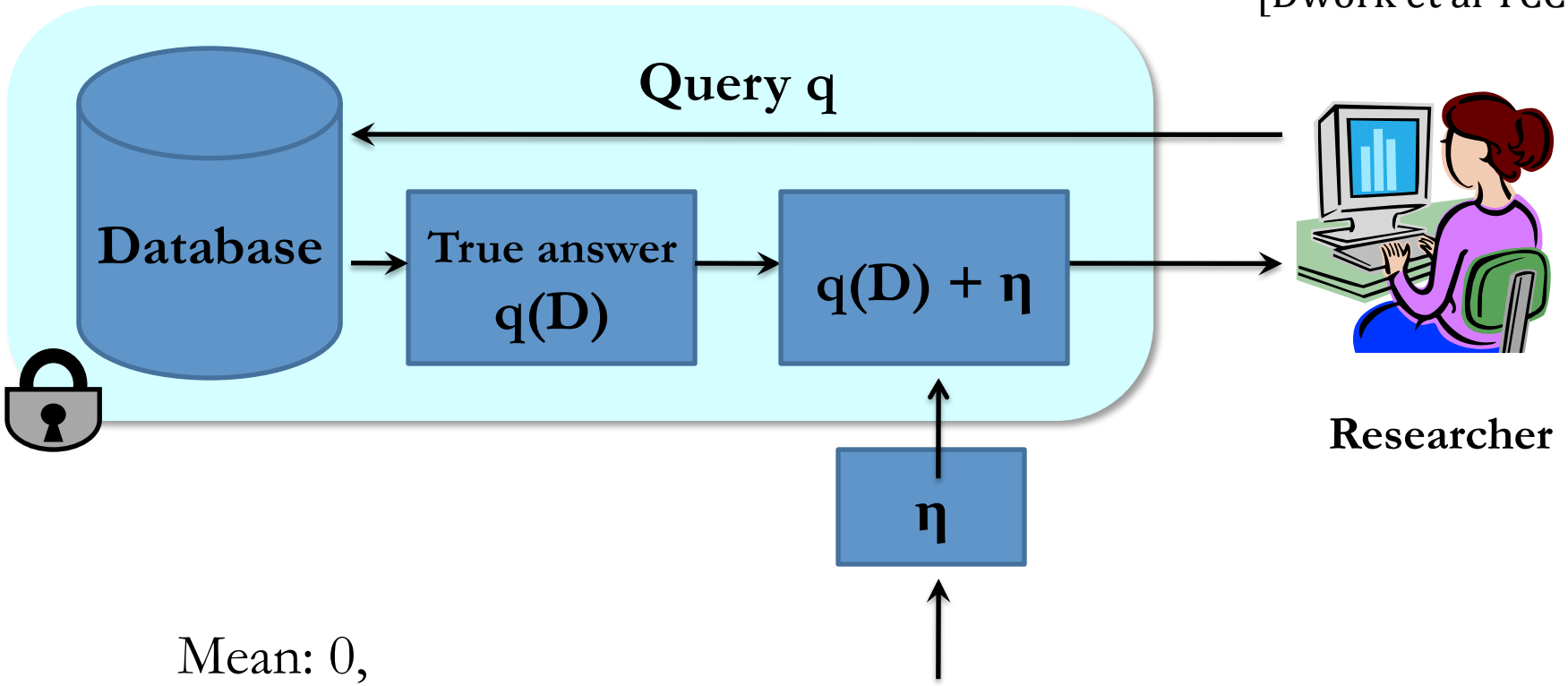
- Provides a differential privacy type privacy guarantee for all employees
  - Algorithm output is insensitive to addition or removal of one employee

- Appropriate privacy for establishments
  - Can learn whether an establishment is large or small, but not exact workforce counts.

# A case study on US Census Bureau Data

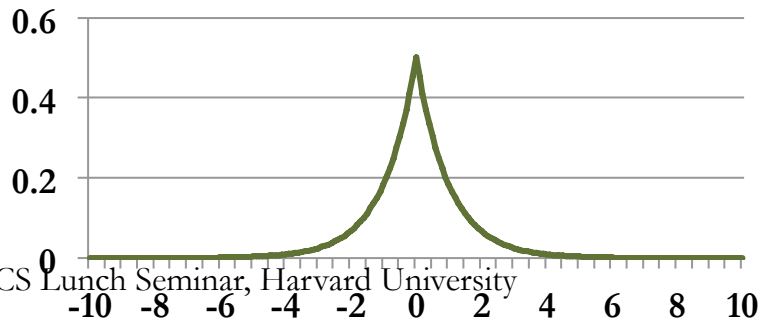
# Laplace Mechanism

[Dwork et al TCC 06]




Mean: 0,
Variance: $2\lambda^2$

**Laplace Distribution – Lap(λ)**



# Global Sensitivity

- S(q): Maximum change in the query answer over any pair of neighboring databases.

- Theorem: Adding Laplace noise with $\lambda = S(q) / \varepsilon$ satisfies $\varepsilon$-DP.

# Sensitivity for LODES queries

- q: Count of jobs in NYC

- If M is the *maximum possible* workforce size for an establishment,

$$S(q) = M(1+\alpha)$$

*Laplace mechanism will result in too much noise*

# Sensitivity for LODES queries

- q: log ( Count of jobs in NYC )

$$S(q) = \log (1+\alpha)$$

*Log-Laplace Algorithm*

# Smooth sensitivity [Nissim et al STOC 07]

- q: Count of jobs in NYC

- Local Sensitivity LS(q, D):
  Maximum change in query answer among neighbors of D.

$$LS(q,D) = \alpha\, q(D)$$

# Smooth sensitivity

- q: Count of jobs in NYC

$$LS(q,D) = \alpha\, q(D)$$

- LS(q, D) in our case is *smooth*, so we can add noise proportional to LS(q, D)

*Smooth-Gamma and Smooth-Laplace* Algorithms*

* Smooth Laplace has a small probability with which privacy is violated

# Composition

**Sequential Composition**:

Answering k queries under Employer-Employee privacy each with parameter $\varepsilon$ ensures privacy with parameter k $\varepsilon$ .

**Parallel Composition**:

Answering queries on disjoint sets of employers does not degrade the privacy parameter.

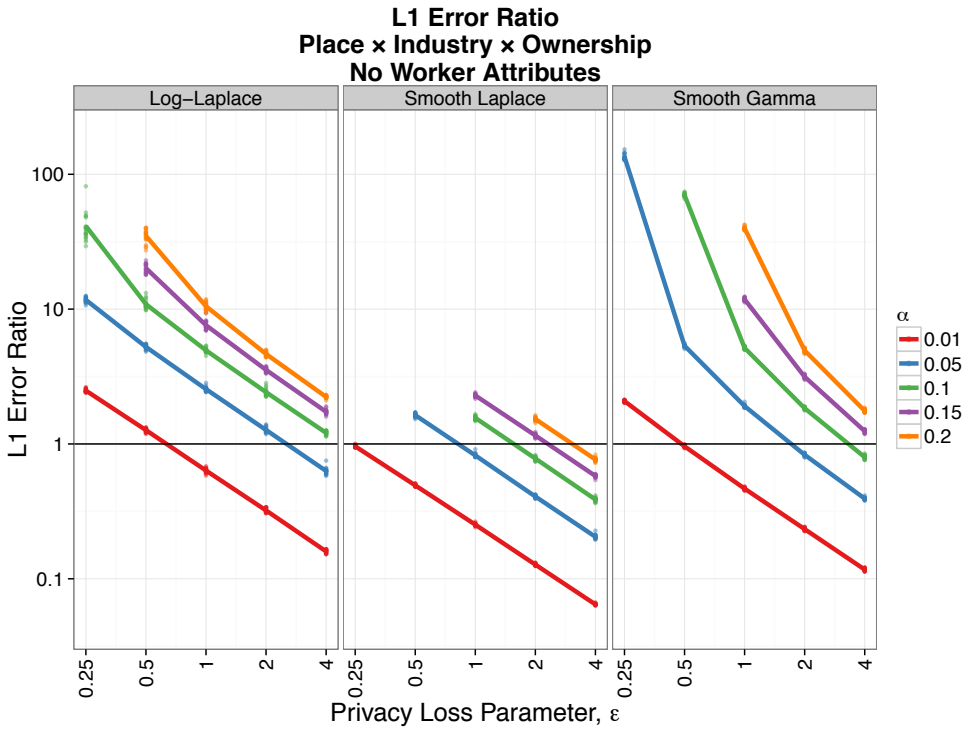
# This talk

- A case study on US Census Bureau Data
  - Current algorithm for data release with *no provable guarantees* and parameters used have to be kept *secret*
  - Our new algorithms provably satisfy a strong privacy guarantee (like differential privacy)
  - Can release tabular summaries with *comparable or better utility* than current techniques!

# Empirical Results on Utility

- Sample constructed from 3 states in US
  - 10.9 million jobs and 527,000 establishments
- Q1: Marginal counts over all establishment characteristics
- Metric: L1 error (new)/L1 error (current)

# Marginal query

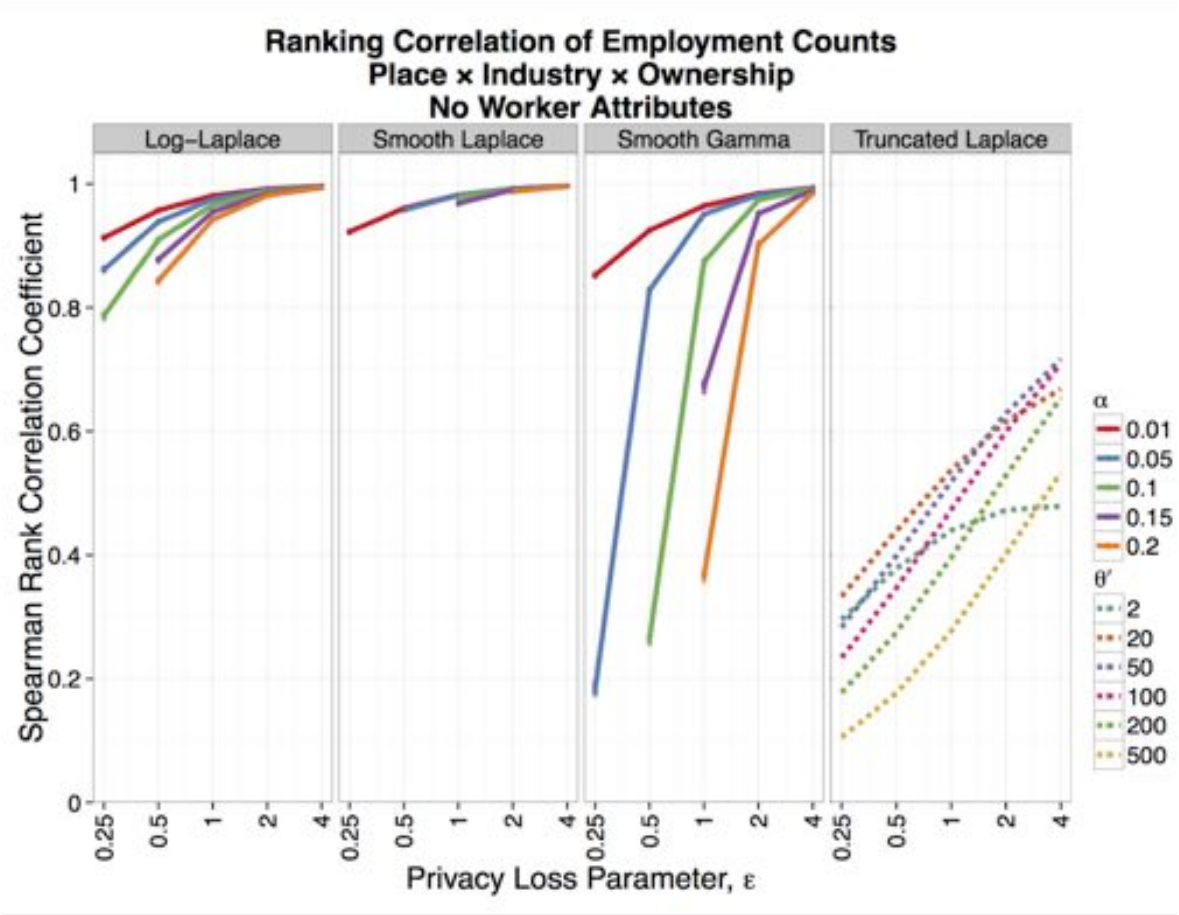



- Smooth-Laplace has the least utility cost
  - Due to small chance of violating privacy
- For $\varepsilon \geq 1$, and $\alpha \leq 5\%$ utility cost is at most a factor of 3.
- DP(employees) algorithm has uniformly high cost (>10) for all epsilon values.

# Empirical Results on Utility

- Sample constructed from 3 states in US
  - 10.9 million jobs and 527,000 establishments
- Q2: Rank cells in the marginal table by total count
- Metric: Spearman rank-ordered correlation

# Ranking Query



# Summary

- Can traditional algorithms for data release and analysis be **replaced with provably private** algorithms while ensuring **little loss in utility**?

# Summary

- A case study on US Census Bureau Data
  - Current algorithm for data release with *no provable guarantees* and parameters used have to be kept *secret*
  - Our new algorithms provably satisfy a strong privacy guarantee (like differential privacy)
  - Our privacy guarantees ensure privacy requirements that we developed by reviewing the current interpretations of the legal regulations
  - Can release tabular summaries with *comparable or better utility* than current techniques!

# Provable Privacy Challenge

- Can traditional algorithms for data release and analysis be **replaced with provably private** algorithms while ensuring **little loss in utility**?

# Challenge 1

- Customizing Privacy to applications

# Challenge 1

- Customizing Privacy to applications

No Free Lunch
**SIGMOD 2011**

One-sided privacy
**SIGIR** Workshop **2016**

Pufferfish
**PODS 2012**

Output constrained DP
**VLDB 2017** (under review)

Blowfish
**SIGMOD 2014**
**VLDB 2015**

Privacy for Census Bureau Da
**SIGMOD 2017**

Xi He Sam Haney

# Challenge 2

- Raising awareness of DP technology

# Challenge 2

## DPComp

**SIGMOD 2016**

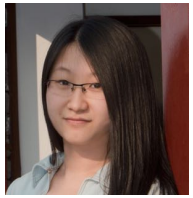

Colgate University

Duke University

UMASS Amherst

Yan Chen

George Bissias



Dan Zhang

Gerome Miklau

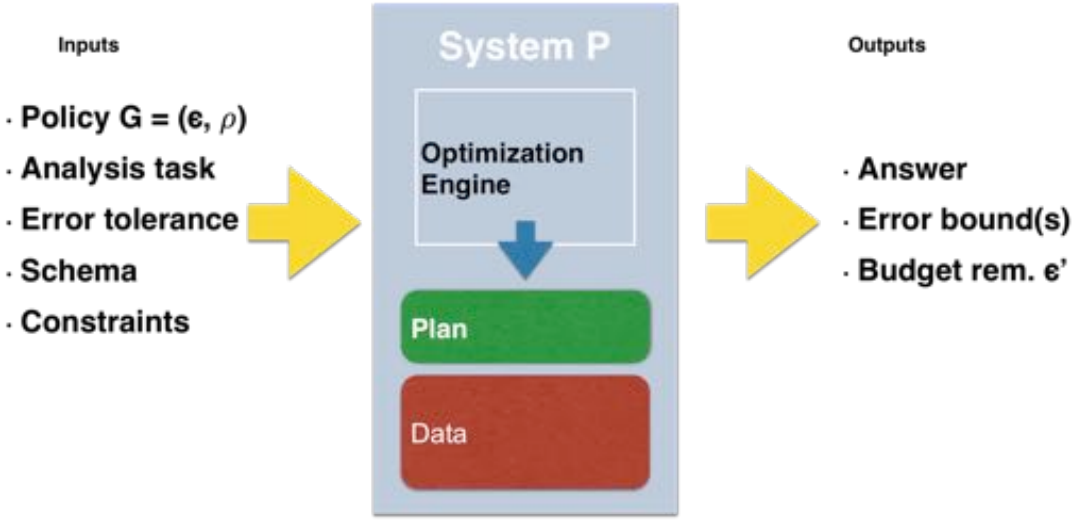
Michael Hay

# Challenge 3

- Building privacy implementations that are provably private and have low error

# Challenge 3

- Building privacy implementations that are provably private and have low error



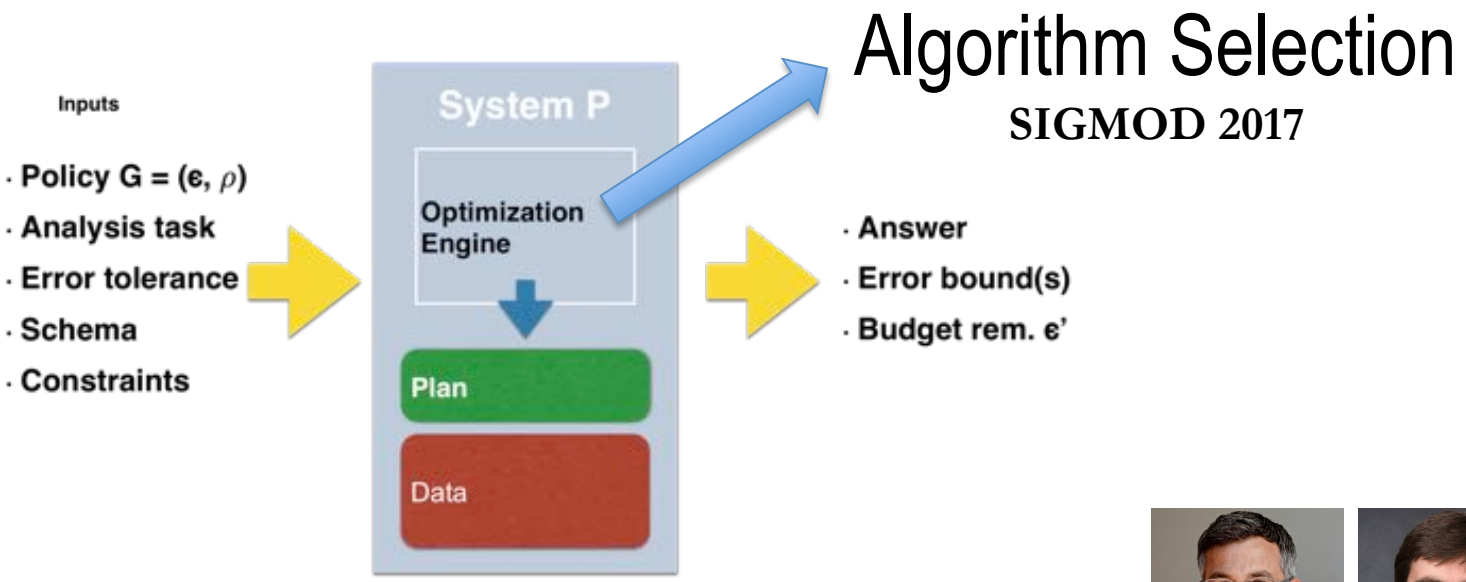



Gerome Miklau

Michael Hay

# Challenge 3

- Building privacy implementations that are provably private and have low error




Ios Kotsogiannis

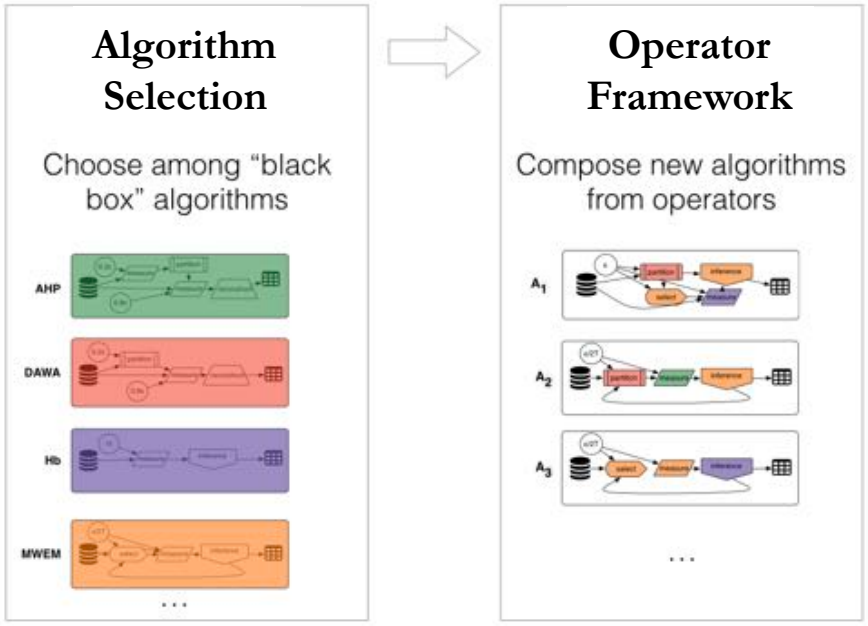
CRCS Lunch Seminar, Harvard University

Gerome Miklau

Michael Hay

# Challenge 3

- Building privacy implementations that are provably private and have low error






Gerome Miklau

Michael Hay

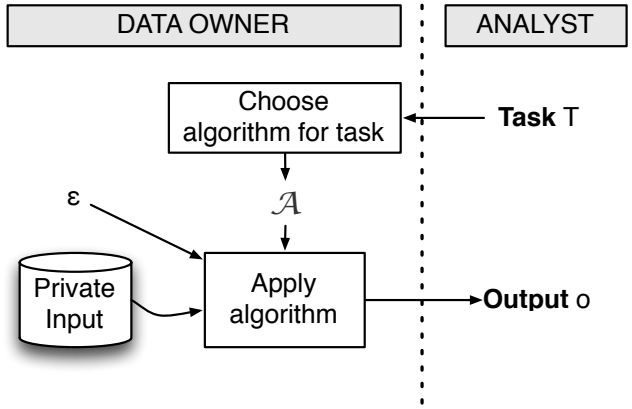
# Challenge 4

- Enabling end-to-end science on private data

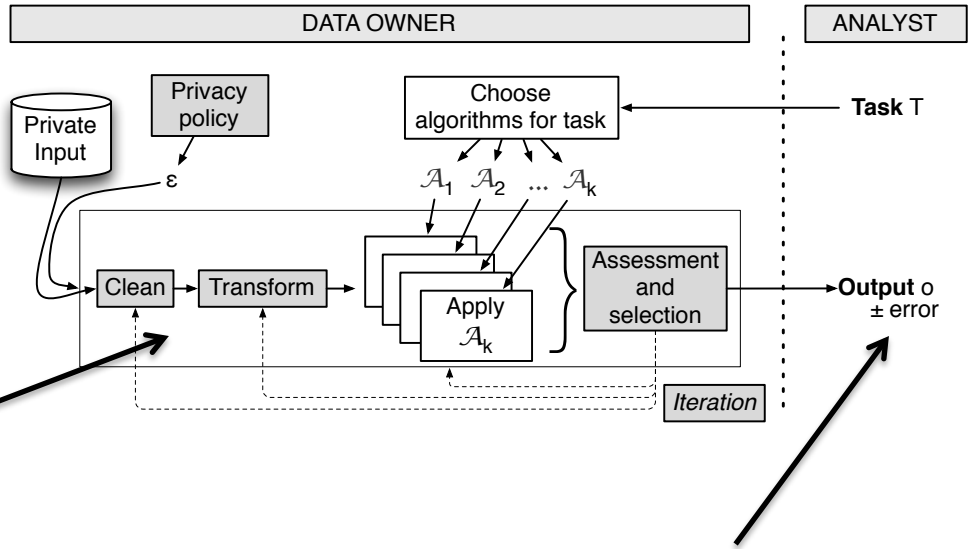
# Challenge 4

## *"Gap between theory and practice"*

*World according to DP research*




*Real World*




*Preprocessing*

*Estimate of Error on dataset*

# Summary

- Goal: Replace traditional algorithms for data release and analysis **with provably private** algorithms while ensuring **little loss in utility**.

- Can be done …

- … number of interesting theoretical and systems research challenges

# Thank you ☺

[Dwork ICALP 06] C. Dwork, "Differential Privacy", ICALP 2006

[Dwork et al TCC 06] C. Dwork, F. McSherry, K. Nissim, A. Smith, "Calibrating noise to sensitivity in private data analysis", TCC 2006

[Nissim et al STOC 07] K. Nissim, S. Raskhodnikova, A. Smith, "Smooth Sensitivity and sampling in private data analysis", STOC 2007

[Machanavajjhala et al ICDE 08] A. Machanavajjhala, D. Kifer, J. Gehrke, J. Abowd, L. Vilhuber, "Privacy: From theory to practice on the map", ICDE 2008

[Kifer-Machanavajjhala SIGMOD11] D. Kifer, A. Machanavajjhala, "No Free Lunch in Data Privacy", SIGMOD 2011

[Kifer-Machanavajjhala PODS 12] D. Kifer, A. Machanavajjhala, "A rigorous and customization framework for privacy", PODS 2012

[Haney et al SIGMOD 2017] S. Haney, A. Machanavajjhala, J. Abowd, M. Graham, M. Kutzbach, L. Vilhuber, "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics" SIGMOD 2017