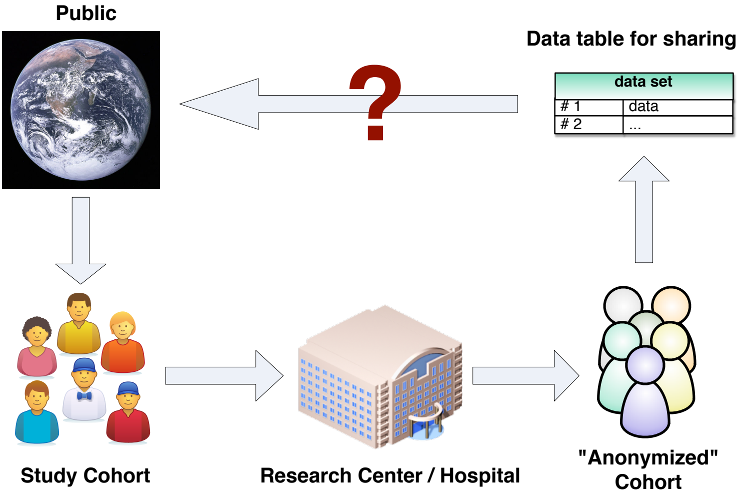
# Algorithms for privacy-preserving machine learning
(and signal processing?)

Anand D. Sarwate

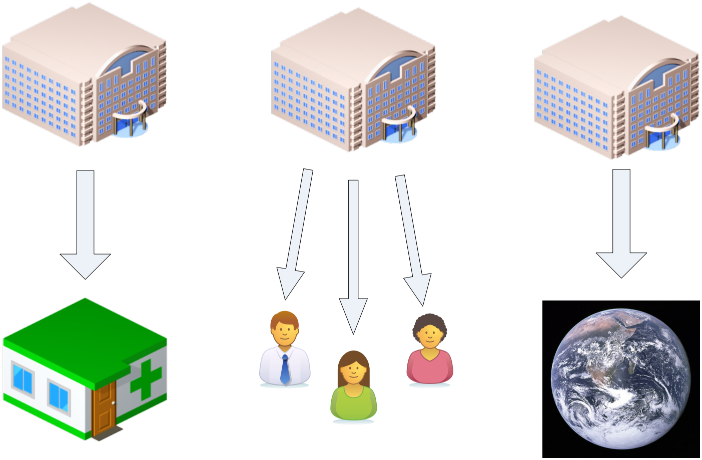
Toyota Technological Institute at Chicago

February 21, 2013

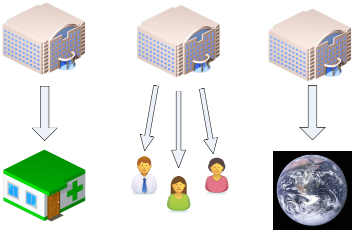
# Data sharing



# Data sharing



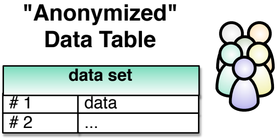
# Concerns about sharing data



There are many issues with sharing sensitive data:

- **Technological** : how do we make information private?
- **Ethical** : what is the harm caused by a breach of privacy?
- **Legal** : what are the obligations of the data holder to protect privacy?

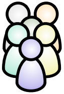
# Linkage and privacy attacks



# Linkage and privacy attacks

**"Anonymized" Data Table**

| data set | |
|---|---|
| # 1 | data |
| # 2 | ... |



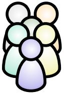
**Other public data**

| public record | |
|---|---|
| A | other data |
| B | ... |

# Linkage and privacy attacks

**"Anonymized" Data Table**

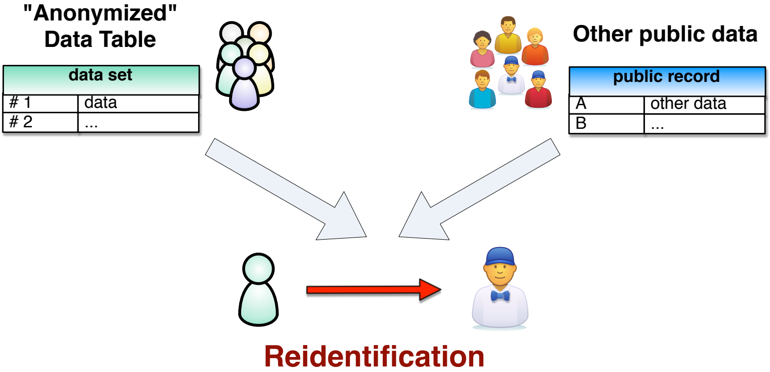
| data set | |
|---|---|
| # 1 | data |
| # 2 | ... |



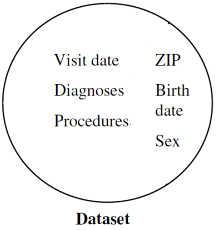
**Other public data**

| public record | |
|---|---|
| A | other data |
| B | ... |

# Linkage and privacy attacks

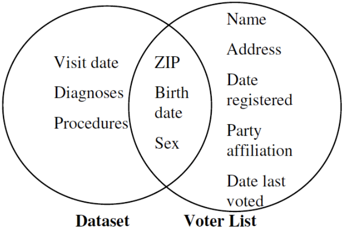


# Cautionary tales




Sweeney 1997

# Cautionary tales




Sweeney 1997

# Cautionary tales

OPEN ACCESS Freely available online PLOS GENETICS

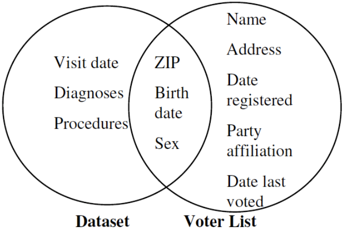
**Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays**

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

**Abstract**

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.




Homer et al. 1998

# Cautionary tales


WIRED SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO » HOW-TOS

Sign In | RSS Feeds

THREAT LEVEL

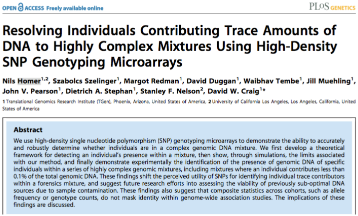
PRIVACY, CRIME AND SECURITY ONLINE

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

By Ryan Singel March 12, 2010 | 2:48 pm | Categories: privacy

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.

Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to expand the amount of information it gave to researchers in hopes that its recommendation system — a key part of Netflix's customer retention strategy — would get even better. That was then followed with a warning by prominent data privacy


OPEN ACCESS Freely available online PLoS GENETICS

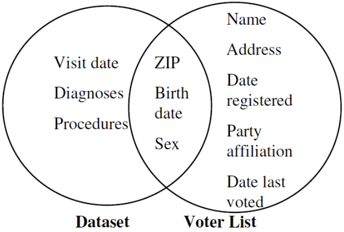
Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

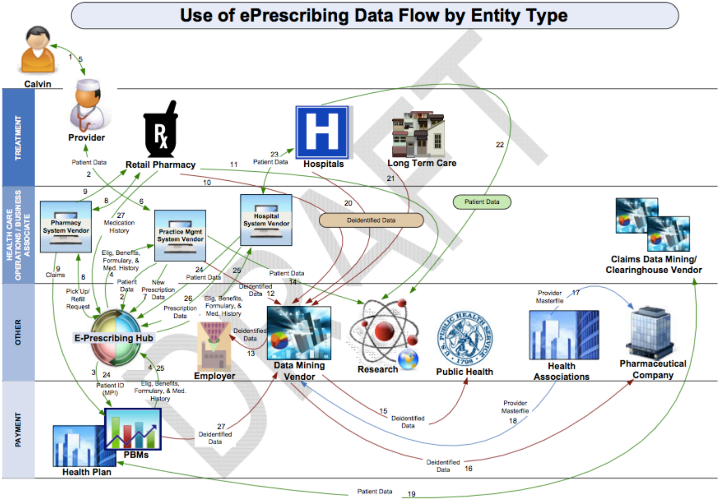
Abstract

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

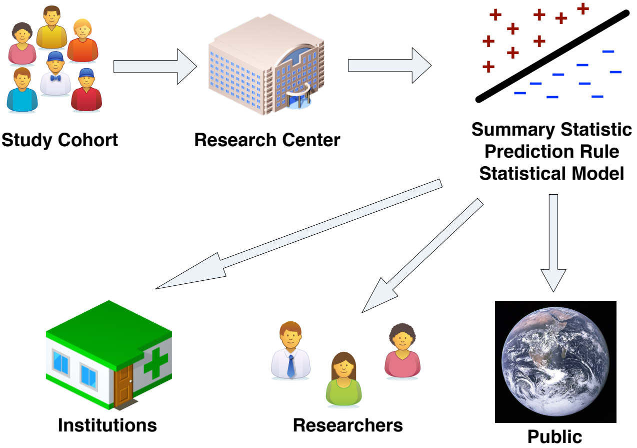



Narayanan and Shmatikov 2008

# Data flows are often invisible

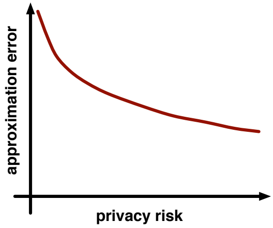


# Share results, not data




**Challenge : design useful algorithms that protect privacy.**

# The statistical setting and privacy-utility tradeoffs

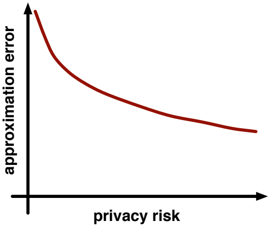



**Less Data**

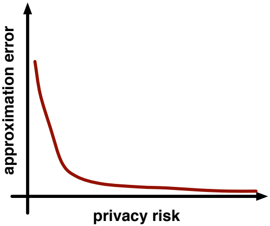
The more data we have the better off we are:

- Stronger evidence for structure → more accuracy
- Less dependence on individuals → more privacy

# The statistical setting and privacy-utility tradeoffs




**Less Data**

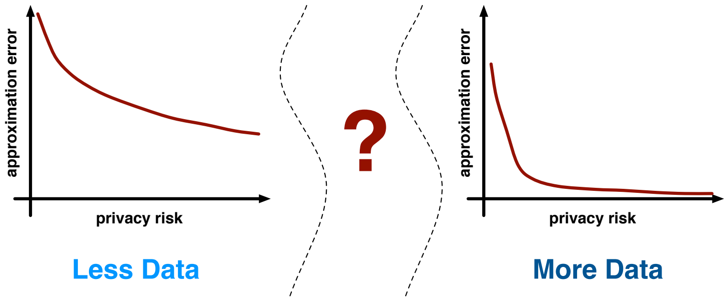



**More Data**

The more data we have the better off we are:

- Stronger evidence for structure → more accuracy
- Less dependence on individuals → more privacy

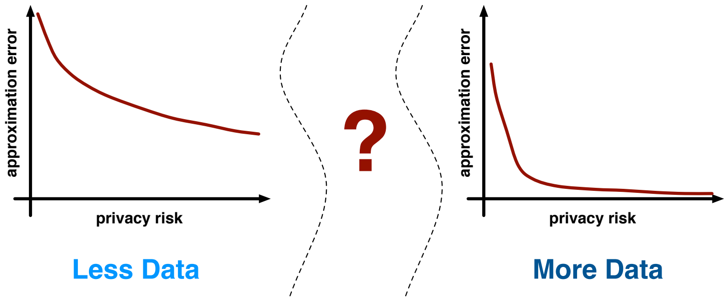
# The statistical setting and privacy-utility tradeoffs




The more data we have the better off we are:

- Stronger evidence for structure → more accuracy
- Less dependence on individuals → more privacy

# The statistical setting and privacy-utility tradeoffs

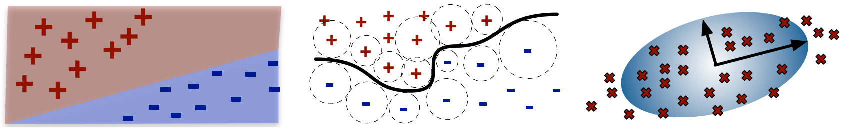



The more data we have the better off we are:

- Stronger evidence for structure → more accuracy
- Less dependence on individuals → more privacy

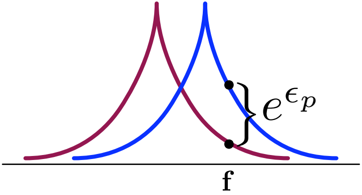
**How much data do we need?**

# This talk



Introduce

1. An introduction to differential privacy
2. Privacy preserving algorithms
3. Algorithms for classification
4. Algorithms for dimension reduction
5. Some thoughts for signal processing




# Defining privacy

# What is privacy?

- Privacy is something that matters to individuals.

# What is privacy?

- Privacy is something that matters to individuals.
- Data is itself inherently identifying.

# What is privacy?

- Privacy is something that matters to individuals.
- Data is itself inherently identifying.
- Privacy depends on what is already "known publicly"

# What is privacy?

- Privacy is something that matters to individuals.
- Data is itself inherently identifying.
- Privacy depends on what is already "known publicly"
- The only way to "maintain privacy" is to release nothing.

# What is privacy?

- Privacy is something that matters to individuals.
- Data is itself inherently identifying.
- Privacy depends on what is already “known publicly”
- The only way to “maintain privacy” is to release nothing.
- Privacy erodes over time.

# What is privacy?

**Privacy is "lost" when we handle the data.**

# What is privacy?

**Protect privacy while processing the data.**

# An example

**ages** **average** **result**

$$\mathcal{D} \longrightarrow \boxed{\frac{1}{n}\sum_i x_i} \longrightarrow 21.3$$

$$\{21, 16, \ldots, 20, 24\}$$

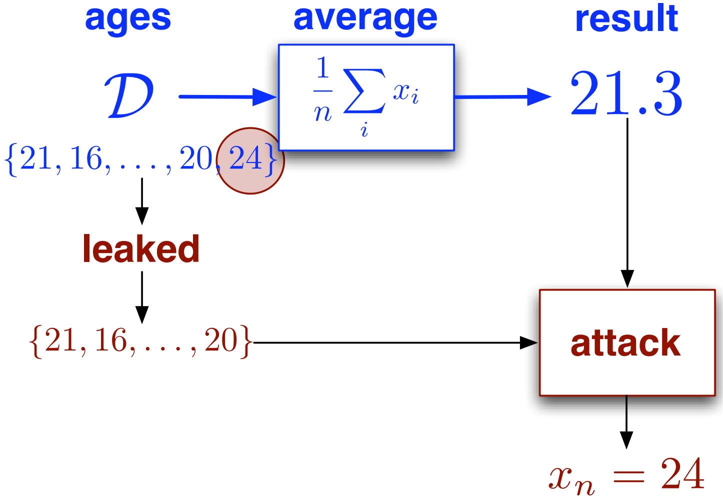
# An example

**ages** **average** **result**

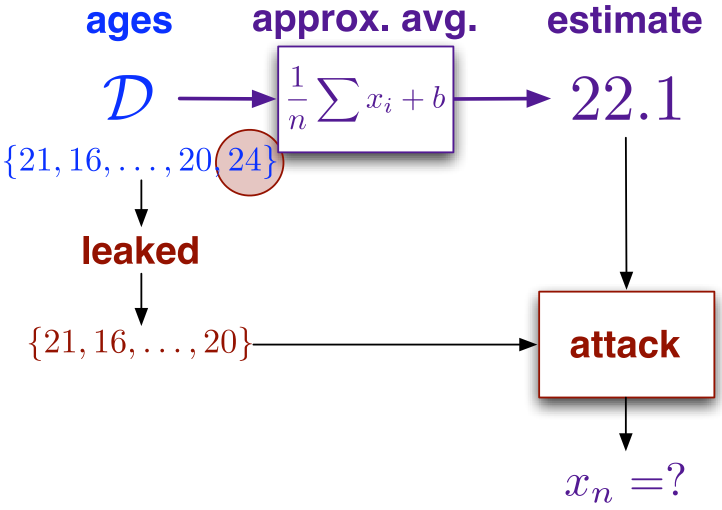
$$\mathcal{D} \longrightarrow \boxed{\frac{1}{n}\sum_i x_i} \longrightarrow 21.3$$

$$\{21, 16, \ldots, 20, 24\}$$

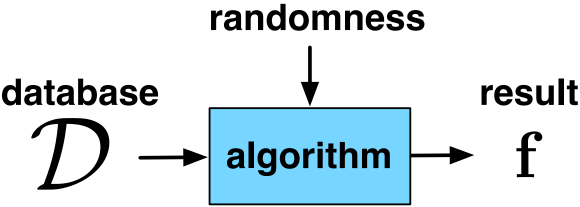
# An example


ages
average
result
$\mathcal{D}$
$\frac{1}{n}\sum_i x_i$
$21.3$
$\{21, 16, \ldots, 20, 24\}$
leaked
$\{21, 16, \ldots, 20\}$
attack
$x_n = 24$

# An example

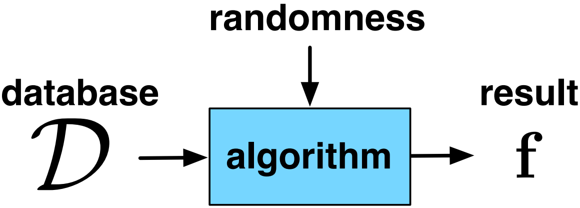

ages
approx. avg.
estimate
$\mathcal{D}$
$\frac{1}{n}\sum x_i + b$
$22.1$
$\{21, 16, \ldots, 20, 24\}$
leaked
$\{21, 16, \ldots, 20\}$
attack
$x_n = ?$

# Privacy via randomization




Algorithms that provide privacy are *randomized*:

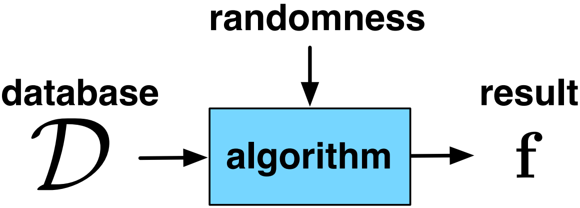
# Privacy via randomization




Algorithms that provide privacy are *randomized*:

- Database $\mathcal{D}$ has $n$ private data points.

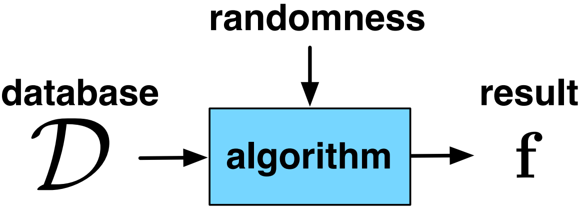
# Privacy via randomization




Algorithms that provide privacy are *randomized*:

- Database $\mathcal{D}$ has $n$ private data points.
- Algorithm $\hat{\mathcal{A}}$ is a randomized approximation to a desired function.

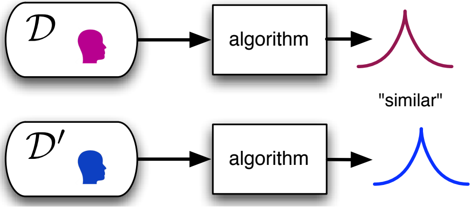
# Privacy via randomization



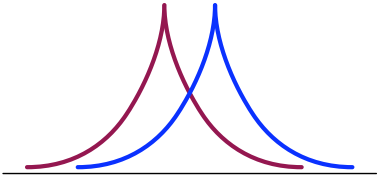

Algorithms that provide privacy are *randomized*:

- Database $\mathcal{D}$ has $n$ private data points.
- Algorithm $\hat{\mathcal{A}}$ is a randomized approximation to a desired function.
- Output $\mathbf{f}$ is a random variable.

# The definition of differential privacy



# The definition of differential privacy

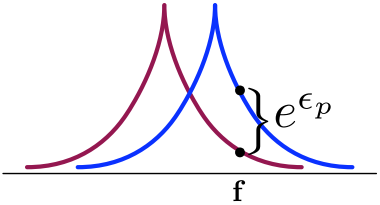


An algorithm $\hat{\mathcal{A}}$ is $\epsilon_p$-differentially private if for any set of outputs $\mathcal{F}$, and all $(\mathcal{D}, \mathcal{D}')$ differing in a single point,

$$\mathbb{P}\left(\hat{\mathcal{A}}(\mathcal{D}) \in \mathcal{F}\right) \leq \exp(\epsilon_p) \cdot \mathbb{P}\left(\hat{\mathcal{A}}(\mathcal{D}') \in \mathcal{F}\right)$$

The distribution of the outputs under neighboring databases is close. (Dwork et al., 2006)

# The definition of differential privacy




An algorithm $\hat{\mathcal{A}}$ is $\epsilon_p$-differentially private if for any set of outputs $\mathcal{F}$, and all $(\mathcal{D}, \mathcal{D}')$ differing in a single point,

$$\mathbb{P}\left(\hat{\mathcal{A}}(\mathcal{D}) \in \mathcal{F}\right) \le \exp(\epsilon_p) \cdot \mathbb{P}\left(\hat{\mathcal{A}}(\mathcal{D}') \in \mathcal{F}\right)$$

The distribution of the outputs under neighboring databases is close. (Dwork et al., 2006)

# Differential privacy and process

1. **Privacy for individuals:** If output $\hat{\mathcal{A}}(\mathcal{D})$ has a density, then

$$\left| \log \frac{p\left(\hat{\mathcal{A}}(\mathcal{D}) = \mathbf{f}\right)}{p\left(\hat{\mathcal{A}}(\mathcal{D}') = \mathbf{f}\right)} \right| \le \epsilon_p.$$

Small LLR means difficulty in disambiguation even when $\mathcal{D} \cap \mathcal{D}'$ is revealed.

# Differential privacy and process

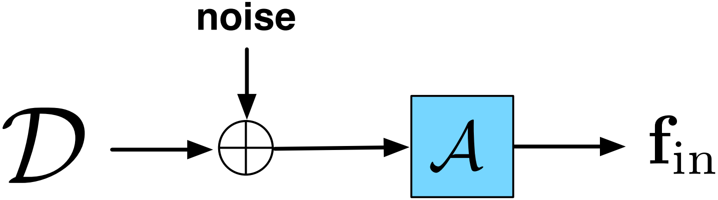
1. **Privacy for individuals:** If output $\hat{\mathcal{A}}(\mathcal{D})$ has a density, then

$$\left| \log \frac{p\left(\hat{\mathcal{A}}(\mathcal{D}) = \mathbf{f}\right)}{p\left(\hat{\mathcal{A}}(\mathcal{D}') = \mathbf{f}\right)} \right| \le \epsilon_p.$$

   Small LLR means difficulty in disambiguation even when $\mathcal{D} \cap \mathcal{D}'$ is revealed.

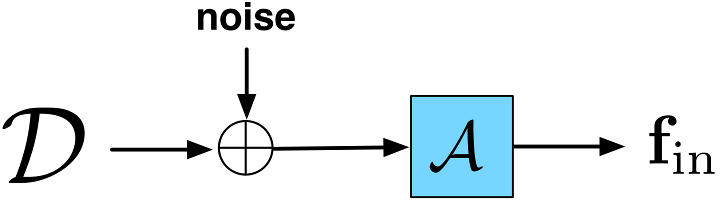
2. **Privacy for data:** No assumption that one can be "lost in the crowd" or that there is a metric on data points to measure "closeness." Distance between databases is Hamming distance.

# Input perturbation



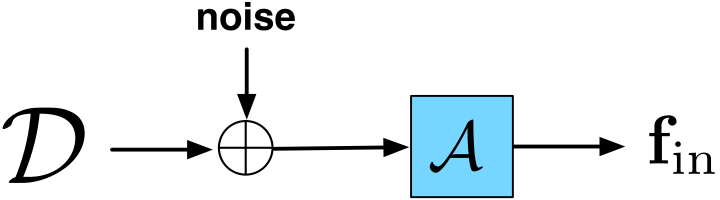

- Target function $\mathcal{A}(\mathcal{D})$ that we want to approximate.

# Input perturbation



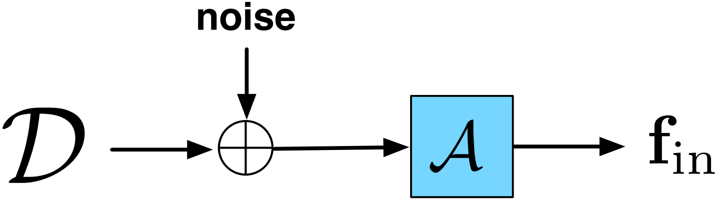

- Target function $\mathcal{A}(\mathcal{D})$ that we want to approximate.
- Add noise to data $\mathcal{D}$ and then compute $\mathcal{A}$.

# Input perturbation



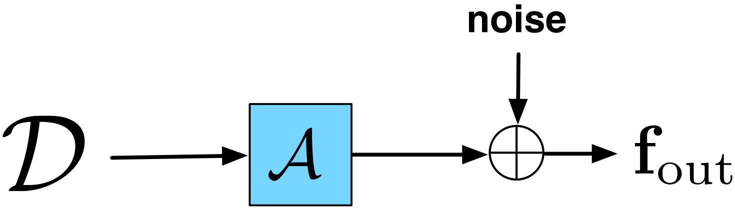

- Target function $\mathcal{A}(\mathcal{D})$ that we want to approximate.
- Add noise to data $\mathcal{D}$ and then compute $\mathcal{A}$.

# Input perturbation




- Target function $\mathcal{A}(\mathcal{D})$ that we want to approximate.
- Add noise to data $\mathcal{D}$ and then compute $\mathcal{A}$.
- Mapping from $\mathcal{D}$ to noisy version has to satisfy differential privacy.

# Output perturbation : adding noise




- Compute desired $\mathcal{A}$, then add noise to output before release.
- Tune noise to the "sensitivity" of $\mathcal{A}$ to changes in its input.

# Some difficulties

There are many technical hurdles to overcome:

# Some difficulties

There are many technical hurdles to overcome:

- Guarantees are different for discrete versus continuous data.

# Some difficulties

There are many technical hurdles to overcome:

- Guarantees are different for discrete versus continuous data.
- Guarantees often scale poorly with data dimension.

# Some difficulties

There are many technical hurdles to overcome:

- Guarantees are different for discrete versus continuous data.
- Guarantees often scale poorly with data dimension.
- Modest changes in $\epsilon_p$ have a large effect empirically.

# Some difficulties

There are many technical hurdles to overcome:

- Guarantees are different for discrete versus continuous data.
- Guarantees often scale poorly with data dimension.
- Modest changes in $\epsilon_p$ have a large effect empirically.
- All computations must be made differentially private (even parameter tuning).

# Other definitions of privacy

Previous privacy approaches enforce ambiguity in the map from data to individuals. Idea is to "quantize" data values so that many individuals have the same data.

- $k$-anonymity (Sweeney, 1998) , $\ell$-diversity (Machanavajjhala et al., 2006) , $t$-closeness (Li et al., 2007) , $m$-invariance (Xiao and Tian, 2007)
- Data can still be combined to re-identify individuals (Dwork et al., 2006) (Ganta et al., 2008)

# Other definitions of privacy

Previous privacy approaches enforce ambiguity in the map from data to individuals. Idea is to “quantize” data values so that many individuals have the same data.

- $k$-anonymity (Sweeney, 1998) , $\ell$-diversity (Machanavajjhala et al., 2006) , $t$-closeness (Li et al., 2007) , $m$-invariance (Xiao and Tian, 2007)
- Data can still be combined to re-identify individuals (Dwork et al., 2006) (Ganta et al., 2008)

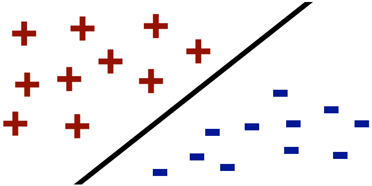
Other approaches to quantifying privacy : information theoretic security (Sankar et al. 2010) or secure multiparty computation (Vaidya and Clifton 2005)

# A perspective from learning theory

Learning theory is concerned with what things *can be learned*:

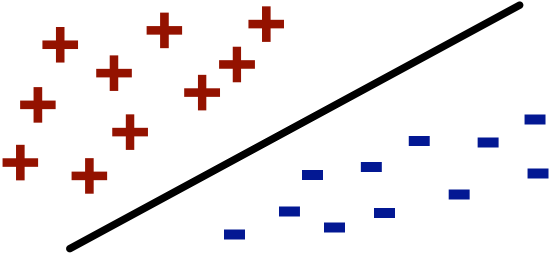
- PAC learning is possible under differential privacy (Kasiviswanathan et al 2008)
- Private learning is not characterized by VC dimension (Beimel et al. 2012)
- Parametric inference is possible (Smith 2011)
- Various learning algorithms will work with enough data (lots of people)

There is a complex interplay between assumptions on the data and the feasibility or efficiency of differentially private learning.



# Differentially private classification

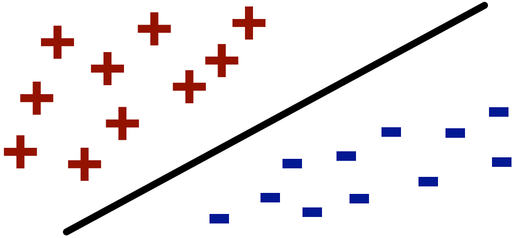
# Classification



**Input** : Data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$.
Data $\mathbf{x}_i \in \mathbb{R}^d$ with $\|\mathbf{x}_i\| \leq 1$ and labels $y_i \in \{-1, +1\}$.

**Output** : Vector $\mathbf{f} \in \mathbb{R}^d$, label points $\mathrm{sgn}(\mathbf{f}^T \mathbf{x})$.

# Using ERM



In empirical risk minimization, we choose $\mathbf{f}$ to minimize

$$J(\mathbf{f}, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{f}^T \mathbf{x}_i, y_i) + \frac{\Lambda}{2} \|\mathbf{f}\|^2$$

Want low empirical risk without overfitting.

# Why is ERM non-private?

Suppose a single point changes in the data set $\mathcal{D}$:

$$\mathcal{D}' = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}'_n, y'_n)\}.$$

# Why is ERM non-private?

Suppose a single point changes in the data set $\mathcal{D}$:

$$\mathcal{D}' = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}'_n, y'_n)\}.$$

Solution of ERM will change:

$$J(\mathbf{f}, \mathcal{D}) = \frac{1}{n}\ell(\mathbf{f}^T \mathbf{x}_n, y_n) + \frac{1}{n}\sum_{i=1}^{n-1} \ell(\mathbf{f}^T \mathbf{x}_i, y_i) + \frac{\Lambda}{2}\|\mathbf{f}\|^2$$

$$J(\mathbf{f}, \mathcal{D}') = \frac{1}{n}\ell(\mathbf{f}^T \mathbf{x}'_n, y'_n) + \frac{1}{n}\sum_{i=1}^{n-1} \ell(\mathbf{f}^T \mathbf{x}_i, y_i) + \frac{\Lambda}{2}\|\mathbf{f}\|^2$$

# Why is ERM non-private?

Suppose a single point changes in the data set $\mathcal{D}$:

$$\mathcal{D}' = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}), (\mathbf{x}_n', y_n')\}.$$

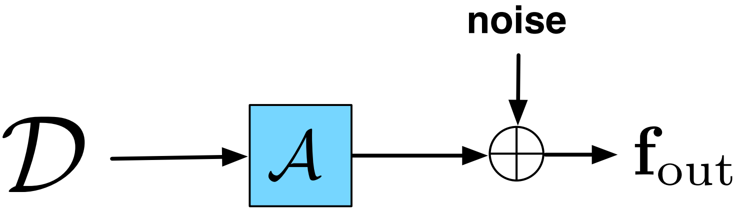
Solution of ERM will change:

$$J(\mathbf{f}, \mathcal{D}) = \frac{1}{n}\ell(\mathbf{f}^T\mathbf{x}_n, y_n) + \frac{1}{n}\sum_{i=1}^{n-1}\ell(\mathbf{f}^T\mathbf{x}_i, y_i) + \frac{\Lambda}{2}\|\mathbf{f}\|^2$$

$$J(\mathbf{f}, \mathcal{D}') = \frac{1}{n}\ell(\mathbf{f}^T\mathbf{x}_n', y_n') + \frac{1}{n}\sum_{i=1}^{n-1}\ell(\mathbf{f}^T\mathbf{x}_i, y_i) + \frac{\Lambda}{2}\|\mathbf{f}\|^2$$

That is, $\mathrm{argmin}_{\mathbf{f}}\, J(\mathbf{f}, \mathcal{D}) \neq \mathrm{argmin}_{\mathbf{f}}\, J(\mathbf{f}, \mathcal{D}')$. Change in the $n$-th individual can be detected if other data are known.

# Output perturbation for ERM

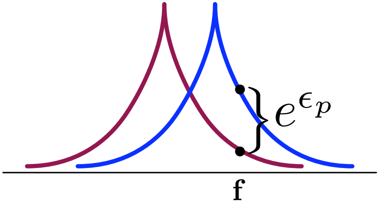



**Sensitivity method** (output perturbation) : add noise to output

$$\mathbf{f}_{\mathrm{out}} = \left( \operatorname*{argmin}_{\mathbf{f}} J(\mathbf{f}) \right) + \mathbf{a}$$

Choose $\mathbf{a}$ with density $\propto \exp(-\alpha \|\mathbf{a}\|)$ to guarantee $\epsilon_p$ privacy.

# Why output perturbation works




Density of output $\mathbf{f}_{\mathrm{out}}$ is just shifted density of $\mathbf{a}$:

$$p(\mathbf{f}_{\mathrm{out}}|\mathcal{D}) = p_{\mathbf{a}}(\mathbf{f}_{\mathrm{out}} - \underset{\mathbf{f}}{\operatorname{argmin}}\, J(\mathbf{f}))$$

Parameter $\alpha$ chosen to match the shift in ERM solution between $\mathcal{D}$ and $\mathcal{D}'$.

# Objective perturbation for ERM

$$\mathcal{D} \longrightarrow \boxed{\arg\min \left( J(\mathbf{f}) + \mathbf{b}^T \mathbf{f} \right)} \longrightarrow \mathbf{f}_{\mathrm{obj}}$$

$$\mathbf{f}_{\mathrm{obj}} = \underset{\mathbf{f}}{\operatorname{argmin}} \left( J(\mathbf{f}) + \color{red}{\mathbf{b}^T \mathbf{f}} \right)$$

# Objective perturbation for ERM

$$\mathcal{D} \longrightarrow \boxed{\arg\min \left(J(\mathbf{f}) + \mathbf{b}^T \mathbf{f}\right)} \longrightarrow \mathbf{f}_{\mathrm{obj}}$$

$$\mathbf{f}_{\mathrm{obj}} = \underset{\mathbf{f}}{\operatorname{argmin}} \left(J(\mathbf{f}) + \mathbf{b}^T \mathbf{f}\right)$$

- Add perturbation inside the objective function.

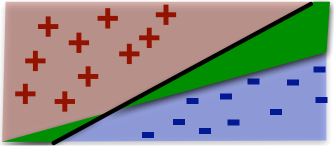
# Objective perturbation for ERM

$$\mathcal{D} \longrightarrow \boxed{\arg\min \left(J(\mathbf{f}) + \mathbf{b}^T \mathbf{f}\right)} \longrightarrow \mathbf{f}_{\mathrm{obj}}$$

$$\mathbf{f}_{\mathrm{obj}} = \operatorname*{argmin}_{\mathbf{f}} \left(J(\mathbf{f}) + \mathbf{b}^T \mathbf{f}\right)$$

- Add perturbation inside the objective function.
- Choose $\mathbf{b}$ with density $\propto \exp(-\beta \|\mathbf{b}\|)$

# Sample complexity for the two methods



For privacy $\epsilon_p$ and generalization error $\epsilon_g$:

1. **Output perturbation**

$$n = \Omega\left\{ \frac{1}{\epsilon_g^2}, \frac{d \log d}{\epsilon_p^{3/2} \epsilon_g} \right\}$$

2. **Objective perturbation**

$$n = \Omega\left\{ \frac{1}{\epsilon_g^2}, \frac{d \log d}{\epsilon_p \epsilon_g} \right\}$$

# Sample complexity for objective perturbation

## Theorem (Excess error of $\mathbf{f}_{\text{obj}}$)

*Let* $\ell$ *be convex, doubly differentiable, and let its derivatives satisfy* $\ell'(\cdot) \leq 1$ *and* $\ell''(\cdot) \leq c$ *and let* $\mathcal{D}$ *be drawn i.i.d. according to* $P$*. For any* $\mathbf{f}_0$ *with expected loss* $L(\mathbf{f}_0) = L^*$*, if*

$$n \geq C \cdot \max \left\{ \frac{||\mathbf{f}_0||^2 \log(1/\delta)}{\epsilon_g^2}, \frac{||\mathbf{f}_0||^2}{\epsilon_g \epsilon_p}, \frac{d \log(\frac{d}{\delta})||\mathbf{f}_0||}{\epsilon_g \epsilon_p} \right\}$$

*we have*

$$\mathbb{P}\left(L(\mathbf{f}_{\text{obj}}) \leq L^* + \epsilon_g\right) \geq 1 - \delta.$$

## Proof sketch

1. Fix a distribution $P$ on the data and define

$$\bar{J}(\mathbf{f}) = \mathbb{E}[\ell(\mathbf{f}^T \mathbf{x}, y)] + \frac{\Lambda}{2} \|\mathbf{f}\|^2 .$$

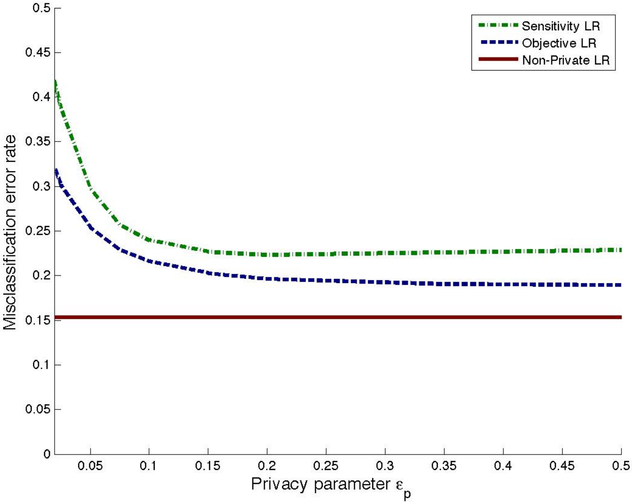
   Let the minimizer be $\mathbf{f}_{\mathrm{rtf}}$

2. For a given "good" $\mathbf{f}_0$, decompose the objective into:

$$\begin{aligned} L(\mathbf{f}_{\mathrm{priv}}) = L(\mathbf{f}_0) + (\bar{J}(\mathbf{f}_{\mathrm{priv}}) - \bar{J}(\mathbf{f}_{\mathrm{rtr}})) + (\bar{J}(\mathbf{f}_{\mathrm{rtr}}) - \bar{J}(\mathbf{f}_0)) \\ + \frac{\Lambda}{2}(\|\mathbf{f}_0\|^2 - \|\mathbf{f}_{\mathrm{priv}}\|^2) \end{aligned}$$

3. Show that the "non-bar" version of the first term is small, then show that the first term is close to the "non-bar" version. Second term is small by standard ERM results.

# Simulation results



# Intuitions

Why is objective perturbation better on real data?

# Intuitions

Why is objective perturbation better on real data?

- Objective function is more convex in some directions in other. Loss is higher in these directions.

# Intuitions

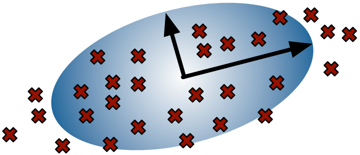
Why is objective perturbation better on real data?

- Objective function is more convex in some directions in other. Loss is higher in these directions.
- Output perturbation is agnostic to this variation : noise affects sensitive directions adversely.

# Intuitions

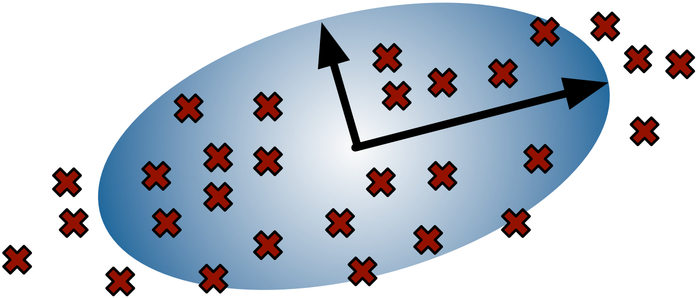
Why is objective perturbation better on real data?

- Objective function is more convex in some directions in other. Loss is higher in these directions.
- Output perturbation is agnostic to this variation : noise affects sensitive directions adversely.
- Objective perturbation allows optimization to smooth out noise in sensitive directions more effectively.



# Differentially private dimension reduction

# Dimension reduction by projection



- Data may be presented in very high dimension.
- Fundamental structure is low-dimensional.
- Other dimensions contain mostly noise.

# The PCA problem

Data in $\mathbb{R}^d$:

$$\{\mathbf{x}_i : i = 1, 2, \ldots, n\}$$

Capture structure by the second moment matrix:

$$A = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$$

The matrix $A$ captures the "geometry" of the data.

# The top-$k$ subspace

If the eigenvalues of $A$ are $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A) \geq 0$, and if

$$A = V \Lambda V^T$$

where $\Lambda$ is diagonal with $\Lambda_{ii} = \lambda_i(A)$ and $V$ is an orthonormal matrix of eigenvectors, then the *rank-$k$ PCA approximation* is

$$\hat{A} = V \Lambda_k V^T$$

where $\Lambda$ is diagonal with $\Lambda_{ii} = \lambda_i(A)$ for $i \leq k$ and $\Lambda_{ii} = 0$ for $i > k$. The Schmidt Approximation Theorem says that $\hat{A}$ minimizes the Frobenius norm:

$$\left\| A - \hat{A} \right\|_F .$$

# Goals

**Goal 0:** approximate the top-$k$ subspace under $\epsilon$ differential privacy.

# Goals

**Goal 0:** approximate the top-$k$ subspace under $\epsilon$ differential privacy.

**Goal 1:** understand the fundamental (distribution-free) limits for PCA.

# Goals

**Goal 0:** approximate the top-$k$ subspace under $\epsilon$ differential privacy.

**Goal 1:** understand the fundamental (distribution-free) limits for PCA.

**Goal 2:** examine the *performance on real data*.

# Main results

We analyze and implement the exponential mechanism (McSherry and Talwar 2007) for this problem.

# Main results

We analyze and implement the exponential mechanism (McSherry and Talwar 2007) for this problem.

- Upper bound on the number of samples needed for our method.

# Main results

We analyze and implement the exponential mechanism (McSherry and Talwar 2007) for this problem.

- Upper bound on the number of samples needed for our method.
- Nearly matching bound on the sample complexity for *any algorithm*.

# Main results

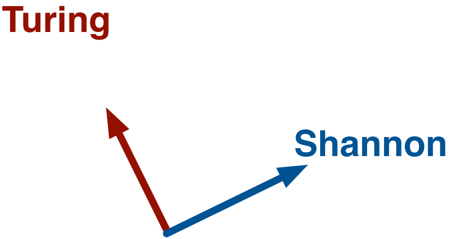
We analyze and implement the exponential mechanism (McSherry and Talwar 2007) for this problem.

- Upper bound on the number of samples needed for our method.
- Nearly matching bound on the sample complexity for *any algorithm*.
- Different lower bound on the sample complexity for input perturbation.

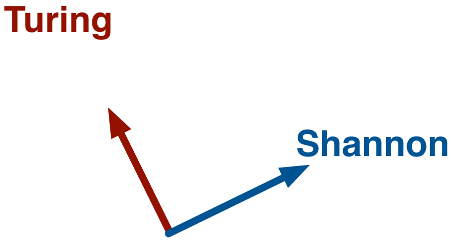
# The past and the future

- Blum, Dwork, McSherry, and Nissim (PODS 2005) : proposed adding noise to the second moment matrix
- Hardt and Roth (STOC 2012) : low rank matrix reconstruction
- Kapralov and Talwar (SODA 2013) : a different approach to this problem
- Hardt and Roth (unpublished) : different model based on matrix coherence

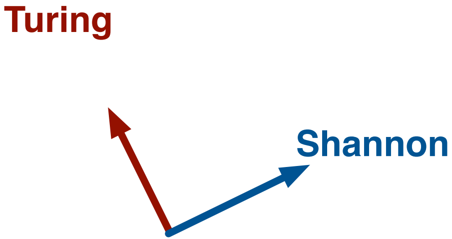
# Privacy concerns



# Privacy concerns

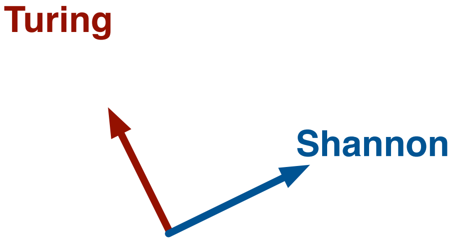


# Privacy concerns



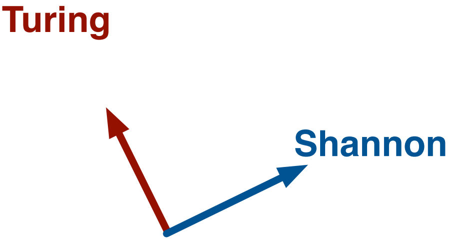

- I don't want my data (or participation) to be revealed.

## Privacy concerns



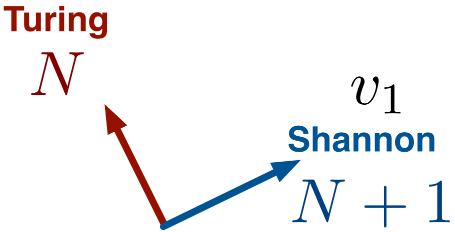

- I don't want my data (or participation) to be revealed.
- I don't trust other people to keep their data secret.

# Privacy concerns



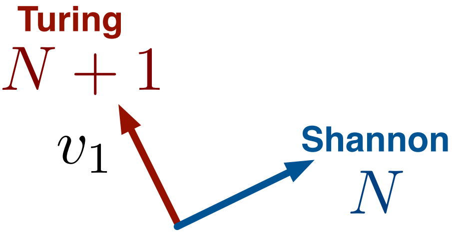

- I don't want my data (or participation) to be revealed.
- I don't trust other people to keep their data secret.

# Privacy concerns



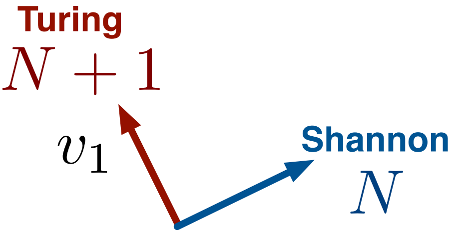

- I don't want my data (or participation) to be revealed.
- I don't trust other people to keep their data secret.

# Privacy concerns




- I don't want my data (or participation) to be revealed.
- I don't trust other people to keep their data secret.

# Privacy concerns




- I don't want my data (or participation) to be revealed.
- I don't trust other people to keep their data secret.
- Can the data holder still publish the subspace?

# PCA and differential privacy

# PCA and differential privacy

The PCA algorithm is not differentially private:

$$\mathcal{D} = \{\underbrace{\mathbf{e}_1, \mathbf{e}_1, \ldots \mathbf{e}_{\mathbf{1}}}_{N/2}, \underbrace{\mathbf{e}_2, \mathbf{e}_2, \ldots \mathbf{e}_{\mathbf{2}}}_{N/2-1}\}$$

Then $v_1(\mathcal{D}) = \mathbf{e}_1$, but change one $\mathbf{e}_1$ to $\mathbf{e}_2$ and $v_1$ changes to $\mathbf{e}_2$.

# PCA and differential privacy

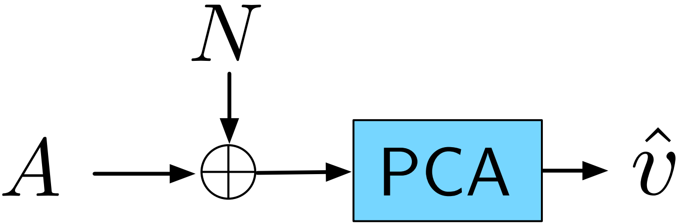
The PCA algorithm is not differentially private:

$$\mathcal{D} = \{\underbrace{\mathbf{e}_1, \mathbf{e}_1, \ldots \mathbf{e}_{\mathbf{1}}}_{N/2}, \underbrace{\mathbf{e}_2, \mathbf{e}_2, \ldots \mathbf{e}_{\mathbf{2}}}_{N/2-1}\}$$

Then $v_1(\mathcal{D}) = \mathbf{e}_1$, but change one $\mathbf{e}_1$ to $\mathbf{e}_2$ and $v_1$ changes to $\mathbf{e}_2$.

Sensitivity depends on the eigenvalue gap.

# SULQ : Input perturbation for PCA




- Add noise to $A$ and then compute PCA on $A + N$.
- This is the SULQ algorithm proposed by Blum et al. (2005)

# Lower bound on sample complexity

**Theorem**

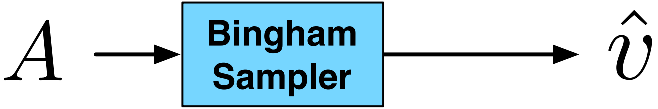
*There are constants $c$ and $c'$ such that for any $\rho$, if*

$$n < C \cdot \frac{d^{3/2}\sqrt{\log(d/\delta)}}{\epsilon_p},$$

*then there is a dataset of size $n$ in dimension $d$, s.t. the top PCA direction $v$ and the output $\hat{v}$ of SULQ satisfy*

$$\mathbb{E}[|\langle \hat{v}_1, v_1 \rangle|] \leq \rho.$$

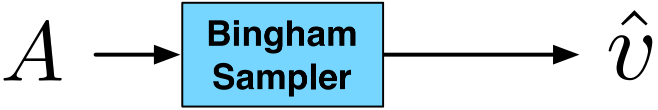
# PPCA : exponential mechanism




Sample a subspace $V$ from the vector Bingham distribution:

$$f(V) \propto \exp\left(n\frac{\epsilon_p}{2} \cdot \mathrm{tr}(V^T A V)\right)$$

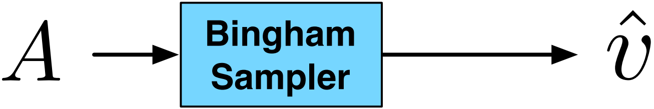
# PPCA : exponential mechanism




Sample a subspace $V$ from the vector Bingham distribution:

$$\begin{aligned} f(V) &\propto \exp\left( n \frac{\epsilon_p}{2} \cdot \mathrm{tr}(V^T A V) \right) \\ &= \frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}d, A\right)} \exp\left( n \frac{\epsilon_p}{2} \mathrm{tr}(V^T A V) \right) \end{aligned}$$

# PPCA : exponential mechanism




Sample a subspace $V$ from the vector Bingham distribution:

$$f(V) \propto \exp\left( n \frac{\epsilon_p}{2} \cdot \mathrm{tr}(V^T A V) \right)$$
$$= \frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}d, A\right)} \exp\left( n \frac{\epsilon_p}{2} \mathrm{tr}(V^T A V) \right)$$

This is the exponential mechanism with score function $\mathrm{tr}(V^T A V)$.
Works for general $k$.

# Performance for our method

## Theorem (PPCA needs less data)

*There exists an absolute constant $C$ such that the following holds. For any $\gamma > 0$, $\epsilon_p > 0$, $t > 0$, if*

$$n > C \cdot \frac{d}{\epsilon_p (1-\rho)(\lambda_1 - \lambda_2)} \cdot \log \frac{1}{(1-\rho^2)(\lambda_1 - \lambda_2)},$$

*then the top PCA direction $v_1$ and the output of our algorithm $\hat{v}_1$ with privacy parameter $\epsilon_p$ satisfy:*

$$\mathbb{P}(|\langle v_1, \hat{v}_1 \rangle| > \rho) \geq 1 - \eta$$

## Proof sketch

The proof is a refined analysis of the Exponential Mechanism (McSherry and Talwar, 2007):

1. Want to upper bound probability of landing in the set

$$\bar{\mathcal{U}}_\rho = \{u : \langle u, v_1 \rangle \leq \rho\}$$

2. An ugly bound:

$$\begin{aligned}\mathbb{P}\left(\bar{\mathcal{U}}_\rho\right) &\leq \frac{\mathbb{P}\left(\bar{\mathcal{U}}_\rho\right)}{\mathbb{P}\left(\mathcal{U}_\sigma\right)} \\ &\leq \frac{\exp(n(\alpha/2)(\rho^2 \lambda_1 + (1-\rho^2)\lambda_2)}{\exp(n(\alpha/2)(\sigma^2 \lambda_1 + (1-\sigma^2)\lambda_d)} \cdot \frac{\mathrm{Surf}\left(\bar{\mathcal{U}}_\rho\right)}{\mathrm{Surf}\left(\mathcal{U}_\sigma\right)}\end{aligned}$$

3. Then do some algebra.

# Lower bound for any method

**Theorem (General lower bound)**

*Fix $d$, $\epsilon_p$, and $\lambda_1 - \lambda_2$. Then there is a constant $C$ such that if*

$$n < C \cdot \frac{d}{\epsilon_p(\lambda_1 - \lambda_2)\sqrt{1-\rho}},$$

*the top PCA direction $v_1$ and the output of our algorithm $\hat{v}_1$ with privacy parameter $\epsilon_p$ satisfy:*

$$\mathbb{E}\left[|\langle v_1, \hat{v}_1 \rangle|\right] < \rho$$

# Proof idea for lower bound

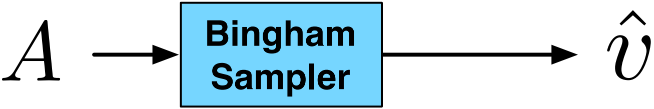
## Lemma

*Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ be $K$ databases which differ in the value of at most $\frac{\ln(K-1)}{\alpha}$ points, and let $u_1, \ldots, u_K$ be the top eigenvectors of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. If $\mathcal{A}$ is any $\alpha$-differentially private algorithm, then,*

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[ |\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \leq K \left( 1 - \frac{1}{16} (1 - \max |\langle u_i, u_j \rangle|) \right).$$

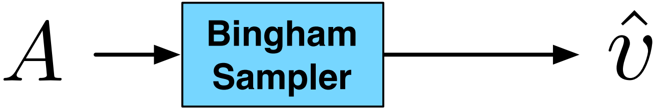
Then construct $K$ databases with this property.

# Implementing the exponential mechanism




A major difficulty is *sampling* from the Bingham distribution:

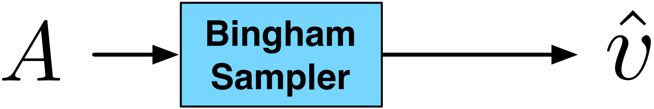
# Implementing the exponential mechanism




A major difficulty is *sampling* from the Bingham distribution:

- "Closed form" involves special functions.

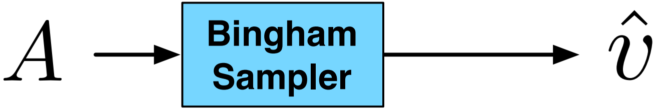
# Implementing the exponential mechanism




A major difficulty is *sampling* from the Bingham distribution:

- "Closed form" involves special functions.
- Markov Chain Monte Carlo (MCMC) sampling.

# Implementing the exponential mechanism




A major difficulty is *sampling* from the Bingham distribution:

- "Closed form" involves special functions.
- Markov Chain Monte Carlo (MCMC) sampling.
- New set of challenges to explore.

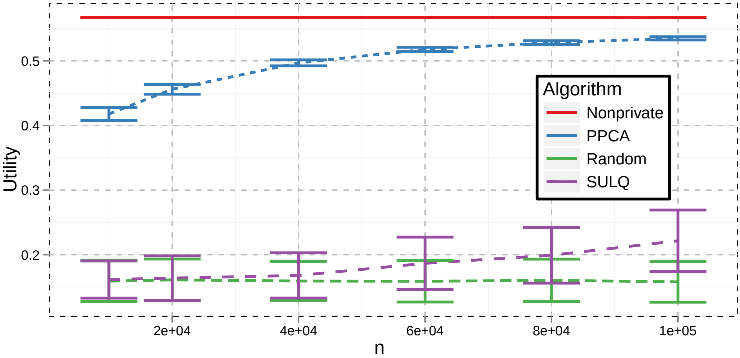
# Error versus data set size

Theoretical guarantees are for $k = 1$ but we can implement the method for general $k$:

| Dataset | #instances | #dimensions | $k$ |
|---|---|---|---|
| `kddcup` | 494,021 | 116 | 4 |
| `census` | 199,523 | 513 | 8 |
| `localization` | 164,860 | 44 | 10 |
| `insurance` | 9,822 | 150 | 11 |

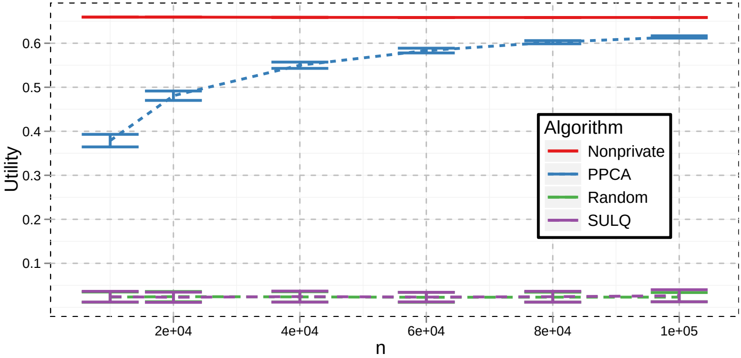
Table : Parameters of each dataset. The second column is the number of dimensions after preprocessing. $k$ is the dimensionality of the PCA, and the fourth column contains $q(U)/\left\|A\right\|_F$ where $U$ is the top $k$ PCA subspace.

# Error versus data set size



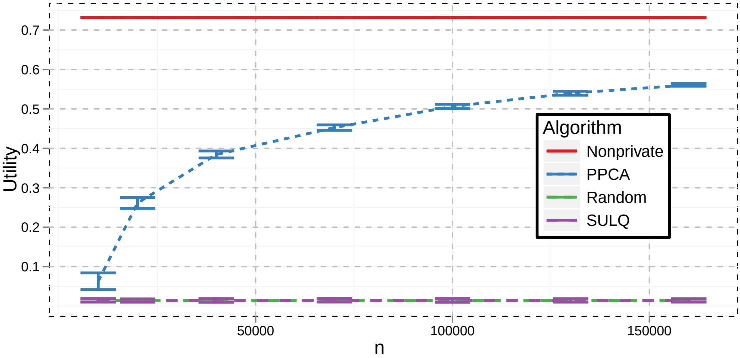

Utility $q(U)$ for `localization` for $d = 44$, $k = 10$.

# Error versus data set size



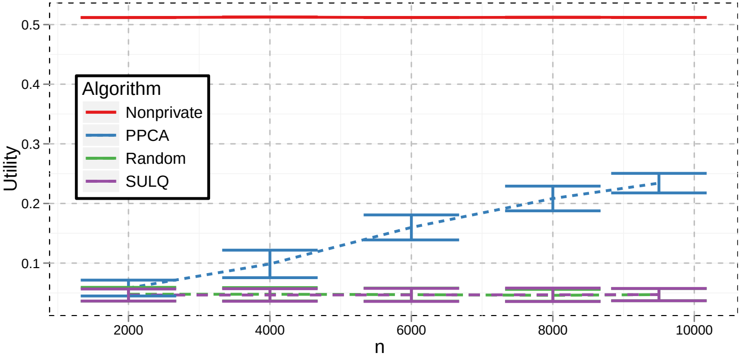

Utility $q(U)$ for `kddcup` for $d = 116$, $k = 4$.

# Error versus data set size




Utility $q(U)$ for census for $d = 513$, $k = 8$.

# Error versus data set size




Utility $q(U)$ for `insurance` for $d = 115$, $k = 11$.

# New challenges

Currently there is lots of active research on the theory side:

- Kernel learning, online learning, convex optimization

# New challenges

Currently there is lots of active research on the theory side:

- Kernel learning, online learning, convex optimization
- New and variant definitions of privacy

# New challenges

Currently there is lots of active research on the theory side:

- Kernel learning, online learning, convex optimization
- New and variant definitions of privacy

but there are important practical issues ahead:

# New challenges

Currently there is lots of active research on the theory side:

- Kernel learning, online learning, convex optimization
- New and variant definitions of privacy

but there are important practical issues ahead:

- Need more algorithms tuned to domain-specific assumptions.

# New challenges

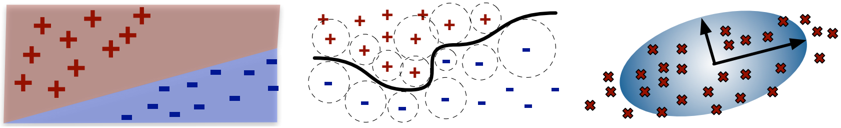
Currently there is lots of active research on the theory side:

- Kernel learning, online learning, convex optimization
- New and variant definitions of privacy

but there are important practical issues ahead:

- Need more algorithms tuned to domain-specific assumptions.
- Extensions to complex data sources (e.g. images)

# Summary



- Privacy-preserving data analysis is a rich and growing research area.
- Demonstrated and evaluated methods for ERM and PCA.
- Incorporating domain knowledge can make a big impact.

# Thank you!