Differential Privacy

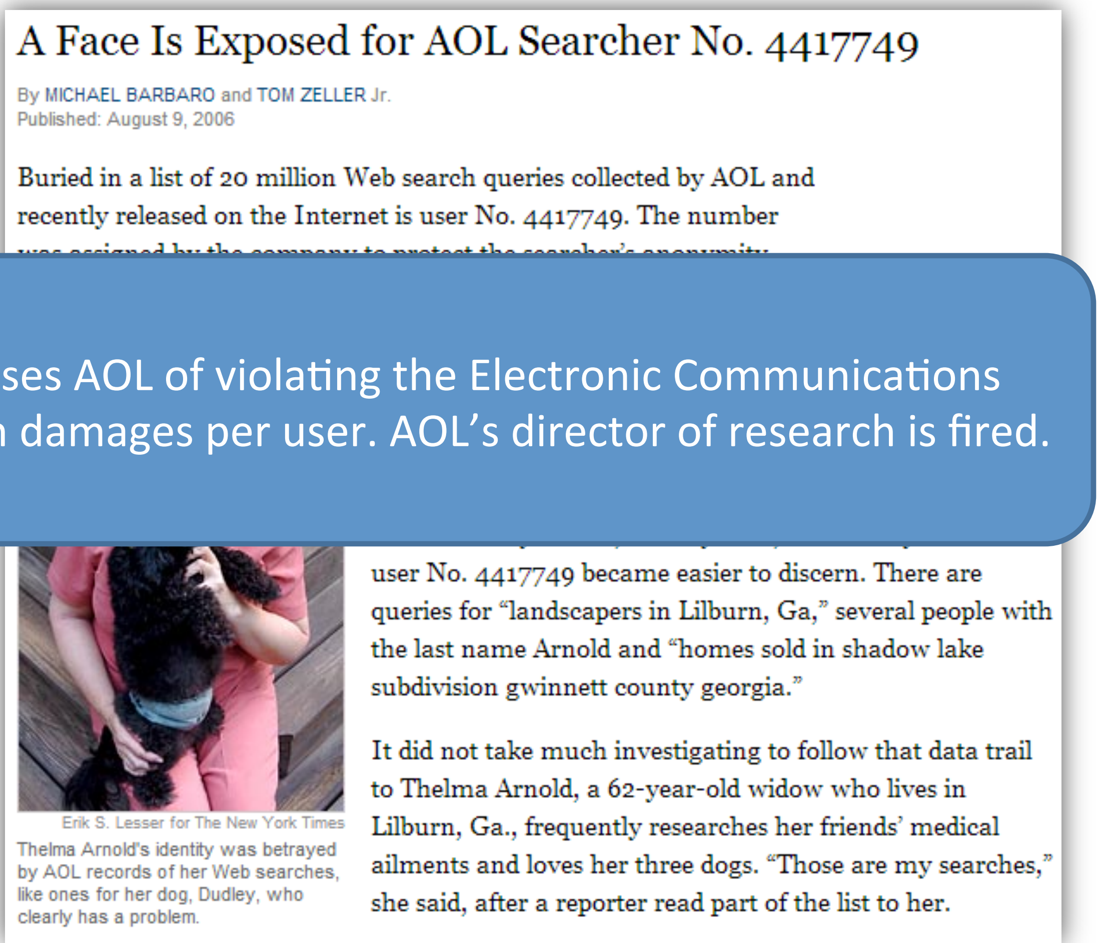
And Data Analysis

Aaron Roth

Penn

May 9, 2017

Protecting Privacy is Important



A Face Is Exposed for AOL Searcher No. 4417749

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

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity

user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

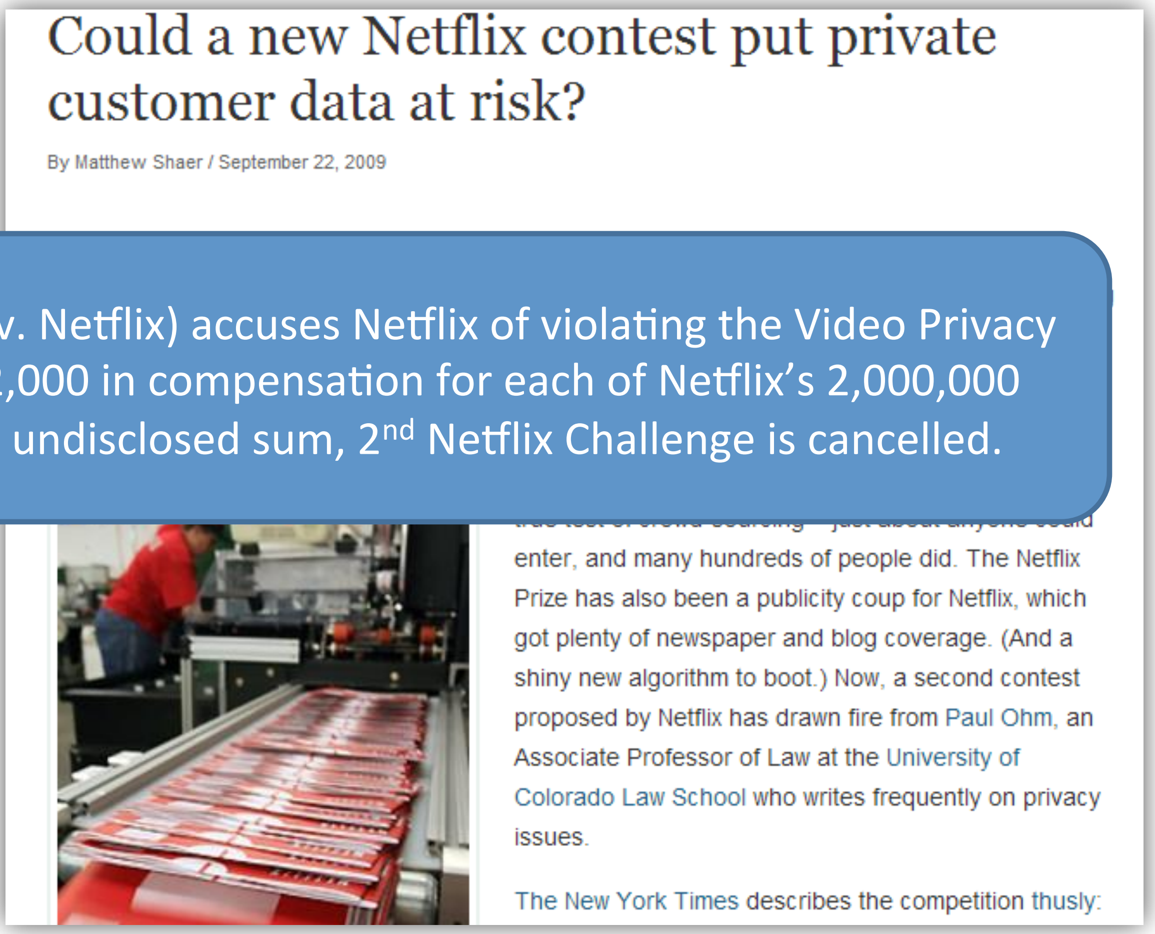
It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Class action lawsuit accuses AOL of violating the Electronic Communications Privacy Act, seeks $5,000 in damages per user. AOL's director of research is fired.

Protecting Privacy is Important



Could a new Netflix contest put private customer data at risk?

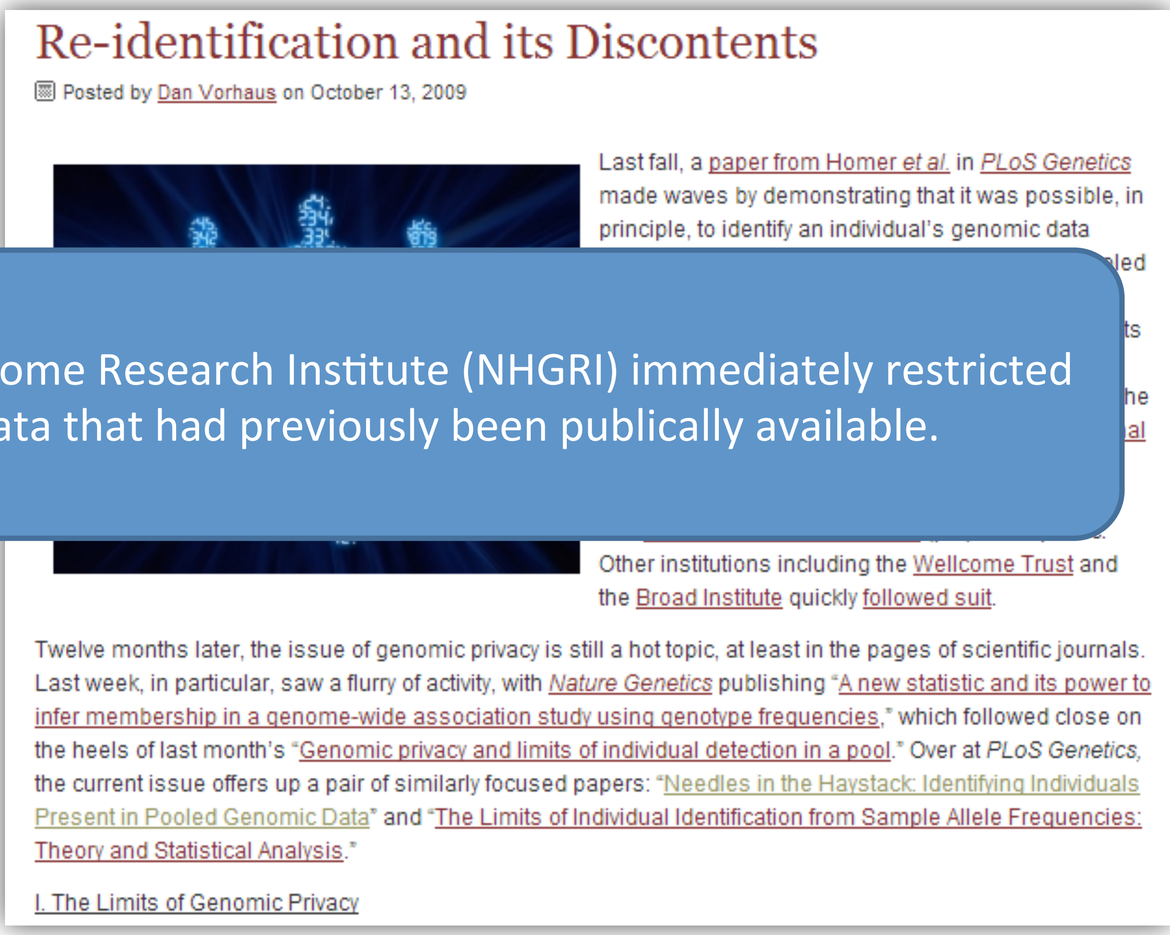
By Matthew Shaer / September 22, 2009

enter, and many hundreds of people did. The Netflix Prize has also been a publicity coup for Netflix, which got plenty of newspaper and blog coverage. (And a shiny new algorithm to boot.) Now, a second contest proposed by Netflix has drawn fire from Paul Ohm, an Associate Professor of Law at the University of Colorado Law School who writes frequently on privacy issues.

The New York Times describes the competition thusly:

Class action lawsuit (Doe v. Netflix) accuses Netflix of violating the Video Privacy Protection Act, seeks $2,000 in compensation for each of Netflix’s 2,000,000 subscribers. Settled for undisclosed sum, 2nd Netflix Challenge is cancelled.

Protecting Privacy is Important



Re-identification and its Discontents

Posted by Dan Vorhaus on October 13, 2009

Last fall, a paper from Homer *et al.* in *PLoS Genetics* made waves by demonstrating that it was possible, in principle, to identify an individual's genomic data

Other institutions including the Wellcome Trust and the Broad Institute quickly followed suit.

Twelve months later, the issue of genomic privacy is still a hot topic, at least in the pages of scientific journals. Last week, in particular, saw a flurry of activity, with *Nature Genetics* publishing "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies," which followed close on the heels of last month's "Genomic privacy and limits of individual detection in a pool." Over at *PLoS Genetics*, the current issue offers up a pair of similarly focused papers: "Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data" and "The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis."

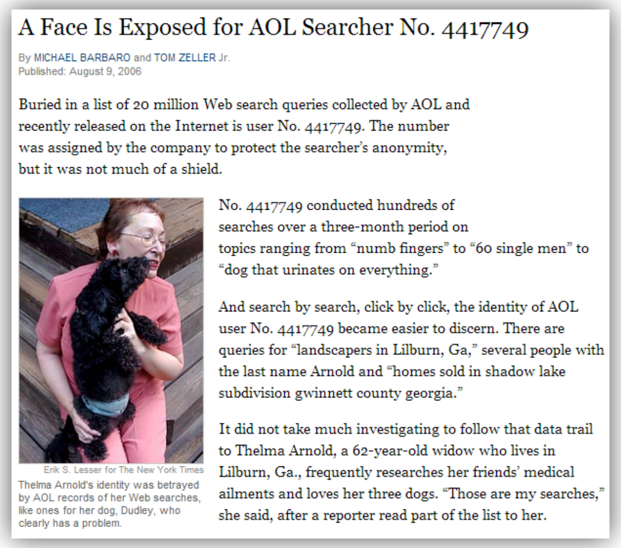
I. The Limits of Genomic Privacy

The National Human Genome Research Institute (NHGRI) immediately restricted pooled genomic data that had previously been publically available.

But what is “privacy”?

But what is “privacy” not?

- Privacy is not hiding “personally identifiable information” (name, zip code, age, etc…)



A Face Is Exposed for AOL Searcher No. 4417749

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

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

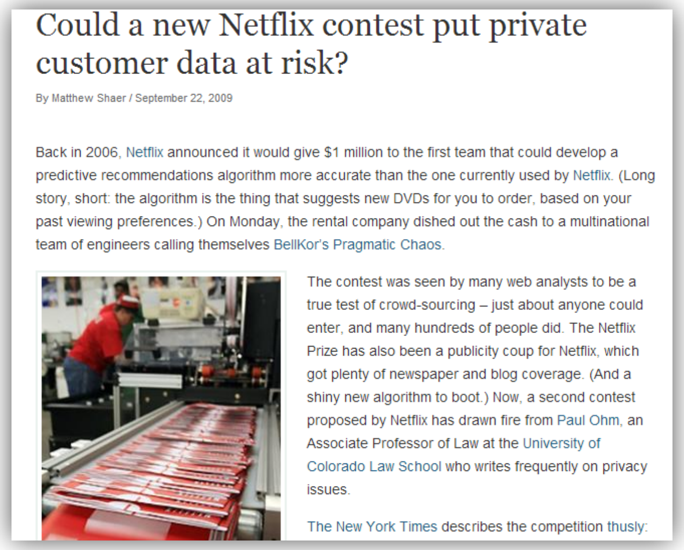
Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.



Could a new Netflix contest put private customer data at risk?

By Matthew Shaer / September 22, 2009

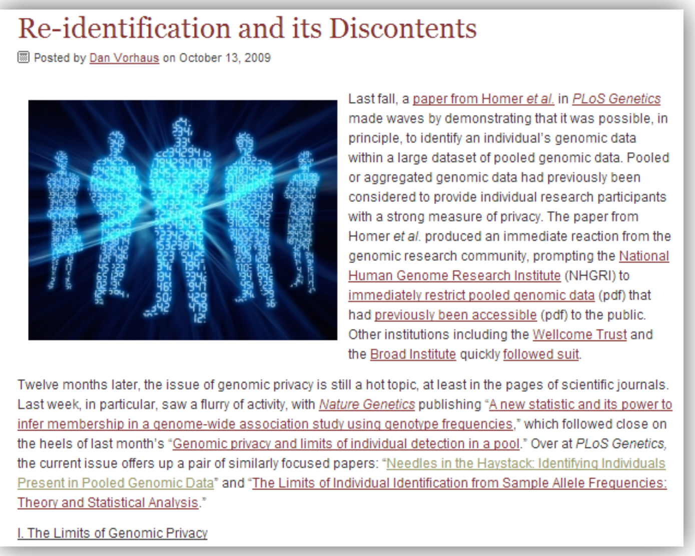
Back in 2006, Netflix announced it would give $1 million to the first team that could develop a predictive recommendations algorithm more accurate than the one currently used by Netflix. (Long story, short: the algorithm is the thing that suggests new DVDs for you to order, based on your past viewing preferences.) On Monday, the rental company dished out the cash to a multinational team of engineers calling themselves BellKor's Pragmatic Chaos.

The contest was seen by many web analysts to be a true test of crowd-sourcing – just about anyone could enter, and many hundreds of people did. The Netflix Prize has also been a publicity coup for Netflix, which got plenty of newspaper and blog coverage. (And a shiny new algorithm to boot.) Now, a second contest proposed by Netflix has drawn fire from Paul Ohm, an Associate Professor of Law at the University of Colorado Law School who writes frequently on privacy issues.

The New York Times describes the competition thusly:

But what is "privacy" not?

- Privacy is not releasing only "aggregate" statistics.



Re-identification and its Discontents

Posted by Dan Vorhaus on October 13, 2009

Last fall, a paper from Homer *et al.* in *PLoS Genetics* made waves by demonstrating that it was possible, in principle, to identify an individual's genomic data within a large dataset of pooled genomic data. Pooled or aggregated genomic data had previously been considered to provide individual research participants with a strong measure of privacy. The paper from Homer *et al.* produced an immediate reaction from the genomic research community, prompting the National Human Genome Research Institute (NHGRI) to immediately restrict pooled genomic data (pdf) that had previously been accessible (pdf) to the public. Other institutions including the Wellcome Trust and the Broad Institute quickly followed suit.

Twelve months later, the issue of genomic privacy is still a hot topic, at least in the pages of scientific journals. Last week, in particular, saw a flurry of activity, with *Nature Genetics* publishing "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies," which followed close on the heels of last month's "Genomic privacy and limits of individual detection in a pool." Over at *PLoS Genetics,* the current issue offers up a pair of similarly focused papers: "Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data" and "The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis."

I. The Limits of Genomic Privacy

So what is privacy?

- Idea: Privacy is about promising people freedom from harm.
  - Attempt 1: “*An analysis of a dataset D is private if the data analyst knows no more about Alice after the analysis than he knew about Alice before the analysis.*”

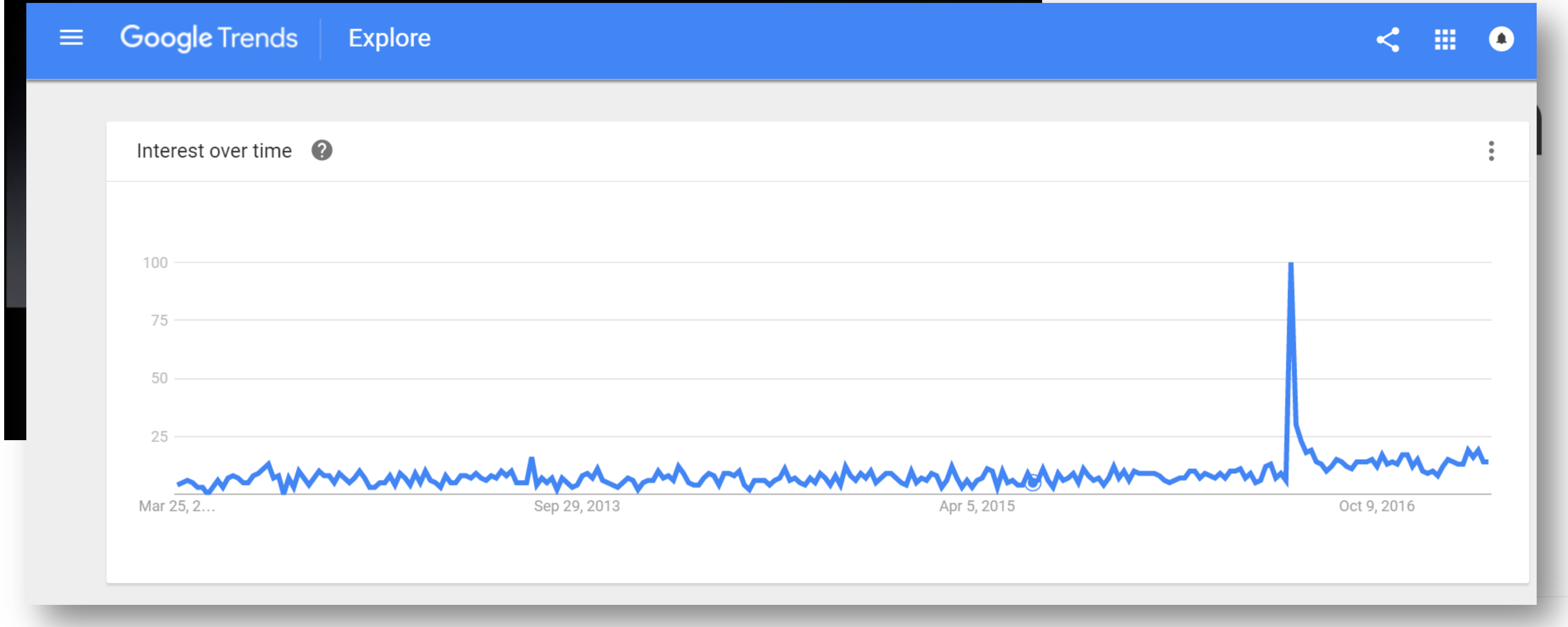
So what is privacy?

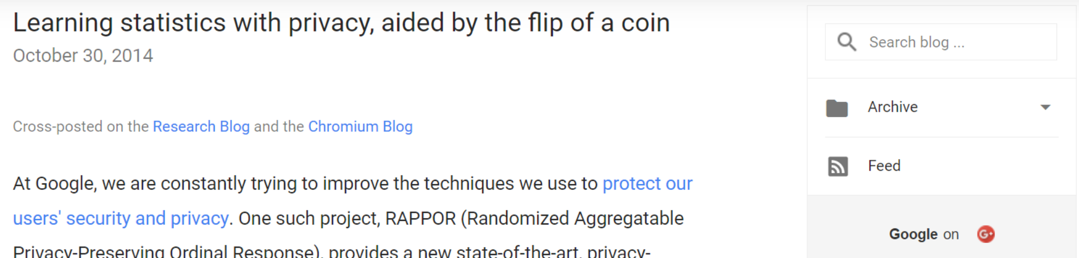
- Problem: Impossible to achieve with auxiliary information.
  - Suppose an insurance company knows that Alice is a smoker.
  - An analysis that reveals that smoking and lung cancer are correlated might cause them to raise her rates!
- Was her privacy violated?
  - This is exactly the sort of information we want to be able to learn…
  - This is a problem *even if Alice was not in the database!*

So what is privacy?

- Idea: Privacy is about promising people freedom from harm.
  - Attempt 2: "*An analysis of a dataset D is private if the data analyst knows almost no more about Alice after the analysis than he would have known had he conducted the same analysis on an identical database with Alice's data removed*."

So What is Differential Privacy?



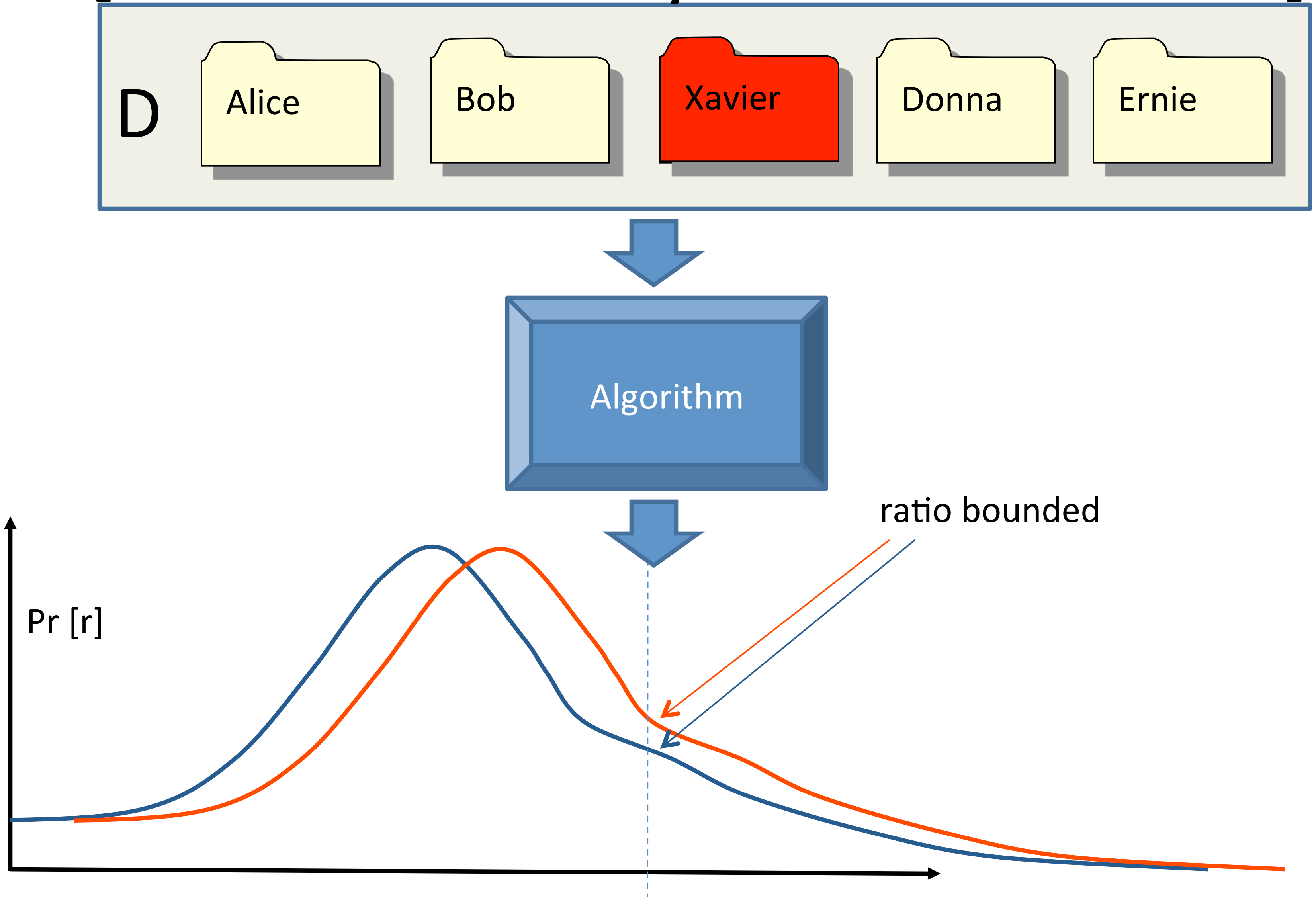

Learning statistics with privacy, aided by the flip of a coin

October 30, 2014

Cross-posted on the Research Blog and the Chromium Blog

At Google, we are constantly trying to improve the techniques we use to protect our users' security and privacy. One such project, RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), provides a new state-of-the-art, privacy-

Differential Privacy [Dwork-McSherry-Nissim-Smith 06]



Differential Privacy

$\mathcal{X}$: The data *universe*.

$D \subset \mathcal{X}$: The dataset (one element per person)

Definition: Two datasets $D, D' \subset \mathcal{X}$ are *neighbors* if they differ in the data of a single individual.

Differential Privacy

X: The data *universe*.

$D \subset X$: The dataset (one element per person)

Definition: An algorithm M is ϵ-differentially private if for all pairs of neighboring datasets D, D', and for all outputs x:

$$\Pr[M(D)=x] \leq (1+\epsilon)\Pr[M(D')=x]$$

Some Useful Properties

Theorem (Postprocessing): If $M(D)$ is ϵ-private, and f is any (randomized) function, then $f(M(D))$ is ϵ-private.

So…

Definition: An algorithm M is ϵ-differentially private if for all pairs of neighboring datasets D, D', and for all outputs x:

$$\Pr[M(D)=x] \leq (1+\epsilon)\Pr[M(D')=x]$$

$x=$

So…

Definition: An algorithm M is ϵ-differentially private if for all pairs of neighboring datasets D, D', and for all outputs x:

$$\Pr[M(D)=x] \le (1+\epsilon)\Pr[M(D')=x]$$

$x=$

CLAIM
DENIED

So…

Definition: An algorithm M is ϵ-differentially private if for all pairs of neighboring datasets D, D', and for all outputs x:

$$\Pr[M(D)=x] \leq (1+\epsilon)\Pr[M(D')=x]$$

$x=$

Some Useful Properties

Theorem (Composition): If $M\downarrow 1$,..., $M\downarrow k$

are ϵ-private, then:

$M(D) \equiv (M\downarrow 1\ (D),\ldots,M\downarrow k\ (D))$

is $k\epsilon$-private.

So…

You can go about designing algorithms as you normally would. Just access the data using differentially private "subroutines", and keep track of your "privacy budget" as a resource.

Private algorithm design, like regular algorithm design, can be modular.

Some simple operations: Answering Numeric Queries

Def: A numeric function f has sensitivity c if for all neighboring D, D' :

$$|f(D) - f(D')| \leq c$$

Write $s(f) \equiv c$

- e.g. “How many professors are in the building?” has sensitivity 1.
- “What fraction of people in the building are professors?” has sensitivity $1/n$.

Some simple operations: Answering Numeric Queries

The Laplace Mechanism:

$M{\downarrow}Lap\ (D,f,\epsilon)=f(D)+Lap(s(f)/\epsilon\ )$

Theorem: $M{\downarrow}Lap\ (\cdot,f,\epsilon)$ is ϵ-private.

Some simple operations: Answering Numeric Queries

The Laplace Mechanism:

$M{\downarrow}Lap\ (D,f,\epsilon)=f(D)+Lap(s(f)/\epsilon\ )$

Theorem: The expected error is $s(f)/\epsilon$

(can answer “what fraction of people in the building are professors?” with error 0.2%)

Some simple operations: Answering Non-numeric Queries

“What is the modal eye color in the room?”

$R=\{Blue, Green, Brown, Red\}$

- If you can define a function that determines how “good” each outcome is for a fixed input:
  - E.g.

$q(D, Red)=$“fraction of people in D with red eyes”

Some simple operations: Answering Non-numeric Queries

$M_{Exp}(D,R,q,\epsilon)$:

Output $r\in R$ w.p. $\propto e^{2\epsilon\cdot q(D,r)}$

Theorem: $M_{Exp}(D,R,q,\epsilon)$ is $s(q)\cdot\epsilon$-private, and outputs $r\in R$ such that:

$$E\left[\left|q(D,r)-\max_{r^*\in R} q(D,r^*)\right|\right]\le 2s(q)/\epsilon\cdot\ln|R|$$

(can find a color that has frequency within 0.5% of the modal color in the building)

So what can we do with that?

Empirical Risk Minimization:

*i.e. almost all of supervised learning

Find θ to minimize:

$$L(\theta)=\sum_{i=1}^{n} \ell(\theta, (x_i, y_i))$$

Stochastic Gradient Descent

```
Let $\theta^1 = 0^d$
For $t=1$ to $T$:
    Pick $i$ at random. Let $g_t \leftarrow \nabla\ell(\theta^t,(x_i,y_i))$
    Let $\theta^{t+1} \leftarrow \theta^t - \eta\cdot g_t$
```

Convergence depends on the fact that at each round: $\mathbb{E}[g_t] = \nabla \mathrm{L}(\theta)$

Private Stochastic Gradient Descent

```
Let $\theta^1 = 0^d$
For $t=1$ to $T$:
    Pick $i$ at random. Let $g_t \leftarrow \nabla\ell(\theta^t,(x_i,y_i)) + Lap(\sigma)^d$
    Let $\theta^{t+1} \leftarrow \theta^t - \eta\cdot g_t$
```

Still have: $\mathbb{E}[g_t] = \nabla \mathrm{L}(\theta)$!

(Can still prove convergence theorems, and run the algorithm…)

Privacy guarantees can be computed from:

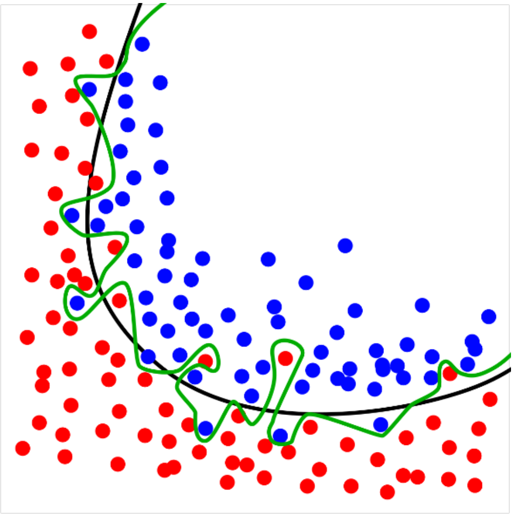
1) The privacy of the Laplace mechanism
2) Preservation of privacy under post-processing, and
3) Composition of privacy guarantees.

What else can we do?

- Statistical Estimation
- Graph Analysis
- Combinatorial Optimization
- Spectral Analysis of Matrices
- Anomaly Detection/Analysis of Data Streams
- Convex Optimization
- Equilibrium computation
- Computation of optimal 1-sided and 2-sided matchings
- Pareto Optimal Exchanges
- …

Differential Privacy ⇒ Learning

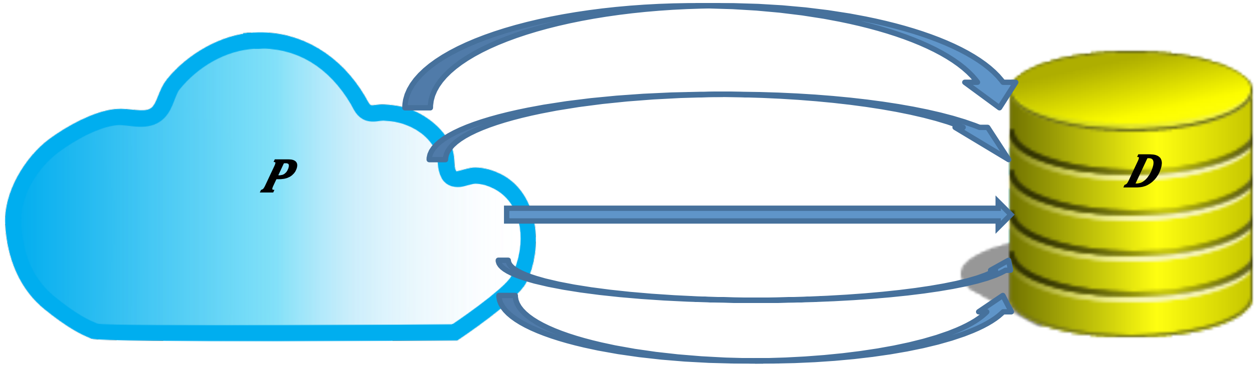
Theorem*: An ϵ-differentially private algorithm cannot overfit its training set by more than ϵ.



*Lots of interesting details missing!

Choosing a Formalism: Statistical Queries

- A data universe X
- A distribution $P \in \Delta X$
- A dataset $D \subseteq X$ consisting of n points $x \in X$ sampled i.i.d. from P.



Choosing a Formalism: Statistical Queries

- A *statistical query* is defined by a predicate

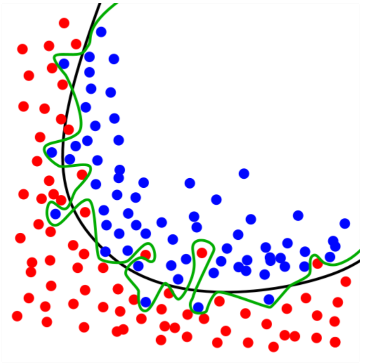
$$\phi : X \to [0,1].$$

- The value of a statistical query is

$$\phi(P)=E\downarrow x\sim P\,[\phi(x)]$$

- A statistical estimator is an algorithm for estimating statistical query: $A\downarrow D\,(\phi)\to[0,1]$

Choosing a Formalism: Statistical Queries



Loses little generality. Captures, e.g.

- Means, variances, correlations, etc.
- Risk of a hypothesis:

$$R(h) = \mathrm{E}{\downarrow}(x,y) \sim P\,[L(h(x),y)]]$$

- *Gradient* of risk of a hypothesis:

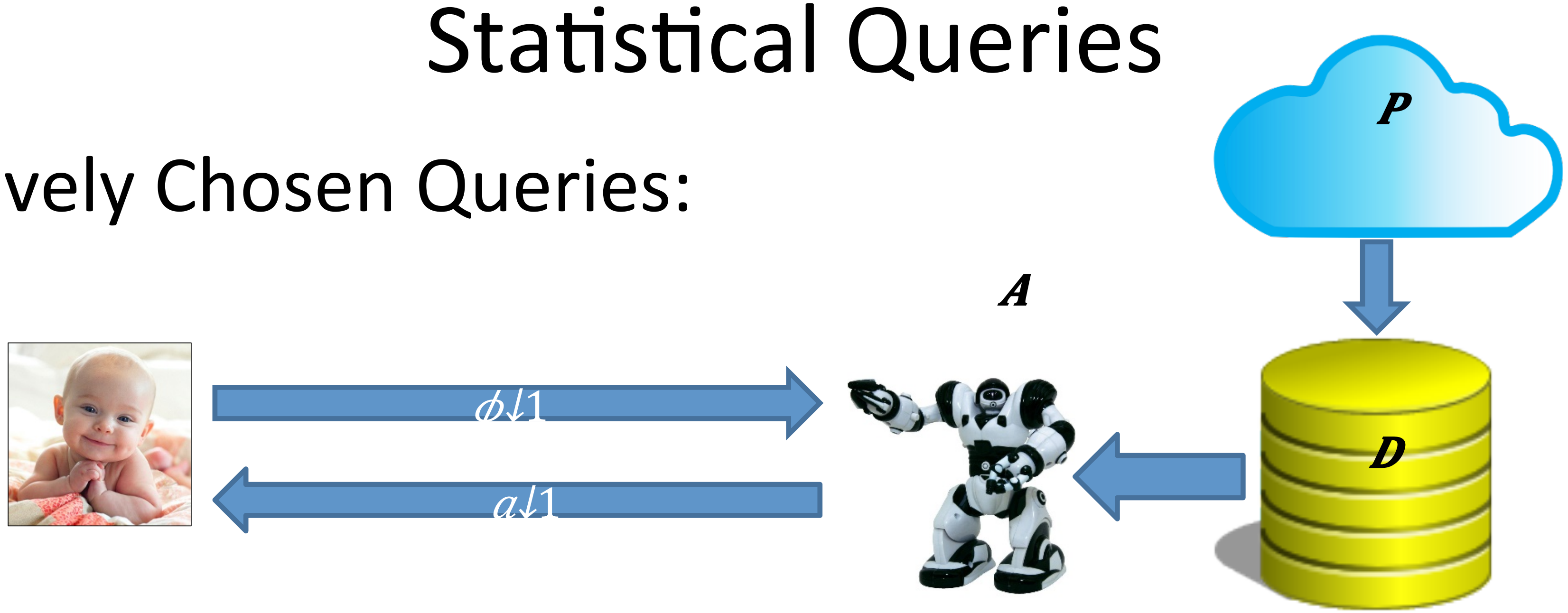
$$\nabla R(h) = \mathrm{E}{\downarrow}(x,y) \sim P\,[\nabla L(h(x),y)]]$$

- *Almost** all of PAC learning

**Except Parity functions*

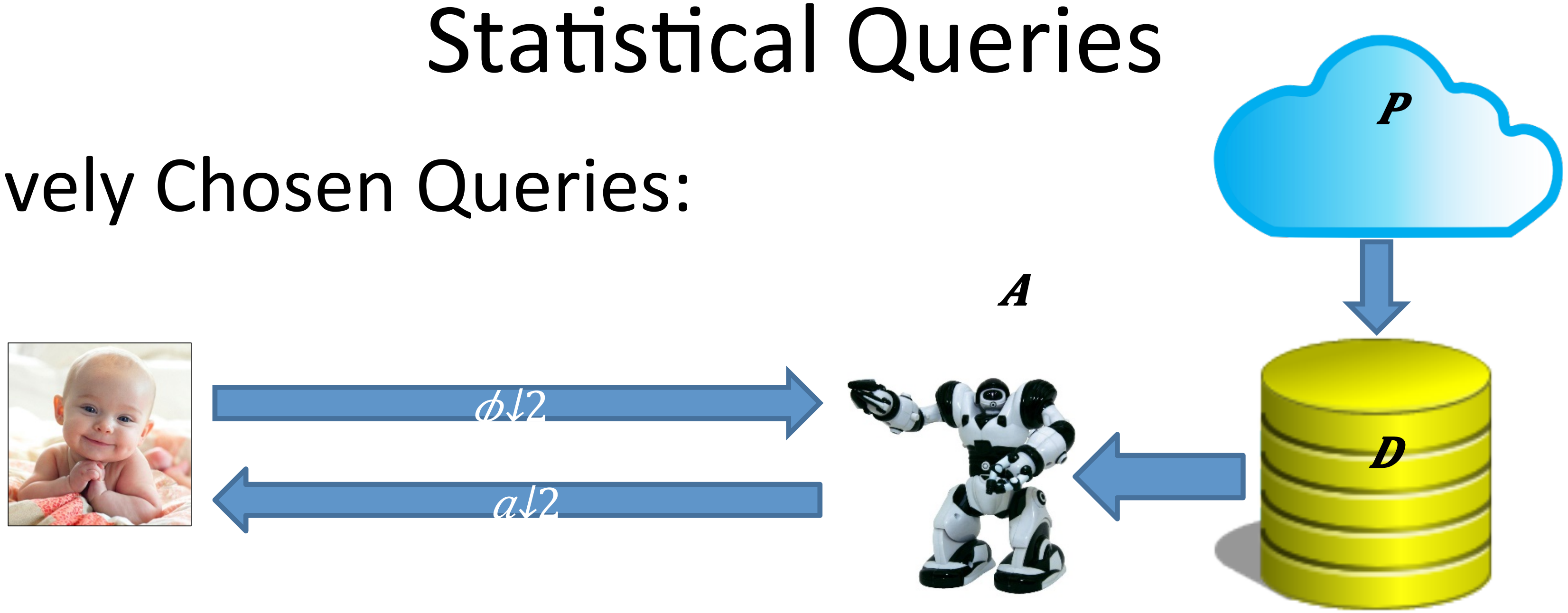
Choosing a Formalism: Statistical Queries

- Adaptively Chosen Queries:



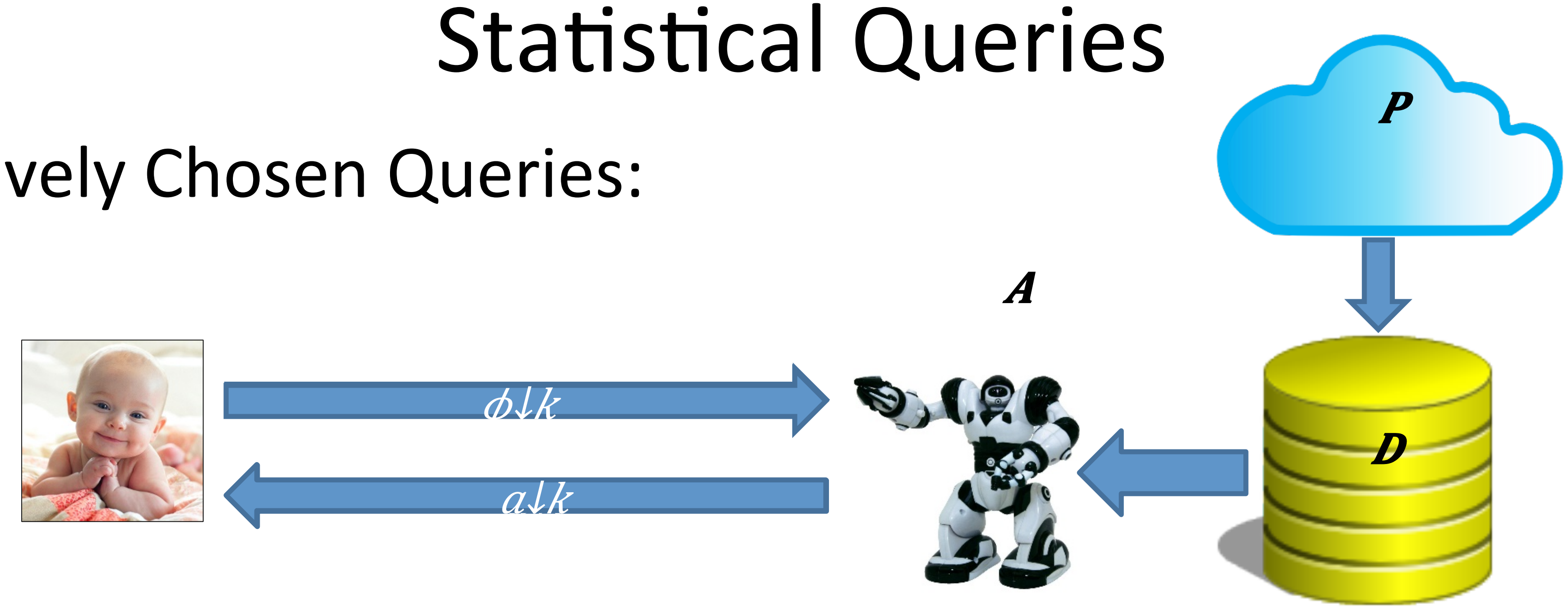
Choosing a Formalism: Statistical Queries

- Adaptively Chosen Queries:



Choosing a Formalism: Statistical Queries

- Adaptively Chosen Queries:

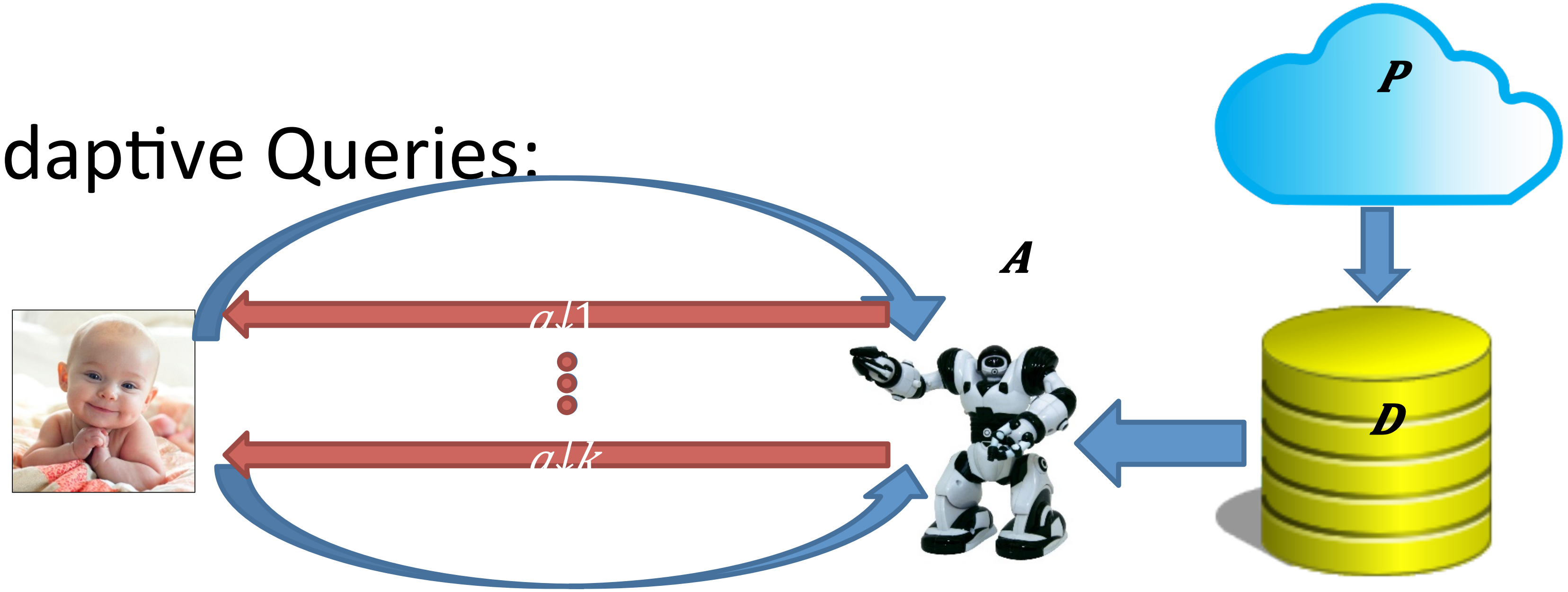


- A statistical estimator A is (ϵ,δ)-accurate for sequences of k adaptively chosen queries $\phi_1,\ldots,\phi_k$ if for and , with probability $1-\delta$:

$$\max_i |A_D(\phi_i) - \phi_i(P)| \le \epsilon.$$

A Baseline

- Non-Adaptive Queries:

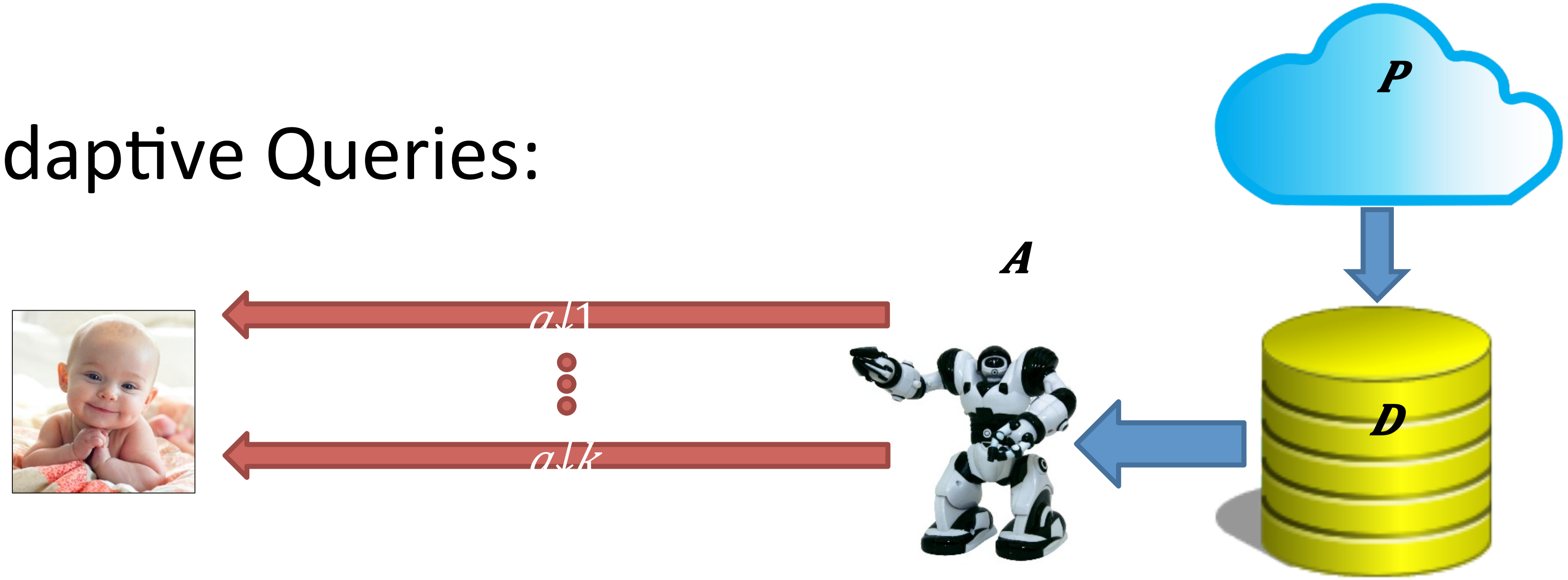


- The "empirical average mechanism": $A_D(\phi) = \phi(D) := \frac{1}{n}\sum_{x \in D} \phi(x)$ can answer k *non-adaptive* queries with $(0.01, 0.01)$-accuracy where:

$$k = e^{\Theta(n)}$$

A Baseline

- Non-Adaptive Queries:



- The “empirical average mechanism”: $A_D(\phi)=\phi(D):=\frac{1}{n}\sum_{x\in D}\phi(x)$ can answer k *adaptive* queries with $(0.01,0.01)$-accuracy where:

$k=0(n)$

Differential Privacy ⇒ Learning

Theorem: [DFHPRR’15,BNSSSU’16]:

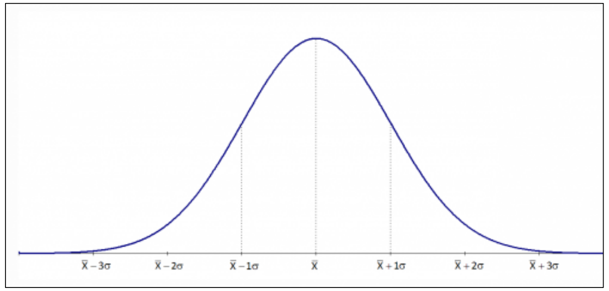
Let A be a statistical estimator for adaptively chosen statistical queries. Let P be any distribution, and let $D \sim P \uparrow n$. If:

1. A is $(\epsilon, \epsilon \cdot \delta)$-differentially private, and

2. A is $(\epsilon, \epsilon \cdot \delta)$-accurate *with respect to the sample* D, then:

A is $(\mathrm{O}(\epsilon), O(\delta))$-accurate *with respect to the distribution* P.

Applications



Using Independent Gaussian Perturbation

Theorem: There exists a simple, computationally efficient statistical estimator that can answer k *adaptive* queries to non-trivial accuracy where:

$$k=\Theta\ (n\uparrow 2\ )$$

A *quadratic* improvement over the empirical average mechanism!

Applications

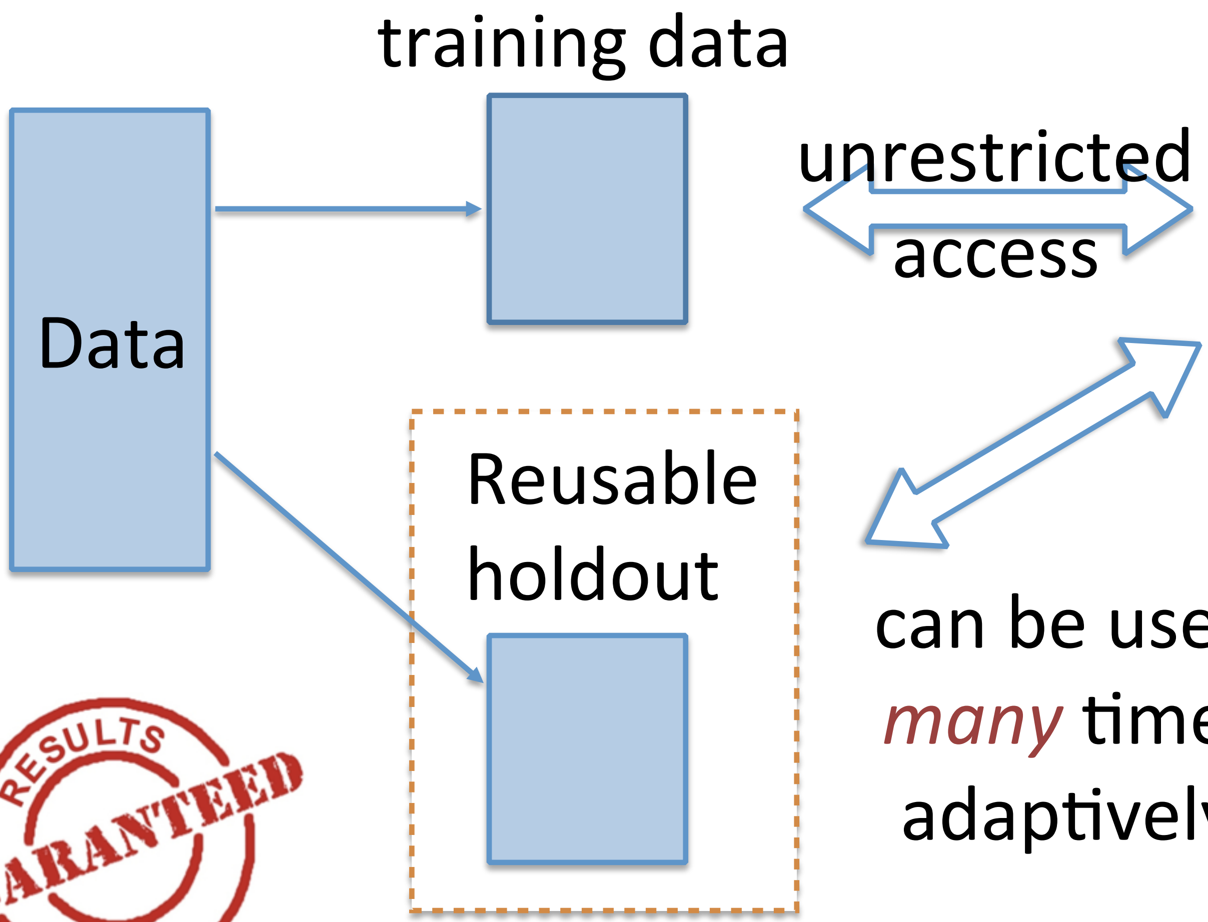
Using State of the Art Differentially Private Mechanisms

Theorem: There exists a statistical estimator that can answer k *adaptive* queries to non-trivial accuracy where:

$$k = e^{\Theta(n/\log|X|)}$$

An *exponential* improvement if the data universe X is finite and $n \gg \log|X|$.

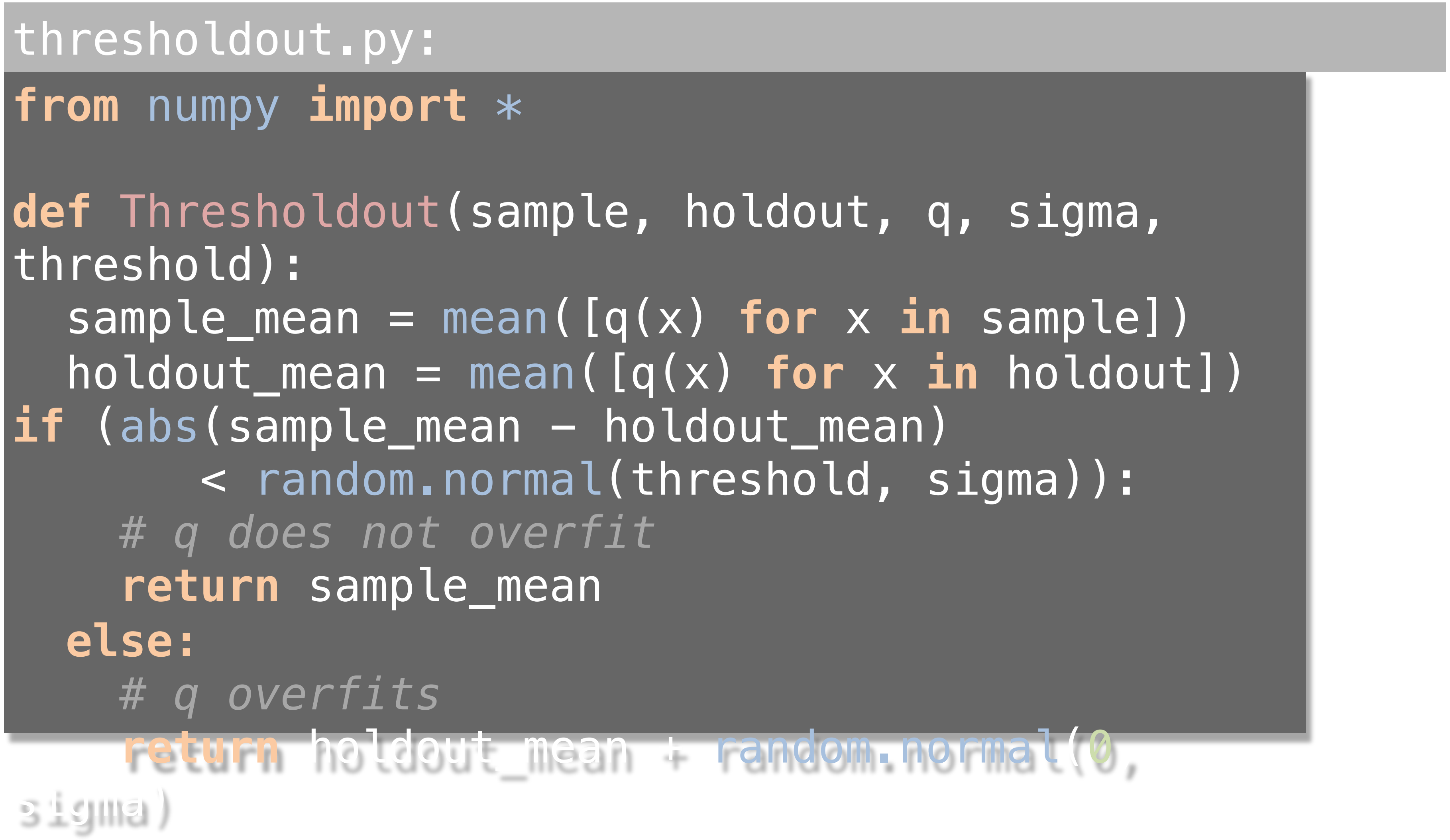
Applications



RESULTS GUARANTEED

valid estimate every time you use the holdout

Thresholdout [DFHPRR15]



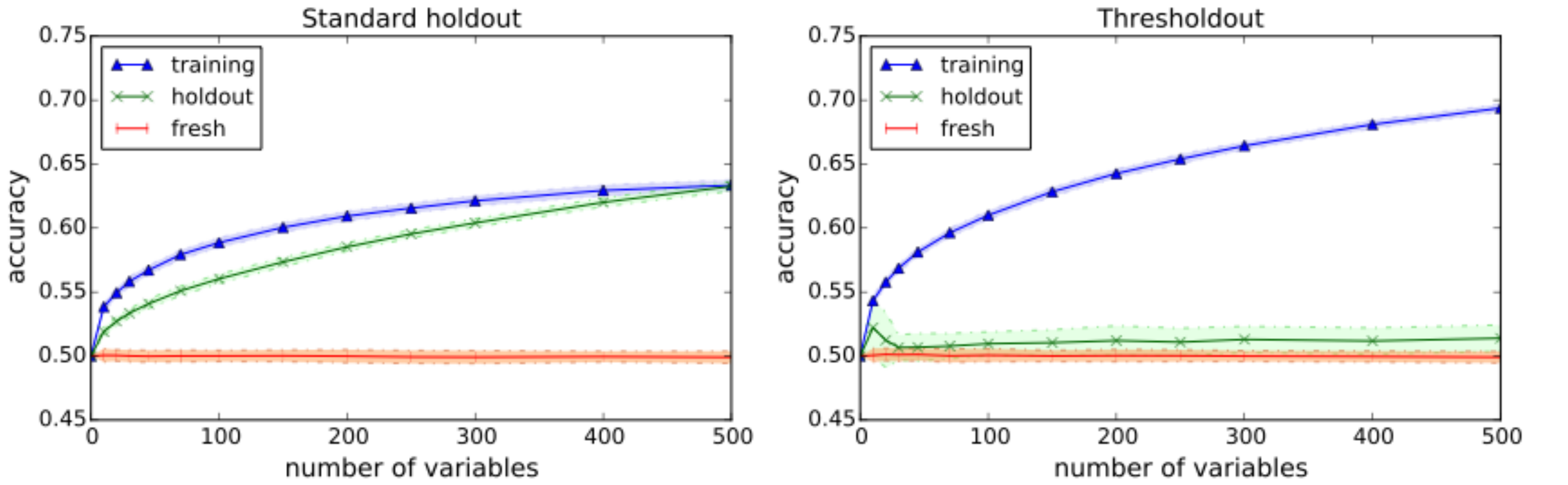
```
thresholdout.py:
from numpy import *

def Thresholdout(sample, holdout, q, sigma,
threshold):
  sample_mean = mean([q(x) for x in sample])
  holdout_mean = mean([q(x) for x in holdout])
if (abs(sample_mean - holdout_mean)
       < random.normal(threshold, sigma)):
    # q does not overfit
    return sample_mean
  else:
    # q overfits
    return holdout_mean + random.normal(0,
sigma)
```

Reusable holdout example

- Data set with $2n = 20{,}000$ rows and $d = 10{,}000$ variables. Class labels in $\{-1,1\}$
- Analyst performs **stepwise variable selection:**
  1. Split data into training/holdout of size n
  2. Select “best” k variables on training data
  3. Only use variables also good on holdout
  4. Build linear predictor out of k variables
  5. Find best $k = 10,20,30,\ldots$

Classification after feature selection

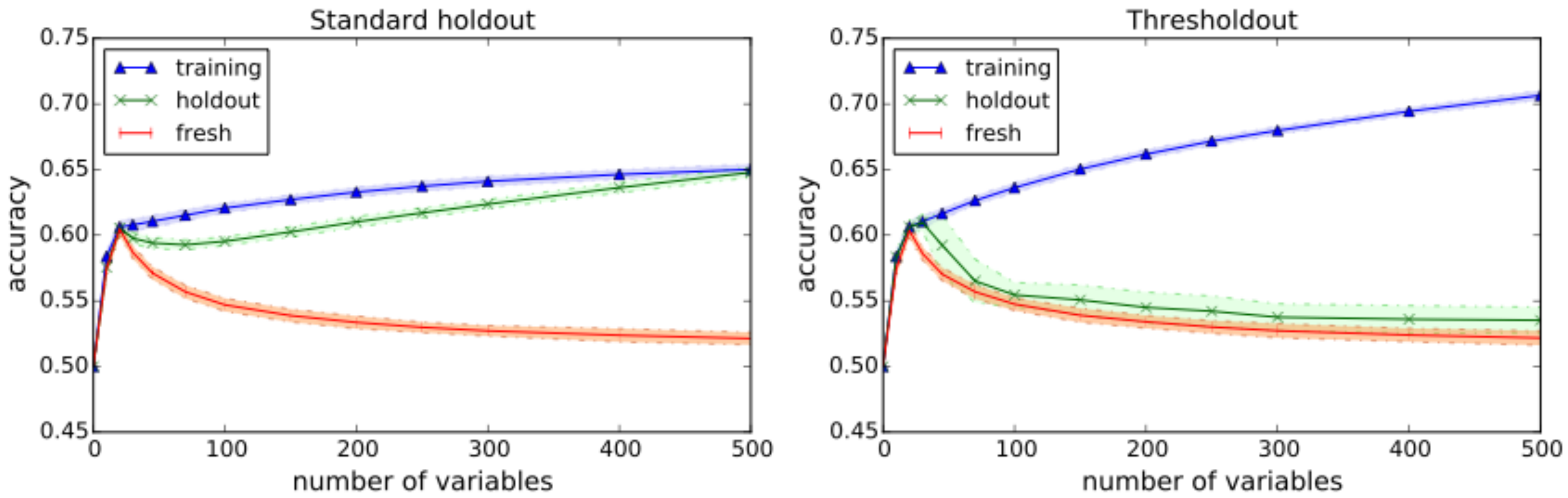
No signal: data are random gaussians
labels are drawn *independently* at random from {-1,1}



Thresholdout correctly detects overfitting!

Classification after feature selection

Strong signal: 20 features are mildly correlated with target
remaining attributes are uncorrelated



Thresholdout correctly detects right model size!

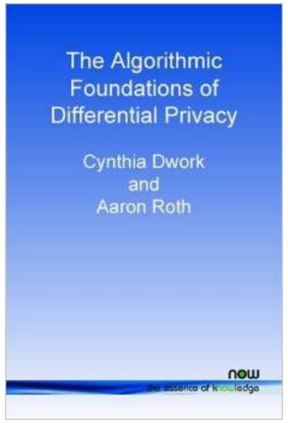
So…

- Differential privacy provides:
  - A rigorous, provable guarantee with a strong privacy semantics.
  - A set of tools and composition theorems that allow for modular, easy design of privacy preserving algorithms.
  - *Protection against overfitting* even when privacy is not a concern.

Thanks!

To learn more:

- Our textbook on differential privacy:



The Algorithmic
Foundations of
Differential Privacy

Cynthia Dwork
and
Aaron Roth

now
the essence of knowledge

  - Available for free on my website: http://www.cis.upenn.edu/~aaroth
- Connections between Privacy and Overfitting:
  - Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, *"The Reusable Holdout: Preserving Validity in Adaptive Data Analysis"*, Science, August 7 2015.
  - Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, *"Preserving Statistical Validity in Adaptive Data Analysis"*, STOC 2015.
  - Bassily, Nissim, Stemmer, Smith, Steinke, Ullman, *"Algorithmic Stability for Adaptive Data Analysis",* STOC 2016.
  - Rogers, Roth, Smith, Thakkar, *"Max Information, Differential Privacy, and Post-Selection Hypothesis Testing"*, FOCS 2016.
  - Cummings, Ligett, Nissim, Roth, Wu, *"Adaptive Learning with Robust Generalization Guarantees",* COLT 2016.