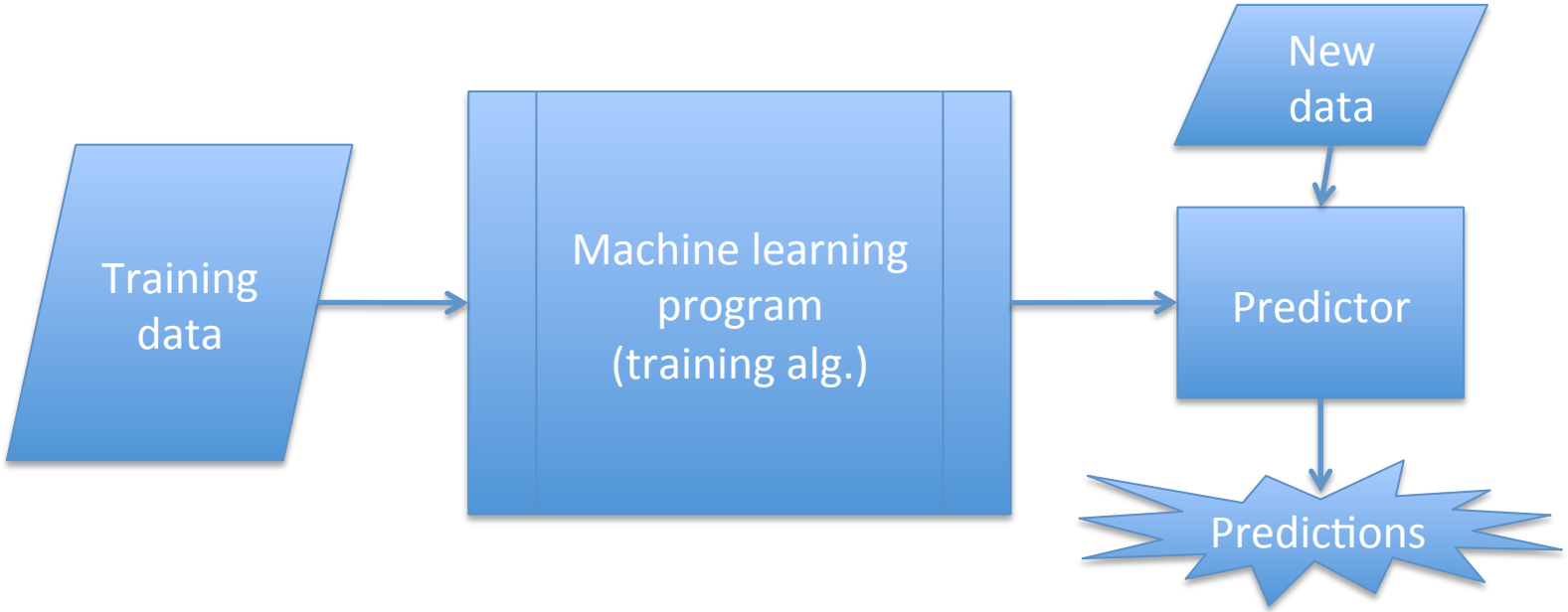
Machine learning and privacy

Daniel Hsu
Columbia University, Department of Computer Science

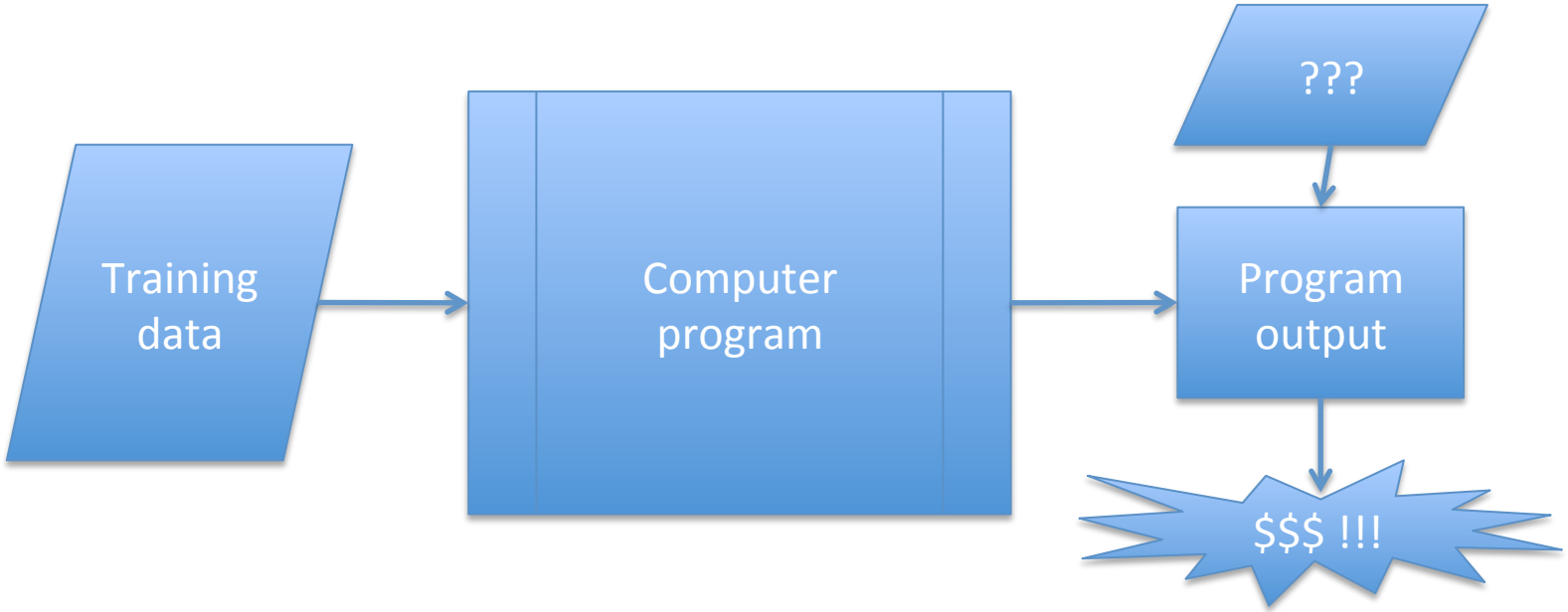
Machine learning

- Learning from data to make accurate predictions in the future.



Data science

- Interesting ways of using data, often with computers, to amaze and delight!



Sensitive training data

- Often, training data is comprised of sensitive information about individuals.

	Sex	Age	BT	BPs	BPd	…
Alice	1	25	0	100	70	…
Bob	0	32	2	140	90	…
Charlie	0	65	1	90	60	…
…	…	…	…	…	…	…

Private database


obamaischeckingyouremail.tumblr.com

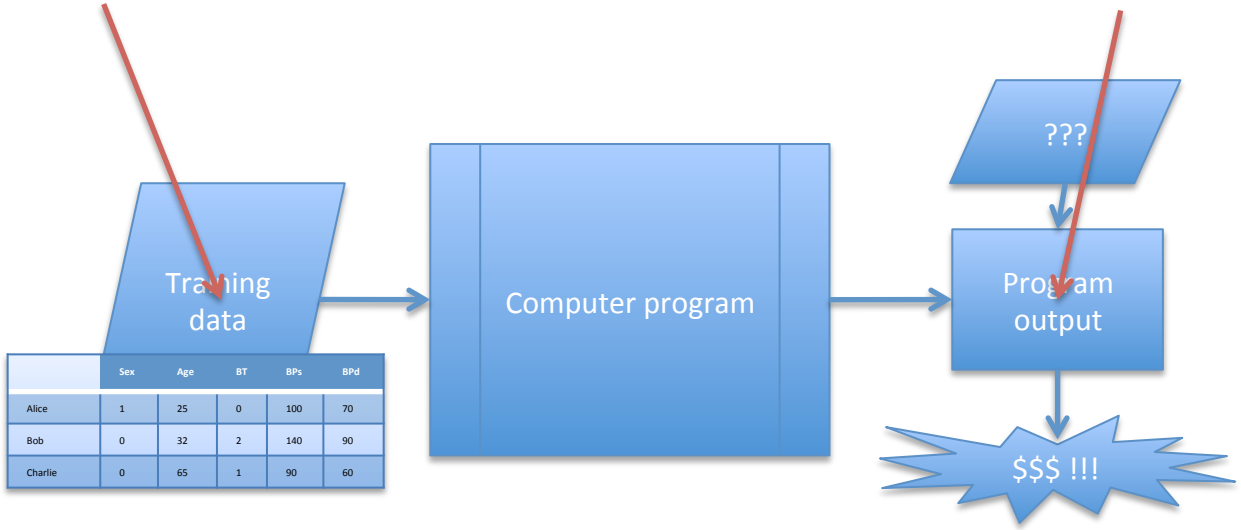

obamaischeckingyouremail.tumblr.com

machine learning + sensitive information = ?

Using sensitive training data

Sensitive information,
want to keep private

f(training data),
publicly released or deployed



Examples: classify e-mails as spam/ham, detect fraud in credit card transactions, predict disease susceptibility from DNA, *etc*.

Want to transform private information into public good!

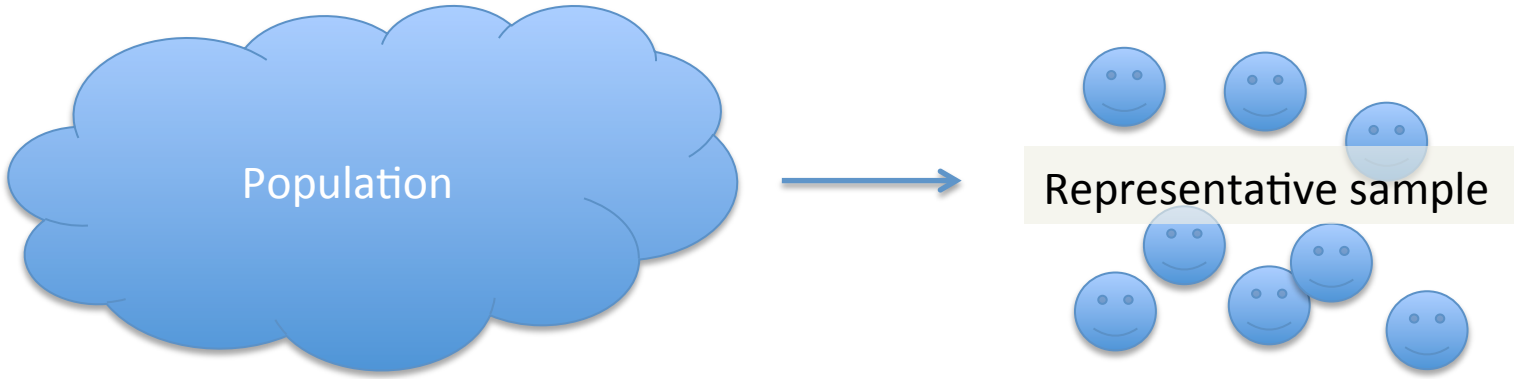
Outline for rest of talk

1. Is machine learning compatible with privacy?
2. Privacy expectations
3. Limits of privacy-preserving statistics and ML
4. Concluding remarks

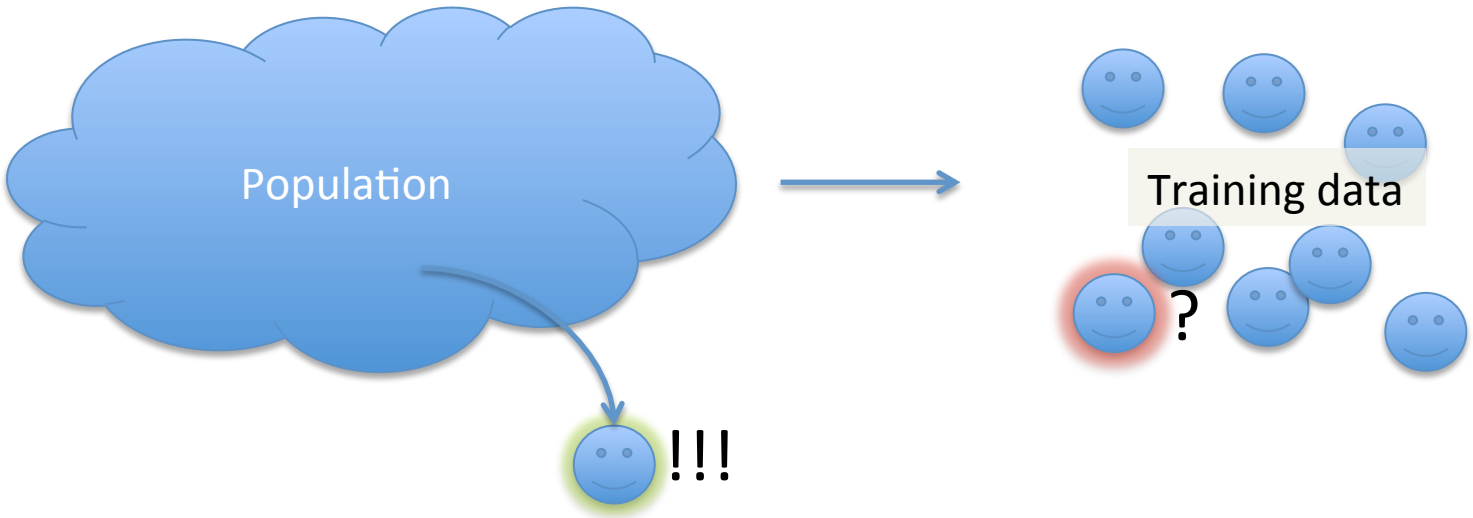
1. IS MACHINE LEARNING COMPATIBLE WITH PRIVACY?

Populations and samples

- **Goal of machine learning**: learn predictive characteristics about a population.
  - Don't have access to data for entire population, so use a representative sample ("training data").
  - Privacy risk is with respect to individuals represented in the sample.



Populations and samples



Generalization: output of learning f (training data) is predictive w.r.t. random individual from population.

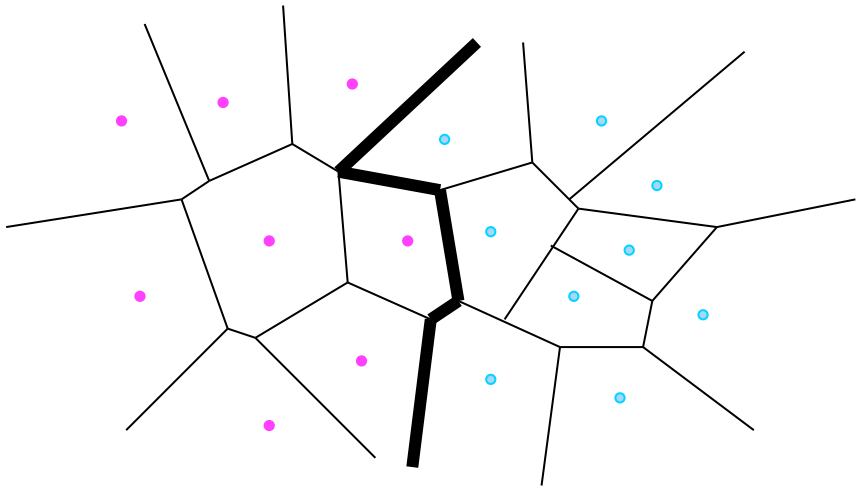
Need not be predictive w.r.t. any particular individual in training sample.

Privacy violations abound

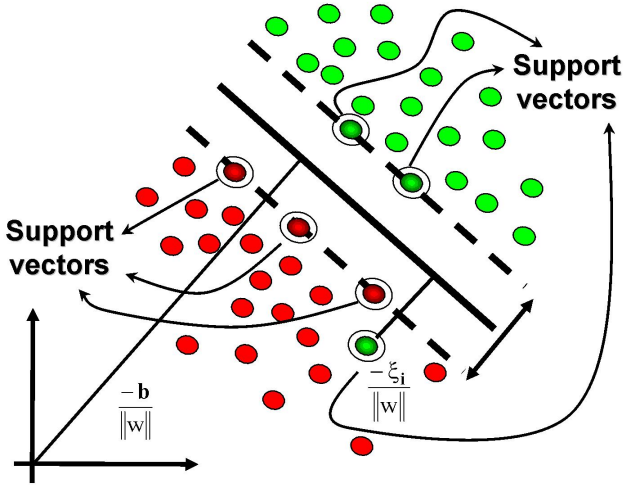
Unfortunately, many standard applications of machine learning and statistics will, by default, compromise the privacy of individuals represented in the sample.

Obvious privacy violations

- If output *includes* training data points…



Nearest neighbor classifier



Kernel-based SVM

Output reveals sensitive information of some (or all) individuals in the training data.

Subtle privacy violations

- If output is sensitive to outliers…

Alice	0.2	0.7	0.8	0.4	0.1	0.2
Bob	0.3	0.8	0.7	0.4	0.5	0.3
Charlie	0.5	0.2	0.3	**250**	0.7	**250**
Dave	0.6	0.1	0.9	0.5	0.4	0.4
Eve	0.3	0.0	0.5	0.3	0.5	0.5
Mean:	**0.38**	**0.36**	**0.64**	**50.32**	**0.44**	**50.28**

Can reveal information about outlying individuals (*e.g.*, uniquely identifying features).

Very subtle privacy violations

Cancer group

1.00										
.190	1.00									
.216	.251	1.00								
.186	.117	.047	1.00							
.154	.011	.170	.083	1.00						
.190	.140	.102	.095	.139	1.00					
.270	.215	.294	.248	.140	.141	1.00				
.101	.085	.170	.056	.234	.099	.175	1.00			
.239	.071	.163	.111	.161	.093	.199	.157	1.00		
.471	.117	.243	.094	.144	.123	.283	.216	.274	1.00	
.179	.202	.132	.094	.087	.159	.207	.108	.092	.294	1.00

Healthy group

1.00										
.141	1.00									
.099	.175	1.00								
.093	.199	.157	1.00							
.123	.283	.216	.274	1.00						
.159	.207	.108	.092	.294	1.00					
.088	.152	.075	.163	.156	.220	1.00				
.046	.161	.092	.072	.157	.143	.147	1.00			
.078	.392	.122	.229	.160	.172	.145	.177	1.00		
.045	.155	.135	.139	.110	.048	.126	.104	.169	1.00	
.178	.135	.102	.258	.314	.165	.147	.158	.131	.074	1.00

Published correlation statistics from GWAS

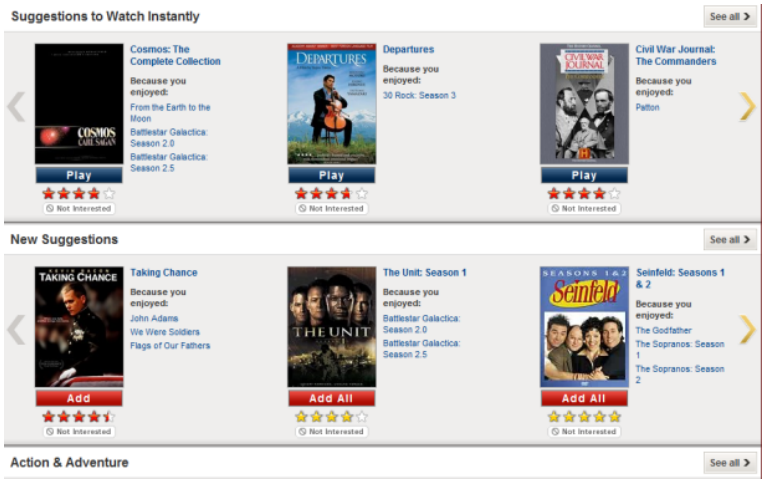
[Wang *et al*, 2009]: shows how published GWAS results reveals whether specific individuals from the study were in cancer group or healthy group.

Why is privacy-preserving ML hard?

- Attackers may have access to side-information
- Attackers may have compromised the training data (*e.g.*, know every row except yours)

In fact, machine learning tools have been used to compromise privacy!

Netflix challenge data set



- "Attackers" [Narayanan & Shmatikov, 2008] obtained access to public IMDb movie ratings.
- Partial information about a user suffices to reconstruct entire row in Netflix data set (all movie ratings).
- Machine learning technique: nearest-neighbor classifiers

2. PRIVACY EXPECTATIONS

What "privacy" can we expect?

- **Motivation**: "big data" is creating unexpected privacy harms
  - Releasing/deploying predictors trained on private DBs that predict sensitive attributes.
- Dalenius (1977): access to f(DB) should not change what an attacker can learn about an individual in the database.

(DB = training data)

Power of side-information

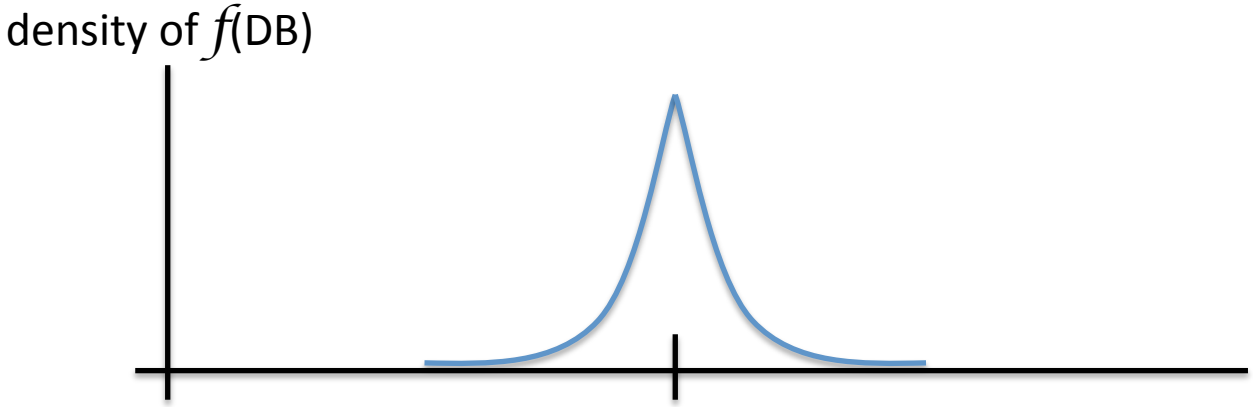
- Many sources of side-information already available to prospective attackers – this is not going to change.
  - e.g., data already available in public records, prior knowledge/beliefs
- Can generate privacy harm from "benign" information + side information
  - Suppose I know that Alice's salary is 2x the average salary in her department.
  - Learning about the average salary gives new information about Alice's salary.
- **Upshot**: Dalenius' criterion is not feasible.

Differential privacy

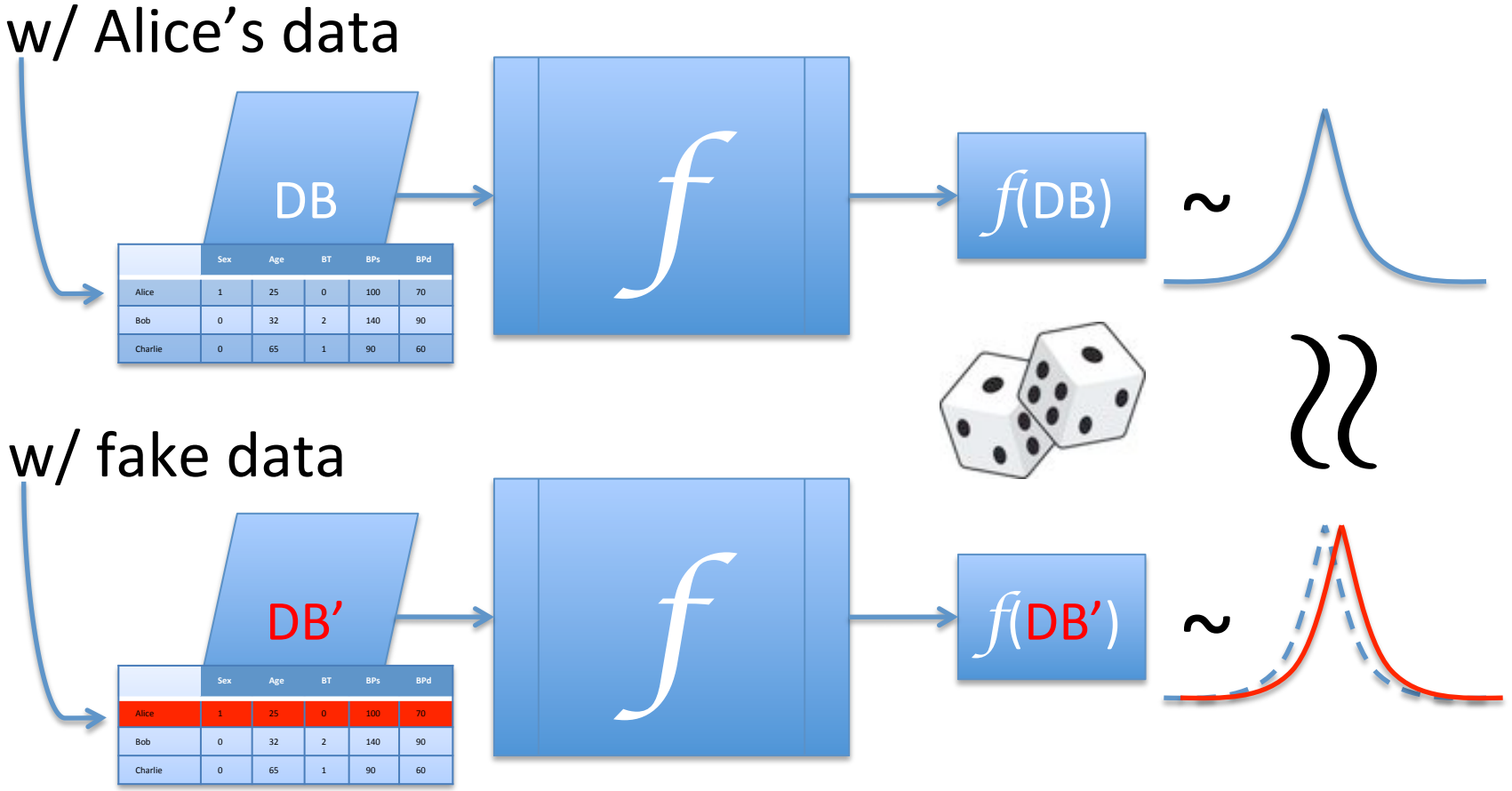
- [Dwork, McSherry, Nissim, & Smith, 2006]
- Similar to Dalenius' criterion, but feasible.
  - Attacker should not learn (much) more from f(DB) than from f(DB – {Alice}).
  - **Limit increase in potential "privacy harm"**.
  - How "much" is quantifiable.

Use of randomness

- **Critical idea**: *require* f to use randomization.
  - On any input DB, f(DB) is a random variable: specify a distribution over possible outputs, then randomly draw from it.



Differential privacy (schematic)



Differential privacy (definition)

- **Formal definition**:
  Say f guarantees ε-differential privacy if:
  for all possible databases DB and DB' differing in a single row,

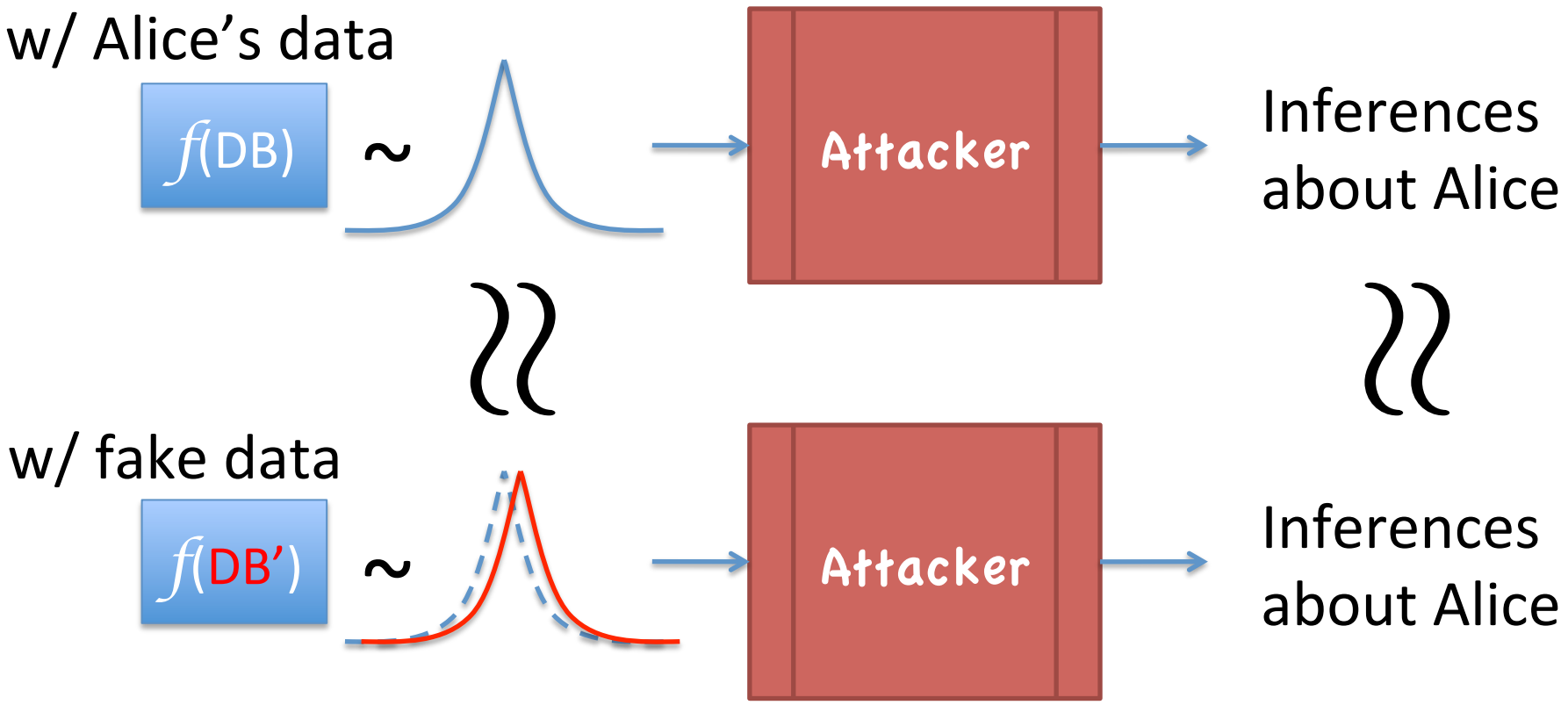
$$(1-\varepsilon) \leq \frac{\Pr[\, f(\mathrm{DB}) = t \,]}{\Pr[\, f(\mathrm{DB}') = t \,]} \leq (1+\varepsilon)$$

  for all outputs $t \in \mathrm{range}(f)$.

[Technically, use exp(-ε) and exp(ε).]

Attacker's perspective

If f guarantees differential privacy:



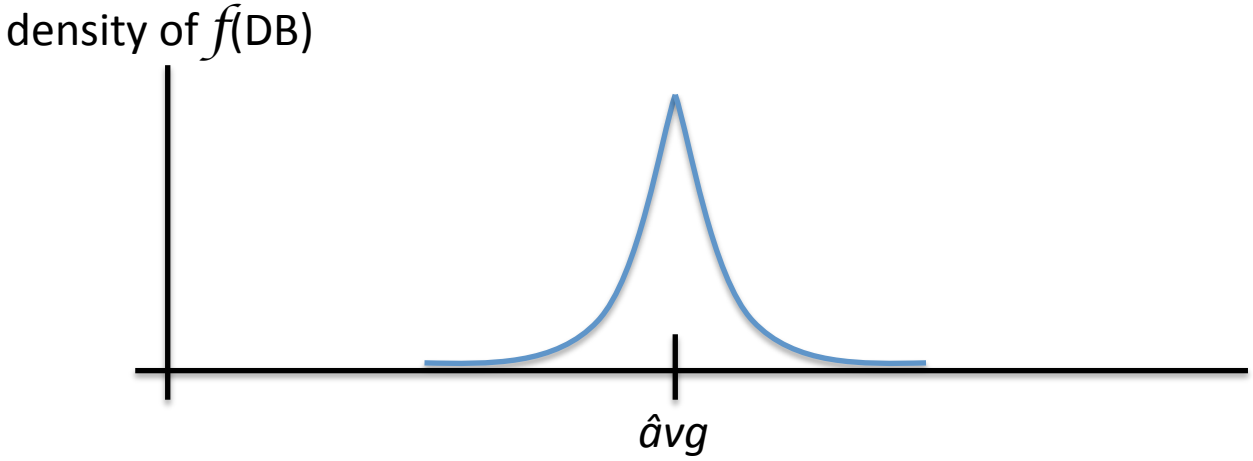
Salary example

- DB := { employee salaries in Alice's dept., each between $0 and $1M }.
- $\hat{avg}$:= average of salaries in DB.
- f(DB) := $\hat{avg}$ not differentially private, trivially because it is deterministic.

Adding random noise [DMNS,'06]

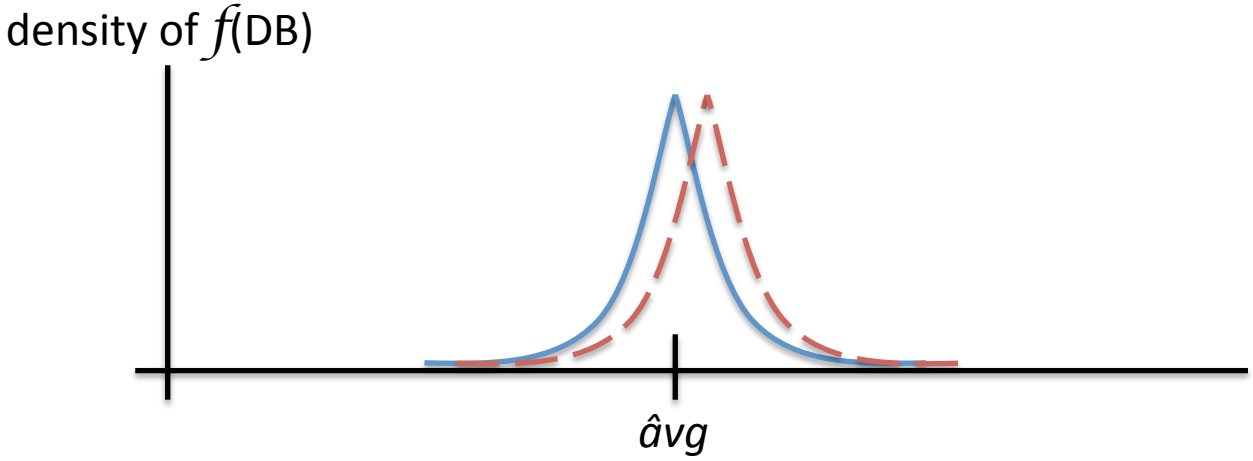
- What about $f(\text{DB}) := \hat{a}vg + \text{random noise}$?
  - Use noise w/ stddev $\sigma \approx \$1\text{M} / (\varepsilon n)$ ($n = |\text{DB}|$)
  - Convenient noise density ("Laplace dist."):

$$p(z) \propto \exp(-|z|/\sigma)$$



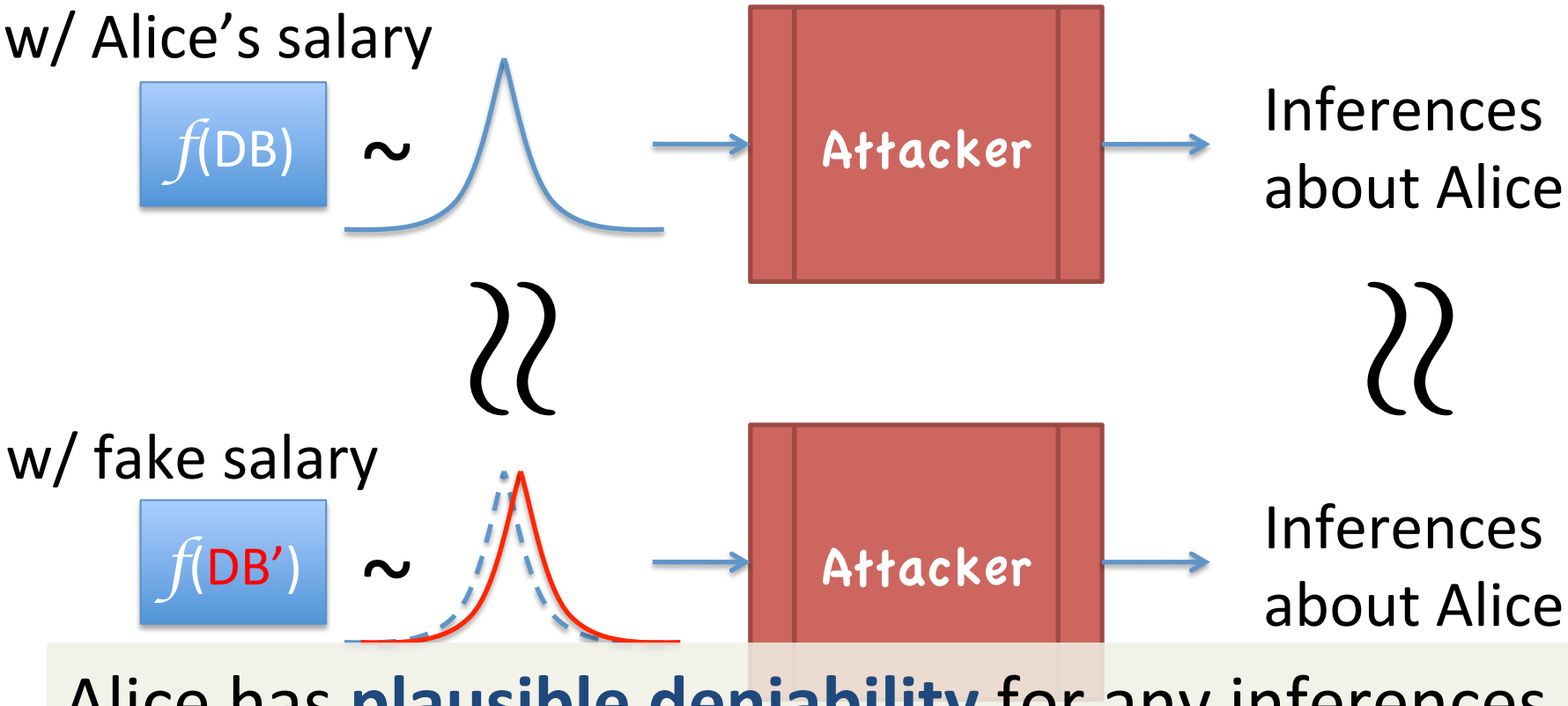
Adding random noise [DMNS,'06]

- This guarantees ε-differentially private!
  - Replacing Alice's salary by arbitrary value in range [0, 1M] can shift mean of f(DB) by ≤ \$1M/n.
  - Density value can change by factor ≤ exp(ε) ≈ 1+ε.



Alice's perspective

If f guarantees differential privacy:



Alice has **plausible deniability** for any inferences an attacker can make about her salary!

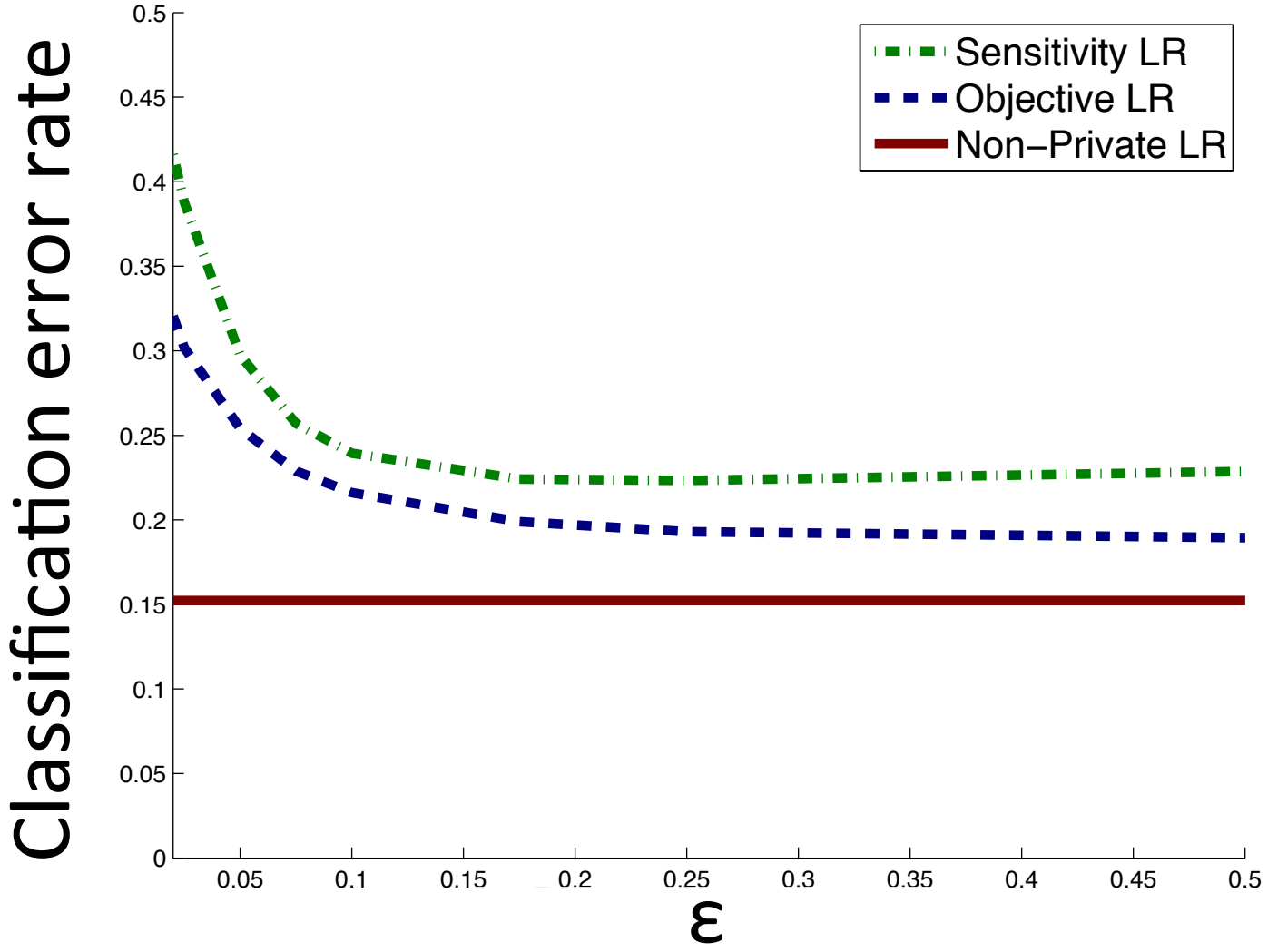
Statistical utility

- **Privacy vs. statistical utility trade-offs**:
  - Best for privacy: just return noise!
  - Best for utility: add no noise.
- If DB is itself a random sample, expect sampling error to be $O(1/\sqrt{n})$.
- Extra noise for privacy has stddev = $O(1/n)$.
  - Lower-order than sampling error!

Is privacy essentially free?

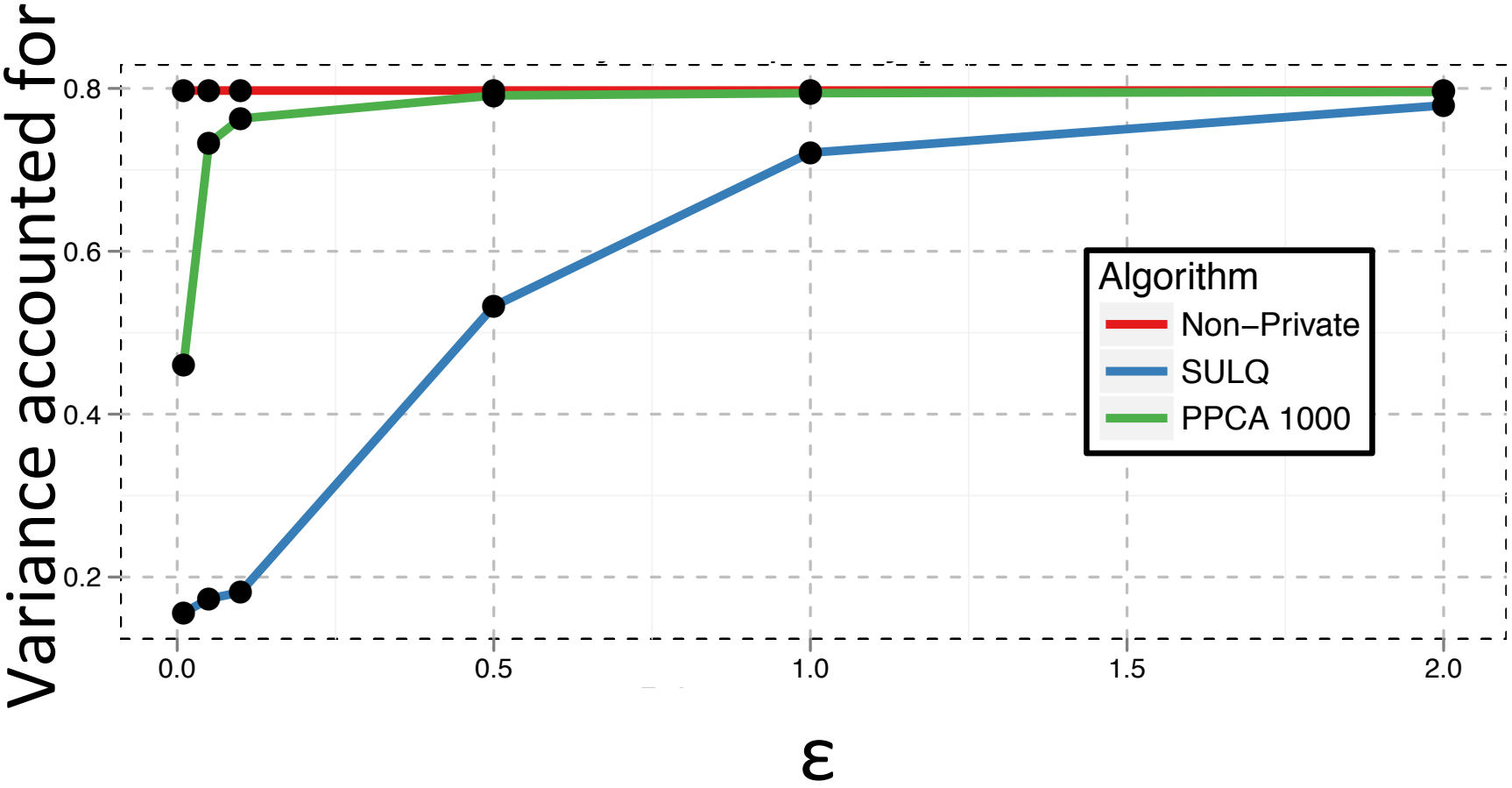
Estimating linear classifiers

[Chaudhuri, Monteleoni, & Sarwate '11]



Estimating principal components

[Chaudhuri, Sarwate, & Sinha '12]



3. LIMITS OF PRIVACY-PRESERVING STATISTICS & ML

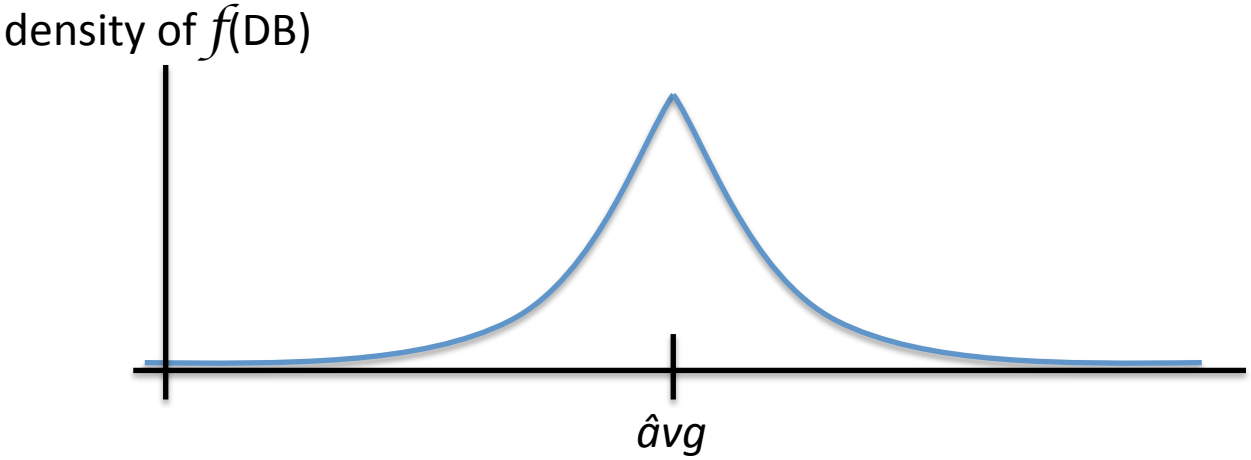
Unfortunately, many limits to what can be done privately

1. Statistical estimation
2. Computational learning
3. Data release

Revisiting the use of random noise

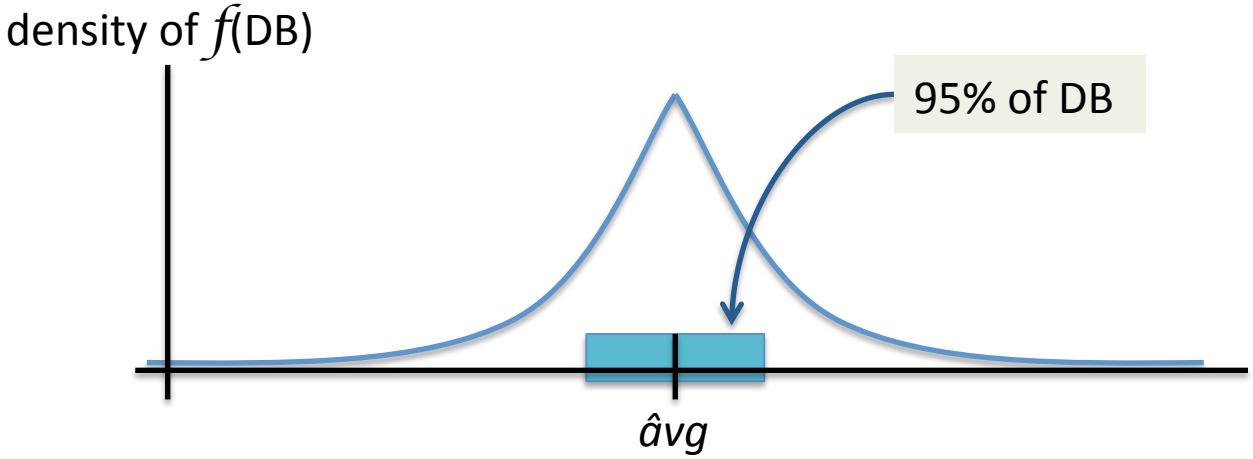
- Recall:
  - DB := { salaries, each between $0 and $1M };
    âvg := average of salaries in DB.;
    f(DB) := *âvg* + random noise.
    - Noise stddev σ ≈ $1M / (ε *n*)
    - Noise density: p(z) ∝ exp(-|z|/σ)

What if 95% of DB is ≤ $50K ?



Too much noise

- Noise dependence on range can wash out accuracy of original estimator.



- **Some recent progress**: "smooth sensitivity" method, but only for weaker privacy guarantee
  [Nissim, Raskhodnikova, & Smith, '07; Chaudhuri & H., '12]

Robustness is critical

- Many estimators (*e.g.*, sample mean) are not very robust.
- Rich literature on robust statistics provides quantification of estimator robustness:

Gross Error Sensitivity

Theorem [Chaudhuri & H., '12]:
Every ε-differentially private estimator f(DB) has expected error at least

GES / (ε n) + inherent statistical error

Robust M-estimators

- Certain classes of *M-estimators* can be made differentially-private [Rubenstein *et al*, '09; Chaudhuri *et al*, '11; Chaudhuri & H., '12].
- Examples: median, trimmed-mean
  - Convergence rates of [Chaudhuri & H., '12] nearly match lower bounds.

Computational learning

- **Probably Approximately Correct (PAC) Learning** [Valiant, '84]: theoretical framework for statistical / computational learning
  - Together with Vapnik-Chervonenkis theory, provides basis for almost all modern machine learning theory
- **Basic question**: Is PAC Learning possible w/ differential privacy?

Limits of computational learning

- **Q**: **Is PAC Learning possible w/ differential privacy**?
  - In some special cases, yes
    [Kasiviswanathan *et al*, '08; Blum, Ligett, & Roth, '08; Chaudhuri & Monteleoni, '08; …]
  - **No, in general**! [Chaudhuri & H., '11]
- **Key**: being insensitive to individual data points (for the purpose of protecting privacy) harms general learning capabilities.

No private PAC learning in general

Task: learn a threshold function

$h_z(x) = 1$ if $x > z$, $h_z(x) = 0$ if $x \leq z$.

Two possible data distributions with
very different optimal threshold functions.

A differentially private learning algorithm with limited data must behave similarly in both cases:
∴ fails in at least one of the cases.

Special cases where learning is possible

- **Low-dimensional problems**: many near-optimal methods for "nice" data distributions.
- **High-dimensional linear problems**: Some recent progress [Rubinstein, Bartlett, Huang, and Taft, '09; Chaudhuri, Monteleoni, & Sarwate, '11; Kifer, Smith, & Thakurta, '12; …]
  - But accuracy suffers significantly for high-dimensional problems (unless ε large).

Private data release

- **Goal**: release "sanitized" version of training data set (*e.g.*, Netflix)
- In some cases, **information-theoretically possible** [Blum, Ligett, & Roth, '08].
- Many **computational intractability** results [Dwork, Naor, Reingold, Rothblum, & Vadhan, '09; Ullman & Vadhan, '11]
  - Recent progress: [Hardt, Ligett, McSherry, '12]
- **Inherent limitation**: must anticipate the types of learning algorithms someone might want to run!

4. CONCLUDING REMARKS

Privacy as a first-order criterion

- Many privacy violations discovered in (previously) unexpected scenarios.
  - Lots of work trying to reform privacy law / regulations to prevent and react to privacy harm.
- **Also need technical solutions**:
  data analysis tools guaranteed to limit privacy harm.
  - Here, mathematical proofs of privacy are essential.
  - **Challenge**: need to understand semantics!
    (*e.g.*, how to interpret ε?)

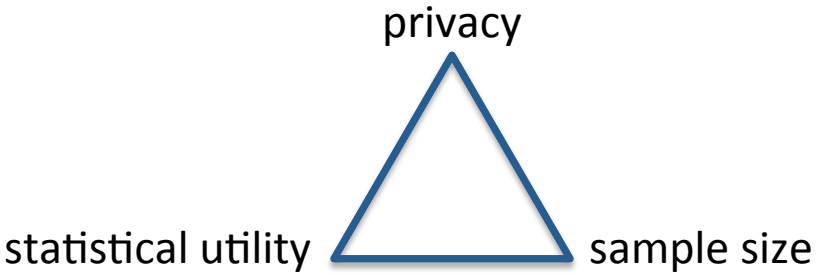
Stronger notions of privacy

- **Pan privacy** [Dwork, Naor, Pitassi, Rothblum, & Yekhanin, '10]
  - Worried that memory of learning algorithm can be compromised and exposed.
- **Local privacy** [Kasiviswanathan et al, '08; Duchi, Jordan, & Wainwright, '12]
  - Don't even trust the learning algorithm / "data curator".

What about weaker (but still non-broken) notions of privacy?

Privacy and machine learning

- Statistical utility vs. privacy trade-offs
- But there is a third-component: **data set size**



  - More data → smaller noise → more stat. utility
  - How to get more data?
    Incentives w/ privacy guarantees?

Questions?