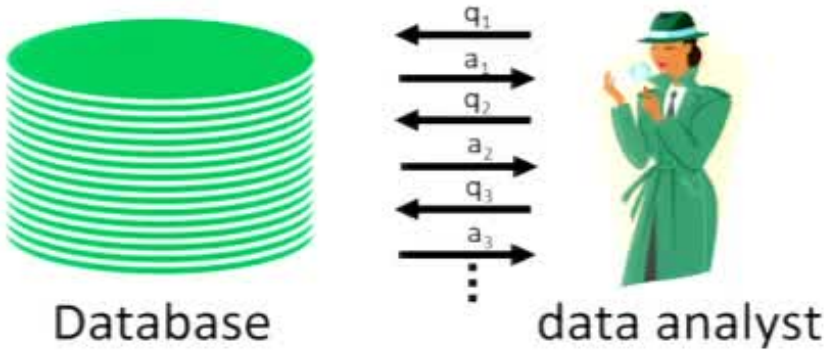
Privacy and Validity in the Land of Plenty

Cynthia Dwork, Microsoft Research

Helen Nissenbaum

Privacy-Preserving Data Analysis



- Census, epidemic detection based on OTC drug purchases; analysis of loan application data for evidence of discrimination,...
- 50+ year old problem

“De-Identification”?

Original Database

“De-Identified” Data Set

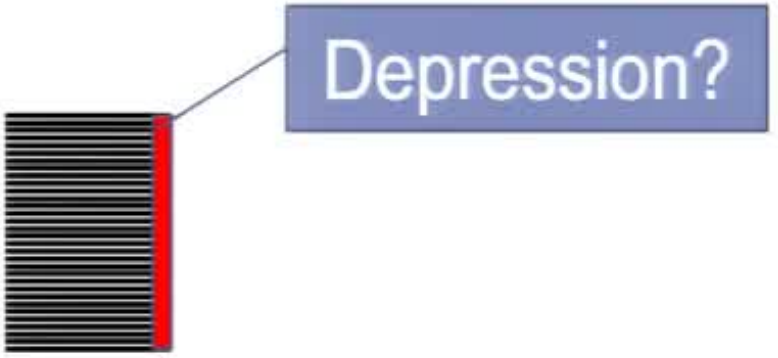
DE-IDENTIFIED DATA ISN’T.

The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*

The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*
- Needle in a Haystack
  - Determine presence of an individual's genomic data in GWAS case group
- The Big Bang attack
  - Reconstruct "depression" bit column



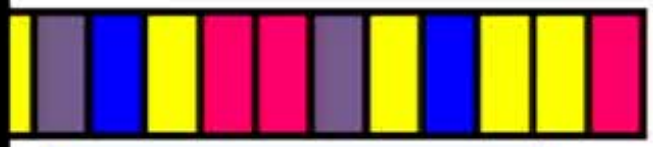
Dinur and Nissim'03 Homer+'08

Fundamental Law of Info Recovery

- "Overly accurate" estimates of "too many" statistics is blatantly non-private.

Information Flows and Combines

Voter Weld: M, DOB, zip



HapMap

2014

2012

2013

Billing

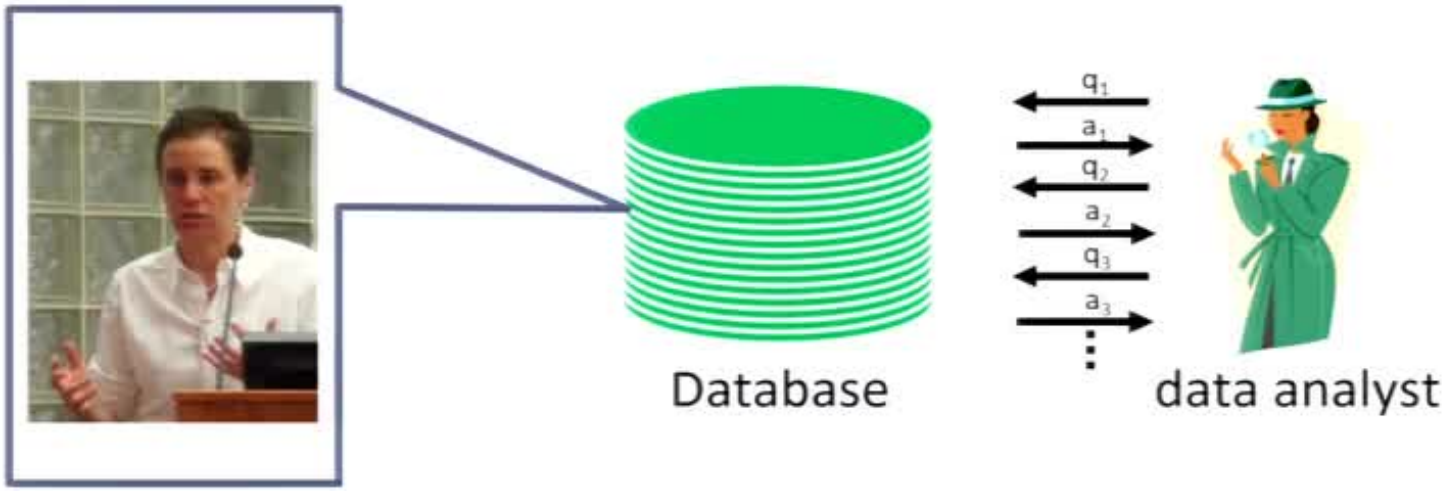
Blog

People who bought this…

"Computer science got us into this mess. Can computer science get us out of it?"

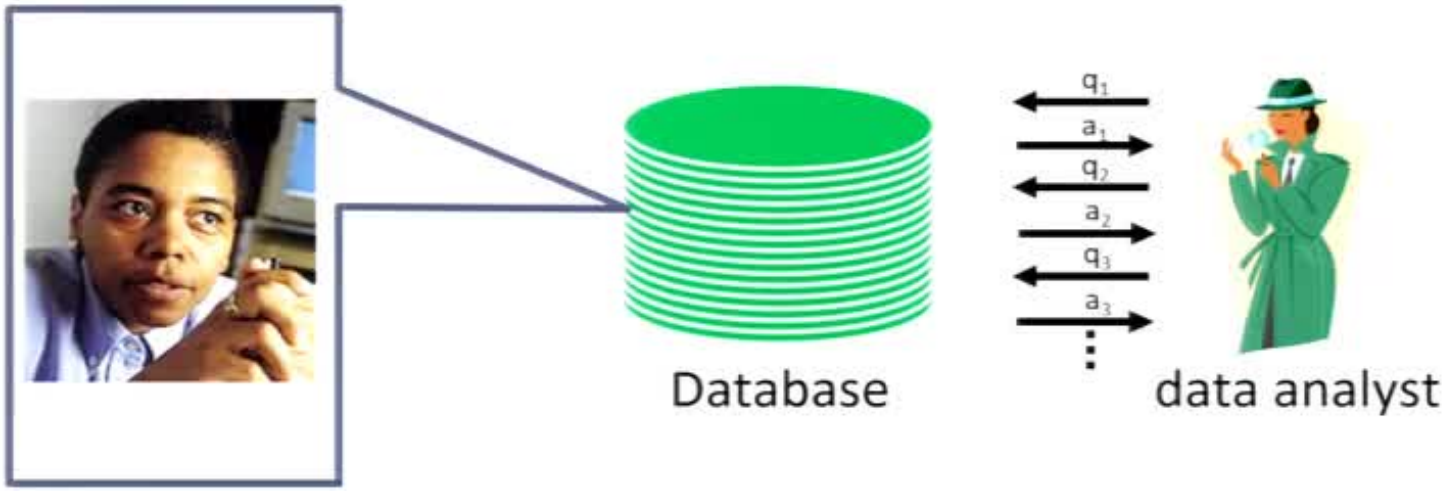
Latanya Sweeney, 2012

Privacy-Preserving Data Analysis?



- "Can't learn anything new about Nissenbaum"?
- Then what is the point?

Privacy-Preserving Data Analysis?

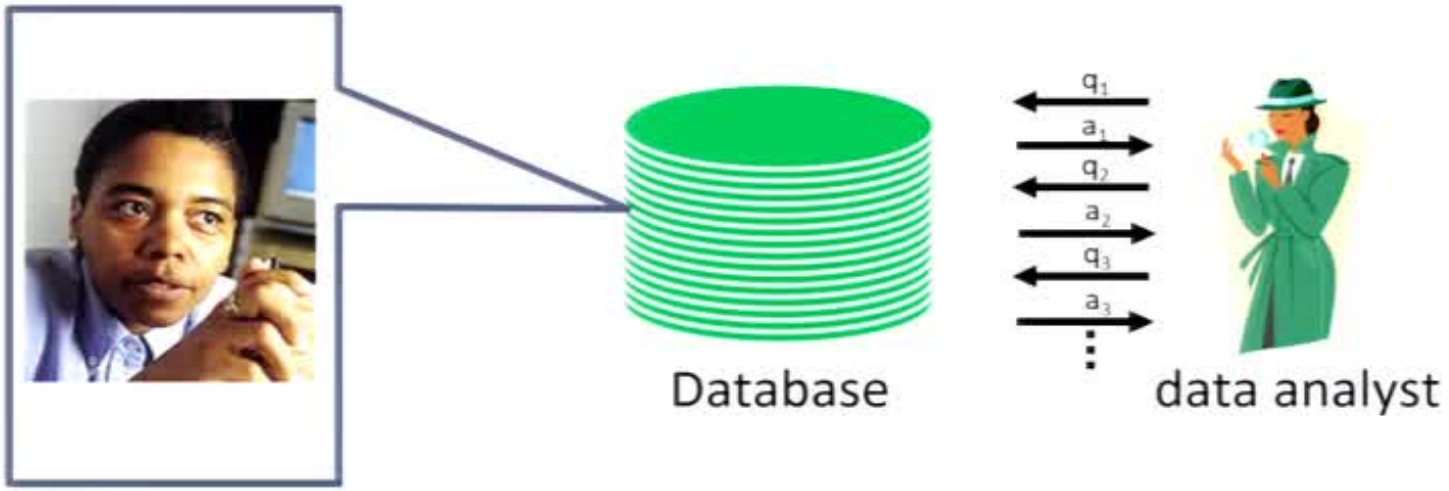


- Ideally: learn same things if Nissenbaum is replaced by another random member of the population ("stability")

Differential Privacy

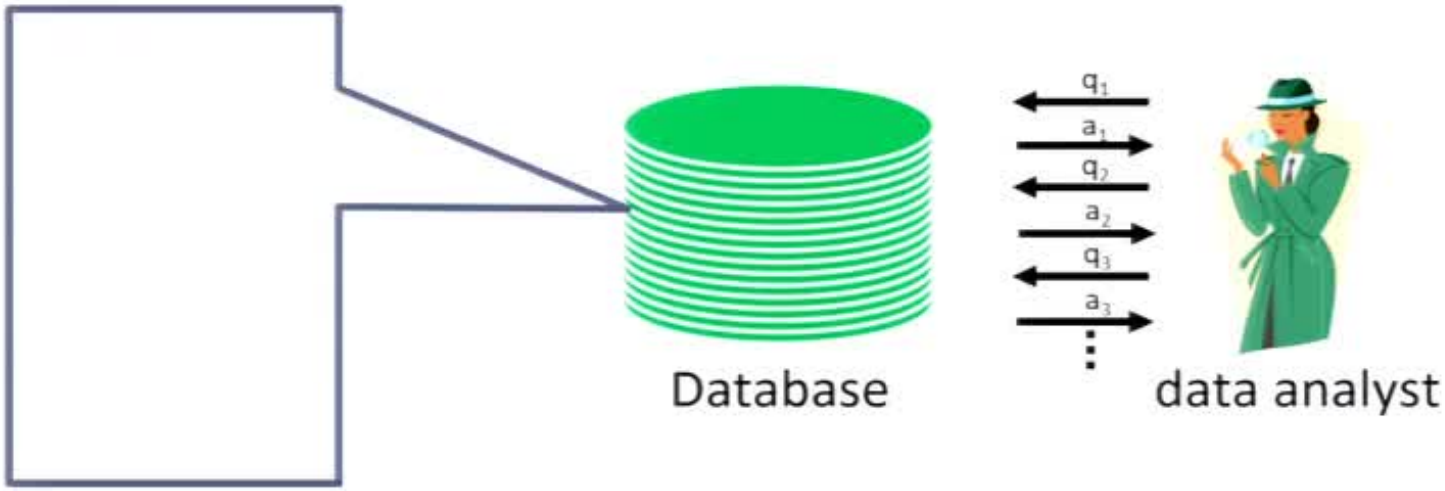
- The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset.
  - Nissenbaum goes away, Sweeney joins, Nissenbaum is replaced by Sweeney

Privacy-Preserving Data Analysis?



- Stability preserves Nissenbaum's privacy AND prevents over-fitting
- Privacy and Generalization are aligned!

Teachings vs Participation



SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, and May Complicate Pregnancy.

Differential Privacy

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

If a bad event is very unlikely when I'm not in dataset (y)
then it is still very unlikely when I am (x)

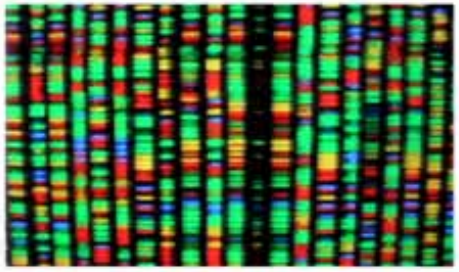
Bounded Ratio

Differential Privacy

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

If a bad event is very unlikely when I'm not in dataset (y) then it is still very unlikely when I am (x)



"Bounded Ratio"

Differential Privacy

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

"Privacy Loss"

If a bad event is very unlikely when I'm not in dataset (y)
then it is still very unlikely when I am (x)

Impossible to know the actual probabilities of bad events.
Can still control change in risk due to joining the database.

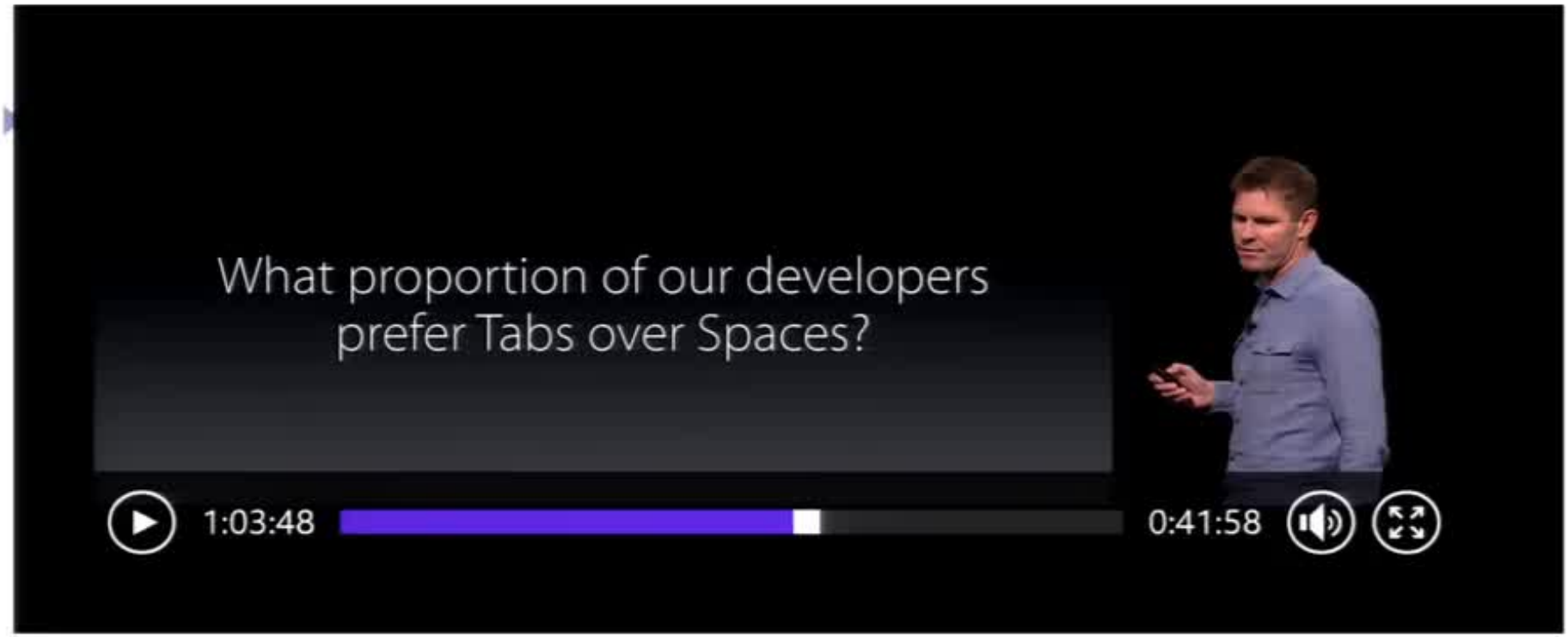
Bounded Ratio

Properties

- Future-proof
  - Current and future(!) side information, post-processing
- Automatically yields group privacy
  - $k\epsilon$ for groups of size k
- Understand behavior under composition
  - Can bound cumulative privacy loss over multiple analyses
    - "The epsilons add up"
- Programmable
  - Complicated private analyses from simple private building blocks

Did You XYZ Last Night?

Did You XYZ Last Night?

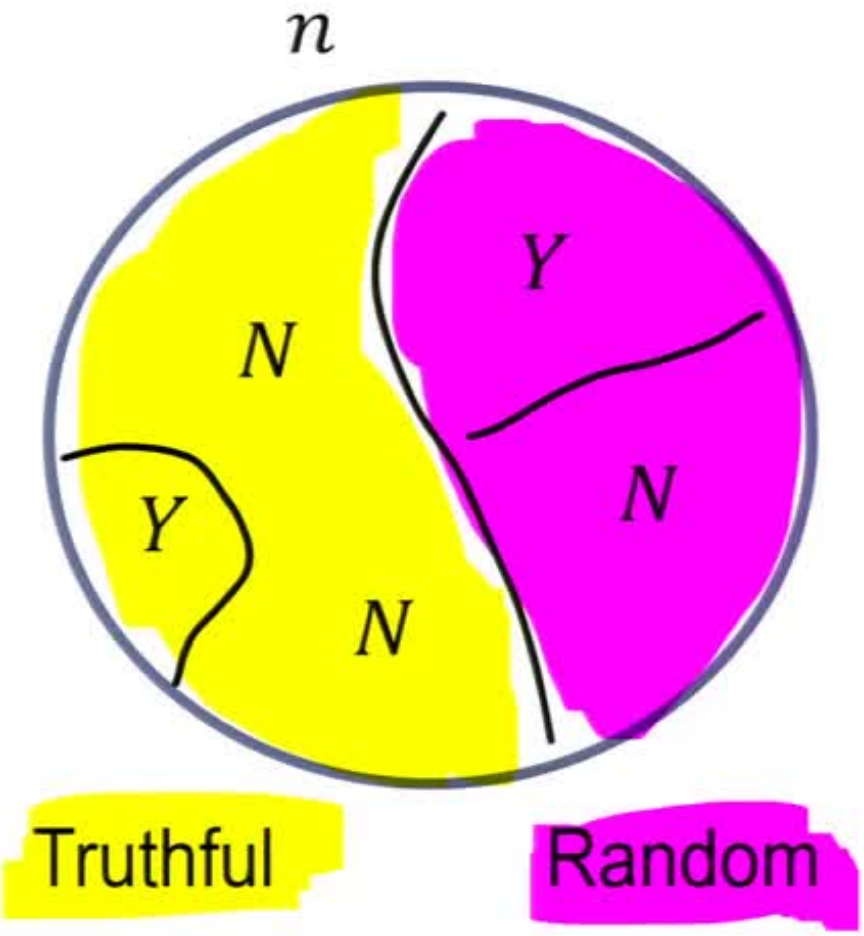


Tabs Over Spaces?

- Flip a coin.
  - Heads: Flip again and respond "Yes" if heads, "No" if otherwise
  - Tails: Answer honestly
- Analysis:
  - Pr [say "Y" given that truth = Y] / Pr [say "Y" given that truth = N] = 3
    - If truth is Y, will say "Y" if first coin is tails (probability ½) or first coin is heads and second coin is heads (probability ½ × ½ = ¼), total probability ¾
    - If truth is N, will say "Y" only if first and second coins are heads, probability ¼
  - Pr [say "N" given that truth = N] / Pr [say "N" given that truth = Y] = 3
  - $\epsilon \approx 1.098$

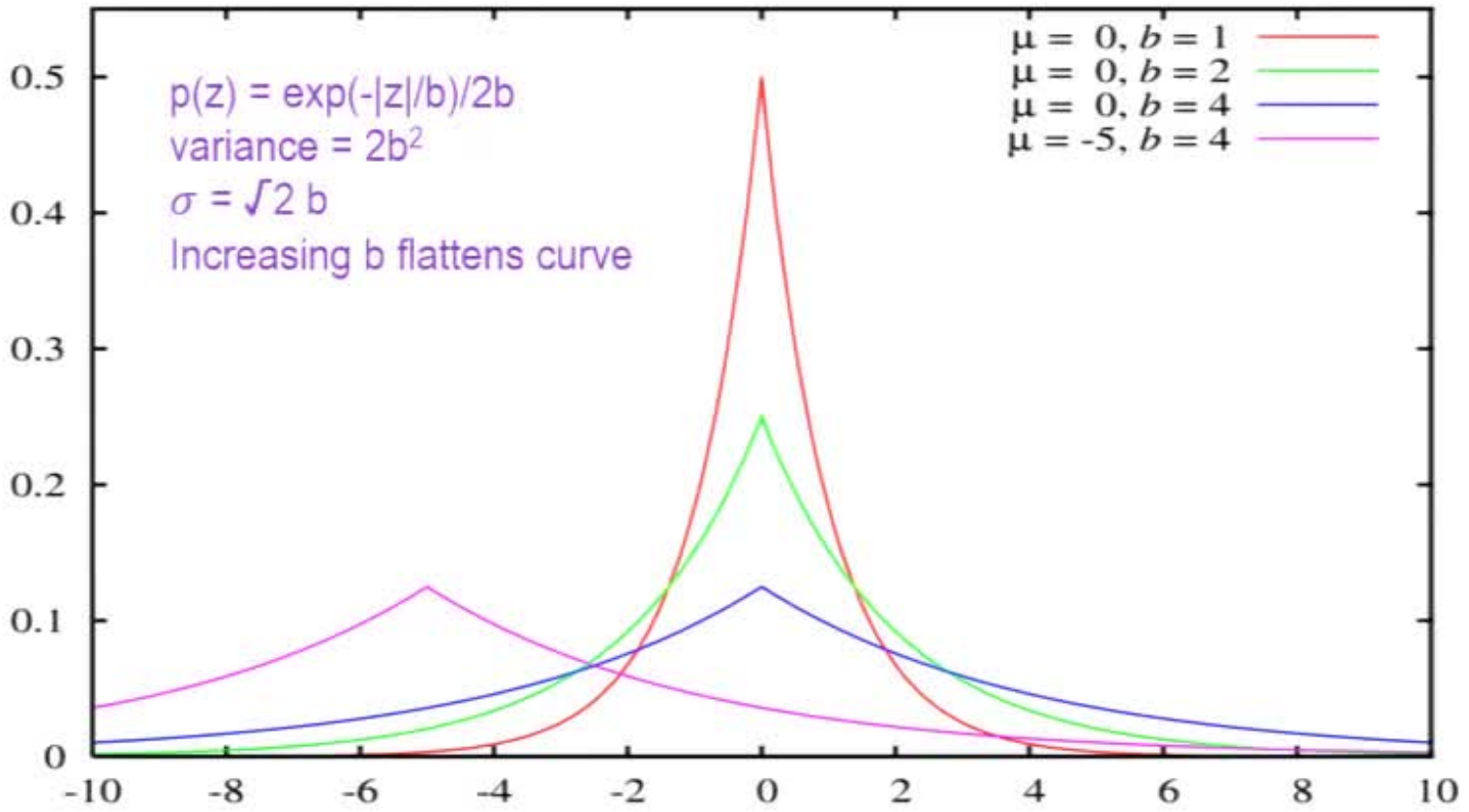
Warner 1965

(Fractional) Estimation Error



- # Random: $n' \sim \frac{n}{2} + c_1\sqrt{n}$
- # Random Y's: $\sim \frac{n'}{2} + c_2\sqrt{n'}$
- Estimate fraction true Y's:
  - (#Y answers $-\frac{n}{4}$) / ($\frac{n}{2}$)
  - Expected fractional error: $O(\frac{1}{\sqrt{n}})$
  - Comparable to sampling error

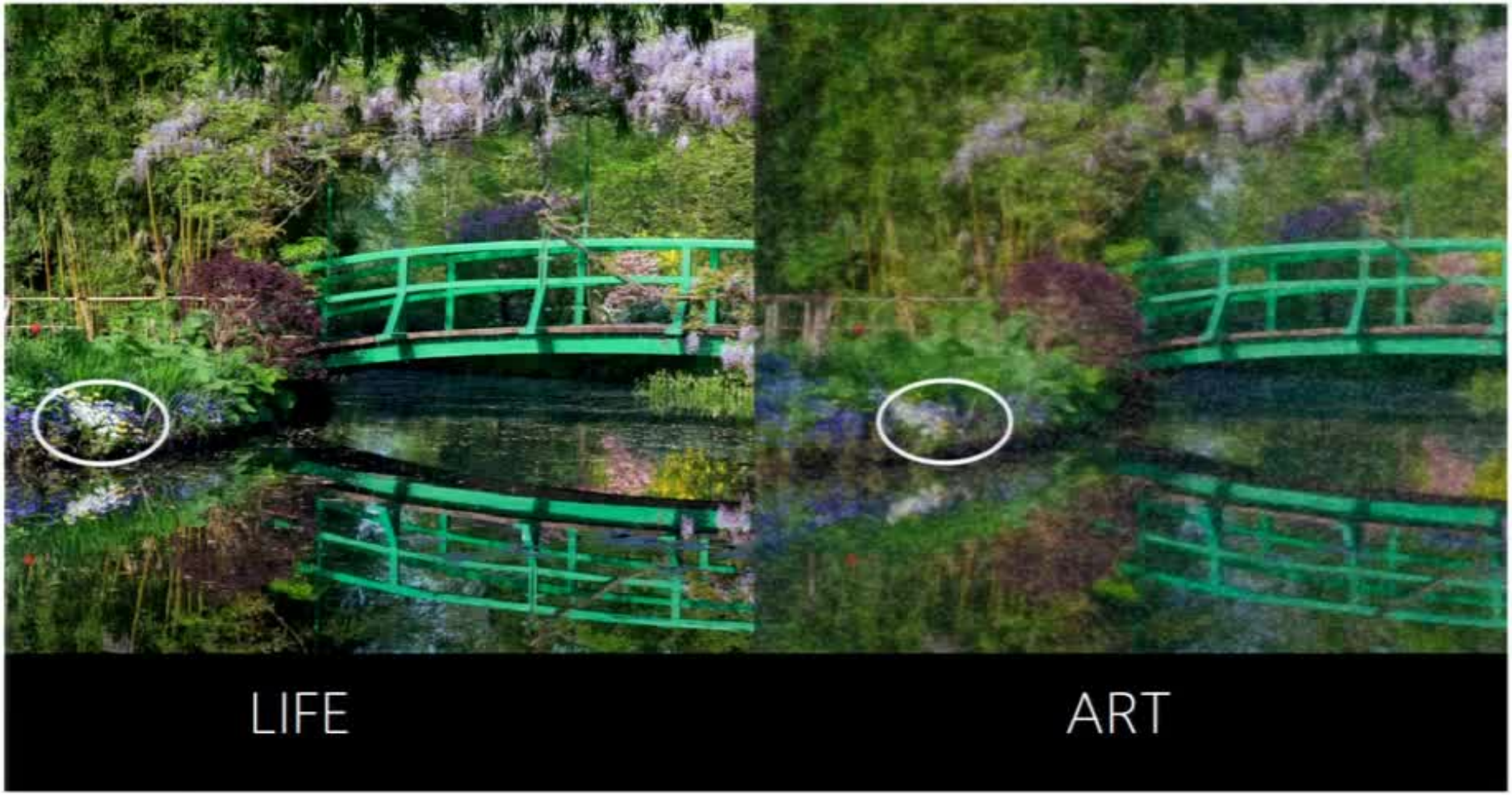
Specifically: Lap($\frac{\Delta f}{\epsilon}$) Noise



Rich Algorithmic Literature

- Counts, linear queries, histograms, contingency tables (marginals)
- Location and spread (eg, median, interquartile range)
- Dimension reduction (PCA, SVD), clustering
- Support Vector Machines
- Sparse regression/LASSO, logistic and linear regression, gradient descent
- Boosting, Multiplicative Weights
- Combinatorial optimization, mechanism design
- Privacy Under Continual Observation, Pan-Privacy
- Kalman filtering
- Statistical Queries learning model, PAC learning
- False Discovery Rate control in multiple hypothesis testing
- ...
- *The Algorithmic Foundations of Differential Privacy,* Dwork and Roth, August 2014

Which is “Truth”?



A Surprising Application of DP

Statistical Validity in Adaptive Data Analysis

Great Efforts to Control False Discovery

- Benjamini-Hochberg's "BHq" *et sequelae* for controlling the false discover rate (FDR) in multiple hypothesis testing
- Sophisticated cross-validation techniques
- Holdout sets for checking conclusions drawn from training data
- Pre-registration

- (Most) theory is for the static case
  - But science is by nature an adaptive process
  - It's going to get worse
  - Validity in the Land of Plenty

Intuition

Intuition

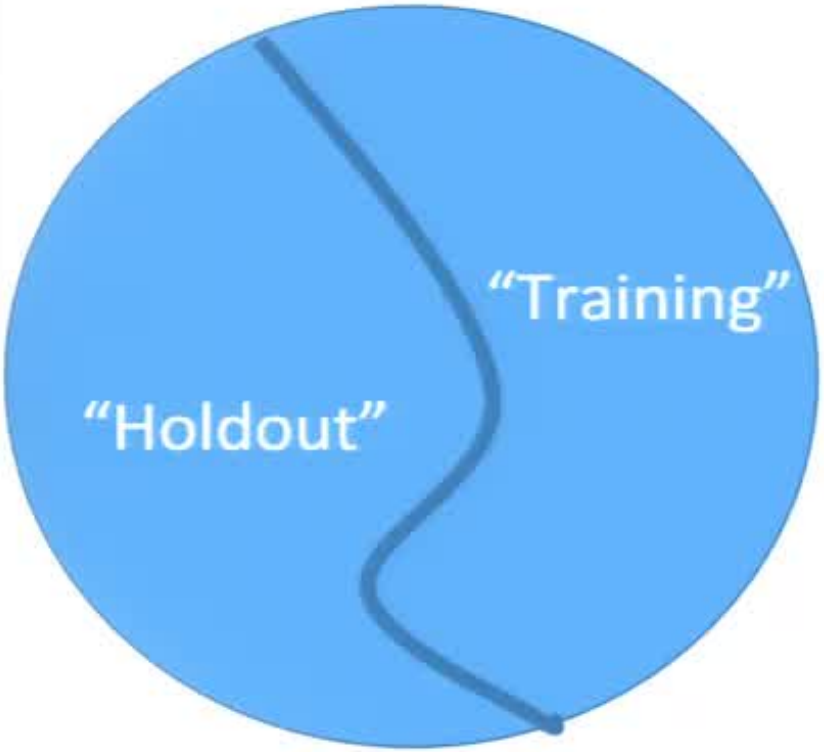
- Fix a query, eg, "What fraction of population is over 6 feet tall?"
- Almost all large datasets will give an approximately correct reply
  - Most datasets are representative with respect to this query
- If in the process of adaptive exploration, the analyst finds a query for which the dataset is not representative, then she must have "learned something significant" about the dataset.
  - Preserving the "privacy" of the data, may prevent over-fitting.

We want to do things the way we always have!

Down with DP interference!!!

Go Ahead. Make My Day.

The Re-Usable Holdout



- Learn on the training set
- Check against holdout via a differentially private mechanism
- Future exploration does not significantly depend on *H*
  - *H stays fresh!*

[D., Feldman, Hardt, Pitassi, Reingold, Roth '14]

Conclusion

- Problem studied for at least 50 years
- DP: General solution concept, robust to the networked world
- There is no competing theory of privacy-preserving data analysis
- There are jobs!

- The approach also tells us something fundamental about computing with stability
  - A general technique for statistical validity in adaptive data analysis

Thank you!

SIAM General Meeting, Boston, July 11, 2016