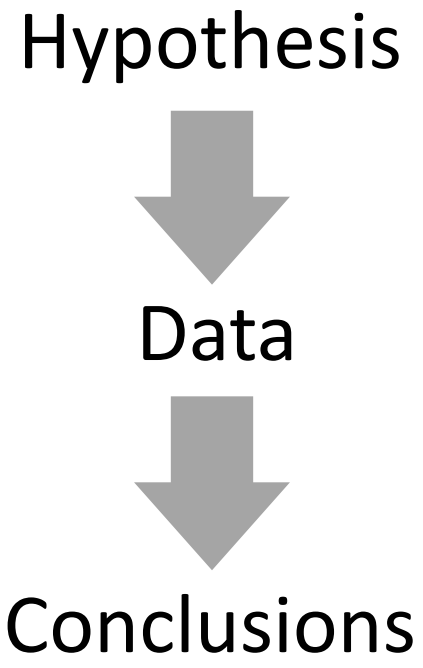
Preventing False Discovery via Differential Privacy

Jonathan Ullman, Northeastern University

Based on work and conversations with many people.
From this room: Cynthia Dwork, Kobbi Nissim, Adam Smith.

Privacy Tools Workshop, Nov 30, 2017.

Statistical Theory: One-Way Streets



Statistical analysis guarantees that your conclusions **generalize** to the population

And Yet…

PLOS | MEDICINE

OPEN ACCESS

ESSAY

1,140,912 VIEWS

1,413 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

NATURE | NEWS

Irreproducible biology research costs put at $28 billion per year

Study calculates cost of flawed biomedical research in the United States.

Monya Baker

09 June 2015

And Yet…

Trouble at the Lab – The Economist

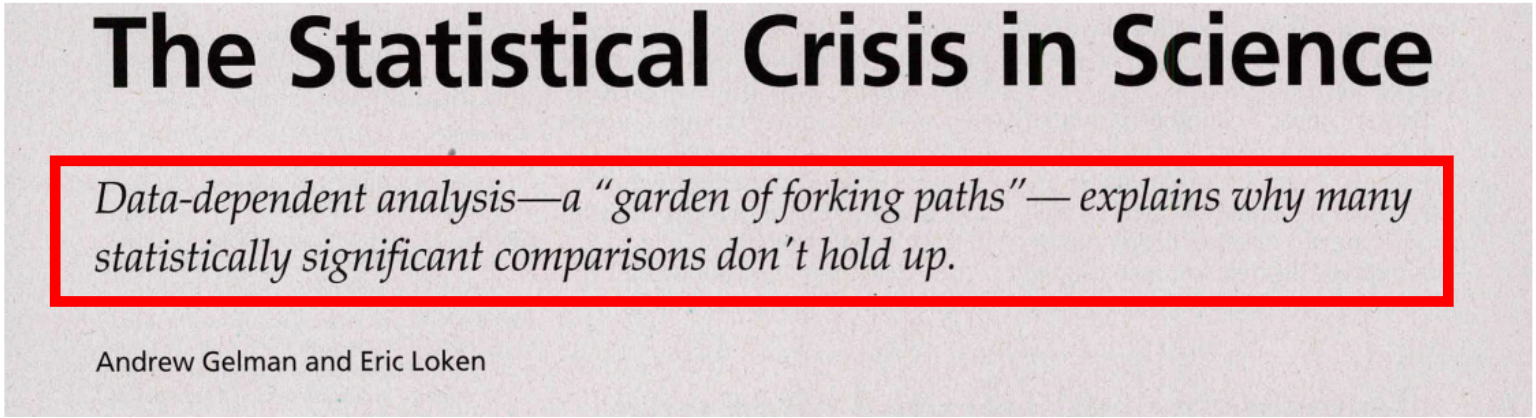
And Yet…

And Yet…

- Several causes of false discovery
  - Multiple Comparisons (fishing, dredging, p-hacking)
  - Misapplication/Misinterpretation of Statistical Methods
  - Misaligned Incentives (publication bias)
  - Outright Fraud

- But I won’t talk about these today

And Yet…

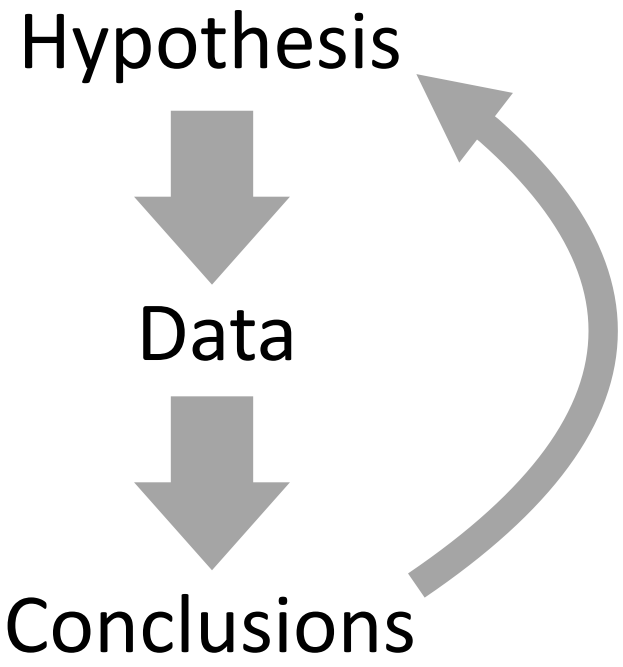


The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

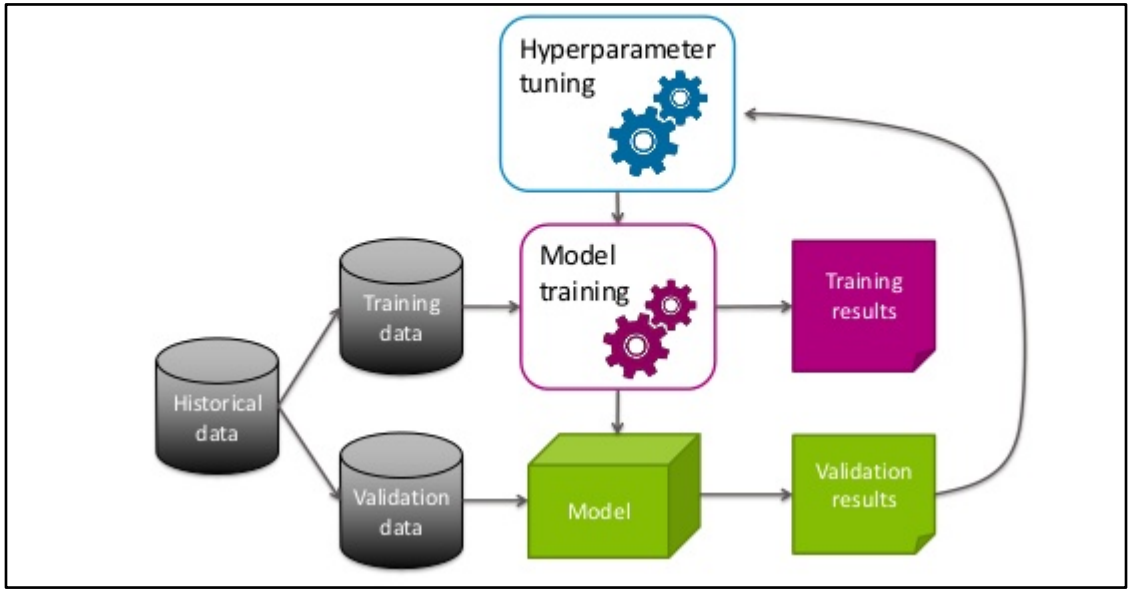
Statistical Practice: Traffic Circles



Statistical guarantees no longer apply
when the dataset is re-used **adaptively**

Adaptive Data Analysis: Examples

- Well specified adaptive algorithms
  - Select features then fit a model (Freedman's Paradox)
  - Hyperparameter tuning (sometimes)



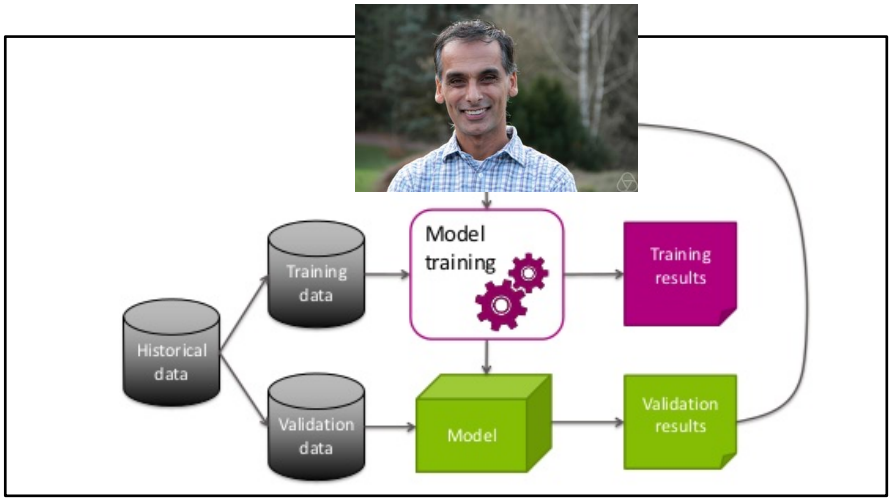
Alice Zheng. "Evaluating Machine Learning Models."

Adaptive Data Analysis: Examples

- "Researcher degrees of freedom"

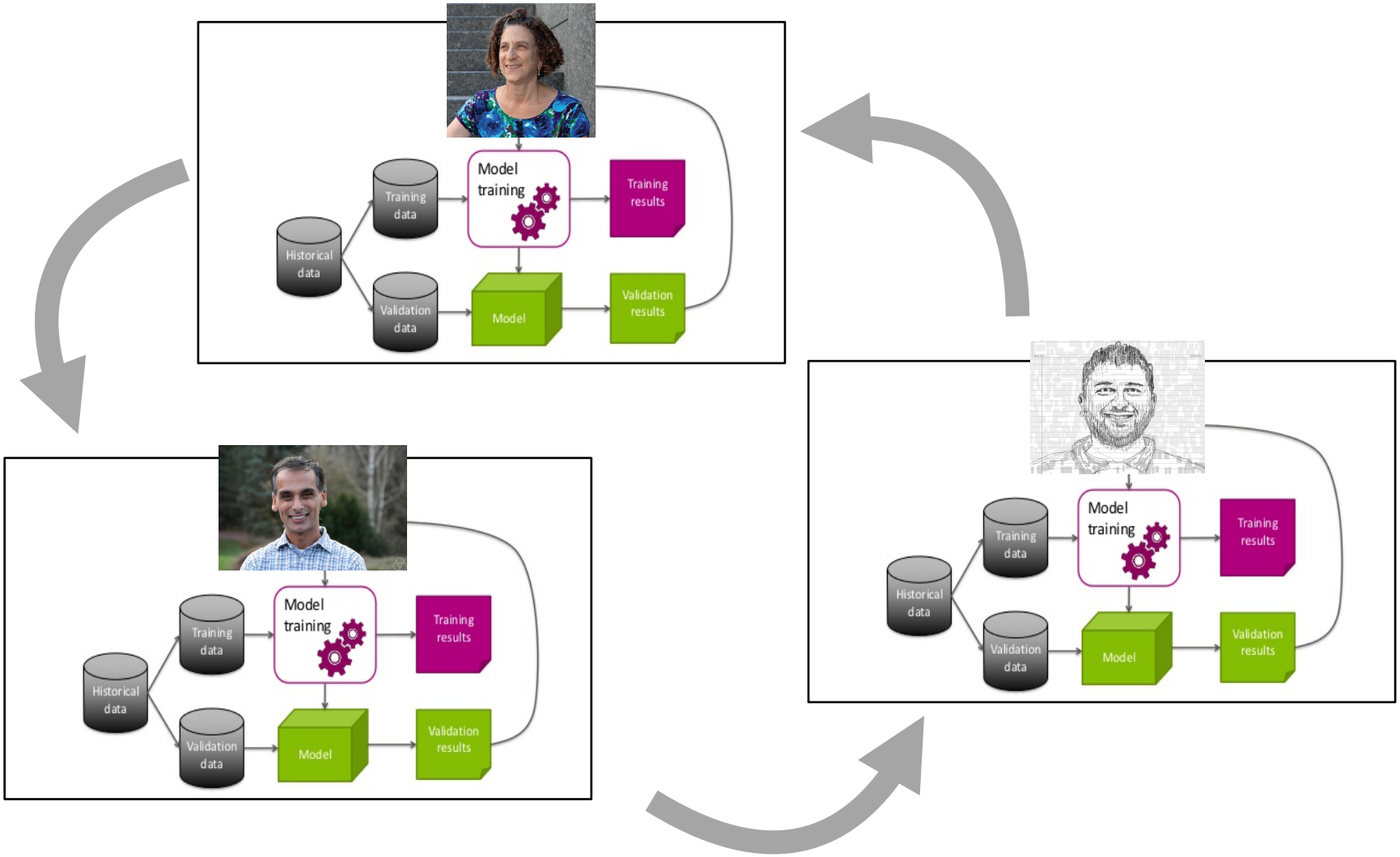
> The interaction effect is not significant when the scale from the Danish study are used to gauge the US subjects' support for redistribution. This arises because two of the items are somewhat unreliable in a US context. Hence, for items 5 and 6, the inter-item correlations range from as low as .11 to .30. These two items are also those that express the idea of European-style market intervention most clearly and, hence, could sound odd and unfamiliar to the US subjects. When these two unreliable items are removed (α after removal = .72), the interaction effect becomes significant.

A. Gelman, E. Loken. "The Garden of Forking Paths."



Adaptive Data Analysis: Examples

- Reuse of datasets by multiple researchers



Previous Approaches

- Hypothesis testing
  - Assumes hypotheses are independent of the data
- Holdout sets / data splitting
  - Need a new holdout set to validate each analysis
- Explicit post-selection inference
  - Only tractable for simple, well specified tasks
- Pre-registration
  - Does not allow exploratory data analysis

Desiderata for Adaptive Data Analysis

- Methods for preventing overfitting that…
  - …are not tied to any specific algorithm or problem
  - …degrade gracefully under composition
  - …allow arbitrary post-processing

- Flashback to the very first project meeting c.2012:

Desiderata for Private Data Analysis

- Methods for protecting privacy that…
  - …are not tied to any specific algorithm or problem
  - …degrade gracefully under composition
  - …allow arbitrary post-processing

Adaptive Data Analysis via DP

- A flexible approach to adaptive data analysis
  - Introduced in [DFHPRR'15, H**U**'14] (also [M'08])
  - Key Tool: Differential Privacy
  - Key Message: Differential Privacy can increase utility!

- My Results
  - Improving accuracy and generality [BNSSS**U**'16, N**U**'17]
  - New composition properties of DP [RR**U**V'16]
  - Novel bottlenecks to preventing overfitting in adaptive data analysis [H**U**'14,S**U**'15,NSSS**U**'??]

Case Study: ML Competitions

- Goal: solve a binary classification task
  - Examples are y_i, binary labels are z_i
  - Classifier c maps examples y to labels z
  - Score of c is the fraction of labels it predicts correctly

Training dataset

$y_1 = 010$	$z_1 = 0$
$y_2 = 011$	$z_2 = 1$
$y_3 = 010$	$z_3 = 0$
…	…
$y_n = 111$	$z_n = 0$

Case Study: ML Competitions

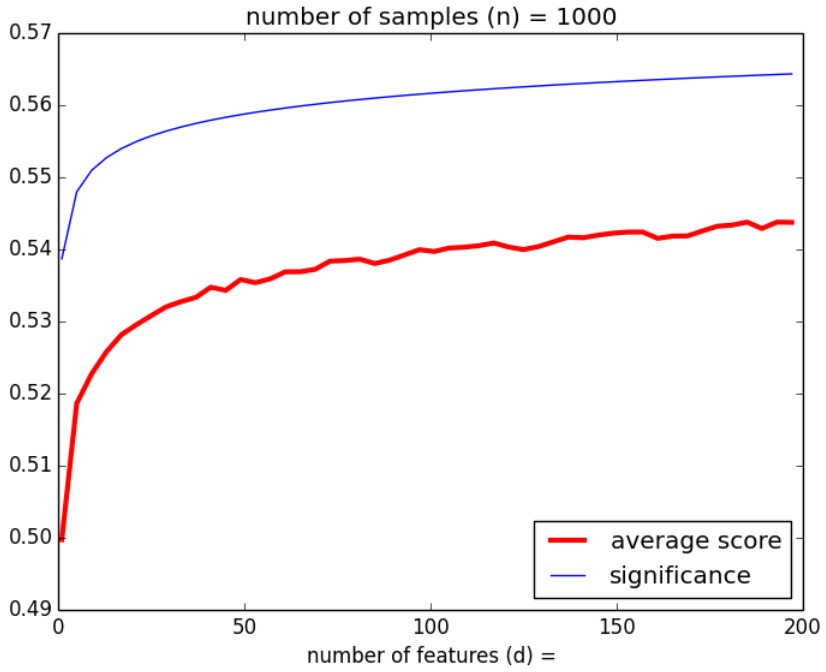
- Unsure how to begin, you try random classifiers
  - Choose $c_1, \ldots, c_k$ randomly
  - Select the "best" one using your test data
  - Test for statistical significance (using Bonferroni correction)

Scores

c_1	.4885
c_2	.5420
c_3	.4850
…	…
c_k	.5055

Case Study: ML Competitions

- Unsure how to begin, you try random classifiers
  - Choose $c_1, \ldots, c_k$ randomly
  - Select the "best" one using your test data
  - Test for statistical significance (using Bonferroni correction)



Score increases with d but remains statistically insignificant

Case Study: ML Competitions

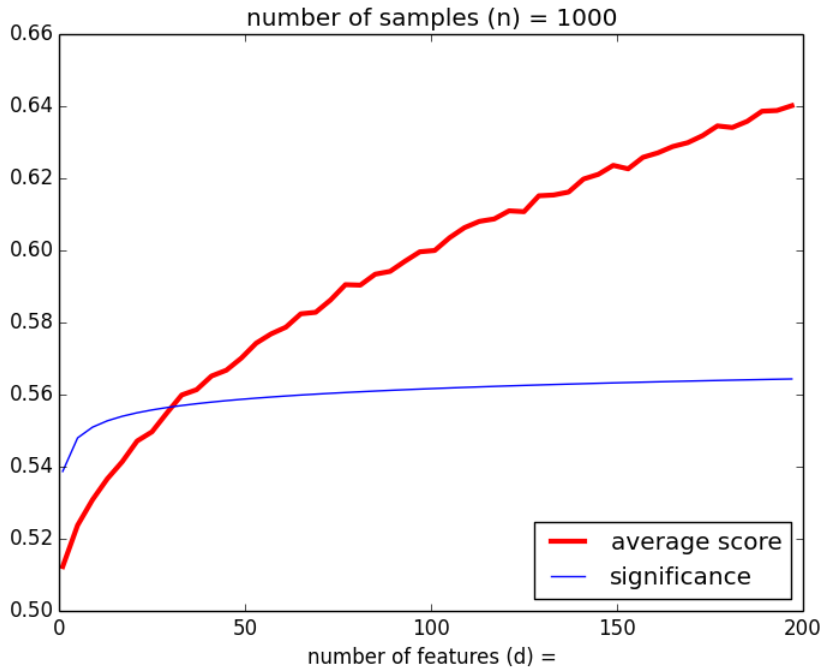
- Frustrated, you try slightly harder
  - Choose $c_1, \ldots, c_k$ randomly
  - Take the classifiers $c_2, c_5, c_6, \ldots$ that are better than random and let c_{k+1} be the majority vote of these classifiers (boosting)
  - Test for statistical significance (using Bonferroni correction)

Scores

c_1	.4885
c_2	.5420
c_3	.4850
…	…
c_k	.5055

Case Study: ML Competitions

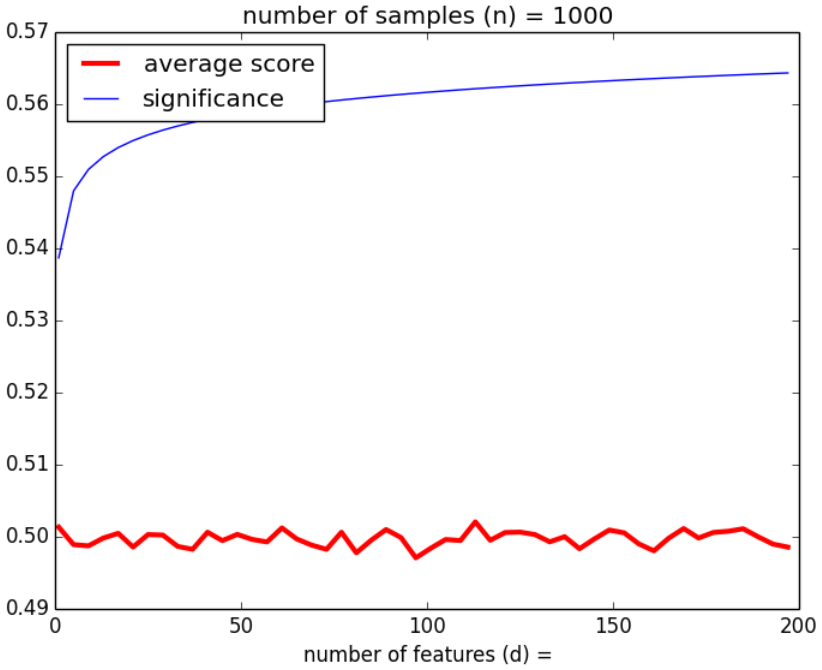
- Frustrated, you try slightly harder
  - Choose $c_1, \ldots, c_k$ randomly
  - Take the classifiers $c_2, c_5, c_6, \ldots$ that are better than random and let c_{k+1} be the majority vote of these classifiers (boosting)
  - Test for statistical significance (using Bonferroni correction)



Majority vote statistically significantly better than a random classifier

Case Study: ML Competitions

- Excited, you call your friend to brag about winning
- She informs you that this was a prank, the labels are random
- Your classifier c_{k+1} gets roughly 50% correct on fresh data



You look awfully foolish when tested on a fresh dataset

Case Study: ML Competitions

- Excited, you call your friend to brag about winning
- She informs you that this was a prank, the labels are random
- Your classifier c_{k+1} gets roughly 50% correct on fresh data

- You came up with c_{k+1} **adaptively** after seeing the scores of $c_1, \ldots, c_k$ on the same dataset!
  - "Freedman's Paradox"
  - Classical statistical hypothesis testing no longer applies
  - Closely related to reconstruction attacks in privacy [DN'03]

How Can We Prevent Overfitting?

- What went wrong?
  - The scores of the initial classifiers $c_1, \ldots, c_k$ reveal too much about the labels z
- What do we do about it?
  - Limit the information revealed about the labels
- How would we do that?
  - Use differential privacy!
  - Analyst can learn about the distribution but cannot learn too much about specific labels z_i

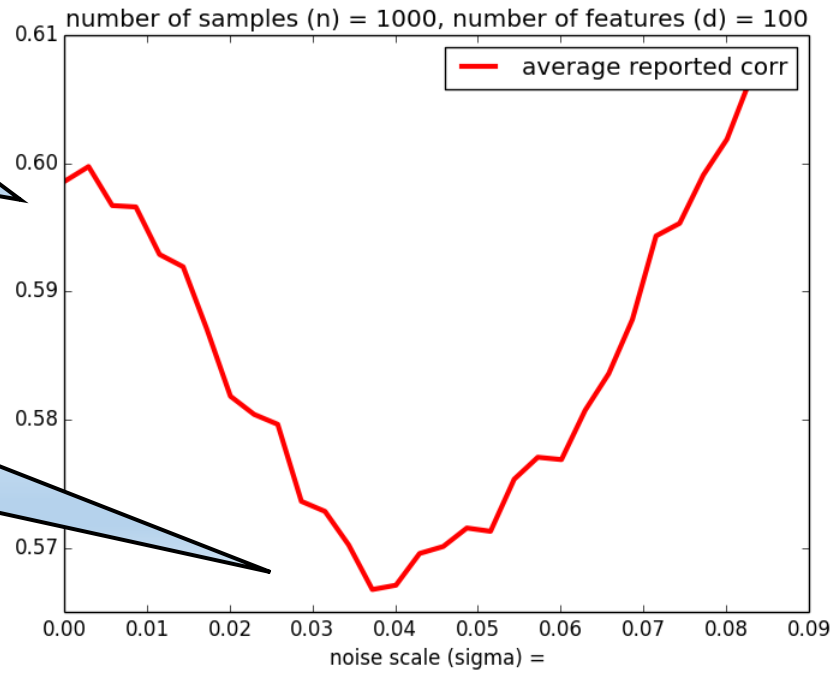
Case Study: ML Competitions

- You try adding Gaussian noise $N(0, \sigma^2)$ to the estimated score of each classifier to prevent overfitting
  - The best choice of σ is not 0!
  - Some privacy comes for free!

Simple way to ensure DP

No noise: overestimate score by ~10%

Some noise: overestimate score by ~6%



Case Study: ML Competitions

kaggle

- You try adding Gaussian noise $N(0, \sigma^2)$ to the estimated score of each classifier to prevent overfitting
  - The best choice of σ is not 0!
  - Some privacy comes for free!

Simple way to ensure DP

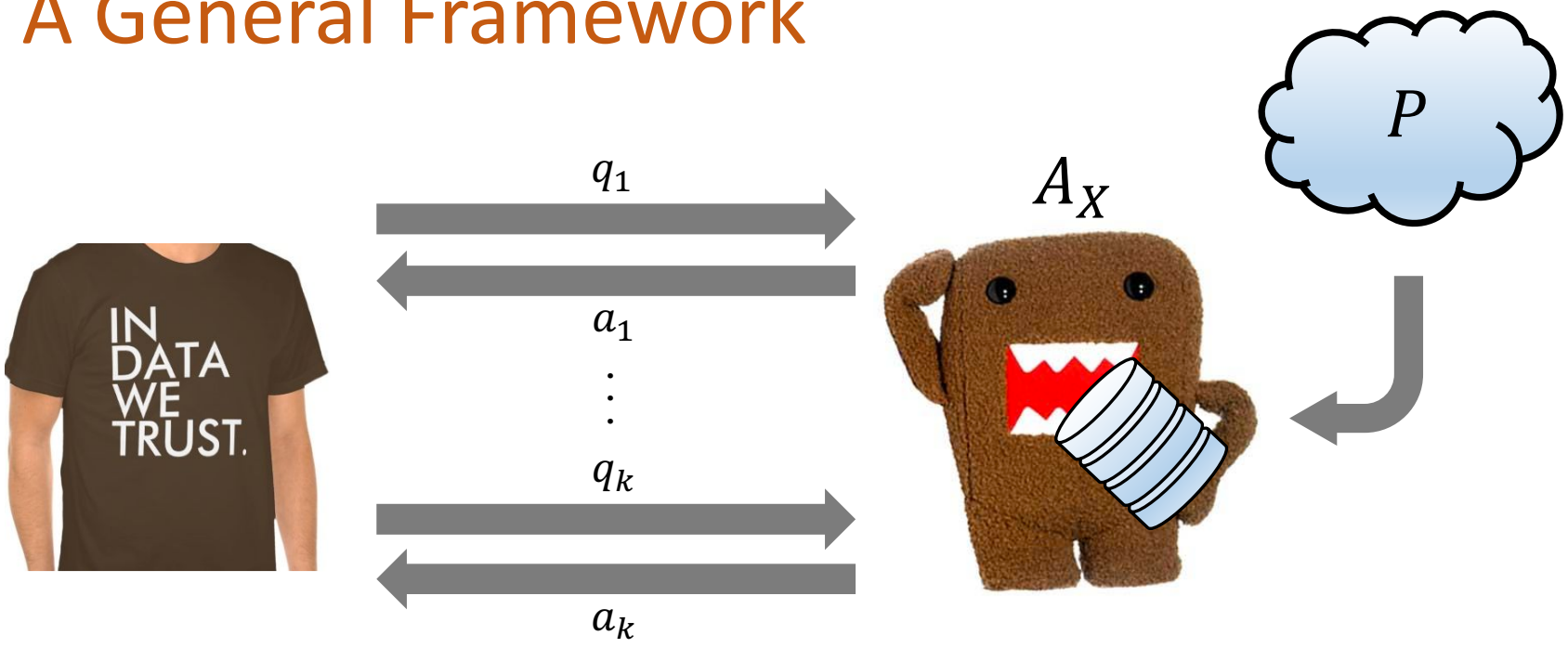
- Theoretical analysis: for any sequence of classifiers

Bias $\sqrt{\frac{k}{n}}$ → Bias $\sqrt{\frac{\sqrt{k}}{n}}$

No noise

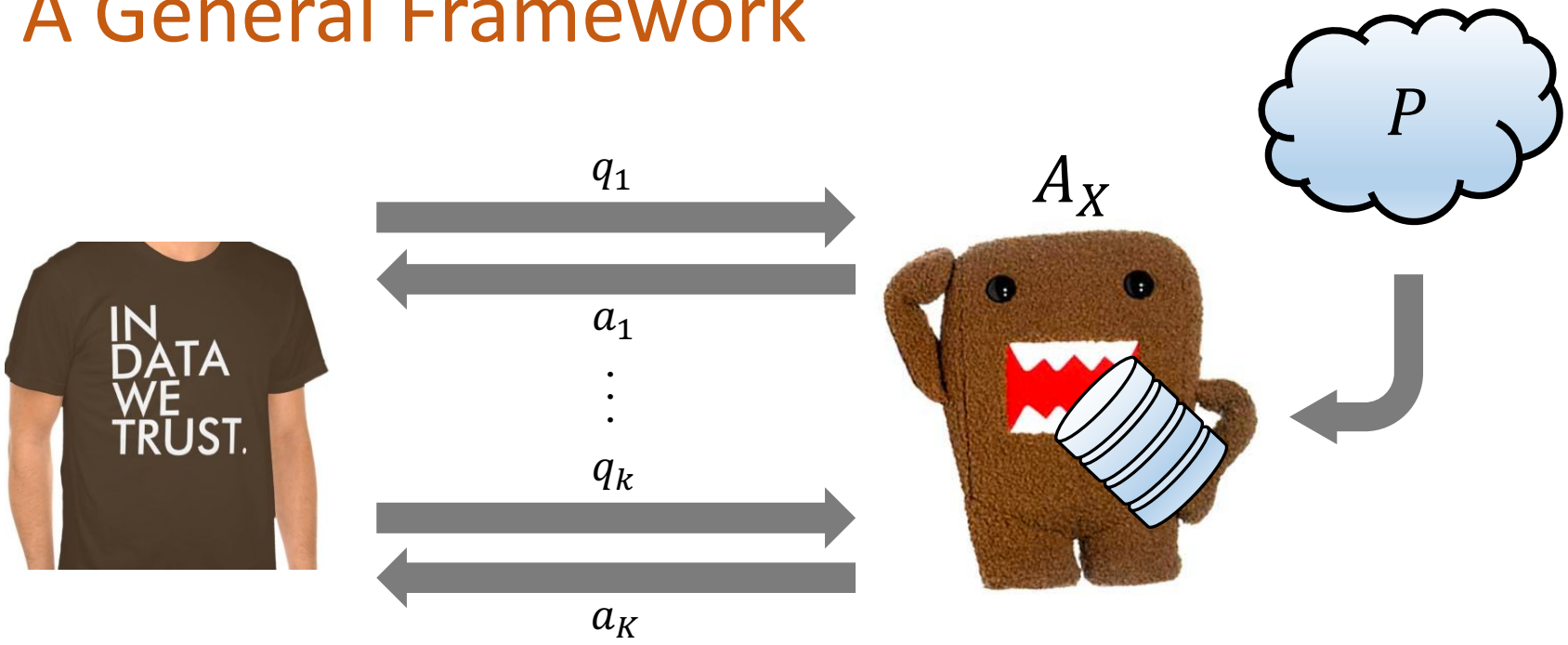
Optimally calibrated noise

A General Framework



- Goal: protect against overfitting while allowing arbitrary adaptive analysis of a dataset
  - Introduce some layer of noise into the information you release about the dataset
  - Give bounds on overfitting by tracking some measure of information revealed about the dataset

A General Framework



- Informal Theorem [DFHPRR'15, BNSSS**U**'16]: In a wide variety of settings, using a differentially private estimator A_X prevents overfitting
  - Often gives the best known estimator
  - Sometimes gives asymptotically optimal estimator

Example: Statistical Queries (SQs)

- Given a bounded function
$$\phi: U \to [\pm 1]$$
- The statistical query q_ϕ is defined as
$$q_\phi(P) = \mathbb{E}\left[\phi(P)\right]$$
- An answer a is α-accurate if $\left|a - q_\phi(P)\right| \leq \alpha$

- Highly useful and general family of queries
  - Mean, variance, covariance
  - Score of a classifier
  - Gradient of the score of a classifier
  - Almost all PAC learning algorithms
  - …

Captures Freedman's Paradox

Future Directions

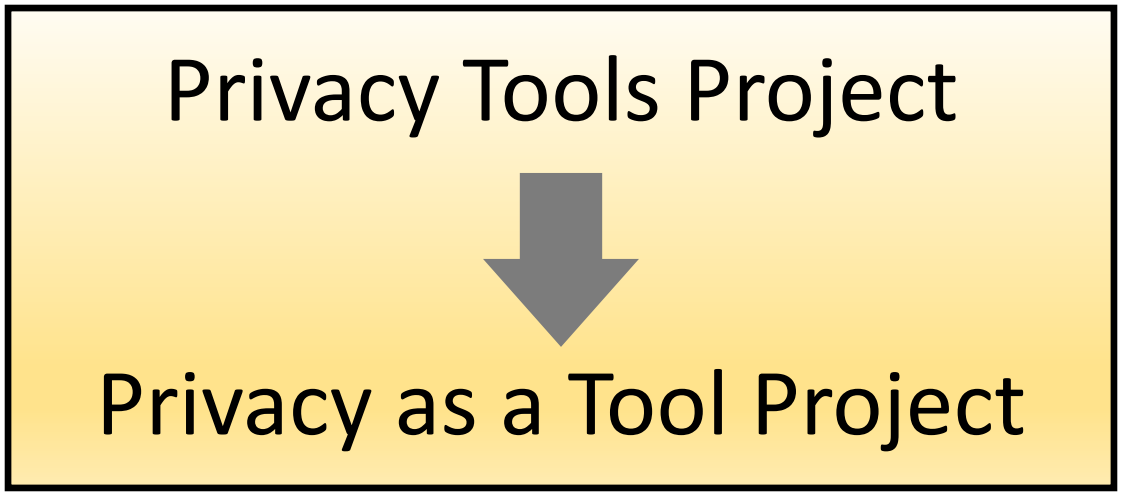
- Theory
  - Does privacy come “for free” in adaptive analysis
  - Expanding and optimizing our toolkit
  - Unify with classical approaches in statistics
    - Realistic models of data analysts

- Practice
  - Tune existing tools for good practical performance
  - Provide useful user interfaces
  - Identify canonical use cases
    - Paradoxical? Adaptive data analysis is inherently ill specified.

Subject of a current NSF grant with Marco Gaboardi

Future Directions

- Increasingly, differential privacy is a useful way to think about problems outside of privacy
  - Digital watermarking [B**U**V'14,S**U**'15,DSS**U**V'15]
  - Online learning [ALMT'17]
  - Mechanism design [MT'07,KRS**U**'14,LST'15]
  - Fairness [DHPRZ'12]



Thank you!