BU CS 591 Spring 2025
Privacy in ML and Statistics

Adam Smith (BU)
Jonathan Ullman (NEU)

Lecture 24: Adaptive Data Analysis

Today

- Adaptive validity in statistical analysis
  - Example setting: ML competitions
  - What can go wrong
  - Nothing about privacy!
- Privacy prevents overfitting
  - Single query case
  - Extension to multiple queries
  - General transfer theorem

Statistical Theory

Method

↓

Sample (from population)

↓

Conclusions

Statistical analysis guarantees that your conclusions generalize to the population

Statistical Practice

PLOS | MEDICINE

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

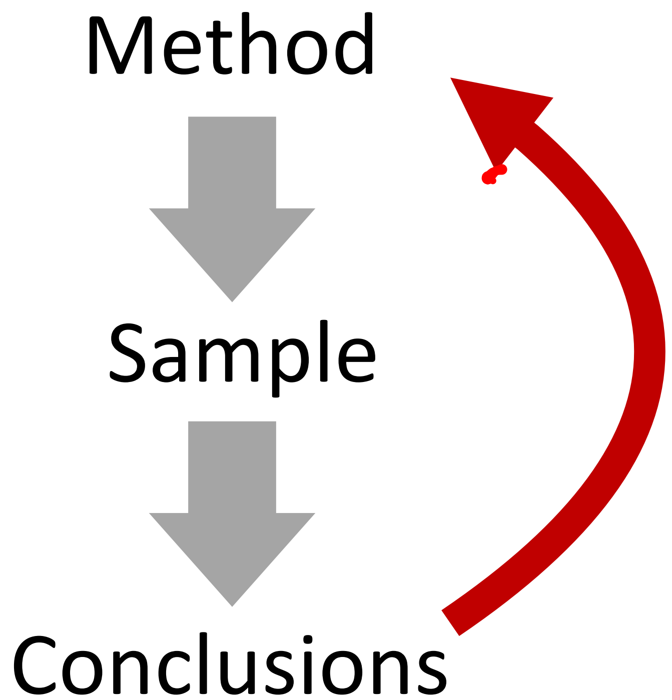
Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

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

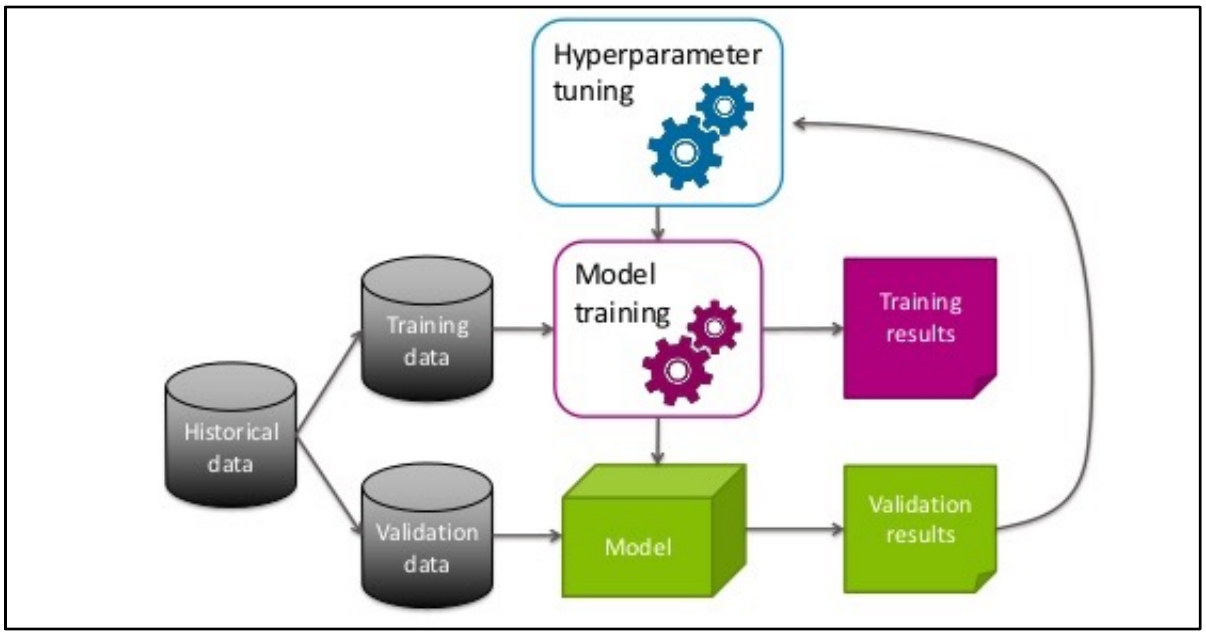


Statistical guarantees no longer apply
when the method and sample are correlated

Examples of Adaptive Data Analysis

Well-specified adaptive algorithms

- Select features then fit a model (Freedman's Paradox)
- Hyperparameter tuning (sometimes)
- **Data science competitions**



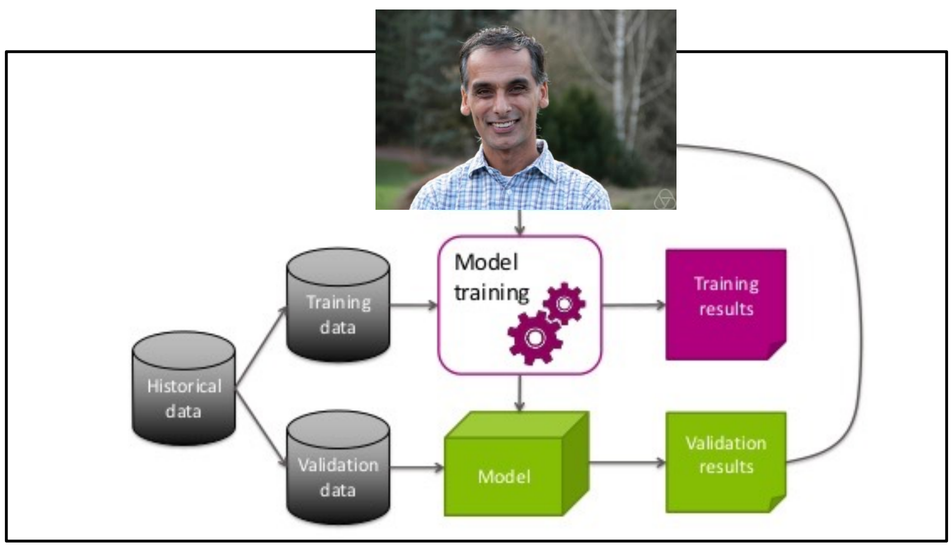
Alice Zheng. "Evaluating Machine Learning Models."

Examples of Adaptive Data Analysis

Researcher degrees of freedom

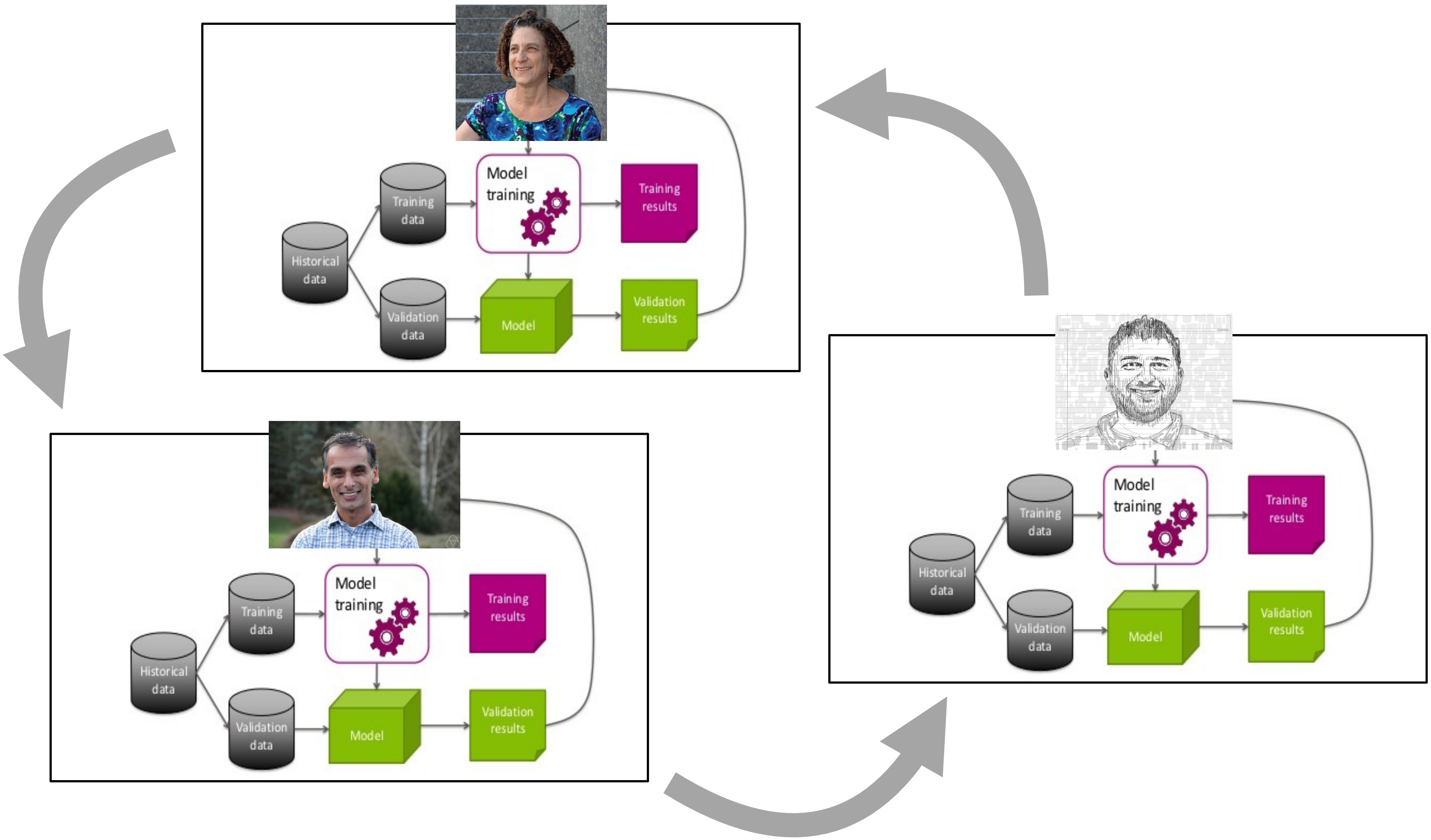
> The interaction effect is not significant when the scale from the Danish study are used to gauge the US subjects' support for redistribution. This arises because two of the items are somewhat unreliable in a US context. Hence, for items 5 and 6, the inter-item correlations range from as low as .11 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."



Examples of Adaptive Data Analysis

Reuse of datasets by multiple researchers



Case Study: ML Competitions

kaggle

secret **competition data** X

z_1	$s_1 = 0$
z_2	$s_2 = 1$
z_3	$s_3 = 1$
…	…
z_n	$s_n = 1$

Classifier φ

Data Gladiator

Answer a

$$a \approx \text{score}_X(\varphi) = \frac{1}{n}\sum_i \mathbf{1}\{\varphi(z_i) = s_i\} = \mathbb{E}_X(\mathbf{1}\{\varphi(z_i) = s_i\})$$

where φ is a classifier

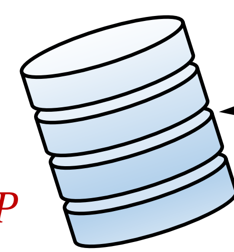
Competition: find a classifier φ^* with large score **on the distribution**

Needed: a method for estimating the score **on the underlying distribution**

$$\text{score}_P(\varphi) = \mathbb{E}_P(\mathbf{1}\{\varphi(z_i) = s_i\})$$

score on the underlying distribution

Secret Prize Distribution P



Competition distribution drawn from P

Case Study: ML Competitions

kaggle

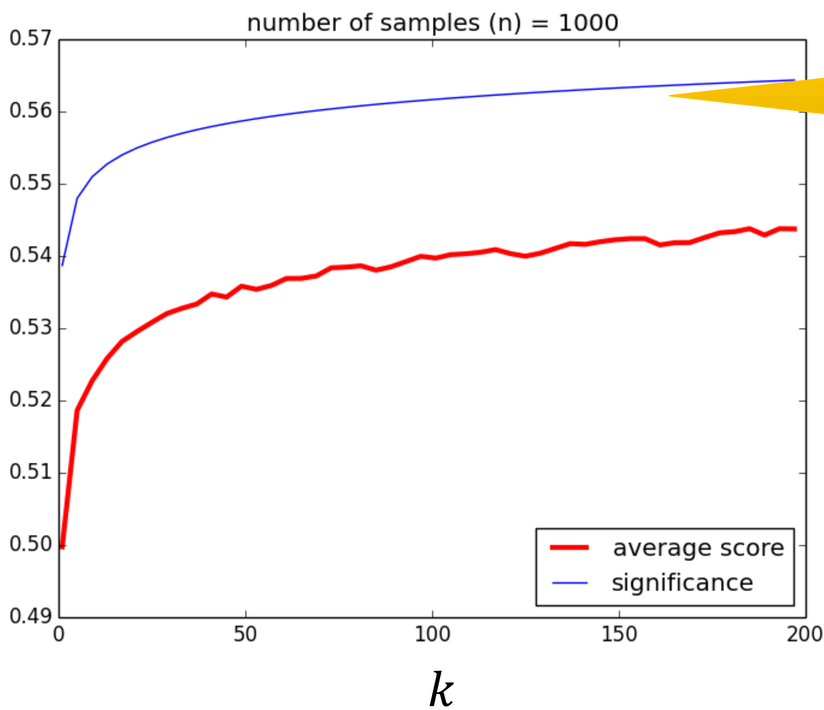
- Suppose prize and competition data have **random labels**
  - Any classifier will have $\text{score}_P(\varphi) \approx \frac{1}{2}$ on the prize distribution P
  - If $\text{score}_X(\varphi) \gg \frac{1}{2}$ then we have overfit

- **How can we prevent the competitors from overfitting to the competition data?**

- **Naïve algorithm:**
  - answer $a = \text{score}_X(\varphi) = \frac{1}{n}\sum_i \mathbf{1}\{\varphi(z_i) = s_i\}$
  - Let's see how well this algorithm does at preventing overfitting

Non-adaptive analysis

- **Competitor's strategy (non-adaptive):**
  - Choose k random classifiers $\varphi_1, \ldots, \varphi_k$
  - Receive $a_1, \ldots, a_k$ where $a_j = score_X(\varphi_j)$
  - Output $\varphi^* = \operatorname{argmax} \operatorname{score}_X(\varphi_j)$



95% significance threshold

Theorem (nonadaptive accuracy):

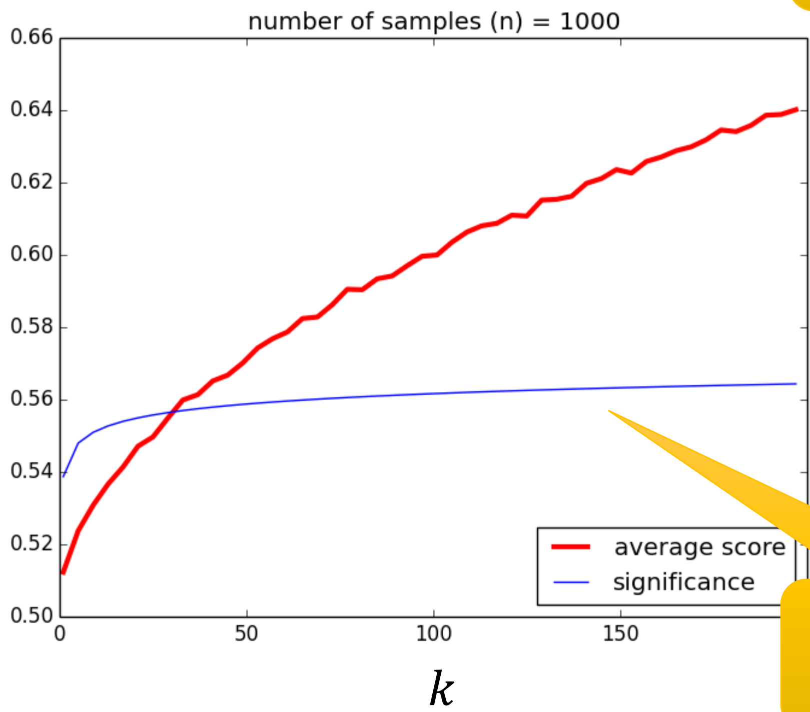
$$\mathbb{E}\left(\max_j \mathrm{sc}_X(\varphi_j) - \mathrm{sc}_P(\varphi_j)\right) \leq \sqrt{\frac{C \cdot \ln k}{n}}$$

1/2

Overfitting with adaptive analysis

- **Competitor's strategy (adaptive):**
  - Choose $k-1$ random classifiers $\varphi_1, \ldots, \varphi_{k-1}$
    Receive scores $a_1, \ldots, a_{k-1}$
  - Define $\varphi_k(z) = \mathrm{sign}\left(\sum_j \left(a_j - \frac{1}{2}\right) \cdot \varphi_j(z)\right)$

Deviation from population mean



95% significance threshold

Theorem (adaptive attack on raw scores):

$$\mathbb{E}\left(\mathrm{sc}_X(\varphi_k) - \mathrm{sc}_P(\varphi_k)\right) = \Omega\left(\sqrt{\frac{k}{n}}\right)$$

Overfitting with adaptive analysis

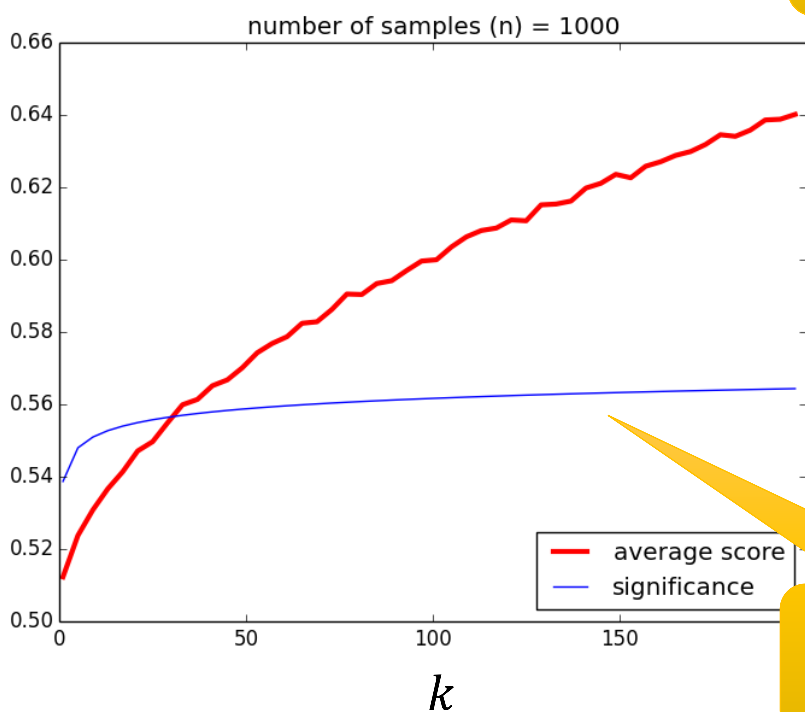
kaggle

- **Competitor's strategy (adaptive):**
  - Choose $k-1$ random classifiers $\varphi_1, \dots, \varphi_{k-1}$
    Receive scores $a_1, \dots, a_{k-1}$
  - Define $\varphi_k(z) = \text{sign}\left(\sum_j \left(a_j - \frac{1}{2}\right) \cdot \varphi_j(z)\right)$ $= \text{sign}\langle \vec{a} - \vec{\tfrac{1}{2}}, \vec{y} \rangle$

$\vec{y} = (\varphi_1(z), \dots, \varphi_{k-1}(z))$

φ_k "runs" a membership inference attack!

Deviation from population mean



95% significance threshold

Theorem (adaptive attack on raw scores):

$$\mathbb{E}\left(\text{sc}_X(\varphi_k) - \text{sc}_P(\varphi_k)\right) = \Omega\left(\sqrt{\frac{k}{n}}\right)$$

What Happened in This Example?

Case Study: ML Competitions

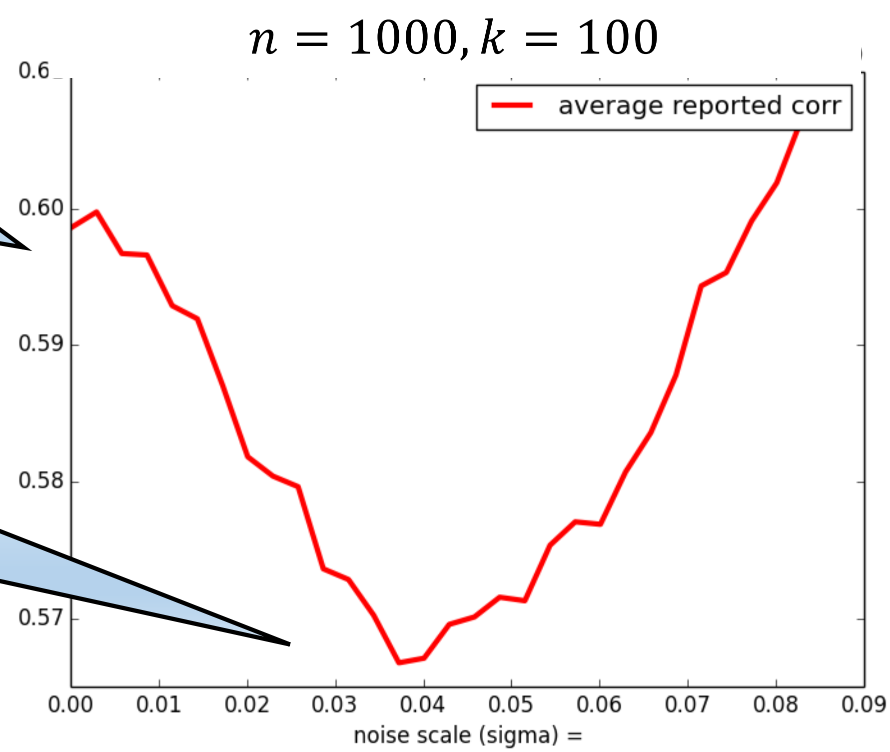
- **Improved estimator:** Add Gaussian noise $N(0, \sigma^2)$ to the estimated score of each classifier
  - Give answers $a_j = \mathrm{score}_X(\varphi_j) + N(0, \sigma^2)$

Case Study: ML Competitions

- **Improved estimator:** Add Gaussian noise $N(0, \sigma^2)$ to the estimated score of each classifier
  - Give answers $a_j = \text{score}_X(\varphi_j) + N(0, \sigma^2)$
  - The best choice of σ is not 0!

No noise: overestimate score by ≈0.10

Some noise: overestimate score by ≈0.06



Case Study: ML Competitions

- **Improved estimator:** Add Gaussian noise $N(0, \sigma^2)$ to the estimated score of each classifier
  - Give answers $a_j = \mathrm{score}_X(\varphi_j) + N(0, \sigma^2)$
  - The best choice of σ is not 0!

Minimized by
$\sigma =$,
achieving value

Theorem [DFHPRR'15, BNSSS**U**'16]**:** for appropriate $\sigma > 0$,

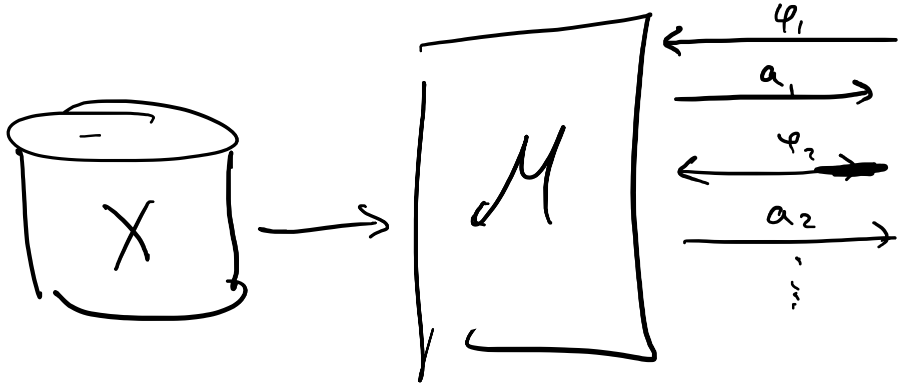
$$\mathbb{E}\left[\max_j a_j - \mathrm{score}_P(\varphi_j)\right] \lesssim \frac{\sqrt{k}}{n\sigma} + \sigma$$

overfitting

noise

- Compare to $O\left(\sqrt{k/n}\right)$ when $\sigma = 0$

General Setting



Queries: $\varphi : \mathcal{U} \to [0,1]$ (data universe)

Desired answer: $\mathop{\mathbb{E}}_{X' \sim P}(\varphi(X'))$

Goal: minimize $\left\{ \max_j \left| a_j - \mathop{\mathbb{E}}_{X^* \sim P}(\varphi(X')) \right| \right\}$

Proof Overview

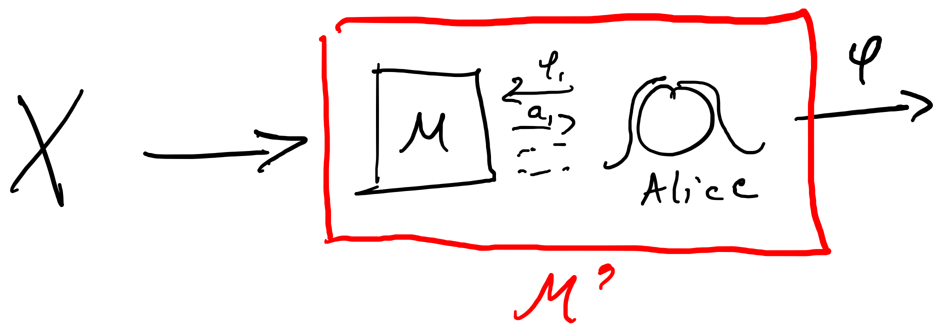
Key Claim: If M' is an (ε, δ)-DP mechanism that maps X to a classifier, then $\mathbb{E}_{X,M}\left[\text{score}_X\left(M'(X)\right)\right] - \mathbb{E}_{X,M}[\text{score}_P(M'(X))] = O(\varepsilon + \delta)$

How will we use this?

Proof Overview

Key Claim: If M' is an (ε, δ)-DP mechanism that maps X to a classifier, then $\mathbb{E}_{X,M}\left[\text{score}_X\left(M'(X)\right)\right] - \mathbb{E}_{X,M}[\text{score}_P(M'(X))] = O(\varepsilon + \delta) + \frac{1}{\sqrt{n}}$

How will we use this?



Say Alice is trying to find φ s.t. $\text{score}_X(\varphi) \gg \text{score}_P(\varphi)$

By POST-PROCESSING (!), M' is (ε, δ)-DP. if M is (ε, δ)-DP.

Proof Overview

Key Claim: If M' is an (ε, δ)-DP mechanism that maps X to a classifier, then $\mathbb{E}_{X,M}\left[\text{score}_X\left(M'(X)\right)\right] - \mathbb{E}_{X,M}[\text{score}_P(M'(X))] = O(\varepsilon + \delta)$

- Proof Sketch:
  - Consider $\left(i, X_i, M'(X)\right)$ and $\left(i, Z, M'(X)\right)$ where $i \sim [n]$, $X \sim P^n, Z \sim P$ independently, and M' is the mechanism

Proof Overview

Key Claim: If M' is an (ε, δ)-DP mechanism that maps X to a classifier, then $\mathbb{E}_{X,M}\left[\mathrm{score}_X\left(M'(X)\right)\right] - \mathbb{E}_{X,M}[\mathrm{score}_P(M'(X))] = O(\varepsilon + \delta)$

- Proof Sketch:
  - Consider $\left(i, X_i, M'(X)\right)$ and $\left(i, Z, M'(X)\right)$ where $i \sim [n]$, $X \sim P^n, Z \sim P$ independently, and M is the mechanism
    - **Sub-claim:** $\left(i, X_i, M'(X)\right) \approx_{\varepsilon,\delta} \left(i, Z, M'(X)\right)$
  - Observe that
    - $\mathbb{E}_{X,M}\left[\mathrm{score}_X\left(M'(X)\right)\right] = \mathbb{E}\left(f\left(i, X_i, M'(X)\right)\right)$
    - $\mathbb{E}_{X,M}[\mathrm{score}_P(M'(X))] = \mathbb{E}\left(f\left(i, Z, M'(X)\right)\right)$
    - Where $f(i, y, m) =$ ____________
  - **Fact:** If $A, B \in [0,1]$ satisfy $A \approx_{\varepsilon,\delta} B$, then $\mathbb{E}(A) \leq e^{\varepsilon}\mathbb{E}(B) + \delta.$

Transfer Theorem

Theorem: Let M be an (ε, δ)-DP mechanism for answering a sequence of k queries that is accurate on the sample, i.e.,

$$\Pr\left(\max_j \left|a_j - \text{score}_X(\varphi_j)\right| \leq \alpha\right) \geq 1 - \beta.$$

Then it is also accurate on the population:

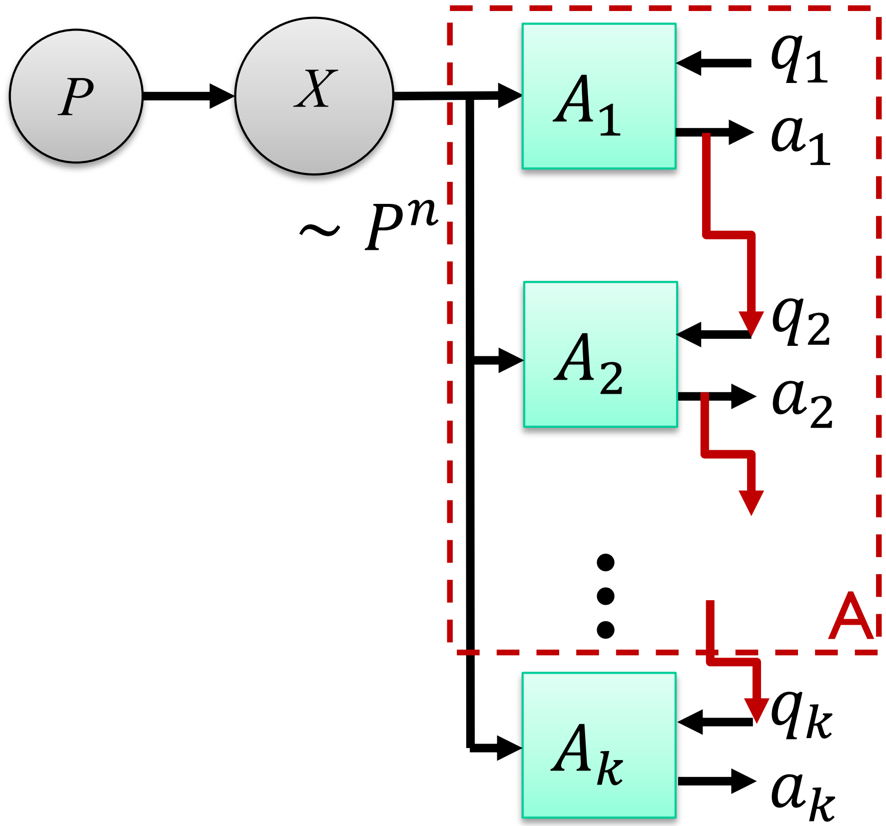
$$\Pr\left(\max_j \left|a_j - \text{score}_P(\varphi_j)\right| \leq \alpha + \varepsilon + \sqrt{\beta} + \sqrt{\delta}\right) \gtrsim 1 - \sqrt{\beta} - \sqrt{\delta}.$$

This result is sufficient to analyze the Gaussian mechanism, as well as others based on MW-EM, for example

See Jung, Ligett, Neel, Roth, Sharifi-Malvajerdi, Moshe Shenfeld, *ITCS 2020* for a nice proof.

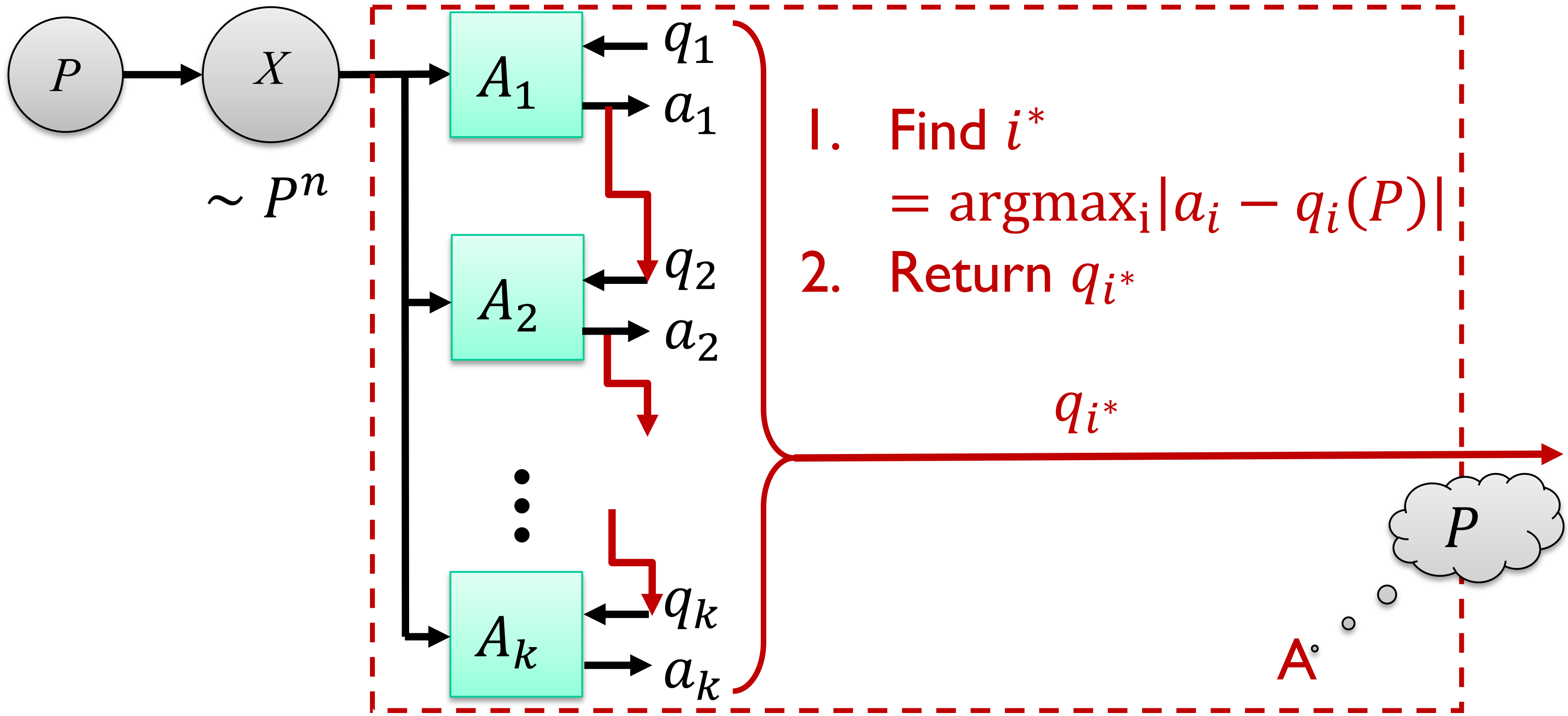
What happens with Many Queries?

From 2 to k stages: Induction [DFHPRR'15]



- Apply overfitting lemma at each round
  - Probability of overfitting adds up over rounds

"Monitor Argument" [BNSSSU'16]



Observation:

$$\epsilon \geq Score_X(q_{i^*}) - Score_P(q_{i^*}) \geq \max_i |a_i - q_{i(P)}| - \alpha$$

- Stronger bounds
- Generalizes beyond linear queries

Versions based on ...

- other variants of DP.

- measures of information

$$I(X; M(X))$$

$\approx \varepsilon\sqrt{n} + (\ldots)$ when M is DP and X i.i.d.

(also other measures).

→ gives results for arbitrary hypothesis tests.