NEU CS 7880 / BU CS 591: Privacy in ML and Statistics

Adam Smith (BU)
Jonathan Ullman (NEU)

Lecture 22: Adaptive Data Analysis
April 16 & 17, 2021

Statistical Theory

Method

↓

Sample

↓

Conclusions

Statistical analysis guarantees that your conclusions generalize to the population

Statistical Practice

PLOS | MEDICINE

OPEN ACCESS

ESSAY

1,140,912 VIEWS

1,413 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

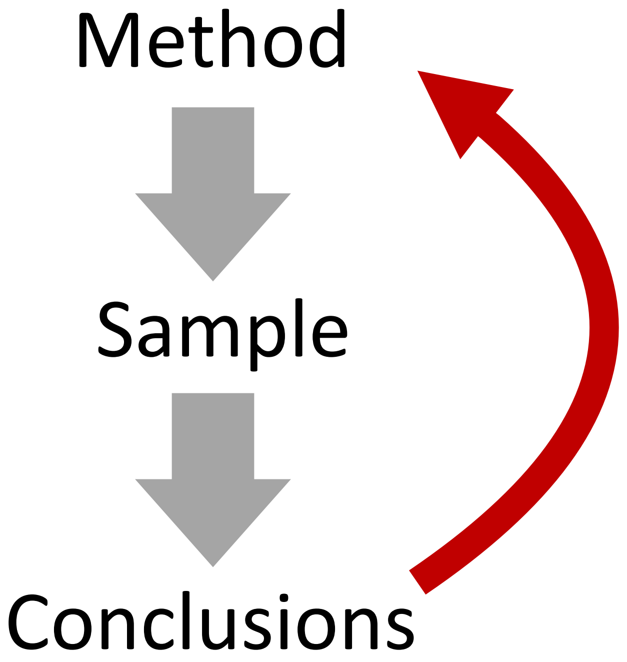
Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

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

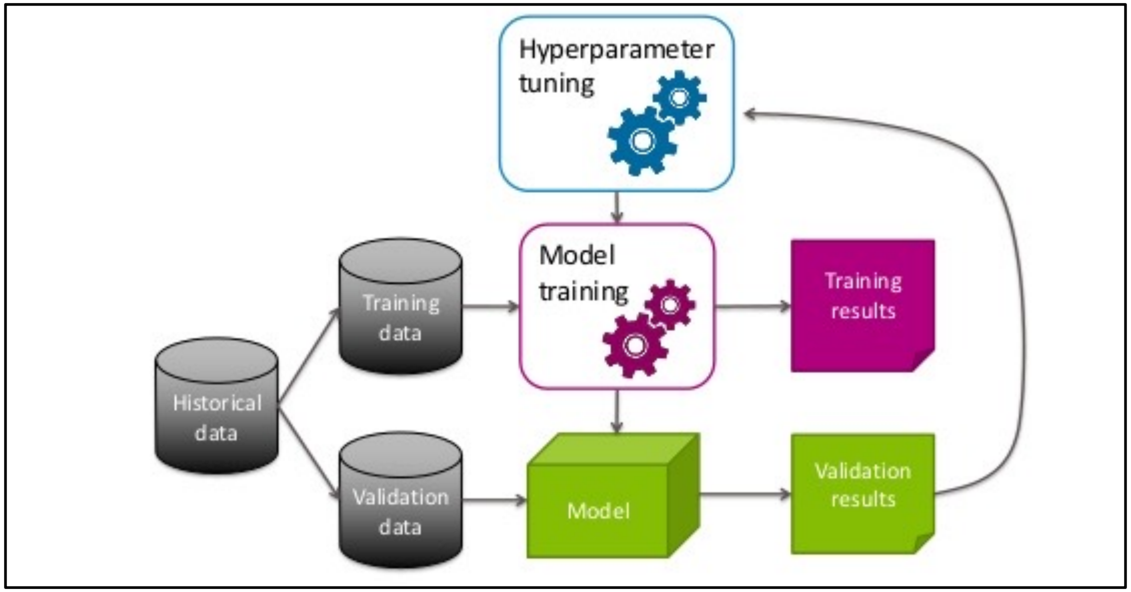


Statistical guarantees no longer apply
when the method and sample are correlated

Examples of Adaptive Data Analysis

Well specified adaptive algorithms

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



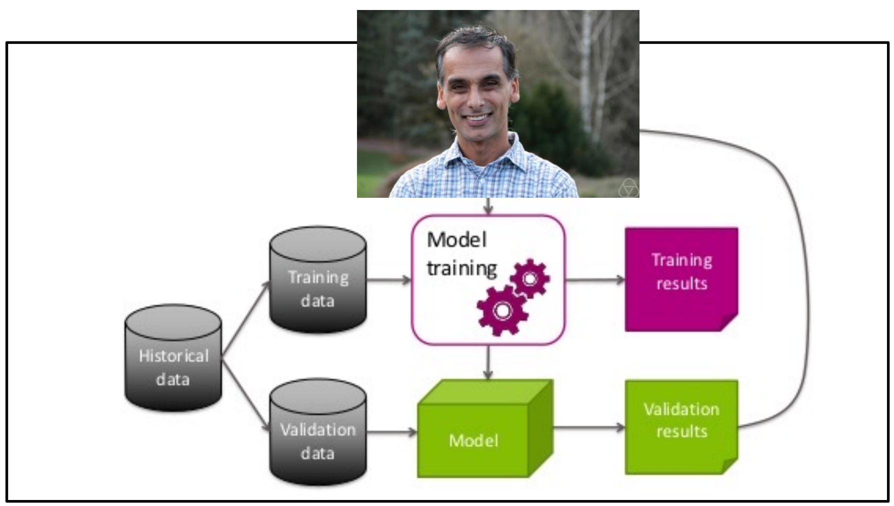
Alice Zheng. "Evaluating Machine Learning Models."

Examples of Adaptive Data Analysis

Researcher degrees of freedom

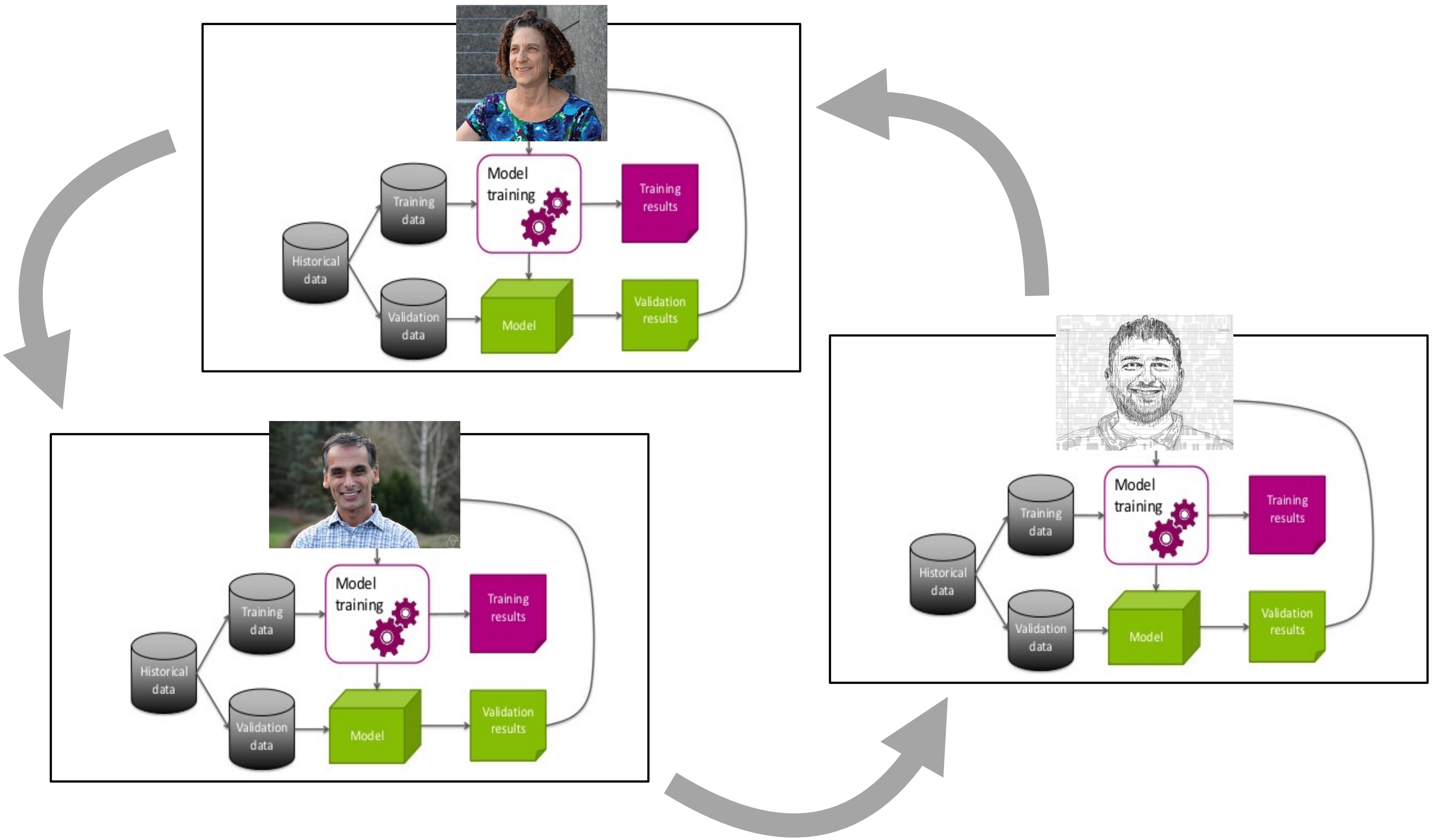
The interaction effect is not significant when the scale from the Danish study are used to gauge the US subjects' support for redistribution. This arises because two of the items are somewhat unreliable in a US context. Hence, for items 5 and 6, the inter-item correlations range from as low as .11 to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

secret **validation data** X

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

Classifier φ

Repeat k times

Answer a

Data Gladiator

Public Data

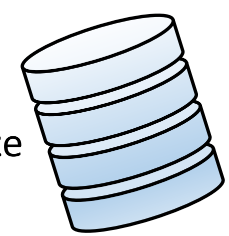
$$\text{score}_X(\varphi) = \frac{1}{n}\sum_i \mathbf{1}\{\varphi(z_i) = s_i\}$$

Goal: design a method for estimating the score **on the prize data**

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

$\text{score}_P(\varphi) =$ score on the prize data

Secret Prize Data P



Same distribution as validation data

Case Study: ML Competitions

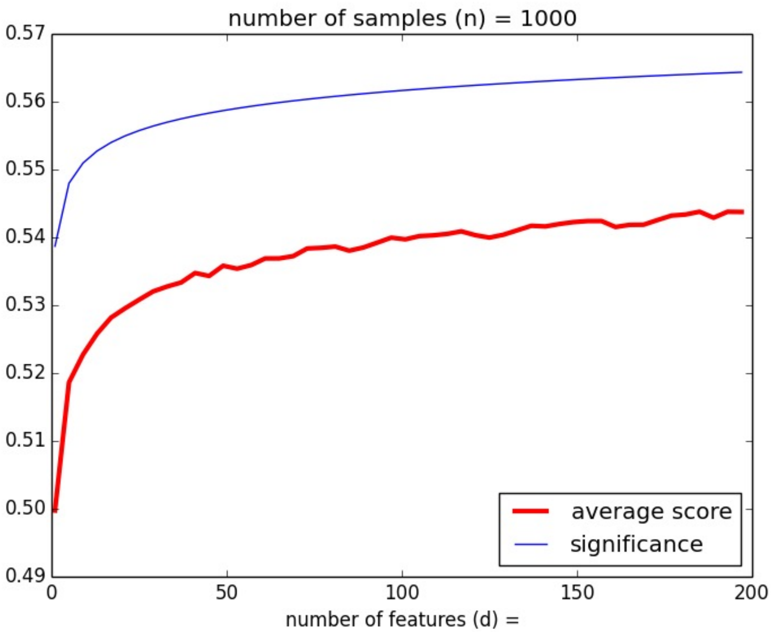
- Suppose prize and validation data have **random labels**
  - Any classifier will have $\mathbb{E}[\mathrm{score}_P(\varphi)] = \frac{1}{2}$ on the prize data
  - If $\mathrm{score}_X(\varphi) \gg \frac{1}{2}$ then we have overfit

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

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

Non-adaptive analysis

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

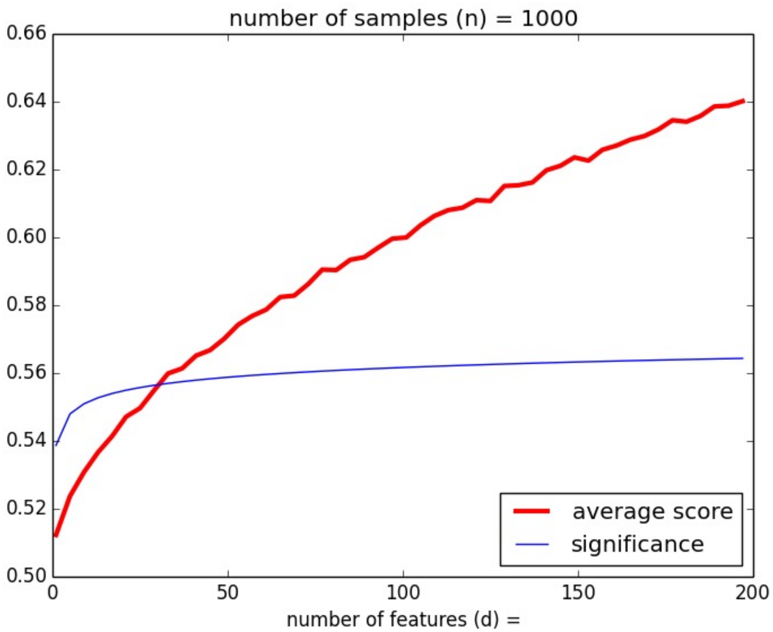


Theorem:

$$\max_j \operatorname{sc}_X(\varphi_j) - \operatorname{sc}_P(\varphi_j) \leq \sqrt{\frac{C \cdot \ln k}{n}}$$

Overfitting with adaptive analysis

- **Competitor's strategy (adaptive):**
  - Choose k random classifiers $\varphi, \dots, \varphi_{k-1}$ get scores $\mathrm{score}_1, \dots, \mathrm{score}_{k-1}$
  - Define $\varphi_k(z) = \mathrm{sign}\left(\sum_j \left(\mathrm{score}_j - \frac{1}{2}\right) \cdot \varphi_j(z)\right)$



Theorem:

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

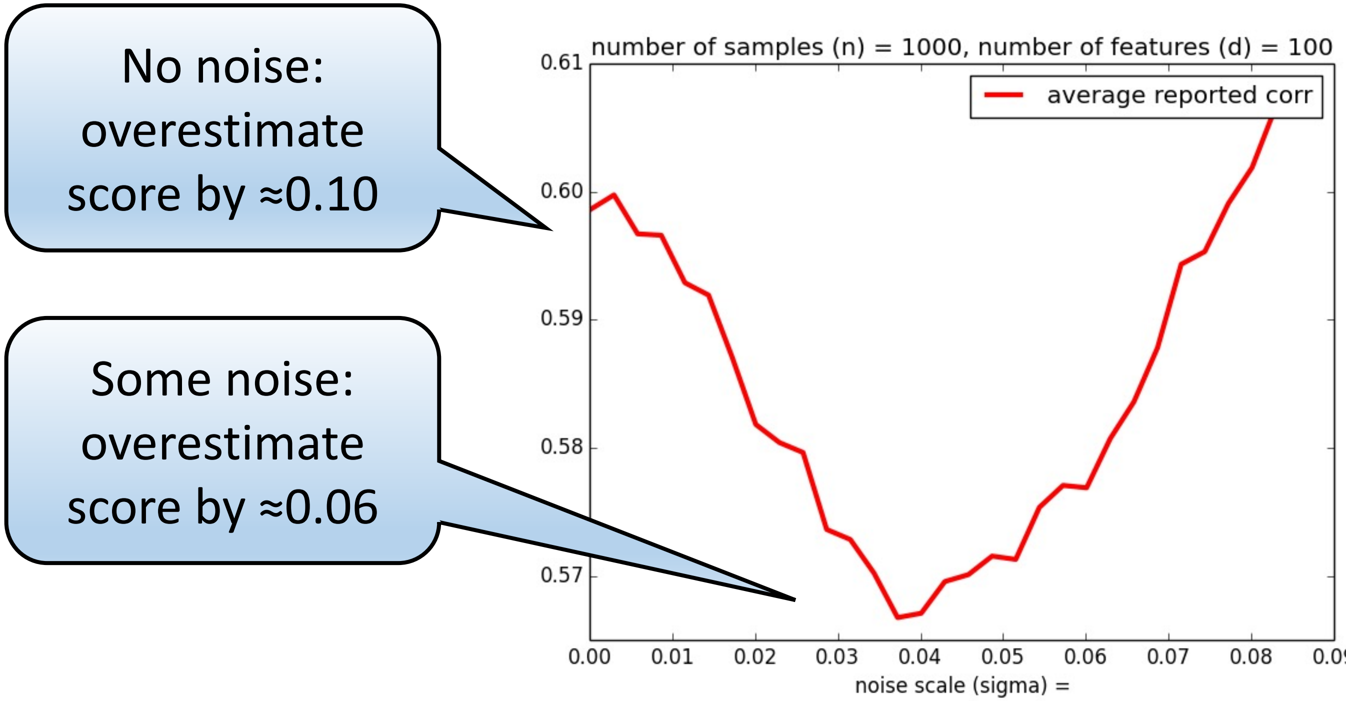
What Happened in This Example

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 $a_j = \text{score}_X(c_j) + N(0, \sigma^2)$
  - The best choice of σ is not 0!



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!

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

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

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

Proof Overview

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

- 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

$$\left(i, X_i, M(X)\right)$$

$$\approx_\varepsilon \left(i, X_i, M(Z||X_{-i})\right)$$ Differential Privacy

$$= \left(i, Z, M(X_i||X_{-i})\right)$$ Symmetry

$$= \left(i, Z, M(X)\right)$$