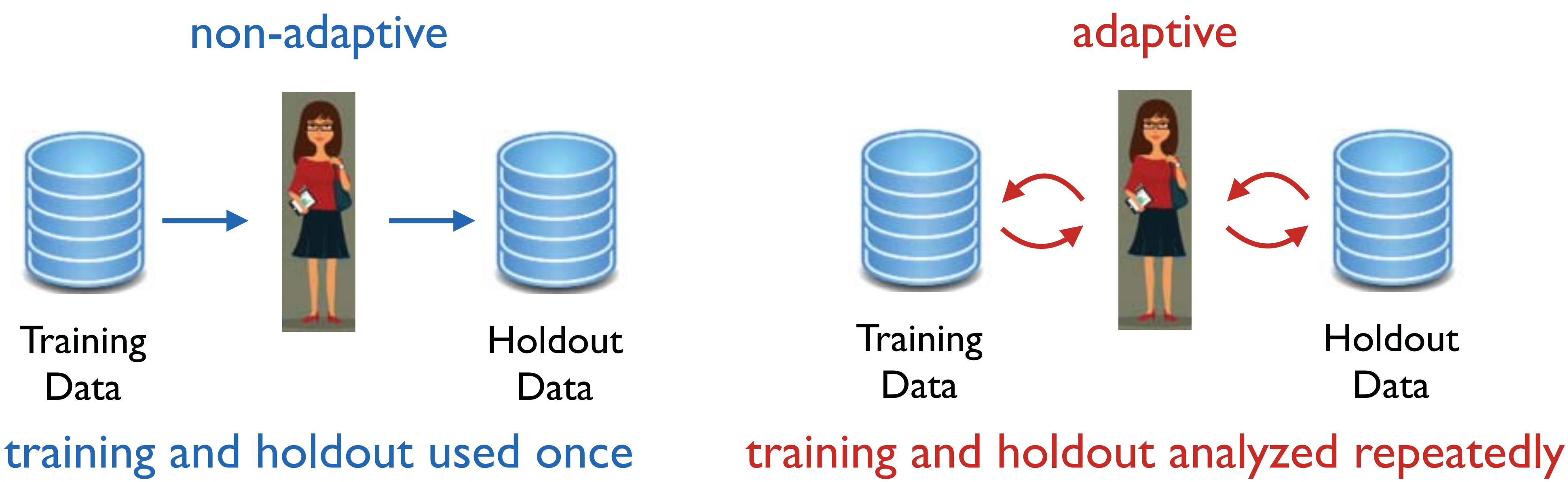
Preventing Overfitting in Adaptive Data Analysis

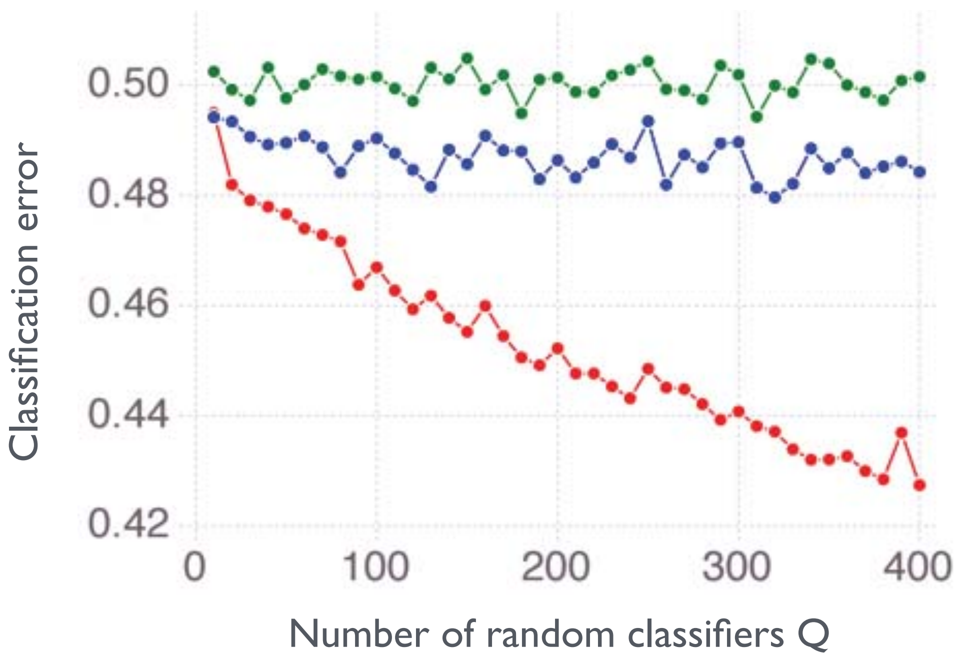


Jonathan Ullman
Northeastern University
College of Computer and Information Science

Based on a body of work by Raef Bassily, Avrim Blum, Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Kobbi Nissim, Toniann Pitassi, Omer Reingold, Aaron Roth, Adam Smith, Thomas Steinke, Uri Stemmer, and myself.

(Not) Overfitting in Adaptive Data Analysis

- Classical techniques for preventing overfitting do not work when the same training data is analyzed adaptively.
- This problem is pervasive, robust, and hard to solve.
- Perturbing the results of the analysis can help.
- Work may be adaptable to preventing over-training when testing automatic threat recognition (ATR) algorithms



Suppose you try to classify purely random data…

1. Submit Q completely random classifiers
2. Receive training error for each one
3. New classifier is the majority of the random classifiers that did "well"

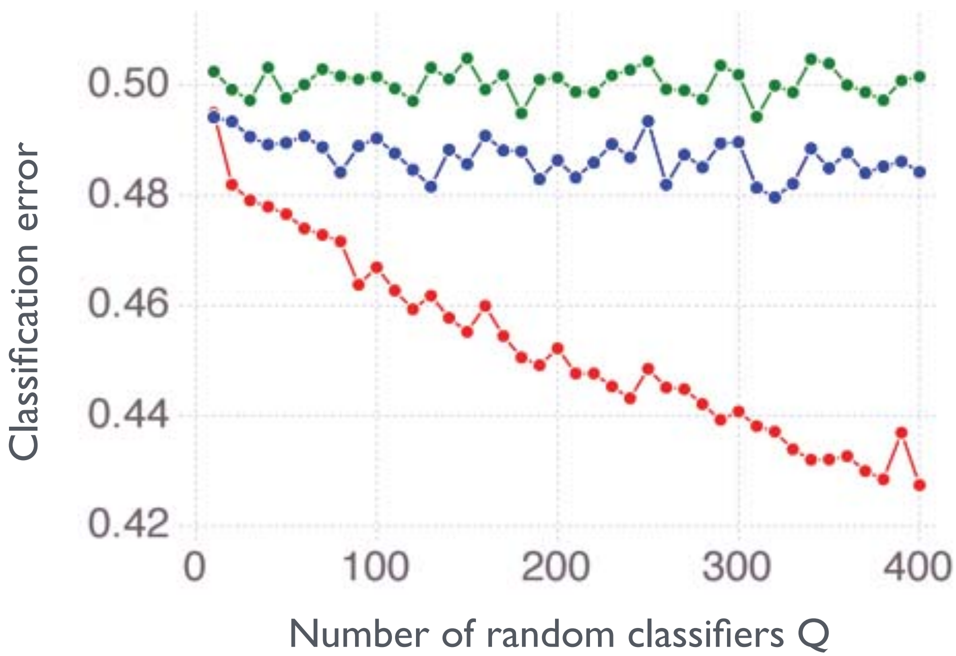
How well will you classify…

…the training data? …fresh data?

Graph reproduced from
Moritz Hardt, "Competing in a data science contest without reading the data."

(Not) Overfitting in Adaptive Data Analysis

- Classical techniques for preventing overfitting do not work when the same training data is analyzed adaptively.
- This problem is pervasive, robust, and hard to solve.
- Perturbing the results of the analysis can help.
- Work may be adaptable to preventing over-training when testing automatic threat recognition (ATR) algorithms



Suppose you try to classify purely random data…

1. Submit Q completely random classifiers
2. Receive noisy training error for each one
3. New classifier is the majority of the random classifiers that did "well"

How well will you classify…

…the training data? …fresh data?

Graph reproduced from
Moritz Hardt, "Competing in a data science contest without reading the data."

Problem is pervasive, hard to solve

- Not the result of malice, error, or p-hacking
  - Overfitting occurs even if you faithfully, correctly apply classical statistical significance testing and control for multiple comparisons.
- "Garden of Forking Paths" (Gelman, Lokem)
  - Identifies numerous examples of unreproducible studies where data was analyzed in an adaptive fashion
- Competitors in Kaggle competitions (classification contests) frequently report that the "leaders" of the competition perform substantially in final test
  - Leaders are based on a single holdout set used for all submissions
  - Final scores are based on a new holdout set

Problem is pervasive, hard to solve

- Injecting noise into what you reveal about the dataset can help prevent overfitting
  - For example, don't report exact performance on the training data
- But, there won't be a one size fits all solution
  - In many scenarios, can overfit even with very noisy responses*
- Rest of the talk: combatting overfitting in "classification competitions"
  - We have found many more scenarios where we can (or can hope to) prevent overfitting

*Moritz Hardt and **U**, "Preventing false discovery in interactive data analysis is hard."
*Thomas Steinke and **U**, "Interactive fingerprinting codes and the hardness of preventing false discovery."

Classification Competitions and the Ladder

- How does kaggle run machine learning competitions?

public data

secret
"leaderboard data"

unseen
"fresh data"

submit classifier,
receive score

public leaderboard

#	Δ1w	Team Name * in the money	Score	Entries
1	↑2	Shen(Deep)-Jing(Whale) *	3.84733	11
2	↓1	Whalearning *	3.85244	25
3	↓1	deepsense.io *	3.87751	12

prizes are awarded based fresh data
scores typically worse

*Avrim Blum, Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions"
*Bassily, Nissim, Smith, Stemmer, Steinke, **U**, "Algorithmic Stability for Adaptive Data Analysis"
*Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, "Preserving Statistical Validity in Adaptive Data Analysis"

Classification Competitions and the Ladder

- How does kaggle run machine learning competitions?

public data

secret
"leaderboard data"

unseen
"fresh data"

submit classifier,
receive noisy score

noisy public leaderboard

#	Δ1w	Team Name *in the money	Score	Entries
1	↑2	Shen(Deep)-Jing(Whale) *	3.84733	11
2	↓1	Whalearning *	3.85244	25
3	↓1	deepsense.io *	3.87751	12

only report if classifier is better than all previous classifiers

scores on fresh data <u>guaranteed</u> to be similar (± $\log(k)/n^{1/3}$) to leaderboard scores

*Avrim Blum, Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions"
*Bassily, Nissim, Smith, Stemmer, Steinke, **U**, "Algorithmic Stability for Adaptive Data Analysis"
*Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, "Preserving Statistical Validity in Adaptive Data Analysis"

Classification Competitions and the Ladder

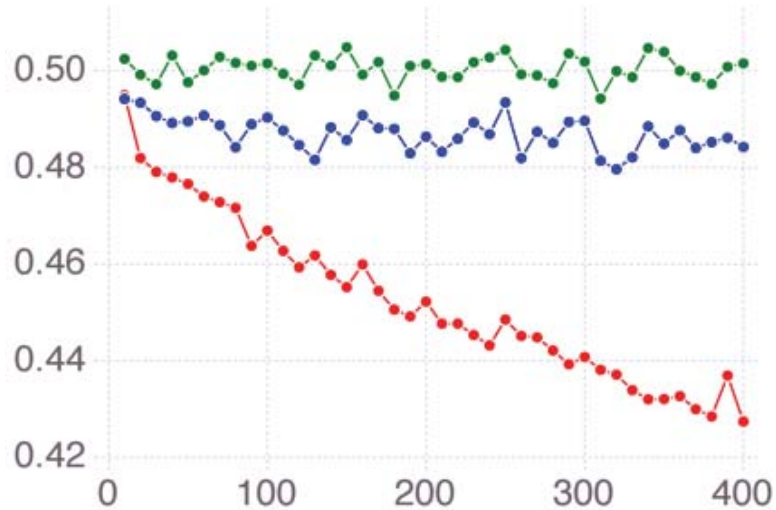
- How does kaggle run machine learning competitions?

public data

secret
"leaderboard data"

unseen
"fresh data"

submit classifier,
receive noisy score



fresh data

leaderboard data,
noisy scores

leaderboard data,
exact scores

only report if classifier is better than all previous classifiers

Scores on fresh data <u>guaranteed</u> to be similar to leaderboard scores

*Avrim Blum, Moritz Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions"
*Bassily, Nissim, Smith, Stemmer, Steinke, **U**, "Algorithmic Stability for Adaptive Data Analysis"
*Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, "Preserving Statistical Validity in Adaptive Data Analysis"

Conclusion: fixing adaptive analysis

- Classical techniques for preventing overfitting do not work when the same training data is analyzed adaptively.
- This problem is pervasive, robust, and hard to solve.
- Perturbing the results of the analysis can help.
- Work may be adaptable to preventing over-training when testing automatic threat recognition (ATR) algorithms

Thank you.