TEST DATA REUSE FOR EVALUATION OF ADAPTIVE MACHINE LEARNING ALGORITHMS: OVER-FITTING TO A FIXED "TEST" DATASET AND A POTENTIAL SOLUTION

ALEXEJ GOSSMANN (TULANE UNIVERSITY), ARIA PEZESHK, AND BERKMAN SAHINER (U.S. FOOD AND DRUG ADMINISTRATION)

FEBRUARY 11, 2018

Are machine learning algorithms and statistical models ever trained independently of previous (exploratory) analyses on the same data?

In medical research?

Doesn't that inadvertently lead to overfitting or false discoveries?

Can machine learning algorithms be allowed to evolve after deployment?

What if the data is partially reused to re-train and to re-test the algorithm as it evolves?

What if the ML algorithm is utilized in clinical practice in medicine?

SOME KEYWORDS

- Adaptive data analysis
- Adaptive machine learning
- Continuous machine learning
- Online machine leaning
- Life-long machine learning
- "Researcher degrees of freedom"
- "A garden of forking paths" (Gelman and Loken, 2014)

To avoid overfitting and false findings, **ideally** one would use a **fresh new dataset** each time...

- ...a machine learning algorithm is trained or re-trained or fine-tuned based on a previously trained algorithm or model
- ...a new data analysis step is informed by previous data analysis steps
- ...the performance of a trained algorithm or model is evaluated

$\Longrightarrow$ But that's **impractical** in most cases!

In Machine Learning practice, generally, usage of two independent datasets — *"training"* data and *"test"* data.

- **Training data**: exploratory analysis, model fitting, parameter tuning, comparison of different machine learning algorithms, feature selection, etc.

  $\Longrightarrow$ Adaptive machine learning, risk of overfitting.

- **Test data**: Performance evaluation *after the trained machine learning algorithm has been "frozen"*.

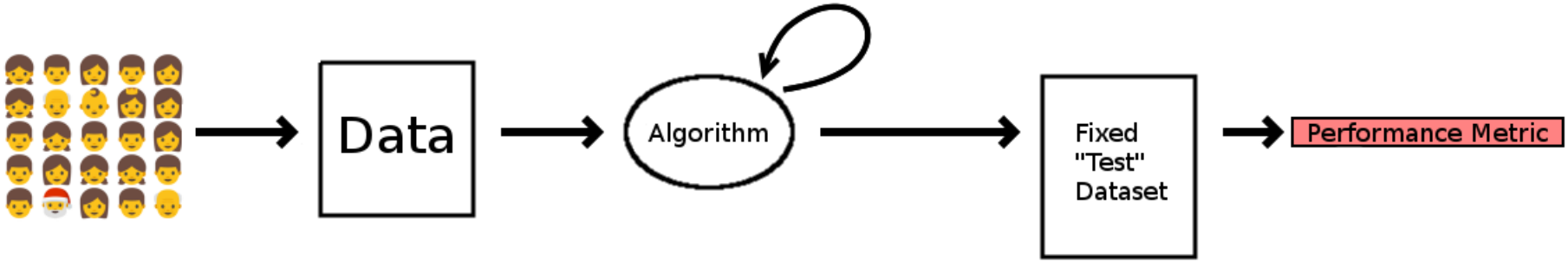
  $\Longrightarrow$ Accurate performance measures of the final model, **if the test data is used only once**.

CONTINUOUS MACHINE LEARNING

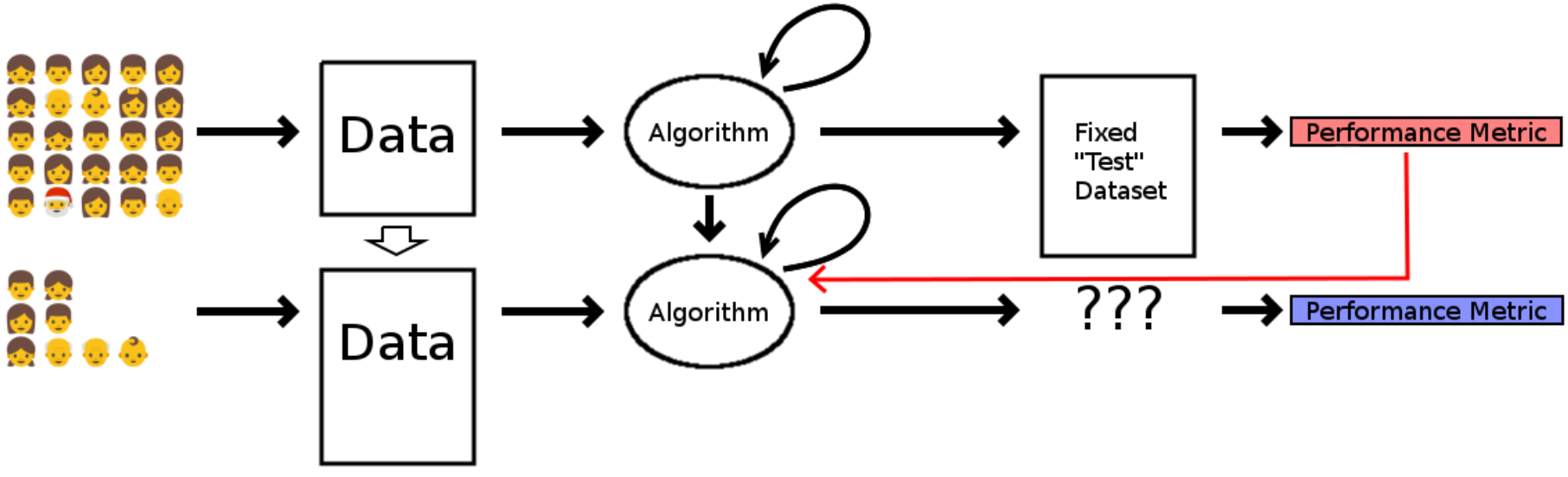
What if we want to *continue training and testing* the machine learning algorithm, *after a performance measure has been obtained* from the test dataset?

- Availability of new training samples after deployment in clinical practice. But no new test samples.
- Low quality data useful for training of ML alg, but test data needs to meet high quality requirements (representative distribution, labeling by experts, etc.).

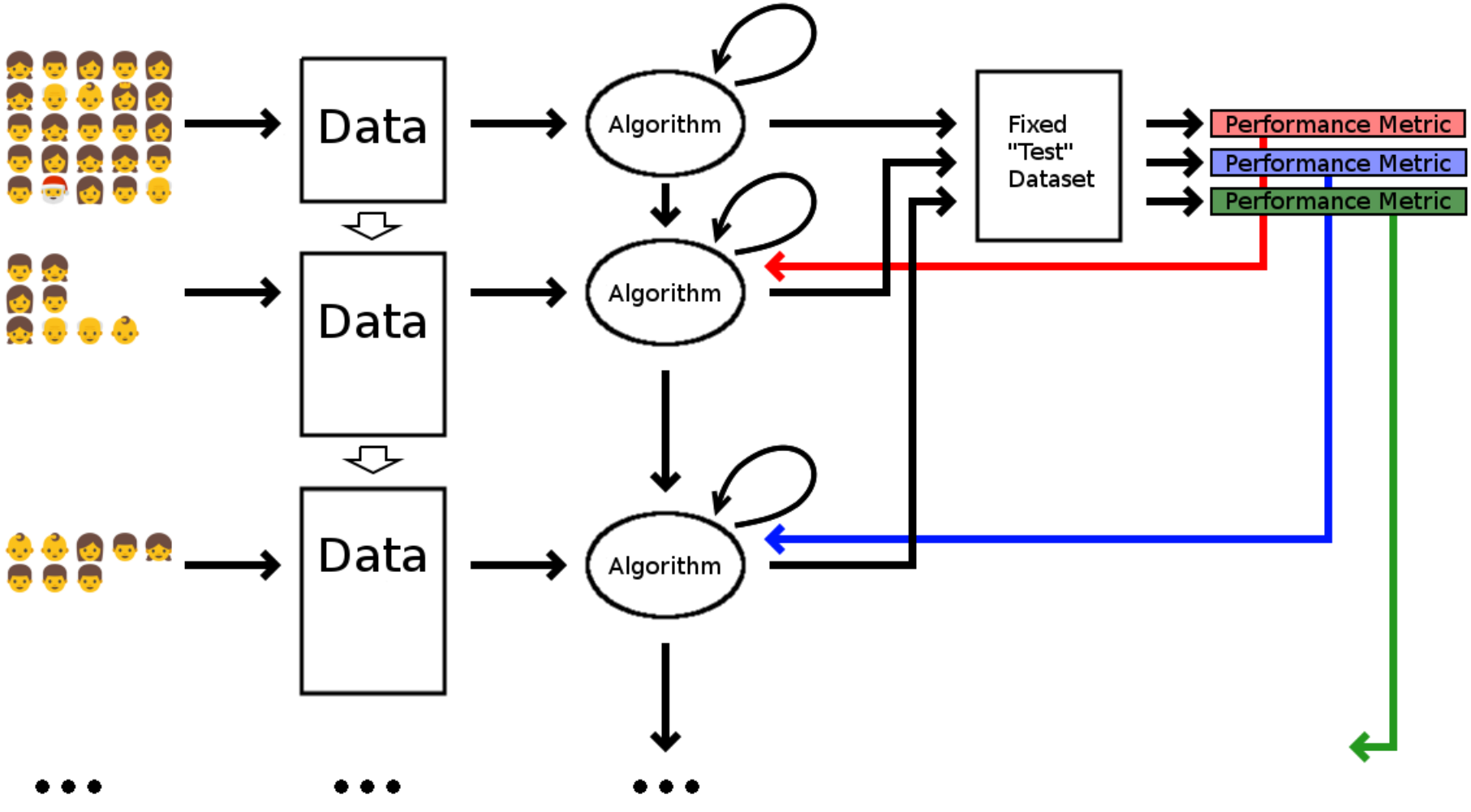
General machine learning process



"Adaptive" machine learning



"Adaptive" machine learning with test data reuse



PERFORMANCE ASSESSMENT IN ADAPTIVE MACHINE LEARNING WITH TEST DATA REUSE

Reuse of test data inadvertently leads to **problems**:

- **Overly optimistic** performance assessments
- **Loss of generalization** — i.e., ML alg. that performs much better on the available test data than on the population from which the data were drawn, a.k.a. **overfitting** to the test dataset.

PERFORMANCE ASSESSMENT IN ADAPTIVE MACHINE LEARNING WITH TEST DATA REUSE

Can we **obfuscate** the test data to avoid overfitting?

$\Longrightarrow$ Recent techniques based on *differential privacy* or *bounded description length* are taking that approach.

(Dwork, Feldman, Hardt, Pitassi, Reingold, Roth, *NIPS* 2015, *STOC* 2015, *Science* 2015; Bassily, Nissim, Smith, Steinke, Stemmer, Ullman, *STOC* 2016; Blum, Hardt, *ICML* 2015; + several follow-up papers since then)

DIFFERENTIAL PRIVACY (DWORK, MCSHERRY, NISSIM, SMITH, 2006)

- A mathematically rigorous definition of data privacy.
- **Intuition:** An individual data point has little impact on the value reported by a differentially private data-releasing mechanism.
- **Intuition:** An adversary cannot learn an individual data point from querying a differentially private data-releasing mechanism.

DIFFERENTIAL PRIVACY (DWORK, MCSHERRY, NISSIM, SMITH, 2006)

Let $\mathcal{M}$ be a (randomized) data access mechanism. $\mathcal{M}$ is (ε, δ)-**differentially private** if for any two datasets D and D' *differing in one observation*, and for all sets $S \in \mathrm{Range}(\mathcal{M})$, it holds that

$$P[\mathcal{M}(D) \in S] \leq e^{\varepsilon} P[\mathcal{M}(D') \in S] + \delta.$$

(Probability is taken with respect to randomness in $\mathcal{M}$.)

DIFFERENTIAL PRIVACY (DWORK, MCSHERRY, NISSIM, SMITH, 2006)

- A mathematically rigorous definition of data privacy.
- **Intuition:** An individual data point has little impact on the value reported by a DP mechanism.
- **Intuition:** An adversary cannot learn an individual data point from querying a DP mechanism.
- **Properties:** DP is *preserved* under *post-processing* and under *adaptive composition*.

DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

A possible solution to the test data reuse problem?

- **Intuition:** If the test dataset can be accessed only via a DP mechanism, then the ML alg. will have no way to extract information about individual dataset records, but it can only learn characteristics of the underlying distribution.

DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

- Limits/randomizes/obfuscates the information learned about the test data in each analysis.
- The ML alg. is less likely to overfit because it cannot adapt to individual records in the test dataset.
- Under (restrictive) theoretical conditions such techniques have provable **generalization guarantees**, even if the test dataset is reused thousands of times (e.g., Dwork et. al., Science, 2015).
- In practice the reported performance metrics can be much more accurate even when the theoretical conditions are not met (e.g., this work).

DIFFERENTIALLY PRIVATE ACCESS TO TEST DATA

- Currently available literature focuses on theory.
- Available theoretical requirements too restrictive for most of applied data analysis and machine learning.
- Computational experiments available in the literature consider only simple instances of adaptivity, simple performance metrics, and simple machine learning algorithms — not capturing the reality of current data analysis practices adequately.

IN THIS WORK:

1. Combining the *Thresholdout* procedure (DFHPRR, Science, 2015) with *AUC* (area under the ROC curve) as the reported performance metric.

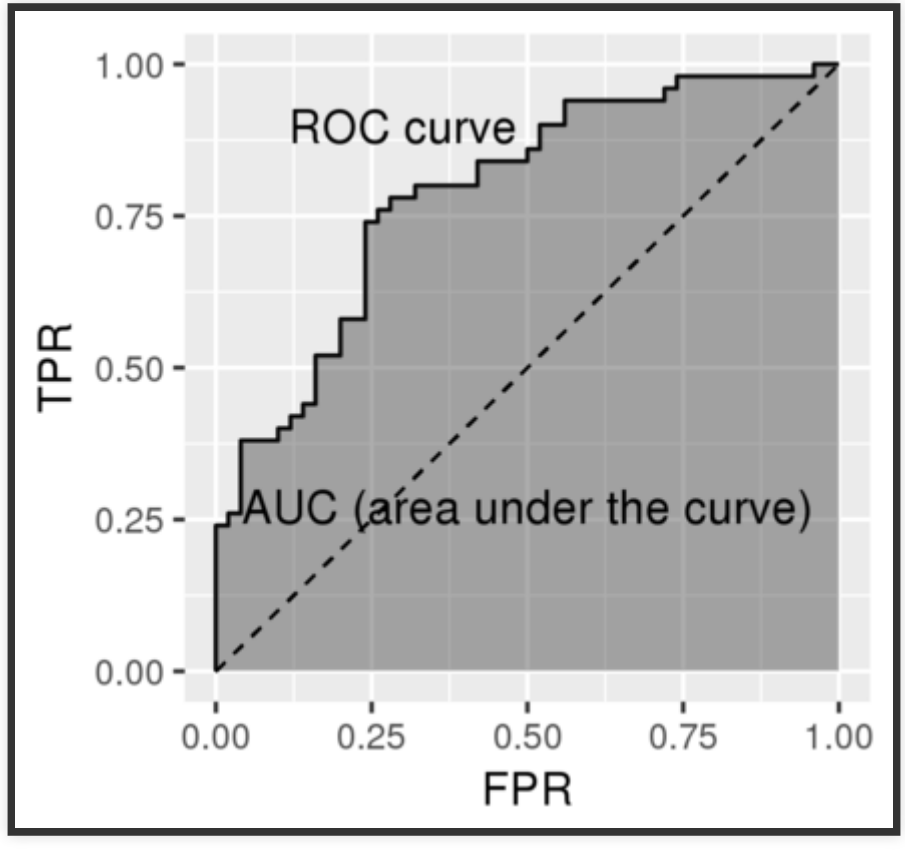
   $\rightsquigarrow$ Thresholdout$_{\mathrm{AUC}}$.

2. Empirical investigation of Thresholdout$_{\mathrm{AUC}}$ by simulation of realistic adaptive data analysis practices.

WHAT IS AUC?

- Dataset S: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \{0, 1\}$.
- A classification algorithm, trained on S, assigns a score to any $\mathbf{x}_* \in \mathbb{R}^p$ signifying how likely $y_* = 1$.
- A threshold or (*operating point*) can be chosen to assign class labels (0 or 1), to balance how many false positives and false negatives will be obtained.
- Plotting the true positive rate (TPR) against the false positive rate (FPR) as the operating point changes from its minimum to its maximum value yields the *receiver operating characteristic (ROC) curve*.

WHAT IS AUC?



Example of an (empirical) ROC curve.

WHY AUC?

Many advantages compared to other "single number" performance measures:

- Invariance to prior class probabilities or class prevalence in the data.
- Meaning: probability of a correct ranking of a random "positive"-"negative" pair of observations.
- Independence of the decision threshold.
- Can choose/change a decision threshold based on cost-benefit analysis after model training.
- Extensively used in the medical field, including medical imaging.

THRESHOLDOUT + AUC = 

Thresholdout$_{\mathrm{AUC}}$ combines the original **Thresholdout (DFHPRR, Science, 2015)** with **AUC** as the reported performance metric on test data.

Algorithm 1 $\text{Thresholdout}_{\text{AUC}}$

Require: Training dataset S_{train}, test dataset S_{test}, noise rate σ, budget B, threshold T.

Sample $\gamma \sim \text{Lap}(2\sigma)$ ▷ where $\text{Lap}(2\sigma)$ denotes the Laplace distribution with mean 0 and scale parameter 2σ

$\widehat{T} \leftarrow T + \gamma$

for each scoring function $\phi : \mathcal{X} \to [0, 1]$ **do**

  if $B < 1$ **then**

    OUTPUT($\perp$) ▷ i.e., the test data access budget B is exhausted

  else

    Sample $\xi \sim \text{Lap}(\sigma), \gamma \sim \text{Lap}(2\sigma), \eta \sim \text{Lap}(4\sigma)$

    if $\left|\widehat{\text{AUC}}_{S_{\text{test}}}(\phi) - \widehat{\text{AUC}}_{S_{\text{train}}}(\phi)\right| > \widehat{T} + \eta$ **then**

      $B \leftarrow B - 1$

      $\widehat{T} \leftarrow T + \gamma$

      OUTPUT($\widehat{\text{AUC}}_{S_{\text{test}}}(\phi) + \xi$)

    else

      OUTPUT($\widehat{\text{AUC}}_{S_{\text{train}}}(\phi)$)

    end if

  end if

end for

$\text{Thresholdout}_{\text{AUC}}$ combines the original Thresholdout (DFHPRR, Science, 2015) with AUC as the reported performance metric on test data.

THEOREM

THEOREM 2.1. *Let $\beta, \tau > 0$ and $m \geq B > 0$. Set $T = \frac{3\tau}{4}$ and $\sigma = \frac{\tau}{\left(96 \ln\left(\frac{4m}{\beta}\right)\right)}$. Let $\mathbf{S}$ denote a test dataset of size n drawn i.i.d. from a distribution $\mathcal{P}$ over $\mathcal{X}$, and let S_{train} be any additional dataset over $\mathcal{X}$. Assume that the class balance within $\mathbf{S}$ is such that*

$$\max\left\{\frac{n_{neg}}{n_{pos}}, \frac{n_{pos}}{n_{neg}}\right\} \leq M - 1,$$

for some $M \geq 2$. Consider an algorithm that is given access to S_{train} and adaptively chooses functions $\phi_1, \ldots, \phi_m : \mathcal{X} \to [0,1]$ while interacting with $\text{Thresholdout}_{\text{AUC}}$ which is given $\mathbf{S}, S_{train}, \sigma, B, T$. For every $i = 1, 2, \ldots, m$, let $\mathbf{a}_i$ denote the answer of $\text{Thresholdout}_{\text{AUC}}$ on function ϕ_i, and let $\mathbf{Z}_i$ be the counter of overfitting defined by

$$\mathbf{Z}_i := \left|\left\{j \leq i : \left|\text{AUC}(\phi_j) - \widehat{\text{AUC}}_{S_{train}}(\phi_j)\right| > \frac{\tau}{2}\right\}\right|.$$

It holds that if $n \geq \max\left\{\frac{16BM^2}{\sigma\tau}, \frac{64M^2}{\tau^2}\ln\left(\frac{12m}{\beta}\right)\right\} = \mathcal{O}\left(\frac{M^2}{\tau^2}\ln\left(\frac{m}{\beta}\right)\right) \cdot B$, then

$$\begin{aligned}\mathbb{P}\left[\exists i \in \{1, \ldots, m\} : \mathbf{Z}_i < B \,\&\, |\mathbf{a}_i - \text{AUC}(\phi_i)| \geq \tau\right] &\leq \mathbb{P}\left[\exists i \in \{1, \ldots, m\} : \mathbf{a}_i \neq \perp \,\&\, |\mathbf{a}_i - \text{AUC}(\phi_i)| \geq \tau\right] \\ &\leq \beta.\end{aligned}$$

$\text{Thresholdout}_{\text{AUC}}$ generalization guarantees (proof similar to DFHPRR '15 arXiv:1506.02629).

THEOREM – ROUGH SUMMARY

1. **If** a test dataset, which is used for performance evaluation repeatedly, is only accessed via Thresholdout$_{\mathrm{AUC}}$. $\Longrightarrow$ **Then** with a high probability $(1-\beta)$ the reported AUC estimates will be correct up to a small tolerance τ.
2. **Restriction**: Test data access "budget" B, which is linear in the size of the test data n, and also depends on β, τ, and the class balance.

THEOREM – DRAWBACKS

1. Required test data size, n, too large for most applications.
2. Thresholdout is designed for the worst case of an adversarial analyst.

THEOREM – DRAWBACKS

Will the Thresholdout$_{\mathrm{AUC}}$ procedure still work, if the test data is small, but the analyst is not adversarial, and is in fact interested in the avoidance of overfitting?

SIMULATION STUDIES ON SMALL SAMPLES

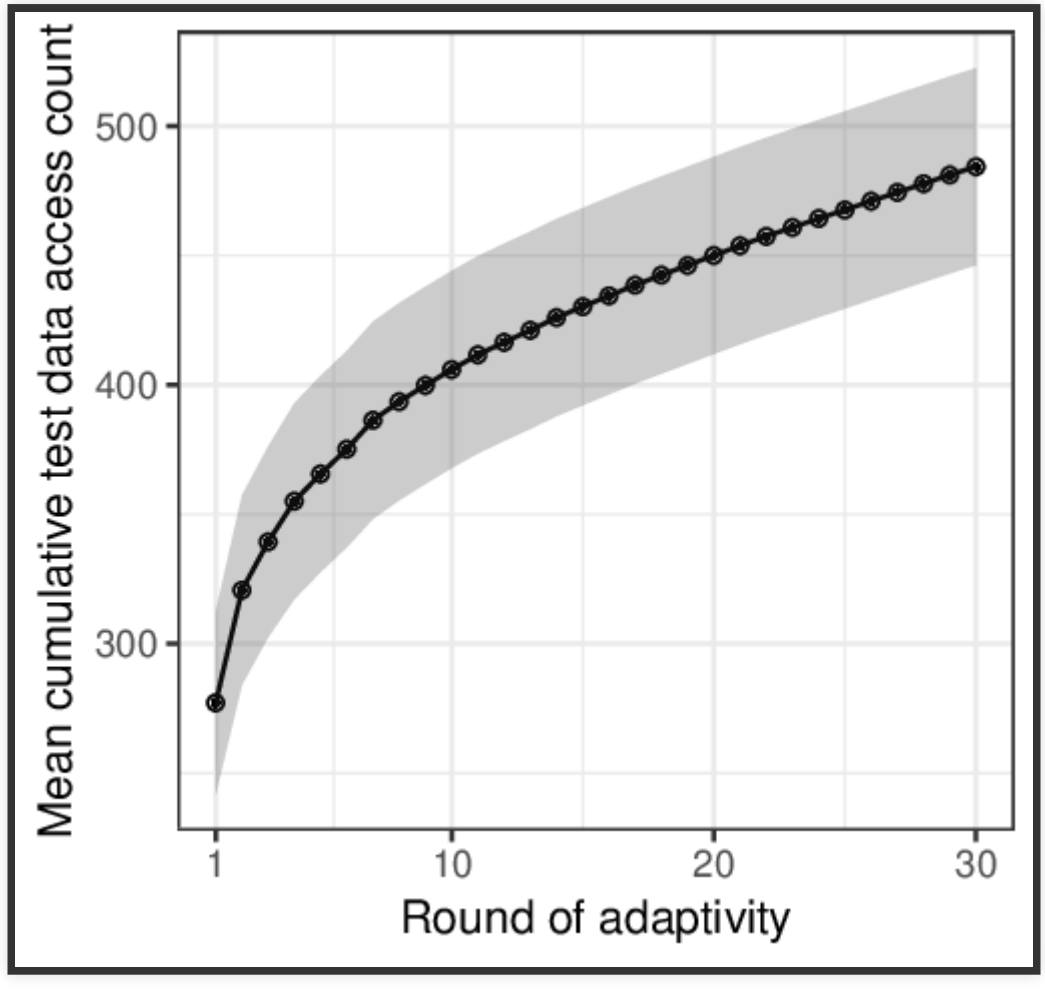
SIMULATION STUDIES ON SMALL SAMPLES

Goal: Compare Thresholdout$_{\mathrm{AUC}}$ to a "naive" test data reuse approach under conditions that mimic realistic adaptive data analysis practices.

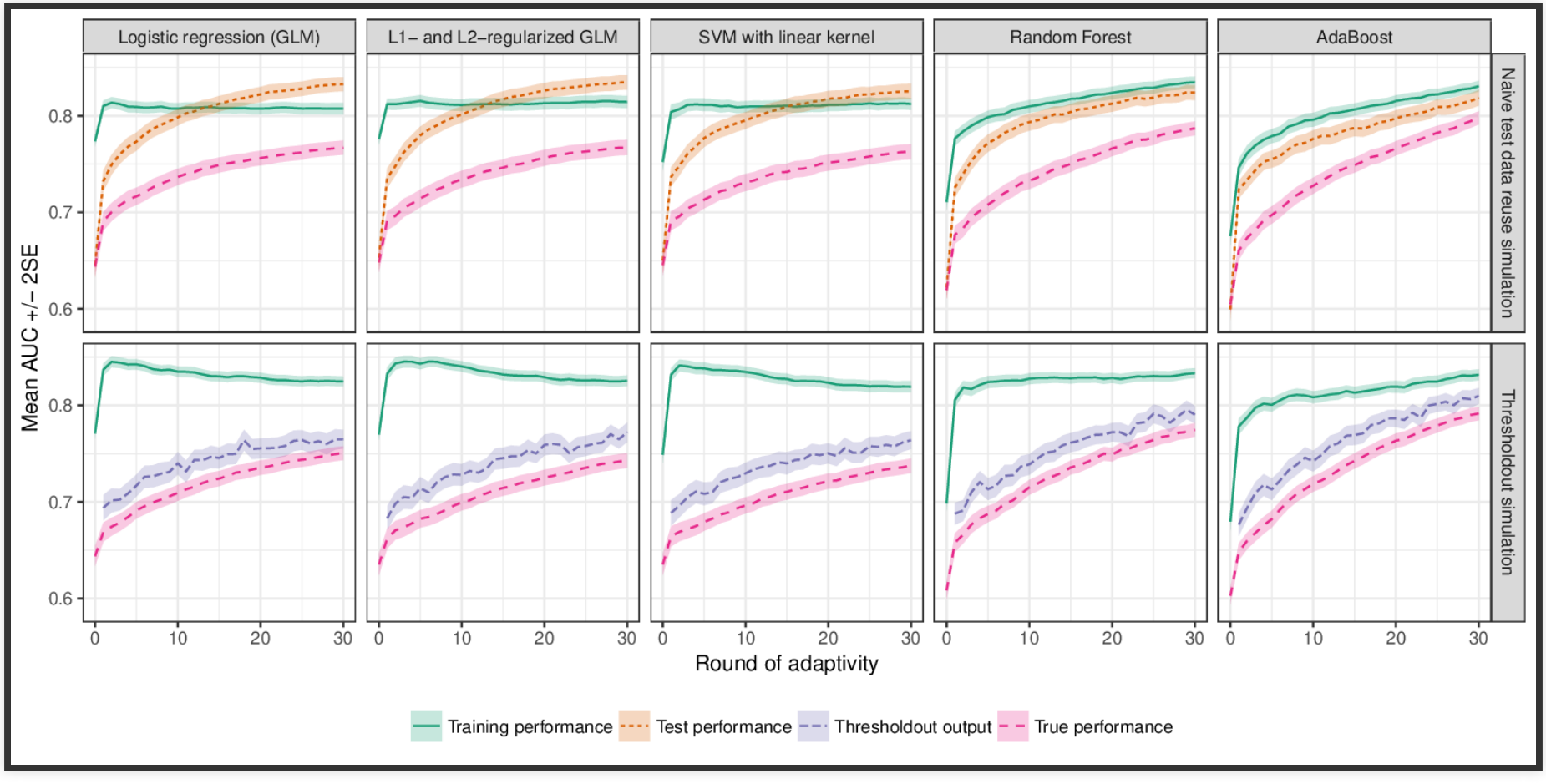
BINARY CLASSIFICATION PROBLEM

- Simulated data:
  - Small training and test sets, $n_{\text{test}} = n_{\text{train}} = 100$.
  - High-dimensional feature space, $p = 300$.
  - Non-linear outcome variable.
  - Only $s = 10$ variables are predictive of the outcome.
- Classification algorithms: logistic regression (GLM), regularized GLM (elastic net), linear SVM, random forest, and AdaBoost.

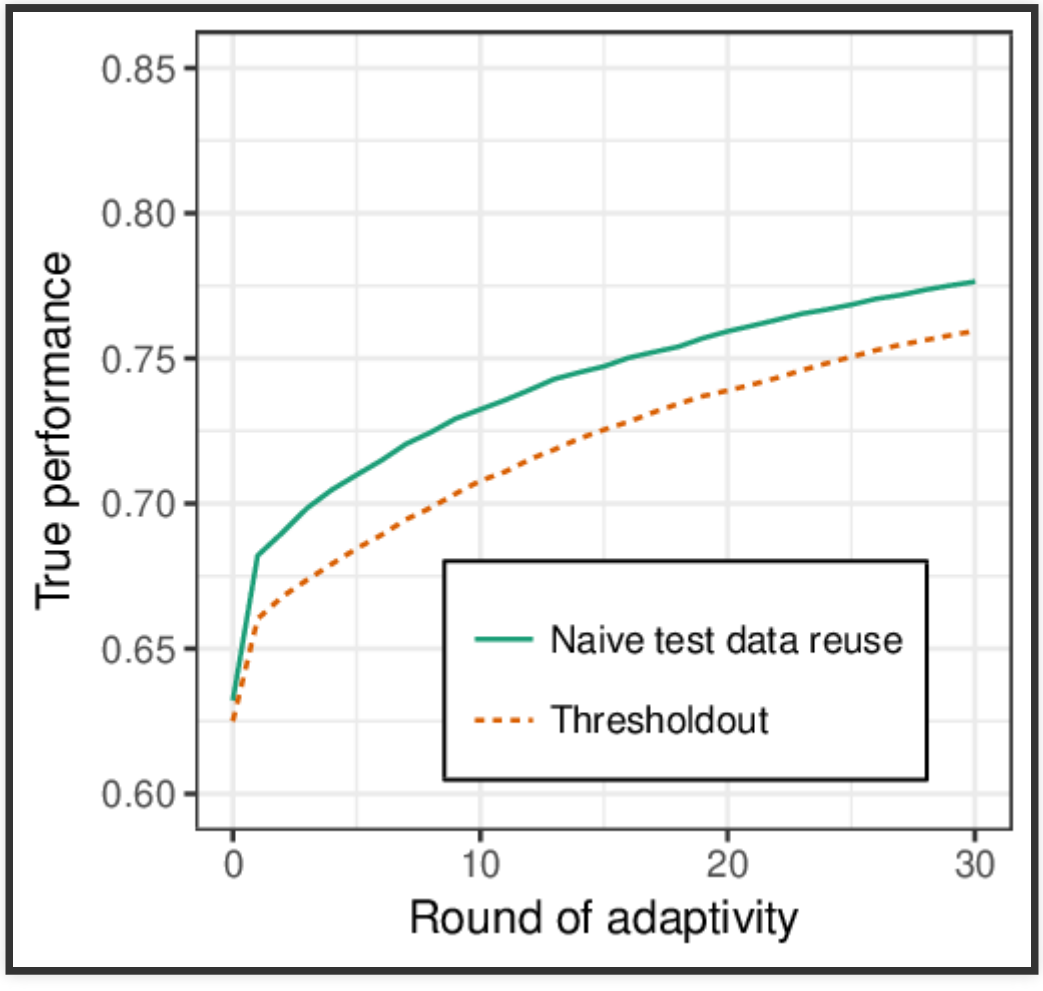
- **30 rounds of adaptive learning.**
- *Only AUC estimates can be reported from test data.*
- **For** round $r = 1, 2, \ldots, 30$ **do**:
  - $n_{\text{train}} \leftarrow n_{\text{train}} + 10$.
  - Candidate variables for addition to the trained classifier from round $(r-1)$ determined by t-tests at significance level $\alpha = 0.01$.
  - A new classifier trained using each subset of candidate variables, with cross-validation on the training data used for parameter tuning.
  - The best among the classifers from previous step is chosen based on test data AUC estimates.



Average number of Thresholdout$_{\mathrm{AUC}}$ queries by round.



Accuracy of reported AUC values is improved, at the cost of slightly higher uncertainty in the reported AUC, and slightly worse predictive performance.



Average true performance of the trained classifier by round with either test data reuse approach.

CONCLUSION

- Machine learning algorithms may continue to evolve after deployment as new data becomes available for training but not for testing. $\rightsquigarrow$ Test data reuse.
- **Theory & simulation**: Thresholdout and similar procedures reduce...
  - ...the upward bias in the reported performance measures.
  - ...overfitting to the test data.
- **Simulation studies**: promising results even on small samples.

REMAINING ISSUES

- Unclear how to choose parameters in practice, outside of the worst case, when the analyst is not actively trying to overfit to the test data.
- Average behavior vs. behavior in a specific execution (deviation from the average).
- Many possible improvements to the simulation protocol to achieve more realistic conditions, and greater range of conditions.