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

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

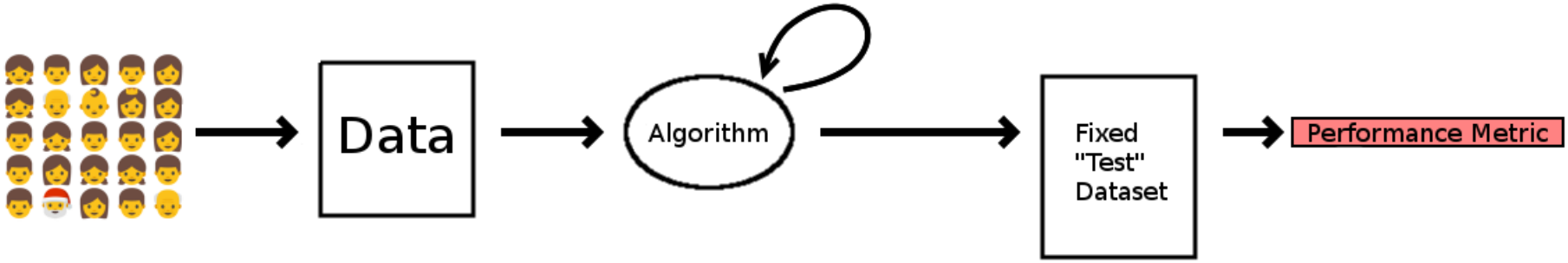
- **Training data**: exploratory analysis, model fitting, parameter tuning, comparison of different machine learning algorithms, feature selection, etc.

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

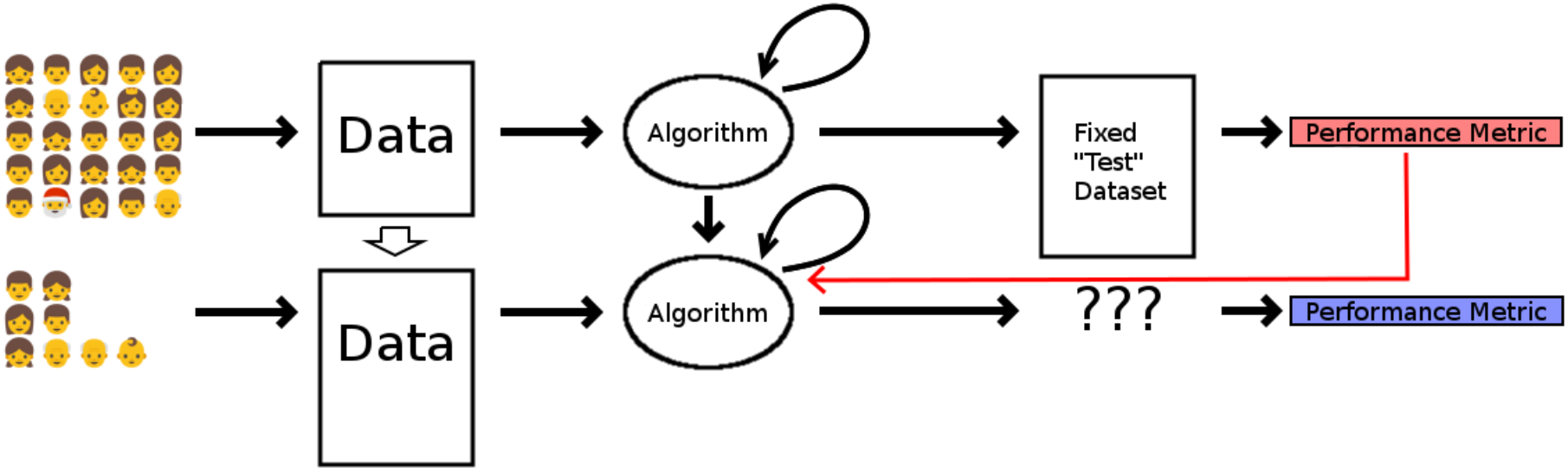
- **Test data**: Performance evaluation *after the trained machine learning algorithm has been "frozen"*.

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

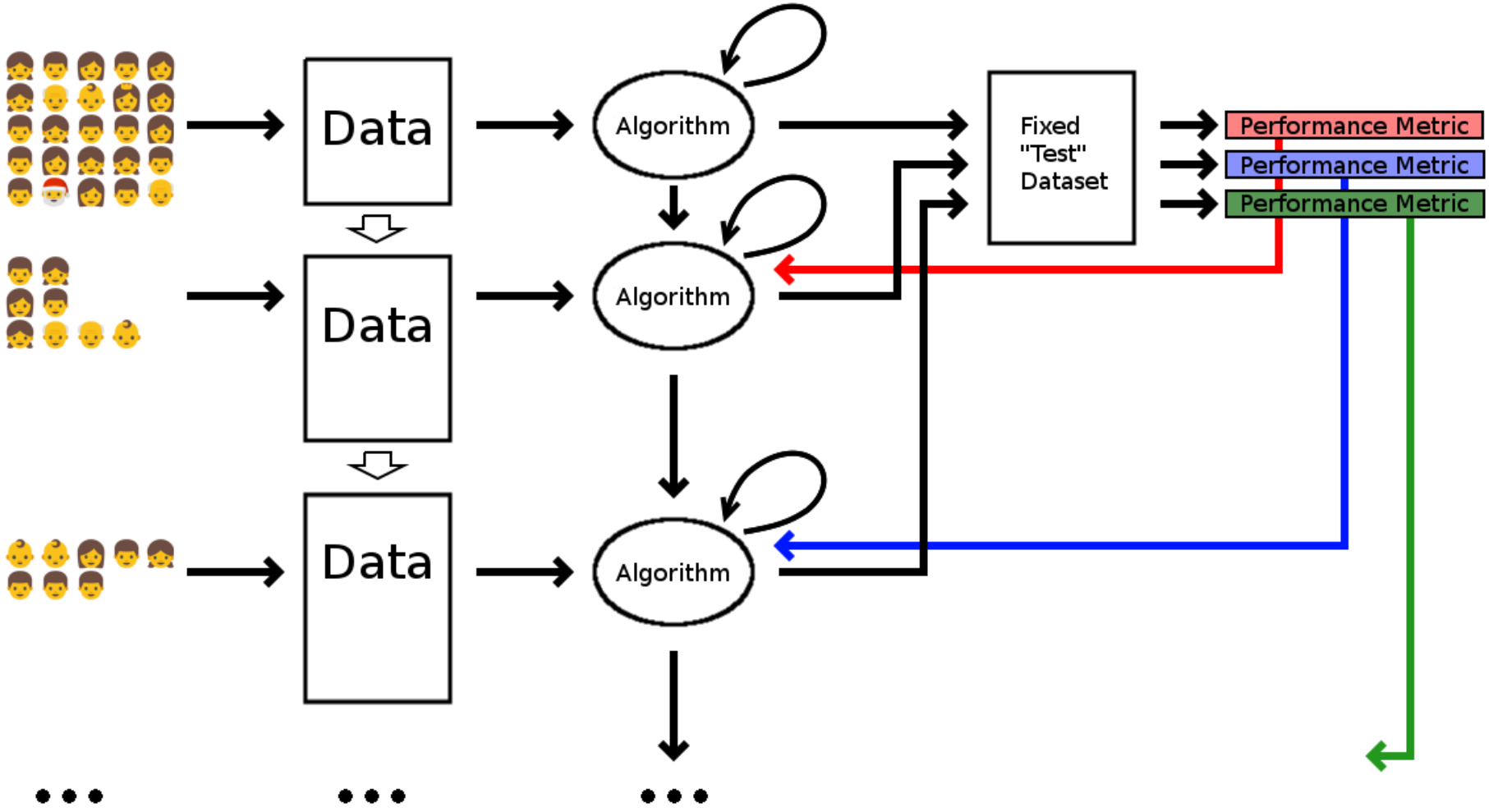
General machine learning process



"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?

- **Idea:** Access test data only through a DP mechanism

  $\Longrightarrow$ ML alg. is prevented from extracting information about individual test data records

  $\Longrightarrow$ only characteristics of the underlying distribution are learned

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.

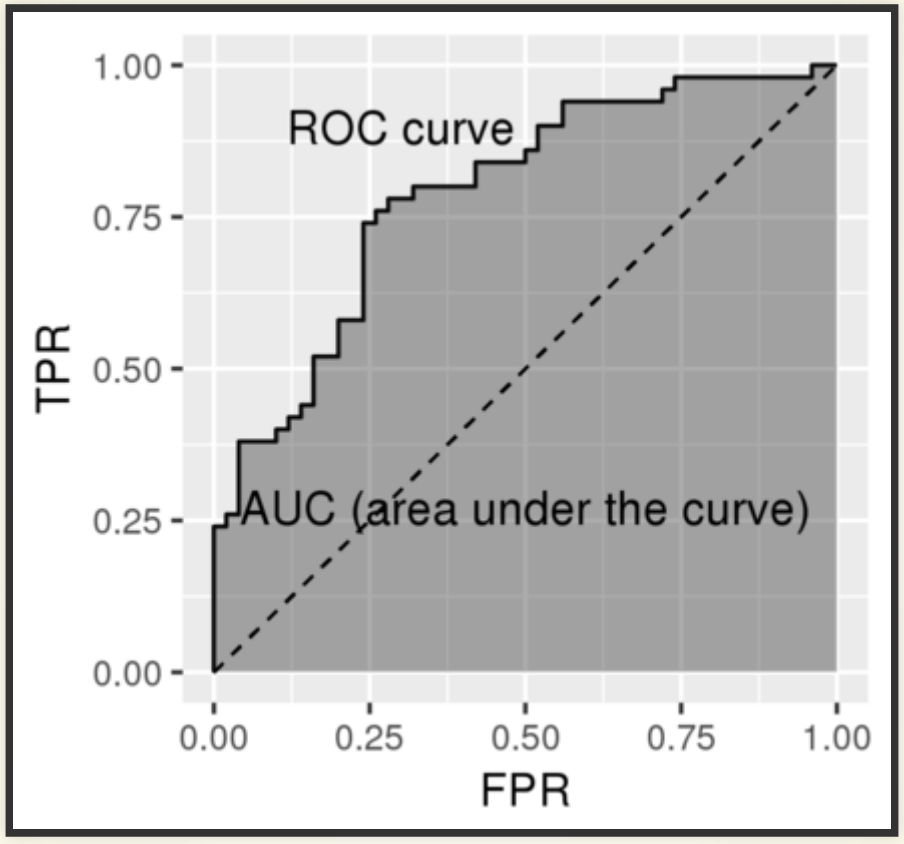
IN THIS WORK:

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?



Example of an (empirical) ROC curve.

WHY AUC?

- Invariance to prevalence.
- Independence of the decision threshold (can be chosen later).
- Probabilistic meaning.
- 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.

THRESHOLDOUT$_{\mathrm{AUC}}$ – ROUGH SUMMARY

Trained classifier $\phi(x) \in [0, 1]$

If $|\mathrm{AUC}_{\mathrm{training}}(\phi) - \mathrm{AUC}_{\mathrm{test}}(\phi)| > \tilde{T}$:
output $\mathrm{AUC}_{\mathrm{test}}(\phi)+$ "a little noise"
Else:
output $\mathrm{AUC}_{\mathrm{training}}(\phi)$

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 – 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.

(Proof similar to DFHPRR '15 arXiv:1506.02629)

THEOREM – DRAWBACKS

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

THEOREM – DRAWBACKS

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

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?

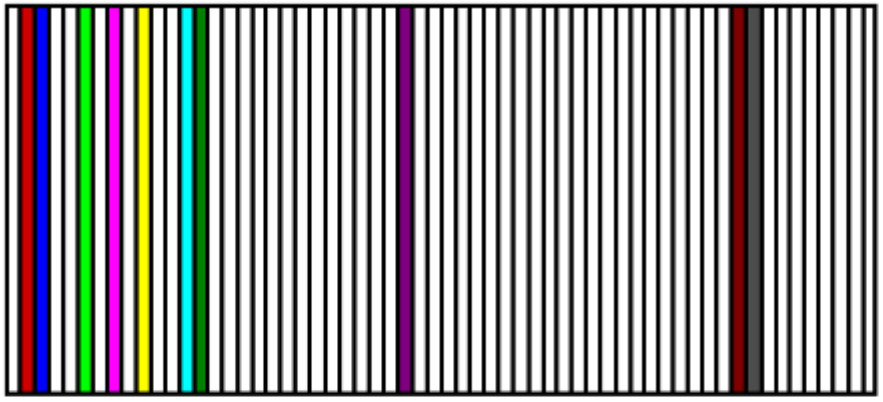
SIMULATION STUDIES ON SMALL SAMPLES

SIMULATION STUDIES ON SMALL SAMPLES

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

("naive" means: use exact test AUC values)

SIMULATED DATA



i.i.d. Gaussian

p = 300 features
s = 10 significant

$$\begin{bmatrix} p1 \\ p2 \\ p3 \\ \dots \end{bmatrix} = f(\text{  })$$

with linear, quadratic, exponential, and interaction effects

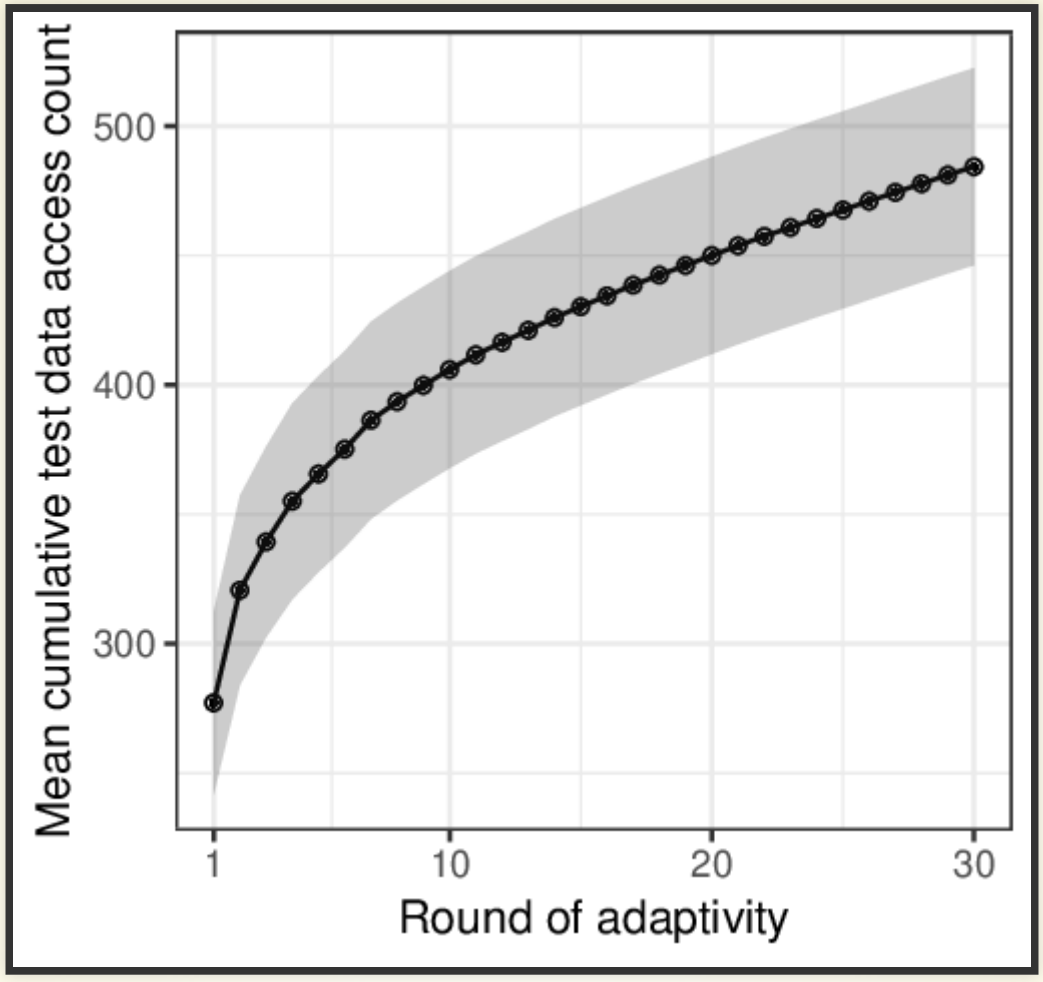
$$\begin{bmatrix} y1 \\ y2 \\ y3 \\ \dots \end{bmatrix} \sim \text{Ber}\left(\begin{bmatrix} p1 \\ p2 \\ p3 \\ \dots \end{bmatrix}\right)$$

binary outcome variable

BINARY CLASSIFICATION PROBLEM

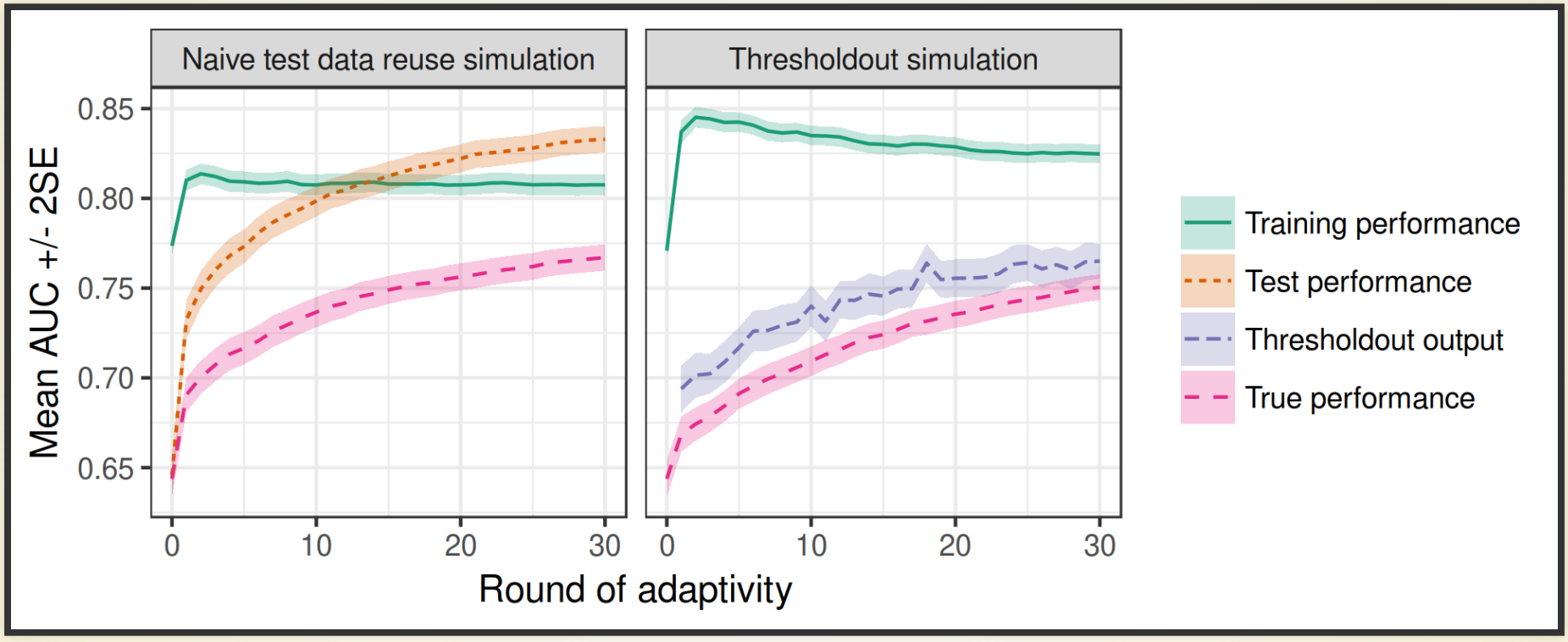
- Initially $n_{\text{test}} = n_{\text{train}} = 100$.
- In each round of adaptive learning:
  $n_{\text{train}} \leftarrow n_{\text{train}} + 10$.
- Classification algorithms: logistic regression (GLM), regularized GLM (elastic net), linear SVM, random forest, and AdaBoost.

- **30 rounds of adaptive learning.**
- *Only AUC estimates can be reported from test data.*
- **For** round $r = 1, 2, \ldots, 30$ **do**:
  1. $n_{\text{train}} \leftarrow n_{\text{train}} + 10$.
  2. Identify new candidate variables on training data.
  3. New classifiers trained with each subset of candidate variables added to the classifier from round $(r-1)$; 5-fold CV for parameter tuning.
  4. Estimate AUC on test data for each classifier from step 3.
  5. Pick the best classifier based on step 4.



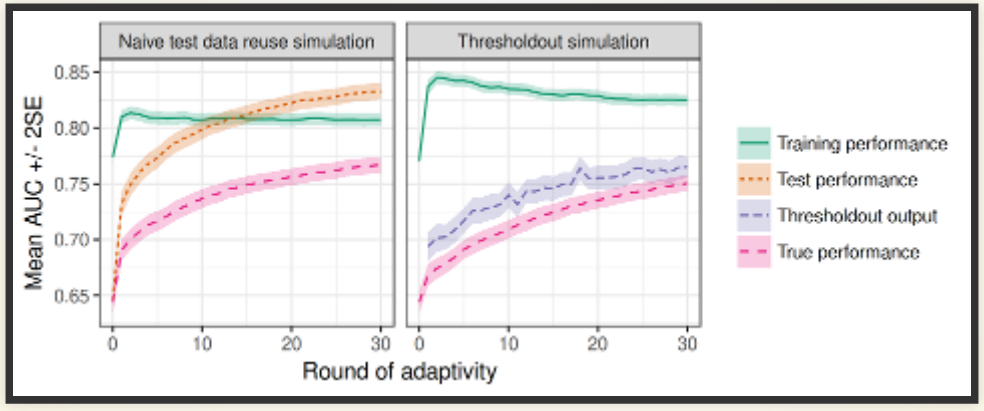
Average number of Thresholdout$_{\mathrm{AUC}}$ queries by round.

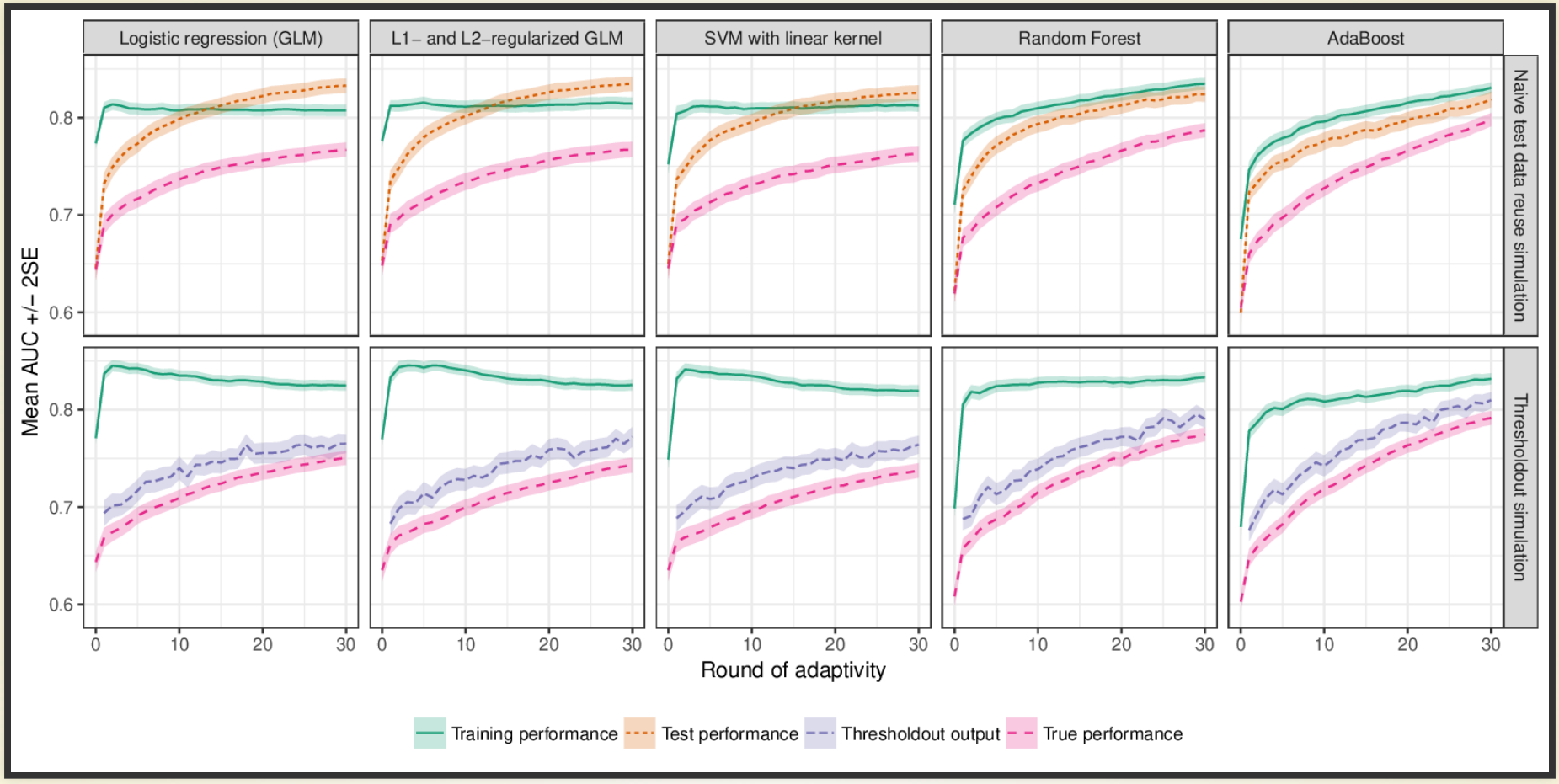
LOGISTIC REGRESSION



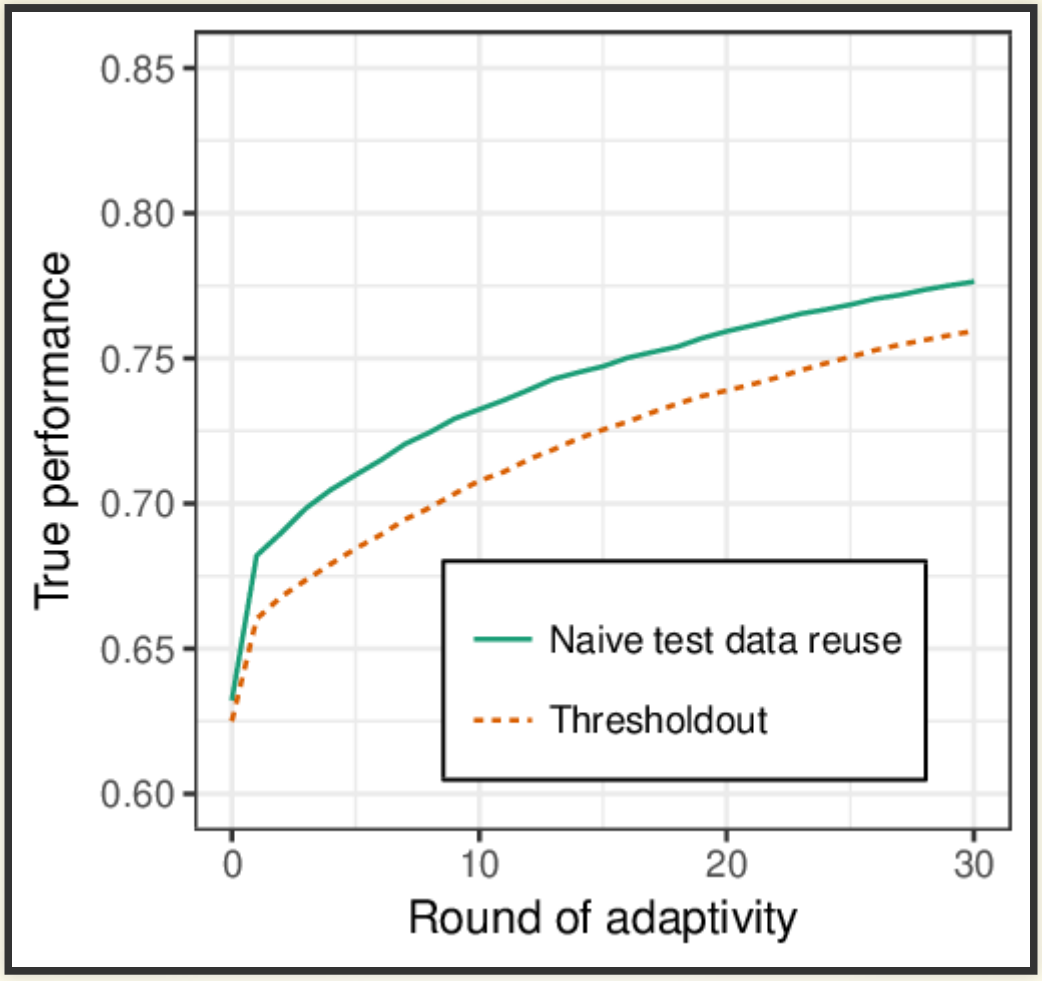
LOGISTIC REGRESSION

- **Naive approach**: Classifier learns the effect of local noise in the test data (overfitting).
- **Thresholdout approach**: The gap between the reported and the true AUC is much narrower!





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.