# Defining the AIM: An Abstraction for Improving Machine Learning Prediction

Victoria Stodden
University of Illinois at Urbana-Champaign

Symposium on Data Science and Statistics
Reston, VA
May 18 2018

# Imagine: Querying the Scholarly Record

1. Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;

2. Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;

3. List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;

4. Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;

5. Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the 2003 and list the trial name and histogram side by side.

Donoho & Gavish, “Three Dream Applications of Verifiable Computational Results,” CiSE, 2012

# The Acute Lymphoblastic Leukemia Dataset

Introduced in Golub et al. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring” (1999): “cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias [to] discover the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)"

In joint work with Xiaomian Wu and April Tang, we tried query 3.

# The ALL Dataset Query

We wanted:

- A list of all classifiers applied to the Golub dataset (with citations);
- A comparison of their misclassification rates.

A literature search produced 30 articles, but they did not give comparable misclassification rates.

Our next step was to create the table of comparable misclassification rates. We identified 5 articles for which this seemed possible.

# Our (Naive) Expectation

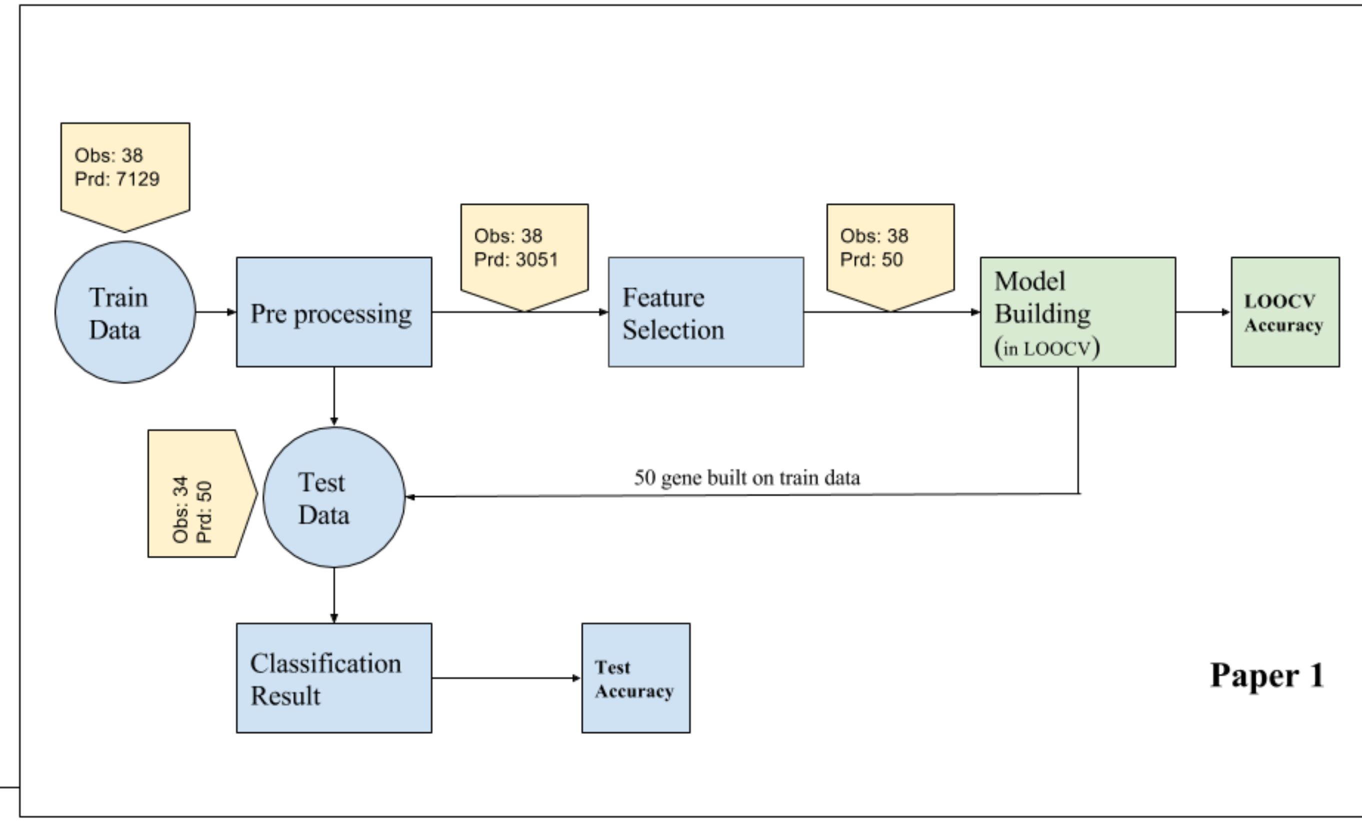
We found that the articles implemented (at least) three steps, each varying from one article to the next:

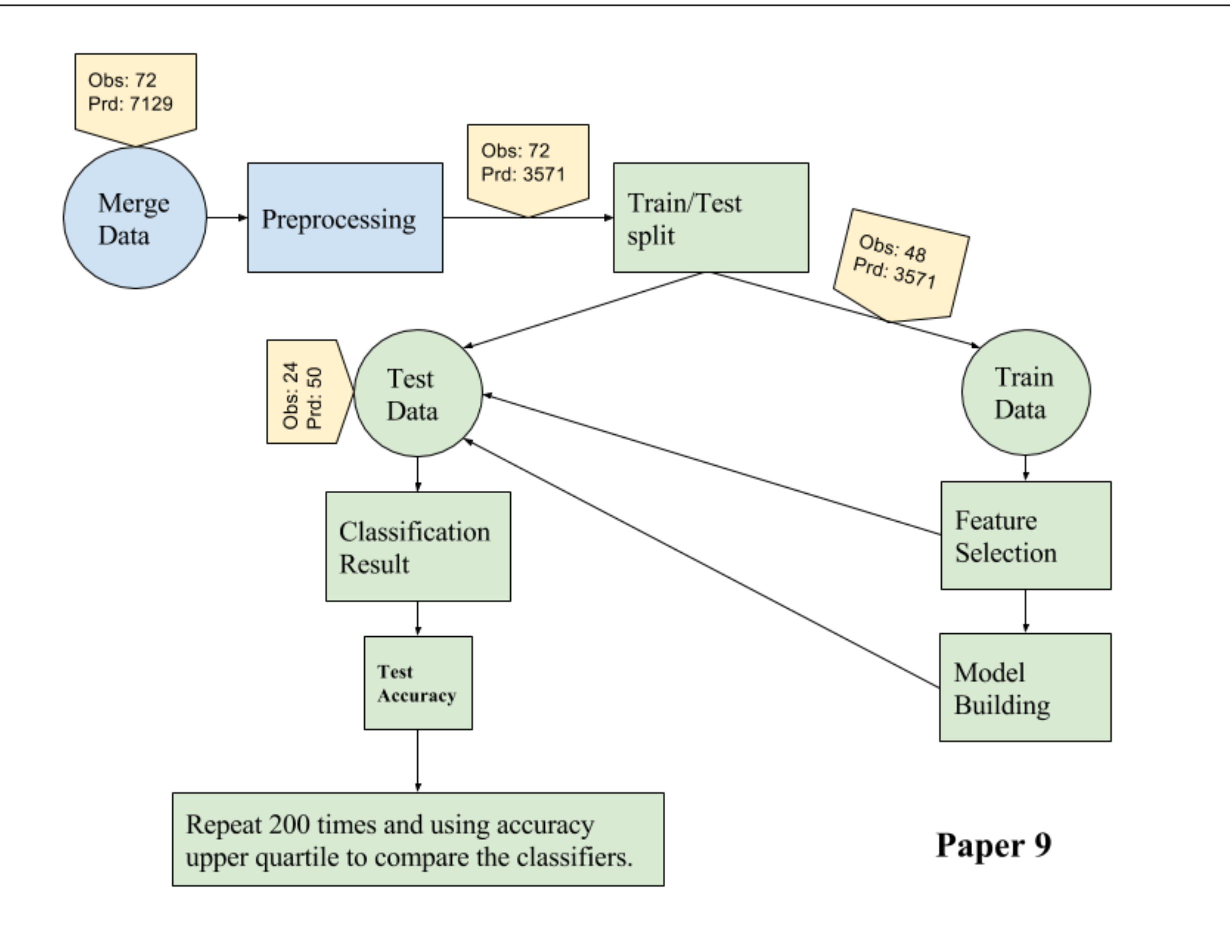
1. data preprocessing,

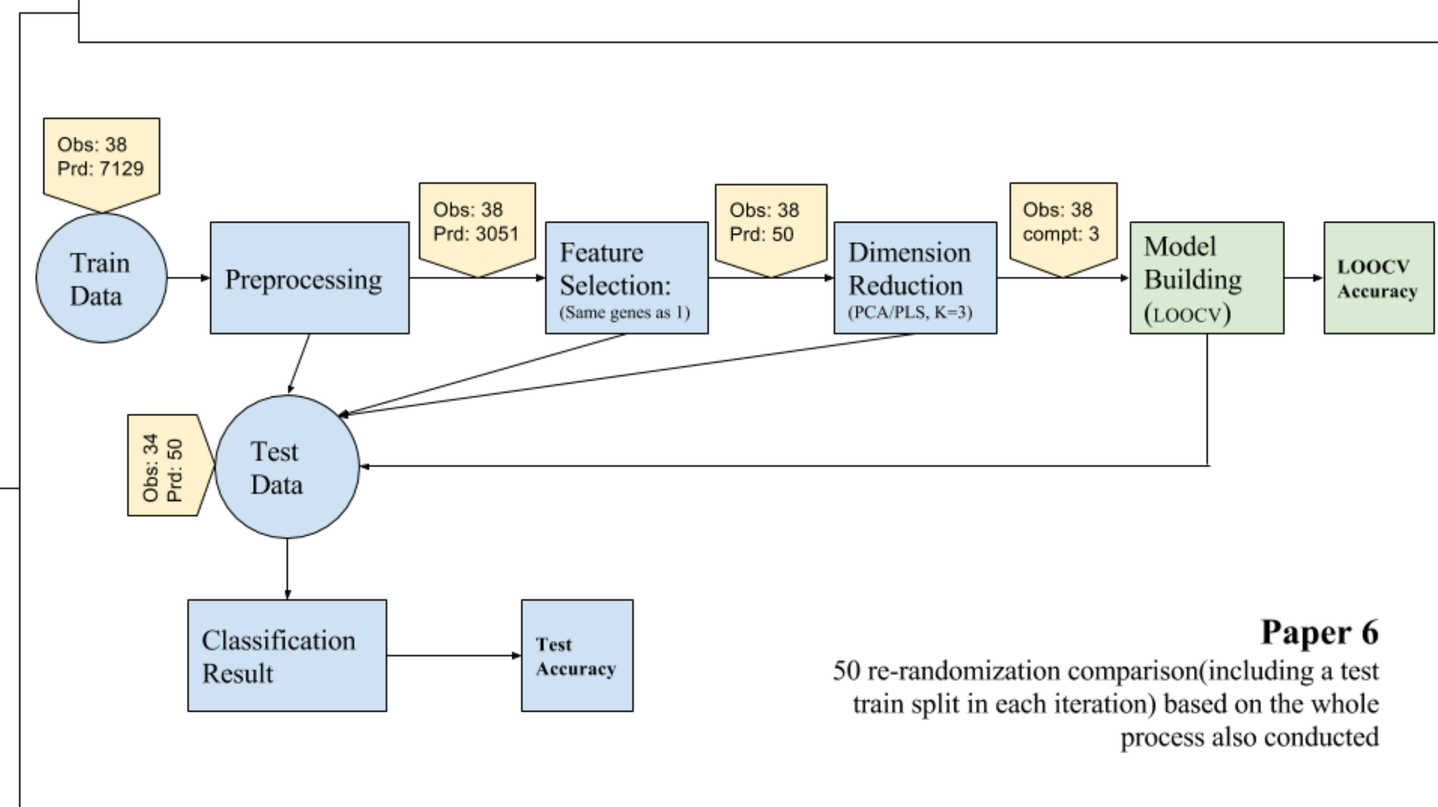
2. feature selection,

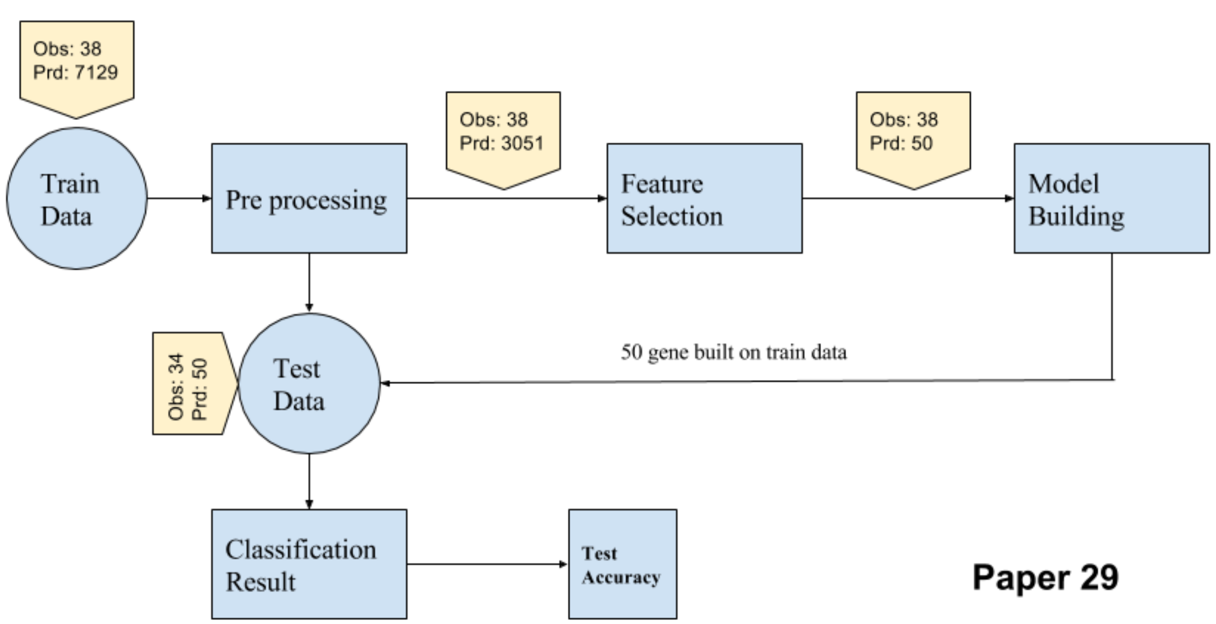
3. application of machine learning algorithm.

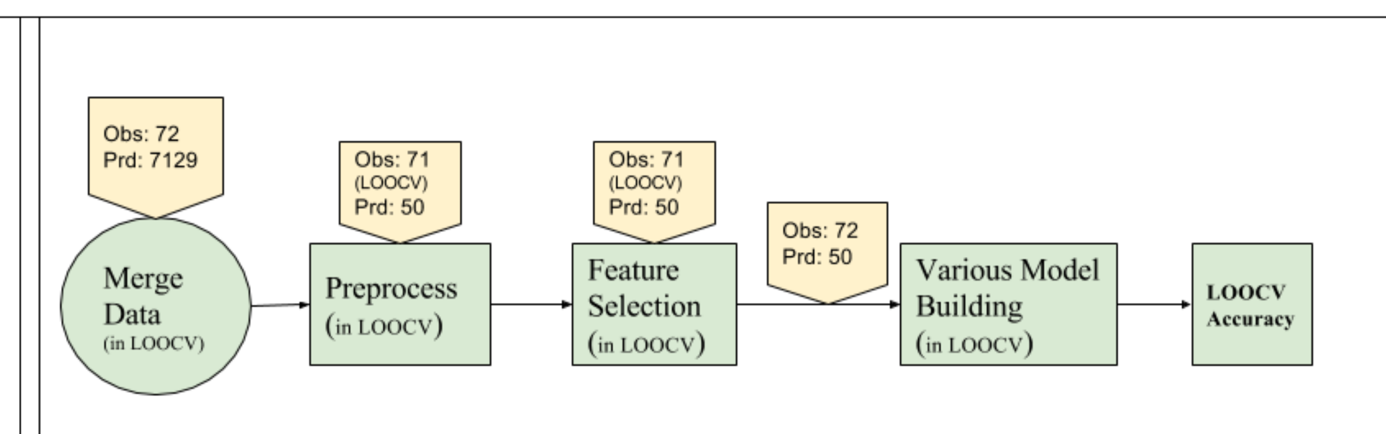
# Analysis Steps



# Learning Algorithms Applied (typically 47ALL; 25AML)

| Paper | Data Size | Algorithm(s) Applied |
|---|---|---|
| 1 | 72 x 6817 | Golub Classifier: informative genes+weighted vote |
| 2 | 72 x 6817 | Golub Classifier: informative genes+weighted vote |
| 3 | 72 x 7129 | Nearest Neighbor; SVM(linear kernel, quadratic kernel); |
| 4 | 72 x 7129 | SVM(top 25, 250, 500, 1000 features) |
| 5 | 72 x 7070 | MVR(median vote relevance); NBGR(naive bayes global |
| 6 | 72 x 6817 | Logistic and Quadratic discriminant analysis |
| 7 | 72 x 7129 | SVM |
| 9 | 72 x 6817 | Linear and Quadratic discriminant analysis; Classification trees; |
| 10 | 72 x 7129 | Decision Trees; AdaBoost |
| 11 | 72 x 7129 | MAVE-LD, DLDA, DQDA, MAVE-NPLD |
| 12 | 72 x 7129 | SIMCA classification |
| … | … | … |

# Comparable Classification Rates

| | **Feature Selection Method** | | | | | |
|---|---|---|---|---|---|---|
| **Classifier** | **Paper1** | **Paper3** | **Paper6a** | **Paper6b** | **Paper9** | **Paper29** |
| **Paper1 classifier** | **0.912** | 0.941 | 0.971 | 0.971 | 0.882 | 0.735 |
| **Paper 3 Adaboost** | 0.912 | **0.912** | 0.971 | 0.971 | 0.912 | 0.912 |
| **Paper 3 NN** | 0.971 | **0.941** | 0.912 | 0.941 | 0.971 | 0.971 |
| **Paper 3 SVM Linear** | 0.971 | **0.971** | 0.941 | 0.971 | 0.971 | 0.765 |
| **Paper 3 SVM Quadratic** | 0.971 | **0.882** | 0.971 | 0.971 | 0.971 | 0.912 |
| **Paper 6 logit** | 0.971 | 0.971 | **0.971** | **0.971** | 0.971 | 0.882 |
| **Paper 6 qda** | 0.941 | 0.912 | **0.941** | **0.971** | 0.971 | 0.853 |
| **Paper 9 bagging** | 0.941 | 0.912 | 0.971 | 0.971 | **0.912** | 0.765 |
| **Paper 9 bagging with CPD** | 0.735 | 0.853 | 0.8235 | 0.912 | **0.765** | 0.677 |
| **Paper 9 decision tree** | 0.912 | 0.912 | 0.971 | 0.971 | **0.912** | 0.735 |
| **Paper 9 DLDA** | 0.971 | 0.941 | 0.971 | 0.971 | **0.971** | 0.882 |
| **Paper 9 DQDA** | 0.971 | 0.941 | 0.971 | 0.971 | **0.971** | 0.882 |
| **Paper 9 FLDA** | 0.882 | 0.882 | 0.971 | 0.971 | **0.882** | 0.882 |
| **Paper 9 nn** | 0.971 | 0.912 | 0.853 | 0.971 | **0.941** | 0.941 |
| **Paper 29 Bayesian Network** | 0.735 | 0.882 | 0.971 | 0.971 | 0.824 | **0.618** |

# Learnings..

- Classification rates are hard to synthesize (200+ student hours)
- Many points of variability: starting dataset; preprocessing steps; feature selection methods; algorithm choice; tuning of algorithm and parameters...
- Details not well-captured in the traditional article, making comparisons difficult or impossible.
- Would be easier if:
    - ➡ there was prior agreement on the dataset,
    - ➡ prior agreement on hold-out data for testing,
    - ➡ full disclosure of feature selection steps,
    - ➡ full disclosure of algorithm application and parameter tuning.

# Conclusions

- *Serious* problems in reporting standards
- Framework for Analysis:
    - ➡ Agreement on datasets prior to analysis, conferences around those datasets,
    - ➡ Hold-out data held by a neutral third party (e.g. NIST), not seen by researchers,
    - ➡ Researchers distinguish and specify feature selection and preprocessing vs learning algorithm application,
    - ➡ Send code to the third party who returns your misclassification rate on the test data.

  Side effect: training data and code/algorithm shared.

- Results and analysis at https://github.com/AIM-Project/AIM-Manuscript/