DRAPER

Do results generalize in clinical practice?

Presentation to the Imagine 2017 Workshop

September 6, 2017

John M. Irvine, Ph.D.
Chief Scientist for Data Analytics
jirvine@draper.com
Phone: + 1-617-258-4957

Charles Stark Draper Laboratory, Inc
555 Technology Square, Cambridge, MA 02139

Statistical Methods

From Wikipedia:

- "**Statistics** is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data."
- "… inferential statistics… draw conclusions from data that are **subject to random variation** (e.g., observational errors, sampling variation)

General approach:

- Probabilistic model is the process that generates the observed data
- Seek to make inferences about the underlying model:
  - *Estimate parameters*
  - *Test hypotheses*
  - *Predict future observation*

Make decisions in the face of uncertainty

Machine Learning

From *SAS:*

- **"Machine learning** is a method of data analysis that **automates analytical model building.**
- Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights **without being explicitly programmed where to look."**

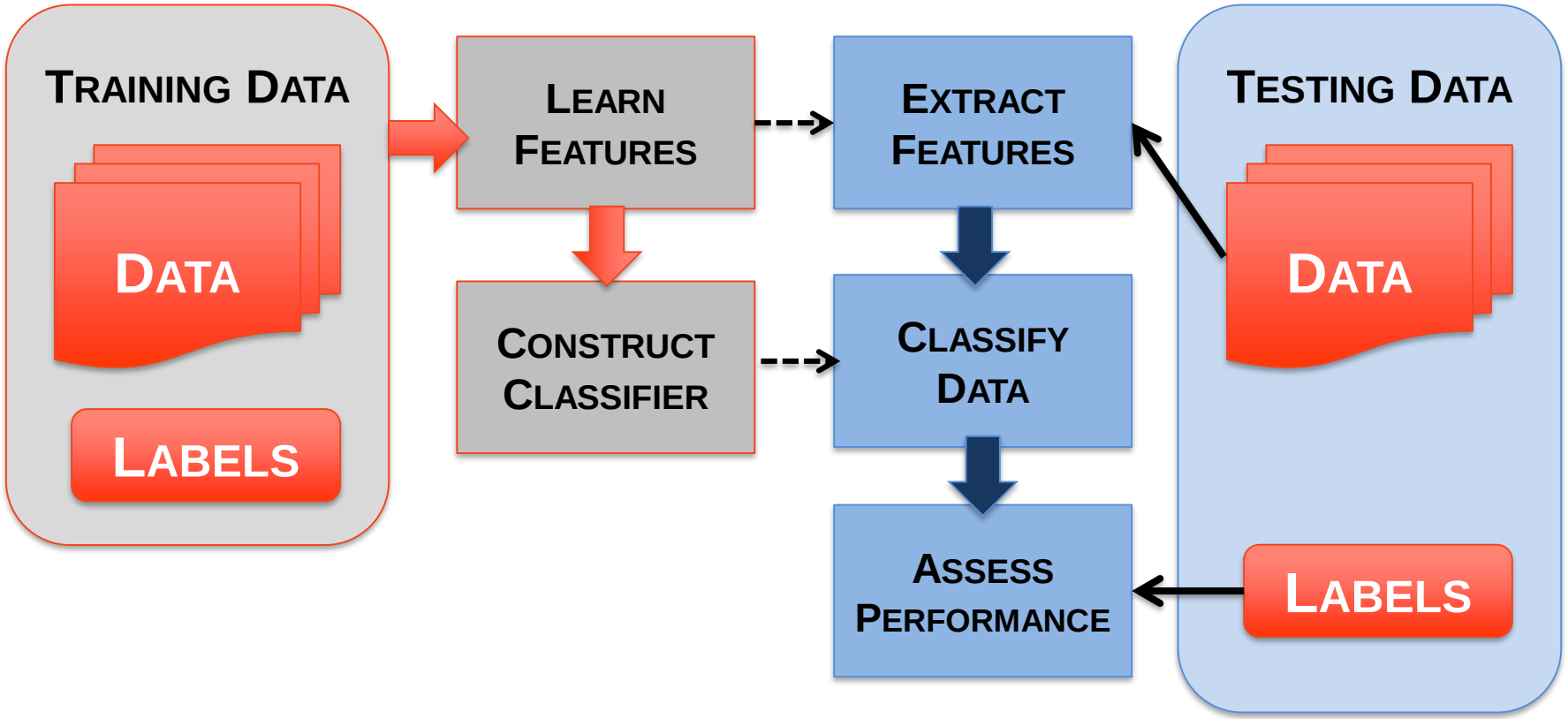
General Approach:

- Apply machine learning algorithms to a set of "labeled" training data to discover relationships:
  - *Discover features*
  - *Develop classification rules*
  - *Validation – with new data, if possible*

Discover predictive relationships in the data

Generic Machine Learning Overview

- Training is the **discovery** part of the process
- Testing is the **validation** of the method



Measures of Performance

Statistical Hypothesis Testing

$$X_i \sim g(X, \theta)$$

$$H_0: \theta \in A \qquad H_1: \theta \in B$$

	Truth	
Decision	**H_0 true**	**H_1 true**
Fail to reject H_0	**Correct**	**Type 2 error**
Reject H_0	**Type 1 error**	**Correct**

Type 1 and Type 2 error quantify expected performance

Classifier Performance

Decision: Classifier assigns individual observations to class (A or B)

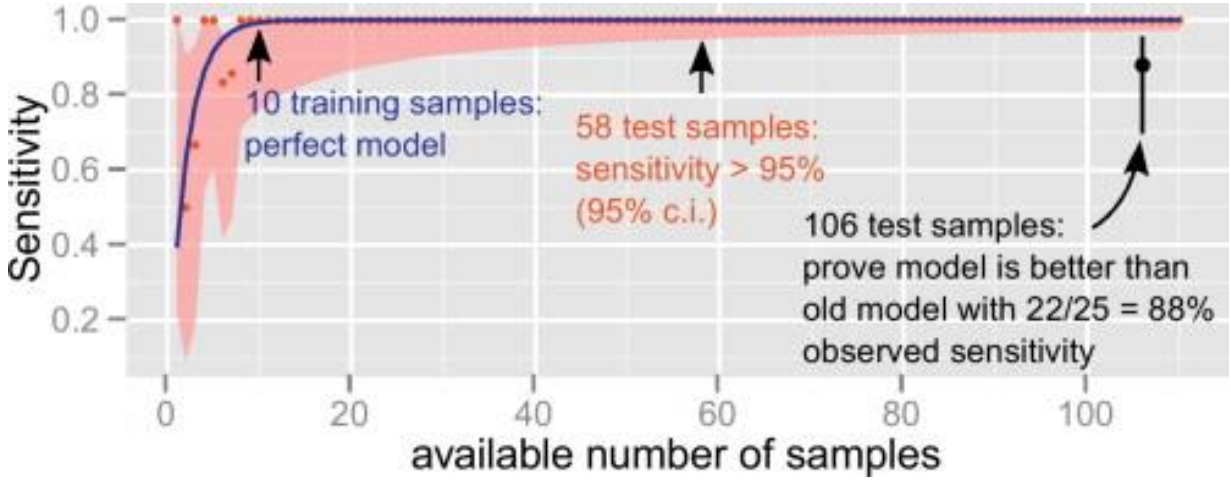
	Truth	
Decision	**Class A**	**Class B**
Class A	**Correct**	**error**
Class B	**error**	**Correct**

Performance measures:

- Probability of Correct Classification (PCC)
- ROC curve
- Confusion matrix (multi-class problems)

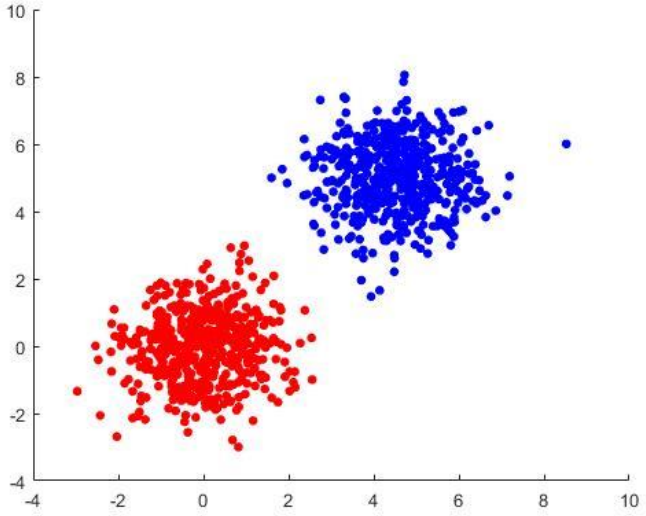
Training Sample Requirements

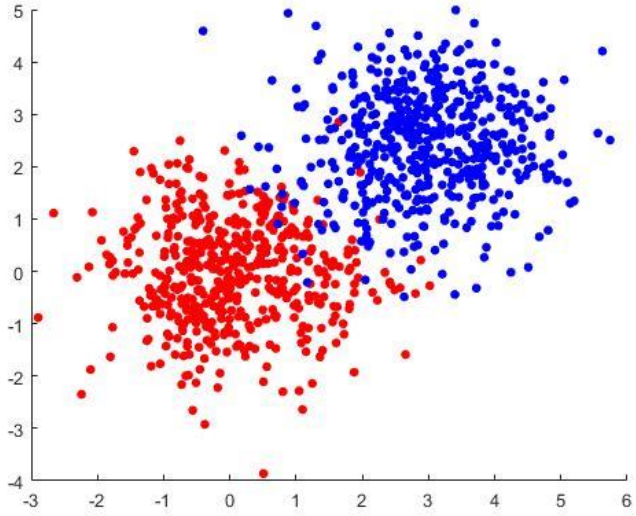
- Average model performance as function of the training sample size – called the ***learning curve*** – depends on:
  - *Classification method*
  - *Complexity of the classifier*
  - *Class separability*

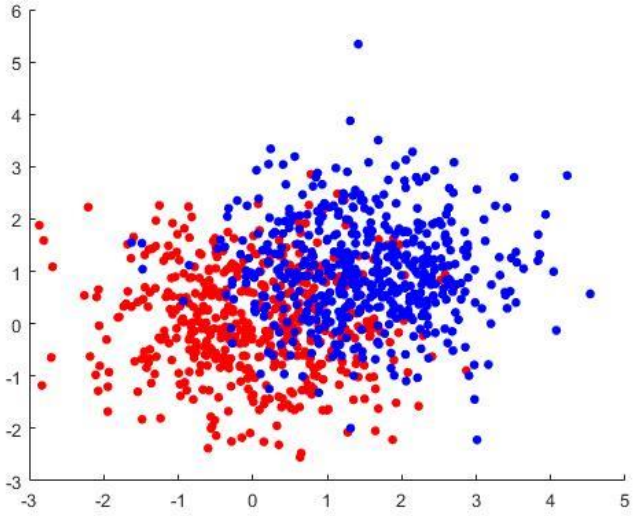


Beleites, C. and Neugebauer, U. and Bocklitz, T. and Krafft, C. and Popp, J.: Sample size planning for classification models. Anal Chim Acta, 2013, 760, 25-33. DOI: 10.1016/j.aca.2012.11.007

Separability

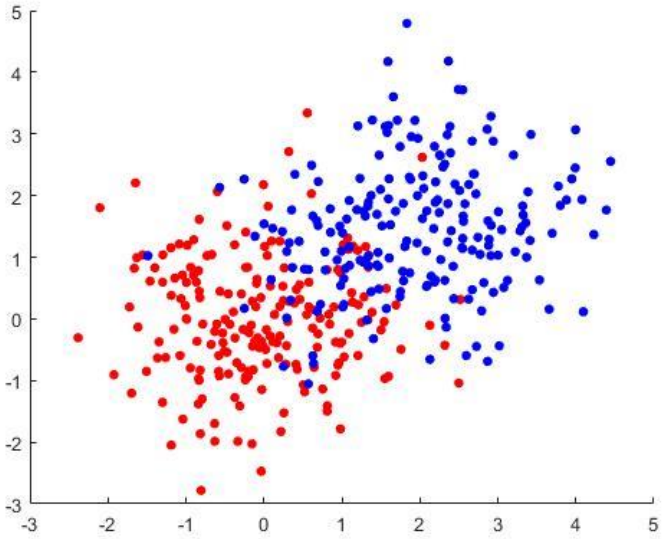


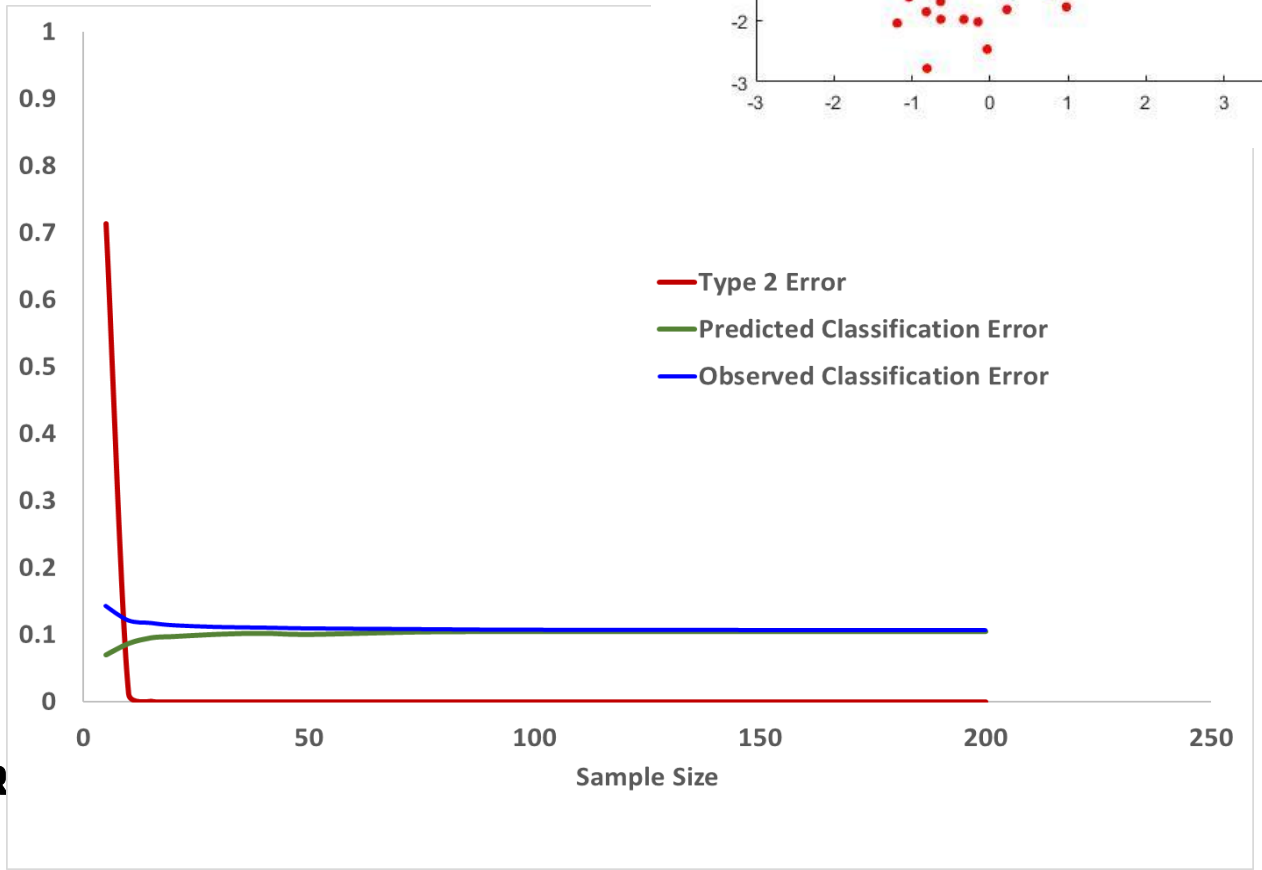




Sample Size

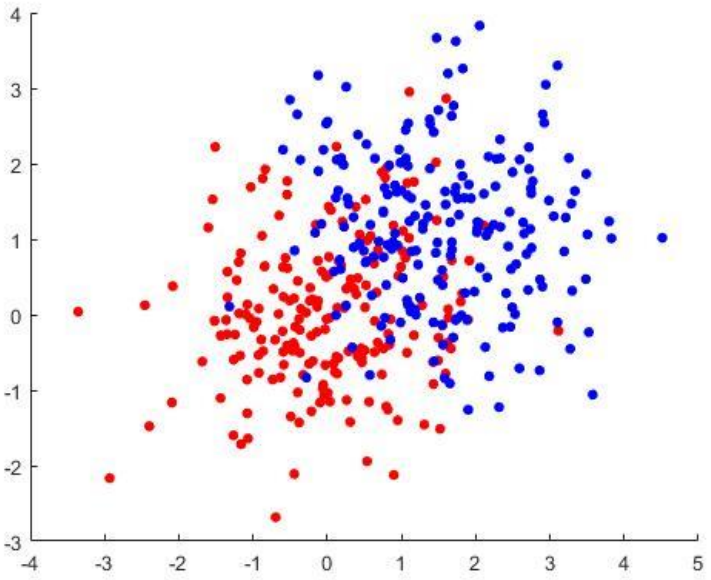
- Type 2 error is driven to zero very quickly (N=20)
- Classification error for new subjects persists, independent of N
  - *Depends on class separability*

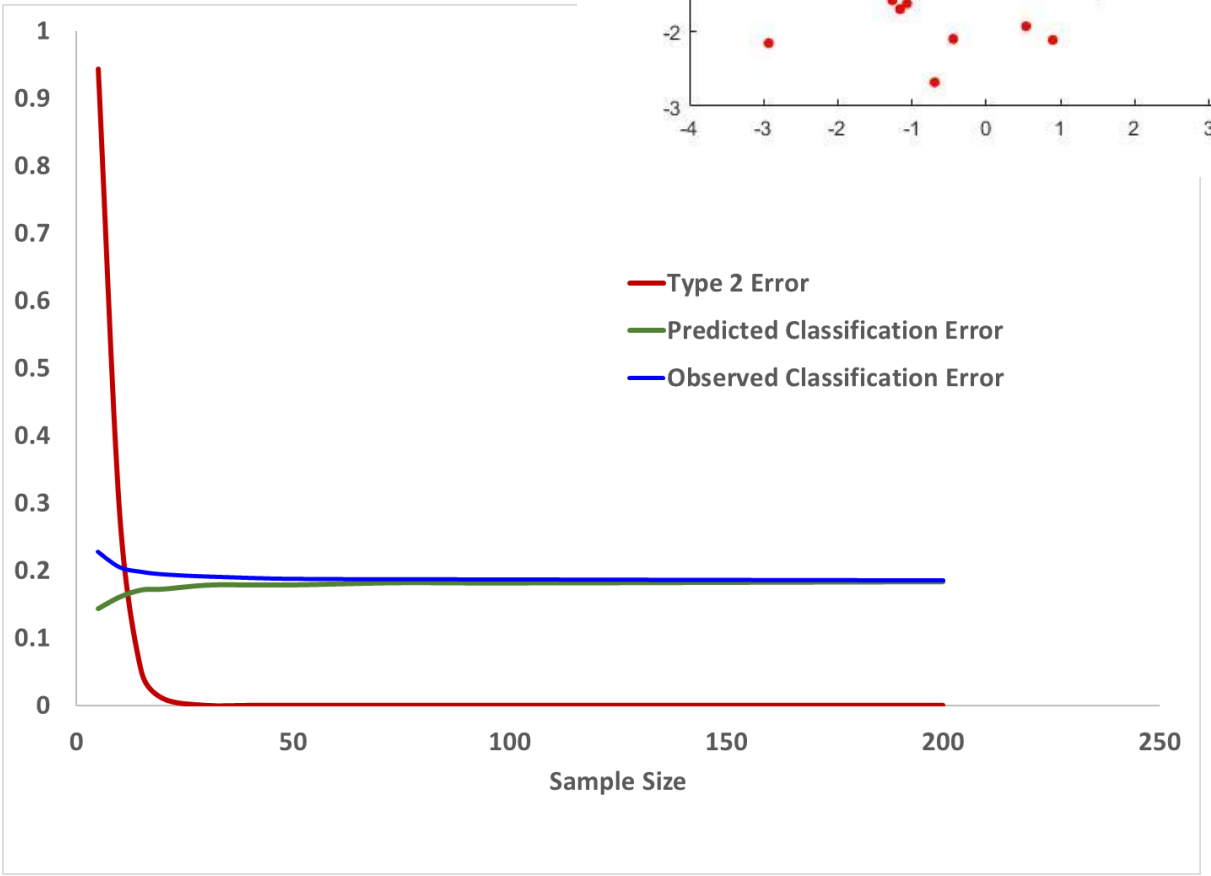




Sample Size

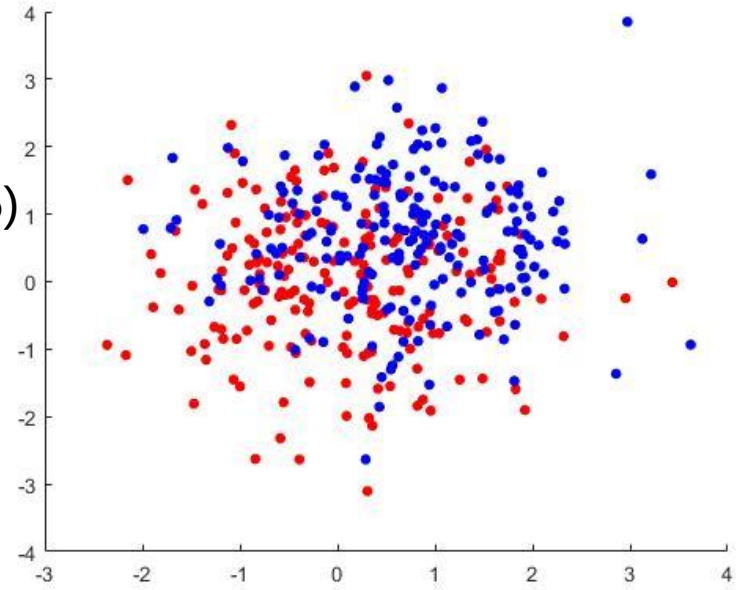
- Type 2 error is driven to zero quickly (N=30)
- Classification error for new subjects persists, independent of N
  - *Depends on class separability*

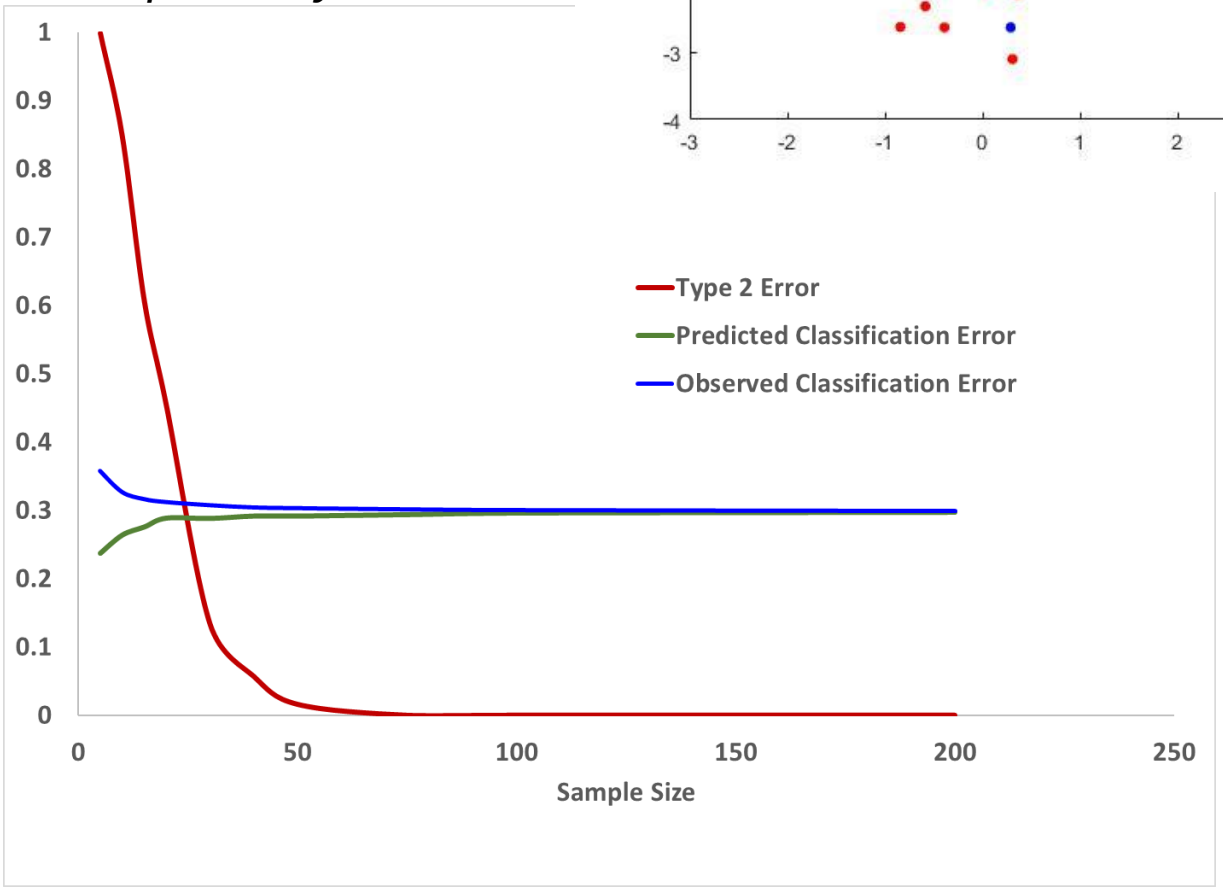




Sample Size

- Type 2 error is driven to zero less quickly (N=75)
- Classification error for new subjects persists, independent of N
  - *Depends on class separability*





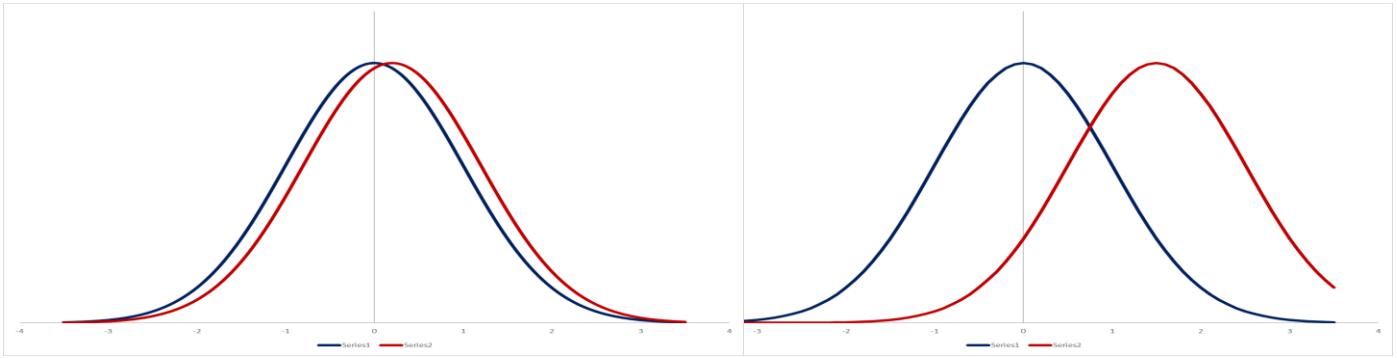
DRAPER

Statistical Power Analysis

- The **power** of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true

$$\text{Power} = \Pr\{ \text{reject } H_0 \mid H_1 \text{ true} \}$$

Statistical Consideration	Machine Learning
Sample Size	Size of the Training Set
Effect size	Separability



Statistical Modeling

"Essentially, all models are wrong, but some are useful"

George Box & Norman Draper

Empirical Model-Building and Response Surfaces

- Classical linear model:

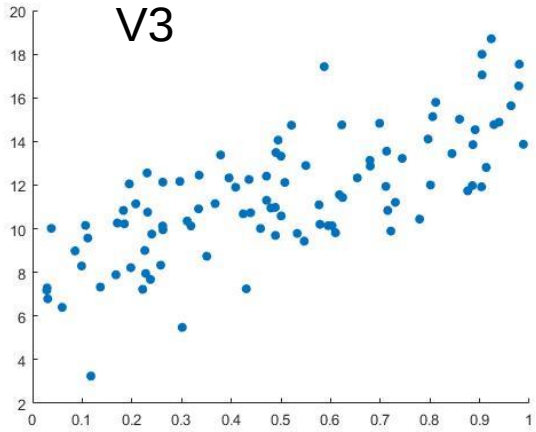
$$Y_i = \underline{X}_i \beta + \varepsilon_i$$

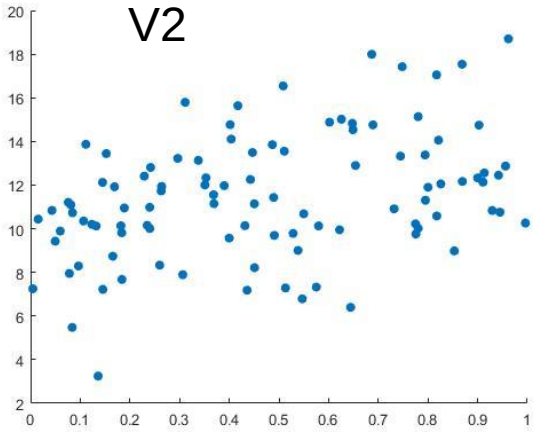
where ε_i are i.i.d. with mean 0 and variance σ^2

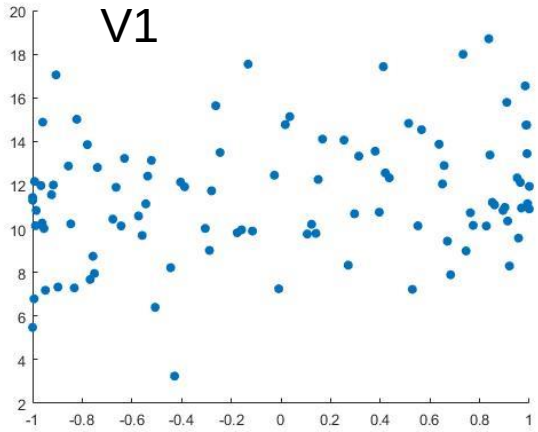
- Anything not modeled by $\underline{X}$ is in the error term

Unmodeled Terms

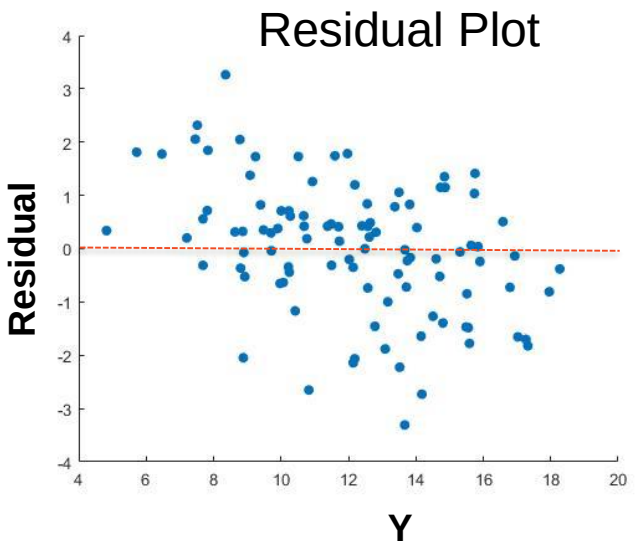
- True model: Y = 5 + v1 + 5*v2 + 8*v3







- Model using only v2 and v3
  $R^2 = 0.990$
- Model using all terms
  $R^2 = 0.994$



DRAPER

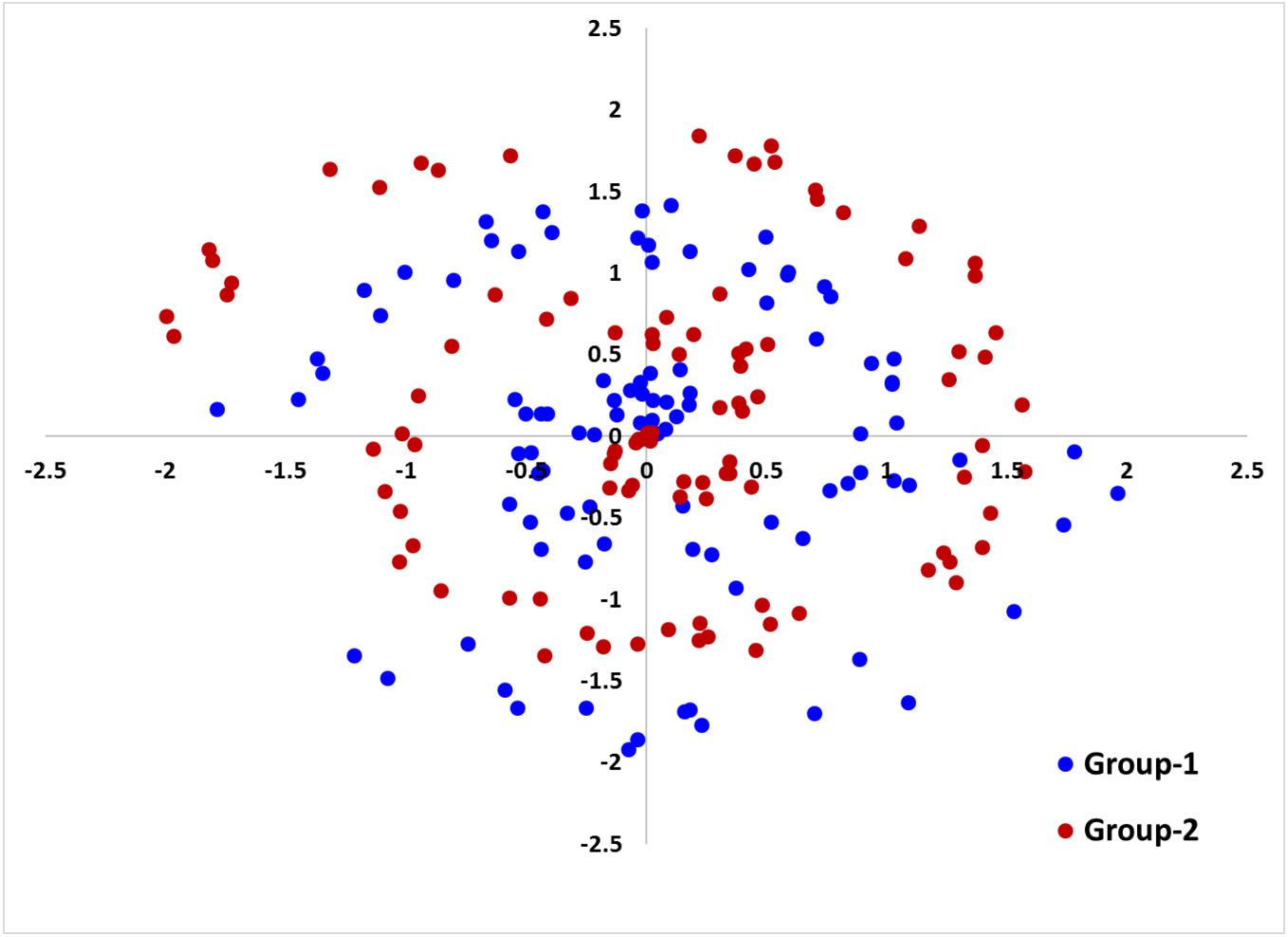
Feature Selection

- Statistical methods:
  - *Stepwise procedures*
  - *Tests based on significance (improvement in the model)*
  - *Candidate features are explicitly presented and evaluated*

- Machine learning:
  - *Large and growing literature on features selection*
  - *Popular methods based on mutual information (MI) maximize relevance and minimize redundancy:*
    - Relevance: High MI between candidate feature and target class
    - Redundancy: Low MI between candidate feature and other features in the model
  - *Stepwise selection process*
  - *Candidate features are explicitly presented and evaluated*

Two Spiral Problem

- Classes defined by two intertwined spirals

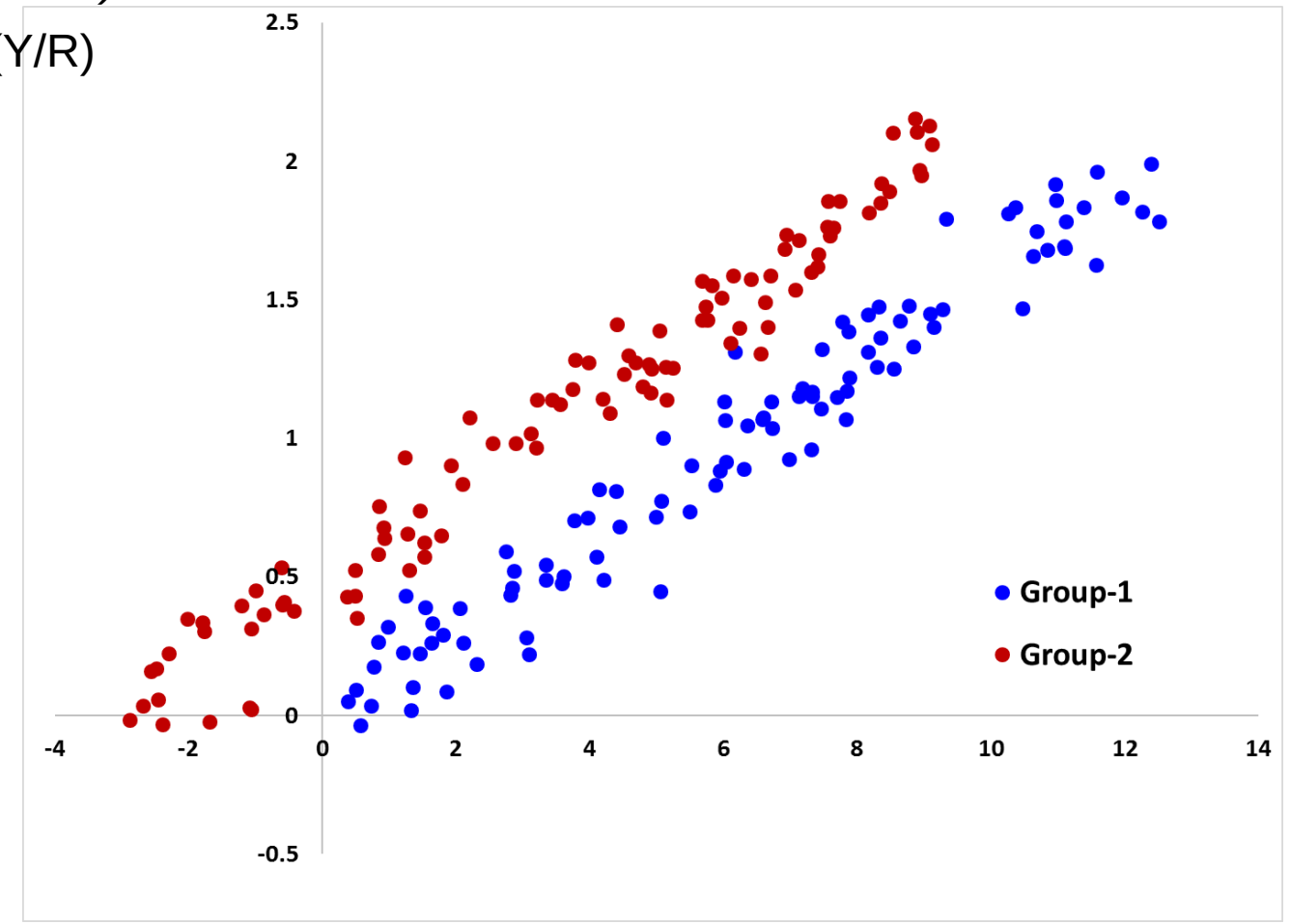


Finding the Right Features

- Transform to polar coordinates

  $R = sqrt(X^2 + Y^2)$

  $\theta = \arcsin(Y/R)$

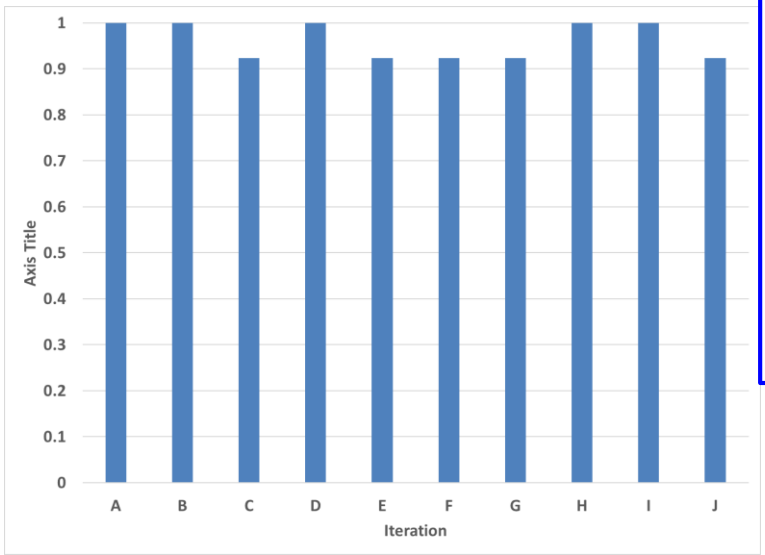


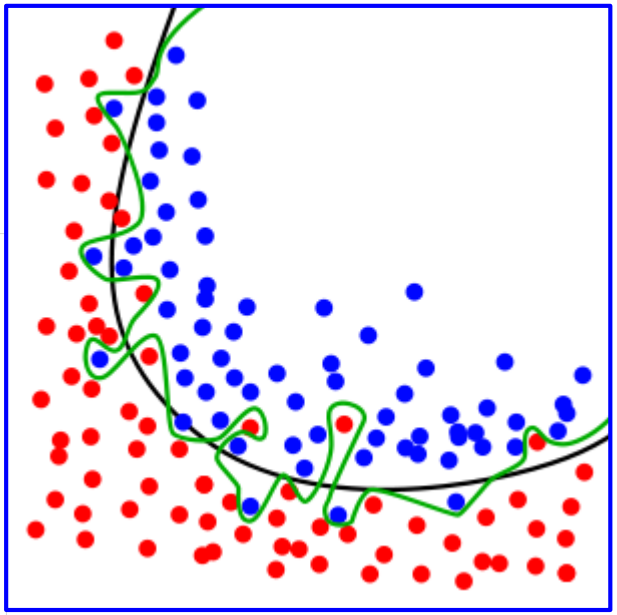
Overfitting

- The model f approximates the relationship between features (explanatory variables) X and the target (dependent variable) Y:

  $$Y = f(X)$$

- Overfitting occurs when the model is driven by the noise or random behavior, rather than the fundamental relationship
- Example:
  - *Two-class problem with N = 25*
  - *Candidate features: All possible pairs of 100 realizations of random Gaussian data*





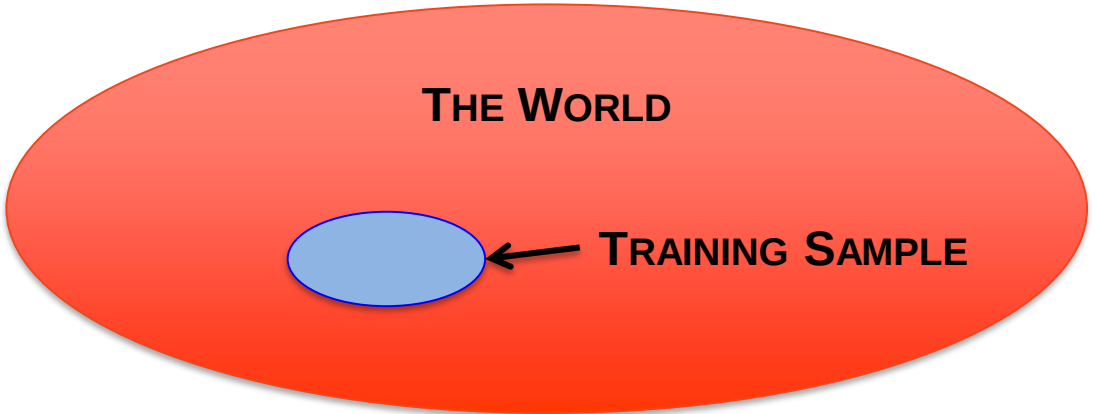
https://en.wikipedia.org/wiki/Overfitting

DRAPER

Validation

- How well does the classifier perform on new data?
  - ***Ideal**: Sequestered set of validation data*
  - ***Limited** data: Multiple methods for partitioning data into training and testing: k-fold validation, leave-one-out, repeated random sub-sampling*

- Generalization: Training sample and the population:
  - *Is the training representative of the relevant population?*
  - *How do selection criteria for inclusion in the clinical study affect the training sample? And, hence, the extensibility of the findings?*



Conclusions

- Value of classical statistical methods and machine learning depend on:
  - *Amount of available data*
  - *Strength of the "signal" (effect size, separability)*
- Many other factors affect the results:
  - *Feature selection*
  - *Overfitting*
- Relationship of the training sample to the larger population affects generalizability of the findings