Materials Informatics

Lecture 2: Introduction to Statistical Learning

Ulisses M. Braga-Neto, Ph.D.

Associate Professor
Genomic Signal Processing Laboratory (GSPLab)
Center for Bioinformatics and Genomic Systems Engineering (CBGSE)
Center for Translational Environmental Health Research (CTEHR)
Department of Electrical and Computer Engineering
Texas A&M University

Modern Science

- Since at least Galileo Galilei, modern science has been established on two principles
  - A **mathematical model** relating the quantities of interest in a process or phenomenon
  - Careful experimental design and **test of predictions** made by the model
- Prediction, and its accurate assessment, are thus critical to the genuine progress of science.

E.R. Dougherty and U.M. Braga-Neto **"Epistemology of Computational Biology: Mathematical Models and Experimental Prediction as the Basis of Their Validity,"** Vol. 14, No. 1, March 2006, pp. 65–90.

Science is Based on Prediction

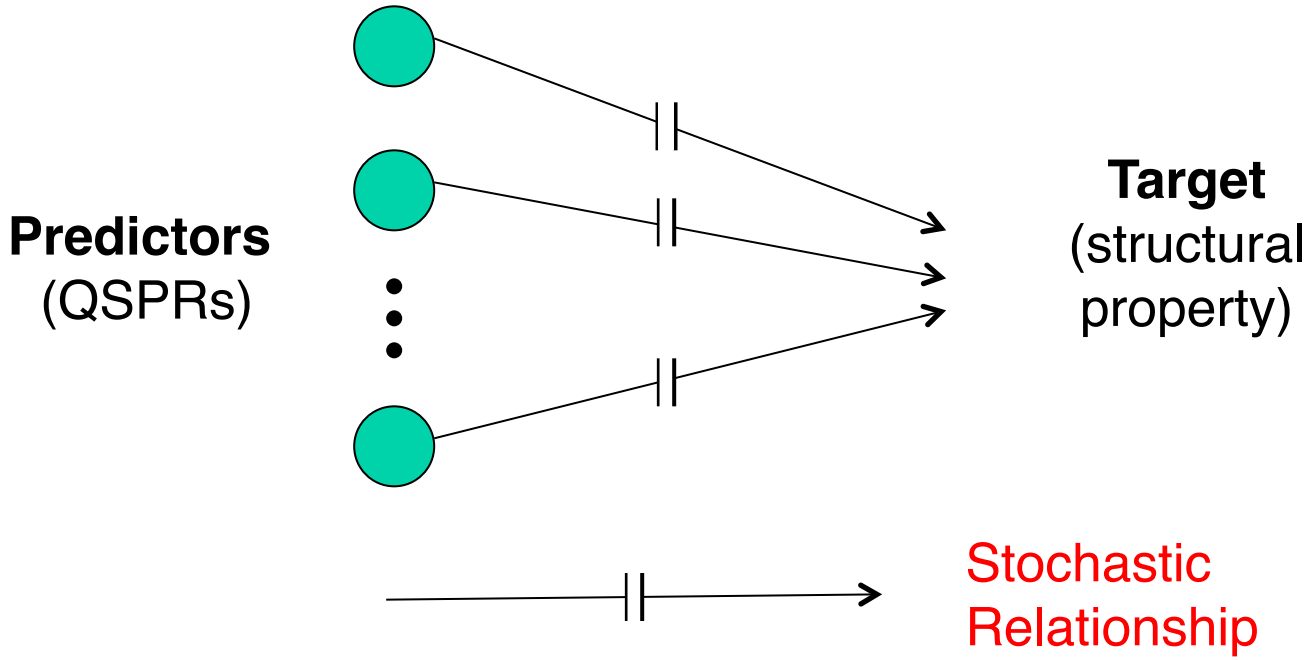
- **Hans Reichenbach** (*Rise of Scientific Philosophy*): "A mere report of relations observed in the past cannot be called knowledge. If knowledge is to reveal objective relations of physical objects, it must include reliable predictions."

Empirical and Mechanistic Models

- **Empirical models** are derived by observation of phenomena and model fitting.
- **Mechanistic models** are founded on basic physical principles instead.
- A good example is provided by the two models for planetary motion (both are deterministic): Kepler's empirical model vs. Newton's mechanistic model .
- Deterministic and mechanistic modeling are largely ineffective due to the presence of significant unexplained variability.

Stochastic Prediction

- Goal: **stochastic empirical models.**



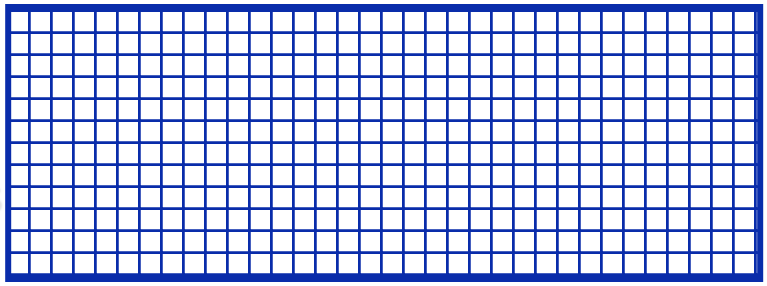
Small-Sample High-Dimensional Data

The prediction model must be inferred from the sample data

Classical Application

100's-1000's cases

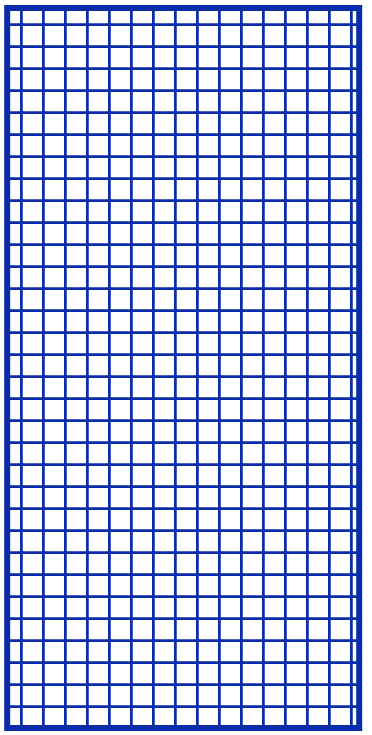
10's-100's variables



Modern Application

10's-100's cases

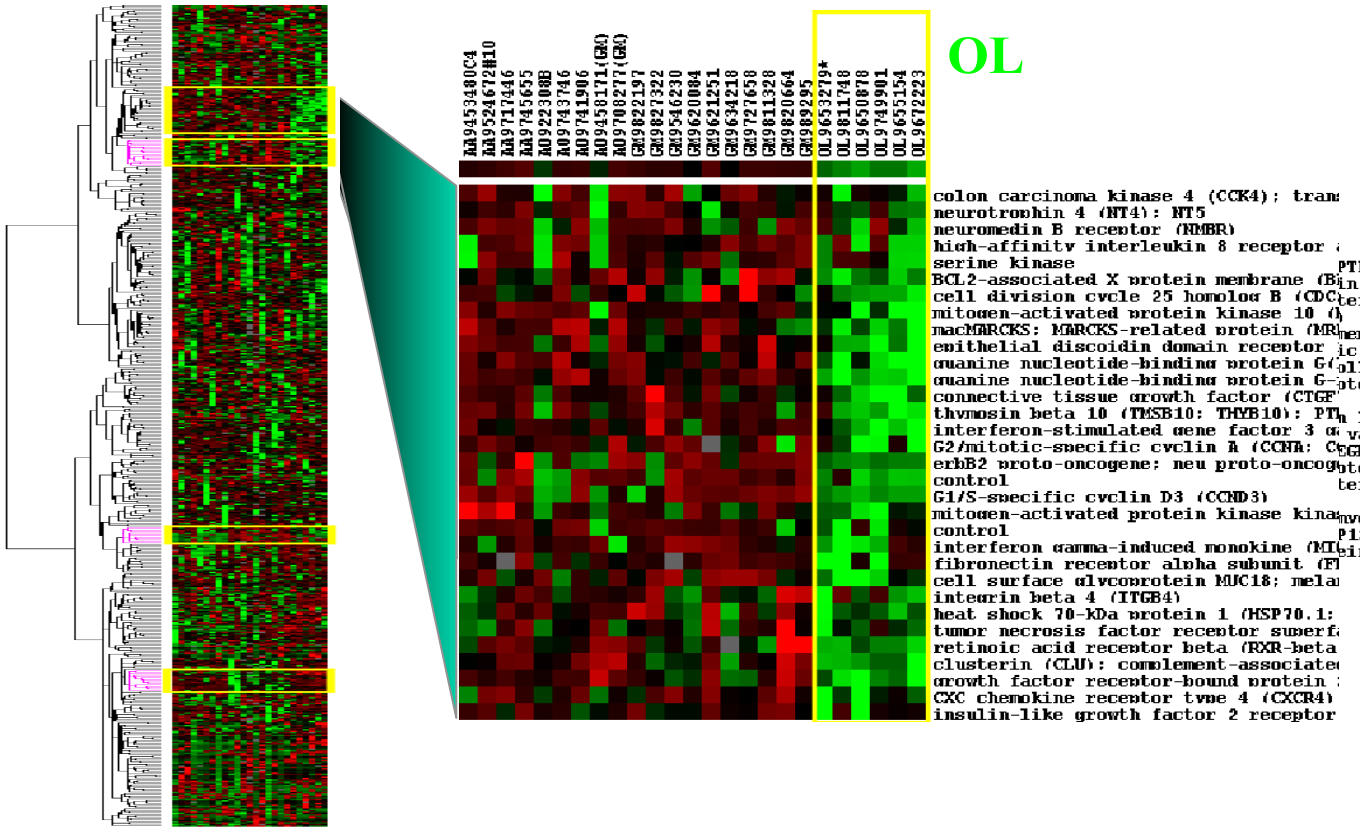
100's-1000's variables



From the statistical point of view, the ratio sample size/dimensionality is fundamental.

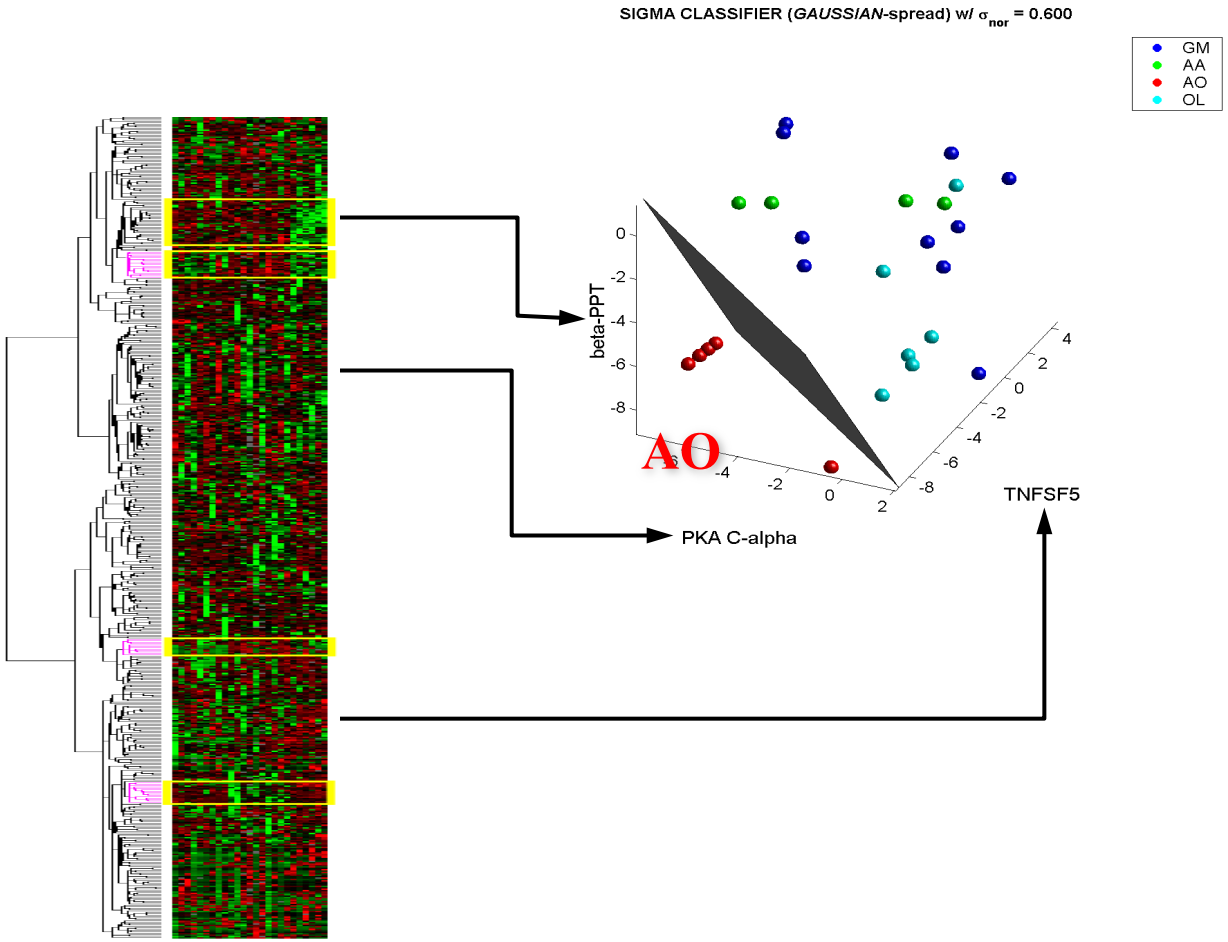
TEXAS A&M UNIVERSITY

Classification



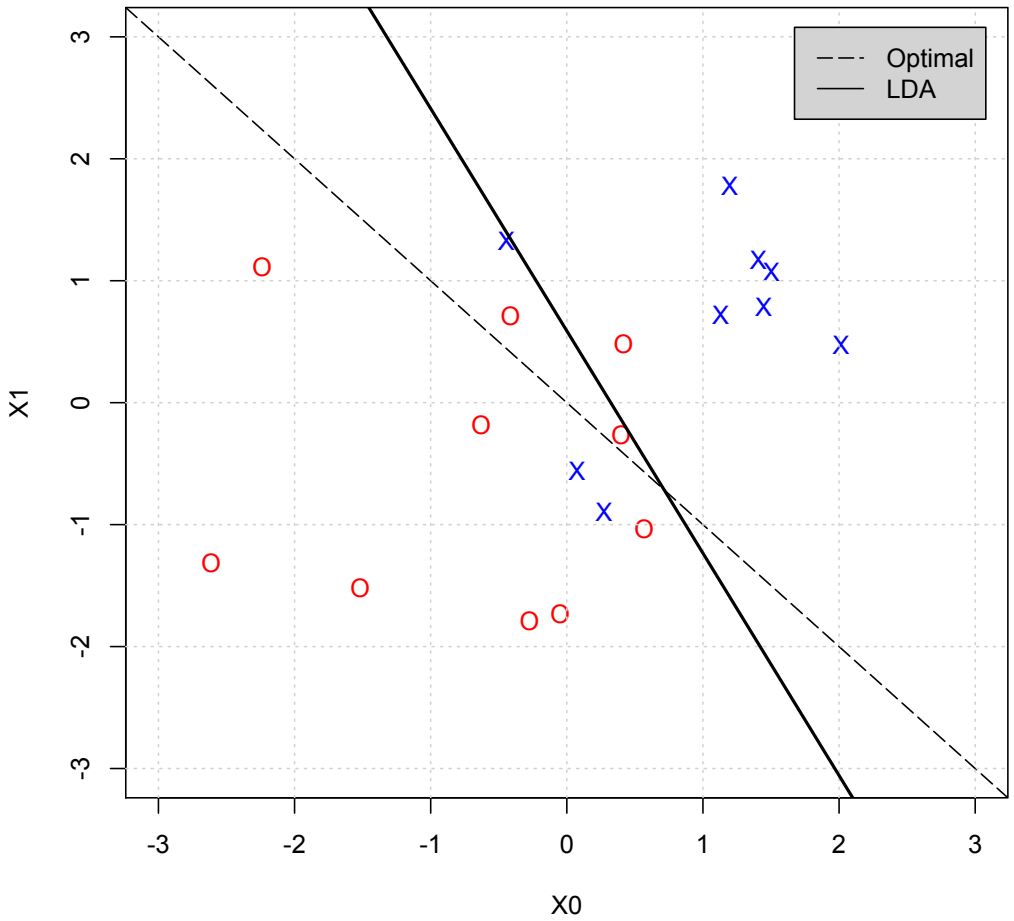
Kim, S. *et al*. "Identification of Combination Gene Sets for Glioma Classification," *Molecular Cancer Therapeutics*, Vol. 1, No. 13, 1229-1236, 2002

Classification



Kim, S. *et al*. "Identification of Combination Gene Sets for Glioma Classification," *Molecular Cancer Therapeutics*, Vol. 1, No. 13, 1229-1236, 2002

Optimal and Designed Classifier



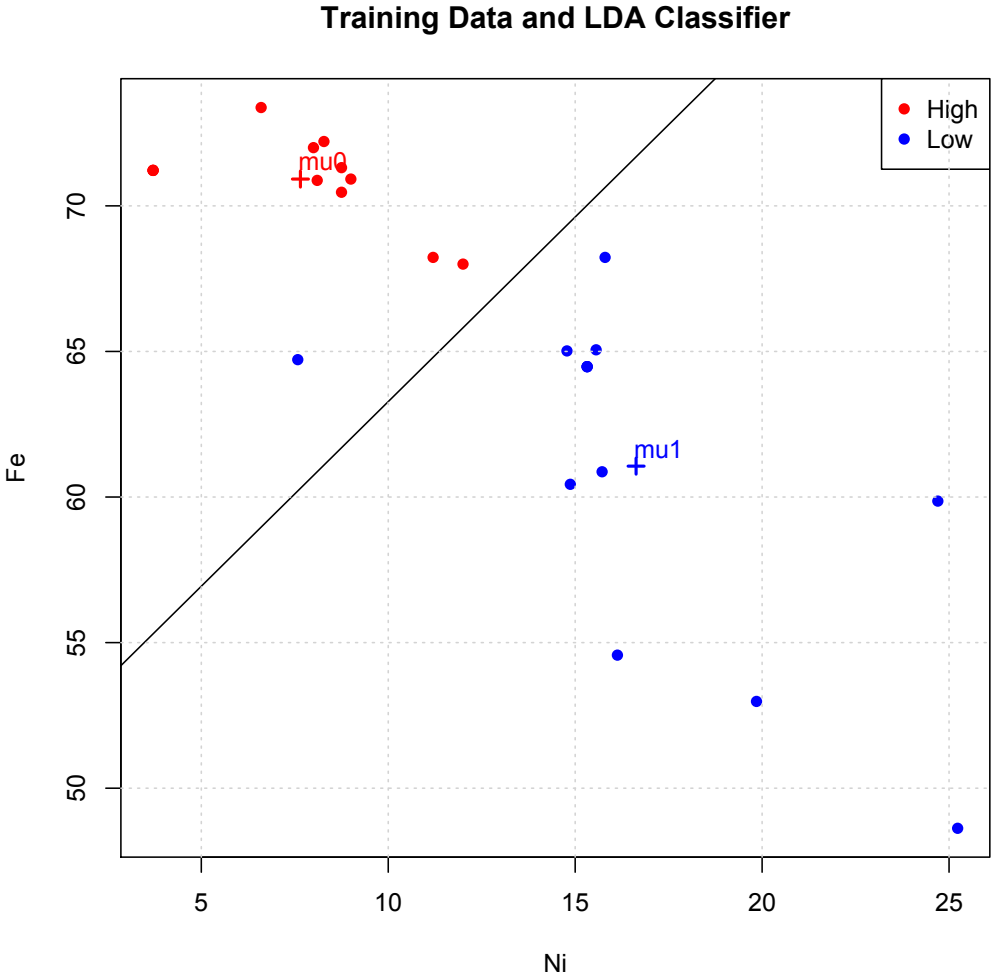
Error of Optimal Classifier = 20%

Error of Designed Classifier = 26%

Test-Error of Designed Classifier = 27%

Apparent Error of Designed Classifier = 4/20 = 20% (optimistic)

Materials Data



Target Property = Stacking Fault Energy (SFE)

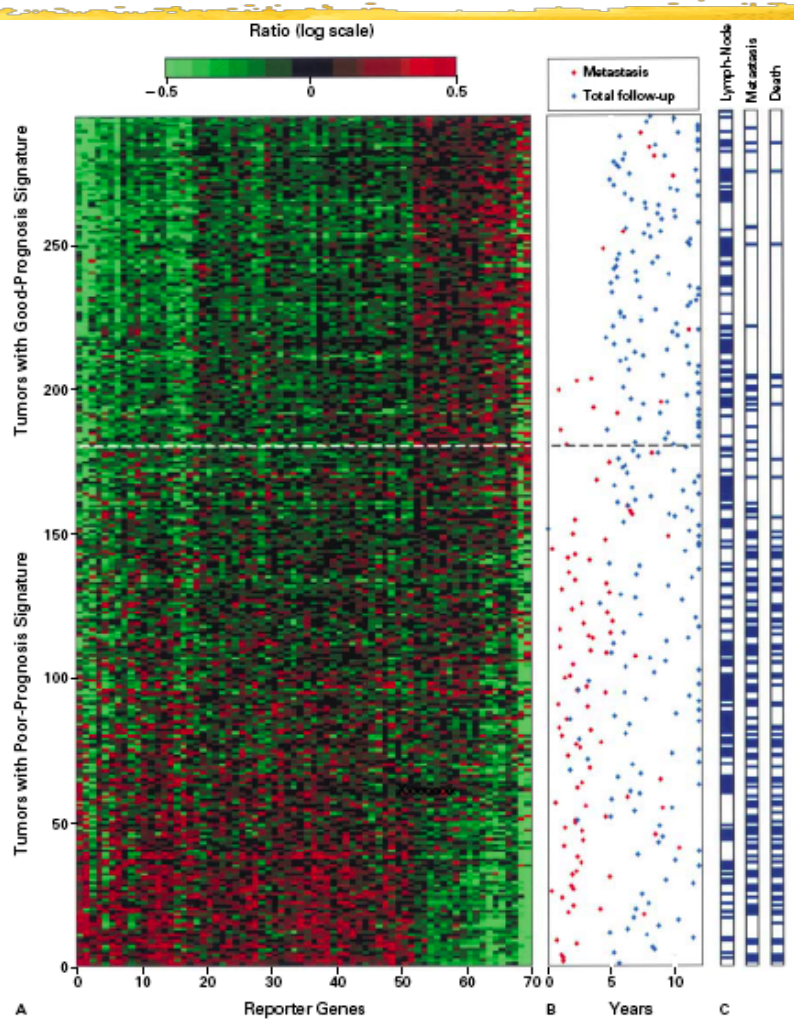
Predictors = Atomic Compositions

Categorize SFE into two classes, low (SFE ≤ 35) and high (SFE ≥ 45).

Test-Set Error of Designed Classifier = 5/82 = 6.1%

Apparent Error of Designed Classifier = 1/21 = 4.8% (optimistic)

Another Genomics Example

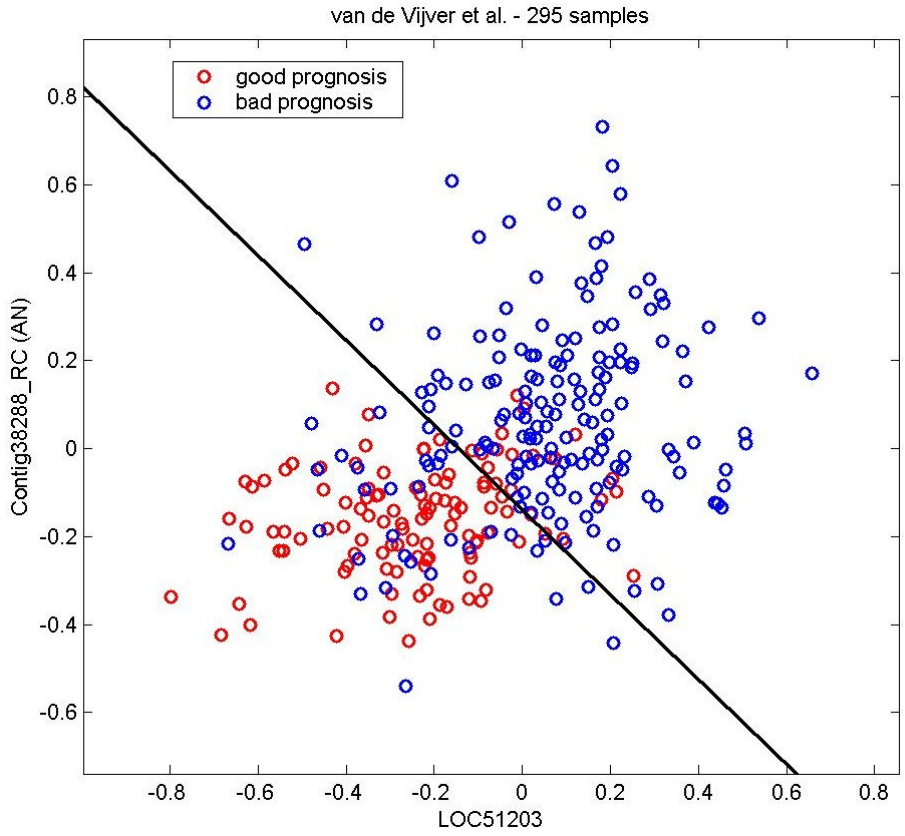


van de Vijver, et al. (2002) "A gene-expression signature as apredictor of survival in breast cancer." New England Journal of Medicine, Vol. 347, 1999–2009.

Originally published 70-gene classifier:

Test-set error = 68/180 = 37.7%

More Accurate Classifier (2 genes)



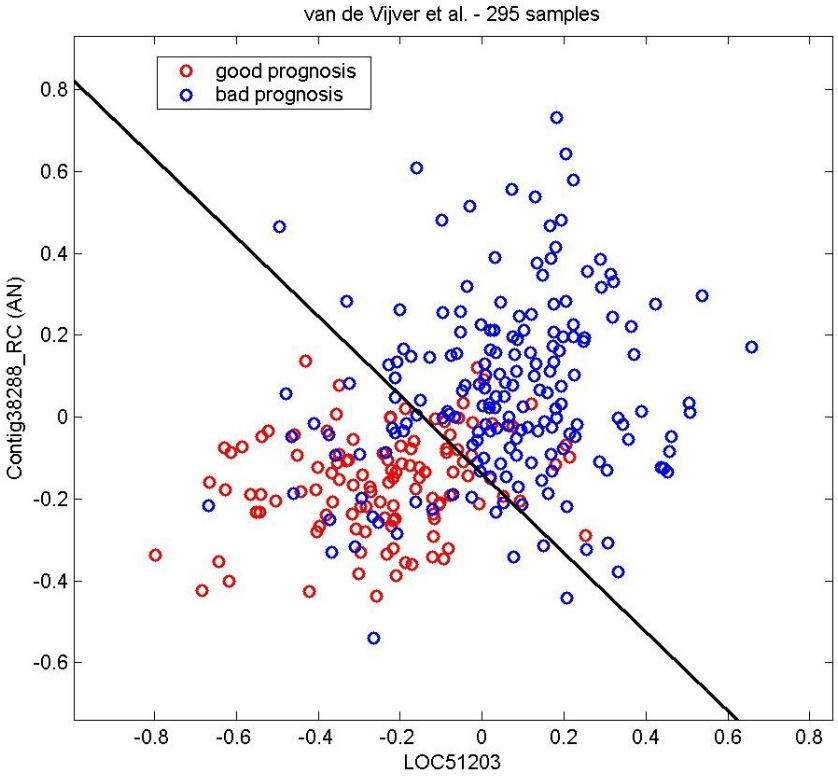
U.M. Braga-Neto, Fads and Fallacies in the Name of Small-Sample Microarray Classification. IEEE Signal Processing Magazine, Special Issue on Signal Processing Methods in Genomics and Proteomics, Vol. 24, No. 1, January 2007, pp. 91-99.

Apparent Error ≈ 52/295 = 17.6%

Feature Selection

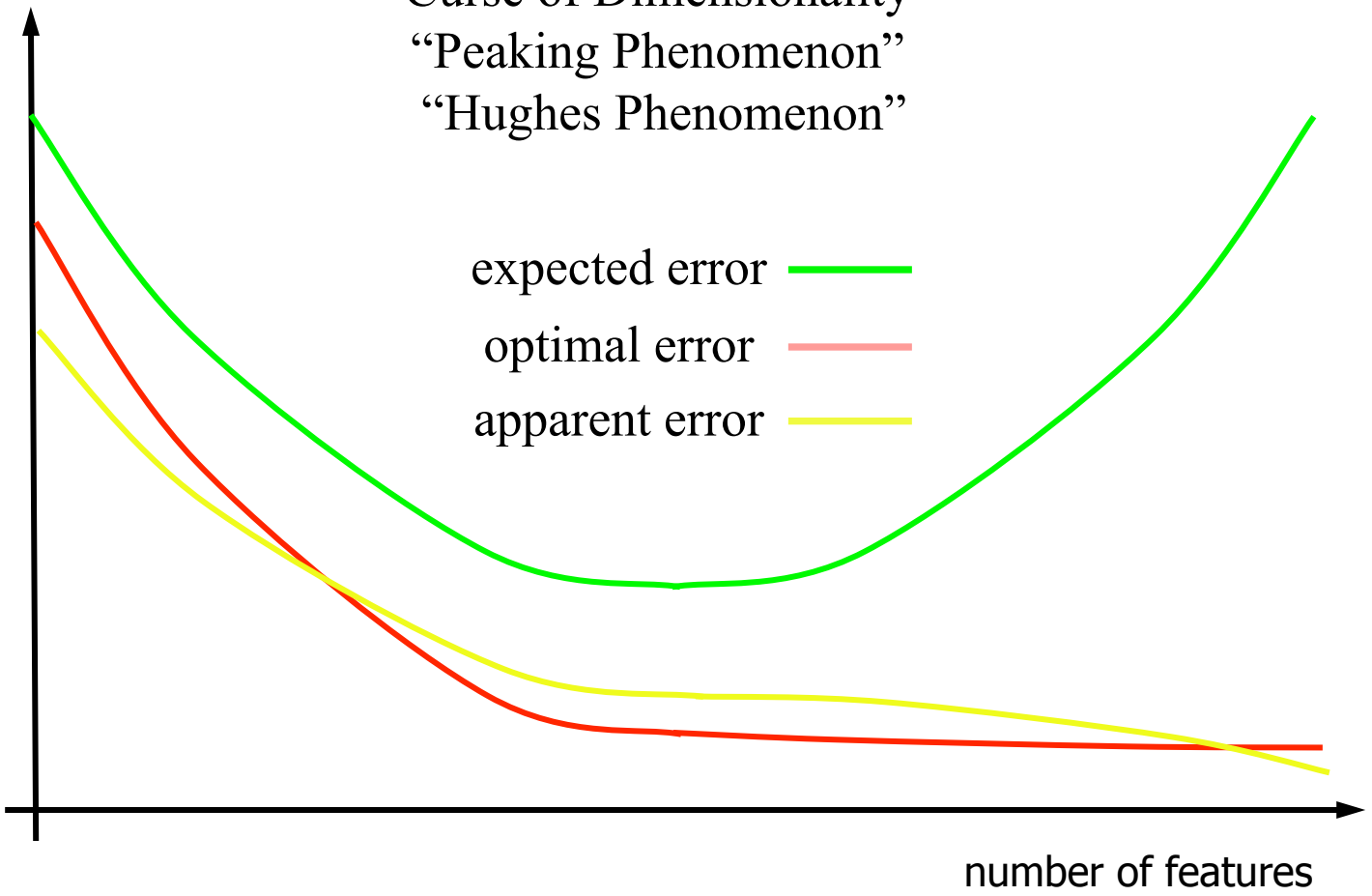
What predictors should I use to classify these data with a small error?

Breast Cancer Data



Feature Selection

"Curse of Dimensionality"
"Peaking Phenomenon"
"Hughes Phenomenon"



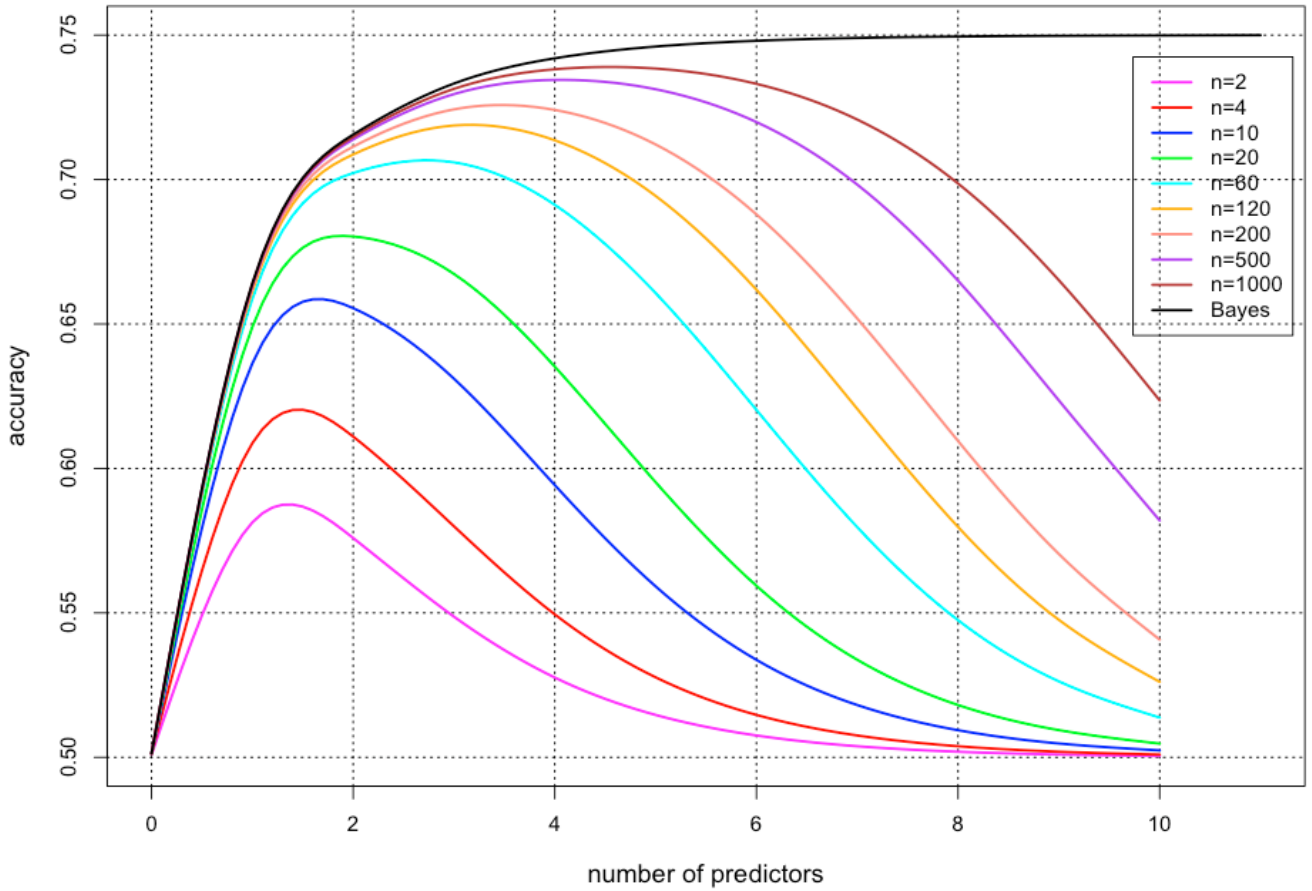
Feature selection can improve accuracy.

"Curse of Dimensionality"

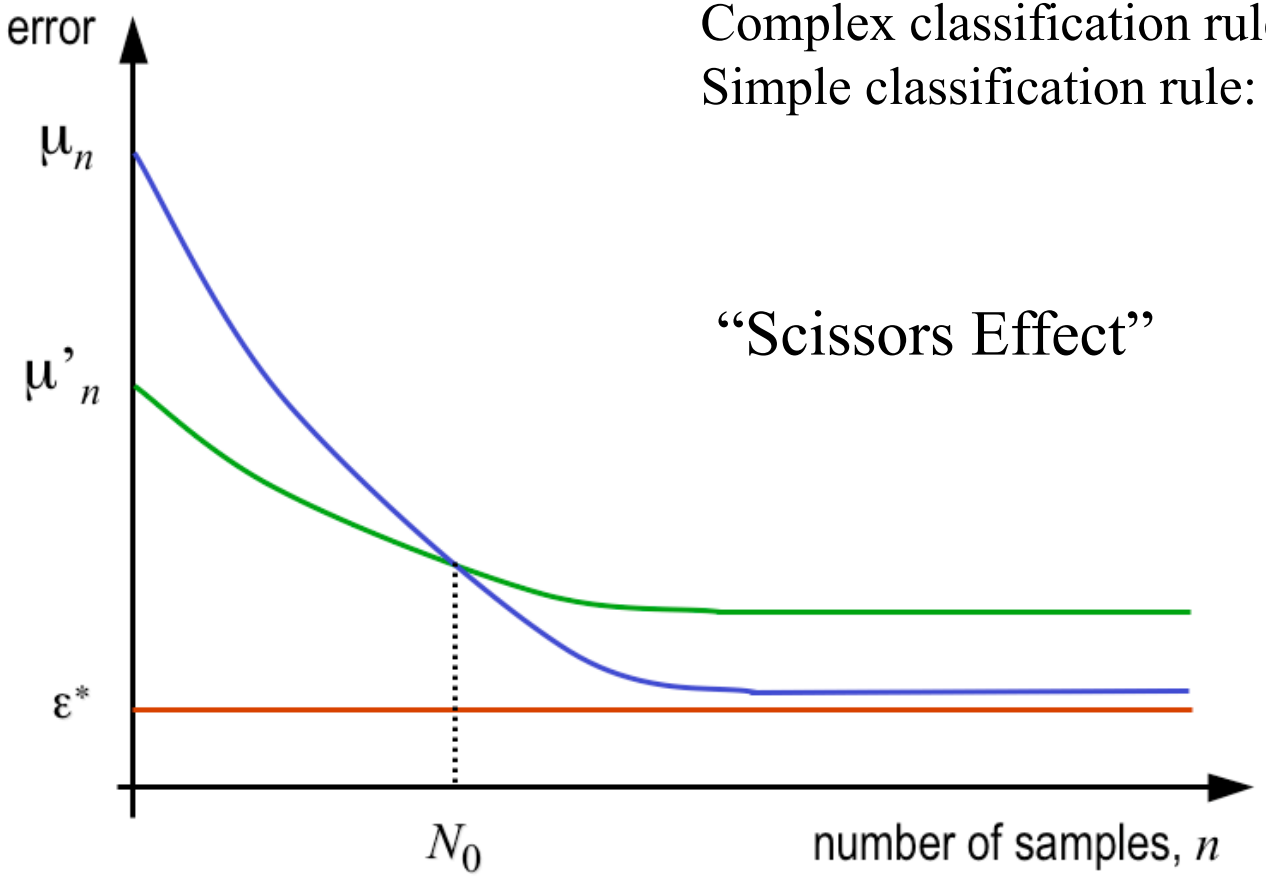
"Peaking Phenomenon" / "Hughes Phenomenon"

G. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, 1968, IT-14, 55-63.

U.M. Braga-Neto, "Classification and Error Estimation for Discrete Data," Current Genomics, Vol. 10, No. 7, November 2009, pp. 446-462.



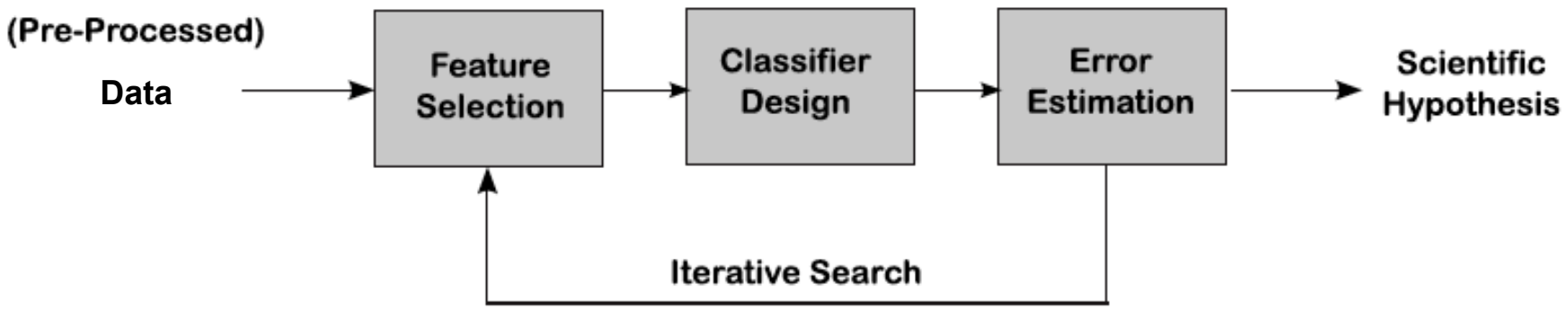
Simple Predictors are Better



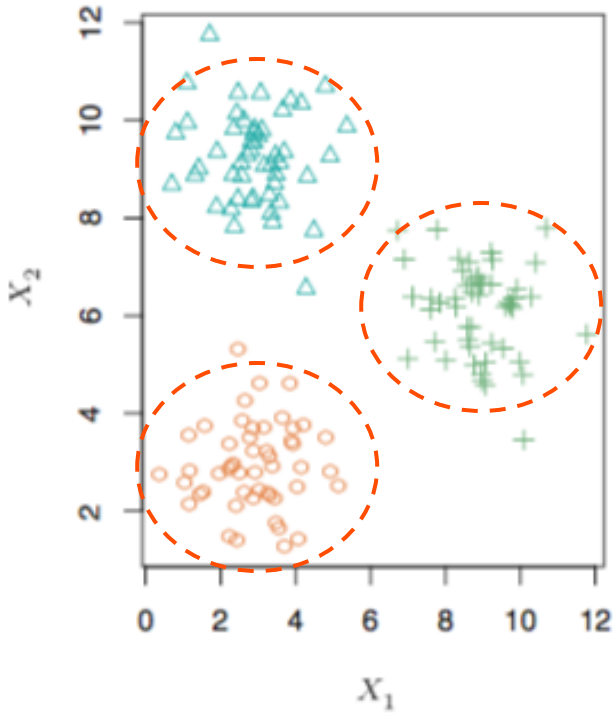
Complex classification rule: blue
Simple classification rule: green

"Scissors Effect"

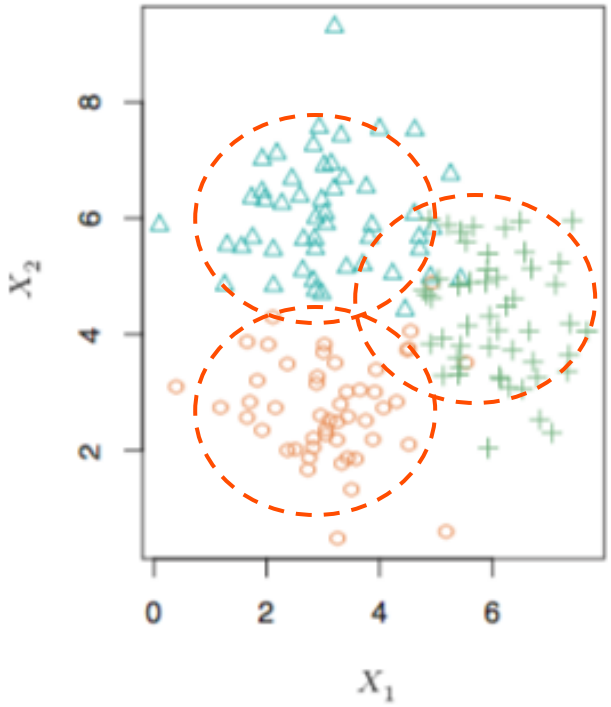
Basic Pipeline



Clustering (Unsupervised Learning)



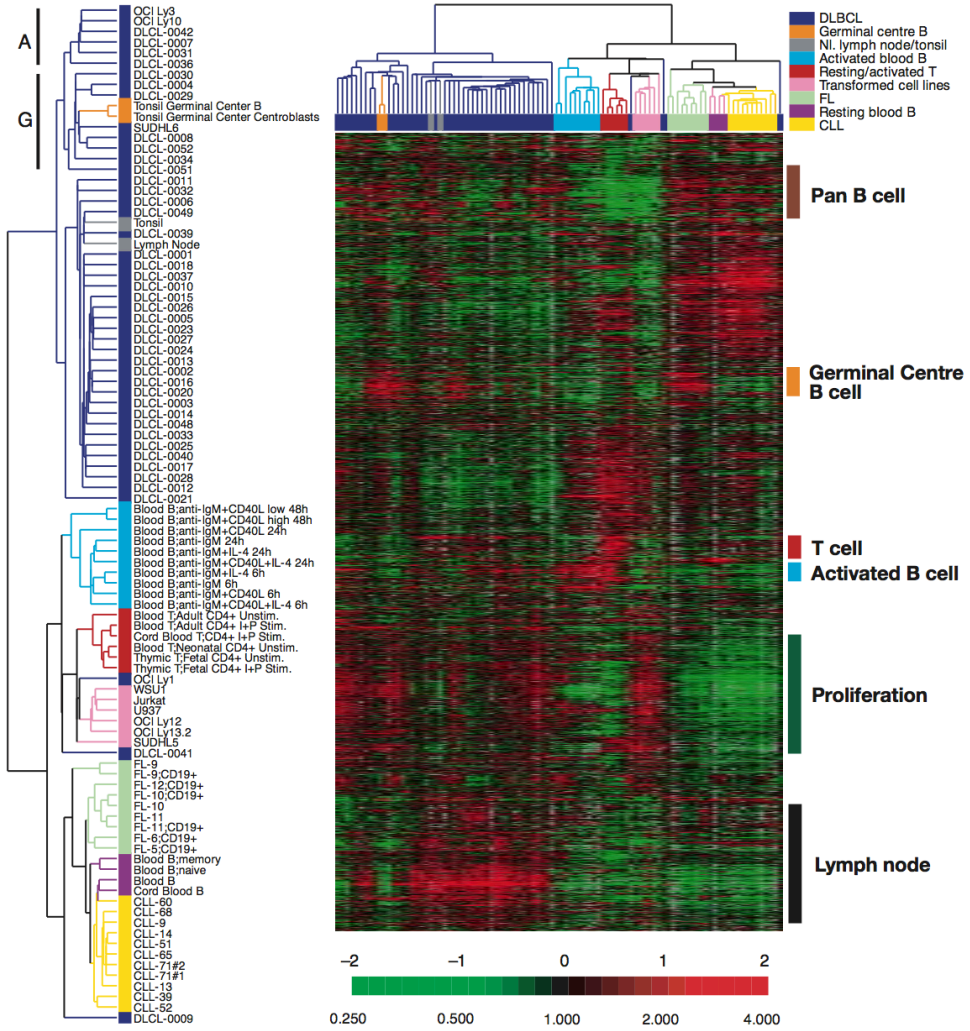
Easy

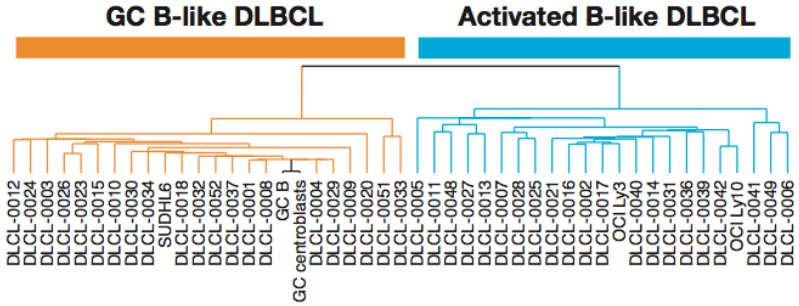


Hard

(Adapted from James et al. "Introduction to Statistical Learning")

Hierarchical Clustering

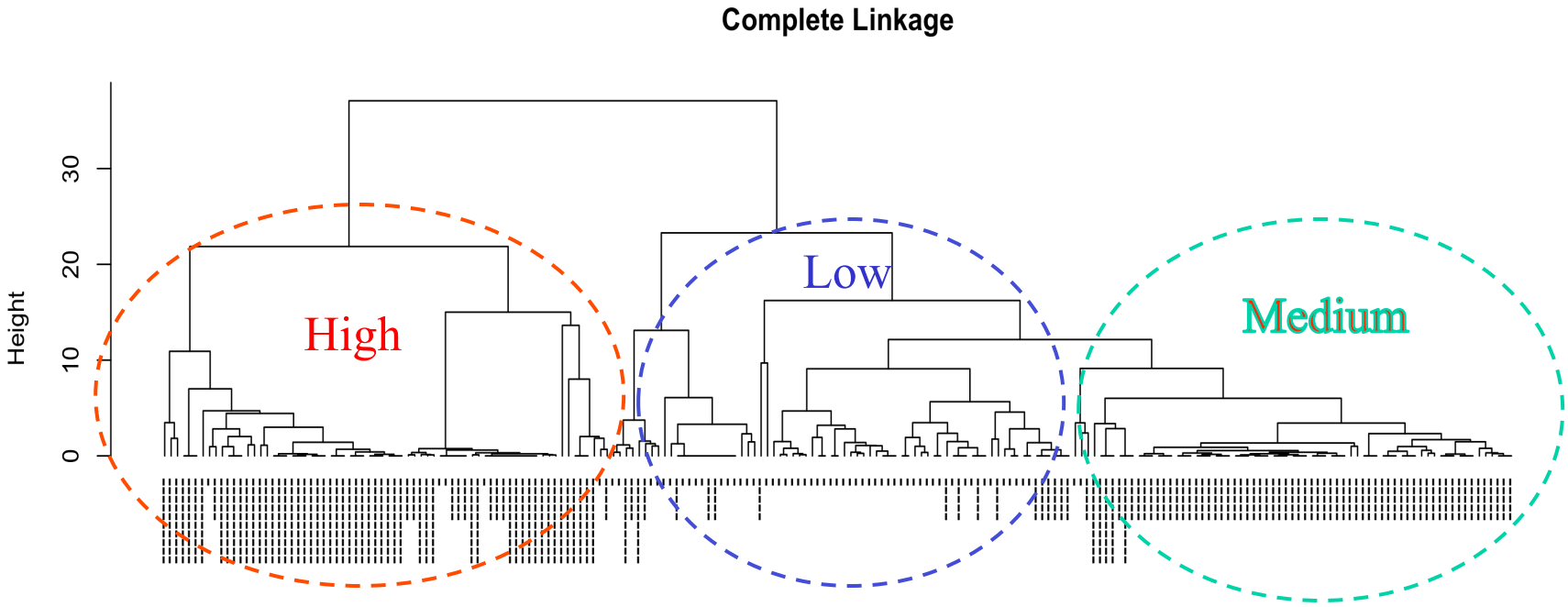




Two new cancer subtypes discovered by clustering

(Adapted from Alizadeh et al. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profling”, Nature, 403, 2000)

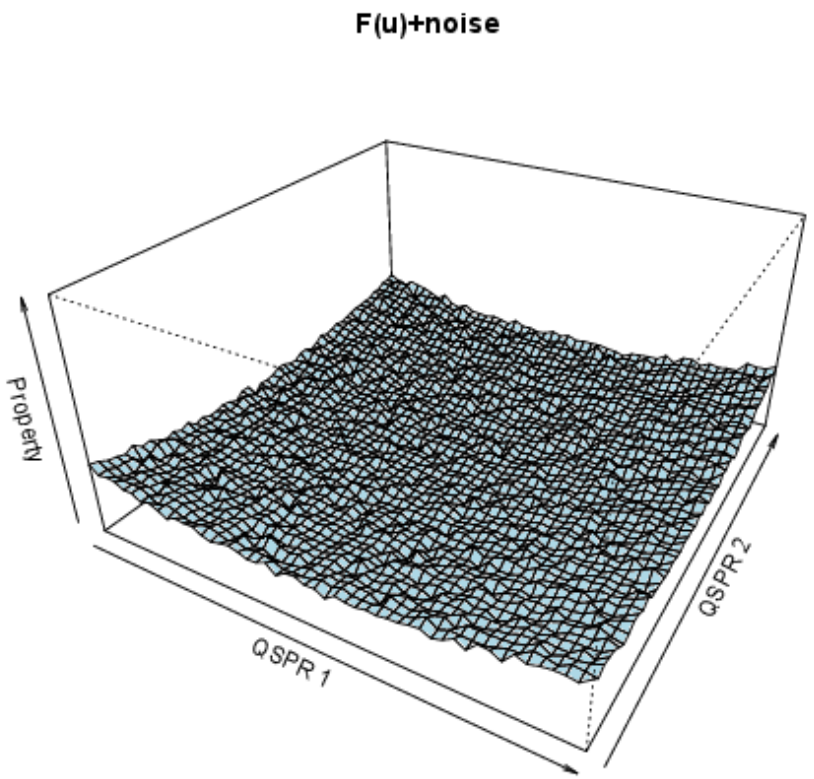
Materials Data



"High," "middle," and "low" SFE responses are indicated by the length of the stick at each leaf (terminal node).

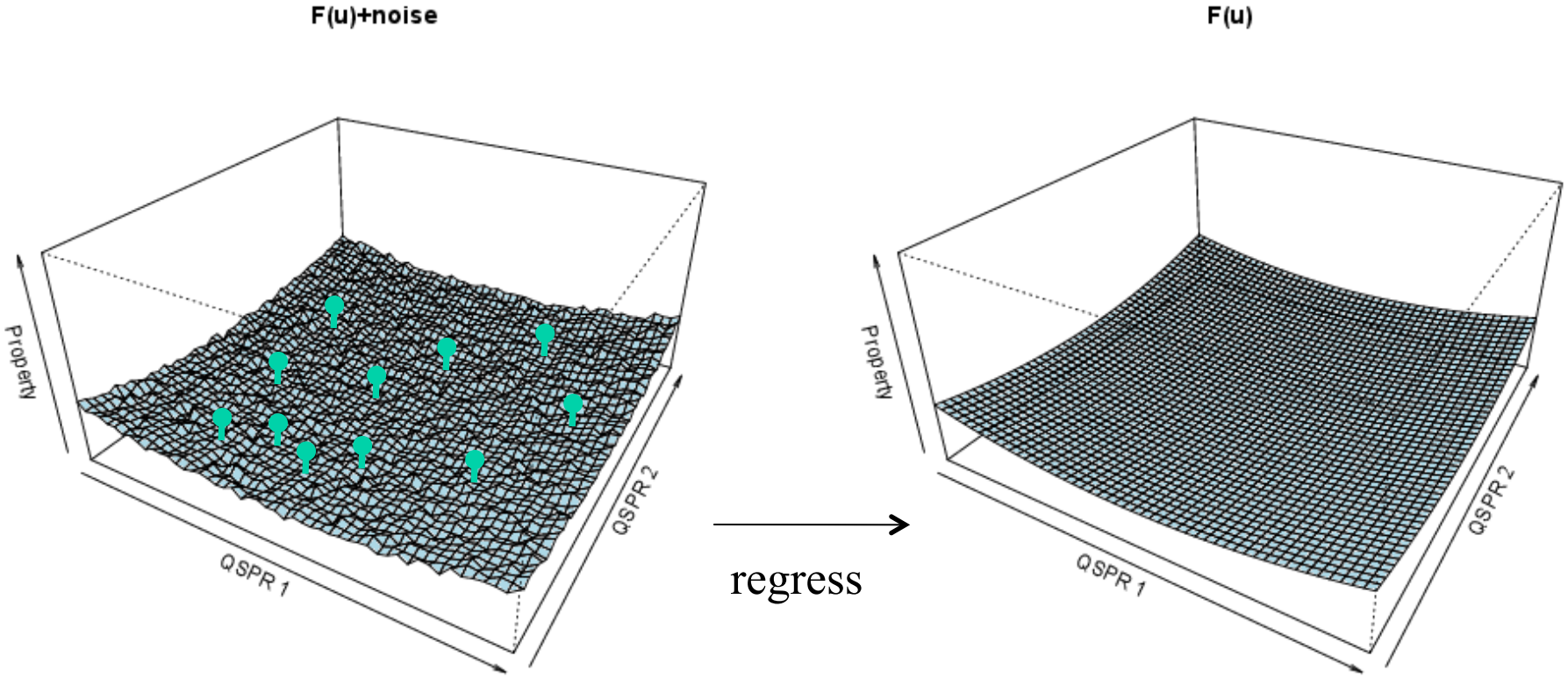
Regression

- General Regression Model: $Y = f(X) + \varepsilon$



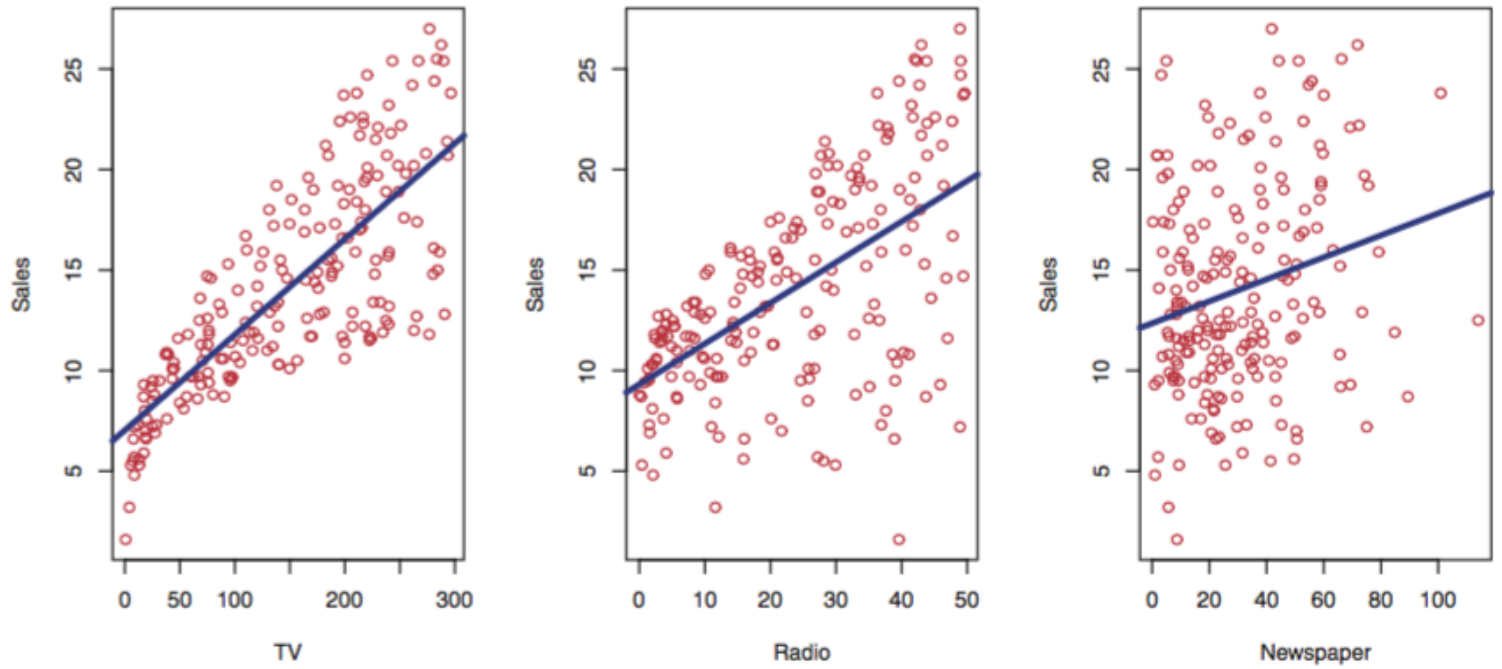
Nonparametric Regression

- Use sample data to obtain a smoothed approximation to the ideal response F(u)



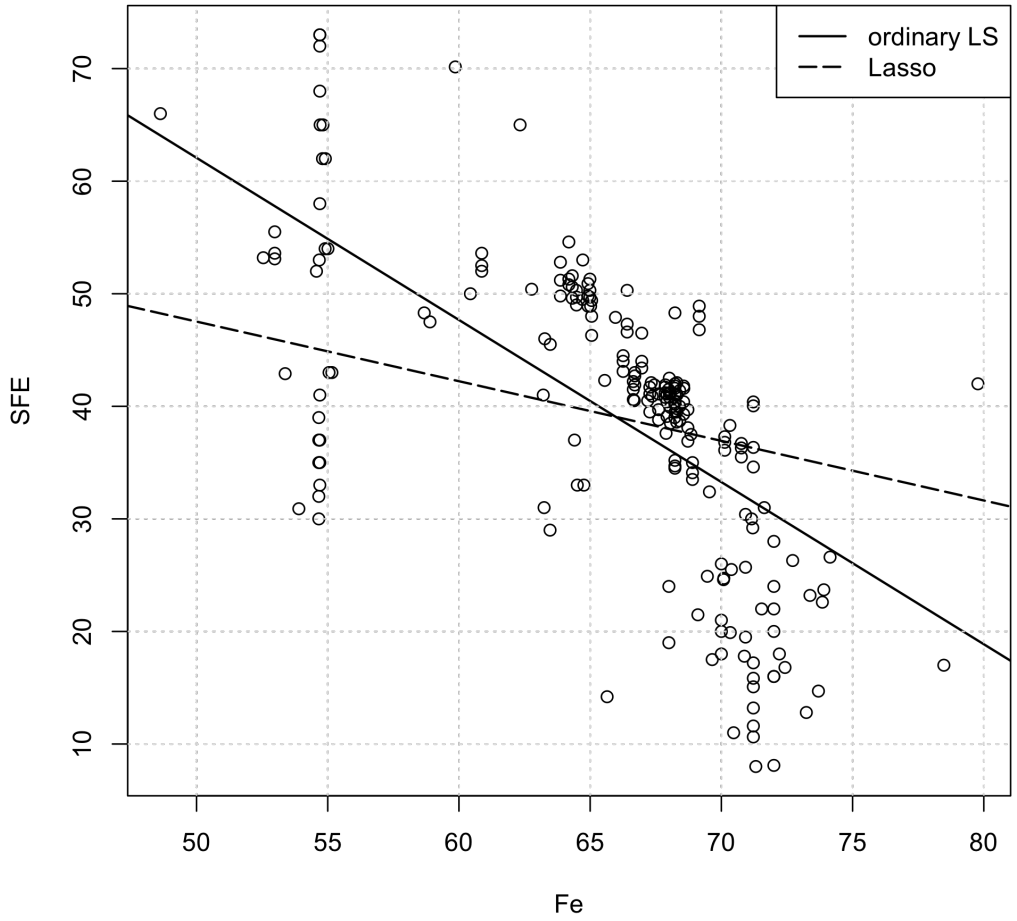
Parametric Regression

- Impose model. E.g. $Y = a^T X + b$ (linear regression).



(Adapted from James et al. "Introduction to Statistical Learning")

Materials Data



Predict SFE values via linear regression models using ordinary least squares, ridge regression, and LASSO.