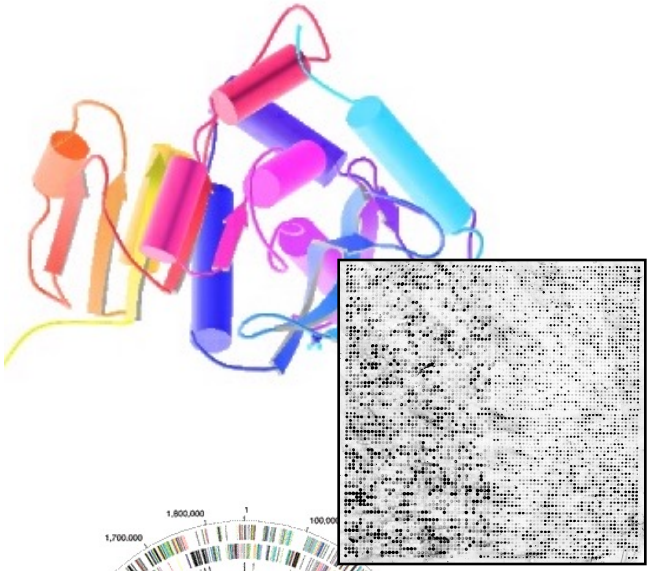
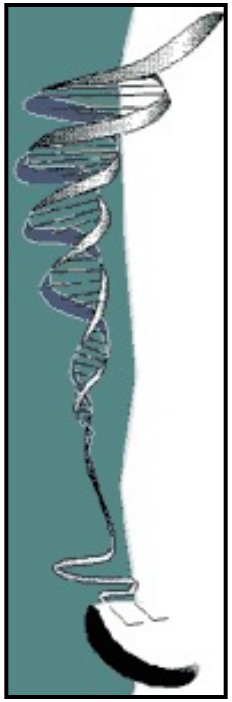
Biomedical Data Science:
Supervised Datamining #1 – Decision Trees





Mark Gerstein, Yale University

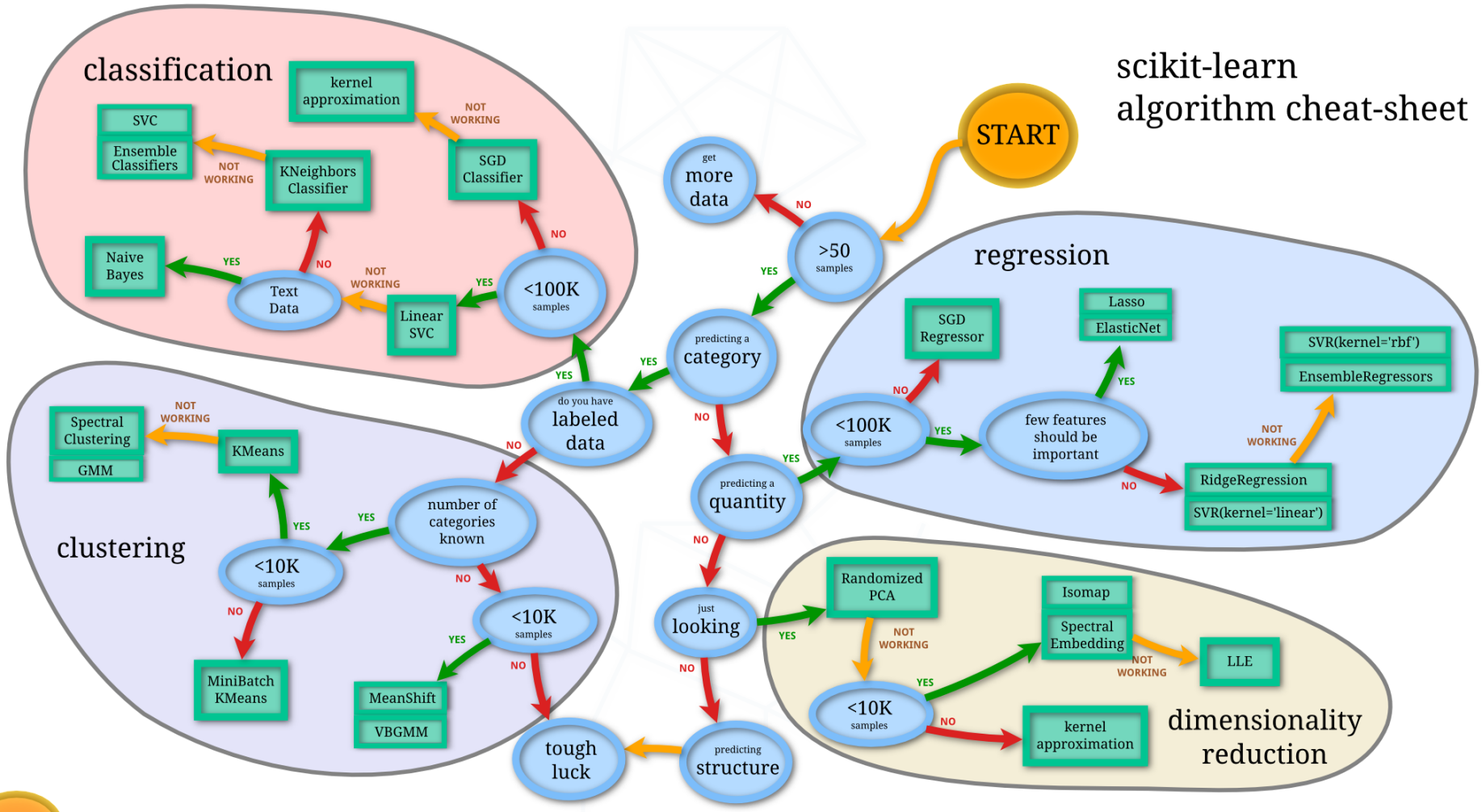
GersteinLab.org/courses/452

(last edit in spring ’21, final)

Supervised Mining:

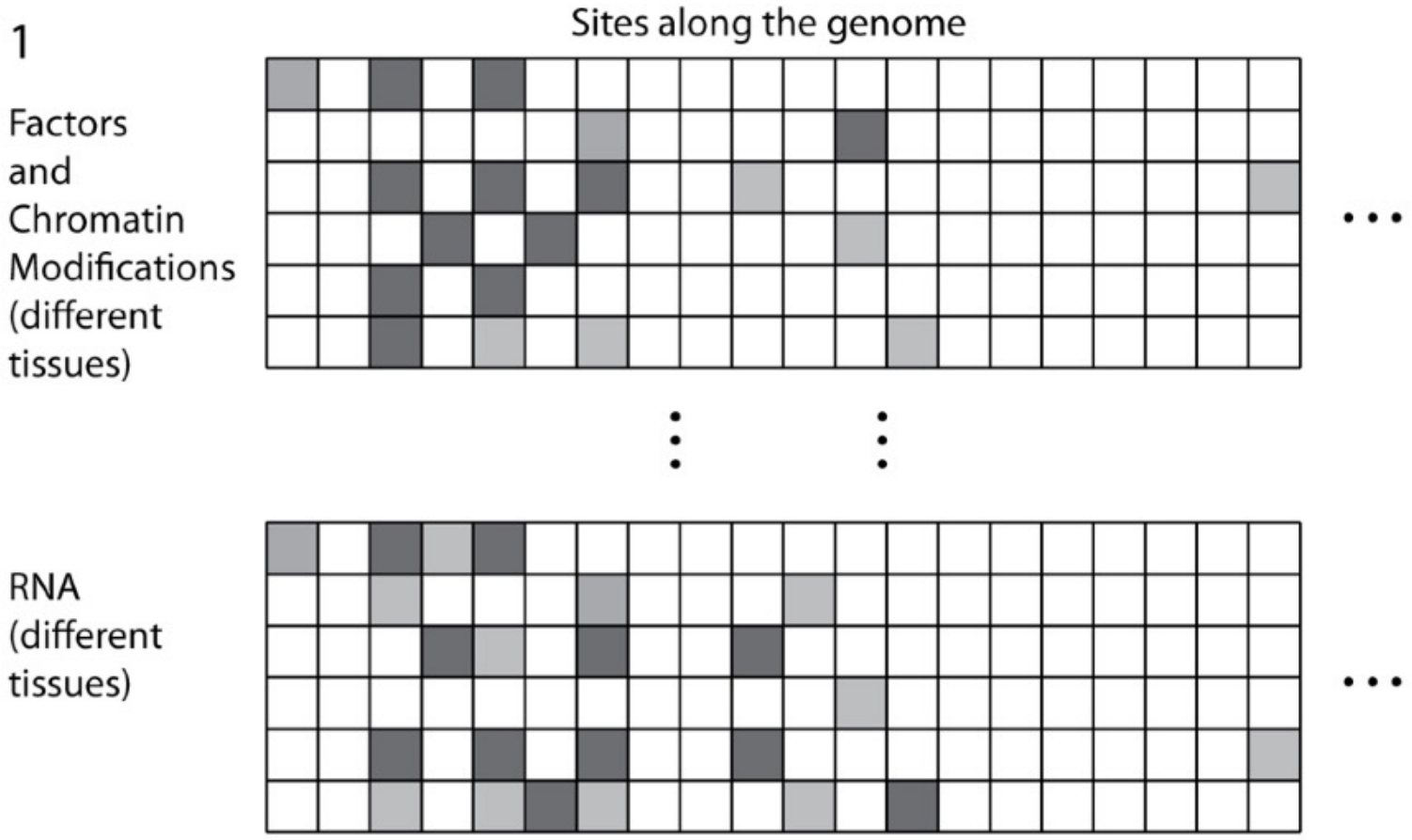
Overview

The World of Machine Learning



SciKit learn: http://scikit-learn.org/stable/tutorial/machine_learning_map/

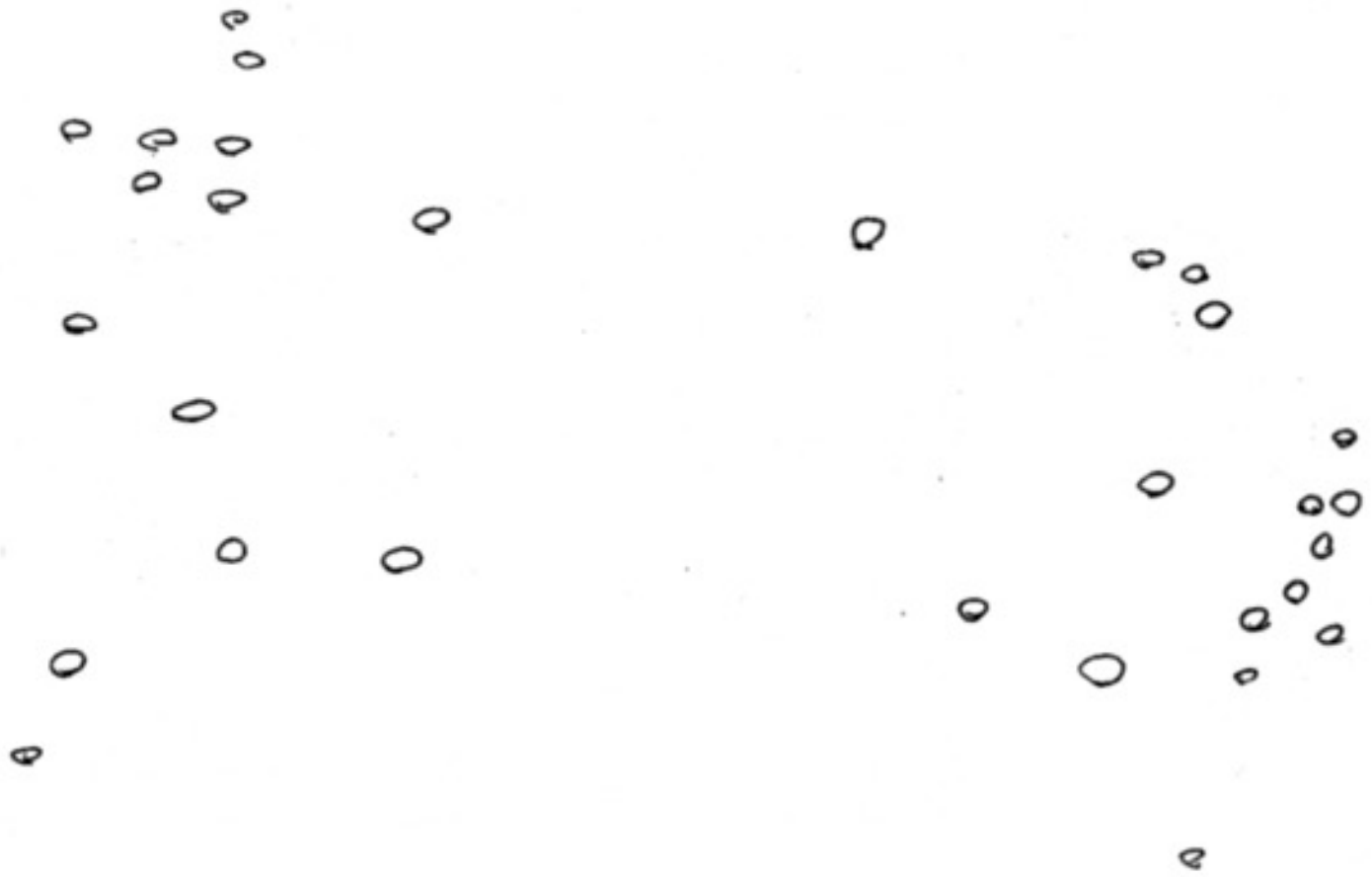
Structure of Genomic Features Matrix



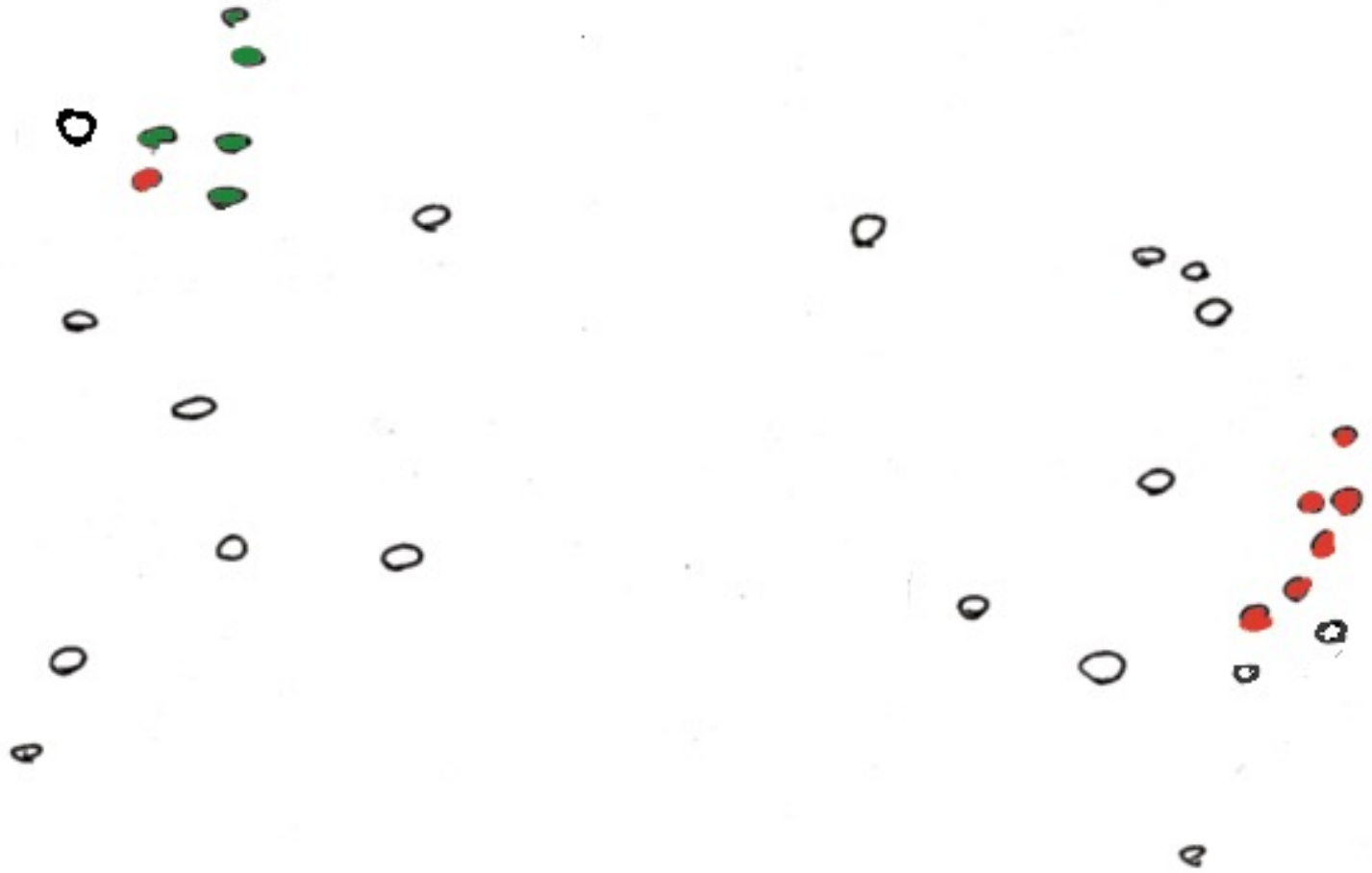
Arrange data in a tabulated form, each row representing an example and each column representing a feature, including the dependent experimental quantity to be predicted.

	predictor1	Predictor2	predictor3	predictor4	response
G1	A(1,1)	A(1,2)	A(1,3)	A(1,4)	Class A
G2	A(2,1)	A(2,2)	A(2,3)	A(2,4)	Class A
G3	A(3,1)	A(3,2)	A(3,3)	A(3,4)	Class B

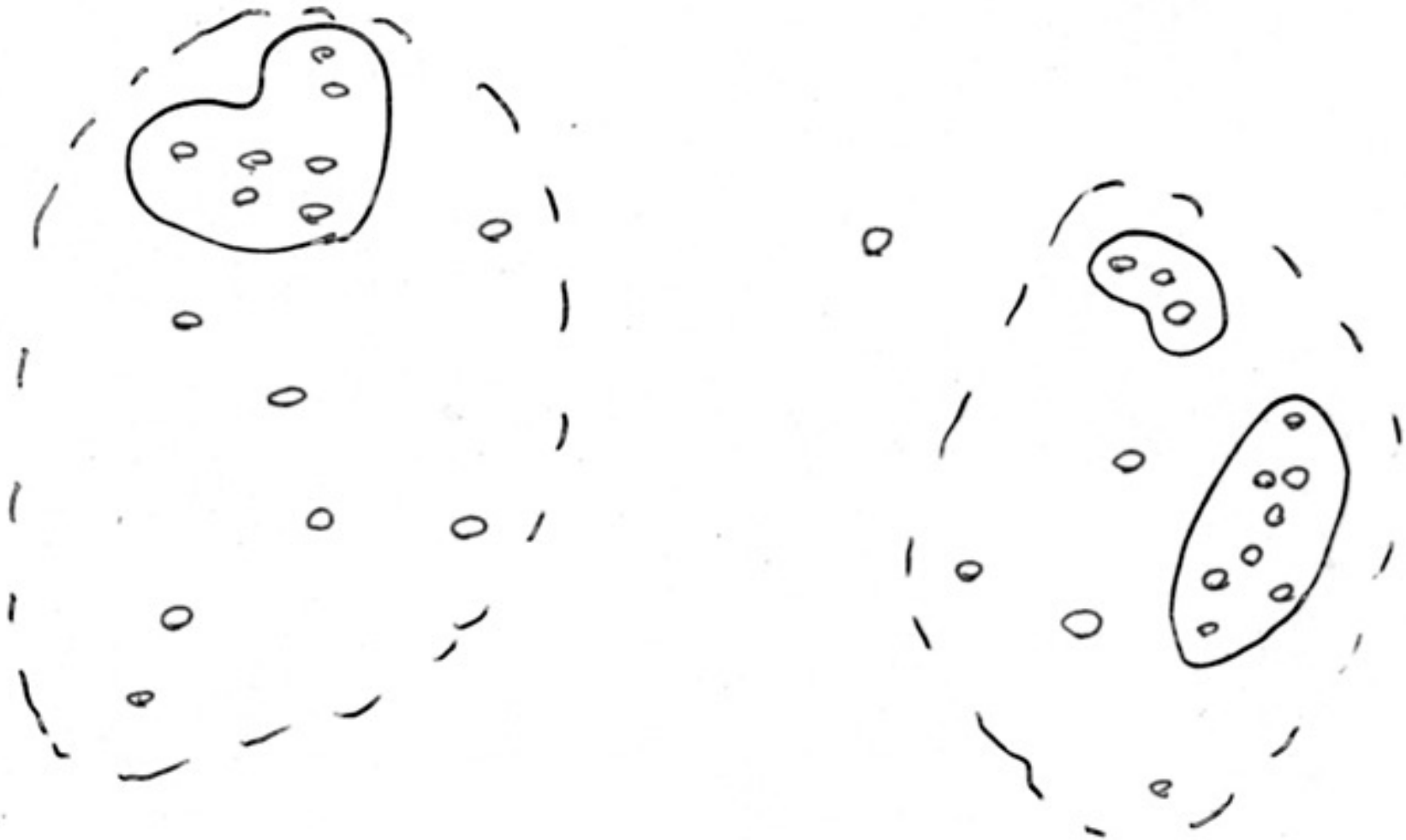
Represent predictors in abstract high dimensional space



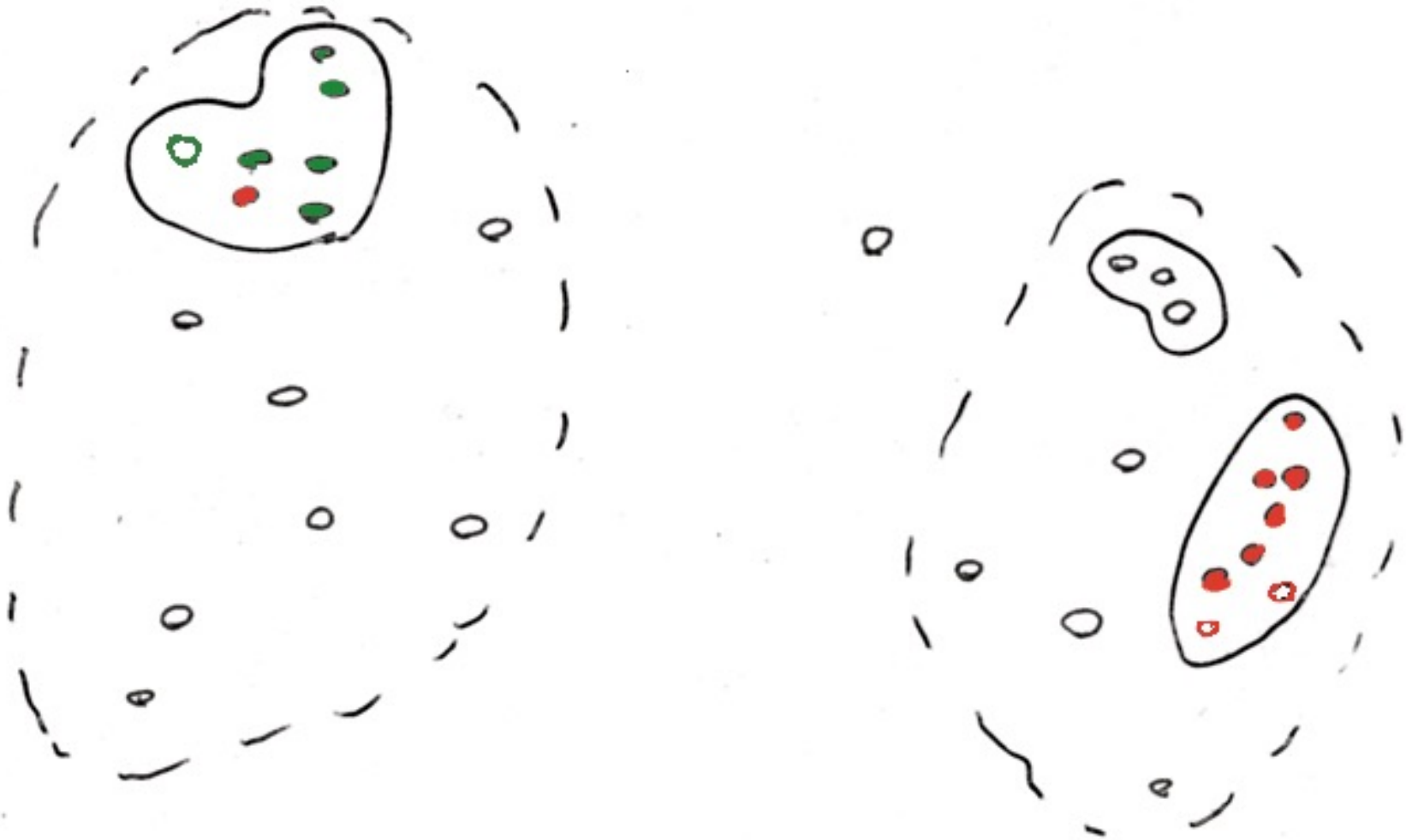
“Label” Certain Points



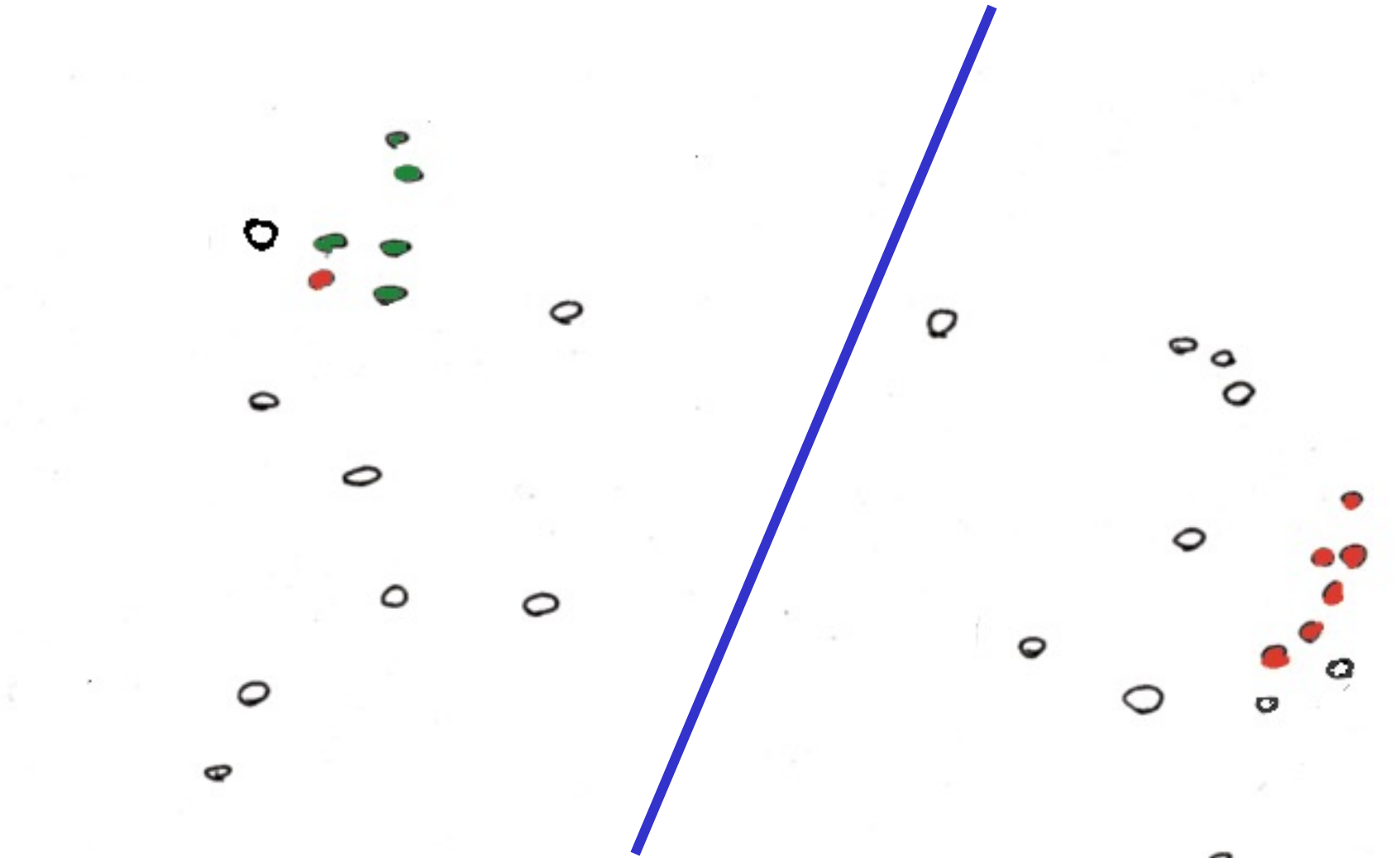
“Cluster” predictors (Unsupervised)



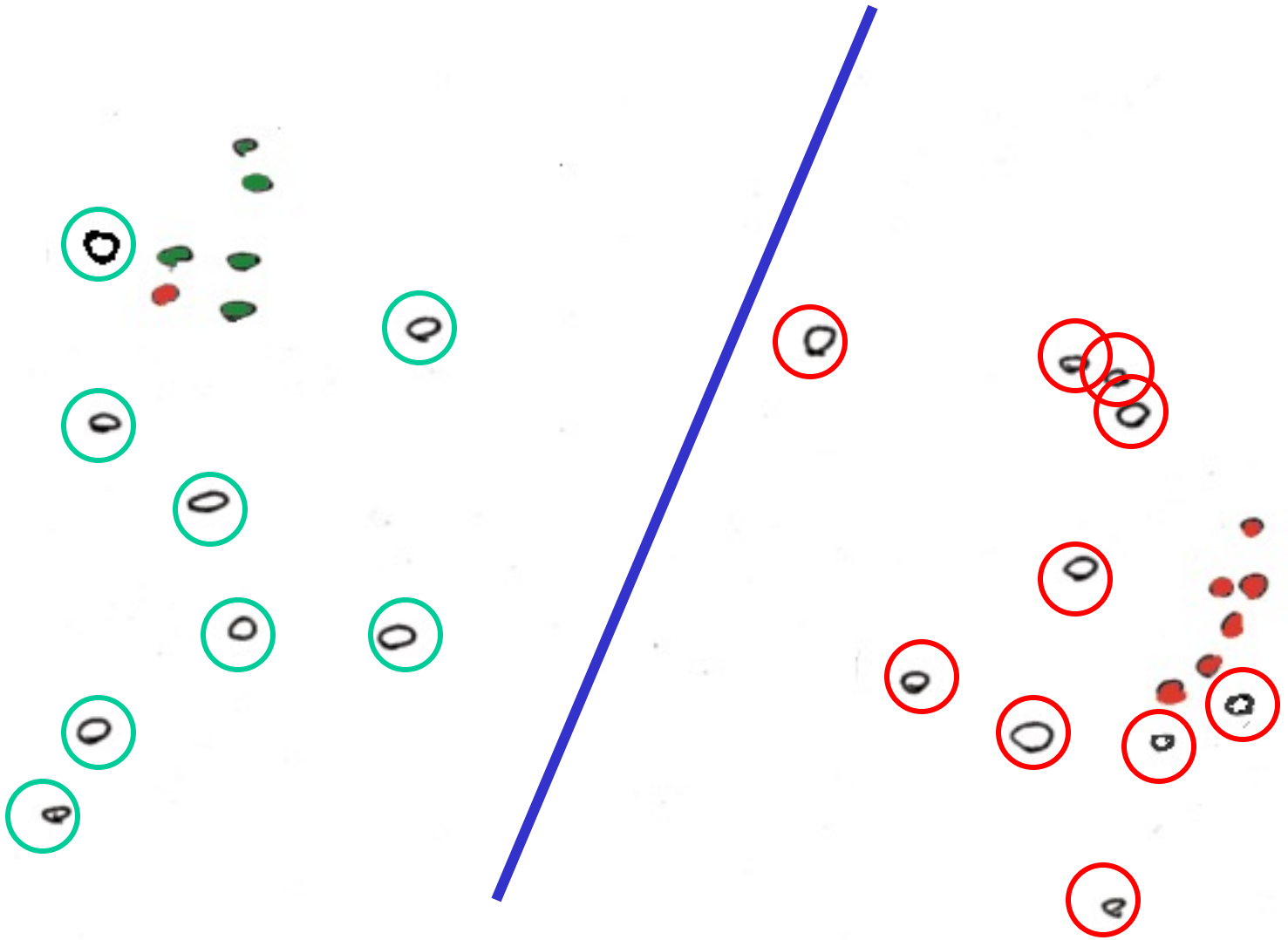
Use Clusters to predict Response
(Unsupervised, guilt-by-association)



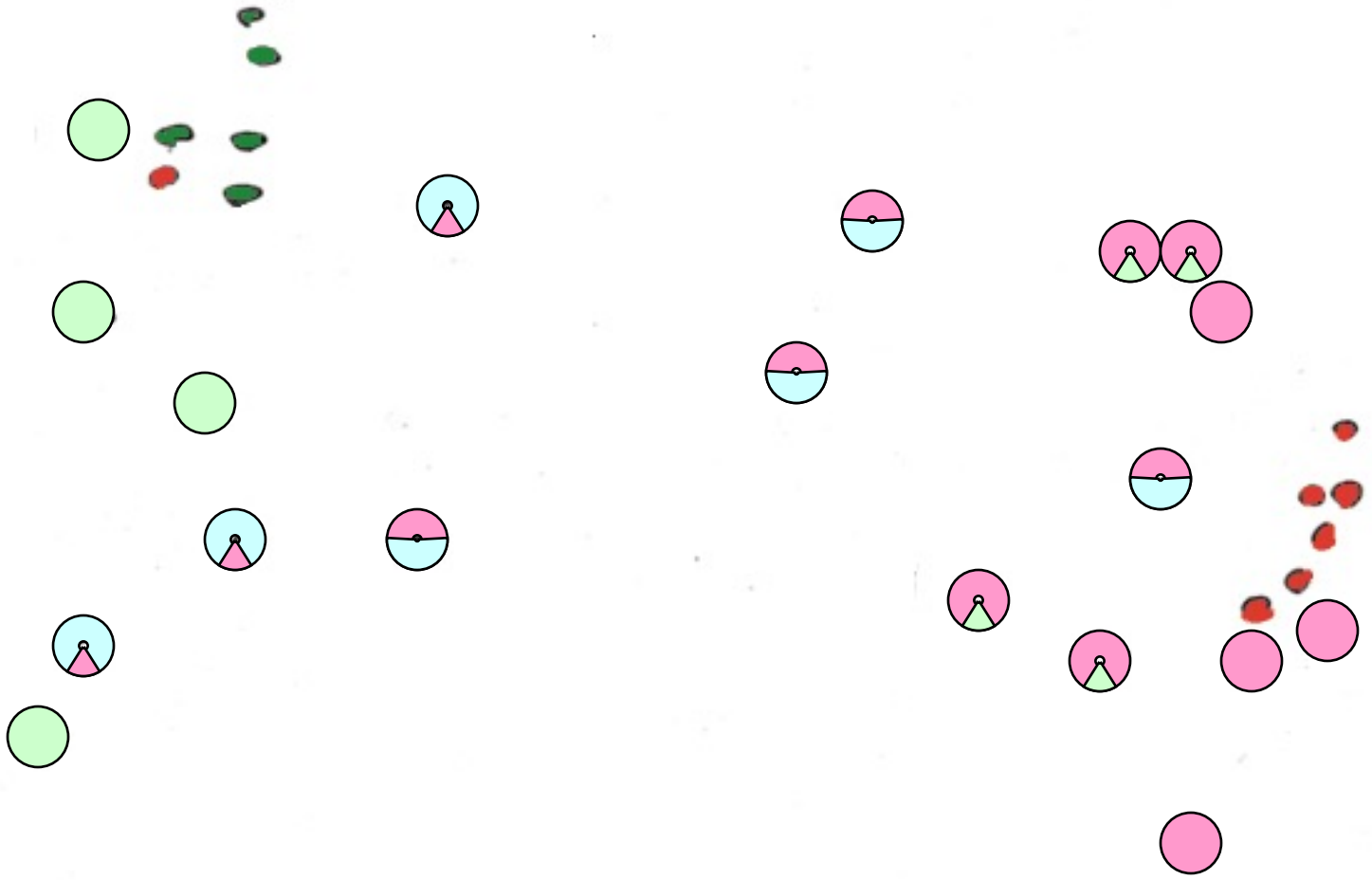
Find a Division to Separate Tagged Points



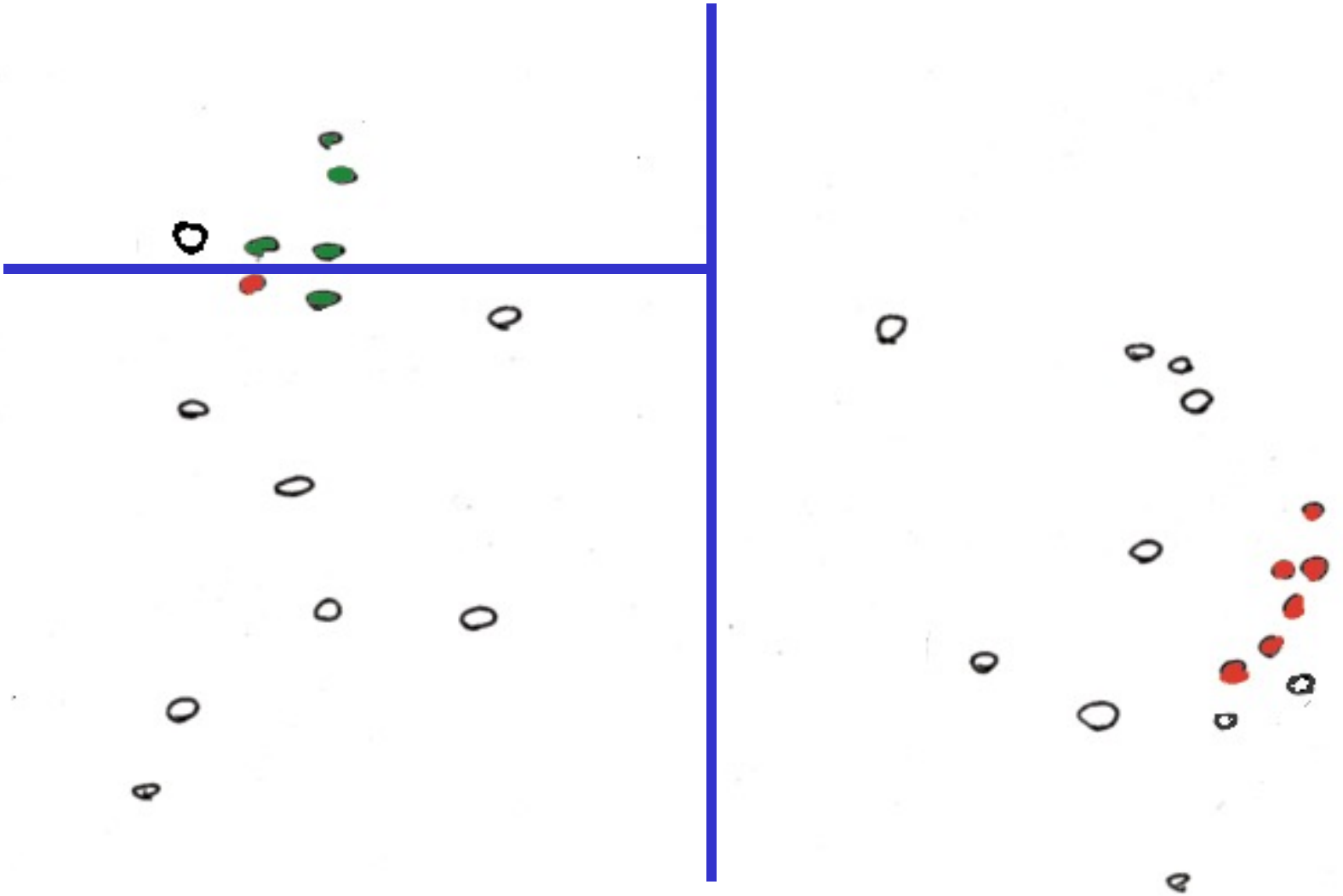
Extrapolate to Untagged Points



Probabilistic Predictions of Class



Find a Division to Separate Tagged Points



Distinctions in Supervised Learning

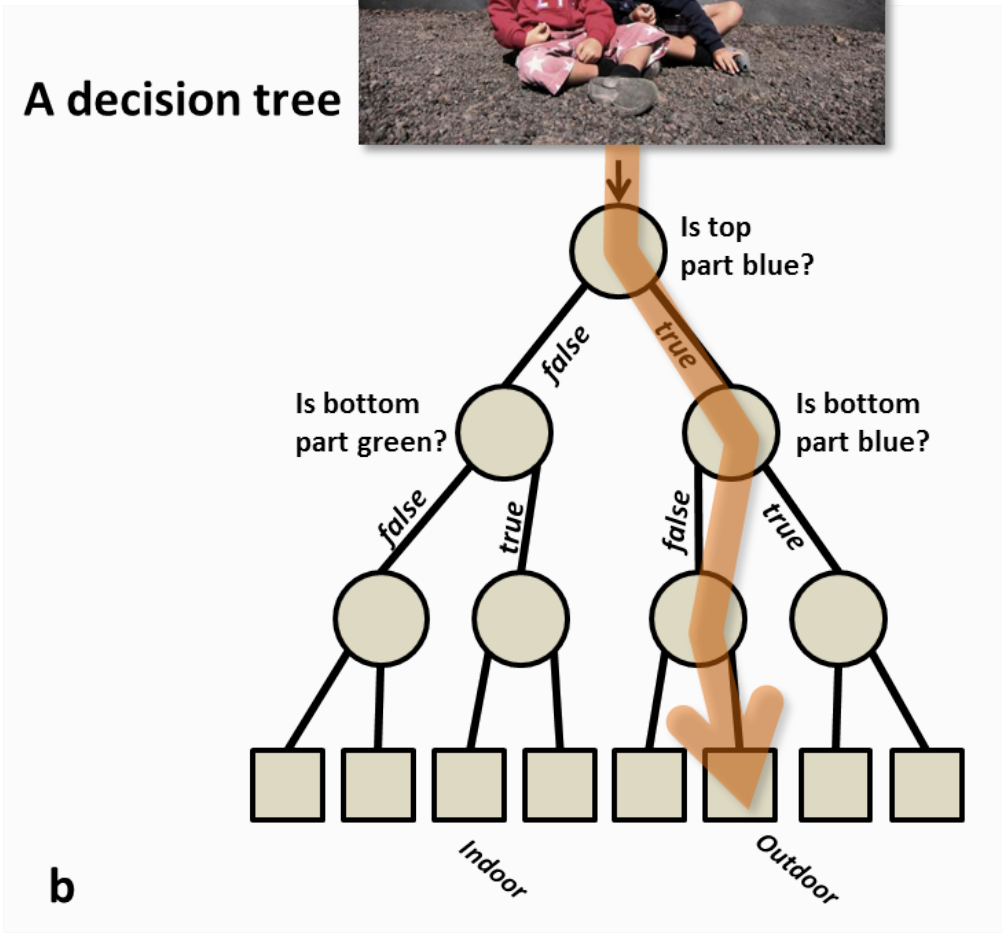
- Regression vs Classification
  - Regression: labels are quantitative
  - Classification: labels are categorical
- Regularized vs Un-regularized
  - Regularized: penalize model complexity to avoid over-fitting
  - Un-regularized: no penalty on model complexity
- Parametric vs Non-parametric
  - Parametric: an explicit parametric model is assumed
  - Non-parametric: otherwise
- Ensemble vs Non-ensemble
  - Ensemble: combines multiple models
  - Non-ensemble: a single model

Supervised Mining:

Decision Trees

Decision Trees

- **Classify data by asking questions** that divide data in subgroups
- Keep asking questions until subgroups become homogenous
- Use **tree** of questions to make predictions

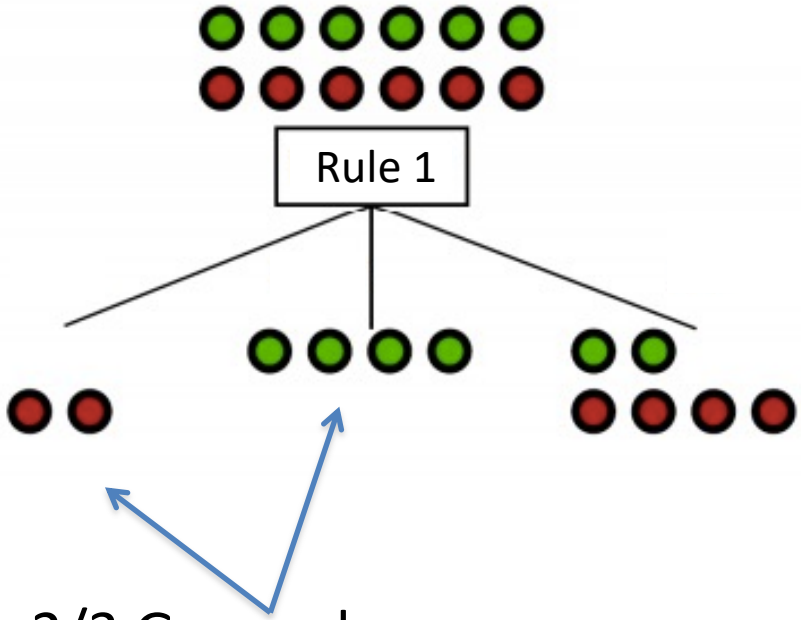


- Example: Is a picture taken inside or outside?

Criminisi, Shotton, and Konukoglu *Microsoft Technical Report* 2011

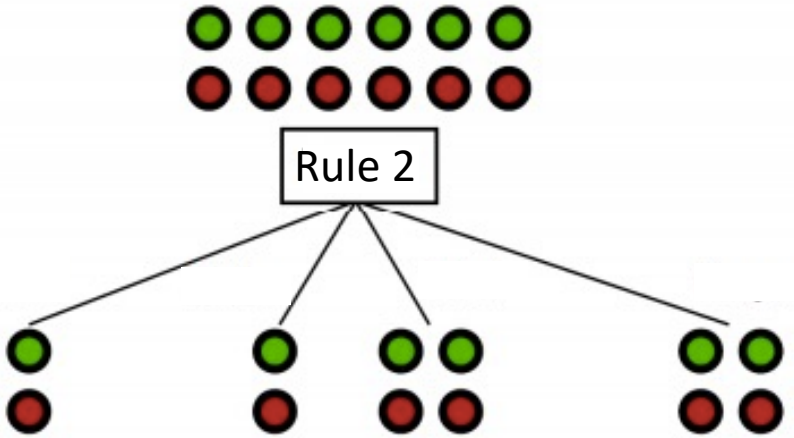
What makes a good rule?

- Want resulting groups to be as homogenous as possible



2/3 Groups homogenous

→Good rule



All groups still 50/50

→ Unhelpful rule

Quantifying the value of rules

- Decrease in inhomogeneity
  - Most popular metric: Information theoretic entropy

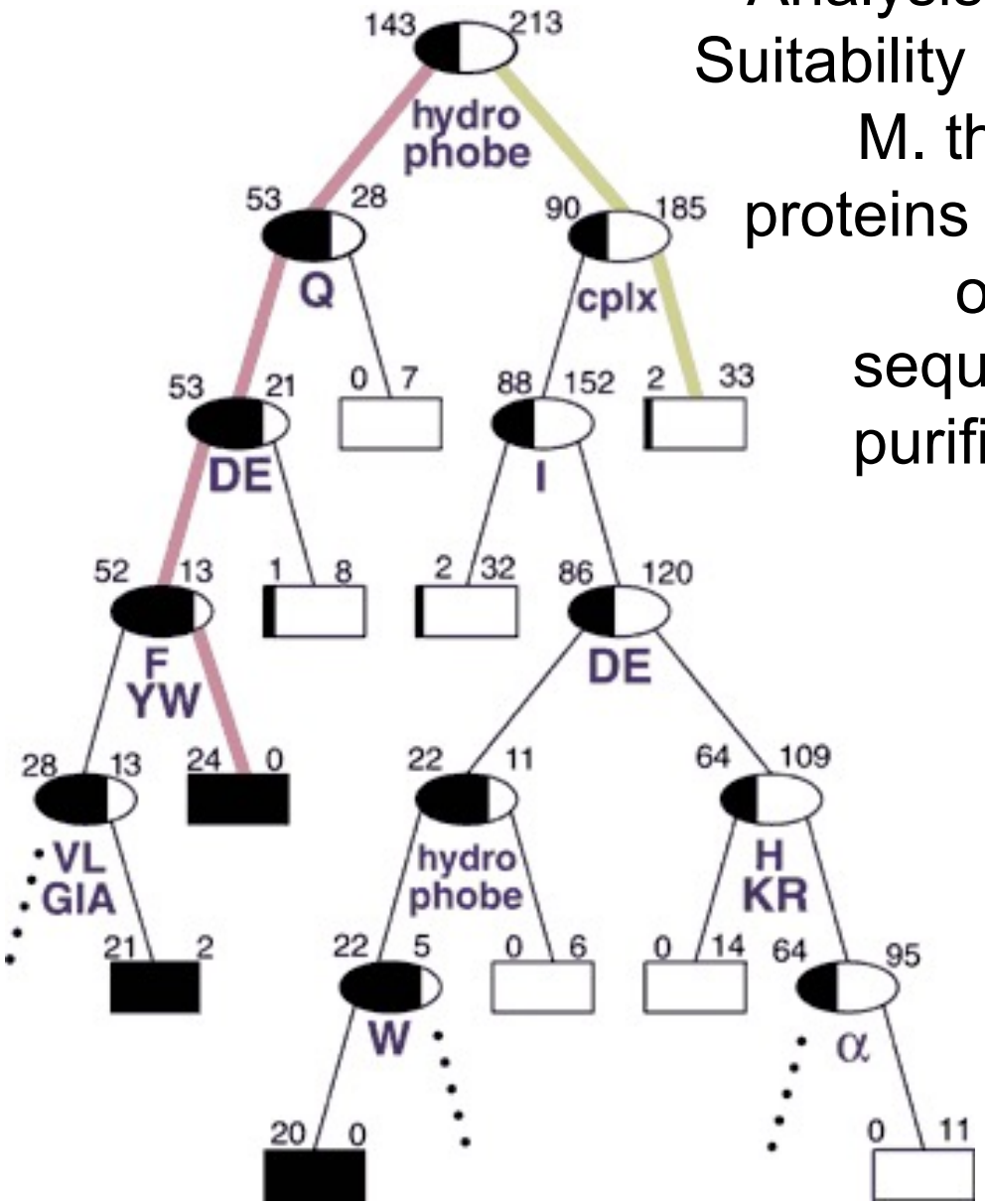
$$S = -\sum_{i=1}^{m} p_i \log p_i$$

  - Use frequency of classifier characteristic within group as probability
  - Minimize entropy to achieve homogenous group

Algorithm

- For each characteristic:
  - Split into subgroups based on each possible value of characteristic
- Choose rule from characteristic that maximizes decrease in inhomogeneity
- For each subgroup:
  - if (inhomogeneity < threshold):
    - Stop
  - else:
    - Restart rule search (recursion)

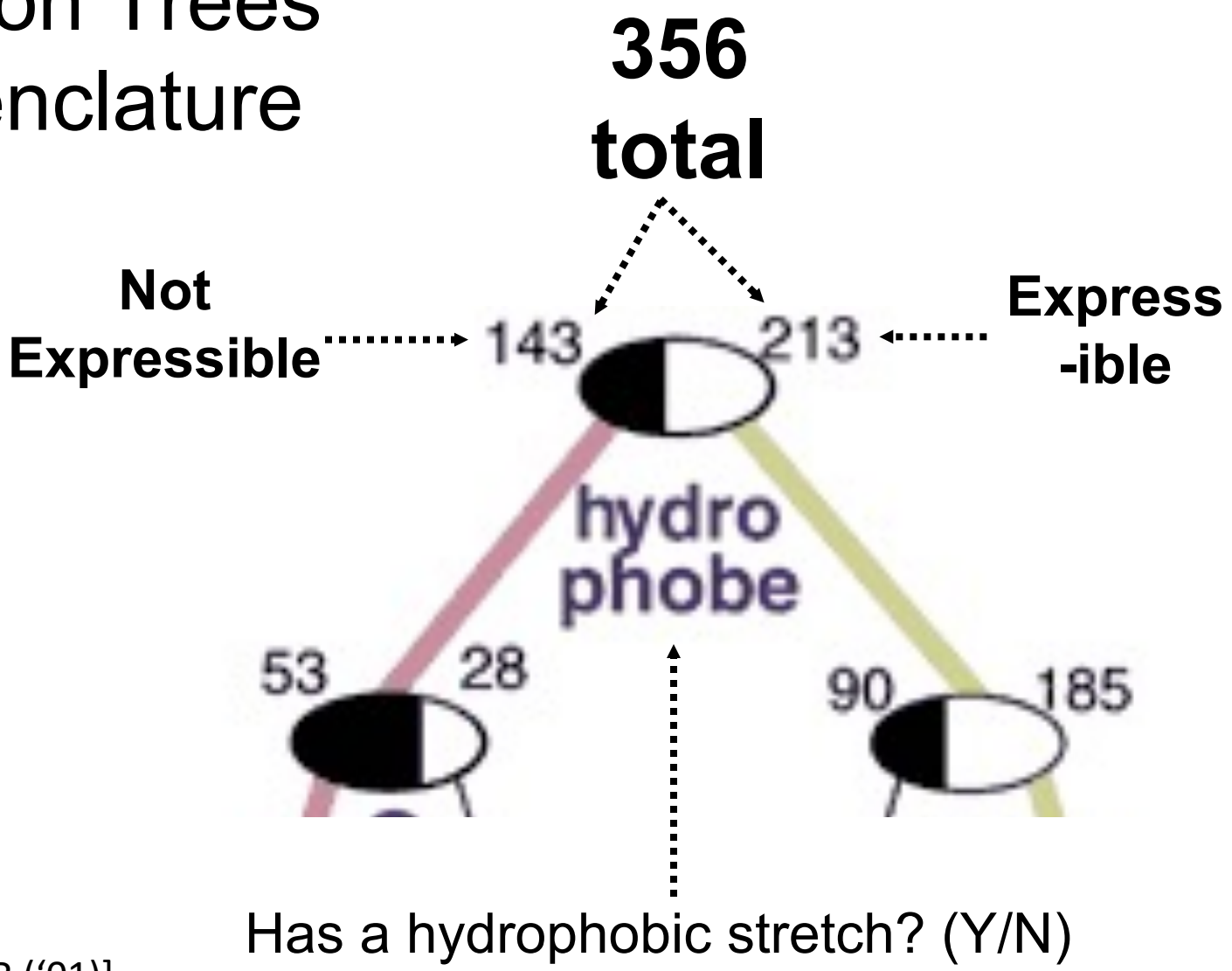
Retrospective Decision Trees



Analysis of the Suitability of 500 M. thermo. proteins to find optimal sequences purification

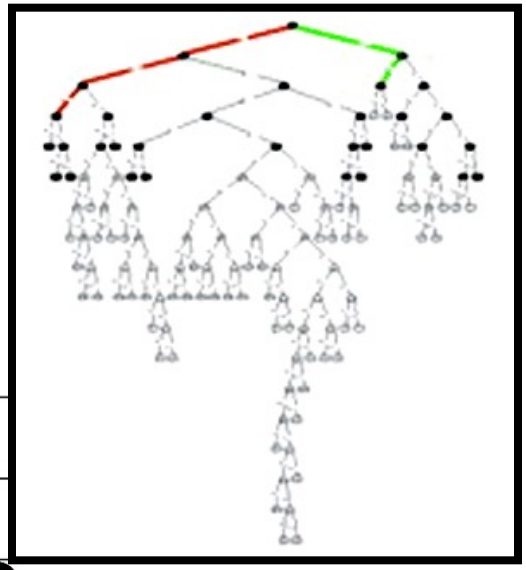
[Bertone et al. NAR (‘01)]

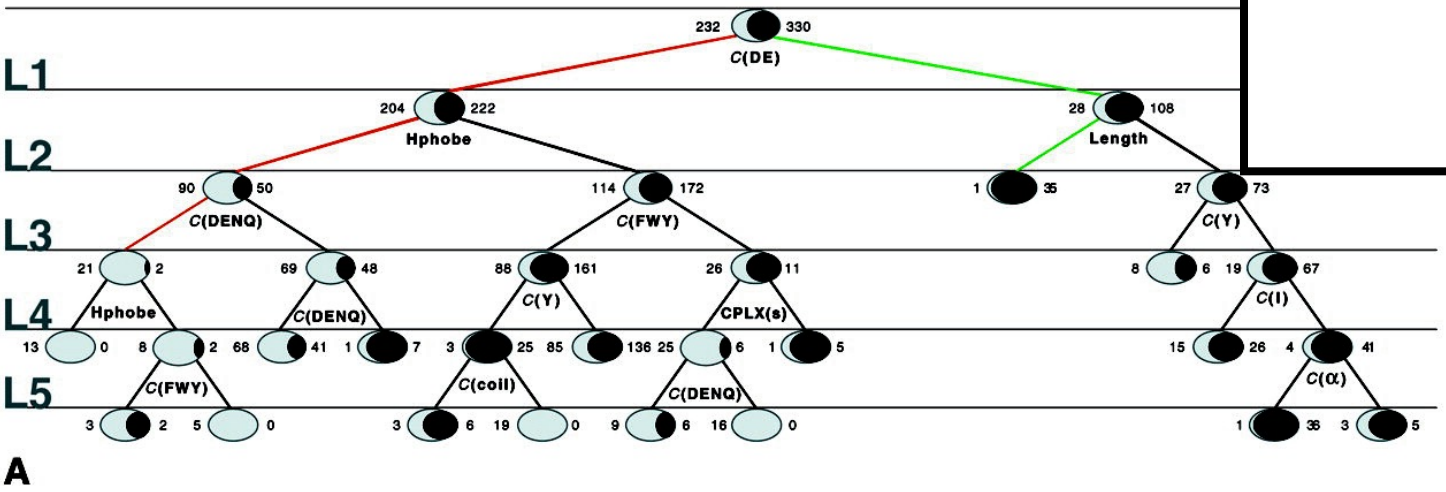
Retrospective Decision Trees Nomenclature



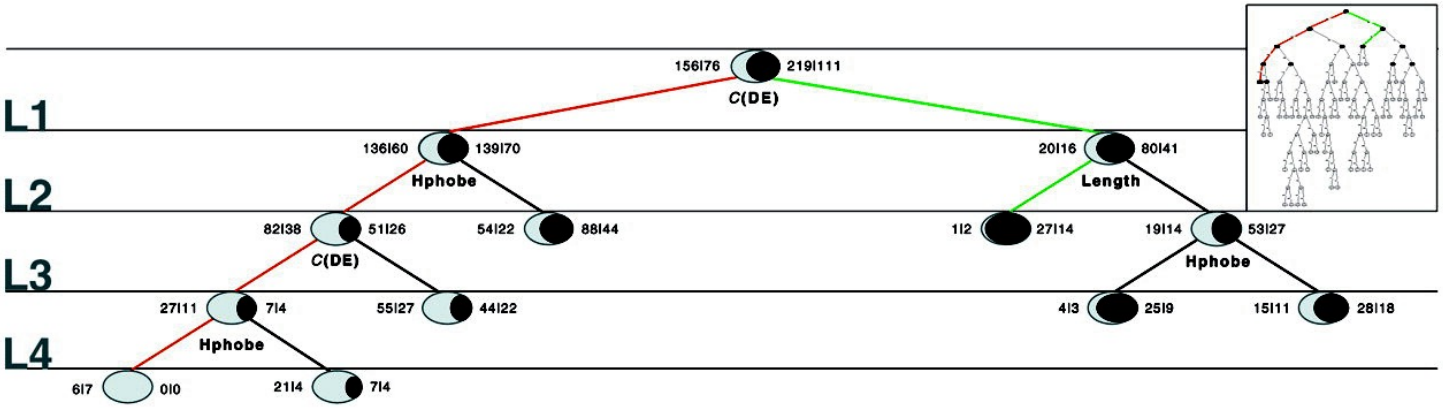
[Bertone et al. NAR (‘01)]

Overfitting, Cross Validation, and Pruning





A



Extensions of Decision Trees

- Decision Trees method is very sensitive to noise in data
- Random forests is an ensemble of decision trees and is much more effective.