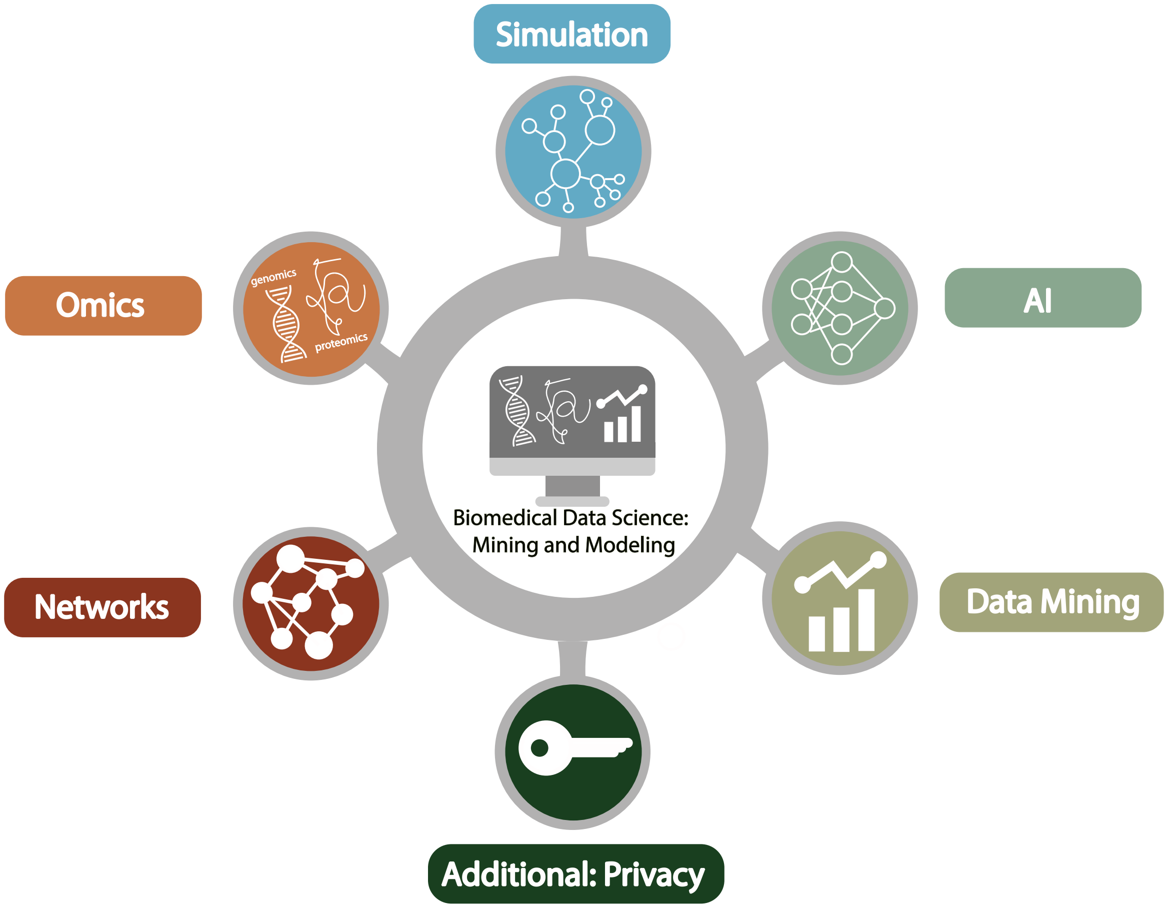
Biomedical Data Science (GersteinLab.org/courses/452)

Supervised Datamining – Decision Trees (23m8a)



Simulation

Omics

AI

Biomedical Data Science: Mining and Modeling

Networks

Data Mining

Additional: Privacy

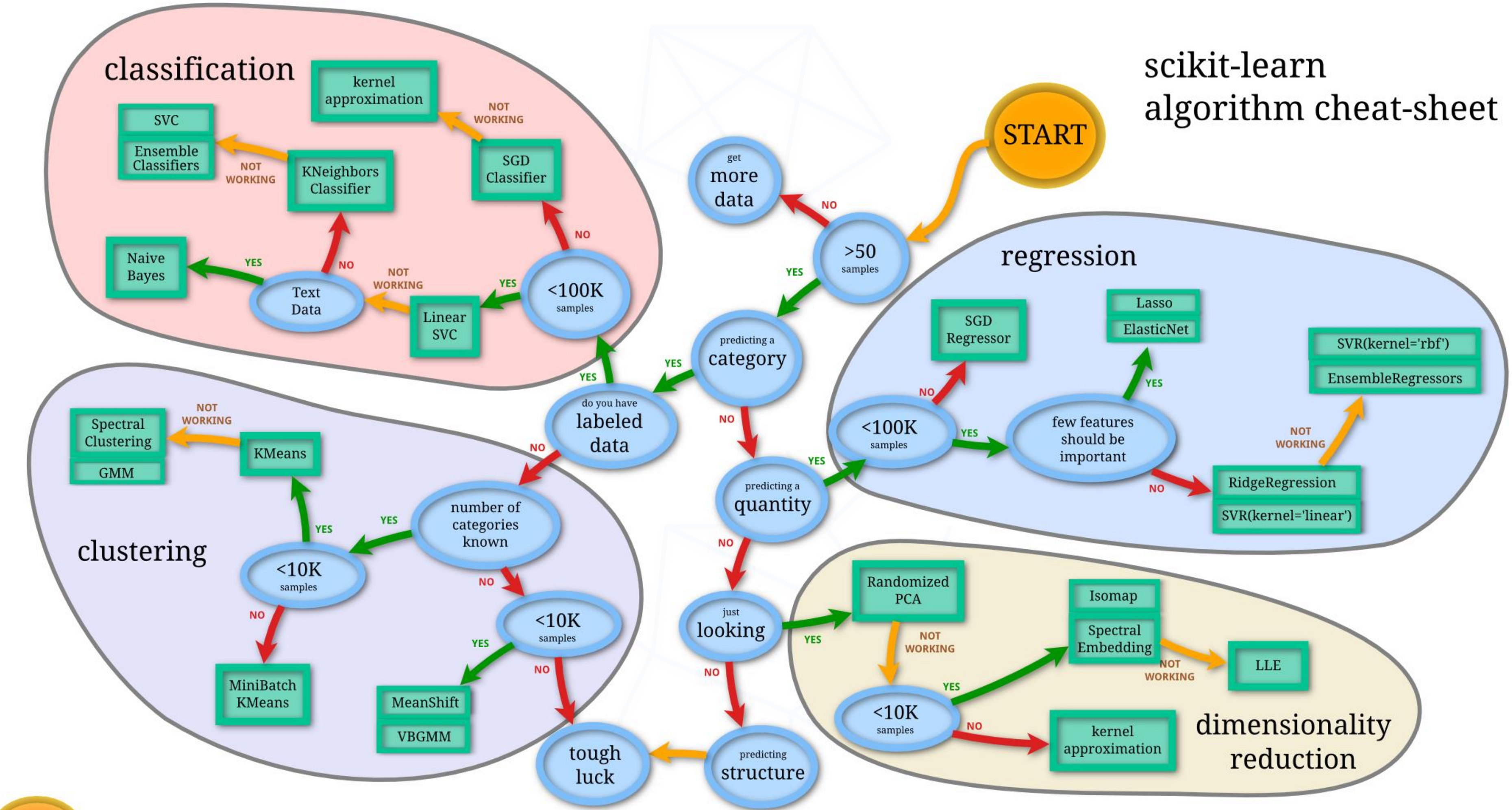
Mark Gerstein
Yale U.

Last edit in spring ’23. Condensed (with ~2 slide deletions) from 22m8a, which is similar to 2021’s M8a [which has a video].

Supervised Mining:

Overview

The World of Machine Learning



scikit-learn
algorithm cheat-sheet

START

>50 samples

get more data

predicting a category

do you have labeled data

predicting a quantity

just looking

predicting structure

tough luck

classification

<100K samples

SGD Classifier

kernel approximation

Linear SVC

Text Data

Naive Bayes

KNeighbors Classifier

SVC

Ensemble Classifiers

clustering

number of categories known

<10K samples

KMeans

Spectral Clustering

GMM

MiniBatch KMeans

<10K samples

MeanShift

VBGMM

regression

<100K samples

SGD Regressor

few features should be important

Lasso

ElasticNet

RidgeRegression

SVR(kernel='linear')

SVR(kernel='rbf')

EnsembleRegressors

dimensionality reduction

Randomized PCA

<10K samples

Isomap

Spectral Embedding

LLE

kernel approximation

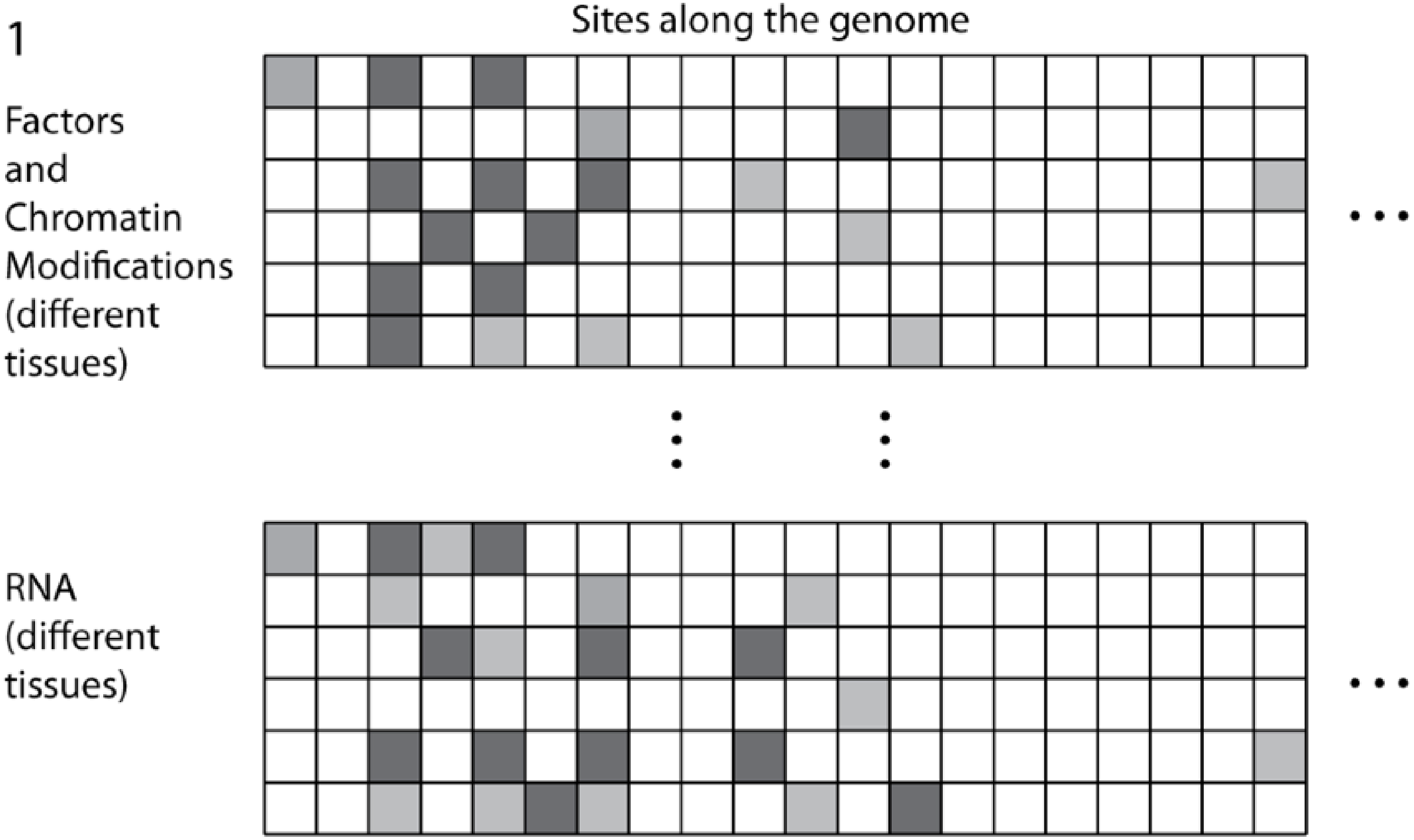
Back

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

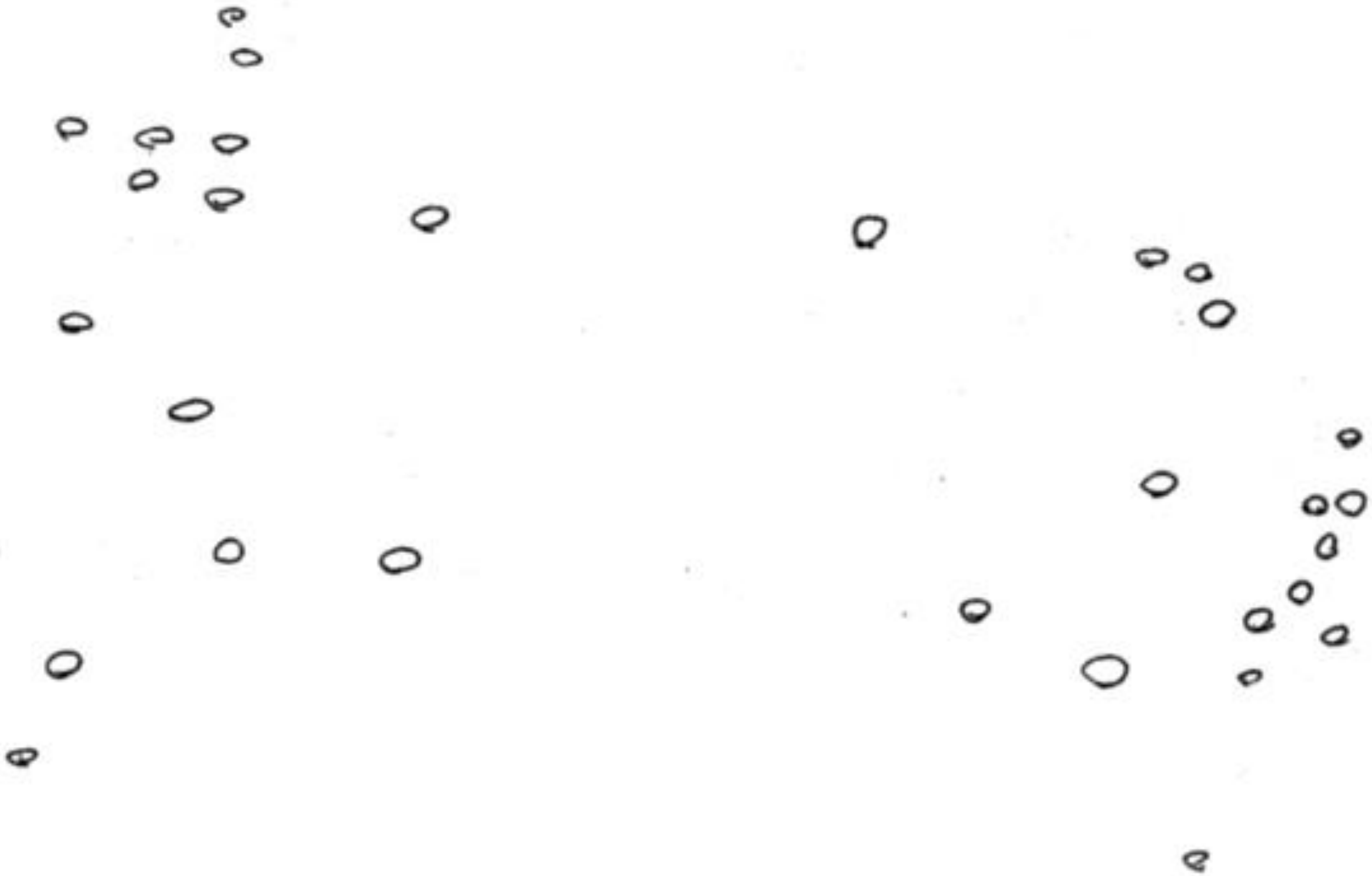
Distinctions in Supervised Learning

- 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

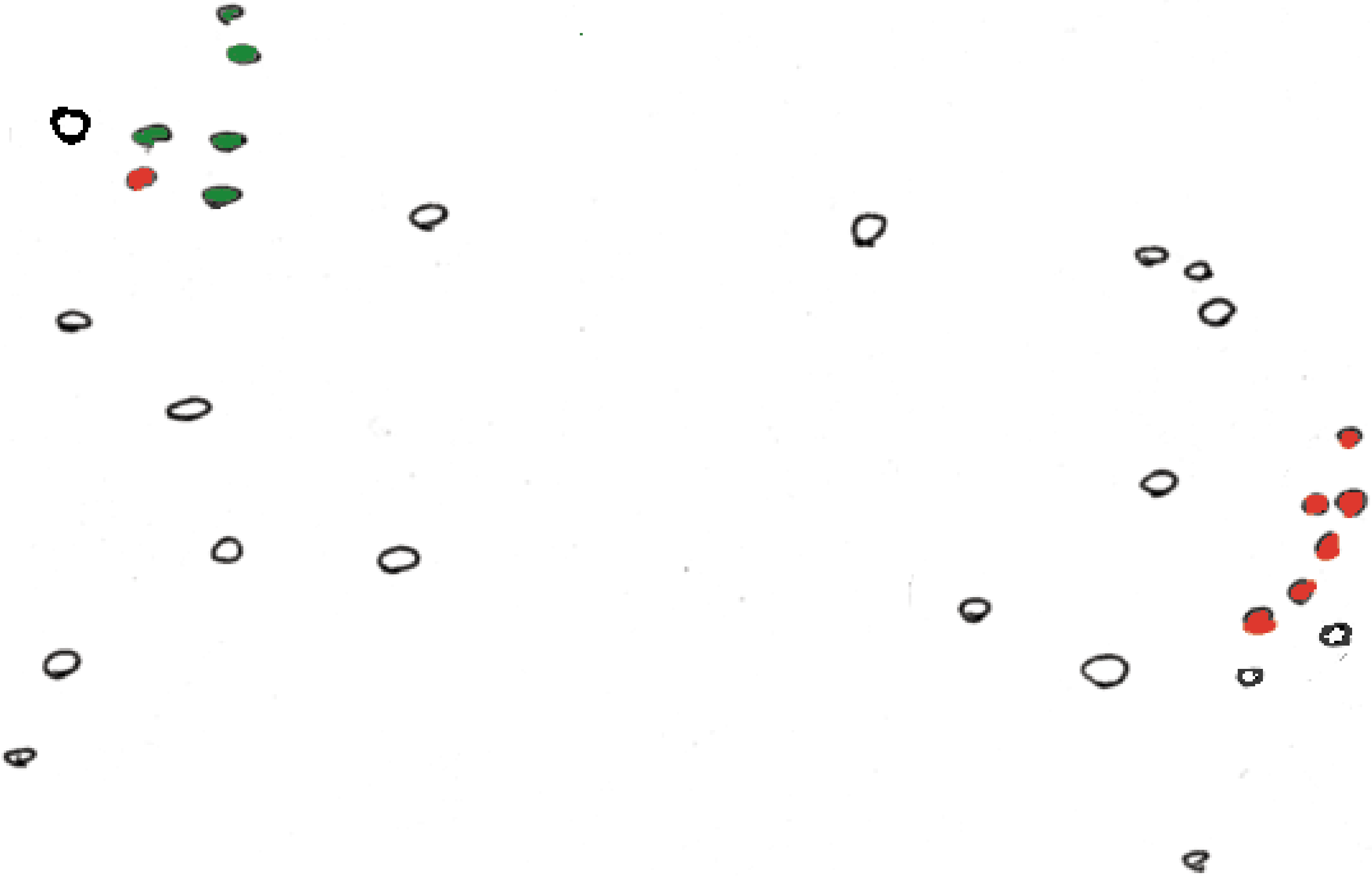
Structure of Genomic Features Matrix



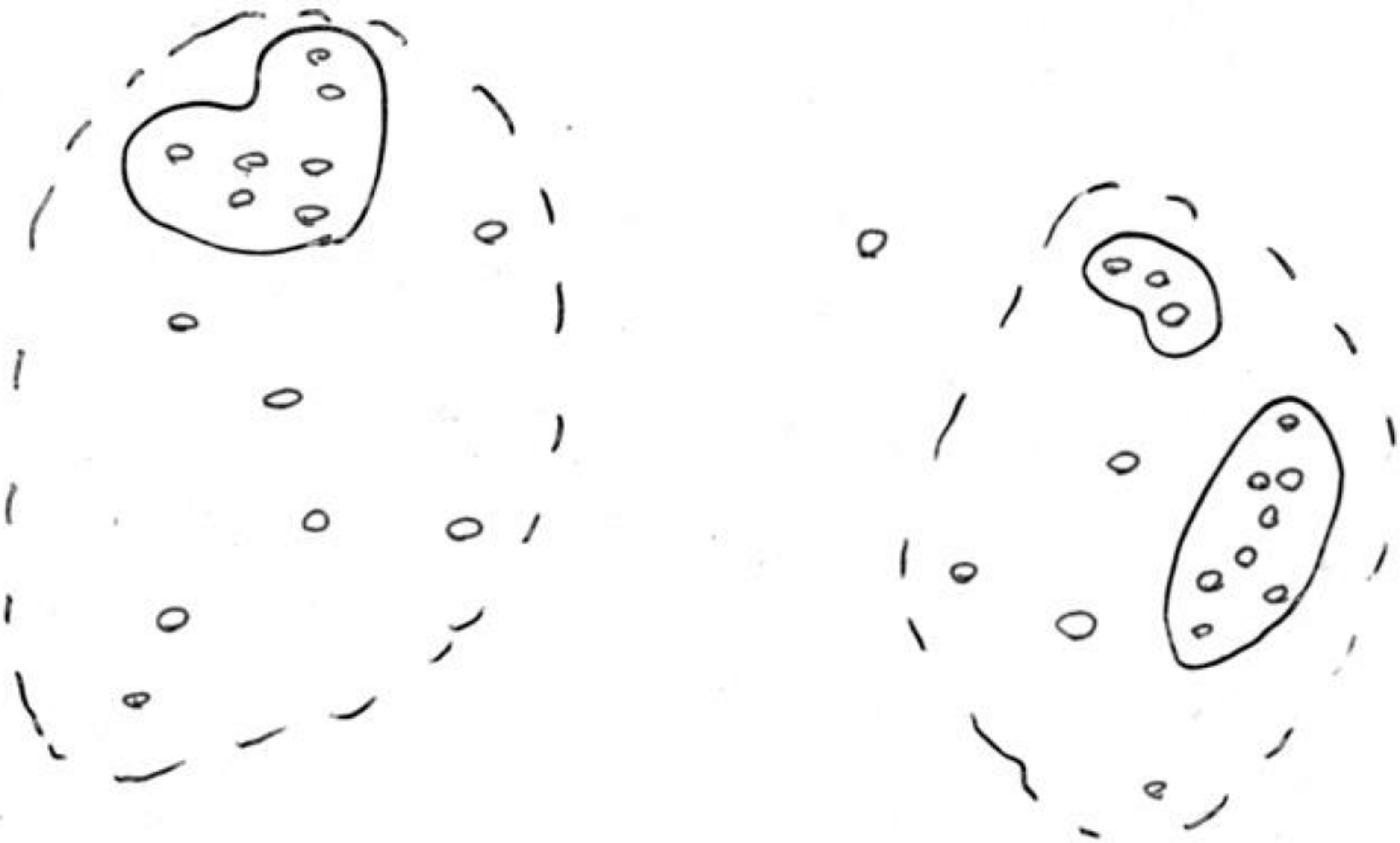
Represent predictors in abstract high dimensional space



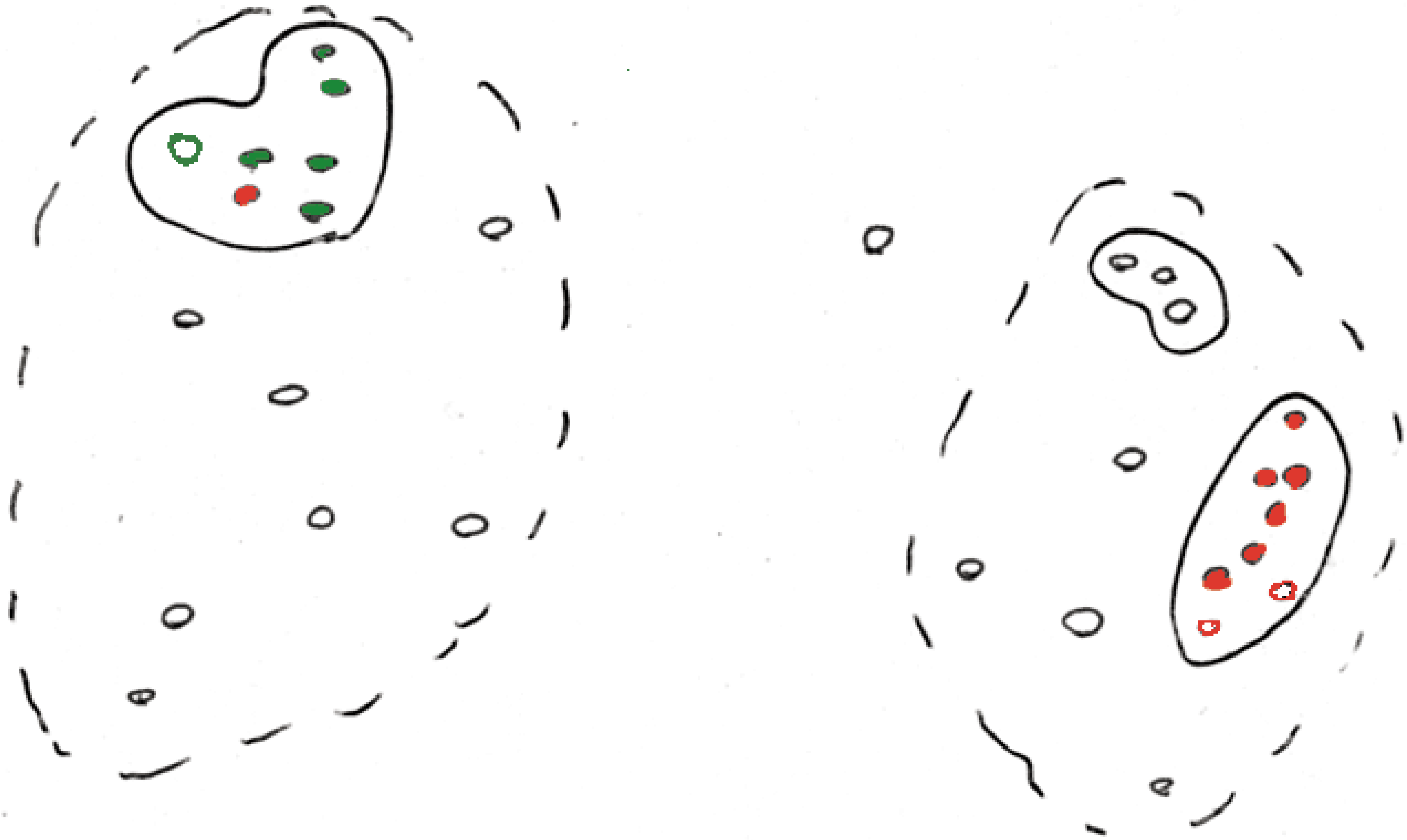
“Label” Certain Points



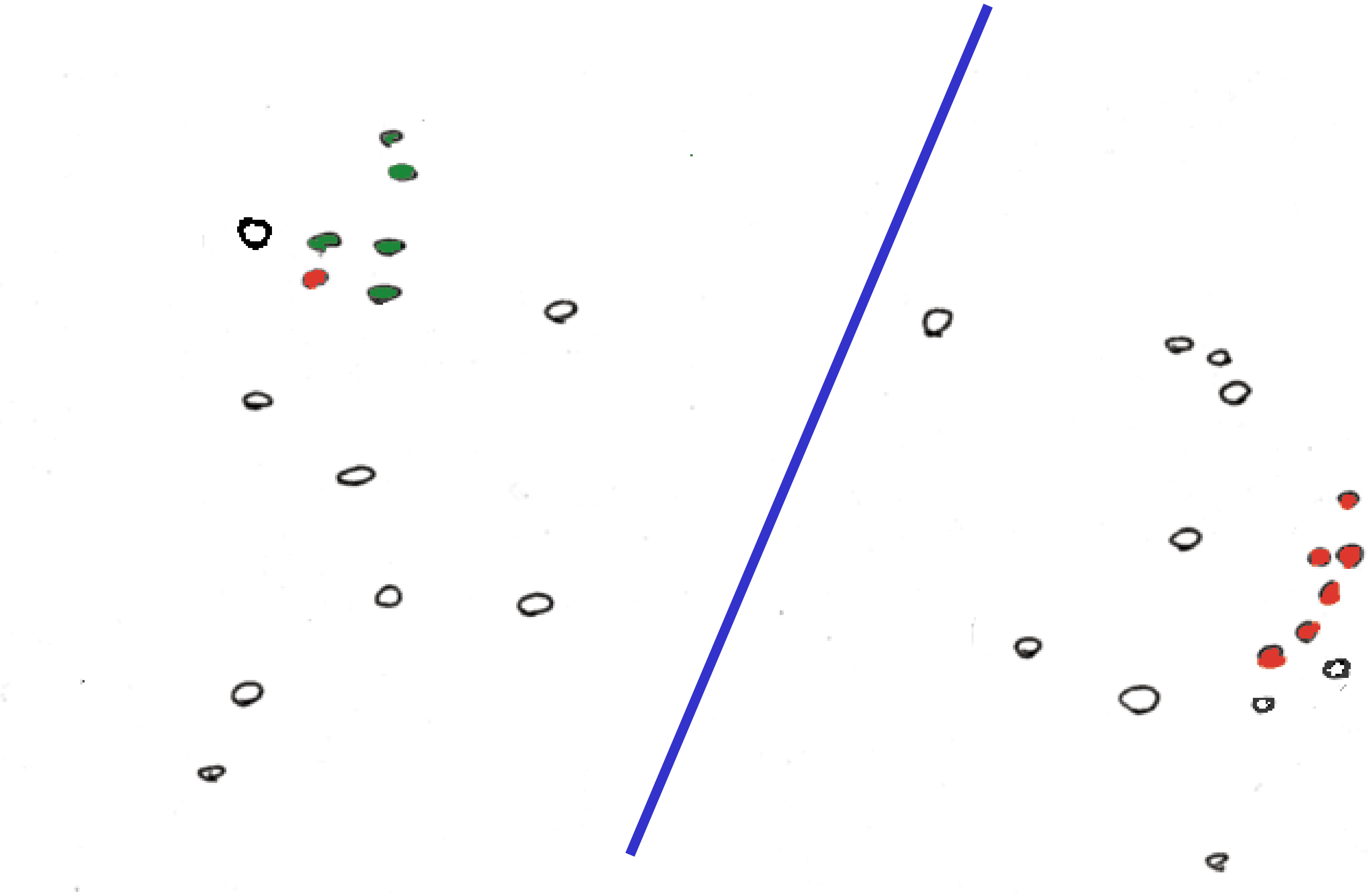
“Cluster” predictors (Unsupervised)



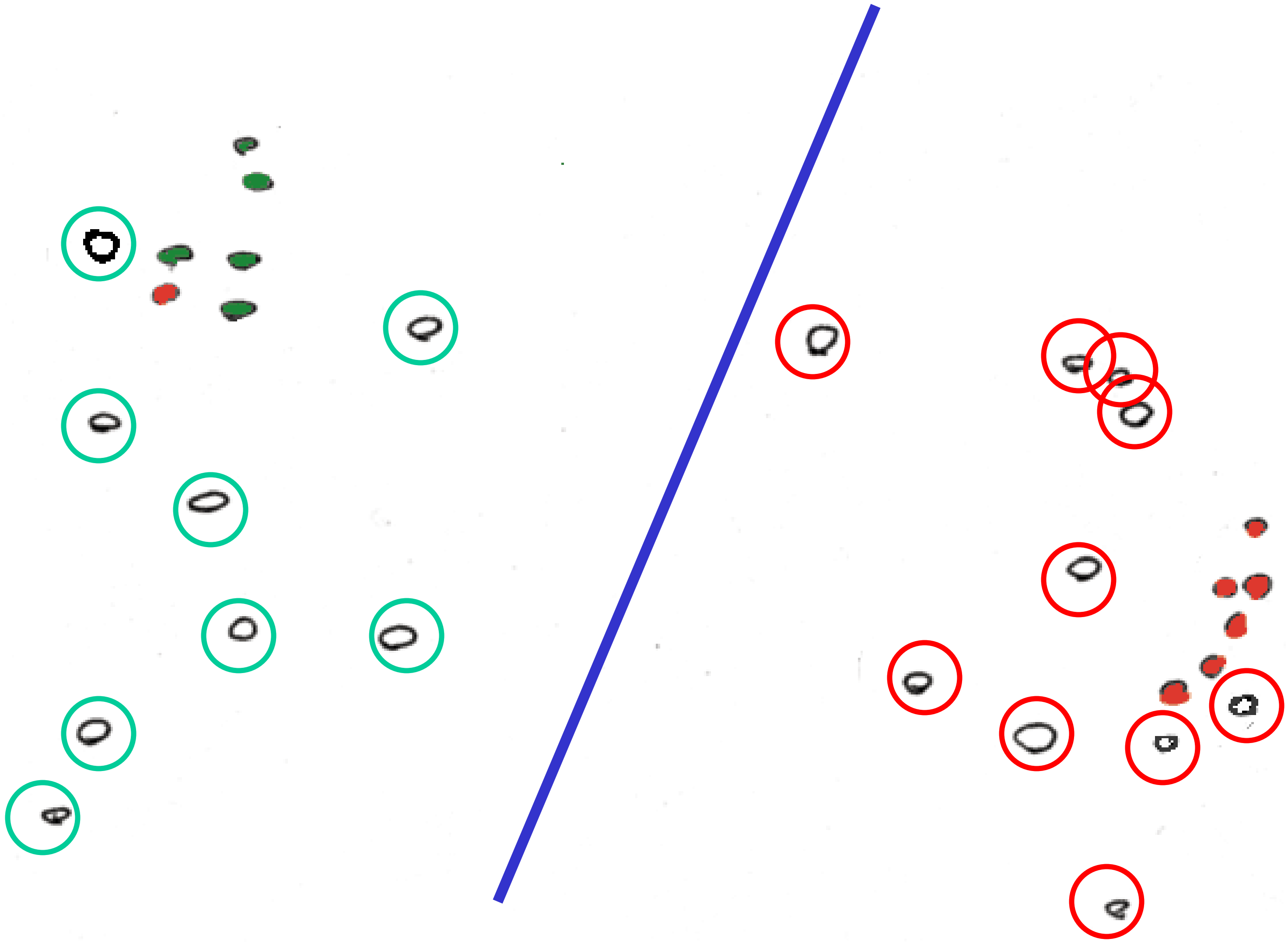
Use Clusters to predict Response (Unsupervised, guilt-by-association)



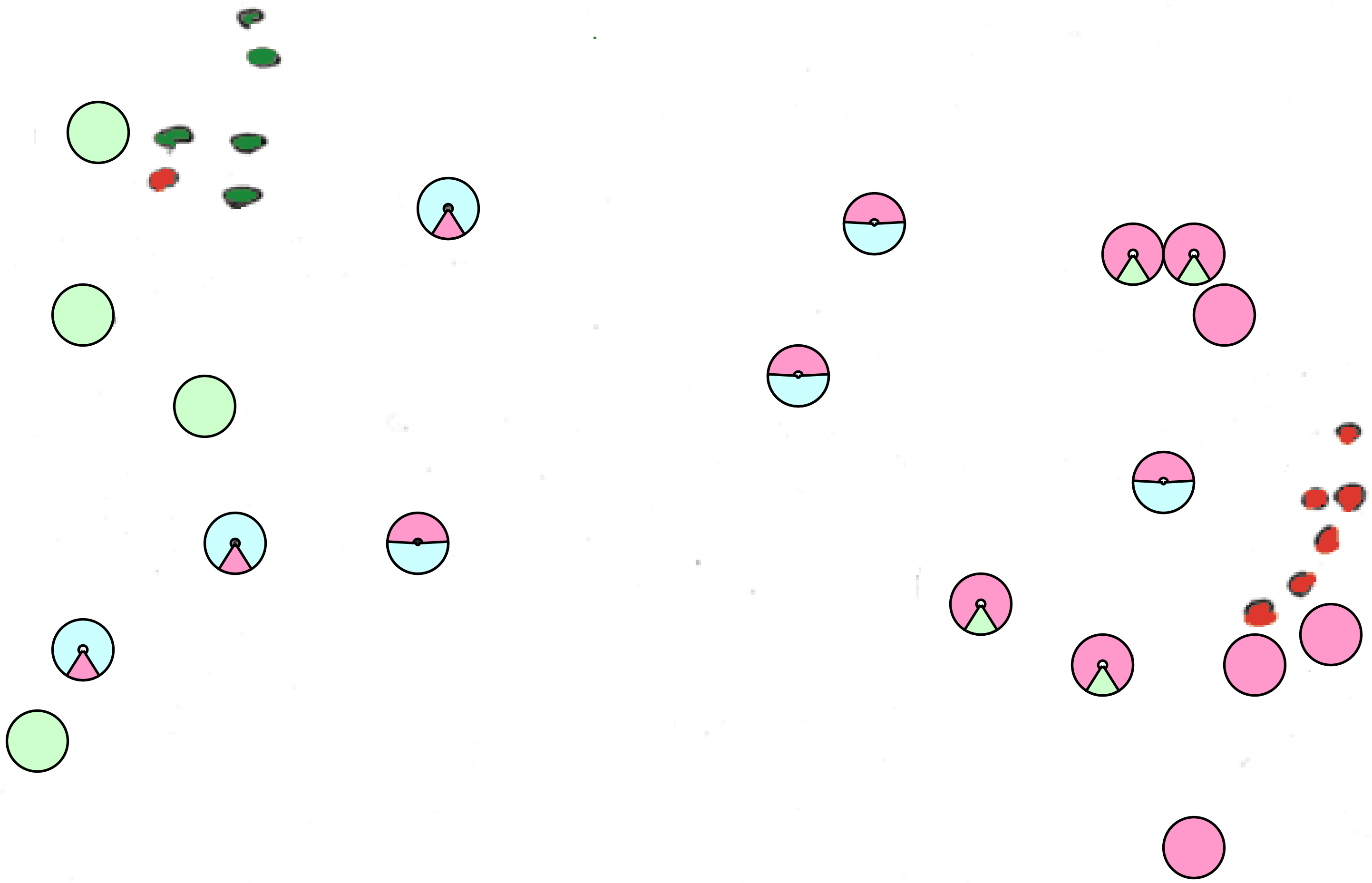
Find a Division to Separate Tagged Points



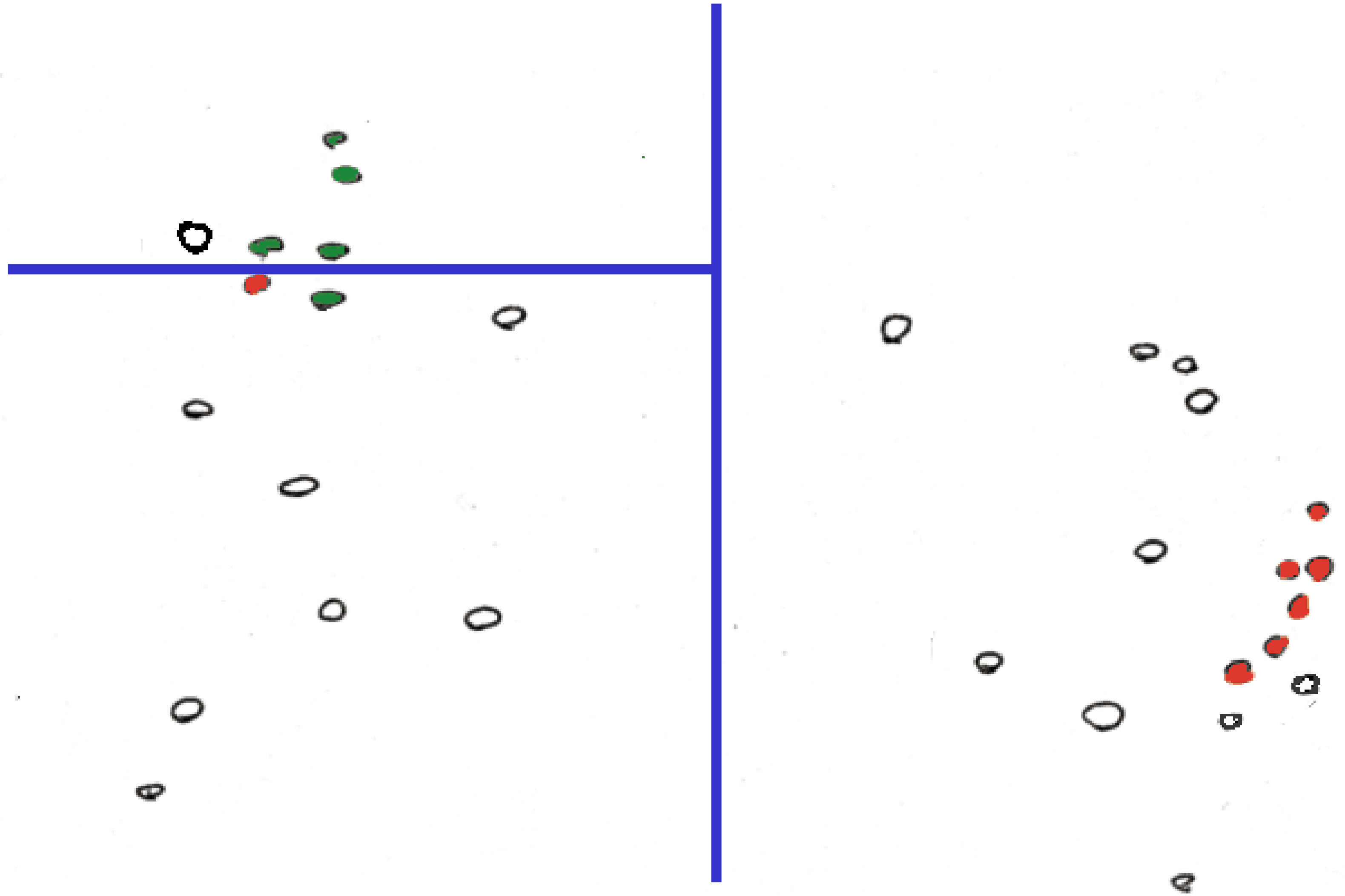
Extrapolate to Untagged Points



Probabilistic Predictions of Class



Find a Division to Separate Tagged Points

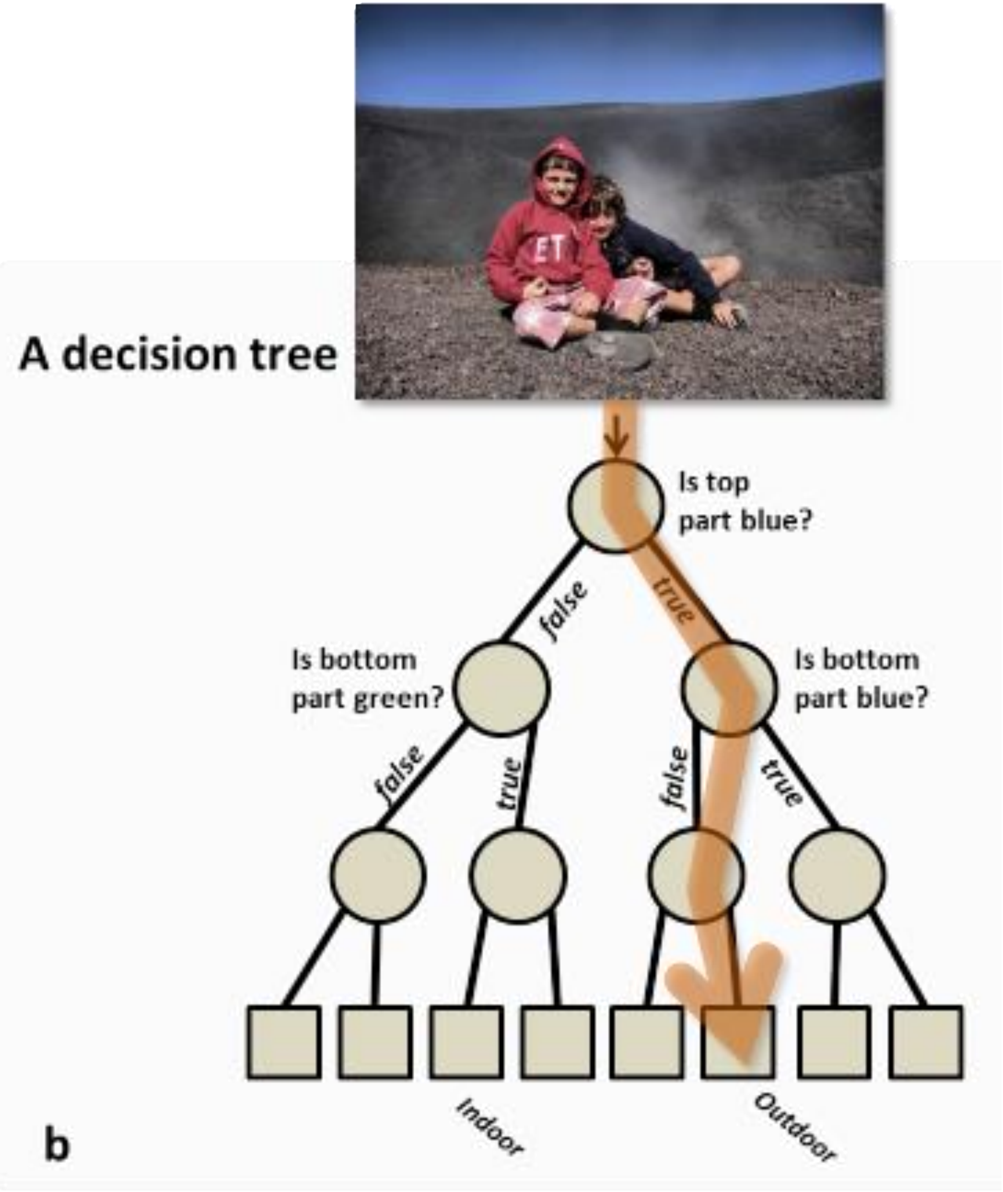


Supervised Mining:

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

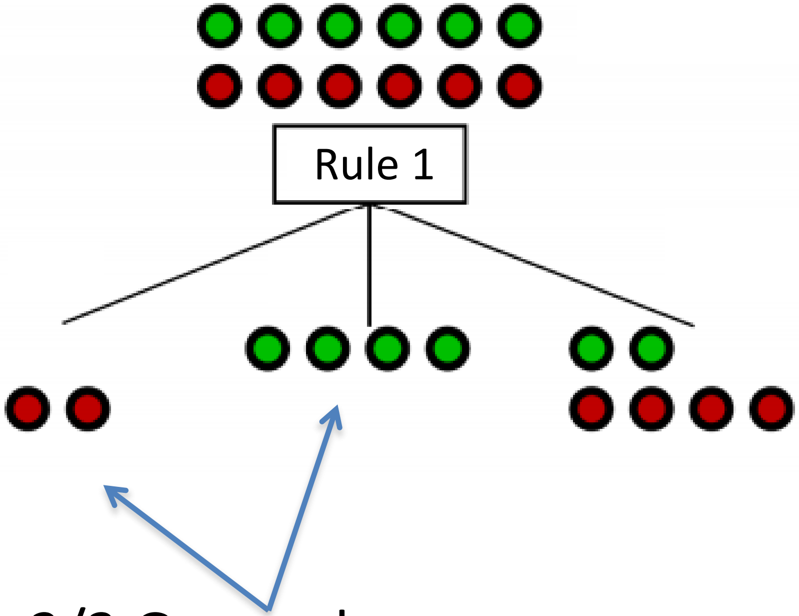


- Example: Is a picture taken inside or outside?

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

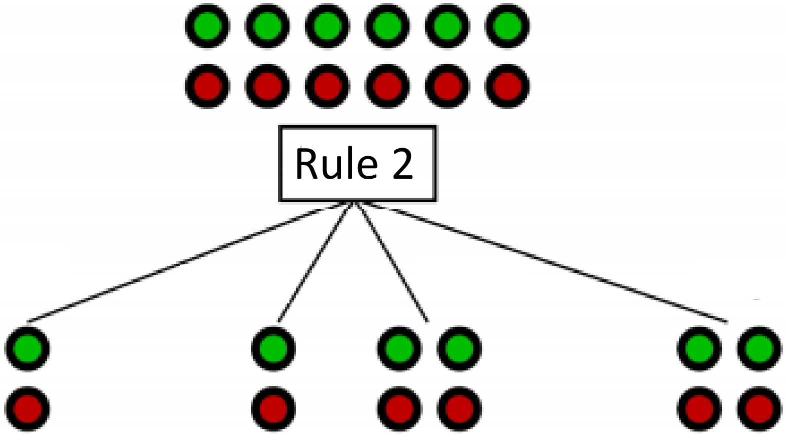
What makes a good rule?

- Want resulting groups to be as homogenous as possible



2/3 Groups homogenous

→Good rule



All groups still 50/50

→ Unhelpful rule

Nando de Freitas 2012 University of British Columbia CPSC 340

Quantifying the value of rules

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

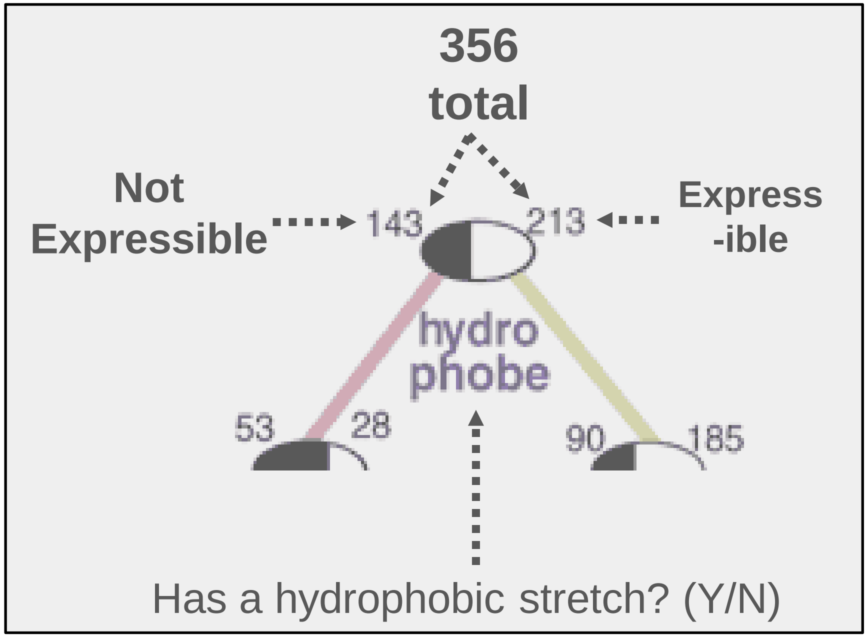
$$S = -\sum_{i=1}^{m} p_i \log p_i$$

  - 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

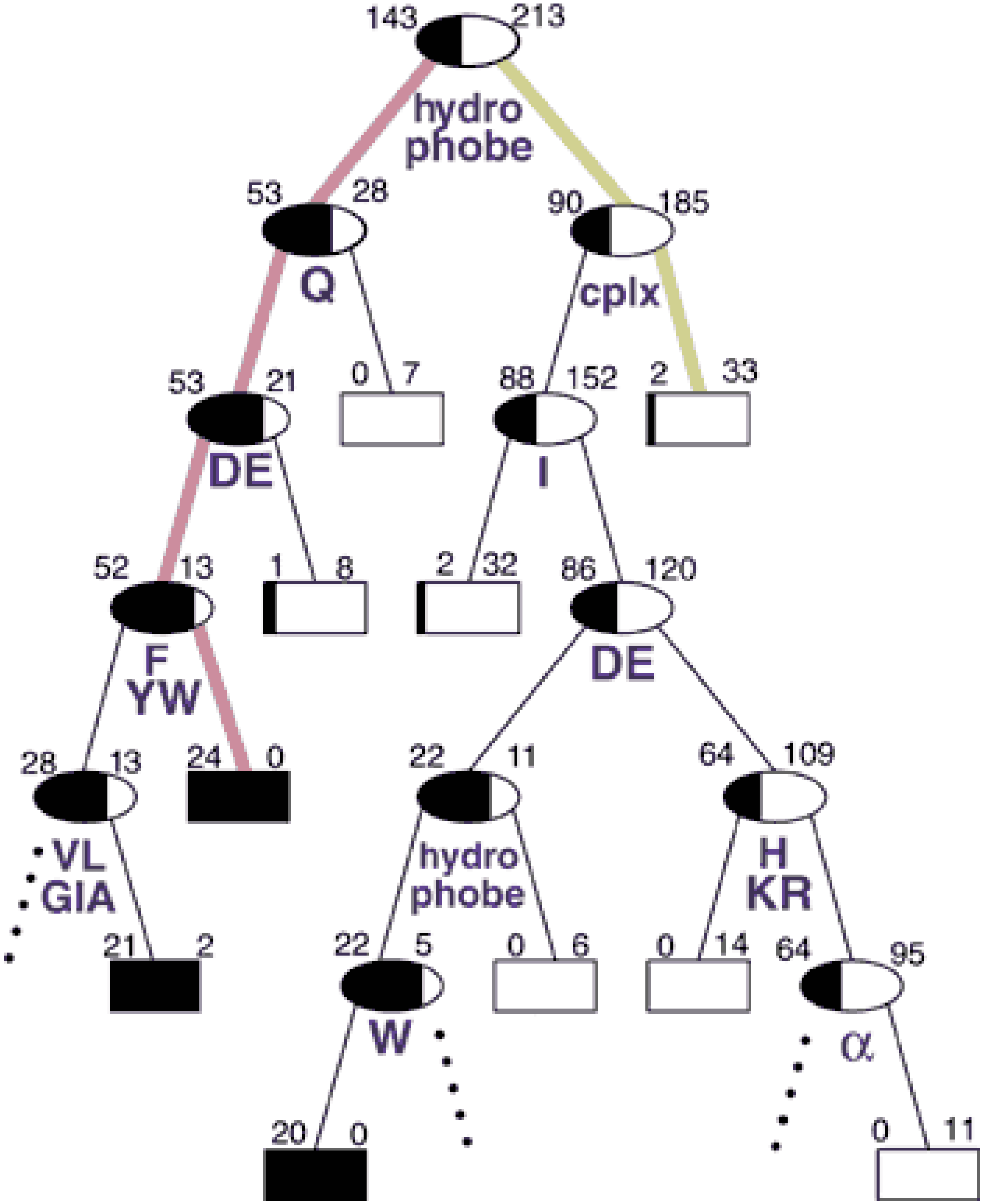


356 total

Not Expressible

Express-ible

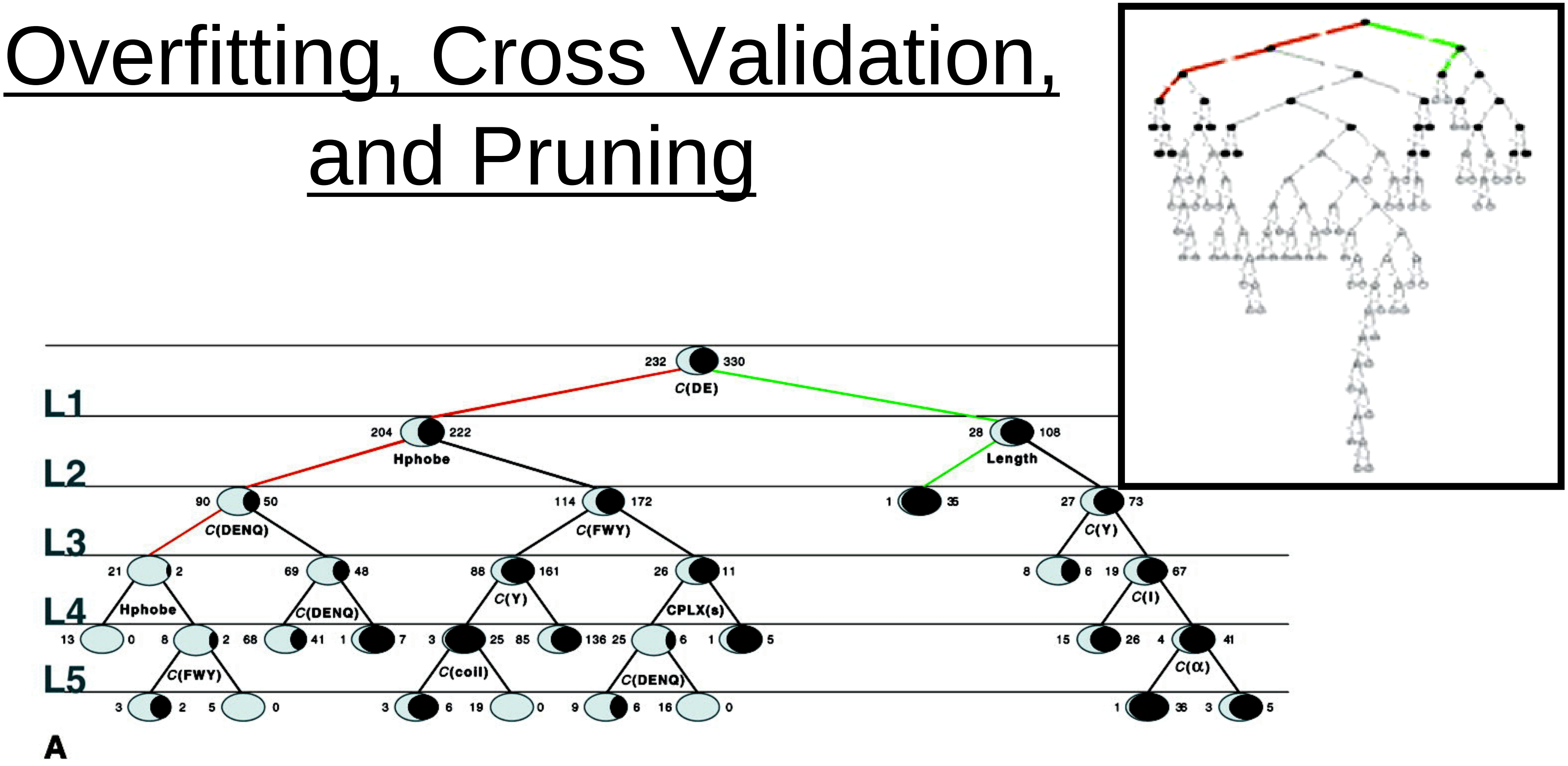
Has a hydrophobic stretch? (Y/N)

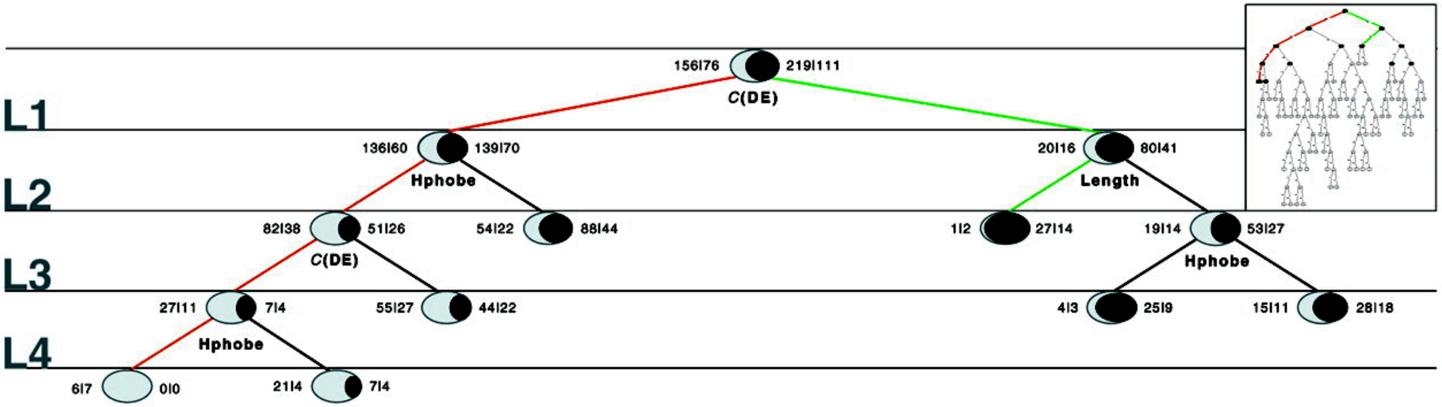


Analysis of the Suitability of 500 M. thermo. proteins to find optimal sequences purification

[Bertone et al. NAR ('01)]

Overfitting, Cross Validation, and Pruning





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.