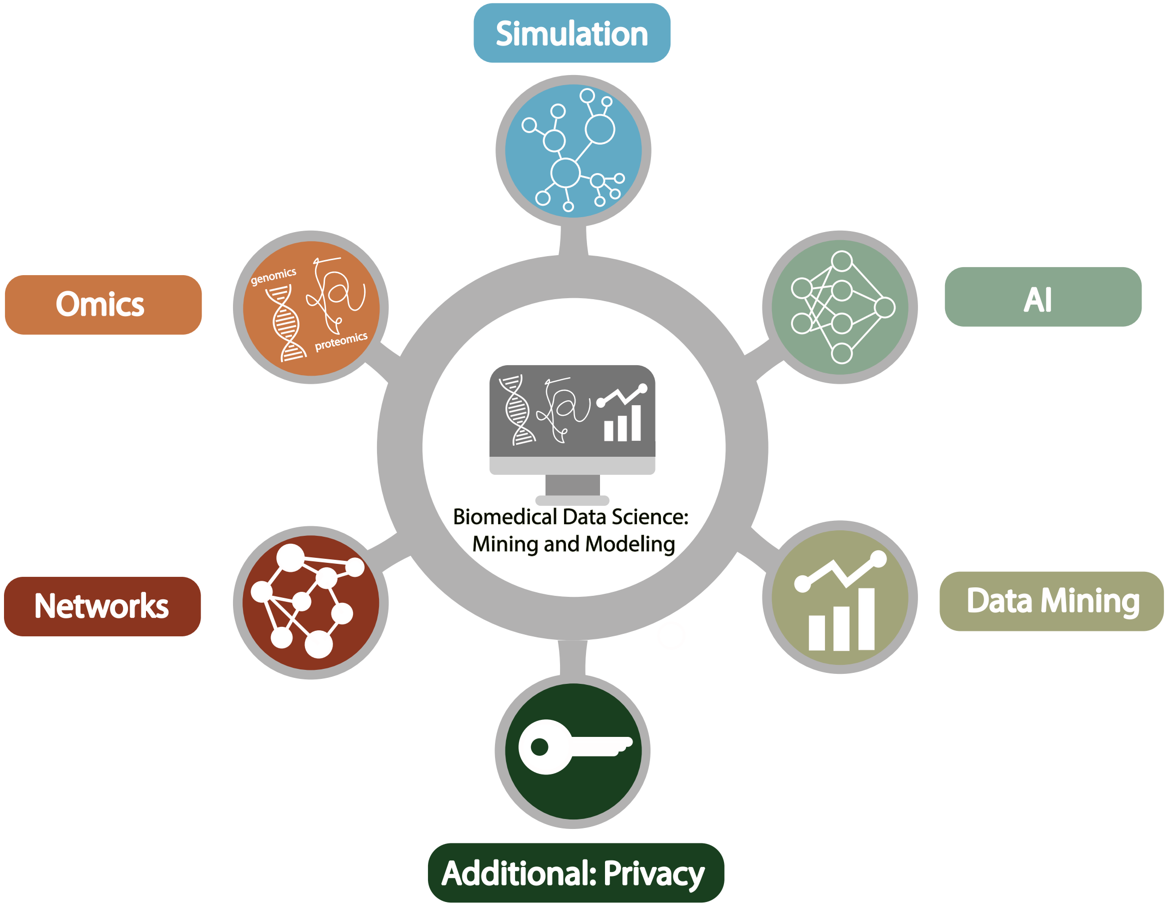
Biomedical Data Science (GersteinLab.org/courses/452)

Supervised Mining: Preliminaries + Decision Trees

(25m8a+25m8b)



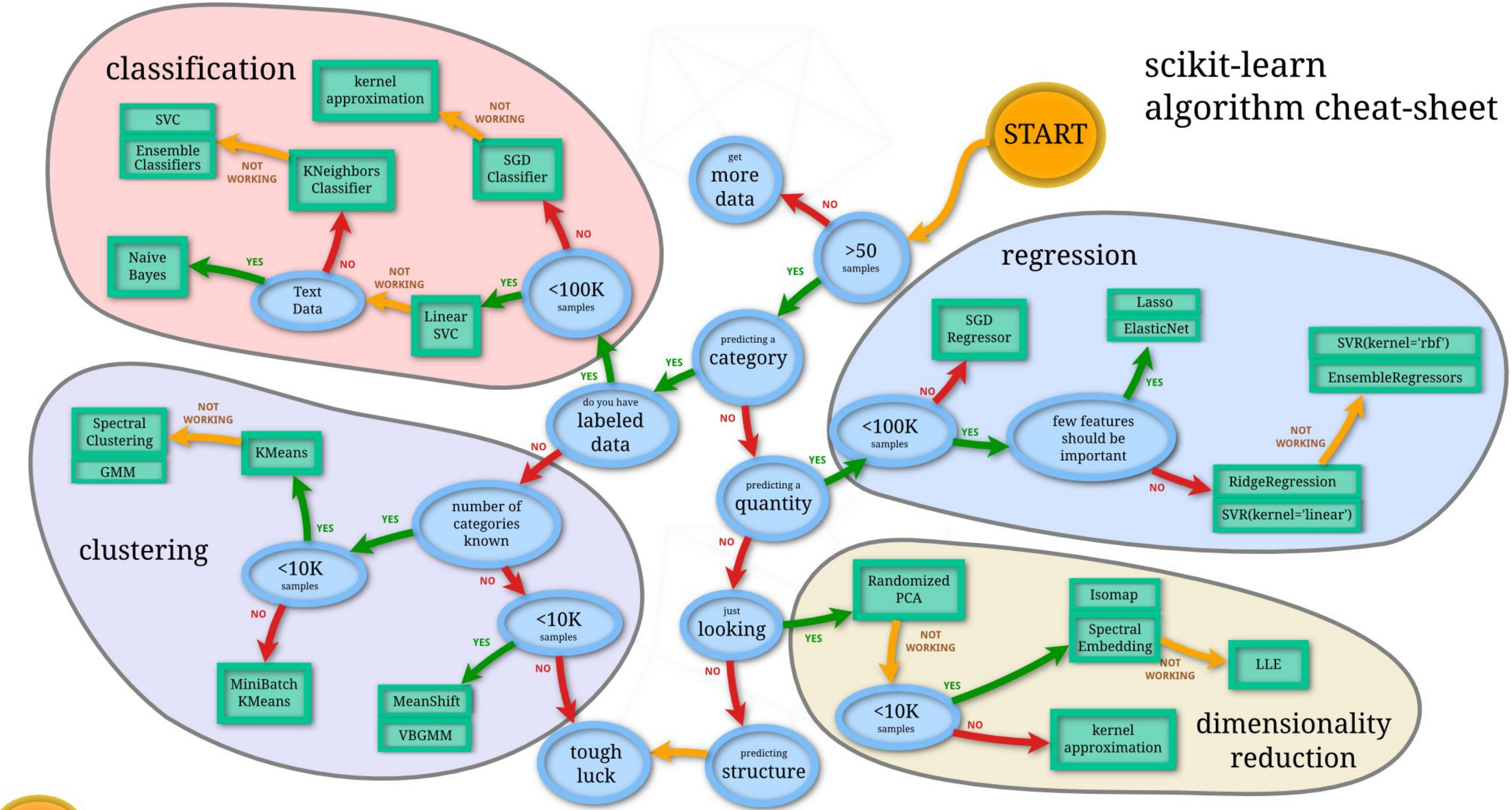
Last edit in spring ’25. Condensed (with ~2 slide deletions + a little more on RF) from 22m8a, Merging in 8b before DTs. Collectively similar to 2021’s M8a & M8b [which have a videos].

Mark Gerstein
Yale U.

Supervised Mining:

Overview

The World of “Classic” ML



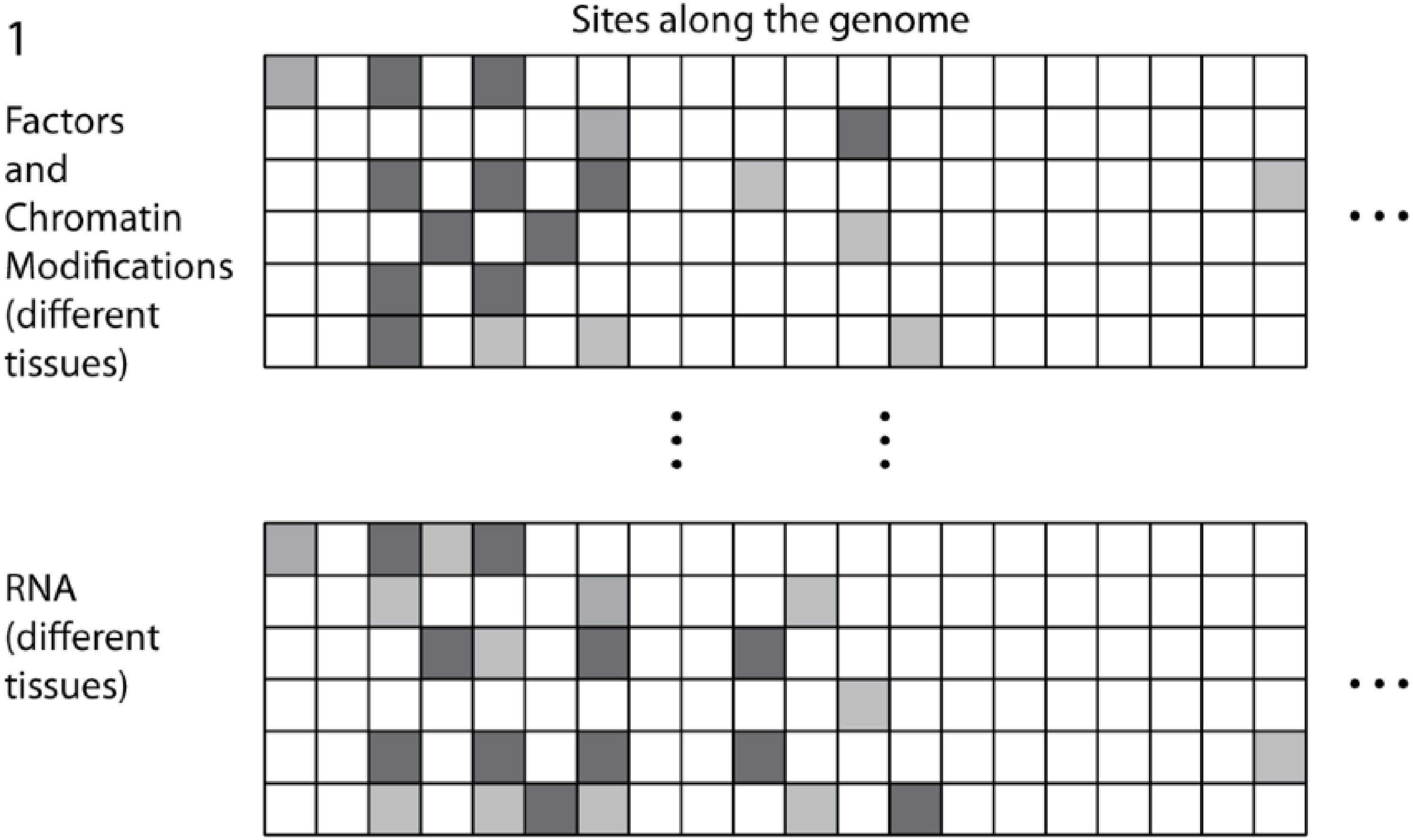
Back

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

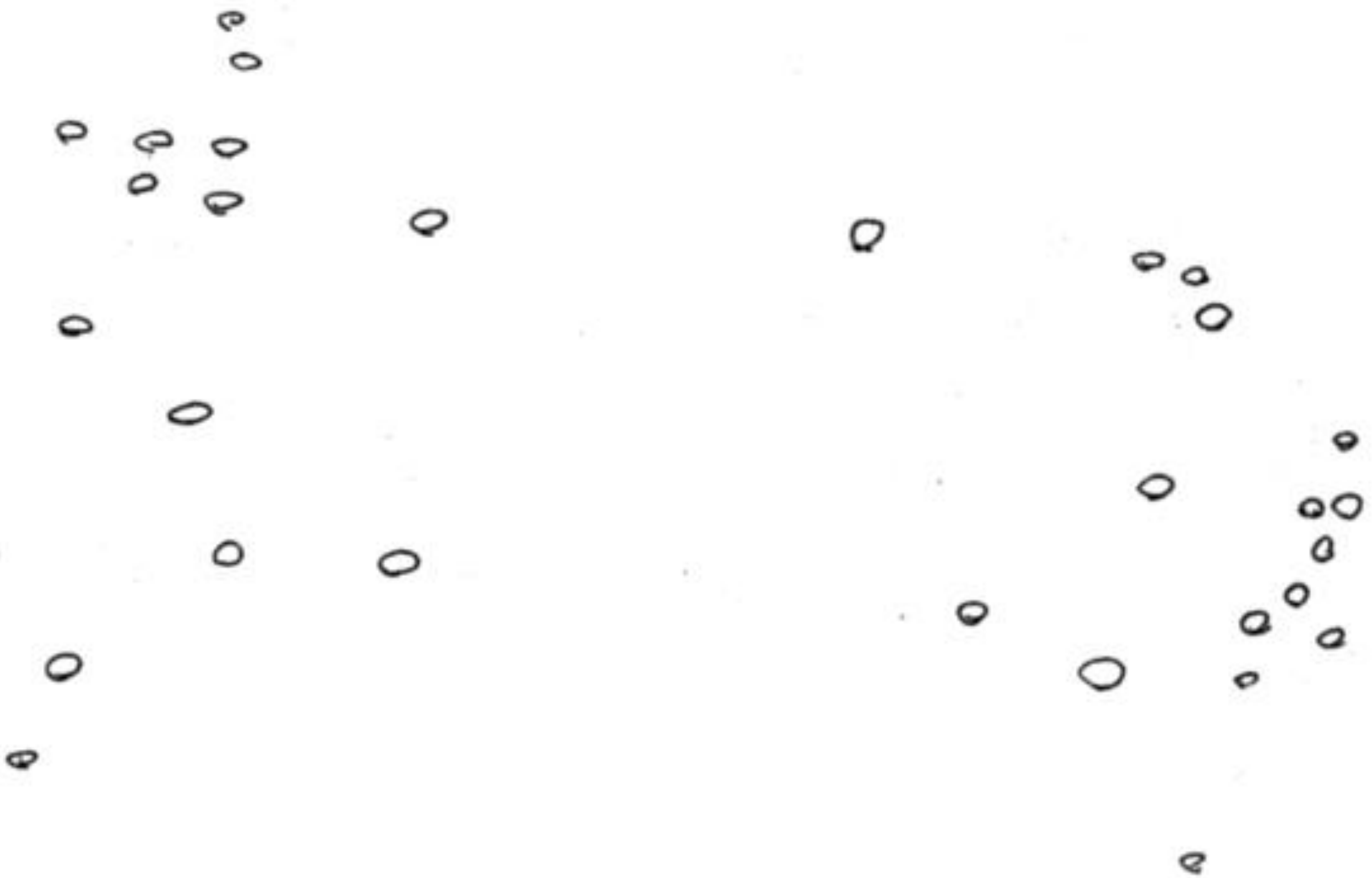
Distinctions in Supervised Learning

- 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

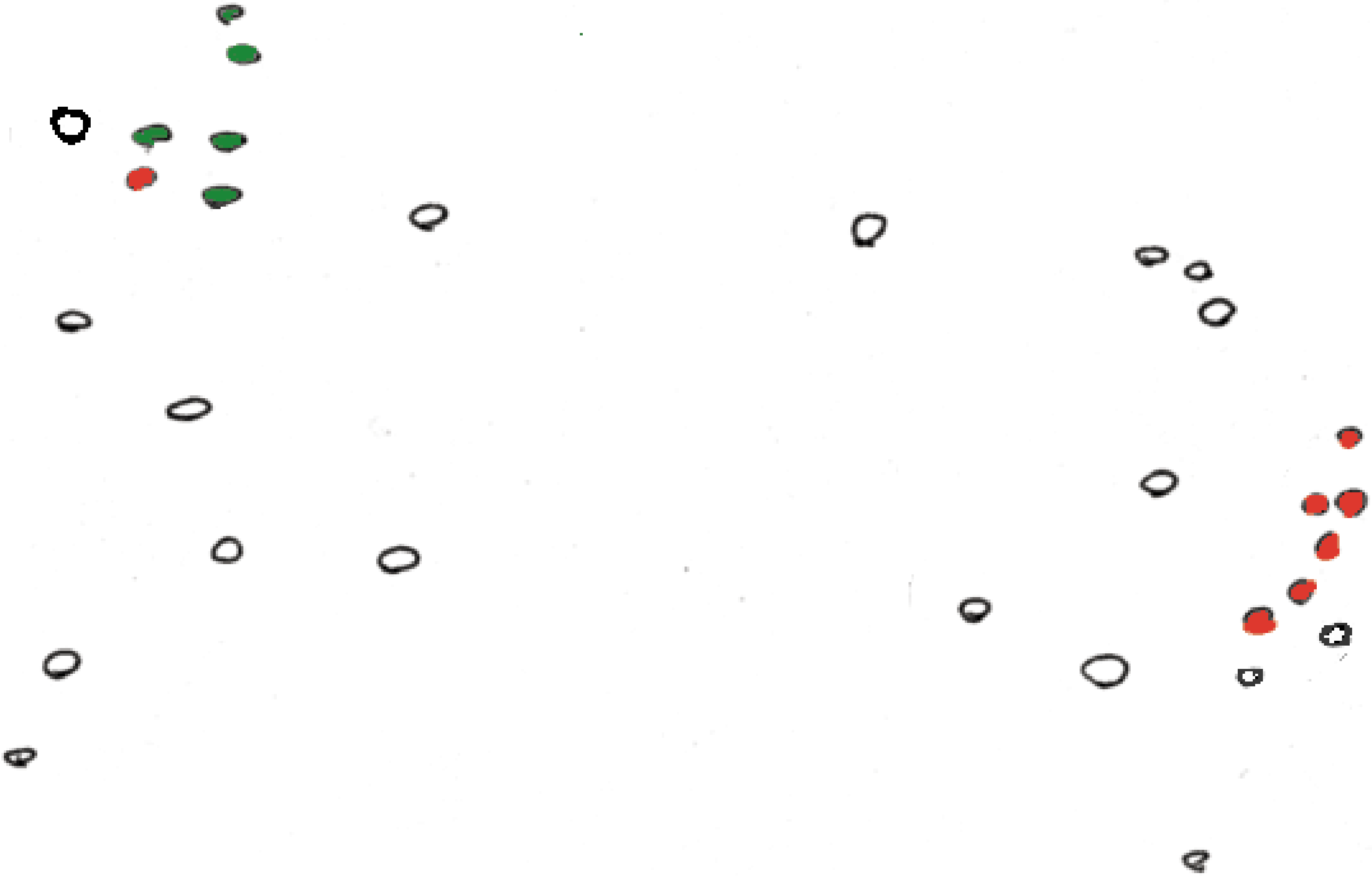
Structure of Genomic Features Matrix



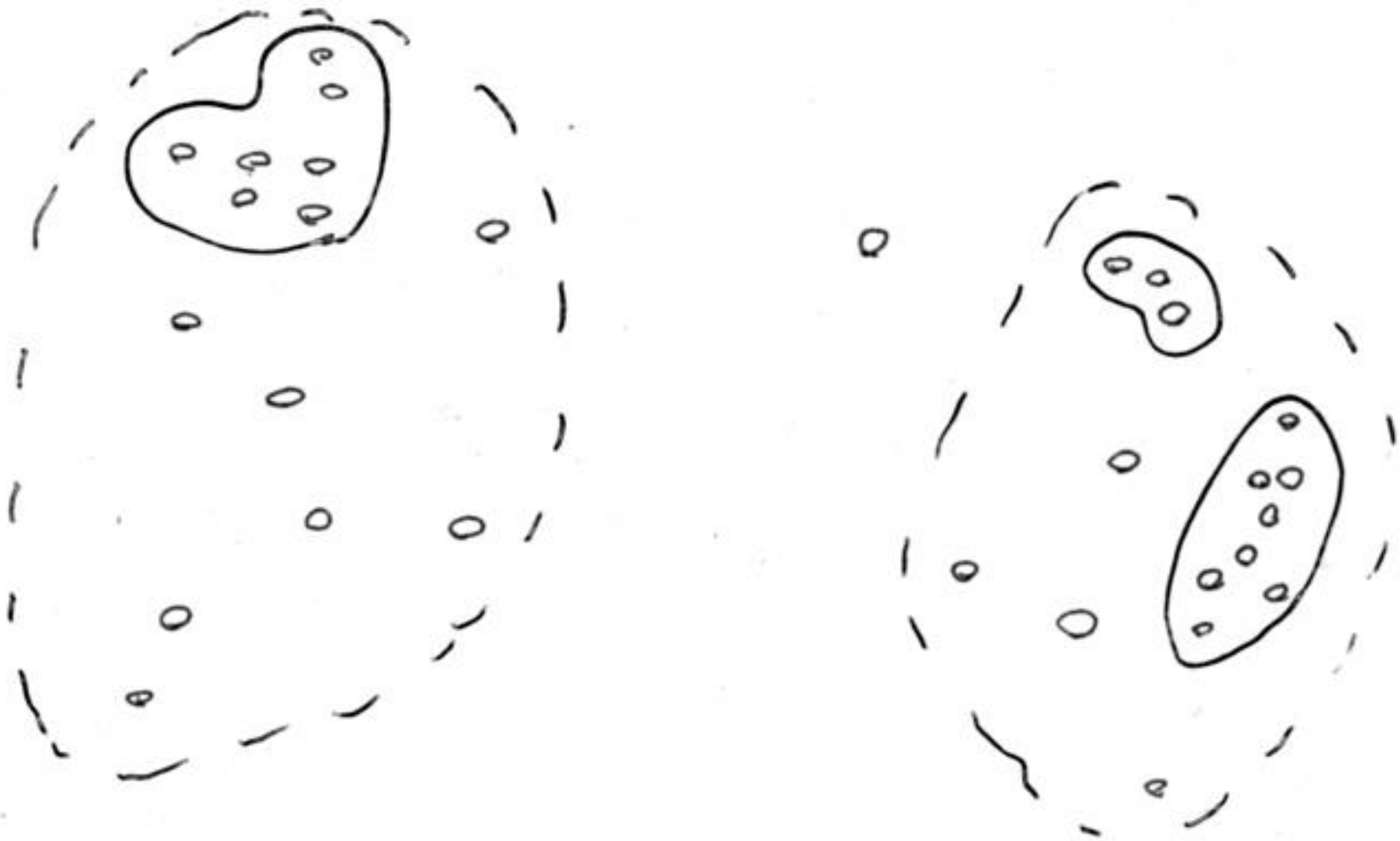
Represent predictors in abstract high dimensional space



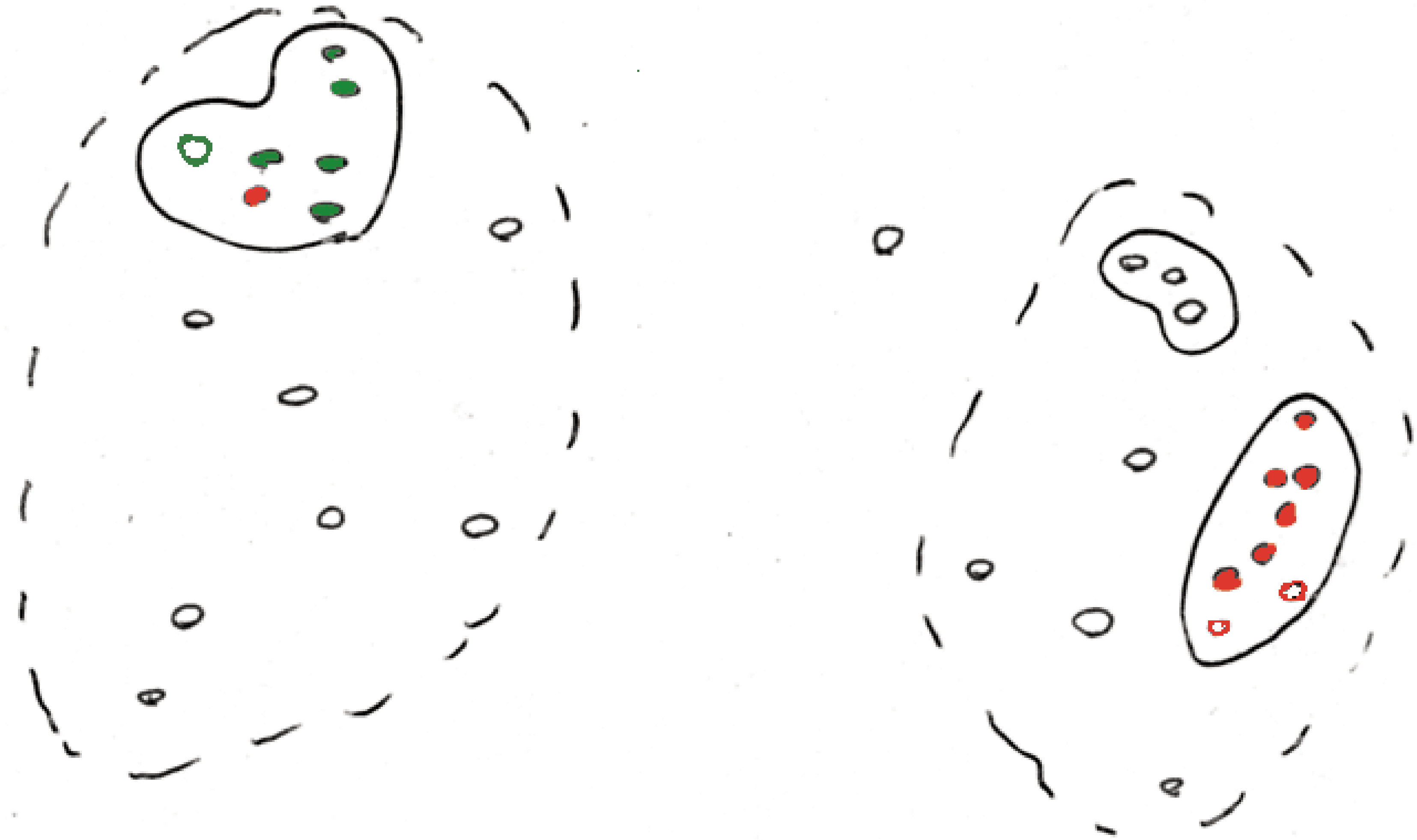
“Label” Certain Points



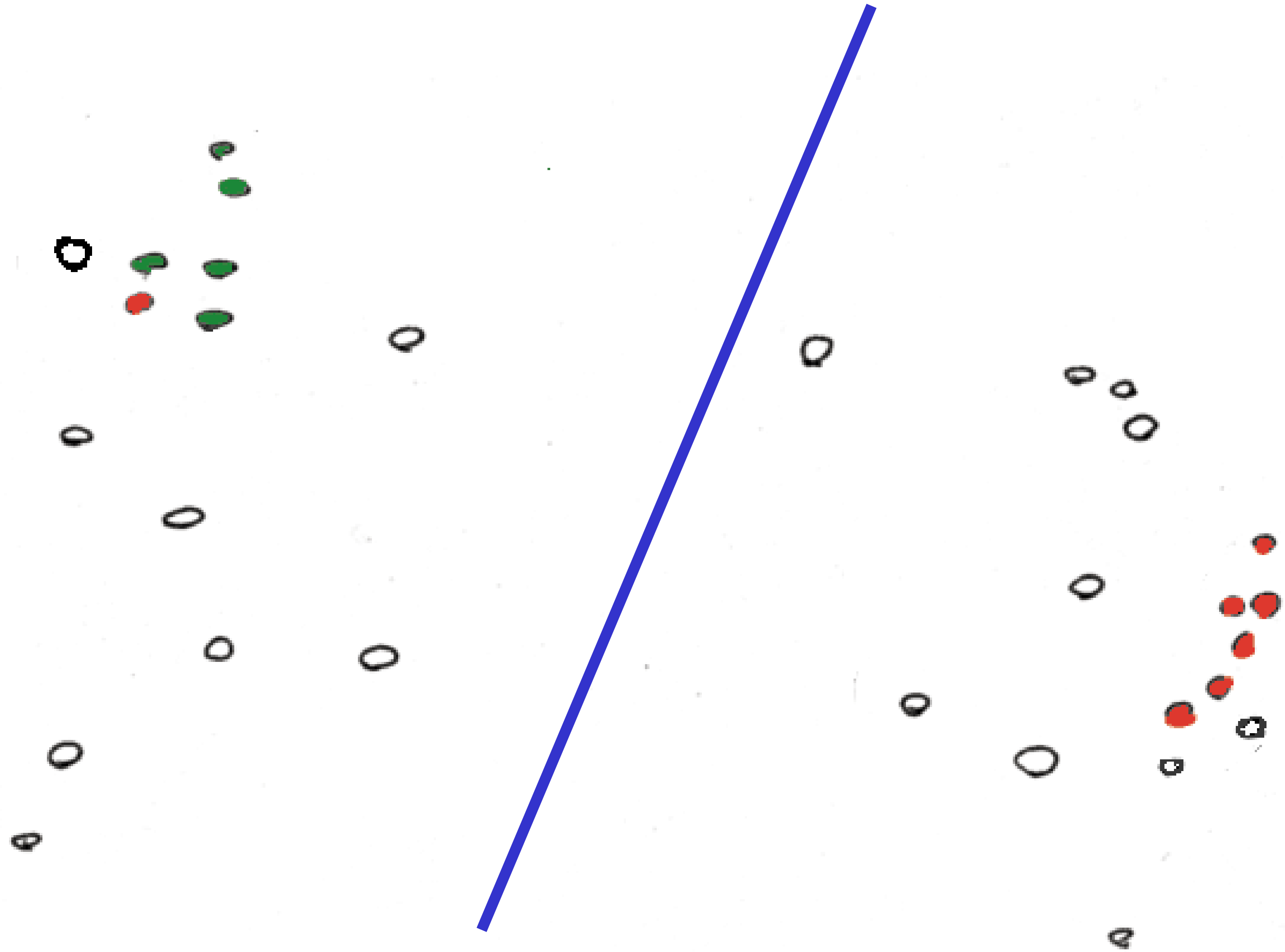
“Cluster” predictors (Unsupervised)



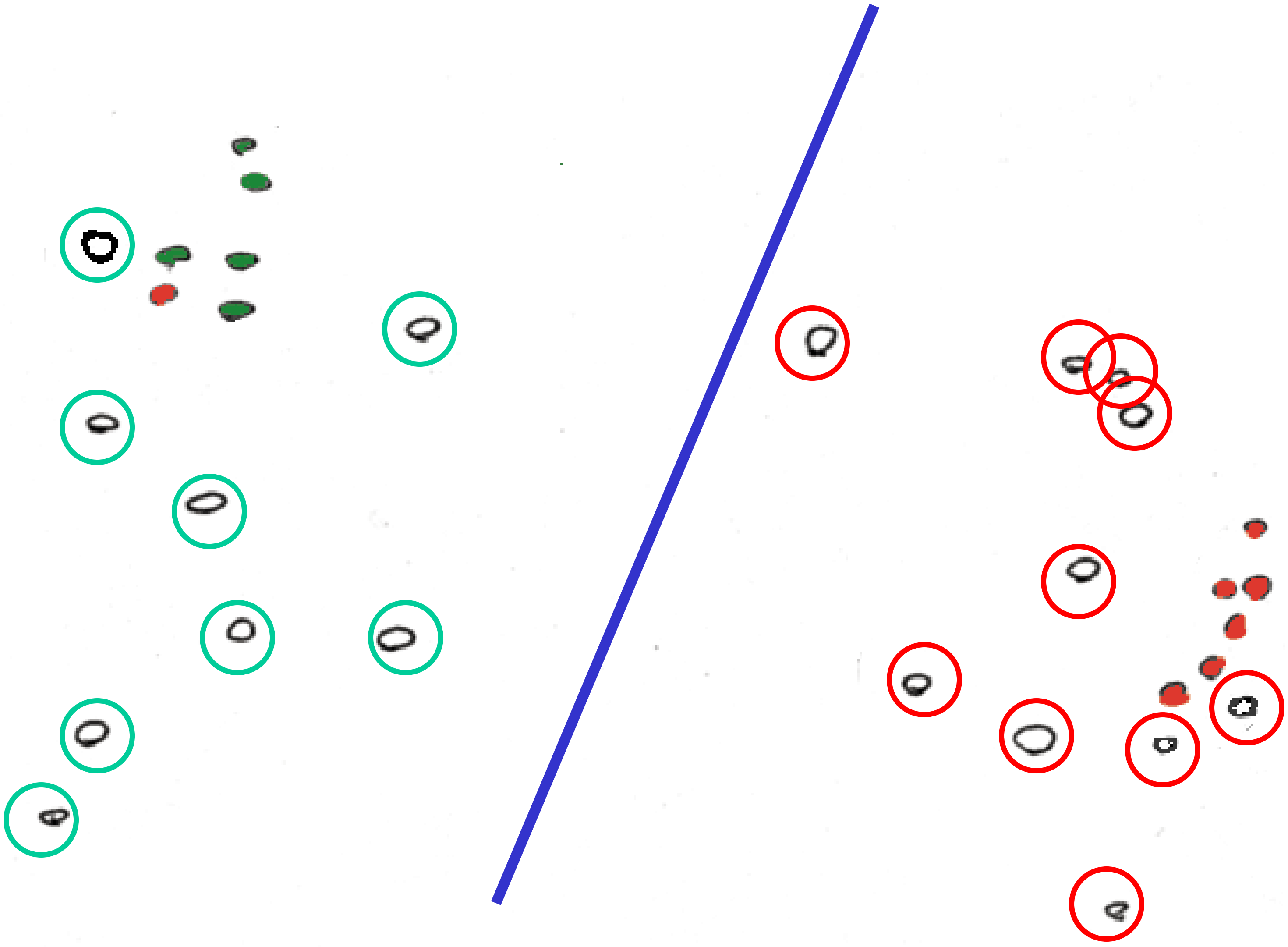
Use Clusters to predict Response
(Unsupervised, guilt-by-association)



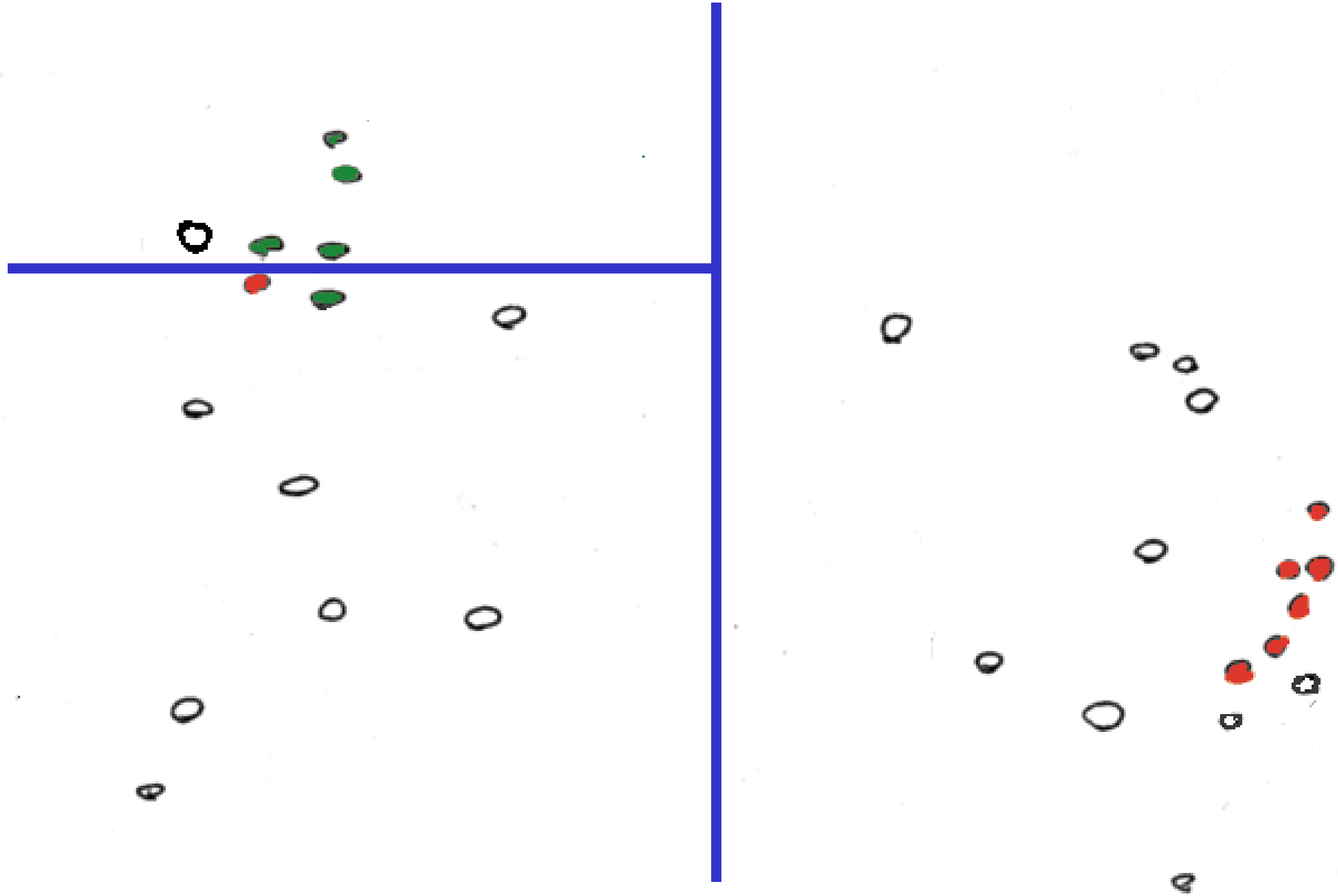
Find a Division to Separate Tagged Points



Extrapolate to Untagged Points



Find a Division to Separate Tagged Points



Supervised Mining:

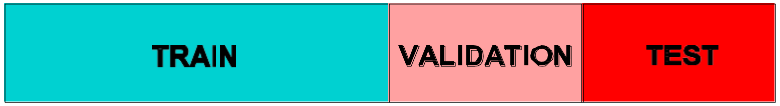
Assessment, Cross-Validation & ROC Curves

Evaluating performance: What? How?

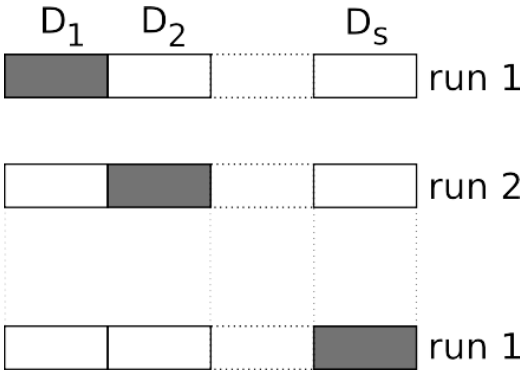
A. *<u>What</u> do we want to evaluate?*

GENERALIZATION

Therefore, it is mandatory to divide your dataset:



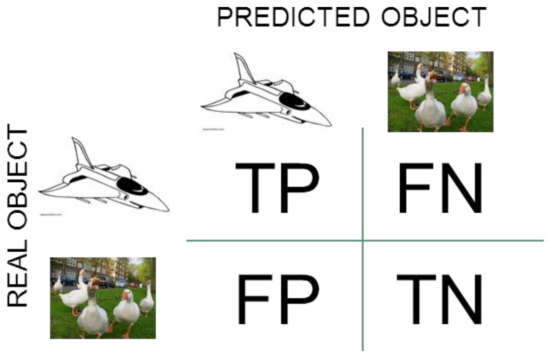
Alternatively, use Cross Validation:



Slide from Alberto Paccanaro

B. How do we evaluate performance?

1. Classification problems



TP	FN
FP	TN

Accuracy

TP+TN/(TP+FP+FN+TN)

Sensitivity (or TPR)

TP/P=TP/(TP+FN)

Specificity

TN/N=TN/(TN+FP)

Positive predictive value (PPV)

TP/(TP+FP)

False positive rate (FPR)

FP/N=FP/(FP+TN)

False discovery rate (FDR)

FP/(FP+TP)

2. Regression problems Sum of squares error

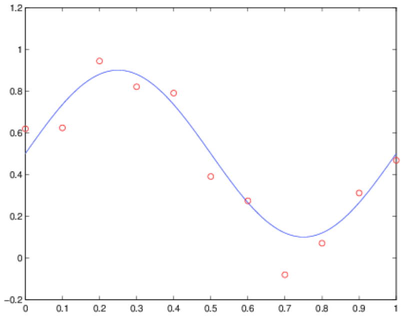
Root Mean Square error

ROC analysis is good for comparing binary classifiers

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

Slide from Alberto Paccanaro

Model dimensionality and overfitting

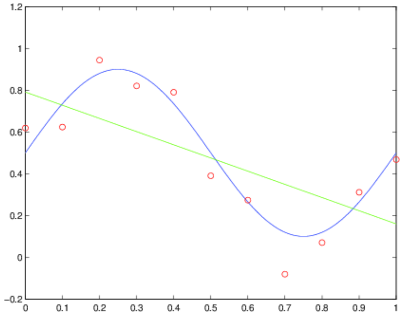


We are given the red dots.

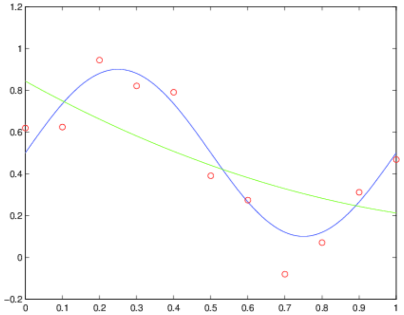
We assume that they are noisy samples from a signal/(function) – the blue curve – which we do not have (we only have the red dots).

We want to predict new points, i.e. the *y* coordinates for other values of *x* (e.g. *x > 1*)

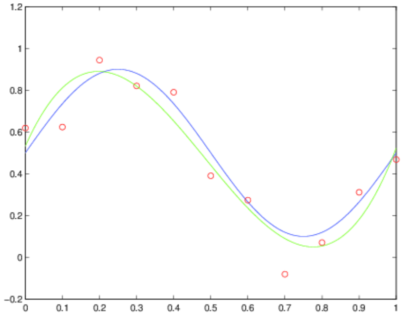
Our model needs to approximate the blue function. We decide to do it with polynomials.



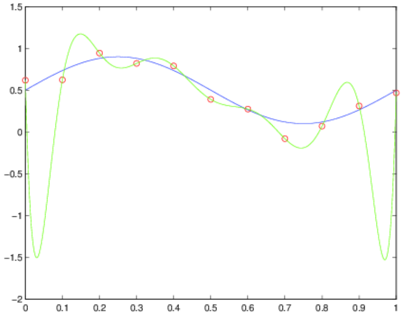
Degree 1 polynomial



Degree 2 polynomial



Degree 3 polynomial

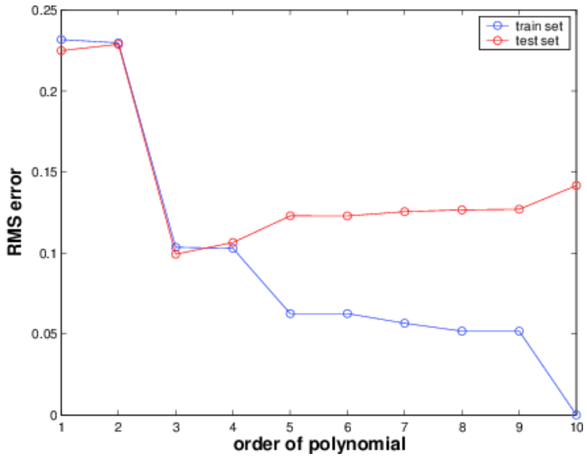


Degree 10 polynomial

Which one is best? And why?

Slide from Alberto Paccanaro

How does the GENERALIZATION performance vary, as we increase the complexity of the polynomial?



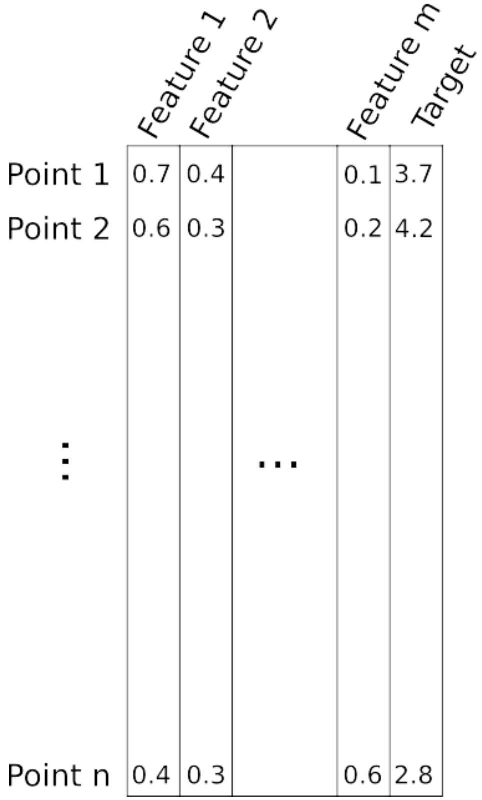
- Occam's razor *(William of Occam, ~1300):* Accept the simplest explanation that fits the data.

We should prefer simpler models to more complex models, and this preference should be traded off against the extent to which the model fits the data.

Related to “Bias-Variance” tradeoff.

Slide from Alberto Paccanaro

- IMPORTANT: increasing the number of features may lead to a reduction in performance if the number of datapoints is not increased. Why?

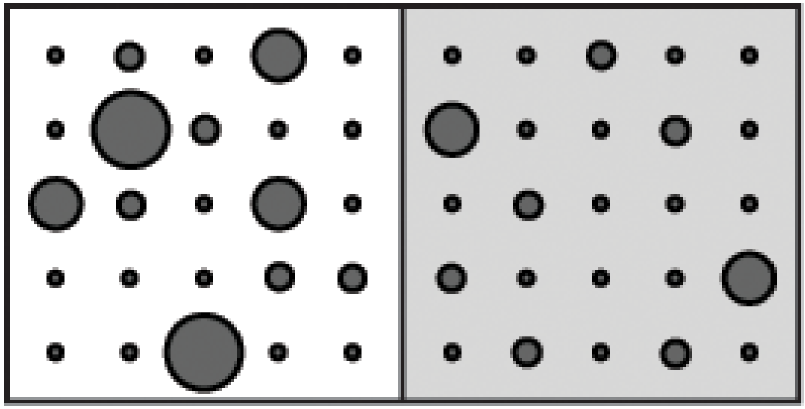


	Feature 1	Feature 2		Feature m	Target
Point 1	0.7	0.4		0.1	3.7
Point 2	0.6	0.3		0.2	4.2
⋮			…		
Point n	0.4	0.3		0.6	2.8

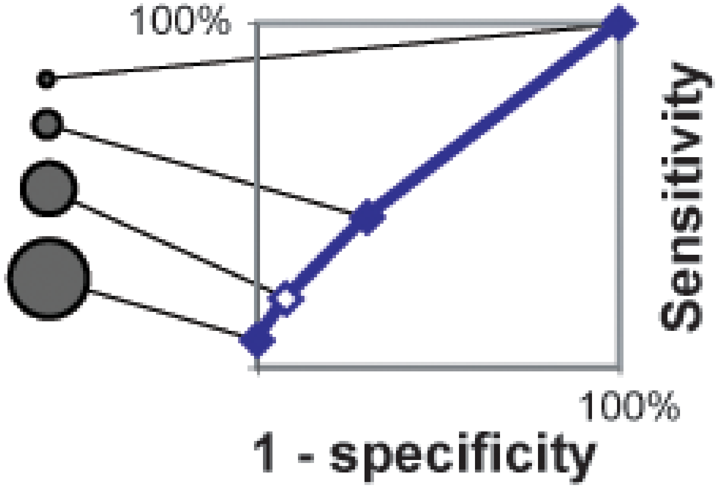
This is related to the "Curse of Dimensionality" Bellman, 1961.

Slide from Alberto Paccanaro

ROC plots & Comparison of Predictions against a Positive & Negative Gold Standard



Threshold "predictions" (strength of positive score is represented by circle size) at different levels and compare to + and - gold standards (represented by white & gray squares). 50 total instances, half + and half -. A concrete example would be doing cancer prediction 50 individuals with known cancer status.

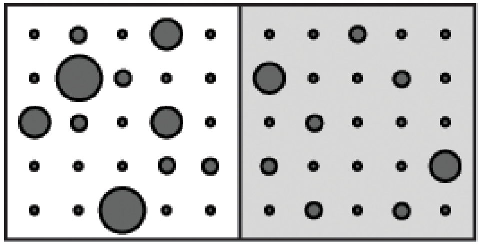


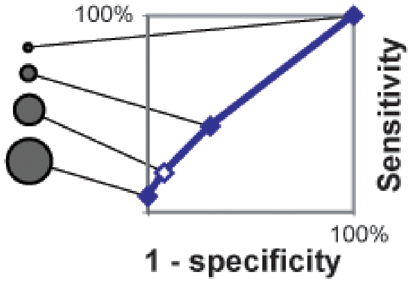
"Coverage" (TP/P)

"Error Rate" (FP/N)

ROC plot
(cross validated)

R Jansen, M Gerstein (2004). *Curr Opin Microbiol 7*: 535





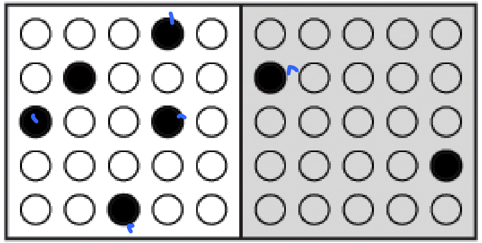
Effect on Predictions of Large Number of Negatives

(e.g. terrorist identification or breast cancer screening)

Sensitivity

1- specificity

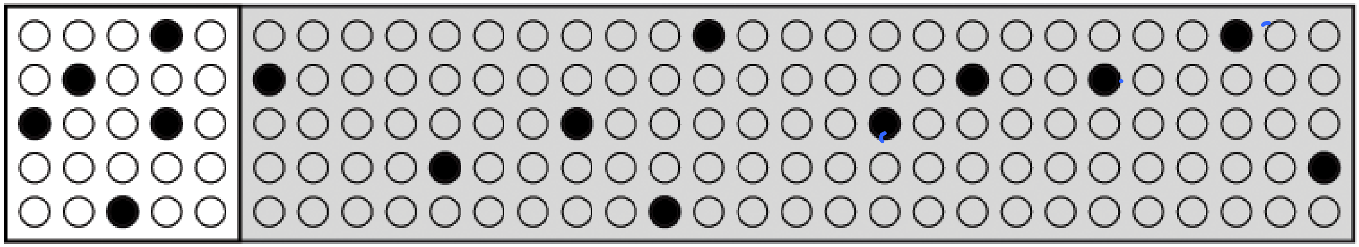
Positive predictive value



$$\frac{5}{25} = \mathbf{20\%}$$

$$\frac{2}{25} = \mathbf{8\%}$$

$$\frac{5}{5+2} \approx \mathbf{71\%}$$



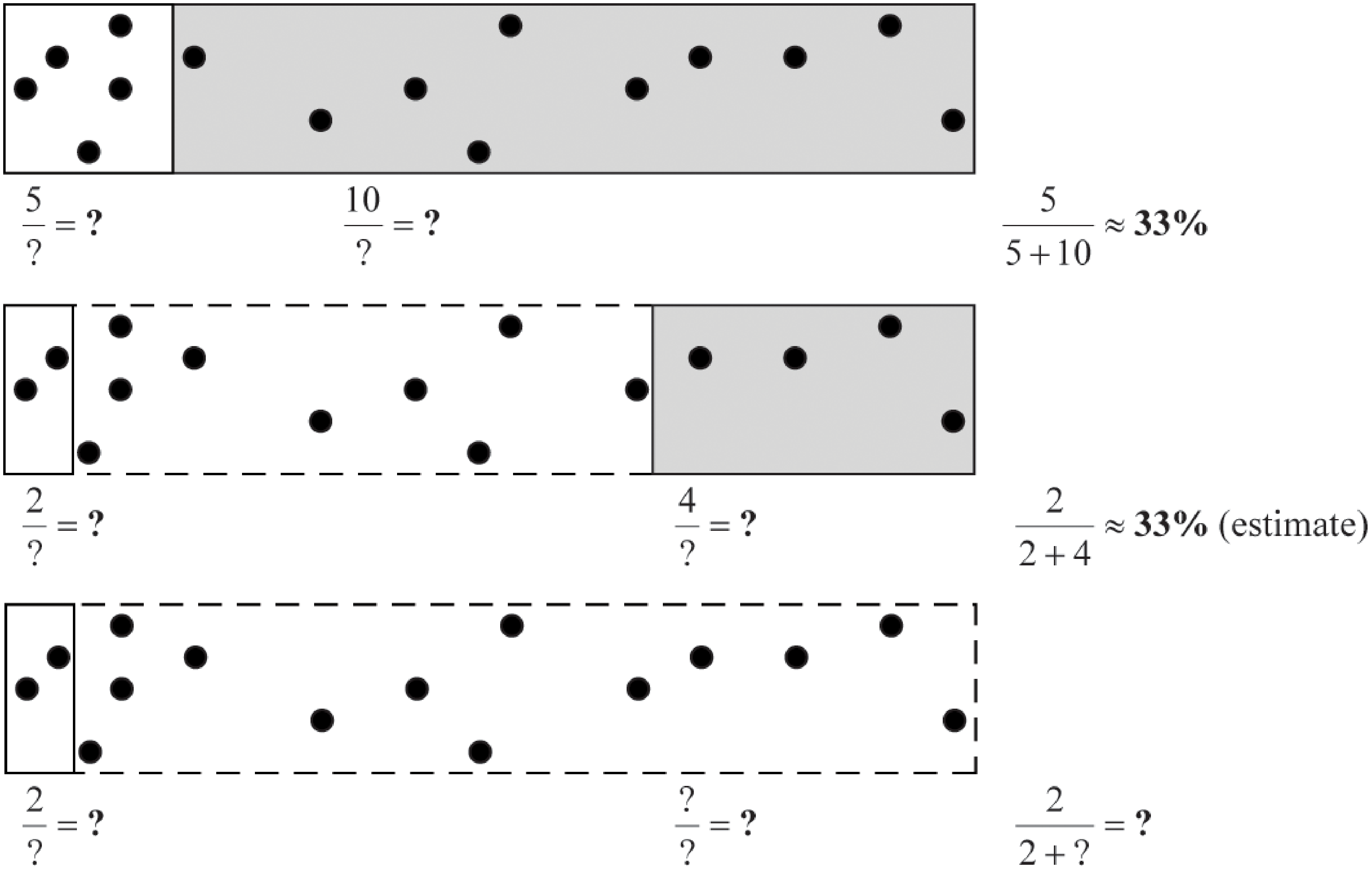
$$\frac{5}{25} = \mathbf{20\%}$$

$$\frac{10}{125} = \mathbf{8\%}$$

$$\frac{5}{5+10} \approx \mathbf{33\%}$$

R Jansen, M Gerstein (2004). *Curr Opin Microbiol 7*: 535

Importance of Balanced Positive and Negative Examples



$\frac{5}{?} = \mathbf{?}$ $\frac{10}{?} = \mathbf{?}$ $\frac{5}{5+10} \approx \mathbf{33\%}$

$\frac{2}{?} = \mathbf{?}$ $\frac{4}{?} = \mathbf{?}$ $\frac{2}{2+4} \approx \mathbf{33\%}$ (estimate)

$\frac{2}{?} = \mathbf{?}$ $\frac{?}{?} = \mathbf{?}$ $\frac{2}{2+?} = \mathbf{?}$

R Jansen, M Gerstein (2004). *Curr Opin Microbiol* 7: 535

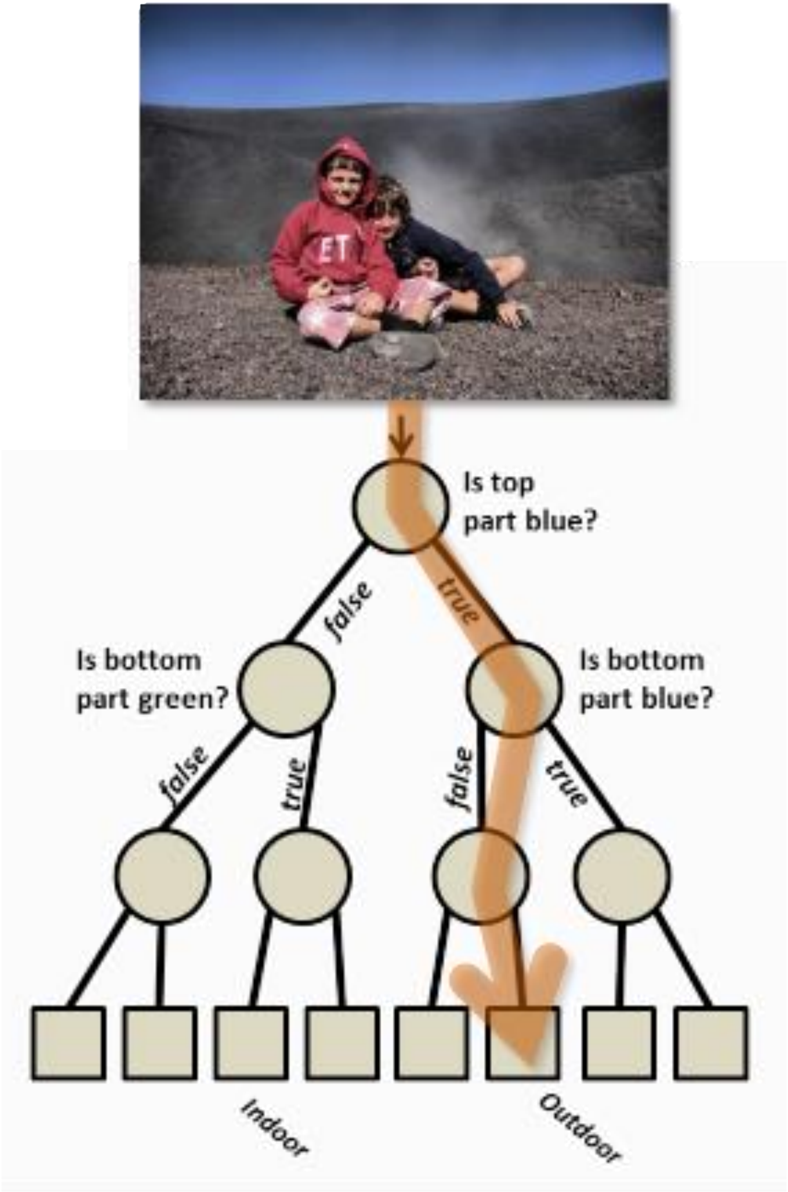
Supervised Mining:

Decision Trees

(“Jumping to first method”)

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

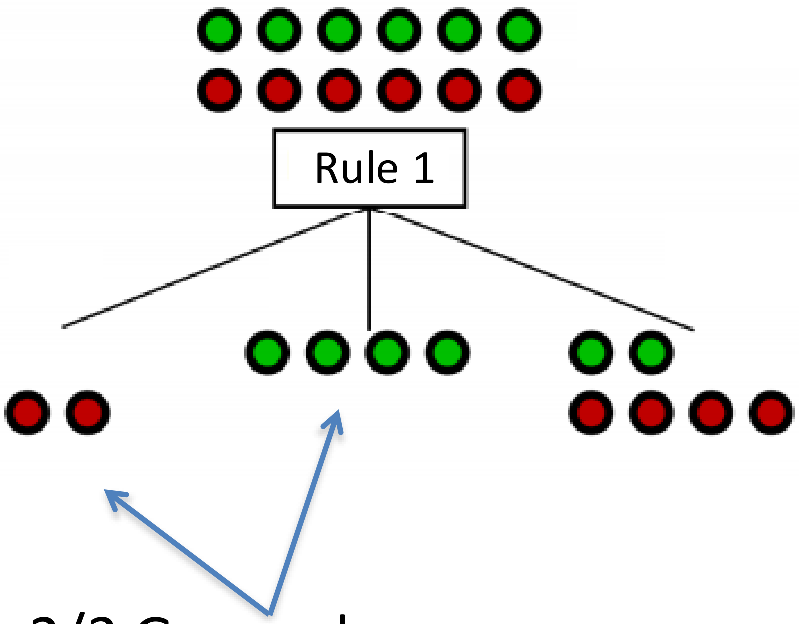


- Example: Is a picture taken inside or outside?

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

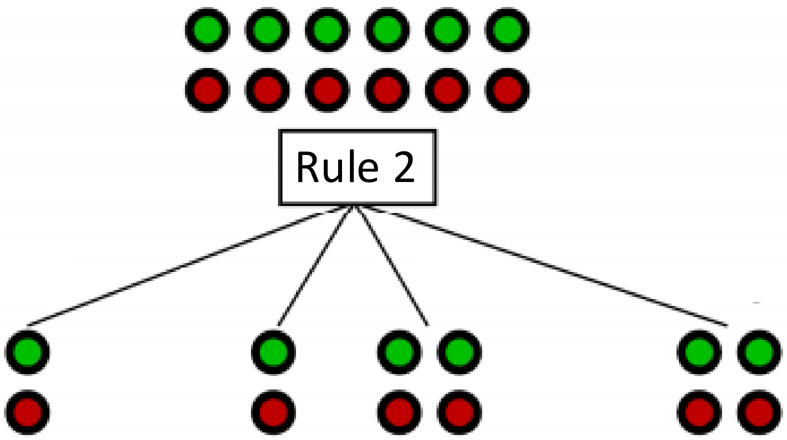
What makes a good rule?

- 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
  (or increase in homogeneity)
  - Most popular metric: Information theoretic entropy

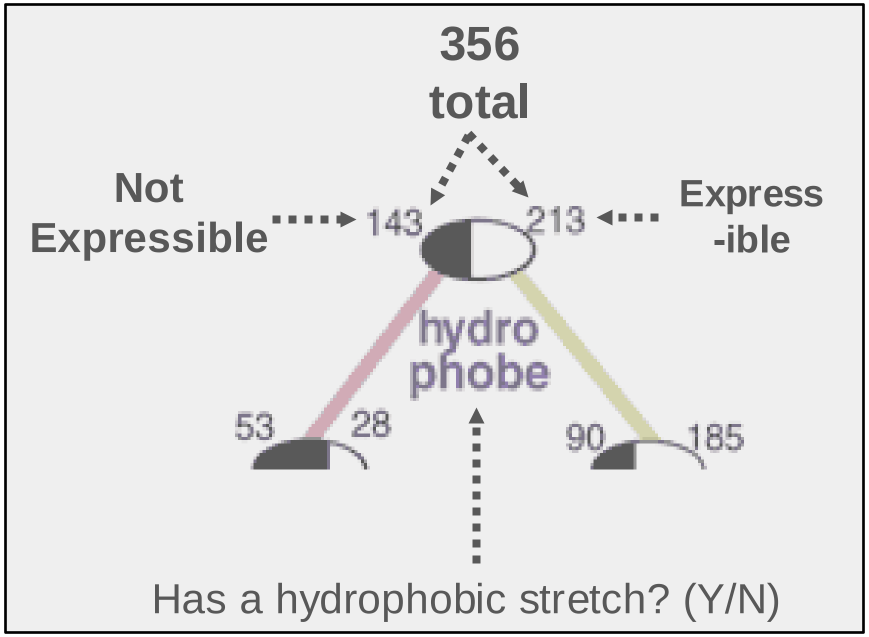
$$S = -\sum_{i=1}^{m} p_i \log p_i$$

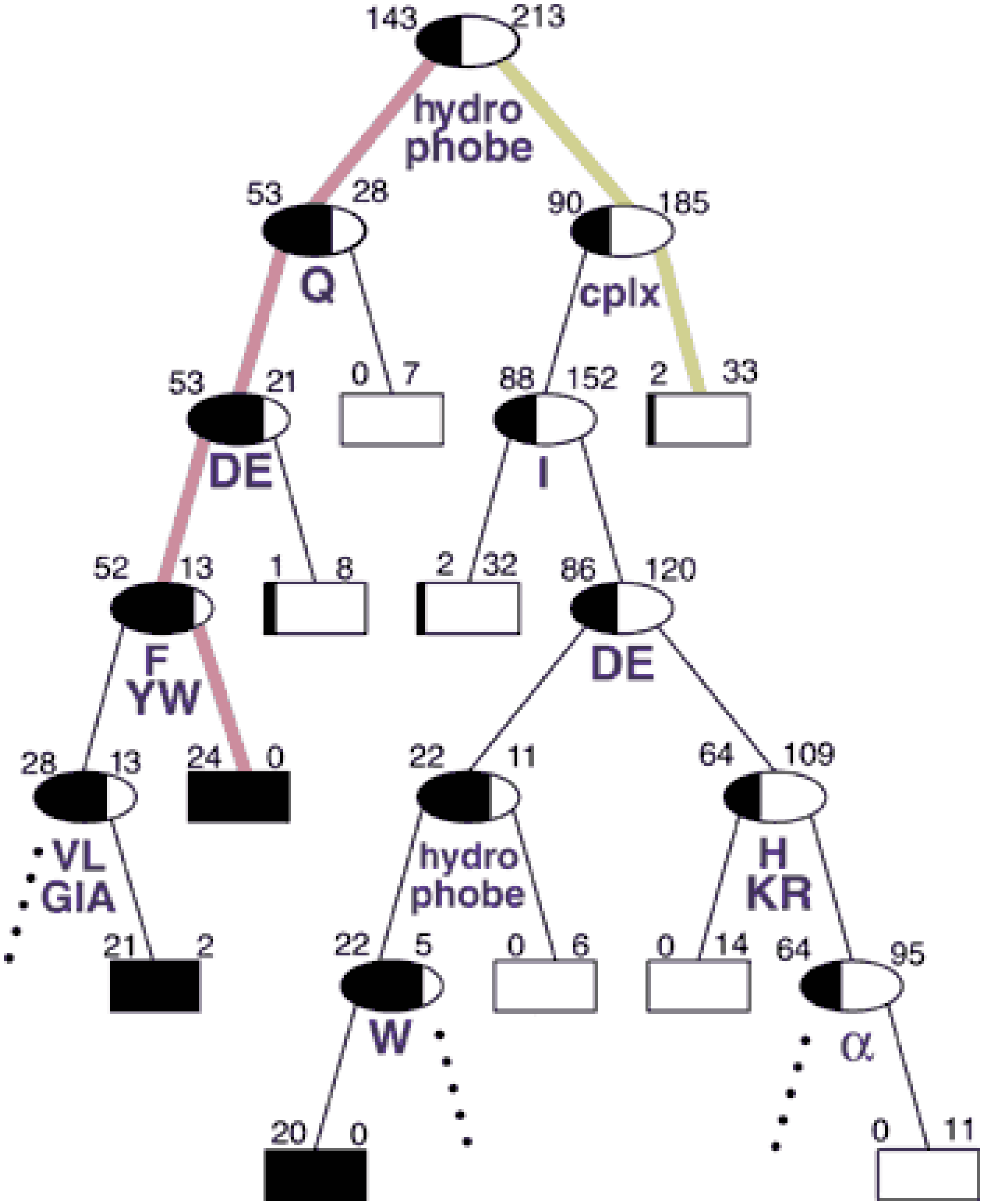
  - Use frequency of classifier characteristic within group as probability
  - Minimize entropy to achieve homogenous group

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

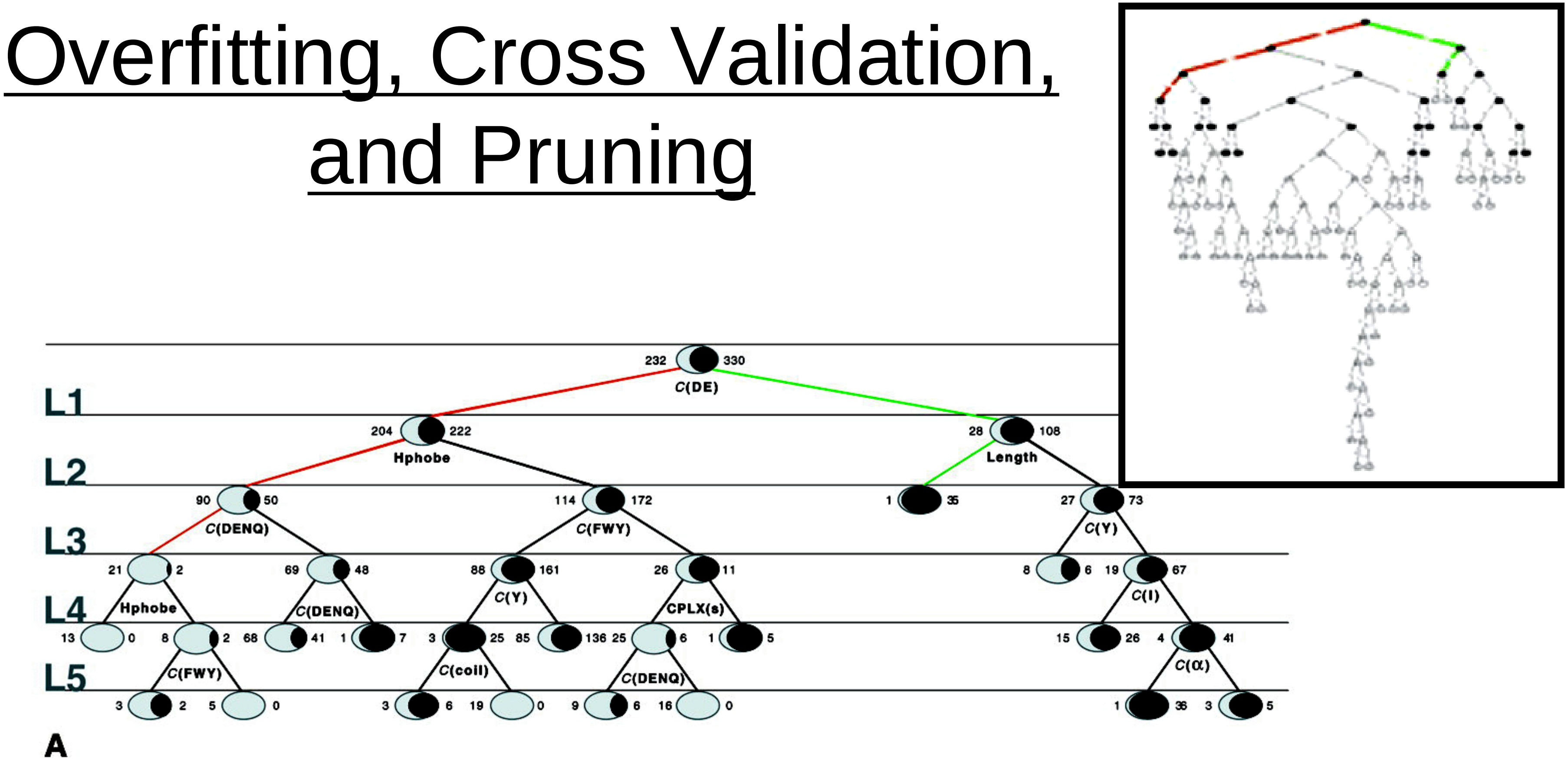


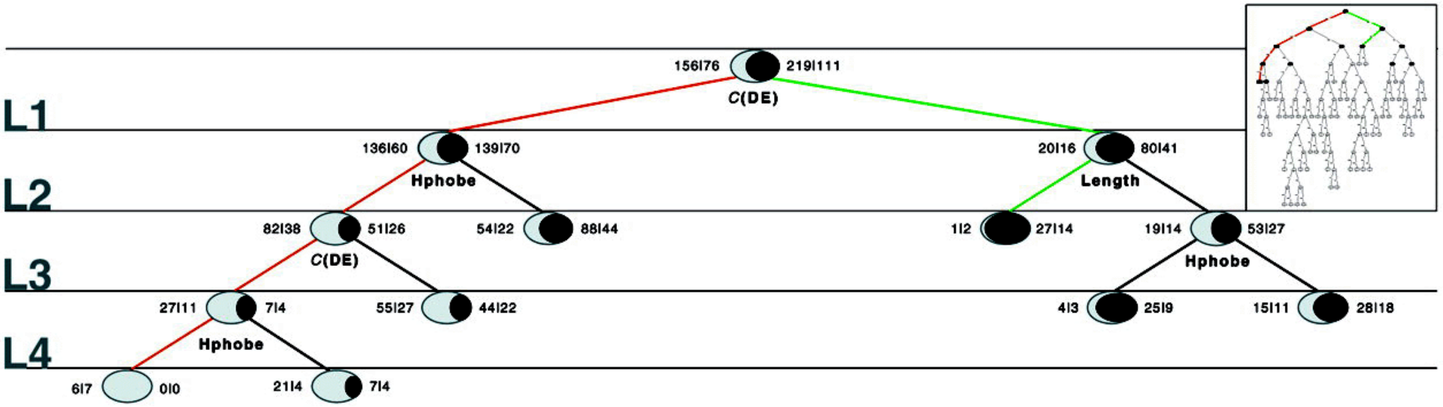


Analysis of the Suitability of 500 M. thermo. proteins
to find optimal sequences purification

[Bertone et al. NAR ('01)]

Overfitting, Cross Validation, and Pruning





Random Forest (RF)

- Basic decision tree (DT) method is very sensitive to dataset selection & noise in the data
- RFs are ensemble of DTs; address this issue
  - Build many DTs on bootstrapped training samples. (Reduces sensitivity to noise.)
  - Each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of predictors. (Decorrelates “bagged” DTs.)
  - Finally, we average or vote amongst the trees

[ISLR, chap 8]

References

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
  An Introduction to Statistical Learning: with Applications in R
  [ISLR (2nd edition)]
  https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ + https://www.statlearning.com
  (Chap 2 gives a nice overview on key concepts in ML.
  Chapter 8 to 8.2.2 gives background on DTs.)

- Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021).
  A guide to machine learning for biologists.
  Nature Reviews Molecular Cell Biology, 23(1), 40–55.
  https://doi.org/10.1038/s41580-021-00407-0
  (Good reference, but for this pack just go to up to section on “key concepts.”)

- Agarwal, R. (2024, March 29).
  ROC Curves and AUC: The Ultimate guide. Built In.
  https://builtin.com/data-science/roc-curves-auc
  (Optional extra background on ROC)