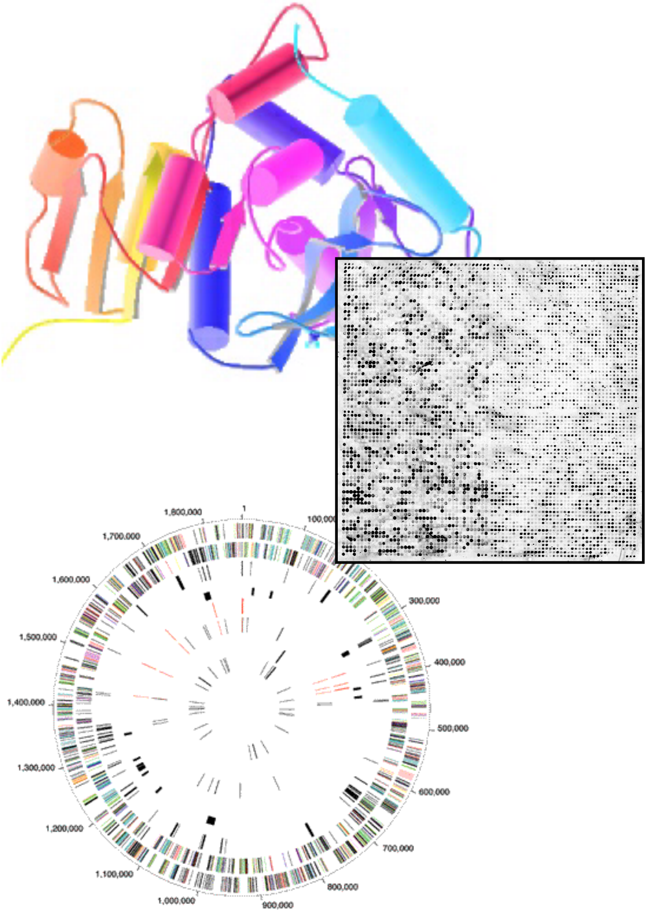
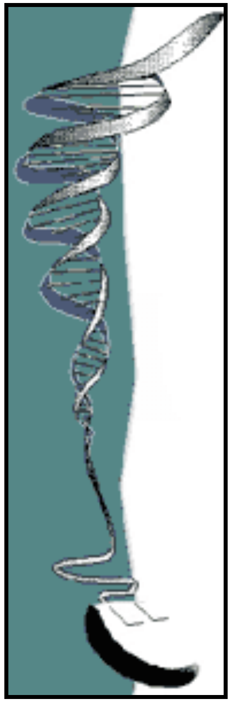
Biomedical Data Science:
Supervised Datamining





Mark Gerstein, Yale University

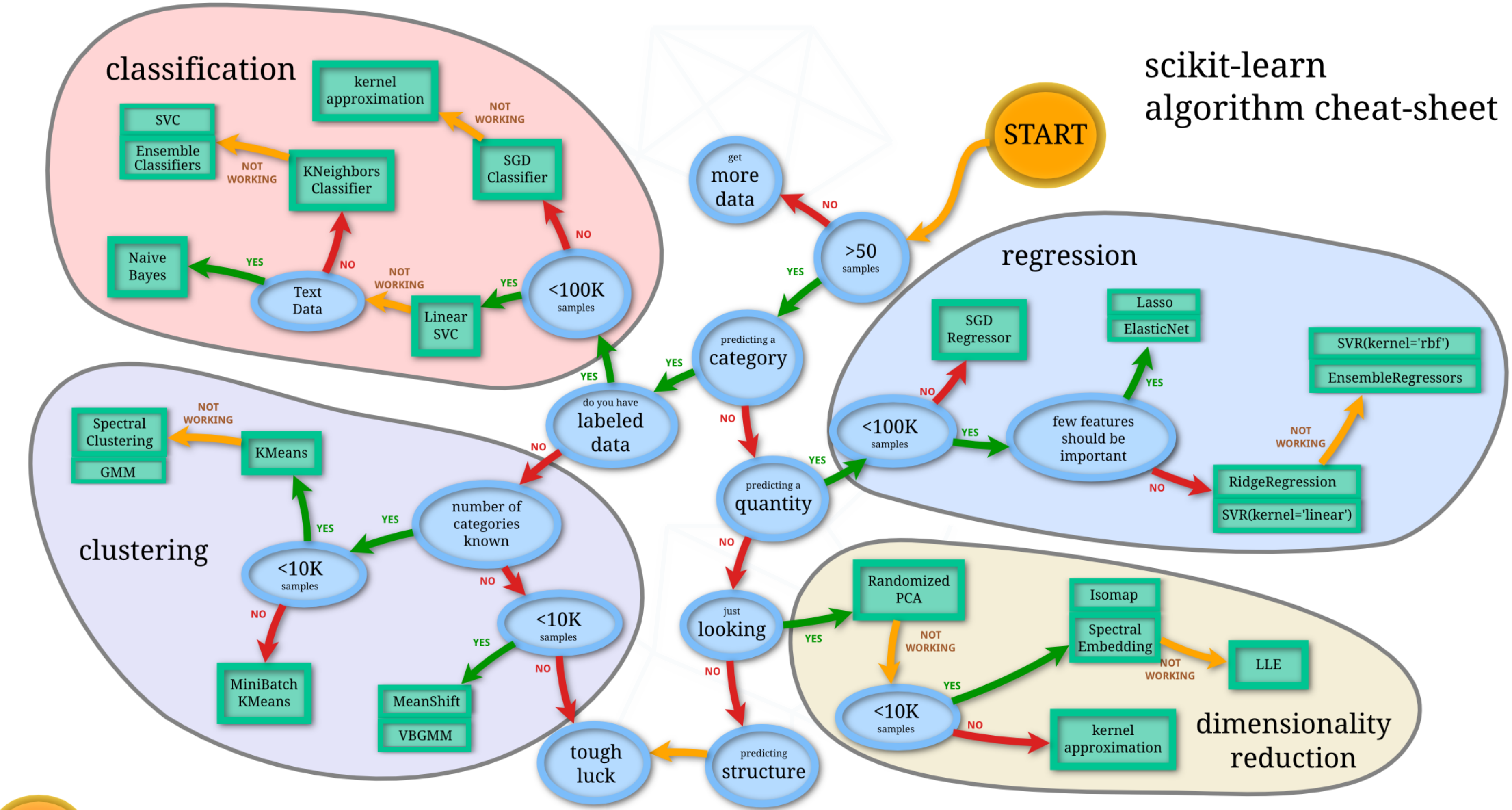
GersteinLab.org/courses/452

(last edit in spring ‘19, pack #9)

Supervised Mining:

Overview

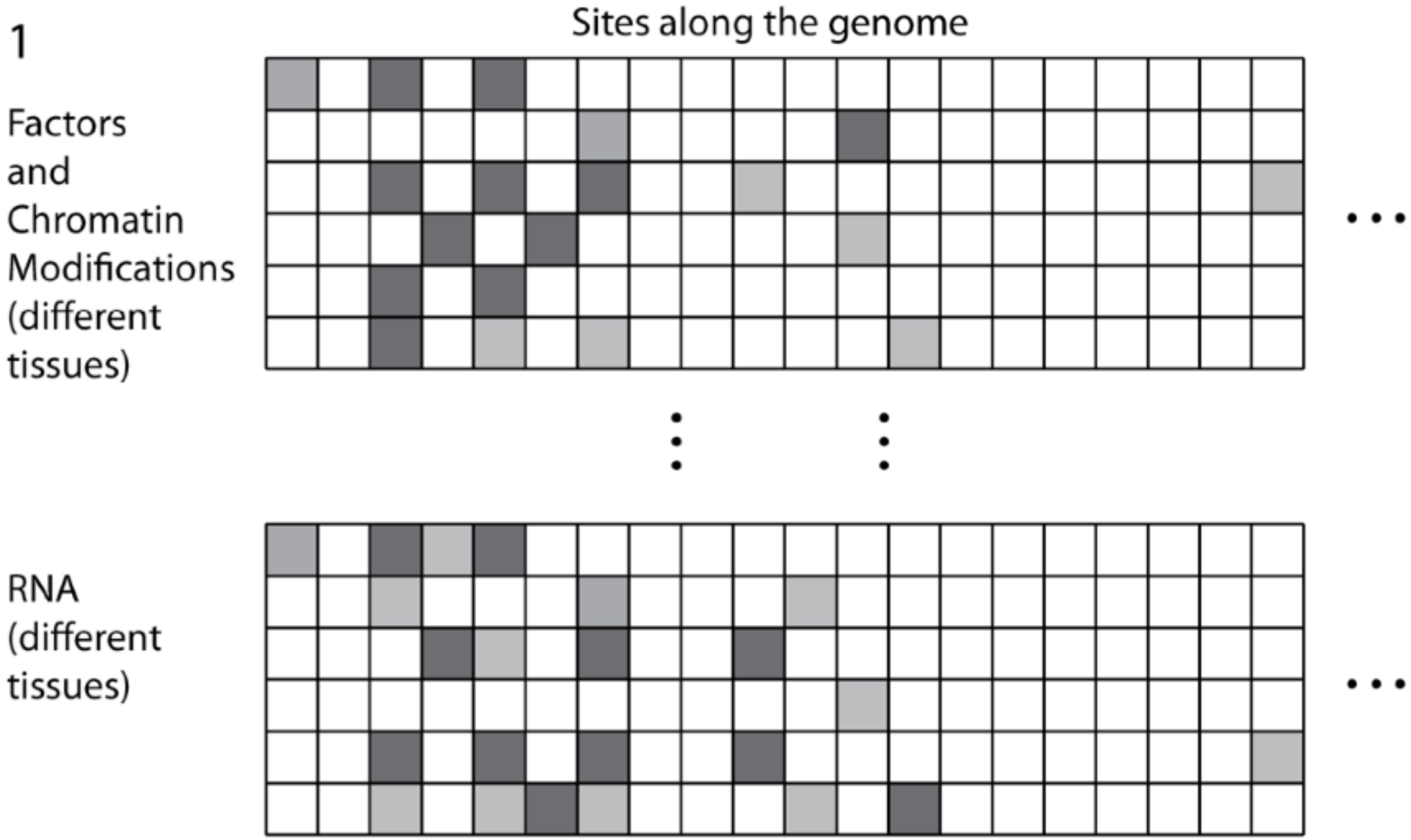
The World of Machine Learning



Back

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

Structure of Genomic Features Matrix



1

Sites along the genome

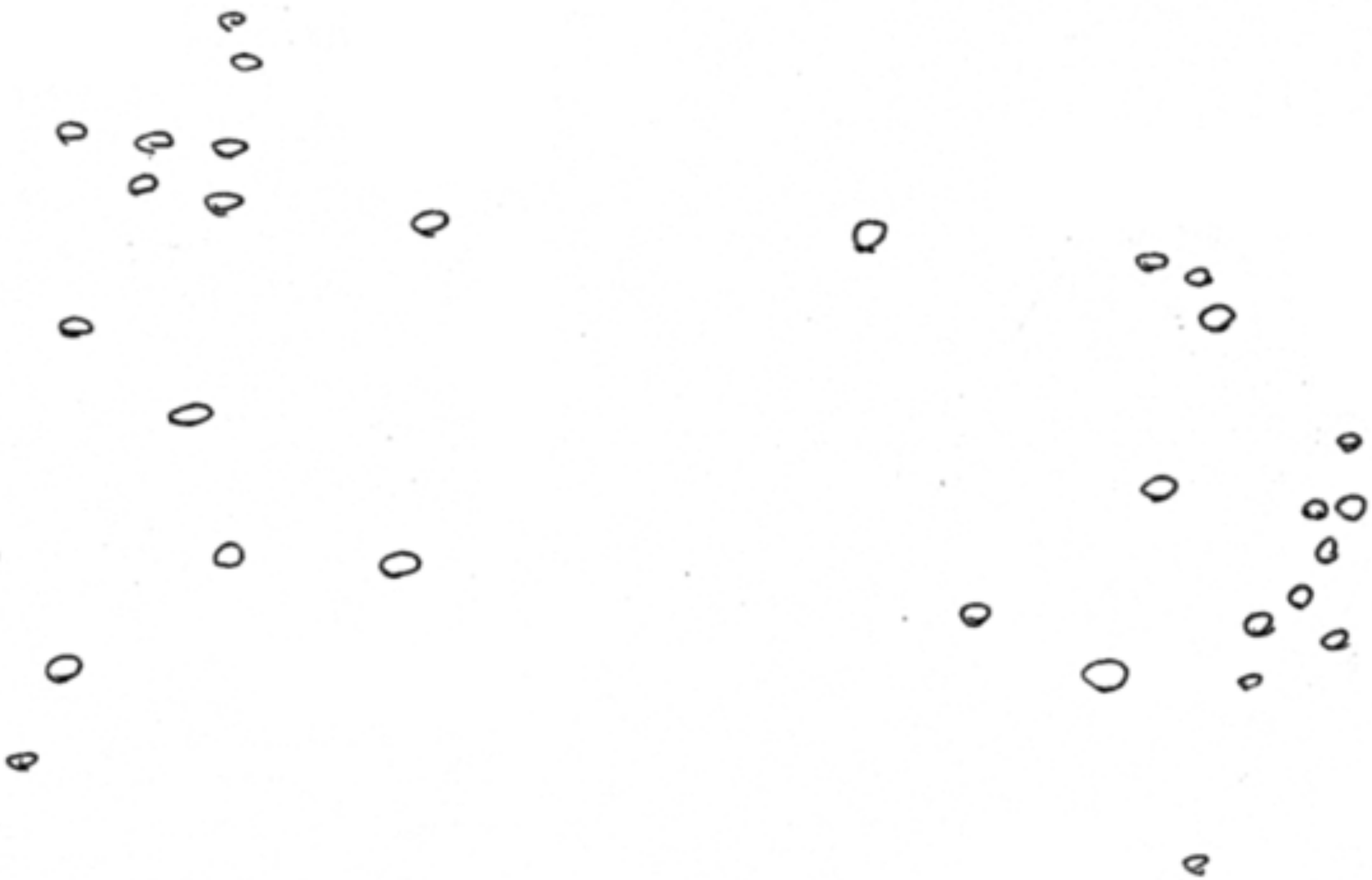
Factors and Chromatin Modifications (different tissues)

RNA (different tissues)

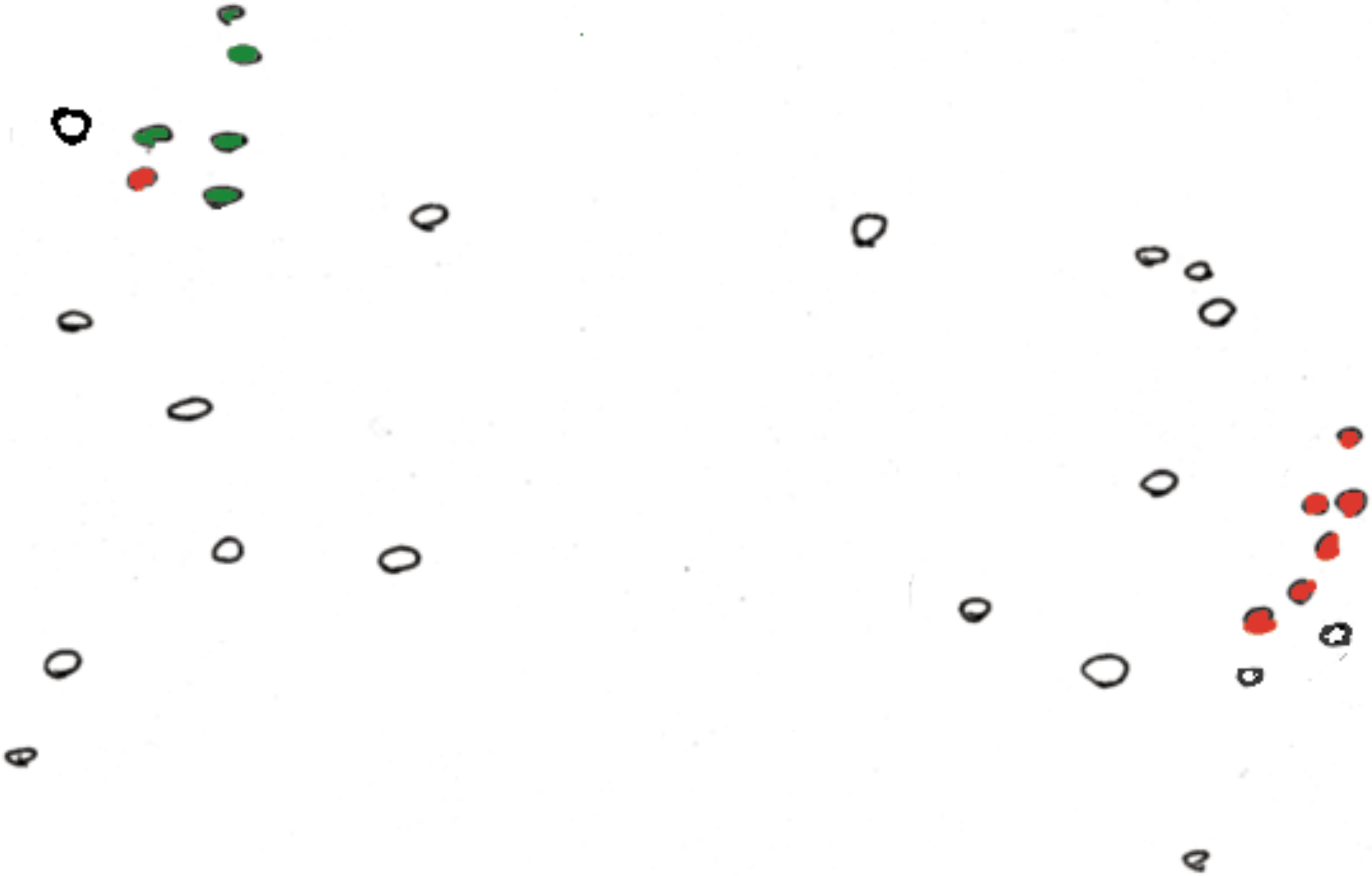
Arrange data in a tabulated form, each row representing an example and each column representing a feature, including the dependent experimental quantity to be predicted.

	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

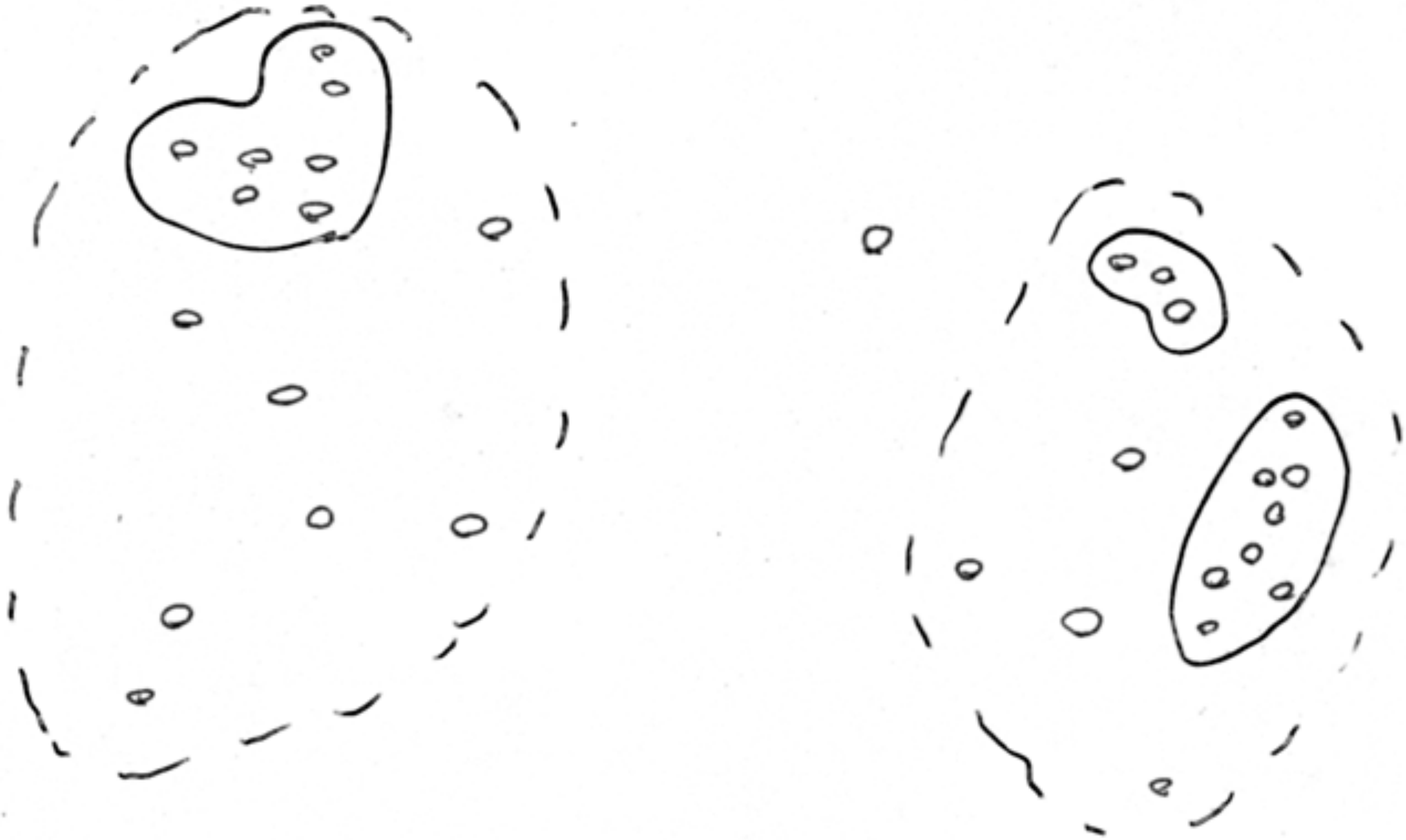
Represent predictors in abstract high dimensional space



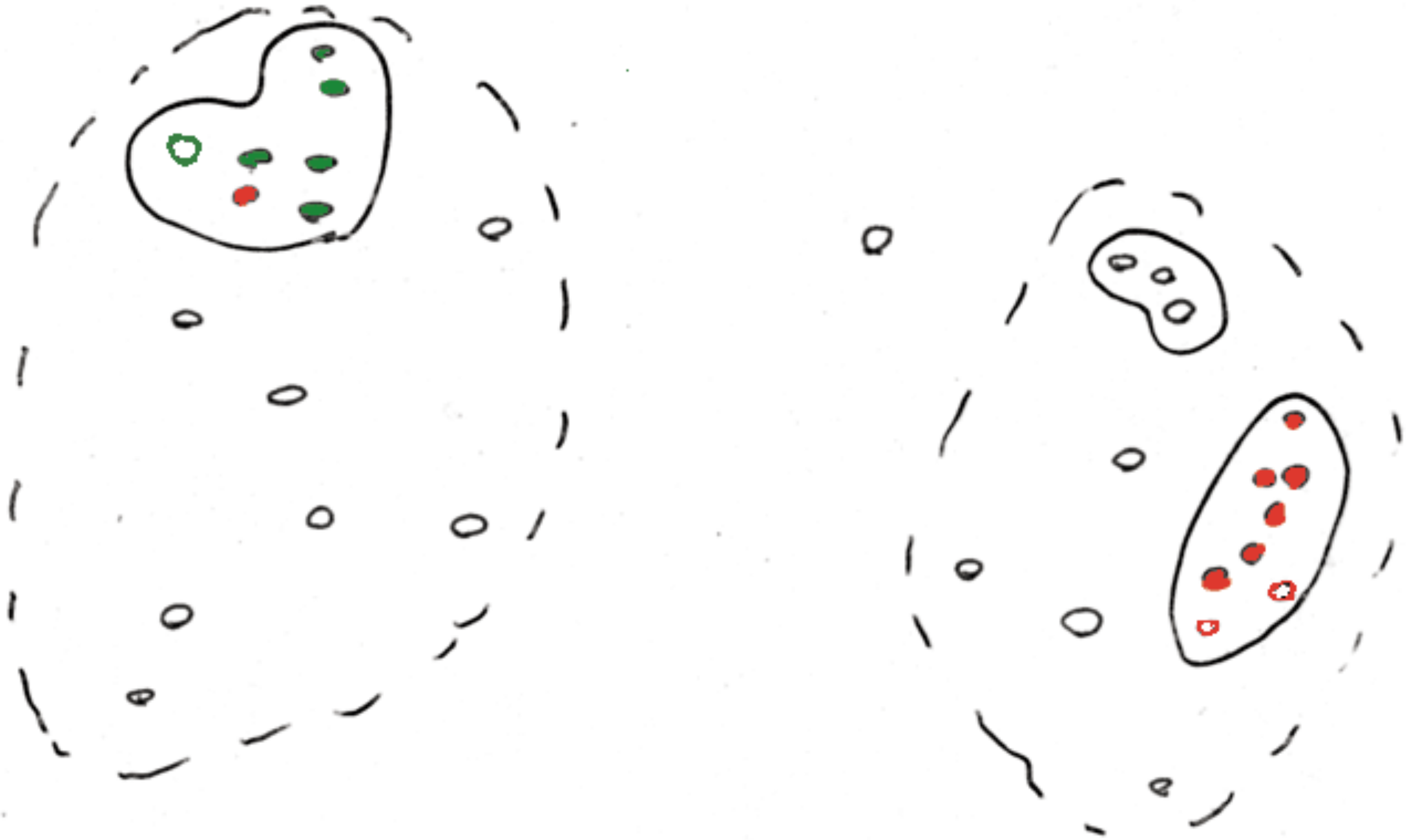
“Label” Certain Points



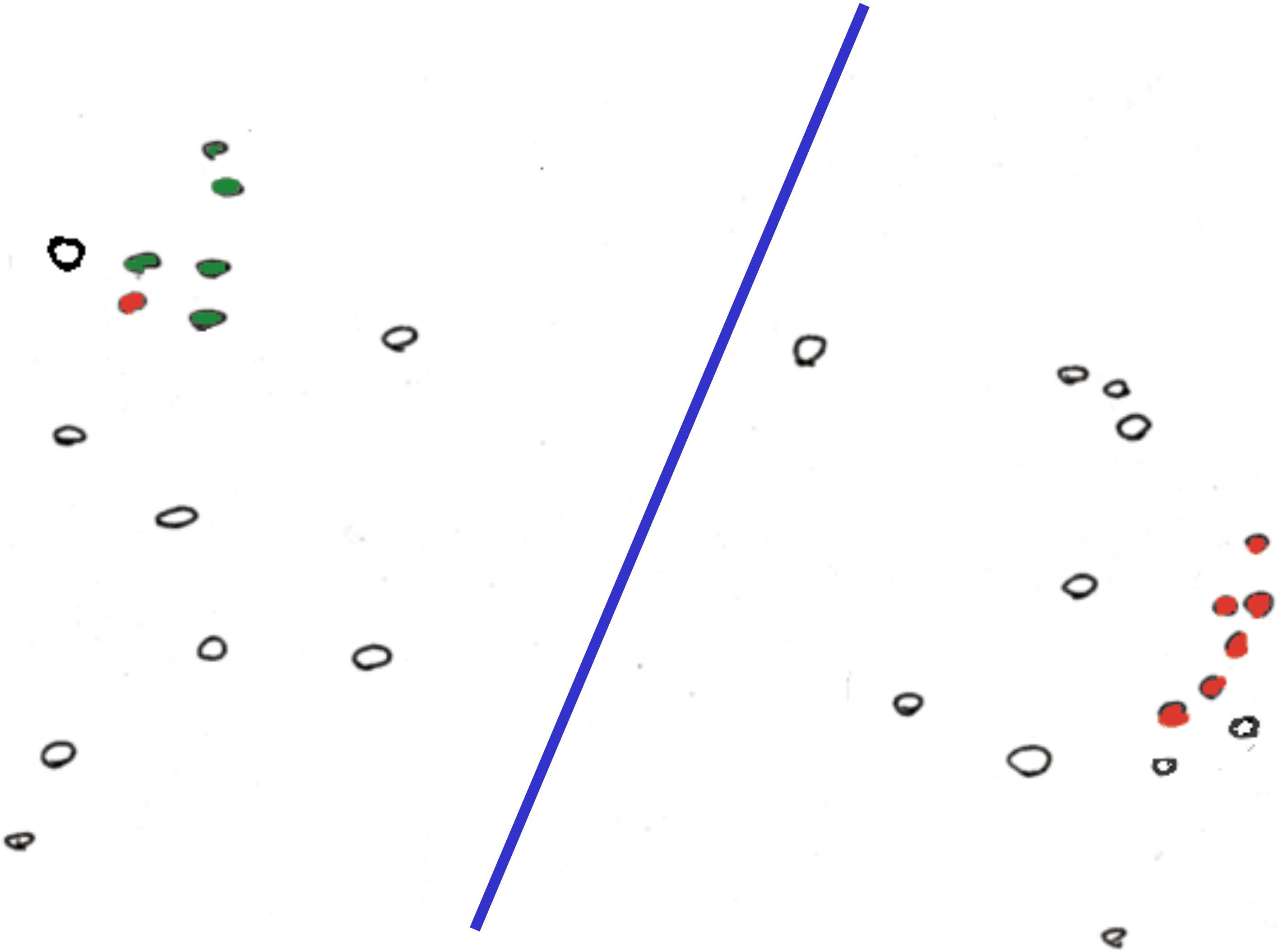
“Cluster” predictors (Unsupervised)



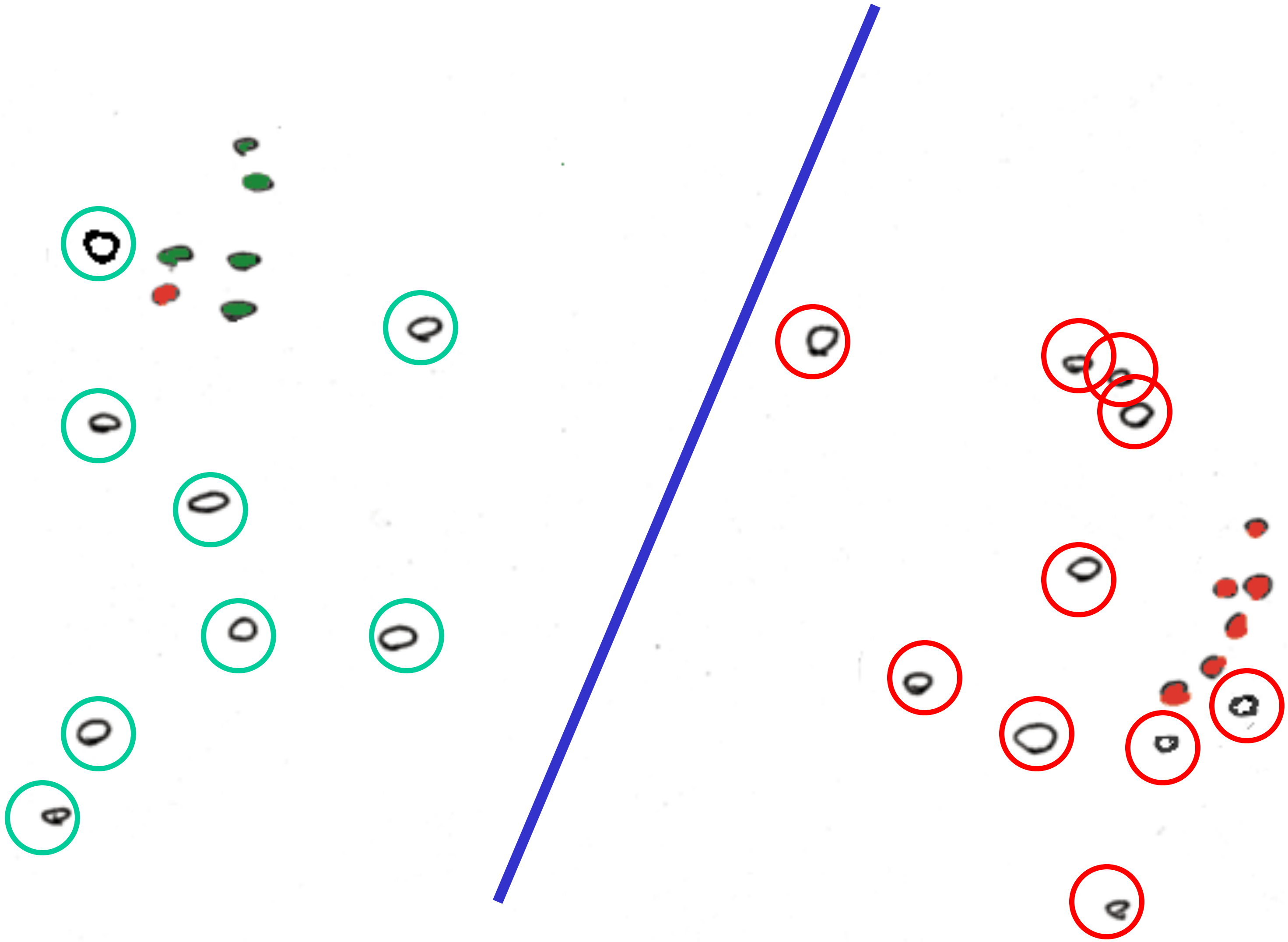
Use Clusters to predict Response (Unsupervised, guilt-by-association)



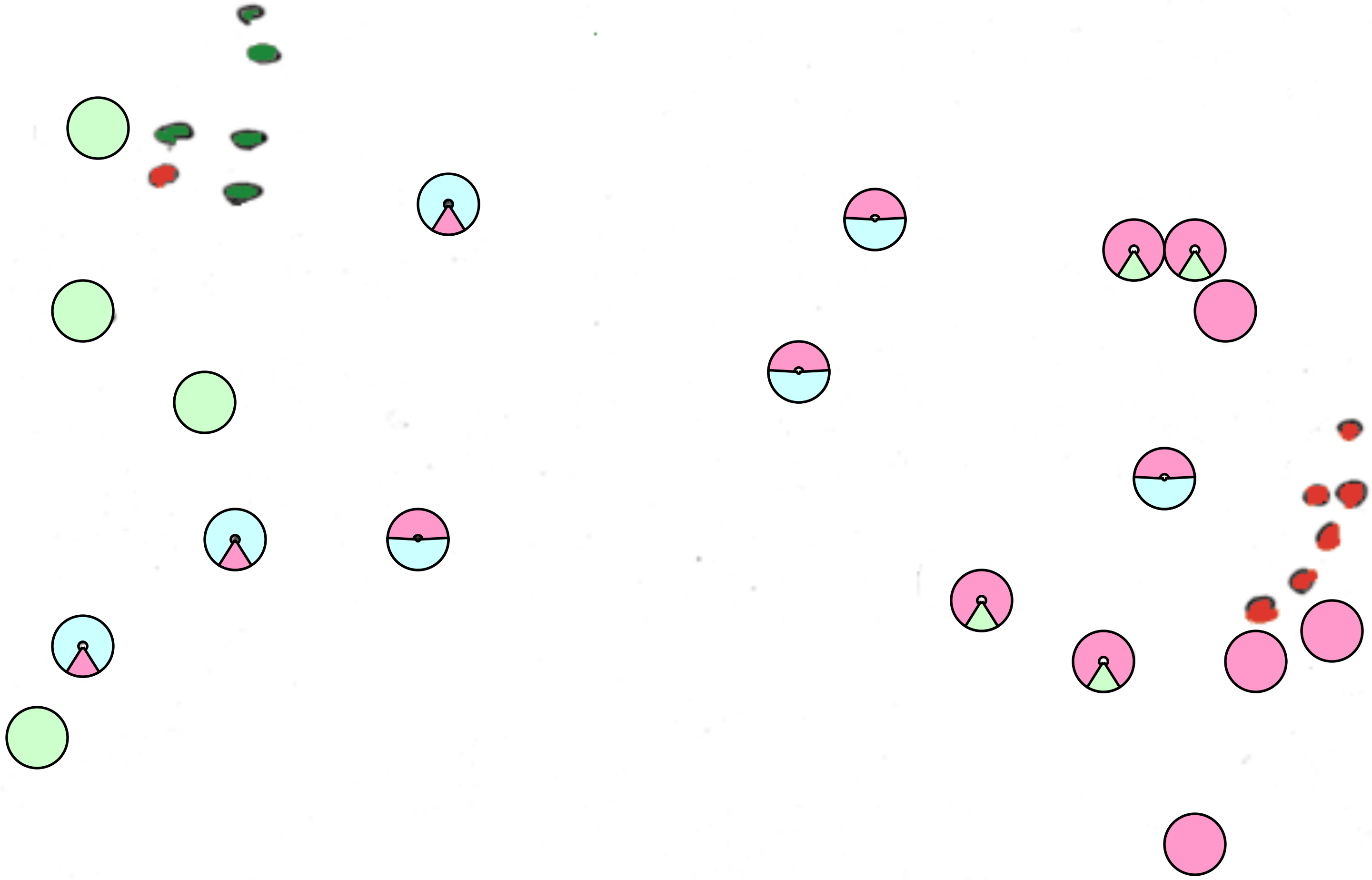
Find a Division to Separate Tagged Points



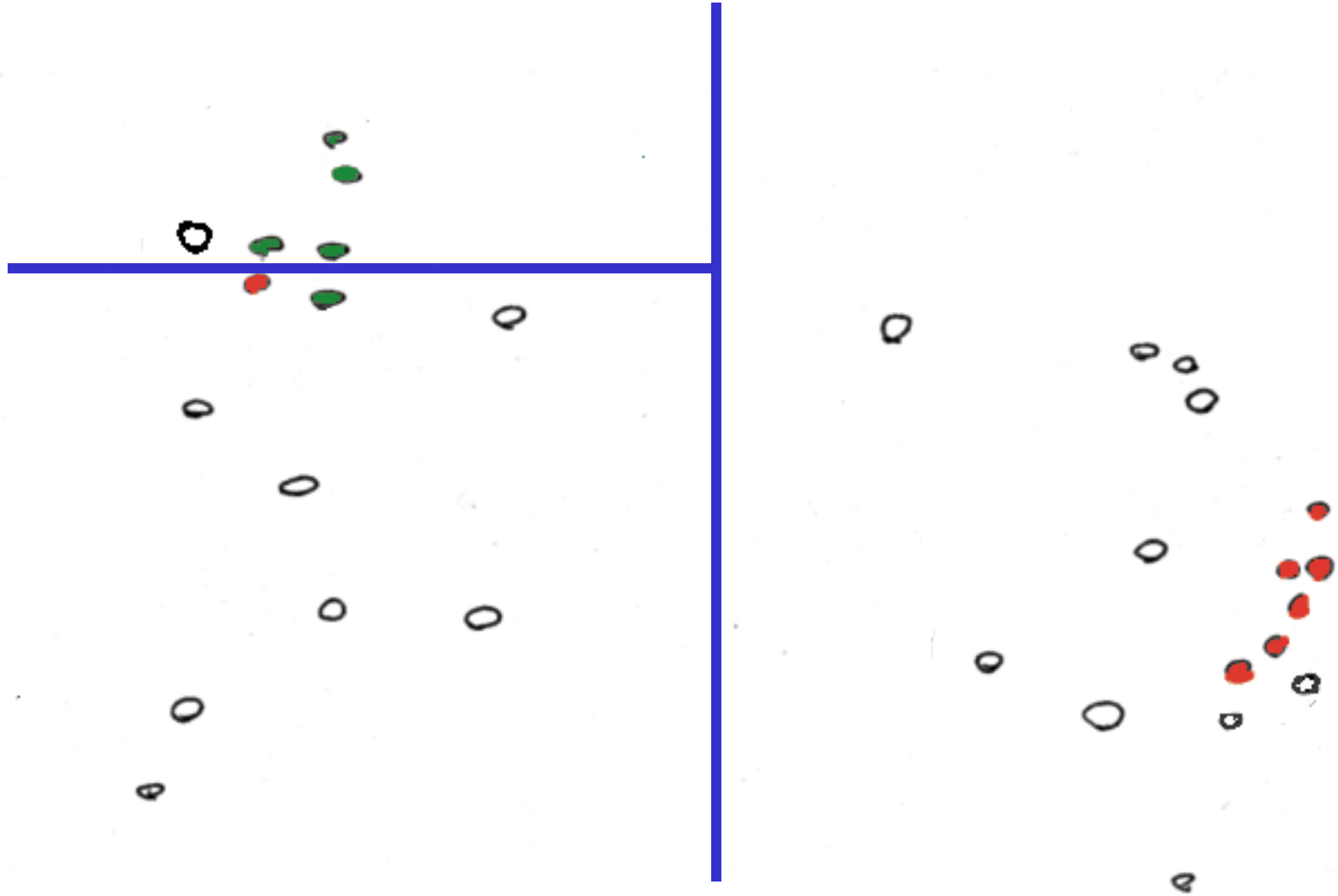
Extrapolate to Untagged Points



Probabilistic Predictions of Class



Find a Division to Separate Tagged Points



Distinctions in Supervised Learning

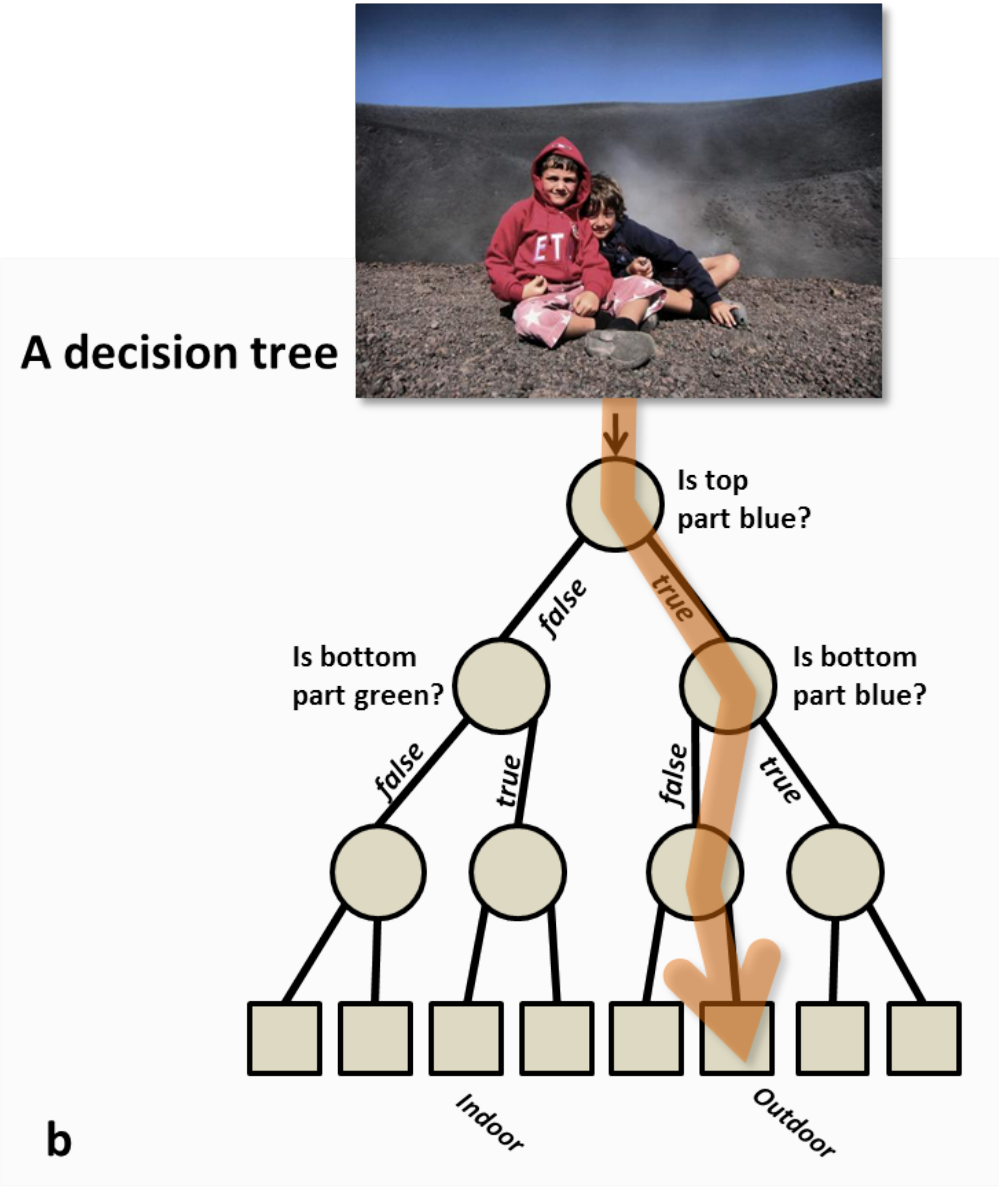
- Regression vs Classification
  - Regression: labels are quantitative
  - Classification: labels are categorical
- Regularized vs Un-regularized
  - Regularized: penalize model complexity to avoid over-fitting
  - Un-regularized: no penalty on model complexity
- Parametric vs Non-parametric
  - Parametric: 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

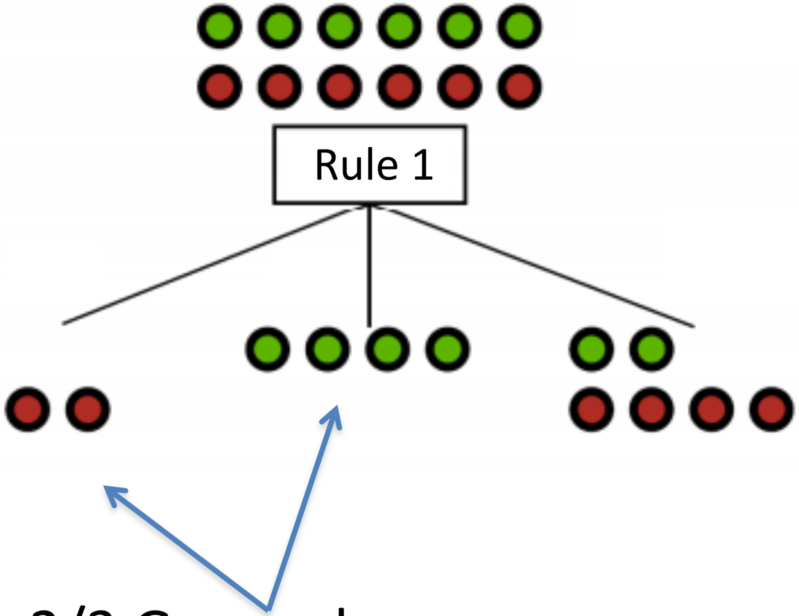


- Example: Is a picture taken inside or outside?

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

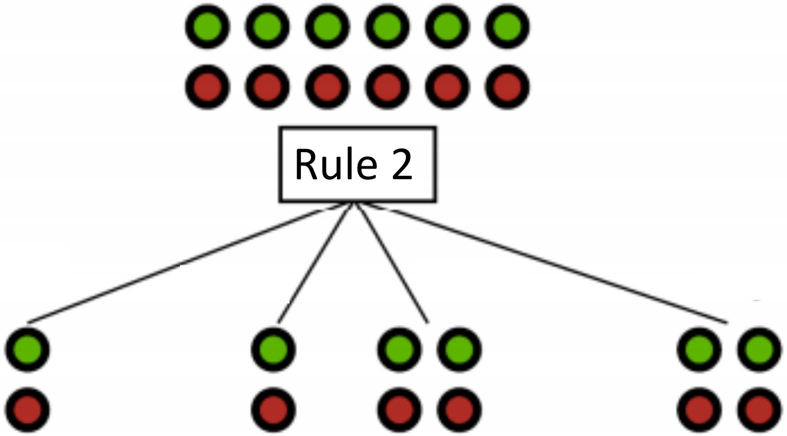
What makes a good rule?

- 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

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

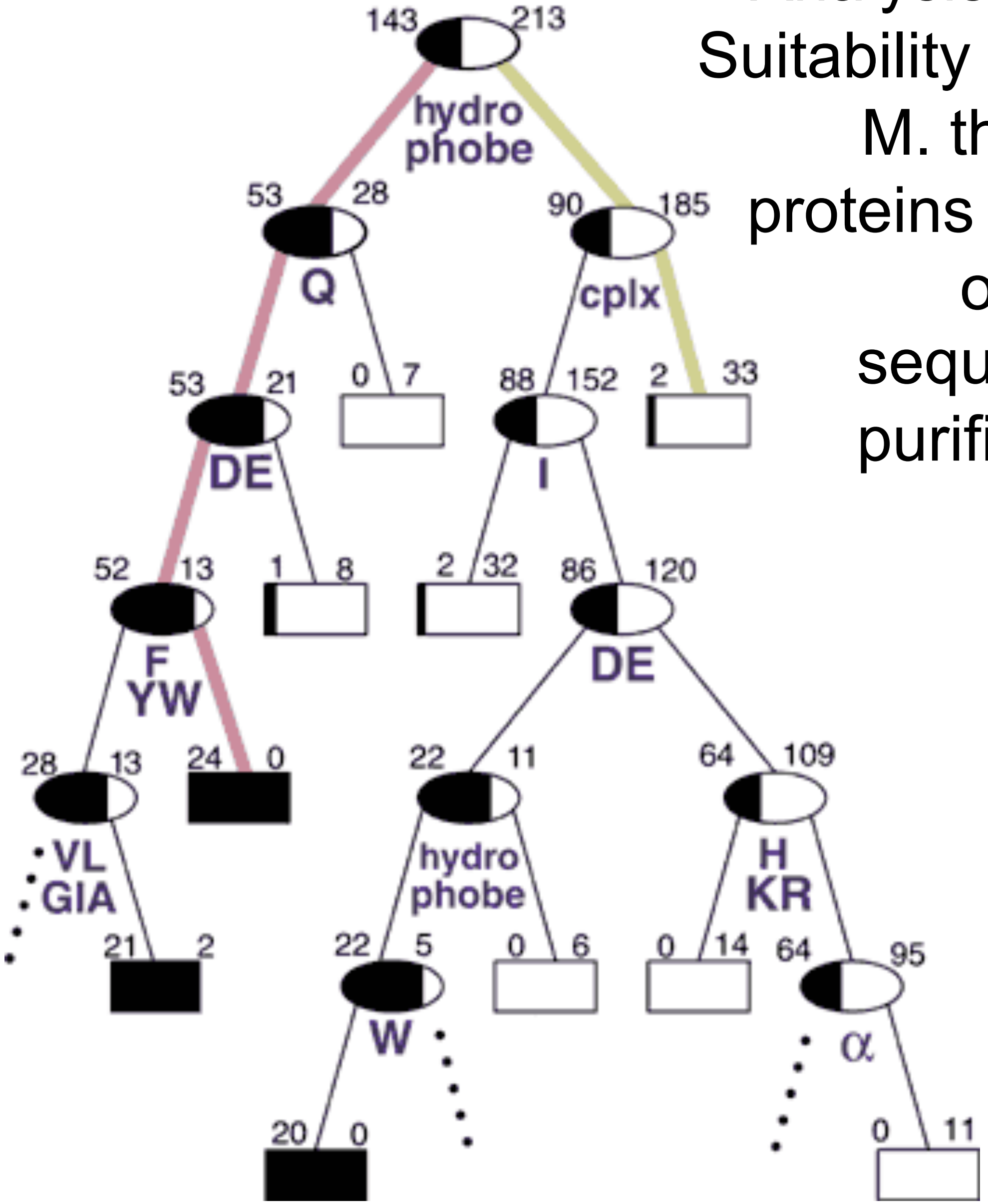
  - Use frequency of classifier characteristic within group as probability
  - Minimize entropy to achieve homogenous group

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



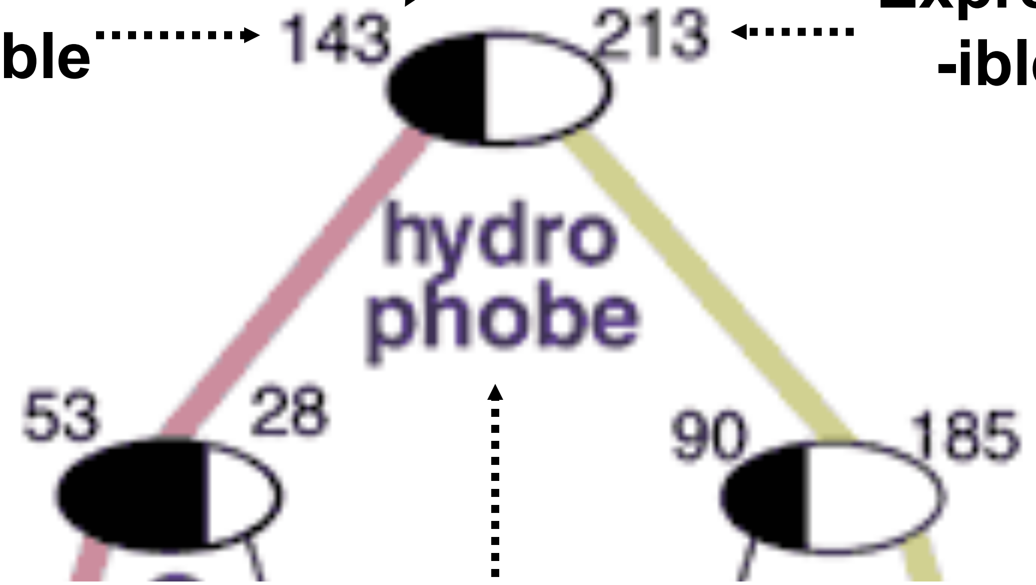
[Bertone et al. NAR ('01)]

Retrospective Decision Trees Nomenclature

356 total

Not Expressible

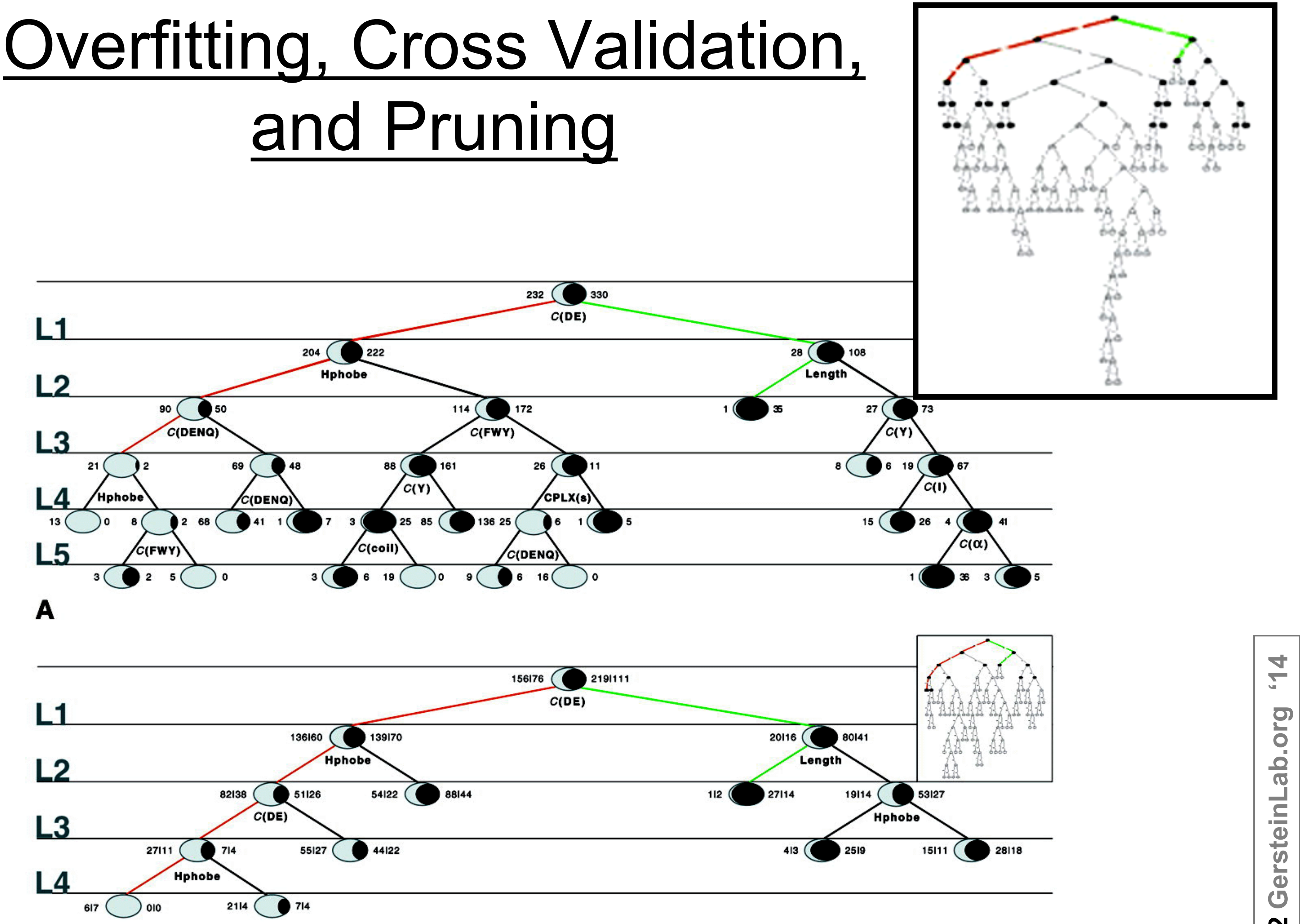
Express -ible



Has a hydrophobic stretch? (Y/N)

[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.

Supervised Mining:

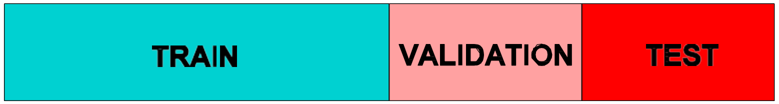
Assessment, Cross-Validation & ROC Curves

Evaluating performance: What? How?

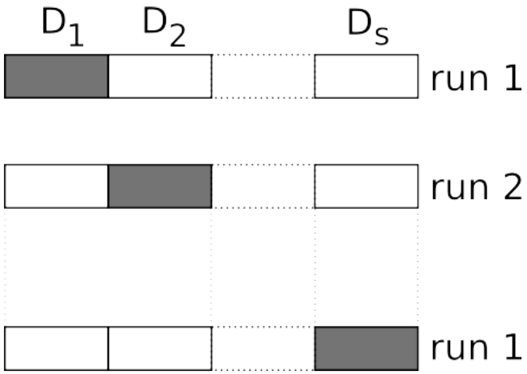
A. <u>What</u> do we want to evaluate?

GENERALIZATION

Therefore it is mandatory to divide your dataset:



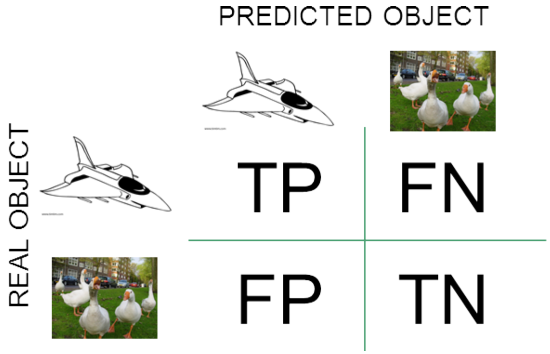
Alternatively, use Cross Validation:



Slide from Alberto Paccanaro

B. <u>How</u> do we evaluate performance?

1. Classification problems



Accuracy

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

Sensitivity (or TPR)

TP/P=TP/(TP+FN)

Specificity

TN/N=TN/(TN+FP)

True positive rate

TP/(TP+FP)

False positive rate

FP/N=FP/(FP+TN)

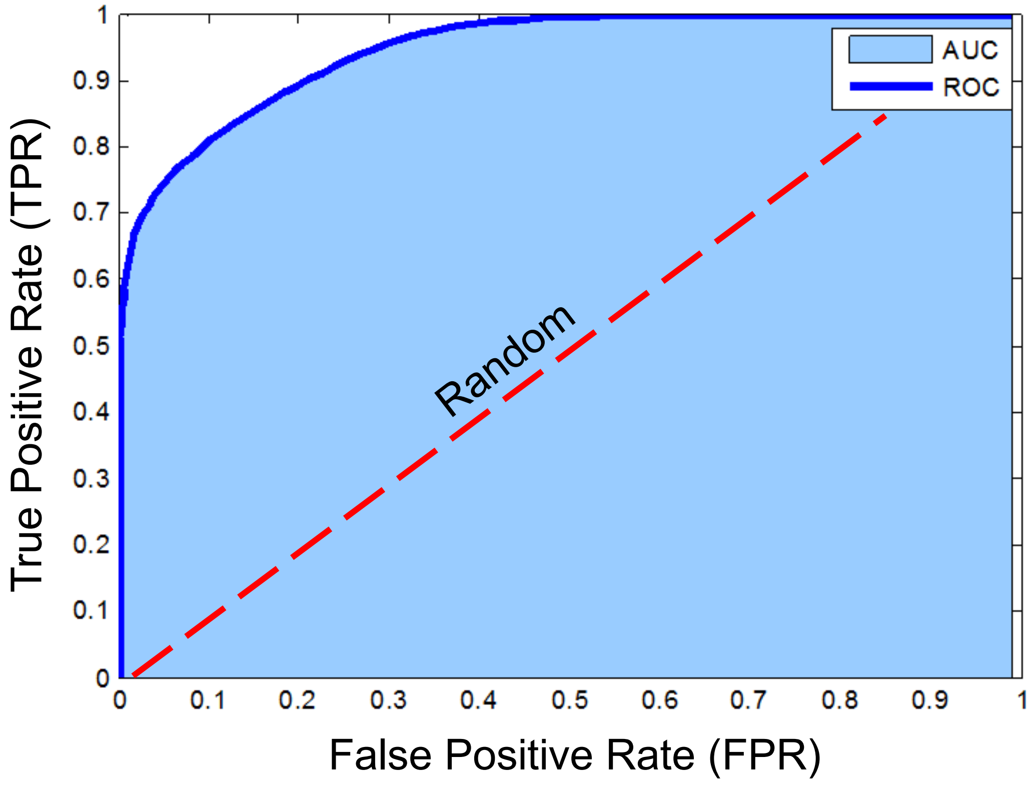
ROC analysis is good for comparing binary classifiers

2. Regression problems Sum of squares error

Root Mean Square error

Slide from Alberto Paccanaro

Intuition :
ROC Curve

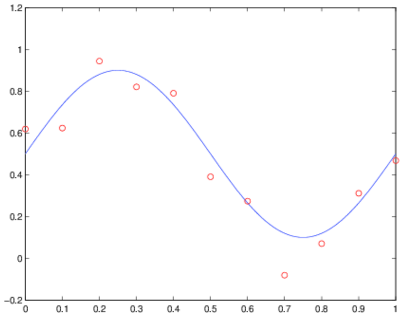


$$TPR = TP / P = TP / (TP + FN)$$

$$FPR = FP / N = FP / (FP + TN)$$

[From Biometrical Fusion - input statistical distribution]

Model dimensionality and overfitting

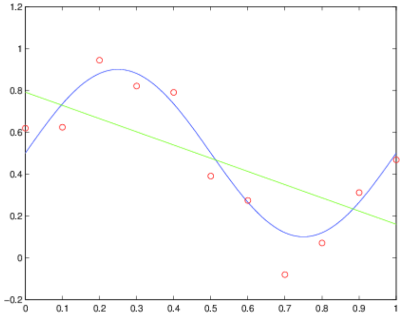


We are given the red dots.

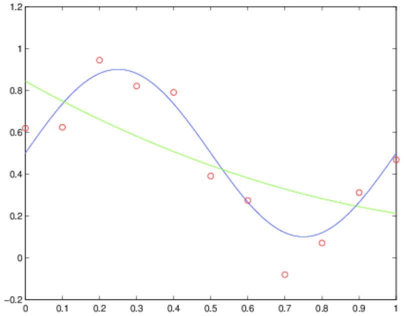
We assume that they are noisy samples from a signal/(function) – the blue curve – which we do not have (we only have the red dots).

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

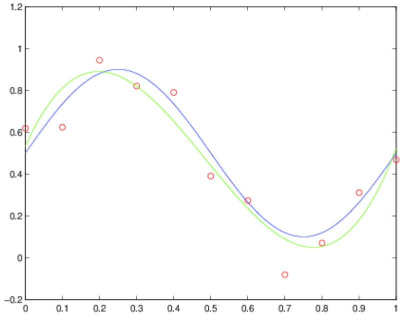
Our model needs to approximate the blue function. We decide to do it with polynomials.



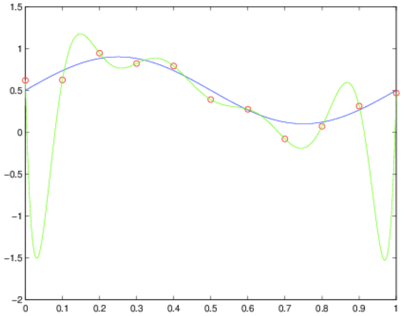
Degree 1 polynomial



Degree 2 polynomial



Degree 3 polynomial

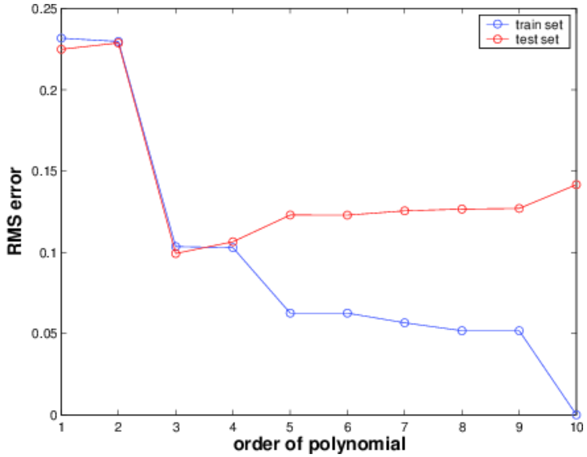


Degree 10 polynomial

Which one is best? And why?

Slide from Alberto Paccanaro

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

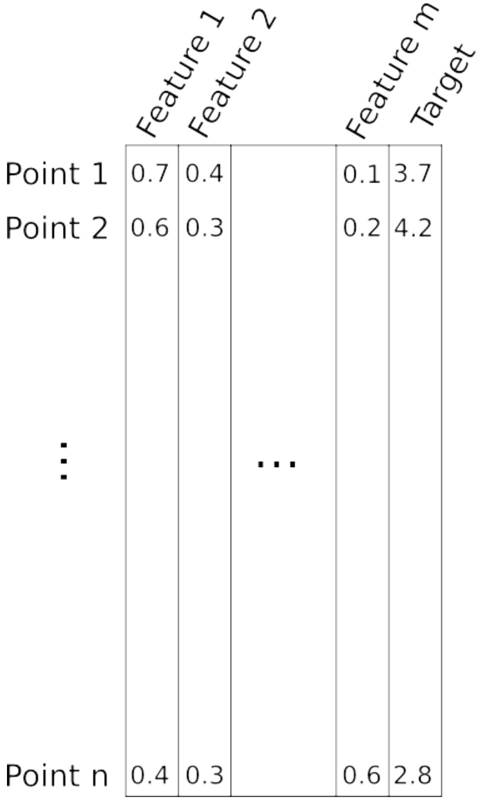


- **Occam's razor** *(William of Occam, ~1300):* Accept the simplest explanation that fits the data.

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.

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?

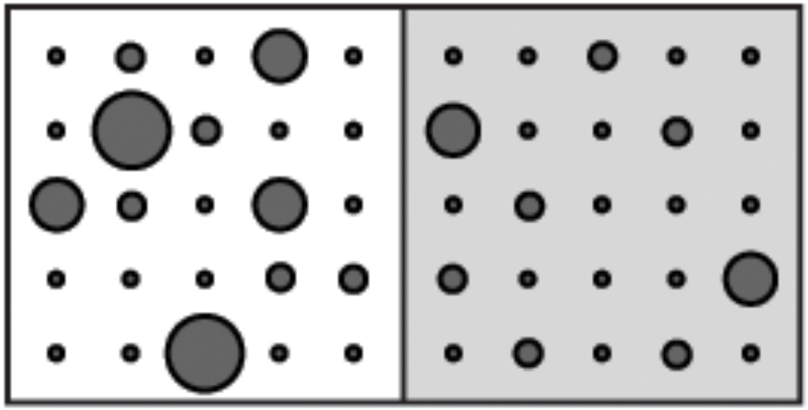


	Feature 1	Feature 2		Feature m	Target
Point 1	0.7	0.4		0.1	3.7
Point 2	0.6	0.3		0.2	4.2
⋮			…		
Point n	0.4	0.3		0.6	2.8

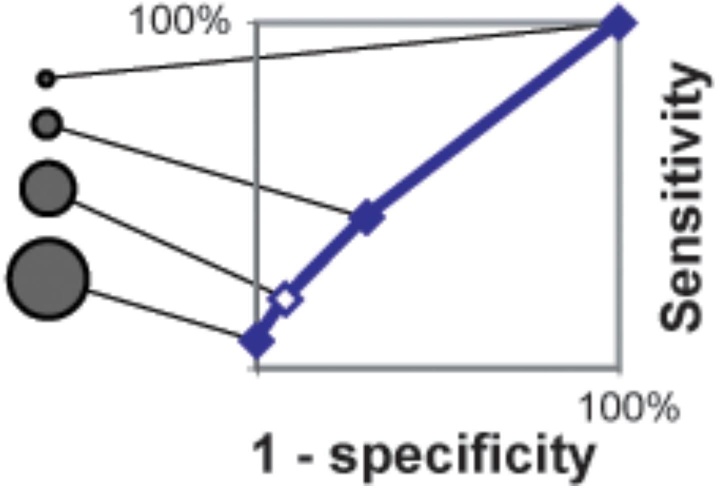
This is related to the "Curse of Dimensionality" Bellman, 1961.

Slide from Alberto Paccanaro

Comparison of Predictions against a Positive and Negative Gold Standard



Threshold "predictions" at different levels and compare to + and - gold standards



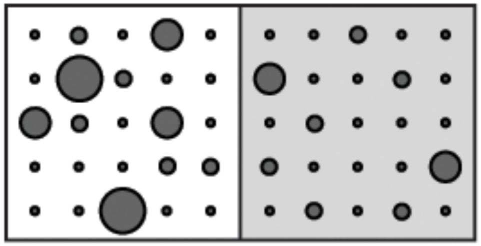
"Coverage"

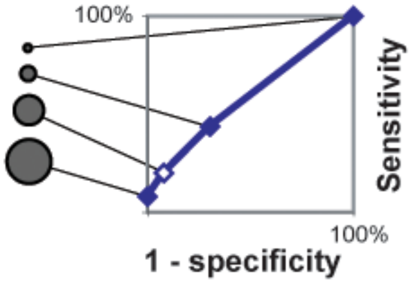
"Error Rate"

ROC plot
(cross validated)

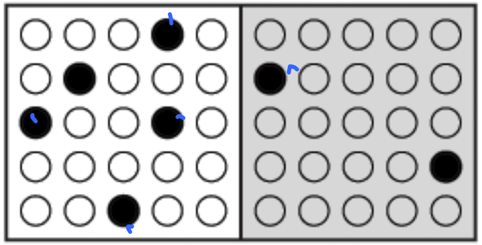
R Jansen, M Gerstein (2004). *Curr Opin Microbiol 7*: 535

Effect on Predictions of Large Number of Negatives

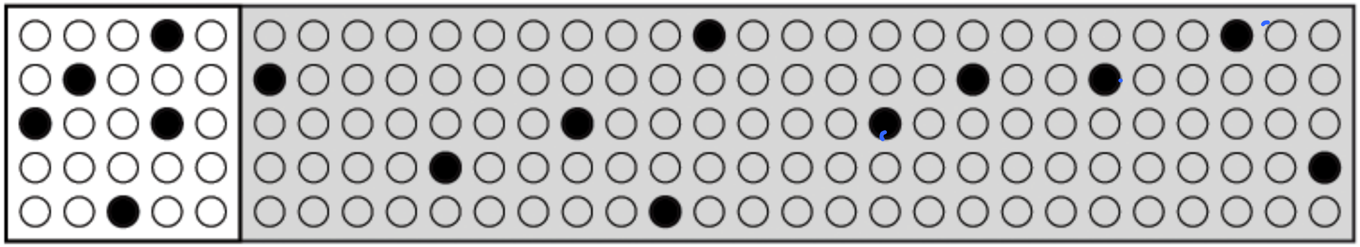




Sensitivity **1- specificity** **Positive predictive value**



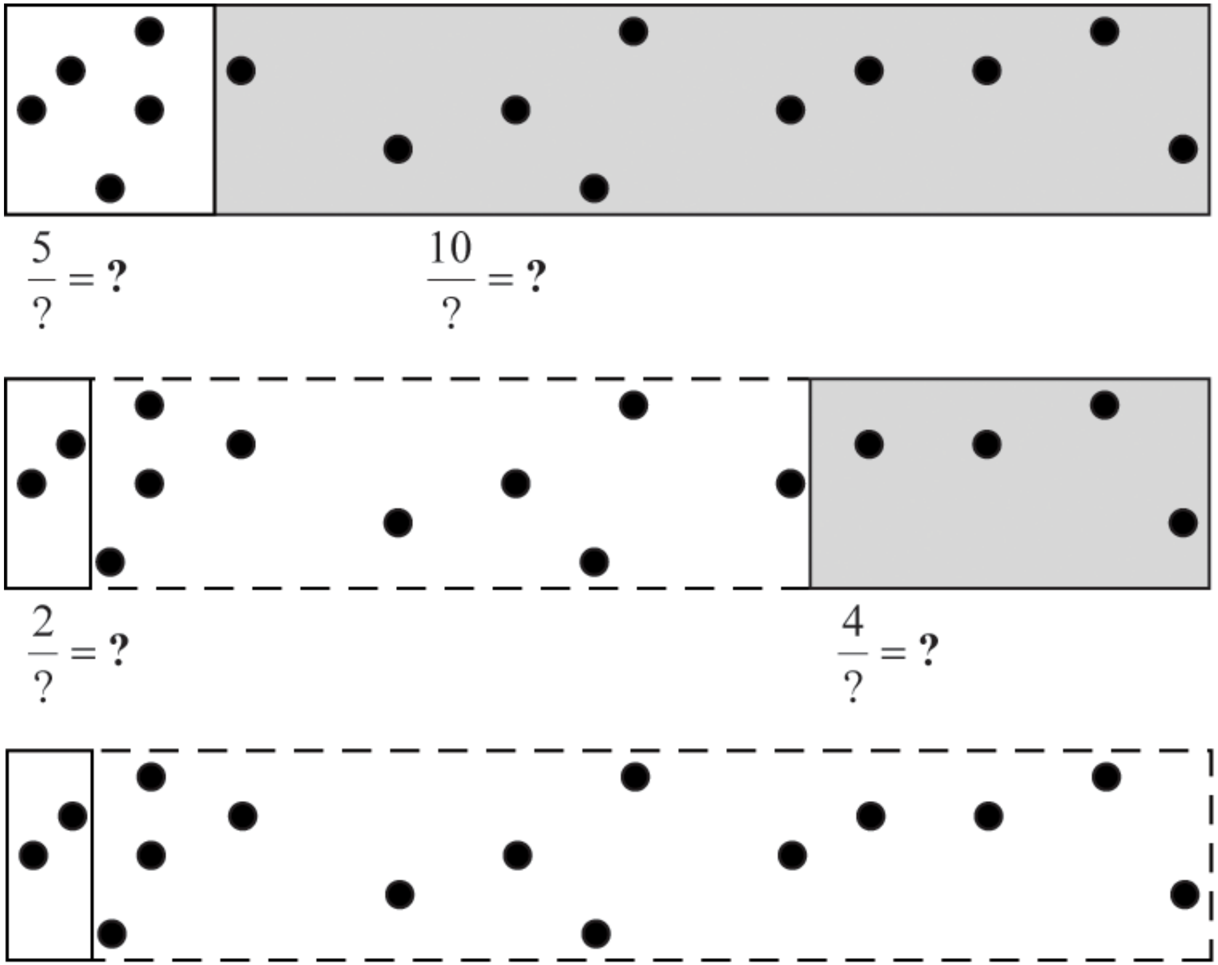
$$\frac{5}{25} = \mathbf{20\%} \qquad \frac{2}{25} = \mathbf{8\%} \qquad \frac{5}{5+2} \approx \mathbf{71\%}$$



$$\frac{5}{25} = \mathbf{20\%} \qquad \frac{10}{125} = \mathbf{8\%} \qquad \frac{5}{5+10} \approx \mathbf{33\%}$$

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

Importance of Balanced Positive and Negative Examples



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

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

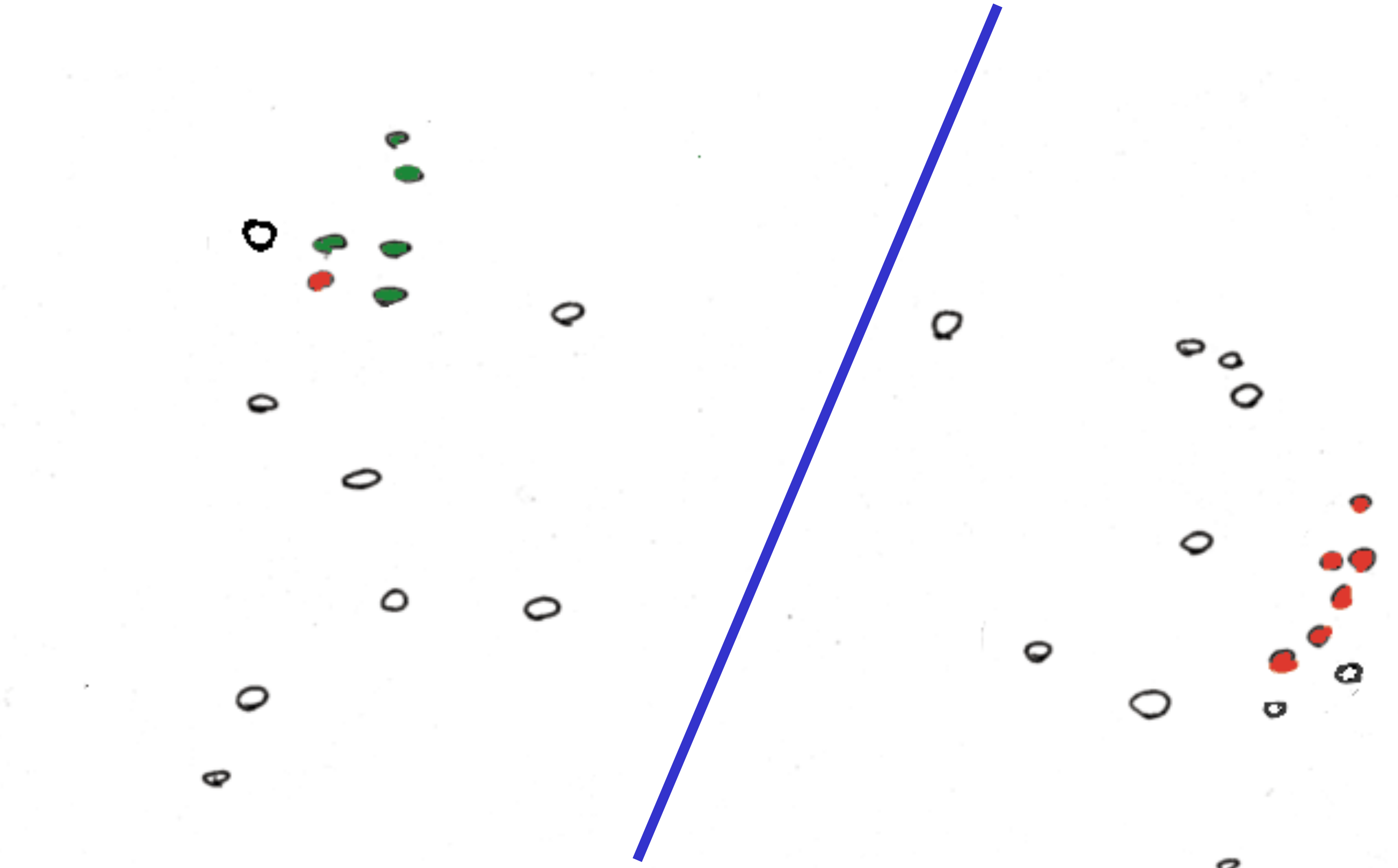
$\frac{2}{?} = \mathbf{?}$ $\quad \frac{?}{?} = \mathbf{?}$ $\quad \frac{2}{2+?} = \mathbf{?}$

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

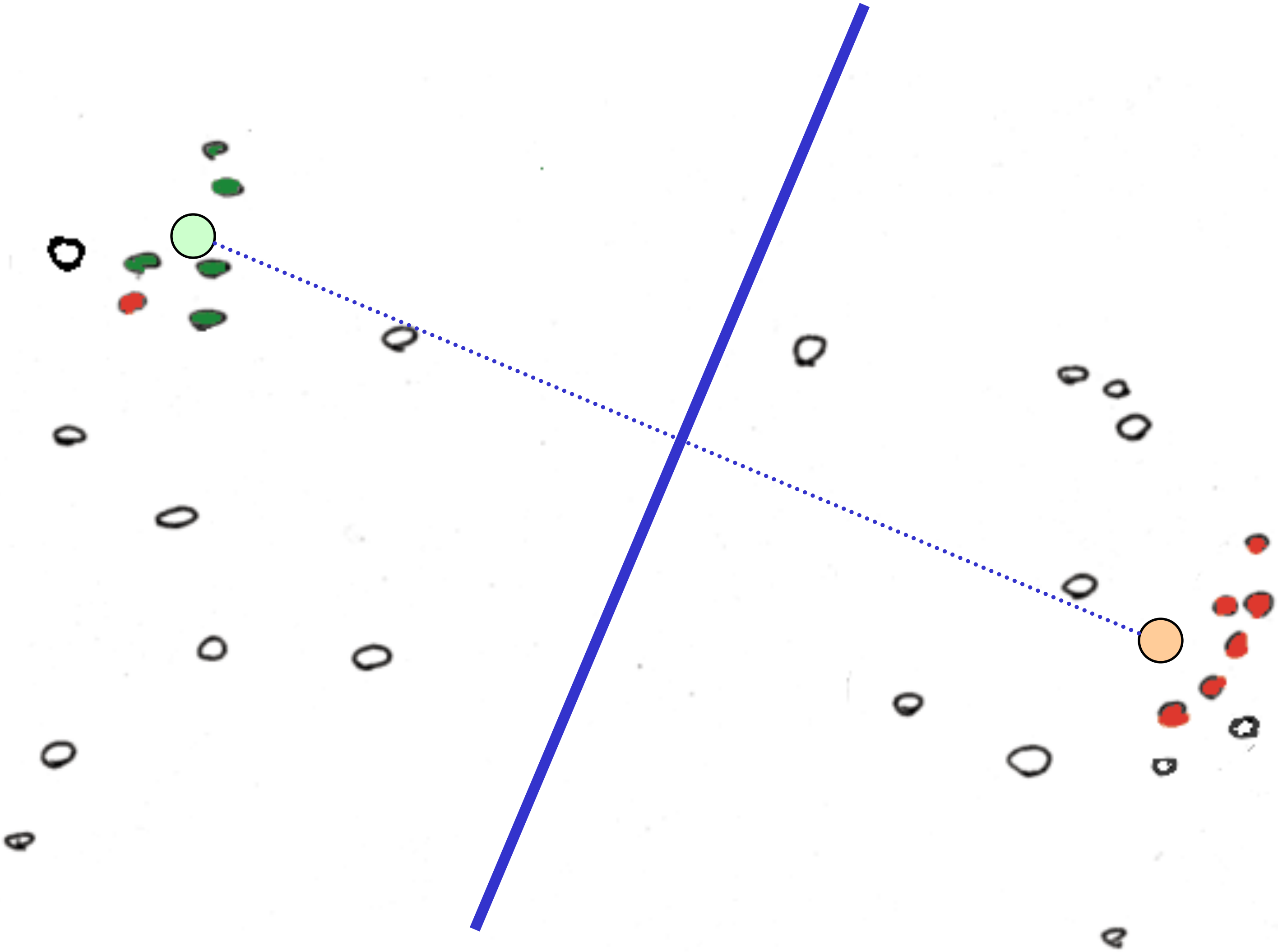
Supervised Mining:

SVMs

Find a Division to Separate Tagged Points



Discriminant to Position Plane



Fisher discriminant analysis

- Use the training set to reveal the structure of class distribution by seeking a linear combination
- $y = w_1x_1 + w_2x_2 + ... + w_nx_n$ which maximizes the ratio of the separation of the class means to the sum of each class variance (within class variance). This linear combination is called the first linear discriminant or first canonical variate. Classification of a future case is then determined by choosing the nearest class in the space of the first linear discriminant and significant subsequent discriminants, which maximally separate the class means and are constrained to be uncorrelated with previous ones.

$$s_i^2 = \sum_{y \in Y_i} (y - m_i)^2$$

$m_i = \vec{w} \cdot \vec{m}_i$

Solution of 1st variate

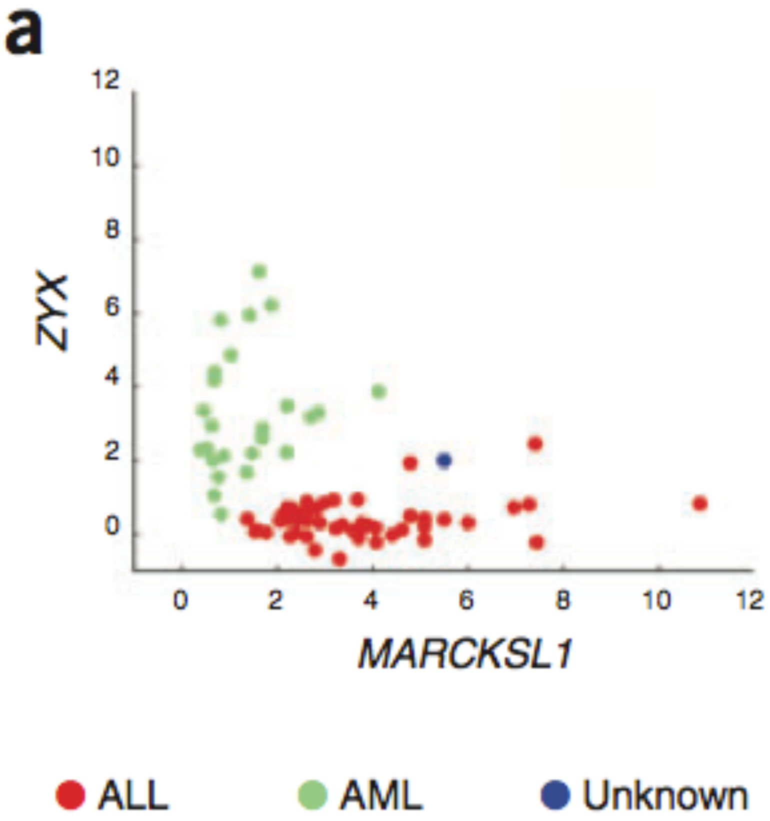
$$\vec{w} = S_W^{-1} (\vec{m}_1 - \vec{m}_2)$$

Support Vector Machines

- A very powerful tool for classifications
- Example Applications:
  - Text categorization
  - Image classification
  - Spam email recognition, etc

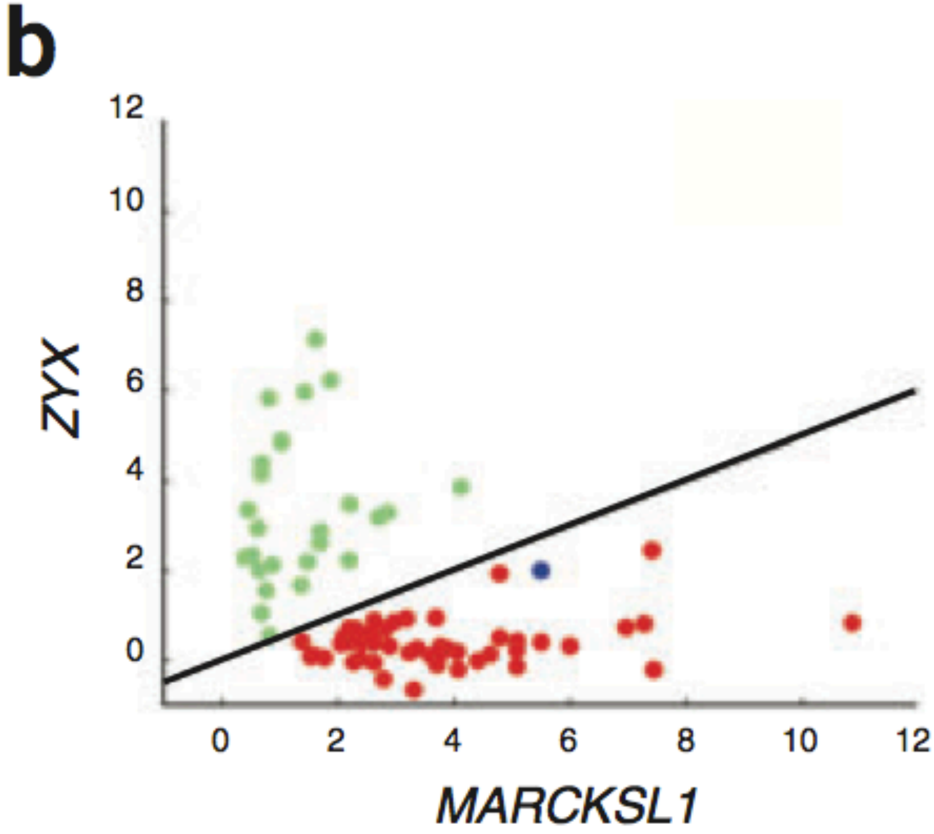
- It has also been successfully applied in many biological problems:
  - Disease diagnosis
  - Automatic genome functional annotation
  - Prediction of protein-protein interactions
  - and more…

- Example: Leukemia patient classification



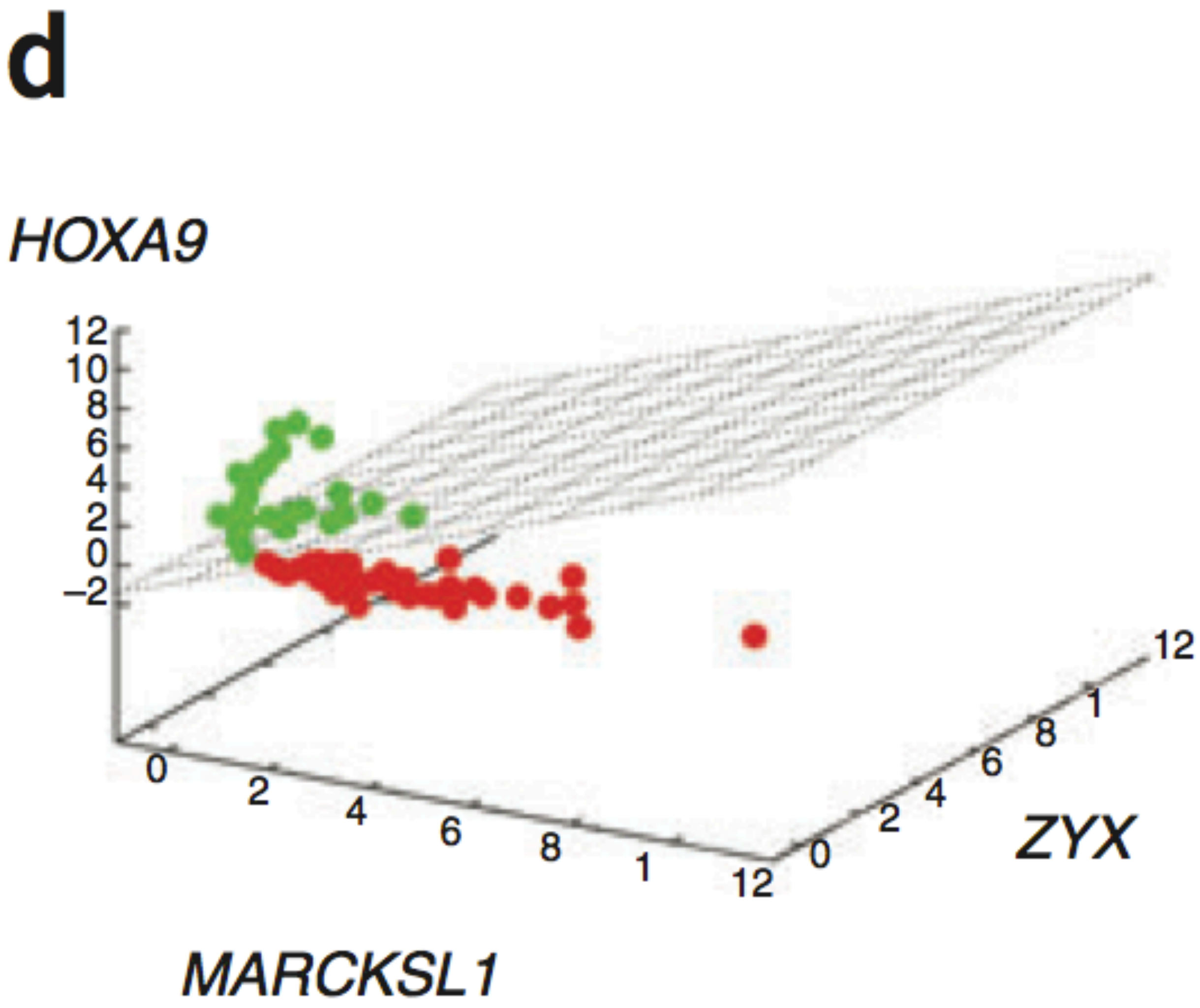
ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia

William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



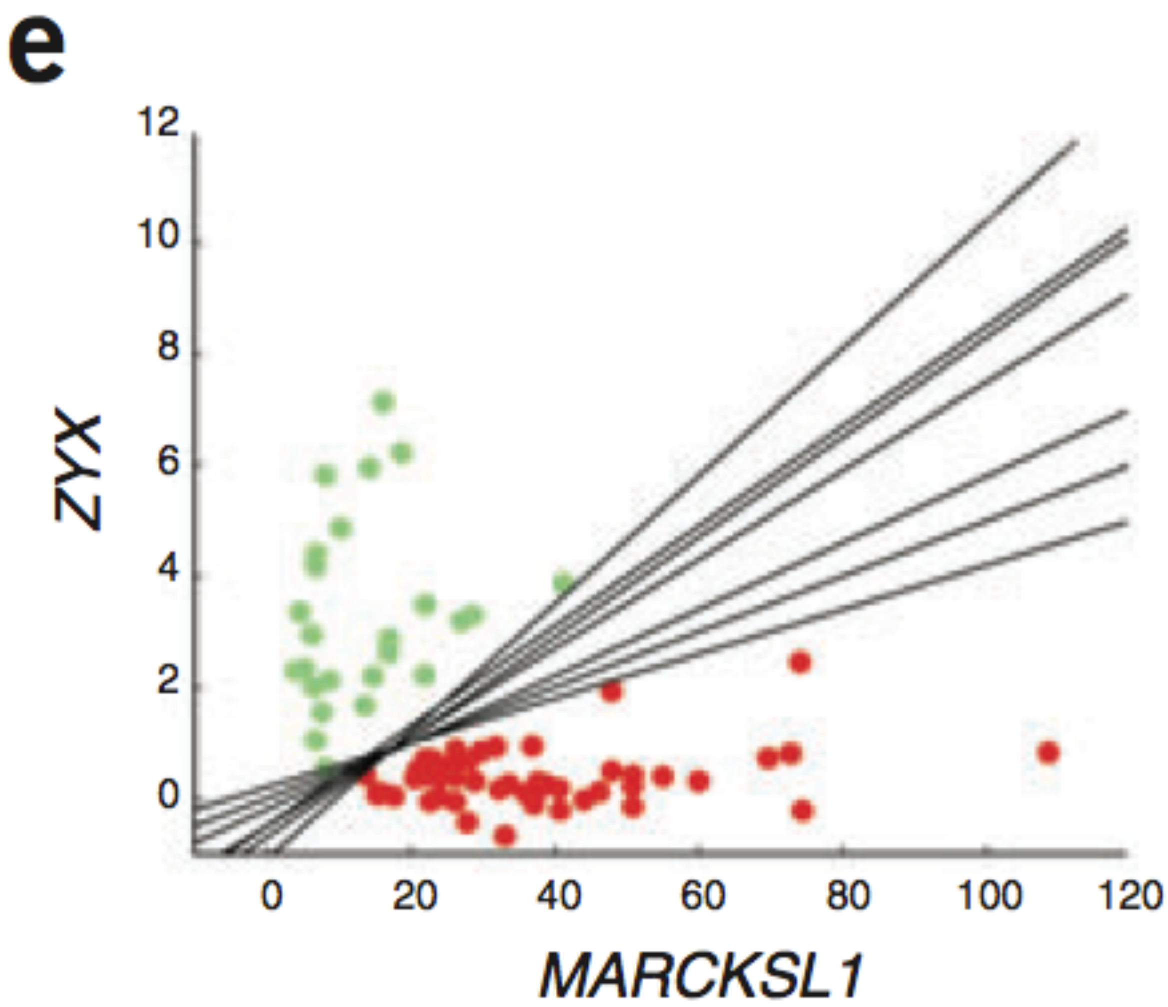
- A simple line suffices to separate the expression profiles of ALL and AML

William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



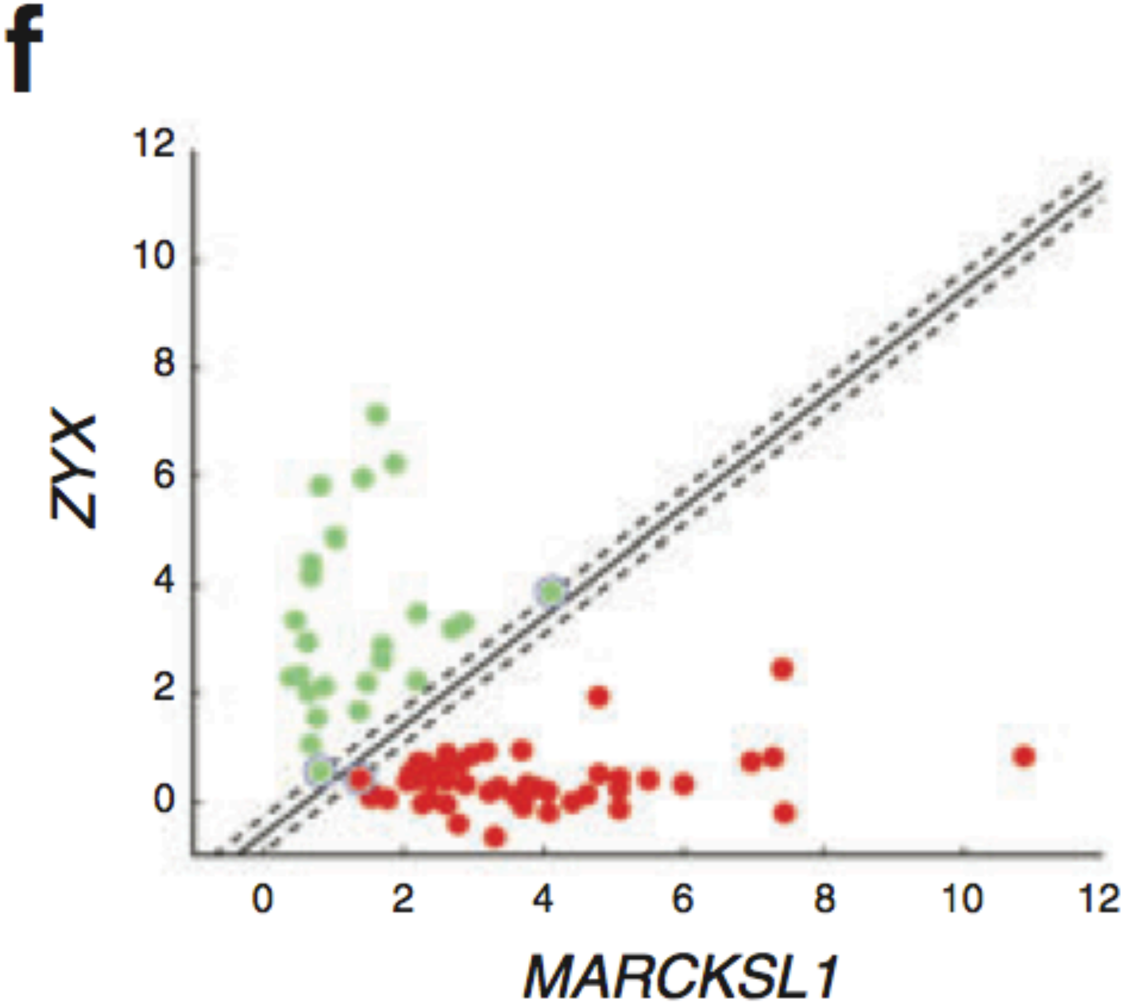
- In the case of more than two genes, a line generalizes to a plane or “hyperplane”.
- For generality, we refer to them all as “hyperplane”

William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



- Is there a “best” line?

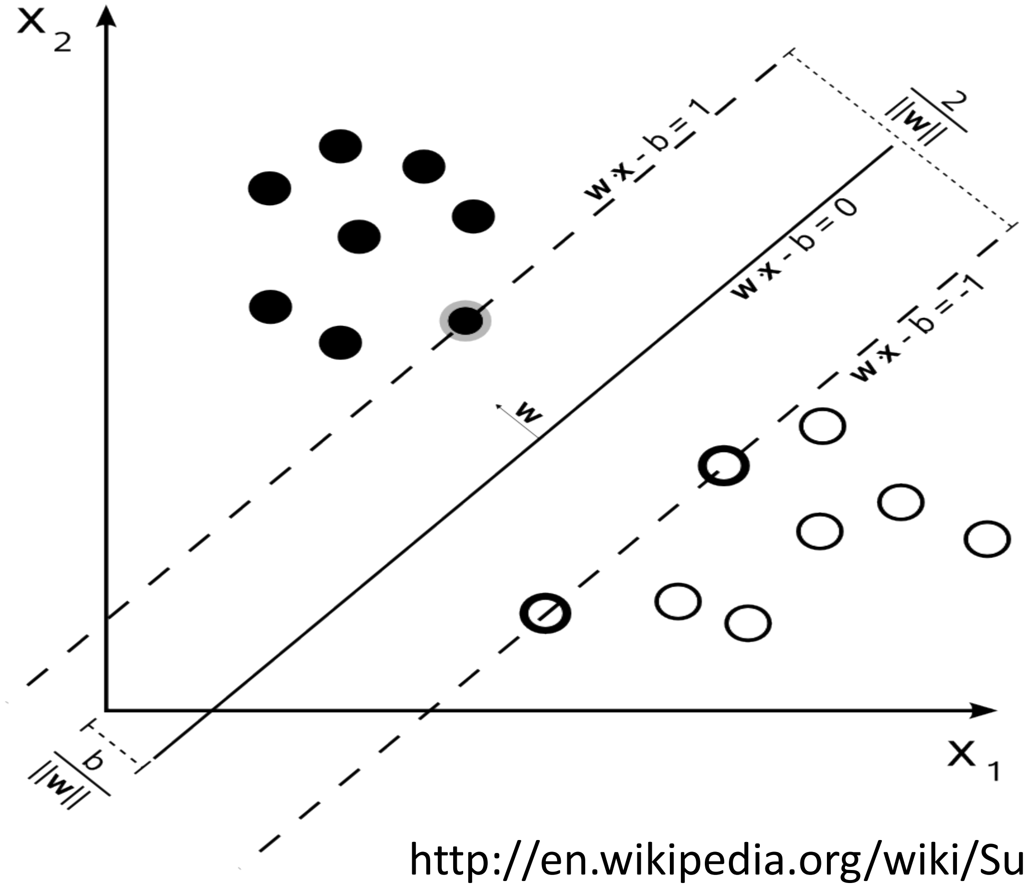
William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



- The maximum margin hyperplane

William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

- Denote each data point as (x_i, y_i)
- x_i is a vector of the expression profiles
- y_i = -1 or 1, which labels the class
- A hyperplane can be represented as: w*x + b = 0
- The margin-width equals to: $2/ \| w \|, \| w \| = \sqrt{w \bullet w}$



http://en.wikipedia.org/wiki/Support_vector_machine

- Find a hyperplane such that:
  - No data points fall between the lines $w \bullet x + b = -1$and $w \bullet x + b = +1$
  - The margin 2/||w|| is maximized

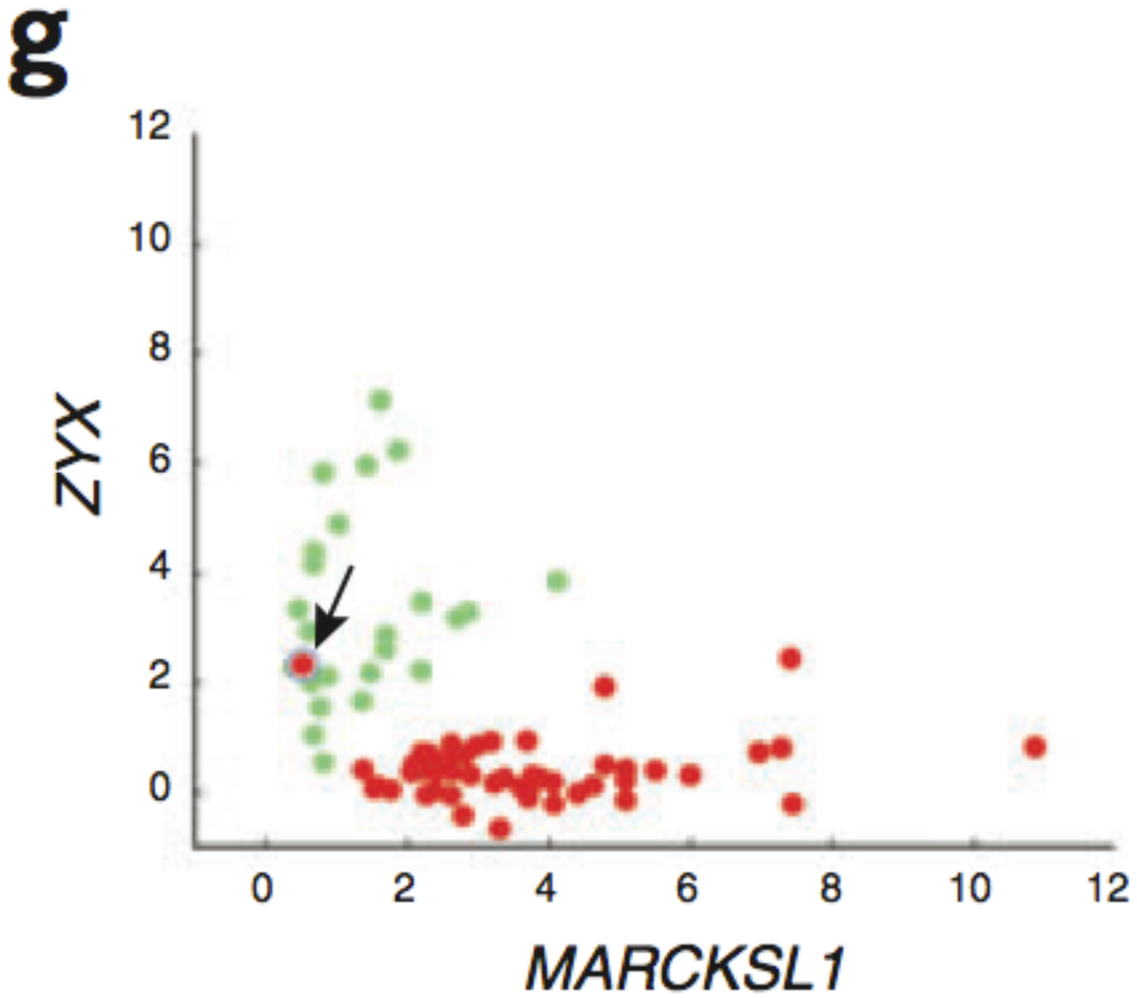
- Mathematically,
  - Minimize$_{w,b}$ 1/2 $||w||^2$, subject to:
  - for y_i = 1, $w \bullet x_i + b \geq 1$
  - for y_i = -1, $w \bullet x_i + b \leq -1$
  - Combining them, for any i, $y_i(w \bullet x_i + b) \geq 1$

- The solution expresses *w* as a linear combination of the x_i

- So far, we have been assuming that the data points from two classes are always easily linearly separable. But that's not always the case

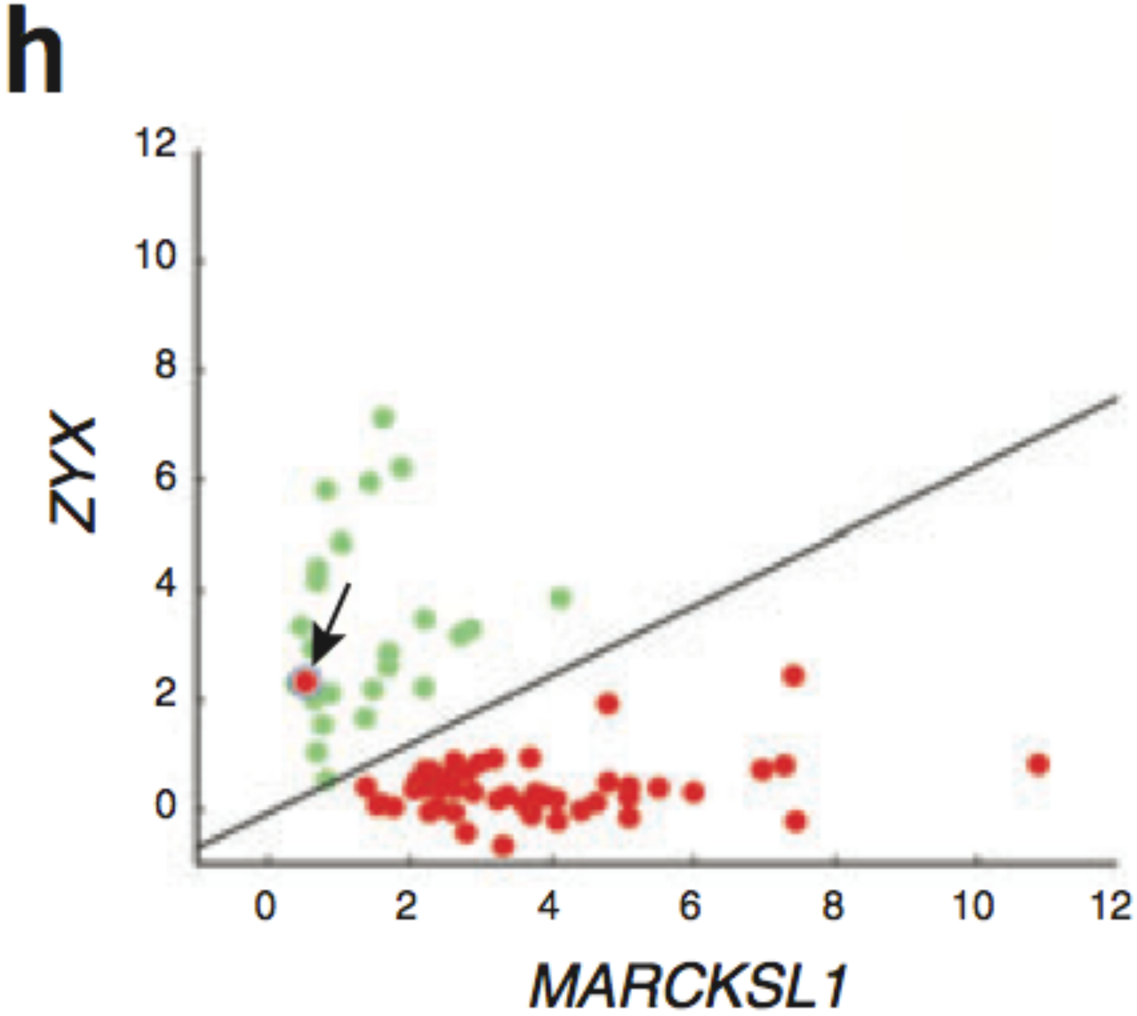
http://en.wikipedia.org/wiki/Support_vector_machine

- What if…



William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

- Allow a few anomalous data points

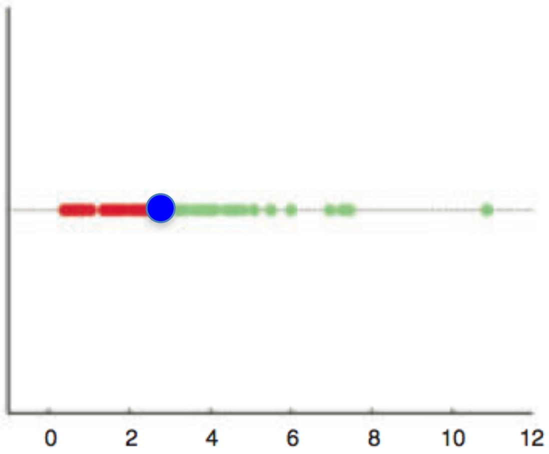


William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

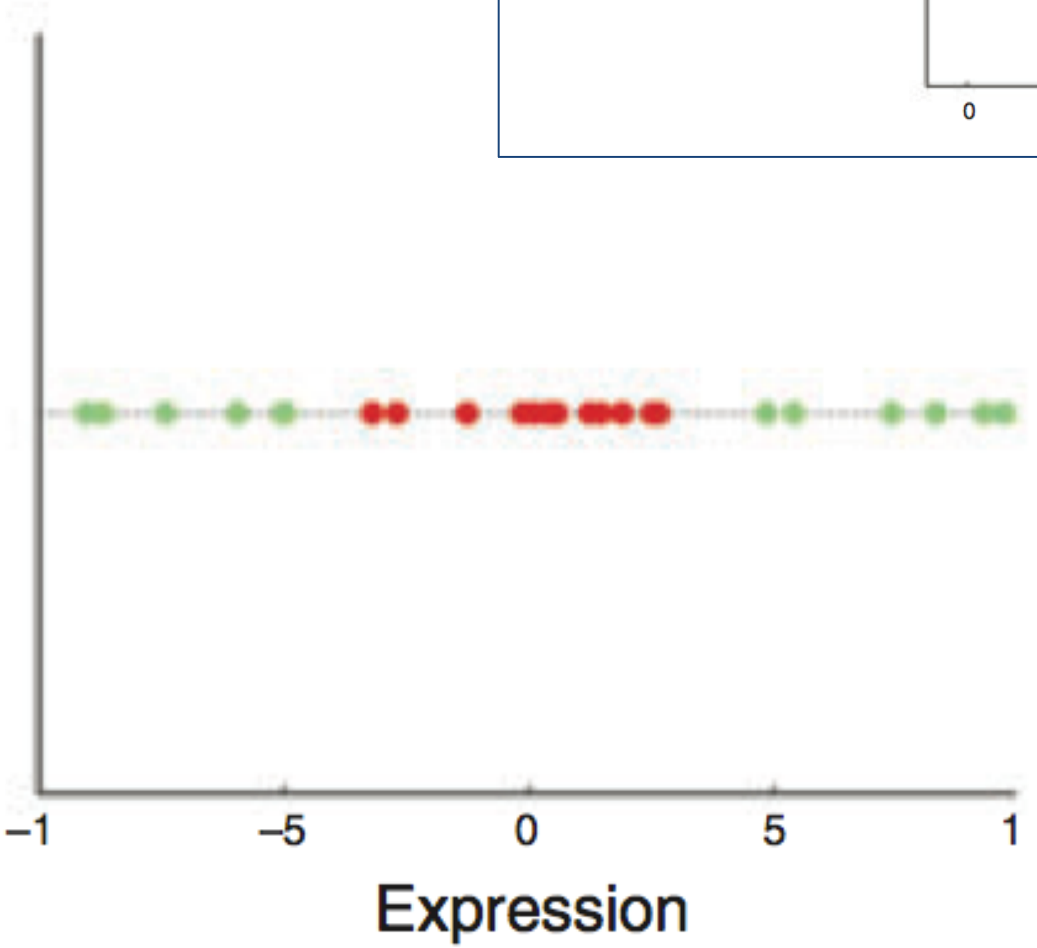
- The soft-margin SVM
  - $\underset{w,b,s}{\text{minimize}} \frac{1}{2} \| w \|^2 + C \sum_i s_i$
  - subject to, for any i, $y_i(w \bullet x_i + b) \geq 1 - s_i, s_i \geq 0$
  - S_i are the slack variables
  - C controls the number of tolerated misclassifications
    (It's effectively a regularization parameter on model complexity)
  - A small C would allow more misclassifications
  - A large C would discourage misclassifications
  - Note that even when the data points are linearly separable, one can still introduce the slack variables to pursue a larger separation margin

http://en.wikipedia.org/wiki/Support_vector_machine

- Are linear separating hyperplanes enough?



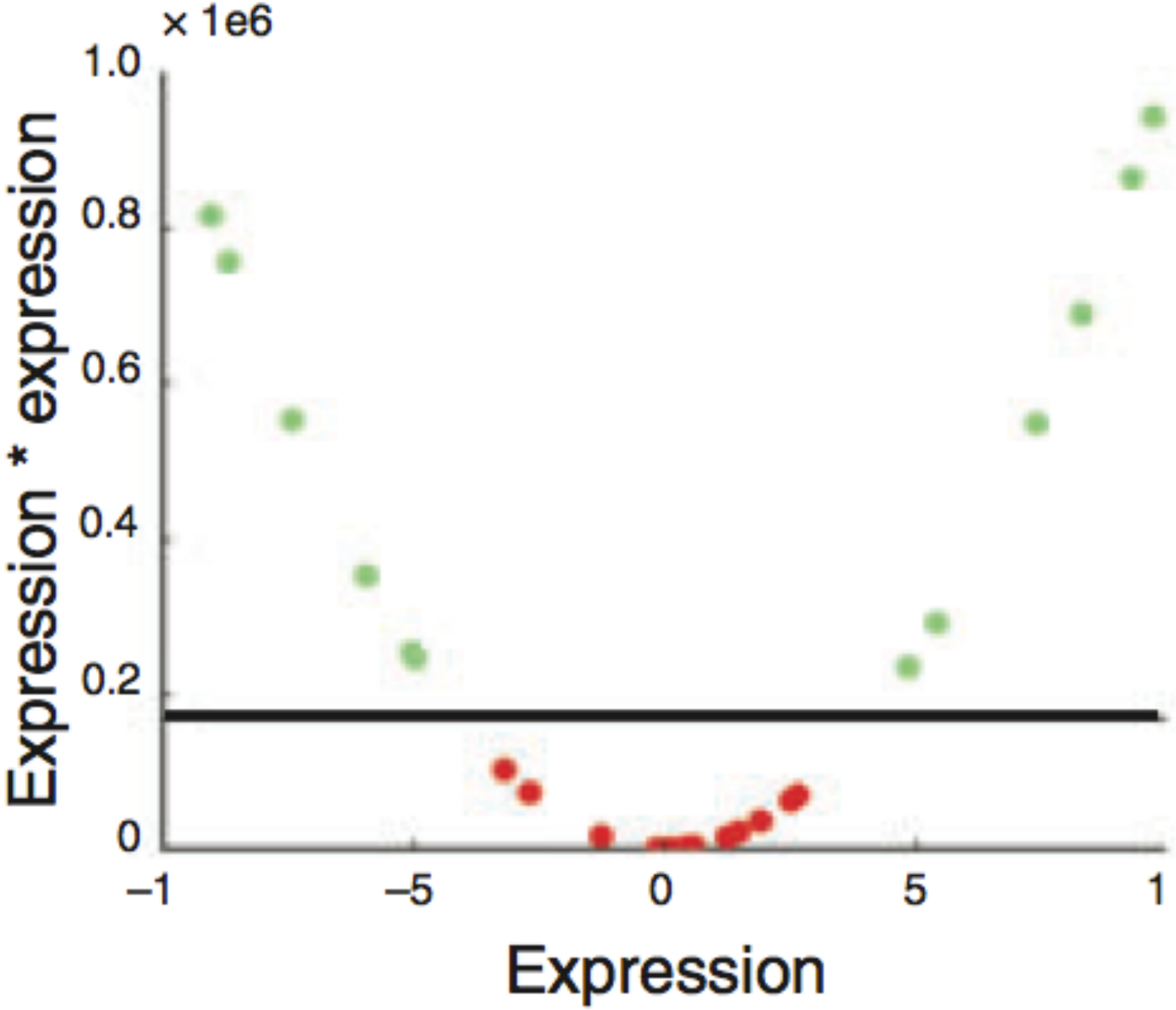
Yes
(by a 1D-hyperplane = dot)



NO

William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

- Transform (x_i) into (x_i, x_i^2)



William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

- Non-linear SVM
  - In some cases (e.g. the above example), even soft-margin cannot solve the non-separable problem
  - Generally speaking, we can apply some function to the original data points so that different classes become linearly separable (maybe with the help of soft-margin)
  - In the above example, the function is f(x) = (x, x^2)
  - The most import trick in SVM: to allow for the transformation, we only need to define the "kernel function", $k(x_i, x_j) = f(x_i) \bullet f(x_j)$
  - The above example essentially uses a polynomial kernel

- Math behind the “kernel trick”
  - In optimization theory, a constrained optimization problem can be formulated into its dual problem (the original problem is called primal problem)

  - The dual formulation of SVM can be expressed as:

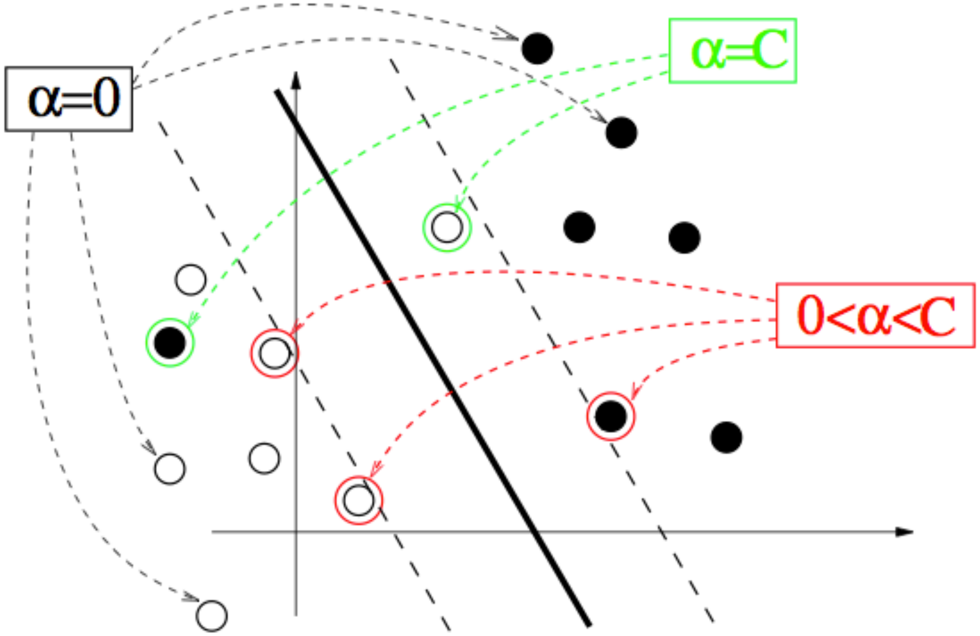
  - $\underset{\alpha}{\text{Maximize}} \sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i \bullet x_j$, subject to

    $\sum_{i} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C$

Complicated!

  - The “Kernel”: $x_i \bullet x_j$, can be replaced by more sophisticated kernels: $k(x_i, x_j) = f(x_i) \bullet f(x_j)$

- “Support vector machine”, where does the name come from?



  - The x_i for which $\alpha_i > 0$ are called support vectors
  - They fall between or right on the separating margins

from http://cbio.ensmp.fr/~jvert/talks/110401mines/mines.pdf

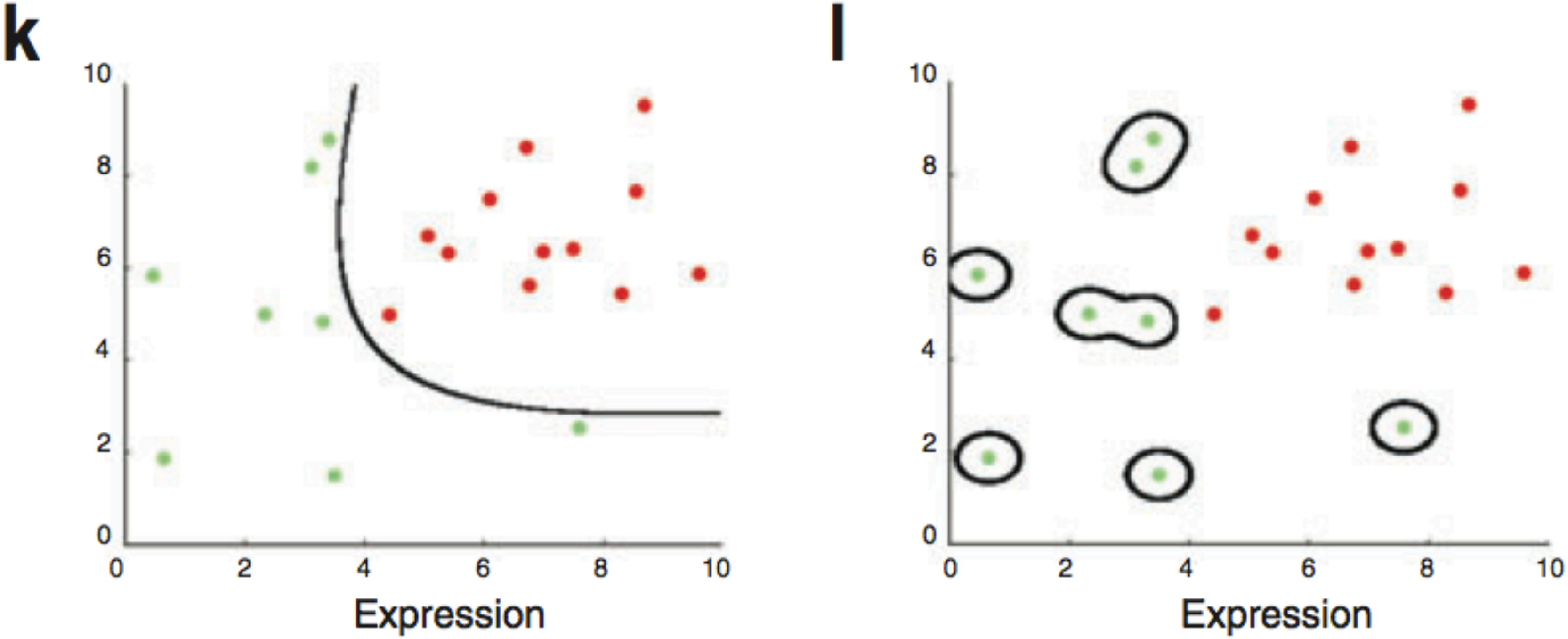
Key idea in the **Kernel Trick**

- Original SVM optimization for refining the hyperplane parameters *w* & *b* in terms of a linear combination of x_i can be replaced by a different optimization problem using "Lagrange multipliers" α_i
  - One only optimizes using the product of x_i*x_j, now expressing the solution in terms of α_i which are non-zero for x_i that function as support vectors
- In a non-linear SVM x_i*x_j is replaced by f(x_i)*f(x_j), so you don't need to know f(x_i) itself only the product
  - This is further formalized in the kernel trick where f(x_i)*f(x_j) is just replaced by k(x_i, x_j). That is, one only has to know the “distance” between x_i & x_j in the high-dimensional space -- not their actual representation

- Two commonly used kernels (and there are more)
- Polynomial kernel:
  - $k(x_i, x_j) = (x_i \bullet x_j + a)^d$
  - a = 1 (inhomogeneous) or 0 (homogenous)
  - d controls the degree of polynomial and henceforth the flexibility of the classifier
  - degenerates to linear kernel when a = 0 and d = 1
- Gaussian kernel:
  - $k(x_i, x_j) = (-1/\sigma \| x_i - x_j \|^2)$
  - σ controls the width of the Gaussian and plays a similar role as d in the polynomial kernels

- More about kernels
  - With kernels, non-vector data can be easily handled – we only need to define the kernel function between two objects
  - Examples of non-vector biological data include: DNA and protein sequences (“string kernels”), nodes in metabolic or protein-protein interaction networks, microscopy images, etc
  - Allows for combining different types of data naturally – define kernels on different data types and combine them with simple algebra
- Questions for practitioners: Which kernel to use? How to choose parameters?
  - Trial and error
  - Cross-validation
- High-degree kernels always fit the training data well, but at increased risks of over-fitting, i.e. the classifier will not generalize to new data points
  - One needs to find a balance between classification accuracy on the training data and regularity of the kernel (not allowing the kernel to be too flexible)

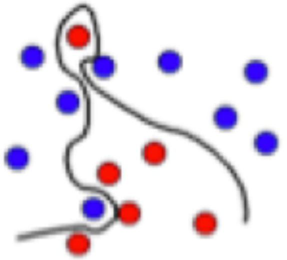
- A low-degree kernel (left) and an over-fitting high-degree kernel (right)



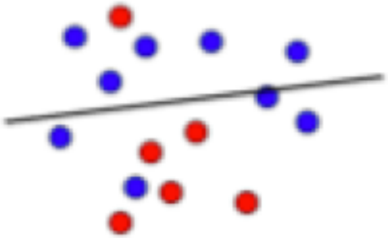
William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

- The parameter C has a similar role
  - Large C will make few classification errors on the training data
  - But this may not generalize to the testing data
  - Small C pursues a large separating margin at the expenses of some classification errors on the training data.
  - The accuracy more likely to generalize to testing data

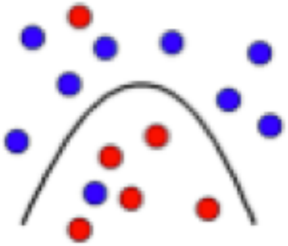
Large C

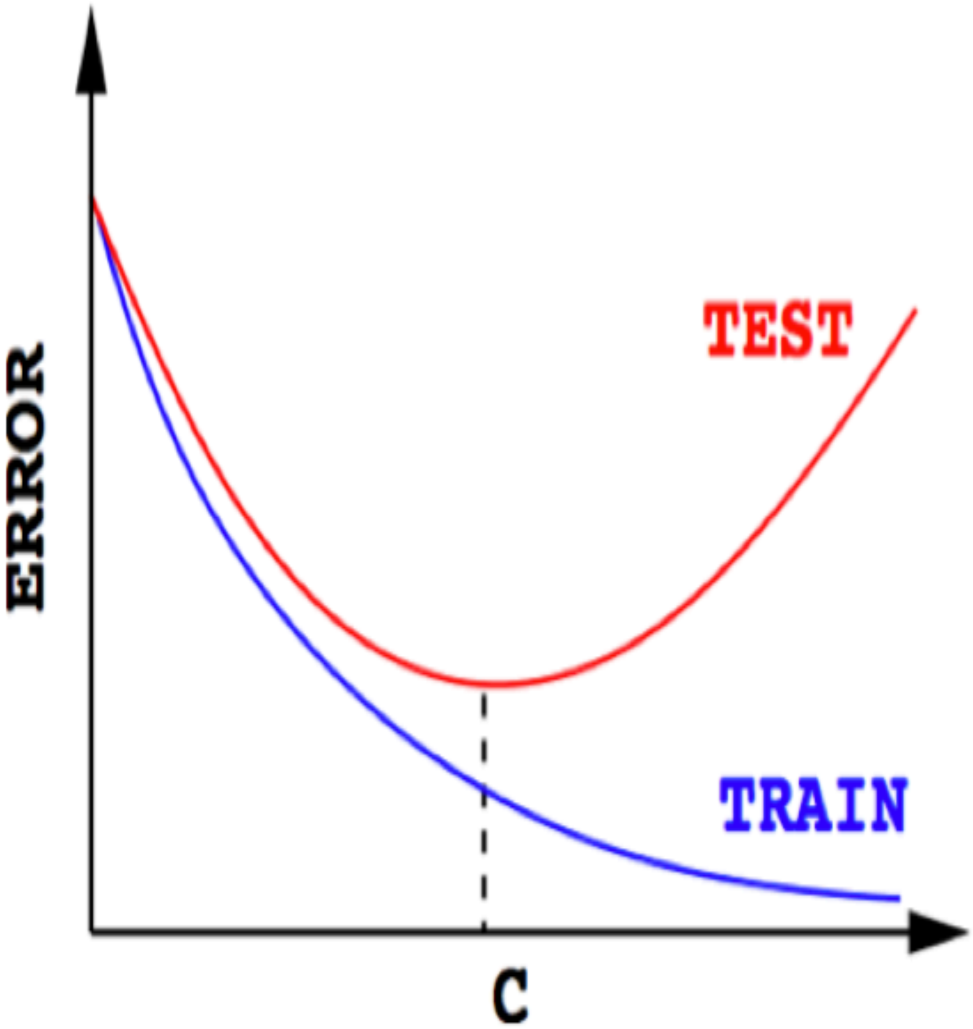


Small C



Intermediate C



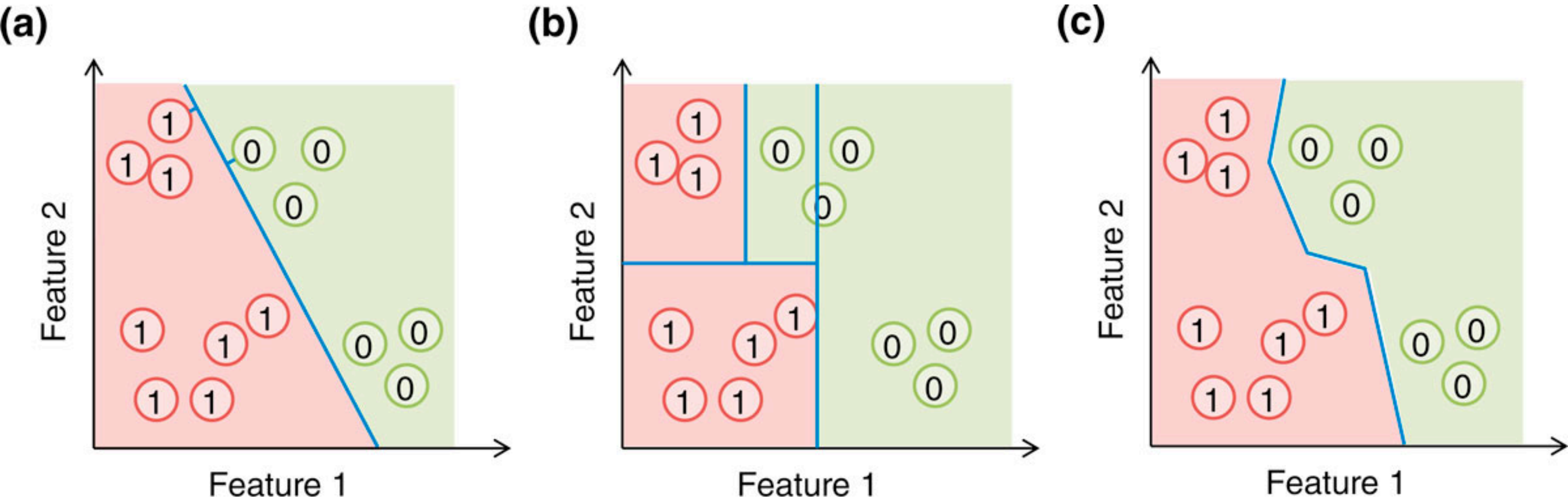


http://cbio.ensmp.fr/~jvert/talks/110401mines/mines.pdf

Supervised Mining:

Decision Boundary & Semi-supervised Approaches

Decision boundaries: SVM v Tree v Nearest NBR



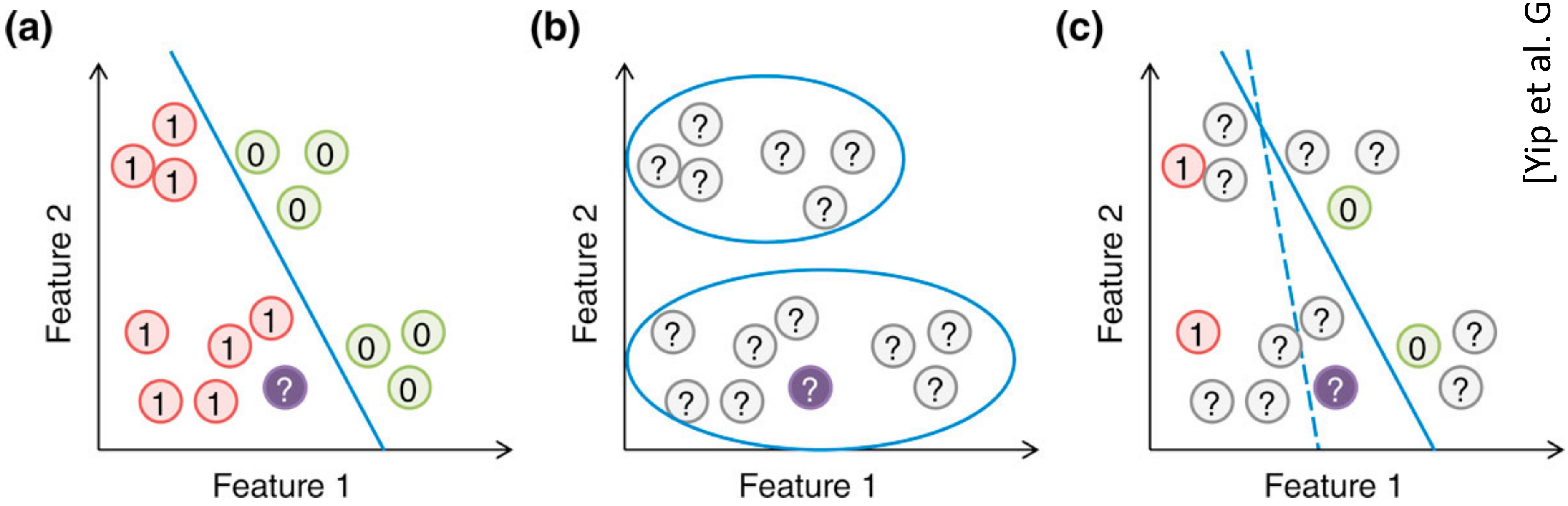
(a) A support vector machine (SVM) forms an affine decision surface (a straight line in the case of two dimensions) in the original feature space or a vector space defined by the similarity matrix (the kernel), to separate the positive and negative examples and maximize the distance of it from the closest training examples (the support vectors, those with a perpendicular line from the decision surface drawn). It predicts the label of a genomic region based on its direction from the decision surface. In the case a kernel is used, the decision surface in the original feature space could be highly non-linear. (b) A basic decision tree uses feature-parallel decision surfaces to repeatedly partition the feature space, and predicts the label of a genomic region based on the partition it falls within. (c) The one-nearest neighbor (1-NN) method predicts the label of a genomic region based on the label of its closest labeled example. In all three cases, the areas predicted to be positive and negative are indicated by the red and green background colors, respectively.

[Yip et al. Genome Biology 2013 14:205   doi:10.1186/gb-2013-14-5-205]

Semi-supervised Methods

- Supervised & Unsupervised:
  Can you combine them? YES
  - RHS (c) shows modifying the optimum decision boundary in (a) by "clustering" of unlabeled points

[Yip et al. Genome Biol. ('13)]



Supervised, unsupervised and semi-supervised learning. (a) In supervised learning, the model (blue line) is learned based on the positive and negative training examples, and the genomic region without a known class label (purple circle) is classified as positive according to the model. (b) In unsupervised learning, all examples are unlabeled, and they are grouped according to the data distribution. (c) In semi-supervised learning, information of both labeled and unlabeled examples is used to learn the parameters of the model. In this illustration, a purely supervised model (dashed blue line) classifies the purple object as negative, while a semi-supervised model that avoids cutting at regions with a high density of genomic regions (solid blue line) classifies it as positive.