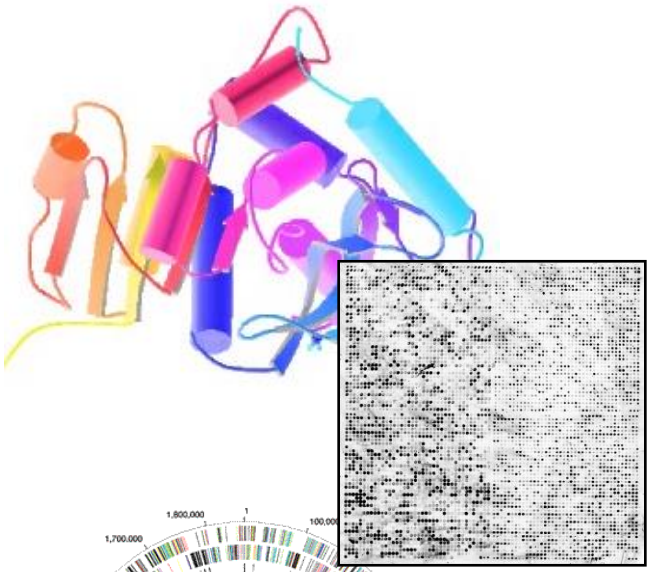
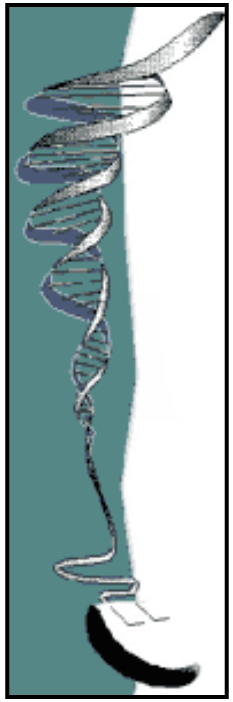
Biomedical Data Science:
Supervised Datamining – SVMs





Mark Gerstein, Yale University

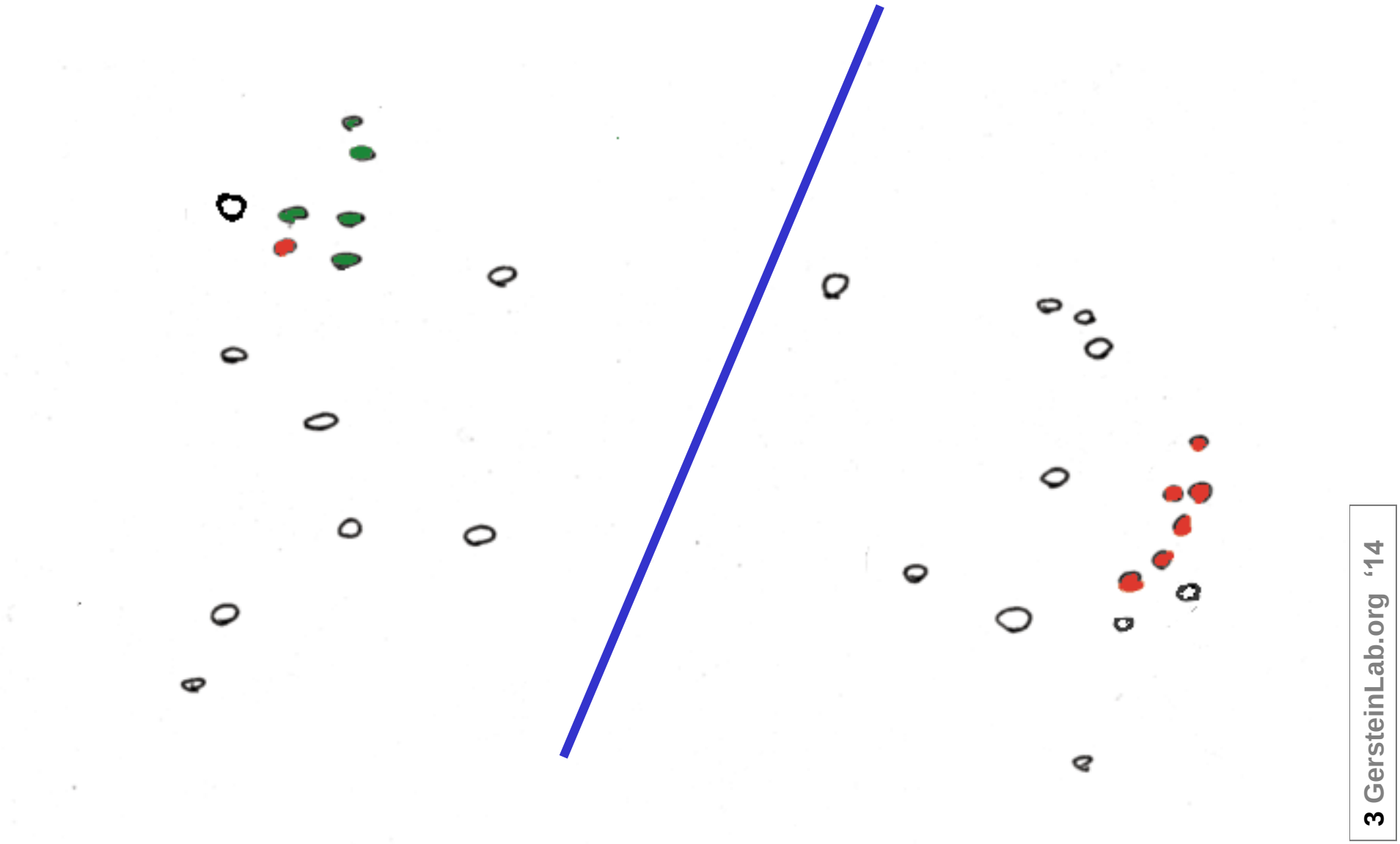
GersteinLab.org/courses/452

(Last edit in spring ’22, final.
This year’s 22m8c pack is essentially the same as last year’s M8c one.)

Supervised Mining:

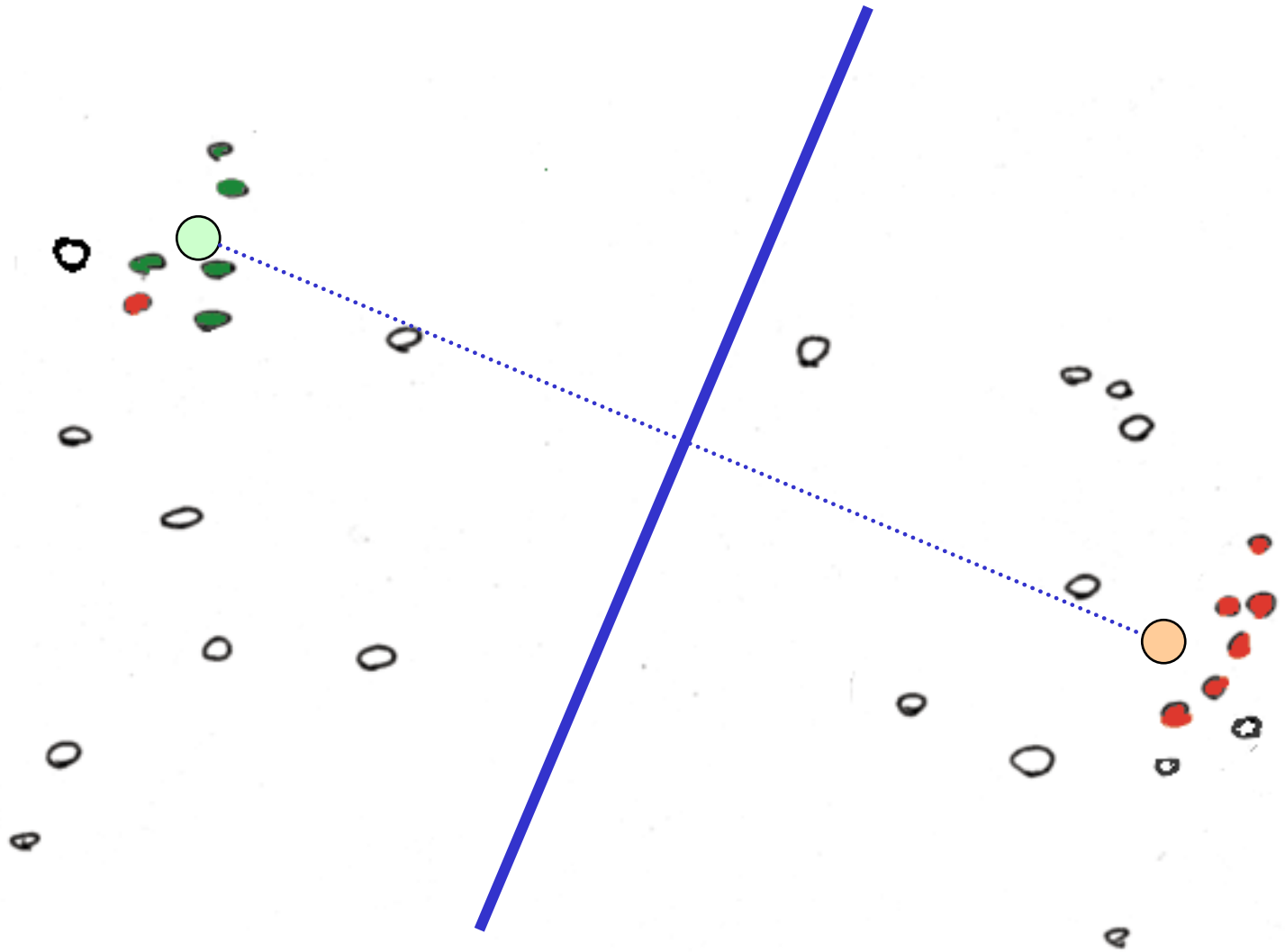
SVMs

Find a Division to Separate Tagged Points



3 GersteinLab.org '14

Discriminant to Position Plane



4 GersteinLab.org '14

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

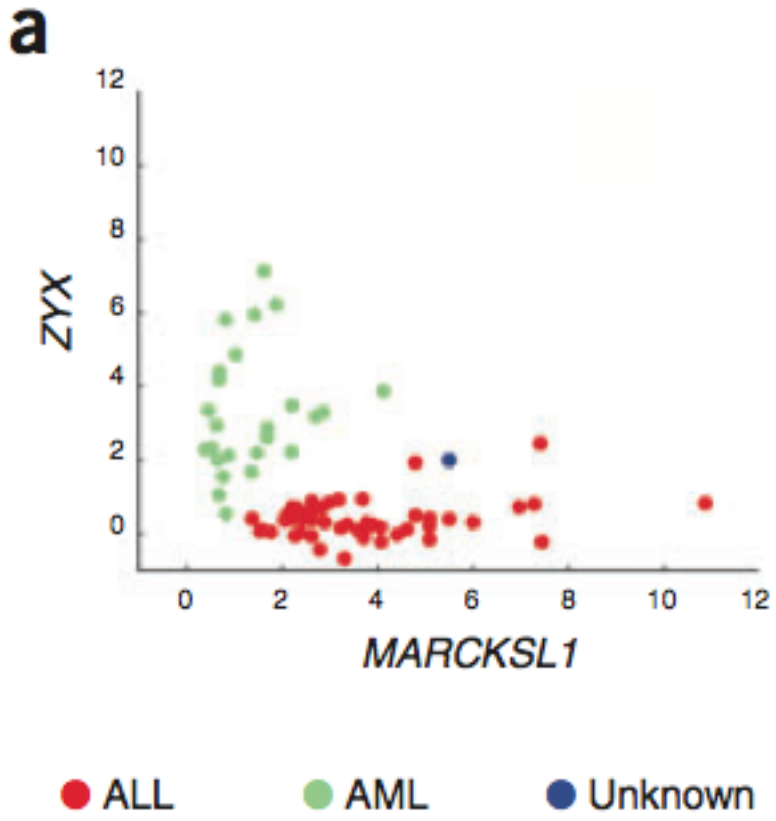
$$\vec{w} = S_W^{-1}(\vec{m}_1 - \vec{m}_2)$$

Support Vector Machines

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

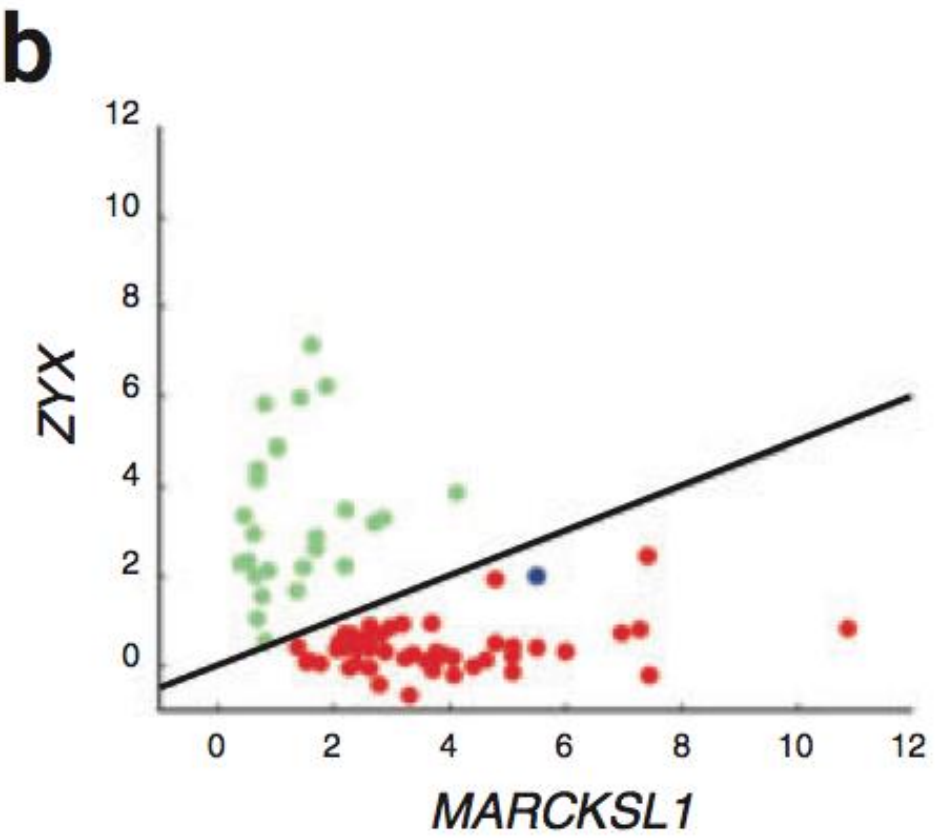
- It has also been successfully applied in many biological problems:
  - Disease diagnosis
  - Automatic genome functional annotation
  - Prediction of protein-protein interactions
  - and more…

- Example: Leukemia patient classification



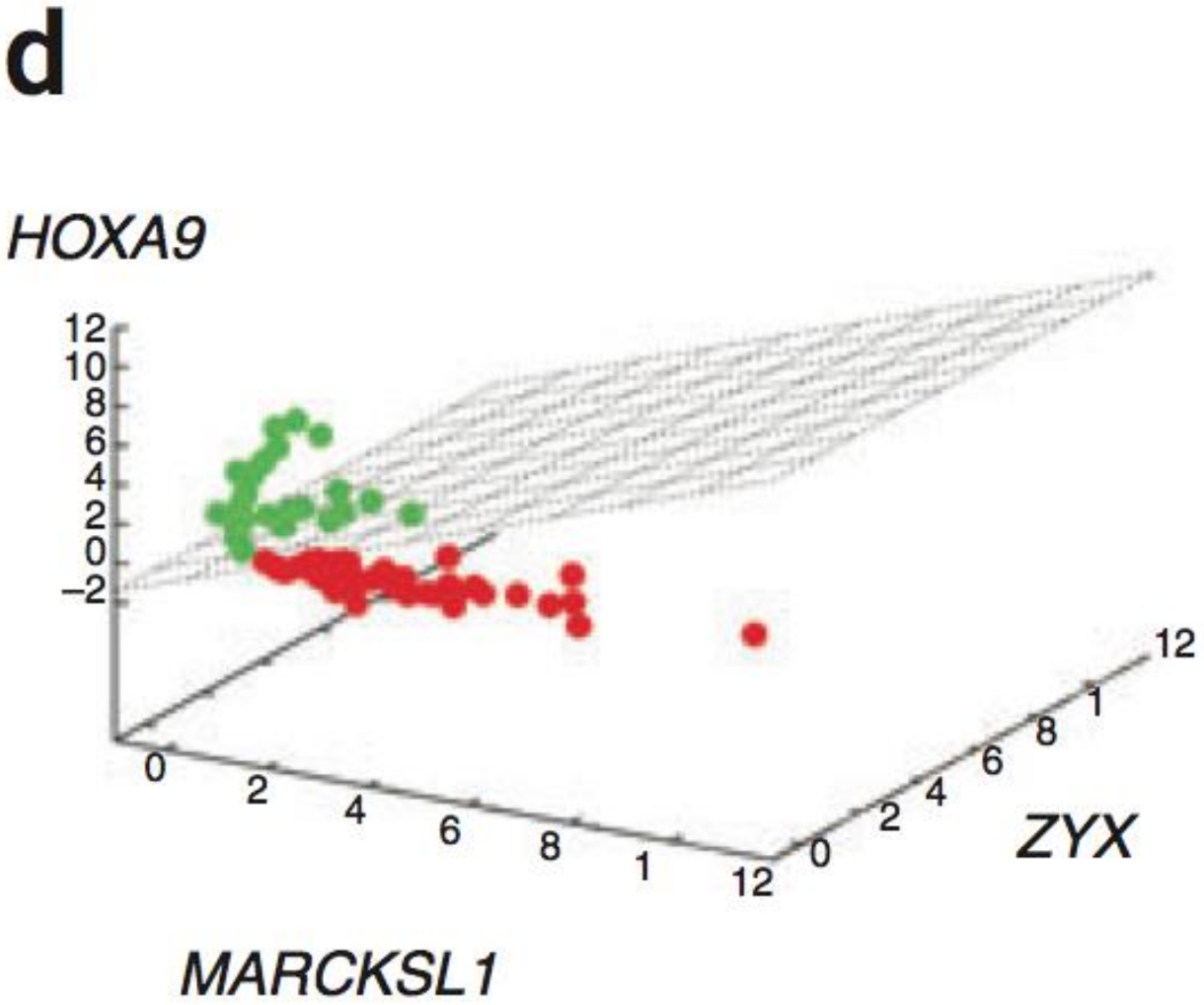
ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia

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



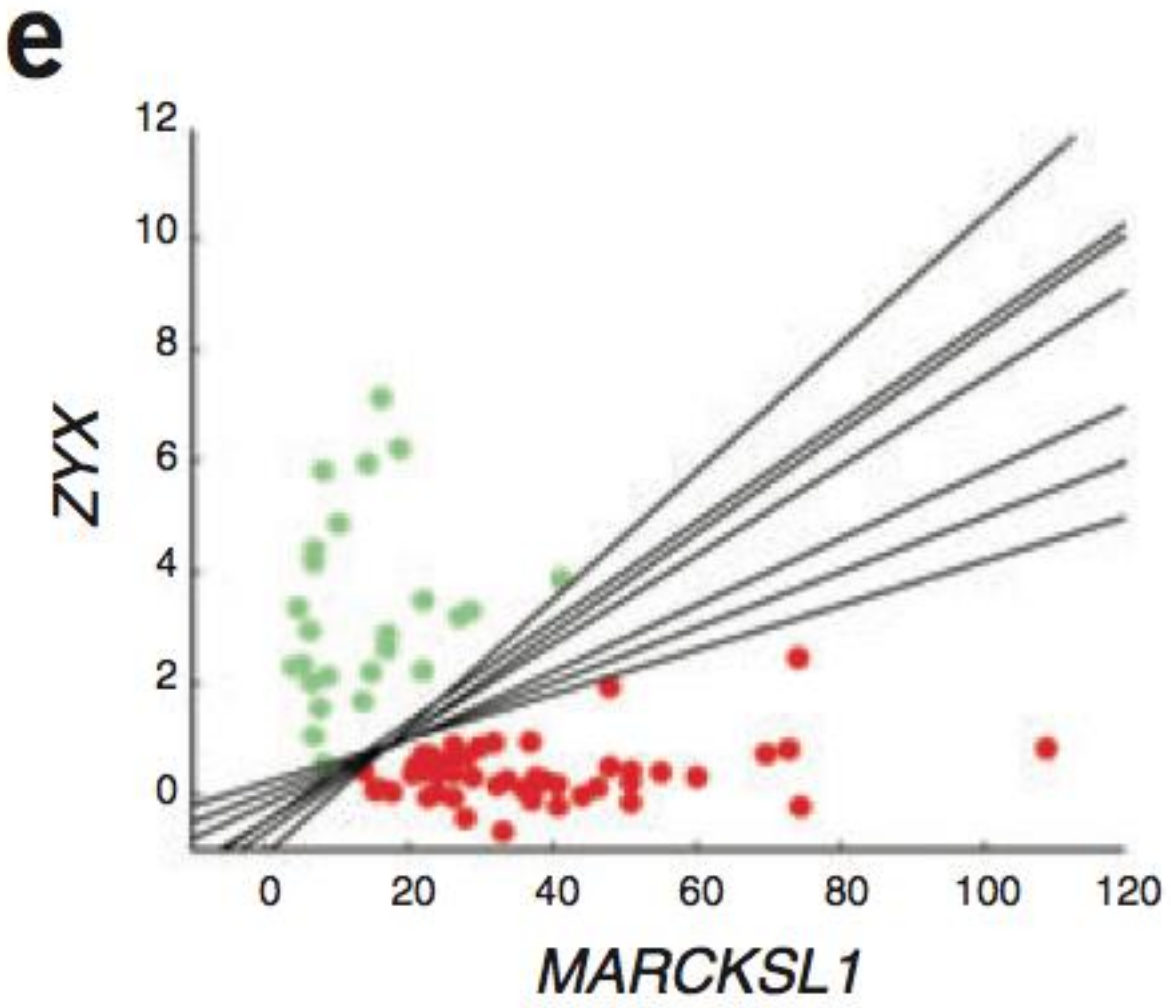
- A simple line suffices to separate the expression profiles of ALL and AML

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



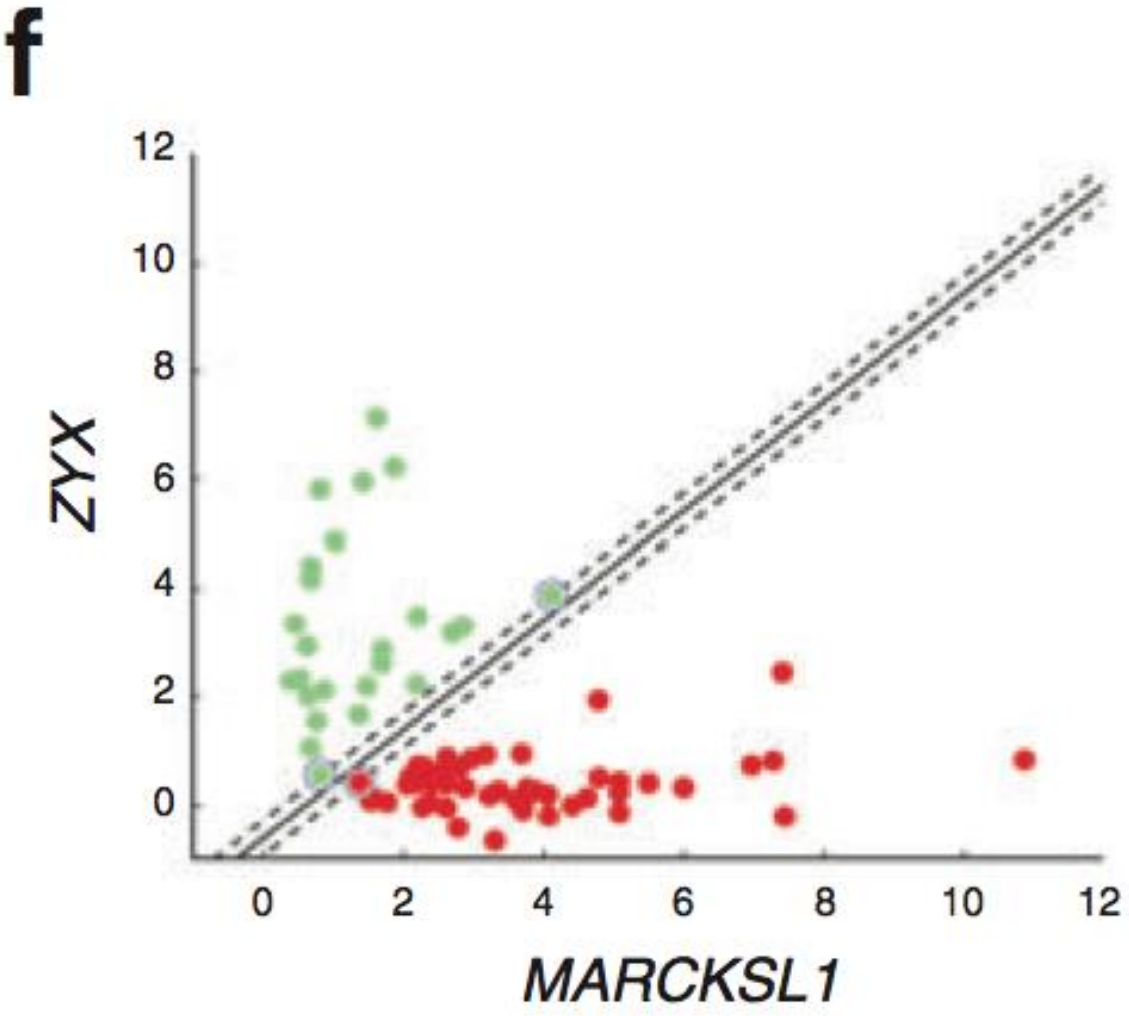
- In the case of more than two genes, a line generalizes to a plane or "hyperplane".
- For generality, we refer to them all as "hyperplane"

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



- Is there a “best” line?

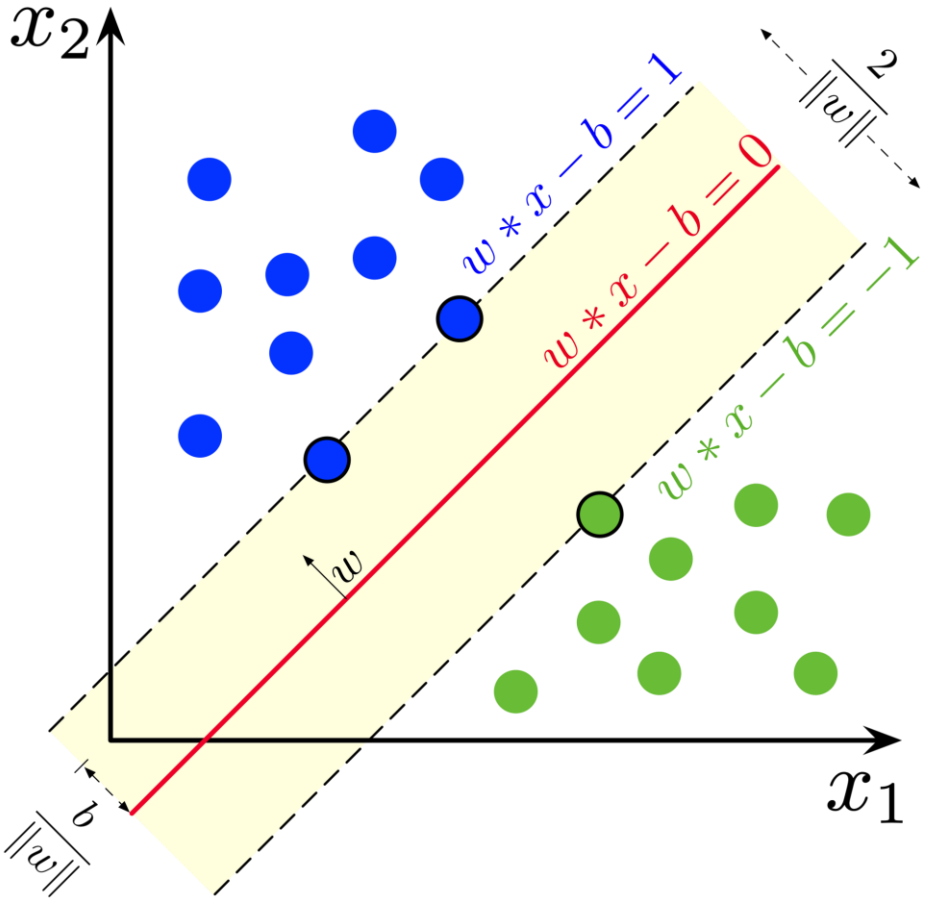
William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



- 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: $2/\|w\|, \|w\| = \sqrt{w \cdot w}$



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

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

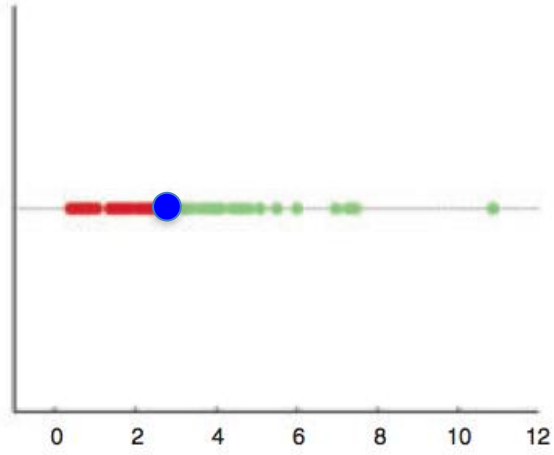
- Mathematically,
  - Minimize$_{w,b}$ 1/2 ||w||^2, subject to:
  - for $y_i = 1$, $w \cdot x_i + b \geq 1$
  - for $y_i = -1$, $w \cdot x_i + b \leq -1$
  - Combining them, for any i, $y_i(w \cdot x_i + b) \geq 1$

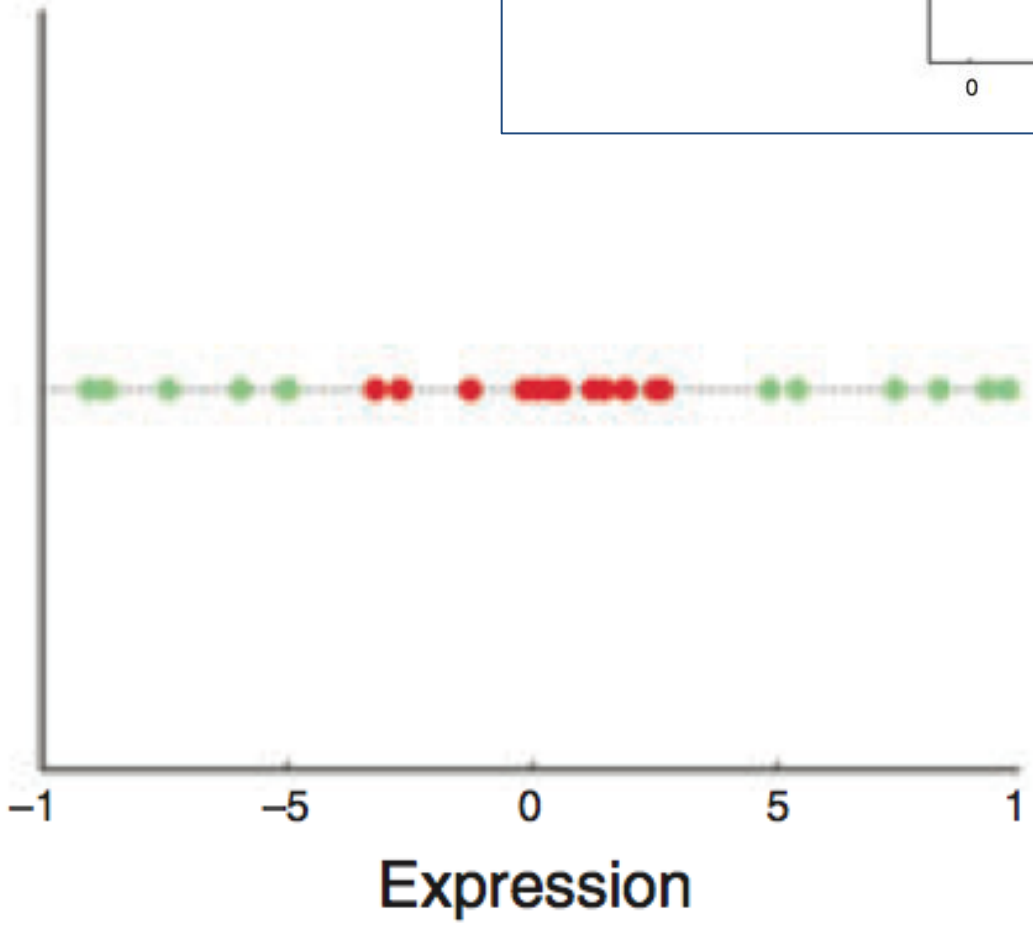
- The solution expresses *w* as a linear combination of the x_i
- So far, we have been assuming that the data points from two classes are always easily linearly separable. But that's not always the case. (Also, could allow a "soft margin" via slack parameters – something we are not covering.)

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

- Are linear separating hyperplanes enough?

Yes
(by a 1D-hyperplane = dot)

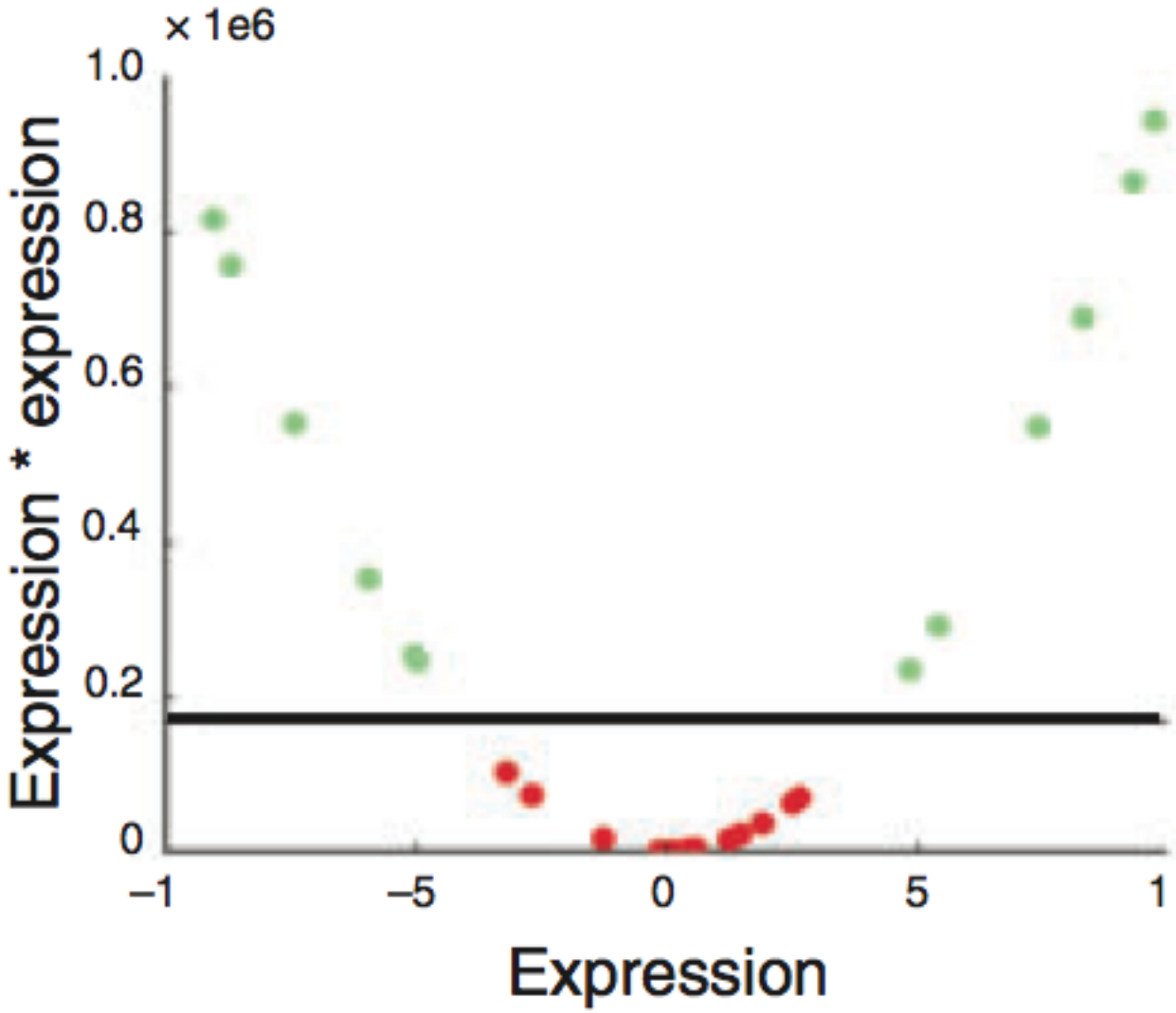




NO

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", $k(x_i, x_j) = f(x_i) \cdot f(x_j)$
  - 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)

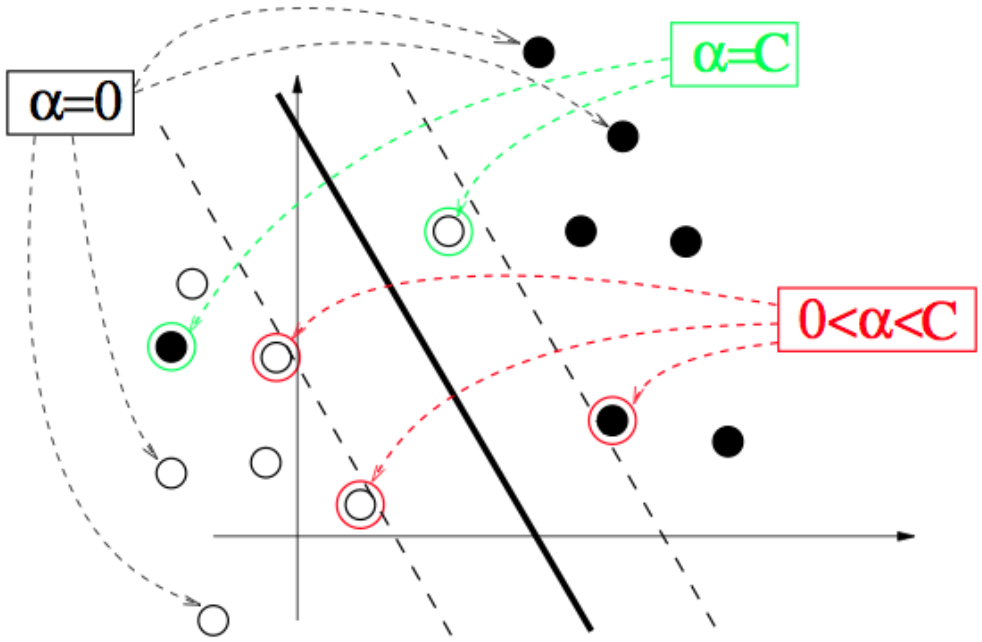
  - The dual formulation of SVM can be expressed as:
  - $\underset{a}{\text{Maximize}} \sum_i a_i - \frac{1}{2} \sum_{i,j} y_i y_j a_i a_j x_i \cdot x_j$, subject to

    $$\sum_i y_i a_i = 0, 0 \le a_i \le C$$

    Complicated!

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

- "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

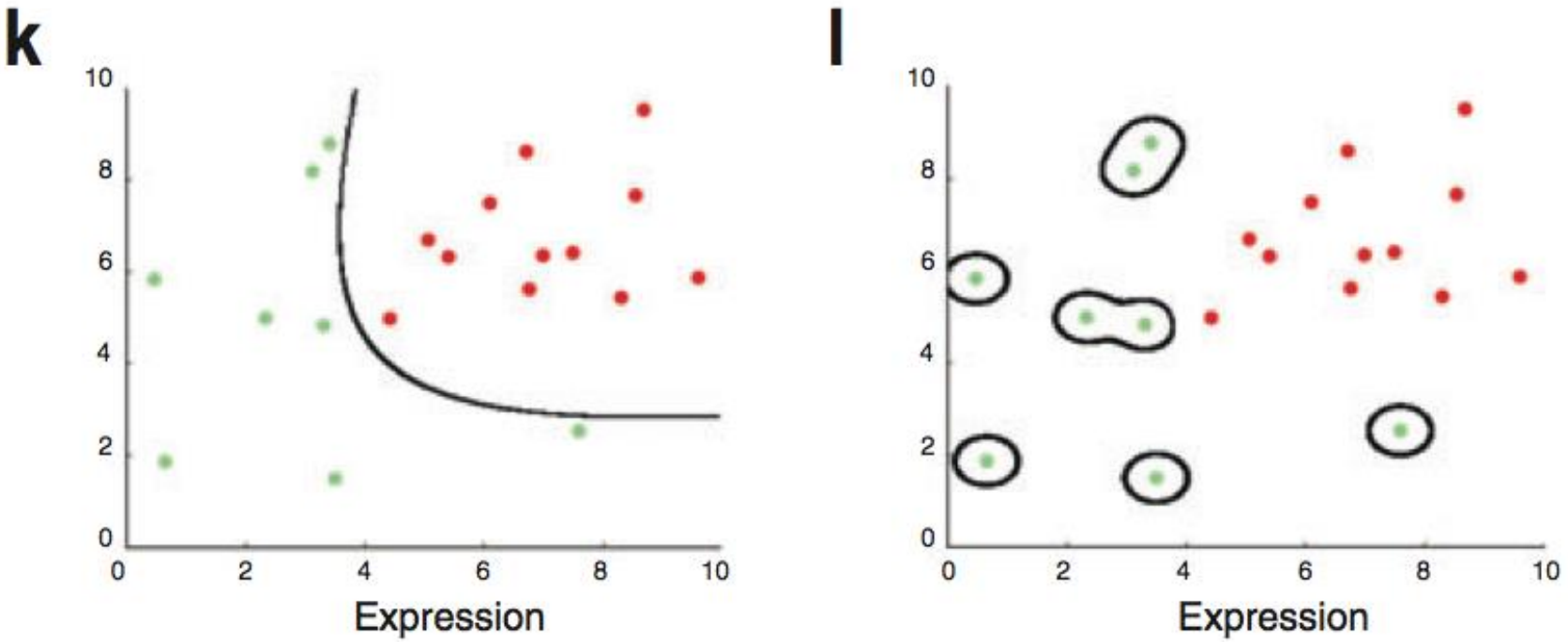
Key idea in the **Kernel Trick**

- 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:
  - $k(x_i, x_j) = (x_i \cdot x_j + a)^d$
  - 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:
  - $k(x_i, x_j) = (-1/s \| x_i - x_j \|^2)$
  - 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)

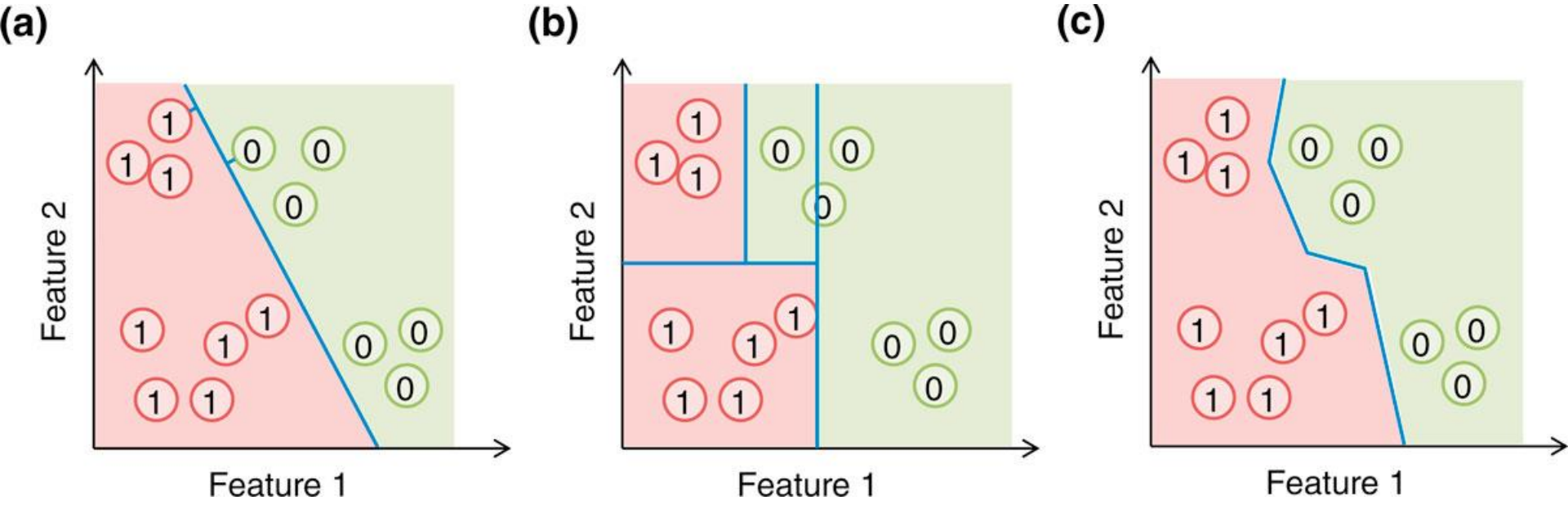


William S Nobel. What is a support vector machine? Nature Biotechnology. 2006

Supervised Mining:

Decision Boundary & Semi-supervised Approaches

Decision boundaries: SVM v Tree v Nearest NBR



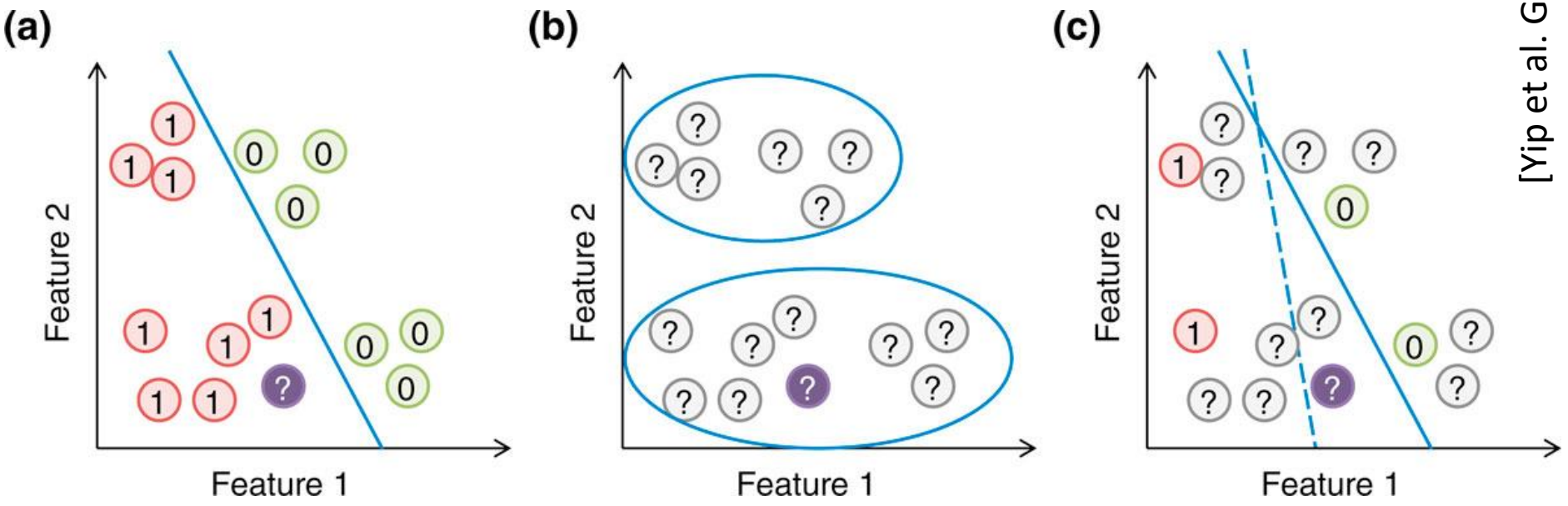
(a) A support vector machine (SVM) forms an affine decision surface (a straight line in the case of two dimensions) in the original feature space or a vector space defined by the similarity matrix (the kernel), to separate the positive and negative examples and maximize the distance of it from the closest 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.