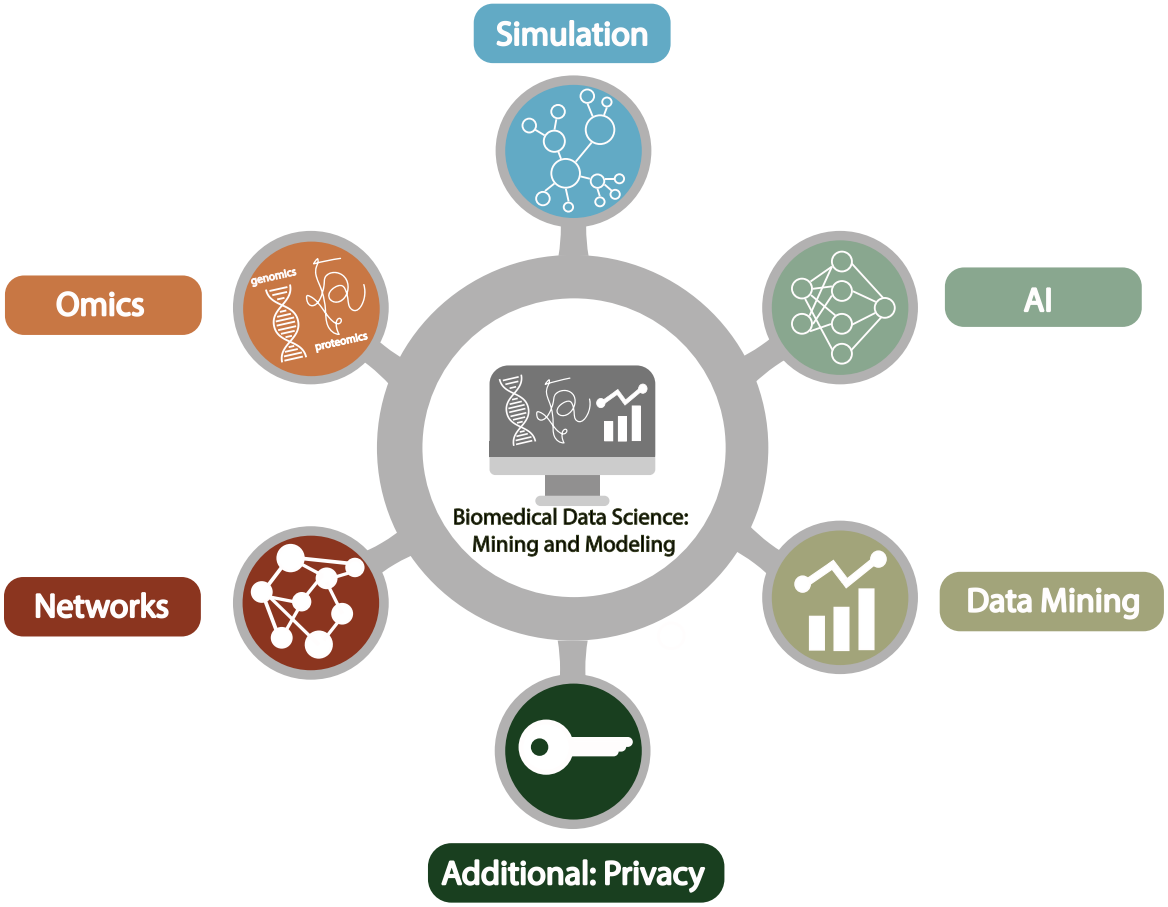
Biomedical Data Science (GersteinLab.org/courses/452)

Supervised Mining – SVMs (25m8c)



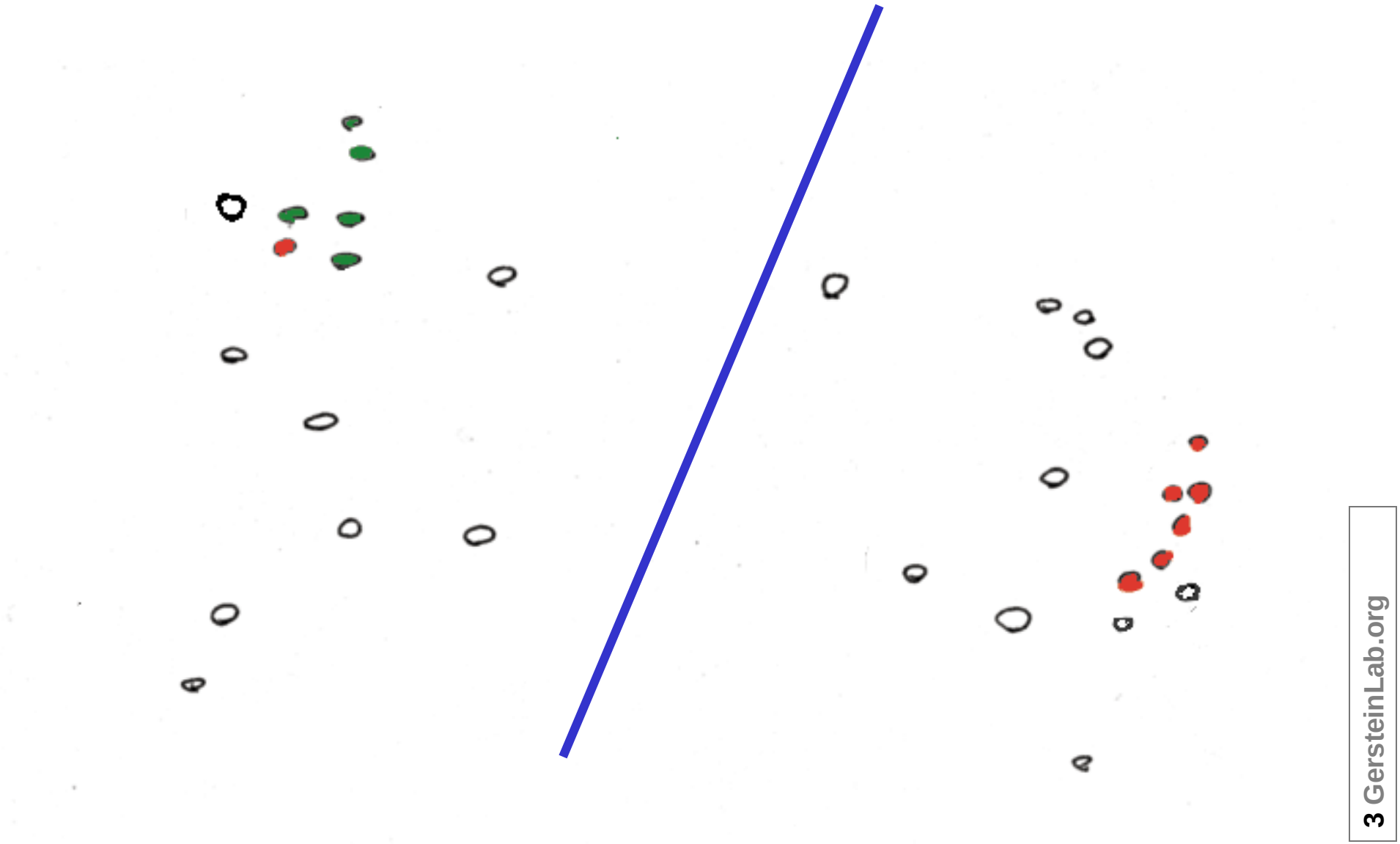
Mark Gerstein
Yale U.

Last edit in spring '25. Condensed (with ~2 slide deletion) and adding connection to ISLR. Similar to 2021's M8c [which has a video].

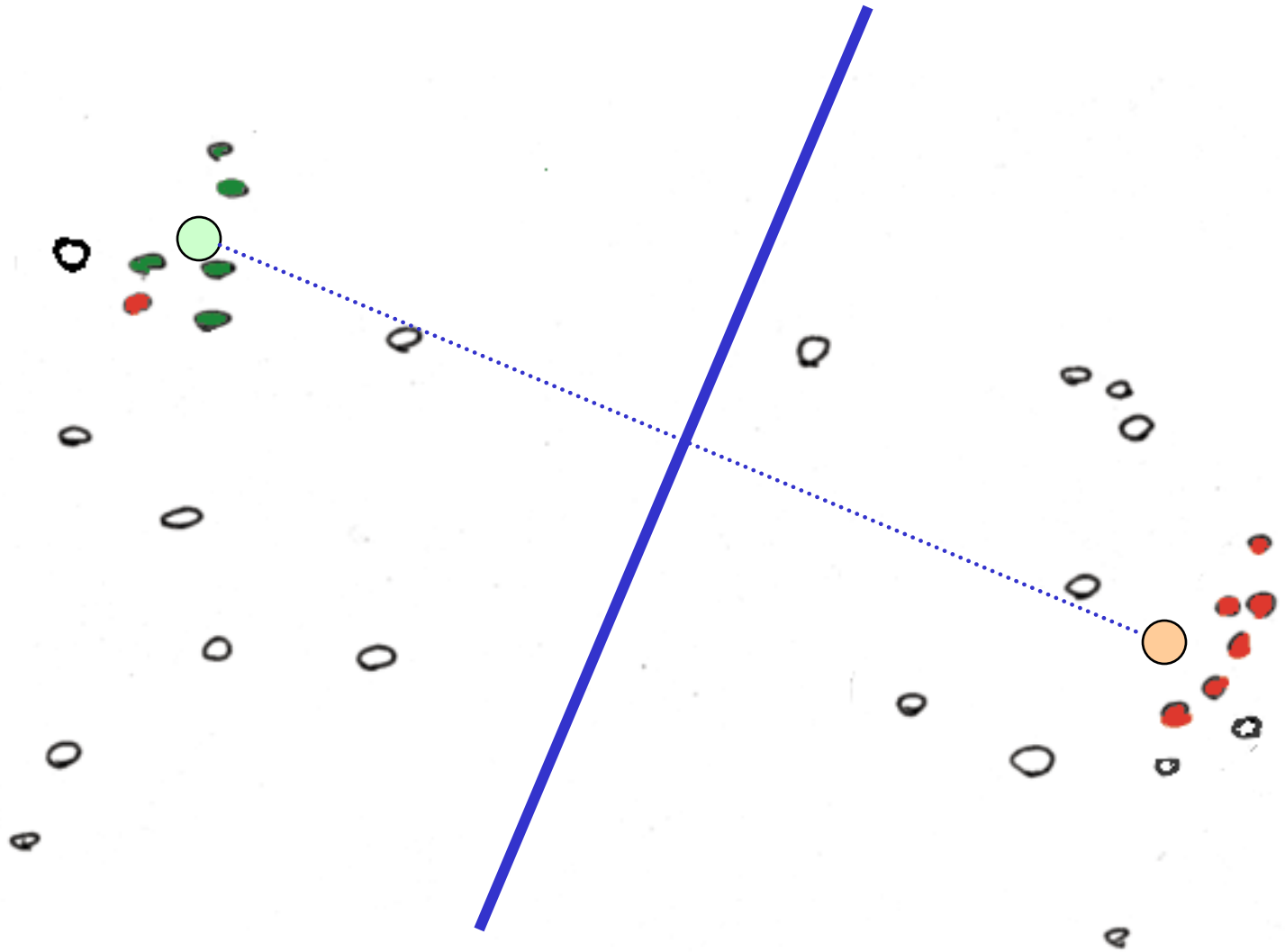
Supervised Mining:

LDA & Linear Methods

Find a Division to Separate Tagged Points



Discriminant to Position Plane



Fisher discriminant analysis & LDA

- $y = w_1x_1 + w_2x_2 + ... + w_nx_n$ 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.
- LDA is a generalization of Fisher's linear discriminant

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

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

Solution of 1st variate

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

Other Classic Linear Methods

- We have already covered lin. regression for GWAS & QTL earlier
- Not going to cover more here (covered in a basic stats class) but **very good to know!!!**

Linear Regression to fit a line &
Logistic Regression "Classifier"

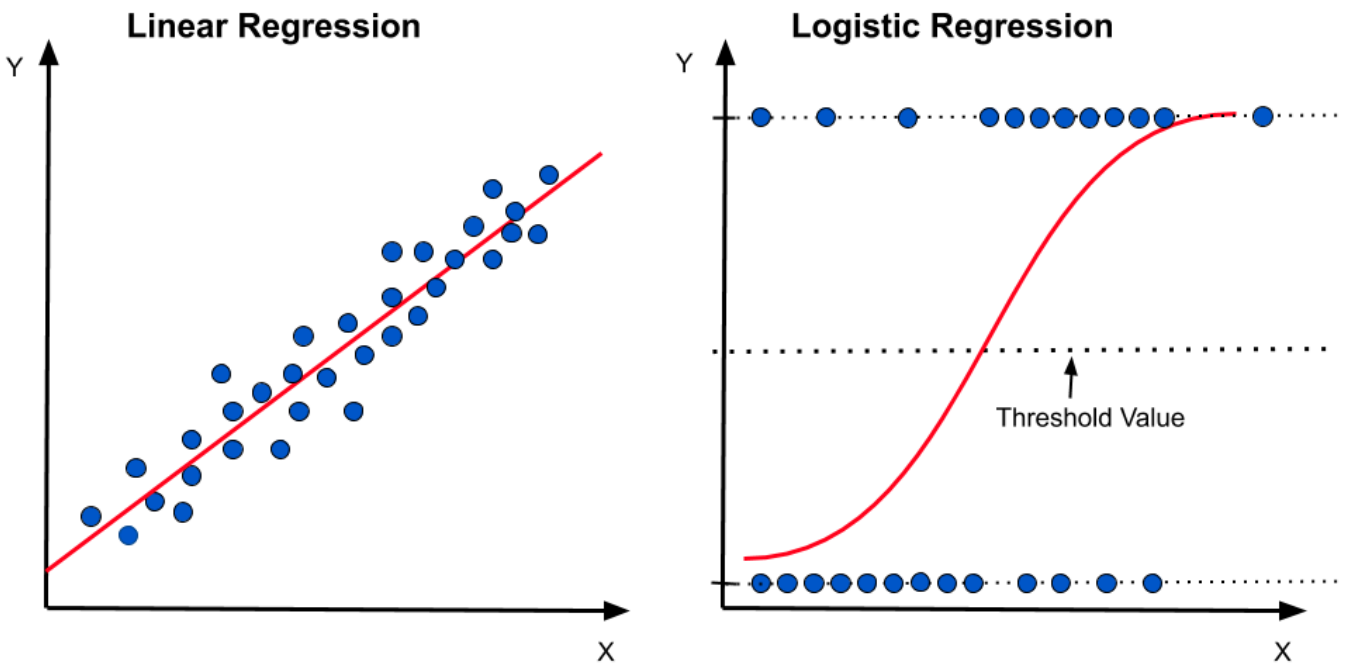


Image from https://medium.com/@praveenraj.gowd/why-is-logistic-regression-called-logistic-regression-and-not-a-logistic-classification-5a4182930 40d

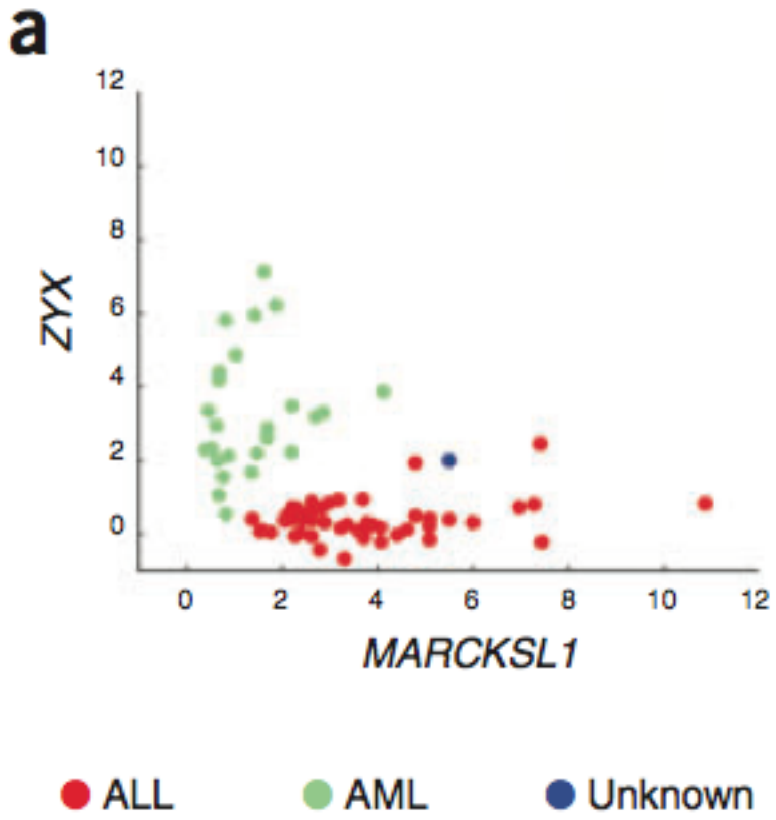
Supervised Mining:

SVM

Support Vector Machines

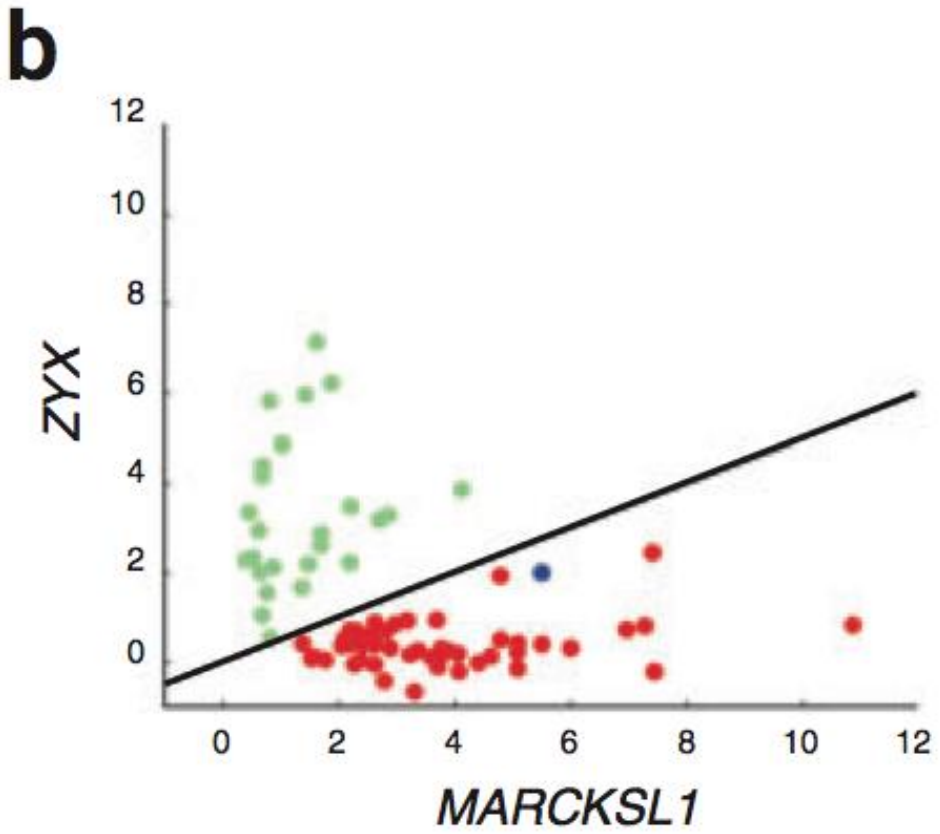
- A very powerful tool for classifications. More flexible than LDA.
- Example Applications:
  - Text categorization
  - Image classification
  - Spam email recognition, etc
- It has also been successfully applied in many biological problems:
  - Disease diagnosis
  - Automatic genome functional annotation
  - Prediction of protein-protein interactions
  - and more…

- Example: Leukemia patient classification



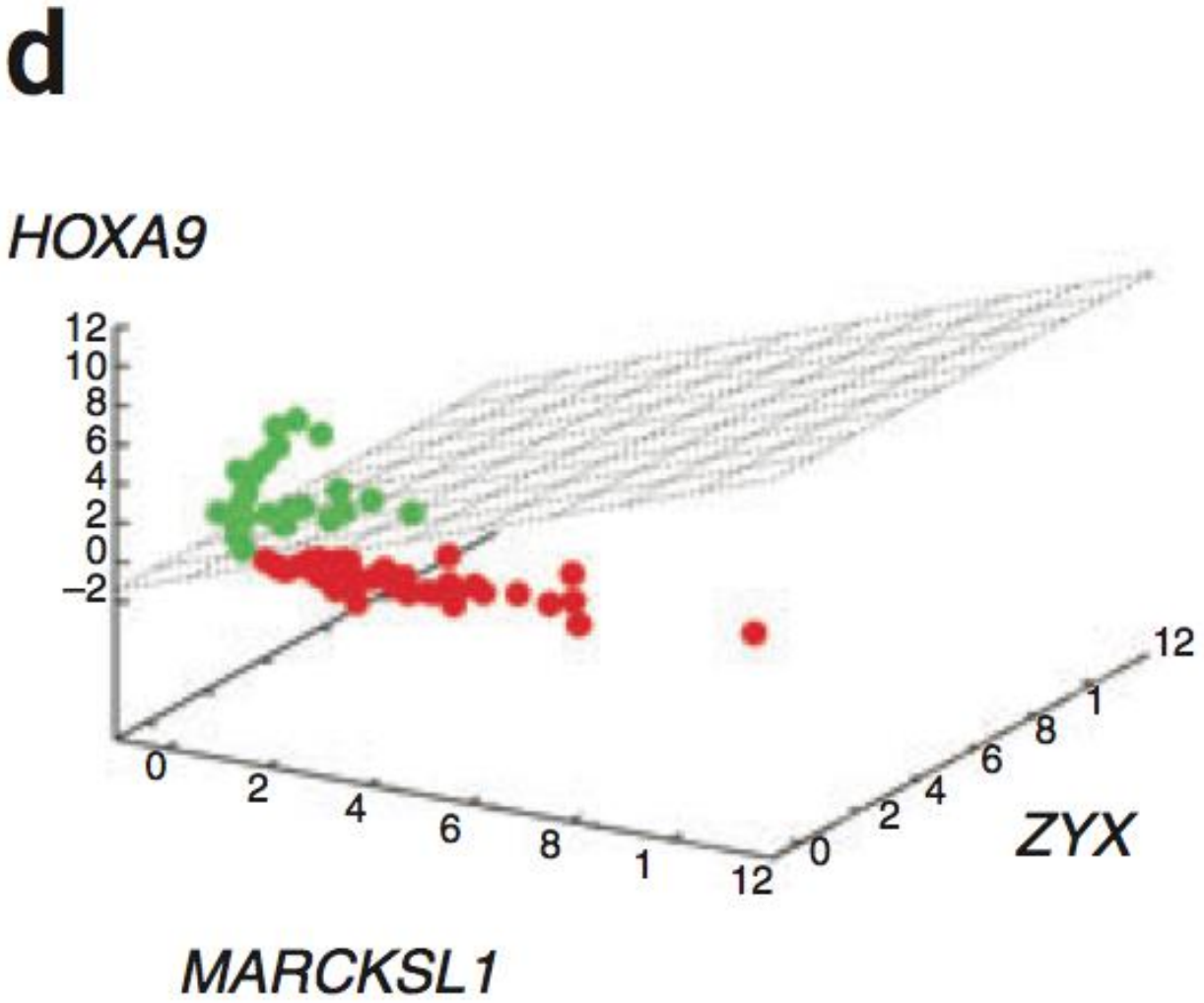
ALL: acute lymphoblastic leukemia
AML: acute myeloid leukemia

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



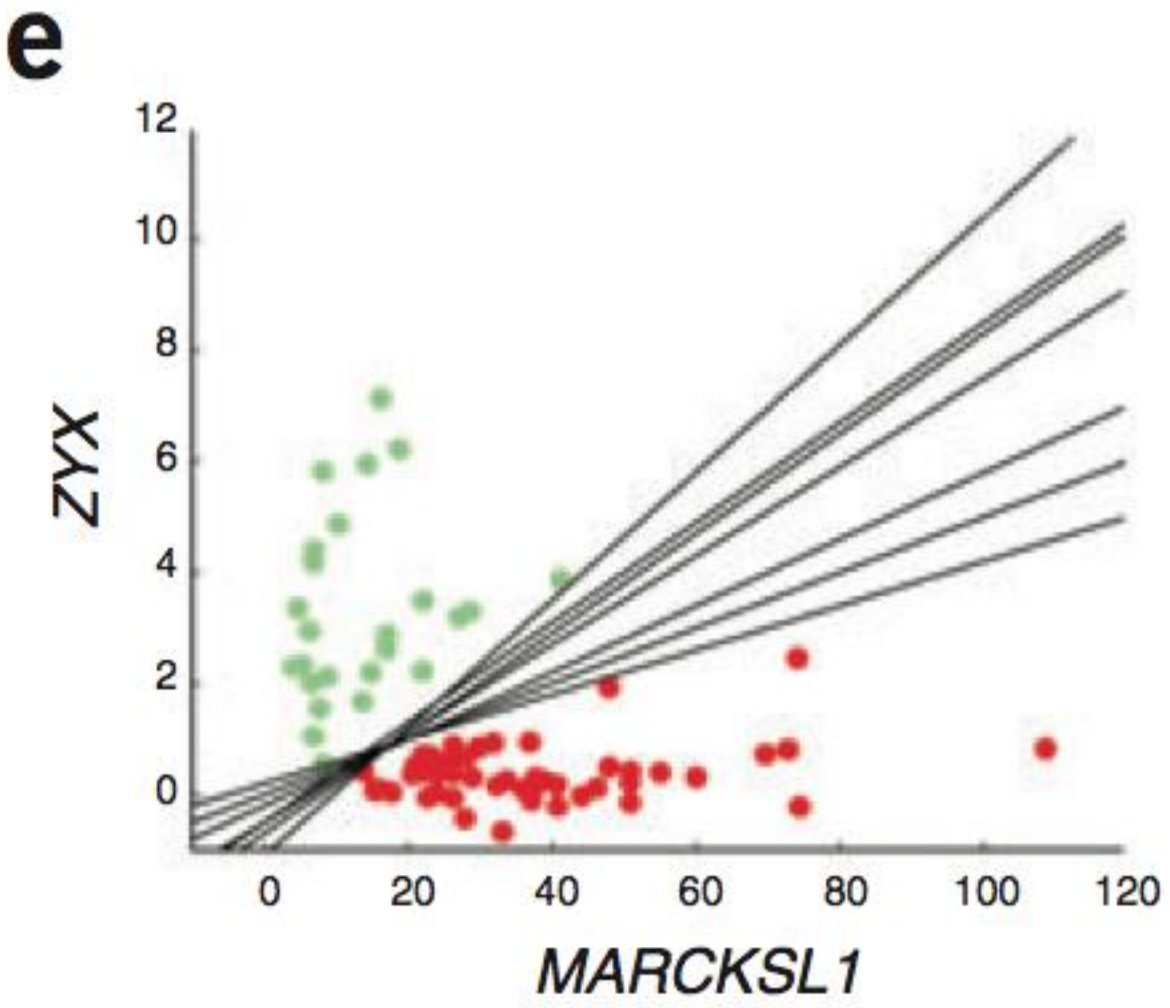
- A simple line suffices to separate the expression profiles of ALL and AML

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



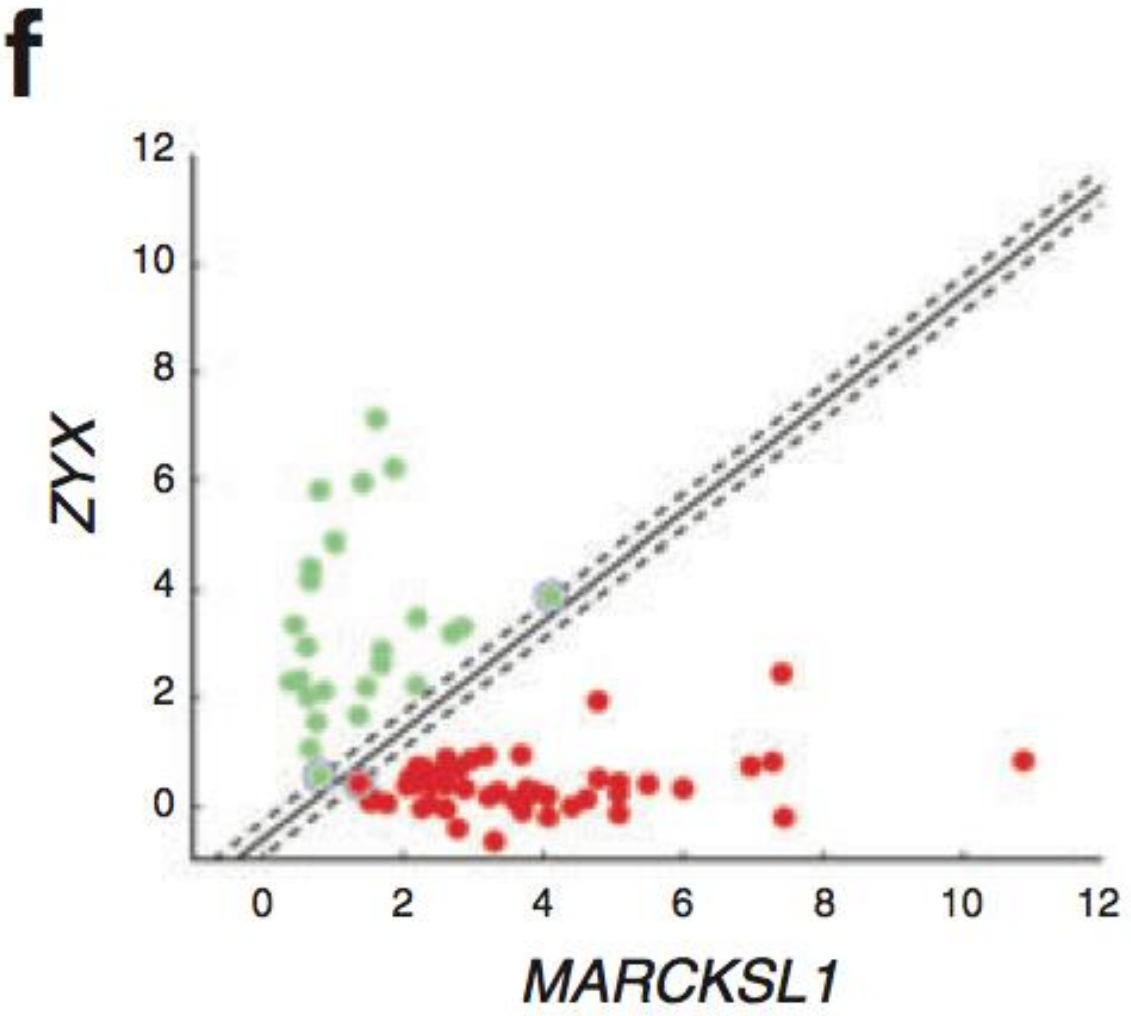
- In the case of more than two genes, a line generalizes to a plane or "hyperplane".
- For generality, we refer to them all as "hyperplane"

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



- Is there a “best” line?

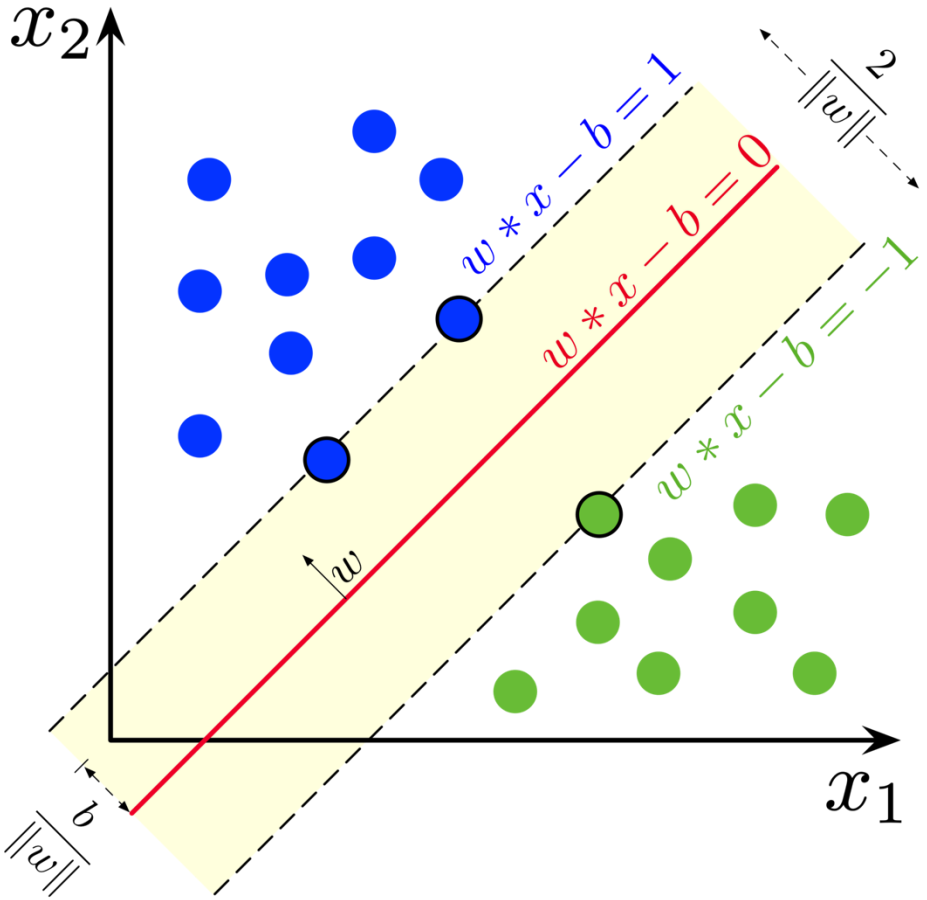
William S Nobel. What is a support vector machine? Nature Biotechnology. 2006



- 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) – e.g. x_i is a vector of the expression profiles & y_i = -1 or 1, which labels the class
- hyperplane: w*x + b = 0
- The margin-width equals to: $2/\|w\|, \|w\| = \sqrt{w \cdot w}$
- We can maximize this.



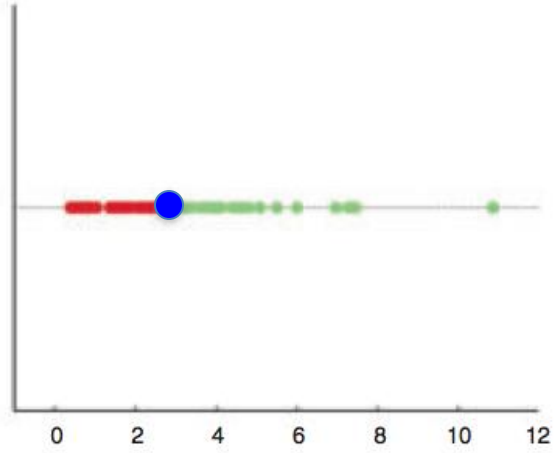
Closest points define the margin & are called support vectors

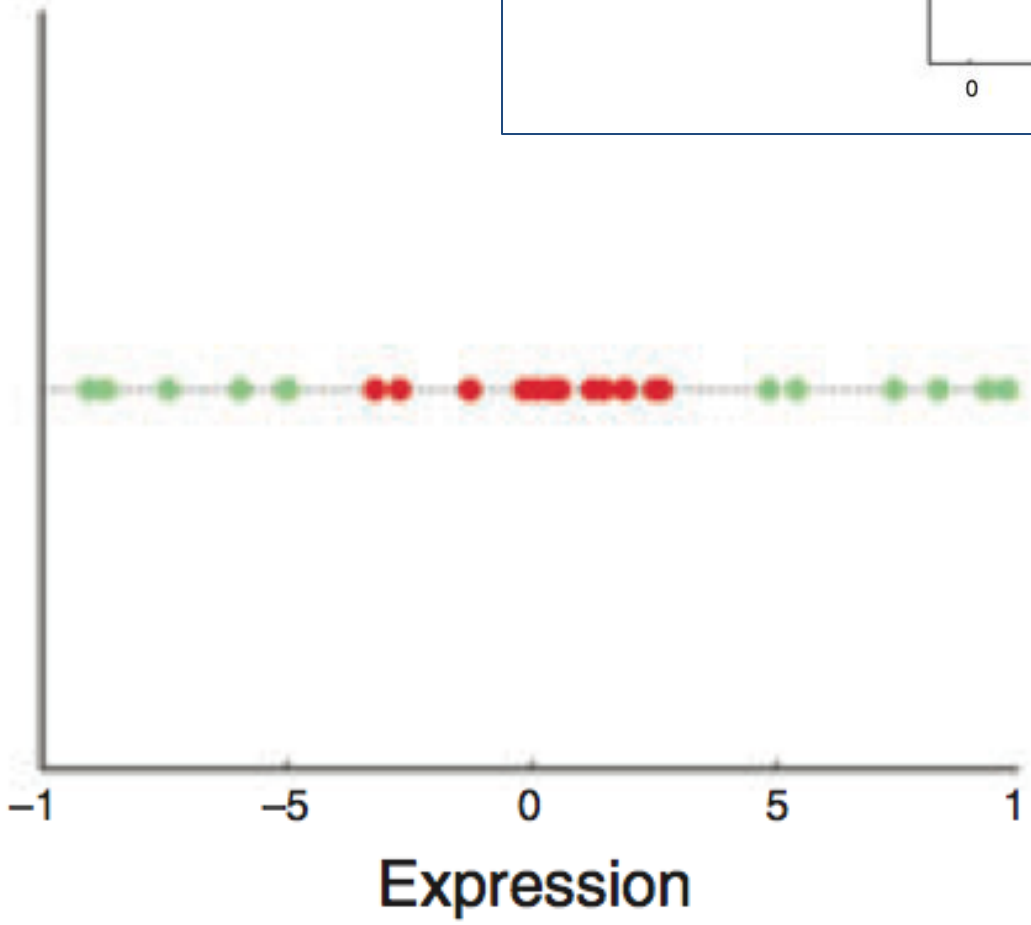
We're assuming the points are linearly separable. (Also, we could allow a "soft margin" via slack parameters – something we are not covering.)

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

- Are linear separating hyperplanes enough?

Yes
(by a 1D-hyperplane
= dot)

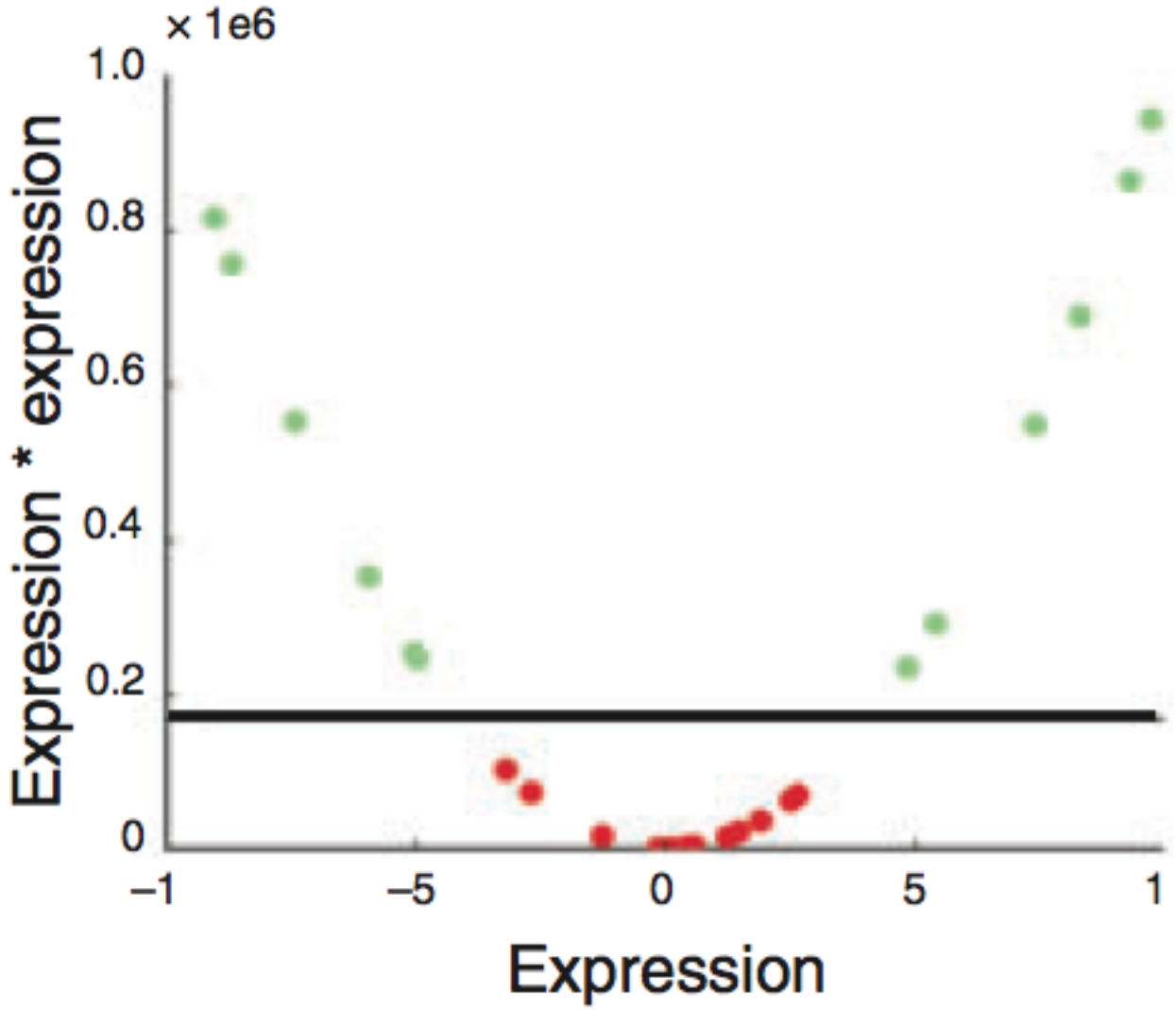




NO

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
  - 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",

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

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

Common Kernels

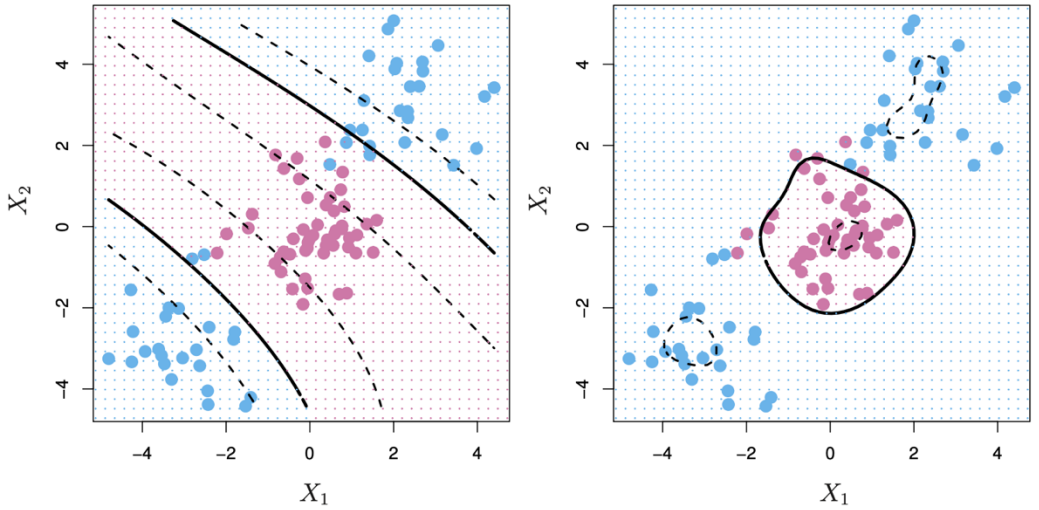


FIGURE 9.9. Left: *An SVM with a polynomial kernel of degree 3 is applied to the non-linear data from Figure 9.8, resulting in a far more appropriate decision rule.* Right: *An SVM with a radial kernel is applied. In this example, either kernel is capable of capturing the decision boundary.*

Radial Kernel K

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

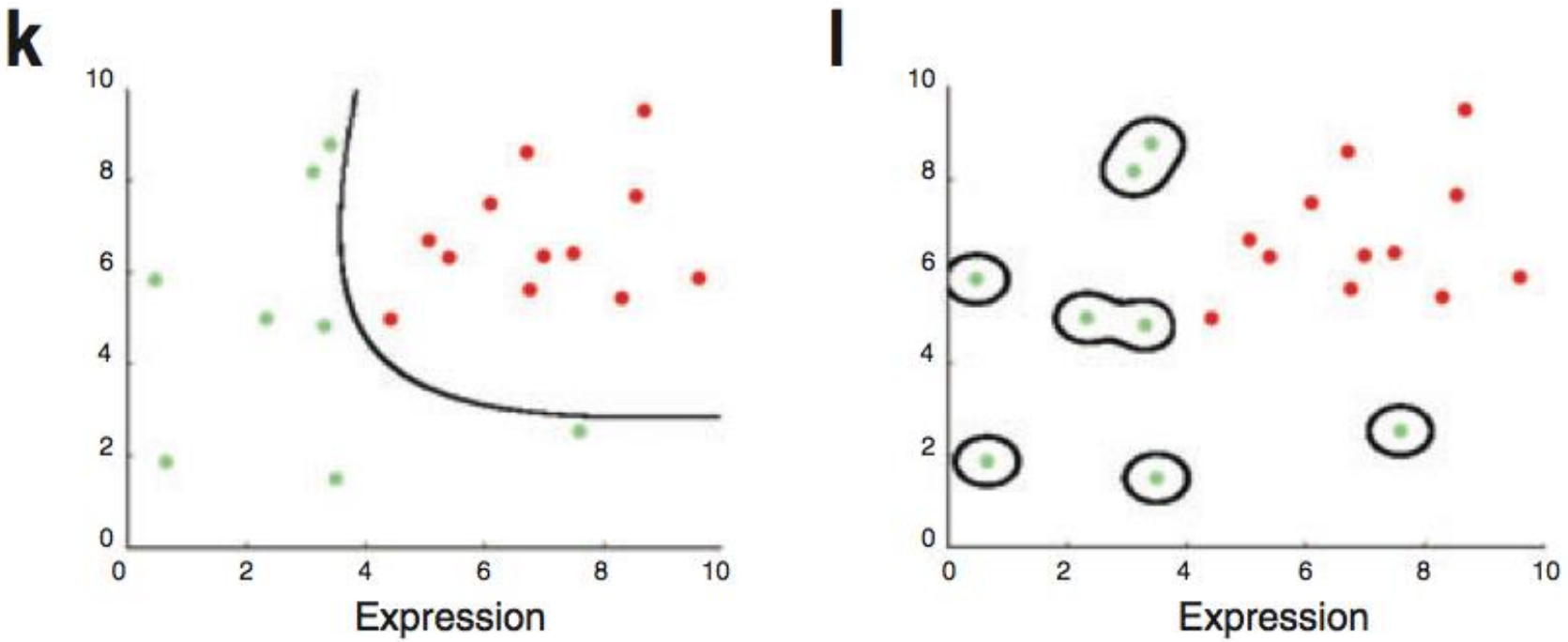
Polynomial Kernel K of degree d (d = 1 just gives a linear classifier)

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d.$$

[ISLR chap. 9]

- 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

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
  (Chapter 9 to 9.4 gives background on SVMs.
  Optional:
  Chapters 3.1 and 3.2 goes over linear regression and
  4.1 to 4.4 gives an overview of classification with logistic regression & LDA.)

- "An Idiot's guide to SVM"
  https://web.mit.edu/6.034/wwwbob/svm.pdf
  (Optional: extra background on SVMs)