# Classifiers!!!

**BCH394P/364C Systems Biology / Bioinformatics**

**Edward Marcotte, Univ of Texas at Austin**

1

**Clustering** = task of grouping a set of objects in such a way that objects in the same group (a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

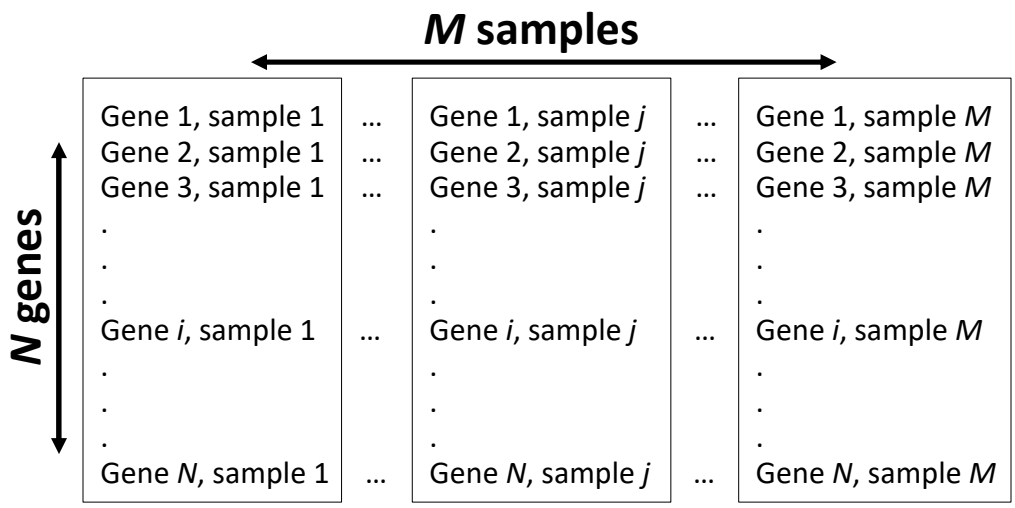
**VS.**

**Classification** = task of categorizing a new observation, on the basis of a training set of data with observations (or instances) whose categories are known

Adapted from Wikipedia

2

# Remember, for clustering, we had a matrix of data…

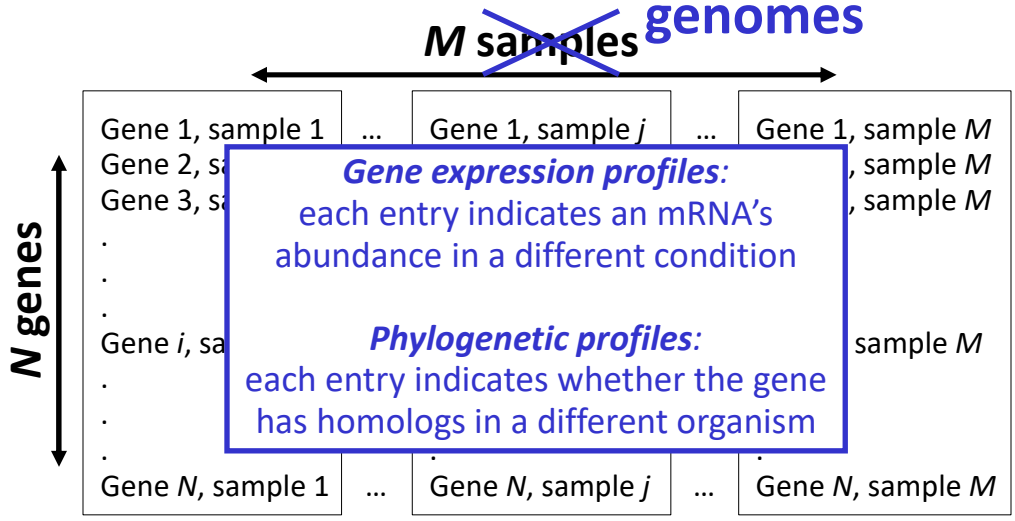



For yeast, N ~ 6,000
For human, N ~ 22,000

*i.e.,* a matrix of *N* x *M* numbers

3

# We discussed gene expression profiles. Here's another example of gene features.




***Gene expression profiles:***
each entry indicates an mRNA's abundance in a different condition

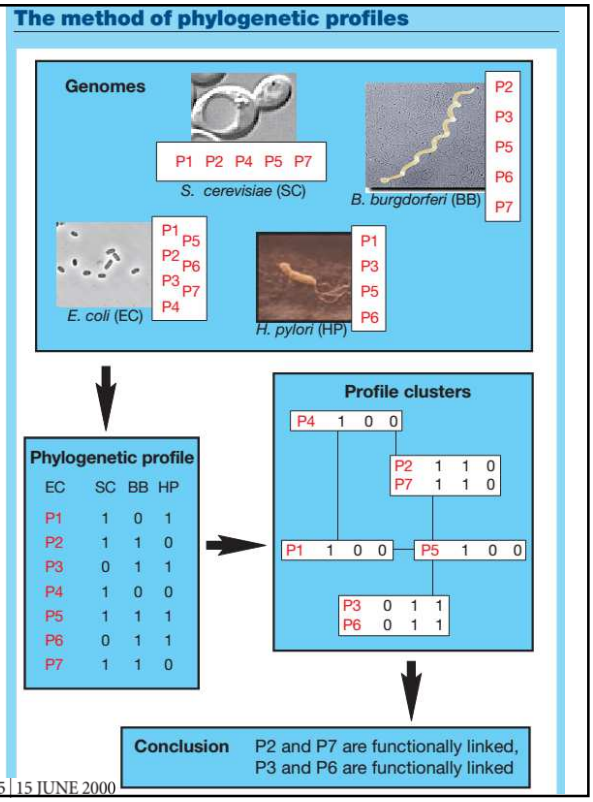
***Phylogenetic profiles:***
each entry indicates whether the gene has homologs in a different organism

For yeast, N ~ 6,000
For human, N ~ 22,000

4

**This is useful because biological systems tend to be modular and often inherited intact across evolution.**

**(e.g. you tend to have a flagellum or not)**




5

## Many such features are possible…

***M* samples**

***N* genes**

| | | |
|---|---|---|
| Gene 1, sample 1 | Gene 1, sample *j* | Gene 1, sample *M* |
| Gene 2, sample 1 | Gene 2, sample *j* | Gene 2, sample *M* |
| Gene 3, sample 1 | Gene 3, sample *j* | Gene 3, sample *M* |
| . | . | . |
| . | . | . |
| . | . | . |
| Gene *i*, sample 1 | Gene *i*, sample *j* | Gene *i*, sample *M* |
| . | . | . |
| . | . | . |
| . | . | . |
| Gene *N*, sample 1 | Gene *N*, sample *j* | Gene *N*, sample *M* |

For yeast, N ~ 6,000
For human, N ~ 22,000

*i.e.,* a matrix of *N* x *M* numbers

6

## We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in clustering.

| Names | Formula |
|---|---|
| Euclidean distance | $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$ |
| Manhattan distance | $\|a-b\|_1 = \sum_i \|a_i - b_i\|$ |
| cosine similarity | $\frac{a \cdot b}{\|a\|\|b\|}$ |

Wikipedia

7

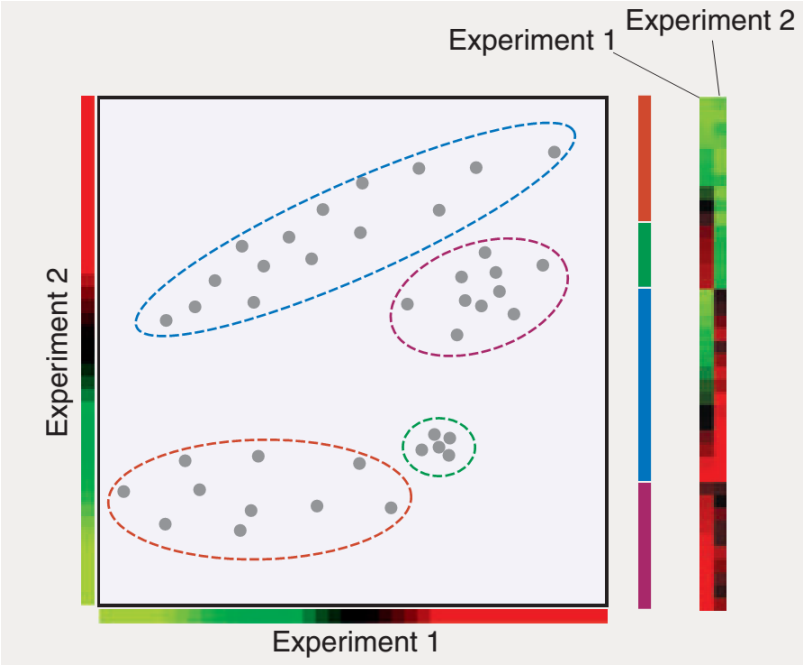
## We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in ~~clustering~~ classifying.

| Names | Formula |
|---|---|
| Euclidean distance | $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$ |
| Manhattan distance | $\|a-b\|_1 = \sum_i \|a_i - b_i\|$ |
| cosine similarity | $\frac{a \cdot b}{\|a\|\|b\|}$ |

Wikipedia

8

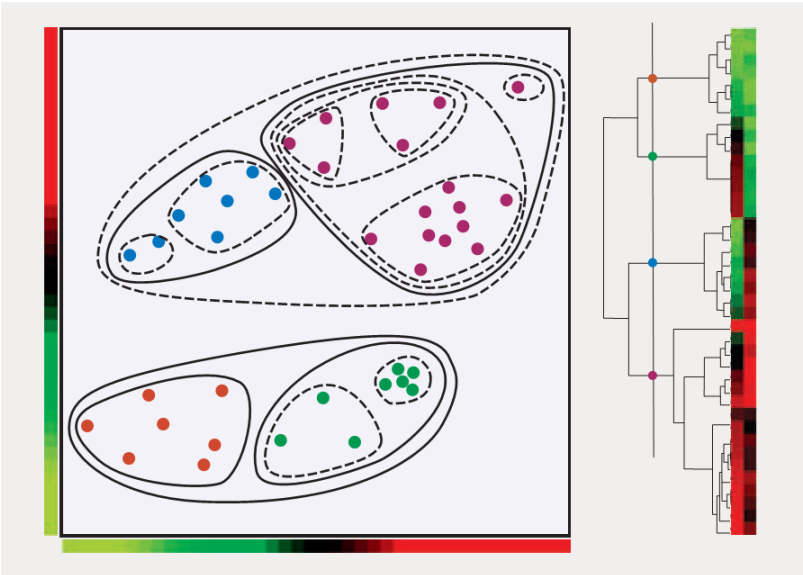
# Clustering refresher: 2-D example




*Nature Biotech* 23(12):1499-1501 (2005)

9

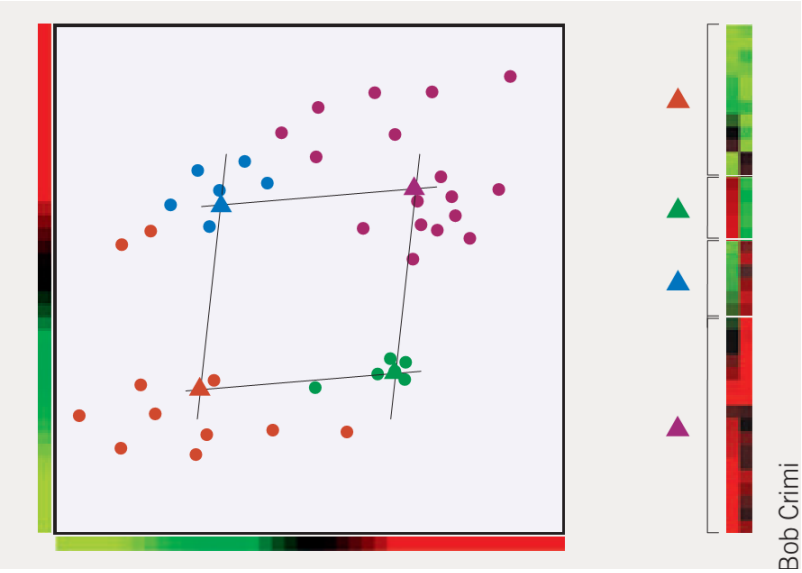
# Clustering refresher: hierarchical



*Nature Biotech* 23(12):1499-1501 (2005)

10

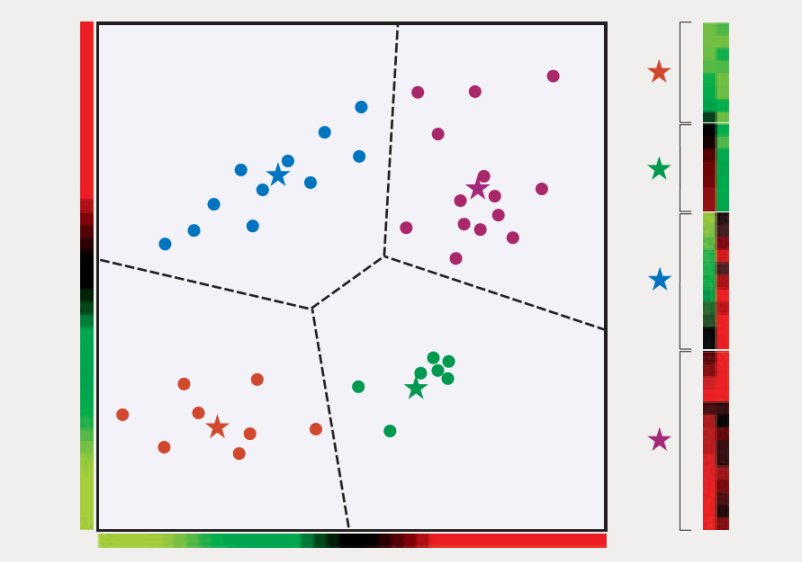
# Clustering refresher: SOM




*Nature Biotech* 23(12):1499-1501 (2005)

11

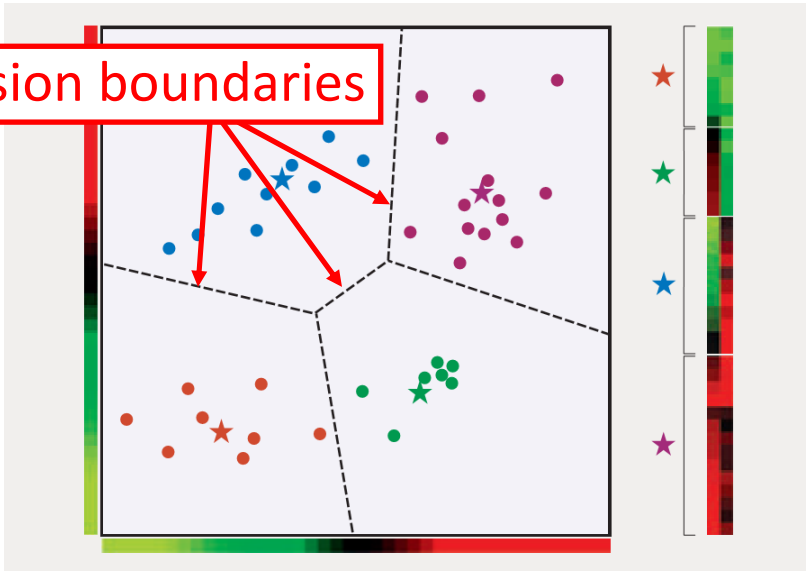
# Clustering refresher: *k*-means



*Nature Biotech* 23(12):1499-1501 (2005)

12

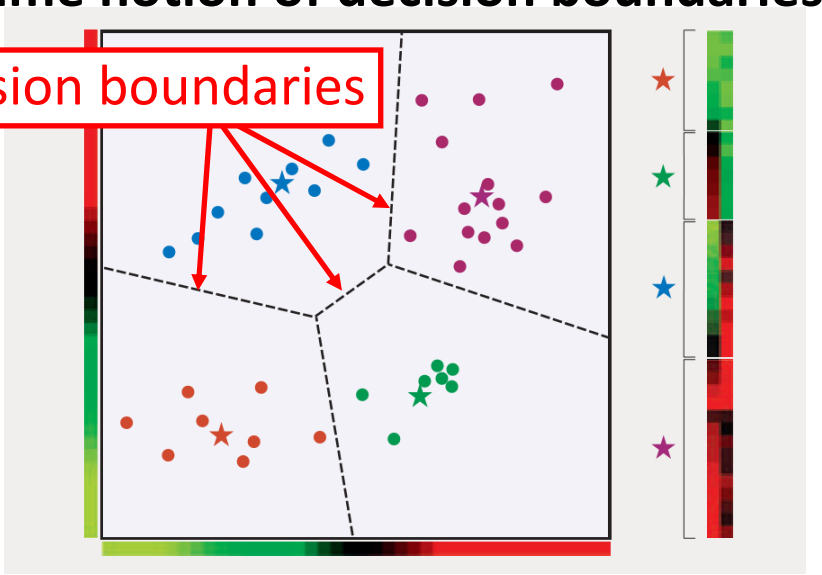
## Clustering refresher: *k*-means




*Nature Biotech* 23(12):1499-1501 (2005)

13

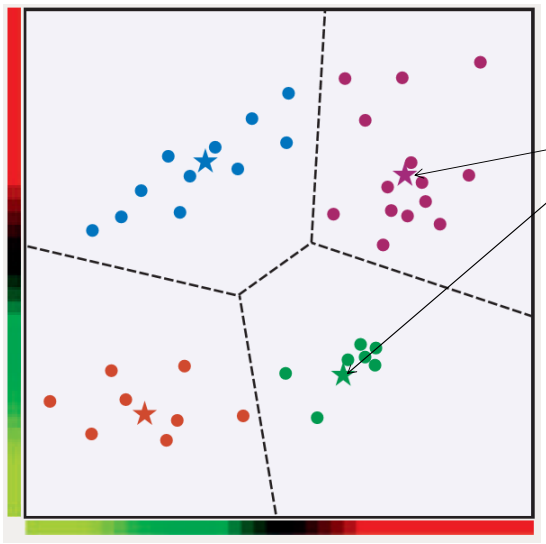
## One of the simplest classifiers uses the same notion of decision boundaries.




*Nature Biotech* 23(12):1499-1501 (2005)

14

# One of the simplest classifiers uses this notion of decision boundaries.



Rather than first clustering, calculate the centroid (mean) of objects with each label.

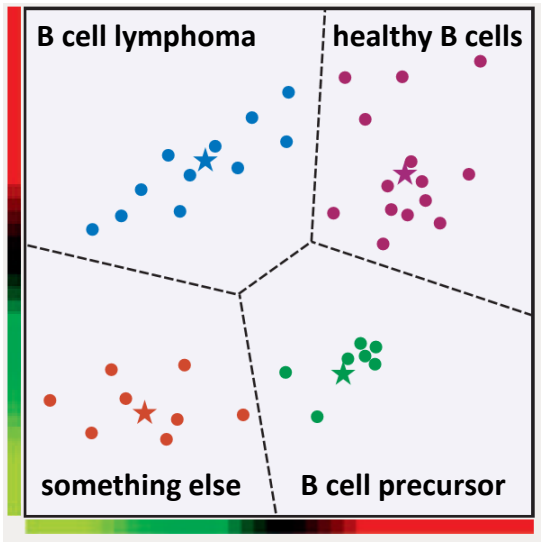
***New observations are classified as belonging to the group whose mean is nearest.***

**="minimum distance classifier"**

*Nature Biotech* 23(12):1499-1501 (2005)

15

# One of the simplest classifiers uses this notion of decision boundaries.




**For example….**

*Nature Biotech* 23(12):1499-1501 (2005)

16

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Let's look at a specific historic example:

"Enzyme-based histochemical analyses were introduced in the 1960s to demonstrate that **some leukemias were periodic acid-Schiff positive, whereas others were myeloperoxidase positive…**

This provided the first basis for classification of acute leukemias into those arising
from lymphoid precursors (acute lymphoblastic leukemia, ALL), or from myeloid precursors (acute myeloid leukemia, AML)."

15 OCTOBER 1999 VOL 286 SCIENCE

17

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Let's look at a specific historic example:

**"Distinguishing ALL from AML is critical for successful treatment…**

chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas

most AML regimens rely on a backbone of daunorubicin and cytarabine (8).

Although remissions can be achieved using ALL therapy for AML (and vice versa), **cure rates are markedly diminished**, and unwarranted toxicities are encountered."

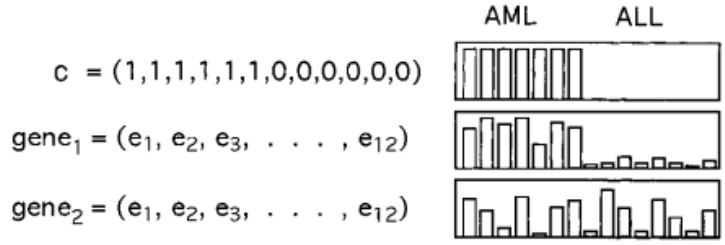
15 OCTOBER 1999 VOL 286 SCIENCE

18

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Let's look at a specific historic example:




Take labeled samples, find genes whose abundances separate the samples…

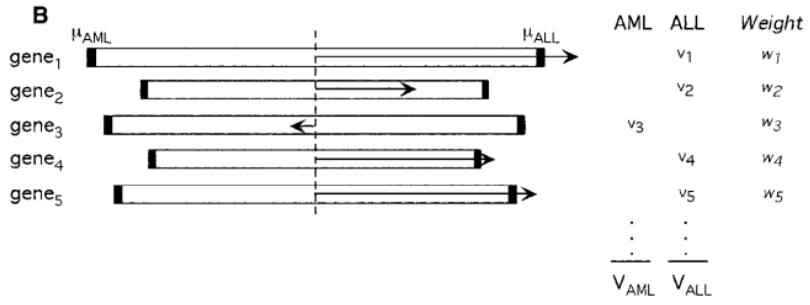


19

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Let's look at a specific historic example:




Calculate weighted average of indicator genes to assign class of an unknown



20

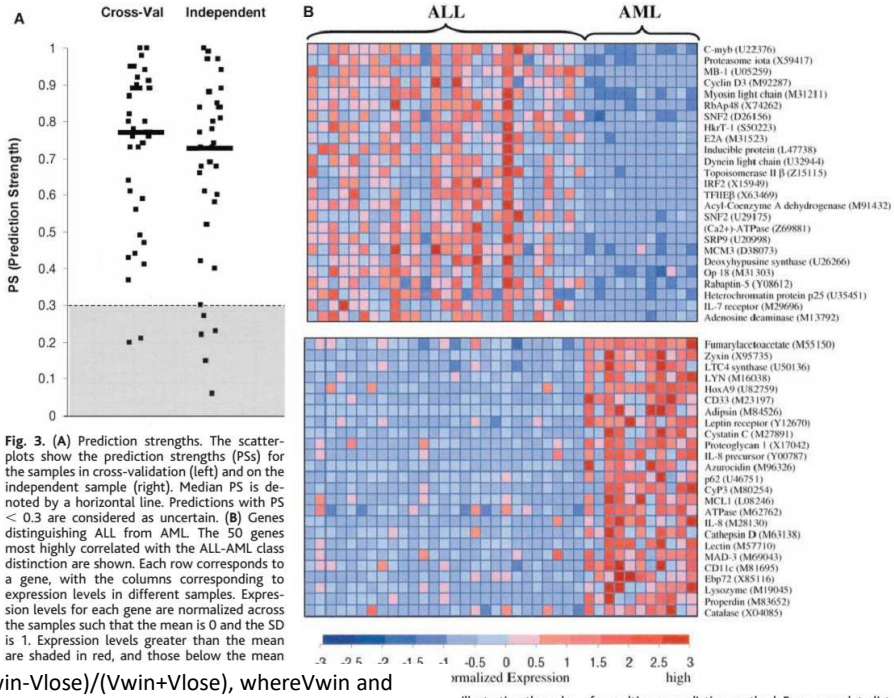



**Fig. 3.** (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.

PS=(Vwin-Vlose)/(Vwin+Vlose), whereVwin and VLose are the vote totals for the winning and losing classes.



21

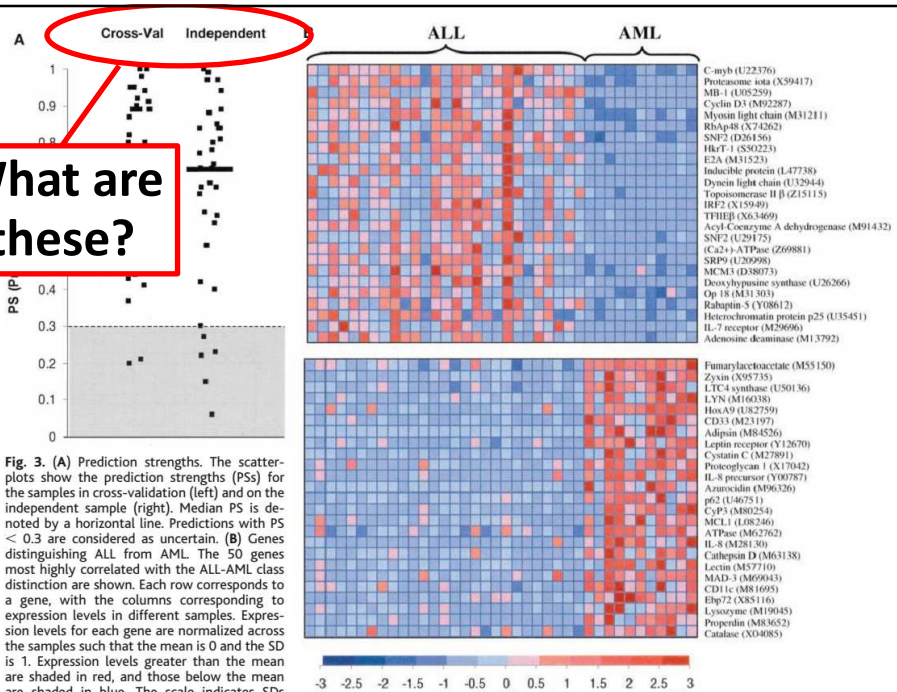



**Fig. 3.** (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.



22

## Cross-validation

Withhold a sample, build a predictor based only on the remaining samples, and predict the class of the withheld sample.

Repeat this process for each sample, then calculate the cumulative or average error rate.



23

## X-fold cross-validation
## e.g. 3-fold or 10-fold

Can also withhold 1/X (e.g. 1/3 or 1/10) of sample, build a predictor based only on the remaining samples, and predict the class of the withheld samples.

Repeat this process X times for each withheld fraction of the sample, then calculate the cumulative or average error rate.



24

## Independent data

Withhold an entire dataset, build a predictor based only on the remaining samples (the training data).

Test the trained classifier on the independent test data to give a fully independent measure of performance.

15 OCTOBER 1999 VOL 286 SCIENCE

25

You already know how to measure how well these algorithms work (way back in our discussion of gene finding!)…

| | | True answer: Positive | True answer: Negative |
|---|---|---|---|
| **Algorithm predicts:** | Positive | True positive | False positive |
| | Negative | False negative | True negative |

Specificity = TP / (TP + FP)

Sensitivity = TP / (TP + FN)

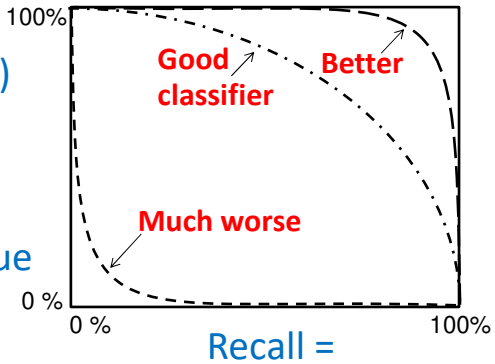
26

You already know how to measure how well these algorithms work (way back in our discussion of gene finding!)…

Sort the data by their classifier score, then step from best to worst and plot the performance:

Precision = TP / (TP + FP)

also called positive predictive value (PPV)




Recall = TP / (TP + FN) (= sensitivity)

Precision-recall curve

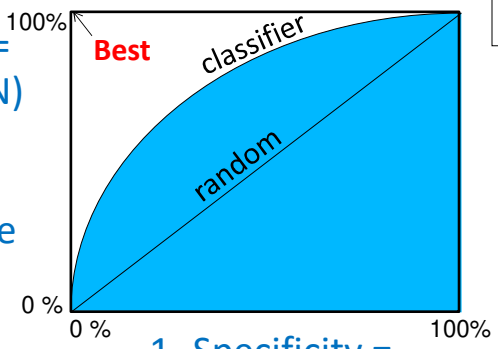
27

Another good option:

Sort the data by their classifier score, then step from best to worst and plot the performance:

Sensitivity = TP / (TP + FN)

also called True Positive Rate (TPR)




1- Specificity = FP / (FP + TN)

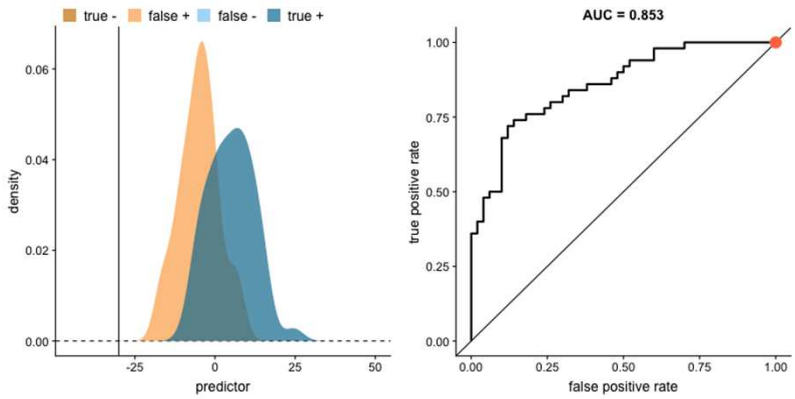
also called False Positive Rate (FPR)

First used in WWII to analyze radar signals (e.g., after attack on Pearl Harbor)

ROC curve (receiver operator characteristic)

28

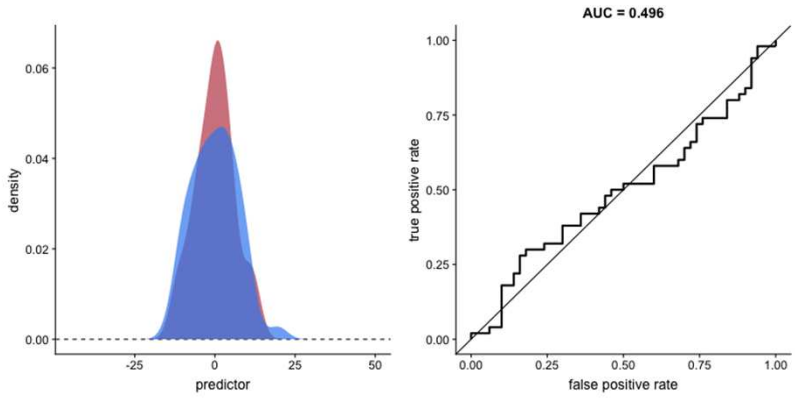
# ROC curve, as you go from stronger to weaker predictions




Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

29

# ROC curve, as you go from stronger to weaker classifiers

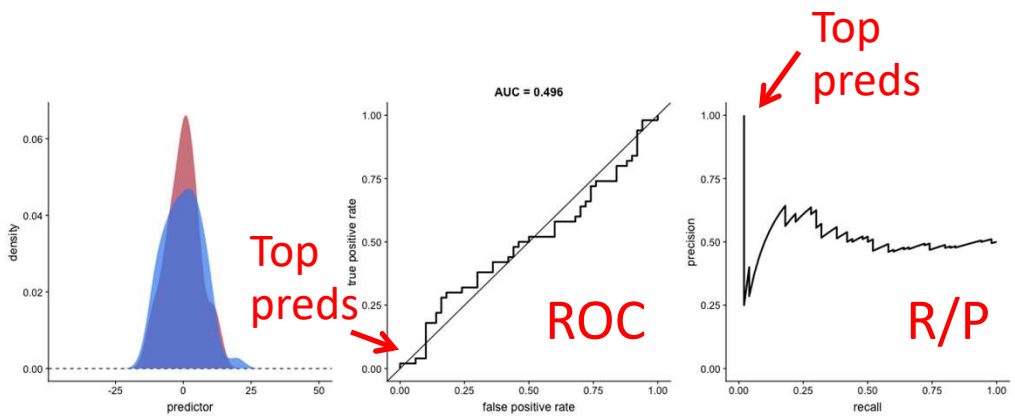



Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

30

# ROC versus Recall/Precision

The 2 measures are related and both useful. They differ strongly in performance as proportions of positive and negative classes change.



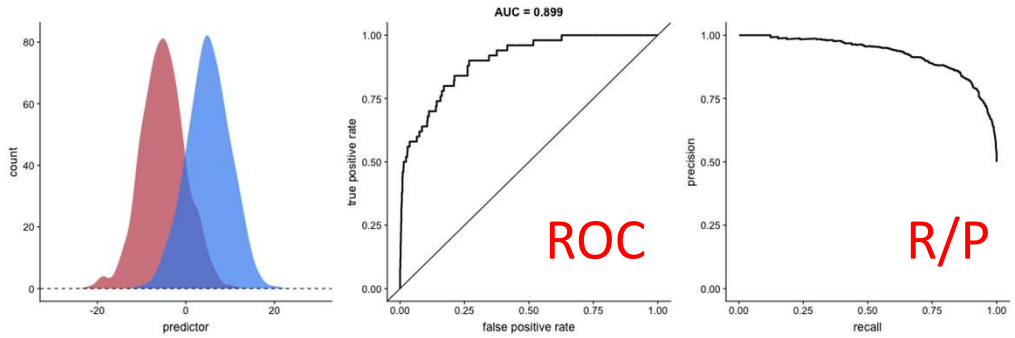

Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

31

# ROC versus Recall/Precision

- R/P depends <u>strongly</u> on relative rates of the 2 classes
- ROC performance is <u>independent</u> of their relative rates

(It may be important or not for your particular problem…)

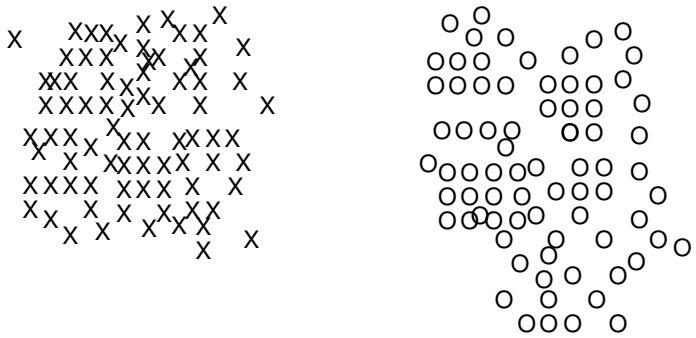



Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

32

Back to our minimum distance classifier…

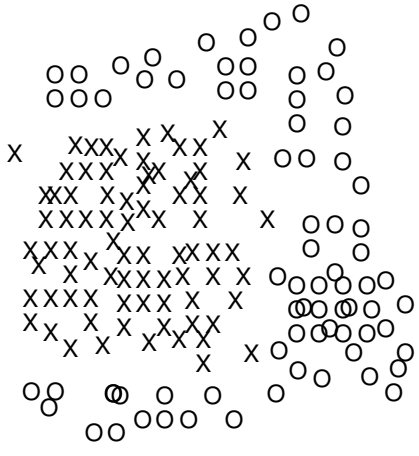
Would it work well for this data?



33

Back to our minimum distance classifier…

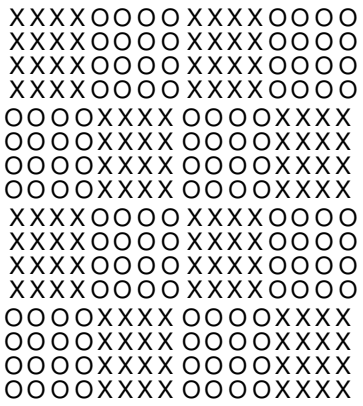
How about this data? What might?



34

Back to our minimum distance classifier…

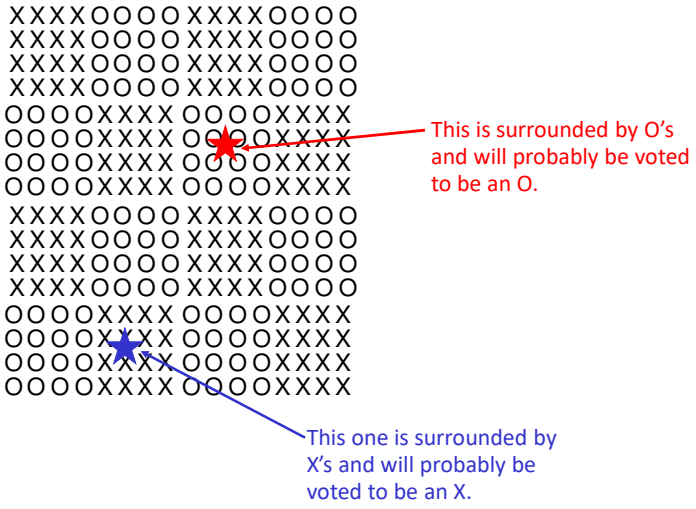
How about this data? What might?



35

This is a great case for something called a ***k-nearest neighbors classifier:***

**For each new object, calculate the *k* closest data points. Let them vote on the label of the new object.**




36

Back to leukemias. There was a follow-up study in 2010:

Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group

*Torsten Haferlach, Alexander Kohlmann, Lothar Wieczorek, Giuseppe Basso, Geertruy Te Kronnie, Marie-Christine Béné, John De Vos, Jesus M. Hernández, Wolf-Karsten Hofmann, Ken I. Mills, Amanda Gilkes, Sabina Chiaretti, Sheila A. Shurtleff, Thomas J. Kipps, Laura Z. Rassenti, Allen E. Yeoh, Peter R. Papenhausen, Wei-min Liu, P. Mickey Williams, and Robin Foà*

- **Tested clinical use of mRNA expression profiling to subtype leukemias into myeloid/lymphoid**
- **Meta-analysis of 11 labs, 3 continents, 3,334 patients**
- **Stage 1 (2,096 patients):**
  **92.2% classification accuracy for 18 leukemia classes (99.7% median specificity)**
- **Stage 2 (1,152 patients):**
  **95.6% median sensitivity and 99.8% median specificity for 14 subtypes of acute leukemia**
- **Microarrays outperformed routine diagnostics in 29 (57%) of 51 discrepant cases**

**Conclusion: "Gene expression profiling is a robust technology for the diagnosis of hematologic malignancies with high accuracy"**

*J Clin Oncol 28:2529-2537. © 2010*

37

In practice, if you want to explore classifiers, I also strongly recommend always testing these classifiers:

**Random forests**
**Support vector machines (SVM)**

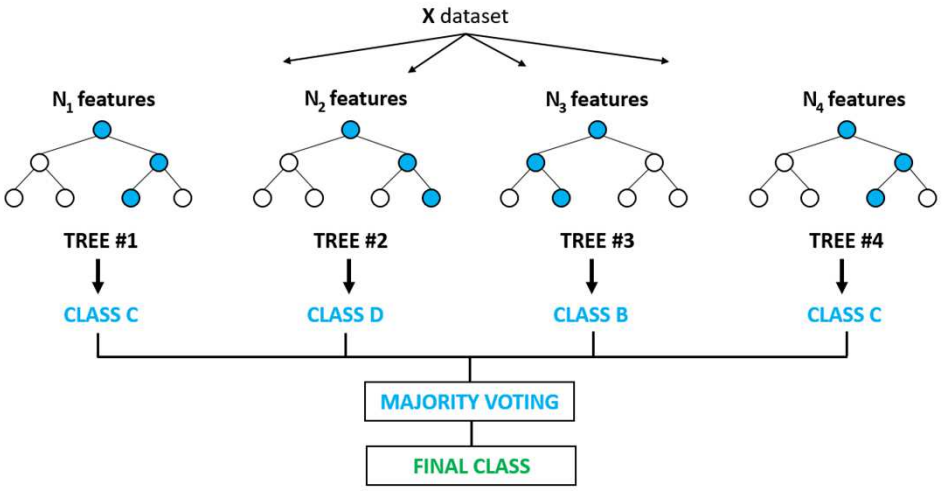
These two are surprisingly often the best for many biological classification problems. Weka can do both of them.

38

The two slide overview of **Random forest classifiers:**

(1) Construct many decision trees from random subsets of your features. Because the features vary across trees, trees tend to be weak but uncorrelated

(2) All the trees "vote" on the answer, majority wins.



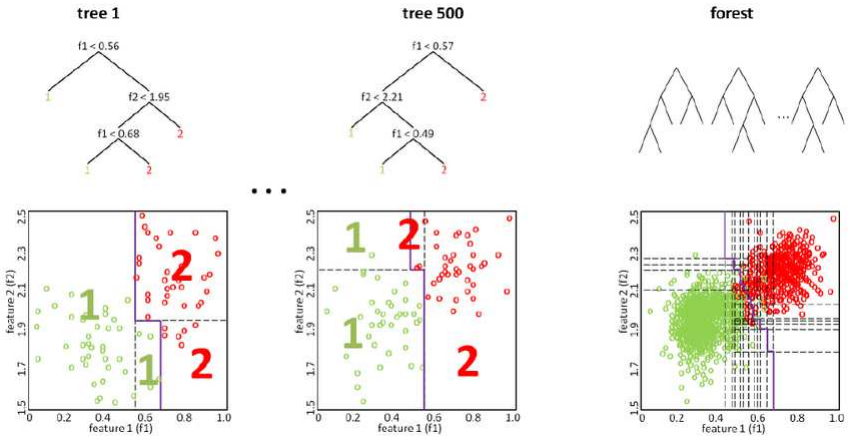

https://www.globalsoftwaresupport.com/random-forest-classifier-bagging-machine-learning/

39

The two slide overview of **Random forest classifiers:**

(1) Construct many decision trees from random subsets of your features. Because the features vary across trees, trees tend to be weak but uncorrelated

(2) All the trees "vote" on the answer, majority wins.

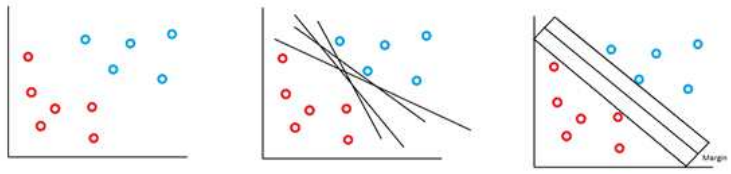



https://www.researchgate.net/figure/The-Random-Forest-classifier-is-an-ensemble-of-decision-trees-where-the-single-trees-are_fig1_228540194

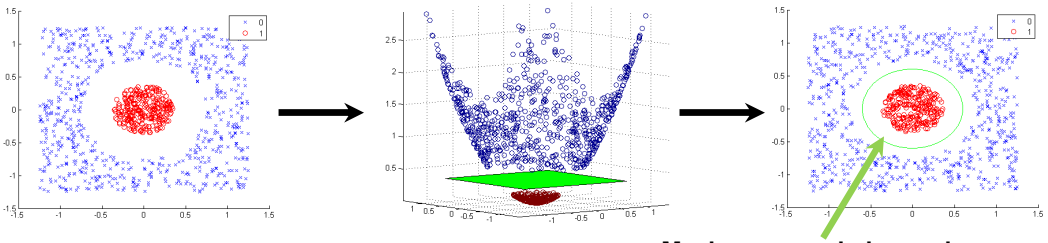
40

The one slide overview of **Support vector machines:**

(1) Goal: make a linear classifier, choosing a decision boundary that *maximizes* the *distance margin* between classes



(2) But what if the boundary is non-linear? Use ***kernels*** to implicitly map the data to higher dimension where a linear decision can be made




41

In practice, if you want to explore classifiers, I <u>strongly</u> recommend the Weka package:

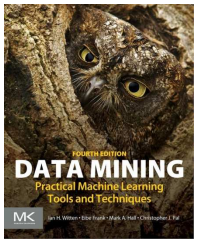
http://www.cs.waikato.ac.nz/ml/weka/


It's free, and easy to install, use, & troubleshoot. It lets you quickly test many alternative (well-vetted) classifiers, all in a proper cross-validated/precision-recall framework.

Here's a nice step-by-step intro for biologists :
**Introducing Machine Learning Concepts with WEKA**, in *Statistical Genomics, Methods in Molecular Biology,* v. 1418, p. 353-378, 24 March 2016

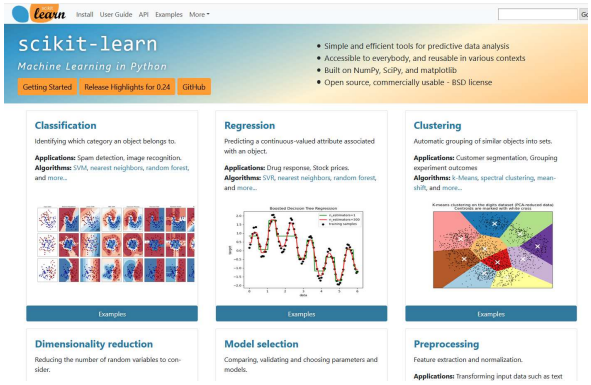
**http://link.springer.com/content/pdf/10.1007%2F978-1-4939-3578-9_17.pdf**

There's also a great book to walk you through the entire process. Highly recommended!!!




42

In Python, you can also use the scikit-learn library:
https://scikit-learn.org/stable/
Like Weka, it's free, easy to install & use, and very powerful




I recommend combining it with the Pandas library for data analysis to make it easy to work with big, tabular datasets:
https://pandas.pydata.org/

pandas

43