Classifiers!!!

BCH339N Systems Biology / Bioinformatics

Edward Marcotte, Univ of Texas at Austin

Clustering = task of <u>grouping</u> 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).
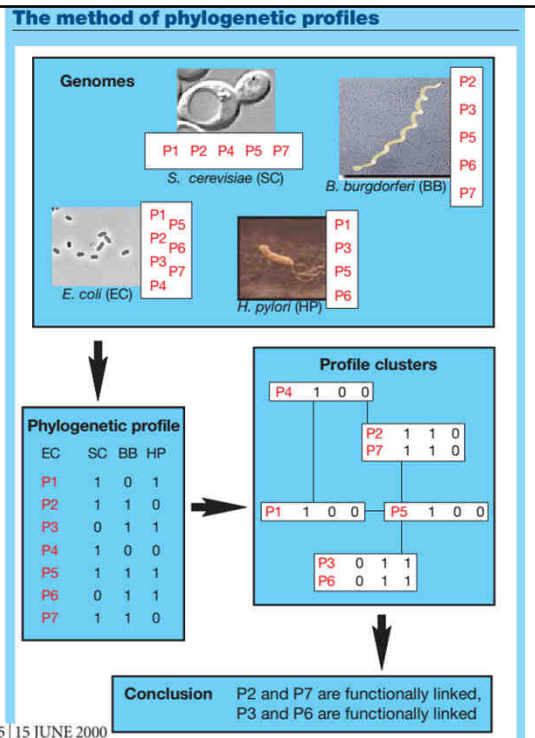
VS.

Classification = task of <u>categorizing</u> a new observation, on the basis of a training set of data with observations (or instances) whose categories are known
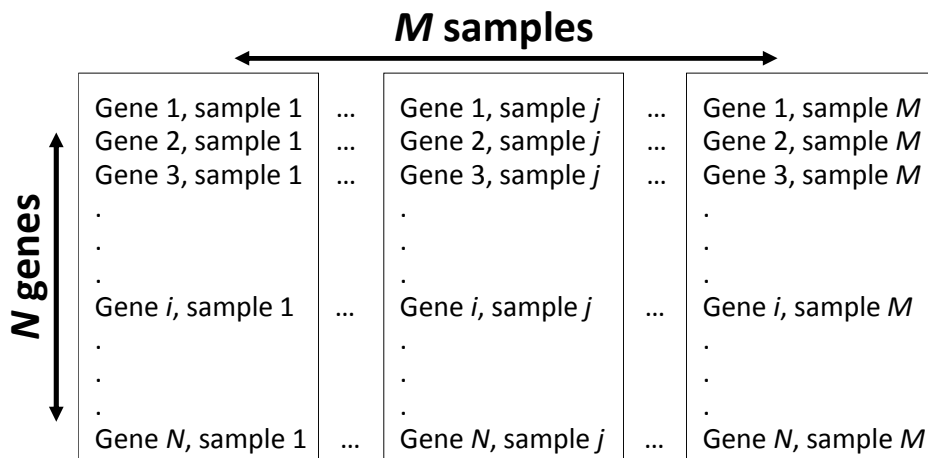
Adapted from Wikipedia

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

_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

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

genomes

_M_ samples

_N_ genes

Gene 1, sample 1	…	Gene 1, sample _j_	…	Gene 1, sample _M_
Gene 2, sa				, sample _M_
Gene 3, sa				, sample _M_
.				
.				
.				
Gene _i_, sa				sample _M_
.				
.				
.				
Gene _N_, sample 1	…	Gene _N_, sample _j_	…	Gene _N_, sample _M_

> *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

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)

The method of phylogenetic profiles

Genomes

P1 P2 P4 P5 P7
S. cerevisiae (SC)

B. burgdorferi (BB)
P2
P3
P5
P6
P7

E. coli (EC)
P1
P2
P3
P4
P5
P6
P7

H. pylori (HP)
P1
P3
P5
P6

Phylogenetic profile

	EC	SC	BB	HP
P1	1	0	1	
P2	1	1	0	
P3	0	1	1	
P4	1	0	0	
P5	1	1	1	
P6	0	1	1	
P7	1	1	0	

Profile clusters

P4 1 0 0

P2 1 1 0
P7 1 1 0

P1 1 0 0 — P5 1 0 0

P3 0 1 1
P6 0 1 1

Conclusion   P2 and P7 are functionally linked, P3 and P6 are functionally linked

NATURE | VOL 405 | 15 JUNE 2000

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

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	$\dfrac{a \cdot b}{\|a\|\|b\|}$

Wikipedia

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	$\dfrac{a \cdot b}{\|a\|\|b\|}$

Wikipedia

Clustering refresher: 2-D example
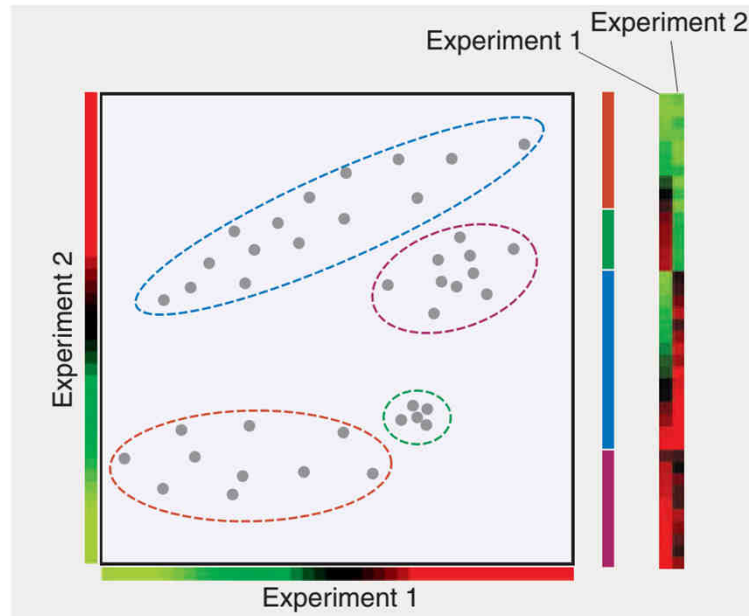
Clustering refresher: hierarchical
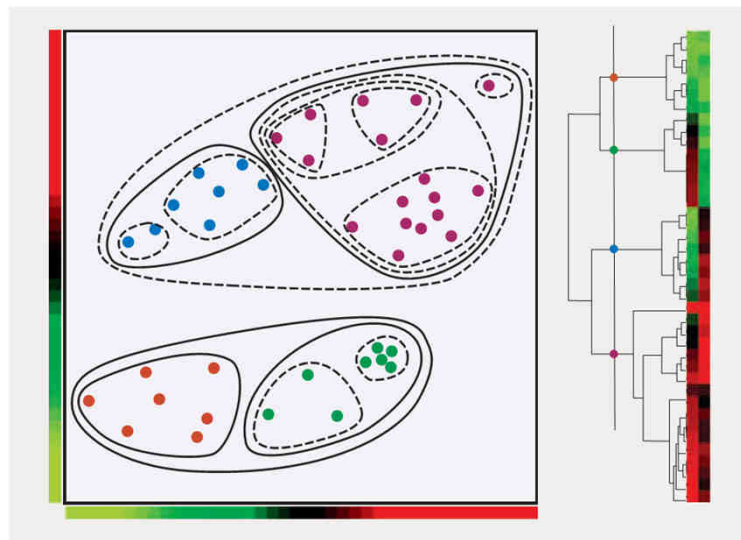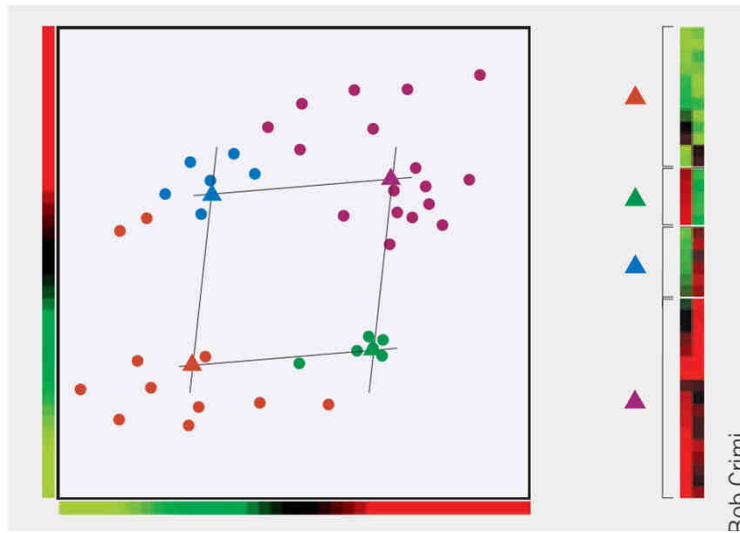
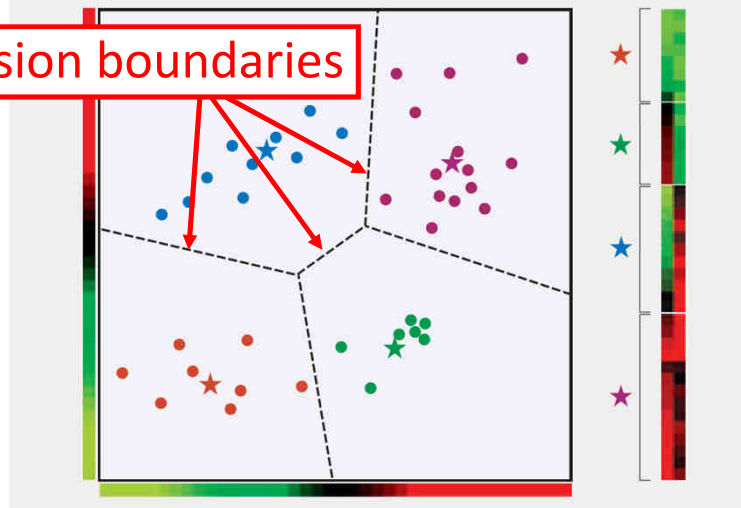# Clustering refresher: SOM
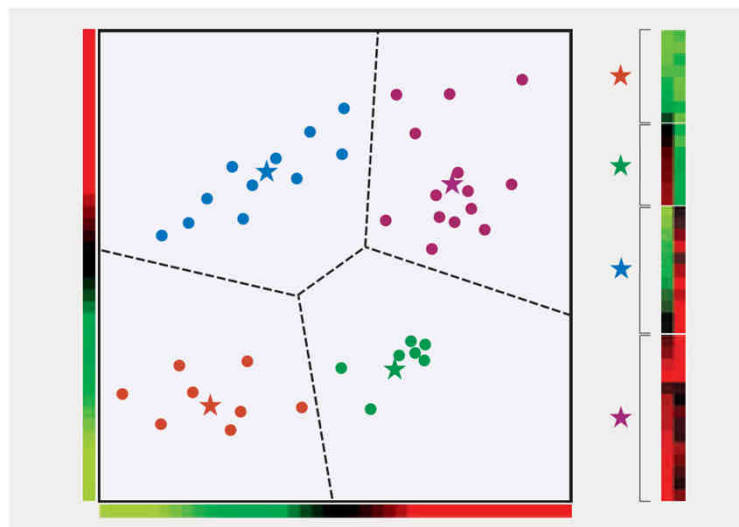
Clustering refresher: *k*-means

Clustering refresher: *k*-means

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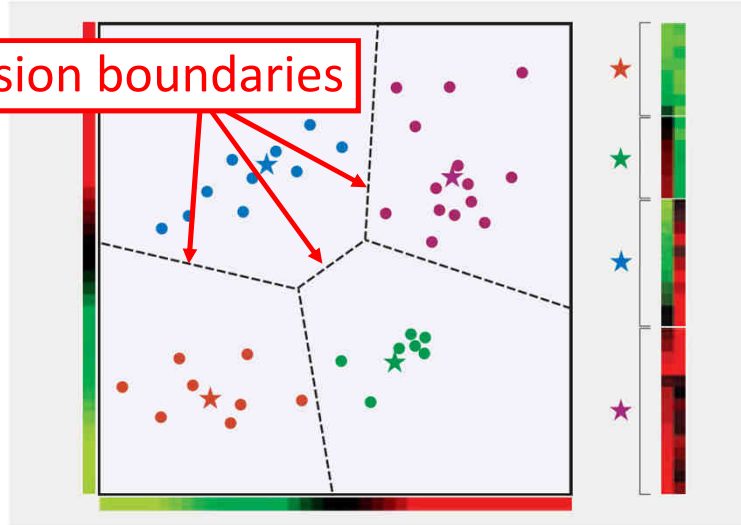
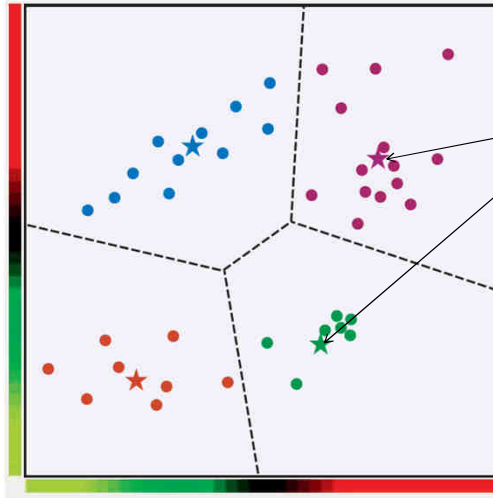# One of the simplest classifiers uses the same notion of decision boundaries.

Decision boundaries

# 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"

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

B cell lymphoma   **healthy B cells**

something else   **B cell precursor**
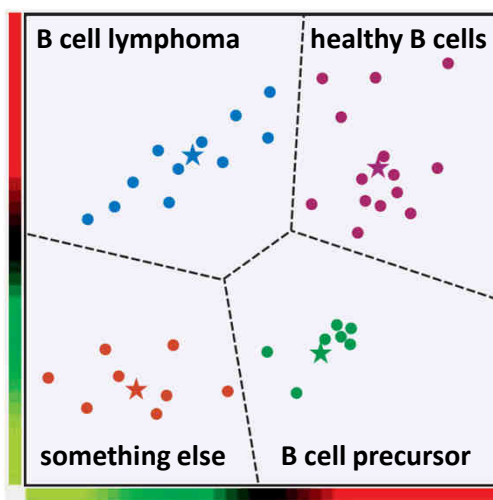
For example….

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 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 <u>lymphoid</u> precursors (acute lymphoblastic leukemia, ALL), or from <u>myeloid</u> precursors (acute myeloid leukemia, AML)."

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 example:

"Distinguishing ALL from AML is critical for successful treatment…

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

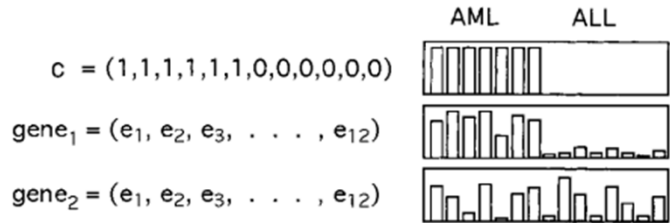most AML regimens rely on a backbone of daunorubicin and cytarabine (8).

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

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 example:

$$c = (1,1,1,1,1,1,0,0,0,0,0,0)$$

$$gene_1 = (e_1, e_2, e_3, \ldots, e_{12})$$

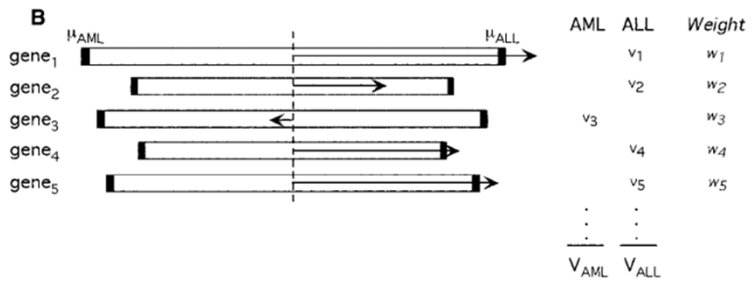$$gene_2 = (e_1, e_2, e_3, \ldots, e_{12})$$

AML   ALL

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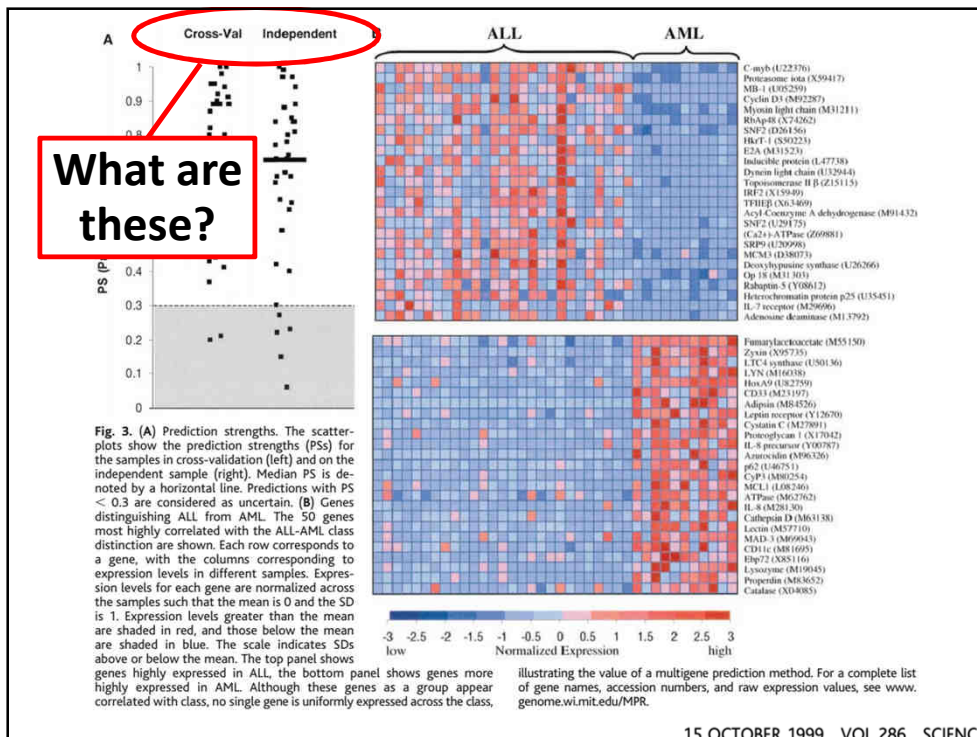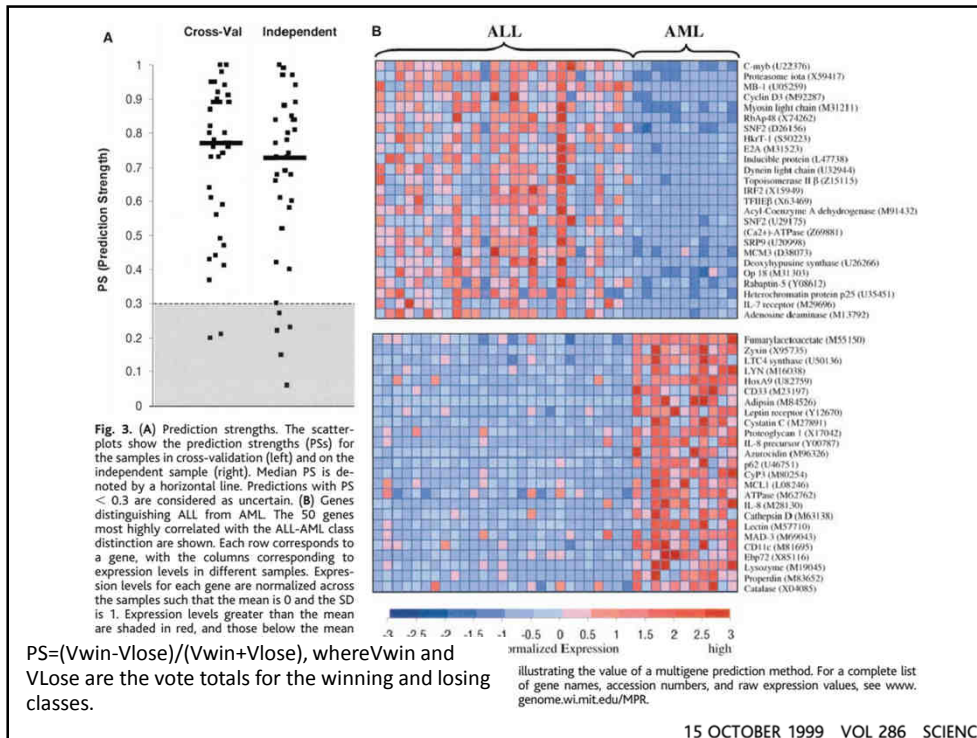

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

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 example:

B

	μ_{AML}	μ_{ALL}	AML	ALL	Weight
gene$_1$				v_1	w_1
gene$_2$				v_2	w_2
gene$_3$			v_3		w_3
gene$_4$				v_4	w_4
gene$_5$				v_5	w_5

V_{AML} V_{ALL}

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

Fig. 3. (A) Prediction strengths. The scatter-plots 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.

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

What are these?

Fig. 3. (A) Prediction strengths. The scatter-plots 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.

11

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.

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.

Independent data

Withhold <u>an entire dataset</u>, build a predictor based only on the remaining samples
<p style="text-align:center">(the training data).</p>

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

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

True answer:

	Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

Algorithm predicts:

Specificity = TP / (TP + FP)

Sensitivity = TP / (TP + FN)

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:

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

Sensitivity =
TP / (TP + FN)

also called
True Positive
Rate (TPR)

100%

Best   classifier

random

0 %

0 %   100%

1- Specificity =
FP / (FP + TN)
also called False Positive Rate (FPR)

ROC curve
(receiver operator characteristic)

Another good option:

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)

100%

Good classifier   **Better**

Much worse

0 %

0 %   100%

Recall =
TP / (TP + FN)
(= sensitivity)
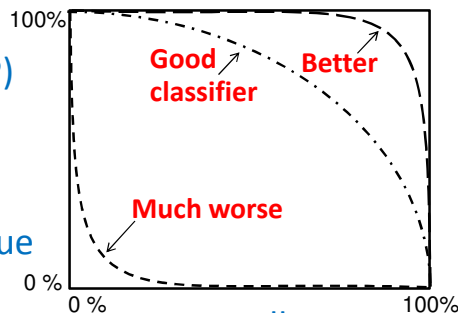
Precision-recall curve

Back to our minimum distance classifier…

Would it work well for this data?
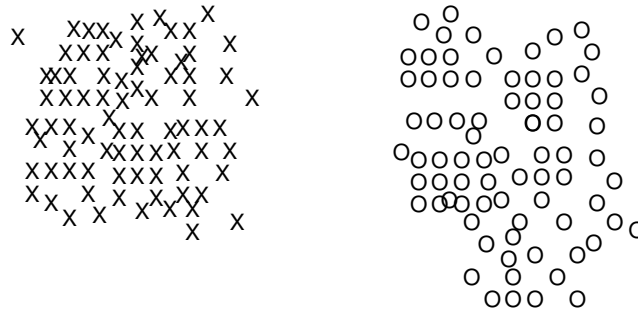
Back to our minimum distance classifier…

How about this data? What might?
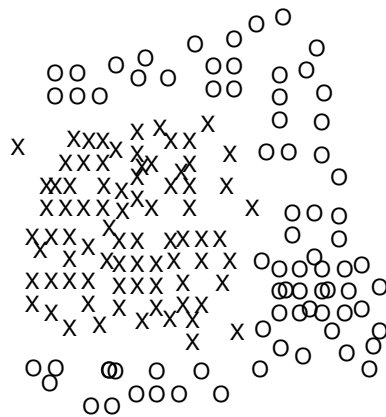
Back to our minimum distance classifier…

How about this data? What might?

```
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
```

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.**

```
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXXOOOOXXXX
OOOOXXXXO★OOXXXX        This is surrounded by O's
OOOOXXXXOOOOXXXX        and will probably be voted
OOOOXXXXOOOOXXXX        to be an O.
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXXOOOOXXXX
OOOOX★XXOOOOXXXX
OOOOXXXXOOOOXXXX
OOOOXXXXOOOOXXXX
```

This one is surrounded by
X's and will probably be
voted to be an X.

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"

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!!!
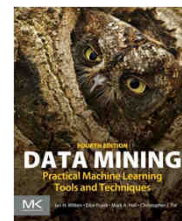
In practice, if you want to explore classifiers, I also <u>strongly</u> recommend always testing these classifiers:

Random forests
Support vector machines (SVM)

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