Classifiers!!!

BCH364C/391L Systems Biology / Bioinformatics – Spring 2015

Edward Marcotte, Univ of Texas at Austin

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

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.

***M* ~~samples~~ genomes**

***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*

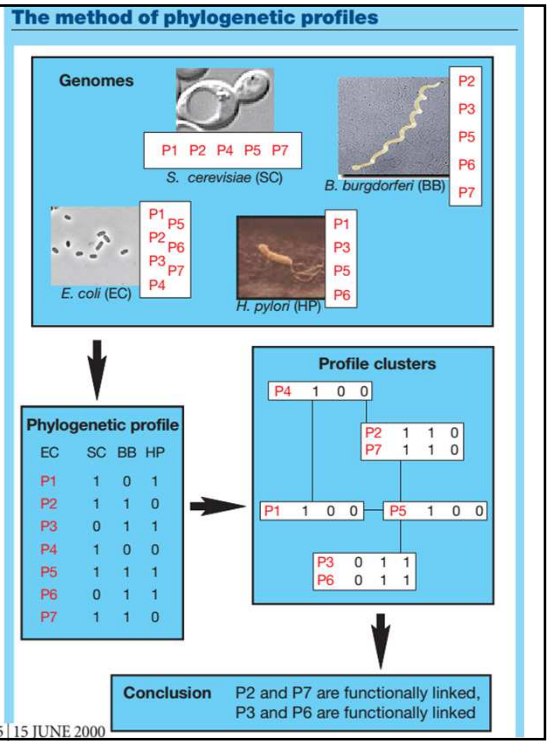
***Gene expression profiles*:**
each entry indicates an mRNA's abundance in a different condition

***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)



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

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

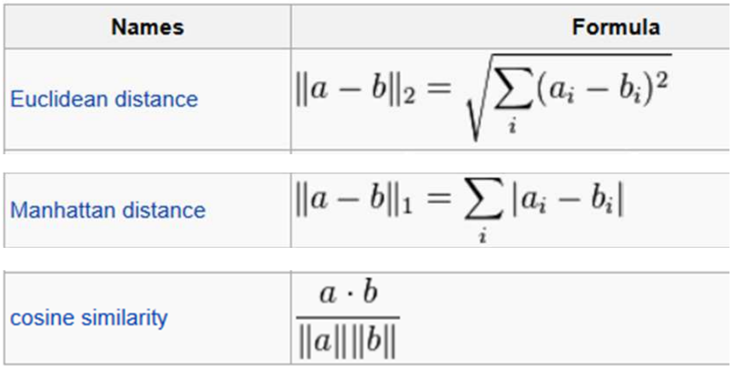
***N* genes**

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

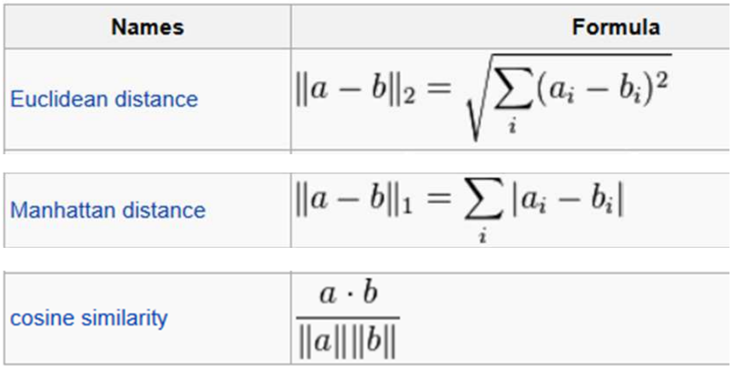
We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in clustering.



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

Wikipedia

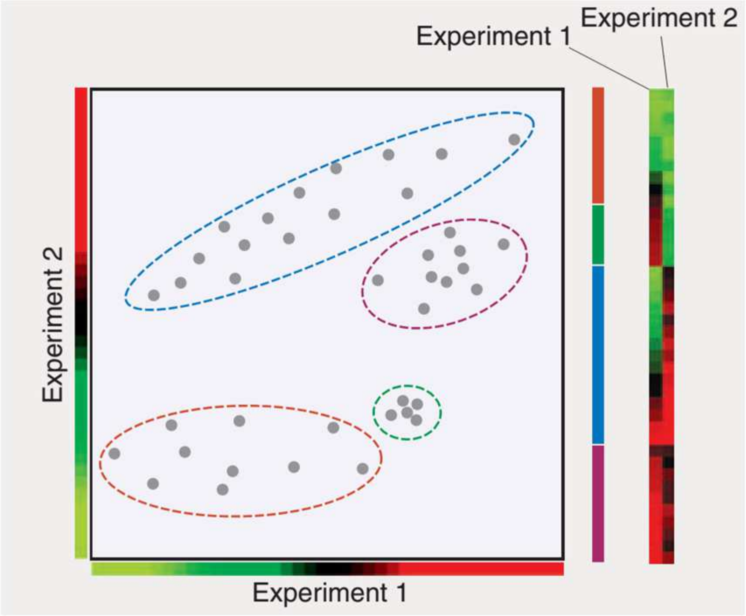
We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in ~~clustering~~. classifying



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

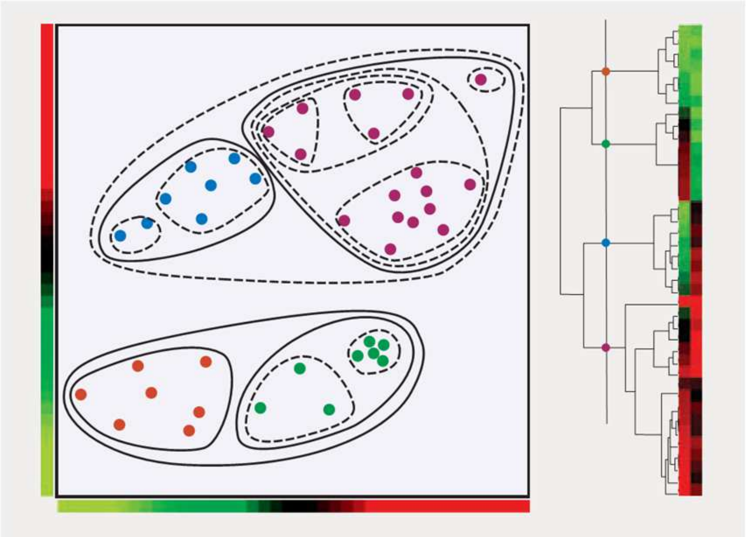
Wikipedia

Clustering refresher: 2-D example



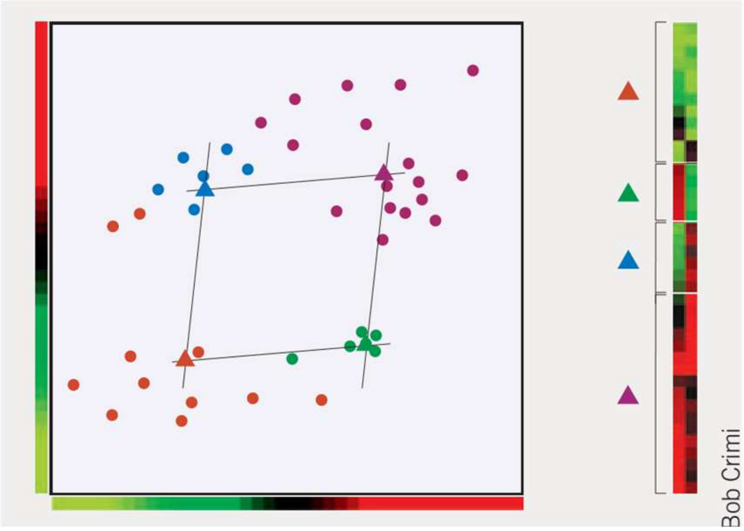
Nature Biotech 23(12):1499-1501 (2005)

Clustering refresher: hierarchical



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

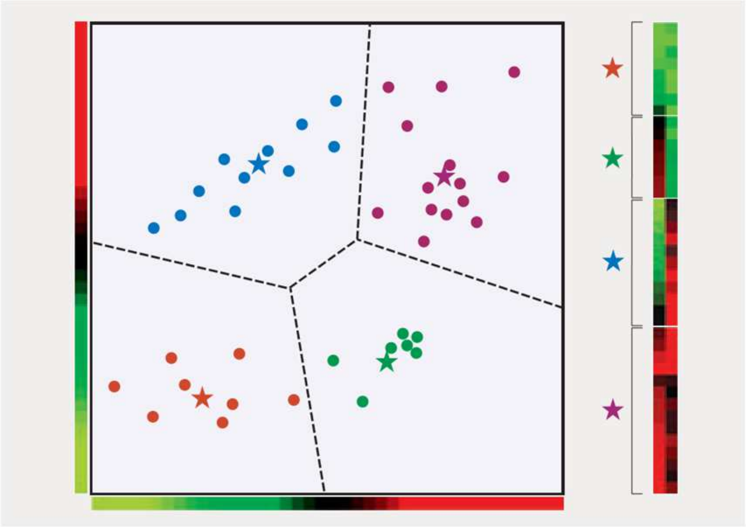
Clustering refresher: SOM



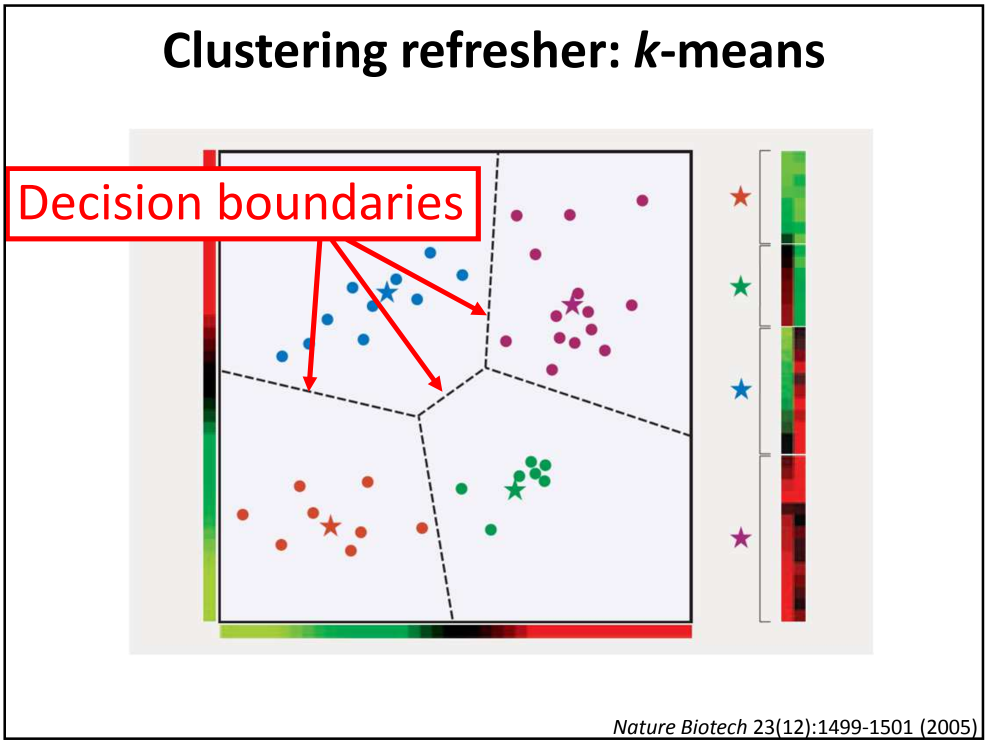
Bob Crimi

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

Clustering refresher: *k*-means



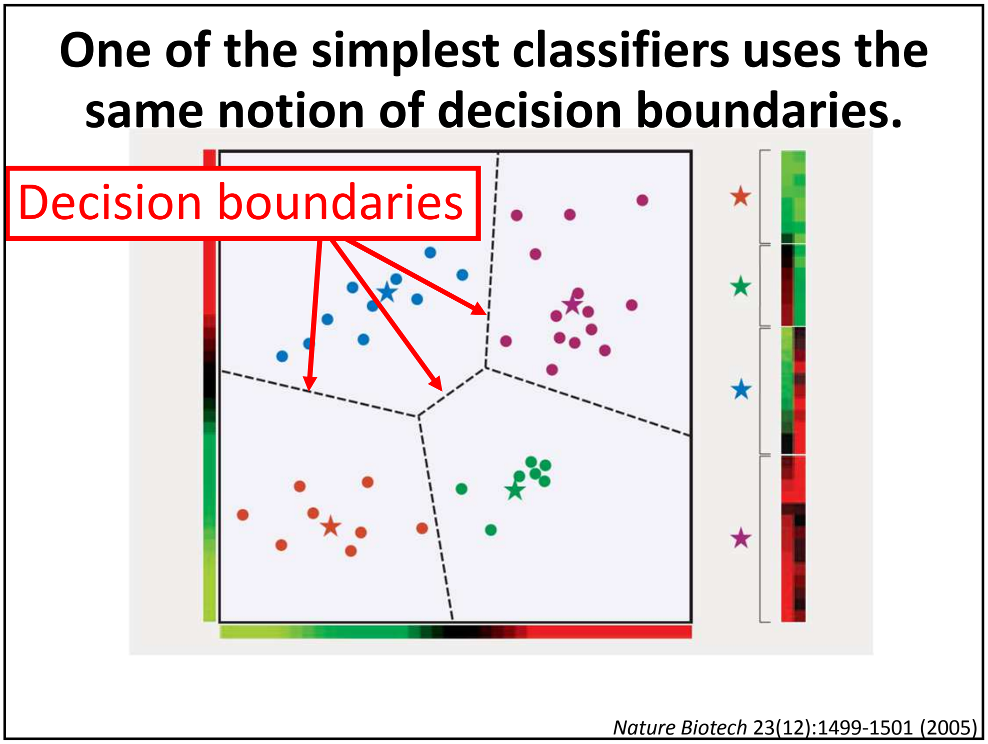
Nature Biotech 23(12):1499-1501 (2005)



Clustering refresher: *k*-means

Decision boundaries

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

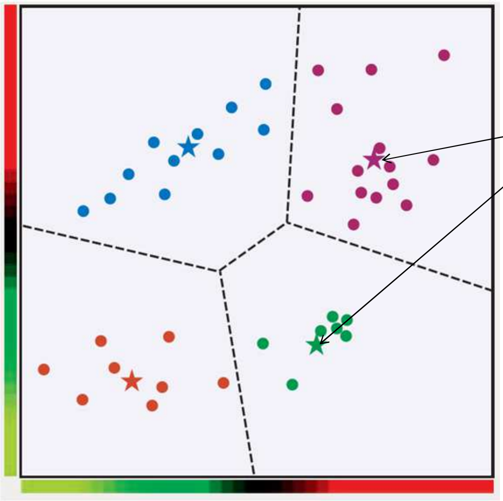


One of the simplest classifiers uses the same notion of decision boundaries.

Decision boundaries

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

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



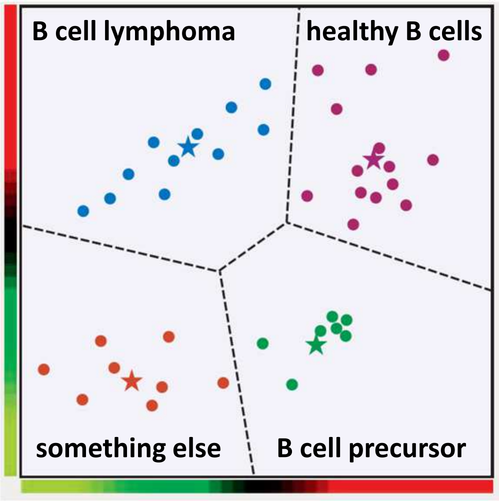
Rather than first clustering, calculate the centroid (mean) of objects with each label.

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

="minimum distance classifier"

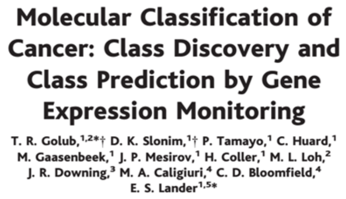
Nature Biotech 23(12):1499-1501 (2005)

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



For example….

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



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

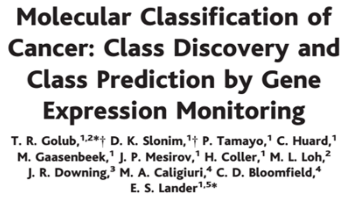
T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

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 lymphoid precursors (acute lymphoblastic leukemia, ALL), or from myeloid precursors (acute myeloid leukemia, AML)."

15 OCTOBER 1999 VOL 286 SCIENCE



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:

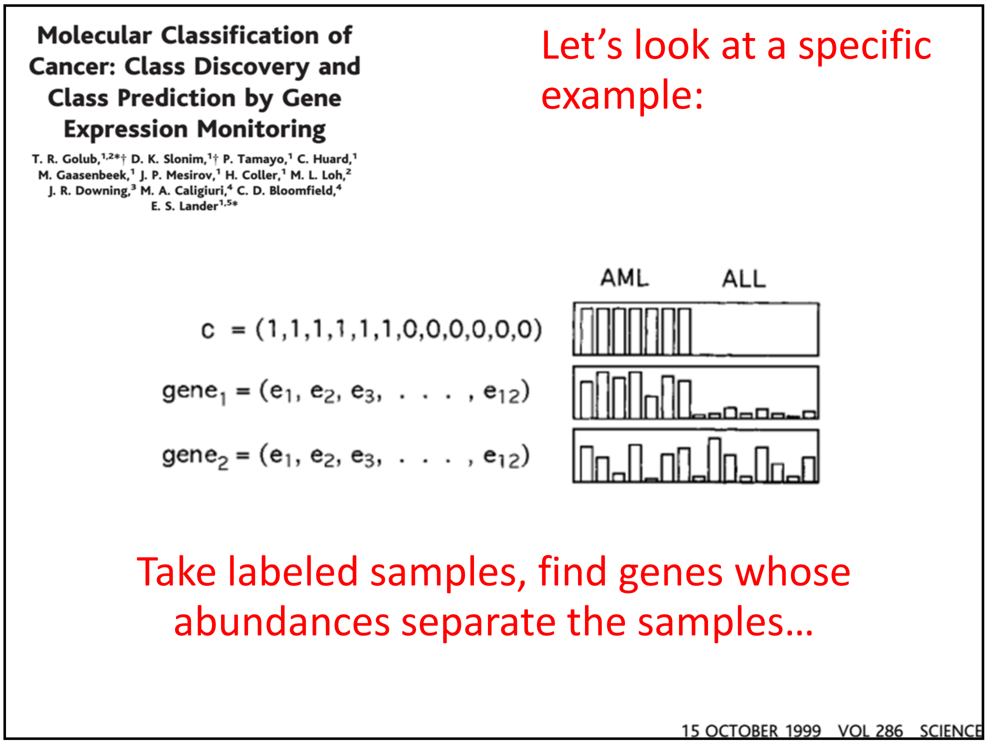
"Distinguishing ALL from AML is critical for successful treatment…

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), **cure rates are markedly diminished**, and unwarranted toxicities are encountered."

15 OCTOBER 1999 VOL 286 SCIENCE



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

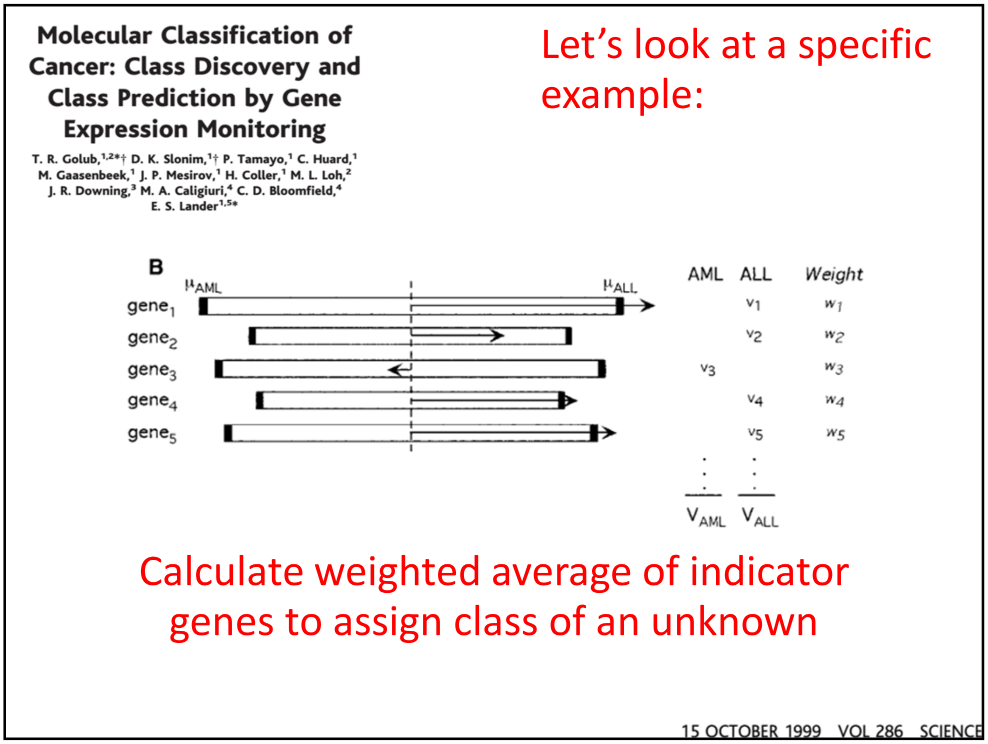
T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Let’s look at a specific example:

AML
ALL
c = (1,1,1,1,1,1,0,0,0,0,0,0)
gene$_1$ = (e_1, e_2, e_3, ... , e_{12})
gene$_2$ = (e_1, e_2, e_3, ... , e_{12})

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

15 OCTOBER 1999 VOL 286 SCIENCE



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:

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

15 OCTOBER 1999 VOL 286 SCIENCE

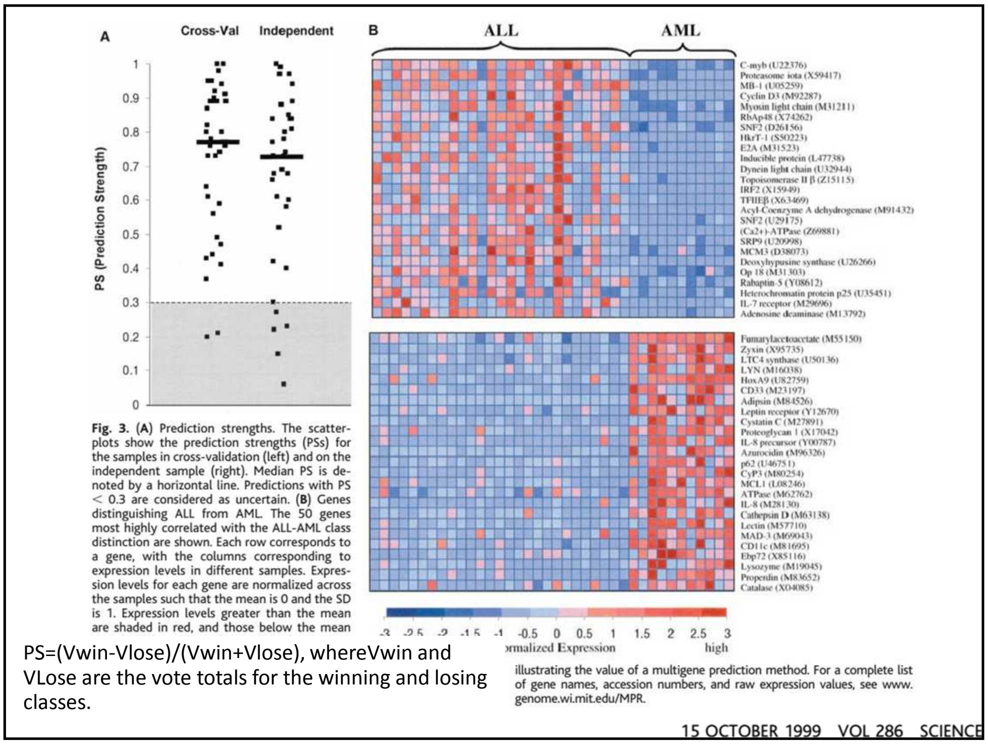
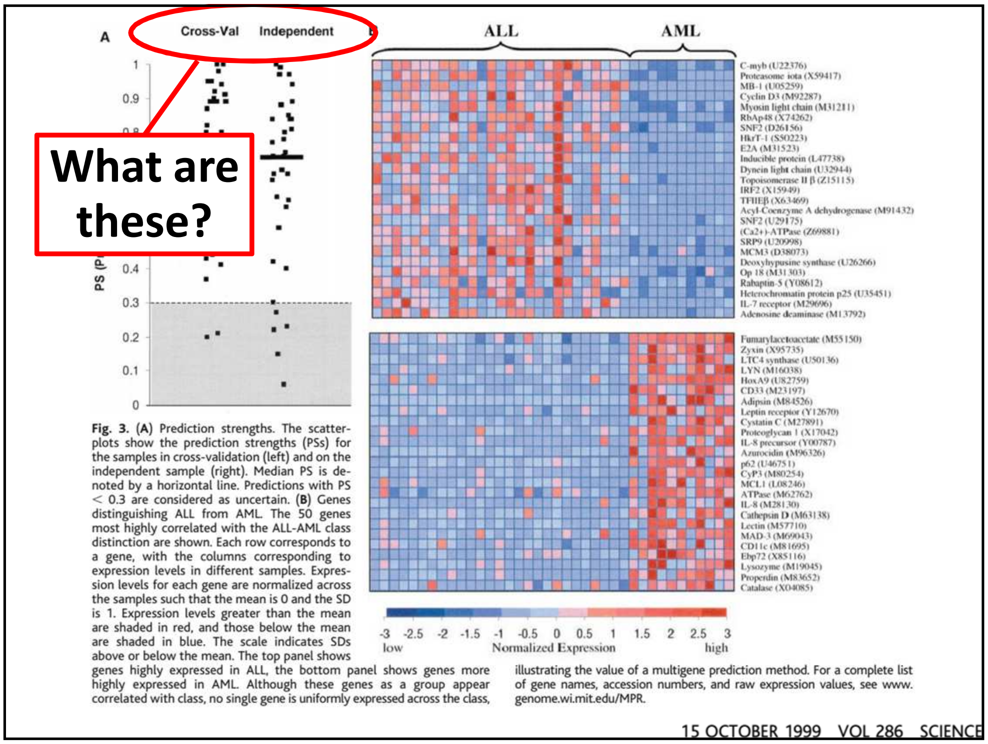


Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL 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.

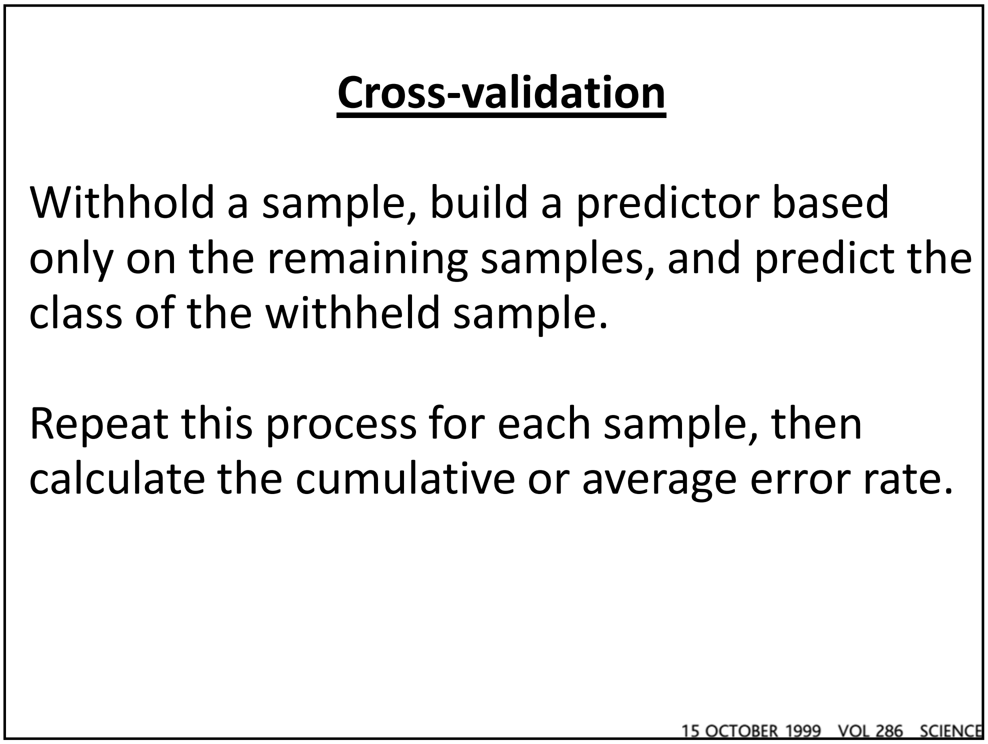
15 OCTOBER 1999 VOL 286 SCIENCE



What are these?

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.

15 OCTOBER 1999 VOL 286 SCIENCE

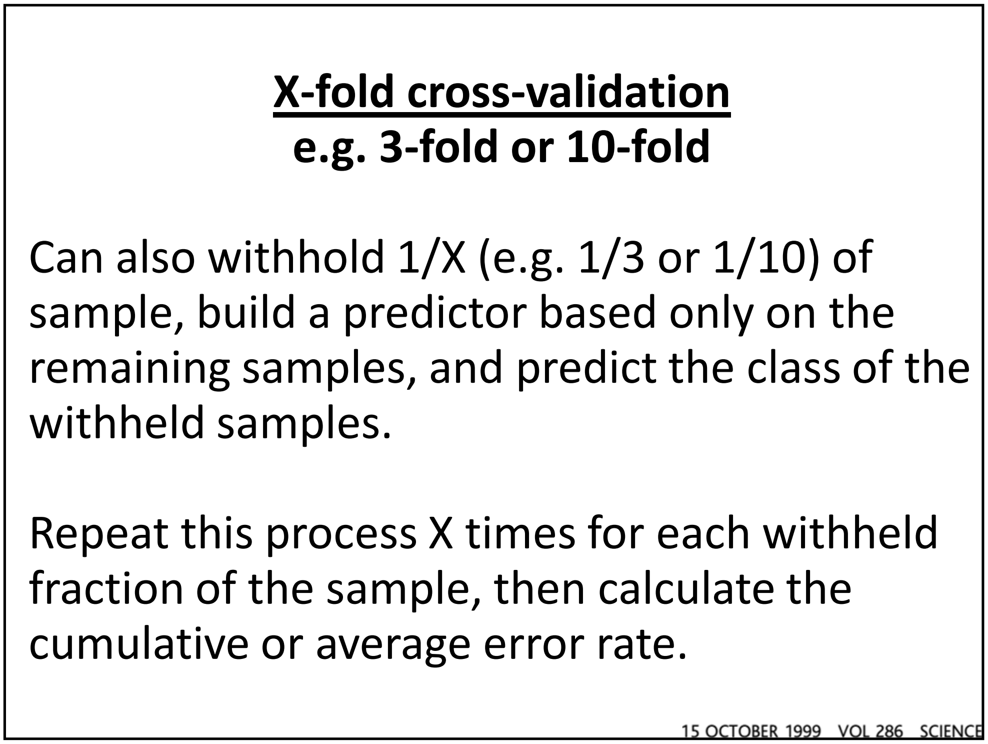


Cross-validation

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.

15 OCTOBER 1999 VOL 286 SCIENCE



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.

15 OCTOBER 1999 VOL 286 SCIENCE

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

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

True answer:

		Positive	Negative
Algorithm predicts:	Positive	True positive	False positive
	Negative	False negative	True negative

Specificity = TP / (TP + FP)

Sensitivity = TP / (TP + FN)

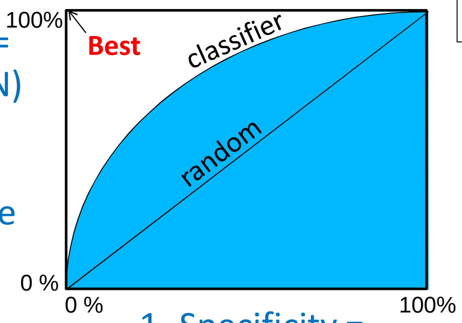
You already know how to measure how well these algorithms work (way back in our discussion of gene finding!)…

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)



1- Specificity = FP / (FP + TN)

also called False Positive Rate (FPR)

ROC curve

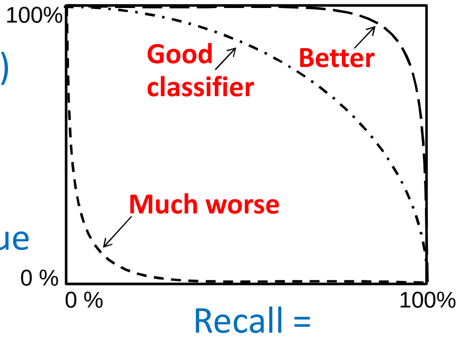
(receiver operator characteristic)

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)

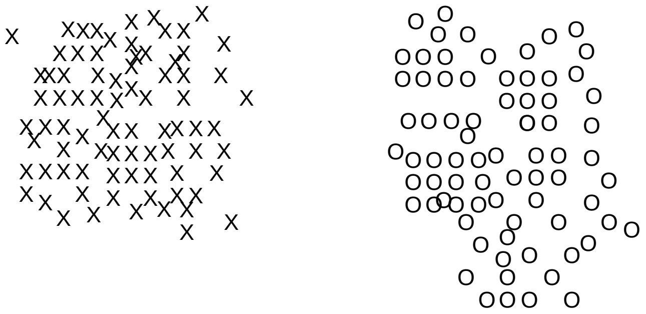


Recall = TP / (TP + FN) (= sensitivity)

Precision-recall curve

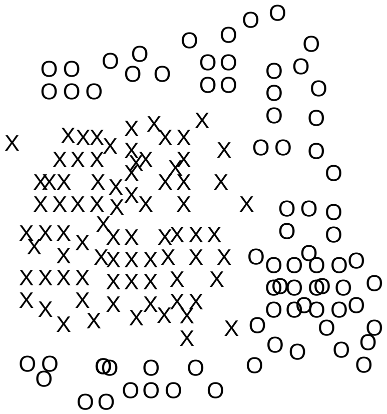
Back to our minimum distance classifier…

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…

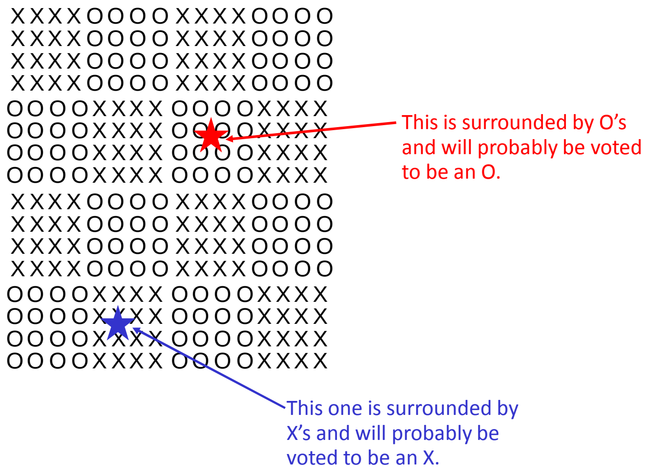
How about this data? What might?

```
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
```

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.

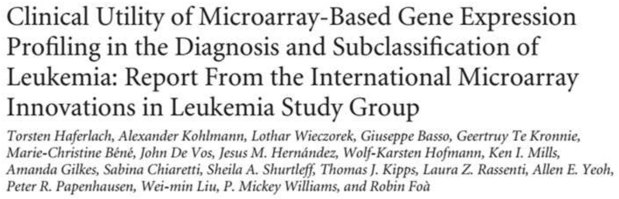


```
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
XXXXOOOOXXXXOOOO
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
OOOOXXXX OOOOXXXX
```

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

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

& back to the leukemia samples. 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

- Assessed clinical utility of gene expression profiling to subtype leukemias into myeloid and 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 diagnostic methods 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