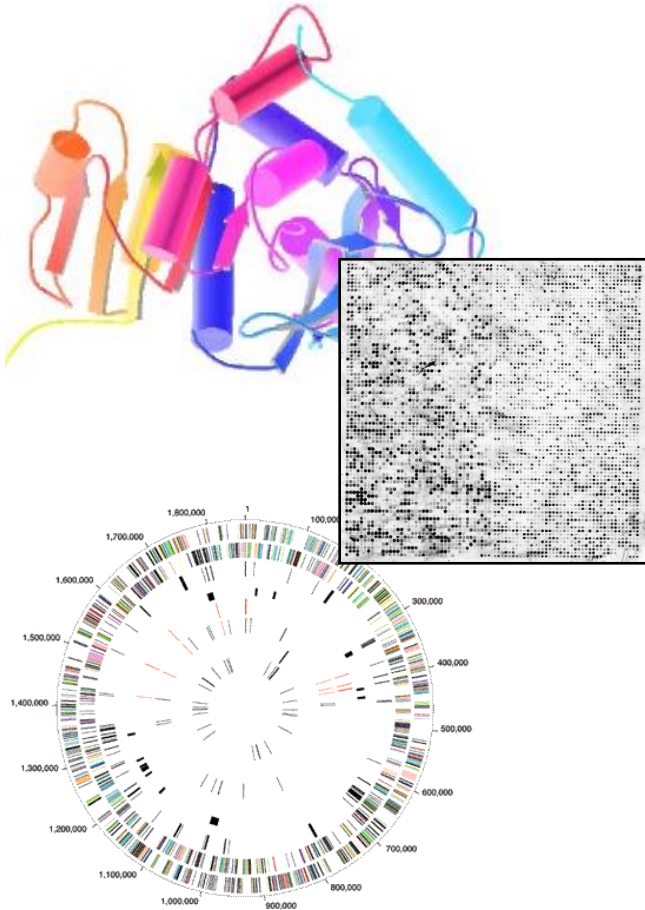# Biomedical Data Science:
Supervised Datamining –
ROC Curves & Cross-validation

Mark Gerstein, Yale University

GersteinLab.org/courses/452

(Last edit in spring '22. Comparing this year to last, 22m8b is essentially the same as M8b.
Checked that error on slide 4 is fixed.)

Supervised Mining:

Assessment, Cross-Validation & ROC Curves
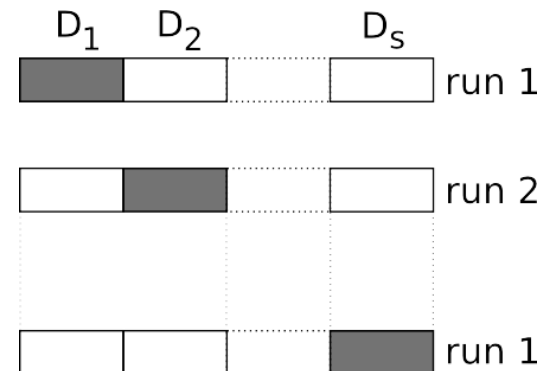
Evaluating performance: What? How?

A. _**What** do we want to evaluate?_

GENERALIZATION

Therefore, it is mandatory to divide your dataset:

Alternatively, use Cross Validation:

B. _How do we evaluate performance?_

1. Classification problems

Accuracy
TP+TN/(TP+FP+FP+TN)

Sensitivity (or TPR)
TP/P=TP/(TP+FN)

Specificity
TN/N=TN/(TN+FP)

Positive predictive value (PPV)
TP/(TP+FP)

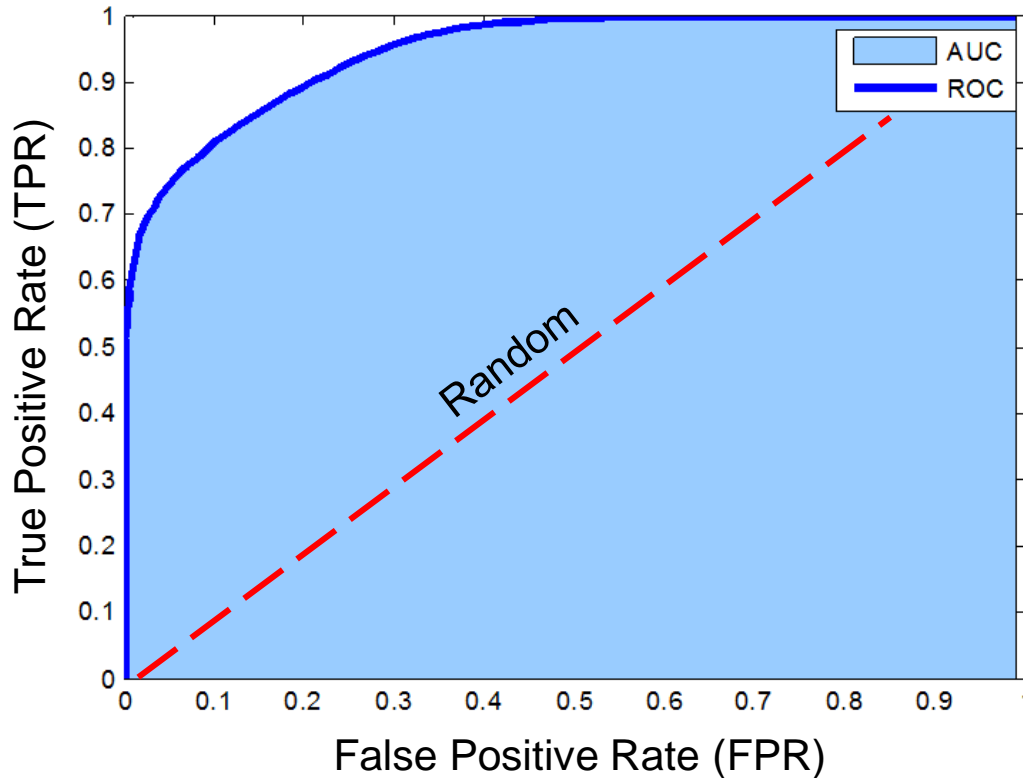*False positive rate (FPR)*
FP/N=FP/(FP+TN)

False discovery rate (FDR)
FP/(FP+TP)

2. Regression problems Sum
of squares error
Root Mean Square error

ROC analysis is good for comparing binary classifiers

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Intuition : ROC Curve

$$TPR = TP / P = TP / (TP + FN)$$

$$FPR = FP / N = FP / (FP + TN)$$

[From Biometrical Fusion - input statistical distribution]

Model dimensionality and overfitting
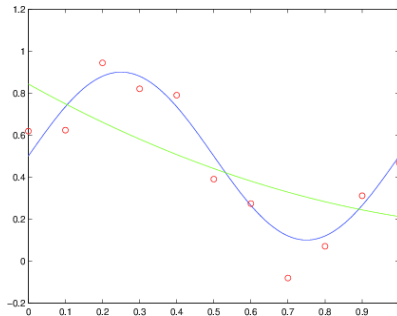

We are given the red dots.

We assume that they are noisy samples from a signal/(function) – the blue curve – which we do not have (we only have the red dots).

We want to predict new points, i.e. the *y* coordinates for other values of *x* (e.g. *x > 1*)
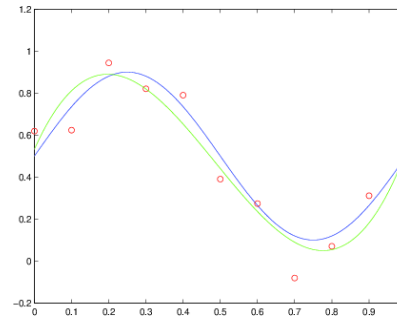
Our model needs to approximate the blue function. We decide to do it with polynomials.
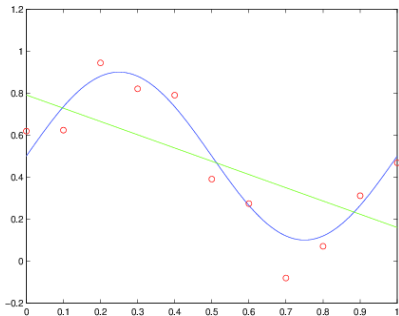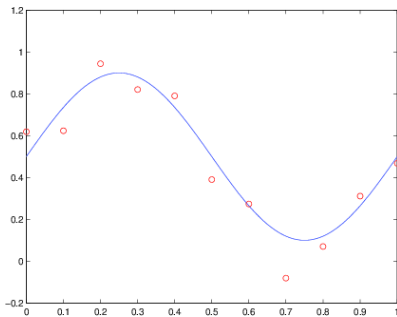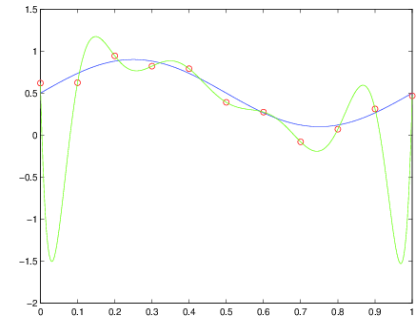

Degree 1 polynomial

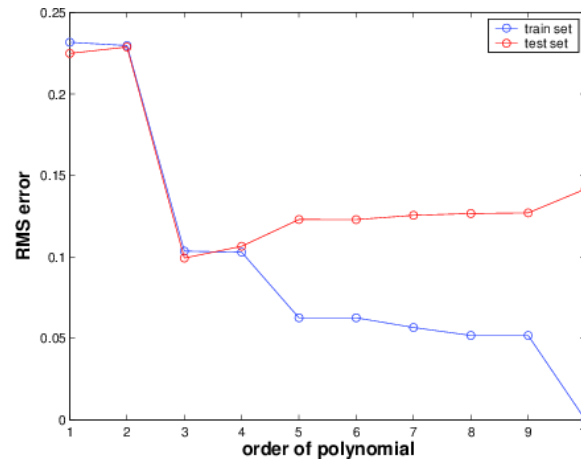

Degree 2 polynomial

Degree 3 polynomial

Degree 10 polynomial

Which one is best? And why?

How does the GENERALIZATION performance vary, as we increase the complexity of the polynomial?

- **Occam's razor** *(William of Occam, ~1300):* Accept the simplest explanation that fits the data.

We should prefer simpler models to more complex models, and this preference should be traded off against the extent to which the model fits the data.

- IMPORTANT: increasing the number of features may lead to a reduction in performance if the number of datapoints is not increased. Why?

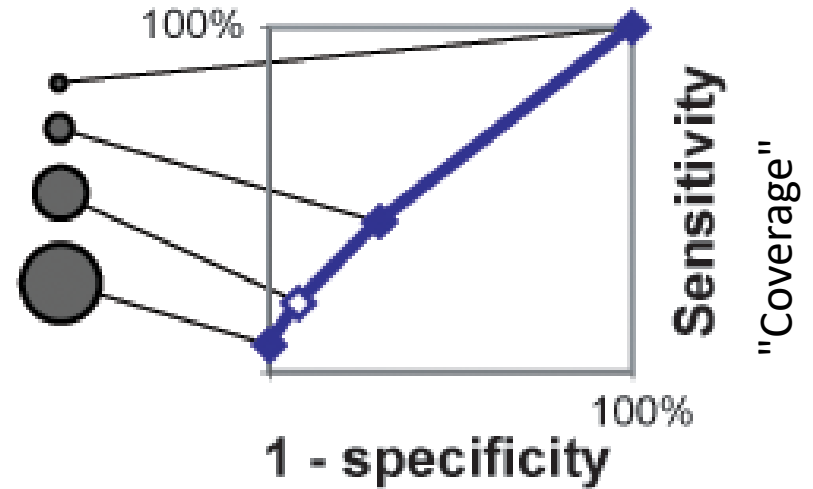|	Feature 1	Feature 2		Feature m	Target
Point 1	0.7	0.4		0.1	3.7
Point 2	0.6	0.3		0.2	4.2
⋮		...			
Point n	0.4	0.3		0.6	2.8

This is related to the "Curse of Dimensionality" Bellman, 1961.

Comparison of Predictions against a Positive and Negative Gold Standard
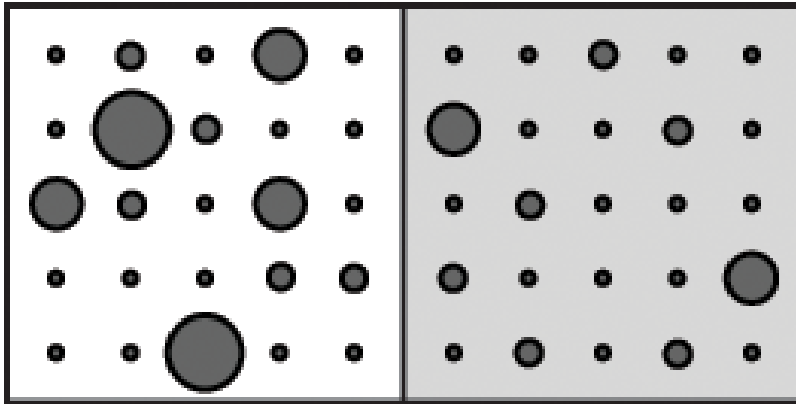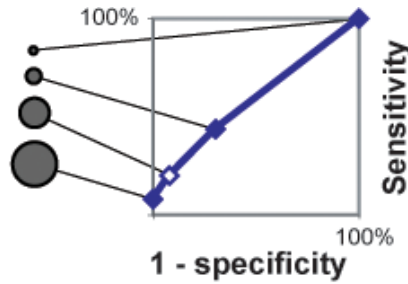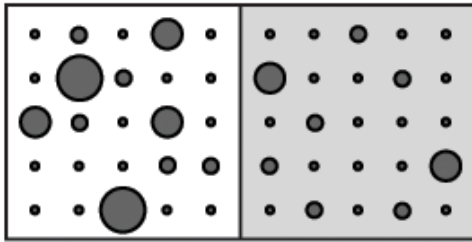
Threshold "predictions" at different levels and compare to + and - gold standards
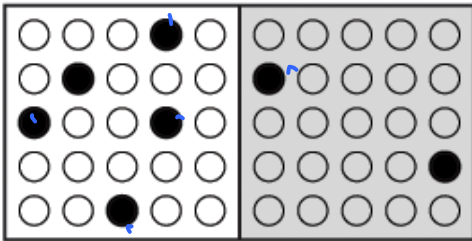
"Error Rate"

ROC plot
(cross validated)

Effect on Predictions of Large Number of Negatives
(e.g. terrorist identification or breast cancer screening)

Sensitivity

$$\frac{5}{25} = 20\%$$
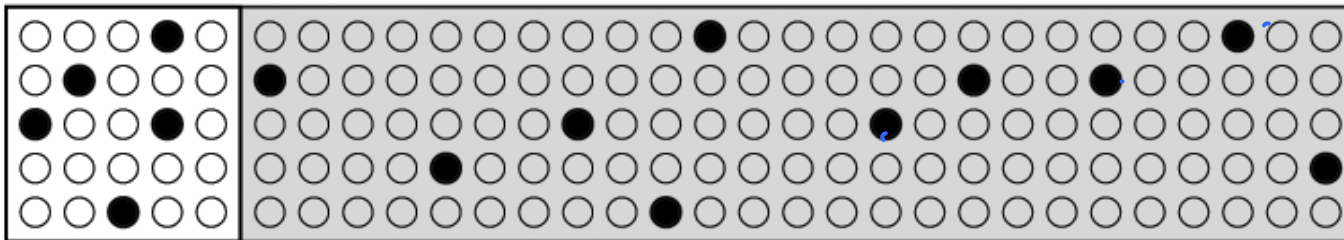
1- specificity

$$\frac{2}{25} = 8\%$$

Positive predictive value

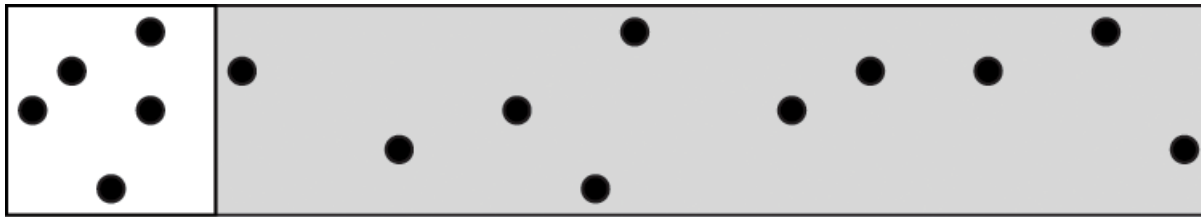$$\frac{5}{5+2} \approx 71\%$$

$$\frac{5}{25} = 20\%$$

$$\frac{10}{125} = 8\%$$
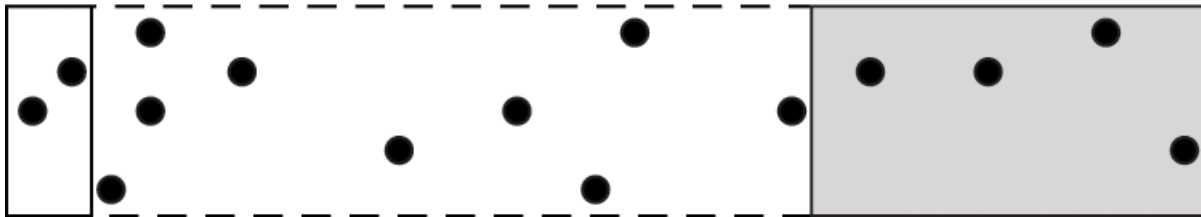
$$\frac{5}{5+10} \approx 33\%$$

Importance of Balanced Positive and Negative Examples
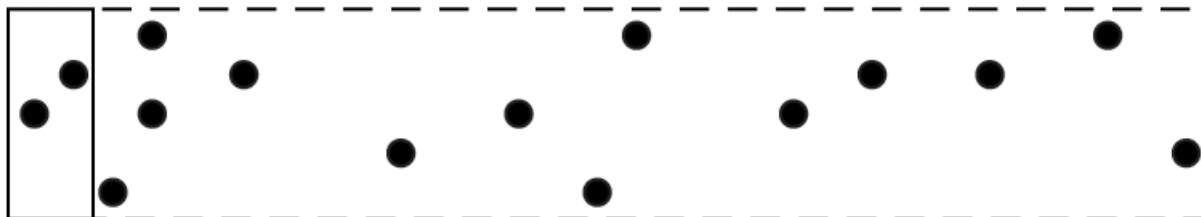
$$\frac{5}{?} = ?$$

$$\frac{10}{?} = ?$$

$$\frac{5}{5+10} \approx \textbf{33\%}$$

$$\frac{2}{?} = ?$$

$$\frac{4}{?} = ?$$

$$\frac{2}{2+4} \approx \textbf{33\%} \text{ (estimate)}$$

$$\frac{2}{?} = ?$$

$$\frac{?}{?} = ?$$

$$\frac{2}{2+?} = ?$$