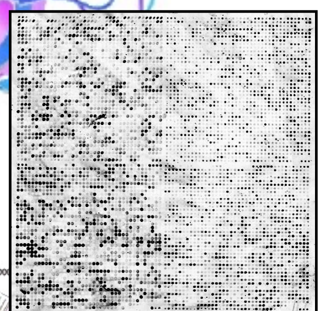
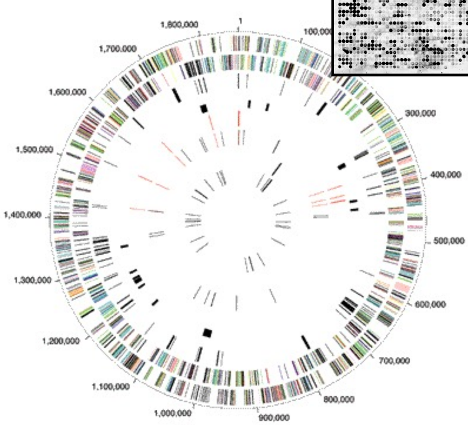
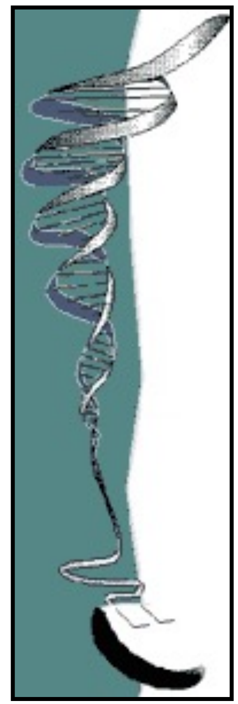
# Biomedical Data Science:
## Supervised Datamining B –
## ROC Curves & Cross-validation


Mark Gerstein, Yale University

GersteinLab.org/courses/452

(last edit in spring ’21, final)

Supervised Mining:

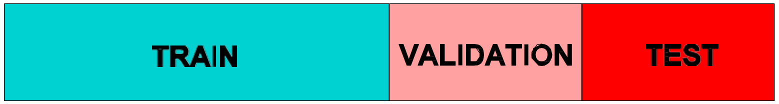
# Assessment, Cross-Validation & ROC Curves

# Evaluating performance: What? How?

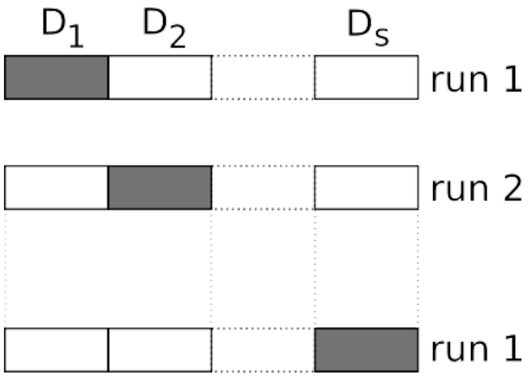
## A. *<u>What</u> do we want to evaluate?*

**GENERALIZATION**

Therefore, it is mandatory to divide your dataset:

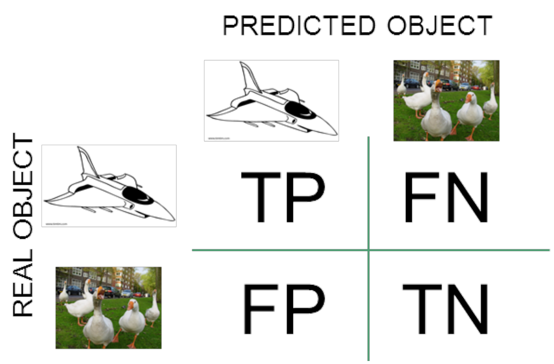



Alternatively, use Cross Validation:



# 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

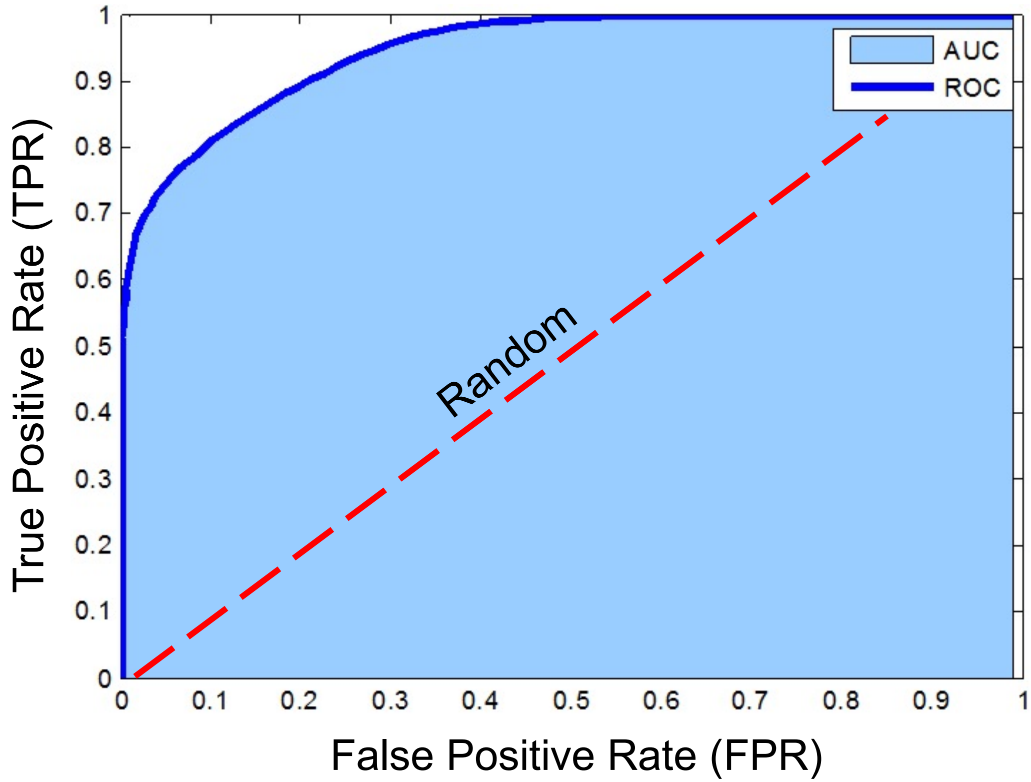
Root Mean Square error

*ROC analysis is good for comparing binary classifiers*

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

Slide from Alberto Paccanaro

# Intuition : ROC Curve



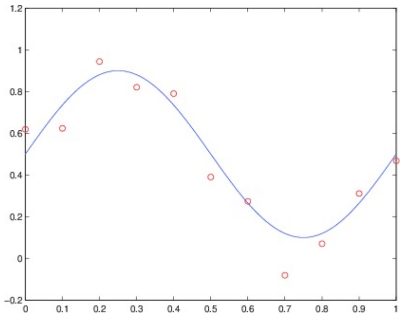

$$TPR = TP / P = TP / (TP + FN)$$

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

[From Biometrical Fusion - input statistical distribution ]

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

# Model dimensionality and overfitting

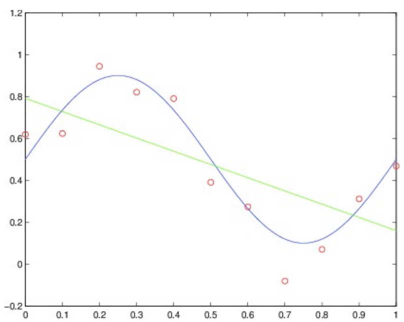


We are given the red dots.

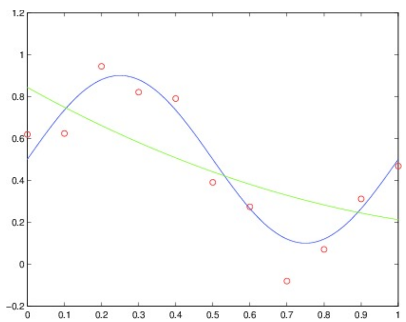
We assume that they are noisy samples from a signal/(function) – the blue curve – which we do not have (we only have the red dots).

We want to predict new points, i.e. the *y* coordinates for other values of *x* (e.g. $x > 1$)

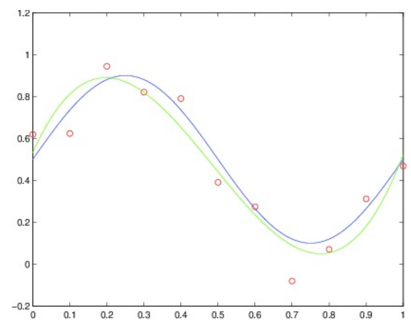
Our model needs to approximate the blue function. We decide to do it with polynomials.



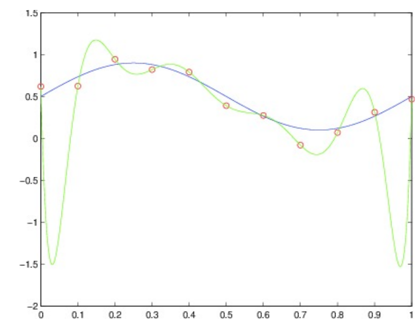
Degree 1 polynomial



Degree 2 polynomial



Degree 3 polynomial

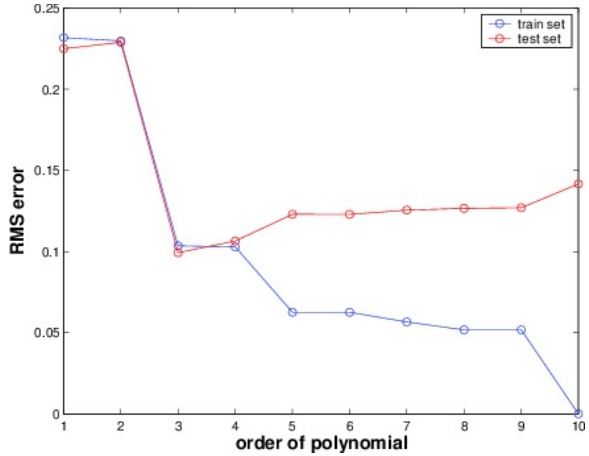


Degree 10 polynomial

Which one is best? And why?

Slide from Alberto Paccanaro

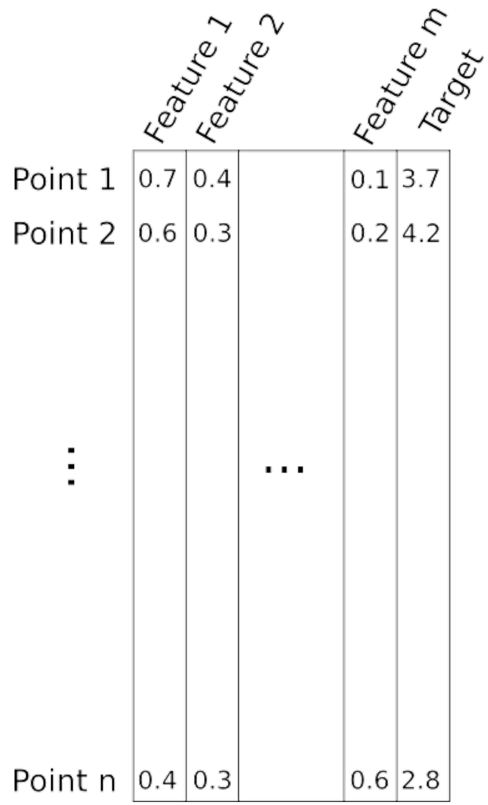
How does the GENERALIZATION performance vary, as we increase the complexity of the polynomial?




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

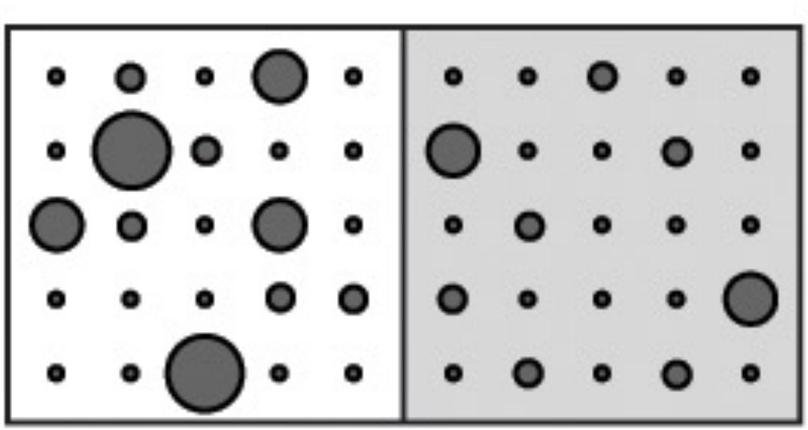
We should prefer simpler models to more complex models, and this preference should be traded off against the extent to which the model fits the data.

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

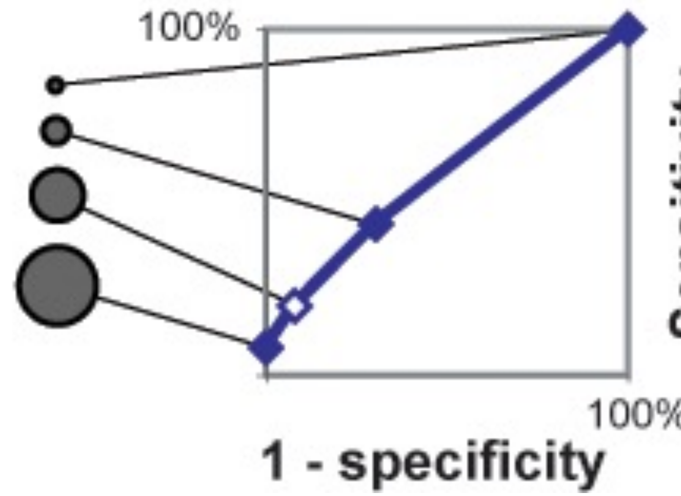



This is related to the “Curse of Dimensionality” Bellman, 1961.

# Comparison of Predictions against a Positive and Negative Gold Standard



Threshold "predictions" at different levels and compare to + and - gold standards

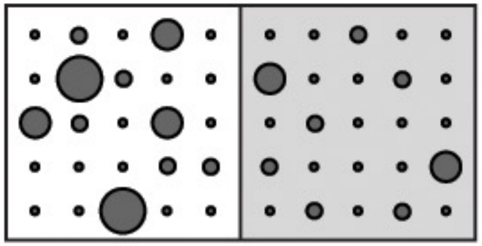


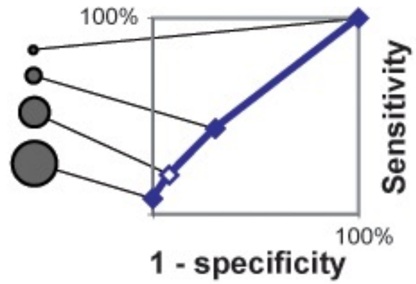

"Coverage"

"Error Rate"

ROC plot
(cross validated)

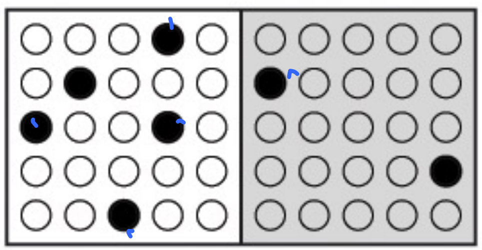
R Jansen, M Gerstein (2004). *Curr Opin Microbiol* 7: 535



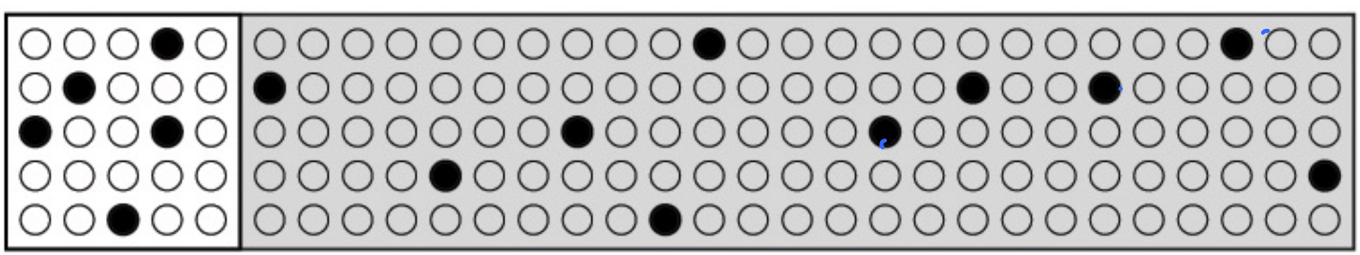



# Effect on Predictions of Large Number of Negatives

**Sensitivity** | **1- specificity** | **Positive predictive value**



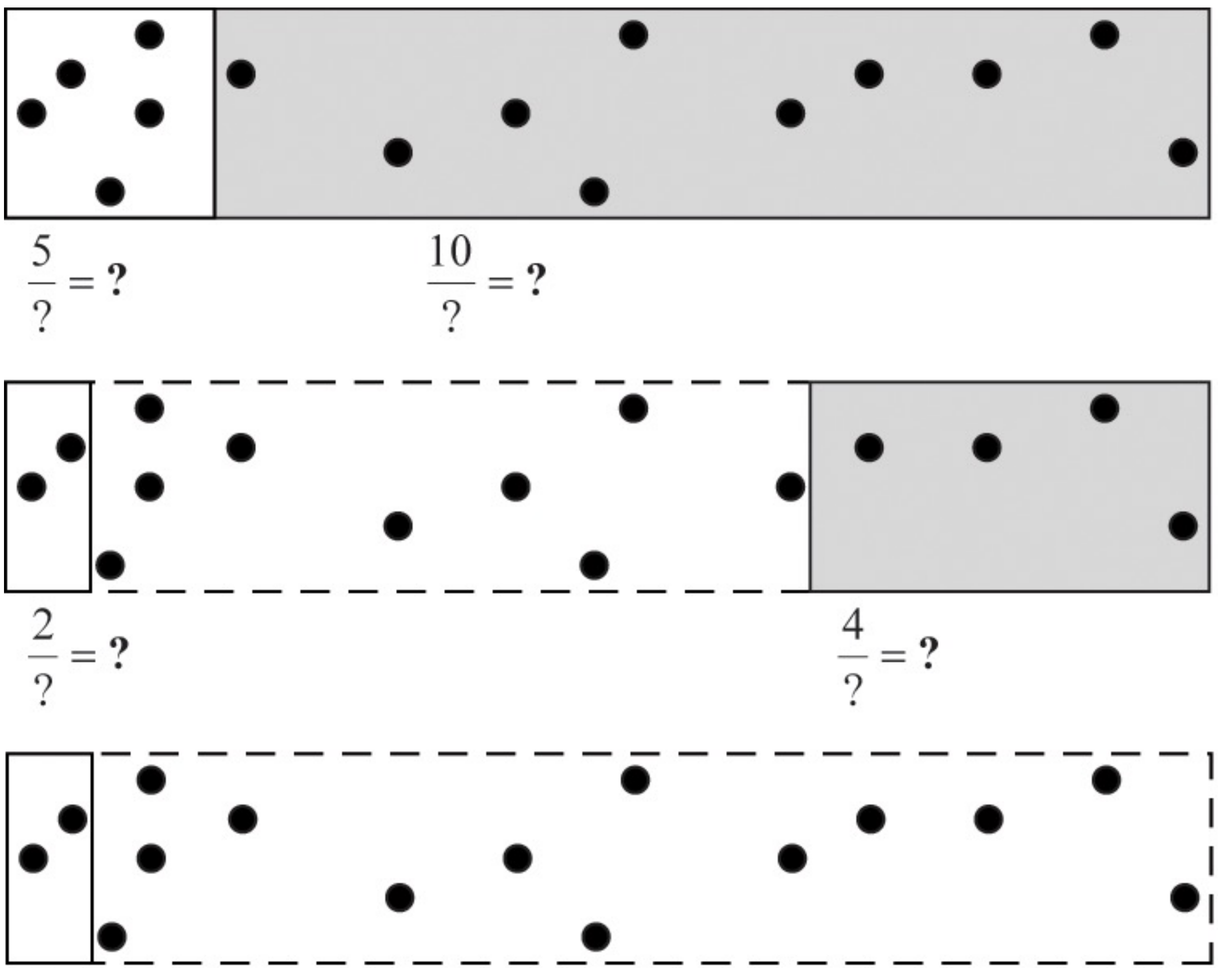
$$\frac{5}{25} = \mathbf{20\%} \qquad \frac{2}{25} = \mathbf{8\%} \qquad \frac{5}{5+2} \approx \mathbf{71\%}$$



$$\frac{5}{25} = \mathbf{20\%} \qquad \frac{10}{125} = \mathbf{8\%} \qquad \frac{5}{5+10} \approx \mathbf{33\%}$$

R Jansen, M Gerstein (2004). *Curr Opin Microbiol* 7: 535

# Importance of Balanced Positive and Negative Examples



$\frac{5}{?} = ?$ $\frac{10}{?} = ?$ $\frac{5}{5+10} \approx \mathbf{33\%}$

$\frac{2}{?} = ?$ $\frac{4}{?} = ?$ $\frac{2}{2+4} \approx \mathbf{33\%}$ (estimate)

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

R Jansen, M Gerstein (2004). *Curr Opin Microbiol* 7: 535