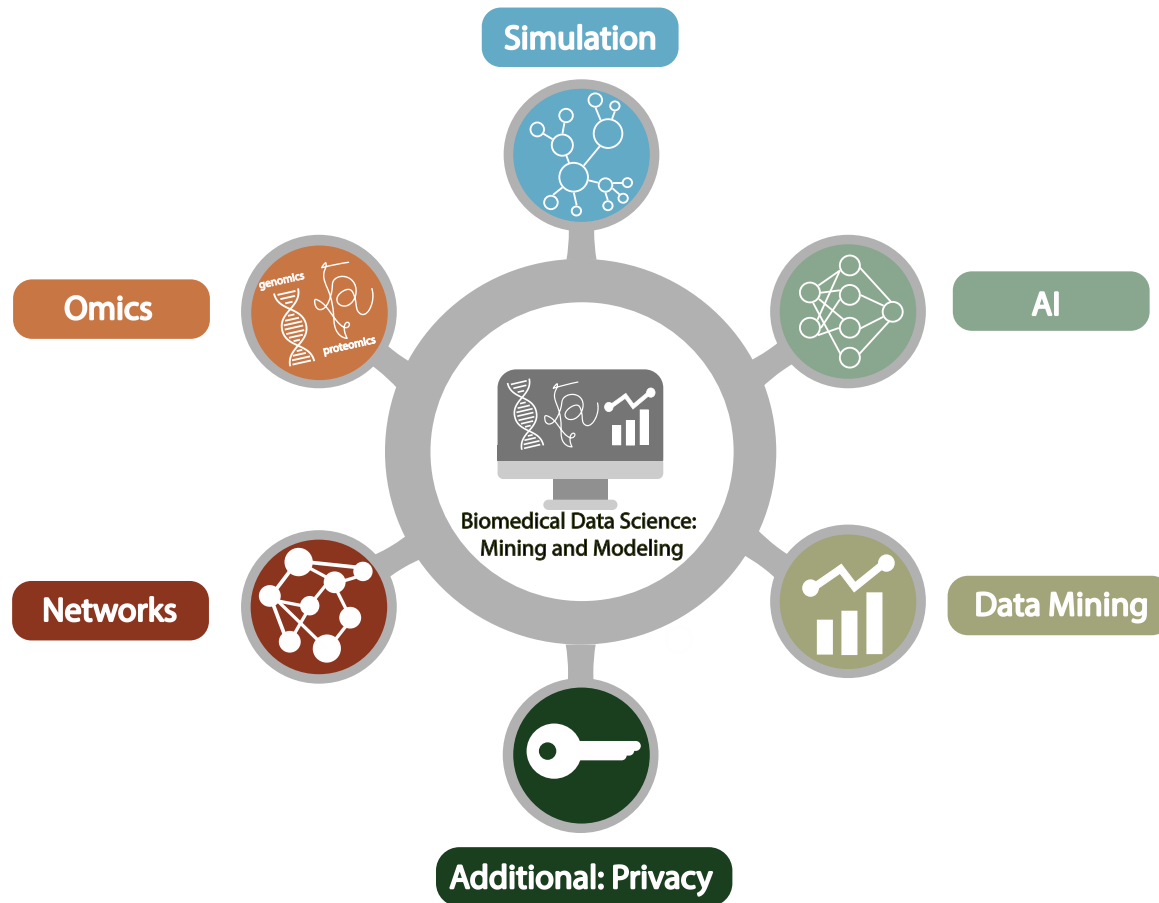


Biomedical Data Science (GersteinLab.org/courses/452)

Supervised Datamining – ROC Curves & Cross-validation (23m8b)



Supervised Mining:

**Assessment, Cross-
Validation & ROC Curves**

Evaluating performance: What? How?

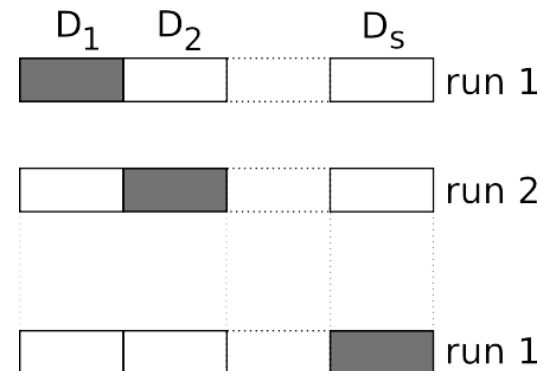
A. What do we want to evaluate?

GENERALIZATION

Therefore, it is mandatory to divide your dataset:







Alternatively, use Cross Validation:



B. How do we evaluate performance?

1. Classification problems

	PREDICTED OBJECT	
		
REAL OBJECT	 TP	 FN
	FP	TN

Accuracy

$$\frac{TP+TN}{(TP+FP+FN+TN)}$$

Sensitivity (or TPR)

$$\frac{TP}{P} = \frac{TP}{(TP+FN)}$$

Specificity

$$\frac{TN}{N} = \frac{TN}{(TN+FP)}$$

Positive predictive value (PPV)

$$\frac{TP}{(TP+FP)}$$

False positive rate (FPR)

$$\frac{FP}{N} = \frac{FP}{(FP+TN)}$$

False discovery rate (FDR)

$$\frac{FP}{(FP+TP)}$$

2. Regression problems

Sum of squares error

Root Mean Square error

ROC analysis is good for comparing binary classifiers

- **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.

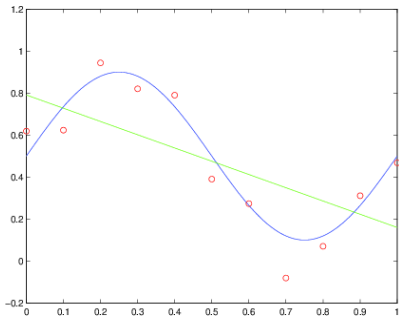
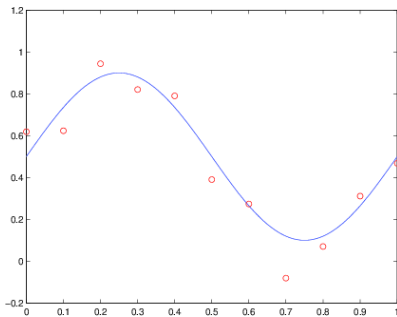
Model dimensionality and overfitting

We are given the red dots.

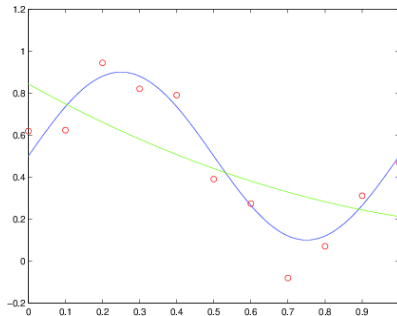
We assume that they are noisy samples from a signal/(function) – the blue curve – which we do not have (we only have the red dots).

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

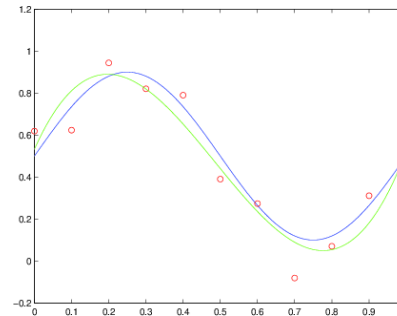
Our model needs to approximate the blue function.
We decide to do it with polynomials.



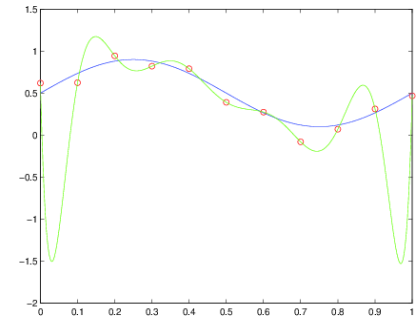
Degree 1 polynomial



Degree 2 polynomial



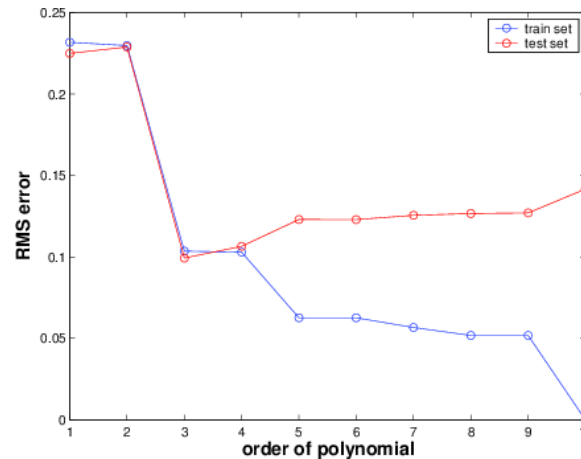
Degree 3 polynomial



Degree 10 polynomial

Which one is best? And why?

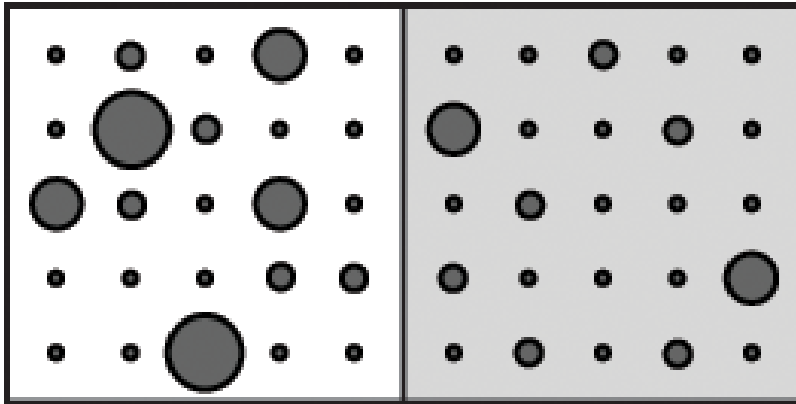
How does the GENERALIZATION performance vary, as we increase the complexity of the polynomial?



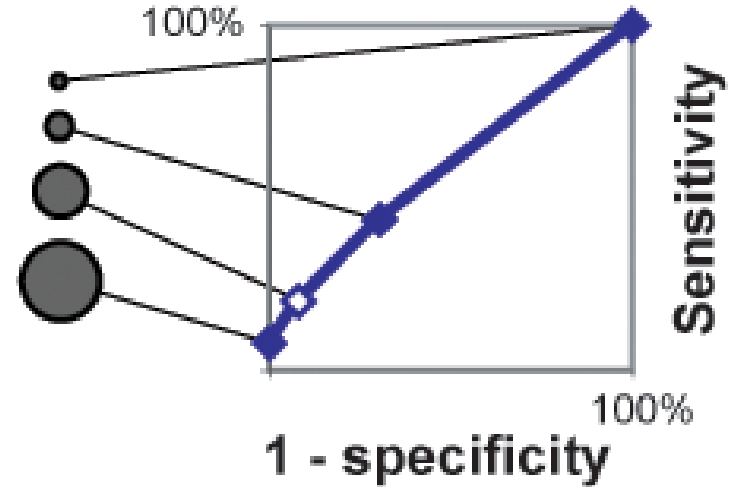
- Occam's razor (*William of Occam, ~1300*): Accept the simplest explanation that fits the data.

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.

Comparison of Predictions against a Positive and Negative Gold Standard

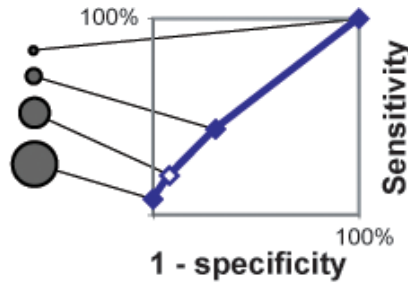
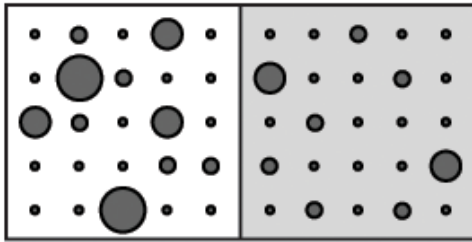


Threshold "predictions" at different levels and compare to + and - gold standards



"Error Rate" (FP/N)

ROC plot
(cross validated)

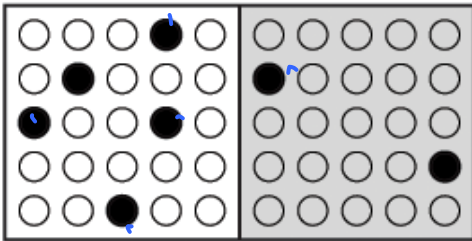


Effect on Predictions of Large Number of Negatives
(e.g. terrorist identification or breast cancer screening)

Sensitivity

1 - specificity

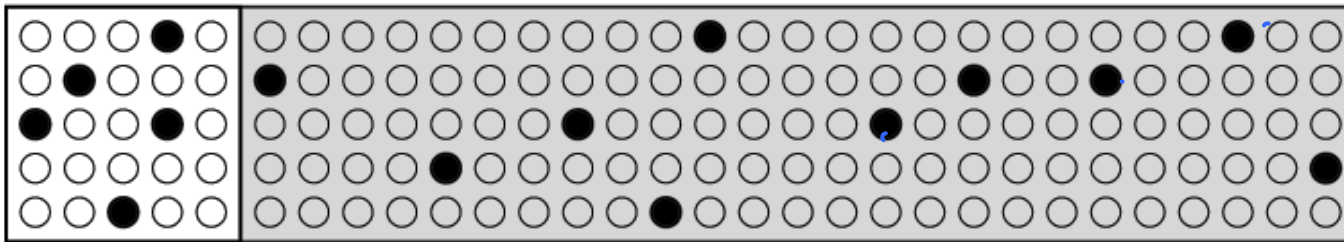
Positive predictive value



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

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

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

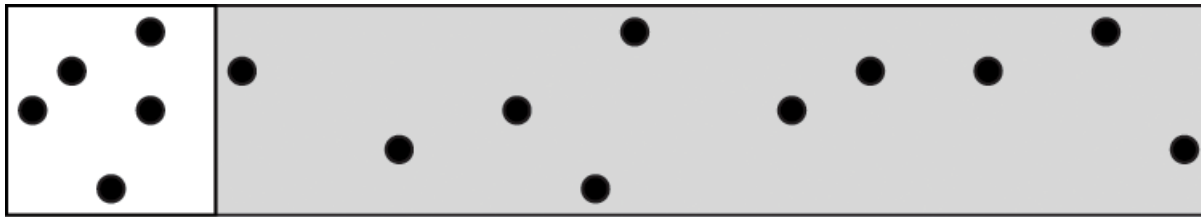


$$\frac{5}{25} = 20\%$$

$$\frac{10}{125} = 8\%$$

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

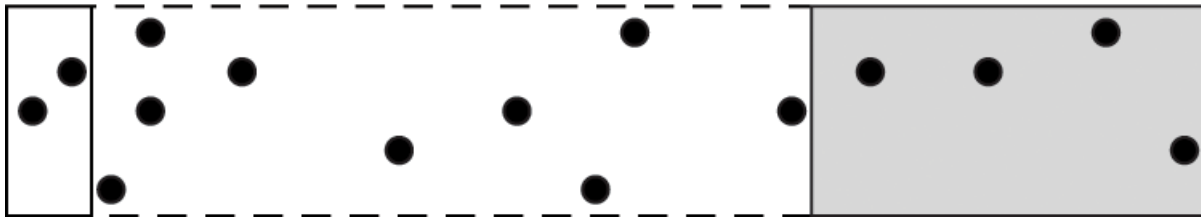
Importance of Balanced Positive and Negative Examples



$$\frac{5}{?} = ?$$

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

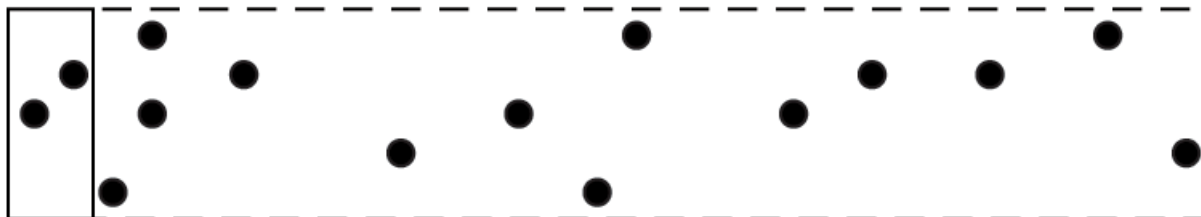
$$\frac{5}{5+10} \approx 33\%$$



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

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

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



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

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

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