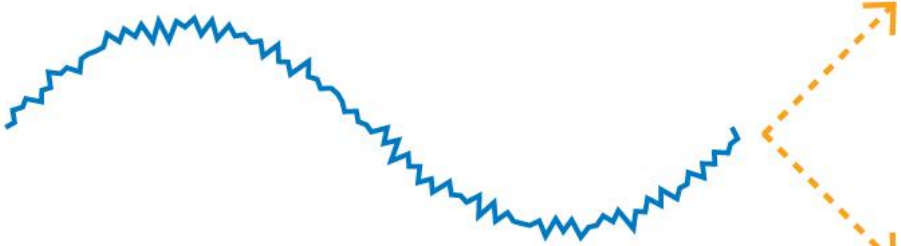
Overfitting, Cross-validation

Week 04 - Day 03

Signal vs. Noise

What we observe

SIGNAL + NOISE



Isolated noise from signal

SIGNAL

NOISE

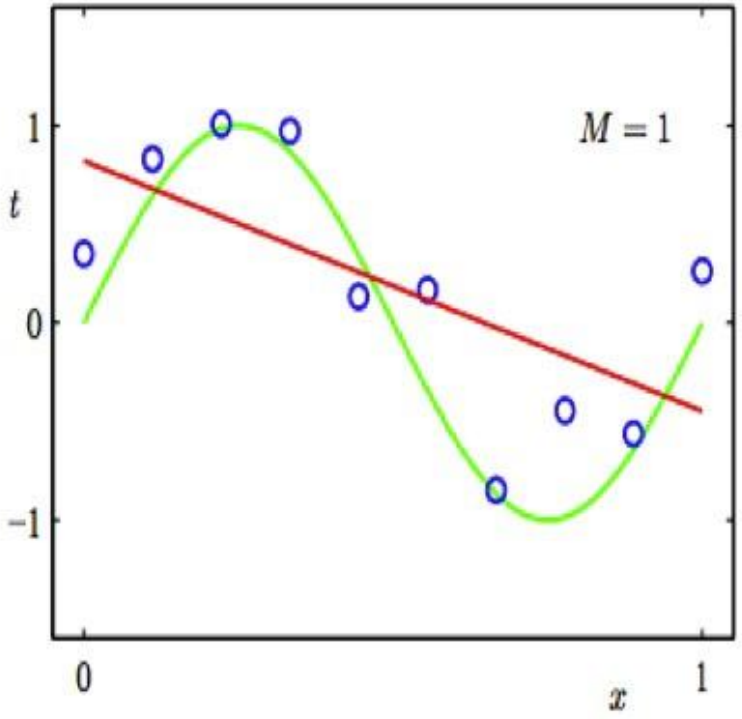
There always is noise in the data!

(otherwise you don’t need ML)

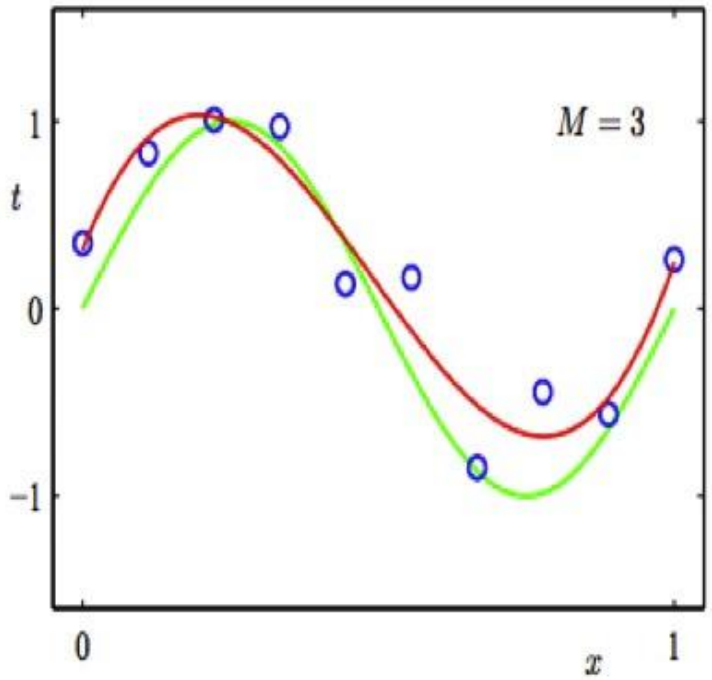
In ML we want to learn the signal, not the noise!

Overfitting
+
Underfitting

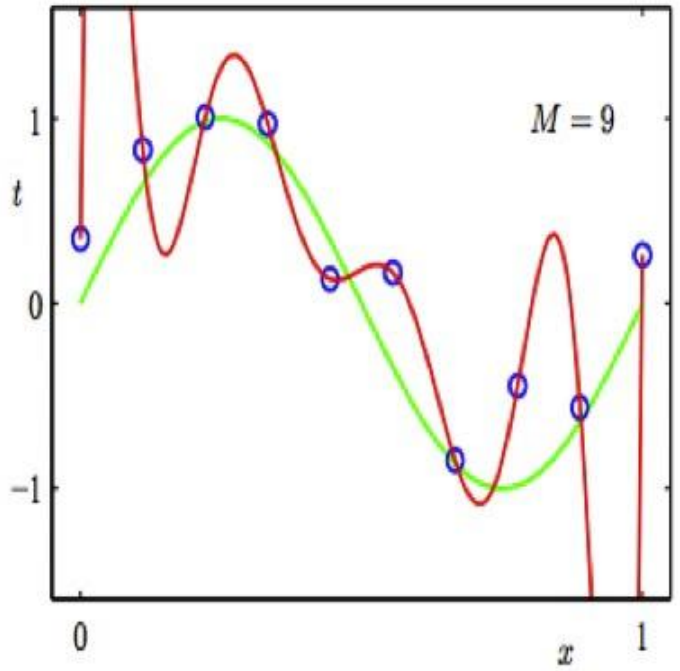
Is this model (red line) good?



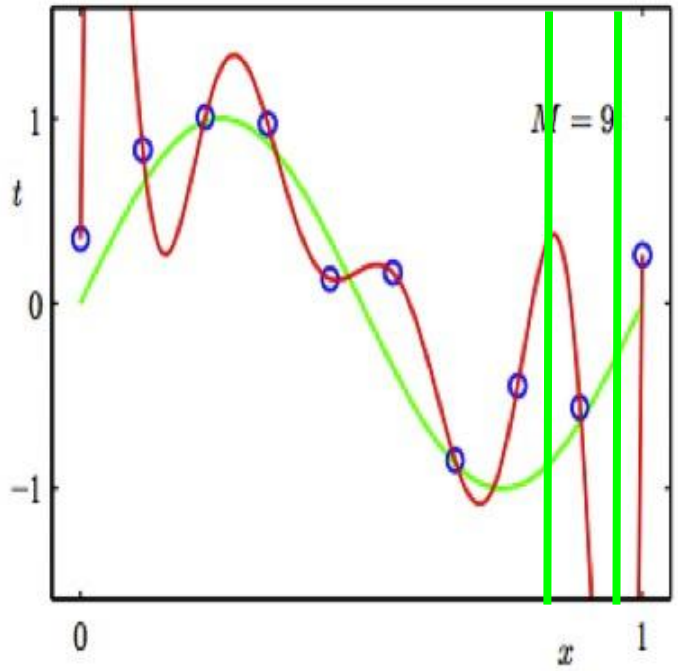
Better now?



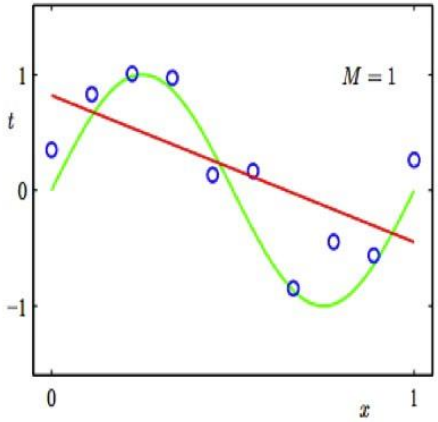
Even better!



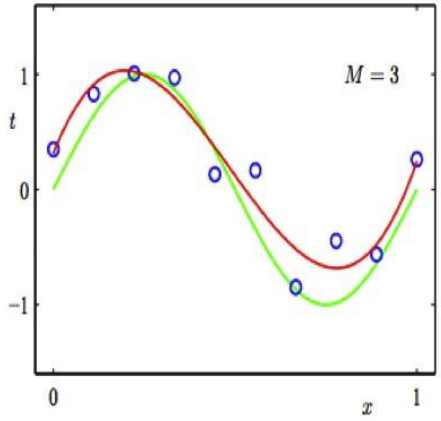
What's the prediction?

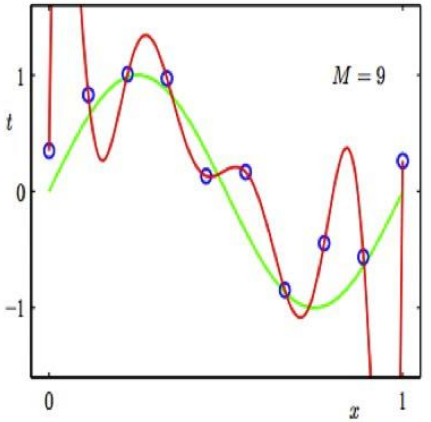


Under- and Over-fitting examples



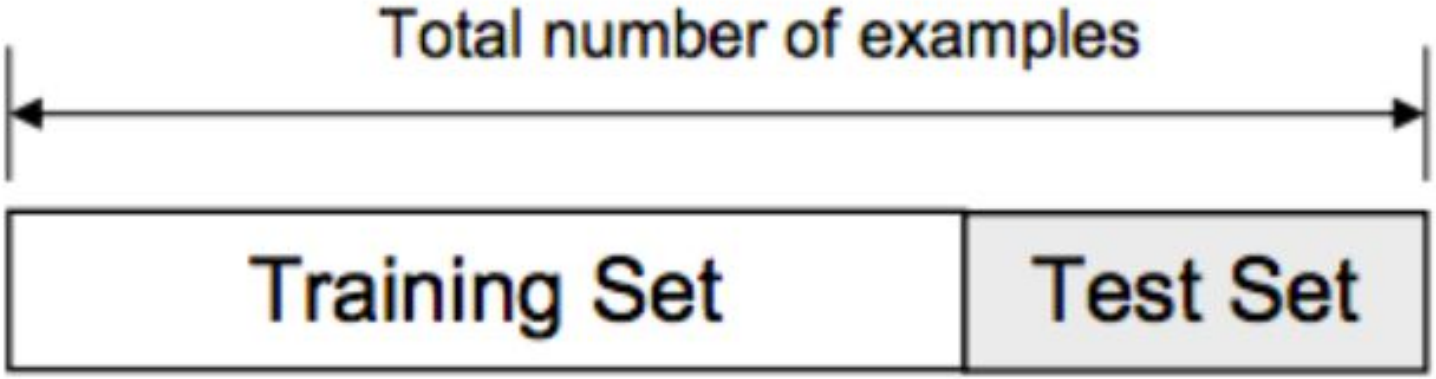
predictor too inflexible:
cannot capture pattern





predictor too flexible:
fits noise in the data

What's the solution? (answers?)


Total number of examples
Training Set
Test Set

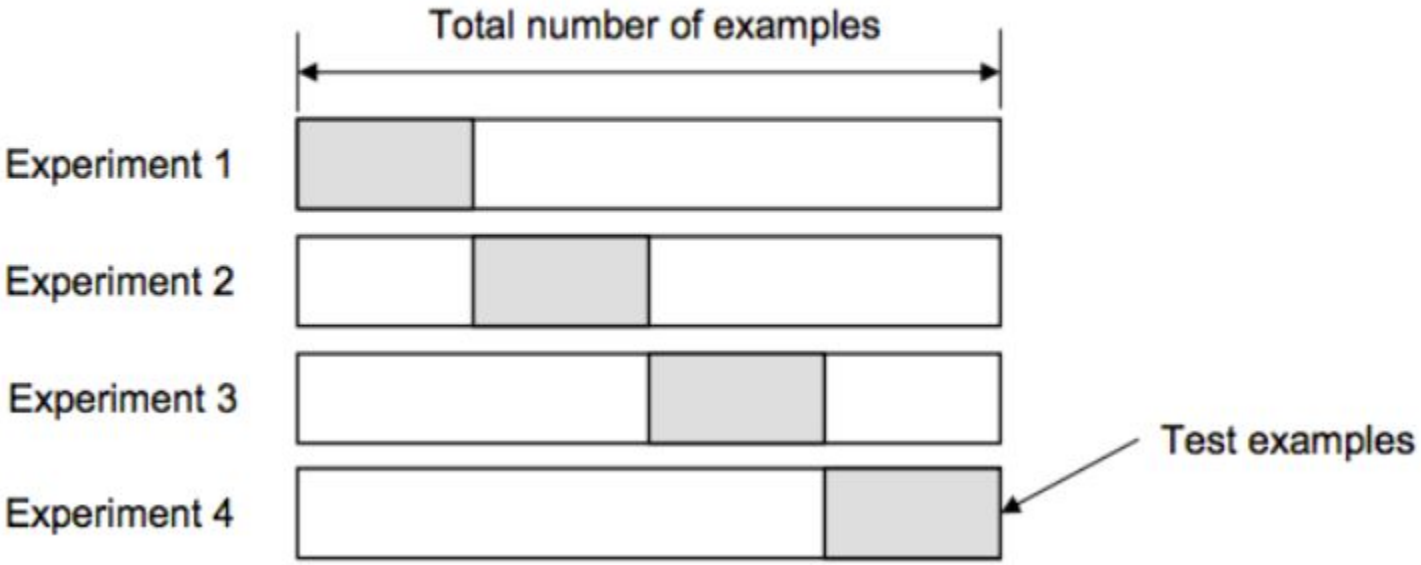
- **Training Set**: Used to train the classifier
- **Testing Set:** Used to estimate the error rate of the trained classifier
- **Advantages?** Fast! Simple! Computationally inexpensive!
- **Disadvantages?**

- **Training Set**: Used to train the classifier
- **Testing Set:** Used to estimate the error rate of the trained classifier
- **Advantages?** Fast! Simple! Computationally inexpensive!
- **Disadvantages?** Eliminating data! Imperfect splits!

- **How can we use the maximum amount of our data points while still ensuring model integrity?**

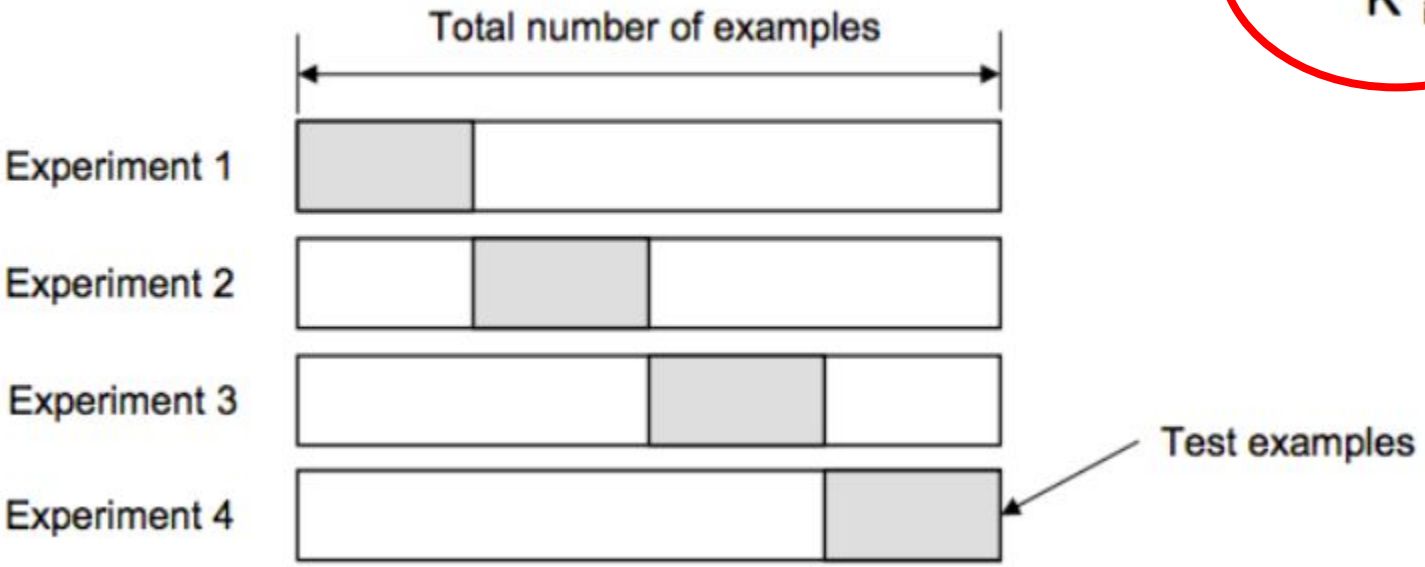
- Toss out answers – your answers are valuable parts of being an inquisitive data scientist wanting to test your assumptions

(K-Fold) Cross Validation!



Cross Validation!

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$



How many folds?
2-5-10-100-1000?

Aspects to consider:

Training time

Variance

Using a big/small % of the data

What happens with just 2 folds?

What happens with 100 folds?

Ideal = 10

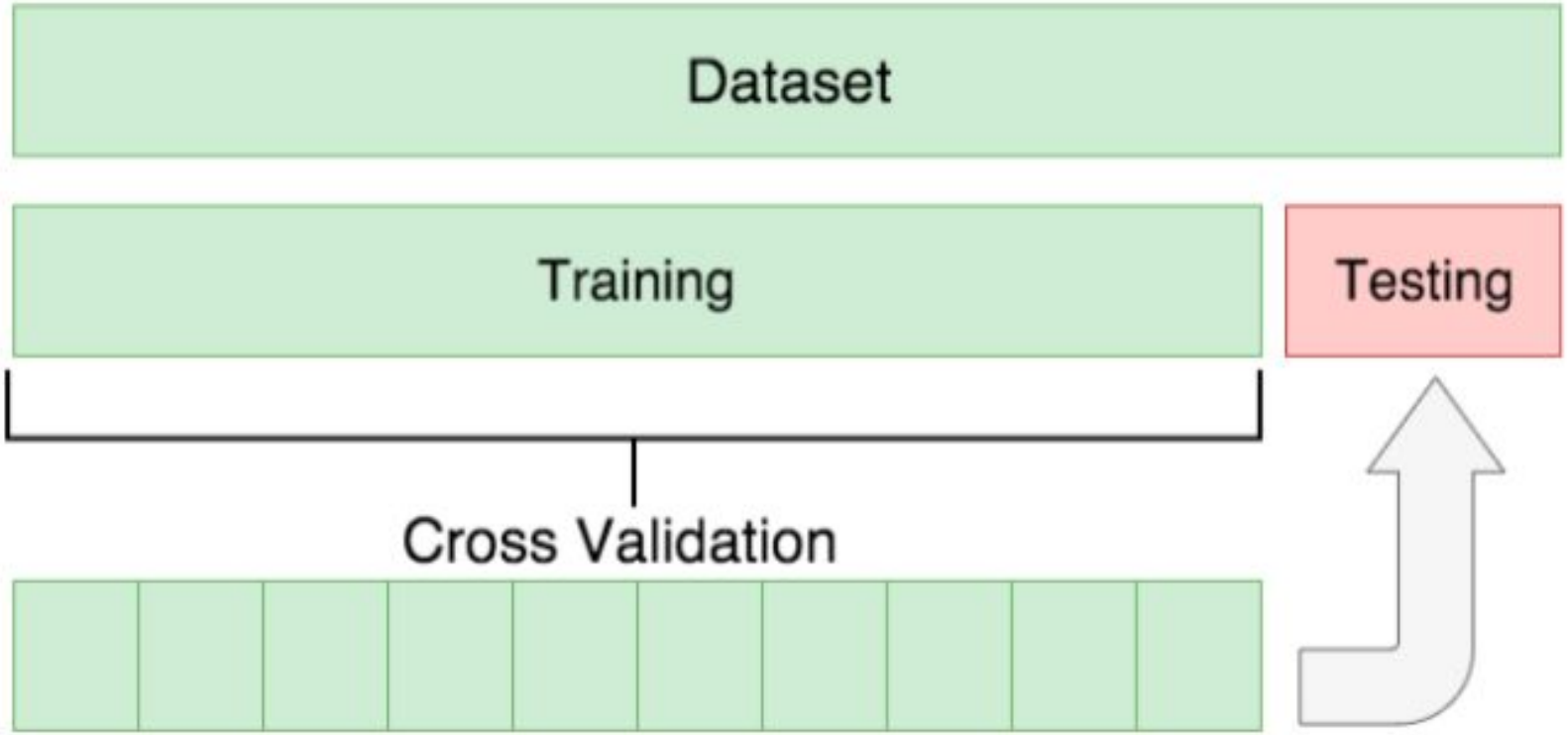
Good enough = 5

For big datasets = 3

Extreme case: k=n

Leave-one-out cross validation!

Ideal case: train+validation+test


Dataset
Training
Testing
Cross Validation

PROCEDURE

- 1. Divide data into training, validation, testing sets
- 2. Select architecture (model type) and training parameters (k)
- 3. Train the model using the training set
- 4. Evaluate the model using the training set
- 5. Repeat 2-4 selecting different architectures (models) and tuning parameters
- 6. Select the best model
- 7. Assess the model with the final testing set