Data Analytics
EEE 4774 & 6777

Module 1

Introduction to Data Science & Machine Learning

Spring 2022

First Day Attendance

- Please complete the survey on Canvas until Friday to keep your course registration
- Those who won’t complete will be automatically dropped from the course
- You can re-register if dropped unintentionally

Who am I?

- Yasin Yilmaz – born and raised in Turkey
- Graduated from Columbia University in 2014 with my Ph.D. in Electrical Engineering.
- I've been with USF since 2016.
- My expertise is Machine Learning and Statistical Signal Processing.

What do I do?

Secure & Intelligent Systems Lab

sis.eng.usf.edu

- **Artificial Intelligence (AI), Deep Learning, Data Science, Big Data**
- Computer vision (real-time video analysis)
- Cybersecurity
- Smart power grid and energy systems
- Internet of Things (IoT)
- Intelligent transportation systems
- Wireless communications

How will the course work?

- Canvas: https://usflearn.instructure.com
  - Syllabus
  - Quizzes
  - Announcements
- Lectures on Teams
  - Same meeting link for all lectures
- To contact me:
  - Office hours: Monday 1:00-3:00 PM
  - Email: yasiny@usf.edu
  - Office: Virtual on MS Teams

How will the course work?

- Python homework assignments due in 1 week
  - Assigned on Friday, due next Friday
  - See the schedule on the syllabus
- Quizzes on Canvas due in 3 days
  - Question from a recent topic
  - Assigned on Friday, due Monday
  - See the schedule on the syllabus

How will the course work?

- **Final project** includes a comprehensive data analysis problem
  - Topic either chosen by you or provided by me
  - If you want to choose your own project, you have to discuss it with me to finalize your selection
  - See the Project module for details
  - Preliminary report due 3/10
  - Oral project presentations during the last three weeks 4/12-4/28
    - **Sign up at Project module**
  - Written report due 5/2

Grading

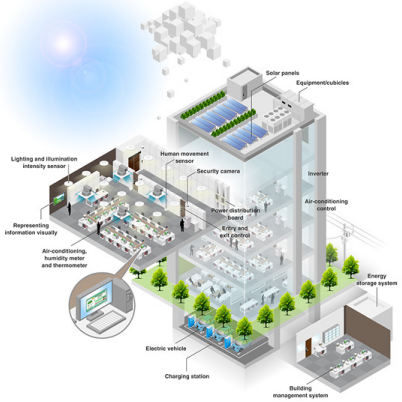
- 5 Python homework assignments: 35% (7% each)
- 5 Quizzes:
  - 25% (5% each) for EEE 6777
  - 35% (7% each) for EEE 4774
- Final project:
  - 40% for EEE 6777

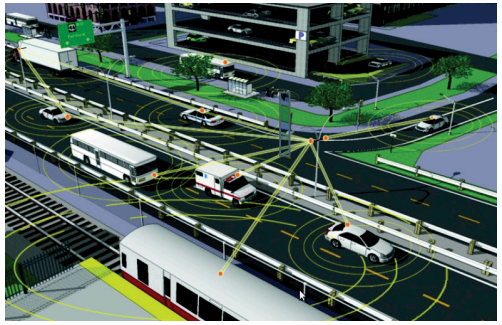
  (5% Prelim. Report + 15% Presentation + 20% Final Report)
  - 30% for EEE 4774

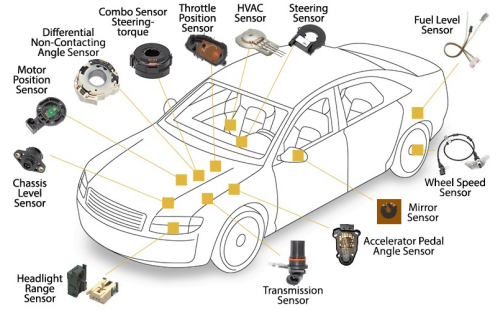
  (5% Prelim. Report + 10% Presentation + 15% Final Report)

Homework Assignments

- Unique submission
  - Not a group assignment
  - You should write your own code. Changing the variable names is not enough to have a unique submission. Plagiarism software for codes will catch it.
  - See academic integrity policy in syllabus for penalties
- Submit the code and a report explaining results



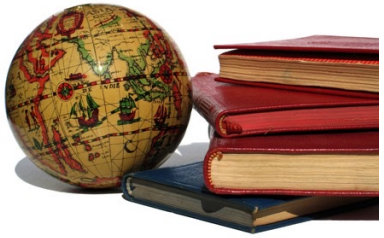




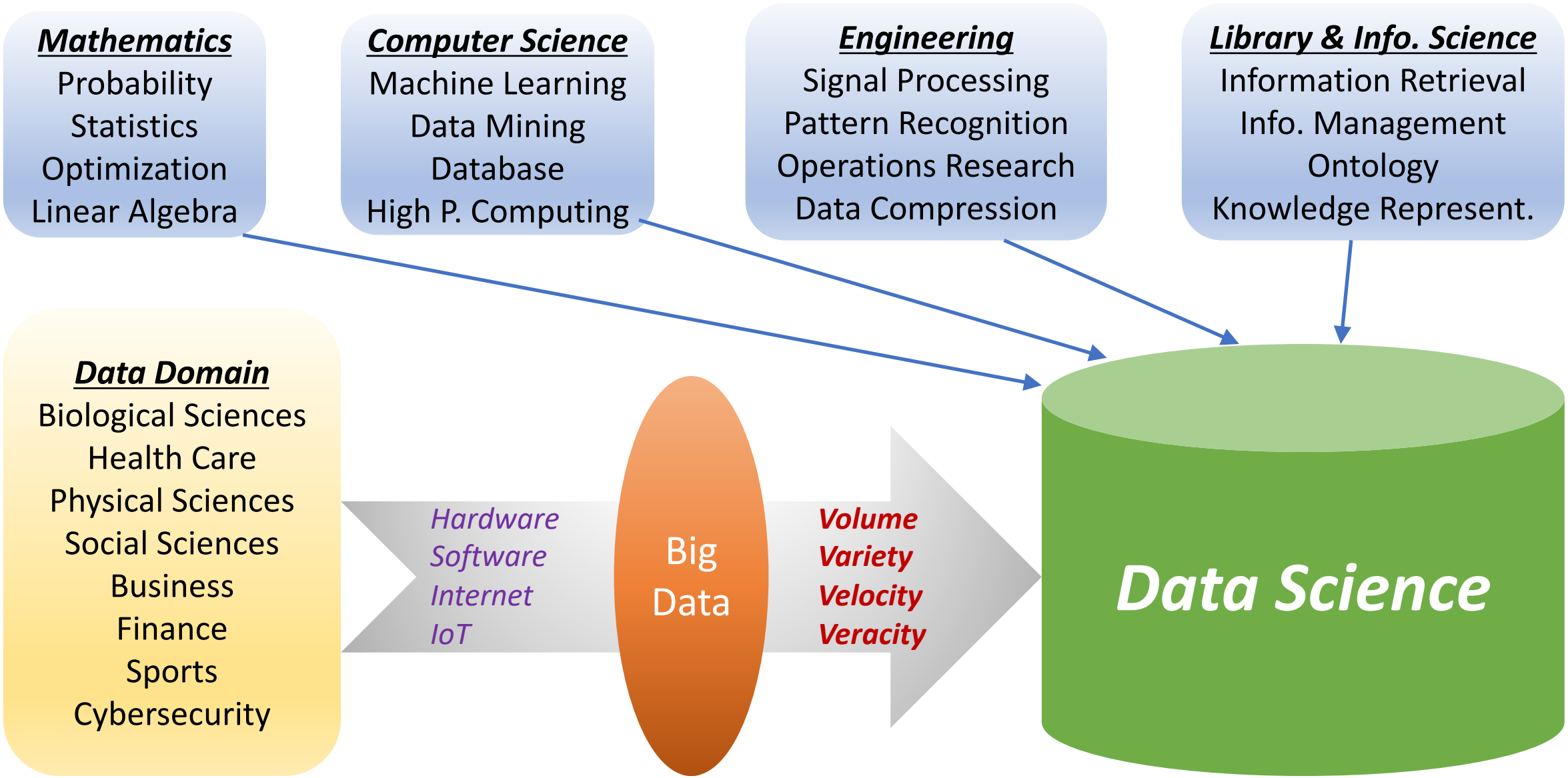
Data Science

- ***Highly interdisciplinary***
- ***No. 1 job on the market ****

* McKinsey, Forbes, Harvard Business Review, Glassdoor, CareerCast



What is Data Science?



What does a Data Scientist do?

- Understands the ***physical process*** that generates **data**
  - e.g., how a transmitted signal travels in air – ***wireless communications***, how people behave in stock market – ***economics***, how DNA transcribes RNA – ***genetics***, how a planet moves on its orbit – ***astronomy***

- Models data using **probability & statistics**

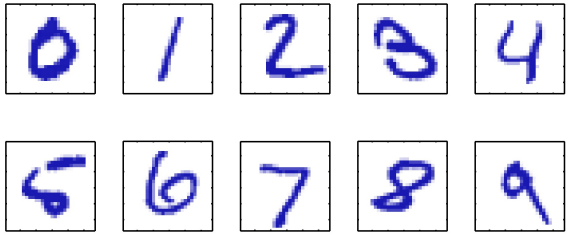
- Develops algorithms that
  - **learn** from data
  - **infer** about the data source

  (i.e., ***generalize*** the information contained in data to the data source)

- Discovers ***patterns/regularities*** in data

What is Machine Learning?

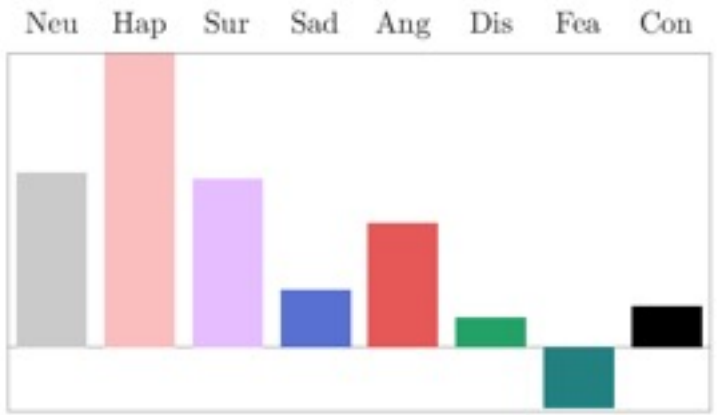
- Through algorithms, ***discover patterns*** in data, and use them to ***infer about the data source***, e.g.,
  - Handwritten digit recognition

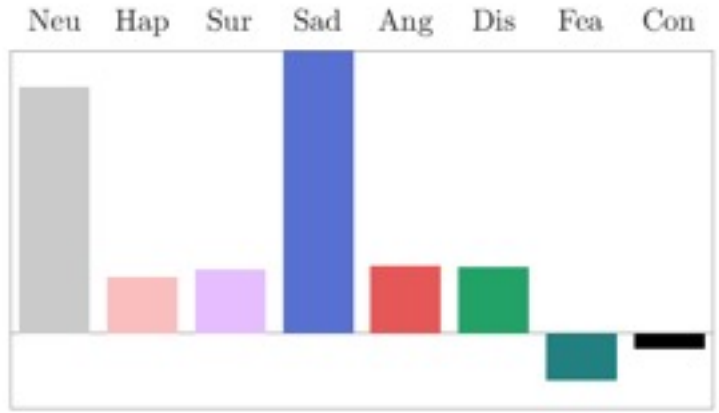


  - Object recognition

https://www.youtube.com/watch?v=n5uP_LP9SmM

  - Emotion recognition

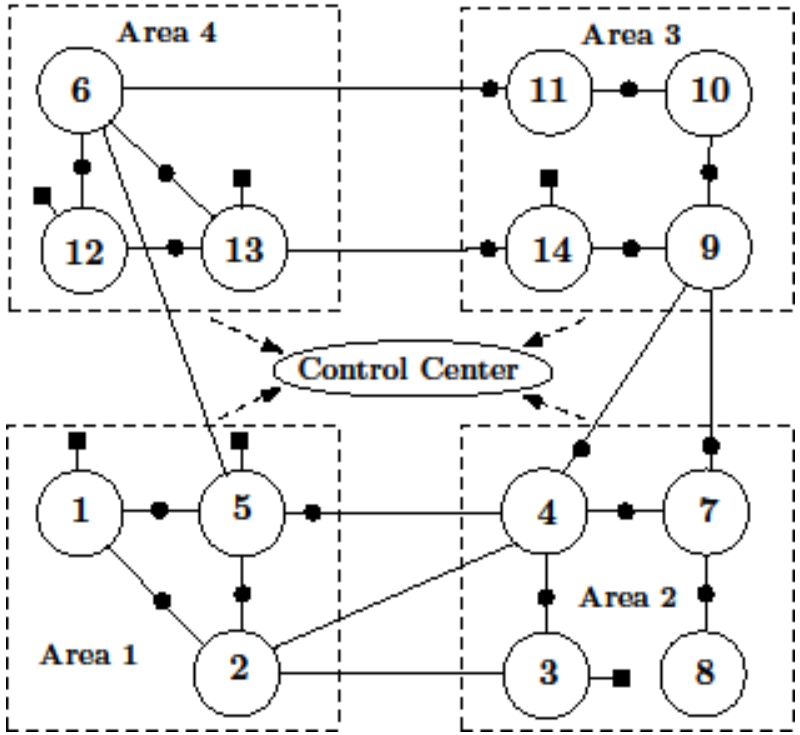


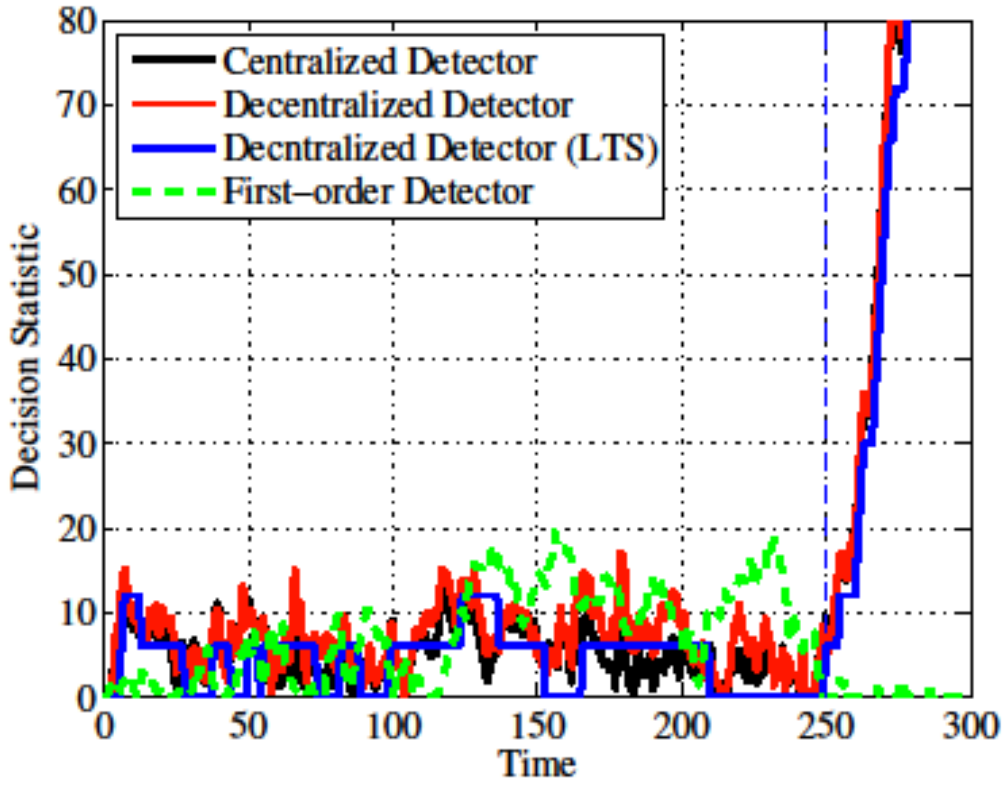


Example: Video Surveillance

Doshi, K. and Yilmaz, Y., 2020. Continual Learning for Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 254-255).

Example: Cyberattack Detection in Smart Power Grid

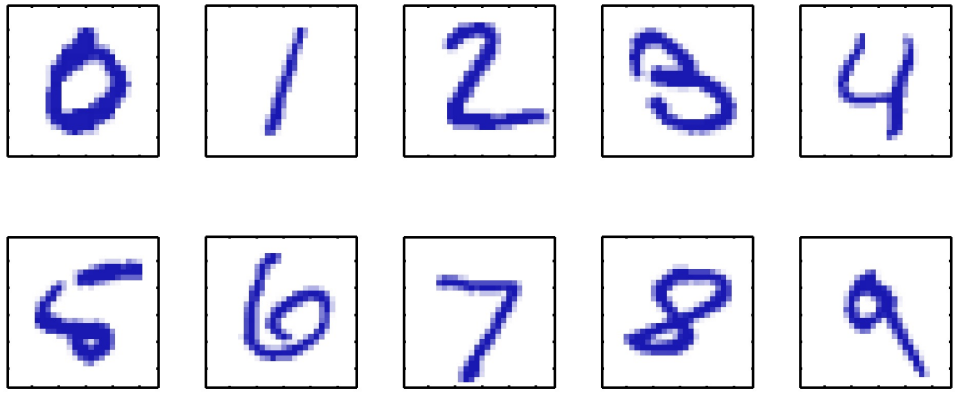


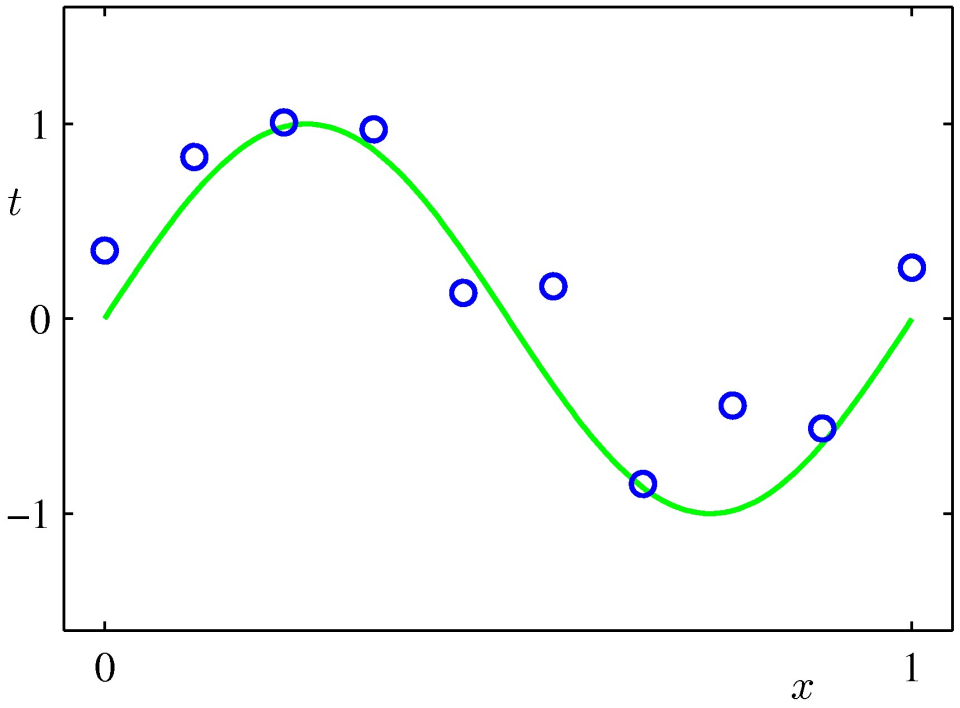


Li, S., Yılmaz, Y., & Wang, X. (2015). Quickest detection of false data injection attack in wide-area smart grids. *IEEE Transactions on Smart Grid*, 6(6), 2725-2735.

What is Machine Learning?

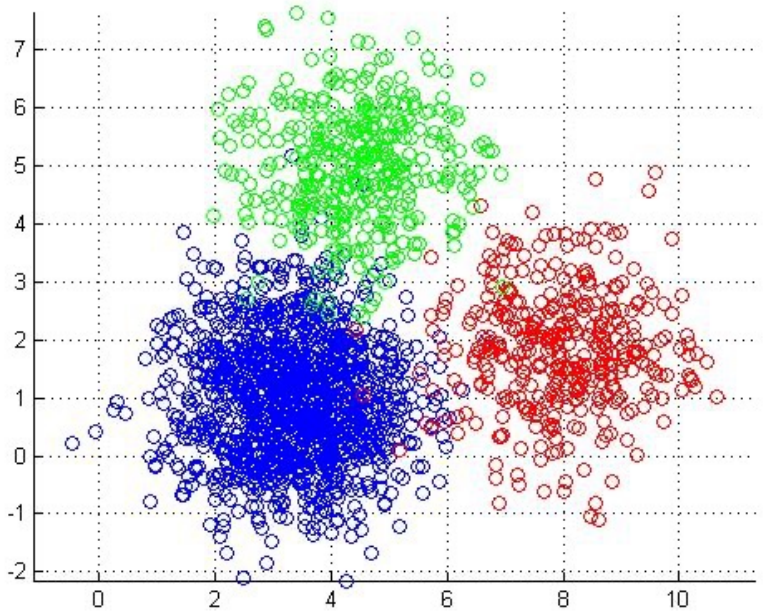
- Through algorithms, ***discover patterns*** in data, and use them to ***infer about the data source***, e.g.,
- ***Feature extraction:*** extract the *meaningful* part from each *object/instance* in data
  - *hand-designed* for a specific application OR
  - *learned from data* in an *unsupervised* fashion
  - ***very important!*** active research area, hardest part in big data problems
- ***Supervised Learning:*** learn a model from ***training data*** *with ground truth* available and use the learned model for ***new/test data***
  - ***Classification:*** assign each object to a category, e.g., handwritten digit recognition, face recognition
  - ***Regression:*** estimate relationships between response and explanatory variables, e.g., prediction of travel times in traffic, estimation of class probabilities

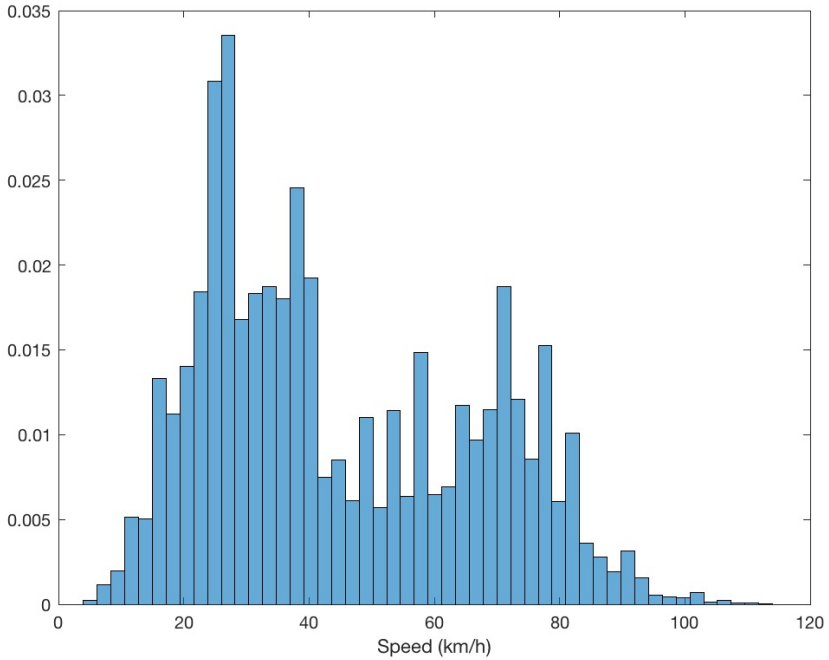




What is Machine Learning?

- ***Unsupervised Learning:*** *no ground truth* in ***training data***





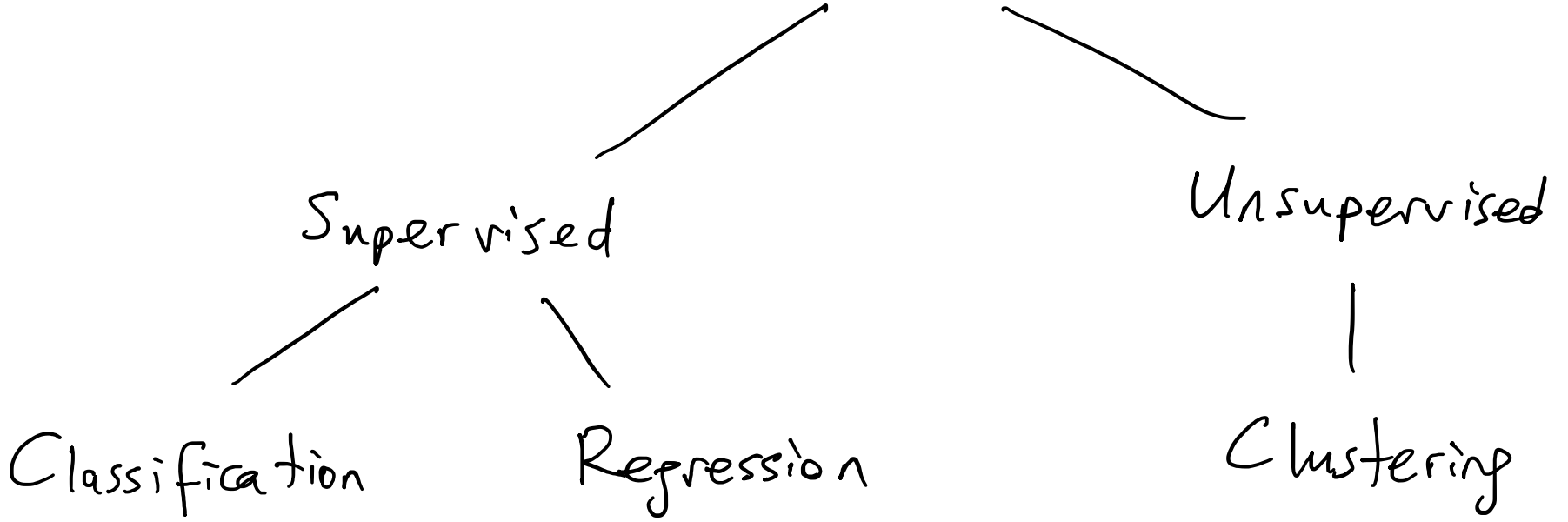
  - ***Clustering:*** group similar objects together
  - ***Density estimation:*** estimate the distribution of data within the space of possible values

- ***Semi-supervised Learning:*** *labeled and unlabeled* data *together* in training

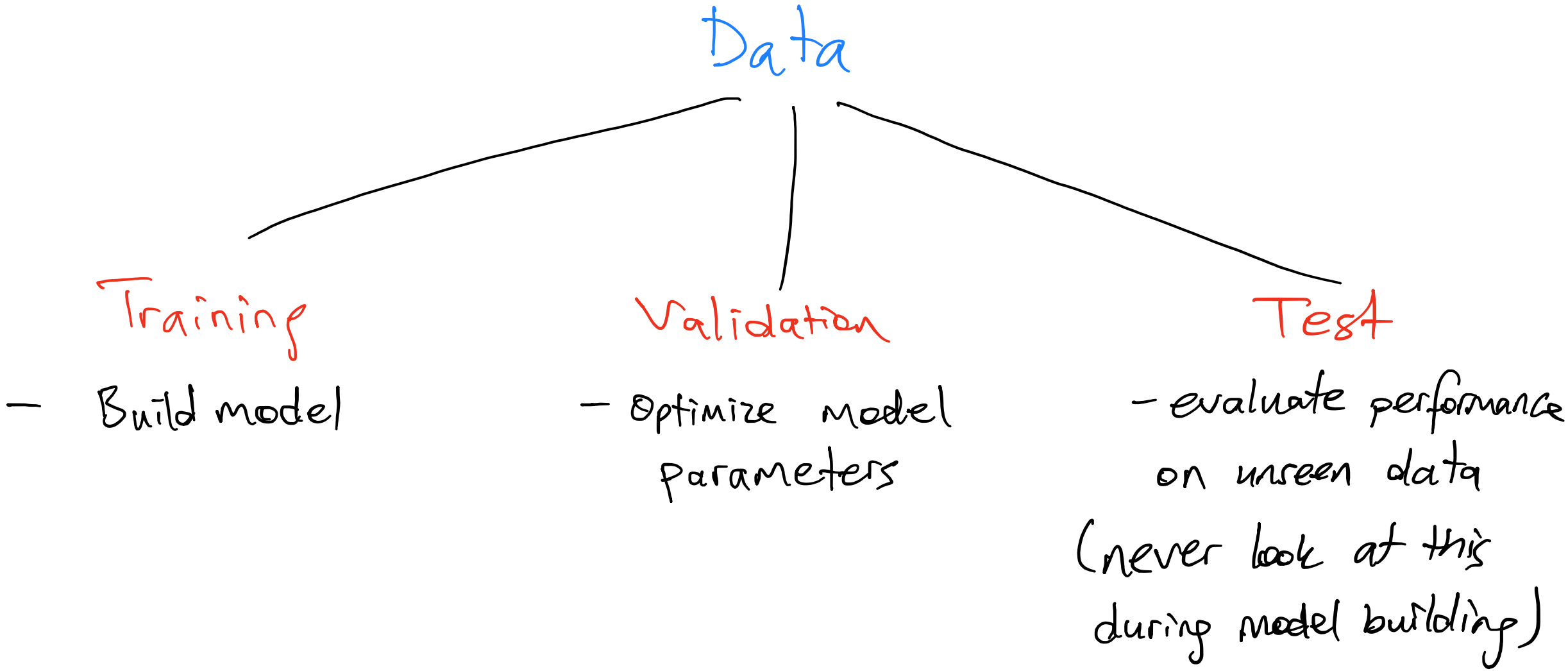
  - ***Anomaly detection:*** detect instances that significantly deviate from standard patterns

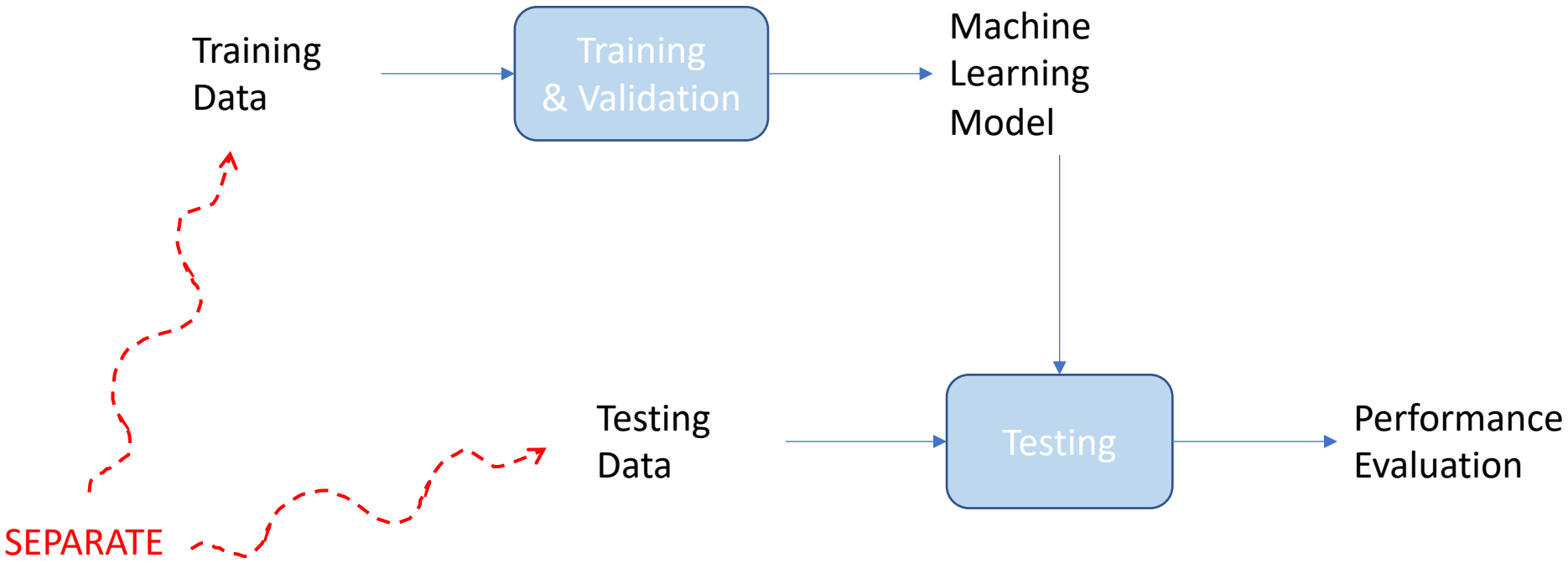
Input : Data

ML



Output : Labels, Var. values, Density, Simpler Data, ...


Data
Training
- Build model
Validation
- Optimize model parameters
Test
- evaluate performance on unseen data
(never look at this during model building)

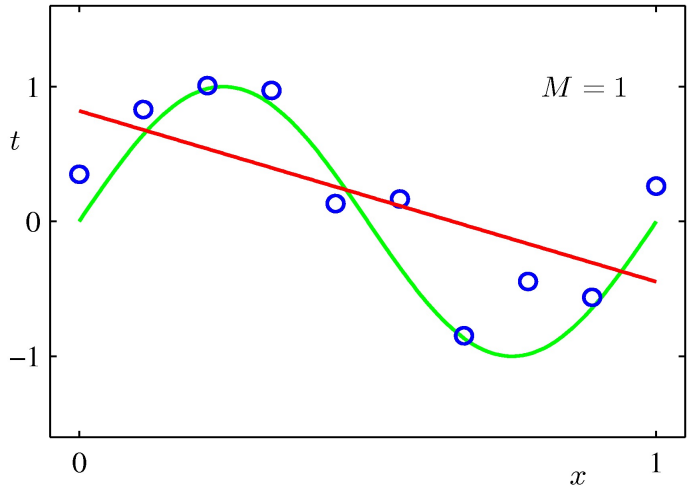

Training
Data
Training
& Validation
Machine
Learning
Model
Testing
Data
Testing
Performance
Evaluation
SEPARATE

Important : Generalize, do NOT memorize the training dataset } overfitting

Overfitting :
- Regularization
- Validation (Cross-validation)

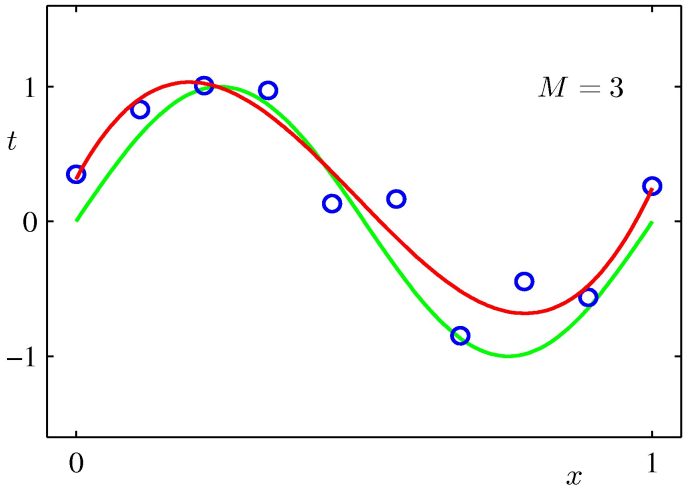
What is Machine Learning?

- ***Objective:*** Select model that *generalizes* well to *unseen possible data*



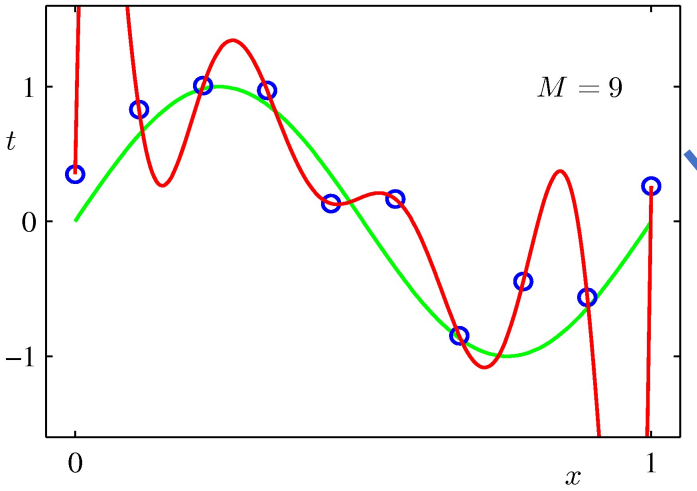
Poor fit & generalization

Model too simple!



Good fit & generalization

Model good enough!



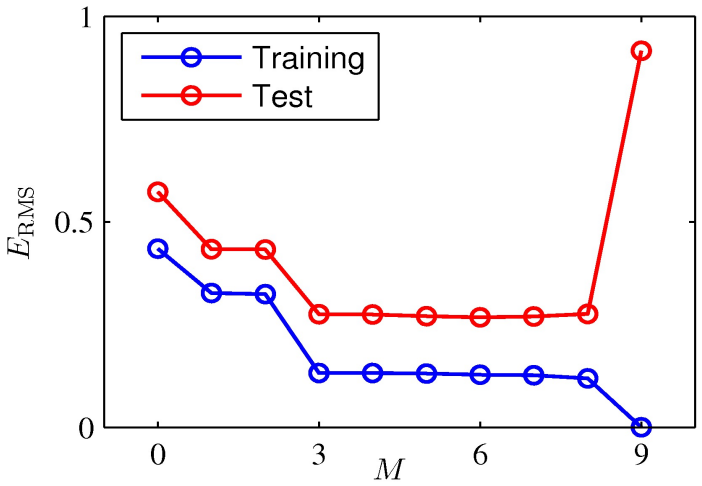
Perfect fit, Poor generalization

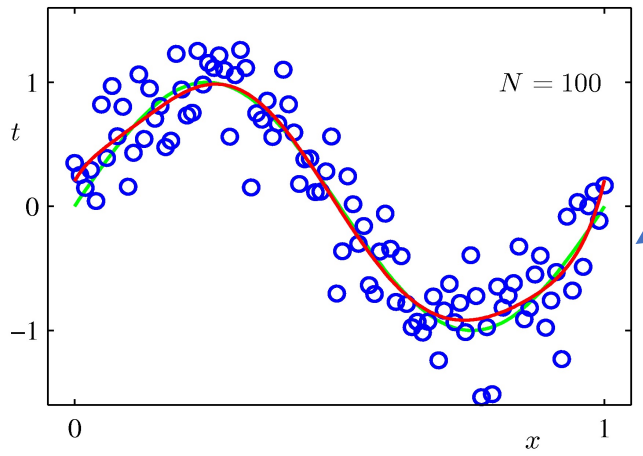
Model too complex, fits noise!

OVER-FITTING !

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2}$$

$$y(x_n,\mathbf{w}) = \sum_{j=0}^{M} w_j x_n^j$$

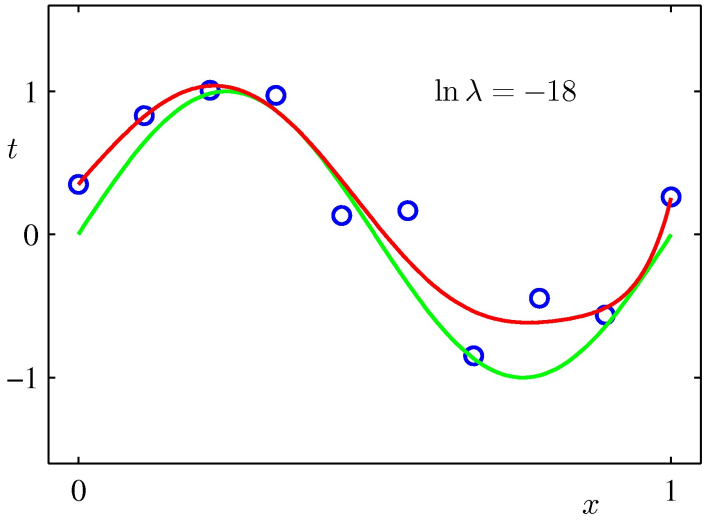


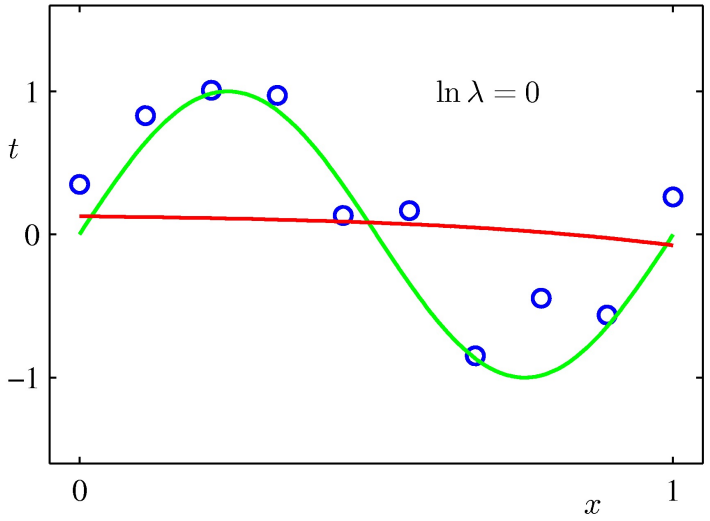


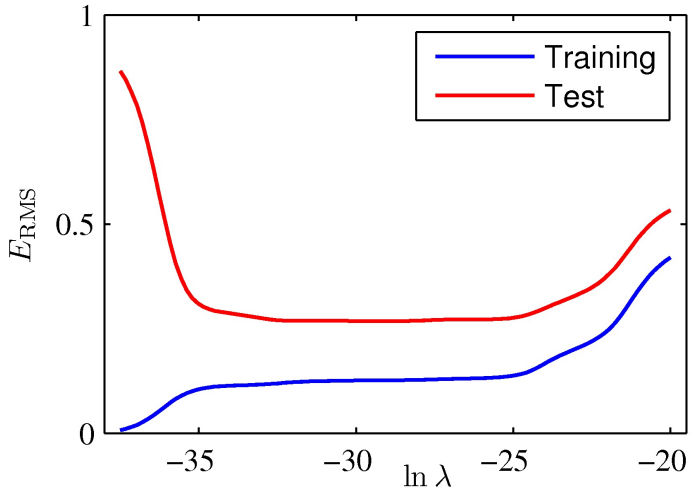
What is Machine Learning?

- ***Regularization:*** avoid over-fitting by adding a *penalty term to error function* to *shrink* coefficients (***shrinkage***)

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2}$$



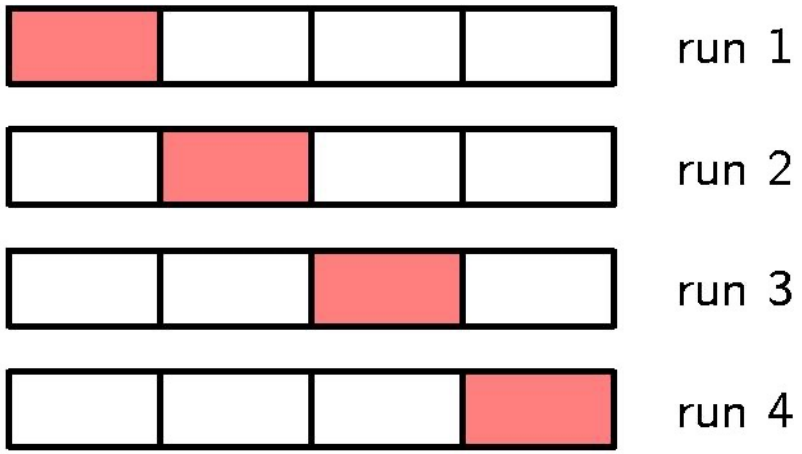




- ***Validation set:*** partition available data into a training set and a validation set to optimize model complexity (M in previous slide)

What is Machine Learning?

K-fold Cross Validation



Python Basics

- Open source
  - python.org
  - dominant and fast growing programming language for data science
- **Anaconda** bundles all the essential Python packages for data science
  - Spyder
  - Jupyter
  - …
- Most popular libraries
  - NumPy
  - SciPy
  - matplotlib
  - pandas
  - scikit-learn
- Tutorial on Thursday – Install Python 3 to work together on the tutorial

Next Class: Python Tutorial

- Install Anaconda or another Python 3 medium (Not Python 2)

…and I look forward to a great semester together

Thanks!