POLITECNICO
MILANO 1863

Machine Learning Methods for Communication Networks and Systems

Francesco Musumeci

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)

Politecnico di Milano, Milano, Italy

Course Introduction

Welcome to the course!

- The lecturer: Francesco Musumeci
  - Office: DEIB building 20, 3rd floor, room 329
  - Contact: francesco.musumeci@polimi.it
  - Web page: https://musumeci.faculty.polimi.it/
  - Main research interests:
    - Machine-Learning-assisted networking
    - 5G and beyond networking
    - Software Defined Networks (SDN) and Network Function Virtualization (NFV)
    - Optical networks architectures
    - Network disasters resilience

Course schedule

- Week 1
  - Dec. 13th h. 10-13 + 14-16 (Room Alpha, Bd. 24)
  - Dec. 14th h. 9-13 (Seminar room N. Schiavoni, Bd. 20)
  - Dec. 16th h. 9-13 (Seminar room N. Schiavoni, Bd. 20)
  - Dec. 17th h. 9-13 (Seminar room N. Schiavoni, Bd. 20)
- Week 2
  - Dec. 20th h. 9-13 (Seminar room N. Schiavoni, Bd. 20)
  - Dec. 21st h. 9-13 (Seminar room N. Schiavoni, Bd. 20)

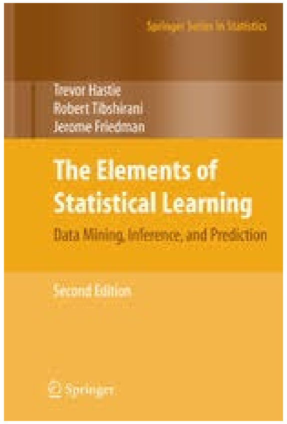
Covered topics

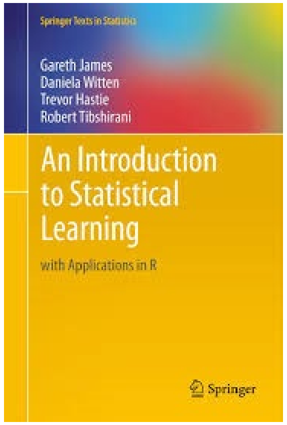
- The course is organized into two main parts
- Part 1: overview on Machine Learning methodologies
  - Basic concepts (supervised/unsupervised learning, bias/variance trade-off, etc.)
  - Linear and logistic regression
  - Neural Networks
  - Support Vector Machine
  - Clustering
  - …
- Part 2: applications of ML to communication nets & systems
  - Part 2a): Physical layer domain use cases
    - QoT estimation, optical power control, modulation format recognition…
  - Part 2b): Network layer domain use cases
    - Traffic prediction, pattern analysis extraction, failure management, virtual topology design,…

Note: this is NOT a “pure” Machine Learning course.
The objective is to learn **how to apply** ML to **your** research problems in comm nets and systems

Course material

- Lecture slides
- Suggested research papers
- Books (general refs. for ML):
  - T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning”, (ESL) Ed. Springer
  - G. James, D. Witten, T. Hastie, R. Tibshirani, “An Introduction to Statistical Learning with Applications in R”, (ISLR) Ed. Springer
- Prof. Andrew Ng lectures (Stanford University)
- … Google it!





Course objectives & evaluation

- At the end of the course you should be able to:
  - identify communication networks & systems use cases where ML can be useful
  - apply proper ML techniques to the use cases
  - evaluate/compare the performance of various ML strategies
  - understand how to select *important* data to use in a ML algorithm
- Two alternatives for the evaluation (student's choice)
  1. Research overview: discuss 2 different research papers on the course subject (not seen during the course)
     - Act as a reviewer: present the papers with criticism highliting pros/cons
  2. Project to be agreed with the instructor (can be individual or in groups of max 2/3 students)
     - Deliverables: source code and datasets, short report, ppt presentation

Before we start…

- Any question?

What is Machine Learning?

- *"Field of study that gives computers the ability to learn without being explicitly programmed" (A. Samuel, 1959)*
- *"A type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed"*
- *"Teaching a computer to automatically learn concepts through data observation"*
- *…*
- For our purposes: An *instrument* to build models which allow us to make decisions and to infer statistical properties on our data

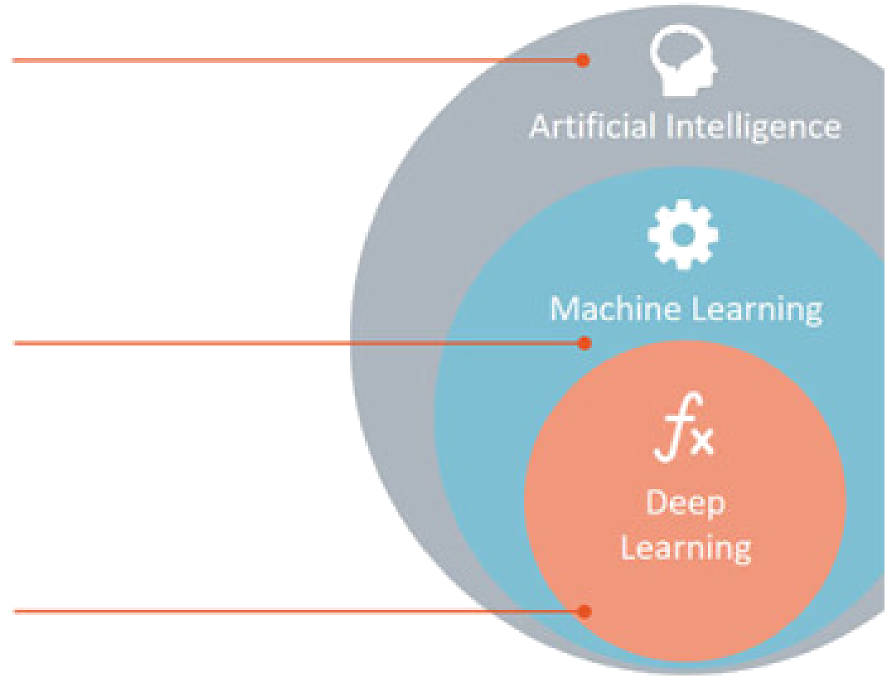
  …in the context of communication networks and systems
- Why all this attention?
  - Huge availability of data
  - Improved efficiency in computational capabilities
- Sometimes confused with other terms: AI, Deep Learning, Data Analytics, Data Mining, etc.

Many definitions with blurred borders

Artificial Intelligence
Any technique which enables computers to mimic human behavior.

Machine Learning
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning
Subset of ML which make the computation of multi-layer neural networks feasible.



https://www.kdnuggets.com/2017/07/rapidminer-ai-machine-learning-deep-learning.html

Main categories of ML algorithms (1)

- Supervised learning: we are given "labeled" data (i.e., "ground truth" input/output relationship)
  - Main objective: given a new set of input(s), predict a corresponding output response
  - Regression: output value is continuous
  - Classification: output value is discrete or "categorical"
- Unsupervised learning: available data is not "labeled"
  - Main objective: derive structures (patterns) from the available data
  - Clustering: finding "groups" in our data, according to a similarity measure
  - Anomaly detection (sometimes seen as a semi-supervised method)

Main categories of ML algorithms (2)

- Semi-Supervised learning
  - Hybrid of previous two categories
  - Most of the training samples are unlabeled, only few are labeled
  - Main objective: exploit information from unlabeled data to improve accuracy in supervised learning problems
    - Self-training: start with labeled data, then label unlabeled data based on first phase
    - Common when labeled datasets are limited or expensive

- Reinforcement learning
  - Available data is not "labeled"
  - Main objective: learn a *policy*, i.e., a mapping between inputs/states and actions performed over a certain ***environment***
  - Behavior is refined through **rewards** coming from the system

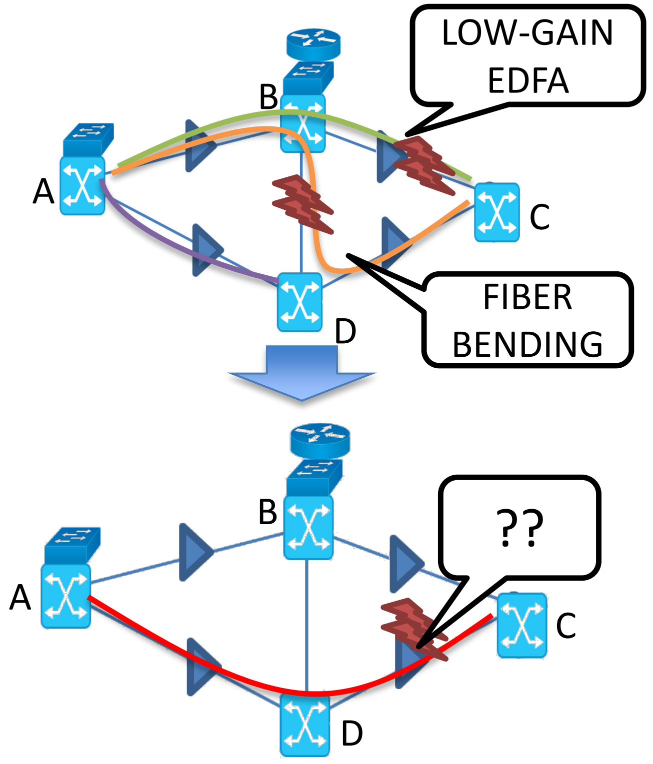
- Supervised learning: some examples

Example in the optical network domain

Supervised learning: discriminate failure types (failure *identification*)

TRAINING PHASE

Sample no. (lightpath)	Wavelength	Route	Modulation format	BER	Failure type
1	1550 nm	A-B-C	BPSK	Sharp Increase	**Faulty EDFA**
2	1553 nm	A-B-D-C	QPSK	Gradual drift	**Fiber bending**
3	1556 nm	A-D	8-QAM	flat	**None**
…	…	…	…	…	**…**



VALIDATION/TEST PHASE

New fault: wavelength= 1559, route= A-D-C, modulation format= QPSK, BER= cyclic drift
➔ failure type=?

Supervised learning: other examples

1. Given traffic exchanges to/from a Data Center during last week/month/year
   – Predict traffic for the next period (regression)
   – Predict if available resources will be sufficient (classification)
2. Given SNR observed at a receiver
   – Predict if quality of transmission will be degraded (e.g., due to some occurring failure)
3. Other domains
   – Speech recognition
   – Spam classifier
   – House prices prediction/estimation

Terminology – Regression

- Different terms for the same concepts

- Model
- Hypothesis (*h*)
- Function (*f*)

- $\underline{x}$
- Input
- Predictor
- Variable
- Feature

- y
- Output
- Prediction
- Response

- **Labeled** data-set, where ($\underline{x}$,y)-s are called:
  - observations
  - examples
  - samples

  y is a <u>real</u> (continuous) value

Terminology – Classification

- Different terms for the same concepts

Note: in both regression and classification, some (or all) of the **inputs** X can be categorical!

- $\underline{x}$
- Input
- Predictor
- Variable
- Feature

→

- Model
- Hypothesis (*h*)
- Function (*f*)
- Classifier

Classification Algorithm

→

- y
- Output
- Prediction
- Response
- Class
- Group

- **Labeled** data-set, where ($\underline{x}$,y)-s are called:
  - observations
  - examples
  - samples

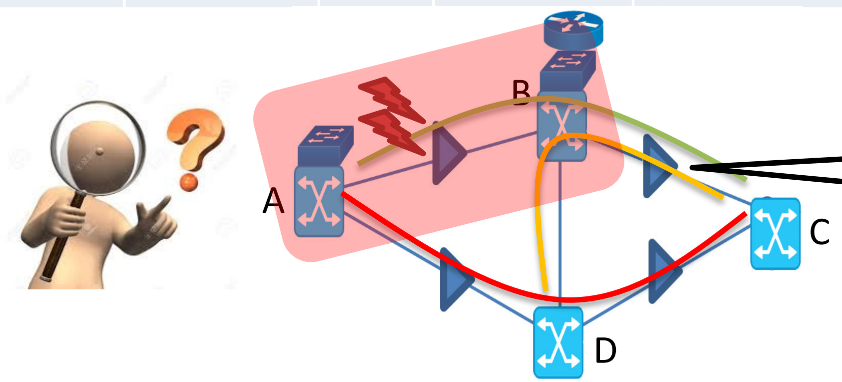
  y is a discrete value (or even "categorical")

- Unsupervised learning: some examples

Example in the optical network domain

Unsupervised learning (anomaly detection): detect/localize failures in optical networks

Sample no. (lightpath)	Wavelength	Route	Modulation format	OSNR
1	1550 nm	A-B-C	BPSK	Fluctuating w/ high spikes
2	1553 nm	D-B-C	BPSK	Fluctuates within [mean +/- std]
3	1556 nm	A-D-C	QPSK	Fluctuates within [mean +/- 2*std]
…	…	…	…	…



1) Most of the lightpaths have OSNR within a certain range (e.g., lightpaths 2 and 3)

2) Lightpath 1 has similar features (hop-length, modulation, adjacent spectrum) wrt lightpaths 2 and 3, but different OSNR behaviour
→ **ANOMALY (=FAILURE)**

3) Lightpath 1 shares link B-C with another **failure-free** lightpath (lightpath 2)
→ **FOCUS ON LINK A-B TO IDENTIFY FAILURE**

Unsupervised learning: other examples

1. Given traffic profiles in different mobile cell sites
   – Understand if some cells provide similar behaviour (patterns)
     - They might cover same type of urban areas (theatre, cinema, stadium…)
     - This information can be used to make network resources planning
2. Given failures in a certain radio link
   – Group similar failures to define new classes of problems (e.g., due to rain, due to large obstacles, hardware failures, etc.)
3. Other domains
   – Group people according to their interests to improve advertisement
   – Group together different genes if they provide similar information

Terminology – Clustering

- Different terms for the same concepts

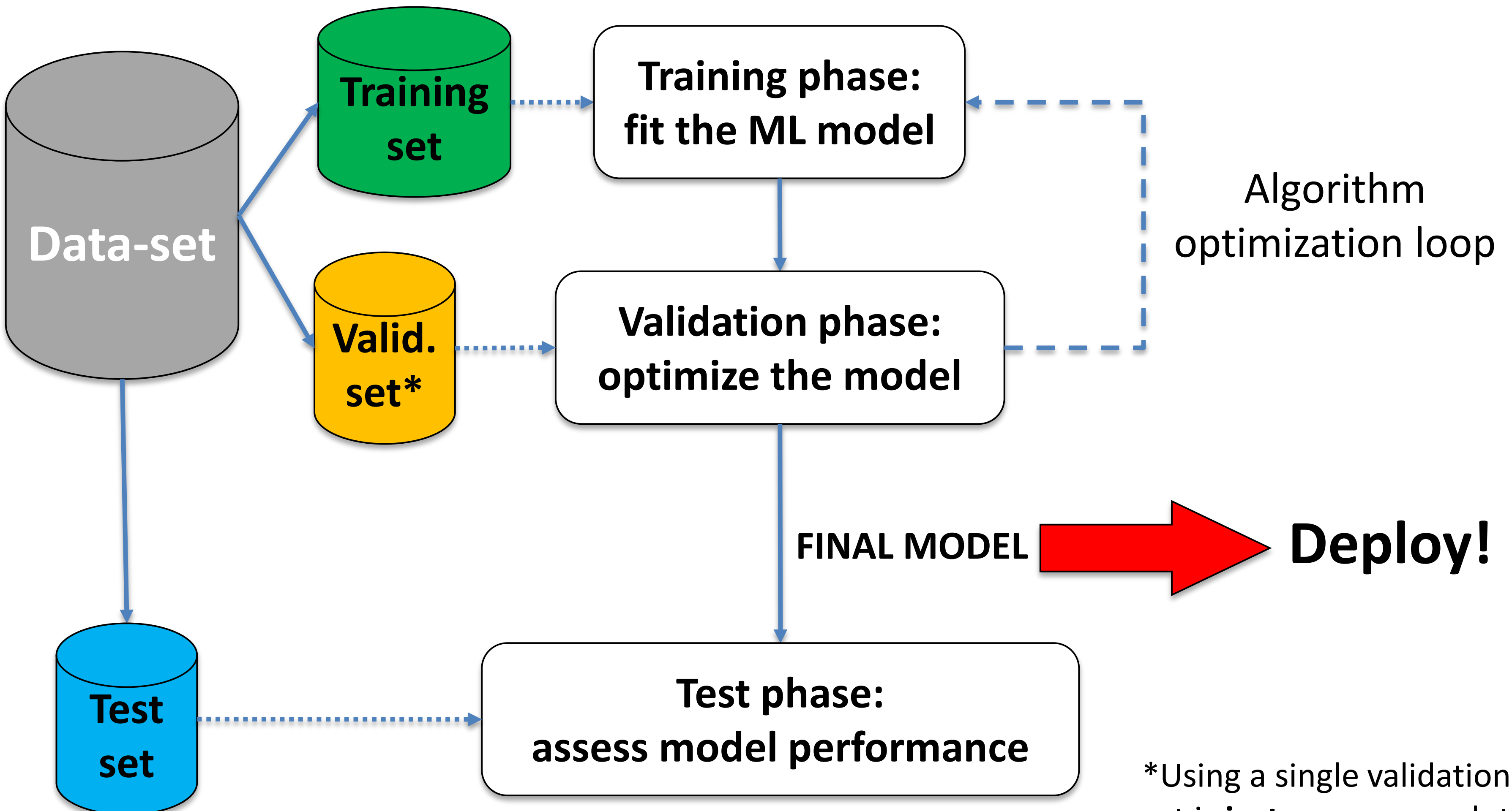
- $\underline{x}$
- Input
- Predictor
- Variable
- Feature

- Model
- Hypothesis (*h*)
- Function (*f*)

- Trends
- Statistical properties
- Outputs
- Partitions
- Groups

- **UN-labeled** data-set, only containing $\underline{x}$-s, known as:
  - observations
  - examples
  - samples

A «big picture» on a ML-based framework



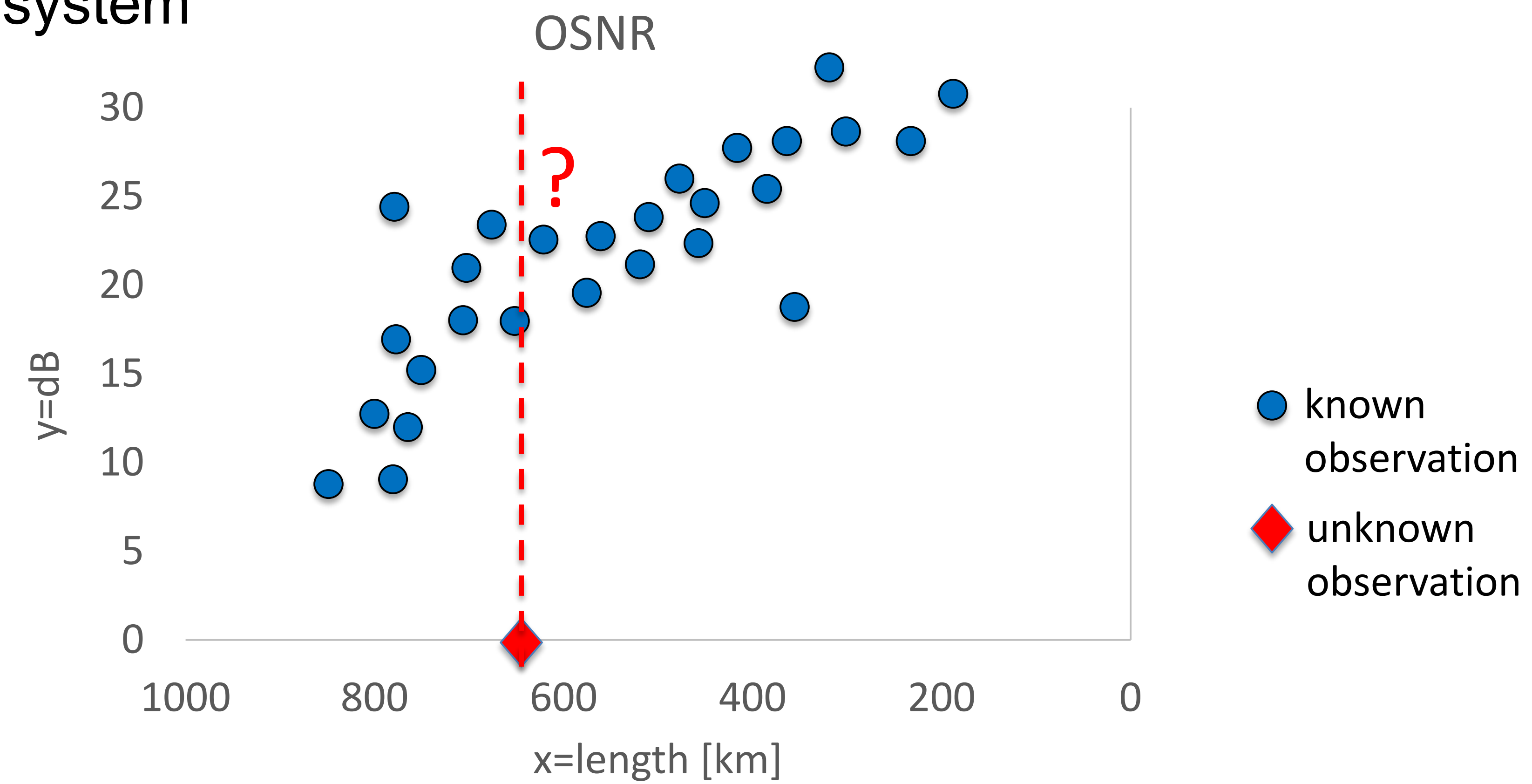
*Using a single validation set is **just one** approach to optimize a ML algorithm

Some basic concepts in ML

- Why do we want to estimate the behaviour $y=f(\underline{x})$?
  - Prediction
    - we want to actually know the "exact" (as accurate as possible) value of y given a new input $\underline{x}$
  - Inference
    - we want to understand how "in general" a certain quantity y behaves as $\underline{x}$ varies
    - which of the x in $\underline{x}$ is the most relevant?
    - is the relation between any of the *x-s* and y linear or is it more complex?
- Trade-off: prediction accuracy vs model interpretability
  - Flexible (more complex) models have high prediction accuracy but low interpretability
  - Simple models (e.g., linear) have high interpretability but low accuracy

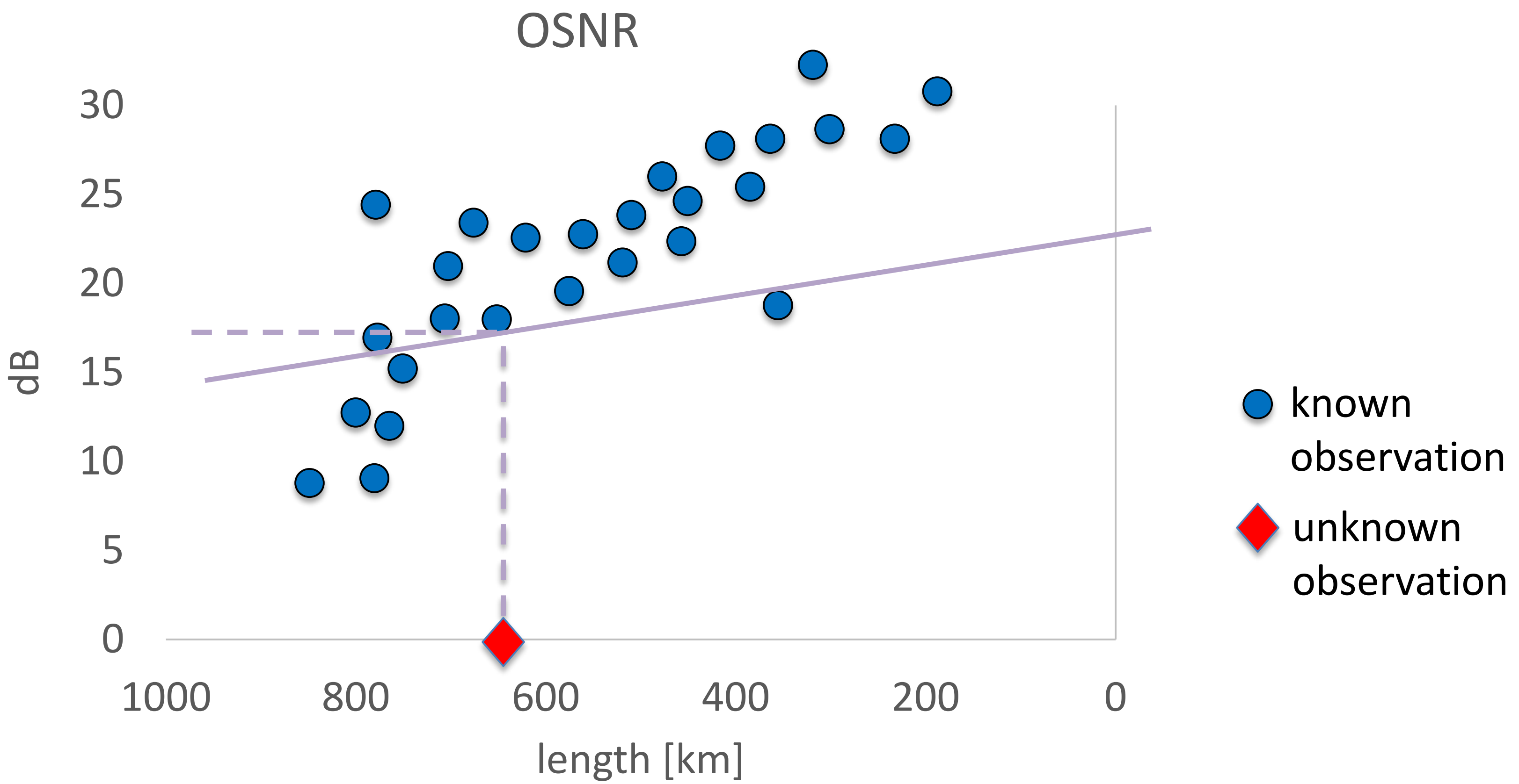
An example of fitting a model

- Suppose we want to predict the OSNR (optical signal-to-noise ratio, i.e., "quality") given the length of a transmission system



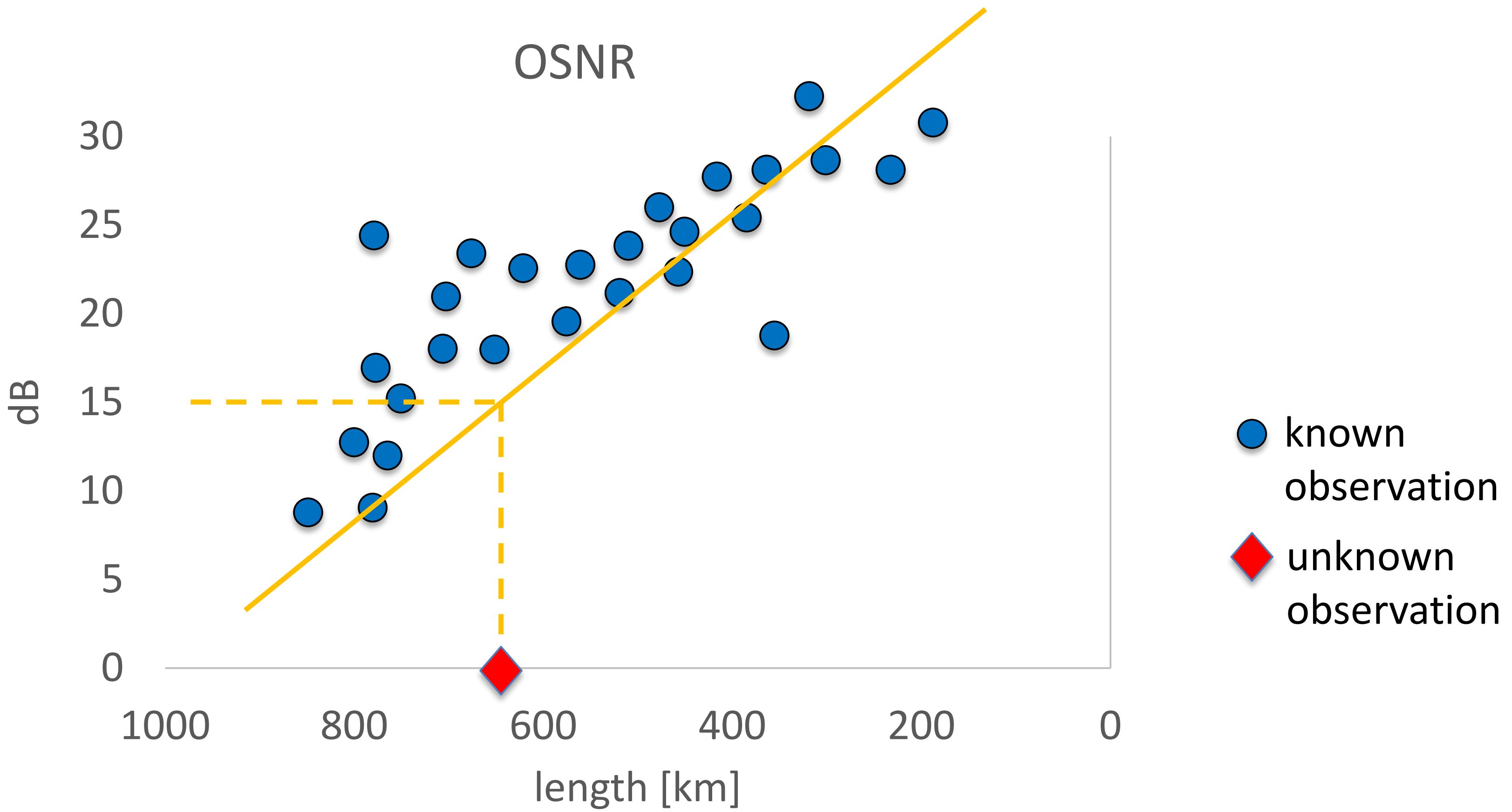
An example of fitting a model

- Trend #1: 17 dB



An example of fitting a model

- Trend #2: 15 dB



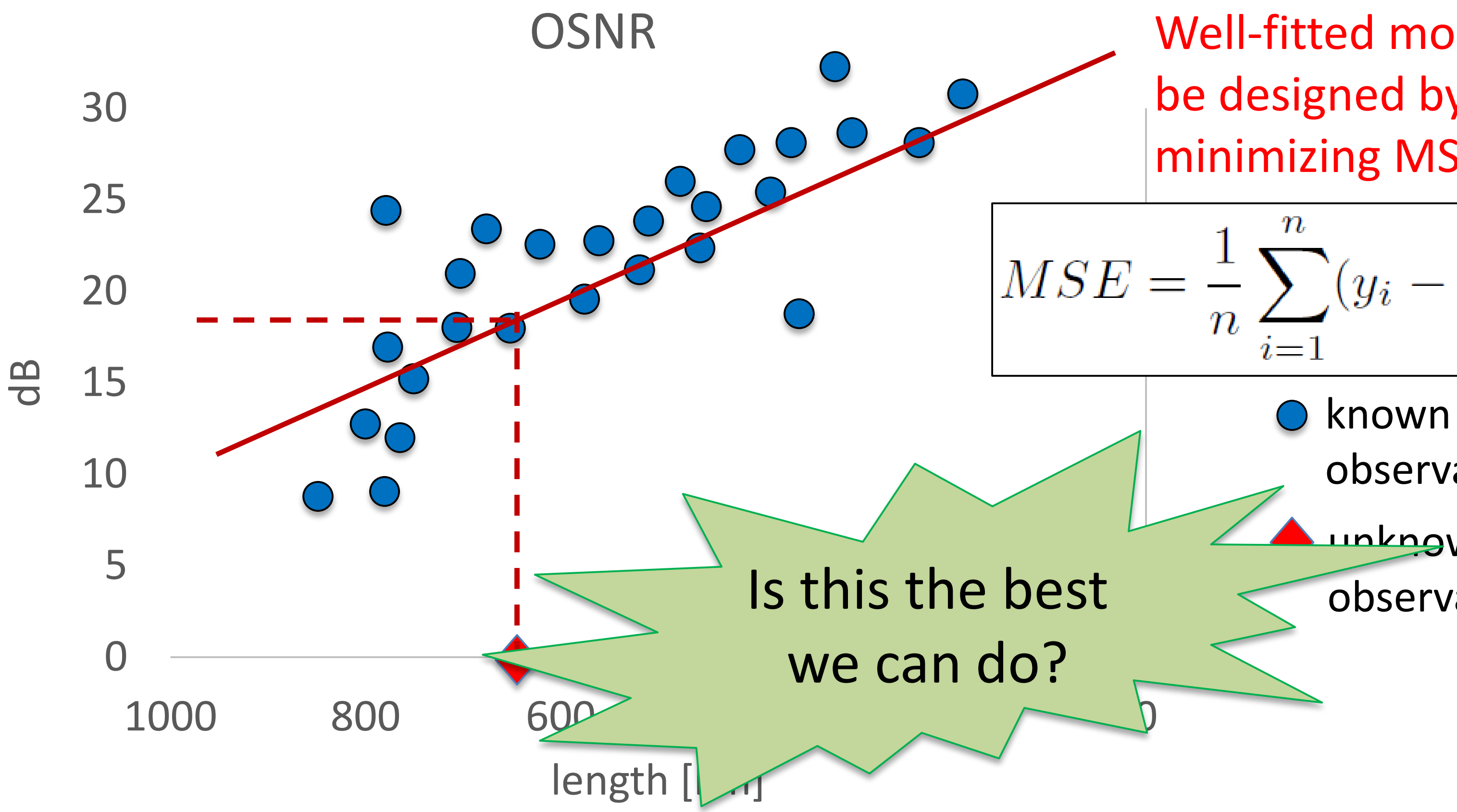
An example of fitting a model

- Trend #3: 18.5 dB

Which of the 3 predictions is correct (most appropriate)?

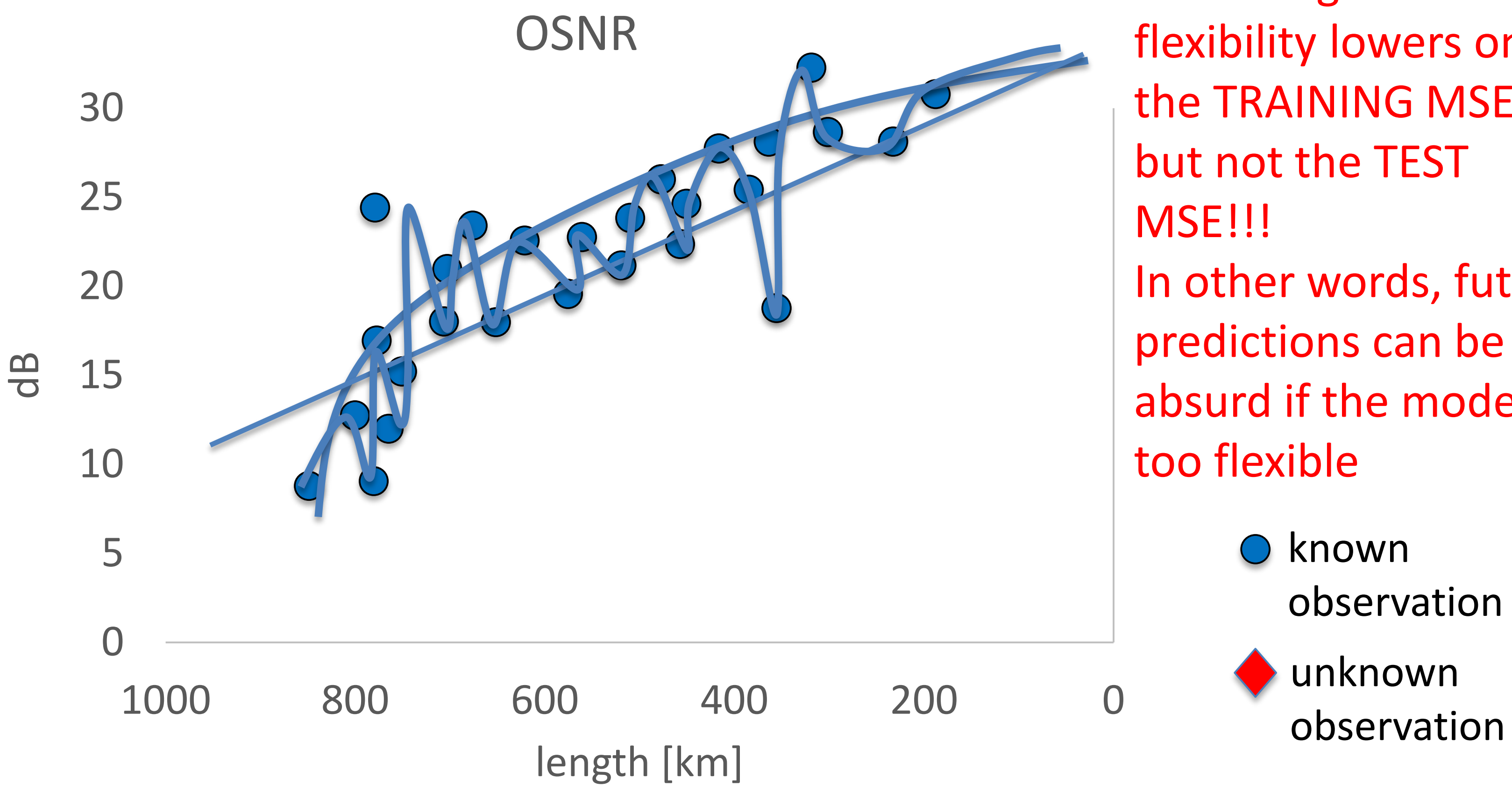
Well-fitted models can be designed by (e.g.) minimizing MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$



An example of fitting a model

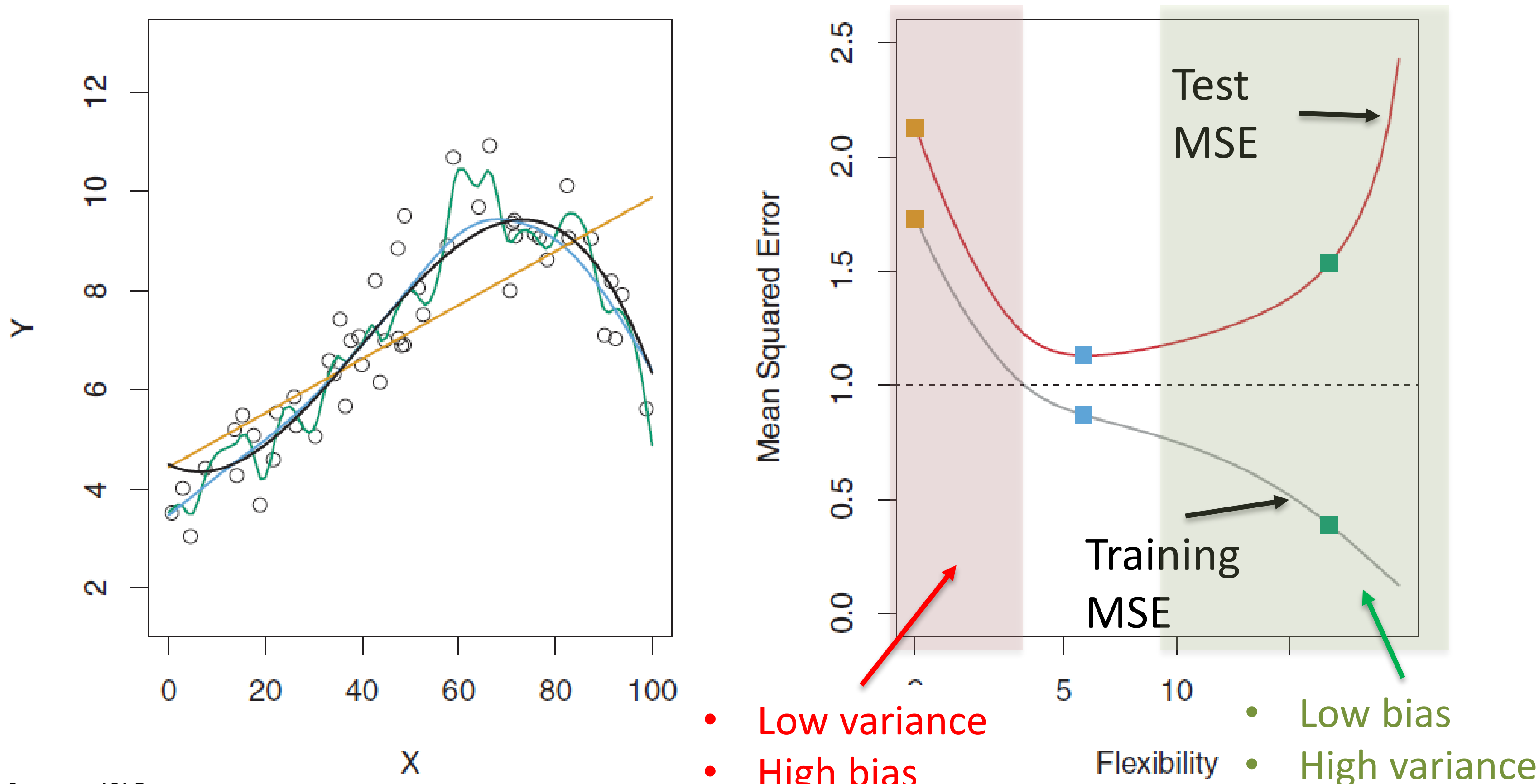
- Suppose trend #3 is the best (lowest MSE) linear model
- Why linear?



Increasing model flexibility lowers only the TRAINING MSE, but not the TEST MSE!!!
In other words, future predictions can be absurd if the model is too flexible

Bias/variance trade-off (1)

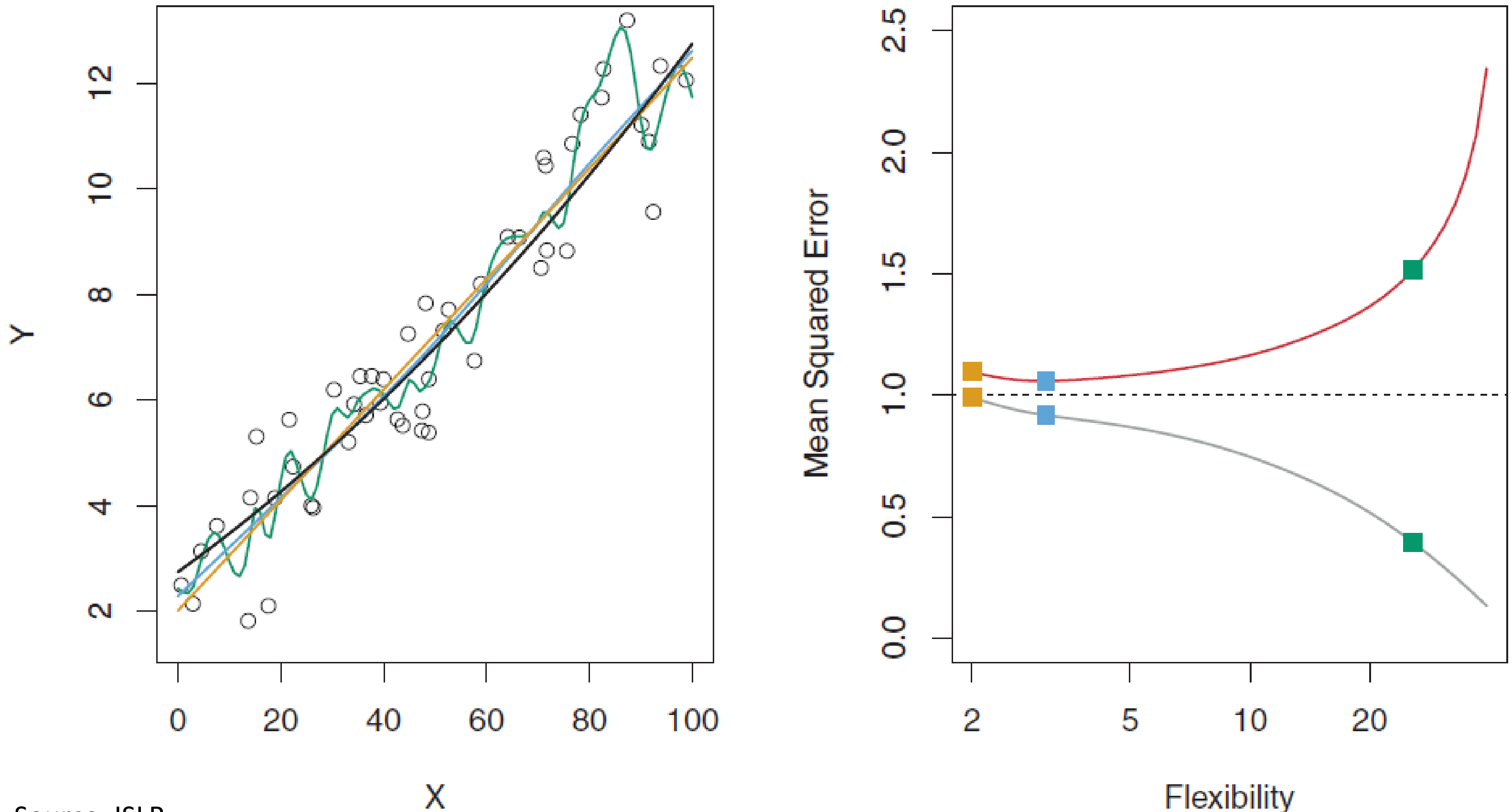
- Bias vs variance in regression problems
  - Increasing model flexibility (e.g., via higher-degree polinomials)



Source: ISLR

Bias/variance trade-off (2)

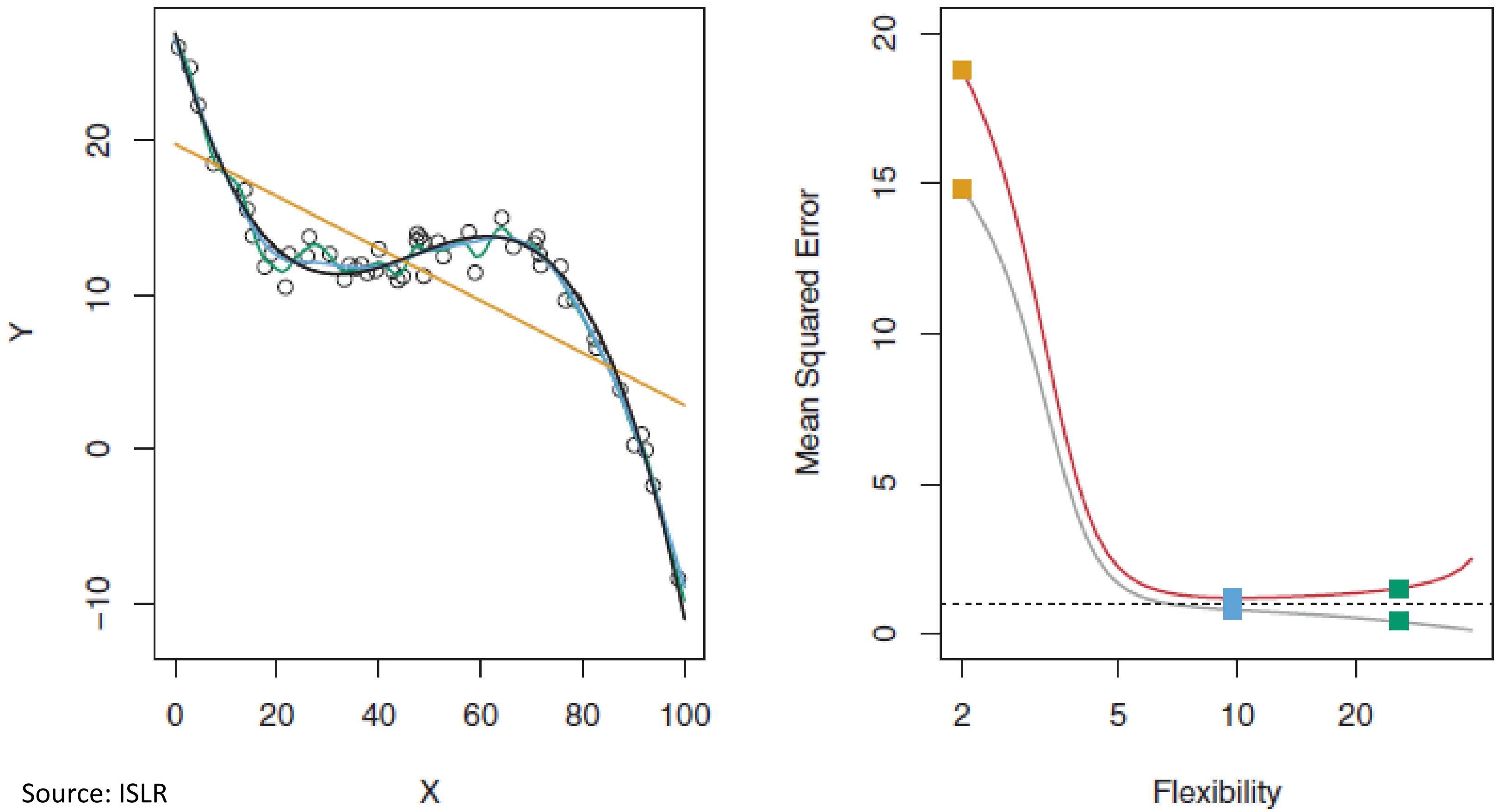
- Bias vs variance in regression problems
  - Increasing model flexibility (e.g., via higher-degree polinomials)



Source: ISLR

Bias/variance trade-off (3)

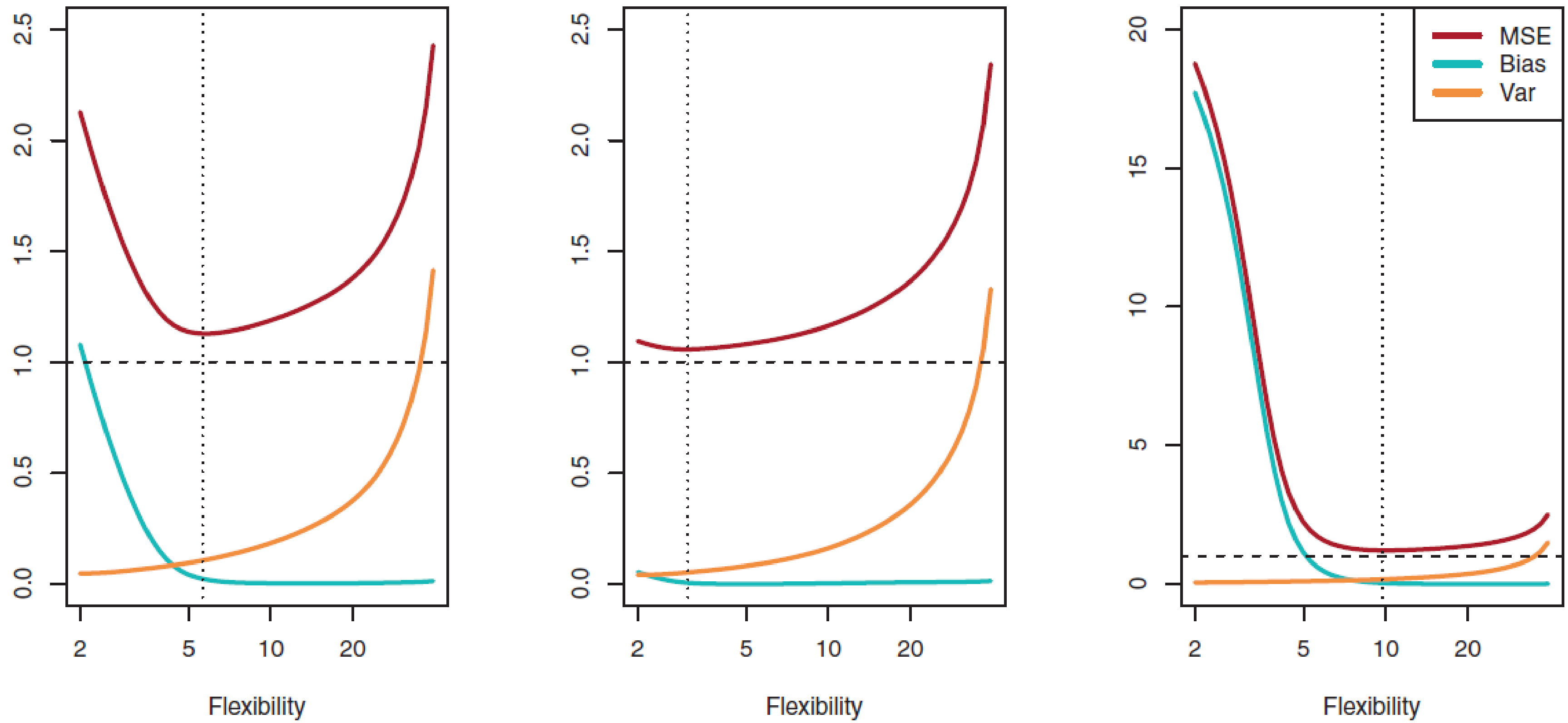
- Bias vs variance in regression problems
  - Increasing model flexibility (e.g., via higher-degree polinomials)



Source: ISLR

Bias/variance trade-off (4)

- Bias vs variance in regression problems
  - Increasing model flexibility (e.g., via higher-degree polinomials)



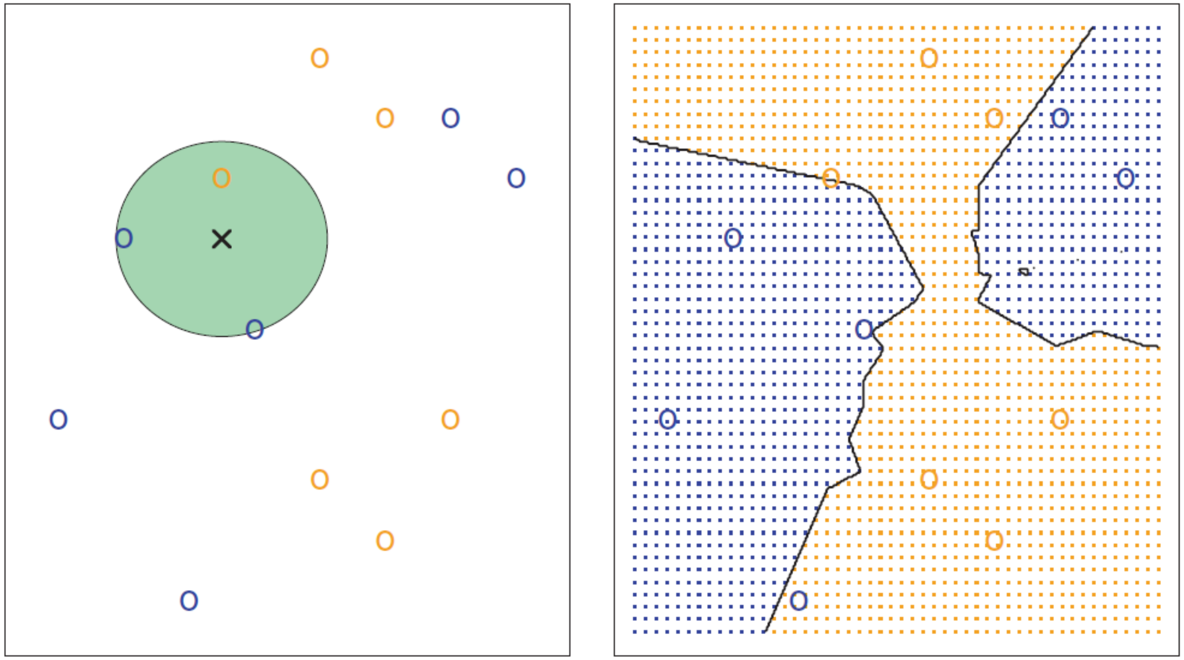
Source: ISLR

Bias/variance trade-off (5)

- Bias vs variance in classification problems
- How can we measure the classification error?
  - MSE is not applicable
  - ERROR RATE: fraction of misclassified points: $$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$
  - *I(condition)=1* if «condition» is true, =0 otherwise

- Example:

K-nearest neighbors (K=3)



Source: ISLR

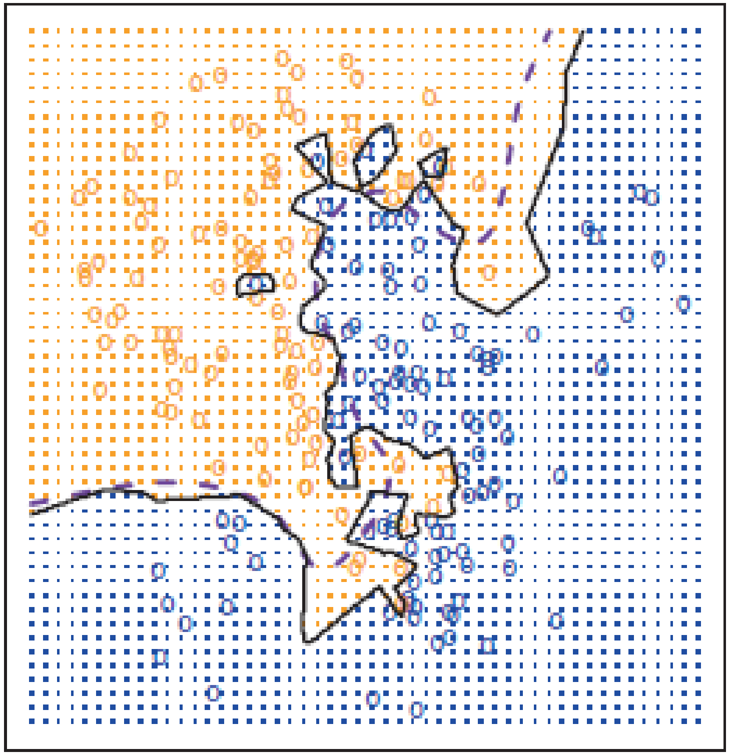
Bias/variance trade-off (5)

- Bias vs variance in classification problems
  - Varying the number of neighbors, K

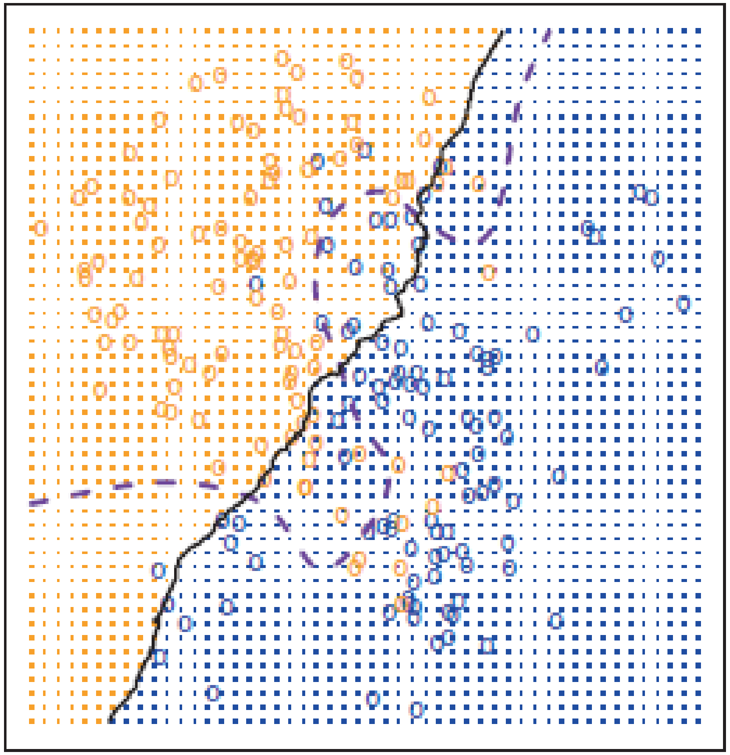
- Low bias
- High variance

- Low variance
- High bias

KNN: K=1



KNN: K=100



Source: ISLR

Bias/variance trade-off: summary

- **Bias**: the error that is introduced by modeling a real life problem by a much simpler problem

- **Variance**: says how much the model would change if using a different training set

- **Challenge**: striking a “good” trade-off between bias and variance

The problem of data availability

- In real problems we do not always have large amount of real data
  - We need lots of monitors/sensors
  - Monitors can be expensive
  - Long periods of data collection are needed
  - Labeling data is costly/time-consuming
  - Real data may produce many «outliers»
- Alternative: **synthetically generate your data**
  - Define certain set of predictors *X*
  - Guess a hypothesis *f(X)* for the behavior of your data
  - Generate synthetic data by:

$$Y=f(X) + \varepsilon$$

    - Random error ε is with zero-mean and is independent from X

Conclusion

- Some questions we want to answer at the end of the course
  - Which ML algorithm best describes our problem?
  - Which data should we consider to make predictions and/or decisions?
  - Is it worth collecting as much data as possible? Is there any irrelevant parameter we can (or should) neglect?
  - What is the performance of our learning algorithm?
  - And what is its complexity?
  - …