IMMUNE CODING INSTITUTE

Supervised Machine Learning

Module 3

Let me introduce myself...

juan-benavente

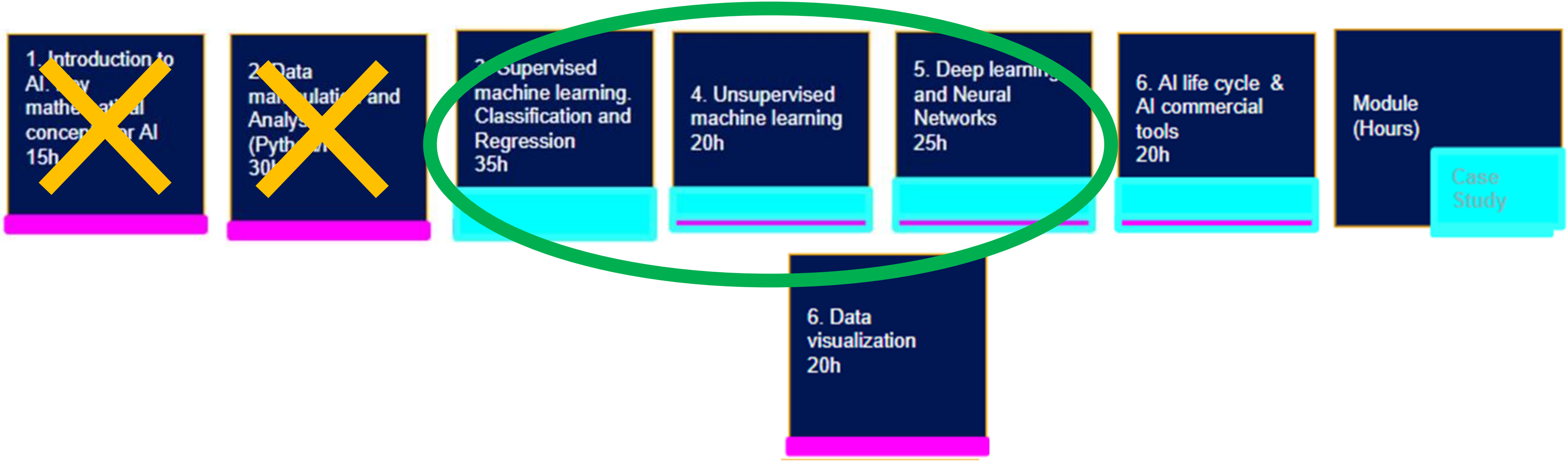
JUAN BENAVENTE

- Industrial engineer
- Computer engineer

- Smarter Cities Technical Presales (IBM)
- Internet of Things Consultant and Team Lead (IBM)
- Industry 4.0 & Blockchain Solutions Architect (CEPSA)
- Sectors: Energy, Industrial, Public

- 1st Runner Up Non-supervisory IT Employee of the year – Mubadala's IT Award
- WPC O&G Spanish Youth Award - Impact of Blockchain in O&G
- Outstanding Technical Achievement Award: Treasure Wild Ducks

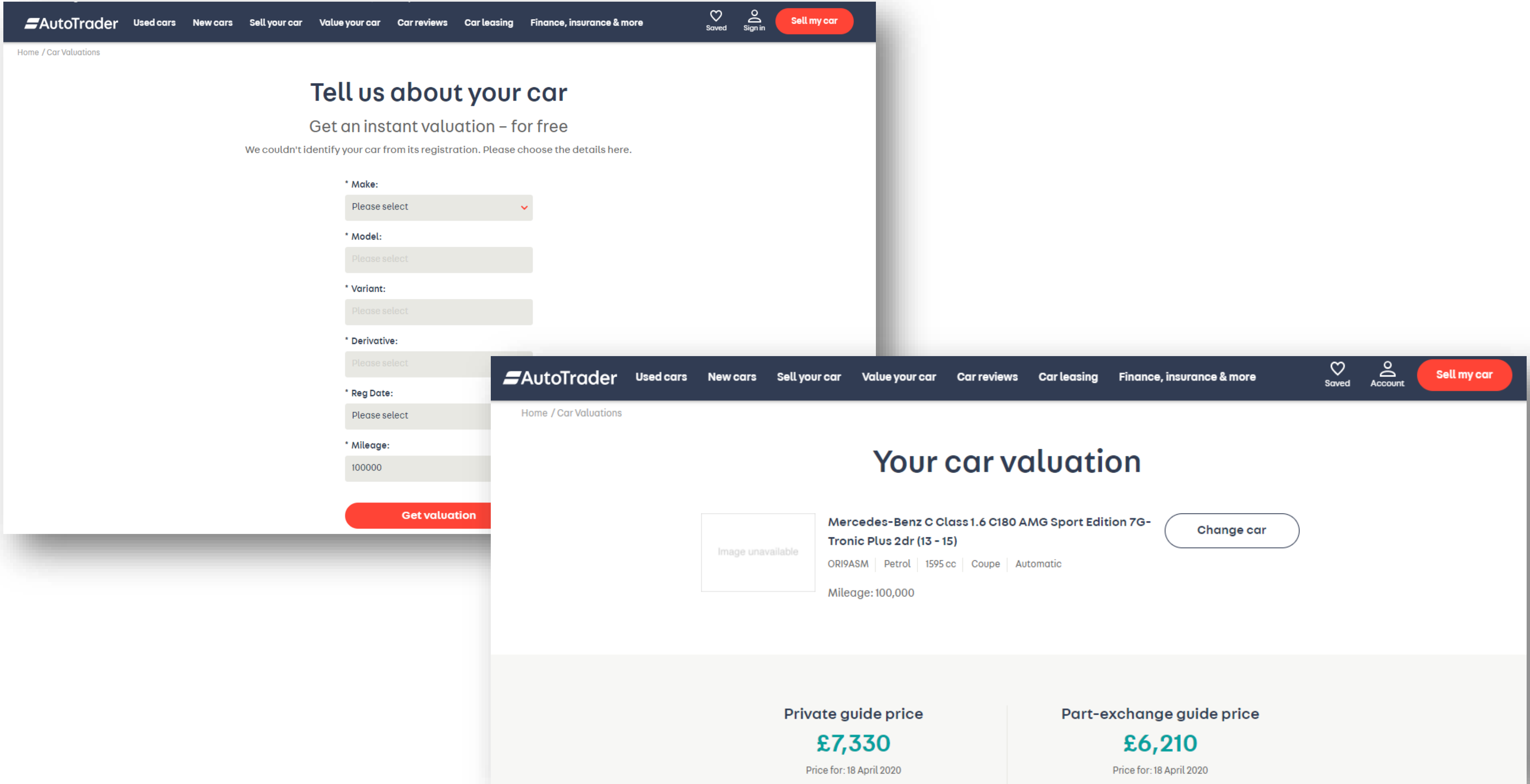
Where are we?



1. Introduction to
15h
Supervised machine learning. Classification and Regression 35h
4. Unsupervised machine learning 20h
5. Deep learning and Neural Networks 25h
6. AI life cycle & AI commercial tools 20h
Module (Hours)
Case Study
6. Data visualization 20h

Why are we here?

IMMUNE CODING INSTITUTE



IMMUNE CODING INSTITUTE

AutoTrader | Used cars | New cars | Sell your car | Value your car | Car reviews | Car leasing | Finance, insurance & more | Saved | Account | Sell my car

How do we value cars?

— What determines a car's condition?

The condition your car is in wi
There are generally 3 types o

Fair:

Bodywork needs minor rep

Minor or cosmetic damage

Minor or cosmetic damage

Interior has wear and tear,

Some interior equipment d

Tyres are safe and legal, b

Documentation and keys p

Large gaps in service histo

Current MOT

No outstanding finance

The car has never been wri

Good:

Minor or cosmetic damage

Minor or cosmetic damage to lights, glass and mirrors

All tyres are safe and legal

Documentation and keys present

Partial, documented service history

Current MOT

No outstanding finance

It is not a write-off (category

Excellent:

No damage to bodywork, whe

No damage to lights, glass an

All tyres are safe and legal

Documentation and keys pres

Full, documented service histo

Current MOT with at least six r

No outstanding finance

It is not a write-off (category

— What affects a car's value?

The guide prices we provide give
or selling it. However, many fact

Overall condition

Cars of similar ages can have vastly different levels of wear and damage.

Number of owners

Typically, cars with more previous owners are worth less, especially if they lack documentation too.

Optional extras

Features like sat navs, leather seats and panoramic sunroofs can bump up the value.

Colour

Certain colours can be more desirable, depending on current trends and fashions.

Service history

Cars are worth more with a fully-documented service history.

MOT

Buyers often pay more for a long MOT.

Modifications

Additions like body kits, spoilers and so on can increase or decrease the value.

Rarity of model

Some cars are desirable because there aren't so many of them around.

Extras included in the deal

Such as extended warranties

Why ML?

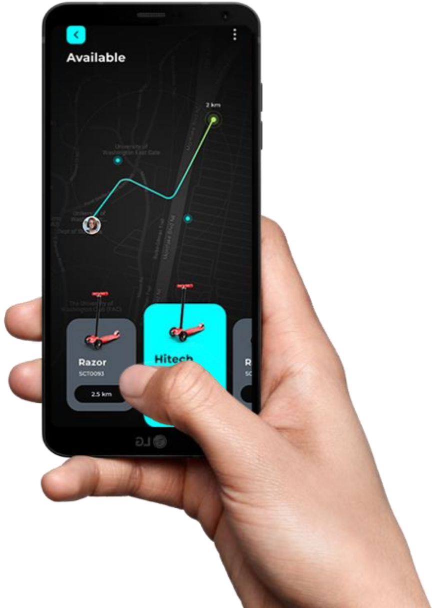
- **Problem type:** inference or prediction/estimation

- **Resolution process:** complex or unkown formulas and/or rules -> solution based on previous experience (data)

- **Other characteristics:**
  - Complexity -> many variables
  - Large amounts of data
  - Repetitive problem

Is this useful?

If you don't dress every day the same, why don't you move around differently?



IMMUNE CODING INSTITUTE

Wheelz Design Thinking

- You have just joined Wheelz startup as the **DS Team Leader**
- As a Young company, **ML has not been implemented so far**, but will be the focus during the next 3-5 years
- You organize a **Design Thinking** with your teammates in order to better understand which **data is available** at the company and **identify possible initiatives** to overcome with your team

IMMUNE
CODING INSTITUTE

Wheelz Design Thinking

- You will be working in **teams**:
    - **TEAM 1 (Juan):**
        - Daniel Rey
        - Laura Martín
        - Samuel Carballo
        - **Mauricio Asperti**
        - **Marcelo Araujo**
        - Isabel Hita
    - **TEAM 2 (Mónica):**
        - Marcos García
        - Ignacio Cifuentes
        - **María Dolores Carmena**
        - Fernando Rodríguez
        - **Ayose Sosa Guerra**
    - **TEAM 3 (Miguel):**
        - Vittoria Reale
        - **Rubén Farias**
        - José Pascual
        - Ángel Moya
        - Kay Kozaronek
        - **Miguel García**

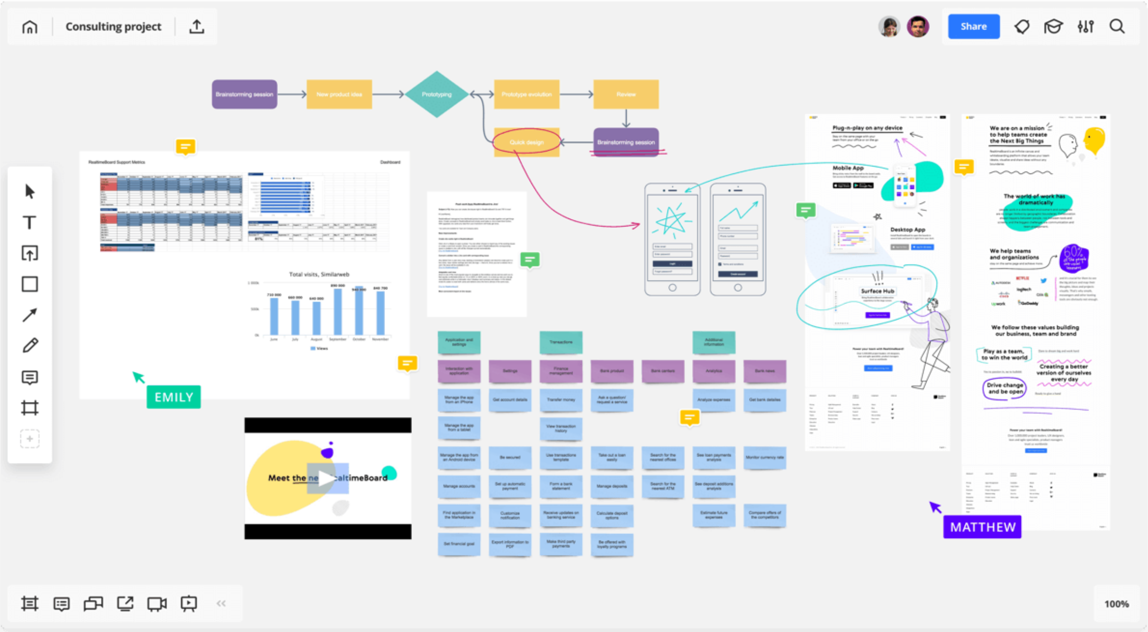
facilitator -> timing, everybody speaks, go,go,go!

presenter -> summarizes results

IMMUNE
CODING INSTITUTE

Wheelz Design Thinking

- You will be using Miro app



https://miro.com/

Wheelz Design Thinking

6 min

Discuss which sources of data are present in the company

- 2': individual creation
- 4': put together, discuss, enrich, and cluster

6 min

Discuss which initiatives we could work at

- 2': individual creation
- 4': put together, discuss, enrich, and cluster

IMMUNE
CODING INSTITUTE

Wheelz Design Thinking

ENTERPRISE DATA

- Inventario patinetes
- Ubicación tiempo real
- Estado patinetes
- Capacidad / Zona
- Demanda / zona
- Inventario clientes
- Ubicación clientes
- Trayectos realizados
- Transacciones realizadas
- Subscripciones
- Utilización servicios
- Inventario proveedores
- Estado pedidos
- Pedidos realizados
- Información empleados
- Nóminas empleados
- GIS
- Meteorología
- Tráfico

ML INITIATIVES

- Tiempo de batería restante
- Capacidad futura / zona
- Mantenimiento predictivo
- Tiempo de vida restante (i.e. batería)
- Demanda futura / zona
- Detección fraude
- Segmentación Clientes
- Capacidad trabajo
- Ruptura stock
- Predicción tráfico
- Predicción precio eléctricidad

Some basic definitions

ML is the science and art of programming computers so that they can **learn from data**.

– Aurélien Géron

ML is the field of study that gives computers the ability to **learn without being explicitly programmed**.

– Arthur Samuel, 1959

A computer program is said to learn from experience E with respect to some task T and some **performance measure P**, it its performance on T, as measured by P, improves with experience E.

– Tom Mitchell, 1997

IMMUNE
CODING INSTITUTE

Module 3 Summary

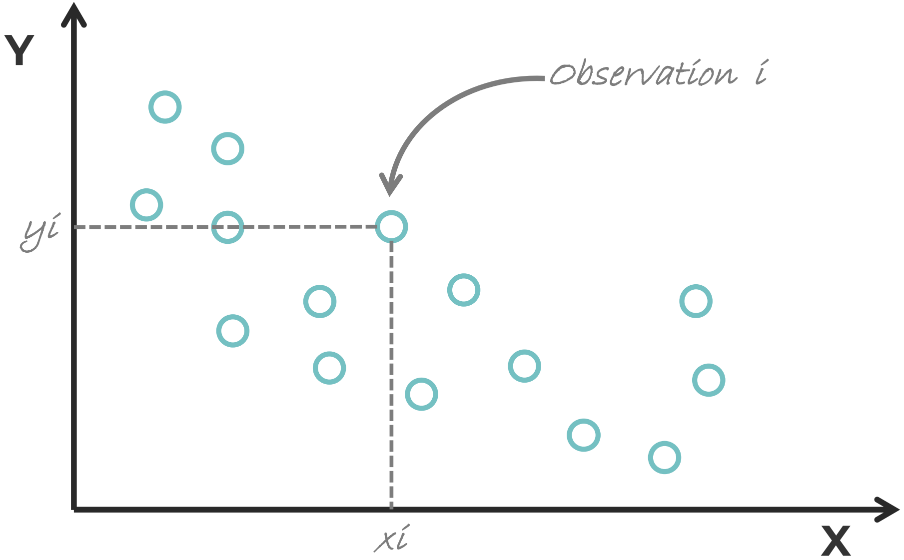
SESSION	TITLE	TEACHER
1	ML Foundations	Juan
2	Regression Introduction and Practice	Juan
3	Classification Introduction and Practice	Carlos
4	Feature Engineering and Selection for ML	Carlos
5	Advanced Supervised Models 1	Carlos
6	Advanced Supervised Models 2	Carlos
7	Hands-on Practice	Carlos

Supervised ML

Module 3.1

IMMUNE
CODING INSTITUTE

What is ML?

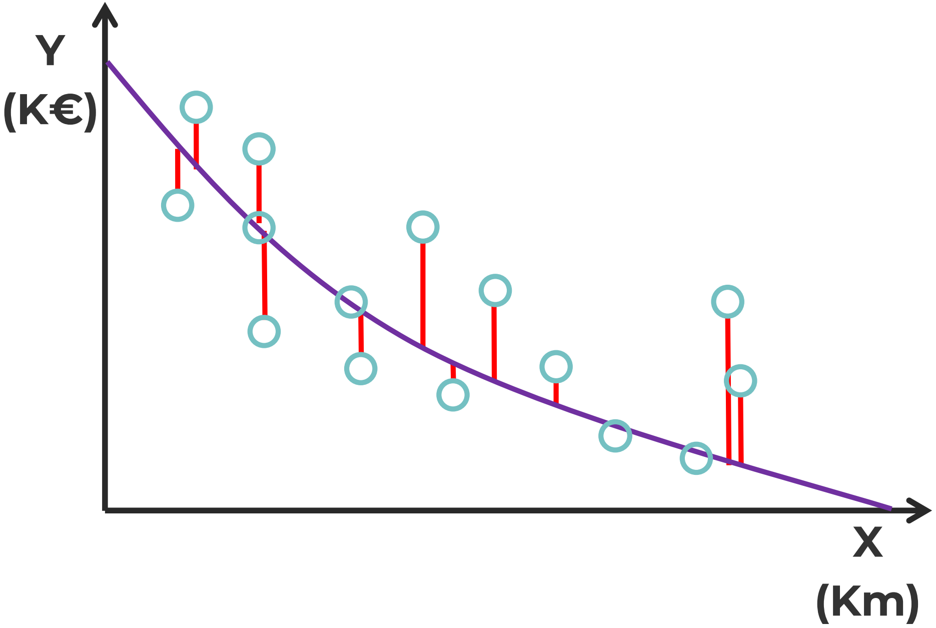


X > input variables – predictors, independent variables, features
$X = (X_1, X_2, \ldots, X_p)$

Y > output variable – response, dependent variable

Observations > $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$

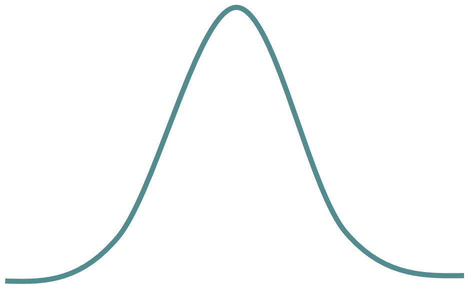
What is ML?



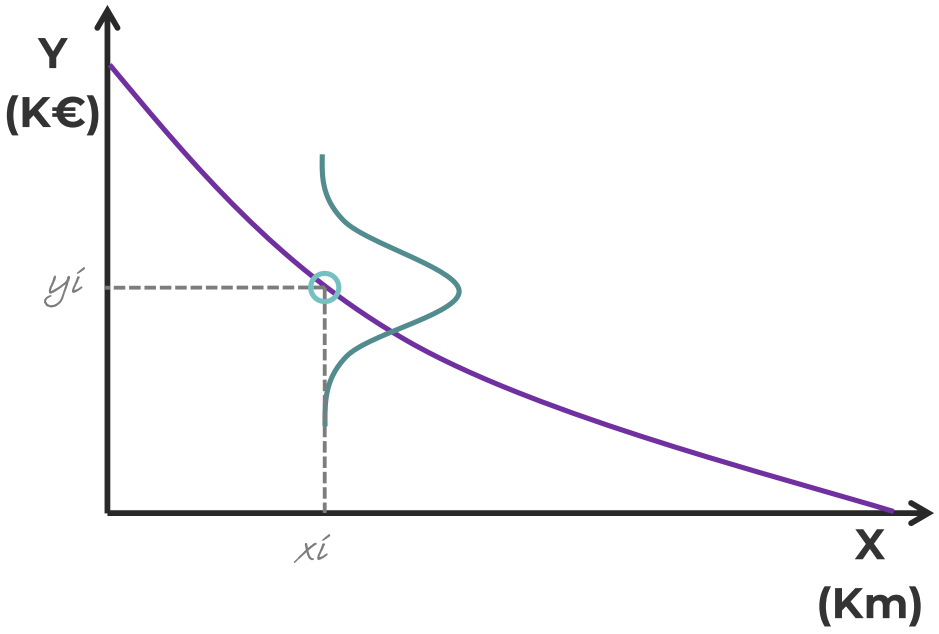
$$Y = f(X) + \epsilon$$

f(X) > fixed but unknown function that relates X and Y.

E > error term, noise, independent of X and mean 0 (a priori)



What is ML?

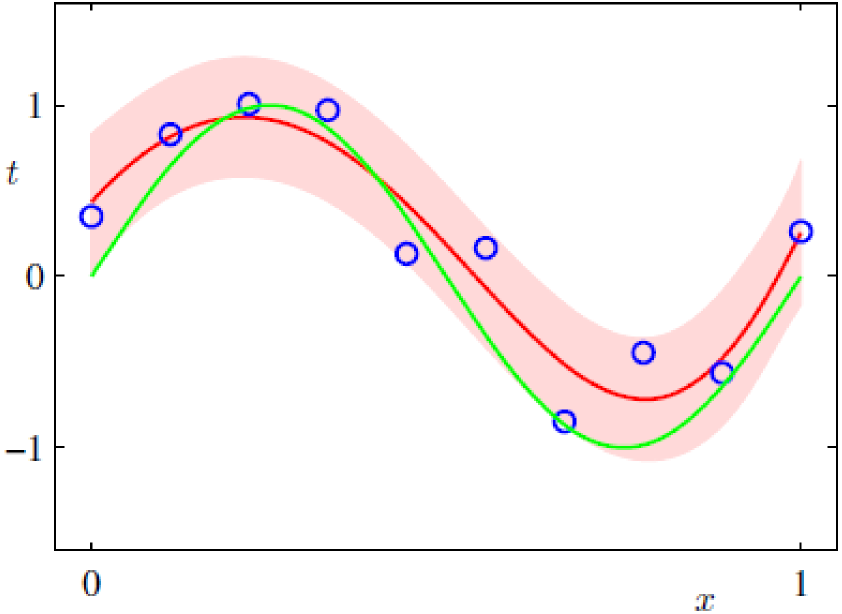


$$Y = f(X) + \epsilon$$

f(X) > fixed but unknown function that relates X and Y.

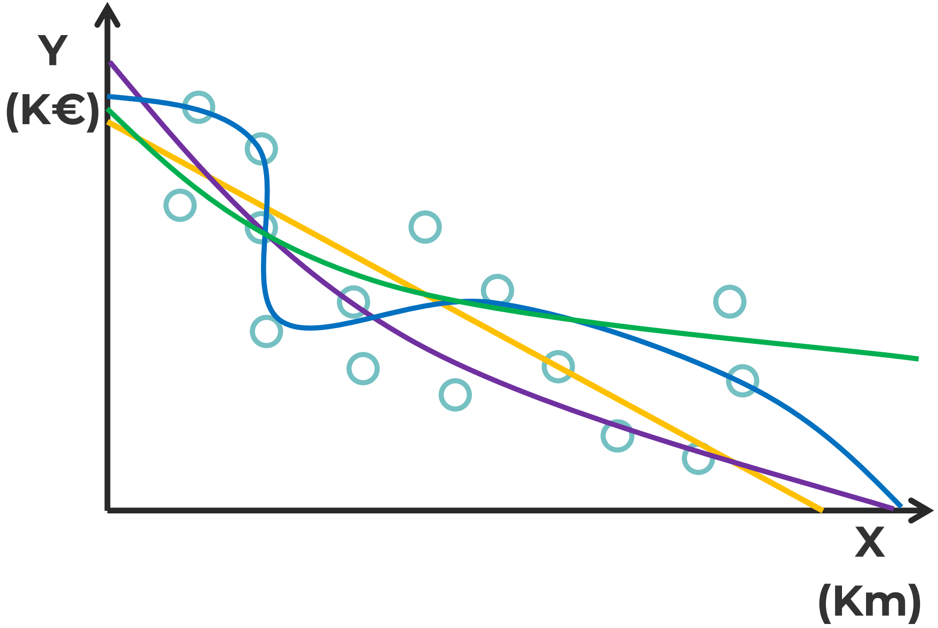
E > error term, noise independent of X and mean 0 (a priori)

What is ML?



[2]

What is ML?



$$\hat{Y} = \hat{f}(X)$$

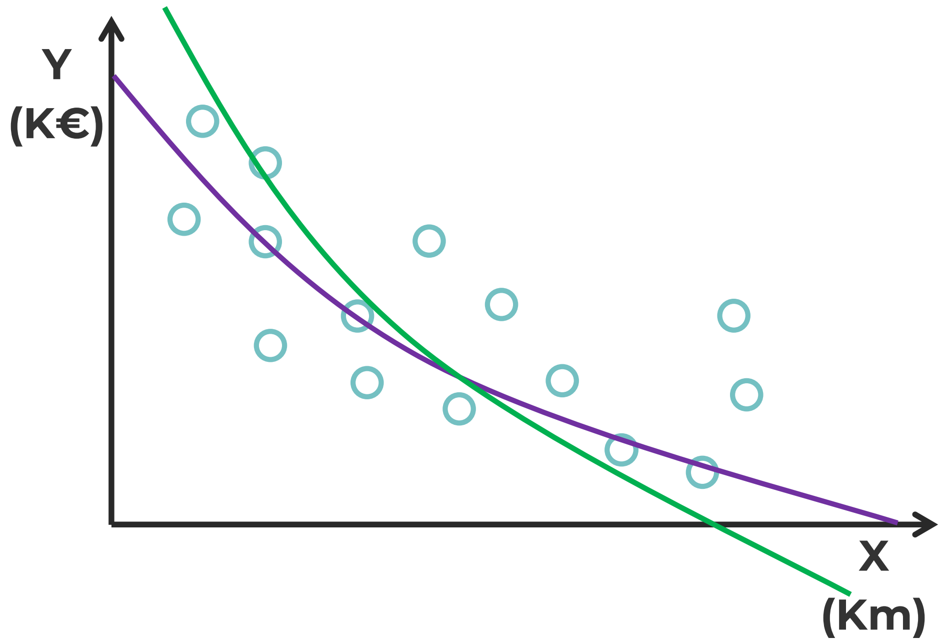
$\hat{f}$ (X) > estimate for f

Ŷ > resulting prediction for Y

IMMUNE CODING INSTITUTE

What is ML?

"Essentially, all models are wrong, but some are useful."
- GeorgeBox, 1987.



$$\hat{Y} = \hat{f}(X)$$

f̂ (X) > estimate for f

Ŷ > resulting prediction for Y

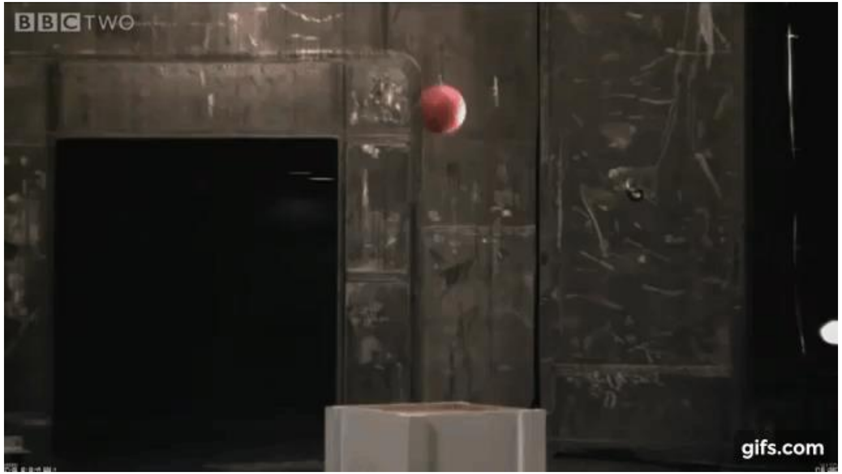
reducible error > prediction function does not suit perfectly to real function

irreducible error > unmeasured variables and unmeasurable variation

Let's model vertical movement

https://giphy.com/gifs/deadpool-dad-peter-1APafwnHNjbwU5A48H

Let’s model vertical movement



IMMUNE
CODING INSTITUTE

What is ML?

$$E\left[\left(Y-\hat{Y}\right)^2\right] = E\left[\left(f(X)+\epsilon-\hat{f}(X)\right)^2\right] = \underbrace{\left(f(X)-\hat{f}(X)\right)^2} + \underbrace{Var(\epsilon)}$$

Reducible:

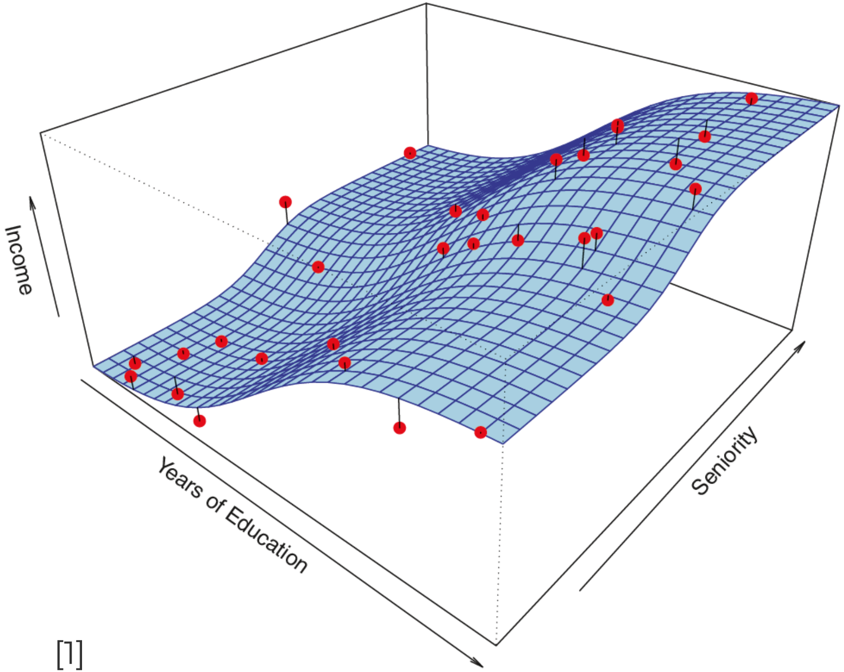
- dependent on input variables
- higher while higher is the diferente between the real function and our estimation
- we can improve our model to reduce this term

Irreducible:

- not dependent in input variables
- upper bound for our model's accuracy

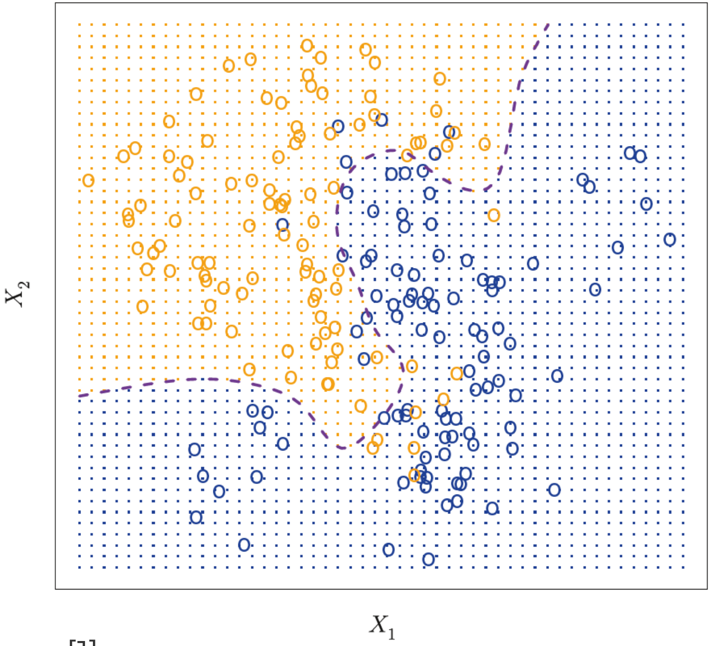
What is ML?

Different dimensionality....



[1]

Different nature....

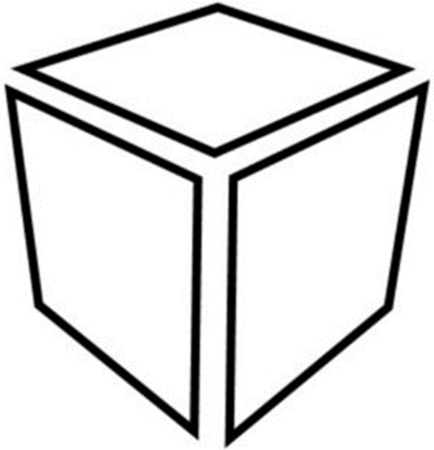


[1]

IMMUNE CODING INSTITUTE

To explain or to predict?

White-box modelling

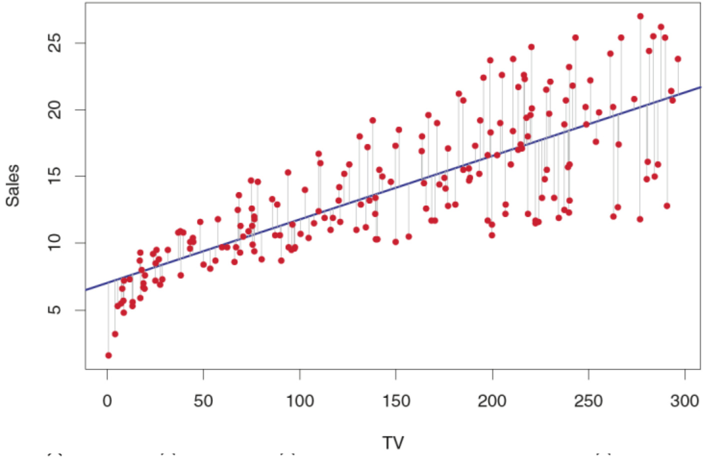


or

Black-box modelling

https://www.linkedin.com/pulse/white-box-black-choosing-machine-learning-model-your-vidyadhar-ranade

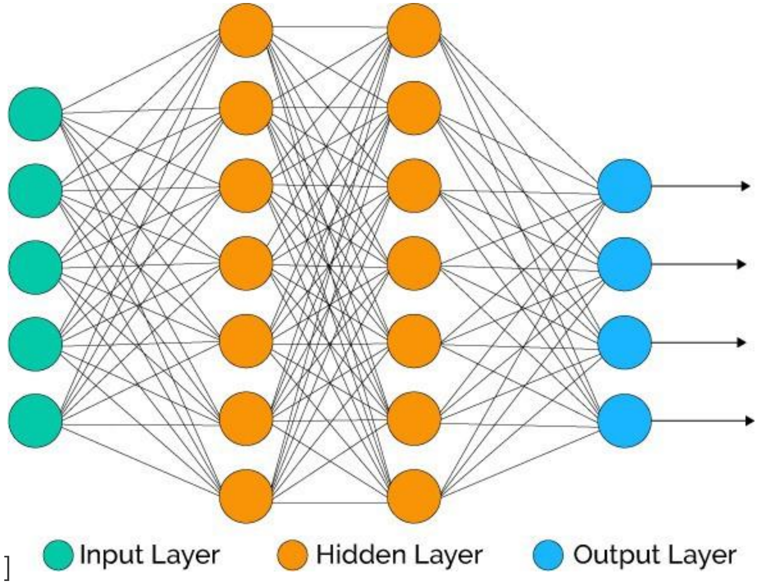
To explain or to predict?



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

[1]

[3]

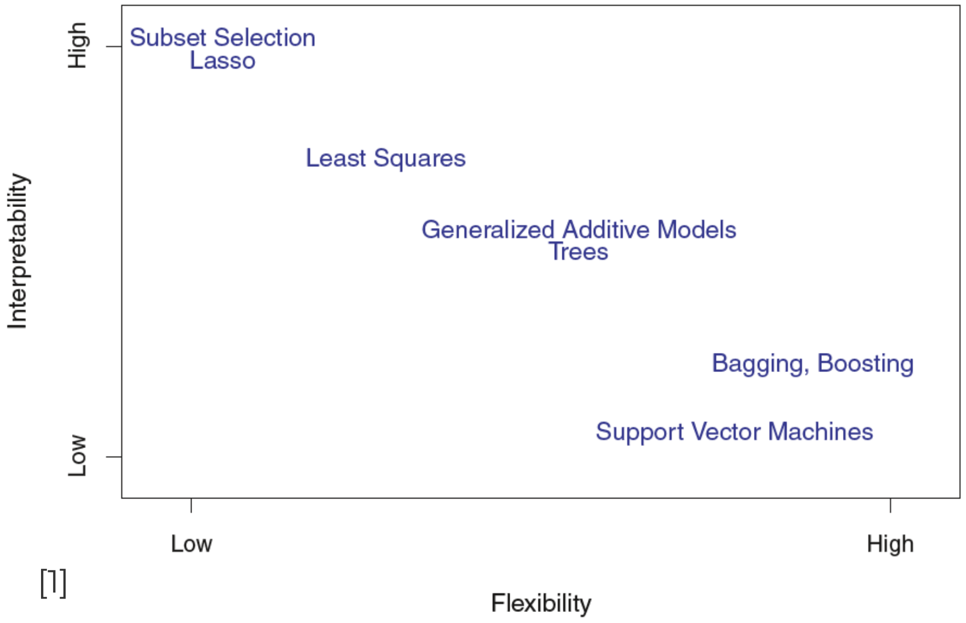


Is interpretability important? Why?

Interpretability

IMPORTANCE

- Debugging
- Informing feature engineering
- Directing future data collection
- Informing human decision-making
- **Building Trust**



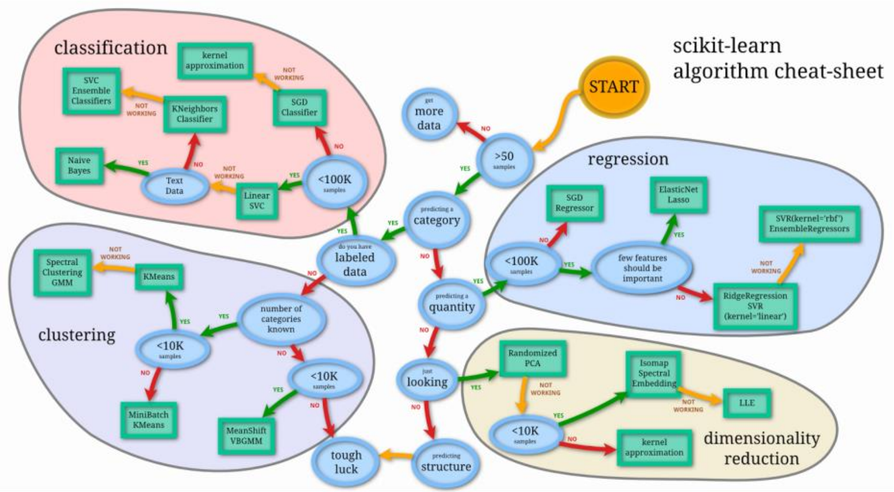
[1]

"The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks."
- Doshi-Velez and Kim, 2017

IMMUNE CODING INSTITUTE

Assessing Model Accuracy

One model to rule them all...



https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

https://www.looper.com/189208/saurons-entire-backstory-explained/

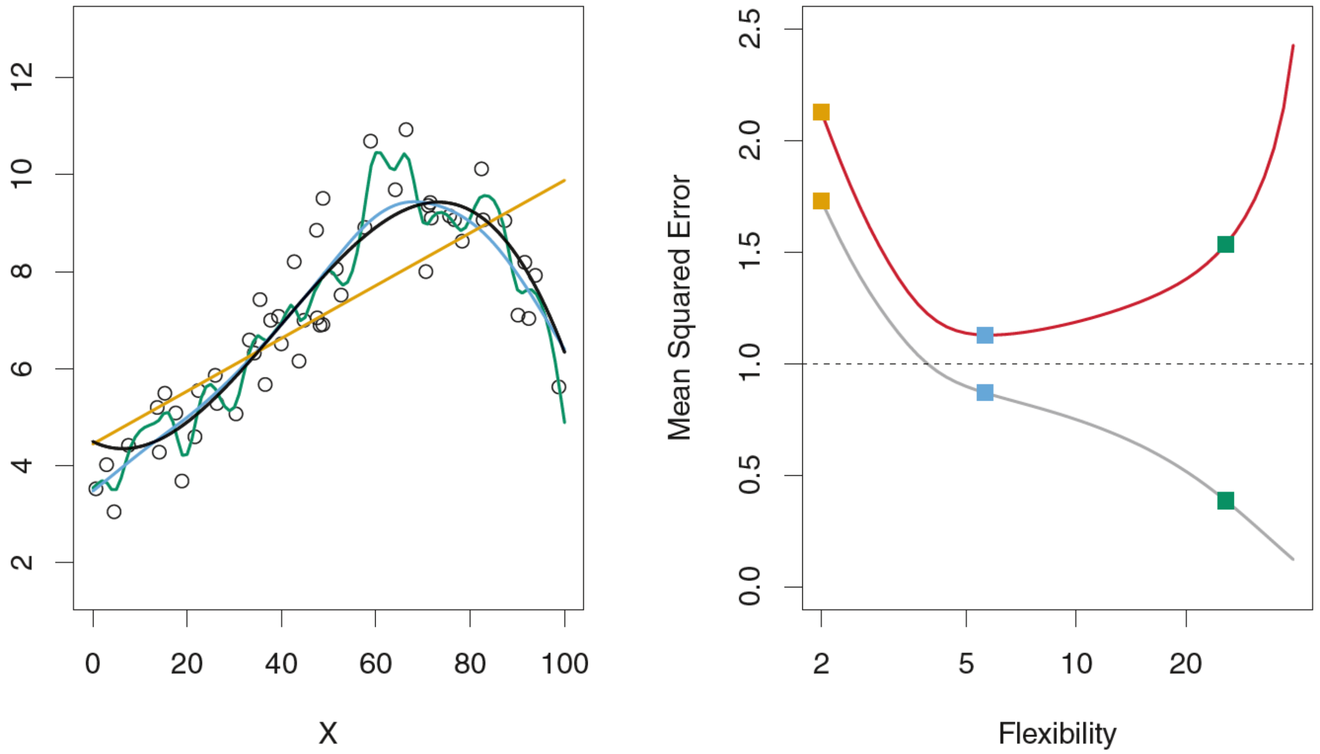
...no

Mean Squared Error

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

- Evaluate the quality of fit, model performance, model skill
- Small for well fitted model
- Training MSE vs Test MSE

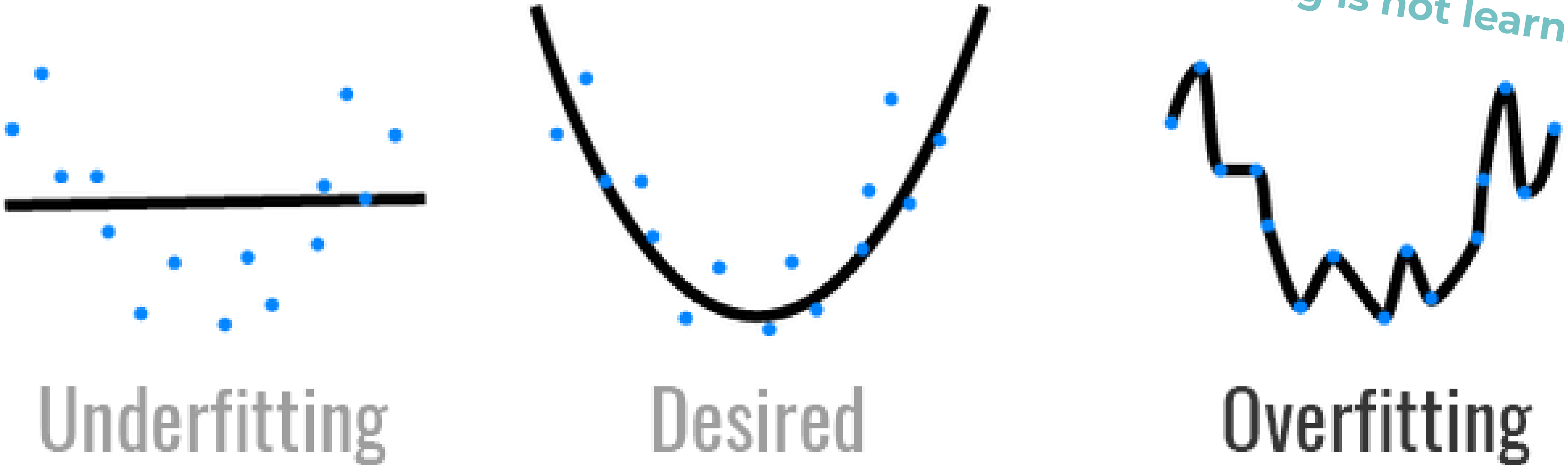
Mean Squared Error



[1]

Overfitting

Memorizing is not learning!



https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Overfitting

	Low *Training* Error	High *Training* Error
Low *Testing* Error	1	4
High *Testing* Error	3	2

https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Overfitting

	Low *Training* Error	High *Training* Error
Low *Testing* Error	The model is learning!	Probably some error in your code. Or you've created a *psychic* AI.
High *Testing* Error	OVERFITTING	The model is not learning.

https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Kahoot!

IMMUNE
CODING INSTITUTE

Bias-Variance Trade-off

$$E\left[\left(Y-\hat{Y}\right)^2\right] = E\left[\left(f(X)+\epsilon-\hat{f}(X)\right)^2\right] = \underbrace{\left(f(X)-\hat{f}(X)\right)^2}_{} + \underbrace{Var(\epsilon)}_{}$$

$$E\left[\left(Y-\hat{Y}\right)^2\right] = \left(E[\hat{f}(X)]-f(X)\right)^2 + E\left[\left(\hat{f}(X)-E[\hat{f}(X)]\right)^2\right] + Var(\epsilon)$$

$$= \underbrace{\left(Bias\left(\hat{f}(X)\right)\right)^2 + Var\left(\hat{f}(X)\right)}_{} + \underbrace{Var(\epsilon)}_{}$$

Bias-Variance Trade-off

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \left(Bias\left(\hat{f}(x_0)\right)\right)^2 + Var\left(\hat{f}(x_0)\right) + Var(\epsilon)$$

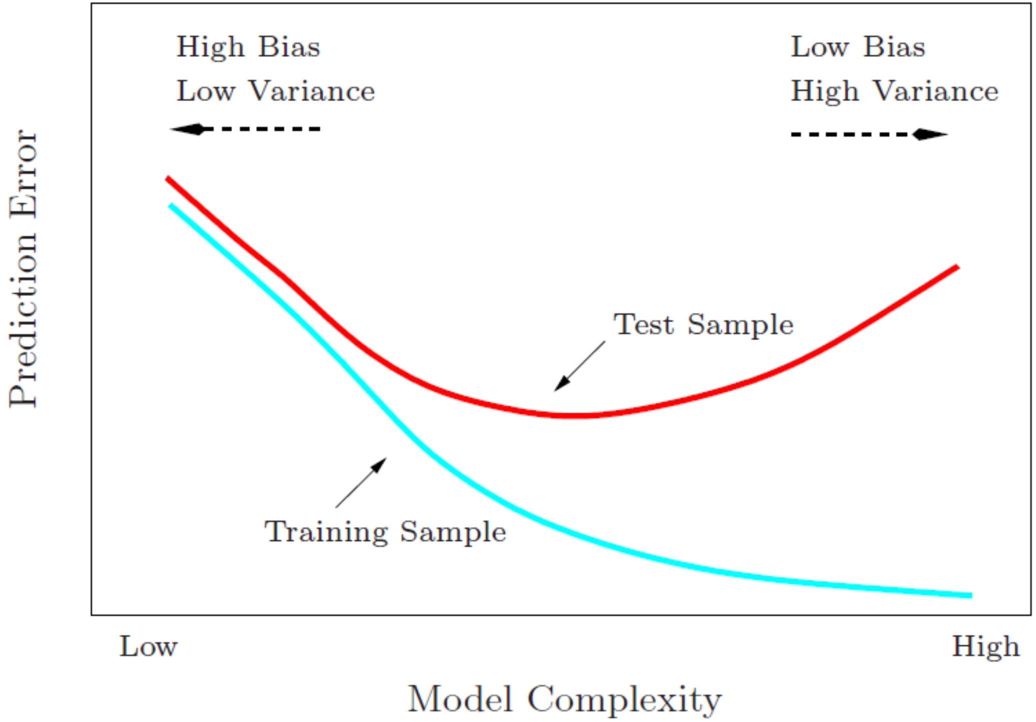
VARIANCE:

- Overcomplex model assumptions
- Amount by which $\hat{f}$ would change if we estimated it using a different training data set

BIAS:

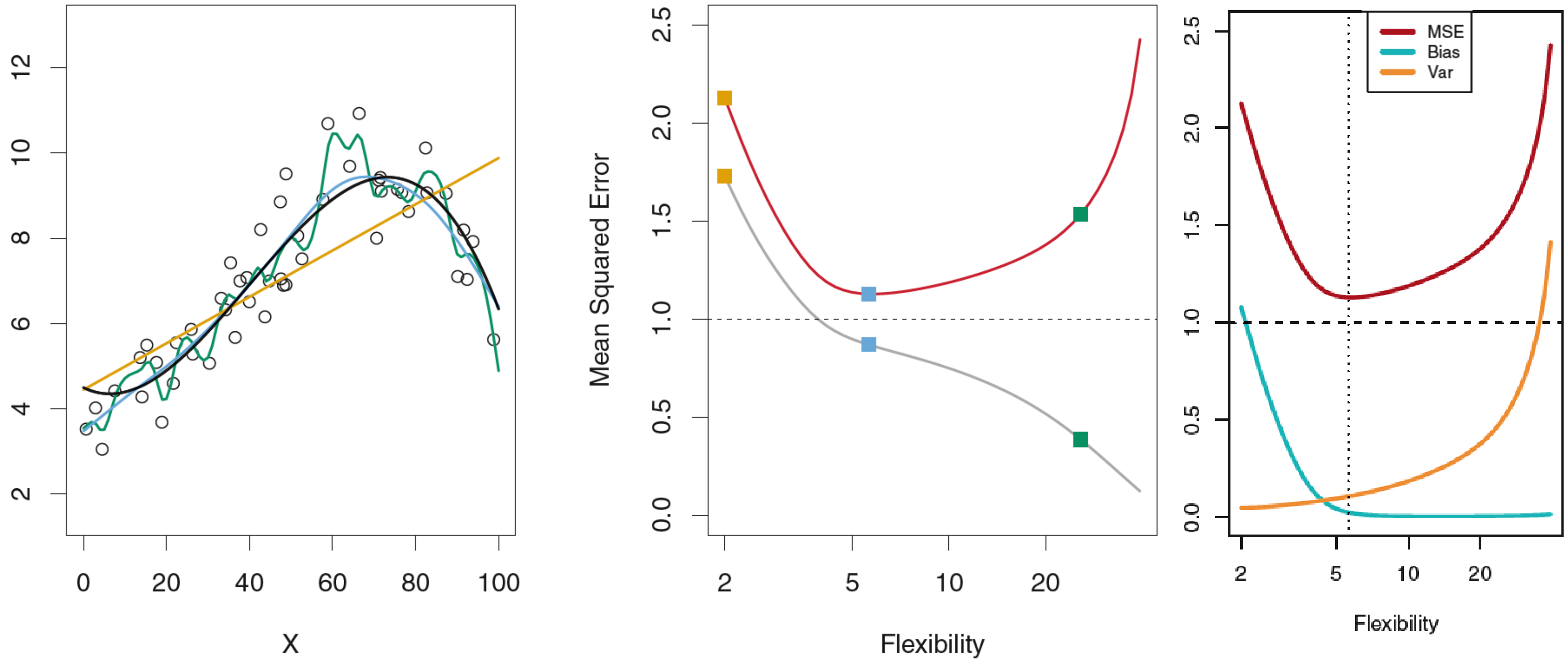
- Oversimplifying model assumptions
- Difference between the average prediction of our model and the correct value which we are trying to predict

Bias-Variance Trade-off



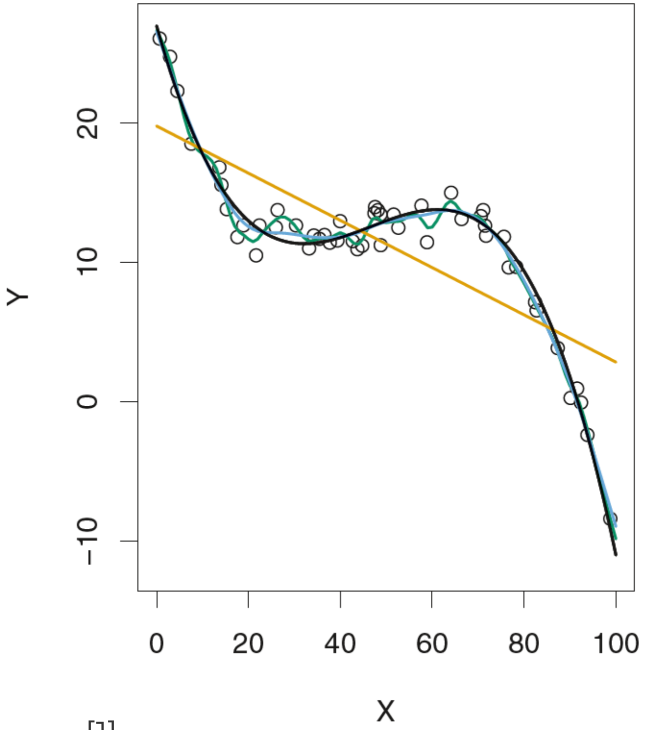
[2]

Bias-Variance Trade-off



[1]

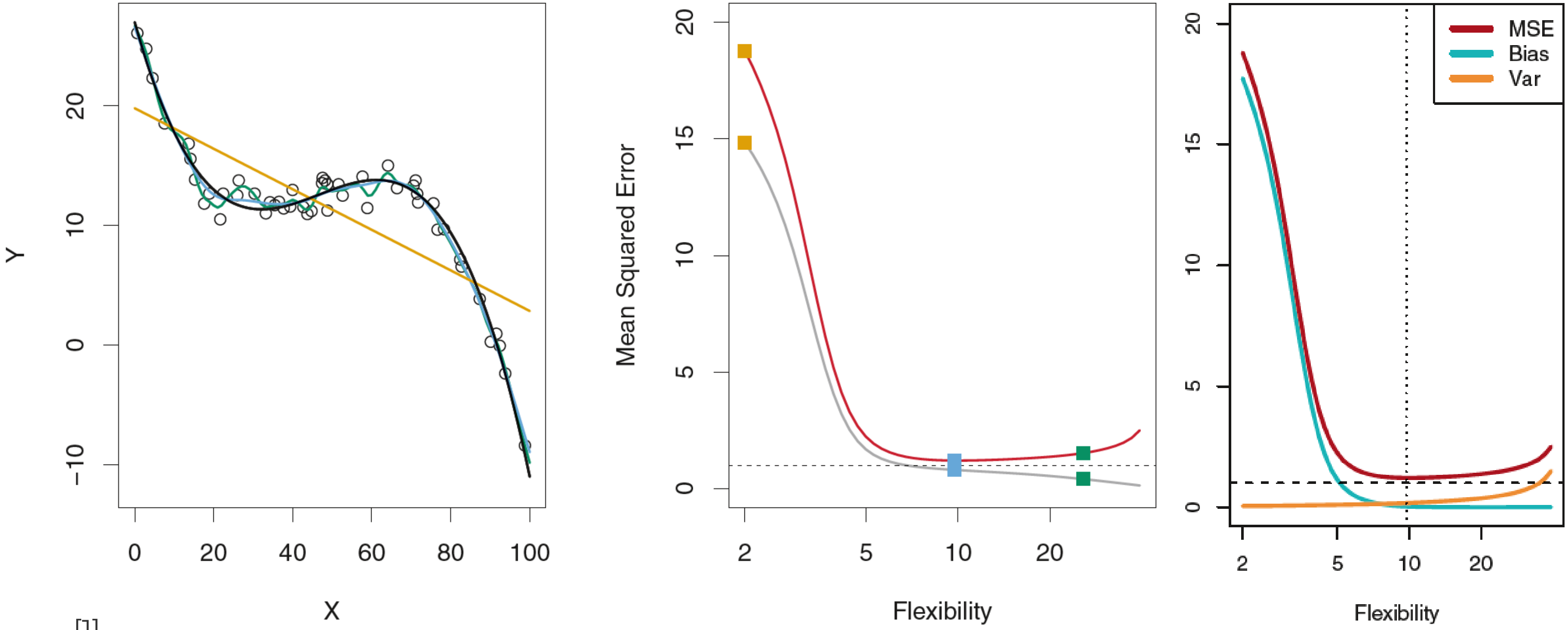
Bias-Variance Trade-off



[1]

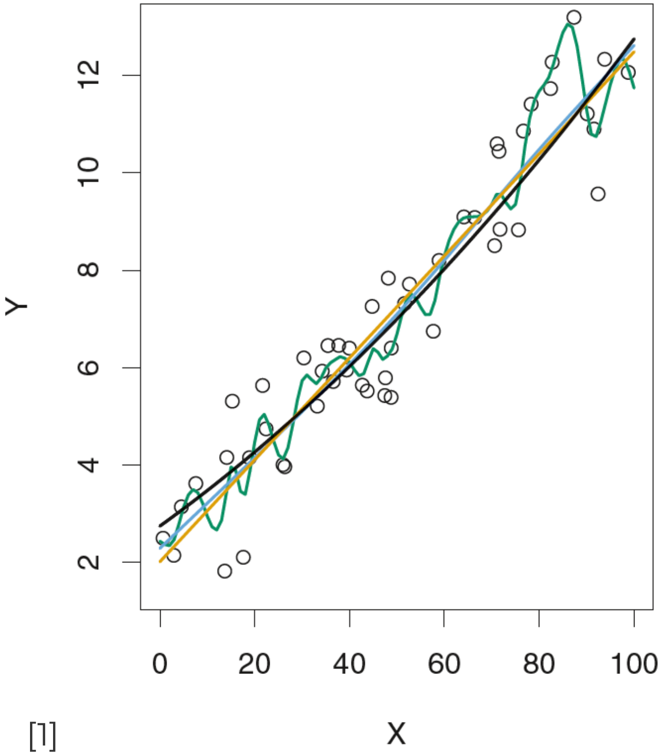
?

Bias-Variance Trade-off



[1]

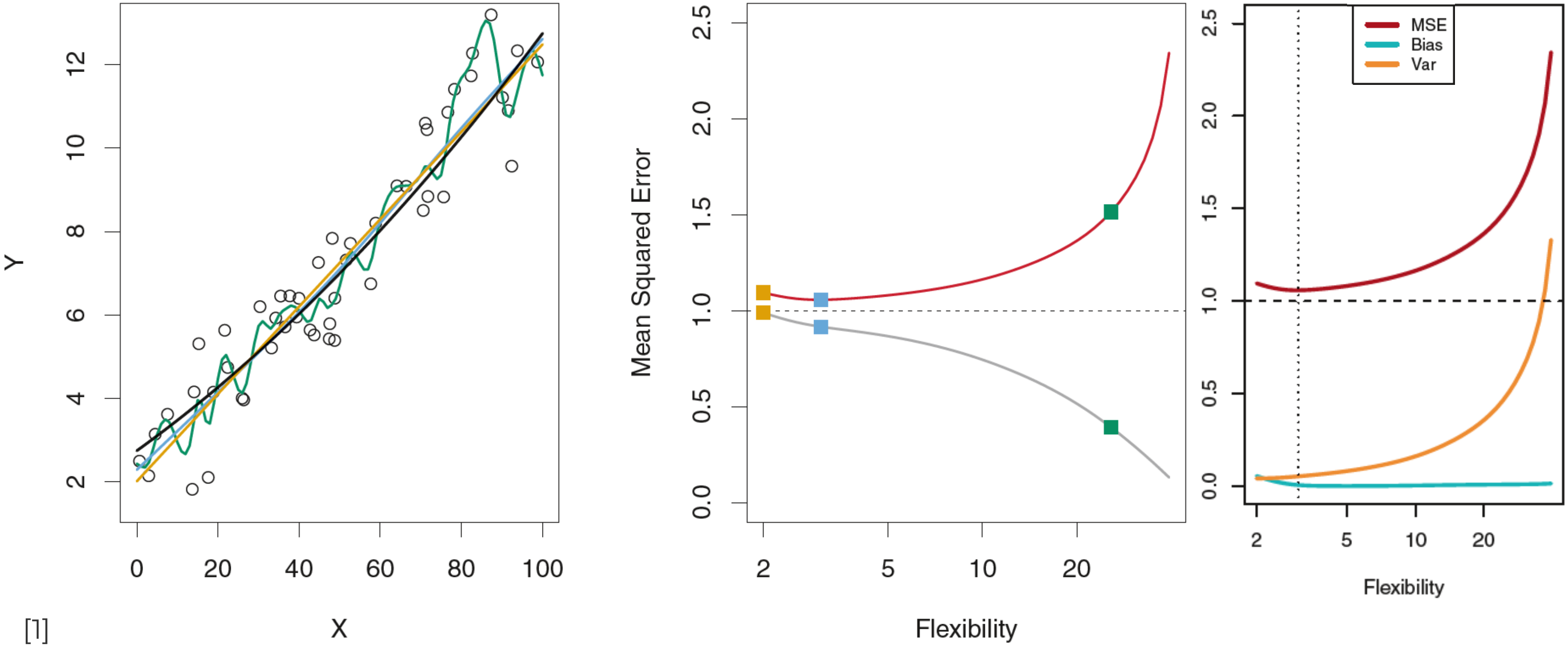
Bias-Variance Trade-off



[1]

?

Bias-Variance Trade-off



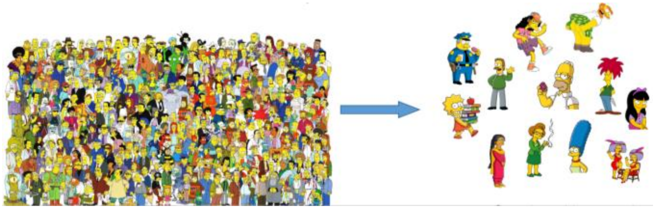
[1]

What is sampling?

IMMUNE
CODING INSTITUTE

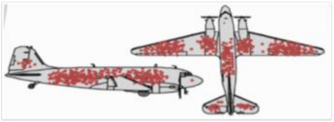
Sampling

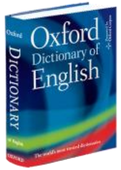
- **WHAT:** Process of selecting subsets of examples from a population with the objective of estimating properties of the population



- **WHY:** manageability, time, computational cost

- **HOW:**
    - Information redundancy has no business value
    - Removing Data vs. Removing Information
    - Consider:
        - Population
        - Sample goal and size
        - Selection criteria





IMMUNE CODING INSTITUTE

Sampling techniques

- **SIMPLE RANDOM SAMPLING:** select a subset of a population in which each member of the subset has an **equal probability of being chosen**. With or without replacement

- **STRATIFIED SAMPLING:** **objects are drawn from each group** even though the groups are of different sizes. Either same amount form each group or proportional amounts

Resampling methods

WHAT:

- Repeatedly drawing samples from a training set and refitting a model of interest on each sample
- Might be computationally expensive

WHY:

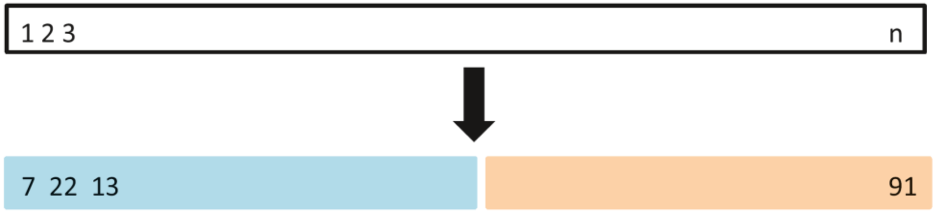
- Extract as much information as possible from a finite size dataset, information that could not be available from fitting the model only once
- Estimate the skill of a machine learning model on unseen data
- Model selection vs. model assessment

HOW:

- Cross-validation
- Bootstrap

Validation set

- Split observations randomly into two sets:
    - Training set
    - Test/Validation/hold-out set

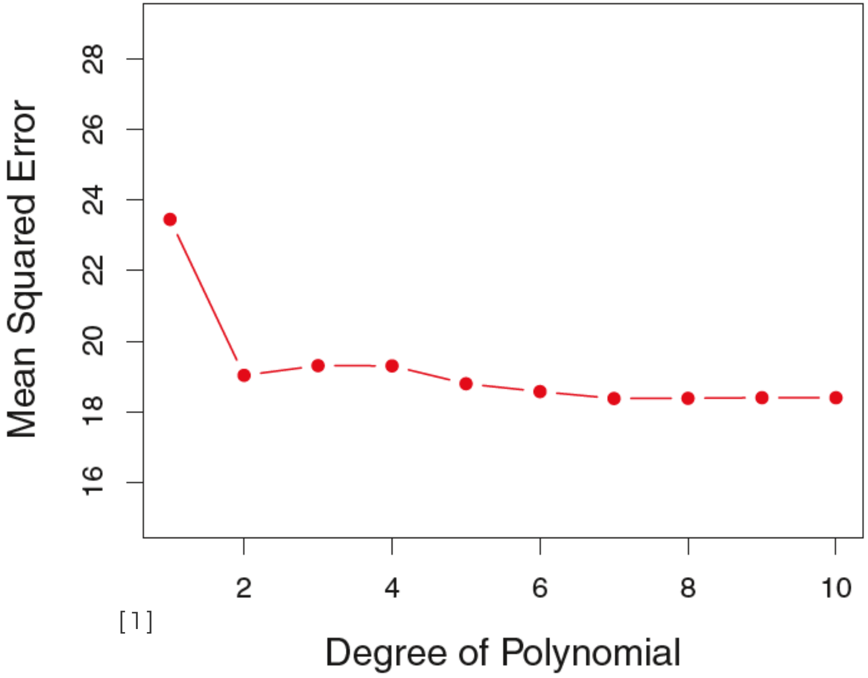


[1]

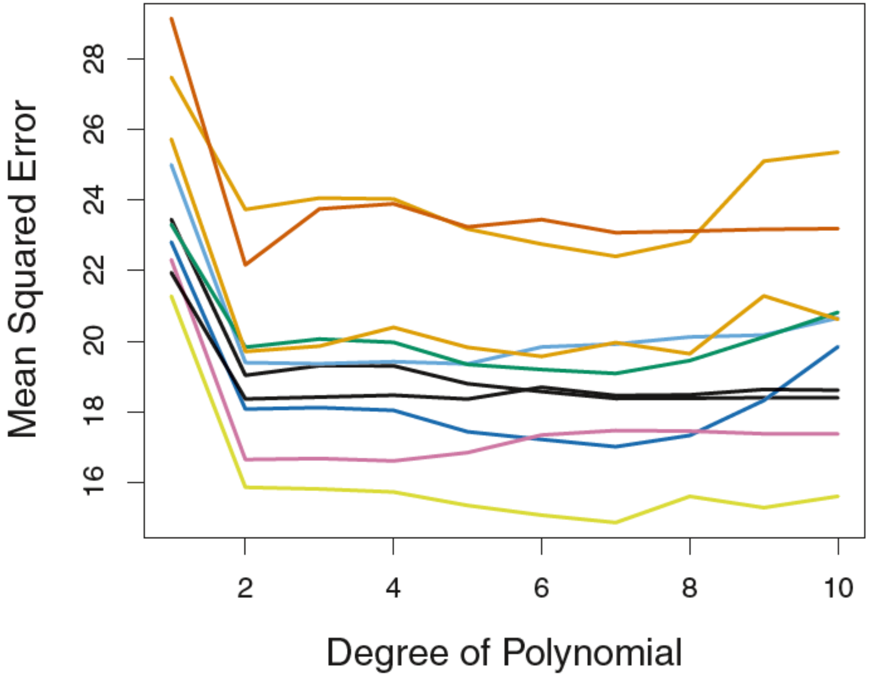
- Repeat the process several times (if necessary)
- Drawbacks:
    - Validation estimate of the test error can be **highly variable** depending on which observations are included
    - Only a subset of observations used to fit model -> **overestimate error**

IMMUNE
CODING INSTITUTE

Validation set



[1]

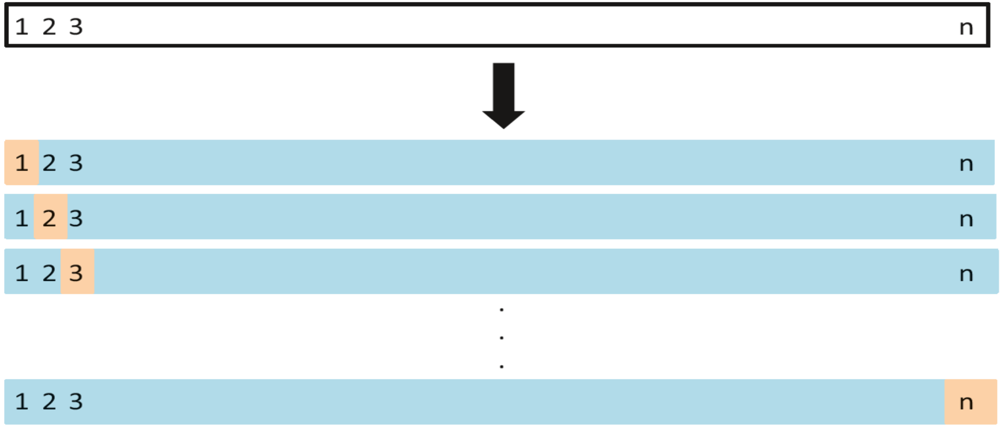


Leave-one-out cross validation (LOOCV)

Split observations into two sets:

- Training set: n-1 observations **>** $\{(x_2, y_2), \dots, (x_n, y_n)\}$
- Test/Validation/hold-out set: 1 observation **>** (x_1, y_1)

$$MSE_1 = (y_1 - \hat{y}_1)^2$$



[1]

Repeat the process n times:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Leave-one-out cross validation (LOOCV)

ADVANTAGES:

- **Less bias** >
  - we use almost every data we have in every fit
  - test error less overestimated
- **No randomness** > executing it multiple times, leads to same results
- **General method** > can be used with any kind of predictive modeling

DRAWBACKS:

Expensive to implement, since the model has to be fit n times. This can be very time consuming if *n* is large, and if each individual model is slow to fit

IMMUNE
CODING INSTITUTE

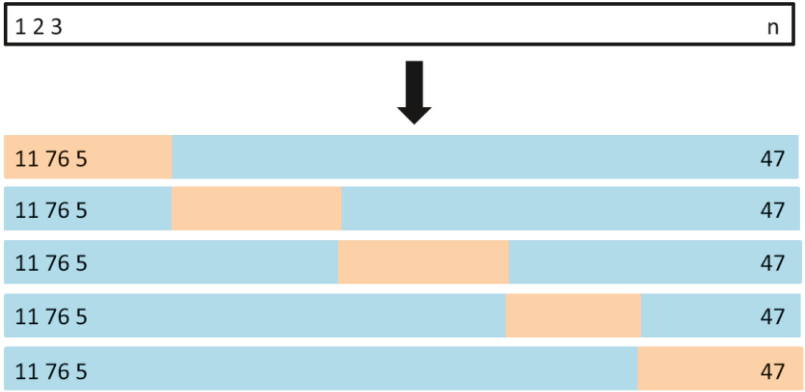
K-Fold Cross Validation

Split observations into k sets (*folds*):

- Observations are randomly allocated to each fold
- Each fold is of the same size (or similar)
- First fold is used for validation and the rest for training

Repeat the process n times using different folds for validation:

$$CV_{(n)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$



[1]

K-Fold Cross Validation

k=5 or k=10

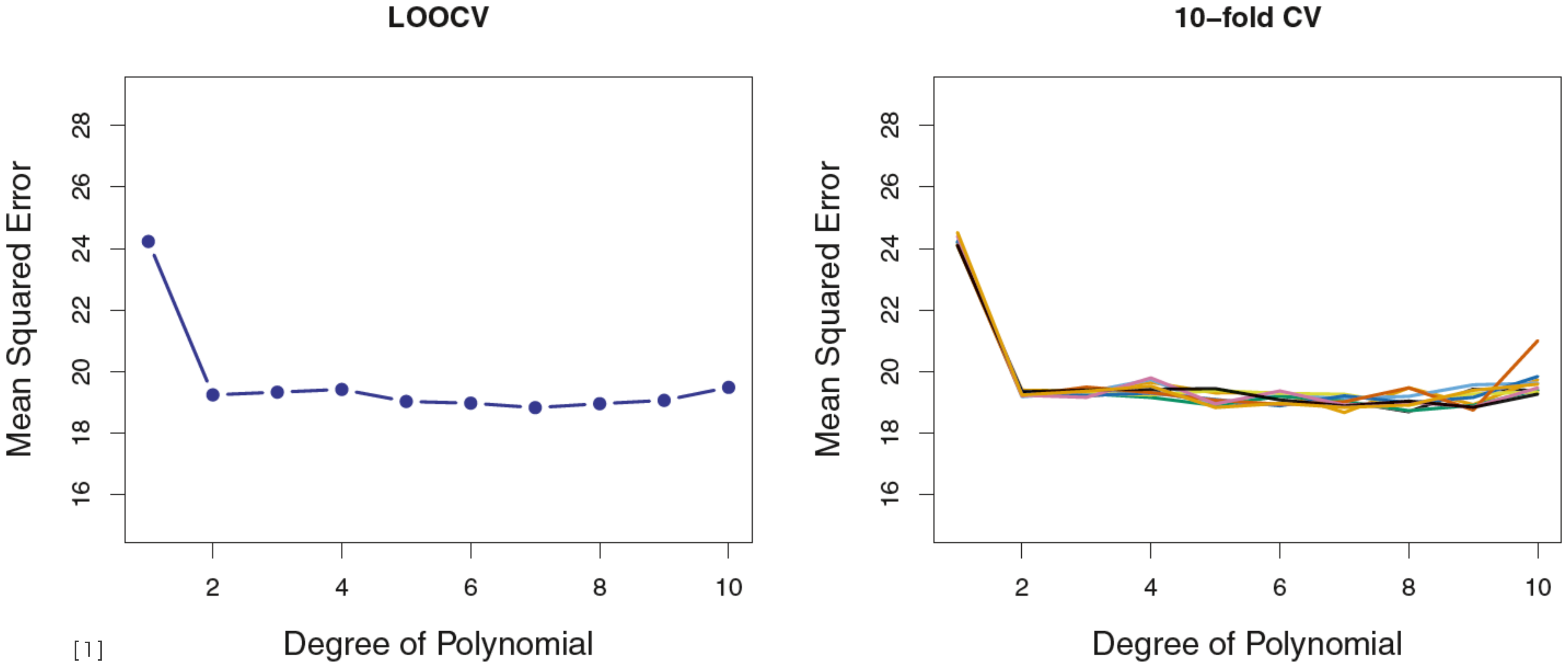
ADVANTAGES:

- Computationally cheaper than LOOCV
- Some variability **>** much lower than validation set approach
- Often gives better results than LOOCV due to bias-variance trade-off

BIAS-VARIACE TRADE-OFF (k<n)

- **BIAS:**
    - LOOCV is almost unbiased as it takes into account n-1 observations
    - K-Fold CV has intermediate level of bias as it considers less observations
- **VARIANCE:**
    - LOOCV has the highest variance as model is fitted on almost identical sets of observations highly correlated with each other > higher variance
    - K-Fold CV models are less correlated > lower variance

K-Fold Cross Validation



[1]

Let's try it ourselves...

IMMUNE CODING INSTITUTE

References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2017.

[2] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical: Data Mining, Inference and Prediction. Springer, 2009.

[3] https://towardsdatascience.com/

IMMUNE

CODING INSTITUTE