Primeros pasos en Machine Learning y regresión lineal

En breve comenzamos...

NEOLAND

Primeros pasos en Machine Learning y regresión lineal

NEOLAND

PONENTE

Josep Miquel Porcar

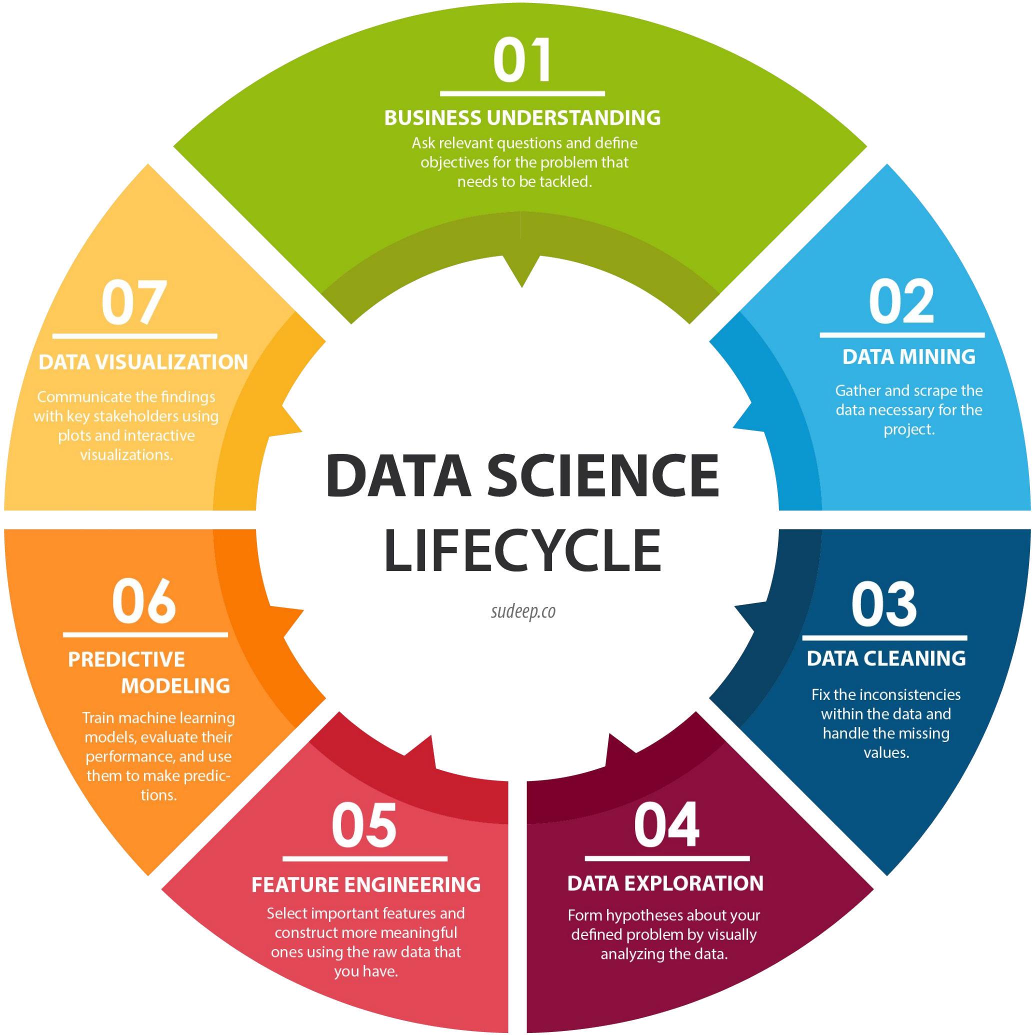
Data Scientist, matemático y estadístico con varios años de experiencia en proyectos de investigación de Data Science.

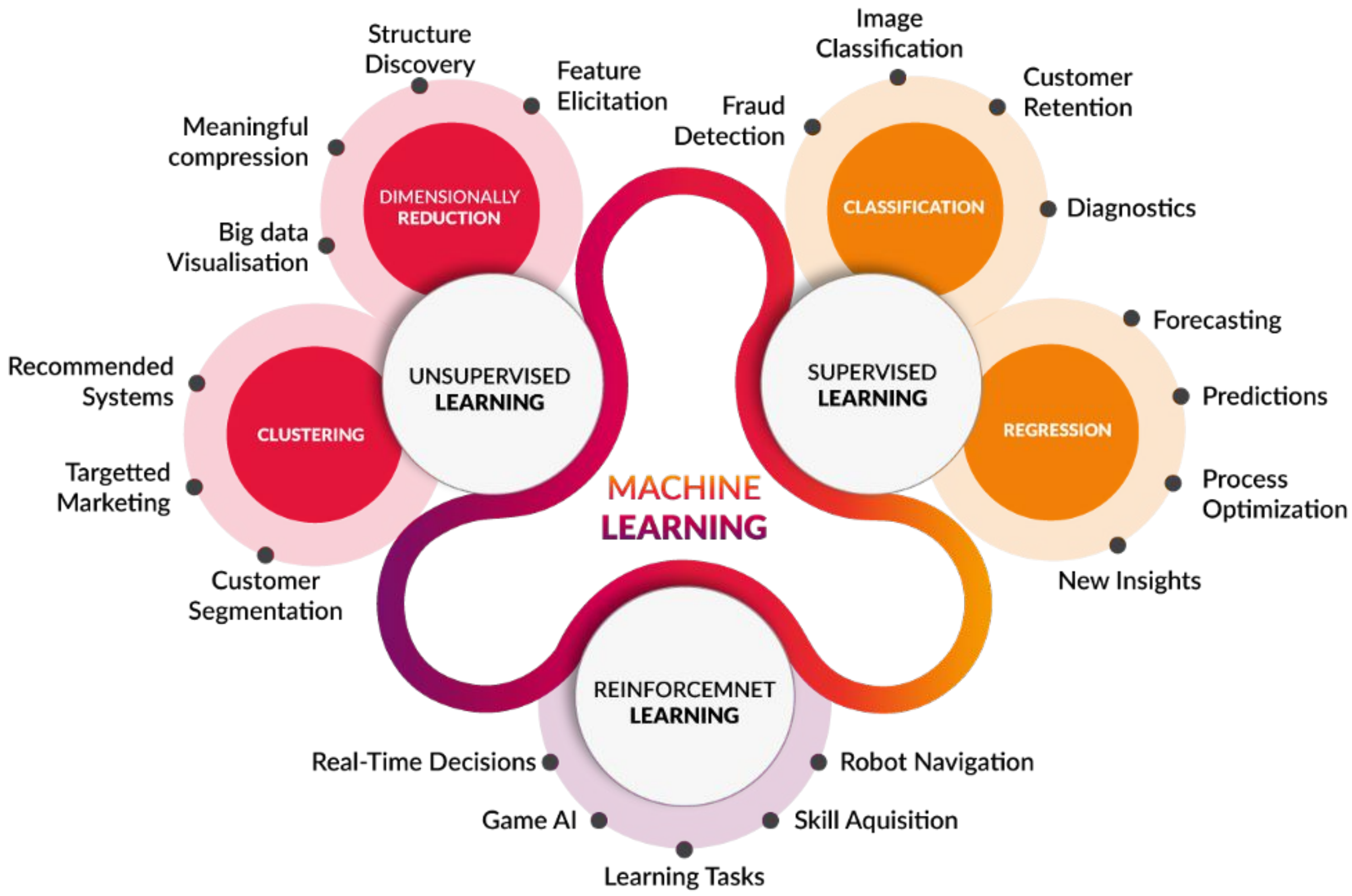
Actualmente es **Head Teacher del Data Science Bootcamp de Barcelona en NEOLAND**.

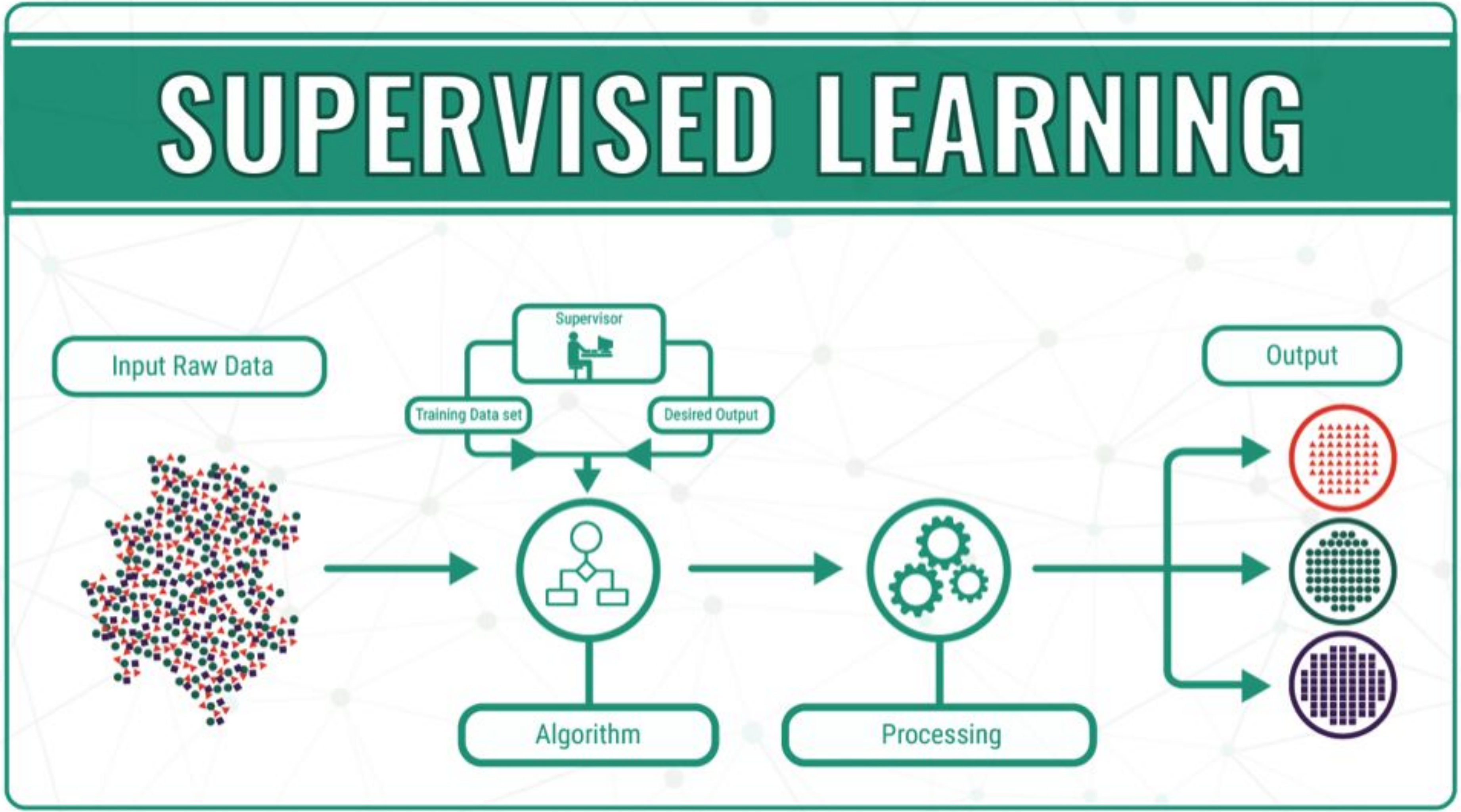
DATA SCIENCE

¿Qué veremos?

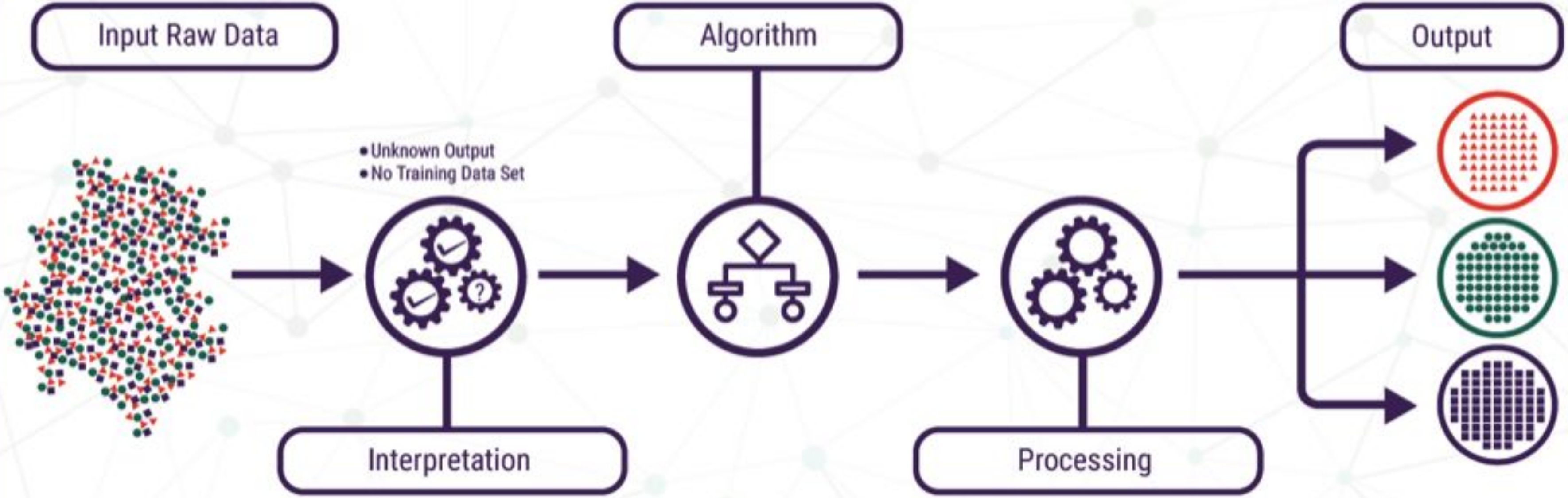
1. Statistical learning
2. Types of algorithms
3. Supervised learning
4. Linear Regression
5. Features types
6. Metrics
7. Polynomial regression
8. Validation


DATA SCIENCE LIFECYCLE
sudeep.co
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predic-tions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

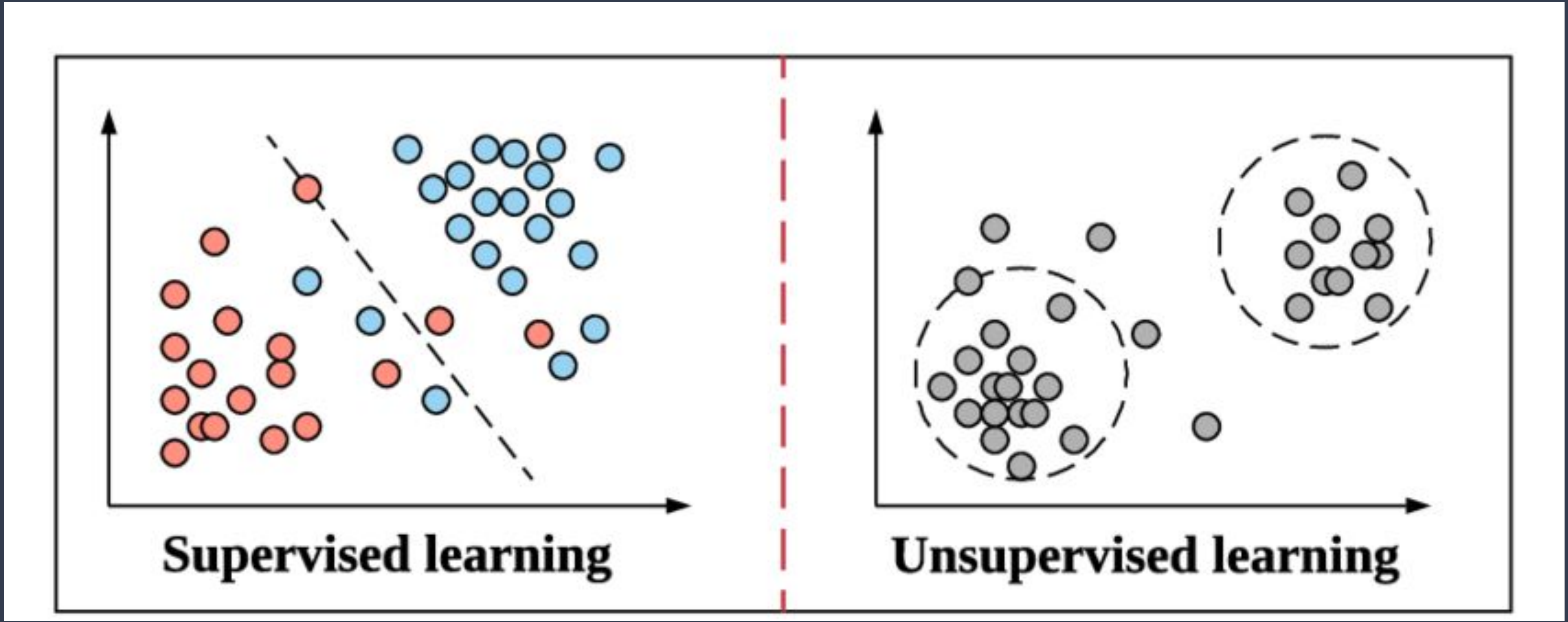

MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

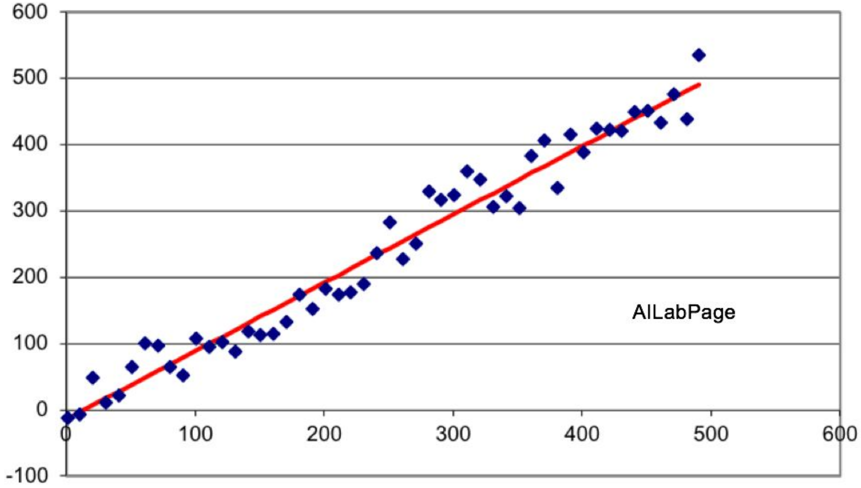

SUPERVISED LEARNING
Input Raw Data
Supervisor
Training Data set
Desired Output
Algorithm
Processing
Output

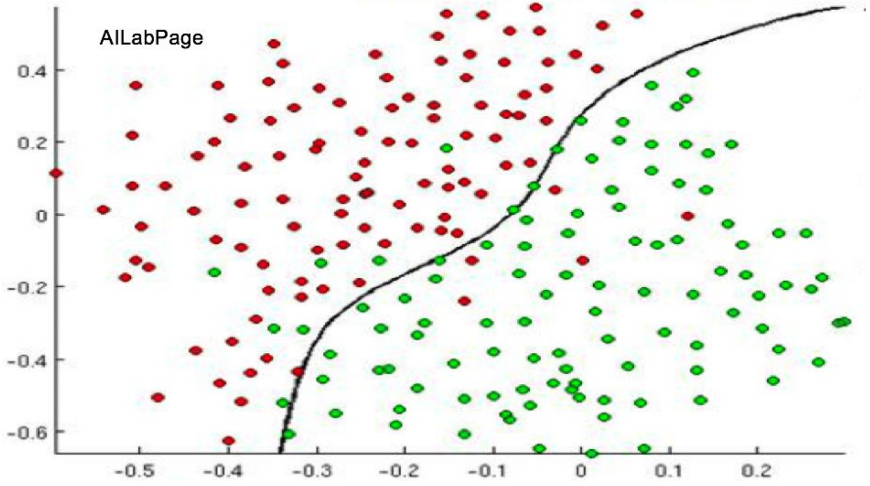
UNSUPERVISED LEARNING



Types of algorithms







Regression

The system attempts to predict a value for an input based on past data.

Example – 1. Temperature for tomorrow

Classification

In classification, predictions are made by classifying them into different categories.

Example – 1. Type of cancer   2. Cancer Y/N

Supervised learning

Supervised learning means learning from data:

- We have a quantitative outcome (regression) or categorical outcome (classification)
- We want to predict the *outcome* based on a set of *features* (supervised)
- We have a *training set*
- We build a prediction model for new unseen objects. The objective is to predict accurately

Vocabulary

Outcome, target, response: Usually denoted by Y

Features, columns, variables: Usually denoted by X (X is a vector of k features)

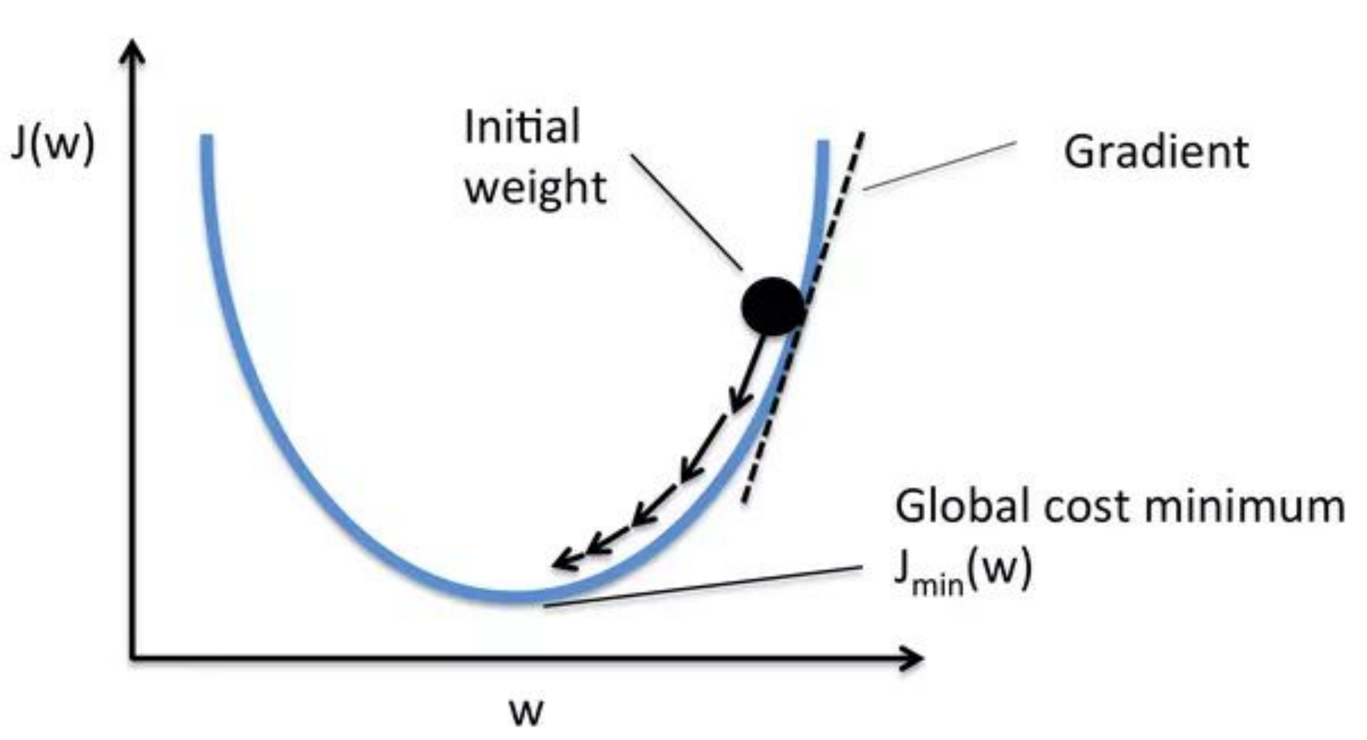
Training set: $(x_1, y_1), \ldots, (x_n, y_n)$

<u>Objective</u>: get a good prediction of Y called $\hat{Y} = f(X)$.

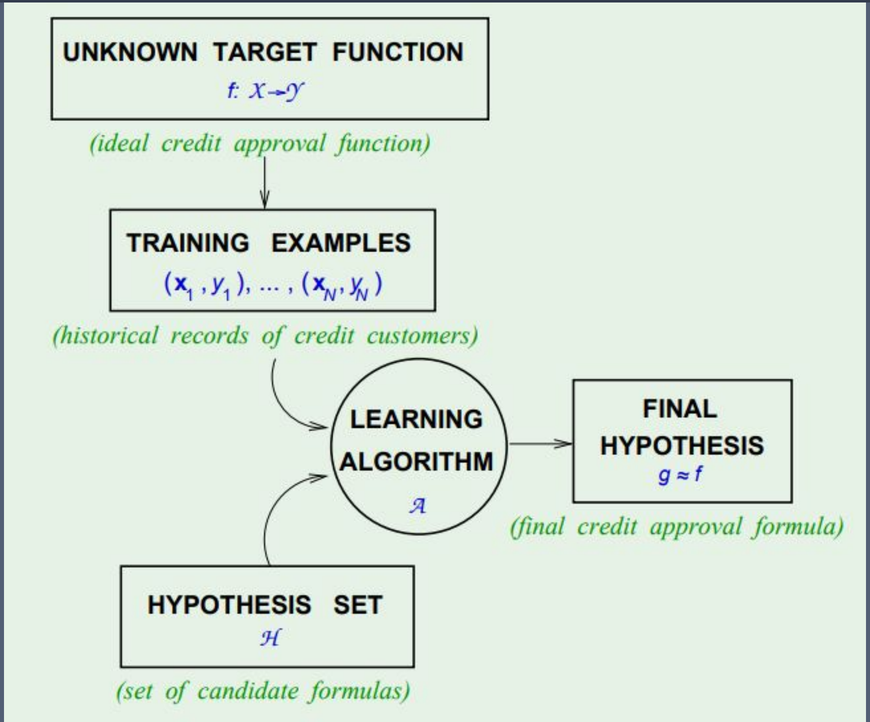
<u>LOSS FUNCTION</u> for penalizing errors (**cost function**)

Squared loss error $(Y - f(X))^2$

Loss function



Learning process



Linear Regression

A linear regression model assumes that Y is linear in the inputs X:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

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

Basic assumptions on errors: $\varepsilon \sim N(0, \sigma^2)$

- Independent
- Mean zero
- Constant variance

Predicting new data

Given a new set of features (X_nuevo), we can predict the outcome as:

$$\hat{Y}_{nuevo} = \hat{\beta} X_{nuevo} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,nuevo} + \cdots + \hat{\beta}_p X_{p,nuevo}$$

Features types in Linear regression

Quantitative - Continuous variables:

- Transformations: log, square root...
- Expansions: ^2, ^3, ...
- Interactions: X3 = X1*X2

Qualitative - Categorical variables:

- Dummy coding of the levels. 1 variable with K categories -> K dummy variables

Metrics

R2:

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma\left(y - \bar{y}\right)^2}$$

RMSE:

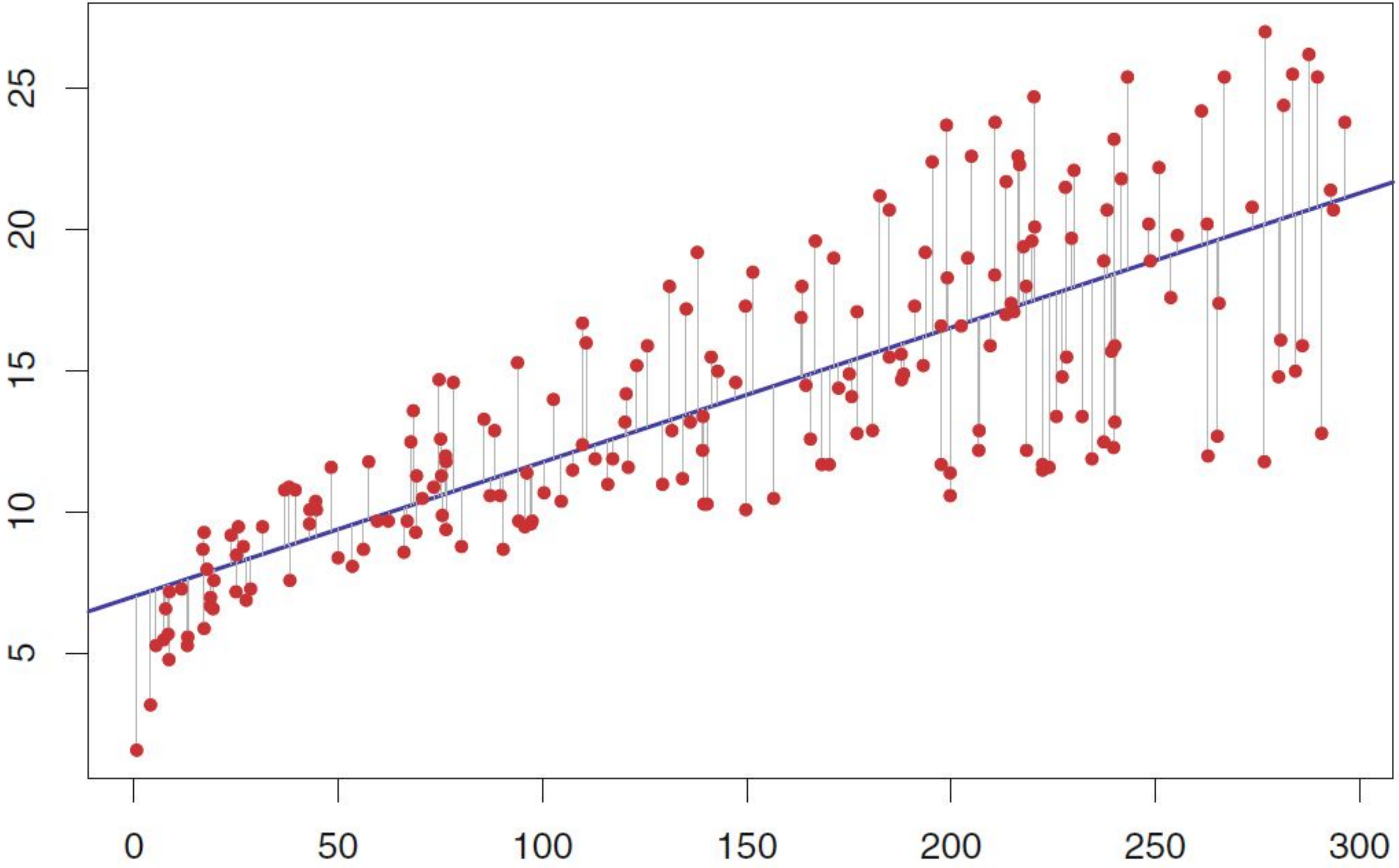
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

AIC:

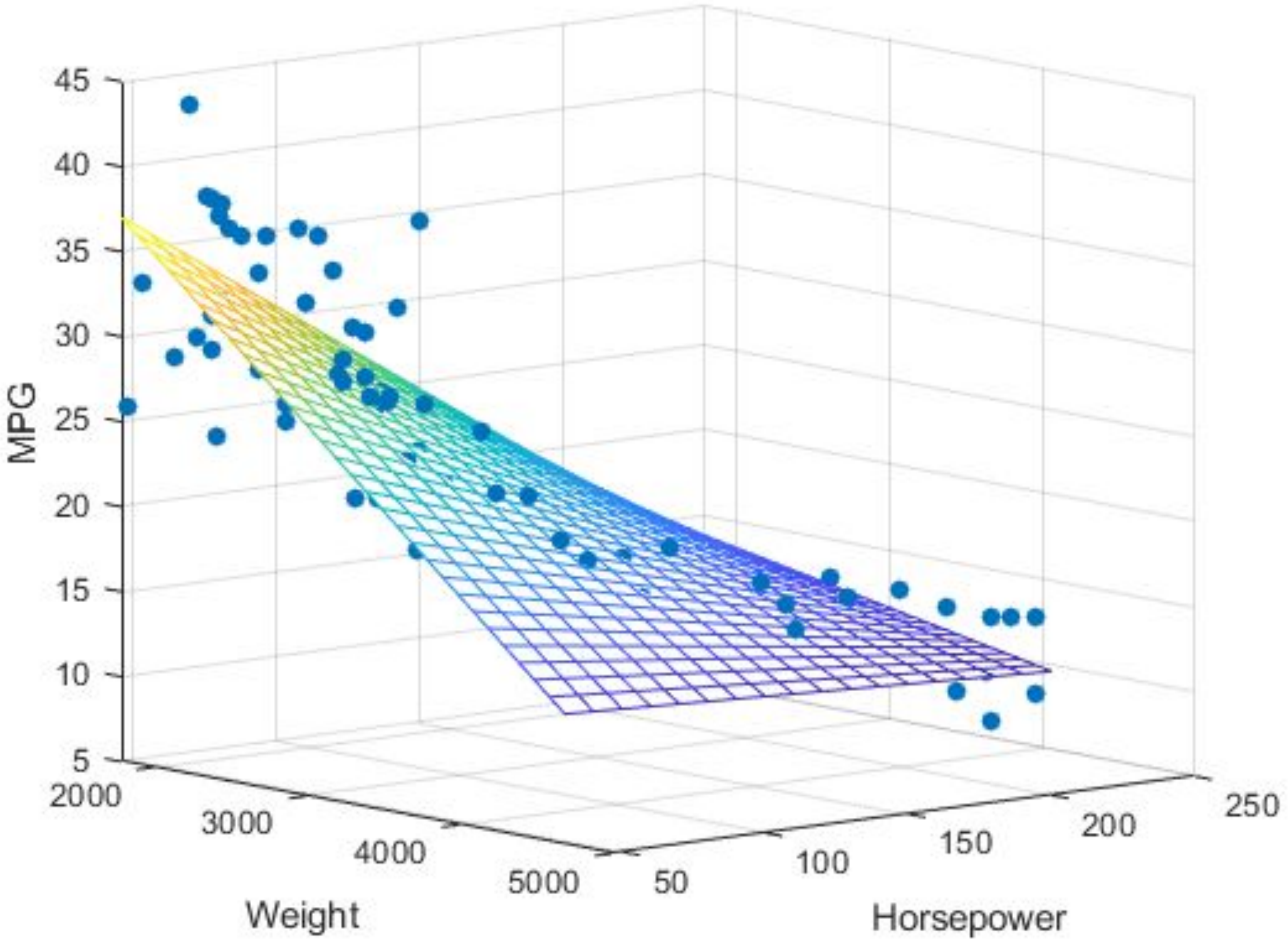
$$AIC = -2logL + 2q$$

BIC:

$$BIC = 2log(L) + qlog(N)$$



25
20
15
10
5
0
50
100
150
200
250
300


MPG
45
40
35
30
25
20
15
10
5
2000
3000
4000
5000
Weight
50
100
150
200
250
Horsepower

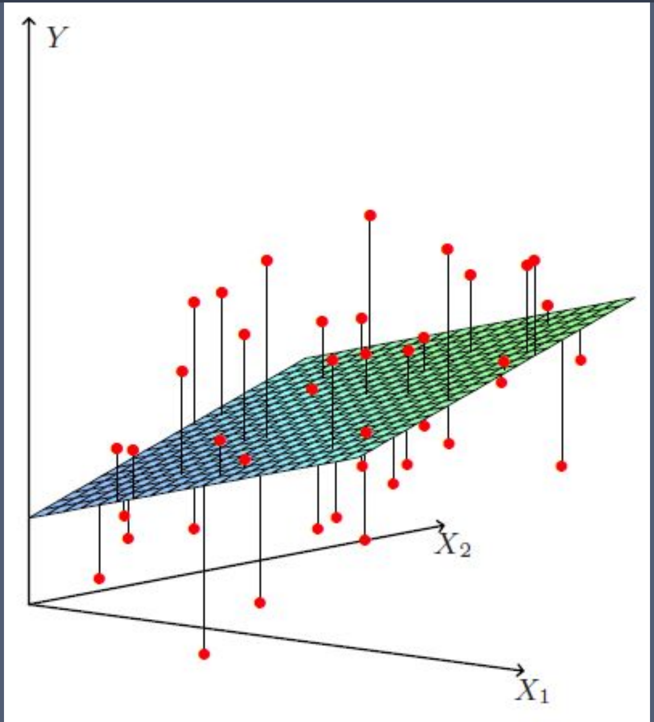
Estimation: Least Squares in LR

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)$$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$



Interpreting estimators

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 D_1 + \hat{\beta}_3 D_2$$

X1 is a continuous variable:

- sign
- size
- marginal effect

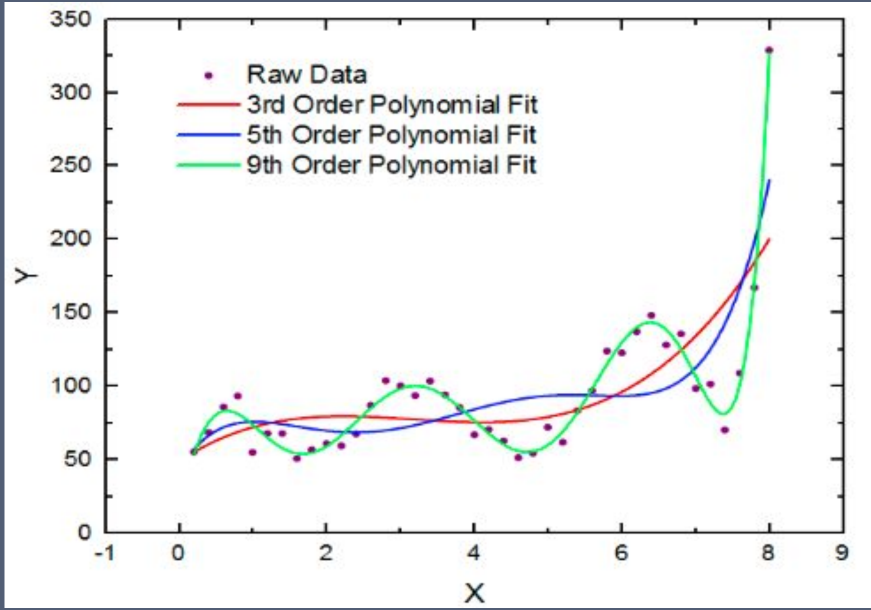
D1, D2 are the dummies of a categorical variable with 3 levels:

- reference category D3

Polynomial regression

f(x) is a polynomial of order k:

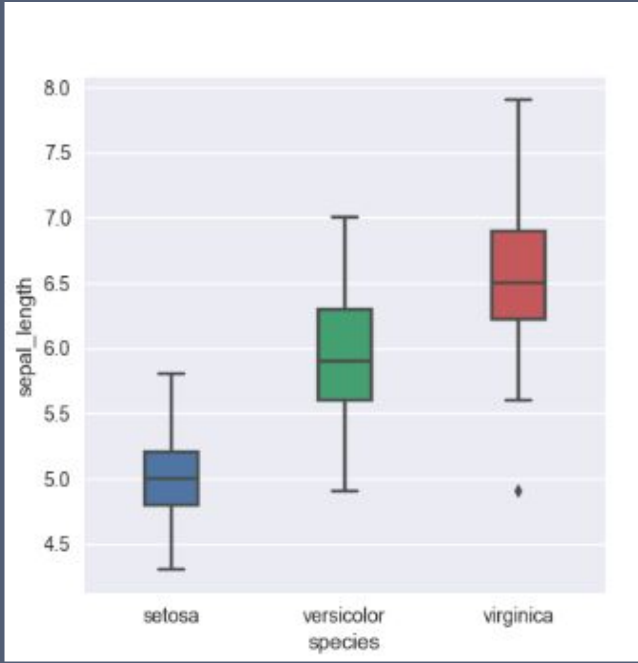
$$f(X) = \sum_{j=0}^{k} \beta_j X^j$$



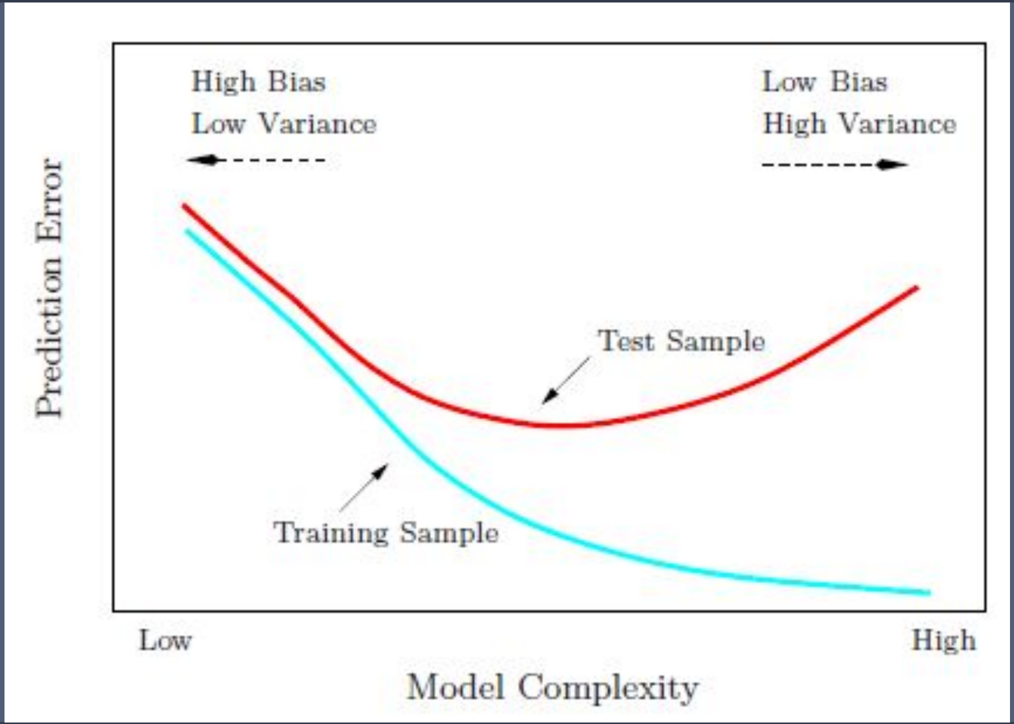
Outliers

In words of Hawkins, 1980:

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"



Validation



High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity

Conectar | Enviar mensaje | Más...

Joseba Elcano Esparza · 2º

Data Scientist | Telecom engineer

Barcelona, Cataluña / Catalunya, España · 159 contactos ·

Información de contacto

Institute for Bioengineering of Catalonia (IBEC)

Universitat Pompeu Fabra

Interés por oportunidades

Cargos de Research Assistant, Data Scientist y Data Analyst

Ver todos los detalles

Enviar mensaje

Más...

Javier Fernández · 1er

Data Scientist en inAtlas

Barcelona, Cataluña, España · **Más de 500 contactos** ·

Información de contacto

inAtlas

University of Havana


Conectar
Enviar mensaje
Más...
Gemma Labraña · 2º
MartiDerm
Data Scientist at Martiderm®
Barcelona, Cataluña, España · Más de 500 contactos ·
Información de contacto

Enviar mensaje

Más...

Ángel Molina Nogueras · 1er

Data Analyst en CEDEC S.A. Consultoría de Organización Estratégica

Barcelona y alrededores, España · **Más de 500 contactos** ·

Información de contacto

CEDEC S.A. Consultoría de Organización Estratégica

NEOLAND