FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Grado en Inteligencia Artificial

1. Introduction and Data Quality

Prof. Dante Conti

Prof. Sergi Ramírez

Credits: Tomàs Aluja-Banet

Guia Docente & Web Site / Atenea

https://www.fib.upc.edu/es/estudios/grados/grado-en-inteligencia-artificial/plan-de-estudios/asignaturas/PMAAD-GIA

https://ramia-lab.github.io/AdvancedModelling/

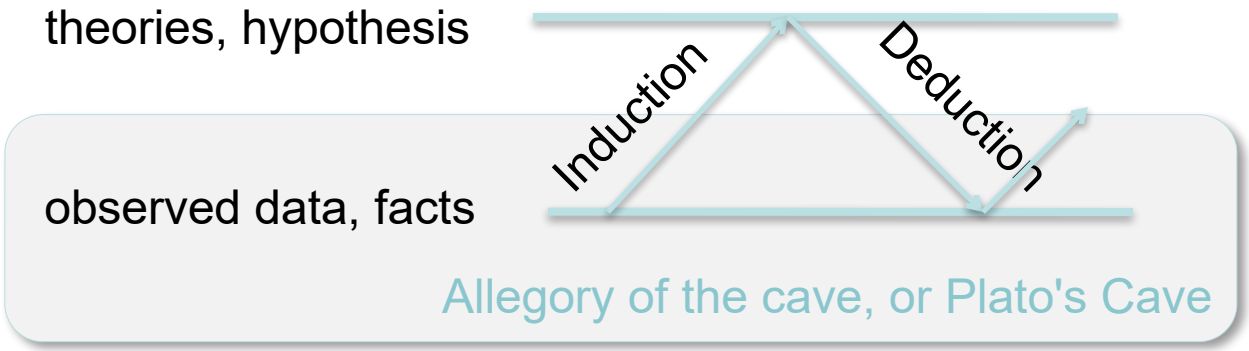
Proyecto

- Grupos de 5/6 personas para un total de 4 grupos por Laboratorio
- Practical work
  - Escoja un “real-world” problema o caso de estudio
  - Implemente algoritmos y métodos.
  - Redacte un informe técnico/gerencial
  - Defensa oral

El lenguaje de soporte será R, no obstante puede usarse Python.

Why we need data?

Learning is an iteration process between the real-world facts and the hypothesized world of theories



https://en.wikipedia.org/wiki/Allegory_of_the_cave

Deduction moves from idea/theory/hypothesis to observation.
Induction moves from observation to idea/theory/hypothesis .

Statistics/Data mining/Data Science concerns the **inductive** phase of learning

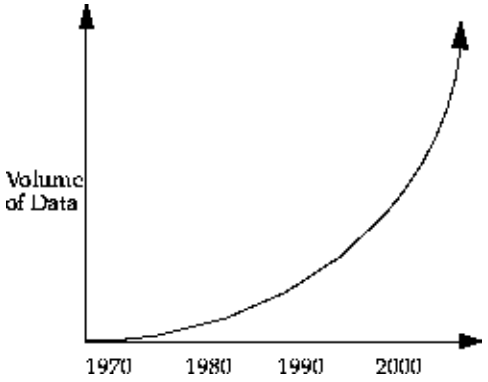
Data = Fit + Noise

Trends leading to data flood (Big Data)

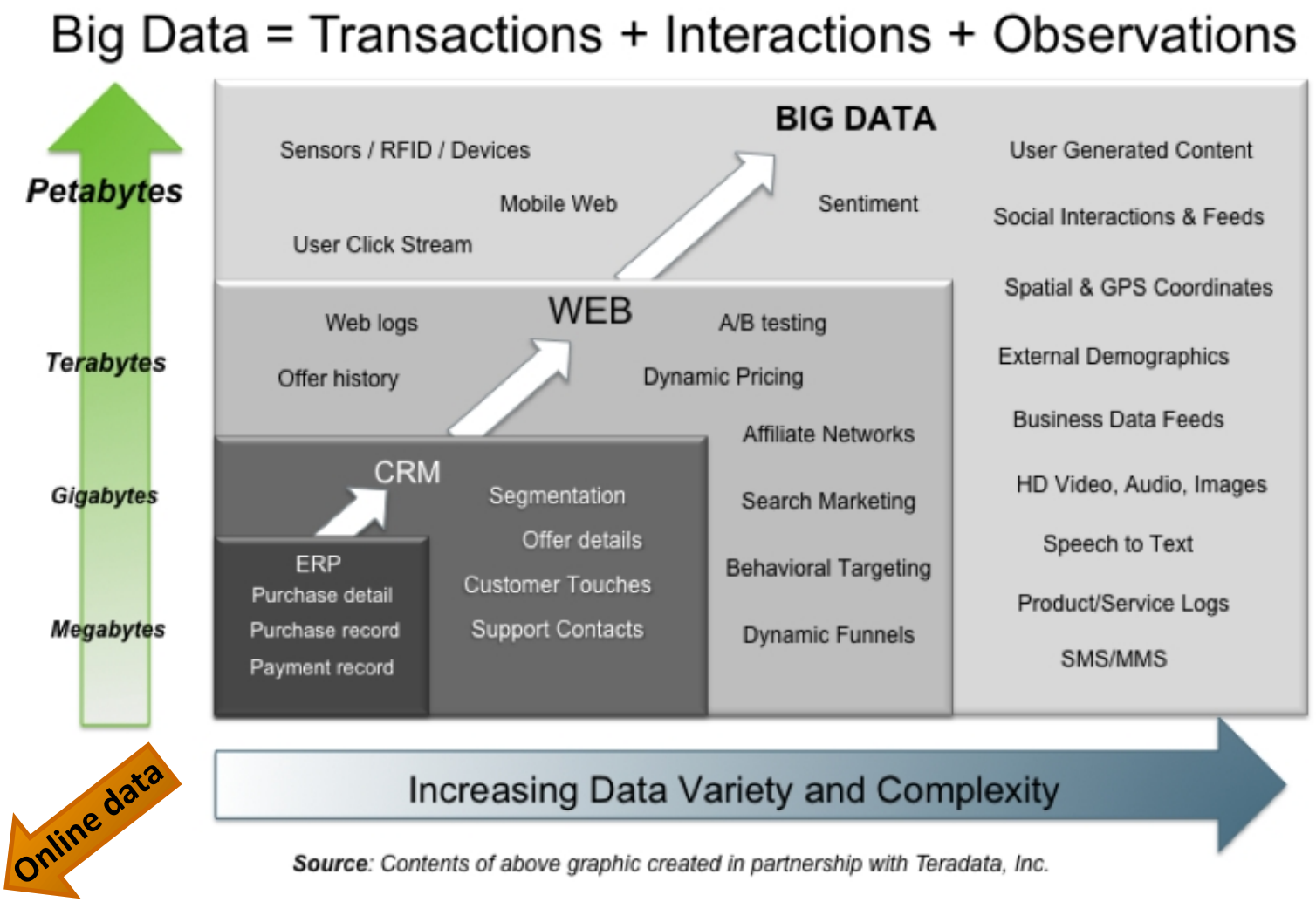
FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Exponential increase of data generation and storage
  - Bank, telecom, other business transactions ...
  - Scientific data: genomics, astronomy, health, …
  - Web, text, and e-commerce
  - Social networks, …

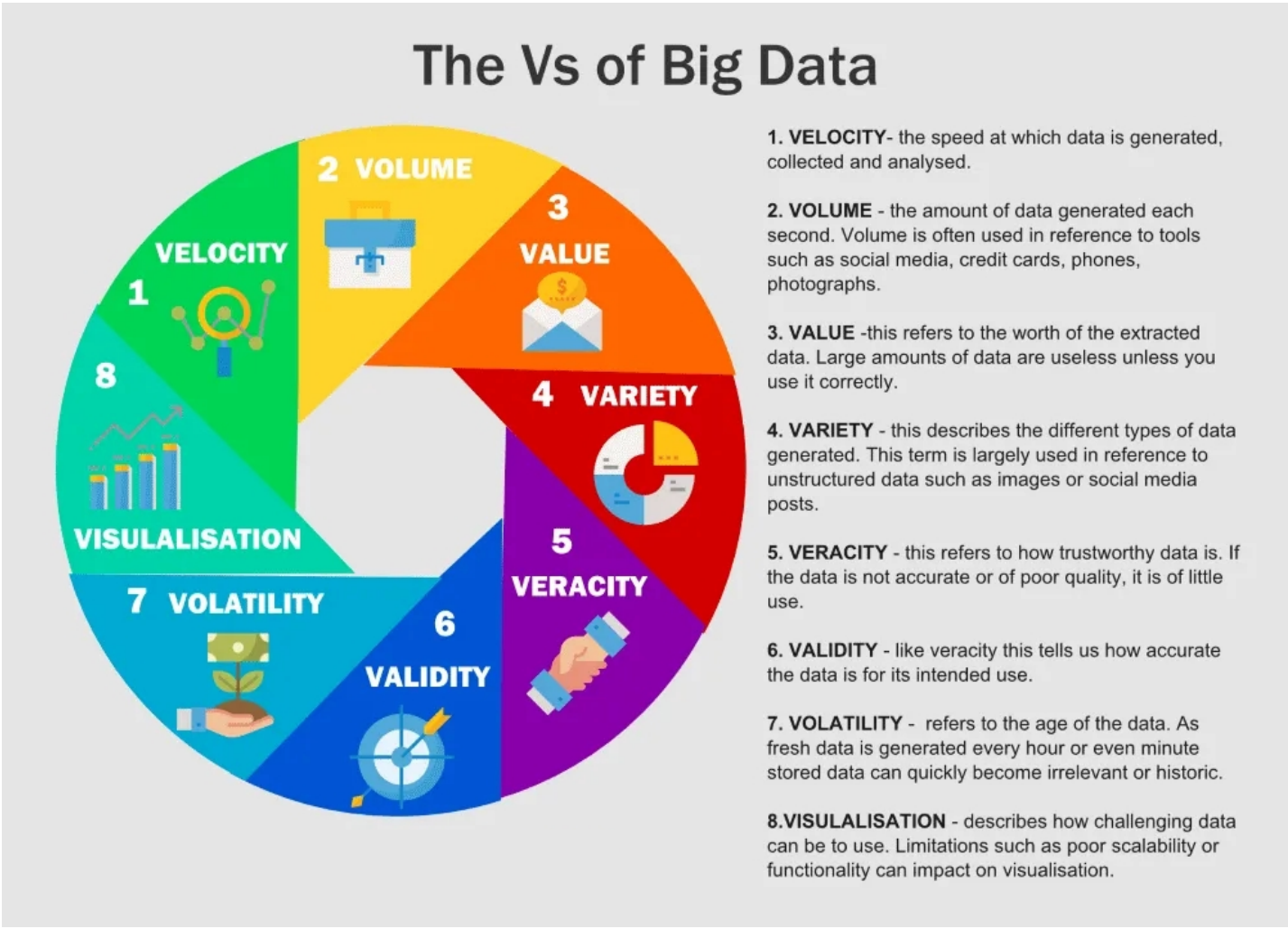
- Increase in data formats
  - Relational tables
  - Non structured tables, log files, …
  - Textual data, image data, …



- Real time, streaming data



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*



Source: https://algorithmxlab.com/blog/big-data/

My data, my treasure

Any stored data from any process always contains information about the generating phenomenon (**statistical regularity**).

Goal: **To reveal the information** (model, patterns, associations, trends, clusters, … hidden in the data

Data are routinely stored (and most will never be analyzed)

Data is a treasure for organizations (be aware of the data quality)

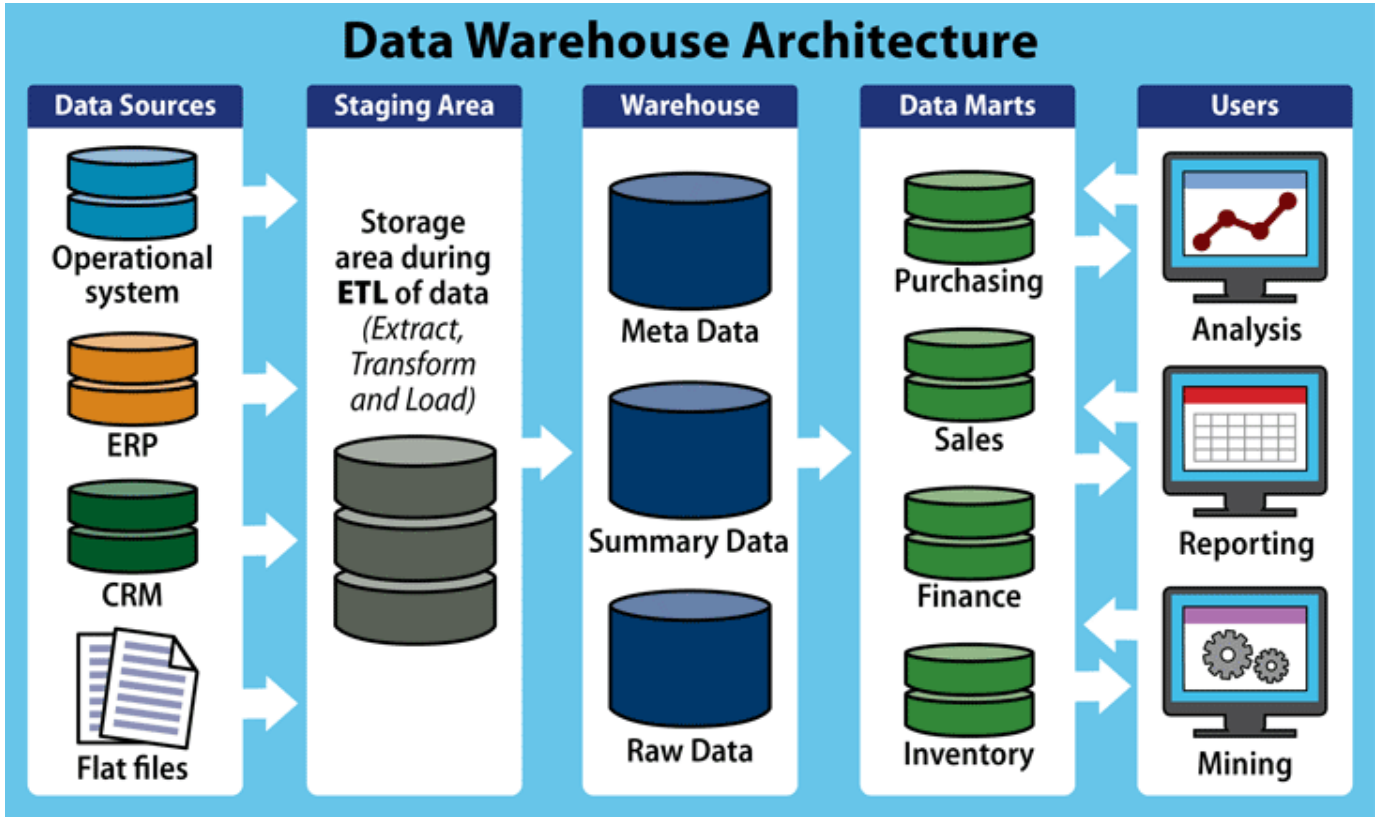
Any digital interaction is a potential valuable source of data.
Any process con be enhanced by analysis of its collected data.

How? *Selecting and reporting what is interesting*

SQL queries are NOT ENOUGH.

Assembling historical data in a consistent manner from their transactional processes is what is called *data warehousing or datalakes*. Data warehouse and similars are the memory of the company.
But that is not enough. We need to learn from the data.

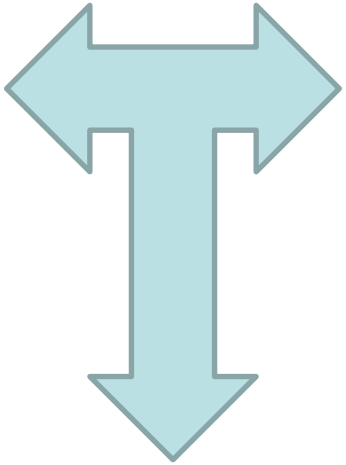
My data, my treasure



Source: https://todobi.com/comparativa-de-las-15-mejores-soluciones-de-data-warehouse/

Interdisciplinarity

Computer Science
Develop machines/algorithms that solve problems



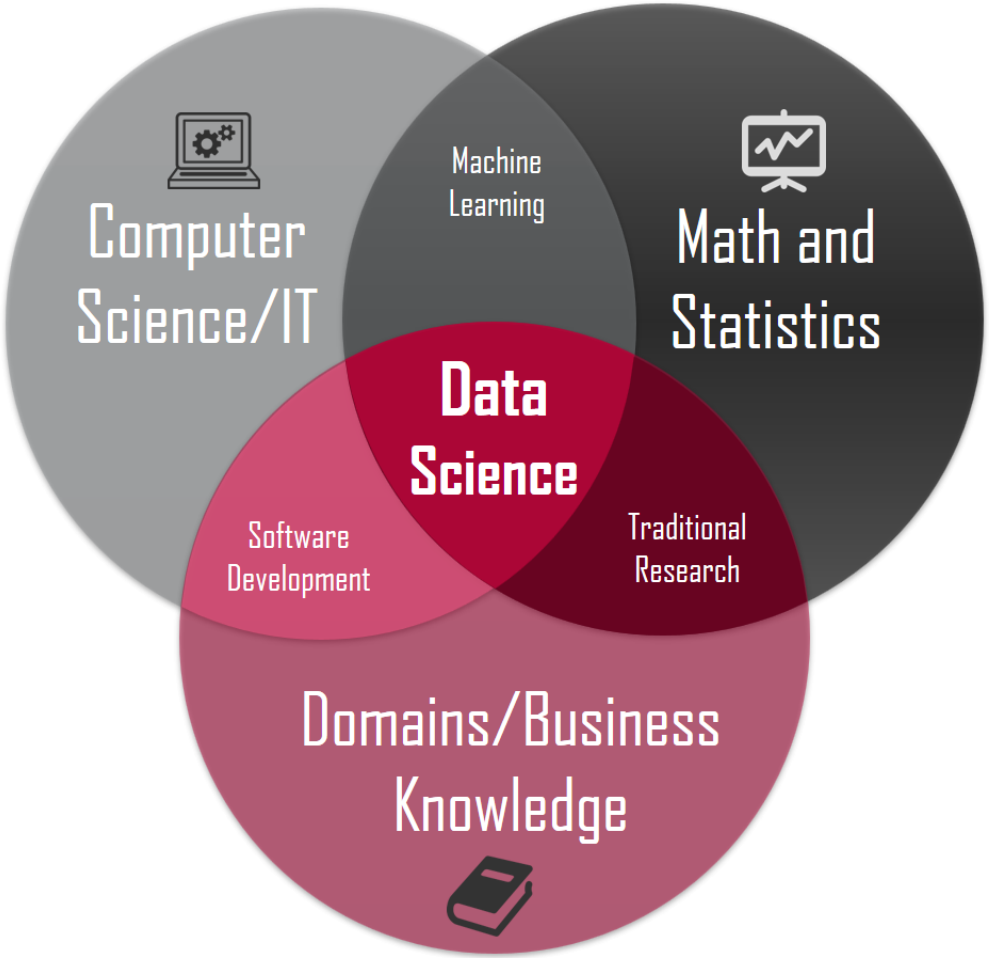
Statistics
Make inferences with confidence measures

Machine learning
Based on Statistics, make machines (algorithms) that program themselves to solve tasks.

Roseta stone (from L. Lebart, 1995)

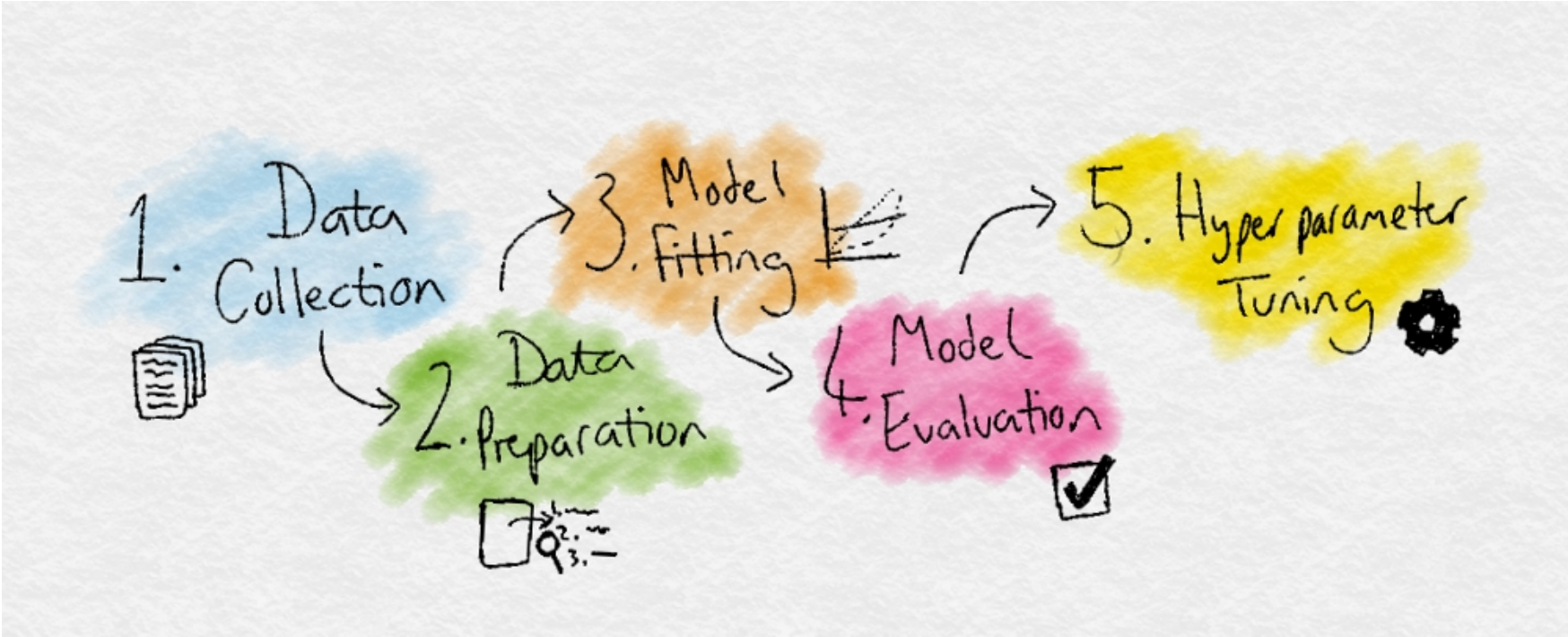
STATISTICS	MACHINE LEARNING
VARIABLES	ATTRIBUTES, FEATURES (DB: FIELDS)
INDIVIDUALS	INSTANCES (DB: REGISTRES)
EXPLANATORY VARIABLES , PREDICTORS,	INPUT
RESPONSE VARIABLES	OUTPUT (TARGET), CONCEPT
MODEL	NETWORK, TREE, ...
COEFFICIENTS	WEIGHTS
FIT CRITERION (OLS, WLS, ML)	COST FUNCTION
ESTIMATION	LEARNING (TRAINING)
CLASSIFICATION ("CLUSTERING")	UNSUPERVISED CLASSIFICATION
DISCRIMINATION	(SUPERVISED) CLASSIFICATION

Data Science



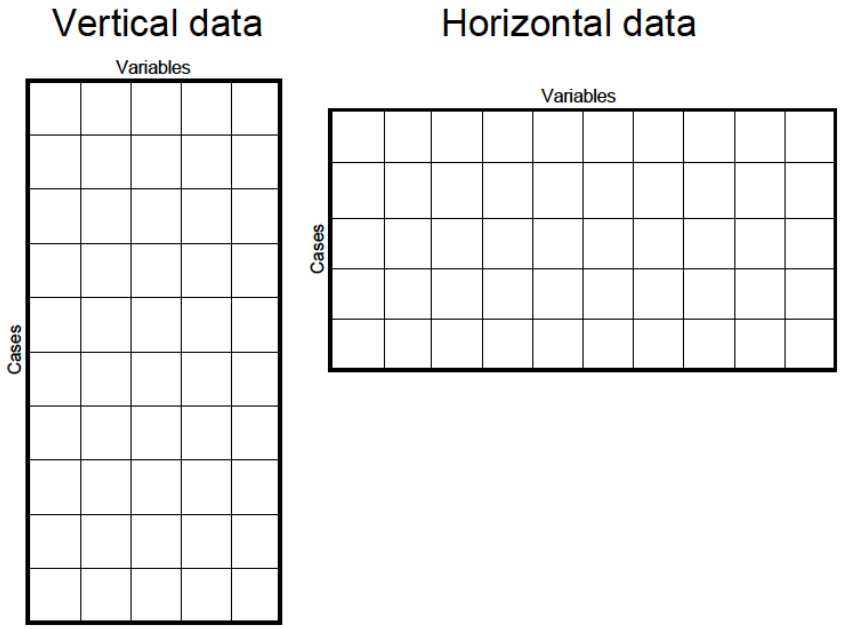
Source: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Process


1. Data Collection
2. Data Preparation
3. Model Fitting
4. Model Evaluation
5. Hyper parameter Tuning

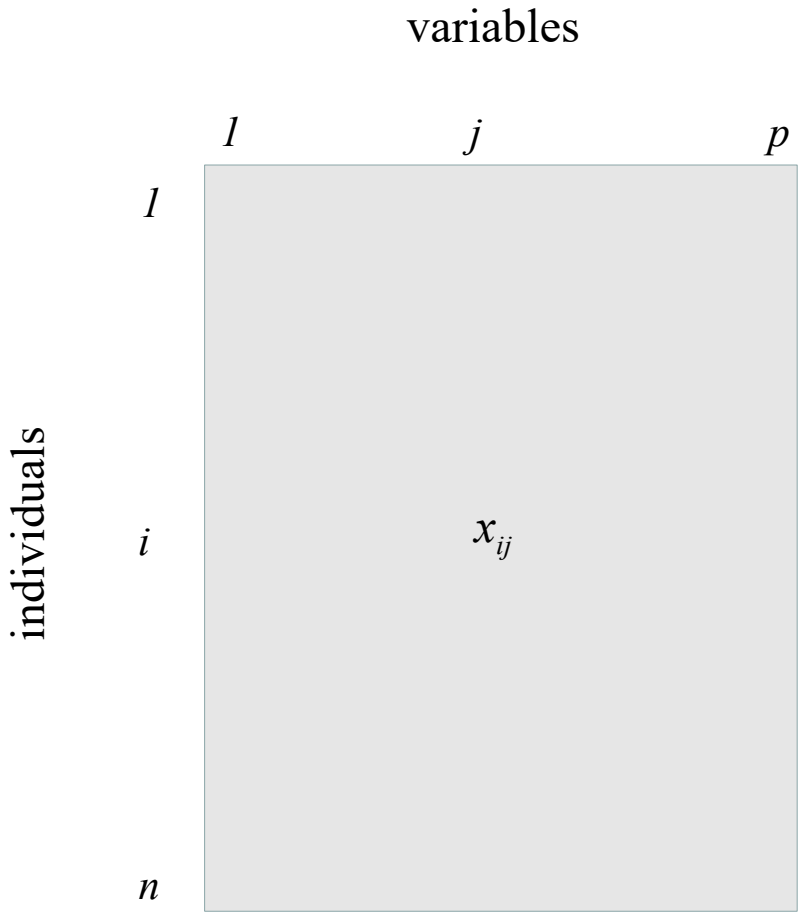
What are multivariate data?

- **Multivariate data** arise when researchers/users record the values of **several variables/attributes** on a set of **units** in which they are interested.
- This leads to a **vector-valued** or **multidimensional** observation for each unit.



Data file

Data is multivariate



Tables:
- Individual by variables (cont. or categ.)
- Counts

Transaction's data

Graphs
- Similarity matrices
- Link data

Textual data:
- Documents
- Html/Xml

Stream data
- Sensors
- Podcasting

Image data
- Medical
- Instagram
- …

The data matrix

- A multivariate data matrix, $\mathbf{X} \in \mathbb{R}^{n \times p}$, will have the form

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

where the element x_{ij} is the value of the jth variable for the ith unit.

- The theoretical entities describing the univariate distributions of each of the p variables and their joint distribution are denoted by random variables $X_1, \ldots, X_p$.

The data matrix (2)

- The rows (columns) of $\mathbf{X}$ will be written $\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top$ $(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)})$.

- That is, we may write

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)}) \ ,$$

where

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \quad (i = 1, \ldots, n) \ , \quad \mathbf{x}_{(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad (j = 1, \ldots, p).$$

Sample mean vector

- The sample mean of the jth variable is

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \ .$$

- The sample mean vector is

$$\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^\top \ .$$

- We also have

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n}\mathbf{X}^\top \mathbf{1}_n \ ,$$

  where $\mathbf{1}_n$ is a column vector of n ones.

Sample covariance matrix

- The (unbiased) sample variance of the jth variable is

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \ , \quad (j = 1, \ldots, p) \ .$$

- The sample covariance of the jth and kth variables is

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \ .$$

- The $p \times p$ matrix $\mathbf{S} = (s_{jk})$ is called the sample covariance matrix.

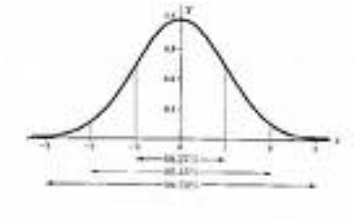
Rows of a data file

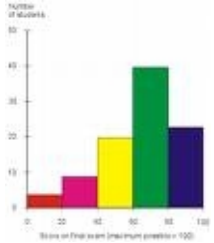
- Represent **individuals** or instances
  - In DS they can be from tens to millions, …
- Also called sample, example, record, …
  - Unit that can be **repeated**, at least theoretically, forming the **population under study**.
  - Thing to be **classified, associated, or clustered**
  - **Characterized by** a predetermined set of **attributes**
- We use **all available data**.

Sampling betrays the spirit of DS, where few individuals may keep the most precious information and only can be justified for alleviating the computational cost

Columns of the data file

- Each instance (row) is described by a predefined **set of features**, its variables or “attributes”
- A variable is a measure of individuals which can take different values (according a probabilistic function)

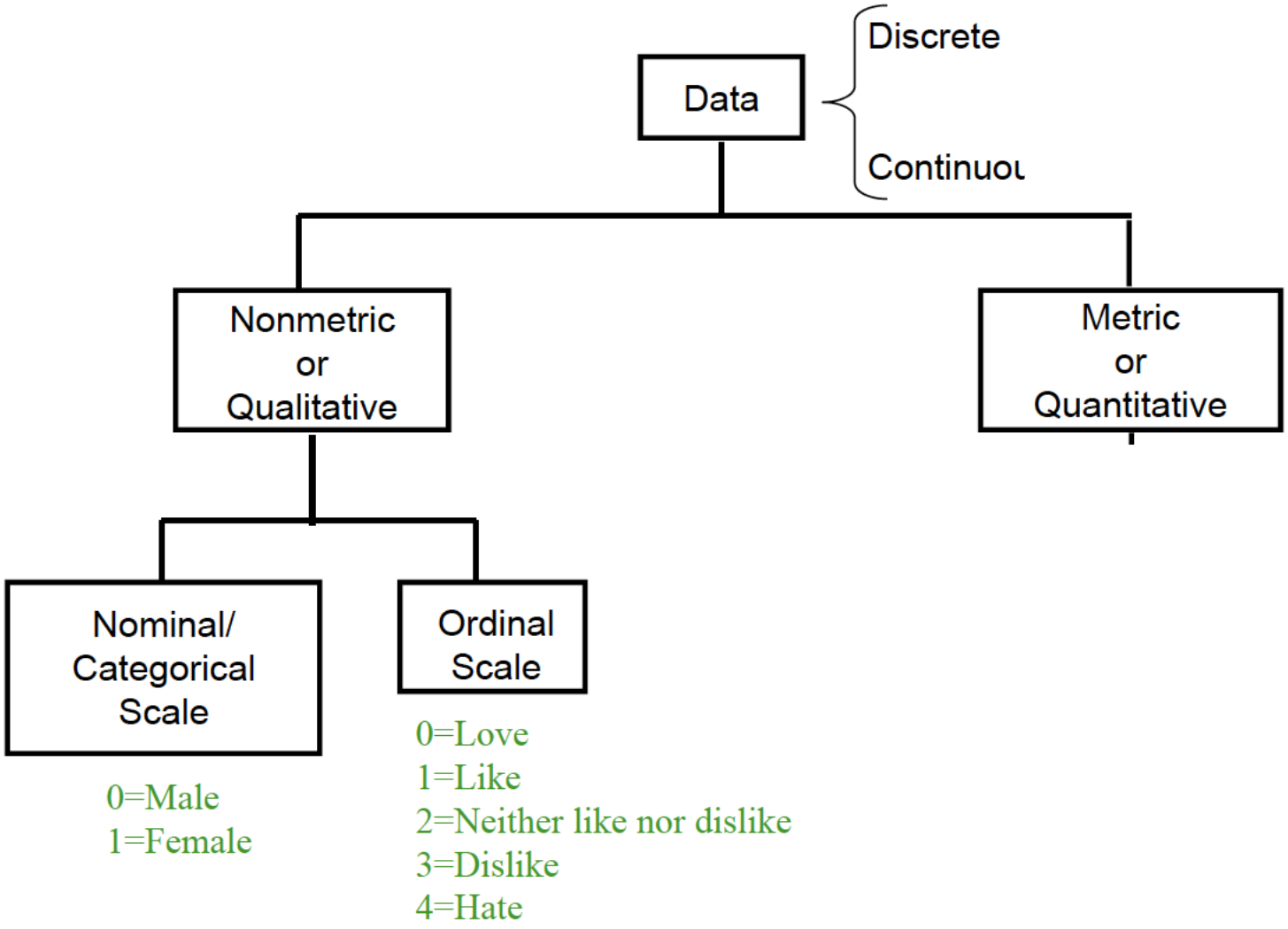




- Possible attribute types (“levels of measurement”):
  Binary, nominal, ordinal, interval, ratio, textual,…

- **Restriction**: Same variables measured in all individuals and in the same order.
- Different formats are possible (fixed, csv, …).
- First rows usually contain the dictionary of variables (var. labels)

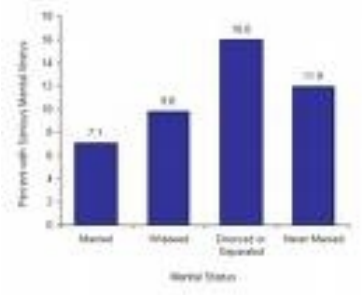
Type of variables



Nominal variables

- Values are **distinct categories** represented by symbols
  - Values themselves serve only as labels or names
- Example1: attribute *citizenship*: "Mexican", "German", "French",…
- No relation is implied among nominal values (no ordering or distance measure)
- Example2: attribute *marital status*: "single","married", "divorced", and "widow"

- Only percentages and tables can be calculated
- Bar plots (among many other plots) for graphical representation



Special case: Binary/Dichotomous data (Yes/No) (Boolean attribute)

- Some DS algorithms **cannot operate** on nominal data directly. They require all input variables and output variables to be numeric.

  Binarisation (one-hot-encoding)

single	1	0	0	0
divorced	0	0	1	0

https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Ordinal variables

- **Impose order** on values, but: no distance between values defined
- Example: attribute *size* in clothes:
  - Values: “XXL” > “XL” > “L” > “M” > “S”
- Or *social status*
  - Values: “upper class” > “middle/high” > “middle” > “middle/lower” > “lower class”
- Arithmetic calculations not possible (as in nominal variables).
- Tables and percentages and bar plots but emphasizing the ordering of values.
- **Internal encoding** of ordinal variables preserving the order. For example:
  “lower class” -> 1
  “middle/lower” -> 2
  “middle -> 3
  …

Count (or discrete) data

- Very often a variable is a result of a count.
- Examples: attribute “number of words of a sentence”, “number of students of a course”, “number of products bought”, “number of bugs per program”, “number of unemployed per country”, …

- Usually modeled by the Poisson distribution.

Poisson Distribution Formula

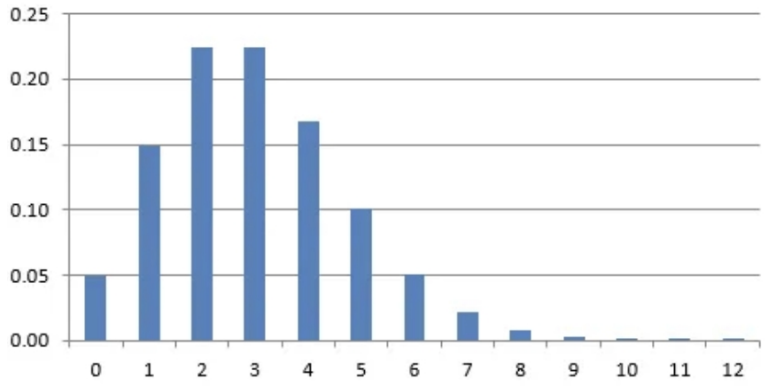
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

x = 0, 1, 2, 3, …

λ = mean number of occurrences in the interval

e = Euler’s constant ≈ 2.71828



Understanding the data

- Variables have meaning
  - Normally expressed in the **metadata file**

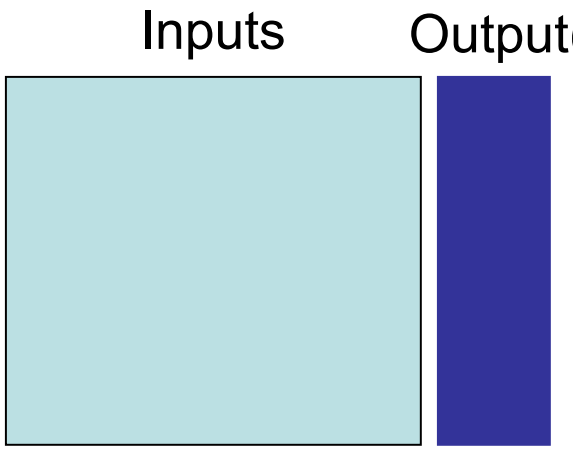
- **Role of variables**
  - **Response**, output or target. They are the variables we want to study, model, predict, … (Y)
  - **Explanatory**, input or predictors. They are the variables used to predict the former. (X)

- **Data origin**
  - **Primary** source: we collect the data (sampling): surveys, health studies
  - **Secondary** source: existing data (public webs, web scrapping,…)

Two types of data files

Depending on a framework **with or without response(s)** variable

e.g. transactions, ecological, survey,…



Data to explore, to describe, to find associations (i.e., itemsets), …

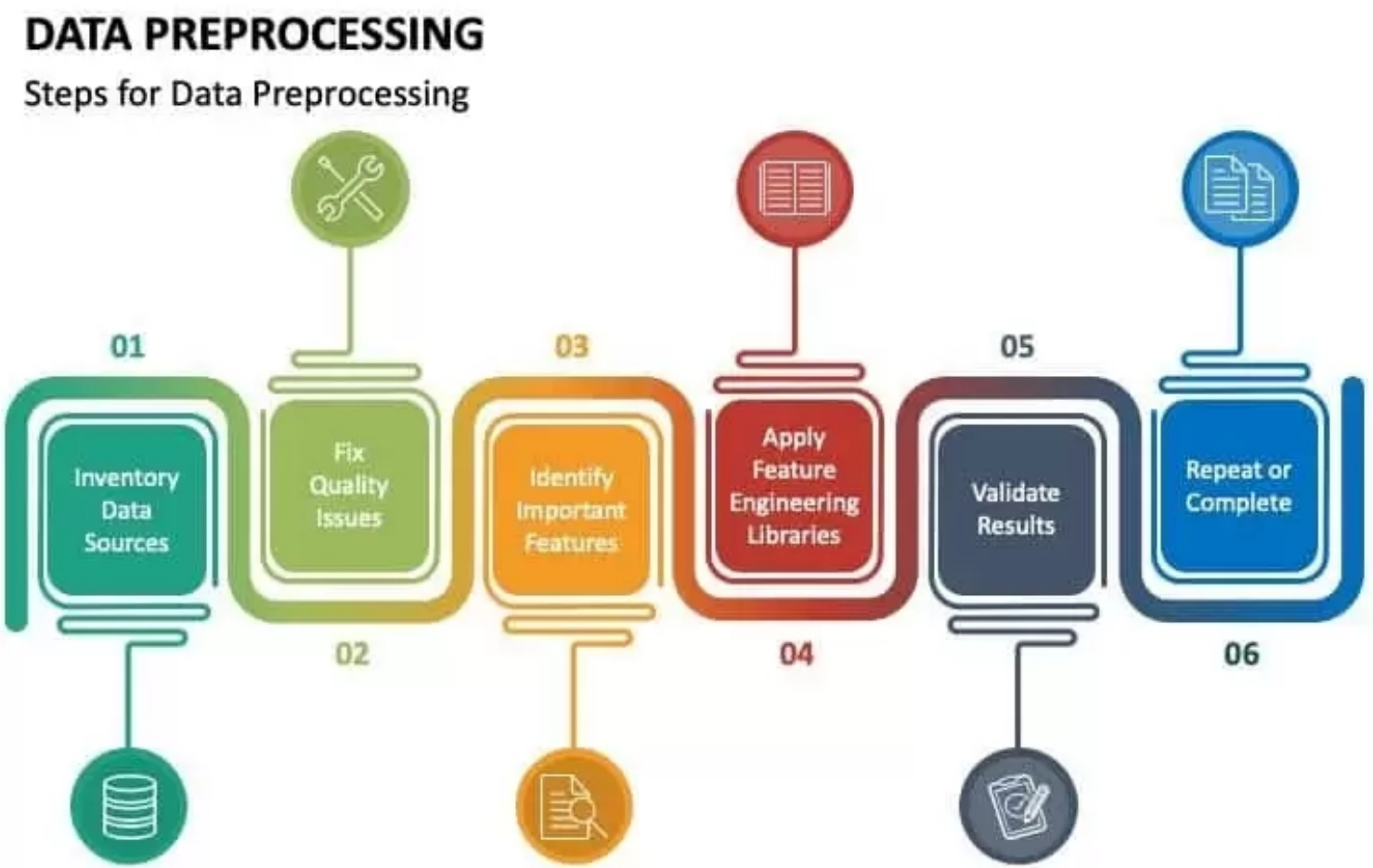
Idem, but data **to find a model to predict the response**

The Data Mining chain (batch mode)

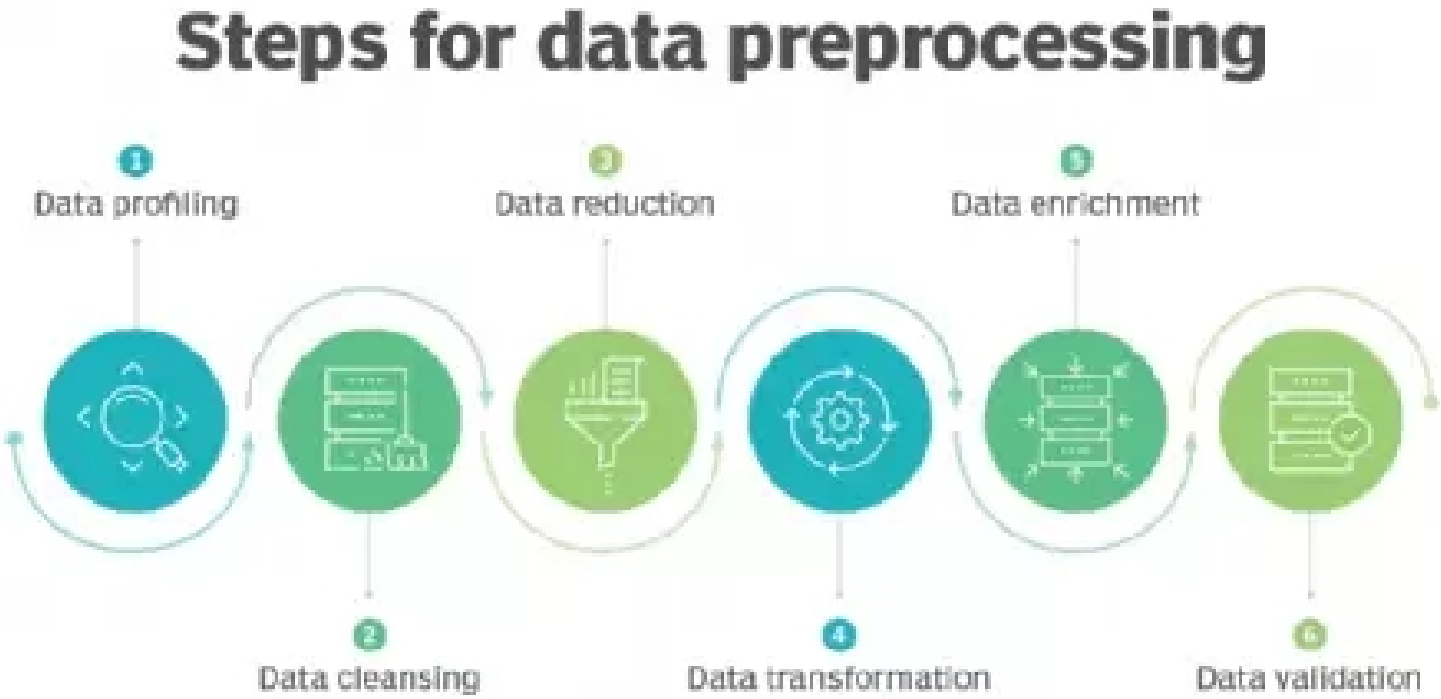
FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

1. Preprocessing
   - First summary of data: measures of central tendency and dispersion
   - Cleaning
   - Preparing the data for the analysis
2. Summary
   - Univariate description, bivariate description of the data, EDA,…
3. Multivariate exploration
   - Visualization
   - Clustering
   - Profiling
4. Modeling
   - Finding the optimal model for prediction
   - Obtaining honest estimates of the prediction error
5. Deployment/communication
   - Using the model in a real context and storing the results

Activity # 1 Advanced Preprocessing

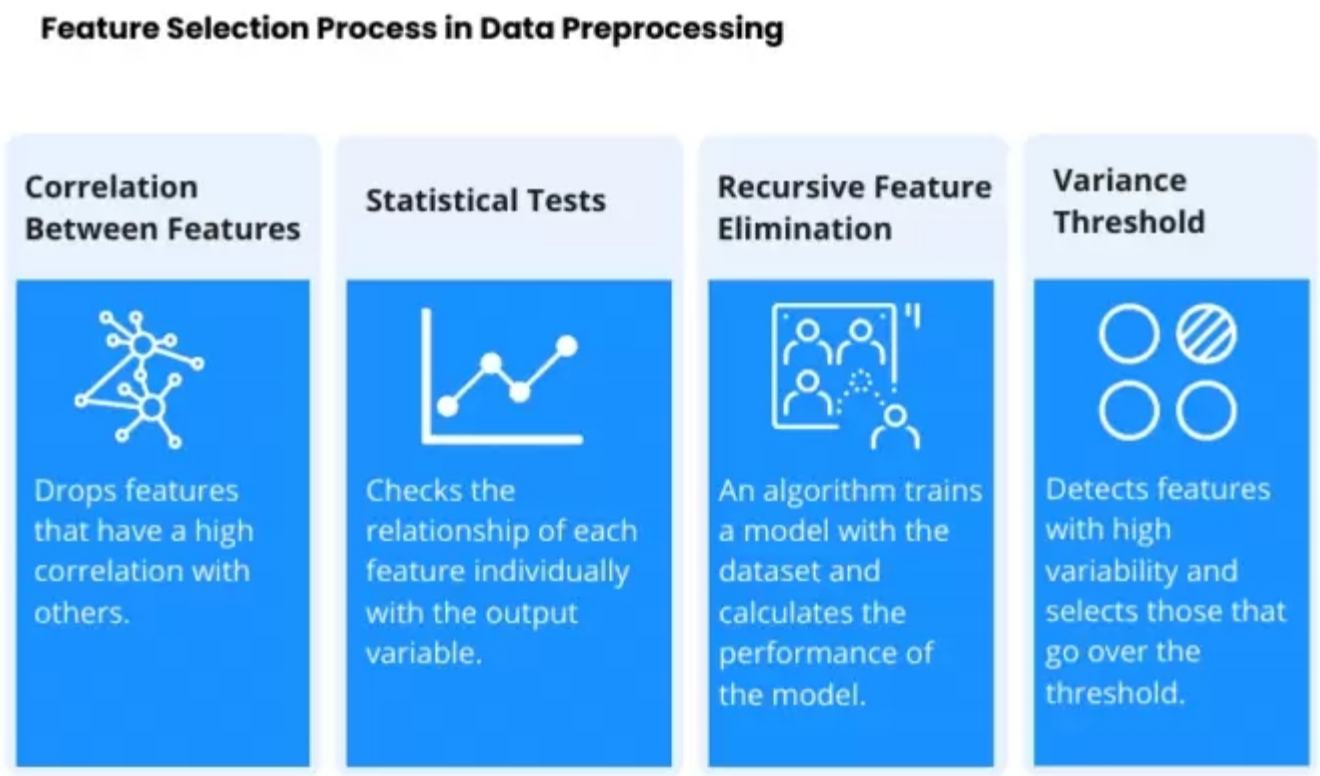


Activity # 1 Advanced Preprocessing Check

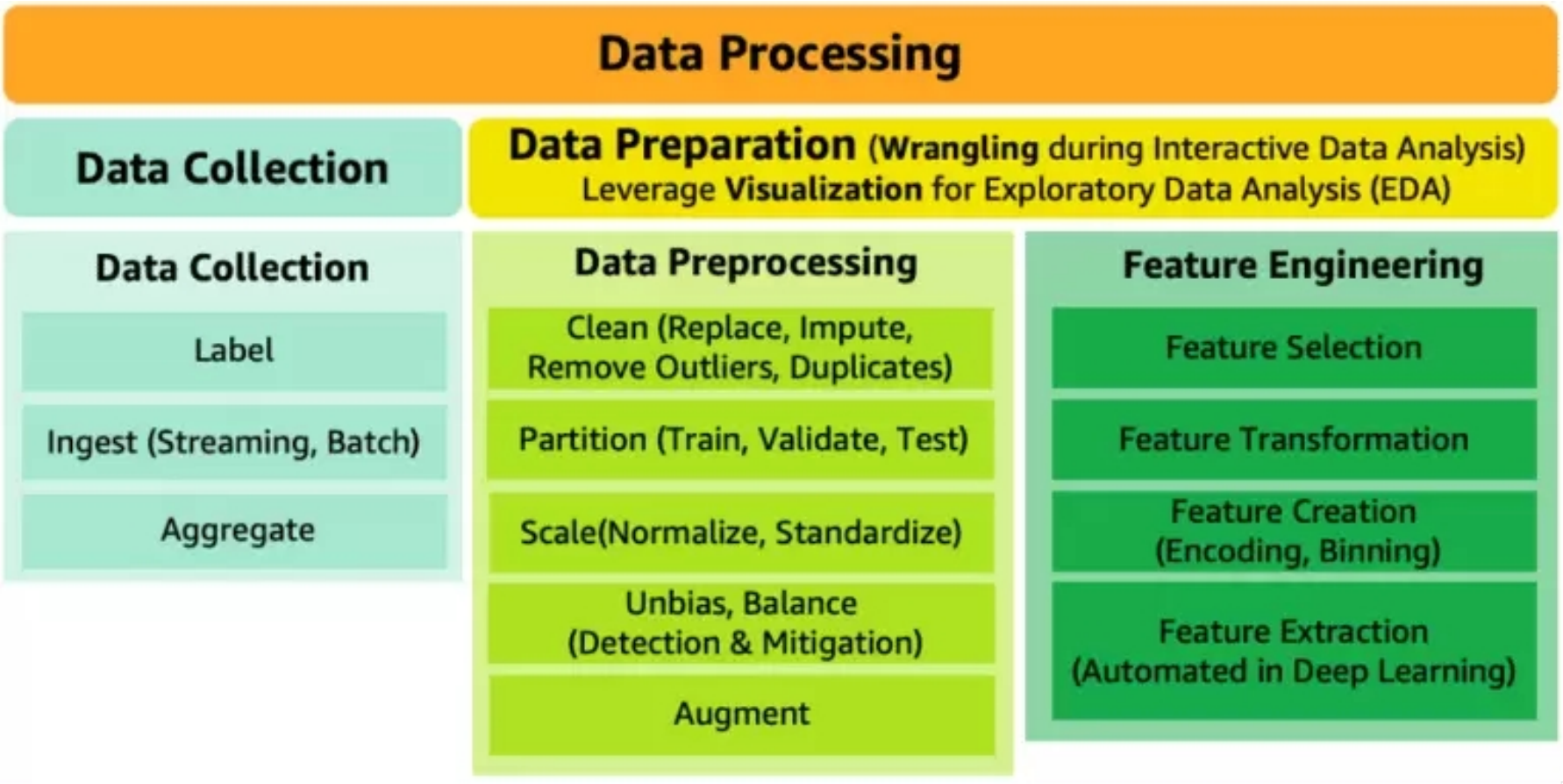


- https://dataexpertise.in/data-preprocessing-techniques-for-data-scientists/
- https://www.kaggle.com/code/nkitgupta/advance-data-preprocessing

Activity # 1 Advanced Preprocessing Check



Activity # 1 Advanced Preprocessing Check

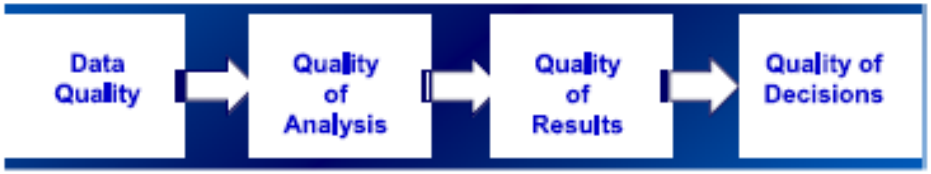


Task # 1 – Survey on Preprocessing

AI Communications 29 (2016) 627–663
DOI 10.3233/AIC-160710
IOS Press

A survey on pre-processing techniques: Relevant issues in the context of environmental data mining

Karina Gibert [a,*], Miquel Sànchez–Marrè [b] and Joaquín Izquierdo [c]
[a] *Knowledge Engineering and Machine Learning Group, Department of Statistics and Operation Research, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Catalonia, Spain*
[b] *Knowledge Engineering and Machine Learning Group, Computer Science Department, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Catalonia, Spain*
[c] *Fluing-IMM Universitat Politècnica de València, Valencia, Spain*



AIC710def.pdf (Available at Atenea)

Task # 1 – Survey on Preprocessing

REVIEW and CURRENT STATUS (PMAAD COURSE, 2024-25/QII) for PREPROCESSING ANALYSIS			
TASK	**Tools/Methods**	**Current Status**	**Suggestions**

Review-Preprocessing-PMAAD.pdf (Available at Atenea)

- **Preparing the data for the analysis**
  - Feature selection: *filtering the uninteresting variables*
  - Feature extraction: *deriving new variables*
  - Transformations
    - Recoding (numeric $\rightarrow$ categorical)
    - Quantifying a nominal variable (categ. $\rightarrow$ numeric)
    - Normalizing

$$z = \frac{x - \bar{x}}{s_x}, \quad \frac{x}{\max(x)}, \quad \log(x), \quad \ldots$$

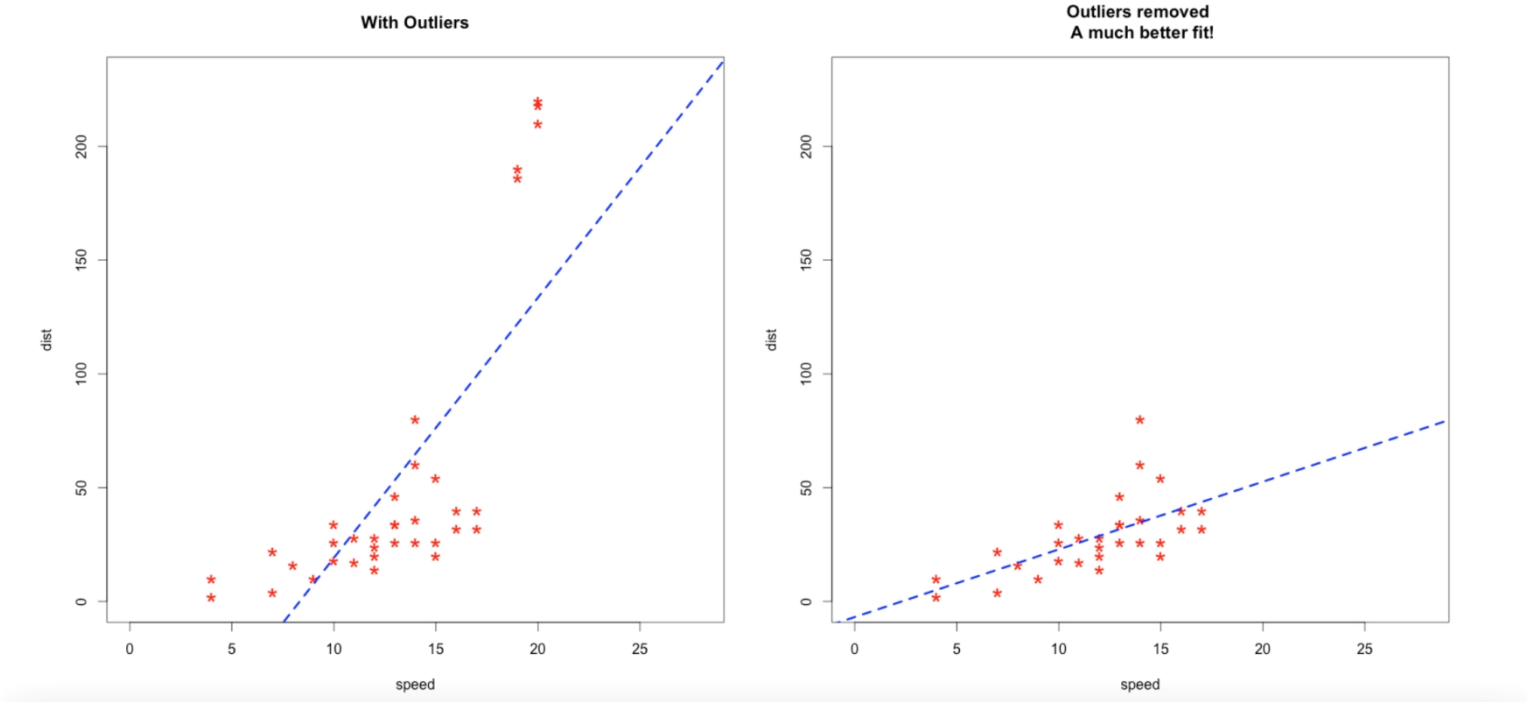
- **Data Cleaning**
  - Errors: Typos. Detect them and correct them
  - Missing values
  - Outliers

  Missing values can bias the results. In multivariate data, they may arise for several reasons:

  1. Non-response in sample surveys.
  2. Dropouts in longitudinal data.
  3. Refusal to answer particular questions in a questionnaire.

- **Complete-case analysis**: omit any case with a missing value on any of the variables.
- **Available-case analysis**: use all the cases available to estimate quantities of interest.
- **Imputation**: the practice of "filling in" missing data with plausible values.

- **Data Cleaning**
  - Errors: typos. Detect them and correct them
  - Missing values
  - Outliers: **They can bias the results**. Remove them or treat them as NA.



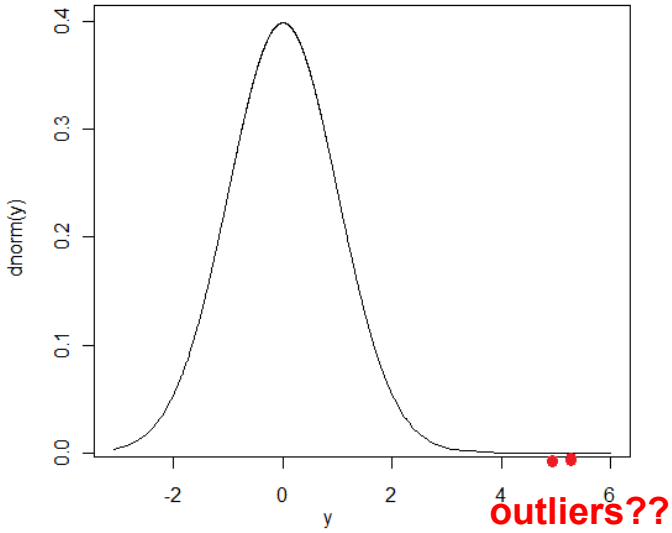
Outlier detection

What is an outlier? Definition of Douglas Hawkins: "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"

https://link.springer.com/book/10.1007%2F978-94-015-3994-4

Statistics-based intuition. Data is always generated by a mechanism that bestow a specific probability distribution (i.e., normal data follow a "normal generating data mechanism"). Outlying data may be a:

- very unlikely events for the current generating mechanism
- data following a different generating mechanism



if X~N(0,1)	Prob(x≥X)
1	0.1586553
2	0.02275013
3	0.001349898
4	3.167124e-05
5	2.866516e-07

Univariate detection of outliers. The Boxplot

- **The Boxplot (Tukey, 1977)** is a graphical display for exploratory data analysis, where the outliers appear tagged.
- Two types of outliers are distinguished: mild outliers and extreme outliers.

➢ An observation x is declared an **extreme outlier** if it lies outside of the interval

$$(Q1-3\times IQR,\ Q3+3\times IQR),$$

where $IQR=Q3-Q1$ is called the Interquartile Range.

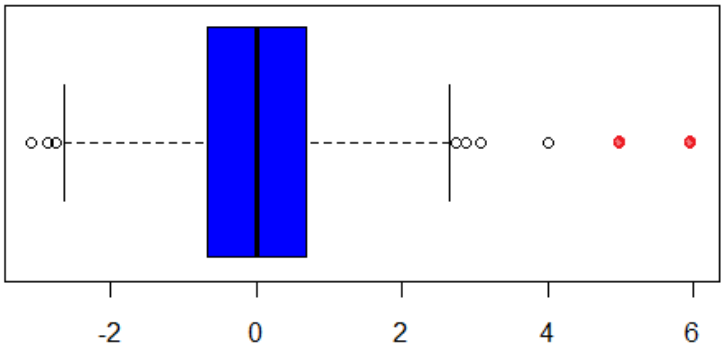
➢ An observation x is declared a **mild outlier** if it lies outside of the interval

$$(Q1-1.5\times IQR,\ Q3+1.5\times IQR).$$

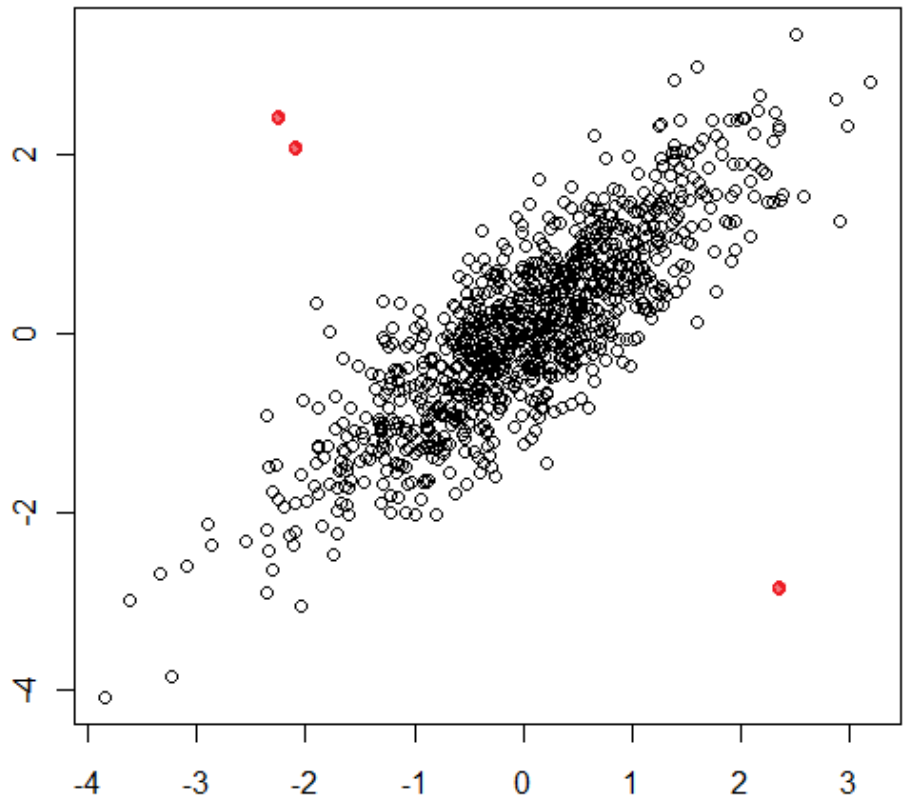
- The numbers 1.5 and 3 are chosen by comparison with a normal distribution.
- If x ~ Normal :

$\text{Prob}(X \geq Q3+1.5\times IQR)= 0.003488302$

$\text{Prob}(X \geq Q3+3\times IQR)= 1.170971e\text{-}06$



Outliers in a multivariate context



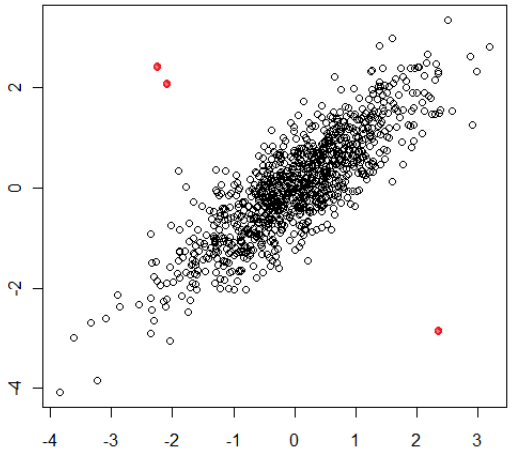
Outliers are multivariate

Univariate detection of outliers doesn't imply multivariate detection

Multivariate outliers

Outliers are multivariate

Univariate detection of outliers **doesn't imply multivariate detection**



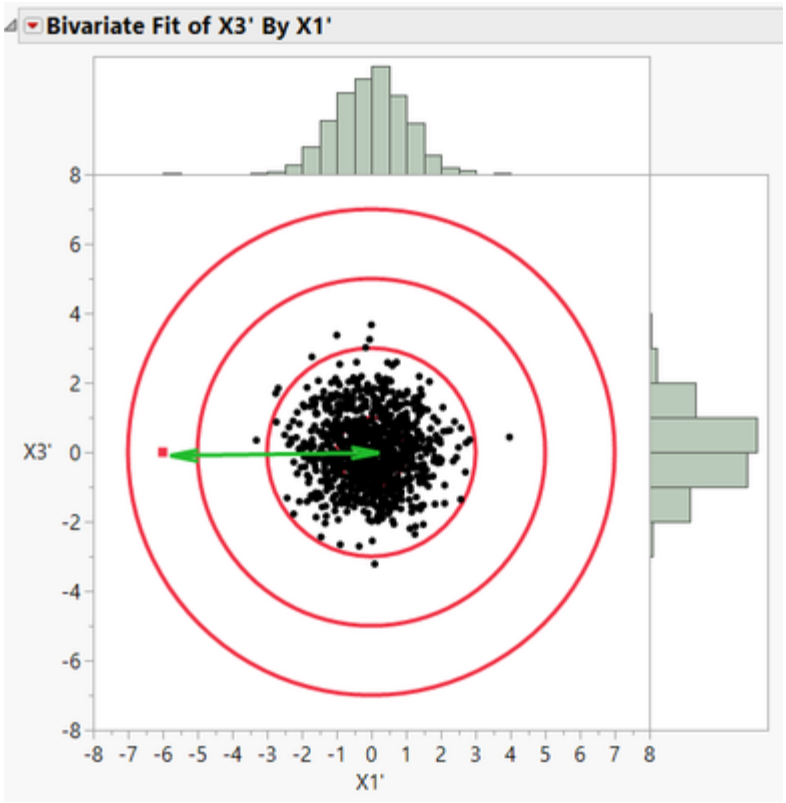
Then, the detection of outliers is based on computing the Mahalanobis distances to the central point of data.

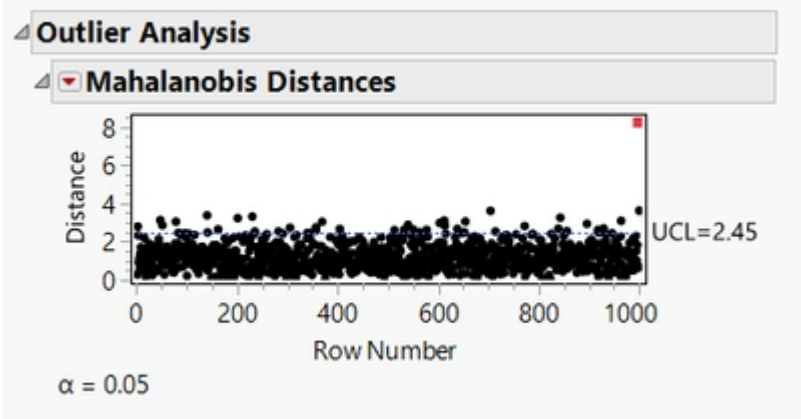
$$D_M^2(i,G) = (x_i - G)' V^{-1} (x_i - G)$$ Mahalanobis distance

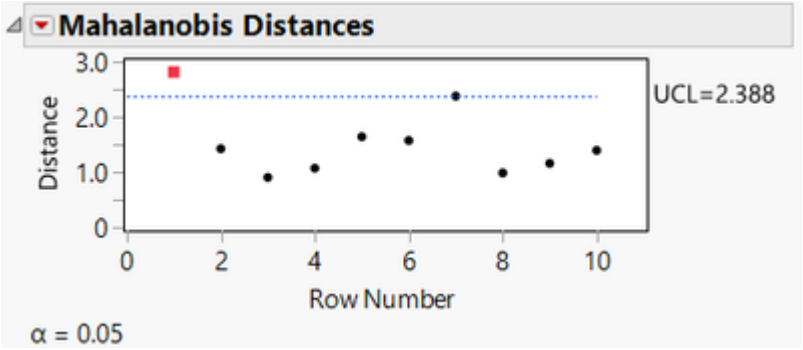
The **Mahalanobis distance** is a measure of the distance between a point *i* and a distribution *G*. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away *i* is from the mean of *G*.

Mahalanobis Distance

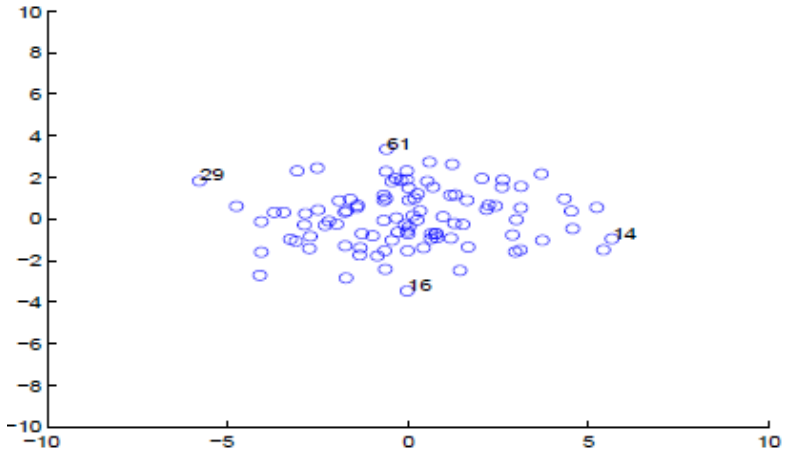
$$D_M(\vec{x}_i) = \sqrt{(\vec{x}_i - \vec{\mu})^T S^{-1} (\vec{x}_i - \vec{\mu})}$$







Mahalanobis Distance

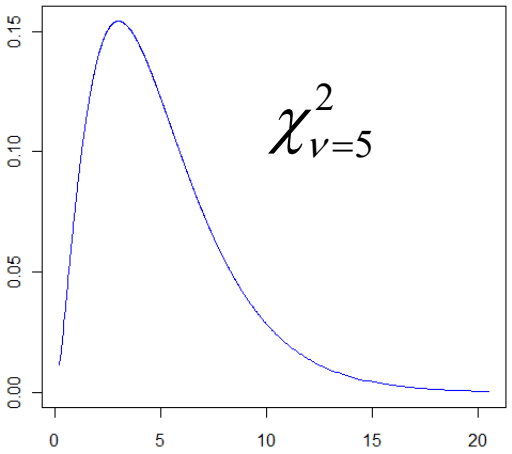


If generating mechanism is Normal distributed:

$$D_M^2(i,G): \quad \chi^2_{\nu=\text{dim space}}$$

It allows to establish a threshold for outlying points:

$$\chi^2_{\nu=\text{dim space}}(0.99)$$



Short distances occur more often

```
> qchisq(0.99,5)
[1] 15.08627
```

Detection of multivariate outliers

Problem: computation of G and V are contaminated by outliers
(G = mean of variables. V = matrix of variances

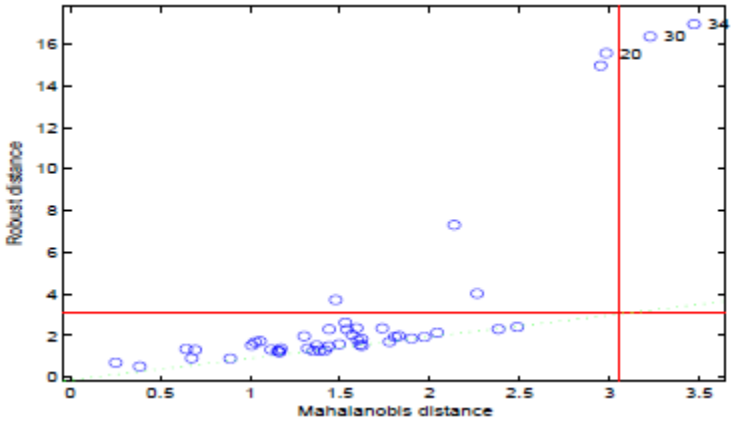
Take a value of h (size of data assumed not containing outliers), *h* must be > p (number of variables). Usual values = $0.75n$

Initialization of an estimation of G and V : Compute the Mahalanobis distances $D^2_M(i,G)$ for all points i.

1. Rank the $D^2_M(i,G)$ and retain the h individuals with lower $D^2_M(i,G)$
2. Update G and V till convergence.

Plot the final "robustified" Mahalanobis distances with the initial Mahalanobis distances to detect the outliers

Minimum Covariance Determinant (MCD)



Detecting outliers in R

Moutlier {chemometrics}

Multivariate outlier detection using the Mahalanobis distance can be used. Plot of the classical and the robust (based on the MCD) Mahalanobis distance is drawn.

```
Moutlier(X, quantile = 0.975, plot = TRUE, ...)
```

Arguments

X	numeric data frame or matrix
quantile	cut-off value (quantile) for the Mahalanobis distance
plot	if TRUE a plot is generated

For multivariate normally distributed data, a fraction of 1-quantile of data can be declared as potential multivariate outliers. These would be identified with the Mahalanobis distance based on classical mean and covariance. For deviations from multivariate normality center and covariance have to be estimated in a robust way, e.g. by the MCD estimator. The resulting robust Mahalanobis distance is suitable for outlier detection. Two plots are generated, showing classical and robust Mahalanobis distance versus the observation numbers.

Values

md	Values of the classical Mahalanobis distance
rd	Values of the robust Mahalanobis distance
cutoff	Value with the outlier cut-off

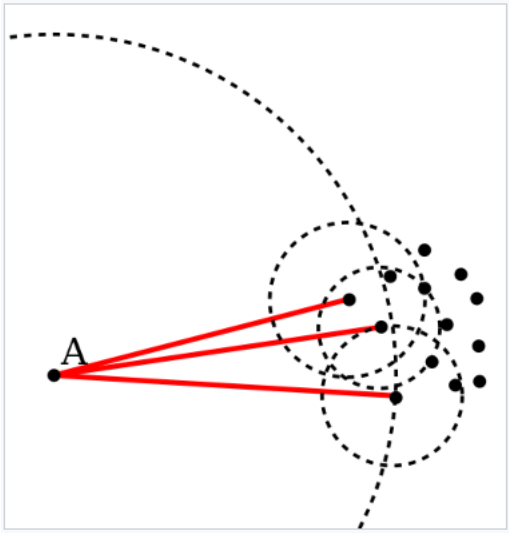
Non parametric detection of outliers

The LOF (Local Outlier Factor)

LOF is an algorithm for identifying density-based local outliers [Breunig et al., 2000]
https://dl.acm.org/doi/abs/10.1145/342009.335388

$$LOF_k(x) = \frac{\sum_{nei_x} \frac{\max dist_k(x)}{\max dist_k(nei_x^k)}}{k}$$

Comparison of the maxdist of the neighborhood of a point, respect to the maxdist of the neighborhood of the neighbors of the point (detection of outliers based on local density).



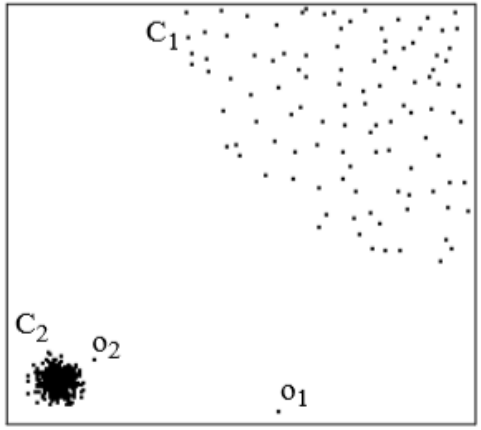
Basic idea of LOF: comparing the local density of a point with the densities of its neighbors. A has a much lower density than its neighbors.

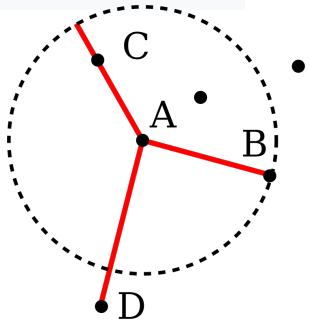
The outcome is an outlying value per individual. Values greater than 1 suggest outliers.

```
library(DMwR)

outlier.scores <- lofactor(data, k=5)
plot(density(outlier.scores))

# pick top 5 outliers
outliers <- order(outlier.scores, decreasing=T
# who are outliers
print(outliers)
```





Outliers with Random Forest

Algoritmo Isolation Tree

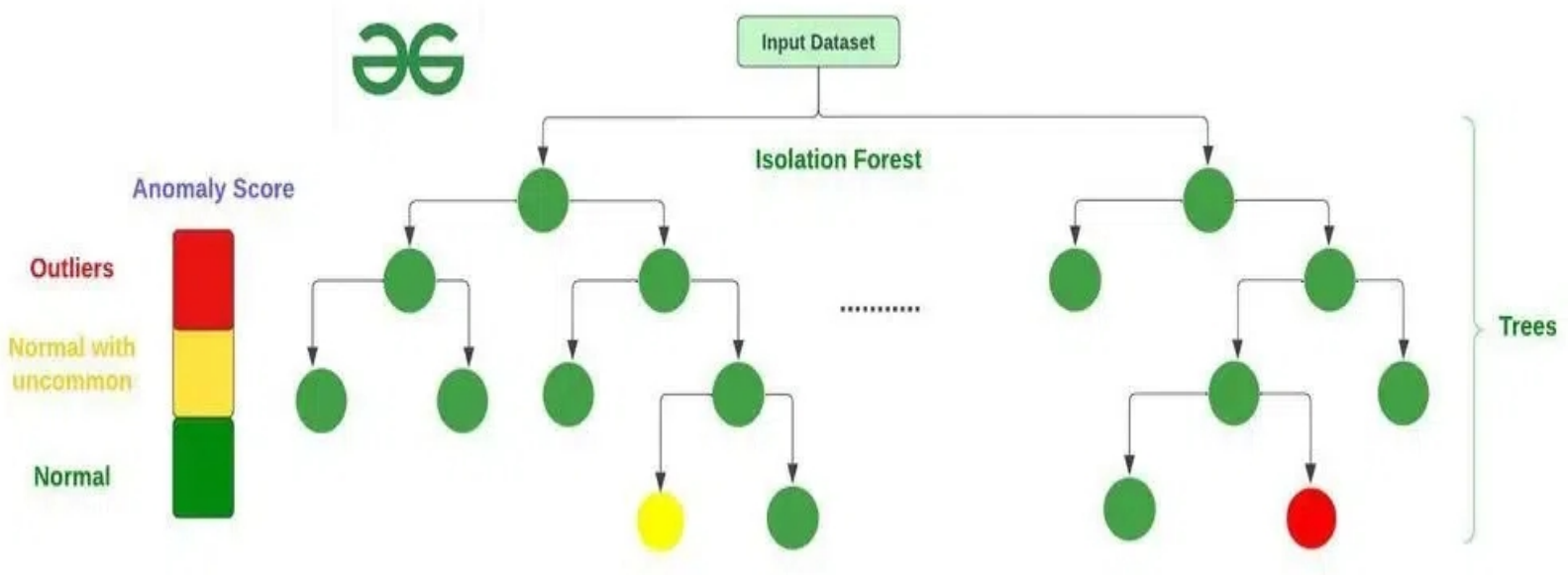
Crear un nodo raíz que contiene las N observaciones de entrenamiento.

Seleccionar aleatoriamente un atributo i y un valor aleatorio a dentro del rango observado de i

Las anomalías, al ser menos numerosas y estar más alejadas de la norma, suelen requerir menos divisiones para aislarlas, lo que las hace más fáciles de detectar.

Crear dos nuevos nodos separando las observaciones acorde al criterio $x_i \leq a$ o $x_i > a$

Repetir los pasos 2 y 3 hasta que todas las observaciones quedan aisladas de forma individual en nodos terminales.



PCA Approach - ENCODING

Reconstrucción

Una vez obtenido el PCA (matriz de eigenvectors, proyecciones y medias), la reconstrucción de las observaciones iniciales se puede obtener empleando la siguiente ecuación:

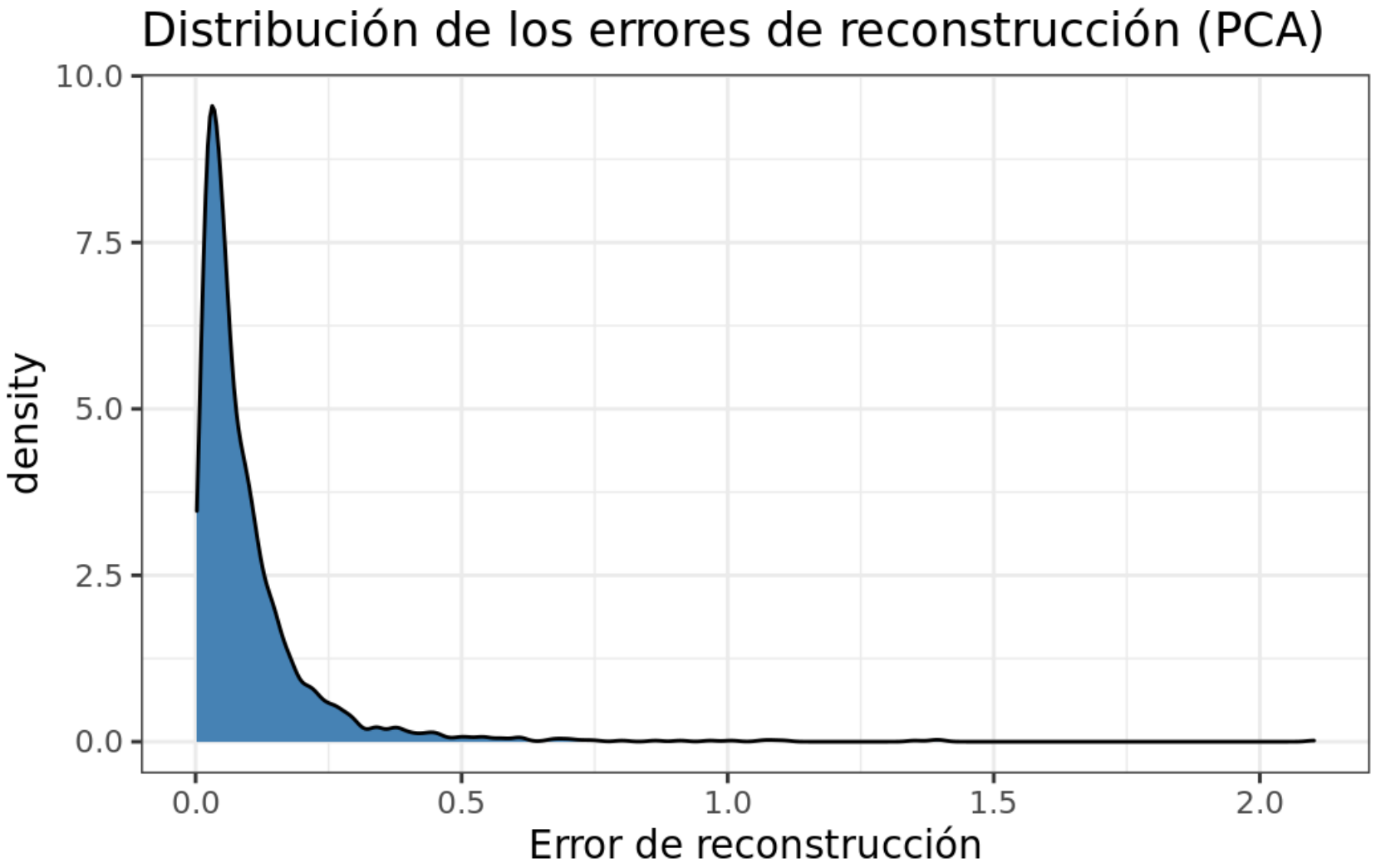
$$\textbf{reconstrucción} = \textbf{PC scores} \cdot \textbf{Eigenvectors}^{\top}$$

Es importante tener en cuenta que, si los datos han sido centrados y escalados (cosa que en términos generales debe hacerse al aplicar PCA) esta transformación debe revertirse.

Error de reconstrucción

El error cuadrático medio de reconstrucción de una observación se calcula como el promedio de las diferencias al cuadrado entre el valor original de sus variables y el valor reconstruido, es decir, el promedio de los errores de reconstrucción de todas sus variables elevados al cuadrado.

PCA Approach - ENCODING



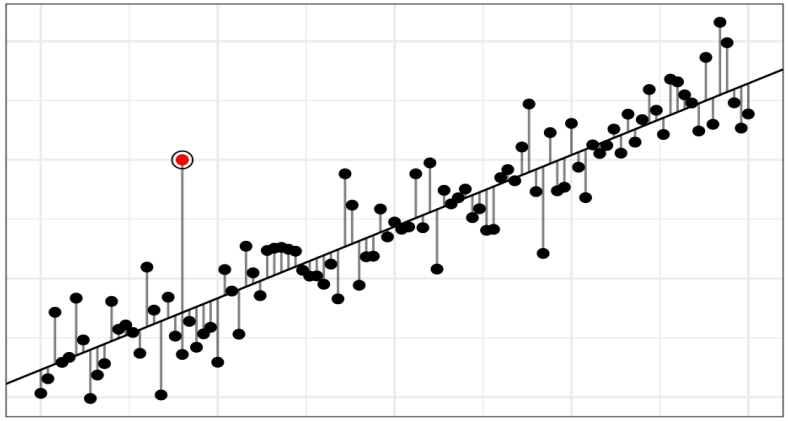
Trimmed K-means

https://cienciadedatos.net/documentos/64_deteccion_anomalias_trimmed_kmeans.html

Attribute wise learning for scoring outliers (ALSO)

Es un algoritmo no supervisado de detección de anomalías (outliers) cuando los datos no están etiquetados. Para cada variable disponible en el set de datos, se entrena un modelo de regresión que predice dicho atributo en función del resto de variables. El grado de anomalía de cada observación se calcula como el agregado del error cuadrático de los modelos al tratar de predecirla.

https://cienciadedatos.net/documentos/67_deteccion_anomalias_also



Practice of detecting outliers

- **Application of outlier detection**: Detecting "rare" events:
  - Fraud detection
  - Detecting network intrusion
  - Detecting changes in the behavior (sales, claims, connections, waiting time,…)

- Once we have detected outliers, what we should do?

1. **Eliminate them** (but we lose the information of the eliminated individuals) and deleting outliers is not the best solution, since outliers are recursive.

2. **Weight the individuals inversely to the outlying degree of individuals**, to diminish its importance (but statistical/learning methods would need to had implemented a weighing option of individuals).

3. Make a robust estimation of the parameters of the "normal generating mechanism", for instance **with a given percentage of the "central" individuals.**

4. Declare outliers as "**missing values**" and treat them as missing data.

Goal # 1 Getting Data Quality Reports
Goal # 2 Detecting and deciding about Outliers

- Automate the exploratory data analysis (EDA) to understand the data faster and easier. Examples:

  EDA Package in R --→ https://rpubs.com/zlzlzl2/749012

  In R, we can use these libraries:
- dataMaid
- DataExplorer
- SmartEDA

  In Python, we can use these libraries:

- ydata-profiling
- dtale
- sweetviz
- Autoviz

  https://medium.com/codex/automate-the-exploratory-data-analysis-eda-to-understand-the-data-faster-not-better-2ed6ff230eed