# Análise de dados em R

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

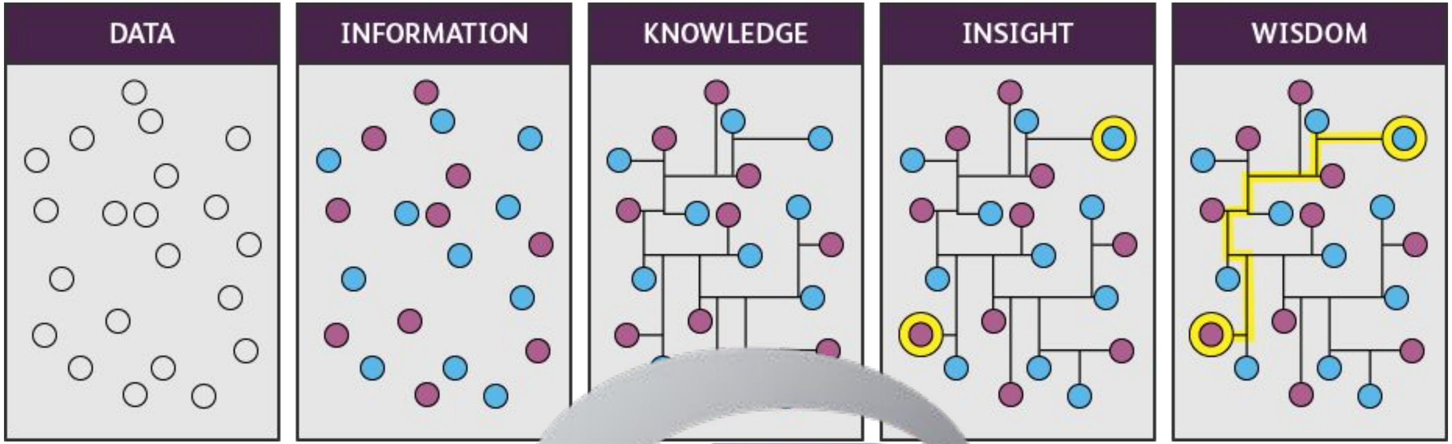
# Important

| | |
|---|---|
| Avaliação | Exame 80% + Projeto 20% |
| Datas-chave | Exame: 8 Julho<br>Apresentações projeto: 30 de Junho e 1 de Julho |
| Projeto | Grupos 2 ou 3 alunos<br>● A cada semana (após a 2ª semana) é definido uma tarefa para entregar |
| Horário | Sextas das 18h até 21h (intervalo 18h50 até 19h10)<br>Sábado das 10h até 13h (intervalo 10h50 até 11h10) |
| Formador | Pedro Abreu (pedabreu@gmail.com) |

# Agenda

- Introdução a análise de dados
- Introdução ao R e Posit
- Conceito básicos de R

# Data to Wisdom



DATA

INFORMATION

KNOWLEDGE

INSIGHT

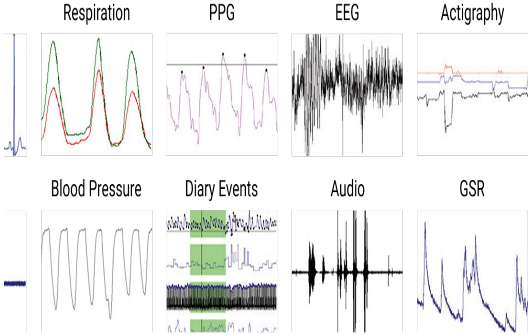
WISDOM

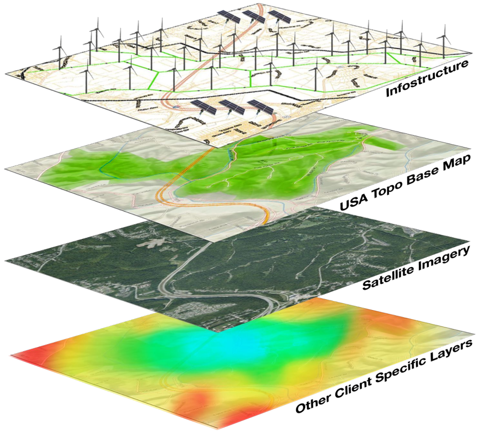
# Data Sources

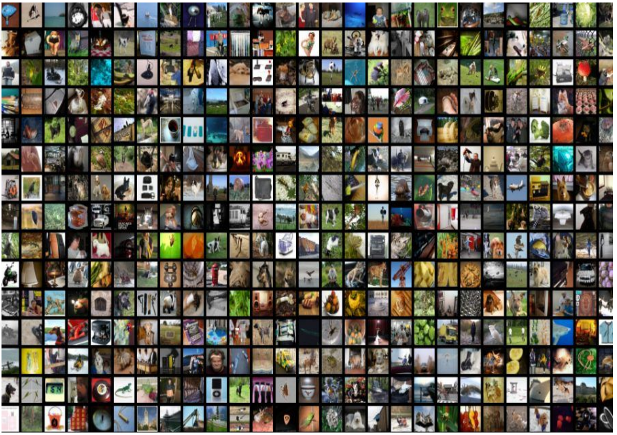
Sensores
Imagens
Sons
Texto
Grafos
Tabelas
Geográficos

|  | A | B | C | D |
|---|---|---|---|---|
| 1 | First Name | Last Name | Age | Salary |
| 2 | Jon | Smith | 36 | 26500 |
| 3 | Helen | Mirren | 22 | 21000 |
| 4 | David | Cameron | 29 | 39000 |
| 5 | Brad | Pitt | 52 | 45000 |
| 6 | Anna | Starolsky | 41 | 22500 |
| 7 | Peter | Piper | 20 | 31500 |
| 8 | David | Duck | 19 | 15700 |
| 9 | Julie | Walters | 33 | 19000 |

ist ein Blindtext. An ihm lässt sich
An abstract for this workshop
Abstract.
moving
RGB-D







# Data tasks and techniques

| Techniques\Tasks | Profiling | EDA | Predictive | Clustering |
|---|---|---|---|---|
| Descriptive Statistic | x | x | | |
| Correlation | | x | | |
| Hypothesis Tests | | x | | |
| Statistical Modelling | | x | x | x |
| Visualization | | x | | |

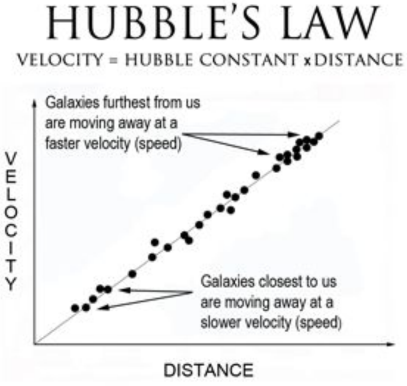
# Tasks - Profiling

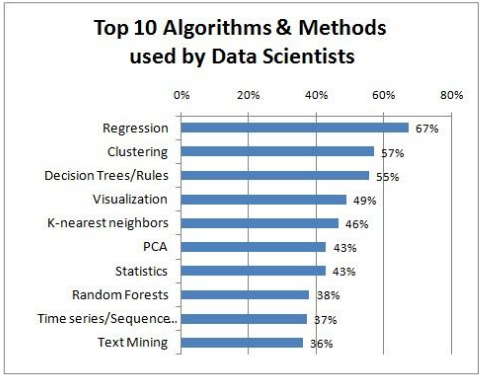
**Data profiling is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.**

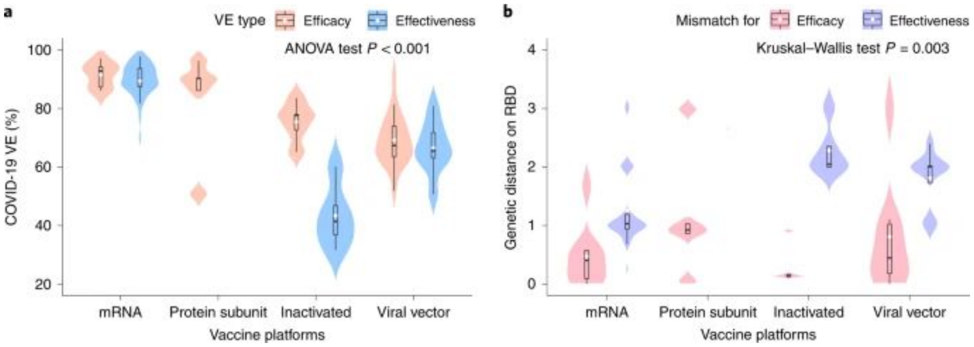
- What percentage of phone numbers do not have the correct number of digits.
- Are the blood type within the values A+, A-, etc
- Does the age field has data to use?
- Does the email as the @ symbol?

# Tasks - Exploratory Data Analysis (EDA)

**Exploratory data analysis (EDA) is used to analyze and investigate data. It helps determine how best to manipulate data sources to get the answers you need, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.**







# Tasks - Predictive

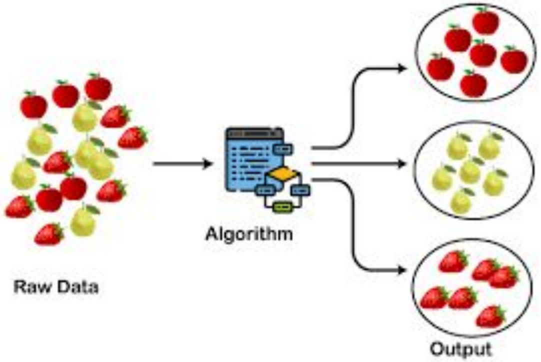
**Make predictions of  a dependent variable using other independent variables.**

- 💡 Forecast sales for next week
- 💡 Predict measure of progression in patients with Parkinson's disease
- 💡 Identifying the bird specie using bird sound
- 💡 Just go see more predictions tasks in https://www.kaggle.com/competitions

# Tasks - Clustering

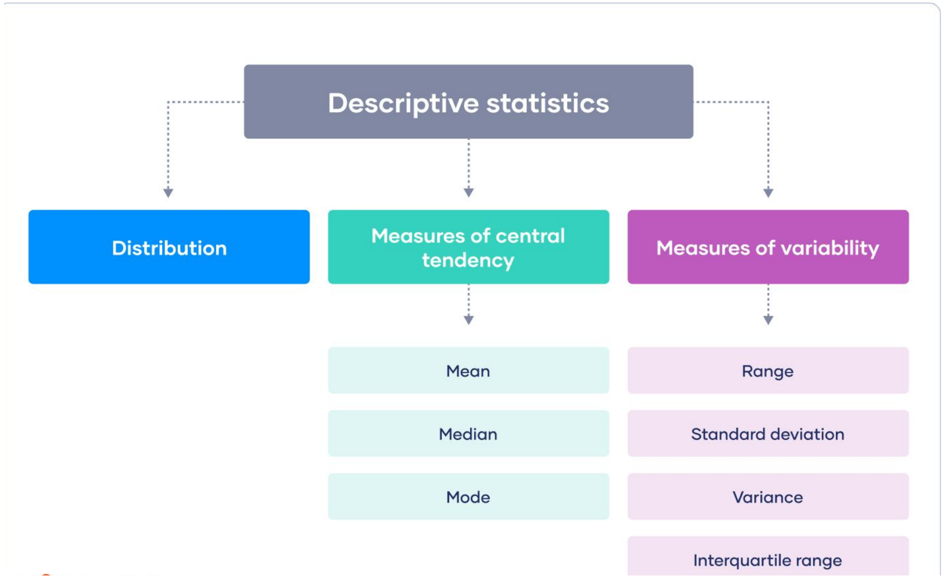
**Grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).**

- Differentiate between different types of tissue in a three-dimensional image for many different purposes
- Create profiles of typical television viewers



# Techniques - Descriptive Statistics

**Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.**



Descriptive statistics

Distribution

Measures of central tendency

Measures of variability

Mean

Median

Mode

Range

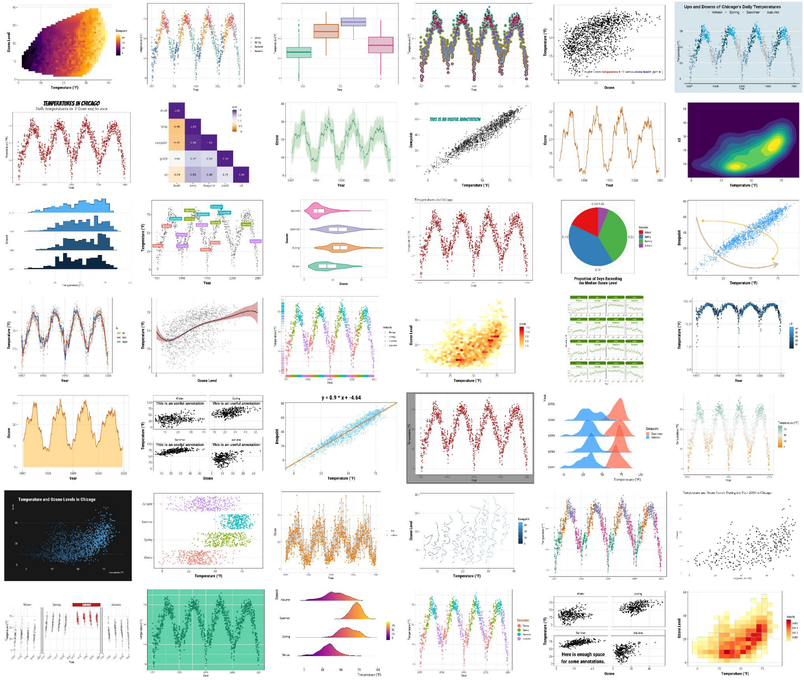
Standard deviation

Variance

Interquartile range

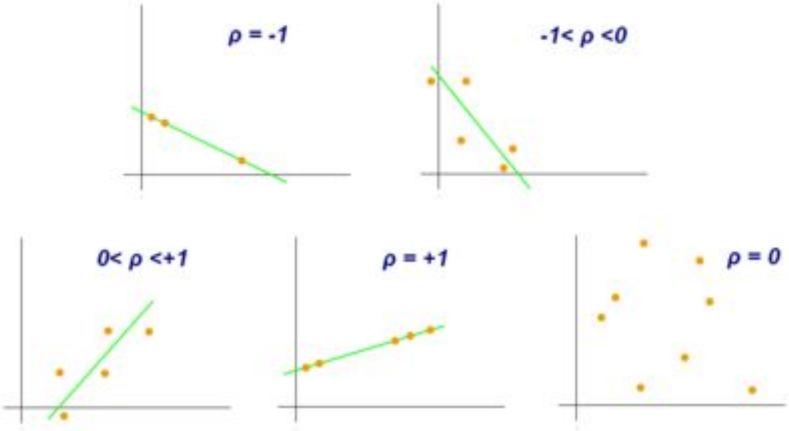
# Techniques - Visualization

**Visualization helps us to have an general view of all data in a single image. This allow us to find some relations and trends in data**



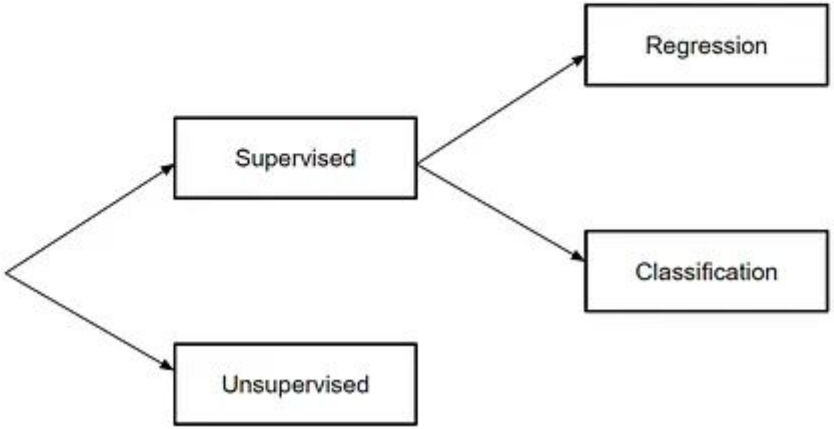
# Techniques - Correlation

**Correlation is a statistical measure that expresses the extent to which two variables are linearly related**



# Techniques - Statistical Modelling

**Relate one or more characteristics (independent variables) to another characteristic of a individual (dependent variables)**

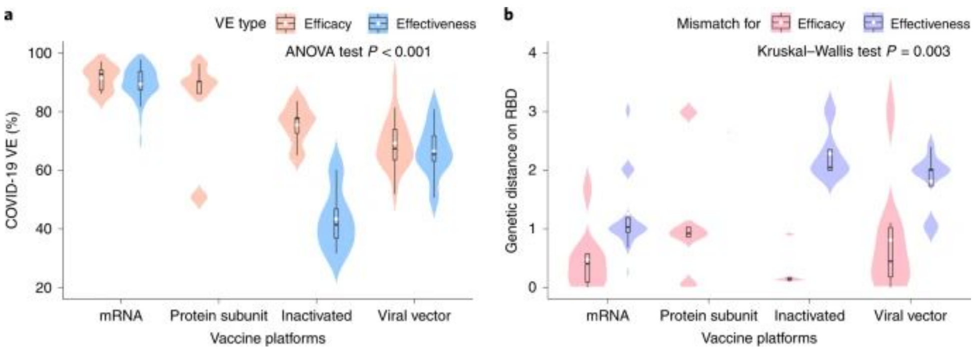


- 💡 Relate CO2 with Earth temperature
- 💡 Relate the characteristics of flowers to their species
- 💡 Group patients with similar symptoms or diseases

# Techniques - Hypothesis Tests

**Verify if data supports a hypothesis about the distribution or parameter**

💡 Does the drug A will provide advantages in the area of lower motor side effects and probably in improved negative symptom treatment?


a
VE type
Efficacy
Effectiveness
ANOVA test $P < 0.001$
COVID-19 VE (%)
mRNA
Protein subunit
Inactivated
Viral vector
Vaccine platforms
b
Mismatch for
Efficacy
Effectiveness
Kruskal–Wallis test $P = 0.003$
Genetic distance on RBD
mRNA
Protein subunit
Inactivated
Viral vector
Vaccine platforms