EI5IS102 Traitement de l'Information

Introduction

Charles Brazier
Postdoctoral researcher
Université de Bordeaux, CNRS, Bordeaux INP, LaBRI
France

Who am I?

Education

- MSc in Acoustics, Signal Processing, and Computer Science Applied to Music (IRCAM, Paris)
- PhD in Machine Learning and Music (Johannes Kepler University, Linz, Austria)
- Postdoc in Speech Recognition (Université de Bordeaux/LaBRI)

Project at LaBRI

- FVLLMONTI Project: Speech Emotion Recognition, Machine Translation
- Autonom Health: Automatic Sleepiness Detection, Speech Biomarkers

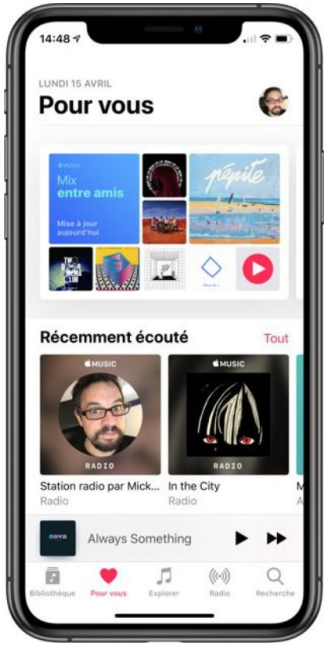
Personal data

- Office: LaBRI - Bureau 356
- E-mail: charles.brazier@u-bordeaux.fr

What is about: data

Data: a decision driver in almost every industry
→ How to best utilize data?

YOLOv10, 2024



Apple Music

What is about: Data Processing & Machine Learning

Data Processing

Automatic extraction of high level information from data (text/images/videos)

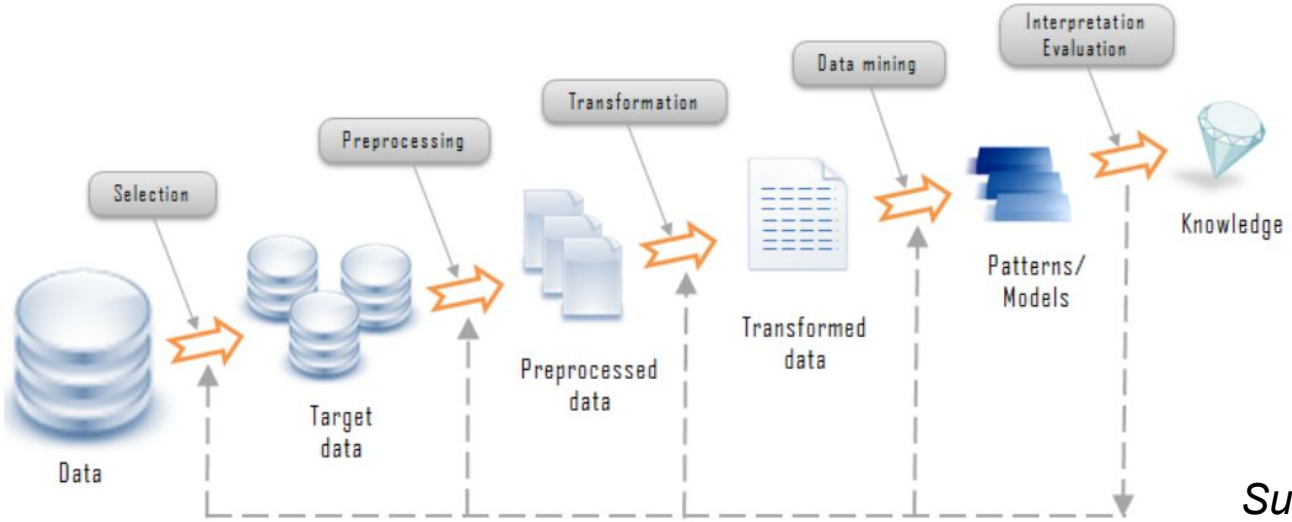
Introduction to **Machine Learning**

Learning from tons of (annotated) examples.

Motivation:

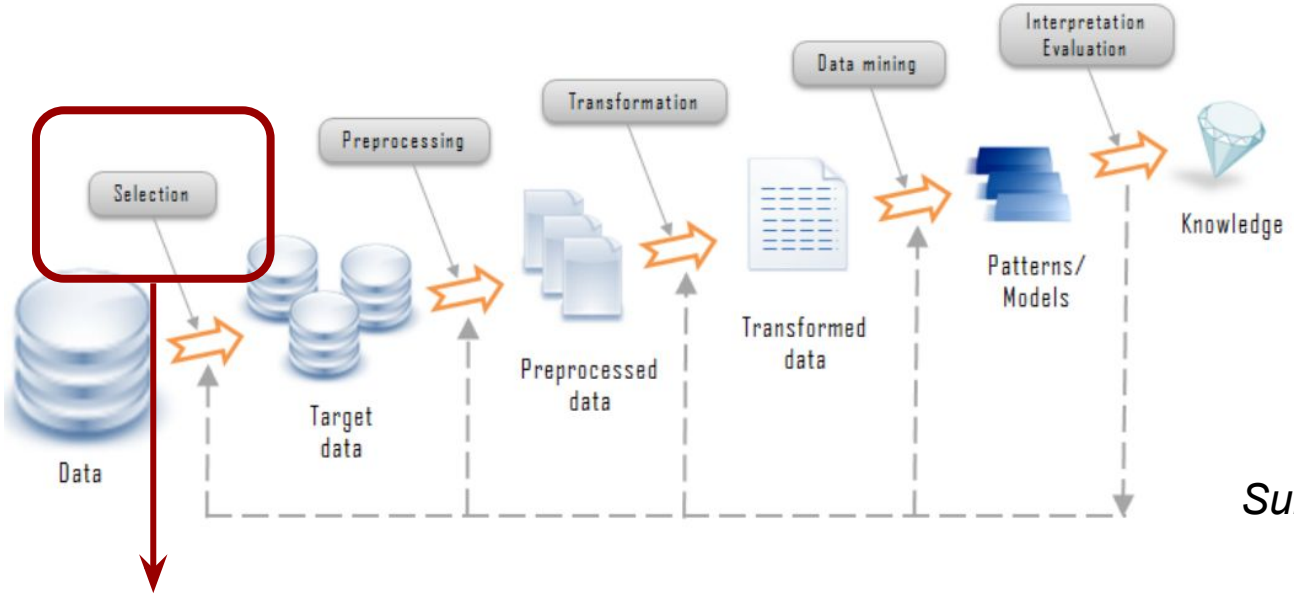
- Automatic data analysis: teaching a computer to listen, to see, to understand, etc.
- Expanding field: Natural Language Processing, Computer Vision, etc.
- Tons of applications: cultural, entertainment, environmental, fashion, medical, linguistic, robotic, sport, urbanism, etc.

The data analysis process



Sumiran, 2018

The data analysis process

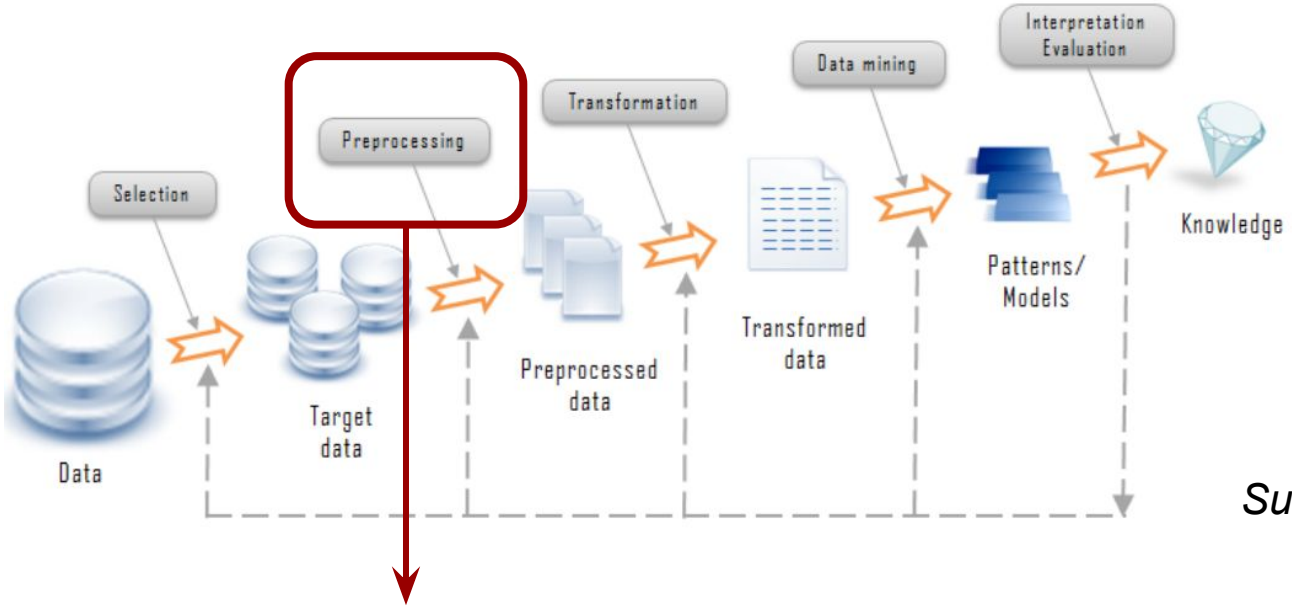


Sumiran, 2018

Data selection: choose relevant data for the analysis.

Example: user listening history, clinical questionnaires, sport game statistics, etc.

The data analysis process

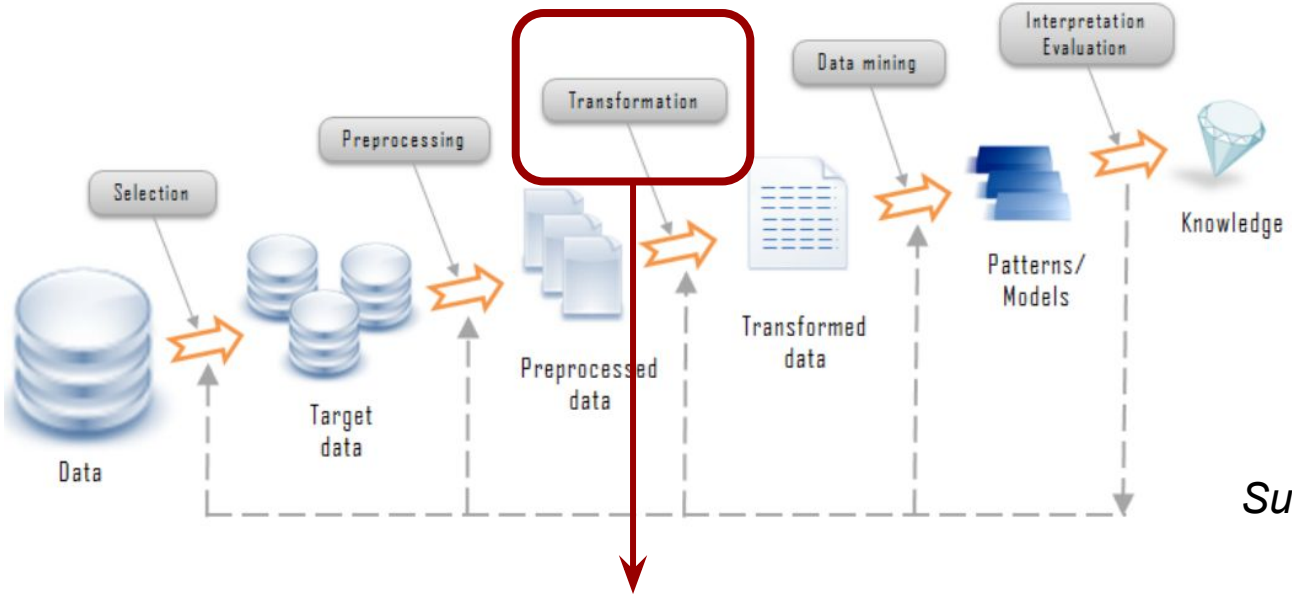


Sumiran, 2018

Data preprocessing: clean and prepare the data for analysis

Example: handling missing data, standardizing formats, removing duplicate entries, etc.

The data analysis process

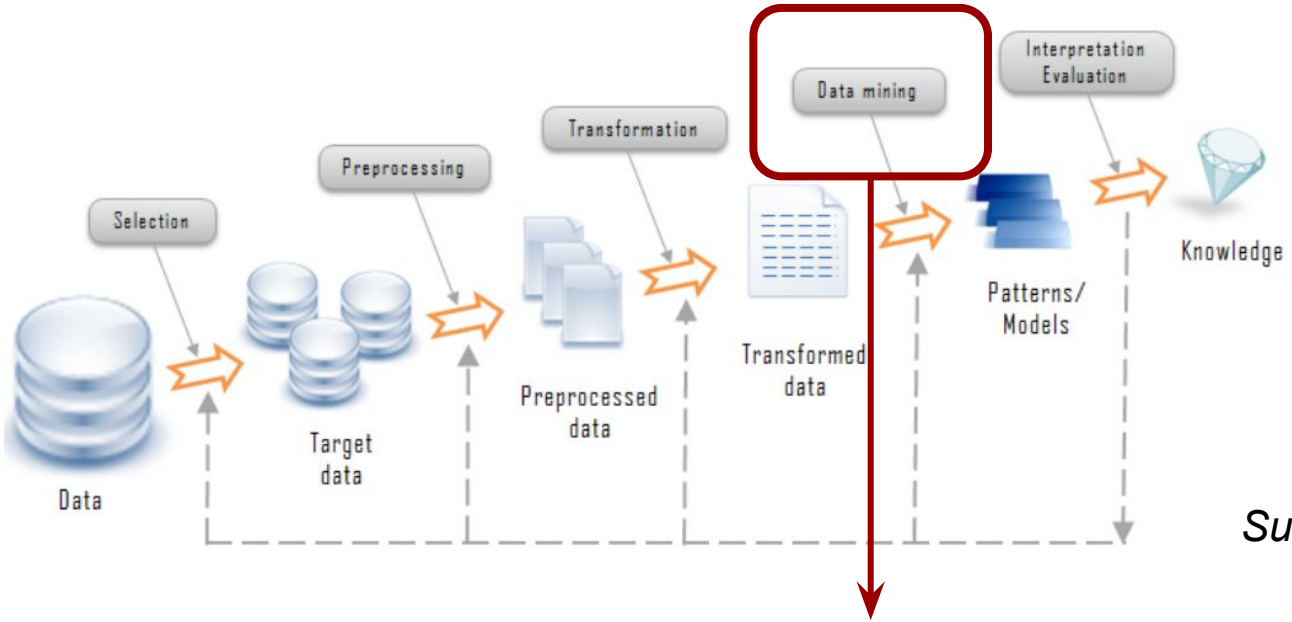


Sumiran, 2018

Data transformation: Transform data into a suitable format for analysis

Example: Converting data tables into numerical values, normalizing, smoothing, etc.

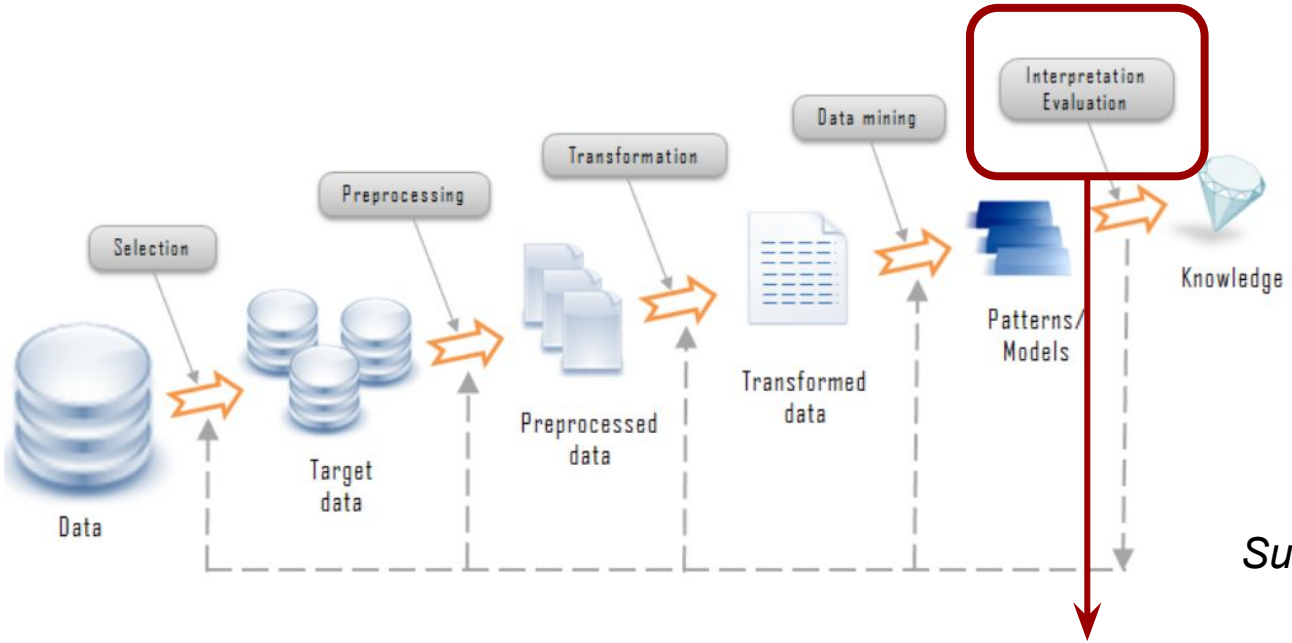
The data analysis process



Sumiran, 2018

Data mining: Apply mining techniques to uncover patterns and relationships in the data.
Example: Principal Component Analysis, Correspondence Analysis, Clustering, etc.

The data analysis process



Sumiran, 2018

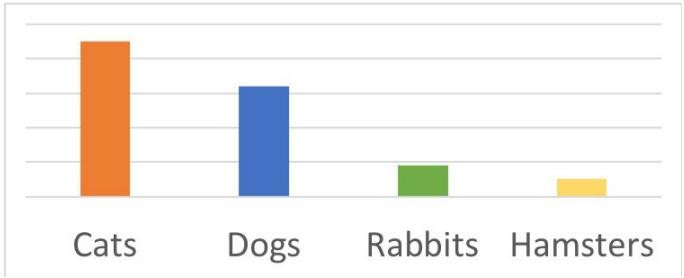
Interpretation & Evaluation: Evaluate the model and interpret the results

Example: Identify distinct musical taste segments or listening patterns to personalize reco.

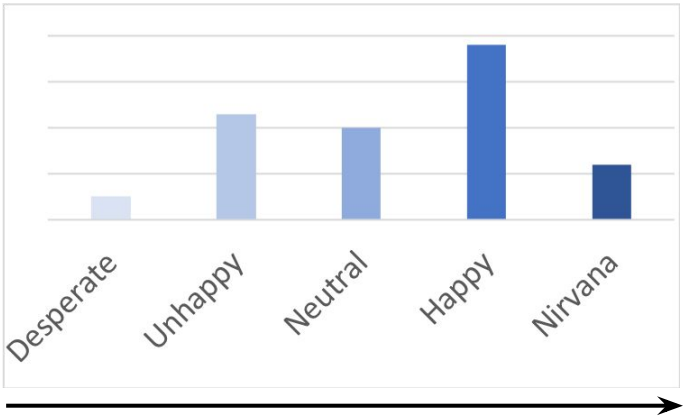
What kind of data?

Qualitative data: data describing categories

- **Nominal:** categories with no specific order



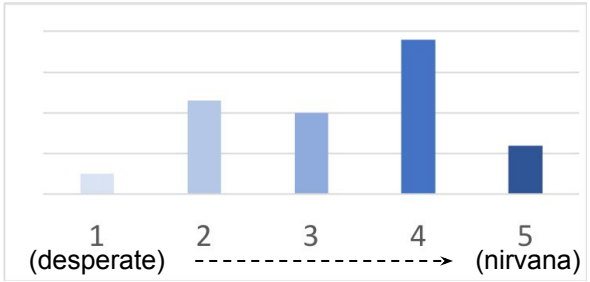
- **Ordinal:** ordered categories

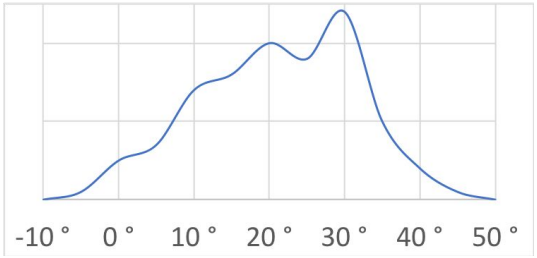


What kind of data?

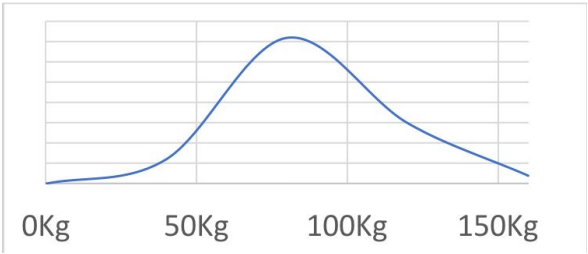
Quantitative data: data expressing quantities or numerical values

- **Interval:** numeric data with equal intervals (discrete or continuous)





- **Ratio:** same with “true zero”
  Zero means the absence of the quantity



What kind of data?

Multinomial data: data consisting of multiple categories with more than two possible values

Game Name	Platform	Year	Genre
F-Zero	NES	1990	Racing
Final Fantasy	NES	1990	RPG
Mario Kart: Double Dash	Wii	2003	Racing
Wii Sports	Wii	2006	Sports
Xenoblade Chronicles	Wii	2010	RPG
Tetris	GB	1989	Puzzle
Pokémon Gold	GB	2000	RPG
Mario Kart DS	DS	2005	Racing
Professor Layton	DS	2007	Puzzle
FIFA 20	DS	2020	Sports

Multidimensional data: data with multiple attributes, each represented as a dimension

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

What kind of data?

Multinomial data: data consisting of multiple categories with more than two possible values

Game Name	Platform	Year	Genre
F-Zero	NES	1990	Racing
Final Fantasy	NES	1990	RPG
Mario Kart: Double Dash	Wii	2003	Racing
Wii Sports	Wii	2006	Sports
Xenoblade Chronicles	Wii	2010	RPG
Tetris	GB	1989	Puzzle
Pokémon Gold	GB	2000	RPG
Mario Kart DS	DS	2005	Racing
Professor Layton	DS	2007	Puzzle
FIFA 20	DS	2020	Sports

Multidimensional data: data with multiple attributes, each represented as a dimension

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

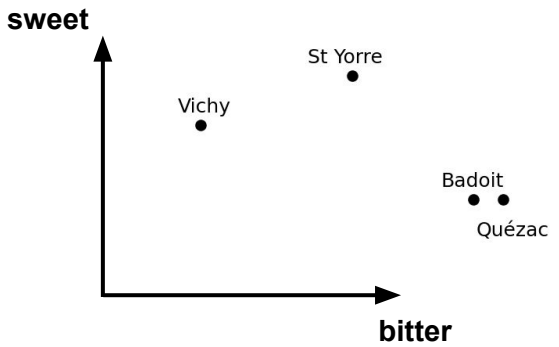
What kind of data?

Multinomial data: data consisting of multiple categories with more than two possible values

Game Name	Platform	Year	Genre
F-Zero	NES	1990	Racing
Final Fantasy	NES	1990	RPG
Mario Kart: Double Dash	Wii	2003	Racing
Wii Sports	Wii	2006	Sports
Xenoblade Chronicles	Wii	2010	RPG
Tetris	GB	1989	Puzzle
Pokémon Gold	GB	2000	RPG
Mario Kart DS	DS	2005	Racing
Professor Layton	DS	2007	Puzzle
FIFA 20	DS	2020	Sports

Multidimensional data: data with multiple attributes, each represented as a dimension

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3



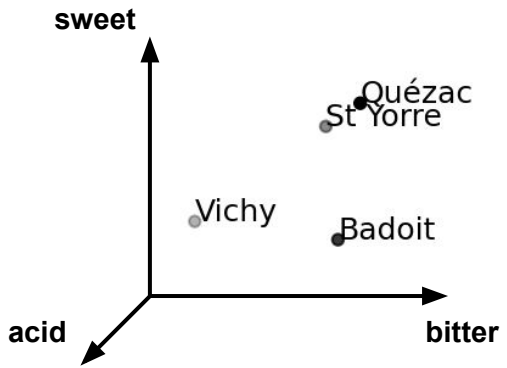
What kind of data?

Multinomial data: data consisting of multiple categories with more than two possible values

Game Name	Platform	Year	Genre
F-Zero	NES	1990	Racing
Final Fantasy	NES	1990	RPG
Mario Kart: Double Dash	Wii	2003	Racing
Wii Sports	Wii	2006	Sports
Xenoblade Chronicles	Wii	2010	RPG
Tetris	GB	1989	Puzzle
Pokémon Gold	GB	2000	RPG
Mario Kart DS	DS	2005	Racing
Professor Layton	DS	2007	Puzzle
FIFA 20	DS	2020	Sports

Multidimensional data: data with multiple attributes, each represented as a dimension

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3



What kind of analysis?

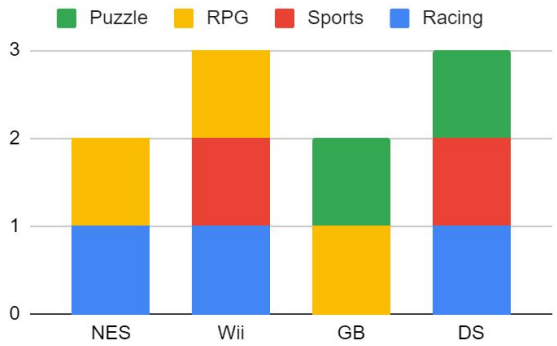
Multinomial analysis:

Game Name	Platform	Year	Genre
F-Zero	NES	1990	Racing
Final Fantasy	NES	1990	RPG
Mario Kart: Double Dash	Wii	2003	Racing
Wii Sports	Wii	2006	Sports
Xenoblade Chronicles	Wii	2010	RPG
Tetris	GB	1989	Puzzle
Pokémon Gold	GB	2000	RPG
Mario Kart DS	DS	2005	Racing
Professor Layton	DS	2007	Puzzle
FIFA 20	DS	2020	Sports

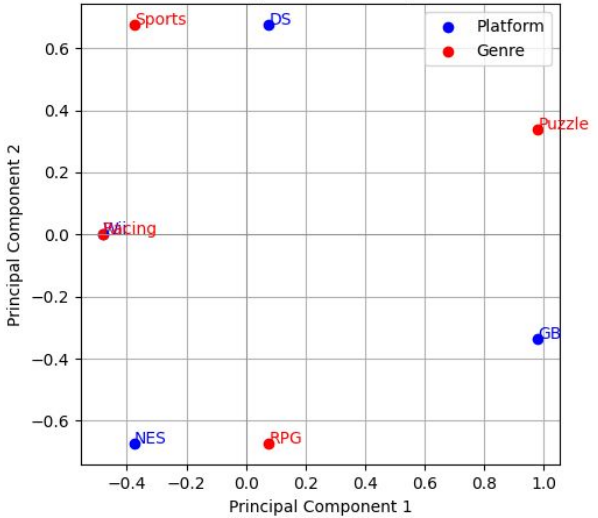
Correspondence Analysis

	Racing	Sports	RPG	Puzzle	Total
NES	1	0	1	0	2
Wii	1	1	1	0	3
GB	0	0	1	1	2
DS	1	1	0	1	3
Total	3	2	3	2	10

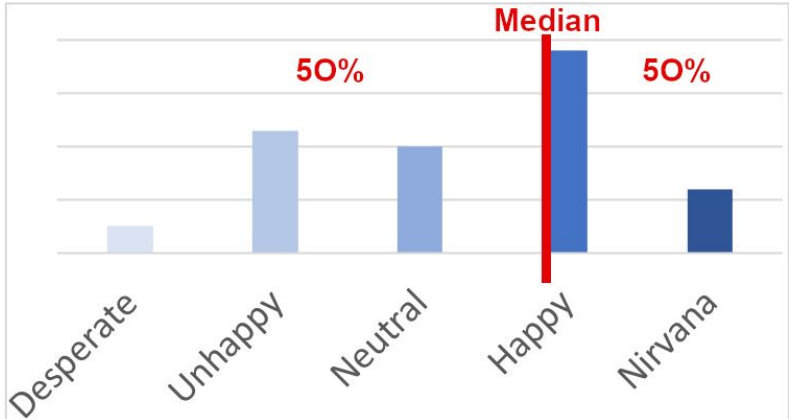
Contingency table



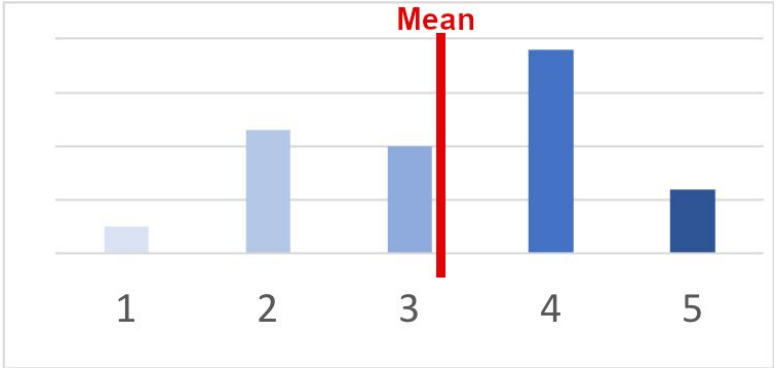
Stacked bar chart



What kind of analysis?



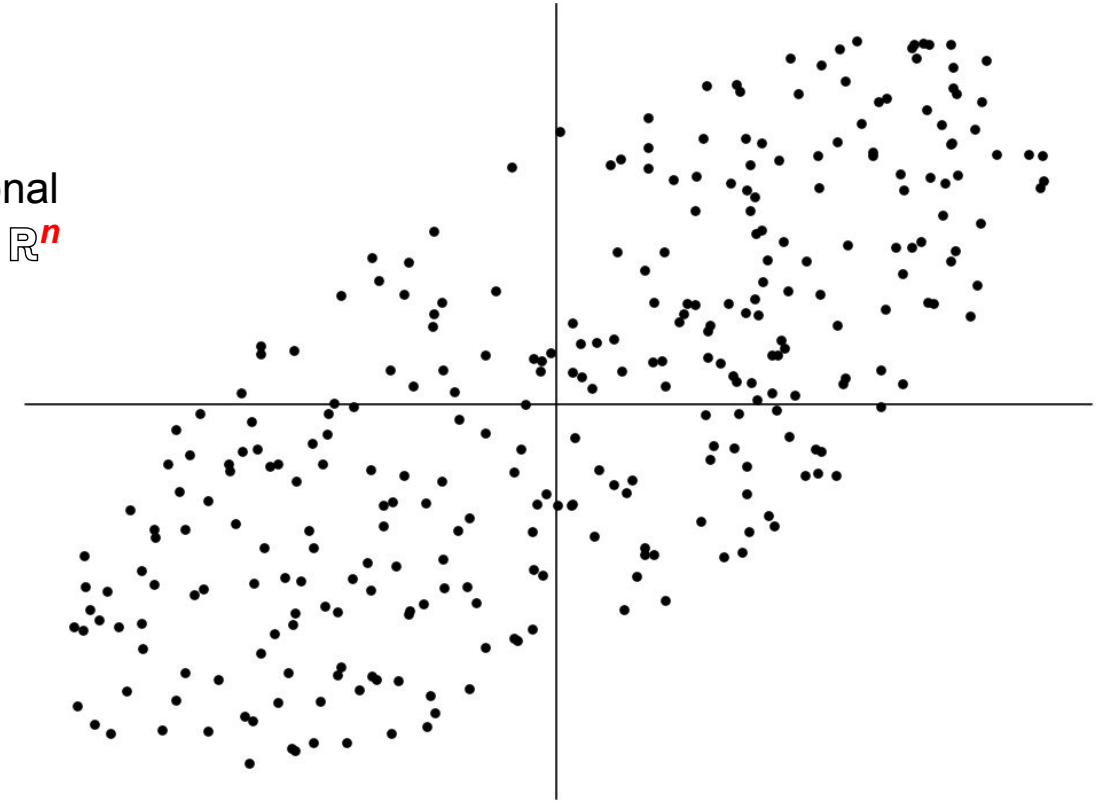
Ordinal data: median, percentiles, etc.



Interval data: mean, std, etc.

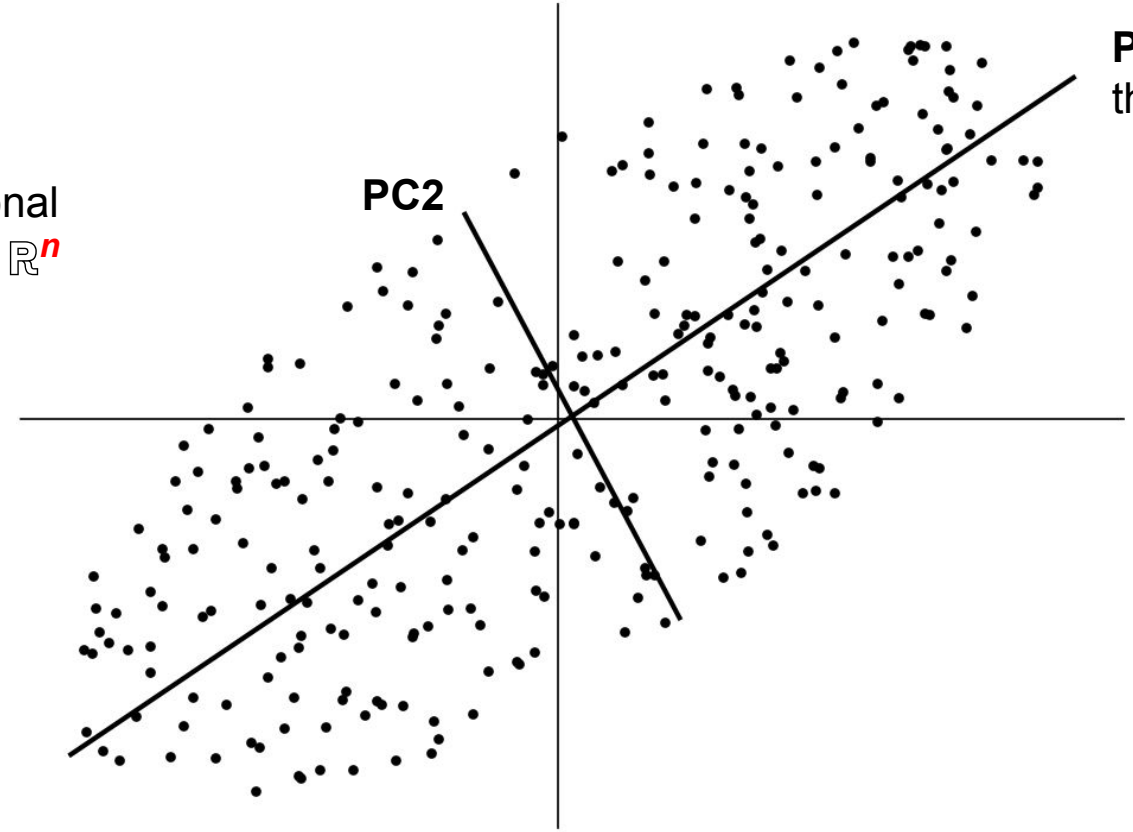
What kind of analysis?

Multidimensional data points in $\mathbb{R}^n$



What kind of analysis?

Multidimensional data points in $\mathbb{R}^n$

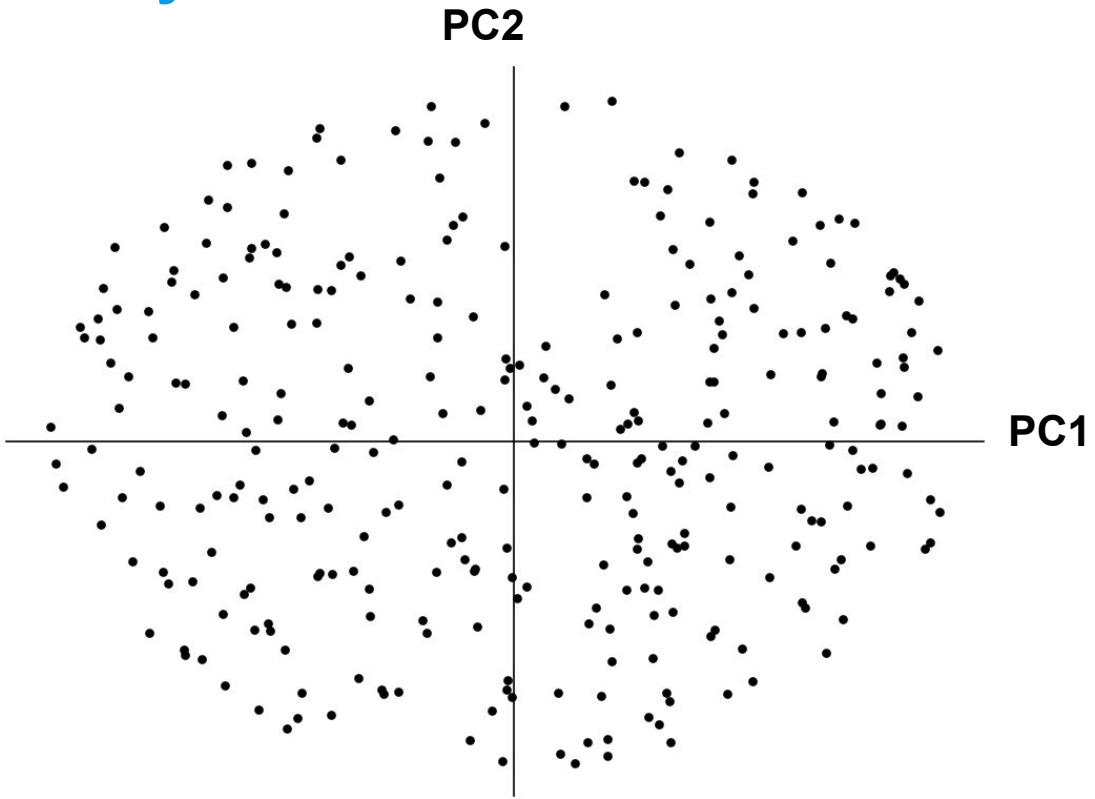


PC1: axis capturing the most variance

Principal Components

What kind of analysis?

Multidimensional data points in $\mathbb{R}^2$

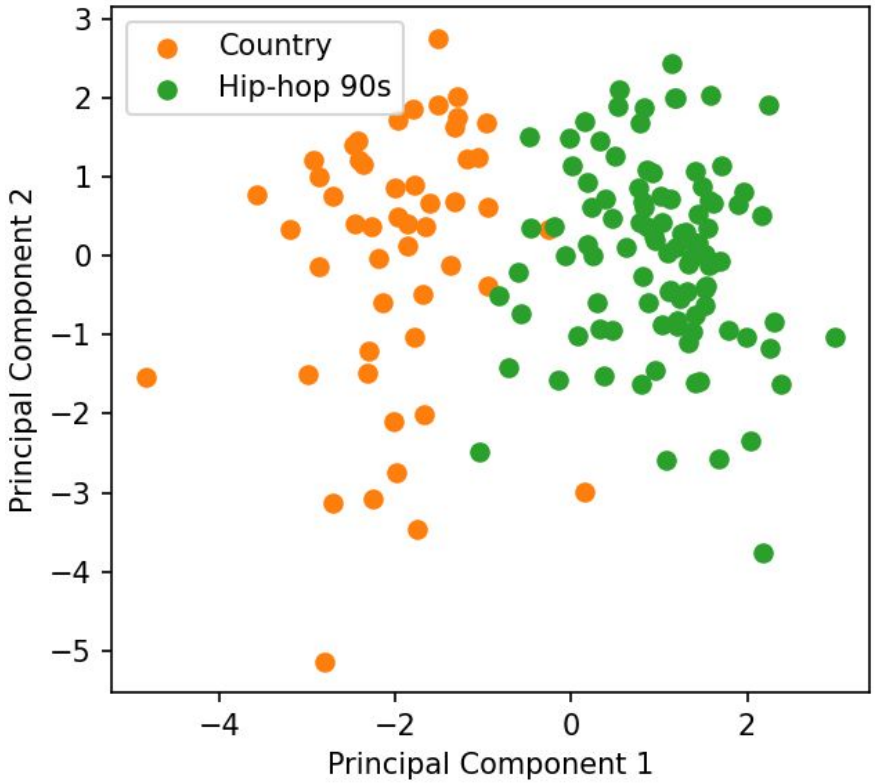


Dimension reduction

What kind of analysis?

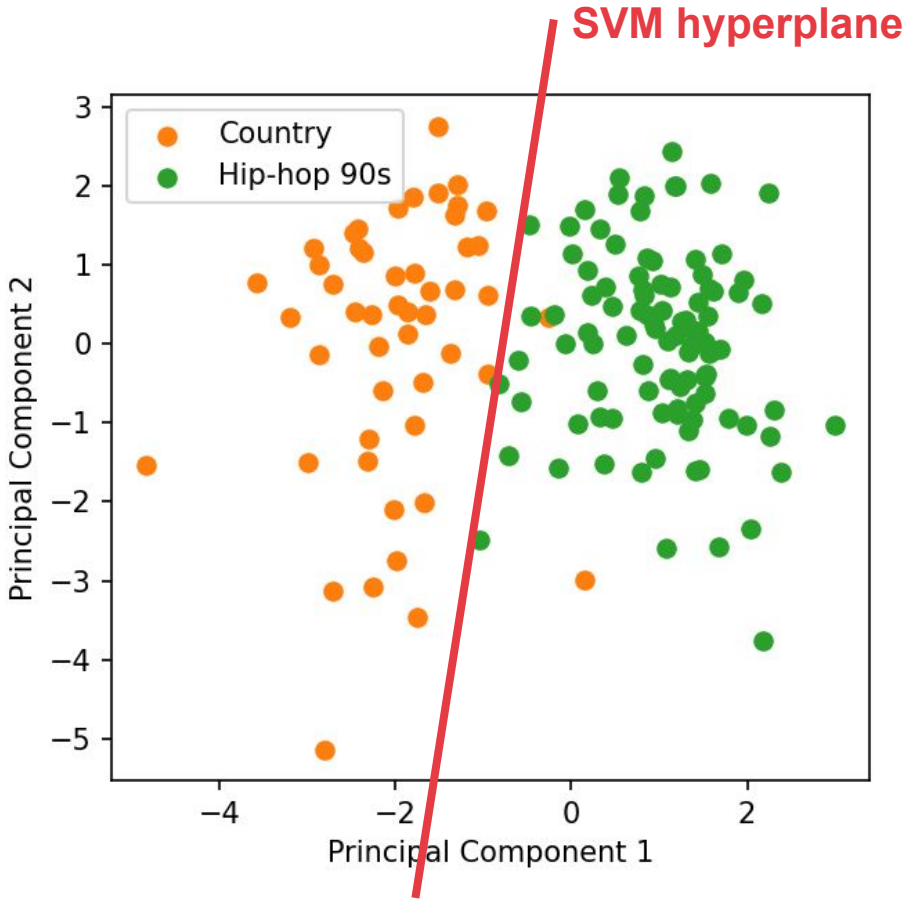
Reduction dimension:
Principal Component Analysis (PCA)

Example with music data



What kind of analysis?

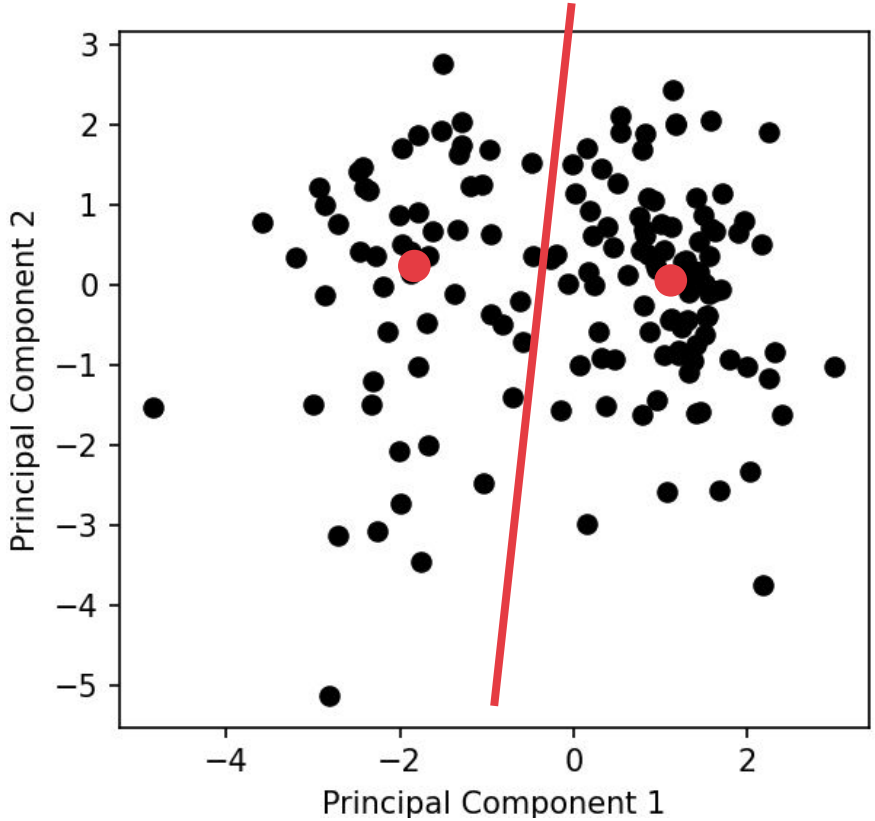
Supervised clustering:
Support Vector Machine (SVM)



What kind of analysis?

Unsupervised clustering:
K-means with 2 classes

Finding the centroids of the classes



Prerequisites

Basics on:

- Linear algebra (matrix/vector multiplication)
- Statistics (median, mean, variance)
- Python programming

Syllabus - engineering skills acquired:

- **Axis 1** Fundamentals
  - Knowledge and understanding of data science analysis tools
- **Axis 2** Tools
  - Capacity to choose and use adequate methods
- **Axis 5** Project management
  - Capacity of presenting efficiently solutions, summarizing results

Content

Introduction to data processing:

- Manipulation of data tables
- Extraction of information
- Dimension reduction

Data analysis methods:

- Quantitative data: PCA (Principal Component Analysis)
- Qualitative data: CA (Correspondence Analysis)

Basics on machine learning: Data clustering and classification

- Unsupervised learning: K-means
- Supervised learning: SVM (Support Vector Machine)

Assignment and project using Python (Numpy, scikit-learn)

How ?

Schedule:

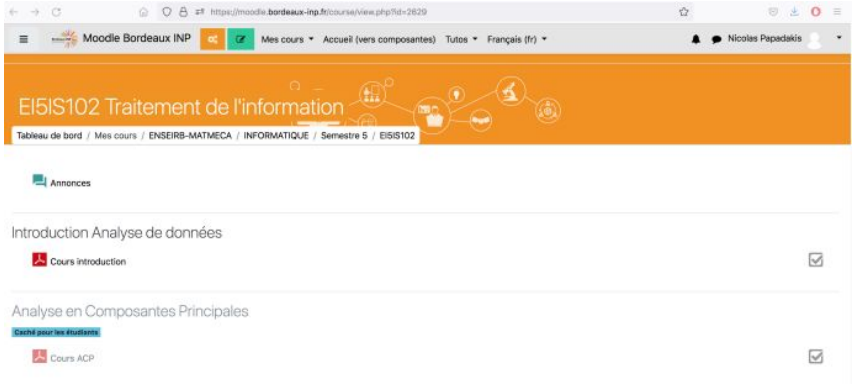
- **Lectures (1h20)** x3
- **Practice (2h40)** x3, 2 groups

Evaluation:

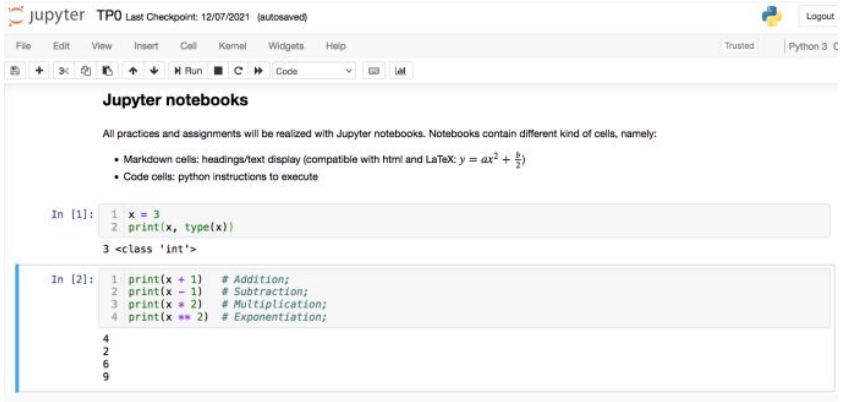
- **1 assignment** in Python (individual or groups of 2) …….. ≈50%
  → Manipulation of data and implementation of a simple data analysis method, notebook release
- **1 project** (groups of 2 or 3) on the application to real data …….. ≈20%
  → Understanding of data analysis method, write a scientific report
- **1 final quiz** (individual, ≈ 15 mins), no documents allowed …….. ≈30%
  → Check acquired notions

How ?

Moodle



Jupyter notebook



What assignment?

Assignment: Implement the **Principal Component Analysis (PCA)** method

- Carbon footprint data

Pays	Region	Terres cultivees	Paturages	Forets	Zones de peches	Terrains batis	Carbone	PIB/habitant (en millier de dollars)	Indice de developpement humain
Afghanistan	Middle East/Central Asia	0.3	0.2	0.1	0.0	0.0	0.2	0.56	0.509
Albania	Other Europe	0.5	0.2	0.2	0.0	0.0	0.9	5.05	0.792
Algeria	Africa	0.6	0.2	0.2	0.0	0.0	1.3	4.76	0.746
Angola	Africa	0.3	0.1	0.1	0.1	0.0	0.2	3.23	0.582
Argentina	South America	0.9	0.7	0.3	0.1	0.1	1.2	10.08	0.842

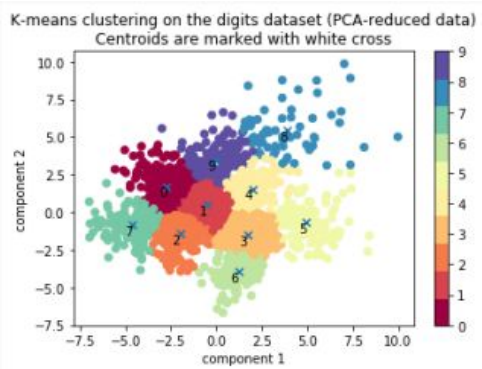
- Find and visualize correlations between modalities
  - Matrix representation
  - Line/column similarities
  - Eigenvalue problem

What project?

Project: Analyze a dataset using Python Machine Learning libraries: dimension reduction, study source of variability, clustering, etc.

Dataset

PCA



Clustering

Assignment and Project Deadlines

Calendar:

1. Lecture 1 – Principal Component Analysis ..... Nov. 6
2. Practice 1 – Start assignment ..... Nov. 13
3. Lecture 2 – Correspondence Analysis ..... Nov. 20
4. Practice 2 – Finish assignment ..... Nov. 27
5. Lecture 3 – Introduction to machine learning and project presentation ..... Dec. 4
6. Practice 3 – Work on project ..... Dec. 11

Deadlines:

1. Assignment (commented notebook) ..... Nov. 27
2. Online quiz ..... Dec. 11
3. Project report (pdf) ..... Dec

Questions?

Sources, images courtesy and acknowledgment:
N. Papadakis

Charles Brazier
charles.brazier@u-bordeaux.fr