Session 1:
Presentation of the Course

Statistical Modelling and Inference

Master in Data Science.

Prof. Lídia Montero and Josep Franquet

lidia.montero@upc.edu *josep.franquet@upc.edu*

FIB

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Departament d'Estadística i Investigació Operativa

Levels of corporate decision

senior management

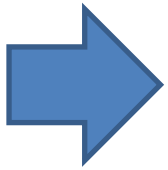
middle management
business areas

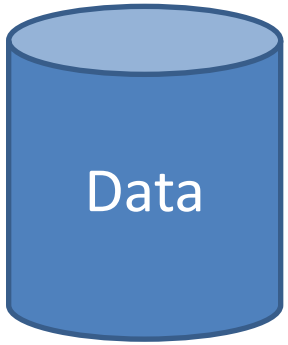
production and service workers
business processes

From Business Process to Data

Business processes are concrete workflows of material, information and knowledge – sets of activities. Each business process generates its own application (or part of it).

The output of the application is stored in DDBB (or files).





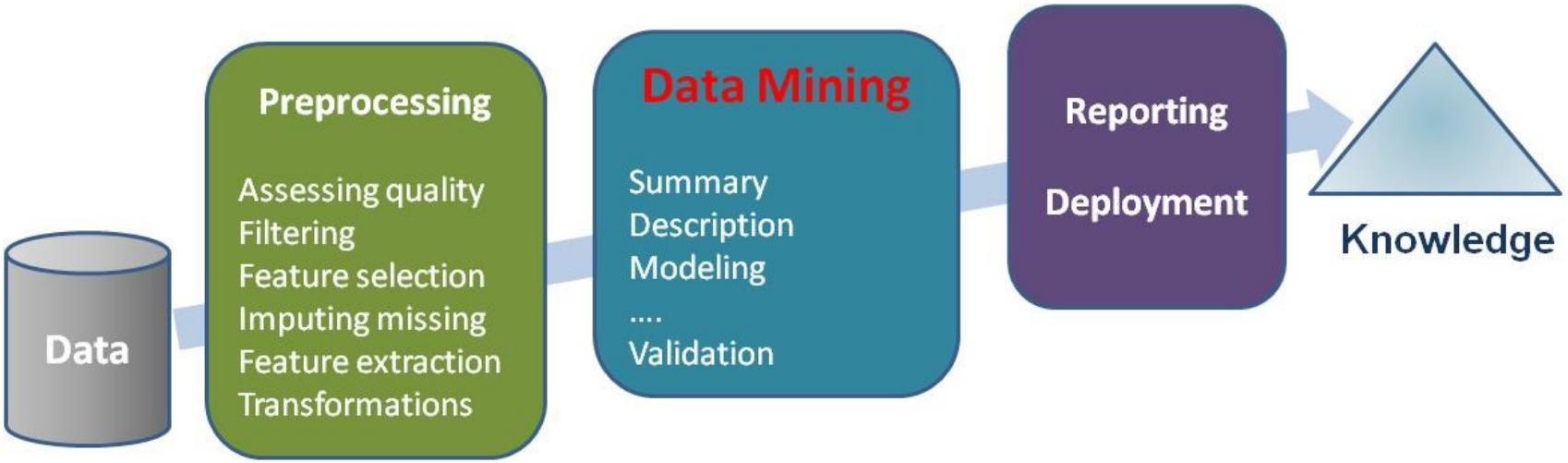
- All levels of management need to take decisions
- Many times in a “context of uncertainty”
- But supported by “experimental data”

The value of Data

Paradigm of the information era:

Data is the new driving force of businesses and governments.

Data is a key value for organizations



Program

Unit		Weeks
1	Classical vs Fisherian Inference	2
2	Quality of data. Profiling	1
3	Normal response linear models	3
4	Binary response linear models	3
5	Polytomous response linear models	2
6	Linear models for counting data	2
7	Design of Experiments	1

Planning of the course

MDS-SIM SCHEDULE. Course 2024-25. Term 1

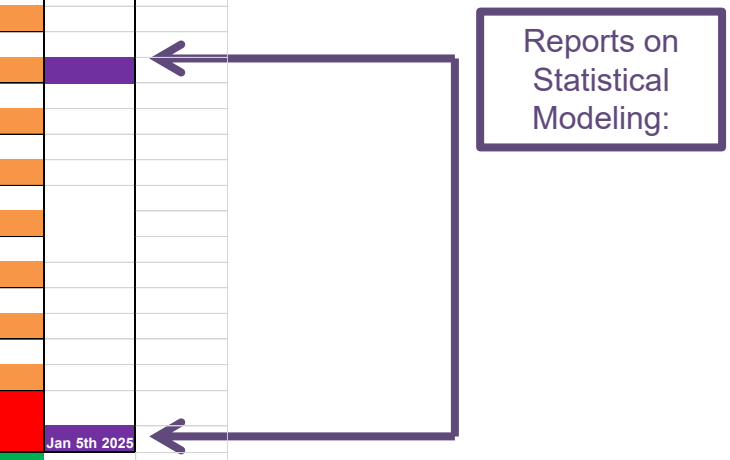
Week		Monday 15-17h (11) & 17-19h (12) Laboratory - A5S108 / 109	Wednesday 15-17h Theory - A6E01	Duedates
1	Presentation of Subject	9-set	11-set	
	Fest		National Fest	
2	Topic 1. Classical vs Fisherian inference	16-set	18-set	
3	Topic 1. Classical vs Fisherian inference	23-set	25-set	
		Local Fest		
4	Topic 1. Classical vs Fisherian inference	30-set	2-oct	
5	Topic 2. Data Quality	7-oct	9-oct	
	Topic 3. Normal Linear Models			
6	Topic 3. Normal Linear Models	14-oct	16-oct	
7	Topic 3. Normal Linear Models	21-oct	23-oct	
8	Topic 3. General Linear Model	28-oct	30-oct	
9	Partial Exams	4-nov	7-nov	
		Midterm Exam 4-nov 15:30		
10	Topic 4. Binary Outcome	11-nov	13-nov	
11	Topic 4. Binary Outcome	18-nov	20-nov	
12	Topic 4. Binary Outcome	25-nov	27-nov	
	Topic 5. Polytomous Outcome			
13	Topic 5. Polytomous Outcome	2-des	4-des	
14	Topic 5. Polytomous Outcome	9-des	11-des	
15	Topic 6. Models for Counts	16-des	18-des	
16	Christmas Holidays	23-des	6-gen	
				Jan 5th 2025
		Final Exam	8 Gen 2025 15h	

Lecturers
Lídia Montero Teaching
Josep Franquet Teaching
Midterm/Final Exams

Assessment:

Partial Exam	17%	Partial Exam (T1, 1/3) and Final Exam (T2, 2/3).
Final Exam	33%	T: Theory Note = Max (T2, (T1 + 2T2) / 3).
2 Practical Assignments	50%	FM=0.5T+0.5P

Assignments in groups of 2/3 - Exams individually
Marks: January 24th, 2024 (12h)



Case study – Assignment 1 : Cancer Mortality

- **The Cancer Mortality dataset is for use in data science education. There are 1831 observations in the train dataset and 1216 in the test dataset. The target variable is `target_deathrate`.**
- **It can be found on the data.world website (https://data.world/exercises/linear-regression-exercise-1.**

Practical Deliverables	Deadline
1 Report on Data cleaning, feature selection and profiling and numeric target modeling (D1) (limited to 40 pages)	**Over the course**
1 Report on Data cleaning, feature selection and profiling and categorical target modeling (D2) (limited to 40 pages)	Over the course

Case study – Assignment 2 : Airline Satisfaction

- **The assignment uses data from https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.**
- **The aim is to develop a binary regression model to predict behavior of customers. The raw data contains 5000 rows (customers) and 25 columns (features). Target variable is satisfaction.**

Practical Deliverables	Deadline
1 Report on Data cleaning, feature selection and profiling and numeric target modeling (D1) (limited to 40 pages)	Over the course
1 Report on Data cleaning, feature selection and profiling and categorical target modeling (D2) (limited to 40 pages)	**Over the course**

Evaluation

The evaluation of the course integrates the three phases of learning process: knowledge, skills and competencies.

- **The knowledge is assessed by two exams**, *in the middle and last week of the course. Partial (T1, 1/3) and Final Exam (T2, 2/3). (score T).*
- **The skills assessed from several deliverables (2) related to the course**. *Each of the blocks involve a practice that students will perform by groups of 2/3 (Score P, average) and should be posted on Atenea tasks.*
- **Final Mark for Theory:** T: Theory Note = Max (T2, (T1 + 2T2) / 3).
- **The final grade will obtained weighing the two scores: Final Mark = 0.5P + 0.5T.**
- **You have to get T > 3.5 otherwise Final Mark = T.**

Software

- The software to be used during the course will be R and RStudio.
- Each block will use its specific packages and functions.
- ***cran.r-project.org/***
- https://www.r-project.org/nosvn/conferences/useR-2013/Tutorials/Kuhn.html
- **A Complete Tutorial to learn Data Science in R from Scratch**
- https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/

Recommended books

✓Fox, J. ***Applied Regression Analysis and Generalized Linear Models***. Sage Publications, Edition 2015.

✓Fox and Weisberg ***An R Companion to Applied Regression***. Sage Publications, Edition 2011.

✓Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. URL: https://data.princeton.edu/wws509/notes/

✓Wickham, H. ***ggplot2: Elegant Graphics for Data Analysis***. Springer New York, 2009.

✓Montgomery, Douglas ***, Design and Analysis of Experiments*** , Wiley , 2020 , ISBN:1119722106.

✓Box, George E. P; Hunter, J. Stuart; Hunter, William Gordon , Statistics for experimenters : design, innovation, and discovery , John Wiley & Sons , cop. 2005 , ISBN:0471718130

✓Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome , The Elements of statistical learning : data mining, inference, and prediction , Springer , cop. 2009 , ISBN:0387848576.

✓Trivedi, K.S, , Probability and statistics with reliability, queuing and computer science applications , John Wiley and Sons , 2016 , ISBN:1119285429.