

Chapter 1

Course plan

In the following, I present the course plan for PO-235 and CMC-16.

Any questions about the classes should be sent via Google Classroom. If your question is of general interest, please use the main stream. If your question is personal and about a specific assignment or grade, please use the private stream.

PO-235 Data science project

Course plan (2024)

Prof. Filipe A. N. Verri

Important: Only graduate students are allowed to take this course.

Number of students: Approx. 20

Course load: 3-0-0-4

Requirements: Advanced programming skills, strong statistics background, and beginner-level machine-learning skills.

Course program: Brief history of data science. Fundamental data concepts. Stages in a data science project. Data Infrastructure. Data integration from multiple sources. Data engineering and shaping. Inductive learning and principles of statistical learning theory. Application of machine learning models in real-world problems. Experimental planning for data science. Model evaluation and Bayesian analysis. Documentation and deployment. Ethical and legal issues in data science. Privacy-preserving computational approaches.

Goals: Providing the theoretical background and the practical concepts to develop an end-to-end data science project for an inductive task.

Teaching methodology: Expository classes in common classroom, using whiteboard, slide presentations, coding examples, books and scientific papers. Supplementary didactic materials will be available in Google Classroom. The development of the case study will happen during home study hours, including programming and scientific paper writing. All classes will be given in English. Students are encouraged to ask questions in English, but Portuguese is also allowed. All written and oral assignments must be in English.

Grading: Two individual written tests in the 1st quarter (T_1 and T_2) and another in the 2nd quarter (T_3). Also, a group activity that includes writing a scientific paper (optional), developing a data science product, and a 30 minutes presentation (L).

Final grades will be calculated as

$$\text{1st Q} = \sqrt{T_1 T_2}, \quad \text{2nd Q} = \sqrt{T_3 L}, \quad \text{Exam} = L.$$

Case study: At most 6 groups will be formed. Each group will be responsible for a case study. Students must choose a real-world problem and develop a data science project, including data collection, data handling, inductive learning, validation, documentation, and deployment. The results must be presented in a 30 minutes presentation. Extra points will be given to groups that write a scientific paper about the case study. The trained models must be incorporated in a data science product, such as a web application, a mobile application, or a web service.

Bibliography:

- Nina Zumel and John Mount (2019). *Practical Data Science with R*. 2nd ed. Manning.
- Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Golemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd ed. O'Reilly Media.
- John D. Kelleher and Brendan Tierney (2018). *Data science*. The MIT Press.

The first two books — Zumel and Mount (2019) and Wickham, Çetinkaya-Rundel, and Golemund (2023) — are available online for free.

Any required extra material will be made available in Google Classroom.

Calendar: The expected schedule is presented below.

1st Quarter	
Week	Topics
1	Brief history of data science Preliminaries
2	Written test
3	Fundamental data concepts
4	Stages in a data science project
5	Statistical learning theory
6	
6	Data handling
7	
8	Written test
2nd Quarter	
Week	Topics
1	Evaluation
2	Experimental planning
3	Ethics, privacy and legal issues
4	Written test
5	Project discussion
6	
7	Presentations
8	

Case studies will be presented during exam weeks. At most 3 case studies will be presented per day, with 30 minutes for each presentation and 20 minutes for questions.

CMC-16 Data science practices

Course plan (2024)

Prof. Filipe A. N. Verri

Important: Only ITA's undergraduate students are allowed to take this course.

Number of students: Approx. 20 (no more than 40 students)

Course load: 2-0-1-5

Requirements: CMC-13 or CMC-15

Course program: Brief history of Data Science. Stages in a Data Science project. Tidy Data. Data integration from multiple sources. Data engineering and shaping. Inductive learning and statistical learning theory. Experimental planning for Data Science. Model evaluation and Bayesian Analysis. Documentation and deployment. Privacy-preserving computational approaches.

Goals: Further studying the practical aspects of Data Science (in relation to CMC-13) and providing the mathematical foundations to ensure the correct usage of Data Science techniques.

The specific goals are:

- Understanding the steps and people involved in a Data Science project;
- Developing an end-to-end case study, including data collection, data transformation, inductive learning, validation, documentation, and deployment; and
- Critically evaluate the results and implications of the case study.

Teaching methodology: Expository classes in common classroom, using whiteboard, slide presentations, coding examples, books and scientific papers. Supplementary didactic materials will be available in Google Classroom. The development of the case study will happen during laboratory classes and home study hours.

Grading: One individual written test in the 1st and another in the 2nd quarter. Software development and oral presentation about the case study (in groups) for the final exam.

Case study: At most 6 groups will be formed. Each group will be responsible for a case study. Students must choose a real-world problem and develop a data science project, including data collection, data handling, inductive learning, validation, documentation, and deployment. The results must be presented in a 30 minutes presentation. The trained models must be incorporated in a data science product, such as a web application, a mobile application, or a web service.

Bibliography:

- Nina Zumel and John Mount (2019). *Practical Data Science with R*. 2nd ed. Manning.
- Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd ed. O'Reilly Media.
- John D. Kelleher and Brendan Tierney (2018). *Data science*. The MIT Press.

The first two books — Zumel and Mount (2019) and Wickham, Çetinkaya-Rundel, and Grolemund (2023) — are available online for free.

Any extra material will be made available in Google Classroom.

Calendar: The expected schedule is presented below.

1st Quarter	
Week	Topics
1	Brief history of Data Science
2	Stages in a Data Science project
3	Tidy Data and data integration from multiple sources
4	Data engineering and shaping
5	Inductive learning and statistical learning theory
6	
7	Case study discussion and definitions
8	Written test

2nd Quarter	
Week	Topics
1	Model evaluation
2	Experimental planning for Data Science
3	Bayesian Analysis
4	Documentation and deployment
5	Privacy-preserving computational approaches
6	Written test
7	
8	Presentations and discussions
