DATA MINING 2
Course Overview

Riccardo Guidotti

Università di Pisa

Teachers

- **Riccardo Guidotti**
  - Computer Science Department
  - Email: riccardo.guidotti@unipi.it

- **Alessio Cascione (Assistant)**
  - Computer Science Department
  - Email: alessio.cascione@phd.unipi.it

Classes

- Classes
  - Monday, 9-11, Room C
  - Wednesday, 11-13, Room C

- Office Hours
  - Riccardo Guidotti's office
  - Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

- Teaching Assistant
  - Alessio Cascione [DM2 Meeting] at alessio.cascione@phd.unipi.it

No Classes and Recovery Classes

No Class

- Mon 06/04/2026 (Easter Monday)
- Thu 30/04/2026 ? (in Denmark)

Recovery Classes

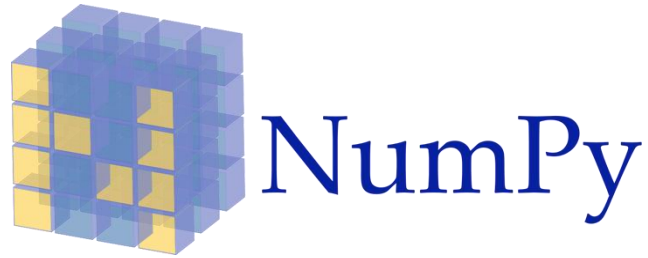
- Tue 06/05/2026 ? (Room C1) 9-11 (change with SMDS)

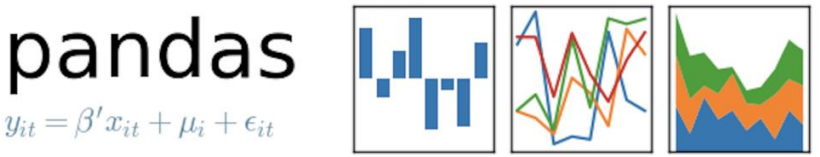
Topics

- **Module 1: Advanced Data-Preprocessing**
  - Imbalanced Learning
  - Dimensionality Reduction
  - Anomaly Detection
- **Module 2: Advanced ML & XAI**
  - Logistic Regression
  - Support Vector Machines
  - Neural Networks
  - Ensemble Methods
  - Gradient Boosting
  - Explainable AI
- **Module 3: Time Series Analysis**
  - Time Series Similarity
  - Approximation
  - Motif, Shapelets
  - Classification, Clustering
- **Module 4: Transactional Data**
  - Sequential Pattern Mining
  - Transactional Clustering

Laboratory

- Python
- Jupyter Notebook

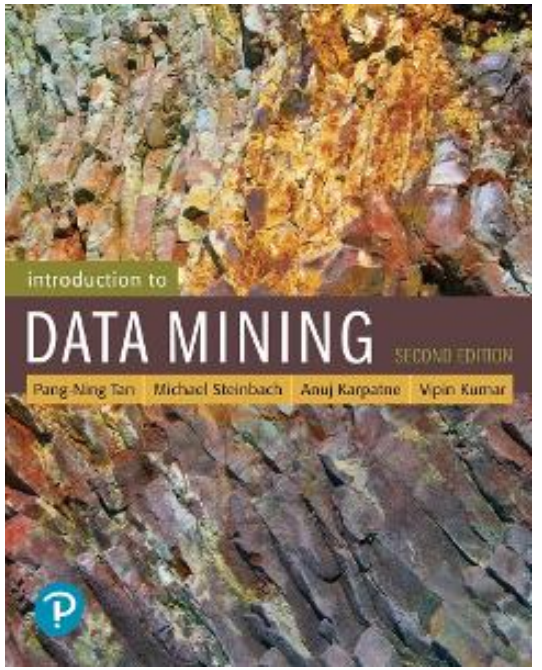




Material

- Web Site: http://didawiki.cli.di.unipi.it/doku.php/dm/start

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. **Introduction to Data Mining**. Addison Wesley, ISBN 0-321-32136-7, 2006, 2° Edition (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)



- Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. Guide to Intelligent Data Analysis. Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7
- Laura Igual et al. Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.
- Slides, Exercises and Notebook

Exam

- Project
  - Topics presented during the classes
  - A single report to be sent periodically and one week before the oral exam
  - Groups composed of up to 3 people (DM1), people (DM2)
- Oral Exam
  - Short discussion of the project (group presentation, where possible), plus
  - Questions on all topics presented during the classes
  - Exercises and questions about all topics

DM1 Mark = 0.6*Oral + 0.4*Project
DM2 Mark = 0.6*Oral + 0.4*Project
DM Mark = (DM1 + DM2) /2

Homework and Suggestions

Homework

- Declare Project Groups by next Thursday adding your information at https://docs.google.com/spreadsheets/d/1JX3VRwcZZFcTdpiguEwPsR_p4gDyRd7J89O84J7AeyY/edit?gid=251564882#gid=251564882
- **Suggestions**
- Download and start to play with the dataset and perform data understanding.
- Use a Github repository for python and ipython files.
- Use a shared Overleaf project (LaTex) for the report.

Dataset

- **Child Mind Institute (CMI) – Problematic Internet Use**
- Can you predict the level of problematic internet usage exhibited by children and adolescents, based on their physical activity?
- The goal of this competition is to develop a predictive model that analyzes children's physical activity and fitness data to identify early signs of problematic internet use. Identifying these patterns can help trigger interventions to encourage healthier digital habits.
- The goal of the competition is to predict from this data a participant's Severity Impairment Index (sii), a standard measure of problematic internet use.
- The CMI dataset to be used can be found on the web page of the course.
- Detailed guidelines for the project will be presented and made available on the web page of the course.

Questions?

riccardo.guidotti@unipi.it

alessio.cascione@phd.unipi.it

Let’s start!