DATA MINING 2
Course Overview

Riccardo Guidotti

UNIVERSITÀ DI PISA

Teachers

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

- **Andrea Fedele (Assistant)**
  - Computer Science Department
  - Email: andrea.fedele@phd.unipi.it

Classes

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

- Office Hours
  - Tuesday 15-17, Riccardo Guidotti's office
  - Appointment [DM2 Meeting] at riccardo.guidotti@unipi.it

- Teaching Assistant
  - Andrea Fedele [DM2 Meeting] at andrea.fedele@phd.unipi.it

No Classes and Recovery Classes

No Class

- Mon 03/03/2025
- Mon 01/04/2024 (Easter Monday)

Recovery Classes

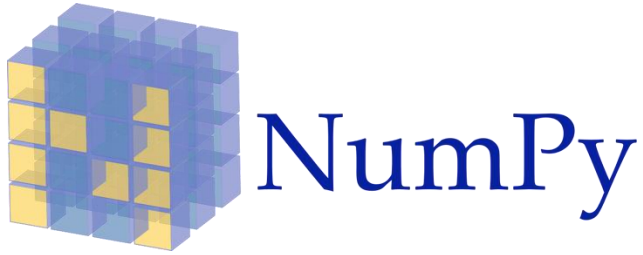
- Tue 04/03/2025 (Room D3)
- Tue 02/04/2025 (Room D3)

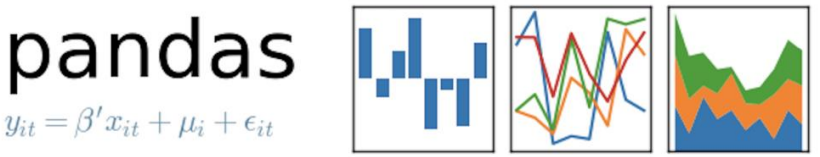
Topics

- **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

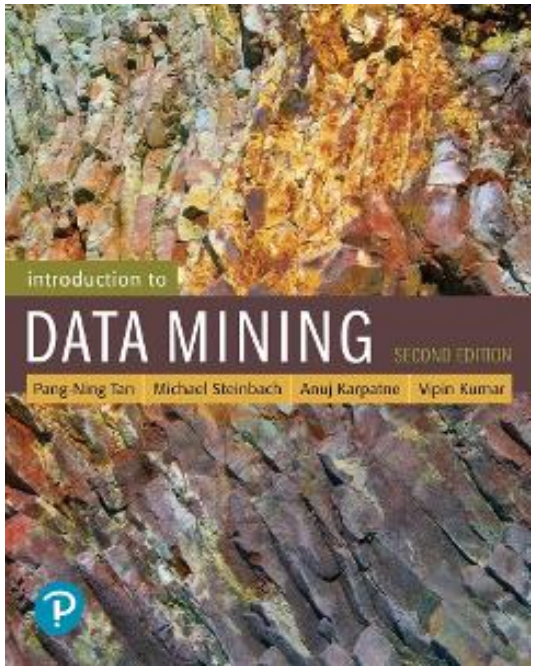




Material

- 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 Tuesday 28th February adding your information at https://docs.google.com/spreadsheets/d/1RFWIwKM5Myaehh4tHceaf3olMYm_CktGvoNOFX2Oovc/edit?usp=sharing
- **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

- **IMDb (extended) dataset + box office income time series**
- The IMDb Dataset contains data about movies, TV shows, and other forms of visual entertainment, along with their ratings, which is generated by the internet community. Each record includes key information such as the original title, release year, runtime, and the number of user votes. Additionally, the dataset provides insights into critical aspects like awards, nominations, and user reviews, as well as statistical ratings, from the best and worst ratings to the total number of critic reviews. It also includes metadata like country of origin, number of images and videos, and the title's genre.
- The IMDb dataset for the project 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

andrea.fedele@phd.unipi.it

Let’s start!