Introduction to Modern Data Science (24h)

Course Outline

This course introduces students to the modern data science ecosystem, combining scientific computing, data engineering, machine learning, and software engineering best practices. Students learn how to build reproducible data science projects using contemporary Python tools while developing the practical skills required to acquire, process, analyze, visualize, and model real-world data.

The course covers the complete data science workflow, from environment setup and collaborative development to exploratory data analysis, statistical inference, machine learning, REST APIs, containerization, and the fundamentals of MLOps. Emphasis is placed on reproducibility, scalability, and industry-standard tools widely used in professional data science.

1. Computing Environment and Python Fundamentals Overview of Linux, Windows, and macOS

File systems, paths, and permissions

Essential shell commands and scripting

Python refresher: scripts, modules, and packages

Command-line workflows for data science

2. Modern Python Tooling

The modern Python ecosystem

Virtual environments and dependency management with uv

Code quality using ruff (linting and formatting)

Development with Visual Studio Code

AI-assisted programming and productivity tools

Reproducible Python environments

3. Data Manipulation and Processing

Working with tabular data using NumPy and Pandas

Data formats: CSV, Parquet, and Apache Arrow

Data cleaning and preprocessing

Filtering, joins, aggregations, and missing values

Performance considerations

Introduction to scalable data processing with Polars, DuckDB, Dask, and Spark

4. Exploratory Data Analysis and Statistical Thinking

Principles of exploratory data analysis (EDA)

Descriptive statistics and data summarization

Statistical inference with SciPy

Data visualization using Matplotlib and Seaborn

Interactive dashboards with Plotly and Streamlit

Communicating insights through visualizations

5. Version Control and Project Organization

Git fundamentals

GitHub and collaborative development workflows

Branching, merging, and pull requests

Organizing reproducible data science projects

Notebooks versus Python scripts

Literate programming with Marimo

6. Building Data Applications

Consuming REST APIs with requests

Designing REST APIs with FastAPI

Introduction to containerization with Docker (or Podman)

Building portable and reproducible applications

Deploying simple data services

7. Introduction to Machine Learning and Artificial IntelligenceThe machine learning workflow

Supervised and unsupervised learning

Feature engineering and model evaluation

Introduction to deep learning

Large Language Models (LLMs)

Generative AI, prompt engineering, and AI agents

8. MLOps and Production Machine Learning

From notebooks to production

Model serving: batch and real-time inference

Experiment tracking with MLflow

CI/CD for machine learning projects

Introduction to cloud-native ML workflows

Model monitoring, drift detection, and retraining strategie

From:
https://wiki.centrale-med.fr/msct-cse/ -

Permanent link:
https://wiki.centrale-med.fr/msct-cse/csds

Last update: **2026/09/01 12:06**