Master 2
Institut de physique du globe de Paris
Academic year 2025-2026

Earth Data Science

Léonard Seydoux (seydoux@ipgp.fr), Antoine Lucas,
Éléonore Stutzmann, Alexandre Fournier & Geneviève Moguilny

Course objectives

This course aims at introducing the fundamental concepts of data science and machine learning applied to Earth sciences, with a strong emphasis on (1) understanding the underlying concepts rather than using methods as black boxes, (2) practical implementation and hands-on experience on real-world datasets, and (3) critical thinking regarding the results produced by statistical and machine learning models.

Examples and practical sessions are mainly drawn from geophysics (seismology, lidar, environmental sensors), but the methods are generic and transferable to other domains. The course is designed to be accessible to students with a basic background in programming (Python) and statistics. Practical sessions will provide opportunities to deepen understanding through direct application, and will make use of the Scikit-Learn and PyTorch libraries on the S-CAPAD platform.

General organization

The course consists of **6 sessions of 4 hours each**, plus **one final examination** at the end of the course. Each session typically follows this structure: approximately **1 hour of lecture** (theory, intuition, examples), followed by approximately **3 hours of hands-on practical work** on computers. Practical sessions are carried out individually and form a central component of the course. Session dates are indicated in the online course schedule, at this link.

Session	Lecture (1 hour)	Practical work (3 hours)
1	Introduction to machine learning: regression	**River sensor calibration: can a cheap sensor replace an expensive one?** Notions seen in the practical session include data inspection, linear and non-linear regression, model evaluation, and physical interpretation of results.

Session	Lecture (1 hour)	Practical work (3 hours)
2	Classification and feature extraction	**Classification of seismo-volcanic signals**. Introduction to feature extraction, model selection, and error analysis with confusion matrices.
3 & 4	Representation, classification and clustering	**Lidar data: supervised classification and unsupervised exploration of cloud of points.** Notions seen include feature extraction, clustering algorithms, and evaluation metrics. We will perform an in-depth analysis, comparison of methods, and critical interpretation of results.
5	Deep learning	**Neural networks applied to digit recognition.** This includes fully-connected networks and convolutional neural networks. For students willing to go further: inference using PhaseNet, a deep learning model for seismic phase picking.
6	Course summary and Q&A	Synthesis of lectures and practical sessions, discussion of results, and possible improvements. Introduction to recent AI approaches (foundation models, language models, attention mechanisms, frugal learning). Critical reading of a recent scientific paper.

Assessment

The final assessment will take the form of an **individual mini-project**, inspired by a lightweight hackathon format over four hours. The project will be based on a novel dataset to classify, where students will be asked to compare different classification approaches introduced during the course. The goal is **not** to achieve the best possible predictive performance.

Grading will primarily be based on the quality of reasoning, the ability to compare and interpret models, the clarity of result presentation, and critical thinking with respect to the methods used.

The final exam will take place during the last session of the course. The document to be submitted will be a single Jupyter notebook, containing code, comments, and visualizations.

Important remarks

Active participation during practical sessions is strongly encouraged. Questions, discussions, and personal initiatives are welcome, especially during synthesis sessions.

This planning may be subject to minor adjustments depending on the progress of the course.