

Machine Learning in Physics

Six weeks of machine-learning foundations, then a semester-long research project in mechanistic interpretability, two presentations and a final paper.

INSTRUCTOR

Prof. Dr. Emre Onur Kahya

CONTACT

eokahya@itu.edu.tr

INSTITUTION

Istanbul Technical University

SCHEDULE

Timetable, room and office hours to be announced

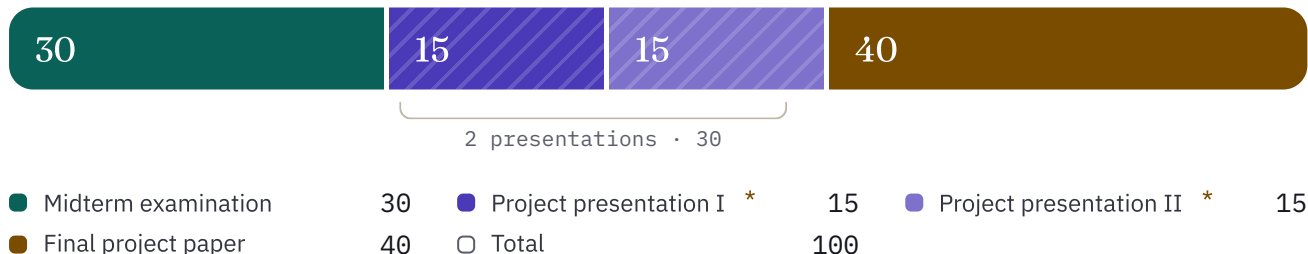
Source version 2026-fall-v1 · updated 2 October 2026

Overview

MYZ 310E introduces the foundations of machine learning through mathematical reasoning, implementation, and examples relevant to physics. The first six weeks provide an intensive foundation in statistical learning, supervised and unsupervised methods, and neural networks. The remainder of the semester is organised around a research project in mechanistic interpretability, with two presentations and a final paper.

The foundations component draws inspiration from the publicly available IFT 6390 course outline at Université de Montréal/Mila. It is a selective adaptation for this course, not a claim of equivalent coverage. The mechanistic interpretability project is a distinct component of MYZ 310E.

ASSESSMENT AT A GLANCE



Final-paper eligibility: at least **12/30** (40% of the midterm itself) **and** completion of both project presentations.

Syllabus

Plan status. The assessment weights and final-assessment eligibility rule are instructor-specified. The equal presentation split, use of the same project for both presentations, presentation milestones, indicative weekly schedule, detailed rubrics, and proposed paper length are provisional teaching choices. This plan does not claim institutional syllabus approval.

Recommended background

Working knowledge of linear algebra, multivariable calculus, elementary probability and statistics, and basic Python programming. These are recommended preparation, not newly imposed official registration prerequisites. Short refresher materials will be identified where appropriate.

Learning outcomes

By the end of the course, students should be able to:

1. Formulate learning problems and evaluate models using appropriate data splits, baselines, and metrics.
2. Implement and compare basic regression, classification, and neural-network methods.
3. Explain the distinction between predictive performance, representation analysis, and causal evidence about a model's computation.
4. Design a focused mechanistic interpretability experiment with suitable interventions and controls.
5. Analyse uncertainty, limitations, and reproducibility, and communicate findings in a scientific paper.

Weekly schedule

An indicative 14-week plan (provisional). The midterm is planned for week 7; presentation rounds are weeks 8–9 and weeks 12–13. Calendar dates will be announced.

WEEK	FOUNDATIONS
1	Statistical learning and scientific data Learning tasks, data representation, a probability/likelihood refresher, loss functions, training/validation/test separation, and data leakage. Illustrative physics application: a small, controlled measurement or classification dataset.
WEEK	FOUNDATIONS
2	Regression, optimisation, and generalisation Linear regression, regularisation, gradient-based optimisation, overfitting, model selection, and uncertainty in evaluation. Core implementation: a transparent regression baseline.
WEEK	FOUNDATIONS
3	Classification and decision boundaries Nearest neighbours, logistic/softmax classification, probabilistic classification, and evaluation metrics. Introduce the motivation for margins and kernels without requiring a full treatment of every method. Begin provisional project-topic selection.
WEEK	FOUNDATIONS
4	Neural networks Multilayer perceptrons, computational graphs, backpropagation, stochastic optimisation, regularisation, and practical training diagnostics. Connect the mathematical description to a small PyTorch implementation.
WEEK	FOUNDATIONS
5	Unsupervised learning and representations PCA, clustering, representation geometry, and interpretable low-dimensional examples. Briefly position trees and ensemble methods within the broader ML landscape. Deeper treatments may be supplementary.
WEEK	FOUNDATIONS
6	From neural computation to mechanistic interpretability An introductory account of attention, residual streams, and transformer components. Introduce activation inspection, ablation and patching, the limits of attention visualisation, and the distinction between probes and causal tests. Use one small worked demonstration rather than claiming to cover the whole field.

WEEK 7	MIDTERM Midterm examination and project scoping The examination date will be announced. Confirm a research question, a feasible model/task pair, and an initial experimental plan.
WEEK 8	PRESENTATION Project presentation I — first round The first presentation round is provisionally distributed over weeks 8–9 as needed. Present motivation, related work, a testable hypothesis, baseline results, and planned controls. Every student/project presents once across this round.
WEEK 9	PRESENTATION Project presentation I — continued Continue the first presentation round and feedback. Every student/project presents once across weeks 8–9; this is one presentation requirement, not two separate presentations.
WEEK 10	CLINIC Controlled experiments and ablations Project clinic: intervention design, comparison conditions, confounding factors, and held-out evaluation.
WEEK 11	CLINIC Robustness and reproducibility Project clinic: uncertainty estimates, repeated runs where feasible, robustness checks, and preparation of a reproducible experiment package.
WEEK 12	PRESENTATION Project presentation II — second round The second presentation round is provisionally distributed over weeks 12–13 as needed. Present results, causal evidence, limitations, and the structure of the final paper. Every student/project presents once across this round.
WEEK 13	PRESENTATION Project presentation II — continued Continue the second presentation round and feedback. Every student/project presents once across weeks 12–13; this is the second presentation requirement.
WEEK 14	WRITING Paper development and feedback Scientific writing, figure quality, methodological review, and reproducibility checks. This is not an additional graded presentation.
AFTER 14	FINAL ASSESSMENT PERIOD Final project paper The final project paper is submitted in the final assessment period. The submission deadline and submission channel will be announced.

Detailed CNN/RNN architectures, full SVM derivations, probabilistic graphical models, and advanced ensemble methods are supplementary unless explicitly assigned. Their presence in IFT 6390 does not imply that they are all taught in depth within six weeks here.

Lecture notes

Instructor-provided weekly PDF notes are published on the course page as they are released:
<https://eokahya.github.io/teaching/myz-310e/2026-fall/>.

Project guide

Research area: Mechanistic interpretability

Both presentations provisionally develop the same research project in two stages: the research question and initial evidence, followed by results and critical interpretation.

Investigate a clearly stated question about the internal operation of a small neural network or language model. Begin with a model/task combination that demonstrably works before interpreting its internal computation. Do not depend on a small language model solving advanced physics problems it cannot reliably solve.

The physics connection should be substantive: for example a controlled dataset, a symmetry or invariance question, a perturbation-response hypothesis, or an experimentally testable constraint. A physics metaphor alone is insufficient.

Include a baseline, an appropriate negative or matched control, a clearly defined evaluation procedure, and an honest discussion of uncertainty and limitations. Where feasible, use repeated seeds and held-out examples. Report the actual computational budget; expensive model training is not a prerequisite.

Small-model results should not be presented as proof that frontier systems are safe. Include a short discussion of the relevance and limits of the work for robustness, reliability, or AI safety.

A carefully executed negative result is acceptable. Assessment rewards scientific quality, not positive results or acceptance at a conference. The aspiration is rigorous research methodology, not guaranteed publication.

Possible starting points

- Controlled representation analysis
- Counterfactual activation patching
- Sensitivity to task-preserving perturbations
- Tests distinguishing probe information from causal use

These are starting points, not claims that a topic is novel or that a result is already known.

Proposed final-paper format

Provisionally, 4–6 pages of main text, excluding references and appendices. Suggested sections: Abstract, Introduction, Related Work, Methods, Experimental Setup, Results, Discussion and Limitations, Safety/Relevance, Conclusion.

Provide code and reproducibility instructions as supporting material, not as an additional grading category.

Team arrangements and detailed submission instructions will be confirmed in class. No student names, private submissions, or grades are published on this website.

Research integrity and tool use

Cite sources and borrowed code. Do not fabricate data, results, or references. Disclose substantive AI assistance and verify the resulting text and code. Students remain responsible for their work and should be able to explain their methods and results. Any exam-specific tool rules will be announced separately; this statement does not grant permission to use AI during the midterm.

Assessment

COMPONENT	POINTS	STATUS
Midterm examination	30	Fixed
Project presentation I	15	Provisional split of the 30 presentation points
Project presentation II	15	Provisional split of the 30 presentation points
Final project paper	40	Fixed
Total	100	The two presentations together carry 30 points

Final-assessment eligibility

Midterm $\geq 12/30$ **AND** Both presentations completed $\Rightarrow$ Final paper assessed

“To be eligible for the final project-paper assessment, a student must achieve at least 40% on the midterm examination AND complete both project presentations.”

“For a midterm marked out of 30, the minimum qualifying score is 12/30. This threshold applies to the midterm itself, not to the overall course grade.”

The final assessment is a paper based on the project. There is no additional written final examination specified in this course plan. Practice exercises and project clinics do not introduce additional grading categories.

Proposed rubrics

Project presentation I · 15 points
(provisional) · Weeks 8–9

CRITERION	POINTS
Research question and relevant background	4
Experimental design and controls	5
Baseline implementation or initial evidence	3
Clarity and responses to questions	3

Project presentation II · 15 points
(provisional) · Weeks 12–13

CRITERION	POINTS
Experiments and controls	6
Interpretation, uncertainty, and limitations	5
Clarity and responses to questions	4

CRITERION	POINTS
Research question and related work	6
Methodological correctness and controls	10
Experimental evidence and evaluation	10
Interpretation, limitations, and safety relevance	6
Reproducibility documentation	5
Clarity of writing and figures	3

Resources

Instructor-provided weekly PDF notes are linked under Lecture notes as they are released. Suggested supplementary references — third-party resources are accessed on their original websites:

Probabilistic Machine Learning: An Introduction

Kevin P. Murphy — supplementary reference

<https://probml.github.io/pml-book/book1.html>

Deep Learning

Ian Goodfellow, Yoshua Bengio and Aaron Courville — supplementary reference

<https://www.deeplearningbook.org/>

IFT 6390 public course outline

Université de Montréal/Mila — foundation-course inspiration; selective adaptation

<https://mitliagkas.github.io/ift6390-ml-classes/>

TransformerLens documentation

Tools and introductory material for mechanistic interpretability

<https://transformerlensorg.github.io/TransformerLens/>

Additional readings will be announced.

Announcements

No announcements yet. Course announcements will appear here.