**Master Internship Proposal** contact `edwige.cyffers@ens-lyon.fr`
Dauphine Lamsade 2026

# Principled Performative Prediction Datasets

## Context

**Performative Learning** [4] addresses distribution changes due to model deployment, aiming to minimize

$$\mathrm{PR}(\theta) = \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \ell(Z;\theta).$$

Unlike classical machine learning where $\mathcal{D}$ is fixed, here it is parameterized by the same parameter $\theta$ used in the loss function. Finding a performatively optimal solution requires accounting for the model's performative effect on the distribution. Typical examples include loan applications or hiring, where applicants may modify their features to increase their probability of being classified in the positive class, and the optimal classifier should thus take into account the extent to which a given feature may be modified. This problem has also been studied in terms of regret minimization [2], to ensure that the whole trajectory minimizes the risk.

In order to minimize the performative risk, one can try to optimize the classical minimization problem without taking into account the dependency of the data distribution on the parameter. On the contrary, one can try to steer the distribution toward a more favorable regime. The optimal algorithm depends on the type of feedback loop induced by the distribution.

**Performative datasets and existing use-cases** The need for better benchmarks in performative learning has been identified [1]. Existing works mainly focus on toy datasets, where the distribution shift is synthetically encoded according to a simple parametric form. An interesting example of a real-world performative dataset is Wisconsin's Dropout Early Warning System [5], but the data has not been released in a reproducible manner, nor benchmarked with performative algorithms. Interactive playgrounds have also been developed [3], but with limited success, and without discussion of the desirable properties required for performative datasets.

## Goal of the internship

The goal of the internship is to propose better datasets to study performative learning, by designing realistic distribution shifts that depend on the deployed models. In particular, good models for performative datasets should satisfy some properties such that the hardness of the task varies with the deployed distribution, and this change satisfies some regularity properties. One objective will be to establish these properties, to find diverse data generators satisfying them, and, if time permits, to benchmark existing algorithms on the new proposed datasets. Another possible outcome is to develop sampling method over a large and heterogenous dataset to simulate performative distribution shifts.

## For the candidate

The ideal candidate is finishing a Master's degree in Machine Learning, Applied Mathematics, or a related field, and is curious about the social impacts of machine learning. The candidate should have a solid background in basic probability and optimization. Prior knowledge of long-term fairness, performative learning, or online learning is a plus, but not required. The balance between theoretical and empirical work may depend on the candidate. The project will likely start with a survey and reimplementation of existing datasets to gain familiarity with the field, and will then move toward more complex models and proofs of interesting properties.

The internship will take place in Paris intra-muros, at PariSantéCampus or Université Paris Dauphine, within the LAMSADE team, with regular meetings. The internship offers the standard "gratification", around 700 euros per month.

**Starting date** The expected starting date is between mid-April and June 2026 (flexible).

**Extension into a PhD** is possible and shall be decided upon mutual agreement. A dedicated funding is available.

**To apply** send an email to `edwige.cyffers@ens-lyon.fr` with:

1. A short description of your motivation: what you understand from the project, why you believe you are a good fit (maximum 500 words)
2. A CV
3. Your transcript (first year of the Master's and available grades for the second year)

## References


[1] Moritz Hardt and Celestine Mendler-Dünner. *Performative Prediction: Past and Future*. 2025. arXiv: `2310.16608 [cs.LG]`.

[2] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. "Regret Minimization with Performative Feedback". In: *ICML*. 2022.

[3] John Miller, Chloe Hsu, Jordan Troutman, Juan Perdomo, Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt, and Moritz Hardt. *WhyNot*. 2020.

[4] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. "Performative Prediction". In: vol. 119. Proceedings of Machine Learning Research. PMLR, 2020.

[5] Juan Carlos Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. "Difficult Lessons on Social Prediction from Wisconsin Public Schools". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '25. Association for Computing Machinery, 2025.