ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Data Attribution

As generative machine learning models become increasingly ubiquitous across various industries, understanding the influence of specific data points on model outputs has become an important challenge. Data attribution in machine learning addresses this need by enabling us to trace model outputs back to the data found in the training set that shaped them. This not only enhances transparency and interpretability but also aids in identifying biases, ensuring compliance with data protection regulations, and safeguarding against data poisoning attacks.

In this project, we aim to explore innovative approaches to address the attribution challenge across various generative applications (audio, image, language, etc.) by integrating both empirical and theoretical perspectives. A key challenge lies in establishing reliable attribution methods that can generalize across diverse generative models, which remains an open problem in both complexity and applicability.

Requirements: Strong Python programming skills and knowledge of various ML libraries. Preferably with interest in writing up a conference paper. An interest in the underlying theory of machine learning is a plus.

Weekly meetings will be scheduled to address questions, discuss progress, and brainstorm future ideas.

Contact

In a few short sentences, please describe your interest in this project and any relevant coding experience or background (e.g., projects or coursework).

- Luca Lanzendörfer: lanzendoerfer@ethz.ch, ETZ G93
- Frédéric Berdoz: fberdoz@ethz.ch, ETZ G60.1