# A Foundation Model and Graph Transformer Networks for Big Data Analytics

Prof. Yanlei Diao (École Polytechnique)

| | |
|---|---|
| **Contact**: | Prof. Yanlei Diao (École Polytechnique, yanlei.diao@polytechnique.edu) |
| | http://www.lix.polytechnique.fr/~yanlei.diao/ |
| **Duration:** | 36 months, available immediately |
| **Deadline:** | Application is open until the position is filled |
| **Team:** | CEDAR team joint between École Polytechnique and Inria |
| **Location:** | École Polytechnique, Palaiseau, France |
| **Application:** | Email a CV and two names of references to the contact |

## Project Description

Global digital transformation increasingly relies on cloud computing for big data analytics, where a growing blend of SQL and Machine Learning (ML) workloads is becoming common. Achieving effective cost-performance optimization in this context requires precise models of SQL and ML workloads. However, a key challenge arises: How can we develop a universal modeling approach that effectively generalizes across millions of cloud users? This challenge is compounded by significant workload diversity (with each user exhibiting a unique workload), data variability (ranging from sparse to extensive telemetry data), and cluster heterogeneity (hundreds of cloud instances with diverse software and hardware configurations). While numerous modeling techniques have been developed, they typically train individual models for *each* user workload, leveraging tree-based methods or, more flexibly, graph neural networks. However, these models often suffer from high error rates, largely due to insufficient training data.

To tackle this challenge, this PhD project aims to pioneer the *development of a foundational model for Relational Algebra*, the core representation underpinning SQL queries in big data analytics. Building on our extensive experience with Graph Transformer Networks (GTNs) from the previous ERC project, our research will take a deep dive into GTN architectures, systematically deconstructing them across various design dimensions. For each dimension, we will innovate new methods specifically tailored to the unique requirements of relational algebra. In addition, we will develop self-supervised learning techniques adapted to relational algebra to scale training effectively. We also plan to explore novel optimizations to accelerate training while reducing computational complexity.

The second objective of this project is to extend our modeling solution to encompass non-SQL tasks, such as machine learning workloads. We will introduce a novel representation learning paradigm for non-SQL workloads, designed to capture workload characteristics from incomplete and noisy telemetry data. This paradigm will integrate seamlessly with our SQL-based modeling approach, utilizing a *generalized Graph Transformer Network that accommodates both SQL and non-SQL tasks in a unified framework.*

The third objective of this project is to fine-tune the foundation model or the generalized graph transformer network for different user workloads and apply them to assist in multi-objective optimization of cloud analytics workloads, thereby enabling the best cost performance for each cloud user.

## Research Environment

École Polytechnique is a French public institution of higher education and research, located in Palaiseau 30 minutes southwest of Paris. It is considered the most prestigious engineering school in France, with

well-known educational programs in science and engineering. Among its alumni are three Nobel prize winners, one Fields Medalist, three Presidents of France, etc. École Polytechnique is a member school of Institut Polytechnique de Paris, a larger university composed of 5 leading engineering and statistics school in close vicinity. Our research team is further a joint team between École Polytechnique and Inria (French Institute for Research in Computer Science and Automation). The student will conduct research in a highly dynamic, collaborative environment, with experts on big data, machine learning, statistics, and distributed systems colocated in the Palaiseau area.

**Desired Background**: Applicants should hold an M.S. degree in Computer Science or Data Science, with a strong background in data analytics and machine learning. Qualified candidates are expected to have proficiency in SQL queries, a good understanding of neural networks, and solid implementation skills in Python. Prior research experience is preferred. The project will be conducted in English.