Visualisation and Analysis of Workflow Provenance Data for Energy-Efficient Artificial Intelligence

Silvina Caino-Lores

November 14, 2024

1 Contact Information and Supervisory Team

Main Contact	Silvina Caino-Lores (ISFP, Inria) silvina.caino-lores@inria.fr https://scainolo.github.io/
Laboratory	Institut national de recherche en sciences et technologies du numérique (Inria) Inria Centre at Rennes University, France
Research Group	KerData (Scalable Storage for Clouds and Beyond; leader: Gabriel Antoniu) https://www.inria.fr/en/kerdata https://team.inria.fr/kerdata/
Supervisory Team	Silvina Caino-Lores, PhD (Inria, France) Jakob Luettgau, PhD (Inria, France) Alexandru Costan, PhD, HDR (INSA Rennes, France)

2 Context and Overview

Artificial Intelligence (AI) is driving scientific discovery and economic growth in all kinds of application domains while impacting from routine daily tasks to societal-level challenges. However, research communities, industry players and social actors are expressing increasing concern about the potential ethical and practical implications of the pervasive presence of AI. Of particular concern is the environmental impact of the GPU-, CPU-, and memory-intensive tasks in the machine learning life-cycle (e.g., large-scale model training), which to high energy consumption [GMRRG19].

The ideal of Responsible and Trustworthy AI (RTAI) garnered attention from academia, industry, national laboratories, and government agencies worldwide [LWM24, Oak24b, IBM24, Mic24, LZXW23, SLD+22]. Particularly, principles of environmental sustainability in AI focus on the environmental and societal impact of developing and using AI models that are energy-efficient, and how this impact can be assessed in the short and long terms. This includes analyzing features such as the carbon footprints and the computational power required for training algorithms [VW21].

To address these challenges, the FAIR principles (i.e., findability, accessibility, interoperability, and reuse of digital assets) have emerged as a valuable framework [WDA+16]. However, FAIRness in AI goes beyond the mere organization and sharing of data and code, encompassing the entire workflow that shapes AI models and applications. Recent works suggest that workflow provenance (i.e., the documentation and tracking of all processes within AI development) might hold the key to supporting FAIR and Responsible AI [SAL+22, KNHJ+23]. Workflow provenance refers to capturing detailed information about all activities, processes, and transformations applied to data and code during AI development and operations. It includes information about data sources, data preprocessing, model selection, hyperparameter tuning, and evaluation metrics, among others. Capturing this provenance could provide a holistic view of the AI workflow, making it transparent and reproducible. However, a challenging aspect of working with AI workflow provenance data is that there are no comprehensive

frameworks or tools to navigate and analyse the complex relationships in these workflows [BBFM23, KRCLJT22].

3 Research Objectives and Envisioned Approach

This project aims to investigate methods to retrieve, inspect, analyse and visualise provenance metadata in AI-powered workflows. We will build upon previous work and active research of the members in the supervisory team in the USA and France, thus this work will be complementary to our ongoing collaboration with the Workflows and Ecosystems Group from Oak Ridge National Laboratory (ORNL). In this collaboration we aim to bring together multiple sources of system telemetry data (e.g., edge devices, supercomputing facility, cloud resources), ML-specific provenance metadata taxonomies, and workflow analysis methods to enable the comprehensive assessment of the energy efficiency of AI workflows. We build upon three state-of-the-art tools:

- Flowcept [SSW+23], a data integration system that captures, stores, and queries workflow provenance in high-performance computing systems, developed by our ORNL collaborators.
- E2CLab [RCAV21], our solution for reproducible workflow execution with support for capturing provenance and monitoring metadata, which extends the provenance capturing capabilities of Flowcept to edge devices.
- IOBAT [LSC+18, LSR+23], a framework to enable the programmatic and interactive analysis and exploration of workflow, performance, and system artifacts for scientific applications in distributed and high-performance computing environments.

As a first step, the student shall become familiar with the interfaces and data layout of Flowcept and IOBAT with the goal of integrating the provenance metadata captured by Flowcept into the workflow topology and analysis framework of IOBAT. This will facilitate subsequent analysis and exploration of the provenance metadata with the existing capabilities of IOBAT. Together with our collaborators, we will refine our existing visualisations and energy-efficiency analyses for a large language model (LLM) fine tuning case study running on the Frontier [Oak24a] supercomputer at ORNL. We will explore how different hyperparameter configurations, datasets, model architectures, layer configurations and behavioural outputs (e.g., accuracy, loss and performance over time) affect system metrics like temperature and resource utilisation (e.g., GPU, CPU, disk), and assess the impact in fine-grained and end-to-end energy consumption. Finally, we will investigate the addition of new visualisation and analysis methods to allow the vertical (i.e., from lower to higher level of detail) and horizontal (i.e., different tasks at the same level of detail) exploration of provenance metadata, with a focus on discovering the relationships between models, data, outcome, performance, system behaviour and energy footprint.

References

[BBFM23] Elisa Bertino, Suparna Bhattacharya, Elena Ferrari, and Dejan Milojicic. Trustworthy ai and data lineage. *IEEE Internet Computing*, 27(6):5–6, 2023.

[GMRRG19] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75–88, 2019.

[IBM24] IBM Research. Trustworthy AI at IBM. https://research.ibm.com/topics/trustworthy-ai, 2024.

[KNHJ+23] Amruta Kale, Tin Nguyen, Frederick C Harris Jr, Chenhao Li, Jiyin Zhang, and Xiaogang Ma. Provenance documentation to enable explainable and trustworthy ai: A literature review. *Data Intelligence*, 5(1):139–162, 2023.

[KRCLJT22] Ariel Keller Rorabaugh, Silvina Caíno-Lores, Travis Johnston, and Michela Taufer. Building high-throughput neural architecture search workflows via a decoupled fitness

prediction engine. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2913–2926, 2022.

[LSC+18] Jakob Luettgau, Shane Snyder, Philip Carns, Justin M. Wozniak, Julian Kunkel, and Thomas Ludwig. Toward Understanding I/O Behavior in HPC Workflows. In *2018 IEEE/ACM 3rd International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISCS)*, pages 64–75, Dallas, TX, USA, November 2018.

[LSR+23] Jakob Luettgau, Shane Snyder, Tyler Reddy, Nikolaus Awtrey, Kevin Harms, Jean Luca Bez, Rui Wang, Rob Latham, and Philip Carns. Enabling Agile Analysis of I/O Performance Data with PyDarshan. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, pages 1380–1391, New York, NY, USA, November 2023. Association for Computing Machinery.

[LWM24] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

[LZXW23] Qinghua Lu, Liming Zhu, Xiwei Xu, and Jon Whittle. Responsible-ai-by-design: A pattern collection for designing responsible artificial intelligence systems. *IEEE Software*, 40(3):63–71, 2023.

[Mic24] Microsoft Research. FATE: Fairness, accountability, transparency, and ethics in ai. https://www.microsoft.com/en-us/research/theme/fate, 2024.

[Oak24a] Oak Ridge National Laboratory. Frontier at OLCF. https://www.olcf.ornl.gov/frontier, 2024.

[Oak24b] Oak Ridge National Laboratory. Oak Ridge National Laboratory initiative in secure, trustworthy, and energy-efficient AI. https://www.ornl.gov/ai-initiative, 2024.

[RCAV21] Daniel Rosendo, Alexandru Costan, Gabriel Antoniu, and Patrick Valduriez. E2clab: Reproducible analysis of complex workflows on the edge-to-cloud continuum. In *IPDPS 2021-35th IEEE International Parallel and Distributed Processing Symposium*, 2021.

[SAL+22] Renan Souza, Leonardo G Azevedo, Vítor Lourenço, Elton Soares, Raphael Thiago, Rafael Brandão, Daniel Civitarese, Emilio Vital Brazil, Marcio Moreno, Patrick Valduriez, et al. Workflow provenance in the lifecycle of scientific machine learning. *Concurrency and Computation: Practice and Experience*, 34(14):e6544, 2022.

[SLD+22] Conrad Sanderson, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle. Towards implementing responsible AI. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, December 2022.

[SSW+23] Renan Souza, Tyler J Skluzacek, Sean R Wilkinson, Maxim Ziatdinov, and Rafael Ferreira da Silva. Towards lightweight data integration using multi-workflow provenance and data observability. In *IEEE International Conference on e-Science*, 2023.

[VW21] Aimee Van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218, 2021.

[WDA+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.