Raphaël Azorin

Machine Learning PhD candidate with industry experience and published research.

Born 15 June 1993
French
raphaaal.github.io

EXPERIENCE

PhD Student — *Huawei Research / EURECOM*

Since June 2021. PhD program between industry and academia in Paris.

- Task affinity characterization for Multi-Task Learning.
- Representation learning for networking data using language models.
- Learned sketches: using ML to design probabilistic data structures.
- Using Transformers for tabular Deep Learning.

Data engineer — *Openfield (xCo Analytics)*

Sept. 2019 - Oct. 2020. Work/study program in Paris.

- Implemented and maintained Big Data analytics pipelines (Spark, Hive, Cassandra).
- Created a ML lifecycle management tool that cut model update delay by 90%.
- Built data quality dashboards that cut customer support time by half for related issues.

Data miner — *Marionnaud*

Sept. 2018 - Sept. 2019. Work/study program in Paris.

- Modelled customer behavior with scoring, forecasts, segmentations and recommendations.
- Created a clustering tool that supported a product line launch for 600K customers.
- Analyzed and reported marketing campaigns performances to management.

Technical marketing consultant — *Markentive*

Feb. 2016 - Sept. 2018. Permanent contract in Paris.

- Before transitioning to Computer Science, I was in marketing and analytics.
- I helped build the leading French marketing automation agency (from 5 to 30 employees).

EDUCATION

PhD in Machine Learning — *EURECOM (Sorbonne University)*

Since June 2021. Co-advised by Prof. P. Michiardi and Principal Engineer M. Gallo.

MSc in Machine Learning — *University PSL (Dauphine / ENS / Mines)*

September 2019 - September 2020. Salutatorian.

BSc in Computer Science — *University PSL (Dauphine)*

September 2017 - September 2019. Valedictorian.

MSc in Management — *Grenoble School of Business / Universität Mannheim*

September 2012 - September 2015. With honors.

SKILLS

Programming experience

- Python
- PySpark
- SQL

ML / DL libraries

- PyTorch, Sickit-Learn
- Pandas, Numpy
- PyPlot, Seaborn

Certifications

- Azure Fundamentals
- Microsoft Excel Specialist
- Google Analytics IQ
- HubSpot Software
- Marketo Certified Expert

Additional skills

- Java
- Azure and AWS Cloud
- Spark, Hive, Cassandra

ACTIVITIES

- Boxing (3 years)
- Climbing (3 years)
- Swimming (lifeguard qual.)
- Summer camp counselor

LANGUAGES

- French: native
- English: advanced
- Spanish: intermediate
- German: beginner

PUBLICATIONS

"It's a Match!" - A Benchmark of Task Affinity Scores for Joint Learning.
Azorin, R., Gallo, M., Finamore, A., Rossi, D., & Michiardi, P.
Accepted at *AAAI's 2nd International Workshop on Practical Deep Learning in the Wild arXiv:2301.02873*. 2023.

Learned data structures for per-flow measurements.
Monterubbiano, A., Azorin, R., Castellano, G., Gallo, M., & Pontarelli, S.
In *Proceedings of the 3rd International CoNEXT Student Workshop* (pp. 42-43). 2022.

Towards a systematic multi-modal representation learning for network data.
Houidi, Z. B., Azorin, R., Gallo, M., Finamore, A., & Rossi, D.
In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks* (pp. 181-187). 2022.

A Reproducible Approach for Mining Business Activities from Emails for Process Analytics.
Azorin, R., Grigori, D., & Belhajjame, K.
In *Proceedings of the 19th Service-Oriented Computing–ICSOC 2021 Workshops (pp. 77-91).* 2021.

Towards a generic deep learning pipeline for traffic measurements.
Azorin, R., Gallo, M., Finamore, A., Filippone, M., Michiardi, P., & Rossi, D.
In *Proceedings of the 2nd CoNEXT Student Workshop* (pp. 5-6). 2021

SELECTED PROJECTS

Automatic clustering – *Operational machine learning*

A scalable clustering lifecycle management tool deployed in a Big Data production environment at Openfield. Enables data scientists to update a clustering model by automatically taking into account the latest available data. The data is processed in a distributed fashion and models are deployed through ML Flow validation flow. Using PySpark, Hive and Cassandra, this project resulted in a 90% decrease in the time spent on updating and deploying new clustering models

Featuring prediction – *Graph analytics*

A musical collaborations predictive model using a graph database of hundreds of thousands nodes ("artists") and millions of relationships ("featurings") scraped from the Spotify API. The ML pipeline includes the computations of several graph metrics (community, centrality and similarity measures), then fed to a classification algorithm learning relationships between pairs of nodes. Using Python and Neo4J, this project resulted in a model with 86% accuracy for link prediction.

Pandemic – *Board game*

An implementation of the popular cooperation-based board game Pandemic. The game engine is playable by multiple humans through a GUI and CLI. Using Java, this project resulted in a game played by several symbolic AIs that ranked second in a university contest.

REFERENCES

Available upon request.