+XX XX XX XX XX
Paris, France
riccardo [dot] cappuzzo [at] gmail [dot] com

Riccardo CAPPUZZO

Postdoctoral researcher

GitHub: rcap107
LinkedIn: riccardo-cappuzzo/

Postdoctoral researcher with 6+ years of experience working in academia on data science and engineering subjects seeking to employ the acquired skills for more practical applications. Well versed in the Python data libraries, thorough and independent-minded in the development and presentation of experimental results.

Skills

Tools and Languages	Python (Polars, Pandas, scikit-learn, Matplotlib), Git, LaTeX, shell
Fields of research	Tabular learning, data engineering, benchmarking, feature selection
Communication	English (professional level), Italian (native), French (conversational level)

Technical Experience

Postdoctoral researcher — **Oct 2022 — September 2024**
SODA Team – INRIA Saclay, Dataiku — *Saclay, France*

- GB-scale data cleaning and engineering.
- Development of YADL, a synthetic benchmarking data lake.
- Development of "Retrieve, Merge, Predict", a pipeline for analyzing methods for automating table augmentation.

Doctoral Student — **Oct 2018 — April 2022**
EURECOM — *Sophia Antipolis, France*

- Development of EmbDI, a data integration algorithm based on embeddings of tabular data.
- Development of GRIMP, a data imputation algorithm based on graph neural networks and multi-task learning.
- Supervision of several master students as Teaching Assistant.

Software Developer — **Mar 2018 — Aug 2018**
SAP Labs — *Mougins, France*

- Integration of the eROCK algorithm in the Toreador framework.
- Development of software for the simulation of corporate data logs.

Internship — **July 2017 — Feb 2018**
SAP Labs — *Mougins, France*

- Development of eROCK, an enhanced version of the ROCK clustering algorithm.

Education

PhD in Automated Methods for Data Cleaning, *Sorbonne Université, France* 2018–2022
Master degree in Computer Security, *EURECOM, France* 2016–2018
Master degree in Communication and computer networks engineering, *Politecnico di Torino, Italy*, 110/110 2015–2018
Bachelor degree in Computer Engineering, *Politecnico di Torino, Italy*, 107/110 2012–2015

Activities

Retrieve, Merge, Predict, an experimental pipeline to benchmark table augmentation from data lakes. 2022–2024
YADL (Yet Another Data Lke), a synthetic data lake for stress-testing SOTA augmentation methods. 2022–2024
GRIMP (Graph embeddings for Relational data IMPutation), an imputation algorithm based on GNNs 2021–2022
EmbDI (Embeddings for Data Integration), a system for generating table embeddings for data integration 2019–2022
EDBT Summer School (Extracting Hidden Knowledge from Heterogeneous Massive Data) September 2019
eROCK, a clustering algorithm for categorical data Fall 2017
Semester project: studying the effect of Wifi networks on network protocols for IoT devices Spring 2017
Marco Poli Spring 2015

Publications

- Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2020, June). **Creating embeddings of heterogeneous relational datasets for data integration tasks** – 2020 ACM SIGMOD.
- Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2021, September). **EmbDI: Generating Embeddings for Relational Data Integration** – 2021 SEBD.
- Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2024, March). **Relational Data Imputation with Graph Neural Networks** – 2024 EDBT.
- Cappuzzo, R., Varoquaux, G., Coelho, A., & Papotti, P. (2024, February). **Retrieve, Merge, Predict: Augmenting Tables with Data Lakes** – Arxiv preprint.