Sara Pidò

Mobile:
+39 3484624263
Email:
sarapid@mit.edu
sara.pido95@gmail.com
Linkedin:
https://www.linkedin.com/in/sara-pido/

About Me

- Ph.D in Data Analytics and Decision Science with a focus on ML and NLP for democratising Data Science
- Developed research qualitative and analytical skills
- Proven skills as a team player willing to share knowledge and feedback with colleagues

Academic History

2020-2023 PhD cum Laude in Data Analytics and Decision Science, Politecnico di Milano
Thesis: Exploiting AI and NLP methods for Empowering Naive Users in Solving Data Science Problems
Fellowship offered by ERC Advanced Grant (693174)

2022-2022 Visiting PhD, Data to AI Lab, Massachusetts Institute of Technology
Fellowship offered by Rocca Project

2017-2019 M.Sc. Computer Science Engineering, Politecnico di Milano
Thesis: Network integration algorithms for the analysis of biological data

2018-2019 Erasmus+ Computer Science, KU Leuven

2014-2017 B.Sc. Computer Engineering, Politecnico di Milano
Thesis: Development of the game Lorenzo il Magnifico

Work Experience

Post Doctoral Associate

Massachusetts Institute of Technology (MIT) | Jan 2024 - December 2024

Data Scientist Intern

Calvin Risk, Zurich | May 2023 - August 2023

- Work on AI incidents: designing the integration and the quantification of AI incidents in a risk management platform

Visiting Research Assistant

Massachusetts Institute of Technology (MIT) | March 2022 - December 2022

- Independently carry out a project on time series data and autoML, designing and developing the related algorithms
- Work and collaborate with a company on another project on wind turbine data by performing data scientists tasks

Research Assistant

Politecnico di Milano | 2020-curr

- Design and develop independently algorithms
- Work collaboratively within a team
- Write scientific papers to present results
- Supervise and lead master's students thesis

Teaching assistant

Politecnico di Milano | 2020-2022

- Prepare practical lectures on Python programming for the Master's Course in Bioinformatics
- Prepare exercises and help students on C and MATLAB programming for the Bsc' students

Skills

- Advanced-level **Python**, **R, Matlab** and **C** languages
- Good knowledge of **Java**
- Good knowledge of **SQL**
- Experience using **Git** to manage programming code
- Experience in Kaggle challenges for **Machine Learning** and **Deep Learning** application
- Experience working in a **Linux** environment
- Advanced knowledge of **MS Office**
- Good knowledge of **BPMN, ER, UML**
- Familiar with **Adobe Photoshop, Lightroom**
- Advanced written and verbal **English** communication skills (my **TOEFL** level of English is currently **C1).**

Projects

Exploiting AI and NLP methods for Empowering Naive Users in Solving Data Science Problems

The main project of my Ph.D. thesis has the objective of **making data science accessible** to users without data science background by creating new algorithms for the automatic generation, labeling and filtering of **prediction tasks** for **time-series data,** designing new tools for assisting users in developing machine learning pipelines through **natural language processing, conversational agents** and **autoML.**

Computational analysis of fused co-expression networks for the identification of candidate cancer gene biomarkers

Developing a general framework to **infer relevant gene biomarkers** using multiple **gene co-expression networks** for each cancer type, specifically, of genes from kidney renal clear cell carcinoma, liver hepatocellular carcinoma, and prostate adenocarcinoma data sets of **TCGA database**. The gene communities are extracted through a **data-driven pipeline** and then evaluated through both functional analyses and literature findings. Furthermore, I provide a computational validation of their relevance for each cancer type.

A non-negative matrix tri-factorization based method for predicting antitumor drug sensitivity

Developing an enhancement of **Non-Negative Matrix Tri-Factorization** method to integrate different data types for the prediction of missing associations. Specifically, the method is tested on a dataset from the **Cancer Cell Line Encyclopedia (CCLE)**, containing the connections among cell lines and drugs by means of their **IC50 values**, and integrating it by linking cell lines to their respective tissue of origin and genomic profile. The method is proved through two different kind of experiments: a) prediction of missing values in the matrix, b) prediction of the complete drug profile of a new cell line.

Publications

Most relevant publications:

- **Pidò, Sara,** Pinoli P., Crovari P., Ieva F., Garzotto F. and S. Ceri. Ask Your Data—Supporting Data Science Processes by Combining AutoML and Conversational Interfaces. IEEE Access, 11(1):45972-45988, 2023.
- **Pidò, Sara,** Ceddia, G., and Masseroli, M. (2021a). Computational analysis of fused co-expression networks for the identification of candidate cancer gene biomarkers**.** NPJ systems biology and applications, 7(1):1–10
- Crovari, P., **Pidò, Sara**, Pinoli, P., Bernasconi, A., Canakoglu, A., Garzotto, F., and Ceri, S. (2021). Gecoagent: a conversational agent for empowering genomic data extraction and analysis. ACM Transactions on Computing for Healthcare (HEALTH), 3(1):1–29
- **Pidò, Sara,** Crovari, P., and Garzotto, F. (2021b). Modelling the bioinformatics tertiary analysis research process. BMC bioinformatics, 22(13):1–27
- **Pidò, Sara,** Testa, C., and Pinoli, P. (2021c)**.** A non-negative matrix tri-factorization based method for predicting antitumor drug sensitivity. bioRxiv
- **Pidò, Sara**, Pinoli, P., Crovari, P., Ieva, F., Garzotto, F., and Ceri, S. Ask your data. supporting data science processes by combining automl and natural language interfaces. IEEE Access. Accepted, 2022

Check out my google scholar profile for more!

https://scholar.google.com/citations?user=M_SrptEAAAAJ&hl=en

Conferences, Presentations, Networking

Dec. 2022 NEURIPS - Human in the Loop Learning New Orleans, LA (US)
Learning from Data through Human-Machine Collaboration

Oct. 2022 CIKM - Human-in-the-loop Data Curation Atlanta, GA (US)
A Paradigm to Reintegrate the User into the AutoML Loop through Natural Language

July 2021 DeepLearn 2021 Las Palmas, ES
What if Data Science is accessible to everyone?

Nov. 2020 Conceptual Modeling for Life Science Virtual
Towards an Ontology for Tertiary Bioinformatics Research Process

Sep. 2019 CIBB Bergamo (IT)
Computational analysis and comparison of gene networks from TCGA normal and cancer data

Referees

Kalyan Veeramachaneni
Massachusetts Institute of Technology
32 Vassar St, Cambridge, MA 02139, (US)
kalyan@csail.mit.edu

Stefano Ceri
Politecnico di Milano
Via Giuseppe Ponzio, 34, Milano 20133, (ITALY)
stefano.ceri@polimi.it