

## María José Basgall, PhD

[majo.basgall@gmail.com](mailto:majo.basgall@gmail.com) | +41 76 496 2084

Rue Beau-Site 1A, 1203, Geneva

Citizenship: argentine, holding a [swiss Ci work permit](#) (no sponsorship is required)

[linkedin.com/in/mjbasgall](https://www.linkedin.com/in/mjbasgall) | [github.com/majobasgall](https://github.com/majobasgall)

**Summary:** Data Scientist with a background in machine learning, statistical analysis and data visualization, working with large datasets. I finished my PhD in Computer Science in May 2022, with a strong technical focus on Big Data preprocessing using technologies such as **Spark**, **Scala**, **Python**. I am looking forward to starting a Data Science or related roles career in industry, able to leverage my previous experience to provide actionable insights and recommendations to improve decision-making. I would also like to contribute by sharing my skills as well as learning new ones in a teamwork environment.

### Professional Experience:

*PhD Researcher* | University of Granada, Spain / National University of La Plata, Argentina | 06/2018 - 05/2022

- Conducted data analysis on large datasets using **Python** statistical functions and **visualization** tools to detect patterns and trends
- Applied the complete data science lifecycle, with emphasis on **data preparation**, including exploratory data analysis, cleaning, feature engineering, and balancing techniques
- Worked with 20+ structured **Big Data sets**, mostly of real-world problems (e.g: bioinformatics, particle physics, humanitarian emergencies, e-commerce)
- Developed a **data reduction** tool that reduced dataset size (instances and attributes) by up to 70% and 98%, respectively, while maintaining original data's predictive power
- Enabled widely-used **data preprocessing** algorithms for large datasets and distributed environments
- Developed strategy for scalable big data characterisation, focusing on **data overlapping** analytics
- Algorithms implemented using **Scala** and **Spark**, running under a 10+ nodes **Hadoop** cluster, resulting in the publication of 5 peer-reviewed articles

*Graduate Researcher* | National University of La Plata, Argentina | 11/2014 - 05/2018

- Developed strategy for data stream (tweets) clustering, using **Java + Spark Streaming**
- Collaborative work with economist on cryptocurrency, resulting in a scientific paper with 500+ citations

*University Teacher* | National University of La Plata, School of Computer Science, Argentina | 11/2014 - 11/2022

- Taught classes on introduction to programming for beginners, and Big Data for advanced students

### Education:

- Double PhD degree in Computer Science | University of Granada, Spain / National University of La Plata, Argentina | Graduated 05/2022 | *Cum Laude* mention
- Systems Engineering degree | National Technological University, Argentina | Graduated 10/2014
- Systems Analyst degree | National Technological University, Argentina | Graduated 09/2013

### Technical Skills:

- Programming languages and libraries: Scala, Python, Java, SQL, Spark, Hadoop, Scikit-Learn, ImbLearn, Pandas, Matplotlib, NumPy
- Cloud and BI tools: Google Cloud Platform (GCP), Google BigQuery, Tableau, Power BI
- Others: Github, JIRA, Jupyter Notebook, GNU/Linux

### Certifications:

- Google Cloud; **From Data to Insights with Google Cloud** Specialization
- DeepLearning.AI; **Neural Networks and Deep Learning** course
- École Polytechnique Fédérale de Lausanne; **Functional Programming Principles in Scala** course
- École Polytechnique Fédérale de Lausanne; **Big Data Analysis with Scala and Spark** course

**Languages:** Spanish (native) - English (fluent) - French (elementary, *learning*)