

# María José Basgall, PhD

[majo.basgall@gmail.com](mailto:majo.basgall@gmail.com) | Geneva, Switzerland  
[linkedin.com/in/mjobasgall](https://www.linkedin.com/in/mjobasgall) | [github.com/majobasgall](https://github.com/majobasgall) | [majobasgall.github.io](https://majobasgall.github.io)

**SUMMARY: Data Scientist** with expertise in machine learning, statistics, and data visualization, proficient in handling large datasets using Spark, Scala, and Python. Currently working as a **Data Engineer** at a fast-paced Swiss company, using Microsoft Azure, Docker, and GitLab CI/CD, Bash, Python, and SQL. Additionally, I participated as a **Data Science Lead** in a Swiss volunteering program, collaborating with an agile team while working on NLP techniques.

## PROFESSIONAL EXPERIENCE:

**Data Engineer** | SOPHiA GENETICS, Rolle, Switzerland | 10/2023 - Present

- Applied hands-on expertise in **Microsoft Azure, Docker, GitLab CI/CD, Bash, Python** and **SQL**.

**Data Science Lead** | *deploy(impact)* volunteering program, remote, Switzerland | 09/2023 - 11/2023

- Applied **NLP** techniques for candidate/job matching mechanism within an **agile**, multidisciplinary team.

**PhD Researcher with Data Science Proficiency - Specializing in Big Data Analysis** | University of Granada, Spain and National University of La Plata, Argentina | 06/2018 - 05/2022

- Performed **statistical analysis** using **Python** and **data visualization** for pattern and trend detection.
- 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.
- **Developed a data reduction tool** that decreased dataset size by up to 70% (instances) and 98% (attributes) while maintaining original data's predictive power.
- Enabled widely-used **data preprocessing** algorithms for large datasets and distributed environments.
- Implemented algorithms in **Scala** and **Spark**, running on 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, Argentina | 11/2014 - 11/2022

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

## EDUCATION:

- **Double PhD in Computer Science** (2022) | University of Granada, Spain / University of La Plata, Argentina | Cum Laude
- **Systems Engineering** (2014) and **Systems Analyst** (2013) degrees | National Technological University, Argentina

## TECHNICAL SKILLS:

- **Programming:** Scala, Python, SQL, Spark, Hadoop, Scikit-Learn, Pandas, Matplotlib, NumPy
- **Cloud and BI tools:** Microsoft Azure, Google Cloud Platform, Google BigQuery, Tableau, Power BI
- **Others:** Docker, Gitlab CI/CD, Bash scripting, JIRA, Jupyter Notebook, GNU/Linux, Scrum

## CERTIFICATIONS:

- Google Cloud; **From Data to Insights with Google Cloud** Specialization | 2023
- DeepLearning.AI; **Neural Networks and Deep Learning** course | 2023
- EPFL; **Functional Programming Principles in Scala** course | 2019
- EPFL; **Big Data Analysis with Scala and Spark** course | 2019

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