

# Cainã Max Couto da Silva

Senior Data Scientist & Researcher

Madison - WI, USA

+1 (608) 395 8543 | [cmcouto.silva@gmail.com](mailto:cmcouto.silva@gmail.com) | [cmcouto-silva.github.io](https://github.com/cmcouto-silva) | [cmcouto-silva](#) | [cmcouto-silva](#) | [cmcouto-silva](#)

## Current Experience

### UW-Madison

Madison - WI, USA

DATA SCIENTIST RESEARCHER | TOP 15 PUBLIC UNIVERSITIES IN THE USA

Aug 2024 - Current

- Automation of multiple database development processes, reducing manual intervention time.
- An ML classifier to assess data quality at scale, allowing for the first-time scalable quality check.
- Enhance a deep Learning model to count cattle in protected areas using high-resolution imagery. In collaboration with Google, this project empowers Brazilian prosecutors to act against non-compliant farmers, thereby reducing Amazon's deforestation.
- Applied technologies: Python, Pytorch, AWS, Postgres, Docker, deep learning, computer vision, etc.

## Industry Experience

### Schlumberger

Houston - TX, USA (remote)

DATA SCIENTIST | WORLD'S LARGEST OFFSHORE DRILLING COMPANY

Jan 2023 - Jul 2024

- Code refactoring and predictive modeling for machine failures, saving huge amounts of US\$.
- Q&A chatbot powered by generative AI for the company data (RAG).
- Idea for innovation using AI ranked the **top 12** out of 237 teams worldwide. I was responsible for the pitch to the CEOs.
- Applied technologies: Dataiku, GCP, Python, SQL, Docker, machine learning, computer vision, langchain, llms, etc.

### Escola DNC

Sao Paulo, Brazil

DATA SCIENCE CONSULTOR (FREELANCER)

Oct 2021 - Jul 2024

- Provide group mentorship and Q&A sections. Prepare test exercises and assignments from core Python to model deployment.
- Covered material: Python, PySpark, regression, classification, clustering, recommender systems, model evaluation, and model deployment.

### Ambev Tech

Sao Paulo, Brazil

DATA SCIENTIST | WORLD'S LARGEST BEER BREWER COMPANY

May 2022 - Dec 2022

- Build the data pipeline for automating pricing and promotion policies.
- Advanced forecasting modeling for multiple products with hierarchical reconciliation.
- Applied technologies: Databricks, Python, Pyspark, SQL, MLFlow, Scikit-learn, and specific forecasting and data visualization libraries.

### Remessa Online

Sao Paulo, Brazil

DATA SCIENTIST | FINTECH

Oct 2021 - Apr 2022

- Exploratory data analysis, time series forecasting, modeling customer retention.
- Meetings with commercial and marketing teams to present insights obtained from statistical, graphical, and machine learning analysis.
- Applied technologies: Databricks, Python, Pyspark, SQL, MLFlow, Scikit-learn, Pycaret, Matplotlib/Seaborn, Plotly.

### Eli Lilly and Company

Indianapolis - IN, USA (remote)

SAFETY DATA SCIENCES ASSOCIATE

Jun 2021 - Oct 2021

- Work on queries and reports for teams worldwide.

## Academic Experience

### M.B.A. in Data Sciences & Analytics

May 2021 - Aug 2023

ESALQ - UNIVERSIDADE DE SÃO PAULO

São Paulo, Brazil

- In-depth study of machine learning models.
- Developed an end-to-end hybrid ML model for churn prediction (available [here](#))

### Ph.D. in Genetics and Evolutionary Biology

Jul 2016 - Apr 2021

UNIVERSIDADE DE SÃO PAULO

São Paulo, Brazil

- Analysis, visualization, and reporting of genomic data using R, Python, and bash scripting.
- Non-supervised algorithms (e.g. PCA), descriptive and inferential statistics, Bioconductor R packages.
- I provided all the code and instructions for replicating my thesis [in this repository](#) (Brazilian Portuguese).

### M.Sc. in Biological Sciences

Apr 2014 - Mar 2016

UNIVERSIDADE DE SÃO PAULO

São Paulo, Brazil

### B.A. in Biological Sciences

Feb 2011 - Dec 2013

UNIVERSIDADE GUARULHOS

São Paulo, Brazil

- Best academic performance's Award

## Publications

---

- **Couto-Silva, C. M.**, Shetty, S., Olid-Gonzalez, A., Wallez, G., Chatelet, C., Kohar, A. (2024). Mitigating Nonproductive Time: A Novel Algorithm for Dsl Fault Detection. OnePetro. <https://doi.org/10.2523/ITPC-24515-MS>.
- **Couto-Silva, C. M.**, Nunes, K., Venturini, G., Araújo Castro e Silva, M., Pereira, L. V., Comas, D., Pereira, A., Hünemeier, T. (2023). Indigenous people from Amazon show genetic signatures of pathogen-driven selection. Science Advances, 9(10). <https://doi.org/10.1126/sciadv.abo0234>.
- Castro e Silva, M. A., Ferraz, T., **Couto-Silva, C. M.**, Lemes, R. B., Nunes, K., Comas, D., Hünemeier, T. (2021). Population Histories and genomic diversity of South American natives. Molecular Biology and Evolution, 39(1). <https://doi.org/10.1093/molbev/msab339>.
- Jacovas, V. C., **Couto-Silva, C. M.**, Nunes, K., Lemes, R. B., de Oliveira, M. Z., Salzano, F. M., Bortolini, M. C., Hünemeier, T. (2018). Selection scan reveals three new loci related to high altitude adaptation in native Andeans. Scientific Reports, 8(1). <https://doi.org/10.1038/s41598-018-31100-6>.

## Technical Skills

---

Using the following tools, I can perform data cleaning, wrangling, and visualization, build supervised and unsupervised models, and evaluate and optimize **machine learning** models. I care a lot about storytelling and reproducible research.

### PROGRAMMING

• Python • R • SQL • Bash scripting

### FURTHER TOOLS

• Machine Learning libraries • PySpark • MLFlow • GCP • Git/GitHub • Virtual environments • Docker

## Talks

---

### Nubank Meetup

Brazil (remote)

DATA SCIENCE APPLIED TO MULTIPLE SECTORS: CHALLENGES, SOLUTIONS, AND LESSONS LEARNED

Oct 2024

- Nubank is the world's largest digital bank outside Asia
- I talked about my journey as a data scientist in different sectors and how I used various techniques, from linear regressions to more complex methods like neural networks, in a creative and practical way.

### PyData Global 2023

Global (remote)

INTRODUCTION TO MACHINE LEARNING PIPELINES: HOW TO PREVENT DATA LEAKAGE AND BUILD EFFICIENT WORKFLOWS

Dec 2023

- A 2-hours workshop explaining how pipelines help to prevent data leakage and ensure model stability by allowing for proper separation of training, validation, and test data. Through a blend of theory and practice, it walked the audience through code chunks in Python using well-known open-source packages to ensure a complete understanding of the machine learning pipelines.
- Slides and code available [here](#).

### PyData SP

Sao Paulo - Brazil (remote)

ASSOCIATION ANALYSIS: HOW TO EXTRACT VALUE FROM CATEGORICAL DATA (TRANSLATED FROM PORTUGUESE)

May 2022

- Workshop highlighting data science techniques for analyzing categorical data, such as chi-square, Cramér's V, CA, MCA, entropy, information gain, and so on.
- Code available [here](#)

## Additional Information

---

### Languages

• Portuguese (native) • English (professional) • Spanish (intermediate)

### Open-source contribution

I have been contributing to the [feature-engine](#) Python package, an open-source tool for feature engineering in machine learning models.

### Conferences

Throughout my professional career, I have had the opportunity to participate in various training courses and international conferences, including internships in Argentina, Spain, and the USA.