Cainã Max Couto da Silva

Senior Data Scientist & Researcher

Madison - WI, USA

+1 (608) 395 8543 | cmcouto.silva@gmail.com | cmcouto-silva.github.io | cmcouto-silva | cmcouto-silva | cmcouto-silva

Current Experience

UW-Madison
Madison - WI, USA

Data Scientist & Postdoctoral Researcher — *Aug 2024 - Current*

- Developed SQLDeps, a Python package leveraging LLMs to extract table and column dependencies from complex SQL scripts automatically, streamlining database change management and improving communication with third-party integrations.
- Engineered a pioneering machine learning solution to assess property data quality at scale, implementing feature engineering on geospatial and entity data to enable scalable data integrity verification for the first time.
- Developed and optimized cattle mapping in protected areas using deep learning on high-resolution satellite imagery, providing actionable intelligence to reduce Amazon deforestation
- Lead UW-Madison students in research projects, mentoring on machine learning and computer vision techniques
- Automated database development processes through SQL/Python integration, reducing manual intervention and increasing team productivity
- Applied technologies: Python, Pytorch, AWS, SQL, Postgres, Docker, deep learning, computer vision, LLMs, 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 U$.
- 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.

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/IPTC-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

I develop end-to-end data science solutions following best practices, emphasizing reproducible research, model interpretability, and impactful storytelling.

Programming

- **Python** • **R** • **SQL** • **Bash scripting**

Frameworks & Tools

- **Machine learning** • **Deep learning** • **Data Visualization** • **Big Data (PySpark & Postgres)** • **Cloud computing (GCP & AWS)** • **Statistics** • **Git/GitHub** • **Docker** • **MLOps best practices**

Talks

PyTorch Workflow Mastery: A Guide to Track and Optimize Model Performance

Brazil (remote)

PyData Global - One of the biggest Python events worldwide *Dec 2024*

- A 90-minute tutorial to teach how to tune deep learning models while tracking their performance with data and plots.

Data Science Applied to Multiple Sectors: Challenges, Solutions, and Lessons Learned

Brazil (remote)

Nubank Meetup - Largest digital bank outside Asia *Oct 2024*

- 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.

Introduction to Machine Learning Pipelines: How to Prevent Data Leakage and Build Efficient Workflows

Global (remote)

PyData Global - One of the biggest Python events worldwide *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.

Association Analysis: How to extract value from categorical data

Sao Paulo - Brazil (remote)

PyData SP *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 contributed to Feature-engine, an open-source Python package implementing a variety of feature engineering techniques.

I developed ProjectLens, an open-source Python tool that generates comprehensive project snapshots optimized for AI consumption, allowing developers to leverage LLMs for code analysis, refactoring, and documentation generation with zero external dependencies.

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.