Ricardo S. Carvalho

Canada – Timezone: From UTC-8 to UTC-3 (eligible to work in Canada)
rsilvaca@sfu.ca | ricardosc@gmail.com | +1 778-776-1186 | ricardocarvalho.ca | github.com/ricardocarvalhods

Hands-on **Data Scientist** with **10 years** of experience building **data-centric products and APIs** in collaboration with **cross-functional teams**, at big tech companies, such as **AWS**, and **startup** environments. Graduating from a **Ph.D. in Differential Privacy**, authored an **award-winning algorithm for privacy preservation,** and published work on top-tier AI venues, such as AAAI and UAI. Proactive team player, led **implementations of NLP** and **ML solutions from ideation to deployment to production**, with a strong focus on business goals and KPIs.

EDUCATION

Ph.D. Computer Science — **Simon Fraser University,** Canada — **2019 - 2023**

- Developed an algorithm for **ranking queries to a database** to add Differential Privacy on top of any database without internal modification, with formally proven better utility than previous similar work.
- Designed a mechanism to enforce Differential Privacy for **vocabulary building** and **SQL "group by" queries**, obtaining up to 25% better utility compared to previous algorithms proposed by Microsoft Research.
- Implemented from scratch and open-sourced an optimizer algorithm to generate private synthetic data with Differentially Private Conditional **Generative Adversarial Networks (DP-CGAN)** in Python + **TensorFlow 2**.
- Co-authored work on differentially private **deep learning** using the lottery ticket hypothesis, designed to improve the privacy-utility trade-off in **private neural networks**, outperforming current methods by 20%.

PG Certificate in Data Science — **University of Washington,** USA — **Mar–Dec/2016**

- Professional Education Program on Data Science with 3 courses to analyze real-life data scenarios, using R.
- Reports are available, e.g., Time Series Analysis, Bayesian Inference, and Hypothesis Testing of Auto Data.

M.Sc. Computer Science — **University of Brasilia,** Brazil — **2013 - 2015**

- Dissertation focused on ML with imbalanced data + adaptive regularization, applied to corruption prediction.

B.Eng. Computer Engineering — **Aeronautics Institute of Technology,** Brazil — **2007 - 2011**

WORK EXPERIENCE

Applied Scientist Intern — **Amazon Web Services (AWS),** Remote, Canada — **Apr–Jul/2020**

- Designed & implemented **NLP** algorithms to **generate differentially private text** from users' sensitive texts, which **reduced** execution time by 68% and storage requirements by 98% compared to Amazon's existing algorithm, by using binary embeddings and optimized nearest-neighbors search implemented in **Python**. The corresponding paper won **Best Paper Award** at Amazon's internal Machine Learning Conference.
- Developed a differentially private mechanism to **aid NLP models via domain adaptation** on users' sensitive texts, getting 32% better test accuracy than Amazon's existing solution on a sentiment analysis problem. The solution was implemented in **PyTorch** and executed using **Amazon SageMaker**.

Data Privacy Consultant — **Bank of Canada / SFU,** Remote, Canada — **2022**

- Designed from scratch and led a 5-day **workshop** on Privacy Enhancing Technologies (PETs), for 10 senior scientists from the data science team to incentivize the use of privacy technologies at the Bank. Covered k-anonymity, differential privacy, homomorphic encryption, privacy-preserving machine learning, and private synthetic data generation. Included theory + **code tutorials** in Python with, e.g., Opacus, diffprivlib, OpenDP.
- Built a **web app**, in Python + Streamlit, that evaluates different privacy technologies given a dataset and a set of goals. It will be used internally by the Bank to **disseminate privacy technologies** and related use cases.

Sr. Machine Learning Engineer **Federal Court of Accounts,** Remote, Brazil **2020 - Current**

- Built a text similarity model with Random Forest, RegEx, and TF-IDF to suggest documents that are related.
- Designed **data pipelines** for various apps ETL, with Python, Docker, and Airflow + Kubernetes Pod Operator.
- **Deployed** a Named Entity Recognition (NER) model **to production** on Azure using **Docker** containers. Built complete CI/CD on **Azure pipelines**, with ACR integration, unit tests with PyTest.
- Implemented REST APIs with FastAPI as the backend of an app that gathers data from individuals cited in documents. Data is extracted using a NER model, RegEx, ElasticSearch queries, and microservices.

Lead Data Scientist | Data Scientist **Comptroller General of the Union,** Brazil **2012 - 2018**

- For one year worked as Lead Data Scientist, leading 15 engineers and data scientists on projects in Data Science, Research, and Auditing. Coordinated national data analysis projects with cooperation between more than 20 states in the country, centralizing data, and reports. Led multiple works on Government data, including identifying anomalies in IT purchases using autoencoders, finding relational databases with similar structures using gradient boosting machines, and reducing mobile consumption cost with outliers detection.
- Built an app using Django to verify federal **suppliers' defaults**, by extracting, processing, and merging data from multiple microservices & databases updated daily, reducing the analysis time from hours to real-time.
- Implemented an end-to-end system to **process federal denunciations**, with complete data pipeline.
- Developed and deployed in production a classification model to **rank civil servants** to be investigated based on the likelihood of being corrupt, using political party affiliation data. Compared to the assessment made by human experts, the resulting model had the same **precision** of 86% while increasing the recall by 88%.
- Web scraping project (Selenium + Python) to gather data to build a data profile of individuals and companies.
- Created a **statistical analysis of insurance death benefits**, finding multiple possible fraudulent schemes.

OTHERS

Project Reviewer **Udacity,** Remote **2017 - Current**

- Reviewer of 7 projects on the Data Analyst and Data Scientist nanodegrees.

Technical Writer **Medium.com | Towards Data Science,** Remote **2021 - Current**

- Writer of technical content related to data science. The most popular post has almost 40,000 views.

Open Source Contributor **Privacy libraries,** Remote **2021 - Current**

- Contributor to PipelineDP by Google and OpenMined.

Conference Reviewer (PC Member) **Multiple AI/ML/DM conferences,** Remote **2019 - Current**

- Reviewer or PC member for AI/ML/DM conferences. Latest venues: KDD 2022, AISTATS 2022, WSDM 2022.

SKILLS

- **Languages, Frameworks and Tools:** Python | R | SQL | ElasticSearch | MySQL | MSSQL | Docker | Airflow | Spark | Cloud computing | Azure | Microservices | REST | CI/CD | OOP | Unit testing | Backend | Git
- **Python libraries:** scikit-learn | numpy | pandas | tensorflow | pytorch | spacy | nltk | transformers | re | matplotlib | seaborn | plotly | streamlit | pytest | selenium | scrapy | beautifulsoup | requests | pydantic
- **Data science / Machine learning techniques:** A/B Testing | ETL | Data pipelines | Data-flow | Experimental design | Hypothesis testing | Statistics | Transformers | Embeddings | Anomaly detection | Federated Learning

→ Complete list of publications (Google Scholar Profile): https://scholar.google.com/citations?user=V8v6VckAAAAJ