

Rafik Chemli – Sr Data Scientist

[Email](#) • [LinkedIn](#) • [GitHub](#) • [Website](#)

Skills: Python, SQL, JS, AWS, Data Visualization, Interactive Dashboards, Machine learning Cycle, Synthetic Data Generation, Federated Learning, Model attacks.

Profile

Senior Data Scientist with extensive experience in health data, specializing in data visualization, machine learning, data pipelines, and privacy. Proven track record of leveraging data-driven insights to drive meaningful decisions. Proficient in transforming complex datasets into intuitive visualizations and developing real-world predictive models. Dedicated to ensuring data privacy compliance and making quantifiable impact through state-of-the-art data-driven solutions.

Education

- Master's degree, Bioinformatics (2018 - 2019) | UQAM | Université du Québec à Montréal | Montréal, Canada.
- Bachelor of Science - BS, Physiology and Toxicology (2015 - 2018) | UQAM | Université du Québec à Montréal | Montréal, Canada.

Experience

2022 – Present

Senior Data Scientist | Government of Canada | Statistics Canada - Data Science Division

- Contributed to the development and unit testing of Hydra benchmark module, a cutting-edge pipeline for automated information extraction from diverse data sources.
- Collaborating with CIFAR to investigate the implementation of privacy-preserving techniques to ensure security and integrity of sensitive data and prevent machine learning model inversion attacks (membership inference attack, reconstruction attacks, attribute inference attack).
- Member of an international team as part of the UnitedNation PETLab project to enhance the applications of Privacy Enhancing Technologies.
- Conducted project scoping and cost estimation to guide strategic decision-making for incoming projects.
- Supervision and management experience, providing mentorship to junior data engineer.
- Active member of the Responsible ML community of expertise and contributing to a forthcoming blog post on Model Poisoning on the Data Science Network.
- Attended prominent AI and Health conferences such as MILA and MAIN 2022, representing Statistics Canada and gaining insights from industry leaders and researchers.

Technologies: Python, bash, git, data engineering, Differential Privacy, Federated Learning, Homomorphic Encryption, Secure Multi-Party Computation, Kubeflow, Pydantic data structure.

2021 – 2022

Senior Compliance Analyst | Government of Canada | Canadian Revenue Agency

- Leveraged advanced analytical tools and techniques including classification, clustering, and text analysis to analyze, interpret, and visualize large, complex tax-related data sets.
- Conducted rigorous data cleaning, pre-processing, and feature engineering requiring a deep understanding of the complex database structure of the CRA, relationships, and integrity constraints to ensure reliable data inputs for modeling.
- Managed the end-to-end development of data pipelines in a complex relational database environment, encompassing data collection, extraction, transformation, loading (ETL), analysis, and reporting.
- Collaborated with various regional stakeholders (such as tax auditors, tax analysts, and policy makers) to understand their needs and requirements and translated these into practical data science solutions.

Technologies: R, SQL, SPSS, SAS.

2020 – 2021

Data Scientist | Galenvs Sciences

- Optimisation of nanoparticles—specifically, magnetic beads used for nationwide COVID-19 testing in Canada. Employed a variety of machine learning techniques, such as neural networks and XGBoost. This project required sophisticated data preprocessing, feature engineering, outlier detection and interpretability processes, which ultimately enhanced the efficiency of the testing product.
- Developed a flexible and interactive web application dashboard with ML prediction capabilities. This application enabled the visualization and effective analysis of over 100,000 data points in real time.
- Collaborated with a cross-disciplinary R&D team consisting of chemists, biologists, and biochemists, which resulted in an improved understanding and optimization of various products.
- Gained valuable insight into cloud technologies and hosting platforms. This experience contributed to the overall project execution and maintenance, particularly in the context of web server hosting.

Technologies: Python, bash, git, XGBoost, SVM, Keras, tensorflow, scikit, Dash, Plotly, Flask App Factory, live Web application, CSS, HTML, AWS, T-SNE, feature engineering (hot encoding, correlations, power transformations), Monte Carlo simulation, model interpretability.

2019 – 2020

Analyst Developer | Sphera solutions

- Spearheaded the development and maintenance of "Intelligent Authoring," a tailored compliance management software for Environment, Health, and Safety (EHS) management solutions. The project heavily utilized technologies such as .NET, SQL, and Microsoft Azure Cloud.
- Played a pivotal role in understanding and implementing changes to compliance regulations. This responsibility ensured that the software remained current with the latest requirements, thereby providing the most accurate and relevant information to users.

- Helped drive the adoption of the software by leading entities in the energy sector, including Shell, BP, and Exxon. The software aids these organizations in providing safe, compliant labeling and instructions for their products at an international level.
- Collaborated closely with domain knowledge experts to develop a compliance and EHS management solution, effectively translating their requirements into structured pipelines.

Technologies and Tools: .NET, SQL, Microsoft Azure Cloud, Compliance Regulations, EHS Management.

2019

Data Scientist Intern | My Intelligent Machines

- Primary responsibilities revolved around detecting new therapeutic targets against basal-like and triple-negative breast cancer, using multi-omics data points.
- Conducted extensive statistical analysis and computation optimization of over a million multi-omics data points. Employed tools such as RStudio Server, Python, SQL, and Bash on a Linux local server to ensure rigorous and efficient data analysis.
- Gained hands-on experience in predictive modeling by designing and implementing multiple regression and classification models using Python3, Scikit, Jupyter Notebook, and R Studio.
- Developed a solid foundation in data science within the healthcare domain, particularly in the context of breast cancer research.

Technologies: RStudio Server, Python, SQL, Bash, Git, Scikit, Jupyter Notebook.

Core skills

- Critical thinking
- Communication.
- Leadership & collaboration.
- Adaptability.
- Ethical considerations.
- Continuous learning.

Languages

- French | Native proficiency.
- English | Professional proficiency.
- C/C/C Profile (Statistics Canada).

Recent training and courses

- Cyber Security for Citizen Developer – CCCS (2022).
- Data or Specimens only Research certification – CITI program (2023).
- Intensive English learning program at the English Language Institute (UBC) (2018).

Recent publications and conferences

Santos, B., et al. (2023). Insights into Privacy-Preserving Machine Learning. To be presented in the *Expert Meeting on Statistical Data Confidentiality*. United Nations Economic Commission for Europe, Wiesbaden.

United Nations. (2023). United Nations Guide on Privacy-Enhancing Technologies for Official Statistics. Case studies 9, 15, and 16. *United Nations Committee of Experts on Big Data and Data Science for Official Statistics*, New York. <https://unstats.un.org/bigdata>

Mila Partner Symposium 3rd edition 2023.

Montreal AI and Neuroscience conference (Main) 2022.