

Ayrton San Joaquin

🔬 **SCIENTIST - TRUSTWORTHY AI** | MLOPS ENGINEER | ✍️ **WRITER**

✉️ ayrton@aya.yale.edu | 📍 Singapore | 🌐 [ajsanjoaquin](#) | 📺 [ajsanjoaquin](#) | 📖 [Values](#)

Education

Yale-NUS College

Singapore

BSC. (HONORS) IN DATA SCIENCE, MINOR IN PHILOSOPHY. (**SCHOLAR, WITH HIGH DISTINCTION**)

August 2018 - May 2023

Focus: Computer Vision (CV) and Natural Language Processing (NLP). **Semester Abroad at University of Copenhagen.**

Experience

French National Centre for Scientific Research (CNRS)

Singapore

NLP RESEARCHER

September 2023 - Present

- Working on Dialogue Summarization for critical decision making as part of the French-Singapore Governments' DesCartes program.

Machine Learning Safety Scholars Program, Center for AI Safety

Palo Alto, United States

SCHOLAR

June 2022 - August 2022

- Studied model failures (CV and NLP), and led research on analyzing Large Language Models (LLMs) using few-shot learning.
- Implemented various strategies in **robustness** (PGD, adversarial training), **anomaly detection** (AUROC, ViM), **calibration** (RSME, Brier scores), and **trojan attacks** (data poisoning).
- Received a grant of US\$4500 to complete the inaugural 2-month program.

Data Privacy and Trustworthy Machine Learning Lab, NUS

Singapore

UNDERGRADUATE RESEARCHER

May 2021 - March 2022

- Collaborated with Google DeepMind on privacy and adversarial machine learning research for my bachelor's thesis in a team across 4 time zones. **Published in a top security conference (ACM CCS) as the youngest and only undergraduate co-author.**

Arterys (Freelance)

San Francisco, United States

DEEP LEARNING ENGINEER

March 2020 - June 2020

- Created a COVID-19 Pneumonia classifier **4 days after pandemic declaration in collaboration with A.I. Singapore.**
- Collaborated with Arterys to **deploy the model in their platform** for use by American hospitals and researchers. Model engineer in a team of 4 across 3 time zones.

Open-Source Projects & Contributions

What's Cooking? Multilingual Recipe Search Engine

Semantic Search, NLP

- Developed an application that allows a user to find a recipe from a 100K recipe-database given a list of ingredients in any of the 170 supported languages using Cohere's Embed model.

Explaining Neural Networks with Meaningful Perturbations

Explainable AI, CV

- For explaining an image classifier's prediction, I implemented the algorithm described in *Explanations of Black Boxes by Meaningful Perturbation* (Fong, et. al., 2018).

Equitable Valuation of Data Using Shapley Values

Data Governance

- Implemented the training data valuation algorithm from *What is your data worth? Equitable Valuation of Data* (Ghorbani and Zou., 2019).

Open-Source AI

DevOps

- Added new features for major machine learning projects including Pytorch, HuggingFace Transformers, and YOLOv4 (object detection model).

Publications

December
2022

San Joaquin, A., Skubacz, F. , Applying Multilingual Models to Question Answering (QA)

[arxiv link](#)

November
2022

Tramer, F., ..., **San Joaquin, A.**, et.al. , Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets

[ACM CCS 2022 link](#)

March 2020

San Joaquin, A., *Using Deep Learning to Detect Pneumonia caused by COVID-19*

[Towards Data Science \(Editor's Choice\) link](#)

Skills

Programming: Python, Java

Machine Learning: Pytorch, Sickit-Learn, JAX, Keras, Tensorflow, Transformers, Langchain, NLTK, Spacy

Data: Pandas, Querying (SQL, MySQL, PySpark, MongoDB), Vector Database (Qdrant, Pinecone)

MLOps: AWS, AzureML, Snowflake, Docker, Flask, Continuous Integration, Kubernetes, MLFlow