

Ayrton San Joaquin

RESEARCHER - TRUSTWORTHY MACHINE LEARNING (PRIVACY, SECURITY) | WRITER

✉ ayrton@yale-nus.edu.sg | 📍 Copenhagen, Denmark | [in ajsanjoaquin](#) | [🔗 ajsanjoaquin.github.io](#)

Education

Yale-NUS College

Singapore

BSC. (HONORS) IN DATA SCIENCE, MINOR IN PHILOSOPHY

August 2018 - February 2023

Scholarship recipient. Focus: Computer Vision (CV) and Natural Language Processing (NLP). Currently at Københavns Universitet.

Experience

Machine Learning Safety Scholars Program, Center for AI Safety

Palo Alto, United States

TEACHING ASSISTANT (FALL), SCHOLAR (SUMMER)

June 2022 - December 2022

- Led a 5-person grading team handling 97 students worldwide, ranging from pre-university students to professionals.
- Led NLP research on analyzing large language models' adaptability to new word definitions using few-shot learning.
- Received a grant of \$4500 to complete the inaugural program.

Data Privacy and Trustworthy Machine Learning Lab, NUS

Singapore

UNDERGRADUATE RESEARCHER

May 2021 - August 2021

- Pitched and led a project to analyze **Unlearnable Data** as a data protection method.
- Collaborated with Google Brain on privacy and adversarial machine learning research for my bachelor's thesis in a team across 4 timezones. Published 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 timezones.

Open-Source Projects & Contributions

Equitable Valuation of Data Using Shapley Values

Data Protection

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

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).

COVID-19 Pneumonia Classifier for Diagnosis Triage

Medical Imaging, CV

- Trained a Resnet-34 Convolutional Neural Network (CNN) on ~ 26,000 images with Resampling to detect Pneumonia caused by COVID-19 on xray scans ultimately to triage patients for urgent diagnosis.

Miscellaneous

Machine Learning, NLP, CV

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

Publications

December 2022 **San Joaquin, A.,** Haroen, A. , Understanding How Model Size Affects Few-shot Instruction Prompting

[arXiv](#)

December 2022 **San Joaquin, A.,** Skubacz, F. , Applying Multilingual Models to Question Answering (QA)

[arXiv](#)

November 2022 Tramer, F., ..., **San Joaquin, A.,** et.al. , Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets

ACM CCS 2022

March 2020 **San Joaquin, A.,** Using Deep Learning to Detect Pneumonia caused by COVID-19

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

Press

April 2022 **Machine learning models leak personal info if training data is compromised**, *The Register*

Skills

Programming Languages: Python, Java, R

Machine Learning in Python: Pytorch, Pytorch Lightning, NumPy, Sickit-Learn, Tensorflow, Keras, Jax

Data Management: Pandas, SQL, MS Excel

Application Deployment &

Version Control: Docker, Google Cloud, Git, Singularity