

Ayrton San Joaquin

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

✉ ajsanjoaquin@gmail.com | 📍 Copenhagen, Denmark | 📄 ajsanjoaquin | 🔗 ajsanjoaquin.github.io

Education

Yale-NUS College

Singapore

BACHELOR OF SCIENCE (HONORS) IN DATA SCIENCE, MINOR IN PHILOSOPHY

August 2018 - February 2023

Awarded Scholarship to attend Full-time. Currently on exchange at Københavns Universitet.

Experience

Machine Learning Safety Scholars Program, Center for AI Safety

Palo Alto, United States

SUMMER SCHOLAR

June 2022 - August 2022

- Supported by the FTX Future Fund to attend the first iteration of the program.
- Led research on analyzing large language models' adaptability to new word definitions using few-shot learning.

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 attack research for my bachelor's thesis. Resulted in a publication as the only undergraduate co-author.

NUS-Tsinghua Center For Extreme Search (NeXT++)

Singapore

DEEPPAKE DETECTION RESEARCH INTERN

May 2020 - August 2020

- Automated training of models on Face++ Dataset (>175,000 images). Role included adapting various defences against adversarial examples (e.g. Adversarial Training, Randomized Smoothing)

Arterys (Freelance)

San Francisco, United States

DEEP LEARNING ENGINEER

March 2020 - June 2020

- Created a COVID-19 Pneumonia classifier four days after pandemic declaration in collaboration with A.I. Singapore.
- Contacted by Arterys, and Deployed model in the Arterys platform, alongside models from NVIDIA and Ping An Technology, for use by American hospitals and researchers.

Open-Source Projects & Contributions

Twitter Algorithmic Bias Challenge 2021

- Identified unintended sexualization of non-sexual images involving nudity by the Twitter Image Cropper Algorithm. Finished 9th out of 40 teams worldwide.

Explaining Neural Networks with Meaningful Perturbations

- 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

- 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

- Contributed improvements to major machine learning projects including Pytorch, HuggingFace Transformers, and YOLOv4 (object detection model).

Publications

*No name indicates first or sole authorship.

November 2022 Tramer, F., ..., San Joaquin, A., et.al. , Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets, To appear in ACM CCS 2022.

March 2020 , Using Deep Learning to Detect Pneumonia caused by COVID-19 Towards Data Science

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