

# Shea Cardozo

 SheaCardozo |  shea-cardozo |  sheacardozo.com |  sacardoz@cs.toronto.edu

## RESEARCH INTERESTS

---

I am broadly interested in creating generalizable autonomous agents and systems. This encompasses model-based reinforcement learning techniques for planning and reasoning, creating robust representations of dynamic scenes, causal learning and understanding, and behaviour prediction in multi-agent scenarios.

## EDUCATION

---

2023 - 2027	<b>PhD in Computer Science</b> Waterloo Intelligence Systems Lab Supervisor: Professor Krzysztof Czarnecki	University of Waterloo
2021 - 2022	<b>Master of Science in Applied Computing</b> Data Science Concentration	University of Toronto
2017 - 2021	<b>Bachelor of Mathematics in Statistics</b> Co-operative Program. Dean's Honours.	University of Waterloo

## SKILLS

---

<b>Programming Languages</b>	Proficiency with Python, R, and C++. Experience with Java, C#, DART, Haskell, Scheme (Racket), SQL, VBA, HTML/CSS.
<b>Deep Learning Frameworks</b>	Proficiency with PyTorch. Experience with Jax, TensorFlow
<b>Development and Deployment</b>	Git, Linux, AWS, GCP, VSCode, RStudio, Jupyter Notebooks

## PUBLICATIONS

---

- Shea Cardozo, Gabriel Islas Montero, et al. 2022. *Explainer Divergence Scores (EDS): Some Post-Hoc Explanations May be Effective for Detecting Unknown Spurious Correlations*. Presented at the AIMLAI workshop at CIKM 2022. Available at [link](#).

## INDUSTRY EXPERIENCE

---

**Machine Learning Scientist** May 2022 - Sept. 2023  
Tenyks

- Chief Machine Learning Scientist at Tenyks (YC'S21) reporting directly to the CTO, researching and prototyping new methods for classification, detection, and segmentation tasks. Projects include:
  - Created a synthetic data pipeline for zero-shot anomaly detection via inserting photorealistic synthetic anomalies into clean data using RGB pixel masks and neural image inpainting.
  - Implemented a scalable system for open-set object retrieval based on visual and text prompting using 'Segment Anything' proposals and 'OpenCLIP' data embeddings. Fine-tuned 'OpenCLIP' embeddings on domain-specific data to improve retrieval performance beyond baseline.
  - Prototyped multiple active learning and core set approaches to suggest an efficient data annotation strategy for settings with minimal annotated data and expensive annotation costs.
  - Implemented a sequence of automated 'data quality checks' to detect common dataset issues in computer vision such as a high prevalence of occlusion or inconsistent class definitions.
  - Formulated a novel evaluation criteria for post-hoc neural network explanations to detect dependence on spurious correlations. Verified and submitted our work as a workshop paper.

## **Data Scientist - Claims AI Team**

Sept. 2020 - Dec. 2020

Intact Insurance

- Constructed a pipeline to automatically classify insurance documents from image and text data using an ensembled 'ResNet' convolutional neural network and 'BERT' transformer neural network.
- Experimented with multi-objective non-gradient optimization methods such as the 'NSGA-II' genetic algorithm to optimize model prediction thresholds to mark unclassified documents for manual review.

## **Data Scientist - Analytics**

May 2020 - Aug. 2020

Noom Inc.

- Specified and fit an autoregressive time series model with seasonal effects to predict the influx of user support tickets to ensure sufficient resource availability.
- Trained and benchmarked a set of 'GloVe' vector embeddings constructed from internal food data to improve user meal recommendation and tracking.

## **Actuarial Analyst - DataLab Division**

Sept. 2019 - Dec. 2019

Intact Insurance

- As part of the 'Rating Revolution' team, trained 'XGBoost' gradient-boosted decision tree models to replace the existing generalized linear models used in home insurance pricing
- Created a Python visualization tool to analyze how different pricing models impact wider financials.

## **Associate Actuarial Programmer**

Jan. 2019 - Apr. 2019,

Moodys Analytics

May 2018 - Aug. 2018

- Implemented highly performance sensitive financial calculations into our insurance software platform using the C++ programming language, with focus on long-term maintainability.
- Expanded UI functionality to more transparently display to clients how financials are calculated.

# **PERSONAL PROJECTS**

---

## **Graddy**



- Created a basic deep learning framework entirely from scratch in Python, including an implementation of AutoGrad full capable of handling the specification, training and inference of neural networks.
- Implemented data structures for generic modules and tensor computation, popular deep learning layers and activation functions, and optimization algorithms such as SGD and SGD with Momentum.

## **Adversarial Conditional UNET**



- Trained a UNET model to generate adversarial examples using the CIFAR-10 dataset to fool a variety of state-of-the-art image classification models, inspired by existing work in adversarial denoising.
- Verified trained model can conduct targeted attacks on models not used in training with only marginal decrease in success rate, indicating successful generalization of the generated adversarial examples.

## **This JoJo Does Not Exist**



- Trained a StyleGAN2 image generation model to generate faces that resemble characters from the manga "JoJo's Bizarre Adventure", using a custom dataset scraped using a 'Selenium' Python bot.
- Employed a Google Cloud Platform VM to train model for 72 hours on a Linux GPU instance.

## **Trump Tweet Generator**



- Trained the GPT-2 338M language model on former US President Donald Trump's twitter feed.
- Built a web app using the 'Flask' Python package to display generated tweets.