

# Khoi Nguyen

nguyentienkhai01@gmail.com · (438) 938-0432

linkedin.com/in/khoi-tien-nguyen · github.com/khoitiennguyen · khoitiennguyen.github.io

## Work Experience

---

### Intact Insurance, Data Lab

Montreal, Canada

DATA SCIENTIST INTERN

May 2023 – Aug 2023

- Performed comprehensive exploratory data analysis on client flood claims, geospatial maps, and satellite imagery using **Python** (Pandas, Matplotlib, NumPy) and **R**, uncovering critical insights for flood risk assessment.
- Increased flood claim identification rate by 85%, leveraging weather claims classification **LLM** and feature engineering.
- Discovered key patterns in claim distribution affecting 700,000+ customers through rigorous hypothesis testing.
- Reduced data inconsistencies by 60% through extensive data cleaning and preparation using **SQL**, ensuring data reliability and validity for model evaluation.

### McGill University, Distributed Digital Music Archives & Libraries Lab

Montreal, Canada

SOFTWARE DEVELOPER

May 2022 – Apr 2023

- Developed a workflow for user-based image labeling and optical recognition model training using **Python** and **Docker**.
- Implemented image processing functionalities with **TensorFlow** and **OpenCV** to accelerate training data preparation efficiency, saving annotators 80+ hours per quarter.
- Designed a data loading scheduler for user-labeled datasets, increasing model training speed by 200%.
- Collaborated with researchers for iterative modeling changes, leading to significant experimental advancements.

### Accreon

Fredericton, Canada

SOFTWARE DEVELOPER CO-OP

Sep 2018 – Jan 2019

- Refined **ETL** processes for improved data handling, introducing new features for enhanced flexibility and efficiency in data processing using **Java** and **SQL**.
- Streamlined client data extraction into bilingual formats and implemented compression techniques, improving data storage by 30% and increasing browser accessibility for developer use.

## Projects

---

### ML Reproducibility Challenge 2022: BERT Data Size Probing [Paper](#) [GitHub](#)

- Reproduced and expanded on experiments from an ACL 2022 research paper: [arxiv.org/pdf/2203.09627v1.pdf](https://arxiv.org/pdf/2203.09627v1.pdf)
- Fine-tuned and evaluated BERT models on SuperGLUE tasks across data size intervals, leveraging **HuggingFace** and **PyTorch** to analyze semantic and syntactic knowledge and assess linguistic recoverability.

### Automatic Music Transcription [GitHub](#)

- Developed an automatic music transcription system, obtaining over 90% accuracy for sheet music conversion on four out of seven instruments, using a Convolutional Recurrent Neural Network (CRNN) in **TensorFlow**.
- Adapted the CRNN model from an [OCR paper](#) to handle time series audio data, avoiding pooling on the time axis, incorporating additional convolutional layers, and tailoring tensor dimensions to support variable batch sizes.

## Skills

---

**Programming** Python · SQL · R · Java · C · C++ · JavaScript · HTML · CSS · Bash · Unix · Linux

**Libraries** NumPy · Pandas · Matplotlib · TensorFlow · PyTorch · Scikit-learn · HuggingFace · XGBoost · Flask

**Tools** Git · Jupyter · AWS · PostgreSQL · Spark · Snowflake · Kafka · Tableau · Power BI · Grafana · Airflow

## Education

---

### McGill University

Montreal, Canada

B.Sc. COMPUTER SCIENCE, GPA: 3.53/4.0

Sep 2019 – May 2023

- Awarded the J.W. McConnell Major Entrance Scholarship for outstanding academic achievement.
- Courses: Data Science, Applied Machine Learning, Deep Learning, Natural Language Processing, Artificial Intelligence.
- Extracurriculars: McHacks Hackathon (2020), McWiCS Hackathon (2022), McGill Orientation Leader (2020 – 2022).