Henry W. Leung, Ph.D.

Machine Learning & Data Scientist

henrysky.github.io | henryskyleung@gmail.com | github.com/henrysky | henry-leung-2664b3259

Bilingual in English & Chinese | Canadian & Hong Konger | Toronto, Canada

SUMMARY

Machine Learning and Data Scientist with 6+ years of experience adapting ML techniques to large-scale datasets and 9+ years of software development in Python/C. Expertise in delivering data-driven solutions and insights for real-world applications.

PROFESSIONAL EXPERIENCE

Data Science Institute, University of Toronto Sept. 2023 — Oct. 2024
Data Science Doctoral Fellow

- Developed applications of **Transformers** architecture with a **denoising diffusion** probabilistic head as density function emulator for tabular data to improve non-Gaussian uncertainty estimation. Applicable to uncertainty quantification in finance and healthcare.
- Curated multiple datasets for a 100 TB machine learning-ready dataset hosted on `huggingface`, for the next-generation large ML models in physical science and industry.

University of Toronto Sept. 2019 — Oct. 2024
Graduate Researcher

- Developed a **self-supervised** foundation model in `PyTorch` with a novel token-scalar embedding technique, outperforming XGBoost by 6% in multiple predictive tasks. Presented at NeurIPS and ICML for advancing structured tabular data learning.
- Designed an **unsupervised** encoder-decoder in `TensorFlow` to address a data-rich label-scarce challenge, by reducing data dimensionality by 1,000x through physics-guided techniques and label regression in the low-dimensional latent space.
- Public product release of ML-derived parameters estimation models with uncertainty quantification, achieving over a 10% accuracy improvement for **low signal-to-noise** data and a 100x speedup compared to non-ML pipelines.
- Built a Flask web app integrating an LLM for seamless natural language interaction with scientific models, demonstrating LLM-driven interaction with specialized tools in workflows.
- Implemented supervised **CNNs with dropout variational inference** for spectral data analysis, contributing to one of the **first direct mappings** of the Milky Way's inner structure.
- Developed and maintained open-source software that are **well-documented** with rich set of examples and **thoroughly tested** with more than 80% code coverages, as well as contributed to other open-source projects.

TECHNICAL CONTRIBUTIONS

Below is a summary of key research contributions, highlighting those in machine learning conferences:
NeurIPS (**Oral** in 2023, Collaboration Poster in 2024), **ICML** (Poster in 2024), AI for Astronomy (**Oral** in 2019)

- Research in **ML adaptation** and **foundation models** for scientific and structured datasets with Transformers and diffusion models. ML-driven predictive modeling robust to noise to extract insights from large-scale datasets with applications in finance, healthcare, and physical sciences
- Discovery enabled by **data-driven insights**, including anomaly detection and ML-derived parameter estimation, enabling analyses otherwise difficult with traditional methods.

First/Second author on 9 refereed papers with 580+ citations among 16 refereed papers with 2840+ citations (h-index=11).

EDUCATION

Ph.D., Astronomy & Astrophysics, University of Toronto Sept. 2020 — Oct. 2024
M.Sc., Astronomy & Astrophysics, University of Toronto Sept. 2019 — Aug. 2020
H.B.Sc., Physics & Astronomy, University of Toronto Sept. 2014 — Aug. 2019

SKILLS & SOFTWARE

Programming Languages: Python, C (proficient), C++, SQL (intermediate), Rust (beginner).
Frameworks/Packages: Proficient in PyTorch, TensorFlow, Scikit-Learn, NumPy, SciPy, Pandas, PySpark, Jupyter Notebook, Matplotlib, SQLite, PostgreSQL, Docker, SSH, Git, Bash, Slurm, Hugo, SCSS, Node.js. Some experiences in ArcGIS and SolidWorks.

My open-source software packages used by the community that are well tested with continuous integration and well documented with docstrings and user guides, includes:

- **astroNN** (195 stars) — **Deep learning framework** with `Keras` for astronomical data, supporting `TensorFlow` and `PyTorch`.
- **MyGaiaDB** — **SQLite-based database** package for managing and querying of astronomical data locally.
- **Galaxy10** — **Benchmark dataset** for CNN-based galaxy classification, used in ML research and education.