Winnie Winnie

linkedin.com/in/winniewi

Data Scientist | ML Engineer | livnlearns.com

High impact Data Scientist / ML Engineer with a First Class BSc in Physics and Astronomy and a Silver Medal in Computer Programming. 6+ years of experience productionizing complex ML systems, distributed geospatial pipelines at multi-TB scale. Proven record of reducing cloud costs for large scale ML by 85% & scaling demographic inference from 6% to 80% coverage. Experience in spatiotemporal forecasting, ML and physical systems modeling

SKILLS / TOOLS

Core

Python | Java | SQL

Geospatial / Science

Pandas | numpy | PyG | scipy | Jax
GeoPandas | Astropy | Rasterio | OSM

ML / Deep Learning

Tensorflow | Keras | PyTorch | Sklearn
CNNs | Transformers | GNNs | ANNs

Data Processing

ApacheBeam | PySpark | Ray | Dask

Data Visualisation

Matplotlib | Seaborn | Plotly | Looker

Other Tools

Airflow | Conda | Docker | Git

Cloud Services

GCP : Big Query | Colab | VertexAI
Google Earth Engine (GEE) | GCS
Cloud Function | Dataproc | Dataflow
AWS : Redshift | Beanstalk | Snowflake
ElasticSearch | Dynamo | S3 | Lambda

Generative AI

OpenAI | Langchain | HuggingFace
VectorDB | RAG | Gemini | Llama

EDUCATION

Sheridan Institute of Tech.
ON, Canada (2017-2019)
Computer Programmer (2yr)
Silver Medalist (Top of Class)

York University
ON, Canada (2022-2025)
BSc Physics & Astronomy
First Class with Distinction

EXPERIENCE

Pelmorex Weather Solutions (The Weather network)

Oakville, ON, Canada

DATA SCIENTIST — Mar 2020 – Current

- Promoted to Senior Data Scientist in 2026 after 2+ years of operating in a senior capacity, establishing the title within the company

Mobility Analytics & Geospatial Intelligence

- Designed and productionized **distributed geospatial pipelines** transforming raw GPS telemetry into structured mobility events at multi terabyte scale
- Engineered **trajectory reconstruction** , stop detection algorithms extracting movement signals from noisy sensor data
- Developed **statistical normalization algorithms** (IPW) correcting variable device density, extrapolating sparse data to match census distributions
- Architected **hexagonal spatial layers** (H3) and boundary masked modeling units to standardize heterogeneous telemetry into consistent geospatial representations
- Optimized **large scale spatiotemporal joins** through region/time partitioning, tile indexing etc. significantly reducing cost & latency
- Implemented high dimensional **vector similarity search** (cosine) on aggregated behavioral logs to cluster semantic phenotypes & identify lookalike patterns
- Designed **causal inference frameworks** using synthetic control groups and lookalikes isolating behavioral signals from seasonal noise and trend fluctuations

Spatiotemporal ML & Forecasting

- Built multivariate spatiotemporal forecasts linking weather anomalies to demand signals using **lag features, rolling stats & leakage safe walk forward validation**
- Designed & evaluated **ensemble forecasting** - integrating multiple Numerical Weather Prediction sources with region aware backtesting & verification pipelines
- Led experiments **comparing global vs localized modeling** strategies, rolling origin evaluation & spatially aware splits to quantify regional generalization tradeoffs
- Built **rare event prediction pipelines** (grid outages) using cost sensitive learning, calibrated thresholds, and PR-AUC optimized evaluation under extreme imbalance

Machine Learning Systems & Infrastructure

- Scaled **distributed training** for 100k–1M localized models (Ray + Dataflow), reducing cloud costs ~85% through parallelization & optimized resource allocation
- Contributed to graph based spatiotemporal forecasting prototype (**GNN encoder decoder** , GraphCast style architectures) supporting production weather prediction workflows
- Architected **heterogeneous graph neural network** prototypes fusing online & physical behavior signals & link prediction inferring attributes across disjoint datasets
- Implemented **semi supervised classifiers** predicting demographic attributes for millions of devices, scaling inference from a 6% ground truth to 80% population

RELEVANT COURSES

Object Oriented Programming
Database Design (SQL & NoSQL)
Systems Analysis & Design
Java Enterprise & C# .NET core
Linux System Admin
Multivariate Calculus
Differential Equations
Advanced Calculus
Modern Physics
Techniques in Theoretical Physics

Online

ML (Andrew Ng) - Coursera
ML Engineer - Google
Deep Learning A-Z - Udemy
LLM Engg. RAG & LoRA - Udemy
Graph NN - CS224W Stanford

- Implemented **ensemble classification** (Weighted Voting, Stacking) with imbalance correction, utilizing **weighted sampling** increasing overall accuracy to ~70%
- Helped develop **prediction intelligence** pipelines to monitor model performance, detect **data drift** ensuring continuous model optimization

Generative AI & Semantic Retrieval

- Helped prototype **agentic data interfaces** & **RAG pipelines** with functional tool calling converting natural language into executable SQL, democratizing complex database insights
- Implemented **Named Entity Recognition (NER)** workflows to extract spatial entities and operational parameters from unstructured text streams
- Engineered **semantic search systems** with vector dbs clustering & retrieving high-dimensional data based on latent feature similarities rather than keyword match

FULL STACK DEVELOPER - DATA — Sep 2019 – Feb 2020

- Developed modular RESTful APIs for **real time** KPI tracking, data ingestion, serving high frequency geospatial metrics to downstream monitoring dashboards
- Built automated **data integrity ETL pipelines** auditing external geospatial data providers with spatial interpolation standardizing varying resolutions, detecting data drift - reducing acquisition costs by 30%
- Designed end-to-end data visualization systems, managing the full lifecycle from **schema design** and API endpoints to frontend for decision-support tools

DATA INTERN — 4mo.

- Engineered **metadata-driven orchestration workflows** , automating complex data dependency chains and reducing manual intervention in production pipelines
- Refactored legacy monolithic systems into a **microservices architecture** , improving scalability and fault isolation for data processing services
- Hardened API security protocols against SQL injection and data leaks, implementing **parameterized queries** and strict input validation standards

Canadian Hydrogen Intensity Mapping Experiment (CHIME)

Remote, BC, Canada

INTERN — 8mo.

- Developed a real-time **Convolutional Neural Network (CNN)** to classify Sidelobe Fast Radio Burst (FRB) candidates, achieving millisecond-latency inference to trigger high-fidelity buffer dumps on outrigger telescopes
- Characterized telescope **beam patterns** to differentiate astronomical mainlobe events from **sidelobe artifacts** , significantly reducing false positive triggers in the transient search pipeline
- Analyzed **Radio Frequency Interference (RFI)** environments to optimize model precision-recall trade offs, implementing mitigation strategies for non-astrophysical noise sources
- Benchmarked classifier performance against cataloged FRB events, tuning hyperparameters to maximize sensitivity while maintaining a constrained data transmission budget

github.com/w-winnie
https://livnlearns.com/