SAMUEL JOHNSTON

johnston.samuelj@gmail.com · https://sjjohnst.github.io
Montréal, H2S 3G5

I am a Geospatial Data Engineer who specializes in developing Earth Observation (EO) data pipelines for analytics and machine learning. My practical experience as a data scientist has taught me the value of quality, reliabilty and scalability, which I carry into my work as an engineer.

PROFESSIONAL EXPERIENCE

Data Scientist & Engineer Sep 2024 - present
H2O Geomatics *Kitchener, ON*

- Optimized EO data pipeline which handles 1.5 PB of MODIS reflectance data, achieving a 12x speedup and 16x reduction in memory consumption.
- Developed a data pipeline for production of high-quality training and testing samples for cloud-gap-filling algorithm development, incorporating quality assessment and unit tests, and scaling to parallel workflows.
- Leveraged a distributed High-Performance Computing (HPC) enbironment to deploy a 3D Convolutional Neural Network (CNN) for cloud gap-filling, resulting in the production of a 20+ year daily gap-free Lake Ice Cover dataset spanning 1391 lakes globally.
- Conducting geospatial data harmonization of lake ice cover datasets across mutliple sensor types and data formats, to enable large-scale algorithm validations.
- **Technologies used :** Dask, Xarray, Linux, Python, Zarr

Data Scientist May 2024 - Aug 2024
University of Waterloo *Waterloo, ON*

- Discovered patterns and trends in agricultural choices across the Canadian prairies using the Annual Crop Inventory dataset, such as the growing popularity of grain - legume rotations.
- **Technologies used :** Google Earth Engine, Python, GeoPandas

Visiting Scientist Oct 2023 - Dec 2023
ECMWF *Reading, UK*

- Leveraged High-Performance Computing environment to run experiments on the parameterization of a high-resolution Numerical Weather Prediction model.
- Improved the lake model within ECLand (the land component of the IFS), by parameterizing snow over lake ice to capture its insulating effect, resulting in the successful correction of over-estimated lake ice thickness.
- **Technologies used :** Fortran, Python, HPC

Machine Learning Researcher Jan 2023 - Sep 2023
H2O Geomatics *Kitchener, ON*

- Distributed model training across HPC clusters using a Distributed Data Parallel paradigm to accelerate development by over 60%.
- Developed data ingestion and harmonization pipeline, which combined ERA5 and IMS (spatial ice cover dataset) into harmonized NetCDF format, before partitioning into individual training and testing examples for deep learning.
- Monitored and tracked model development and experiments using Weights and Biases (WandB), allowing for optimized resource allocation and rapid response to failures.
- Developed novel lake ice forecasting algorithm using spatiotemporal transformer networks, outperforming the predominant one-dimensional physics based model FLake at the timing of key ice phenology events.
- **Technologies used :** Python, Linux, HPC

Data Engineer Sep 2022 - Dec 2022
Lakes Environmental Software *Waterloo, ON*

- Built an ingestion and harmonization data pipeline to handle Dynamic World V1 expert labels and Sentinel-1 SAR observations, which produced samples for training downstream machine learning models to perform land-use classification from SAR.

· **Technologies used :** Google Earth Engine, Rasterio, Python, Dask

Machine Learning Researcher Sep 2021 - Apr 2022
Lakes Environmental Software *Waterloo, ON*

· Benchmarked various deep learning architectures (FNO, ConvLSTM, CNN-LSTM) on the simulation of classical lid-driven cavity flow problem, achieving strong performance (RMSE $0.0061 m \cdot s^{-1}$), and resulting in a publication.
· Incorporated mass balance into model training, resulting comparable mass balance magnitudes (10^{-5}) to the Direct Numerical Solution produced by OpenFOAM.
· **Technologies used :** Python, PyTorch, Keras, Dask

Research Assistant Jan 2021 - Aug 2021
University of Waterloo *Waterloo, ON*

· Developed random forest model to predict nutrient loads from land-use characteristics, achieving strong R^2 scores between 0.6-0.8, and resulting in a publication.
· **Technologies used :** Python, Scikit-Learn, Pandas

EDUCATION

Bachelor of Mathematics, Computer Science *2019 - 2024*
Diploma in Sustainability
University of Waterloo

CERTIFICATIONS

DeepLearning.AI Data Engineering Professional Certificate *Sept 2025 - present*
Comprehensive course teaching the framework of data engineering, taught by Joe Reis, author of the Fundamentals of Data Engineering book. Learning the principles of good data architecture, and applying them in hands-on labs to build data systems on the AWS Cloud.
Key Concepts Learned : Data Architecture, Batch & Streaming Pipelines, Infastructure as Code, Orchestration, DataOps [Automation, Observability, Incident Response]
Technologies Applied : S3, Airflow, Terraform, Kinesis Data Streams, PostgreSQL, DynamoDB, REST API

Land in Focus – Basics of Remote Sensing 2025
EO College

TECHNICAL SKILLS

Cloud & Orchestration	S3, Airflow, Terraform
Databases	PostgreSQL, Amazon RDS, DynamoDB
Big Data & Distributed Processing	Dask, Xarray, Apache Spark
Data Formats	Zarr, NetCDF COG, HDFS
Languages	Python, C, C++, Bash, SQL, JavaScript, R
Geospatial	QGIS, STAC, Zarr, GDAL
CI/CD	GitHub Actions
Machine Learning	PyTorch, PyTorch Lightning, Keras, Scikit-Learn

SOFT SKILLS

· Excellent communication skills in English and French
· Working autonomously with minimal direction
· Putting customer requirements first

PROJECTS

Automated Boulder and Cliff Detection Platform (In Development)
Platform which supports local rock climbing development, by allowing users to detect potential boulders and cliffs. Serves 1m resolution rasters of pre-processed DEM data, leveraging edge detection and machine learning to highlight relevant features.
Technologies used : Titiler, Cloud-Optimized GeoTiff, GDAL, Vite, Python.

PUBLICATIONS AND PRESENTATIONS

Samuel J. Johnston, Justin Murfitt, Claude Duguay, and Clément Albergel. 2025. "Addressing Cloud Contamination in Satellite Derived Lake Ice Cover Products." Remote Sensing of Environment, submitted.

Samuel J. Johnston, Justin Murfitt, and Claude Duguay. 2025. "A Deep Learning Approach to Lake Ice Cover Forecasting" GMD, submitted.

Costa Rocha, Paulo A., **Samuel J. Johnston**, Victor Oliveira Santos, Amir A. Aliabadi, Jesse V.G. Thé, and Bahram Gharabaghi. 2023. "Deep Neural Network Modeling for CFD Simulations: Benchmarking the Fourier Neural Operator on the Lid-Driven Cavity Case." Applied Sciences, 13, (5): 3165-65. `https://doi.org/10.3390/app13053165`

Basu, Nandita B., J. Dony, K.J. Van Meter, **Samuel J. Johnston**, and Anita T. Layton. 2023. "A Random Forest in the Great Lakes: Stream Nutrient Concentrations across the Transboundary Lake Basin." Earth's Future, 11, (4). `https://doi.org/10.1029/2021EF002571`