SAMUEL JOHNSTON

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

Je suis ingénieur en données géospatiales spécialisé dans le développement de pipelines de données d'observation de la Terre (Earth Observation), pour les analyses et l'apprentissage automatique. Mon expérience pratique en tant que scientifique des données m'a appris la valeur de la qualité, de la fiabilité et de la scalabilité, que je mets à profit dans mon travail d'ingénieur.

PROFESSIONAL EXPERIENCE

Scientifique/Ingénieur de données — Septembre 2024 - présent
H2O Geomatics — *Kitchener, ON*

- Optimisation du pipeline de données EO qui traite 1.5 PB de données de réflectance MODIS, permettant ainsi un gain de vitesse de 12 fois et une réduction de 16 fois de la consommation de mémoire.
- Développement d'un pipeline de données pour la production d'échantillons de formation et de test de haute qualité pour le développement d'algorithmes pour supprimer la couverture nuageuse, intégrant de la qualité et les tests unitaires, et s'adaptant à des flux de travail parallèles.
- Utilisation dùn environnement informatique haute performance (HPC) distribué pour déployer un 3D Convolutional Neural Network (CNN) pour supprimer la couverture nueageuse, ce qui a permis de produire un ensemble de données quotidien sans lacunes sur plus de 20 ans concernat la couverture de glace lacustre de 1,391 lacs à travers le monde.
- Harmonisation des données géospatiales relatives à la couverture de glace des lacs à partir de plusieurs types de capteurs et formats, afin de permettre la validation d'algorithmes à grande échelle.
- **Technologies utilisées :** Dask, Xarray, Linux, Python, Zarr

Data Scientist — Mai 2024 - Août 2024
Université of Waterloo — *Waterloo, ON*

- Découverte de modèles et de tendences dans les choix agricoles à travers les Prairies canadienns à l'aide de l'ensemble de données Annual Crop Inventory, telles que la popularité croissante de la rotation des culture cérélières et légumineuses.
- **Technologies utilisées :** Google Earth Engine, Python, GeoPandas

Chercheur Invité — Octobre 2023 - Décembre 2023
ECMWF — *Reading, Royaume-Uni*

- Utilisation d'un environment informatique haute performance (HPC) pour mener des expériences sur la paramétrisation d'un modèle numérique de prévision météorologique haute résolution.
- Amélioration du modèle lacustre dans ECLand (la composante terrestre de l'IFS) grâce à la paramétrisation de la neige sur la glace, afin de tenir compte de son effet isolant, ce qui a permis de corriger avec succès la surestimation de l'epâisseur de la glace.
- **Technologies utilisées :** Fortran, Python, HPC

Chercheur en apprentissage automatique — Janvier 2023 - Septembre 2023
H2O Geomatics — *Kitchener, ON*

- Formation de modèles distribués sur des clusters HPC à l'aide d'un paradigme de parallélisme de données distribué afin d'accélérer le développement de plus de 60/
- Développement d'un pipeline d'ingestion et d'harmonisation des données, qui combine ERA5 et IMS (ensemble de données spatiales sur la couverture de neige) dans un format NetCDF harmonisé, avant de les partitionner en exemples individuel d'entraînement et de test pour l'apprentissage automatique.
- Suivi et contrôle du développement des modèles et des expériences à l'aide de Weights and Biases (WandB), permettant une allocation optimisée des ressources et une réponse rapide aux défaillances
- Développement d'un nouvel algorithme de prévision de la glace sur lacs basé sur des réseaux de transformateurs spatiotemporels, surpassant le modèle physique prédominant FLake.
- **Technologies utilisées :** Python, Linux, HPC

Ingénieur de données Sep 2022 - Dec 2022
Lakes Environmental Software *Waterloo, ON*

· Conception d'un pipeline de données d'ingestion et d'harmonisation pour traiter les étiquettes d'experts Dynamic World V1 et les observations SAR Sentinel-1, qui ont produit des échantillons pour l'entraînement de modèles d'apprentissage automatique en aval afin d'effectuer la classification de l'utilisation des sols à partir du SAR
· **Technologies utilisées :** Google Earth Engine, Rasterio, Python, Dask

Chercheur en apprentissage profond Septembre 2021 - Avril 2022
Lakes Environmental Software *Waterloo, ON*

· Évaluation de diverse architecture d'apprentissage profond (FNO, ConvLSTM, CNN-LSTM) sur la simulation d'un problème classique d'écoulement dans une cavité à couvercle, obtenant d'excellente performances (RMSE $0.0061m \cdot s^{-1}$ et donnant lieu à une publication.
· Intégration du bilan massique dans l'entraînement du modèle, ce qui a permis d'obtenir des valeurs de bilan massique comparables (10^{-5} à celles de la solution numériques directre produite par OpenFOAM.
· **Technologies utilisées :** Python, PyTorch, Keras, Dask

Assistant de recherce Janvier 2021 - Août 2021
Université of Waterloo *Waterloo, ON*

· Développement d'un modèle de forêt aléatoire pour prédire les charges et nutriments à partir des caractéristique d'utilisation de sols, obtenant des scores R^2 entre 0.6 et 0.8, et donnant lieu à une publication.
· **Technologies utilisées :** Python, Scikit-Learn, Pandas

ÉDUCATION

Licence en informatique *2019 - 2024*
Diplôme en développement durable
Université of Waterloo

CERTIFICATIONS

Certificat professionnel en ingénierie des données (DeepLearning.AI) *Septembre 2025 - aujourd'hui*
Cours complet sur le cadre de l'ingénierie des données, dispensé epar Joe Reis, auteur du livre 'Fundamentals of Data Engineering'. Apprentissage des principes d'une bonne architecture des données et application de ces principes dans des laboratories pratiques pour créer des systèmes de données sur le cloud AWS.
Concepts clés appris : Architecture des données, pipelines par lots et en continu, infrastructure en tant que code, orchestration, DataOps [Automatisation, Observabilité, Réponse aux incidents]
Technologies Applied : S3, Airflow, Terraform, Kinesis Data Streams, PostgreSQL, DynamoDB, REST API

Land in Focus – Basics of Remote Sensing 2025
EO College

COMPÉTENCES TECHNIQUES

Cloud & Orchestration	S3, Airflow, Terraform
Bases de données	PostgreSQL, Amazon RDS, DynamoDB
Big Data & Traitement distribué	Dask, Xarray, Apache Spark
Formats de données	Zarr, NetCDF COG, HDFS
Langages	Python, C, C++, Bash, SQL, JavaScript, R
Géospatial	QGIS, STAC, Zarr, GDAL
CI/CD	GitHub Actions
Apprentissage automatique	PyTorch, PyTorch Lightning, Keras, Scikit-Learn

COMPÉTENCES TRANSVERSALES

· Excellentes compétences en communication en français et en anglais
· Capacité à travailler de mannière autonome avec un minimum d'encadrement
· Priorité aux exigences des clients

PROJECTS

Plateform automatisée de déetection des rochers et falaise (en cours de développement)
Plateforme qui soutient le développement local de l'escalade en permettant aux utilisateurs de détecter les rochers et les falais potentiels. Elle fournit des raster d'une résolution de 1 mètre à partir de données DEM prétraitées, en utilisant la détection des contours et l'apprentissage automatique pour mettre en évidence les caractéristique perinentes.
Technologies utilisées : Titiler, Cloud-Optimized GeoTiff, GDAL, Vite, Python.

PUBLICATIONS ET PRÉSENTATIONS

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, soumis.

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

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`