Joshua **Greenhalgh**

joshuadouglasgreenhalgh@gmail.com | josh-gree

Experience

Senior Software Engineer — *UK*

IMU Biosciences — *Feb 2024 - Current*

- Architected a **data catalogue and discovery layer** for 100k+ Flow Cytometry samples, unifying scattered Parquet and WSP sources into a structured hierarchy with lineage-aware indexing. This enabled faster query performance and reproducible ML training datasets.
- Designed and deployed **hierarchical gating tree ontologies**, linking semantic labels with raw gating dimensions. This standardisation streamlined cell population tagging, automated QC, and improved interpretability of ML models across diverse panels.
- Introduced **Zarr- and Parquet-based storage strategies** for single-cell intensity data, benchmarking access patterns across S3 and Snowflake. Delivered a hybrid design balancing whole-sample retrieval with cross-sample aggregation for model training.
- Developed a **synthetic data generation framework** for Flow Cytometry, layering immune lineage trees with activation states to produce realistic marker distributions. Used for stress-testing pipelines and validating ML model generalisation.
- Implemented a **QC and monitoring framework** in Dagster that automatically validates staining consistency across donor ”cone” reference samples, ensuring reproducibility and flagging anomalies in experimental workflows.
- Partnered with immunologists to embed **computational QC and metadata standards** (e.g. lineage, phenotype, activation states) directly into pipeline outputs, bridging experimental and computational perspectives.

Senior Data Engineer — *UK*

Mobkoi — *Jan 2022 - Oct 2023*

- Introduced **Infrastructure as Code across the organisation**, migrating all source code to GitHub with version control. Established clear visibility into data processing workflows and deployment practices.
- Rebuilt the company’s data pipelines from **ad-hoc cloud functions to Prefect orchestration** on GCP. Implemented CI/CD processes that increased deployment speed and safety whilst reducing operational overhead.
- Led migration from aggregated partner data to **impression-level data ingestion**, interfacing with AWS ad-serving infrastructure and extracting large volumes from CloudWatch logs. Enabled more granular analytics and increased client confidence.
- Designed and implemented a **new data warehouse** with well-modelled internal operational systems. Created robust change data capture processes that improved data freshness and reliability for the analytics team.
- Overhauled revenue estimation processes with **improved FX handling**, reducing monthly variance against P&L from ~20% to <1%. Significantly improved financial forecasting accuracy.

Senior Data Engineer — *UK*

TransitionZero — *Jan 2021 - March 2021*

- Implemented **Infrastructure as Code using Terraform on GCP**, establishing revision-based deployment workflows that minimised costs from wasted compute and cloud resources.
- Replaced Cloud Composer with **Prefect pipelines**, providing analysts with greater flexibility and control over data transformations whilst reducing operational complexity.
- Built a **CI-based framework** enabling analysts to create and deploy data pipelines independently, accelerating delivery and reducing engineering bottlenecks.
- Developed satellite imagery processing pipelines using **Google Earth Engine** for vision-based ML models supporting climate analytics.

Senior Data Engineer — *Berlin*

HeyJobs — *Aug 2019 - Nov 2020*

- Built a **QA tool for Redshift data warehouse** implementing property-based testing of SQL transformations to validate business invariants. Integrated with Airflow pipelines, DataDog monitoring, and Slack alerts as part of CI process, significantly reducing post-deployment issues.
- Tech lead and architect for redesigning **cost processing infrastructure**, migrating from legacy undocumented SQL to functional Python with full unit test coverage. Reduced marketing overspend by ~5% (hundreds of euros daily savings).
- Implemented **Airflow best practices** with fully incremental, idempotent, and deterministic pipelines. Leveraged AWS Glue catalogue and Spectrum for cost-effective storage, reducing reliance on expensive Redshift for intermediary data.

Data Analyst/Engineer — *Southampton & Berlin (remote)*

Stiftelsen Flowminder (Flowminder Foundation) — *Jun 2017 - Jun 2019*

- **Inter-American Development Bank**: Led project estimating commuting flows and road usage in Haiti’s two largest cities using 3 years of mobile phone location data (10 million subscribers). Built interactive dashboards with PostGIS and React for the Haitian transport ministry.
- **Nepal National Statistics Office**: Designed and delivered a two-week training programme on Python, R, and mobile phone data analysis for government statisticians. Deployed JupyterHub learning environment on GKE for hands-on exercises.
- **GSMA Disaster Response Innovation Fund**: Led project end-to-end including budgets and contractor management. Standardised Flowminder’s **ETL processes using Airflow** across multiple countries, ingesting 50+ million daily mobile phone records. Built post-crisis mobility analysis tool using Kafka and TimescaleDB for humanitarian aid targeting.

Ph.D. student - Computational Modelling

Southampton UK

University of Southampton

Sep 2015 - Jun 2017

- Investigated **deep learning, classical optimisation, and Bayesian approaches** to non-linear inverse problems in x-ray tomography. Focused on reducing reconstruction artefacts using TensorFlow, PyMC3, and parallel computing (OpenMP/MPI) on HPC infrastructure.
- Taught undergraduate courses in Python programming, Linear Algebra, and Calculus. Delivered postgraduate training in parallel programming, advanced Python, and research software engineering best practices (CI, testing).
- Completed three-month internship with Flowminder, received permanent position offer, and withdrew from Ph.D. programme to pursue applied data engineering.

High School Mathematics Teacher

Buxton UK

Buxton Community School

Sep 2013 - Sep 2014

- Taught Mathematics and ICT to students aged 11-18, including GCSE and A-Level examination classes. Led after-school programme introducing talented mathematics students (ages 16-18) to numerical methods in Python.

Research Assistant

Manchester UK

Manchester Metropolitan University

June 2012 - Sep 2012

- Built a **GUI tool for automated solution of large differential equation systems** in astrochemistry using Fortran, C++, and QT. Implemented text parsing and matrix reordering methods (LAPACK) to improve solution efficiency.

Education

MSc in Computational Modelling

Southampton UK

University of Southampton

Sep 2014 to Sep 2015

- Simulation and Modelling (Molecular Dynamics, Monte Carlo simulation, Agent-based modelling, Finite Elements and Stochastic Differential Equations), Numerical Methods, Advanced Computational Methods I/II (OpenMP + MPI), Statistics for Computational Modelling
- Dissertation: Application of deep learning to the problem of signal deconvolution. Technologies used: Tensorflow, Keras.

Qualified Teacher Status

Buxton UK

Buxton Community School

Sep 2013 to Sep 2014

Post Graduate Certificate of Education (Mathematics)

Manchester UK

Manchester Metropolitan University

Sep 2012 to Sep 2013

BSc Mathematics (1st Class Hons)

Manchester UK

Manchester Metropolitan University

Sep 2009 to Jun 2012

- Mathematics Fundamentals (Calculus and Discrete Mathematics), Probability Theory and Statistics, Numerical Methods and Modelling, Number Theory and Cryptography, Numerical Methods for Partial Differential Equations, Computational Methods in Ordinary Differential Equations, Dynamical Systems and Chaos.
- Dissertation: Investigation of computational approaches to problems in Group Theory. Technologies used: Python, GAP, MATLAB

Skills

Languages, Python, R, Javascript, SQL

Data Engineering, Airflow, Prefect, Dagster, Kafka, Docker, Kubernetes

Cloud & Infrastructure, GCP, AWS, Terraform, CI/CD

Databases, Postgres, PostGIS, Redshift, Snowflake, TimescaleDB

Web & APIs, Flask, React

Testing & QA, Pytest, property-based testing