GAVIN THOMSON

Building Scalable and Secure Big Data Pipelines with Apache Spark

@ gavin.r.thomson@gmail.com ☎ 07548887028 ✉ 6 Woolmer Close, Canterbury CT1 3BU
United Kingdom gavin-r-thomson grthomson

EXPERIENCE

Methodologist

Office for National Statistics

May 2023 – Present Manchester

- Higher Executive Officer role
- Researching, implementing and delivering statistical methods as software to meet stakeholder requirements
- Member of the Record Linkage Expert Group with a focus on entity resolution / deduplication for Big Data
- Expertise in coding with Python and SQL, using AWS S3 and Apache Spark for efficient and scalable architectures
- Self-directed design and build of modules for all stages of ETL/ELT, including pipelines for record linkage and analytics
- Adhering to best practice Quality Assurance using GitLab/Github for version control and CI/CD, and Pytest unit test framework
- Co-organiser for acceptance testing, upskilling and code refactoring for migration from on-premises servers to AWS/CDP cloud services
- Presenting with live software demonstrations at stakeholder meetings, cross-governmental events and office-wide calls

Assistant Lecturer

University of Kent

Sept 2018 – July 2023 Kent

- Lectures and Seminar Teaching
- Essay and Exam Marking
- Creating slides, exercises and supporting teaching materials
- Liaising with colleagues to deliver a consistent learning experience in line with intended outcomes and student needs

Courses Taught:

- Logic and Reasoning (Stage 1)
- Ethics in Cognitive Science and Artificial Intelligence (Stage 3)

Research Assistant (UKRI Placement)

Imperial College London – Department of Mathematics

May 2020 – Nov 2020 London

- Multi-institutional collaboration with senior academics at Imperial College London and University of Wyoming
- Formalising and verifying undergraduate mathematics (Real Analysis) using Microsoft Lean Language and Proof Assistant
- Designing and coding content for a Real Number Game
- Codebase at: **Imperial College Github**

MOST PROUD OF

Granted UK Research Council Funding for a PhD programme

Completed BSc (Hons) Mathematics
Entered higher education as a mature student

STRENGTHS

Self-Motivated Learner | Analytical Thinker | Strong Communication Skills

Expert in communicating complex ideas to audiences ranging from non-specialists and undergraduates to experts, in fields from Mathematics and Computing to Linguistics and Philosophy

TECHNICAL SKILLS

Linux/Bash | Python | Apache Spark | AWS Lambda | AWS S3 | GitHub Actions | SQL | Docker | Kubernetes | MS Lean

My undergraduate degree included programming in C and Bash scripting. Since then I am a committed user of Linux operating systems.

In my own time I continuously build knowledge of Cloud and DevOps practices with practical projects using AWS Education, Docker Hub and Kode–Kloud Pro. I maintain personal projects at **my GitHub** .

EDUCATION

Ph.D. Candidate – Foundations of Mathematics

University of Kent

Sept 2017 – July 2023

Thesis title: *Logic, Types and Expressivism*

B.Sc. (Hons) Mathematics (2:i)

University of Glasgow

Sept 2012 – June 2016

• Mathematics and Physics modules • Programming in C Under Linux • Numerical Methods in MATLAB and Mathematica

Final Year Project: *"Quantum Groups and Hopf Algebras"*

PROJECTS

QUAIL (QUality Analyser for Interpreting Linkage)

Office for National Statistics

November 2023 – Present

An ETL pipeline for quality evaluation of data linkage / deduplication tasks (a form of binary classification). Written in Python using PySpark, QUAIL takes as input a dataset of linked records and automatically performs efficient stratified sampling using a novel Bayesian (Markov Chain Monte Carlo) approach to sample size determination. QUAIL also produces precision and recall estimates with bootstrap confidence intervals. My work in this project leverages my knowledge in Spark-native design (tested up to 1 trillion rows).

I am a core developer on this project, contributing to design and build of several modules. This involved liaising with project stakeholders, key contributions to design and implementation of sampling and evaluation modules, mermaid.js for diagramming, and live “show and tell" demonstrations of work-in-progress code at stakeholder meetings. I also presented novel research on Bayesian methods at ONS Bayesian Expert Group and cross-governmental events.

Scalelink

Office for National Statistics

November 2023 – Present

Implementing a cutting-edge algorithm for record linkage based on Multiple Correspondence Analysis and introduced by Goldstein et al. (2017). Unlike standard approaches to large-scale record matching based on a Fellegi-Sunter model, Scalelink is an unsupervised machine learning algorithm which makes no use of computationally expensive Expectation Maximisation. The work of this project has been i) to translate the original Scalelink R package to a Python/PySpark implementation for Big Data and ii) to investigate and implement a method for handling missing data and Null values. I have presented this project at multiple events including a cross-governmental Data Engineering workshop.

I am one of two current maintainers for this project and am responsible for implementing missingness handling, designing and coding elements of the analysis framework, and automating outputs such as data visualisations. The Scalelink Git repository is due to be open-sourced via ONSDigital GitHub in 2025.

GLADIS

Office for National Statistics

May 2023 – November 2023

The *Generalisable Linkage of Administrative Demographic Index Service* is a robust and complete ETL pipeline for linking administrative datasets. The pipeline consists of a sequence of modules which manage data ingest, cleaning and formatting, deterministic and probabilistic linkage processes, and analysis. AWS S3 is used for Data Lake storage and Pandas/PySpark for transformations.

I contributed to multiple modules of this high priority project including SQL code for operations on Spark DataFrames, optimisation of PySpark code for “blocking” or search space minimisation, and Quality Assurance using GitLab merge requests for code review and analysis.

Logic, Types and Expressivism

Arts & Humanities Research Council

September 2017   University of Kent

- Doctoral Research (On hold as of July 2023)

Thesis Outline:

A thesis in formal and philosophical questions in foundations of mathematics. In the philosophical part I develop a general concept of semantic theory applicable to formal logics, natural languages, and programming languages. Drawing on proof-theoretic constructions, I argue that semantic theories are in general independent of the standard semantic concepts of denotation and truth-conditions.

In the technical part I develop a novel categorical semantics for substructural logics presented as sequent calculi. I use a 2-categorical framework to track sequents for which Weakening holds under context transformation. Some elementary properties of the 2-categorical system are proven, making use of Zeilberger’s (2016) and Mellies’ (2015) treatments of functors as type refinements. An novel Hopf algebra model is described, drawing on Chomsky et al. (2023). These results point to a link between logics for defeasible reasoning and systems of typed lambda calculi via a form of Curry-Howard-Lambek correspondence.

COURSES

Inferential Statistics Using R

(University of Kent Graduate Training School)

Algorithms Part I, II

Coursera (Sedgewick)
Java-based course on algorithms and data structures using IntelliJ IDEA IDE

Python Core

JetBrains Academy
Practical Python course using PyCharm IDE

DevOps Learning Path

Ongoing practical learning with KodeKloud Pro
Completed: • DevOps Prerequisite Course
• Terraform for Beginners

MORE ACTIVITIES

ONS Data Linkage Festival

Co-organiser

November 2023   Newport/Online
July 2024   Fareham/Online

Practical and Foundational Aspects of Type Theory – Workshop

Chief Organiser

September 2017   University of Kent

Conference overview online at the *nLab*.