# Callum McMahon

mcmahon.callum@gmail.com ✉
07576 630365 ☎
callummcmahon.github.io

## Education

**University College London** 2019–2020
M.Sc in Machine Learning, Distinction with 89% average
Modules: Supervised learning, Unsupervised learning, Machine vision, Reinforcement Learning, Statistical natural language processing, Inverse problems, Applied ML, Intro to Deep Learning

**University of St Andrews** 2016–2019
B.Sc in Mathematics, First Class Honours
Dean's list for academic excellence awarded for all three years of study

## Work Experience - GSK (November 2020 - present)

**Semantic Information Extraction/Knowledge Graph** (Pytorch, Dask, Self-Supervised Learning, SQL, Kafka, Parquet, Arrow, Docker, Neo4J/Cypher)

Entity recognition/normalisation/relation extraction of biological concepts from internal sources and scientific literature for building a knowledge graph

- Deduplicated internal corpus based on a combination of metadata and locality-sensitive hashing using Spark, as part of a wider effort building an internal 70B GPT model, reducing training times by 10% while maintaining performance.
- Replaced multiple NER models with a unified multi-head classifier, improving performance by 4% F1 while drastically improving inference times.
- Architected and implemented database caching for intermediary pipeline steps, improving development iteration times, and enabling querying for business insights.

**Pathway Enrichment Analysis** (RDF, SPARQL, OWL, Pytorch Geometric)

- Developed a heterogeneous GNN with pathway-dependent readout layer, improving AUROC by 0.1 when benchmarked across both python and production R implementation.

**Multi-Agent Research Assistant** (LangGraph, FastAPI, OpenAI, SQLAlchemy, Technical lead, ↗)

- Integrating tabular data from multiple DBMSs, APIs to various internal predictive models and all unstructured internal+literature documents.
- Synthetic data generation pipeline tackling the cold-start problem of collecting example NL question/SQL query pairs per database/schema

**Histopathology Companion Diagnostics** (Pytorch, openCV, HDF5, Clinical Data)

Using H&E stain slides to predict HRD deficiency for targeted cancer treatments.

- Implemented custom loss for multiclass cell type classification allowing integration of data from multiple sources at varying levels of granularity.
- Replaced hand-crafted background/pen-mark CV filters with U-Net segmentation model, reducing engineer time manually adjusting filters for new out-of-distribution staining techniques/artefacts.
- Removed inference stochasticity by systematically processing all image tiles in batches, accumulating intermediary results. Deterministic results needed for clinical setting, with 1% F1 improvement.

## Projects - University

**M.sc Thesis (Pytorch, Style transfer, 2020)**
Translation of 2D style techniques to 3D meshes, allowing for style embeddings for arbitrary meshes. Generalisation of manifold embedding methods to arbitrary sized meshes using sparse matrices.

**B.sc Dissertation (Pytorch, openCV, 2019)**
Bounding-box detector and icon classifier for the card game Dobble, from data collection, annotation, model implementation, training, and GUI frontend. Super-human performance at over 95% accuracy at 25 FPS.

## Skills

- **Deep Learning:** Pytorch, minimal JAX
- **Scalability:** Spark, Dask, kafka, profiling, testing
- **Visualisations:** Streamlit, tkinter, HTML/CSS/Javascript
- **Linux:** HPC (Slurm), Git, SSH, Vim, Docker, CI/CD, Cloud (GCP), Nix, LaTeX typesetting

## Languages

- **Python (8 yrs experience):** Pytorch, Numpy, Scikit-learn, matplotlib, pandas, openCV
- **R (3 yrs experience):** tidyverse (dplyr, ggplot2)
- **Working knowledge of:** SQL, Rust, Julia, Java
- **Bilingual:** Fluent in both English and French