Ana Pengelly, PhD

LinkedIn | anapengelly.info

PROFILE

Data Scientist and engineer with experience in Machine Learning (mainly NLP) research, engineering and development, passionate about using data for the greater good and transforming society for the better. My main goal is to use data and technology ethically in order to solve global socio-economic and health problems. I have broad international research experience and I have a unique combination of skills in Machine Learning/Data Science and Genetics/molecular biology. I am curiosity driven, and enjoy learning from people with different skill-sets, as well as sharing my skills and collaborating as a proactive team player.

SKILLS

LLMs: **Prompt engineering**, **Langchain, RAG**.

Programming: **Python**, Unix, **SQL**, Git, Cypher (Neo4J), R, basic Html5 and CSS3. Linting: Black, Flake8. Database systems: **DataBricks**, **Pyspark**, Mysql, Pysql, Hive, Hadoop.

Machine Learning: **Tensorflow, Pytorch**, **NLP** (NER, NEL, Transformers), **Computer Vision**, Advanced Time-series Analysis, **XGBoost**, MLflow, KubeFlow.

Engineering: app design and implementation with **Streamlit**.

MLOps: Kubernetes, Docker, MLFlow, Weights and Biases, AWS, Airflow.

Omics: Transcriptomics, Proteomics, Metabolomics, GWAS, Biochemistry, Advanced Statistics.

Mentoring and management: **collaboration**, **enthusiastic team player**, **student mentoring & management,** university teaching, public speaking and article and report writing.

EXPERIENCE

Advanced Data Scientist, BenevolentAI, London– 08/2022-present

•rNN Model training, hyperparameter optimisation and deployment using Weights and Biases.
•Using LLMs to generate text from a knowledge graph to train and evaluate target identification language models.
•Developed a drug discovery copilot using generative AI: LLMs, Langchain, RAG.
•Lead and developed new LLM (GPT-4/Claude) based agents to: augment datasets, assess data quality, create new ontologies & conduct NER.
•Re-built and improved model pipelines to save thousands in computational costs.
•Lead and contributed to the creation of an LSTM-classifier to detect False Positive mentions in text.
•Lead a data-centric evaluation of distinct biological/text datasets in transformer model performance, which lead to the creation of a new leaner and cheaper product.
•Evaluated & improved NER-NEL systems by improving the user feedback service and cleaning source data. This resulted in computational cost savings.
•Enhanced data using ML-NER model inference and added thousands of data points to the internal database.

Data Scientist, Visa Europe & Imperial College, London— 09/2020-07/2022

• Created and conducted a deep-learning analysis on Visa transaction data to model crowdedness and COVID-19 spread in the UK.
• Organised and carried out explainable machine learning on high dimensionality health data (UK Biobank) to understand the impact of environmental exposures on cardiovascular diseases and construct health scores.
• Found that Visa transactions model behavioural traits and are associated with CVD incidence.
• Mentored and led four Machine Learning Imperial College students in their Health Data Analytics and Machine Learning Master's thesis. Three out of four students graduated with honours.

Student analyst, data science lab, visa, London — 05/2020-09/2020

Used Visa transaction data and XGBoost to model COVID-19 mortality in London boroughs.

POSTDOCTORAL FELLOW, THE FRANCIS CRICK INSTITUTE, LONDON — 2016-2019

Functional analysis of transcriptional regulation by dFOXO during stress response.

EDUCATION/CONFERENCES

HarvardX: Fundamentals of TinyML **(verify certificate here)**

NeurIPS 2023— Attendance.

DeepLearning.ai (via Coursera) — Natural Language Processing with classification and vector spaces, 2023, **(verify certificate here)**

Imperial College, London — MSc Health Data Analytics and Machine Learning, 2020, **with distinction**. **Full scholarship from the School of Public Health. (With Distinction)**

Max-Planck Institute of Biochemistry, Munich — PhD Molecular Biology and Genetics, 2015. **European Commission, Marie-Curie FPN7 fellowship. (With Distinction)**

Université Paris Diderot, Paris 7 (now Université de Paris) — BSc & MRes Genetics, 2010. **Excellence-Major fellowship from the French Government. (With Distinction)**

PUBLICATIONS

Developmental diet regulates *Drosophila* lifespan via lipid autotoxins, *Nature Communications,* 2017

Transcriptional repression by PRC1 in the absence of H2A monoubiquitination, *Genes & Development,* 2015

A histone mutant Reproduces the Phenotype Caused by Loss of Histone-Modifying Factor Polycomb, *Science,* 2013

LECTURES

I have given the following lectures to Imperial College, London, students as part of the MSc Health Data Analytics and Machine Learning:

- Introduction to Molecular and Cellular Biology for computational scientists as part of the Molecular Epidemiology.
- Recurrent Neural Networks as part of the Machine Learning module.

PERSONAL INTERESTS

TinyML computer vision models on microcontroller units for environmental/health projects. Contemporary African dance, Contact Improvisation & photography. I am also interested in giving back to society, that is why I created a crowdfunding campaign to fund solar panels for an indigenous family in Talamanca Costa Rica that didn't have any electricity (more info here).

LANGUAGES

- English (mother tongue)
- Spanish (mother tongue)
- French (full proficiency, as went to French School in Costa Rica and lived in France for 5 years)
- German (intermediate level)