Julio Gamboa

Preferred name: Julen Gamboa

GitHub: https://github.com/evoclock *(Note: most applied project work is not public due to sensitive classification and data governance restrictions)*

j.a.r.gamboa@gmail.com | (+44) 7584 122253 | Swindon, United Kingdom

---

Profile

Data Scientist with extensive experience in government operational data science, applied ML, NLP/transformers, forecasting, fraud detection, and scalable pipeline development. Lead developer for multiple production-grade models and analytical assurance pipelines in regulated environments. Experienced in stakeholder engagement, cross-functional delivery, and technical mentoring.

Current Role — Department for Education (DfE), UK

HEO Data Scientist — Analysis & Modelling Assurance (Jan 2024 – Present)

- Lead developer for sensitive risk assessment tools, rebuilding Python pipelines on Databricks with MLflow, hyperparameter tuning, feature stores, automation, and LLM-based reporting integration to support investigations, fund recovery, and fraud prevention.
- Delivered mid-cycle forecasting model improvements, enhancing predictive stability to business-critical forecasting models.
- Led analytical assurance development for the National Funding Formula (Schools Block) for 2025-26, delivering rapid model revisions under major policy changes.
- Led ILR data analysis identifying Qualification Achievement Rate (QAR) risks; analysis directly informed policy revisions and ILR data model changes.
- Developed fraud detection approaches using graph-based network models.
- Mentor/coach to Fast Stream/HEO colleagues; active contributor to cross-government data science communities and inclusivity networks.

Academic Research — Texas A&M University, USA

PhD Candidate – Biology / Computational Genomics (on compassionate leave Jun 2023 – Dec 2024) - Projected completion May/2026

- Developed scalable models linking genomic variation to behavioural differences across rodent strains, addressing gaps in detecting functionally relevant genetic divergence.
- Built pipelines combining sequence clustering (CD-HIT), structural feature extraction, and unsupervised clustering (HDBSCAN) of behavioural data.
- Applied dnaBERT2 transformers fine-tuned with LoRA to generate embeddings capturing complex sequence patterns beyond standard alignments.
- Integrated embeddings with structural features using Dynamic Time Warping to improve detection of subtle inter-strain variation.

Independent NLP & ML Pipeline Development

- Built NLP bias simulation pipelines using synthetic data generation, data augmentation, oversampling, and demographic bias injection for fairness analysis.
- Trained BERT and GPT-2 models; delivered explainability via LIME after resolving multi-class masking challenges with SHAP on HuggingFace models.
- Extended NLP pipelines to Banking77, developing and training a stabilized Mamba-inspired sequence encoder for sequence classification tasks as an alternative method.

Education

- Texas A&M University — PhD Biology (paused 2023–24) Ongoing
- University College London — MSci Cell & Developmental Biology (2:1)

Brief Technical Summary

Python, R, SQL, Databricks (including MLflow, feature stores, production pipelines, cloud deployment, CI/CD), Bash, Git, Azure DevOps.

Supervised (XGBoost, ElasticNet, Boruta) and unsupervised learning (Isolation Forest and surrogate trees, HDBSCAN, BIRCH); NLP (BERT, GPT-2, DNABERT-2, LoRA); explainability (LIME, SHAP); sequence embeddings (Mamba-SSM-like custom model, Dynamic Time Warping); graph networks (graph-tool, NetworkX, igraph).