Cosmo Santoni
Email: santonicosmo@gmail.com | cosmo.santoni@imperial.ac.uk | Tel: +44 7444 403542

Machine Learning Systems Engineer specialising in extreme-scale acceleration and production reliability. I design neural surrogates and deep learning pipelines and data systems that turn hour-scale models into millisecond services.

SELECTED IMPACT

- Reduced agent-based runtime from **24 minutes to 10.5 ms** (+125,000x); enabled interactive what-if analysis.
- Deployed **production ML systems used internationally** for live government decision support.
- Built analytics over **6.89B values** with sub-second queries; reduced time-to-insight from days to seconds.
- **SAGE Award** from UK Government Chief Scientific & Medical Officers for COVID-19 response tooling.

PROFESSIONAL EXPERIENCE

Machine Learning Systems Engineer & Technical Analyst – *Imperial College London, U.K., 2023 – Present*

- Owned end-to-end redesign of national-scale policy simulation infrastructure, replacing 24-minute HPC agent-based computations with sub 10.5ms neural inference (over 125,000x speedup, 99.8% fidelity). Enabled real-time what-if analysis previously impossible at decision-making speed deployed for government policy across multiple countries.
- For typical funded modelling requests (~1M scenarios), cut energy from ~22–43 MWh on consumer CPUs (≈£5.9k–£11.3k) and ~140–281 MWh HPC (PUE-adjusted; ≈£37k–£74k) to ~0.13 kWh with 10.5ms neural inference.
- Architected MINTverse production ML platform processing 100+ scenarios per second on consumer GPU. Designed neural emulation architecture (PyTorch + CUDA + AMP), built data infrastructure (DuckDB handling 574M records with sub-second queries), implemented model serving APIs, and deployed with deterministic versioning. System enabled policy decisions at millisecond latency where hours were previously required.
- Engineered high-throughput analytics layer processing 6.89B data points with sub-second query performance. Eliminated multi-hour data bottlenecks and reduced researcher time-to-insight from weeks to hours through optimised data architecture.
- Built open-source ML infrastructure (MINTverse: segMINT, estiMINT, MINTe, MINTer, and more) reducing research iteration cycles from months to days. Modular architecture adopted by internal teams lacking ML infrastructure capacity.
- Partnered cross-functionally with Global Fund and WHO to translate neural emulation outputs into deployment guidance. Bridged technical performance and policy requirements for international malaria intervention programs.

Machine Learning Research Engineer – *University of Cambridge, Department of Computer Science, UK., & German Centre for Artificial Intelligence (DFKI), DE., 2022 – 2023*

- Engineered neural ODE framework for modelling complex biological systems (tumour growth, epidemic dynamics) with hard physical constraints. Implemented physics-informed loss functions in PyTorch maintaining interpretability while achieving high predictive accuracy on irregular time-series with missing observations.
- Built production-ready deep learning infrastructure handling multi-scale dynamical systems. Designed training pipelines for sparse temporal data, implemented uncertainty quantification through Bayesian inference, and integrated sensitivity analysis ensuring model predictions met clinical safety requirements.
- Drove architecture decisions for Neural Universal Differential Equations balancing expressiveness vs interpretability trade-offs. Framework enabled modelling previously intractable biological systems where pure data-driven approaches failed physical consistency checks.

Research Assistant – *COVID-19 Real Time Modelling, Imperial College London, U.K., 2021 – 2023*

- Owned real-time inference pipeline delivering daily briefings to UK government under extreme time pressure (<24hr from raw data to policy recommendations). Particle MCMC implementation in C++/R processed national-scale epidemic data with sub-hour turnaround when decisions affected millions. Zero tolerance for errors—every output directly informed UK lockdown policy during national emergency.
- Maintained mission-critical modelling software (sircovid, spimalot, MCState) under active use by UK SAGE during pandemic. Managed breaking changes, backward compatibility, and emergency bug fixes while ensuring reproducibility across distributed teams working under crisis conditions. Software became foundation for international epidemic response efforts.
- Engineered automated reporting infrastructure generating statistical analyses, visualisations, and forecasts with <24hr latency end-to-end. System ran continuously for 18+ months supplying weekly briefings to UK Chief Scientific Advisors with zero missed deadlines.
- Received SAGE Award for Modelling and Data Support from UK Government Chief Scientific and Medical Officers recognising technical contributions enabling evidence-based crisis response at national scale.

TECHNICAL SKILLS & LANGUAGES

- **Languages:** Python, R, Julia, C++, C#, SQL, Bash (Linux)
- ***ML / DL Systems:*** *PyTorch, TensorFlow/Keras, XGBoost, Optuna, NumPy, Pandas*
- ***Data & Storage:*** *DuckDB, HDF5*
- ***HPC & Acceleration:*** *CUDA, Automatic Mixed Precision (AMP); CPU/GPU clusters*
- ***Tooling:*** *Git/GitHub, CI/CD (GitHub Actions), Linux*

EDUCATION

- **Full-time PhD in Applied Mathematics,** *Imperial College London, U.K., January 2025 – December 2027*
  Thesis: "*Large-Scale Acceleration of Real-Time Agent-Based System Simulations with Neural Surrogates and Graph Neural Networks"*

- **MSc. Epidemiology (Merit),** *Imperial College London, U.K., 2020 – 2021*

- **BSc. (Hons.) Mathematics with Economics (2:1),** *Aston University, U.K., 2015 – 2019*

PUBLICATIONS & REFERENCES

Publications:

- *Morgenstern, C., et al. (including Cosmo Santoni) "Severe acute respiratory syndrome (SARS) mathematical models and disease parameters: a systematic review and meta-analysis"* ***The Lancet Microbe***
- *Imai, N., Rawson, T., et al. (including Cosmo Santoni.) "Quantifying the impact of delaying the second COVID-19 vaccine dose in England: a mathematical modelling study"* ***The Lancet Public Health***
- *Perez-Guzman, P. N., Knock, E., et al. (including Cosmo Santoni.) "Epidemiological drivers of transmissibility and severity of SARS-CoV-2 in England."* ***Nature Communications***

Manuscripts under Review:

- *McCabe, R., et al. (including Cosmo Santoni) "The impact of ambiguously reported epidemiological parameters for infectious disease modelling and recommended best practices"* ***The Lancet Infectious Disease***
- *Cosmo Santoni., et al. "Deep Neural Universal Differential Equations: A Novel Approach for Tumour Volume Growth in Complex Mathematical Systems"* ***Nature Machine Intelligence***

Published Software & Tools:

- *MINTverse: A modular Python/R toolkit—DuckDB data layer + PyTorch emulators + XGBoost for real-time prevalence/case forecasting. Open-source. Production ready.*
- *RADAR (in-dev): risk diagnostics & quantile calibration for time-series emulators (private; results on request)*
- *ALETHIA: Metacognitive control & selective acceptance for LMs*
- *KAIROS: Weakly-supervised selector with target coverage + calibration; full audit artifacts*
- *Epireview: A tool to obtain the latest data, figures and tables from the Pathogen Epidemiology Review Group (PERG). PERG is an internationally recognised World Health Organization collaborative collective.*
- *Sircovid: Tools for Bayesian analysis of stochastic models using adaptive Metropolis-Hastings and particle MCMC.*
- *Spimalot: The models in this package can be used to estimate key epidemic parameters and predict the course of the epidemic under different intervention scenarios.*
- *MCState: Parameter inference for stochastic, compartmental models from data, using Monte Carlo methods.*

CONFERENCES & PRESENTATIONS

- **Epireview: Hands-on Workshop for Public Health & Epidemiology Researchers,** *Infectious Disease Modelling Conference, Bangkok, Thailand, November 2024*
- **Applying Neural Network Emulation to Assess the Impact of Pyrethroid-Pyrrole Bed Nets on Malaria in Africa,** *9th International Conference on Infectious Disease Dynamics, Bologna, Italy, November 2023*
- **Investigating Parameterisation and Inference Trade-Offs in Stochastic and Deterministic Epidemic Models**, *9th International Conference on Infectious Disease Dynamics, Bologna, Italy, November 2023*

ACADEMIC SERVICE & VOLUNTARY WORK

- **Curator, Amphibian & Malaria Collections**, *Museum of Life Sciences, King's College London*, U.K., 2025 –
- **Departmental MRC GIDA Seminar Series Co-Organiser**, *Imperial College London, U.K.*, 2023 –
- **MSc. Epidemiology Graduate Teaching Assistant**, *Imperial College London, U.K.*, 2021 – 2022
- **Lay Grant Reviewer**, *University College London & Parkinson's UK*, U.K., 2019 – 2022