Charles T. Gray, PhD

Applied Mathematician | Analytics & Data Engineer | Statistical Software Specialist
Canberra (ACT) · Australian Citizen
GitHub – softloud | Publications | ctgray.inbox@gmail.com | Data storytelling gallery

Profile

Applied mathematician and statistical engineer with over a decade of implementation experience who **thinks in graphs**. I specialise in **entity resolution** and the design of **graph-structured data systems**, ensuring that real-world entities remain coherent, deduplicated, and traceable across large, distributed, heterogeneous data stores. Whether reconciling identities, modelling uncertainty with Bayesian DAGs, or designing lineage-sound analytical workflows, I approach data problems through **structural reasoning first**.

My mathematics honours, completed with the Australian Algebra Group, trained me in general algebra, topological chaos, model theory, and complexity — foundations that now directly inform how I formalise entity identity, equivalence, and constraints across data systems. My PhD in statistical science, large-scale data simulation, and research software engineering gave me deep expertise in validation, reproducibility, and test-driven analytical development.

I design fast, reliable, testable systems from ingestion to decision, specialising in:

- **entity resolution** across distributed and graph-based data stores
- **graph-based data modelling** and semantic identity architecture
- **algorithmic optimisation** for efficient search, storage, and enrichment
- **workflow orchestration** with dependency and test architecture
- **statistical modelling and data storytelling** that remains verifiable and explainable
- **semantic modelling and observable validation layers** that keep analysis aligned with intent

I build systems that can be trusted — auditable, constraint-driven, and observable — and I advocate for collaborative environments where minimum requirements, documentation, and rigorous testing are standard practice.

Professional Experience

Independent Data Architect, Strategist, Engineer | Australia | Denmark | UK (2018 – Present)

dbt · Dagster · Python · R · SQL · Snowflake · AWS Redshift · Fivetran · Tableau · PowerBI · pandas · scikit-learn · numpy · ggplot2 · Shiny · Git · GitHub · GitLab · CI/CD · ESG · Agile · JIRA · Monday · Clickup

Design and implement **entity-resolution frameworks** within large analytical ecosystems, ensuring real-world entities are reconciled across disparate sources and preserved through graph-structured transformations. Architect graph-based data models and validation frameworks for lineage reliability and semantic consistency across distributed systems. Lead data migration strategy, ESG governance, and data quality. Proficient in interpretable machine learning (scikit-learn, classical ML models), with a strong focus on uncertainty, data leakage mitigation, and model validation. I approach ML through mathematical structure and reproducibility — ensuring models remain testable, explainable, and aligned with operational constraints.

Build end-to-end learning pipelines integrating SQL and Python with dbt and Dagster, embedding automated tests for lineage validation, identity coherence, and auditability. Create internal documentation frameworks and CI/CD validation layers surfacing pipeline and entity-level health. Mentor data teams in FAIR principles, ESG reporting, and workflow design. Clients included a national telco, neuromarketing consultancy, video games studio, and consultancies such as PwC.

Impact: Spearheaded data migration, governance validation, and analytical traceability across high-stakes environments handling millions of daily events.

Research Software Engineer / Data Scientist · Academic & Policy Institutions (Australia & EU, 2012 – 2020)

R · ggplot2 · Shiny · Targets · academic/policy writing · Git · GitHub

Developed **entity-aware FAIR data pipelines** and reproducible analytical systems using Targets, with explicit modelling of identity, uncertainty, and equivalence across biomedical and ecological datasets. Led creation of research-grade R software packages for simulation, estimation, and validation. Collaborated with multidisciplinary teams to resolve data interoperability issues under ethical and regulatory constraints. Published methods in reproducible data science and open-source tooling. Extensively lectured and coordinated subjects including **graph theory**, linear programming, linear algebra, **automata**, and generalised linear models. Organisations included the Walter and Eliza Institute of Medical Research, University of Melbourne, Evidence Synthesis Hackathon, and La Trobe University.

Impact: Delivered validated large-scale statistical learning simulations and reproducible entity-aligned pipelines adopted in peer-reviewed research and policy reporting.

Music Career to STEM Transition (1997 – 2017)

Pianist → Data Scientist: Transitioned from music pedagogy and performance to mathematics and data science beginning 2011, developing strong communication, discipline, and ensemble problem-solving skills. Supported myself through mathematics training as an independent musician. Specialised in music theory and music as formal systems, which informed my pivot to mathematics. Performed as emerging mathematician of interest at the opening ceremony of the 2018 Heidelberg Laureate Forum.

Education & Technical Projects

Qualification	Institution	Focus & Dissertation
2020 **PhD (Data Science)**	La Trobe	Reproducible Analytics · FAIR Data Science · Statistics · *Toward a Measure of `code::proof`: A toolchain walkthrough for computationally developing a statistical estimator*
2015 **BSc (Hons I, Mathematics)**	La Trobe	Mathematical Graph Theory · Quasi-primal Algebra · thesis fully published in *Order — The Homomorphism Lattice Induced by a Finite Algebra*
2007 **BA/BMus**	UoM	Critical Theory & Music · *Disney and Mulan: Familiar strangers in the Disneyian Orient*

Academic distinctions: Graduated with First Class Honours in Mathematics (87% aggregate), with High Distinction results across **universal algebra**, **topology and dynamics (mathematical chaos)**, **advanced calculus**, **model theory**, and **computational complexity**. Awarded multiple research grants; published a new mathematical result in quasi-primal algebras; invited to join the Australian Algebra Group.

Selected Invited Talks & Outreach

- **Aarhus Comedy Club (2025)** - *If shrimp, then towel* - Standup comedy on data system governance
- **DBT Meetup Copenhagen (2024/2023)** – *A dbt Grrl in a dbt World* — Invited speaker
- **Copenhagen useR Group (2024)** – *Animating a Melody as a Mathematical Object* — Invited speaker
- **PyData Copenhagen (2023)** – Presented analysis of open video-game player data
- **Heidelberg Laureate Forum (2018)** – *Tesselations from Eppalock to Heidelberg* — Featured early-career mathematician; Opening Ceremony keynote
- **Code Like a Girl (2018)** – *Data Storytelling with Charles Gray* — national speaking tour
- **World Science Festival (2018)** – Keynote.
- **Melbourne Writers Festival (2018)** – Chair
- **R-Ladies (2018–2021)** – Presenter and organiser across multiple international chapters
- **AMSI Choose Maths Campaign (2018)** — National speaking tour

Selected Publications

- **Davey, B.A., Gray, C.T., Pitkethly, J.G.** (2018). *The Homomorphism Lattice Induced by a Finite Algebra*. **Order**, 35(2), 193–214.
  Published Honours thesis; presents a new mathematical theorem in quasi-primal algebra.

- **Gray, C.T., Marwick, B.** (2019). *Truth, Proof, and Reproducibility: There's No Counter-attack for the Codeless*. In **Research School on Statistics and Data Science** (pp. 111–129). Springer.
  Framework for reproducible computational research and code-validated inference.

- **Gray, C.T.** (2019). *code::proof: Prepare for Most Weather Conditions*. In **Research School on Statistics and Data Science** (pp. 22–41). Springer.
  Tools and principles for testable statistical workflows and computational validation.

- **Haddaway, N.R., Gray, C.T., Grainger, M.** (2021). *Novel Tools and Methods for Designing and Wrangling Multifunctional, Machine-Readable Evidence Synthesis Databases*. **Environmental Evidence**, 10(1), 5.
  Methods for large-scale data integration, structuring, and semantic interoperability.

- **Haddaway, N.R., Grainger, M.J., Gray, C.T.** (2022). *Citationchaser: A Tool for Transparent and Efficient Forward and Backward Citation Chasing in Systematic Searching*. **Research Synthesis Methods**, 13(4), 533–545.
  High-impact R package and computational workflow for graph-based citation searching.

- **Haddaway, N.R., Feierman, A., Grainger, M.J., Gray, C.T., et al.** (2019). *EviAtlas: A Tool for Visualising Evidence Synthesis Databases*. **Environmental Evidence**, 8(1), 22.
  Interactive graph-based visualisation tool for large heterogeneous evidence datasets.