Natalie Zelenka

natalie.zelenka@bristol.ac.uk | natalie-zelenka.com

Research Interests

- **Applied Data Science** - using statistics, AI, and mathematical modelling to solve research problems from diverse disciplines across science and arts.
- **Research Improvement** - reducing the barrier to creating better research and a more accessible research environment through projects in reproducibility, Open Research, and AI ethics.
- **Molecular Biology** - investigating mechanisms between phenotype and protein function, and creating the Open resources needed to make this possible.

Research Experience

University of Bristol

Data Scientist, *Jean Golding Institute* 2018 - present

- Served the community of researchers who use or want to use data science or data intensive methods in their research, through community-building, performing Data Science and Software Engineering, and Data science competency-building
- Supported management through providing data science expertise at meetings, in grant applications, and in building new partnerships.

PhD Student, *Bristol Centre for Complexity Sciences* with Julian Gough's *Computational Genomics* group 2013 - 2021

- Investigated the relationship between genotype and phenotype using data from publicly available tools and databases such as FATHMM, DcGO, the gene expression atlas, SUPERFAMILY.
- Contributing to collaborative group-led research and materials, such as the SUPERFAMILY update and Proteome Quality Index papers.

National Institute of Informatics, Tokyo

Research Intern working with Michael Houle 2015 - 2016

- Applied high-dimensional clustering and outlier detection methods to genotype data to find genetic outliers in human populations

University of Manchester

MMath/Phys Master's projects Masters projects in: 2011 - 2012

- Stochastic Turing patterns of predators and prey on Complex networks
- Modelling the spread of infectious diseases over networks: simulating the spread of bubonic plague in the historic village of Eyam

Research and Software Engineering Projects (selected)

(See "Awards and Grants" for further projects.)

Applied Data Science

Machine-learning aided translation (PI: Lucas Nunes Vierra) 2018-2021

- Provided statistical consultancy on hierarchical modelling of the role of self-revision in Machine-learning aided translation
- Provided statistical consultancy on hierarchical modelling of AI-assisted translations of science fiction in multiple languages

Linear Optimisation of Research Excellence Framework (REF) Outputs 2018
(for: University of Bristol Senior Management)

- Created a command line interface, which optimises which REF outputs should be submitted based on the results of scoring exercises, and creates automatic reporting of chosen outputs.

Research Improvement

Data Hazards, *Co-PI* 2021 - present

- Created a community-led Open Source classification of 11 different types of Data Ethics hazards, in the form of hazard labels (similar to COSHH chemical hazard labels).
- Ran two workshops to collect community feedback on the best way to present and use the Hazard Labels.
- Used as part of Data Science seedcorn funding, in undergraduate lessons (in Bristol and the US),

Multiverse analysis, *Co-PI* 2020 - 2021

- Created a prototype Python package to simulate the "multiverse" of different plausible analytical choices for Generalised Linear Models in Python.

Computational Biology

Ontolopy 2020

- Conceived of and created a small Python package for extracting information from OBO ontologies in Python

Filip 2018

- Conceived of and created a method that applies to structure-based protein function predictions, removing predictions where the protein is never found in the tissue of interest.

Teaching, workshops, coaching, and mentoring (selected)

Ask-JGI Data Science Support 2018 - present

- Answered over 200 queries to the "Ask JGI" (ask-a-data-scientist) mailbox, ranging from signposting to detailed code-review or debugging which supported staff around the University in applying a variety of data science techniques in their work, from neural networks to PCA to multilevel modelling.

Data Science Specialist advising on BA Talent Development Grant December 2022 - present

- Providing data science support, bespoke training, and competency-building to develop the NLP skills of a senior researcher.

Data Science Advisor for Widening Participation Intern June - August 2021

- Provided data science supervision to an undergraduate computer science intern, supporting them in applying NLP to Arabic and English versions of Hezbollah political speeches.

Carpentries instructor 2019 - present

- Carried out official Carpentries Instructor training to develop pedagogy specific to teaching programming, and ran a number of Carpentries events including a week of events as part of Strathmore University's (Nairobi) *Data Week* 2019.
- Developed and repeatedly led a Carpentries-style lesson for Reproducible and Open Analysis in R.

ACRC training 2019 - present

- Supported training in R, Python, Linux, and Deep Learning, provided by the *Advanced Computing Research Centre* (ACRC).

UKRN training 2020 - present

- Provided training on behalf of the UK Reproducibility Network (UKRN), including supporting their train-the-trainer initiative for Open Source.

Pygame teaching 2014

- Designed a 12 week after-school course syllabus which taught 11-16 year olds to program in Python on Raspberry Pi's, and coached them through designing and creating their own video game using `pygame`.

Community-building (selected)

Data Ethics Club Jan 2021 - present

- Co-founded a discussion group for data science and ethics, building an international community of over 300 people.
- Ran the group through GitHub to maintain a list of suggested readings and credit contributors.
- Ran over 20 well-attended events in the first year, including a movie screening and an invited talk, as well as paper discussions.

Weekly co-working March 2020 - September 2021

- Ran weekly drop-in calls during national lockdowns to maintain and build the Data Science community at the University.

Turing Data Study Group August 2019

- Liaised with the Alan Turing institute to co-organise and facilitate the first Data Study Group outside of London for over 50 Data Scientists from around the UK, working on 5 challenges.
- Worked with challenge owners (companies and researchers) to ensure challenges were well-posed, and data was suitable for AI challenges.

Data Visualisation Interest Group 2018-2020

- Worked with a small team to organise monthly data visualisation talks and training.

Our Data Bristol hack-a-thons 2018-2020

- Both contributed to as a participant and facilitated a series of four themed hack-a-thons (e.g. "Art and Culture", "Transport") for Bristol City Council, supporting a mixture of civic, design, and data science attendees to use Open data to, e.g.:
  - Apply style transfer to Museum of Bristol archive images
  - Create an app that found the most eco-friendly route between to a location

Data Competitions 2018 - present

- Liaised with charities and companies to create the Jean Golding Institute's annual data science and beauty of data competitions, including cleaning data and designing the challenge brief.

Awards and grants

Data Hazards: Train-the-trainer 2022

PI (Co-I: Nina Di Cara), *Research England*, Enhancing Research Culture Funding £20,000

- Create train-the-trainer materials to support data science and AI researchers in considering the ethical and societal outcomes of their work.

Multiverse analysis 2020

Co-PI (Co-PI: Marcus Munafo), *Elizabeth Blackwell Institute*, Health Data Science Strand Seedcorn Award £15,000

- Create a prototype Python package for multiverse analysis.

What Aren't You Seeing? 2020

Co-I (PI: Greg McIrnery, Co-I: James Tripp), *Alan Turing Institute*, Tools, Practices and Systems Award £100,000

- Contributing to the "technical fieldwork" strand: observing what kinds of decisions data scientists make by analysing their data visualisations and what decisions they made to get to that point (using the Grapho R library).

Flood Risk 2019

Co-I Researcher (PI: Laurence Hawker) *Brigstow Institute*, Seed corn award £5,000

- Developed prototype interactive maps which were embedded inside a survey, as part of a project to investigate how to best represent flood data for NGOs involved in flood response.

Chemspeed Data Capture, 2019

Co-I (PI: Natalie Fey, Co-I: Ella Gale), *Jean Golding Institute*, Seed corn award £5,000

- Developed a strategy for capturing and sorting the data collected by the Chemspeed digital chemistry robot, enabling usability for future Machine Learning methodologies, leading to a larger grant.

MapHis 2019

Co-I Researcher (PI: Yanos Zylberberg), *Brigstow Institute*, Seed corn award £5,000

- Applied image recognition algorithms to recognise industrial chimneys on historical maps, leading to a larger grant.

PURE Data Science Competition 2017

2nd place (team-mate: Emily Pole) *Jean Golding Institute, UoB* £250

- Applied NLP and clustering to research output abstracts to identify disciplines, compare to School and Faculty assignments, and cre ate an interactive visualisation.

PhD Student Seed corn Award 2017

Co-PI (Co-PIs: Louis MacGregor, Emily Pole), *South West Research Cooperative, GW4* and *Brigstow Institute, UoB* £375

- Worked with a small local charity (PRISM) to write a successful grant, carry out a survey, and analyse the results.
- The results were presented at Bristol City Hall and used by the benefiting charity in decision-making and to evidence need in funding applications.

New Enterprise Competition 2016-2017

Successful (rounds 1 and 2) (co-founders: Louis MacGregor and Emily Pole), *Research and Enterprise Development* £600

- Successfully wrote and pitched a business plan for a data science consultancy run as a social enterprise.

Publications

Note: Last name changed from Thurlby to Zelenka in 2021

1 İbrahim Sönmez, Cinta Folch, Nicolas Lorente, Rigmor C Berg, Natalie Thurlby, and Axel J Schmidt. Influence of internalised homonegativity on sexual risk behaviour of men who have sex with men in spain. *Sexuality & Culture*, pages 1–19, 2021.

2 Julian Gough, Jan Zaucha, and Natalie Thurlby. Determining phenotype from genotype. June 4 2020. US Patent App. 16/070,920.

3 Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, and others. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23, 2019.

4 Matt E Oates, Jonathan Stahlhacke, Dimitrios V Vavoulis, Ben Smithers, Owen JL Rackham, Adam J Sardar, Jan Zaucha, Natalie Thurlby, Hai Fang, and Julian Gough. The superfamily 1.75 database in 2014: a doubling of data. *Nucleic acids research*, 43(D1):D227–D233, 2015.

5 Jan Zaucha, Jonathan Stahlhacke, Matt E Oates, Natalie Thurlby, Owen JL Rackham, Hai Fang, Ben Smithers, and Julian Gough. Ap roteome q uality i ndex. *Environmental microbiology*, 17(1):4–9, 2015.

Talks (selected)

Data Hazards initial results Tech Ethics Bristol Meetup: AI ethics best practices and the future of innovation 2021

Open Code and Software CARQI (Centre for Academic Research Quality and Improvement) Launch Event 2020

Education

PhD Computational Biology *University of Bristol* 2021

- Nominated for a thesis prize (awaiting prize decision), for interactive reproducible thesis in Jupyter Book with illustrations.
- Taught masters year in Complexity Sciences (2012-2013)

MMath/Phys Undergraduate masters in Mathematics and Physics (Joint Honours), 2:1, *University of Manchester* group 2012

Technical Skills

- 7+ years experience developing research software (scripts, CLI tools, packages) in Python, using a range of packages, e.g. `scikit-learn`, `statsmodels`, `pandas`, `numpy`, `pytest`, `numba`, `multiprocessing`, `sphinx`, `jupyter-book`, `plotly`, `pytorch`, `tensorflow`, `venv`
- Git, GitHub, GitHub Actions
- 3+ years experience analysis and statistical modelling in R using the tidyverse ecosystem
- Some experience with AWS, Matlab, C++, Godot

Non-Science Interests

I play the cajon and ukulele in a pirate-punk sea shanty band, attempt to build video games in impossibly short time-frames during game jams like Ludum Dare, and occasionally draw things.