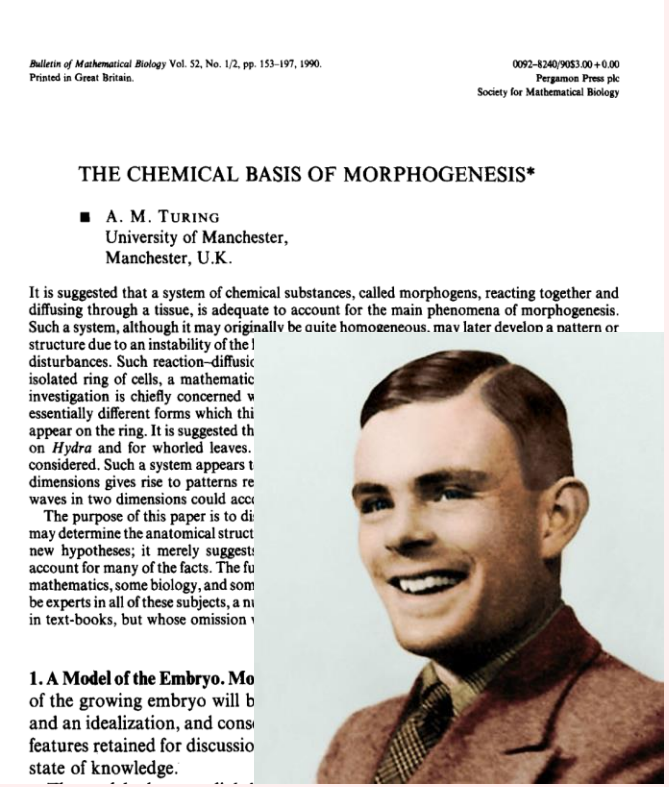
How I tricked myself into writing my thesis

by making it as ethical and reproducible as I could

Dr Natalie Zelenka,
Data Scientist,
University of Bristol

My PhD…



Bulletin of Mathematical Biology Vol. 52, No. 1/2, pp. 153–197, 1990.
Printed in Great Britain.

0092-8240/90$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

THE CHEMICAL BASIS OF MORPHOGENESIS*

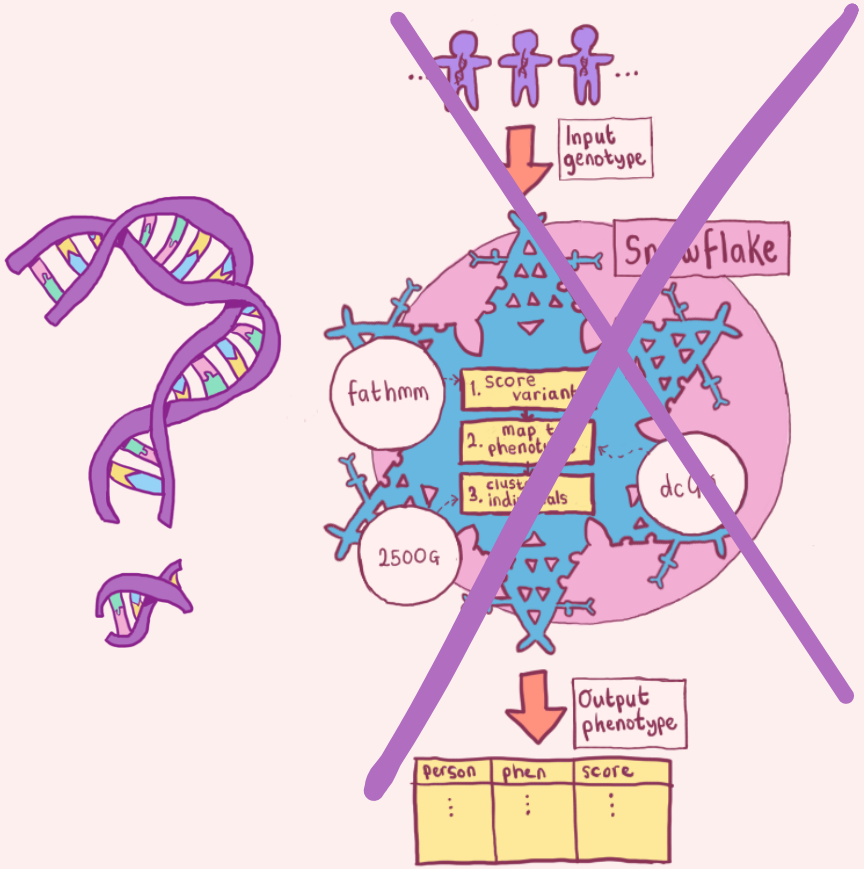
A. M. TURING
University of Manchester,
Manchester, U.K.

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the

1. A Model of the Embryo. Mo

MANCHESTER 1824

The University of Manchester



University of BRISTOL

But…

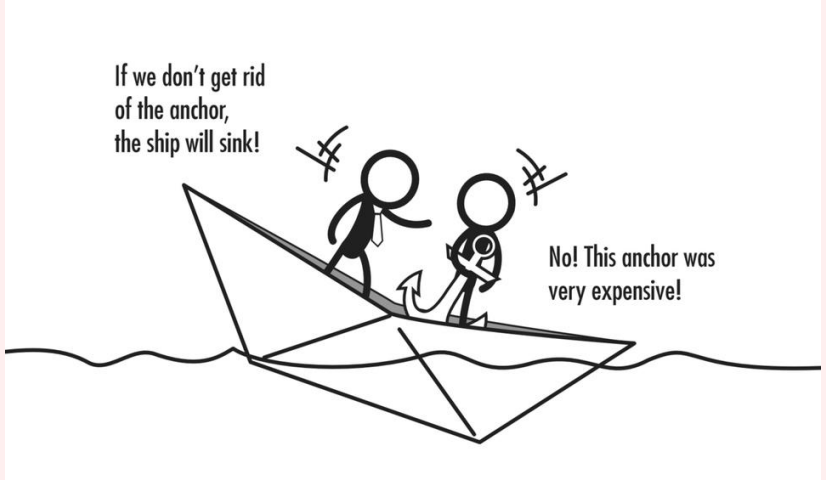
- The predictor I built didn't work
- It was impossible to say why
- It didn't seem like it would be interesting or useful for anyone to read.
- I was starting to find the whole concept a bit eugenics-y
- Both my supervisors left the university.
- I didn't write it up as I was going.

A glimmer of hope?

1. The sunk cost fallacy was preventing me from quitting.

2. During my PhD, I had the chance to learn about and participate in Reproducibility and Open Science.

3. There was one small part of my research that I was proud of:
   - Contributing back to improve some of the open-source resources that I'd used in my work.



Structured procrastination

structuredprocrastination.com

“. . . anyone can do any amount of work, provided it isn't the work he is supposed to be doing at that moment.”

-- Robert Benchley, in Chips off the Old Benchley, 1949

Reproducibility

- Literate programming with R and Python code using `JupyterBook` (100% inspired by The Turing Way!)
- Version control of the code and text using `MyST Markdown` and `jupytext` on GitHub.
- Testing with `pytest`
- Tests and website built using `GitHub Actions`.
- A simple reproducible environment with `requirements.txt` created with my pinned versions using `pigar` and `renv` for R.

8.3.2. Mapping to UBERON

Mapping from samples to Uberon tissue required the development of a small Python package Ontolopy. To create input to this package, informal tissue names (e.g. blood, kidney) were taken from the experimental design files (or the human sample information file for FANTOM) to create a map of samples to informal tissue names. For FANTOM, the FANTOM ontology could also be used to create a more fine-grained mapping of samples to tissues based on FANTOM sample identifiers and/or cell type (CL) identifiers.

HPA The HPA samples were mapped using exact matches to Uberon names. Three types of sample did not have exact matches: *transformed skin fibroblast*, *suprapubic skin*, and *ebv-transformed lymphocyte*. I manually mapped *suprapubic skin* to `UBERON:0001415` *Skin of pelvis*, and excluded the other two (corresponding to excluding 869 samples).

HDBR For HDBR, tissue names from the "organism part' column of the column data file were matched to Uberon names and synonyms from the Uberon extended ontology. The 96 unmatched terms corresponding to mixed brain tissues and brain fragments were defaulted to the more general Uberon Brain term.

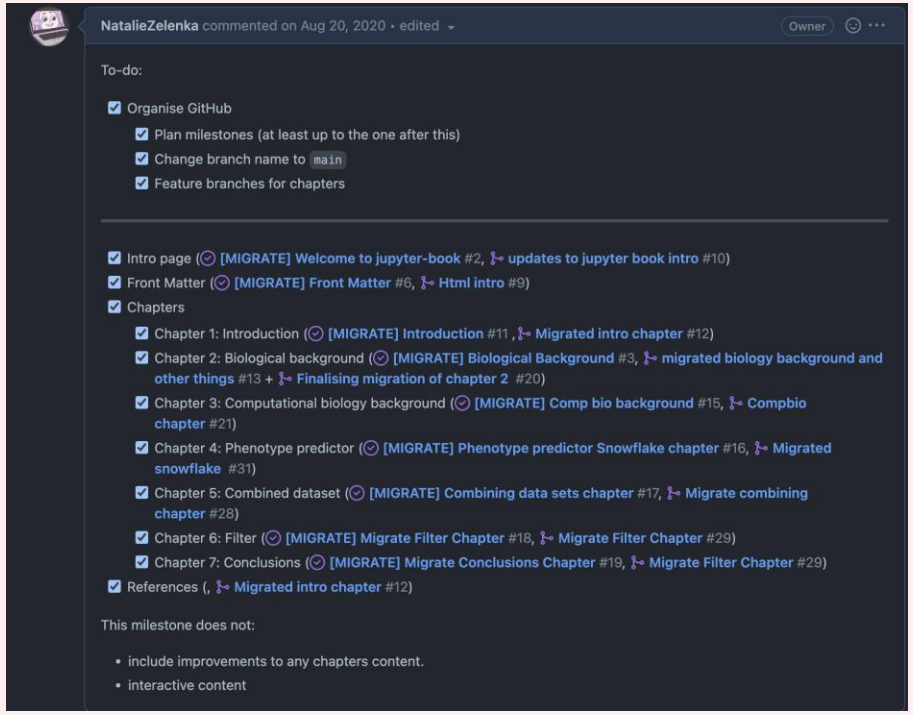
FANTOM Since an experimental design file could not be obtained for FANTOM via GxA, additional sample information was obtained via the FANTOM5 website, namely the human sample information file and the FANTOM5 ontology.

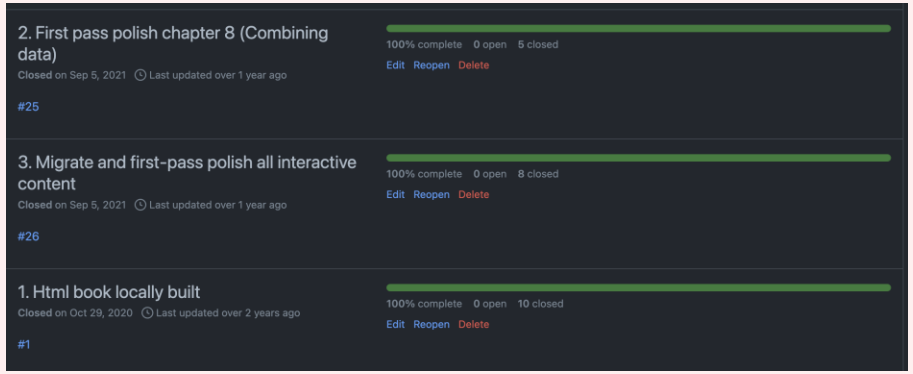
FANTOM also contains time courses of cell differentiation (cells changing from one type to another) as well measures of perturbed cells. Since these samples do not have a well-defined locality in the body given by cell or tissue type, they were not used in the combined dataset. Such samples were filtered out using the human sample information file.

Since the FANTOM data had both an ontology file and the human sample information file, both were used to map to Uberon. The disagreements between the two mappings revealed some inconsistencies with the data

Reproducibility bonus: project management

- You can have a list of all the things that you need to do (issues).
- You can sort them into categories and into small “sprints” (milestones).
- You can have a list of all the things you want to do, but won’t:
  - you don’t have to move things from the “maybe” list to the ”no” list until you’re ready (you will become ready).





Including interactive images

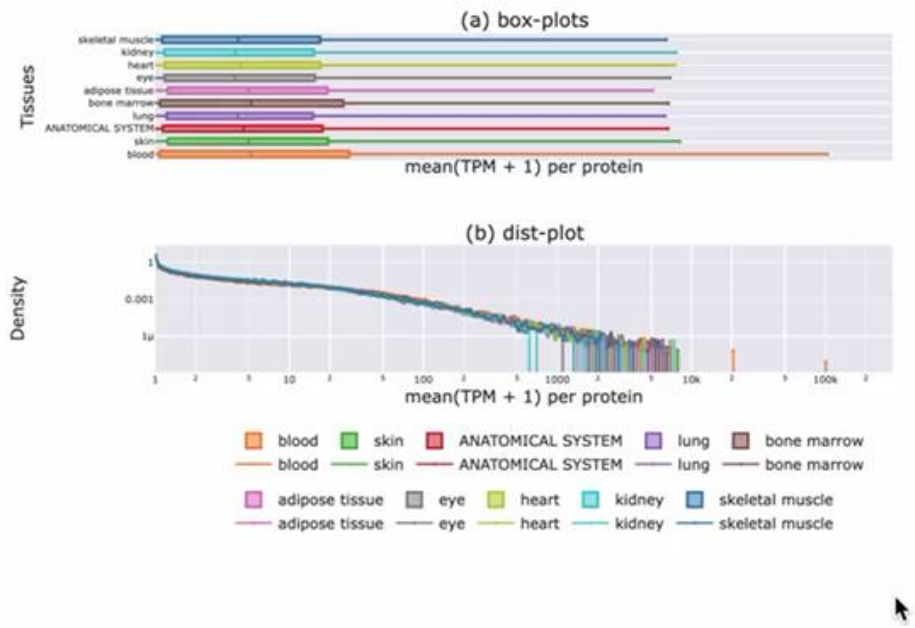


Fig. 6.4 (a) box-plots showing the distribution of mean (TPM+1) values (note: logarithmic x axis) for the top 10 most common tissue and primary cell samples in the FANTOM5 human data. (b) density-plots showing the distribution of mean (TPM+1) values for the top 10 most common tissue and primary cell samples in the FANTOM5 human data on log-log axes.

- I used `plotly` to include interactive images: so readers can look at specific ranges of values or compare different data points.
- Had to do a slightly weird workaround to give the Figure a caption!

```
<!-- ../images/blank.png This is a workaround to put a 1x1px blank image after an interactive image so that it appears to have a figure label -->
```

Documentation

- Documentation using `sphinx`

- Write docstrings in your code and they automatically create documentation.

- Could use in parts of my thesis as well as on the page that held the Python package I released.

7.2.3.1. The `Relations` class

The `Relations` class finds relationships of certain types between sources and targets. It subclasses a Pandas DataFrame since that is a convenient and familiar format for the relationship information to be returned.

class `ontology.relations.Relations`(*allowed_relations: list, ont, sources=None, targets=None, source_targets=None, excluded=None, col_names=None, mode='any'*)

`__init__`(*allowed_relations: list, ont, sources=None, targets=None, source_targets=None, excluded=None, col_names=None, mode='any'*)

Pandas Dataframe containing relationships between *sources* and *targets* terms according to *ont*. Finds relationships that do not pass through *excluded* terms and uses only *allowed_relations*. We keep looking until we find a relation to a target (if mode == 'any') or we run out of leads.

Parameters:

- **allowed_relations** – a list of allowed relations, e.g. ['is_a', 'part_of']
- **sources** – list of sources. For mode *all* must be a list of source-target tuple airs.
- **mode** – 'any' or 'all' - 'all is looking for specific term1-term2 pairs, while 'any' is looking for any relationship between something in specific source and anything in targets.
- **targets** – list of targets.
- **source_targets** – list of tuples of source-target pairs. Do not provide source or targets if using this parameter. Only runs in "all" mode.
- **ont** – Obo ontology object.
- **excluded** – a list/set of terms which are explicitly not being searched for (which may otherwise match the targets). Useful e.g. if we want to look for any tissue targets with prefix 'UBERON', except for very general ones. Does not allow relationships that pass through this term.
- **col_names** – Alternative column names for the output of Relations Data Frame, by default is ['from', 'relation_path', 'relation_text', 'to']

Fun and formatting

Contributions in this section

The Proteome Quality Index paper was created as a joint project between the Computational Biology group (then) at Bristol. I contributed to ideas for metrics, code to calculate some of these metrics, and paper editing.

Humans and bananas

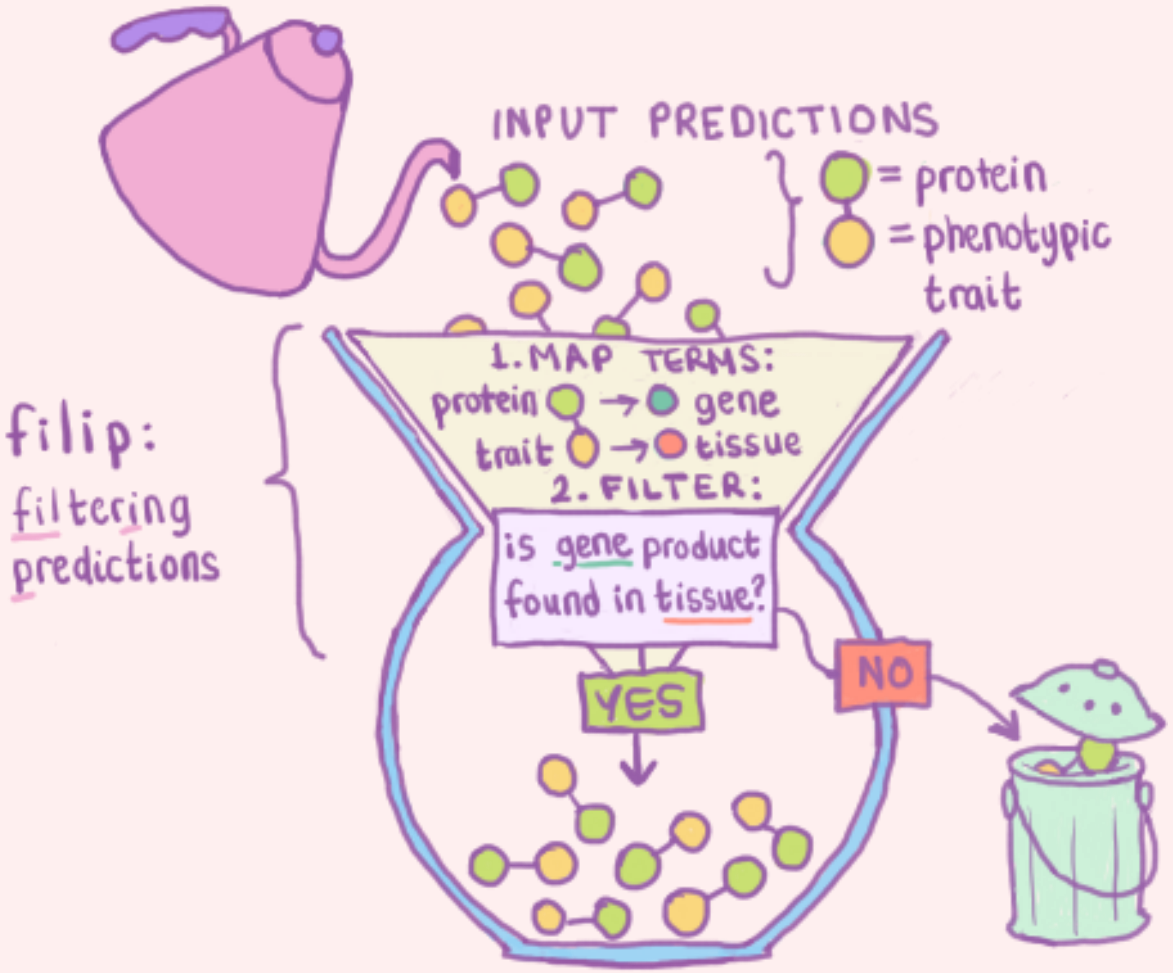
Humans share 50% of their *protein-coding* dna with bananas, but only 1% of their genome.

In science consensus is irrelevant. What is relevant is reproducible results.

—Michael Crichton

Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should. – Dr Ian Malcolm

Illustrations as science communication



99.973% correct

All incorrect filters were developmental traits

Low coverage

Decolonisation

Future Learn Subjects Courses FutureLearn for business

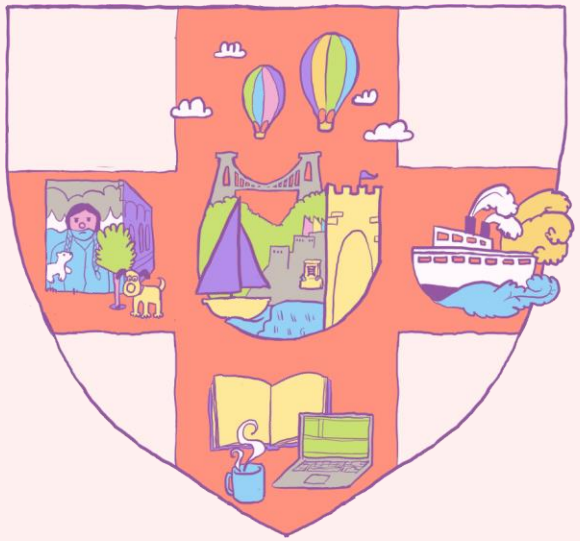
Online Courses / Teaching

University of BRISTOL

Decolonising Education: From Theory to Practice

Get to grips with the nature of the colonial legacy on our current state of knowledge and learning practices.

4.4 (61 reviews) 5,372 enrolled on this course



Charles Darwin and racism

Darwin used his theory of natural selection to argue that women and ite races were inferior to

Ronald Fischer, racism and eugenics

full title of *On the* Origin

Fischer has a legacy of scie racism. For example, ac campaigning for st tenth of the p eugenics

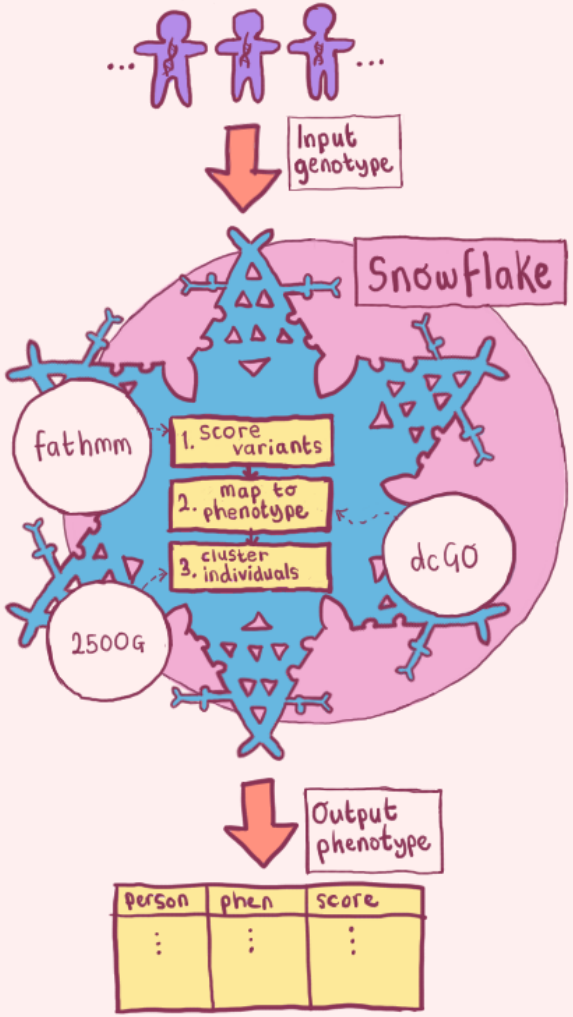
James Watson and racism

Watson's has publicly asserted that he lieves differences in average r between blacks and etic

Linneus and scientific racism

Linneus' classifications included a racist hierarchical classification of human beings[92].

Data Hazards ethics self-assessment



Label name	Label description	Label image	Reason for applying	Relevant safety precautions
Contains Data Science	Data Science is being used in this output, and any negative outcome of using this work are not the responsibility of "the algorithm" or "the software", but the people using it.		Snowflake uses data, makes predictions, and uses unsupervised learning.	When snowflake is deployed in new contexts (e.g. patent licenses sold), it should be done with the understanding that the licensee becomes accountable for using it responsibly.
Reinforces existing biases	Reinforces unfair treatment of individuals and groups. This may be due to for example input data, algorithm or software design choices, or society at large.		Project does not check that the algorithm works just as well for non-white races, and we would expect it to work less well for them since they are less represented in the input data linking variants and diseases[189].	Snowflake's efficiacy should be tested separately for each demographic that any deployment may effect.

What happened?

- It helped me submit:
  - I figured out my original contributions.
  - I was really proud of the *way* I'd written up my thesis
  - Those things combined (+ being sick of it!) gave me the courage to hand it in even though my supervisors hadn't read it and the main parts weren't published.
- I passed my Viva with minor corrections, which was a huge relief, and they even nominated it for a thesis prize!
- I got to use all the cool skills I'd learned on future projects.

Lessons learned

- You have to do some WEIRD work-arounds to adopt new technology.
- Writing things helps get your thoughts straight.
- PhDs are an opportunity to learn stuff that you want to!
- Find communities and pockets of research that inspire you.
- There’s a lot of different ways to do your PhD

Thank you

Twitter: @NatZelenka
GitHub: NatalieZelenka
Email: Natalie.Zelenka@Bristol.ac.uk