Data horrors & fighting ghosts

Vrije Universiteit Amsterdam – Data Conversations

Dr. Veronika Cheplygina

@drveronikach

https://www.veronikach.com

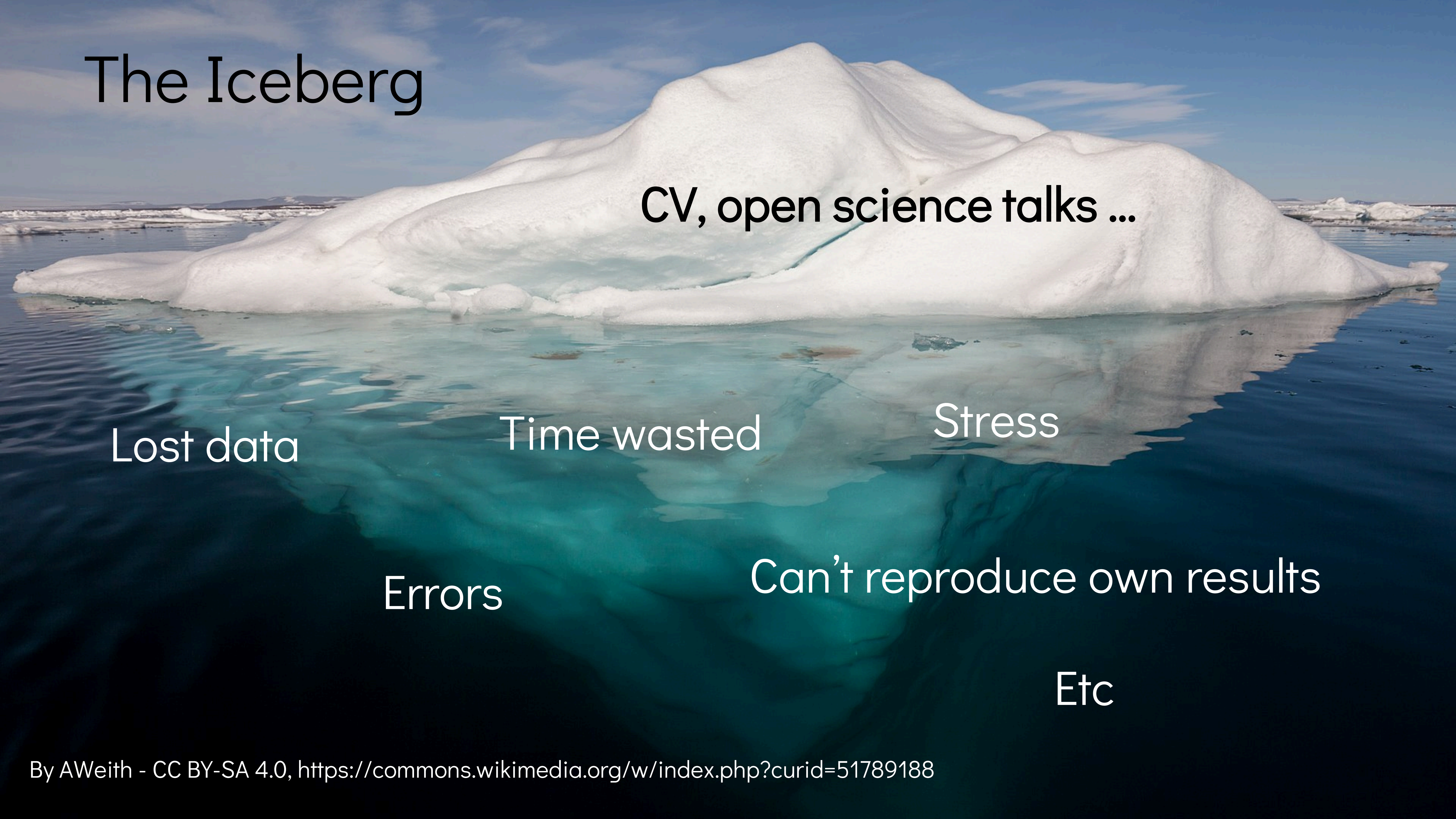
My Science CV

- Area: machine learning / medical imaging

- 2011-2015 PhD TU Delft
- 2015-2016 postdoc Erasmus Medical Center
- 2017- 2020 assistant professor TU Eindhoven

- Publications, invited talks (incl. open science…)

The Iceberg



CV, open science talks …

Lost data

Time wasted

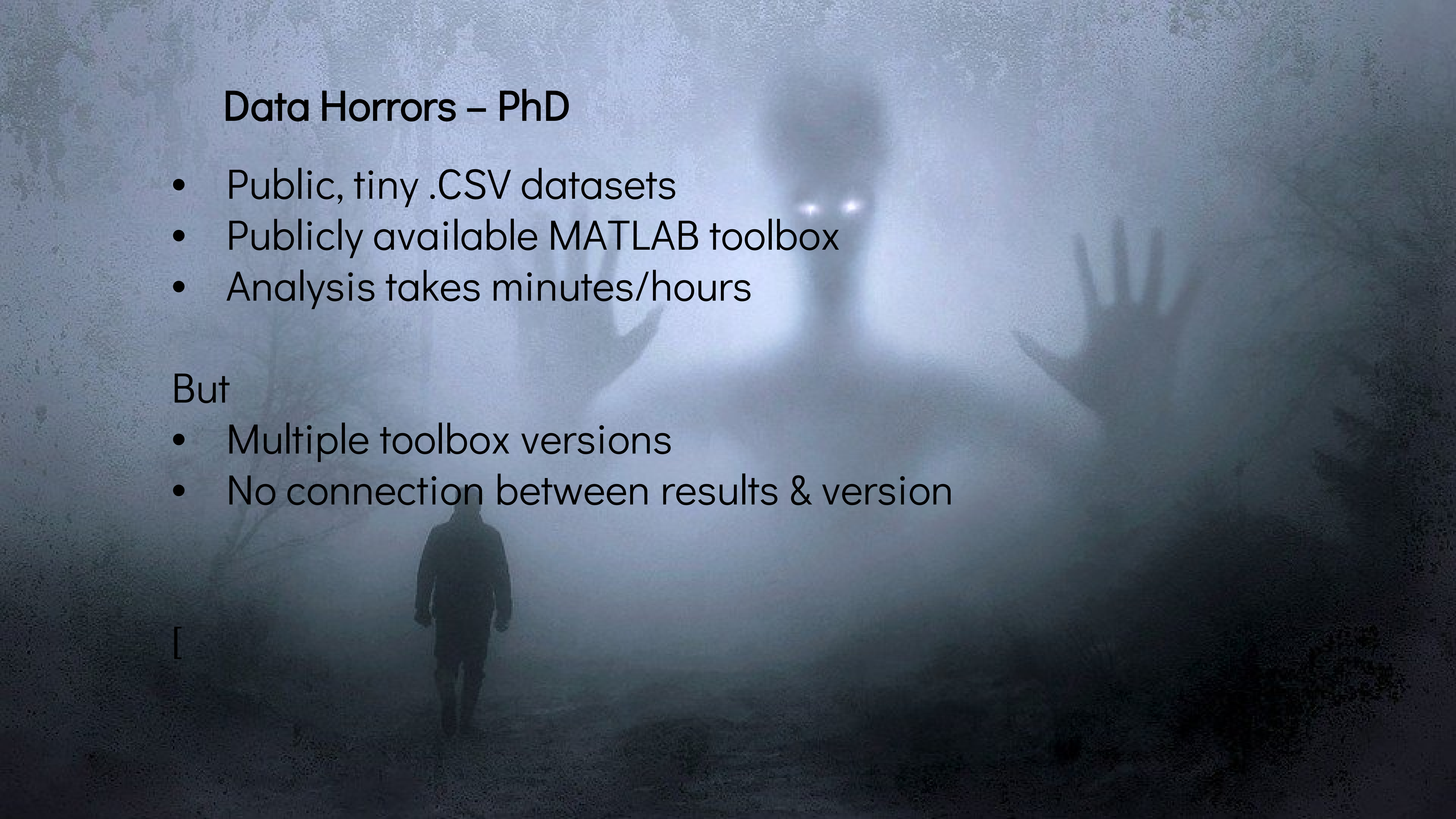
Stress

Errors

Can’t reproduce own results

Etc

By AWeith - CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=51789188



Data Horrors – PhD

- Public, tiny .CSV datasets
- Publicly available MATLAB toolbox
- Analysis takes minutes/hours

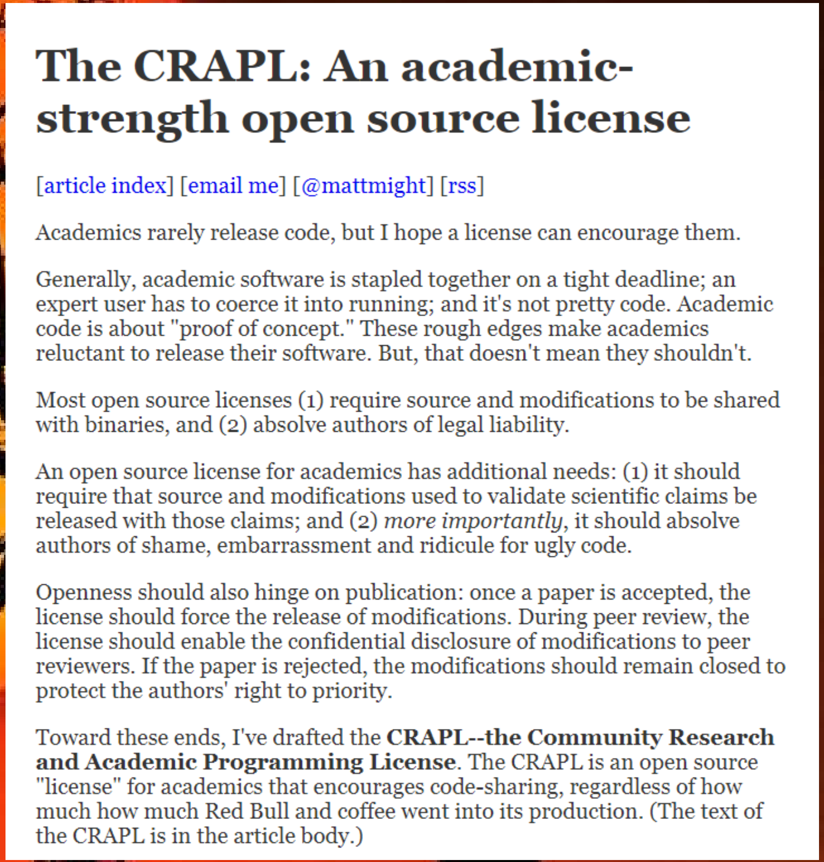
But

- Multiple toolbox versions
- No connection between results & version

[

PhD writing (& procrastinating)

- Preprints
- Share data & code "badly"
- Still a bit effective



The CRAPL: An academic-strength open source license

[article index] [email me] [@mattmight] [rss]

Academics rarely release code, but I hope a license can encourage them.

Generally, academic software is stapled together on a tight deadline; an expert user has to coerce it into running; and it's not pretty code. Academic code is about "proof of concept." These rough edges make academics reluctant to release their software. But, that doesn't mean they shouldn't.

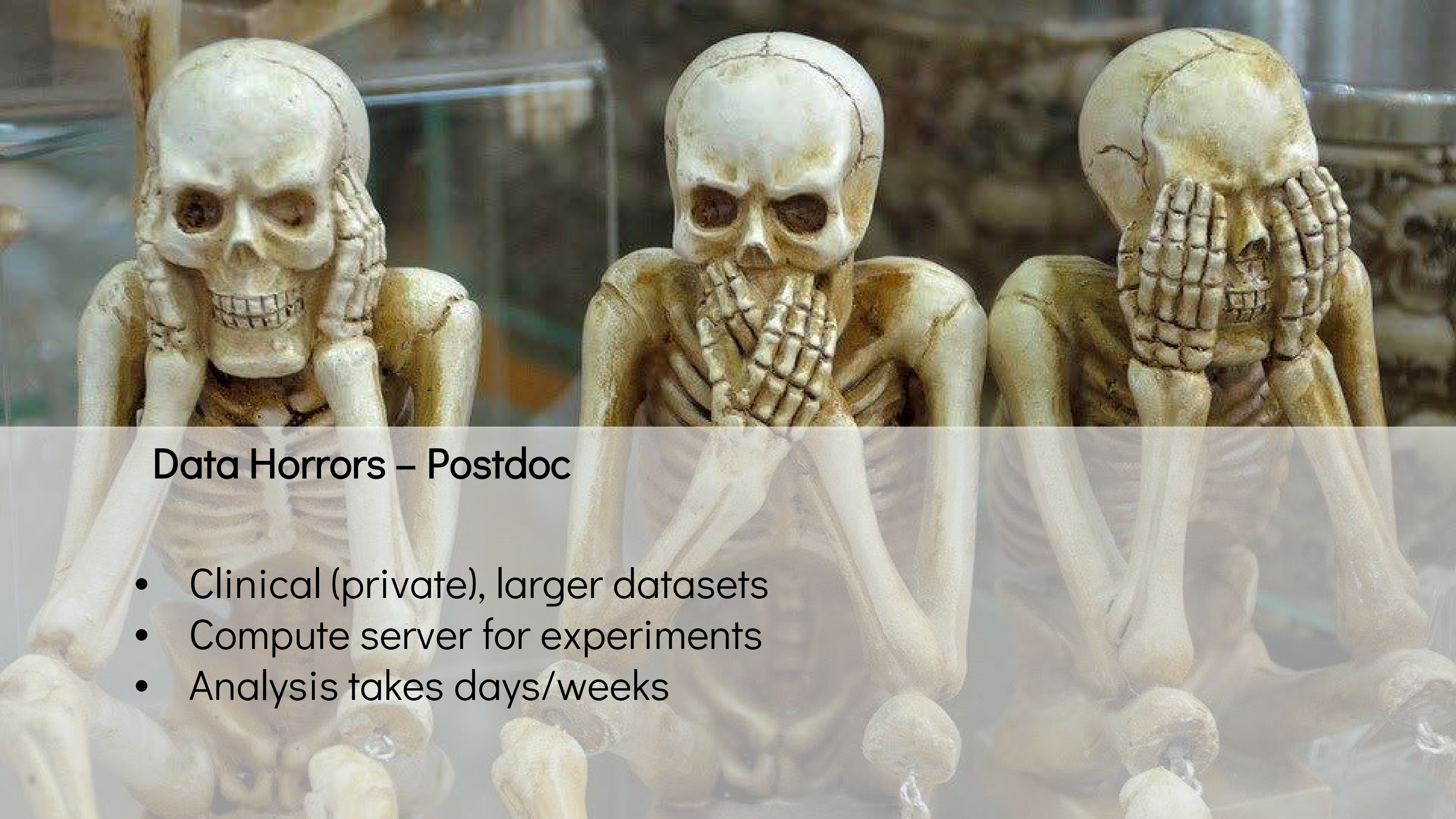
Most open source licenses (1) require source and modifications to be shared with binaries, and (2) absolve authors of legal liability.

An open source license for academics has additional needs: (1) it should require that source and modifications used to validate scientific claims be released with those claims; and (2) *more importantly*, it should absolve authors of shame, embarrassment and ridicule for ugly code.

Openness should also hinge on publication: once a paper is accepted, the license should force the release of modifications. During peer review, the license should enable the confidential disclosure of modifications to peer reviewers. If the paper is rejected, the modifications should remain closed to protect the authors' right to priority.

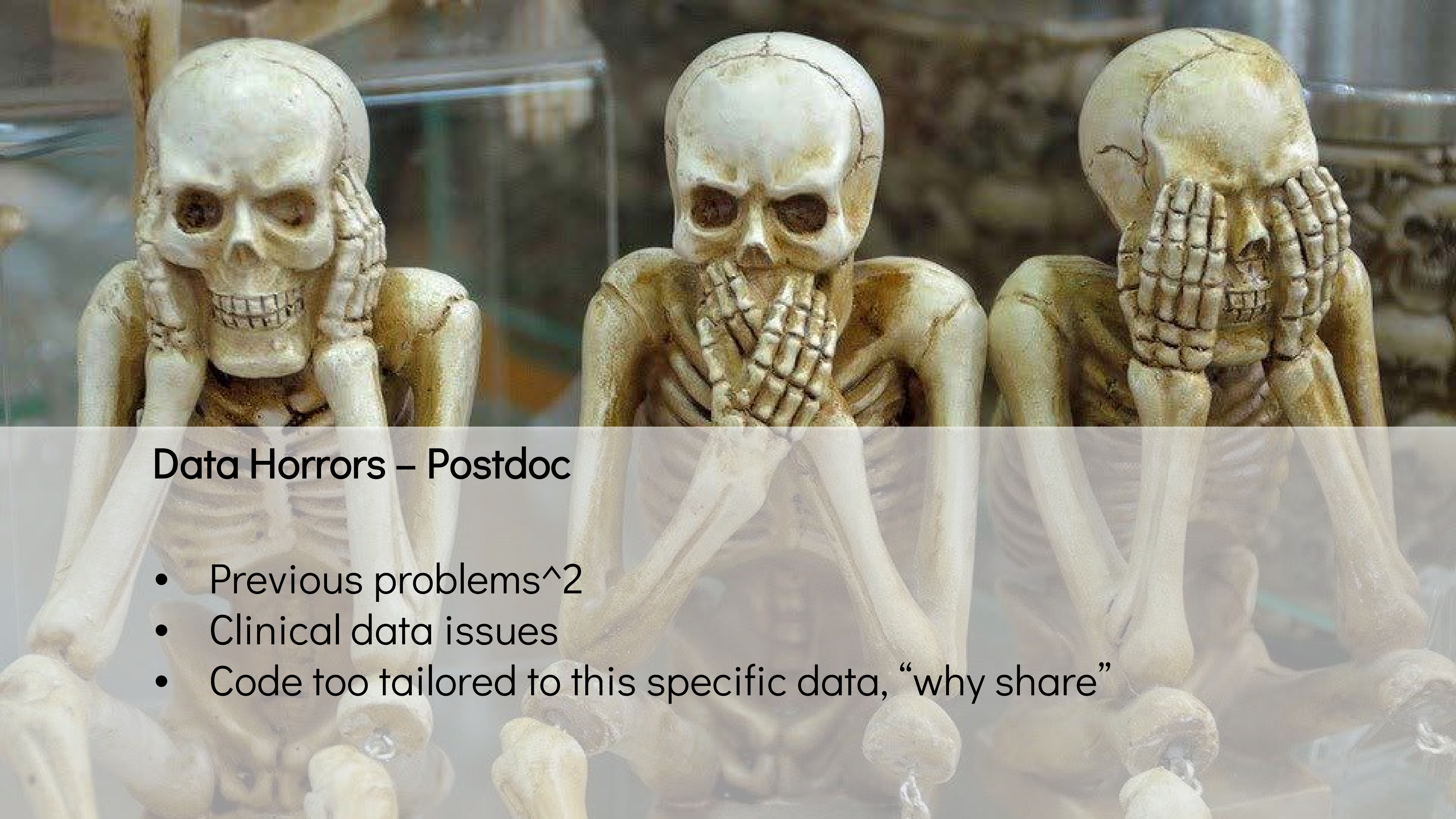
Toward these ends, I've drafted the **CRAPL--the Community Research and Academic Programming License**. The CRAPL is an open source "license" for academics that encourages code-sharing, regardless of how much how much Red Bull and coffee went into its production. (The text of the CRAPL is in the article body.)

http://matt.might.net/articles/crapl/



Data Horrors – Postdoc

- Clinical (private), larger datasets
- Compute server for experiments
- Analysis takes days/weeks



Data Horrors – Postdoc

- Previous problems^2
- Clinical data issues
- Code too tailored to this specific data, “why share”

Current situation

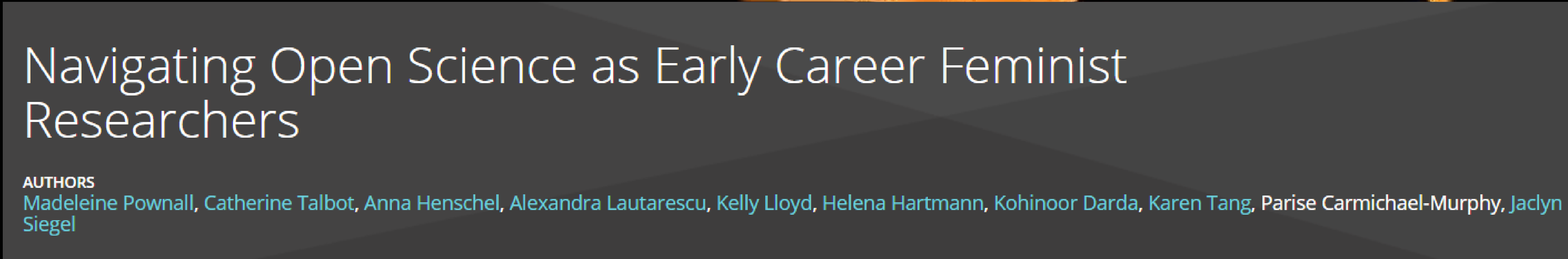
- Work with others
- Public data, share generated data
- "Lazy Github"

Still difficult:

- Procedures/conventions
- Responses

"Shouldn't you do more real research instead?"

"THAT's not how to do open science!"



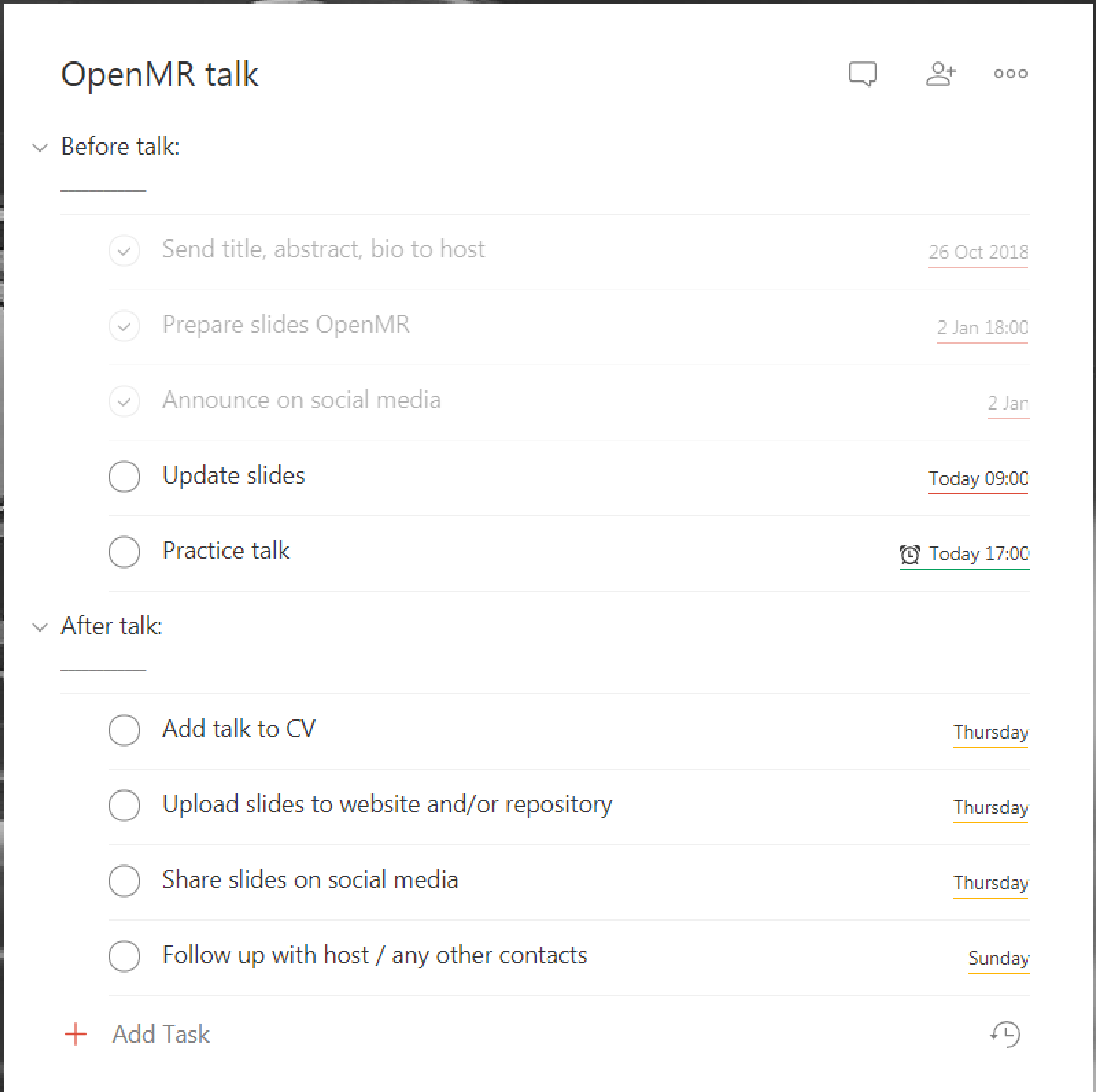
Navigating Open Science as Early Career Feminist Researchers

AUTHORS

Madeleine Pownall, Catherine Talbot, Anna Henschel, Alexandra Lautarescu, Kelly Lloyd, Helena Hartmann, Kohinoor Darda, Karen Tang, Parise Carmichael-Murphy, Jaclyn Siegel

Fighting ghosts

1. Start somewhere, but show up



OpenMR talk

Before talk:

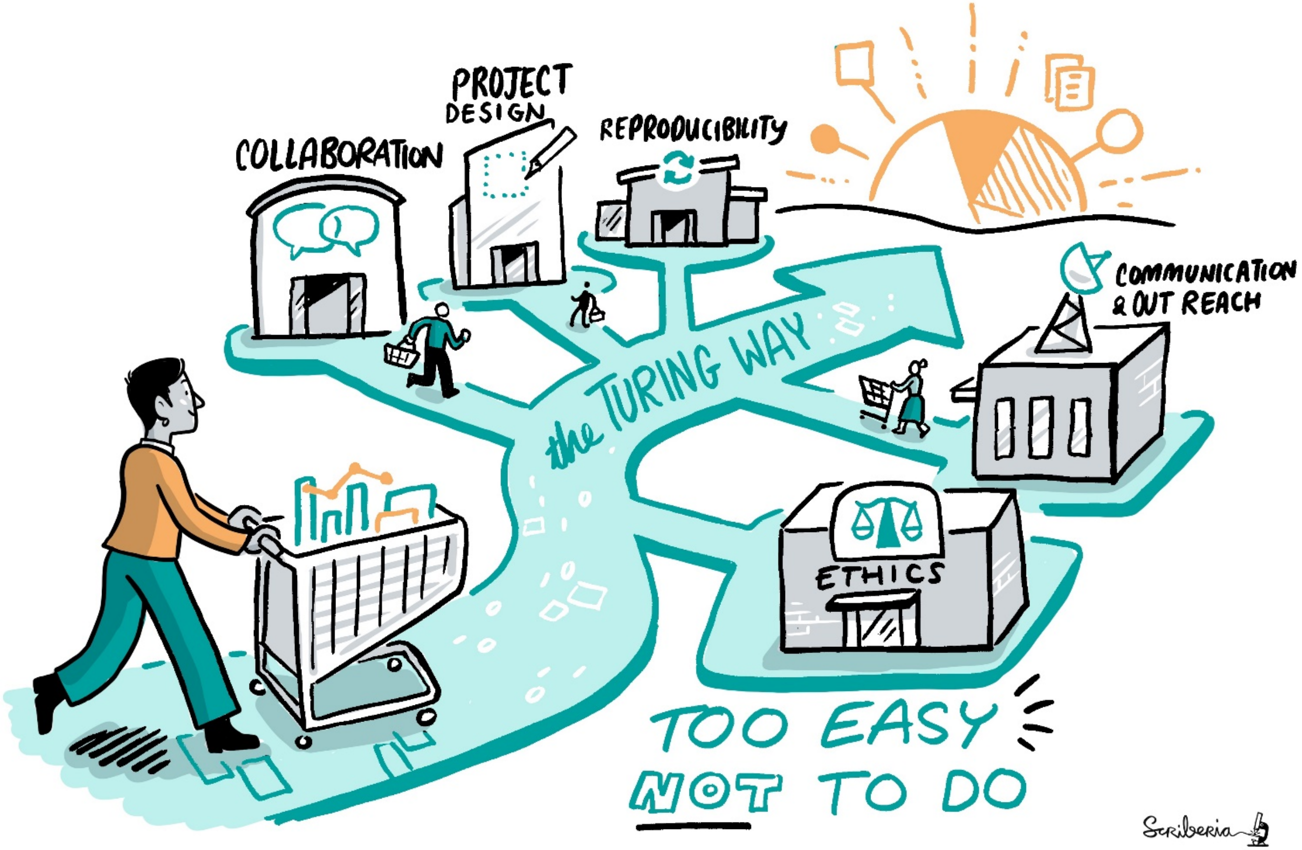
- ✓ Send title, abstract, bio to host — 26 Oct 2018
- ✓ Prepare slides OpenMR — 2 Jan 18:00
- ✓ Announce on social media — 2 Jan
- ○ Update slides — Today 09:00
- ○ Practice talk — Today 17:00

After talk:

- ○ Add talk to CV — Thursday
- ○ Upload slides to website and/or repository — Thursday
- ○ Share slides on social media — Thursday
- ○ Follow up with host / any other contacts — Sunday

2. Find accountability & support

2. Find accountability & support

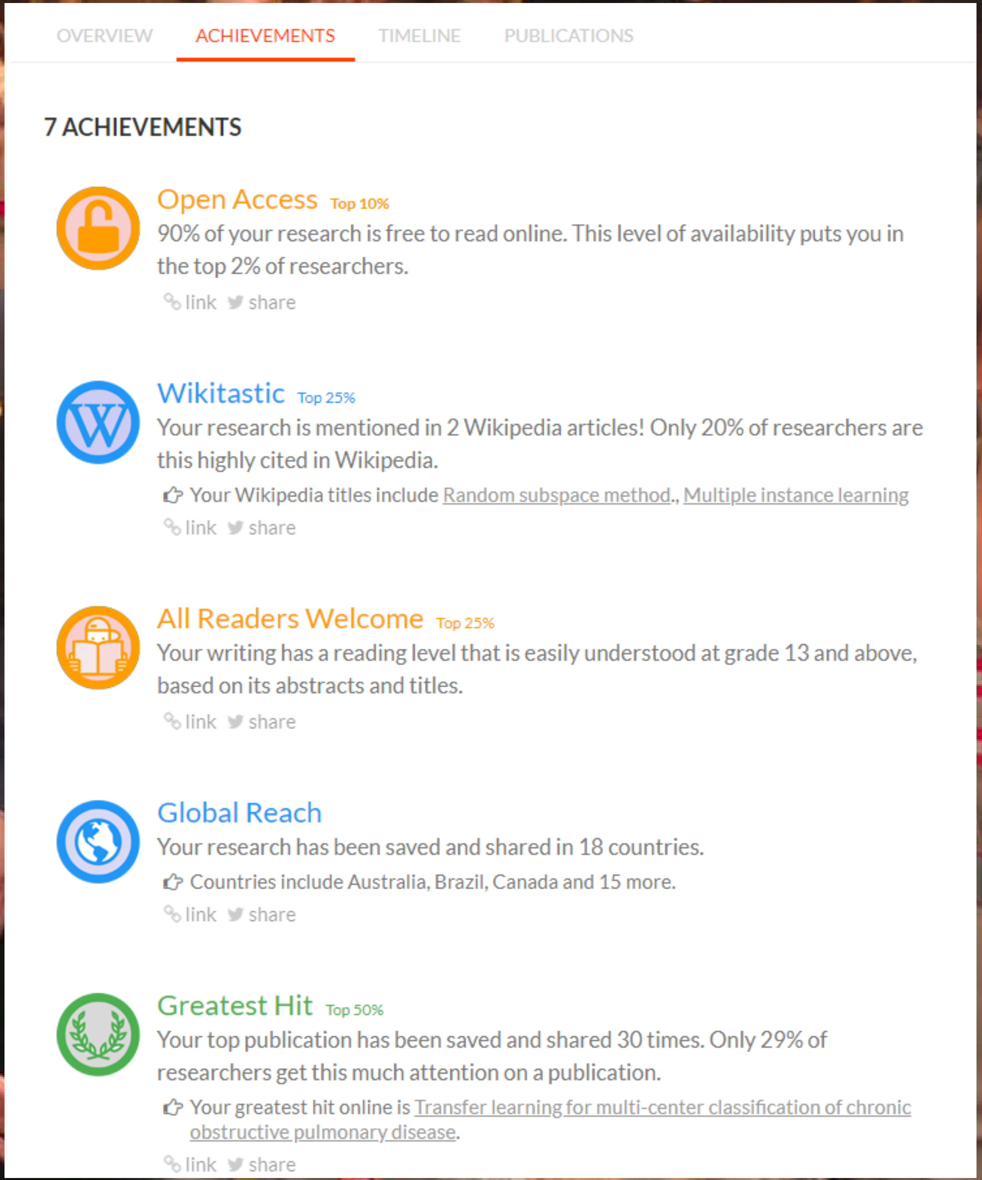


https://the-turing-way.netlify.app/welcome

Open Life Science

https://openlifesci.org/

3. Reward yourself



7 ACHIEVEMENTS

Open Access Top 10%

90% of your research is free to read online. This level of availability puts you in the top 2% of researchers.

link share

Wikitastic Top 25%

Your research is mentioned in 2 Wikipedia articles! Only 20% of researchers are this highly cited in Wikipedia.

Your Wikipedia titles include Random subspace method., Multiple instance learning

link share

All Readers Welcome Top 25%

Your writing has a reading level that is easily understood at grade 13 and above, based on its abstracts and titles.

link share

Global Reach

Your research has been saved and shared in 18 countries.

Countries include Australia, Brazil, Canada and 15 more.

link share

Greatest Hit Top 50%

Your top publication has been saved and shared 30 times. Only 29% of researchers get this much attention on a publication.

Your greatest hit online is Transfer learning for multi-center classification of chronic obstructive pulmonary disease.

link share

THANKS!

@drveronikach

https://www.veronikach.com

Image credits

- https://pixabay.com/photos/fantasy-spirit-nightmare-dream-2847724/
- https://pixabay.com/photos/tree-cat-silhouette-moon-full-moon-736877/
- https://pixabay.com/photos/skeletons-funny-hear-no-evil-1617539/
- https://pixabay.com/photos/ufo-alien-guy-pozaziemianin-2413965/
- https://pixabay.com/photos/moon-night-full-moon-gespenstig-703537/

- https://pixabay.com/photos/lego-ghostbusters-peter-4418653/
- https://pixabay.com/photos/phone-old-year-built-1955-bakelite-3594206/
- https://pixabay.com/photos/pumpkin-halloween-face-autumn-2892303/
- Ghostbusters 2016 screen captures from https://movie-screencaps.com/ghostbusters-2016/