REMODEL- A Community Poll on Reproducibility in Computational Geosciences

Robert Reinecke

May 18, 2021

Report last updated: 17th May, 2021

This document summerizes a full analysis for detailed results. **It is not a full publication but rather an automated representation of the data present in this repository**
It summarizes the extensive results and builds the foundation for the publication of a journal paper. It will also be distributed along with the data in the course of a journal and data publication.

1 Aim of this survey

Software development has become an integral part of the geosciences[1] as models and data processing get more sophisticated. Paradoxically, it poses a threat to scientific progress as the pillar of science, reproducibility, is seldomly reached[2]. Software code tends to be either poorly written and documented or not shared at all; proper software licenses are rarely attributed. This is especially worrisome as scientific results have potential controversial implications for stakeholders and policymakers and may influence the public opinion for a long time[3].
In recent years, progress towards open science has led to more publishers demanding access to data and source code alongside peer-reviewed manuscripts4,5. Still, recent studies find that results can rarely be reproduced [6,7].
In this project, we conduct a poll among the geoscience community which is advertised via scientific blogs (AGU, EGU), research networks (researchgate.net and mailing lists), and social media. Therein, we strive to investigate the causes for that lack of reproducibility. We take a peek behind the curtain and unveil how the community develops and maintains complex code and what that entails for reproducibility[8] . Our survey includes background knowledge, community opinion, and behaviour practices regarding reproducible software development. We postulate that this lack of reproducibility[9] might be rooted in insufficient reward within the scientific community, insecurity regarding proper licencing of software and other parts of the research compendium as well as scientists' unawareness about how to make software available in a way that allows for proper

attribution of their work. We question putative causes such as unclear guidelines of research institutions or that software has been developed over decades[10] , by researchers' cohorts without a proper software engineering process[1] and transparent licensing.
To this end, we also summarize solutions like the adaption of modern project management methods from the computer engineering community[11] that will eventually reduce costs while increasing the reproducibility of scientific research[8] .

1 A comment to "Most Computational Hydrology is not Reproducible, so is it Really Science?" R.W. Hut, N.C. van de Giesen, N. Drost, Water Resources Research, 2017
2 Hutton, C., Wagener, T., Freer, J., Han, D., Du_y, C., and Arheimer, B., Most computational hydrology is not reproducible, so is it really science? Water Resources Research, 2016
3 Munafò, M., Nosek, B., Bishop, D. et al., A manifesto for reproducible science. Nat Hum Behav, 2017
4 Executive editors, G. Editorial: The publication of geoscientifc model developments v1.2. Geoscientifc Model Development, 2019
5 Katz, D. S., Niemeyer, K. E., and Smith, A. M., Publish your software: Introducing the journal of open source software (joss), Computing in Science Engineering, 2018
6 Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., and James, R., Assessing data availability and research reproducibility in hydrology and water resources. Scientific data, 2019
7 Añel, J. A., García-Rodríguez, M., and Rodeiro, J.: Current status on the need for improved accessibility to climate models code, Geosci. Model Dev., 2021
8 Stodden, V., The reproducible research standard: Reducing legal barriers to scientific knowledge and innovation. IEEE Computing in Science & Engineering, 2009
9 https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
10 Muller, C., Schaphoff, S., von Bloh, W., Thonicke, K., and Gerten, D., Going open-source with a model dinosaur and establishing model evaluation standards. EGU, 2018
11 https://software.rajivprab.com/2019/11/25/the-birth-of-legacy-software-how-change-aversion-feeds-on-itself

Our larger questions:

- Is eproducibility is an issue in the geosciences? Are bad code and documentation the root cause of that issue?
- Is model software too complex? Does that hinder reproducibility?
- Are researchers missing the tools and know-how (methods, licenses etc.) to build good model code?

- Is missing funding and missing time preventing researchers from making their models more accessible?

We define reproducibility as:
"Reproducibility in the context of modeling in the geosciences means that results obtained by a modeling experiment should be achieved again with a high degree of agreement when the study is replicated with the same model design, inputs, and general methodology by different researchers.
We explicitly exclude the retracing of results by means of using a different modeling environment (including variations in model concept, algorithms, input data or methodology)."

2 Data processing

We designed the survey according to standards from psychology research. We apply descriptive statistics to analyse demographic background and basic analysis. Further, we apply inferential statistical methods to test the unerlying hypotheses.
The raw data of the survey is stored in the folder **LiveData**. **The raw data has not been modified or cleaned in any way.** To run some basic cleanup run the follwoing script:

```
import process_data.py as p
p.process()
```

3 Results

All data processing and plotting (including building this document) can be executed by running `python run.py`. Plotting details and addtional processing can be found in th script `plot_all.py`.
Our main hypothesis for this analysis where the following:

- **H1** Young scientists develop software more actively than established researchers.
- **H2** Young scientists are more familiar with software licenses than established researchers.
- **H3** Young scientists are more familiar with modern development methods than established researchers.
- **H4** Software has gotten so complex that senior researchers are not able to comprehend it anymore.
- **H5** Software has gotten so complex that senior researchers are not capable in teaching the right methods.

- **H6** Researchers code frequently but without knowledge about proper engineering methods, licences and tools.
- **H7** The most frequently used language is still C/Fortran. Younger scientists tend to use Python and R; this is consistent throughout fields
- **H8** Most researchers are autodidacts when it comes to coding.
- **H9** Most researchers have never reproduced code with the original model. Only with their own model. This differs between fields.
- **H10** Practitioners and researchers perceive the issue of reproducibility differently. Scientists are more aware (?).
- **H11** There are more researchers that apply software than they are ones that develop it. This differs between activities and fields.
- **H12** Most researcher do not know if their software belongs to them.
- **H13** Models that are available are hard to use. Causes?: Bad code, no documentation, no input data
- **H14** Senior researchers are convinced their work is reproducible. (much more at least than young scientists)
- **H15** The smaller the scale the more reproducible and accessible.
- **H16** Researchers think that their software is bug free and always correspond to their intended implementation.
- **H17** Most senior scientists think investment in FOSS doesn't pay off.
- **H18** Senior researchers think that their code/project is easier to understand but that conflicts with reality. And younger researchers have the opposite understanding.
- **H19** We need more funding to enable reproducible computational science.
- **H20** New research software tends to be FOSS, big players are often legacy software grown over decades which is hard to reimplement as FOSS.

3.1 Sample Characteristics (demographics)

Who were our participants? Here we present characteristics of the participants in our poll, i.e. their current career stage, their years of research experience, their geo-scientific field and scale as well as their current focus of work. This is purely descriptive statistics. As we welcomed everyone to our poll, we did not form any assumptions regarding sample characteristics. Also, we tried, but might not have reached a representative sample of the population of geo-scientists.
Here, we report basic sample characteristics. Also, we can check for and report any salient sample properties.
Corresponding survey questions:

- DM01 - What career stage are you in?
- DM02 - For how long have you been working in your research field?
- DM06 - To which field within the geosciences does your research mainly belong?
- DM05 - What geographic scale are you working on?
- DM07 - What is the focus of your work?

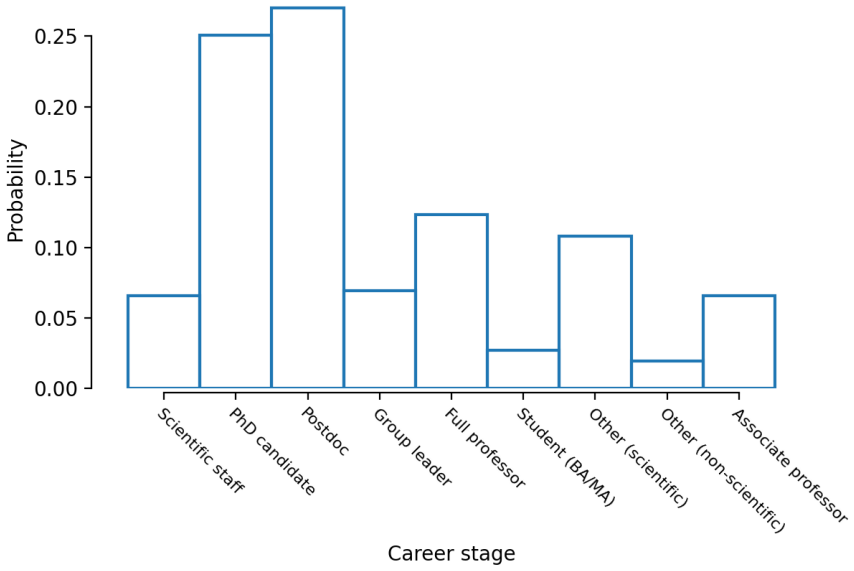

Figure 1: DM01 - Career stage of participants

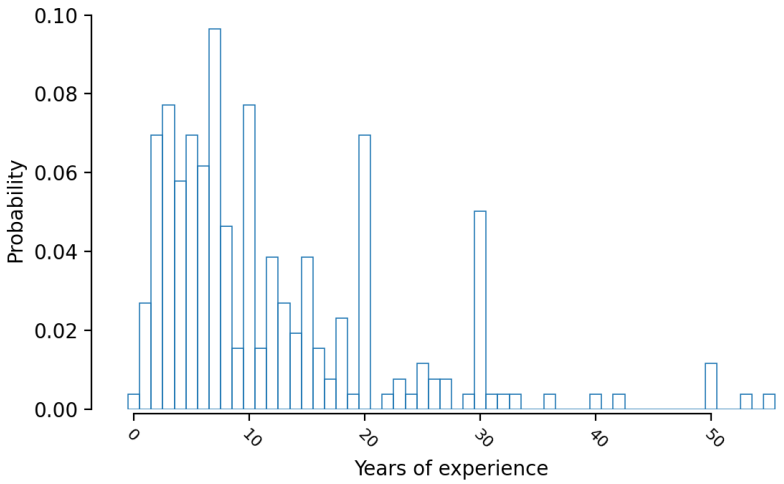

Figure 2: DM02 - For how long have you been working in your field?

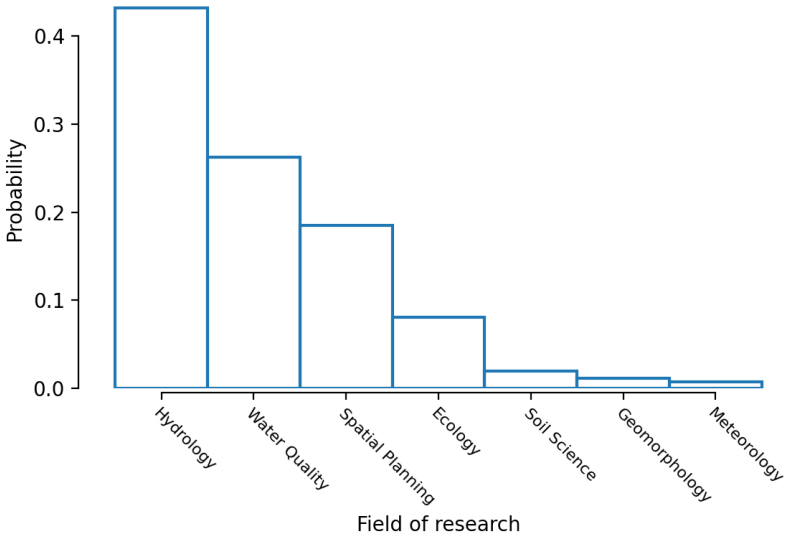

Figure 3: DM06 - Which field do you belong to?

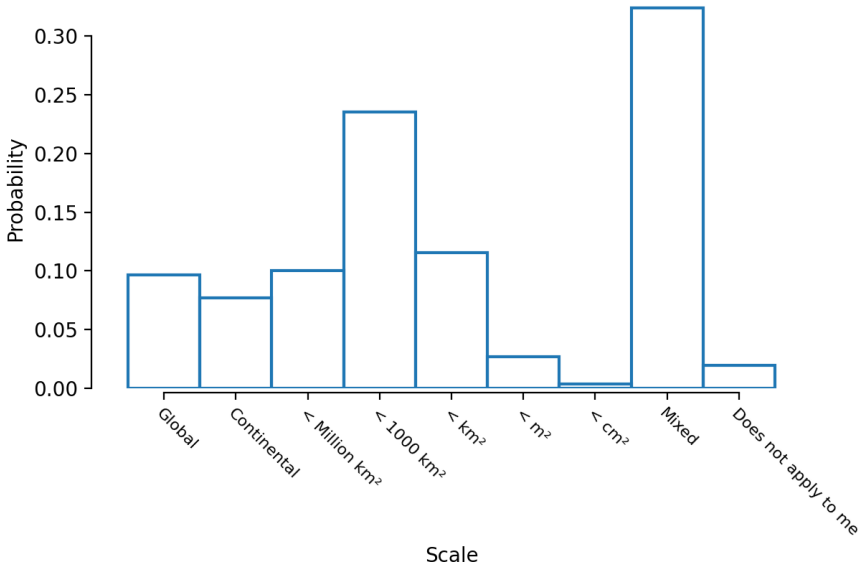

Figure 4: DM05 - What scale are you working on?

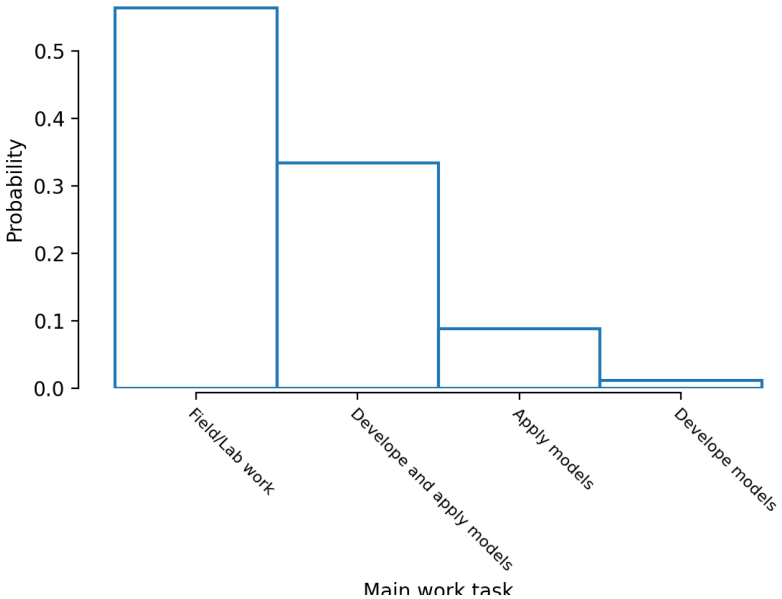

Figure 5: DM07 What is the focus of your work?

3.2 Community Opinion on Reproducibility

This part covers people's view and opinion. We assume that our defintion of reproducibility was used (annotated at several points in the poll).
Analysis Steps:

- plot corresponding questions (across full sample)
- perform statistical testing of our above stated hypotheses (across subgroups, e.g. early career vs senior researcher)
- perform further data exploration (NOT hypothesis testing) - this might inspire future research, e.g. correlation analysis

Corresponding survey questions:

- O101 - How strongly do you agree with the following statements?
- O103 - How often do you use research software in your research practice?
- S113 - How long do you think does it take for an average PhD student to efficiently work with your research software?

Corresponding hypotheses

- H4 Software has gotten so complex that senior researchers are not able to comprehend it anymore.
- H5 Software has gotten so complex that senior researchers are not capable in teaching the right methods.
- H9 Most researchers have never reproduced code with the original model. Only with their own model. This differs between fields.
- H10 Practitioners and researchers perceive the issue of reproducibility differently. Scientists are more aware (?).
- H13 Models that are available are hard to use. Causes?: Bad code, no documentation, no input data.
- H14 Senior researchers are convinced their work is reproducible. (much more at least than young scientists).
- H16 Researchers think that their software is bug free and always correspond to their intended implementation.

O101: Opinion on Reproducibility in Geo-Sciences: How strongly did participants generally agree with statements? Do they consider it a problem at all? Do they think that their work is reproducible?

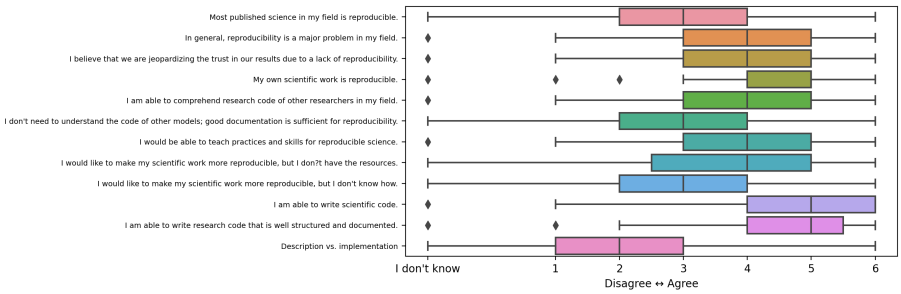

Figure 6: O101 How strongly do you agree with the following questions?

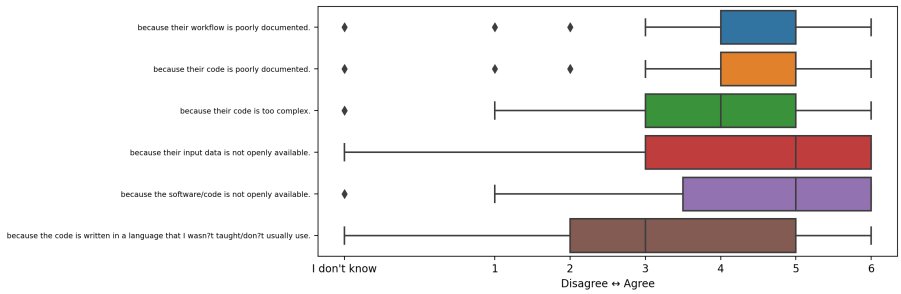

Figure 7: O103 How often do you use research software?

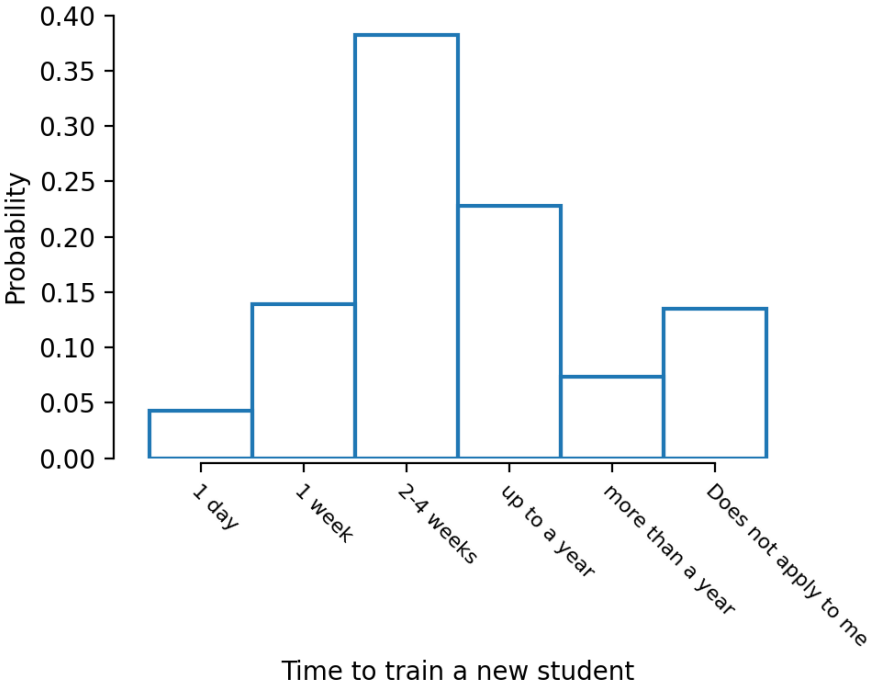

Figure 8: S113 How long does it take to train a new student in your software?

3.2.1 H4

Differences in Opinion on Reproducibility in Geo-Sciences: Do early career researcher differ from senior researchers in their agreement?
Established researchers 32 Young researchers 177 Mean agreement to reproducibility 3.1185770750988144
Mean agreement among Y 3.023121387283237
Mean agreement among O 3.6129032258064515
Median agreement to reproducibility 3.0
Median agreement among Y 3.0
Median agreement among O 4.0 Career has no normal distribution
No statistical test possible -¿ much to less Prof. to do a proper Wilcoxon or even t-test

3.2.2 H5

Reasons for lack of reproducibility: Rating of agreement to reasons
perform hypothesis test
explore data even further
etc...

3.2.3 H9

XX

3.2.4 H10

XX

3.2.5 H13

XX

3.2.6 H14

XX

3.2.7 H15

XX

3.3 Reproducibility Practices and Skills

This part covers "actual" behaviour. (still only self-report assessment, but we can't change that)
Start here with summary of our hypotheses
We expected to see that... (formulate in a neutral tone)
Analysis Steps: plot corresponding questions (across full sample) perform statistical testing of our above stated hypotheses (across subgroups, e.g. early career vs senior researcher) perform further data exploration (NOT hypothesis testing) - this might inspire future research, e.g. correlation analysis
corresponding hypotheses (see Robert's evaluation plan): H6 H9 H12 H5 H2
Corresponding survey questions: O102 S103 S110 S202 S112 S101 S111 S104 S105 S106

3.4 Hurdles against and Solutions towards Reproducibility

This part covers "actual" behaviour. (still only self-report assessment, but we can't change that)
Start here with summary of our hypotheses
We stated that... (formulate in a neutral tone)
Analysis Steps: plot corresponding questions (across full sample) perform statistical testing of our above stated hypotheses (across subgroups, e.g. early career vs senior researcher) perform further data exploration (NOT hypothesis testing) - this might inspire future research, e.g. correlation analysis
corresponding hypotheses (see Robert's evaluation plan): H7 H13 H3
Corresponding survey questions: S203 S201 S204 (open text)

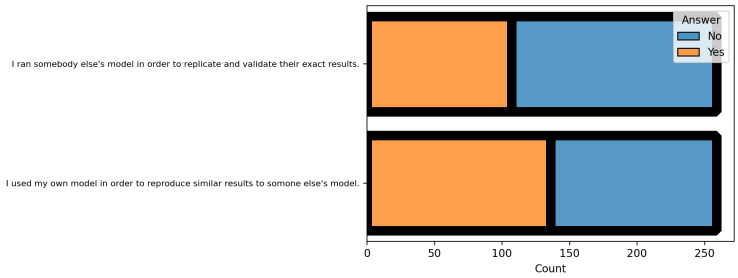

Figure 9: O102

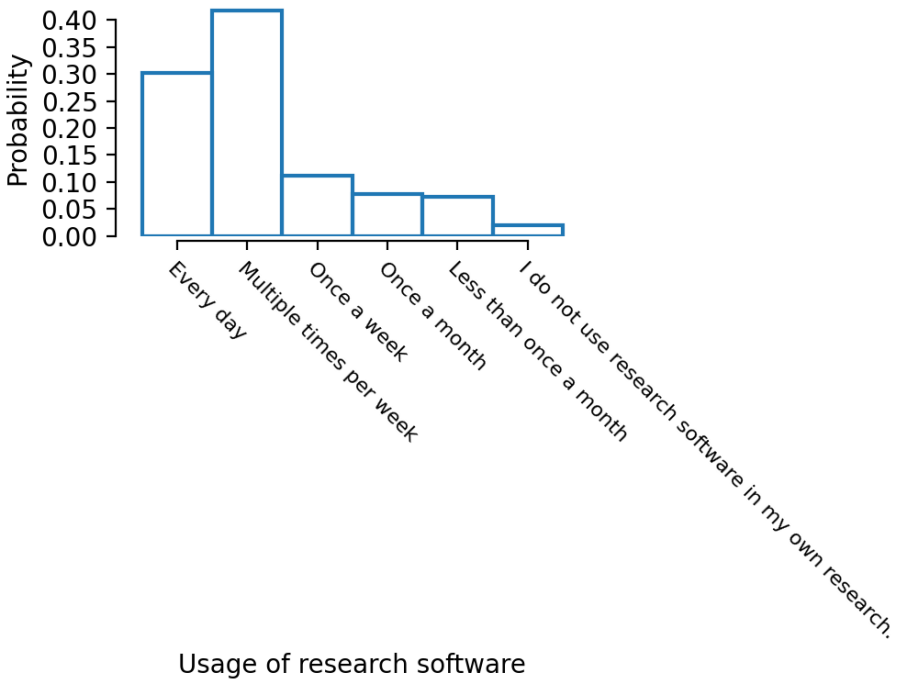

Figure 10: S103

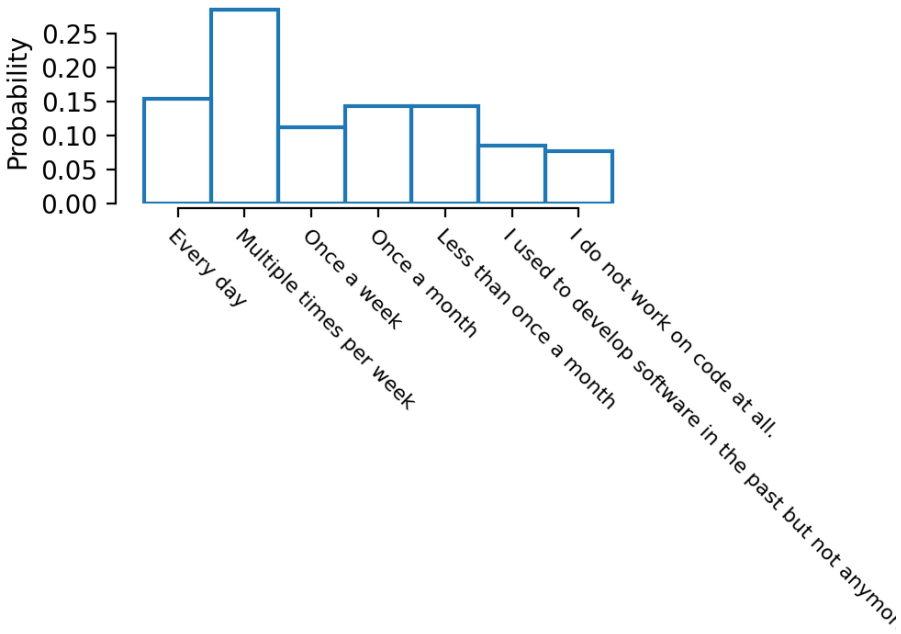

Figure 11: S110

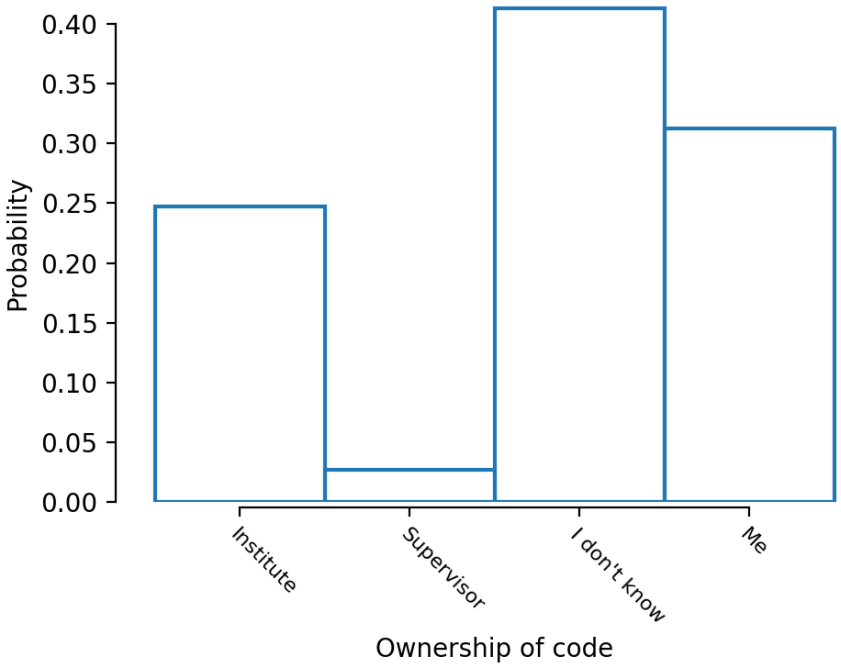

Figure 12: S202

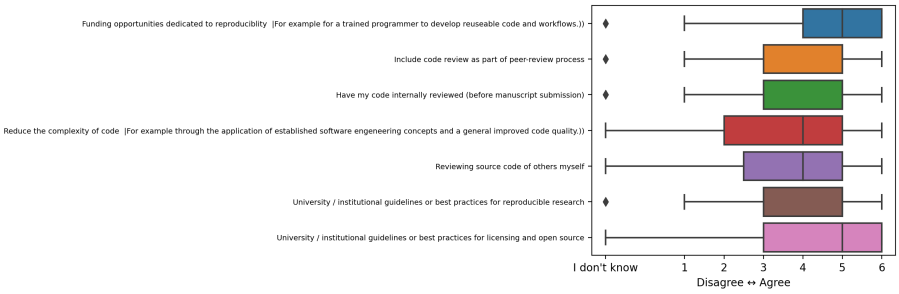

Figure 13: S201

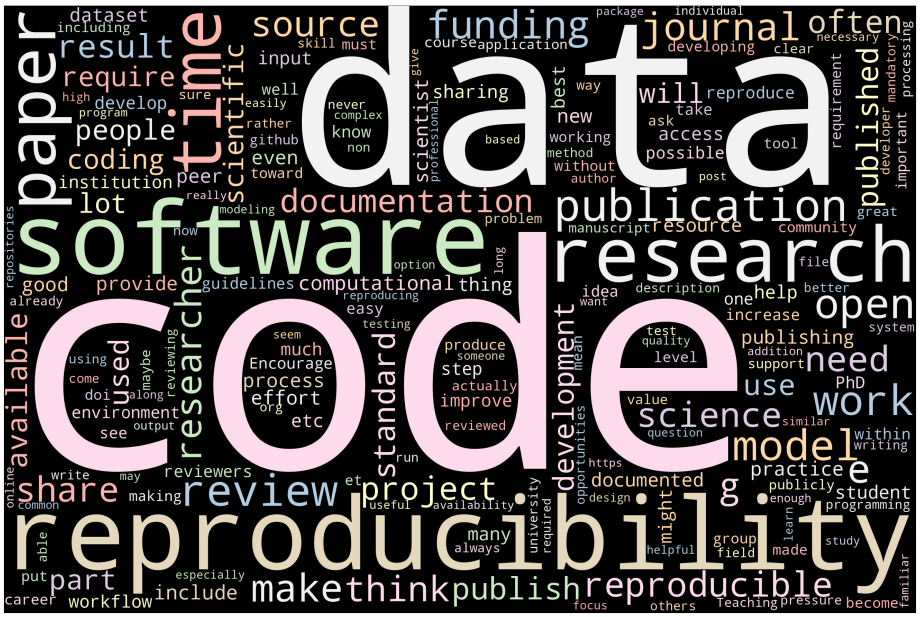

Figure 14: Full text answers as word cloud