Feedback on coursework and how to make your PhD reproducible

Points from your writing

Levels of reproducibility:

working with closed data means reproduction from raw is not possible -- often raw data are processed and summarized and from that summary we do statistics/make models. Is it possible to share those summary data? Ie split your analysis into to parts - one script can be re-used or at least one can infer/see what was doe on raw, the other is fully reproducible using intermediate results

The data are so big that it is not possible to share, or (most) others won't be able to recompute -- 1st, let me tell you want took 2 days to compute on my server 10 years ago, take 2 hours on m current laptop, so might very well be fine soon -- 2nd this is fine :-) LHC throw away 99% of their data and yet this is considered one of the highest standard in data sharing and analysis (transparency is what we are after)

Points from your writing

I will consolidate, clean-up, etc when archiving -- even while you are using version control you accept that the code is not clean, which is fine because that's the code used to produce your research paper ; but then what is it that you are archiving? It the code is now different this is not a 'true' archive to reproduce the paper.

-- perfection is the enemy of done -- once you have a working code and that produce the results you use for publication, maybe tidy up the comments and add some plain text where needed (ie literate programming) but that's the only thing that should change (also also archive the .git history)

Points from your writing

the authors were surprisingly very responsive and helped to set up a correct analysis.

Our experience as ‘old’ researchers is that most people are genuine, and once you explain what you want to do (unless your plan is to destroy their reputation) they usually help - even to set up competing hypotheses (see adversarial collaboration)