Reproducible Data Analysis with conda

Mika Pflüger
Motivation

• We publish our methods / algorithms
• We publish our data
⇒ In principle, our research can be reproduced

However:
• A lot of work
• Method description incomplete
⇒ Nobody actually builds on other people’s work
Publish your code!

Data $\Rightarrow$ Conclusions

often via code

Not enough to publish data, also publish code
Plan

• My code is python
• Publish all the code from raw data to graphic as found in the paper
• Put together with data in one .zip
Easily reproducible

• Add README.txt
• Better: add Makefile

• Test: my own paper from two years ago\(^1\), try to reproduce my own graphics

Easily reproducible?

$ make gisaxs_length_comparison.eps
python3 gisaxs_length_comparison.py
Traceback (most recent call last):
  File "gisaxs_length_comparison.py", line 3, in <module>
    import matplotlib
ModuleNotFoundError: No module named 'matplotlib'
make: *** [Makefile:4: gisaxs_length_comparison.eps] Fehler 1
Easily reproducible?

- Install missing libraries

New `matplotlib` completely breaks layout
Explicit dependencies with conda

- conda is a package manager
- Multiple self-contained environments
- Export your environment with
  
  ```bash
  $ conda list --export [--explicit]
  ```
- Can be imported with
  
  ```bash
  $ conda create --file
  ```
Easily reproducible!

- Library versions from 2017

⇒ Success!
Recipe

- Data and scripts together
- One .zip
- Add Makefile
- Export your conda environment
- Test your setup in clean environment (other computer, virtual machine)
Not enough

The best I could come up with

Problems:

• Not a standard, relies on README.txt
• Hard to discover as supplementary material
• Relies on conda.io to archive old library versions

We need a standard way to declare this
Conclusion

- Publish your code!
- Just publishing code, I could not reproduce my own work from 2 years ago
- With python and conda, I could
- Makes it easier to write a thesis
- We need to standardize this, and make it discoverable