Teaching Reproducibility to Data Science for Social Good Fellows

Valentina Staneva
Senior Data Scientist, eScience Institute

UNIVERSITY *of* WASHINGTON

UW Data Science for Social Good Program

- established in summer 2015
- ~16 students working in teams of 4 or 5
- mainly graduate students with a few advanced undergrads
- coming from diverse backgrounds working on problems in transportation, public health, equity, etc.
- working for 10 weeks, handing deliverables to project lead afterwards
- tutorials on git, programming, data analysis, stakeholder analysis, blogging, public speaking, etc.

https://escience.washington.edu/dssg/

Data and Code Challenges

- Data can be big
  - laptops, AWS, Hyak
- Data can be private and very sensitive
- Data are noisy: a lot of preprocessing
  - hard to set up tests
- Code is written under deadline pressure
- Different components may be in different languages

Reproducibility Learning Exercises

- **Testing**

  Teams exchange pieces of work, and try to reproduce results based on instructions

- **Self-Evaluation**

  Teams go through a set of questions of self evaluation of the state of their deliverables (questionnaire)

- **Learning**

  Tutorial on technical aspects of writing reproducible code: tests, virtual environments, CI

- **Elevating Status**

  A student on the team takes the responsibility to review and test th reproducibility of the code

Assessment

- Students fill out an anonymous survey to evaluate the tutorials of the program
- Discussions with students and project leads regarding the state of their code and data, and ‘final deliverables’
- Can the project be continued on after the end of the program?

Lessons Learnt

- Students need context instead of preaching:
  - integrating with project helps
- Community pressure helps
  - if all teams are doing it we will do it
- Reproducibility should be taught continuously
  - the concepts should be introduced early on, but some of the motivation may be missing in the beginning

Resources

Tutorial on tools for Reproducible Research in Python:

https://github.com/valentina-s/mobility-index

Assessing Reproducibility Book Chapter:

https://osf.io/preprints/socarxiv/gne3w/

Self-evalutation Questionnaire:

questionnaire link

Examples of Project Websites:

https://uwescience.github.io/DSSG2016-UnsafeFoods/

https://dds-lab.github.io/disaster-damage-detection/

Summary of Tools for Reproducible Research:

https://github.com/valentina-s/presentations/blob/master/ReproducibleResearchTools.pdf