COSCS494/594 Fundamentals of Digital Archeology
Course Tools and Practices

Audris Mockus
University of Tennessee
audris@utk.edu

Data science is about working with domain experts

- You can be on any part of the team
- A team will typically have at least three kinds of expertise
    - Problem domain
    - Quantitative
    - Implementation
        - Database
        - UI
        - Server

Why use tools to communicate?

Many of your coworkers you will never meet in person

- Need to be effective with tools
  - Artifact-mediated communication
  - Share artifacts: email/documents/stack traces
  - Communicate via issues, pull requests
- Use IM/Audio/Video conferencing

Why Reproducible Research?

- Big Data analysis has
  - Many steps
  - Is implemented as scripts/databases/presentatiobns/essays
  - Takes a long time and a lot of effort
  - There will always be some error on the first try
- How do you fix such errors?
  - Start from scratch again?
    - Too much time
    - Will likely fail again
  - Record steps in an easy-to-reproduce way

Rules/Good Practices for Reproducible Research

- Keep the scripts
- Keep the data
- Keep the context
- Break analysis into parts and levels
- Record every step
- Keep track of past states of data, scripts, essays, ..

Tools supporting reproducible research

VCS, org-mode, IPython notebook, Virtual machine

- Python is lingua franca of Big Data
- Notebook is a way to combine essay, scripts, and data into a single reproducible environment
- Virtual machine
    - Preserves full operating environment
    - Can deploy to Amazon or other cloud

Why GitHub?

- It is like LinkedIn but with substance (actual work)
- Provides version control so you can reproduce the past states of your notebooks/data/essays
- Provides means to collaborate
  - Share artifacts
  - Merge work (via pull-requests)
  - Workflow and communication
    - Issues
    - Wikies
    - Event notifications
  - Organizations and teams

Everything on GitHub

- Projects
- Homework
- Issues
- Class participation includes issues resolved