ISTM 6212 - Week 2
the shell, pipelines, csvkit, data types

Daniel Chudnov, dchud@gwu.edu

2016-09-06

Agenda

- Reproducibility
- Exercise 01
- Git / GitHub review (if needed)
- The shell: input, output, pipelines
- csvkit and data types
- Exercise 02

What is "reproducibility"?

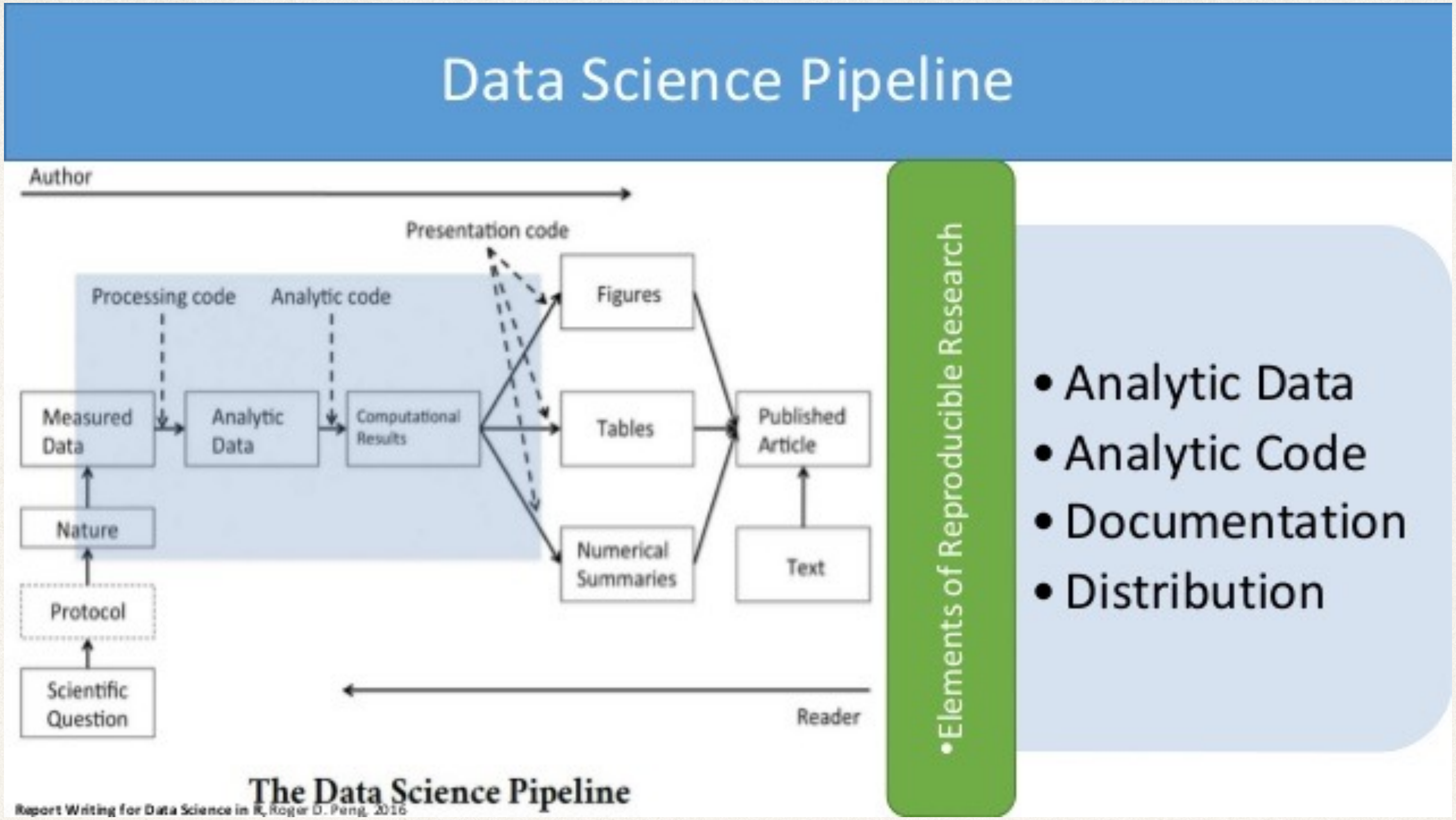
Reproducibility

- Ability to produce the same outputs from same inputs using same methods
- Sharing of analytical data and code
- Documentation of data, processing, and statistical methods
- Attention to code style, file formats, packaging
- Allows verification of methods

Why not focus on "replication"?

Why not "replication"?

- Expense
- Impracticality
- Uniqueness
- Ethics



Data Science Pipeline

The Data Science Pipeline

Report Writing for Data Science in R, Roger D. Peng, 2016

•Elements of Reproducible Research

- Analytic Data
- Analytic Code
- Documentation
- Distribution

The research/data science pipeline

(c) 2016 by Roger D. Peng

Reproducible by whom?

- Your colleagues
- Your clients / customers
- Your bosses, their bosses, and their bosses
- Your professors
- Future you

Literate programming

- Look it up in Wikipedia (really!)
- Human readable, and Machine readable
- R: sweave, RMarkdown, knitr
- For us: Jupyter!

- make it a habit
- develop it as your craft

Exercise 01

You did well!

- Setting up our pipeline:
  - Jupyter and Git/GitHub (for you)
  - datanotebook.org, nbgrader, scripts (for me)
- A few missed answers, but largely on target

Good answers (part 1)

- "The use of the capital 'F' means that the contents of the current directory will be shown in a sorted order, and directories will be marked with a trailing '/'. The lower-case 'f' means the contents will be displayed without marking directories and will be unsorted. The single dot '.' references the current directory."

Good answers (part 2)

- "The double dots are a shortcut to refer to the parent directory. In this case, the double dots, used with 'ls', show the contents of the parent directory."
- "The double dot is the directory above the current working directory. In my case, it is '/home/vagrant/Data Warehousing'."

¡For your next exercise!

- Python 3
- Submit your .ipynb file to Blackboard/GitHub
- Your "pwd" might not be my "pwd". Same with executed cell counts.
- Always write complete sentences w/proper capitalization and punctuation.
- Limit long cell output

- Give datanotebook.org a try

Git / GitHub review (if needed)

- Add / edit / move / delete your files how you like
- Tell Git what you did; Git tracks **changes**
- Publish to GitHub; GitHub facilitates **exchanges**

Git tracks **changes**

Git basics - one local repository

- init, add, commit
- status, diff
- reset
- log
- mv, rm

Git basics - more than one

- clone, remote
- branch, checkout
- fetch, merge
- push, pull
- .gitignore

GitHub facilitates **exchanges**

GitHub

- Publish your code for others to read / use / modify
- Clone others' code for you to read / use / modify
- Send code changes to other people
- Review and incorporate changes from other people
- Store your code as a remote backup

"future you" $\subseteq$ "other people"

Use GitHub's docs (they're good!)

- help.github.com esp. Bootcamp, Setup, Using Git
- Setting up your keys: laptop, in VM, desktop, etc.
- Following other people
- Tracking other projects
- Pull requests and forks