Topics in Data Science = Өгөгдлийн шинжлэх ухаан

Reproducible Research

Ryan Womack

Data Librarian, Rutgers University, https://ryanwomack.com

National Statistics Office of Mongolia, Ulaanbaatar, Mongolia, May 9, 2017

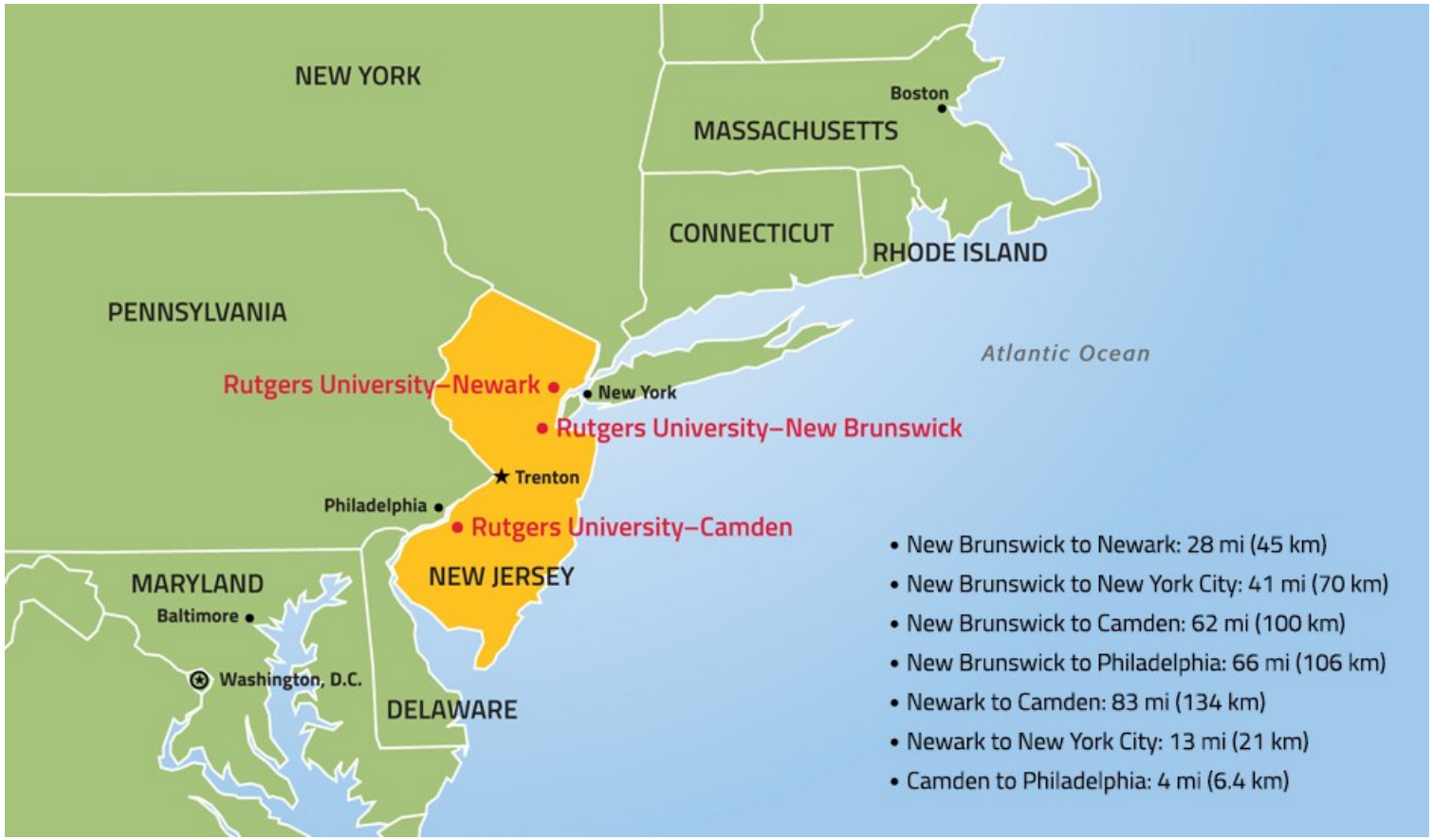
About Me

- Data Librarian
- Masters in Economics, Library Science, and Statistics
  - working in the intersection of these fields
  - help people find and use data
  - manage data for the Libraries
- interaction with advanced scholarship and research

About Rutgers

- Rutgers, The State University of New Jersey
- New Jersey, 8.8 million people, between New York and Philadelphia
- Rutgers founded in 1766, over 250 years old
- Carnegie Classification: Research - Very High Research Activity, the highest classification (R1)
- more than 100 major programs of study
- over 68,000 students (7,500 international from over 120 countries)
- Top 100 university in the world (Times Higher Education, Shanghai Ranking, CWUR, and others)

About Rutgers



About IASSIST

- IASSIST is the International Association for Social Science Information Services and Technology
- A community of over 300 data professionals working in information technology, libraries, data services, research & higher education, government, non-profit and private research sector
- IASSIST seeks to
  - Foster and promote a network of excellence for data service delivery
  - Advance infrastructure in the social sciences and beyond
  - Provide opportunities for collegial exchange of sound professional practices
- Annual conference is a forum for presentation and networking (US, Canada, Europe, next Asia?)
- IASSIST is seeking increased membership from Asia

Reproducibility: What do we mean?

- Credibility in Science
  - Scientific fraud is on the rise
- Duke "starter set" and article
  - Research misconduct is a problem, but so is human error
- Reputation, Prestige, and Funding are all affected
  - "Set the default to Open"
- Replication (redoing the experiment from scratch) is expensive, and may not be possible due to the passage of time. (see Validation)
- Science on Replication and Reproducibility
- Victoria Stodden, Friedrich Leisch, and Roger D. Peng (eds.). *Implementing Reproducible Research*. CRC Press, 2014.

"In all research that utilizes a computer, instructions for the research are stored in software and scientific data are stored digitally. A typical publication in computational research is based foundationally on data, and the computer instructions applied to the data that generated the scientific findings. The complexity of the data generation mechanism and the computational instruction is typically very large, too large to capture in a traditional scientific publication. Hence when computers are involved in the research process, scientific publication must shift from a scientific article to the triple of scientific paper, and the software and data from which the findings were generated. This triple has been referred to as a "research compendia" and its aim is to transmit research findings that others in the field will be able to reproduce by running the software on the data. Hence, data and software that permits others to reproducible the findings must be made available."
– Victoria Stodden -
http://blog.stodden.net/2014/09/28/my-input-for-theostp-rfi-on-reproducibility/

Exemplars from the Social Sciences

- ICPSR has been in operation for over 50 years, with well-established archiving practices and data documentation via codebooks and metadata
- IPUMS is reformatting and making data compatible across many decades and different projects, to enable international comparisons of microdata [*Mongolia: 1989 and 2000 data*]
- Coming from the world in the social sciences where long-term is, if not always routine, at least well-established
- Disciplinary separation of practices is diminishing when similar computational techniques can be applied to physical sciences or digital humanities

Outline

Introduction

Individuals (Everyone)

Collaborate

Team Science

Mongolia

Conclusion

- We will illustrate some practices in a few contexts
  - an individual researcher
  - a team or research group
  - ongoing, large-scale collaboration

The Data In Itself

Some basic practices:

- Keep raw data pristine and separate from any working data
- Document your variables and data collection
  - anything you yourself would forget when revisiting the project 3 years later in response to a query
  - that will be the same thing other users need too!
- Don't work in Excel [if you can] or other manual editing environment
  - you should write down all your steps if you are doing this
  - better to use code or an environment that will at least record your steps

DOI

DOI, the Digital Object identifier, is the great success story

- makes it easy to have a permanent reference and good citation practice
- usually associated with quality data repositories
- encapsulates a lot of good stuff
- moral: defined standards and centralized tools make adoption and use easy
- Treat your local data as if you were pulling it from a DOI, and you will be baking in reproducibility

Code/Data/Document Integration

We will discuss examples in R, but other programming environments support this as well (Mathematica notebooks, iPython/Jupyter)

- originally implemented in LaTeX + Sweave
- can embed R code and run it as the document is generated
- Code that is "tangled" in with text can be extracted, and formatted documents can be "woven" from the literate program.
- always ensures that the latest data and results are actually incorporated
- helps to document and explain code in context (literate programming)
- PDF, document, and HTML formats are easy to obtain

knitr and Markdown

- Markdown + knitr has become a popular, lightweight replacement
- Simple syntax and implementation
- Integrated into RStudio
- Publish documents with one click at RPubs
- Can fall back on LaTeX/Sweave for more complex document formatting

more document/program tools

- notedown converts Rmd to ipython notebook and back
- bookdown to write books in Rmd
- Zeppelin allows for interactive R sessions in the browser, and can work with Apache Spark
- R notebooks now built into RStudio
- Online notebooks provide an easy way to incorporate interactive visualizations

A short Rmd file

— title: "A short literate programming exercise" author: "Ryan Womack" date: "October 10, 2016" output: pdf_document —
“‘{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) “‘
Read in the data
Let's read in the data with the following commands:
“‘{r load} library(readxl)
download.file("https://ryanwomack.com/data/PharmaDemo.xls", "mydata.xls")
mydata<-read_excel("mydata.xls")
names(mydata)
attach(mydata)
“‘
Describe the Data
Then we will get some summary statistics on the Age and Weight variables:
“‘{r summary} summary(Age)
summary(Weight) “‘
Now plot the data:
“‘{r plot, echo=FALSE} library(ggplot2)
ggplot(mydata, aes(Weight, Age))+ geom_point() “‘
Regression
“‘{r regression} summary(lm(Age~Weight))
ggplot(mydata, aes(Weight, Age))+ geom_point()+ stat_smooth() “‘
All done!

The computing environment

- Open source is an important enabler of reproducibility
- Anyone can grab copies of the software to execute
- And can get older versions if necessary for compatibility
- You can also record information about your computing environment (session.Info() in R)
- The checkpoint package automates this process in R

Other issues

- Don't save output. Where did it come from? This should be done in the code.
- Clean, well-formatted data (tidyr) and code (formatR) are a plus
- If using "readme" files approach, document everything

Other platforms

- Jupyter grew out of iPython
  - now over 40 languages supported
- Mathematica Notebooks, cloud support
- Cloud services making sharing much easier
- Becoming an expectation

Collaboration

- The same forces (cloud computing, shared platforms, standards) are making collaboration easier than ever
- Github, Bitbucket, and others enable easy collaboration on programming
  - with significant side benefits for reproducibility due to availability of code
- The Open Science Framework provides a more data-specific approach
- A key feature is that the same platform is used for private work and then public sharing
- Psychology reproducibility study uses OSF.
  - See the Science article for a start

Enabling Collaboration

- Collaborative projects must/should agree on:
  - data practices and sharing platform
  - coding practices and sharing platform
- This is made easier by already existing platforms and practices
  - ropensci
  - ProjectTemplate
  - Data and code can be distributed through packages, start with the package.skeleton() command
  - Reprozip and others at Reproducible Science

Yale ISPS

The Yale Institute for Social and Policy Studies is an example of a research group that enables reproducibility.

- Data and papers archived together onsite
- Handles (not DOIs) for data
- Code and documentation archived
- Code review for correct execution
- Good example of providing explanatory metadata for studies
- Possible because the Institute requires compliance as a condition of grants

Open Data and Replication

- Open Data allows for discovery, sharing, and reuse
- (Open Data also often refers to government data)
- Landscape of research data repositories
- Many layers of Open
- Open Data in a Big Data World
- Replication (based on Open Data), validates and confirms science
- Replication Wiki

Big Science

Multi-year, multi-institutional projects that may continue beyond original PIs *require* reproducibility.

- Many people will be coming on and off of the project over time
- Many unanticipated uses are anticipated ("known unknowns" or something like that)
- Collaboration and continuity must be consciously planned for
- Decisions should be made with more consideration for future use than current convenience
- But disciplinary expertise is building in these areas
- Protein Data Bank
- Also, https://www.teamsciencetoolkit.cancer.gov/

Big Science - Infrastructure

- In big science, the main node(s) are enablers of future reuse
- They provide basic infrastructure (OOI - Ocean Observatories Initiative)
- But also provide a clearinghouse for other projects that link to and build on the central node
- One major future goal is to have more generic, all-purpose collaborative infrastructure
- Rutgers is developing a Virtual Data Collaboratory for this purpose
- Open infrastructure like Zenodo, Dataverse, OpenICSPR and the Open Science Data Cloud have been developed

Big Science - Practices

- Standards such as DOI and ORCID enable the broader community to coalesce around good practice
- Data repositories are developing standards
- Re3data.org is a directory of repositories
- The Data Seal of Approval is awarded to repositories using sound data practices
- ISO 16363 (Trusted Digital Repositories) is a more stringent standard
- One important step is to plan for what happens when the project winds down (expectedly or unexpectedly)
- What is the equivalent standard for computing and reproducibility?

Big Science - Genomic Data Sharing

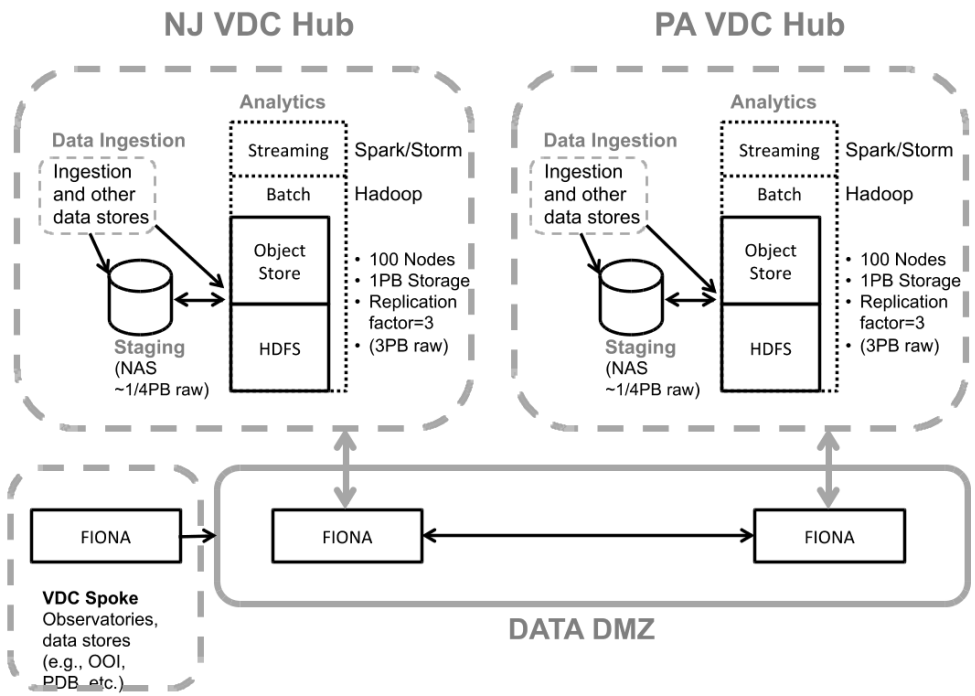
- Massive investments, massive amounts of data
- Many repositories too (NIH GDS - National Institutes of Health Genomic Data Sharing)
- Existing repositories are useful and aggregate many software tools
- But researchers want even larger pooled databases, especially for human genome
- Technical issues are complex, but the rights and permissions involved are equally complex
- How can data be federated for maximum discoverability?

Virtual Data Collaboratory

- A virtual environment for sharing and analyzing big data, enabling reuse, discovery and collaboration [Rutgers and Penn State leads]

RUTGERS RDI²

Computing Platform



Open Data in Mongolia

- Chinggis Khan - First Census in 1206
- Mongolia ranks 3rd in Openness for Government Data (of 125 developing countries) using OpenDataWatch information.
- Open Data Initiative of Mongolia
- NUM and NITP collaborating on Open Data
- Mongolia GIS Data
- A good foundation for data sharing and reproducible research!

Open Data -> Reproducible Research

- Increasing openness is a long-term trend
  - Internet, Government, Data, Software, Cloud Computing
- Pressure for Reproducible Research can only increase as these trends intensify
- Good news is...
  - Benefits are clear for society and for knowledge creation!
  - Tools to enable this are getting easier all the time!
  - Eventually it will be a standard we will take for granted!

Төгсгөл

- Асуулт байна уу?
- Маш их баярлалаа!
- Намайг дагаарай...
  - https://youtube.com/librarianwomack
  - https://www.linkedin.com/in/ryanwomack
  - https://twitter.com/ryandata
  - https://ryandata.wordpress.com