



W04: Pragmatic Reproducible Research for Analysis, Dissemination and Publication

Luke Rasmussen

Northwestern University

@lrasmus

Eric Whitley

Northwestern University

Instructor Disclosures



Luke Rasmussen - I and my spouse/partner have no relevant relationships with commercial interests to disclose.

Eric Whitley – I and my spouse/partner have no relevant relationships with commercial interests to disclose.

Acknowledgements



This represents joint work with:

- Leah J. Welty, Project Director
- Abigail S. Baldridge, Biostatistician



Supported in part through a Clinical Translational Sciences Award presented by National Institutes of Health to Northwestern University Clinical and Translational Sciences Institute (UL1TR001422).

The content is solely the responsibility of the instructors and does not necessarily represent the official views of the National Institutes of Health.

We want you to enjoy this workshop!



Let's make sure what you expect to get out of this workshop matches what we're planning to cover.

What you'll learn...

- What is (and isn't) reproducible research
- How to apply the concept of reproducible research to your projects
- How to assess tools, data sets and publication options when planning a reproducible project

What we won't have time to cover...

- Every step needed to clean your data
- All of the tools out there that help with reproducible research (*we will discuss a few*)
- How to master all of the tools we mention (*you will get an idea how a few can be used*)

Agenda



1. Overview of reproducible research (20 minutes)
2. Thinking pragmatically about reproducible research (20 minutes)
3. Exercise: Teams and Data (15 minutes)
4. Use Case – EHR-Based Phenotyping (Part 1) (25 minutes)
5. Exercise: Analysis (10 minutes)
6. Break (30 minutes)
7. Use Case – EHR-Based Phenotyping (Part 2) (25 minutes)
8. Exercise: Documentation (15 minutes)
9. Use Case – EHR-Based Phenotyping (Part 3) (30 minutes)
10. Exercise: Dissemination (15 minutes)
11. Wrap-up (5 minutes)

Exercises



- Goal: Applying the concepts learned
- Methods and Concepts (Primary)
 - Guide – provides questions and prompts to consider
 - Worksheet – spaces to jot down your notes
- Example Software (Secondary)
 - Examples available on GitHub repo

Think of a past / current / future project that you'd like to make reproducible. After each section, think through the prompts in the guide and apply what we covered. Are there new considerations you found that we didn't cover?



Overview of Reproducible Research



Introduction



Origins lie in the inconvenience of irreproducible research

“In the mid-1980s, we realized that our laboratory’s researchers often had difficulty reproducing their own computations without considerable agony.”

Schwab M, Karrenbach M, Claerbout J.
“Making Scientific Computations Reproducible”. Computing in Science & Engineering 2000 2:6, 61-67

MAKING SCIENTIFIC COMPUTATIONS REPRODUCIBLE

To verify a research paper’s computational results, readers typically have to recreate them from scratch. ReDoc is a simple software filing system for authors that lets readers easily reproduce computational results using standardized rules and commands.

In the mid-1980s, we realized that our laboratory’s researchers often had difficulty reproducing their own computations without considerable agony. We also noticed that junior students, who typically build on the work of more advanced students, frequently spent a great deal of time and effort just to reproduce their colleagues’ computational results.

Reproducing computational research poses challenges in many environments. Indeed, the problem occurs wherever people use the traditional methods of scientific publication to describe computational research. For example, in a

recent progress in electronic publishing) they can only recompute the results by invoking the various programs exactly as the author invoked them; such information is something that is usually undocumented and difficult to reconstruct.

To address these problems, we developed ReDoc, a system for reproducing scientific computations in electronic documents. Since implementing it in the early 1990s, ReDoc has become our principal means for organizing and transferring our laboratory’s scientific computational research.

ReDocs are best defined operationally: After

Insanity

"doing the same thing over and over and expecting a **different** result"

Irreproducible research

"doing the 'same' thing over and over and getting a **different** result"

Reproducible research

"doing the same thing over and over and getting the **same** result"

What Isn't Reproducible Research?



There are many best practices and concepts that can be confused with it:

Open Science

- Open data
- Open source
- Preregistration of studies
- Prepublication servers

Reproducible research can be disseminated openly.

Replication

- “...*the chances other experimenters will achieve a consistent result...*” – Leek & Peng (2015)

Reproducible research can help others replicate your work.

Tools / Software Engineering

- Unit tests
- Bash scripts
- Any specific programming language (Python, C++)
- Any specific tool (Jupyter Notebooks, Docker)

Reproducible research can be automated.

Scientific Rigor

- Are the methods used the right ones?
- Quality of the data collected and used
- “...*invalid reports can do more harm than irreproducible reports*” – Shiffren, Borner & Stigler 2018

Reproducible research can help us evaluate our methods.

Kinds of Reproducible Research



- Computational Reproducibility
 - Have all of the information about the code, software and hardware.
 - Can re-run it on the same data set
- Empirical Reproducibility
 - Able to reproduce the non-computational experiments
 - Requires very clear methods
- Statistical Reproducibility
 - Full details about statistical tests
 - Pre-registration of studies to prevent p-value hacking

Stodden, 2014 - <https://www.edge.org/response-detail/25340>

The journey to transparency, reproducibility, and replicability FREE

Suzanne Bakken

Journal of the American Medical Informatics Association, Volume 26, Issue 3, 1 March 2019,
Pages 185–187, <https://doi.org/10.1093/jamia/ocz007>

Published: 25 January 2019

■ Split View PDF Cite Permissions Share ▾

Regardless of the type of biomedical and health informatics research conducted (eg computational, randomized controlled trials, qualitative, mixed methods), transparency, reproducibility, and replicability are crucial to scientific rigor, open science, and advancing the knowledge base of our field and its application across practice domains. These principles are also essential to high-quality publications in

Journal of the American Medical Informatics Association, Volume 26, Issue 3, 1 March 2019,
Pages 185–187, <https://doi.org/10.1093/jamia/ocz007>

reflected by explicit, clear, and open communication procedures used to obtain the research results (ability to repeatedly obtain the same results from other investigators to observe the same results following paragraphs, I summarize key strategies well as my own thoughts in 4 categories (data, applicable, describe their relationship to published apply across types of research, the relevance of some strategies varies.



Importance to our field:

- Scientific rigor
- Open science
- Advancing knowledge

Does health informatics have a replication crisis? ORCID iD

Enrico Coiera , Elske Ammenwerth, Andrew Georgiou, Farah Magrabi

Journal of the American Medical Informatics Association, Volume 25, Issue 8, 1 August 2018, Pages 963–968, <https://doi.org/10.1093/jamia/ocy028>

Published: 13 April 2018 Article history ▾

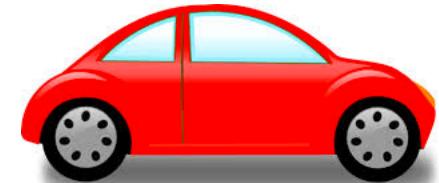
Pragmatic Benefits



- Allows us to go back and know exactly what we did
 - 2015 Luke – “I’ll totally remember where I put that data set”
 - 2019 Luke – “AGGGH, 2015 LUKE!!!”
- Guides us to being better scientists
 - Forces us to think through parts of the process
 - Still doesn’t guarantee accuracy
- A little can go a long way
 - Not an all-or-nothing endeavor

The Hard Truth

- Up-front investment in time
- Not everyone finds it “fun”
- ROI comes over time
- When you need it, you’ll be happy you did it



Definition For Today



“...the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline...”

Leek & Peng, PNAS February 10, 2015; 112 (6)

<https://doi.org/10.1073/pnas.1421412111>

What Could Change?



- People
- Data
- Project-specific analysis code
- Supporting software – statistical package, compiler, interpreter
- Operating system
- Hardware

It is about reducing **variation**

Summary



- Many definitions of “reproducible research” exist
- For today: the ability to recompute results given the data and code
- Goal is reducing variation
- Start small, and work your way up
- It’s okay to do this just for you!

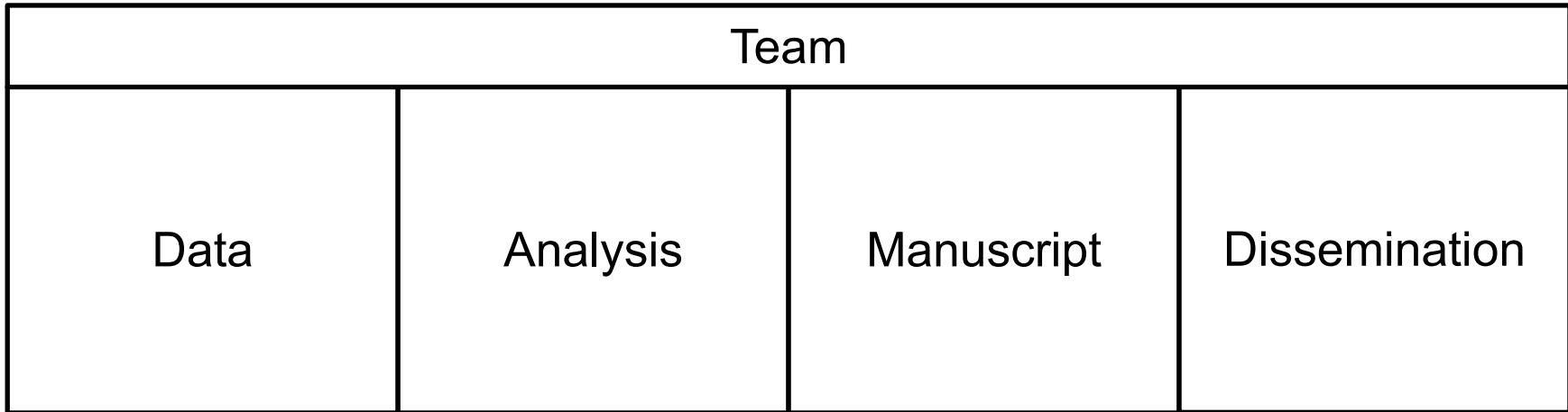


Thinking Pragmatically About Reproducible Research

Assessing and planning for your projects



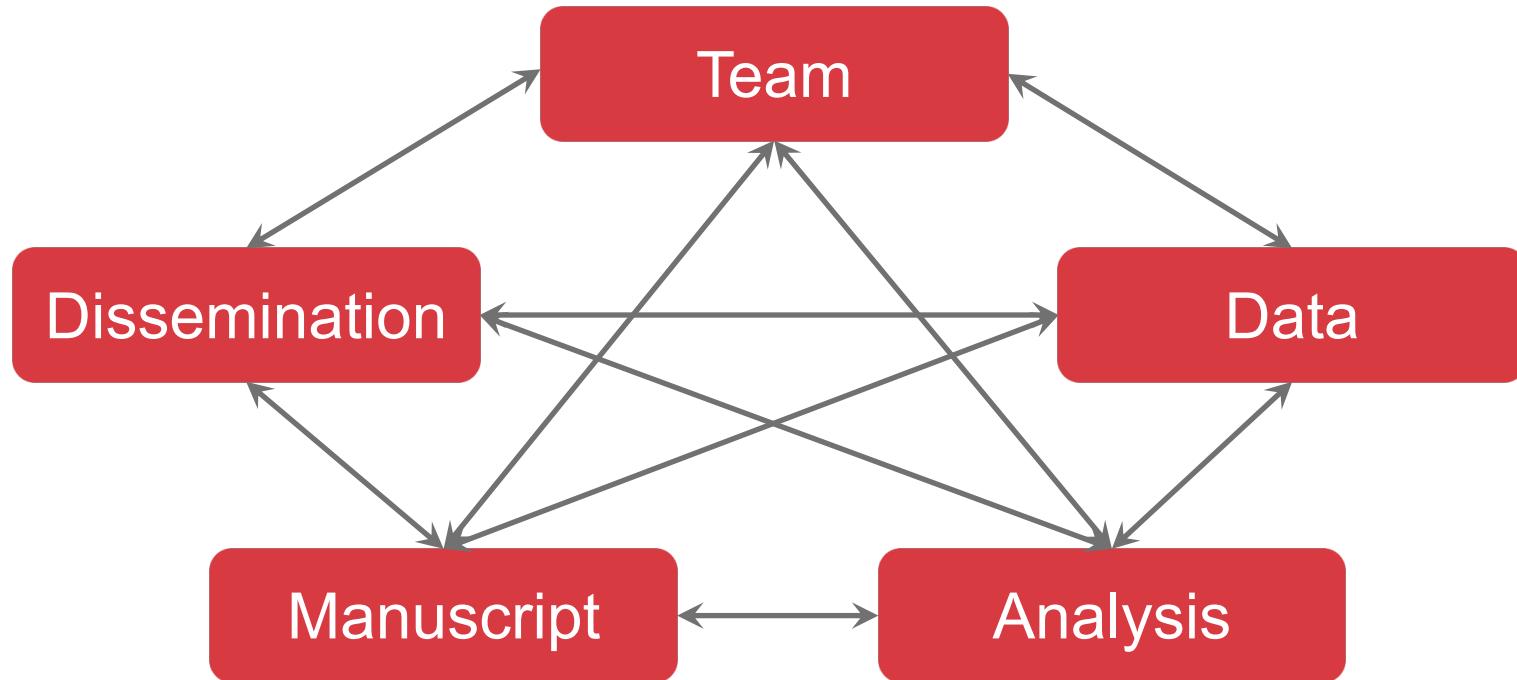
Reproducing the Research Workflow



Tools for reproducible research

Makefiles (Automation), Jupyter Notebook (Electronic Lab Notebook), Git (Source Control), StatTag, R Markdown (Dynamic Documents), Open Science Framework (Documentation)

Thinking Through the Process



Teams



- Who is on the team at any given time can change
 - “Hit by a bus” contingency plan
- Expertise, experience, and willingness to change can differ
 - Preferred tools, technologies, workflow
 - Reproducible research is a cultural change
- Collaboration is great, coordinated collaboration is better
 - Mutual understanding on use of reproducible research
 - Roles and responsibilities
 - Project file structure

Structuring Your Project

Having a central storage location and systematic file/path structure: (1) assists BCC faculty and staff who are collaborating on the same project; (2) protects against transitions in a project team; and (3) provides a long term record of work provided to investigators.

Location:

BCC Project Folders should reside in the following location:

- <\\fsmresfiles fsm.northwestern.edu\fsmresfiles\PrevMed\Projects\BCC\Projects>

Project Folder Naming Convention:

Within the above folder, the Project Folder should follow this naming convention:

PI LastNameFirstInitial_REDCap Project Number

Example: For PI Jane Doe, and REDCap Project Number 123, the file name should be DoeJ_123

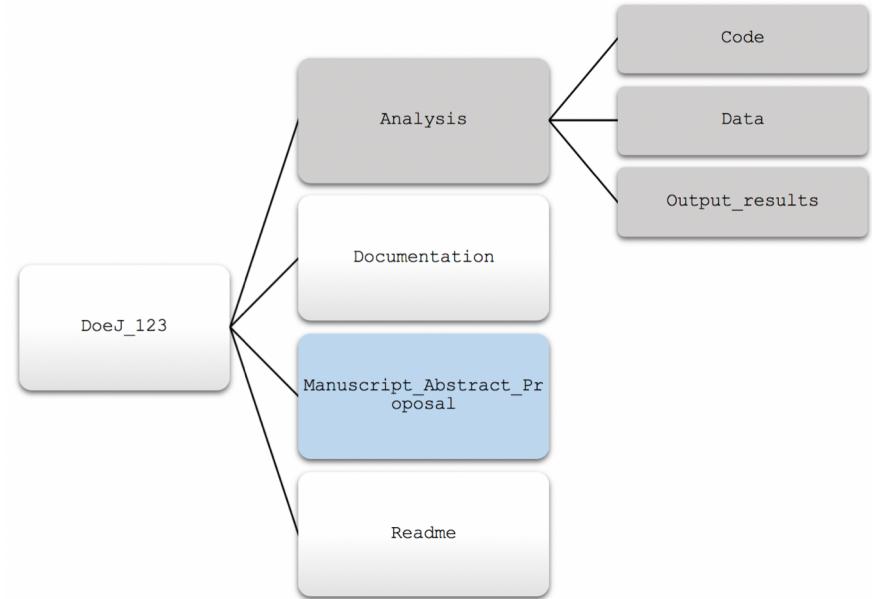
- The folder name does not include the project description - this avoids lengthy path/folder names. Project descriptions can be found in the Readme subfolder.
- Do not include leading zeros in the Project Number.

The diagram which follows illustrates the recommended folder structure and substructure.

Additional folders may certainly be added as needed.

Date Requirements for Code/Programs/Manuscripts

Code/programs/manuscript files should include date information either inside the programs or as part of the file names. The code header should also include at the minimum the authors' names and the date last updated.



Northwestern Biostatistics Collaborative Core
L. Welty, personal communication, March 21, 2019

- Sensitivity of the data will impact how you approach reproducible research
 - PHI
 - Institutional policies on data handling
 - Considerations for dissemination
- Technology for collecting and storing your data
 - Who can access it?
 - How is it accessed?
 - Flat file export
 - Real-time query / API call

Data Sensitivity

Additional Reference(s):

- * [NUIT Data Access Policy](#)
- ** [HIPAA Privacy Protected Health Information](#)
- *** [HIPAA Privacy Limited Data Sets \(LDS\)](#)

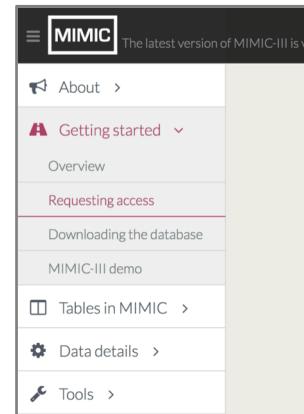
3) Identify the level(s) of data sensitivity that will be collected/maintained during the research. Example categories of sensitivity include HIPAA PHI and/or non-PHI personally identifiable information (PII) <select one or more options>:

- Legally/Contractually Restricted (FERPA, Illinois Personal Information Protection Act) Information*
- NU Internal Information (see NUIT Data Access Policy*)
- Protected Health Information (HIPAA defined)**
- LDS as a subset of Protected Health Information***
- De-identified information**
- Public Information*
- Other <please explain>

Source: Northwestern University
Feinberg School of Medicine, Central IT
Data Security Plan Template v1.4

Data

- Public data sets need consideration too
 - You may not be able to redistribute, but can give instructions on accessing
 - Structure analysis pipeline with a placeholder for the data
- Is the data source changing?
 - e.g., EHR, surveys
- Accessing data in the future
 - Is the data set versioned?
- What version of data did I use?
- How do you know it's the same?
 - Checksum (e.g., MD5 hash)



The screenshot shows the MIMIC-III website homepage. At the top, it says "The latest version of MIMIC-III is v1.4". Below that is a navigation menu with the following items:

- About >
- Getting started > (highlighted in red)
- Overview
- Requesting access
- Downloading the database
- MIMIC-III demo
- Tables in MIMIC >
- Data details >
- Tools >

Requesting access

The latest version of MIMIC is MIMIC-III v1.4, which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. The data spans June 2001 - October 2012. The database, although de-identified, still contains detailed information regarding the clinical care of patients, so must be treated with appropriate care and respect.

Researchers seeking to use the database must formally request access with the steps below.

<https://mimic.physionet.org/gettingstarted/access/>

Teams

- Variation – who is on the team at any given time
- Willingness to try reproducible research
 - Even on a large team, you can still make your work reproducible!
 - Project structure may be a place to start

Data

- Variation – what is in the data at any given time
- Consider licenses, agreements, and regulations in storage and sharing
- How can I be reasonably sure each time my data is the “same”?

Exercise: Applying the Concepts to Your Project

Team and Data



Use Case – EHR-Based Phenotyping (Part 1)

Technologies we will touch on:

- Electronic lab notebooks (Jupyter Notebooks)

Use Case Overview



You want to implement a Type 2 Diabetes Mellitus phenotype at your institution.

Your plan is to approach this in as reproducible a way as you can.

Use Case - The Team & Work



1) Data Warehouse Team

- Data extraction experts

→ *Generate patient data*

2) Ontologist

- Medical terminology expert

→ *Map source data to required target standard(s)*

3) Data Analyst – focus on phenotyping

- Loves Python

→ *Implement phenotype algorithm*

4) Data Analyst 2 – focus on statistics

- Loves R

→ *Analyze phenotype results*

5) Physician – PI

- Medical expert, but largely non-technical

→ *Author paper*

The Tools



- 1) Data set provided by source system
 - 1) Synthetic patients generated by MITRE Synthea
- 2) Data set modified (cleaned)
 - 1) Mapped from SNOMED CT to ICD-9 where required using UMLS and Python
- 3) Data set processed (phenotype)
 - 1) Jupyter Notebook
- 4) Phenotype-annotated data analyzed
 - 1) R / R Markdown
- 5) Paper authored
 - 1) Microsoft Word



The Tools



- 1) Data set provided by source system → *patient_data/*.csv*
 - 1) Synthetic patients generated by MITRE Synthea
- 2) Data set modified (cleaned) → *1_generate_dx_terms.py*
(UMLS not provided)
 - 1) Mapped from SNOMED CT to ICD-9 where required using UMLS and Python
- 3) Data set processed (phenotype) → *2_process_algorithm.ipynb*
 - 1) Jupyter Notebook
- 4) Phenotype-annotated data analyzed → *synthea_analysis.rmd*
 - 1) R / R Markdown
- 5) Paper authored → *synthea_paper.docx*
 - 1) Microsoft Word

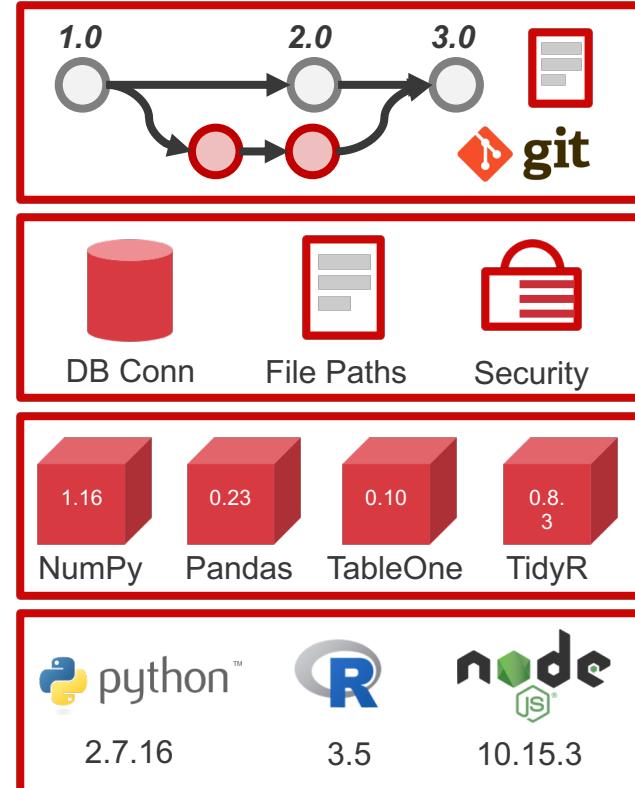
Analysis

Think of your work as a series of **variables**

- + Source code revision
- + System / application / user configuration
- + Module / package versions
- + Language / framework versions

The variables are **combined** to create a single overall analysis pipeline

A change to even a *single* variable can influence reproducibility



The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild , Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: June 1, 2012 • <https://doi.org/10.1371/journal.pone.0038234>



Article	Authors	Metrics	Comments	Media Coverage
▼				

Abstract

Introduction

Materials and Methods

Results

Discussion

Supporting Information

Acknowledgments

Author Contributions

References

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average $8.8 \pm 6.6\%$ (range 1.3–64.0%) (volume) and $2.8 \pm 1.3\%$ (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

Reader Comments (5)

Media Coverage (1)

Figures

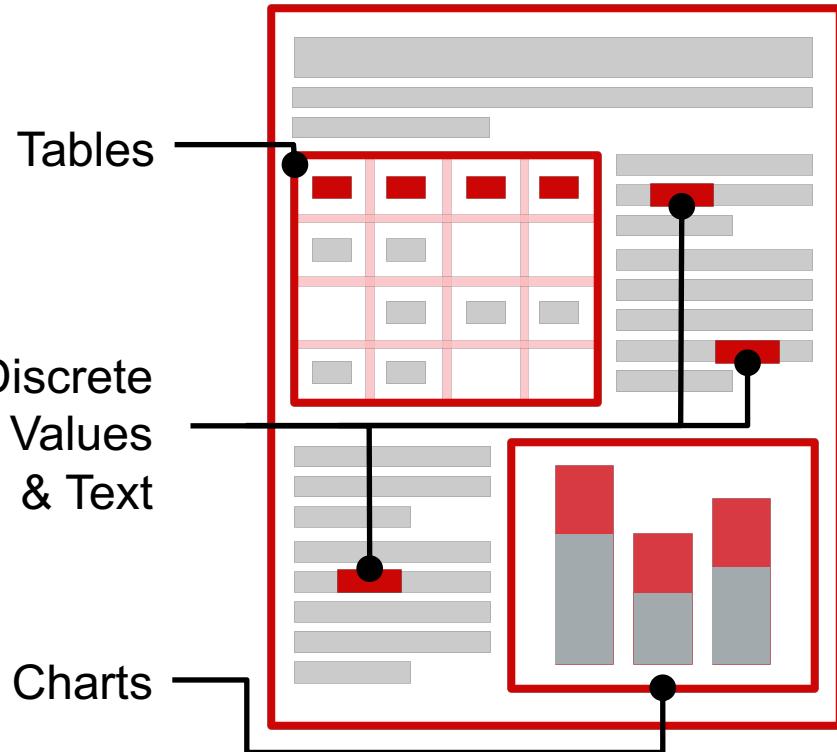
The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Lab Notebook Platforms

“Weave” code chunks and text into a dynamic document



These tools provide an interactive environment that allows you to run multiple languages in a single session – with several forms of visual output



Lab Notebook Platforms

Acts as a Process Pipeline

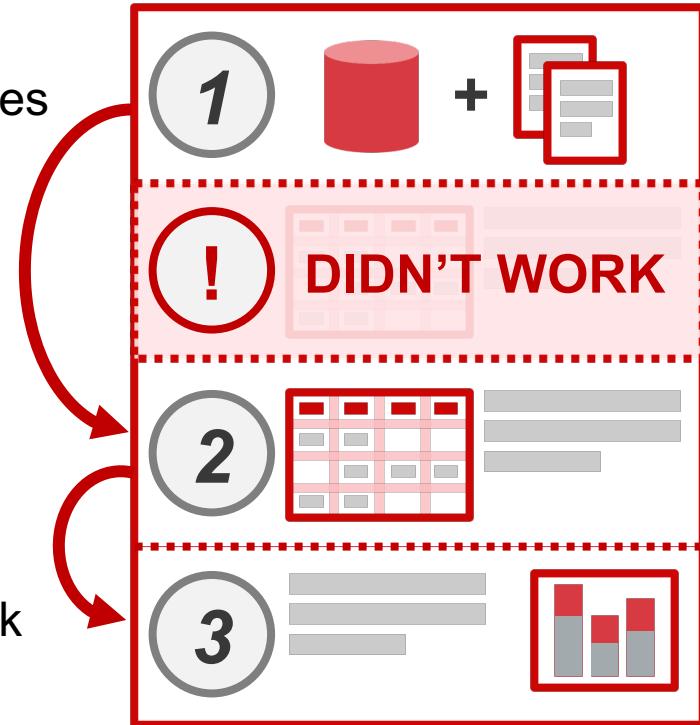
- Automates & consolidates discrete processes
- Easy to review – highly visual

Shows Provenance

- Trace your work
- See unsuccessful paths
- Trace the ultimate right path

Supports Exploration and Creativity

- Interactive – feedback for faster, easier work
- Doesn't mandate “perfection” – keep *everything* - that's part of the value



Demo



Exercise: Applying the Concepts to Your Project

Analysis



Break

We will start up again at 10:30



Use Case – EHR-Based Phenotyping (Part 2)

Technologies we will touch on:

- R / R Markdown
- StatTag

Manuscripts



- Copy / Paste to bring results into paper
 - Easy to do
 - Time consuming & error prone
- “Weaving” approach can be extended beyond notes to manuscripts
 - Develop code specific to what you need in the manuscript

Dynamic Documents: R Markdown



```
1 ---  
2 title: 'Association of Education with Anthropometrics in US Adults: National H  
3 and Nutrition Examination Study 2013-2014'  
4 geometry: margin=.75in  
5 output: html_document  
6 sansfont: Calibri Light  
7 fontsize: 11pt  
8 ---  
9  
10 <style type="text/css">  
11 h1.title {  
12   font-size: 11pt;  
13   font-weight: bold;  
14 }  
15 </style>  
16  
17 ````{r setup, include=FALSE}  
18 knitr::opts_chunk$set(echo = FALSE)  
19 setwd("R:/NUCATS/NUCATS_Shared/BERDShared/Analysis Manager/Data and Programs/R/  
20 analysis<-read.csv("Analysis.csv")  
21 attach(analysis)  
22 library(tableone)  
23 library(knitr)  
24 options(digits=2)  
25  
26  
27 \usepackage[utf8]  
28  
29 **Introduction:** Education level has been shown to be associated with body ma  
metabolic characteristics may alter this association.  
30  
31 **Methods:** This study included adult ( $\geq 30$  years) participants from the  
Examination Study (NHANES). Education and demographic information were assesse  
measurements were taken by study personnel. Education was dichotomized based o  
Associations were estimated using T-tests or wilcoxon rank sum tests for conti  
data. We examined the association between BMI and education level using multiv  
32  
33 ````{r, echo=FALSE}  
34 ps<- table(analysis$PostSecondary)  
35 percents<-100*prop.table(ps)  
36 model <- lm(BMXBMI ~ PostSecondary, data = analysis)  
37 Betaunivariable <- summary(model)$coefficients[2, 1]  
38 LBunivariable <- (summary(model)$coefficients[2, 1]) - 1.96 * (summary(model)$  
39 UBunivariable <- (summary(model)$coefficients[2, 1]) + 1.96 * (summary(model)$  
40 myvars <- c("Gender", "Race", "RIDAGEYR", "Married", "BMXBMI", "LBXTC", "LBXGL  
41 Tableone <- createTableone(data = analysis, vars=myvars, strata= "PostSecondar
```

Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014

Introduction: Education level has been shown to be associated with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

Methods: This study included adult (≥ 30 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were assessed by questionnaire and anthropometric measurements were taken by study personnel. Education was dichotomized based on matriculation to post-secondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate linear regression.

Results: Among 4808 participants, 2649 (55.1%) self-reported any post-secondary education. Post-secondary education was associated with lower BMI (Beta: -0.62, 95% CI: -1.03 to -0.21). After adjusting for gender, race, age, marital status, fasting glucose and total cholesterol, post-secondary education was no longer significantly associated with BMI (Beta: -0.19, 95% CI: -0.79 to 0.41).

Table 1. Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	
Non-Hispanic Asia	161 (7.5)	411 (15.5)	

Demo



Dynamic Documents: R Markdown



```
NHANES Example.Rmd x
1 ---  
2 title: 'Association of Education with Anthropometrics in US Adults: National H  
3 and Nutrition Examination Study 2013-2014'  
4 geometry: margin=.75in  
5 output: html_document  
6 sansfont: Calibri Light  
7 fontsize: 11pt  
8 ---  
9  
10 <style type="text/css">  
11 h1.title {  
12   font-size: 11pt;  
13   font-weight: bold;  
14 }  
15 </style>  
16  
17 ```{r setup, include=FALSE}  
18 knitr::opts_chunk$set(echo = FALSE)  
19 setwd("R:/NUCATS/NUCATS_Shared/BERDShared/Analysis Manager/Data and Programs/R/  
20 analysis<-read.csv("Analysis.csv")  
21 attach(analysis)  
22 library(tableone)
```

Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014

Introduction: Education level has been shown to be associated with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

Methods: This study included adult (≥ 30 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were assessed by questionnaire and anthropometric measurements were taken by study personnel. Education was dichotomized based on matriculation to post-secondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate regression analysis.

Do you have non-technical collaborators who are willing to work this way? Think of a clinical expert working on a phenotype algorithm paper.

31 **Methods:** This study included adult (≥ 30 years) participants from the National Health and Nutrition Examination Study (NHANES). Education and demographic information were assessed by questionnaire and anthropometric measurements were taken by study personnel. Education was dichotomized based on matriculation to post-secondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate regression analysis.

32
33 ```{r, echo=FALSE}
34 ps<- table(analysis\$PostSecondary)
35 percents<-100*prop.table(ps)
36 model <- lm(BMXBMI ~ PostSecondary, data = analysis)
37 Betaunivariable <- summary(model)\$coefficients[2, 1]
38 LBUnivariable <- (summary(model)\$coefficients[2, 1]) - 1.96 * (summary(model)\$
39 UBUnivariable <- (summary(model)\$coefficients[2, 1]) + 1.96 * (summary(model)\$
40 myvars <- c("Gender", "Race", "RIDAGEYR", "Married", "BMXBMI", "LBXTC", "L BXGL"
41 Tableone <- createTableone(data = analysis, vars=myvars, strata= "PostSecondary")

(Beta: 0.19, 95% CI: 0.77 to 0.71).

Table 1. Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	
Non-Hispanic Asia	161 (7.5)	411 (15.5)	

Dynamic Documents and Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back this

I have two choices:

1. Continue in Word, and loose the dynamic nature of the document.
2. Re-enter all of their changes in my source file.

Association of Education with Anthropometrics in US Adults: Results from the National Health and Nutrition Examination Study 2013-2014

Introduction: Education level ~~may has been shown to be associated~~ with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

Methods: ~~We studied This study included~~ adult (≥ 30 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were ~~self-reported assessed by questionnaire~~ and anthropometric measurements were taken by ~~trained~~ study personnel. Education was dichotomized based ~~on matriculation to~~ post-secondary education (~~yes/no~~). Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate linear regression.

~~4808 participants. We found that 2649 (55.1%) of the 4808 participants~~
~~post-secondary education. Participants with some post-secondary~~
~~secondary education was associated with significantly lower BMI (Beta: -0.3 to -0.21), although. After adjusting for gender, race, age, marital~~
~~status and total cholesterol, the association was no longer statistically~~
~~secondary education was no longer significantly associated with BMI (Beta: 0.9 to 0.41).~~

Table 1: Association of Education with Participant Characteristics among $n = 4808$ NHANES Participants

	No Postsecondary	Postsecondary	p
Gender = Male (%)	2159	2649	
Race (%)	1079 (50.0)	1208 (45.6)	0.003
Mexican American	456 (21.1)	173 (6.5)	<0.001

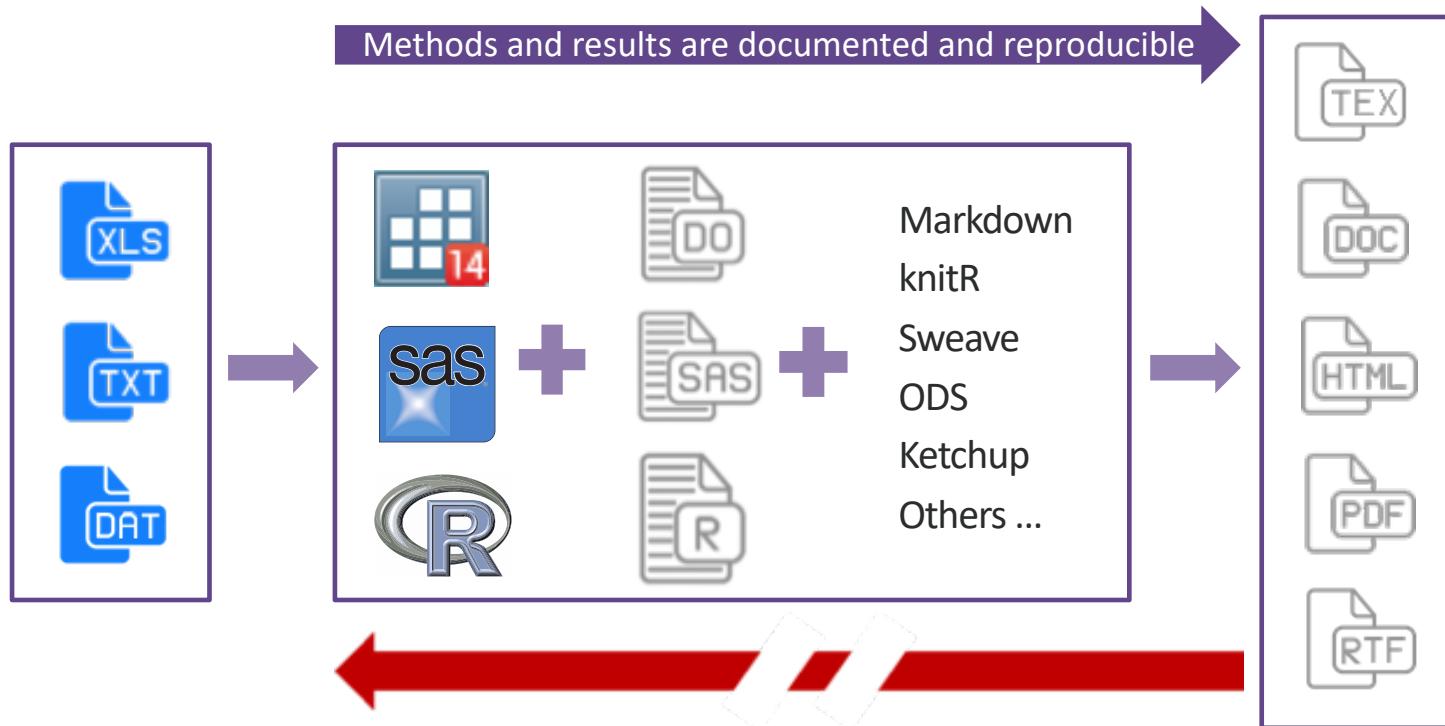
Leah J Welty A few seconds ago Leah, the spacing in this table is too far apart. Can you please make in single-spaced? And make the middle two columns wider?

Leah J Welty Formatted: Font: Bold

Leah J Welty Formatted: Font: Bold

Leah J Welty Formatted: Font: Bold

Limitations: Unidirectional Flow



Alternative Approach: StatTag



- StatTag creates a link between a statistical code file and a Word document
- Embeds output (values, tables, figures, verbatim) in document
- Can work separately on the code and the Word document but retain link
- Software agnostic: connects Stata, SAS or R code and Word document
- Can connect multiple code files (of different types) to the same document

<http://stattag.org>



Demo



Exercise: Applying the Concepts to Your Project

Manuscript



Use Case – EHR-Based Phenotyping (Part 3)

Technologies we will touch on:

- Containers (Docker)

Expanding to NLP



- Expand our phenotype to incorporate natural language processing
- Colleague has some Python 2 code that calculates frequency distribution of top 50 words ("nltk-freq-dist")
 - Very specific / curated set of stop words
- Sends you a ZIP containing the Python script and stop word list
- You try to run it on your machine (you're using Python 3)

```
nltk-freq-dist $ python freq-dist.py
  File "freq-dist.py", line 12
    print "Usage: nltk-freq-dist input-dir/ output-dir/"
                                         ^
SyntaxError: Missing parentheses in call to 'print'. Did you mean print("Usage:
nltk-freq-dist input-dir/ output-dir/")?
```

Finding the variation

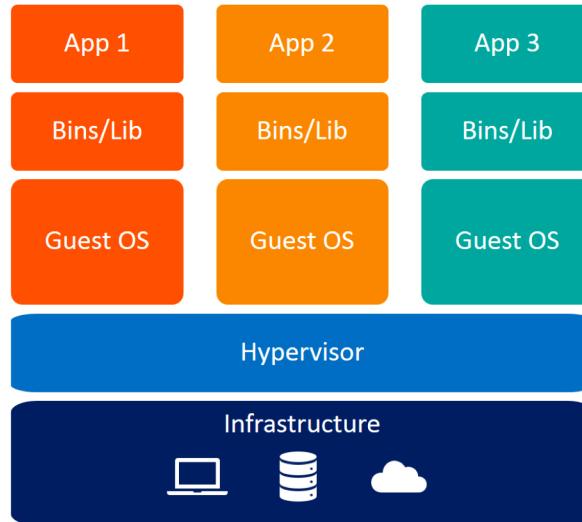
- Differences in Python versions
 - Does this mean you need to install Python 2?
 - Will that mess up any other programs running on your machine?
- How long will Python 2 be available to download?

DEPRECATION: Python 2.7 will reach the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 won't be maintained after that date. A future version of pip will drop support for Python 2.7.

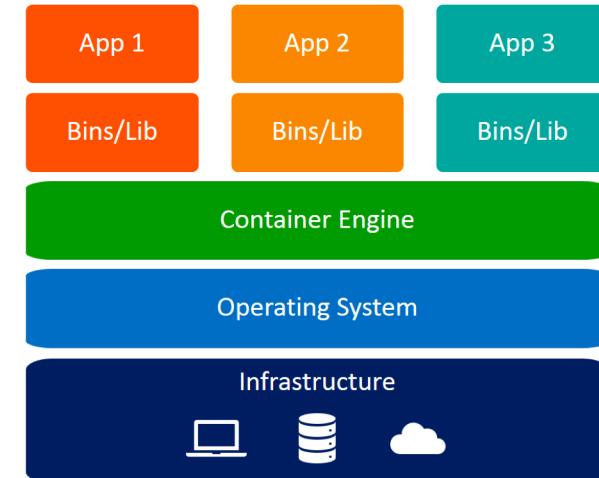
- Same considerations for the underlying dependencies / modules
 - NLTK (<https://www.nltk.org/>)
- We want to have a snapshot of the environment

Virtual Machines and Containers

Abstractions
of physical
hardware



Virtual Machines



Containers

Abstractions
of software

Source: <https://www.bmc.com/blogs/containers-vs-virtual-machines/>

Building a Docker image



- The Docker **image** is the actual snapshot of the stack (immutable)
- A running instance of the image is a Docker **container**
 - Mutable when running, but not stateful
- We want to have an image for the program that we can keep and share
- We can't store any of the data in the image – need data to come in
- We want to save the results of the processing – need to preserve state

Collecting the Pieces



nltk-freq-dist

Our source code

NLTK 3.3

Library used by source code

Python 2.7

Language interpreter

Linux

Operating system

Docker image layer cake



nltk-freq-dist

For our image, we start with our base, then “do stuff” to add our software and dependencies. When done, it’s a static snapshot of the environment.

Python 2.7.16

We’ve started with the Python 2.7.16 image as our base. It has the Alpine Linux image as its base

Alpine Linux

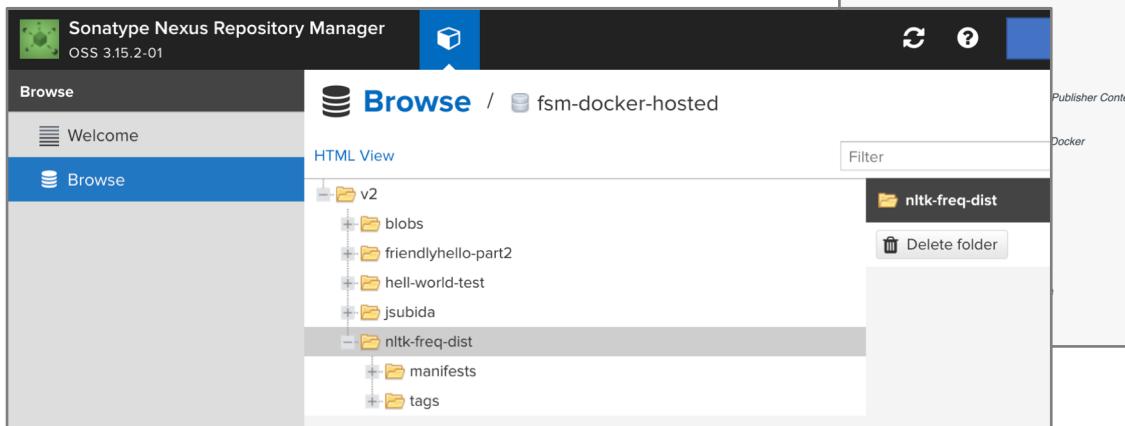
We aren’t particular about the exact OS in this case. We just need it to be consistent.

Demo



Making Images Available

- Docker Hub - <https://hub.docker.com/>
 - Public or private
- Internally hosted repos
- Your machine



The screenshot shows the Docker Hub search results for "python". The search bar at the top has "python" entered. Below it, there are tabs for "Explore", "Repositories", and "Organizations". The main content area shows a list of results with the following details:

1 - 25 of 35,852 results for **python**. [Clear search](#)

Image	Name	Last Updated	Description	Tags
	python	Updated 25 minutes ago	Python is an interpreted, interactive, object-oriented, open-source programming language.	Container, Linux, Windows, ARM, x86-64, IBM Z, PowerPC 64 LE, 386, ARM 64
	pypy	Updated 24 minutes ago	PyPy is a fast, compliant alternative implementation of the Python language.	Container, Linux, ARM, x86-64, 386, Programming Languages

Cloud-Based Solutions



Published Demonstration of soundscape separation by using the Soundscape_Viewer (Tzu-Hao Lin)

Capsule File Edit View Tabs Settings Help

Sign up or login to edit and run

Files

- environment
- code
 - LICENSE
 - LTSA_PCNMF.m
 - LTSA_combine.m
 - LTSA_context_analysis.m
 - LTSA_k_means.m
 - README.md
- ReconstructFromDecomposition.m
- Soundscape_viewer.m
- Unsupervised_classify.m
- ValidateParameters.m
- basis_exchange.m
- button_action.m
- button_action2.m
- clustering_interface.m
- demo.m
- getfile.m
- matrix_mean.m
- matrix_standardization.m
- nmfsc.m
- nmfsc_clustering.m
- projfunc.m
- run
- seminmf.m
- sequential_matrix.m
- source_number.m

data Manage Datasets

- LICENSE 18.21 KB
- ML_S3_20140911-20141012_5min... 6.97 MB

Sign up or login to edit and run

Metadata Edit Capsule Sign up

Re-Run

Timeline

November 16, 2018 Published Version 1.0 Currently viewing

Author ran November 16, 2018 0:00:51

Published Result

- clustering_result.mat
- figure_1.png
- figure_2.png
- figure_3.png
- Output
- pcnmf_model.mat

November 14, 2018 Created capsule

Reproducibility

<https://codeocean.com/capsule/7292152/tree/v1>



Publishing Options



The Journal of Open Source Software

Submit Papers

The Journal of Open Source Software

A developer

Learn more

Accepted papers

District

Yellowbrick diagnostic tool



About this Journal

The *Journal of Open Research Software* (JORS) is a peer-reviewed journal publishing research software with high reuse value. JORS provides a platform for specialist and institutional repositories to describe their software assets. JORS ensures that software is professionally archived, preserved, and is cited. The journal also ensures that software and the papers will be citable, and that the software can be reused.

JORS also publishes full-length research papers describing the development and maintenance of open source research software. The journal promotes the dissemination of best practices for the development, maintenance, and reuse of sustainable research software.

DRYAD About ▾ For researchers ▾ For organizations ▾ C

Overview Governance Our community Funding Annual reports

The organization: Overview

The Dryad Digital Repository is a **curated** resource that makes the data underlying scientific publications **discoverable**, **freely reusable**, and **citable**. Dryad provides a **general-purpose** home for a wide diversity of datatypes.

Dryad's **vision** is to promote a world where research data is openly available, integrated with the scholarly literature, and routinely re-used to create knowledge.

Our **mission** is to provide the infrastructure for, and promote the re-use of, data underlying the scholarly literature.

arXiv.org

Login

Search or Article ID All fields

(Help | Advanced search)

Open access to 1,516,136 e-prints in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Submissions to arXiv should conform to Cornell University academic standards. arXiv is owned and operated by Cornell University, a private not-for-profit educational institution. arXiv is funded by Cornell University, the Simons Foundation and by the member institutions.

Subject search and browse: Physics Search Form Interface Catchup

14 Jan 2019: The annual update from the arXiv team is now available
5 Sept 2018: arXiv looks to the future with move to Cornell CIS
See cumulative "What's New" pages. Read robots beware before attempting any automated download

Summary



- Focus on “How I am going to make this available”
- Consider who you want to be able to access this work
 - Just you
 - Internal project team
 - Public
- Think about what you are allowed to share directly
 - Pointers in documentation for things you can’t

Exercise: Applying the Concepts to Your Project

Dissemination





Wrap-Up



Building on Pragmatism



Focus today was on small steps to get started

Many guides to explore with expanded definitions of reproducible research

- Bakken S. The journey to transparency, reproducibility, and replicability. JAMIA, 26(3) 185-187. <https://doi.org/10.1093/jamia/ocz007>
- McIntosh LD, et al. Repeat: a framework to assess empirical reproducibility in biomedical research. BMC Medical Research Methodology. 2017;17:143. <https://doi.org/10.1186/s12874-017-0377-6>
https://github.com/ripeta/Research_Reproducibility

Continue the Momentum



“The most important tool is the *mindset*, when starting, that the end product will be reproducible.”

– Keith Baggerly



AMIA Workshop Attendee Surveys

AMIA (and ultimately you) benefits from hearing what you thought of this workshop. If you loved it, hated it, or are even just 'meh', please take a few minutes to fill out the survey when you get it.

Thank you!

Luke Rasmussen
luke.rasmussen@northwestern.edu

Eric Whitley
ewhitley@northwestern.edu

Resources



- Karl Broman – Initial Steps Toward Reproducible Research (<https://kbroman.org/steps2rr/>)
- Victoria Stodden – Enabling Reproducible Research: Open Licensing for Scientific Innovation (March 3, 2009). International Journal of Communications Law and Policy, Forthcoming. Available at SSRN: <https://ssrn.com/abstract=1362040>
- Carl Boettiger - An introduction to Docker for reproducible research, with examples from the R environment. (2015) ACM SIGOPS Operating Systems Review, Special Issue on Repeatability and Sharing of Experimental Artifacts. 49(1), 71-79. DOI: 10.1145/2723872.2723882. arXiv:1410.0846 [cs.SE]
- Jeffrey Leek & Richard Peng - Opinion: Reproducible research can still be wrong: Adopting a prevention approach. PNAS February 10, 2015 112 (6) 1645-1646; <https://doi.org/10.1073/pnas.1421412111>
- Suzanne Bakken - The journey to transparency, reproducibility, and replicability. JAMIA, 26(3) 185-187. <https://doi.org/10.1093/jamia/ocz007>
- Leslie McIntosh et al., Repeat: a framework to assess empirical reproducibility in biomedical research. BMC Medical Research Methodology. 2017 17:143. <https://doi.org/10.1186/s12874-017-0377-6>
- Enrico Coiera et al. Does health informatics have a replication crisis? Journal of the American Medical Informatics Association, Volume 25, Issue 8, 1 August 2018, Pages 963–968, <https://doi.org/10.1093/jamia/ocy028>

Resources



- Anaconda (<https://www.anaconda.com/distribution/>)
- Code Ocean (<https://codeocean.com/>)
- Docker (<https://www.docker.com/>)
- Dryad (<https://datadryad.org/>)
- Jupyter Notebook (<https://jupyter.org>)
- NLM UMLS (<https://uts.nlm.nih.gov>)
- Natural Language Toolkit (<https://www.nltk.org/>)
- Python (<https://www.python.org>)
- R (<https://www.r-project.org>)
- R Markdown (<https://rmarkdown.rstudio.com>)
- StatTag (<http://stattag.org>)