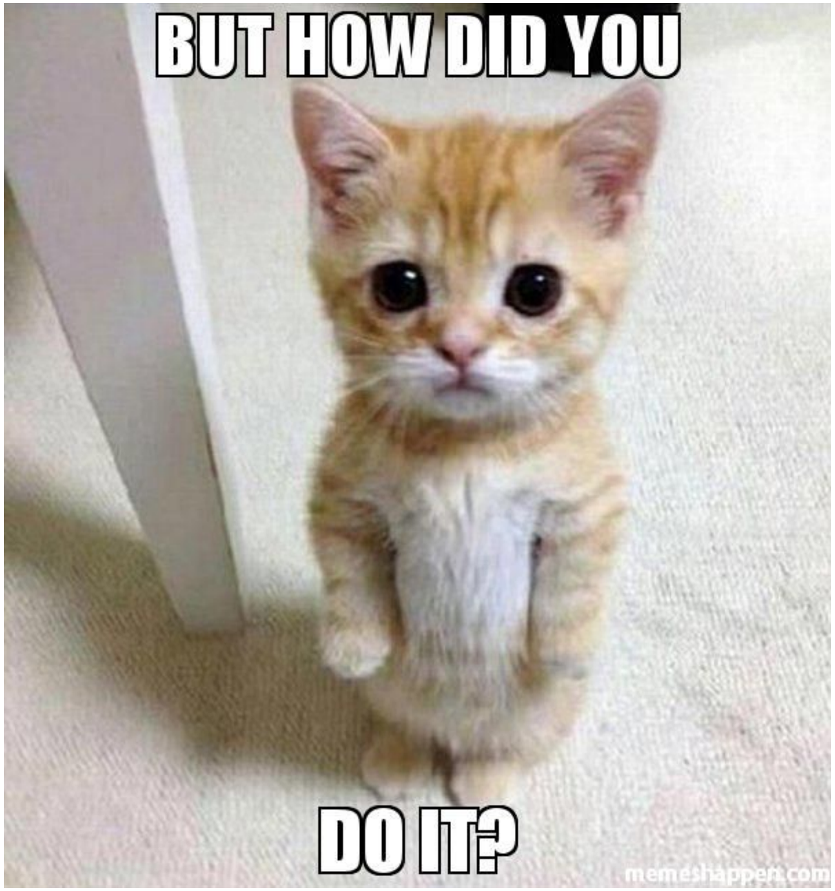
Reproducible Research

Lecture 9
BIOS 6660, Spring 2019
Instructor: Pam Russell



http://memeshappen.com/meme/kitten-timesheet/but-how-did-you-do-it-49245/9

Intro to reproducibility

What do we mean by “reproducibility”

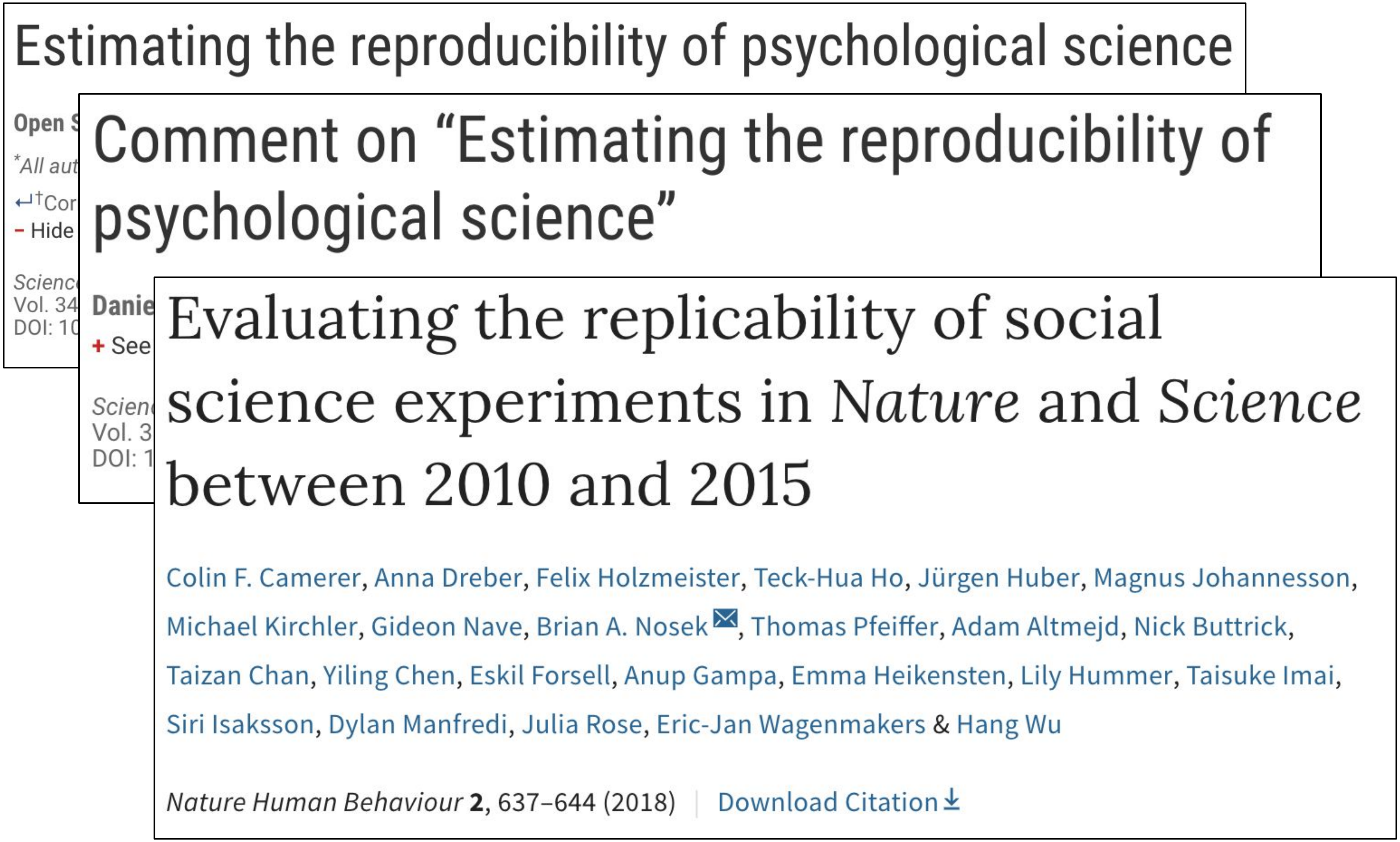
Analysis in a paper can be repeated by independent analyst with same data and methods to obtain same results

Reproducibility vs. replicability

Study replication	Reproducible research
Independent investigators attempt to repeat a study	Ability to repeat the analysis in a paper with the original data and same methods
The ultimate standard, but often difficult or impossible	An attainable minimum standard for assessing the value of scientific claims

http://science.sciencemag.org/content/334/6060/1226

“Replication crisis” in psychology



Estimating the reproducibility of psychological science

Open S

*All aut

↵†Cor

– Hide

Scienc
Vol. 34
DOI: 10

Comment on “Estimating the reproducibility of psychological science”

Danie

+ See

Scien
Vol. 3
DOI: 1

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek ✉, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

Nature Human Behaviour **2**, 637–644 (2018) | Download Citation ⤓

Why is reproducibility important?

> “Reproducibility is important not because it ensures that the results are correct, but rather because it ensures transparency and gives us confidence in understanding exactly what was done.”
>
> - Roger Peng

https://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/

Why is reproducibility difficult?

- Huge datasets
- Complex algorithms
- Complex pipelines
- Software environments

http://www.quickmeme.com/Complicated-situation

The Duke breast cancer saga

2006

Nature Medicine publishes Potti, Nevins et al. showing gene expression arrays can be used to predict treatment responses

Baggerly and Coombes at MD Anderson attempt to reproduce the results (to use the technology). Can't reproduce results, but can do it by introducing specific errors

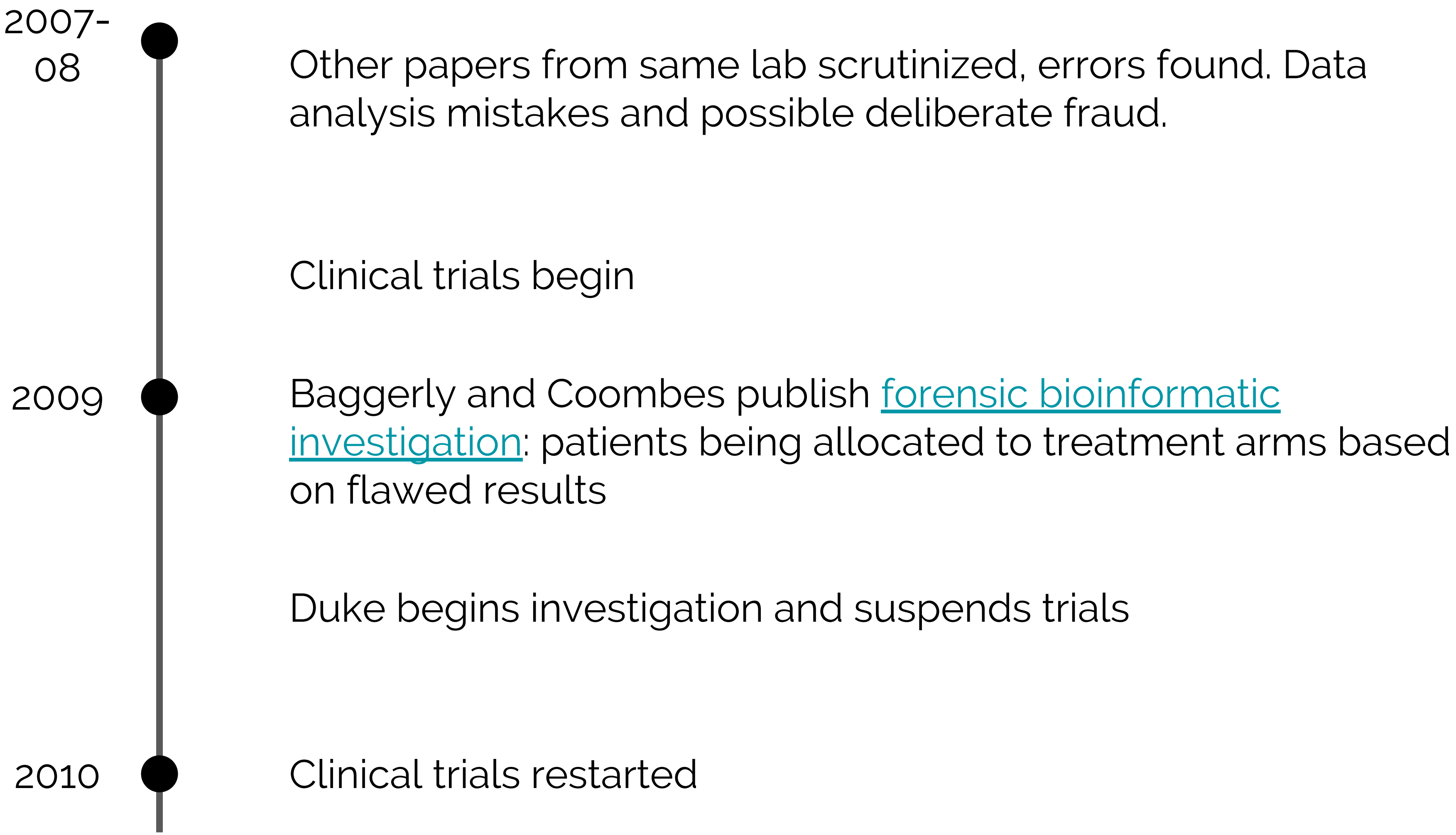
2007

Baggerly and Coombes go back and forth with Potti and Nevins, who continue to insist it works

Baggerly and Coombes publish "Microarrays: retracing steps" in *Nature Medicine*

https://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/StarterSet/baggerly_nebraska12.pdf
https://simplystatistics.org/2018/04/23/what-can-we-learn-from-data-analysis-failures/

The Duke breast cancer saga



2007-08

Other papers from same lab scrutinized, errors found. Data analysis mistakes and possible deliberate fraud.

Clinical trials begin

2009

Baggerly and Coombes publish forensic bioinformatic investigation: patients being allocated to treatment arms based on flawed results

Duke begins investigation and suspends trials

2010

Clinical trials restarted

The Duke breast cancer saga

2010 Report that Potti lied about being a Rhodes scholar on an application for federal funding

Clinical trials terminated, Potti resigns

2011 Patient lawsuits filed

2012 60 Minutes episode

2015 *The Cancer Letter* publishes a memo by a former trainee exposing shoddy work he recognized at the time; memo was originally ignored

The Duke saga: reproducibility

> “We spent approximately 1500 person-hours on this issue, mostly because we could not tell what data were used or how they were processed. Transparently available data and code would have made checking results and their validity far easier. Because transparency was absent, an understanding of the problems was delayed, trials were started on the basis of faulty data and conclusions, and patients were endangered.”
>
> - Baggerly & Coombes

http://clinchem.aaccjnls.org/content/57/5/688.long

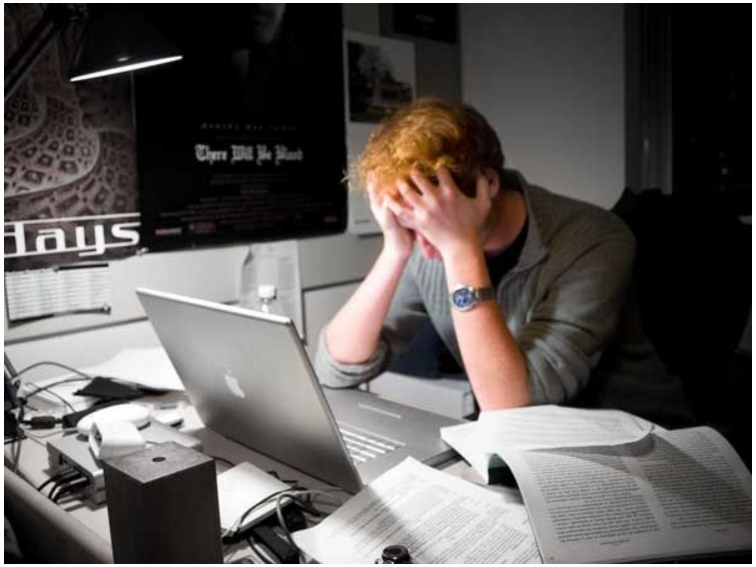
Roger Peng: reproducibility would not have prevented the problem

> “Yes, genomic analyses are ‘hard to do’ but clearly there was expertise in the lab to recognize that difficulty and to recognize when statistical methods were being incorrectly applied… The problem was a breakdown in communication and a total lack of trust between investigators and members of the data analytic team.”

https://simplystatistics.org/2018/04/23/what-can-we-learn-from-data-analysis-failures/

Limits of reproducibility

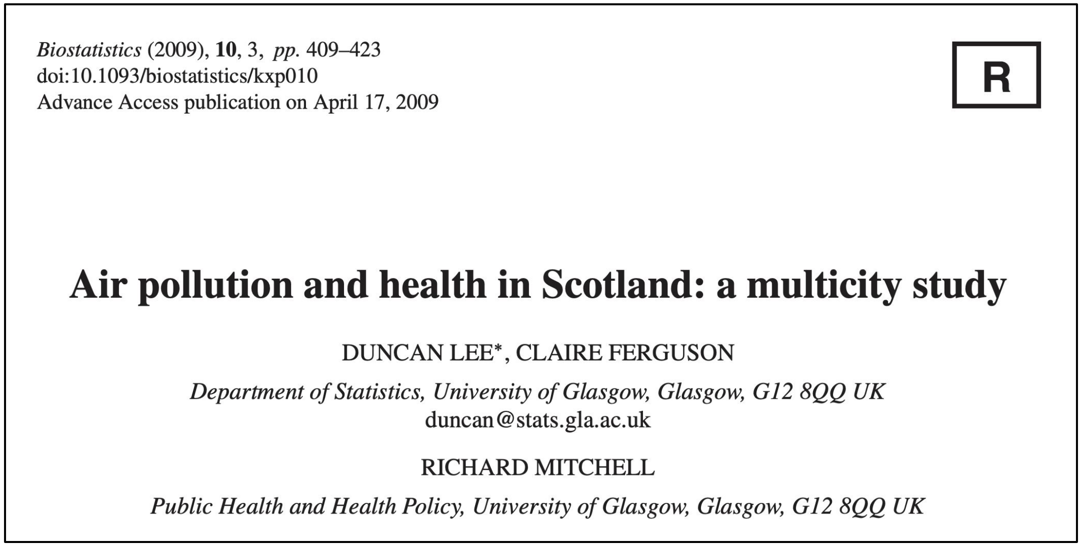
- Claims of study can still be wrong
- Still challenging for readers to put pieces together



Journal policies: *Biostatistics*

Badge on article PDF

- “D”: data provided
- “C”: code provided
- “R”: results were reproduced during review, implies D and C



Biostatistics (2009), **10**, 3, *pp.* 409–423
doi:10.1093/biostatistics/kxp010
Advance Access publication on April 17, 2009

Air pollution and health in Scotland: a multicity study

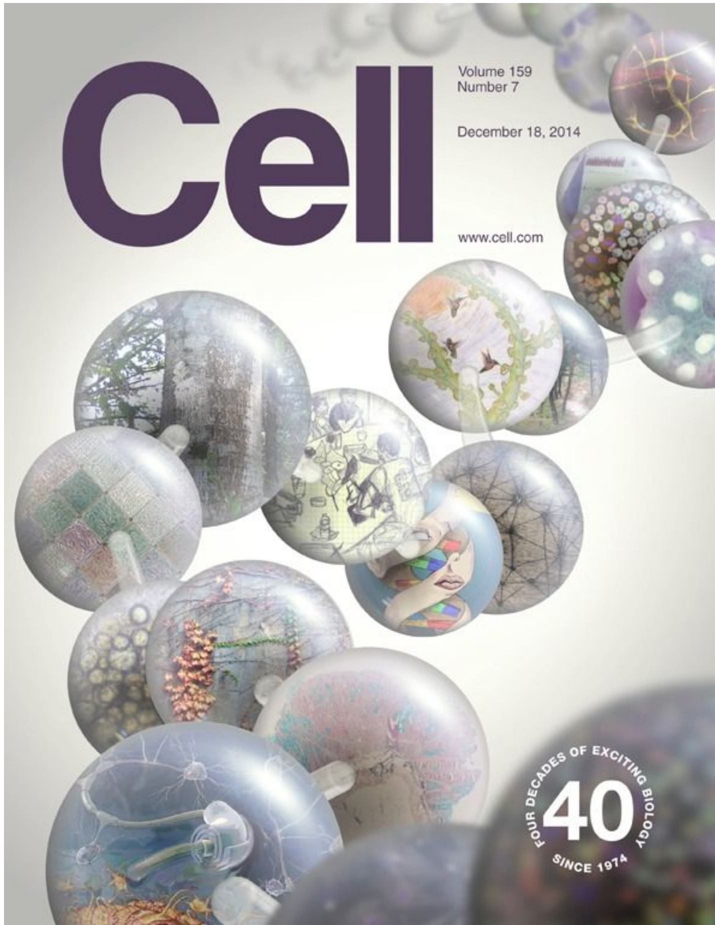
DUNCAN LEE*, CLAIRE FERGUSON
Department of Statistics, University of Glasgow, Glasgow, G12 8QQ UK
duncan@stats.gla.ac.uk

RICHARD MITCHELL
Public Health and Health Policy, University of Glasgow, Glasgow, G12 8QQ UK

https://academic.oup.com/biostatistics/article/10/3/405/293660

Journal policies: *Cell*

- Sharing policies emphasize data and experimental methods
- “Software and data resources should be reported by providing a short description of the software or custom script/data resource and the URL to obtain them unless it is provided as a supplemental file.”
    - Seemingly not enforced

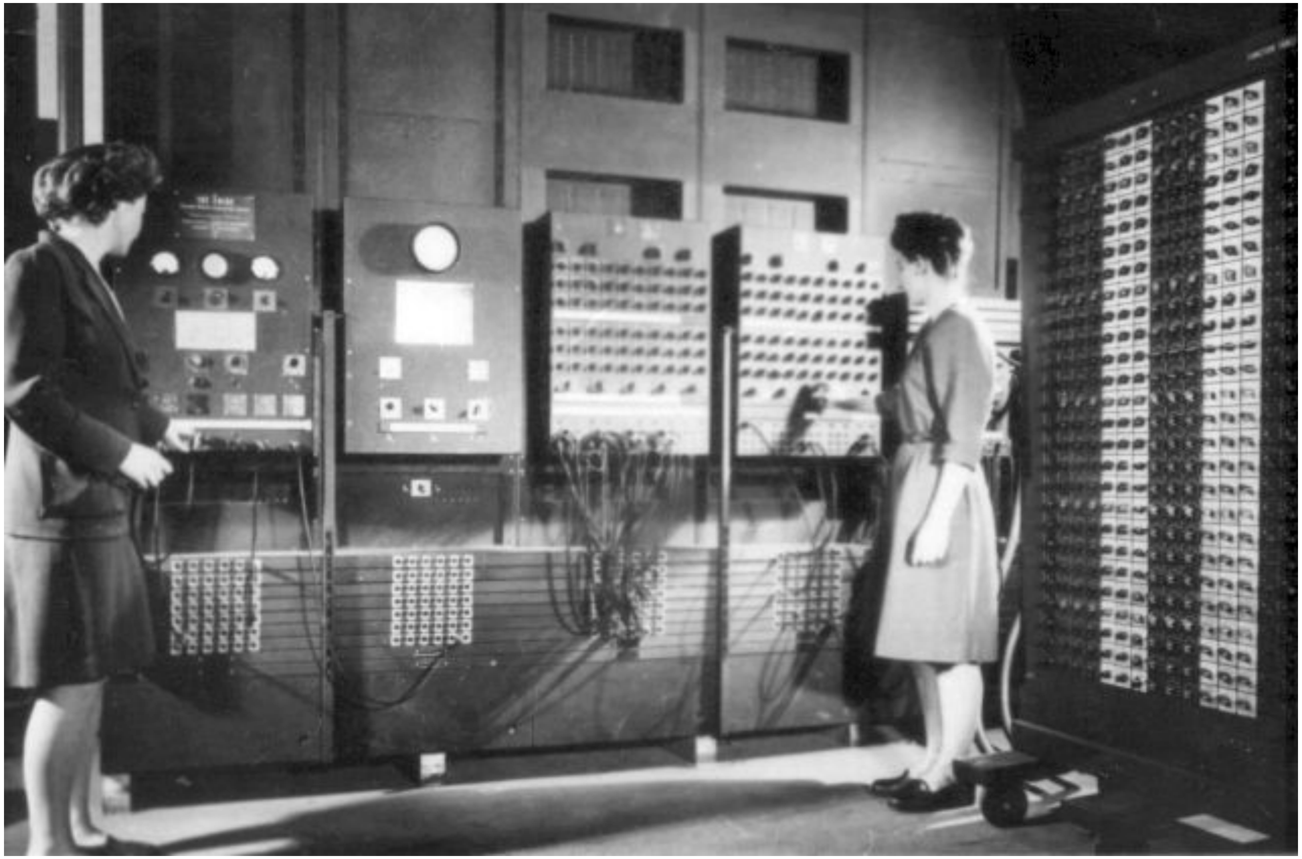


Basic approach to reproducibility

The key to reproducibility

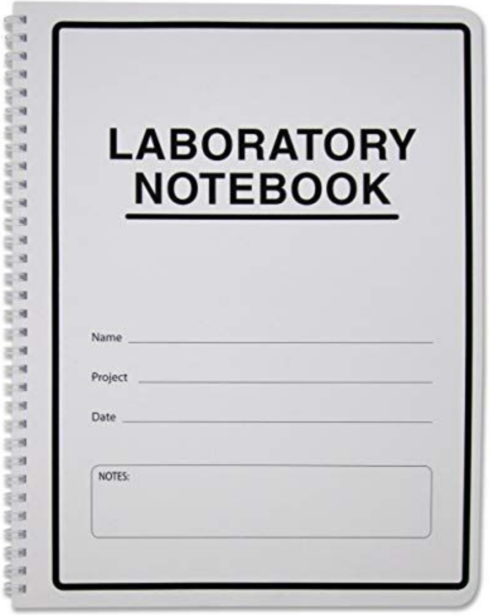
https://medium.com/@abdulazeezidris28/write-it-down-d2b9dbda0b65

How we write down instructions



https://en.wikipedia.org/wiki/ENIAC

MIT Mechanical Engineering Dept. (2007)



"Your laboratory notebook is a permanent record of what you did and what you observed in the laboratory… Your notebook should be like a diary, recording what you do, and why you did it… A good test of your work is the following question: could someone else, with an equivalent technical background to your own, use your notebook to repeat your work, and obtain the same results? For that matter, could you come back six months later, read your notes, and make sense of them? If you can answer yes to these two questions, you are keeping a good notebook."

http://web.mit.edu/me-ugoffice/communication/labnotebooks.pdf

Ideal form of a reproducible workflow

Can run from start to finish using the raw data only.

Test: can you delete everything except raw data and run the entire workflow?

Script everything

Automate your “polished” workflow:

- Data import
- Data processing
- Analysis
- Products e.g. figures

But also:

- Exploratory analysis
- Tests
- Data download if possible

Don’t do anything by hand.
No pointing and clicking!

Version control

Treat like lab notebook:

- Exploratory analysis
- Dead ends
- All versions of “actual” analysis
- Small outputs
- Documentation

Just be careful when making repositories public! All previous versions are visible.

Share all code

- Keep all code for project in a single repo
- Don’t let scripts creep into other directories
- Document as you go along

Why would anyone not share all code?

Usually not deliberately hiding code.

- Didn’t adhere to an organizational structure
- Scripts strewn around file system
- Crunch time before paper submission: standards go out the window

Making your environment reproducible

Software environment

Reproducibility depends not only on data and code but also on the computational environment.

Most frustrating: software dependencies and versions

Smoothing the process

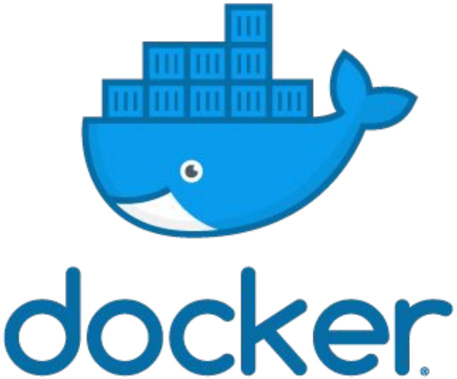
Capturing and sharing the environment

Containers: package up code and all dependencies

Runs consistently anywhere

BioContainers

https://biocontainers.pro/



R tools for reproducibility

R Markdown and knitr

“Literate programming”:

- Presenting a program for a human reader
- Code follows the program logic
- Human readable explanations are interspersed

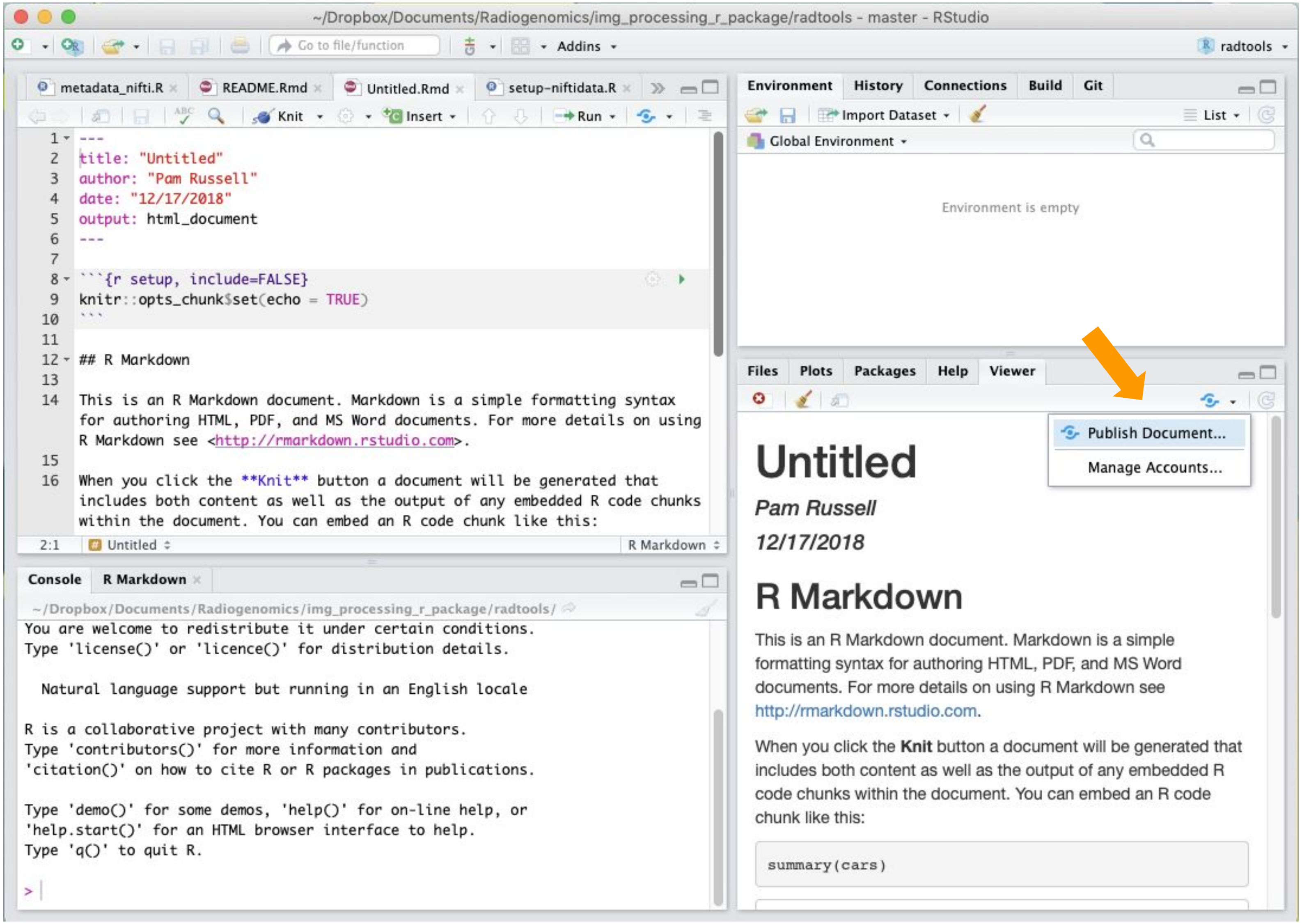
You would have written the code anyway; knitr makes it easy to make it reproducible

Rpubs

Write R Markdown documents in RStudio

Share on rpubs.com for free

Rpubs



~~~
---
title: "Untitled"
author: "Pam Russell"
date: "12/17/2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax
for authoring HTML, PDF, and MS Word documents. For more details on using
R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that
includes both content as well as the output of any embedded R code chunks
within the document. You can embed an R code chunk like this:
~~~

```
~/Dropbox/Documents/Radiogenomics/img_processing_r_package/radtools/
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Untitled

Pam Russell

12/17/2018

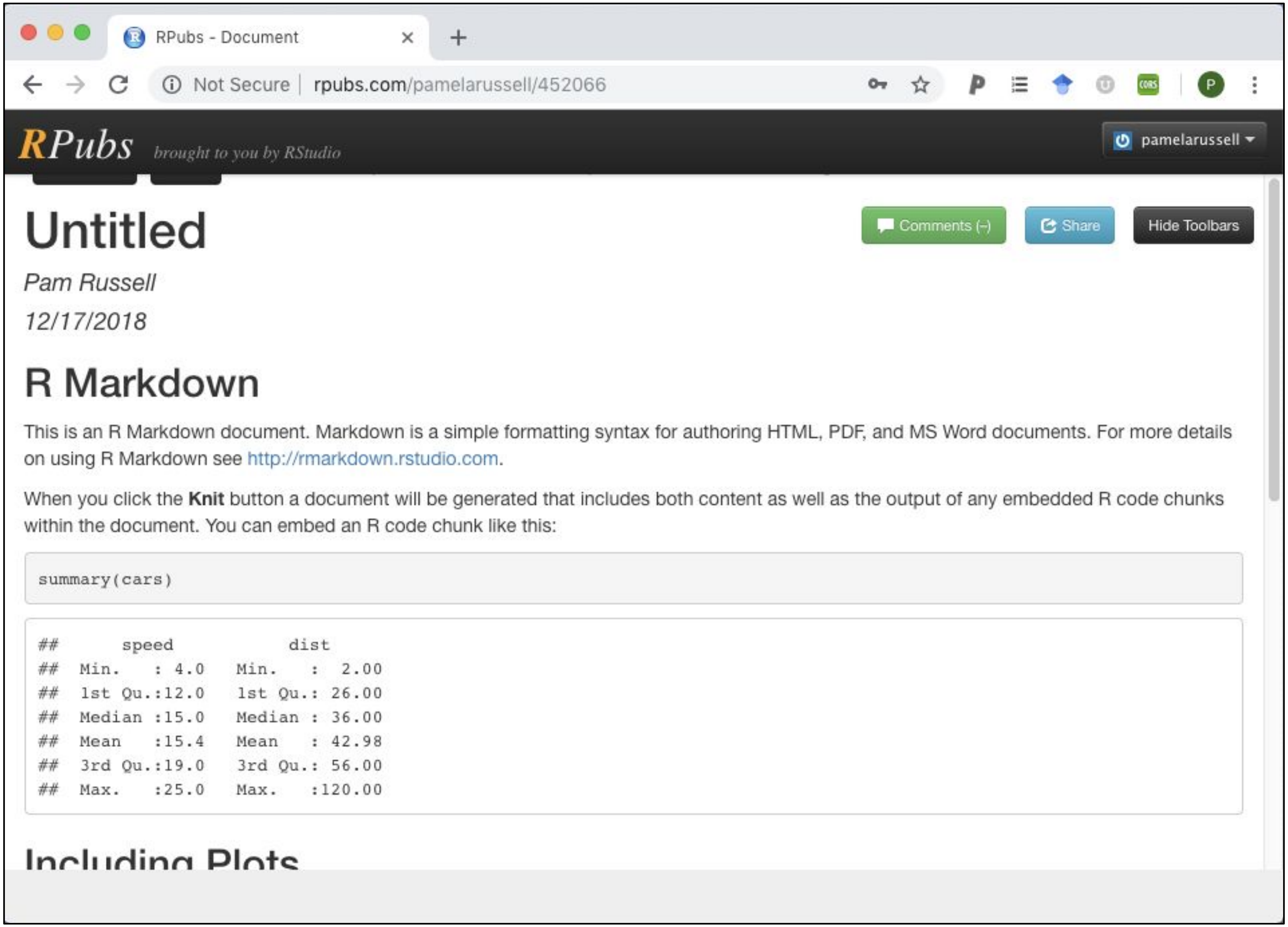
R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

Rpubs


RPubs - Document
Not Secure | rpubs.com/pamelarussell/452066
RPubs brought to you by RStudio
pamelarussell
Comments (-)
Share
Hide Toolbars

Untitled

Pam Russell

12/17/2018

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

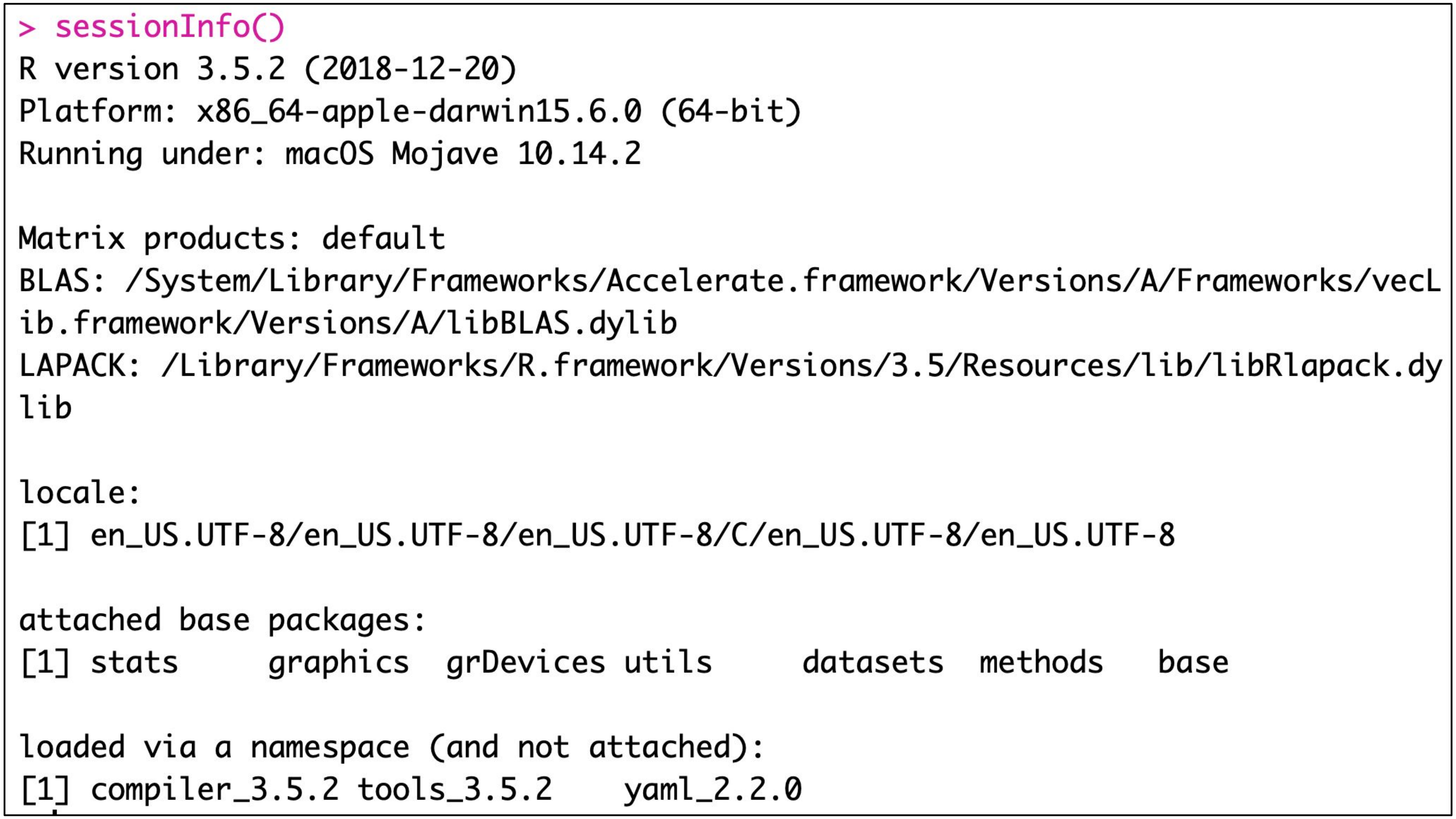
```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

sessionInfo()

Prints R version, system info, attached packages and versions



```
> sessionInfo()
R version 3.5.2 (2018-12-20)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Mojave 10.14.2

Matrix products: default
BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecL
ib.framework/Versions/A/libBLAS.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dy
lib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] compiler_3.5.2 tools_3.5.2    yaml_2.2.0
```

Genomic workflows

Genomic analysis workflows

Data for multiple samples

Data processing

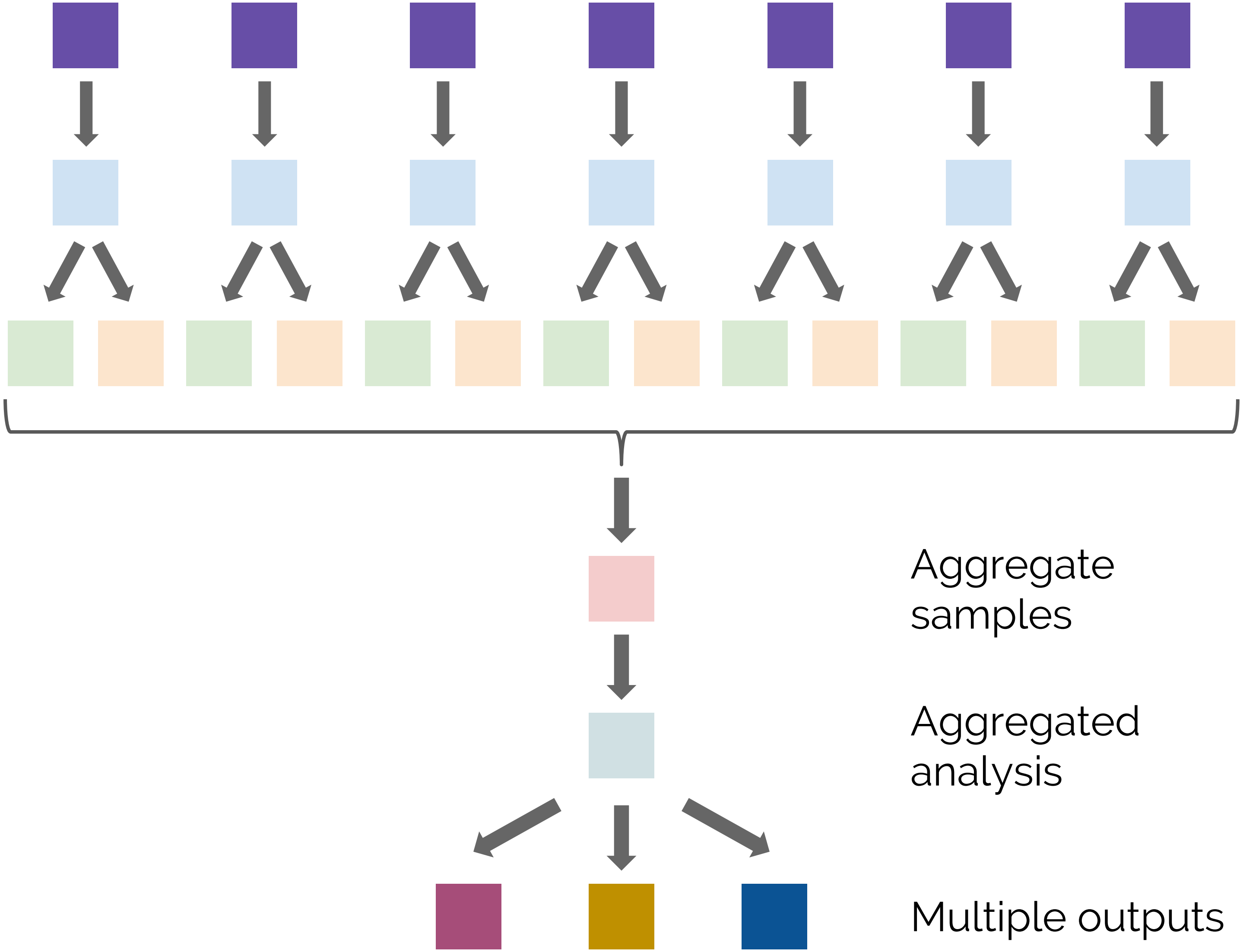
Analysis with one or more tools / packages

Presentation of results

What does a workflow actually look like?

Simple workflow	Moderate workflow	Complex workflow
• A few data files • A few steps that can all be done in R	• A few data files or many, but fairly uniform • More complex steps but can still be done in R	• Many data files • Multiple tools • Some steps run on each sample, some in aggregate
Single R script	Multiple R files with a master script	Bash script or workflow management tool

Shape of a typical workflow



Raw data for each sample

Processed data

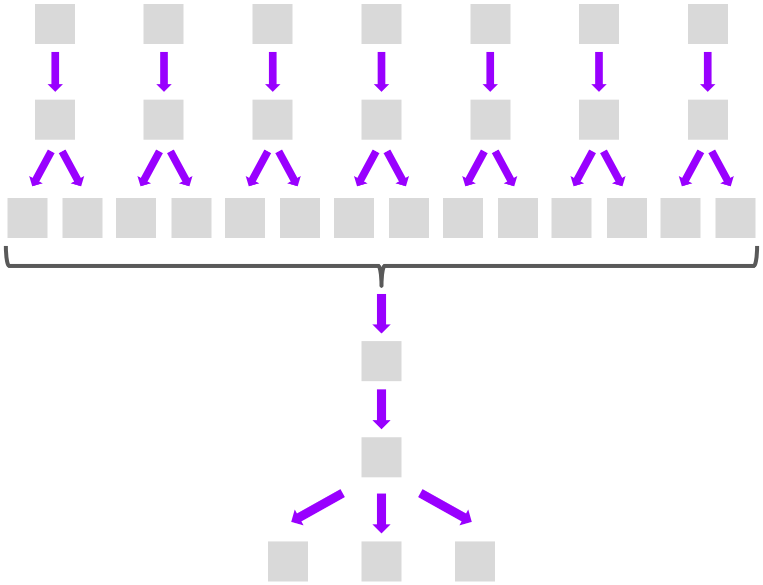
Per-sample analysis

Aggregate samples

Aggregated analysis

Multiple outputs

Workflow management tools



You specify the files and how they relate to each other through analysis steps

Workflow manager figures out which steps to run and in what order; runs and manages the jobs for you

You publish the workflow specification

A great one: [Snakemake]

Galaxy

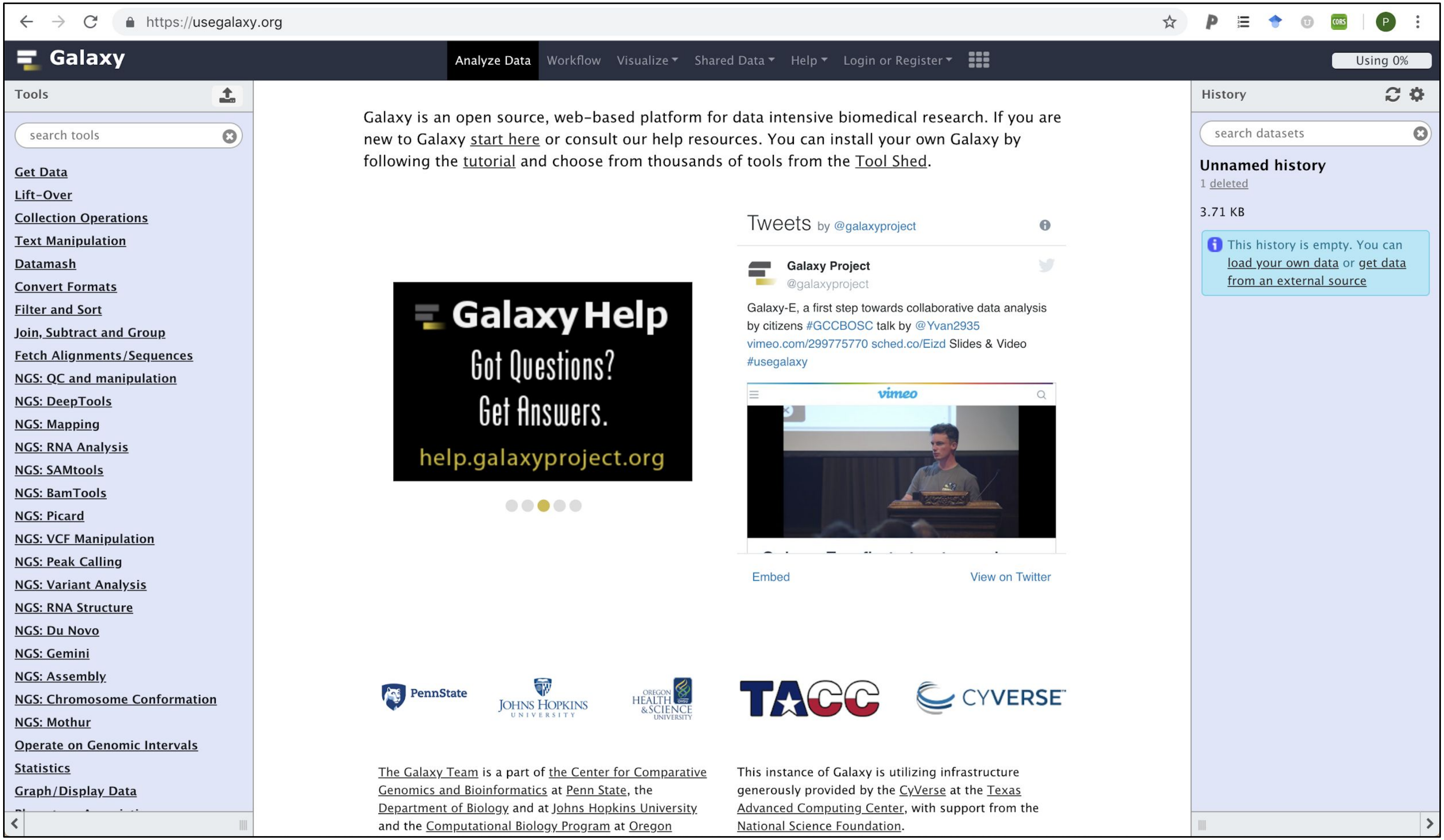
Web platform with thousands of publicly available tools

Reproducible bioinformatic analysis without writing code

Don’t need access to a Linux server

https://usegalaxy.org/

Galaxy



Get Data
Lift-Over
Collection Operations
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
NGS: RNA Analysis
NGS: SAMtools
NGS: BamTools
NGS: Picard
NGS: VCF Manipulation
NGS: Peak Calling
NGS: Variant Analysis
NGS: RNA Structure
NGS: Du Novo
NGS: Gemini
NGS: Assembly
NGS: Chromosome Conformation
NGS: Mothur
Operate on Genomic Intervals
Statistics
Graph/Display Data

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Tweets by @galaxyproject

Galaxy Project
@galaxyproject

Galaxy-E, a first step towards collaborative data analysis by citizens #GCCBOSC talk by @Yvan2935 vimeo.com/299775770 sched.co/Eizd Slides & Video #usegalaxy

Unnamed history
1 deleted

3.71 KB

This history is empty. You can load your own data or get data from an external source

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology and at Johns Hopkins University and the Computational Biology Program at Oregon

This instance of Galaxy is utilizing infrastructure generously provided by the CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

Steps for a reproducible analysis

Developing a basic reproducible analysis

Today: steps of reproducible analysis

Thursday: live demo of small complete analysis

Homework 5: another complete reproducible analysis

Reproducibility checklist

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

RULE #2—AVOID MANUAL DATA MANIPULATION STEPS

RULE #3—ARCHIVE THE EXACT VERSIONS OF ALL EXTERNAL PROGRAMS USED

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

RULE #5—RECORD ALL INTERMEDIATE RESULTS, WHEN POSSIBLE IN STANDARDIZED FORMATS

RULE #6—FOR ANALYSES THAT INCLUDE RANDOMNESS, NOTE UNDERLYING RANDOM SEEDS

RULE #7—ALWAYS STORE RAW DATA BEHIND PLOTS

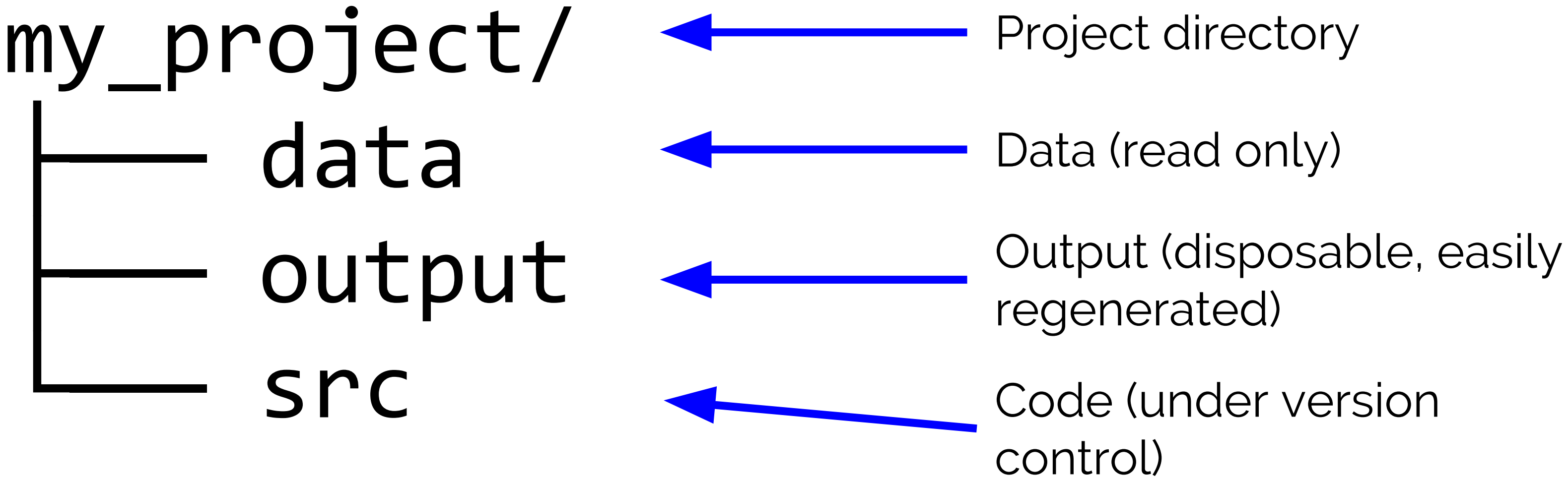
RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

RULE #9—CONNECT TEXTUAL STATEMENTS TO UNDERLYING RESULTS

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

https://dataconomy.com/2017/07/10-rules-results-data-science/

Organization



my_project/

data

output

src

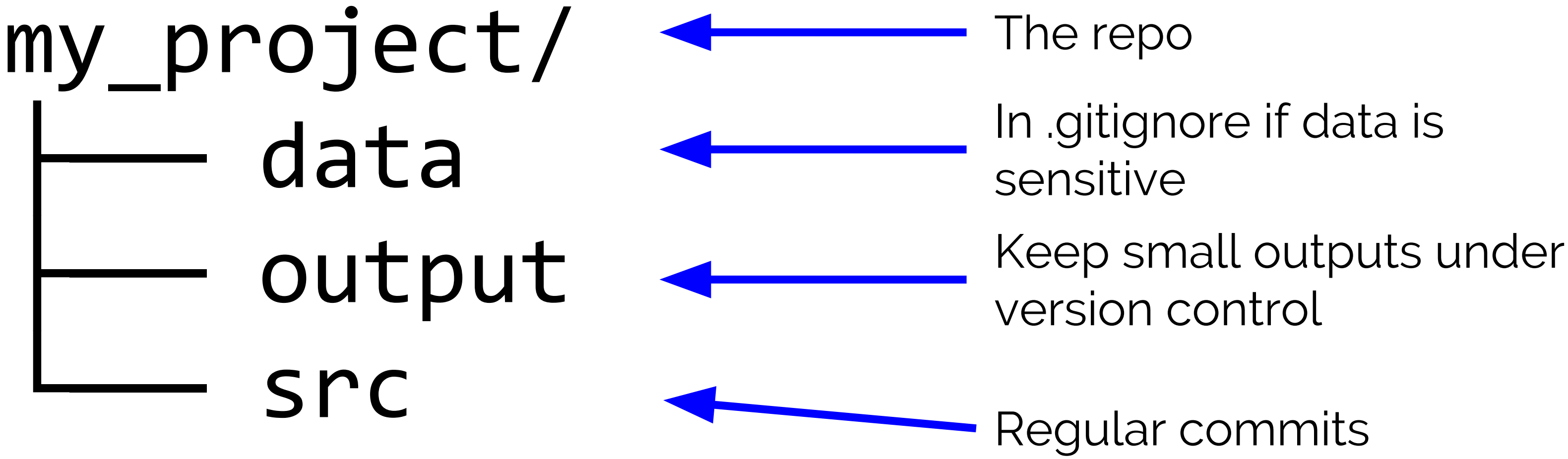
Project directory

Data (read only)

Output (disposable, easily regenerated)

Code (under version control)

Version control



```
my_project/
├── data
├── output
└── src
```

The repo

In .gitignore if data is sensitive

Keep small outputs under version control

Regular commits

New data from project

- Full data management and sharing practices from last week

Public data

- From data repository: document DOI
- From public database: document version
- Paper supplemental data: document paper
- Small dataset with open license: can go on GitHub

- Record a digital fingerprint (more on this later)
- Put in **`data`** directory with documentation in **`README.txt`**
- Remove write permissions from file(s)

R Markdown

We use R Markdown for complete analyses on Thursday and on Homework 5

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

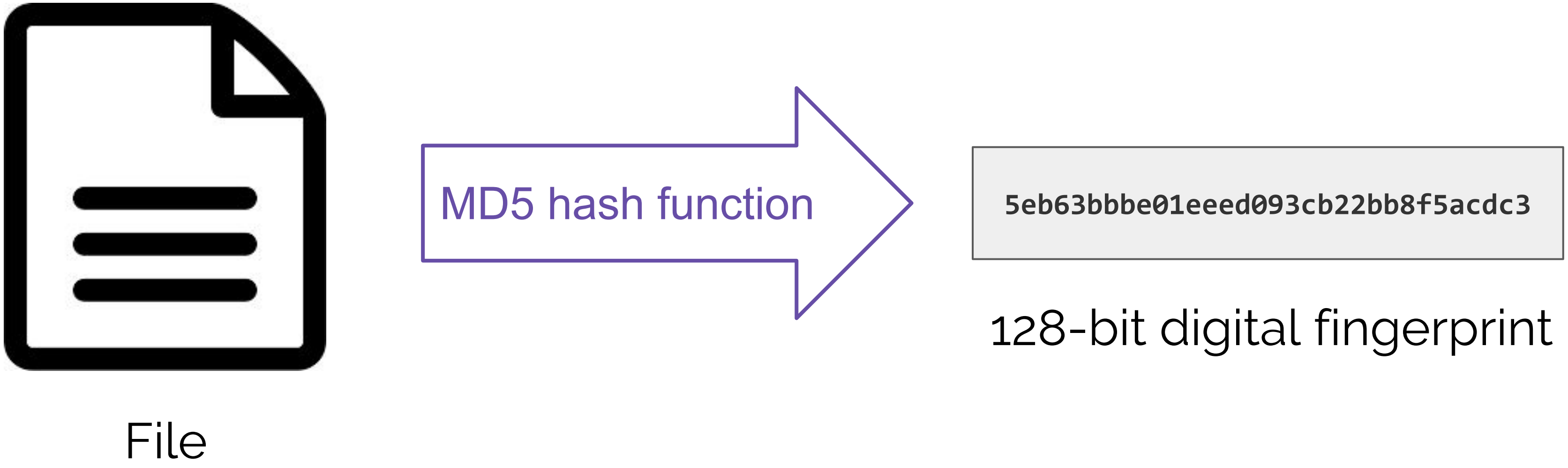
RULE #7—ALWAYS STORE RAW DATA BEHIND PLOTS

RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

RULE #9—CONNECT TEXTUAL STATEMENTS TO UNDERLYING RESULTS

Data fingerprint

Capture a digital fingerprint of the data so future users can verify their copy of the data



File

MD5 hash function

`5eb63bbbe01eeed093cb22bb8f5acdc3`

128-bit digital fingerprint

Work from raw data

Workflow starts with loading raw data

Should always be able to delete any intermediate data and run entire workflow from raw data

RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

RULE #2—AVOID MANUAL DATA MANIPULATION STEPS

Exploratory analysis

Exploratory plots

Helps make decisions about future analysis

Keep under version control

Mostly for you to come back to

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

Main analysis

Funding agency and journal requirements:
Mostly stop at theoretical reproducibility.
No requirement of practical reproducibility.

Make a good faith effort toward practical reproducibility.
Put yourself in the user’s shoes!

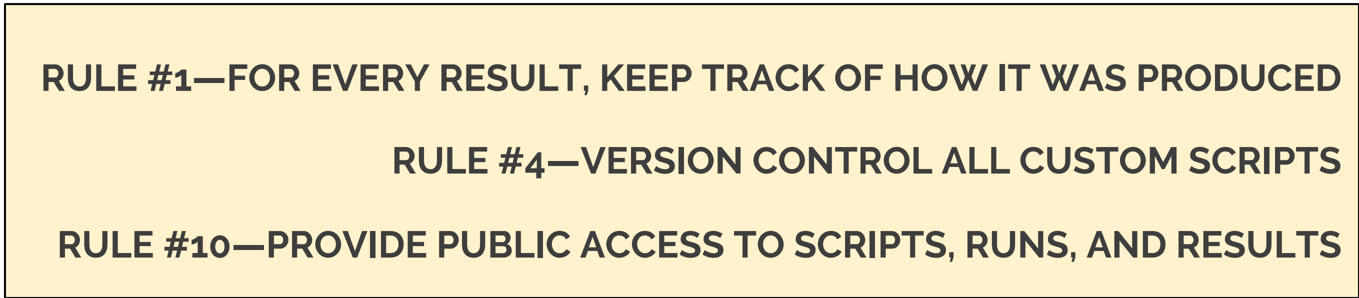
Main analysis

Basic requirements

- Keep code under version control
- Share repo publicly

“Good faith” requirements:

- Documentation in GitHub README
    - Repo contents
    - Mapping between paper results and scripts
    - How data is imported and moves through pipeline
- Code comments to guide new users



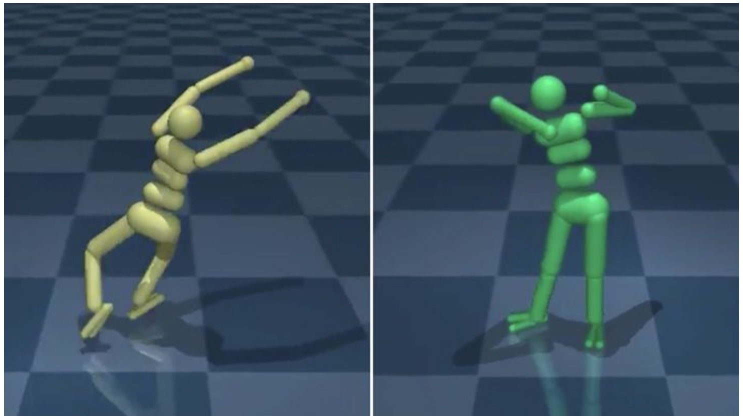
RULE #1—FOR EVERY RESULT, KEEP TRACK OF HOW IT WAS PRODUCED

RULE #4—VERSION CONTROL ALL CUSTOM SCRIPTS

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

Analyses with randomness

- Any analysis with randomness: machine learning, simulations, ...
- Provide pseudo-random number generator with an initial value
- Subsequent runs will get same sequence of “random” numbers
- In R: **`set.seed()`**

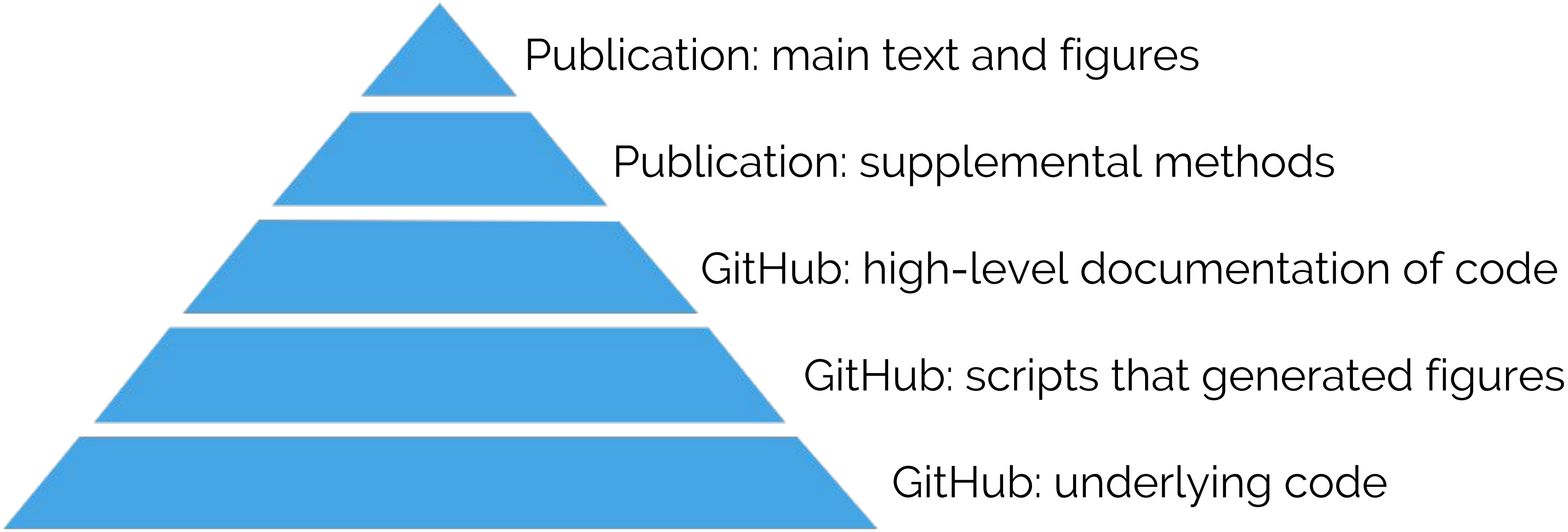


Algorithm learning to walk differently with different initial conditions

https://doi.org/10.1126/science.aat3298

RULE #6—FOR ANALYSES THAT INCLUDE RANDOMNESS, NOTE UNDERLYING RANDOM SEEDS

Output



Publication: main text and figures

Publication: supplemental methods

GitHub: high-level documentation of code

GitHub: scripts that generated figures

GitHub: underlying code

RULE #8—GENERATE HIERARCHICAL ANALYSIS OUTPUT, ALLOWING LAYERS OF INCREASING DETAIL TO BE INSPECTED

RULE #10—PROVIDE PUBLIC ACCESS TO SCRIPTS, RUNS, AND RESULTS

Software environment

- On Linux
  - Minimum requirement: record all program versions and system information
  - Better: use a container
- In R: **`sessionInfo()`**

RULE #3—ARCHIVE THE EXACT VERSIONS OF ALL EXTERNAL PROGRAMS USED

RULE #5—RECORD ALL INTERMEDIATE RESULTS, WHEN POSSIBLE IN STANDARDIZED FORMATS

- Necessary when:
    - Limited resources to run from scratch
    - Need to support users with limited ability to run some of the tools
- For relatively small projects: better to ensure everything can be run from scratch