Towards more reproducible + open research in building science:

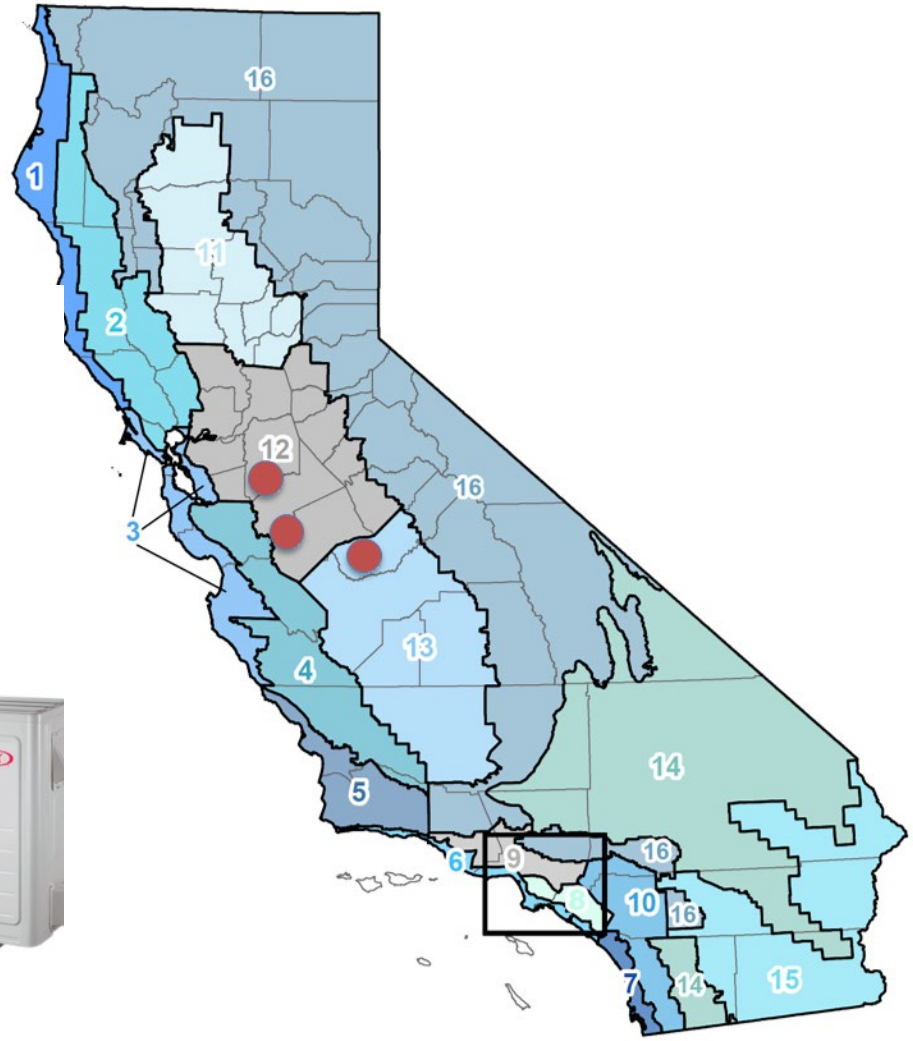
why, what, and some tools for how

Dana Miller
September 27, 2018
Arch 241

Current research – field study

In existing buildings,
can high-efficiency ceiling fans
with automatic controls
reduce energy consumption
while maintaining comfort?

+



software carpentry

ASHRAE

B E R C Berkeley Energy & Resources Collaborative

pyladies

R Ladies

AIA San Francisco

Goals: after this class, able to answer:

- Why is reproducible science important?
- What does reproducible mean anyway?
- Why is this hard?
- What tools can we start using now?

- Hands-on practice
  - How to create a project in RStudio?
  - (If time) What is R Markdown?

The Guardian

US edition

Attempt to replicate major social scientific findings of past decade fails

Scientists and the design of experiments under scrutiny after a major project fails to reproduce results of high profile studies

https://www.theguardian.com/science/2018/aug/27/attempt-to-replicate-major-social-scientific-findings-of-past-decade-fails

UNITED STATES FEDERAL RESERVE SYSTEM

Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"

https://www.federalreserve.gov/econres/data/feds/2015/files/2015083pap.pdf

Mother Jones

This Cornell Food Researcher Has Had 13 Papers Retracted. How Were They Published in the First Place?

The social sciences have a problem.

KIERA BUTLER SEPTEMBER 25, 2018 6:00 AM

https://www.motherjones.com/food/2018/09/cornell-food-researcher-brian-wansink-13-papers-retracted-how-were-they-published/

https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking#.qs6A0K3DW

Hi Ozge,

Glad you had a chance to take an initial look at the data.

I don't think I've ever done an interesting study where the data "came out" the first time I looked at it. The interesting stories come from seeing when things -- like the 1/2 price buffet -- works and when it doesn't.

I would like you to really dig into this to find a number of situations or people for which this relationship does hold -- that is where the 1/2 price buffet did result in a difference.

Here's some things to do.

Males

Females

Lunch goers

Dinner goers

People sitting alone

People eating with groups of 2

People eating in groups of 2+

People who order alcohol

People who order soft drinks

People who sit close to buffet

People who sit far away

and so on . . .

Second, think of all the different ways you can cut the data and analyze subsets of it to see when this relationship holds. For instance, if it works on men but not women, we have a moderator. Here are some groups you'll want to break out separately:

The Washington Post

Wonkblog

Is the evidence for austerity based on an Excel spreadsheet error?

Thomas Herdon, PhD student
UMass Amherst

"Growth in a time of debt"
Reinhart and Rogoff 2011

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5		7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

https://www.washingtonpost.com/news/wonk/wp/2013/04/16/is-the-best-evidence-for-austerity-based-on-an-excel-spreadsheet-error
https://www.bbc.com/news/magazine-22223190
http://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646

News

Attempt to replicate major social scientific findings of past decade fails

https://www.theguardian.com/science/2018/aug/27/attempt-to-replicate-major-social-scientific-findings-of-past-decade-fails

Publication

Letter | Published: 27 August 2018

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

https://www.nature.com/articles/s41562-018-0399-z

The actual data

"The replications follow analysis plans **reviewed by the original authors** and pre-registered prior to the replications. The replications are high powered, with **sample sizes on average about five times higher** than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size."

News

Attempt to replicate major social scientific findings of past decade fails

https://www.theguardian.com/science/2018/aug/27/attempt-to-replicate-major-social-scientific-findings-of-past-decade-fails

Publication

Letter | Published: 27 August 2018

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

https://www.nature.com/articles/s41562-018-0399-z

"I don't think it's a crisis, it's a reformation.
We're in the midst of a dramatic increase in the rigour and transparency of research in the social sciences."

- Brian Nosek, U Virginia (corresponding author)

Why reproducible research? Robustness

PHYSICS TODAY

22 Aug 2018 in Research & Technology

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top theorists.

"Limmer maintains that he and his mentor weren't trying to hide anything. "I had and was very willing to share the code," he says. What he didn't have, he says, was the time or personnel to prepare the code in a form that could be useful to an outsider"

"Suddenly it made sense that the Berkeley researchers hadn't seen a second liquid phase; they were effectively running their simulations at temperatures well above the critical point."

https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/

Why reproducible research? Records

Email from grad student to Berkeley professor in 2017:

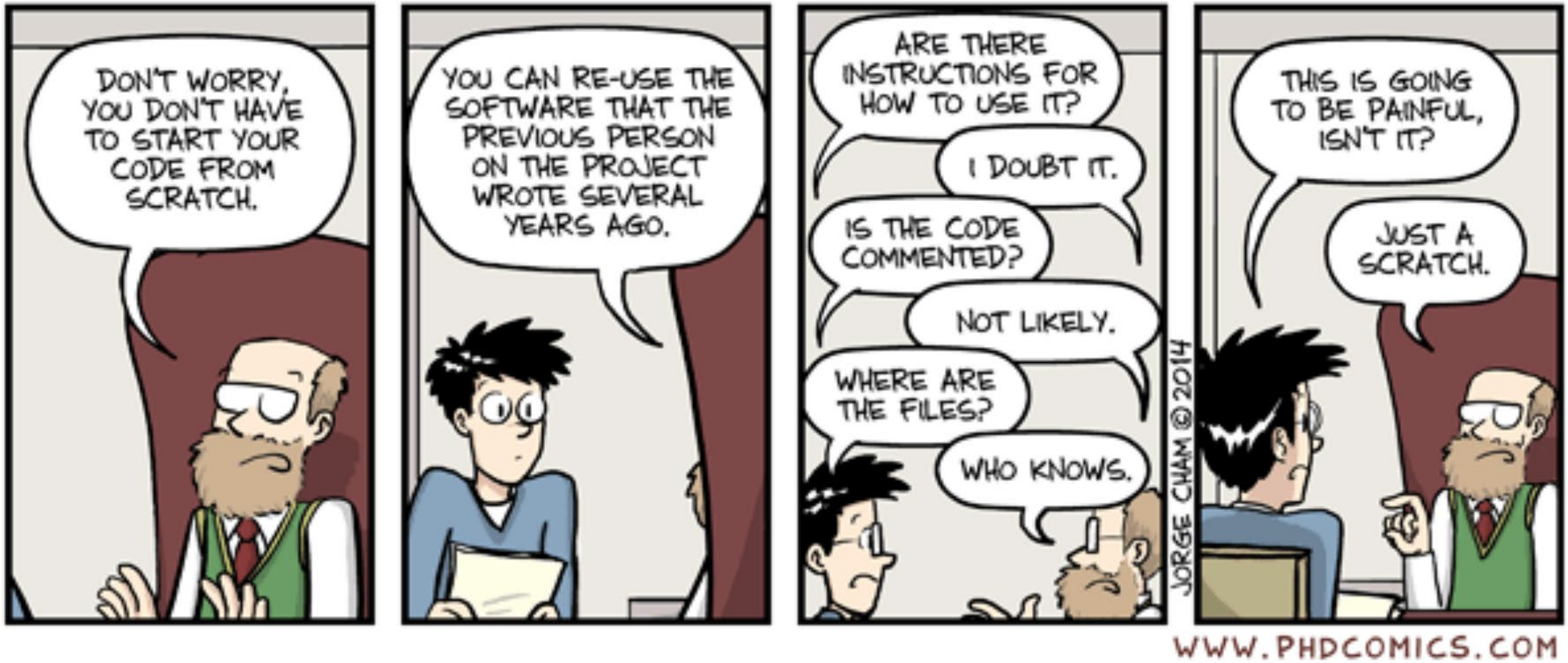
> "…The citation on the slide is [your report on a state government website].
>
> I downloaded that PDF and do not see [important statistic] anywhere in it.
> Table 3 seems to have [different value] , Table 15 has [related value] in California…
>
> …The closest match that I've found is [other statistic] from Table 15...
> Is it possible there was an error in transcription? If not, can you help me figure out what's going on here? What does [important statistic] on [slide] represent?"

Response from Berkeley professor:

> "I don't remember the details. [Former student], do you? Thanks in advance."

> [Former student] never responded

Why reproducible research? Reusability



http://phdcomics.com/coamics/archive.php?comicid=1689

Why is this important?

- Show evidence of the correctness of our results
- Enable others to make use of our methods and results (including future you)
- Increasingly encouraged or required by funders and journals

http://ropensci.github.io/reproducibility-guide/sections/introduction/

https://www.stat.berkeley.edu/~stark/Seminars/reproNE16.htm

https://cos.io/our-services/open-science-badges

“reproducibility”

What does reproducibility mean? Three ideas to distinguish

1. Reproducibility (sometimes called repeatability or even 'preproducibility') [6]

- *Data, and/or statistical methods, code, software environment are available so data analysis can be repeated and get a similar result [1, 2, 3]*

2. Replicability (sometimes called reproducibility)

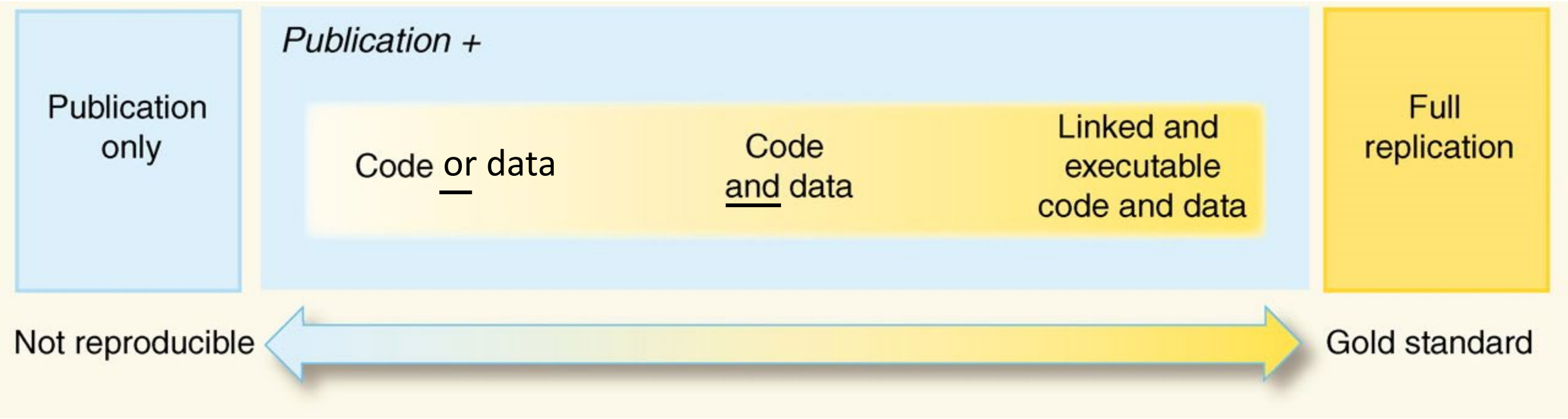
- *Another researcher is able to conduct the same experiment and get similar results*

3. Correctness

- *Is your analysis appropriate? Is your hypothesis correct?*
  → *Reproducible research can still be wrong! [4]*

1 - Peng *Science* 2011, DOI: 10.1126/science.1213847
2 - http://ropensci.github.io/reproducibility-guide/sections/introduction/
3 - https://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/, Peng 2014
4 - Reproducible research can still be wrong, Leek, Peng PNAS 2015, DOI: 10.1073/pnas.1421412111
5- https://www.nature.com/news/muddled-meanings-hamper-efforts-to-fix-reproducibility-crisis-1.20076#/b1, citing https://www.ascb.org/wp-content/uploads/2015/11/How-can-scientist-enhance-rigor.pdf citing Schmidt, Stefan. (2009). Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. Review of General Psychology. 13. 90-100. 10.1037/a0015108.
6 – Barba 2018 https://arxiv.org/pdf/1802.03311.pdf

A spectrum of practice (depends on project)



Adopted from:
Reproducible Research in Computational Science
By Roger D. Peng
Science 02 Dec 2011 : 1226-1227
http://science.sciencemag.org/content/334/6060/1226.full

Why might reproducibility be hard?

Stodden (2010) Survey of NIPS:

Code	Complaint/Excuse	Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	N/A
34%	Legal Barriers (ie. copyright)	41%
N/A	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

I.e., fear, greed, ignorance, & sloth.

Philip Stark,
https://www.stat.berkeley.edu/~stark/Seminars/reproNE16.htm

What's possible – interactive publications!

New paper
(open access)

ECOLOGY LETTERS

Ecology Letters, (2018) doi: 10.1111/ele.13085

REVIEW AND SYNTHESIS

From noise to knowledge: how randomness generates novel phenomena and reveals information

Carl Boettiger*

Dept of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, CA 94720-3114, USA

*Correspondence: E-mail: cboettig@berkeley.edu

Abstract

Noise, as the term itself suggests, is most often seen a nuisance to ecological insight, a inconvenient reality that must be acknowledged, a haystack that must be stripped away to reveal the processes of interest underneath. Yet despite this well-earned reputation, noise is often interesting in its own right: noise can induce novel phenomena that could not be understood from some underlying deterministic model alone. Nor is all noise the same, and close examination of differences in frequency, colour or magnitude can reveal insights that would otherwise be inaccessible. Yet with each aspect of stochasticity leading to some new or unexpected behaviour, the time is right to move beyond the familiar refrain of "everything is important" (Bjørnstad & Grenfell 2001). Stochastic phenomena can suggest new ways of inferring process from pattern, and thus spark more dialog between theory and empirical perspectives that best advances the field as a whole. I highlight a few compelling examples, while observing that the study of stochastic phenomena are only beginning to make this translation into empirical inference. There are rich opportunities at this interface in the years ahead.

Keywords

Coloured noise, demographic noise, environmental noise, quasi-cycles, stochasticity, tipping points.

Ecology Letters (2018)

INTRODUCTION: NOISE THE NUISANCE

To many, stochasticity, or more simply, noise, is just that – something which obscures patterns we are trying to infer (Knape & de Valpine 2011); and an ever richer batteries of statistical methods are developed largely in an attempt to strip away this undesirable randomness to reveal the patterns beneath (Coulson 2001). Over the past several decades, literature in stochasticity has transitioned from thinking of stochasticity in such terms; where noise is a nuisance that obscures the deterministic skeleton of the underlying mechanisms, to the recognition that stochasticity can itself be a mechanism for driving many interesting phenomena (Coulson *et al.* 2004). Yet this transition from *noise the nuisance* to *noise the creator* of ecological phenomena has had, with a few notable exceptions, relatively little impact in broader thinking about stochasticity. One of the most provocative of those exceptions has turned the classical notion of noise the nuisance on its head: recognising that noise driven phenomena can become a tool to reveal underlying processes: to become *noise the informer*. Here I argue that this third shift in perspective offers an opportunity to better bridge the divide between respective primarily theoretical and primarily empirical communities by seeing noise not as mathematical curiosity or statistical bugbear, but as a source for new opportunities for inference.

In arguing for this shift, it essential to recognise this is a call for a bigger tent, not for the rejection of previous paradigms. What I will characterize as 'noise the nuisance' reflects a predominately statistical approach, in which noise, almost by definition, represents all the processes we are not interested in that create additional variation which might obscure the pattern of interest. By contrast, an extensive literature has long explored how noise itself can create patterns and explain processes from population cycling to coexistence. These broad categories should be seen as a spectrum and not be mistaken for either a sharp dichotomy nor a reference to a strictly empirical-theoretical divide. Each paradigm expands upon rather than rejects the previous notion of noise: the recognition that noise can create novel phenomena does not mean that noise cannot also obscure the signal of some process of interest. Likewise, seeking to use noise as a novel source of information about underlying processes will be informed by both previous paradigms, as our discussion will illustrate.

Accompanying this discussion, I provide concise and commented code for simulating each of the models we will discuss as Appendix S1, and more mathematical background and derivations in Appendix S2. Numerical simulations permit poking and prodding investigation unencumbered by either experimental design or mathematical formalism. A copy of these appendices is maintained at https://github.com/cboettig/noise-phenomena.

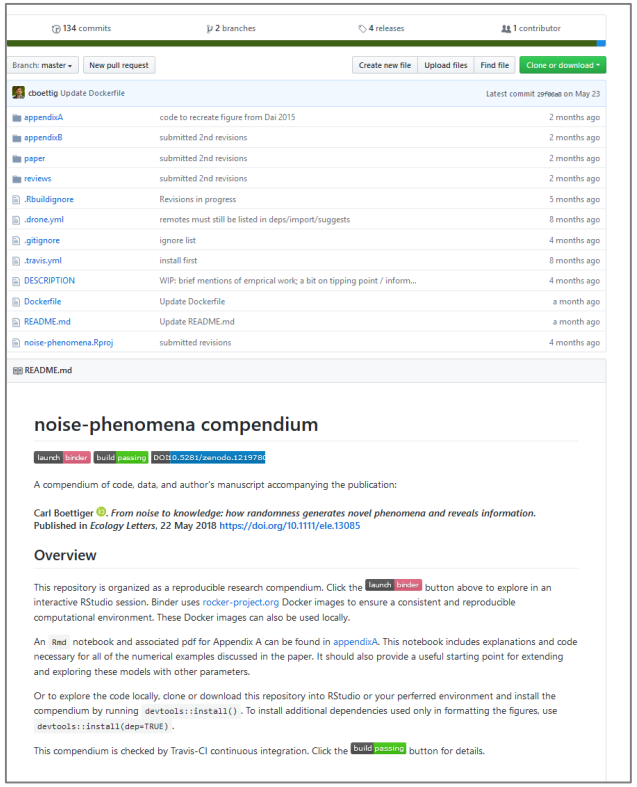
To emphasise the underlying trend in the changing roles in which we see and understand noisy processes, I will also restrict my focus to relatively simple models primarily from population ecology context. Simplicity not only makes examples (in equations and in code) more tractable but also allows us to focus on aspects that are germane to many contexts rather than unique to particular complexities (Bartlett 1960;

© 2018 John Wiley & Sons Ltd/CNRS

https://onlinelibrary.wiley.com/doi/epdf/10.1111/ele.13085

+

Corresponding "compendium"
(with text, code, data)



134 commits | 2 branches | 4 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

cboettig Update Dockerfile — Latest commit 29f6ad0 on May 23

appendixA	code to recreate figure from Dai 2015	2 months ago
appendixB	submitted 2nd revisions	2 months ago
paper	submitted 2nd revisions	2 months ago
reviews	submitted 2nd revisions	2 months ago
.Rbuildignore	Revisions in progress	5 months ago
.drone.yml	remotes must still be listed in deps/import/suggests	8 months ago
.gitignore	ignore list	4 months ago
.travis.yml	install first	8 months ago
DESCRIPTION	WIP: brief mentions of emprical work; a bit on tipping point / inform...	4 months ago
Dockerfile	Update Dockerfile	a month ago
README.md	Update README.md	a month ago
noise-phenomena.Rproj	submitted revisions	4 months ago

README.md

noise-phenomena compendium

launch binder | build passing | DOI 10.5281/zenodo.121978

A compendium of code, data, and author's manuscript accompanying the publication:

Carl Boettiger. *From noise to knowledge: how randomness generates novel phenomena and reveals information.* Published in *Ecology Letters*, 22 May 2018 https://doi.org/10.1111/ele.13085

Overview

This repository is organized as a reproducible research compendium. Click the launch binder button above to explore in an interactive RStudio session. Binder uses rocker-project.org Docker images to ensure a consistent and reproducible computational environment. These Docker images can also be used locally.

An Rmd notebook and associated pdf for Appendix A can be found in appendixA. This notebook includes explanations and code necessary for all of the numerical examples discussed in the paper. It should also provide a useful starting point for extending and exploring these models with other parameters.

Or to explore the code locally, clone or download this repository into RStudio or your perferred environment and install the compendium by running devtools::install(). To install additional dependencies used only in formatting the figures, use devtools::install(dep=TRUE).

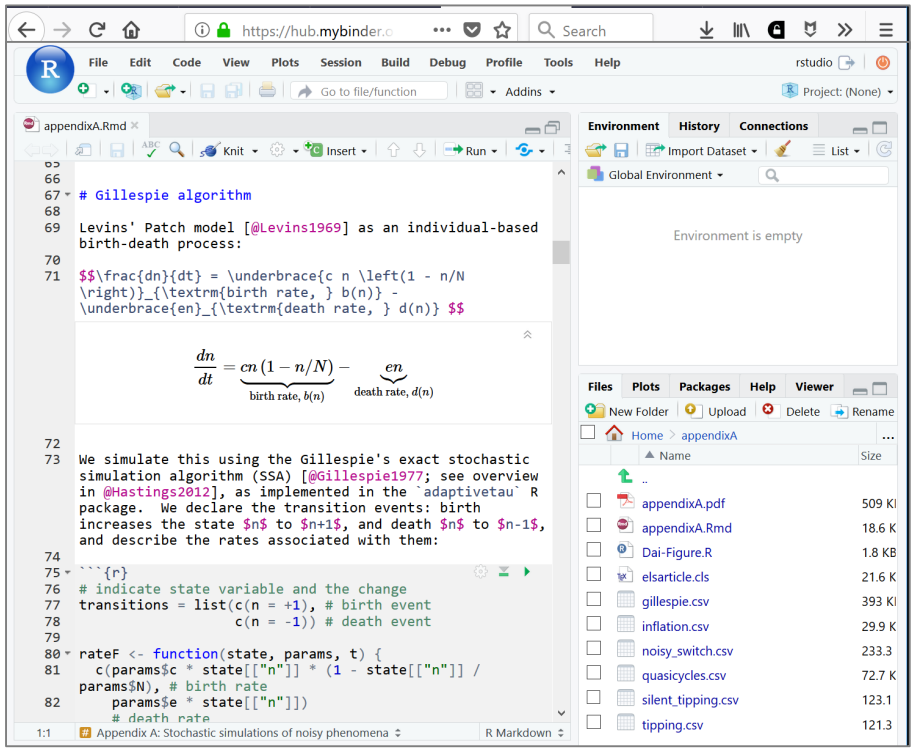
This compendium is checked by Travis-CI continuous integration. Click the build passing button for details.

https://github.com/cboettig/noise-phenomena

+

Interactive R notebook to explore code

(with all required files and software ready to run in your browser with no installation, thanks to Binder)



https://hub.mybinder.c

File Edit Code View Plots Session Build Debug Profile Tools Help rstudio

appendixA.Rmd

```
# Gillespie algorithm

Levins' Patch model [@Levins1969] as an individual-based
birth-death process:

$$\frac{dn}{dt} = \underbrace{c n \left(1 - n/N
\right)}_{\textrm{birth rate, } b(n)} -
\underbrace{en}_{\textrm{death rate, } d(n)} $$
```

$$\frac{dn}{dt} = \underbrace{cn\left(1 - n/N\right)}_{\text{birth rate, } b(n)} - \underbrace{en}_{\text{death rate, } d(n)}$$

```
We simulate this using the Gillespie's exact stochastic
simulation algorithm (SSA) [@Gillespie1977; see overview
in @Hastings2012], as implemented in the `adaptivetau` R
package.  We declare the transition events: birth
increases the state $n$ to $n+1$, and death $n$ to $n-1$,
and describe the rates associated with them:

```{r}
# indicate state variable and the change
transitions = list(c(n = +1), # birth event
                   c(n = -1)) # death event

rateF <- function(state, params, t) {
  c(params$c * state[["n"]] * (1 - state[["n"]] /
params$N), # birth rate
    params$e * state[["n"]])
    # death rate
```

Environment | History | Connections

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

New Folder | Upload | Delete | Rename

Home > appendixA

Name	Size
..	
appendixA.pdf	509 KI
appendixA.Rmd	18.6 K
Dai-Figure.R	1.8 KB
elsarticle.cls	21.6 K
gillespie.csv	393 KI
inflation.csv	29.9 K
noisy_switch.csv	233.3
quasicycles.csv	72.7 K
silent_tipping.csv	123.1
tipping.csv	121.3

1:1 | Appendix A: Stochastic simulations of noisy phenomena | R Markdown
```

Interactive publications – in python too!

New paper
(in this case, a preprint)

+

Website explaining paper

(has code to replicate experiments + plots and explicit software versions)

+

Interactive Jupyter notebook with code to apply their method to your own images

(with all required files and software ready to run in your browser with no installation required, thanks to Binder)

Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning

David Mascharka*[1] Philip Tran[2] Ryan Soklaski[1] Arjun Majumdar*[1]
[1]MIT Lincoln Laboratory[†] [2]Planck Aerosystems[‡]
{first.last}@ll.mit.edu, phil@planckaero.com

arXiv:1803.05268v1 [cs.CV] 14 Mar 2018

Abstract

Visual question answering requires high-order reasoning about an image, which is a fundamental capability needed by machine systems to follow complex directives. Recently, modular networks have been shown to be an effective framework for performing visual reasoning tasks. While modular networks were initially designed with a degree of model transparency, their performance on complex visual reasoning benchmarks was lacking. Current state-of-the-art approaches do not provide an effective mechanism for understanding the reasoning process. In this paper, we close the performance gap between interpretable models and state-of-the-art visual reasoning methods. We propose a set of visual-reasoning primitives which, when composed, manifest as a model capable of performing complex reasoning tasks in an explicitly-interpretable manner. The fidelity and interpretability of the primitives' outputs enable an unparalleled ability to diagnose the strengths and weaknesses of the resulting model. Critically, we show that these primitives are highly performant, achieving state-of-the-art accuracy of 99.1% on the CLEVR dataset. We also show that our model is able to effectively learn generalized representations when provided a small amount of data containing novel object attributes. Using the CoGenT generalization task, we show more than a 20 percentage point improvement over the current state of the art.

1. Introduction

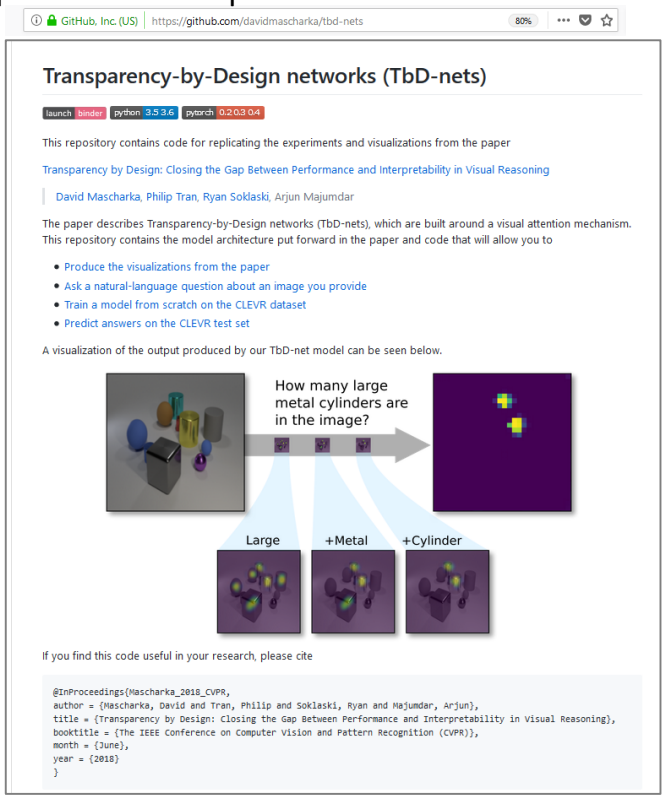
A visual question answering (VQA) model must be capable of complex spatial reasoning over an image. For example, in order to answer the question "What color is the cube to the right of the large metal sphere?", a model must identify which sphere is the large metal one, understand what it means for an object to be to the right of another, and apply this concept spatially to the attended sphere. Within this new region of interest, the model must find the cube and determine its color. This behavior should be compositional to allow for arbitrarily long reasoning chains.

While a wide variety of models have recently been proposed for the VQA task [8, 12, 23, 26, 35, 37], neural module networks [2, 3, 12, 18] are among the most intuitive. Introduced by Andreas *et al.* [2], neural module networks compose a question-specific neural network, drawing from a set of modules that each perform an individual operation. This design closely models the compositional nature of visual reasoning tasks. In the original work, modules were designed with an attention mechanism, which allowed for insight into the model's operation. However, the approach did not perform well on complex visual reasoning tasks such as CLEVR [17]. Modifications by Johnson *et al.* [18] address the performance issue at the cost of losing model transparency. This is problematic, because the ability to inspect each step of the reasoning process is crucial

Figure 1. A diagram of a visual question answering task, in which our proposed Transparency by Design network (TbD-net) composes a series of attention masks that allow it to correctly count two large metal cylinders in the image.

*Indicates equal contribution.
[†]This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.
[‡]This work conducted while Philip was at MIT Lincoln Laboratory.

https://arxiv.org/abs/1803.05268



GitHub, Inc. (US) https://github.com/davidmascharka/tbd-nets

Transparency-by-Design networks (TbD-nets)

launch binder python 3.5 3.6 pytorch 0.2 0.3 0.4

This repository contains code for replicating the experiments and visualizations from the paper

Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning

David Mascharka, Philip Tran, Ryan Soklaski, Arjun Majumdar

The paper describes Transparency-by-Design networks (TbD-nets), which are built around a visual attention mechanism. This repository contains the model architecture put forward in the paper and code that will allow you to

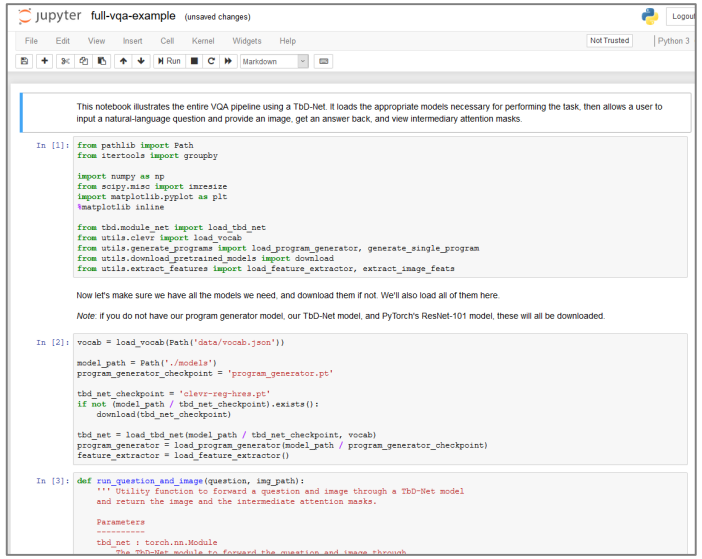
- Produce the visualizations from the paper
- Ask a natural-language question about an image you provide
- Train a model from scratch on the CLEVR dataset
- Predict answers on the CLEVR test set

A visualization of the output produced by our TbD-net model can be seen below.

If you find this code useful in your research, please cite

```
@InProceedings{Mascharka_2018_CVPR,
author = {Mascharka, David and Tran, Philip and Soklaski, Ryan and Majumdar, Arjun},
title = {Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning},
booktitle = {The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2018}
}
```

https://github.com/davidmascharka/tbd-nets



jupyter full-vqa-example (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

This notebook illustrates the entire VQA pipeline using a TbD-Net. It loads the appropriate models necessary for performing the task, then allows a user to input a natural-language question and provide an image, get an answer back, and view intermediary attention masks.

In [1]:
```
from pathlib import Path
from itertools import groupby

import numpy as np
from scipy.misc import imresize
import matplotlib.pyplot as plt
%matplotlib inline

from tbd.module_net import load_tbd_net
from utils.clevr import load_vocab
from utils.generate_programs import load_program_generator, generate_single_program
from utils.download_pretrained_models import download
from utils.extract_features import load_feature_extractor, extract_image_feats
```

Now let's make sure we have all the models we need, and download them if not. We'll also load all of them here.

Note: if you do not have our program generator model, our TbD-Net model, and PyTorch's ResNet-101 model, these will all be downloaded.

In [2]:
```
vocab = load_vocab(Path('data/vocab.json'))

model_path = Path('./models')
program_generator_checkpoint = 'program_generator.pt'

tbd_net_checkpoint = 'clevr-reg-hres.pt'
if not (model_path / tbd_net_checkpoint).exists():
    download(tbd_net_checkpoint)

tbd_net = load_tbd_net(model_path / tbd_net_checkpoint, vocab)
program_generator = load_program_generator(model_path / program_generator_checkpoint)
feature_extractor = load_feature_extractor()
```

In [3]:
```
def run_question_and_image(question, img_path):
    ''' Utility function to forward a question and image through a TbD-Net model
    and return the image and the intermediate attention masks.

    Parameters
    ----------
    tbd_net : torch.nn.Module
```

https://mybinder.org/v2/gh/davidmascharka/tbd-nets/binder?filepath=full-vqa-example.ipynb

Additional reference: Link to tweet by first author publicizing this work

Demo – reproduce a figure from paper on the previous slide

See the paper: https://onlinelibrary.wiley.com/doi/epdf/10.1111/ele.13085

- Note the PDF includes a link to a GitHub repo

2. Inspect the code

- *Open the associated GitHub repository containing the code and data used to produce the figures for the paper*
  - *Repo: **https://github.com/cboettig/noise-phenomena***
  - *Code for the paper: https://github.com/cboettig/noise-phenomena/blob/master/paper/paper.Rmd*

3. Reproduce a figure without installing anything on your local computer!

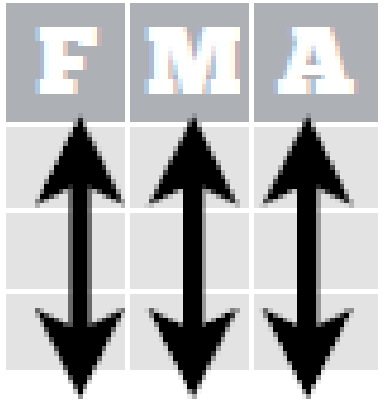
- *Click the* launch binder *button on the main page of the repo*

What are tools and practices we can start using now?

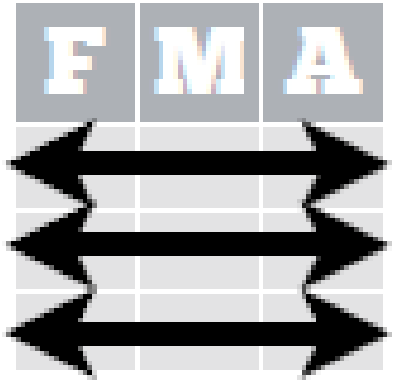
Data collection

1. Back up your data regularly

2. “Tidy” spreadsheet formats



&



Each **variable** is saved in its own **column**

Each **observation** is saved in its own **row**

www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

Is this a tidy dataset?

Subject	Thermal sensation votes
Person1	0.5, 0.7
Person2	-0.3
Person3	1.2, 1.5, 2

No, has more than one observation per row

Based on https://swcarpentry.github.io/visualization-novice/12-tidyr/

Ok, is it a tidy dataset now?

Subject	TS_vote_1	TS_vote_2	TS_vote_3
Person1	0.5	0.7	
Person2	-0.3		
Person3	1.2	1.5	2

No, has more than one observation per row,
and more than one column per type of information

Based on https://swcarpentry.github.io/visualization-novice/12-tidyr/

A tidy dataset

Subject	ThermalSensationVote
Person1	0.5
Person1	0.7
Person2	-0.3
Person3	1.2
Person3	1.5
Person3	2

Based on https://swcarpentry.github.io/visualization-novice/12-tidyr/

Is this weather data tidy?

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1 - 23.doi:http://dx.doi.org/10.18637/jss.v059.i10

No, has more than one observation per row

What improved?

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1 - 23.doi:http://dx.doi.org/10.18637/jss.v059.i10

What might be a problem in these spreadsheets?

A

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102		95.3
4	103	2015–06–18	97.5
5	104		117.0
6	105		108.0
7	106	2015–06–20	149.0
8	107		169.4

B

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

How would you enter this data into R?

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

How would you enter this data into R?

A

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102	2015–06–14	95.3
4	103	2015–06–18	97.5
5	104	2015–06–18	1.1
6	105	2015–06–18	108.0
7	106	2015–06–20	149.0
8	107	2015–06–20	169.4

B

	A	B	C	D
1	id	date	glucose	outlier
2	101	2015–06–14	149.3	FALSE
3	102	2015–06–14	95.3	FALSE
4	103	2015–06–18	97.5	FALSE
5	104	2015–06–18	1.1	TRUE
6	105	2015–06–18	108.0	FALSE
7	106	2015–06–20	149.0	FALSE
8	107	2015–06–20	169.4	FALSE

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

More on tidy data and spreadsheet organization

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

- Be consistent
- Write dates as YYYY-MM-DD
- No empty cells
- Put just one thing in a cell
- Make it a rectangle
- Create a data dictionary
- No calculations in the raw data files
- Font color or highlighting isn't data
- Make backups
- Use data validation to avoid errors
- Save the data in plain text files

Data analysis

1. File organization

- Suggest one folder per project, eg:

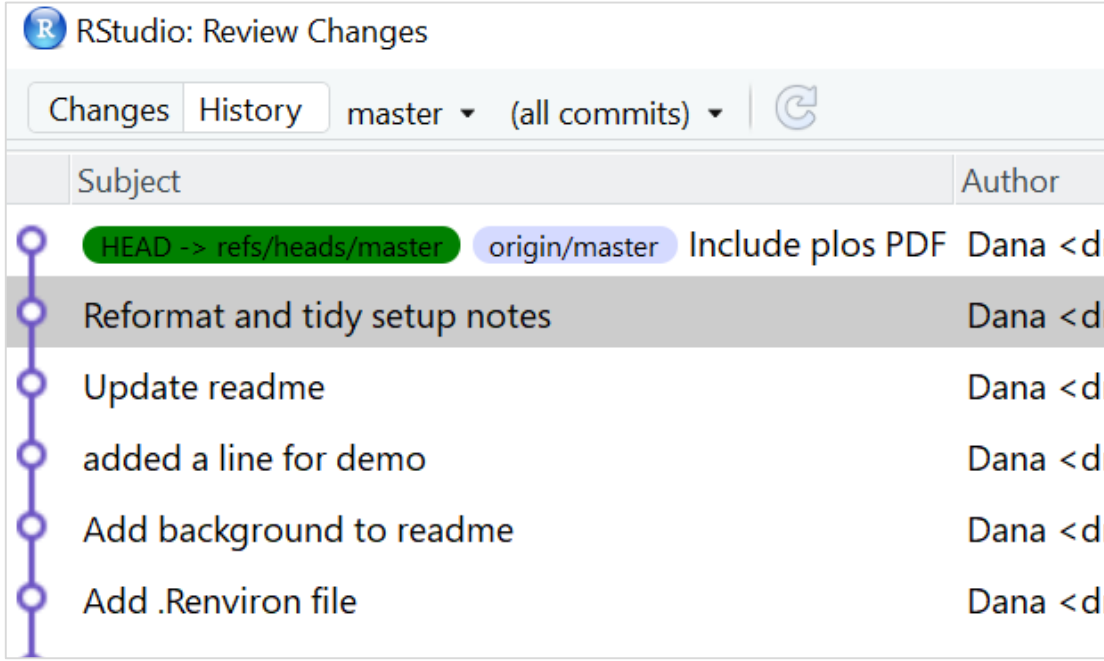
```
My_project
        Data
            raw_data.csv
        Docs
            Manuscript.doc
        Results
            Processed_results.csv
        Scripts
            clean_data.R
            fit_model_and_graph.R
```

Resources:

- A Quick Guide to Organizing Computational Biology Projects
- Good enough practices in scientific computing

2. Version control (for code)

- Like MS Word's Previous Versions → Restore, or unlimited undo for your R scripts



Resources:

- http://happygitwithr.com/
- A quick introduction to version control with Git and GitHub

Sharing + publication of results

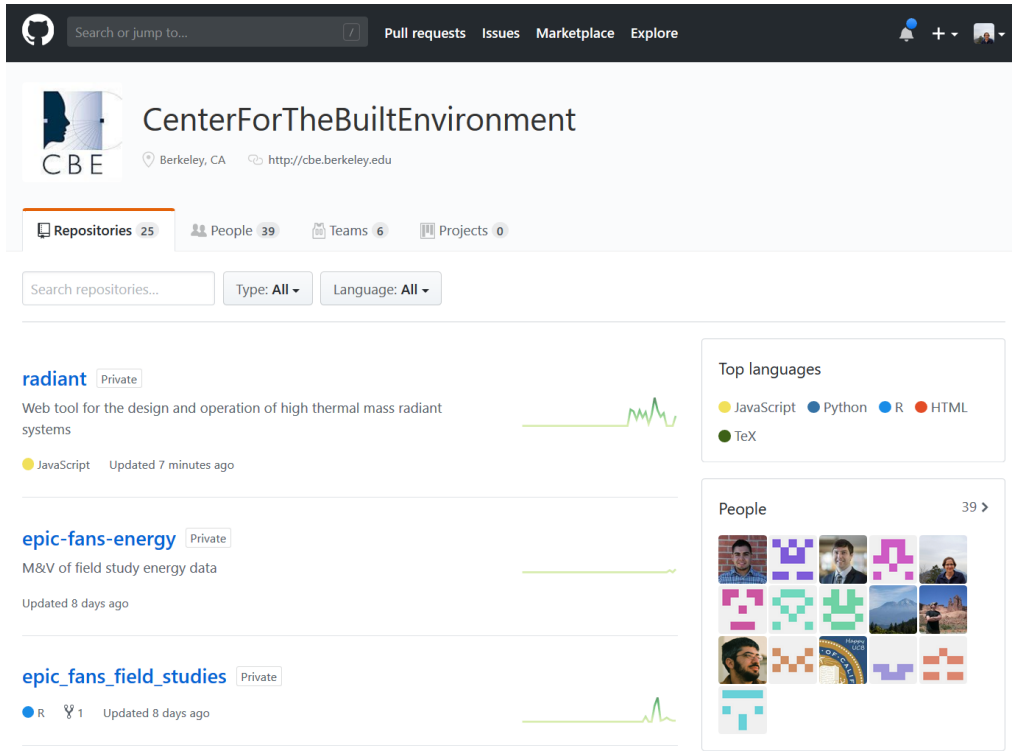
1. Data repositories

- UC system – dash
- Example: ASHRAE Global Thermal Comfort Database II

Resources:

- Berkeley Research Data Management librarians!
- Other repositories include Zenodo, Open Science Framework, DataDryad, Figshare

2. Code repositories



Resources:

- Software Carpentry's Intro to Version Control for Novices
- A quick introduction to version control with Git and GitHub

Demo – make a new project in RStudio

1. Open RStudio (this is a good time for questions)

2. File → New project → Choose directory → New project

   (Detailed instructions: RStudio "Using Projects")

3. What are benefits of Rstudio Projects?

- All files for one project in one place
- Easily switch between projects
- Starts you in a fresh R session each time
- Integration with git (version control)

4. If time allows:

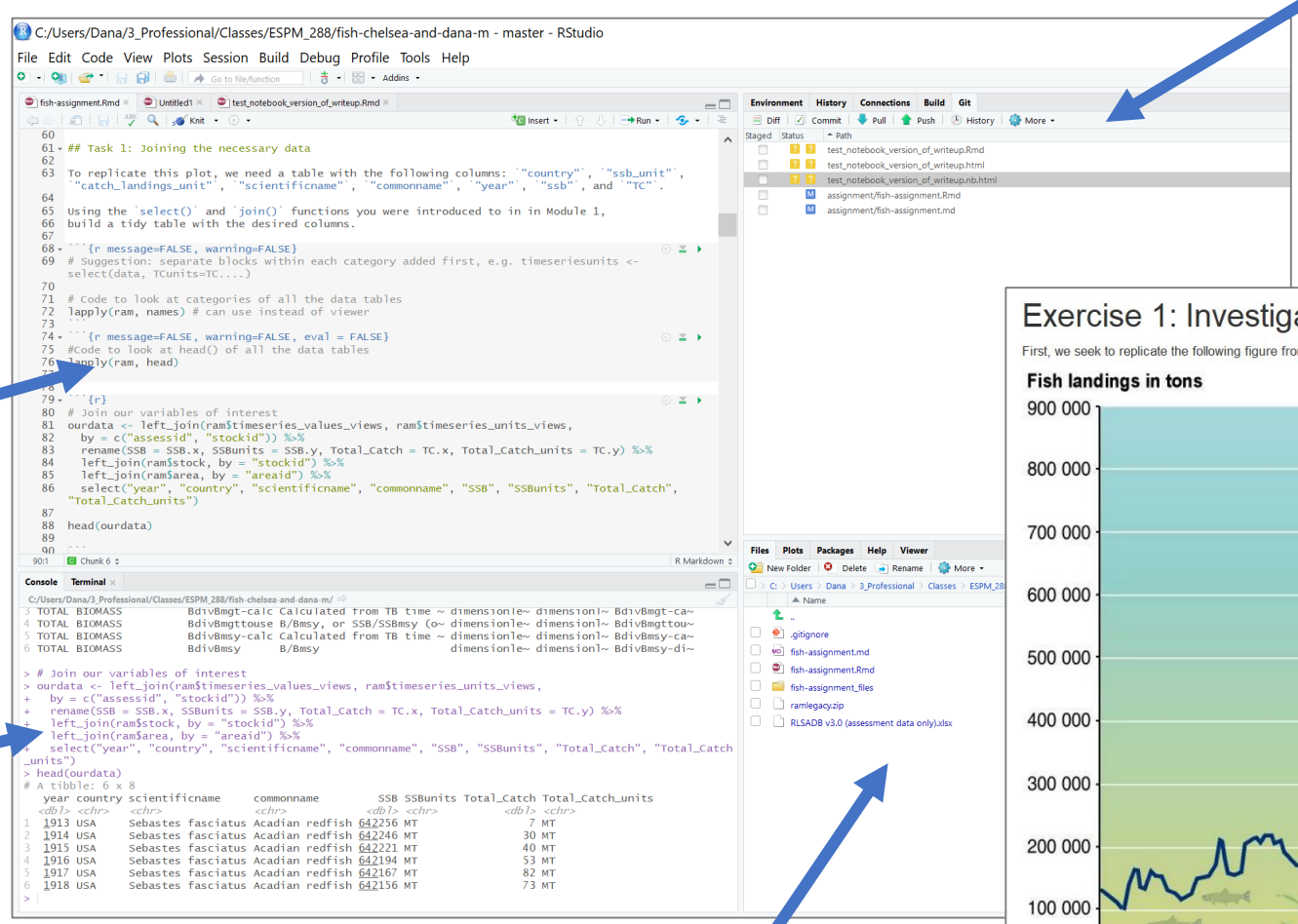
- R Markdown files
- Look at example of paper being written in Markdown
- Using git inside RStudio

R Markdown documents

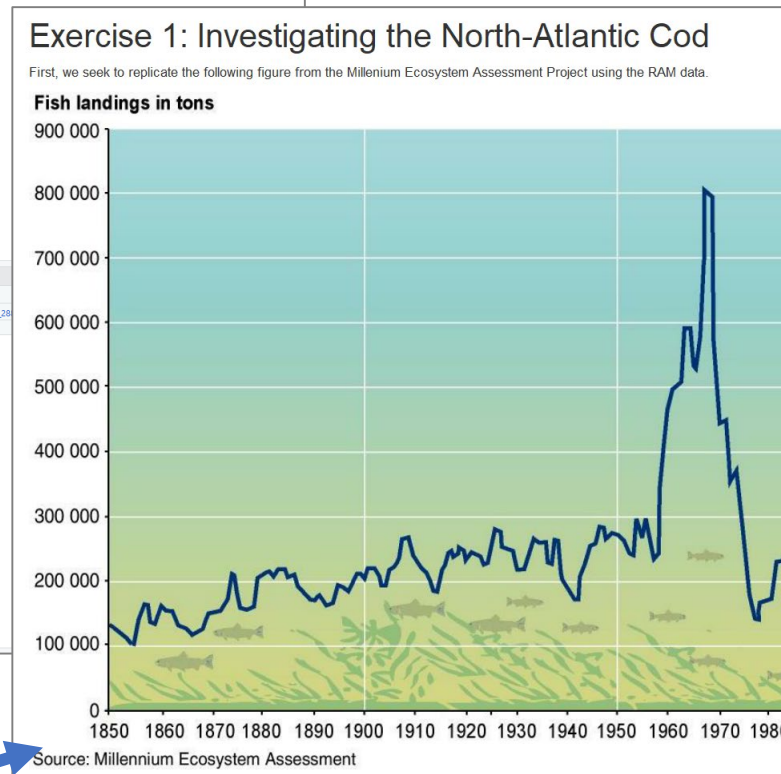
Version control tracker

See happygitwithr

R Markdown (.Rmd) document mixing code and text



Code can be executed in-line (above) or in console (below)



See file explorer, help documentation, plots, packages

Convert file to .md, .doc, .pdf, html, html notebook, slides, book, blog, dashboards, Shiny app, etc

See R Markdown output formats

Literate programming with R Markdown

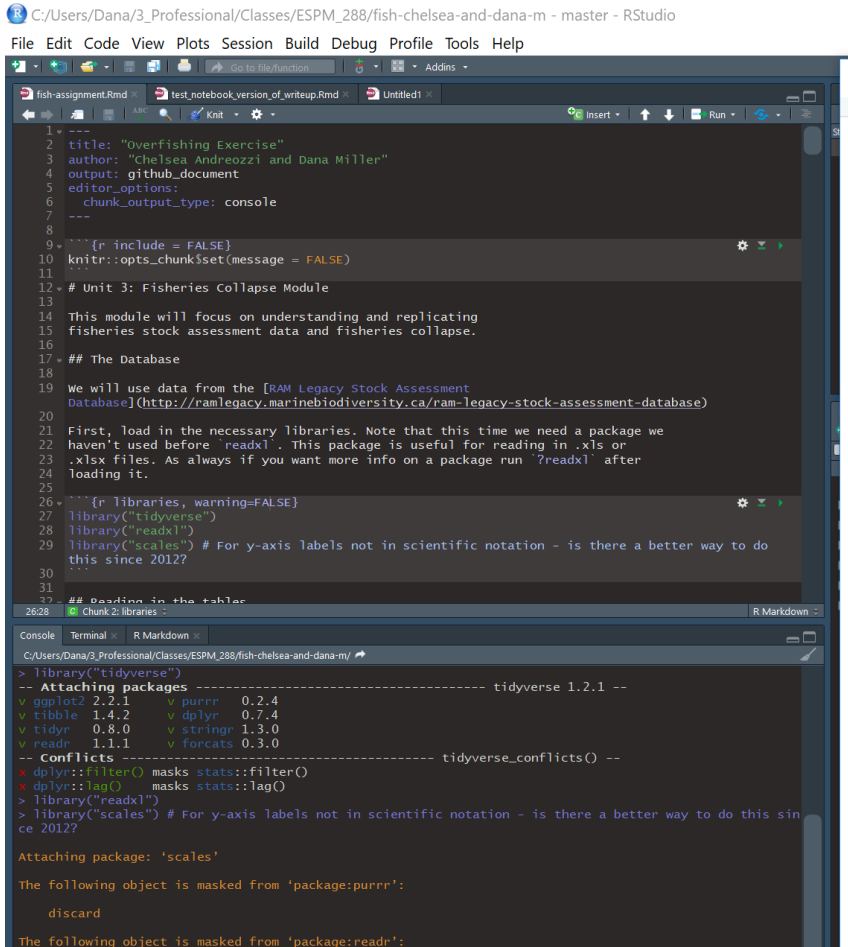
Starting with one .Rmd file with both R code and text

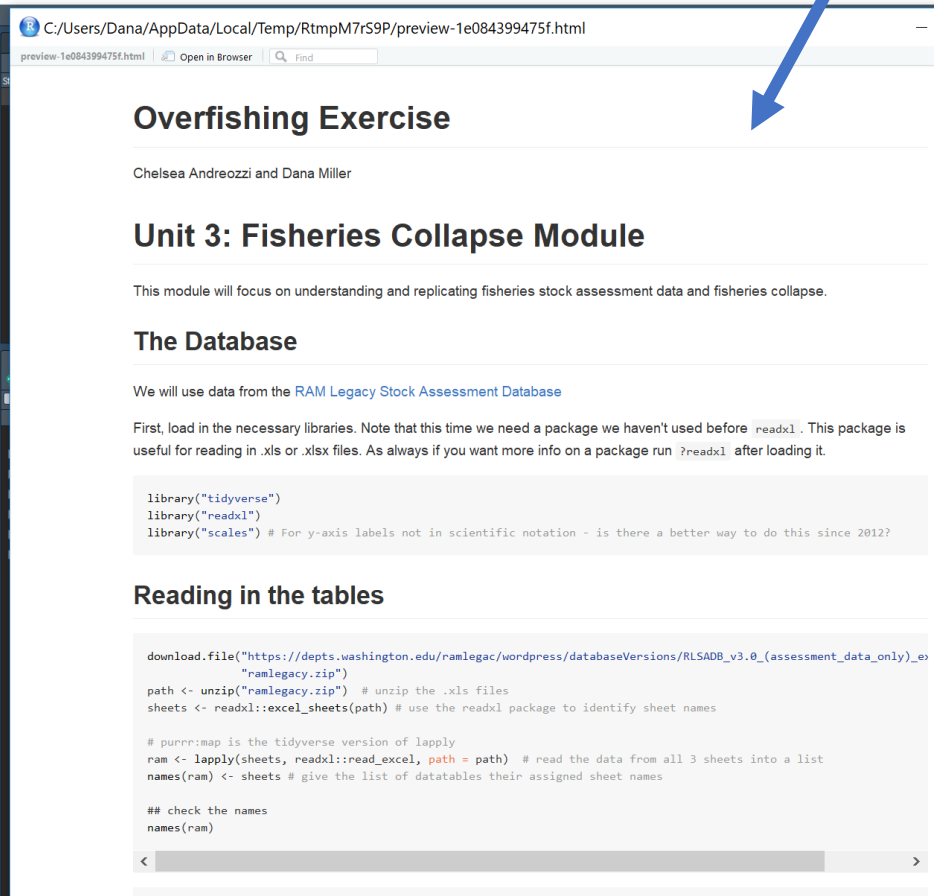
This button or command:

executes all R code (output shown in console)

This button or command:

renders formatted document with executed code and text (.md, .pdf, .doc etc)





What next?

UCB Prof Phillip Stark on reproducibility:

"Science is *show me*, not *trust me*"
"Perfection is impossible but improvement is easy"

There are many communities who can help!

On campus: BIDS Data Analysis Tool Series, D-Lab, Librarians, Research Computing, Software Carpentry Workshops…
Beyond: R for Data Analysis (R4DS), RBloggers, PyData (also has R videos), RLadies, #rstats…

And we can learn a lot from each other ☺

https://www.stat.berkeley.edu/~stark/Seminars/reproNE16.htm

Some more R resources

R for Data Science: *https://r4ds.had.co.nz*

From Excel to R: http://rpubs.com/acolumbus/how-to-use-r-with-excel (includes list of common Excel functions in R)

Markdown guide: https://bookdown.org/yihui/rmarkdown/

Free resources aimed at beginners:

1. introduction to programming in R

2. Reproducible scientific analysis using RStudio and R

Useful but not fully free:

Datacamp - Each of these video tutorial series + hands on exercises is free to start and then unfortunately requires a paid subscription to continue. If you are in a class with more than 10 students, the instructor can sign up and then all the students can get access for the whole semester for free

•Introduction to the **RStudio IDE**

•Introduction to the tidyverse: https://www.datacamp.com/courses/introduction-to-the-tidyverse

•Generating reports with R markdown: **https://www.datacamp.com/courses/reporting-with-r-markdown**

Bonus slides!

(for reference)

Where do we go from here?

ROpenSci's perspective:

- **Train students** by putting homework, assignments & dissertations on the reproducible research spectrum
- **Publish examples** of reproducible research in our field
- **Request code & data** when reviewing
- **Submit to & review for journals** that support reproducible research
- **Critically review & audit** data management plans in grant proposals
- Consider reproducibility wherever possible in **hiring, promotion & reference letters**.

http://ropensci.github.io/reproducibility-guide/sections/introduction/ - Leveque et al

https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking#.qs6A0K3DW

From: Brian Wansink
To: David Just
Cc: Collin Payne; Sandra Cuellar
Subject: Can Branding Improve School Lunches?
Date: Saturday, January 7, 2012 7:17:42 AM
Attachments: Elmo Icon-AJPH - 1-7-12.doc
ATT00001.htm

Hi David,

Here's the Elmo study we are going to spin off and submit.
I think we start with the AJPH as a Brief (80 word abstract and 800 word paper), and go from there. I'll give Sandra a list of the journals and the priority order we should consider. Let's consider these two first:

Brief -- American Journal of Public Health

Research Letter – Archives of Pediatric and Adolescent Medicine

One sticking point is that although the stickers increase apple selection by 71%, for some reason this is a p value of .06. It seems to me it should be lower. Do you want to take a look at it and see what you think. If you can get the data, and it needs some tweeking, it would be good to get that one value below .05.

BuzzFeed News

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Breaking the prejudice habit: Mechanisms, timecourse, and longevity☆,☆☆,★

Patrick S. Forscher*, Chelsea Mitamura, Emily L. Dix, William T.L. Cox, Patricia G. Devine*

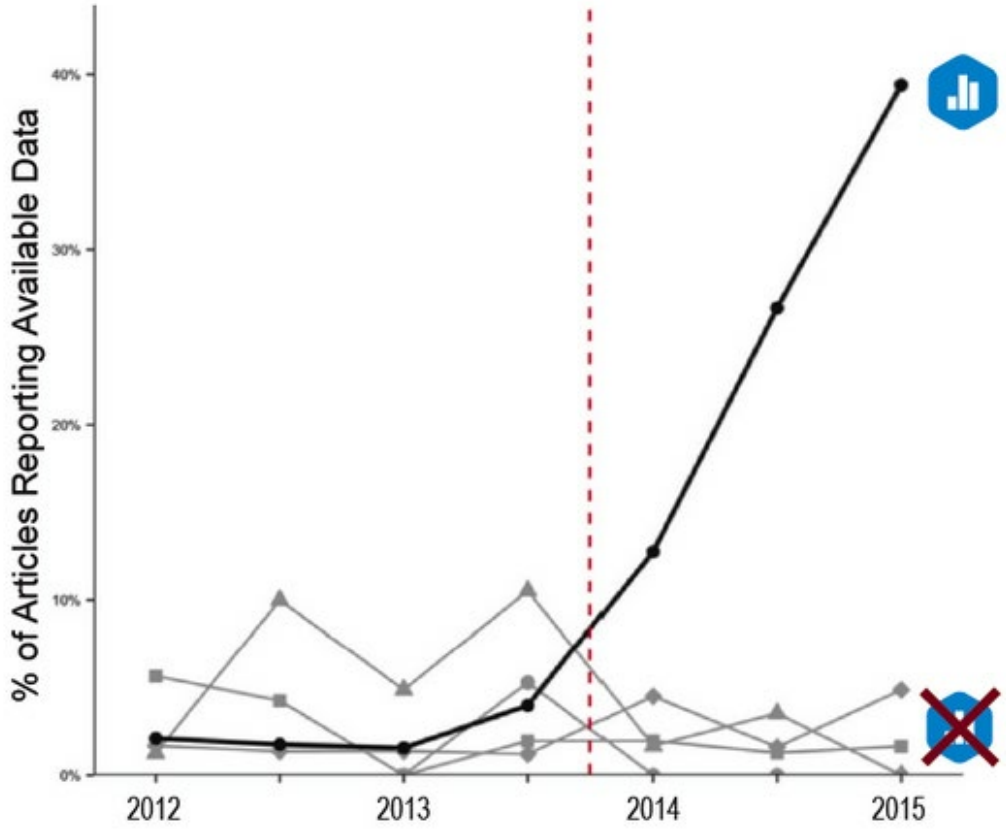
Department of Psychology, University of Wisconsin, Madison, United States

"Badges seem silly. Do they work?

•Yes. Implementing these badges dramatically increases the rate of data sharing (Kidwell et al, 2016).

•A recent systematic review identified this badging program as the only evidence-based incentive program that is associated with increased data sharing (Rowhani-Farid et al., 2017).

•View a list of journals and organizations that have adopted badges here."



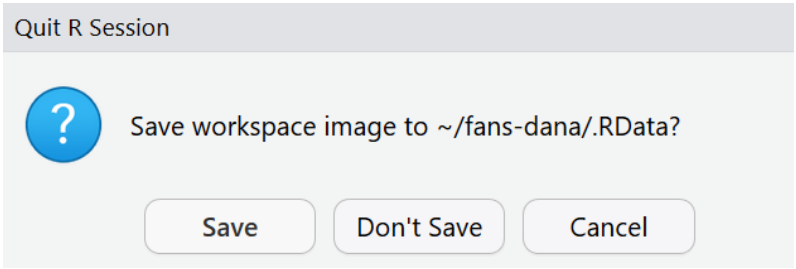
https://cos.io/our-services/open-science-badges/

Three tips you can start using today

- here::here

```
library(here)
db <- read_csv(here("subfolder_name", "file_name.csv"))
```

- "Save workspace image?"
  - Don't save!



- read_csv
  - tidyverse version of read.csv
  - won't coerce strings to factors, outputs a dataframe

What is markdown?

Text using Markdown syntax

```
Heading
=======

## Sub-heading

Paragraphs are separated
by a blank line.

Two spaces at the end of a line
produces a line break.

Text attributes _italic_,
**bold**, `monospace`.

Horizontal rule:

---

Bullet list:

  * apples
  * oranges
  * pears
```

Corresponding HTML produced by a Markdown processor

```
<h1>Heading</h1>

<h2>Sub-heading</h2>

<p>Paragraphs are separated
by a blank line.</p>

<p>Two spaces at the end of a line<br />
produces a line break.</p>

<p>Text attributes <em>italic</em>,
<strong>bold</strong>,
<code>monospace</code>.</p>

<p>Horizontal rule:</p>

<hr />

<p>Bullet list:</p>

<ul>
<li>apples</li>
<li>oranges</li>
```

Text viewed in a browser

Heading [edit]

Sub-heading [edit]

Paragraphs are separated by a blank line.

Two spaces at the end of a line
produces a line break.

Text attributes *italic*, **bold**, `monospace`.

Horizontal rule:

Bullet list:

- apples
- oranges
- pears

Numbered list:

1. wash
2. rinse
3. repeat

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

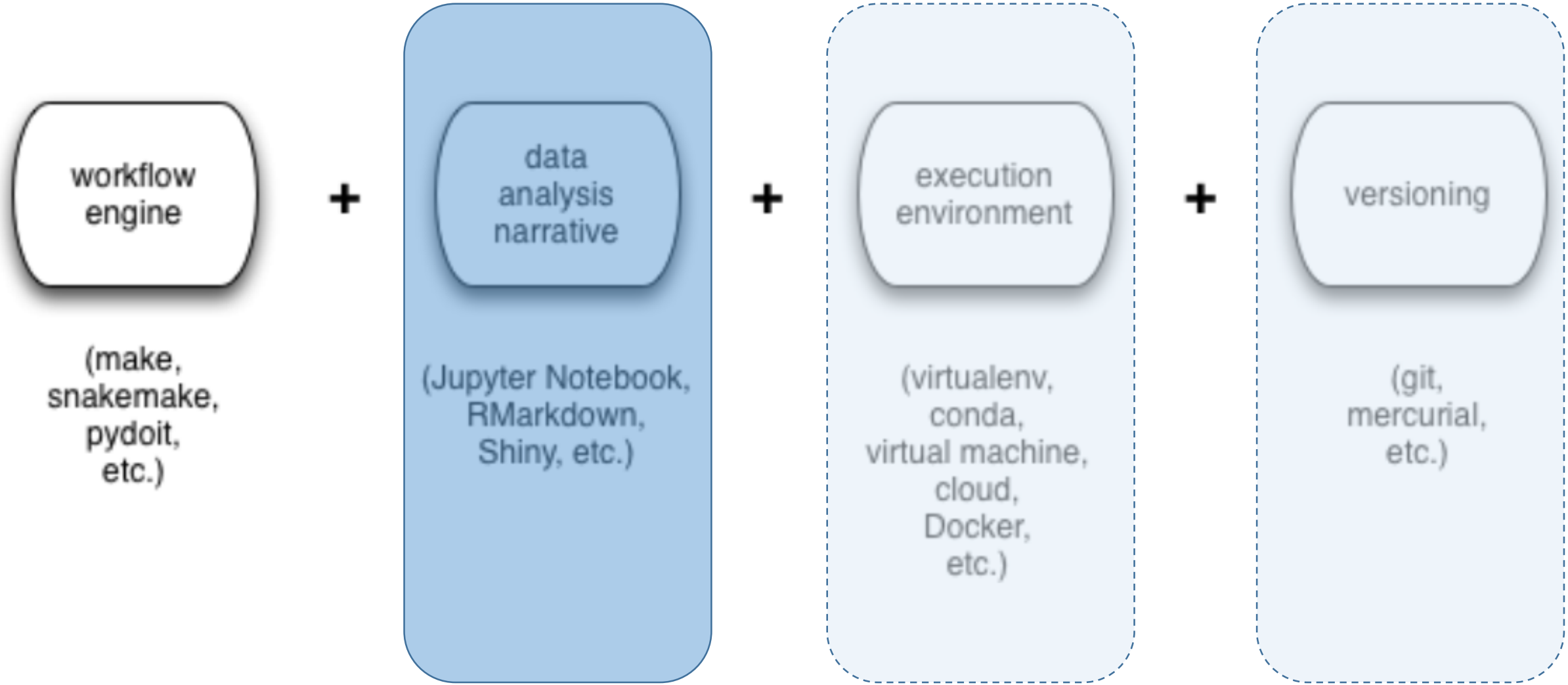
A reproducibility spectrum…example for a concept or theory

Classification created by the American Society for Cell Biology [5]:

1. **analytic replication -** reproduce results by reanalysing original data
2. **direct replication** – repeat original experiment
3. **systematic replication** - repeat with different experimental conditions (eg different cell line)
4. **conceptual replication–** repeat validity of a concept
5. (eg in different organisms)

5- https://www.nature.com/news/muddled-meanings-hamper-efforts-to-fix-reproducibility-crisis-1.20076#/b1, citing https://www.ascb.org/wp-content/uploads/2015/11/How-can-scientist-enhance-rigor.pdf citing Schmidt, Stefan. (2009). Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. Review of General Psychology. 13. 90-100. 10.1037/a0015108.

Key categories of tools



C Titus Brown, UC Davis
https://2016-oslo-repeatability.readthedocs.io/en/latest/overview-and-agenda.html

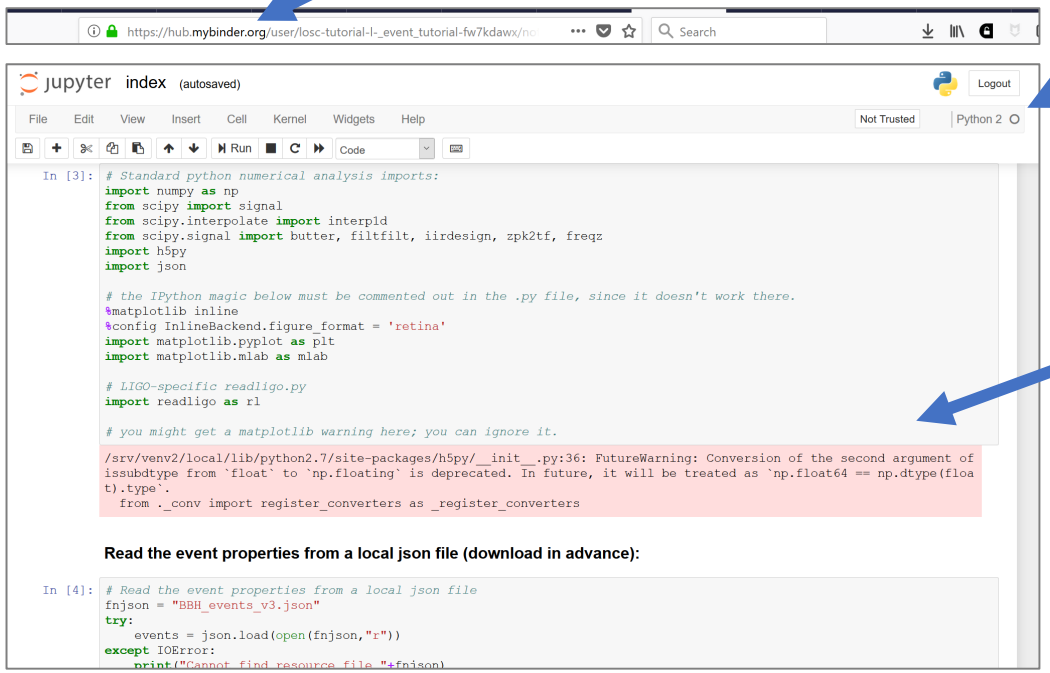
How does Binder work?

"Binder allows you to create custom computing environments that can be shared and used by many remote users"

https://mybinder.readthedocs.io/en/latest/

Visible in your browser,
but running on a remote machine

Your browser

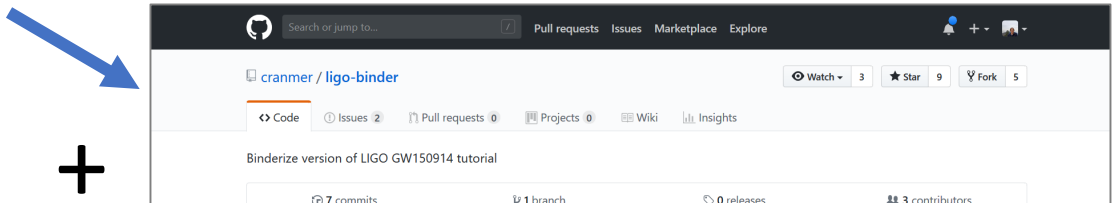


Specified version
of Python
(or other kernels eg R)

You see the code in
the .ipynb file

How does Binder work?

Code for notebook and list of dependencies and files available on GitHub (right now Binder only works with public repositories)



+

Your browser

How does Binder work?

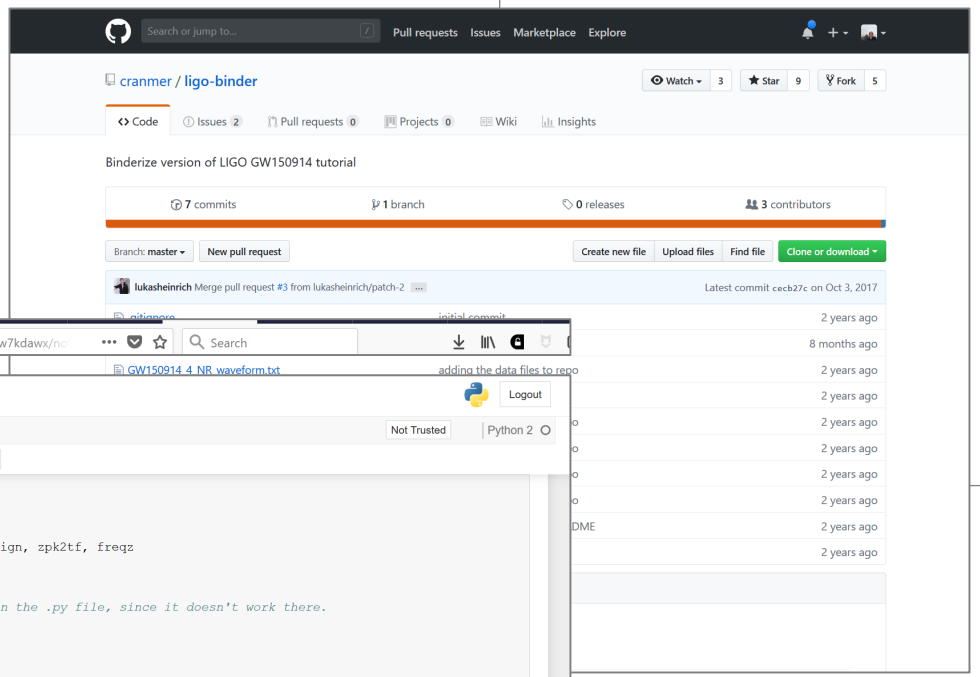
binder (beta) builds Docker image based on repo and generates URL for public access

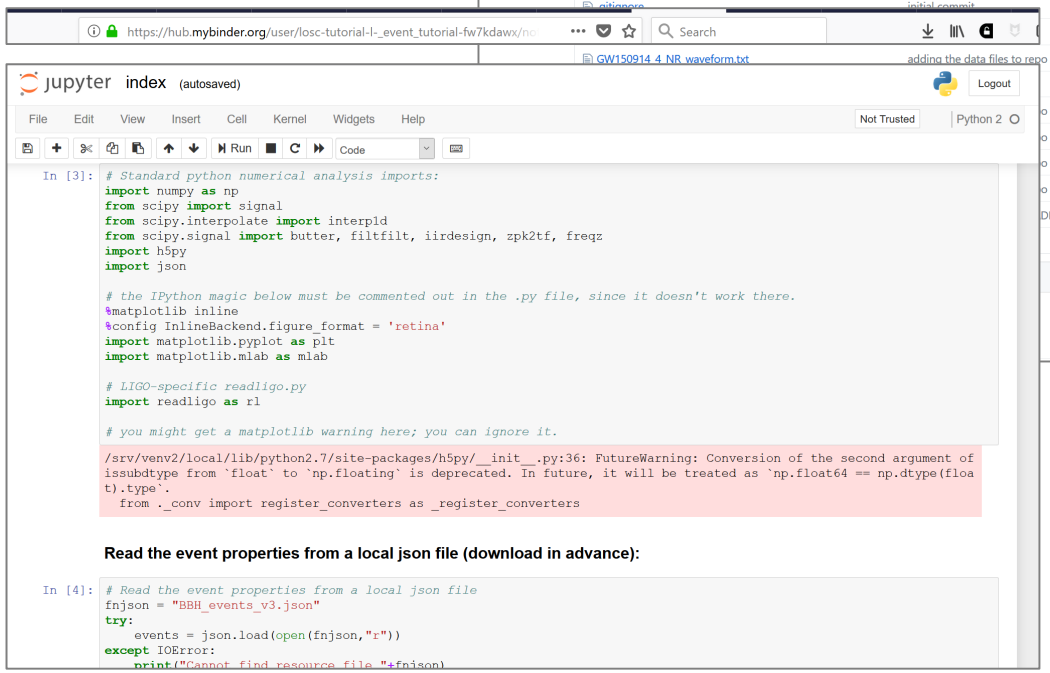
+

Code and dependencies in repo

+

Your browser





required software + packages

included data

How does Binder work?

Cloud hosting + management

+

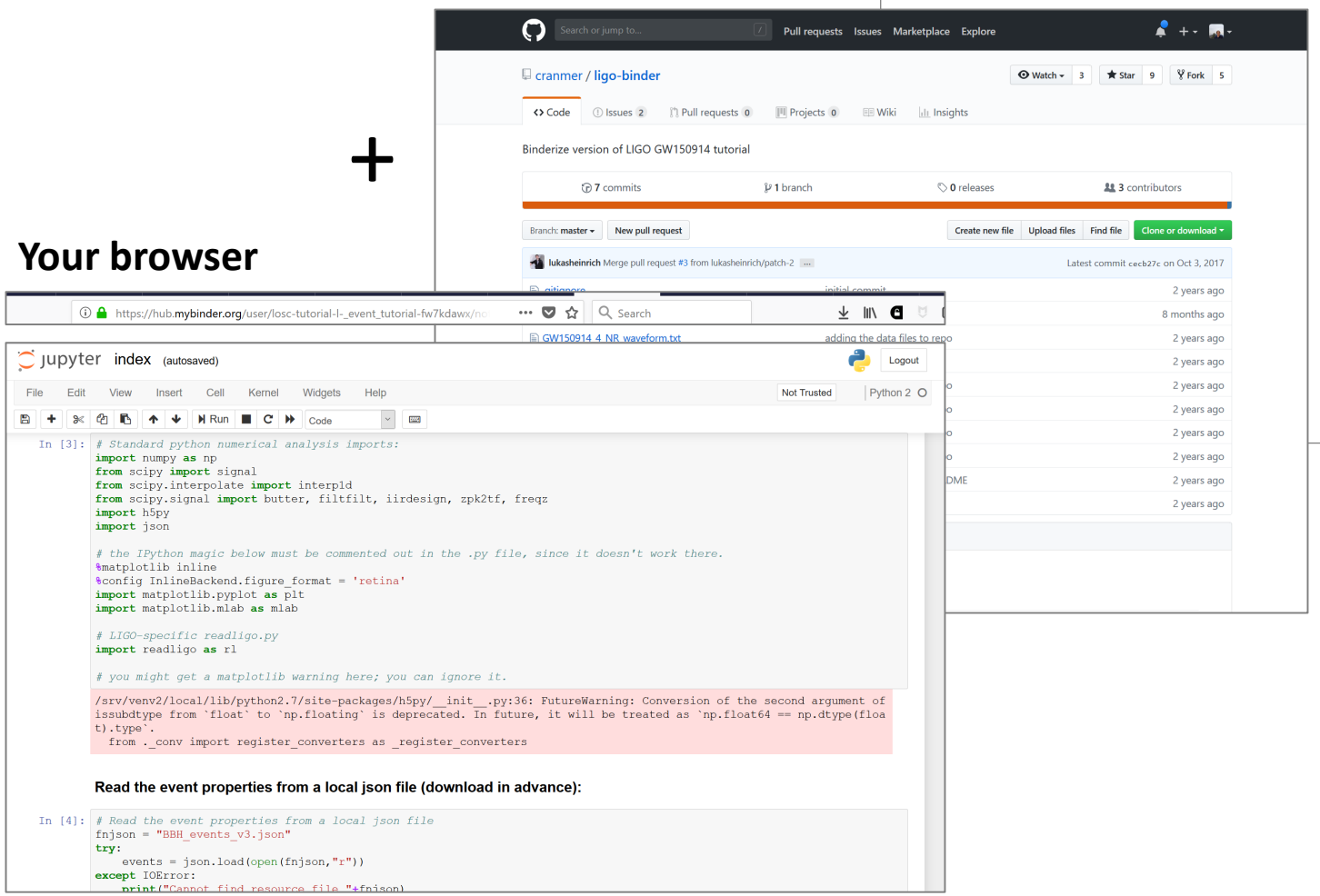
Docker image

+

Code and dependencies in repo

+

Your browser



JupyterHub

+

Kubernetes

required software + packages

included data

Generated binders are hosted by a hub that deploys, scales, and manages compute resources using Kubernetes

(can run on any cloud platform provider)
More details: BinderHub

Is this a tidy dataset?

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are `wk4`, `wk5`, ..., `wk75`.

No, has more than one observation per row

http://vita.had.co.nz/papers/tidy-da

What improved?

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6	98^0	Give Me Just One Nig...	3:24
7	A*Teens	Dancing Queen	3:44
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
1	2000-04-01	94
1	2000-04-08	99
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66