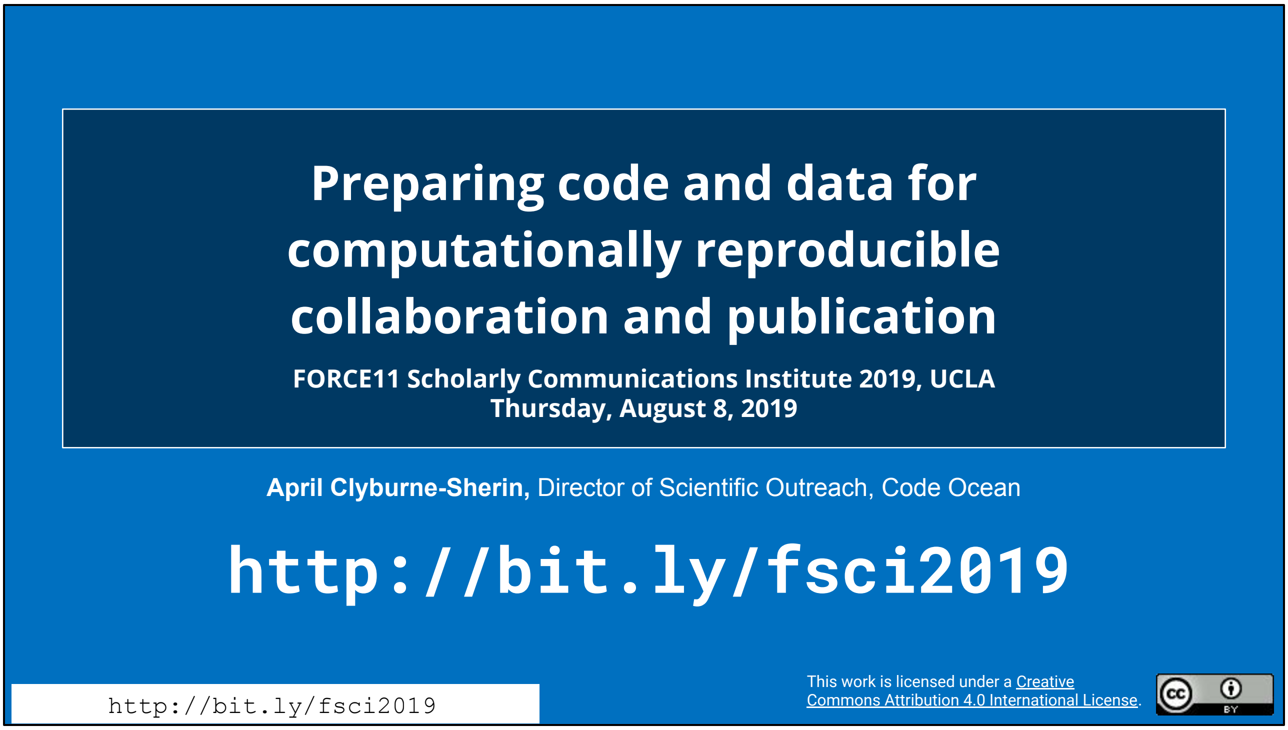


Preparing code and data for computationally reproducible collaboration and publication

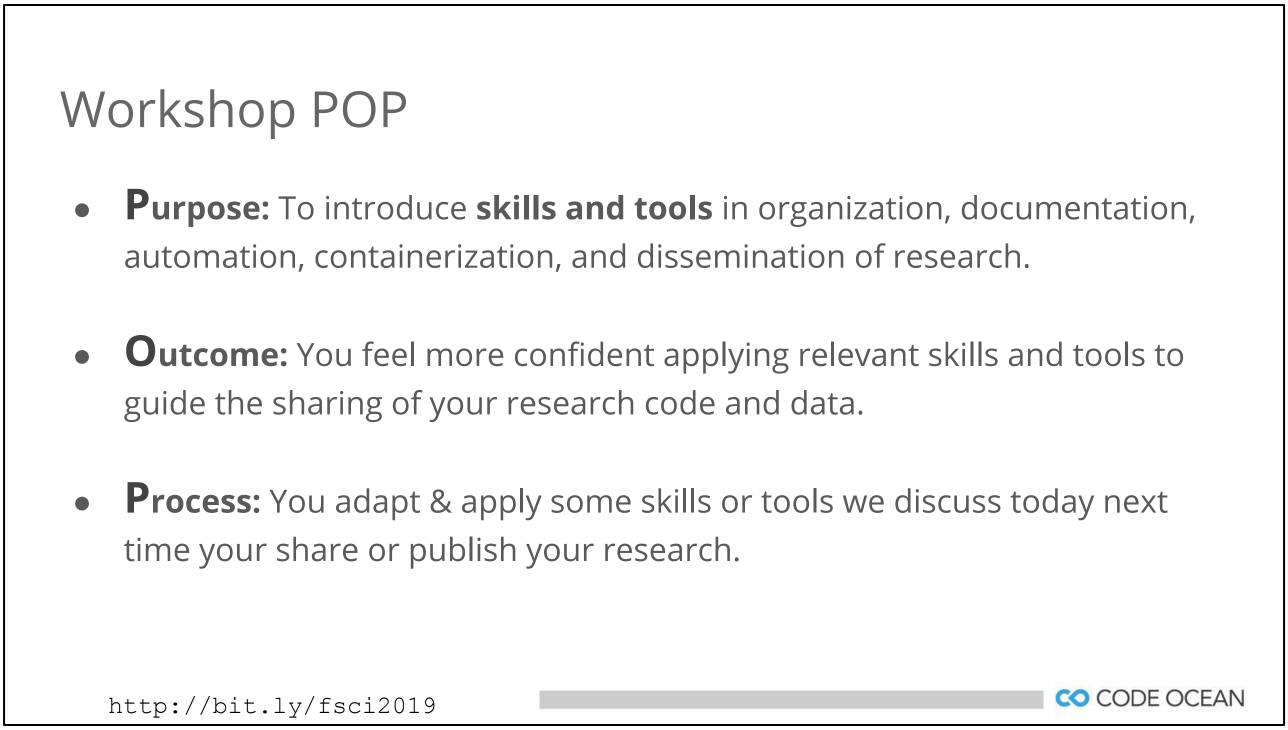
FORCE11 Scholarly Communications Institute 2019, UCLA
Thursday, August 8, 2019

April Clyburne-Sherin, Director of Scientific Outreach, Code Ocean

http://bit.ly/fsci2019

http://bit.ly/fsci2019

- Participants should have:
  - A laptop or other computer.
  - A supported browser (not IE or Edge).
  - 1 x pink post-it note.
  - 1 x green post-it note.
  - 1 x workshop survey.
  - A handful of mixed candy.
- Participants should follow along with the slide deck at the bitly link.

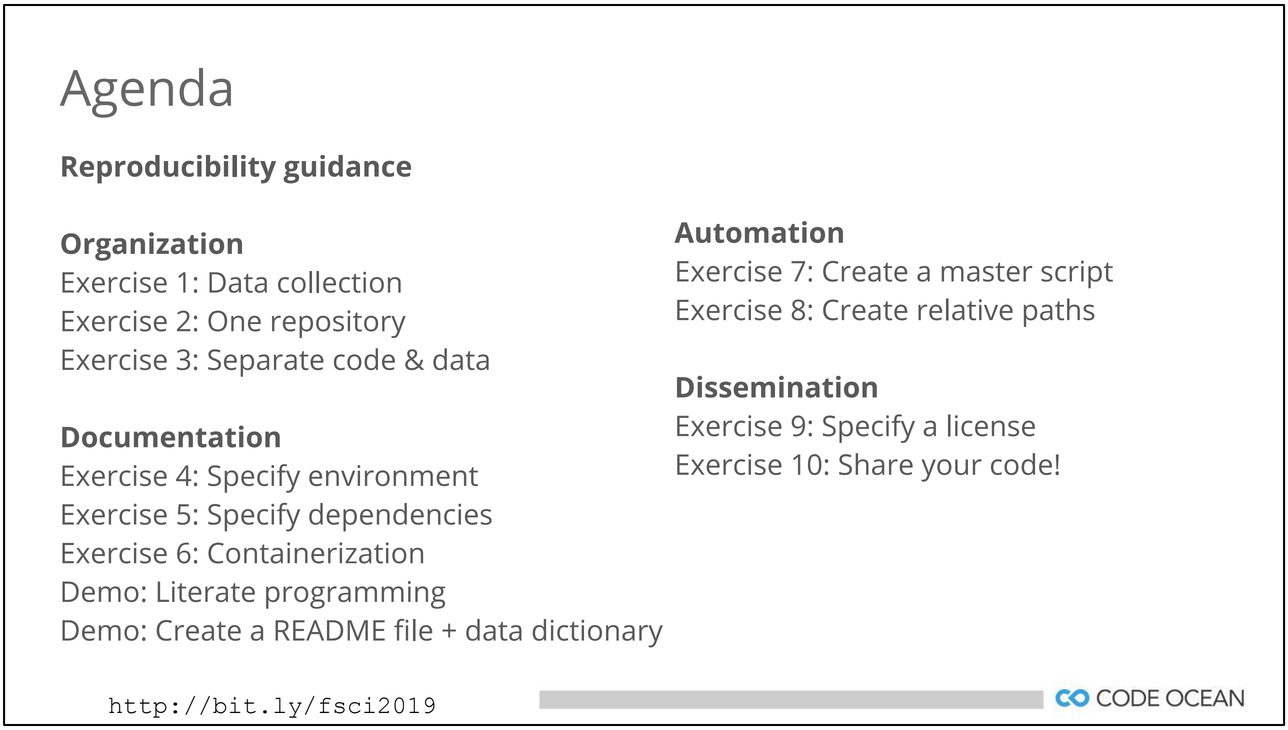


Workshop POP

- **Purpose:** To introduce **skills and tools** in organization, documentation, automation, containerization, and dissemination of research.
- **Outcome:** You feel more confident applying relevant skills and tools to guide the sharing of your research code and data.
- **Process:** You adapt & apply some skills or tools we discuss today next time your share or publish your research.

http://bit.ly/fsci2019

- Why are we here?
- What do we hope to accomplish together?



Agenda

Reproducibility guidance

Organization
Exercise 1: Data collection
Exercise 2: One repository
Exercise 3: Separate code & data

Documentation
Exercise 4: Specify environment
Exercise 5: Specify dependencies
Exercise 6: Containerization
Demo: Literate programming
Demo: Create a README file + data dictionary

Automation
Exercise 7: Create a master script
Exercise 8: Create relative paths

Dissemination
Exercise 9: Specify a license
Exercise 10: Share your code!

http://bit.ly/fsci2019

CODE OCEAN

- How will we try to accomplish our Workshop POP?
- This is an adaptable agenda:
    - We will take two breaks together.
    - Participants should feel free to take breaks whenever they wish.

Icebreaker

3:00

http://bit.ly/fsci2019

CODE OCEAN

Participants are asked to stand in a line ordered from **most recently coded** to **never coded**:

- How long has it been since you last touched research code?
- Participants must discuss with each other when they last touched research code.

Once the line is created:

- Facilitator makes these points:
  - There is a diversity of coding experience in the room - take a look who is near you who has less coding experience than you.
  - There are too many people in the room for one facilitator to troubleshoot alone, and there are lots of skilled people in the room.
  - We will use the post-it notes to signal when we have completed an exercise.
  - When you finish an exercise and switch your post-it note to green, look to your neighbors with a pink sticky - see if they could use some help!

Your thoughts?

IS THERE A REPRODUCIBILITY CRISIS?

http://bit.ly/fsci2019

CODE OCEAN

With a show of hands, ask the participants whether they think there is a reproducibility crisis in their discipline:

1. Yes, a significant crisis.
2. Yes, a slight crisis.
3. Don’t know.
4. No, there is no crisis.

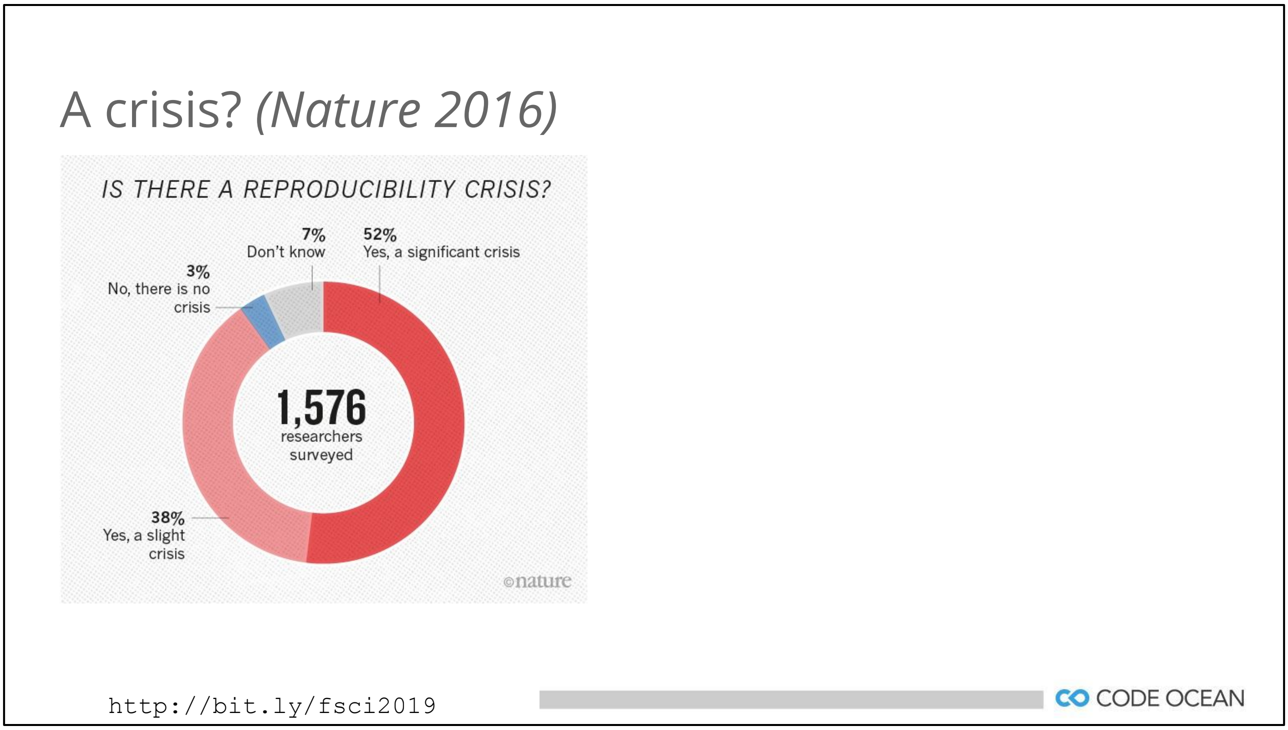
Source information:

1,500 scientists lift the lid on reproducibility

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 (2016)

“A minority of respondents reported ever having tried to publish a replication study. When work does not reproduce, researchers often assume there is a perfectly valid (and probably boring) reason. What's more, incentives to publish positive replications are low and journals can be reluctant to publish negative findings. In fact, several respondents who had published a failed replication said that editors and reviewers demanded that they play down comparisons with the original study.”

“The survey — which was e-mailed to Nature readers and advertised on affiliated websites and social-media outlets as being 'about reproducibility' — probably selected for respondents who are more receptive to and aware of concerns about reproducibility.”

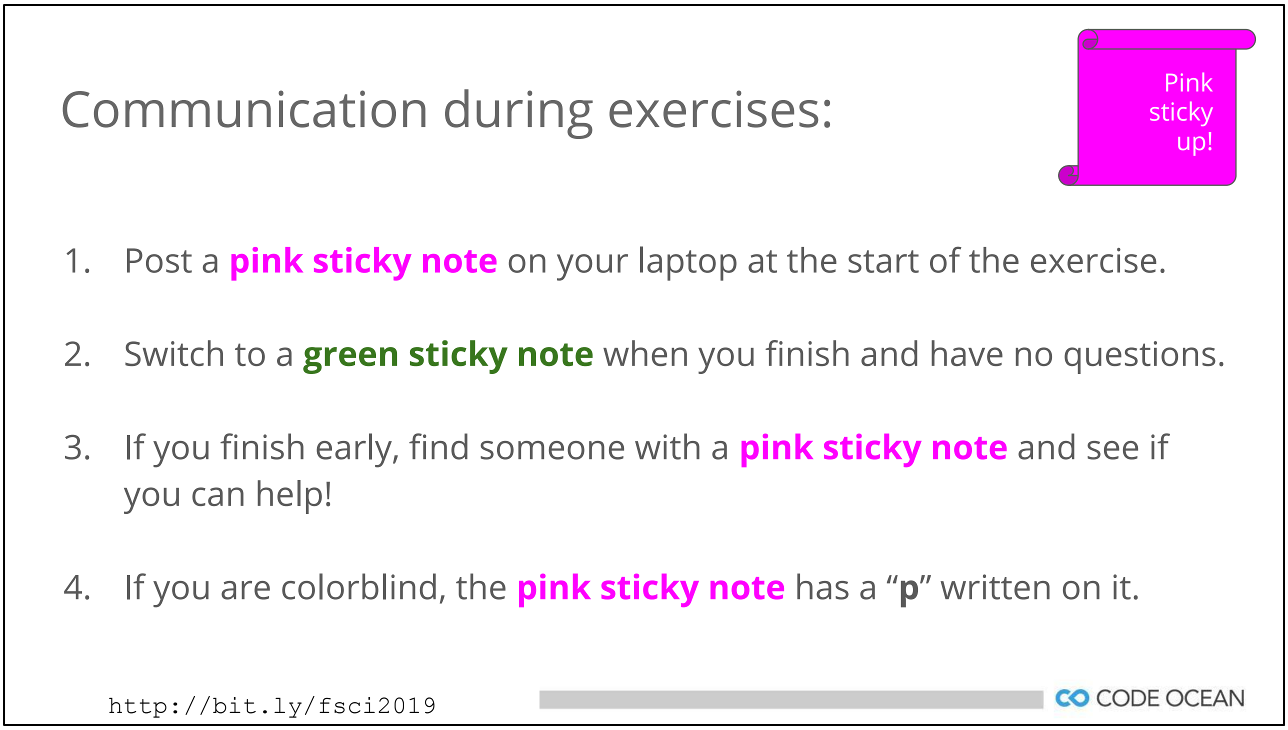


A crisis? *(Nature 2016)*

http://bit.ly/fsci2019

CODE OCEAN

- Most researchers think that reproducibility is an issue in their discipline, but may disagree about the urgency.



Communication during exercises:

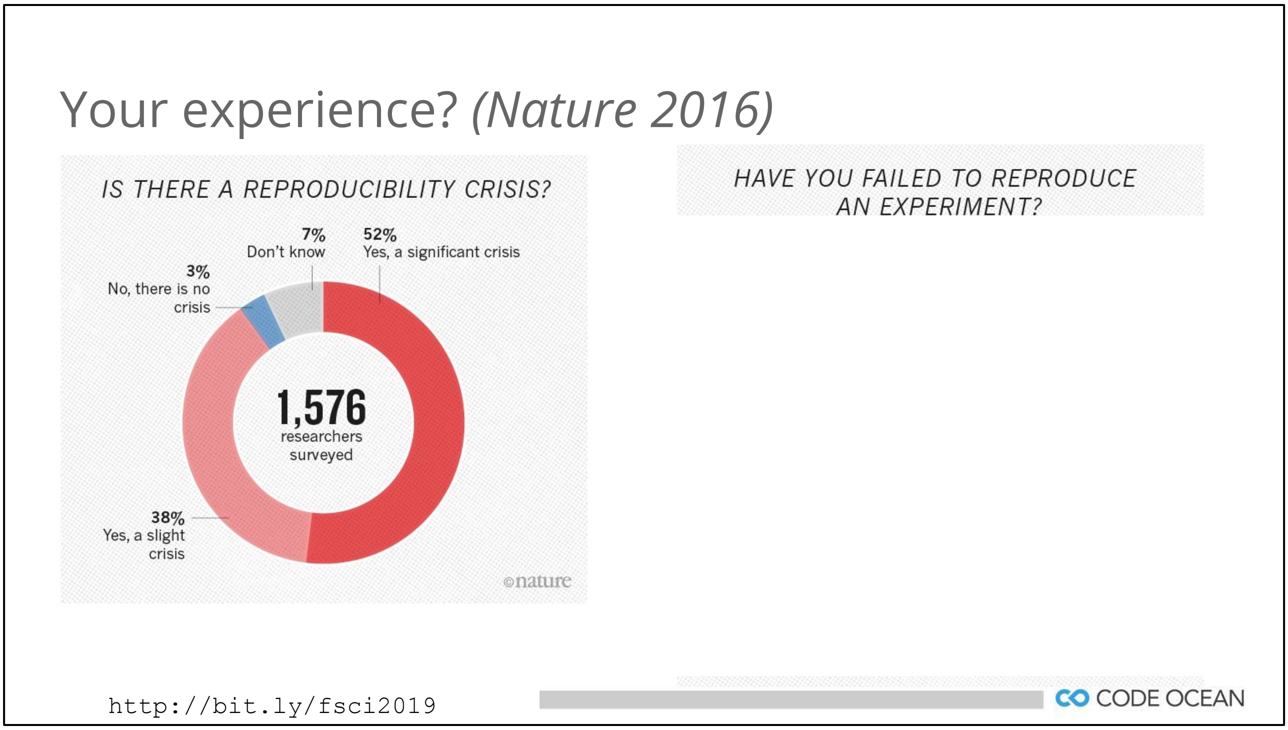
Pink sticky up!

1. Post a **pink sticky note** on your laptop at the start of the exercise.
2. Switch to a **green sticky note** when you finish and have no questions.
3. If you finish early, find someone with a **pink sticky note** and see if you can help!
4. If you are colorblind, the **pink sticky note** has a “**p**” written on it.

http://bit.ly/fsci2019

CODE OCEAN

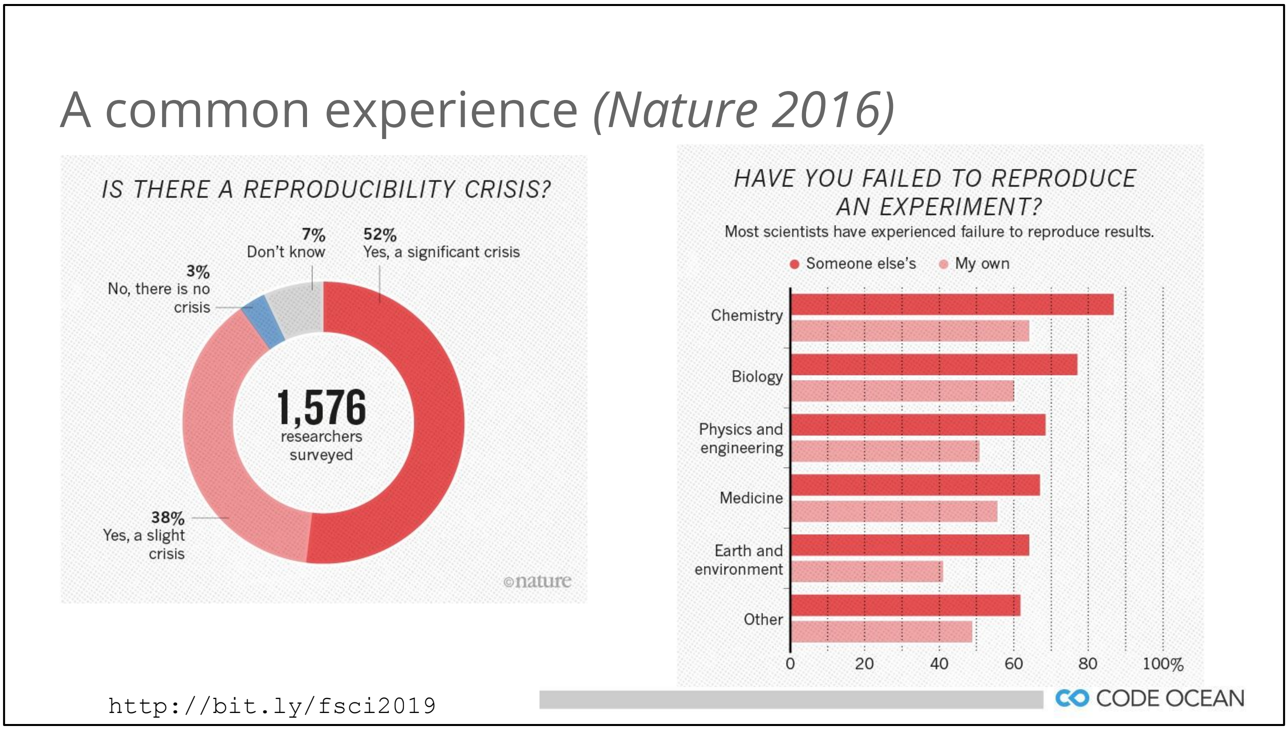
- At the beginning of each exercise, I will remind you to put a pink sticky note on your laptop.
- When you finish the exercise and have no questions, switch to a green sticky note.
- Since we are a large group and we can learn from each other, see if a neighbor who is still working is interested in some help.



Your experience? *(Nature 2016)*

http://bit.ly/fsci2019

- However, we are going to frame this question in a different way.
  - With a show of hands, how many participants have had difficulty reproducing someone else's work?
  - And how many participants have had difficulty reproducing your own work a few weeks, months, or years later?

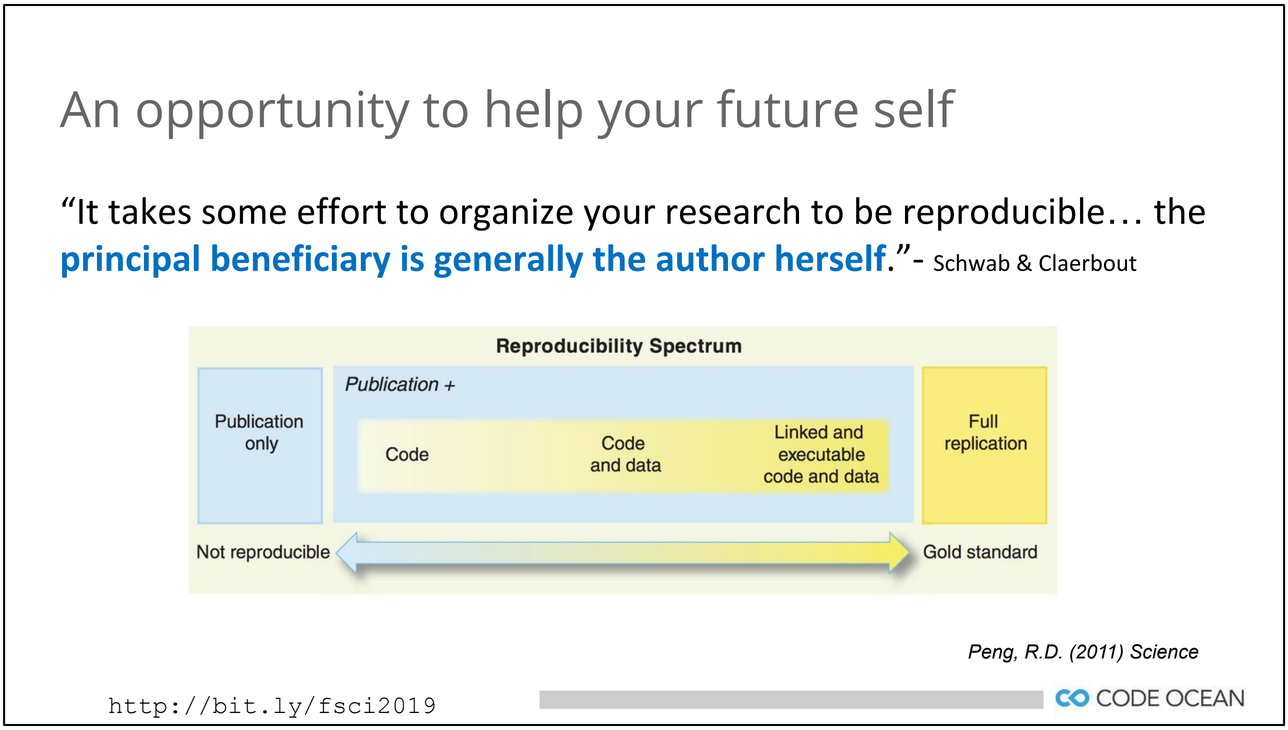


A common experience *(Nature 2016)*

`http://bit.ly/fsci2019`

CODE OCEAN

- This is how I prefer to frame issues about reproducibility.
- Difficulty reproducing our own work or the work of peers is very, very common in research.
- Many of the steps that a researcher can take to address irreproducibility of published research also improves the reusability of their research for themselves, labmates, and close collaborators.



An opportunity to help your future self

"It takes some effort to organize your research to be reproducible… the **principal beneficiary is generally the author herself**."- Schwab & Claerbout

Reproducibility Spectrum

Publication only

Publication +

Code

Code and data

Linked and executable code and data

Full replication

Not reproducible

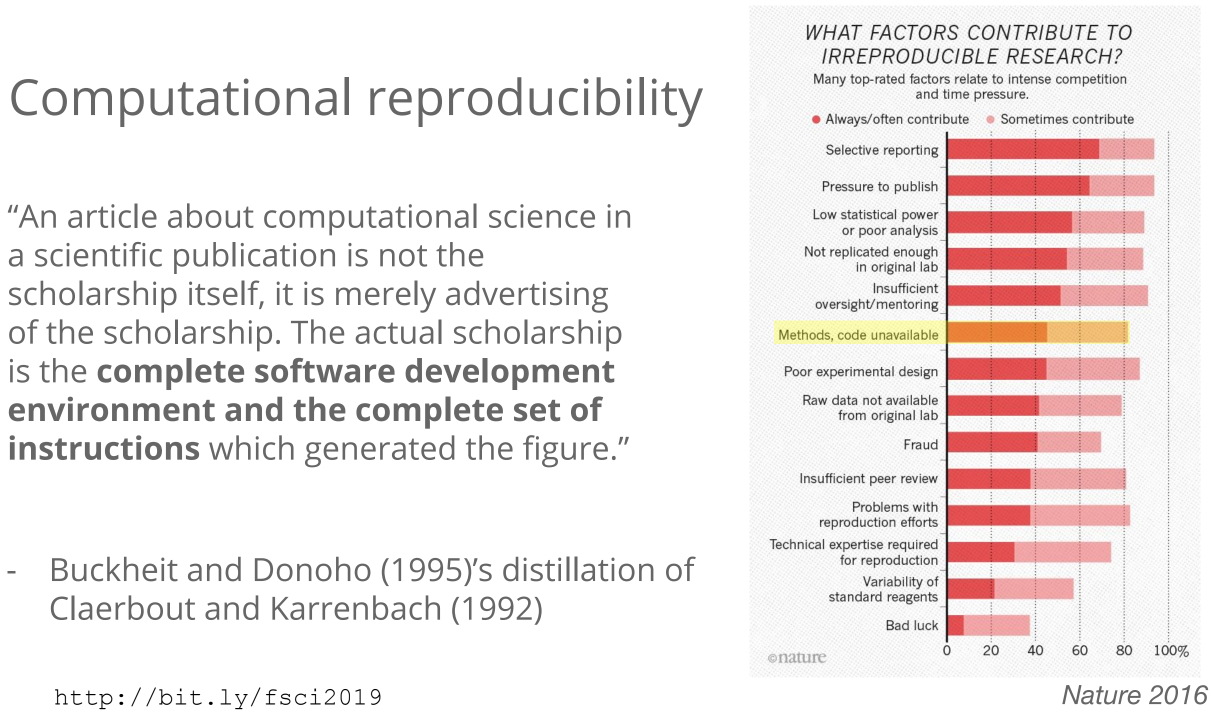
Gold standard

Peng, R.D. (2011) Science

http://bit.ly/fsci2019

CODE OCEAN

- Therefore, when thinking about adopting best practices to tackle reproducibility, adopt first those practices that will benefit yourself first - as your future self is the most frequent reuser of your research.
- Remember that, like reproducibility generally, computational reproducibility is a spectrum. Integrating one or two new practices into your research will make your research more reproducible - it is not all or nothing. And each of these steps will benefit yourself first!



Computational reproducibility

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the **complete software development environment and the complete set of instructions** which generated the figure."

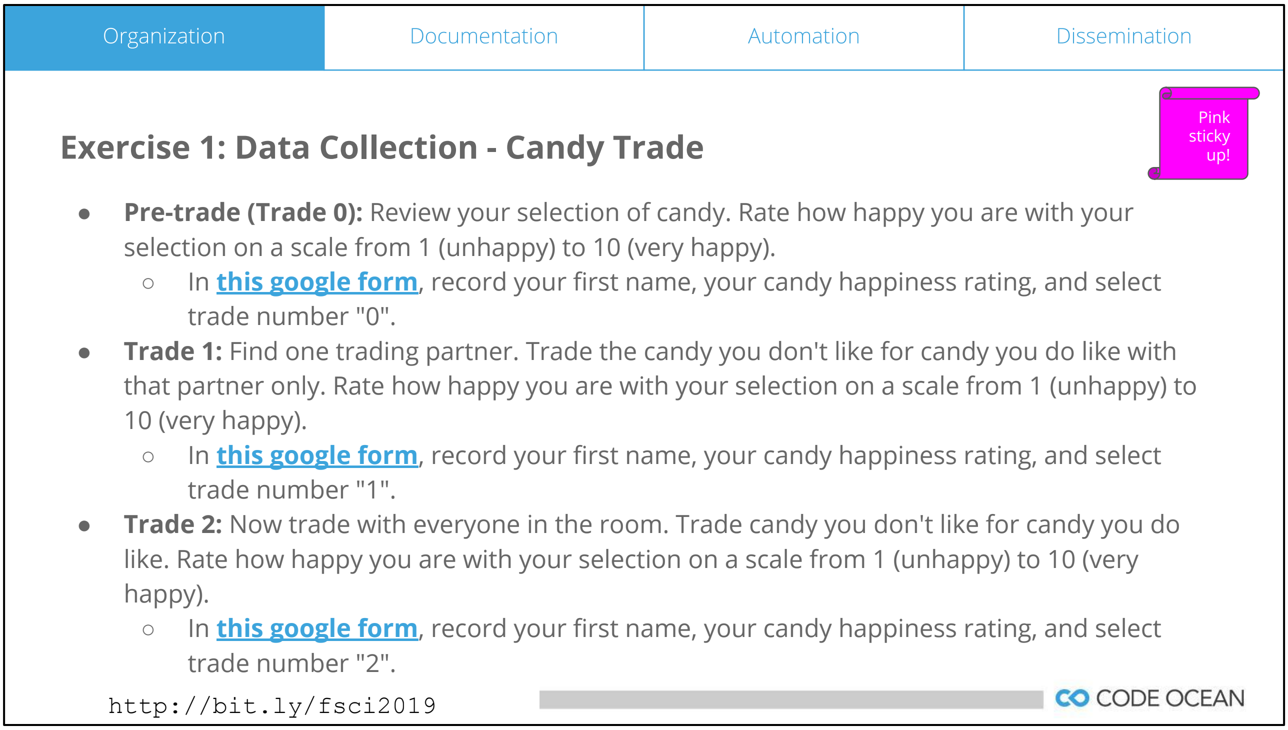
- Buckheit and Donoho (1995)'s distillation of Claerbout and Karrenbach (1992)

`http://bit.ly/fsci2019`

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
● Always/often contribute
● Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
20
40
60
80
100%
©nature

Nature 2016

- Professor Stodden et al. talked a lot about reproducibility yesterday and I just assume you were there
- For our purposes, computational reproducibility is the ability for a new researcher to rerun the data and code of the original research to create the same figures, tables, and values shared by the original researcher.
  - Almost all research relies on software and code for some aspect of its methods. Computational reproducibility relies on the sharing of that software and code in addition to a published narrative of the methods. This is why the founders of our idea of reproducibility describe a published article as just an advertising of the research. The actual research includes our research software and code.
- Unavailable code and methods is a barrier to computational reproducibility. So to improve computational reproducibility, we want to know how best to share our methods and code.
-



Organization | Documentation | Automation | Dissemination

Pink sticky up!

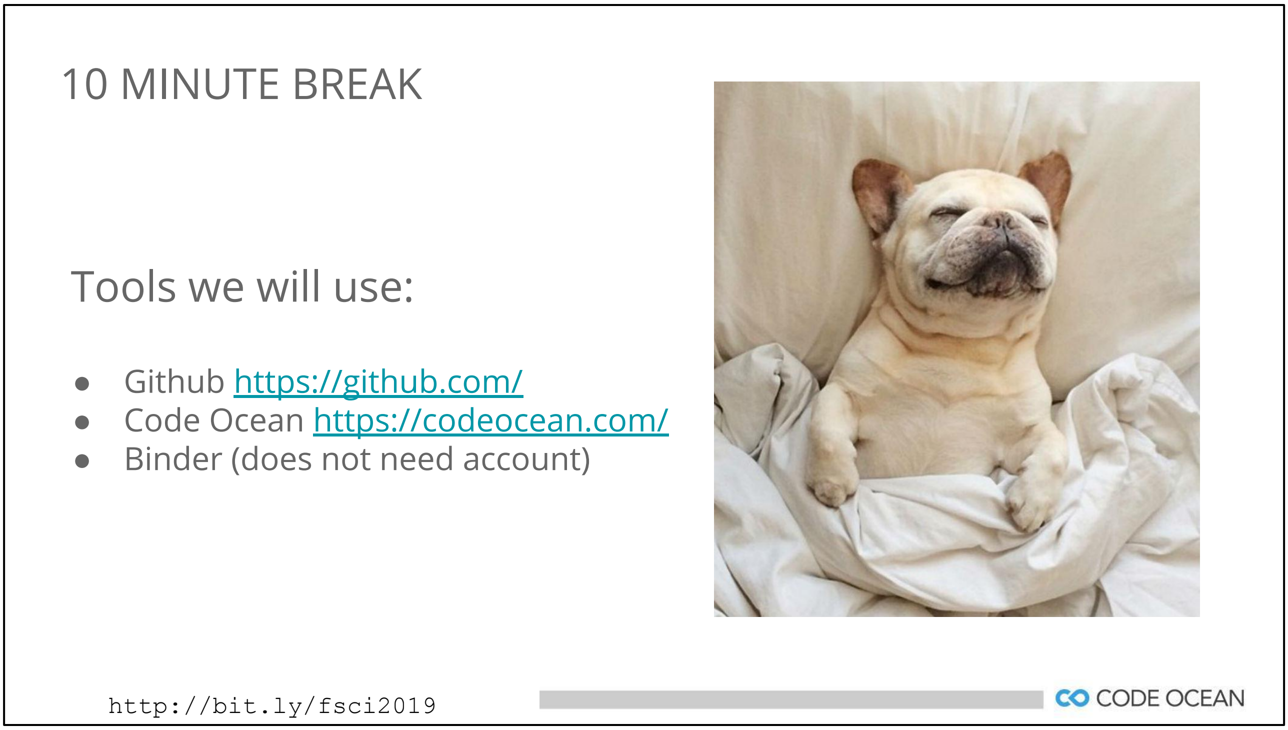
Exercise 1: Data Collection - Candy Trade

- **Pre-trade (Trade 0):** Review your selection of candy. Rate how happy you are with your selection on a scale from 1 (unhappy) to 10 (very happy).
  - In **this google form**, record your first name, your candy happiness rating, and select trade number "0".
- **Trade 1:** Find one trading partner. Trade the candy you don't like for candy you do like with that partner only. Rate how happy you are with your selection on a scale from 1 (unhappy) to 10 (very happy).
  - In **this google form**, record your first name, your candy happiness rating, and select trade number "1".
- **Trade 2:** Now trade with everyone in the room. Trade candy you don't like for candy you do like. Rate how happy you are with your selection on a scale from 1 (unhappy) to 10 (very happy).
  - In **this google form**, record your first name, your candy happiness rating, and select trade number "2".

http://bit.ly/fsci2019

CODE OCEAN

- Because some participants did not bring their own data and code, we will run a short data collection exercise.
- This exercise is followed by a 10 minute break.
- Instructor: During the break:
  - Download the responses: https://docs.google.com/forms/d/13_i1jKlyTwhl0UaUkmqhfZuG-M9i4I5_6Q81vz54Ypw/edit#responses.
  - Remove the “Timestamp” column.
  - Save as a CSV file.
  - Upload to CSV file, named “data.csv”, to this github repo: https://github.com/aprilcs/candy_trade. Be sure to delete the existing “data.csv” file if there already is one in the repo.



10 MINUTE BREAK

Tools we will use:

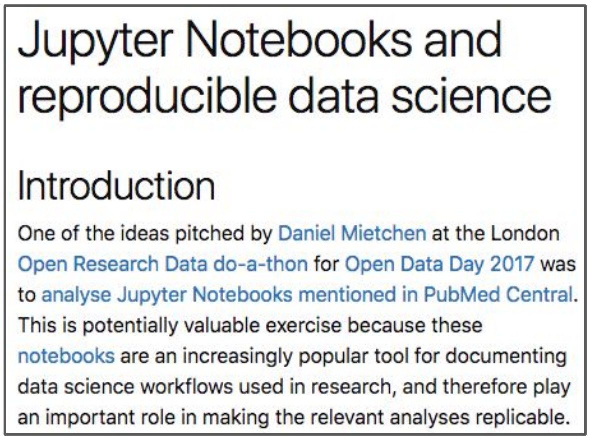
- Github https://github.com/
- Code Ocean https://codeocean.com/
- Binder (does not need account)

http://bit.ly/fsci2019

- Instructor: During the break:
    - Download the responses: https://docs.google.com/forms/d/13_i1jKlyTwhl0UaUkmqhfZuG-M9i4l5_6Q81vz54Ypw/edit#responses.
    - Remove the “Timestamp” column.
    - Save as a CSV file.
    - Upload to CSV file, named “data.csv”, to this github repo: https://github.com/aprilcs/candy_trade. Be sure to delete the existing “data.csv” file if there already is one in the repo.

Lessons learned: testing computational reproducibility

- PMC “jupyter OR ipynb” -> 107 papers
- “My initial thought was that analysing the validity of the notebooks would simply involve searching the text of each article for a notebook reference, then downloading and executing it … **It turned out that this was hopelessly naive**…”



Jupyter Notebooks and reproducible data science

Introduction

One of the ideas pitched by Daniel Mietchen at the London Open Research Data do-a-thon for Open Data Day 2017 was to analyse Jupyter Notebooks mentioned in PubMed Central. This is potentially valuable exercise because these notebooks are an increasingly popular tool for documenting data science workflows used in research, and therefore play an important role in making the relevant analyses replicable.

Mark Woodbridge, Daniel Sanz, Daniel Mietchen, & Ross Mounce (2017). Jupyter Notebooks and reproducible data science, https://markwoodbridge.com/2017/03/05/jupyter-reproducible-science.html.

http://bit.ly/fsci2019

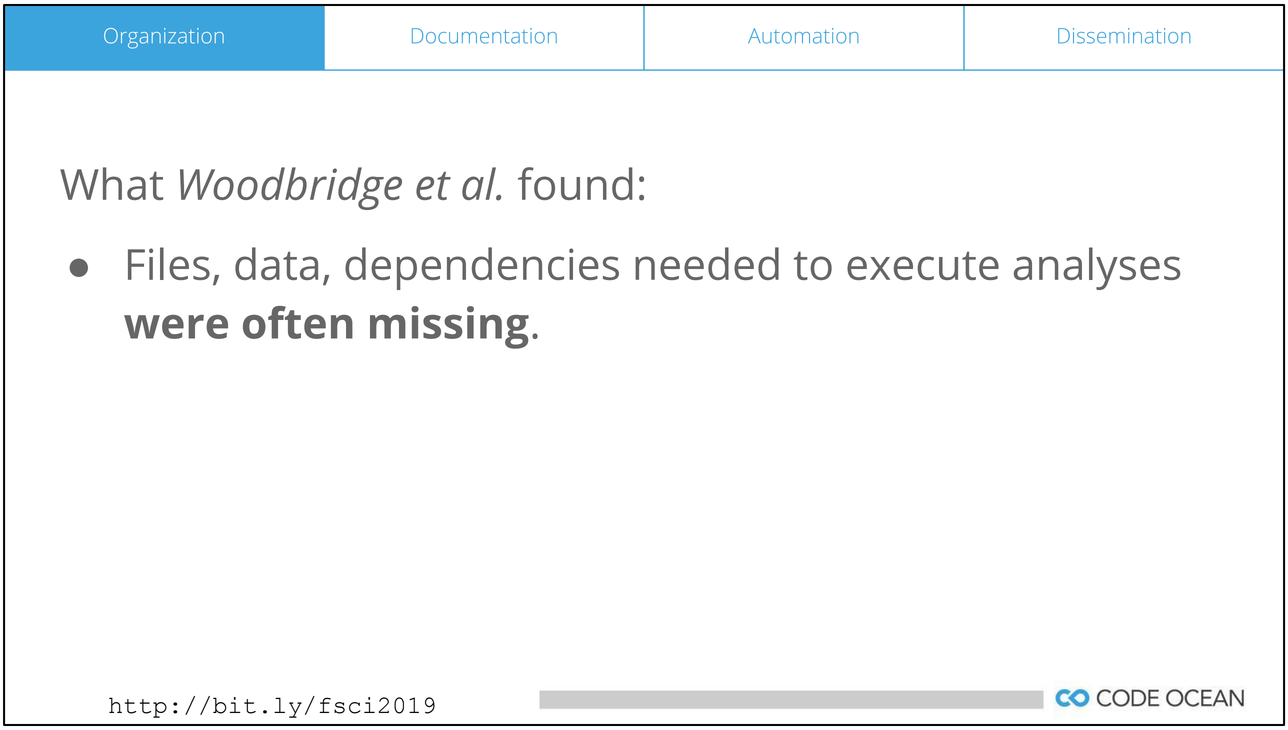
CODE OCEAN

- We mentioned before the break that sharing code, data, and methods is necessary for computational reproducibility.
- However, sharing just the code and methods does not ensure reproducibility. This was demonstrated by an informal study by Woodbridge et al. to reproduce code from published papers in the form of Jupyter Notebooks.
- They were able to successfully execute only one of the ~25 notebooks that we downloaded.
- We don’t have to be naive like them - we can learn from their attempt.

If people do not know what a jupyter notebook is, you can talk about the uses of notebooks:

- Documentation of analyses
  - A Modularized Efficient Framework for Non-Markov Time Series Estimation: https://codeocean.com/2018/01/16/a-modularized-efficient-framework-for-non-markov-time-series-estimation/code
- Programming or statistical education
  - Fractal Generation with L-Systems: https://codeocean.com/2017/12/08/fractal-generation-with-l-systems/code

- Executable article
    - On Writing Reproducible and Interactive Papers: https://codeocean.com/2018/06/28/on-writing-reproducible-and-interactive-papers/code



Organization | Documentation | Automation | Dissemination

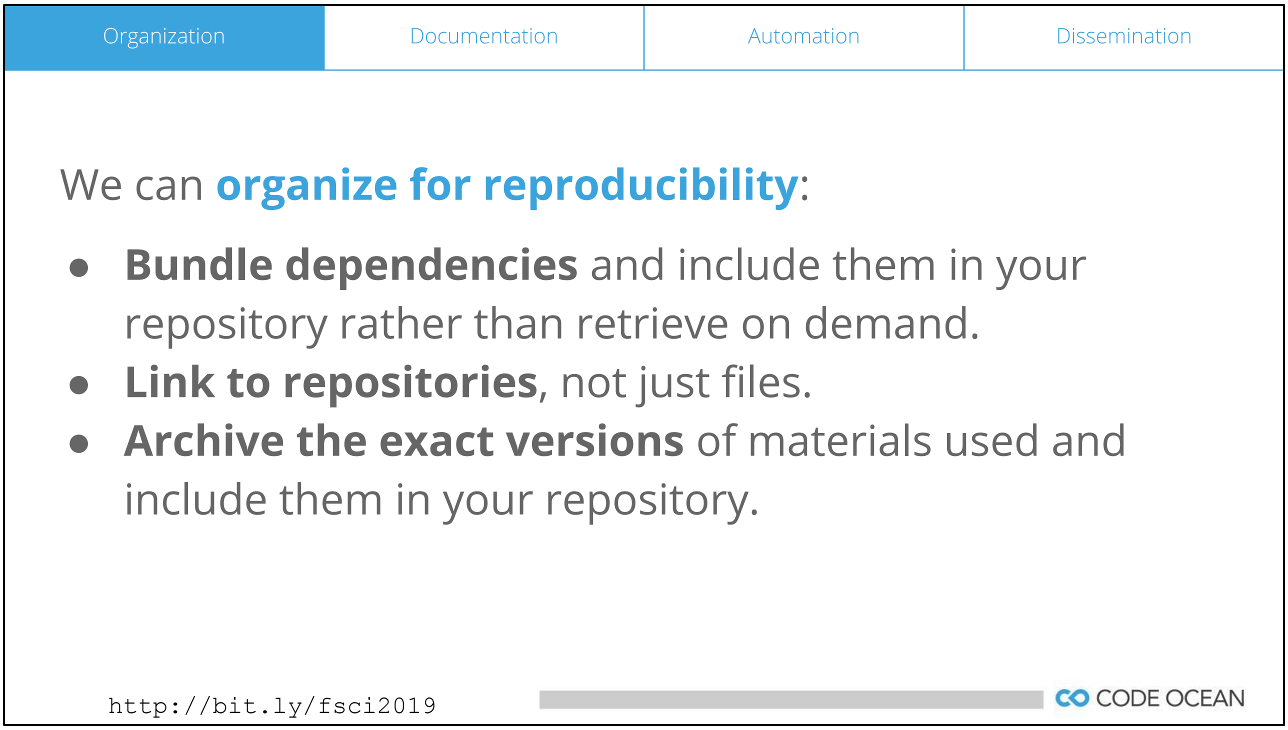
What *Woodbridge et al.* found:

- Files, data, dependencies needed to execute analyses **were often missing**.

http://bit.ly/fsci2019

CODE OCEAN

- The first thing that Mark Woodbridge and his colleagues learned was that the files, data, and dependencies needed to execute analyses were often missing from the publication or accompanying repository.



Organization
Documentation
Automation
Dissemination

We can **organize for reproducibility**:

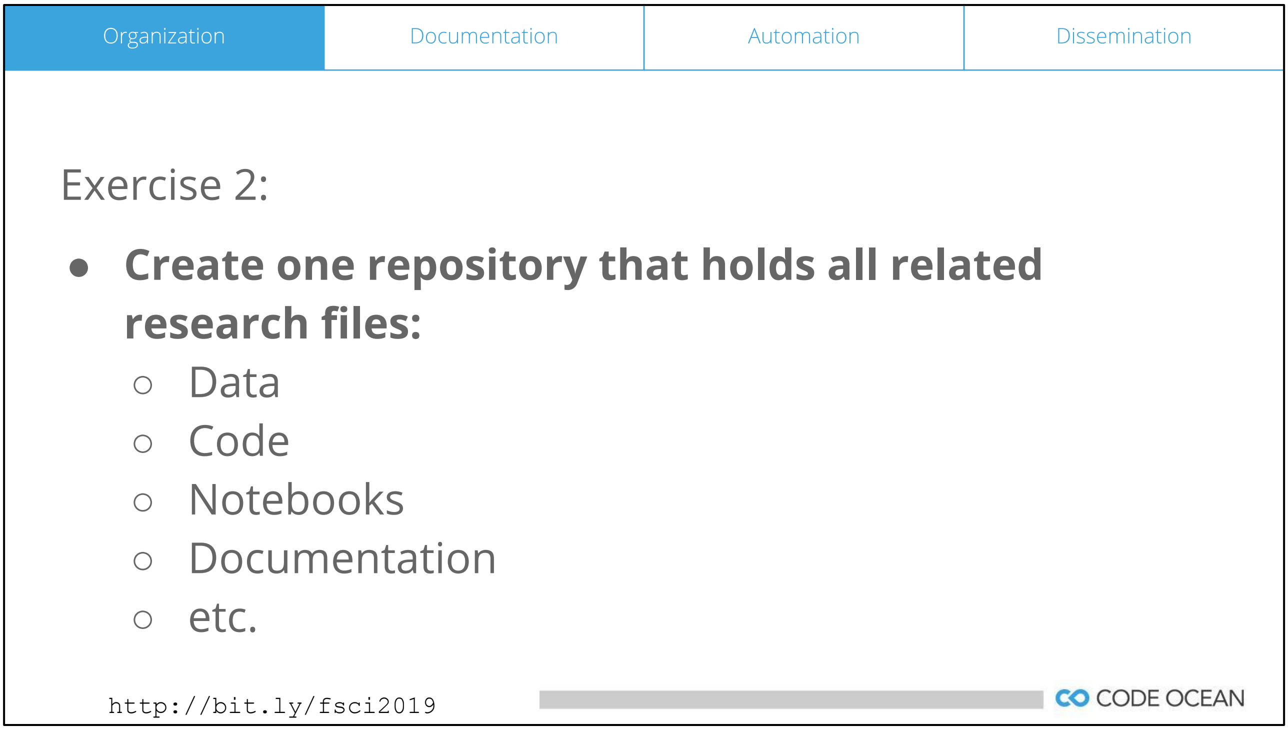
- **Bundle dependencies** and include them in your repository rather than retrieve on demand.
- **Link to repositories**, not just files.
- **Archive the exact versions** of materials used and include them in your repository.

http://bit.ly/fsci2019

CODE OCEAN

Learning from Woodbridge’s finding, we can organize for reproducibility:

- Archive the exact versions of data used and include them in your repository.
- Bundle dependencies and include them in your repository rather than retrieve on demand.
- Link to repositories rather than individual code files or data files.



Organization | Documentation | Automation | Dissemination

Exercise 2:

- **Create one repository that holds all related research files:**
  - Data
  - Code
  - Notebooks
  - Documentation
  - etc.

http://bit.ly/fsci2019

CODE OCEAN

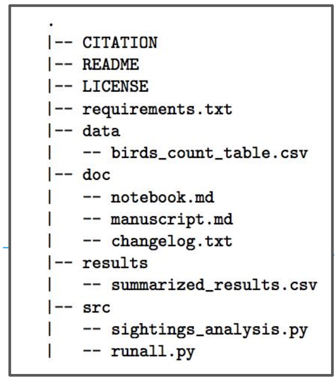
So, the first way to overcome the risk of missing files, data, or dependencies that Woodbridge found is to put everything in one repository.

Therefore, exercise 2 will be to create one repository for all our research materials. We will use Code Ocean in this workshop, but you could do this locally with a file folder or using a git repository.

Pink sticky up!

Exercise 3:

- **Organize your research to separate code from data.**



```
.
|-- CITATION
|-- README
|-- LICENSE
|-- requirements.txt
|-- data
|     -- birds_count_table.csv
|-- doc
|     -- notebook.md
|     -- manuscript.md
|     -- changelog.txt
|-- results
|     -- summarized_results.csv
|-- src
|     -- sightings_analysis.py
|     -- runall.py
```

Resource on reproducible organization:

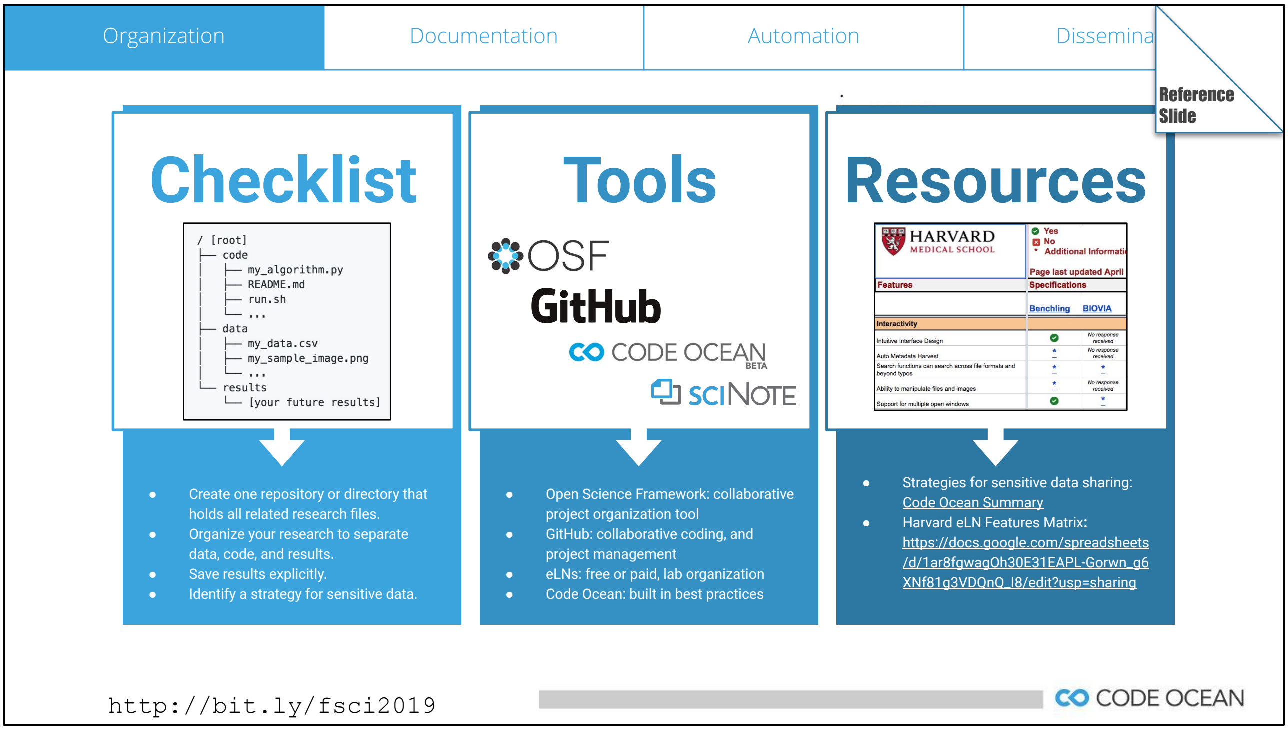
- Karl Broman: http://kbroman.org/steps2rr/pages/organize.html

Join our Candy Swap project

Pink
sticky
up!

R: https://github.com/aprilcs/sips-workshop

Python: https://github.com/aprilcs/sips-workshop-py


OSF
GitHub
CODE OCEAN BETA
sciNOTE
CODE OCEAN

Organization | Documentation | Automation | Dissemina

Reference Slide

Checklist

```
/ [root]
├── code
│   ├── my_algorithm.py
│   ├── README.md
│   ├── run.sh
│   └── ...
├── data
│   ├── my_data.csv
│   ├── my_sample_image.png
│   └── ...
└── results
    └── [your future results]
```

- Create one repository or directory that holds all related research files.
- Organize your research to separate data, code, and results.
- Save results explicitly.
- Identify a strategy for sensitive data.

Tools

- Open Science Framework: collaborative project organization tool
- GitHub: collaborative coding, and project management
- eLNs: free or paid, lab organization
- Code Ocean: built in best practices

Resources

HARVARD MEDICAL SCHOOL

✓ Yes
☒ No
* Additional Informati

Page last updated April

Features	Specifications	
	Benchling	BIOVIA
Interactivity		
Intuitive Interface Design	✓	No response received
Auto Metadata Harvest	*	No response received
Search functions can search across file formats and beyond typos	*	*
Ability to manipulate files and images	*	No response received
Support for multiple open windows	✓	*

- Strategies for sensitive data sharing: Code Ocean Summary
- Harvard eLN Features Matrix: https://docs.google.com/spreadsheets/d/1ar8fgwagOh30E31EAPL-Gorwn_g6XNf81g3VDQnQ_I8/edit?usp=sharing

http://bit.ly/fsci2019

How do we organize for reproducibility?

- For your reference, we have created these reference slides within the slide deck with key points, tools, and resources.
- We won’t spend time on these reference slides as they are intended for your future reference, and will just go over the main points.

Reproducible organization includes:

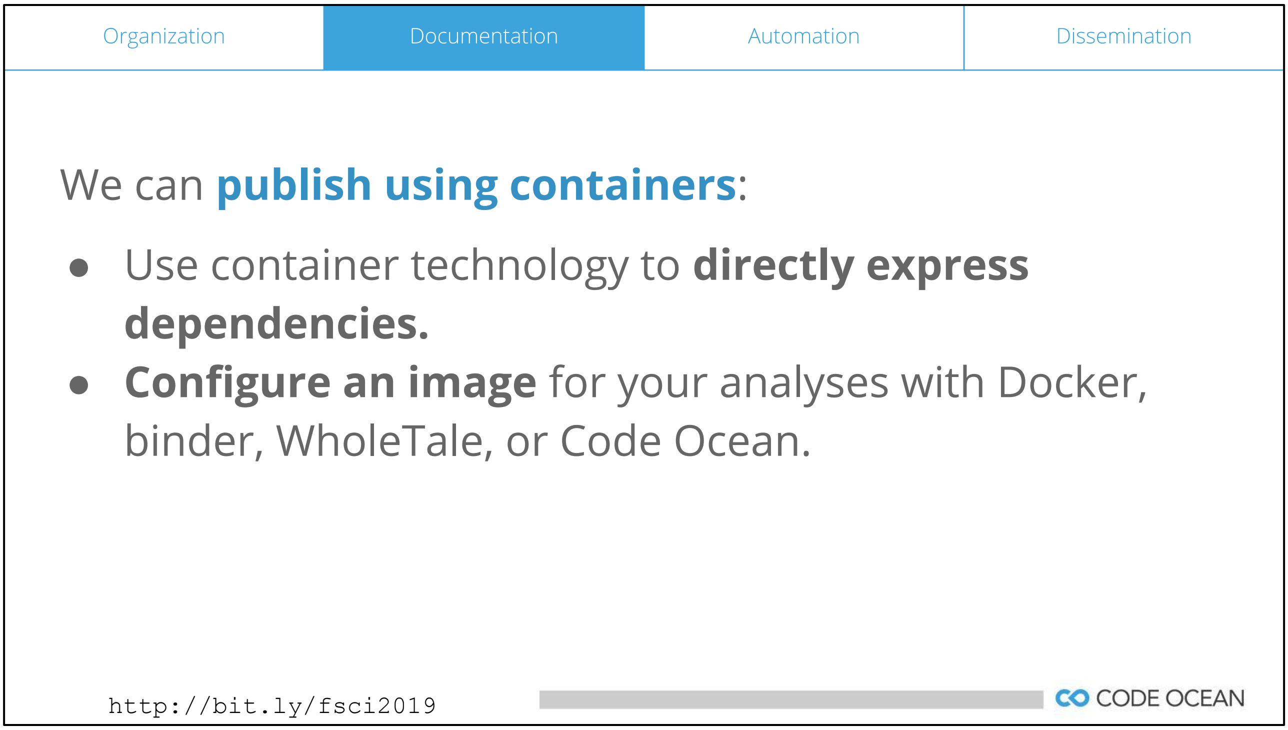
- Creating one repository for all your research materials.
- Separating data and code.
- And saving your results explicitly as a function of your data and code.

Free tools are available to help organize your research materials:

- The OSF is a free, open source, collaboration tool.
- Github provides free public repositories for collaborative coding and project management.
- Electronic lab notebooks help to organize a lab, and Harvard’s Features Matrix is a great place to start comparing them.
- Code Ocean structures their repository to separate data and code and save results explicitly.

What *Woodbridge et al.* found:

- There is no way to **directly express dependencies** of published code.



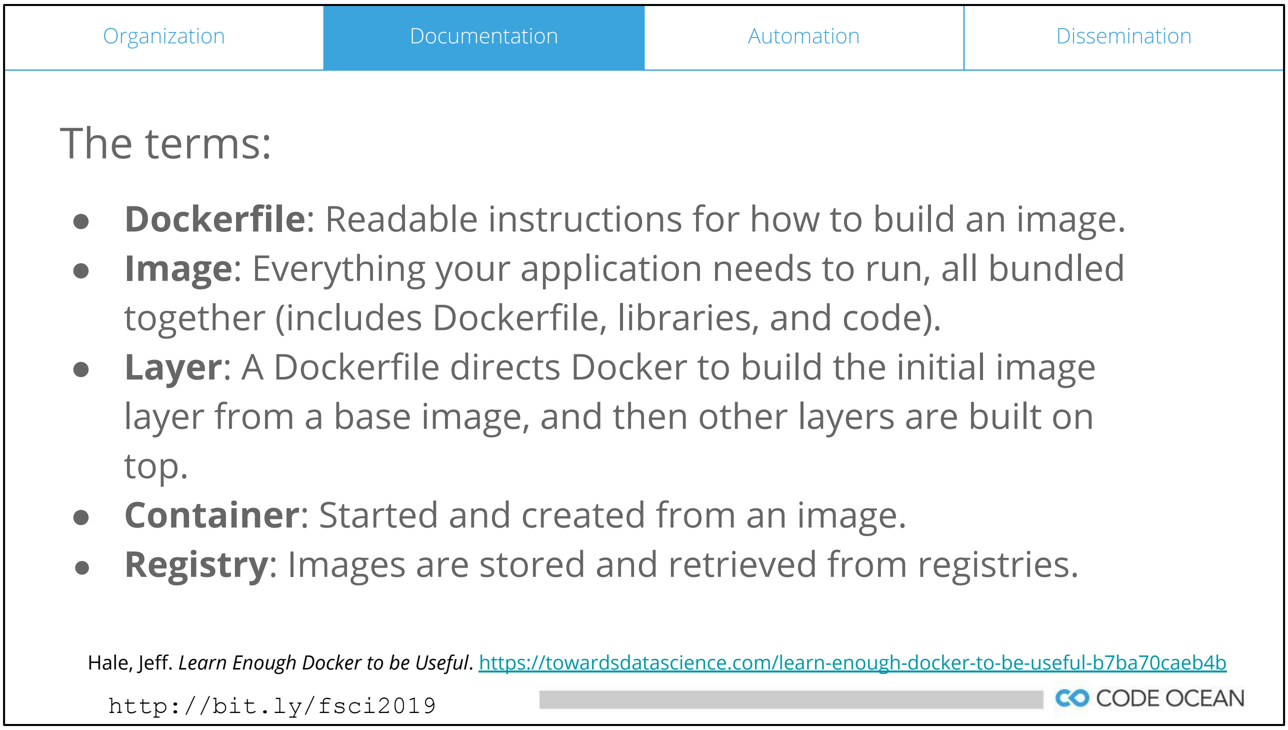
Organization | Documentation | Automation | Dissemination

We can **publish using containers**:

- Use container technology to **directly express dependencies.**
- **Configure an image** for your analyses with Docker, binder, WholeTale, or Code Ocean.

http://bit.ly/fsci2019

CODE OCEAN

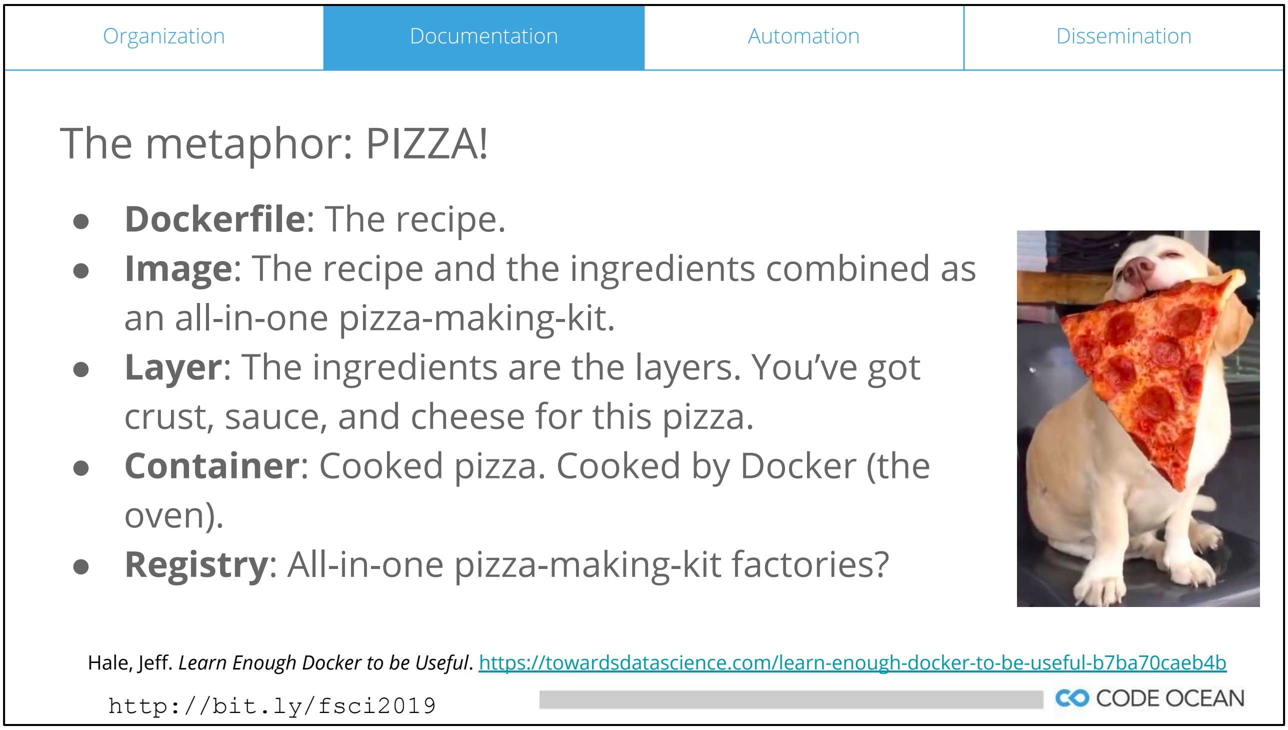


Organization | Documentation | Automation | Dissemination

The terms:

- **Dockerfile**: Readable instructions for how to build an image.
- **Image**: Everything your application needs to run, all bundled together (includes Dockerfile, libraries, and code).
- **Layer**: A Dockerfile directs Docker to build the initial image layer from a base image, and then other layers are built on top.
- **Container**: Started and created from an image.
- **Registry**: Images are stored and retrieved from registries.

Hale, Jeff. *Learn Enough Docker to be Useful*. https://towardsdatascience.com/learn-enough-docker-to-be-useful-b7ba70caeb4b



Organization | **Documentation** | Automation | Dissemination

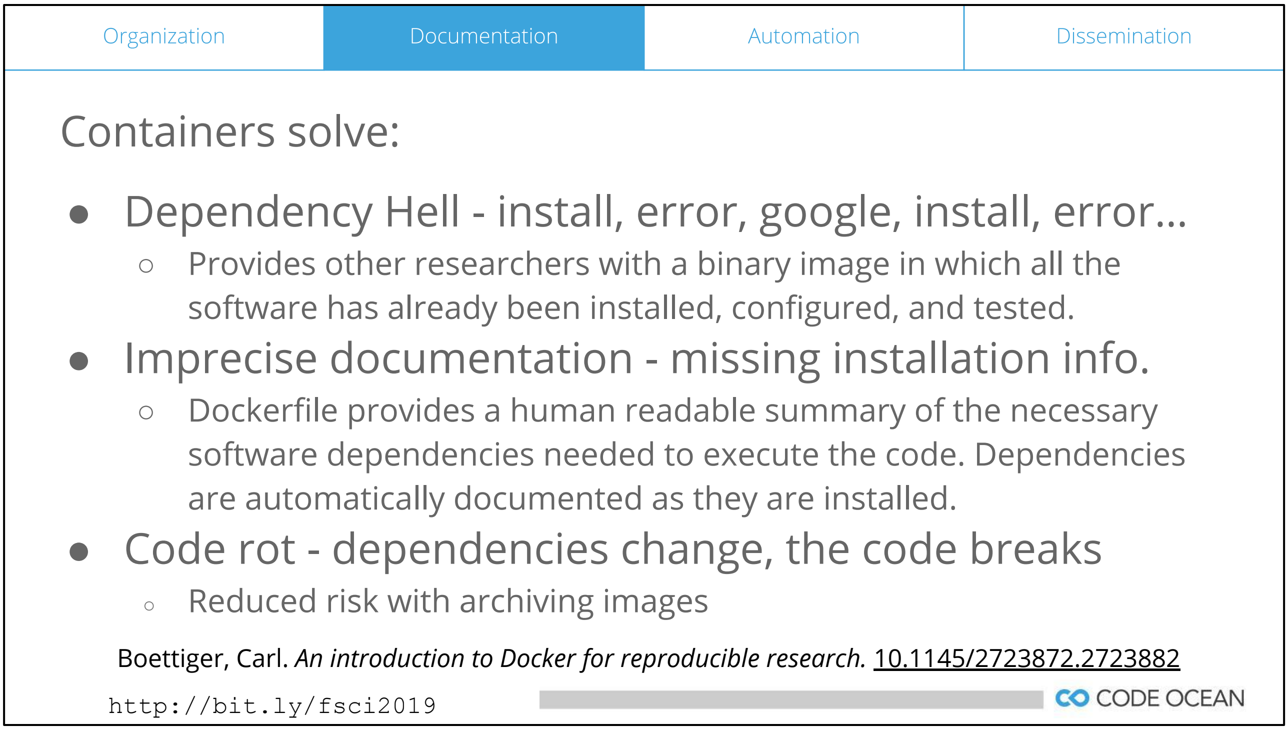
The metaphor: PIZZA!

- **Dockerfile**: The recipe.
- **Image**: The recipe and the ingredients combined as an all-in-one pizza-making-kit.
- **Layer**: The ingredients are the layers. You've got crust, sauce, and cheese for this pizza.
- **Container**: Cooked pizza. Cooked by Docker (the oven).
- **Registry**: All-in-one pizza-making-kit factories?

Hale, Jeff. *Learn Enough Docker to be Useful*. https://towardsdatascience.com/learn-enough-docker-to-be-useful-b7ba70caeb4b

http://bit.ly/fsci2019

CODE OCEAN



Organization | Documentation | Automation | Dissemination

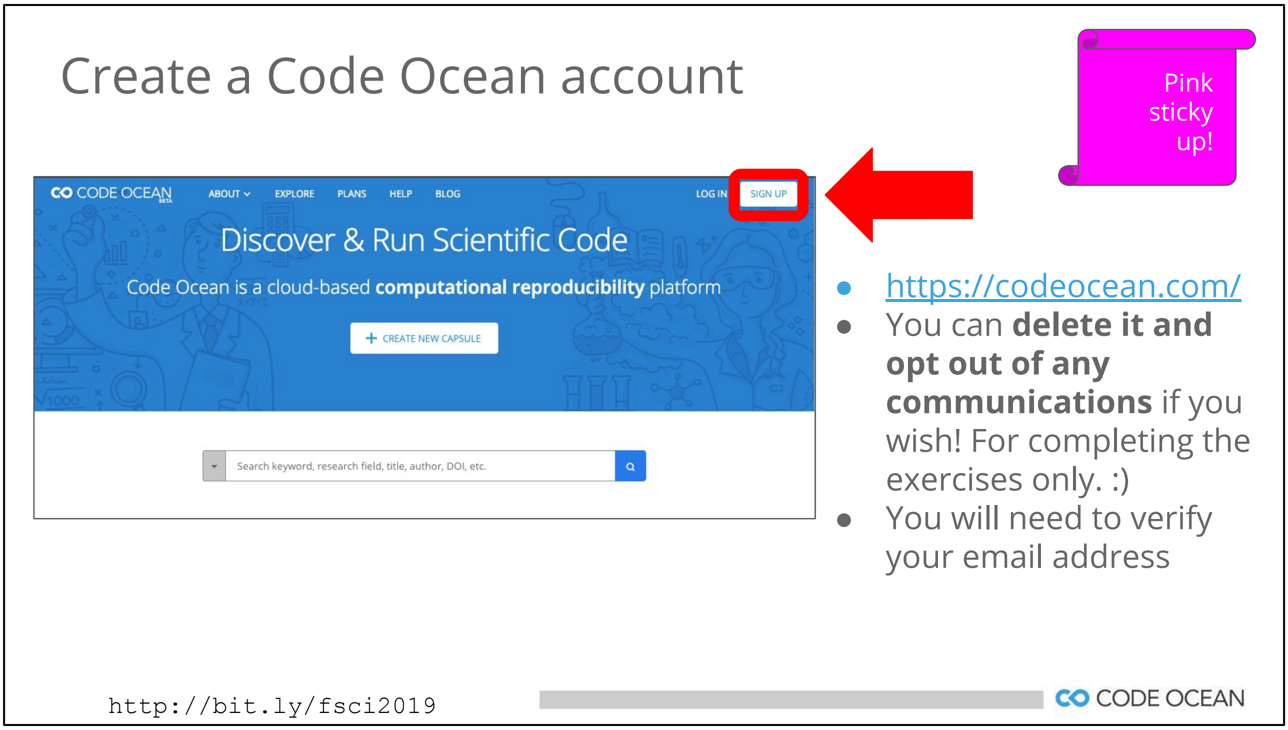
Containers solve:

- Dependency Hell - install, error, google, install, error…
  - Provides other researchers with a binary image in which all the software has already been installed, configured, and tested.
- Imprecise documentation - missing installation info.
  - Dockerfile provides a human readable summary of the necessary software dependencies needed to execute the code. Dependencies are automatically documented as they are installed.
- Code rot - dependencies change, the code breaks
  - Reduced risk with archiving images

Boettiger, Carl. *An introduction to Docker for reproducible research.* 10.1145/2723872.2723882

http://bit.ly/fsci2019

CODE OCEAN



Create a Code Ocean account

Discover & Run Scientific Code

Code Ocean is a cloud-based **computational reproducibility** platform

- https://codeocean.com/
- You can **delete it and opt out of any communications** if you wish! For completing the exercises only. :)
- You will need to verify your email address

http://bit.ly/fsci2019

- Everyone put your pink sticky up!
- We are going to create a Code Ocean account to create a repository, called a “capsule”, on Code Ocean for all research materials.
- Participants should:
  - Go to https://codeocean.com/
  - Sign up
  - They will need to verify their email address (sometimes they will need to resend this several times).
- Participants can delete their account and opt out of any communications if you wish! There is a check box on the survey where they should select “No” to receiving news if they wish.
- Code Ocean is for completing the exercises only, but the exercises can make code more reproducible no matter which platform participants which to use for themselves in the future.
- Once they have successfully created an account, they should switch to a green sticky.

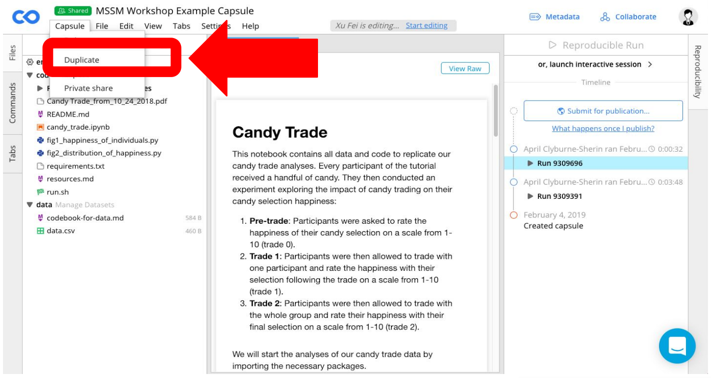
Duplicate this capsule:

R: http://bit.ly/r-example

Python: http://bit.ly/py-example

Pink sticky up!

- Click “Capsule”
- Select “Duplicate”


Shared
MSSM Workshop Example Capsule
Capsule
File
Edit
View
Tabs
Settings
Help
Xu Fei is editing... Start editing
Metadata
Collaborate
Files
Commands
Tabs
Duplicate
Private share
Candy Trade_from_10_24_2018.pdf
README.md
candy_trade.ipynb
fig1_happiness_of_individuals.py
fig2_distribution_of_happiness.py
requirements.txt
resources.md
run.sh
data Manage Datasets
codebook-for-data.md 584 B
data.csv 460 B
View Raw
Reproducible Run
or, launch interactive session
Timeline
Submit for publication...
What happens once I publish?
April Clyburne-Sherin ran Febru... 0:00:32
Run 9309696
April Clyburne-Sherin ran Febru... 0:03:48
Run 9309391
February 4, 2019
Created capsule
Reproducibility

Candy Trade

This notebook contains all data and code to replicate our candy trade analyses. Every participant of the tutorial received a handful of candy. They then conducted an experiment exploring the impact of candy trading on their candy selection happiness:

1. **Pre-trade**: Participants were asked to rate the happiness of their candy selection on a scale from 1-10 (trade 0).
2. **Trade 1**: Participants were then allowed to trade with one participant and rate the happiness with their selection following the trade on a scale from 1-10 (trade 1).
3. **Trade 2**: Participants were then allowed to trade with the whole group and rate their happiness with their final selection on a scale from 1-10 (trade 2).

We will start the analyses of our candy trade data by importing the necessary packages.

http://bit.ly/fsci2019

CODE OCEAN

- Everyone put your pink sticky up!
- Participants should
  - Ensure they are in “Classic” Code Ocean mode for the workshop.
  - Select the ellipses [...].
  - Click Duplicate
- Duplicating this capsule creates a copy owned by you, and you know you are successful when you have a capsule named “____ Workshop (copy)”
- Once they have successfully duplicated the capsule, they should switch to a green sticky.

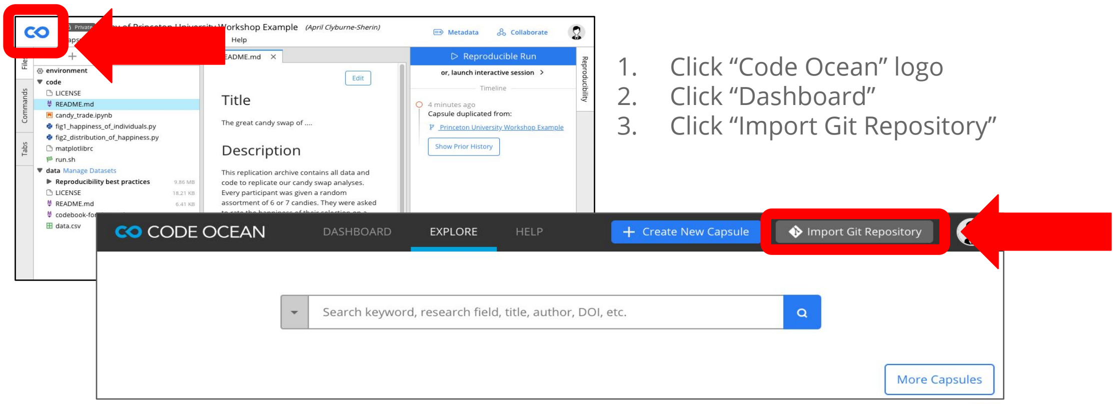
About this capsule:

- This is an example and reference capsule for the workshop.
- It is a skeleton example of where are headed with our exercises.
- It also includes 3 papers that can be great resources for computational reproducibility if you wish to read further.

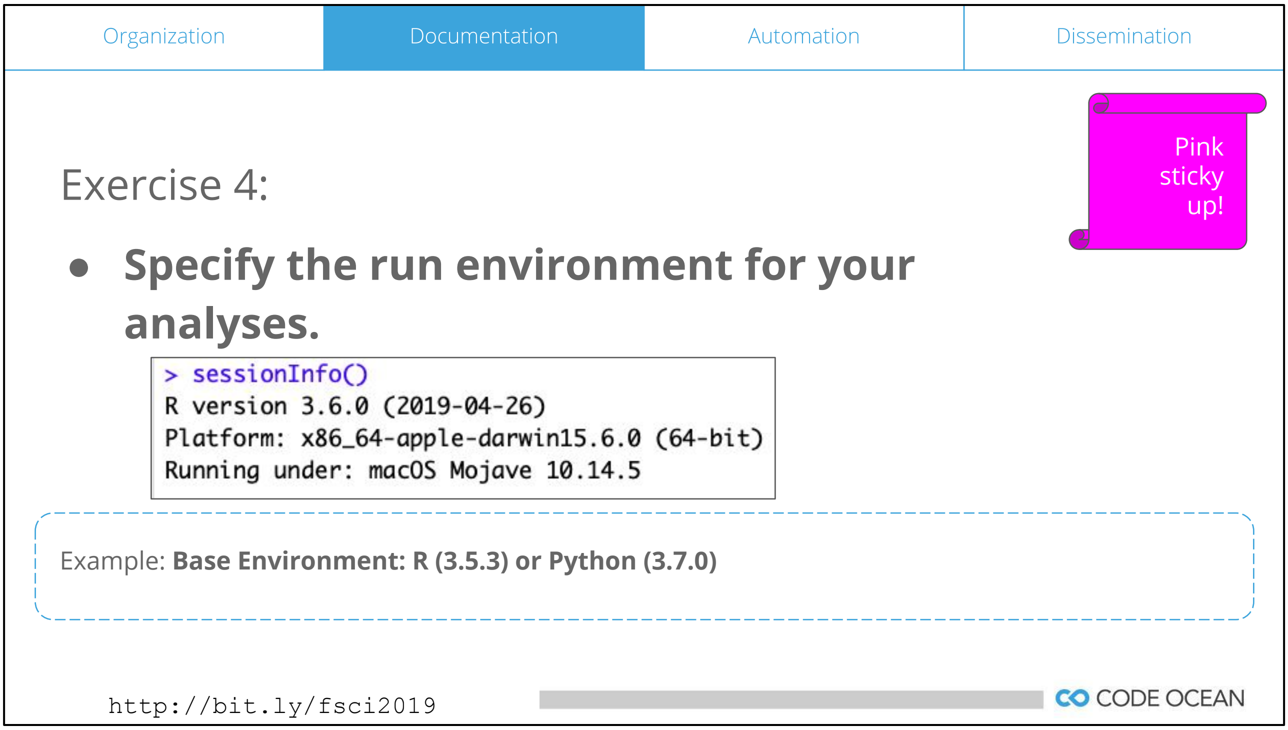
Create a new compute capsule

Import Git Repository:
R: https://github.com/aprilcs/sips-workshop
Python: https://github.com/aprilcs/sips-workshop-py

Pink
sticky
up!


Private
...y of Princeton University Workshop Example (April Clyburne-Sherin)
Help
Metadata
Collaborate
File
Commands
Tabs
environment
code
LICENSE
README.md
candy_trade.ipynb
fig1_happiness_of_individuals.py
fig2_distribution_of_happiness.py
matplotlibrc
run.sh
data Manage Datasets
Reproducibility best practices
LICENSE
README.md
codebook-fo...
data.csv
...EADME.md
Edit
Title
The great candy swap of ....
Description
This replication archive contains all data and code to replicate our candy swap analyses. Every participant was given a random assortment of 6 or 7 candies. They were asked
Reproducible Run
or, launch interactive session
Timeline
4 minutes ago
Capsule duplicated from:
Princeton University Workshop Example
Show Prior History
Reproducibility
CODE OCEAN
DASHBOARD
EXPLORE
HELP
Create New Capsule
Import Git Repository
Search keyword, research field, title, author, DOI, etc.
More Capsules

1. Click “Code Ocean” logo
2. Click “Dashboard”
3. Click “Import Git Repository”



Organization | Documentation | Automation | Dissemination

Pink sticky up!

Exercise 4:

- **Specify the run environment for your analyses.**

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Mojave 10.14.5
```

Example: **Base Environment: R (3.5.3) or Python (3.7.0)**

http://bit.ly/fsci2019

CODE OCEAN

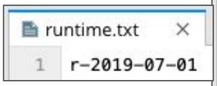
This comes with miniconda and is called conda.

Python --version, when you open a terminal and type python

Pink sticky up!

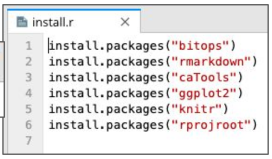
Exercise 5:

- **Specify your packages and dependencies with versions.**
  - ○ Python: pip freeze > /requirements.txt
  - ○ R: install.r and runtime.txt



runtime.txt

```
r-2019-07-01
```



install.r

```
install.packages("bitops")
install.packages("rmarkdown")
install.packages("caTools")
install.packages("ggplot2")
install.packages("knitr")
install.packages("rprojroot")
```

R packages: **apt-get** pandoc; **CRAN** bitops, markdown, caTools, ggplot2, knitr, rprojroot

Python packages: **conda** matplotlib, pandas, numpy, jupyter

Resource on documenting dependencies:

- Binder: https://mybinder.readthedocs.io/en/latest/config_files.html

Pink sticky up!

Exercise 6:

- **Use container technology to create an image of your complete computational environment.**
  - Code Ocean
  - Binder

Export your capsule to see how an image and Dockerfile were created through your specifications.

Inspect the Dockerfile.

We will demonstrate building a container with repo2docker using mybinder and github.

Pink sticky up!

Demo:

- **Consider using literate programming to document the analysis narrative with the code.**
    - Jupyter Notebooks
    - RMarkdown

Explore Jupyter notebooks in this example capsule: http://bit.ly/uiuc-example

Explore RMarkdown in this example capsule: http://bit.ly/rmarkdown-example

- **Create a README file and data dictionary.**

Documenting your file overview and dependencies in your README:

- AJPS Replication Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EZSJ1S

Documenting your data in a codebook or data dictionary:

- DataONE: https://www.dataone.org/best-practices/create-data-dictionary

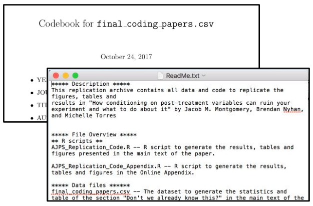
Resource on using markdown:

- GitHub: https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet

Organization | Documentation | Automation | Dissemina

Reference Slide

Checklist



Codebook for final_coding_papers.csv

October 24, 2017

- YE
- JOU
- TIT
- AU

ReadMe.txt

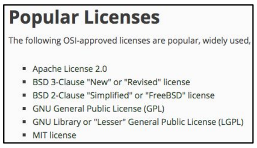
```
***** Description *****
This replication archive contains all data and code to replicate the
figures, tables and
results in "How conditioning on post-treatment variables can ruin your
experiment and what to do about it" by Jacob M. Montgomery, Brendan Nyhan,
and Michelle Torres


***** File Overview *****
** R scripts **
AJPS_Replication_Code.R -- R script to generate the results, tables and
figures presented in the main text of the paper.

AJPS_Replication_Code_Appendix.R -- R script to generate the results,
tables and figures in the Online Appendix.

***** Data files ******
final_coding_papers.csv -- The dataset to generate the statistics and
table of the section "Don't we already know this?" in the main text of the
```

- Consider literate programming.
- Document each element or variable in your dataset with a data dictionary / codebook.
- Create a README file.

Tools

- Version control: git and GitHub tracks changes to documents and metadata
- Literate programming: knits documentation with code (Jupyter)
- Document & share metadata: Code Ocean renders documentation, notebooks, and records metadata

Resources



Popular Licenses

The following OSI-approved licenses are popular, widely used,

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license

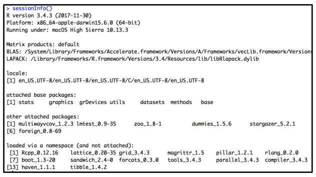
- DataONE: https://www.dataone.org/best-practices/create-data-dictionary
- Cornell: https://data.research.cornell.edu/content/readme
- Digital Curation Center: http://www.dcc.ac.uk/resources/how-guides/license-research-data
- OSI: https://opensource.org/licenses

http://bit.ly/fsci2019

Organization | Documentation | Automation | Dissemina

Reference Slide

Checklist



```
> sessionInfo()
R version 3.4.3 (2017-11-30)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS High Sierra 10.13.3

Matrix products: default
BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Version
LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib

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

other attached packages:
[1] multiwayvcov_1.2.3 lmtest_0.9-35      zoo_1.8-1          dummies_1.5.6      stargazer_5.2.1
[6] foreign_0.8-69

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.16    lattice_0.20-35 grid_3.4.3      magrittr_1.5    pillar_1.2.1    rlang_0.2.0
 [7] boot_1.3-20     sandwich_2.4-0  forcats_0.3.0   tools_3.4.3     parallel_3.4.3  compiler_3.4.3
[13] haven_1.1.1     tibble_1.4.2
```

- Specify your computational environment and package versions.
- Configure a container to make your analysis portable and reusable.

Tools

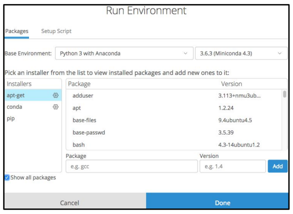
CODE OCEAN

docker

binder (beta)

- Container technology: packages data, code, metadata, & computational environment for portable analyses
- Docker: container technology for devs
- Code Ocean: easy configuring, preservation, & reuse of containers for researchers
- Binder: configure & share containers

Resources



Run Environment

Base Environment: Python 3 with Anaconda | 3.6.3 (Miniconda 4.3)

Pick an installer from the list to view installed packages and add new ones to it:

Installers
apt-get
conda
pip

Package	Version
adduser	3.113+nmu3ub...
apt	1.2.24
base-files	9.4ubuntu4.5
base-passwd	3.5.39
bash	4.3-14ubuntu1.2

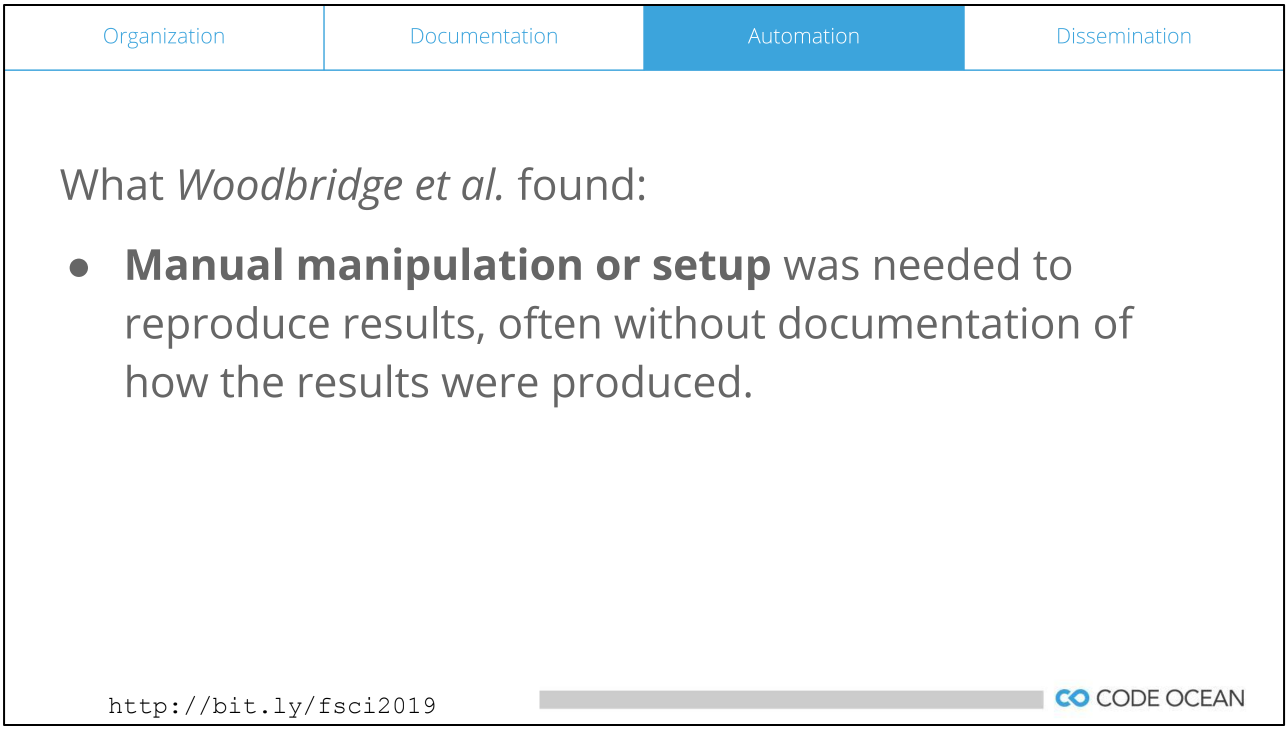
Package: e.g. gcc | Version: e.g. 1.4

☑ Show all packages

- Documenting dependencies: http://mybinder.readthedocs.io/en/latest/using.html#preparing-a-repository-for-binder
- Specifying environments: https://help.codeocean.com/getting-started/the-computational-environment/selecting-a-base-environment

http://bit.ly/fsci2019

CODE OCEAN



Organization | Documentation | Automation | Dissemination

What *Woodbridge et al.* found:

- **Manual manipulation or setup** was needed to reproduce results, often without documentation of how the results were produced.

http://bit.ly/fsci2019

CODE OCEAN

We can **automate the execution of our analyses**:

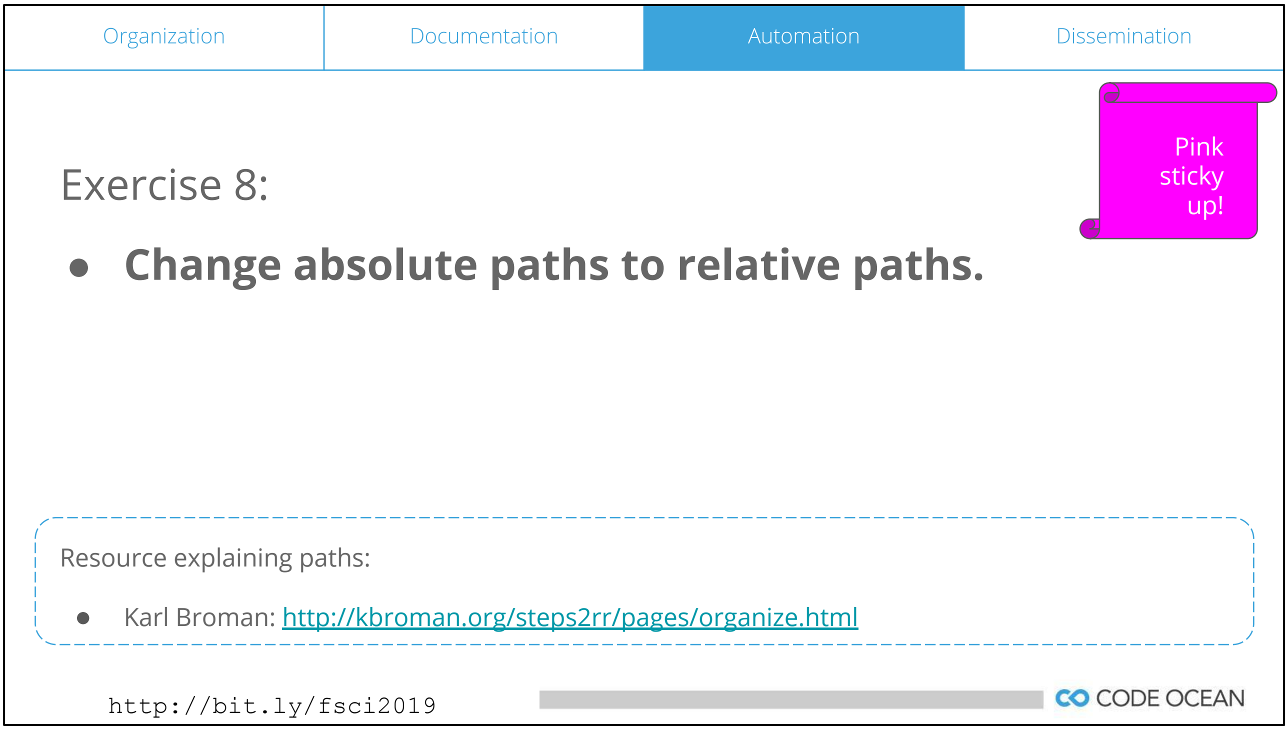
- Create a master script to execute all analyses.
- Reproduce results automatically as a function of the data & the code; Save results explicitly.
- Use relative paths.

Pink sticky up!

Exercise 7:

- **Create a master script to execute your code.**

- In R, use a run.r or main.r master script
  - Use source() to run your scripts
  - Run your install.r script
- In Python, use a main.py or run.sh master script
  - In your run.sh script, use nbconvert to execute your notebook into the results directory.
- Case study: https://www.practicereproducibleresearch.org/core-chapters/3-basic.ht



Organization | Documentation | Automation | Dissemination

Pink sticky up!

Exercise 8:

- **Change absolute paths to relative paths.**

Resource explaining paths:

- Karl Broman: http://kbroman.org/steps2rr/pages/organize.html

http://bit.ly/fsci2019

CODE OCEAN

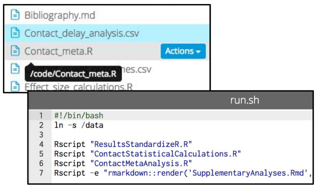
Show an example of changing a path:

- In the python code “fig1_happiness_of_individuals”
  - Change “data = pd.read_csv('../data/data.csv')” to “data = pd.read_csv('C:/Users/SomeOne/Projects/data/data.csv')”

Organization | Documentation | **Automation** | Dissemina

Reference Slide

Checklist



Bibliography.md
Contact_delay_analysis.csv
Contact_meta.R
/code/Contact_meta.R
Effect_size_calculations.R

run.sh

```
#!/bin/bash
ln -s /data

Rscript "ResultsStandardizeR.R"
Rscript "ContactStatisticalCalculations.R"
Rscript "ContactMetaAnalysis.R"
Rscript -e "rmarkdown::render('SupplementaryAnalyses.Rmd',
```

- Use relative rather than absolute paths.
- Create a master script that runs your scripts in sequence.

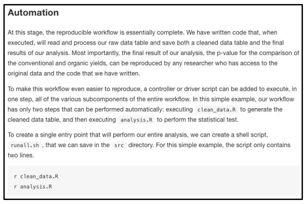
Tools

CODE OCEAN   docker

binder (beta)

- Docker: share automated code for devs
- Code Ocean: easy configuring, preservation, & reuse of automated code
- Binder: share automated code for using containers

Resources



Automation

At this stage, the reproducible workflow is essentially complete. We have written code that, when executed, will read and process our raw data table and save both a cleaned data table and the final results of our analysis. Most importantly, the final result of our analysis, the p-value for the comparison of the conventional and organic yields, can be reproduced by any researcher who has access to the original data and the code that we have written.

To make this workflow even easier to reproduce, a controller or driver script can be added to execute, in one step, all of the various subcomponents of the entire workflow. In this simple example, our workflow has only two steps that can be performed automatically: executing `clean_data.R` to generate the cleaned data table, and then executing `analysis.R` to perform the statistical test.

To create a single entry point that will perform our entire analysis, we can create a shell script, `runall.sh`, that we can save in the `src` directory. For this simple example, the script only contains two lines.

```
r clean_data.R
r analysis.R
```

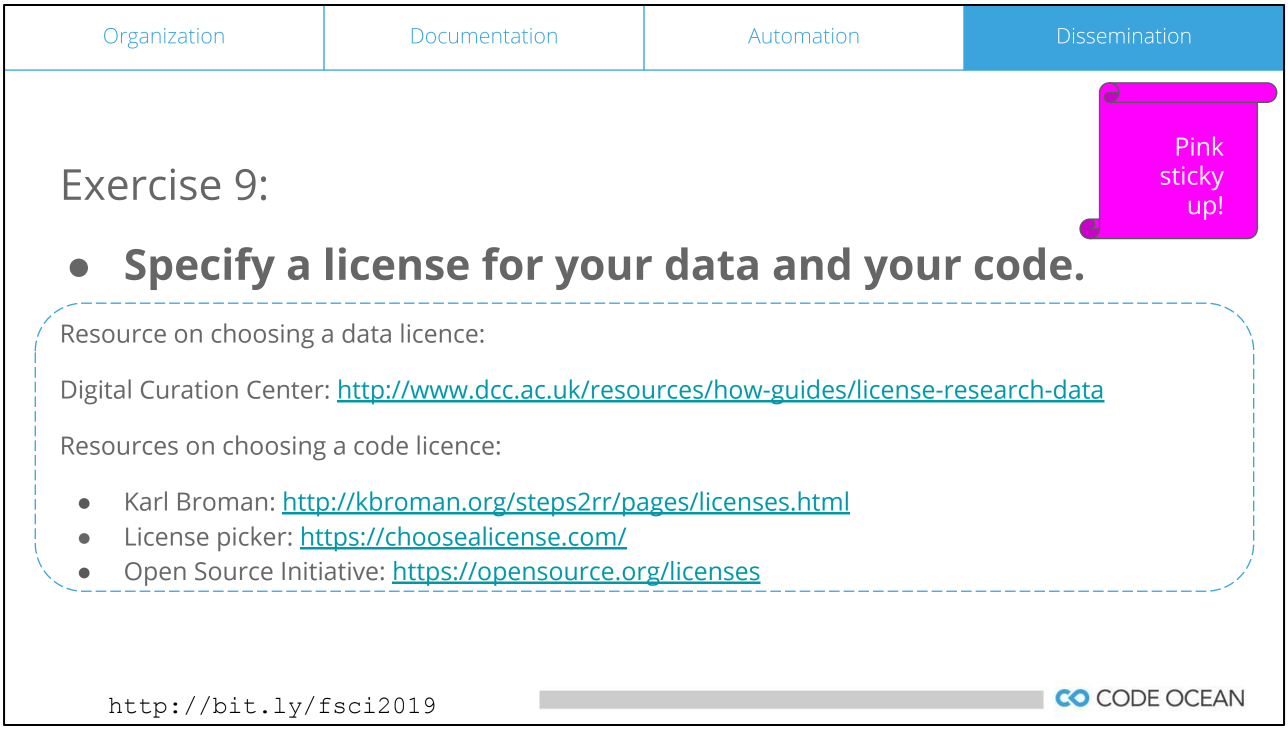
- Karl Broman on paths: http://kbroman.org/steps2rr/pages/organize.html
- Resource on automation using a master script: https://www.practicereproducibleresearch.org/core-chapters/3-basic.html

What *Woodbridge et al.* found:

- There is no standardized way of **attaching code to published articles**.
- Therefore it is difficult to **discover and retrieve** code.

We can **embed or link code persistently**:

- **Obtain a DOI for your repository and use this link throughout your article.**
  - Example: [Github -> Binder -> Zenodo -> DOI linked in article]
  - Example: [CodeOcean -> DOI in article]
- **Cross link repository with published article in metadata of each.**
- **Embed executable capsule within the article.**
  - Example: https://doi.org/10.1017/bpp.2018.25



Organization | Documentation | Automation | Dissemination

Pink sticky up!

Exercise 9:

- **Specify a license for your data and your code.**

Resource on choosing a data licence:

Digital Curation Center: http://www.dcc.ac.uk/resources/how-guides/license-research-data

Resources on choosing a code licence:

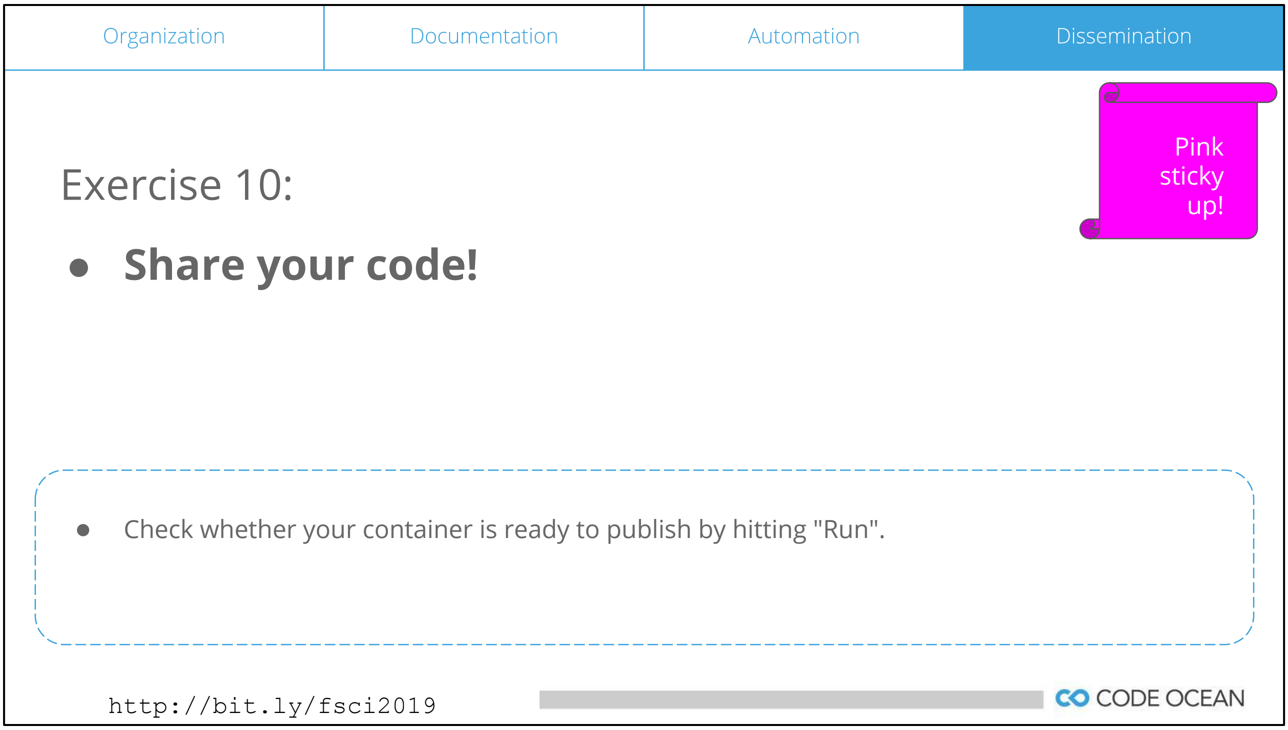
- Karl Broman: http://kbroman.org/steps2rr/pages/licenses.html
- License picker: https://choosealicense.com/
- Open Source Initiative: https://opensource.org/licenses

http://bit.ly/fsci2019

CODE OCEAN

Add a licence.txt file to your project or select one in the metadata section (CO or GitHub)

- Consider Creative Commons licenses for data and text, either CC-0 or CC-BY.
- For software, we recommend a permissive open source license such as the MIT, BSD, or Apache license



Organization | Documentation | Automation | **Dissemination**

Pink sticky up!

Exercise 10:

- **Share your code!**

- Check whether your container is ready to publish by hitting "Run".

http://bit.ly/fsci2019

CODE OCEAN

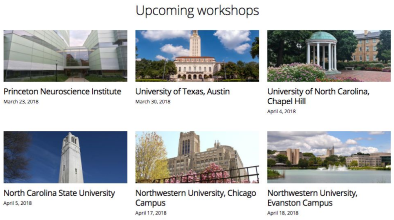
Reproducibility support

Workshops & Webinars

- Theory or hands-on
- Customized to researcher needs
- Request a workshop or webinar at https://codeocean.com/events

1:1 Computational Reproducibility Consult

- In person
  - Lab meeting
  - Office visit
- Virtual
- Request at april@codeocean.com or https://doodle.com/codeocean



Upcoming workshops

Princeton Neuroscience Institute
March 23, 2018

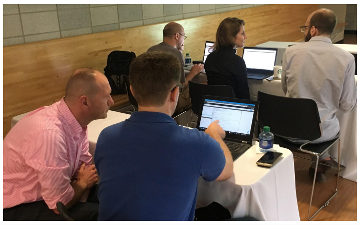
University of Texas, Austin
March 30, 2018

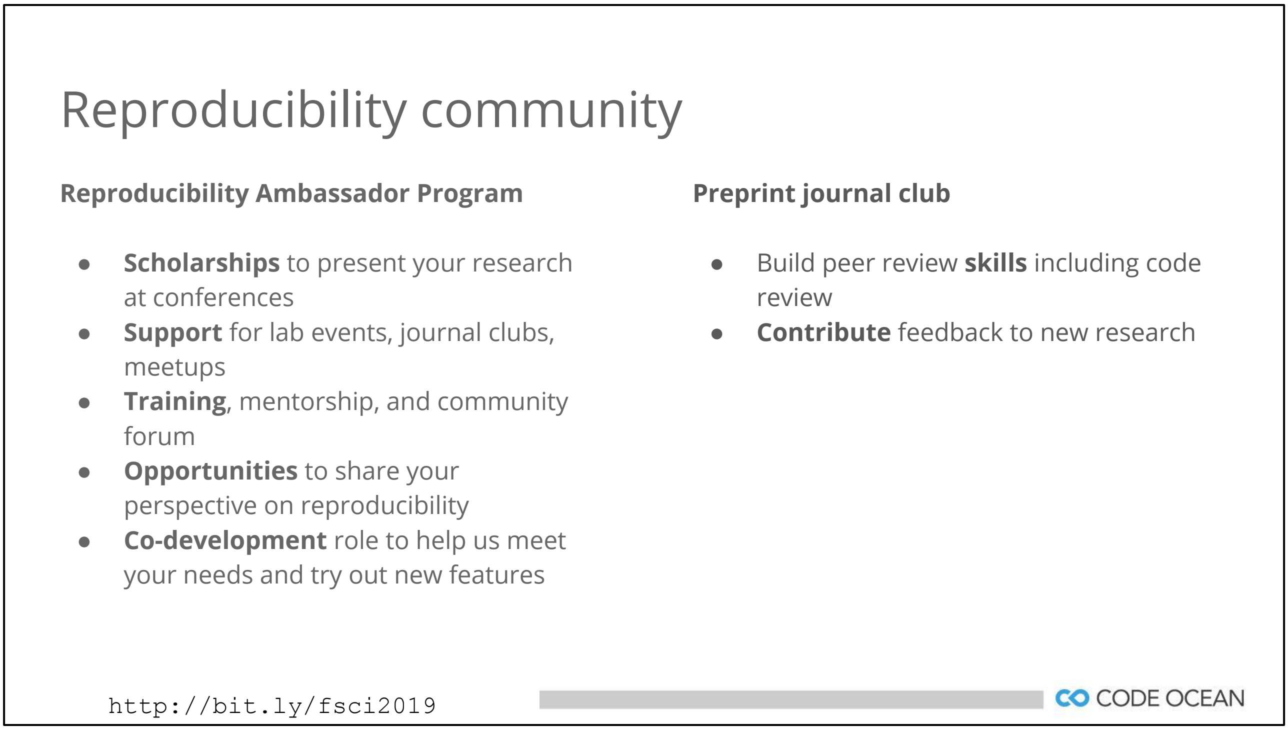
University of North Carolina, Chapel Hill
April 4, 2018

North Carolina State University
April 5, 2018

Northwestern University, Chicago Campus
April 17, 2018

Northwestern University, Evanston Campus
April 18, 2018





Reproducibility community

Reproducibility Ambassador Program

- **Scholarships** to present your research at conferences
- **Support** for lab events, journal clubs, meetups
- **Training**, mentorship, and community forum
- **Opportunities** to share your perspective on reproducibility
- **Co-development** role to help us meet your needs and try out new features

Preprint journal club

- Build peer review **skills** including code review
- **Contribute** feedback to new research

http://bit.ly/fsci2019

CODE OCEAN

Thank you for your time :)

Please fill out an evaluation so we can keep improving!
http://bit.ly/workshop-survey-2019

April Clyburne-Sherin
Code Ocean

april@codeocean.com