Code Ocean eScience presentation

Seth Green

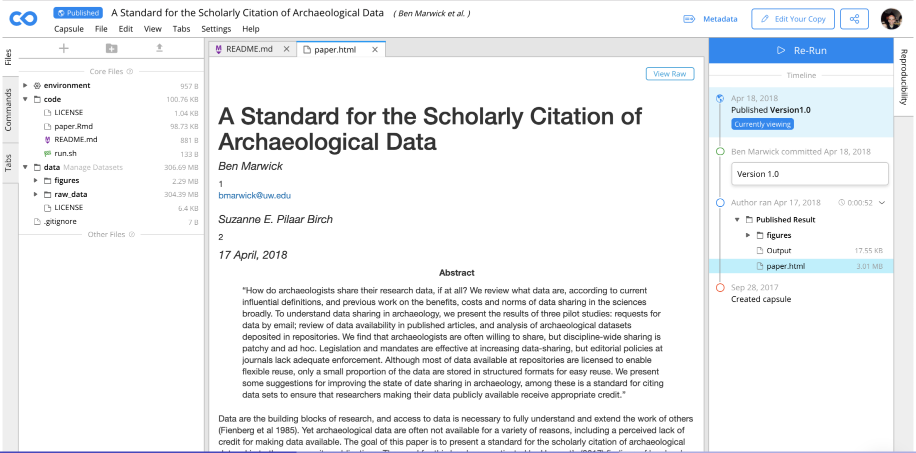
June 10, 2019

A really important quote we should all think about

"[A]n article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." David Donoho, paraphrasing John Claerbout

Code Ocean

- Code Ocean is, more or less:
  - JupyterLab IDE + modificationss
  - A robust dependency management system
  - A publishing platform (DOIs & stable URLs)
  - A sharing platform (embed your 'compute capsules' on webpages)



Core Files

environment	957 B
code	100.76 KB
LICENSE	1.04 KB
paper.Rmd	98.73 KB
README.md	881 B
run.sh	133 B
data Manage Datasets	306.69 MB
figures	2.29 MB
raw_data	304.39 MB
LICENSE	6.4 KB
.gitignore	7 B

Other Files

A Standard for the Scholarly Citation of Archaeological Data

Ben Marwick

1

bmarwick@uw.edu

Suzanne E. Pilaar Birch

2

17 April, 2018

Abstract

"How do archaeologists share their research data, if at all? We review what data are, according to current influential definitions, and previous work on the benefits, costs and norms of data sharing in the sciences broadly. To understand data sharing in archaeology, we present the results of three pilot studies: requests for data by email; review of data availability in published articles, and analysis of archaeological datasets deposited in repositories. We find that archaeologists are often willing to share, but discipline-wide sharing is patchy and ad hoc. Legislation and mandates are effective at increasing data-sharing, but editorial policies at journals lack adequate enforcement. Although most of data available at repositories are licensed to enable flexible reuse, only a small proportion of the data are stored in structured formats for easy reuse. We present some suggestions for improving the state of date sharing in archaeology, among these is a standard for citing data sets to ensure that researchers making their data publicly available receive appropriate credit."

Data are the building blocks of research, and access to data is necessary to fully understand and extend the work of others (Fienberg et al 1985). Yet archaeological data are often not available for a variety of reasons, including a perceived lack of credit for making data available. The goal of this paper is to present a standard for the scholarly citation of archaeological

Apr 18, 2018
Published Version1.0
Currently viewing

Ben Marwick committed Apr 18, 2018
Version 1.0

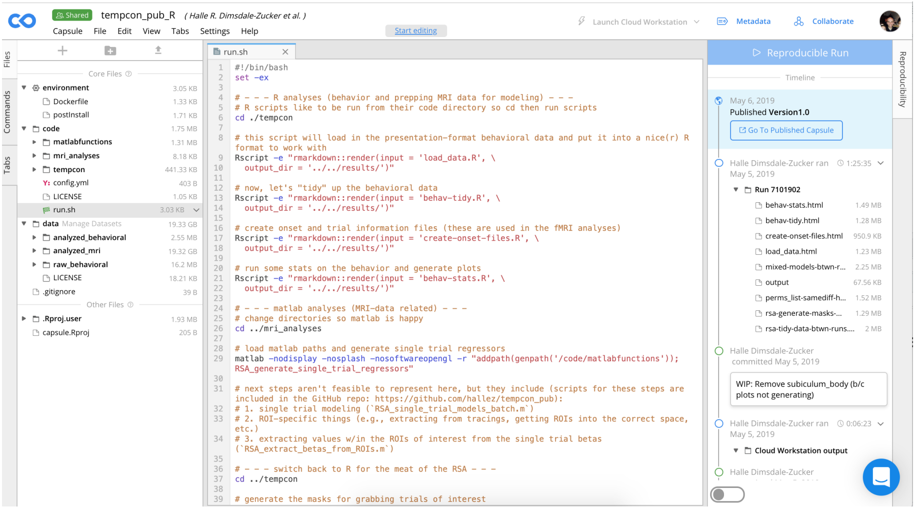
Author ran Apr 17, 2018 0:00:52

Published Result	
figures	
Output	17.55 KB
paper.html	3.01 MB

Sep 28, 2017
Created capsule

Publishing reproducible R notebooks

- https://codeocean.com/capsule/5777882/tree/v1 for .Rmd
- https://codeocean.com/capsule/0129473/tree for *.R.



Core Files

environment	3.05 KB
Dockerfile	1.33 KB
postInstall	1.71 KB
code	1.75 MB
matlabfunctions	1.31 MB
mri_analyses	8.18 KB
tempcon	441.33 KB
config.yml	403 B
LICENSE	1.05 KB
run.sh	3.03 KB
data Manage Datasets	19.33 GB
analyzed_behavioral	2.55 MB
analyzed_mri	19.32 GB
raw_behavioral	16.2 MB
LICENSE	18.21 KB
.gitignore	39 B

Other Files

.Rproj.user	1.93 MB
capsule.Rproj	205 B

```
#!/bin/bash
set -ex

# - - - R analyses (behavior and prepping MRI data for modeling) - - -
# R scripts like to be run from their code directory so cd then run scripts
cd ./tempcon

# this script will load in the presentation-format behavioral data and put it into a nice(r) R
format to work with
Rscript -e "rmarkdown::render(input = 'load_data.R', \
  output_dir = '../../results/')"

# now, let's "tidy" up the behavioral data
Rscript -e "rmarkdown::render(input = 'behav-tidy.R', \
  output_dir = '../../results/')"

# create onset and trial information files (these are used in the fMRI analyses)
Rscript -e "rmarkdown::render(input = 'create-onset-files.R', \
  output_dir = '../../results/')"

# run some stats on the behavior and generate plots
Rscript -e "rmarkdown::render(input = 'behav-stats.R', \
  output_dir = '../../results/')"

# - - - matlab analyses (MRI-data related) - - -
# change directories so matlab is happy
cd ../mri_analyses

# load matlab paths and generate single trial regressors
matlab -nodisplay -nosplash -nosoftwareopengl -r "addpath(genpath('/code/matlabfunctions'));
RSA_generate_single_trial_regressors"

# next steps aren't feasible to represent here, but they include (scripts for these steps are
included in the GitHub repo: https://github.com/hallez/tempcon_pub):
# 1. single trial modeling (`RSA_single_trial_models_batch.m`)
# 2. ROI-specific things (e.g., extracting from tracings, getting ROIs into the correct space,
etc.)
# 3. extracting values w/in the ROIs of interest from the single trial betas
(`RSA_extract_betas_from_ROIs.m`)

# - - - switch back to R for the meat of the RSA - - -
cd ../tempcon

# generate the masks for grabbing trials of interest
```

May 6, 2019
Published Version1.0
Go To Published Capsule

Halle Dimsdale-Zucker ran 1:25:35
May 5, 2019

Run 7101902

behav-stats.html	1.49 MB
behav-tidy.html	1.28 MB
create-onset-files.html	950.9 KB
load_data.html	1.23 MB
mixed-models-btwn-r...	2.25 MB
output	67.56 KB
perms_list-samediff-h...	1.52 MB
rsa-generate-masks-...	1.29 MB
rsa-tidy-data-btwn-runs....	2 MB

Halle Dimsdale-Zucker
committed May 5, 2019

WIP: Remove subiculum_body (b/c plots not generating)

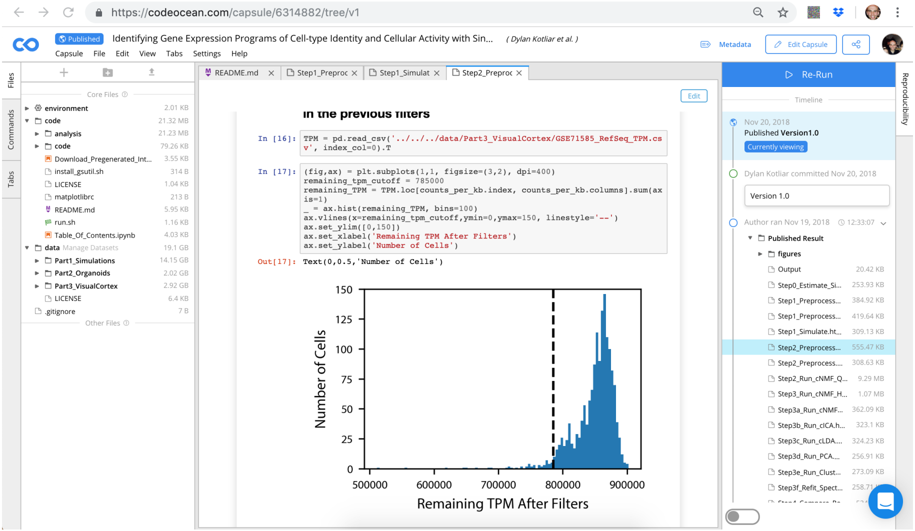
Halle Dimsdale-Zucker ran 0:06:23
May 5, 2019

Cloud Workstation output

Halle Dimsdale-Zucker

Publishing reproducible Jupyter Notebooks

- Notebook + environment + nbconvert = a rendered HTML
- https://codeocean.com/capsule/6314882/tree/v1



environment	2.01 KB
code	21.32 MB
analysis	21.23 MB
code	79.26 KB
Download_Pregenerated_Int...	3.55 KB
install_gsutil.sh	314 B
LICENSE	1.04 KB
matplotlibrc	213 B
README.md	5.95 KB
run.sh	1.16 KB
Table_Of_Contents.ipynb	4.03 KB
data Manage Datasets	19.1 GB
Part1_Simulations	14.15 GB
Part2_Organoids	2.02 GB
Part3_VisualCortex	2.92 GB
LICENSE	6.4 KB
.gitignore	7 B

in the previous filters

In [16]:
```
TPM = pd.read_csv('../../../data/Part3_VisualCortex/GSE71585_RefSeq_TPM.cs
v', index_col=0).T
```

In [17]:
```
(fig,ax) = plt.subplots(1,1, figsize=(3,2), dpi=400)
remaining_tpm_cutoff = 785000
remaining_TPM = TPM.loc[counts_per_kb.index, counts_per_kb.columns].sum(ax
is=1)
_ = ax.hist(remaining_TPM, bins=100)
ax.vlines(x=remaining_tpm_cutoff,ymin=0,ymax=150, linestyle='--')
ax.set_ylim([0,150])
ax.set_xlabel('Remaining TPM After Filters')
ax.set_ylabel('Number of Cells')
```

Out[17]:
```
Text(0,0.5,'Number of Cells')
```

Nov 20, 2018
Published **Version1.0**
Currently viewing

Dylan Kotliar committed Nov 20, 2018

Version 1.0

Author ran Nov 19, 2018 12:33:07

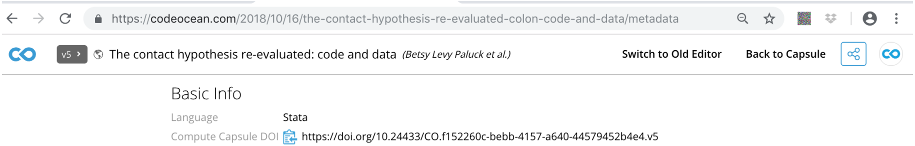
Published Result	
figures	
Output	20.42 KB
Step0_Estimate_Si...	253.93 KB
Step1_Preprocess...	384.92 KB
Step1_Preprocess...	419.64 KB
Step1_Simulate.ht...	309.13 KB
Step2_Preprocess...	555.47 KB
Step2_Preprocess....	308.63 KB
Step2_Run_cNMF_Q...	9.29 MB
Step3_Run_cNMF_H...	1.07 MB
Step3a_Run_cNMF...	362.09 KB
Step3b_Run_cICA.h...	323.1 KB
Step3c_Run_cLDA....	324.23 KB
Step3d_Run_PCA....	256.91 KB
Step3e_Run_Clust...	273.09 KB
Step3f_Refit_Spect...	258.71

Interactive sessions

- https://codeocean.com/capsule/8962292/tree/v2
- Jupyter
- JupyterLab
- Time for a live demonstration

Questions?

- How is this different than Binder?
- What is the uploading process like?
- How are dependencies managed?
- Is this exportable?
- Whom is Code Ocean for? (All scientists)
- What are the most pressing issues in computational reproducibility? (e.g., big data, confidential data, inferring system level dependencies from listed scientific libraries, adjudicating between many competing worthy aims, what language should we all be using. . . .) ## Reference Slide 1: Publishing on Code Ocean:
- https://codeocean.com/2018/10/16/the-contact-hypothesis-re-evaluated-colon-code-and-data/code
- Will have a DOI and link to your article's metadata


https://codeocean.com/2018/10/16/the-contact-hypothesis-re-evaluated-colon-code-and-data/metadata
v5
Switch to Old Editor
Back to Capsule

The contact hypothesis re-evaluated: code and data *(Betsy Levy Paluck et al.)*

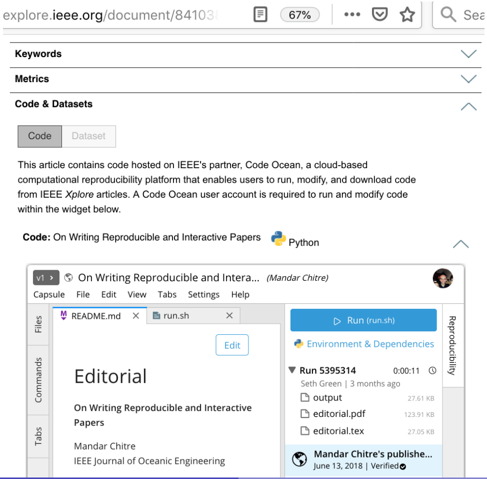
Basic Info

Language Stata

Compute Capsule DOI https://doi.org/10.24433/CO.f152260c-bebb-4157-a640-44579452b4e4.v5

Reference Slide 2: Embedding on webpages & within articles

- You can also embed your published capsule in your article's HTML page or on your personal webpage, a la https://ieeexplore.ieee.org/document/8410389/algorithms#algorithms:



explore.ieee.org/document/84103
67%
Sea
Keywords
Metrics
Code & Datasets
Code
Dataset

This article contains code hosted on IEEE's partner, Code Ocean, a cloud-based computational reproducibility platform that enables users to run, modify, and download code from IEEE *Xplore* articles. A Code Ocean user account is required to run and modify code within the widget below.

Code: On Writing Reproducible and Interactive Papers Python

v1
On Writing Reproducible and Intera... (Mandar Chitre)
Capsule File Edit View Tabs Settings Help
Files
Commands
Tabs
README.md
run.sh
Edit
Run (run.sh)
Environment & Dependencies
Reproducibility

Editorial

On Writing Reproducible and Interactive Papers

Mandar Chitre
IEEE Journal of Oceanic Engineering

Run 5395314 0:00:11
Seth Green | 3 months ago

output	27.61 KB
editorial.pdf	123.91 KB
editorial.tex	27.05 KB

Mandar Chitre's publishe...
June 13, 2018 | Verified