

Code Ocean eScience presentation

Seth Green

June 10, 2019

Code Ocean

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

The screenshot displays the Code Ocean web interface. At the top, a navigation bar includes the Code Ocean logo, a 'Published' status badge, the capsule title 'A Standard for the Scholarly Citation of Archaeological Data' by Ben Marwick et al., and buttons for 'Metadata', 'Edit Your Copy', and a user profile icon. Below the navigation bar is a menu with 'Capsule', 'File', 'Edit', 'View', 'Tabs', 'Settings', and 'Help'.

The main workspace is divided into three panels:

- Files Panel (Left):** A file explorer showing the capsule's structure. It includes a 'Core Files' section with 'environment' (957 B), 'code' (100.76 KB) containing 'LICENSE', 'paper.Rmd', 'README.md', and 'run.sh', and 'data' (306.69 MB) containing 'figures', 'raw_data', and 'LICENSE'. There is also an 'Other Files' section with a '.gitignore' file (7 B).
- Editor Panel (Center):** Displays the 'paper.html' file. The title is 'A Standard for the Scholarly Citation of Archaeological Data' by Ben Marwick. The author's email is bmarwick@uw.edu. The paper is dated 17 April, 2018. The abstract discusses the challenges of data sharing in archaeology and the need for a standard. The full text of the paper is visible below the abstract.
- Reproducibility Panel (Right):** A 'Re-Run' section showing the capsule's timeline. It includes a 'Published Version 1.0' badge, a 'Currently viewing' status, and a list of actions: 'Ben Marwick committed Apr 18, 2018', 'Author ran Apr 17, 2018' (00:00:52), and 'Sep 28, 2017 Created capsule'. A 'Published Result' section shows the outputs: 'figures' (17.55 KB), 'Output' (3.01 MB), and 'paper.html' (3.01 MB).

Publishing reproducible R notebooks

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

The screenshot displays the Code Ocean web interface for a capsule named 'tempcon_pub_R' by Halle R. Dimsdale-Zucker et al. The interface is divided into several sections:

- Files:** A sidebar on the left showing the file structure. It includes 'Core Files' (environment, code, data) and 'Other Files' (.Rproj.user, capsule.Rproj). The 'code' folder is expanded, showing 'run.sh' as the active file.
- Commands:** A central pane showing the contents of 'run.sh'. The script is a bash file that sets up an R environment, loads behavioral data, and generates trial regressors and masks. It includes comments explaining the steps and the purpose of the scripts.
- Reproducible Run:** A right-hand pane showing the execution timeline. It lists the published version (1.0) and the run history. The most recent run is by Halle Dimsdale-Zucker on May 5, 2019, with a duration of 1:25:35. Below this, a list of generated files is shown, including 'behav-stats.html', 'behav-tidy.html', 'create-onset-files.html', 'load_data.html', 'output', 'mixed-models-btwn-r...', 'perms_list-samediff-h...', 'rsa-generate-masks...', and 'rsa-tidy-data-btwn-runs...'. A note indicates that the 'WIP: Remove subiculum_body (b/c plots not generating)' is still in progress.

```
1 #!/bin/bash
2 set -ex
3
4 # --- R analyses (behavior and prepping MRI data for modeling) ---
5 # R scripts like to be run from their code directory so cd then run scripts
6 cd ../tempcon
7
8 # this script will load in the presentation-format behavioral data and put it into a nice(r) R
9 # format to work with
10 Rscript -e "rmarkdown::render(input = 'load_data.R', \
11   output_dir = '../results/')"
12
13 # now, let's "tidy" up the behavioral data
14 Rscript -e "rmarkdown::render(input = 'behav-tidy.R', \
15   output_dir = '../results/')"
16
17 # create onset and trial information files (these are used in the fMRI analyses)
18 Rscript -e "rmarkdown::render(input = 'create-onset-files.R', \
19   output_dir = '../results/')"
20
21 # run some stats on the behavior and generate plots
22 Rscript -e "rmarkdown::render(input = 'behav-stats.R', \
23   output_dir = '../results/')"
24
25 # --- matlab analyses (MRI-data related) ---
26 # change directories so matlab is happy
27 cd ../mri_analyses
28
29 # load matlab paths and generate single trial regressors
30 matlab -nodisplay -nosplash -nosoftawareopen -r "addpath(genpath('../code/matlabfunctions'));
31   RSA_generate_single_trial_regressors"
32
33 # next steps aren't feasible to represent here, but they include (scripts for these steps are
34 # included in the GitHub repo: https://github.com/hallez/tempcon_pub):
35 # 1. single trial modeling ('RSA_single_trial_models_batch.m')
36 # 2. ROI-specific things (e.g., extracting from tracings, getting ROIs into the correct space,
37 # etc.)
38 # 3. extracting values w/in the ROIs of interest from the single trial betas
39 ('RSA_extract_betas_from_ROIs.m')
40
41 # --- switch back to R for the meat of the RSA ---
42 cd ../tempcon
43
44 # generate the masks for grabbing trials of interest
```

Publishing reproducible Jupyter Notebooks

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

← → ↻ <https://codeocean.com/capsule/6314882/tree/v1> 🔍 ☆ 🌐 📄 👤 ⋮

CO Published Identifying Gene Expression Programs of Cell-type Identity and Cellular Activity with Sin... (Dylan Kotliar et al.)

Capsule File Edit View Tabs Settings Help

Files

- Core Files
- environment 2.01 KB
- code 21.32 MB
 - analysis 21.23 MB
 - code 79.26 KB
 - Download_Pregenerated_Int... 3.55 KB
 - install_utils.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
 - Part1_Simulations 14.15 GB
 - Part2_Organoids 2.02 GB
 - Part3_VisualCortex 2.92 GB
 - LICENSE 6.4 KB
 - .gitignore 7 B
- Other Files

Commands

Tabs

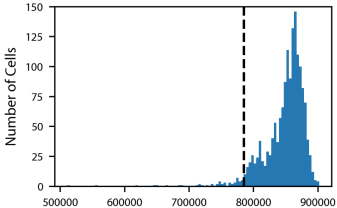
README.md x Step1_Preproc x Step1_Simulat x Step2_Preproc x

in the previous filters

```
In [16]: TPM = pd.read_csv('../.../data/Part3_VisualCortex/GSE71585_RefSeq_TPM.csv', 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')
```



Re-run

Timeline

Nov 20, 2018 Published Version 1.0 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_dCA.h... 323.1 KB
 - Step3c_Run_cLDA... 324.23 KB
 - Step3d_Run_PCA... 256.91 KB
 - Step3e_Run_Clust... 273.09 KB
 - Step3f_RefSeq_Spect... 258.71 KB

Cloud Workstations

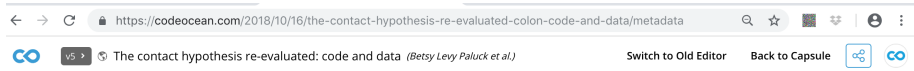
- <https://codeocean.com/capsule/8962292/tree/v2>
- Jupyter
- JupyterLab
- Rstudio

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/capsule/8235972/tree/v6>
- Will have a DOI and link to your article's metadata



Basic Info

Language **Stata**

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

License Info

Software License [MIT license](#)

Data License [No Rights Reserved \(CC0\)](#)

Associated Publication

DOI <https://doi.org/10.1017/bpp.2018.25>

Title [The contact hypothesis re-evaluated](#)

Publication Date **July 2018**

Journal/Conference **Behavioural Public Policy**

Funded by **National Science Foundation**

Grant Number **1322356**

Citation **PALUCK, ELIZABETH LEVY, SETH A. GREEN, DONALD P. GREEN. "The contact hypothesis re-evaluated." Behavioural Public Policy (2018): 1-30**

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>:

[//ieeexplore.ieee.org/document/8410389/algorithms#algorithms](https://ieeexplore.ieee.org/document/8410389/algorithms#algorithms):

explore.ieee.org/document/8410389 67% ...

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

The screenshot shows the Code Ocean interface for a capsule titled "On Writing Reproducible and Interactive Papers" by Mandar Chitre. The interface includes a file explorer on the left with "README.md" and "run.sh" files. A central panel displays the "Editorial" content. On the right, there is a "Run" button and a list of files: "output" (27.61 KB), "editorial.pdf" (123.91 KB), and "editorial.tex" (27.05 KB). A "Reproducibility" sidebar is also visible on the far right.