Using Cloud Computing Resources for Big Data and Code Reproducibility

—

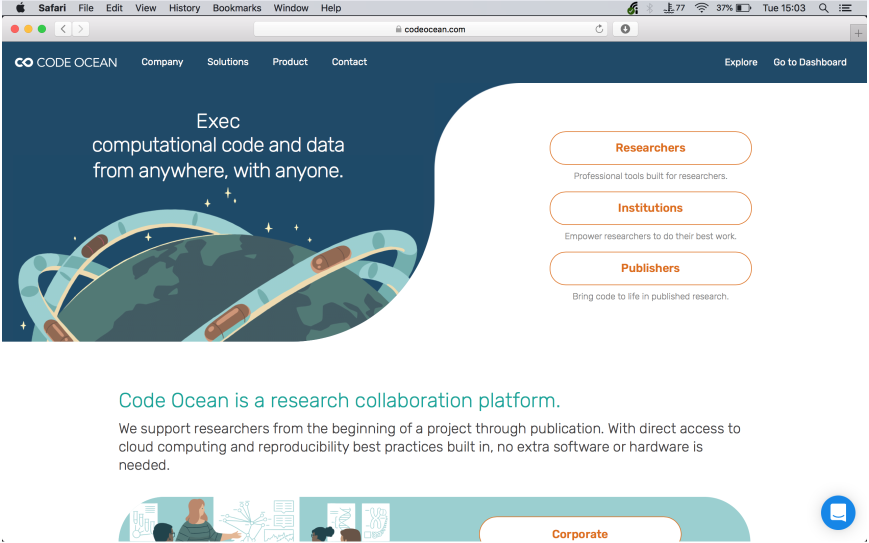
Code Ocean

Winter Institute in Data Science and Big Data

Simon Heuberger

9 January 2020

American University
Washington, DC



Exec
computational code and data
from anywhere, with anyone.

Professional tools built for researchers.

Empower researchers to do their best work.

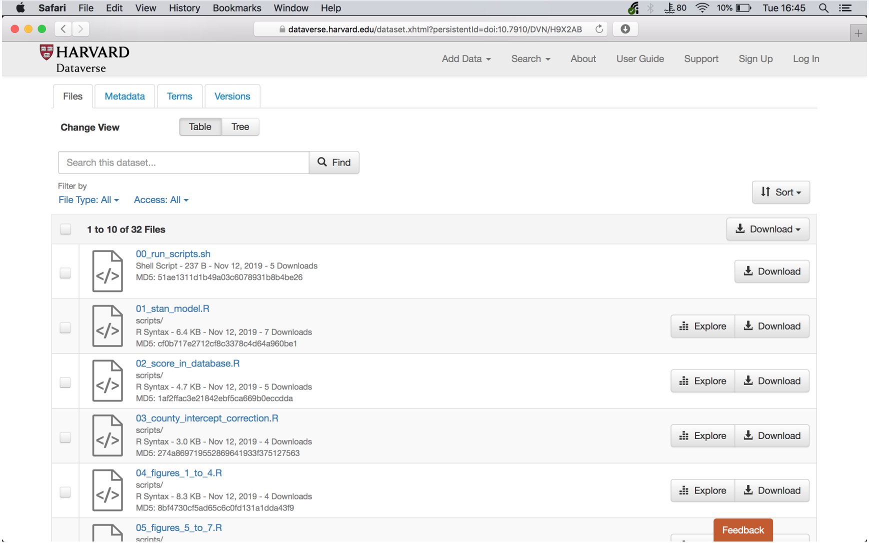
Bring code to life in published research.

Code Ocean is a research collaboration platform.

We support researchers from the beginning of a project through publication. With direct access to cloud computing and reproducibility best practices built in, no extra software or hardware is needed.

What is a capsule?

- Code + (optional) Data + Computational environment = The minimum required for computational reproducibility
- At the core of the capsule: Docker image
- Potential to revolutionize data reproduction and replication



dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/H9X2AB

□ 1 to 10 of 32 Files

□ 00_run_scripts.sh
Shell Script - 237 B - Nov 12, 2019 - 5 Downloads
MD5: 51ae1311d1b49a03c6078931b8b4be26

□ 01_stan_model.R
scripts/
R Syntax - 6.4 KB - Nov 12, 2019 - 7 Downloads
MD5: cf0b717e2712cf8c3378c4d64a960be1

□ 02_score_in_database.R
scripts/
R Syntax - 4.7 KB - Nov 12, 2019 - 5 Downloads
MD5: 1af2ffac3e21842ebf5ca669b0eccdda

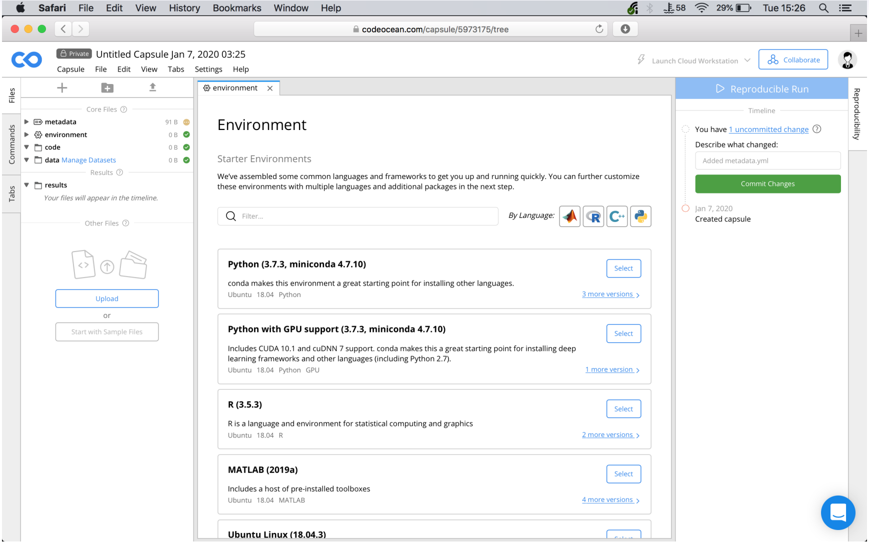
□ 03_county_intercept_correction.R
scripts/
R Syntax - 3.0 KB - Nov 12, 2019 - 4 Downloads
MD5: 274a869719552869641933f375127563

□ 04_figures_1_to_4.R
scripts/
R Syntax - 8.3 KB - Nov 12, 2019 - 4 Downloads
MD5: 8bf4730cf5ad65c6c0fd131a1dda43f9

05_figures_5_to_7.R
scripts/

What is a capsule?

- Code?
- Data?
- Computational environment?



You have 1 uncommitted change

Describe what changed: Added metadata.yml

Commit Changes

Jan 7, 2020
Created capsule

Environment

Starter Environments

We've assembled some common languages and frameworks to get you up and running quickly. You can further customize these environments with multiple languages and additional packages in the next step.

Filter... *By Language:*

Python (3.7.3, miniconda 4.7.10) Select

conda makes this environment a great starting point for installing other languages.

Ubuntu 18.04 Python — 3 more versions

Python with GPU support (3.7.3, miniconda 4.7.10) Select

Includes CUDA 10.1 and cuDNN 7 support. conda makes this a great starting point for installing deep learning frameworks and other languages (including Python 2.7).

Ubuntu 18.04 Python GPU — 1 more version

R (3.5.3) Select

R is a language and environment for statistical computing and graphics

Ubuntu 18.04 R — 2 more versions

MATLAB (2019a) Select

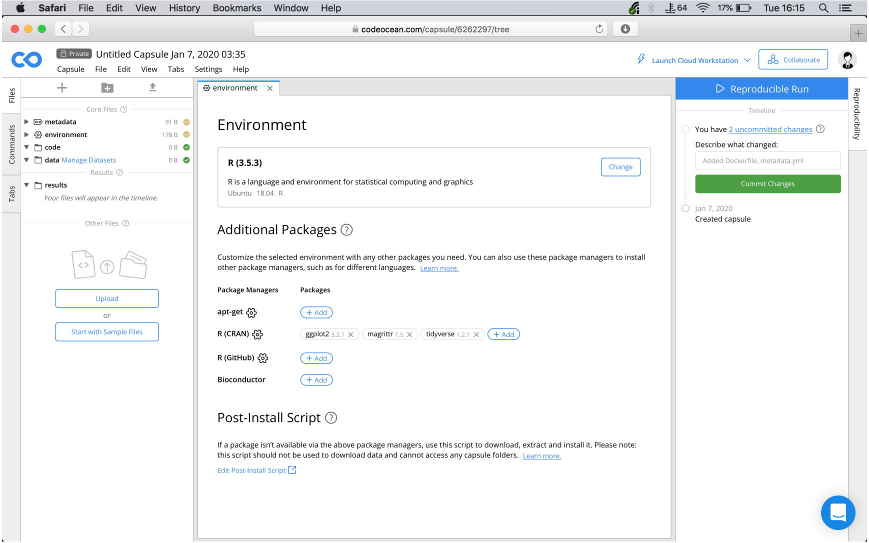
Includes a host of pre-installed toolboxes

Ubuntu 18.04 MATLAB — 4 more versions

Ubuntu Linux (18.04.3)

Computational environment

- Base environment
  - Operating system
  - Programming language
- Packages/Dependencies



Untitled Capsule Jan 7, 2020 03:35

Your files will appear in the timeline.

Environment

R (3.5.3)

R is a language and environment for statistical computing and graphics

Ubuntu 18.04 R

Additional Packages

Customize the selected environment with any other packages you need. You can also use these package managers to install other package managers, such as for different languages. Learn more.

Package Managers	Packages
apt-get	+ Add
R (CRAN)	ggplot2 3.2.1 ✕ magrittr 1.5 ✕ tidyverse 1.2.1 ✕ + Add
R (GitHub)	+ Add
Bioconductor	+ Add

Post-Install Script

If a package isn't available via the above package managers, use this script to download, extract and install it. Please note: this script should not be used to download data and cannot access any capsule folders. Learn more.

Edit Post-Install Script

You have 2 uncommitted changes

Describe what changed:

Added Dockerfile, metadata.yml

Jan 7, 2020

Created capsule

Example: Actual capsule

Exercises

On our GitHub repository: 08-03-cloud-authors_how_to.pdf

1. Create a Code Ocean account at https://codeocean.com
2. Create a new capsule
3. Add R as the base environment
4. Add packages along with their specific versions
5. Upload data
6. Upload one .R file
   - The file should read in the data
   - The file should create and save one plot with ggplot2
7. Create a run script for the uploaded .R file
8. Upload a very rudimentary Readme file
9. Commit the changes
10. Execute one successful reproducible run

Extra exercises

- Add `Python 3` to your `R` base environment via `apt-get`
- Include a full log of the files in `run`
- Save your plot to the subfolder `/figures`
- Create another new capsule by importing your `Git` repository