Open Sourcing Reproducibility: ReproZip

Vicky Steeves | NYU Division of Libraries & Center for Data Science

In collaboration with Remi Rampin. Fernando Chirigati, Juliana Freire, & Dennis Shasha

ReproZip tries to solve…

Workload & Time Challenges

It is a time commitment to get data and code ready to share, and to share it

Otherwise known as…

the Incentive Problem

Reproducibility takes time, and is not always valued by the academic reward structure

“Insufficient time is the main reason why scientists do not make their data and experiment available and reproducible.”
Carol Tenopir, Beyond the PDF2 Conference

“77% claim that they do not have time to document and clean up the code.”
Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

ReproZip tries to solve...

Technical Obsolescence

Technology changes affect the reproducibility

Normative Dissonance[1]

Espoused values don't always match practice

Otherwise known as...

the Pipeline Problem

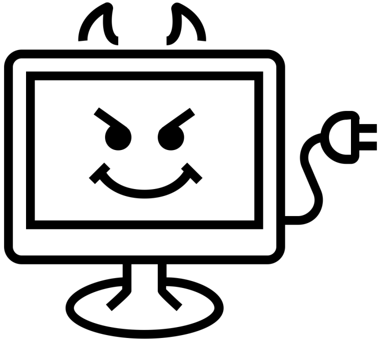
Reproducibility requires skills that are often not included in most curriculums!

"It would require huge amount of effort to make our code work with the latest versions of these tools." Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

[1] https://www.ncbi.nlm.nih.gov/pubmed/19385804

Big Challenge ReproZip solves...

Dependency Hell



You cannot expect people to find all the chains of dependencies!

You cannot expect people to install all the dependencies and run your code smoothly!

Gap: tools that can automatically capture all the dependencies in the original environment and automatically set them up in another environment

Even if runnable, results may differ

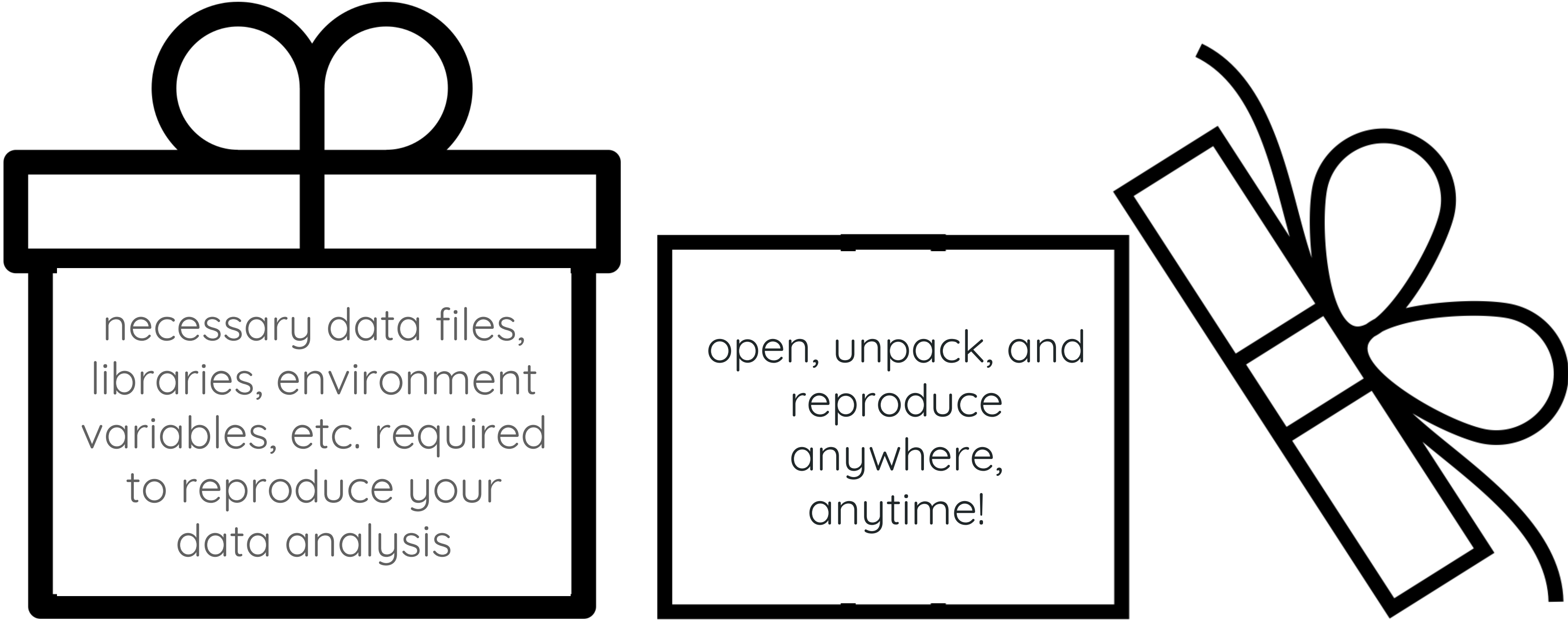
The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements, June 1, 2012, http://dx.doi.org/10.1371/journal.pone.0038234

> We investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). **Significant differences** were revealed between **FreeSurfer version v5.0.0 and the two earlier versions.** [...]
> About a factor two **smaller differences were detected between Macintosh and Hewlett-Packard** workstations and between **OSX 10.5 and OSX 10.6.**

Q from Researchers:

How do I make sure my work is reproducible without **spending all my time on it??**

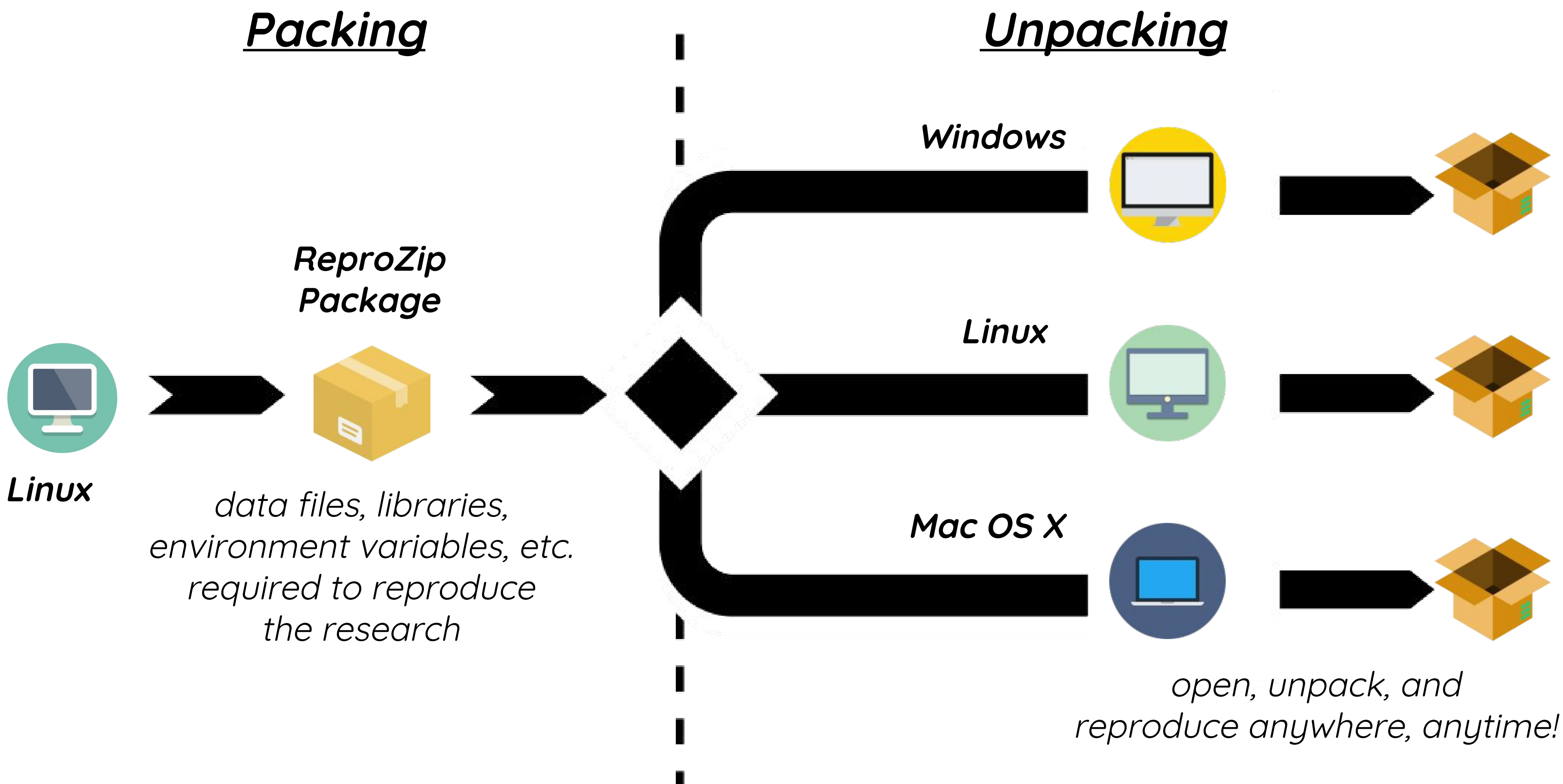
ReproZip, the Reproducibility Packer!



necessary data files, libraries, environment variables, etc. required to reproduce your data analysis

open, unpack, and reproduce anywhere, anytime!

ReproZip: Reproducibility in 2 Steps



Packing

Unpacking

Windows

ReproZip Package

Linux

Linux

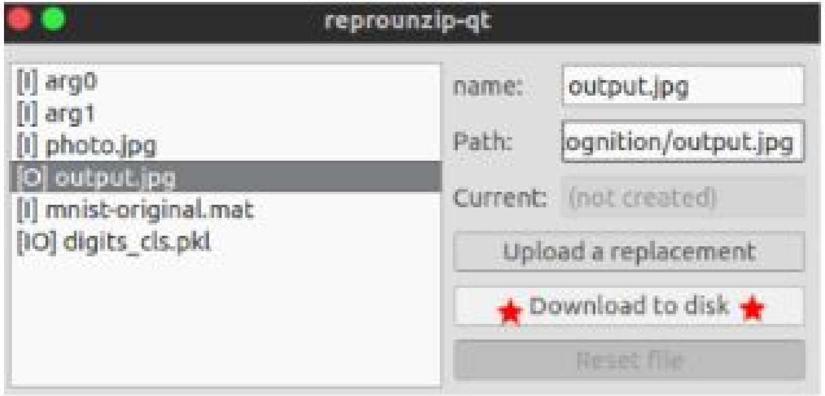
Mac OS X

data files, libraries, environment variables, etc. required to reproduce the research

open, unpack, and reproduce anywhere, anytime!

Extending the original work is also simple!

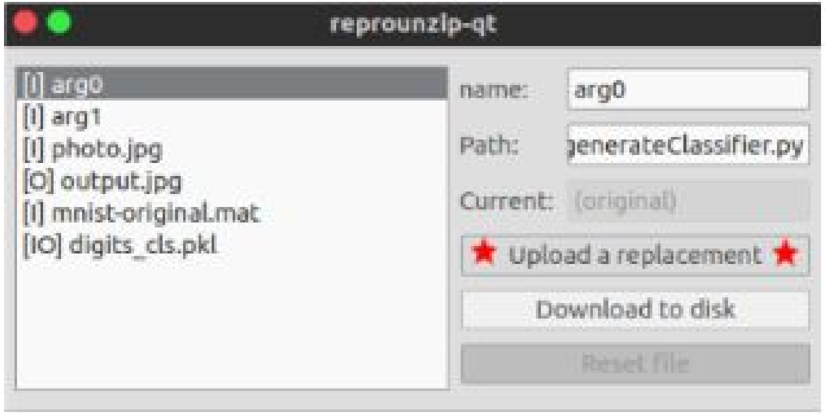
Download Ouput



[I] arg0
[I] arg1
[I] photo.jpg
[O] output.jpg
[I] mnist-original.mat
[IO] digits_cls.pkl

name: output.jpg
Path: ognition/output.jpg
Current: (not created)

Upload New Inputs



[I] arg0
[I] arg1
[I] photo.jpg
[O] output.jpg
[I] mnist-original.mat
[IO] digits_cls.pkl

name: arg0
Path: jenerateClassifier.py
Current: (original)

Unpackers for ReproZip

Unpacking

directory

unpacks and reproduces from a single directory
(Linux)

vagrant

unpacks in a virtual machine using Vagrant
(Linux, Mac OS X, Windows)

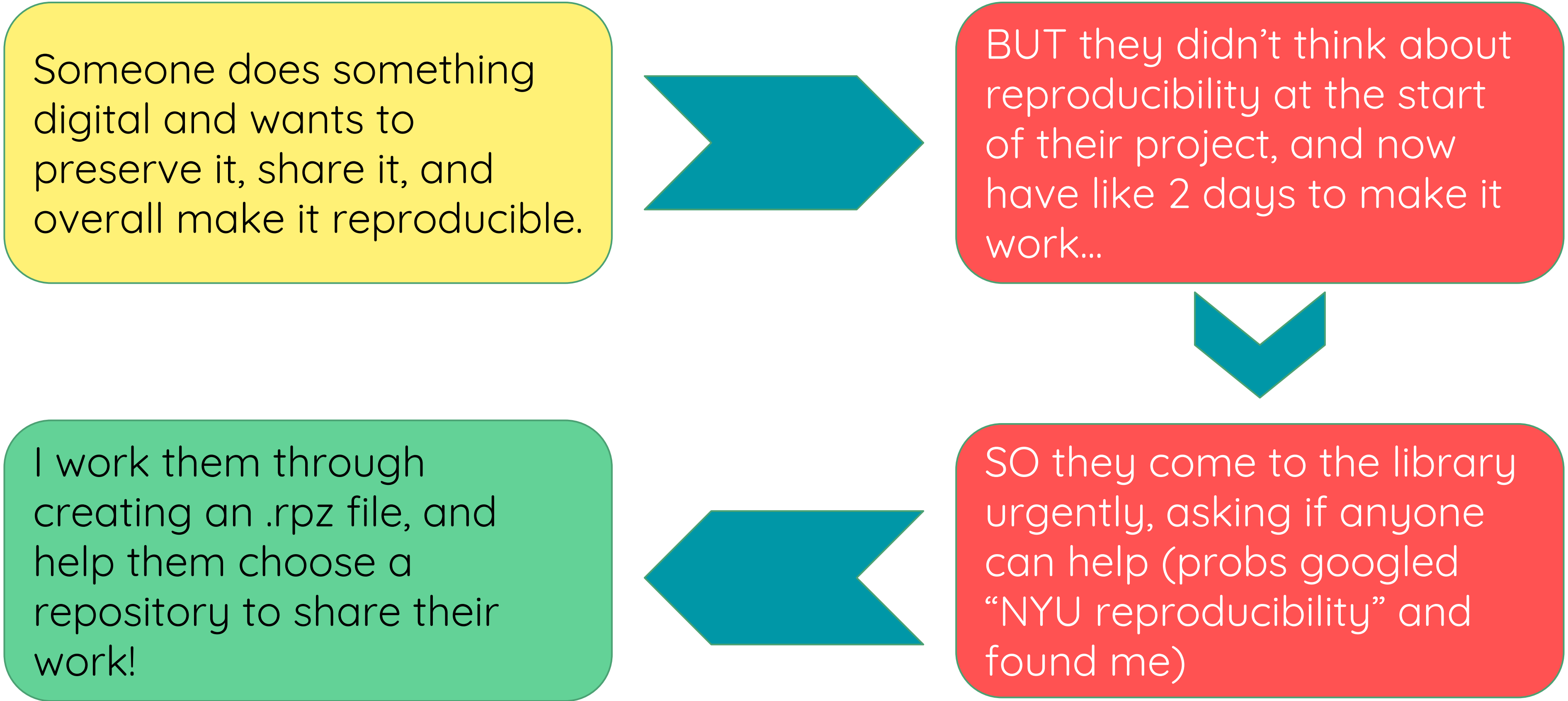
chroot

unpacks in a single directory and builds a full system environment
(Linux)

docker

unpacks in a Docker image
(Linux, Mac OS X, Windows)

Typical Workflow



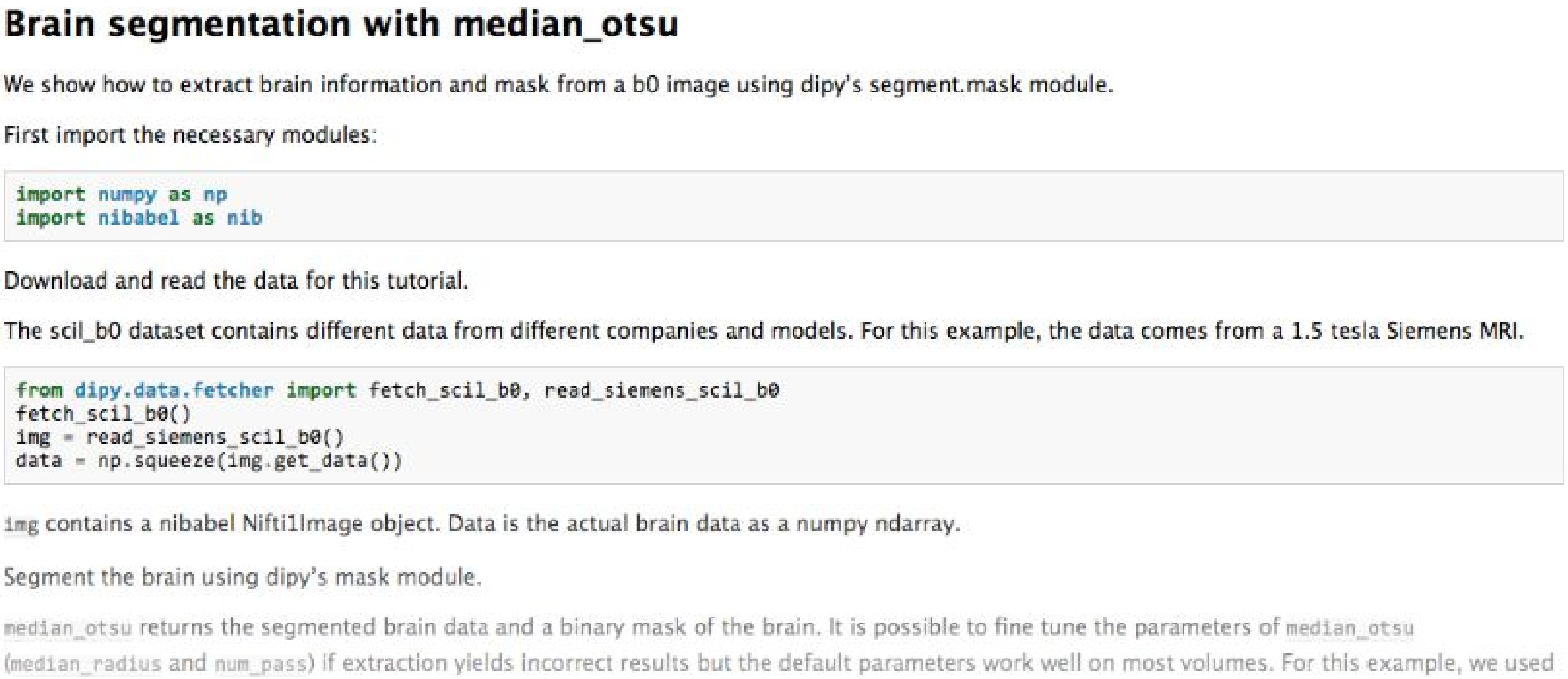
Someone does something digital and wants to preserve it, share it, and overall make it reproducible.

BUT they didn’t think about reproducibility at the start of their project, and now have like 2 days to make it work...

SO they come to the library urgently, asking if anyone can help (probs googled “NYU reproducibility” and found me)

I work them through creating an .rpz file, and help them choose a repository to share their work!

EXAMPLE 1: Image Analysis & Jupyter Notebooks



Brain segmentation with median_otsu

We show how to extract brain information and mask from a b0 image using dipy's segment.mask module.

First import the necessary modules:

```
import numpy as np
import nibabel as nib
```

Download and read the data for this tutorial.

The scil_b0 dataset contains different data from different companies and models. For this example, the data comes from a 1.5 tesla Siemens MRI.

```
from dipy.data.fetcher import fetch_scil_b0, read_siemens_scil_b0
fetch_scil_b0()
img = read_siemens_scil_b0()
data = np.squeeze(img.get_data())
```

img contains a nibabel Nifti1Image object. Data is the actual brain data as a numpy ndarray.

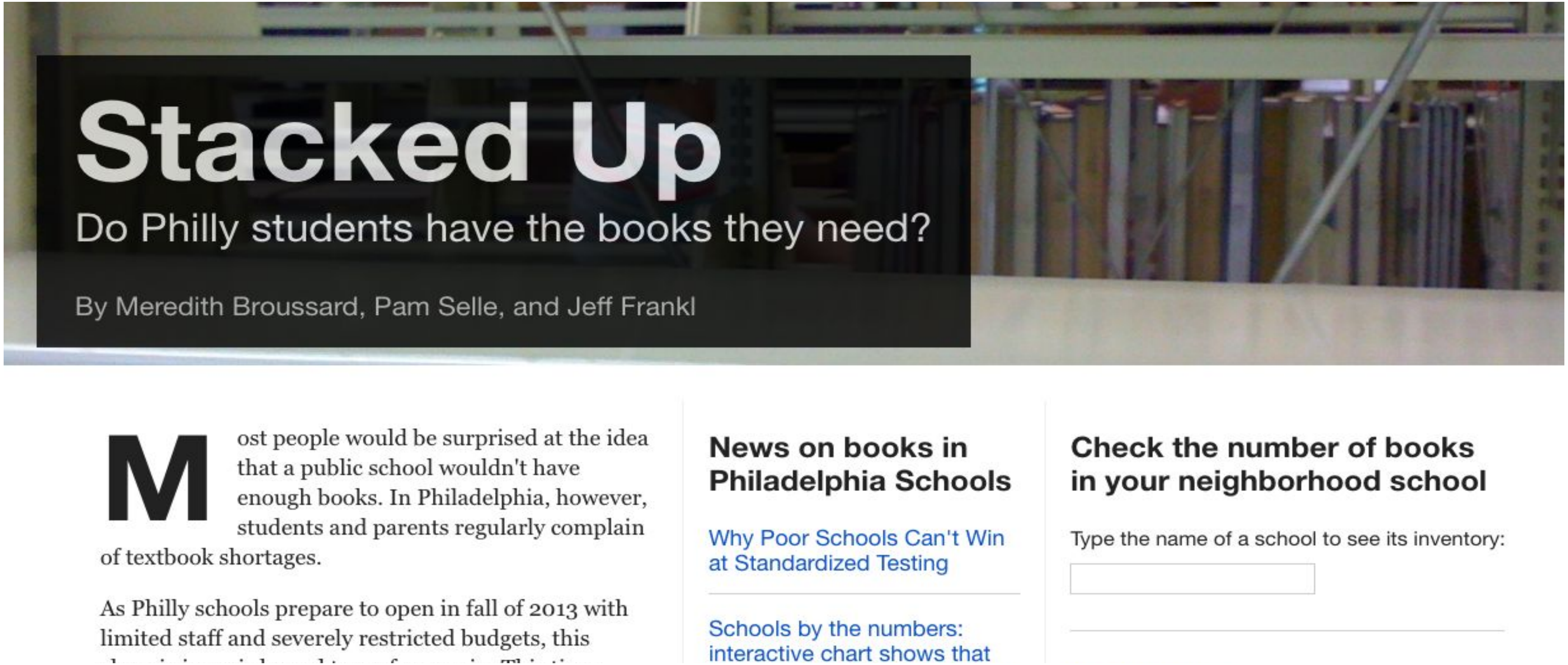
Segment the brain using dipy's mask module.

median_otsu returns the segmented brain data and a binary mask of the brain. It is possible to fine tune the parameters of median_otsu (median_radius and num_pass) if extraction yields incorrect results but the default parameters work well on most volumes. For this example, we used

Original Experiment: http://nipy.org/dipy/examples_built/brain_extraction_dwi.html | 2GB

ReproZip Package: brain_segmentation.rpz | 47 MB

EXAMPLE 2: Packing Research App & Unpacking it to deploy on AWS!



Stacked Up

Do Philly students have the books they need?

By Meredith Broussard, Pam Selle, and Jeff Frankl

Most people would be surprised at the idea that a public school wouldn't have enough books. In Philadelphia, however, students and parents regularly complain of textbook shortages.

As Philly schools prepare to open in fall of 2013 with limited staff and severely restricted budgets, this

News on books in Philadelphia Schools

Why Poor Schools Can't Win at Standardized Testing

Schools by the numbers: interactive chart shows that

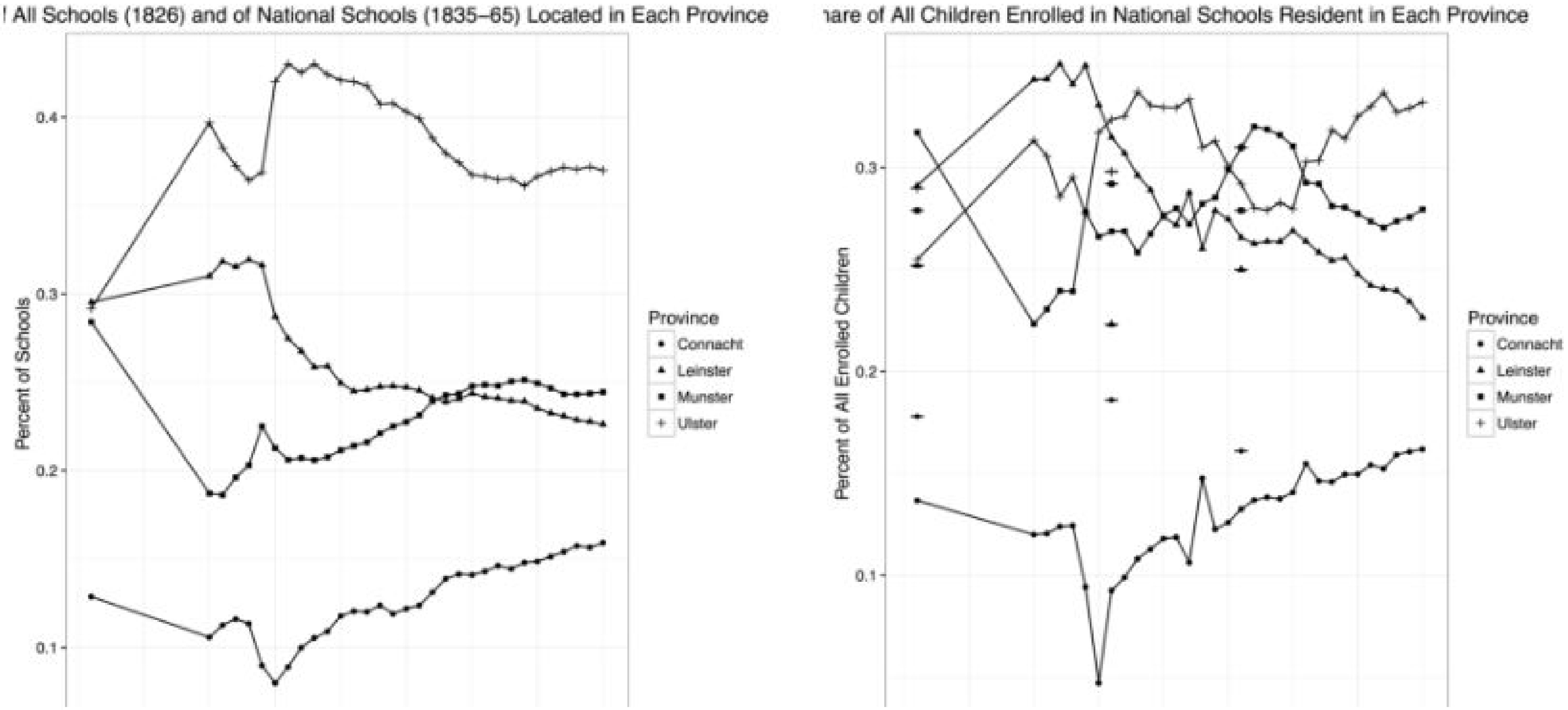
Check the number of books in your neighborhood school

Type the name of a school to see its inventory:

Original Experiment: https://github.com/merbroussard/sdp_curricula

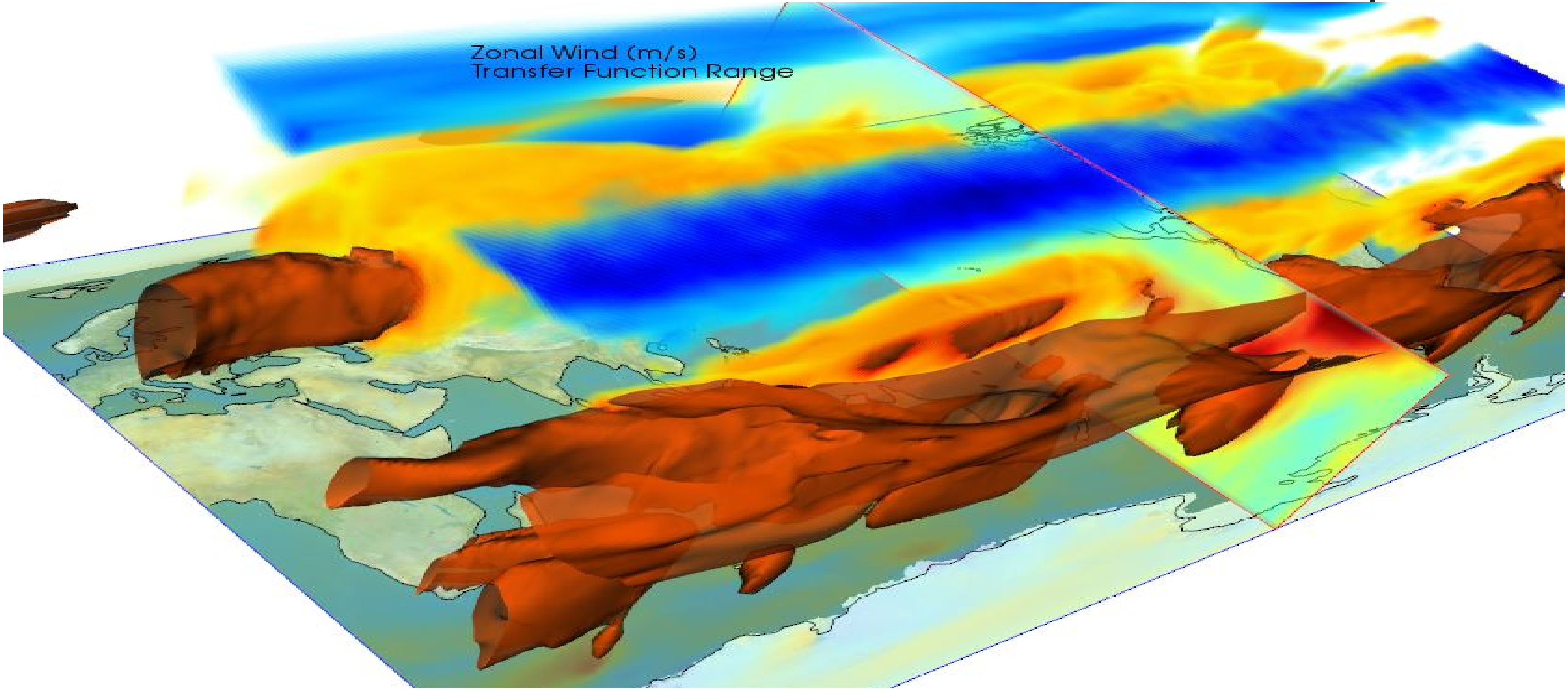
ReproZip Package: stacked-up.rpz

EXAMPLE 3: Publication-Ready Plots with R



Original Work: https://osf.io/uh46c/ & **ReproZip Package:** irish-schools.rpz | 17 MB

BONUS: 3D Visualization of Wind on Map



Original Experiment: https://uvcdat.llnl.gov/examples/vcs3D_multiplot.html

ReproZip Package: https://osf.io/93rvc

ReproZip can pack:

Data analysis scripts / software (any language, you name it!)

Graphical tools

Interactive tools

Client-server applications (including databases)

Jupyter notebooks

MPI experiments (setting up the experiment is involved though...)

... and much more!

Current Use Cases:

Academic Use Cases

- Recommended by the Information Systems Journal, Reproducibility Section
- Recommended by the ACM SIGMOD Reproducibility Review
- Listed on the ACM Artifact Evaluation Process Guidelines

Other Use Cases

- Integrated as a component of CoRR
- Archiving data journalism apps, e.g.: Stacked Up

... and many more!

Other Resources for ReproZip

ReproZip Website:
https://reprozip.org

ReproZip Examples:
https://examples.reprozip.org

ReproZip GitHub:
https://github.com/ViDA-NYU/reprozip

ReproZip Mailing list:
reprozip-users@vgc.poly.edu

ReproZip YouTube Demos:

- General Demo: https://goo.gl/o1Hqrx
- Website packing: https://goo.gl/yMEOZJ
- Jupyter notebook: https://goo.gl/NvMHnw

ReproZip on Twitter:
https://goo.gl/d6NXoH

Thank You:

Rémi Rampin, main ReproZip developer who lets me add to his dev queue constantly.

Fernando Chirigati, ReproZip team member who let me poach some of his slides.

Juliana Freire, ReproZip PI and reproducibility master.

Questions?

Get this Presentation:
https://osf.io/z3yfp/

Email us:

- reprozip-users@vgc.poly.edu

  OR

- vicky.steeves@nyu.edu