Preserving and Reproducing Research with ReproZip

Fernando Chirigati

Doctoral Candidate and Research Assistant
New York University

*In collaboration with **Rémi Rampin**, **Vicky Steeves**, **Juliana Freire**, and **Dennis Shasha***

Re

To b

To h

To d

http:

To v

… it

The Annals of Applied Statistics

Info | Current issue | All issues | Search

Ann. Appl. Stat.
Volume 2, Number 2 (2008), 536-549.

← Previous article | TOC | Next article →

Should the Democrats move to the left on economic policy?

Andrew Gelman and Cexun Jeffrey Cai

Full-text: Open access

Enhanced PDF (222 KB)

"The results seemed perfectly reasonable."

Abstract | Article info and citation | First page | References

Abstract

Could John Kerry have gained votes in the 2004 Presidential election by more clearly distinguishing himself from George Bush on economic policy? At first thought, the logic of political preferences would suggest not: the Republicans are to the right of most Americans on economic policy, and so in a one-dimensional space with party positions measured with no error, the optimal strategy for the Democrats would be to stand infinitesimally to the left of the Republicans. The median voter theorem suggests that each party should keep its policy positions just barely distinguishable from the opposition.

In a multidimensional setting, however, or when voters vary in their perceptions of the parties' positions, a party can benefit from putting some

How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011

Nobel Winner Retracts Research Paper

Nobel Laureate Retracts Two Papers Unrelated to Her Prize

By KENNETH CHANG SEPT. 23, 2010

Email

Linda B. Buck

Medicine for d

red the 2004

TWITTER

LINKEDIN

Over half of psychology studies fail reproducibility test

Cancer researcher retracts 19 studies at once

e results.

with 4 comments

A former cancer biologist

retracted 19 papers from

Jin Cheng, who studies h

the Journal of Biological C

corrected another. All of

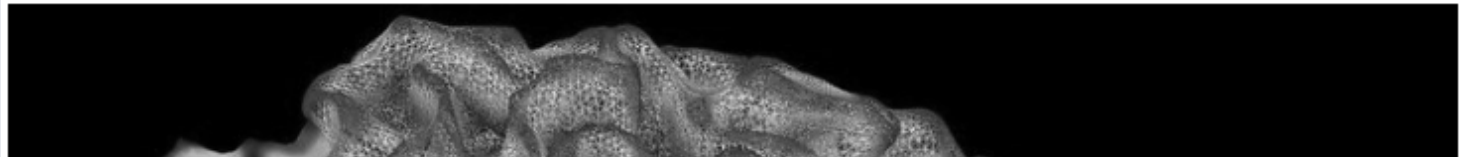
For example, here's the r

kinase/Akt pathway by a

and Src," a paper original

New Study Calls the Reliability of Brain Scan Research Into Question

Three million analyses point to a problem with fMRI brain activity studies



How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011

Nobel Winner Retracts Research Paper

Nobel Laureate Retracts Two Papers Unrelated to Her Prize

By KENNETH CHANG SEPT. 23, 2010

Email

TWITTER

LINKEDIN

Over half of psychology studies fail reproducibility

Cancer re

with 4 commer

A former cance
retracted 19 pa

Jin Cheng, who
the Journal of E
corrected anot

For example, h
kinase/Akt pat
and Src," a pap

results.

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

Author Affiliations

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved May 17, 2016 (received for review February 12, 2016)

A correction has been published

Abstract | Full Text | Authors & Info | Figures | SI | Metrics | Related Content | PDF | PDF + SI

Significance

… but there are a few obstacles!

Privacy / Confidentiality

Cultural Change

It's hard!

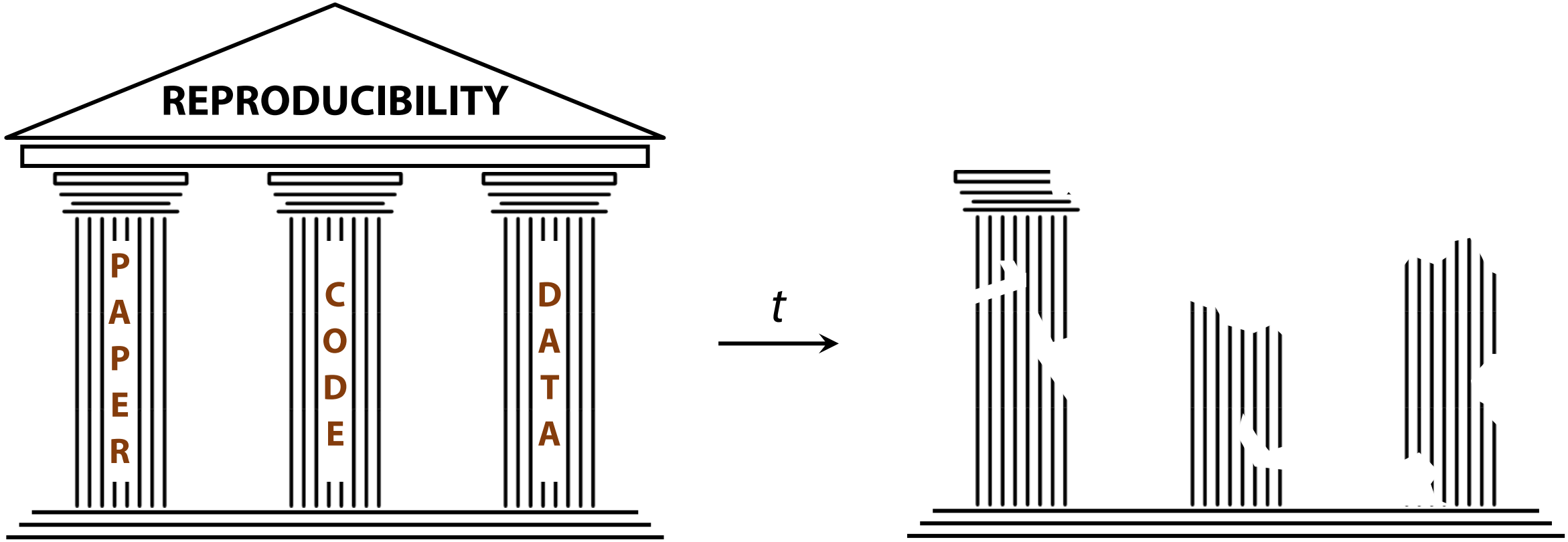
Even if runnable, results may differ!

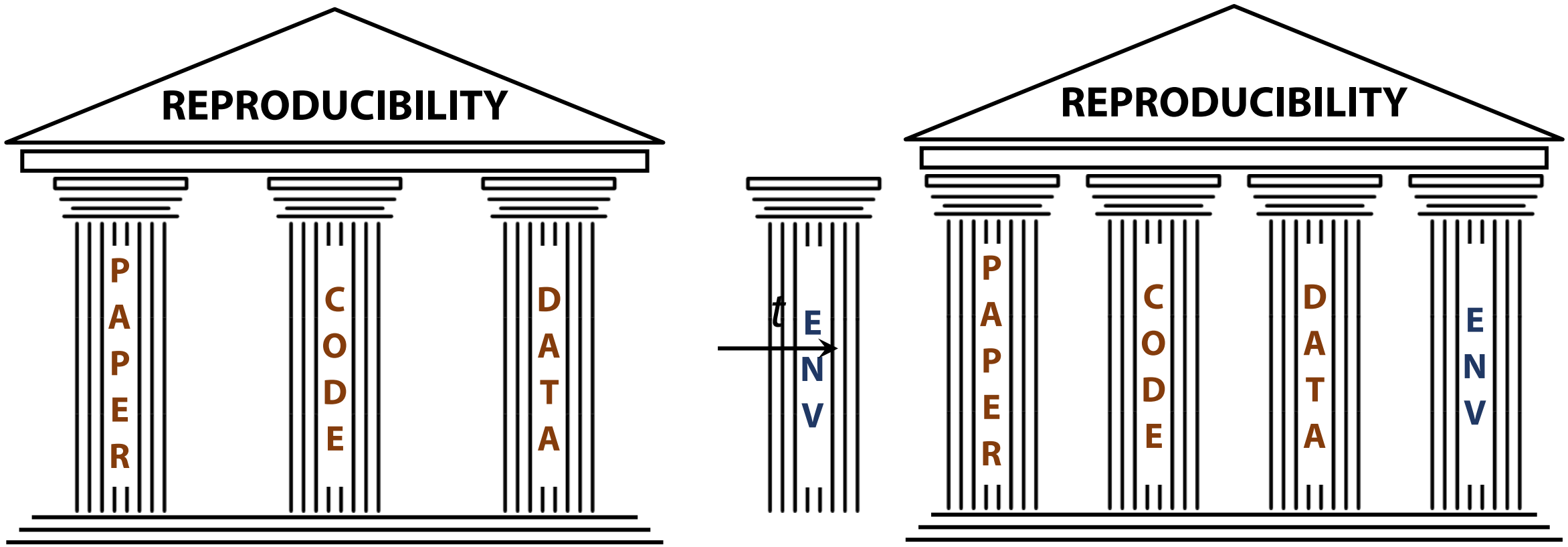
The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements
PLOS ONE, June 1, 2012

"[…] differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6."

Reproducibility of Neuroimaging Analyses Across Operating Systems
Front Neuroinform, April 24, 2015

"A first step to correct these reproducibility issues would be to use more precise representations of floating-point numbers."

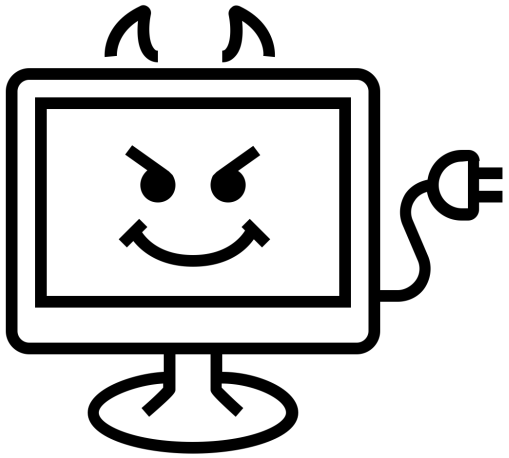

REPRODUCIBILITY
PAPER
CODE
DATA
t



Computational environment is as indispensable as the paper, code, and data for reproducibility!

But environments are hard to capture…

Dependency Hell

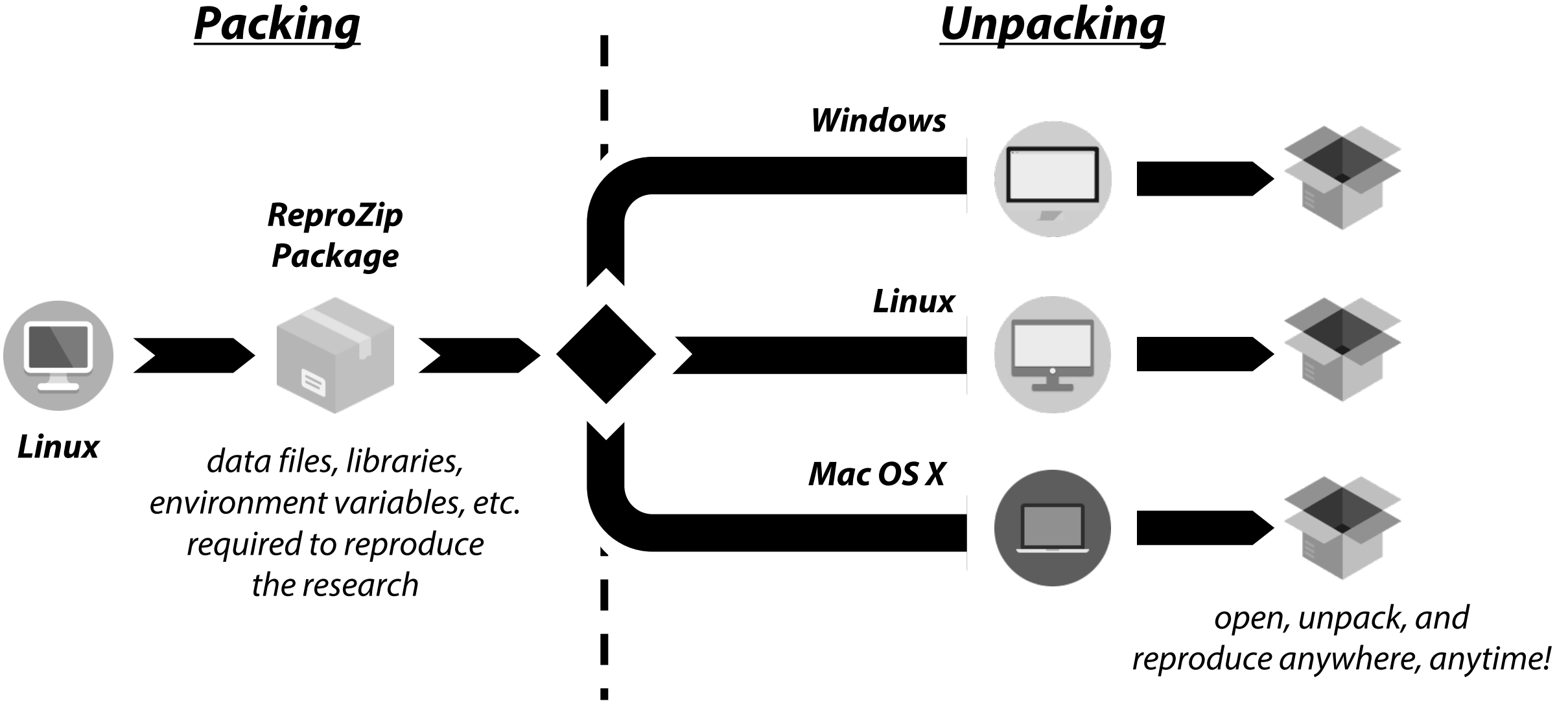


You cannot expect people to find all the chains of dependencies!

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*

ReproZip, the Reproducibility Packer!



1

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.

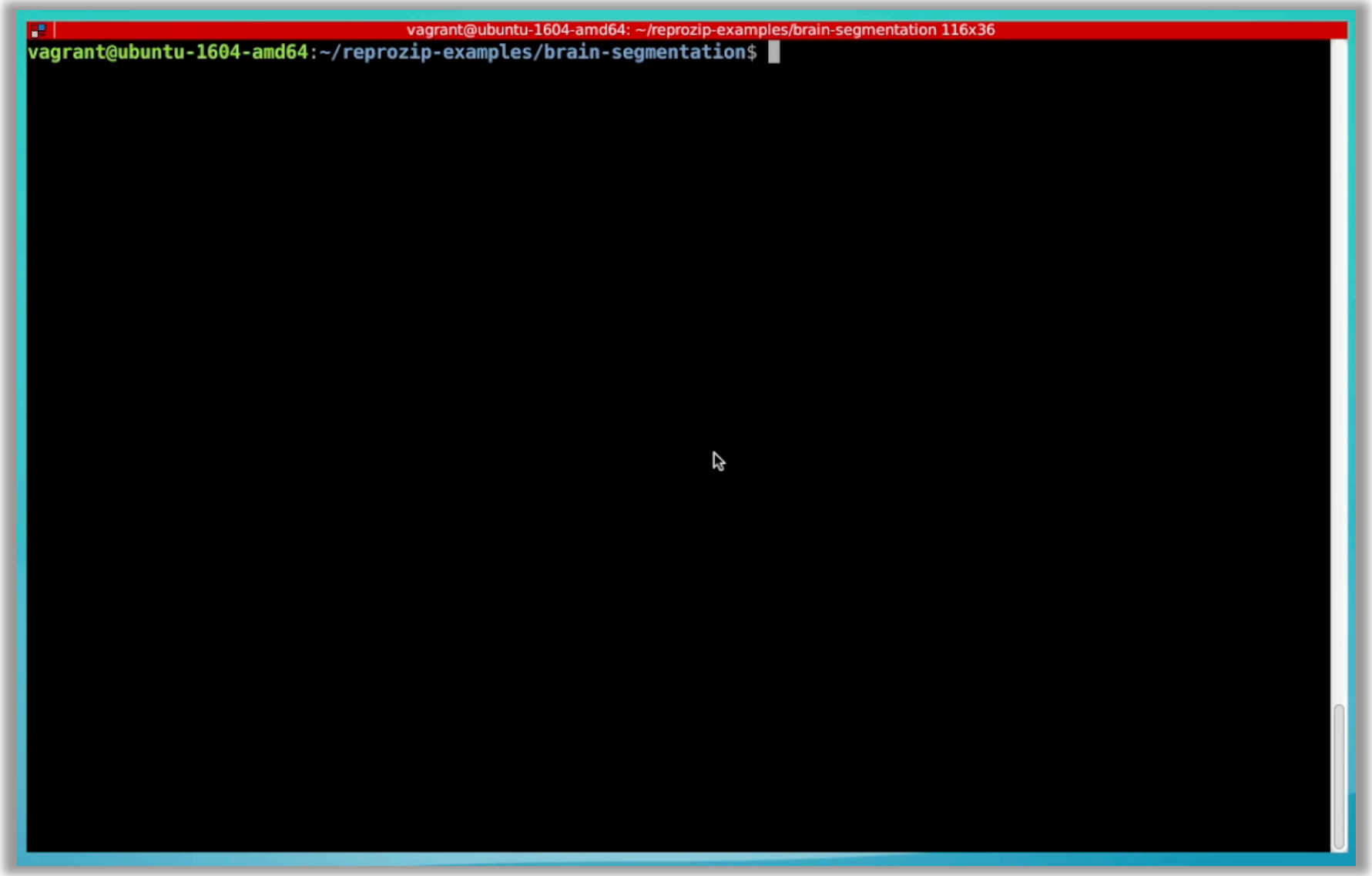
`median_otsu` returns the segmented brain data and a binary mask of the brain. It is possible to fine tune the parameters of `median_otsu` (`median_radius` and `num_pass`) if extraction yields incorrect results but the default parameters work well on most volumes. For this example, we used 2 as `median_radius` and 1 as `num_pass`

```
from dipy.segment.mask import median_otsu
b0_mask, mask = median_otsu(data, 2, 1)
```

Website: *http://nipy.org/dipy/examples_built/brain_extraction_dwi.html*

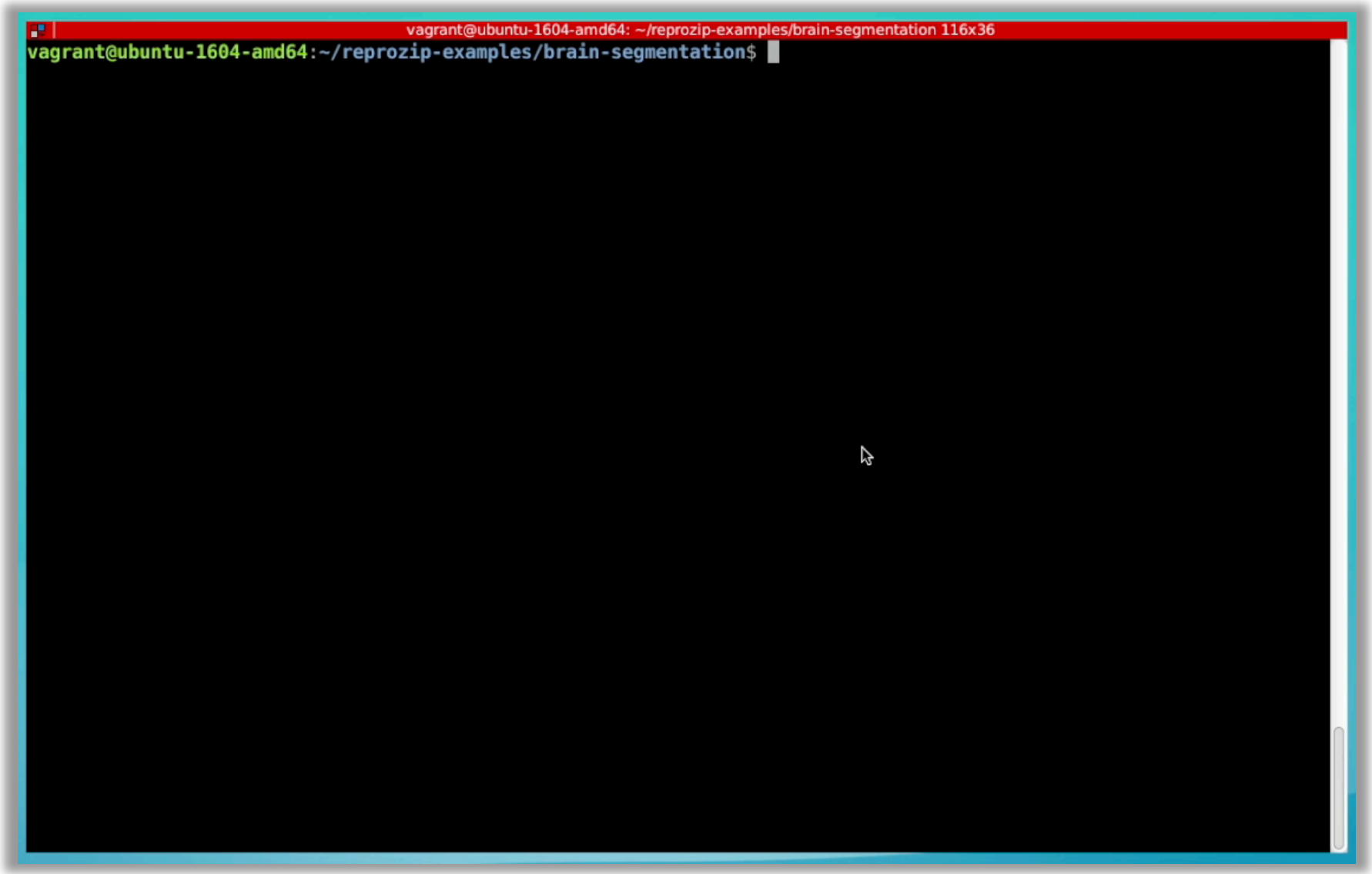
ReproZip: *brain-segmentation*

reprozip trace



Packing

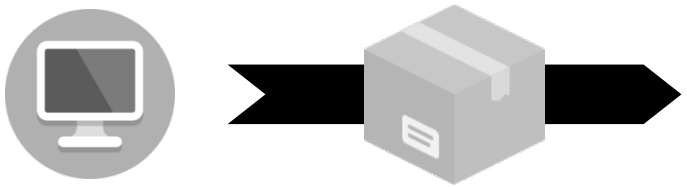
reprozip pack



Packing

brain-segmentation.rpz
47MB

VM
4GB

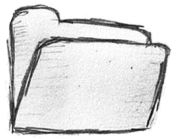


Linux

Mac OS X

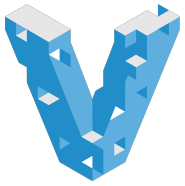
Unpackers

Unpacking



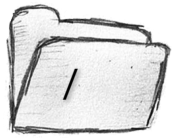
directory

unpacks and reproduces from a single directory ***(Linux)***



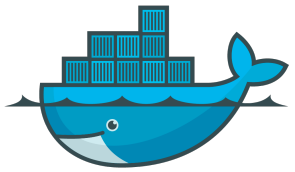
vagrant

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



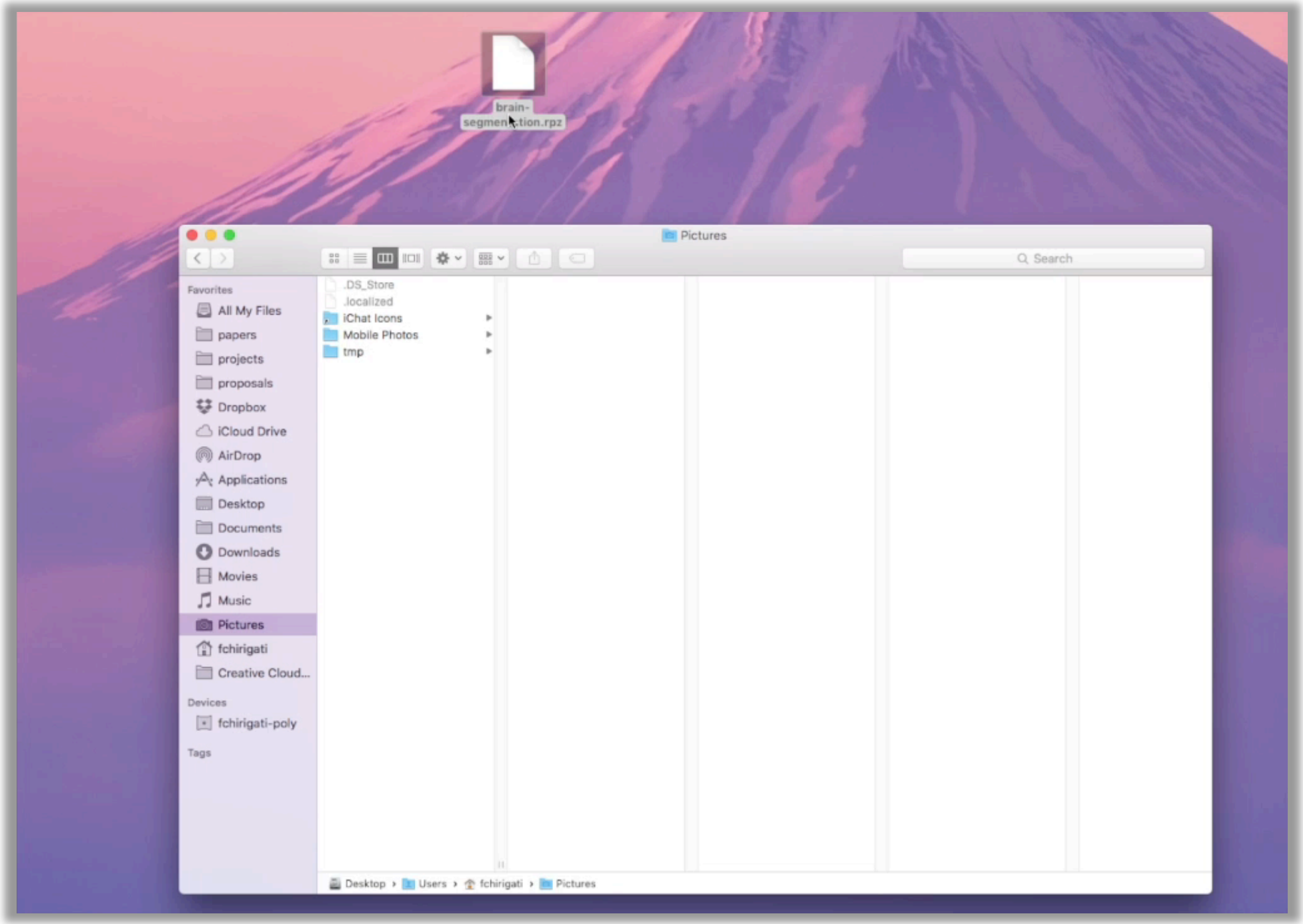
chroot

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

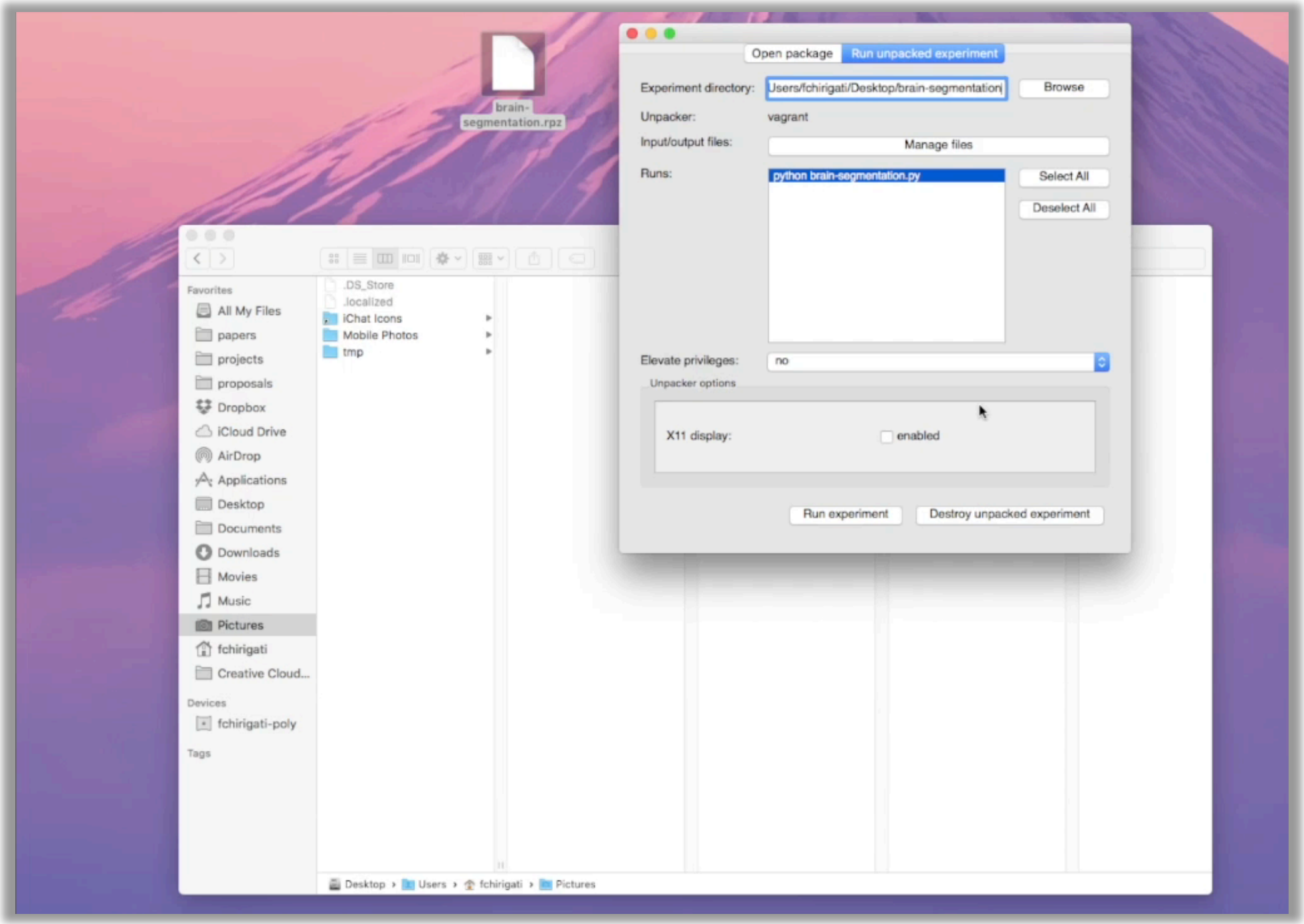


docker

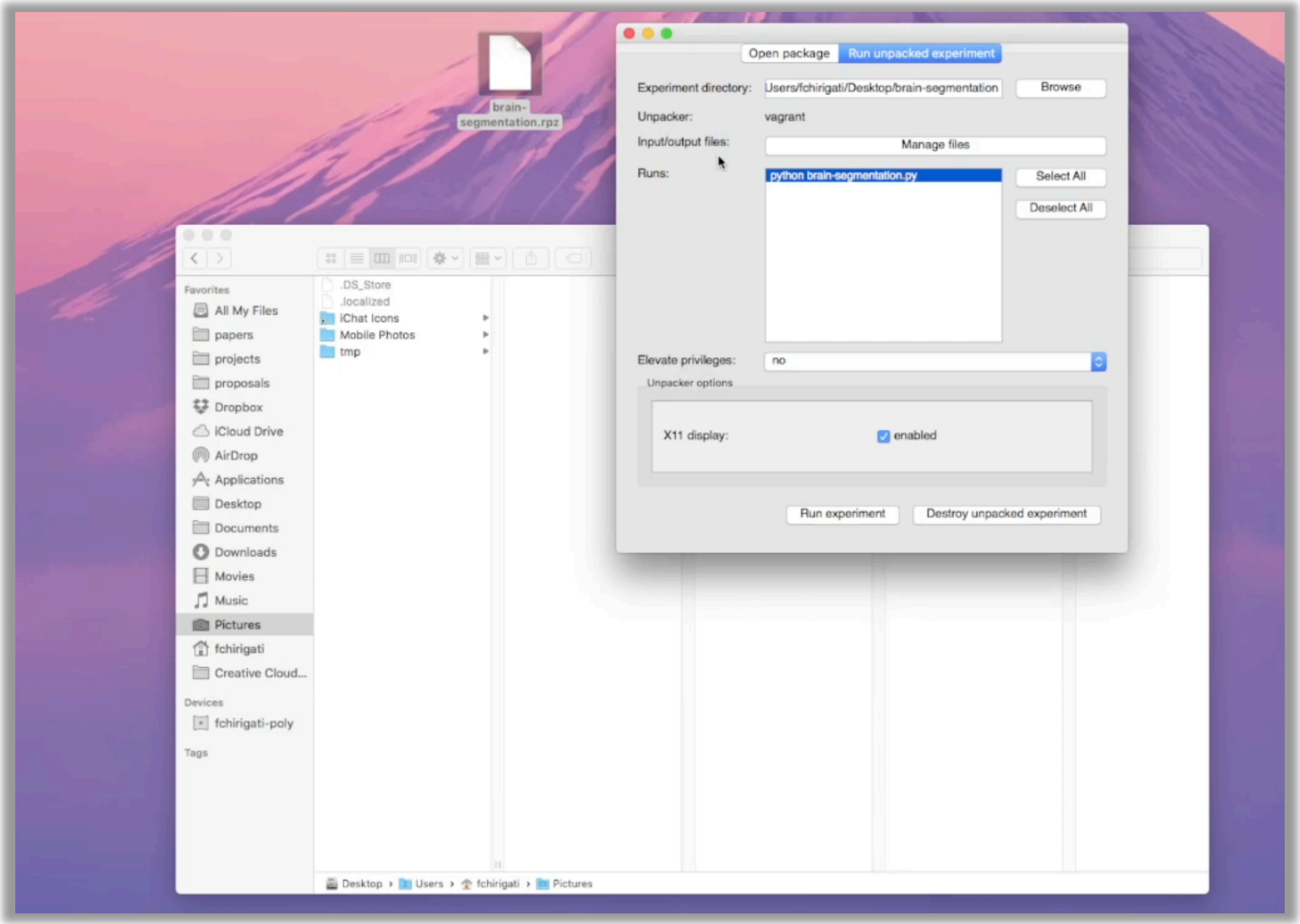
unpacks in a Docker image ***(Linux, Mac OS X, Windows)***


brain-
segmentation.rpz
Pictures
Search
Favorites
All My Files
papers
projects
proposals
Dropbox
iCloud Drive
AirDrop
Applications
Desktop
Documents
Downloads
Movies
Music
Pictures
fchirigati
Creative Cloud...
Devices
fchirigati-poly
Tags
.DS_Store
.localized
iChat Icons
Mobile Photos
tmp
Desktop › Users › fchirigati › Pictures

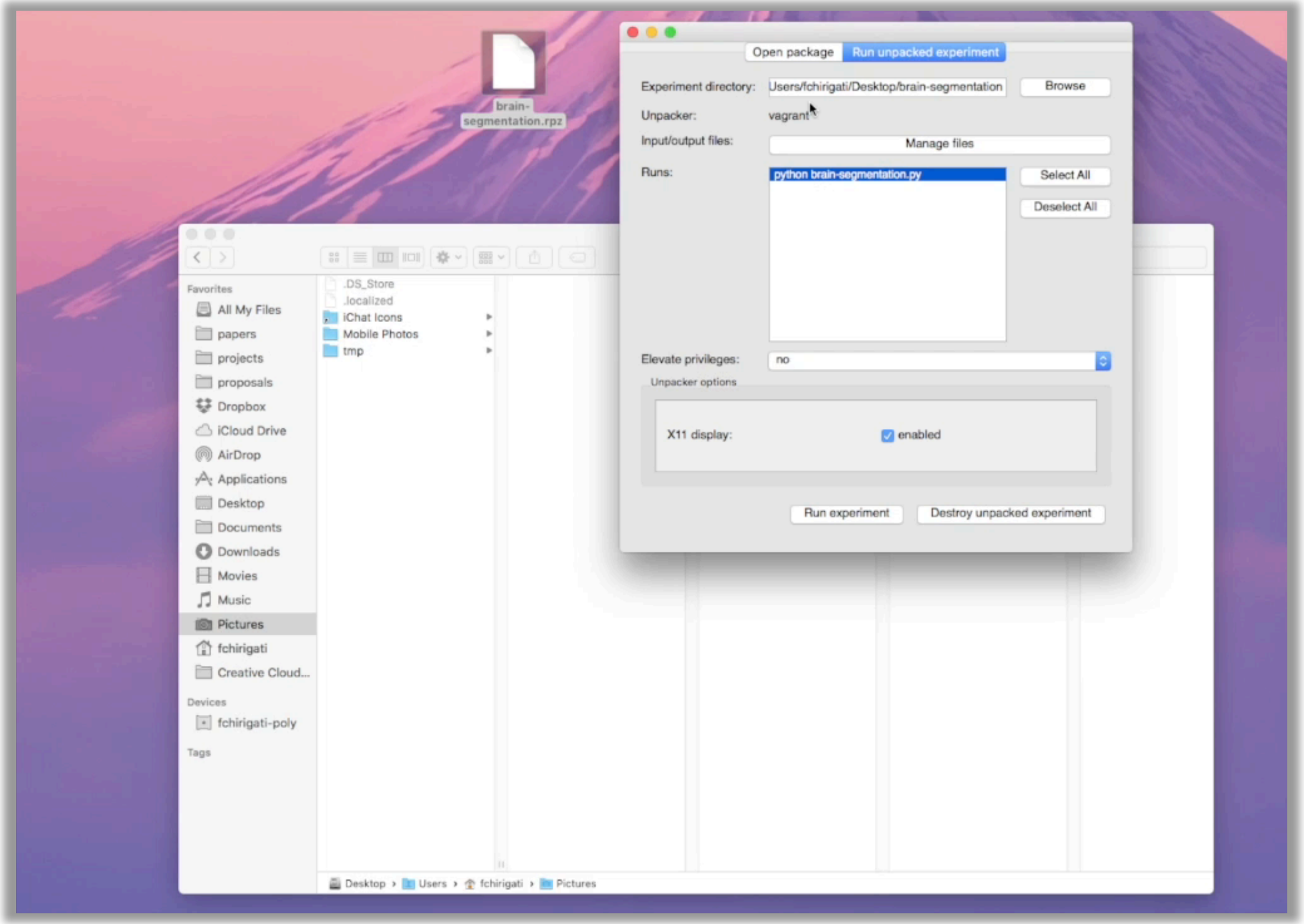
Unpacking


brain-segmentation.rpz
Open package
Run unpacked experiment
Experiment directory:
Users/fchirigati/Desktop/brain-segmentation
Browse
Unpacker:
vagrant
Input/output files:
Manage files
Runs:
python brain-segmentation.py
Select All
Deselect All
Elevate privileges:
no
Unpacker options
X11 display:
enabled
Run experiment
Destroy unpacked experiment
Favorites
All My Files
papers
projects
proposals
Dropbox
iCloud Drive
AirDrop
Applications
Desktop
Documents
Downloads
Movies
Music
Pictures
fchirigati
Creative Cloud...
Devices
fchirigati-poly
Tags
.DS_Store
.localized
iChat Icons
Mobile Photos
tmp
Desktop › Users › fchirigati › Pictures

Unpacking

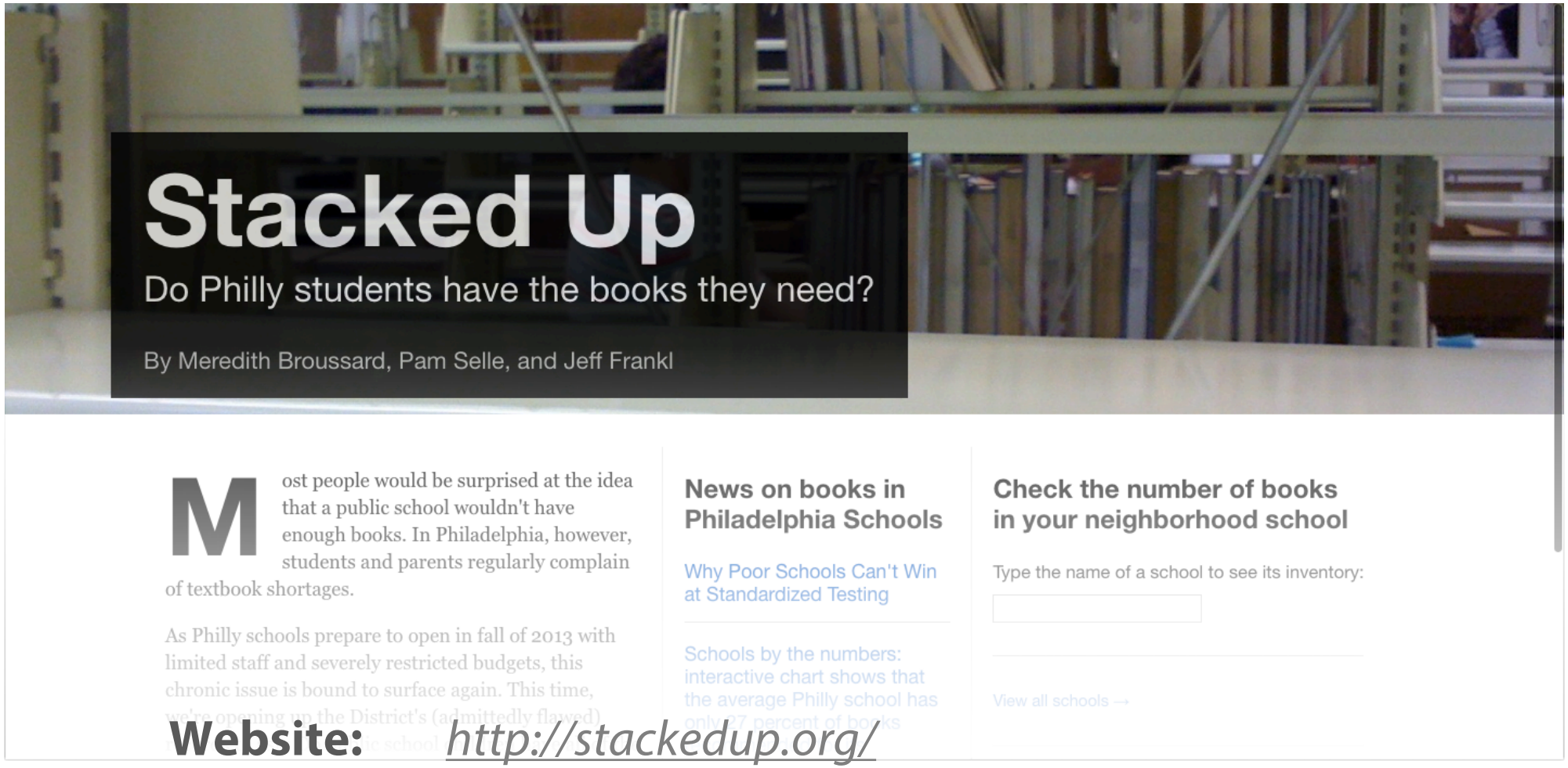

brain-segmentation.rpz
Open package
Run unpacked experiment
Experiment directory:
Users/fchirigati/Desktop/brain-segmentation
Browse
Unpacker:
vagrant
Input/output files:
Manage files
Runs:
python brain-segmentation.py
Select All
Deselect All
Elevate privileges:
no
Unpacker options
X11 display:
enabled
Run experiment
Destroy unpacked experiment
Favorites
All My Files
papers
projects
proposals
Dropbox
iCloud Drive
AirDrop
Applications
Desktop
Documents
Downloads
Movies
Music
Pictures
fchirigati
Creative Cloud...
Devices
fchirigati-poly
Tags
.DS_Store
.localized
iChat Icons
Mobile Photos
tmp
Desktop › Users › fchirigati › Pictures

Unpacking


brain-segmentation.rpz
Open package
Run unpacked experiment
Experiment directory:
Users/fchirigati/Desktop/brain-segmentation
Browse
Unpacker:
vagrant
Input/output files:
Manage files
Runs:
python brain-segmentation.py
Select All
Deselect All
Elevate privileges:
no
Unpacker options
X11 display:
enabled
Run experiment
Destroy unpacked experiment
Favorites
All My Files
papers
projects
proposals
Dropbox
iCloud Drive
AirDrop
Applications
Desktop
Documents
Downloads
Movies
Music
Pictures
fchirigati
Creative Cloud...
Devices
fchirigati-poly
Tags
.DS_Store
.localized
iChat Icons
Mobile Photos
tmp
Desktop › Users › fchirigati › Pictures

Unpacking

2



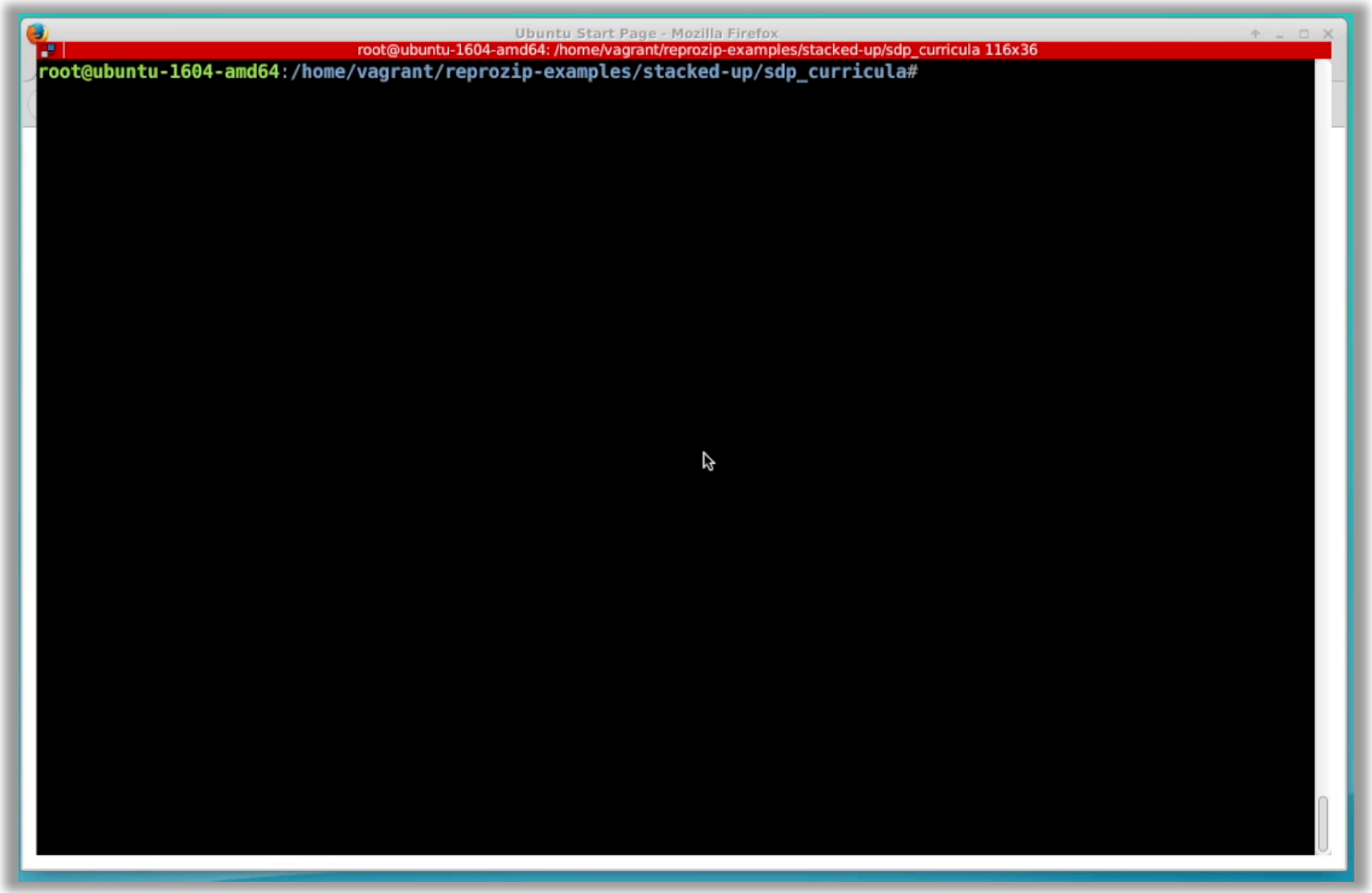
Website: *http://stackedup.org/*

GitHub: *https://github.com/merbroussard/sdp_curricula*

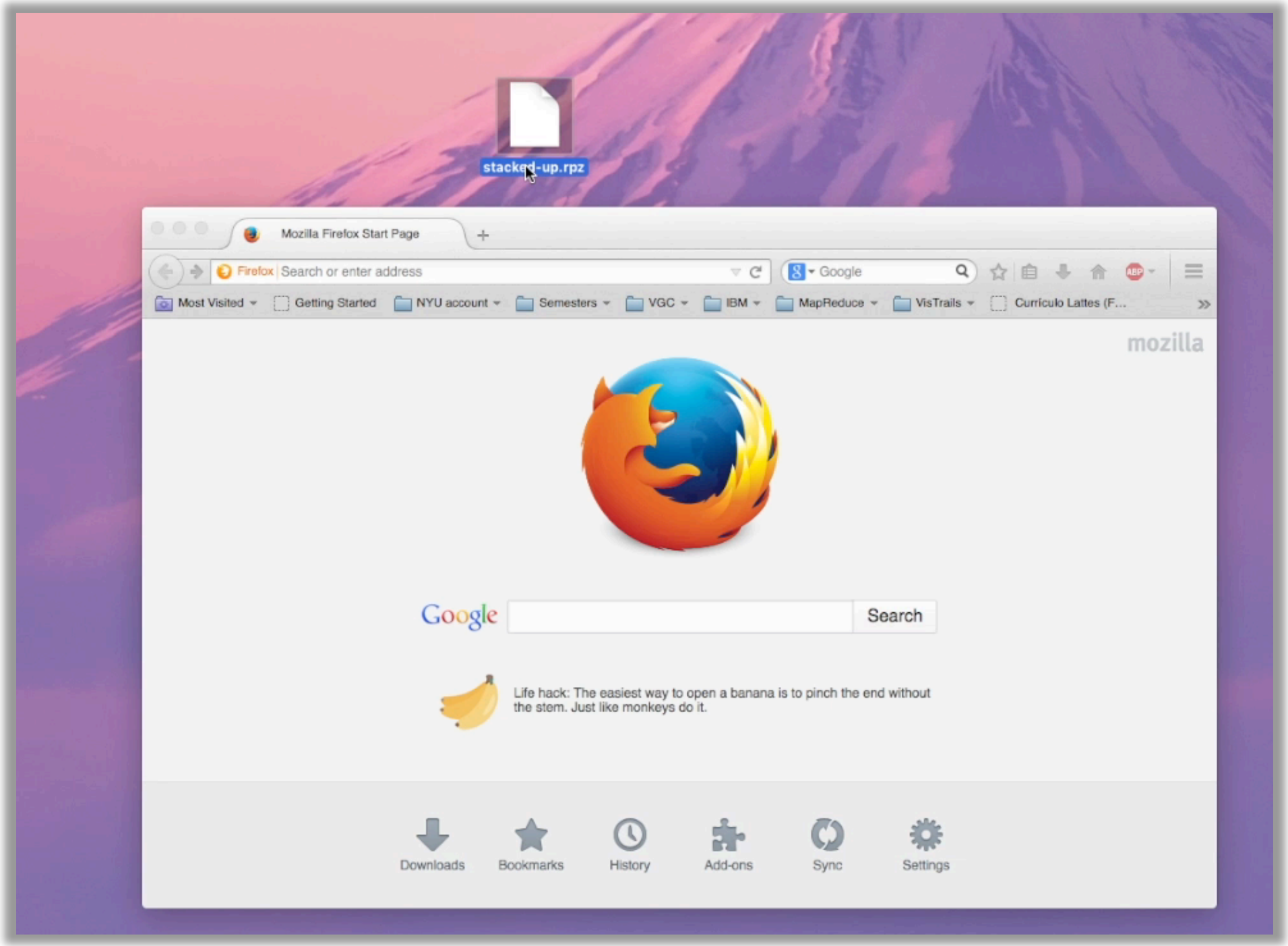
Data: *https://github.com/merbroussard/philasdbudget*

ReproZip: *stacked-up*

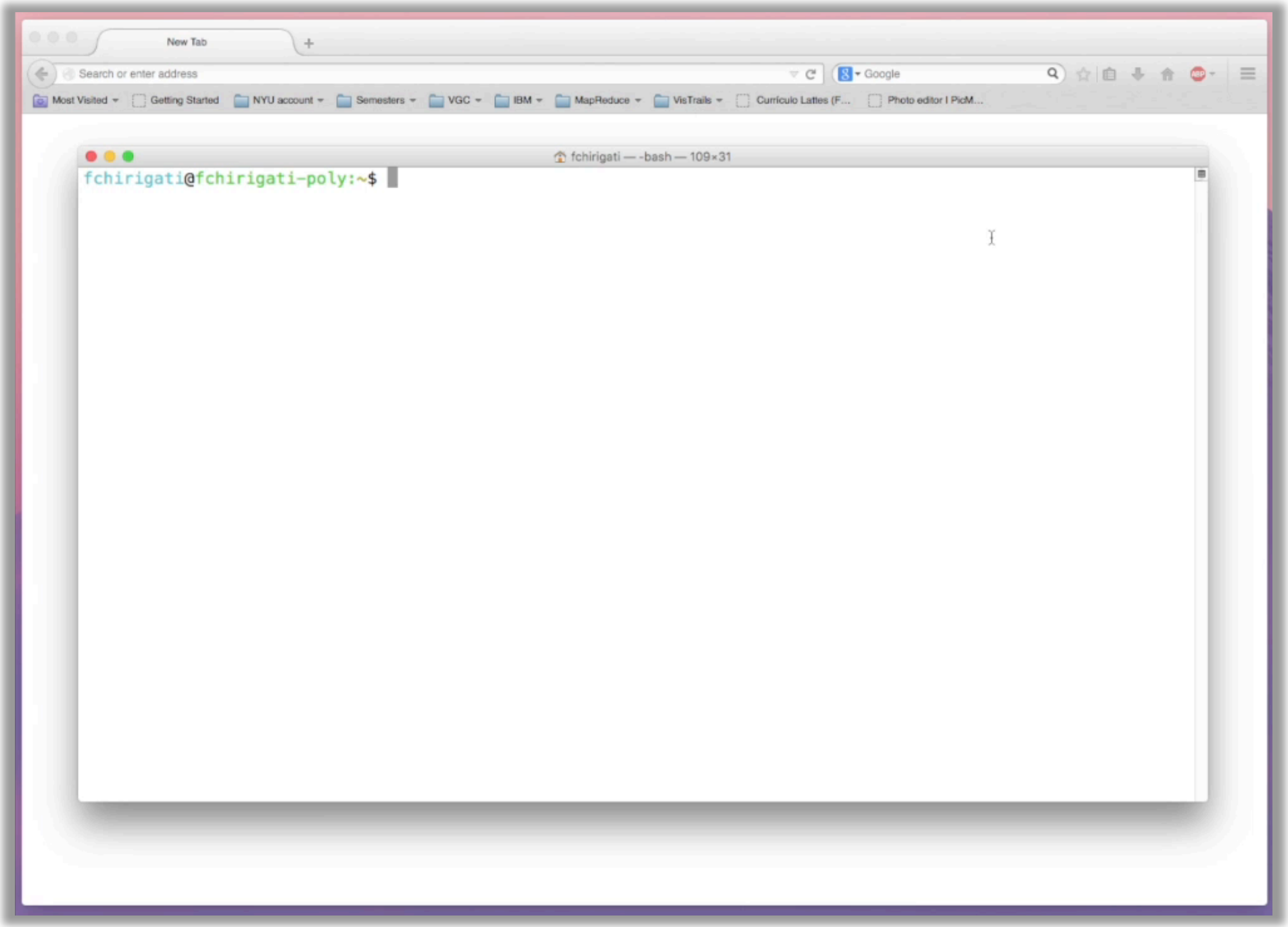
Packing the Web application



Packing


stacked-up.rpz
Mozilla Firefox Start Page
Firefox
Search or enter address
Google
Most Visited
Getting Started
NYU account
Semesters
VGC
IBM
MapReduce
VisTrails
Currículo Lattes (F...
mozilla
Google
Search
Life hack: The easiest way to open a banana is to pinch the end without the stem. Just like monkeys do it.
Downloads
Bookmarks
History
Add-ons
Sync
Settings

Unpacking


New Tab
Search or enter address
Google
Most Visited
Getting Started
NYU account
Semesters
VGC
IBM
MapReduce
VisTrails
Currículo Lattes (F...
Photo editor I PicM...
fchirigati — -bash — 109×31
fchirigati@fchirigati-poly:~$

Unpacking

3

Diffusion MRI - the signal

In this notebook, we examine the raw diffusion-weighted MRI signal. We will load some data, and look at characteristics of the signal. First, let's import some of the elements we will need in the analysis:

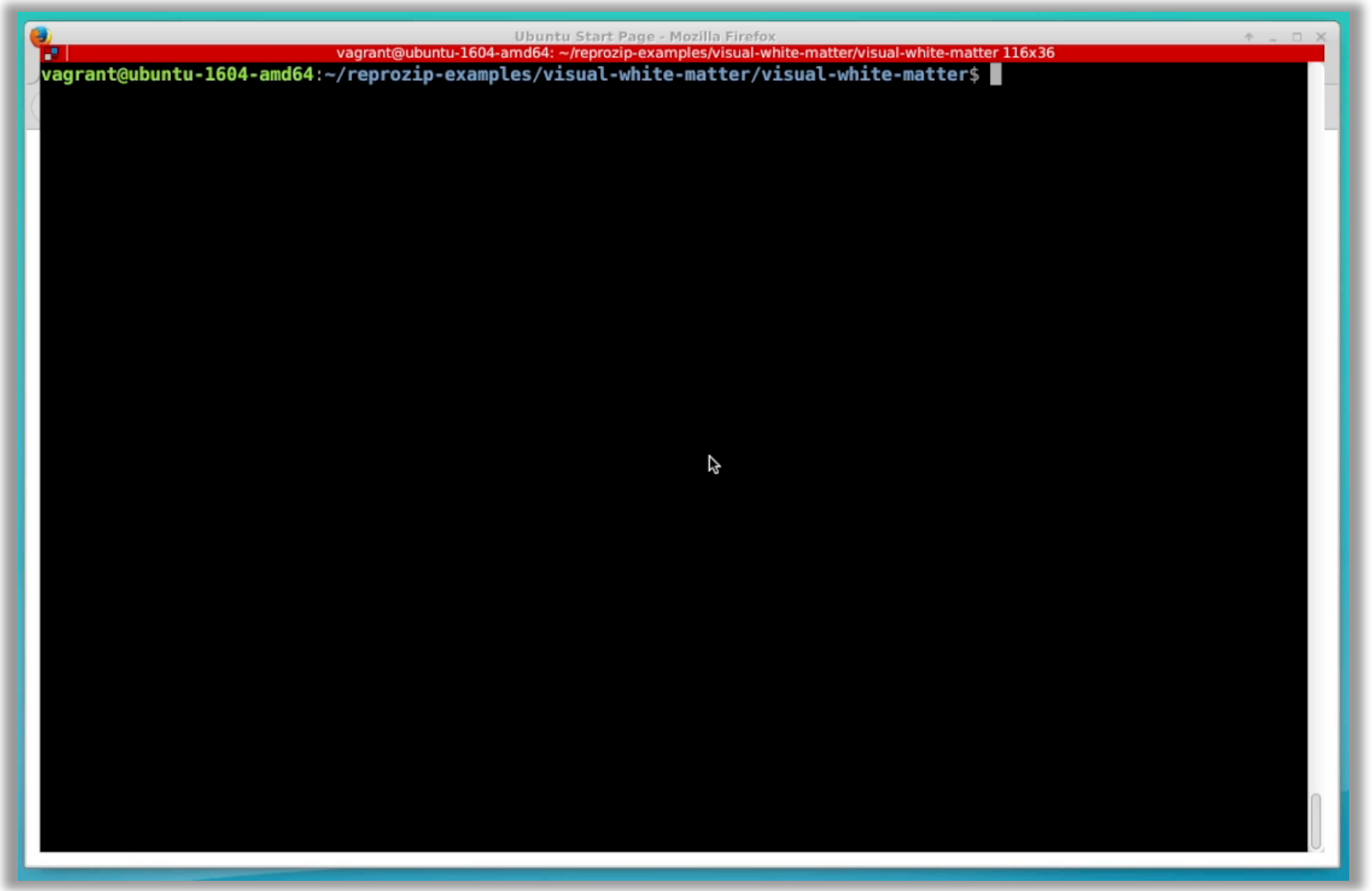
In [1]:
```
import os.path as op
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
import urllib as url
import nibabel as nib
import os.path as op
from mpl_toolkits.mplot3d import Axes3D
import ipywidgets as wdg
import IPython.display as display
from matplotlib.patches import FancyArrowPatch
from mpl_toolkits.mplot3d import proj3d
import dipy.core.geometry as geo
```

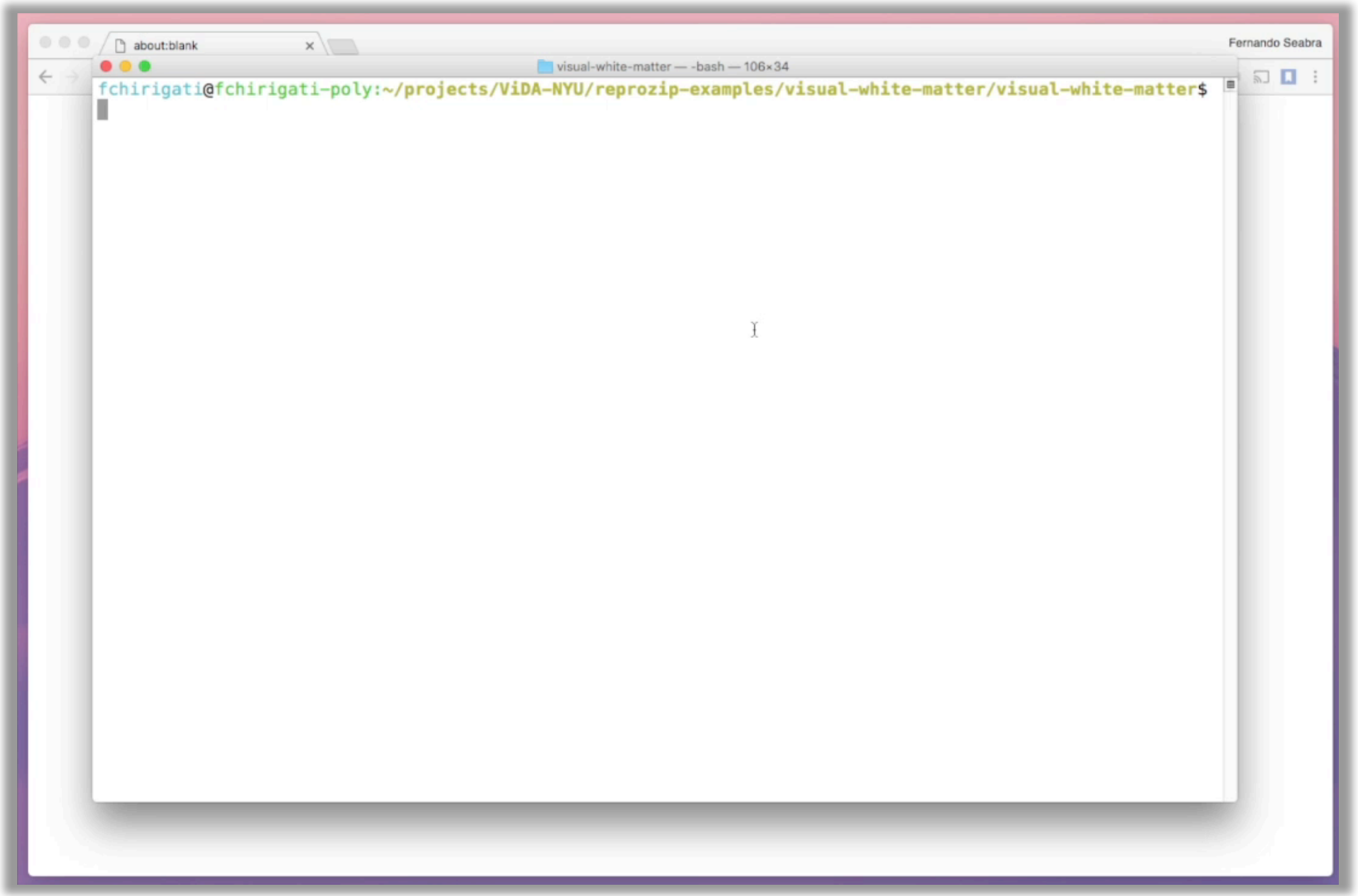
Note

GitHub: *https://github.com/arokem/visual-white-matter*

Packing the notebook

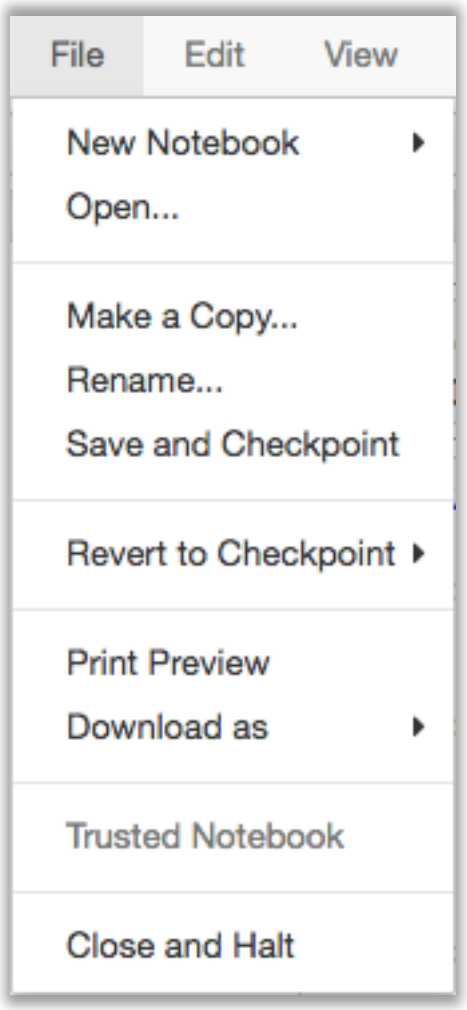


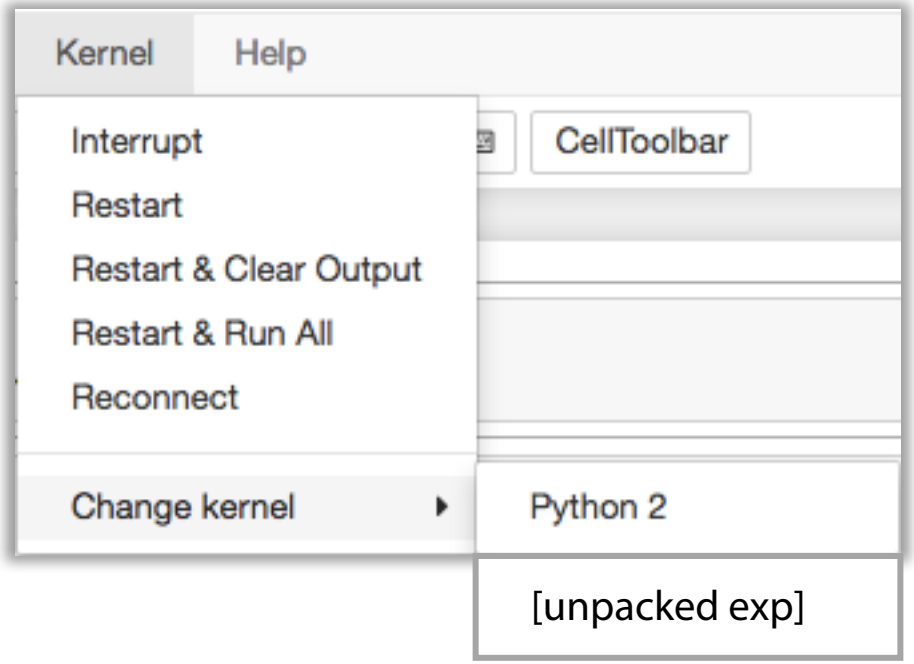
Packing


about:blank
Fernando Seabra
visual-white-matter — -bash — 106×34
fchirigati@fchirigati-poly:~/projects/ViDA-NYU/reprozip-examples/visual-white-matter/visual-white-matter$

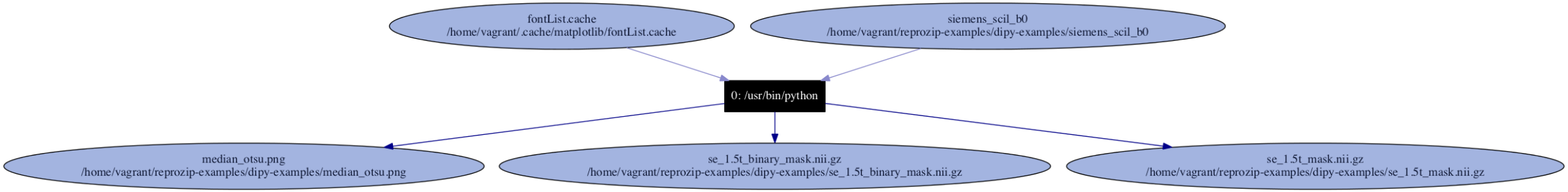
Unpacking

Pack with ReproZip

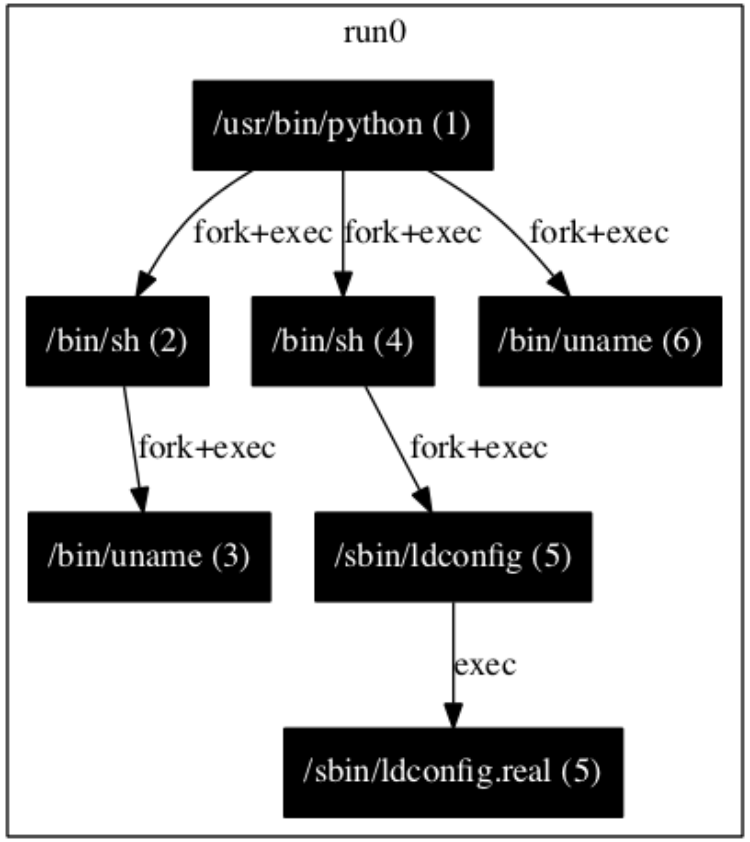




[unpacked exp]




```
    size: 6414336
    packfiles: true
    files:
      # Total files used: 30.70 KB
      # Installed package size: 6.12 MB
      - "/bin/uname" # 30.70 KB
  - name: "dash"
    version: "0.5.8-2.1ubuntu2"
    size: 247808
    packfiles: true
    files:
      # Total files used: 150.46 KB
      # Installed package size: 242.00 KB
      - "/bin/dash" # 150.46 KB
      - "/bin/sh" # Link to /bin/dash
  - name: "fontconfig-config"
    version: "2.11.94-0ubuntu1.1"
    size: 376832
    packfiles: true
    files:
      # Total files used: 53.45 KB
      # Installed package size: 368.00 KB
      - "/etc/fonts/conf.avail/10-antialias.conf" # 223.0 bytes
      - "/etc/fonts/conf.avail/10-hinting-slight.conf" # 229.0 bytes
      - "/etc/fonts/conf.avail/10-hinting.conf" # 212.0 bytes
      - "/etc/fonts/conf.avail/10-scale-bitmap-fonts.conf" # 1.96 KB
      - "/etc/fonts/conf.avail/11-lcdfilter-default.conf" # 526.0 bytes
      - "/etc/fonts/conf.avail/20-unhint-small-vera.conf" # 1.23 KB
      - "/etc/fonts/conf.avail/30-metric-aliases.conf" # 12.50 KB
      - "/etc/fonts/conf.avail/30-urw-aliases.conf" # 701.0 bytes
      - "/etc/fonts/conf.avail/40-nonlatin.conf" # 5.35 KB
      - "/etc/fonts/conf.avail/45-latin.conf" # 4.51 KB
      - "/etc/fonts/conf.avail/49-sansserif.conf" # 545.0 bytes
      - "/etc/fonts/conf.avail/50-user.conf" # 673.0 bytes
      - "/etc/fonts/conf.avail/51-local.conf" # 189.0 bytes
```



Easy Reproduction

Debugging

Dependency Analysis

Easy Virtualization

*"We **do not need** tools like ReproZip!"*

*"We **need** tools like ReproZip!"*

ReproZip can pack…

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

Graphical tools

Interactive tools

Client-server applications (including databases)

Jupyter notebooks (very soon!)

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

... and many more!

Who is using ReproZip?

Academic Publications

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

Other Use Cases

Integrated as a component of CoRR
Archiving data journalism apps, e.g.: Stacked Up
… and many more: https://examples.reprozip.org/

Thank You!

Website: *https://www.reprozip.org/*

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

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

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

Acknowledgements: Ariel Rokem (for sharing awesome examples!)
Moore-Sloan Data Science Environment at NYU
National Science Foundation

Fernando Chirigati
fchirigati@nyu.edu

Rémi Rampin
remi.rampin@nyu.edu

Vicky Steeves
victoria.steeves@nyu.edu

Current and Future Work

Distributed experiments (MPI)

Packing support for OS X

Remote file management

Integration with Jupyter Notebook

Limitations

Only packs experiments in Linux distros

Only detects information about software packages in Debian and Fedora-based environments

… but all the required files are captured regardless of the Linux system!

Does not allow reproducibility of non-deterministic processes

Does not save state

ReproZip vs. Existing Packing Systems

Packing systems: CDE, PTU, CARE

ReproZip adds important features and contributions:

- **Portability:** Linux experiments can be unpacked in different OS'es
- **Extensibility:** Developers can easily implement new unpackers for other environments / systems
- **Reusability:** ReproZip automatically identifies input files, parameters, and output files, allowing users to easily modify these for reuse purposes
- **Easy of use:** Users have control over the collected trace and can customize the reproducible package; ReproZip also provides command-line and graphical interfaces that make it easier to setup, reproduce, and modify the original experiment