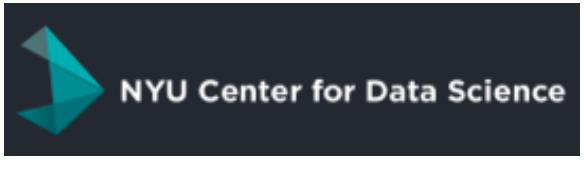
Reproducing and Preserving Research with ReproZip

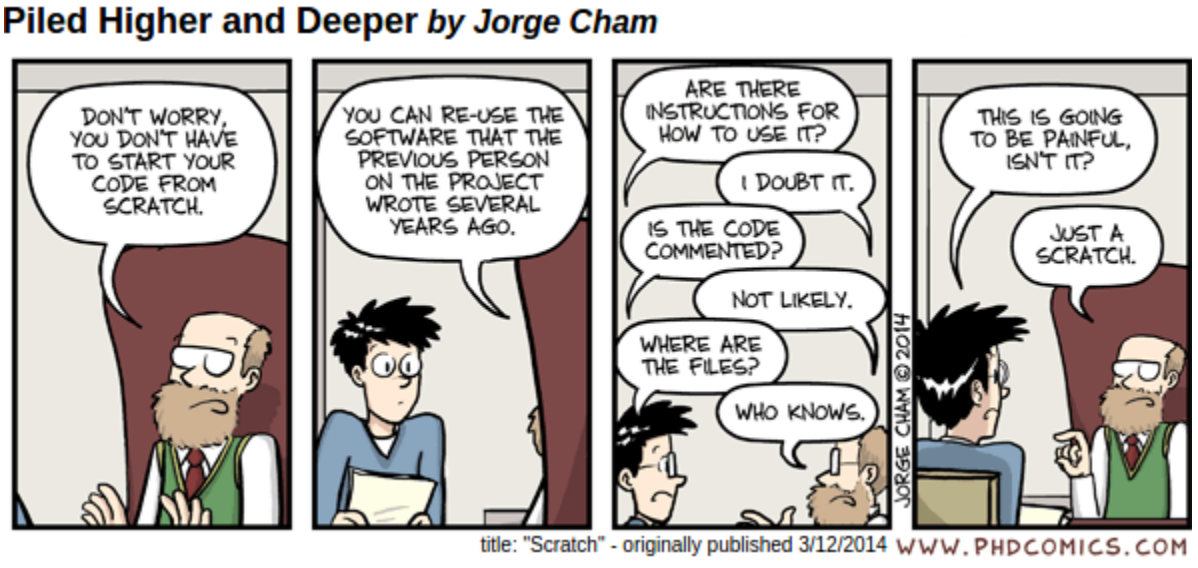
Remi Rampin, Vicky Steeves, Fernando Chirigati

@remram44, @VickySteeves, @fchirigati

IASSIST 2017 | May 25, 2017



Obligatory PhD Comics strip...



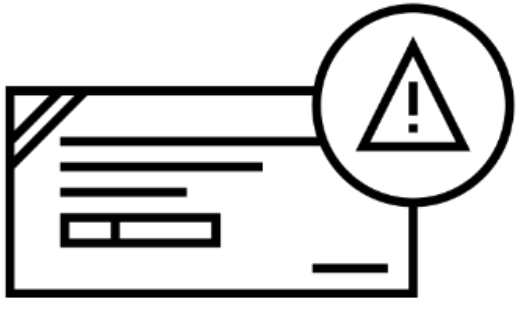
Reproducibility exists on a spectrum

- Reviewable Research: Sufficient detail for peer review & assessment.
- Replicable Research: Tools are available to duplicate the author's results using their data.
- Confirmable Research: Main conclusions can be attained independently without author's software.
- Auditable Research: Process & tools archived such that it can be defended later if necessary.
- **Open/Reproducible Research: Auditable research made openly available.**

Stodden et al ICERM report (2013)

Challenge 1: Everyone Messes Up

Human Errors



People make mistakes--and it impacts their research.

It's good to have other people check out your research and analyses--it's like having a copy editor for your data!

Gap: a tool to seamlessly review whole research projects without the reviewer having to manually install and debug all dependencies, code, and data.

Excel is Terrible

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

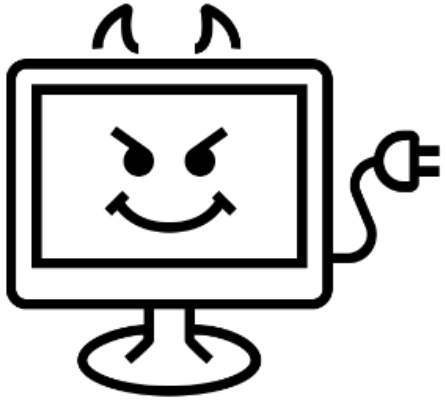
	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

In
histor
searc
public
main
growt
mal"
tries
of GD
wise;
perce
betwe
simila
econo
find
debt
mies
excep
trast,
debt l
Ou
Public
recen
the epicenter countries. This should not be sur-
prising, given the experience of earlier severe

g pop-
s? Are
a man-
pirical,
torical
central
hen M.
009b).
ifficult
f pub-
es, and
arkets.
untries
er, the
rations
, insti-
range-
xternal
ments
arkets,
more
severe threshold for total gross external debt
(public and private)—which is almost exclu-

Challenge 2: 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.

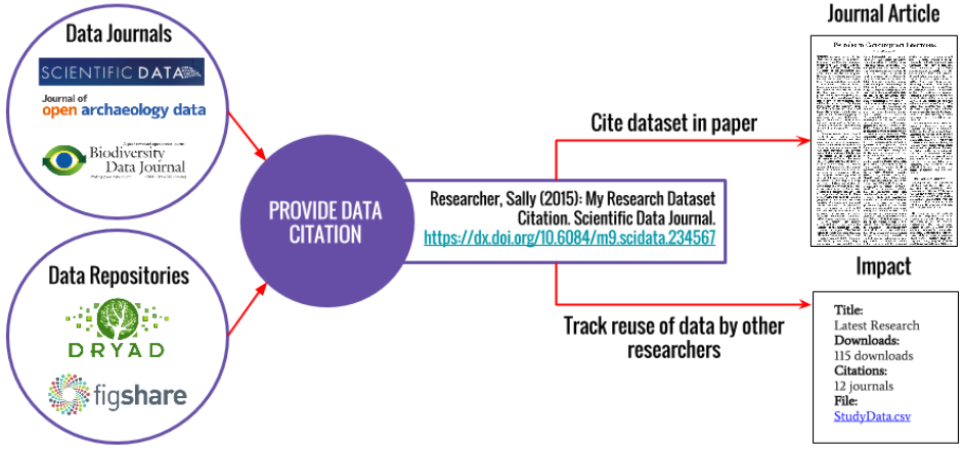
Even if runnable, results may differ

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

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**.

The New Traditional Model

- Publish a paper.
- Publish the underlying code and data.
- Link the paper + code/data.
- Bump up your H-Index.



And I have slides on that model here and here.

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."

"77% claim that they do not have time to document and clean up the code."

Victoria Stodden, Survey of the

Carol Tenopir, Beyond the PDF2 Conference

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

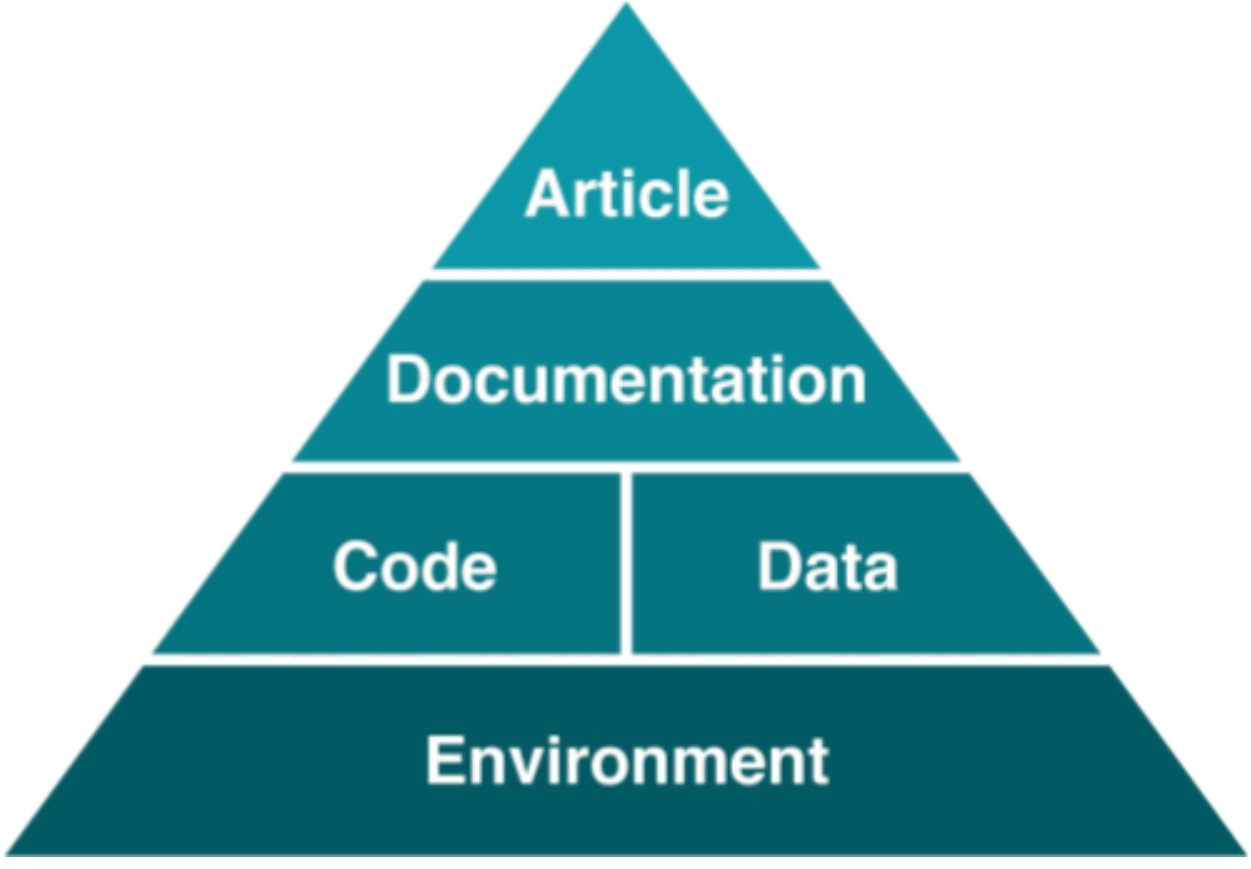
Reproducibility requires skills that are not included in most curriculums!

"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

But the article, code, and data are really just the tip of the iceberg



What does this mean for...

Librarians

We now have to help researchers license code + data + computer environments, select or build repositories to reliably store those objects, and preserve them forever.

Researchers

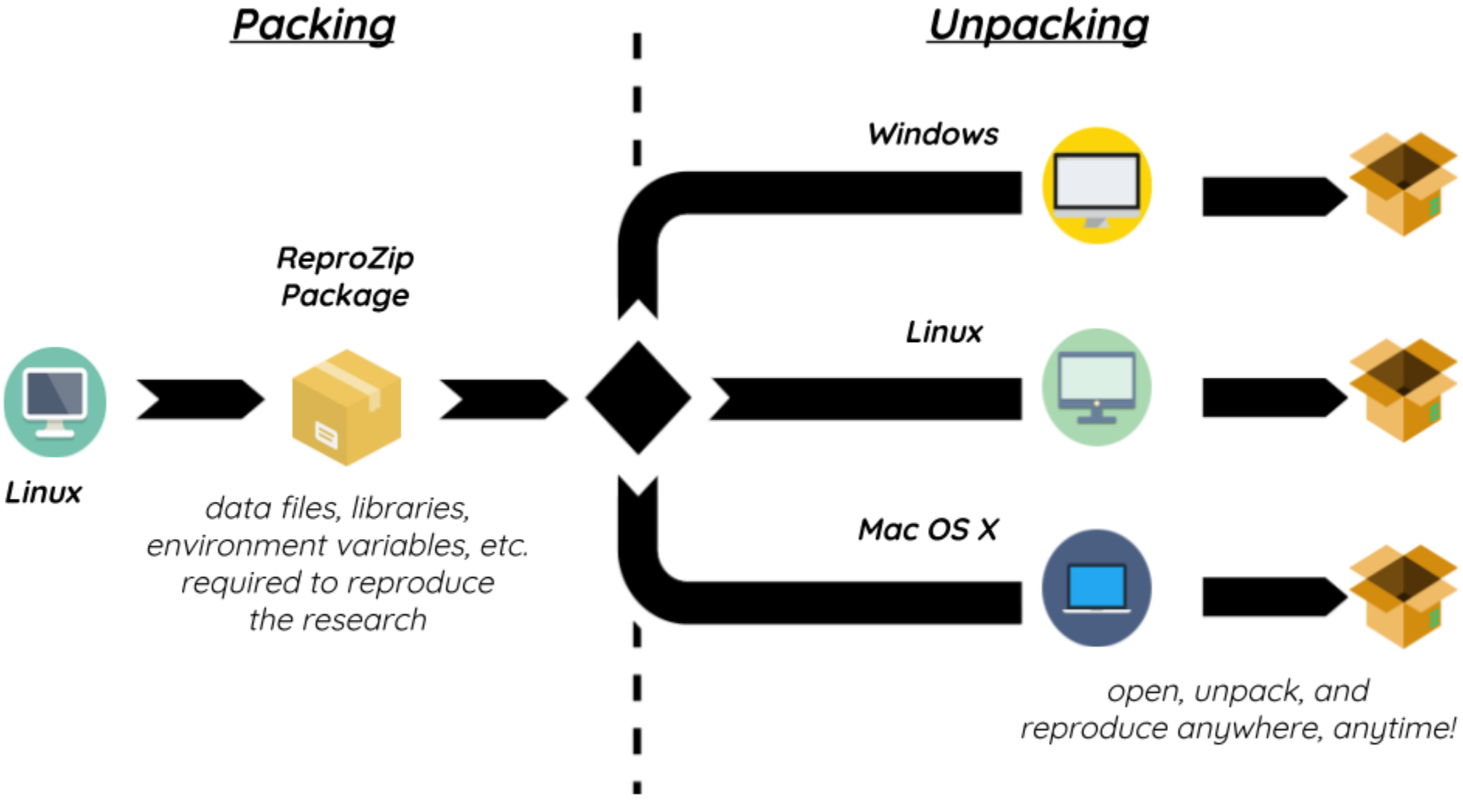
They have to clean up code and code, learn how to capture computer environments and make it shareable, without spending all their time + research budget on it.

But what if I told you that you could put code + data + applications + environment in **one small file that has a ton of automatically captured metadata and provenance**?

Tool to Help: ReproZip!

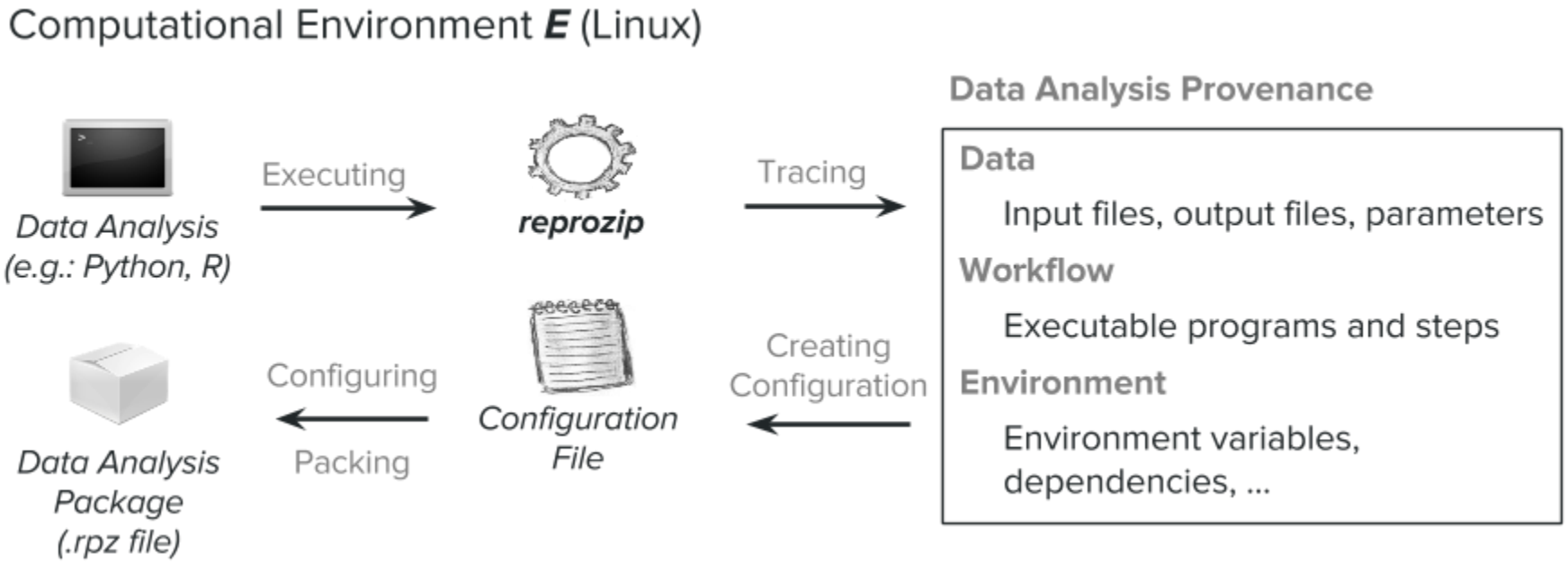
ReproZip is a tool aimed at simplifying the process of creating reproducible...whatever. It can be research, it can be applications, it can be databases, it can be websites...if you can do it on a computer, chances are we can pack it!

2 Steps to Reproducibility



Step 1: Trace & Pack

```
reprozip trace [command]
```

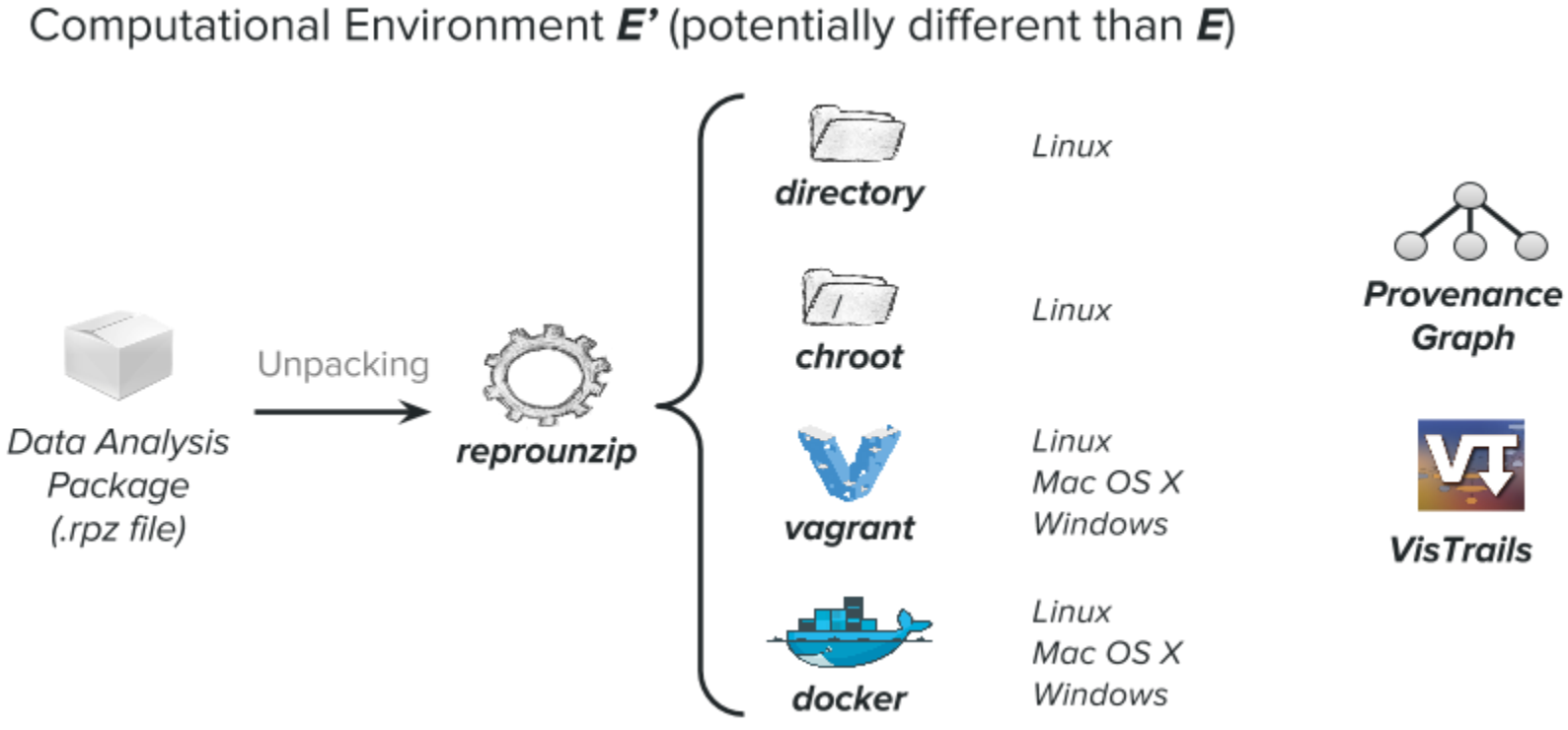



```
reprozip pack package-name.rpz
```

Before you pack, you can edit the config.yml (but it's not recommended). This is what a config looks like.

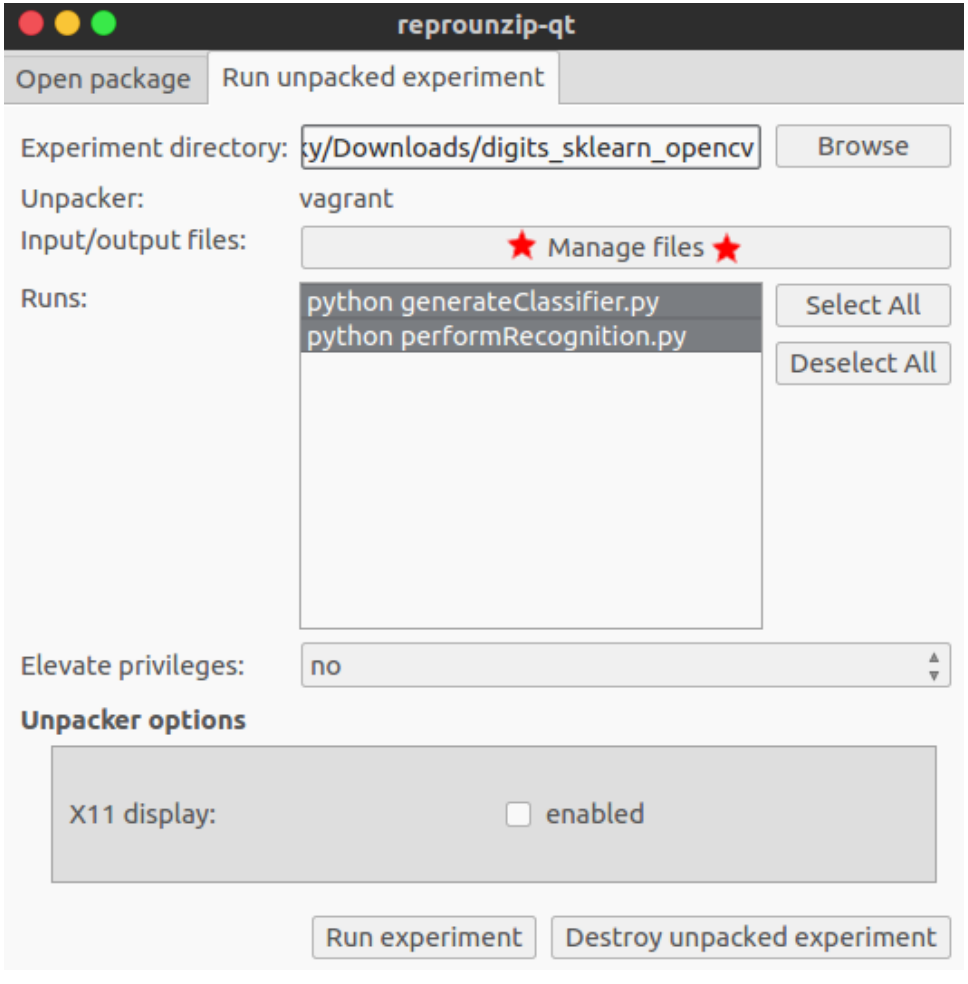
Step 2: Set up & Run

Double click on the RPZ file, and choose your unpacker!



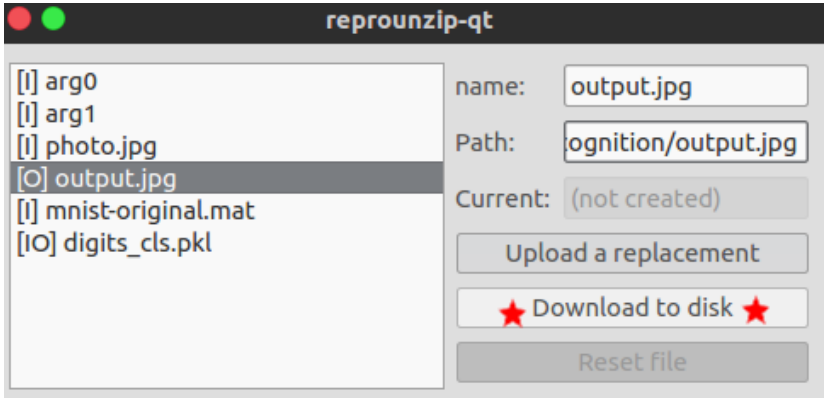
Not just simple reproduction...

When you unpack your .rpz package with the GUI, you'll come to this screen.

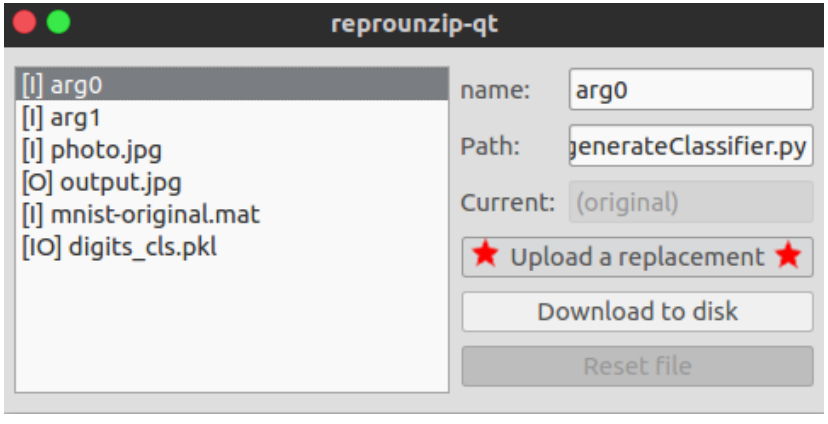


Download the results or add your own input!

Download Ouput



Upload New Inputs



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

Rec. by the Information Systems Journal, Reproducibility Section

Rec. by the ACM SIGMOD Reproducibility Review

Listed on the ACM Artifact Evaluation Process Guidelines

Integrated as a component of CoRR

Archiving data journalism apps like StackedUp

... and many more!

Potential for ReproZip in (Academic)

ReproZip in (Academic) Libraries

Liaison Librarians

- Liaison libs have an excellent opportunity as the first line of contact with patrons to encourage and disseminate information on reproducible practices (e.g. using ReproZip).
- The library gains an excellent way to build collections of diverse, preservation-ready research outputs, and patrons easily adopt reproducible practices.

Data Services

Adding ReproZip to the curriculum of data services classes and workshops, and to the list of supported software, data services teams can become a center on campus for assisting the reproducibility of their patrons' scholarship.

Potential for ReproZip in (Digital) Libraries

Digital Libraries

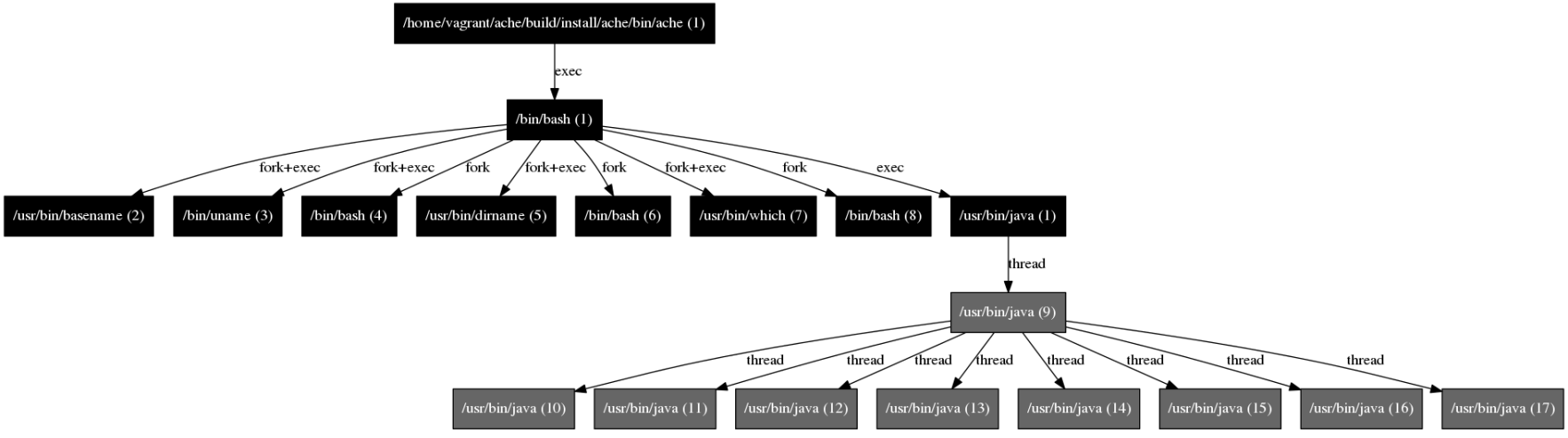
- ReproZip isn't relient on Docker or Vagrant; it uses a plugin model for unpackers, so new ones can be added for forward compatibility.
- If no containers or VMs exist in the future, then the archivist can still read and use the robust technical and administrative metadata in ReproZip's config.yml.

Repository Management

ReproZip contains extensive technical and administrative metadata, which can be an be exported a json file, which allows for extensible models of metadata – such as crosswalking to Resource Description Framework (RDF) or Dublin Core, automating ingest workflows.

Future Development Work

- Packing on macOS (beginning summer 2017)
- Improvements to provenance graph visualizaion (beginning summer 2017), because right now we've got...



- ReproZip plugin for Jupyter Notebooks (beginning summer 2017)
- Better MPI/HPC support

Conclusion

ReproZip is extensible enough to be used for reproducibility across research domains as well as across library services.

Because ReproZip is open-source software, the community drives its development – others can contribute, modify, and reuse it for a variety of purposes.

The library community can leverage ReproZip in instruction, consultation, repository services, digital archiving, and in their own research.

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:

Prof. Juliana Freire, ReproZip PI and reproducibility master.

Dr. Nicholas Wolf for his help in editing our paper.

The Gordon and Betty Moore Foundation & the Alfred P. Sloan Foundation, who support The Moore-Sloan Data Science Environment at NYU, which was vital to the development of ReproZip

Questions?

Get this Presentation:
https://vickysteeves.gitlab.io/2017-IASSIST-ReproZip

Email us:
reprozip-users@vgc.poly.edu
or
vicky.steeves@nyu.edu
or
remi.rampin@nyu.edu
or
fchirigati@nyu.edu