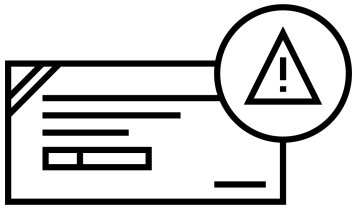
5/03/2016

# ReproZip: Reproducibility with Ease

Rémi Rampin | Fernando Chirigati | Vicky Steeves
Juliana Freire | Dennis Shasha
NYU Tandon School of Engineering

NYU

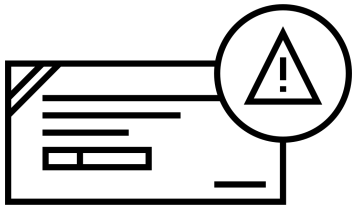
# Challenges of Reproducibility



## Honest, Human Errors

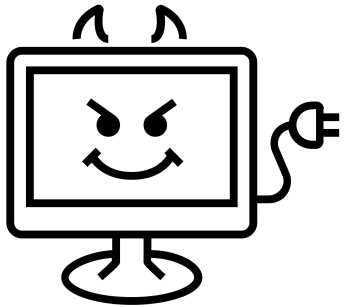
- Drawing wrong conclusions
- Poorly documented processes and setups
- Statistical significance, ...

# Challenges of Reproducibility (computational science)



## Honest, Human Errors

- Drawing wrong conclusions
- Poorly documented processes and setups
- Statistical significance, ...



## Dependency Hell

- It's *hard* to keep track of what was used, what version was used, on what hardware/software config it was used, etc.
- You can't just include all the scripts and the data and expect people to be able to run it!
  - Libraries get updated, even operating system changes can disrupt reproducibility

# 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.

# ReproZip, the Reproducibility Packer!


# Packing Experiments

AUTHORS

Computational Environment ***E*** (Linux)

*Data Analysis*
*(e.g.: Python, R)*

# Packing Experiments

AUTHORS

Computational Environment ***E*** (Linux)

*Data Analysis*
*(e.g.: Python, R)*

Executing

***reprozip***

# Packing Experiments

AUTHORS

Computational Environment ***E*** (Linux)

*Data Analysis*
*(e.g.: Python, R)*

Executing →

***reprozip***

Tracing →

**Data Analysis Provenance**

**Data**

Input files, output files, parameters

**Workflow**

Executable programs and steps

**Environment**

Environment variables, dependencies, ...

# Packing Experiments


Computational Environment ***E*** (Linux)

*Data Analysis (e.g.: Python, R)*

Executing

***reprozip***

Tracing

**Data Analysis Provenance**

**Data**

Input files, output files, parameters

**Workflow**

Executable programs and steps

**Environment**

Environment variables, dependencies, ...

*Configuration File*

Creating Configuration

# Packing Experiments


Computational Environment ***E*** (Linux)

*Data Analysis (e.g.: Python, R)*

Executing

***reprozip***

Tracing

**Data Analysis Provenance**

**Data**

Input files, output files, parameters

**Workflow**

Executable programs and steps

**Environment**

Environment variables, dependencies, ...

Creating Configuration

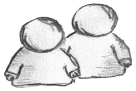
*Configuration File*

Configuring

Packing

*Data Analysis Package (.rpz file)*

# Unpacking Experiments



REVIEWERS
READERS

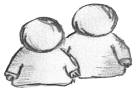
Computational Environment ***E'*** (potentially different than ***E***)

*Experiment Package (.rpz file)*

Unpacking

***reprounzip***

# Unpacking Experiments



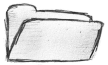
REVIEWERS
READERS

Computational Environment ***E'*** (potentially different than ***E***)

*Experiment Package (.rpz file)*

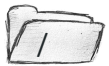
Unpacking

***reprounzip***



***directory***

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



***chroot***

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

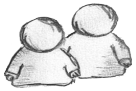
***vagrant***

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

***docker***

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

# Unpacking Experiments



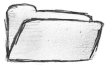
REVIEWERS
READERS

Computational Environment ***E'*** (potentially different than ***E***)

*Experiment Package (.rpz file)*

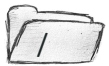
Unpacking

***reprounzip***



***directory***

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



***chroot***

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

***vagrant***

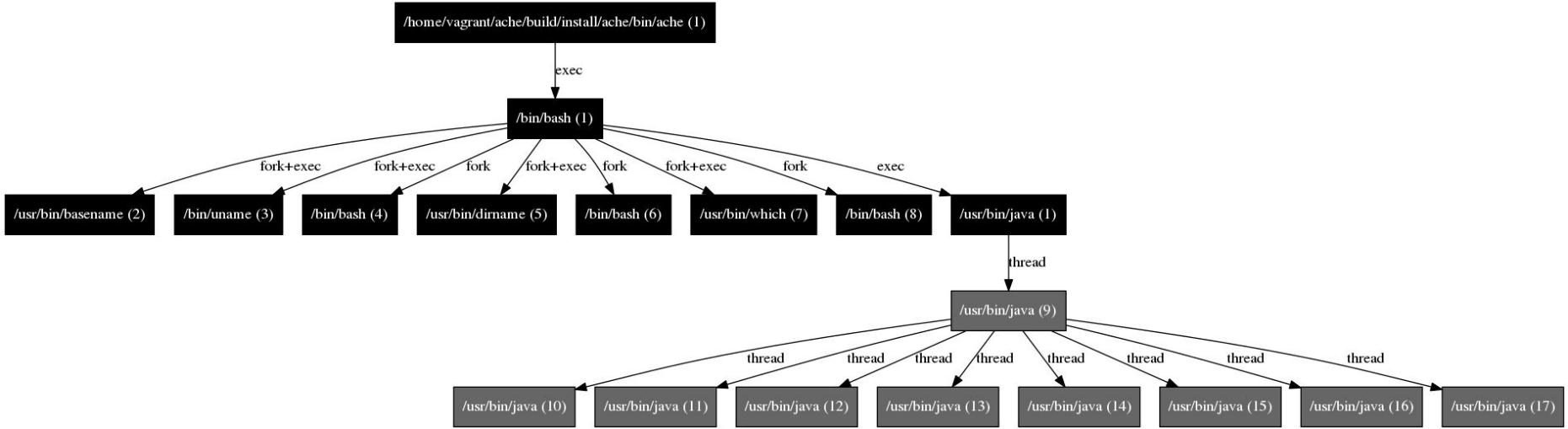
*unpacks in a virtual machine using Vagrant (Linux, Mac OS X, Windows)*

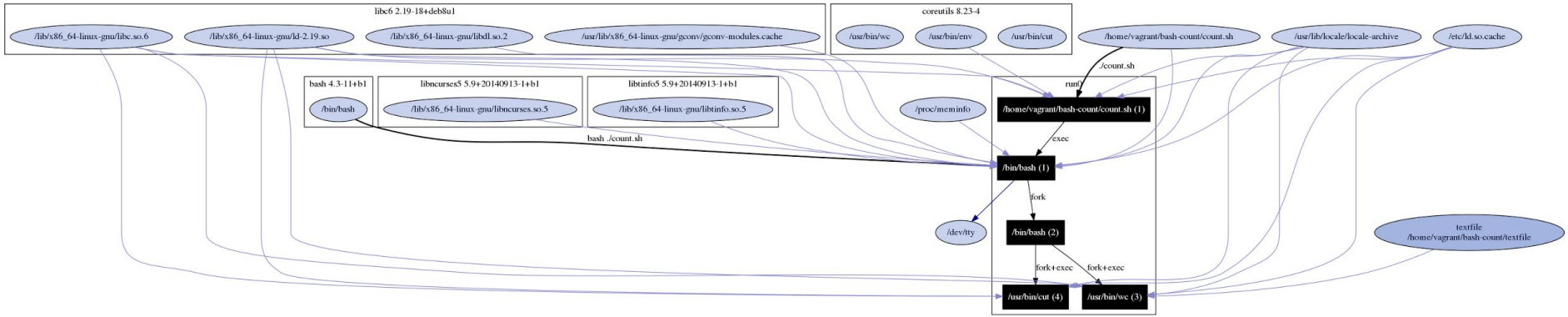
***docker***

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

***Provenance Graph***

***VisTrails***

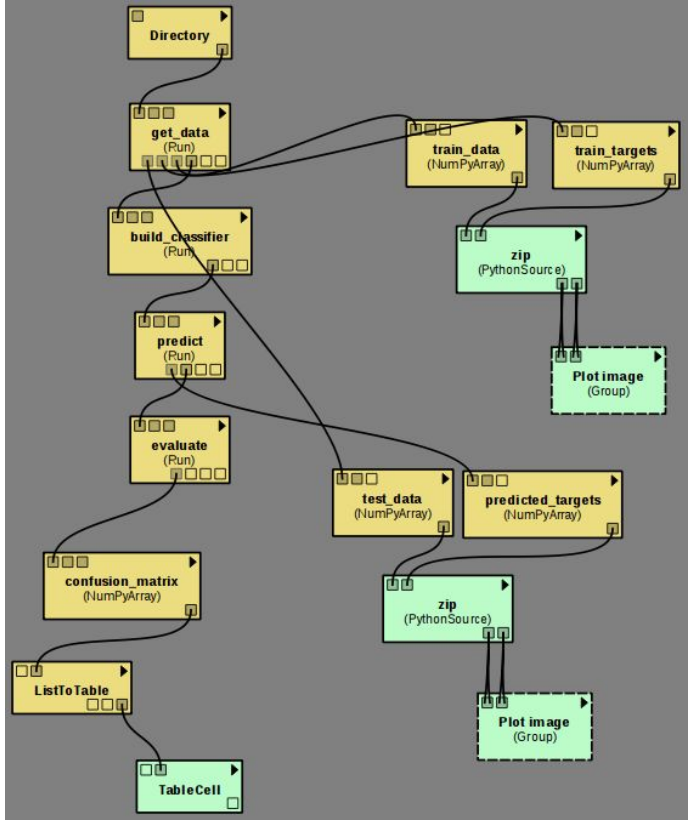

/home/vagrant/ache/build/install/ache/bin/ache (1)
exec
/bin/bash (1)
fork+exec
fork+exec
fork
fork+exec
fork
fork+exec
fork
exec
/usr/bin/basename (2)
/bin/uname (3)
/bin/bash (4)
/usr/bin/dirname (5)
/bin/bash (6)
/usr/bin/which (7)
/bin/bash (8)
/usr/bin/java (1)
thread
/usr/bin/java (9)
thread
thread
thread
thread
thread
thread
thread
thread
/usr/bin/java (10)
/usr/bin/java (11)
/usr/bin/java (12)
/usr/bin/java (13)
/usr/bin/java (14)
/usr/bin/java (15)
/usr/bin/java (16)
/usr/bin/java (17)


libc6 2.19-18+deb8u1
/lib/x86_64-linux-gnu/libc.so.6
/lib/x86_64-linux-gnu/ld-2.19.so
/lib/x86_64-linux-gnu/libdl.so.2
/usr/lib/x86_64-linux-gnu/gconv/gconv-modules.cache
coreutils 8.23-4
/usr/bin/wc
/usr/bin/env
/usr/bin/cut
/home/vagrant/bash-count/count.sh
/usr/lib/locale/locale-archive
/etc/ld.so.cache
bash 4.3-11+b1
/bin/bash
libncurses5 5.9+20140913-1+b1
/lib/x86_64-linux-gnu/libncurses.so.5
libtinfo5 5.9+20140913-1+b1
/lib/x86_64-linux-gnu/libtinfo.so.5
/count.sh
runt
/proc/meminfo
/home/vagrant/bash-count/count.sh (1)
bash ./count.sh
exec
/bin/bash (1)
fork
/dev/tty
/bin/bash (2)
textfile
/home/vagrant/bash-count/textfile
fork+exec
fork+exec
/usr/bin/cut (4)
/usr/bin/wc (3)

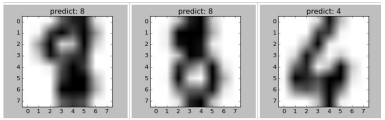
# ReproZip: Workflow & VisTrails

VisTrails Integration provides a graphical view of the unpacked ReproZip package.

- VisTrails drives the execution of this unpacked experiment or environment
  - exposes the input and output files and adds them on ports
  - easier for user to add or subtract input files
    - the user can create any other block--which could be a script they want to run with the packed experiment/environment




| row | col 0 | col 1 | col 2 | col 3 | col 4 |
|---|---|---|---|---|---|
| 0 | 87 | 0 | 0 | 0 | 1 |
| 1 | 0 | 88 | 1 | 0 | 0 |
| 2 | 0 | 0 | 85 | 1 | 0 |
| 3 | 0 | 0 | 0 | 79 | 0 |
| 4 | 0 | 0 | 0 | 0 | 88 |



# Current Use Cases

Academic Publications

ReproZip packages (.rpz) files to be included with each publication and cited as data, no different than other datasets.

1st Case: Information Systems journal (Reproducibility Section)

Authors included a DOI to their .rpz package in a shared Mendeley repository available to be cited as a dataset.

Recommended by the ACM SIGMOD 2015 Reproducibility Review

Listed on the Artifact Evaluation Process Guidelines

Invited to be presented in a workshop on preserving scientific research that will result in a report for the NSF

Bonneau Lab (NYU): Comp. Biology using .rpz to make archival snapshots of research

# Current (and future) work

- Distributed experiments (and HPC)
  - Works, though setting up the experiment for reproduction is involved
- OSX support
- Graphical UI to reproduce without touching the command-line
- Integration with the Jupyter Notebook (tmpnb)

# Thank You!

ReproZip Info

Website: http://vida-nyu.github.io/reprozip/

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

Examples: https://github.com/ViDA-NYU/reprozip-examples

Contact Info

Vicky Steeves: victoria.steeves@nyu.edu
Rémi Rampin: remi.rampin@nyu.edu
Fernando Chirigati: fchirigati@nyu.edu

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