Achieving Reproducibility with ReproZip

Fernando Chirigati and **Rémi Rampin**
ViDA – Visualization and Data Analysis Lab
NYU Tandon School of Engineering

Joint work with: *Juliana Freire (NYU Tandon)*
Dennis Shasha (Courant)

NYU | TANDON SCHOOL OF ENGINEERING

Why Reproducibility?

"If I have seen further, it is by standing on the shoulders of giants."
Isaac Newton

To build on top of previous work – *science is incremental!*

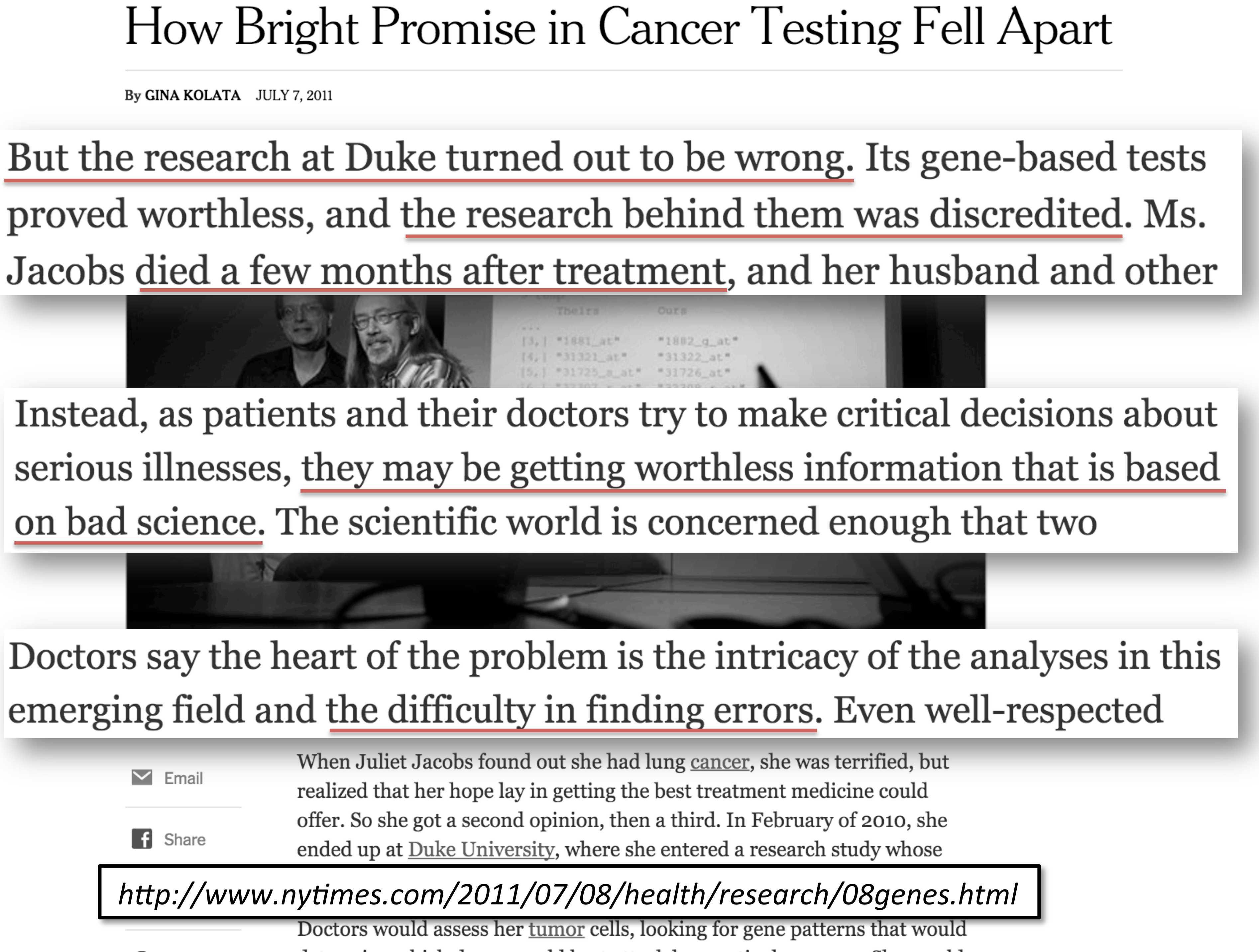
To verify the correctness of results

To defeat self-deception [Nuzzo 2015]

To help newcomers

To increase impact, visibility [Vandewalle et al. 2009] and research quality [Begley and Ellis 2012]

"Without reproducibility, people die!"
John Wilbanks, AMPS Workshop on Reproducibility 2011



How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011

But the research at Duke turned out to be wrong. Its gene-based tests proved worthless, and the research behind them was discredited. Ms. Jacobs died a few months after treatment, and her husband and other

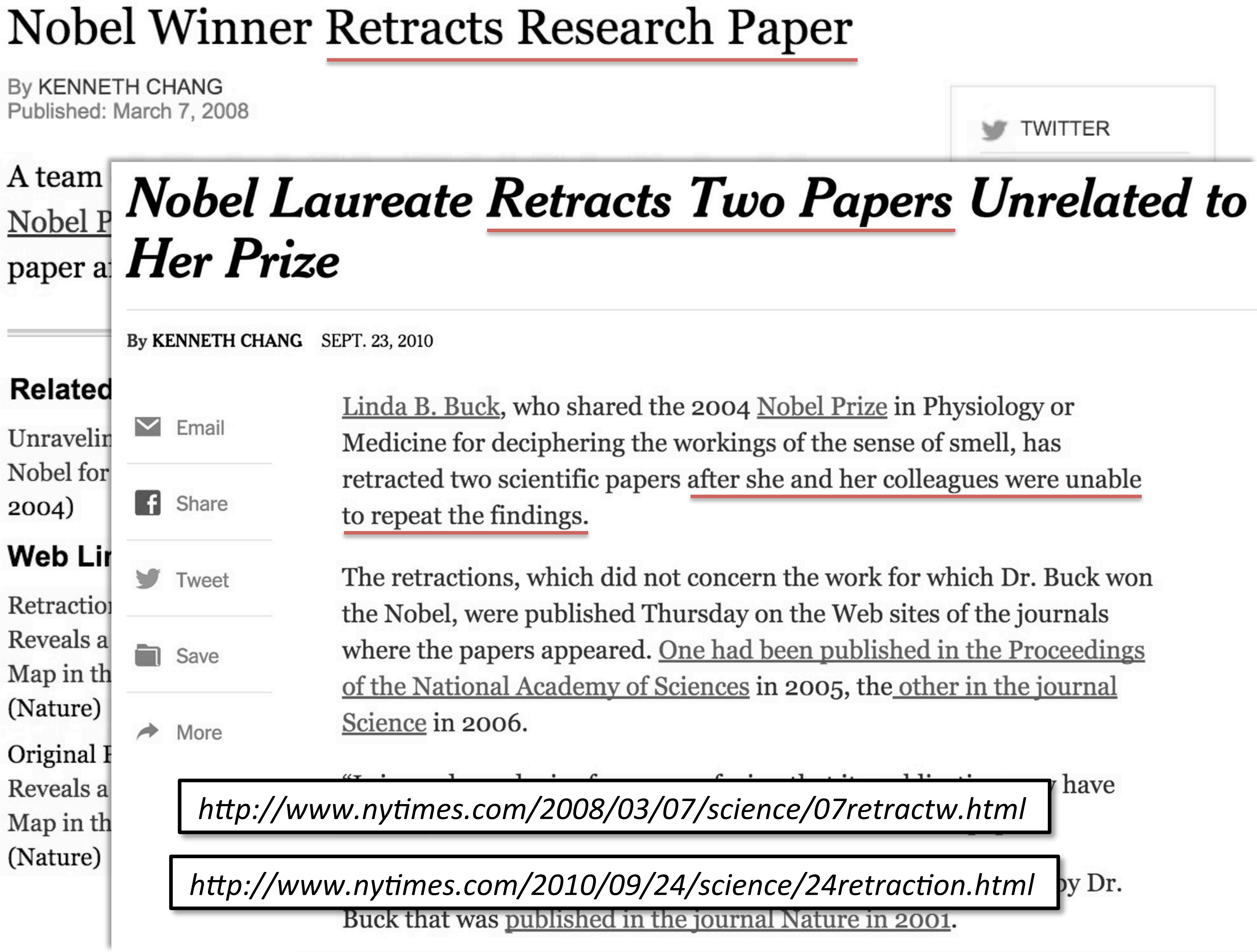
Instead, as patients and their doctors try to make critical decisions about serious illnesses, they may be getting worthless information that is based on bad science. The scientific world is concerned enough that two

Doctors say the heart of the problem is the intricacy of the analyses in this emerging field and the difficulty in finding errors. Even well-respected

When Juliet Jacobs found out she had lung cancer, she was terrified, but realized that her hope lay in getting the best treatment medicine could offer. So she got a second opinion, then a third. In February of 2010, she ended up at Duke University, where she entered a research study whose

http://www.nytimes.com/2011/07/08/health/research/08genes.html

Doctors would assess her tumor cells, looking for gene patterns that would



Nobel Winner Retracts Research Paper

By KENNETH CHANG
Published: March 7, 2008

A team
Nobel P
paper a

Related

Unravelin
Nobel for
2004)

Web Lin

Retractio
Reveals a
Map in th
(Nature)

Original P
Reveals a
Map in th
(Nature)

Nobel Laureate Retracts Two Papers Unrelated to Her Prize

By KENNETH CHANG SEPT. 23, 2010

Linda B. Buck, who shared the 2004 Nobel Prize in Physiology or Medicine for deciphering the workings of the sense of smell, has retracted two scientific papers after she and her colleagues were unable to repeat the findings.

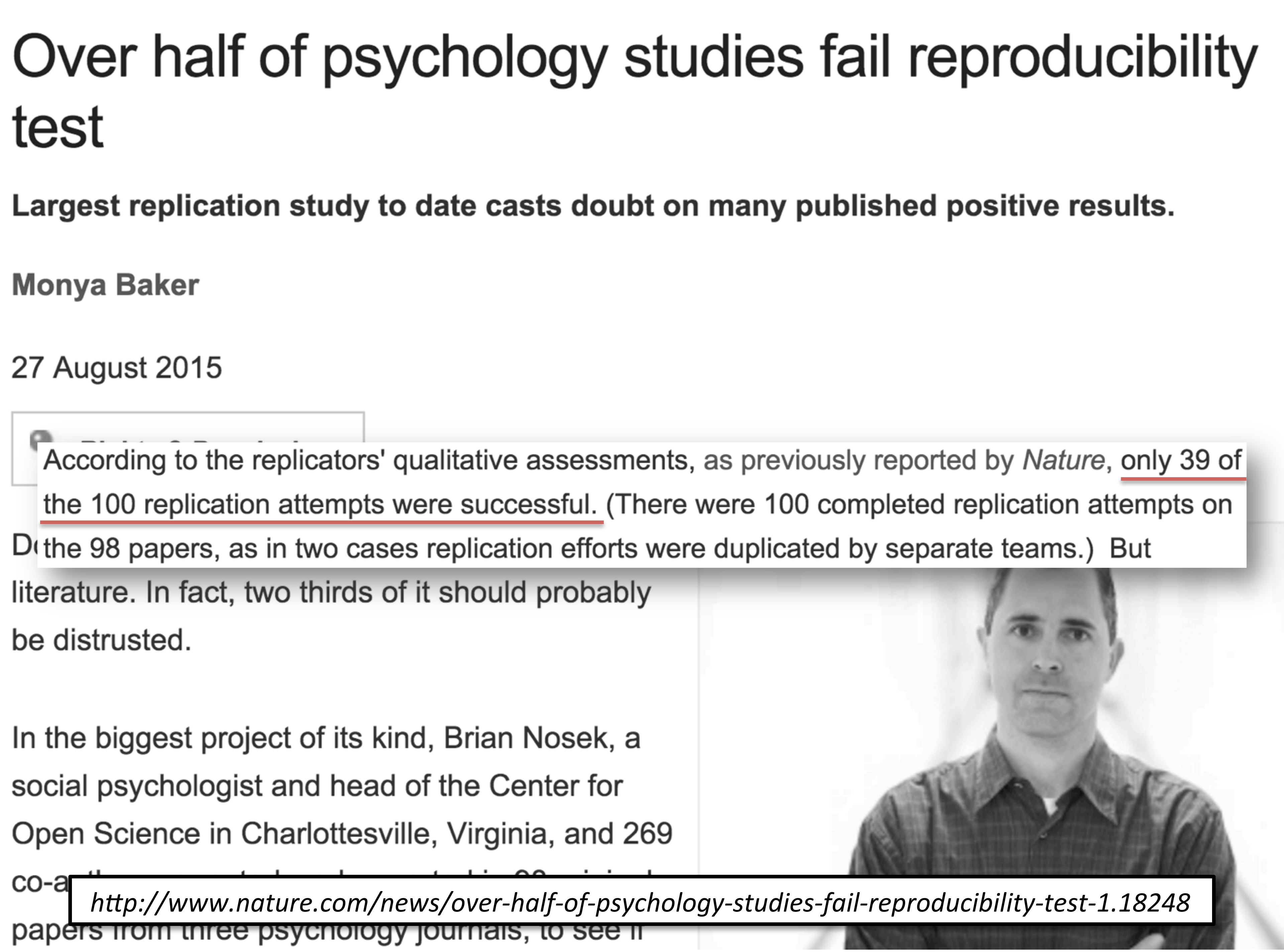
The retractions, which did not concern the work for which Dr. Buck won the Nobel, were published Thursday on the Web sites of the journals where the papers appeared. One had been published in the Proceedings of the National Academy of Sciences in 2005, the other in the journal Science in 2006.

“I … have

… by Dr. Buck that was published in the journal Nature in 2001.

http://www.nytimes.com/2008/03/07/science/07retractw.html

http://www.nytimes.com/2010/09/24/science/24retraction.html



Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

According to the replicators' qualitative assessments, as previously reported by *Nature*, only 39 of the 100 replication attempts were successful. (There were 100 completed replication attempts on the 98 papers, as in two cases replication efforts were duplicated by separate teams.) But

D literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-a

papers from three psychology journals, to see if

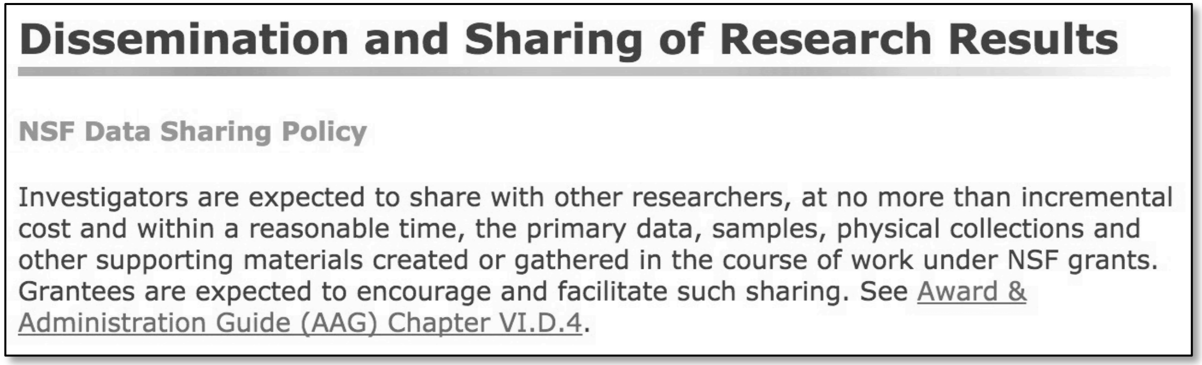
http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248

So… let’s make reproducibility happen!

Encouraged (and sometimes **required**!) by:

- Prestigious conferences
  - E.g.: SIGMOD, VLDB, EuroVis, and IEEE VIS
- Journals
  - E.g.: PNAS, Nature, and Science
- National Academy of Sciences [Cicerone 2005]
- Funding agencies
  - E.g.: NSF and NIH

NSF



Dissemination and Sharing of Research Results

NSF Data Sharing Policy

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI.D.4.

(Computational) Reproducibility is hard.

Why?

Cultural Change

Potential Lack of Attribution

Legal Barriers

Burdensome

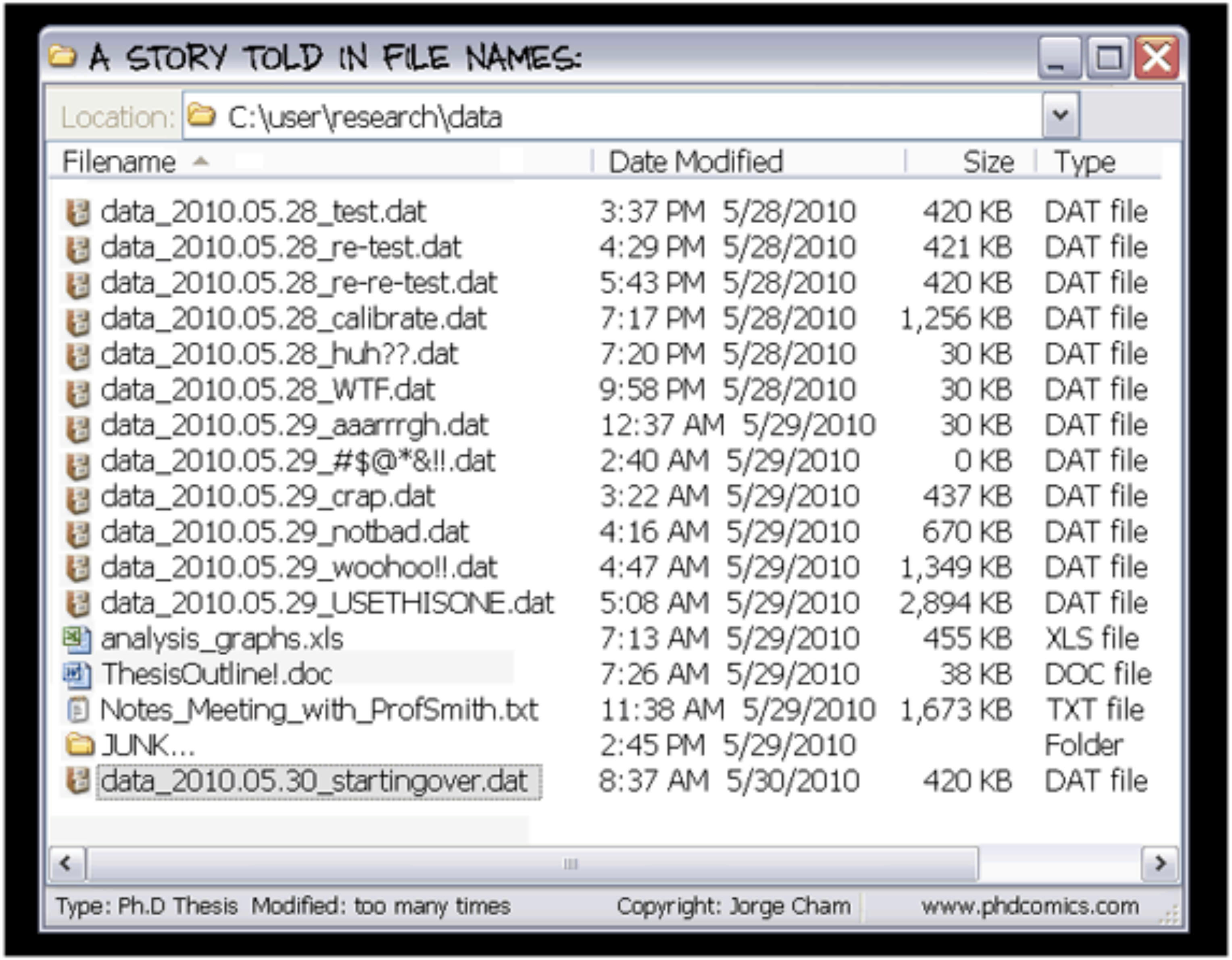
Too many dependencies!

DATA

PROVENANCE

WORKFLOW

ENVIRONMENT



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

Too many different platforms!

From intelerad-insider.com

Too much to do, too little time!

“authors have complained that the process ***requires too much work for the benefit derived****”*

Bonnet et al., SIGMOD Record 2011

“Insufficient time *is the main reason why scientists do not make their data and experiment available and reproducible.”*

Carol Tenopir, Beyond the PDF 2 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

“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

From galleryhip.com

Planning for Reproducibility

Scientific Workflow Systems (VisTrails, Taverna, Kepler, …)

Virtual Machines and Containers (VirtualBox, Vagrant, Docker, …)

Configuration Management Tools (Chef, Puppet, …)

… and many others !

But what about *reproducibility after the fact*?

Again, time-consuming and error-prone!

ReproZip to the Rescue !

Automatically and systematically captures the *provenance* of an existing experiment (Linux only)

Language-independent approach and solution

Creates a self-contained *reproducible package* from captured provenance

Extracts package in another environment, *independent* of the operating system

Provides *easy-to-use* interfaces for replicating and varying the original configuration of the experiment

How does ReproZip work?

ReproZip is a packaging tool

PACKING STEP

From reputablemoving.com

UNPACKING STEP

From wykop.pl

Packing Experiments

AUTHORS

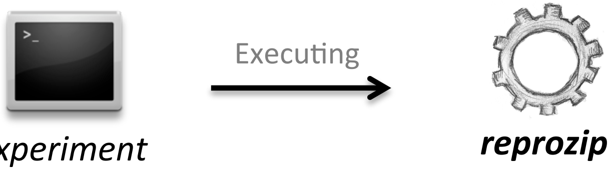
Computational Environment ***E*** (Linux)

Experiment

Packing Experiments

AUTHORS

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



Experiment

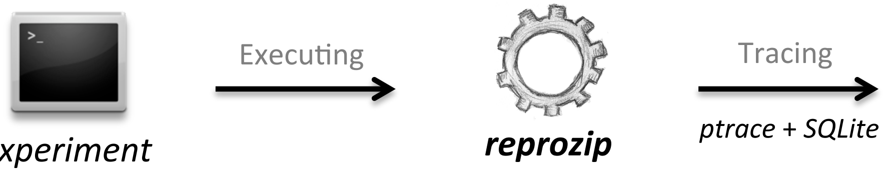
Executing

reprozip

Packing Experiments

AUTHORS

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



Experiment Provenance

Data

Input files, output files, parameters, …

Workflow

Executable programs and steps

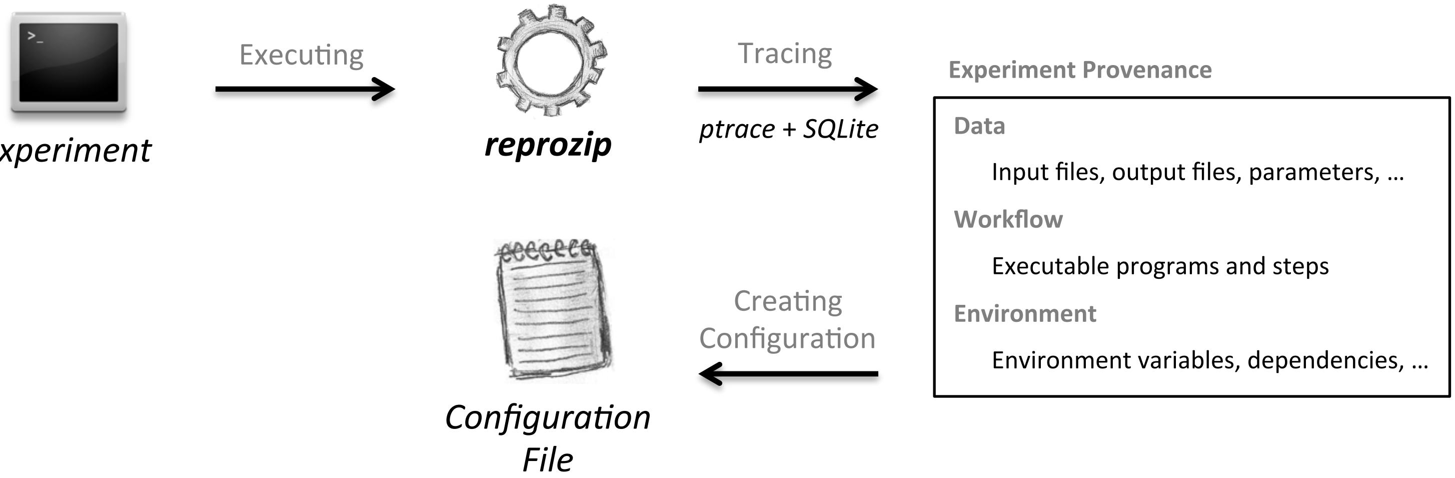
Environment

Environment variables, dependencies, …

Packing Experiments

AUTHORS

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



Experiment

Executing

reprozip

Tracing

ptrace + SQLite

Experiment Provenance

Data

Input files, output files, parameters, …

Workflow

Executable programs and steps

Environment

Environment variables, dependencies, …

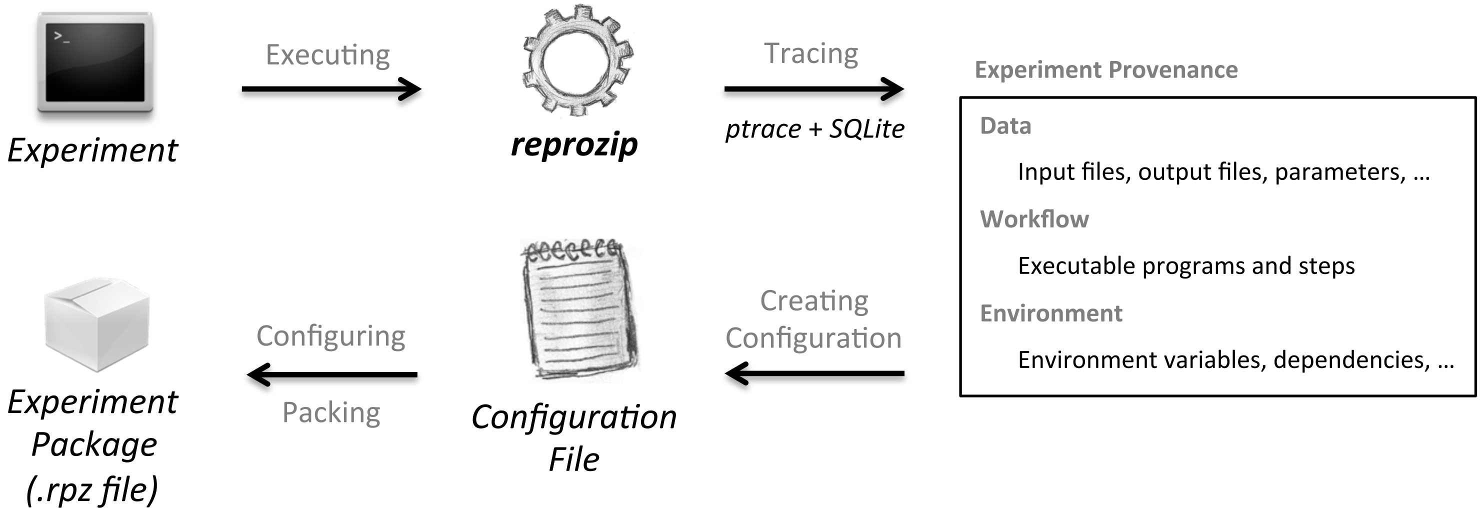
Creating Configuration

Configuration File

Packing Experiments

AUTHORS

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



Experiment

Executing

reprozip

Tracing

ptrace + SQLite

Experiment Provenance

Data

Input files, output files, parameters, …

Workflow

Executable programs and steps

Environment

Environment variables, dependencies, …

Creating Configuration

Configuration File

Configuring

Packing

Experiment Package (.rpz file)

Unpacking Experiments

REVIEWERS
READERS

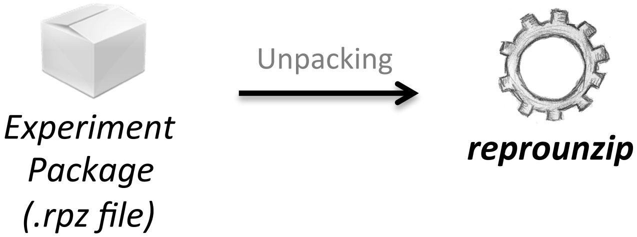
Computational Environment ***E’*** (potentially different than ***E***)

Experiment
Package
(.rpz file)

Unpacking Experiments

Reviewers
Readers

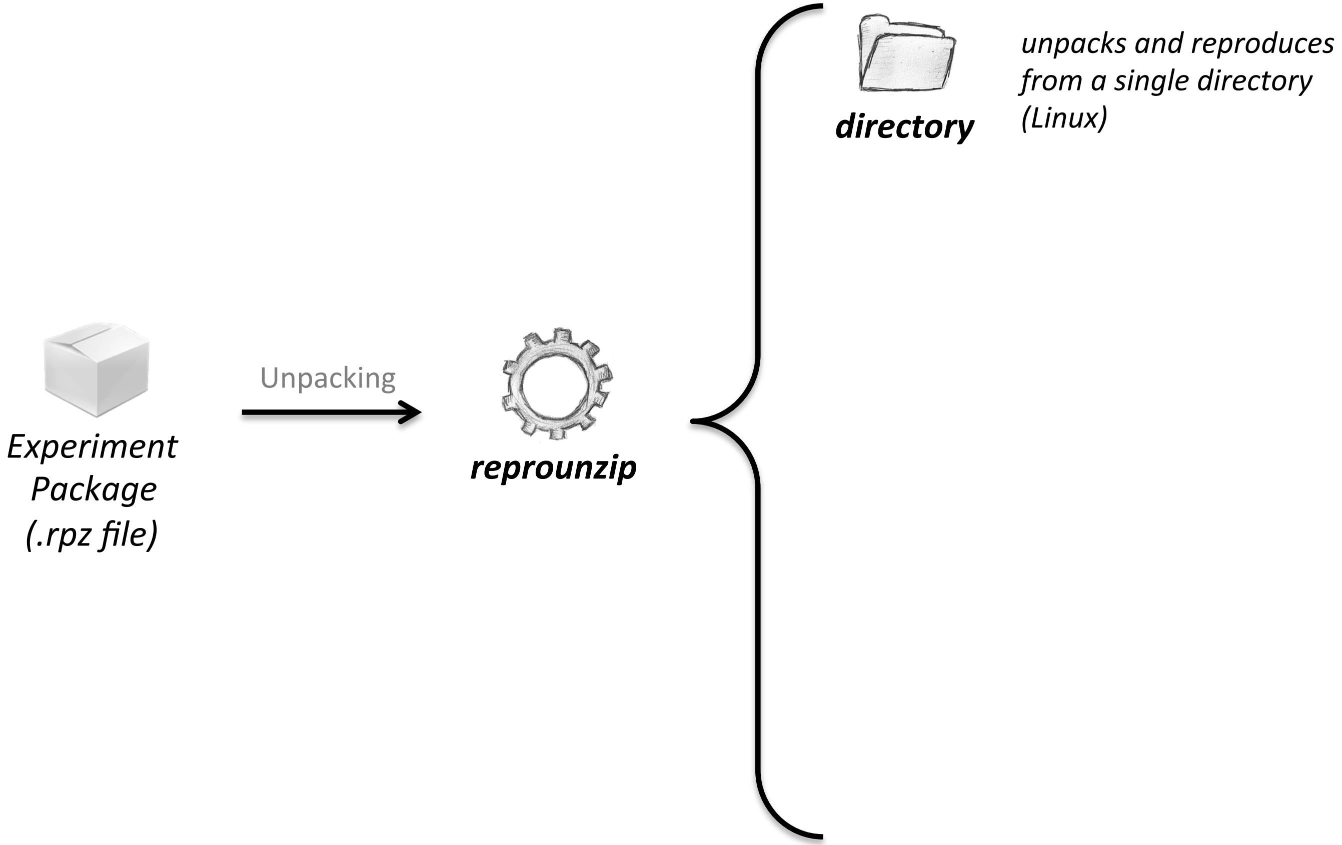
Computational Environment ***E’*** (potentially different than ***E***)



Unpacking Experiments

REVIEWERS
READERS

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



Experiment Package (.rpz file)

Unpacking

reprounzip

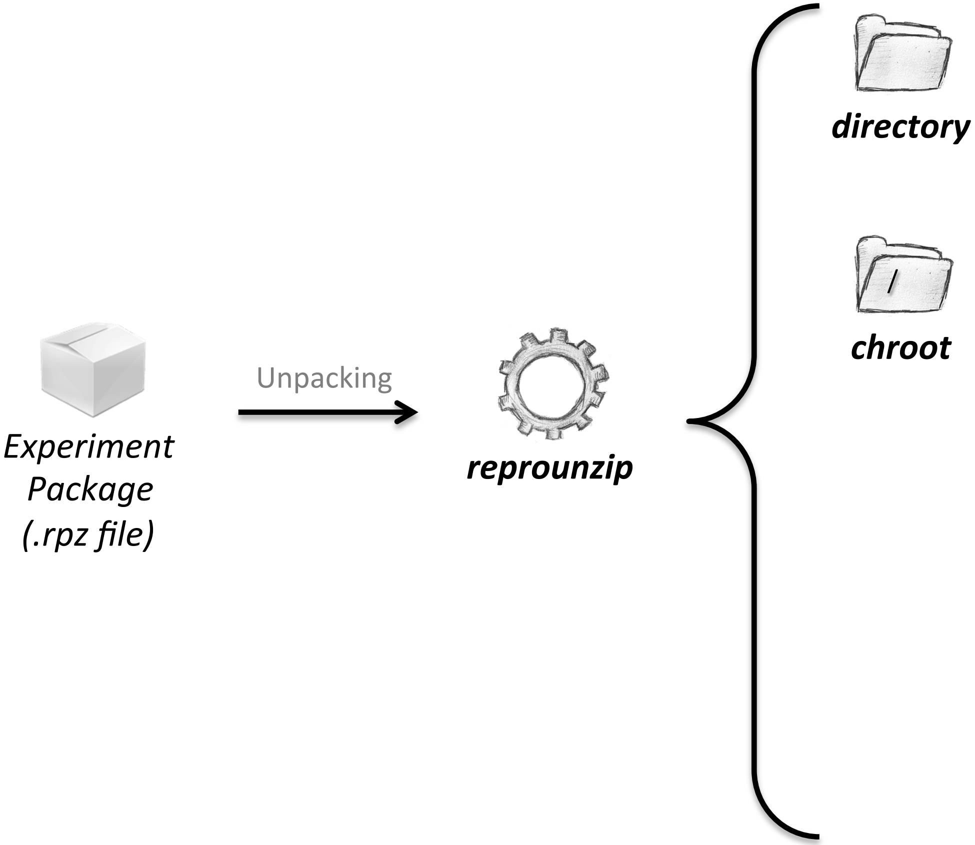
directory

unpacks and reproduces from a single directory (Linux)

REVIEWERS
READERS

Unpacking Experiments

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



Experiment Package (.rpz file)

Unpacking

reprounzip

directory

unpacks and reproduces from a single directory (Linux)

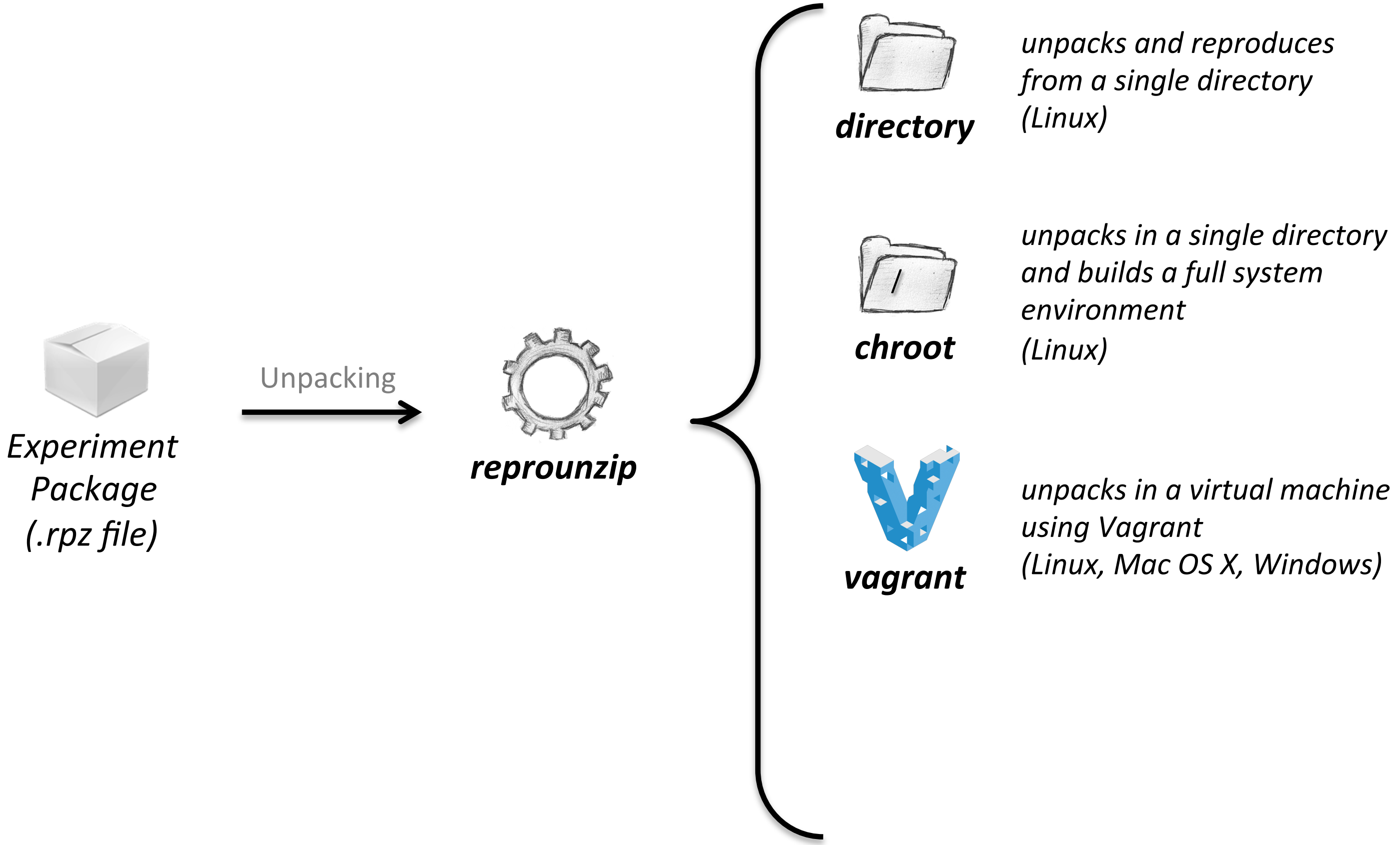
chroot

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

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)

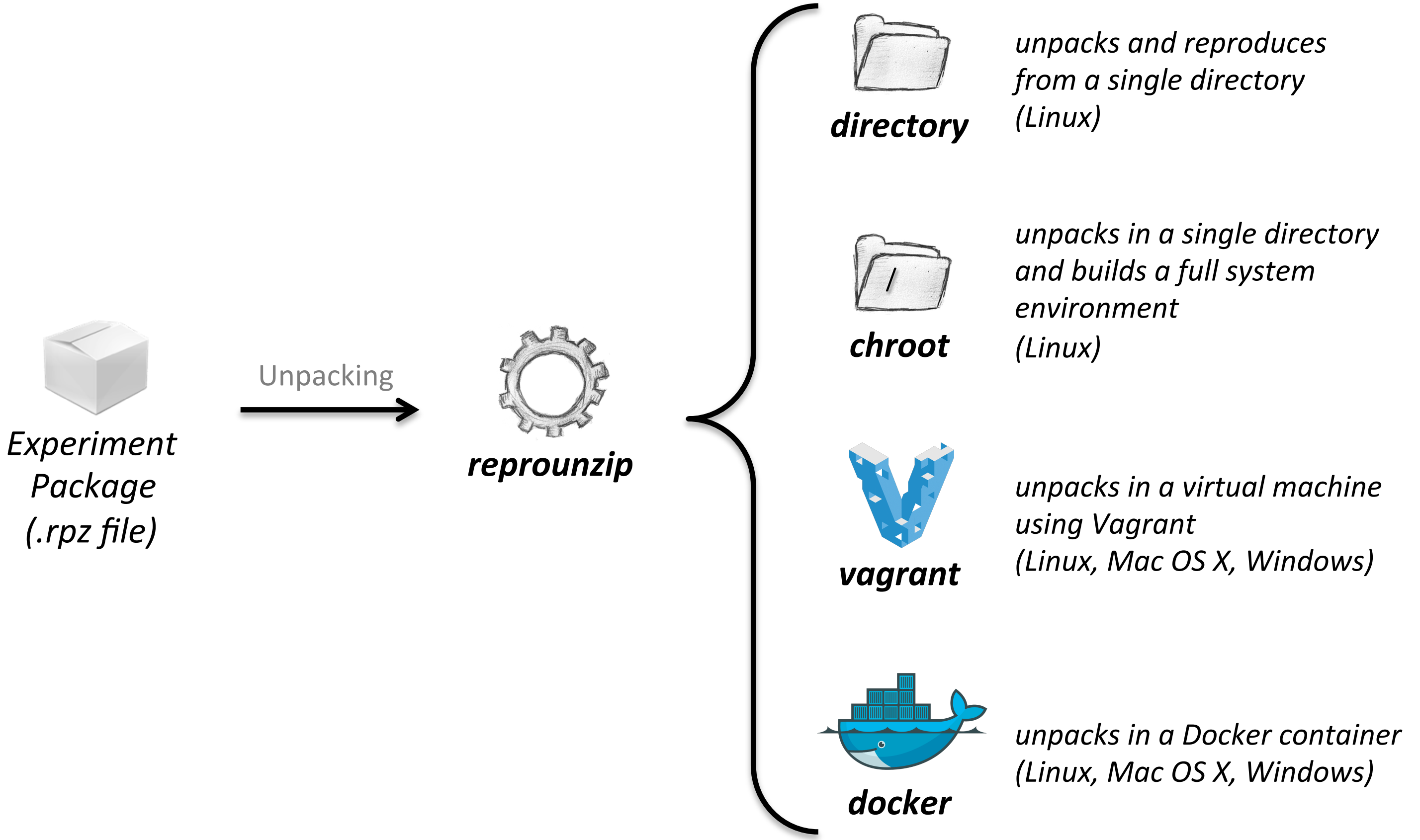
vagrant

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

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)*

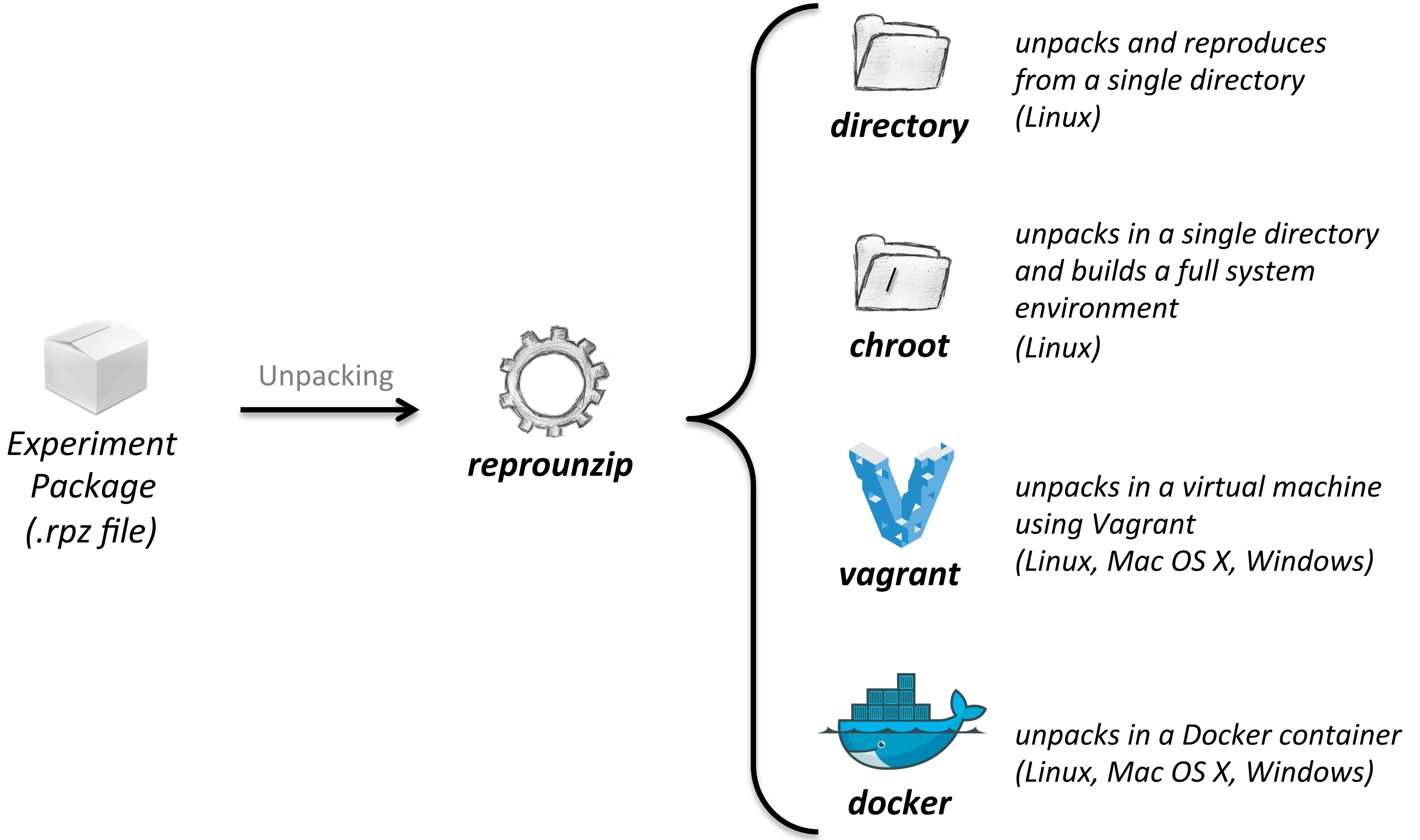
vagrant *unpacks in a virtual machine using Vagrant (Linux, Mac OS X, Windows)*

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

Unpacking Experiments

Reviewers
Readers

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



unpacks and reproduces from a single directory (Linux)

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

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

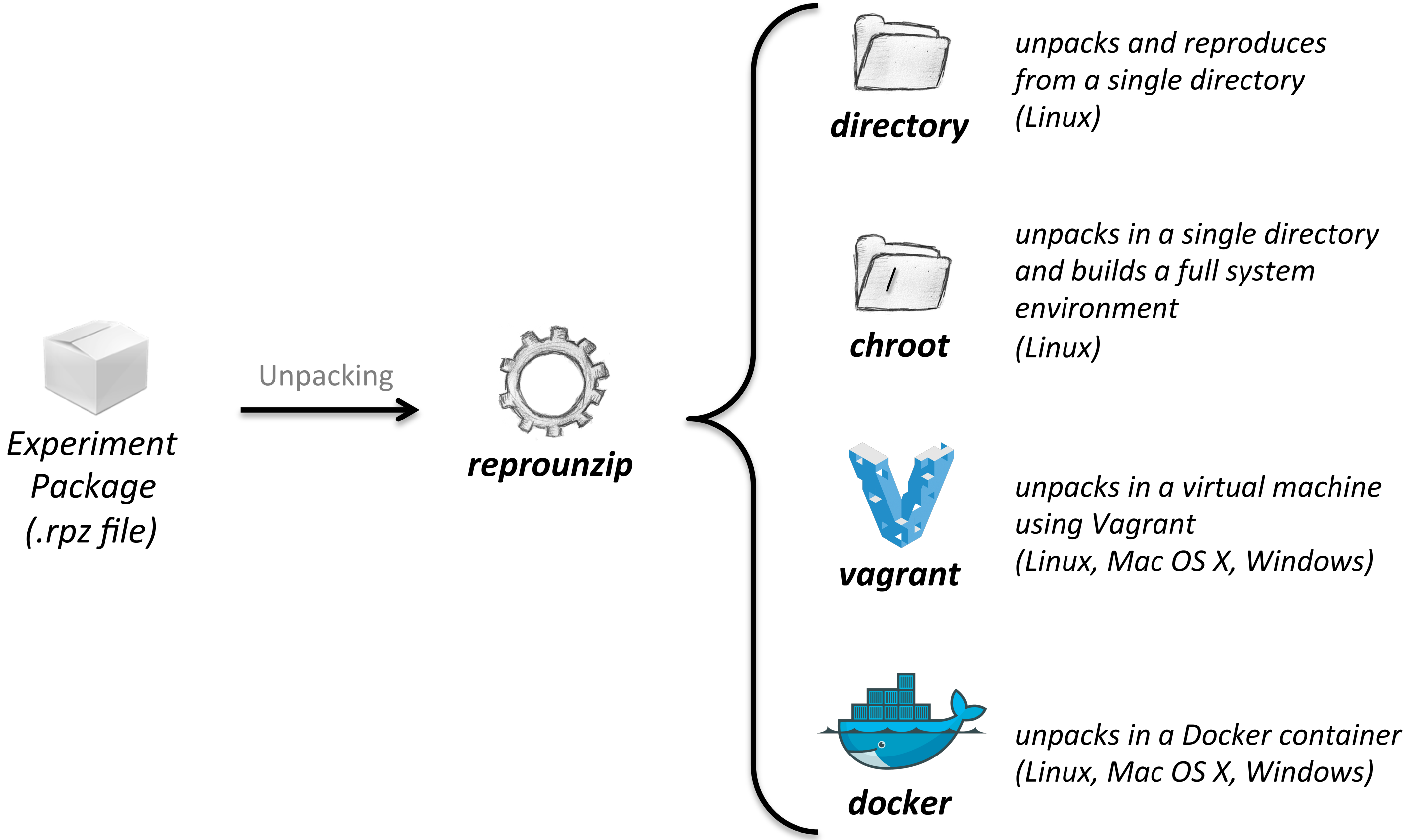
unpacks in a Docker container (Linux, Mac OS X, Windows)

Provenance Graph

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)*

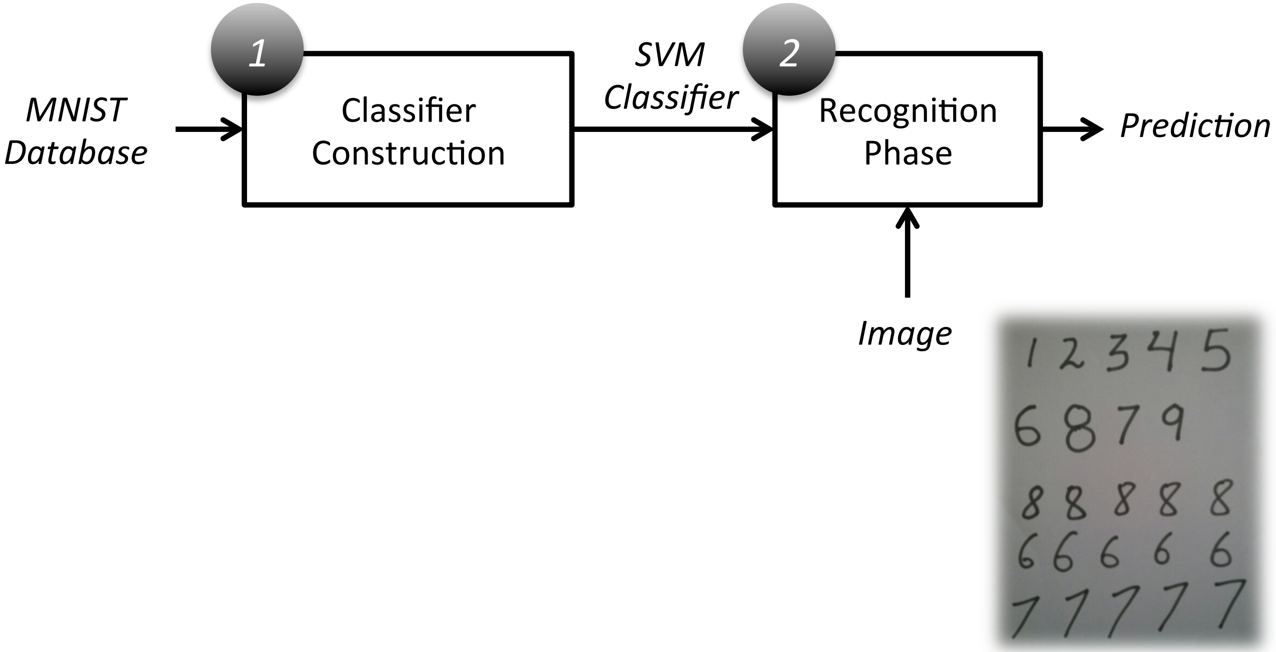
docker — *unpacks in a Docker container (Linux, Mac OS X, Windows)*

Provenance Graph

VisTrails

Example

Predicting the Value of a Handwriting Digit from an Image



MNIST Database

1 Classifier Construction

SVM Classifier

2 Recognition Phase

Prediction

Image

News!

ReproZip …

… has been adopted in the Bonneau Lab (NYU)

http://bonneaulab.bio.nyu.edu/

… has been used by the ACM SIGMOD 2015 Reproducibility Review

http://db-reproducibility.seas.harvard.edu/

… has been used by the Information Systems journal (Reproducibility Section)

http://www.journals.elsevier.com/information-systems/

… has been used for enabling automatic version upgrades of complex systems (work by Dennis Shasha and colleagues)

Limitations

Only packs experiments in Linux distros (yet…)

Does not guarantee reproducibility of distributed applications (yet…)

Only detects software packages in Debian-based environments (yet…)

Does not allow reproducibility of *non-deterministic* processes

Does not save *state*

Future Work

Creating reproducible packages in Mac OS X – *ongoing work*

Reproducibility of distributed applications – *ongoing work*

Identifying software packages in other systems

Proprietary software

Try it!

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

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

Mailing lists: *reprozip-users@vgc.poly.edu*
reprozip-dev@vgc.poly.edu

F. Chirigati, D. Shasha, and J. Freire: *Packing Experiments for Sharing and Publication*. In Proceedings of the 2013 International Conference on Management of Data (SIGMOD), pp. 977-980, 2013

F. Chirigati, D. Shasha, and J. Freire: *ReproZip: Using Provenance to Support Computational Reproducibility*. In Proceedings of the 5th USENIX conference on Theory and Practice of Provenance (TaPP), 2013

Send your feedback and interesting use cases!

Thanks!

Questions?

References

- Begley, C. and Ellis, L. *Drug development: Raise standards for preclinical cancer research*. Nature, 483(7391):531-533, 2012
- Cicerone, R. *Research Reproducibility, Replicability, Reliability*. In http://www.nasonline.org/about-nas/leadership/president/2015-presidents-address.pdf, 2005
- Guo, P. *CDE: run any Linux application on-demand without installation*. In Proceedings of LISA'11. 2011
- Janin, Y., Vincent, C., and Duraffort, R. *CARE, the comprehensive archiver for reproducible execution*. In Proceedings of TRUST '14, Article 1, 2014
- Nuzzo, R. *How scientists fool themselves – and how they can stop*. Nature, 526(7572):182–185, 2015
- Pham, Q., Malik, T., and Foster, I. *Using provenance for repeatability*. In Proceedings of TaPP '13, Article 2, 2013
- Vandewalle, P., Kovacevic, J., and Vetterli, M. *Reproducible research in signal processing*. Signal Processing Magazine, IEEE , vol.26, no.3, pp.37-47, 2009