Sciunits: Reusable Research Objects

(http://sciunit.run)

Tanu Malik

School of Computing,

College of Computing and Digital Media

DePaul University

DOMA Workshop Flatiron Institute Nov 16-17 2017

Share and Repeat an Application

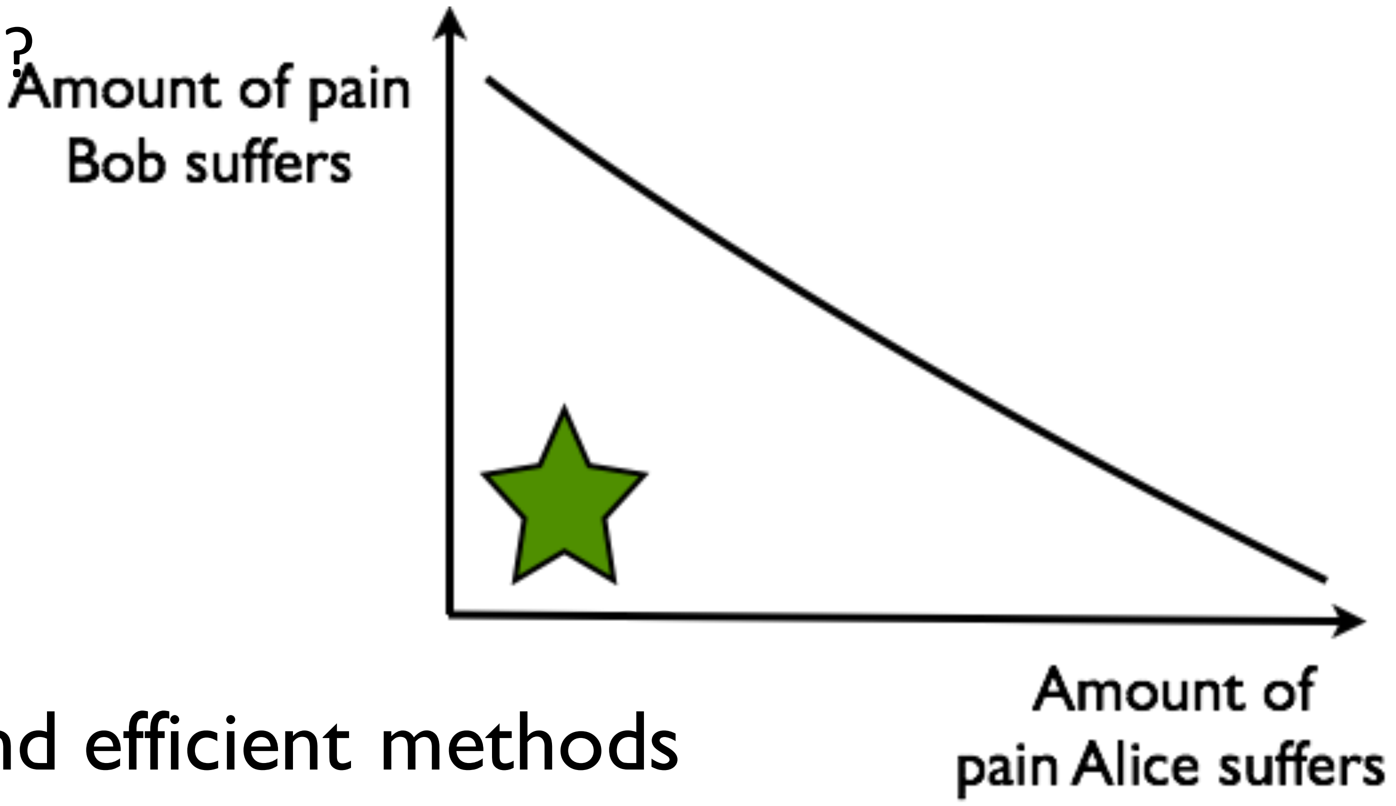
Alice wants to share her input data files and program source code with Bob
Bob wants to repeat Alice's application to validate her inputs and outputs.

Alice’s options

1. A tar and gzip
2. Build a website with model code, parameters, and data
3. Submit to a repository such as GitHub, DockerHub
4. Create a virtual machine

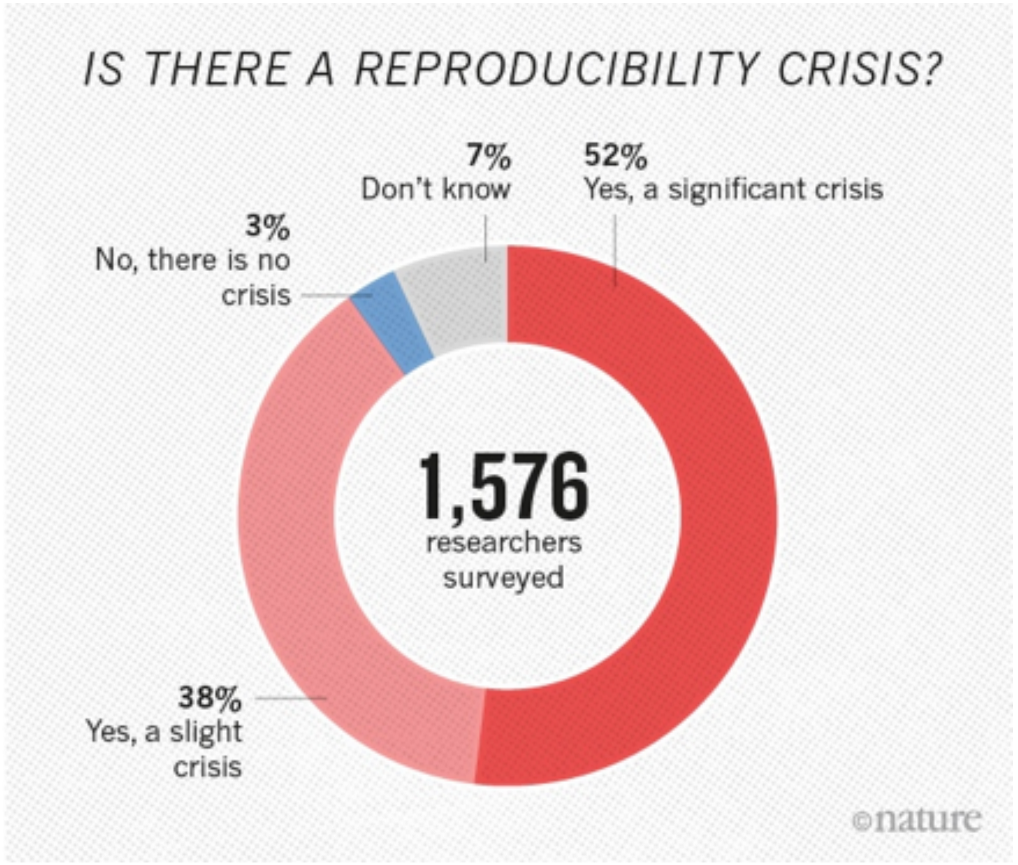
Bob’s frustration

- I do not find the lib.so required for building the model.
- How do I do this?



Lack of easy and efficient methods for sharing and reproducibility

Scientific Reproducibility Crisis



- Source: 1,500 scientists lift the lid on reproducibility. Nature Survey. Corrected 25th May, 2016. Accessed April 13th, 2016

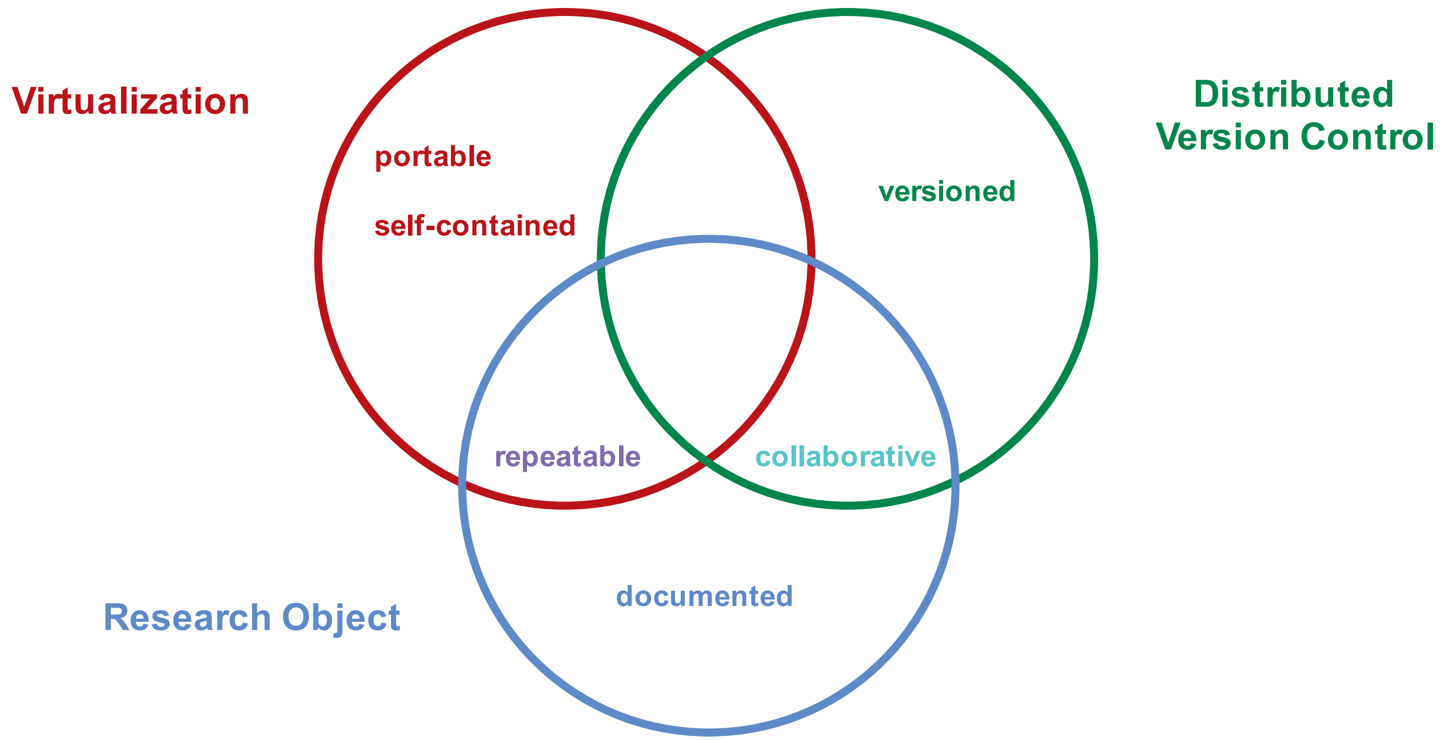
Data Sharing Crisis

Data Sharing		2011	2012
	Required as condition of publication	18	19
	Required but may not affect editorial decisions	3	10
	Encouraged/addressed, may be reviewed and/or hosted	35	30
	Implied	0	5
	No mention	114	106

Code Sharing		2011	2012
	Required as condition of publication	6	6
	Required but may not affect editorial decisions	6	6
	Encouraged/addressed, may be reviewed and/or hosted	17	21
	Implied	0	3
	No mention	141	134

[1]Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

Solution Space



No easily creatable, readily reusable, efficiently versioned, discrete unit of computation exists

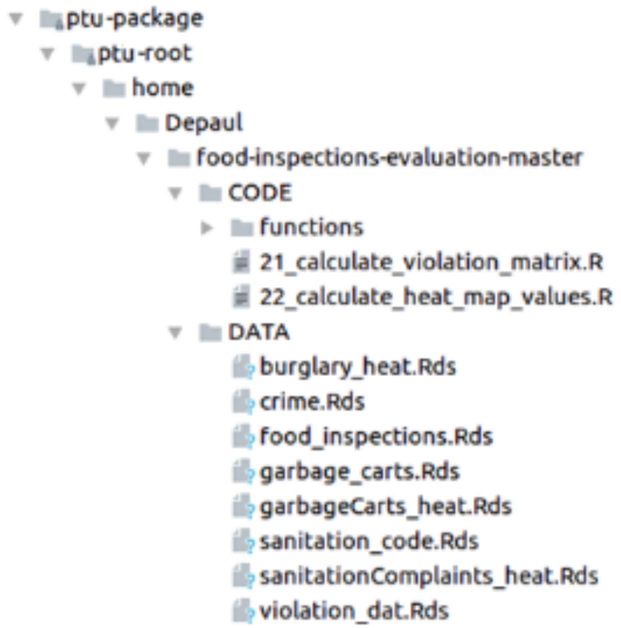
The Sciunit: A reusable research object

- Captures application executions
- Repeats executions
- Reproduces executions, changing input args
- Versioned executions stored as one sciunit
- Uses provenance for self-documentation

Demo

Packaging Details

1) Attach to process
2) Intercept system calls
3) Copy files / executables
4) Log system calls



```
▼ ptu-package
  ▼ ptu-root
    ▼ home
      ▼ Depaul
        ▼ food-inspections-evaluation-master
          ▼ CODE
            ▶ functions
              21_calculate_violation_matrix.R
              22_calculate_heat_map_values.R
          ▼ DATA
              burglary_heat.Rds
              crime.Rds
              food_inspections.Rds
              garbage_carts.Rds
              garbageCarts_heat.Rds
              sanitation_code.Rds
              sanitationComplaints_heat.Rds
              violation_dat.Rds
```

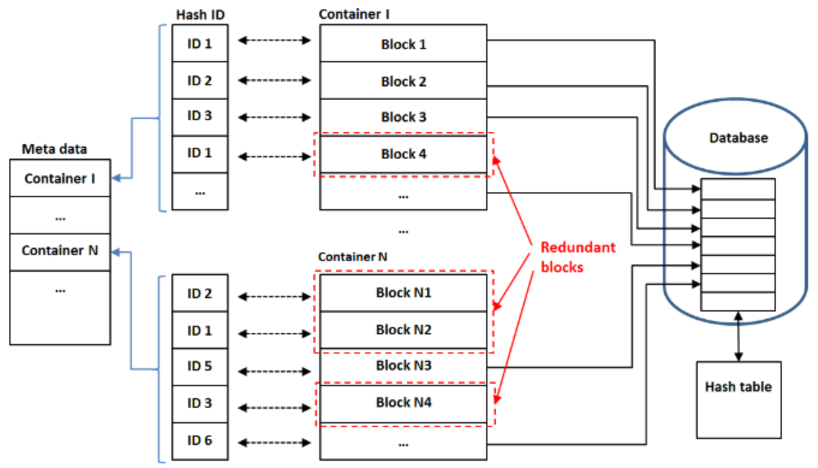
Storage and Retrieval

Store package:

1) Archive package-root
2) CDC on archive
3) Store manifest

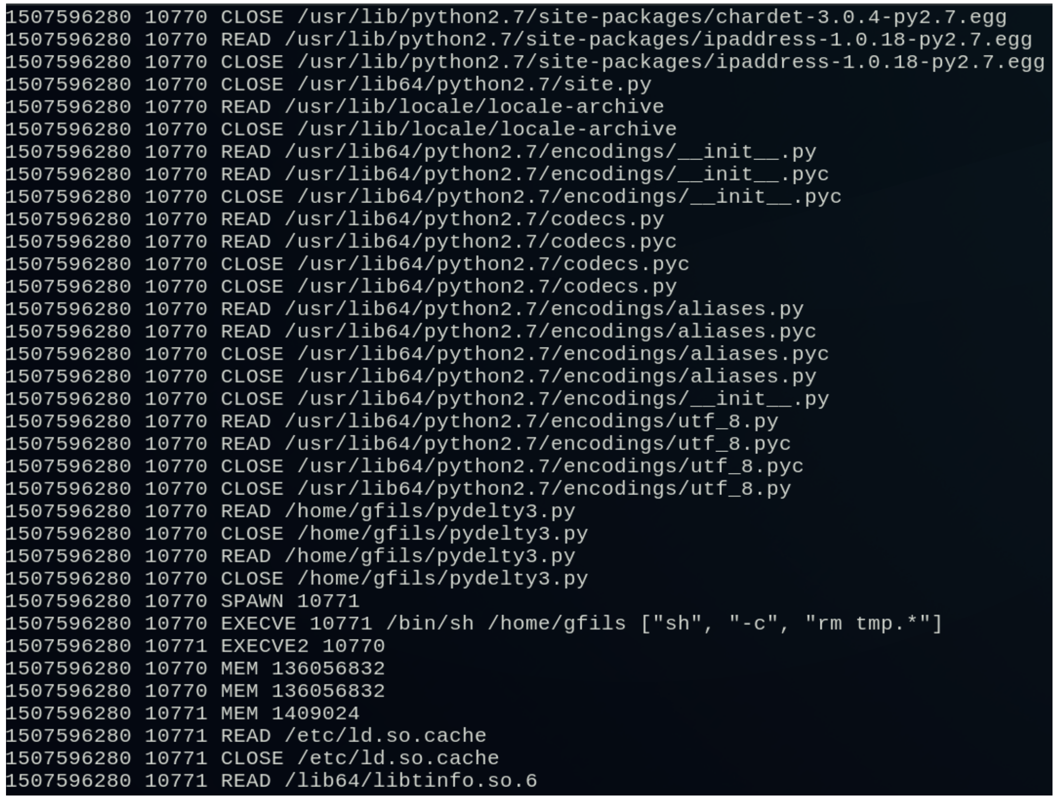
Retrieve package:

1) Retrieve manifest
2) Concatenate chunks
3) Extract archive



Deduplicated Container Storage

Provenance



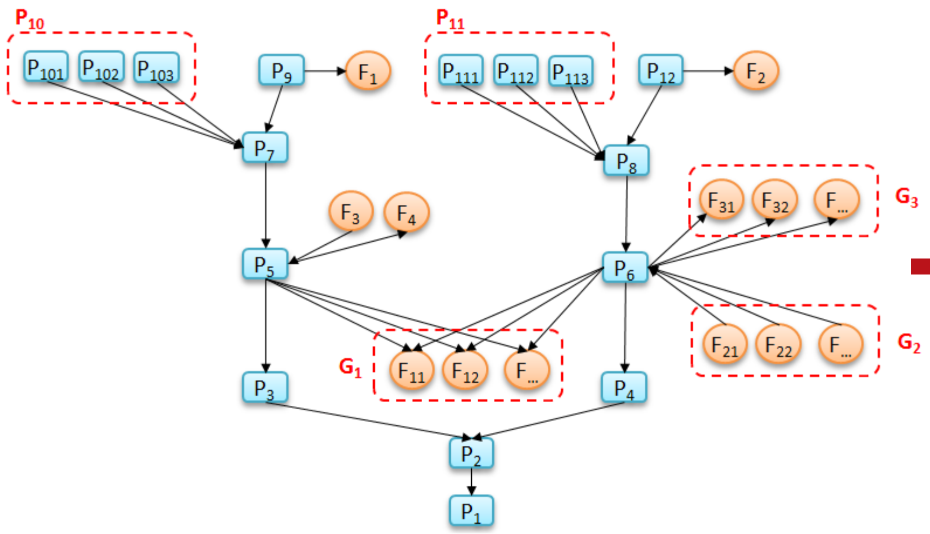
```
1507596280 10770 CLOSE /usr/lib/python2.7/site-packages/chardet-3.0.4-py2.7.egg
1507596280 10770 READ /usr/lib/python2.7/site-packages/ipaddress-1.0.18-py2.7.egg
1507596280 10770 CLOSE /usr/lib/python2.7/site-packages/ipaddress-1.0.18-py2.7.egg
1507596280 10770 CLOSE /usr/lib64/python2.7/site.py
1507596280 10770 READ /usr/lib/locale/locale-archive
1507596280 10770 CLOSE /usr/lib/locale/locale-archive
1507596280 10770 READ /usr/lib64/python2.7/encodings/__init__.py
1507596280 10770 READ /usr/lib64/python2.7/encodings/__init__.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/__init__.pyc
1507596280 10770 READ /usr/lib64/python2.7/codecs.py
1507596280 10770 READ /usr/lib64/python2.7/codecs.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/codecs.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/codecs.py
1507596280 10770 READ /usr/lib64/python2.7/encodings/aliases.py
1507596280 10770 READ /usr/lib64/python2.7/encodings/aliases.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/aliases.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/aliases.py
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/__init__.py
1507596280 10770 READ /usr/lib64/python2.7/encodings/utf_8.py
1507596280 10770 READ /usr/lib64/python2.7/encodings/utf_8.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/utf_8.pyc
1507596280 10770 CLOSE /usr/lib64/python2.7/encodings/utf_8.py
1507596280 10770 READ /home/gfils/pydelty3.py
1507596280 10770 CLOSE /home/gfils/pydelty3.py
1507596280 10770 READ /home/gfils/pydelty3.py
1507596280 10770 CLOSE /home/gfils/pydelty3.py
1507596280 10770 SPAWN 10771
1507596280 10770 EXECVE 10771 /bin/sh /home/gfils ["sh", "-c", "rm tmp.*"]
1507596280 10771 EXECVE2 10770
1507596280 10770 MEM 136056832
1507596280 10770 MEM 136056832
1507596280 10771 MEM 1409024
1507596280 10771 READ /etc/ld.so.cache
1507596280 10771 CLOSE /etc/ld.so.cache
1507596280 10771 READ /lib64/libtinfo.so.6
```

Part Of A Normal (Verbose) Provenance Log

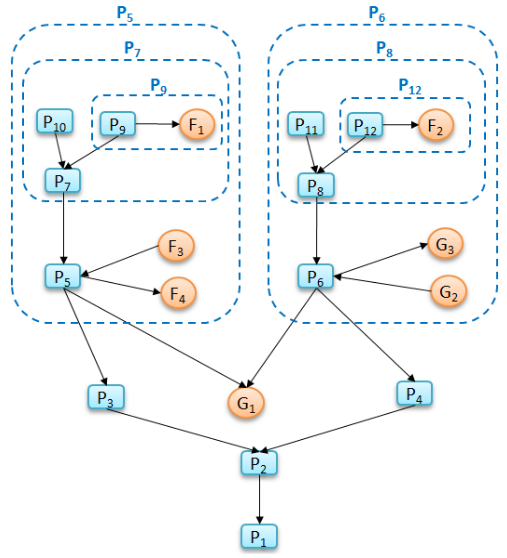
Small Section Of Graph Built From Normal Provenance Log

Summarization: Group By Similarity

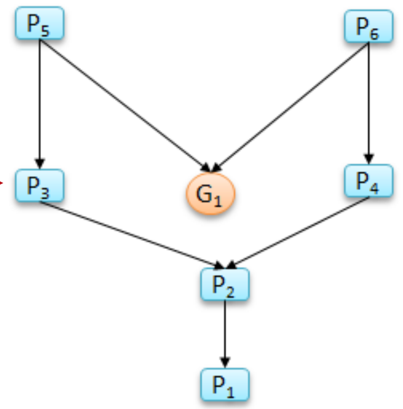
- Group vertices by type/connections
- Find min-connected nodes, pack into hubs


P_{10}
P_{101} P_{102} P_{103}
P_9 F_1
P_{11}
P_{111} P_{112} P_{113}
P_{12} F_2
P_7 P_8
F_3 F_4
F_{31} F_{32} $F_{...}$
G_3
P_5 P_6
F_{21} F_{22} $F_{...}$
G_2
G_1
F_{11} F_{12} $F_{...}$
P_3 P_4
P_2
P_1

Full Graph



Similarity Applied

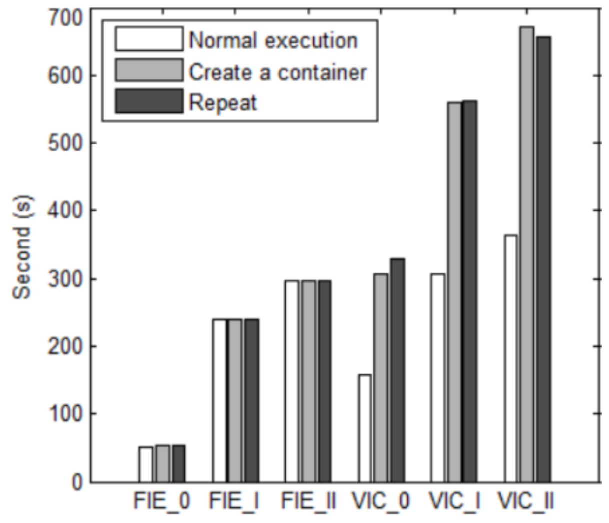


Packability Applied

Package and Repeat

1) Run app normally
2) Run with package
3) Run with repeat

- I/O-intensive apps: VIC
- Non-I/O-intensive apps: FIE

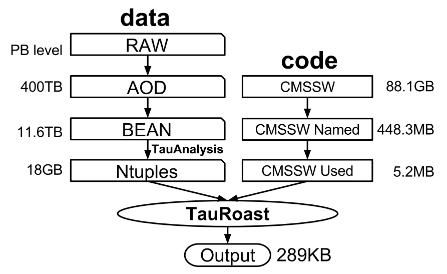


Package/Repeat Runtimes

Use Cases

- City of Chicago Food Inspections Evaluation Model (Data Mining)
- Four applications
- Two languages
- 130 files
- 1580 dependencies
- 908 MB

- Atlas and CMS
- TauRoast and Athena
- Python and C-based event reconstruction and data reduction
- Used code and configuration are dynamic depend-
  ing upon input data,



- Variable Infiltration Capacity
- Four applications
- Five languages
- 7 GB

- Jupyter Notebooks
- December 12-13 at the American Geophysical Union

Conclusions and current work

sciunit is a portable, self-contained, and inherently understandable versioned unit of computation.

- Graph summarization testing
- Database applications
- Exact partial repeatability
- Apps with network-operations
- Parallel HPC applications
- Emerging reusable object formats

Links and Acknowledgements

National Science Foundation grants ICER-1639759, ICER-1661918, ICER-1440327, ICER-1343816

Website:

- https://sciunit.run

Sciunit paper:

- https://arxiv.org
- Search for "*sciunit*"