PROBE4RSE: Provenance Replay/Observation Engine for Research Software Engineers

by Samuel Grayson, Reed Milewicz, Daniel S. Katz, and Darko Marinov

University of Illinois Urbana-Champaign

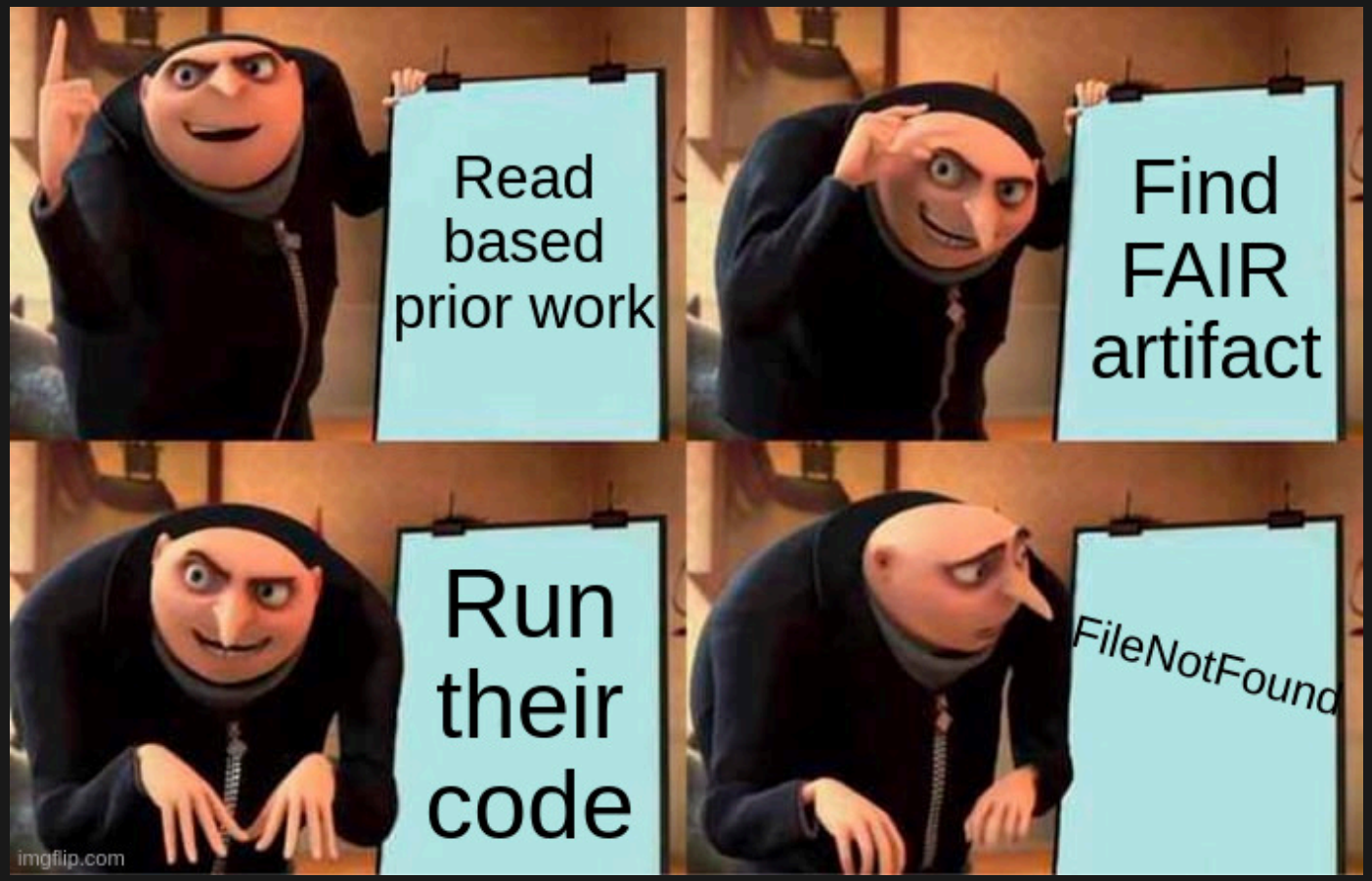
Sandia National Laborato

Takeaways

1. Provenance is useful (record/replay and more!)
2. Consider using **PROBE** to collect provenance.
3. Looking to collaborate on complex use-cases

https://github.com/charmoniumQ/PROBE

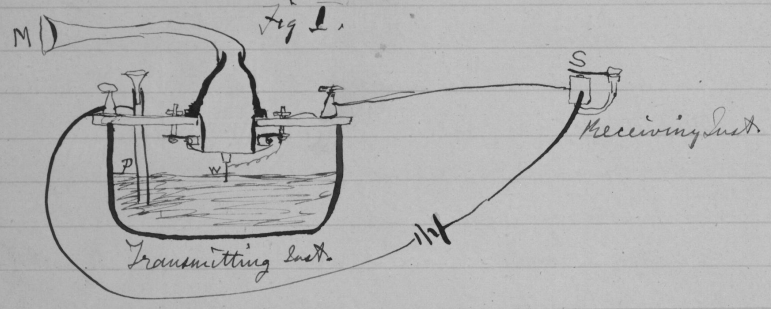
Has this every happend to you?



What is provenance?

March 10th 1876

Fig 1.



1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece — and S the armature of the Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting Instrument was placed in another room and the doors of both rooms were closed.

I then shouted into M the following sentence: "Mr. Watson — Come here — I want to

see you". To my delight he came and declared that he had heard and understood what I said.

I asked him to repeat the words — ~~He said~~ He answered "You said 'Mr. Watson — come here — I want to see you'." We then changed places and I listened at S while Mr. Watson read a few passages from a book into the mouth piece M. It was certainly the case that articulate sounds proceeded from S. The effect was loud but indistinct and muffled.

If I had read beforehand the passage given by Mr. Watson I should have recognized every word. As it was I could not make out the sense — but an occasional word here and there was quite distinct. I made out "to" and "out" and "further"; and finally the sentence "Mr. Bell Do you understand what I say? DO-YOU-UNDER-STAND-WHAT-I-SAY" came quite clearly and intelligibly. No sound was audible when the armature S was removed.

What is computational provenance?

1. Process by which a file was generated
2. The inputs to that process
3. The provenance of those inputs (recursively)

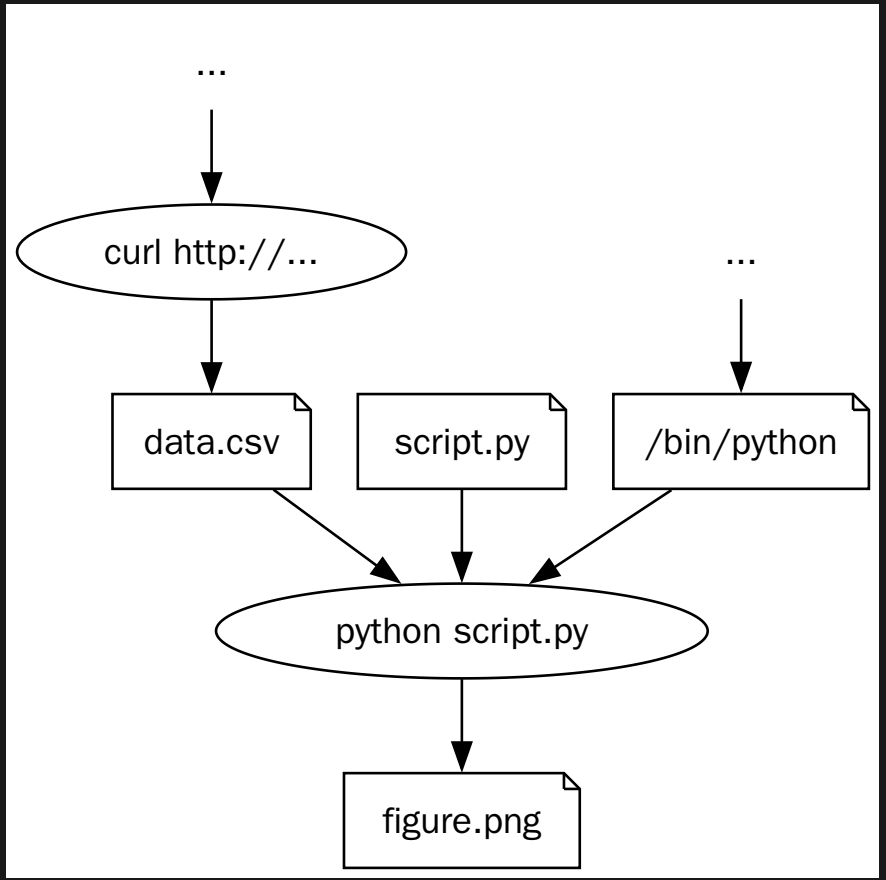
Graph on processes and files

Comp. Provenance Example

```
curl -o data.csv http://data.com
python script.py
```

Comp. Provenance Example

```
curl -o data.csv http://data.com
python script.py
```



Prior works in provenance collection

- Workflow-level provenance
- Language-level provenance
- Provenance standards (PASSv2, OpenLineage)

Prior works in record/replay

- CDE (Guo and Engler 2011), Sciunit (Ton That et al. 2017), RR (O'Callahan et al. 2017), CARE (Yves et al. 2014), ReproZip (Chirigati et al. 2016)
- Speed
- Robustness of reproducibility
- Openness to downstream analysis
- Reuse environment for new exe

PROBE

- `LD_PRELOAD` incomplete! but good enough in practice
- Less than 2x slowdown in most cases (improving)
- No root!
- `$ probe record <your-command-here>`
- 2 implemented and 2 planned features

Understand dataflow in your pile of scripts

Contributed by Shofiya Bootwala (new grad applying for PhD)

Understand dataflow in your pile of scripts

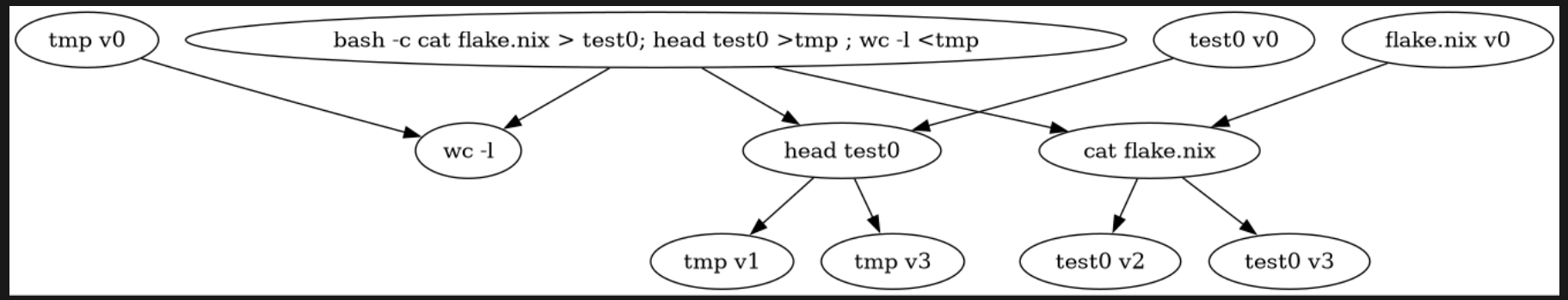
```
$ probe record ./plot.sh
```

Contributed by Shofiya Bootwala (new grad applying for PhD)

Understand dataflow in your pile of scripts

```
$ probe record ./plot.sh
$ probe export dataflow-graph
```

Contributed by Shofiya Bootwala (new grad applying for PhD)


tmp v0
bash -c cat flake.nix > test0; head test0 >tmp ; wc -l <tmp
test0 v0
flake.nix v0
wc -l
head test0
cat flake.nix
tmp v1
tmp v3
test0 v2
test0 v3

Create container automatically

Contributed by Asif Zubayer Palak
(new grad applying for PhD)

Create container automatically

```
$ probe record ./plot.sh
```

Contributed by Asif Zubayer Palak
(new grad applying for PhD)

Create container automatically

```
$ probe record ./plot.sh
$ probe export docker-image experiment:1
```

Contributed by Asif Zubayer Palak
(new grad applying for PhD)

Create container automatically

```
$ probe record ./plot.sh
$ probe export docker-image experiment:1
$ docker run experiment:1.0.1
```

Contributed by Asif Zubayer Palak
(new grad applying for PhD)

Create Makefile automatically (planned feature)

Contributed by Kyrillos Ishak (new grad applying for PhD)

Create Makefile automatically (planned feature)

```
$ probe record ./plot.sh 42
```

Contributed by Kyrillos Ishak (new grad applying for PhD)

Create Makefile automatically (planned feature)

```
$ probe record ./plot.sh 42
$ probe export makefile
```

Contributed by Kyrillos Ishak (new grad applying for PhD)

Create Makefile automatically (planned feature)

```
$ probe record ./plot.sh 42
$ probe export makefile
$ cat Makefile
input.csv: generate_data.py /bin/python
    ./generate_data.py --foo=42
plot.png: input.csv plot.py /bin/python
    ./plot.py
```

Contributed by Kyrillos Ishak (new grad applying for PhD)

What libraries does cmd use? (Planned feature)

What libraries does cmd use? (Planned feature)

```
$ probe record ./plot.sh
```

What libraries does cmd use? (Planned feature)

```
$ probe record ./plot.sh
$ probe export libs
  apt-get:
    - name: python3
      version: 3.12.4
  pip:
    - name: numpy
      version: 2.2.4
```

What libraries does cmd use? (Planned feature)

```
$ probe record ./plot.sh
$ probe export libs
  apt-get:
    - name: python3
      version: 3.12.4
  pip:
    - name: numpy
      version: 2.2.4
```

Heuristic-based, imperfect, **useful**

Performance and portability

Python → Rust by Jenna Fligor (ugrad applying for internships)

Performance analysis by Saleha Muzammil (newgrad applying for PhD)

Performance and portability

- Rust record (statically-linked) + Python extras

Python → Rust by Jenna Fligor (ugrad applying for internships)

Performance analysis by Saleha Muzammil (newgrad applying for PhD)

Performance and portability

- Rust record (statically-linked) + Python extras
- Preliminary results show `LD_PRELOAD` (1.1x) faster than `ptrace` (2x)

Python → Rust by Jenna Fligor (ugrad applying for internships)

Performance analysis by Saleha Muzammil (newgrad applying for PhD)

Future prov applications

- Remote provenance
- Compare two runs (diff of intermediate results)
- Using provenance to generate Spack package
- Interoperable prov representation

Thank you

Please give us feedback and issues

https://github.com/charmoniumQ/PROBE

sam@samgrayson.me