Provenance

Using Provenance to Improve Debugging Support for Data Scientists

CPSC 508 Conference
April 8, 2021

Alison Li and James Yoo

Provenance?

> “ where a piece of data came from
> and the process by which it arrived in
> the database… [Buneman, 2001]

Database Provenance

Figure adapted from “The Fine Line between Bold and Fringe Lunatic” [Seltzer, 2020]

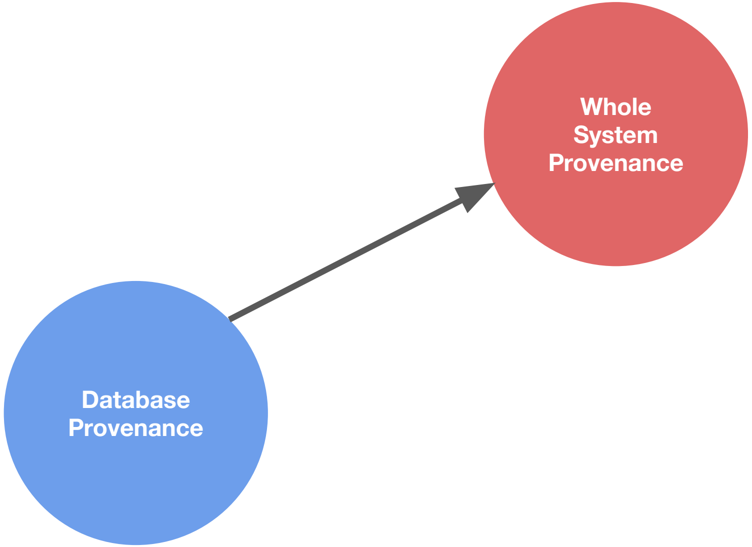


Figure adapted from “The Fine Line between Bold and Fringe Lunatic” [Seltzer, 2020]

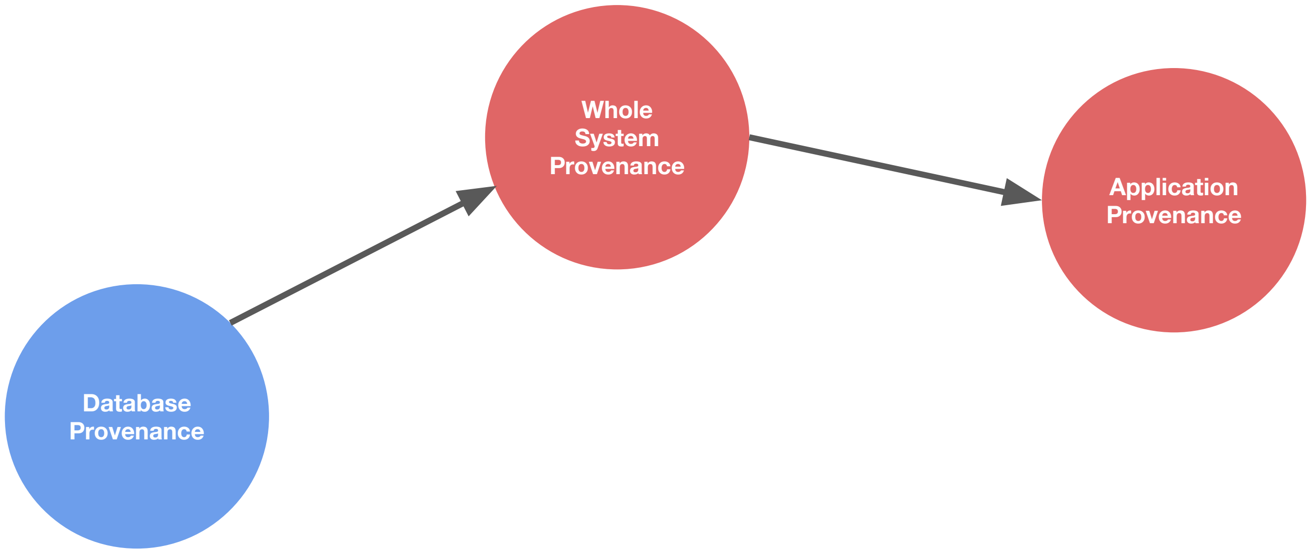


Figure adapted from “The Fine Line between Bold and Fringe Lunatic” [Seltzer, 2020]

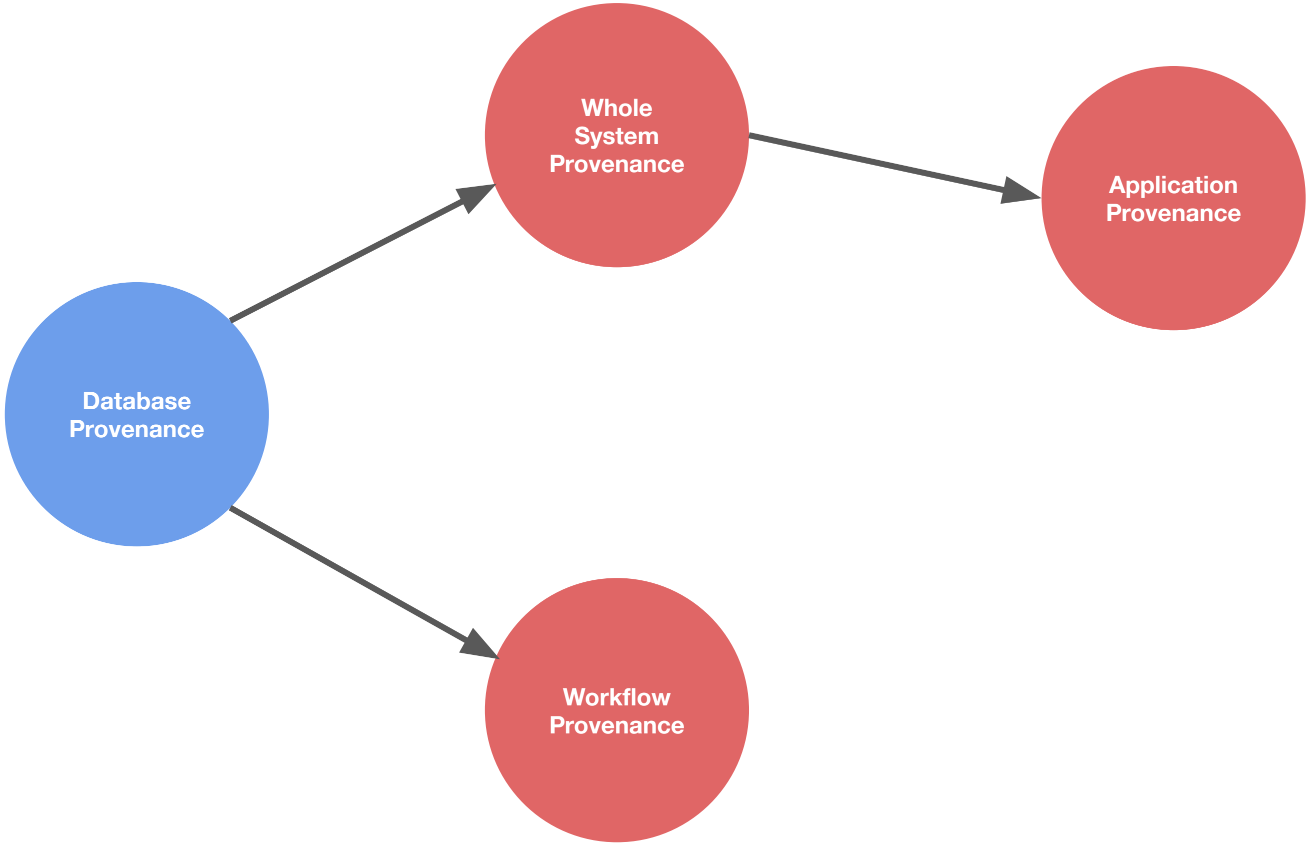


Figure adapted from “The Fine Line between Bold and Fringe Lunatic” [Seltzer, 2020]

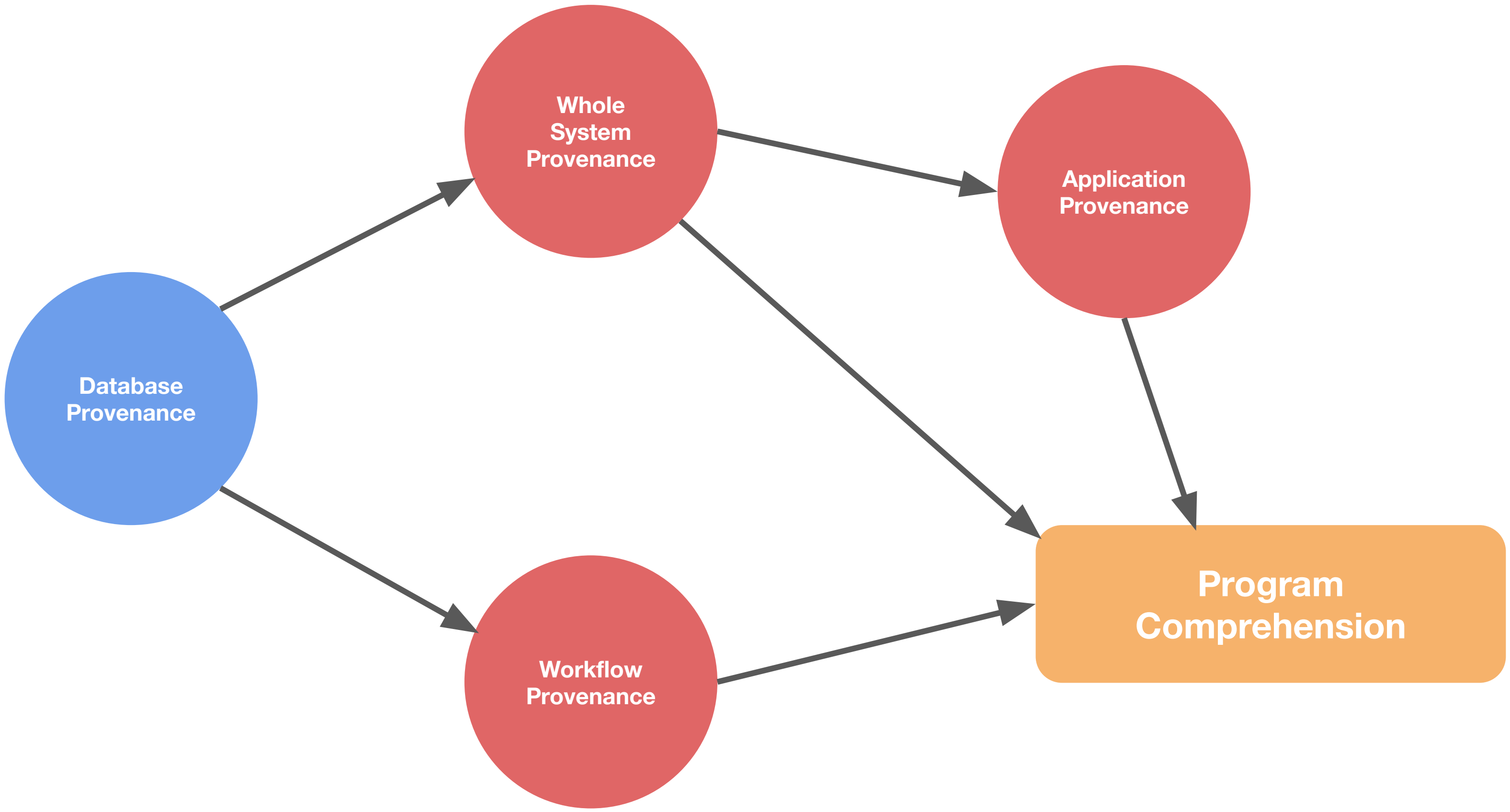


Figure adapted from “The Fine Line between Bold and Fringe Lunatic” [Seltzer, 2020]

Data Science Workflows

Data scientists ...

… engage in exploratory programming. [Subramanian et al. 2019]

… spend a large amount of time re-running scripts that they frequently modify. [Hu et al., 2020]

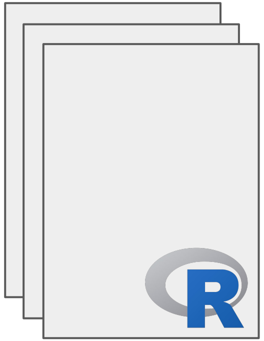
… are cautious of deleting code and prefer using informal versioning methods for storing code, such as in code comments. [Kery et al., 2017]

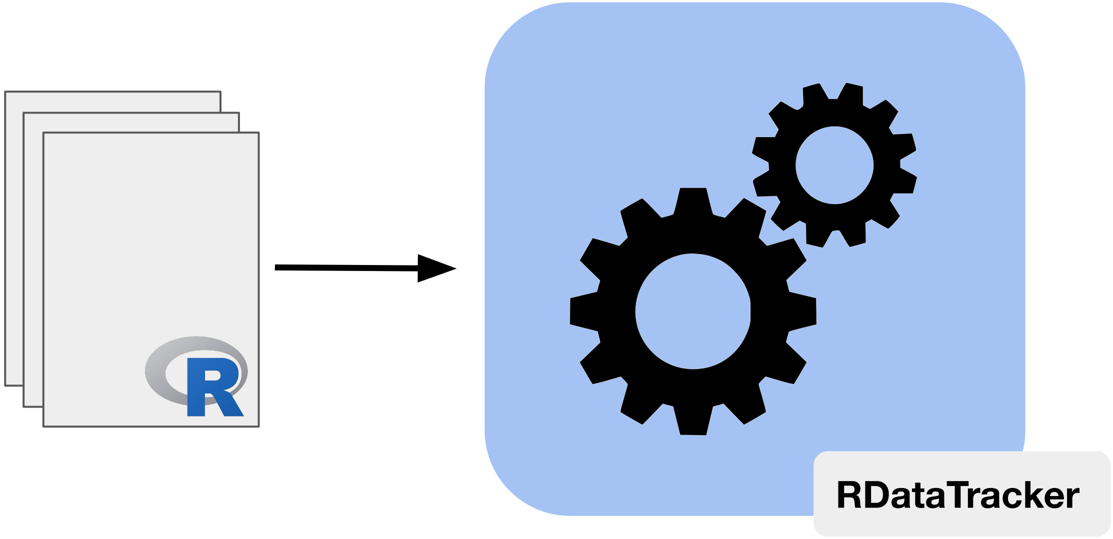
… lack motivation to document lower-level decision-making. [Zhang et al., 2020]

Multilingual Provenance Debugger

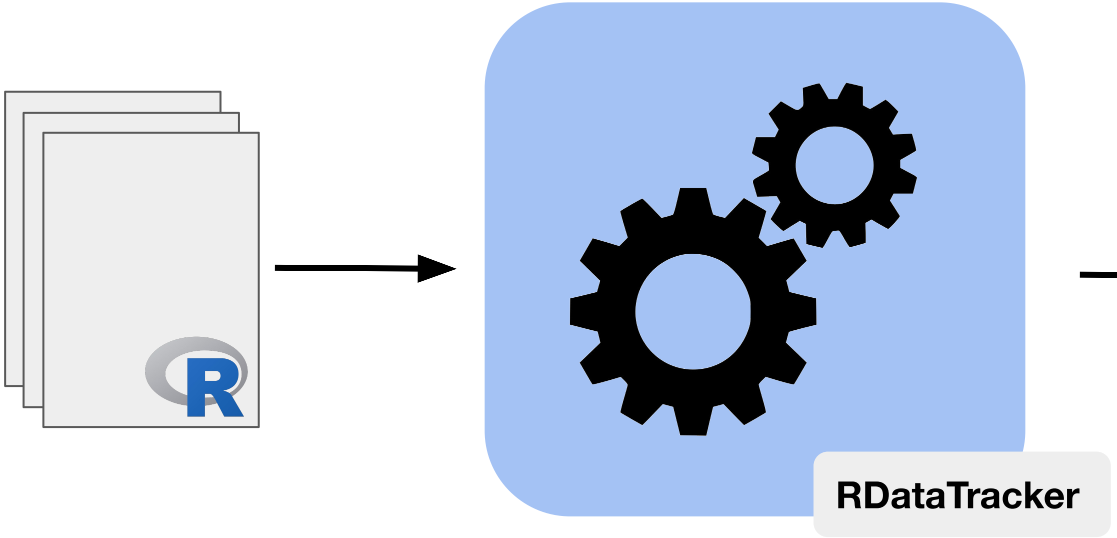
Existing Features

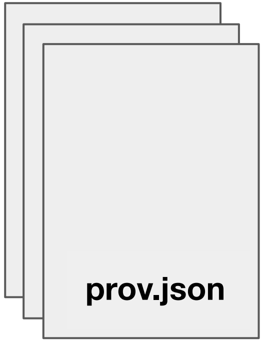
- “time-travelling” debugger
- exploits language-level provenance
- re-constructs the execution of a script
- find intermediate data values used/created/modified
- trace the lineage of variables
- search error messages on Stack Overflow



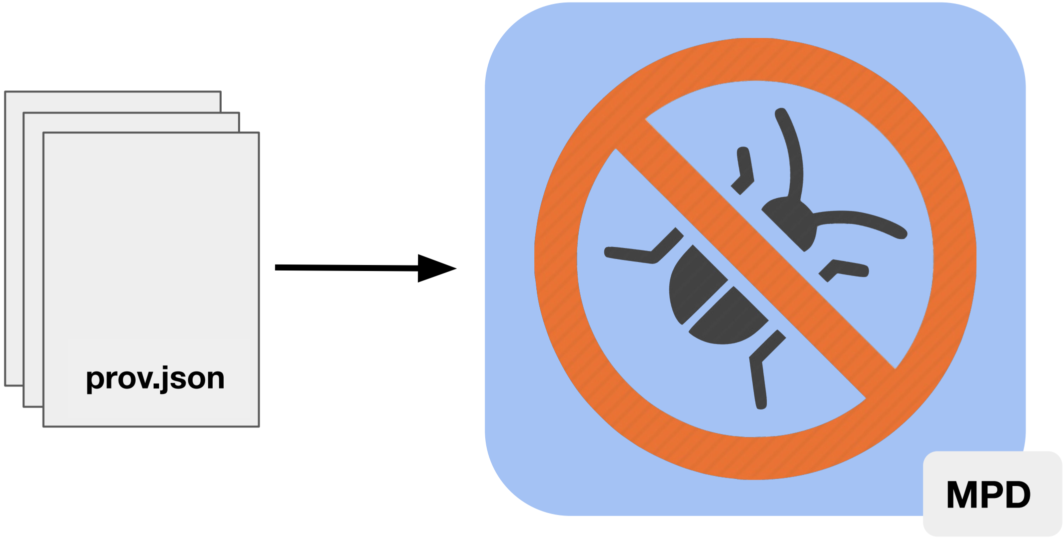


RDataTracker


RDataTracker
prov.json

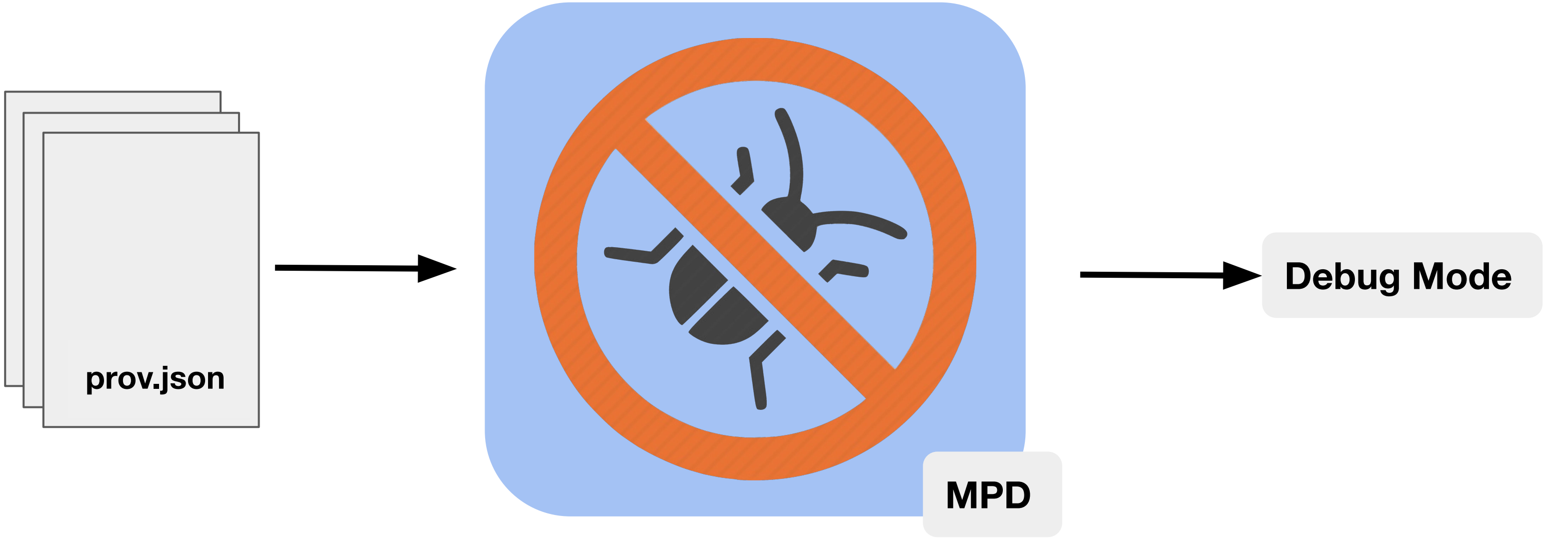


prov.json



prov.json

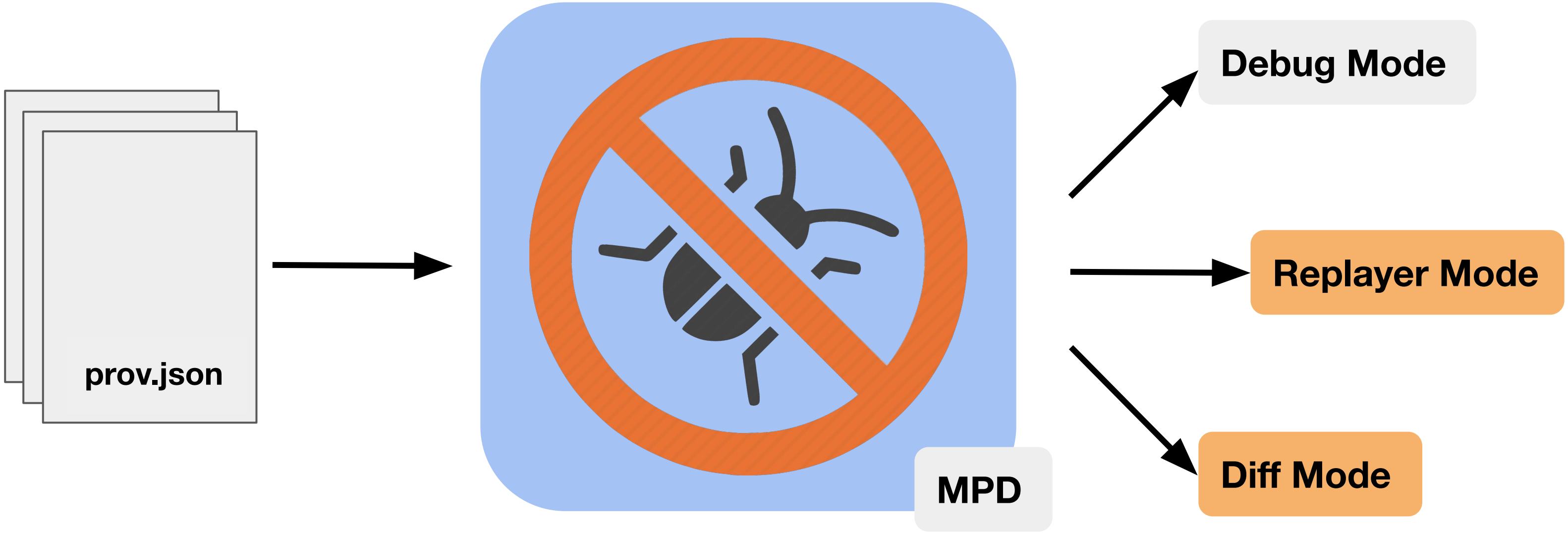
MPD



prov.json

MPD

Debug Mode



prov.json

MPD

Debug Mode

Replayer Mode

Diff Mode

RQ1

Can provenance collected during a debugging session provide an accurate representation of a programmer's debugging workflow?

RQ2

Can the comparison of PROV-JSON files generated for a program aid in defect identification?

Replayer Mode

Implementation

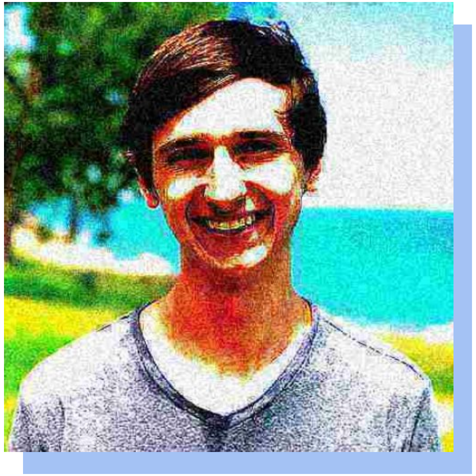
- record debugger interaction and associated provenance nodes
- user annotations for each debugging interaction
- obtain a debugging record + use it to automatically generate issues and communicate debugging steps

Diff Mode

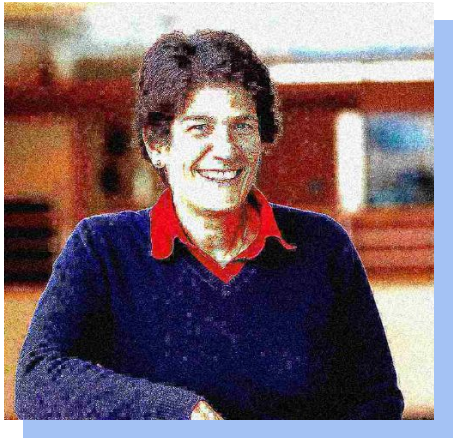
Implementation

- model each script as a set of nodes using data provenance, compute similarities/differences
- let users set a “threshold of similarity” calculated using the longest common subsequence (LCS)
- retrieve information about line number and program content

Acknowledgements



Joe “ProvBoi” Wonsil



Margo “M-Dawg” Seltzer

Reto “R-diddy” Achermann

Chris “Durian-Grey” Chen