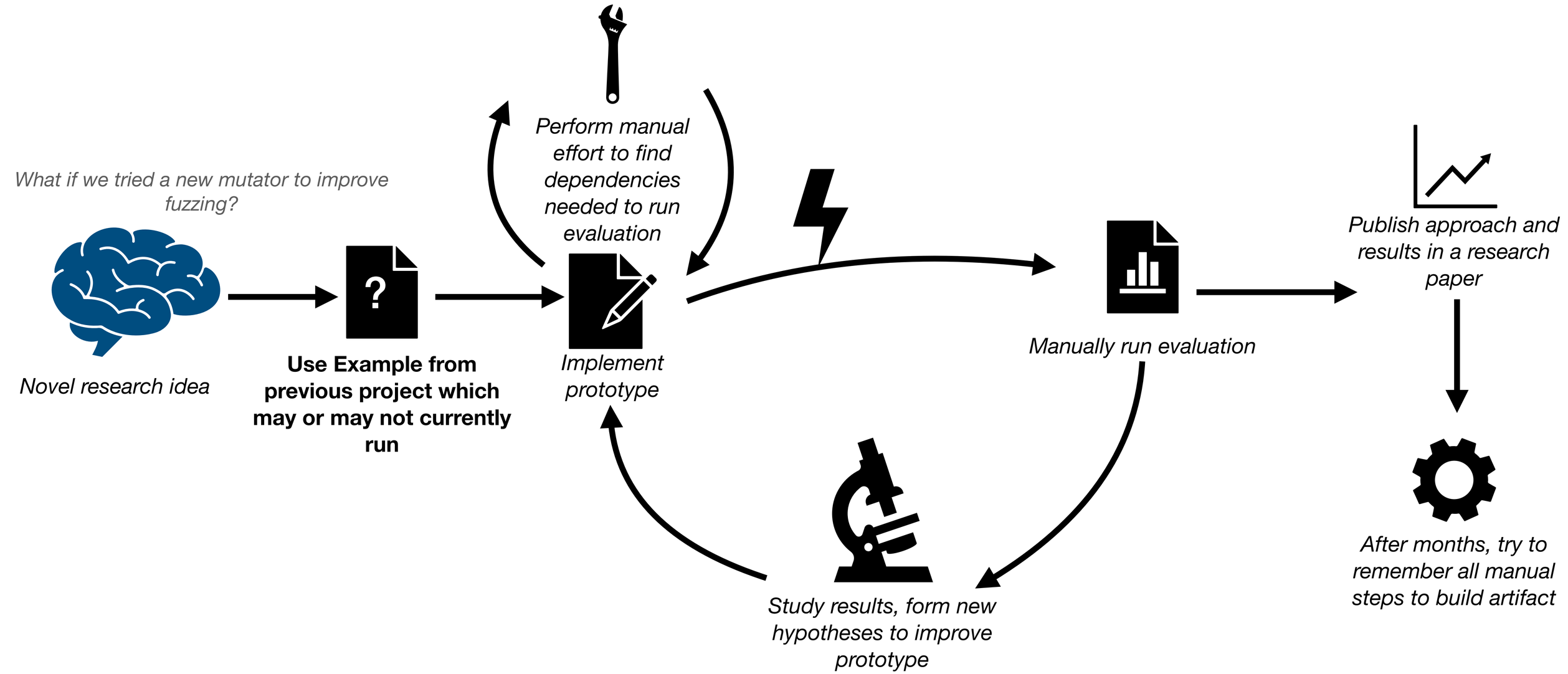
Enabling Continuous Large Scale Software Engineering Experimentation in the Cloud

Jonathan Bell (NEU)
Christopher Timperley, Michael Hilton and Lauren Herckis (CMU)

More resources: https://classee.cloud

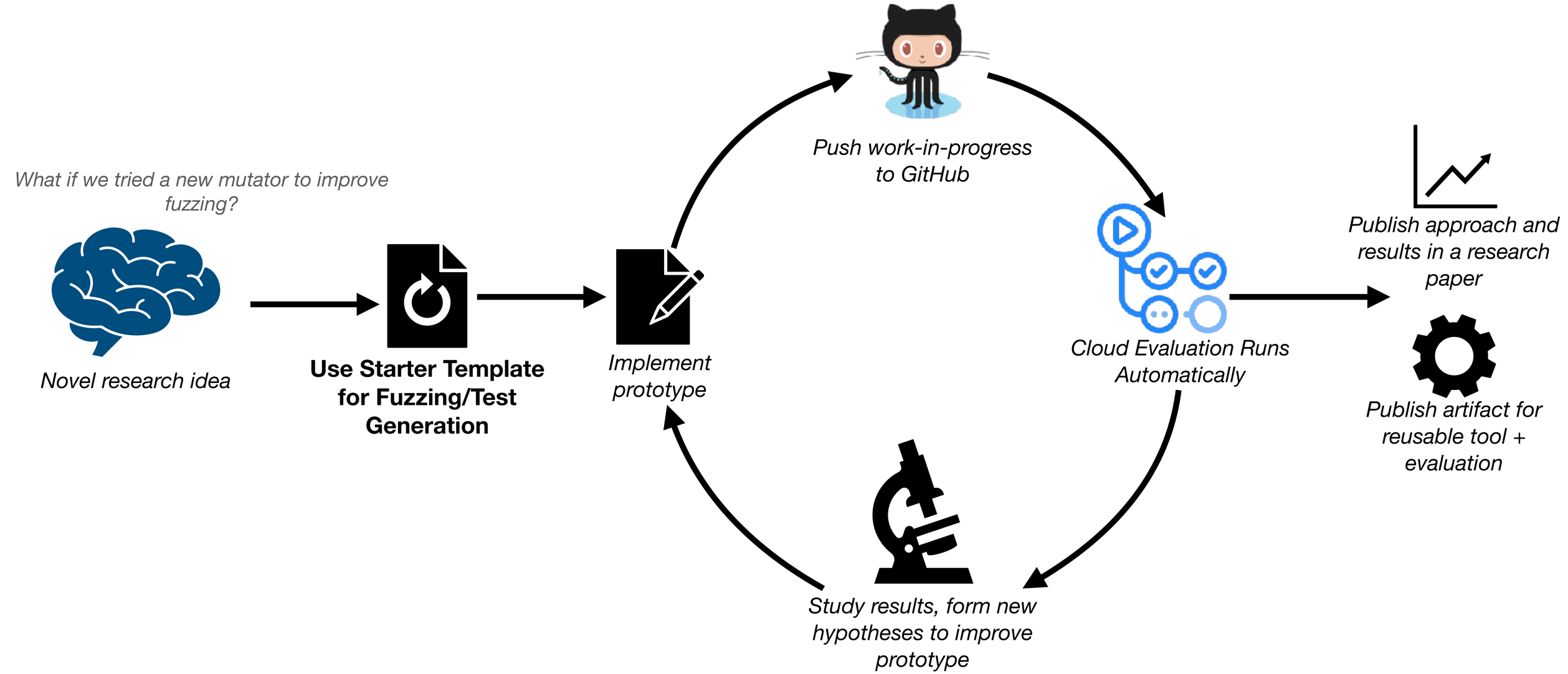
Manual Large-Scale Experiment Execution

State of the practice for building and evaluating software tools



Automating Large-Scale Experiment Execution

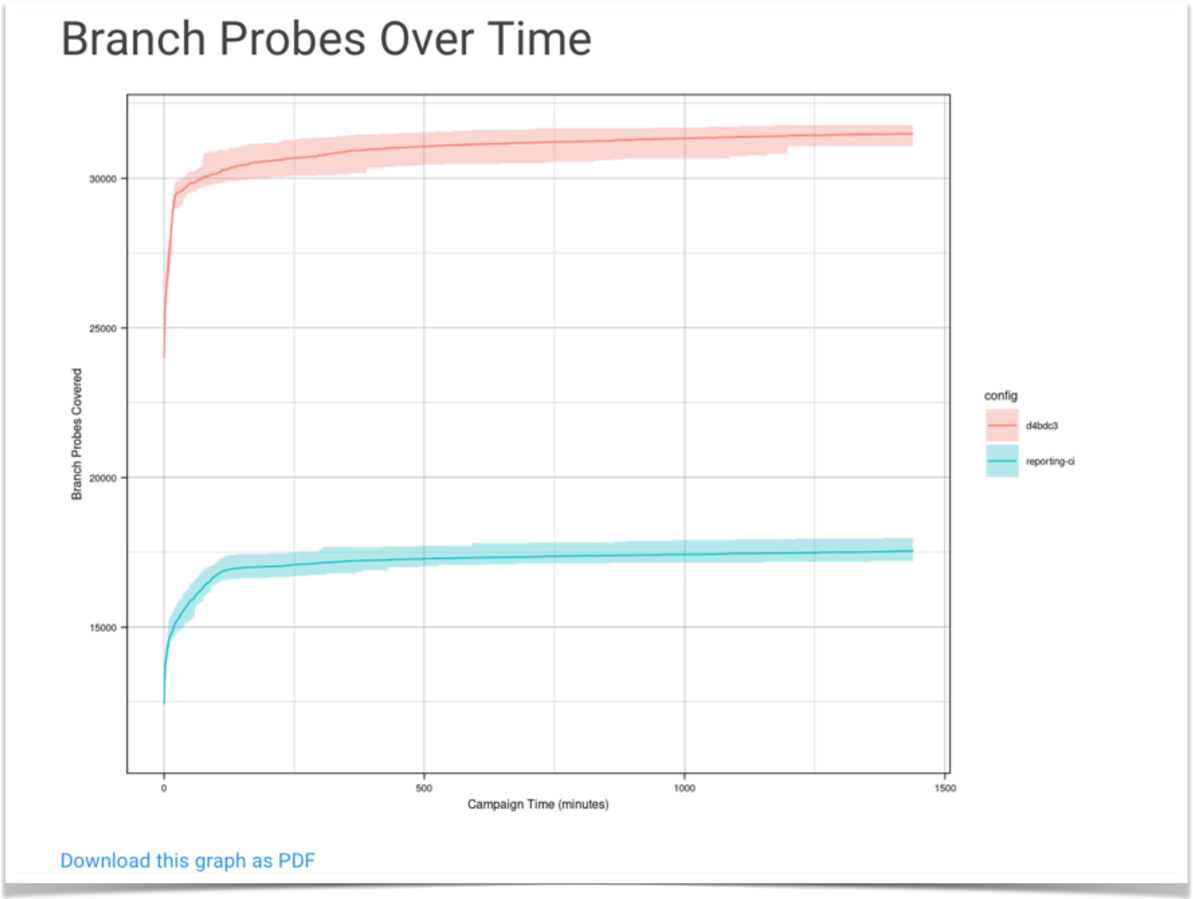
A new workflow for building and evaluating software tools



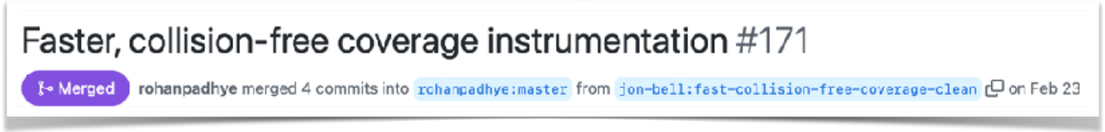
Prototype application: Evaluating Fuzzing in CI

Fully automated performance tests of our fuzzer, in CI, using on-prem HPC cluster

- Our ICSE ’22 “CONFETTI” project needs 100 x 4-core VMs for 24 hrs each for evaluation
- Now: fully automated in CI, 20x4-core VM evaluation runs for 10 mins each commit, full evaluation manually triggered
- Greatly simplifies testing and makes pull-requests far easier



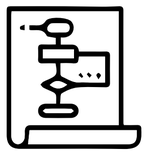
Sample graph generated by the CI workflow - automatically compares triggering commit to baseline



We used this workflow to find and fix a significant performance bug in the upstream tool

Our Research & Infrastructure Agenda

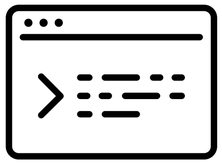
Infrastructure to write evaluations



Workflow templates and reusable actions

Community Research

Virtual & In-Person Training

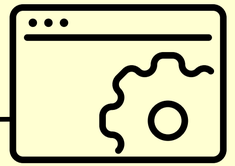


Documentation and resources

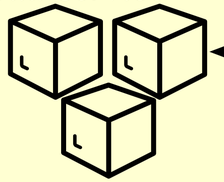
Infrastructure to run evaluations

Evaluation Reports

Hosted Classee.Cloud Service



Dashboard with plugins



Shared Utility Infrastructure

Evaluation scripts utilize

Trigger Evaluation

GitHub Actions

Compute Service

Researcher-provided cloud

CI Runner

CLASSEE.cloud launches; GitHub Actions configures and manages

Open Research Questions

- What are the best practices for running and monitoring large evaluation campaigns in CI?
- How to target specific projects/evaluation targets, while still providing general resources?
- How to design flexible interfaces for integrating tools and datasets?
- Where to store and how to manage artifacts and dependencies?
- How to encourage adoption of best practices for evaluation design?
- …What else haven't we thought of?