Main Objective

- to develop a dataset that allows to evaluate novel debugging techniques w.r.t. humans.
- How do developers explain the bug?
  - Which fault locations do experts point to?
  - Do developers agree on a single explanation?
- How do developers patch the bug?
  - Do patch and fault locations overlap?
  - How many human-generated patches are plausible but incorrect?

Demography

27 real bugs with simplified bug report and test cases
12 software professionals
11 developers + 1 researcher
07 plus years experience
06 countries
Russia, India, Slovenia, Spain, Canada, Ukraine
02 Open Source Software
GNU Grep, GNU Find, each with ~17KLOC
29 working days spent debugging these bugs
About 27 hours per developer

Evaluating debugging aids is difficult and time consuming.

While for most bugs all submitted patches are plausible (i.e., pass the failing test case), for most bugs 30% of patches are incorrect (i.e., fail the code review)!

DBGBENCH allows for effective evaluation of debugging aids.

DBGBENCH helps to evaluate:
- automated fault localisation,
- automated bug diagnosis, and
- automated repair techniques.

DBGBENCH helps to compare:
- how much faster a developer is diagnosing and repairing a bug using a novel debugging aid.