Reproducibility in Systems/Networking Research

Roger Chamberlain, Lance Fiondella, Geoff Kuenning, Ethan Miller, Dan Rubenstein,
Alex Spritson, Peter Steenkiste, Keith Winstein, Erez Zardok, many others...
Ken Calvert, Ann von Lehmen, Darleen Fisher,
Nick Feamster and Alex C. Snoeren

https://tinyurl.com/nsf19reproducibility

In search of reproducible systems research

- Repeatability
  - Need to archive datasets, store and version control code, implement regression testing, etc.
- Replicability
  - Supported by release source code, data sets, traces, etc.
- Reproducibility
  - Requires sufficient description of experiments to understand how results were obtained
  - In other fields, a prerequisite for publication and acceptance of results

- A key factor when considering technology transition
  - Need some sort of quantitative metric of reproducibility to facilitate contracting
- A large amount of effort to do right
  - Many of our results have a limited lifetime
  - Need to keep in mind for an effort composed of 3-4 grad students

Current Best Practices

- ACM has defined a vocabulary and badging system
  - Standardizes terminology and allows broad recognition of researcher effort
  - Some venues (e.g., SC, SOSP) have adopted systematic artifact evaluation
- Reviewers increasingly mandating reproducibility as a requirement
  - Multiple workloads/multiple tools to generate those workloads
  - Pick multiple points on continuum.
  - Support for appendices listing artifacts, including recipes
  - Releasing code / document the simulator
- Explicit focus on technology incubation
  - UC Santa Cruz encourages students to stick around for a few years to polish prototypes
  - Modeled on IUCRC program

Current Impediments

- Conference culture does not lend itself to reproducible results
  - Short timelines, race-to-publish, little time for revision
  - Double-blind reviewing frustrates explicit detail of experimental conditions
- Systems, hardware, software environments change
  - Hard enough to get the same results from the same code n years later
- Students graduate--that's a good thing!
  - Little institutional memory, few resources to support archival storage
  - PI's (increasingly) move around as well
- Releasing data is fraught with regulatory challenges
  - Privacy regulations, IRBs, lawyers, etc.
- Our community doesn't understand/apply/believe statistics
  - We don't "p-hack" because we don't even have confidence intervals! — Henning(?)

Concrete suggestions for the NSF

- Consider changes to proposal/reporting requirements
  - E.g., data management plan as on add-on may send wrong message
  - Unclear the extent to which annual/final reports are effective in collecting artifacts
  - Abortive? attempt at results dissemination plan
- Expand support for the work required to enable reproducibility
  - Providing code is one thing; providing useable code is another entirely
  - Research programmers/analysts are hard to fund in current model
  - Potential data-curation/code-hardening post docs? TTP-style track?
- The more bold the claim, the more important the validation
  - Perhaps “moon-shot” style research engenders greater interest in reproducing results
  - Are we doing science, or engineering? The proof is in the pudding for the latter.
  - Mandate reproducibility support from testbed/infrastructure proposals?

Some crazier thoughts

- Provide a time-limited “reproducibility” bounty
  - Offered to anyone who reproduces, e.g., the results from an SIGCOMM/SOSP best paper
  - Maybe split with the paper authors to incentivize co-operation
  - Maybe the bounty is higher if the results are invalidated?
- NSF could propose a “grand challenge” problem/program
  - Modeled after DARPA’s grand challenge programs
- Workshops and/or formal PhD education on importance of reproducibility
  - Already held a recent Dagstuhl seminar…
  - Stanford networking class a great example; how can we replicate that model?
- Make reproducibility a first-class metric that drives funding
  - A committee of visitors could evaluate success of various sub-disciplines, recommend changes in large-scale budget allocations