General theme: snapshots of environments are necessary for reproducibility and replicability

- Includes datasets and software mentioned above
- But also many other factors that can impact a research experiment

Challenge: Uncertainty about how long artifacts for reproducibility should be functional.

- Capturing software versions, environments, etc is necessary for reproducible artifacts, but how long should infrastructure providers ensure those artifacts are usable? 2 years? 3 years? 5 years? Likely depends on the project
- Reproducibility facilitated by creating a snapshot of full environment
- Android says when things will be deprecated. Are we supposed to make sure things work for a certain period of time?
- Something that aims to support reproducibility should have a longer lifetime
- HPC machines change every few years making some computational reproducibility not really possible
  - Testbeds can hold onto machines for "past the time they're interesting" so that they can enable reproducibility
  - Old media types like tapes still have the hardware hanging around

Challenge: Reproducibility differs project to project/field to field and we need standards

- Simulations show up differently on different architectures; capturing data about the experiment is important
- Some have limited amount of control over environments (e.g. you're sending stuff over the air and there's interference; or a dependency is no longer available)
- Importance of replicability: when talking about systems, they're going to have some randomness and so you can't actually replicate even if you wanted to (nondeterministic algorithms, users, mean we're going to get different results)
- Maybe authors should say what their expectations for reproducibility are—which effects should be "replicated" and to what degree
- Opportunity: As a user and provider, it's important to talk about the three Rs. There's a lack of standards around what sort of details we need to provide the

community. Wants providers to give standards for users to follow; possibly de facto standards rather than mandatory

  - AECs sometimes provide this kind of information
- Some work is novel and innovative enough that we don’t need to worry about reproducibility, right now
  - exploratory work may have less demand for reproducibility; when you want to generalize or specify an effect you might want reproducibility
  - Possibly true: Observation is useful, science should be replicable
- Different communities (e.g. recommender systems) have different standards for what is good/bad replicability (some have no standards)
  - Maybe need to focus on the high quality/rigorous way
  - Best practices aren’t used by default all the time; want publishers to say, “we can’t accept this unless you also include X”
    - Different fields have varying degrees of maturity
  - Part of it’s replication and part of it’s the ability to build on and comparing against the study correctly; want common ways of comparing (validated metrics)
  - Psychology was/is interested in preventing questionable research practices; perhaps agreement on QRPs look like in different research communities would be useful
- How do medical fields and physicists define reproducibility?
  - Preregistration: specify research and analysis methods in advance to improve rigor and ensure all results are shared
    - Show conflicts of interest here as well?
      - Who checks that you did this right? There are standards for this.
    - Is this applicable in CS fields? Yes.

Challenge: Getting the environment to run the experiment and getting all the details is almost always impossible

  - Have to capture hardware and environmental information
  - Papers don’t have enough information
  - Getting others’ code is really difficult; even if you have it you might not be able to run it for a variety of reasons, including skill/domain expertise. Some packages are really well known (e.g. numpy) but to others that’s completely new and they “can’t” learn it.
  - Architectures are changing

    - Want to try and make it easy so that when someone wants to use the environment you built that they can reproduce it easily
        - Capture/snapshot the experiment
    - Maybe these issues mean that rather than reproduce, you replicate
    - Want elements of experiments to be included in a manifest

Opportunity: Publication venues and conferences can lead the way in changing our publishing model to better support reproducibility

- Change can't come unless top publication venues prioritize reproducibility
    - Could do preregistration as a publishing model (but it's not super popular)
        - This might be difficult because it might be difficult to show your as better
        - Comes from "Open Science" principles
        - Full transparency from the start
        - Allows others to review
        - Keep you honest
        - E.g., Center for Open Science registered reports - peer review before results are known
    - COS
        - Provide infra to support this ^^^^
        - Work with journals to push these ideas forward
        - Open data
        - Open sharing
        - Open materials
        - Guarantees for uninteresting results (as long as pre register)
        - Any examples in engineering? CS probably looks more like engr than medical
        - Takes a different publishing model
- Conferences could/some do request source code
    - Currently encouraged, moving toward required
    - Industry won't share code and that means they don't want to go to the conference if you require it
    - Some data can't be shared because privacy, proprietary agreements
        - Supercomputing often happens in classified labs
        - Don't want separate venues for classified vs not; communities need to talk
    - Sometimes you're required unless you have a rationale

- LASER (Learning from Authoritative Security Experiment Results) (formerly LUSSR) Workshop
- Conferences are sometimes paper factories with too many papers; might want to change this model to be more about conferring (vs multiple paper tracks)
    - AI and ML are exemplars
    - Should there be fewer papers at these venues and more in journals?
    - Why should we get together when the presentations are recorded?
        - Inverted classroom model might make more sense: 1 min review, 10 min discussion (or can do group of papers); keeps value of socializing at the conference (New Security Paradigms might be a good example to follow)
        - By the time you're at the conference, the work you're presenting is a little out of date and the audience is already familiar with it
    - In addition, work may be public on Github and already shared within the community
    - Virtual conferences are expensive and difficult and frustrating; how can we keep the convenience but get rid of the expense?
    - General agreement (e.g. CRA letter) that we put too much emphasis on publications and you need to need to get into the conference for your university to even support you going

Opportunity: NSF/funders can encourage reuse of existing datasets and sharing of artifacts by asking for a sharing *and reuse* plan

- Asking the proposer what reproducibility means to them/what will it look like
- If you're using XYZ datasets, you have to evaluate using them
    - Or encourage existing NSF infrastructure generally, e.g. National Hazards Program says you need to use their tools (NSF does this at least sometimes)
        - Possibly better to do this in a progressive manner (let's identify 6-8 that are mature enough for NSF to promote), especially if things cost
        - Ask people to say why they wouldn't use them; can help encourage their use →especially relevant if a field is less mature/for whatever reason doesn't have the standards that you'd like (guidance from NSF on this would be helpful)
        - Possibly needs to be an appendix, especially because the

more ambitious you are the less room you have

    - Want panelists to pay attention to this too
  - Ask in proposals for an assessment of the research in the field, its weaknesses, and how the infrastructure will affect them -> the point of the infrastructure is to do something, not just be there
    - Include discussion of repro/artifacts
    - e.g. “Most work does a lousy job when we have lots of background noise, here’s how we’ll deal with that.”
  - Broaden data management plan to address this; more like a sharing and reuse plan
    - However, enforcement is a big problem!!
      - DARPA has a big stick: one dataset, one benchmark, top couple make it to the next round
      - People do not share their data/code even when they say they will
      - Can CIRC resources be used?

Opportunity: Find champions who can lead disciplines toward reproducibility

- Luminaries in this area?
  - Ron Boisvert, NIST
  - Jack Davidson, UVA
  - Victoria Stodden, USC
  - Brian Nosek, UVA
  - Cultivating champions can be a way to encourage change
    - Workshops like LASER might be good for that
- Run Workshops: Dagstuhl, Japan, others? NSF workshop?

Opportunity: Create a directory of infrastructure to help find tools for reproducibility and for their reuse in general

  - Virtual Organization (VO)
  - Well-trained AI

Debate: If you publish, does that mean you specifically want to share and not keep work private?

  - Not necessarily. Lots of interest in tech transfer/commercialization; mixture between publishing/patenting gets blurry and tricky

Miscellaneous/unsure where to put:

- IEEE BOG initiative on repro
- Saurabh Bagchi part of the committee
- ACM moving to open access
- If you can't repeat your own work, it's probably not replicable
- Can look at other science domains and how they did reproducibility
  - Cold fusion debunked, room temperature superconductor

Attendees:

David Balenson, USC-ISI (Organizer)
Hannah Cohoon, University of Utah (Scribe)
Fitz Elliott, COS
Rob Ricci, University of Utah
Joseph Konstan, University of Minnesota
Stephen Harrell, TACC
Raveen Wijewickrama, UT San Antonio
Joe Cavallaro, Rice
Talha Mehboob, U Mass Amherst
Shengli Fu, University of North Texas
Desh Raj, Johns Hopkins
Amy LaViers, RAD Lab
Jennifer Hoster, Georgia Tech
Harem Deep, Texas A&M (affiliation is correct, need to confirm name)
Saurabh Bagchi, Purdue