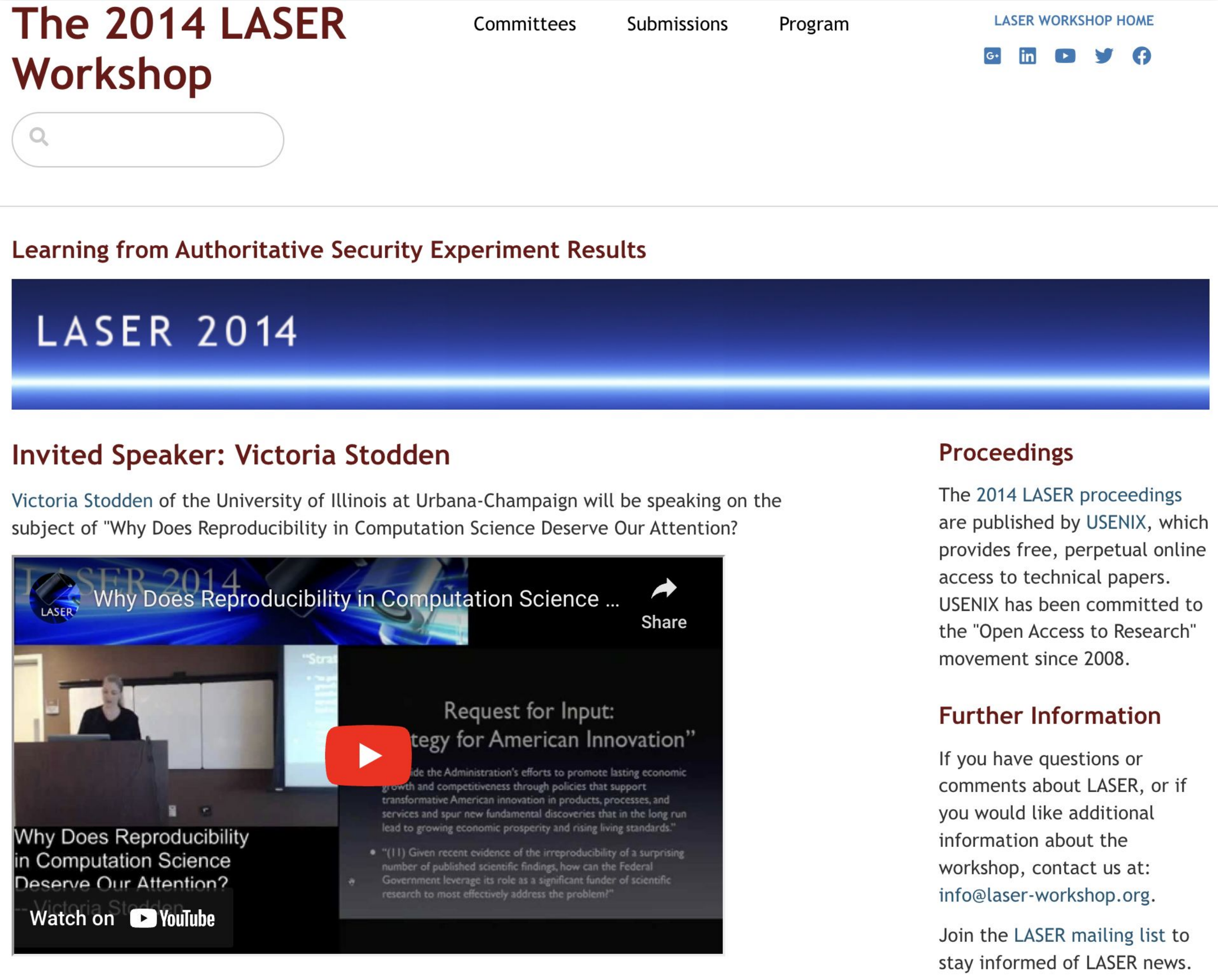
A Decade Later: Reproducibility & Reliability of Research Results

Victoria Stodden
Department of Industrial and Systems Engineering
University of Southern California

Keynote
Learning from Authoritative Security Experiment Results (LASER) workshop
December 5, 2023

USC



The 2014 LASER Workshop

Committees Submissions Program

LASER WORKSHOP HOME

Learning from Authoritative Security Experiment Results

LASER 2014

Invited Speaker: Victoria Stodden

Victoria Stodden of the University of Illinois at Urbana-Champaign will be speaking on the subject of "Why Does Reproducibility in Computation Science Deserve Our Attention?

LASER 2014
LASER
Why Does Reproducibility in Computation Science ...
Share
Request for Input:
"Strategy for American Innovation"
Why Does Reproducibility in Computation Science Deserve Our Attention?
Victoria Stodden
Watch on YouTube

Proceedings

The 2014 LASER proceedings are published by USENIX, which provides free, perpetual online access to technical papers. USENIX has been committed to the "Open Access to Research" movement since 2008.

Further Information

If you have questions or comments about LASER, or if you would like additional information about the workshop, contact us at: info@laser-workshop.org.

Join the LASER mailing list to stay informed of LASER news.

Agenda

1. **My Background**

2. **Reproducibility Redux**
   - NASEM report Reproducibility and Replication in Science
   - Frameworks for Policy
   - Trust and Public Access to Digital Research Objects

3. **(The Next 10 Years?) Reproscreener: Automating Model Checking**

1. My Background

Educational Experience

Ph.D. Sept 2006. Statistics, Stanford University. Advisor: David Donoho,

Committee Chair: Michael Saunders (Management Science and Engineering)

Committee: Michael Saunders, David Donoho, Jerry Friedman, Trevor Hastie, and Rob Tibshirani

M.L.S. Dec 2007. Stanford Law School

M.S. June 2000. Statistics, Stanford University

M.S. July 1996. Economics, University of British Columbia

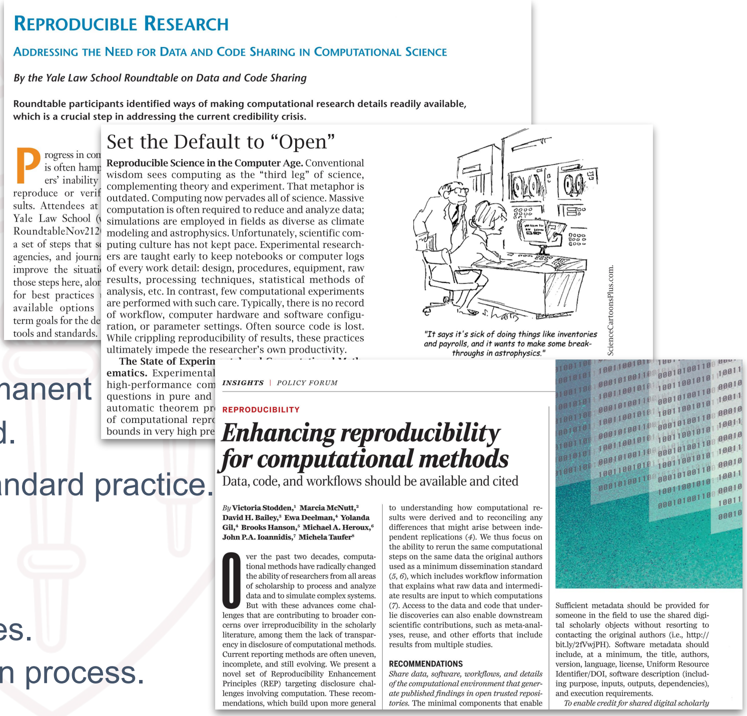
B.Soc.Sci. Dec 1994. Economics (magna cum laude), University of Ottawa

2. Reproducibility Redux

Reproducibility Standards Development

Community Efforts: AAAS 2016 Workshop on Code and Modeling Reproducibility recommended:

- **Share** data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.
- **Persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
- To enable credit for shared digital scholarly objects, **citation** should be standard practice.
- To facilitate reuse, adequately **document** digital scholarly artifacts.
- Use **Open Licensing** when publishing digital scholarly objects.
- Funding agencies should instigate new research programs and pilot studies.
- Journals should conduct a **reproducibility check** as part of the publication process.



REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

Roundtable participants identified ways of making computational research details readily available, which is a crucial step in addressing the current credibility crisis.

Progress in com
is often hamp
ers' inability
reproduce or verif
sults. Attendees at
Yale Law School (v
RoundtableNov212
a set of steps that s
agencies, and journa
improve the situati
those steps here, alon
for best practices
available options
term goals for the de
tools and standards.

Set the Default to "Open"

Reproducible Science in the Computer Age. Conventional wisdom sees computing as the "third leg" of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every work detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While crippling reproducibility of results, these practices ultimately impede the researcher's own productivity.

The State of Experim
ematics. Experimental
high-performance com
questions in pure and
automatic theorem pr
of computational repro
bounds in very high pre

"It says it's sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

ScienceCartoonsPlus.com

INSIGHTS | *POLICY FORUM*

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly

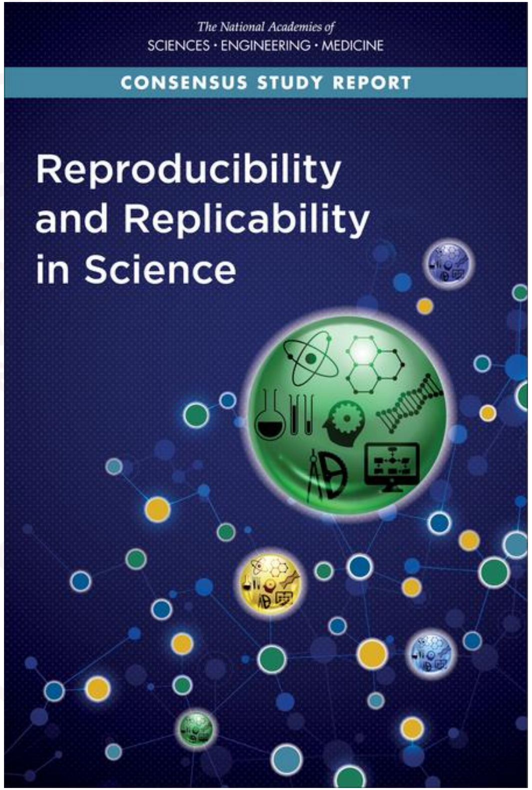
Stodden, McNutt, Bailey, Deelman, Gil, Hanson, Heroux, Ioannidis, Taufer (2016). Enhancing Reproducibility for Computational Methods. Science.

National Academies Consensus Report 2019

“Reproducibility and Replication in Science”

- 15 distinguished members (I was a committee member)
- Chair: Harvey Fineberg, President of Gordon and Betty Moore Foundation
- Stakeholder input: over 50 individuals representing a range of disciplines

⇨ Produced key definitions and several recommendations.



Report and white papers available at https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

Committee Charge

- Define reproducibility and replicability accounting for the diversity of fields in science and engineering.

- Examine state of contemporary science with regard to reproducibility and replication.

- Determine if lack of replication and reproducibility impacts the overall health of science and engineering as well as the public’s perception of these fields.

- Make recommendations for improving rigor and transparency in scientific and engineering research.

Reproducibility Definitions

- ***Reproducibility*** is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "**computational reproducibility**."

- ***Replicability*** is obtaining **consistent results across studies** aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

Recommendation 4-1(Transparency)

To help ensure the reproducibility of computational results, r**esearchers should convey clear, specific, and complete information about any computational methods and data products that support their published results** in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:

- **the input data** used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

- **a detailed description of the study methods (ideally in executable form)** together with its computational steps and associated parameters; and

- **information about the computational environment** where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

Recommendation 6-6 (Coordination)

Many stakeholders have a role to play in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.

- Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.
- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research. Societies have an important role in educating the public about the evolving nature of science and the tools and methods that are used.
- Researchers should collaborate with expert colleagues when their education and training are not adequate to meet the computational requirements of their research.
- In line with its priority for "harnessing the data revolution," the National Science Foundation (and other funders) should consider funding of activities to promote computational reproducibility.

Recommendation 6-7 (Publishers)

Journals and scientific societies requesting submissions for conferences should **disclose** their policies relevant to achieving reproducibility and replicability. The strength of the claims made in a journal article or conference submission should reflect the reproducibility and replicability standards to which an article is held, with stronger claims reserved for higher expected levels of reproducibility and replicability.

Journals and conference organizers are encouraged to:

- set and implement desired standards of reproducibility and replicability
- adopt policies to reduce the likelihood of non-replicability
- require as a review criterion that all research reports include a thoughtful discussion of the uncertainty in measurements and conclusions.

Recommendation 6-8 (Funding Initiatives)

Many considerations enter into decisions about what types of scientific studies to fund, including striking a balance between exploratory and confirmatory research. If private or public funders choose to invest in initiatives on reproducibility and replication, two areas may benefit from additional funding:

- *education and training initiatives* to ensure that researchers have the knowledge, skills, and tools needed to conduct research in ways that adhere to the highest scientific standards; that describe methods clearly, specifically, and completely; and that express accurately and appropriately the uncertainty involved in the research;

- *reviews of published work*, such as testing the reproducibility of published research, conducting rigorous replication studies, and publishing sound critical commentaries.

Recommendation 6-3 (Tools and Training)

Funding agencies and organizations should consider investing in research and development of **open-source, usable tools and infrastructure that support reproducibility** for a broad range of studies across different domains in a seamless fashion.

Concurrently, investments would be helpful in outreach to inform and **train researchers** on best practices and how to use these tools.

Recommendation 6-5 (Repositories)

In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should:

- Develop a set of **criteria for trusted open repositories** to be used by the scientific community for objects of the scholarly record.
- Seek to **harmonize with other funding agencies** the repository criteria and data-management plans for scholarly objects.
- Endorse or consider creating code and data repositories for **long-term archiving** and preservation of digital artifacts that support claims made in the scholarly record based on NSF-funded research. These archives could be based at the institutional level or be part of, and harmonized with, the NSF-funded Public Access Repository.
- Consider extending NSF’s current **data-management plan** to include other digital artifacts, such as software.
- Work with communities reliant on non-public data or code to develop **alternative mechanisms** for demonstrating reproducibility.

Recommendation 6-9 (Proposal Review)

Funders should require a thoughtful discussion in grant applications of how uncertainties will be evaluated, along with any relevant issues regarding replicability and computational reproducibility.

Funders should introduce review of reproducibility and replicability guidelines and activities into their merit-review criteria, as a low-cost way to enhance both.

Recommendation 6-10 (Funding Replication)

When funders, researchers, and other stakeholders are considering whether and where to direct resources for replication studies, they should consider the following criteria:

- The scientific results are important for individual decision-making or for policy decisions.
- The results have the potential to make a large contribution to basic scientific knowledge.
- The original result is particularly surprising, that is, it is unexpected in light of previous evidence.
- There is controversy about the topic.
- There was potential bias in the original investigation, due, for example, to the source of funding.
- There was a weakness or flaw in the design, methods, or analysis of the original study.
- The cost of a replication is offset by the potential value in reaffirming the original results.
- Future expensive and important studies will build on the original scientific results.

Recommendation 7-1 & 7-2 (Communication)

RECOMMENDATION 7-1: Scientists should take care to **avoid overstating** the implications of their research and also exercise caution in their review of press releases, especially when the results bear directly on matters of keen public interest and possible action.

RECOMMENDATION 7-2: Journalists should report on scientific results with as much **context and nuance** as the medium allows. In covering issues related to replicability and reproducibility, journalists should help their audiences understand the differences between non-reproducibility and non- replicability due to fraudulent conduct of science and instances in which the failure to reproduce or replicate may be due to evolving best practices in methods or inherent uncertainty in science. Particular care in reporting on scientific results is warranted when:

- the scientific system under study is complex and with limited control over alternative explanations or confounding influences;

- a result is particularly surprising or at odds with existing bodies of research;

- the study deals with an emerging area of science that is characterized by significant disagreement or contradictory results within the scientific community; and

- research involves potential conflicts of interest, such as work funded by advocacy groups, affected industry, or others with a stake in the outcomes.

Recommendation 7-3 (Context)

Anyone making personal or policy decisions based on scientific evidence should be wary of making a serious decision based on the results, no matter how promising, of a single study.

Similarly, no one should take a new, single contrary study as refutation of scientific conclusions supported by multiple lines of previous evidence.

Developing Frameworks for Policy

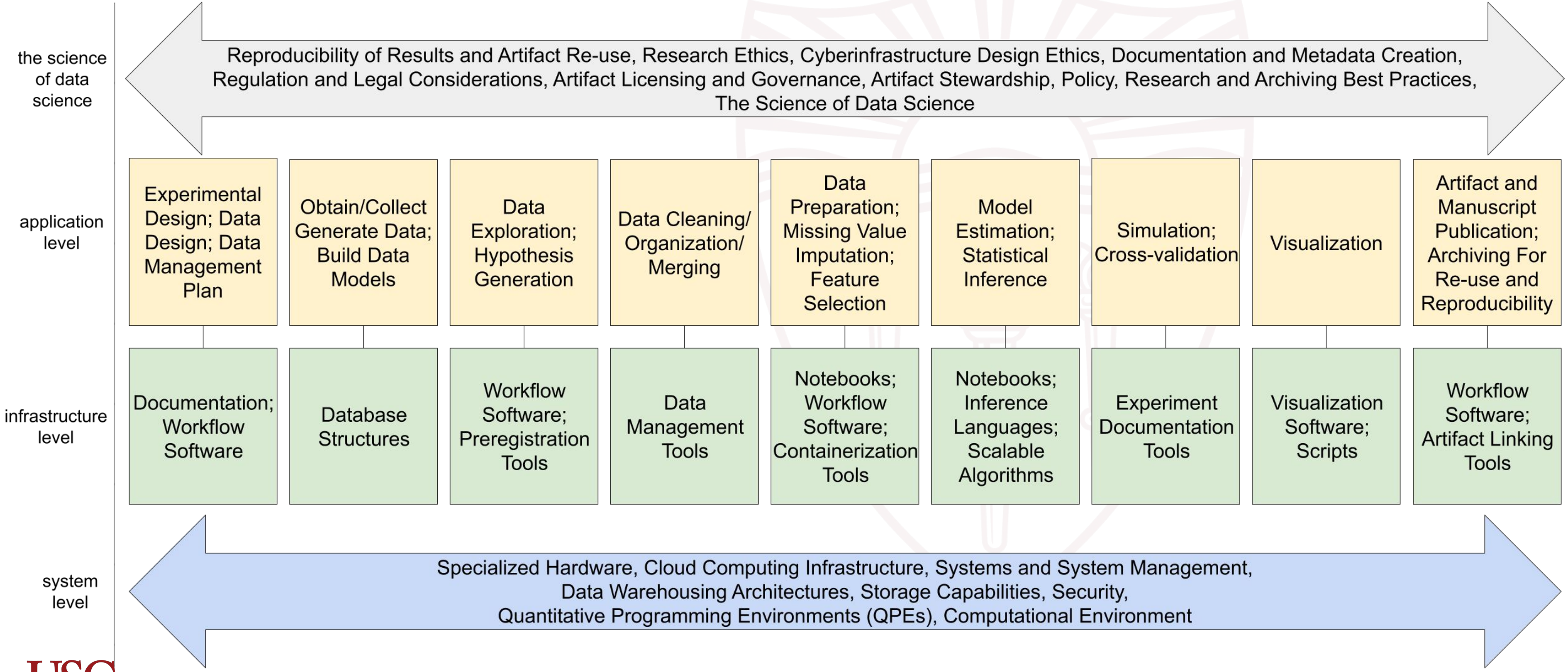
“Lifecycle of Data” is an abstraction from the Information Sciences

- Describes and relates actors in the ecosystem of data use and re-use.

What if we applied this idea to data-enabled science?

- **Clarify steps** in research projects: people/skills involved, tools and infrastructure, and reproducibility through the cycle.
- **Holistically guide implementations**: infrastructure, ethics, reproducibility and sources of uncertainty, curricula, training, and other programmatic initiatives.
- **Develop and reward contributing areas**.

A Proposal: Lifecycle of Data Science



the science of data science

Reproducibility of Results and Artifact Re-use, Research Ethics, Cyberinfrastructure Design Ethics, Documentation and Metadata Creation, Regulation and Legal Considerations, Artifact Licensing and Governance, Artifact Stewardship, Policy, Research and Archiving Best Practices, The Science of Data Science

application level	Experimental Design; Data Design; Data Management Plan	Obtain/Collect Generate Data; Build Data Models	Data Exploration; Hypothesis Generation	Data Cleaning/ Organization/ Merging	Data Preparation; Missing Value Imputation; Feature Selection	Model Estimation; Statistical Inference	Simulation; Cross-validation	Visualization	Artifact and Manuscript Publication; Archiving For Re-use and Reproducibility
infrastructure level	Documentation; Workflow Software	Database Structures	Workflow Software; Preregistration Tools	Data Management Tools	Notebooks; Workflow Software; Containerization Tools	Notebooks; Inference Languages; Scalable Algorithms	Experiment Documentation Tools	Visualization Software; Scripts	Workflow Software; Artifact Linking Tools

system level

Specialized Hardware, Cloud Computing Infrastructure, Systems and System Management, Data Warehousing Architectures, Storage Capabilities, Security, Quantitative Programming Environments (QPEs), Computational Environment

V. Stodden (2020). The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. CACM.

A Proposed Formalism: The "Tale"

What information do we need to reproduce and verify computational findings?

- Manuscript
  - source or reference
- Documentation
  - README, codebook, install instructions, user guide, etc.
  - License, copyright, permissions
- Code
  - Preprocessing, analysis, workflow
- Data
  - By copy, by reference, data access protocol
- Results
  - Output, figures, tables
- Environment
  - Hardware, OS, compilers, dependent software
  - Runtime, image, container
- Provenance
  - Computational, archival
- Metadata
  - Identifiers, related artifacts, Domain metadata
  - Badges
- Version

Chard et al. (2019) Implementing Computational Reproducibility in the Whole Tale Environment. P-RECS '19: Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems

Challenges Across the Community

- Relating data and software e.g. LLMs.
- Upskilling in the era of Data Science / Data Inference / Data Collection / Data Visualization / Data Policy / Data Ethics / Data CI / AI.
- Culture change: potentially enabling bad behaviors e.g. data and software capture, minimal value add, ignoring or quashing disruption.
- Cost/benefit/risk analysis.
- Public perception of science.
- Funding long term curation and archiving.

Challenge: IP and Transparency

Researchers generally don’t resolve IP issues regarding their research products.

- ⇨ Funding agency policy setting (in cooperation with institutions and other stakeholders).

Public access to research artifacts and scholarly information data, support of scholarly norms. “Giving back.”

- ⇨ “Reproducible Research Standard” (Stodden 2008)

Long Term Goals?

An **integrated computable scholarly record** that is queryable e.g.:

- Show a table of effect sizes and p-values in all vaccination/autism studies published after 1997;
- Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;
- List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;
- Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1; and
- Randomly re-assign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the year 2003 and list the trial name and histogram side by side.

M. Gavish, D. Donoho, and A. Onn. (2013) Dream applications of verifiable computational results. XRDS, 19, 3.

3. Reproscreener

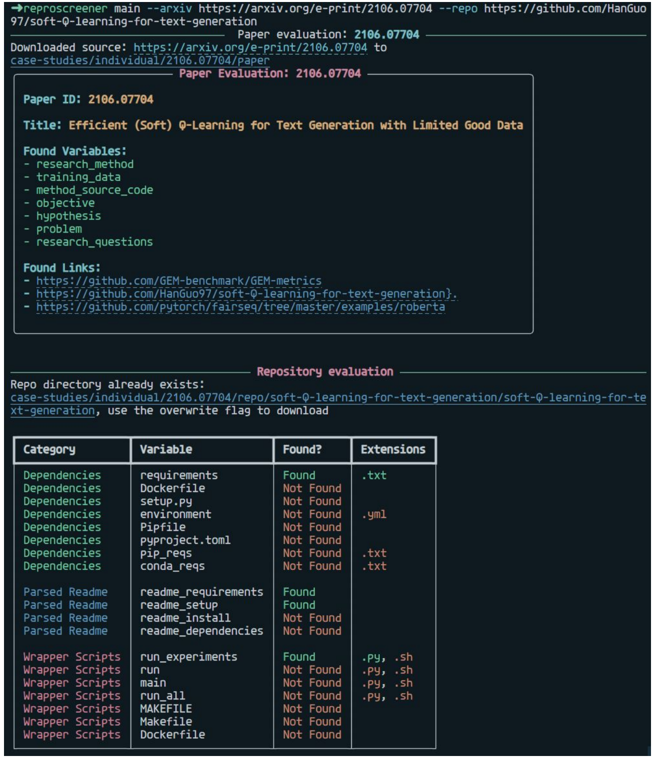
Automating Model Checking: Reproscreener (work in progress)

- Automate Machine Learning model checking *at the point of publication*, to provide guarantees on correctness, scalability, and transparency.

- Reproscreener software tool verifies criteria and provides feedback.

- Available at https://reproscreener.org and https://github.com/Machine-Learning-Pipelines/reproscreener/

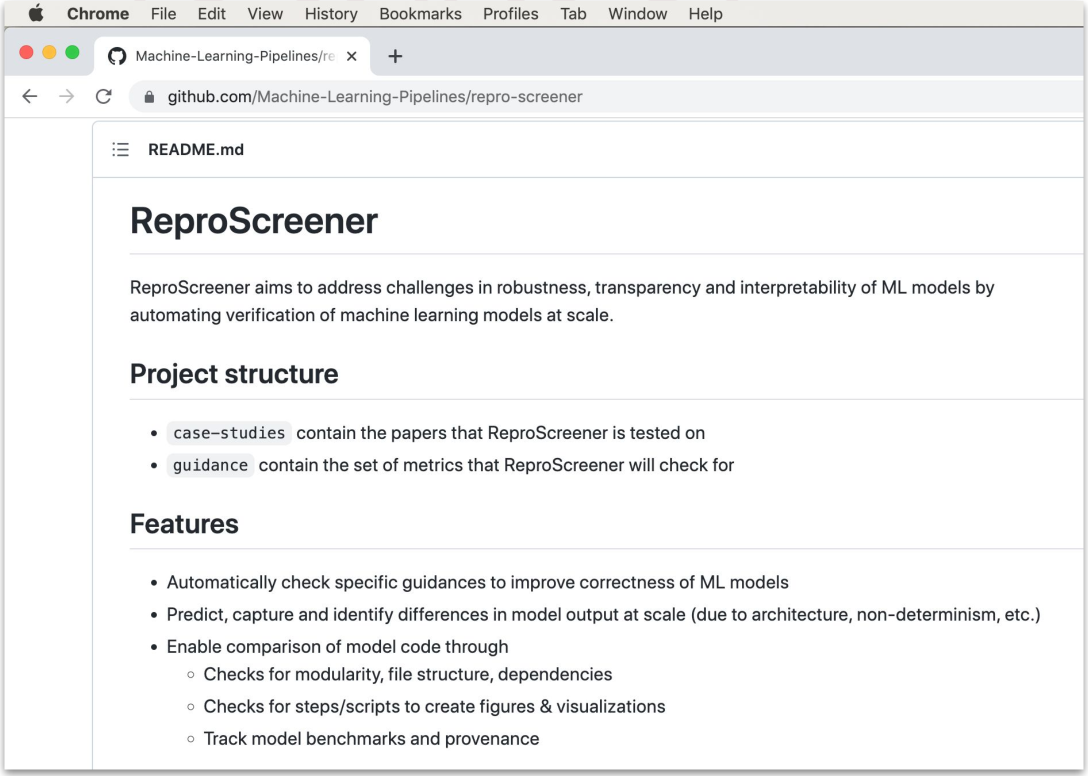
Reproscreener Open Source Development (work in progress)



```
➜reproscreener main --arxiv https://arxiv.org/e-print/2106.07704 --repo https://github.com/HanGuo
97/soft-Q-learning-for-text-generation
────────────────────────────────── Paper evaluation: 2106.07704 ──────────────────────────────────
Downloaded source: https://arxiv.org/e-print/2106.07704 to
case-studies/individual/2106.07704/paper
╭────────────────────────── Paper Evaluation: 2106.07704 ──────────────────────────╮
│                                                                                  │
│ Paper ID: 2106.07704                                                             │
│                                                                                  │
│ Title: Efficient (Soft) Q-Learning for Text Generation with Limited Good Data    │
│                                                                                  │
│ Found Variables:                                                                 │
│ - research_method                                                                │
│ - training_data                                                                  │
│ - method_source_code                                                             │
│ - objective                                                                      │
│ - hypothesis                                                                     │
│ - problem                                                                        │
│ - research_questions                                                             │
│                                                                                  │
│ Found Links:                                                                     │
│ - https://github.com/GEM-benchmark/GEM-metrics                                   │
│ - https://github.com/HanGuo97/soft-Q-learning-for-text-generation}.              │
│ - https://github.com/pytorch/fairseq/tree/master/examples/roberta                │
│                                                                                  │
╰──────────────────────────────────────────────────────────────────────────────────╯


──────────────────────────────────── Repository evaluation ────────────────────────────────────
Repo directory already exists:
case-studies/individual/2106.07704/repo/soft-Q-learning-for-text-generation/soft-Q-learning-for-te
xt-generation, use the overwrite flag to download

┏━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Category        ┃ Variable             ┃ Found?    ┃ Extensions ┃
┡━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ Dependencies    │ requirements         │ Found     │ .txt       │
│ Dependencies    │ Dockerfile           │ Not Found │            │
│ Dependencies    │ setup.py             │ Not Found │            │
│ Dependencies    │ environment          │ Not Found │ .yml       │
│ Dependencies    │ Pipfile              │ Not Found │            │
│ Dependencies    │ pyproject.toml       │ Not Found │            │
│ Dependencies    │ pip_reqs             │ Not Found │ .txt       │
│ Dependencies    │ conda_reqs           │ Not Found │ .txt       │
│                 │                      │           │            │
│ Parsed Readme   │ readme_requirements  │ Found     │            │
│ Parsed Readme   │ readme_setup         │ Found     │            │
│ Parsed Readme   │ readme_install       │ Not Found │            │
│ Parsed Readme   │ readme_dependencies  │ Not Found │            │
│                 │                      │           │            │
│ Wrapper Scripts │ run_experiments      │ Found     │ .py, .sh   │
│ Wrapper Scripts │ run                  │ Not Found │ .py, .sh   │
│ Wrapper Scripts │ main                 │ Not Found │ .py, .sh   │
│ Wrapper Scripts │ run_all              │ Not Found │ .py, .sh   │
│ Wrapper Scripts │ MAKEFILE             │ Not Found │            │
│ Wrapper Scripts │ Makefile             │ Not Found │            │
│ Wrapper Scripts │ Dockerfile           │ Not Found │            │
└─────────────────┴──────────────────────┴───────────┴────────────┘
```


Chrome
File
Edit
View
History
Bookmarks
Profiles
Tab
Window
Help
Machine-Learning-Pipelines/re
github.com/Machine-Learning-Pipelines/repro-screener
README.md

ReproScreener

ReproScreener aims to address challenges in robustness, transparency and interpretability of ML models by automating verification of machine learning models at scale.

Project structure

- `case-studies` contain the papers that ReproScreener is tested on
- `guidance` contain the set of metrics that ReproScreener will check for

Features

- Automatically check specific guidances to improve correctness of ML models
- Predict, capture and identify differences in model output at scale (due to architecture, non-determinism, etc.)
- Enable comparison of model code through
  - Checks for modularity, file structure, dependencies
  - Checks for steps/scripts to create figures & visualizations
  - Track model benchmarks and provenance

Evaluation Criteria used by Reproscreener

1. Machine Learning model criteria for publication based on Gunderson (2018).

2. Code/repo criteria from Krafczyk et al. (2020).

Curated a labelled testbed of arXiv publications: 50 most recent arXiv preprint submissions in stat.ML and CS.GL from October 25 2022.

Reproscreener Testbed Performance

Metric	Proportion Correct (n=50)
Code available	0.82
Hypothesis stated	0.60
Experimental setup	0.54
Dataset available	0.48
Problem stated	0.36
Predicted result	0.30
Research method	0.28
Objective/Goal	0.28
Research question	0.16

Reproscreener Testbed Performance on Code Repositories

Metric	Proportion Correct (n=22)
Readme has dependencies info	0.45
Readme has setup instructions	0.45
Readme has requirements info	0.41
Readme has install instructions	0.41
Wrapper scripts	0.36
Dependency tracking files	0.32

Reproscreener Goals

- Automatically check specific guidances to improve correctness of ML models to predict error bounds, capture and identifies difference in model output at scale (due to architecture, non-determinism, etc.)
- Enable comparison of model code through:
  - Checking for modularity, file structure, dependencies.
  - Checking for steps/scripts to create figures & visualizations.
  - Tracking model benchmarks and provenance.
- Real world case studies to demonstrate ReproScreener’s functionality
- Boundedness guarantees regarding correctness of reproduced results compared to original ML pipeline.

Thank you!

Joint work with **Adhithya Bhaskar**
Ph.D. student
Department of Industrial and Systems Engineering
University of Southern California

USC This material is based upon work supported by the REAL@USC-META Center and National Science Foundation Grant No 2138776