

When are Data Science Results Reproducible?

Victoria Stodden
Associate Professor

Department of Industrial & Systems Engineering
University of Southern California

Stats 285 Guest Lecture
May 10, 2021

Agenda

- 1. Setting the Stage: Reproducibility & Reliability Examples**
 - Boeing; IEEE; National Academies of Science, Engineering, and Medicine report
- 2. A Tour of Three Examples of Recent Work**
 - Reproducible Data Science with the Whole Tale project
 - Improving Outcomes in Machine Learning Tournaments
 - Reproducibility Standards Development
- 3. Future Research Directions (if time)**
 - A “Lifecycle of Data Science”
 - A “Computable Scholarly Record”

1. Research Reproducibility & Reliability Examples

Reproducibility Example 1: Boeing

The NASEM Committee on “Reproducibility and Replication in Science”
hosted a panel entitled **Reproducibility in Industry and Industrial**
Engineering on April 18, 2018.

Bill Lyons presented, the Director for Global Research and Development Strategy on the Global Technology Organization of Boeing’s Advanced Centralized Research and Development Team



Disruption Expands the Need for Reproducibility

Lyons: The ability to replicate ideas and capabilities across the company is what got Boeing to be a 100 year old company “One Boeing”

Aerospace industry is undergoing disruption:

- Digitization; AI; Autonomy; Additive Manufacturing, Electrification...
- Data integrity are critical. “Our customers’ lives depend on it”
- Results of a system, e.g. based on Machine Learning, can be nondeterministic.

Boeing: \$4 billion of \$1.9 trillion in global R&D. They know they don’t have all the answers.

Boeing Leverages Reproducibility

Employs a global model for replication: standards setting and sharing results to validate results in consortia (12 R&D centers) and beyond.

Results may come from a partner in Australia with new materials developments and a lab in St. Louis does high throughput combinatorial analysis of materials to rapidly check the results.

Knowledge management and information sharing to accelerate the pace of change in their industry.

A Replication Award honors teams that “applied existing capability in new ways throughout Boeing, enabling business process or technology improvements.”

Reproducibility Example 2: Biosciences

In 2012, in a watershed publication AmGen claimed its scientists could reproduce only 6 of 53 landmark publications in preclinical life sciences

This lead to journal policy changes and funding agency initiatives, e.g.:

The screenshot shows the homepage of the journal 'nature'. At the top, there's a navigation bar with links for 'Announcement: Reducing our irreproducibility', 'nature.com', 'Sitemap', 'Login', and 'Register'. Below this, the main 'nature' logo is displayed with the subtitle 'International weekly journal of science'. A search bar is present. The main content area features a large heading 'Announcement: Reducing our irreproducibility' followed by the date '24 April 2013'. At the bottom left, there are links for 'PDF' and 'Rights & Permissions'.

The screenshot shows the homepage of 'Science', described as 'The World's Leading Journal of Original Scientific Research, Global News, and Commentary'. It features a red header with the AAAS logo and links for 'AAAS.ORG', 'FEEDBACK', 'HELP', 'LIBRARIANS', 'All Science Journals', 'GUEST ALERTS', and 'ACCESS REGISTRATION'. Below the header, the main content area shows an article titled 'Science' with the subtitle 'The World's Leading Journal of Original Scientific Research, Global News, and Commentary'. The article title is 'Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in Science.' The author is listed as 'Marcia McNutt'.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low¹. Sadly, clinical

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and

investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models² make it difficult for even ▶

29 MARCH 2012 | VOL 483 | NATURE | 531

The screenshot shows the NIH logo and the text 'National Institutes of Health' with the tagline 'Turning Discovery Into Health'. There are links for 'Search NIH' and 'NIH Em'. Below this, there are several menu options: 'Health Information', 'Grants & Funding', 'News & Events', 'Research & Training', and 'In'. The main content area features a section titled 'RIGOR AND REPRODUCIBILITY' with a sub-section 'Rigor and Reproducibility'.



Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of

Reproducibility Example 3: IEEE

IEEE steps to reproducibility and computational transparency:



Report on the First IEEE Workshop on The Future of Research Curation and Research Reproducibility

Download the final workshop report (PDF, 2 MB)

NSF logo

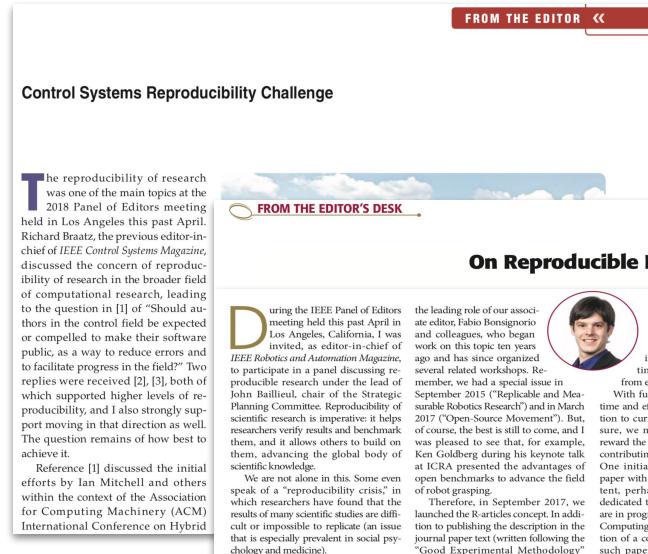
IEEE logo

Marriott Marquis, Washington, DC, USA
5–6 November 2016

National Science Foundation Award # 1641014

2016 workshop

<http://www.ieee.org/researchreproducibility>



Control Systems Reproducibility Challenge

FROM THE EDITOR <>

Control Systems Reproducibility Challenge

The reproducibility of research was one of the main topics at the 2018 Panel of Editors meeting held in Los Angeles this past April. Richard Braatz, the previous editor-in-chief of *IEEE Control Systems Magazine*, discussed the concern of reproducibility of research in the broader field of computational research, leading to the question in [1] of "Should authors in the control field be expected or compelled to make their software public, as a way to reduce errors and to facilitate progress in the field?" Two replies were received [2], [3], both of which supported higher levels of reproducibility, and I also strongly support moving in that direction as well. The question remains of how best to achieve it.

Reference [1] discussed the initial efforts by Ian Mitchell and others within the context of the Association for Computing Machinery (ACM) International Conference on Hybrid

On Reproducible Research

By Bram Vanderborght

During the IEEE Panel of Editors meeting held this past April in Los Angeles, California, I was asked to chair a panel on reproducibility of research in the field of *IEEE Robotics and Automation Magazine*. I participated in a panel discussing reproducible research under the lead of John Baileul, chair of the Strategic Planning Committee. Reproducibility of scientific research is imperative; it helps researchers verify results and build them, and it allows others to build on them, advancing the global body of scientific knowledge.

We are not alone in this. Some even speak of a "reproducibility crisis," in which researchers have found that the results of many scientific studies are difficult to replicate (see for example, Ken Goldberg's excellent keynote talk at ICRA presented the advantages of open benchmarks to advance the field of robot grasping).

Therefore, in September 2017, we launched the Rafticles project. In addition to the Rafticles being used for publishing paper test (written following the "Good Experimental Methodology" guidelines), the required data sets, the leading role of our associate editor, Fabio Bonsignori and colleagues, who began this work a few years ago and has since organized several related workshops. Remember, we had a special issue in September 2015 ("Replicable and Measurable Research") and in March 2017 ("Open-Source Movement"). But, of course, the best is still to come, and I was pleased to see that, for example, Ken Goldberg's excellent keynote talk at ICRA presented the advantages of open benchmarks to advance the field of robot grasping.

With full awareness that it requires time and effort that may be in opposition to current publish-or-perish pressure, we need to increase efforts to encourage authors to make their code available and to contribute to reproducibility. One initiative suggests highlighting papers with reproducible research content, perhaps even granting awards dedicated to such papers. Discussing in progress with the Association for Computing Machinery, we are looking at a common badge system for such papers. A culture shift will be needed. Moreover, along the lines of the



SPOTLIGHT ON TRANSACTIONS

EDITOR RON VETTER
University of North Carolina Wilmington; vetter@uncw.edu

The Reproducibility Initiative

Manish Parashar, Rutgers University

This installment of Computer's series highlighting the work published in IEEE Computer Society journals comes from IEEE Transactions on Parallel and Distributed Systems.

Reproducibility is a foundation of solid scientific and technical research. The ability to repeat research is key to confirming the validity of a publication. Code Ocean sends the EIC a review copy of the compute capsule, which is passed on to the assigned reproducibility associate editor

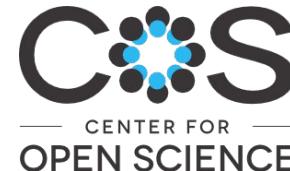
to upload the code to Code Ocean, which generates a "compute capsule" that contains the code, data, results, and computation environment specifications. Code Ocean sends

Editorial Policies and Badging
Pilot Partnerships: Code Ocean

Reproducibility Example 4: Social Psychology

In 2012 an email by Daniel Kahneman was published in Nature revealing reproducibility concerns of “priming” studies in social psychology. A constellation of questions had arisen regarding such studies, and several highly visible cases of fraud

Since then several initiatives in psychology have arisen to take on these challenges



A screenshot of a news article from the journal 'nature'. The header 'nature International weekly journal of science' is at the top. Below it are navigation links: Home, News & Comment, Research, Careers & Jobs, Current Issue, News & Comment, News, 2019, May, Article. The main title is 'Nobel laureate challenges psychologists to clean up their act'. Sub-headlines include 'Social-priming research needs “daisy chain” of replication.' and 'Ed Yong'. The date is 03 October 2012. There is a 'Rights & Permissions' section. To the right is a portrait photo of Daniel Kahneman, a man with glasses and a light blue shirt, with his arms crossed. Below the photo is the caption 'Kahneman, a psychologist at Princeton University in New Jersey, addressed his open e-mail to researchers who work on social priming, the study of how subtle cues can unconsciously' and the name 'Jon Roemer'.

Reproducibility Example 5: National Academies

In 2019 the “Reproducibility and Replication in Science” committee published consensus report (I was a committee member).

Produced key definitions and several recommendations.

- *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.

Report Recommendation Highlights

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers 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.

RECOMMENDATION 6-3: Funding agencies and organizations should consider investing in research and development of ***open-source, usable tools and infrastructure*** that support reproducibility.

RECOMMENDATION 6-9: 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.

2. Three Examples of Recent Work

1. Data Science in the Whole Tale Project

- Building an **open platform for computational reproducibility**
 - Create and publish **executable research objects ("Tales"**)
- Simplify process of creating & verifying reproducible computational artifacts for scientific discovery

Easy-to-access
cloud-based computing
environments



Transparent access to
research **data**



Export and **publish**
executable research
objects



Use case: Ren et al. (2018)

- ML experiments in materials science
- Published in *Science Advances*
- Code in Github
- Data published to Materials Data Facility

The image shows two side-by-side screenshots. On the left is a screenshot of the Science Advances website. At the top, it says "ScienceAdvances" with links for "Contents", "News", "Careers", and "Journals". Below that is a "SHARE" section with social media icons for Facebook, Twitter, LinkedIn, and GitHub. It also includes a "RESEARCH ARTICLE" and "RESEARCH METHODS" link. The main content is an article titled "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments" by Fang Ren, Logan Ward, Travis Williams, Kevin J. Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. It includes a DOI: 10.1126/sciadv.aat1126. On the right is a screenshot of a GitHub repository page for "fang-ren / Discover_MG_CoVZr". The repository has 162 commits and 1 branch. The codebase includes files like README.md, requirements.txt, and to-do.list. The repository is associated with the Materials Data Facility, CHIMeD, and NIST. The README.md file contains a brief description of the project: "Discovering Metallic Glasses through Machine Learning". The GitHub interface shows various commit details such as "Added some more details to README", "Added CoVZr", "Convert some", "Updated some", "Merge branch", "Added Mag", "Improvement", "Added some", "Uploaded some", "Added feature", and "Updated some". The GitHub interface also shows a "Pull requests" tab with 6 open pull requests, a "Issues" tab, and a "Marketplace" tab.

ScienceAdvances

Contents News Careers Journals

SHARE RESEARCH ARTICLE RESEARCH METHODS

Fang Ren^{1,*}, Logan Ward^{2,3}, Travis Williams¹, Kevin J. Laws⁵, Christopher Wolverton¹, Jason Hattrick-Simpers¹ and Apurva Mehta¹

* See all authors and affiliations

Science Adv. Vol. 4, no. 4, DOI: 10.1126/sciadv.aat1126

Search or jump to... Pull requests Issues Marketplace Explore

Pull requests Issues Marketplace Explore

Article

No description, website, or topics provided.

162 commits 1 branch 0 packages 0 releases All 3 contributors

Branch: master New pull request

WardLT Added some more details to README

idea Added CoVZr

figures Convert some

machine-learning Updated some

scripts Merge branch

gitmodules Added Mag

Dockerfile Improvement

README.md Added some

Simple_key.pdf Uploaded some

requirements.txt Added feature

to-do.list Updated some

README.md

Discovering Metallic Glasses through Machine Learning

Materials Data Facility CHIMeD NIST

Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-Throughput Experiments

Fang Ren, Ward, Logan; Williams, Travis; Laws, Kevin J.; Wolverton, Christopher; Hattrick-Simpers, Jason; Mehta, Apurva

Get it with Forge

From mdf_Forge Jupyter Forge
mdf = Forge()
mdf.match_source_names("pub_99_feng_accelerated")
mdf.search()

Forge installation

pip install mdf_forge

Forge documentation

Organizations

CHIMeD DOI

10.1126/sciadv.aat1126

See on Datacite

Year

2018

Source Name

pub_99_feng_accelerated

Tags

metallic-glasses machine-learning
high-throughput experiment
metals-and-alloys

Contact

Logan Ward

Get the Data

Globus

How can we publish the code and data to support computational reproducibility and reuse/exploration?

- Reproducibility implemented in Whole Tale

Elements of a "Tale"

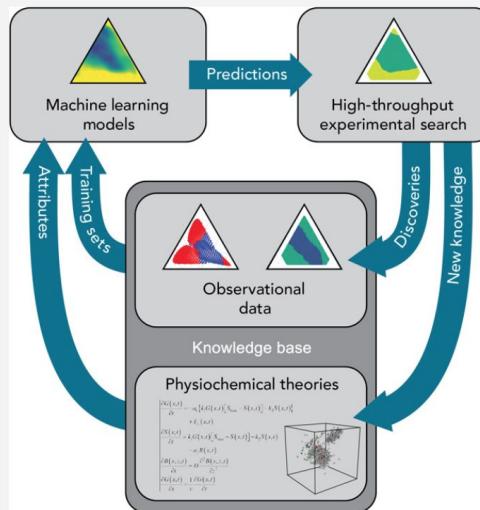
What information do we need to reproduce and verify computational findings?

- 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

[◀ Back](#)[▶ Launch Tale](#)

Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments

By Logan Ward



Access to Underlying Artifacts

The screenshot shows the WholeTale dashboard interface. At the top, there is a navigation bar with links for "WHOLE TALE DASHBOARD", "BROWSE", and "MANAGE". On the right side of the top bar, there is a user profile for "Victoria Stodden" with icons for help, settings, and logout.

In the main content area, there is a header for a "Predicting the Properties of Inorga..." project by "Logan Ward". Below this, there are three tabs: "Interact", "Files" (which is selected), and "Metadata".

The "Tale Workspace" section displays a table of files and folders:

Name	Size	Last Modified
datasets	27 MB	10 months ago
magpie	0	10 months ago
modeling-metallic-glasses	48.3 MB	10 months ago
predicting-band-gap-energies	5.87 MB	10 months ago
docker.bat	186 B	10 months ago
docker.bs	199 B	10 months ago
README.md	2.06 KB	10 months ago
run-all.bs	352 B	10 months ago

Packaging for sharing, dissemination, archiving

- Research Object
 - Beyond PDFs and datasets -- include code, workflows
 - Distributed elements
- Interoperability between systems
 - Archives/repositories
 - Active compute platforms
- BagIt serialized "Research Object" bundle
 - Zip archive + metadata + JSON-LD
 - <https://github.com/ResearchObject/bagit-ro> (=> ro-crate)



researchobject.org

Whole Tale as a Research Environment

By enabling computational transparency, Whole Tale:

- Improves/accelerates discovery e.g. Materials Science compound discovery.
- Facilitates standards development for scholarly object dissemination and evaluation.
- Testbed for understanding stakeholder/community needs to enable improved policy and decision making.
- “Meta science” orchestrations across “Tales” permits meta-science research.
- Creates an environment to study social incentives and pain points.

Winners in ML Tournaments

- *Leaderboard style problem solving structures* are frequently used in ML driven discovery where the “winner” has the lowest error rates on test data.
e.g. Kaggle.com, DrivenData.org, OpenML.org, Netflix Prize..
- A high variance across approaches is generally observed.
e.g. In one challenge, effect sizes varied from 0.89 to 2.93 in odds ratio units with 72% of the analyses using unique feature combinations.
- **Problem:** Given a pre-determined performance metric, there is generally little or no information on why an algorithm performed the best.
- **Proposed Solution:** A structured delivery of the ML pipeline in leaderboard style competitions (Abstraction for Machine learning (AIM)).

AML/ALL Data Example

- A gene expression dataset with each observation one of two cancers, acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) (Golub '99).
- Let $X = (x_{ij})$ be the dataset of genetic predictor variables where x_{ij} is the expression of gene j in sample i .
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is the gene expression profile for sample i .
- y_i is the response or class label, $i = \{1, 2\}$.
- Let \mathcal{X} be the space of all gene expression profiles.
- Let $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{n_{\mathcal{L}}}, y_{n_{\mathcal{L}}})\}$ be the learning set, $\mathcal{T} = \{(\mathbf{x}_{n_{\mathcal{L}}+1}), \dots, (\mathbf{x}_n)\}$ the test set, and $\mathcal{C} = \mathcal{C}(\mathbf{x}, \mathcal{L})$ be our classifier.

Stating the Classification Problem

Given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{n_{\mathcal{L}}}, y_{n_{\mathcal{L}}})\}$ where the \mathbf{x}_i 's are independent p -dimensional gene expression samples, the y_i 's the class labels, and given a test set $\mathcal{T} = \{(\mathbf{x}_{n_{\mathcal{L}}+1}), \dots, (\mathbf{x}_n)\}$,

find a classification function $\mathcal{C} = \mathcal{C}(\cdot, \mathcal{L})$ that maximizes classification accuracy on \mathcal{T} .

We found 30 attempts at this classification problem in the literature. Which gave us the best accuracy?

Results: Exposing the ML Pipeline

Direct comparison of *reported* classifier performance was impossible due to the use of different preprocessing and feature selection steps.

We attempted to reproduce the results in 5 papers, controlling for data preprocessing and feature selection.

We thus revise our classification problem as follows:

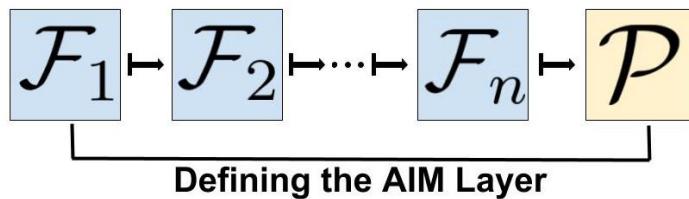
Find a classification function $\mathcal{C} = \mathcal{C}(\cdot, \tilde{L})$ that maximizes classification accuracy on $\tilde{\mathcal{T}}$, where $\mathcal{F}(Z) = \tilde{Z}$ is a function that carries out preprocessing and feature selection steps on input data Z .

Baseline Comparisons (5 articles, n=72 obs)

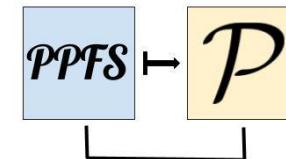
Classifier(Paper)	Preprocessing/Feature Selection Method						Average
	1	3	6a	6b	9	29	
WeightedVote(1)	.91	.94	.97	.97	.89	.74	.90
NN(3)	.97	.94	.91	.94	.97	.97	.95
Linear SVM(3)	.97	.97	.94	.97	.97	.77	.93
Quadratic SVM(3)	.97	.88	.97	.97	.97	.91	.95
Adaboost(3)	.91	.91	.97	.97	.91	.91	.93
Logit(6)	.97	.97	.97	.97	.97	.88	.96
QDA(6)	.94	.91	.94	.97	.97	.85	.93
NN(9)	.97	.91	.85	.97	.94	.94	.93
Decision Trees(9)	.91	.91	.97	.97	.91	.77	.90
Bagging(9)	.94	.91	.97	.97	.92	.77	.91
Bagging CPD(9)	.74	.85	.82	.91	.77	.68	.79
FLDA(9)	.88	.88	.97	.97	.88	.88	.91
DLDA(9)	.97	.94	.97	.97	.97	.88	.95
DQDA(9)	.97	.94	.97	.97	.97	.88	.95
BayesNetwork(29)	.74	.88	.97	.97	.83	.62	.83
Average	.92	.92	.95	.97	.92	.83	

Abstraction for Improving Machine learning (AIM)

Define a formal abstraction layer (AIM) that pre-specifies steps in the ML pipeline.



A cartoon AIM layer showing discrete components $\mathcal{F}_1, \dots, \mathcal{F}_n$ that carry out n data steps to be input into a prediction model P .



The AIM for ALL/AML Cancer Classification

The simple AIM we defined in this example. The workflow was segmented into two discrete components: Preprocessing/Feature Selection (PPFS) and Classifier (P).

Reproducibility Standards Development

Reproducibility requires community adoption and standards development.

Example: a 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.
- Journals should conduct a **reproducibility check** as part of the publication process.
- Funding agencies should instigate new research programs and pilot studies.

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.

Set the Default to “Open”

Progress in science is often hampered by researchers’ inability to reproduce or verify results. Attendees at the Yale Law School Roundtable on Data and Code Sharing identified a set of steps that agencies, journals, and journals can take to encourage those steps here, along with best practices for sharing data, design, procedures, equipment, raw results, processing, analysis, etc. In contrast, workflows are performed with a workflow, computation, or parameter. While this may be ultimately impeded by the tools and standards used.

The State of Experiments. Experimentation is a question in pure and applied mathematics. Computational methods are performed with a workflow, computation, or parameter. While this may be ultimately impeded by the tools and standards used.



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 Eva Deelman,⁴ Yolanda Gil,⁵ Brooks Hanson,⁶ Michael A. Heroux,⁷ John P.A. Ioannidis,⁸ Michael Taufer⁹

In the past two decades, computational methods have radically changed the ability of researchers from all areas of science to generate, analyze, store, and reuse data to simulate complex systems. But with these advances come challenges that are unique to computing in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often incomplete, and this paper presents a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general

understanding how computational results were derived and to reconciling any differences that might arise between independent researchers (Fig. 1). We also focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (S), which includes workflow information and metadata that describe the data and the analytical results are input to which computations (7). Access to the data and code that underlie discoveries can also enable downstream users to validate, reuse, and other efforts that include results from multiple studies.

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

someone in the field to use the same digital scholarly objects without resorting to contacting the original authors (*i.e.*, <http://bit.ly/2tWpJH>). Software metadata should include, at a minimum, license, Uniform Resource Identifier (DOI), software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly

Thematic Synthesis Across Projects

- Testbeds for evaluating actionable social change in the area of computational reproducibility.
- Enabling results comparisons allows quality assessment and improvement in data science pipelines.
- Enabling interoperability and comparisons between results allows modeling and synthesis of results.
- Permits efficiency and cost-effectiveness evaluation: re-use of methods, code, data; technology and infrastructure decision decisions.
- Working across communities and stakeholders.

3. What's Next? Future Directions

Revisit: NASEM Report Recommendations

6-6: **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 **NSF (and other funders) should consider funding of activities to promote computational reproducibility.**

Applying these ideas: The Lifecycle of Data Science

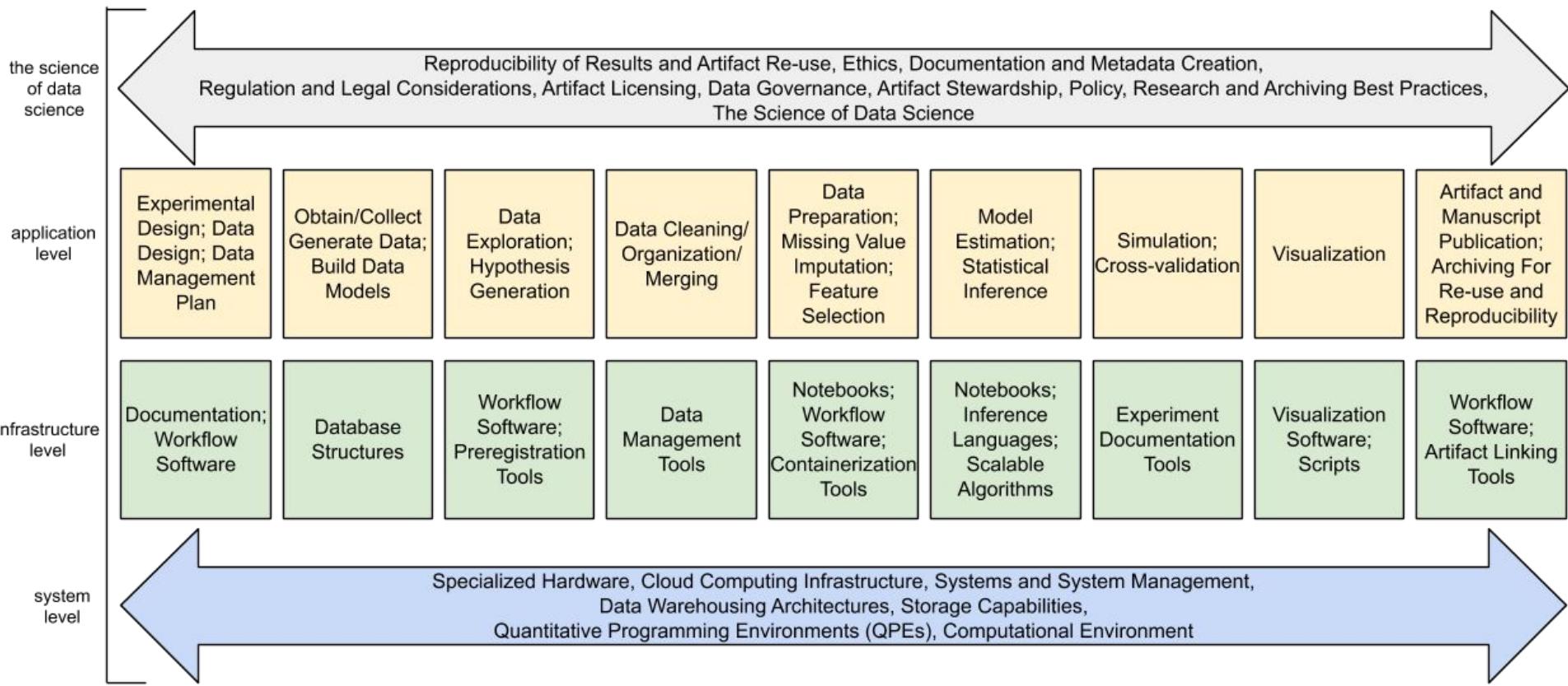
“Lifecycle of Data” is an abstraction from the Information Sciences

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

What if we applied this idea to Data Science?

- **Clarify steps** in data science projects: people/skills involved, tools and infrastructure, and reproducibility through the cycle.
- **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 Lifecycle of Data Science: An Abstraction

An abstraction that organizes the computational pipeline.. and so recognizes different contributions including from e.g.:

- Ethicists
- Knowledge and data managers
- Compute resources and cyberinfrastructure

Goals:

- Improve understanding of Data Science advancement.
- Permit the comparison of results.
- Improve research output and social impact.

Caution! Under construction!



Proposal: A Computable Scholarly Record

- A testbed for studying reproducibility and reliability in data science.
- Acts as a “living lab” that allows development/testing of infrastructure, policies, and statistical inference methods, and studying cultural barriers to reproducibility.
- Entertains meta-research queries such as:
 - Show a table with effect sizes and p-values for all phase-3 clinical trials for Melanoma;
 - List all image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;
 - List all classifiers applied to the famous ALL/AML cancer dataset, with misclassification rates;
 - Create a unified dataset containing all published whole-genome sequences with the BRCA1 mutation;
 - Randomly reassign 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 2003 and list trial name and histogram side by side.

Exposure of computational steps

A dream:

- ◆ Executability/re-executability of pipelines/code (transparency)
 - ◆ Methods application in new contexts
 - ◆ Pooling data and improved experimental power
 - ◆ Improved validation of findings
 - ◆ Comparisons of methods
 - ◆ Organization of discovery pipeline information
- **Structured dissemination** of findings enabling query and meta-analysis
- Organization of the scholarly record around **research questions**
- **Probabilistic models of correctness** in a distributed knowledge production system

A More Modest Proposal: The Knowledge Integrator

- Development of dissemination standards around results (stack agnostic).
- Central deposition of computationally reproducible results: open access, open deposit, to grow the computable scholarly record.
- Integration of results to extend knowledge e.g. systems analytics.
- The scholarly record as a dataset: overall false discovery rate; identify key questions in different fields; meta-science and assessment; benchmarking and algorithm performance..
- Pilot in receptive communities.

Conclusion

Reproducibility questions are emerging in several forms.

Their commonality is the use of computational technology.

Computation engenders a rethinking of the products of the research pipeline as part of a distributed computational system, which admits exciting new opportunities:

- a computable scholarly record as a source of data in itself leveraging analysis, modeling, system analytics and “health checks,”
- greater understanding of norms and social structures for discovery,
- enabling **efficiency, productivity, and discovery**,