

# New Approaches to Reproducibility & Reliability of Research Results

Victoria Stodden  
University of Illinois at Urbana-Champaign

Department of Industrial & Systems Engineering Seminar  
University of Southern California  
January 29, 2020

# Agenda

---

- 1. Setting the Stage: Research 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.

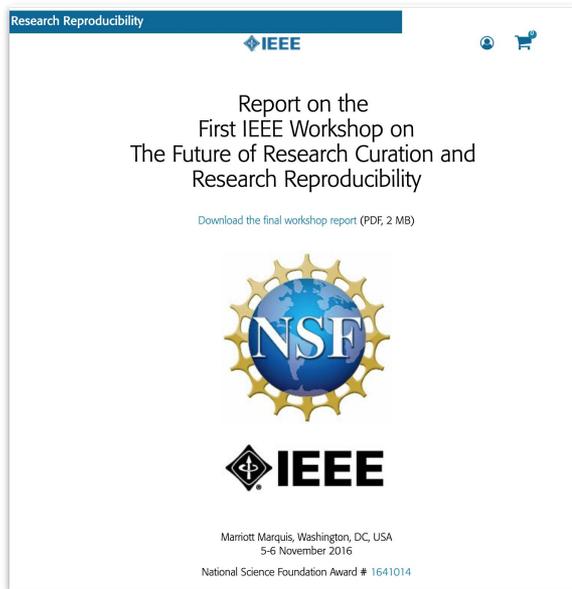
Results may come from a partner in Australia with new materials developments and a lab in St. Louis does high throughput combinatorial analysis of materials to rapidly check the results.

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

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

# Reproducibility Example 2: IEEE

## IEEE steps to reproducibility and computational transparency:



Research Reproducibility

IEEE

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

Download the final workshop report (PDF, 2 MB)



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

**FROM THE EDITOR'S DESK**

#### On Reproducible Research

By Bram Vanderborght

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 Bratz, 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

the leading role of our associate editor, Fabio Bonsignorio and colleagues, who began work on this topic ten years ago and has since organized several related workshops. Remember, we had a special issue in September 2015 ("Replicable and Measurable Robotics 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 during his keynote talk at ICRA presented the advantages of open benchmarks to advance the field of robot grasping.

Therefore, in September 2017, we launched the R-articles concept. In addition to publishing the description in the journal paper text (written following the "Good Experimental Methodology" guidelines), the required data sets, the

During the IEEE Panel of Editors meeting held this past April in Los Angeles, California, I was invited, as editor-in-chief of *IEEE Robotics and Automation Magazine*, to participate in a panel discussing reproducible research under the lead of John Baillieu, chair of the Strategic Planning Committee. Reproducibility of scientific research is imperative: it helps researchers verify results and benchmark 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 or impossible to replicate (an issue that is especially prevalent in social psychology and medicine).

Within IEEE publications, reproduc-



### 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.

to upload the code to Code Ocean, which generates a "compute capsule" that includes the code, data, results, and computational environment specifications. Code Ocean sends the EIC a review copy of the compute capsule, which is passed on to the assigned reproducibility associate editor for the article.

Again, I offer a warm invitation to submit articles with reproducible research. Together with the authors, we need to improve the procedure over time, but we can only learn from experience.

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 reward the positive behavior of authors contributing to reproducible research. One initiative suggests highlighting paper with reproducible research content, perhaps even granting awards dedicated to such papers. Discussions are in progress with the Association for Computing Machinery about the addition of a common badging system for such papers. A culture shift will be needed. Moreover, along the lines of the

Editorial Policies and Badging Pilot Partnerships: Code Ocean

# Reproducibility Example 3: 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

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

- Reproducibility implemented in Whole Tale

The image displays three overlapping screenshots related to the research article. The top screenshot shows the article page on Science Advances, titled "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments" by Fang Ren et al. The middle screenshot shows a GitHub repository for "fang-ren / Discover\_MG\_CovZr", listing files such as "data", "figures", "machine-learning", and "scripts". The bottom screenshot shows a Materials Data Facility (MDF) record for the article, including metadata like DOI (10.18126/M2B06M), year (2018), and tags like "metallic glasses" and "machine learning".

# 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**

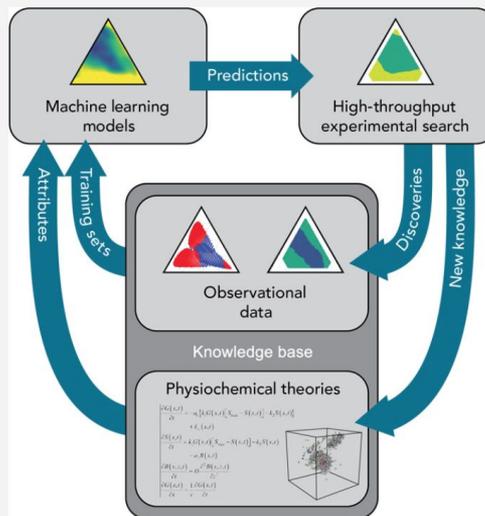
Back

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 Whole Tale dashboard for a workspace titled "Predicting the Properties of Inorga..." by Logan Ward. The interface includes a navigation bar with "WHOLE TALE DASHBOARD", "BROWSE", and "MANAGE". The workspace has tabs for "Interact", "Files", and "Metadata". The "Files" tab is active, showing a "Tale Workspace" with a table of files and folders. The table lists items such as "datasets" (27 MB), "magpie" (0), "modeling-metallic-glasses" (48.3 MB), "predicting-band-gap-energies" (5.87 MB), and several scripts like "docker.bat", "docker.bs", "README.md", and "run-all.bs".

WHOLE TALE DASHBOARD BROWSE MANAGE Victoria Stodden

< Return to Dashboard

Predicting the Properties of Inorga... Logan Ward Run Close

Interact Files Metadata

Tale Workspace

Global directory accessible across all Tales; can be synced to local machine

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

Home

External Data

Tale Workspace

# 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](https://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 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  $\mathcal{C} = \mathcal{C}(\cdot, \tilde{\mathcal{L}})$  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  $Z$ .

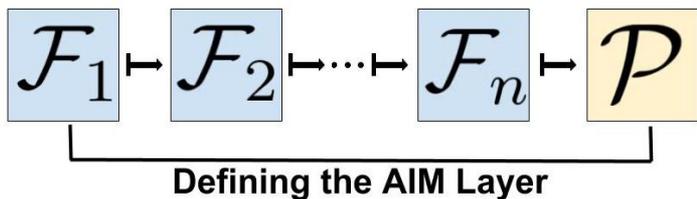
# Baseline Comparisons (5 articles, n=72 obs)

Preprocessing/Feature Selection Method							
Classifier(Paper)	1	3	6a	6b	9	29	Average
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	<b>.96</b>
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
<b>Average</b>	<b>.92</b>	<b>.92</b>	<b>.95</b>	<b>.97</b>	<b>.92</b>	<b>.83</b>	

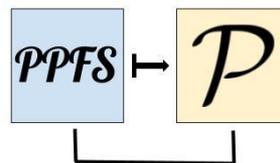
# Abstraction for Improving Machine learning (AIM)

---

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  $\mathcal{P}$ .



**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 ( $\mathcal{P}$ ).

# 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"

**P**rogress in computing is often hampered by researchers' inability to reproduce or verify results. Attendees at the Yale Law School RoundtableNov2) a set of steps that agencies, and four improve the status those steps here, all for best practices available option term goals for the tools and standard

**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 analysis, etc. In contrast, computing is performed with a workflow, compilation, or parameterization. While crippling reporting ultimately impedes the State of Experimentation, high-performance computing questions in pure automatic theorem proving of computational bounds in very high

INSIGHTS | POLICY FORUM

### REPRODUCIBILITY

## Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

by Victoria Stodden,<sup>1</sup> Marcia McNutt,<sup>2</sup> David H. Bailey,<sup>3</sup> Ewa Deelman,<sup>4</sup> Yolanda Gil,<sup>5</sup> Brooke Hanson,<sup>6</sup> Michael A. Heroux,<sup>7</sup> John P.A. Ioannidis,<sup>8</sup> Michela Taufer<sup>9</sup>

**O**ver 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/2PVP1P1>). 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

# 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. Future Research 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

---

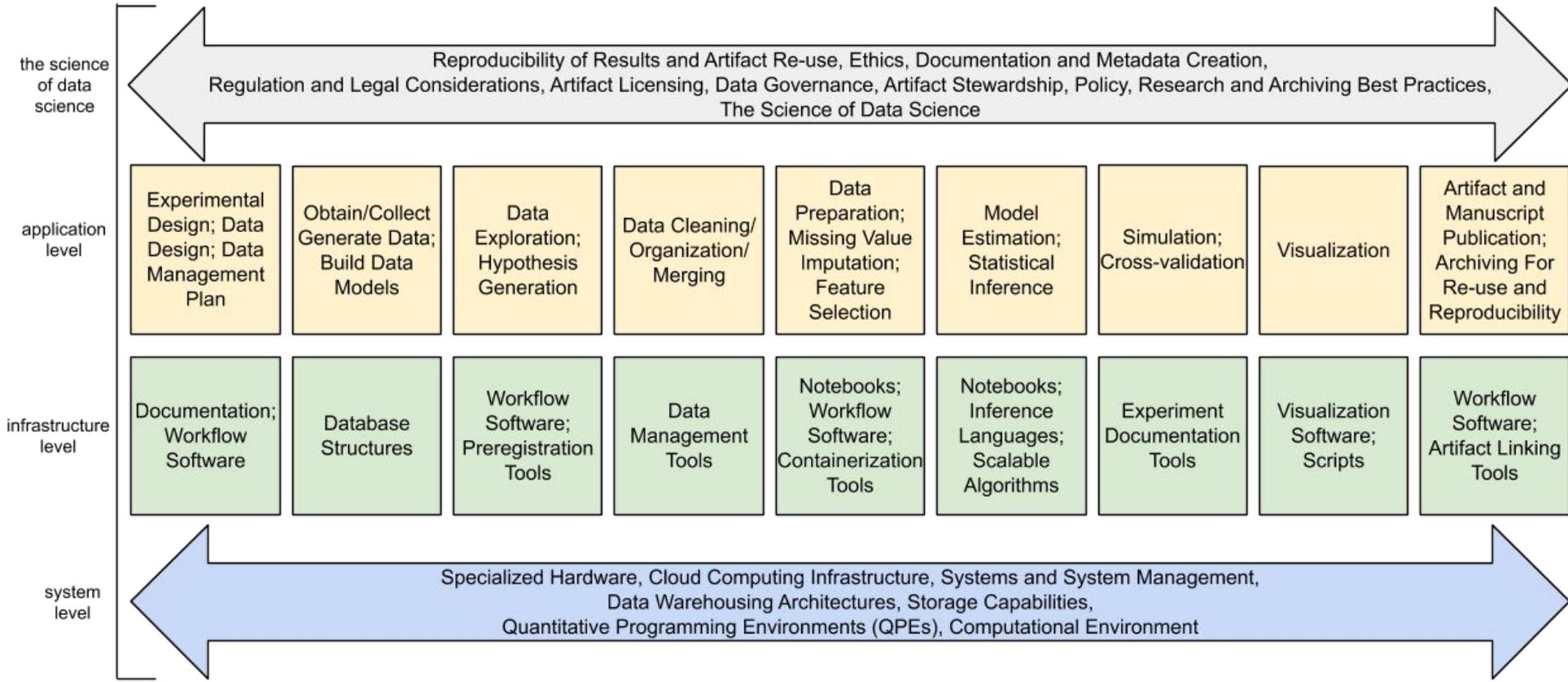
“Lifecycle of Data” is an abstraction from the Information Sciences

- 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

---

Two (ordinarily antagonistic) trends are converging:

Research will become **massively more compute and data intensive**,  
and

Research computing will become **dramatically more transparent**.

These are reinforcing trends, which can admit 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**,