

The Lifecycle of Data Science: A Framework for Advancing Computational and Data-enabled Research

Victoria Stodden

School of Information Sciences
University of Illinois Urbana Champaign
vcs@stodden.net

CS Distinguished Seminar
Department of Computer Science
Northwestern University
November 20, 2019

Agenda

1. Definitions: Unpacking Reproducibility
2. Framing: Introducing the Lifecycle of Data Science
3. Infrastructure: The Whole Tale Project

1. Unpacking Reproducibility

No crisis . . . No complacency.

- Improvements are needed.
- Reproducibility is important but not currently easy to attain.
- Aspects of replicability of individual studies are a serious concern.

Neither are the main or most effective way to ensure reliability of scientific knowledge.

Reproducibility
and Replicability
in Science



8

Harvey Fineberg

Chair, Committee on Reproducibility and Replicability in Science

NASEM Report 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.

Parsing Aspects of Reproducibility

Empirical Reproducibility
(Replicability)

Statistical Reproducibility

Computational Reproducibility

The collage consists of three overlapping screenshots. The top screenshot shows the Nature journal website with a search bar and navigation links. The middle screenshot shows the Science journal website with a sidebar and a main article titled 'Reproducibility'. The bottom screenshot shows the SIAM website with a news article titled '“Setting the Default to Reproducible” in Computational Science Research'.

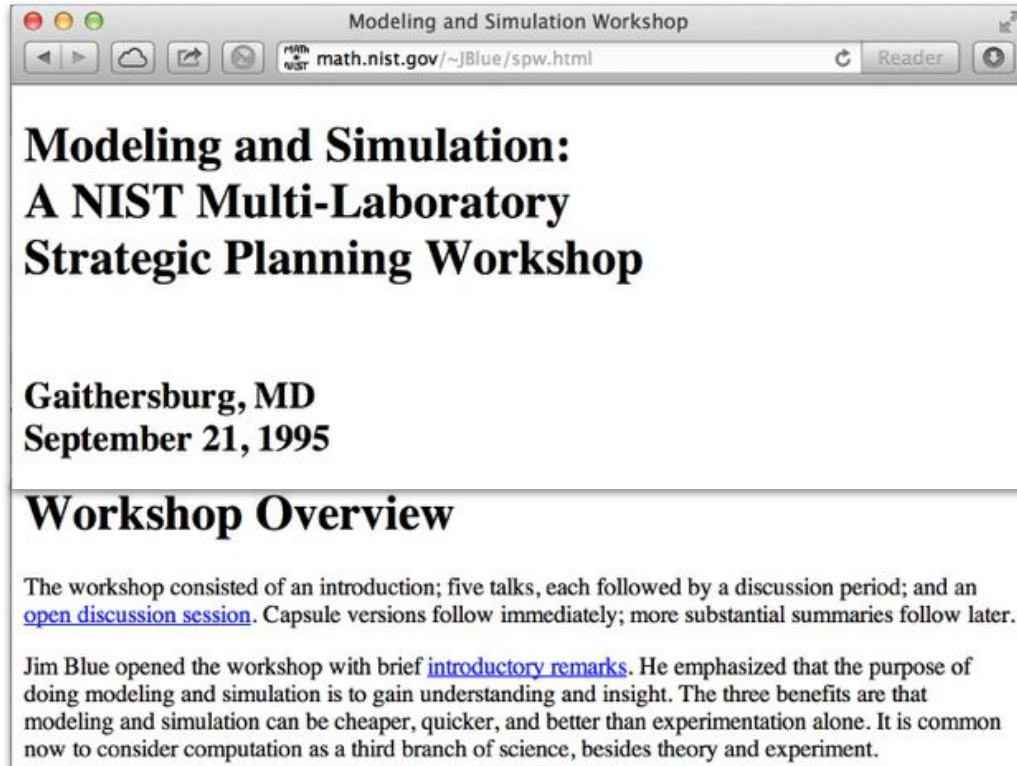
Computational Reproducibility

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic.
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

- Branch 3,4? (computational): large scale simulations / data driven computational science.



Modeling and Simulation Workshop

math.nist.gov/~jBlue/spw.html

Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

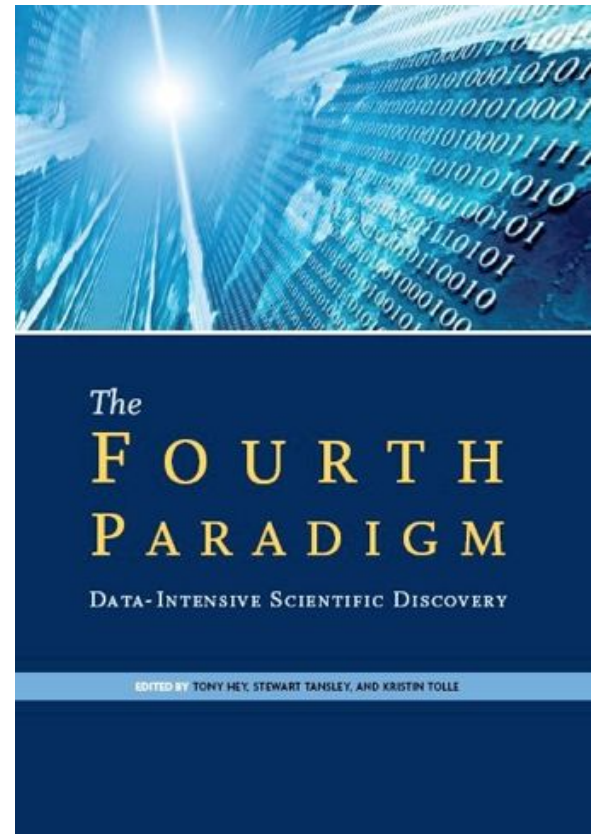
Gaithersburg, MD
September 21, 1995

Workshop Overview

The workshop consisted of an introduction; five talks, each followed by a discussion period; and an [open discussion session](#). Capsule versions follow immediately; more substantial summaries follow later.

Jim Blue opened the workshop with brief [introductory remarks](#). He emphasized that the purpose of doing modeling and simulation is to gain understanding and insight. The three benefits are that modeling and simulation can be cheaper, quicker, and better than experimentation alone. It is common now to consider computation as a third branch of science, besides theory and experiment.

“It is common now to consider computation as a third branch of science, besides theory and experiment.”



“This book is about a new, fourth paradigm for science based on data-intensive computing.”

The Ubiquity of Error

The central motivation for the scientific method is to root out error:

- Deductive branch: the well-defined concept of the proof,
- Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation and Data Science present only *potential* third/fourth branches of the scientific method, until the development of comparable standards.

Community Approach



Researchers
(processes)



Funders
(policy)



**Universities/
institutions**
(hiring/promotion;
programmatic change)



**Universities/
libraries**
(empowering w/tools)



Publishers
(TOP guidelines)



Scientific Societies



Regulatory Bodies
(OSTP)

REPRODUCIBILITY

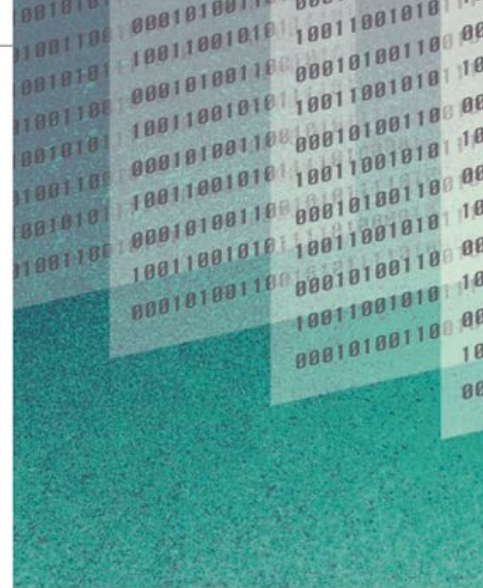
Enhancing reproducibility for computational methods

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 Taufer⁸

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 transpar-

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



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. [**Access to the computational steps taken to process data and generate findings is as important as access to data themselves.**](http://</p>
</div>
<div data-bbox=)

Stodden, Victoria, et al. "Enhancing reproducibility for computational methods." *Science* 354(6317) (2016)

ness Promotion (TOP) guidelines (1) and recommendations for field data (2), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre-

results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or

All data, code, and workflows, including software written by the authors, should be cited in the references section (10). We suggest that software citation include software version information and its unique identifier in addi-

“Reproducibility Enhancement Principles (REPS)”

1. **Share data, software, workflows**, and details of the computational environment that generate published findings in open trusted repositories.
2. **Persistent links should appear in the published article** and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
3. To enable credit for shared digital scholarly objects, **citation should be standard**.
4. To facilitate reuse, adequately **document** digital scholarly artifacts.
5. **Use Open Licensing** when publishing digital scholarly objects.
6. Journals should conduct a **reproducibility check** as part of the publication process and should enact the TOP standards at level 2 or 3.
7. To better enable reproducibility across the scientific enterprise, **funding agencies should instigate new research programs and pilot studies**.

Key Recommendations NASEM Report 2019

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. 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..

Key Recommendations NASEM Report 2019

6-3: **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.

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, as a low-cost way to enhance both.

Key Recommendations NASEM Report 2019

6-5: In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the **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.
- **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.
- 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. Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Key Recommendations NASEM Report 2019

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.

2. Applying these ideas: The Lifecycle of Data Science

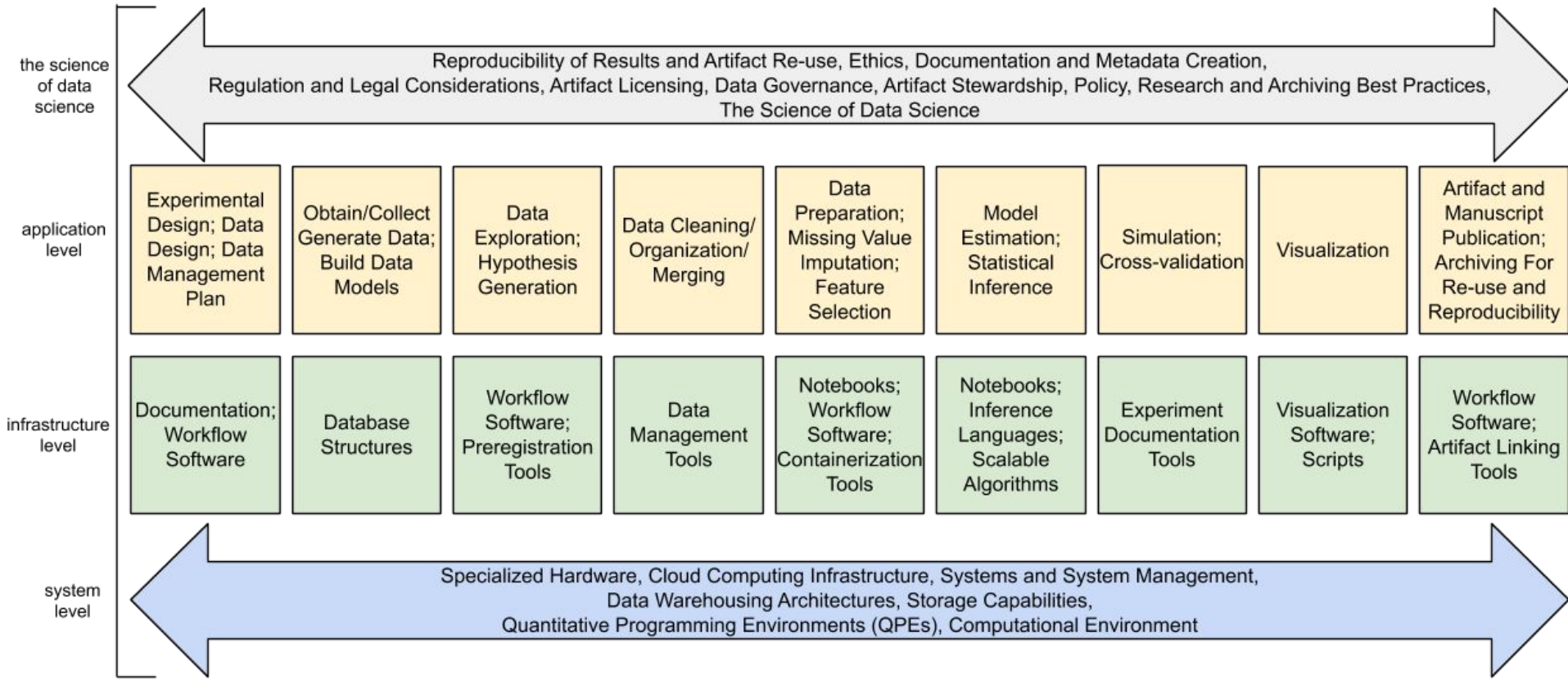
“Lifecycle of Data” is an abstraction from the Information Sciences

- 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, curricula, training, and other programmatic initiatives.
- **Develop and reward contributing areas.**

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
- Data managers
- Compute resources and cyberinfrastructure
- ...

Goals:

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

3. Infrastructure: The Whole Tale Project

5 institutions, NSF funded co-operative project:

U Illinois (NCSA): Bertram Ludäscher, Victoria Stodden, Matt Turk

- overall lead (co-operative agreement)
- reproducibility; provenance; open source software development; outreach

U Chicago (Globus): Kyle Chard

- data transfer & storage; compute; infrastructure

UC Santa Barbara (NCEAS): Matt Jones

- (meta-)data publishing; provenance; repositories

U Texas, Austin (TACC): Niall Gaffney

- compute; HTC; “big tale”; Science Gateways

U Notre Dame (CRC): Jarek Nabrzyski

- UX design; UI design

What is Whole Tale?

A Double Entendre:

- **Whole *tale***: captures the end-to-end scientific discovery story, including computational aspects
- **Long *tail***: includes all computational research, e.g. small scale research

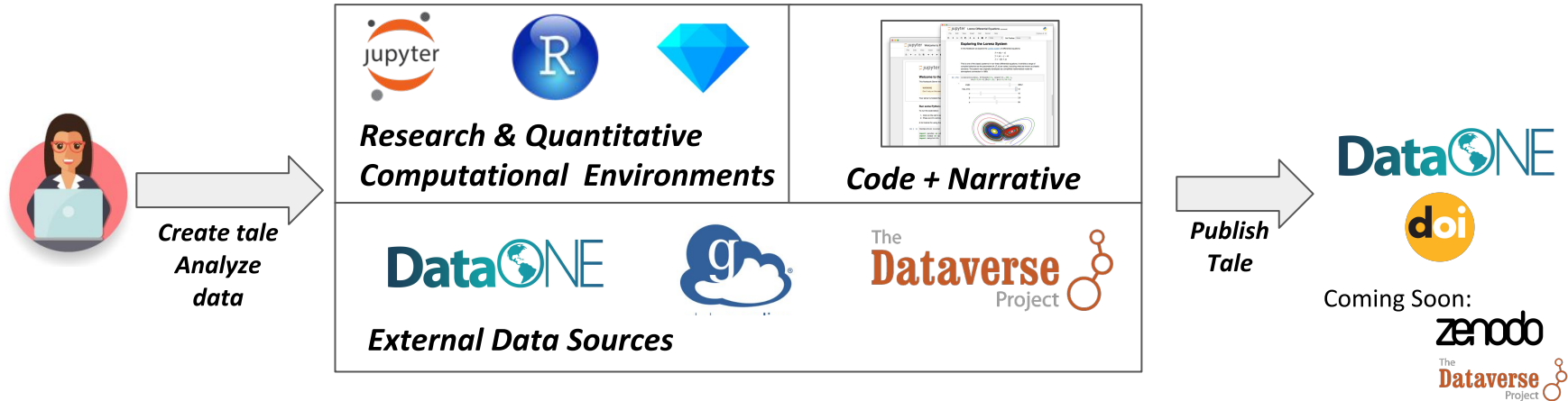
Addresses problems scientists face:

- **Reproducibility** (and re-use) challenges in computational & data-enabled research (*e.g. data+code access, dependency hell, ...*)

Whole Tale Approach:

- Directly respond to community needs and requirements
- Open source project
- Platform to create, publish, and execute reproducible tales
- Simplify process of creating & verifying reproducible computational artifacts
- <https://dashboard.wholetale.org>

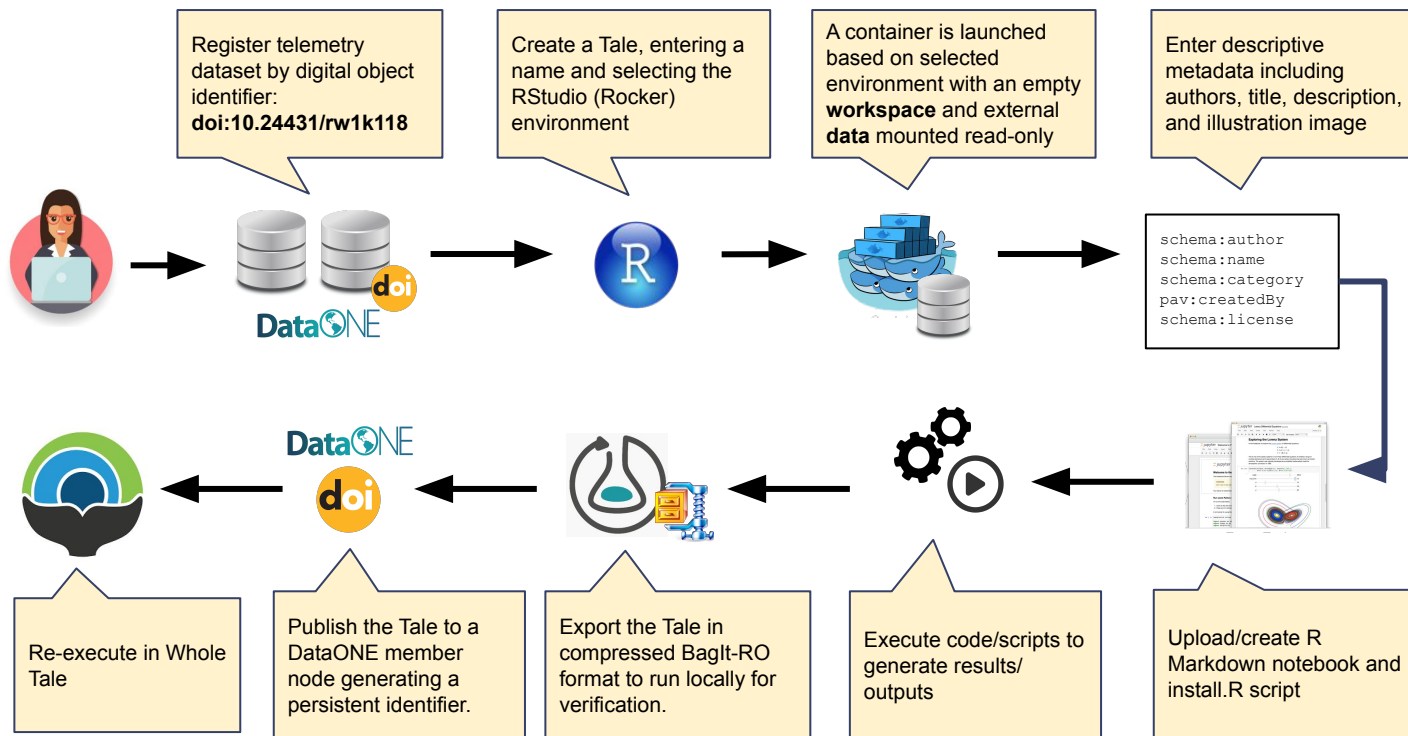
Whole Tale Platform Overview



- **Authenticate** using your institutional identity
- **Access** commonly-used **computational environments**
- Easily **customize** your environment (via repo2docker)
- Reference and access externally **registered data**

- Create or upload **your data and code**
- Add **metadata** (including **provenance** information)
- Submit code, data, and environment to **archival repository**
- Get a **persistent identifier**
- **Share** for **verification** and **re-use**

Tale Creation Workflow



Simplifying Computational Reproducibility

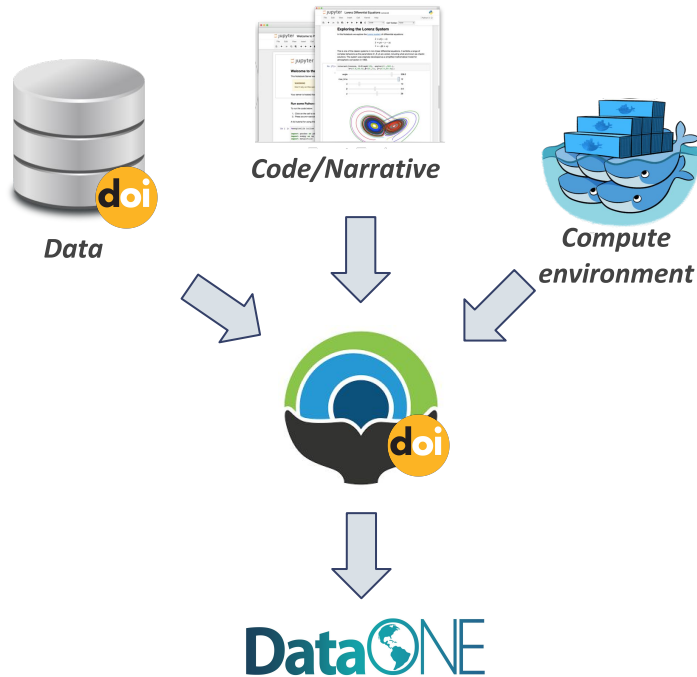
Researchers can easily package and share “Tales”

Data, Code, and Compute Environment including

- Narrative,
- Code, data, workflow information,
- Inputs, outputs, and intermediates to re-create the computational results from a scientific study

Empowers users to verify and extend results with different data, methods, and environments.

What exactly is (in) a Tale?



Tale::Research Object

- ✓ Contains data (by reference), code, narrative, compute environment, meta data including licensing
- ✓ Executable
- ✓ Publishable

Wholetale.org: Browse Existing Tales

The screenshot shows the WholeTale dashboard interface. At the top, there is a navigation bar with the following elements: the WholeTale logo, the text 'WHOLETALE DASHBOARD', and a menu with 'BROWSE', 'RUN', 'MANAGE', and 'COMPOSE'. On the right side of the navigation bar, there is a user profile icon for 'Craig Willis' and a share icon.

The main content area is divided into two panels. The left panel is titled 'Browse Tales' and includes a search bar with the placeholder text 'Search tales...'. Below the search bar, there is a filter dropdown set to 'All' and a link to 'Switch to list view'. The panel displays a grid of three 'Tales' (interactive stories):

- COMPUTATIONAL CHEMISTRY: Anharmonic vibrational structure of...** This tale features molecular diagrams of a carbon dioxide dimer and a 'jupyter' logo, indicating it is an interactive computational chemistry project.
- ARCHAEOLOGY: Climate change stimulated agricultu...** This tale displays a grid of 20 heatmaps showing climate variations over time and space, with an 'R' logo indicating it is an R-based project.
- ECONOMICS: L2-Boosting for Economic Applicatio...** This tale features a population pyramid chart with age groups on the y-axis (ranging from 20-24 to 80+) and population counts on the x-axis. It includes an 'R' logo and a description of a replication package for L2-boosting in economic applications.

The right panel is titled 'Launched Tales' and currently shows one tale: 'L2-Boosting for Economic Applicatio...'. Each tale entry in the 'Launched Tales' panel includes a small 'R' logo icon and a close button (an 'x' in a square).

Compose New Tales

Create New Tale ✕

Tale name:

Compute environment:

Compute Environment ▼

OFFICIAL ENVIRONMENTS

- Jupyter Notebook
- Jupyter with Spark
- JupyterLab
- OpenRefine 2.8
- RStudio

Input data: *Add data after Tale creation using your chosen compute environment, or the Files tab of your running Tale.*

▼

Run and Interact with Tales

The screenshot displays the WholeTale dashboard interface. At the top, navigation tabs include DASHBOARD, BROWSE, RUN (highlighted), MANAGE, and COMPOSE. The user is identified as Craig Willis. The main workspace is titled "L2-Boosting for Economic Applicatio..." by Ye Luo and Martin Spindler. The interface is divided into several panels:

- Interact Panel:** Contains the R script editor with the following code:

```
1 #####
2 # L2-Boosting for Economic Applications
3 #####
4 # Parameter for simulation study
5 rm(list=ls())
6 source("DGP.R")
7 source("helper.R")
8 R <- 500 # number of repetitions
9 set.seed(12345)
10 library(MASS)
11 library(mvtnorm)
12 library(hdm)
13 library(newboost) # can be downloaded from R-Forge or requested by the a
14 #####
15 # IV Estimation
```
- Environment Panel:** Shows the current R environment with the following data objects:

Object	Type	Value
data	List of 3	
ds	num [1:90, 1]	-1.24 -0.974 1.33 -0.154 -0...
ED	List of 6	
ED1	List of 6	
EDB	List of 6	
- Files Panel:** Shows a file explorer for the workspace containing:

Name	Size	Modified
apt.txt	5 B	Mar 6, 2019, 1:43 PM
DGP.R	1.5 KB	Mar 5, 2019, 3:36 PM
helper.R	9.2 KB	Mar 5, 2019, 3:36 PM
install.R	148 B	Mar 5, 2019, 3:36 PM
Readme.pdf	60.7 KB	Mar 5, 2019, 3:36 PM
runtime.txt	13 B	Mar 5, 2019, 3:36 PM
Sim_AER.RData	6.6 MB	Mar 5, 2019, 4:14 PM
Sim_AER_V3.R	5.3 KB	Mar 5, 2019, 3:46 PM
- Console Panel:** Shows the R prompt and the command `load("~/WholeTale/workspace/Sim_AER.RData")` being executed.
- Launched Tales Panel:** Shows a list of active sessions, including the current one.

At the bottom of the dashboard, there is a footer with the text: "© WholeTale (Build: {commit}) Report a problem" and "This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450."

Explore and Use Tale Metadata

The screenshot displays the WholeTale dashboard interface. At the top, a dark navigation bar contains the following elements from left to right: the WholeTale logo, the text 'WHOLETALE DASHBOARD', and navigation links for 'BROWSE', 'RUN', 'MANAGE', and 'COMPOSE'. On the right side of the navigation bar, there is a user profile icon for 'Craig Willis', an information icon, and a share icon.

The main content area is divided into two panels. The left panel shows the metadata for a Tale titled 'L2-Boosting for Economic Applications' by 'Ye Luo and Martin Spindler'. The 'Metadata' tab is selected, showing fields for Title, Authors, Category, Environment, Date Created, and Last Updated. Below these fields are 'Edit' and 'Preview' buttons. A 'Description' section contains a text area with the following text: 'Replication package for: L₂-Boosting for Economic Applications. The authors present the L₂-Boosting algorithm and two variants, namely post-Boosting and orthogonal Boosting. Building on results in Ye and Spindler (2018), they demonstrate how boosting can be used for estimation and inference of low-dimensional treatment effects. In particular, we consider estimation of a treatment effect in a setting with very many controls and in a setting with very many instruments. We provide simulations and analyze two real applications. Based on <https://www.aeaweb.org/articles?id=10.1257/aer.p20171040>'. At the bottom of the metadata panel, there is an 'Illustration' field with a URL and a 'Generate Illustration' button.

The right panel, titled 'Launched Tales', shows a list of launched tales. The first entry is 'L2-Boosting for Economic Applications' with a close button (X) on the right.

Publish to repositories with one click

The screenshot shows the DataONE interface. At the top, there are navigation links: About, News, Participate, Resources, Education, and Data. Below this is a search bar and a 'DATAONE SEARCH' section with 'Search', 'Summary', and 'Jump to: DOI or ID' options. The main content area displays the dataset title 'Daniel White and Lilian Alessa. Humans and Hydrology at High Latitudes: Water Use Information, Arctic Data Center. doi:10.5065/D6862DM8.' and a 'Choose an analysis environment to interactively explore this dataset online using Whole Tale.' prompt. Below the title, there are statistics for Citations (0), Downloads (183), and Views (72). A 'Download All' button is visible. A table lists files in the dataset:

Name	File type	Size	Views	Downloads
Metadata: science_metadata.xml	EML v2.1.1	8 KB	65 views	Download
estimated_use_of_water_in_US_2000.pdf	PDF	6 MB	6 downloads	Download
estimated_use_of_water_in_US_2005.pdf	PDF	5 MB	5 downloads	Download
first_nations_canada_water_and_wastewater_systems.pdf	PDF	365 KB	4 downloads	Download

At the bottom, there is a 'General' section with the identifier 'doi:10.5065/D6862DM8'.

The screenshot shows the Harvard Dataverse interface. At the top, there are navigation links: Search, About, User Guide, Support, Sign Up, and Log In. Below this is the Harvard Dataverse logo and the title 'AMERICAN JOURNAL of POLITICAL SCIENCE'. The main content area displays the dataset title 'Replication Data for: Greater Expectations: A Field Experiment to Improve Accountability in Mali' and the author 'Gottlieb, Jessica, 2015, "Replication Data for: Greater Expectations: A Field Experiment to Improve Accountability in Mali", https://doi.org/10.7910/DVN/29010, Harvard Dataverse, V3, UNF:6:CCpqsMSJGMVbNdHwPM1Q--- [fileUNF]'. Below the title, there are statistics for Downloads (100) and a 'Cite Dataset' button. A 'Description' section contains the following text:

I argue that if citizens systematically underestimate what their government can and should do for them, then they will hold politicians to a lower standard and sanction poor performers less often. A field experiment across 95 localities in Mali in which randomly assigned localities receive a civics course identifies the effect of raising voter expectations of government on their willingness to hold leaders accountable. The course provides information about local government capacity and responsibility as well as how local politicians perform relative to others, effectively raising voter expectations of what local governments can and should do. Survey experiments among individuals in treated and control communities (N=5,560) suggest that people in treated villages are indeed more likely to sanction poor performers and vote based on performance more often. A behavioral outcome—the likelihood that villagers challenge local leaders at a town hall meeting—adds external validity to survey findings.

Subject: Social Sciences
Keyword: Government accountability; Voting behavior; Field experiments
Related Publication: Gottlieb, Jessica, 2016, "Greater Expectations: A Field Experiment to Improve Accountability in Mali," American Journal of Political Science 60 (1): 143-157, doi: 10.1111/ajps.12186

- Enables **turnkey exploratory data analysis** on existing published datasets
- **DataONE** and **Dataverse** networks cover > 90 major research repositories

Whose problems are we addressing?

Researchers, scientists, others may be

- **creators** of tales e.g. share your findings in a tale
- **reviewers** of articles can review tales e.g. reproduce new scientific claims
- **(re-)users** of tales e.g. build upon progress of others

Standards development for research sharing: “Tale” definition

Conclusion

Two (ordinarily antagonistic) trends are converging:

Scientific projects will become **massively more compute and data intensive**,
Research computing will become **dramatically more transparent**.

These are reinforcing trends, which can admit a computable scholarly record, leveraging the central role of infrastructure.

Better transparency will allow people to run much more ambitious computational experiments. And **better** computational experiment **infrastructure** will allow researchers to be **more transparent**.

This approach is used because it enables **efficiency/productivity**, and **discovery**.

Caution! Under construction!



4. 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**

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.