

Telling the Whole Tale via Reproducible Data Reuse



DataONE

Matthew B. Jones

 0000-0003-0077-4738
 @metamattj

ESIP Summer Meeting
July 16-19, 2019
Tacoma, WA



Data Reuse for Reproducible Synthesis

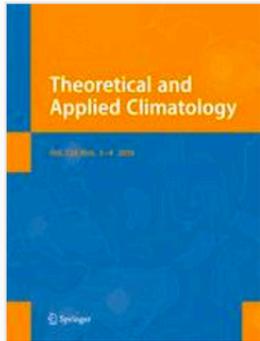


 Fiorenza Micheli, Benjamin Halpern, Shaun Walbridge, Saul Ciriaco, Francesco Ferretti, et al. **Cumulative Human Impacts on Mediterranean and Black Sea Marine Ecosystems, 2013.** Knowledge Network for Biocomplexity. doi:10.5063/F15M63Z8.
    2.9K  369

 Benjamin Halpern, Shaun Walbridge, Kimberly Selkoe, Carrie Kappel, Fiorenza Micheli, et al. **A Global Map of Human Impact on Marine Ecosystems, 2008.** Knowledge Network for Biocomplexity. doi:10.5063/F19C6VN5.
    17.6K  524

 Benjamin Halpern, Shaun Walbridge, Kimberly Selkoe, Carrie Kappel, Fiorenza Micheli, et al. **Transformed Stressor Data: A Global Map of Human Impact on Marine Ecosystems, 2008.** Knowledge Network for Biocomplexity. doi:10.5063/F1F47MCW.
    10.8K  93

DataONE



[Theoretical and Applied Climatology](#)

November 2016, Volume 126, [Issue 3–4](#), pp 699–703 | [Cite as](#)

Learning from mistakes in climate research

[Authors](#)

[Authors and affiliations](#)

Rasmus E. Benestad , Dana Nuccitelli, Stephan Lewandowsky, Katharine Hayhoe, Hans Olav Hygen, Rob van Dorland, John Cook

[Open Access](#) | [Original Paper](#)

First Online: 20 August 2015

3.2k

Shares

103k

Downloads

18

Citations

replicationDemos

help

Meta

demo

html

R

replicationDemos.rdb

replicationDemos.rdx

replicationDemos

data

Rdata.rdx

Rdata.rdb

Rdata.rds

INDEX

NAMESPACE

DESCRIPTION

Ships with an R package



Edzer Pebesma

@edzerpebesma

Follow

Replying to @jhollist @metamattj

It is on CRAN, but in Archived; I could install it after installing a bunch of other Archived packages from source, and could run a number of examples. Another number depended on web resources no longer available.

5:04 AM - 14 Jul 2019



Parsing **Reproducibility**

- **Empirical Reproducibility:**
 - traditional empirical experiments, e.g. at the bench/lab
- **Statistical Reproducibility:**
 - statistical methodology used permits generalizability of data inferences
- **Computational Reproducibility:**
 - transparency of computational steps that produce scientific findings

Simplifying **Computational Reproducibility** in Whole Tale



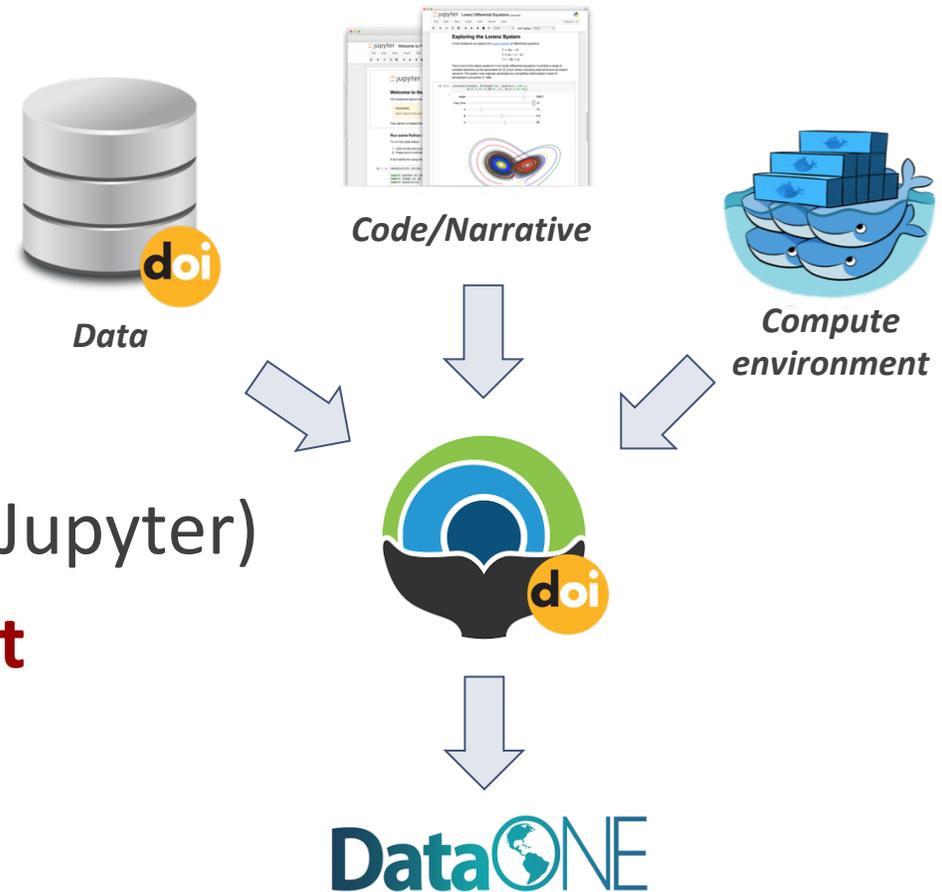
- Researchers can easily package and share *tales*:
 - Data, Code, and Compute Environment
 - to re-create the computational results from a scientific study
 - achieving computational reproducibility
 - thus “setting the default to reproducible.”
- Also empowers users to verify and extend results with **different** data, methods, and environments.

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. (2013). *Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics*, ICERM workshop (2013)



What exactly is (in) a **Tale**?

- **Tale** = executable **research object**, i.e.
 - **data** (references)
 - **+ code** (computational methods)
 - **+ narrative** (traditional science story)
 - **+ compute environment** (e.g. RStudio, Jupyter)
- Captured in a standards-based **tale format** complete with metadata



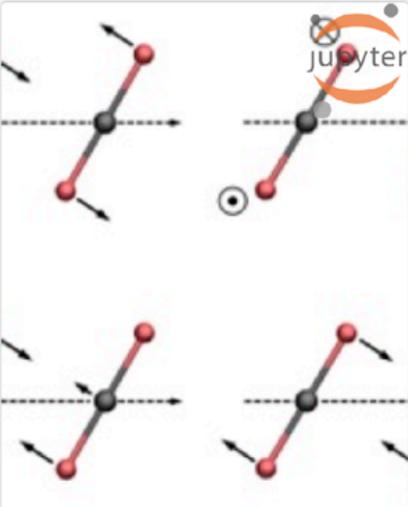


Browse Tales Launch to add to Launched Tales list



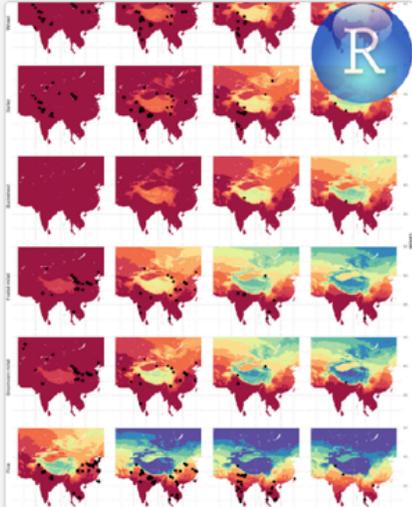
Search tales...

All



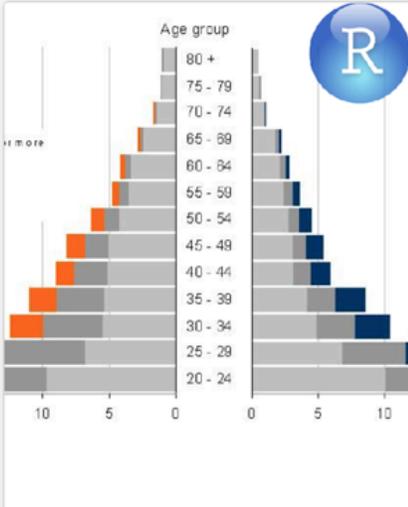
COMPUTATIONAL CHEMISTRY Anharmonic vibrational structure of...

This project produces all of the data from the Anharmonic vibrational structure of the carbon dioxide dimer with a many-body potential energy surface journal article. The project solves the vibrational Schrodinger equation for the CO2 monomer and dimer



ARCHAEOLOGY Climate change stimulated agricultu...

Ancient farmers experienced climate change at the local level through variations in the yields of their staple crops. However, archaeologists have had difficulty in determining where, when, and how changes in climate affected ancient farmers. We



ECONOMICS L2-Boosting for Economic Applicatio...

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 Yu and Qu (2010), they

Launched Tales

L2-Boosting for Economic Applicatio...

Browse Existing Tales ...



Compose Create a new Tale by pairing a compute environment with a dataset

Tale name:

Compute environment:

 RStudio (rocker/geospatial)

Input data:

[Launch New Tale](#)

... Compose New Tales ...

Environments

-  RStudio (rocker/geospatial) i
-  Jupyter Classic i
-  RStudio i
-  Jupyter Lab i





L2-Boosting for Economic Applicatio...
Ye Luo and Martin Spindler

Interact Files Metadata

File Edit Code View Plots Session Build Debug Profile Tools Help

```

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 History Connections Jobs

Global Environment

Data

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 Plots Packages Help Viewer

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 Terminal

```

/WholeTale/workspace/

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> load("/WholeTale/workspace/Sim_AER.RData")
> |
    
```

Launched Tales

L2-Boosting for Economic Applicatio... ✕

...

Run & Interact with Tales

...



DataONE

About News Participate Resources Education Data

DATAONE SEARCH: Search Summary Jump to: DOI or ID Go Sign in or Sign up

Search / Metadata

Daniel White and Lilian Alessa. Humans and Hydrology at High Latitudes: Water Use Information. Arctic Data Center. doi:10.5065/D6862DM8.

Citations 0 Downloads 183 Views 72 Copy Citation Analyze RStudio Jupyter Notebook

Name	File type	Size	Views/Downloads	Download
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

Show 13 more items in this data set

General Identifier doi:10.5065/D6862DM8

Search About User Guide Support Sign Up Log In

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

an Journal of Political Science (AJPS) Dataverse > Replication Data for: Greater Expectations: A Field Experiment to Improve Accountability in Mali

Replication Data for: Greater Expectations: A Field Experiment to Improve Accountability in Mali Version 3.0

Explore Cite Dataset Analyze in WT Learn about Data Citation Standards

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=5,660) 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.

Social Sciences

Government accountability, Voting behavior, Field experiments

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

... Integrate Data Repos with Whole Tale!

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



Publish Tale

Publishing will create an immutable copy of your Tale with a DOI. [i](#)

This process will allow another user to easily rerun your published analysis using the WholeTale platform.

Please choose a target repository.*

DataONE-The Knowledge Network for Biocomplexity ▼

More Details ▼

Your published Tale will include everything that has been uploaded to its associated workspace.

The following required files will be generated and published along with the Tale itself:

Quantifying FAIR: metadata improvement and guidance in the DataONE repository network

- manifest.json [i](#)
- environment.json [i](#)
- LICENSE [i](#)
- README.md [i](#)
- metadata.xml [i](#)

This process will allow another user to easily rerun your published analysis using the WholeTale platform.

For more information about publishing, please consult the [Publishing Guide](#).

Cancel

Publish [✓](#)

... Publish Data, Code, and Environment

- Enables **full circle** reproducibility to **DataONE** repositories that accept API deposits

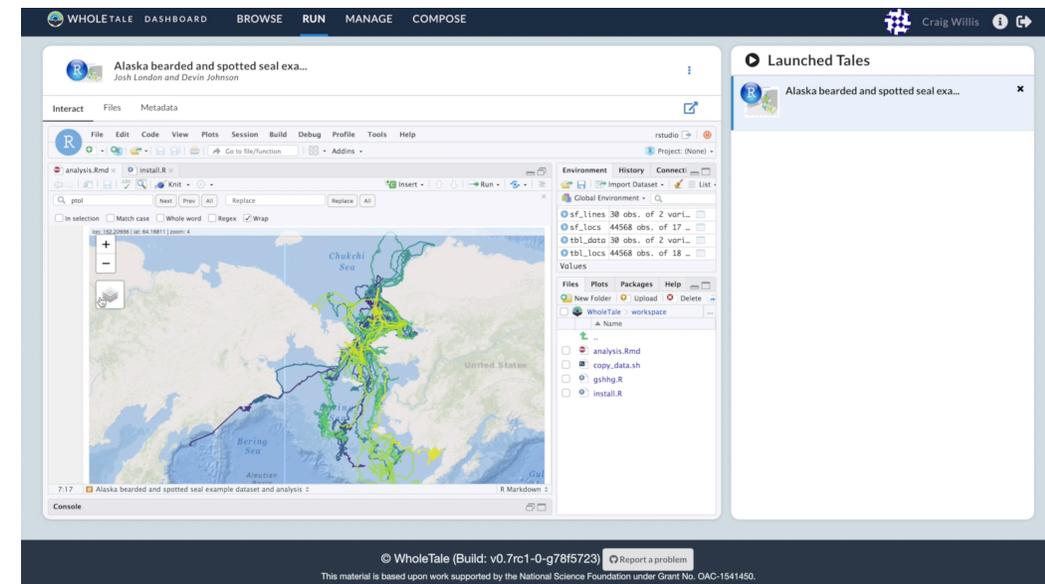


Whole Tale Forecast Demo

Demonstration of a model to predict the movement paths of seals using satellite telemetry data.

Based on analysis and models by:
Josh London and Devin Johnson
NOAA Marine Mammal Laboratory

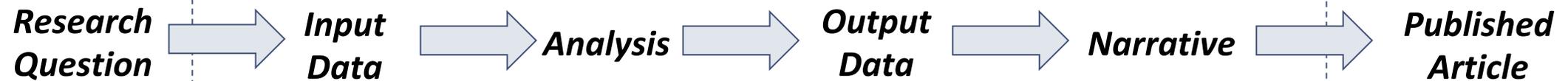
<https://youtu.be/MI5d7r5OtCk>



Accelerating Reproducible Open Science



Verify / Reproduce / Re-use



Accelerate



Whole Tale **Collaboration** (PI Team)

- **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





The Whole **Team** (members **present** in bold)

- **Adam** Brinckman (Notre Dame, Dev)
- **Bertram** Ludäscher (UIUC, PI)
- **Bryce** Mecum (UCSB, former Dev)
- **Craig Willis** (UIUC, Dev, tech project manager)
- **Damian** Perez (Notre Dame, former Dev)
- **Ian** Taylor (Notre Dame, SP, Dev)
- **Jarek Nabrzyski** (Notre Dame, co-PI)
- **Joe** Stubbs (U Texas, Dev)
- **Kacper Kowalik** (UIUC, Dev, Senior Architect)
- **Kandace** Turner (UIUC, former project mgr)
- **Kristina** Davis (Notre Dame, UI, UX)
- **Kyle** Chard (U Chicago, co-PI)
- **MT Campbell** (UIUC, project manager)
- **Matt Jones** (UCSB, co-PI)
- **Matt Turk** (UIUC, co-PI)
- **Michael** Lambert (UIUC, Dev)
- **Mihael** Hategan (U Chicago, Dev)
- **Niall Gaffney** (U Texas, co-PI)
- **Rachel** Volentine (UTK, UX)
- **Sebastian** Wyngaard (Notre Dame, Dev)
- **Sivakumar** Kulasekaran (U Texas, former Dev)
- **Thomas** Thelen (UCSB, Dev)
- **Timothy** McPhillips (UIUC, Dev)
- **Victoria** Stodden (UIUC, co-PI)

+ *WT Summer Interns (7); WT/RDA Fellows (4+4); WG Leads (5); other collaborators*



This material is based upon work supported by the National Science Foundation under Grant No. 1541450 and for DataONE under Grant No. 0830944 and 1430508.