Telling the Whole Tale via Reproducible Data Reuse

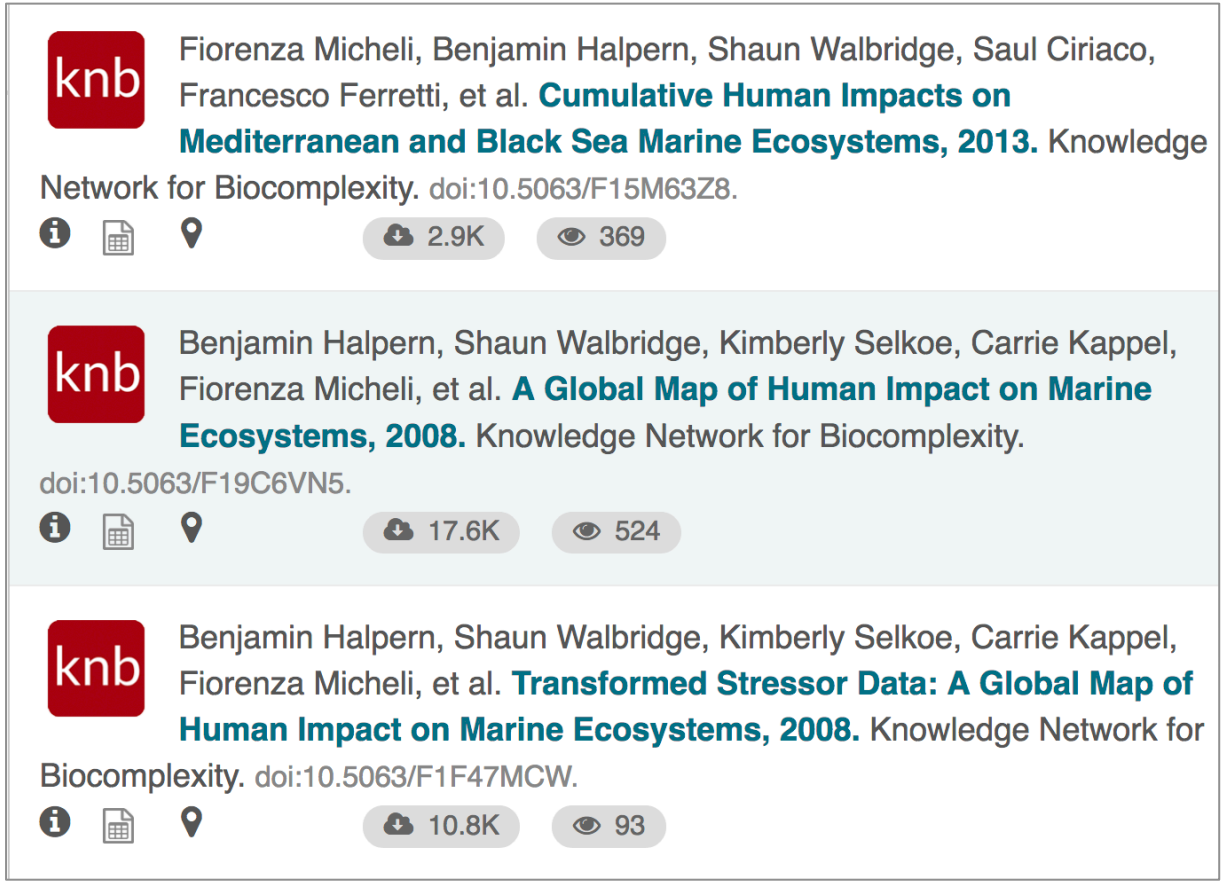
DataONE

Matthew B. Jones
0000-0003-0077-4738
@metamattj

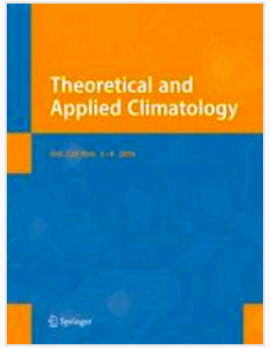
ESIP Summer Meeting
July 16-19, 2019
Tacoma, WA

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
THE UNIVERSITY OF CHICAGO
UCSB UNIVERSITY OF CALIFORNIA SANTA BARBARA
TEXAS The University of Texas at Austin
UNIVERSITY OF NOTRE DAME

Data Reuse for Reproducible Synthesis



DataONE



Theoretical and Applied Climatology

November 2016, Volume 126, Issue 3–4, pp 699–703 | Cite as

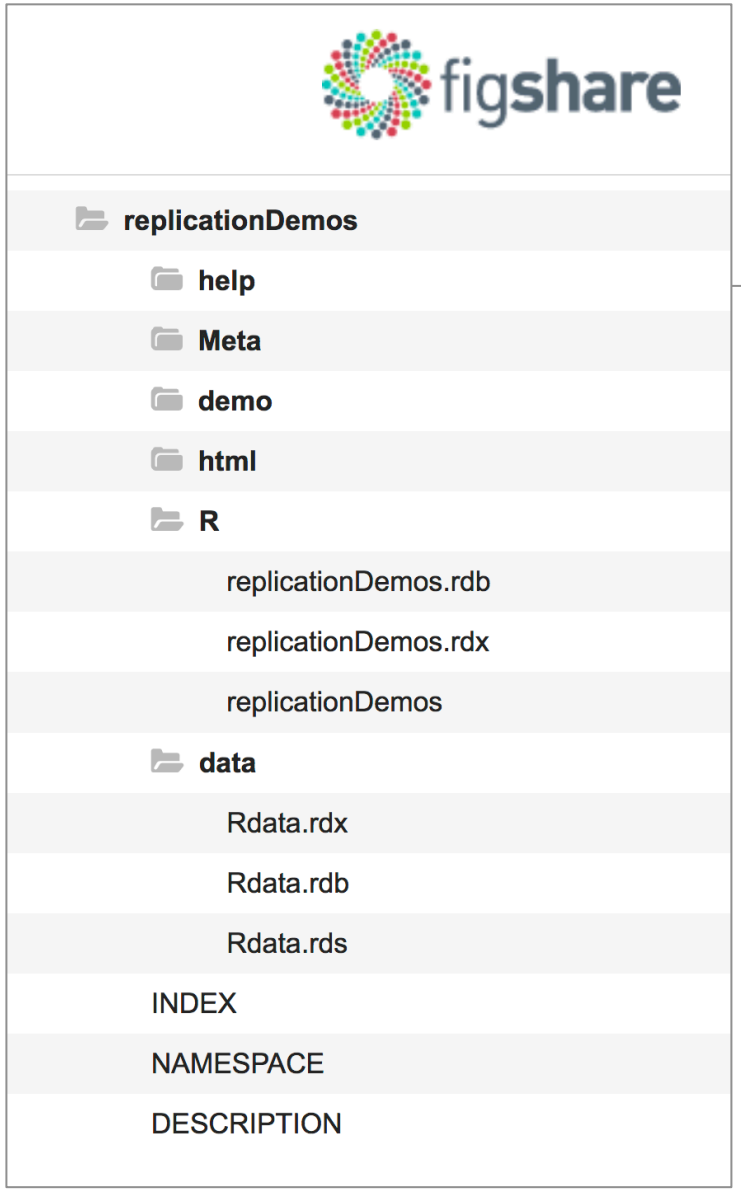
Learning from mistakes in climate research

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



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

V. Stodden. (2013). *Resolving Irreproducibility in Empirical and Computational Research.* IMS Bulletin

Simplifying **Computational Reproducibility** in Whole Tale

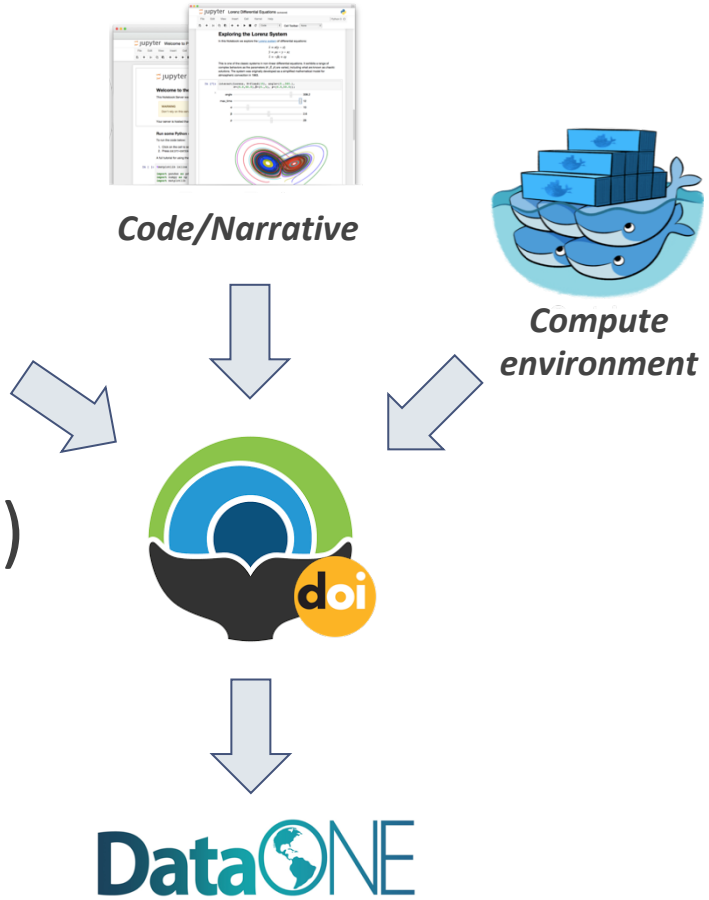
- Researchers can **easily package** and **share *tales***:
  - **Data**, **Code**, and **Compute Environment**
  - to **re-create the computational results** from a scientific study
  - achieving **computational reproducibility**
  - thus **"setting the default to reproducible."**

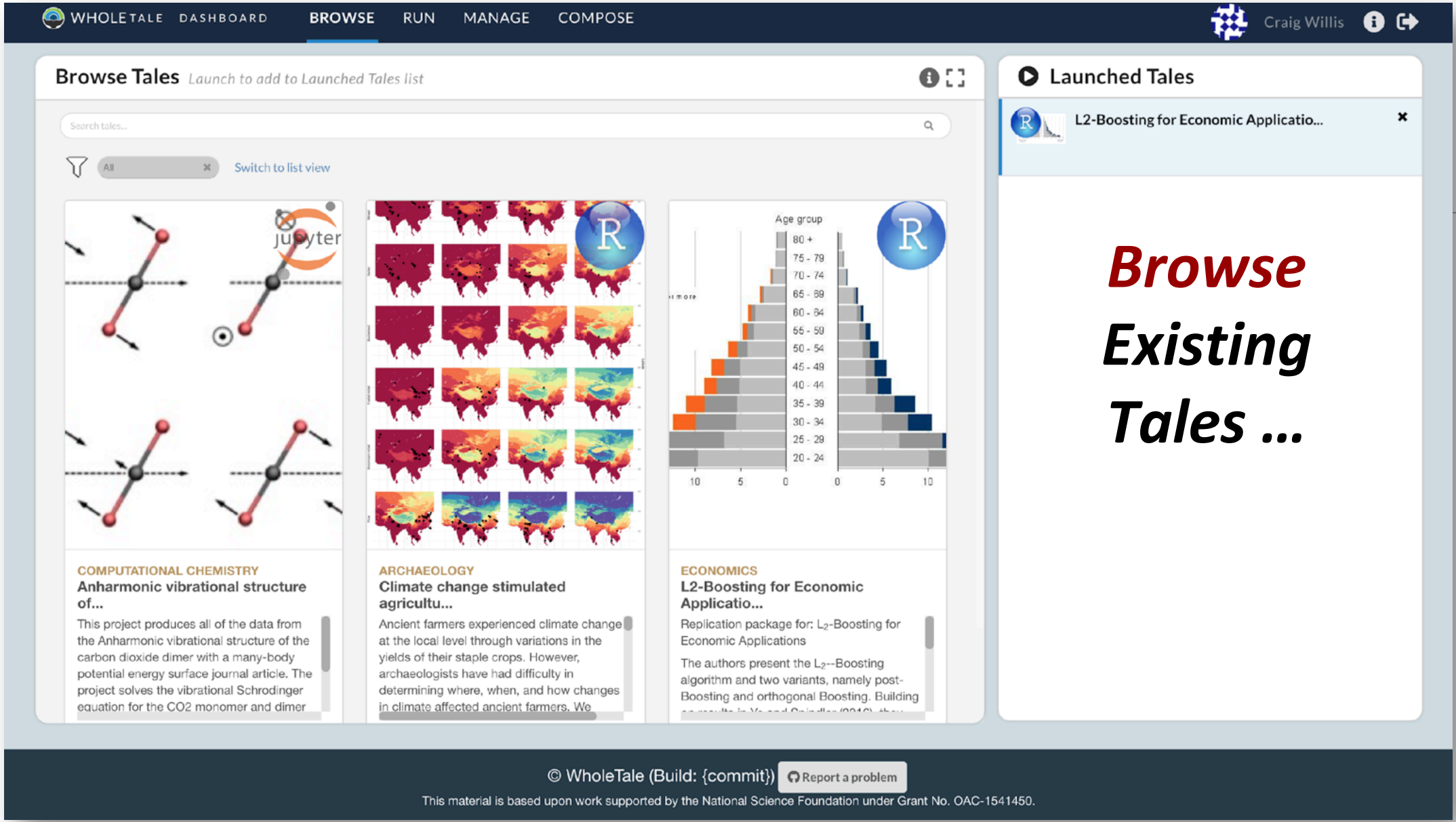
- Also empowers users to verify and extend results with **different** data, methods, and environments.

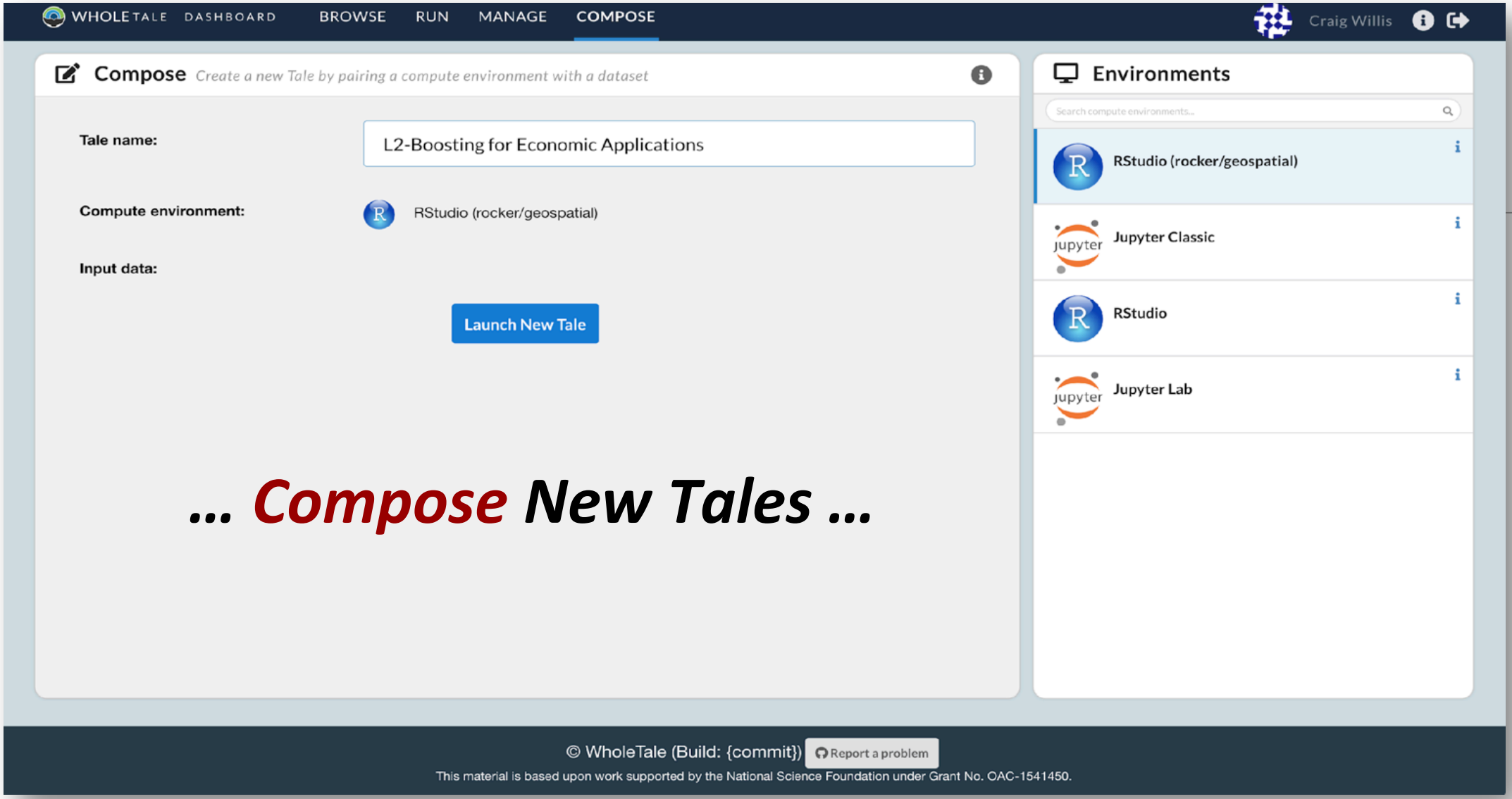
V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. (2013). S*etting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics*, ICERM workshop (2013)

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

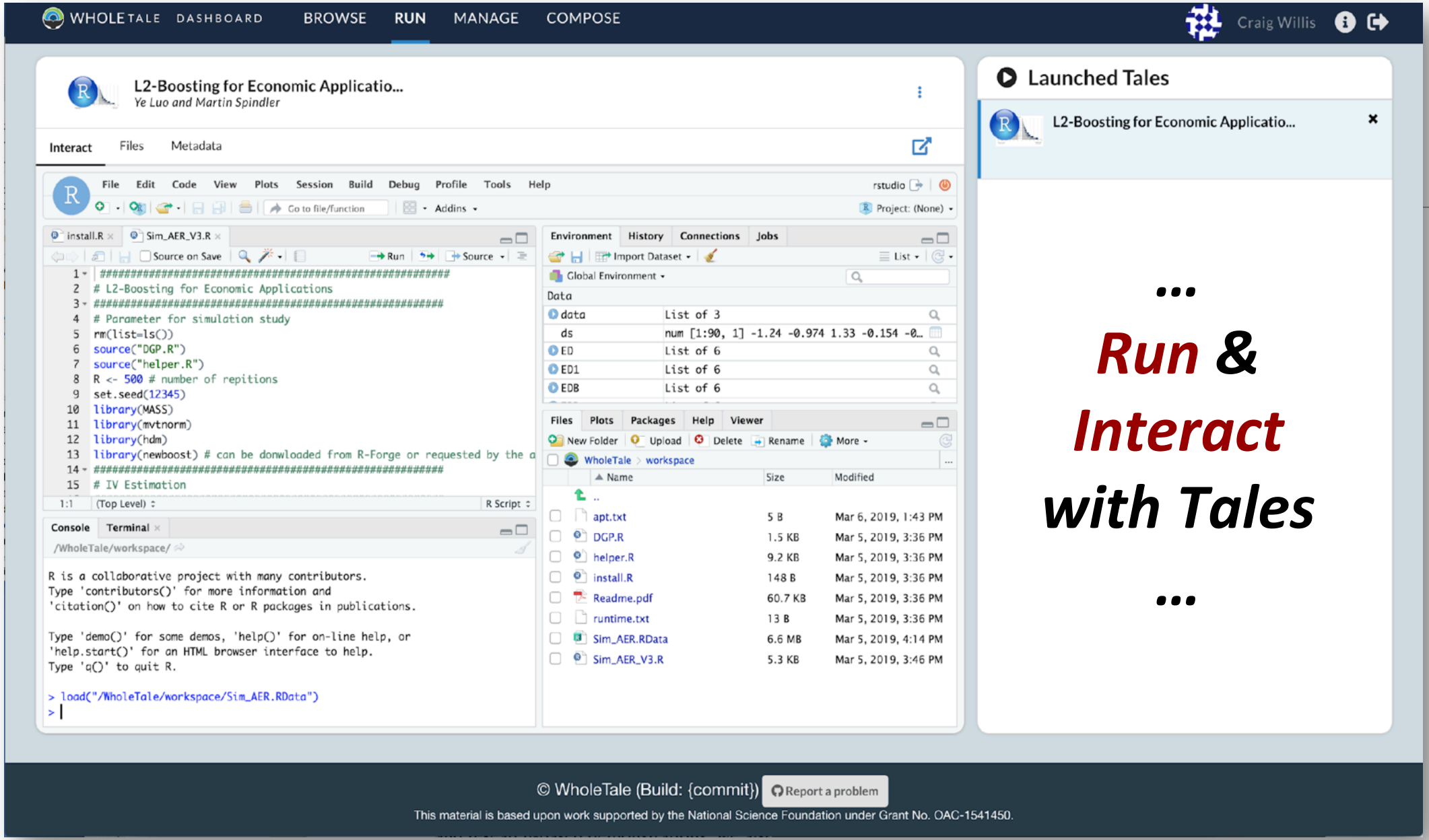
- **Tale** = executable **research object**, i.e.
  - **data** (references)
  - **+ code** (computational methods)
  - **+ narrative** (traditional science story)
  - **+ compute environment** (e.g. RStudio, Jupyter)
- Captured in a **standards-based tale format** complete with metadata

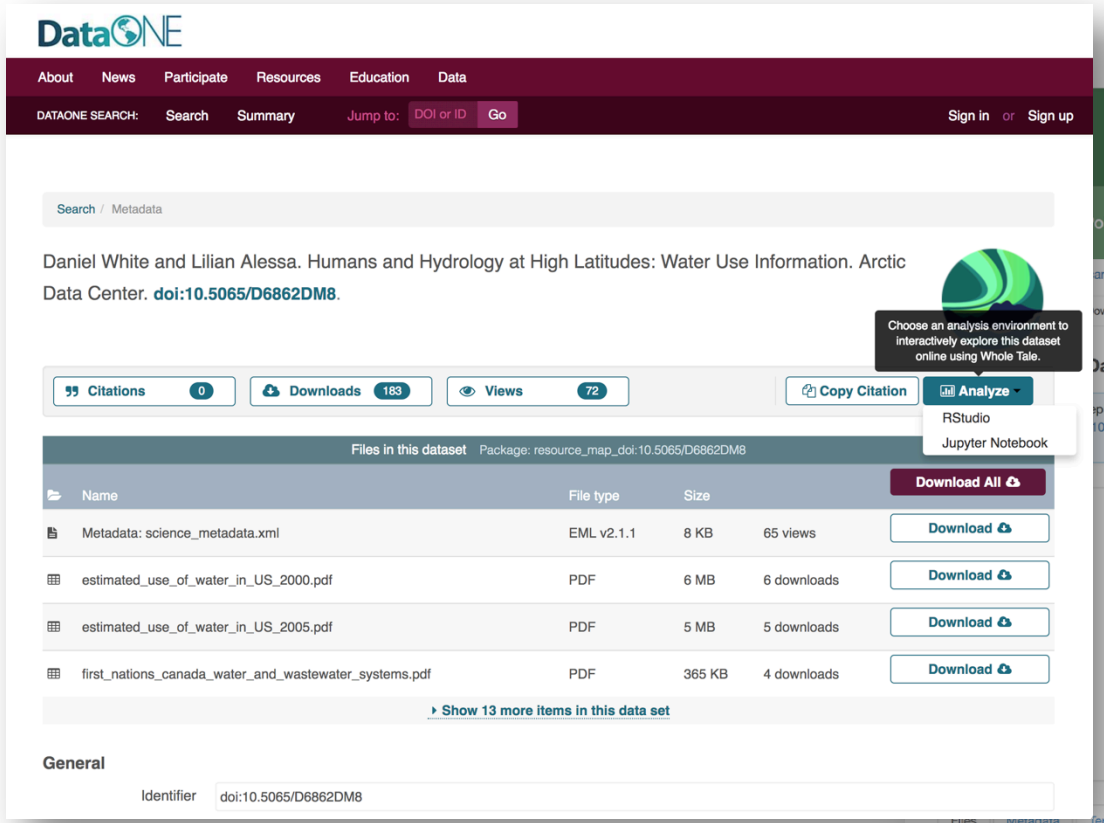


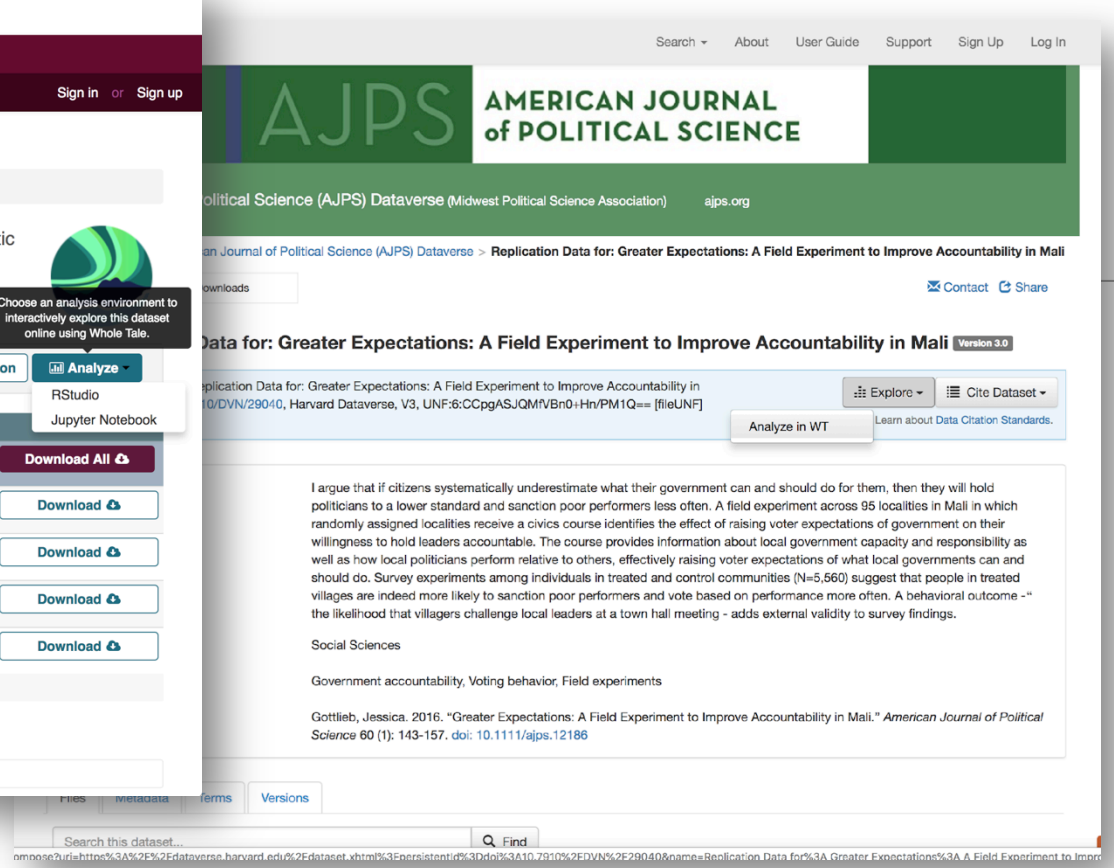

WHOLETALE DASHBOARD
BROWSE
RUN
MANAGE
COMPOSE
Craig Willis
Browse Tales
Launch to add to Launched Tales list
Search tales...
All
Switch to list view
jupyter
Age group
80 +
75 - 79
70 - 74
65 - 69
60 - 64
55 - 59
50 - 54
45 - 49
40 - 44
35 - 39
30 - 34
25 - 29
20 - 24
COMPUTATIONAL CHEMISTRY
Anharmonic 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
Launched Tales
L2-Boosting for Economic Applicatio...
Browse Existing Tales ...
© WholeTale (Build: {commit})
Report a problem
This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450.



… *Compose* *New Tales* …

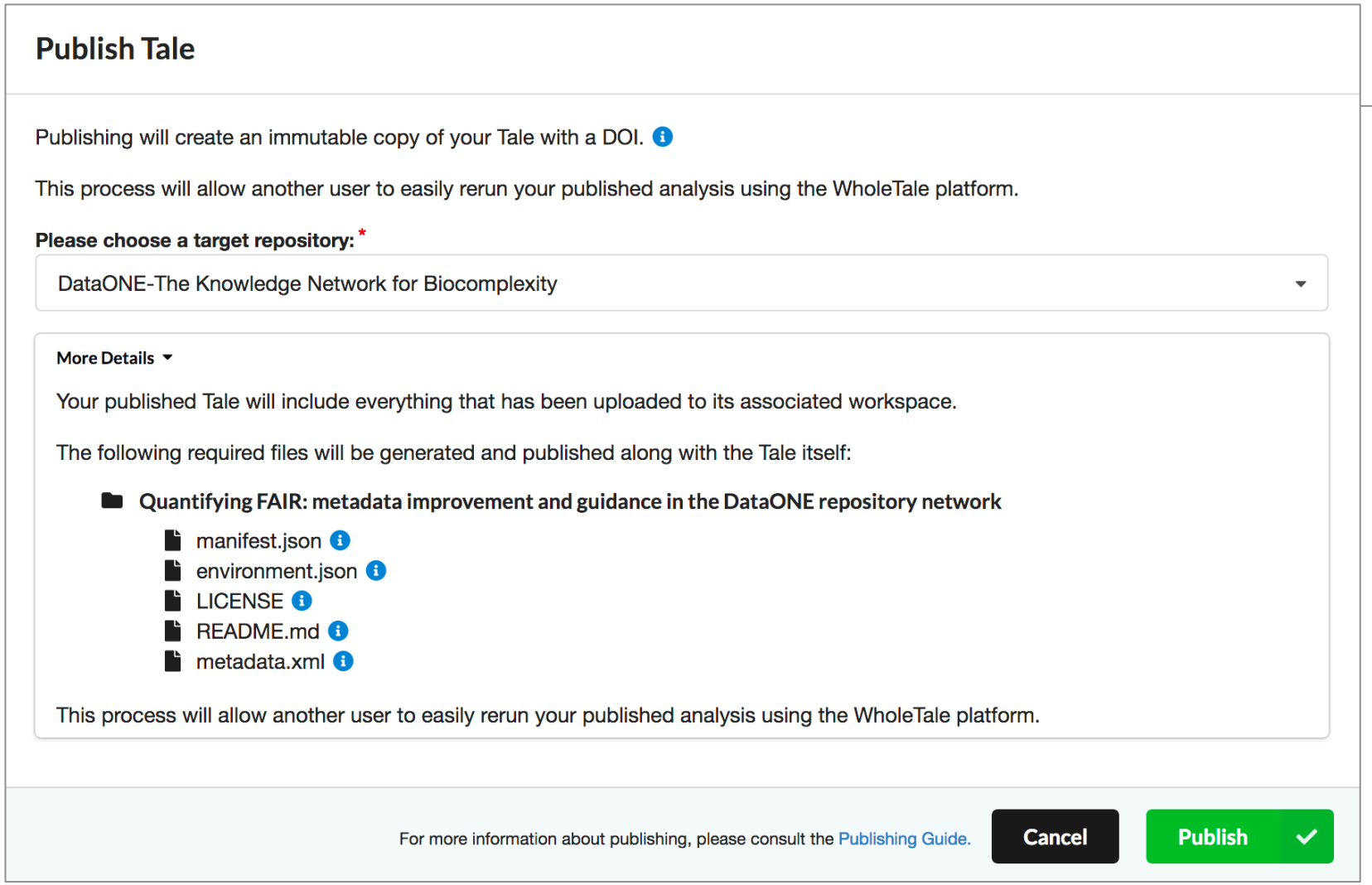

WHOLETALE DASHBOARD
BROWSE
RUN
MANAGE
COMPOSE
Craig Willis
L2-Boosting for Economic Applicatio...
Ye Luo and Martin Spindler
Interact
Files
Metadata
File Edit Code View Plots Session Build Debug Profile Tools Help
rstudio
Go to file/function
Addins
Project: (None)
install.R
Sim_AER_V3.R
Source on Save
Run
Source
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 repitions
9 set.seed(12345)
10 library(MASS)
11 library(mvtnorm)
12 library(hdm)
13 library(newboost) # can be donwloaded from R-Forge or requested by the a
14 ##
15 # IV Estimation
1:1 (Top Level) R Script
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")
>
Environment
History
Connections
Jobs
Import Dataset
List
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
New Folder
Upload
Delete
Rename
More
WholeTale > workspace
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
Launched Tales
L2-Boosting for Economic Applicatio...
...
Run &
Interact
with Tales
...
© WholeTale (Build: {commit})
Report a problem
This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450.





… 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 Data, Code, and Environment

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

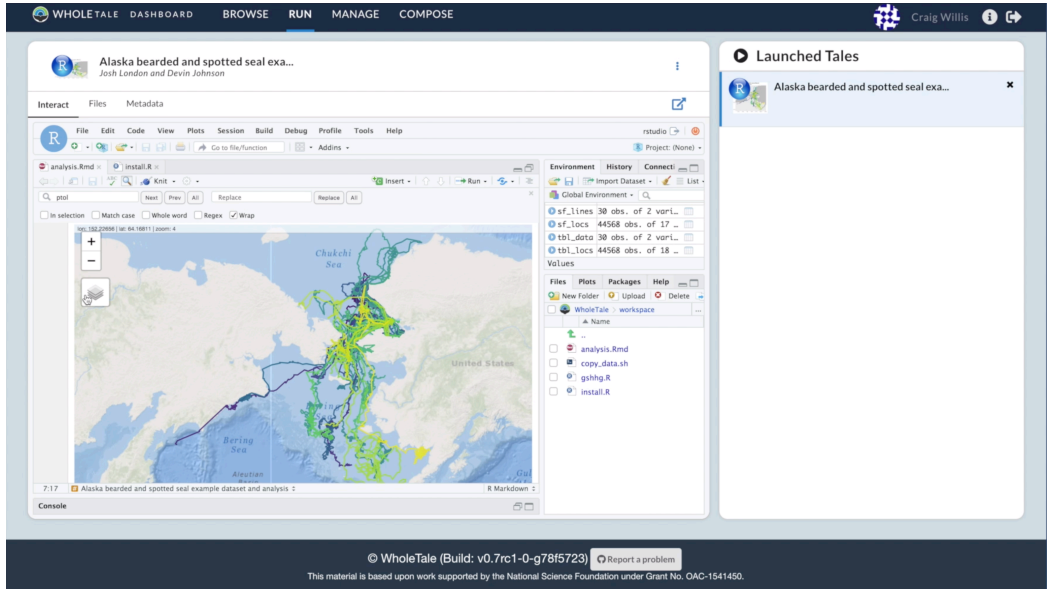
Whole Tale Forecast Demo

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

Based on analysis and models by:

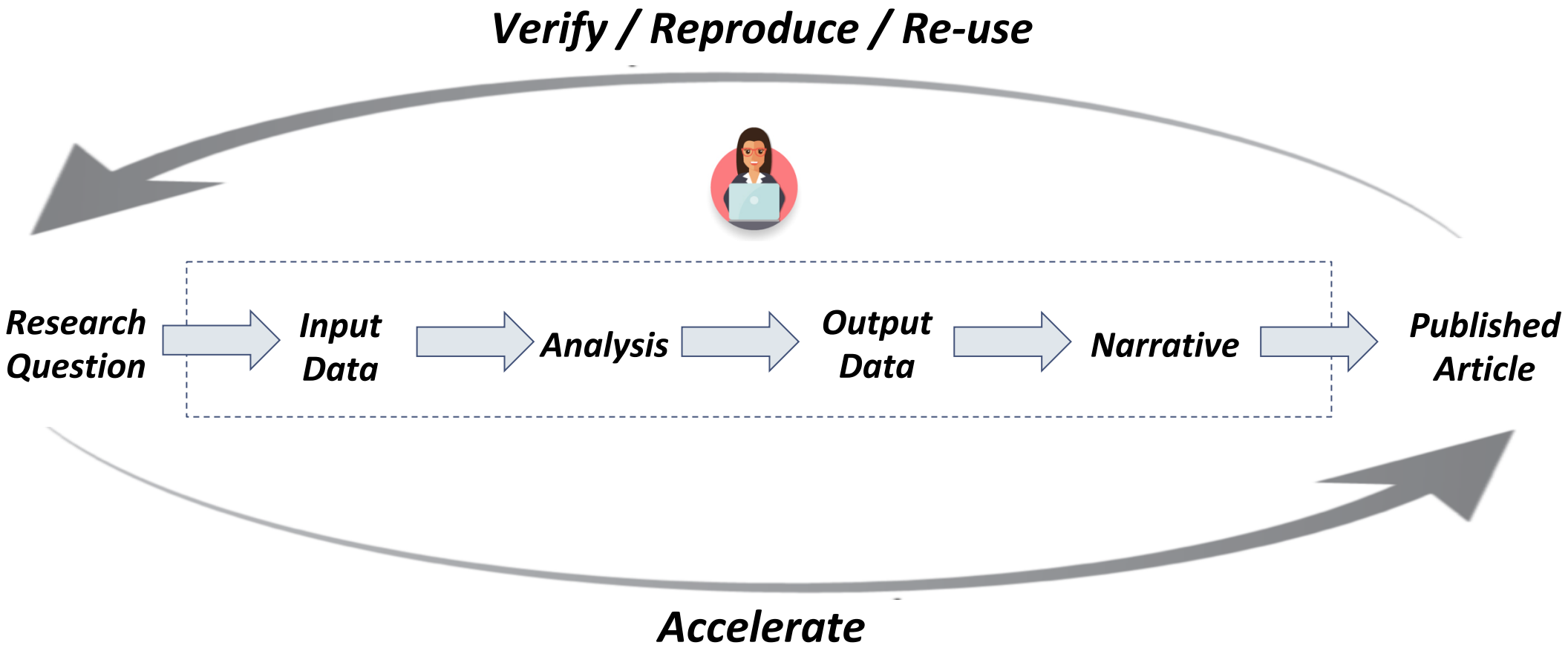
Josh London and Devin Johnson

NOAA Marine Mammal Laboratory



https://youtu.be/Ml5d7r5OtCk

Accelerating Reproducible Open Science



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.