Better science in less time
How open data science tools have improved our science

Julia Stewart Lowndes, PhD
NCEAS, UC Santa Barbara
and our awesome team at NCEAS
and Conservation International

@juliesquid
jules32.github.io
Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes¹*, Benjamin D. Best², Courtney Scarborough¹, Jamie C. Afflerbach¹, Melanie R. Frazier¹, Casey C. O'Hara¹, Ning Jiang¹ and Benjamin S. Halpern¹,³,⁴
Better science is:

- Reproducible
- Collaborative
- Open

Separate things, but deeply interwoven. Your most important collaborator is you!
Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes, Benjamin D. Best, Courtney Scarborough, Jamie C. Afflerbach, Melanie R. Frazier, Casey C. O’Hara, Ning Jiang and Benjamin S. Halpern

Better science is:
- Reproducible
- Collaborative
- Open

Open data science tools:
- Align with data science practices
- Enable better science
- Exist! For your science! (meet scientists where we are)
Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes1*, Benjamin D. Best2, Courtney Scarborough1, Jamie C. Afflerbach1, Melanie R. Frazier1, Casey C. O’Hara1, Ning Jiang1 and Benjamin S. Halpern1,3,4

Lessons learned:

- barriers: exposure & change in mindset
- incorporate incrementally into existing practices (spectrum)
Kurt Vonnegut on the shapes of stories

beginning

ill fortune

good fortune

end

Washington Post 2015; The Atlantic 2016; YouTube
Story arc: ‘boy meets girl’

Beginning → Ill fortune → Good fortune → End

Washington Post 2015; The Atlantic 2016; YouTube
Story arc: The Ocean Health Index
Story arc: The Ocean Health Index

2012
OHI framework & first global assessment published

Halpern et al. 2017, Nature
Story arc: The Ocean Health Index

2012
OHI framework & first global assessment published

2013
second global assessment
Story arc: The Ocean Health Index

beginning

2012
OHI framework & first global assessment published

Good fortune

2013
second global assessment

Today
transparent and repeatable workflow, fifth global assessment, 20+ groups engaged

end

ill fortune

Lowndes et al. 2017 Nature Ecology & Evolution
Halpern et al. 2017, PLOS One
OHI framework
A healthy ocean sustainably delivers a range of benefits to people now and in the future.

Halpern et al. 2012, Nature
Repeatable OHI assessment process

goals

models

scores

inform

planning & data
Scientific process

analyses/models/workflows

core

tailorable
An index to assess the health and benefits of the global ocean

Benjamin S. Halpern¹,², Catherine Longo¹, Darren Hardy¹, Karen L. McLeod³, Jameal F. Samhouri⁴, Steven K. Katona⁵, Kristin Kleisner⁶, Sarah E. Lester⁷,⁸, Jennifer O’Leary¹, Marla Ranelletti¹, Andrew A. Rosenberg⁵, Courtney Scarborough¹, Elizabeth R. Selig⁵, Benjamin D. Best⁹, Daniel R. Brumbaugh¹⁰, F. Stuart Chapin¹¹, Larry B. Crowder¹², Kendra L. Daly¹³, Scott C. Doney¹⁴, Cristiane Elfes¹⁵,¹⁶, Michael J. Fogarty¹⁷, Steven D. Gaines⁸, Kelsey I. Jacobsen⁸, Leah Bunce Karrer⁵, Heather M. Leslie¹⁸, Elizabeth Neeley¹⁹, Daniel Pauly⁶, Stephen Polasky²⁰, Bud Ris²¹, Kevin St Martin²², Gregory S. Stone⁵, U. Rashid Sumaila⁶ & Dirk Zeller⁶
data processing collaboration

data_v1.xls
data_v1b.xls
data_v2_2012_02_26.xls

FWD:Re:data question
coding – reusable record
version control – bookkeeping

collaboration

best practices – now + future
shared – websites, online docs
data processing

- coding – reusable record
- version control – bookkeeping

- best practices – now + future
- shared – websites, online docs

collaboration

Communication. Open data science.
OHI Toolbox

tailorable (GitHub repo)

core (R package)

Lowndes et al. 2015, PeerJ
How we work
How we work

Collaboration
- Record keeping
- File sharing

Communication
- Sharing data, code, methods

Reproducibility
- Code
- Version control

GitHub

R Studio online cloud

local computer
How we work

RStudio – GitHub workflow

GitHub connection, env, build

R code (scripts and console)

File nav, help, plots, packages
How we work

RStudio – GitHub workflow

See what changed line-by-line

...and plot by plot
How we work

Data science theory & tidy data

Import

Tidy

Transform

Visualise

Model

Communicate

Understand

Wickham & Grolemund 2016: R for Data Science
http://r4ds.had.co.nz
How we work
Data science theory & tidy data

In a tidy data set:
- Each **variable** is saved in its own **column**
- Each **observation** is saved in its own **row**

<table>
<thead>
<tr>
<th>country</th>
<th>continent</th>
<th>year</th>
<th>LifeExp</th>
<th>pop</th>
<th>gdpPercap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1952</td>
<td>28.801</td>
<td>8425333</td>
<td>779.4453</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1957</td>
<td>30.352</td>
<td>9240934</td>
<td>820.8530</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1962</td>
<td>31.997</td>
<td>10267083</td>
<td>853.1067</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1967</td>
<td>34.020</td>
<td>11537986</td>
<td>818.1971</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1972</td>
<td>36.068</td>
<td>13079460</td>
<td>739.9811</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1977</td>
<td>38.438</td>
<td>14880372</td>
<td>785.1134</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1982</td>
<td>39.354</td>
<td>12881816</td>
<td>973.0114</td>
</tr>
</tbody>
</table>
How we work
Data science theory & tidy data

In a tidy data set:
- Each variable is saved in its own column
- Each observation is saved in its own row

Subset Variables (Columns)

Subset Observations (Rows)

Group Data
Compute new variables by group.

www.rstudio.com/resources/cheatsheets
How we work
Best community practices for naming & organization

Jenny Bryan: Naming files
machine readable, human readable, plays well with default ordering

RStudio projects as GitHub repos
RMarkdown combines code & text

NO
myabstract.docx
Joe's Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES
2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt
Working openly online - for science and communication: github.com/ohi-science

Ocean Health Index - Science

Global assessment!
Manuscript website!
Training e-book!
OHI+ assessment!
Science website!

Collaborators!
Analyses – R code and text together (R Markdown)

Data Source

Reference: Feely et al.(2009)

Downloaded: March 15, 2016

Description: Aragonite Saturation State $\Omega_{arg}$

Native data resolution: 1 degree cells

Time range: 1880-1889 and 2005-2015, monthly data provided for each year

Format: NetCDF

Notes about the data:

This data was shared with us by Ivan Lima from Woods Hole Institute for Oceanography in December 2014 and again February 2016. The data came as NetCDFs in an irregular grid format with a resolution of about 1 degree. The data values are monthly average surface $\Omega$ aragonite saturation state.

Methods

Setup

The main R libraries needed for this analysis are the `raster` and `ncdf4` packages.

```r
#set options for all chunks in code
knitr::opts_chunk$set(warning=FALSE, message=FALSE, fig.width=6, fig.height=6)
source("~/github/chiprep/src/R/common.R")
```
Ocean Health Index
open data science tools and resources for marine science and management
Interactive websites for published articles:
http://ohi-science.nceas.ucsb.edu/plos_change_in_global_ocean_health

Drivers and implications of change in global ocean health over the past five years

Benjamin S. Halpern1,2,3,*, Melanie Frazier1, Jamie Affierbach1, Casey O’Hara2, Steven Katona1,4, Julia S. Stewart Lowndes1, Ning Jiang1, Erich Pacheco2, Courtney Scarborough1, Johannra Posenberg3

1. National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, California, USA
2. Bren School of Environmental Science & Management, University of California, Santa Barbara, California, USA
3. Imperial College London, Blackwood Park Campus, Ascot, United Kingdom

Drivers and implications of change in global ocean health over the past five years. PLoS ONE 12(7): e0176207, doi:10.1371/journal.pone.0176207

The Ocean Health Index

A healthy ocean sustainably delivers a range of benefits to people now and in the future. The Ocean Health Index is the comprehensive framework used to quantify ocean-derived benefits to humans and to help inform sustainable ocean management using the best available information. Assessments using the OHI framework require synthesizing existing data representing those benefits, using methods that are reproducible and repeatable. Repeated assessments using the same methods enable quantifiable comparison of changes in ocean health through time, which can be used to inform policy and track progress.

Visit (http://www.ohi-science.org/) for more about the science and methods behind the Ocean Health Index, or (http://www.oceanhealthindex.org/) for an overview of the Ocean Health Index project.

Abstract of published paper

Drivers and implications of change in global ocean health over the past five years

Growing international and national focus on quantitatively measuring and improving ocean health has increased the need for comprehensive, scientific, and repeated indicators to track progress towards achieving policy and societal goals. The Ocean Health Index (OHI) is one of the few indicators available for this purpose. Here we present results from five years of annual global assessment for 220 countries and territories, evaluating potential drivers and consequences of changes and presenting lessons learned about the challenges of using composite indicators to measure sustainability goals.

Globally scores have shown little change, as would be expected. However, individual countries have seen notable increases or declines due in particular to improvements in the harvest and management of wild-caught fisheries, the creation of marine protected areas (MPAs), and decreases in natural product harvest. Rapid loss of sea ice and the consequent reduction of coastal protection from that sea ice was also responsible for declines in overall ocean health in many Arctic and sub-Arctic countries. The OHI performed reasonably well at predicting near-term future scores for many of the ten goals measured, but data gaps and limitations hindered these predictions for many other goals.

Ultimately, all indicators face the substantial challenge of informing policy for progress toward broad goals and objectives with insufficient monitoring and assessment data. If countries and the global community hope to achieve and maintain healthy oceans, we will need to dedicate significant resources to measuring what we are trying to manage.
The OHI and open data science

Global assessments
science to global policy

OHI+
science to local policy

Toolbox software
R packages • architecture

Websites
ohi-science.org
oceanhealthindex.org

RStudio
GitHub

git
With open data science, repeating analyses takes less time

Lowndes et al. 2017

OHI Toolbox

Ease of Collaboration (including future self)

Ease of Reproducibility

Reasonably easily

With great difficulty

Less

Greater

Time required

- less
- more

Focus

- science
- data science

2012
2013
2014
2015
2016
2017 (upcoming)
So what can you do?
Resources

So many awesome resources...
academic pubs, webinars, books, blogs, trainings, tutorials, podcasts, etc

A few lists we’re trying to keep updated:
   http://ohi-science.org/betterscienceinlesstime

1. Specific resources that helped us learn
e.g. R for Data Science, RStudio webinars & cheatsheets, Software Carpentry

2. Academic literature + media on the importance of open data science for science
e.g. How open science helps researchers succeed (McKiernan 2016),
   Git can facilitate greater reproducibility and increased transparency in science (Ram 2013)
Learn in an intentional way

Use online resources for self-paced learning
books, tutorials, slide decks, archived webinars, presentations, blogs & more
Learn in an intentional way

Use online resources for self-paced learning
books, tutorials, slide decks, archived webinars, presentations, blogs & more

Learn together by joining and/or creating local communities
academic articles or webinars during lab meetings, journal clubs, or happy hours. PIs can encourage this culture even if they don’t code themselves
Learn in an intentional way

Use online resources for self-paced learning
books, tutorials, slide decks, archived webinars, presentations, blogs & more

Learn together by joining and/or creating local communities
academic articles or webinars during lab meetings, journal clubs, or happy hours.
PIs can encourage this culture even if they don’t code themselves

Ask for help
expect that someone has already had this question/this error (paste errors into Google!)
Learn in an intentional way

Use online resources for self-paced learning
books, tutorials, slide decks, archived webinars, presentations, blogs & more

Learn together by joining and/or creating local communities
academic articles or webinars during lab meetings, journal clubs, or happy hours. PIs can encourage this culture even if they don’t code themselves

Ask for help
expect that someone has already had this question/this error (paste errors into Google!)

Attend in-person workshops and conferences
workshops to learn skills, conferences to see examples “in the wild”
Learn in an intentional way

Use online resources for self-paced learning
books, tutorials, slide decks, archived webinars, presentations, blogs & more

Learn together by joining and/or creating local communities
academic articles or webinars during lab meetings, journal clubs, or happy hours.
PIs can encourage this culture even if they don’t code themselves

Ask for help
expect that someone has already had this question/this error (paste errors into Google!)

Attend in-person workshops and conferences
workshops to learn skills, conferences to see examples “in the wild”

Engage online
see what’s new, how people are working:
#rstats on Twitter, rOpenSci, RStudio, Software Carpentry, etc
Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes1*, Benjamin D. Best2, Courtney Scarborough1, Jamie C. Afflerbach1, Melanie R. Frazier1, Casey C. O’Hara1, Ning Jiang1 and Benjamin S. Halpern1,3,4

Lessons learned:

• barriers: exposure & change in mindset
• incorporate incrementally into existing practices (spectrum)
Better science in less time
How open data science tools have improved our science

Thank you

ohi-science.org // @OHIscience
oceanhealthindex.org // @OceanHealthIndx

Julia Stewart Lowndes
lowndes@nceas.ucsb.edu
twitter: @juliesquid

DataONE Webinar
October 9, 2017
Some of the most useful advice

Essential

Googling the Error Message

O RLY?

The Practical Developer
@ThePracticalDev

Essential

Changing Stuff and Seeing What Happens

O RLY?

How to actually learn any new programming concept

The Practical Developer
@ThePracticalDev
How we work
Best community practices for naming & organization

Jenny Bryan: Naming files
- Machine readable, human readable, plays well with default ordering

<table>
<thead>
<tr>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>myabstract.docx</td>
</tr>
<tr>
<td>Joe's Fnames Use Spaces and Punctuation.xlsx</td>
</tr>
<tr>
<td>figure 1.png</td>
</tr>
<tr>
<td>fig 2.png</td>
</tr>
<tr>
<td>JW7d^2sl@deletethisandyourcareerisoverWx2*.txt</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-06-08_abstract-for-sla.docx</td>
</tr>
<tr>
<td>joes-filenames-are-getting-better.xlsx</td>
</tr>
<tr>
<td>fig01_scatterplot-talk-length-vs-interest.png</td>
</tr>
<tr>
<td>fig02_histogram-talk-attendance.png</td>
</tr>
</tbody>
</table>

RStudio projects as GitHub repos