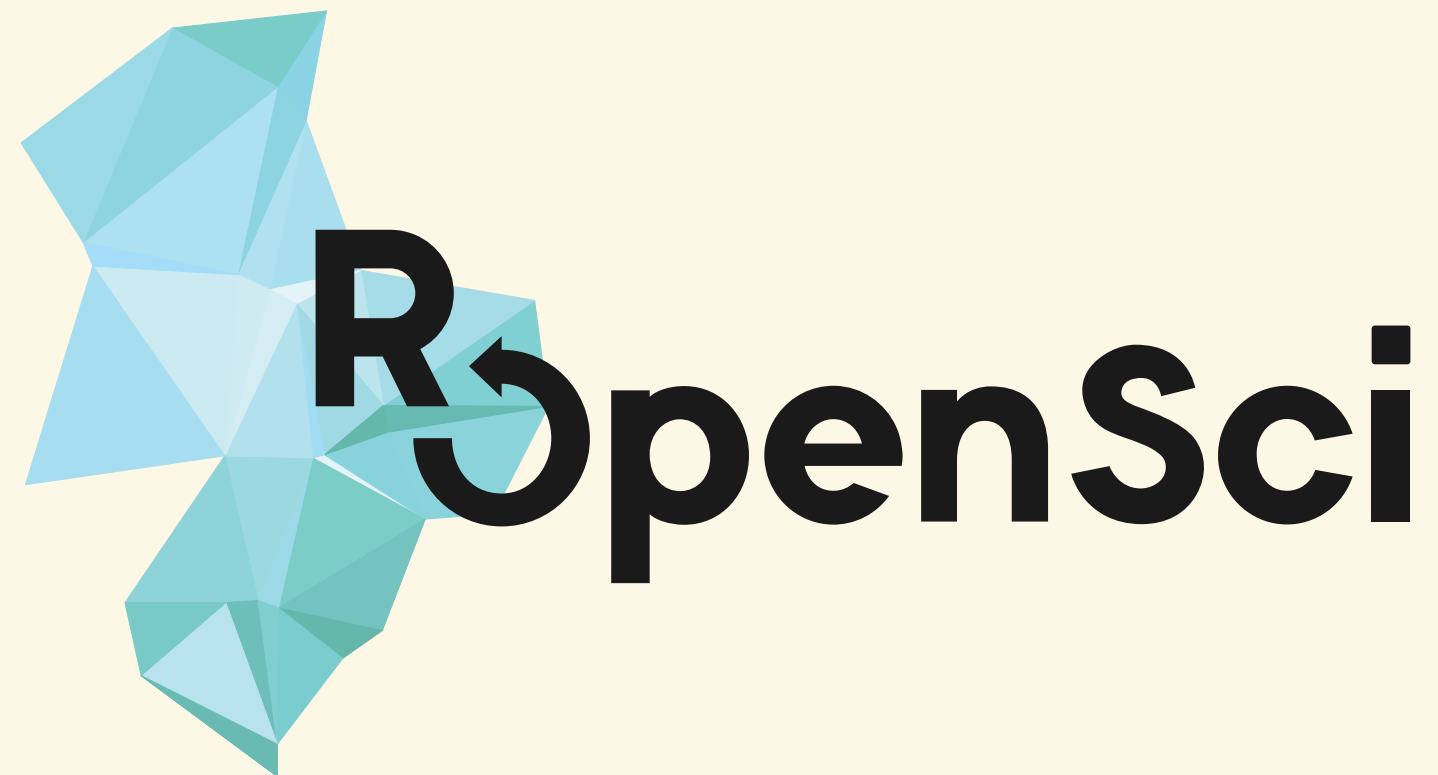


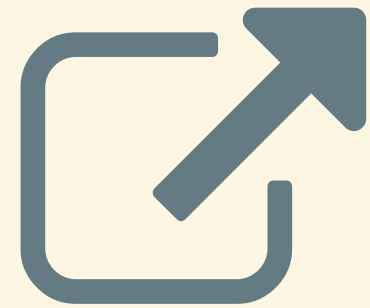
Open science and R

Scott Chamberlain ([@sckott](#)/[@ropensci](#))

UC Berkeley / rOpenSci



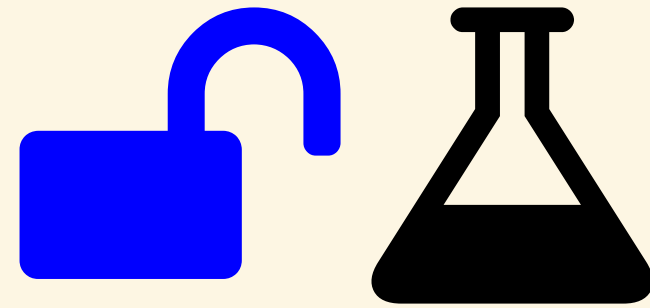
THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST



scotttalks.info/ossps

LICENSE: CC-BY 4.0

open science



open science is badly
needed

Retractions



Duke University is at the center of a whistleblower lawsuit concerning potential research misconduct.

Uschools University
Images/iStockphoto

Whistleblower sues Duke, claims doctored data helped win \$200 million in grants

By **Alison McCook**, **Retraction Watch** | Sep. 1, 2016 , 2:00 PM

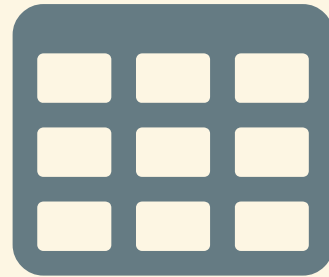


science should be
reproducible!

but doing for real is another issue

Emergent findings

e.g., data



Open science as a lego set



Open science as a lego set

open science may be hard to do

but - you can work on different
components

and - individual components are useful on
their own

Open Data

make your data open

funders/journals often requiring this
anyway

future self will thank you

Open Access

make your papers open

funders often requiring this anyway

talk to your librarians!

Versioning: code/data/text

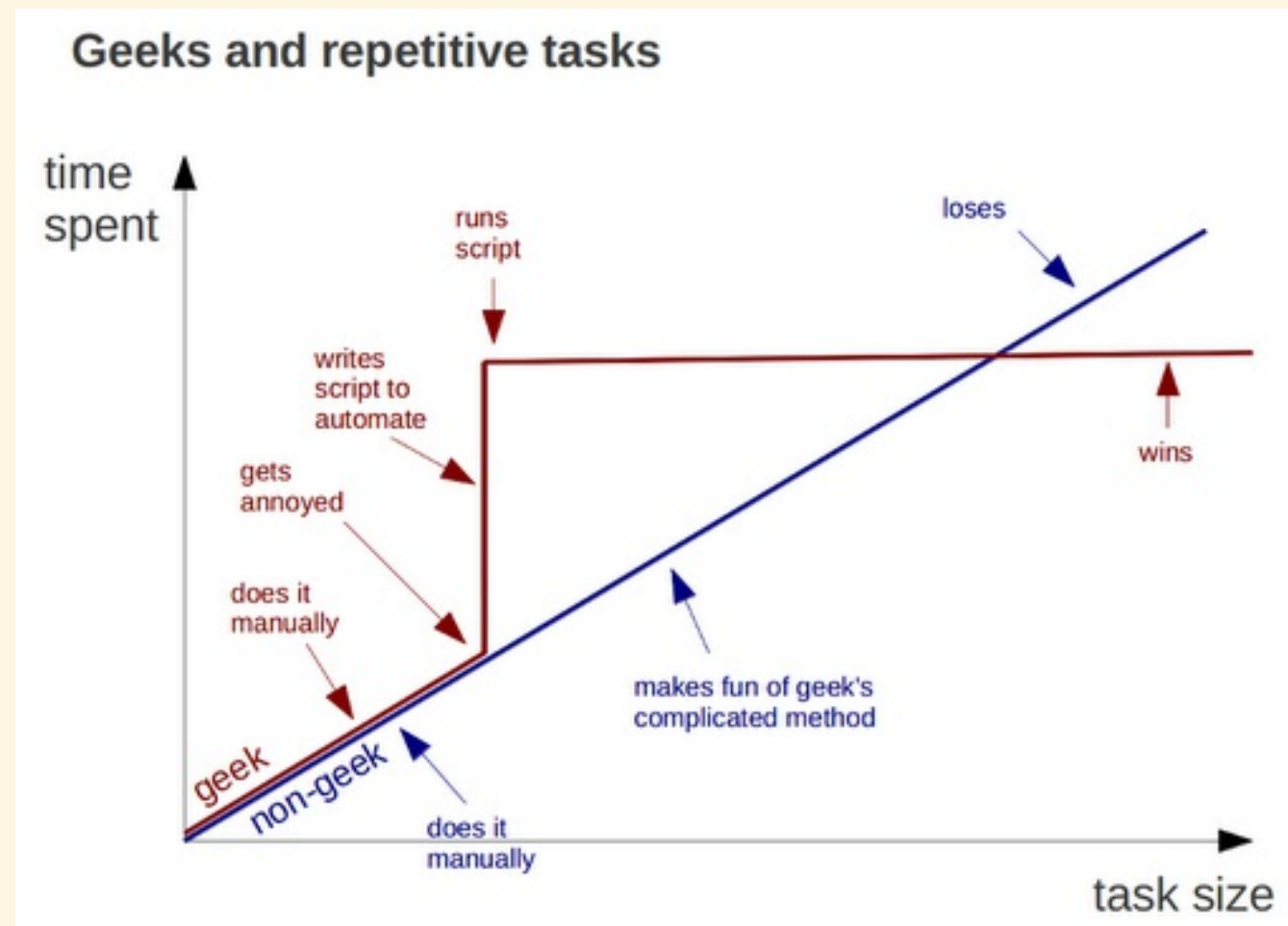


Versioning:
code/data/text

failure proofs your work

experiment freely!

Do all work programmatically



from geeksaresexy.net/2012/01/05/geeks-vs-non-geeks-picture

Do all work programmatically

Key to reproducibility

Most important person that wants to
reproduce your work is you!

Do all work programmatically

you and yourself

- one week from now
- two months from now
- & so on

Wellcome Trust

Towards Open Research

Practices, experiences, barriers and
opportunities

October 2016

Veerle Van den Eynden, Gareth Knight, Anca Vlad, Barry Radler, Carol Tenopir,
David Leon, Frank Manista, Jimmy Whitworth and Louise Corti

N=583 (N=259 ESRC)

[link](#) 

Wellcome Trust: Open Access

OA part of open science held back by impact factors

“As much as I love the idea, my long term career prospects currently depend on obtaining high impact papers, so fully Open Access journals have to be of comparable merit.”

Wellcome Trust: Open Data

"The majority of respondents make datasets available as open access (80%), 19% make data available upon request via an application procedure, 10% restrict access to immediate collaborators and 9% restrict access to registered users."

No!!!

Wellcome Trust: Open Code

"only 12% ... indicated they had a bad experience when sharing code ... BUT the majority of ESRC-funded respondents did not recognise any personal benefits from code sharing activities"

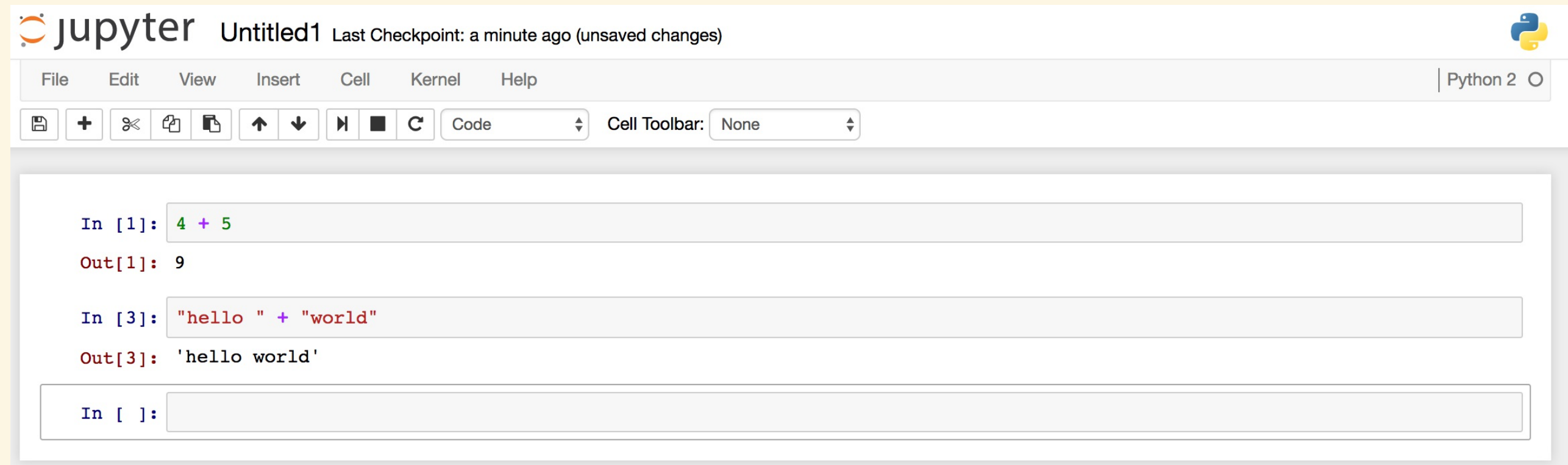
scientific programming languages

are:
the canvas on which to do science

important scientific programming languages



Jupyter Notebooks



reproducing a Jupyter notebook

```
{
  "cells": [
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "## **Reproducible spatial analysis with ArcPy and R using Jupyter Notebook**"
      ]
    },
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "In this example I'm going to crop a large image with a polygon, run a majority"
      ]
    },
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "### Let's start working with R"
      ]
    },
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "I found that cropping an image with R is **much simpler** than doing it with s"
      ]
    },
    {
      "cell_type": "code",
      "metadata": {},
      "source": [
        "# Import arcpy module"
      ]
    }
  ]
}
```

reproducing a Jupyter notebook

[extras](#) / [2016-06-29-reproducibility-arcpy-jupyter-notebook-r](#) / Reproducible spatial analyses with ArcPy and R.ipynb



amsantac renamed notebook for reproducibility r post

4b5f32e on Jun 30, 2016

1 contributor

1.14 MB

No coverage

Download

History



Reproducible spatial analysis with ArcPy and R using Jupyter Notebook

In this example I'm going to crop a large image with a polygon, run a majority filter and then compare frequency of cell values between the cropped image and the filtered image.

Let's start working with R

I found that cropping an image with R is **much simpler** than doing it with some other GIS software programs. Let's define the working directory and load the required package:

```
In [1]: setwd("C:/Users/Public/Documents/amsantac/data")  
library(raster)
```

Loading required package: sp

Let's import the files into R:

something similar in R: Rmarkdown

The screenshot displays the RStudio interface with an R Markdown notebook open. The notebook is titled "Viridis Notebook" and contains the following code:

```
1 ---
2 title: "Viridis Notebook"
3 output: html_notebook
4 ---
5
6 ```{r include = FALSE}
7 library(viridis)
8 ```
9
10 The code below demonstrates two color palettes in the
11 [viridis](https://github.com/sjmgarnier/viridis) package. Each
12 plot displays a contour map of the Maunga Whau volcano in
13 Auckland, New Zealand.
14
15 ## Viridis colors
16
17 ```{r}
18 image(volcano, col = viridis(200))
19 ```
```

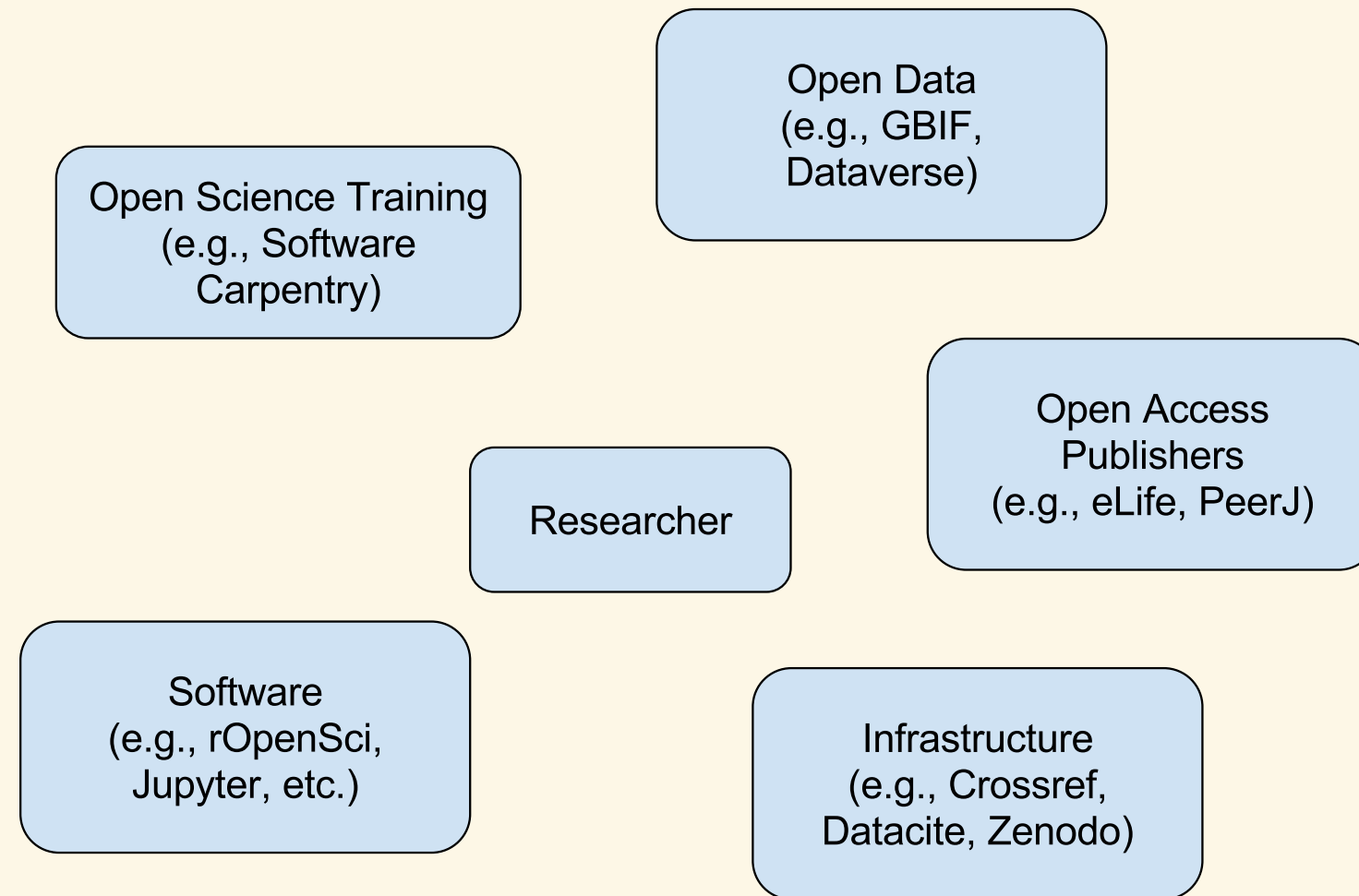
The notebook shows two heatmaps of the Maunga Whau volcano. The first heatmap, titled "Viridis colors", uses the viridis color palette. The second heatmap, titled "Magma colors", uses the magma color palette. Both heatmaps show a contour map of the volcano with a color gradient from dark purple to yellow.

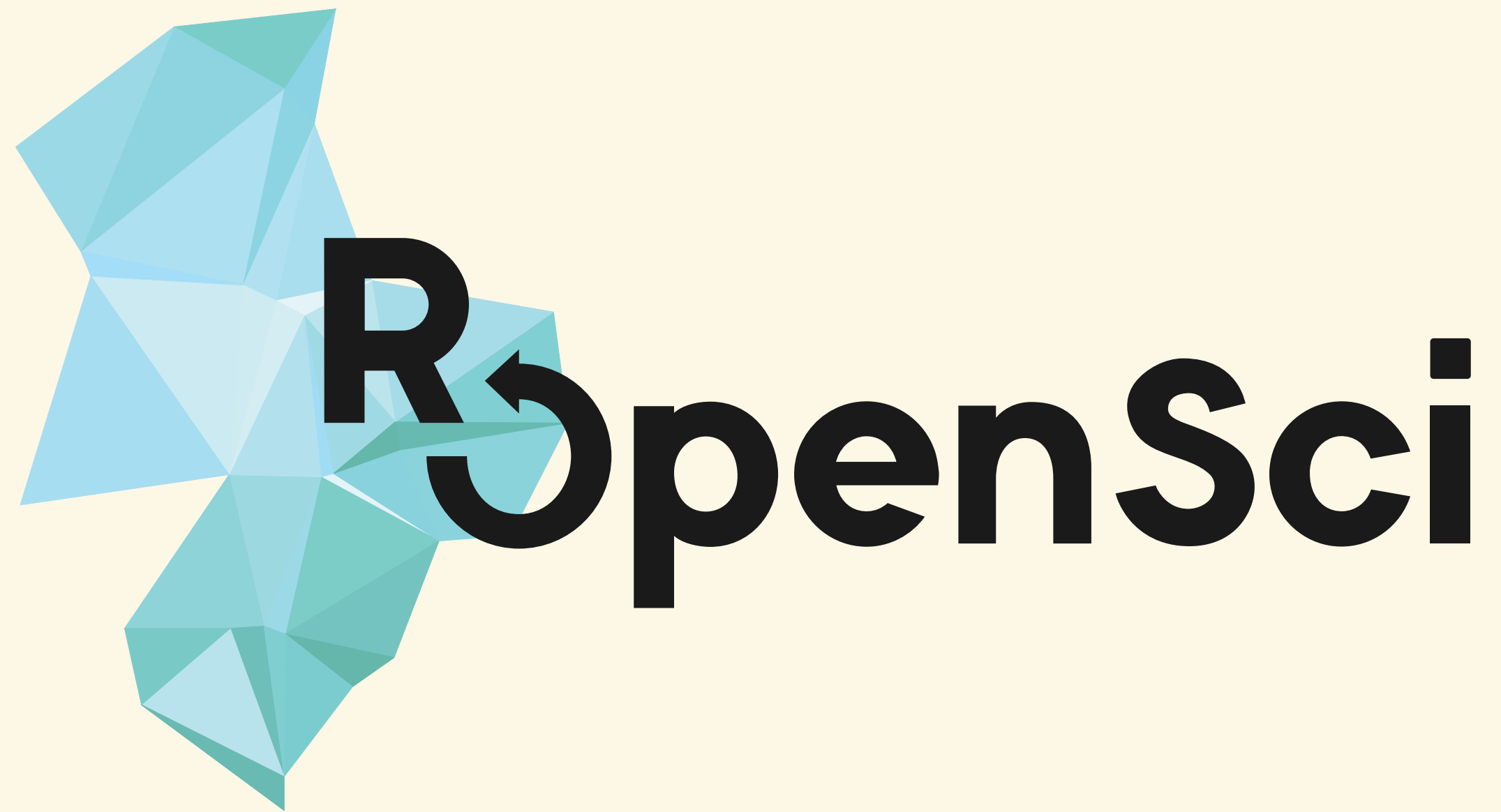
The RStudio interface includes a top toolbar with icons for file operations, a left sidebar with a file explorer, and a right sidebar with tabs for Environment, History, Build, Git, Files, Plots, Packages, Help, and Viewer. The Viewer tab is active, showing the rendered HTML output of the notebook. The console at the bottom shows the execution of the R code.

R language

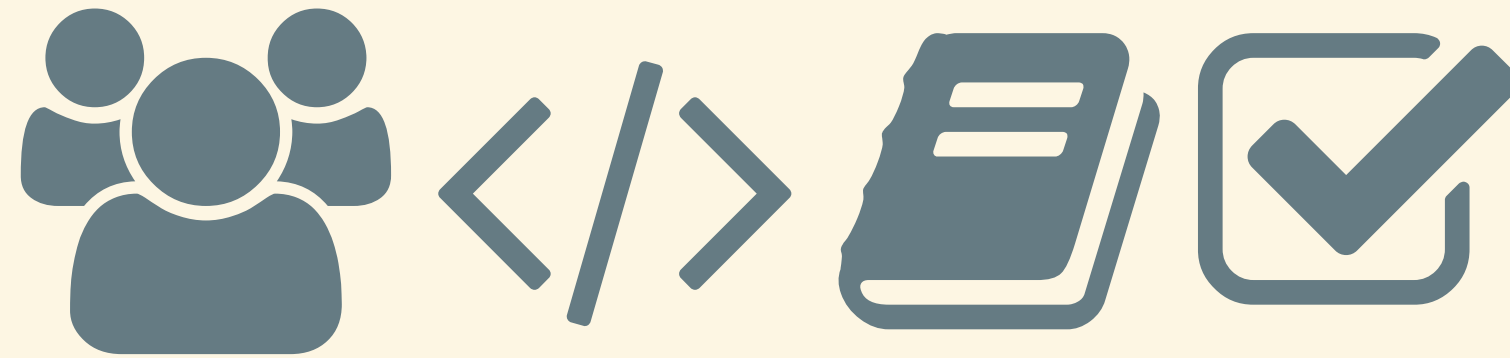
- used widely in biology, psychology, medicine, etc.
- rapidly growing user base, companies surrounding it
- includes all tools for open science workflow
- though work to be done ...

Open science ecosystem





rOpenSci does:



rOpenSci staff

ropensci.org/about/#staff

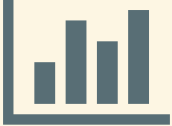
- 4 full time
- now including a community manager!
- leadership team
- advisory board

rOpenSci stats

- ~ 250 code contributors
- ~ 343 Github repositories
- ~ 30,000 commits
- ~ 117 published R packages

the research workflow

Data acquisition  +

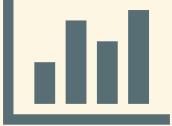
data manipulation/analysis/viz  +

writing  +

publish 

the research workflow

Data acquisition  +

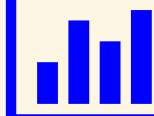
data manipulation/analysis/viz  +

writing  +

publish 

the research workflow

Data acquisition  +

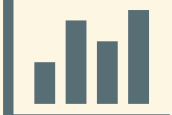
data manipulation/analysis/viz  +

writing  +

publish 

the research workflow

Data acquisition  +

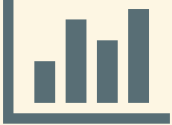
data manipulation/analysis/viz  +

writing  +

publish 

the research workflow

Data acquisition  +

data manipulation/analysis/viz  +

writing  +

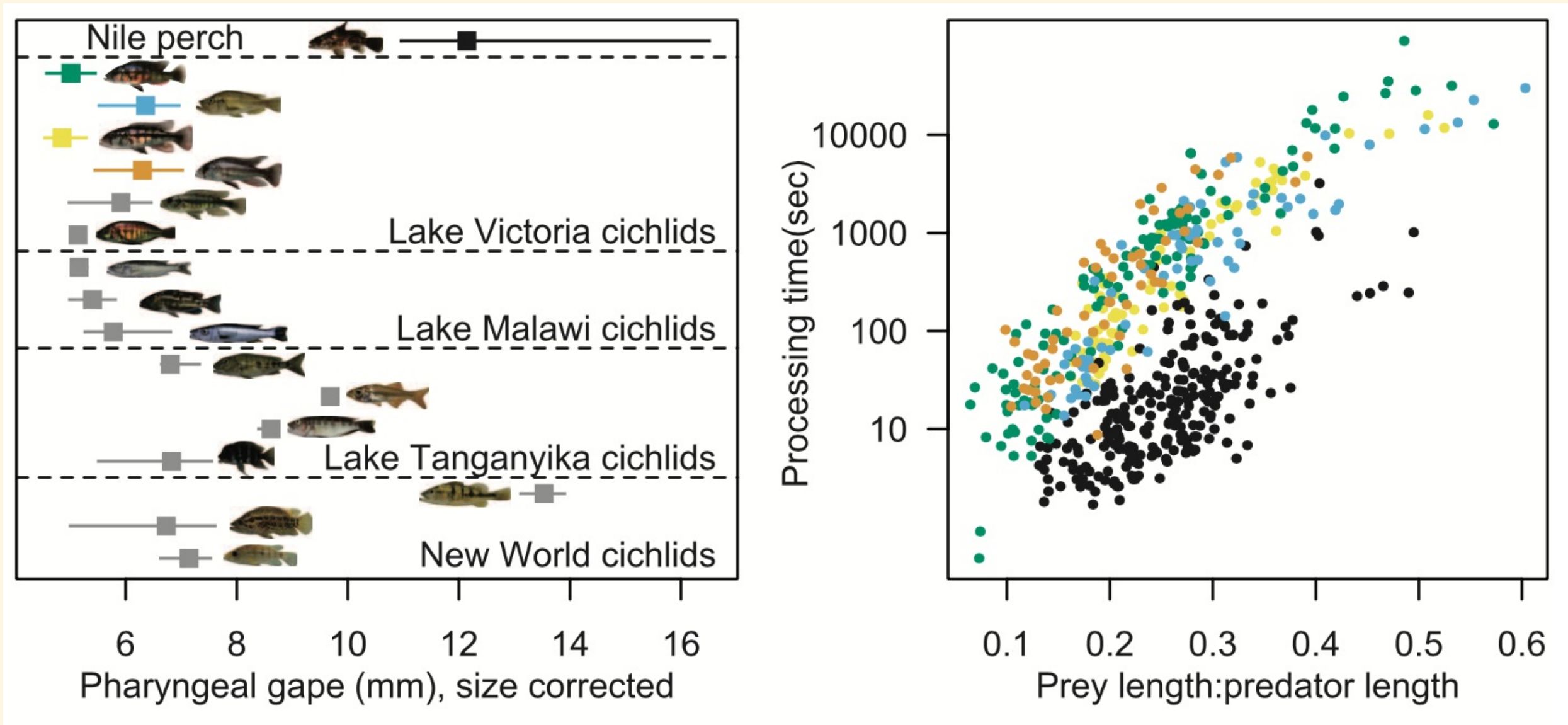
publish 

We make data driven
stories easier to tell

here are some stories ...

use case 1

McGee, et al. (2015). A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. Science [↗](#)





ropensci / rfishbase

<> Code

! Issues 16

🔗 Pull requests 1

📁 Projects 0

📖 Wik

R interface to the fishbase.org database <http://ropensci.org> — Edit

🔄 408 commits

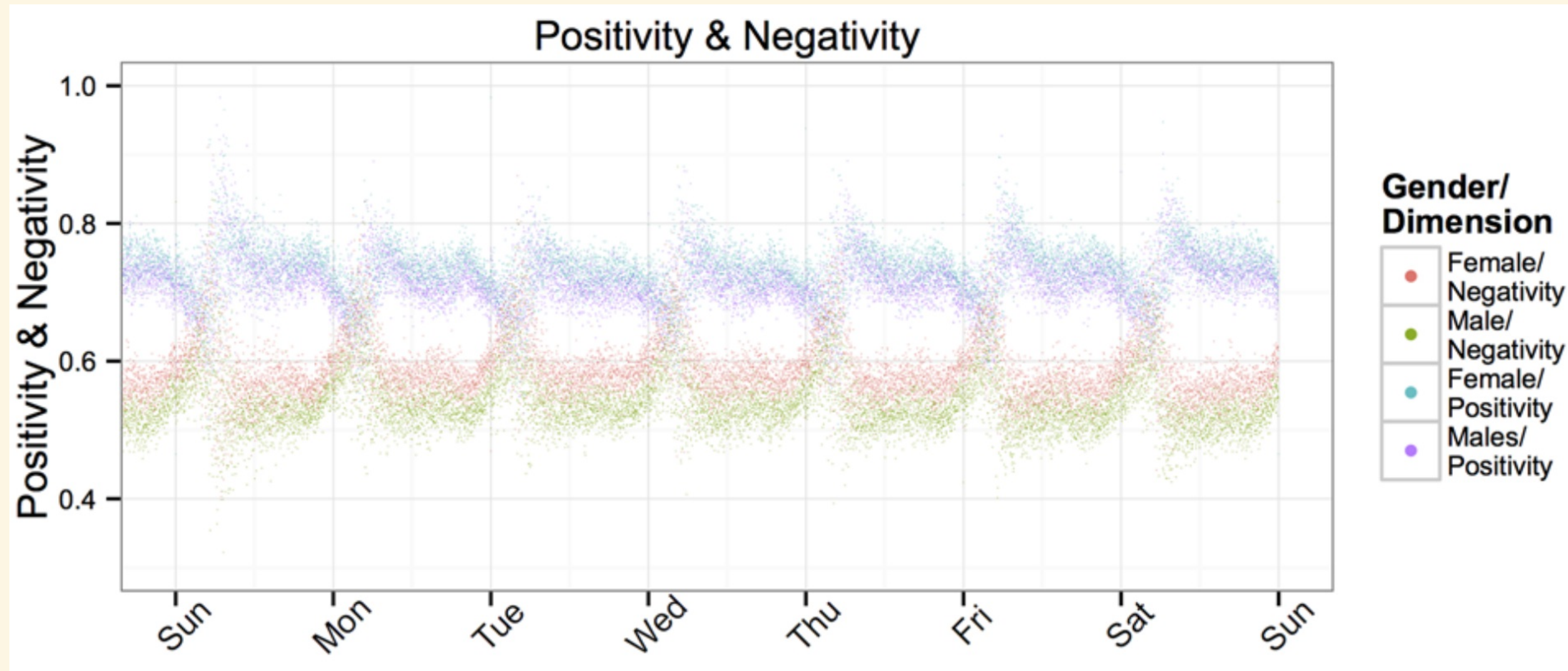
🔗 6 branches


Branch: master ▾

New pull request

use case 2

Serfass, D. G., & Sherman, R. A. (2015). Situations in 140 Characters: Assessing Real-World Situations on Twitter. PLoS ONE [↗](#)



 ropensci / gender

<> Code

! Issues 4

🔗 Pull requests 0

📁 Projects 0

Predict Gender from Names Using Historical Data — Edit

🔄 304 commits

🔗 2 branches

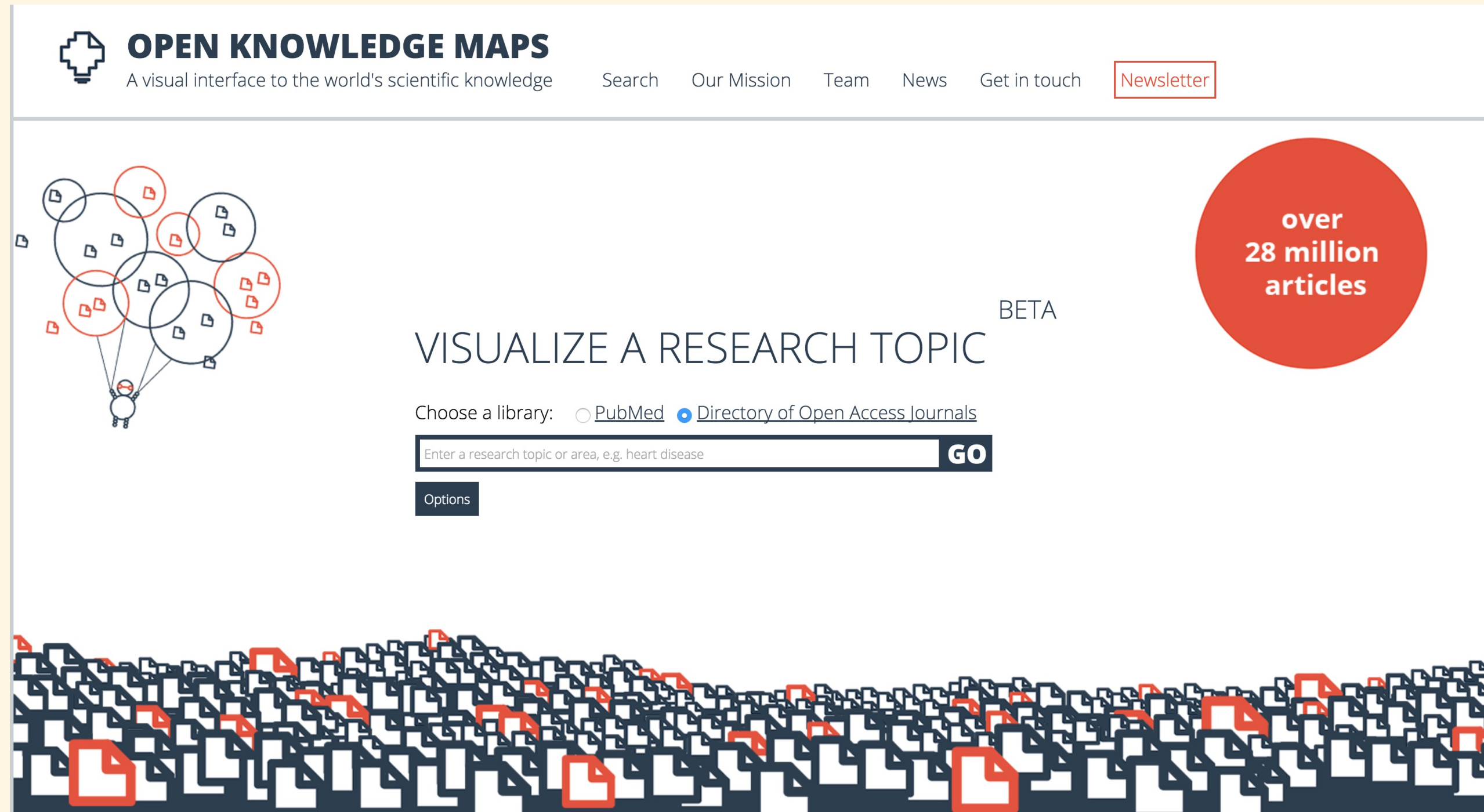
Branch: master ▾

New pull request




use case 3: OKMaps

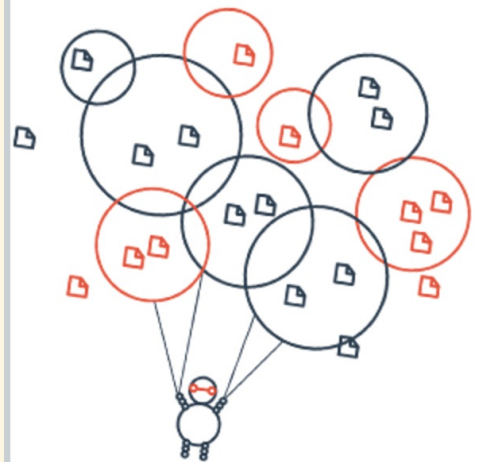
openknowledgemaps.org 



The screenshot shows the Open Knowledge Maps website. At the top left is the logo, a stylized cross made of squares, followed by the text "OPEN KNOWLEDGE MAPS" and the tagline "A visual interface to the world's scientific knowledge". To the right of the tagline is a navigation menu with links: "Search", "Our Mission", "Team", "News", "Get in touch", and "Newsletter" (which is highlighted with a red border). Below the navigation bar, on the left, is a graphic of several overlapping circles, each containing a document icon, connected by lines to a central point. In the center, the text "VISUALIZE A RESEARCH TOPIC" is displayed, with "BETA" to its right. Below this text are two radio buttons for "Choose a library": "PubMed" (unselected) and "Directory of Open Access Journals" (selected). Below the radio buttons is a search input field with the placeholder text "Enter a research topic or area, e.g. heart disease" and a "GO" button. Below the input field is an "Options" button. On the right side, there is a large red circle containing the text "over 28 million articles". At the bottom of the page, there is a decorative border consisting of a dense, repeating pattern of document icons in white and red.

 **OPEN KNOWLEDGE MAPS**
A visual interface to the world's scientific knowledge

[Search](#) [Our Mission](#) [Team](#) [News](#) [Get in touch](#) [Newsletter](#)



over 28 million articles

BETA

VISUALIZE A RESEARCH TOPIC

Choose a library: ☐ PubMed ☒ [Directory of Open Access Journals](#)

Enter a research topic or area, e.g. heart disease **GO**

Options

use case 4: mining gene ontology labels

goldi R package 

goldi: Gene Ontology Label Discernment and Identification


A tool for identifying multiple word key terms in free text with application to Gene Ontology labels.

Version: 1.0.0
Depends: R (≥ 2.15.0)
Imports: [dplyr](#), [Rcpp](#), [tm](#), [SnowballC](#), [magrittr](#), [futile.logger](#)
LinkingTo: [Rcpp](#), [RcppArmadillo](#)
Suggests: [testthat](#), [covr](#), [rmarkdown](#), [knitr](#), [pdftools](#), [RISmed](#)
Published: 2016-10-17
Author: Christopher B. Cole [aut, cre, cph], Sejal Patel [ctb], Jo Knight [ctb]
Maintainer: Christopher B. Cole <chris.c.1221 at gmail.com>
BugReports: <https://github.com/Chris1221/goldi/issues>
License: [MIT](#) + file [LICENSE](#)
URL: <https://github.com/Chris1221/goldi>
NeedsCompilation: yes
Materials: [README](#)
CRAN checks: [goldi results](#)

Downloads:

using our R package [pdftools](#)

use case 5: plant pathogens explained by taxonomic similarity

Bufford, et al. (2016). Taxonomic similarity, more than contact opportunity, explains novel plant-pathogen associations between native and alien taxa. New Phytologist 

Plant-pathogen associations explained by taxonomic similarity

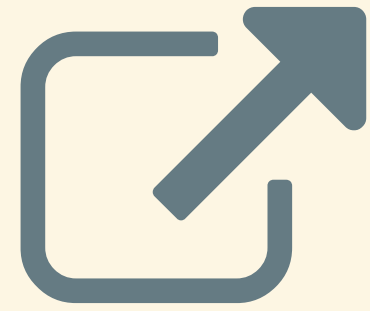
taxonomic data cleaning with our R package **taxize**

Wrap Up

- Open science is essential
- Open science tools are useful on their own
- rOpenSci: one of the tool makers
- Challenges going forward
 - Largely cultural - will slowly change

Wrap Up

- rOpenSci is a community project
- Let us know what you need
- Help us make better tools



scotttalks.info/ossps

Made w/: [reveal.js v3.2.0](#)

Some Styling: [Bootstrap v3.3.5](#)

Icons by: [FontAwesome v4.4.0](#)