Open science and R

Scott Chamberlain (@sckottie/@ropensci)

UC Berkeley / rOpenSci

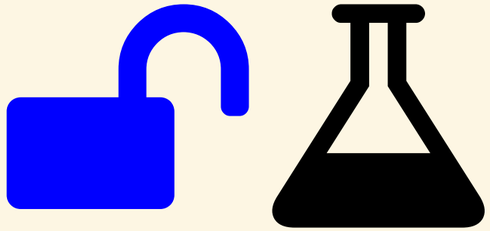
THE LEONA M. AND HARRY B.

HELMSLEY

CHARITABLE TRUST

scotttalks.info/ossps

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!

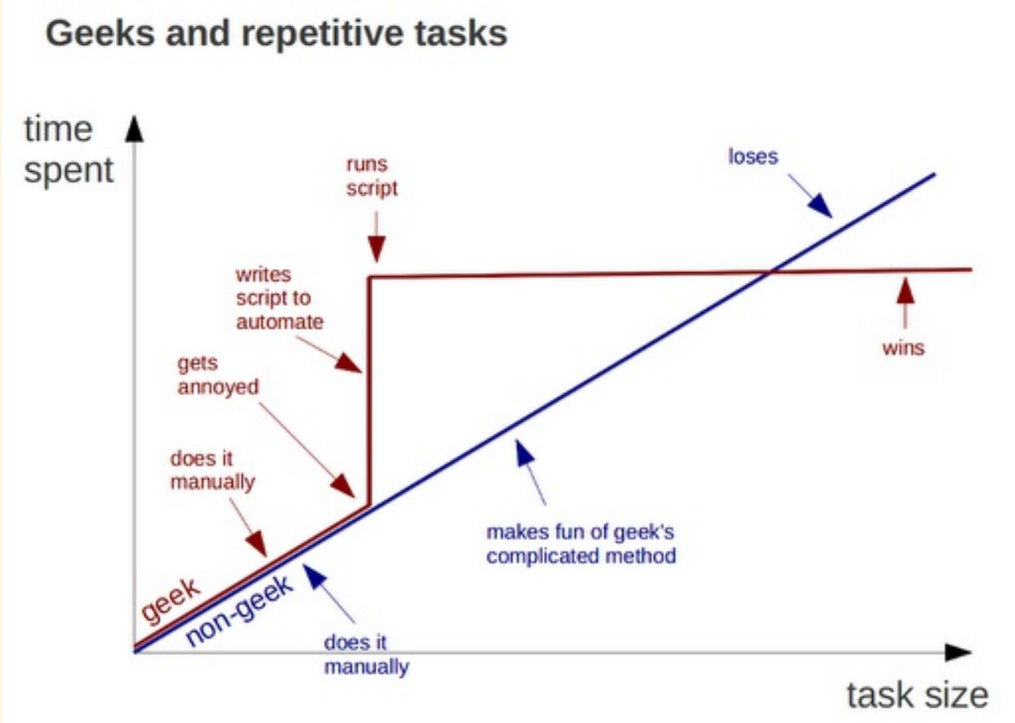
Versioning: code/data/text

Versioning: code/data/text

failure proofs your work

experiment freely!

Do all work programatically



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

Do all work programatically

Key to reproduciblity

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

Do all work programatically

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"

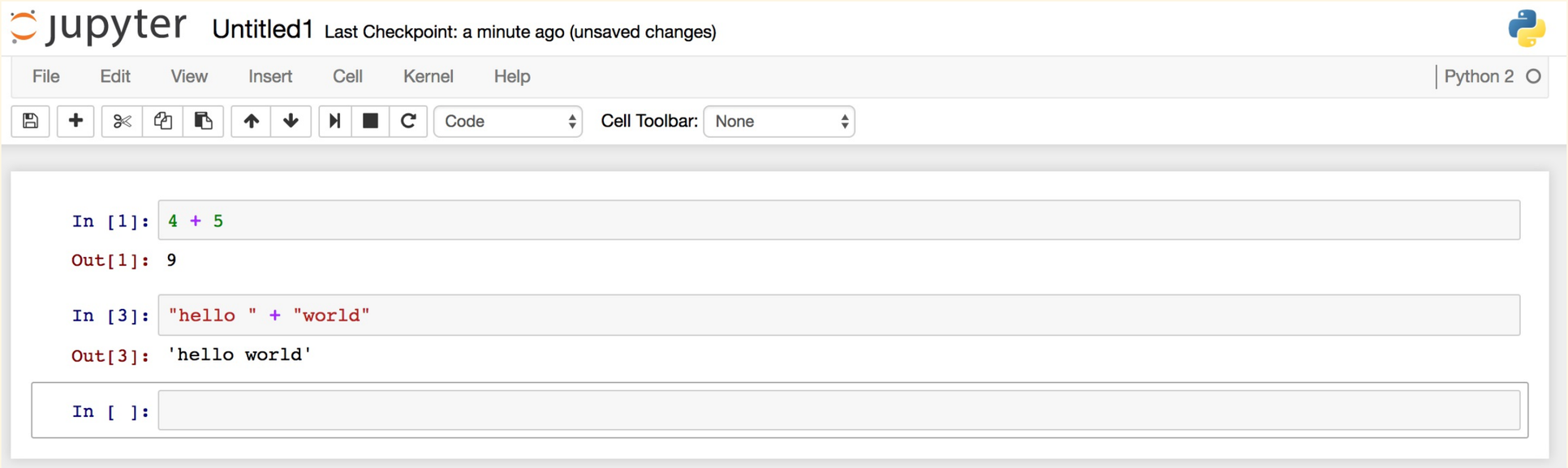
important scientific programming languages

python

R

julia

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",
```

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

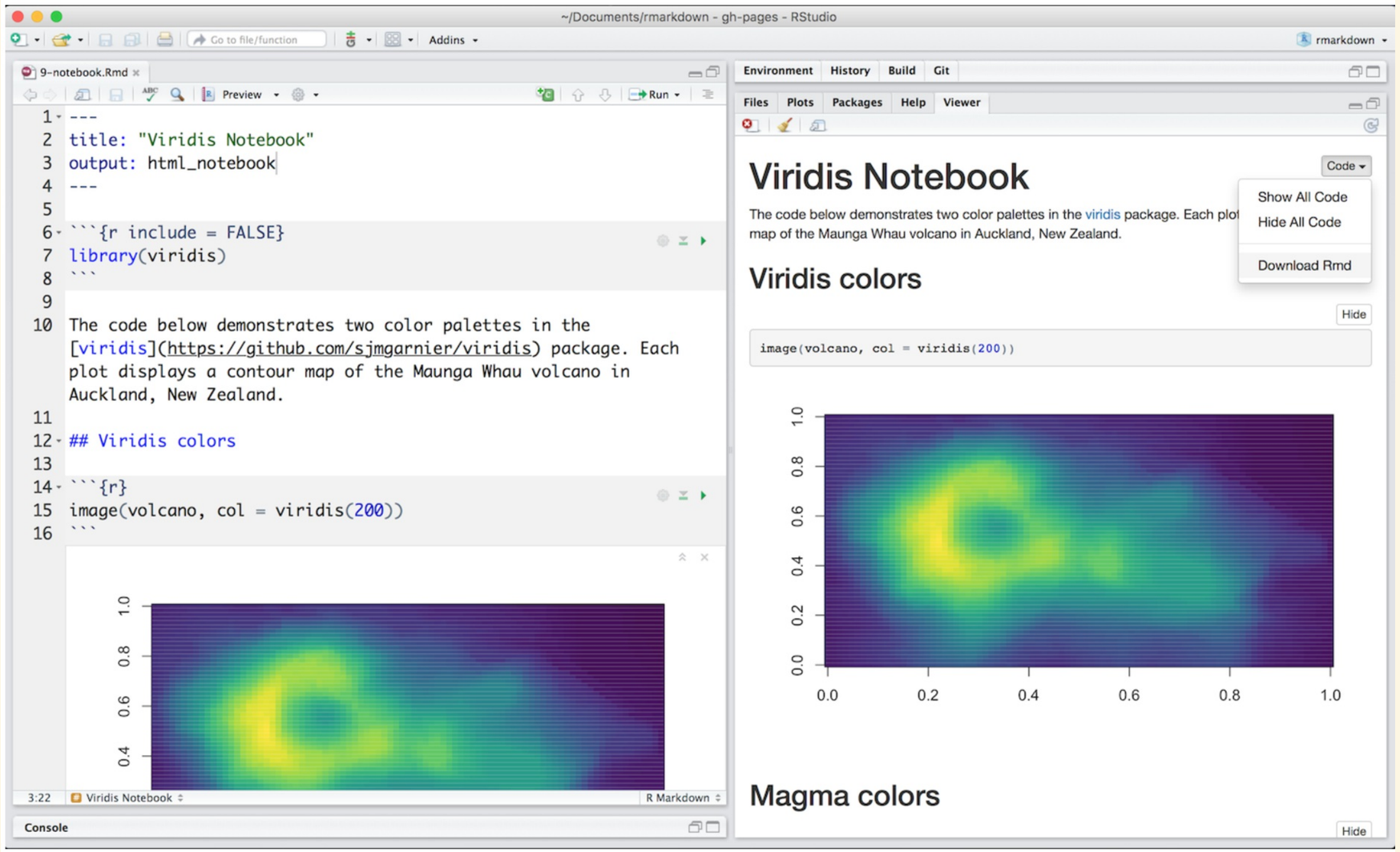
I found that cropping an image with R is **much simpler** than doing it with some other GIS software programs. Let's define the working directory and load the required package:

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

```
Loading required package: sp
```

Let's import the files into R:

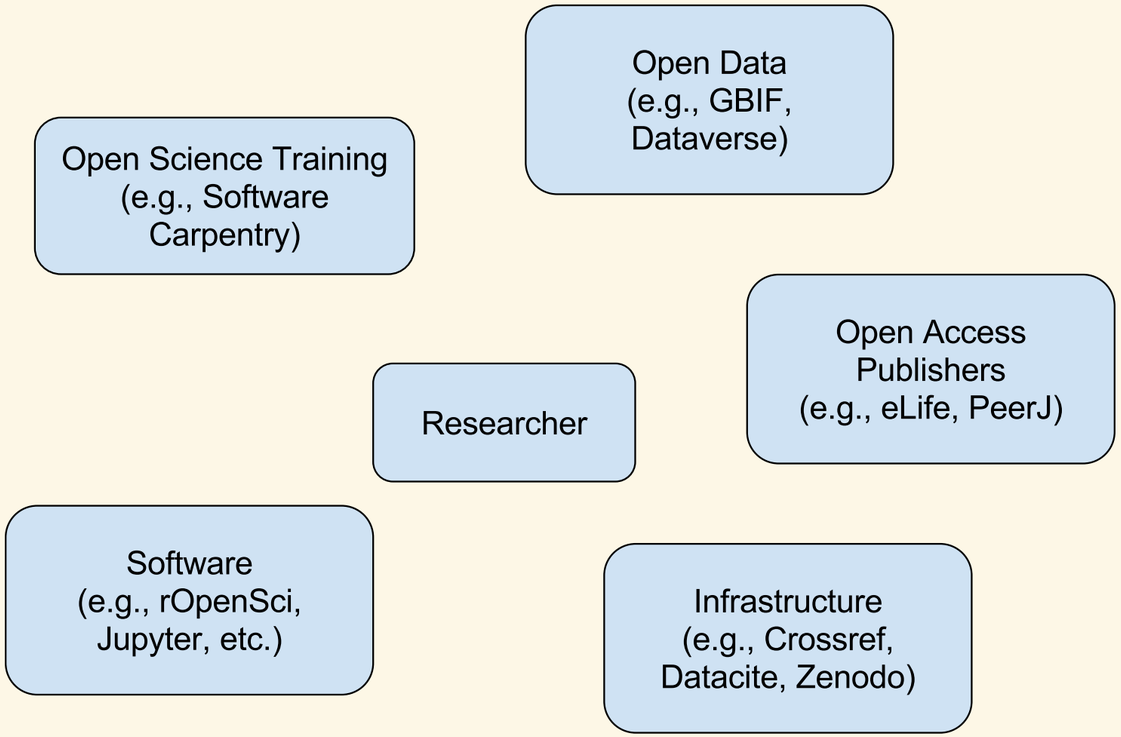
something similar in R: Rmarkdown



R language

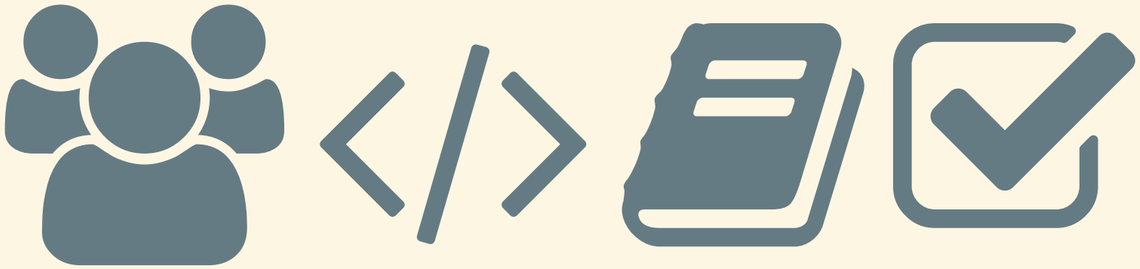
- used widely in biology, psychology, medicine, etc.
- rapidly growing user base, companies surrounding it
- includes all tools for open science workflow
- though work to be done ...

Open science ecosytsem



ROpenSci

rOpenSci does:



rOpenSci Staff

ropensci.org/about/#staff

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

rOpenSci Community

https://ropensci.org/community

Class Thido-Pfaff

Claas-Thido is a doctoral student at the new iDiv (Integrative Biodiversity Research) project in Leipzig. He maintains rbefdata and contributes to EML.

Brian O' Meara

Brian is a professor of ecology at University of Tennessee. He contributes to Reol with Barb

Hilary Parker

Hilary is a data analyst at Etsy, where she focuses on experiment-driven product development. Prior to Etsy, she got her PhD in Biostatistics from Johns Hopkins. In both of these roles she has been focused on reproducibility. She contributes to testdat.

Alyssa Frazee

Alyssa is a biostatistics PhD student at Johns Hopkins University working on computational biology problems involving lots of RNA-seq data. She contributes to testdat.

Ciera Martinez

Ciera is a PhD candidate studying Plant Biology. She's leading an effort to create a science reproducibility guide (code).

Martin Fenner

Martin is a developer with the Public Library of Science (PLOS). He is making an rOpenSci cookbook, a Ruby gem to wrap `knitr` called knitr-ruby and a Jekyll plugin for knitr called jekyll-knitr.

Community 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 +

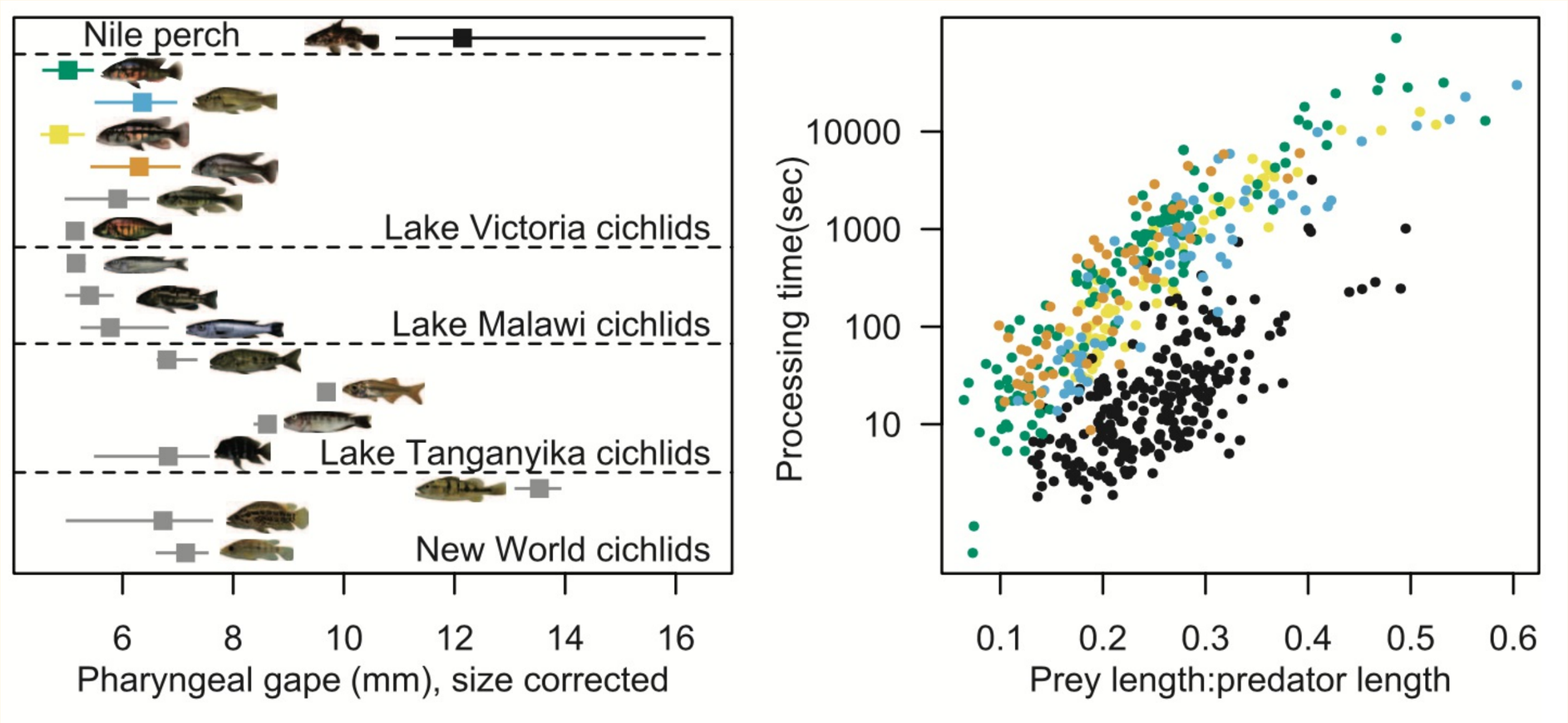
publish

We make data driven stories easier to tell

here are some stories ...

use case 1

McGee, M. D., Borstein, S. R., Neches, R. Y., Buescher, H. H., Seehausen, O., & Wainwright, P. C. (2015). A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. Science, 350(6264), 1077–1079



FishBase

ropensci / rfishbase

Code | Issues 16 | Pull requests 1 | Projects 0 | Wik

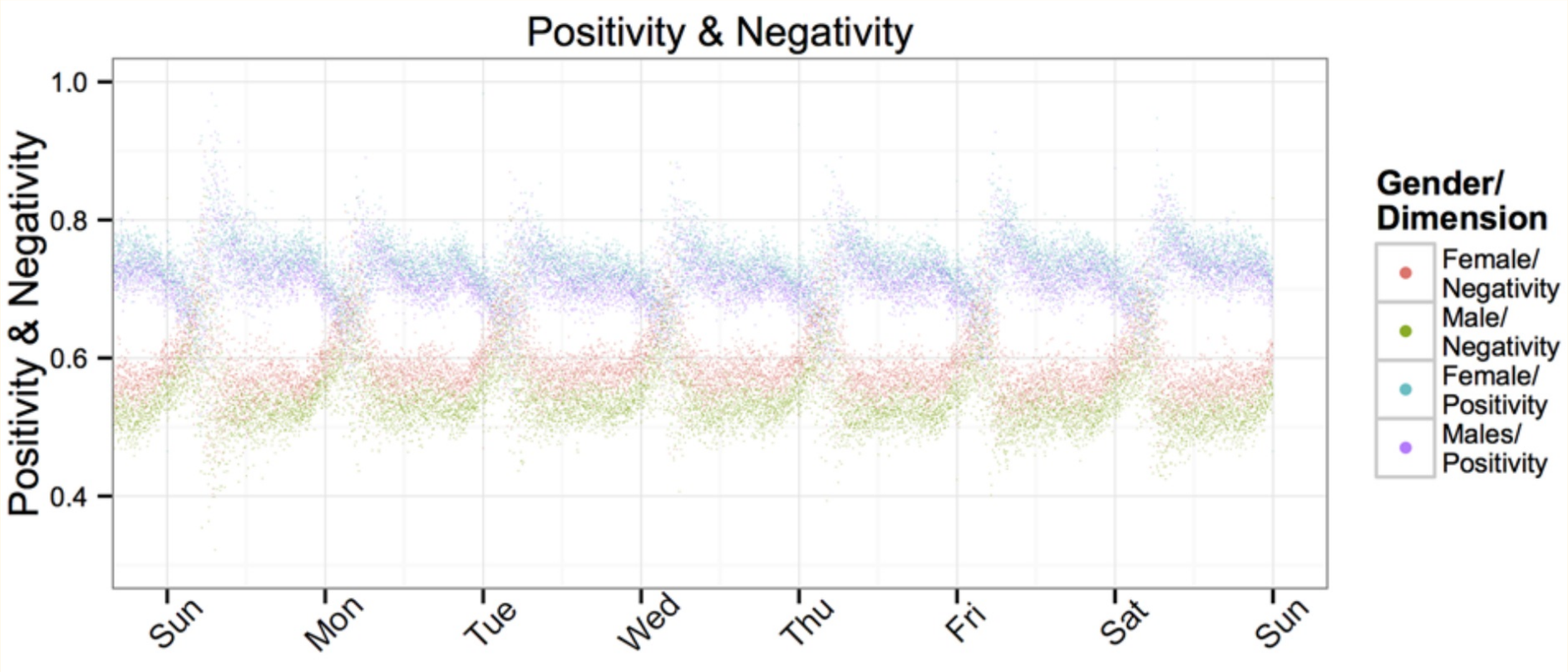
R interface to the fishbase.org database http://ropensci.org — Edit

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, 10(11), e0143051



ropensci / gender

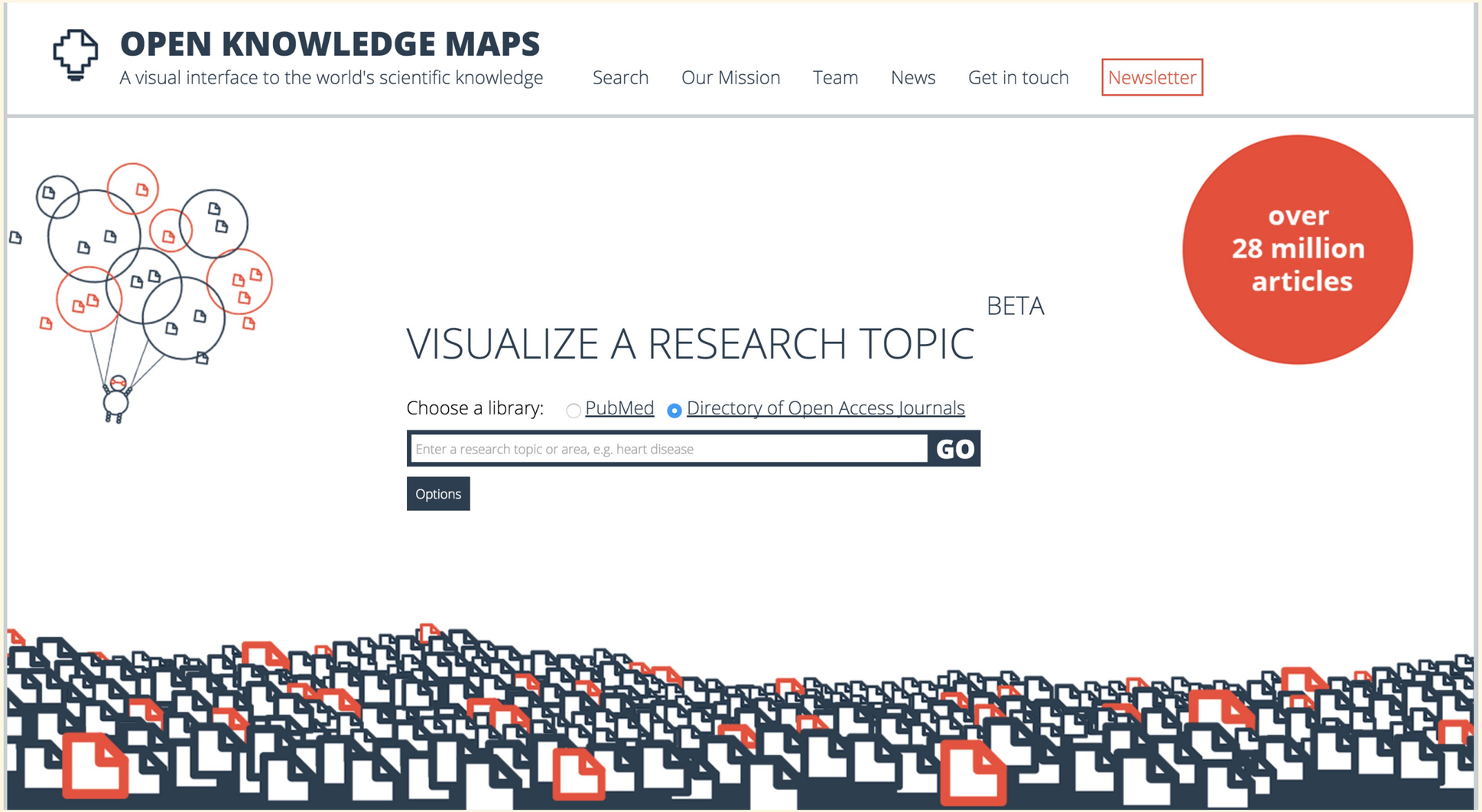
Code | Issues 4 | Pull requests 0 | Projects 0

Predict Gender from Names Using Historical Data — Edit

304 commits | 2 branches

Branch: master | New pull request

use case 3: OKMaps



use case 4: mining gene ontology labels

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

Plant-pathogen associations explained by taxonomic similarity

taxonomic data cleaning with our R package taxize

scotttalks.info/ossps

Made w/: reveal.js v3.2.0

Some Styling: Bootstrap v3.3.5

Icons by: FontAwesome v4.4.0