Open science, R & rOpenSci

Scott Chamberlain (@sckottie/@ropensci)

UC Berkeley / rOpenSci

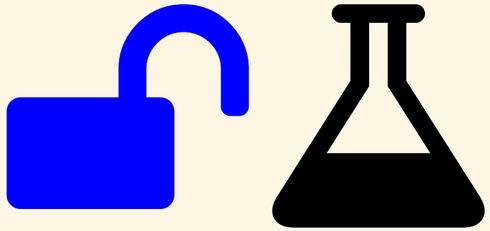
THE LEONA M. AND HARRY B.

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 worth learning

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!

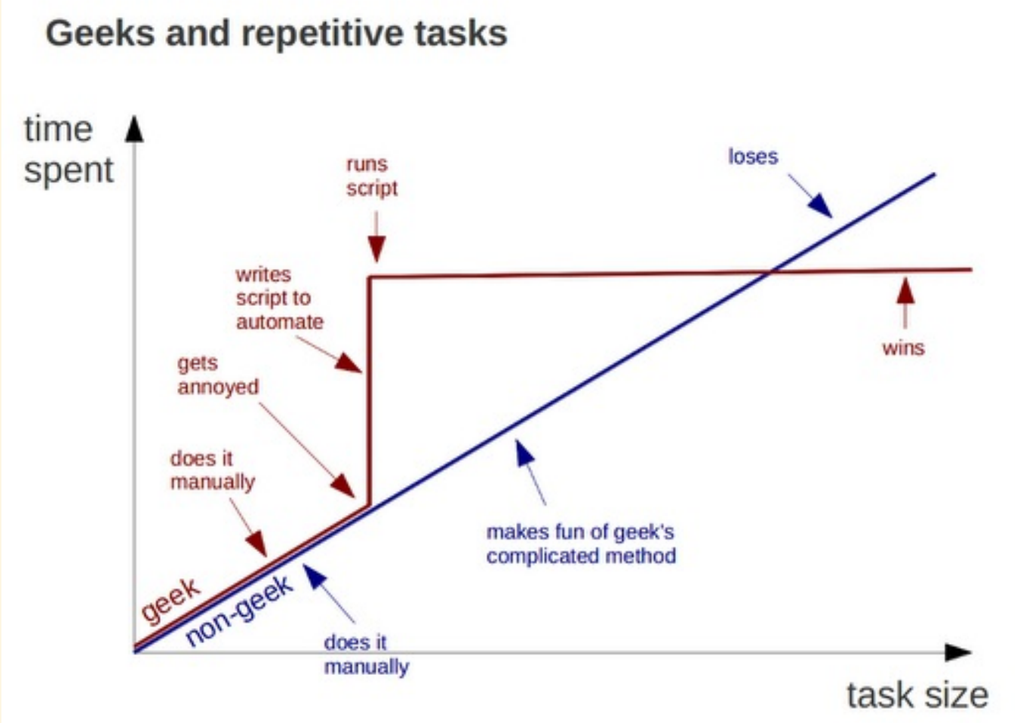
Versioning: code/data/text

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

important scientific programming languages

python

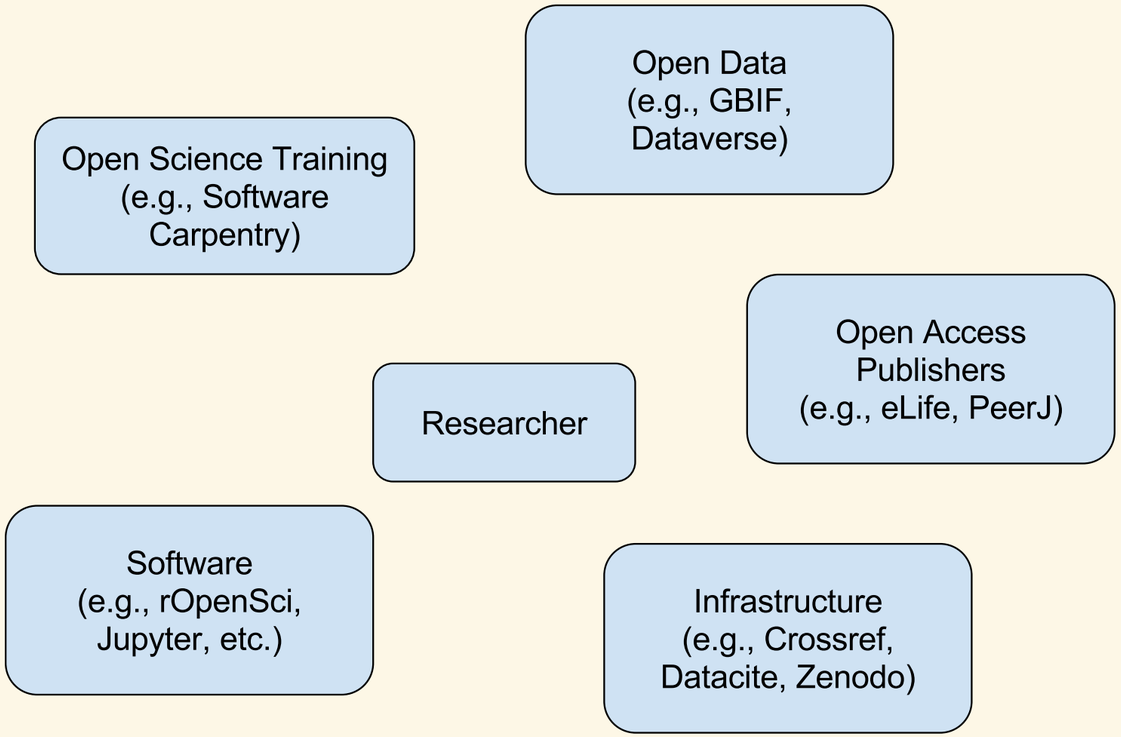
R

julia

R language

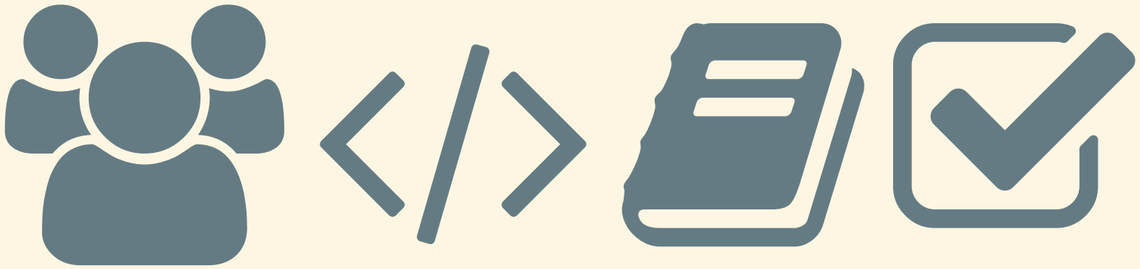
- used widely in biology, psychology, medicine, etc.
- rapidly growing user base, companies surrounding it
- includes all tools for open science workflow
- though work to be done ...

Open science ecosytsem



ROpenSci

rOpenSci does:



rOpenSci Staff

ropensci.org/about/#staff

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

Community stats

- ~ 250 code contributors
- large no. bug reports/feature requests
- ~ 364 Github repositories
- ~ 30,000 commits
- ~ 123 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

rOpenSci Tools

https://ropensci.org/packages

Data Publication | Data Access | Literature | Altmetrics | Scalable & Reproducible Computing | Databases | Data Vizualization | Image Processing | Data Tools | HTTP tools | Geospatial

Data Publication

Packages that not only retrieve data but also allows for data submission.

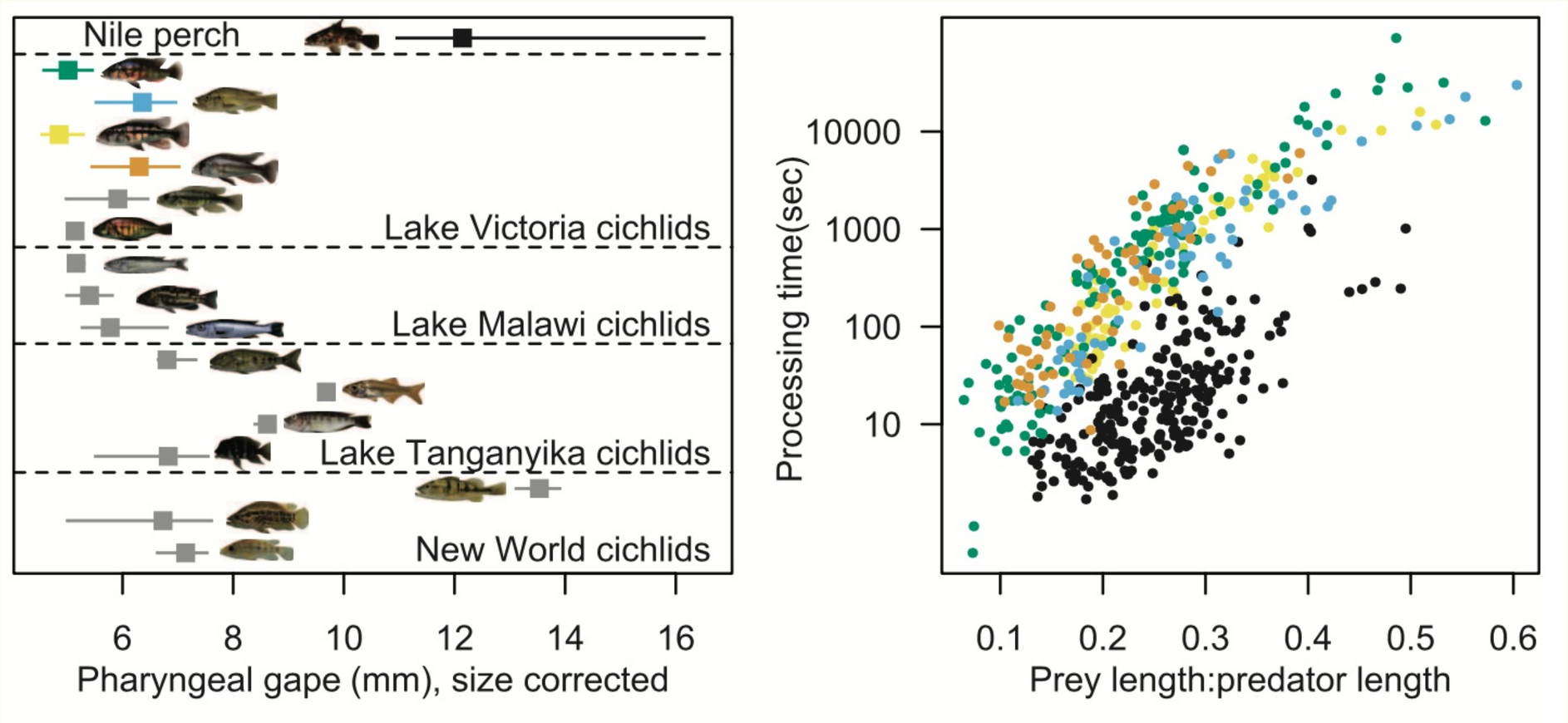
Package	Description	Details
dataone	Search across repositories, and read and write data and metadata from the DataONE federation of data repositories from R. Includes over 30 data repositories such as the KNB and Dryad.	CRAN GITHUB
datapack	A flexible container to transport and manipulate data and associated resources	CRAN GITHUB
dvn	Programmatic interface to the DataVerse Network.	CRAN GITHUB
EML	An R package for reading, writing, integrating and publishing data using the Ecological Metadata Language (EML) format.	CRAN GITHUB
rfigshare	Push data, figures, and text to, and search and retrieve data from, Figshare from R	CRAN GITHUB
RNeXML	Semantically rich NeXML I/O in R - next generation XML for Phylogenetic data.	CRAN GITHUB

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

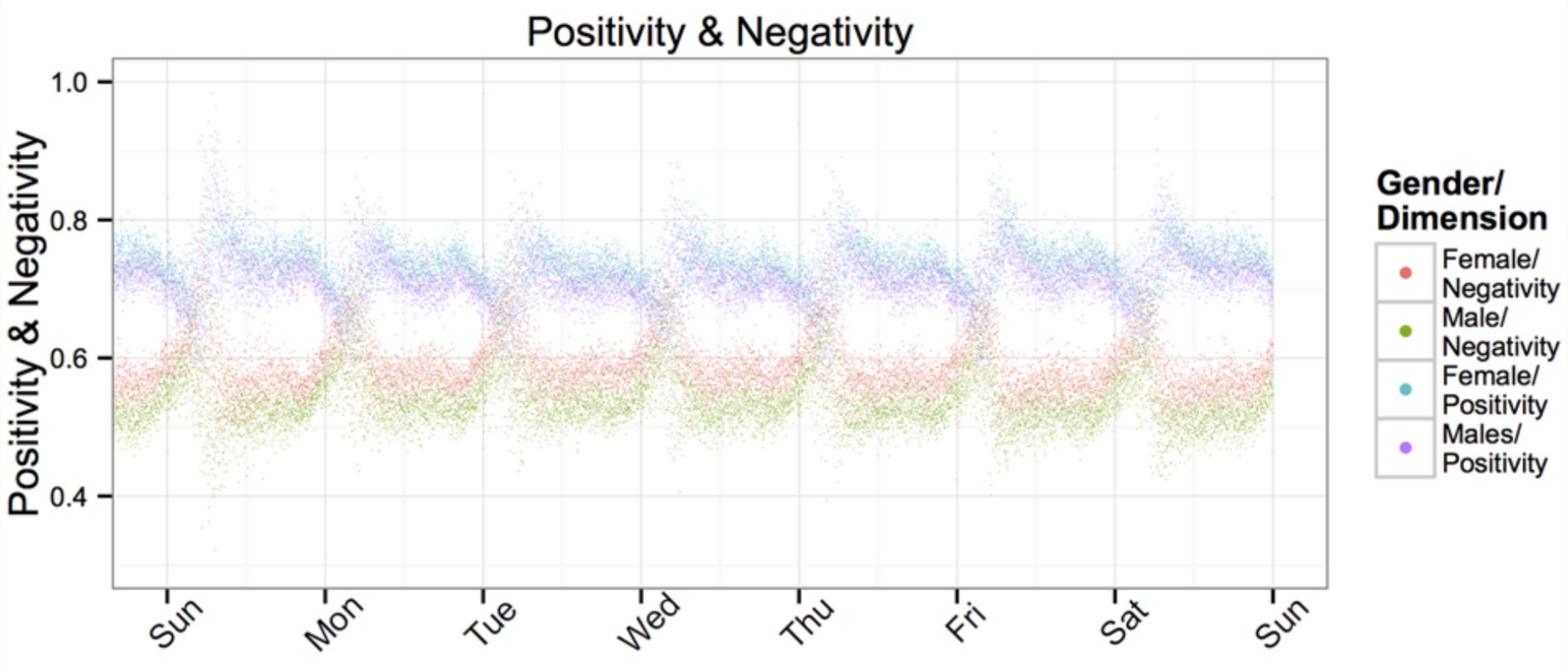
R interface to the fishbase.org database http://ropensci.org — Edit

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

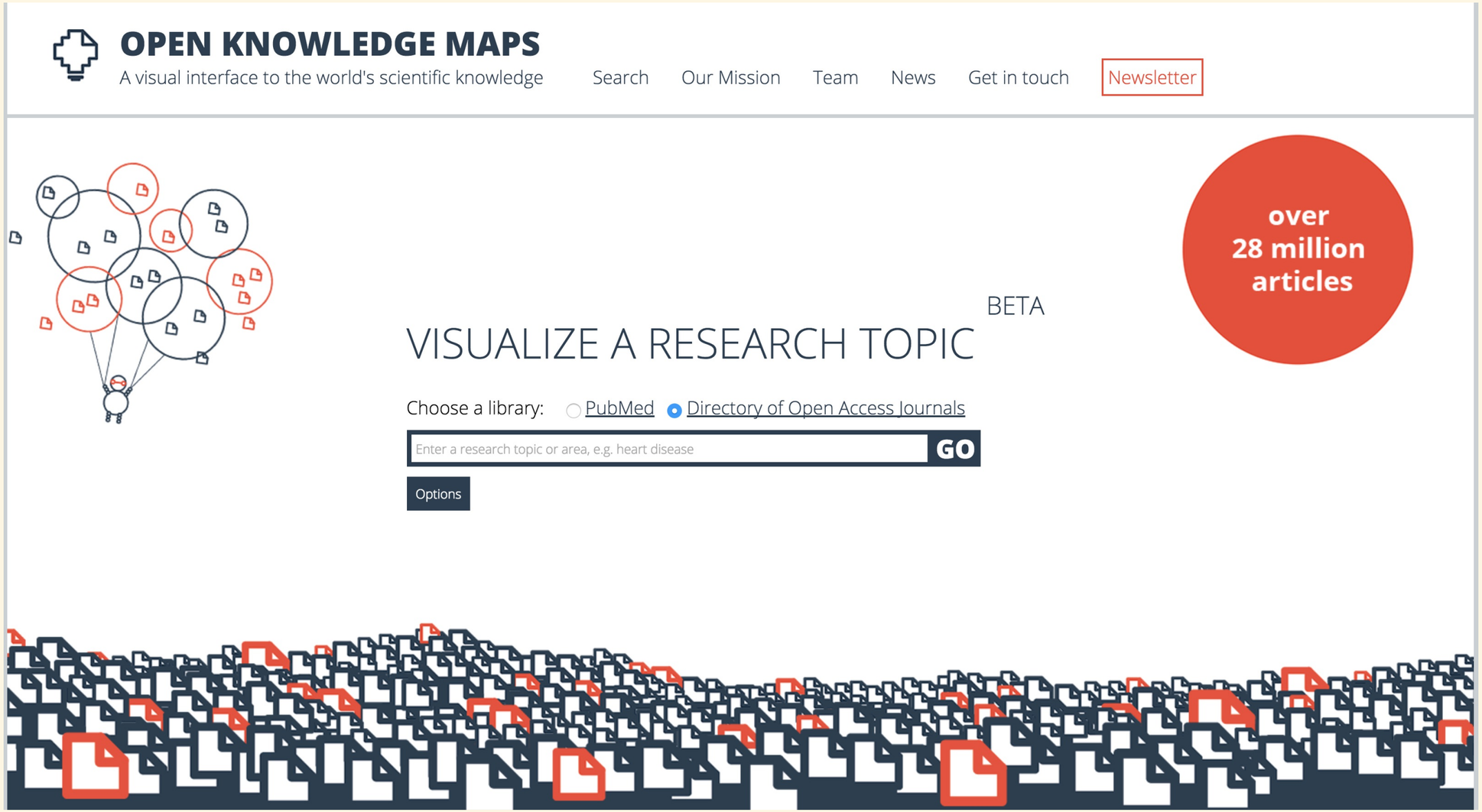
Code | Issues 4 | Pull requests 0 | Projects 0

Predict Gender from Names Using Historical Data — Edit

304 commits | 2 branches

Branch: master | New pull request

use case 3: OKMaps



rOpenSci Biodiversity Tools

Taxonomy

Taxonomy

- taxize - Taxonomic toolbelt
- taxizesoap - Taxonomic toolbelt (SOAP)
- ritis - ITIS client (avail. in taxize)
- taxizedb - Access to SQL dumps
- wikitaxa - Taxonomy from Wiki-pedia/-species/-commons
- worrms - WORMS client (avail. in taxize)
- natserv - Natureserve client (avail. in taxize)
- taxa - Taxonomic classes to be used by other pkgs (coming soon)

Taxonomic data from 20 sources - taxize

Taxonomic hierarchies from NCBI/ITIS/COL/etc

```
library('taxize')
classification("Chironomus riparius", db = "gbif")
```

```
#> $`Chironomus riparius`
#>                name    rank      id
#> 1           Animalia kingdom       1
#> 2          Arthropoda  phylum      54
#> 3             Insecta   class     216
#> 4             Diptera   order     811
#> 5        Chironomidae  family    3343
#> 6          Chironomus   genus 1448033
#> 7 Chironomus riparius species 1448237
```

taxa classes - taxa

```
library('taxa')
taxon_name("Mammalia")
taxon_rank("class")
taxon_id(9681)

mammalia <- taxon(taxon_name("Mammalia"), taxon_rank("class"), taxon_id(9681))
#> <Taxon>
#>   name: Mammalia
#>   rank: class
#>   id: 9681
#>   authority: none

felidae <- taxon(taxon_name("Felidae"), taxon_rank("family"), taxon_id(9681))
panthera <- taxon(taxon_name("Panthera"), taxon_rank("genus"), taxon_id(146712))
tigris <- taxon(taxon_name("tigris"), taxon_rank("species"), taxon_id(9696))

tiger <- hierarchy(mammalia, felidae, panthera, tigris)
#> <Hierarchy>
#>   no. taxon's:  4
#>   Mammalia / class / 9681
#>   Felidae / family / 9681
#>   Panthera / genus / 146712
#>   tigris / species / 9696
```

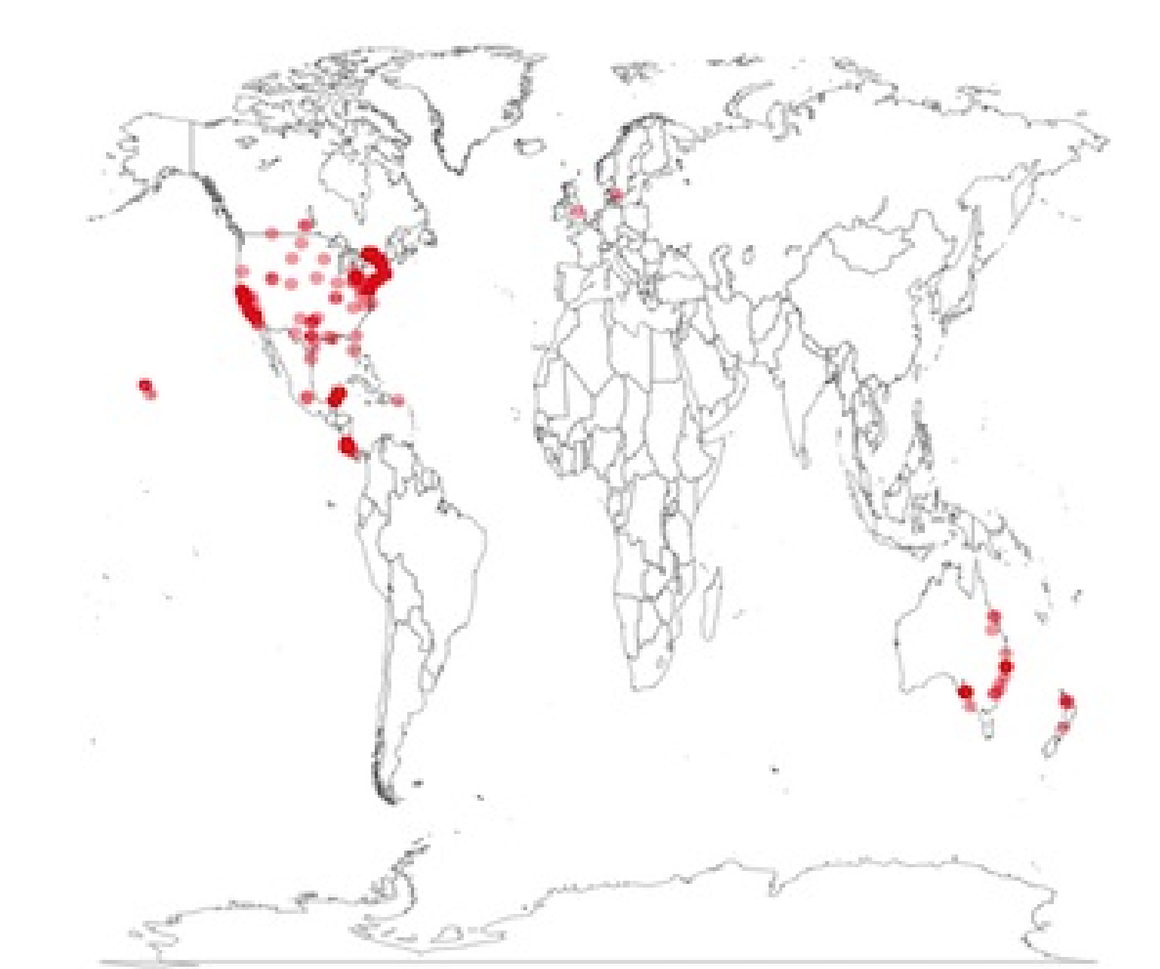
Species occurrence data

Species occurrence data

- spocc - One client to rule them all
- rgbif - GBIF data (avail. in spocc)
- AntWeb - AntWeb ant data (avail. in spocc)
- ecoengine - Berekeley Ecoengine client (avail. in spocc)
- rinat - iNaturalist client (avail. in spocc)
- rbison - USGS BISON client (avail. in spocc)
- rebird - eBird data (avail. in spocc)
- rvertnet - VertNet data (avail. in spocc)
- rfishbase - Fishbase.org data

Mapping biodiversity data - rgbif

```
library(rgbif)
key <- name_backbone(name='Danaus plexippus', kingdom='animals')$speciesKey
out <- occ_search(taxonKey=key, limit=300, return='data')
gbifmap(out)
```



Unified species occurrence data - spocc

```
library(spocc)
spnames <- c('Accipiter striatus', 'Setophaga caerulescens', 'Spinus tristis')
(out <- occ(query = spnames, from = c('gbif', 'ebird')))
```

```
#> Summary of results - occurrences found for:
#>  gbif  : 75 records across 3 species
#>  bison :  0 records across 3 species
#>  inat  :  0 records across 3 species
#>  ebird :  75 records across 3 species
#>  ecoengine :  0 records across 3 species
#>  antweb :  0 records across 3 species
#>  idigbio :  0 records across 3 species
#>  obis :  0 records across 3 species
#>  ala :  0 records across 3 species
```

Occurrence data cleaning

Occurrence data cleaning

- scrubr - general purpose cleaner
- rgeospatialquality - API wrapper for checking occ. data

Species occurrence data cleaning - scrubr

```
library(scrubr)
NROW(sample_data_1)
#> [1] 1500

sample_data_1 %>%
  coord_impossible() %>%
  coord_incomplete() %>%
  coord_unlikely()
```

```
#> # A tibble: 1,294 × 5
#>                name  longitude latitude               date        key
#> *             <chr>      <dbl>    <dbl>             <dttm>      <int>
#> 1  Ursus americanus  -79.68283 38.36662 2015-01-14 16:36:45 1065590124
#> 2  Ursus americanus  -82.42028 35.73304 2015-01-13 00:25:39 1065588899
#> 3  Ursus americanus  -99.09625 23.66893 2015-02-20 23:00:00 1098894889
#> 4  Ursus americanus  -72.77432 43.94883 2015-02-13 16:16:41 1065611122
#> 5  Ursus americanus  -72.34617 43.86464 2015-03-01 20:20:45 1088908315
#> 6  Ursus americanus -108.53674 32.65219 2015-03-29 17:06:54 1088932238
#> 7  Ursus americanus -108.53691 32.65237 2015-03-29 17:12:50 1088932273
#> 8  Ursus americanus -123.82900 40.13240 2015-03-28 23:00:00 1132403409
#> 9  Ursus americanus  -78.25027 36.93018 2015-03-20 21:11:24 1088923534
#> 10 Ursus americanus  -76.78671 35.53079 2015-04-05 23:00:00 1088954559
#> # ... with 1,284 more rows
```

Geospatial

Geospatial

- geojson - GeoJSON classes
- geojsonio - GeoJSON/TopoJSON input/output
- geonames - Geonames API client
- wicket Well-Known Text tools
- wellknown - WKT from R objects - and convert to GeoJSON
- rnaturalearth - NaturalEarth data
- osmplotr - Open Street Maps tools
- opencage - OpenCage geocoding API
- lawn - Turf.js javascript geo client
- geojsonlint - lint GeoJSON

Geospatial: Geonames data - geonames

http://www.geonames.org/

```
library(geonames)
```

Find a contry code

```
GNcountryCode(lat = 47.03, lng = 10.2)
```

Search for nearby streets

```
GNfindNearbyStreets(lat = 37.45, lng = -122.18)
```

Search by place name

```
GNsearch(q = "london", maxRows = 10)
```

Postal code search

```
GNpostalCodeSearch(postalcode = 90210, country = "FI")
```

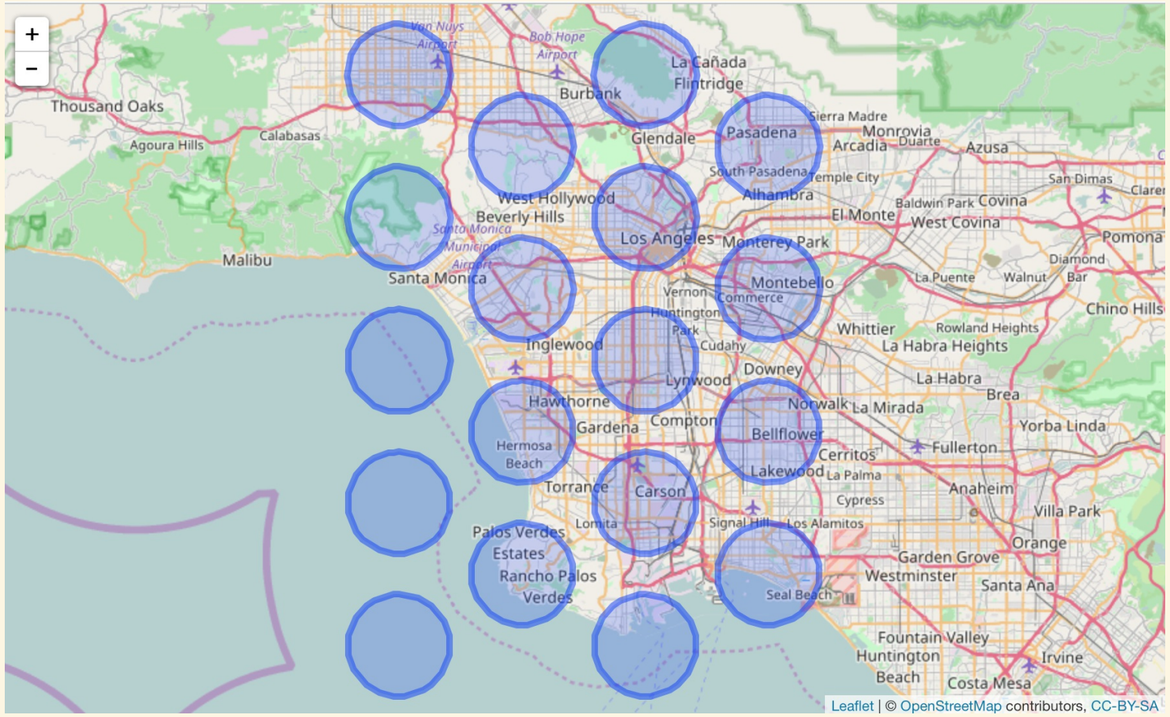
Geospatial: conversion between data/spatial data formats - geojsonio

- geojson_list - convert to GeoJSON as R list
- geojson_json - convert to GeoJSON as JSON
- geojson_read/geojson_write - read/write GeoJSON

from most R object types + many spatial data formats

Geospatial: Spatial ops. w/ GeoJSON & w/o heavy dependencies - lawn

```
library(lawn)
bbox <- c(-118.521, 33.715, -118.145, 34.179)
lawn_hex_grid(bbox, 10, 'miles') %>%
  as_feature(hex_grid) %>%
  purrr::map(lawn_centroid) %>%
  purrr::map(lawn_circle, radius = 5) %>%
  view
```



scotttalks.info/ossps

Made w/: reveal.js v3.2.0

Some Styling: Bootstrap v3.3.5

Icons by: FontAwesome v4.4.0