rOpenSci Data Packages

Scott Chamberlain

sckottie

Why Data Packages?

- Reduce duplicated effort by each researcher
- One best way to get data XYZ
- Reduced user error
- Allow researchers to focus on the science
- One more part of the science workflow that's reproducible

Data Packages: Caveats

- User base (# of people) for data pkgs small relative to utilities
  - ... attracts fewer contributors
  - ... & all the carry on effects of above
- Pkg can get out of sync with data source's API
- Risk leaving out metadata/context

rOpenSci Data Packages

- Biological occurrences
- Taxonomy
- Data utilities

find more: ropensci.org/packages

Data package patterns

Or at least patterns we strive towards ...

- Cache downloaded data for FTP
- Return data.frame's: facilitates downstream processing
- Make it easy to cite data providers
- Incorporate metadata
- Properly tested
    - webmockr/vcr - See also: HTTP testing book
- Helps if maintainer has domain knowledge
    - rOpensci software review facilitating this

Occurrence Data Packages

- spocc - Biodiversity data toolbelt
- rgbif - GBIF data
- ecoengine - Berekeley Ecoengine client
- rinat - iNaturalist client
- rbison - USGS BISON client
- rebird - eBird data via their API
- auk - eBird bulk data
- rvertnet - VertNet data
- rfishbase - Fishbase.org data
- finch - Handle Darwin Core
- seaaroundus - Sea Around Us data
- rglobi - Global Biotic Interactions

Occurrence Data Packages

- spocc - Biodiversity data toolbelt
- rgbif - GBIF data
- ecoengine - Berekeley Ecoengine client
- rinat - iNaturalist client
- rbison - USGS BISON client
- rebird - eBird data via their API
- auk - eBird bulk data
- rvertnet - VertNet data
- rfishbase - Fishbase.org data
- finch - Handle Darwin Core
- seaaroundus - Sea Around Us data
- rglobi - Global Biotic Interactions

Occurrence Data Packages

- spocc - Biodiversity data toolbelt
- rgbif - GBIF data
- ecoengine - Berekeley Ecoengine client
- rinat - iNaturalist client
- rbison - USGS BISON client
- rebird - eBird data via their API
- auk - eBird bulk data
- rvertnet - VertNet data
- rfishbase - Fishbase.org data
- finch - Handle Darwin Core
- seaaroundus - Sea Around Us data
- rglobi - Global Biotic Interactions

rgbif

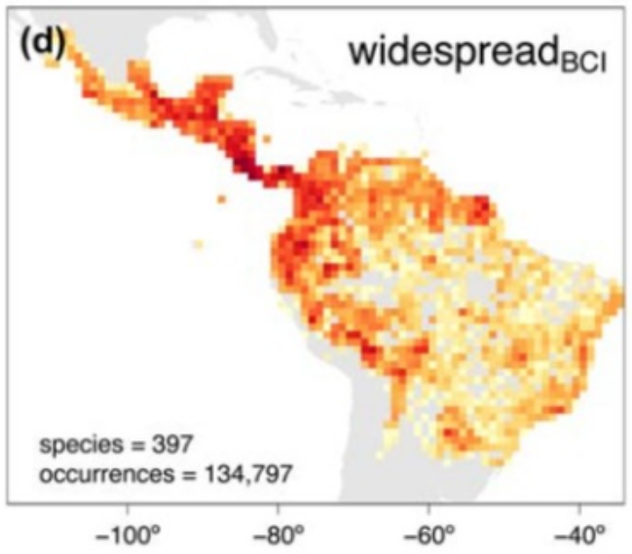
Access > 1 billion occurrence records from GBIF

ropensci/rgbif

- Search for species
- Download occurrence data
- Clean occurrence data
- Make maps

rgbif usage

Bemmels, et al. (2018) -- Filter-dispersal assembly of lowland Neotropical rainforests across the Andes



> *... records were obtained for each species ... using ...* `rgbif` *... Quality controls were applied to ensure that occurrence records were correctly taxonomically identified and accurately georeferenced. In particular, records with the following issues were excluded ... invalid coordinates or geodetic datum; failed or suspicious coordinate reprojection; ... species name with no match ...*

~70 more citations at https://github.com/ropensci/roapi/blob/master/data/citations.csv

rgbif usage: Checklist recipe

TrIAS Project - standardizing species checklist data to Darwin Core using R

```
├── README.md                   : Description of this repository
├── LICENSE                     : Repository license
├── checklist-recipe.Rproj      : RStudio project file
├── .gitignore                  : Files and directories to be ignored by git
|
├── data
|   ├── raw                     : Source data, input for mapping script
|   └── processed               : Darwin Core output of mapping script GENERATED
|
├── docs                        : Repository website GENERATED
└── src
    ├── dwc_mapping.Rmd         : Darwin Core mapping script, core functionality
    ├── _site.yml               : Settings to build website in /docs
    └── index.Rmd               : Template for website homepage
```

rgbif

Taxonomy Packages

- taxize - Taxonomic toolbelt (remote data)
  - taxize book
- taxa - Taxonomic R classes
- taxizedb - Leverage dumps of taxonomic database locally
- ritis - USGS's ITIS
- rotl - Open Tree of Life
- rentrez - NCBI ENTREZ taxonomy database
- worrms - WORMS marine taxonomy
- wikitaxa - Taxonomy data on Wikipedia
- zbank - ZooBank
- rgbif - GBIF taxonomy data

Taxonomy Packages

- taxize - Taxonomic toolbelt (remote data)
  - taxize book
- taxa - Taxonomic R classes
- taxizedb - Leverage dumps of taxonomic database locally
- ritis - USGS's ITIS
- rotl - Open Tree of Life
- rentrez - NCBI ENTREZ taxonomy database
- worrms - WORMS marine taxonomy
- wikitaxa - Taxonomy data on Wikipedia
- zbank - ZooBank
- rgbif - GBIF taxonomy data

taxize

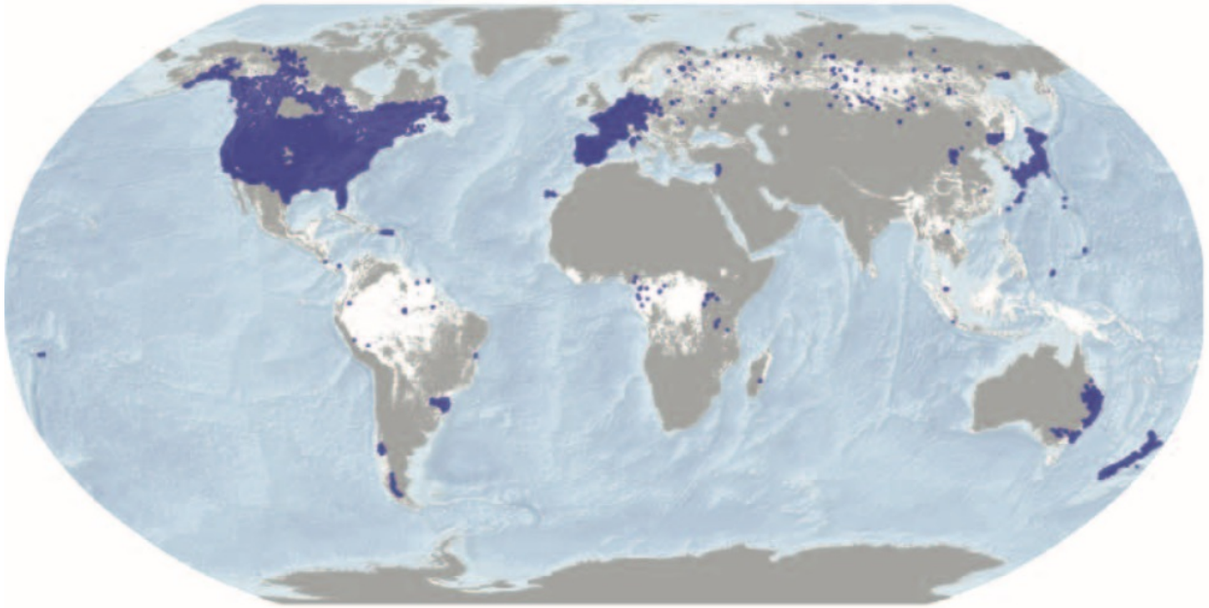
Access taxonomic data from > 20 sources

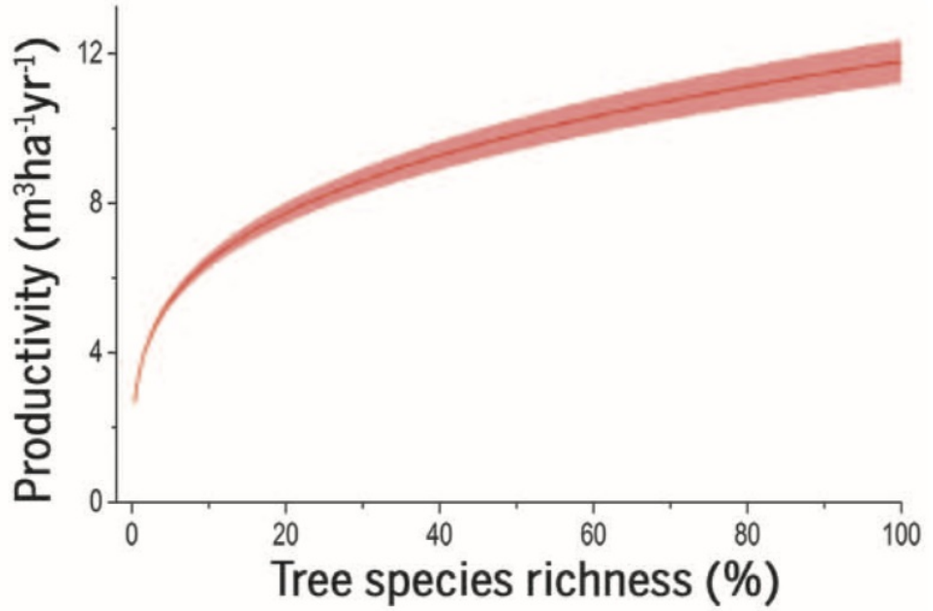
ropensci/taxize

- Resolve misspelled names
- Search for names
- Fetch taxonomic classifications
- Fetch all taxa up- and down-stream
- Fetch taxonomic synonyms
- Common names to taxonomic and vice versa

taxize usage

Liang, J., et al. (2016) -- Positive biodiversity-productivity relationship predominant in global forests





> *there were ... 8,737 species ... We verified all ... names against 60 taxonomic data-bases, including NCBI, GRIN Taxonomy for Plants, Tropicos–Missouri Botanical Garden, and the International Plant Names Index, using the 'taxize' package in R*

~90 more citations at https://github.com/ropensci/roapi/blob/master/data/citations.csv

taxa

Taxonomic classes & manipulate taxonomy+data

ropensci/taxa

- Classes for taxonomic data only
- Classes for taxonomic data + user data
- dplyr-like "taxonomically aware" data manipulation

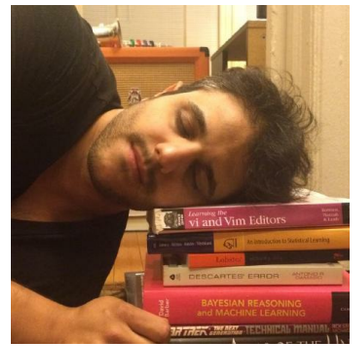
Utility Packages

- jqr - R client for jq, the JSON processor
- jsonld - JSON linked data
- rerddap - ERDDAP server client
- rdflib - RDF pkg, wrapped around Redland
- assertr - Assertions for analysis pipelines

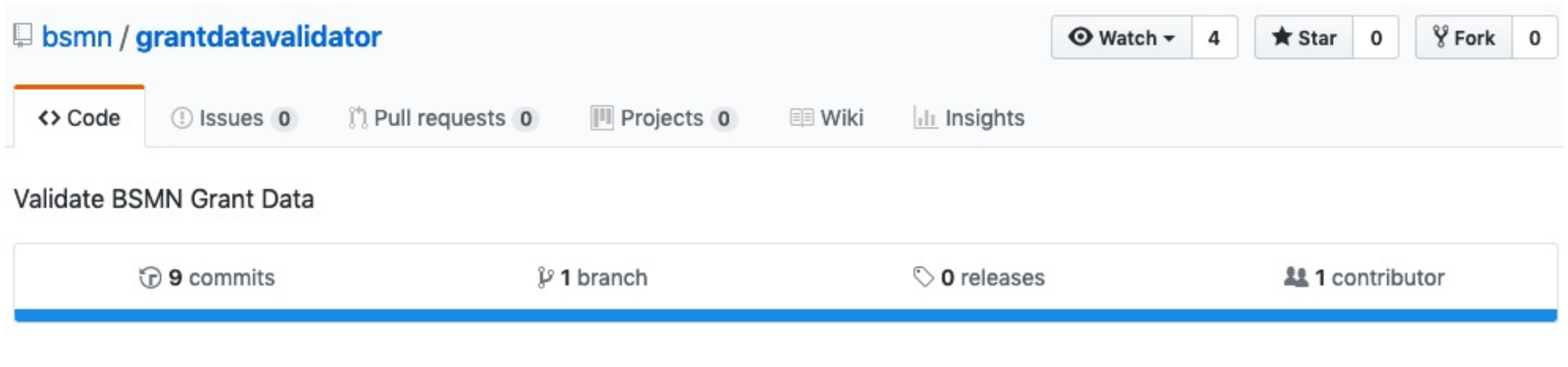
Utility Packages

- jqr - R client for jq, the JSON processor
- jsonld - JSON linked data
- rerddap - ERDDAP server client
- rdflib - RDF pkg, wrapped around Redland
- assertr - Assertions for analysis pipelines

assertr usage: eg



Brain Somatic Mosaicism Network - Validate BSMN Grant Data

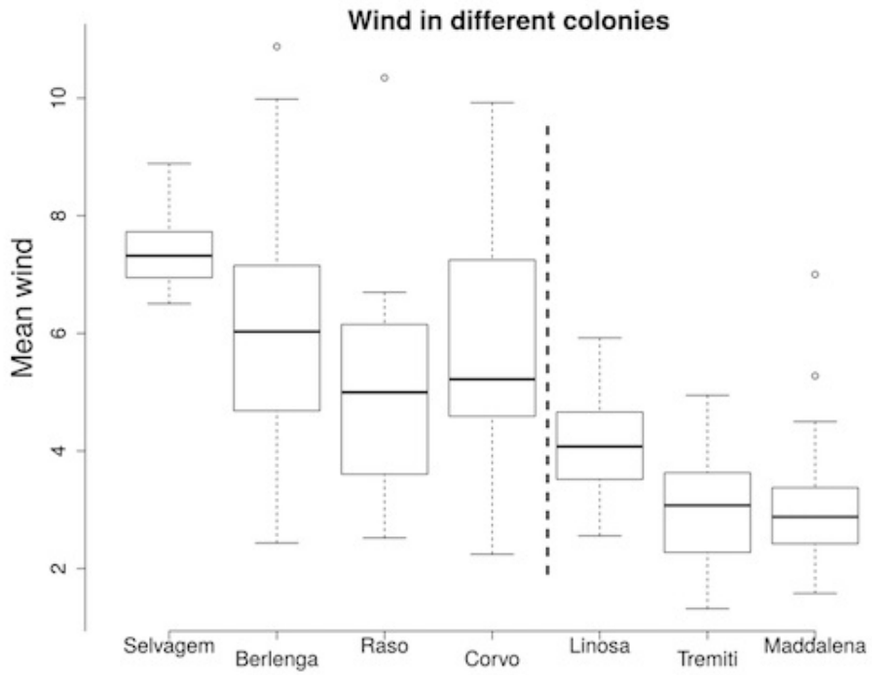



```
data %>%
   assertr::chain_start() %>%
   assertr::verify(nrow(data) == 3) %>%
   assertr::verify(assertr::is_uniq(nda_short_name)) %>%
   assertr::verify(assertr::not_na(grant)) %>%
   assertr::verify(dplyr::n_distinct(grant) == 1) %>%
   assertr::verify(nda_short_name %in% expected_nda_short_names) %>%
   assertr::chain_end() %>%
   tibble::as_tibble()
```

assertr

rerddap usage: eg

Abolaffio, J., et al. (2018) -- Olfactory-cued navigation in shearwaters: linking movement patterns to mechanisms



> *wind data were downloaded from the NOAA web site from the* `rerdapp` *package for R*

rerddap

Data integration: Steps

- Start with a species list
- Clean names with taxize
- Get occurrence data with rgbif
- Clean occurrence data w/ rgbif, scrubr or CoordinateCleaner
- assertr to check data
- Map with mapr

Data integration: code

```
# read in species list
spp <- read.csv("spp_list.txt", header = TRUE,
  stringsAsFactors = FALSE)$bad
# resolve names: fix misspellings
spp2 <- taxize::gnr_resolve(spp, data_source_ids = 11,
  canonical = TRUE)$matched_name2
# fetch GBIF occurrence data
dat <- rgbif::occ_data(scientificName = spp2, limit = 300)
# remove data with issues: COUNTRY_MISMATCH & COORDINATE_ROUNDED
dat <- rgbif::occ_issues(dat, -cum, -cdround)
# make a single data.frame
dat <- dplyr::bind_rows(lapply(dat, "[[", "data"))
# remove records with incomplete lat/lon data
dat <- scrubr::coord_incomplete(dat)
```

Data: spp_list.txt

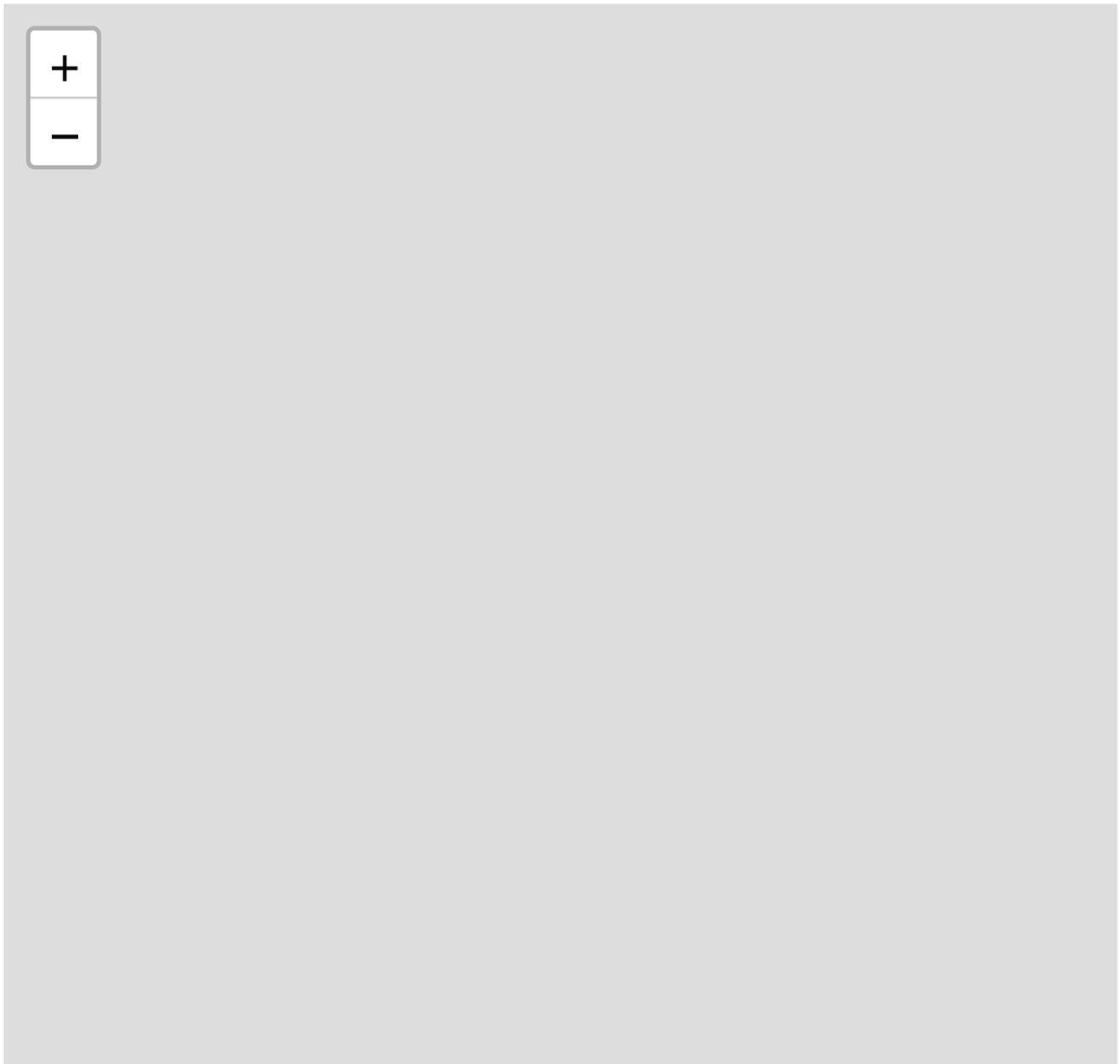
Data integration: code

Test assertions about the data

```
library(magrittr)
dat %>%
  assertr::chain_start() %>%
  # does it have more than 100 rows?
  assertr::verify(NROW(dat) > 100) %>%
  # is the key for the occurrence record unique?
  assertr::verify(assertr::is_uniq(key)) %>%
  # are there any NA's in lat/lon?
  assertr::verify(assertr::not_na(decimalLatitude)) %>%
  assertr::verify(assertr::not_na(decimalLongitude)) %>%
  assertr::chain_end() %>%
  tibble::as_tibble()
```

Data integration: code

```
dat <- dat[,c("name", "decimalLongitude", "decimalLatitude")]
mapr::map_leaflet(dat, lon = "decimalLongitude",
  lat = "decimalLatitude")
```



Thanks!

Scott Chamberlain

sckottie

slides: scotttalks.info/dataone19

Karthik Ram

_inundata

rOpenSci: ropensci.org