ROpenSci

Fostering open science and open data with R

Karthik Ram

Made possible by:

Alfred P. Sloan Foundation

ropensci.github.io/ropensci_intro/

“ Most science is not reproducible or repeatable, even within the same lab group over time.

A reproducibility crisis

The Economist

OCTOBER 19TH–25TH 2013 Economist.com

Washington's lawyer surplus

How to do a nuclear deal with Iran

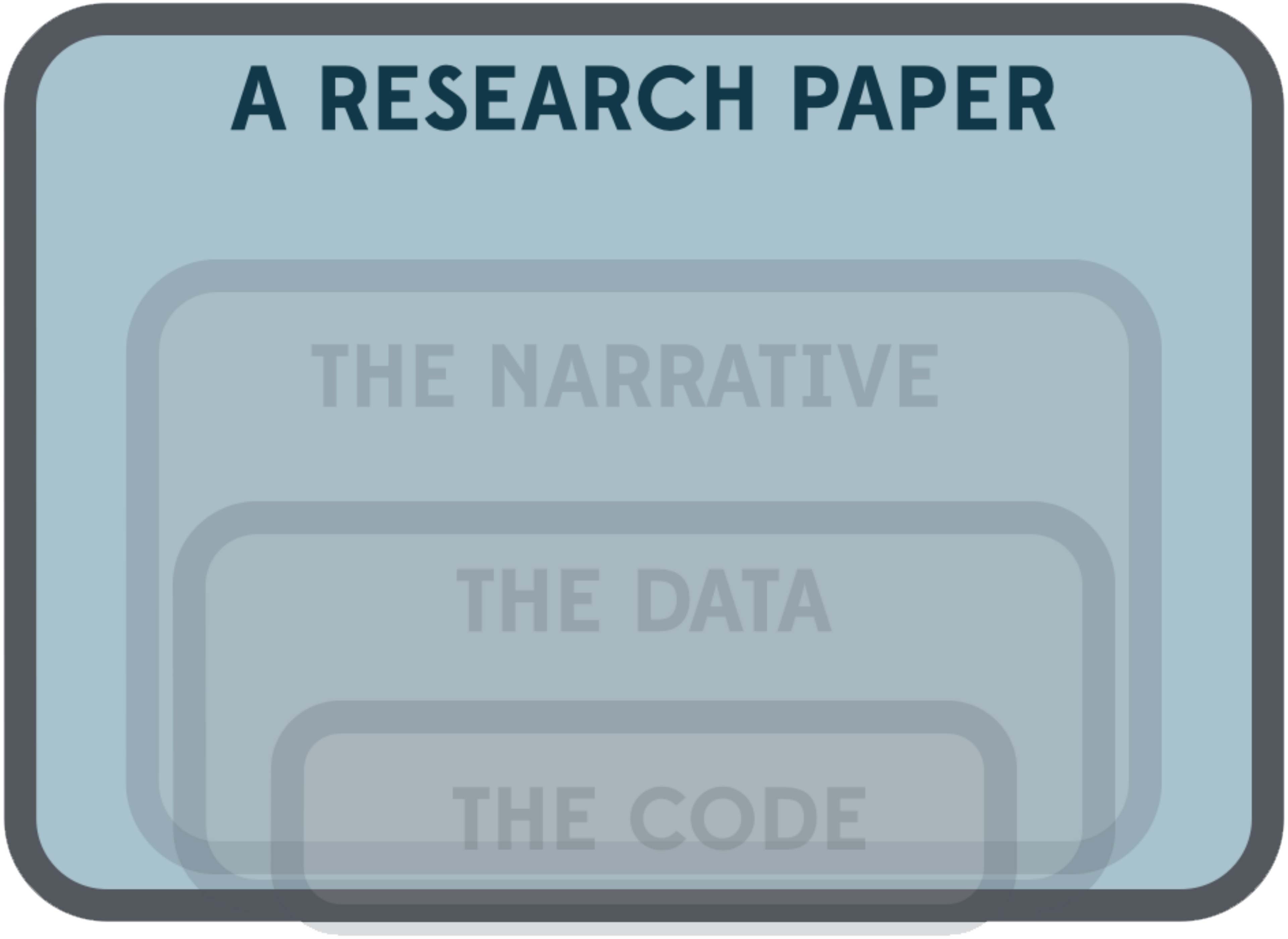
Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

The scientific workflow



A RESEARCH PAPER

THE NARRATIVE

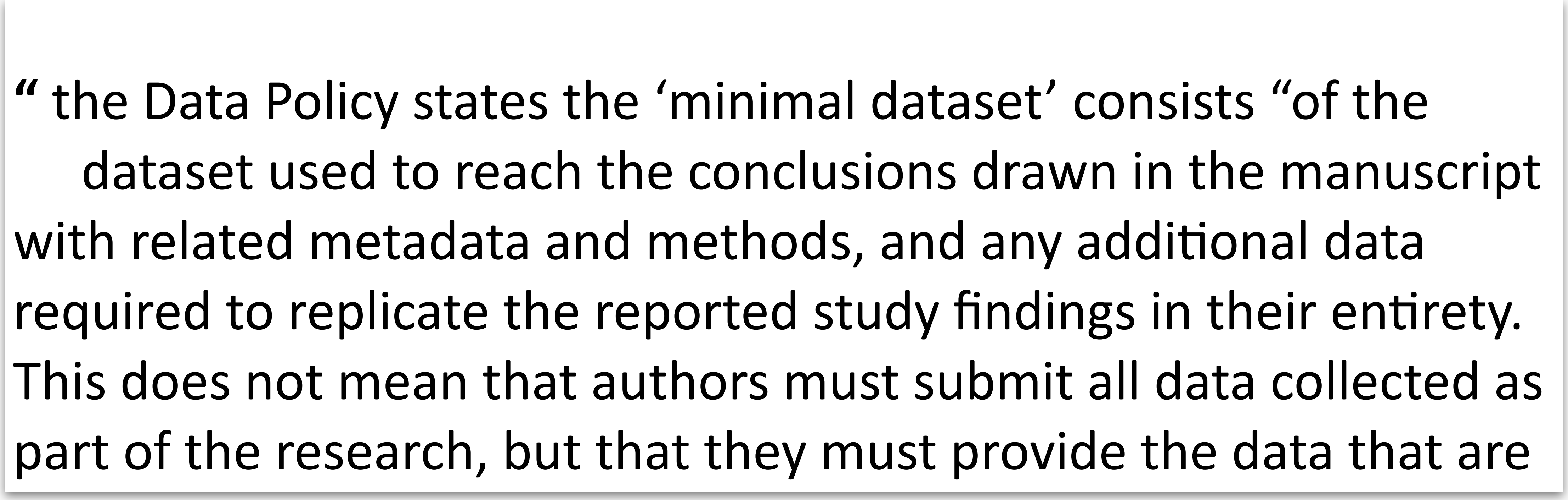
THE DATA

THE CODE

> “ Instructions for preparation of the Biographical Sketch have been revised to rename the Publications section to "Products” and amend terminology and instructions accordingly. This change makes clear that products may include, but are not limited to, publications, data sets, software, patents, and copyrights.

Issuance of a new NSF Proposal & Award Policies and Procedures Guide (October 4th)

PLOS Data Policy



“ the Data Policy states the ‘minimal dataset’ consists “of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. This does not mean that authors must submit all data collected as part of the research, but that they must provide the data that are

ROpenSci

“Enable access to scientific data repositories, full-text of articles, and science metrics and also facilitate a culture shift in the scientific community.

ROpenSci

More info @ ropensci.org/packages

Data

Treebase,
Fishbase,
Flybase,
GBIF, Vertnet
Dryad, ITIS
NPN, Taxize
opensnp

Journals

PLOS
Springer
textmine
pensoft

Data publication

figshare
Mendeley
DataONE,
rAltmetric, EML,
rNEXML

```
ted States","Ca
United States","Calif
tes","California",NA,NA,"h
ates","California",NA,NA,"htt
tates","California",NA,NA,"http
cus","United States","California
cus","United States","California
United States","California",195
tates","California",NA,NA,"htt
tes","California",1953-09-0
 States","Michigan",195
ited States","Col
```

Access all available ecological data

Liberating 400+ million observation records

Full text 100k articles

Data from papers in > 200 journals

Accessing data behind papers – dryad

```
library(rdryad)
dryaddat <- download_url(“10.255/dryad.1759”)
# Get a file given the URL
file  <- dryad_getfile(dryaddat)
dim(file)
```

```
## [1] 131 30
```

Dataset

Biodiversity records from museums

```
library(ecoengine)
pinus_data <- ee_observations(genus = “Pinus”, georeferenced =
TRUE, page = 1:25)
nrow(pinus_data$data)
```

```
## [1] 625
```

Get complete and current taxonomic records

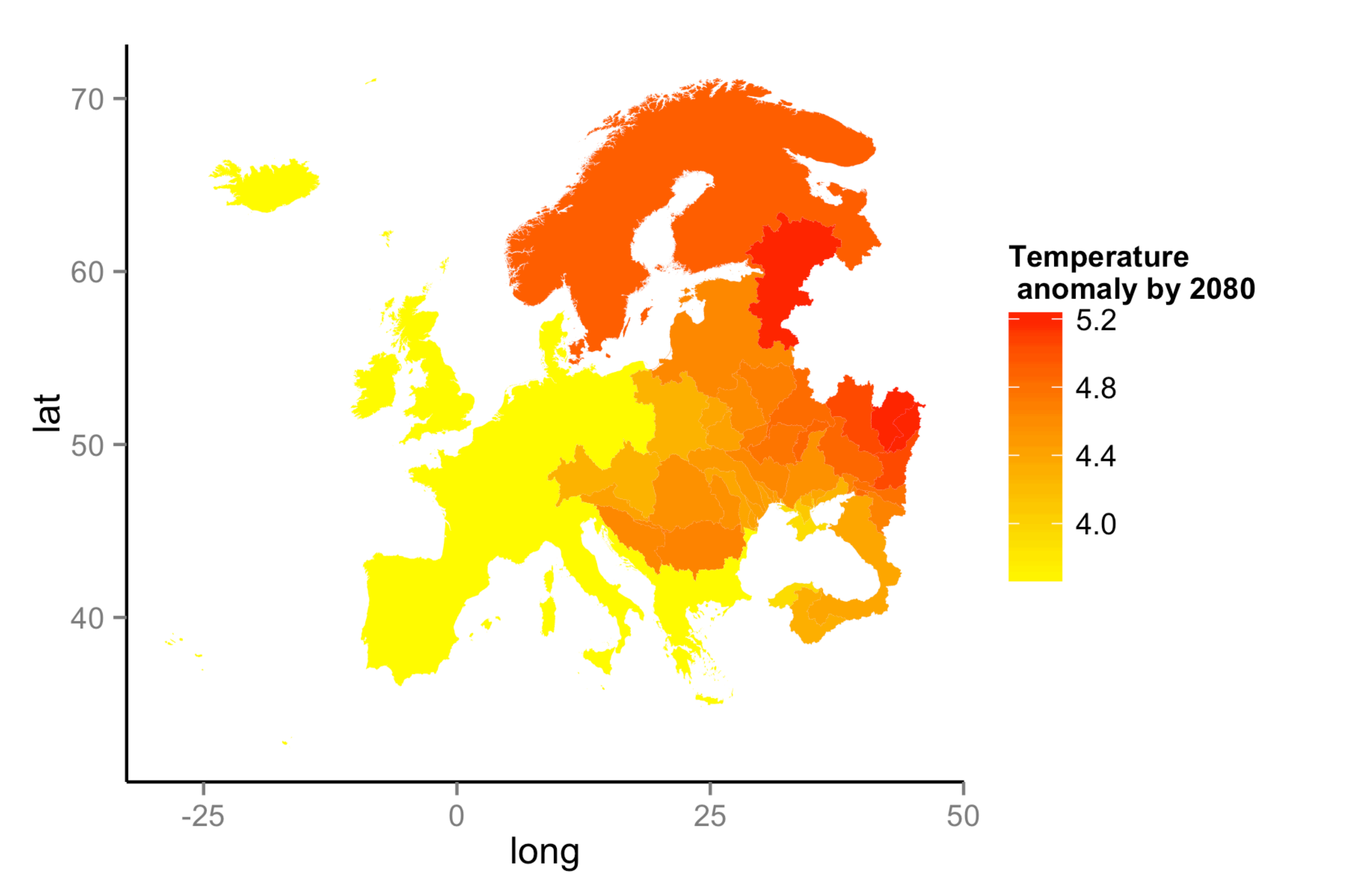
```
classification(c(“Helianthus annuus”), db = “ncbi”)
```

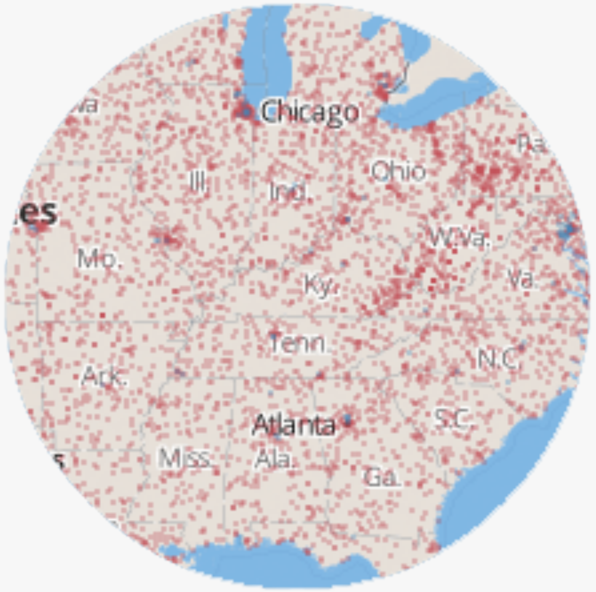
```
## $`Helianthus annuus`
##                  name         rank
## 1   cellular organisms      no rank
## 2             Eukaryota superkingdom
## 3         Viridiplantae      kingdom
## 4          Streptophyta       phylum
## 5        Streptophytina      no rank
... cutoff
## attr(,"class")
## [1] "classification"
## attr(,"db")
## [1] "ncbi"
```

World Bank climate knowledge portal

rWBclimate

```
library(rWBclimate)
eu_basin <- create_map_df(Eur_basin)
eu_basin_dat <- get_ensemble_temp(Eur_basin, “annualanom”, 2080, 2100)
```



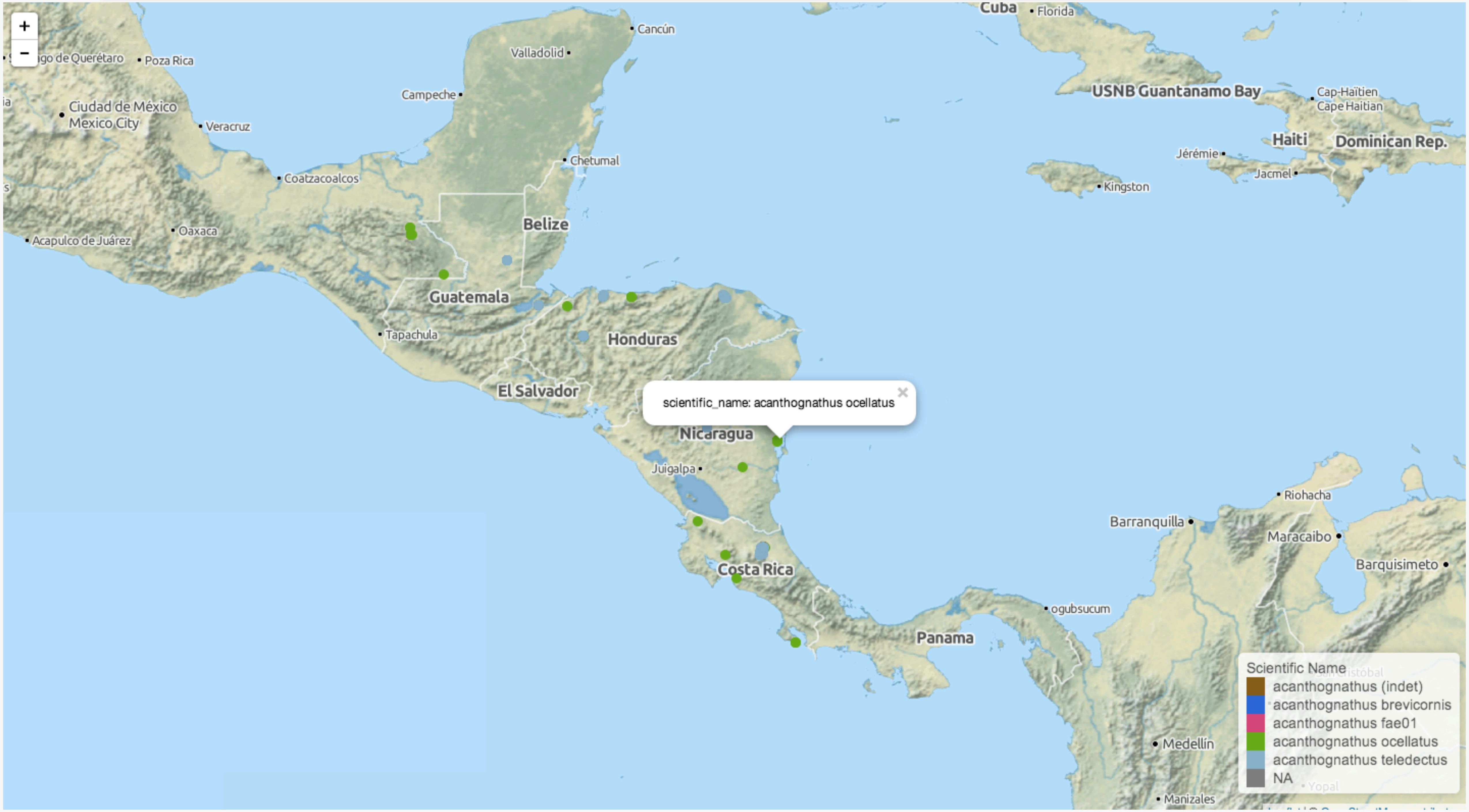


Interactively visualize and analyze data

Explore these data interactively, including any plots you might make

Taxon specific databases - AntWeb

```
library(AntWeb)
acdd <- aw_data(genus = “acanthognathus”)
aw_map(acd)
```

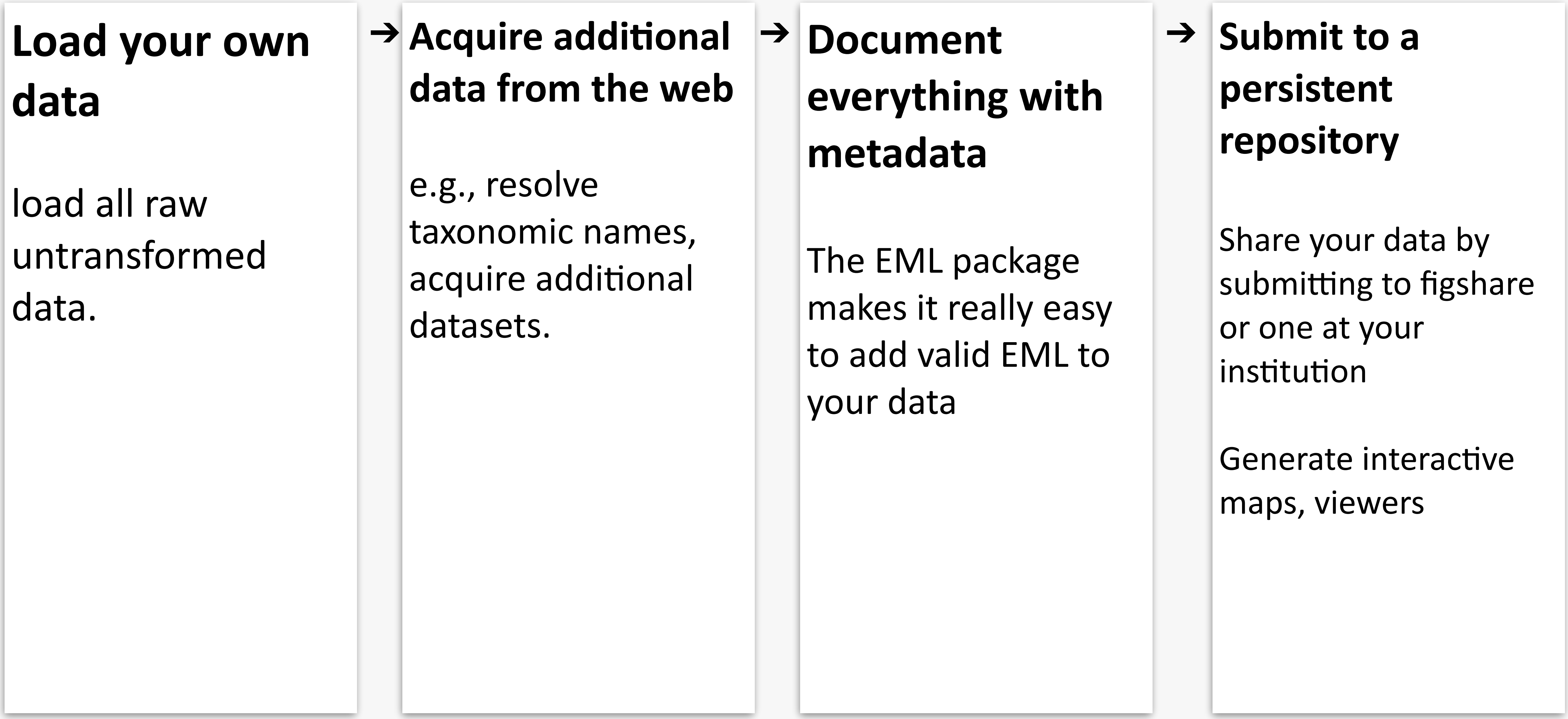


Sharing data – (figshare)

Using figshare's API it is possible to share figures, data and any other object generated in R and obtain a data citation.

```
library(rfigshare)
id <- fs_create(“Fisheries dataset”, “A dataset containing
catch for 4 important commercial fish species”, “dataset”)
fs_upload(id, “dat.csv”)
```

A reproducible workflow (in R)



Load your own data

load all raw untransformed data.

→ **Acquire additional data from the web**

e.g., resolve taxonomic names, acquire additional datasets.

→ **Document everything with metadata**

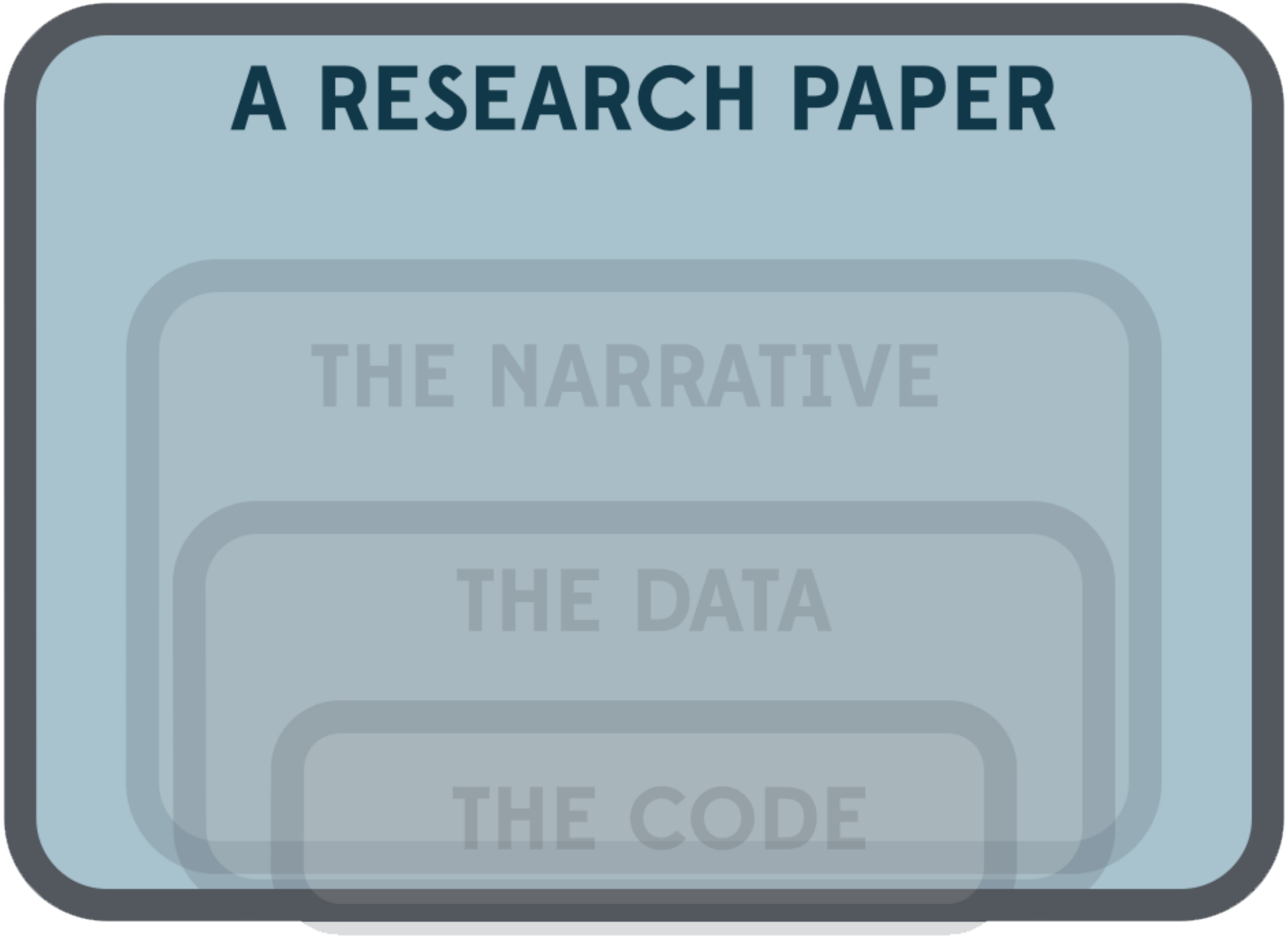
The EML package makes it really easy to add valid EML to your data

→ **Submit to a persistent repository**

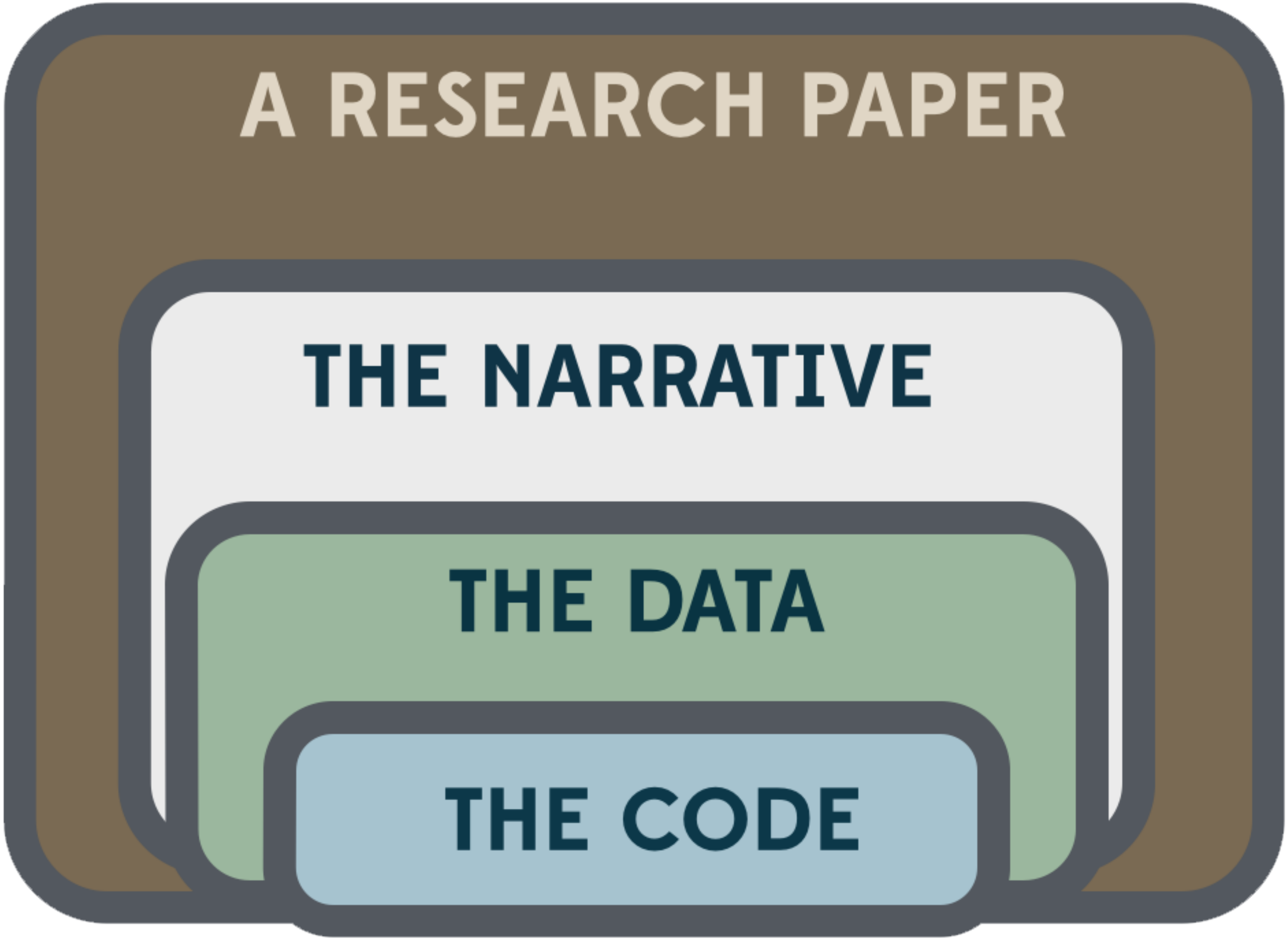
Share your data by submitting to figshare or one at your institution

Generate interactive maps, viewers

The scientific workflow



The open scientific workflow



A RESEARCH PAPER

THE NARRATIVE

THE DATA

THE CODE

ropensci.org

ropensci on GitHub

@ropensci on Twitter

Questions or comments to: info@ropensci.org