Layers of Reproducibility

Jonas Schöley

@jschoeley
0000-0002-3340-8518
schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Benefits of reproducibility and open science

Some personal experiences

LogrankA

Together with

Roland **Rau**

@Demographie

CRAN Task View: Survival Analysis

Maintainer:	Arthur Allignol, Aurelien Latouche
Contact:	arthur.allignol at gmail.com
Version:	2022-03-07
URL:	https://CRAN.R-project.org/view=Survival
Source:	https://github.com/cran-task-views/Survival/

Testing

- The `survdiff` function in survival compares survival curves using the Fleming-Harrington G-rho family of test. NADA implements this class of tests for left-censored data.
- The maxcombo package compares survival curves using the max-combo test, which is often based on the Fleming-Harrington G-rho family of tests and is designed to have higher power than the logrank test in the scenario of non-proportional hazards such as those resulting from delayed treatment effects.
- clinfun implements a permutation version of the logrank test and a version of the logrank that adjusts for covariates.
- The exactRankTests implements the shift-algorithm by Streitberg and Roehmel for computing exact conditional p-values and quantiles, possibly for censored data.
- `SurvTest` in the coin package implements the logrank test reformulated as a linear rank test.
- The maxstat package performs tests using maximally selected rank statistics.
- The interval package implements logrank and Wilcoxon type tests for interval-censored data.
- Three generalised logrank tests and a score test for interval-censored data are implemented in the glrt *(archived)* package.
- survcomp compares 2 hazard ratios.
- The TSHRC implements a two stage procedure for comparing hazard functions.
- The FHtest package offers several tests based on the Fleming-Harrington class for comparing surival curves with right- and interval-censored data.
- The LogrankA *(archived)* package provides a logrank test for which aggregated data can be used as input.
- The short term and long term hazard ratio model for two samples survival data can be found in the YPmodel package.
- The controlTest implements a nonparametric two-sample procedure for comparing the median survival time.
- The survRM2 package performs two-sample comparison of the restricted mean survival time
- The emplik2 package permits to compare two samples with censored data using empirical likelihood ratio tests.
- The KONPsurv package provides powerful nonparametric K-sample tests for right-censored data. The tests are consistent against any differences between the hazard functions of the groups.

Together with

Ilya **Kashnitsky**
@ikashnitsky

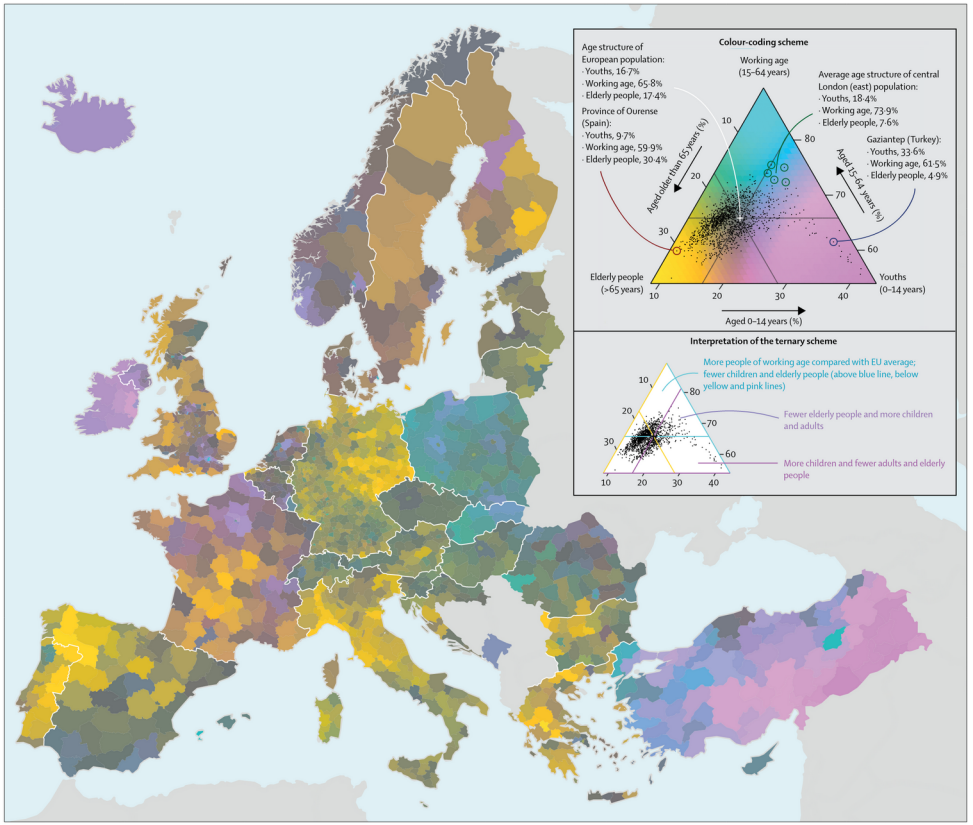


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky

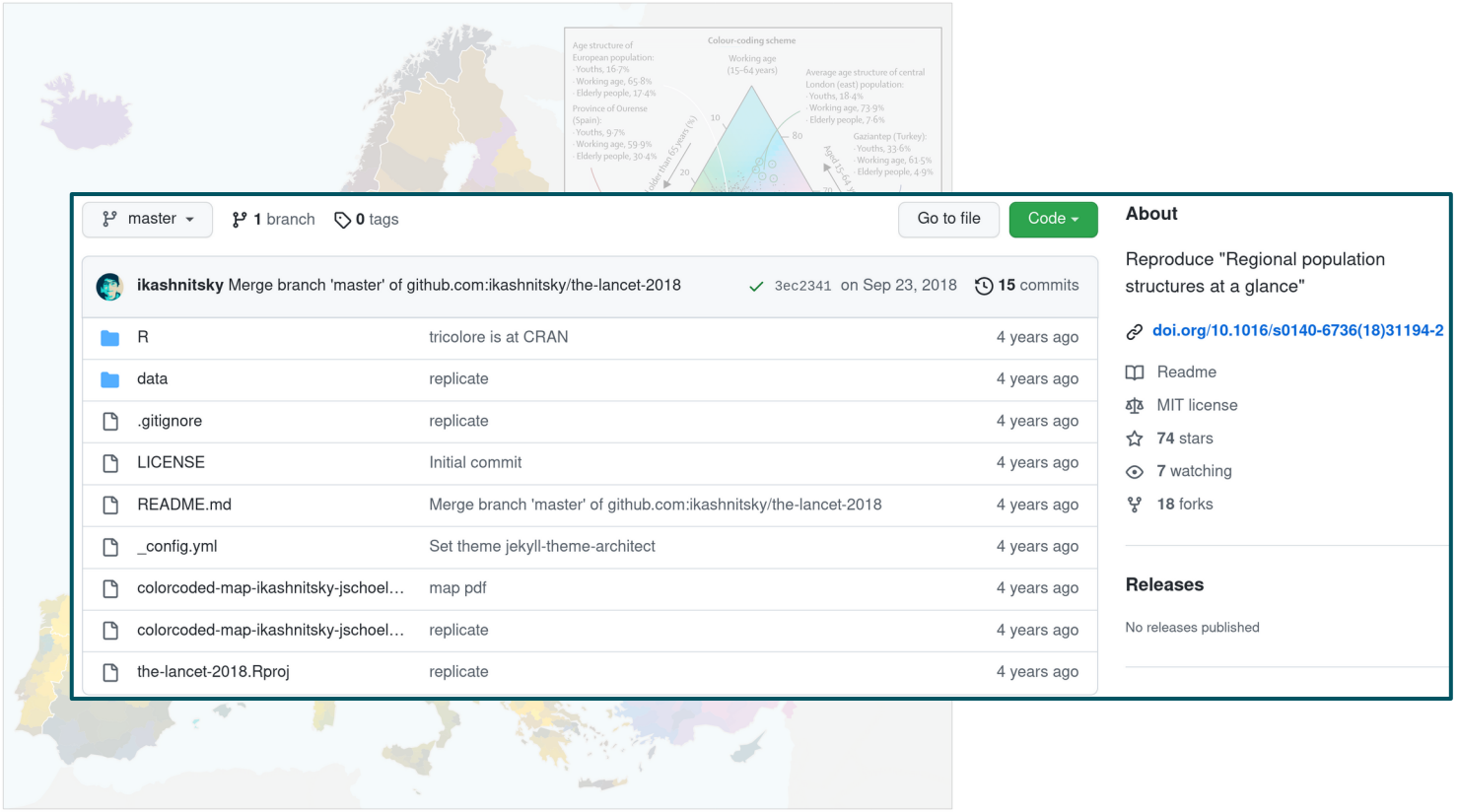


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky

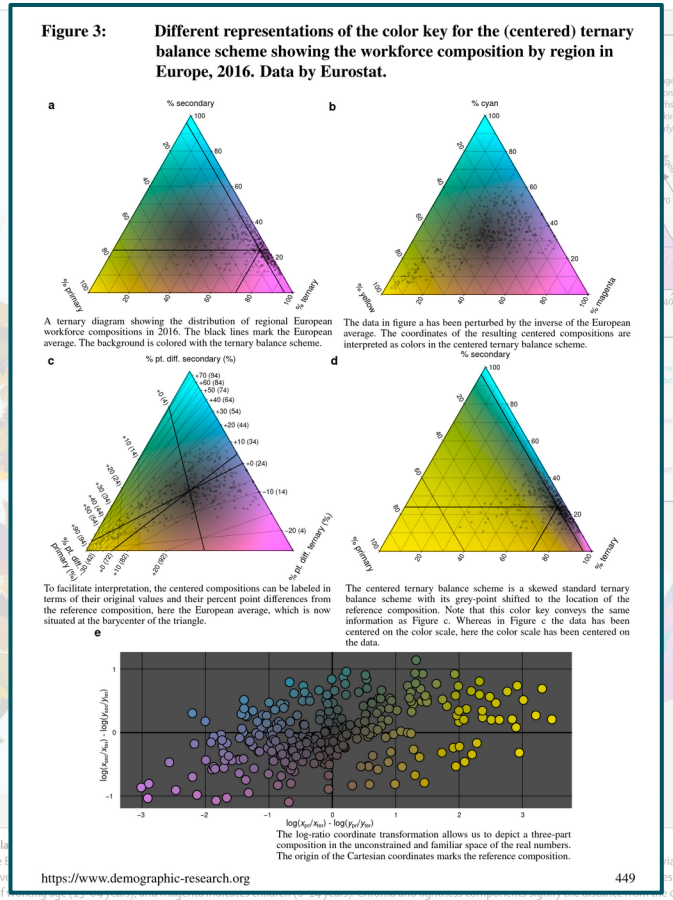


Figure 3: **Different representations of the color key for the (centered) ternary balance scheme showing the workforce composition by region in Europe, 2016. Data by Eurostat.**

a

A ternary diagram showing the distribution of regional European workforce compositions in 2016. The black lines mark the European average. The background is colored with the ternary balance scheme.

b

The data in figure a has been perturbed by the inverse of the European average. The coordinates of the resulting centered compositions are interpreted as colors in the centered ternary balance scheme.

c

To facilitate interpretation, the centered compositions can be labeled in terms of their original values and their percent point differences from the reference composition, here the European average, which is now situated at the barycenter of the triangle.

d

The centered ternary balance scheme is a skewed standard ternary balance scheme with its grey-point shifted to the location of the reference composition. Note that this color key conveys the same information as Figure c. Whereas in Figure c the data has been centered on the color scale, here the color scale has been centered on the data.

e

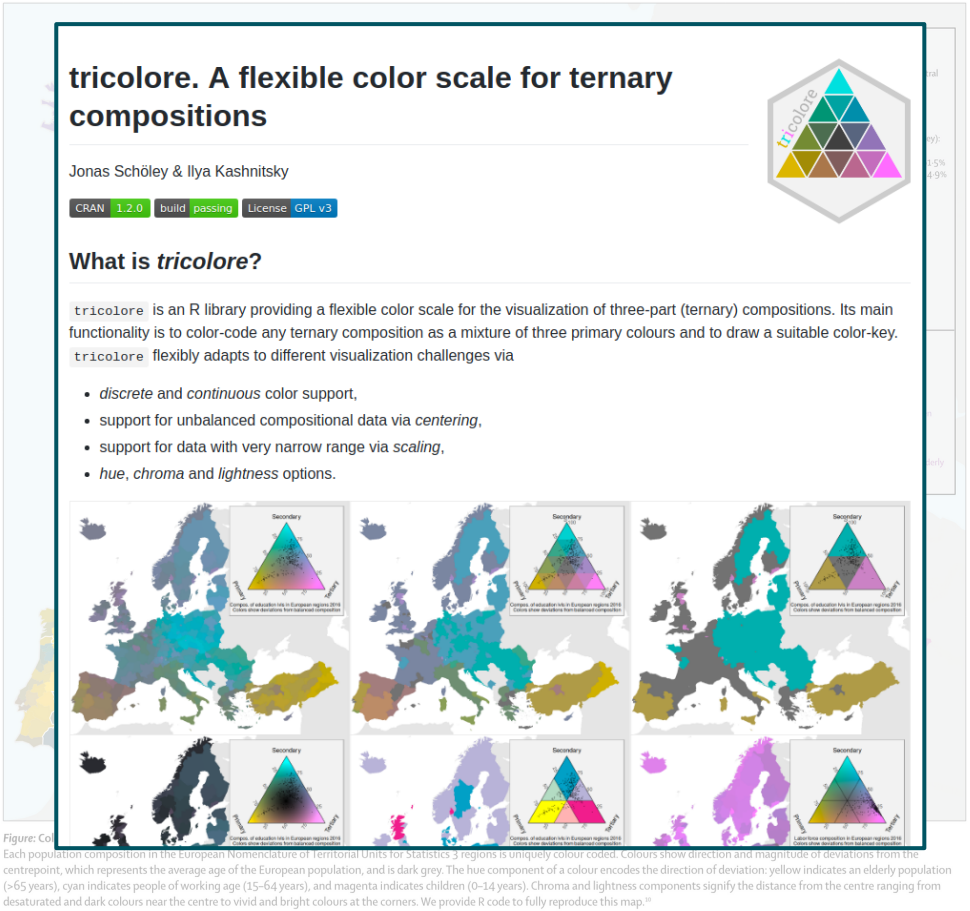
The log-ratio coordinate transformation allows us to depict a three-part composition in the unconstrained and familiar space of the real numbers. The origin of the Cartesian coordinates marks the reference composition.

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky



tricolore. A flexible color scale for ternary compositions

Jonas Schöley & Ilya Kashnitsky

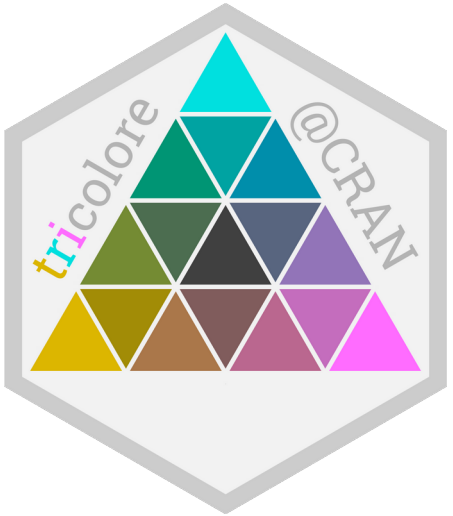
CRAN 1.2.0 | build passing | License GPL v3

What is *tricolore*?

`tricolore` is an R library providing a flexible color scale for the visualization of three-part (ternary) compositions. Its main functionality is to color-code any ternary composition as a mixture of three primary colours and to draw a suitable color-key. `tricolore` flexibly adapts to different visualization challenges via

- *discrete* and *continuous* color support,
- support for unbalanced compositional data via *centering*,
- support for data with very narrow range via *scaling*,
- *hue*, *chroma* and *lightness* options.

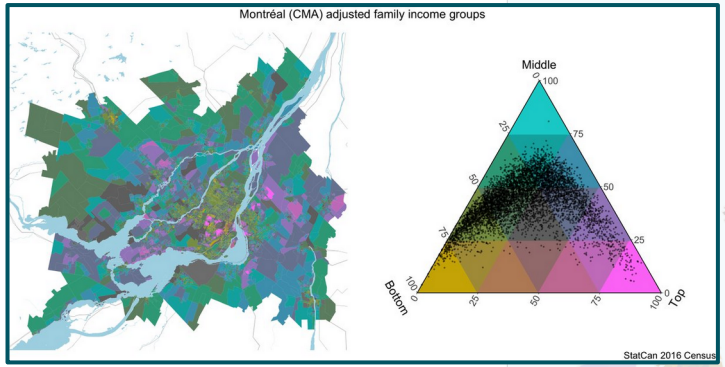
Figure: Co
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of deviations from the centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encodes the direction of deviation: yellow indicates an elderly population (>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightness components signify the distance from the centre ranging from desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.[10]



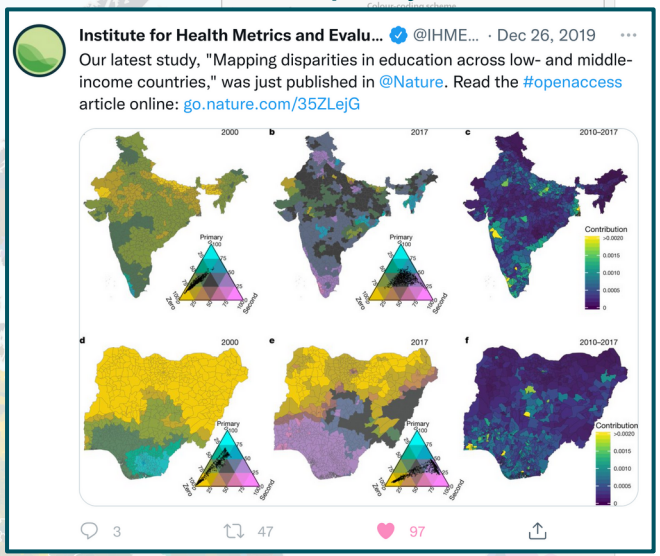
Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

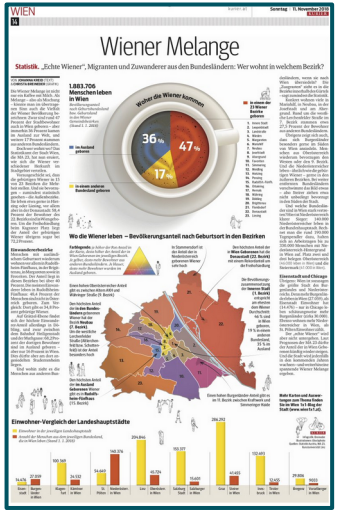
Income distribution in Canadian cities



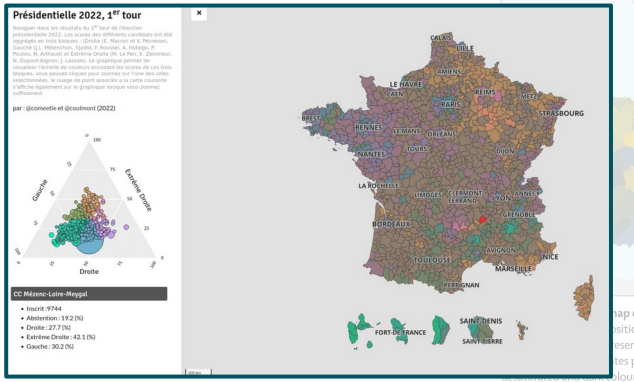
LMIC education disparity



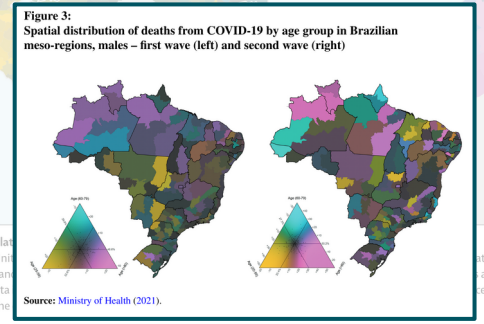
Vienna's population by origin



French election results



Regional age distribution of COVID deaths In Brazil



Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

tricolore

Agricultural and Forest Meteorology

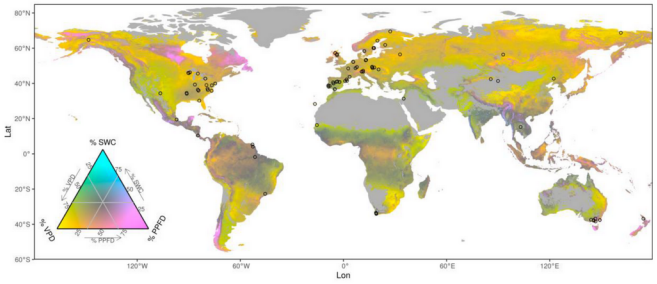


Fig. 3. Relative importance (relative R^2) of the three hydrometeorological drivers of transpiration regulation. Relative R^2 were calculated at each cell dividing the projection of each three drivers partial R^2 from the FULL model by the sum of the three partial R^2 at the same cell. Partial R^2 were projected at the global scale using linear models with climate, soil and vegetation structural variables as explanatory variables (Table S6). Grid colours were calculated using the 'tricolore' package (Schöley and Kashnitsky, 2020) for each cell. Colour gradient indicate the relative importance of the three hydrometeorological constraints. Light grey colour are non-forested areas or with vegetation not taller than 0.5 m (Simard et al., 2011). % VPD: vapour pressure deficit relative importance. % SWC: soil water content relative importance. % PPFD: photosynthetic photon flux density relative importance. Points indicate locations of study sites.

Soil composition

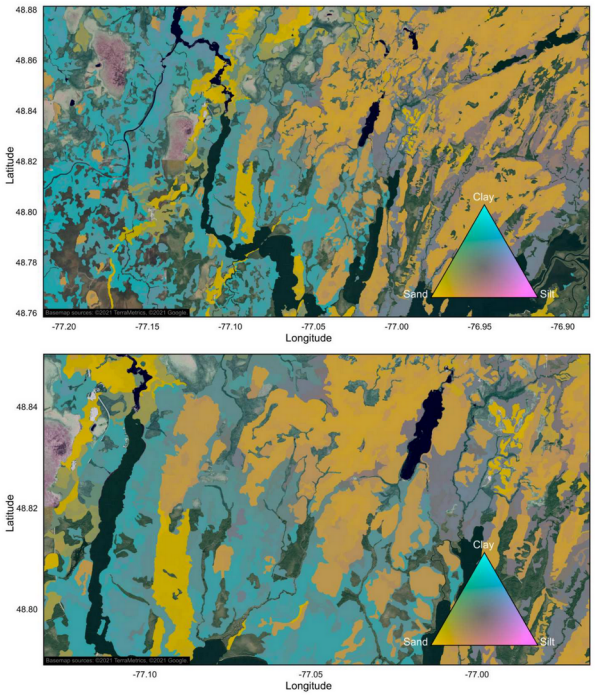


Figure 8 **Maps of soil texture composition at two zoom levels, for a region at the southern edge of the Abitibi and James Bay Lowlands soil province.** Only productive forest land characterized by mineral soils was mapped. Agricultural and unproductive forest land, organic soils, anthropogenic infrastructures, and water areas were excluded. Maps were produced with QGIS software, version 3.4 (*QGIS, 2020*). Basemap credit: ©2021 TerraMetrics, ©2021 Google.

Full-size DOI: 10.7717/peerj.11685/fig-8

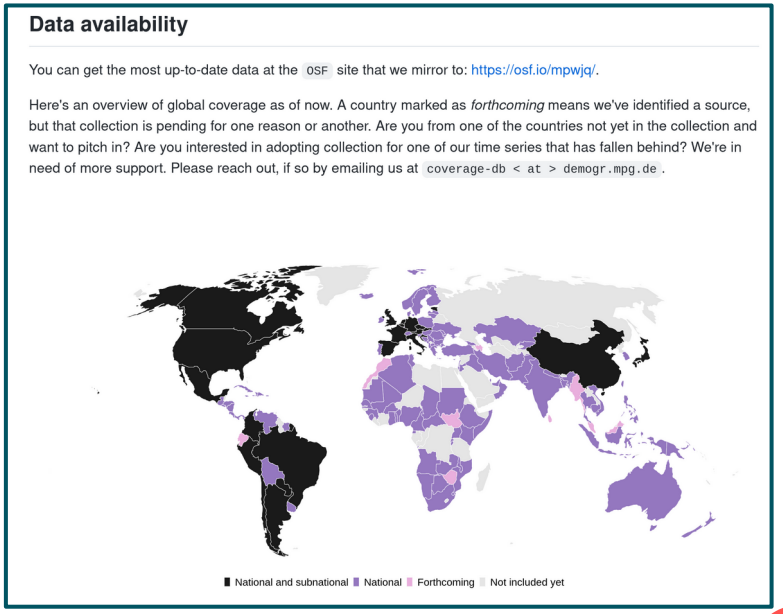
Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 20**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Col
centrepoint, which represents the average age of the European population, and is dark grey. The hue component of a colour encode
(>65 years), cyan indicates people of working age (15–64 years), and magenta indicates children (0–14 years). Chroma and lightnes
desaturated and dark colours near the centre to vivid and bright colours at the corners. We provide R code to fully reproduce this map.

Kashnitsky & Schöley (2018). Regional population structures at a glance. 10.1016/S0140-6736(18)31194-2

Reproducible Demographers

Tim **Riffe** @timriffe1 & Enrique **Acosta** @Acosta_Kike_ &
Manal **Elzalabany** & Maxi **Kniffka** @MaxiKniffka & Jessica **Donzowa** @jdonzowa

Created a fully reproducible data base of age specific COVID-19 statistics.



Reproducible Demographers

Christina **Bohk-Ewald**

- Method00_FreezeRates
- Method01_Hadwiger1940
- Method02_CoaleMcNeil1972
- Method03_CoaleTrussell1974
- Method04_Brass1974
- Method05_Evans1986
- Method06_Chandola1999
- Method07_Schmertmann2003
- Method08_PeristeraKostaki2007M1
- Method09_PeristeraKostaki2007M2
- Method10_MyrskylaGoldstein2013
- Method11_Saboia1977
- Method12_WillekensBaydar1984
- Method13_deBeer1985and1989
- Method14_Lee1993Log
- Method16_HyndmanUllah2007
- Method17_ChengLin2010
- Method18_Myrskyla2013
- Method22_LiWu2003

Bohk et al. (2018). Forecast accuracy hardly improves with methods complexity when completing cohort fertility. 10.1109/5.771073

Implemented, shared and compared 22 fertility forecasting methods

github.com/fertility-forecasting/validate-forecast-methods/tree/master/basic-scripts-forecast-methods

Reproducible Demographers

Rob Hyndman

Author of countless R packages widely applied in demography.

`demography: Forecasting Mortality, Fertility, Migration and Population Data`

Functions for demographic analysis including lifetable calculations; Lee-Carter modelling; functional data analysis of mortality rates, fertility rates, net migration numbers; and stochastic population forecasting.

Version:	1.22
Depends:	R (≥ 3.4), forecast (≥ 8.5)
Imports:	ftsa (≥ 4.8), rainbow, cobs, mgcv, strucchange, RCurl
Published:	2019-04-22
Author:	Rob J Hyndman with contributions from Heather Booth, Leonie Tickle and John Maindonald.
Maintainer:	Rob J Hyndman <Rob.Hyndman at monash.edu>
BugReports:	https://github.com/robjhyndman/demography/issues
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	https://github.com/robjhyndman/demography
NeedsCompilation:	no
Materials:	README ChangeLog
CRAN checks:	demography results

github.com/robjhyndman

Reproducible Demographers

Diego **Alburez-Gutierrez**

Liliana **Calderón-Bernal**

Tom **Theile**

Amanda Martins **de Almeida**

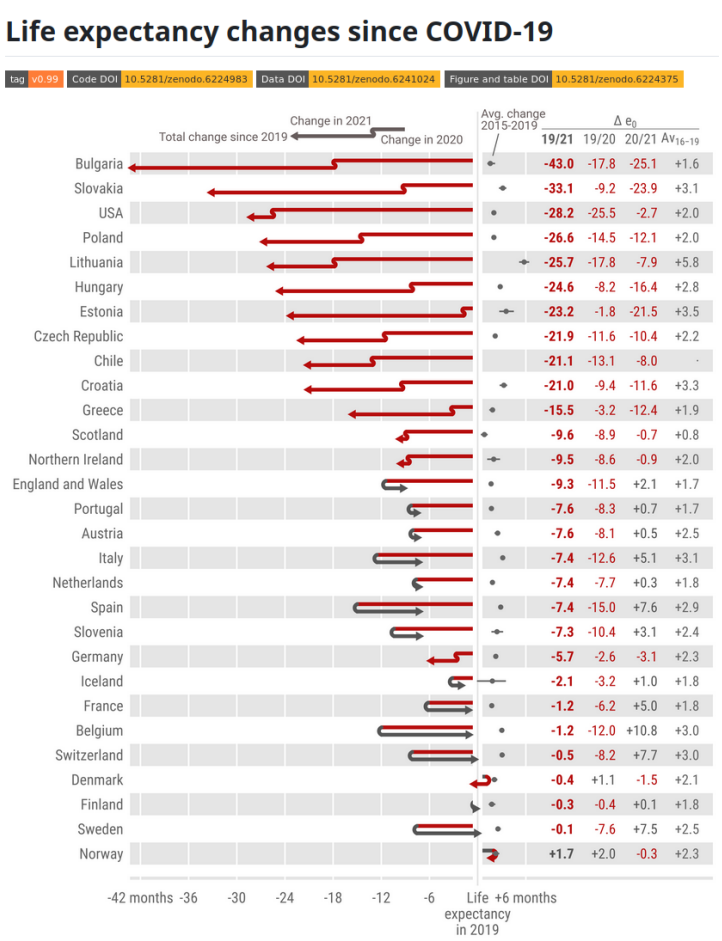
and many more...

High quality R libraries coming out of MPIDR

mpidr.github.io/rsocsim/

github.com/IvanWilli/DemoKin

Consuming open science



2.2 Data quality. *Data: I see that all data were downloaded on February 18, 2022. Can you provide more information about the level of completeness or comparability of completeness across countries as of this download date? I am only familiar with the US data, which have a very long lag time for all-cause mortality and so any counts obtained are likely an underestimate as of this date. Even if they technically cover deaths through the end of 2021, many deaths get reported and processed at later dates. I would also recommend updating these estimates with the most recently available data when doing the next revision.*

Consuming open science

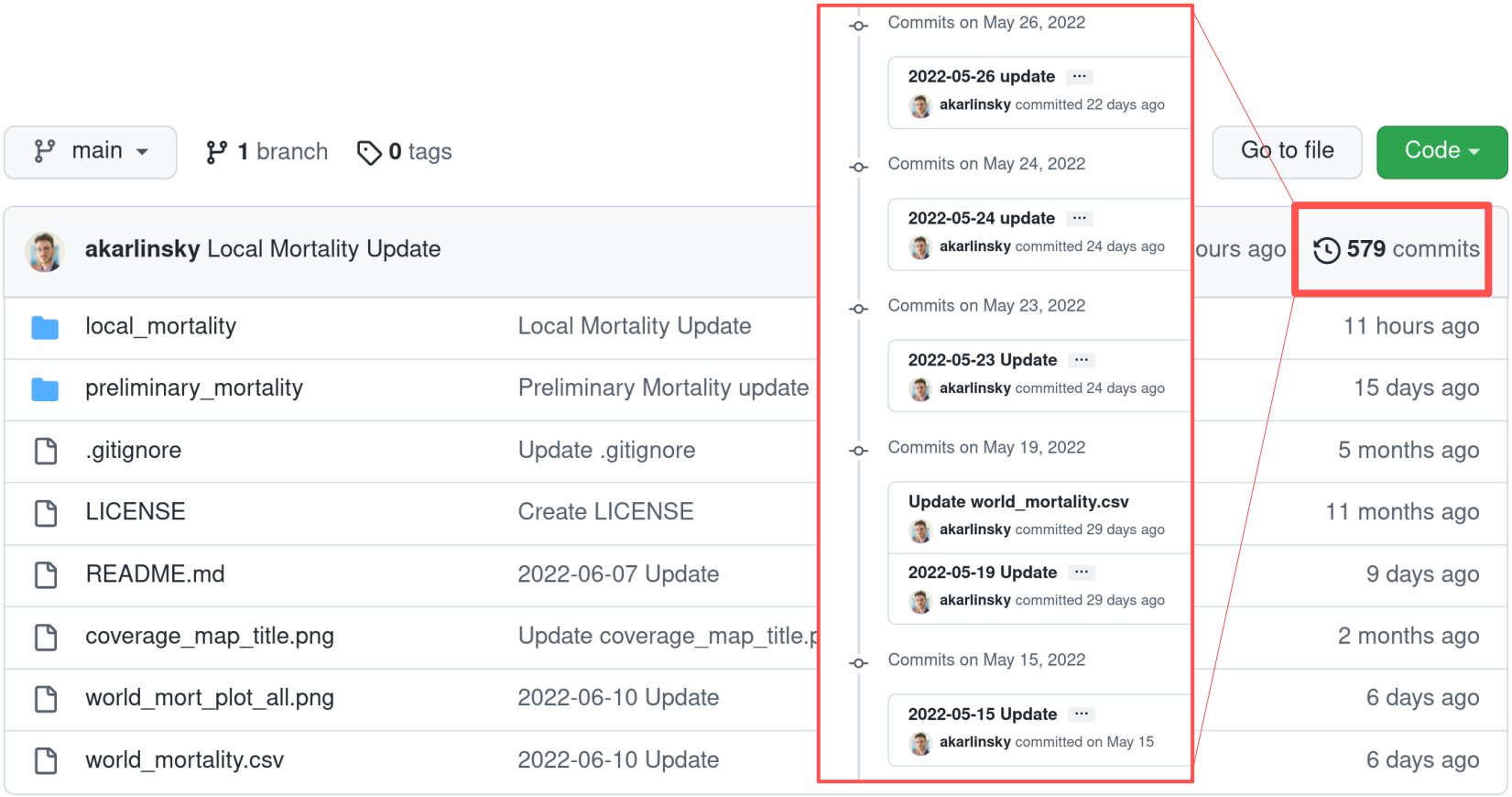
main | 1 branch | 0 tags | Go to file | Code

akarlinsky Local Mortality Update — d3a6d38 11 hours ago — **579** commits

local_mortality	Local Mortality Update	11 hours ago
preliminary_mortality	Preliminary Mortality update	15 days ago
.gitignore	Update .gitignore	5 months ago
LICENSE	Create LICENSE	11 months ago
README.md	2022-06-07 Update	9 days ago
coverage_map_title.png	Update coverage_map_title.png	2 months ago
world_mort_plot_all.png	2022-06-10 Update	6 days ago
world_mortality.csv	2022-06-10 Update	6 days ago

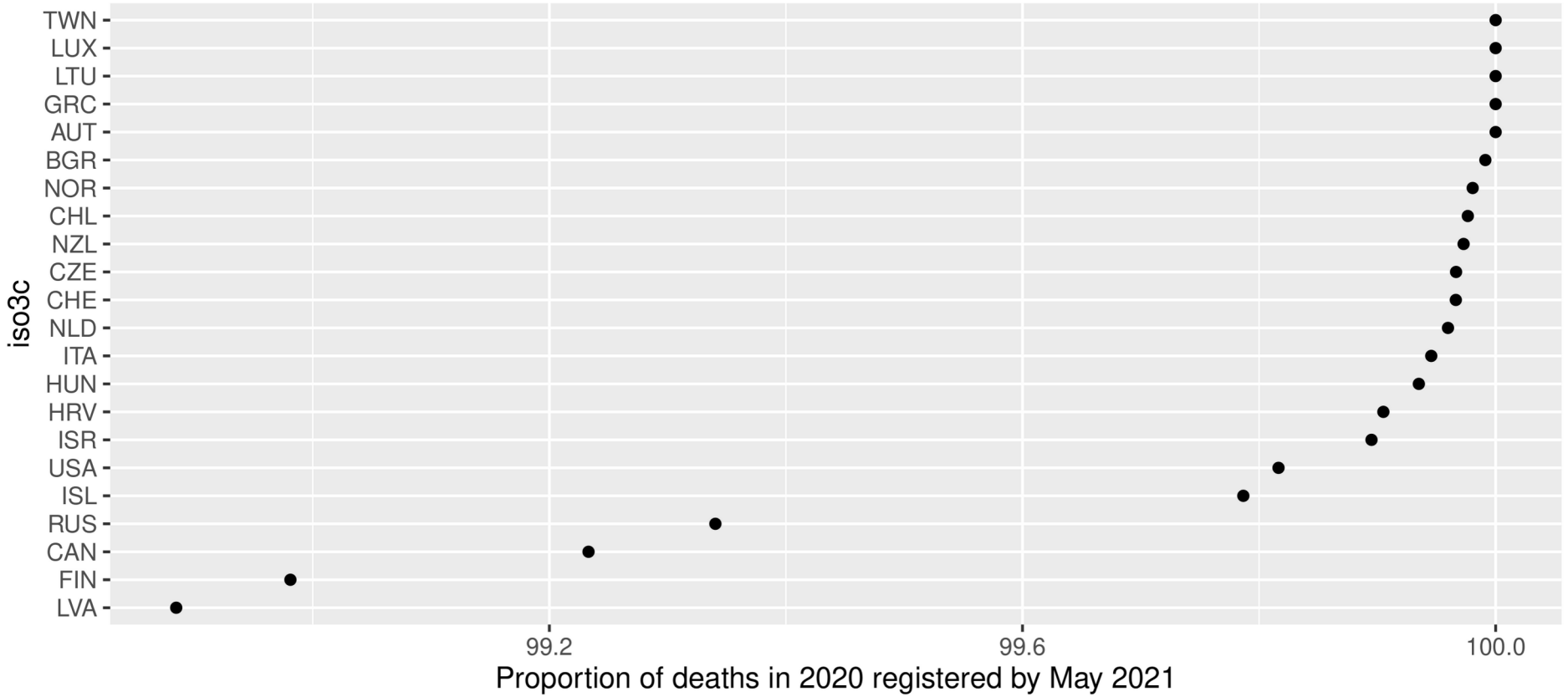
Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



Derived from Karlinsky & Kobak (2022). World Mortality Database.
github.com/akarlinsky/world_mortality

Institutionalization of open science

Aliakbar **Akbaritabar**

OPEN SCIENCE AMBASSADOR

Open science at Max Planck Society. osip.mpdl.mpg.de.

Reproducibility & replicability

Given **your data** and **your analysis**
I arrive at **your results**

– –

Given **your research question**,
my data and **my analysis**
I arrive at **your results**

4 Layers

Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

Layer 3: Computational

4 Layers

Layer 0: Aspirational

Realize you're having a problem

4 Layers

Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

Layer 3: Computational

Reproduce your own work

Project structure
Analysis pipeline
Documentation

4 Layers

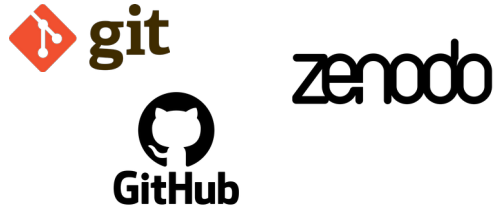
Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

Layer 3: Computational

Let others reproduce your work



4 Layers

Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

Layer 3: Computational

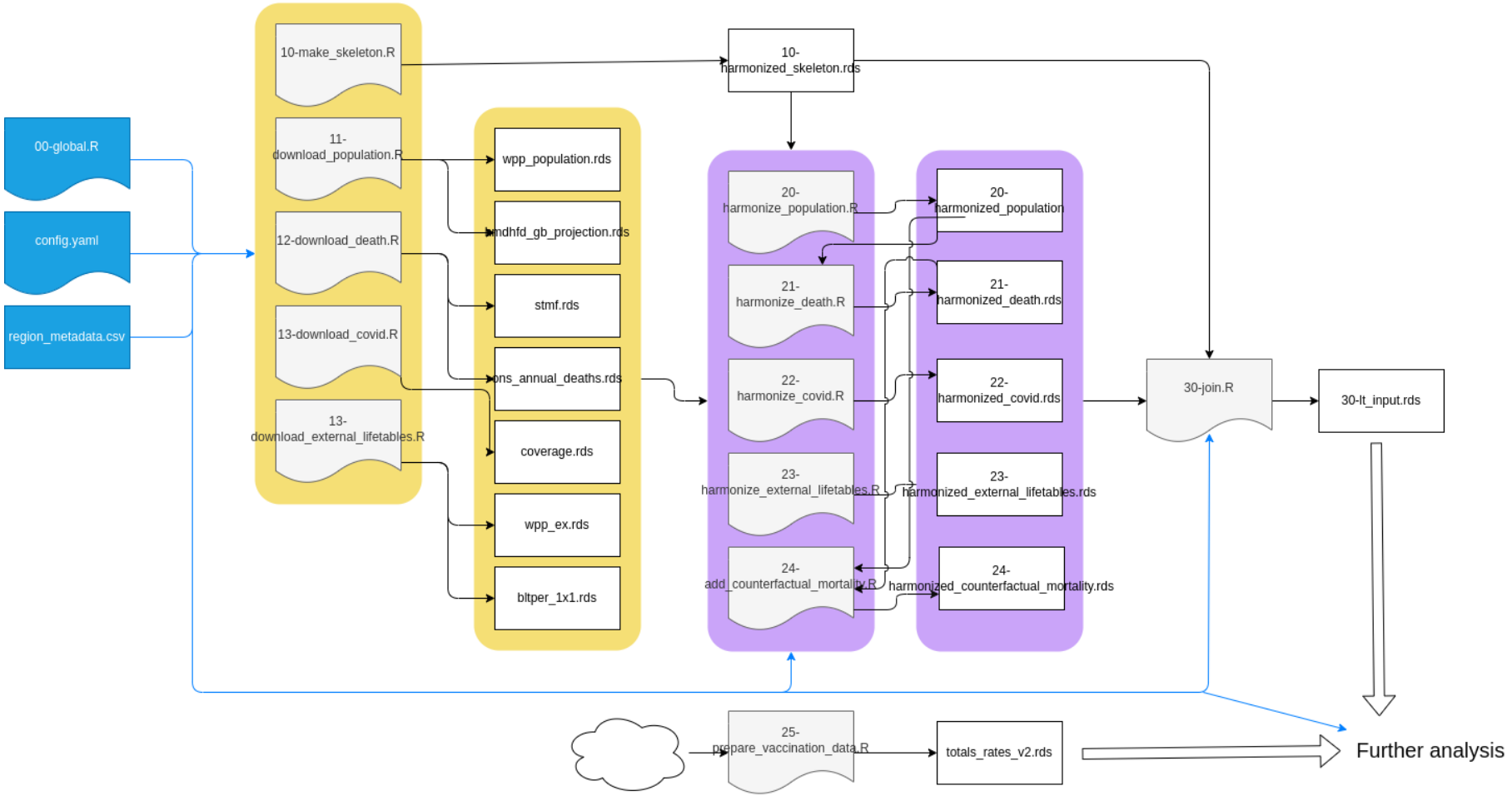
Automatize the reproduction of your work

Personal reproducibility

Project based workflow

Project structure

Code structure



A simple way to log package dependencies in R

```
# list packages used in project
# dput(unique(renv::dependencies()$Package))
packages <- c(
  'dplyr',
  'eurostat',
  'lubridate',
  'readr',
  'sf',
  'stringi',
  'tidyr',
  'yaml',
  'ggplot2',
  'rnaturalearth',
  'rnaturalearthdata',
  'ggtern',
  'scales'
)
# install/update packages
install.packages(packages)

# write out list of dependencies and currently used versions
versions <- installed.packages()[packages, c('Package', 'Version')]
write.csv(versions, './output/00-package_versions.csv', row.names = FALSE)
```

Communal reproducibility

Communal reproducibility

Ensure everyone can run your analysis

Share and version your analysis

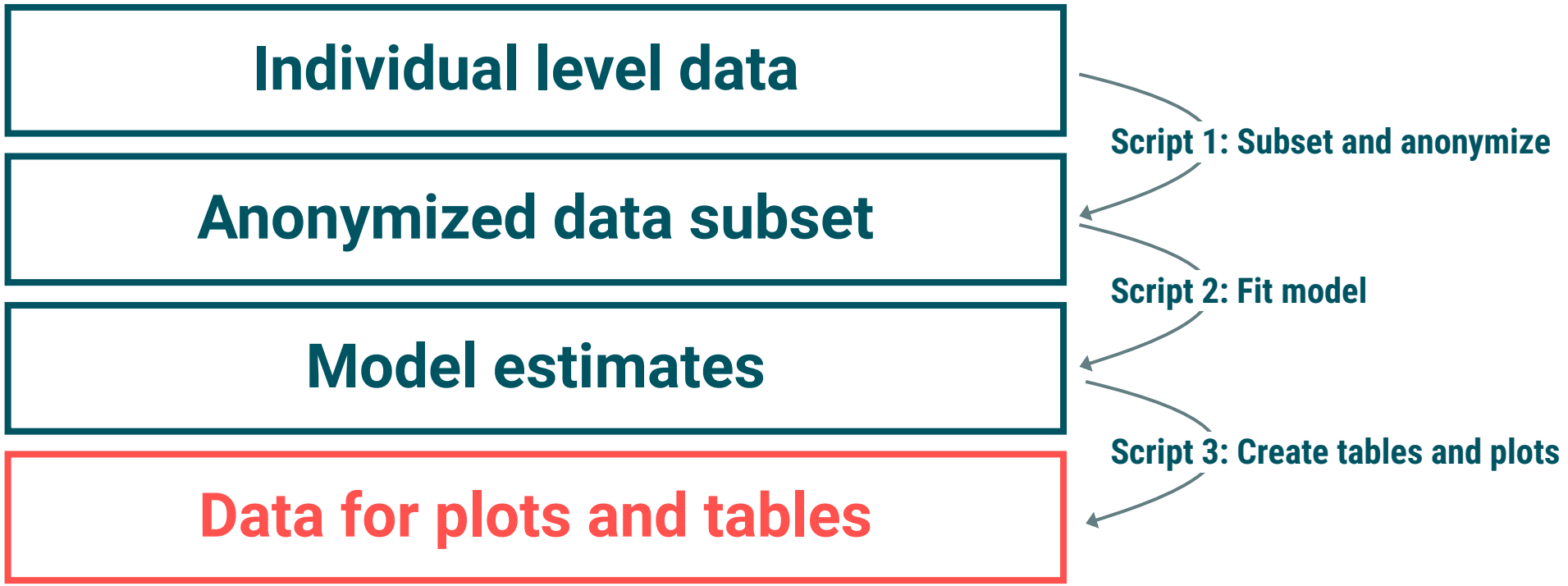
Share your data Archive your data Get DOIs

I’m not allowed to share my data

There is always a
Largest Shareable Derived Dataset

I’m not allowed to share my data

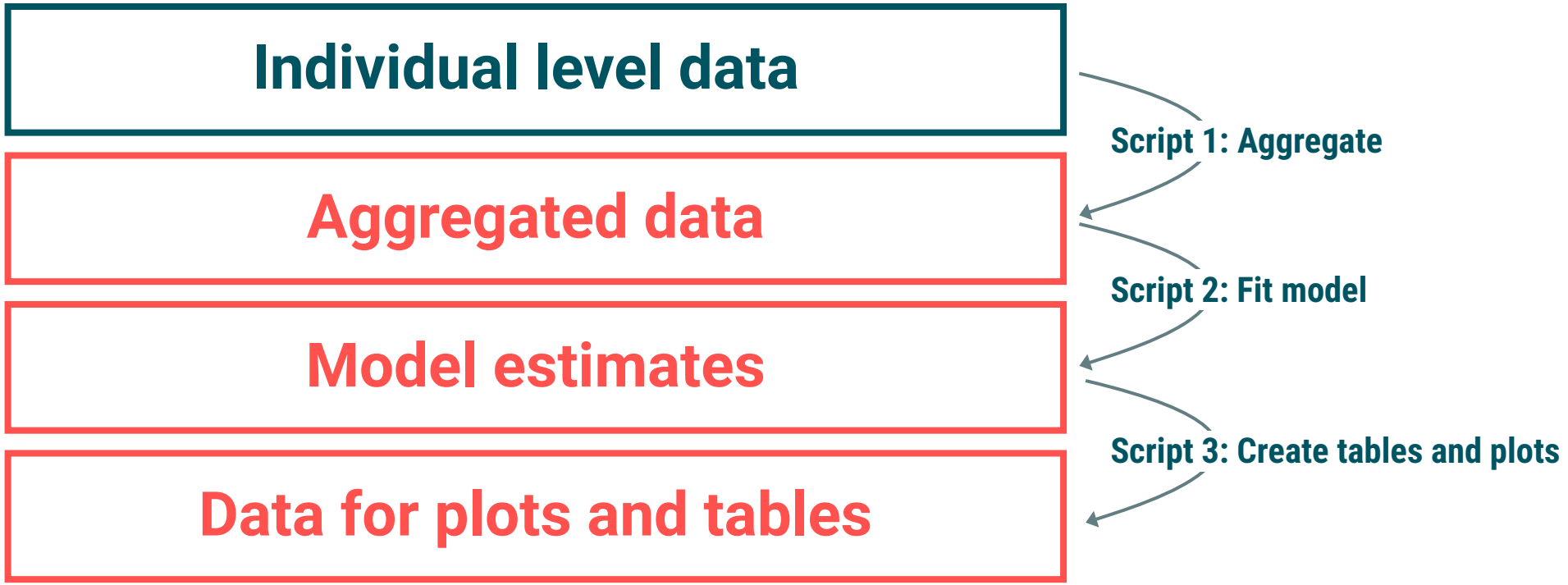
Identify your Largest Shareable Derived Dataset



Example for sharing plot and table data. github.com/jschoeley/e0deficit.

I’m not allowed to share my data

Identify your Largest Shareable Derived Dataset



Example for sharing aggregated data. github.com/jschoeley/inselect.

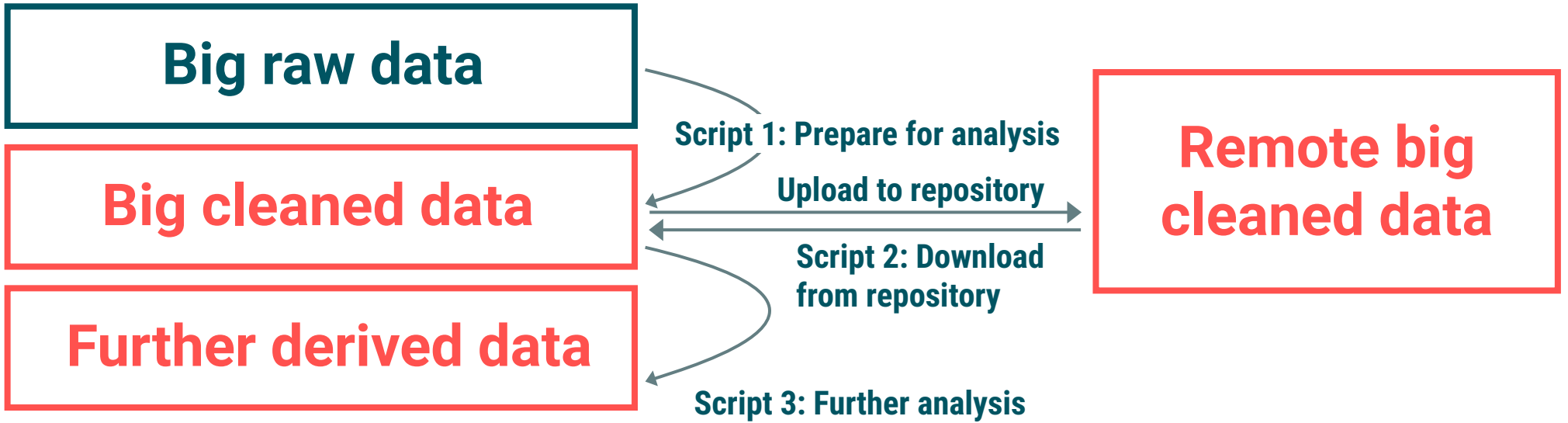
My data is too big to share

Upload to a dedicated data repository
Data outsourcing

My data is too big to share

Upload to a dedicated data repository

Data outsourcing



Example for data outsourcing. github.com/jschoeley/inselect.

Useful git commands

.gitignore

Computational reproducibility

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data

R Libraries

System dependencies

Hardware

git zenodo

renv

docker

The computational reproducibility stack

Ensure everyone can run your analysis

- Scripts & Data
- R Libraries
- System dependencies
- Hardware

git zenodo

renv

docker

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data


```
Warning message:
“`funs()` was deprecated in dplyr 0.8.0.
ℹ Please use a list of either functions or lambdas:

# Simple named list: list(mean = mean, median = median)

# Auto named with `tibble::lst()`: tibble::lst(mean, median)

# Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE
ℹ The deprecated feature was likely used in the dataiku package.
  Please report the issue to the authors.”
```

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data

R Libraries

System dependencies

Hardware

git zenodo

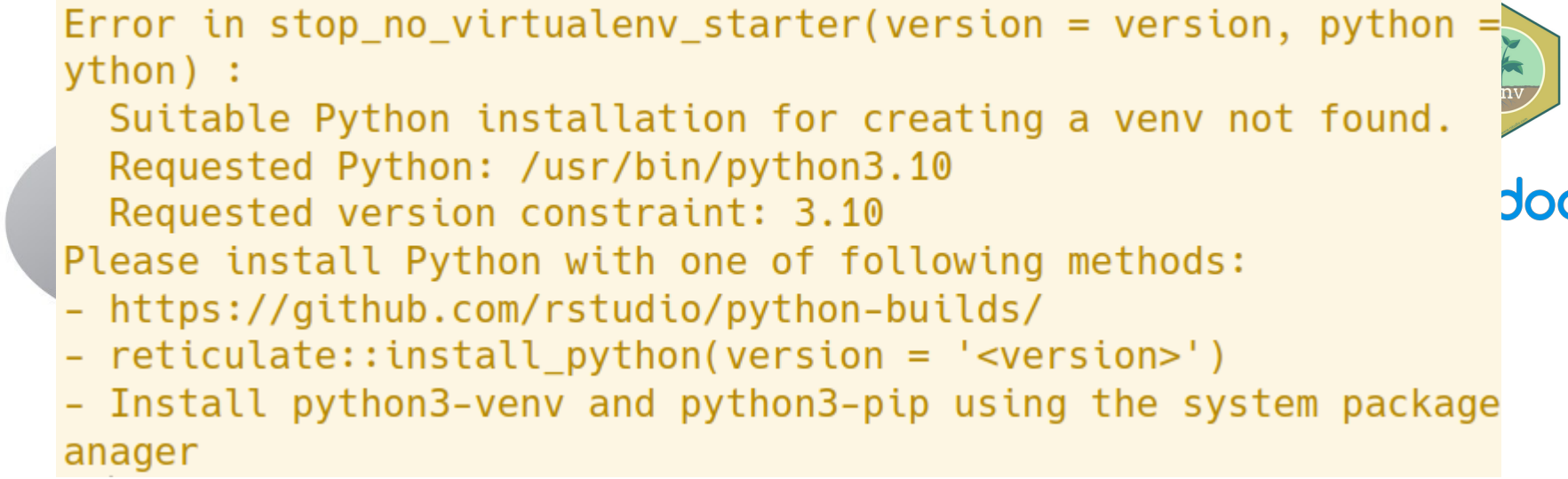
renv

docker

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data



The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data

R Libraries

System dependencies

Hardware

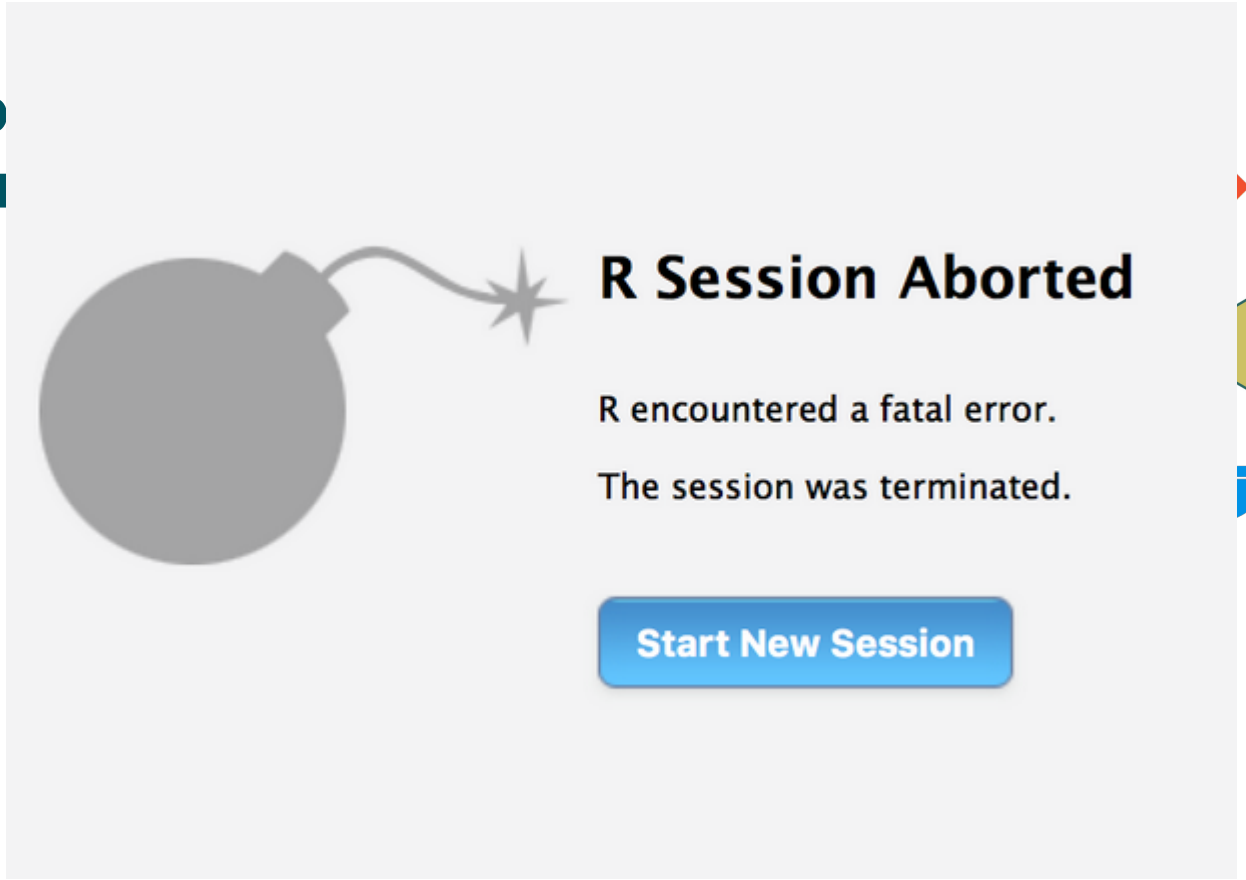
git zenodo

renv

docker

The computational reproducibility stack

Ensure everyo
can run your
analysis



git zenodo

renv

docker

Reproducible analysis

github.com/jschoeley/openscience25

Jonas Schöley

@jschoeley

0000-0002-3340-8518

schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH