Open Science Practices for Demographic Research

Jonas Schöley

schoeley@demogr.mpg.de

@jschoeley.com

jschoeley.com

Aliakbar Akbaritabar

akbaritabar@demogr.mpg.de

@akbaritabar.bsky.social

akbaritabar.github.io

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

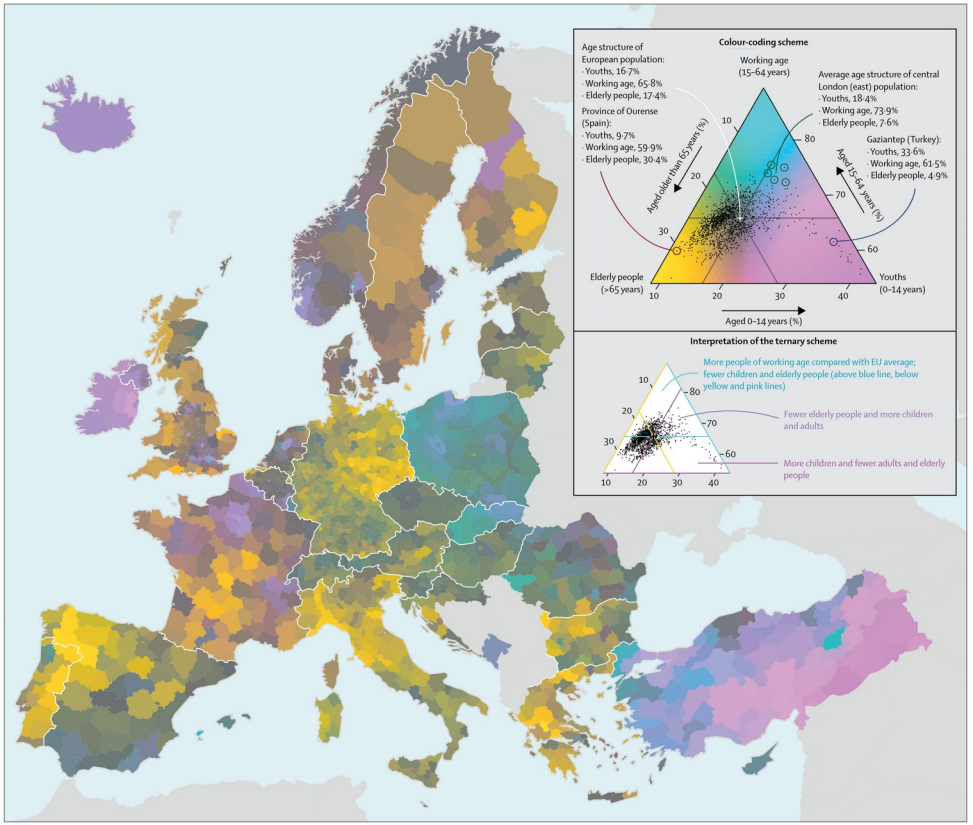


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of 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

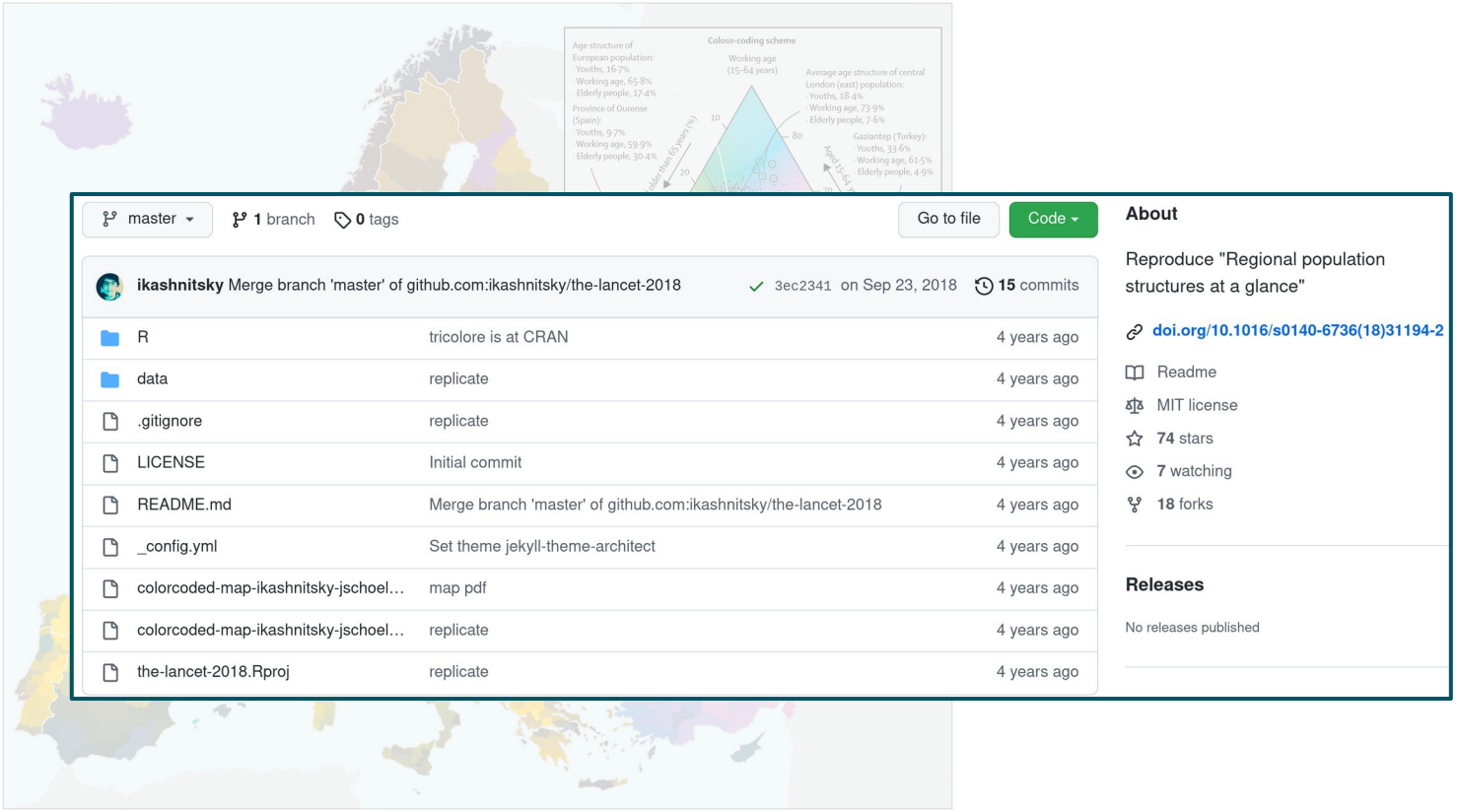


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of 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

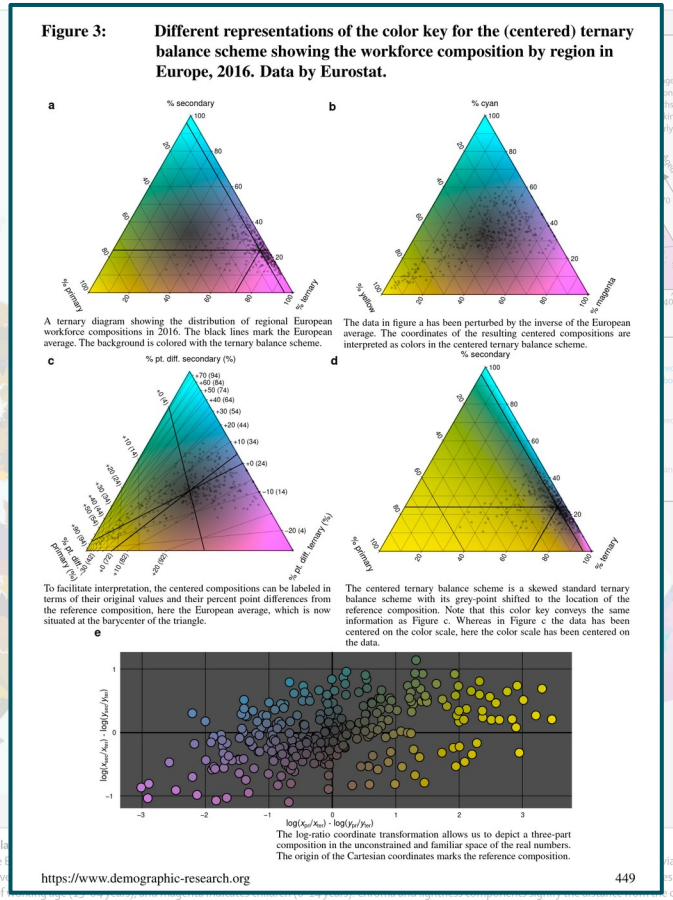


Figure 3: **Different representations of the color key for the (centered) ternary balance scheme showing the workforce composition by region in Europe, 2016. Data by Eurostat.**

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

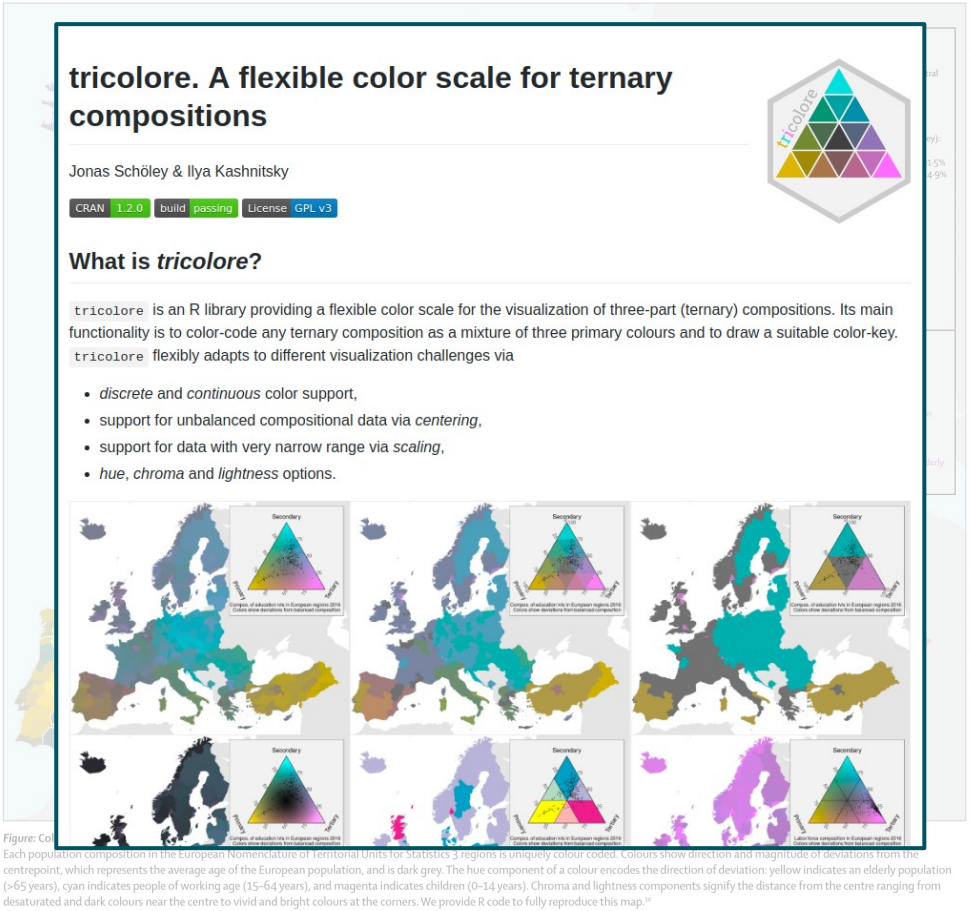
The log-ratio coordinate transformation allows us to depict a three-part composition in the unconstrained and familiar space of the real numbers. The origin of the Cartesian coordinates marks the reference composition.

Schöley (2021). The centered ternary balance scheme. demographic-research.org/volumes/vol44/19/44-19.pdf

tricolore

Together with

Ilya **Kashnitsky**
@ikashnitsky



tricolore. A flexible color scale for ternary compositions

Jonas Schöley & Ilya Kashnitsky

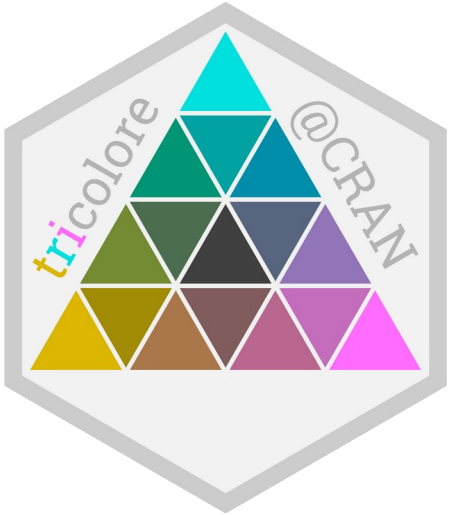
CRAN 1.2.0 | build passing | License GPL v3

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

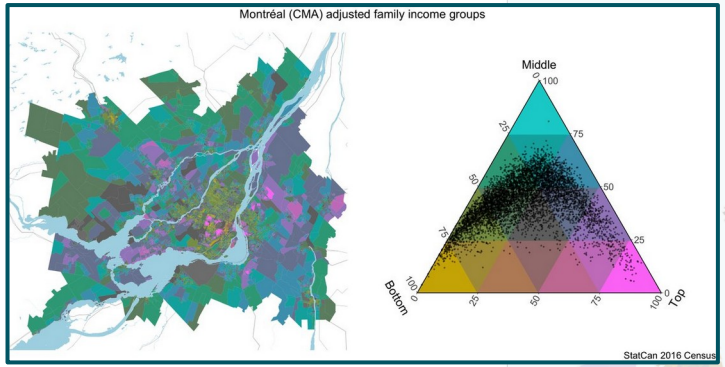
- *discrete* and *continuous* color support,
- support for unbalanced compositional data via *centering*,
- support for data with very narrow range via *scaling*,
- *hue*, *chroma* and *lightness* options.

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]

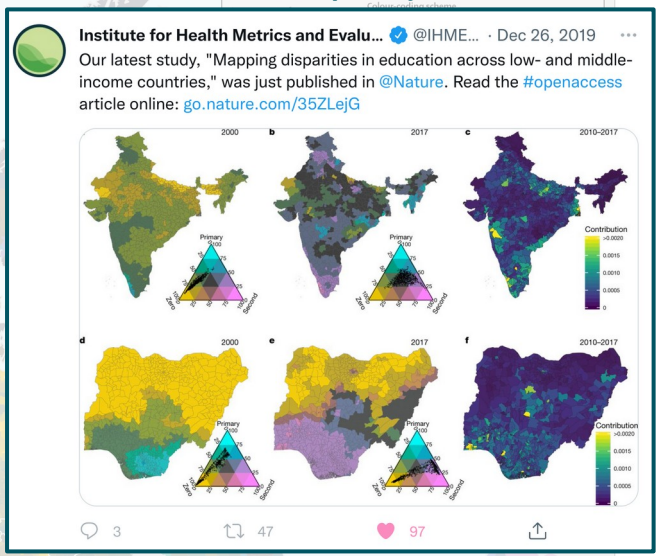


tricolore

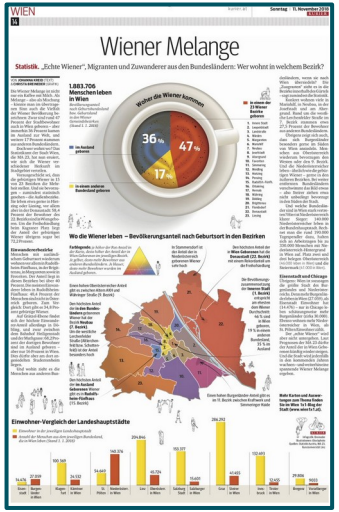
Income distribution in Canadian cities



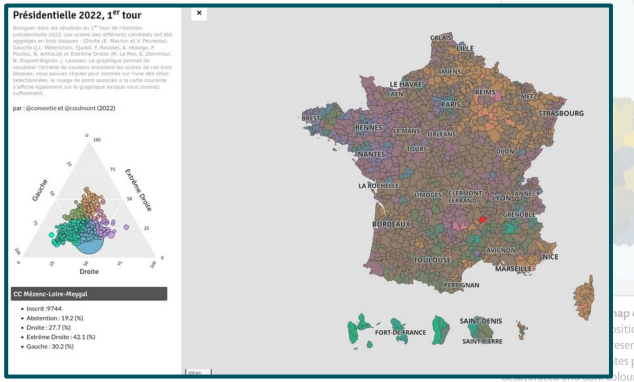
LMIC education disparity



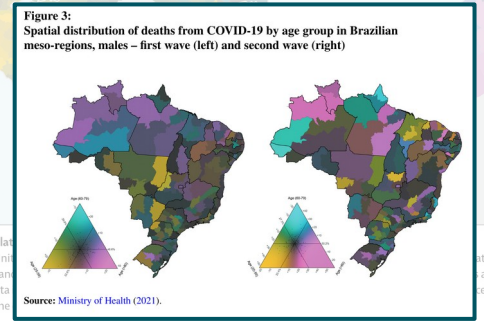
Vienna's population by origin



French election results



Regional age distribution of COVID deaths In Brazil



Agricultural and Forest Meteorology

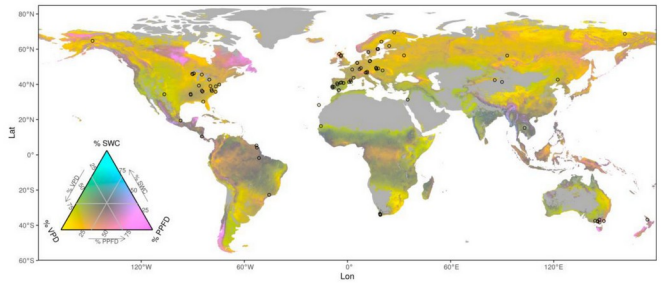


Fig. 3. Relative importance (relative R^2) of the three hydrometeorological drivers of transpiration regulation. Relative R^2 were calculated at each cell dividing the projection of each three drivers partial R^2 from the FULL model by the sum of the three partial R^2 at the same cell. Partial R^2 were 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.

Labor force composition

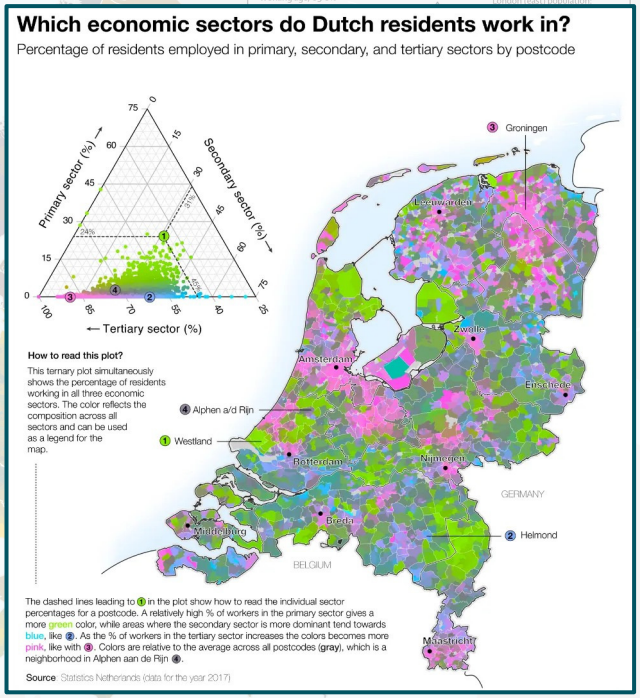


Figure: **Colour-coded map of population structures in European Nomenclature of Territorial Units for Statistics 3 regions in 2015**
Each population composition in the European Nomenclature of Territorial Units for Statistics 3 regions is uniquely colour coded. Colours show direction and magnitude of 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]

Soil composition

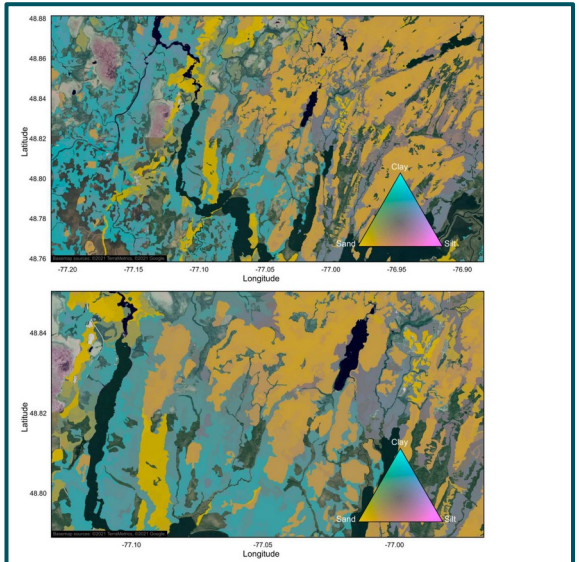


Figure 8 **Maps of soil texture composition at two zoom levels, for a region at the southern edge of the Abitibi and James Bay Lowlands soil province.** Only productive forest land characterized by mineral soils was mapped. Agricultural and unproductive forest land, organic soils, anthropogenic infrastructures,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

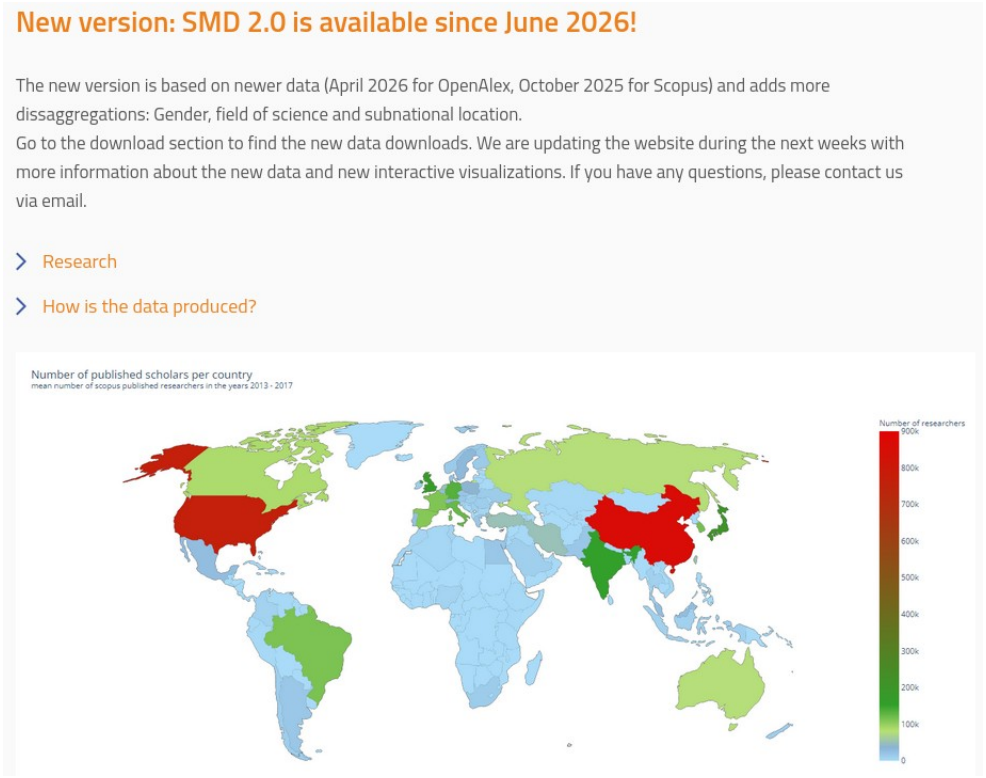
Full-size DOI: 10.7717/peerj.11685/fig-8

Reproducible Demographers

Aliakbar **Akbaritabar** @akbaritabar.bsky.social

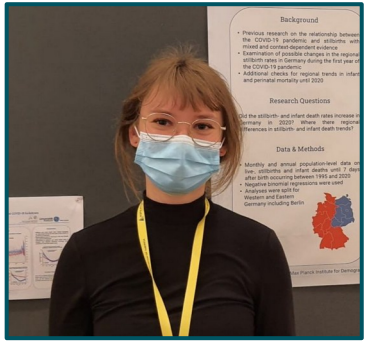
Tom **Theile**

Created a fully reproducible data base of academic migration flows.

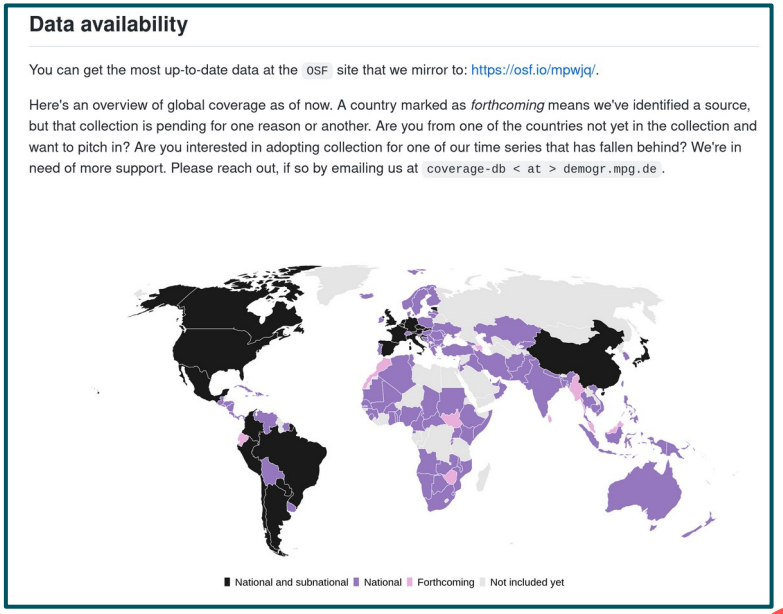


Reproducible Demographers

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

Diego **Alburez-Gutierrez**

Liliana **Calderón-Bernal**

Tom **Theile**

Amanda Martins **de Almeida**

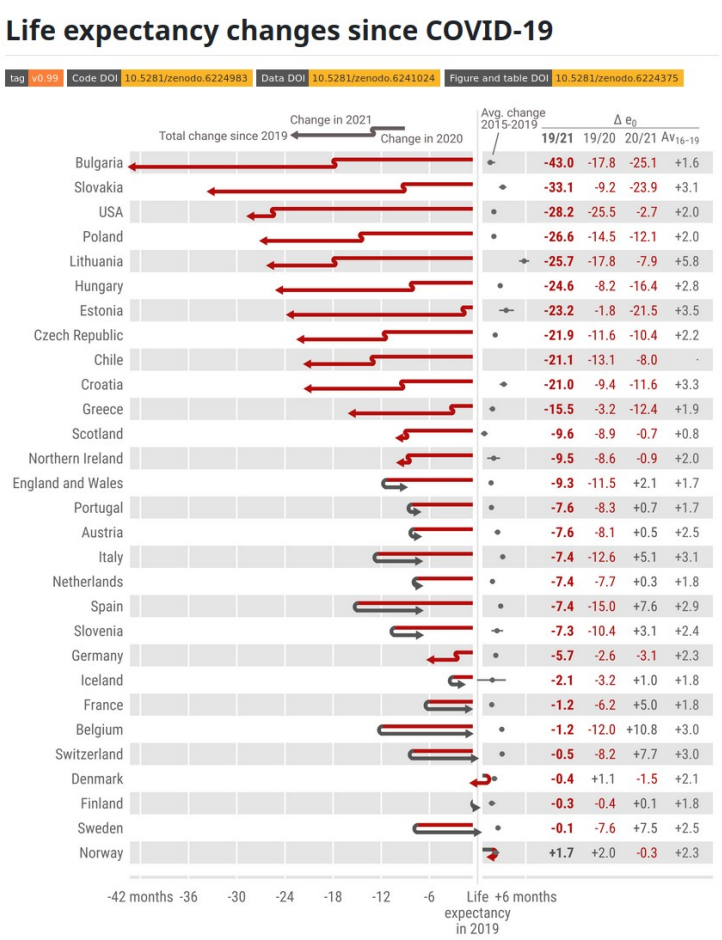
and many more...

High quality R libraries coming out of MPIDR

mpidr.github.io/rsocsim/

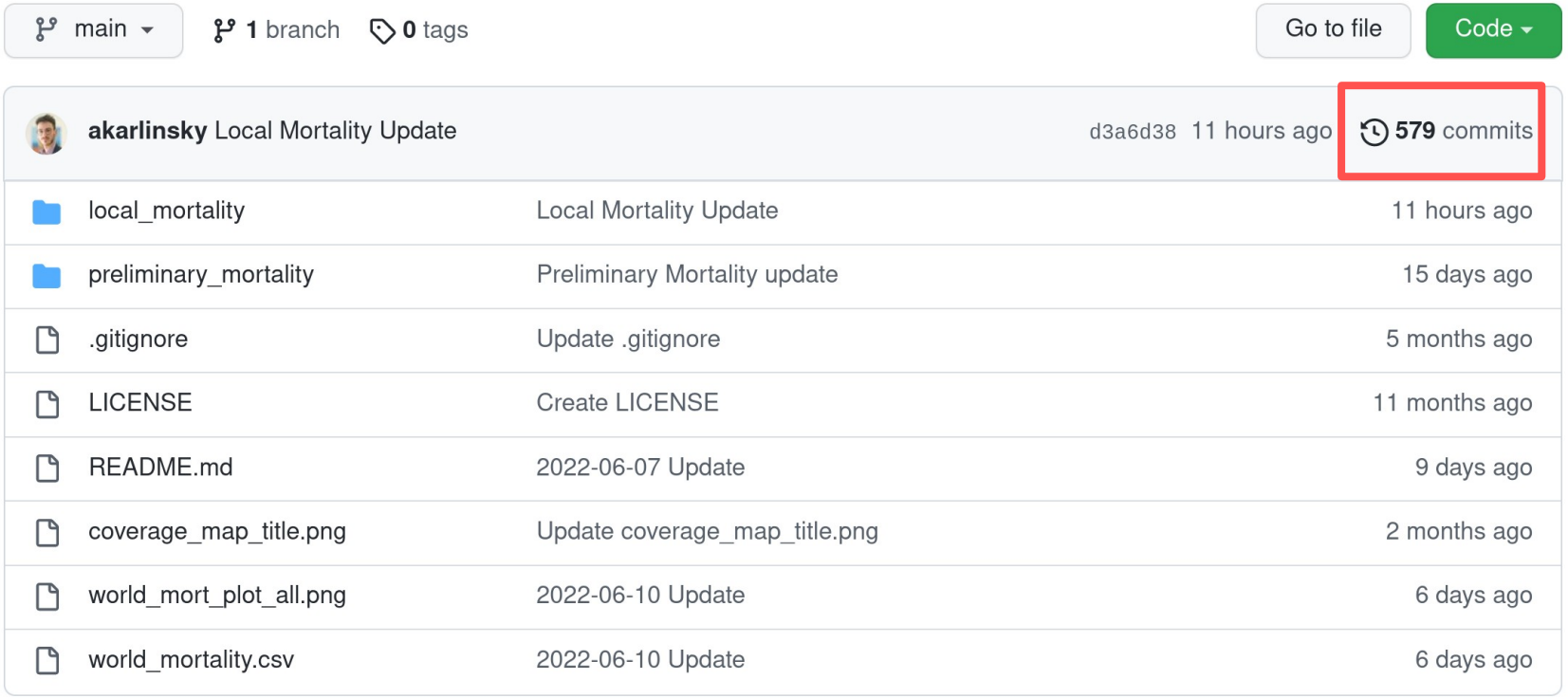
github.com/IvanWilli/DemoKin

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



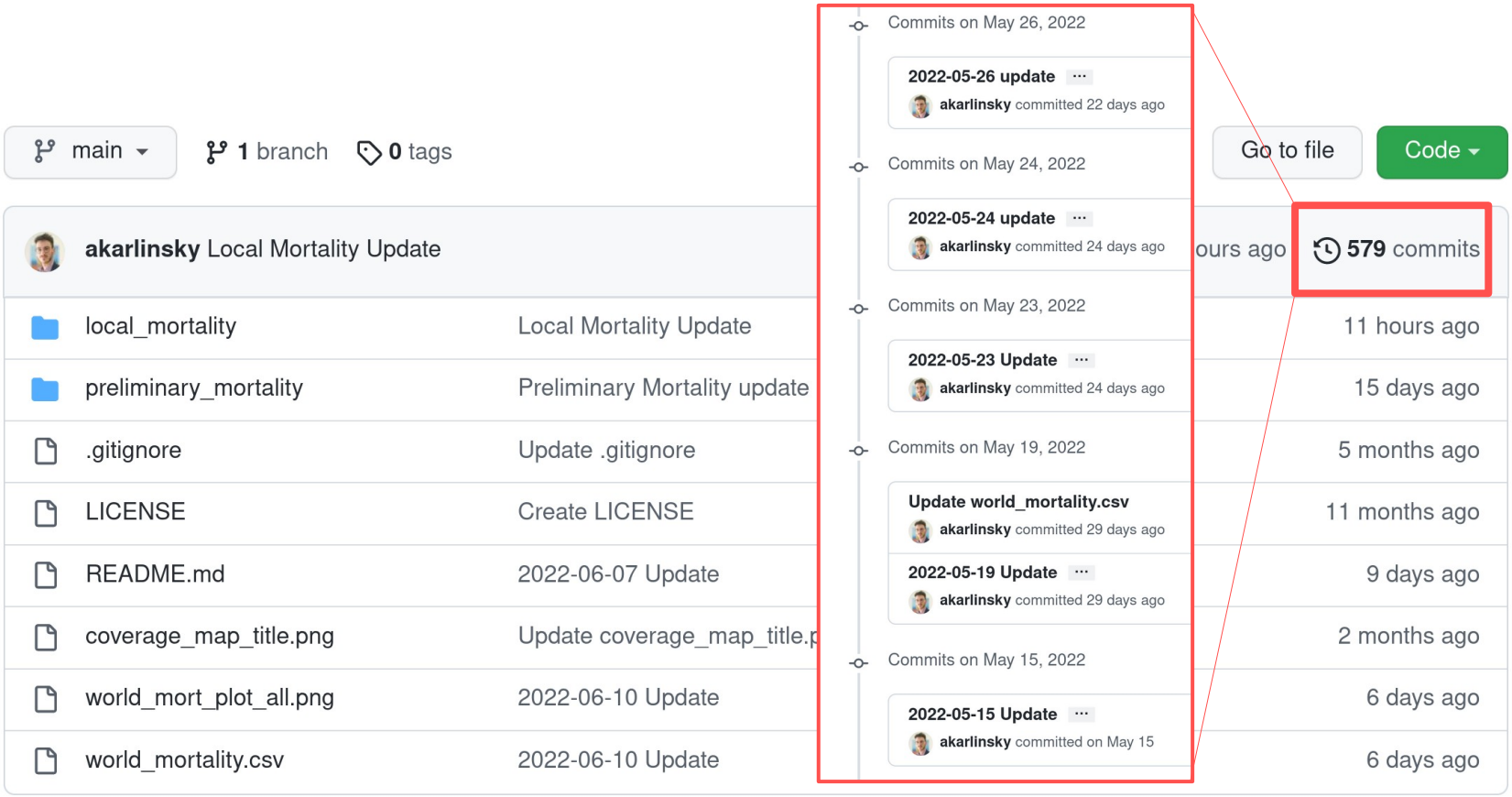
main · 1 branch · 0 tags · Go to file · Code

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

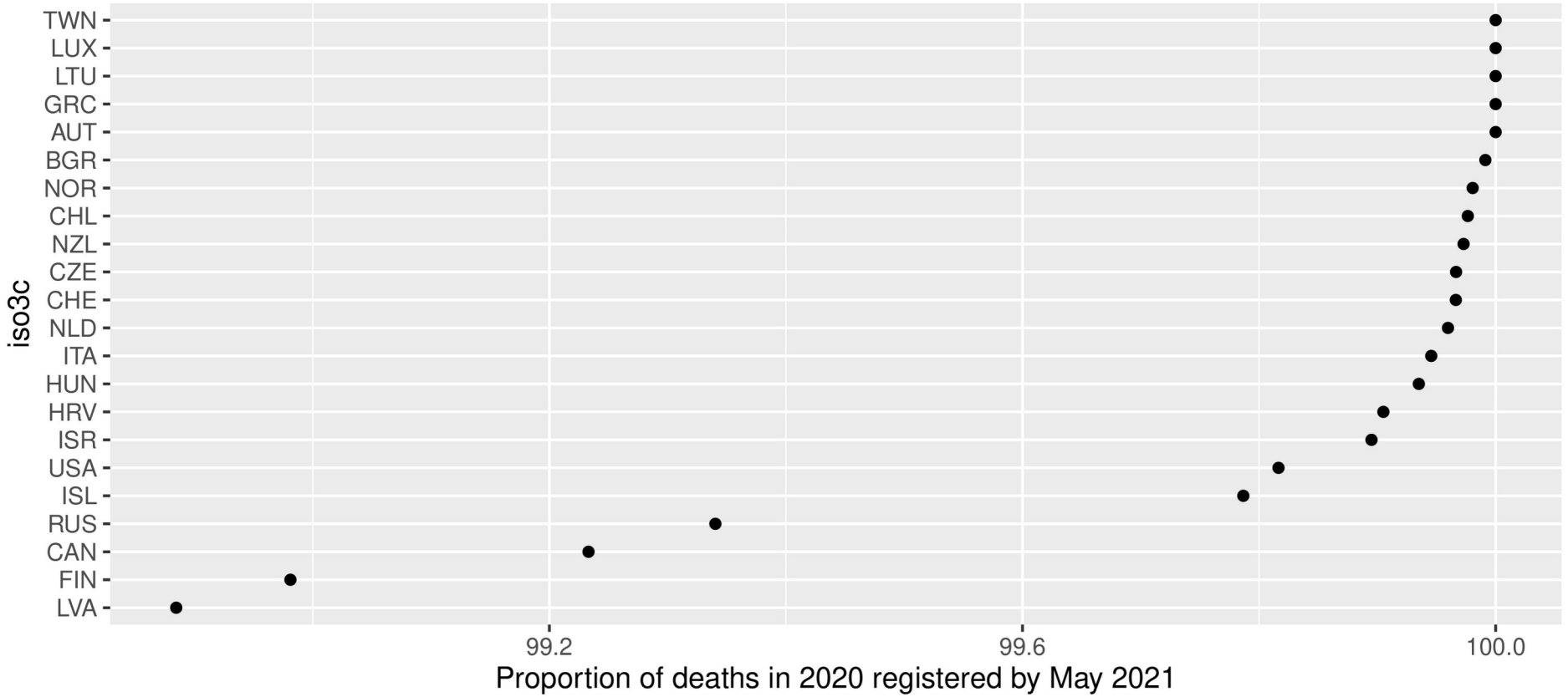
Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



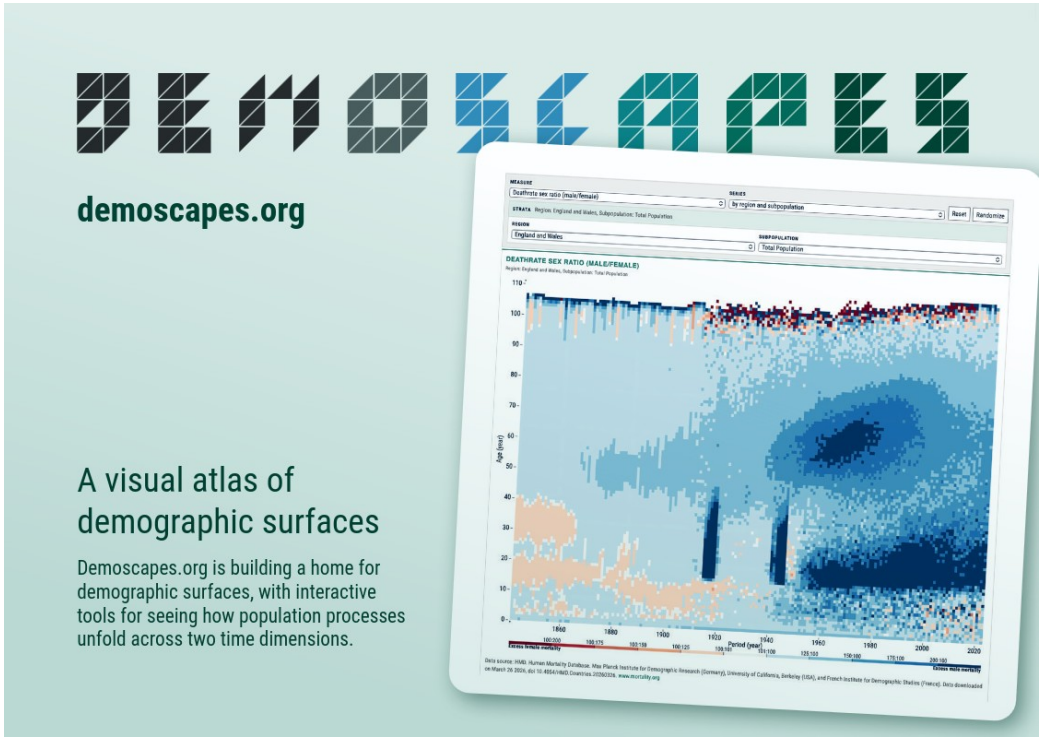
Karlinsky & Kobak (2022). World Mortality Database. github.com/akarlinsky/world_mortality

Consuming open science



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

Open science as win-win



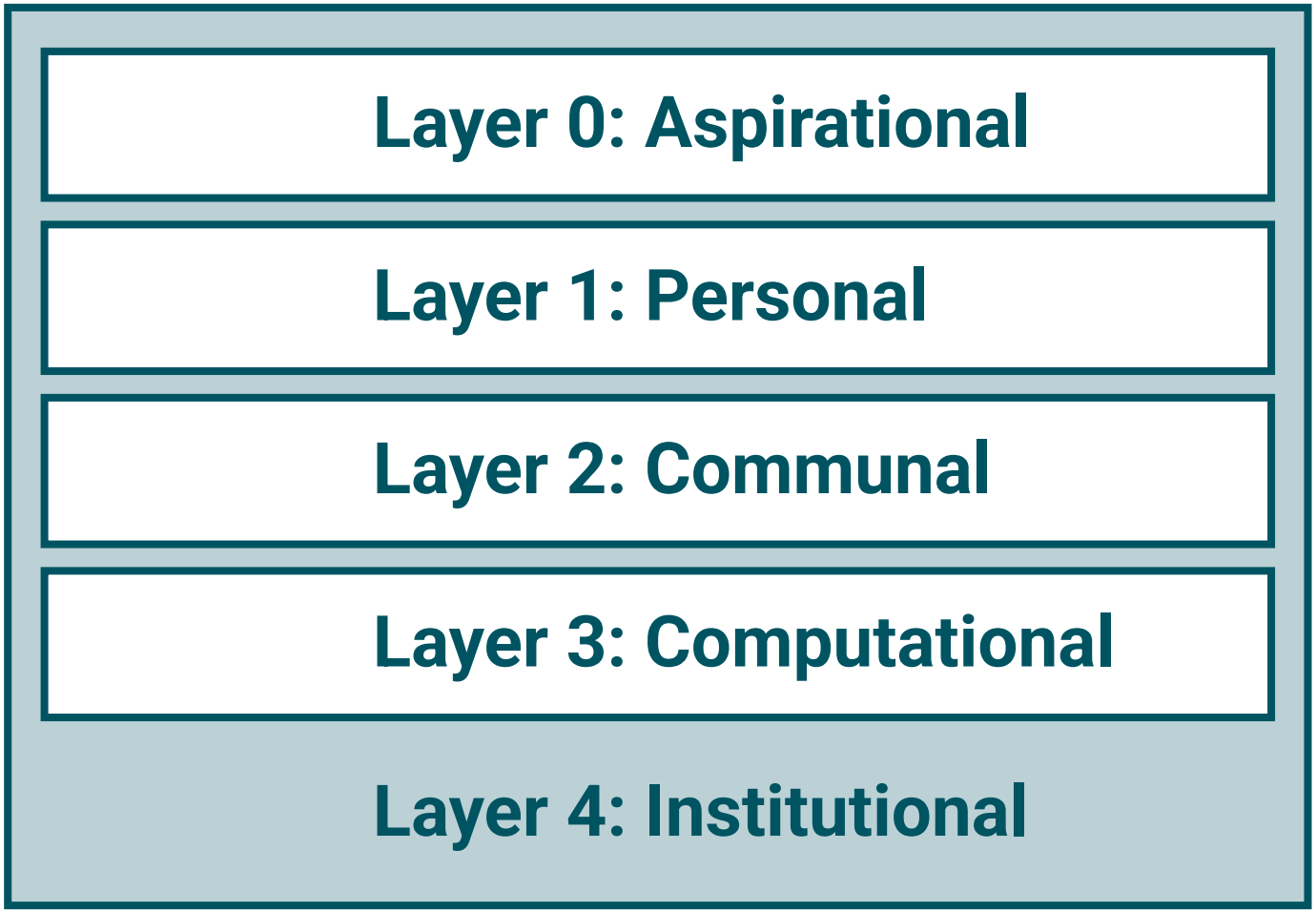
Schöley (2026). Demoscapes.org

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 + 1 Layers

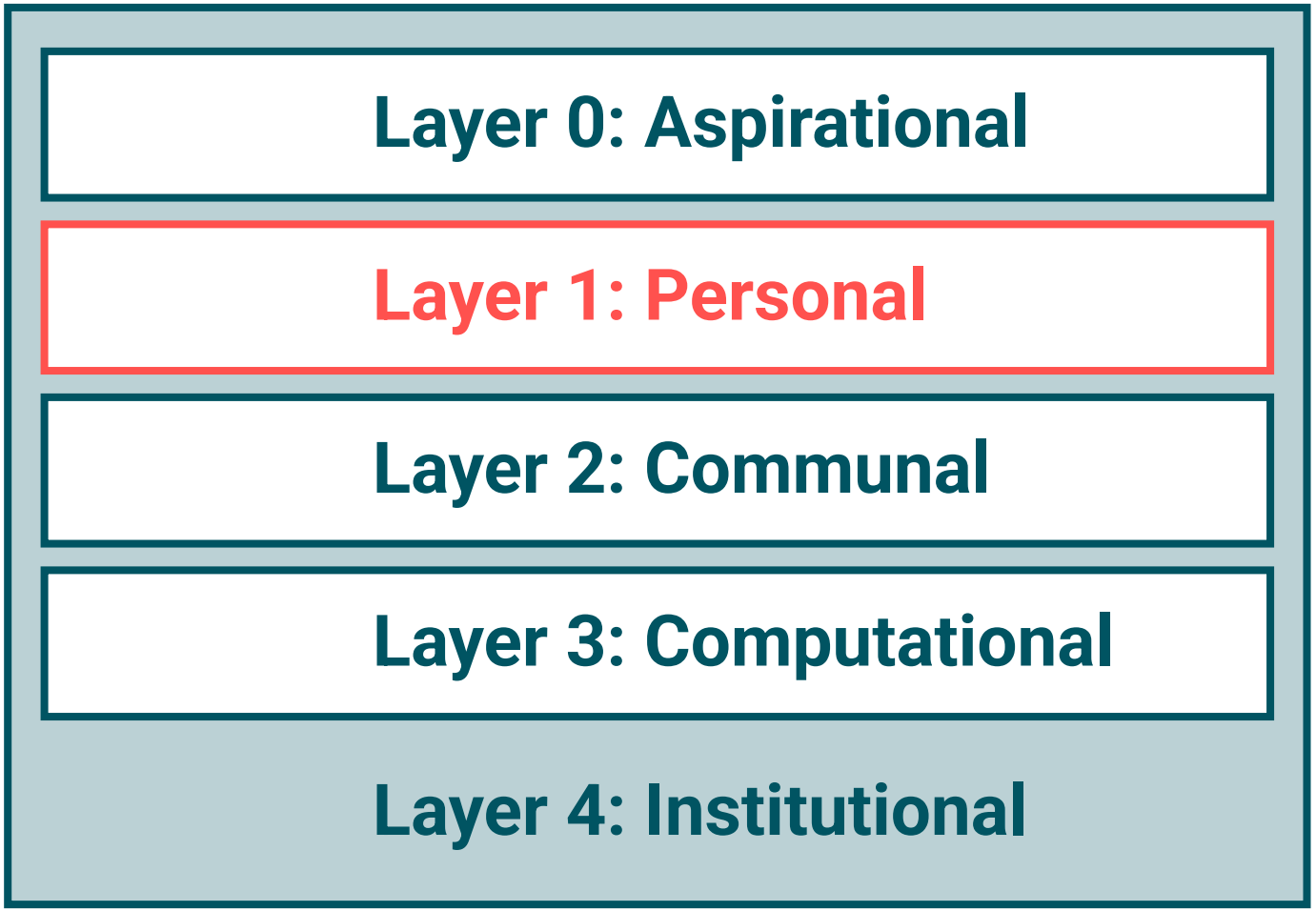

Layer 0: Aspirational
Layer 1: Personal
Layer 2: Communal
Layer 3: Computational
Layer 4: Institutional

4 Layers

Layer 0: Aspirational

Realize you're having a problem

4 Layers



Reproduce your own work

Project structure
Analysis pipeline
Documentation

4 Layers

Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

Layer 3: Computational

Layer 4: Institutional

Let others reproduce your work

4 Layers

Layer 0: Aspirational

Layer 1: Personal

Layer 2: Communal

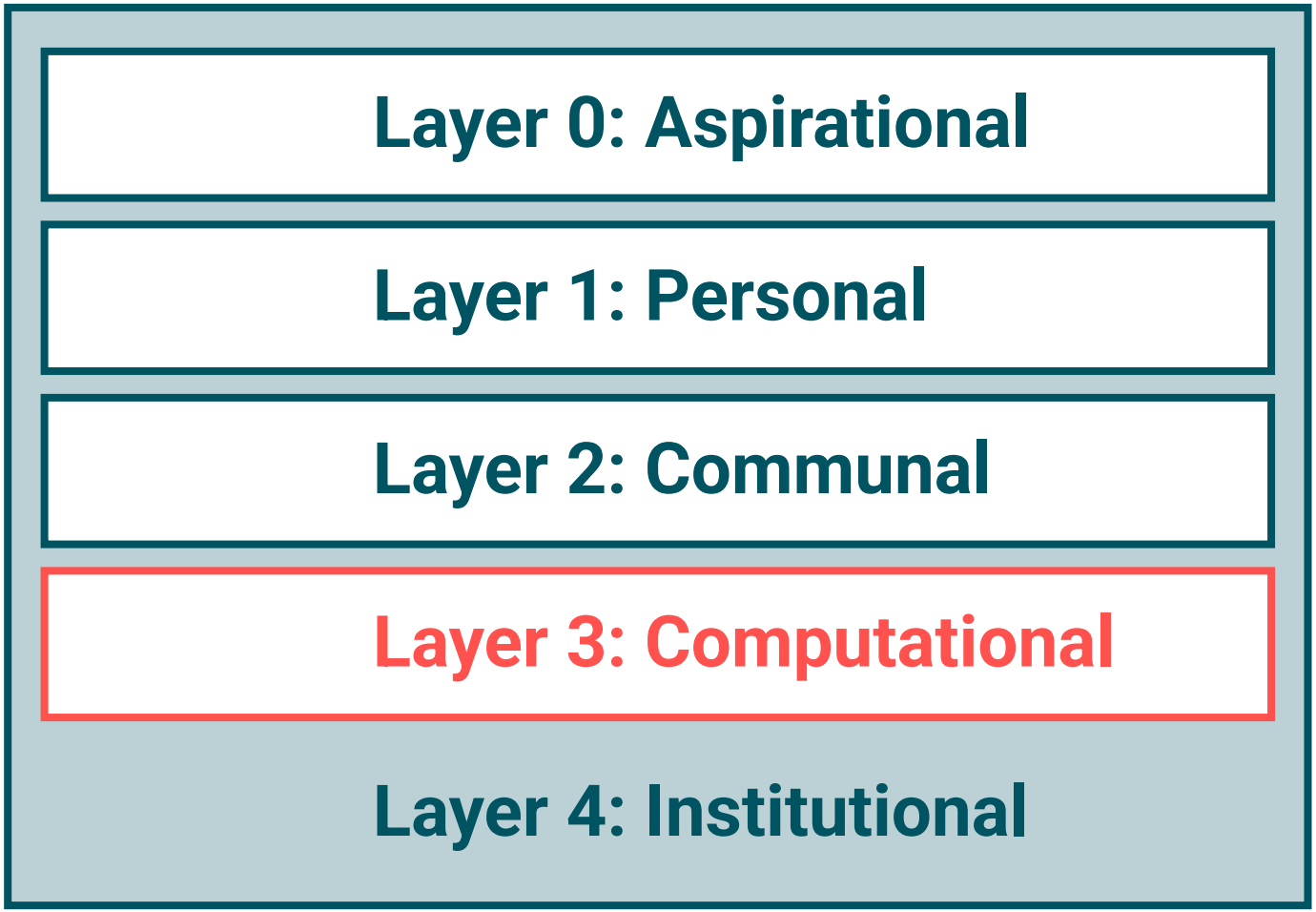
Layer 3: Computational

Layer 4: Institutional

Automatize the reproduction of your work

SnakeMake
(R, Python, Bash, ...)

4 Layers



Embed reproducibility in your organization

Personal reproducibility

Communal reproducibility

Communal reproducibility

Ensure everyone can run your analysis

Share and version your analysis

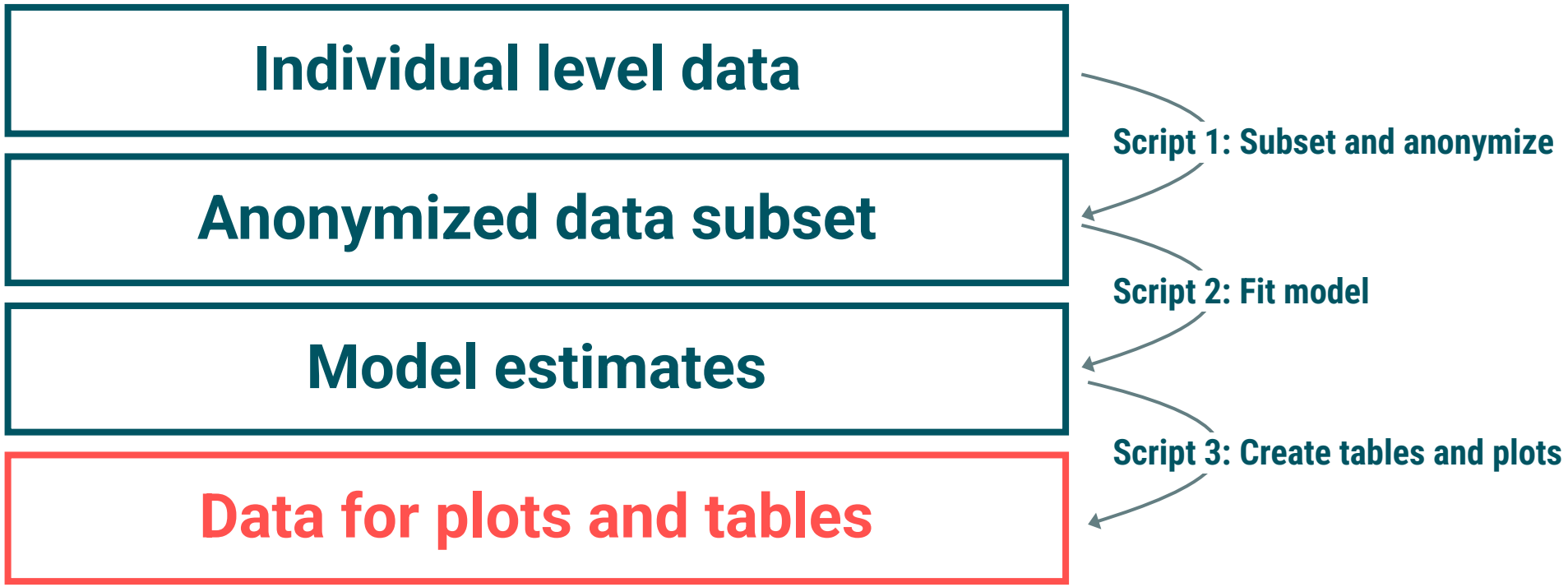
Share your data Archive your data Get DOIs

I’m not allowed to share my data

There is always a
Largest Shareable Derived Dataset

I’m not allowed to share my data

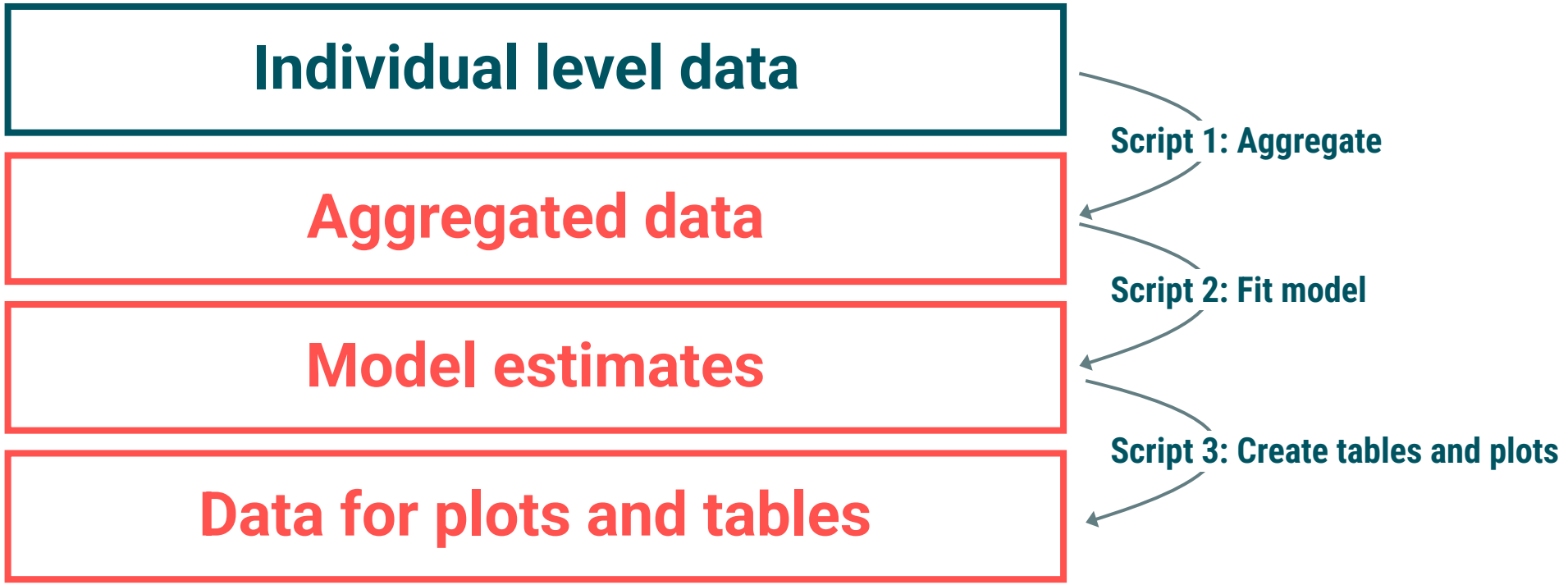
Identify your Largest Shareable Derived Dataset



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.

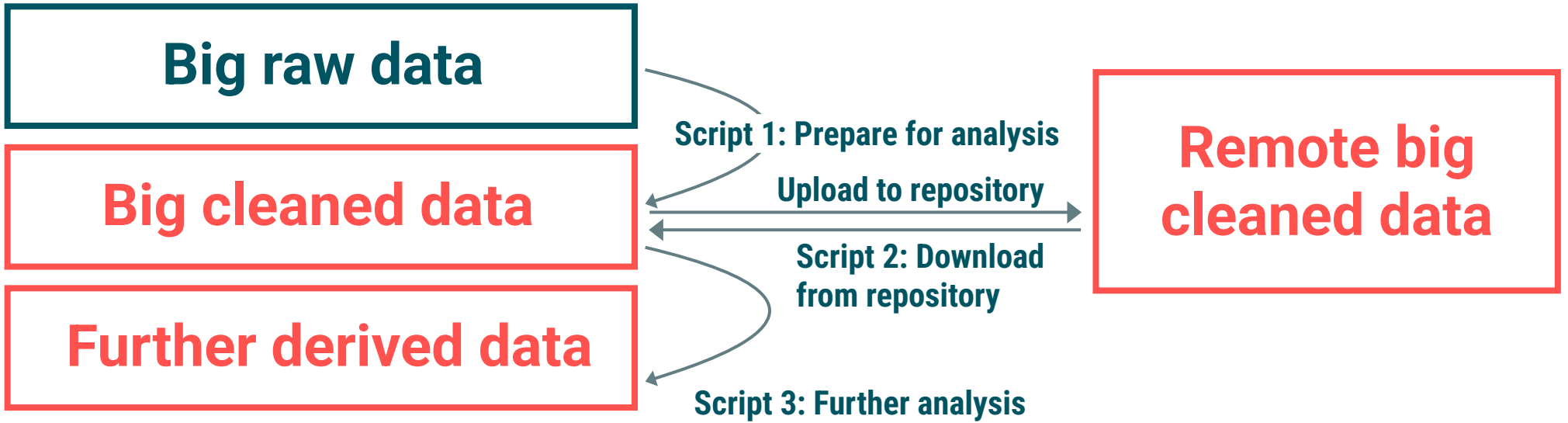
My data is too big to share

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.

Computational reproducibility

The computational reproducibility stack

Ensure everyone can run your analysis

- **Scripts & Data** — git, zenodo
- **R Libraries** — renv
- **System dependencies** — docker
- **Hardware** — podman

The computational reproducibility stack

Ensure everyone can run your analysis

- **Scripts & Data** — git, zenodo
- **R Libraries** — renv
- **System dependencies** — docker
- **Hardware** — podman

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

podman

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data

git zenodo

```
Error in stop_no_virtualenv_starter(version = version, python =
ython) :
  Suitable Python installation for creating a venv not found.
  Requested Python: /usr/bin/python3.10
  Requested version constraint: 3.10
Please install Python with one of following methods:
- https://github.com/rstudio/python-builds/
- reticulate::install_python(version = '<version>')
- Install python3-venv and python3-pip using the system package
anager
```

docker

podman

The computational reproducibility stack

Ensure everyone can run your analysis

Scripts & Data

R Libraries

System dependencies

Hardware

git zenodo

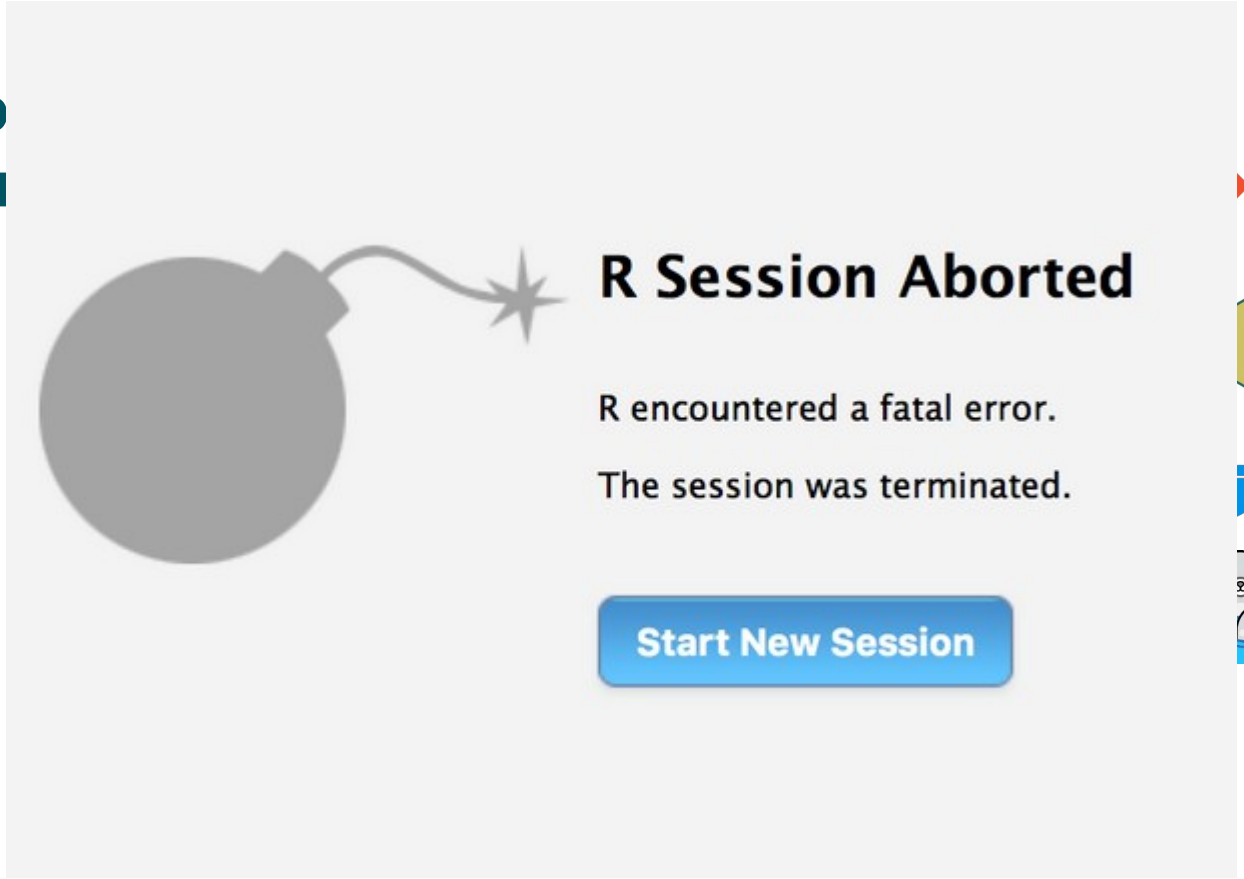
renv

docker

podman

The computational reproducibility stack

Ensure everyo
can run your
analysis



git zenodo

renv

docker

podman

Institutional open science

Institutional open science

Aliakbar **Akbaritabar**

OPEN SCIENCE AMBASSADOR

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

A few questions!

- Who **owns** the scripts/codes developed by researchers?
- Who should receive Open Science **training**?
  - At which **stage** of the **career**?
- Who should provide OS **infrastructure** and **maintain** it?
- What software solutions should be used and encouraged? **Proprietary**?
- How should OS activities be **recognized/rewarded**?

What can institutions do for OS? (1/2)

- Permit and encourage researchers to **share** their research products under an **open license**.
- Particularly, researchers **with limited contracts** should have the right to access and publish all research products they developed during their time with the institute, even after they leave.
- **Train** researchers in OS practices, funded by the institute.
- Create a centralized, in-house **research data infrastructure**, and maintain it.
- **Binding guidelines** for the organization, storage, and retrieval of research data, ensuring continuity of access and management beyond the tenure of individual researchers or teams.

What can institutions do for OS? (2/2)

- Prioritize the use and development of **open-source research software**.
- Move away from **proprietary** software solutions.
- Micro-funds for **OS projects**, and/or **dedicated staff** to support OS.
- Transparent and reproducible research as a standard **criterion** for institute's **evaluation**.
- **Recognize** OS contributions by researchers and **reward** them, especially for **early career** researchers.

Hands-on part

What to expect in the Hands-on part?

- An interactive, self-paced, course on Git, GitHub, and version control with 4 steps:
  - Git and version control basics
  - GitHub - remote repositories and online collaboration
  - Branches, collaboration, and open science
  - Advanced Git, GitHub tools, and reproducibility
- **Requirement**:
  - Having a GitHub account, or opening one! Otherwise, follow on a neighbor's screen.
- **Link**:
  - https://tinyurl.com/EPC26OS

Slides

github.com/jschoeley/epc26os

Hands-on materials

github.com/akbaritabar/Git-GitHub-Version-Control-Interactive-Self-paced-Course

Jonas Schöley

schoeley@demogr.mpg.de

@jschoeley.com

jschoeley.com

Aliakbar Akbaritabar

akbaritabar@demogr.mpg.de

@akbaritabar.bsky.social

akbaritabar.github.io

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH