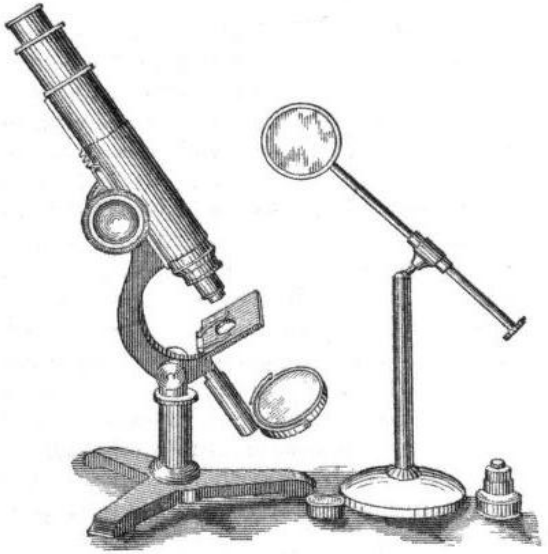
Avoin datatiede

Leo Lahti
Turun yliopisto



Turun yliopisto
University of Turku

Avoin tiede

Kansallinen koordinaatio

Asiasanahaku

Suomi Svenska English

ASIANTUNTIJARYHMÄT LINJAUKSET JA AINEISTOT KOORDINAATIO MITÄ ON AVOIN TIEDE? VERKOSTOT

Toimintakulttuuri

Julkaiseminen

Tutkimusaineistot

Oppiminen

Työryhmä ehdottaa toimenpiteitä julkishallinnon tietojen tehokkaammaksi hyödyntämiseksi

Opetus- ja kulttuuriministeriö, Valtiovarainministeriö 15.6.2020 9.00

Tavoitteena on saattaa julkishallinnon aineistot paremmin tutkimuskäyttöön ja parantaa tietoon perustuvan päätöksenteon tukea.

Valtiovarainministeriö ja opetus- ja kulttuuriministeriö asettivat maaliskuussa 2020 työryhmän, jonka tavoitteena oli laatia suunnitelma julkishallinnon aineistojen hyödyntämisen parantamiseksi. Työssä keskityttiin luvanvaraisiin aineistoihin.

Akatemian linjaukset avoimesta tieteestä

Tieteellisten julkaisujen avoin saatavuus

Tutkimusaineistojen hallinnointi ja avoimuus

Tutkimusmenetelmien avoimuus

Tutkimustuotosten metatiedot

Hallitus syventää tietopolitiikan johtamista. Julkisen tiedon avoimuudesta tehdään koko tietopolitiikan kantava periaate. Hallitus edistää avoimen lähdekoodin ensisijaisuutta julkisissa tietojärjestelmissä ja niiden hankinnoissa. Hallitus säätää lailla velvoitteen edellyttää avoimia rajapintoja julkisia tietojärjestelmiä hankittaessa, ellei painavasta syystä muuta johdu. Hallitus jatkaa määrätietoista julkisten tietovarantojen avaamista ja laaditaan niille hyödyntämistä helpottavat sitovat laatukriteerit. Lisäksi julkisuuslain periaatteet ja vaatimus tietovarantojen avaamisesta ulotetaan koskemaan myös julkisomisteisia yhtiöitä.

Avoin tieto

eurostat
Your key to European statistics

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

OpenStreetMap
The Free Wiki World Map

Tilastokeskus
Statistikcentralen
Statistics Finland

MAANMITTAUSLAITOS
LANTMÄTERIVERKET

1640
KANSALLIS
KIRJASTO

documents

data

code

review

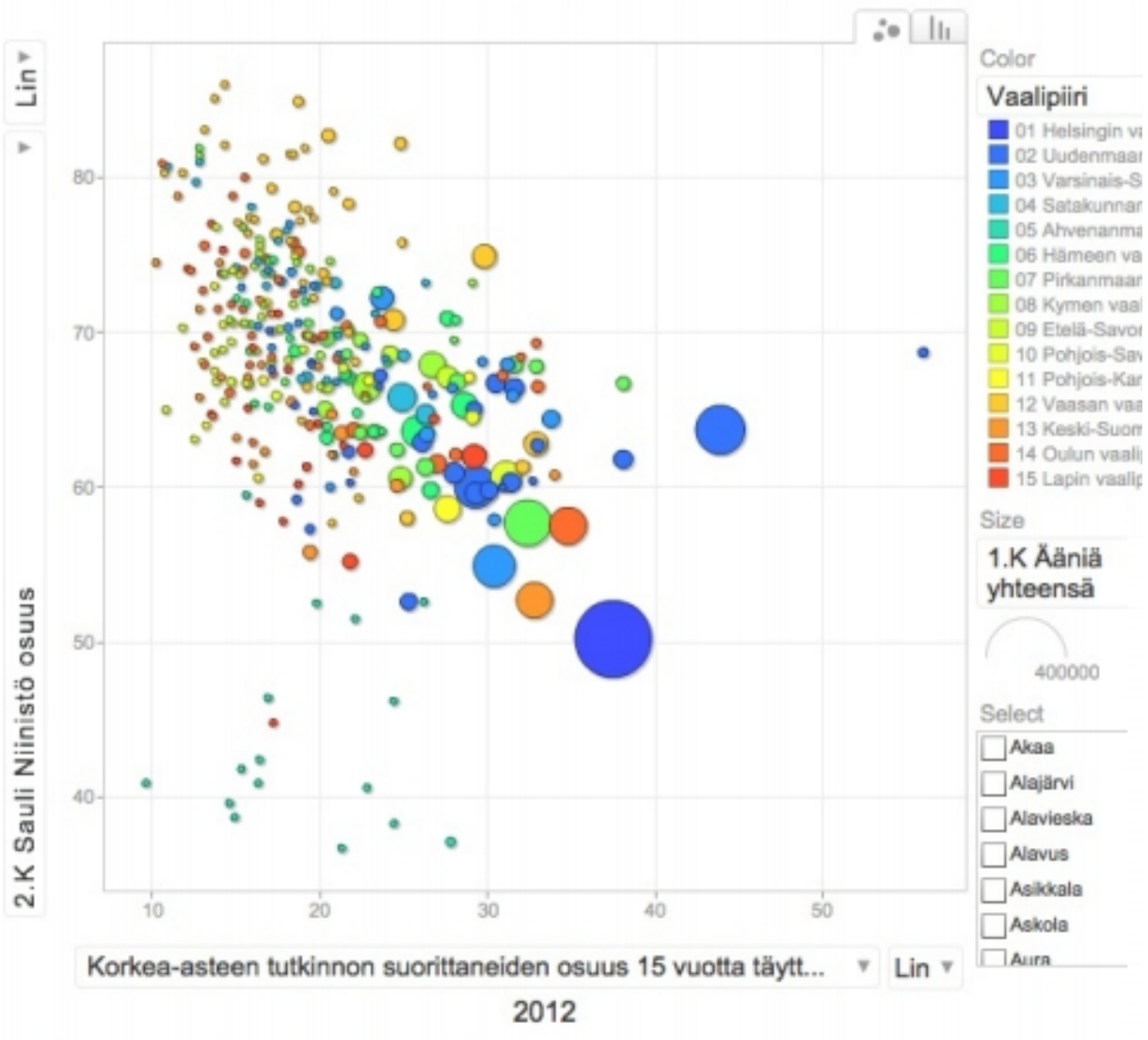
community

infrastructure

Presidenttiehdokkaiden kannatus ja suomalaisten hyvinvointi

Julkaistu helmikuu 16, 2012 by antagomir

louhos.wordpress.com



Data:
- MML
- TIlastokeskus
- YLE / HS

Avoimet kehittäjäverkostot

rOpenGov

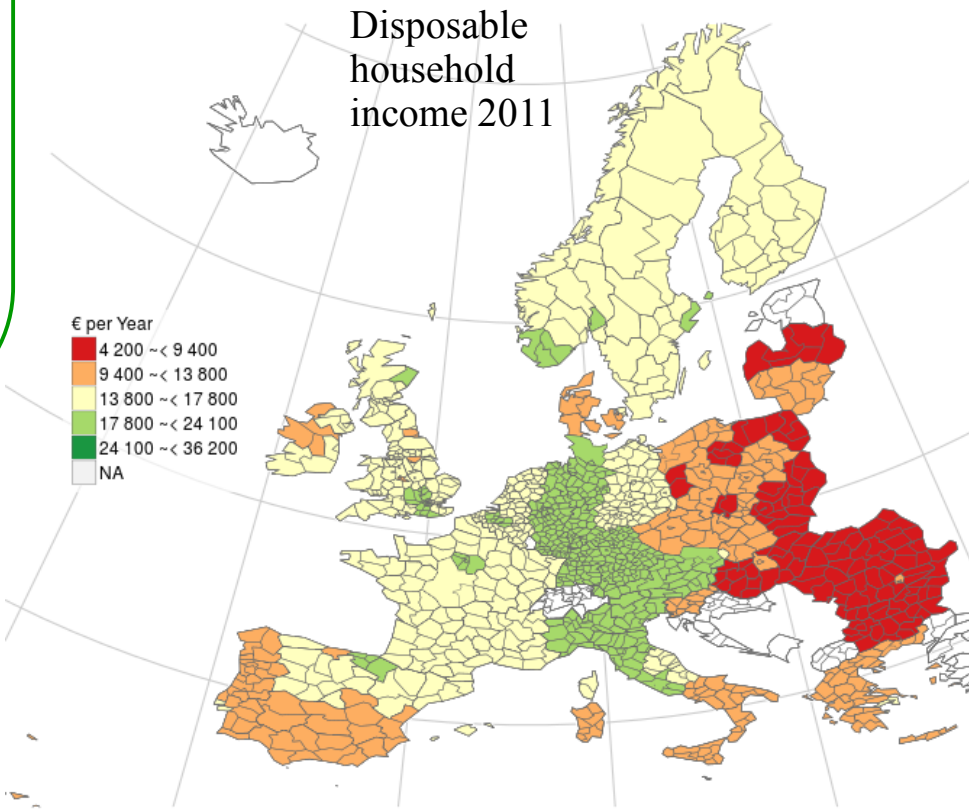
International developer network
for open government data analytics

Officially launched at
NIPS'13 Machine Learning
Open Source Software workshop

20+ R packages | ~100,000 dl/year
Eurostat, Statistics Finland, THL, FMI,
Land Survey Finland, Open Street Map, ...

Retrieval and Analysis of Eurostat Open Data with the eurostat Package

by Leo Lahti, Janne Huovari, Markus Kainu, and Przemysław Biecek



Awards, Seed funding & Collaboration 2009-2019

Apps4Finland
doing good with open data

HS.fi

datavaalit

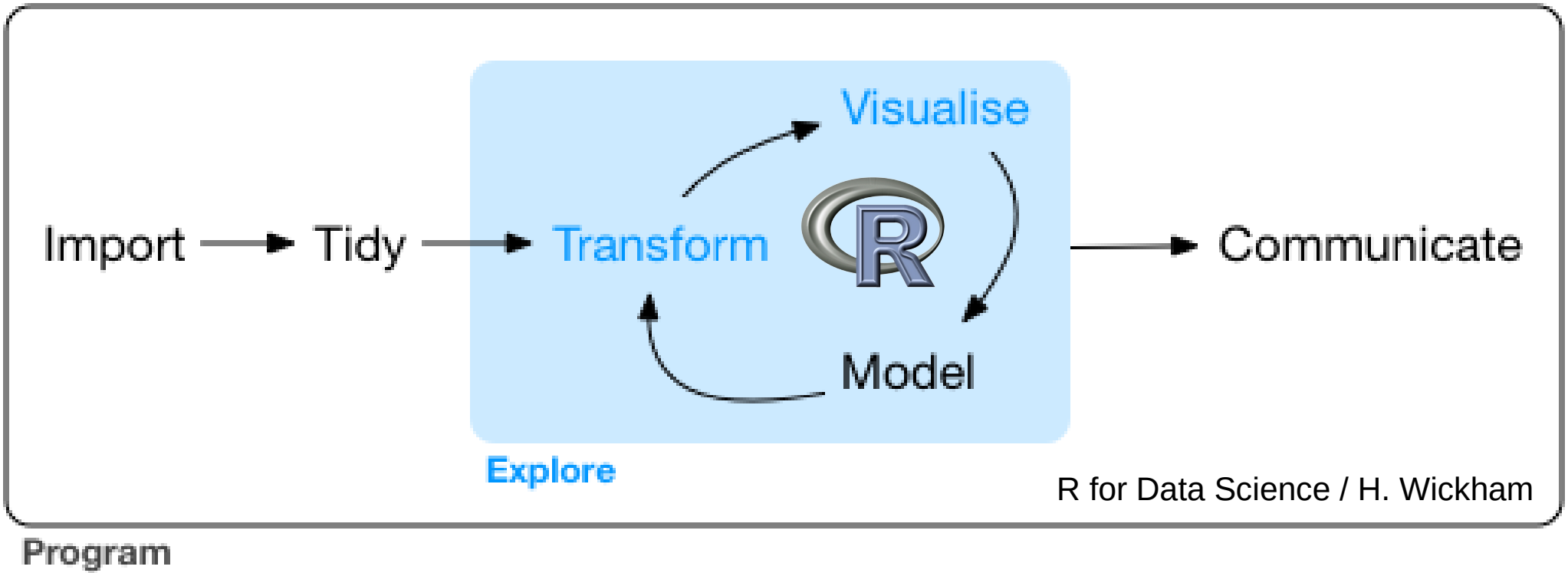
yle

ROpenSci

SITRA

Workshops, Tutorials,
Education material

Datasta tietoon?



Computational workflows have an increasingly central role in research & decision-making

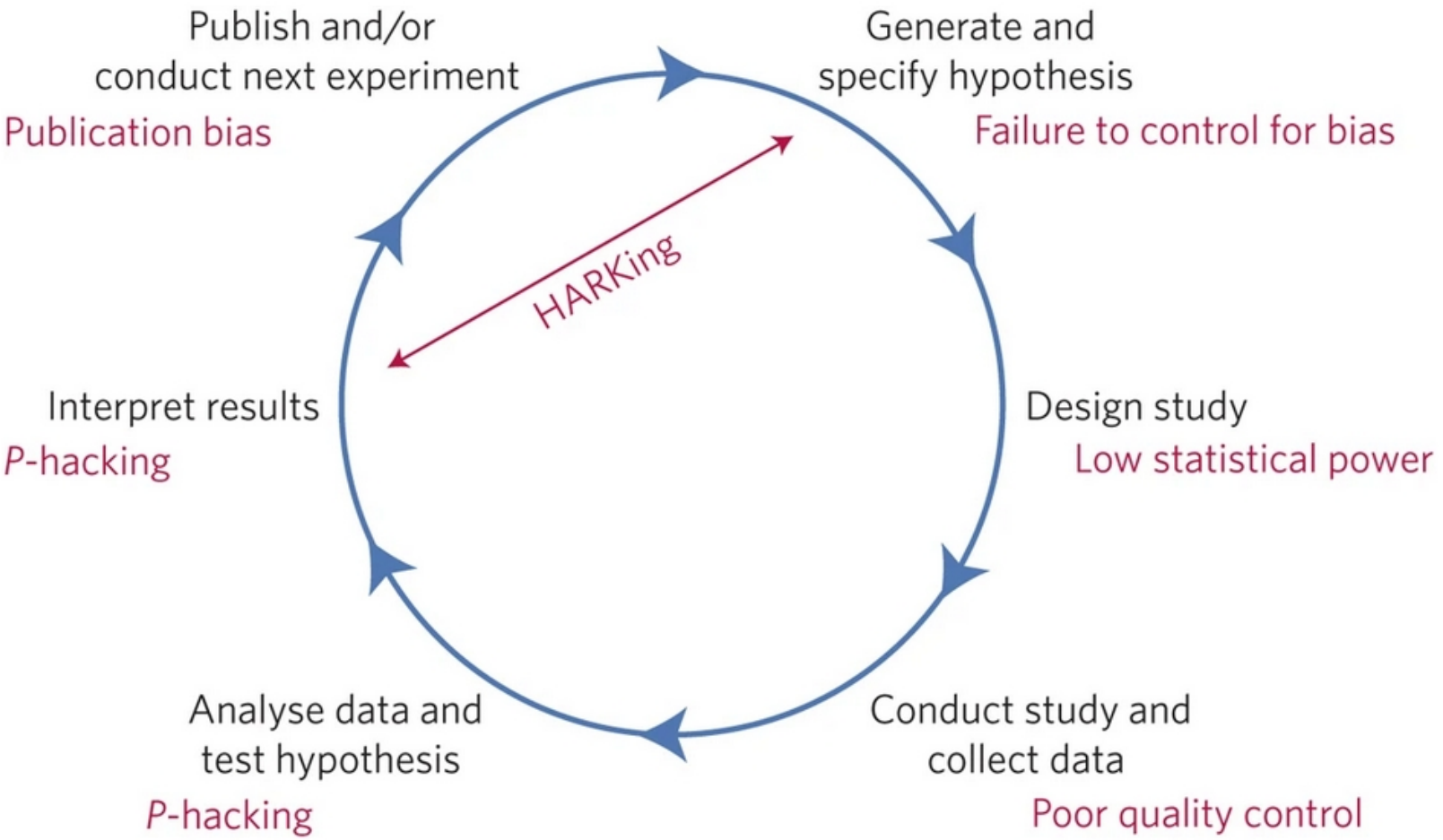
A manifesto for reproducible science

Marcus R. Munafò ✉, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

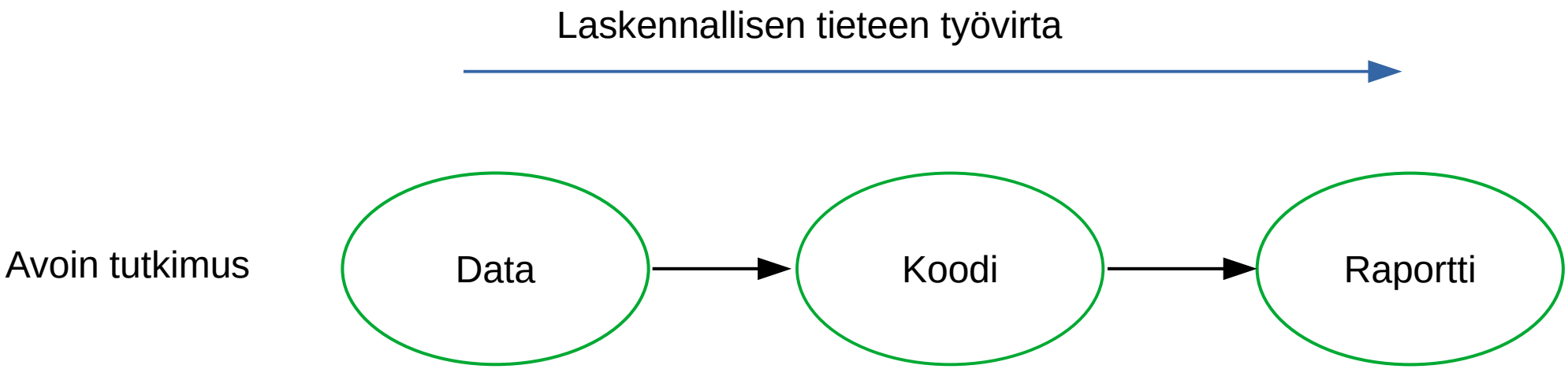
Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

Figure 1: Threats to reproducible science.

From: A manifesto for reproducible science



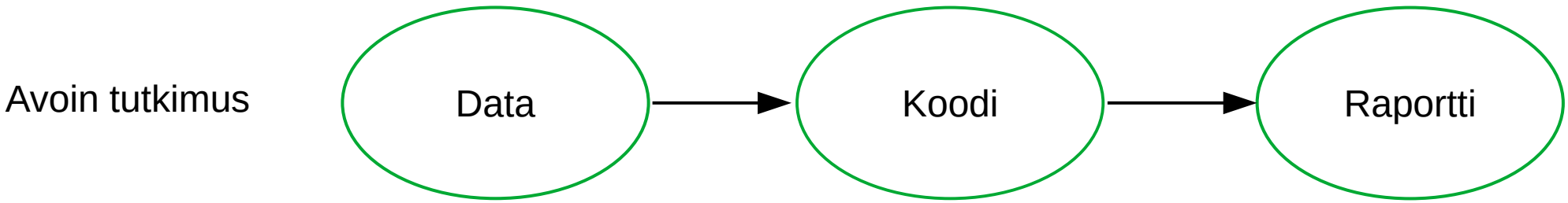
An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.



"*I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place whereit should be accessible, under reasonable restrictions, to those who desire to verify his work.*"

Francis Galton (1901), *Biometrika* 1:1, pp. 7-10.

Laskennallisen tieteen työvirta



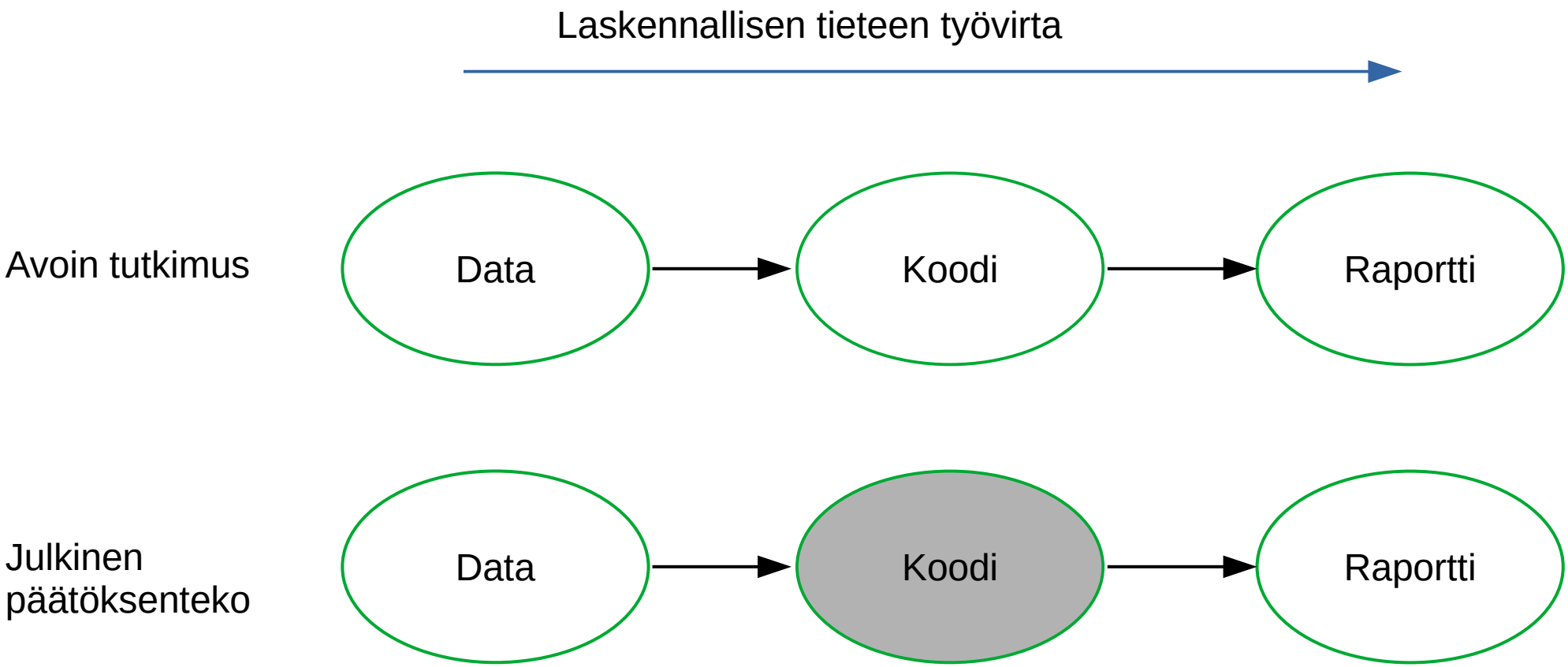

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```

http://web.stanford.edu/class/cs109l/unrestricted/images/

RESEARCH PRIORITIES

Shining Light into Black Boxes

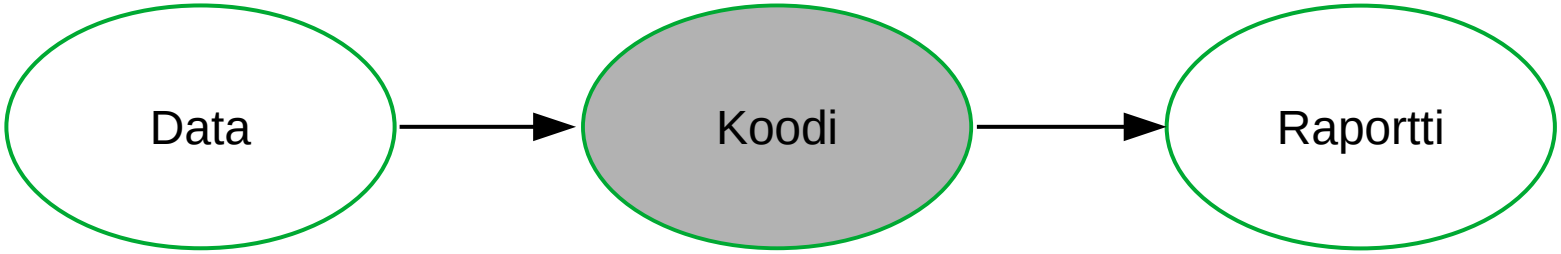
A. Morin[1], J. Urban[2], P. D. Adams[3], I. Foster[4], A. Sali[5], D. Baker[6], P. Sliz[1,*]

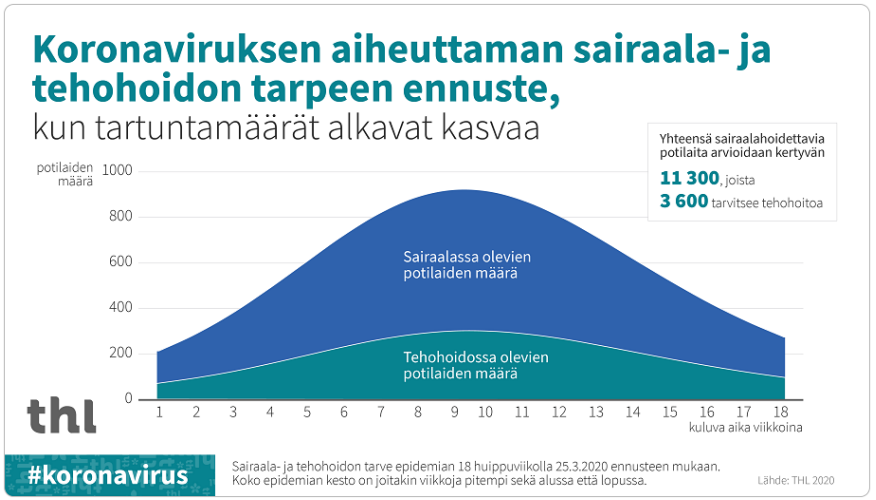


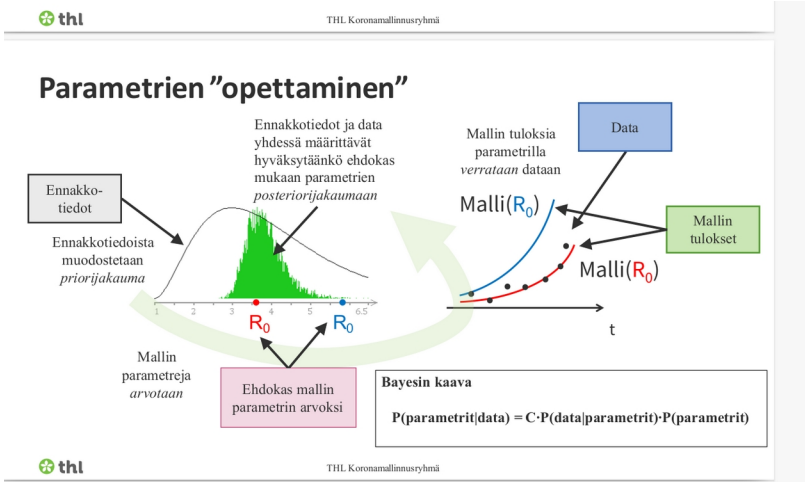
Kuva 1: Datasta jalostetaan tietoa tutkimuksen ja päätöksenteon raportteihin laskennallisten työvirtojen avulla. Avoimen tutkimuksen käytännöt ovat korostaneet koko päättelyketjun avoimuutta, mutta julkisessa päätöksenteossa datan tulkintaan käytettävät menetelmät muodostavat harmaan laatikon. Harmaan laatikon malli voidaan kuvata yleisellä tasolla, mutta lähdekoodin sisältämät keskeiset yksityiskohdat salataan.

Laskennallisen tieteen työvirta

Julkinen päätöksenteko







Oikeuskansleri pyytää selvitystä koronatietojen panttaamisesta – ”Peruslähtökohtana on viranomais-toiminnan avoimuus”

Sosiaali- ja terveysministeriö joutuu vastaamaan myös siihen, onko valtioneuvoston periaatepäätöksen taustalla piiloon jääneitä perusteita.

HS 14.5.2020: Oikeuskanslerinvirasto viittaa selvityspyynnössään uutisiin ja kommentteihin, joiden mukaan **THL:n tuottamia epidemian kulkua kuvaavia mallinnuksia ei ole julkaistu kaikilta osin eikä niistä siten ole voitu käydä julkista keskustelua.** Myöskään ihmisten oikeuksiin voimakkaasti vaikuttavien toimenpiteiden perusteena olevia parametreja, taustaoletuksia ja laskelmia ole julkaistu kaikilta osin. ”Myös valtioneuvoston valmisteluun liittyvissä valmisteluasiakirjoissa on todettu muun muassa, ettei THL ole julkaissut kaikkia käyttämänsä mallin parametreja, jolloin **epidemian leviämisen ennusteita ei voida esimerkiksi tutkimuksessa toisintaa**.”

3.5.2020: “*Hallitus katsoo, että kaikki päätöksenteon perusteena olleet taustatiedot ja laskelmat oletuksineen ja parametreineen noudattaen avoimen tieteen ja tutkimuksen periaatteita tulee julkaista.*”

Mielipide | Lukijan mielipide

Epidemiologiset mallit tulisi julkistaa

Vaihtoehtoisten mallinnusten keskinäinen vertailu on mahdollista vain, jos kilpailevien mallien oletukset ovat läpinäkyviä.

"Vahva julkinen hallinto on koko oikeusvaltiomme toimivuuden perusta"

4.5 Maailman paras julkinen hallinto Hallitus syventää tietopolitiikan johtamista. Julkisen tiedon avoimuudesta tehdään koko tietopolitiikan kantava periaate. Hallitus edistää avoimen lähdekoodin ensisijaisuutta julkisissa tietojärjestelmissä ja niiden hankinnoissa. Hallitus säätää lailla velvoitteen edellyttää avoimia rajapintoja julkisia tietojärjestelmiä hankittaessa, ellei painavasta syystä muuta johdu. Hallitus jatkaa määrätietoista julkisten tietovarantojen avaamista ja laaditaan niille hyödyntämistä helpottavat sitovat laatukriteerit. Lisäksi julkisuuslain periaatteet ja vaatimus tietovarantojen avaamisesta ulotetaan koskemaan myös julkisomisteisia yhtiöitä.

26.05.2020 / JAAKKO KUORIKOSKI JA SAMULI REIJULA

Laskennalliset mallit voivat lisätä julkisen päätöksenteon avoimuutta

Koronaviruspandemian mallinnukseen tarvitaan avointa dataa ja yhteistyötä

Kaiken EU:n tietosuoja-asetuksen salliman datan sekä viranomaisten käyttämien mallien tulisi olla saatavilla avoimesti ja viiveettä täydentävää mallinnusta varten.

Virus osuu pelkojemme ytimeen, ja avoin tieto on siihen parasta lääkettä

…elänyttä ajatella, että päättäjien ja asiantuntijoiden …istua kriisissä tiedon päällä, jottei syntyisi …sta.

Mielipide | Lukijan mielipide

Laskentamallit eivät lähtökohtaisesti ole salassa pidettävää tietoa

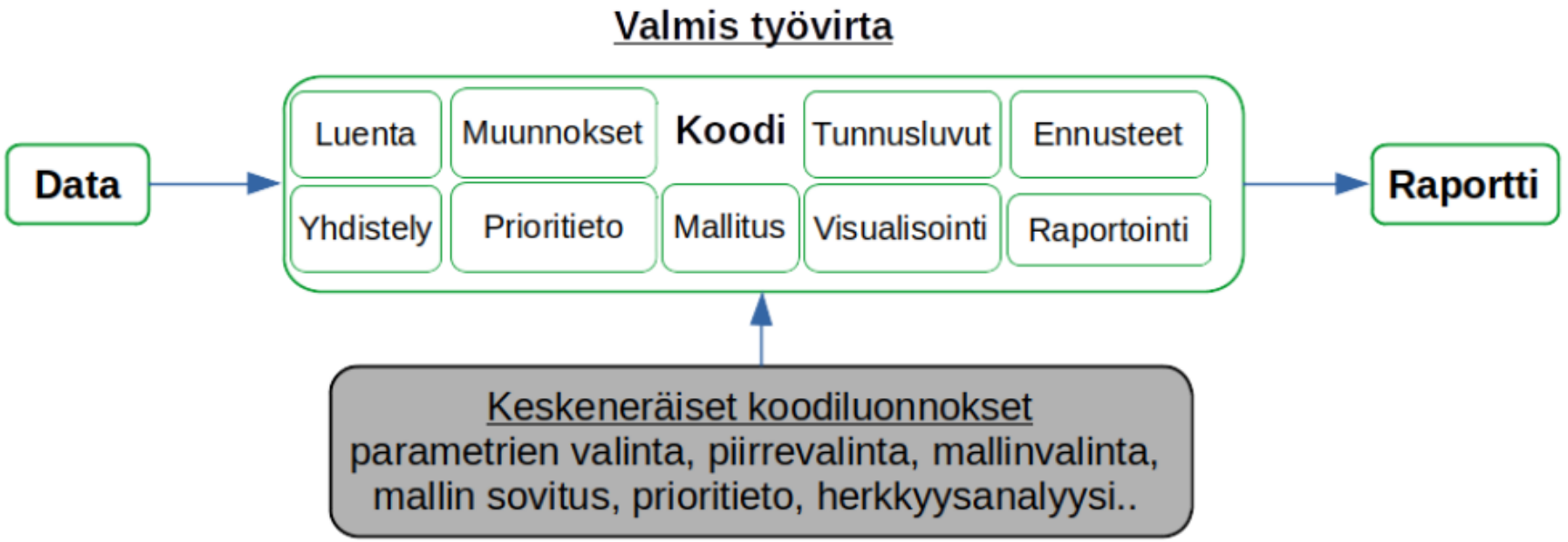
Epidemialaskelmien avoimuus on poliittinen valinta. Laskelmien avointa kehitystyötä tukemalla hallitus voisi edistää päätöksenteon läpinäkyvyyttä.

Mielipide | Lukijan mielipide

Keskustelu THL:n epidemiamalleista oli tärkeä oppi viranomaisille

Julkisuuslain soveltaminen algoritmeihin on juridisesti monin tavoin epäselvää.

21.7. 2:00



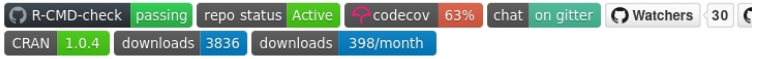
Kuva 2 Valmis työvirta kuvaa prosessin, jolla data tulkitaan raportoitavaksi tiedoksi.
Tulkinnan toteuttava *työvirta* sisältää monia vaiheita. Vakiintuneen määritelmän mukaan työvirta on avoin, kun se on jaettu julkisesti avoimella lisenssillä. Toimintaperiaatteiden tai työvirtaan sisältyvien yksittäisten mallien kuvailu yleisellä tasolla ei tee työvirrasta *avointa*. Tulkinnan jokainen vaihe ja virhe vaikuttaa lopullisiin johtopäätöksiin. Valmistelun aikana syntyy myös keskeneräisiä luonnoksia (harmaa laatikko), mutta *lopullinen työvirta lähdekoodeineen (valkea laatikko) on yhtä valmis kuin sen avulla laadittu raportti.*

The influence of hidden researcher decisions in applied microeconomics

Nick Huntington-Klein ✉, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

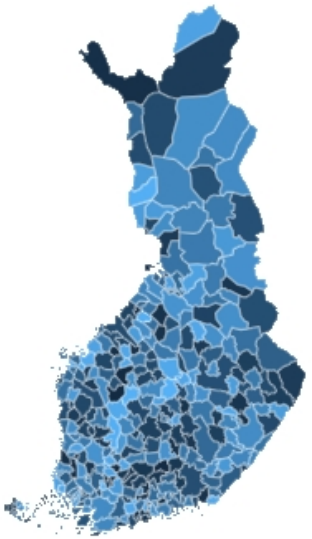
First published: 22 March 2021
https://doi.org/10.1111/ecin.12992

Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error.

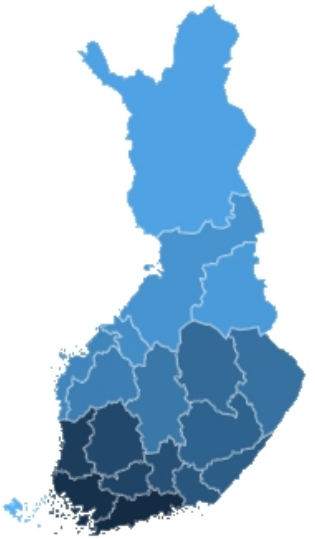


geofi - Access Finnish Geospatial Data

municipalities

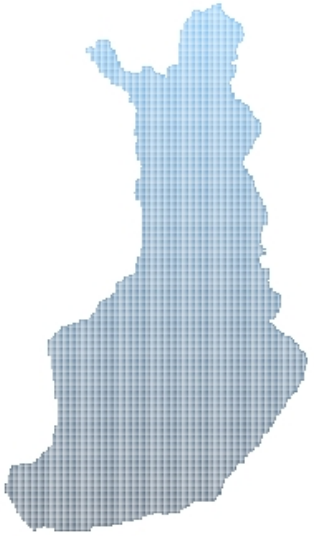


Aggregated municipality data at region (maakunta) level (one of many!)

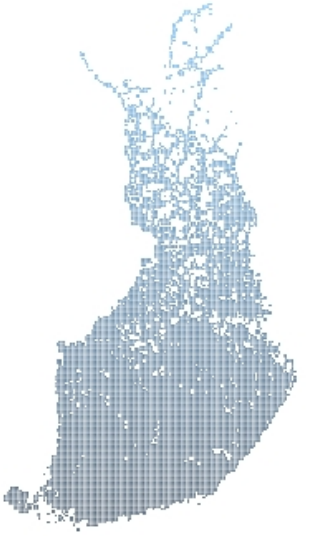


zipcodes

statistical grid

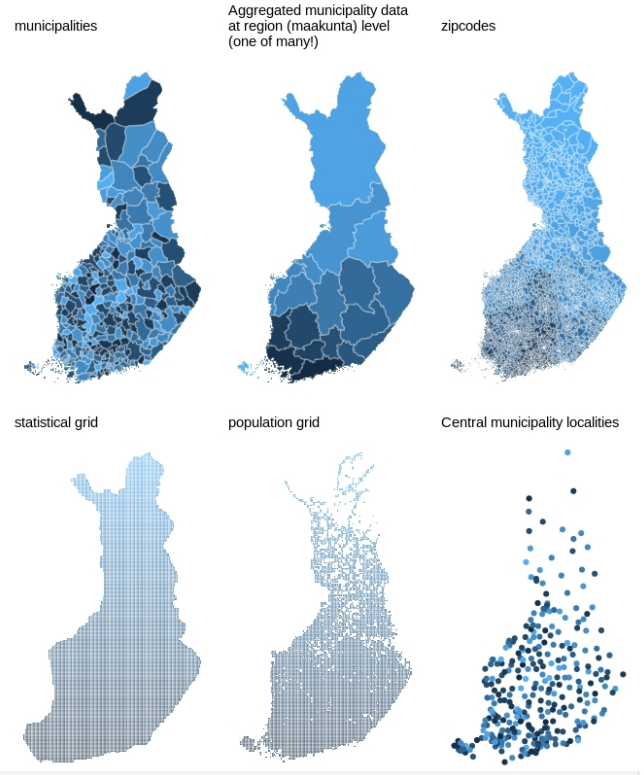


population grid



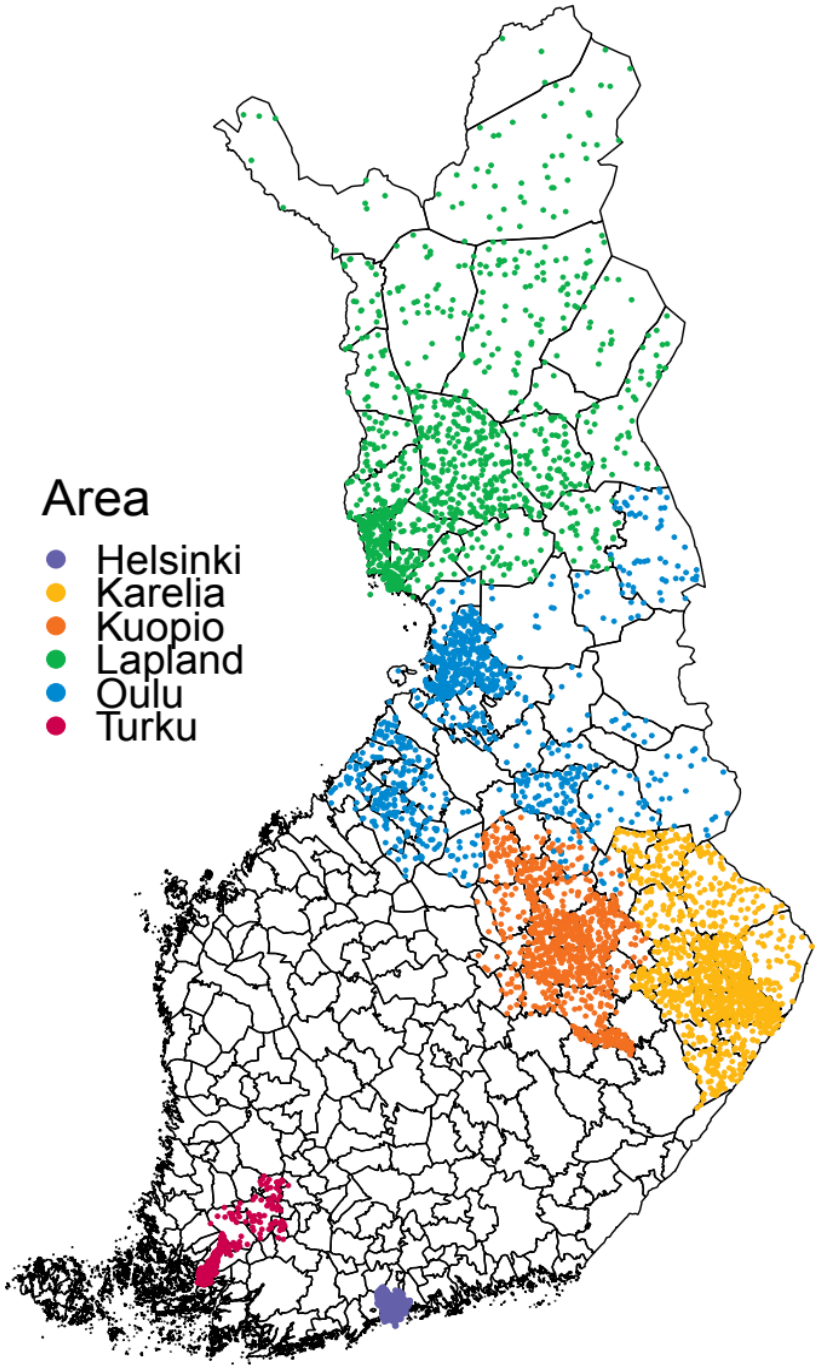
Central municipality localities

Full source code




```
library(geofi)
d1 <- get_municipalities(year = 2020)
d2 <- get_zipcodes(year = 2020)
d3 <- get_statistical_grid(resolution = 5)
d4 <- get_population_grid(resolution = 5)

library(ggplot2)
library(dplyr)
theme_set(
  theme_minimal(base_family = "Arial") +
  theme(legend.position= "none",
        axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank()
        )
)
p1 <- ggplot(d1, aes(fill = kunta)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "municipalities")
p2 <- ggplot(d1 %>% count(maakunta_code), aes(fill = maakunta_code)) + geom_sf(colour = alpha("white", 1/3)) + l
abs(subtitle = "Aggregated municipality data \nat region (maakunta) level \n(one of many!)")
p3 <- ggplot(d2, aes(fill = as.integer(posti_alue))) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "
zipcodes")
p4 <- ggplot(d3, aes(fill = nro)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "statistical grid")
p5 <- ggplot(d4, aes(fill = id_nro)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "population gri
d")
p6 <- ggplot(municipality_central_localities, aes(color = as.integer(kuntatunnus))) + geom_sf() + labs(subtitle
= "Central municipality localities")

library(patchwork)
wrap_plots(list(p1,p2,p3,p4,p5,p6), ncol = 3) +
  patchwork::plot_annotation(title = "Spatial data in geofi-package")
```

N=7231



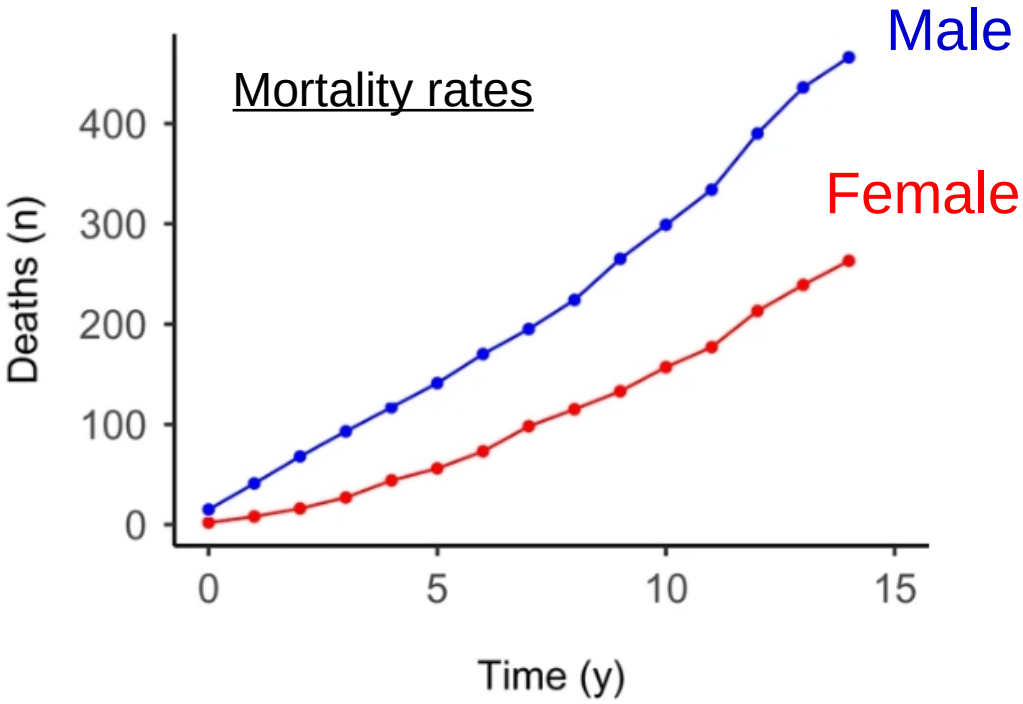
Article | Open Access | Published: 11 May 2021

Taxonomic signatures of cause-specific mortality risk in human gut microbiome

Aaro Salosensaari, Ville Laitinen, Aki S. Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo A. Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti ✉ & Teemu Niiranen ✉

Nature Communications **12**, Article number: 2671 (2021) | Cite this article

9060 Accesses | **1** Citations | **349** Altmetric | Metrics

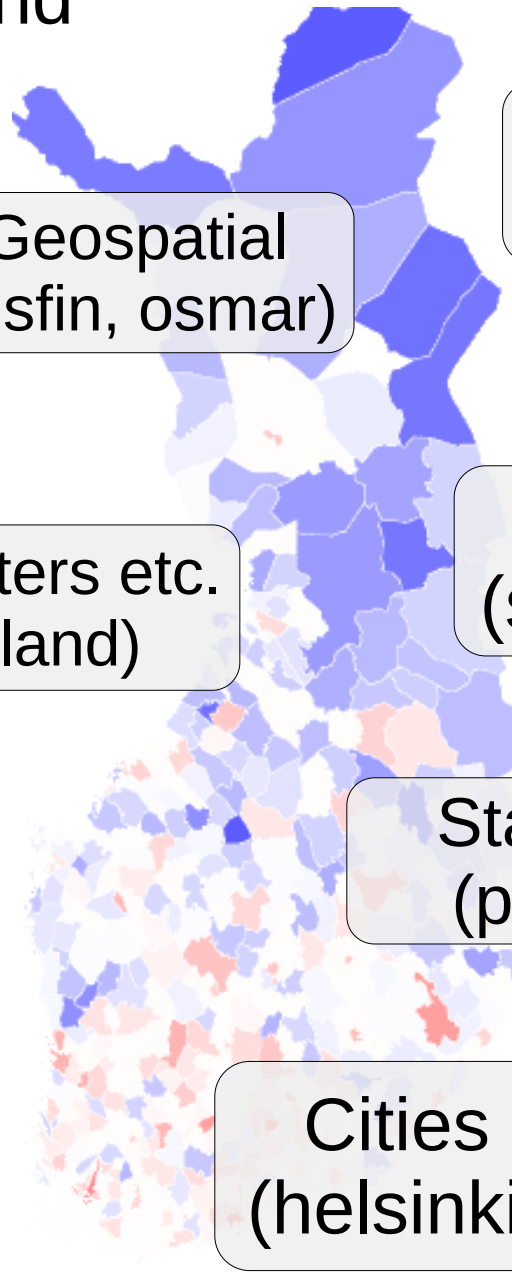


NATIONAL INSTITUTE FOR HEALTH AND WELFARE

From specific packages to package ecosystems

Open Street Map
Helsinki (osmar)

Algorithms for open data in Finland



Weather (fmi)

Geospatial (gisfin, osmar)

Health (sotkanet)

Registers etc. (finland)

Statistics (pxweb)

Cities (helsinki)

datavaalit

MAANMITTAUSLAITOS
LANTMÄTERIVERKET

16 40 KANSALLIS KIRJASTO

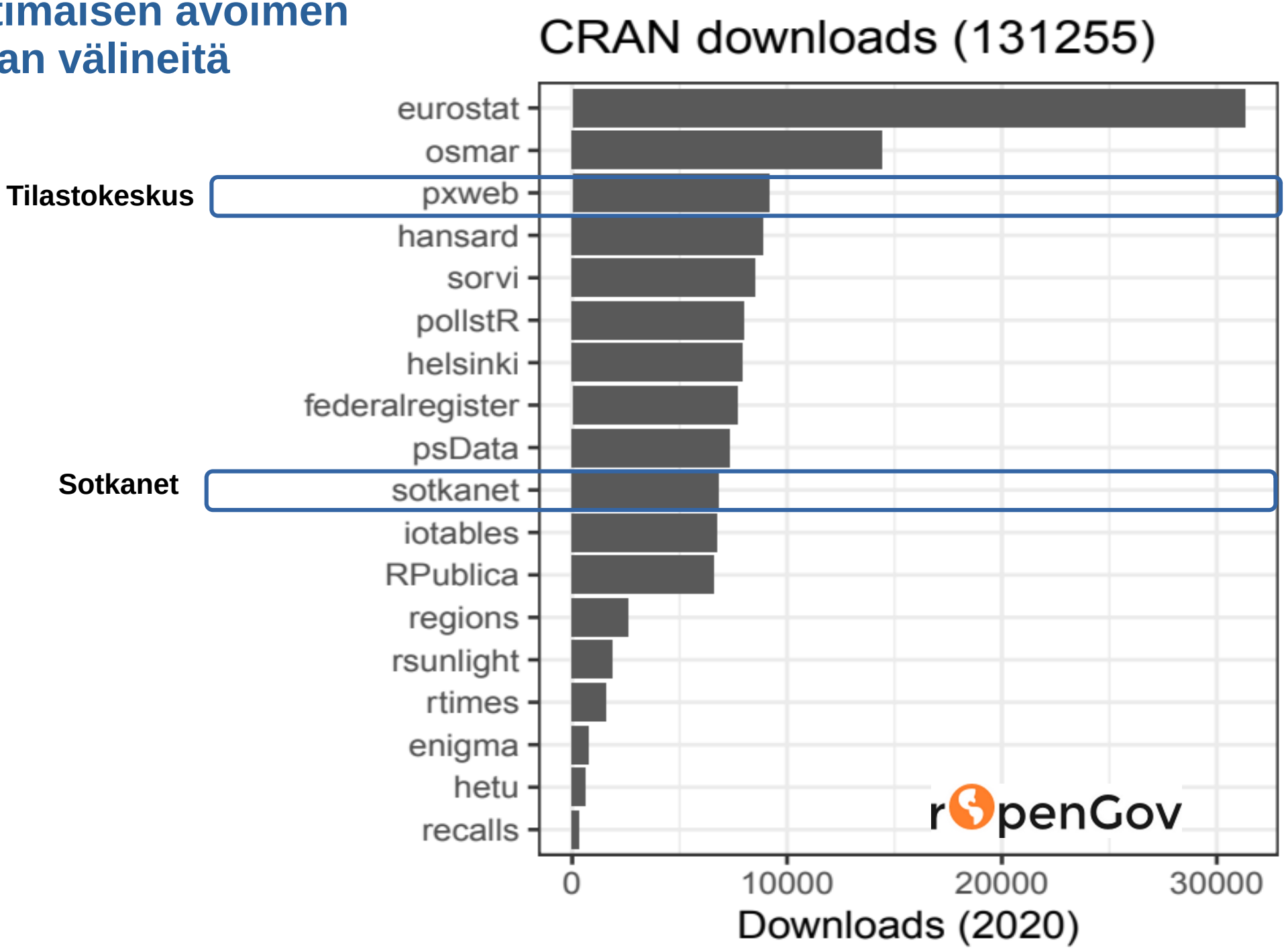
Tilastokeskus
Statistikcentralen
Statistics Finland

SITRA

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

pxweb for PX-Web/PC-Axis data from stats authorities in: Denmark, Finland, Greenland, Iceland, Latvia, Norway, Sweden.. **world bank**, **FAO**

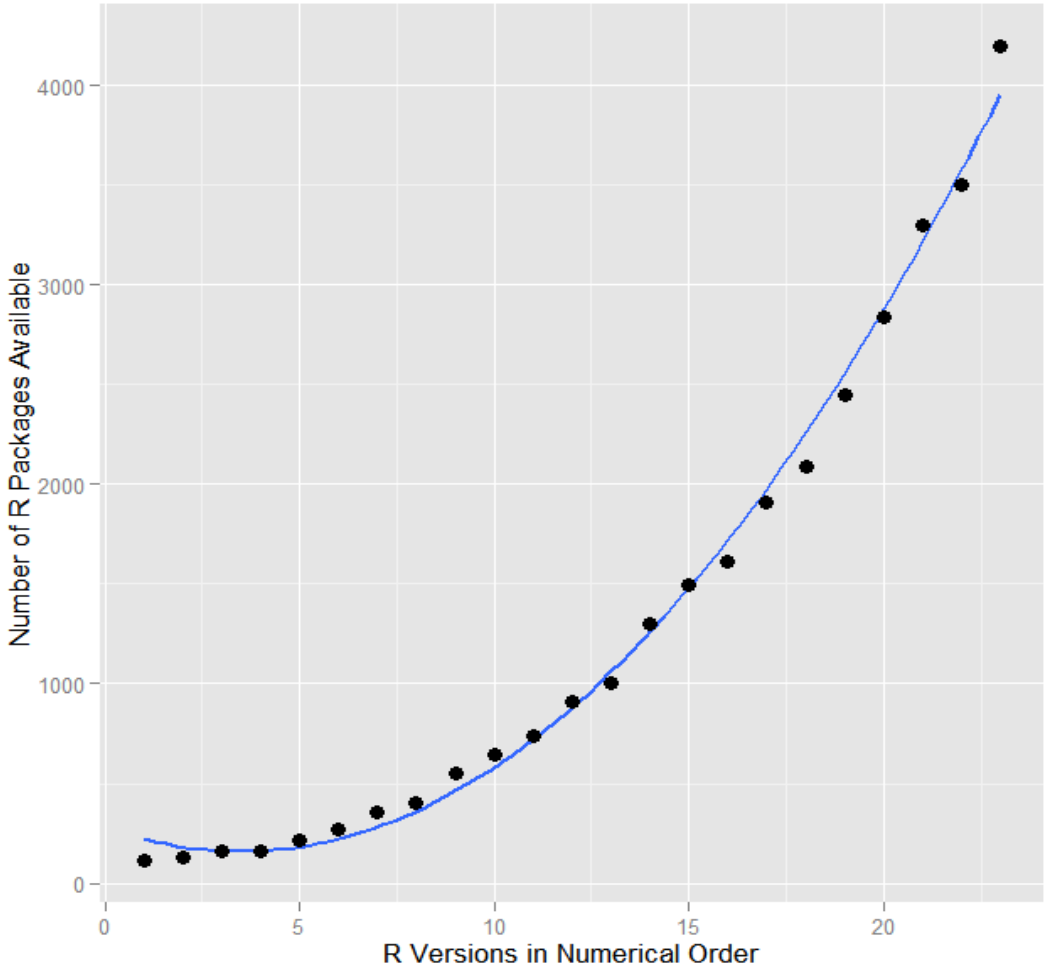
Kotimaisen avoimen datan välineitä



Tilastokeskus

Sotkanet

Number of open analysis tools has grown exponentially



github SOCIAL CODING

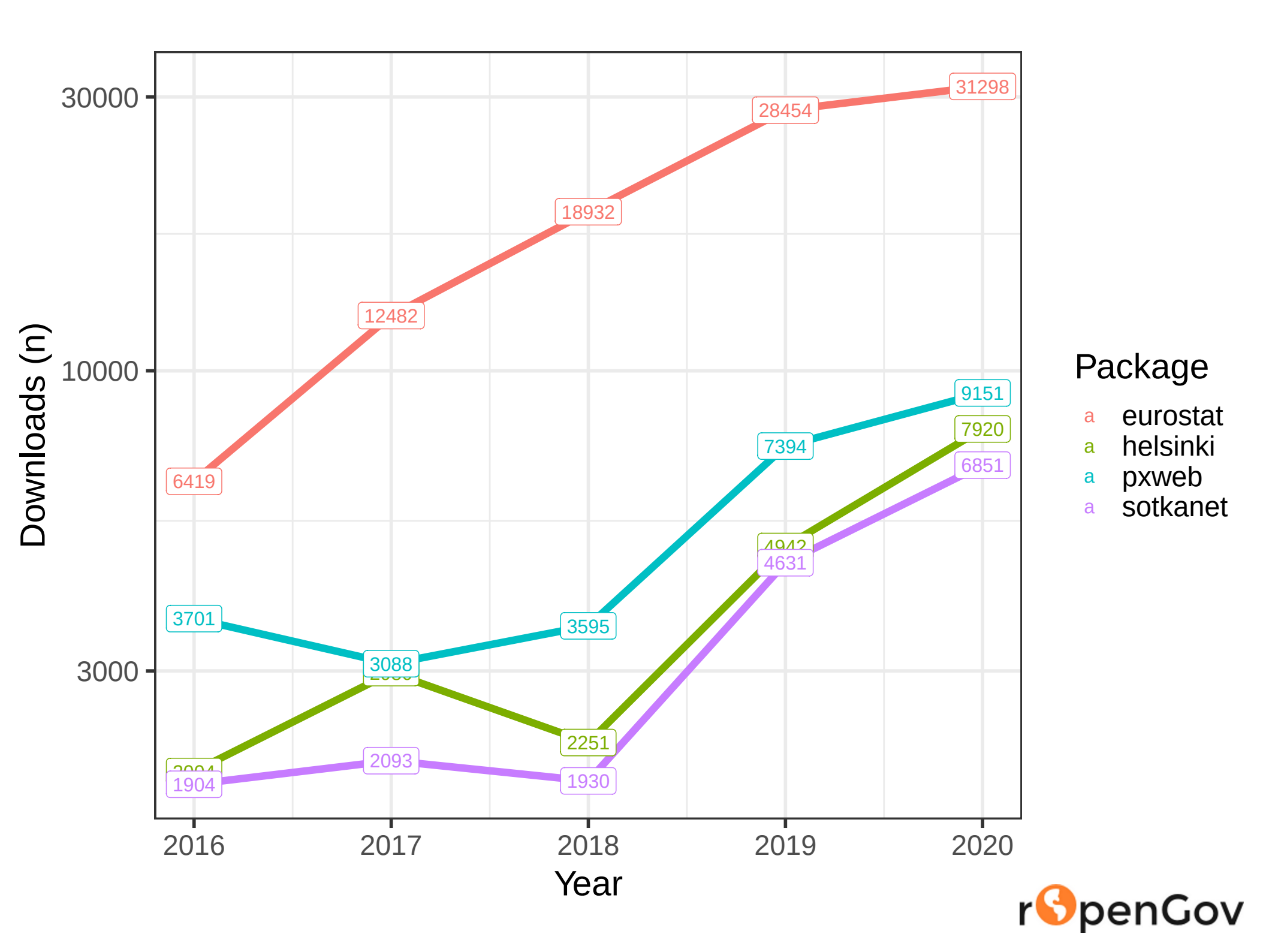
R

rOpenGov

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ROpenSci

Value of data can increase through sharing & use


30000
10000
3000
Downloads (n)
2016
2017
2018
2019
2020
Year
6419
12482
18932
28454
31298
3701
3088
3595
7394
9151
1904
2093
1930
4631
6851
2251
7920
Package
eurostat
helsinki
pxweb
sotkanet
rOpenGov

Tutorials, vignettes, training material..

The *eurostat* package

R tools to access open data from Eurostat database

Search and download

Data in the Eurostat database is stored in tables. Each table has an identifier, a short table_code, and a description (e.g. *tsdtr420* - People killed in road accidents).
Key eurostat functions allow to find the table_code, download the eurostat table and polish labels in the table.

Find the table code

The **search_eurostat(pattern, ...)** function scans the directory of Eurostat tables and returns codes and descriptions of tables that match pattern.

```
library("eurostat")
query <- search_eurostat("road", type = "table")
query[1:3,1:2]
##                                  title     code
## 1          Goods transport by road ttr00005
## 2   People killed in road accidents tsdtr420
## 3 Enterprises with broadband access tin00090
```

Download the table

The **get_eurostat(id, time_format = "date", filters = "none", type = "code", cache = TRUE, ...)** function downloads the requested table from the *Eurostat bulk download facility* or from *The Eurostat Web Services JSON API* (if **filters** are defined). Downloaded data is cached (if **cache=TRUE**). Additional arguments define how to read the time column (**time_format**) and if table dimensions shall be kept as codes or converted to labels (**type**).

```
dat <- get_eurostat(id="tsdtr420", time_format="num")
head(dat)
##     unit    sex    geo  time values
## 1     NR      T     AT  1999   1079
## 2     NR      T     BE  1999   1397
## 3     NR      T     CZ  1999   1455
## 4     NR      T     DK  1999    514
## 5     NR      T     EL  1999   2116
## 6     NR      T     ES  1999   5738
```

Add labels

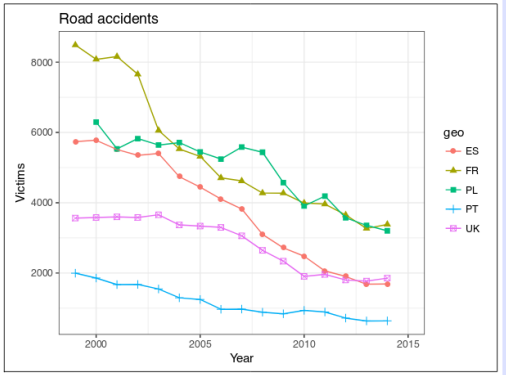
The **label_eurostat(x, lang = "en", ...)** gets definitions for Eurostat codes and replace them with labels in given language ("en", "fr" or "de").

```
dat <- label_eurostat(dat)
head(dat)
##     unit    sex            geo  time values
## 1 Number  Total         Austria  1999   1079
## 2 Number  Total         Belgium  1999   1397
## 3 Number  Total Czech Republic  1999   1455
## 4 Number  Total         Denmark  1999    514
## 5 Number  Total          Greece  1999   2116
## 6 Number  Total           Spain  1999   5738
```

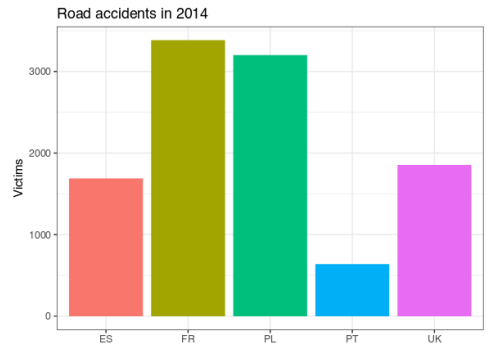
eurostat and plots

The **get_eurostat()** function returns tibbles in the long format. Packages **dplyr** and **tidyr** are well suited to transform these objects. The **ggplot2** package is well suited to plot these objects.

```
t1 <- get_eurostat("tsdtr420", filters =
        list(geo = c("UK", "FR", "PL", "ES", "PT")))

library("ggplot2")
ggplot(t1, aes(x = time, y = values, color = geo,
               group = geo, shape = geo)) +
  geom_point(size = 2) +
  geom_line() + theme_bw() +
  labs(title="Road accidents", x = "Year", y = "Victims")
```




```
library("dplyr")
t2 <- t1 %>% filter(time == "2014-01-01")
ggplot(t2, aes(geo, values, fill=geo)) +
  geom_bar(stat = "identity") + theme_bw() +
  theme(legend.position = "none")+
  labs(title="Road accidents in 2014", x="", y="Victims")
```



eurostat and maps

Fetch and process data

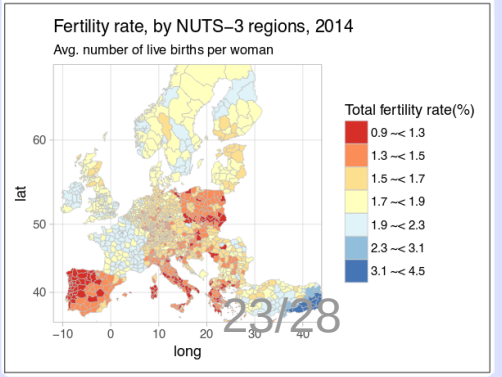
There are three function to work with geospatial data from GISCO. The **get_eurostat_geospatial()** returns preprocessed spatial data as sp-objects or as data frames. The **merge_eurostat_geospatial()** both downloads and merges the geospatial data with a preloaded tabular data. The **cut_to_classes()** is a wrapper for **cut()** - function and is used for categorizing data for maps with tidy labels.

```
library("eurostat")
library("dplyr")

fertility <- get_eurostat("demo_r_frate3")  %>%
  filter(time == "2014-01-01") %>%
  mutate(cat = cut_to_classes(values, n=7, decimals=1))

mapdata <-  merge_eurostat_geodata(fertility,
              resolution = "20")

head(select(mapdata,geo,values,cat,long,lat,order,id))
##     geo values        cat     long      lat order id
## 1 AT124   1.39 1.3 –< 1.5 15.54245 48.90770   214 10
## 2 AT124   1.39 1.3 –< 1.5 15.75363 48.85218   215 10
## 3 AT124   1.39 1.3 –< 1.5 15.88763 48.78511   216 10
## 4 AT124   1.39 1.3 –< 1.5 15.81535 48.69270   217 10
## 5 AT124   1.39 1.3 –< 1.5 15.94094 48.67173   218 10
## 6 AT124   1.39 1.3 –< 1.5 15.90833 48.59815   219 10
```

Draw a cartogram

The object returned by **merge_eurostat_geospatial()** are ready to be plotted with ggplot2 package. The **coord_map()** function is useful to set the projection while **labs()** adds annotations o the plot.

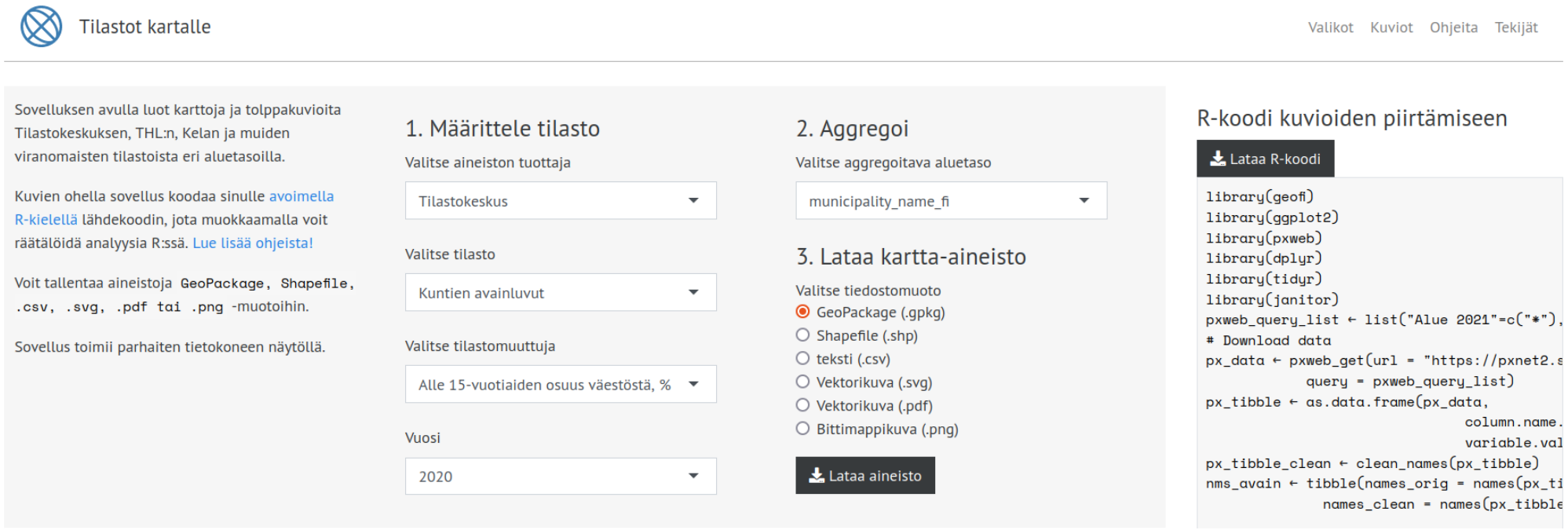
```
library("ggplot2")
ggplot(mapdata, aes(x = long, y = lat, group = group))+
  geom_polygon(aes(fill=cat), color="grey", size = .1)+
  scale_fill_brewer(palette = "RdYlBu") +
  labs(title="Fertility rate, by NUTS-3 regions, 2014",
       subtitle="Avg. number of live births per woman",
       fill="Total fertility rate(%)") + theme_light()+
  coord_map(xlim=c(-12,44), ylim=c(35,67))
```

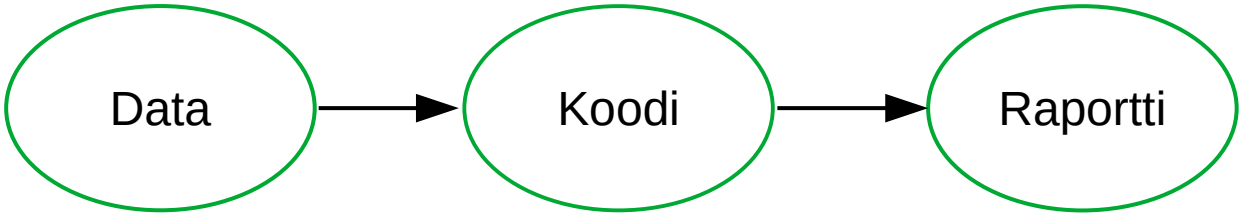


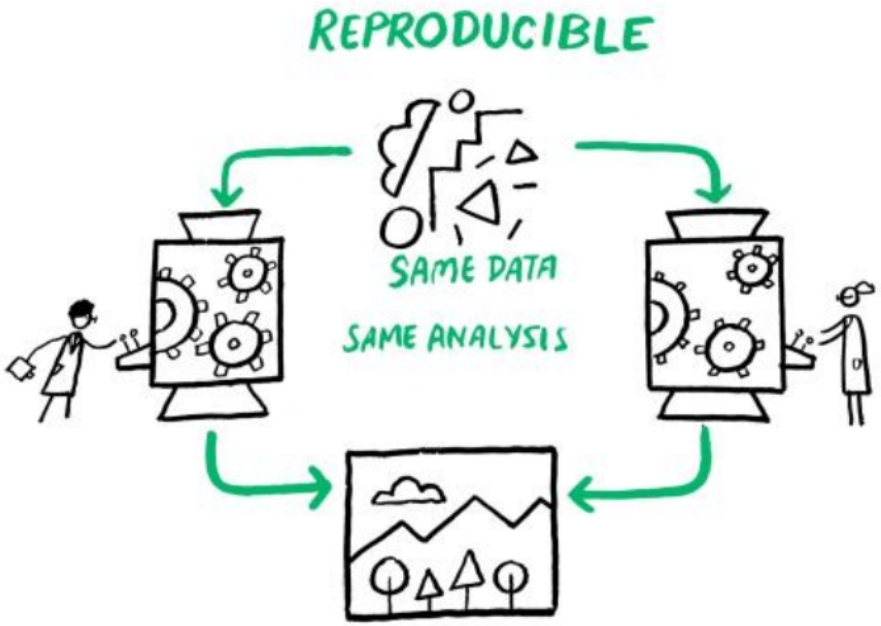
This onepager presents the **eurostat** package
Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek 2014-2017 package version 2.2.43 URL: *https://github.com/rOpenGov/eurostat*

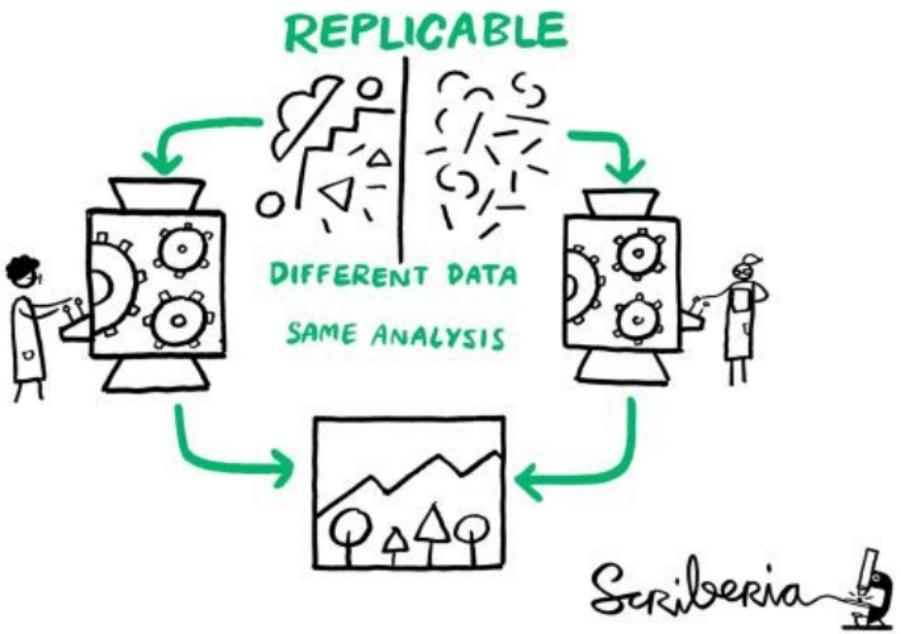
Demo https://ropengov.shinyapps.io/tilastot_kartalle/

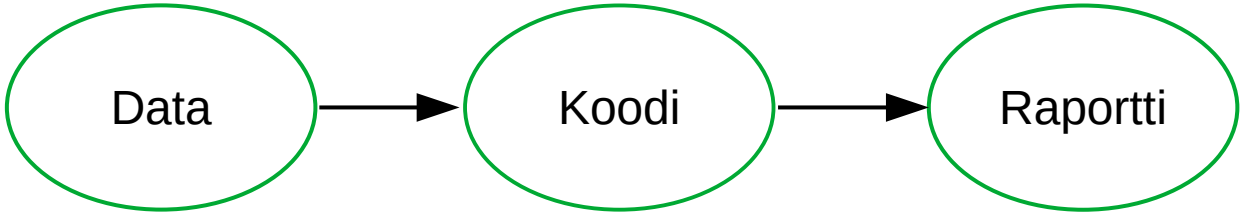
Havainnollistaa kotimaisia R-paketteja (väestöryhmittäiset terveyserot)









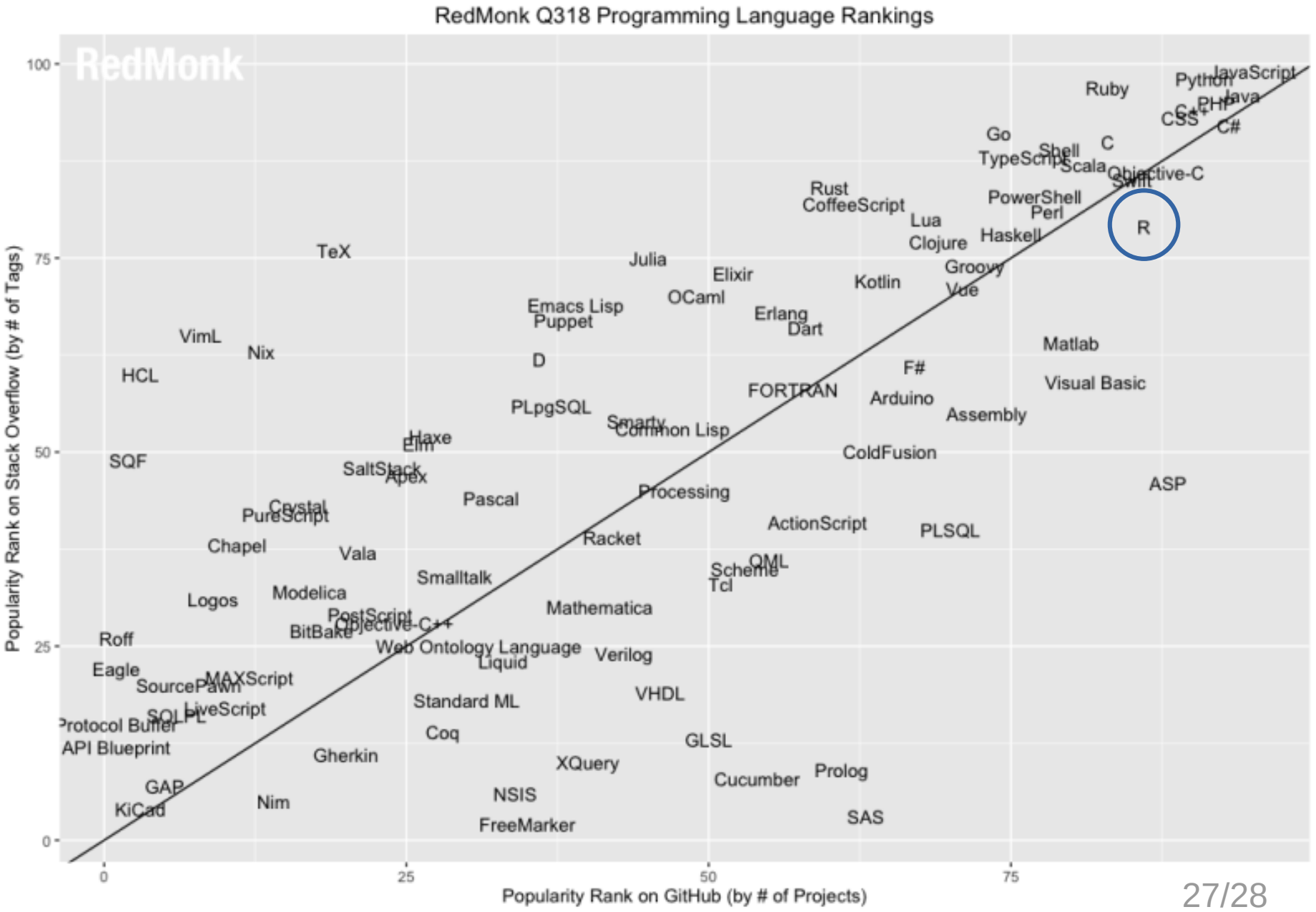


Beyond Open Access - The Changing Culture of Producing and Disseminating Scientific Knowledge

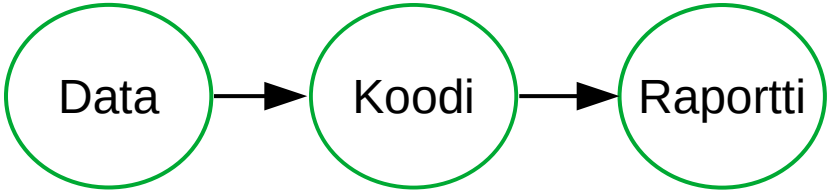
Heidi Laine Leo Lahti Anne Lehto

Pyry Kantanen

Cultures of open data science collaboration



rOpenGov



Welcome to rOpengov-project!

We have been waiting for you!

rOpenGov is a community of R package developers on open government data analytics and related topics.

The open collaboration network was initiated in 2010 and has since then led to many R packages and other fruitful outcomes. A number of independent authors have contributed through github and written to the rOpenGov blog.

You are welcome to check out our projects and join us. Proposals for new collaborations are also always welcome! See the community page for info on the people behind rOpenGov and get in touch.

Recent Posts

1. Regions package for Eurostat sub-national statistics
2. Economic and environmental impact analysis with iotables
3. Visualizing City of Helsinki procurements with geofi-package
4. Hetu-package for handling of Finnish personal identity codes
5. geofi R-package for accessing Statistics Finland spatial data

Categories

data latex news paikkatieto poster r r-package research sweave tiedonlouhinta tikz visualisointi

Tags

dataviz digital-humanities eu-datathon eurostat finland foi geofi ggplot2