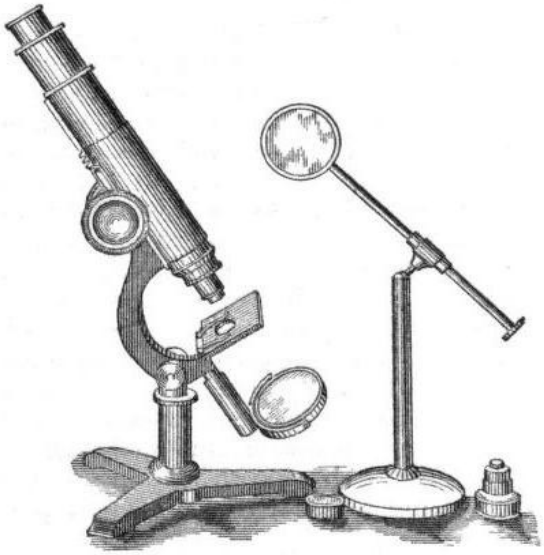
Open Data Science

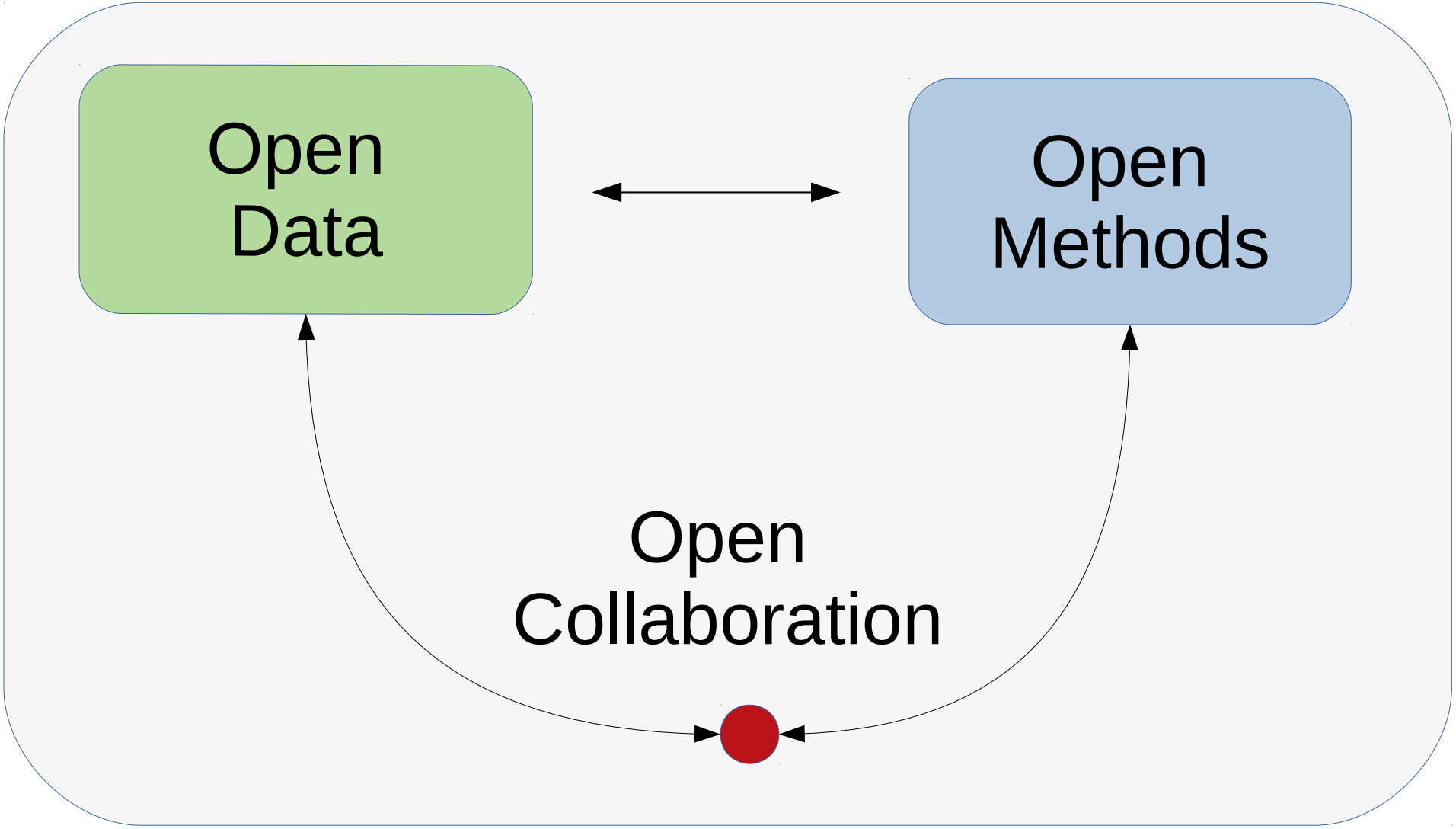
Leo Lahti
University of Turku, Finland

@openreslabs

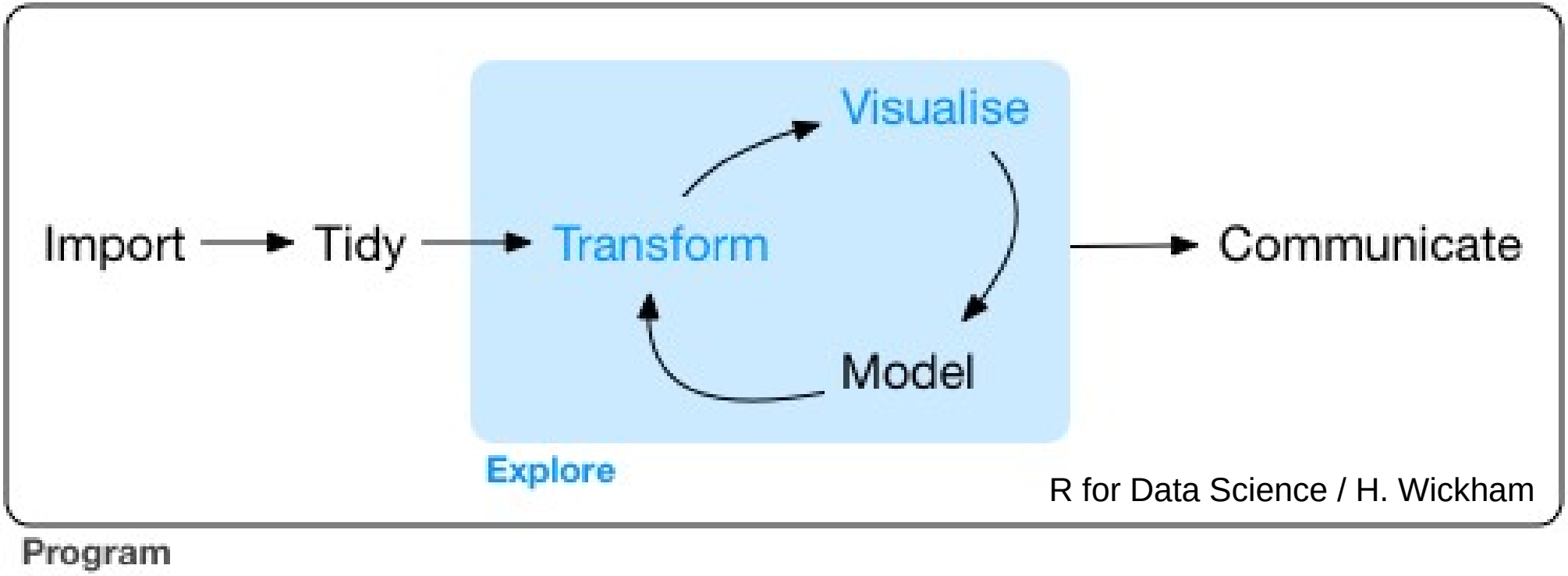


Turun yliopisto
University of Turku

Elements of open data science



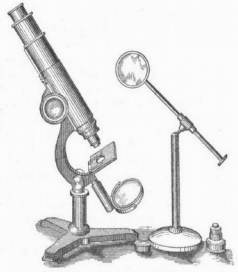
Data Science Workflow



Data
- raw
- supporting

Open data science ecosystem

IP[y]:
IPython



Outputs
- understanding
- reporting
- reuse

1. Open Data

Human genome project:
a prime example on research data sharing
enabled timely genome sequencing and assembly

Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

Reuse of public genome-wide gene expression data

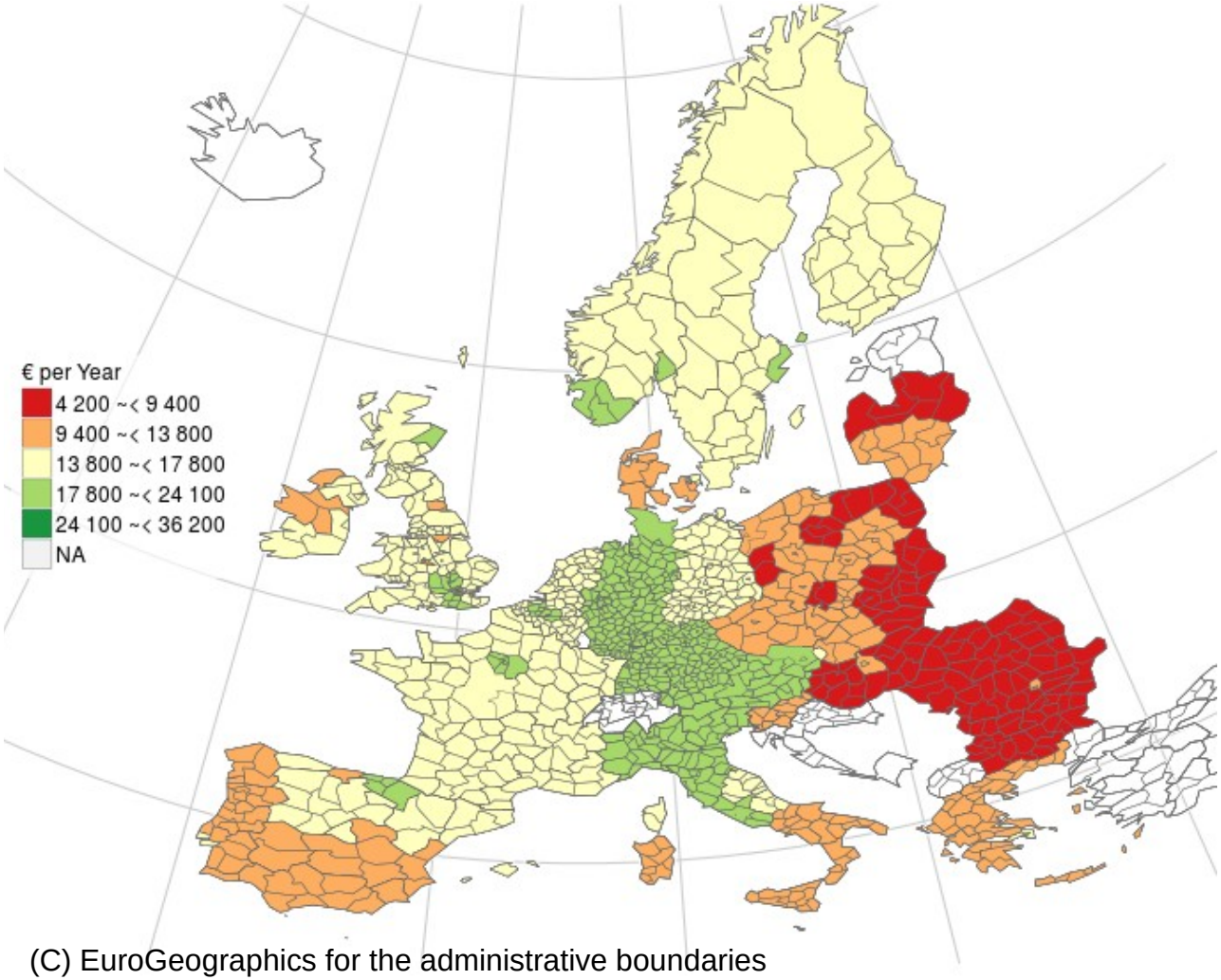
Johan Rung[1] & Alvis Brazma[1] About the authors

DRYAD

Open data now supporting research & enabling cross-disciplinary collaboration and discovery at an unprecedented scale from natural to social sciences and humanities.

Open data: from natural sciences to humanities

Eurostat: average household expenditure in 2011



(C) EuroGeographics for the administrative boundaries

Science e.g. human genome project, EBI

Health National Institutes of Health

Populations & demography Eurostat, FAO, National statistical authorities

Economics World Bank

Cultural heritage Digitized collections of books, artwork etc.

Weather Finnish Meteorological Institute

Geospatial Open Street Maps; Geonames; Land Survey Finland

OpenStreetMap
The Free Wiki World Map

Tilastokeskus
Statistikcentralen
Statistics Finland

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

EEBO
EARLY ENGLISH BOOKS ONLINE

ECCO TCP
Eighteenth Century Collections Online
Text Creation Partnership

THE WORLD BANK

MAANMITTAUSLAITOS
LANTMÄTERIVERKET

1640
KANSALLIS KIRJASTO

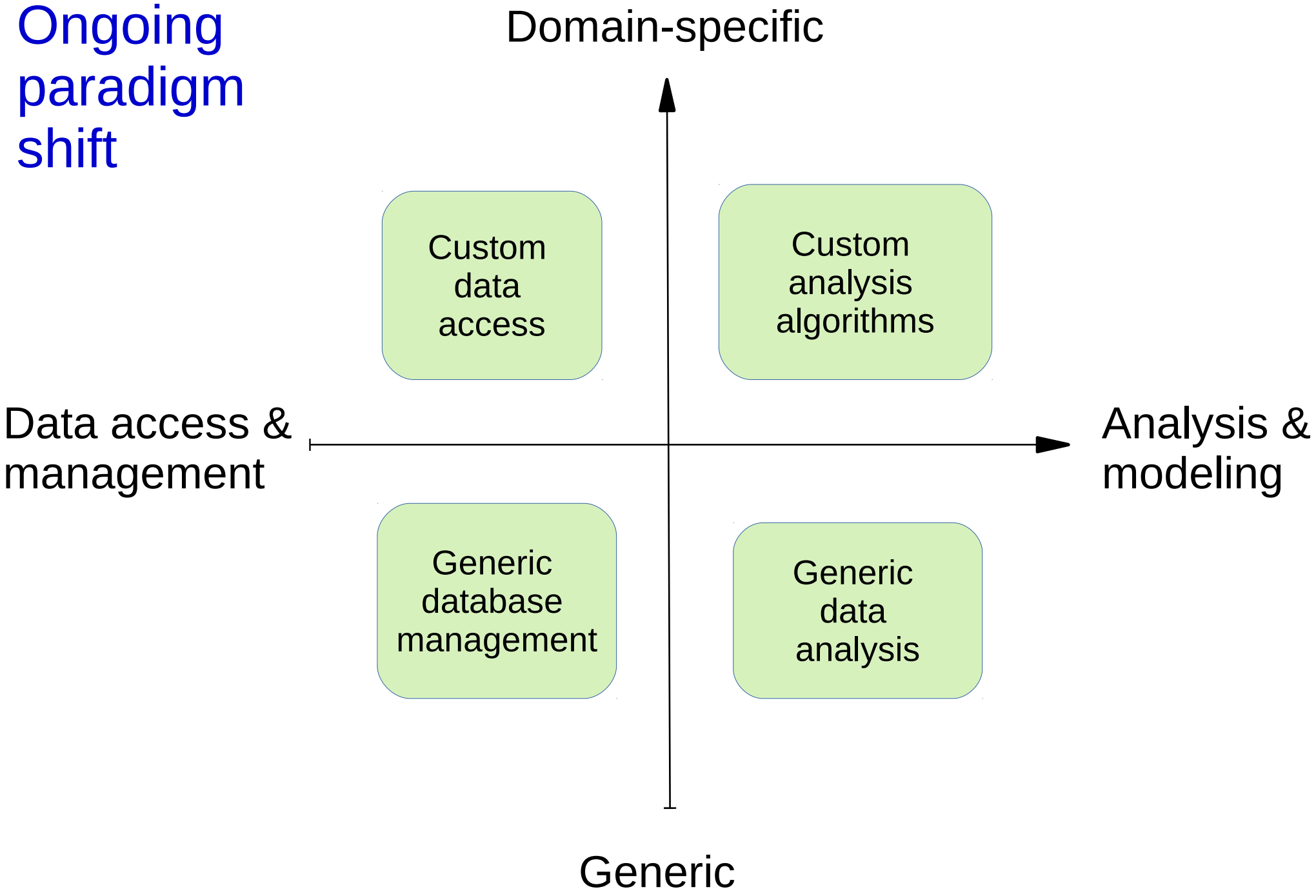
rOpenGov

Bottlenecks in data access

- **F**indability
- **A**ccessibility (scattered, noisy, incomplete, non-machine-readable)
- **I**nteroperability
- **R**eusability (quality, rights, life span, documentation..)

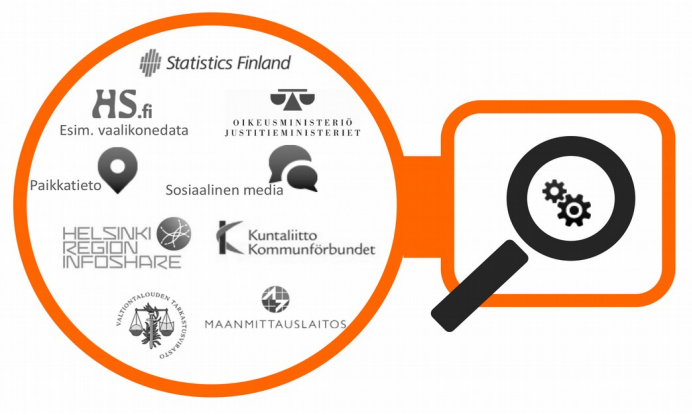
Technical, cultural, and historical challenges still forming **bottlenecks for research code and data sharing**

Ongoing paradigm shift

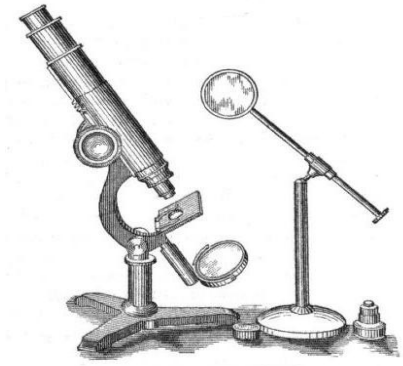


2. Open Methods

Computational workflows have an increasingly central role in research



Science 13 April 2012:
Vol. 336 no. 6078 pp. 159-160
DOI: 10.1126/science.1218263



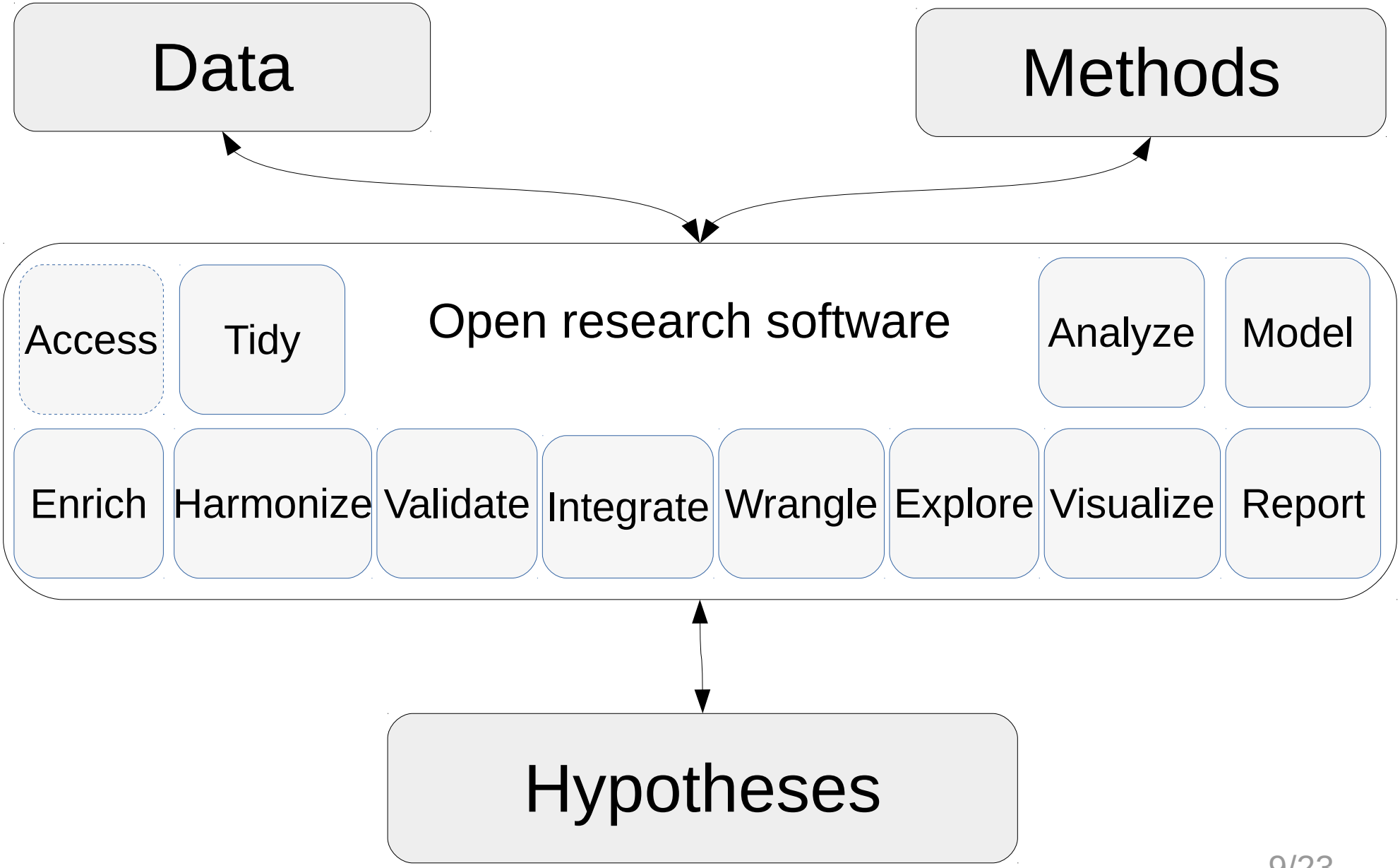
POLICY FORUM

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,*]

Open methods development can complement FAIR data sharing



PLOS | MEDICINE

Browse | For Authors | About Us | S

OPEN ACCESS

ESSAY

898,944 VIEWS | 1,119 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

How to Make More Published Research True

John P. A. Ioannidis

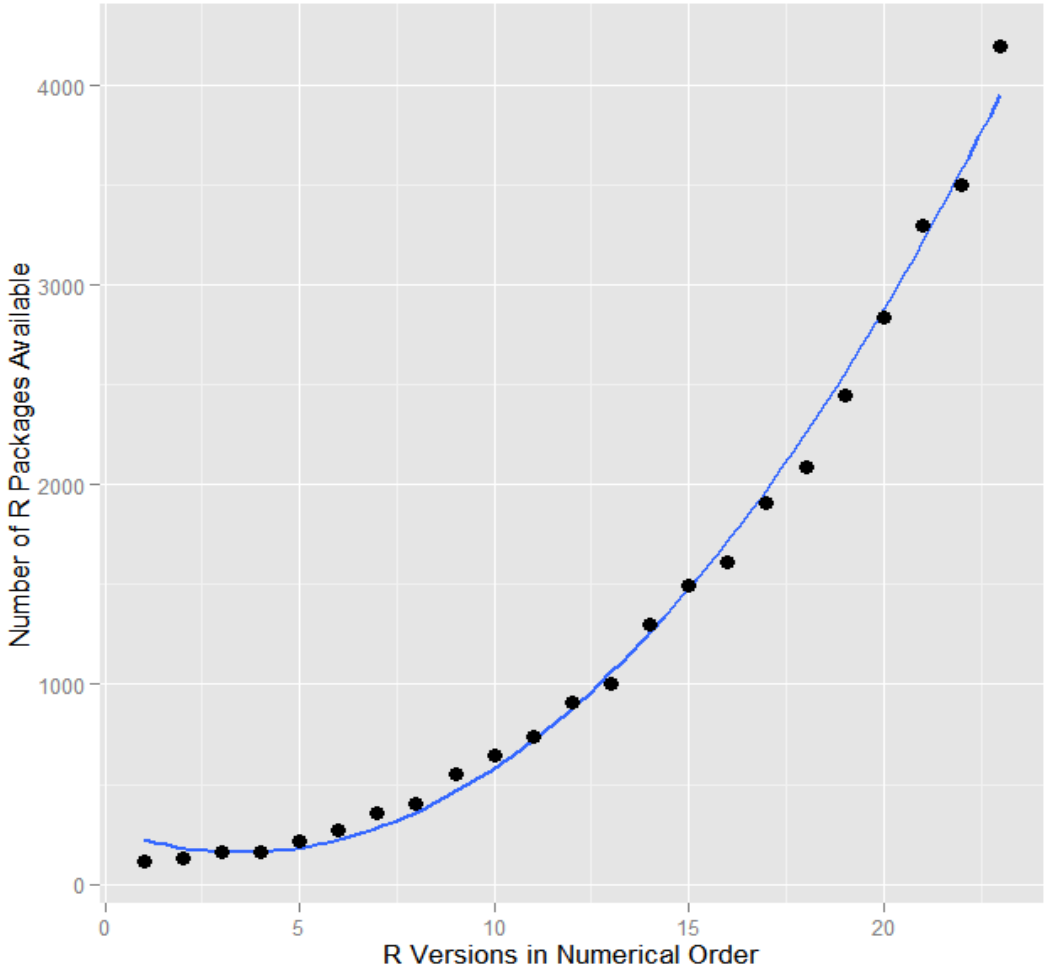
Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

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,*]

Number of open analysis tools has grown exponentially



3. Open Collaboration

Open collaboration through
open research software ecosystems
can improve efficiency & methods base

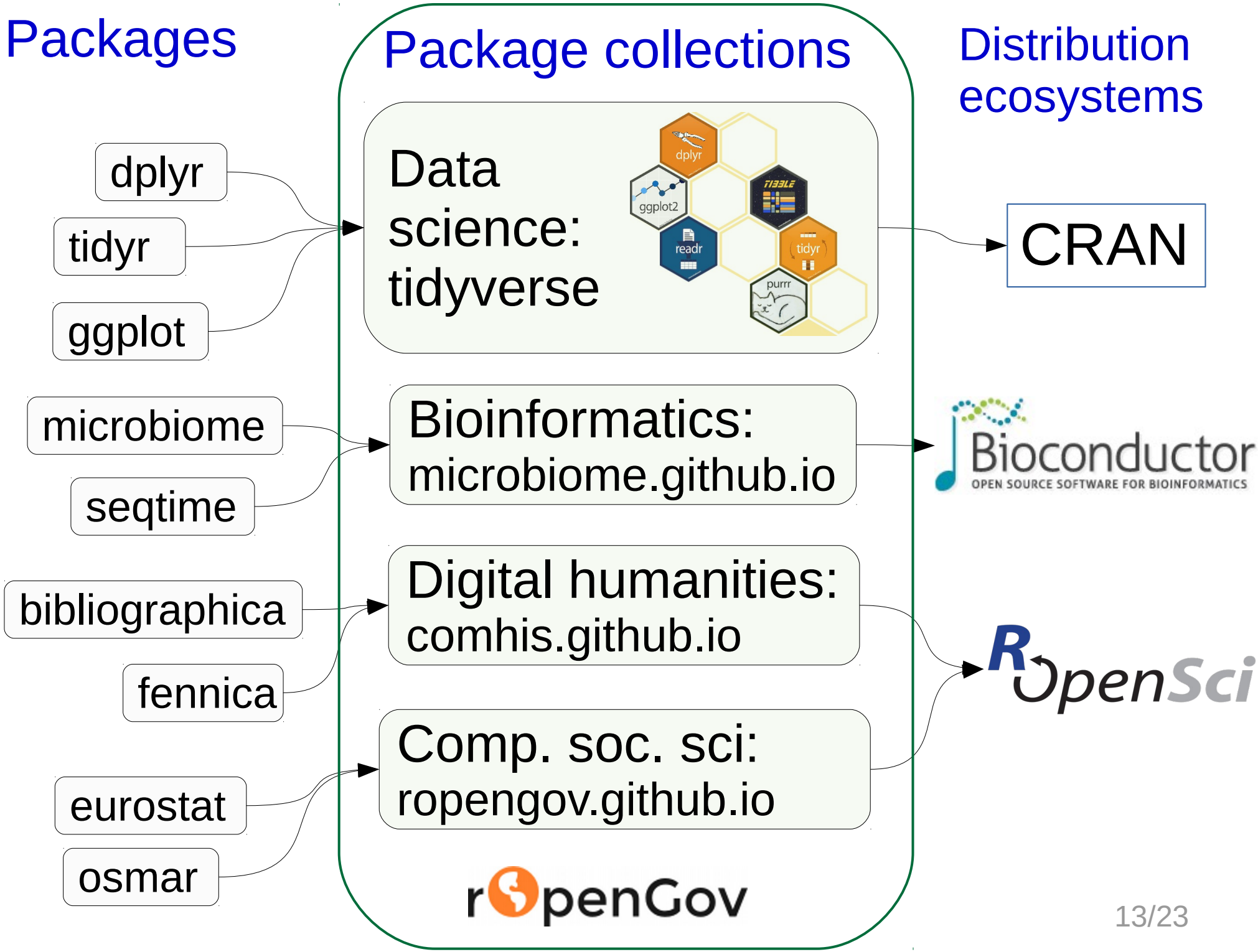
Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas

Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang & Adam A Margolin

Affiliations | Contributions | Corresponding authors

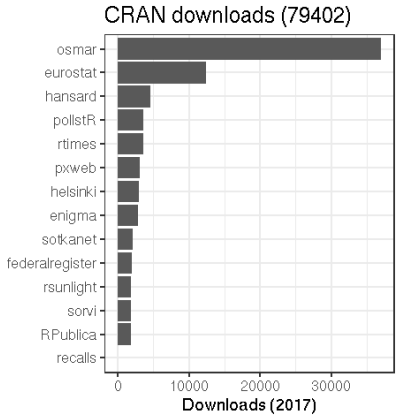
Nature Genetics **45**, 1121–1126 (2013) | doi:10.1038/ng.2761
Published online 26 September 2013

Social coding revolution

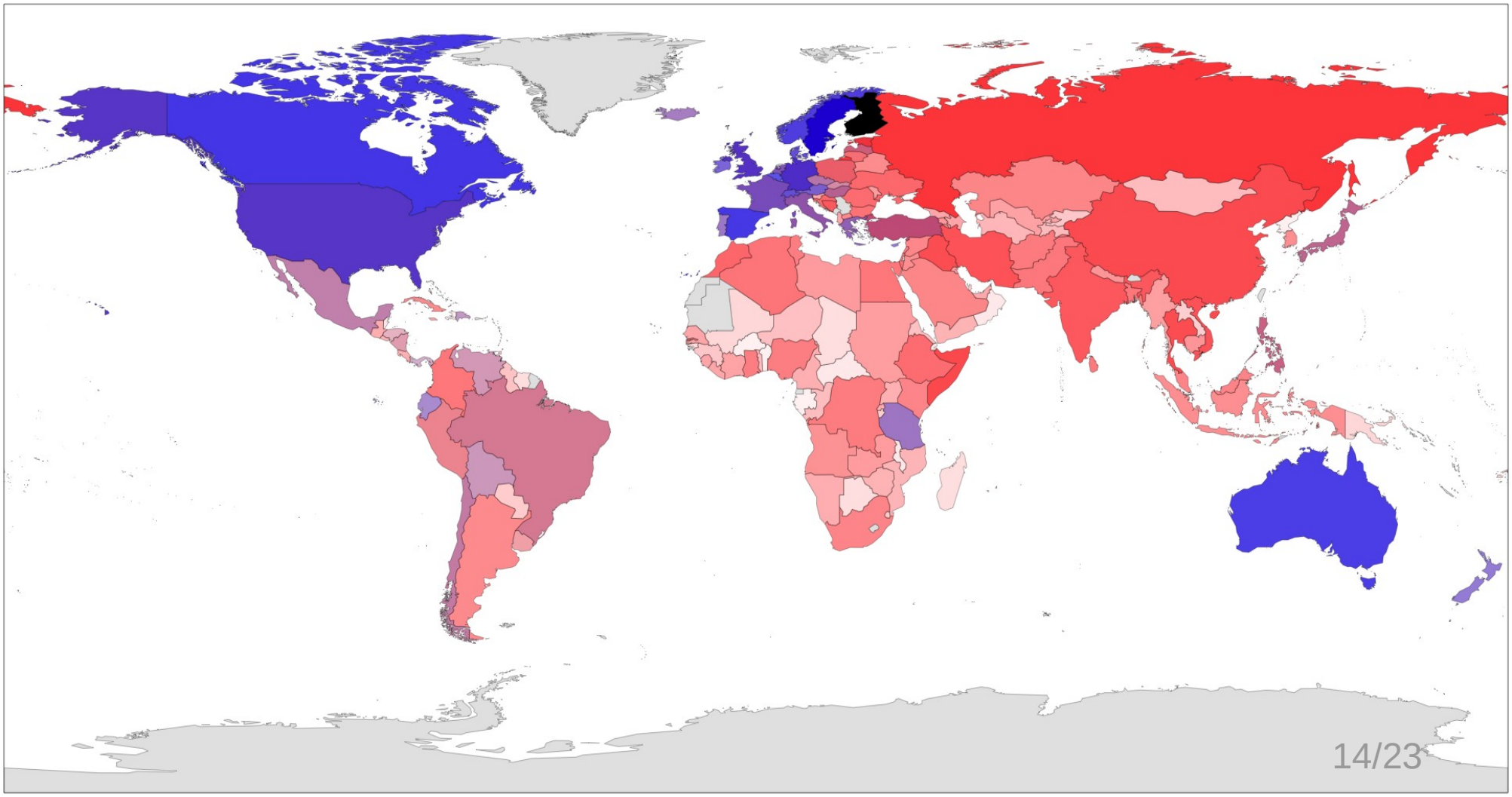

Packages
Package collections
Distribution ecosystems
dplyr
tidyr
ggplot
microbiome
seqtime
bibliographica
fennica
eurostat
osmar
Data science: tidyverse
Bioinformatics: microbiome.github.io
Digital humanities: comhis.github.io
Comp. soc. sci: ropengov.github.io
CRAN
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
ROpenSci
rOpenGov

rOpenGov

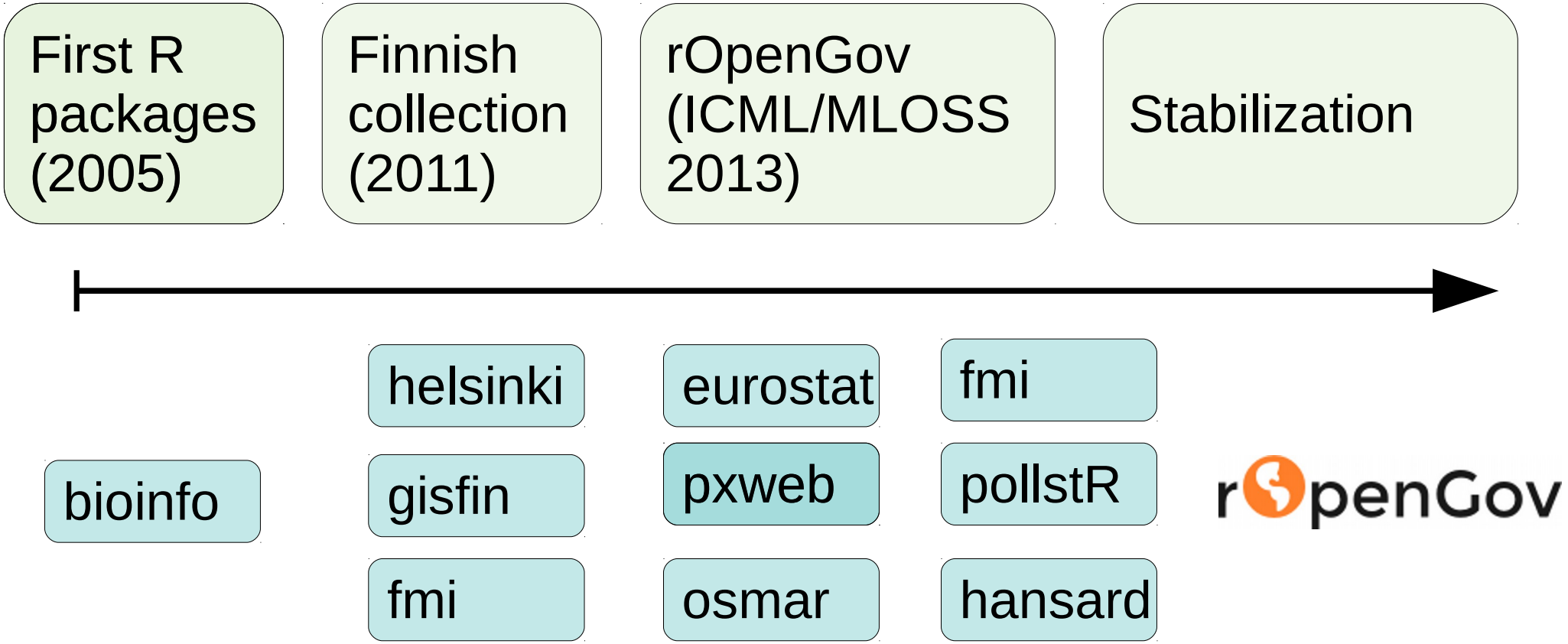
- 20+ pkgs (14 in CRAN)
- 80k downloads 2017



open data science network



Timeline on community building



Retrieval and Analysis of Eurostat Open Data with the eurostat Package

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

R packages:
- eurostat
- eurostat_geospatial

Documentation & cheat sheets

Online tutorials & blog posts

Issue tracker

Automated unit tests

Project homepage

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

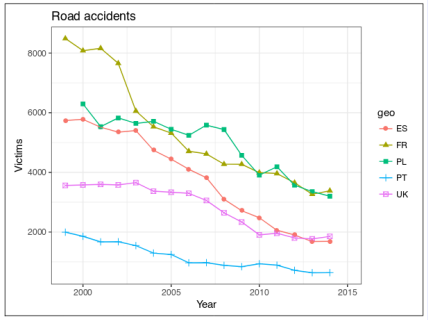
The **label_eurostat(x, lang = "en", ...)** gets definitions for Eurostat codes and replace them with labels in given language ("en", "fr" or "de").

```
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
```

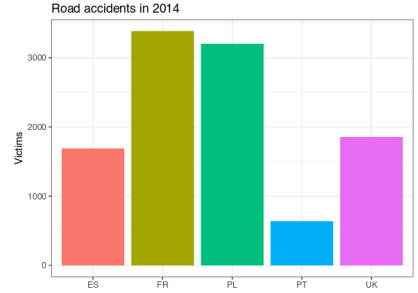
eurostat and plots

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

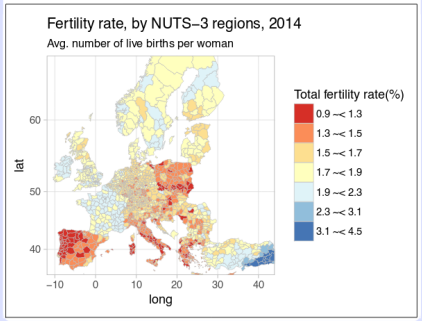
There are three function to work with geospatial data from GISCO. The **get_eurostat_geospatial()** returns preprocessed spatial data as sp-objects or as data frames. The **merge_eurostat_geospatial()** both downloads and merges the geospatial data with a preloaded tabular data. The **cut_to_classes()**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

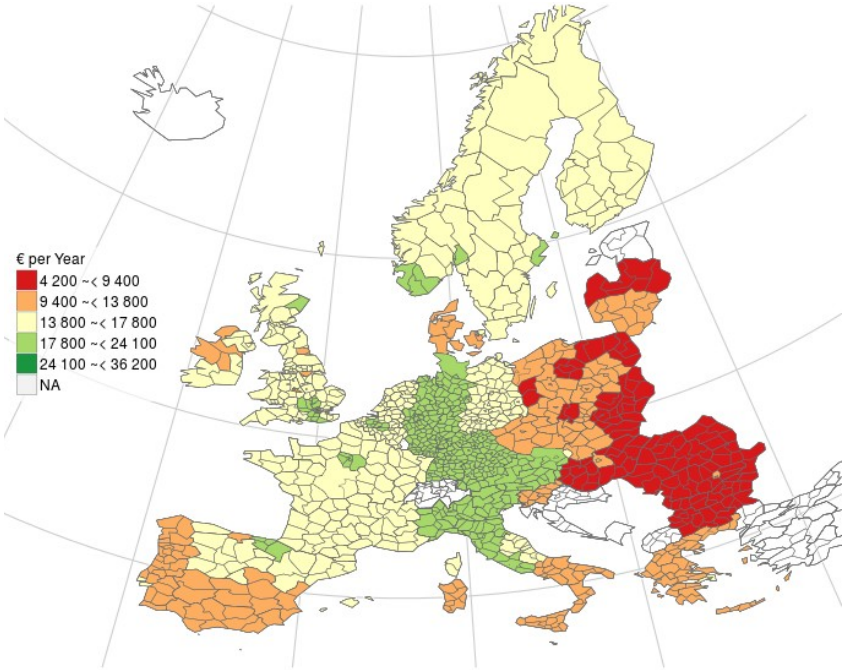
```
library("ggplot2")
ggplot(mapdata, aes(x = long, y = lat, group = group))+
  geom_polygon(aes(fill=cat), color="grey", size = .1)+
  scale_fill_brewer(palette = "RdYlBu") +
  labs(title="Fertility rate, by NUTS-3 regions, 2014",
       subtitle="Avg. number of live births per woman",
       fill="Total 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*

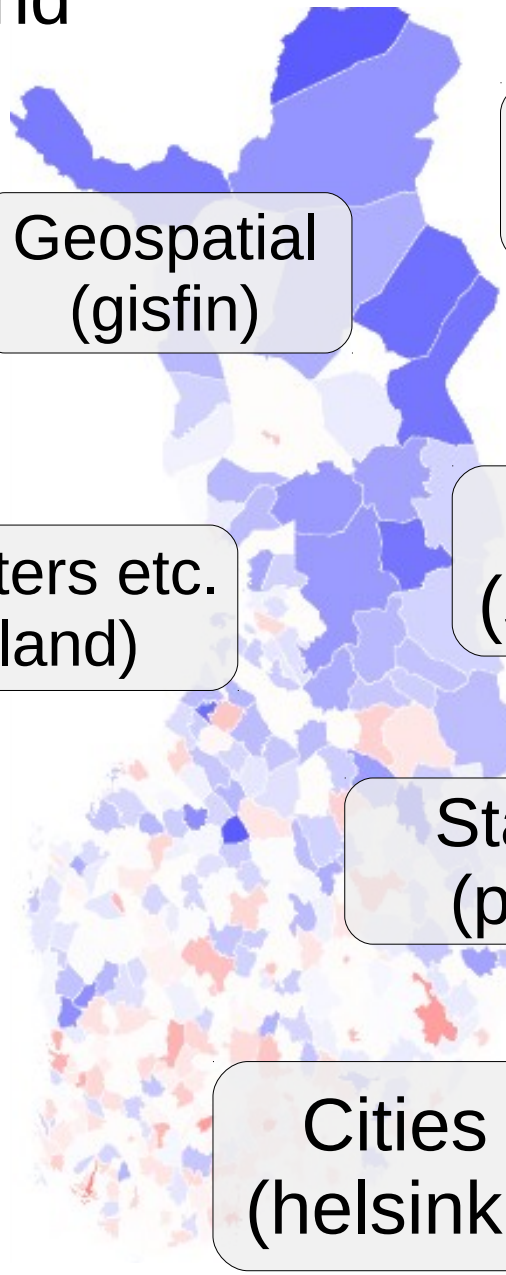
From specific packages to package ecosystems

eurostat &
eurostat_geospatial



pxweb for PX-Web/PC-Axis data from stats authorities in: Denmark, Finland, Greenland, Iceland, Latvia, Norway, Sweden, Various cities...

Algorithms for open data in Finland



Weather (fmi)

Geospatial (gisfin)

Health (sotkanet)

Registers etc. (finland)

Statistics (pxweb)

Cities (helsinki)

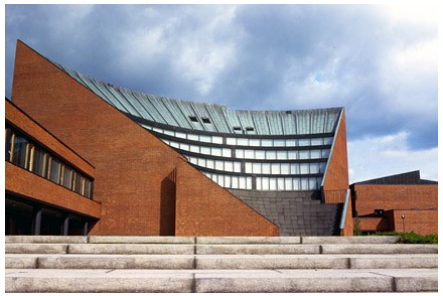
Unexpected outcomes from open collaboration

Aalto University Library

lib.aalto.fi

The future is made today.

Donate to Aalto University

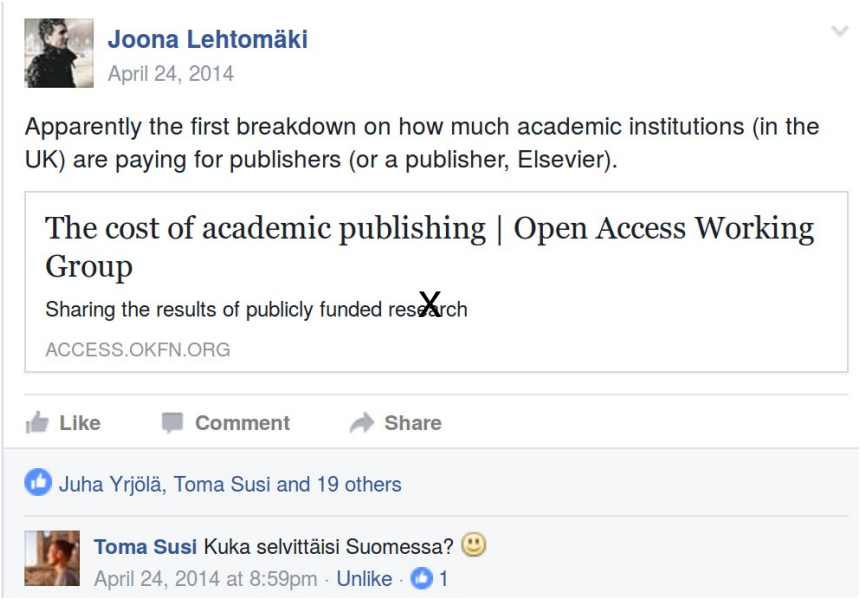


A?

ELSEVIER

Down-right Figure from: https://www.frenchentree.com/wp-content/uploads/2015/11/Waste-of-Money-Euro-exchange-Photo-by-Africa-Studio-via-Fotolia-550x367.jpg

Freedom of Information & Open science activism



OPEN KNOWLEDGE FINLAND

FOI requests to Finnish universities (2014)

Data collection ~ 70 institutions - 2010-2017

Open data released by MoE (2016-2018)

Two years from the initial requests to actual data release

Finland takes leading role in the openness of academic journal pricing

June 13, 2016

Journal subscription prices 2010-2017 for ~70 Finnish institutions as open data from MoE.

Rapid increase in costs (roughly 10% annually).

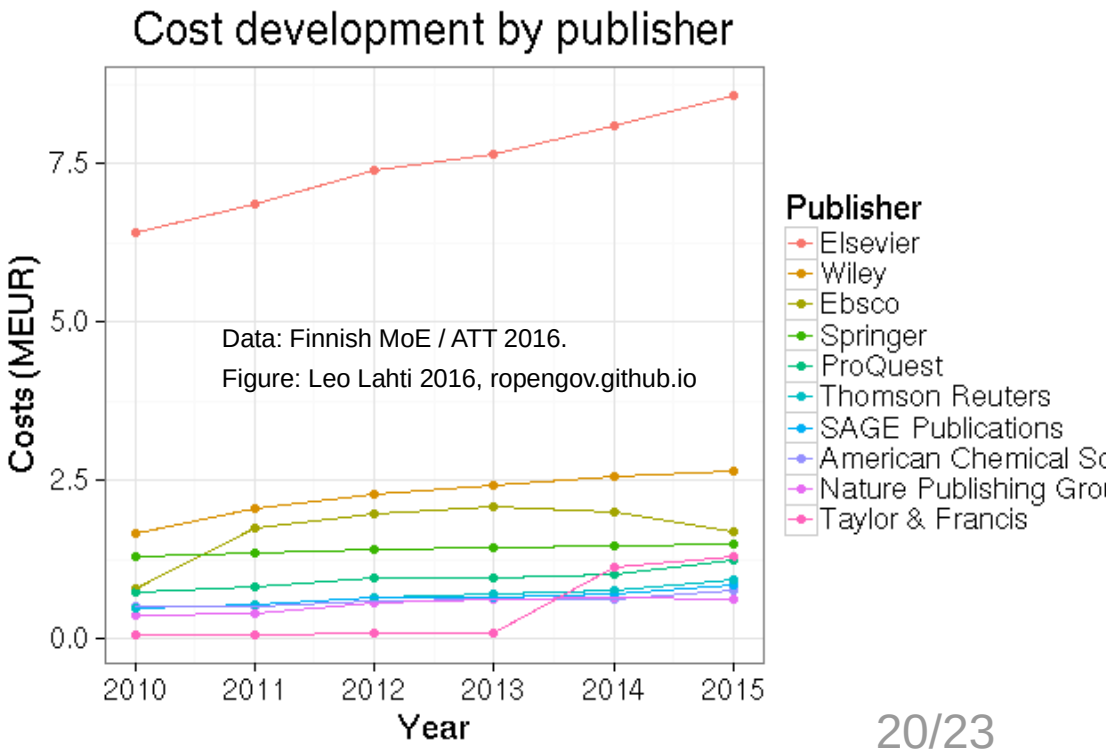
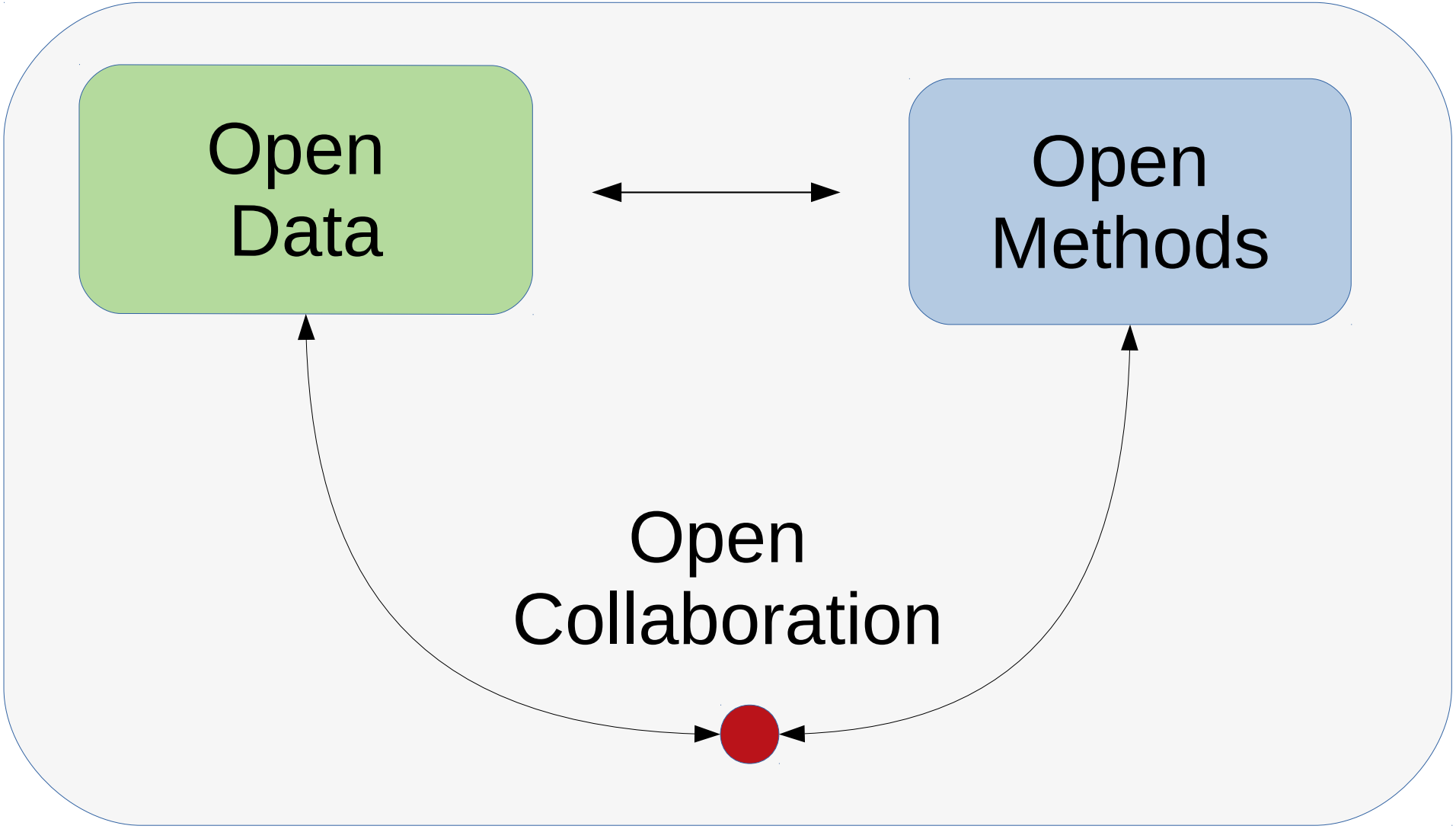


Figure & Data release: Ministry of Education and Culture of Finland / Open Science and Research Initiative 2014–2017

Elements of open data science



Challenges in Open Data Science projects

Community building

Focus & Research questions

Scattered resources

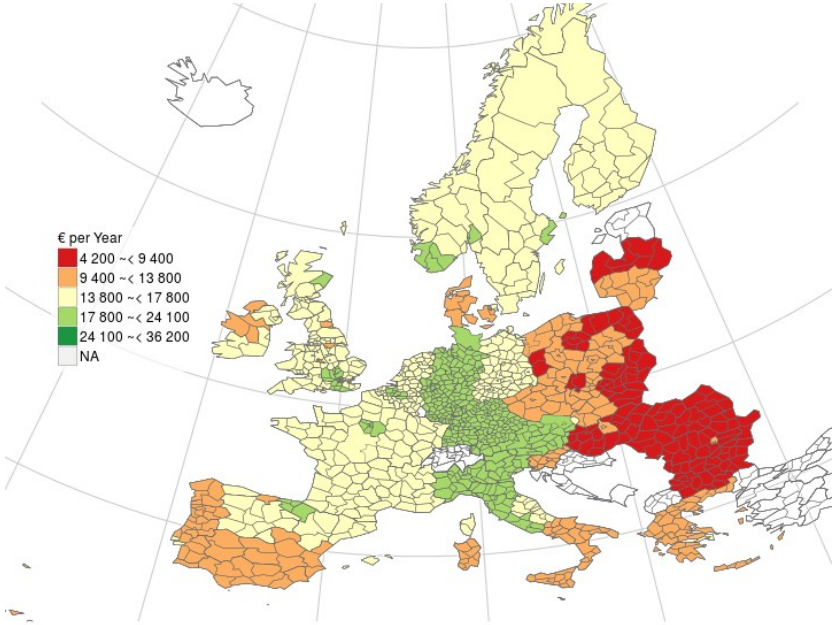
Sustainability
(data, methods, workflows)

Thank You!

Mikko Tolonen
Joona Lehtomäki
Markus Kainu
Juuso Parkkinen
Przemyslaw Biecek
Måns Magnusson
Sudarshan Shetty
Ville Laitinen
Aaro Salosensaari

@openreslabs
openresearchlabs.github.io

Helsinki Computational History Group



Suomen Akatemia
Finlands Akademi • Academy of Finland

Turun yliopisto
University of Turku

rOpenGov