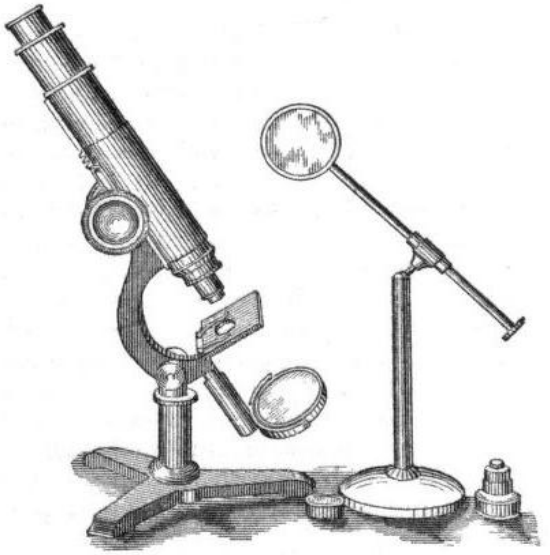
Open Data Science

Leo Lahti
University of Turku, Finland

@openreslabs



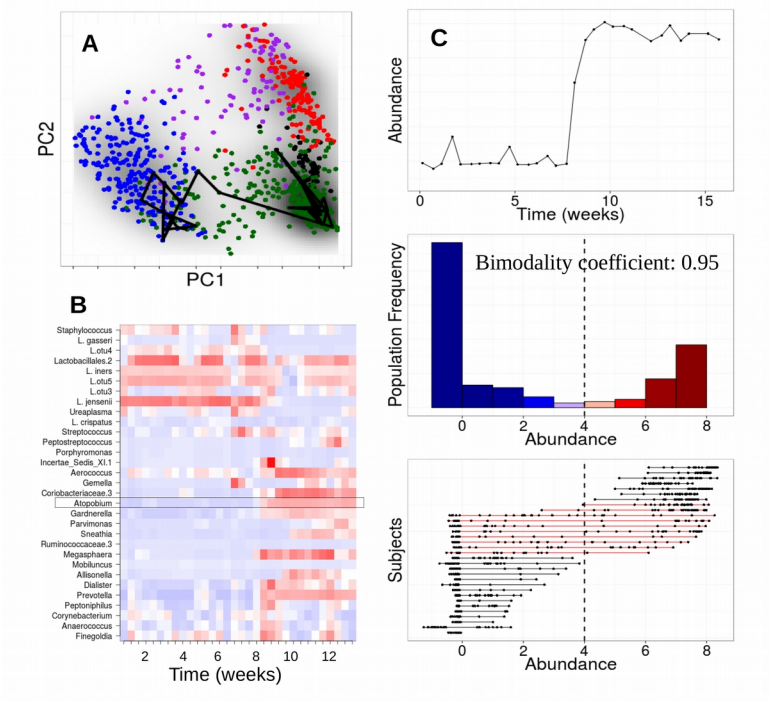
Turun yliopisto
University of Turku

Bioinformatics

Ecosystems, Genomics,
Microbiomics,
Large population studies

Metagenomics meets time series analysis: unraveling microbial community dynamics

Karoline Faust[1,2,3,9], Leo Lahti[4,5,9], Didier Gonze[6,7], Willem M de Vos[4,5,8] and Jeroen Raes[1,2,3]



A fully scalable online pre-processing algorithm for short oligonucleotide microarray atlases

Leo Lahti,[1,2,*] Aurora Torrente,[3,4] Laura L. Elo,[5,6] Alvis Brazma,[3] and Johan Rung[3]

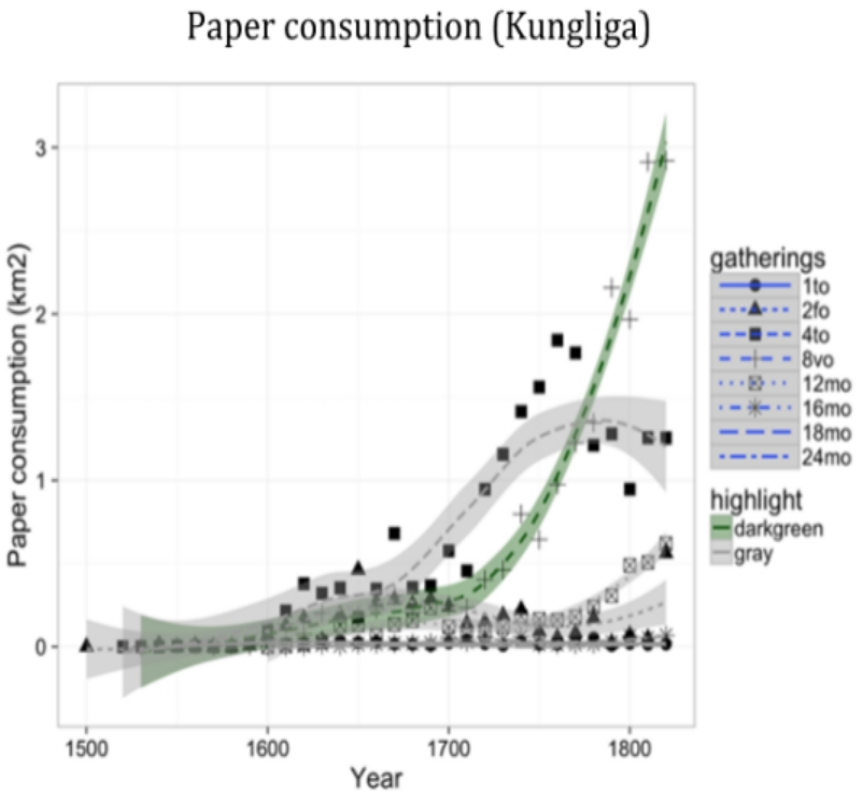
Digital Humanities

16 40 KANSALLIS KIRJASTO

AATEHISTORIA JA DIGITAALISTEN AINEISTOJEN MAHDOLLISUUDET

11.8.2015 MIKKO TOLONEN JA LEO LAHTI

CASE 1: The rise of the octavo-sized book



Transparent reporting and communication were part of academic culture since the early days

Source: Wikimedia Commons / Public domain

Alchemy & algorithms: perspectives on the philosophy and history of open science

Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenoja, Mikko Tolonen

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

Heidi Laine Leo Lahti Anne Lehto

ESSAY

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

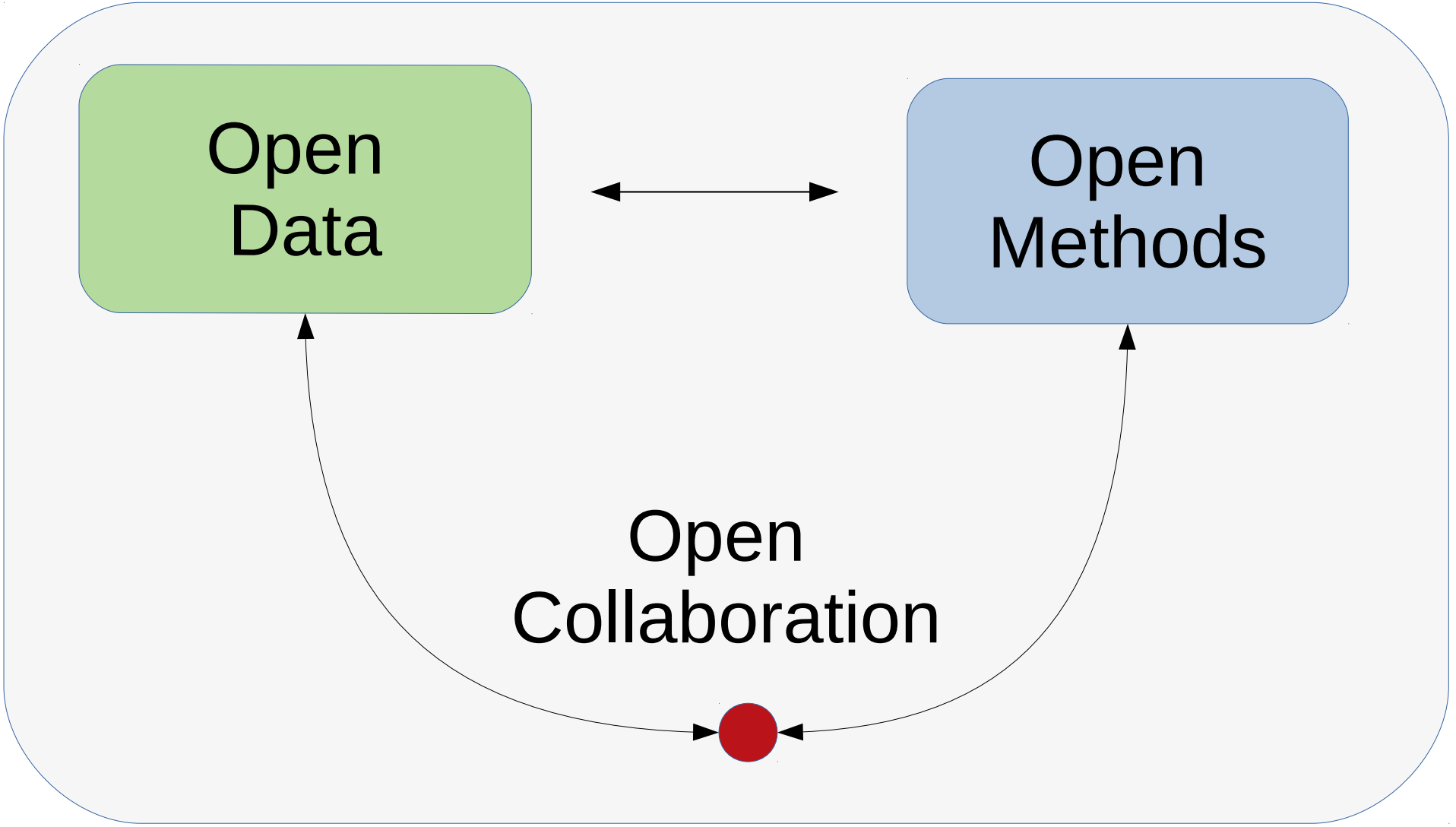
Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

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,*]

Elements of open data science



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.

Review

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

top

Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by tec experiments — microarrays and large amounts of genome-wide archives. Added-value database to make them accessible to eve the gene expression data that a making use of these data. Reus many obstacles in data prepara results. We will discuss these c believe can improve the utility of such data.

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

Automated multidimensional phenotypic profiling using large public microarray repositories

Min Xu[a,1], Wenyuan Li[a,1], Gareth M. James[b], Michael R. Mehan[a], and Xianghong Jasmine Zhou[a,2]

[a]Molecular and Computational Biology, Department of Biological Sciences, and [b]Marshall School of Business, University of Southern California, Los Angeles, CA 90089

Open Methods

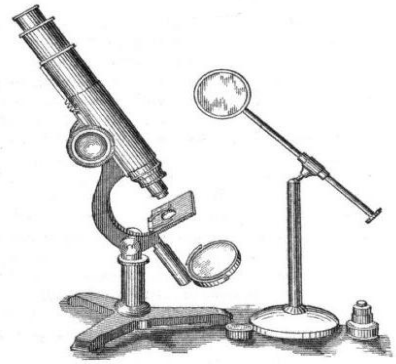
Computational workflows have an increasingly central role in research

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

POLICY FORUM

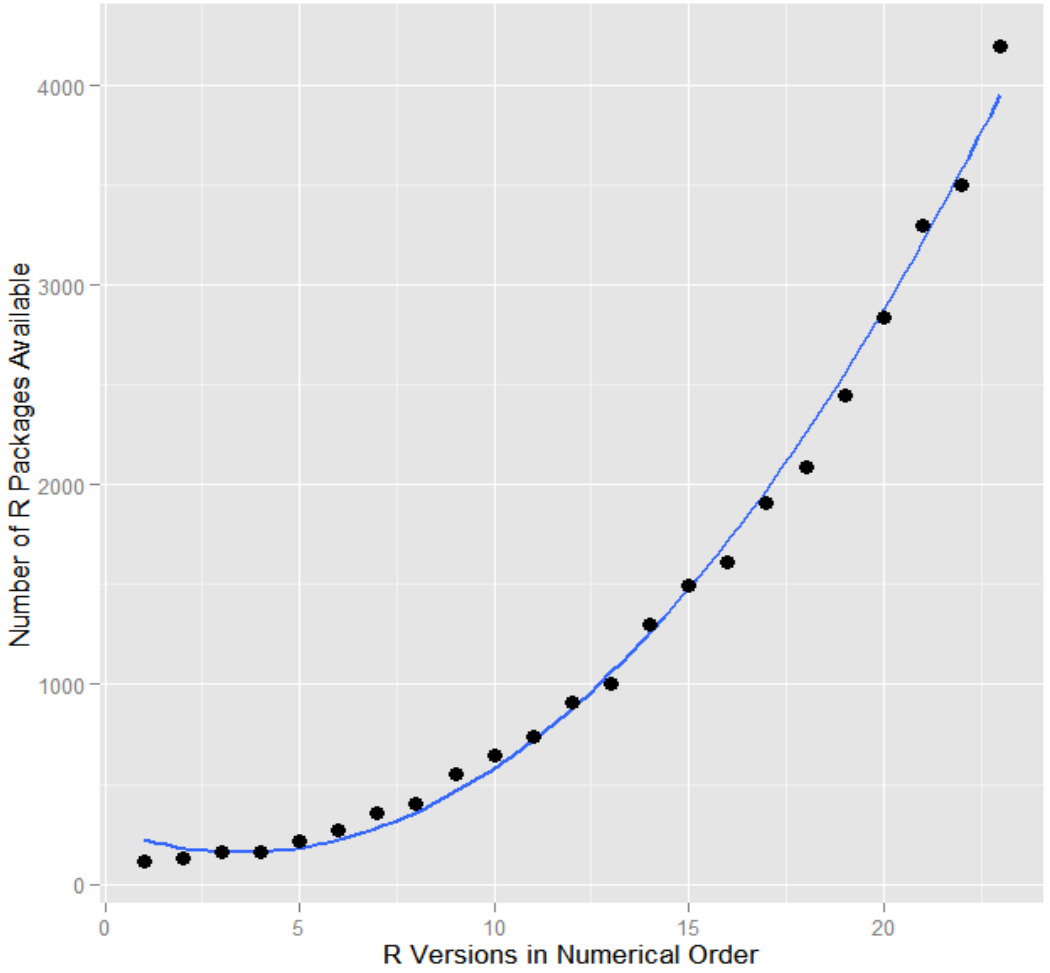
RESEARCH PRIORITIES

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,*]

Number of open analysis tools has grown exponentially



github SOCIAL CODING

R

rOpenGov

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

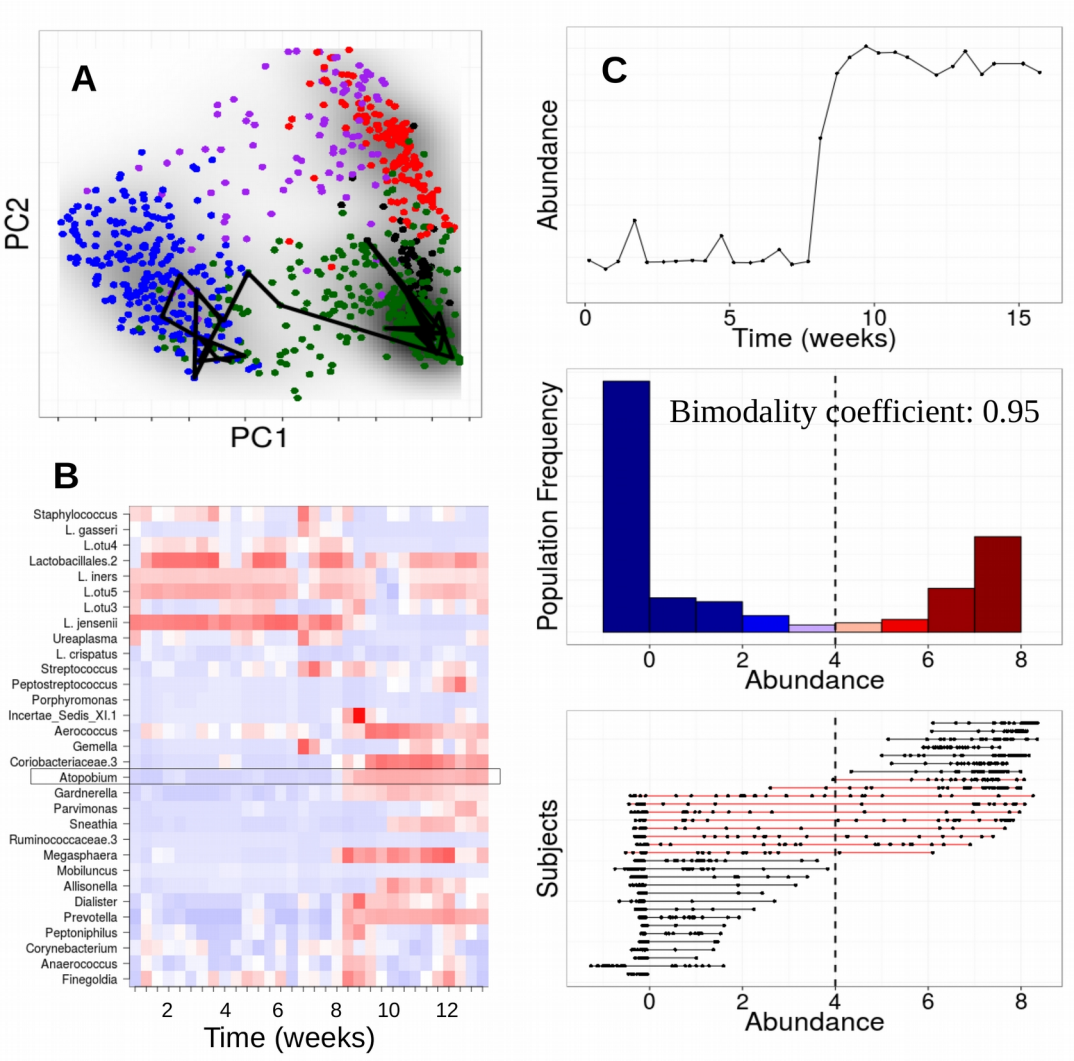
ROpenSci

Value of data can increase through sharing & use

Visualizing & modeling vaginal microbiome dynamics

Variation:
- *cross-sectional*
- spatial
- temporal

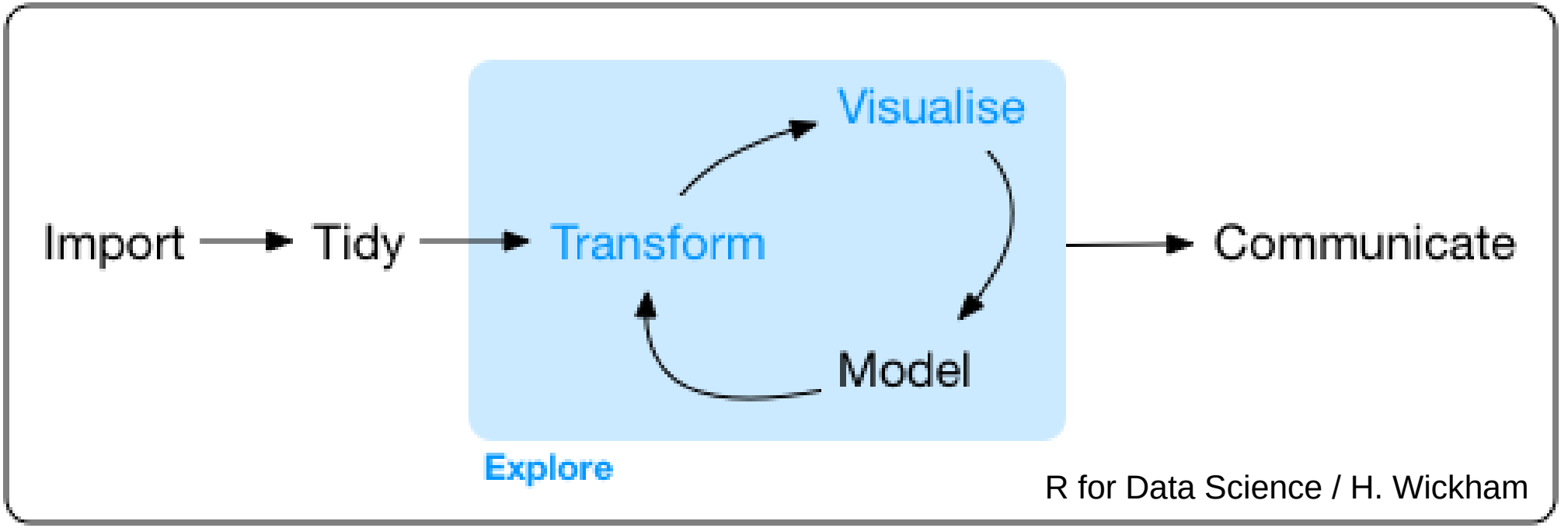
Levels of analysis:
- *ecosystem(s)*
- metagenome
- function (metabolome)
- host interactions



Metagenomics meets time series analysis: unraveling microbial community dynamics
Karoline Faust[1,2,3,9], Leo Lahti[4,5,9], Didier Gonze[6,7], Willem M de Vos[4,5,8] and Jeroen Raes[1,2,3]

Data: Gajer *et al.2012*

Data Science Workflow

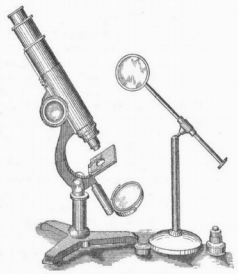


Program

Data
- raw
- supporting

Open data science ecosystem

IP[y]:
IPython



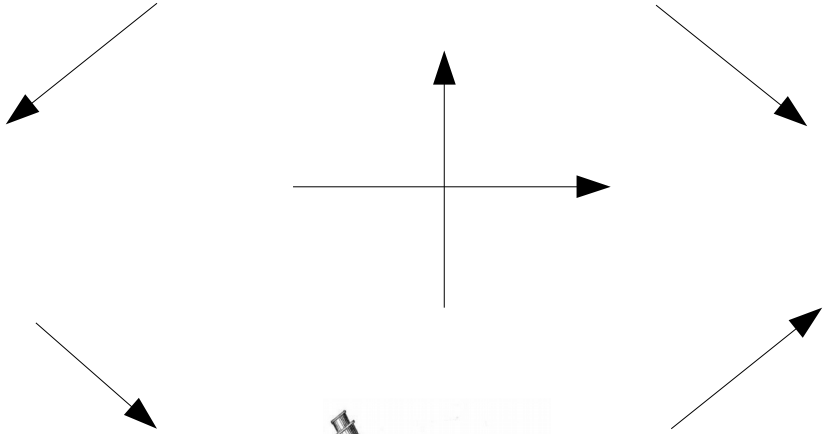
Outputs
- understanding
- reporting
- reuse

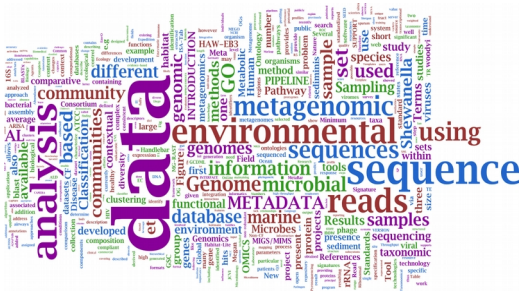
Hypothesis testing vs. hypothesis discovery?

?

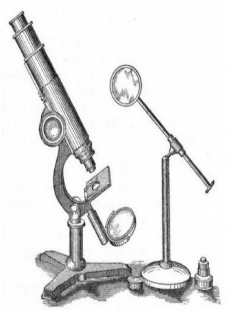
Hypothesis

Method





Data



Tools

Human Emit Over 10 Million Biological Particles per Hour - Personal Cloud

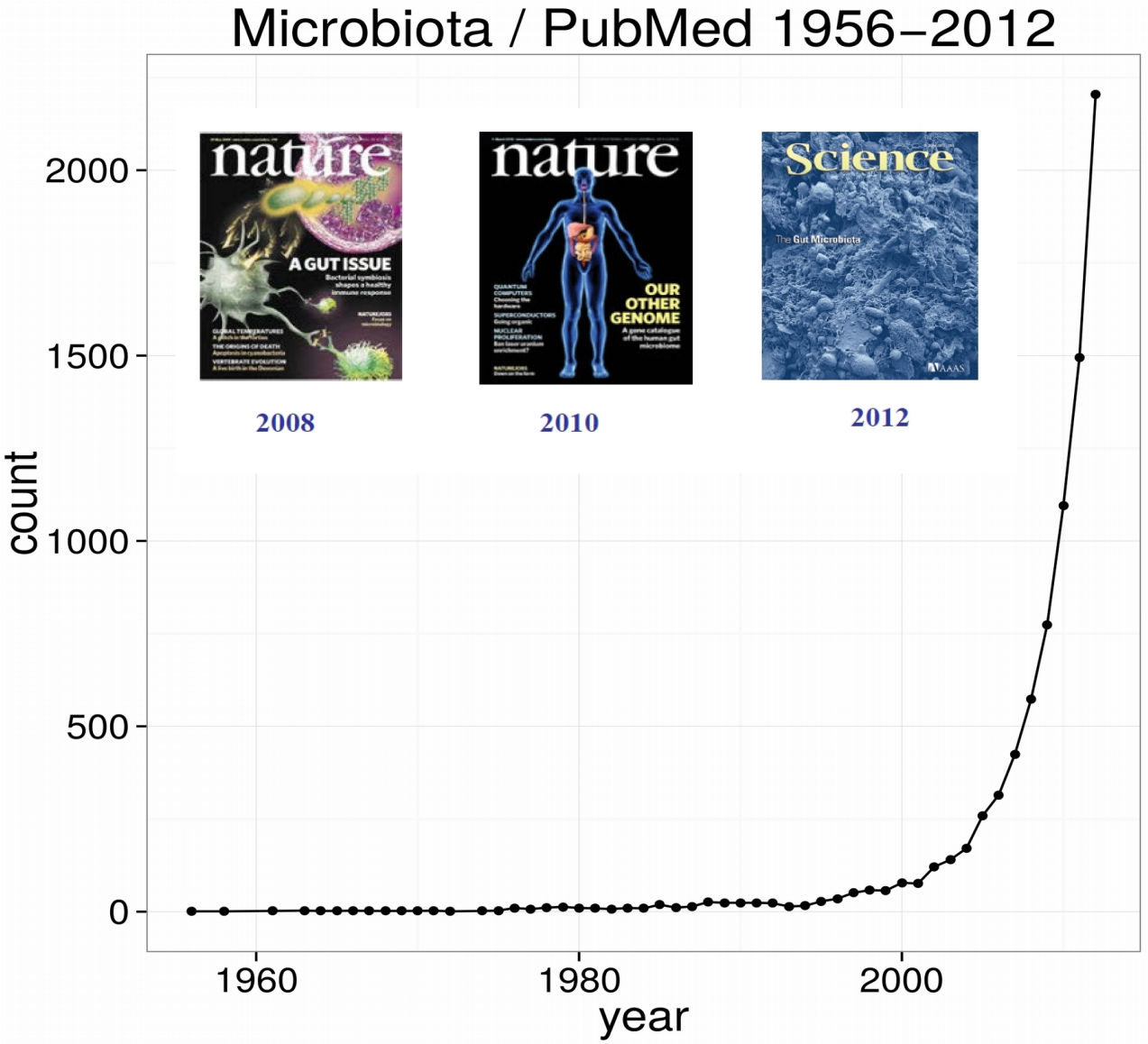
From Cradle to Grave

Meadows et al Peer J 2015 - Metcalf et al Science 2015

Research on human microbiota increasing at an accelerating pace

HUMAN MICROBIOME PROJECT

MetaHIT



Human microbiome: 1-2kg & 10 trillion cells & 10M genes

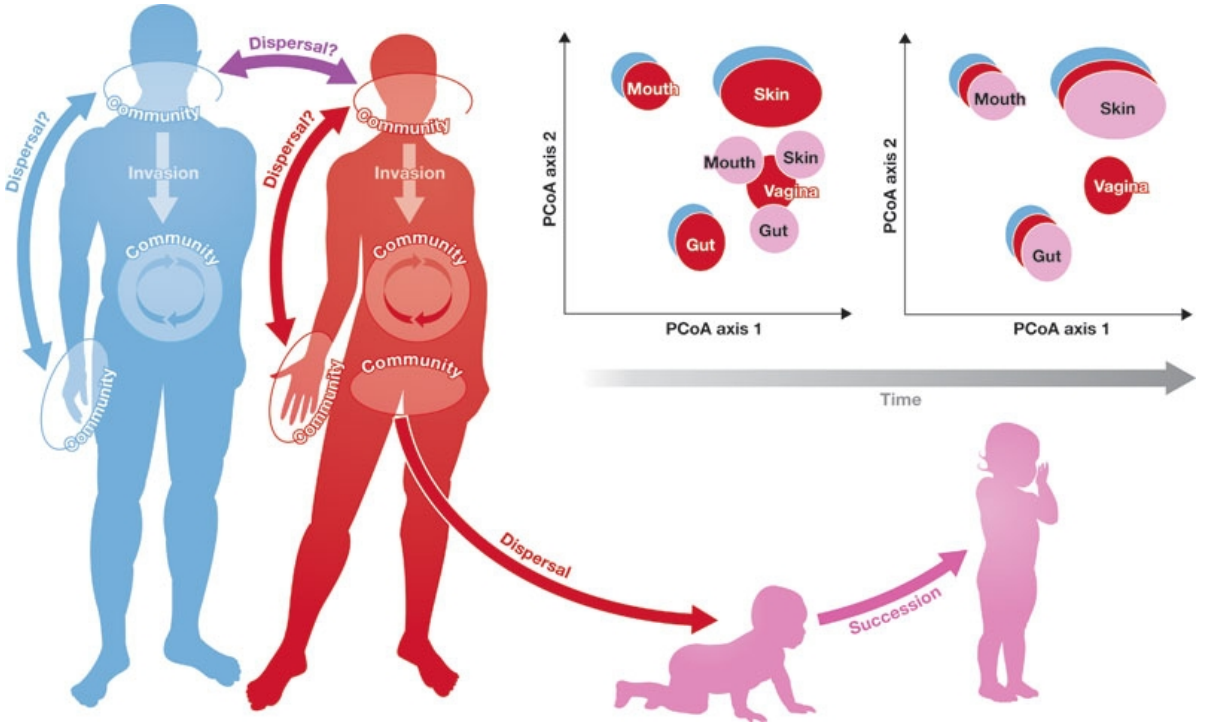
Plastic & easier to manipulate than our genome

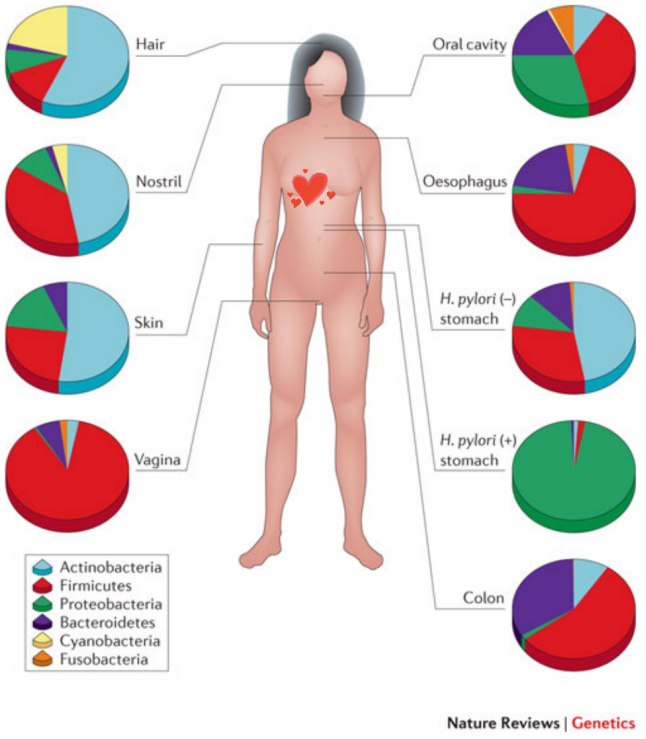
Unique as fingerprint

1.3x more microbial cells

500x more genes (10M~ 20k)

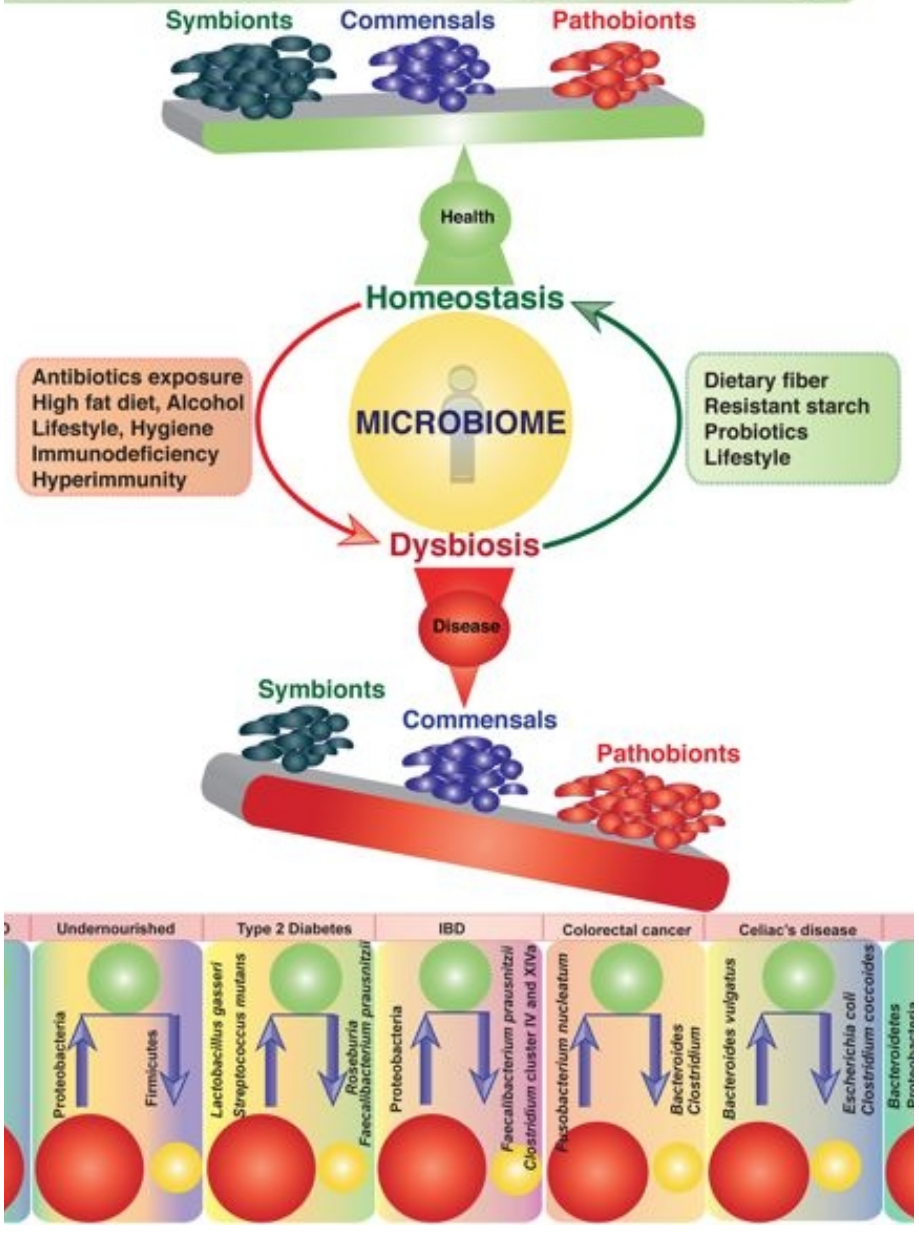
5:1 more viruses than bacteria; in addition eukaryotes, parasites..





Cho & Blaser Nature Rev Genetics 2012
Douillard & De Vos Micr Cell Fact 2014

We need a right balance of bugs



95% of our microbes are in gut

100 billion / gram

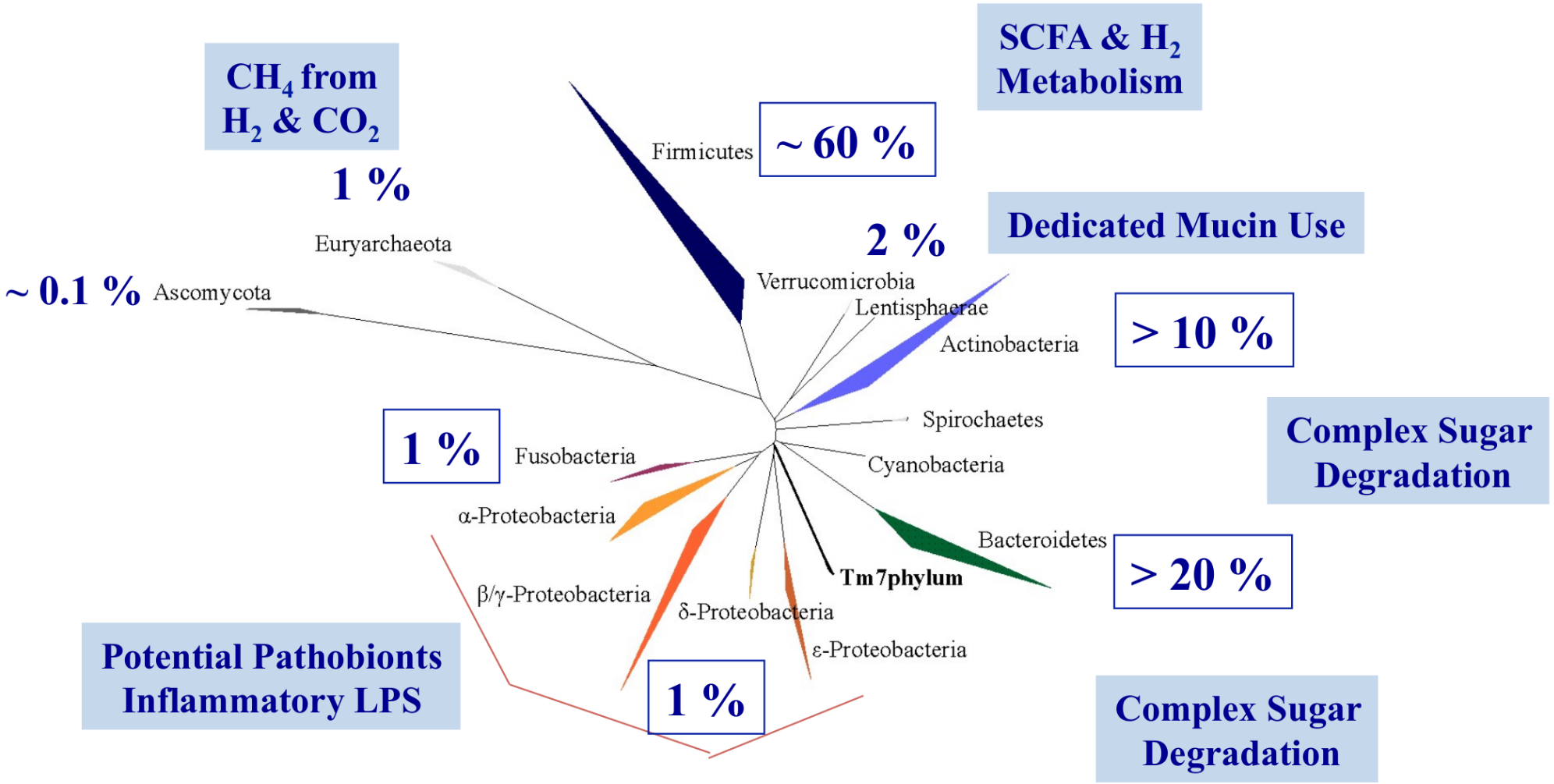
Symbionts: mutual benefit

Commensals: passengers

Pathobionts: harmful

Gut microbiome: 300m$^{2.}$ Brussel central square flower carpet

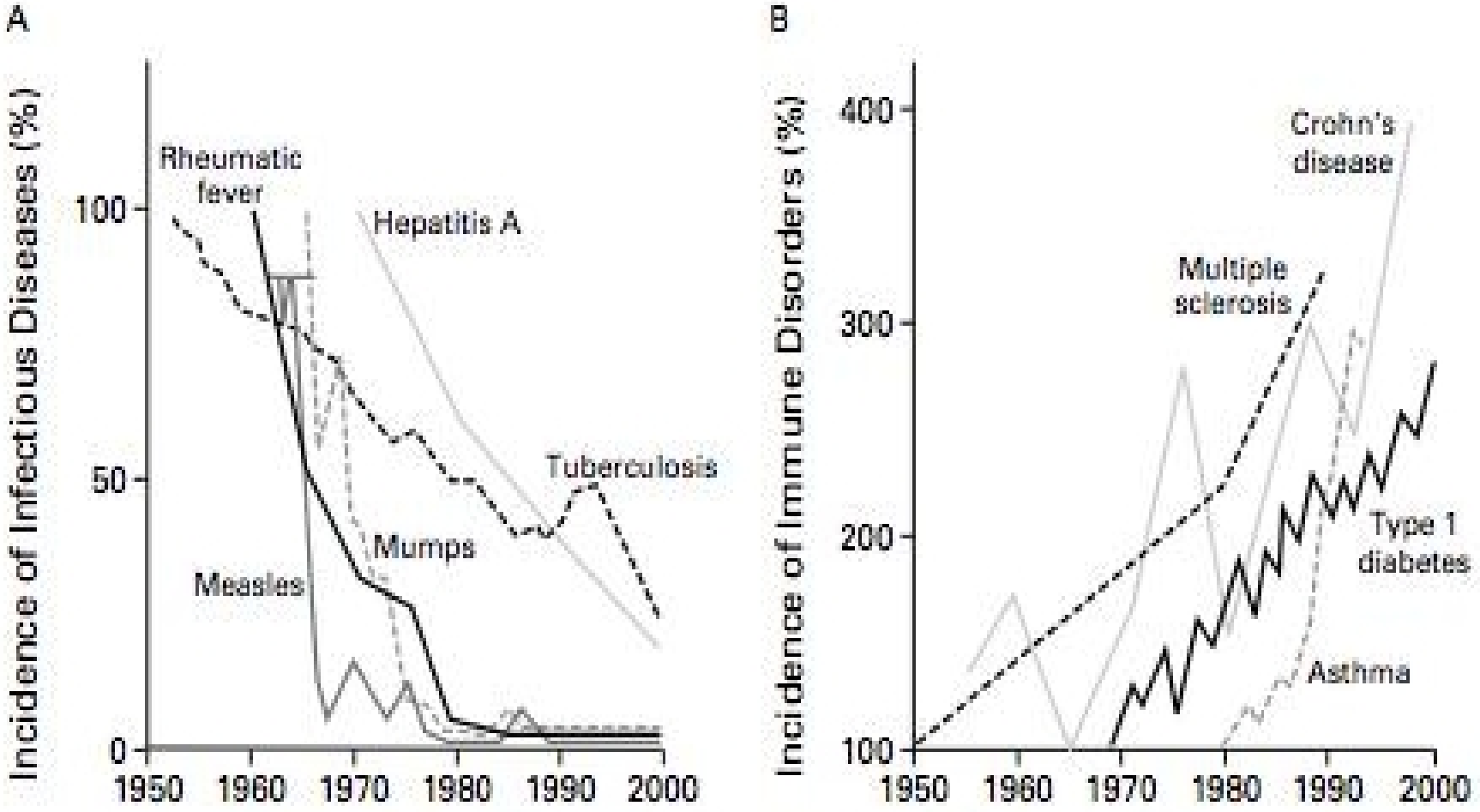
Colonic Climax Community in ~ 100 EU Adults



Zoetendal EG, EE Vaughan & WM de Vos (2006) Mol Microbiol 59: 1639

Lay C, L Rigottier-Gois, K Holmstrom, M Rajilic, EE Vaughan, WM de Vos, MD Collins, R Their, P Namsolleck, M Blaut & J Dore (2005) AEM 71: 4153

Many diseases linked to human microbiome have become more prevalent in the past decades !



Bach NEJM 2002

18 / 81

How do we measure gut microbiome?

Bristol Stool Chart

Type 1		Separate hard lumps, like nuts (hard to pass)
Type 2		Sausage-shaped but lumpy
Type 3		Like a sausage but with cracks on the surface
Type 4		Like a sausage or snake, smooth and soft
Type 5		Soft blobs with clear-cut edges
Type 6		Fluffy pieces with ragged edges, a mushy stool
Type 7		Watery, no solid pieces. **Entirely Liquid**

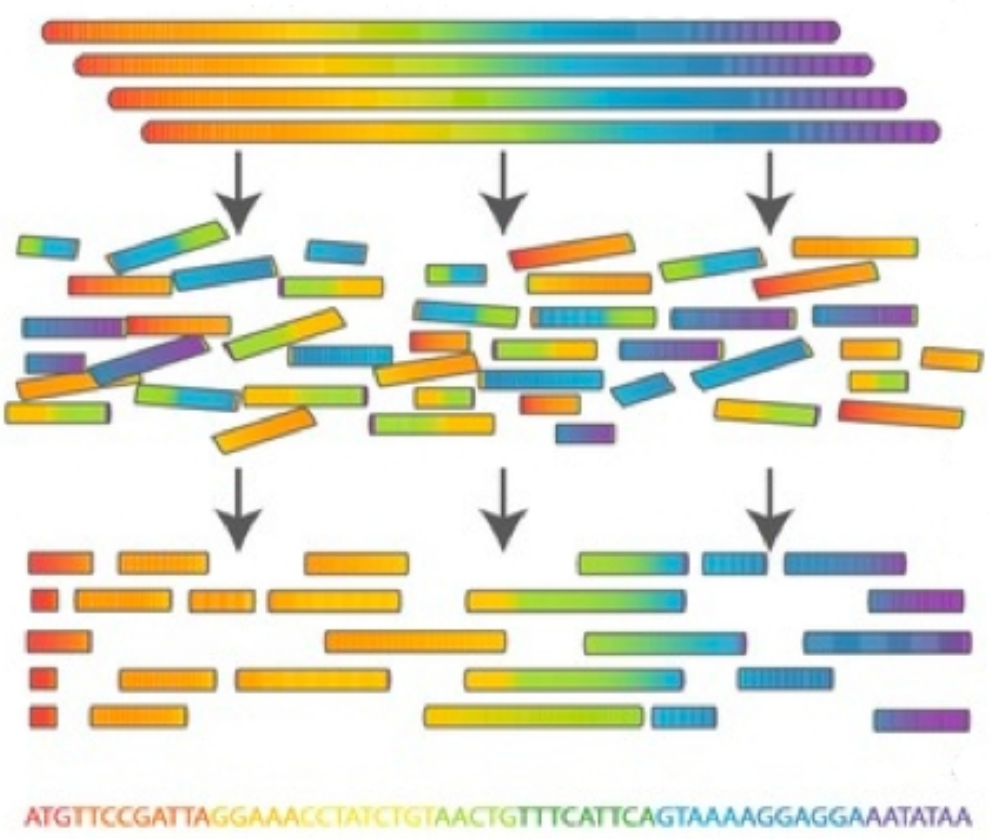
How do we measure microbiome?

Culture-based

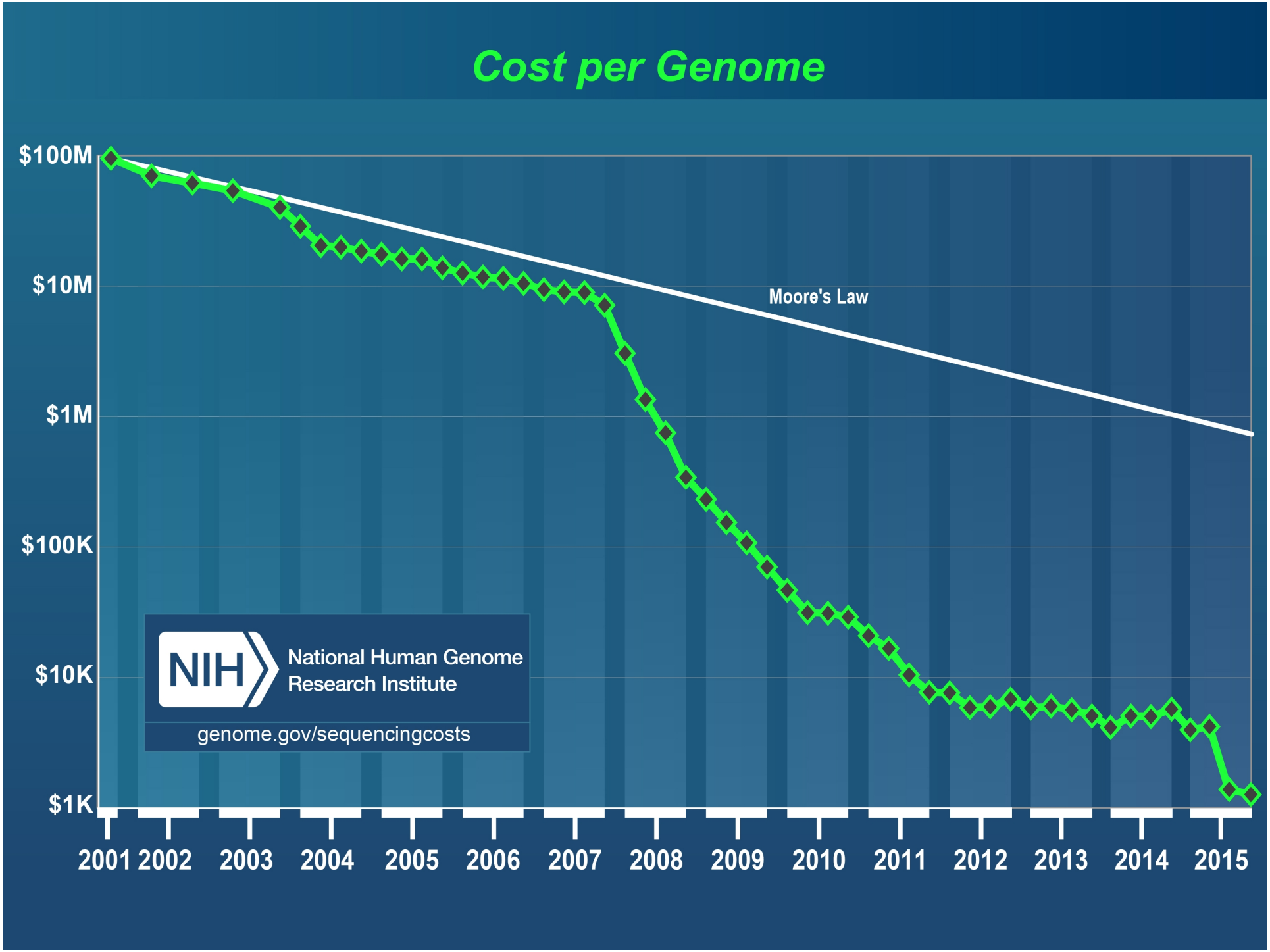
(for *culturable* bugs)

https://www.sott.net/article/309408-A-childs-bacteria-filled-handprint-reveals-the-wonder-of-the-human-microbiome

Sequencing-based
(for *all* bugs; most!)

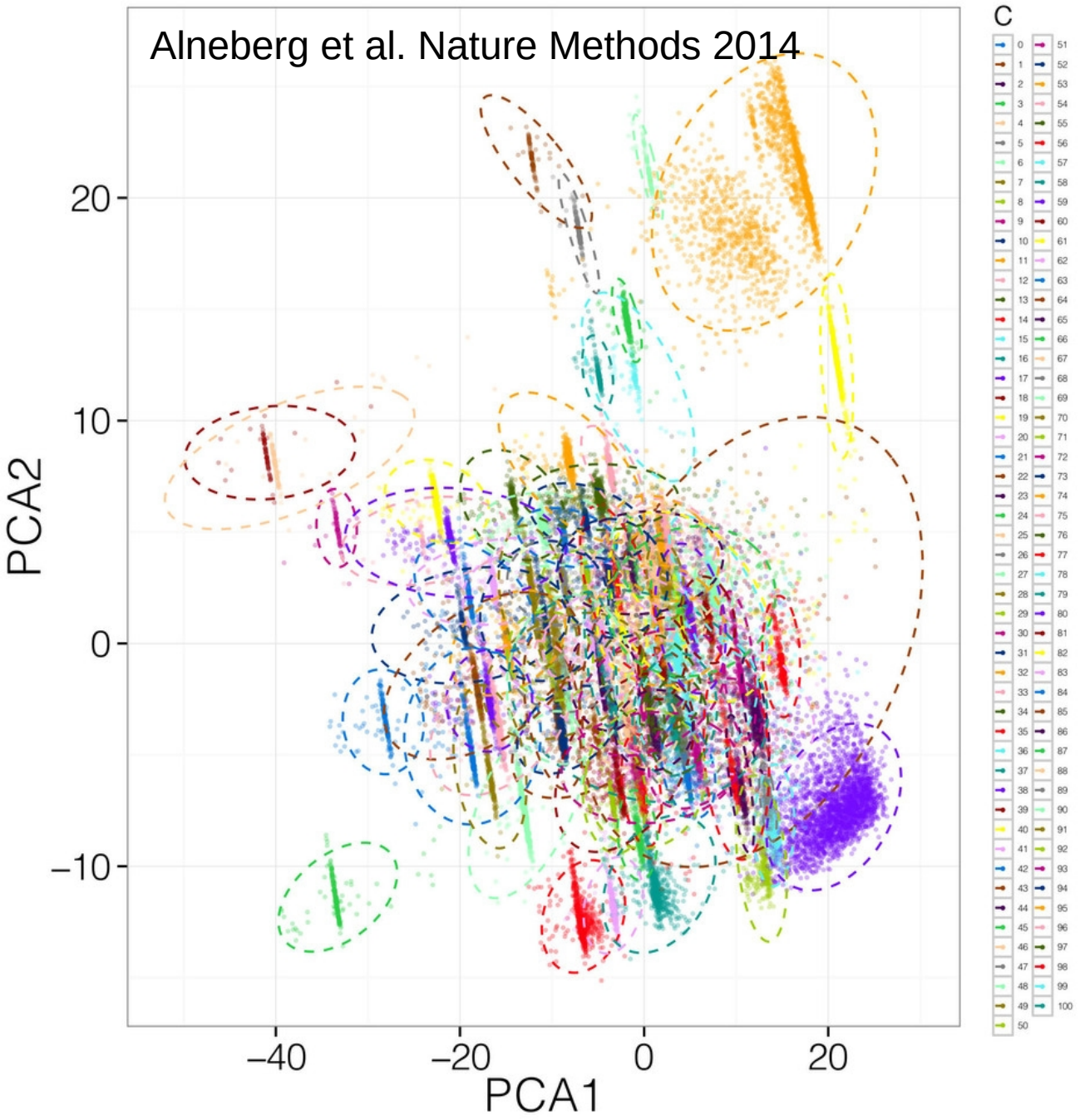


Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology and Evolution 29(1): 51-63


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Clustering contigs by coverage and composition (CONCOCT)

Variational Dirichlet Process multivariate mixture models work very well in practical applications with high-dimensional data



Front. Cell. Infect. Microbiol., 09 August 2012 | https://doi.org/10.3389/fcimb.2012.00104

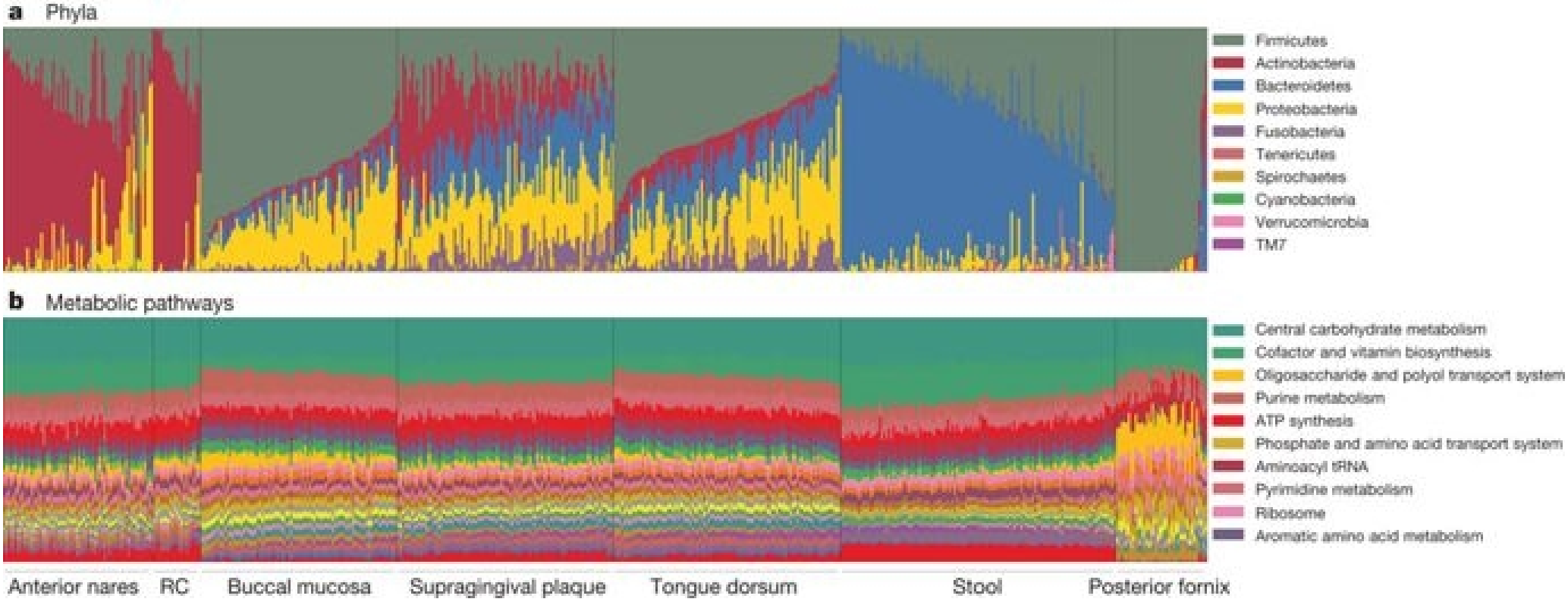
The function of our microbiota: who is out there and what do they do?

Noora Ottman[1], **Hauke Smidt**[1], **Willem M. de Vos**[1,2] **and Clara Belzer**[1]*

[1]Laboratory of Microbiology, Wageningen University, Wageningen, Netherlands

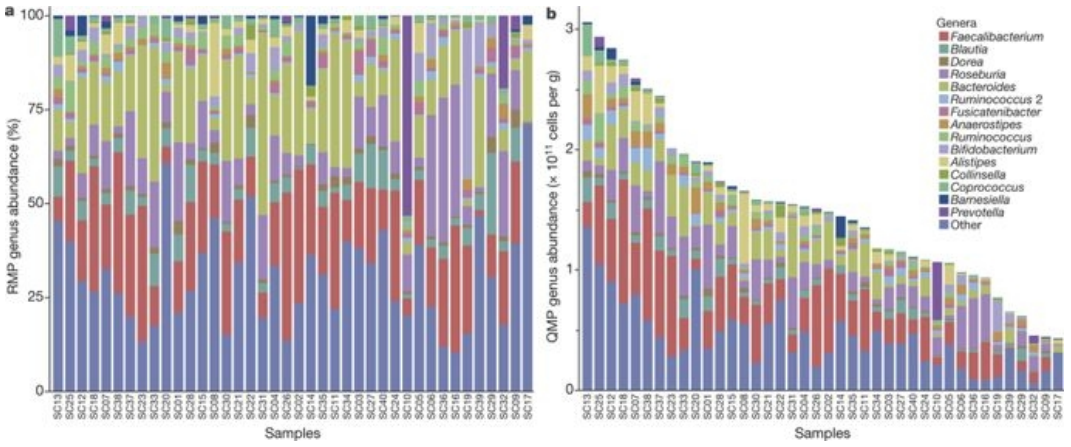
[2]Department of Basic Veterinary Medicine and Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland

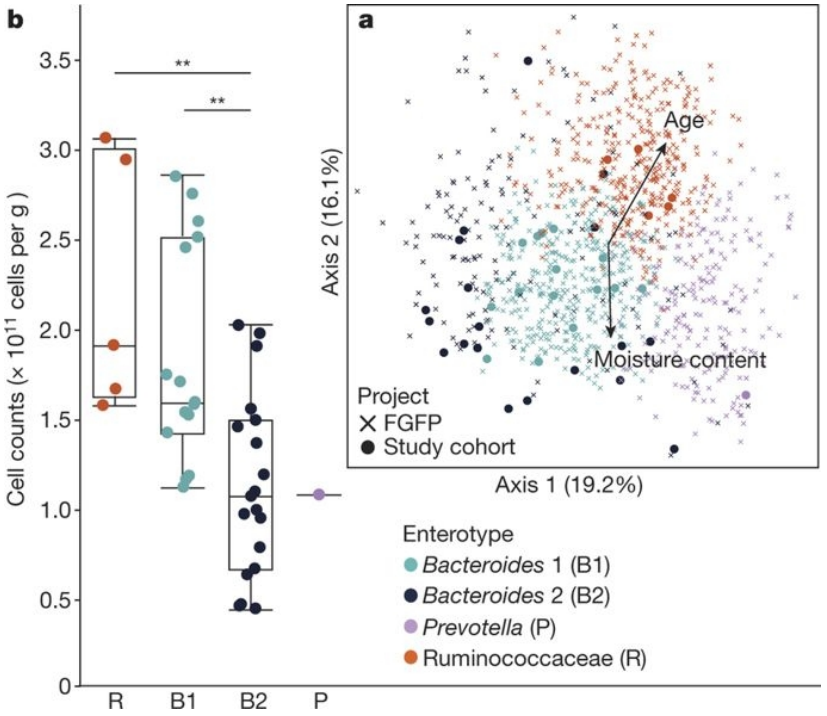
Bacteria vary a lot between individuals but they do similar things!



Human Microbiome Project Consortium. Nature 2012.

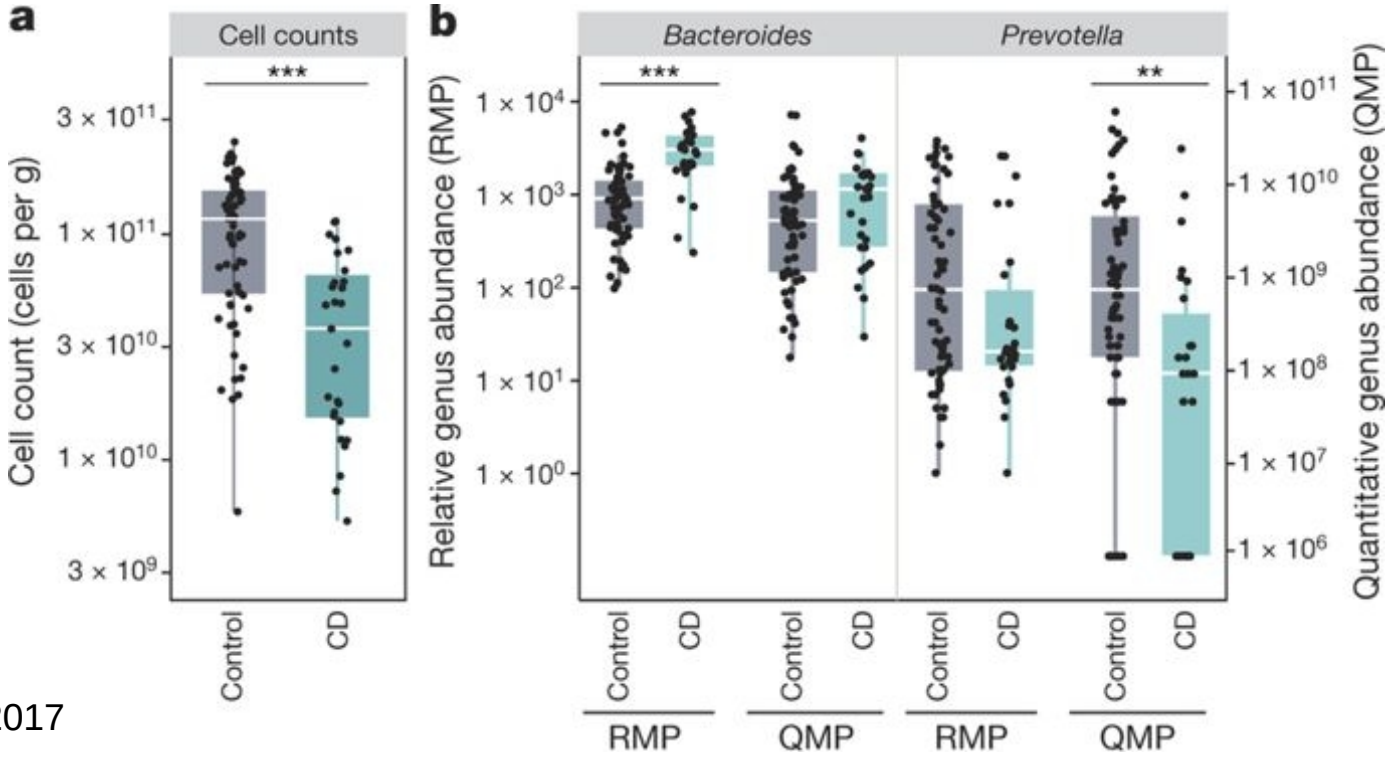
Relative versus absolute abundance: quantitative microbiome profiling





RMP vs. QMP:

drastic effect
on conclusions!



Vandeputte et al. Nature 551:507-511, 2017

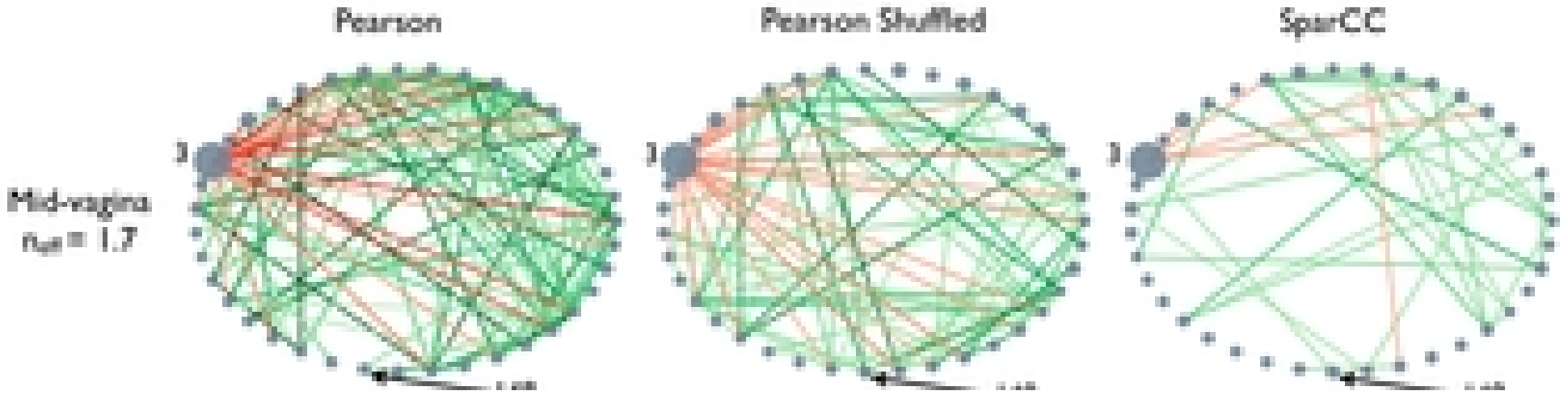
Aitchison transformations remove compositionality

Centered log-ratio transformation (CLR)

g(x): geometric mean

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_{D-1}}{g(x)}\right]$$

Drastic effects on co-occurrence network inference

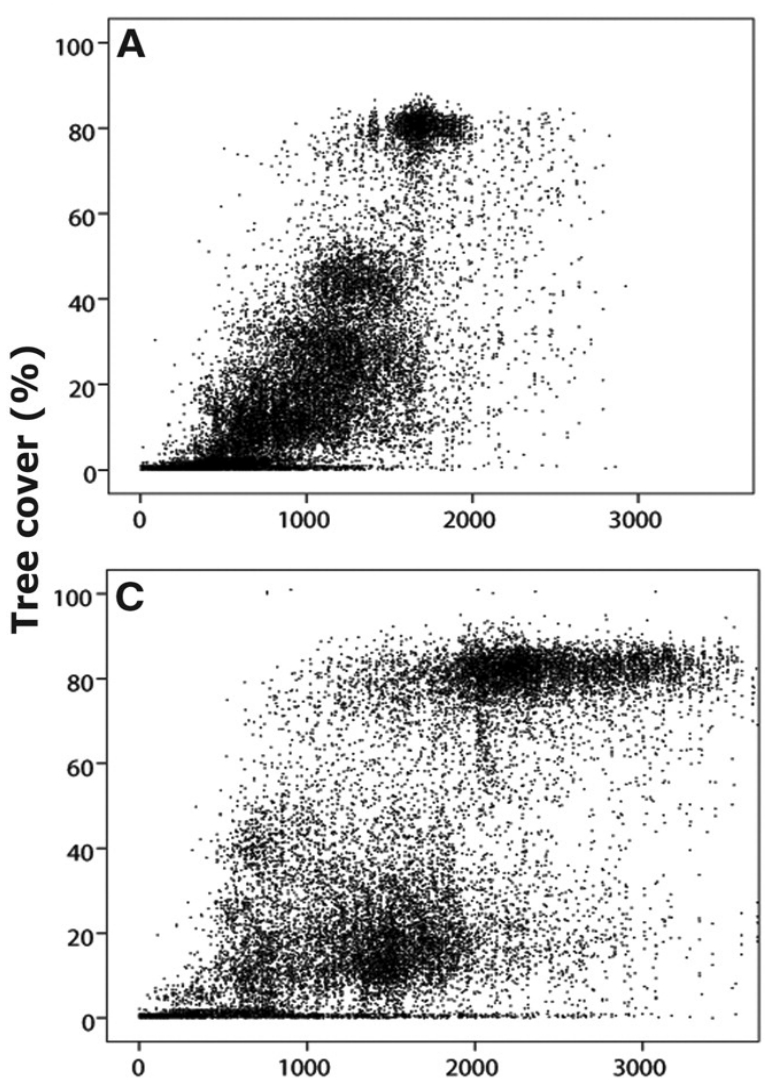


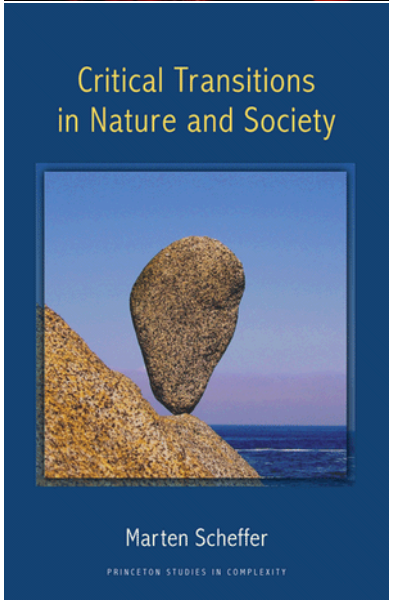
Science 14 October 2011:
Vol. 334 no. 6053 pp. 232-235
DOI: 10.1126/science.1210657

REPORT

Global Resilience of Tropical Forest and Savanna to Critical Transitions

Marina Hirota[1], Milena Holmgren[2,*], Egbert H. Van Nes[1], Marten Scheffer[1]





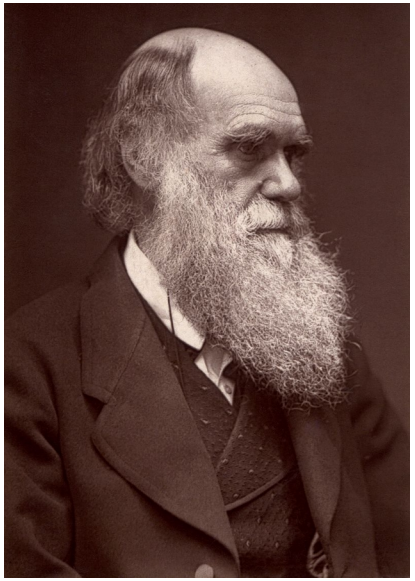
Neutral model of biodiversity: could random chance explain variation in human gut ecosystem?

Niche model

”When we look at the plants and bushes clothing an entangled bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this!” (**Darwin**, *The Origin of Species)*

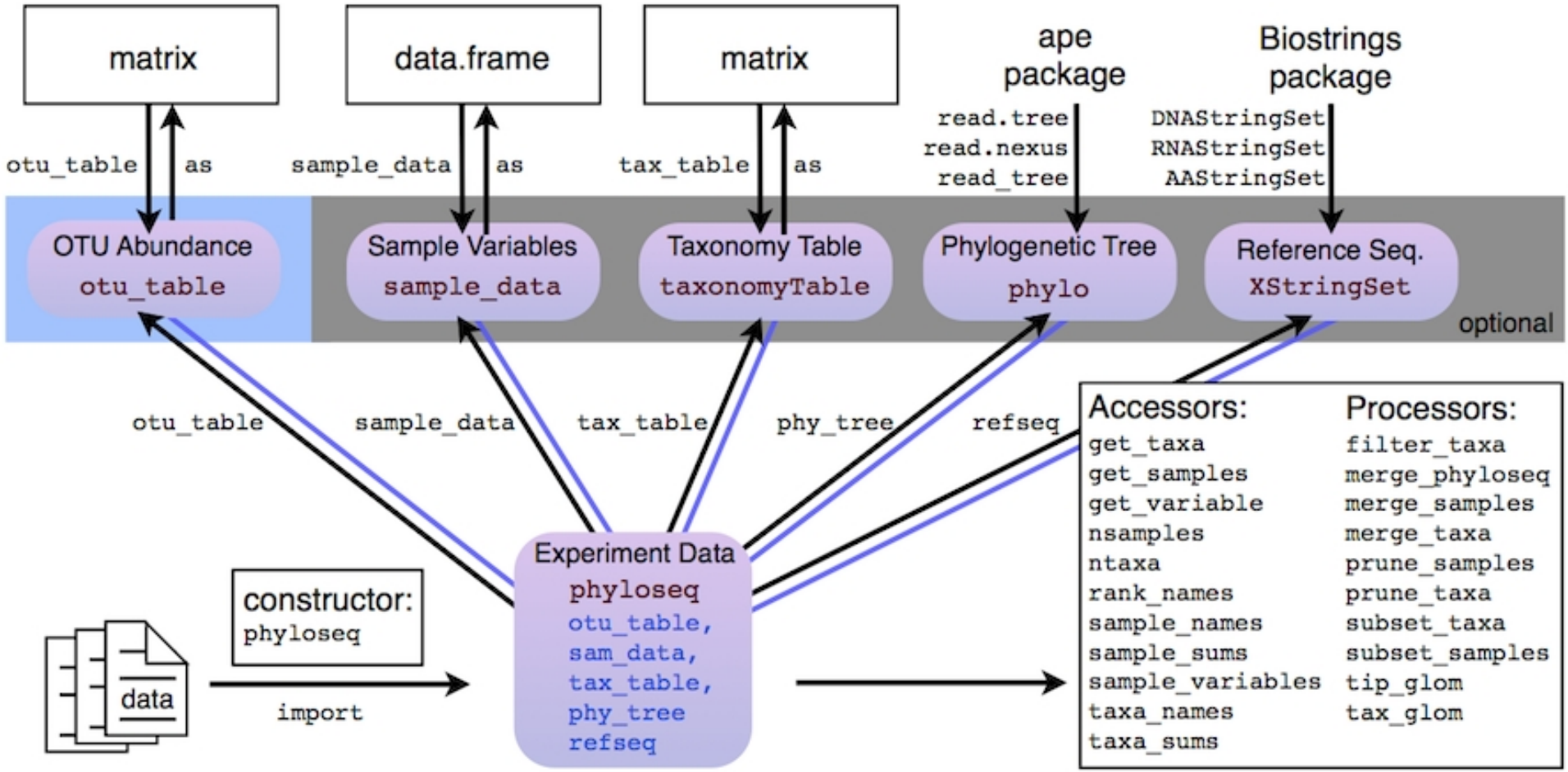
Neutral model

After >25 years on the Barro Colorado Island tropical forests, **Hubbell** proposed that.. random variation could in fact best explain observed biodiversity (Hubbell 2001).



Data standard: *phyloseq*

Standard for (16S) microbiome bioinformatics in R
(J McMurdie, S Holmes *et al.*)



microbiome R package

chat on gitter | build passing | codecov 24% | PRs welcome

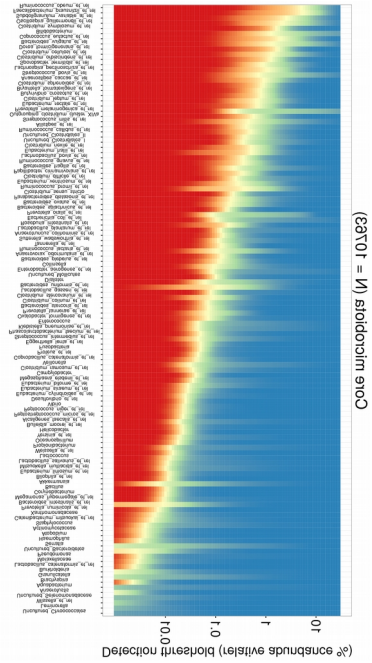
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Complementing *phyloseq*

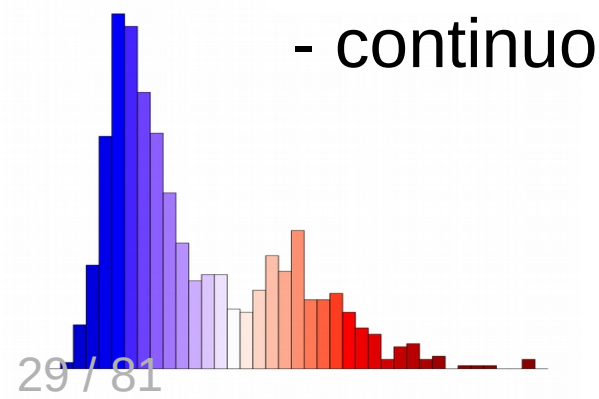
Data
- Population cohorts
- Time series
- Interventions
- Multi-omics

Utilities
- transformations
- alpha & beta diversity
- core microbiota
- visualizations

Analysis & modeling
- stability analysis
- tipping elements



Quality control:
- unit tests
- continuous integration..



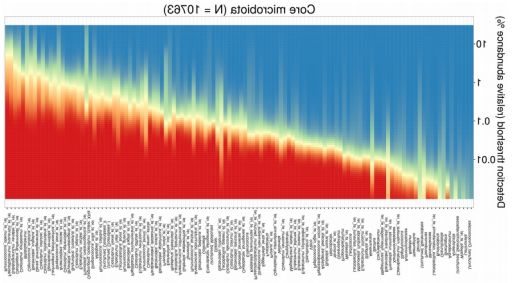
Community:
- mailing list
- gitter
- workshops

http://microbiome.github.io

Some utilities

Core & prevalence

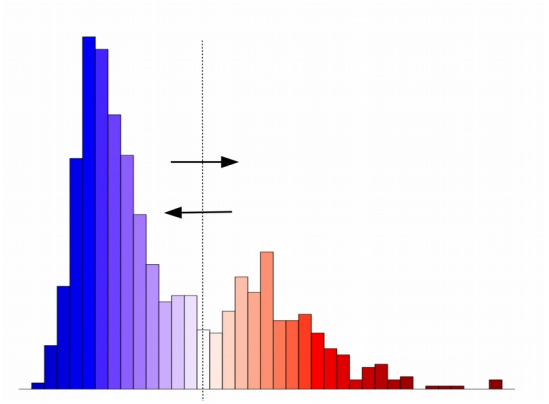
prevalence(x)
core(x)
core_members(x)

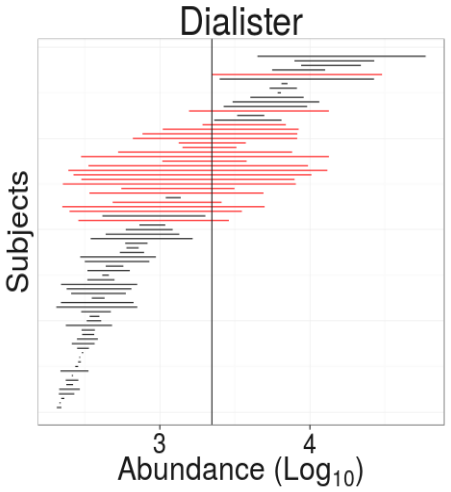


Transformations

transform(x, “compositional”)
transform(x, “clr”)
transform(x, “log10p”)
transform(x, “hellinger”)
transform(x, “identity”)

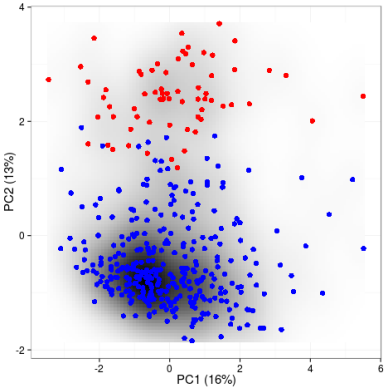
Stability & resilience





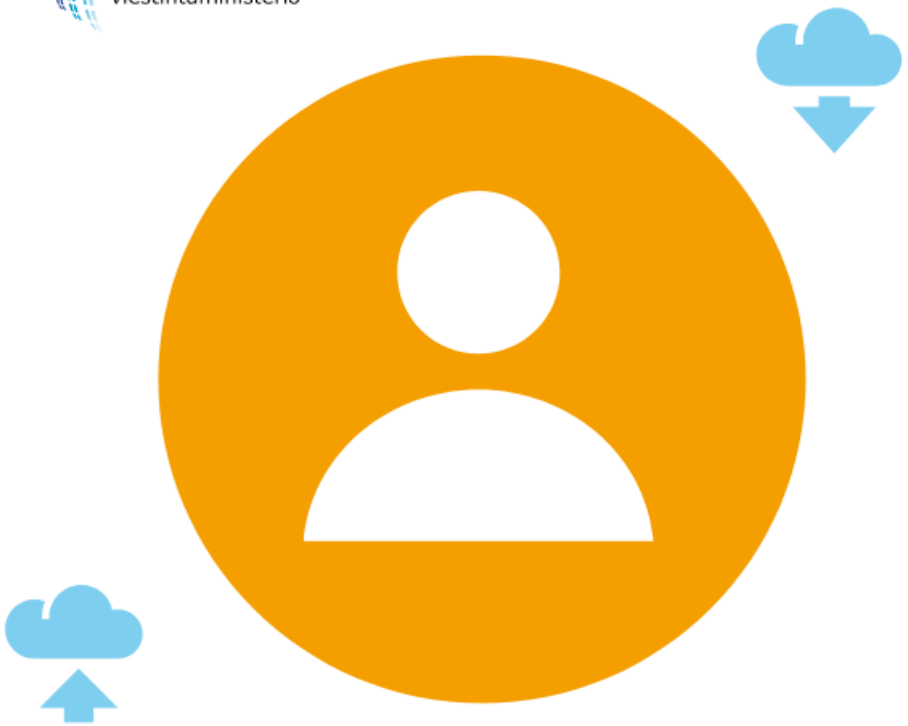
Alpha & beta diversity

alpha(x)
diversity(x)
evenness(x)
dominance(x)
rarity(x)
readcount(x)



Package website
- Tutorials & support
- Data & code sharing

Liikenne- ja viestintäministeriö



Antti Poikola
Kai Kuikkaniemi
Ossi Kuittinen

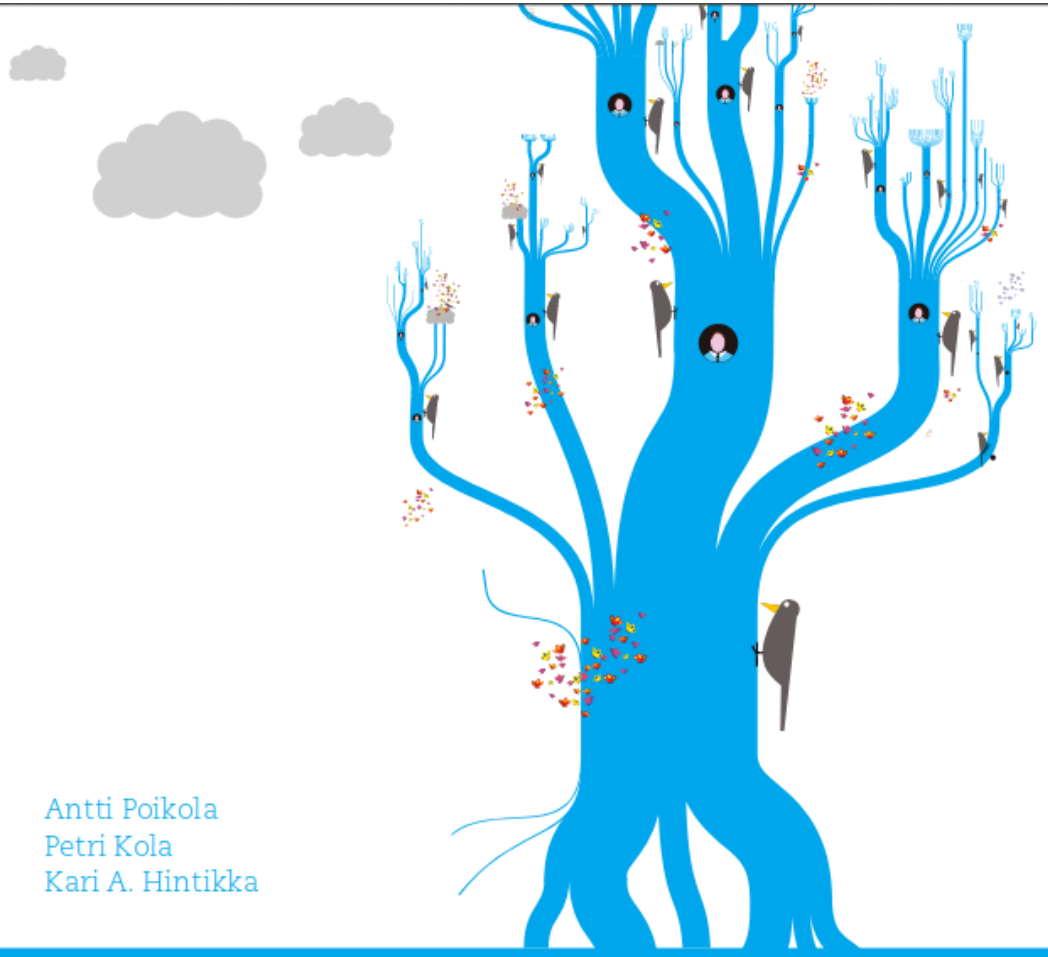
My Data

– johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen

Finland to lead the way in MyData

TIEDOTE 05.07.2018 14.44 fi sv en

Finland to lead the way in MyData (Foto: Shutterstock.com)



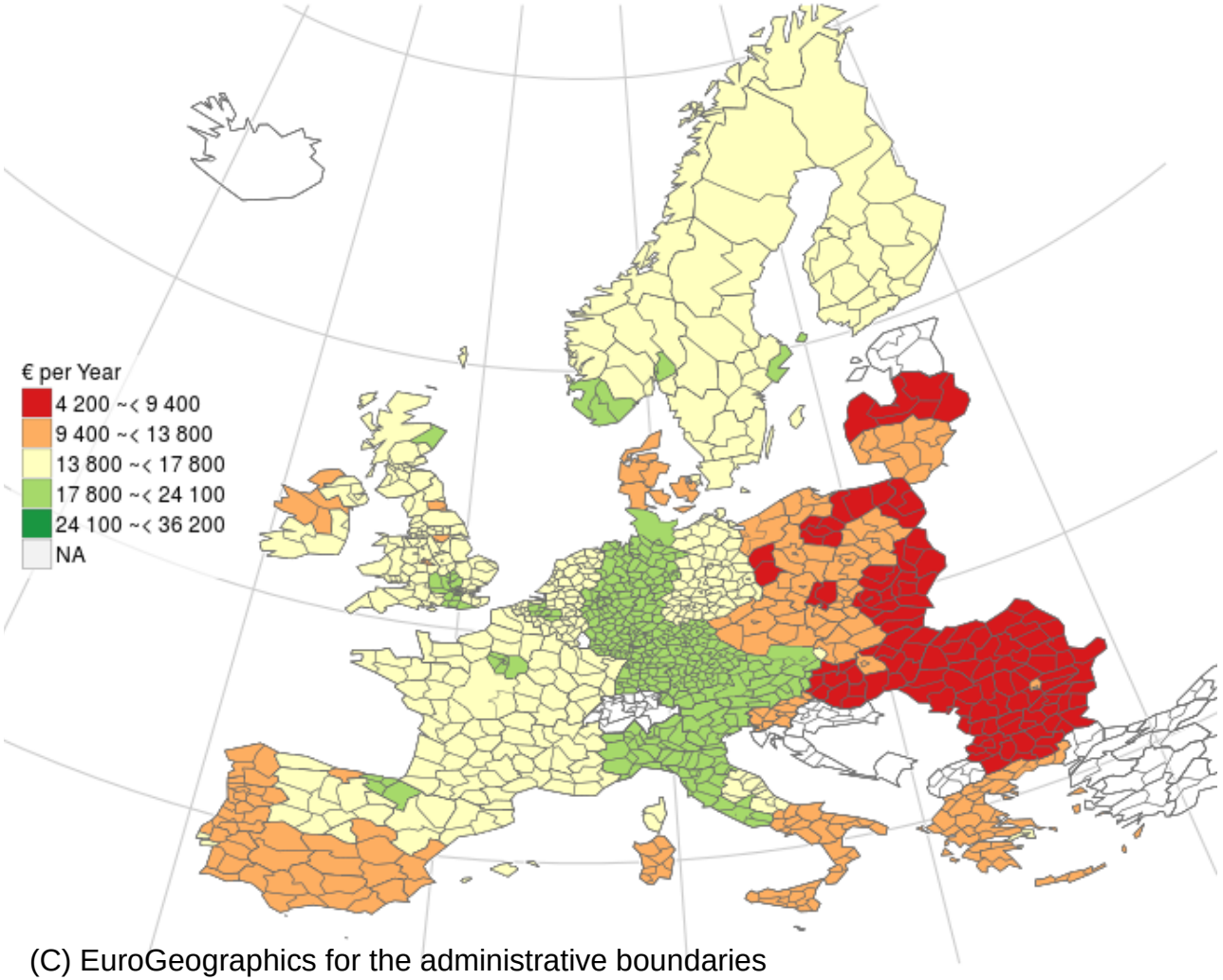
Antti Poikola
Petri Kola
Kari A. Hintikka

Julkinen data©

johdatus tietovarantojen avaamiseen

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..)

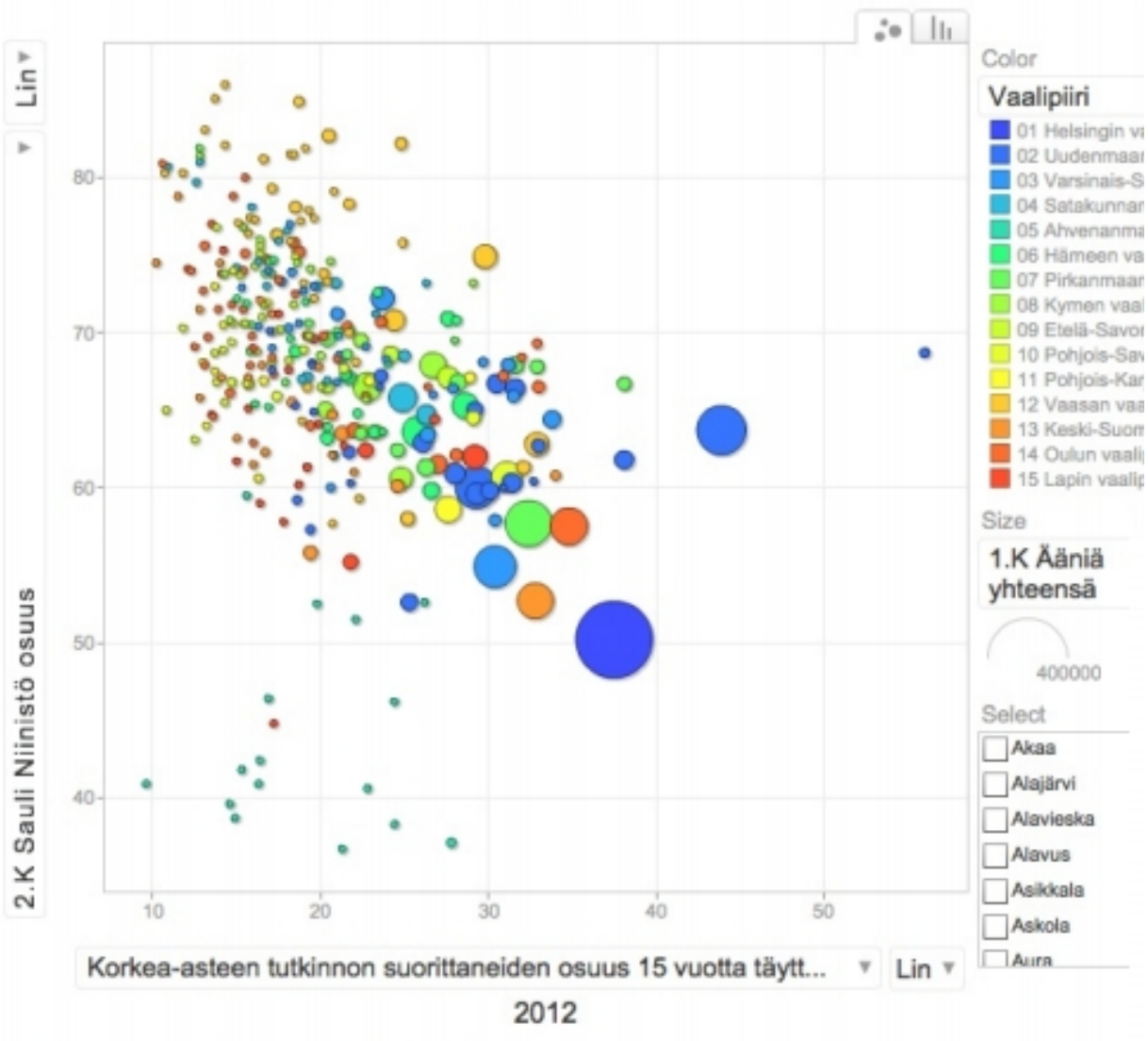
Technical, cultural, and historical challenges still forming **bottlenecks for research code and data sharing**

Open collaboration & social coding revolution

Presidenttiehdokkaiden kannatus ja suomalaisten hyvinvointi

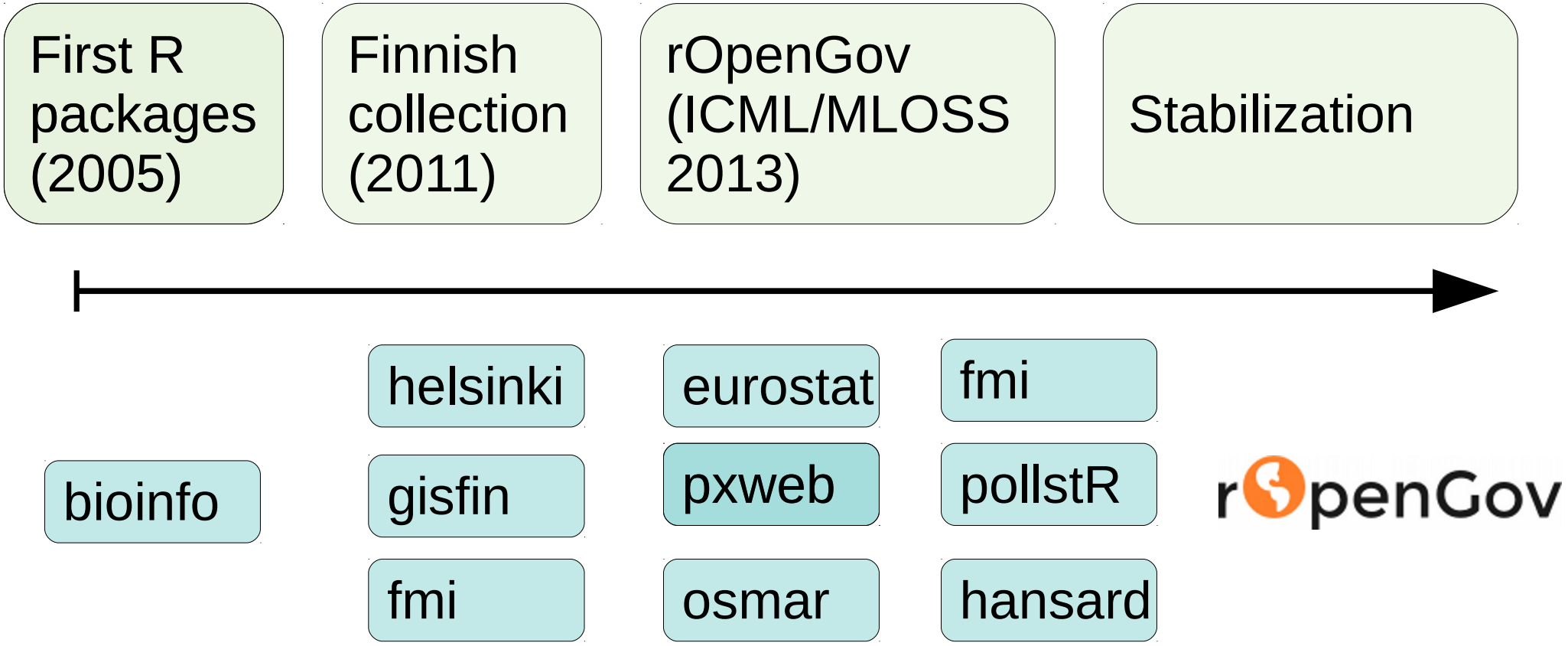
Julkaistu helmikuu 16, 2012 by antagomir

louhos.wordpress.com



Data:
- Land Survey Finland
- Statistics Finland
- YLE / HS

History of an open science community project



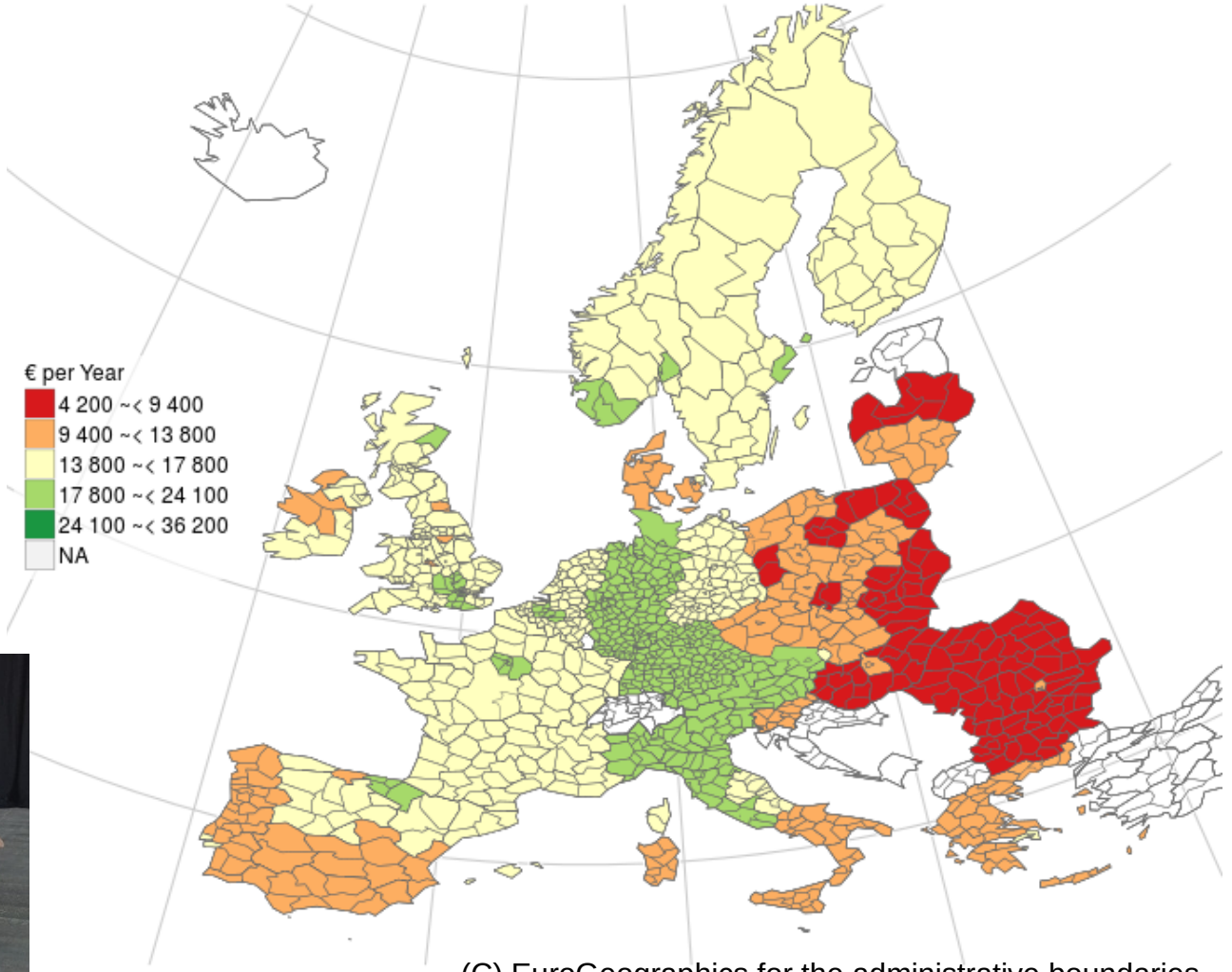
Retrieval and Analysis of Eurostat Open Data with the eurostat Package

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

rOpenGov

International network for open government data analytics

- 20+ R packages
- 100k+ downloads
- open collaboration



(C) EuroGeographics for the administrative boundaries

Eurostat open data: average household expenditure in 2011

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

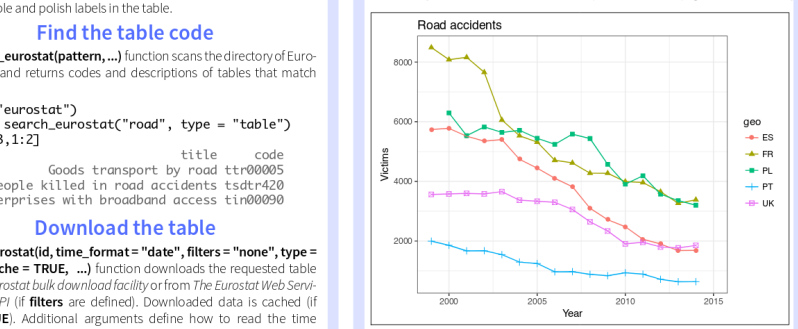
The **label_eurostat(x, lang = "en", ...)** gets definitions for Eurostat codes and replace them with labels in given language ("en", "fr" or "de").

```
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
```

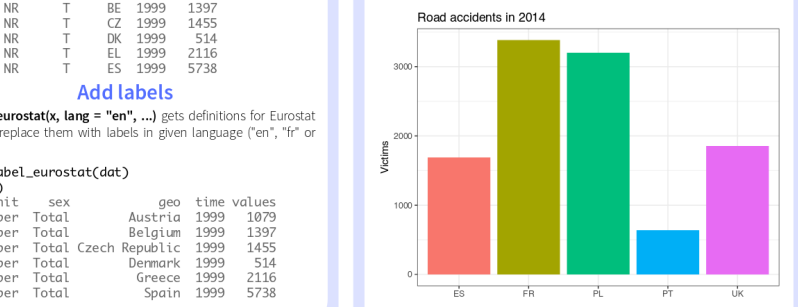
eurostat and plots

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

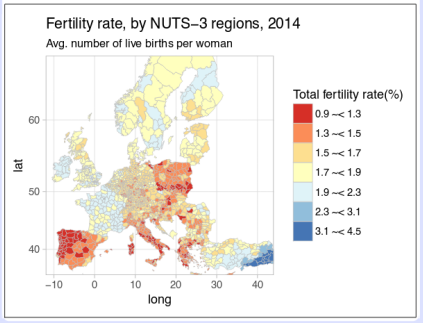
There are three function to work with geospatial data from GISCO. The **get_eurostat_geospatial()** returns preprocessed spatial data as sp-objects or as data frames. The **merge_eurostat_geospatial()** both downloads and merges the geospatial data with a preloaded tabular data. The **cut_to_classes()**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



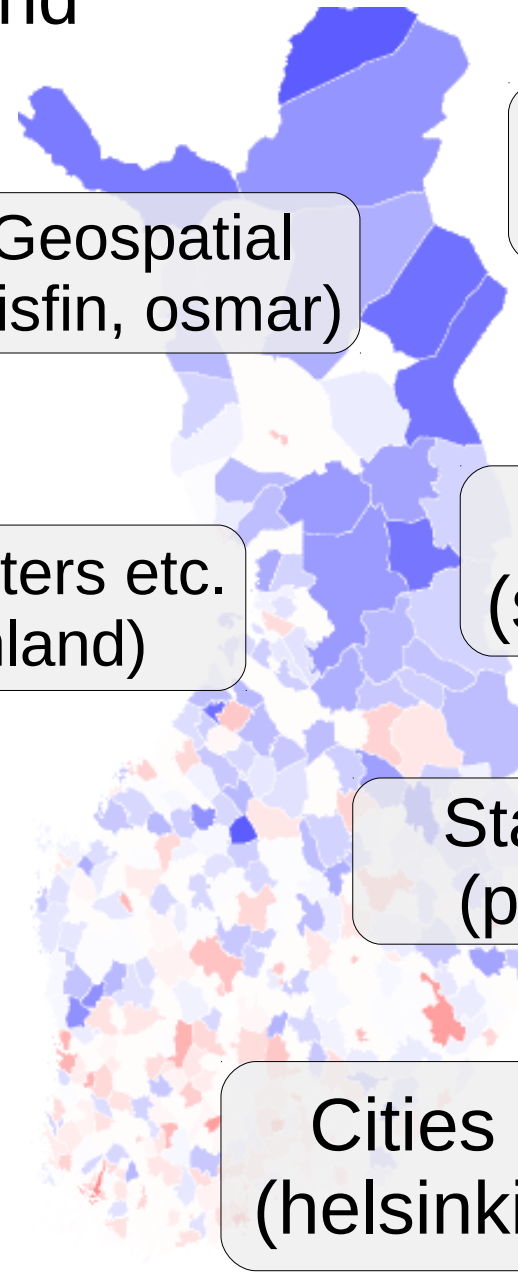
This onepager presents the **eurostat** package
Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek 2014-2017 package version 2.2.43 URL: *https://github.com/rOpenGov/eurostat*

From specific packages to package ecosystems

Open Street Map
Helsinki (osmar)

Algorithms for open data in Finland

KELA Open
Data & Methods?



Weather (fmi)

Geospatial (gisfin, osmar)

Health (sotkanet)

Registers etc. (finland)

Statistics (pxweb)

Cities (helsinki)

datavaalit

SITRA

MAANMITTAUSLAITOS
LANTMÄTERIVERKET

16 40 KANSALLIS KIRJASTO

Tilastokeskus
Statistikcentralen
Statistics Finland

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**

Analysis & Figure: Janne Sinkkonen

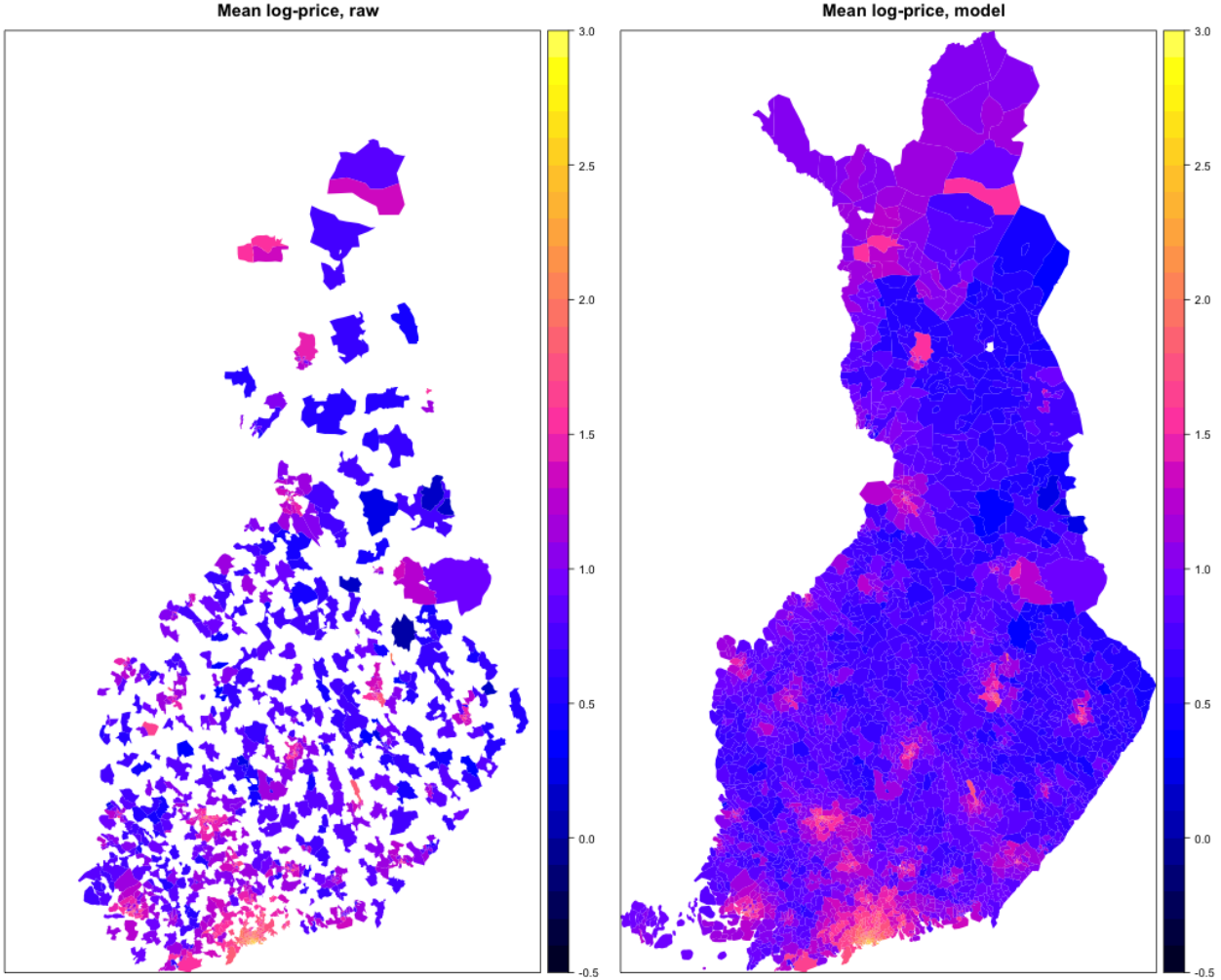
BY JANNE SINKKONEN — JUNE 11, 2015

A hierarchical model of Finnish apartment prices

Basing on open data from Statistics Finland, we at Reaktor modelled Finnish apartment prices and their trends on zip-code level, in the years 2005–2014. Estimates from the model are available as an interactive visualization.

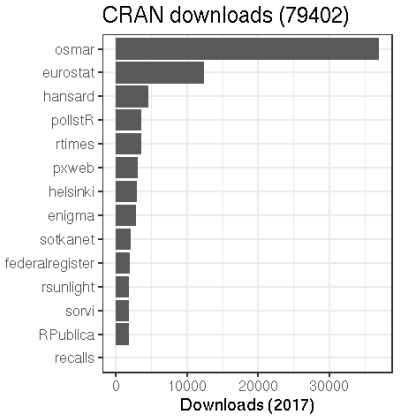
Original data: mean price per postal code area

Probabilistic model (Stan): smooth price estimates



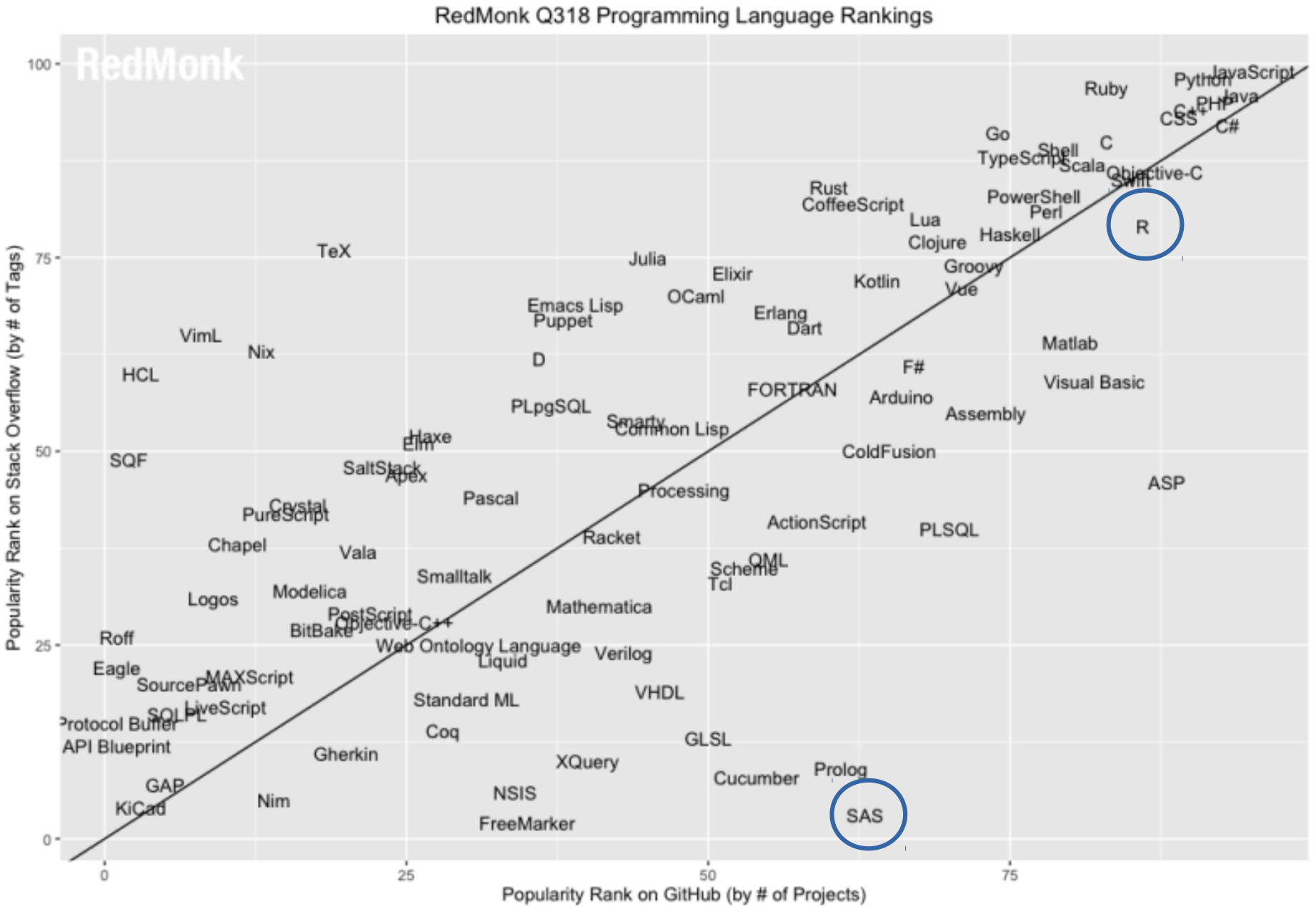
rOpenGov

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



open data science network

Varying cultures of open collaboration



Helsinki Computational History Group

https://comhis.github.io/

Computer scientists researching open workflows, algorithms and interfaces for humanities text and metadata

Linguists exploring the relationship between words and concepts

Historians interested in conceptual and actual historical processes

Bibliographic Data Science and the History of the Book (c. 1500-1800)

Cataloguing & Classification Quarterly, 2019 (in press).

Leo Lahti, Jani Marjanen, Hege Roivainen, Mikko Tolonen

Helsinki Computational History Group

National bibliographies have been identified as a crucial resource for historical research on the publishing landscape, but using them requires addressing challenges of data quality, completeness, and interpretation. We call this approach *bibliographic data science*. In this paper, we briefly assess the development of book formats and the vernacularization process in early modern Europe. The work undertaken paves the way for more extensive integration of library catalogues to map the history of the book.

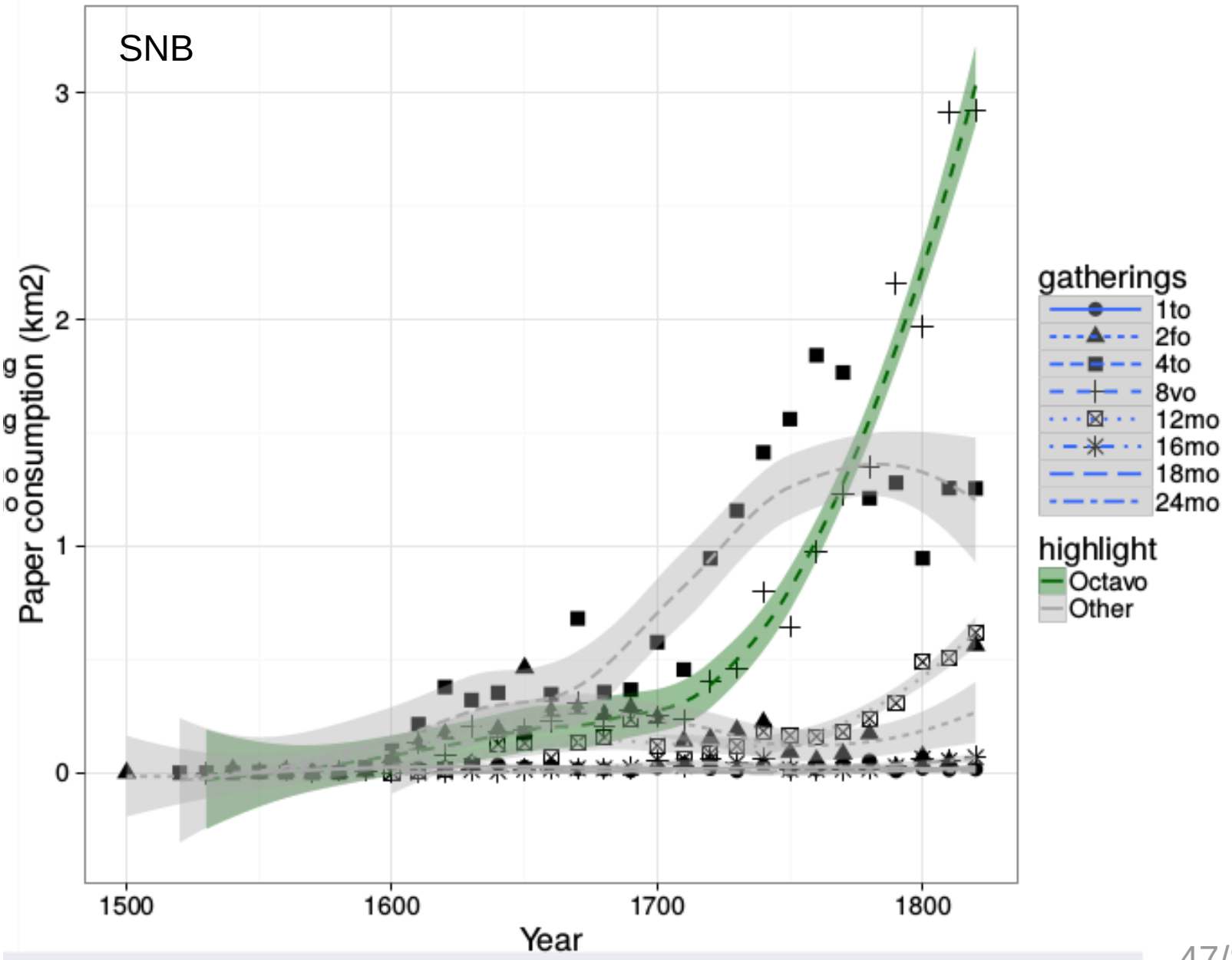
Potential of data science in SSH research?

- New methods, classical questions
- Entirely new scales of quantitative analysis
- Transparent conclusions
- Quality through collaboration

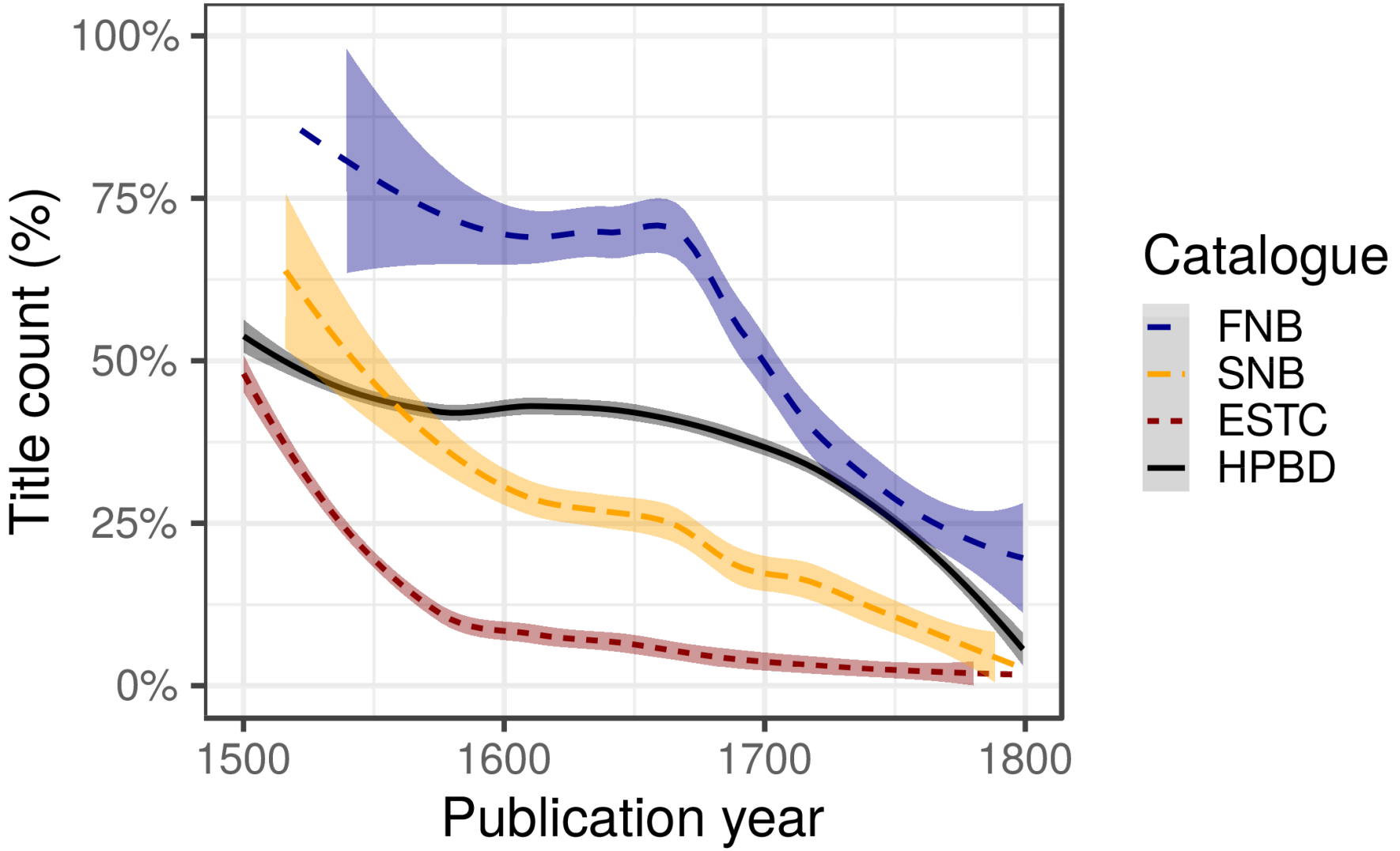
Pitfalls of data science in SSH research?

- Data quality overlooked
- Expertise lacking
- Tools drive research
- Unrealistic expectations

The rise of Octavo: paper consumption



Title count share for books in Latin (primary language)



Data: library catalogues

Catalogue	Earliest Record	Records 1500-1800 (N)	Language available	Publication place available	Page count available	Gatherings available
FNB	1488	16365	100.0%	93.9%	99.9%	98.3%
SNB	1457	46764	100.0%	95.0%	99.9%	84.8%
ESTC	1473	479790	100.0%	99.4%	99.9%	97.0%
HPBD	1446	2095628	100.00%	86.7%	99.5%	45.3%

FNB (Fennica)

Finnish National bibliography

- >900,000 books and monographies (printed and electronic) since 1488
- >70,000 continuous publications (journals or series) since 1771
- Series, maps, audiovisual, and electronic material
- **Open data**

SNB (Kungliga)

Swedish National bibliography

>18 million entries

ESTC

British Library

> 500,000 entries

Heritage of the printed book database (CERL/HPBD)

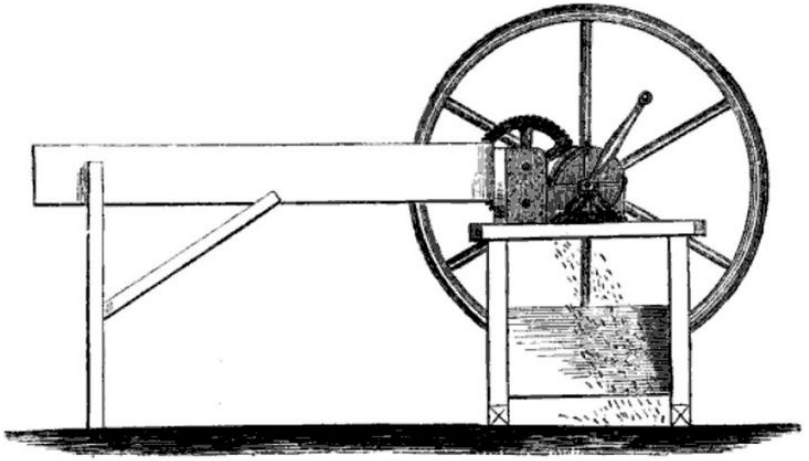
Göttingen.

>2M entries 1470-1800

>6M entries total

Data harmonization: estimating page counts from MARC cataloguing standards

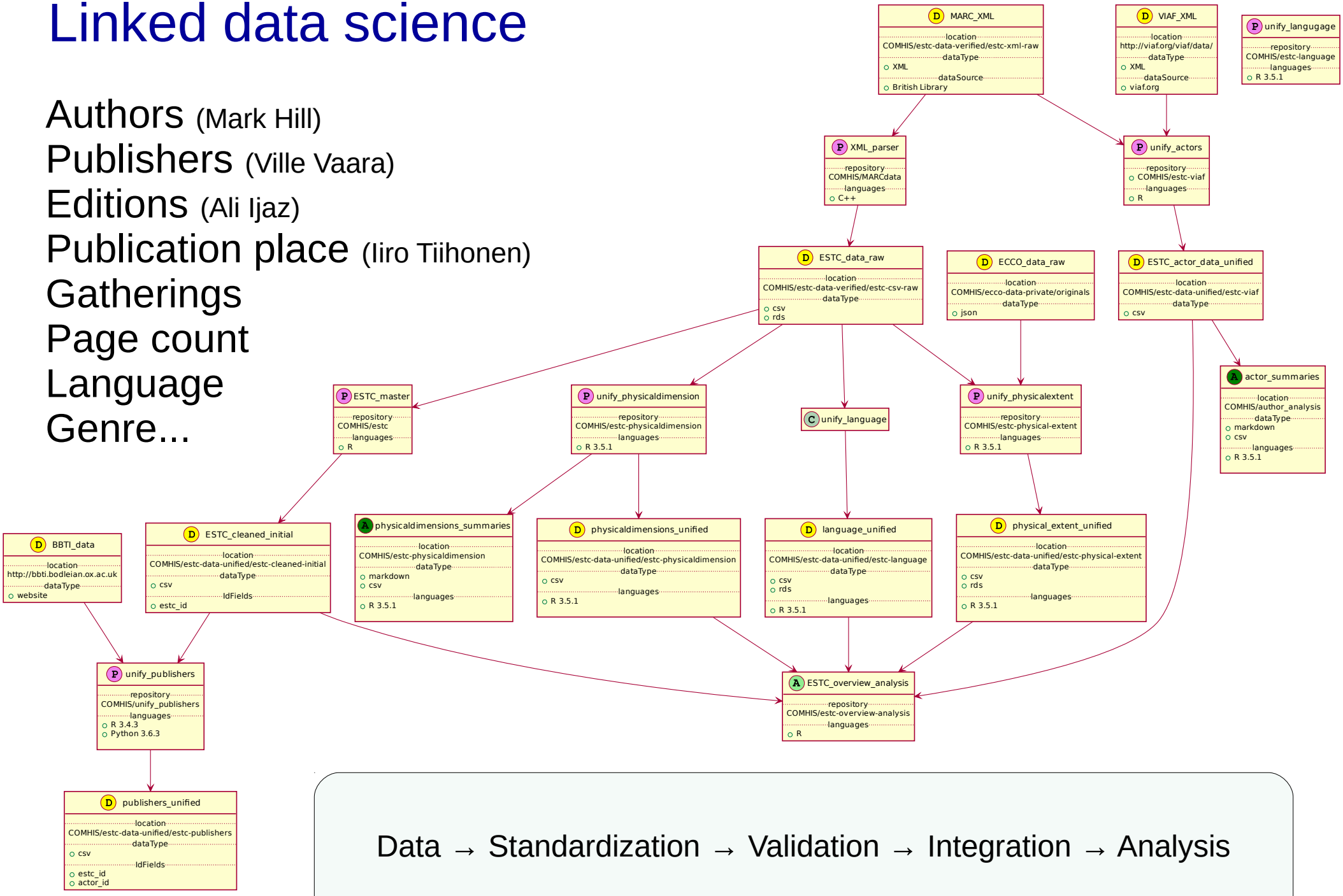
“[4],vii-xii,[4],222p.,plate”



→ 240 pages

Linked data science

Authors (Mark Hill)
Publishers (Ville Vaara)
Editions (Ali Ijaz)
Publication place (Iiro Tiihonen)
Gatherings
Page count
Language
Genre...



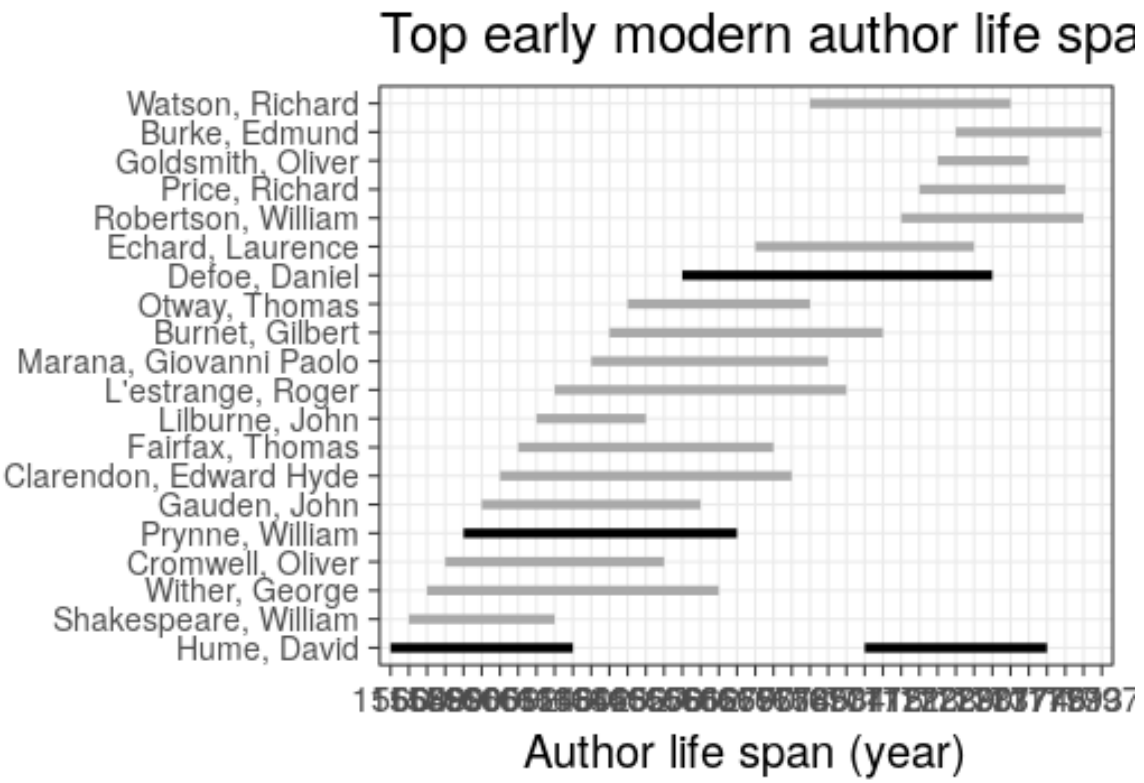
Data → Standardization → Validation → Integration → Analysis

Automated summaries for the unified data

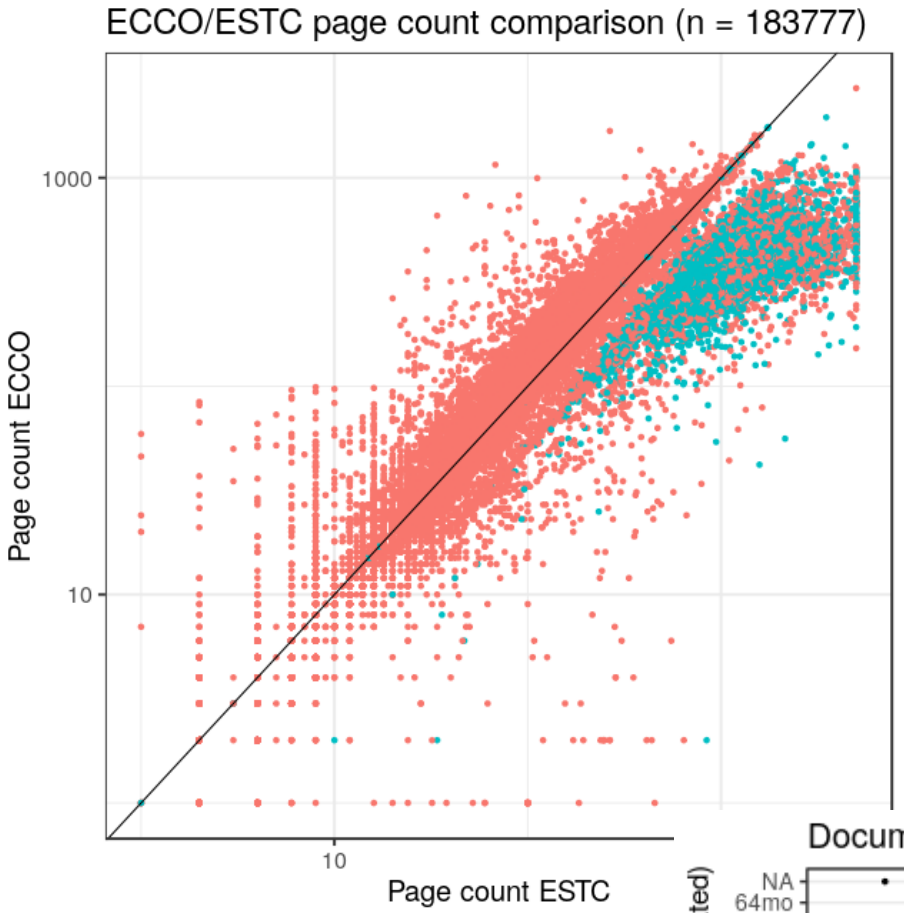
The data spanning years 1488-1955 has been included and contains 70451 documents on the data collection, see the source code for details.

Specific fields

- Author info
- Gender info
- Publisher info
- Publication geography
- Publication year info
- Titles
- Page counts
- Physical dimension
- Document and subject topics
- Languages



Validation: page count (ESTC ~ ECCO)



Good match:
estimated and
curated page counts

pagecount.estimated
FALSE
TRUE

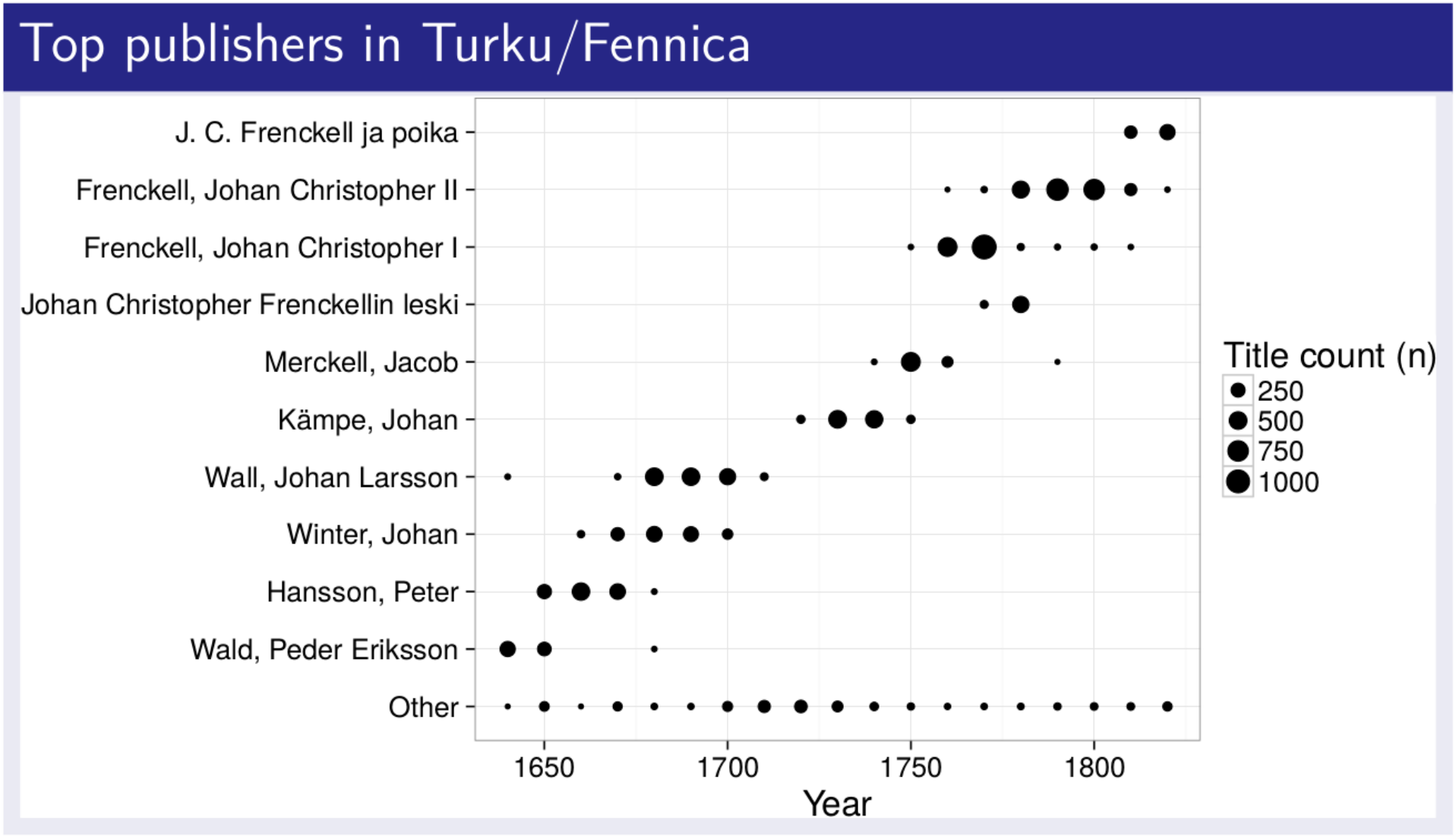
Information for certain book formats and time periods is missing more often → bias?

HISTORICAL METHODS
https://doi.org/10.1080/01615440.2018.1526657

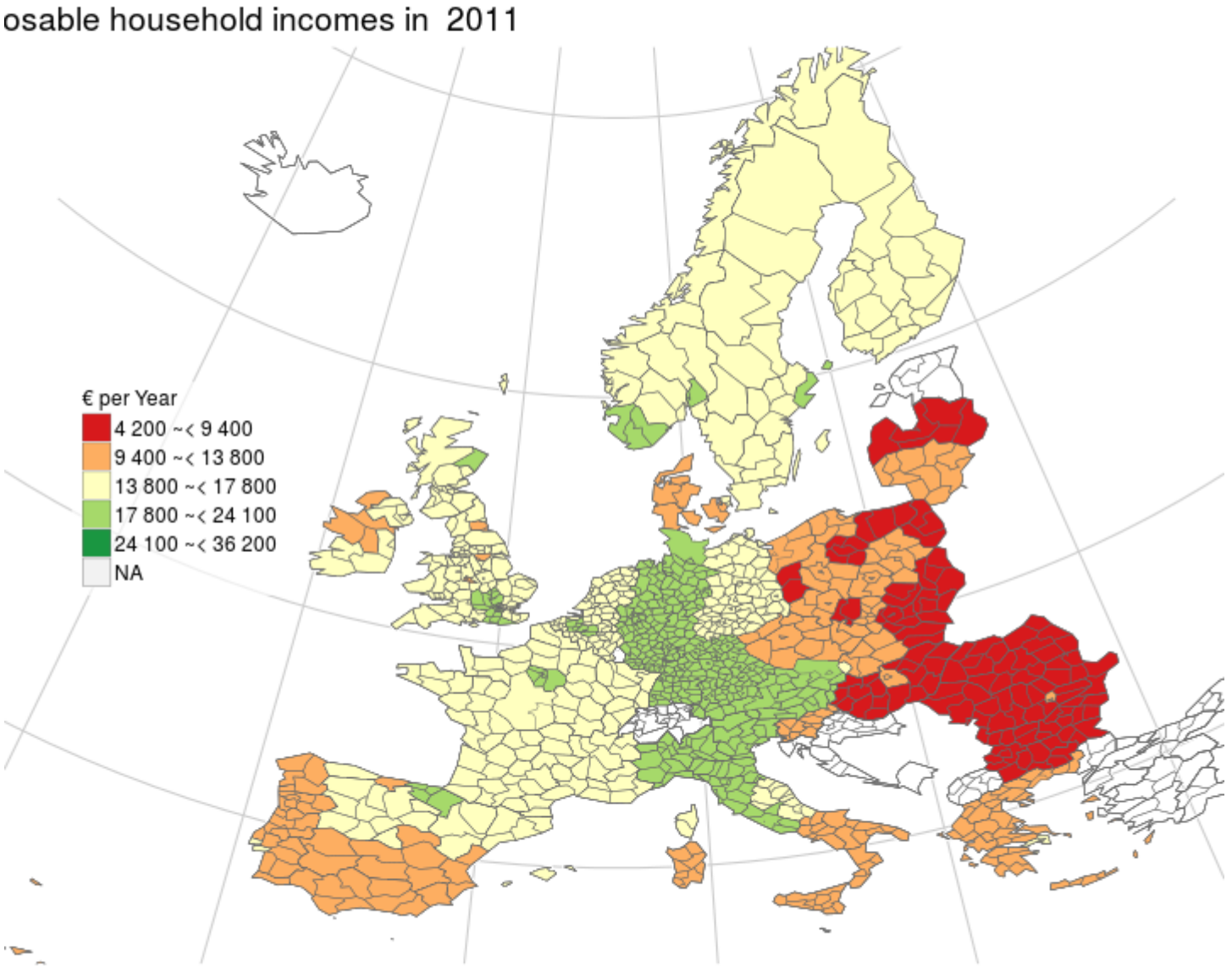
Routledge
Taylor & Francis Group

A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828

Mikko Tolonen[a], Leo Lahti[b], Hege Roivainen[a], and Jani Marjanen[a,*]



Information diffusion & spread



Fennica: analysis of the Finnish national bibliography

This repository contains automated analysis of the Finnish national bibliography, Fennica. Fennica includes bibliographic metadata for over 70,000 documents between 1488-1955, representing the publishing activity in Finland during that period. This is analyzed in parallel with Kungliga, a related collection of bibliographic metadata from the Swedish National library.

The research project is funded by Academy of Finland 2016-2019.

Reproducible analysis

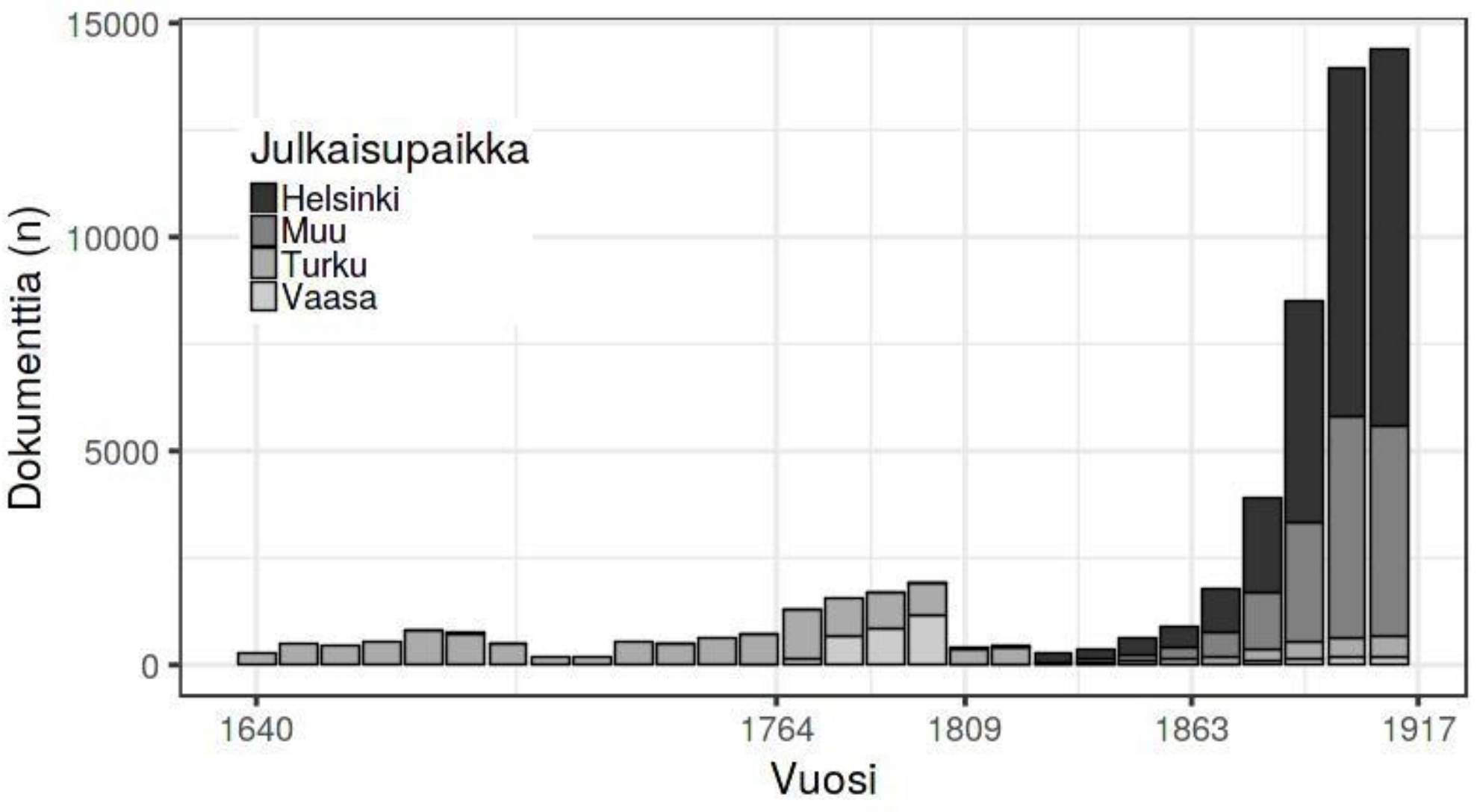
The data is summarized in the following automatically generated files:

- Fennica: a generic overview
- Fennica: a specific overview (Fennica specific preprocessing notes)
- Presentation slide templates (PDF) and code
- A Quantitative Approach to Book Printing in Sweden and Finland, 1640–1828 Source code for the figures
- Knowledge production in Finland 1470-1828: Digital Humanities 2016 conference presentation slides (PDF) and code
- Analyses on specific publication places and other topics (see the .md files)
- Figures and analyses for CCQ2019

The analyses cover several steps including XML parsing, data harmonization, removing unrecognized entries, enriching and organizing the data, carrying out statistical summaries, analysis, visualization and automated document generation. The analyses and full source code are provided in this repository and can be freely reused under the BSD 2 clause (FreeBSD) open source licence. The analyses are based on the R and rely on the custom bibliographica package for bibliographic data analysis, as well as many other R packages. The original raw data is available only on a separate agreement, so we are here publishing only the statistical summaries and our own analysis code.

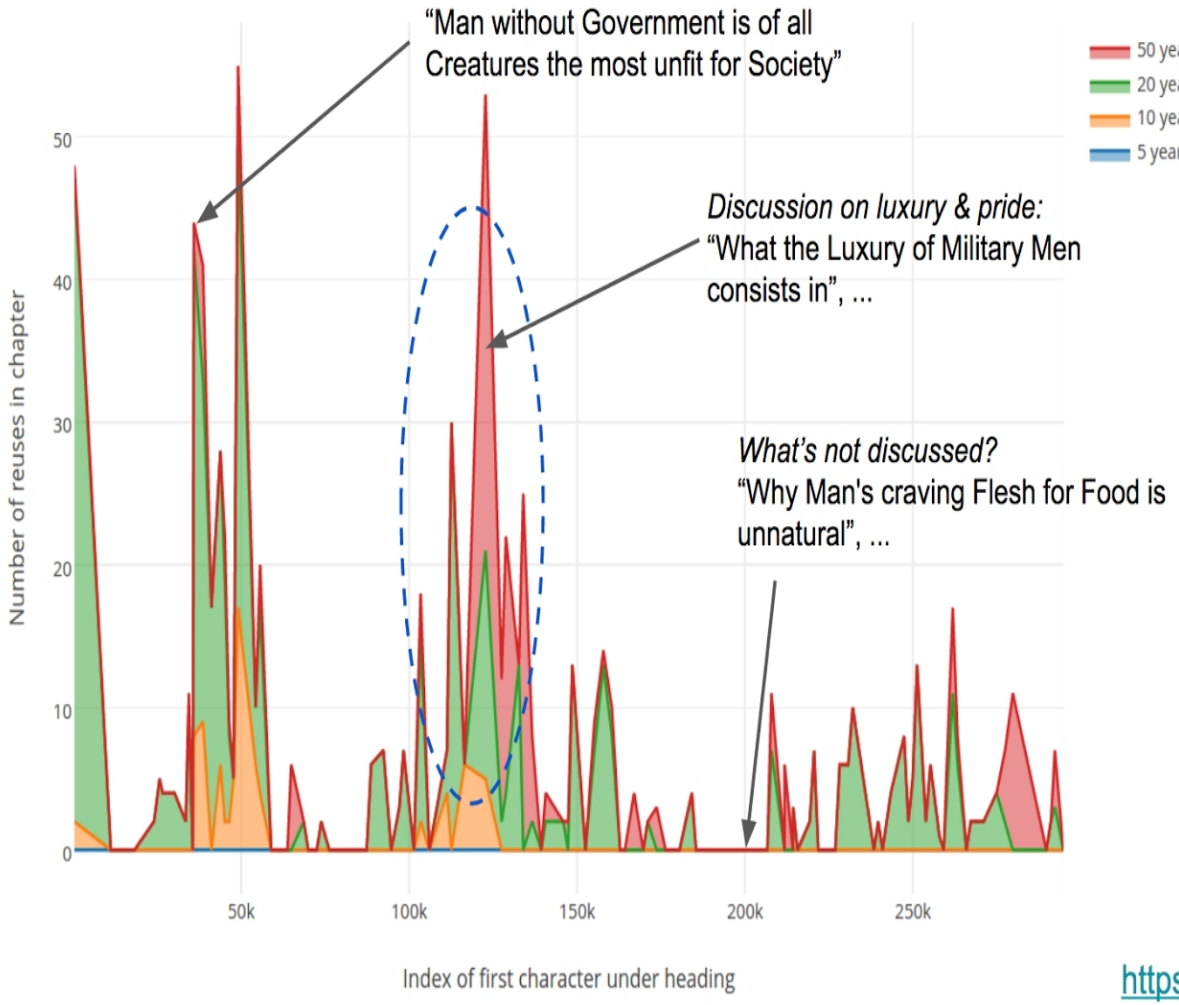
github.com/COMHIS/fennica

Publishing in Finland 1640-1917 (title count)



Mandeville: The Fable of the Bees (1714)

"Fable" reuses by chapter heading, x years from publication



CONTENTS.

https://plot.ly/~villepvaara/7/

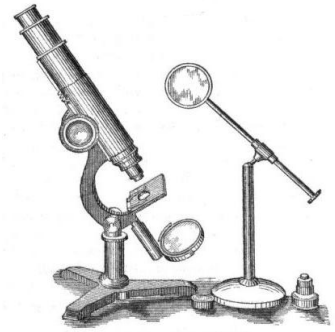
Comhis
Helsinki Computational History Group

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

Towards a greater transparency in academic journal subscription costs

Leo Lahti (@antagomir)
University of Turku & Open Knowledge Finland

Munich Open Access Days Oct 11, 2016



Suomi maailman kärkeen tiedejulkaisujen hintatietojen avoimuudessa

Finland takes leading role in the openness of academic journal pricing

June 13, 2016

Freedom of Information request by open science advocates has revealed academic journal pricing through an administrative court decision. Finland is the first country where the subscription prices paid by practically all universities and research institutions to individual publishers are made available. This strengthens the position of universities in the 2016 contract negotiations, made ever more timely by the recent deep funding cuts. Comparisons between publishers and countries also supports the ongoing discussion of alternative publishing models and directing funding towards open access (OA) publishing.

The costs of academic journals have risen precipitously, but the lack of detailed pricing information has made the overall situation difficult to perceive. There are significant price differences between publishers, universities and countries. While dominant publishing houses have reported profit margins of tens of percent and the industry is ever more concentrated

Affiliatio

D+M

Disorder in Mate
Faculty of Phys
University of Vie
Austria

Finland: public spending & freedom of information

- Open Access vs. Subscription models
- Finland FOI process
- International aspects

Journal subscription costs are increasing ~10% per year

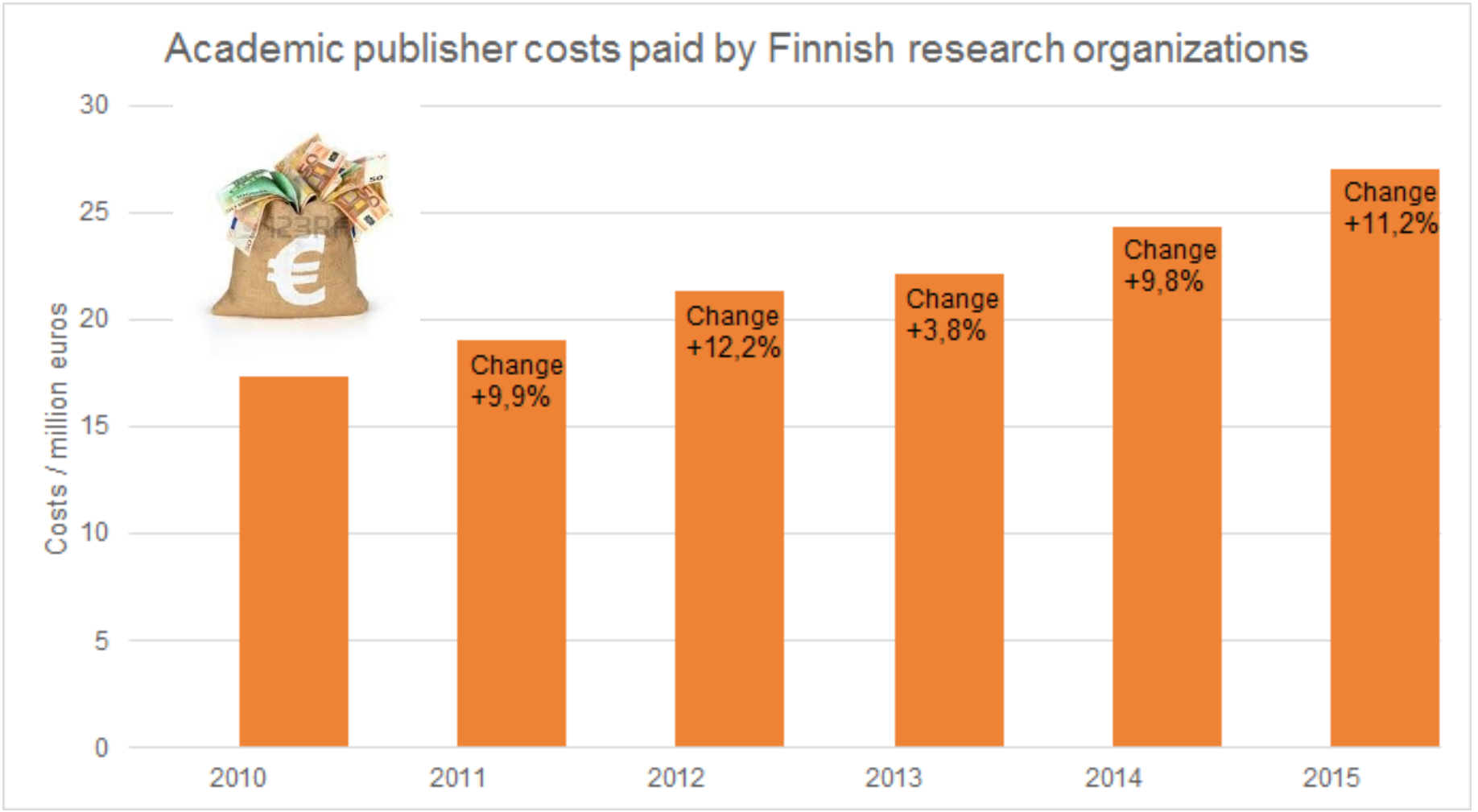
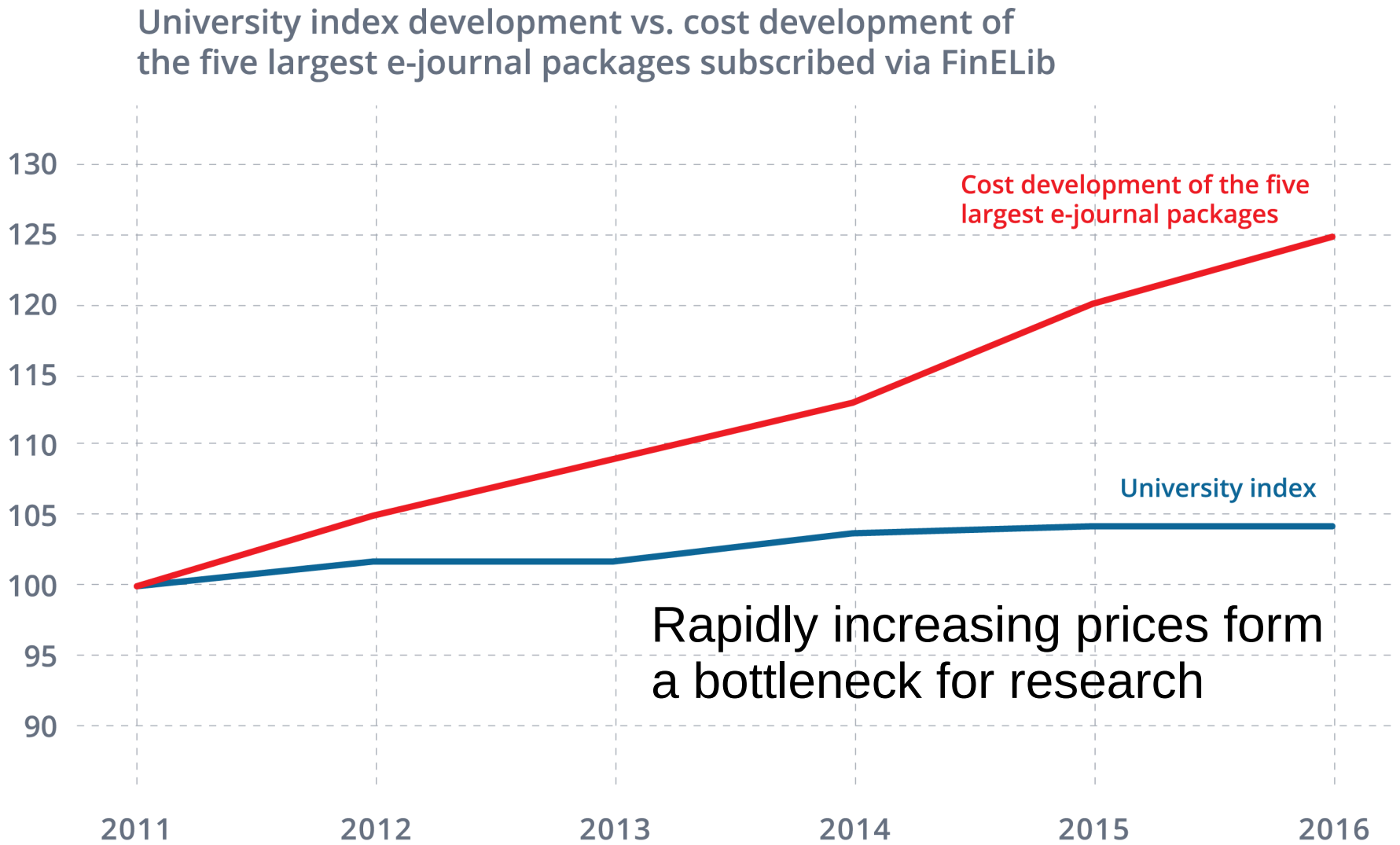


Figure & Data release: Ministry of Education and Culture of Finland / Open Science and Research Initiative 2014–2017

Subscription costs: some observations

- Cost increase 4x faster than inflation 1986-2007 (US Association of Research Libraries).
- Largest publisher profits around ~40% (eg Elsevier 2013 ~£800 million / ~£2 billion).
- High price variation among publishers; for-profit journals 3x more expensive than not-for-profit (Bergstrom et al. PNAS 2014).

Costs for the top-5 journal packages increased 25% while university funding nearly unaltered (Finland 2011-2016)

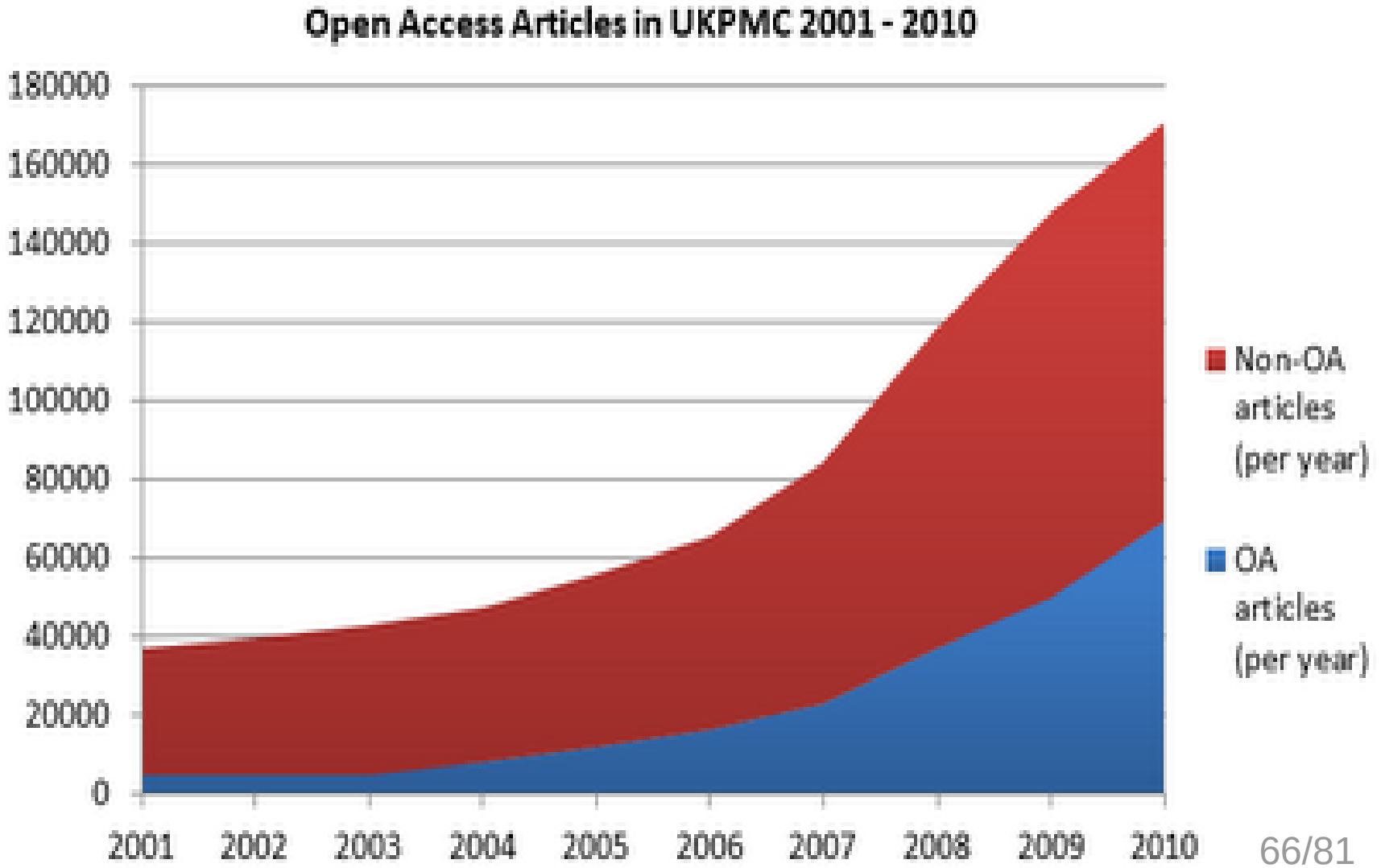


Source: https://www.kiwi.fi/pages/viewpage.action?pageId=64487647#What%27sgoingonwithscholarlyjournalnegotiations?-Accesstoscholarlyjournals%E2%80%93atwhatcost?(19.9.2016)

In 2016, Finland paid ~30 million euros to access academic journals. Third of this went to Elsevier, which has reported 30-40% profit margins.

Dramatic growth in open access publishing 2001-2010

Out of ~35 000 peer-reviewed academic journals (Ware & Mabe, 2015) less than a third (11 000) are open access (DOAJ, 2016).



Problems with the secret disclosure agreements

From the researcher perspective, OA is more expensive

<u>Open Access</u>

- Primary research funds
- Transparent price

→ Costs obvious and high

<u>Subscriptions</u>

- Library funding
- Secret agreements

→ Costs are hidden

Earlier FOI requests (US & UK)

- US 2009: data for 36 institutions collected by Courant/Bergstrom/McAfee ‘Elsevier made strenuous efforts to prevent the disclosures but a judge ruled against them’.
- UK 2014: wildly different subscription fees for the same journal bundles paid by different universities. JISC essentially refused to give the data but FOI request was successful (gowers.wordpress.com/2014/04/24/)

Only limited data available (not all universities or publishers were covered)

Huge variation between universities (UK)

UCL:

~2x King's College London

~6x Exeter

University	Cost	Enrolment	Academic Staff
Birmingham	£764,553	31,070	2355 + 440
Bristol	£808,840	19,220	2090 + 525
Cambridge	£1,161,571	19,945	4205 + 710
Cardiff	£720,533	30,000	2130 + 825
*Durham	£461,020	16,570	1250 + 305
**Edinburgh	£845,000	31,323	2945 + 540
*Exeter	£234,126	18,720	1270 + 290
Glasgow	£686,104	26,395	2000 + 650
Imperial College London	£1,340,213	16,000	3295 + 535
King's College London	£655,054	26,460	2920 + 1190
Leeds	£847,429	32,510	2470 + 655
Liverpool	£659,796	21,875	1835 + 530
§London School of Economics	£146,117	9,805	755 + 825
Manchester	£1,257,407	40,860	3810 + 745
Newcastle	£974,930	21,055	2010 + 495
Nottingham	£903,076	35,630	2805 + 585
Oxford	£990,775	25,595	5190 + 775
* ***Queen Mary U of London	£454,422	14,860	1495 + 565
Queen's U Belfast	£584,020	22,990	1375 + 170
Sheffield	£562,277	25,965	2300 + 460
Southampton	£766,616	24,135	2065 + 655
University College London	£1,381,380	25,525	4315 + 1185
Warwick	£631,851	27,440	1535 + 305
*York	£400,445	17,405	1205 + 285

Table from: gowers.wordpress.com/2014/04/24/

Open Science Finland, April 2014

Joona Lehtomäki
April 24, 2014

Apparently the first breakdown on how much academic institutions (in the UK) are paying for publishers (or a publisher, Elsevier).

The cost of academic publishing | Open Access Working Group

Sharing the results of publicly funded research

ACCESS.OKFN.ORG

Like Comment Share

Juha Yrjölä, Toma Susi and 19 others

Toma Susi Kuka selvittäisi Suomessa?
April 24, 2014 at 8:59pm · Unlike · 1

Joona Lehtomäki Hyvä kysymys. Liisa Siipilehto: mistäs saisi samantyyppisiä lukuja Helsingin yliopistolle?
April 24, 2014 at 9:07pm · Like

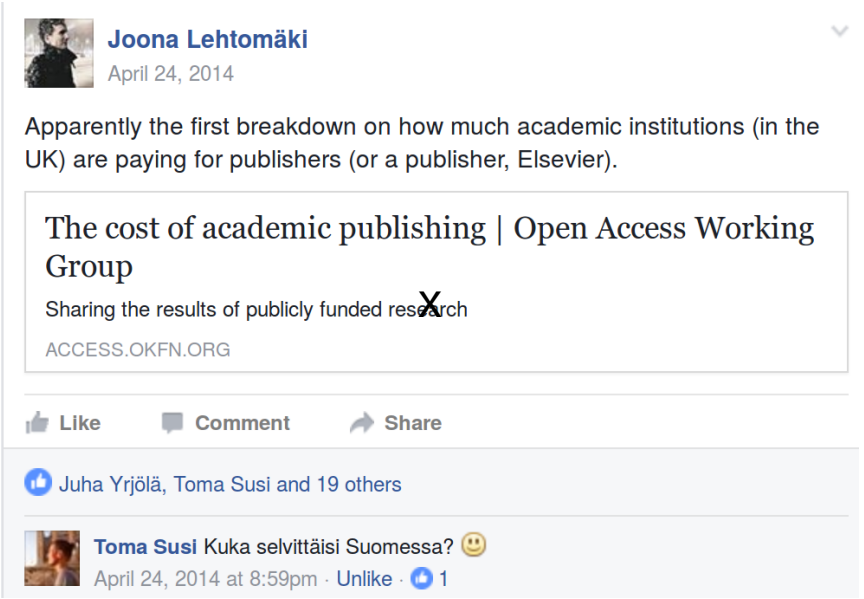
Toma Susi HY:tä sitovat saleen samat klausuulit, onkohan Suomessa FOIA-tyylistä mekanismia?
April 24, 2014 at 9:15pm · Like

Joona Lehtomäki Salettiin. Wikipedia listaa eri maiden lainsäädäntöä asian tiimoilta ja Suomessa on esim. "Laki yleisten asiakirjain julkisuudesta"
://en.wikipedia.org/.../Freedom_of_information_laws...

c-publishing/

- Initial FOI request to all Finnish universities
- According to legal advice, this information is public
- Universities refused

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

Legal precedent: focus on a single university and publisher

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

FOI process: Finland 2014-2016

2014
FOI request &
Appeal in court
(easy & cheap!)

2015
Positive
court
decision

2016
Data release
by MoE

Ample support
from law experts

Laborious data
collection &
harmonization

Two years from the initial requests to actual data release

University viewpoint not clear

Court decision for protection against publishers ?

Serious efforts to prevent information release against their own interests and openness strategy. Reasons not entirely clear.

The largest university considers appropriate to express dislike on public discussion in Facebook.

Finally, even very confidential documents sent to us.

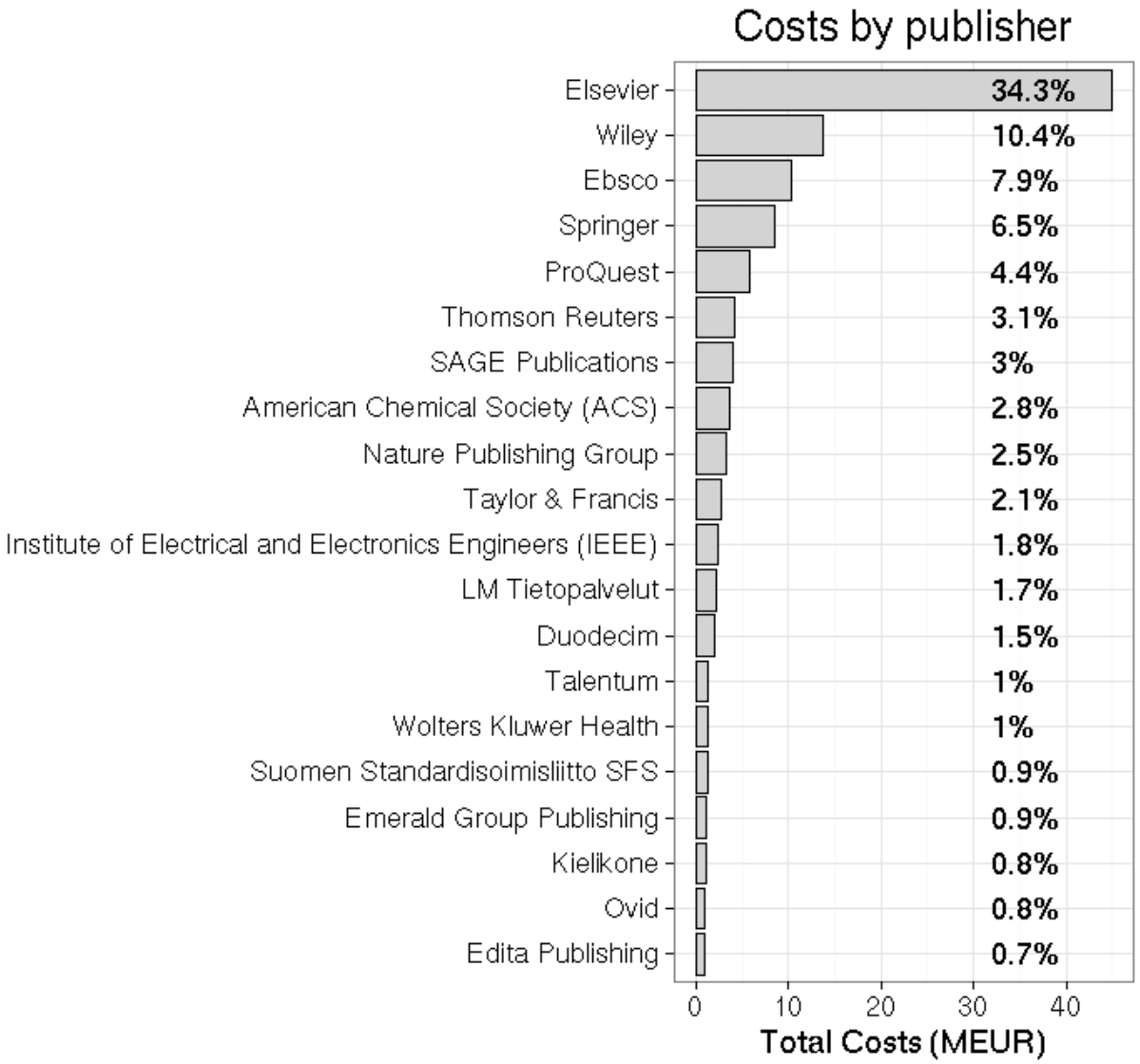
→ Severe problems handling a routine case.

Ministry of Education and Culture stands out to help after court decision

- 2015. Meetings between Aalto University, Ministry of Education Open Science Initiative, and Open Knowledge Finland.
- 2015/16. Data collection and harmonization coordinated by Ministry of Education. A major effort, best carried out by an official body.
- June 2016. Finland became the first country to release detailed data on subscription fees openscience.fi/publisher_costs

Total costs per publisher (2010-2015)

- 244 publishers
- 63 public institutions
- 27 million euros in 2015
- 131.1 million euros 2010-2015
- Costs are per bundle, cant compare individual publishers per article or per citation basis
- Elsevier prices per citation are roughly 3x higher than with non-profit publishers. Emerald, Sage, and Taylor & Francis had ~10x higher price (Bergstrom et al. PNAS 2014).

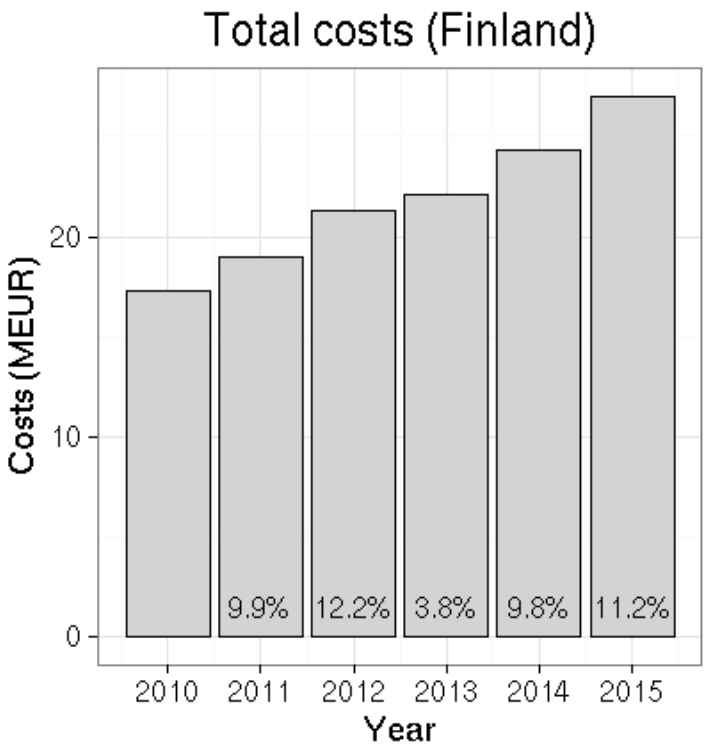


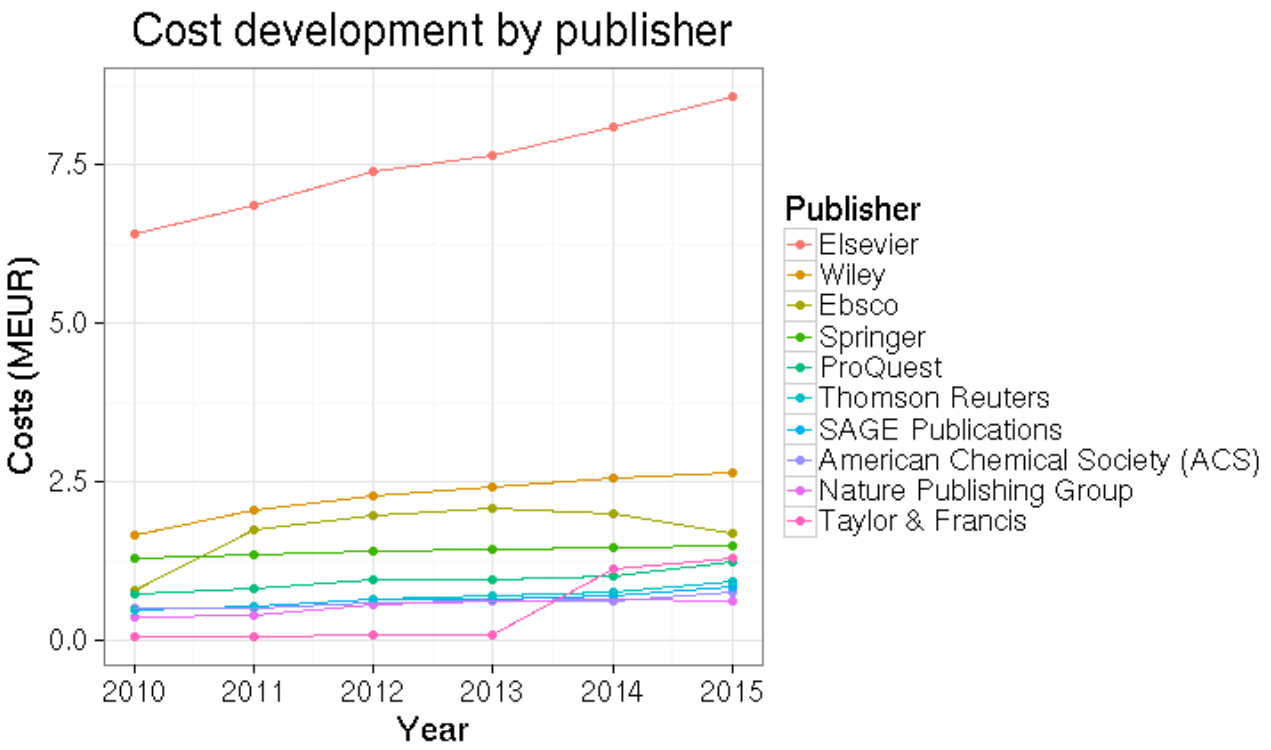
Data: Finnish MoE / ATT 2016.

Figure: Leo Lahti 2016, ropengov.github.io

Costs over time

Top-10 publishers correspond to 77% of the overall costs.





Data: Finnish MoE / ATT 2016.

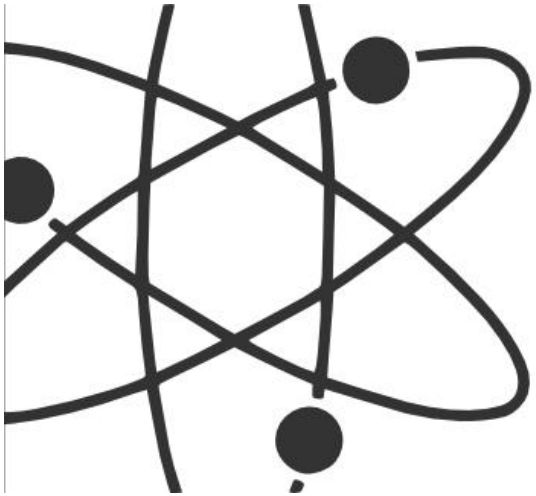
Figure: Leo Lahti 2016, ropengov.github.io

Recent policy recommendations

- PlanS & other policies
- Call on scientific publishers to “adapt their business models to new realities” (Commissioner Moedas and Secretary of State Dekker)
- Subscription costs must be made transparent and OA supported (Amsterdam Call for Actions 2016 & Alliance of Science Organizations in Germany & The League of European Research Universities LERU 2015).
- All publicly funded research publications openly available by 2020 (EU Competitiveness Council target).

THE COST OF SCIENTIFIC PUBLICATIONS MUST NOT GET OUT OF HAND

Heidi Laine



Open science means for instance open release of data, code, protocols, teaching material, publications, and the promotion of principles of openness, inclusivity and transparency in scientific research.

The **Open Science Finland** working group promotes openness in Finnish scientific and academic field.

AVOIN TIEDE JA TUTKIMUS

OPEN KNOWLEDGE FINLAND

Opening Academic Publishing

Development and application of systematic evaluation criteria

tiedonhinta.fi

Publisher	OA Journals	Costs	Licensing	Self-archiving	Open Citations	TDM	Access to info	Total
ACM	0	1	1	3	3	1	3	**12**
ACS	1	0	1	0	0	1	3	**6**
Elsevier	1	0	2	2	0	2	2	**9**
IEEE	1	1	1	3	0	2	3	**11**
LWW	1	2	2	2	0	0	2	**9**
Sage	1	1	2	1	3	2	3	**13**
SpringerNature	1	1	2	2	3	2	3	**14**
Taylor and Francis	1	1	2	3	3	1	3	**14**
Wiley-Blackwell	1	0	1	1	3	2	3	**11**
Max. points								**21**

Thank You!

@antagomir

Mikko Tolonen
Heidi Laine
Joona Lehtomäki
Markus Kainu
Juuso Parkkinen
Antti Poikola
Przemyslaw Biecek
Måns Magnusson
Sudarshan Shetty
Ville Laitinen
Aaro Salosensaari
Willem de Vos

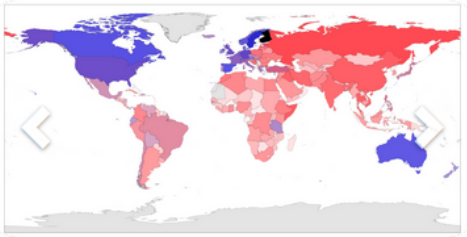
Welcome to Open Research Labs

Modern data analysis from theory to practice

We carry out research on theory, methods and applications of modern statistical and algorithmic data analysis, blending elements from various fields including statistical machine learning/AI, probabilistic programming, numerical ecology, and data science among others.

Our focus is on data-rich applications ranging from molecular life sciences to computational history. The research team is led by PI Leo Lahti who coordinates the Computational biosciences group in University of Turku, Finland. We are also founding members and part of Helsinki Computational History Group.

There is a great demand for targeted algorithmic methods to extract information and insights from data with minimal human intervention to guide modeling and experimentation. By combining information across multiple, complementary sources it is possible to overcome some of the limitations and statistical uncertainties associated with individual data sets. Open research practices play an important role in our all work.



Helsinki Computational History Group

OPEN KNOWLEDGE FINLAND

SUOMEN AKATEMIA
FINLANDS AKADEMI • ACADEMY OF FINLAND

Turun yliopisto
University of Turku

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

rOpenGov