CATALOGING & CLASSIFICATION QUARTERLY
https://doi.org/10.1080/01639374.2018.1543747

Routledge
Taylor & Francis Group

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

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

[a]Department of Mathematics and Statistics, University of Turku, Finland; [b]Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Finland

ABSTRACT

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 article, 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 catalogs to map the history of the book.

ARTICLE HISTORY
Received July 2018
Revised September 2018
Accepted October 2018

KEYWORDS
National bibliography; data ecosystem; publishing history; digital humanities; open science

Computational workflows have now increasingly central role in research: challenge & opportunity

1) data gathering and storage
2) access, documentation
3) harmonization & enrichment
4) quality control
5) custom analysis tools and workflows
6) reporting & dissemination

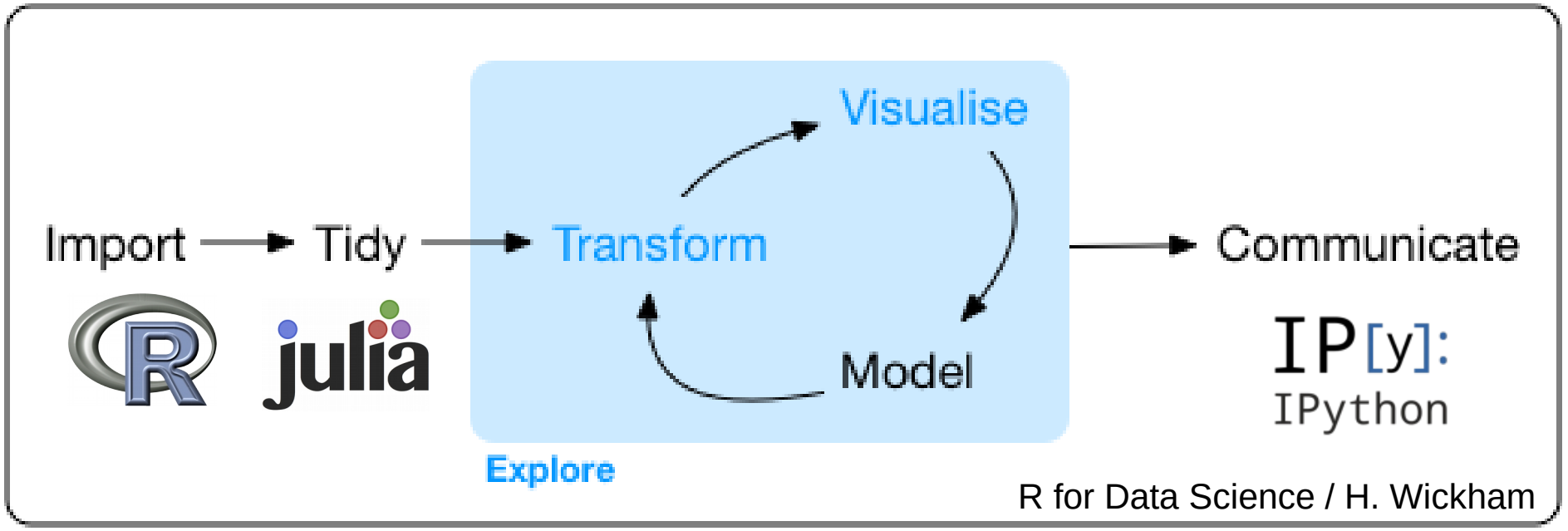
Science 13 April 2012:
Vol. 336 no. 6078 pp. 159-160
DOI: 10.1126/science.1218263

POLICY FORUM

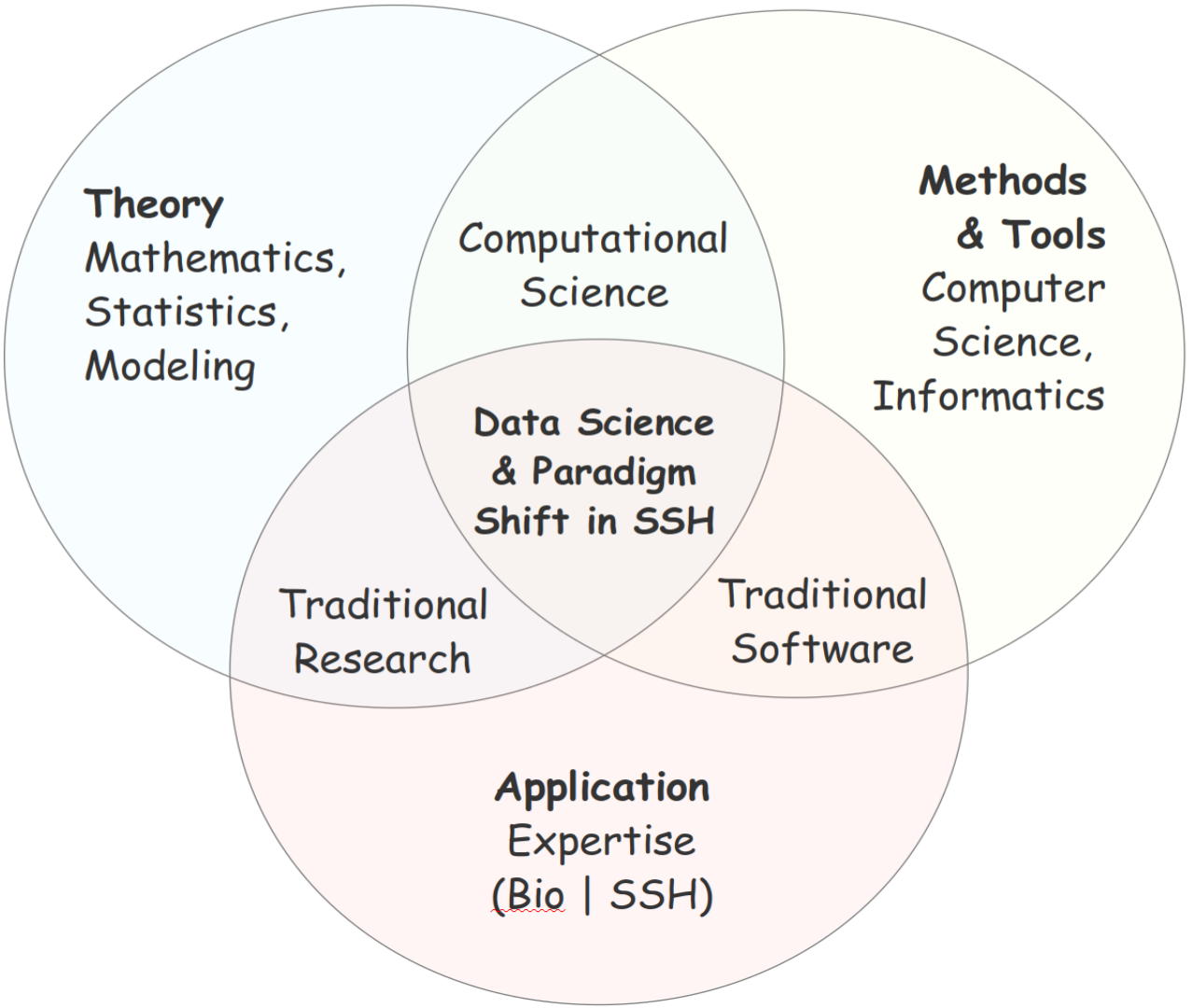
RESEARCH PRIORITIES

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,*]



Data science as a melting point of many fields

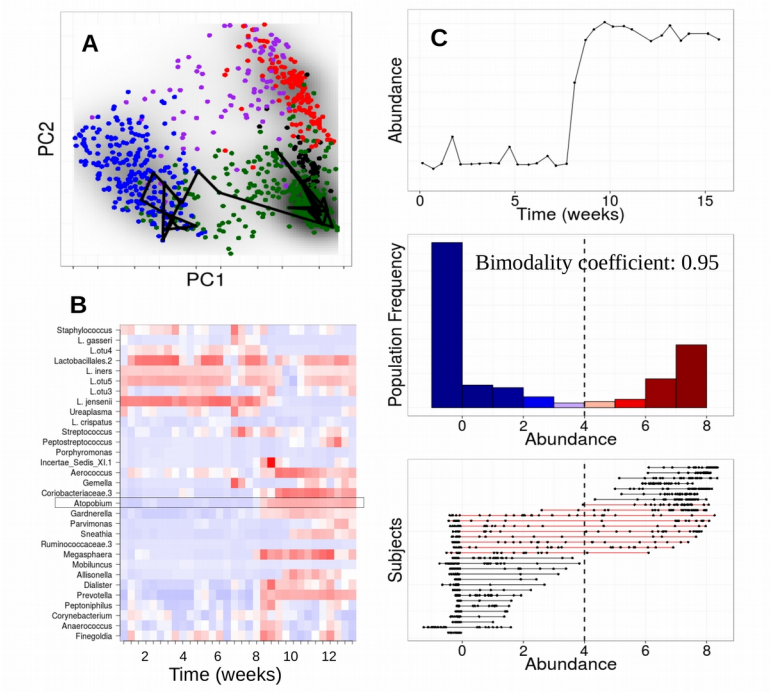


Bioinformatics

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]

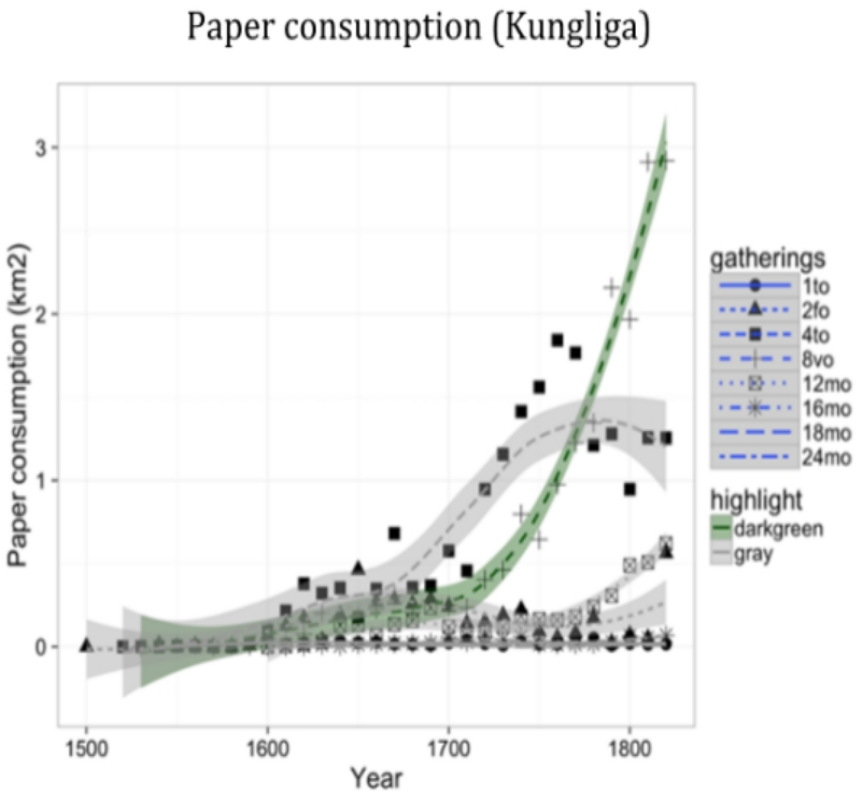
Digital Humanities

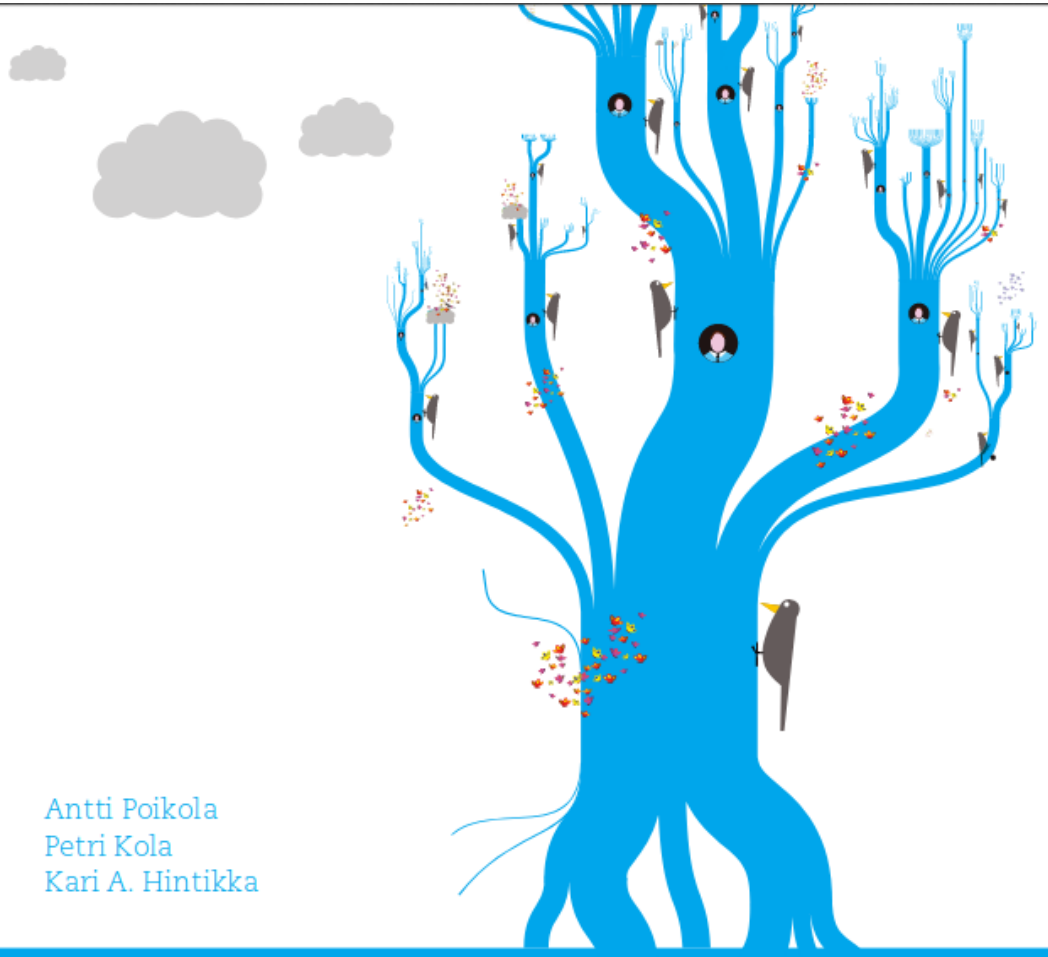
16 40 KANSALLIS KIRJASTO

AATEHISTORIA JA DIGITAALISTEN AINEISTOJEN MAHDOLLISUUDET

11.8.2015 MIKKO TOLONEN JA LEO LAHTI

CASE 1: The rise of the octavo-sized book





Antti Poikola
Petri Kola
Kari A. Hintikka

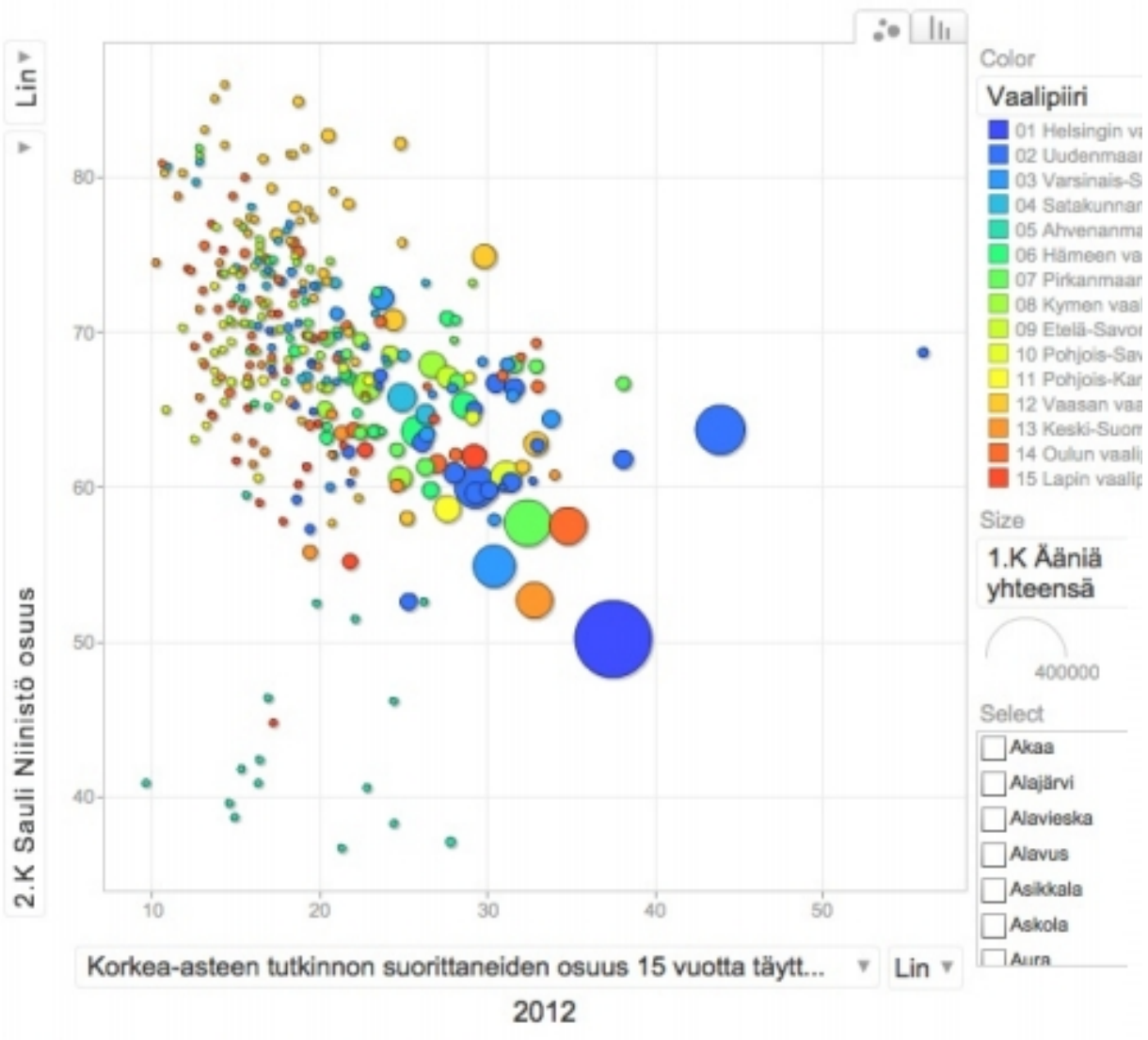
Julkinen data©

johdatus tietovarantojen avaamiseen

Presidenttiehdokkaiden kannatus ja suomalaisten hyvinvointi

Julkaistu helmikuu 16, 2012 by antagomir

louhos.wordpress.com



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

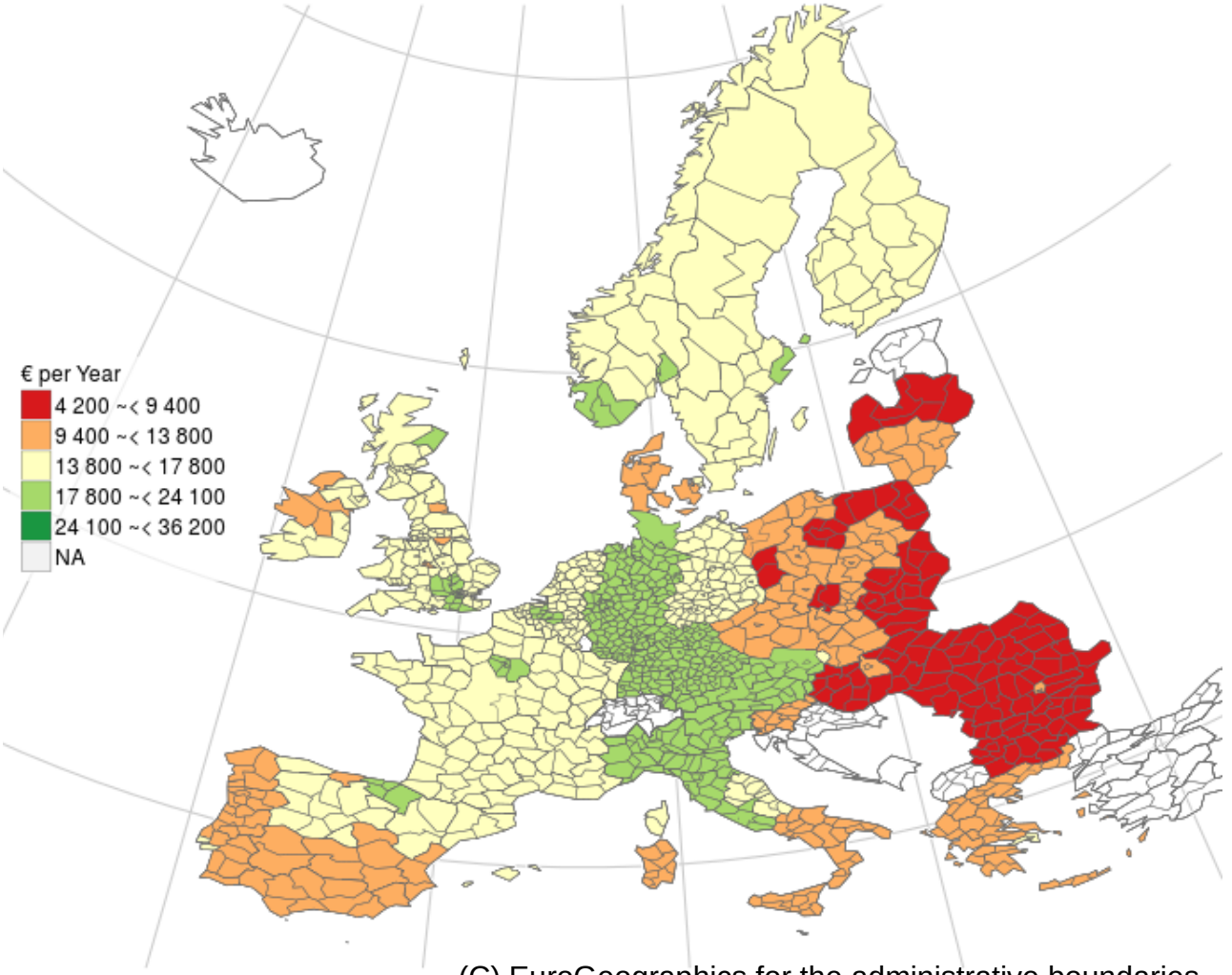
Retrieval and Analysis of Eurostat Open Data with the eurostat Package

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



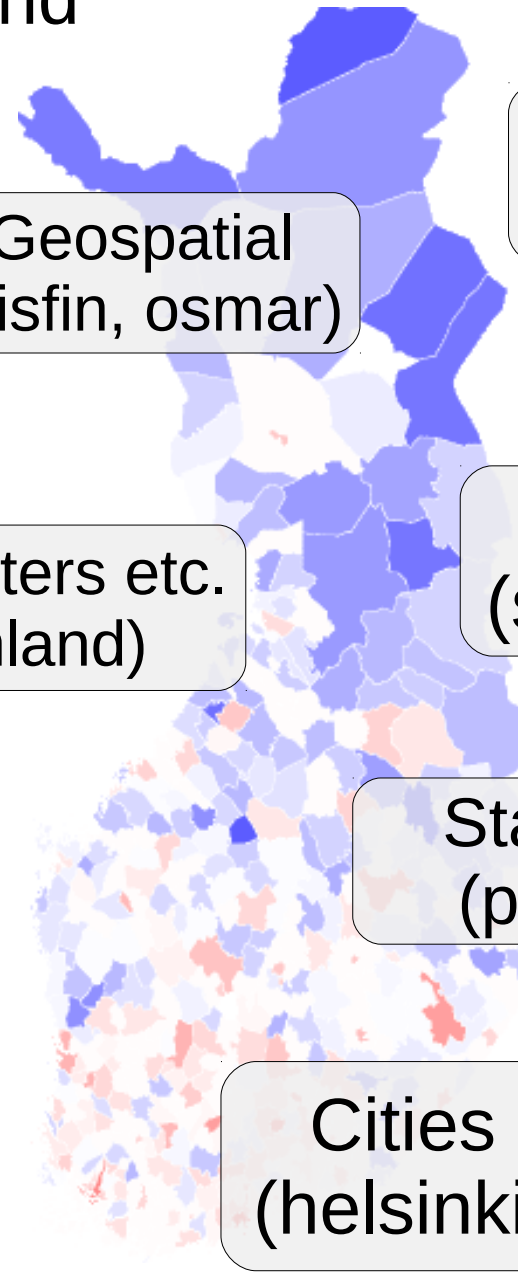
Eurostat open data: average household expenditure in 2011

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

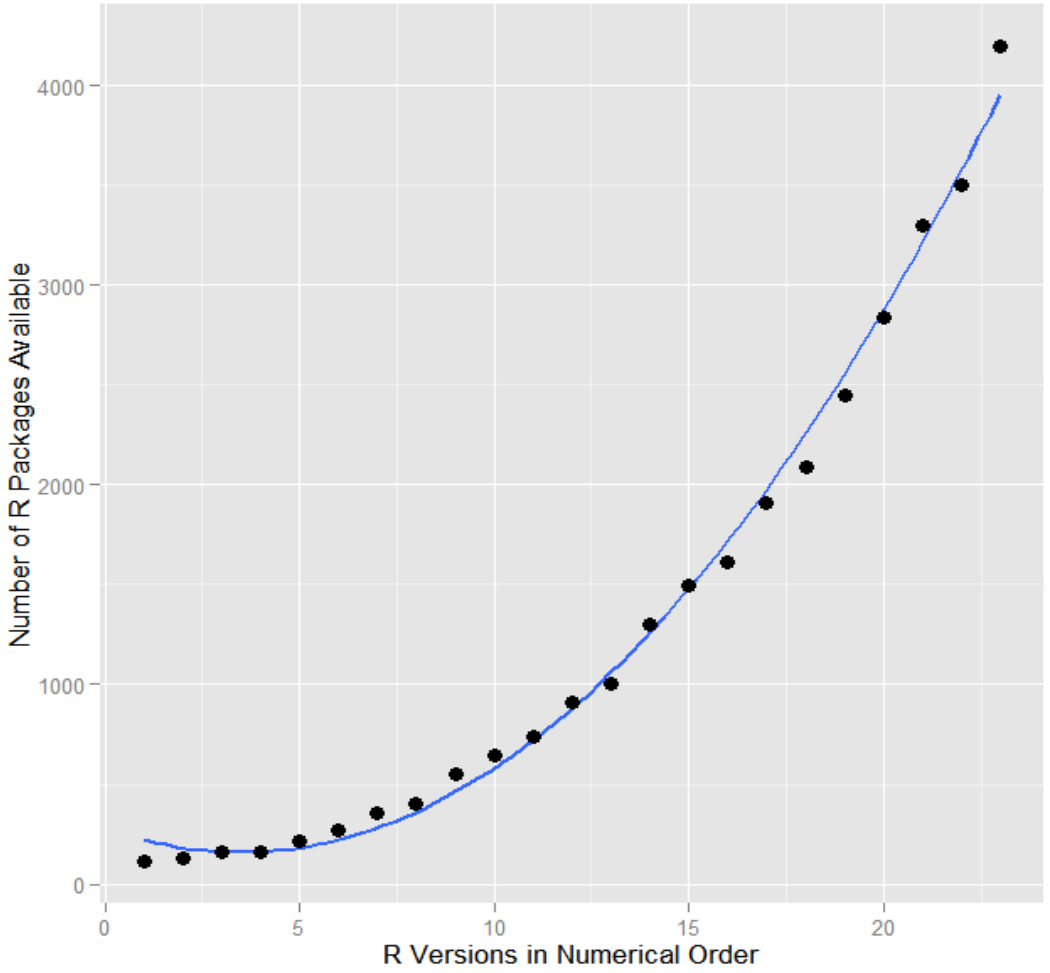
16 40 KANSALLIS KIRJASTO

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**

Number of open analysis tools has grown exponentially



github SOCIAL CODING

R

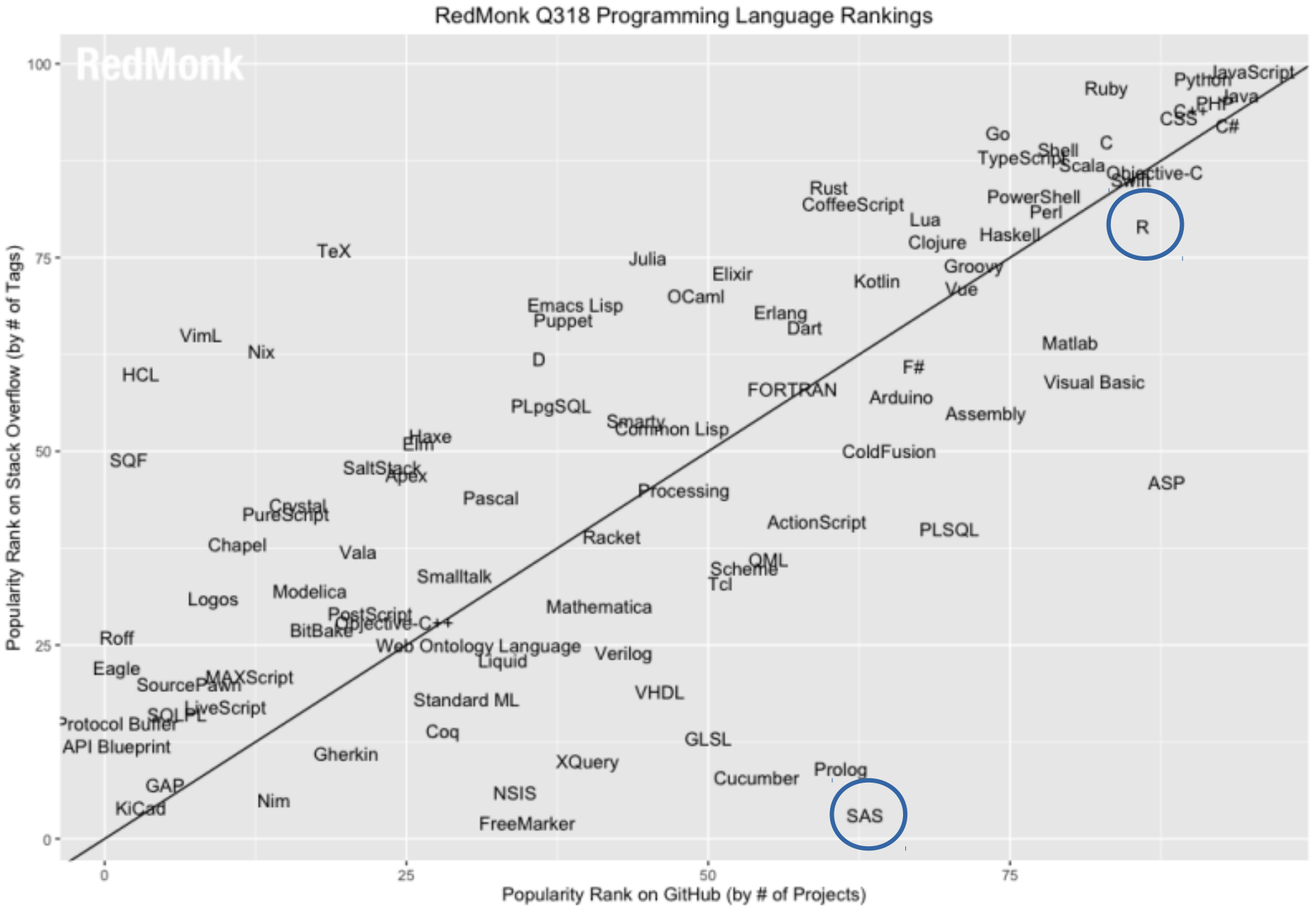
rOpenGov

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ROpenSci

Value of data can increase through sharing & use

Varying cultures of open collaboration



Elements of bibliographic data science

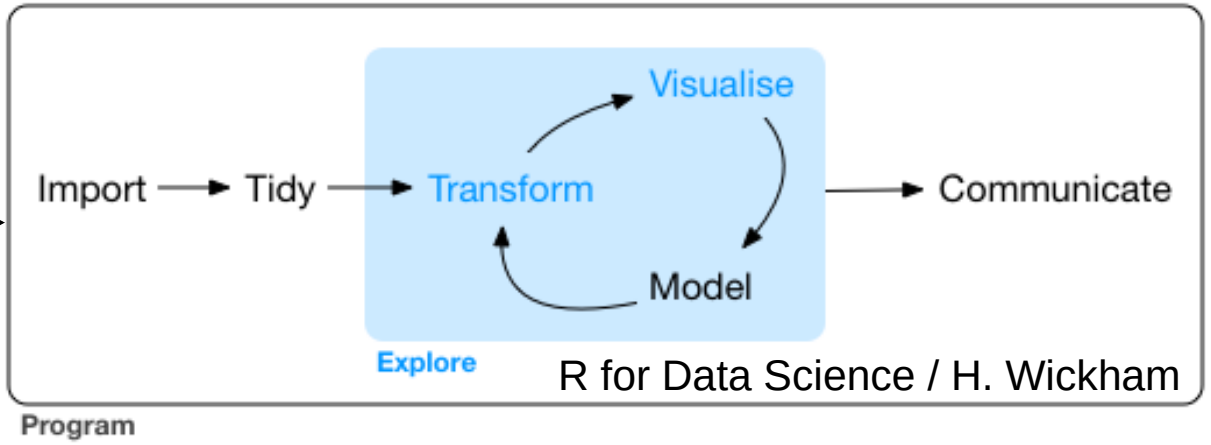
Quantitative frame for qualitative research
Beyond counting titles (volume, time, geography…)

Bibliographic metadata

Full texts: books, newspapers...

Supporting data

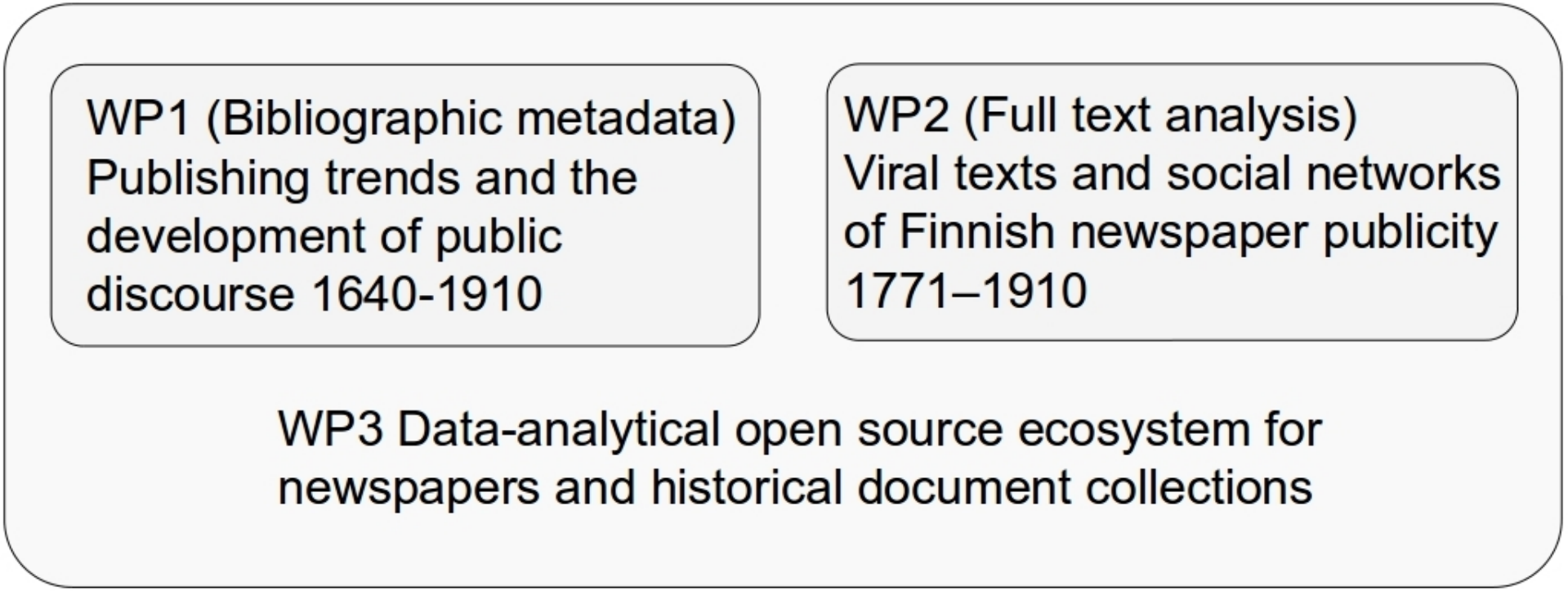
Open data science ecosystem



Academy of Finland COMHIS consortium 2016-2019

University of Helsinki, University of Turku, National Library of Finland

Quantify early-modern knowledge production with massive bibliographic collections, full texts, and open data analytical infrastructure



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

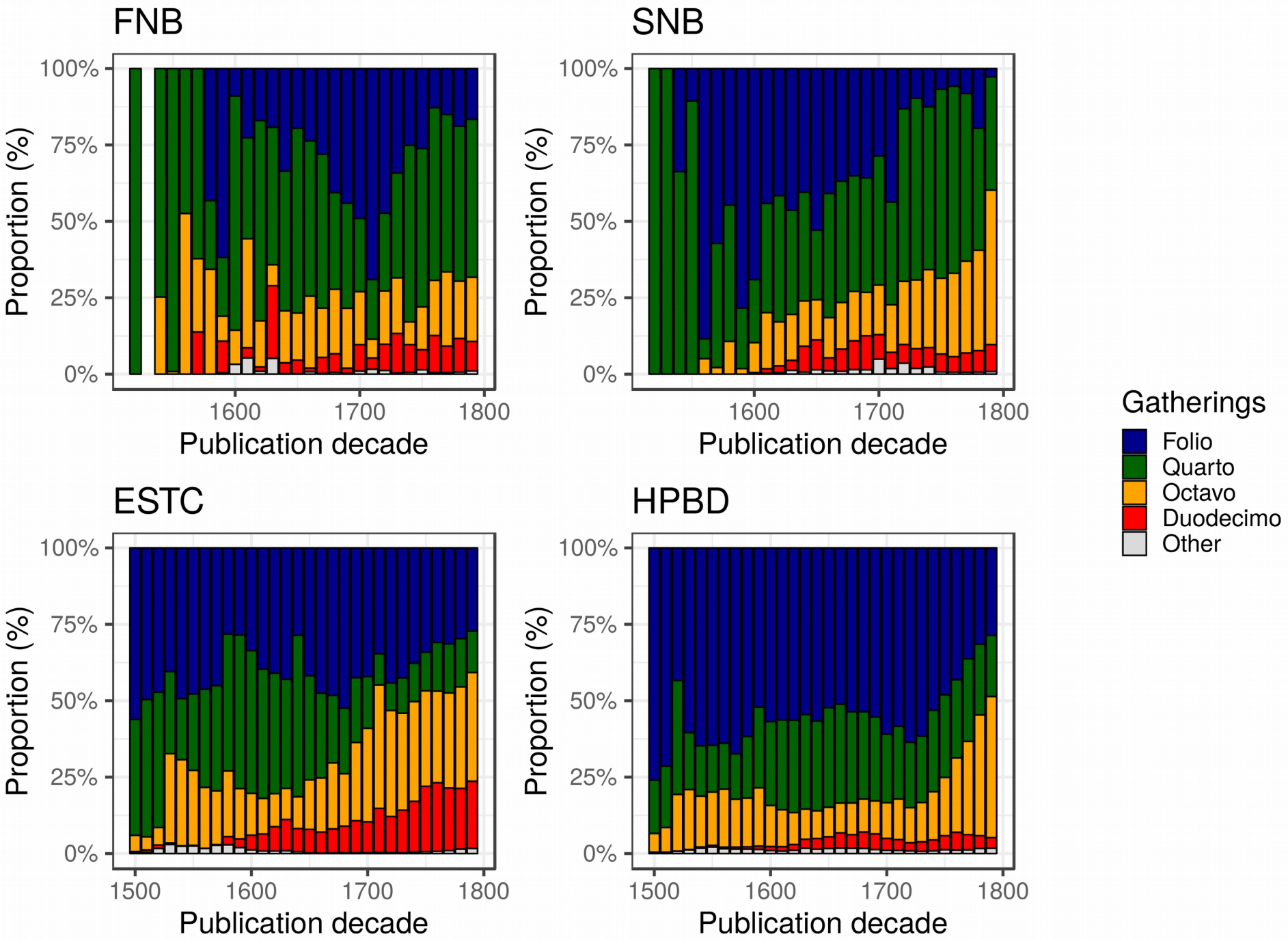
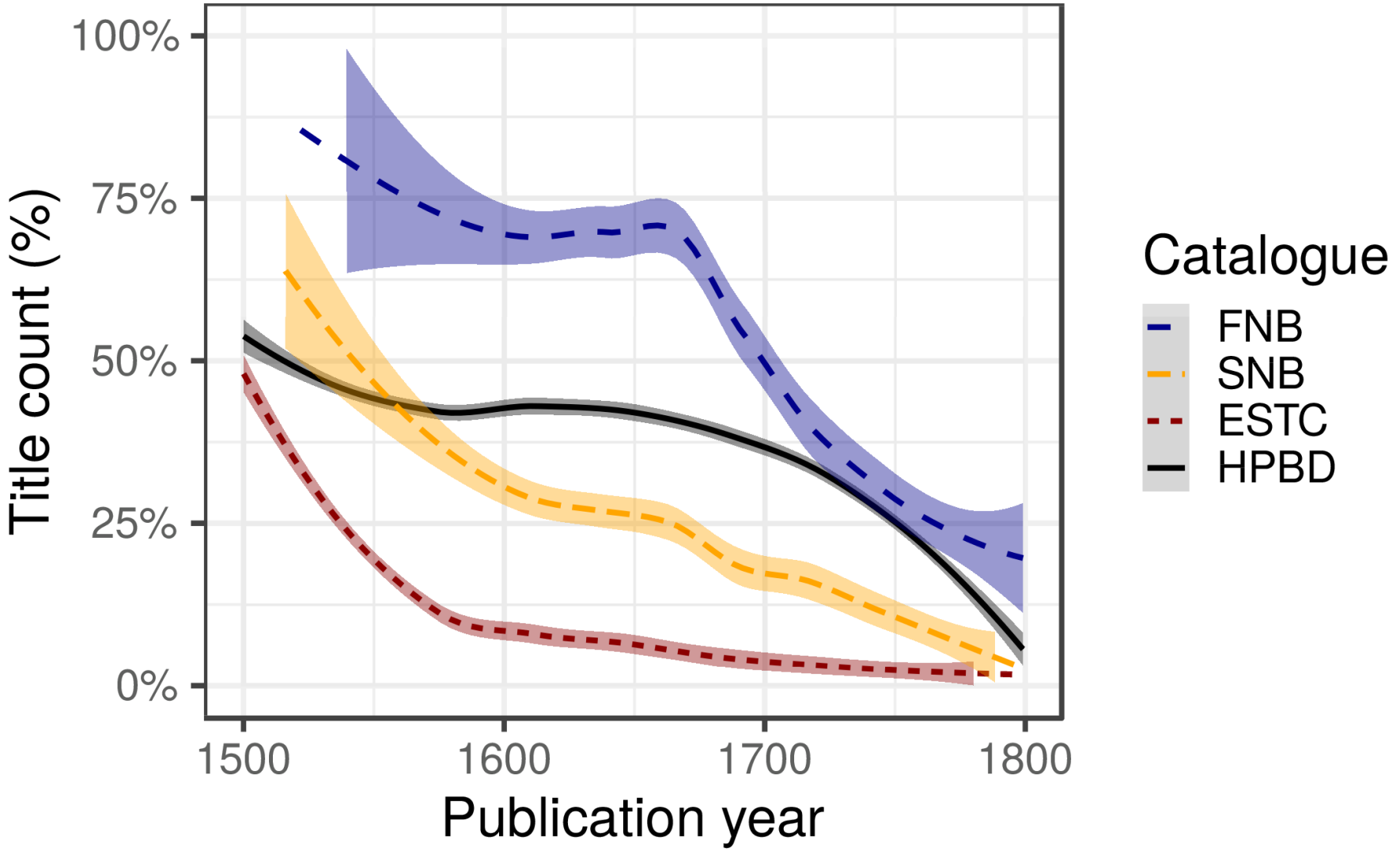
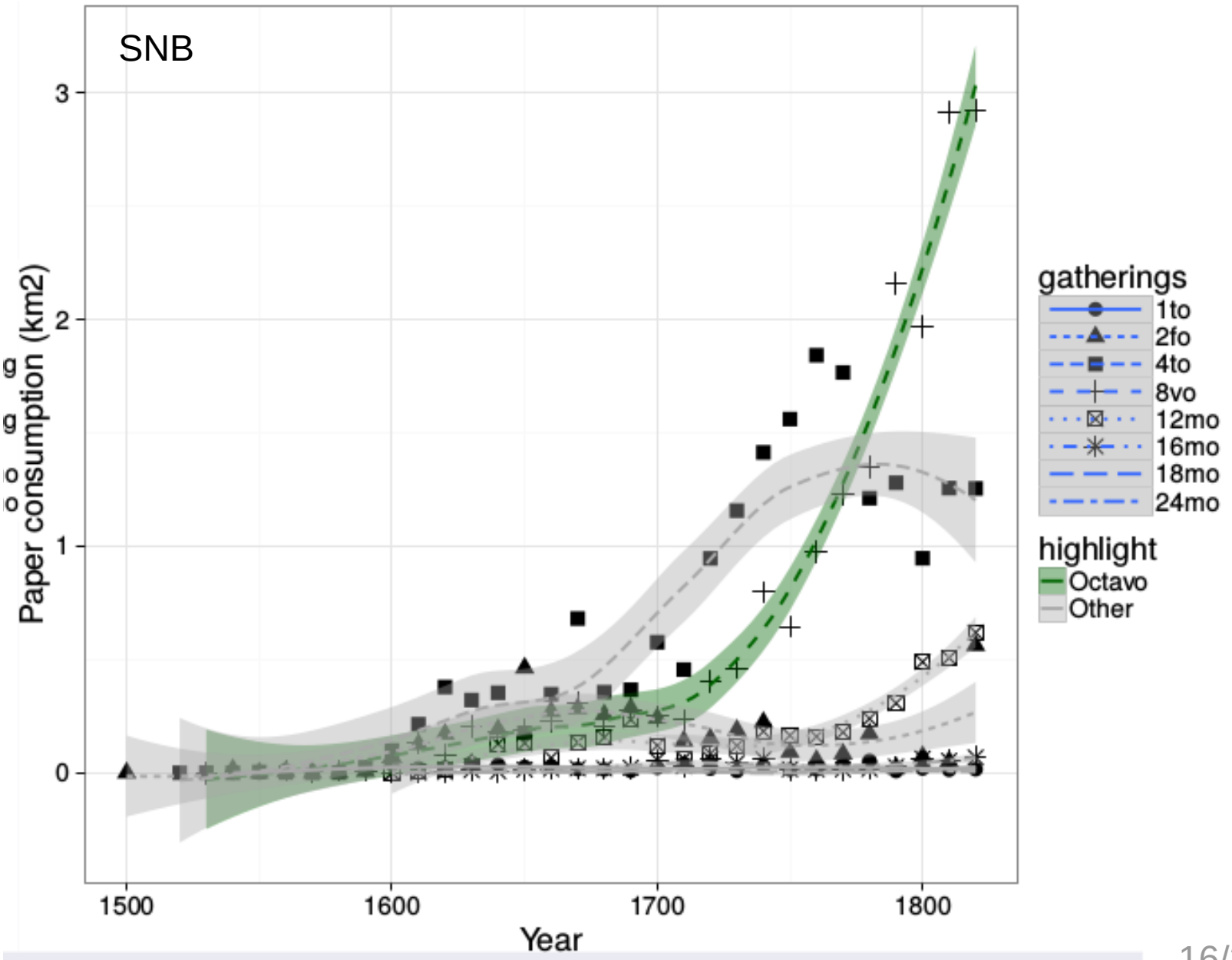


Fig. 1: Annual relative print area for common book formats.

Title count share for books in Latin (primary language)



The rise of Octavo: paper consumption



Print area for top languages

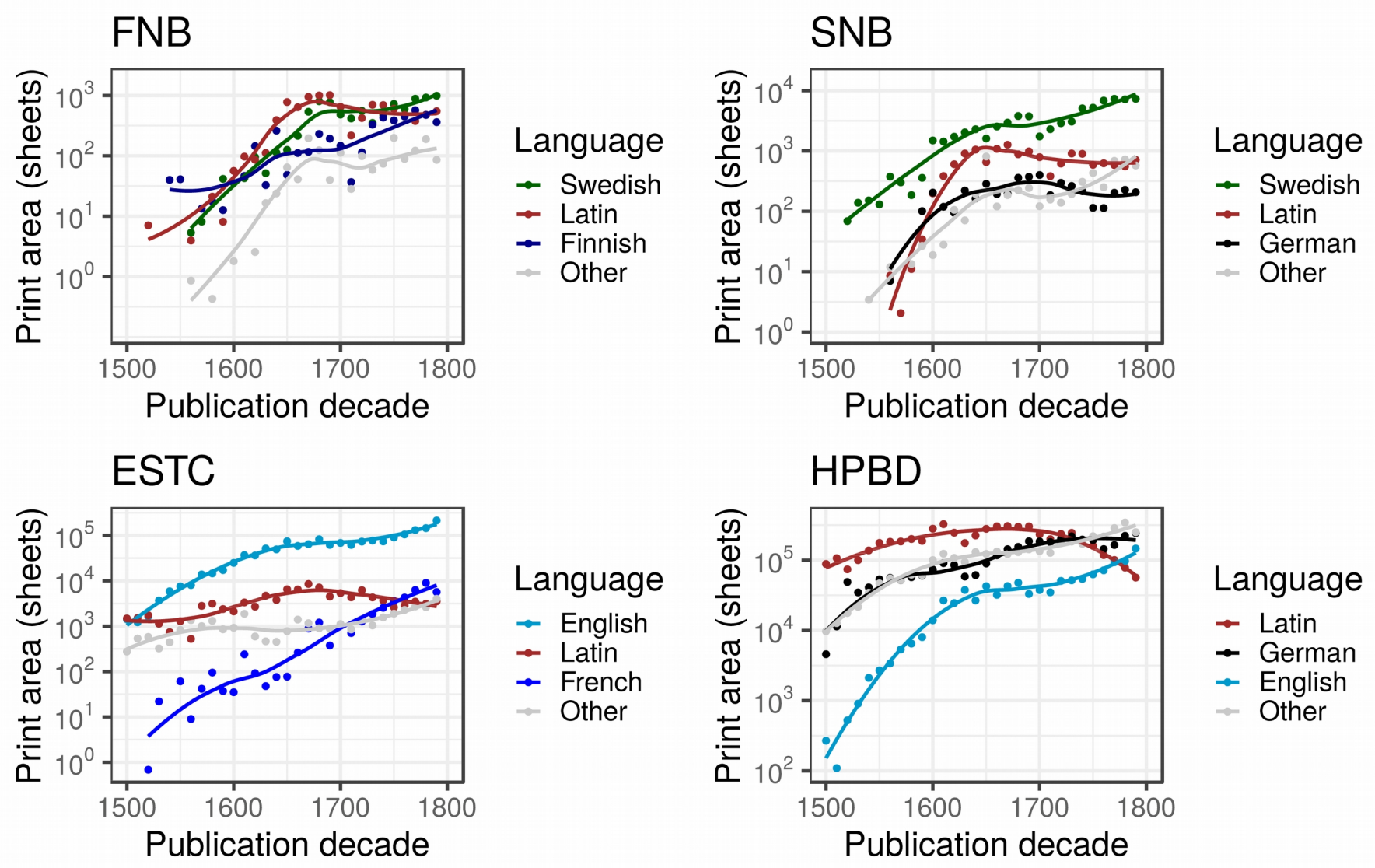
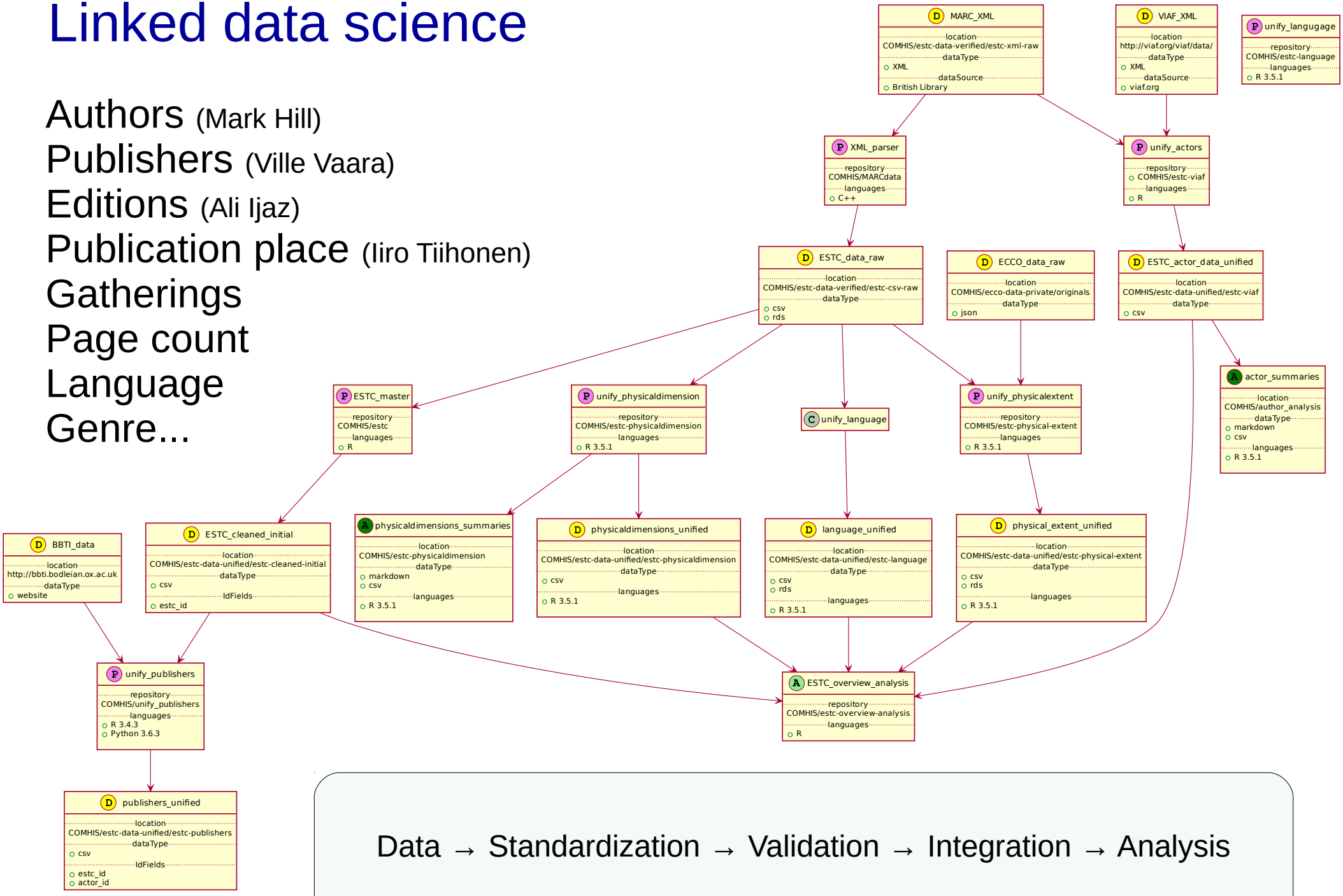


Fig 4. Changes in print area over time. The most common language from each catalogue are included.

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

Preprocess & Enrich

Clean up messy entries

```
polish_physical_extent("iii-xxiv, 118, [2] p.")$pagecount
```

```
## [1] 142
```

Enrich data (geocoordinates, gender, ..)

```
get_country("Porvoo")
```

```
## [1] "Finland"
```

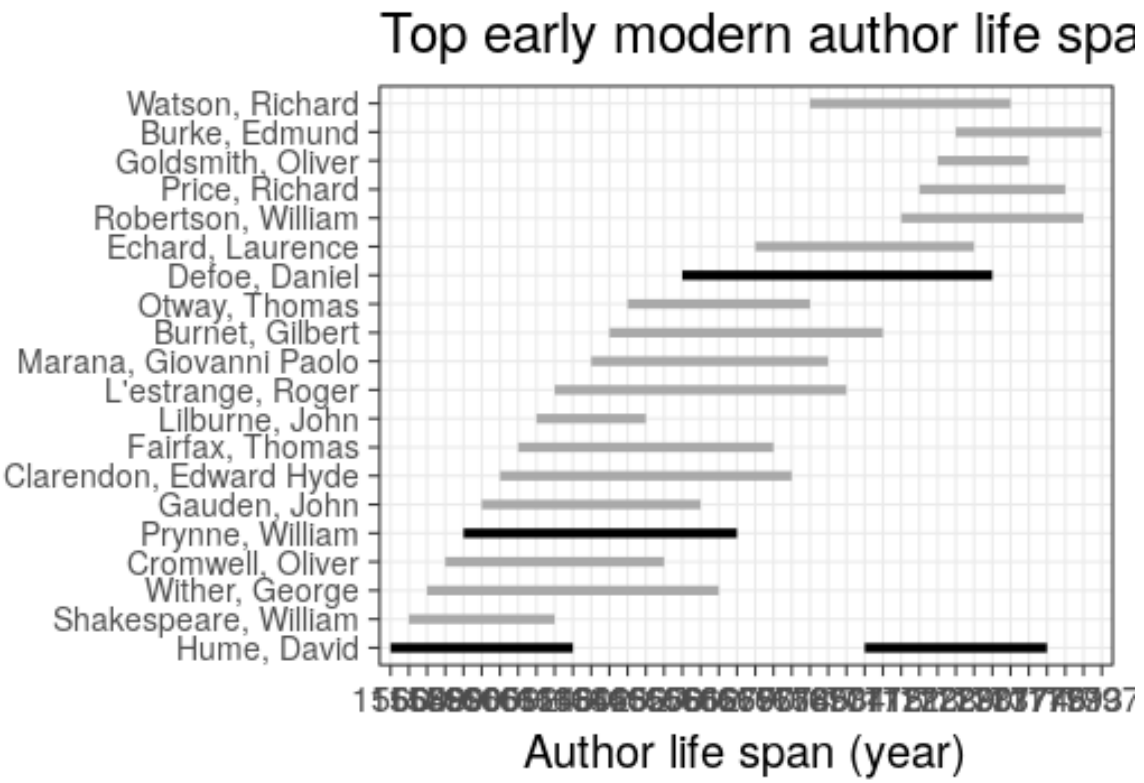
- Parse, clean up, enrich, summarise, analyze, visualize, report..

Automated summaries for the unified data

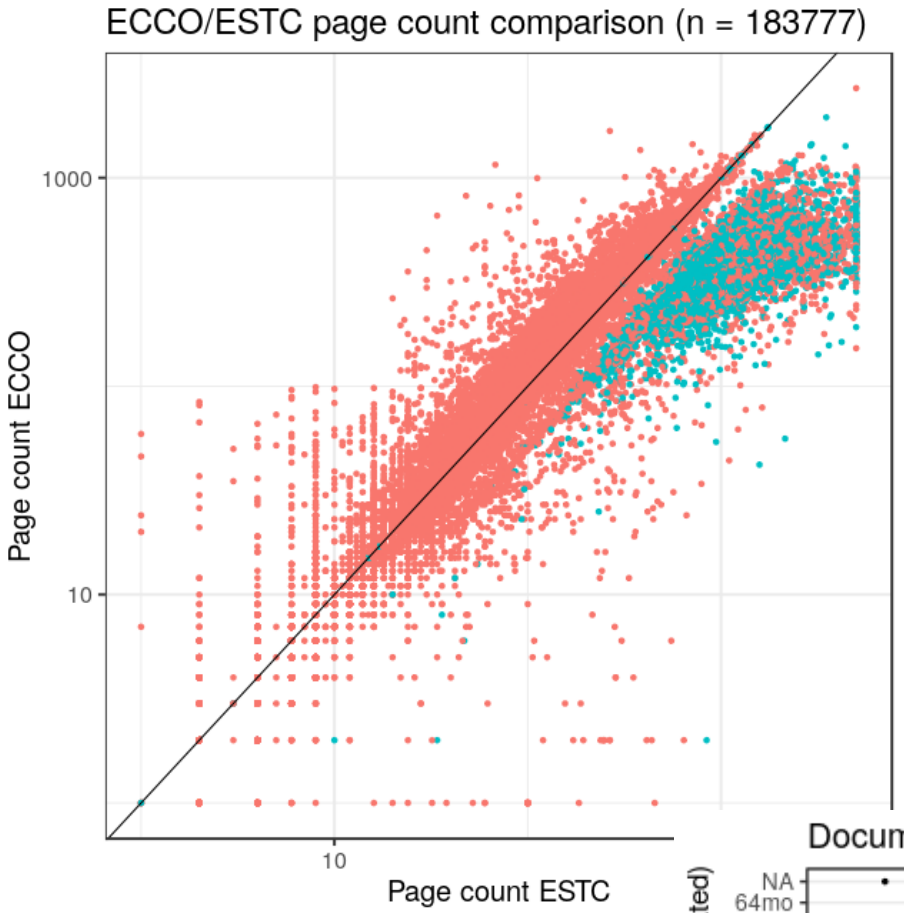
The data spanning years 1488-1955 has been included and contains 70451 documents on the data collection, see the source code for details.

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 (ECCO vs. ESTC)



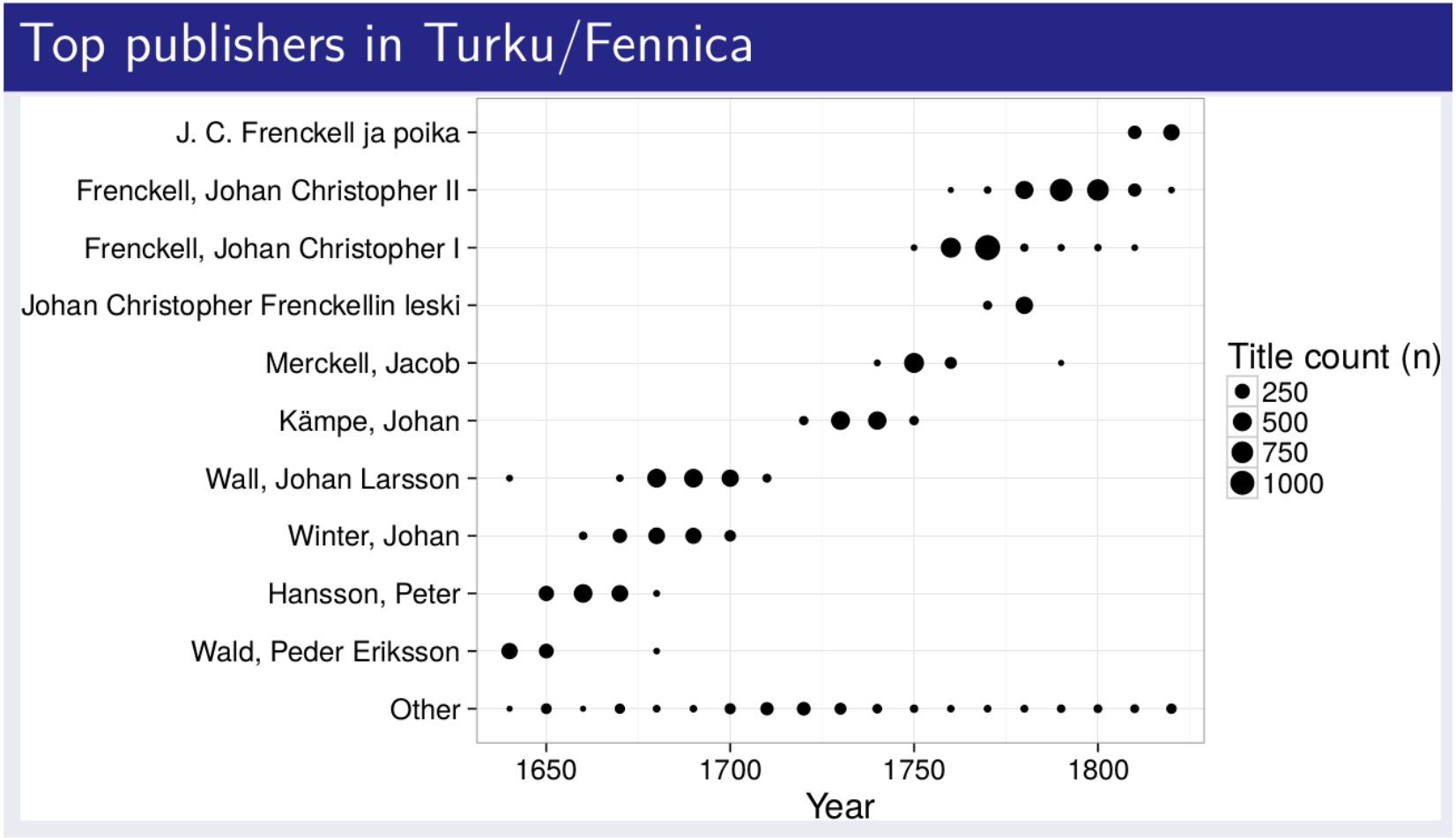
Clean up messy entries

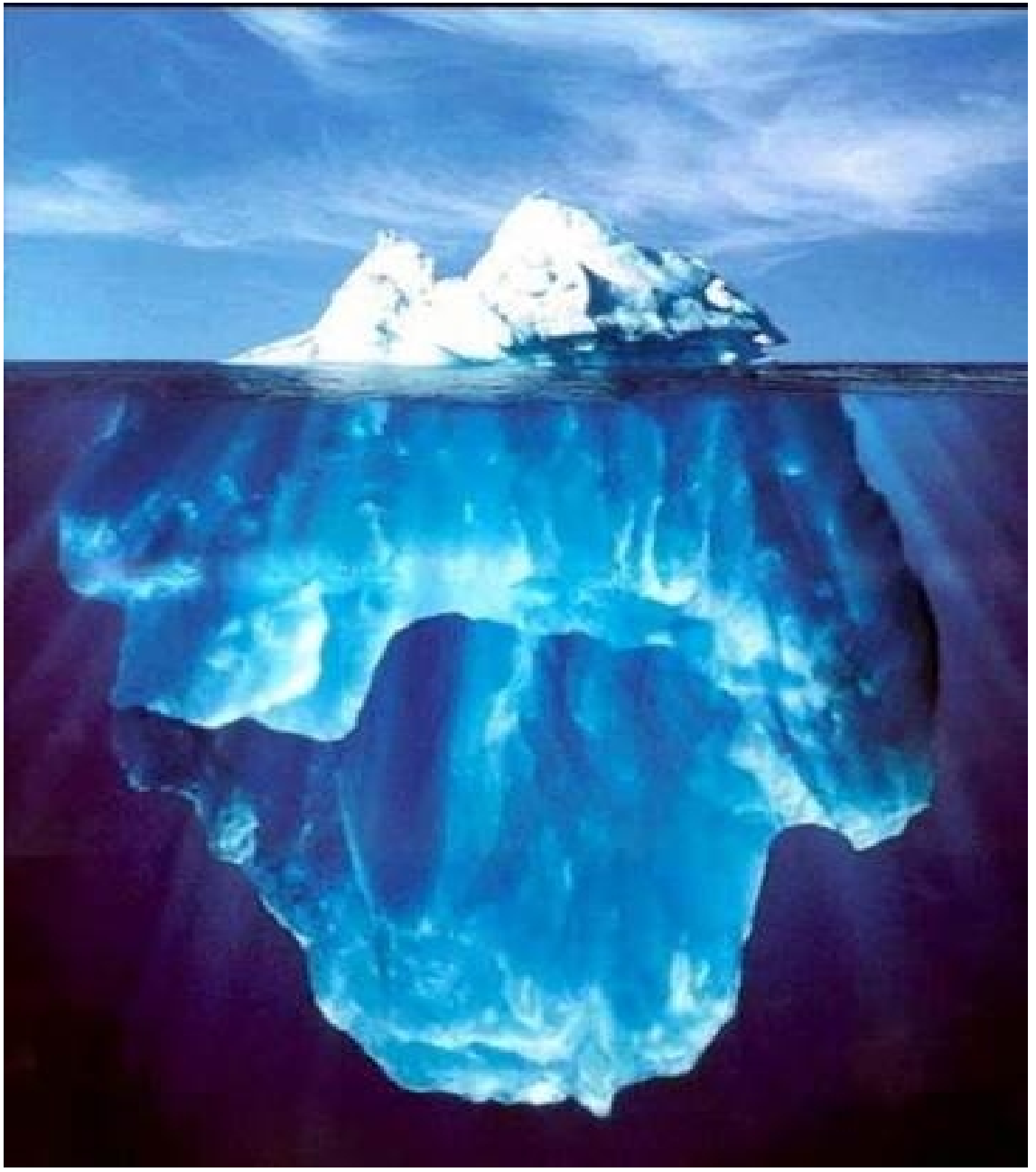
```
polish_physical_extent("iii-xxiv, 118, [2] p.")$
```

```
## [1] 142
```

Library metadata as research object

Author & publisher networks
Spatio-temporal dynamics
History, philosophy, religion, science..





Data
Analysis

Data
Preprocessing

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

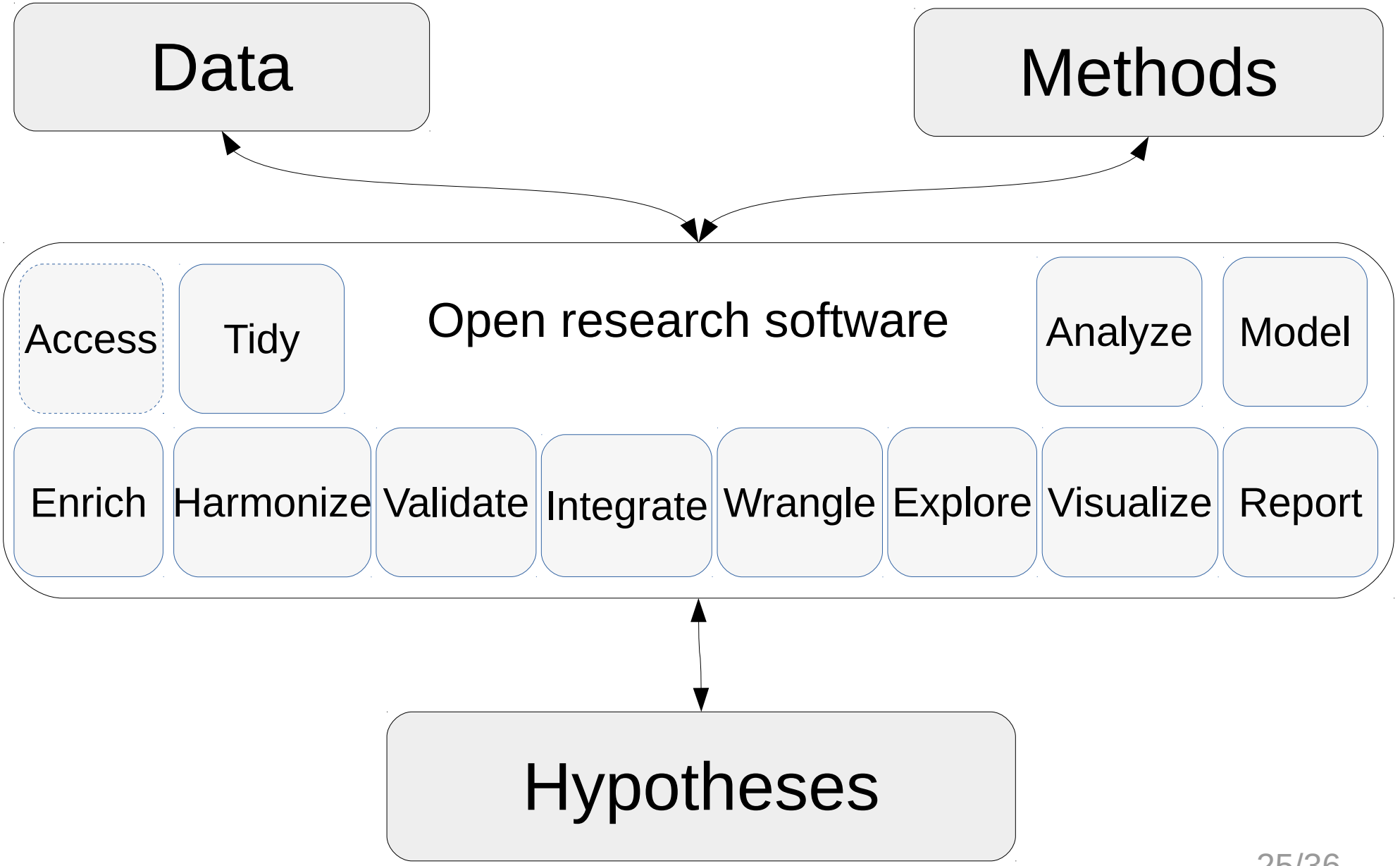
The data is summarized in the following automatically generated files:

- 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

Open methods development can complement FAIR data sharing

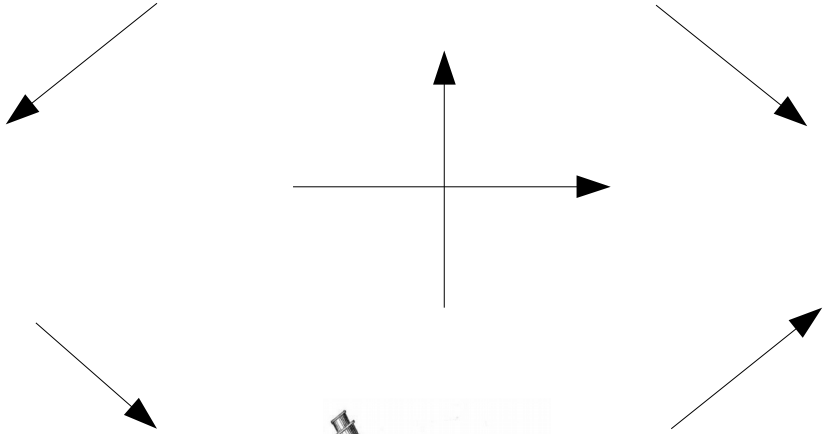


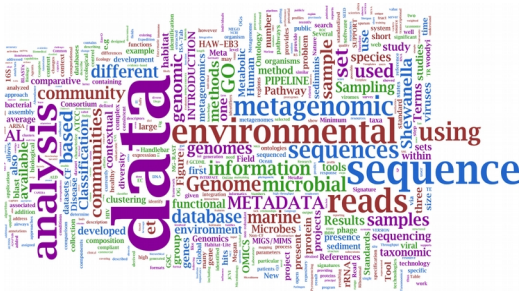
Hypothesis testing vs. hypothesis discovery?

?

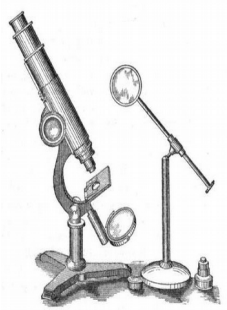
Hypothesis

Method





Data



Tools

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

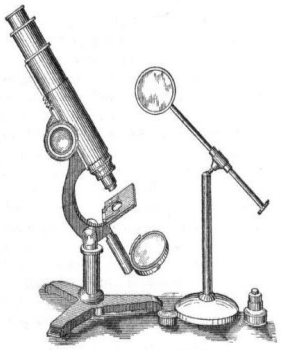
Shining Light into Black Boxes

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

From Data to Knowledge

Wisdom

Knowledge,
understanding



Information,
facts

Observations,
Data

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

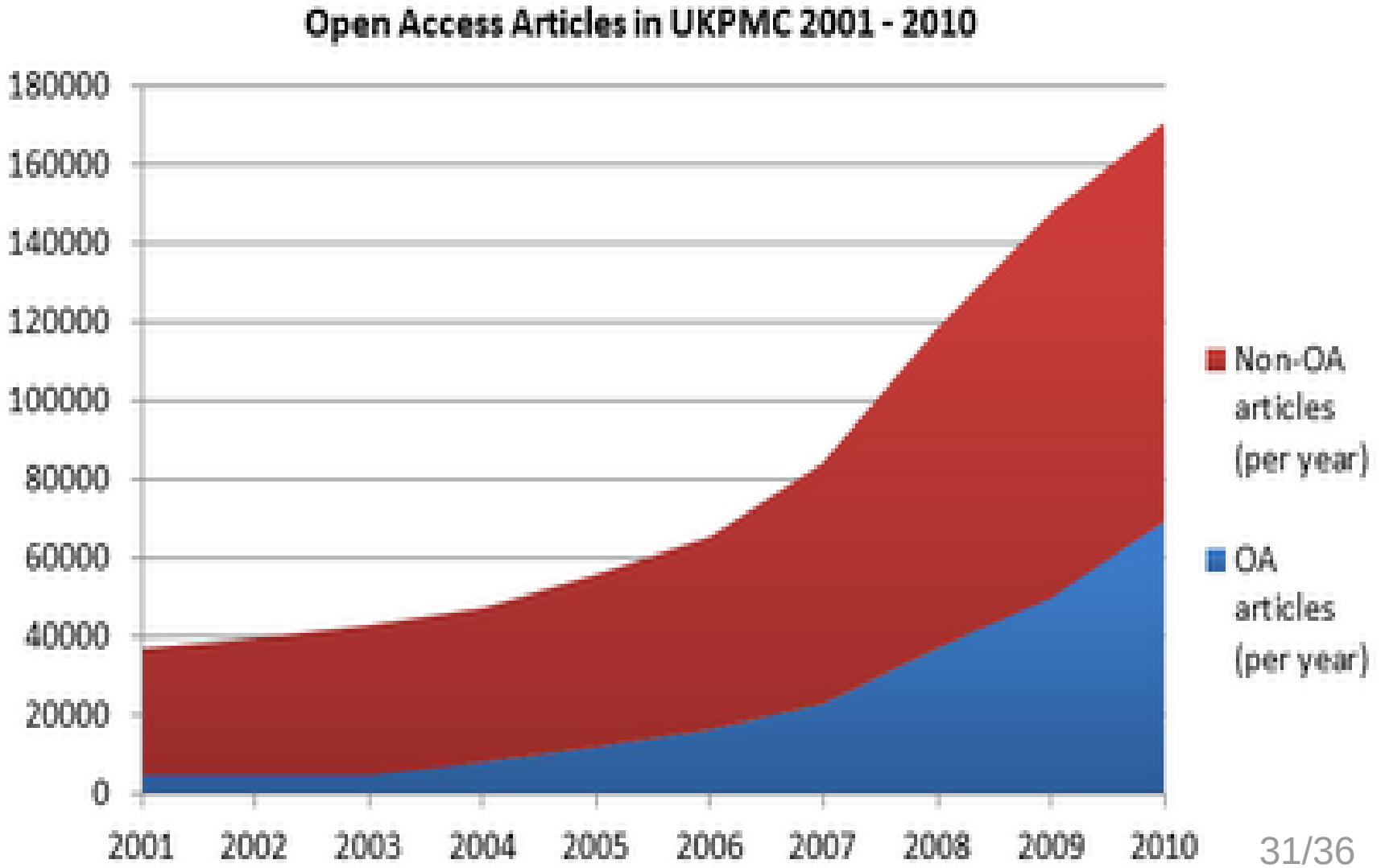
Heidi Laine Leo Lahti Anne Lehto

Key modes of academic publishing

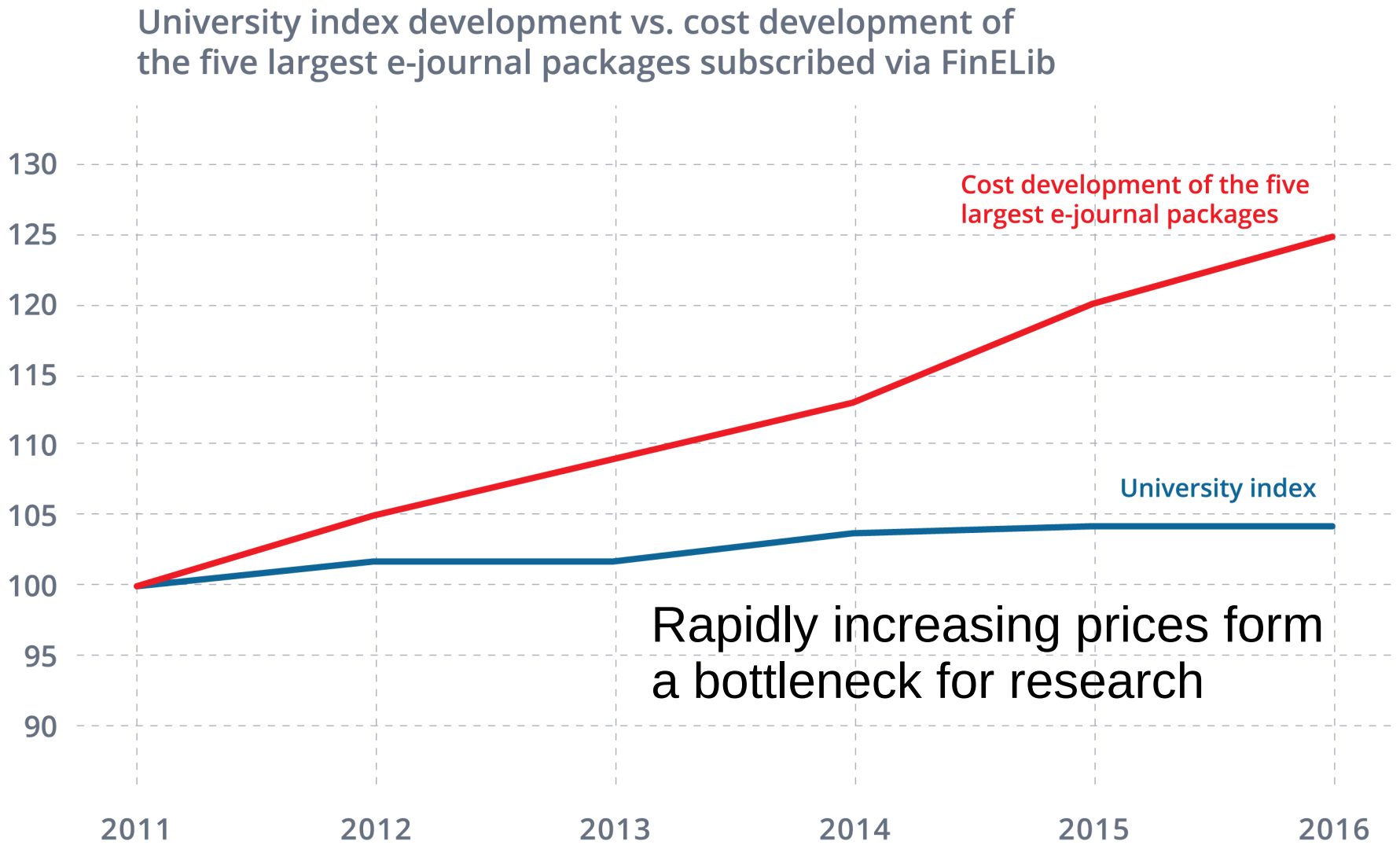
	Subscription	Open access
Novelty	Traditional	Rapidly growing
Access	Behind paywall	Open for ‘all’
Licensing	Proprietary licenses	Open access licenses
Who pays ?	**Library (or reader) pays**	**Researcher pays**
Pricing information	**Confidential pricing**	Open pricing
Publishing costs for a researcher	‘Free’ to publish	‘Expensive’ APCs (0-5000e/pub)

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).

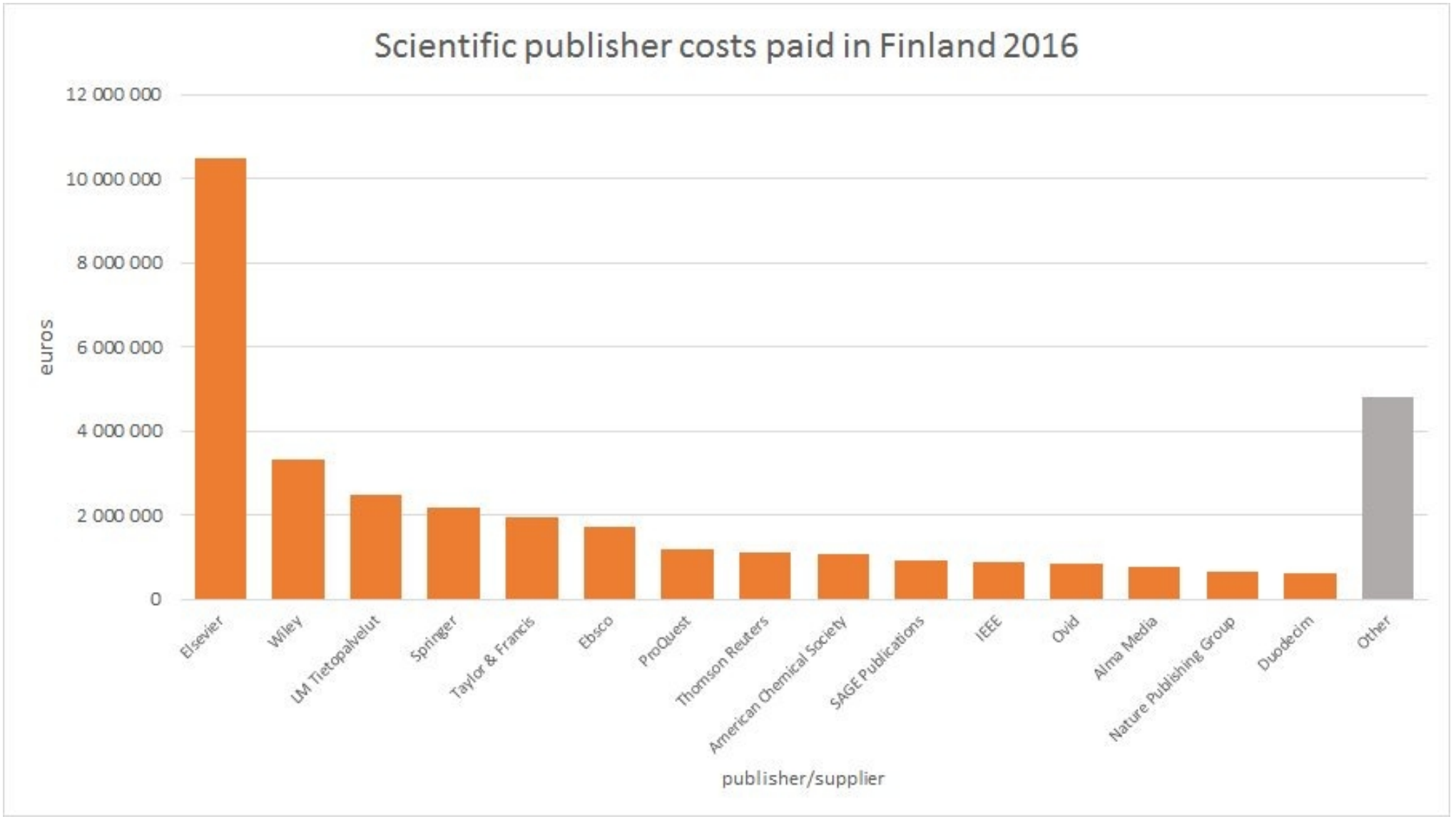


Costs for the top-5 journal packages increased 25% while university funding nearly unaltered (Finland 2011-2016)



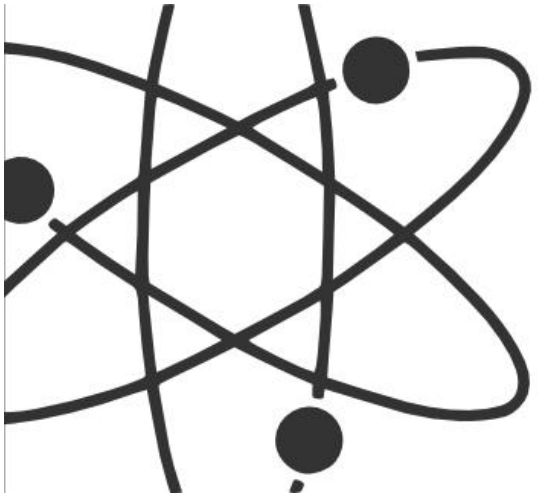
Source: https://www.kiwi.fi/pages/viewpage.action?pageId=64487647#What%27sgoingonwithscholarlyjournalnegotiations?-Accesstoscholarlyjournals%E2%80%93atwhatcost?(19.9.2016)

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



THE COST OF SCIENTIFIC PUBLICATIONS MUST NOT GET OUT OF HAND

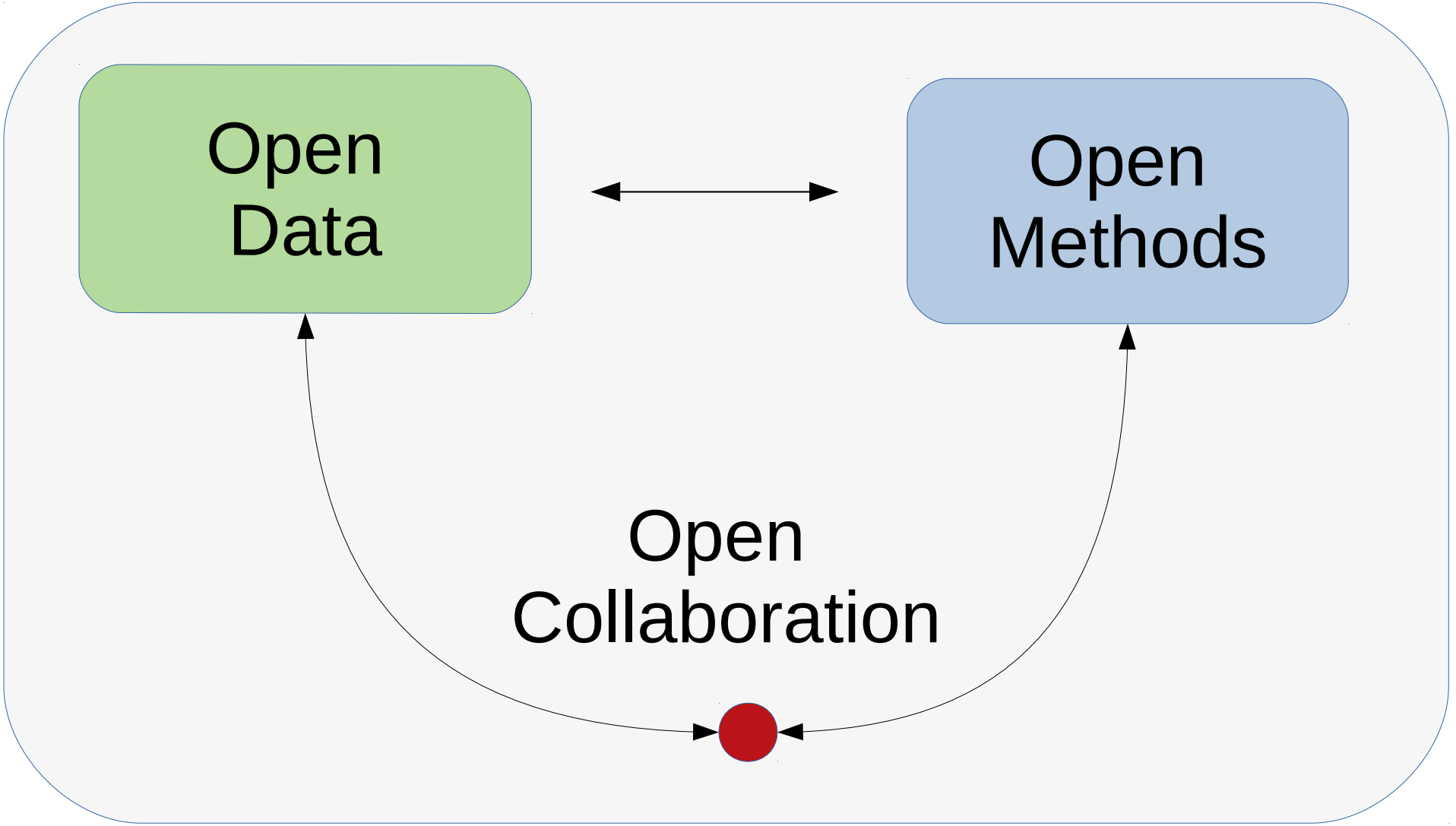
Heidi Laine



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.

Elements of open data science



L Lahti. Advances in Intelligent Data Analysis XVII. Springer, 2018.

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