

Probabilistic programming for Bibliographic Data Science

DARIAH

4 November 2020

Leo Lahti (University of Turku, Finland)

datascience.utu.fi



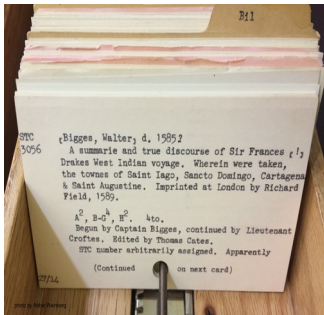
Turun yliopisto
University of Turku



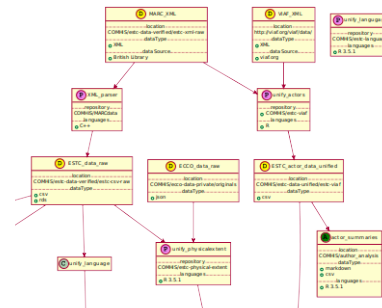
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

From library catalogues to research reports?

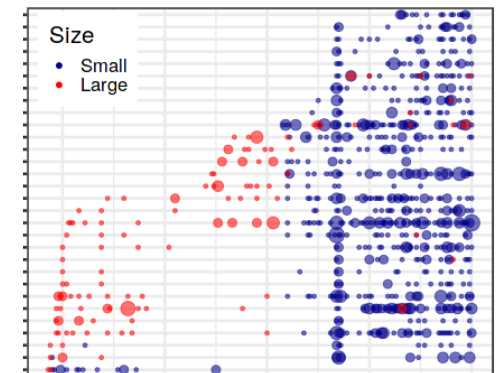
Research potential



Open bibliographic data science ecosystem



Research cases

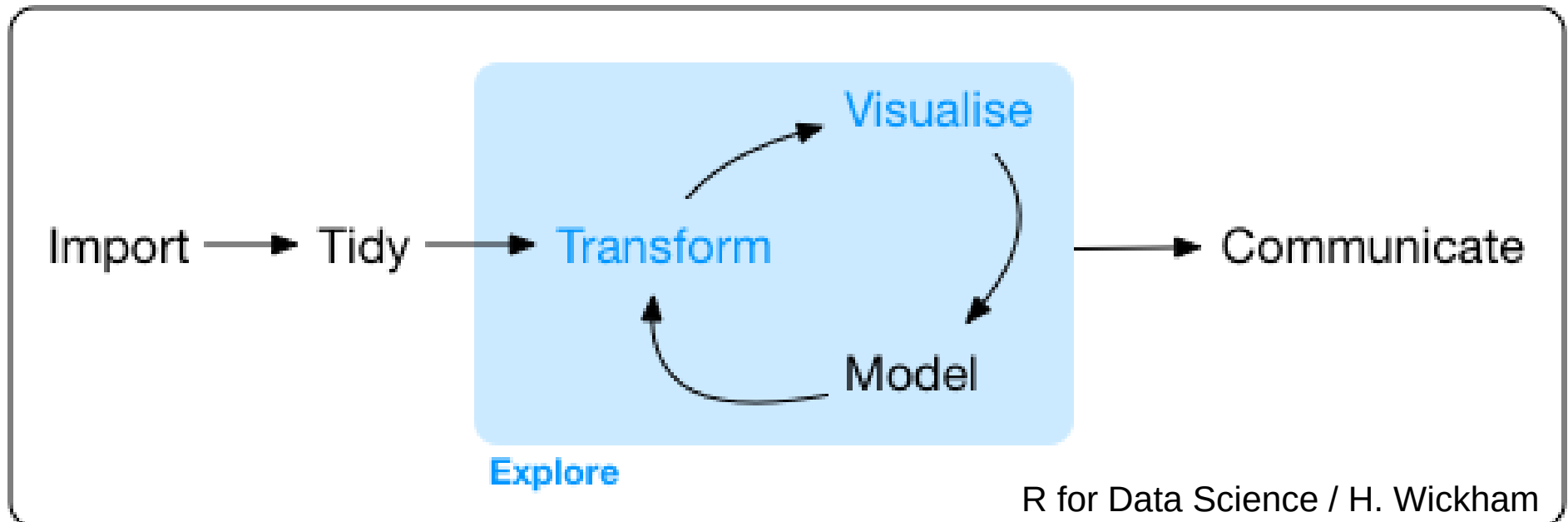


Computational workflows: challenges & opportunities



- 1) data access
- 2) **data analysis**
- 3) data communication

IP[y]:
IPython



R for Data Science / H. Wickham

Probabilistic programming

Implement (Bayesian) probabilistic models

Infer model parameters (and structure)

Criticize & revise the models

Bibliographic data

Catalogue:

English Short Title

Catalogue (ESTC; BL)

Hundreds of thousands of
works catalogued

1701-1800

Bibliographic data

Catalogue:

English Short Title

Catalogue (ESTC; BL)

Hundreds of thousands of
works catalogued

1701-1800



Full texts:

**Eighteenth Century
Collections Online (ECCO;
Gale)**

Same works as full texts

Bibliographic data

Catalogue:

English Short Title

Catalogue (ESTC; BL)

Hundreds of thousands of
works catalogued

1701-1800



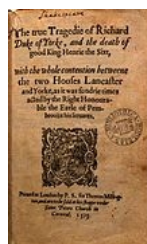
Full texts:

**Eighteenth Century
Collections Online (ECCO;
Gale)**

Same works as full texts



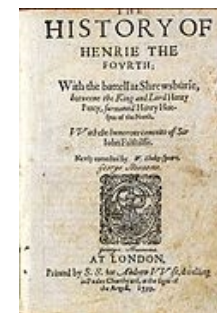
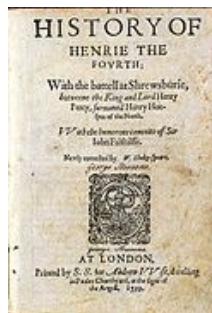
Losses and biases?
~60% coverage



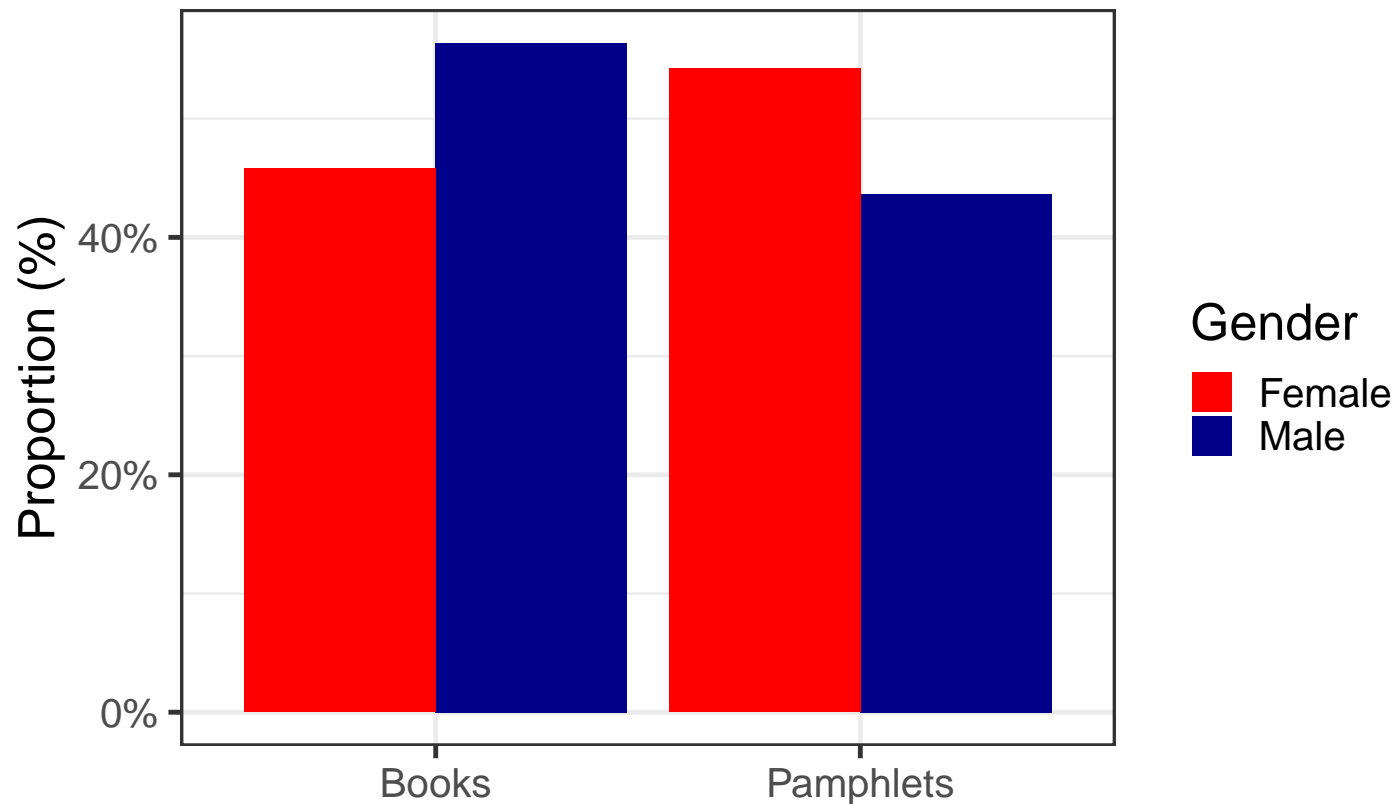
Coin tossing experiment? effect of gender, genre, time..



Losses and biases?
~60% coverage



Women write more pamphlets & pamphlets have a lower coverage



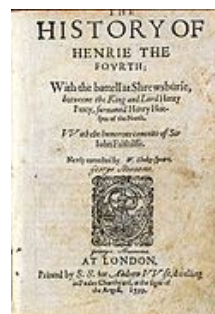
Binomial probability model

$$k \sim \text{Bin}(n, p)$$

$$p \sim \text{logit}^{-1}(\beta_0 + \sum_{i=1}^J \beta_i x_i)$$



Losses and biases?
~60% coverage



Bayes theorem

Posterior

Likelihood

Prior

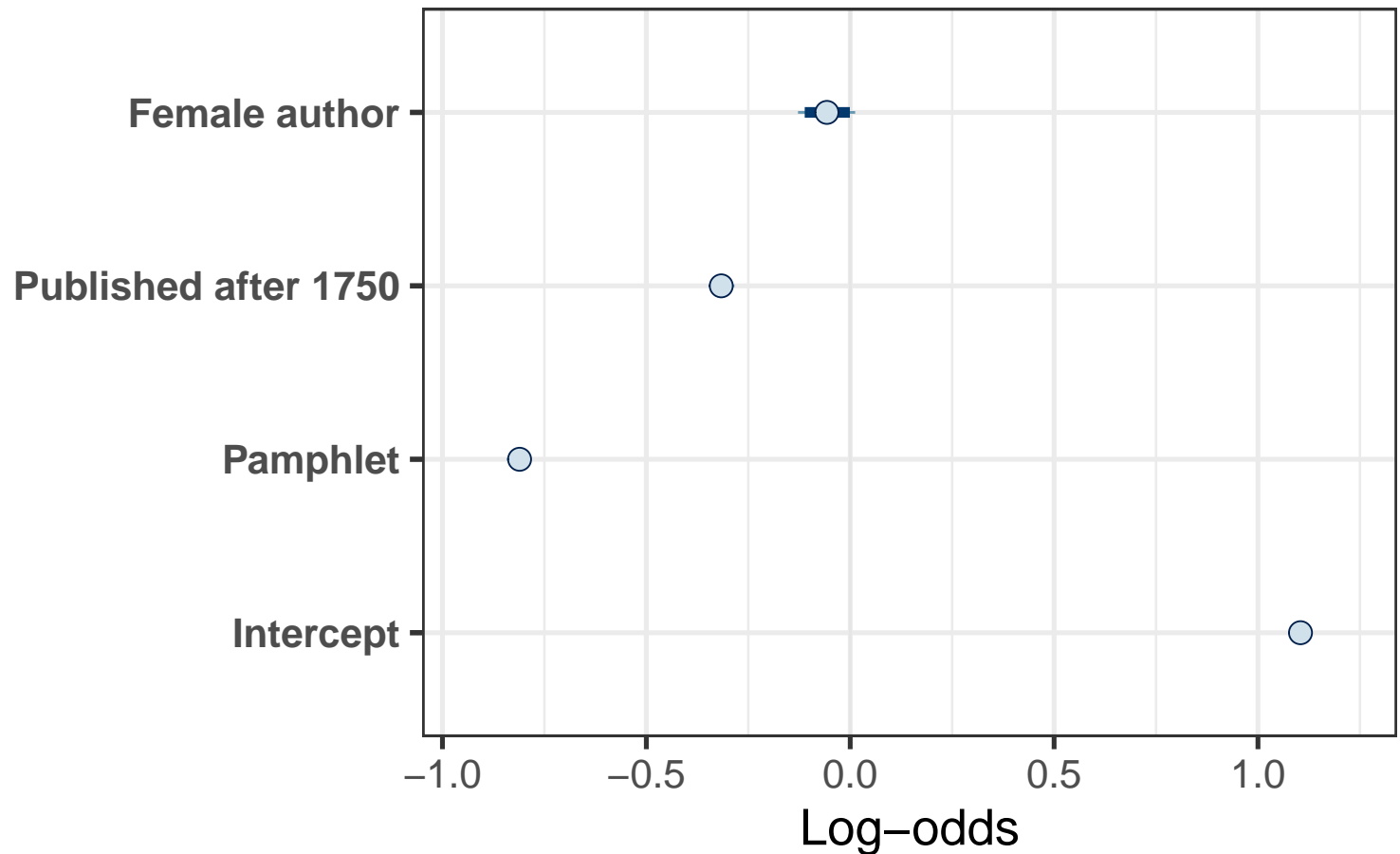
$$P(\beta|x) \sim \frac{P(x|\beta)P(\beta)}{P(x)}$$

Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming

Leo Lahti^a, Eetu Mäkelä^b and Mikko Tolonen^b

^aUniversity of Turku, Turku, Finland

^bUniversity of Helsinki, Helsinki, Finland



CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

✉ leo.lahti@utu.fi (L. Lahti)

🌐 <http://www.iki.fi/Leo.Lahti> (L. Lahti)

📄 0000-0001-5537-637X (L. Lahti); 0000-0002-8366-8414 (E. Mäkelä); 0000-0003-2892-8911 (M. Tolonen)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

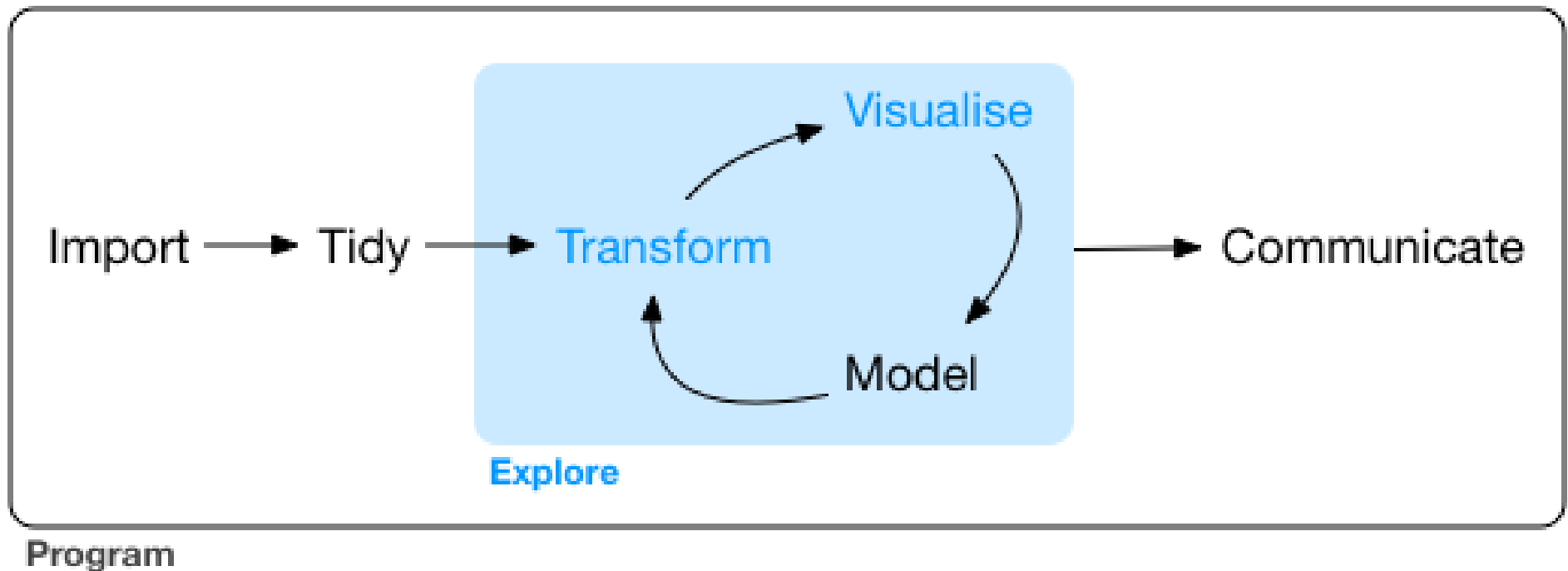
CEUR Workshop Proceedings (CEUR-WS.org)

CHR2020 ONLINE

ONLINE WORKSHOP ON COMPUTATIONAL HUMANITIES RESEARCH
(18–20 NOVEMBER 2020);

<http://ceur-ws.org/Vol-2723/short46.pdf>

From tools towards statistical understanding: enhancing the overall value of data





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 

^aDepartment of Mathematics and Statistics, University of Turku, Finland; ^bHelsinki 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

Thanks!



Turku Data Science Group (Leo Lahti):
datascience.utu.fi

Helsinki Computational History Group (Mikko Tolonen):
helsinki.fi/en/researchgroups/computational-history