Building Scientific Publication Profiles for U.S.-trained Doctorate Recipients

Haoyi Wei[1], Eric Livingston[2], Christina Freyman[3], and Wan-Ying Chang[1]

[1]National Center for Science and Engineering Statistics, National Science Foundation

[2]Elsevier

[3]Evaluation and Assessment Capability Section, National Science Foundation

Disclaimer

Working papers are intended to report ***exploratory*** results of research and analysis undertaken by the National Center for Science and Engineering Statistics at the National Science Foundation (NSF). Any opinions, findings, conclusions, or recommendations expressed in this work do not necessarily reflect the views of NSF. This work is being presented to inform interested parties of ongoing research or activities and to encourage further discussion of the topic.

Brief Summary

Research Goal:

Identify the Scopus publication records for U.S.-trained doctorate recipients in science, engineering, and health fields.

Data:

Clarivate's Survey of Doctorate Recipients (SDR) - Web of Science (WoS) linkage

Elsevier's SDR – Scopus linkage.

Method:

Machine Learning

Results:

Finalized SDR-Scopus linkage.

Contribution:

This work is an important contribution to the development of a U.S. Science & Engineering Enterprise Data Network.

Motivation

- Why do we build publication profiles?
  - To advance understanding of scientific research and the impact made by individual researchers.
  - To investigate the relationships between scientific productivity and various author attributes.
  - To access the nation’s strategic investment in doctoral training to inform science policies.

Citation Bias?

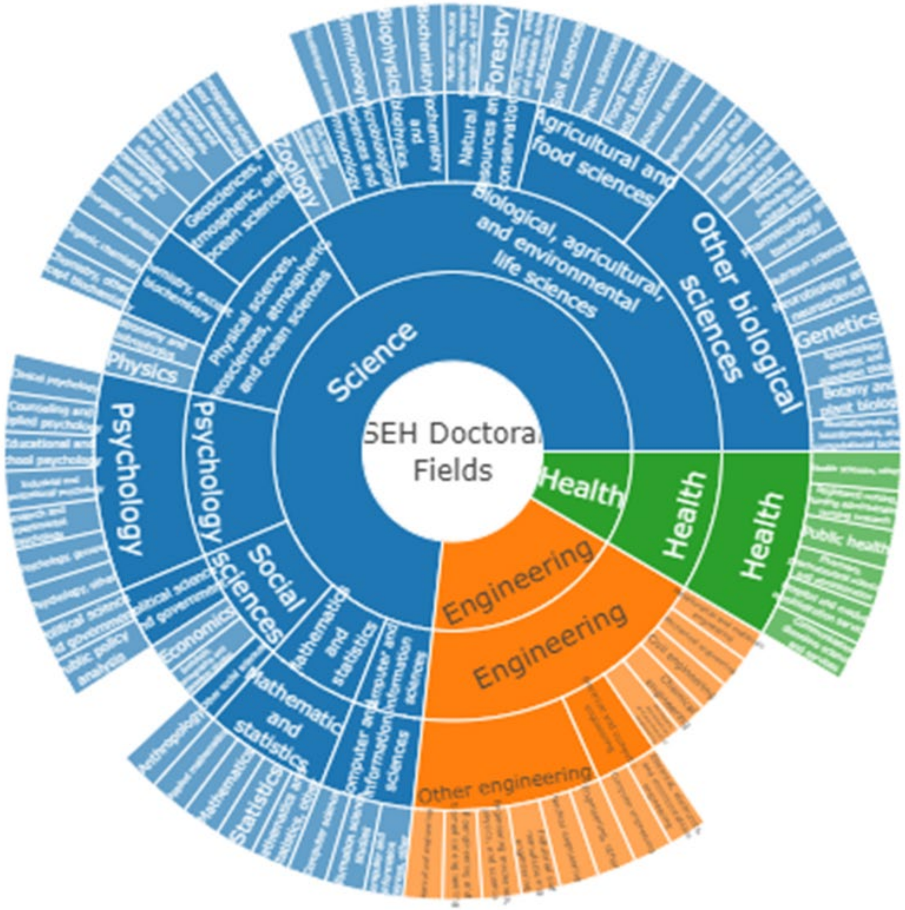
Credit: Biomedical Odyssey

Data

Surveys and bibliometric data

Surveys on U.S.-Trained Doctoral Researchers

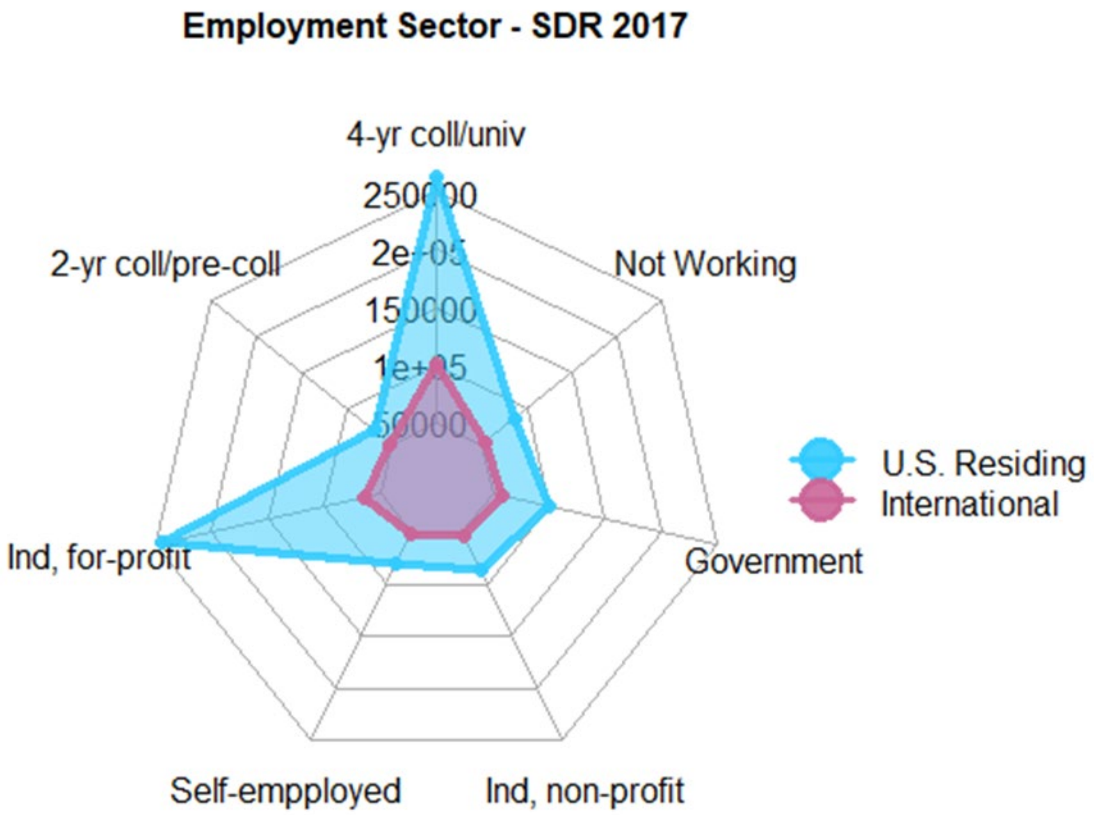
Survey of Earned Doctorates (SED): annual census of research PhDs



SOURCE: National Center for Science and Engineering Statistics, Survey of Earned Doctorates and Survey of Doctorate Recipients, 2017.

Surveys on U.S.-Trained Doctoral Researchers

Survey of Doctorate Recipients (SDR): biennial sample survey on science, engineering, and health doctorate degree holders



SOURCE: National Center for Science and Engineering Statistics, Survey of Earned Doctorates and Survey of Doctorate Recipients, 2017.

Longitudinal data from SED to SDR

Education

1st bachelor, master, research PhD, up to 5 degrees (year, institution, place, field of study); dissertation field, financial support, debt

Postgraduation Plans

Country/State intend to live, taking a postdoc, employment status/commitment, employer type, salary, work activities

Background

Sex, marital status, dependents, parent's educational attainment, birthplace, citizenship, race/ethnicity, disability

Employment Situation

Labor force status, reasons for not working, year retired, principal employer, faculty rank, tenure status, principal job, work activity, salary, benefits, job satisfaction, federal support

Past Employment

Other Work-Related Experiences

Recent Educational Experiences

Demographic Information

Spouse working status, living with children, residing location, citizenship/visa type

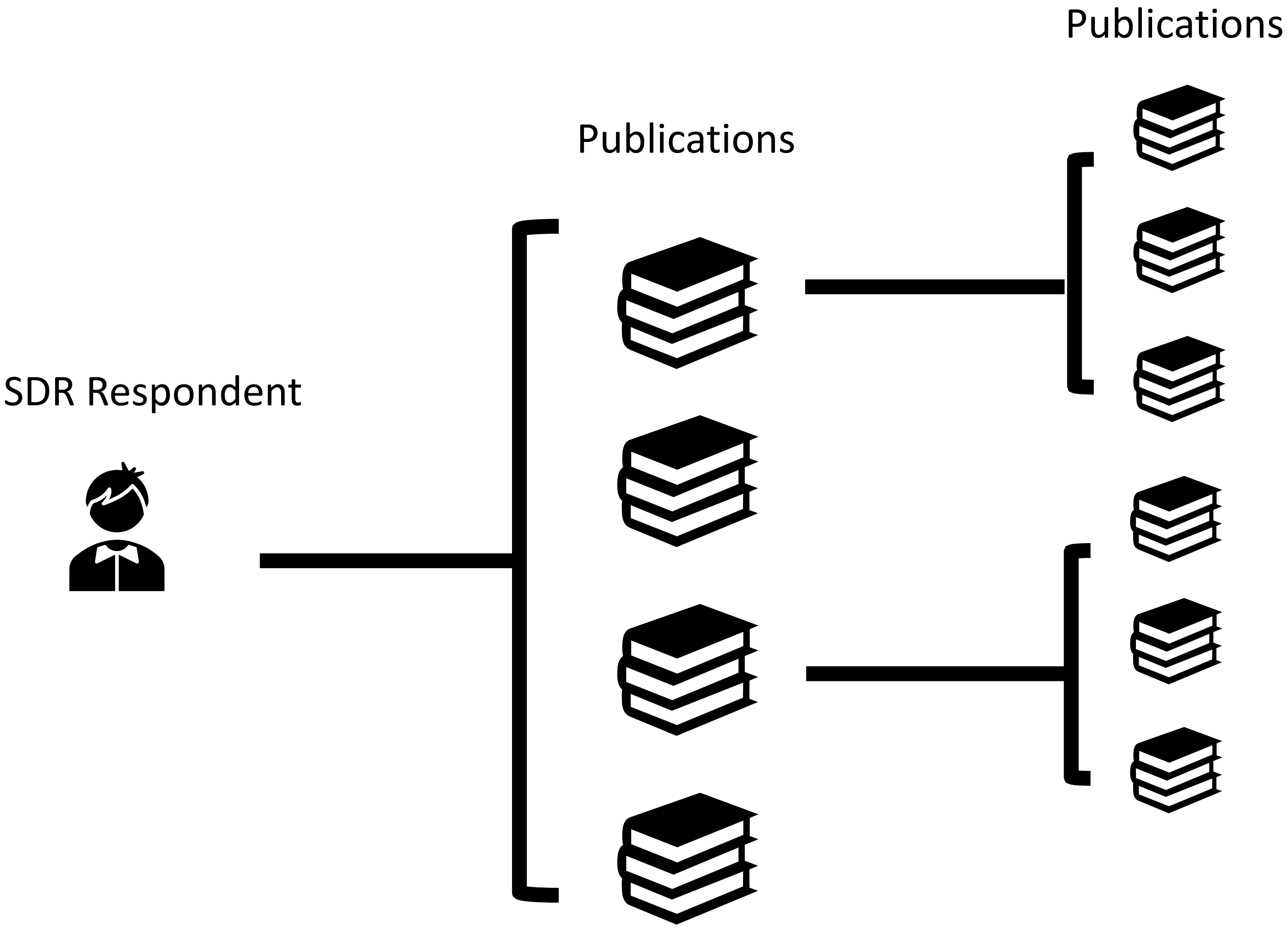
Publication Databases

Clarivate
Web of Science™

Scopus

The Web of Science (WoS) and Scopus are two leading databases providing reference and citation data from academic journals, conference proceedings, and other documents in various academic disciplines.

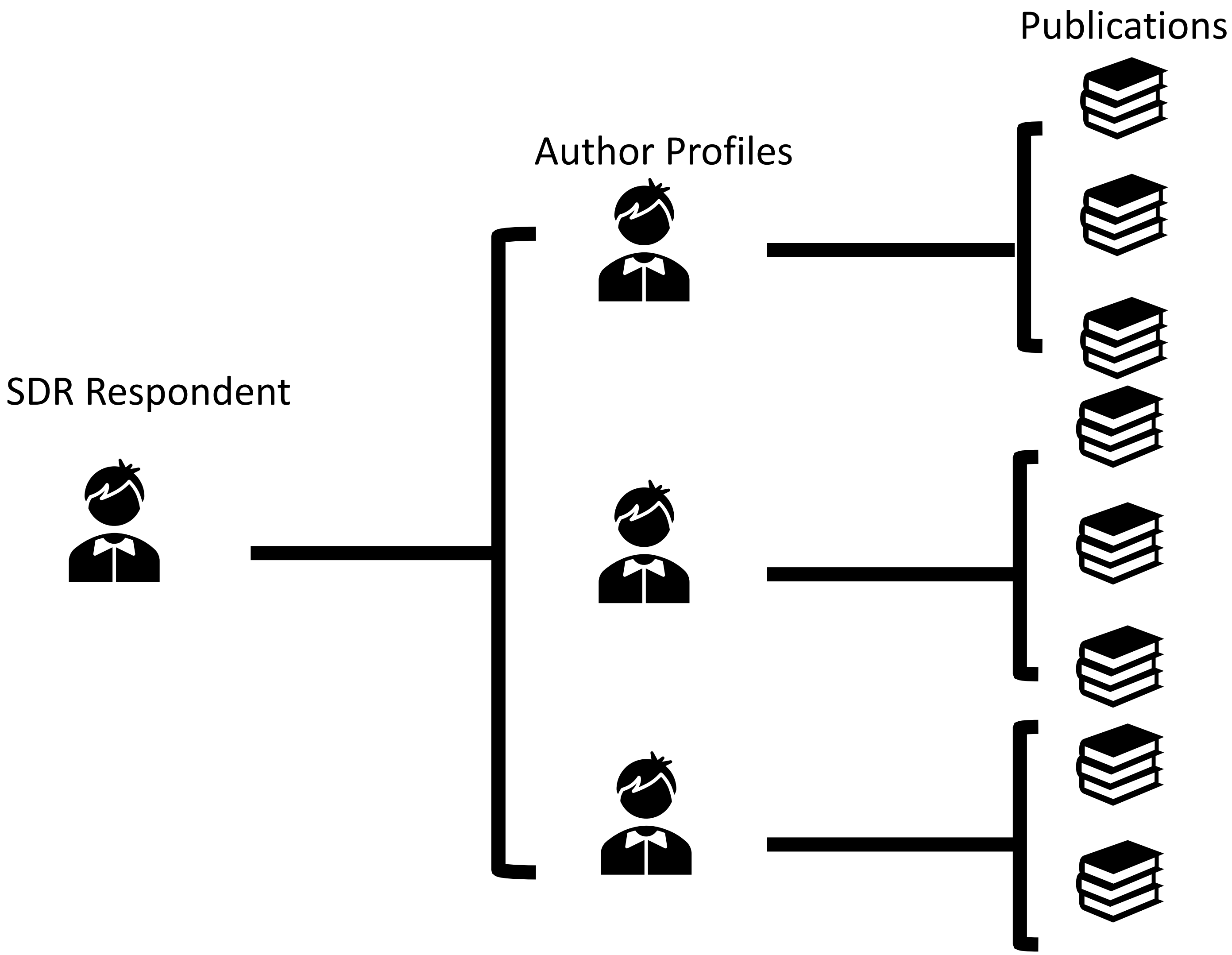
Linkage 1: SDR-WoS



Two stages of machine learning approaches implemented sequentially to address

1. survey-to-publication linking
2. Publication-to-publication linking

Linkage 2: SDR-Scopus (Preliminary)



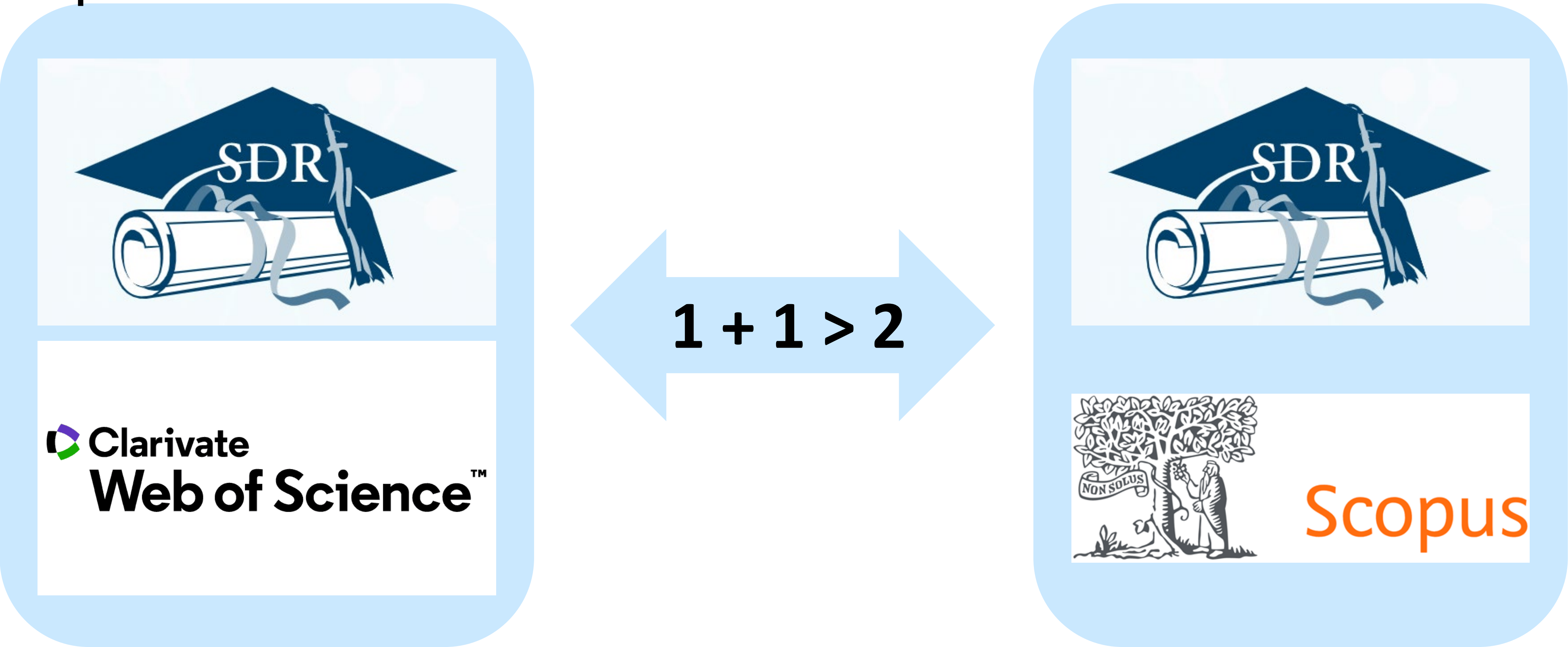
All variations of attributes from publications within a Scopus author profile are matched to comparable SDR respondent attributes, including names, emails, affiliation addresses, doctorate field, and doctorate year.

Methods

Connecting two independent data linkages

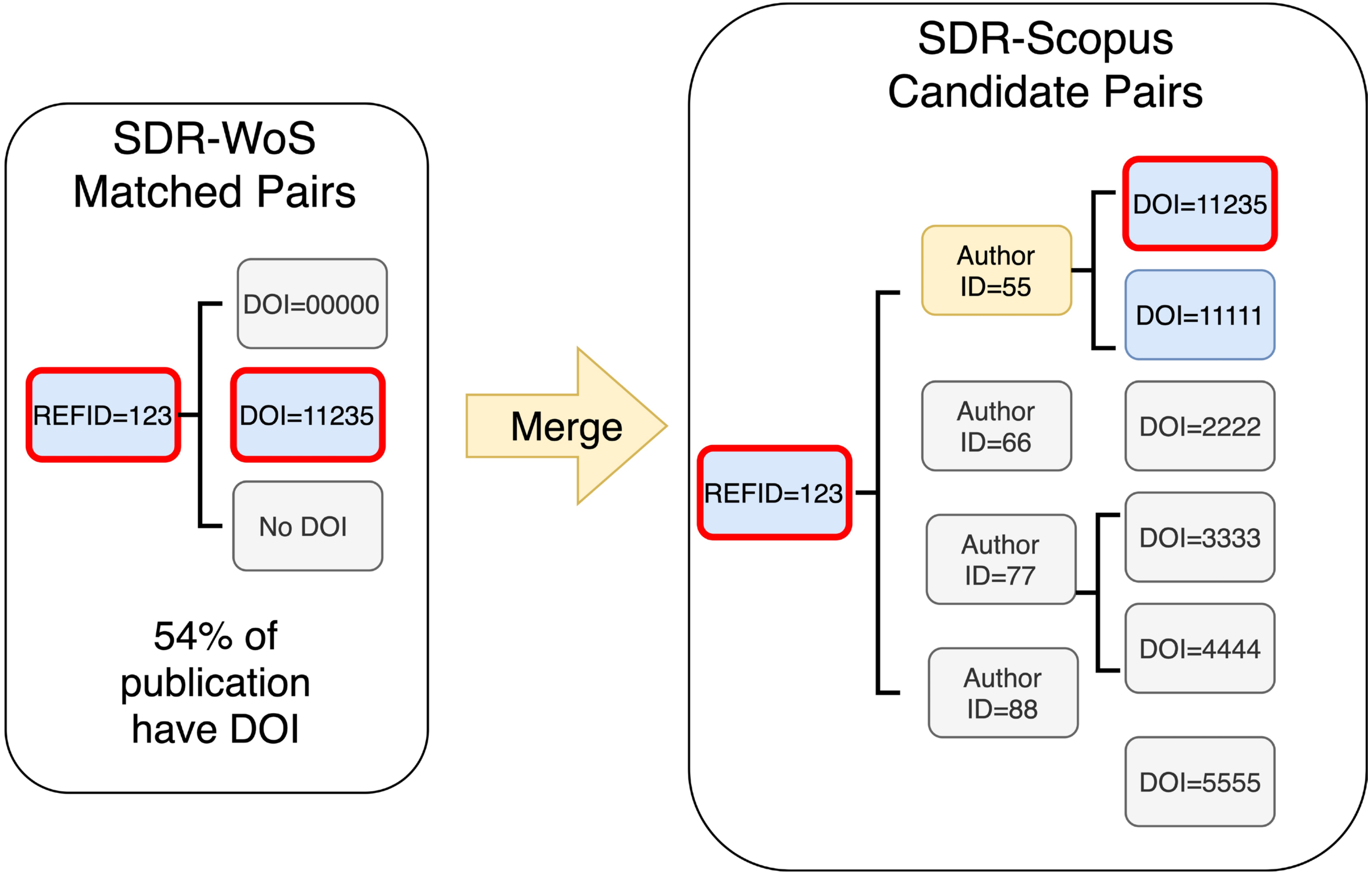
Approach

- Connect two independent large scale data linkage to gain coverage, data quality, and cost-effectiveness for future expansion.



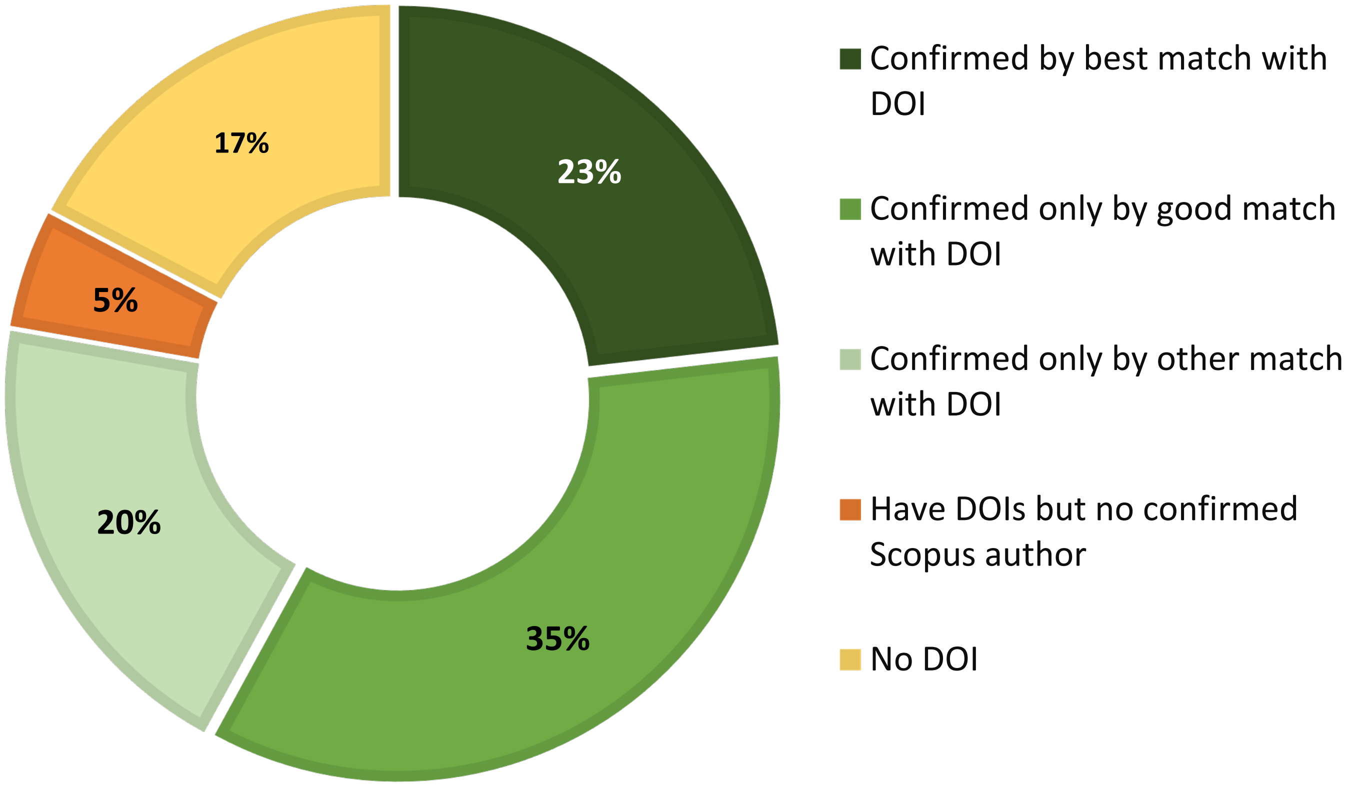
Join SDR-WoS and SDR-Scopus

Digital Object Identifier (DOI) is the bridge to connect matched publications



DOI data can effectively confirm Scopus authors

SDR-WoS authors



Although DOI is known for only half of the SDR-WoS publications, 83% of SDR-WoS authors have publication DOI data and 78% of SDR-WoS authors find DOI-confirmed Scopus authors in the candidate pool.

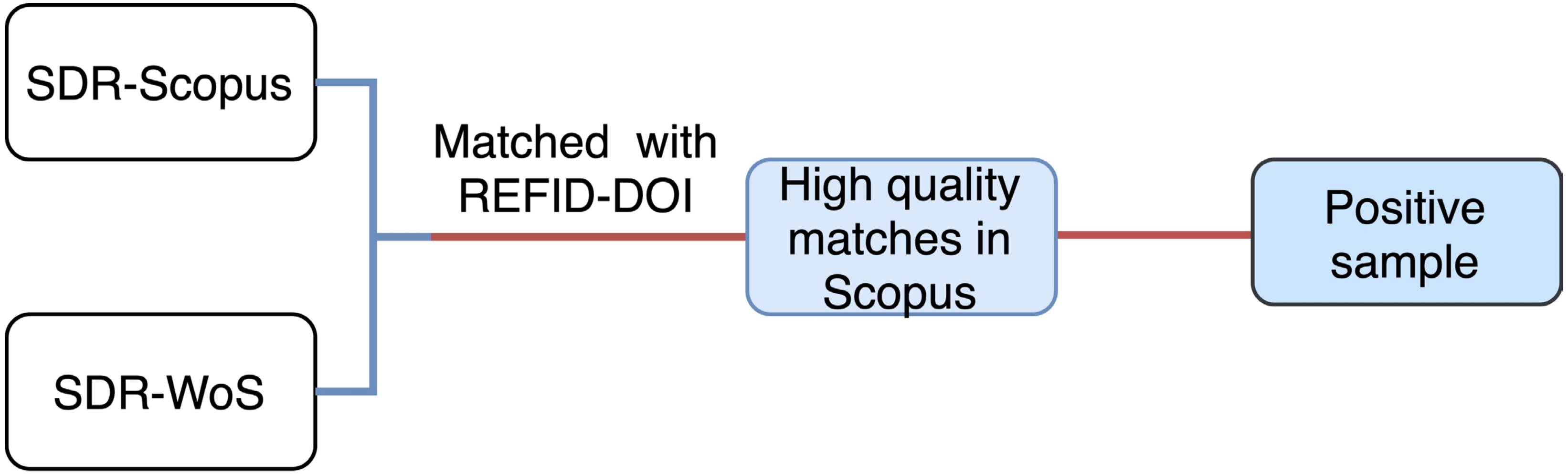
SOURCE: National Center for Science and Engineering Statistics, Survey of Doctorate Recipients linked to Web of Science Bibliometric Database and Scopus Author Profiles, 2021.

Machine learning

Constructing training and evaluation data

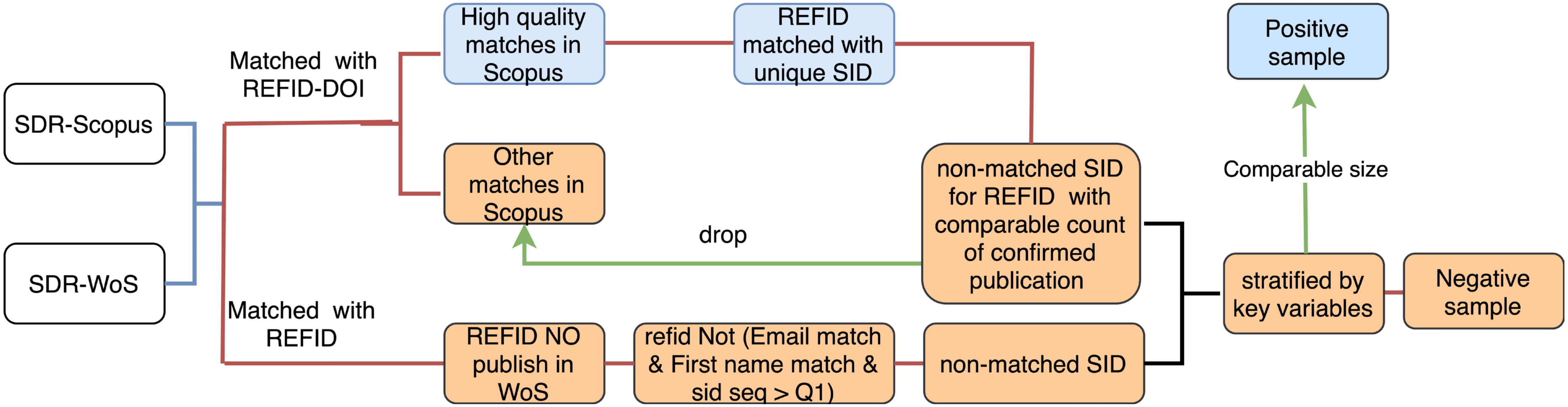
Construct training data – positive sample

- Keep REFID-author ID pairs with confirmed DOIs appeared in high-quality SDR-WoS matches.



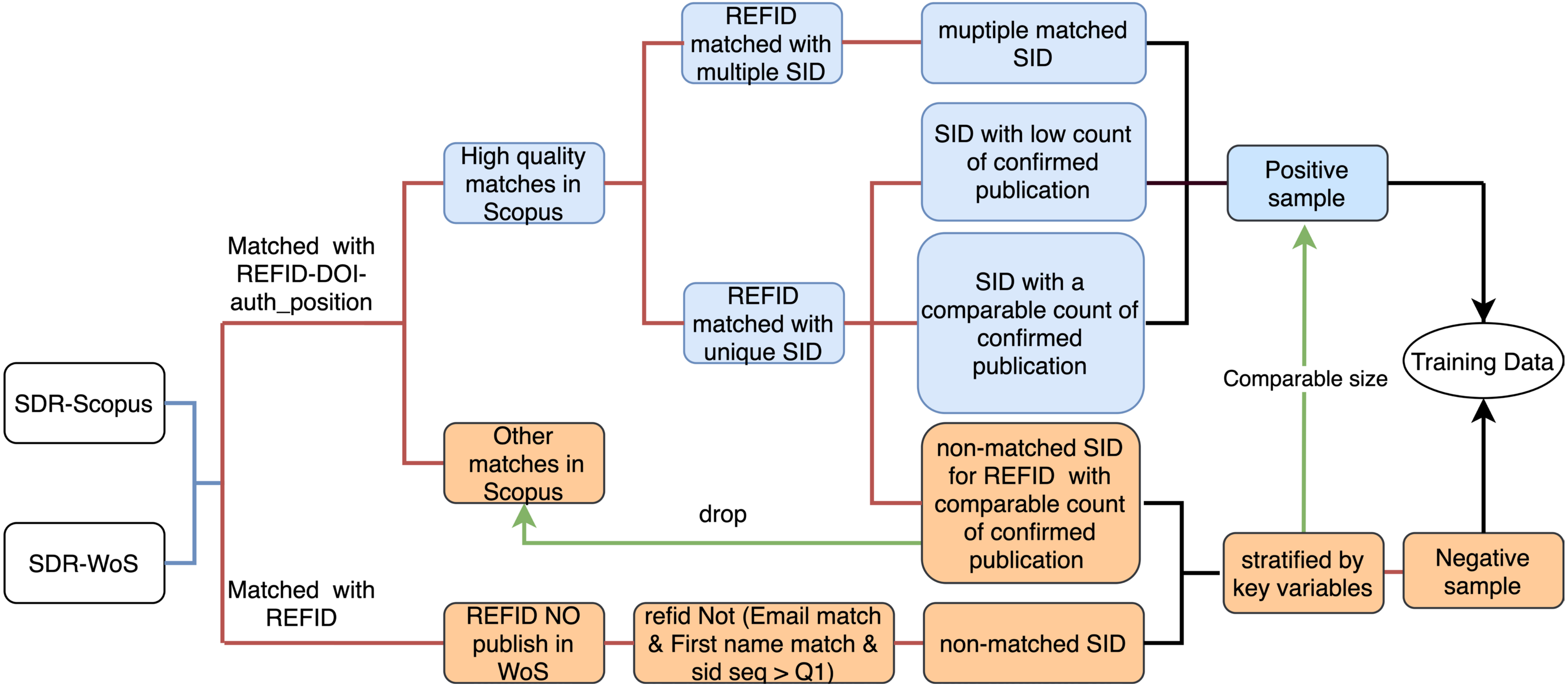
Construct training data – negative sample

- Wrong matches within same refid: remaining (SDR, Scopus author ID) pairs under a respondent having a confirm pair with comparable total matched publications.
- Non-authors: Candidate pairs under a respondent with no matched publications in WoS.



SDR-Scopus
SDR-WoS
Matched with REFID-DOI
Matched with REFID
High quality matches in Scopus
Other matches in Scopus
REFID NO publish in WoS
REFID matched with unique SID
refid Not (Email match & First name match & sid seq > Q1)
drop
non-matched SID for REFID with comparable count of confirmed publication
non-matched SID
Positive sample
Comparable size
stratified by key variables
Negative sample

Constructing the training data


SDR-Scopus
SDR-WoS
Matched with REFID-DOI-auth_position
Matched with REFID
High quality matches in Scopus
Other matches in Scopus
REFID NO publish in WoS
REFID matched with multiple SID
REFID matched with unique SID
refid Not (Email match & First name match & sid seq > Q1)
muptiple matched SID
SID with low count of confirmed publication
SID with a comparable count of confirmed publication
non-matched SID for REFID with comparable count of confirmed publication
non-matched SID
drop
Positive sample
Comparable size
stratified by key variables
Negative sample
Training Data

Machine Learning

Training and evaluation

- Split the training data into training and test sets.

Predictors

- Background and employment outcome: SED and SDR survey data
- Richness of source data for matching: quantity and quality of matching keys
- Scores of similarity: component scores of SOLR query

ML methods tried

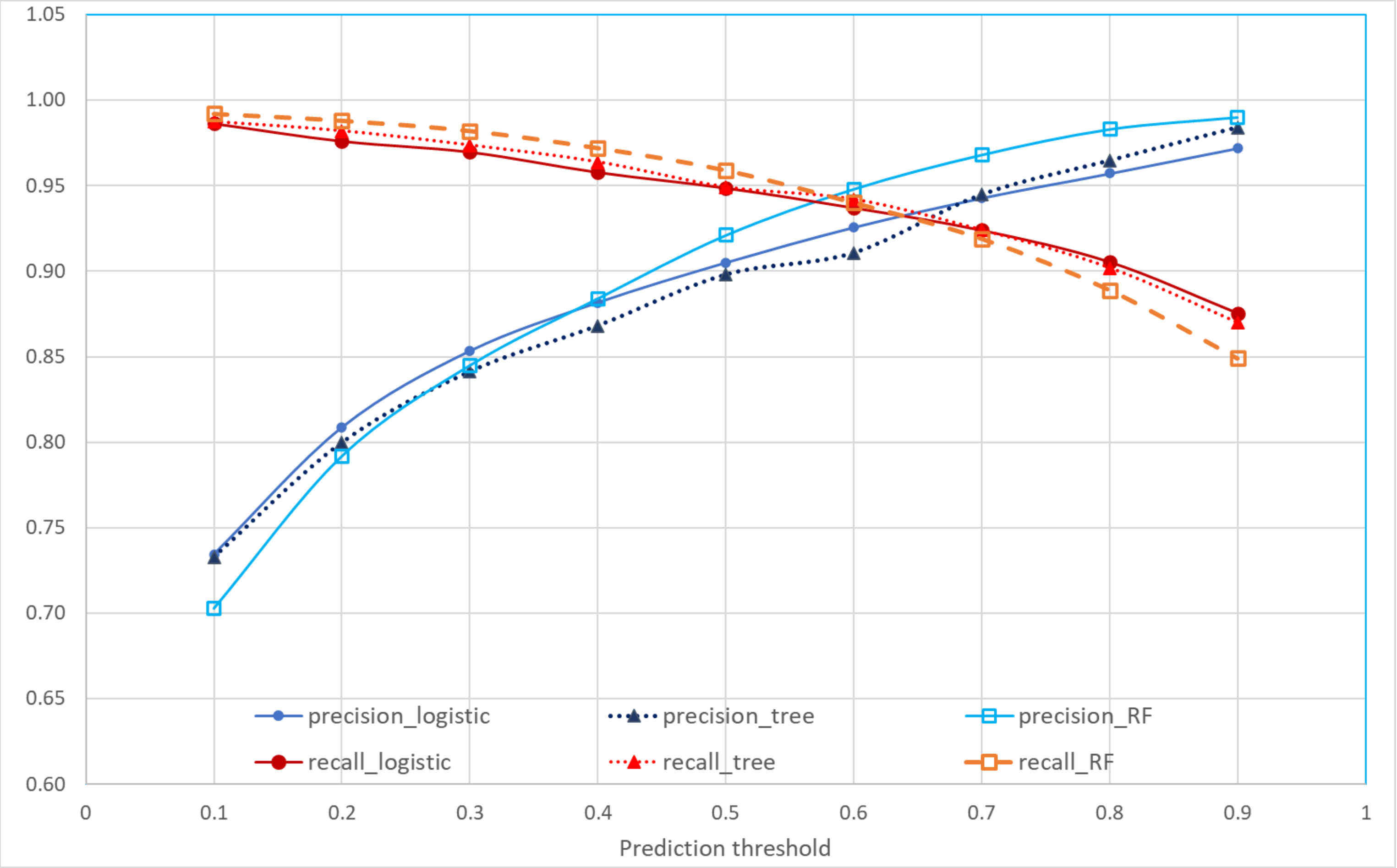
- Logistic regression
- Regression tree
- Random Forest

Precision and Recall

Precision is the ratio between the correct predictions and the total predictions.

Recall is the ratio between the correct predictions and the total number of correct items in the set.

Evaluate ML predictions



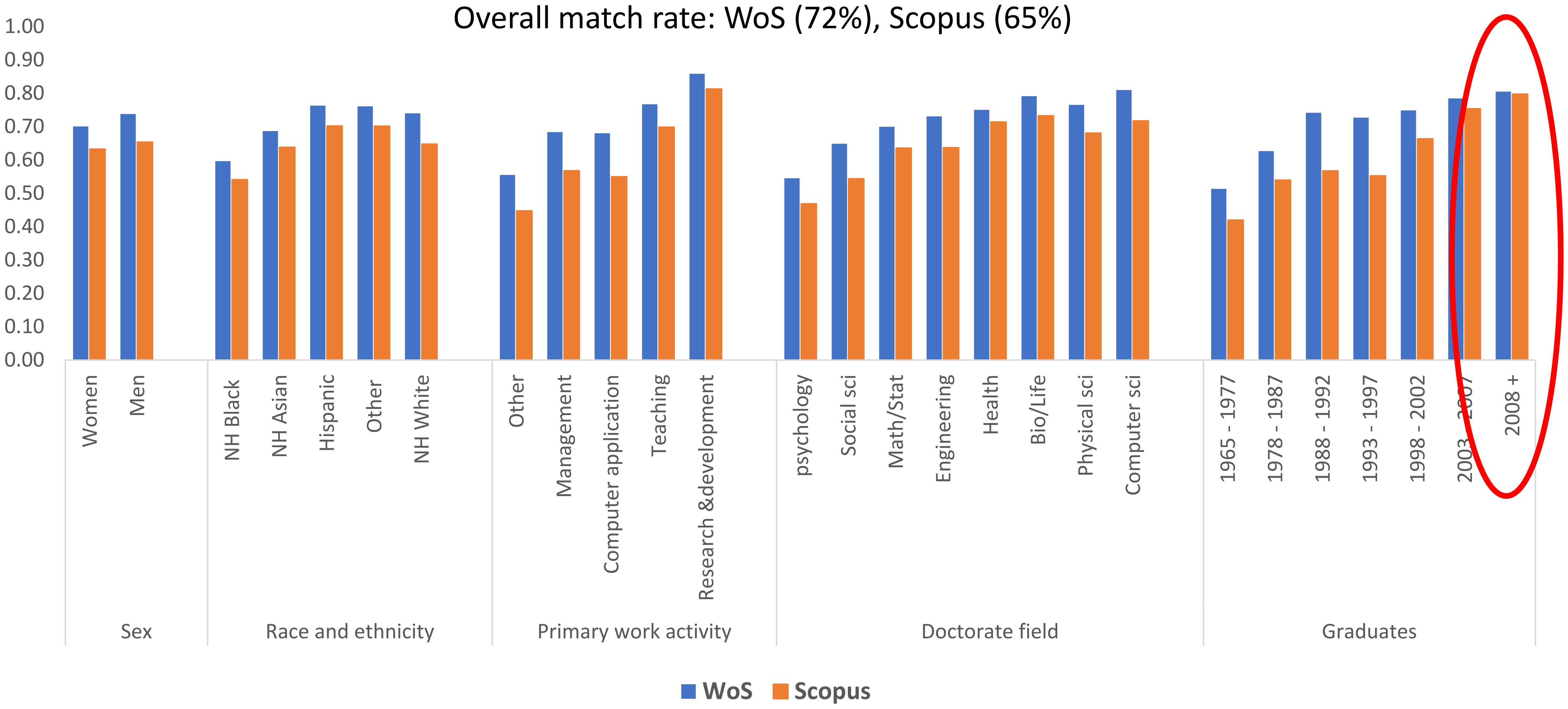
Preliminary results showing comparable performance among the three ML methods. Random Forest model performs slightly better in the prediction threshold of [0.5, 0.8]

SOURCE: National Center for Science and Engineering Statistics, Survey of Doctorate Recipients linked to Web of Science Bibliometric Database and Scopus Author Profiles, 2021.

Preliminary findings

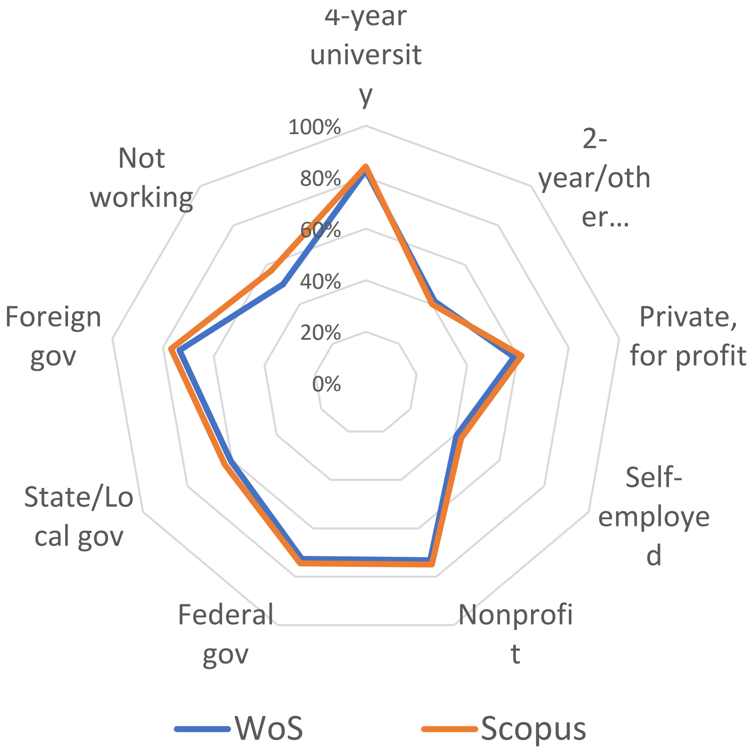
ML predictions

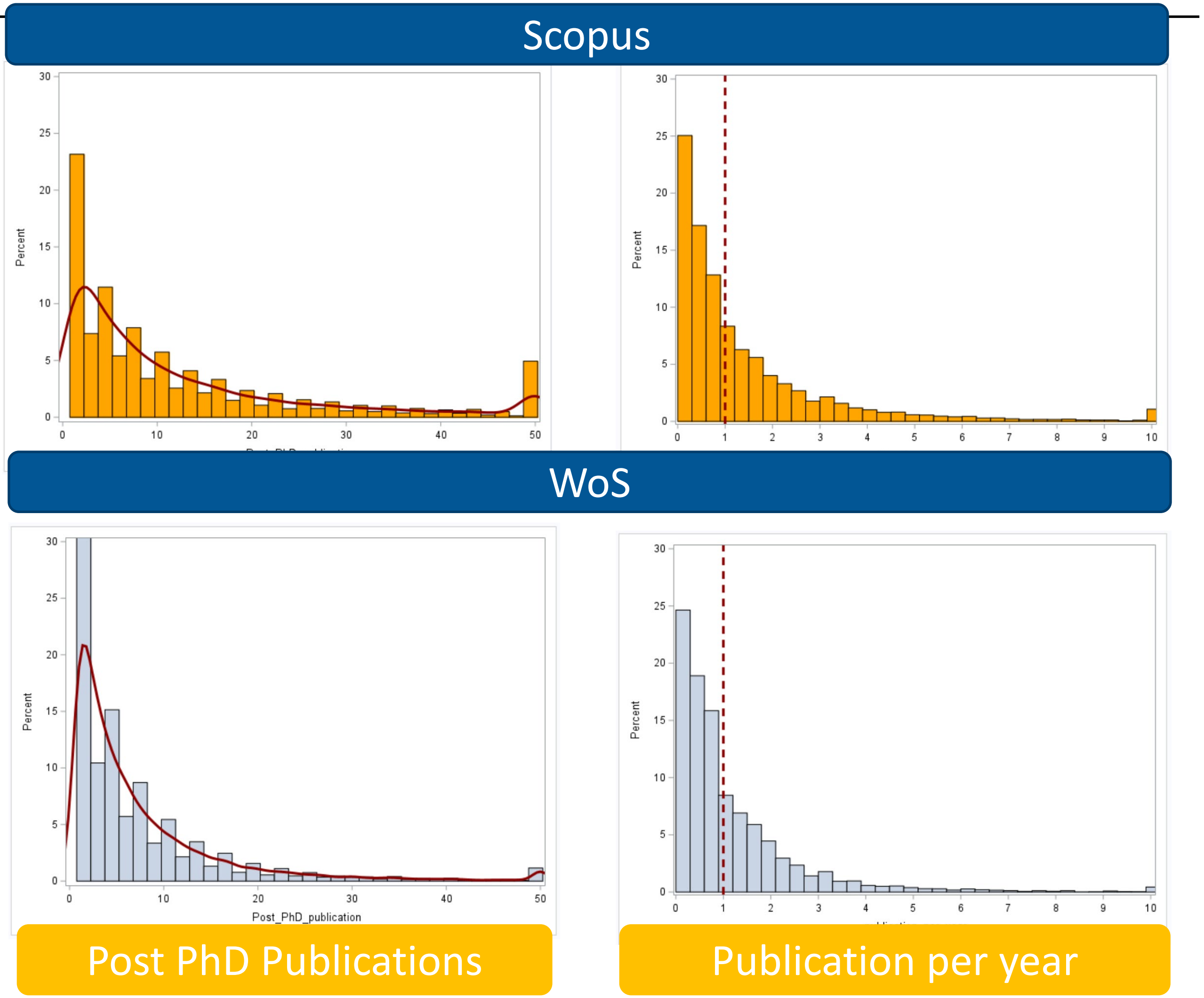
Match rates: WoS vs. Scopus



SOURCE: National Center for Science and Engineering Statistics, Survey of Doctorate Recipients linked to Web of Science Bibliometric Database and Scopus Author Profiles, 2021.

Post-graduation publications: 2008-2012 graduates





Scopus

WoS

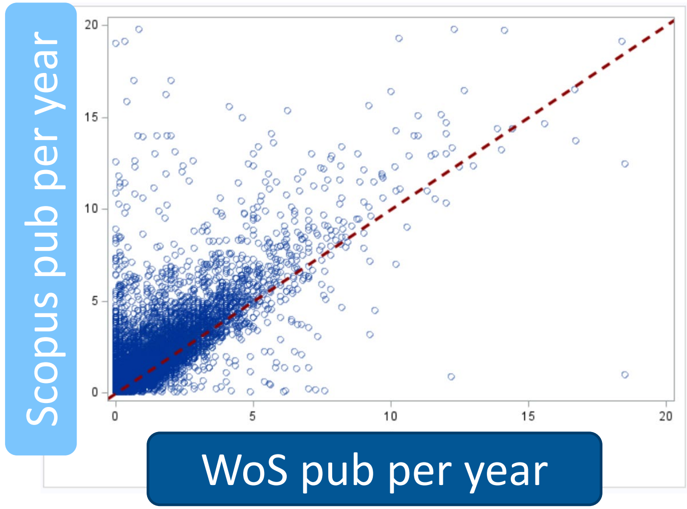
Post PhD Publications

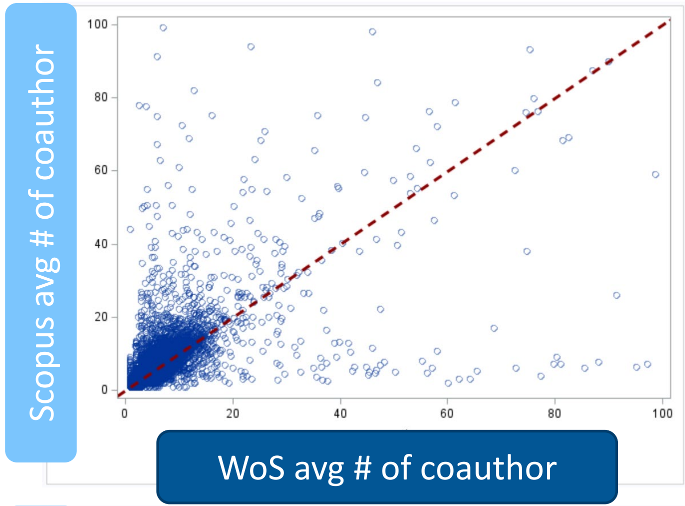
Publication per year

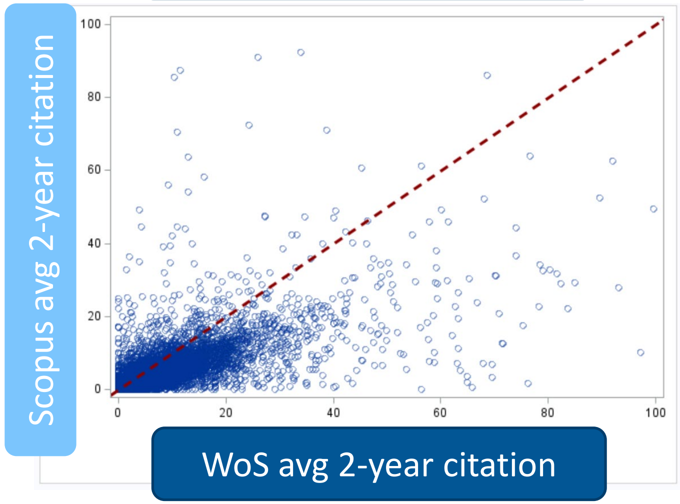
Preliminary Scopus matches show similar rate of post-PhD publications than SDR-WoS matches. Scopus author profiles contain more post-PhD publications

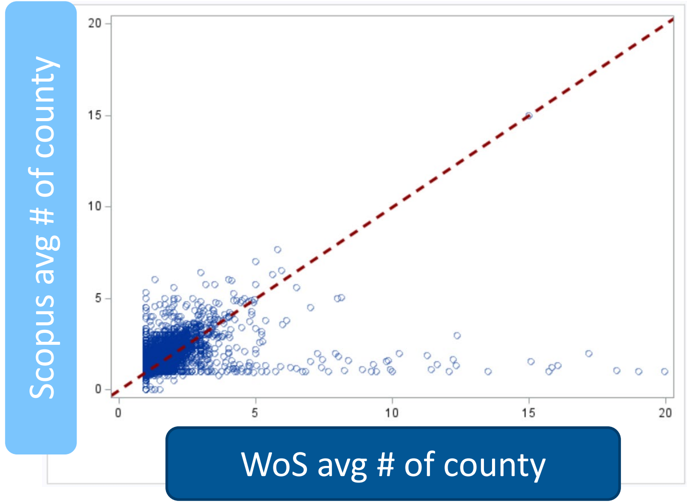
SOURCE: National Center for Science and Engineering Statistics, Survey of Doctorate Recipients linked to Web of Science Bibliometric Database and Scopus Author Profiles, 2021.

Productivity, collaboration, and impact

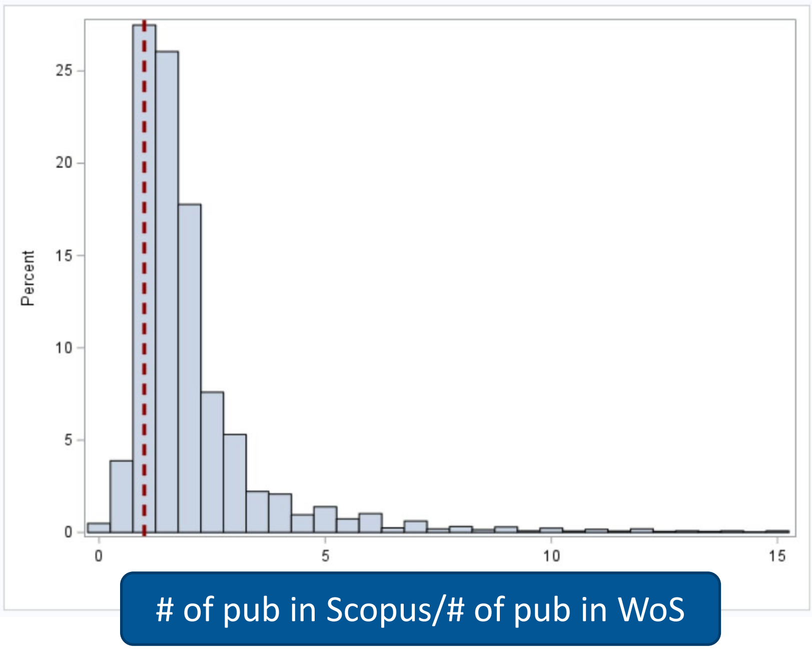








At individual author level, more matched publication per year, more coauthors, more countries of affiliation on average are observed from Scopus matches. Average 2-year citations are higher from WoS matches.



SOURCE: National Center for Science and Engineering Statistics, Survey of Doctorate Recipients linked to Web of Science Bibliometric Database and Scopus Author Profiles, 2021.

Next steps

- ML modeling refinement
- Coverage bias analysis
- Matching quality evaluation
- Linkage updates
- Build use cases

Contact:	Haoyi Wei	howei@nsf.gov
	Wan-Ying Chang	wchang@nsf.gov