LinkBERT: Discovering STEM Researchers' Trajectories through AI-Aided Data Linkage

Eric Livingston, ORISE Fellow to NCSES; Wan-Ying Chang, NCSES

October 23, 2024

Federal Committee on Statistical Methodology (FCSM) Conference

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
U.S. NATIONAL SCIENCE FOUNDATION

Disclaimer

This presentation provides results of exploratory research by the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). This information is shared to inform interested parties of ongoing activities and to encourage further discussion. Views expressed are those of the presenters and not necessarily those of NCSES or NSF.

This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN24-062.

Project Overview

Goal: Develop AI models for superior data linkage between the Survey of Doctorate Recipients (SDR) data and Scopus author bibliometrics to support analysis of research trajectories of STEM PhDs

Two-phase approach

1. Apache Solr search to gather initial candidates for linking
   - Traditional keyword-based database search
   - Generate an initial list of top 10 likely candidate links

2. AI refinement to link records between SDR and Scopus
   - LinkBERT: Custom BERT-like transformer we have developed that uses Solr search scores and other metadata to refine and finalize the list of linked records
     - Bert: Bidirectional Encoder Representations from Transformers
   - Random forest (RF): A machine learning model uses a collection of decision trees to make predictions and determine linked records; used for baseline comparison
   - Goal is to out-perform previous models

Data

Survey of Doctorate Recipients (SDR)

A longitudinal survey of U.S.-trained research PhDs in science, engineering, or health fields. The 2021 SDR sample represents doctorate recipients from degree year: 1968–2019. Survey data include the following:

Education history

Bachelor's, master's, doctoral degrees (year, institution, place, field of study),

Background

Age, sex, marital status, dependents, parent's education level, birthplace, country of citizenship, race and ethnicity, disability status, source of financial support

Demographic information

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

Employment situation

Labor force status, principal job, principal employer, faculty rank, tenure status, work activity, salary, benefits, job satisfaction, federal support

Other work-related experiences

Recent educational experiences

Contact information

Names, e-mails, contact address, affiliation names

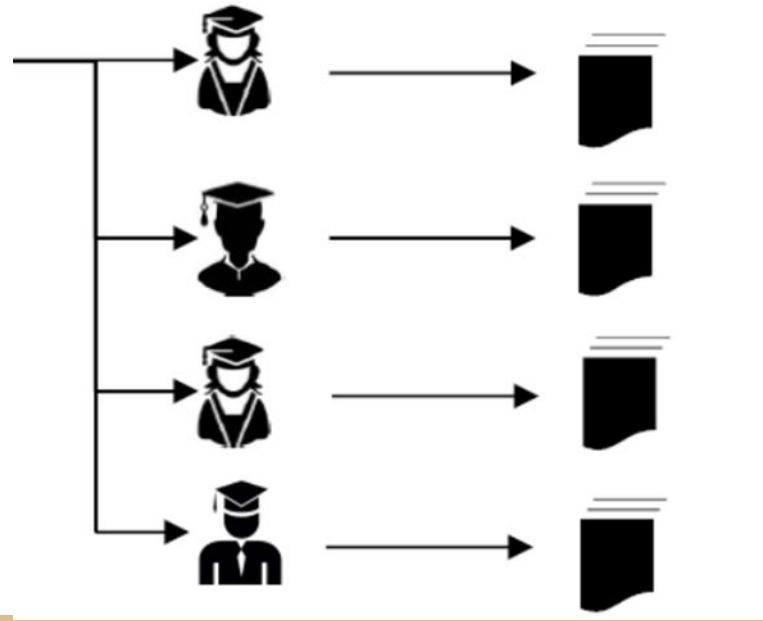
Scientific Publication Databases

Scopus is a leading database providing reference and citation data from academic journals, conference proceedings, and other documents in various academic disciplines.

- More than 16 million disambiguated author profiles
- Author metadata: Elsevier generated two Solr Core files (one with author data and one with publication data) for linking

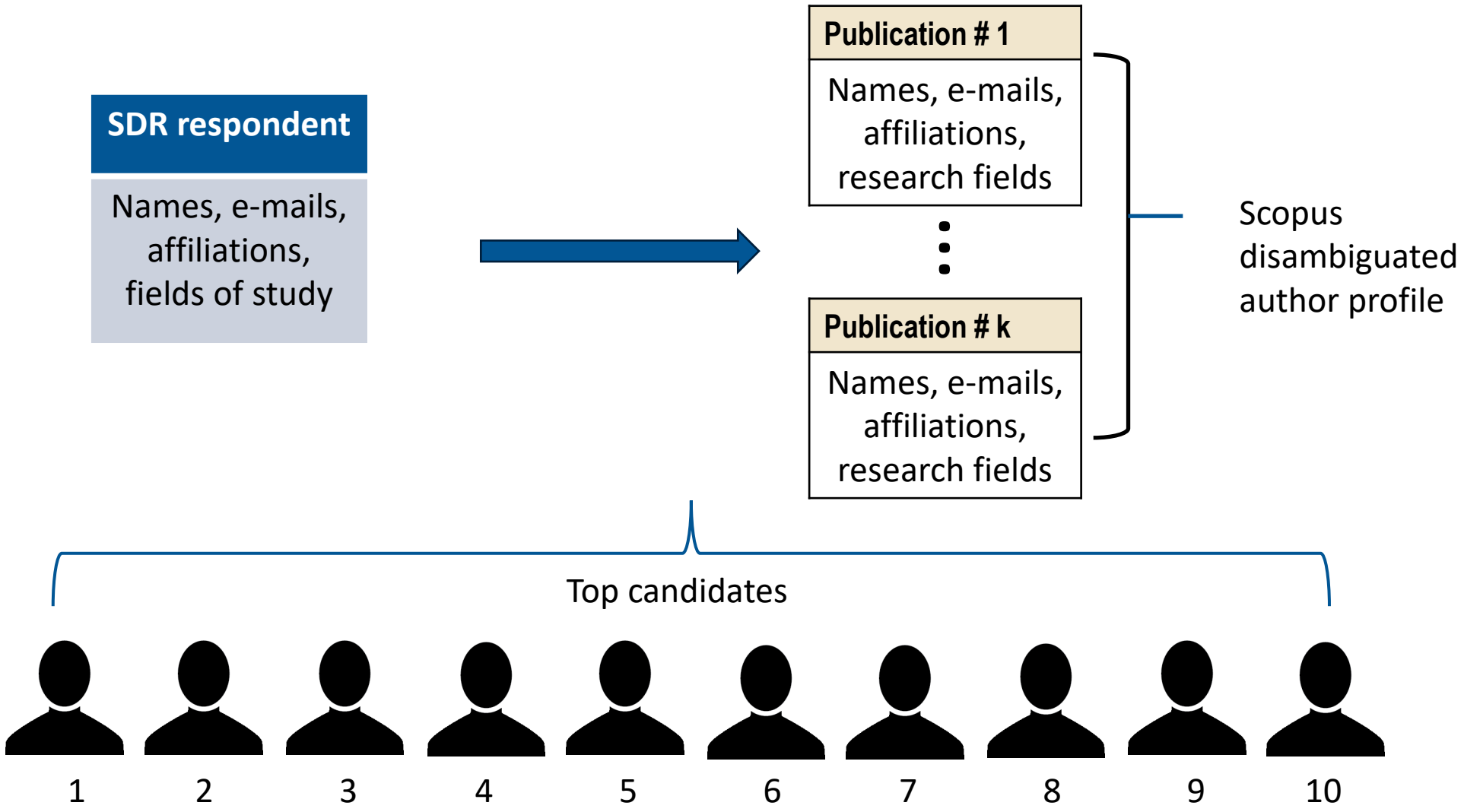
Scopus

Author profiles



Linkage Approach

Phase 1: Finding Candidate Author Profiles



Phase 1: Solr Scopus Search

- Apache Solr search engine to propose top 10 candidates per SDR respondent
- Scopus author profiles database (16+ million profiles)
- Matching scores aggregated from several subscores
  - Names
    - Levenshtein distance, term frequency, inverse document frequency (TF-IDF) scores on first and last names
  - E-mail addresses
  - Affiliations (PhD and employers vs. Scopus publications)
  - Fields of science (PhD fields vs. Scopus journal classification)

Phase 1: SDR and Scopus Data Challenges

- Name variations
  - Multiple spellings (e.g., François vs. Francois)
  - Nicknames, cultural naming conventions
- Multiple author profiles
  - Single SDR respondent may have several Scopus profiles
- Mapping SDR fields to Scopus fields
  - e.g., SDR institutions are coded by the Integrated Postsecondary Education Data System (IPEDS) ID, whereas Scopus uses its own affiliation IDs
  - Not all SDR fields have Scopus equivalents

Phase 2: Machine Learning to Finalize Linkage

- Features for prediction
  - Solr scores summarizing and measuring the level of similarity between names, e-mails, affiliations, and research fields
  - Summary statistics of quantity and quality of personally identifiable information
  - Background and employment outcome: SDR survey data
- Develop labeled data for training
  - Use high-quality previously developed SDR-Web of Science links and a unique identifier (DOI [Digital Object Identifier]), to identify linked publications in Scopus and create a labeled set
  - Select the labeled set into for-training and for-evaluation samples stratified by SDR demographics and quality of key variables for linking
- Prediction and evaluation
  - Explore various cutoff values for predicting links; evaluate the results to balance rates of false links and missed links

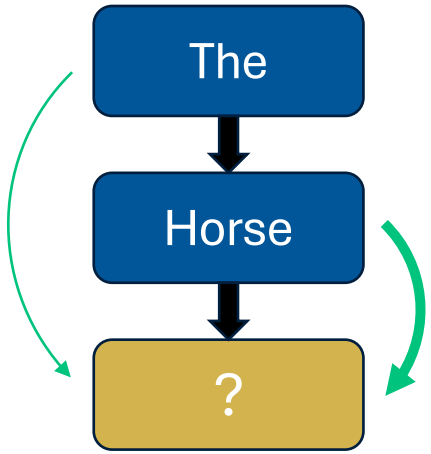
Phase 2: BERT (vs. Large Language Models)

Traditional LLM (Transformer)

- ChatGPT, Claude, Llama3
- Unidirectional
- Predicts next word (class)
- Context: Previous words

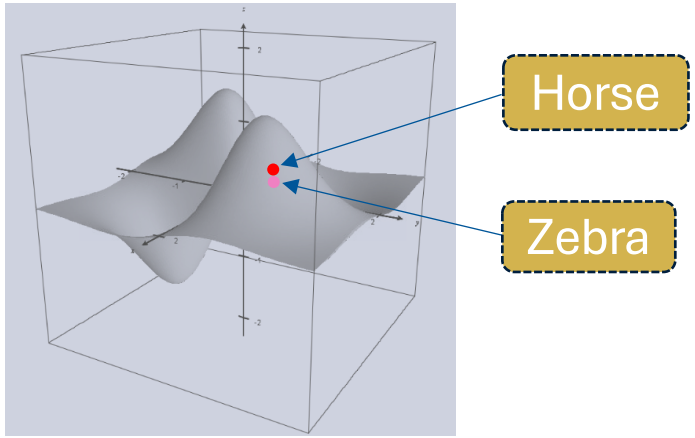
BERT (Bidirectional Encoder Representations from Transformers)

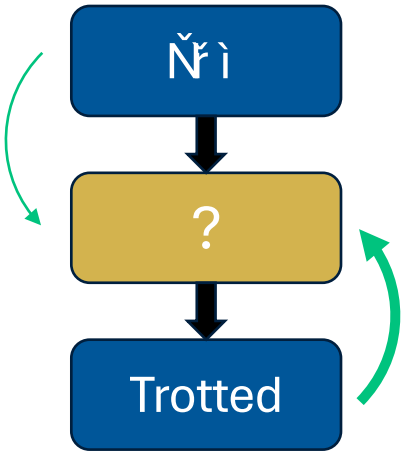
- Bidirectional
- Predicts missing word (class)
- Context: All other words



BERT (for Classification)

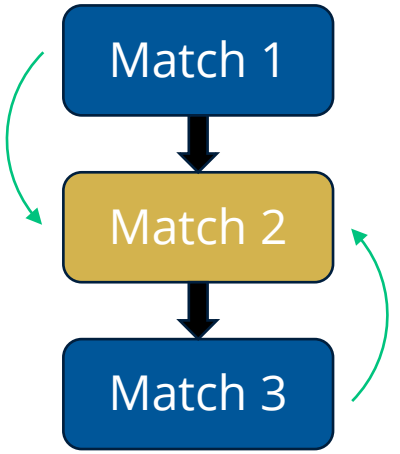
- Can consider the full context
- Perfect for lists of options
- Is basically curve-fitting data

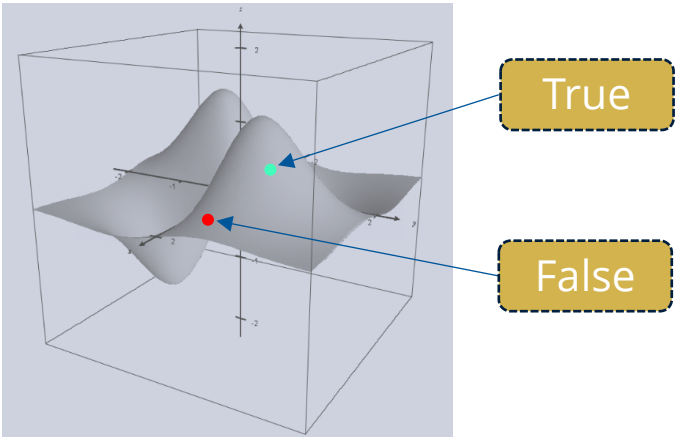




Phase 2: AI Inference with LinkBERT

- Custom AI transformer model based on BERT
  - Examines 10 candidate matches as a group
  - Aims to classify each as a “true” or “false” match
  - Like BERT with a two-word vocabulary
- Input data
  - Phase 1 scores
  - Additional categorical metadata
    - Demographic data (e.g., gender, race)
    - Employment information (e.g., employer type, income level)
    - Temporal data (e.g., age, year of PhD)



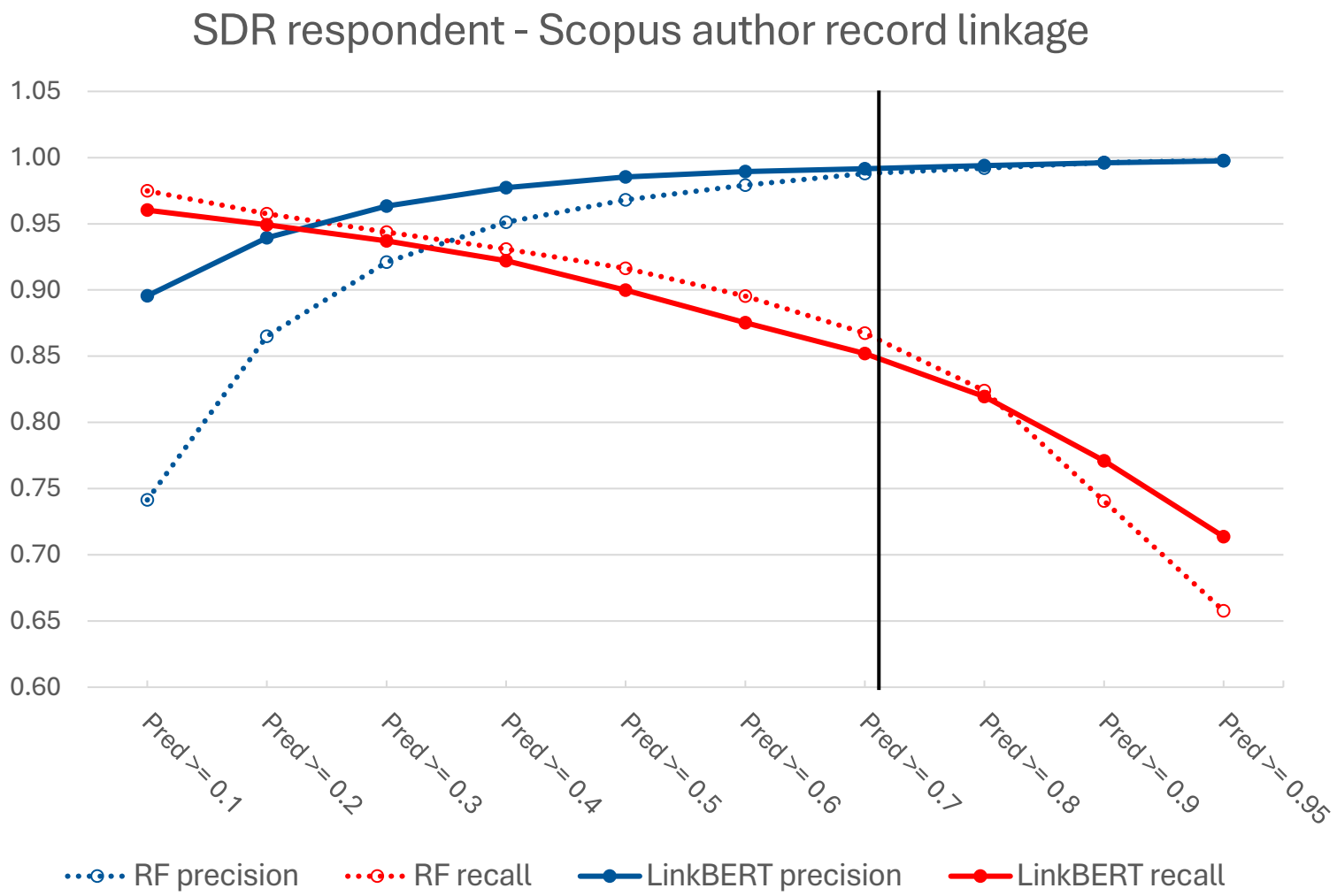


Results

Linkage Errors: Precision and Recall

- Precision
  - What proportion of all links were true matches?
  - Higher precision means fewer false links
- Recall
  - What proportion of all true matches were found in our set of links?
  - Higher recall means fewer missed links

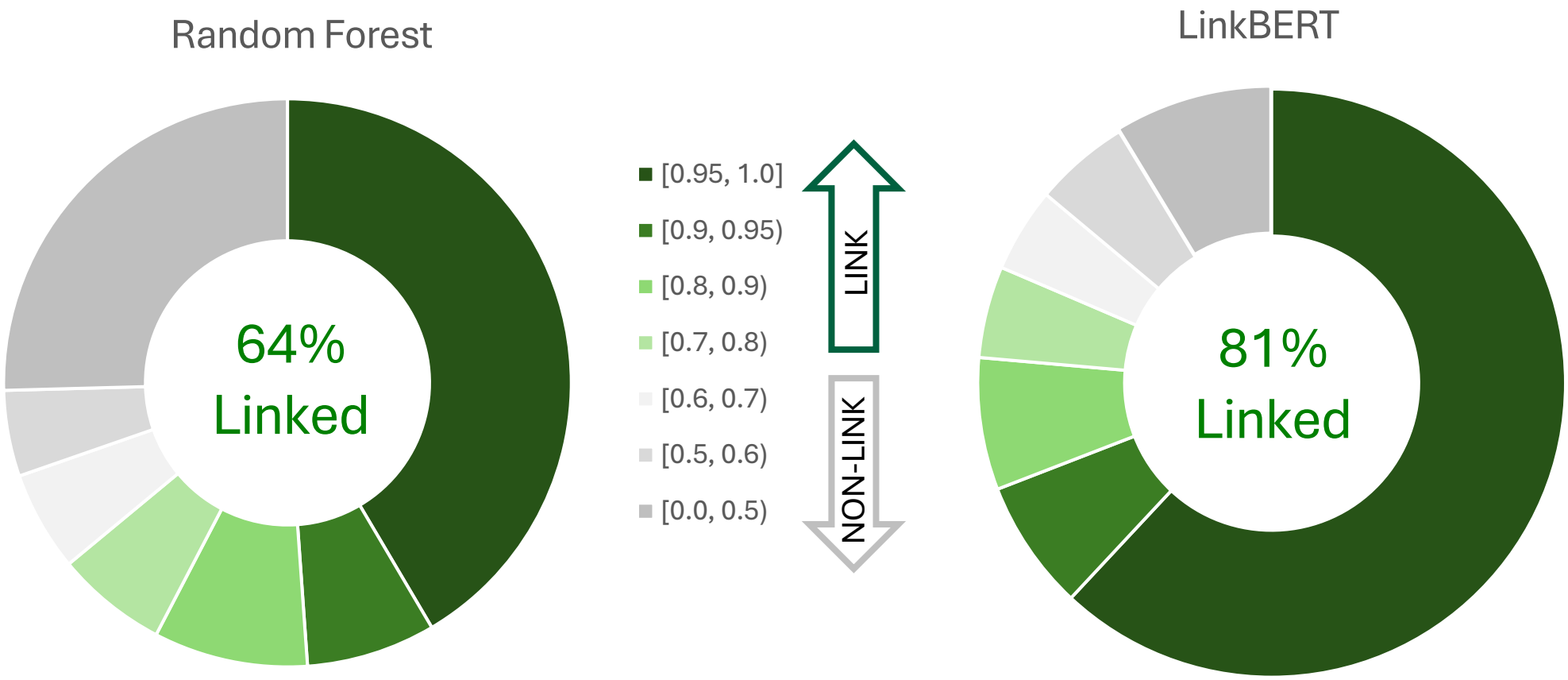
Measure Precision and Recall



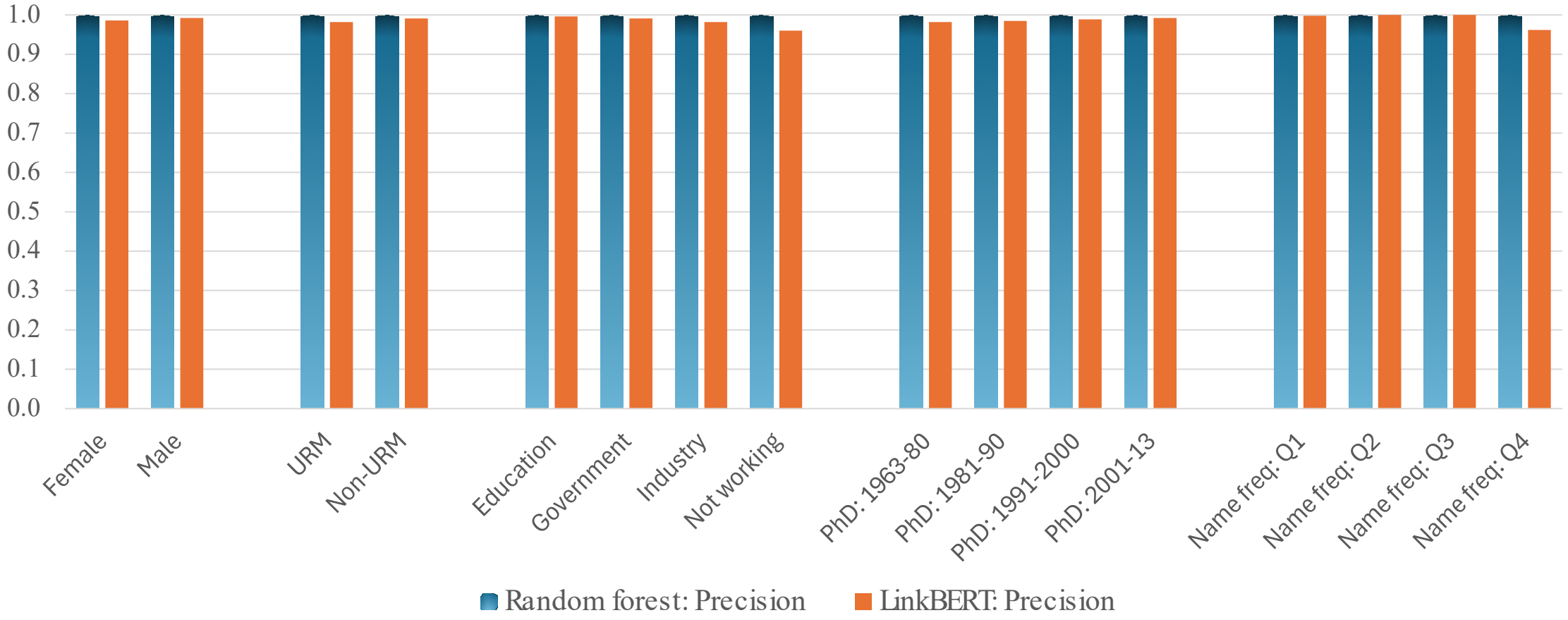
The evaluation sample, ~108,000 respondent-author pairs with known linkage status, is used to estimate precision and recall for each prediction cutoff value ranging from 0.1 to 0.95 and applied to the RF (random forest) and LinkBERT linkage predictions.

Random Forest vs. LinkBERT

Proportion of linked SDR respondents with prediction cutoff at 0.7

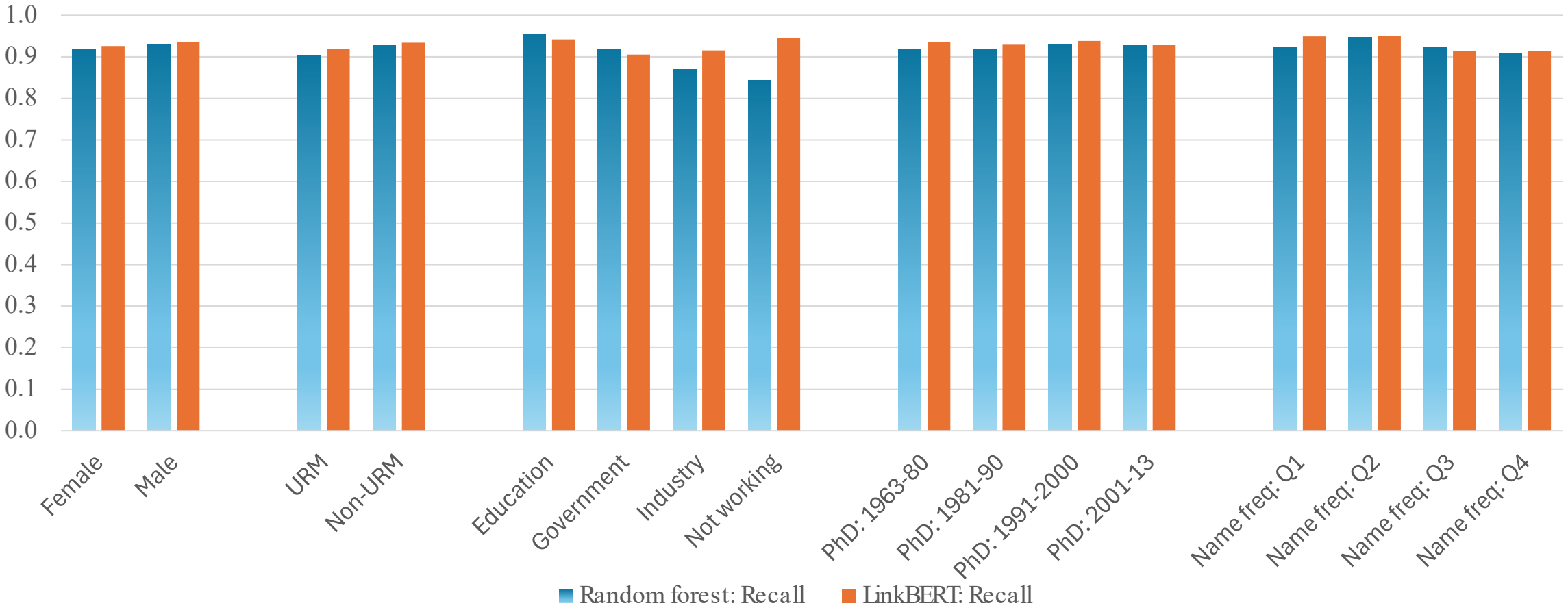


Precision by Subgroups



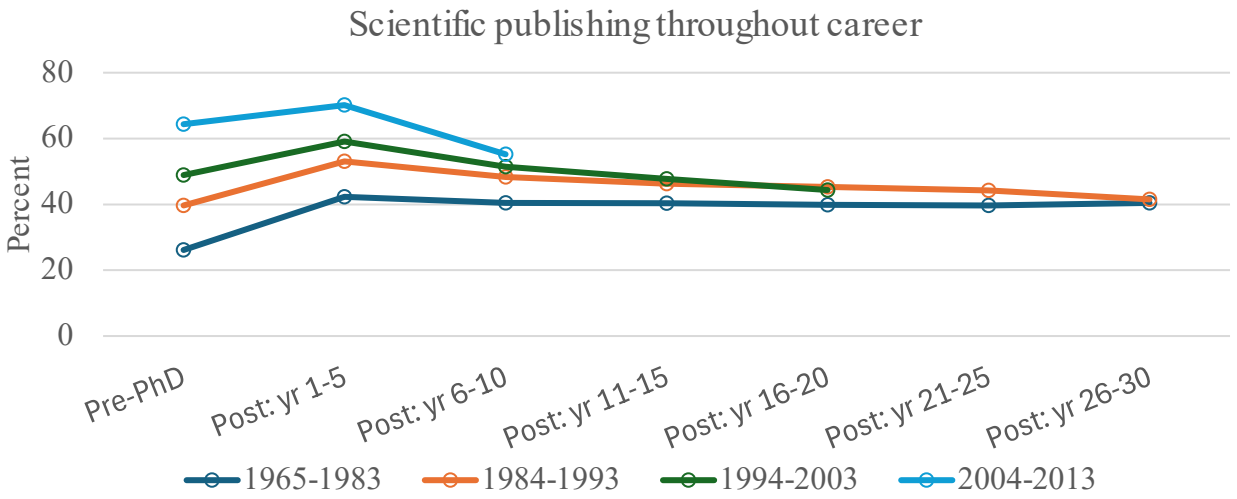
Notes: Prediction cutoff of 0.7 is applied. Underrepresented minority (URM) includes Hispanic or Latino, Black, and American Indian or Alaska Native.

Recall by Subgroups

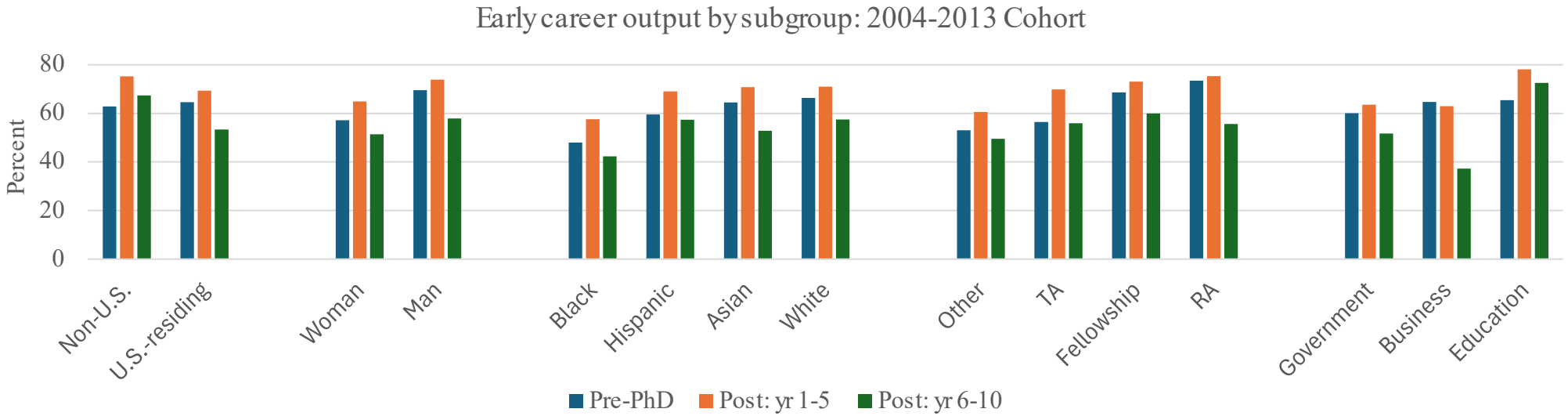


Notes: Prediction cutoff of 0.7 is applied. Underrepresented minority (URM) includes Hispanic or Latino, Black, and American Indian or Alaska Native.

Publication Output of Scientists and Engineers



Proportion of doctoral researchers published at least once during each of the pre- and post-PhD interval is estimated by graduating cohort and demographic subgroup using the LinkBERT results.



Next Steps

Machine learning prediction and evaluation

Refine Methods and Expand Applications

- Improve representation of labeled data
- Adjust Solr search criteria to explore alternative strategies
- Adjust LinkBERT model parameters to increase performance
- Research performance characteristics of the model
- Focus on harder edge cases where data are ambiguous
- Develop a sustainable procedure for linkage updates with new samples and more current Scopus data

Thank you!

Eric Livingston: elivings@nsf.gov
Wan-Ying Chang: wchang@nsf.gov

https://ncses.nsf.gov