DICE @ Rich Context Competition

Prof. Dr. Axel Ngonga
Nikit Srivastava
Rricha Jalota

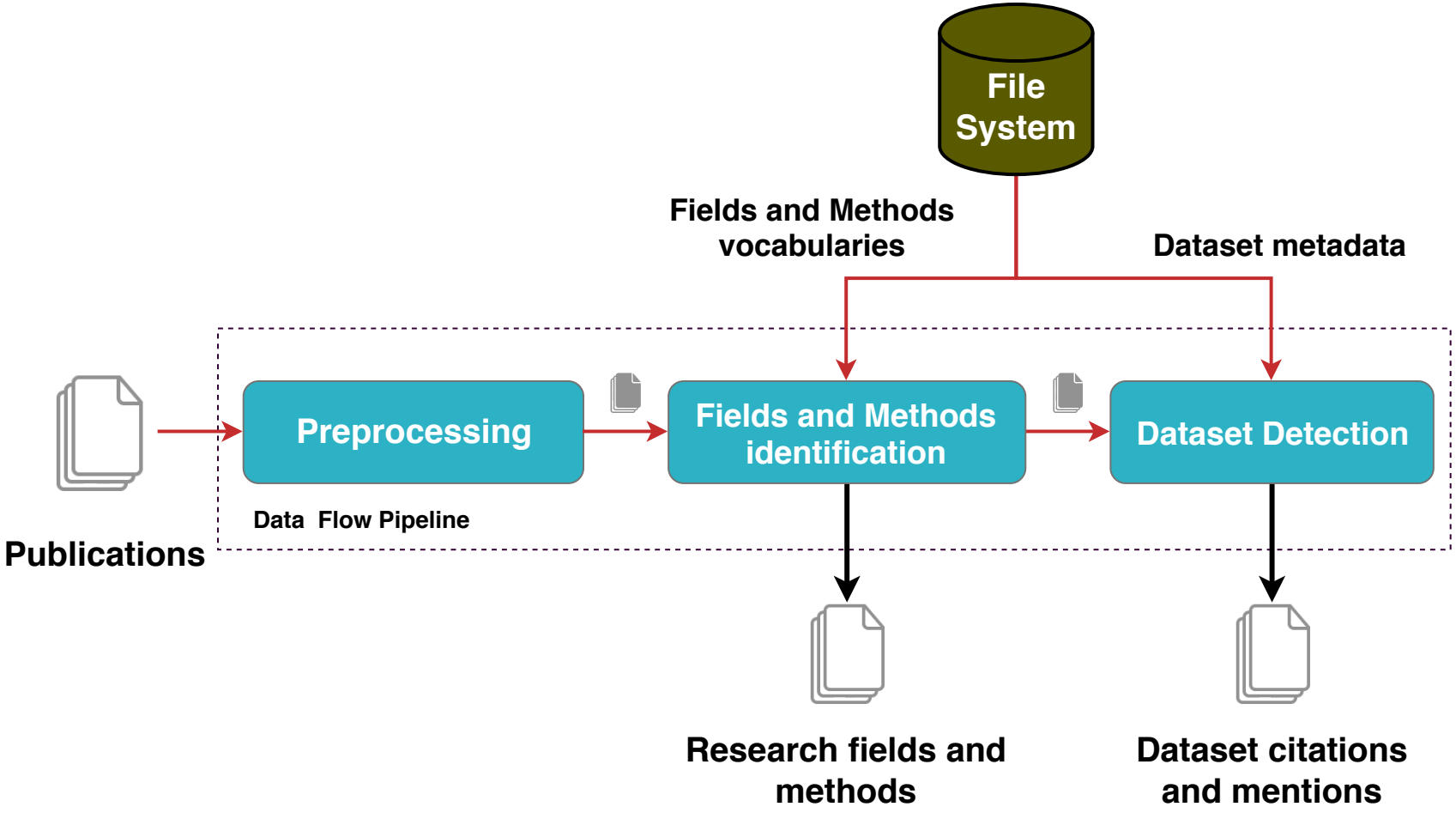
Data Science Group
Paderborn University

February 15, 2019

Section 1

Overview

Project Architecture



Preprocessing

- Generated text files from PDF using pdftotext[a]

```
ISBN 3865581226 (Internetversion)
×ØÖ
Ø
Í× Ò ×ØÓ
 ×Ø
 ×
ÓÙÒØ
ØÓÖ ÔÔÖÓ
 ¸ Û Ö Ú Ø Ü
Ø ×ÓÐÙØ ÓÒ ÓÖ Ö ØÖ ¹
Ö ÓÒ Ý Ð × ÓÖ Ø
 × Ø Ø Ø × ÓÖØ¹Ø ÖÑ ÒØ Ö ×Ø Ö Ø ÓÐÐÓÛ× Ø Ö × ÓÐ ÔÖÓ
 ××
Û Ø Ø ÒØ Ö
 ÔØ ×Û Ø
 Ò Ò Ó ÒÓÙ×ÐÝº Ì Ý Ð ÙÒ
Ø ÓÒ×¸ Ñ ÔÔ Ò Ø ÓÒ ¹ÑÓÒØ
Ö Ø ÒØÓ n¹Ô Ö Ó Ý Ð ×¸ Ü Ø
ÓÒÚ Ü¹
ÓÒ
 Ú × Ô ØÓ Ø Ð Ø Ò Ø Ö Ø Ó Ø
Ø Ö × ÓÐ Ú ÐÙ ¸ Ö ×Ô
Ø Ú ÐÝº Ì × × Ò
ÓÒØÖ ×Ø ØÓ Ð Ò Ö × ÓÖØ¹Ö Ø ÔÖÓ
 ×× × Û
 ÑÔÐÝ
Ò Ò Ý Ð ÙÒ
Ø ÓÒº Ì ÒØ ÖÚ Ð× ÓÖ Û
```

Sample text

Preprocessing

Centrality-based Capital Allocations*

Adrian Alter[†], Ben Craig[‡], Peter Raupach[§]

December 19, 2014

Abstract

We look at the effect of capital rules on a banking system that is connected through correlated credit exposures and interbank lending. The rules, which combine individual bank characteristics and interconnectivity measures of interbank lending, are to minimize a measure of system-wide losses. Using the detailed German Credit Register for estimation, we find capital rules based on eigenvectors to dominate any other centrality measure, followed by closeness. Compared to the baseline case, capital reallocation based on the Adjacency Eigenvector saves 14.6% in system losses as measured by expected bankruptcy costs.

Keywords: Capital Requirements, Centrality Measures, Contagion, Financial Stability

JEL classification: G21, G28, C15, C81.

Sample text

- Generated text files from PDF using pdftotext[a]
- Handled words that got split by hyphens
- Removed irrelevant data (references, acknowledgments...)
- Extracted main sections (Abstract, JEL-Classification code, keywords, methodology/data, summary, discussion)

Preprocessing

Centrality-based Capital Allocations*

Adrian Alter[†], Ben Craig[‡], Peter Raupach[§]

December 19, 2014

Abstract

We look at the effect of capital rules on a banking system that is connected through correlated credit exposures and interbank lending. The rules, which combine individual bank characteristics and interconnectivity measures of interbank lending, are to minimize a measure of system-wide losses. Using the detailed German Credit Register for estimation, we find capital rules based on eigenvectors to dominate any other centrality measure, followed by closeness. Compared to the baseline case, capital reallocation based on the Adjacency Eigenvector saves 14.6% in system losses as measured by expected bankruptcy costs.

Keywords: Capital Requirements, Centrality Measures, Contagion, Financial Stability

JEL classification: G21, G28, C15, C81.

Sample text

- Generated text files from PDF using pdftotext[a]
- Handled words that got split by hyphens
- Removed irrelevant data (references, acknowledgments...)
- Extracted main sections (Abstract, JEL-Classification code, keywords, methodology/data, summary, discussion)
- Extracted noun phrases from these sections

[a] poppler-utils/pdfinfo.1.en.html

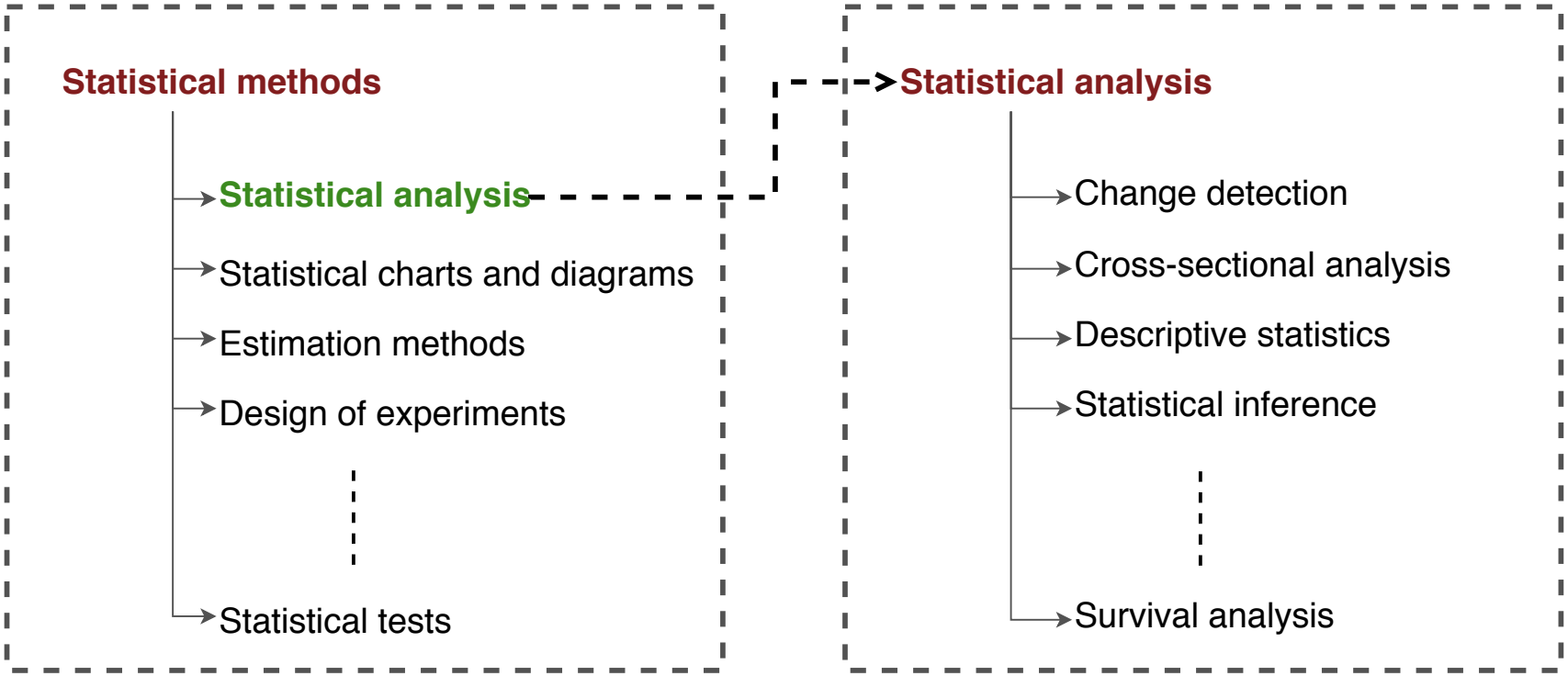
Section 2

Identification of Research Methods and Fields

Identification of Research Methods and Fields

Preprocessing

- Research methods vocabulary from DBpedia
  - Curated statistical methods from Wikipedia[1]
  - Scraped their description from corresponding DBpedia resources
  - Extracted noun phrases from description



Hierarchy of Wikipedia Categories

Identification of Research Methods and Fields

Preprocessing

- Research methods vocabulary from DBpedia
  - Curated statistical methods from Wikipedia[1]
  - Scraped their description from corresponding DBpedia resources
  - Extracted noun phrases from description

Statistical methods

- Bayesian model reduction
- DeMix
- Genetic assignment methods
- Kneser–Ney smoothing
- ⋮
- Synthetic control method

Statistical analysis

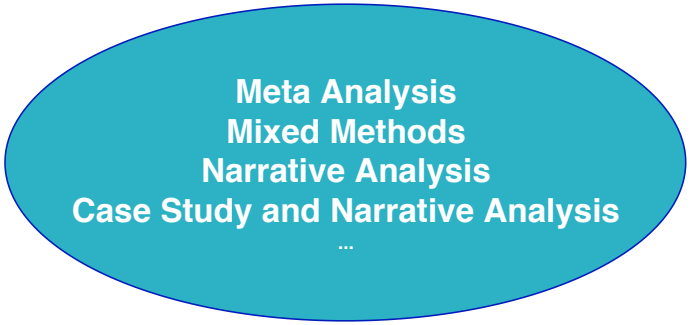
- Boolean analysis
- Fenwick (statistic)
- Functional data analysis
- InfoQ
- ⋮
- Weighting

Resources in Wikipedia Category

Identification of Research Methods and Fields

Preprocessing

- Research methods vocabulary from DBpedia
  - Curated statistical methods from Wikipedia[1]
  - Scraped their description from corresponding DBpedia resources
  - Extracted noun phrases from description
- SAGE Research fields vocabulary
  - Created blacklist of irrelevant fields
  - Extracted noun phrases from description



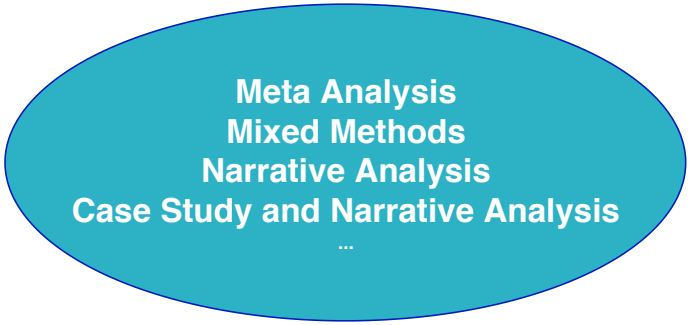
Blacklisted terms

[1] https://en.wikipedia.org/wiki/Category:Statistical_methods

Identification of Research Methods and Fields

Preprocessing

- Research methods vocabulary from DBpedia
  - Curated statistical methods from Wikipedia[1]
  - Scraped their description from corresponding DBpedia resources
  - Extracted noun phrases from description
- SAGE Research fields vocabulary
  - Created blacklist of irrelevant fields
  - Extracted noun phrases from description
- Generated vector models for both Research Fields and Methods

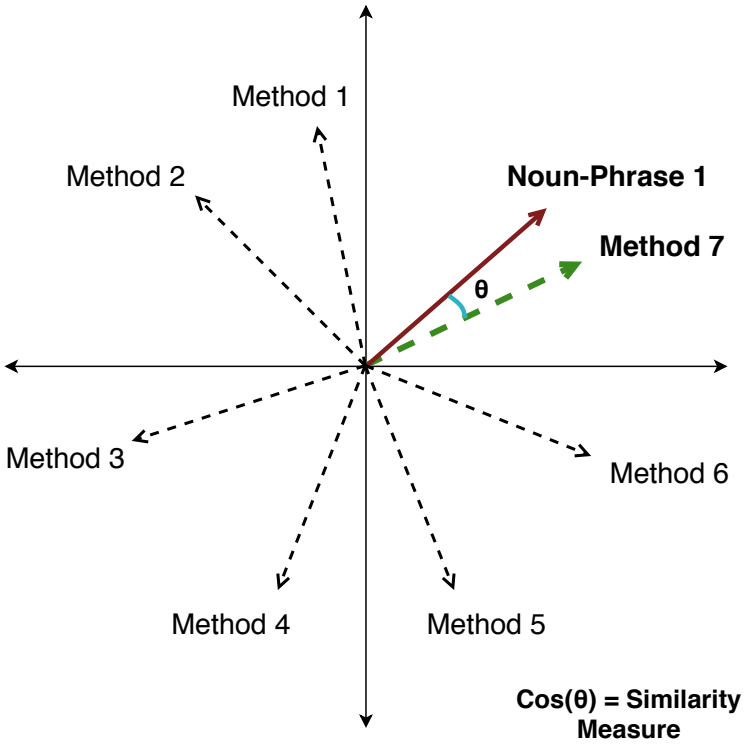


Blacklisted terms

[1] https://en.wikipedia.org/wiki/Category:Statistical_methods

Identification of Research Methods and Fields

Methods Identification - Word2Vec

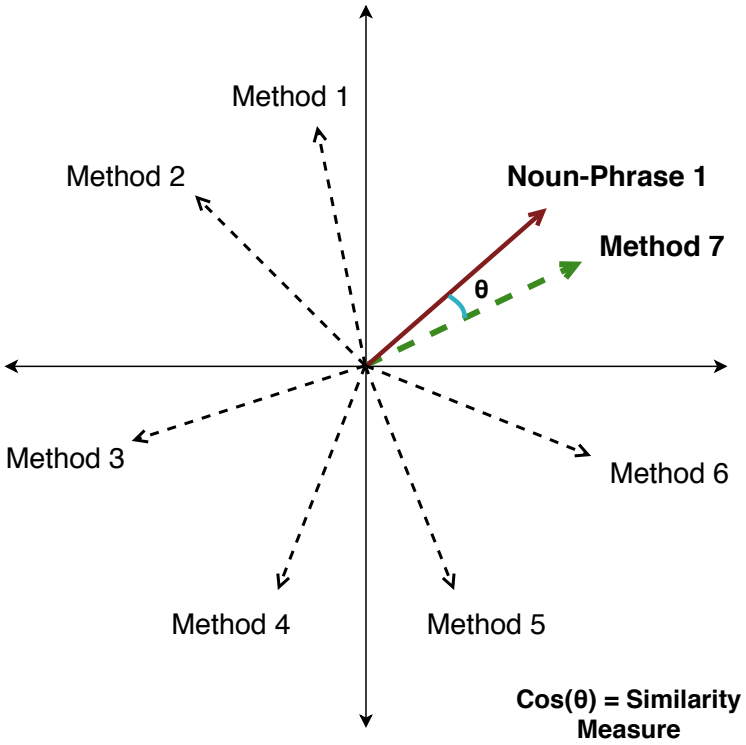


Noun-phrases from text and methods in Embedding Space

- Closest research method vector found by measuring cosine similarity between noun phrase vectors and method vectors

Identification of Research Methods and Fields

Methods Identification - Word2Vec

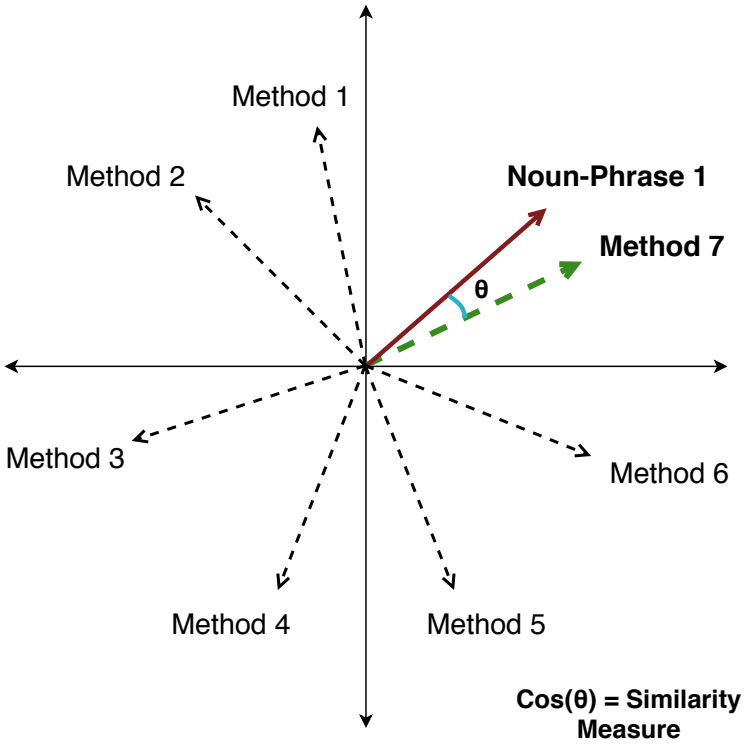


Noun-phrases from text and methods in Embedding Space

- Closest research method vector found by measuring cosine similarity between noun phrase vectors and method vectors
- Computed the significance of recurring methods using IDF
- Each research method assigned a weightage

Identification of Research Methods and Fields

Methods Identification - Word2Vec



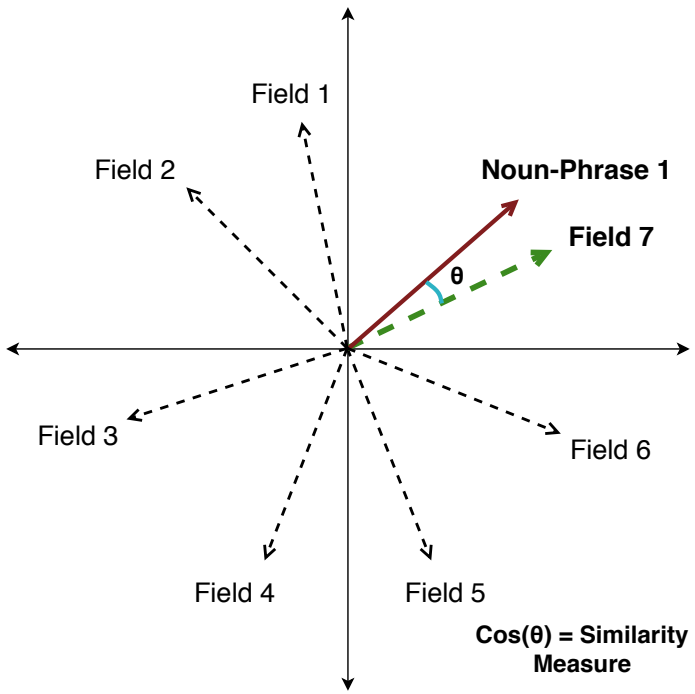
Noun-phrases from text and methods in Embedding Space

- Closest research method vector found by measuring cosine similarity between noun phrase vectors and method vectors
- Computed the significance of recurring methods using IDF
- Each research method assigned a weightage
- Re-ran the algorithm to find closest method vector and then sorted the pairs based on weighted cosine similarity
- Pair with highest cosine similarity chosen

Identification of Research Methods and Fields

Fields Identification - Word2Vec

- Top 10 closest research field vectors found using cosine similarity between noun phrase vectors and field vectors
- Pairs with similarity score < 0.9 filtered out
- If not blacklisted, top-ranked term marked as Research Field



Noun-phrases from text and fields in Embedding Space

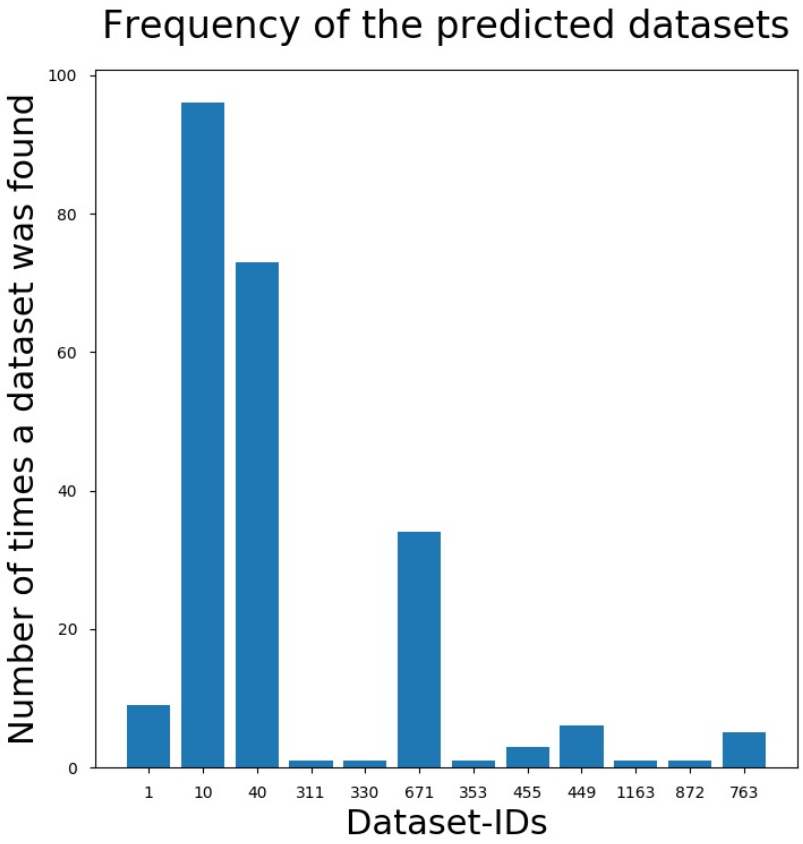
Section 3

Dataset Detection

Dataset Detection

Simple dataset mention search

- Unique datasets selected from Dataset metadata
- Datasets searched in publication text
- Frequency of dataset mentions calculated
- Removed mentions $>$ threshold-value $*$ median of dataset-frequency



Dataset Detection

Rasa-based Dataset Detection

- Trained Entity Extraction model using **Rasa NLU**[2]
  - Training data : 2500 labeled publications (phase-1 training data)
  - Training data : 7500 labeled publications (entire phase-1 corpus)
- Only those entities that
  - Had confidence-score > threshold-value
  - Belonged to the Research Field of the article

  considered as datasets.

[2] https://rasa.com/docs/nlu/

Dataset Detection

Combining the two approaches

- Removed irrelevant mentions from Rasa-identified datasets
- Took a union of results from two approaches

Section 4

Results

Quantitative Evaluation of Datasets against Validation data

	Rasa-based Approach (2500)	**Rasa-based Approach** (7500)	**Combined Approach** (2500)	**Combined Approach** (7500)
Precision	0.382	0.388	0.456	0.456
Recall	0.26	0.26	0.31	0.31
F1	0.309	0.311	0.369	0.369

Numbers in brackets indicate training samples

Improvement - 19.42%

Evaluation against Phase-1 holdout

pub_id	Keywords	Phase-1	Phase-2
10328	Cycling for transport, leisure and sport cyclists	Health evaluation	Public health and health promotion
7270	Older adult drug users, harm reduction	Health Education	Correctional health care
6053	Economic conditions - crime relationship, homicide	Homicide	Gangs and crime

Results

Methods

pub_id	Keywords	Phase-1	Phase-2
10328	Thematic content analysis	Thematic analysis	Sidak correction
7270	interviews conducted face to face, finding systematic patterns or relationships among categories identified by reading the interview transcript	Qualitative interviewing	Sampling design
6053	Autoregressive integrated moving average (ARIMA) time-series model	Methodological pluralism	Multivariate statistics

Section 5

Challenges Encountered and Future Agenda

Challenges Encountered

- Appropriate extraction of text from PDFs
  - Extraction of specific sections from text

Challenges Encountered

- Appropriate extraction of text from PDFs
  - Extraction of specific sections from text
- Not enough uniformity in labeled data to detect datasets

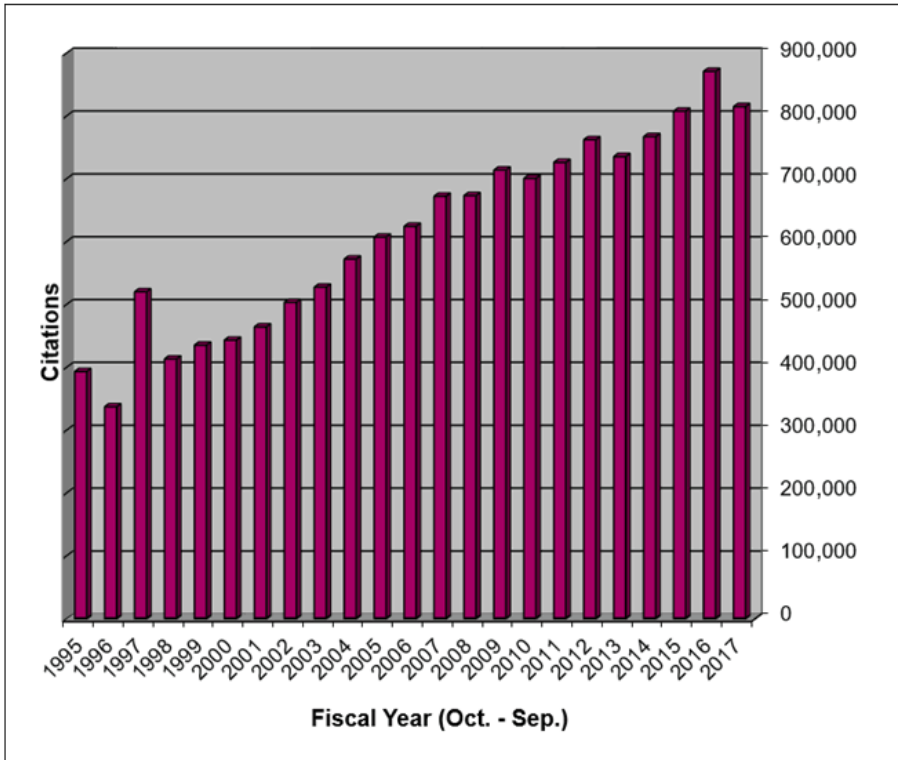
Challenges Encountered

- Appropriate extraction of text from PDFs
  - Extraction of specific sections from text
- Not enough uniformity in labeled data to detect datasets
- Some polysemous and high-level terms in the given SAGE vocabularies

Challenges Encountered

- Appropriate extraction of text from PDFs
  - Extraction of specific sections from text
- Not enough uniformity in labeled data to detect datasets
- Some polysemous and high-level terms in the given SAGE vocabularies
- Unavailability of Open Ontologies for Social Science Fields and Methods

Future Agenda



Number of indexed citations added to MEDLINE during each fiscal year since 1995[3]

[3] https://www.nlm.nih.gov/bsd/stats/cit_added.html

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending
- Distributed

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending
- Distributed
- Self-feeding (focused crawling)

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending
- Distributed
- Self-feeding (focused crawling)
- Self-repairing (introspection)

Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending
- Distributed
- Self-feeding (focused crawling)
- Self-repairing (introspection)
- Humans in the loop (weak supervision)

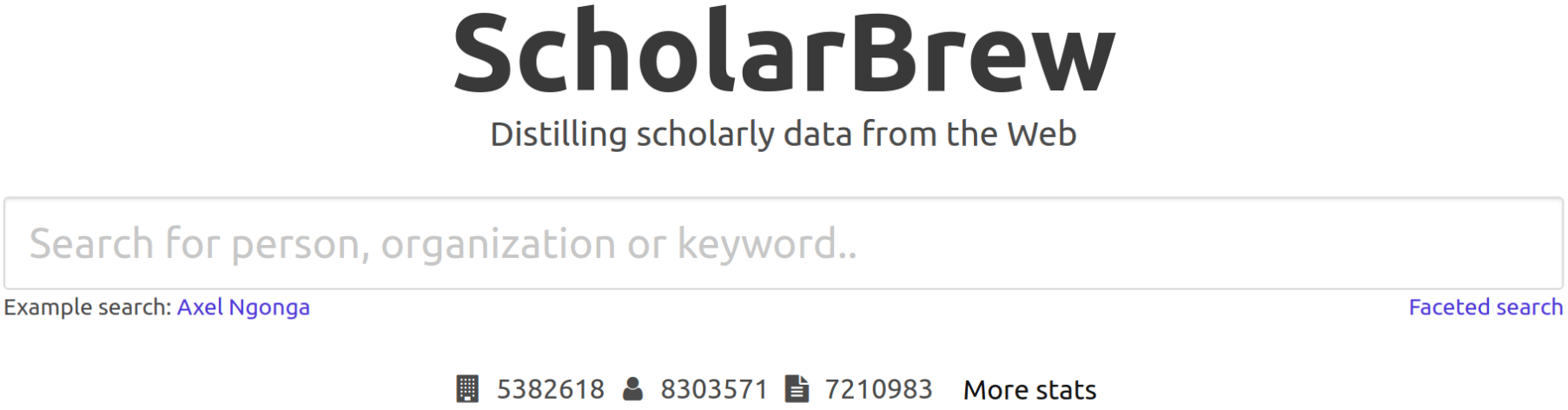
Future Agenda

Vision: Giant Global Scientific Knowledge Graph

- Never ending
- Distributed
- Self-feeding (focused crawling)
- Self-repairing (introspection)
- Humans in the loop (weak supervision)
- Standardized access (SPARQL, Linked Data Fragments, question answering, etc.)

Future Agenda

Exynize



Check out: s2.exynize.com

Our Team

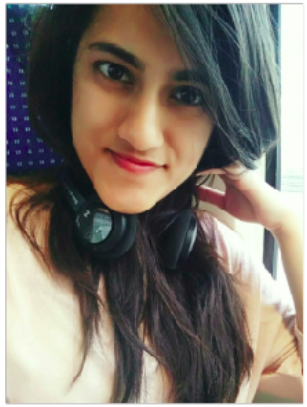
Prof. Dr.
Axel Ngonga

Dr. Ricardo
Usbeck

Michael
R¨oder

Daniel
Vollmers

Nikit
Srivastava



Rricha Jalota

Rene Speck

Connect

Thank You! Questions?

Follow us

@DiceResearch

github.com/dice-group

cs.uni-paderborn.de/ds/