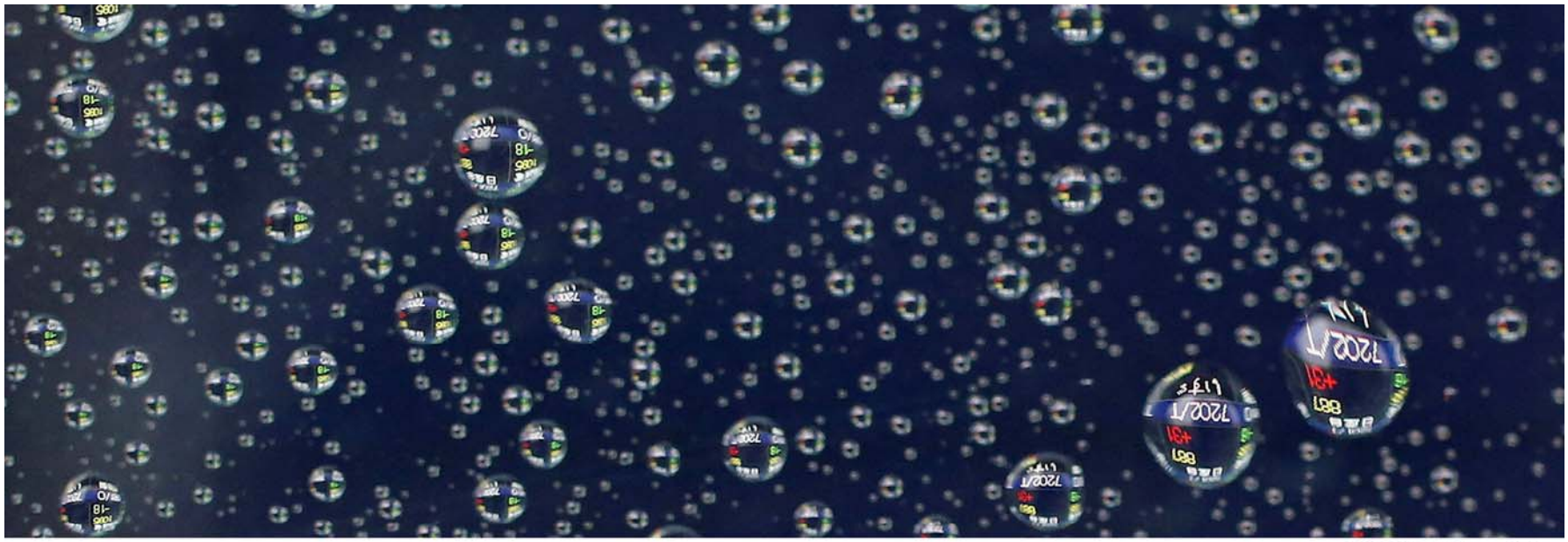


REUTERS/Yuya Shino

People Disambiguation Engine and Analysis of Russian Science in Web of Science (2007-2011)

Ilya Ponomarev, Victor Shyu, Pawel Sulima, Eva Darian, Ronan Sorensen, Brian Lawton, Ciaran Bolger, Etienne Godard, and Joshua Schnell

ilya.ponomarev@thomsonreuters.com

Custom Analytics, Rockville, MD

THOMSON REUTERS

Outline

1. **People disambiguation for research management: why it is important?**
2. **Custom Analytics at Thomson Reuters: PDE solution**
3. **Example: Russian Map of Science Project (Challenges, Solutions, Validation)**
4. **Enabling new analytical capabilities with new bibliometric indicators**

Name Disambiguation

Research Management DBs about events: Publications, Patents, Grants, Documents

Name Disambiguation: the process of detecting and correcting ambiguous *named entities* that represent the same real-world *object*

Synonyms: name unifications, matching, entity resolution, authorship

Approaches: supervised, unsupervised, semi-supervised, ORCHID
How to build scalable approach?

People Disambiguation for Research Management – Why It Is Important?

Improved accuracy

- **Productivity**
- **Citation Impact**
- **Research Area Interests**
- **Affiliations**
- **Co-authors**

Publication profile

- Recommendation engine of reviewers, conferences, grants, collaborations
- Conflict of interests

Research mobility tracking

- Hiring
- Grant success

Matching with other people DBs

Scientometric Indicators

- Normalization on distinct people count
- Collaboration
- Benchmarking

People Disambiguation Engine

Custom Analytics
Rockville, MD

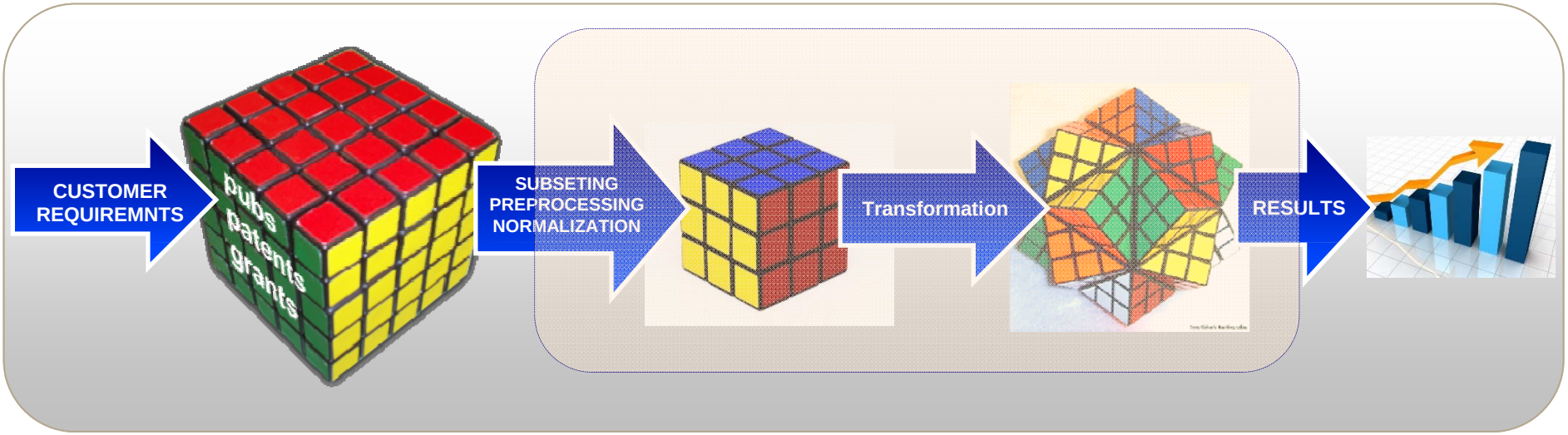
Projects Workflow: Custom Analytics Group

1. Project oriented
2. Scalability
3. Semantic and context-aware
4. DB-centric system

ScienceWire®

PDE Engine

New Knowledge



PDE: Vocabulary

Descriptor - a group of attributes describing a person

- Can be complicated (e.g. – a list of co-authors)
- Includes a link to source

PD1: Petrov A | Journal A | v 5 |p. 221 | Moscow|…

PD2: Polyakov M | Journal A | v 5 |p. 221 | St Ptbg|…

PD3: Smith J | Journal A | v 5 |p. 221 | London|…

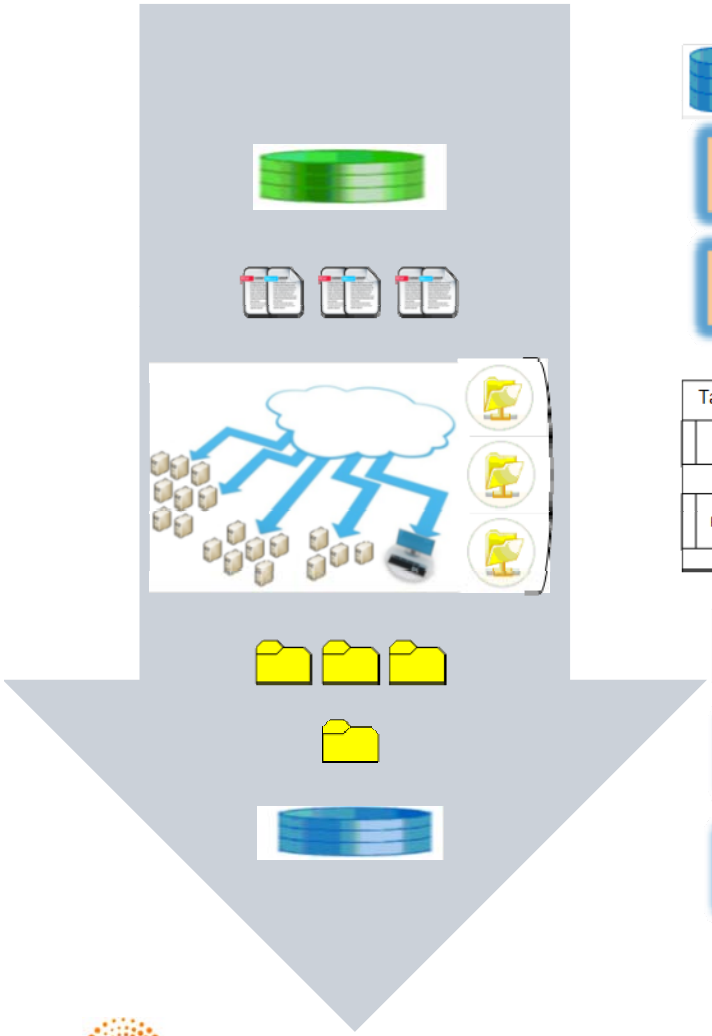
Pub1

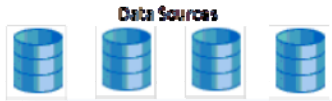
Descriptor pair - two descriptors linked together

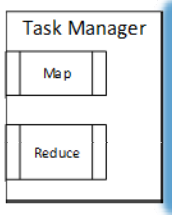
PD1	**PD2**
Bob	Robert
Smith	Smith
geneguy@umd.edu	GeneGuy@umd.edu
etc.	etc.

PD1	**PD4**
Bob	Bobby
Smith	Smith
geneguy@umd.edu	
etc.	etc.

OVERALL STRUCTURE







LOADERS

Transform source data into descriptor records

BLOCKERS

Generate block groups of pairs

EVALUATER

Calculate all similarity scores From descriptor pairs

Adjudicator

Assigns single pair similarity score

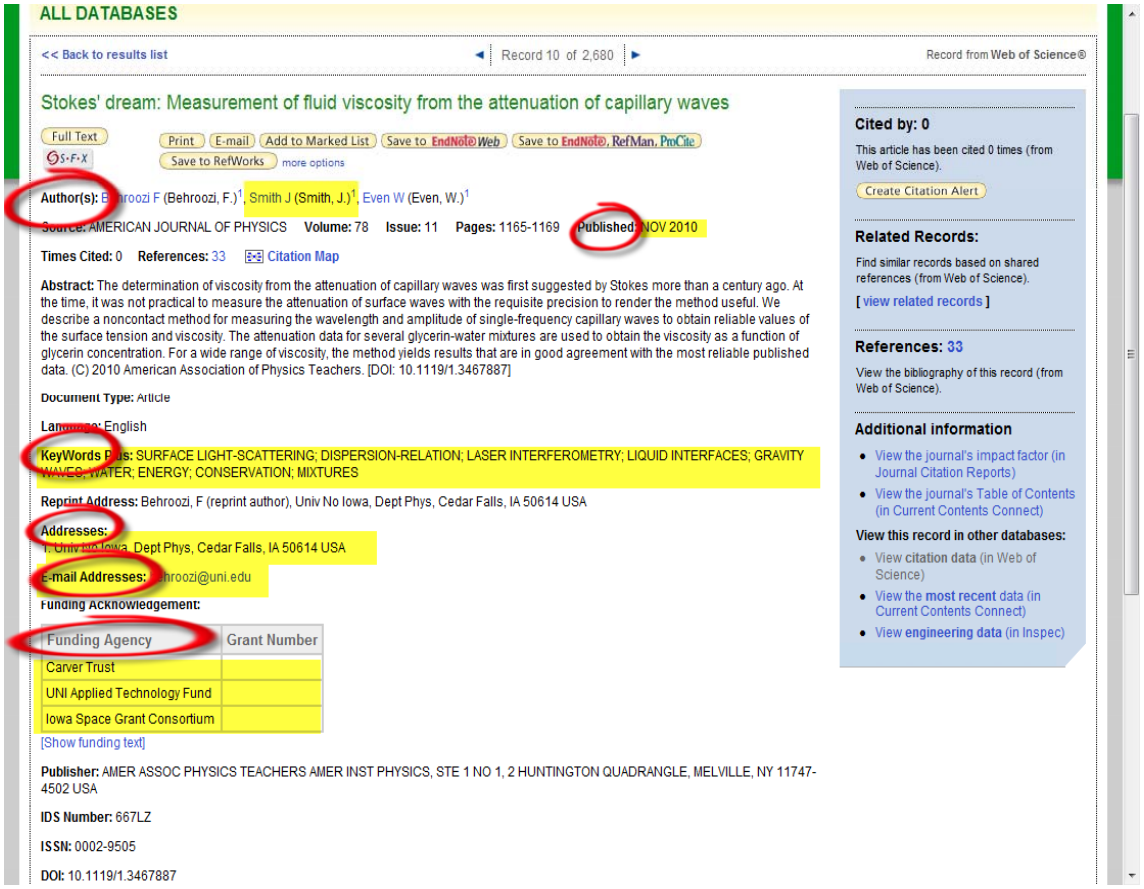
Ranker

Ranks match pairs

Stats Manager

Clustering

EVALUATORS: List of Comparisons Performed



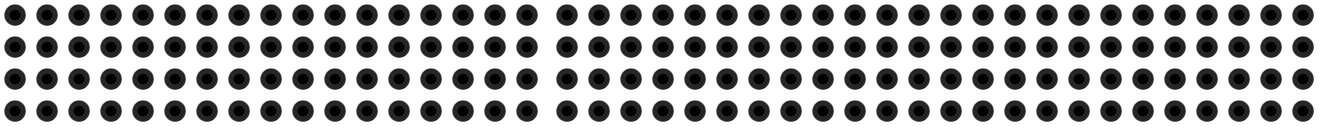
Attributes:

1. Names
2. Emails
3. Co-authors
4. Self-citations
5. Citations
6. References
7. Common words in titles
8. Subject categories (similar journals)
9. Affiliations

Summary of Process

Loader

People
Descriptors



Blocker

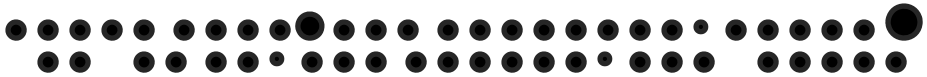
Descriptor Pairs

Evaluator

Descriptor Pairs
w/similarity scores

Validator/Aggregator

Reduced set of Descriptor
Pairs w/similarity scores



Adjudicator

Descriptor Pairs w/similarity
and match scores

Ranker

Descriptor Pairs w/similarity and
match scores in rank order

Stats Manager

PersonItems

RMOS PROJECT

RMOS Project Goals

- Create Map of the Most Productive Research Organizations in Russia

  (Russian Ministry of Edu and Sci, PWC, NICON)

  1. Collect all papers with at least 1 Russian affiliation in 2007-2011
  2. Disambiguate organizations in these papers
  3. Disambiguate people on these pubs

<u>Sounds simple</u>

Challenges

Author disambiguation in Publication DBs

1. Complexity: there are 180,000,000 descriptions of people in WoS (51000 years of calculations by brut force algorithm)
2. Orthographic and spelling name variants (author, publisher, OCR-DB soft)
3. Real name changes over time (marriage)
4. Incomplete data (First Name vs First Initial vs NULL)
5. Parsing errors (First name in MiddleName field)
6. Incorrect or ambiguous links to other attributes, incorrect attributes
7. Name commonality (Li, Wang, Ivanov)

Challenges

Real people, real full source documents

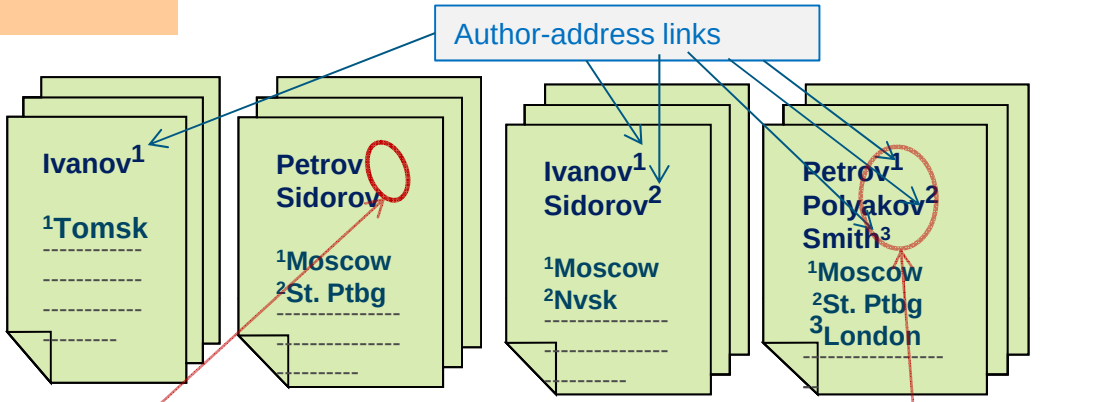

Ivanov #1


Petrov


Ivanov #2


Sidorov

1. Emails are not always good
2. Common name
3. Poison descriptors
4. Incorrect affiliations

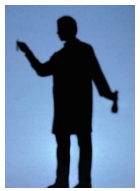


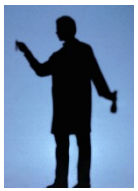
No links

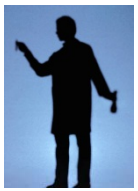
Typos, wrong links

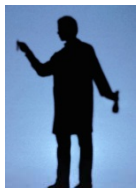
Missing links (90% before 2007)

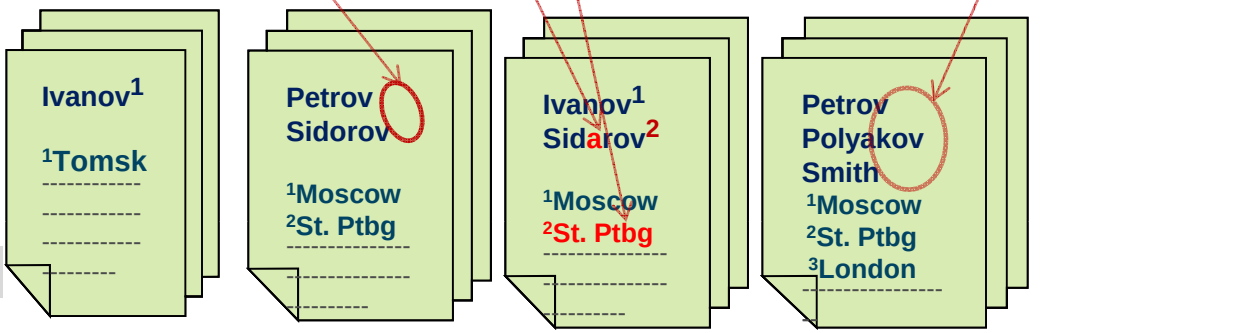
Web of Science


Ivanov


Petrov


Ivanov


Sidorov



Project Specific Challenges: Preprocessing Organizations Unification and Ambiguous links

Dusamb Org	# addresses	# WoS orgs
Moscow MV Lomonosov State Univ	12352	113
St Petersburg State Univ	4072	53

Ambiguous links recovery

Year	Total Number of publications	Number pubs without Author-address link	% without Author-Address link (before)	Number pubs with recovered links	Recovery rate, %
2007	31.0K	24.6K	79%	12.6K	52%
2008	33.3K	5.3K	16%	5.0K	96%
2009	33.3K	5.4K	16%	5.2K	98.8%
2010	31.9K	4.8K	15%	4.7K	98.7%
2011	32.5K	4.6K	14%	4.5K	98.7%
Total	**161.9K**	**44.6K**	**27%**	**32.0K**	**72%**

Strong recovery rate in 2008-2011

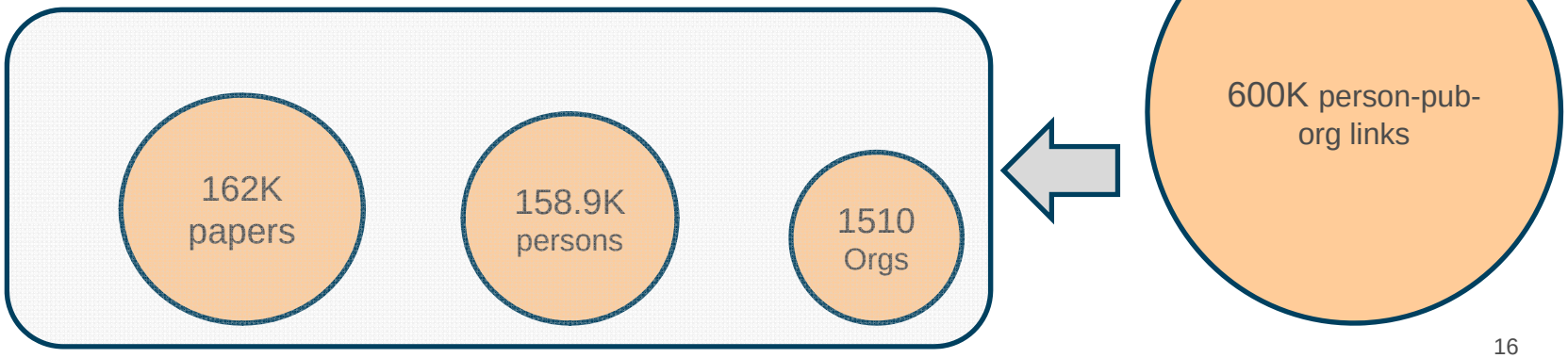
RMOS Project Results

161,870 papers

- Without address links
- 1,687,204 authors
- Brut force would be
- 1647 days

- With address links
- 548,688 Russian authors
- Brut force would be
- 173 days

- 98980 Blocks
- 15M pairs
- Evaluator time
- 4 hours



Validation

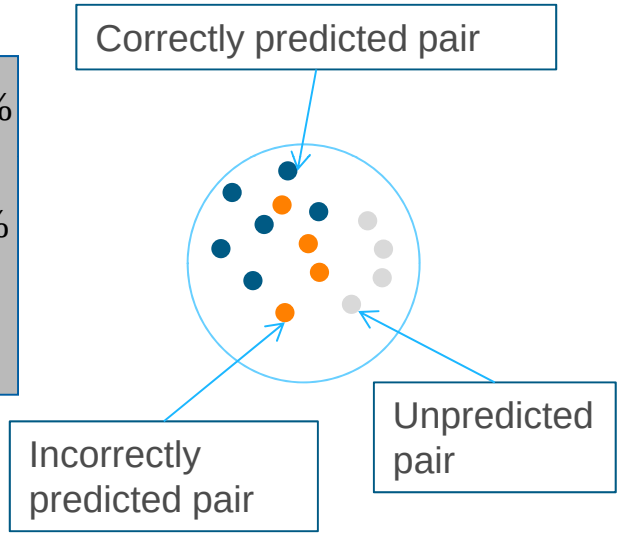
Choosing threshold: precision/recall/accuracy tradeoff

Evaluation measures:

$$\text{Pairwise Precision} = \frac{\#\ pairs\ correctly\ predicted\ to\ the\ author}{\#total\ \ pairs\ \ predicted\ to\ the\ author} = 96\%$$

$$\text{Pairwise recall} = \frac{\#\ pairs\ correctly\ predicted\ to\ the\ author}{\#total\ \ pairs\ \ to\ the\ author} = 90\%$$

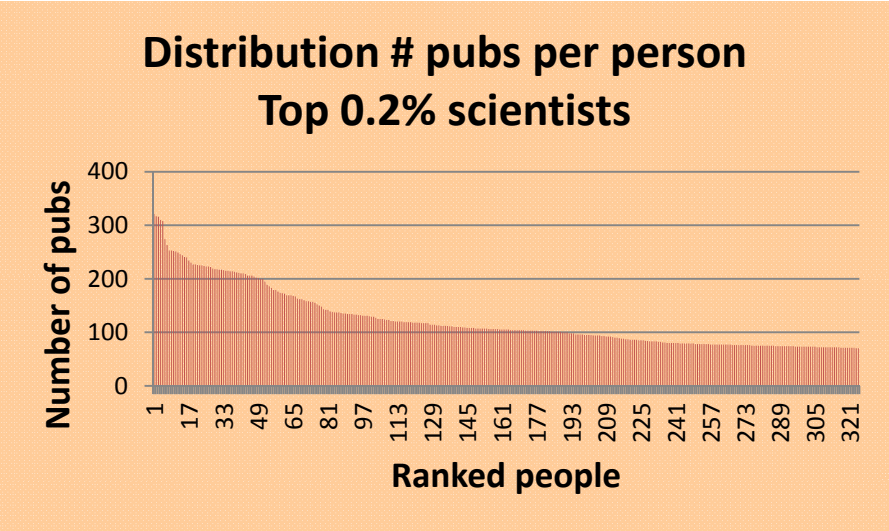
$$\text{Pairwise } F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

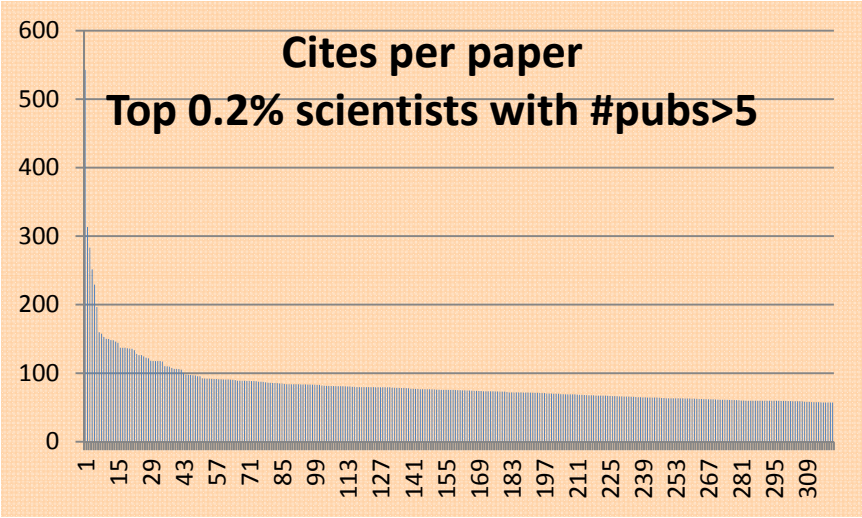


Caveat: Error is larger in common name clusters

- Lumping – assigning articles by different people to the same person
- Splitting – assigning articles by the same person to different people

New Knowledge – Benchmarking the Whole Country Distributions 2007-2011



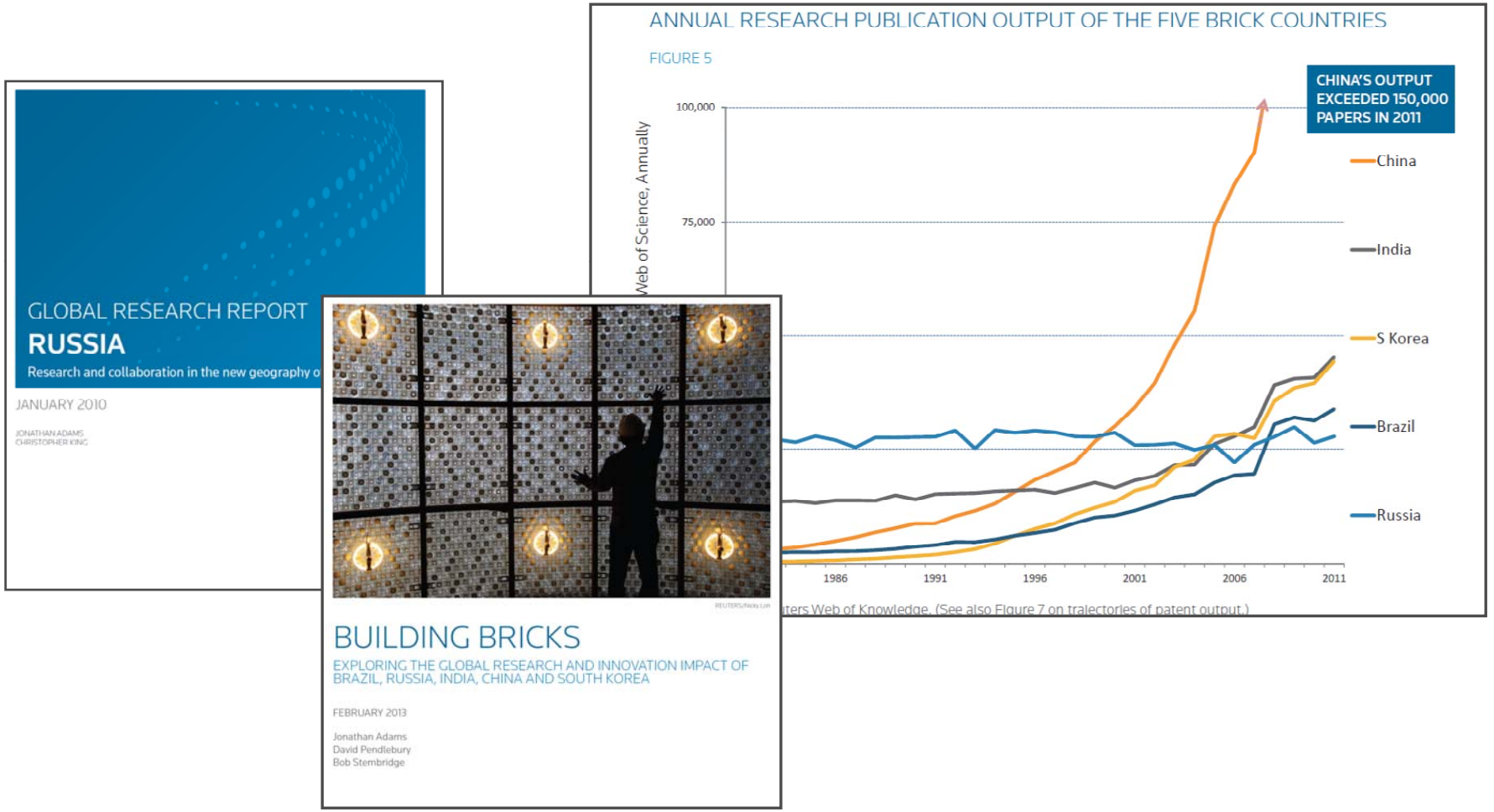


Having 1 pub – 86.5K
Habing 2 pubs – 24.7K
Habing 3 pubs – 12.6K
Habing >3 pubs –35K

Max papers – 321
Top 1% - 29
Average - 3
Median - 1

Max cites_pub – 543
Top 1% - 62.
Average - 4.5
Median - 1.

Thank You!


GLOBAL RESEARCH REPORT
RUSSIA
Research and collaboration in the new geography o
JANUARY 2010
JONATHAN ADAMS
CHRISTOPHER KING
BUILDING BRICKS
EXPLORING THE GLOBAL RESEARCH AND INNOVATION IMPACT OF BRAZIL, RUSSIA, INDIA, CHINA AND SOUTH KOREA
FEBRUARY 2013
Jonathan Adams
David Pendlebury
Bob Stembridge
ANNUAL RESEARCH PUBLICATION OUTPUT OF THE FIVE BRICK COUNTRIES
FIGURE 5
CHINA'S OUTPUT EXCEEDED 150,000 PAPERS IN 2011
Web of Science, Annually
100,000
75,000
China
India
S Korea
Brazil
Russia
1986
1991
1996
2001
2006
2011

Contacts

- Joshua Schnell joshua.schnell@thomsonreuters.com
- Victor Shyu victor.shyu@thomsonreuters.com
- Ilya Ponomarev ilya.ponomarev@thomsonreuters.com