Deduplication and Author-Disambiguation of Streaming Records via Supervised Models based on Content Encoders

Reza Karimi, Elsevier

#EUai2

Outline

- **Introduction: trends in consuming peer-reviewed publications**
- Scopus: deduplication and disambiguation needs
- Supervised Author Disambiguation
- Content Encoders
- Streaming and Delta Algorithm
- Conclusion

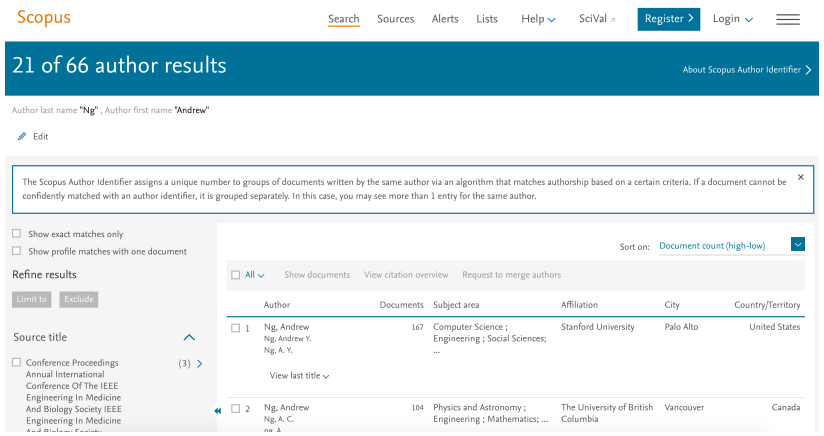
Elsevier: what you may know

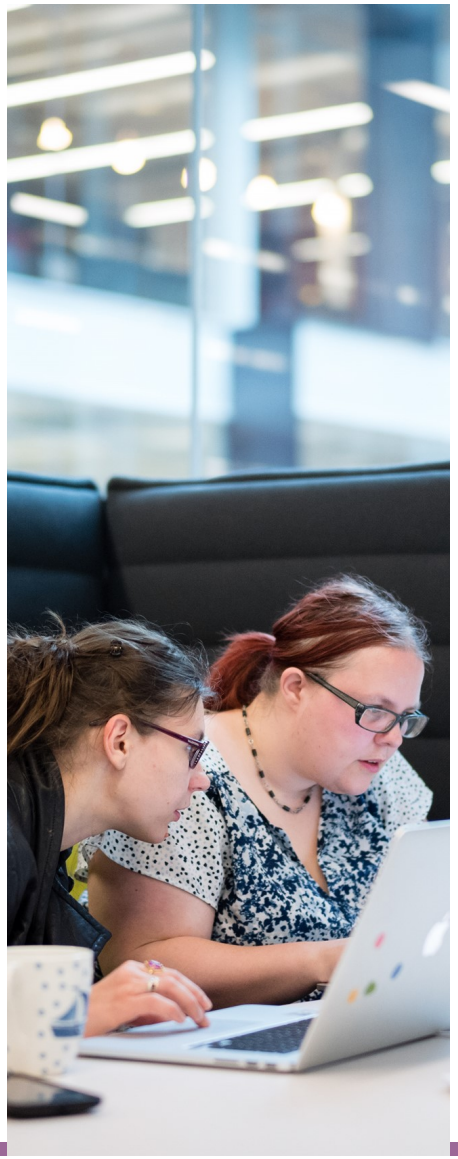
ELSEVIER

ScienceDirect ScienceDirect ScienceDirect ScienceDirect makes sense. ScienceDirect

enceDirect ScienceDirect | Scopus ScienceDirect ScienceDirect Science Direct

ScienceDirect ScienceDirect ScienceDirect





Who We Are

A global

information analytics business

specializing in science and health.

Why We Do It

We help you solve your challenges, for the benefit of humanity

What We Do

We help institutions and professionals progress science, advance healthcare and improve performance.

A Unique Combination

Combine content with technology,

supported by operational efficiency,
to turn information into actionable knowledge.

Elsevier combines content with technology to provide actionable knowledge

Read *this*	**Search** *this*	**Do** *this*
Cell Fundamentals	ScienceDirect	Sherpath
Gray‘s Anatomy	Scopus	Mendeley
	ClinicalKey	Knovel
	Reaxys	

Parallel Big data Processing (in cloud) is a precursor for Advanced ML

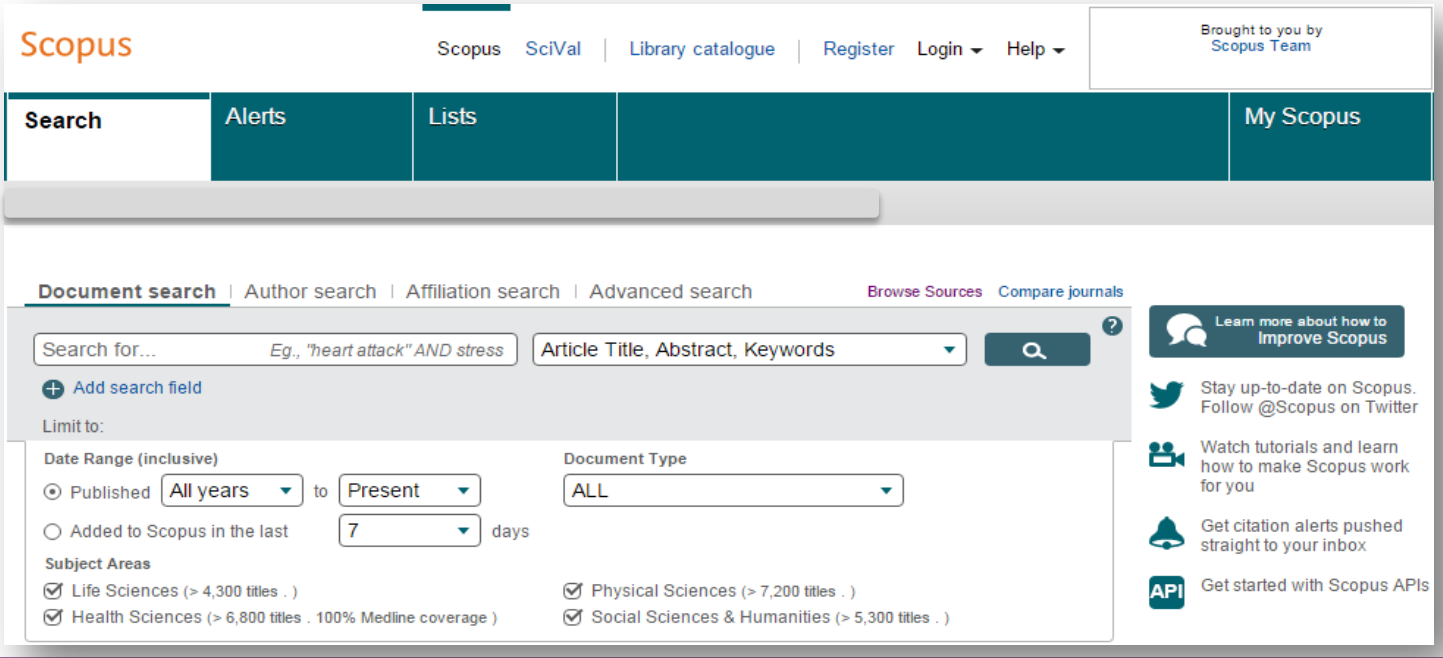
- Jeff-Dean a main innovator of Map-Reduce technique in 2004
- 2006 AWS Elastic Cloud Marketed
- 2008-2012 Google cloud started
- Running the Google Brain Project. This project started in 2011 (now open sourced as Tensorflow)

Outline

- Introduction: trends in consuming peer-reviewed publications
- **Scopus: deduplication and disambiguation needs**
- Supervised Author Disambiguation
- Content Encoders
- Streaming and Delta Algorithm
- Conclusion

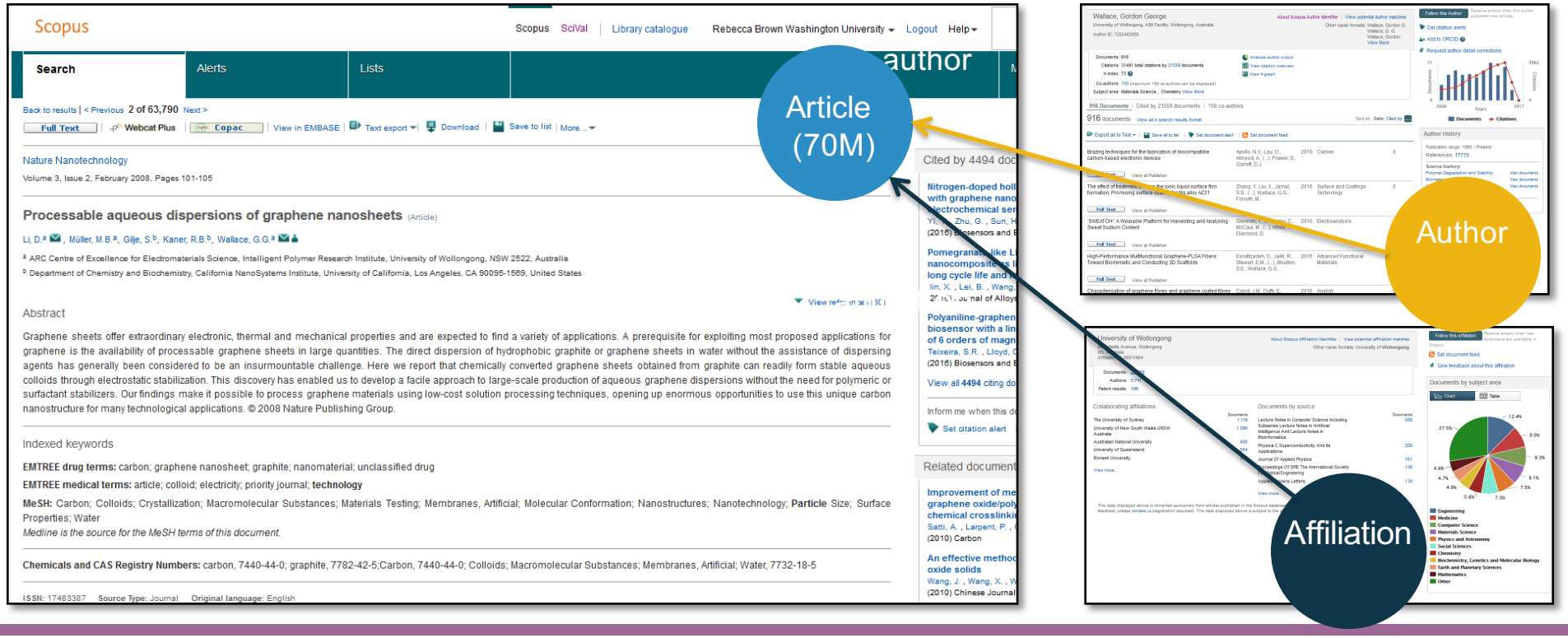
What is Scopus?

Scopus is the largest abstract and citation database of peer-reviewed literature, and features smart tools that allow you to track, analyze and visualize scholarly research.



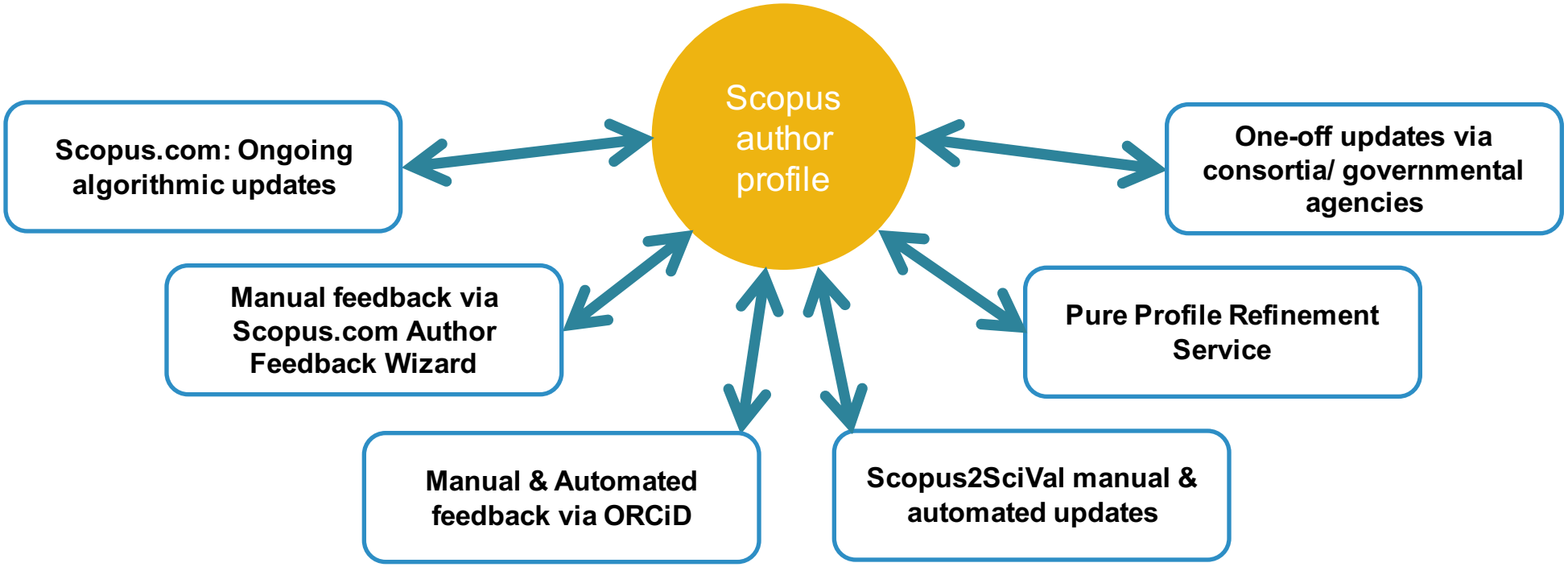
Scopus Data Model

By employing the state of the art disambiguation and deduplication algorithms, entities such as authors, institutes and cited document are disambiguated. This enables us to to analyze trends and to track researchers and institutes.



Author Disambiguation:
AI algorithms enhanced with multi-level feedbacks

Scopus use a combination of automated and curated data to automatically build robust author profiles, which power the Elsevier Research Intelligence portfolio.



Scopus Feedback Scopus | ORCiD SciVal Pure

Outline

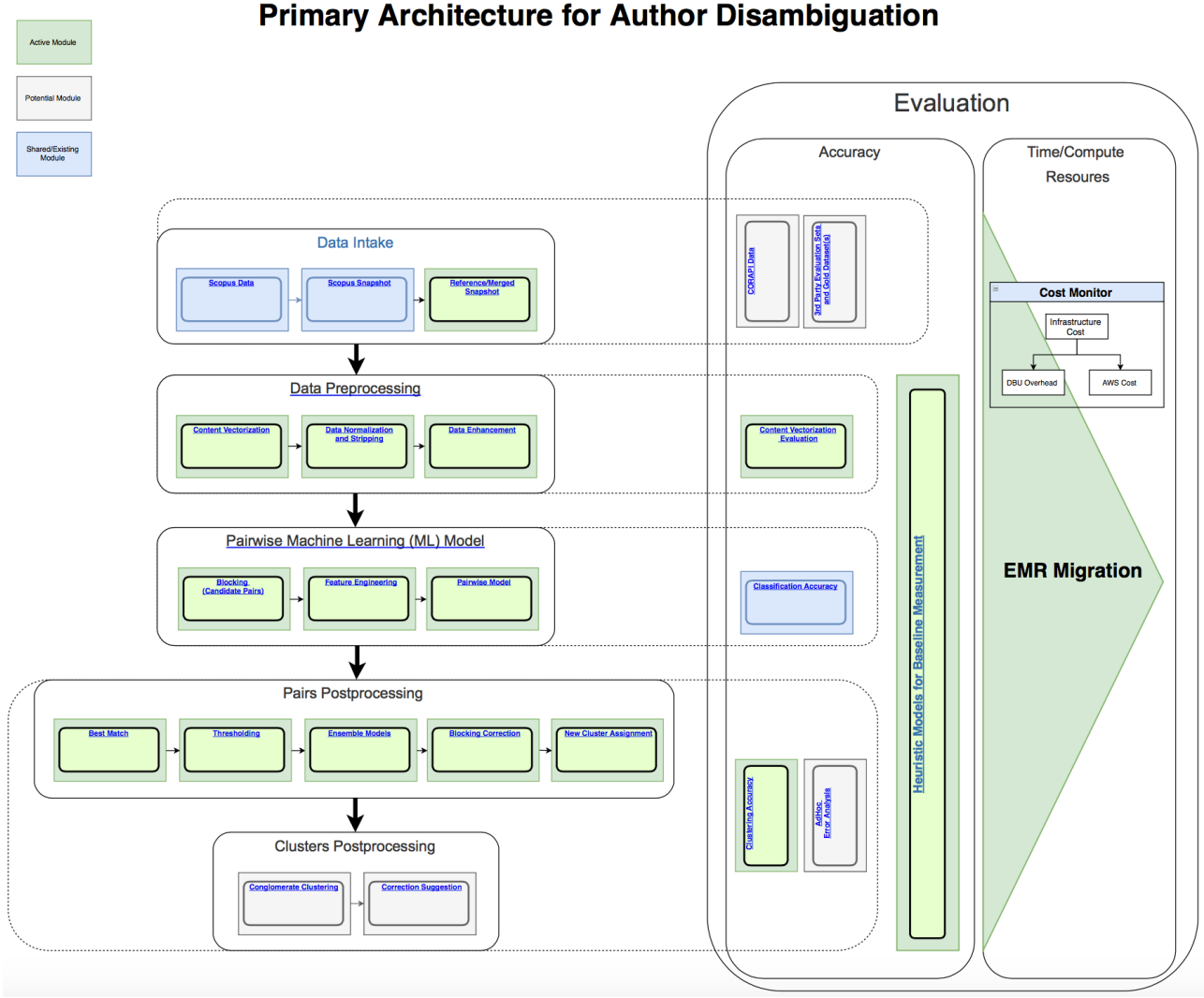
- Introduction: trends in consuming peer-reviewed publications
- Scopus: deduplication and disambiguation needs
- **Supervised Author Disambiguation**
- Content Encoders
- Streaming and Delta Algorithm
- Conclusion

Deduplication or Disambiguation Principles

- Classification problems:
  - deduplication: mostly identical documents except for missing fields
  - author-disambiguation: articles with (some rare) similar fields

- Generic supervised approach (due to N-squared complexity):
  1. create candidate pairs (by blocking rules or min-hash both limiting recall)
  2. in each block apply a pairwise model binary classification
  3. aggregate pair links via connected nodes algorithm to find full-clusters

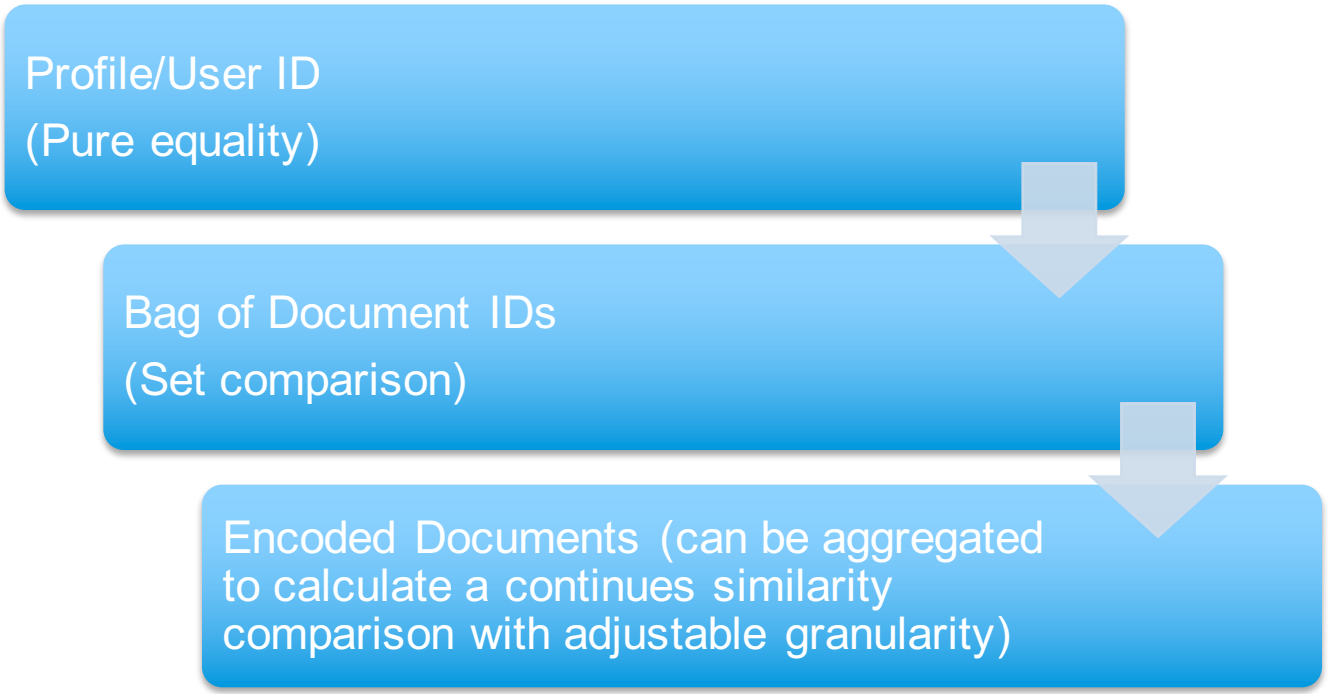
Primary Architecture for Author Disambiguation



Outline

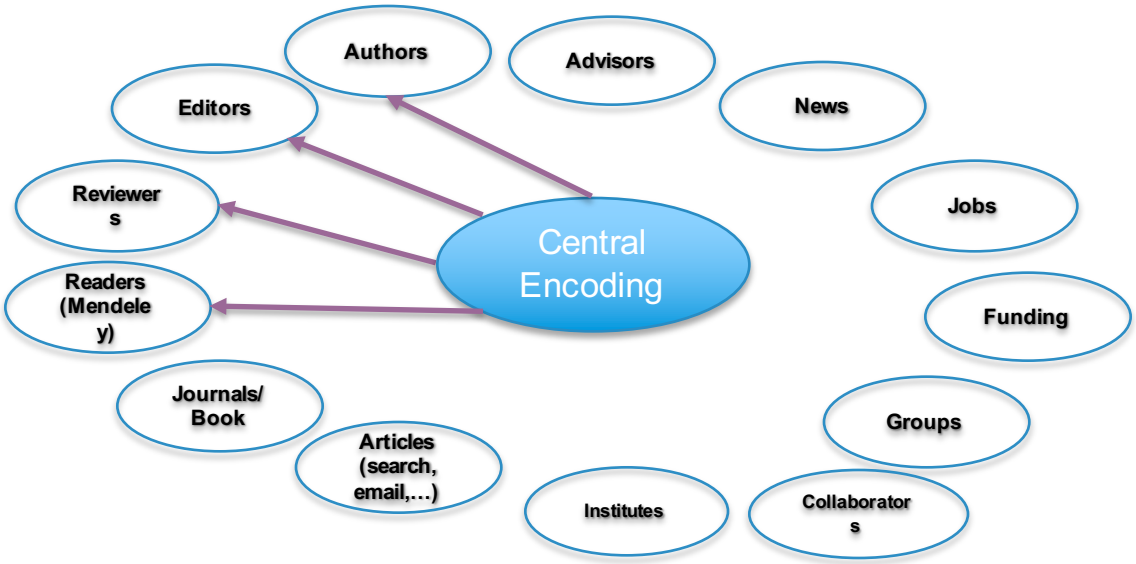
- Introduction: trends in consuming peer-reviewed publications
- Scopus: deduplication and disambiguation needs
- Supervised Author Disambiguation
- **Content Encoders**
- Streaming and Delta Algorithm
- Conclusion

Bag of Documents/Items: Standard Entity Representation



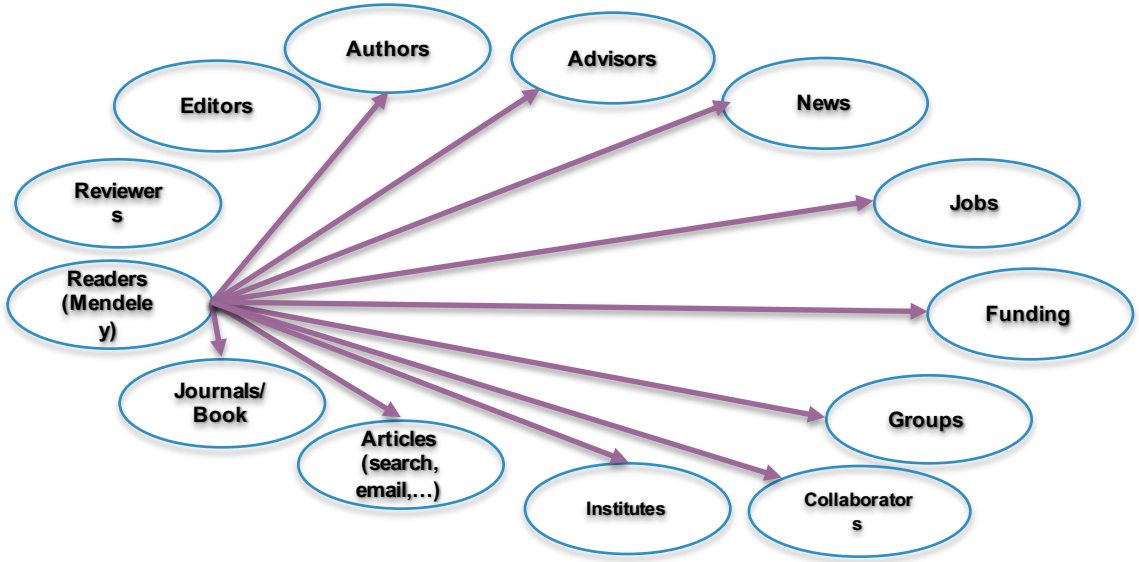
Encoding: Common Data Language across All Entities

- A fingerprint hub can act as the central hub to maintain (semantic) relation across entities.



Recommendation and Personalization

- When you adopt a unified login, you are expected to personalize your offerings further. This requires relating actions in one product to another.



- You may lack recommendation between most of our entities. It is not feasible to maintain or build specific recommendation one at a time
- Instead we can have a relatively general recommender architecture which can connect any two entities based on their standard encoding. ***To achieve that we need a standard entity representation***

Efficient Document Encoding

- Traditionally, TFIDF has been used to build a feature set for text pieces. However:
  - as a hot encoder, lacks semantic similarity
  - not suitable for streaming calculations
  - Sparse representation of nK bits
    - Has to be converted to a dense vector for most ML libraries
    - When aggregated across many documents misses sparsity and creates a big long tail
    - Not efficient for real time comparison
- Another alternative is to use defined ontologies/classes (pre-defined trees). However:
  - Lack of similarity/distance
  - Update
  - Interdisciplinary work
  - Rigid structure and lack of custom hierarchical levels
- Instead, we use word2vec algorithms to a encode each piece of text
- Spark implementation can achieve a parallel and fast gensim, but it would be better to decrease number of partitions or learning rate

Encoders in Action

id_facet	Facet Description	Product	
1	Scopus Authorship	Scopus	
2	Evise Editor	Evise	
3	Evise Reviewer	Evise	
4	Science-Direct FTA/View	Science-Direct	
5	Science-Direct Click	Science-Direct	
6	Science-Direct Search	Science-Direct	
7	Mendeley-Library	Mendeley	
8	Mendeley-Group	Mendeley	
9	Mendeley-Feed Click	Mendeley	
10	Average of Scopus Co-Authors	Scopus	
11	NSF Funding Proposal	FROS	
12	Citation Weighted Scopus-Authorship	Scopus	
13	Mendeley-Library Addition in the past 6 months	Mendeley	
14	Career/Job Search	ROS	
15	Documents mentioned in the news	Newsflow	
17	Twitter mentions	Plum	
18	Patent Authorship	LexisNexis	
…	…		

id_entity	Set_of_known_id_facets	….
00034382943809	{1, 3,10, 12, 15}	
09852308932508	{5, 6, 12, 2}	
….	….	

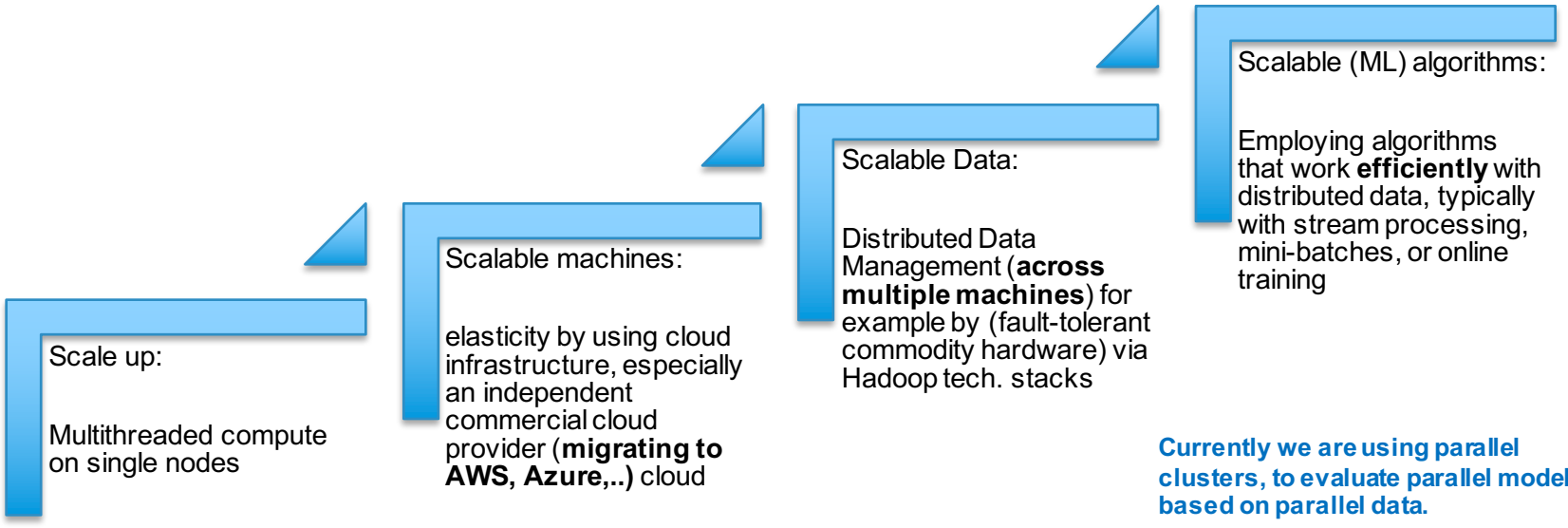
id_entity	id_facet	Vector_of_embedded_facet_value	Set_of_corresponding_document_or_url_id	….
00034382943809	1	[0.02, 0.45, 0.25, 0.03, 0.25]	{…}	
00034382943809	3	[0.18, 0.03, 0.09, 0.48, 0.22]	{…}	
….	….			

A suggested generic schema for our personalization hub with example facets:

Outline

- Introduction: trends in consuming peer-reviewed publications
- Scopus: deduplication and disambiguation needs
- Supervised Author Disambiguation
- Content Encoders
- **Streaming and Delta Algorithm**
- Conclusion

Scalable Design Stages: examine your pipeline maturity level



Currently we are using parallel clusters, to evaluate parallel models based on parallel data.

Models require minimal calculation as they are updated by passing through new data

Delta and Streaming Benefits

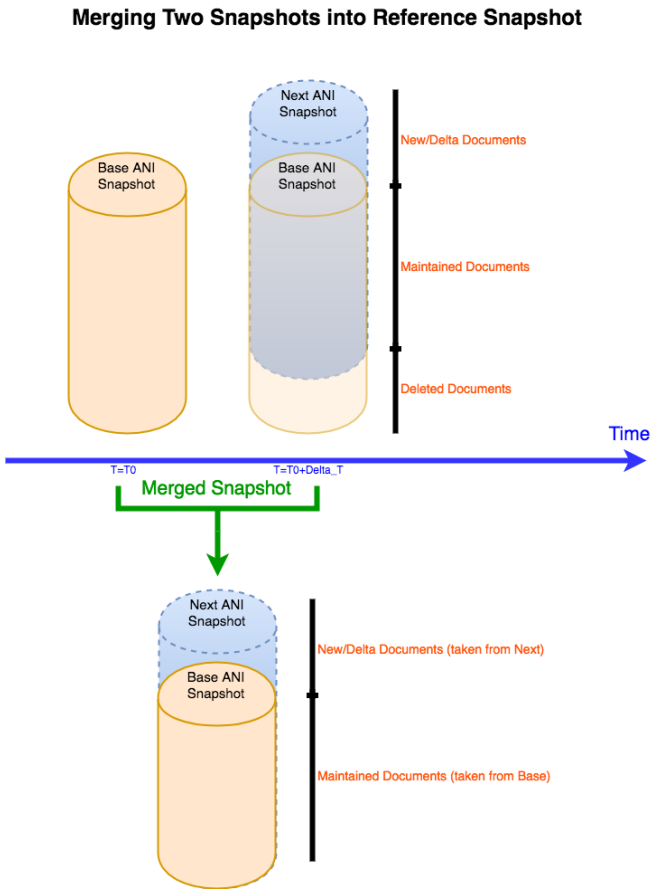
- Business continuity requires a smooth transitions from an algorithm to another (to minimize impact on historically claimed/corrected/used author profiles)

- Not all records are equal:
  - Newer documents offer richer meta-data and new users have higher expectation.
  - A model that is good for all times, will be suboptimal for new documents

- Streaming model saves on compute time. This essentially boils down to disambiguate one document at a time. This enables A/B testing by applying different algorithms to different documents.

Proper Construction of Pairs

- Ideally maintain similarity between Batch and Stream Processing
- Construct training pairs to best simulate information flow and maintain causality: all pairs have one side from delta documents
- Pair=(Document_Base, Document_Delta)
- For training pick random pairs, but for test and end-to-end measurement pick full blocks



Conclusion

- Cloud scalability and smart ML algorithms enables us to learn from millions/billions of data.
- New technologies enable us to literally read and understand collective human knowledge and offer products that go beyond search or simple Q&A.
- Understanding content can help to better serve customers and adds value to the products
- Author disambiguation and deduplications can be cast as a binary classification problem
- Word2vec family of encoders can offer a unified fixed-size content encoder across all products
- An efficient and scalable ML algorithm requires a delta/streaming implementation to maintain business continuity as well as infinite scalability

THANK YOU.

Feel free to reach me for further information or if interested to join our team:

r.karimi@elsevier.com

Acknowledgment:

Curt Kohler,

Bob Schijvenaars, Ronald Daniel

Elsevier members in BOS/Labs/Scopus teams

Who uses Scopus Data? Examples