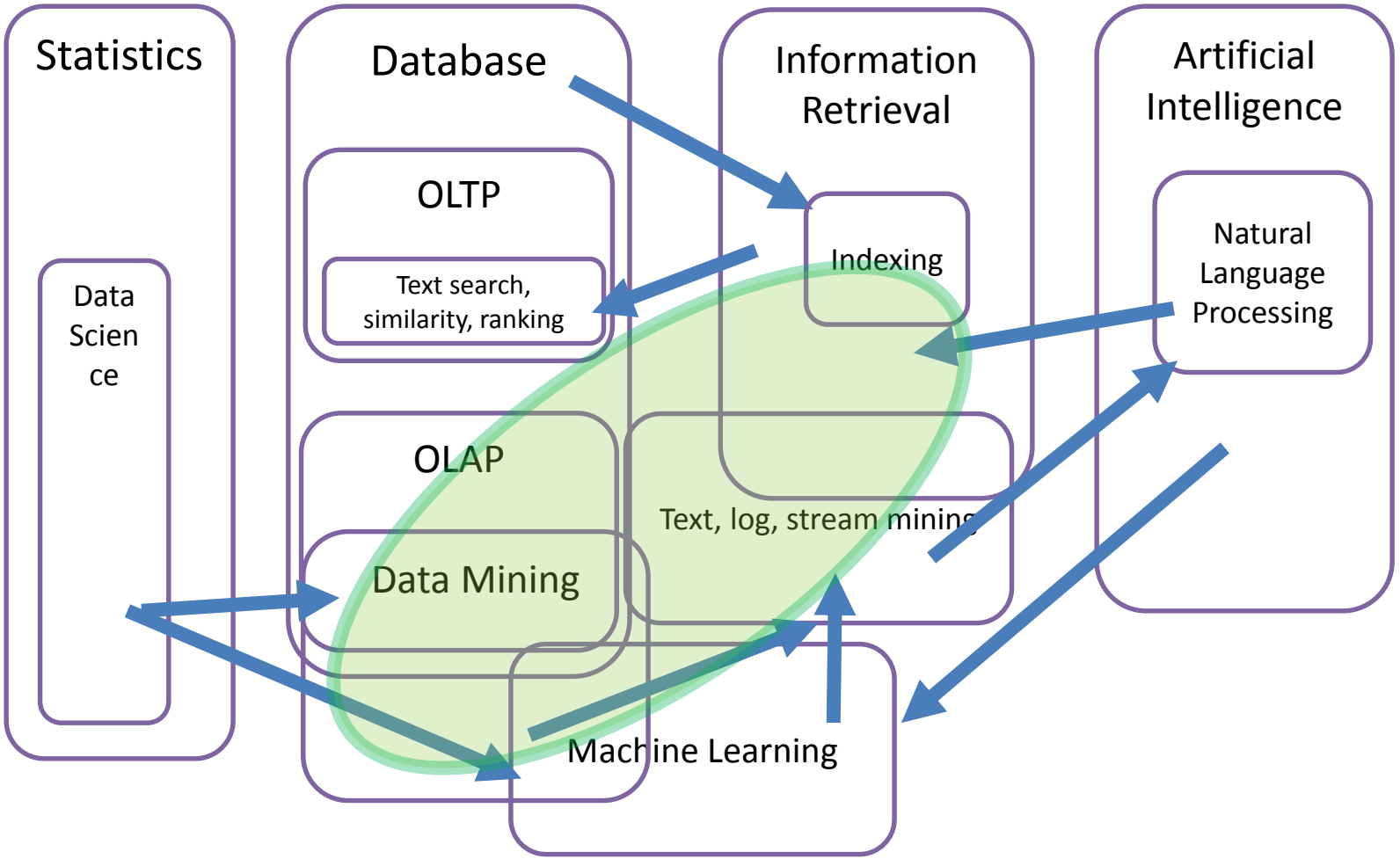
Search and Data Mining: Techniques

Search and Retrieval

Anna Yarygina

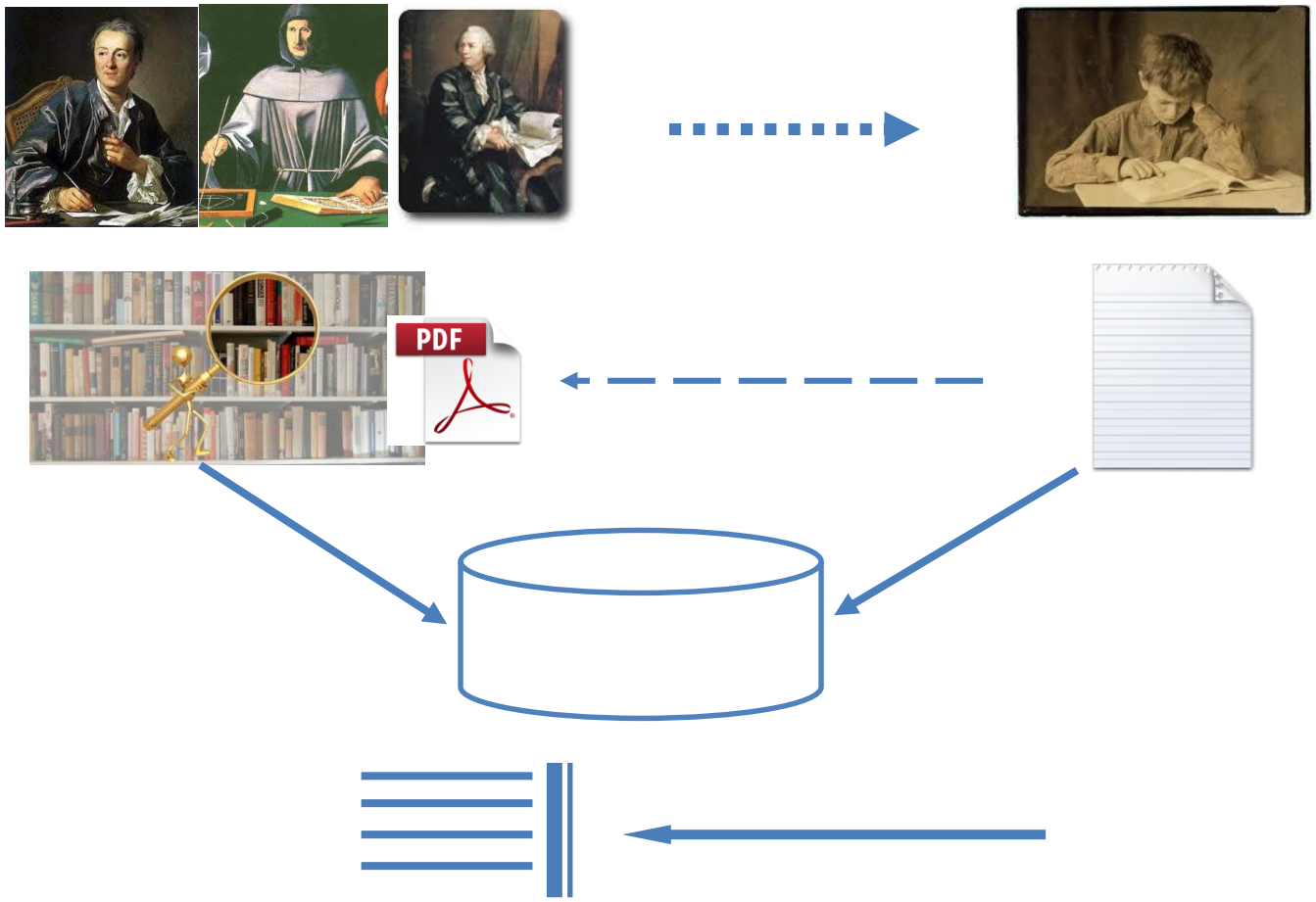
Boris Novikov

Balanced Viewpoint



The Problem of IR

- Document meaning vs. Information needs
- Document collection vs. Query
- Document image vs. Query image
- Index entry



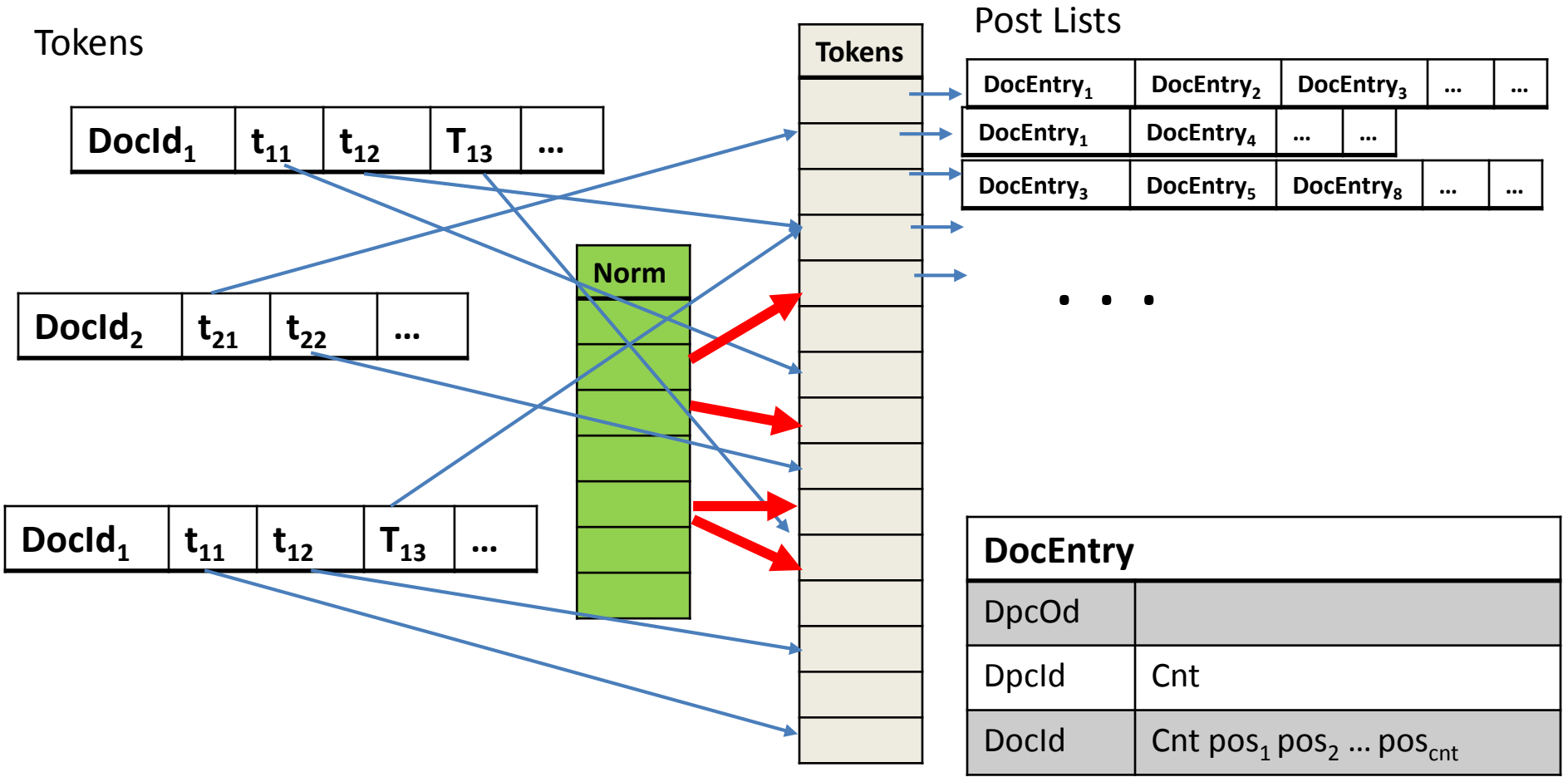
Quality Metrics

- Metrics
  - Precision: found_relevant / found
  - Recall: found_relevant / relevant
  - R-precision
  - More complex for uncertain models where relevance is not boolean
- Ok, but how to measure?
  - Human assessment is expensive and unreliable
  - Surprisingly, IR people trust it (there is nothing better)

IR Models

- Boolean
- Vector space
  - Term weights: term_frequency/document_frequency
  - Similarity metrics
    - ~~Euclidian~~
    - Cosine: $\Sigma x_i y_i$
- Probabilistic
- Ranks and Scores
  - Rank: position in the list of returned documents
  - Score: estimation of relevance

Inverted Files



Using Inverted File: Highlights

- Boolean model
  - Logical operators are implemented as set-theoretic operations on post lists
  - Start from shortest post lists, then use document term lists to finalize
- Vector model
  - Post Lists can be used to calculate DF and TF (if counts are stored)

Superimposed Coding (Signature Files)

- A signature is a bit vector of fixed length (say, 64)
- A term is hashed into a signature with fixed number of bits (say, 4)
- Document or query signature is OR-ed signatures of terms
- Sig(query) $<_{bitwise}$ sig(document)
- False positives

	Cat	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**
	rat	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**
$Document_1$	**Cat rat**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**0**
$Document_2$	**Cat eat**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**
$Query_1$	**rat**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**
$Query_2$	**big**	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**0**

Signature Files: Discussion

- 20+ years of research, several variations and improvements
- The rate of false positives rapidly grows with the size of document collection
- In most cases is worse than inverted files

Text Search in DBMS

- PostgreSQL
- Documents: text type (table column or expression)
- Document image: tsvector type
- Query image: tsquery type
- Search:
  - Matching (Boolean)
  - Actually, you have to write some code
  - Ranking
- Configurations define:
  - Parser and language
  - Dictionaries
- Indexes
  - gin (inverted file)
  - gst (signature file)

Improving the Quality

- Query expansion
  - Synonyms
  - Equivalents
- Relevance feedback
  - Expand the query with documents marked as relevant
  - Implicit: guesses based on user clicks
- Personalization
  - Requires detailed discussion, not in the scope today

Query Segmentation

Dell Latitude 14	Brand = ‘Dell’ Family = ‘Latitude’ Screen_size = 14
Dell Latitude 8 500	Brand = ‘Dell’ Family = ‘Latitude’ Memory = ‘8G’ HDD=‘500G’
Dell Inspiron Win 8	Brand = ‘Dell’ Family = ‘Inspiron’ Operating_system=‘MS Windows 8’

- May be sent directly to a database
- Improved quality of search