Europäisches Patentamt
European Patent Office
Office européen des brevets

Search Matters 2017

Towards semantic search at the European Patent Office



Alexander Klenner-Bajaja

About me

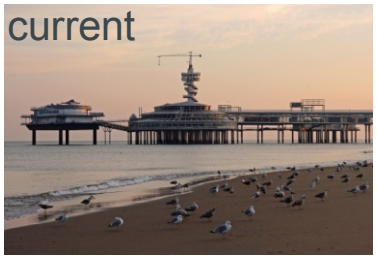
Alexander Klenner-Bajaja

- Bioinformatics at Goethe University Frankfurt

- PhD ETH Zurich & Goethe University –
  in Cheminformatics

- PostDoc at Fraunhofer Society SCAI in Cologne
  - Chemical entity recognition in patents

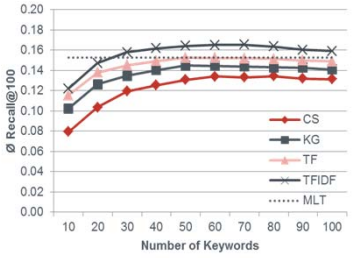
- Data Scientist, Search & Knowledge, DG2, EPO
  - automated search
  - search benchmarking
  - new search technologies



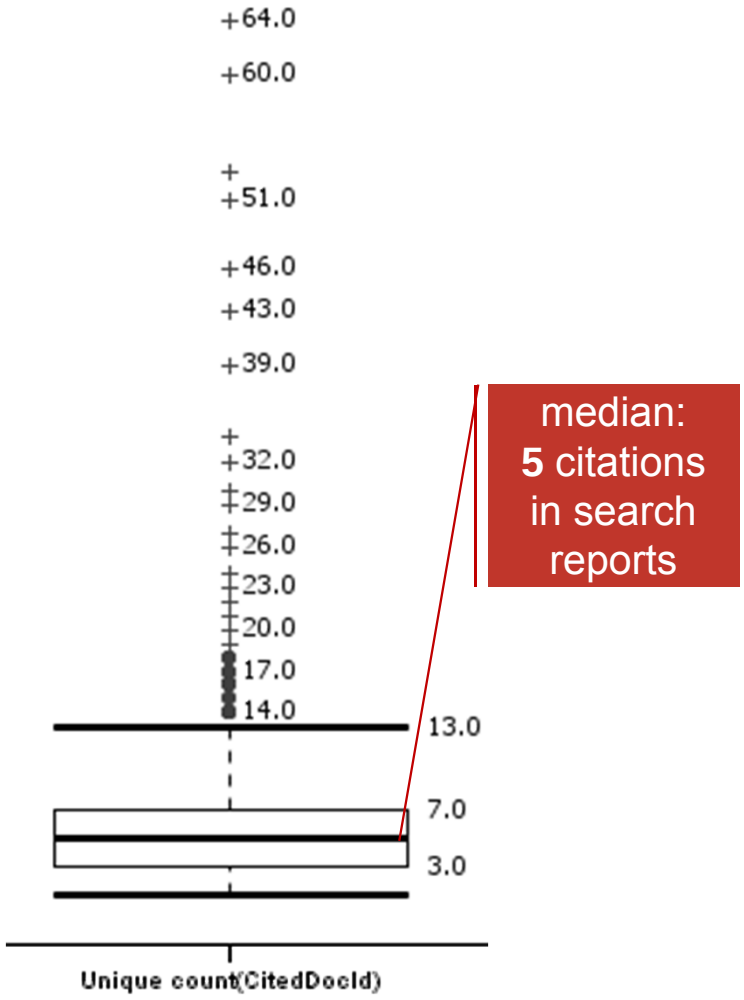
- Anecdotal Evidence and the need for a Benchmarking Environment
- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence

Anecdotal Evidence and Benchmarking

- How can two different search strategies be compared?
- Historically decision are taken based on expert feedback
- Valuable information that can be complemented with measurements

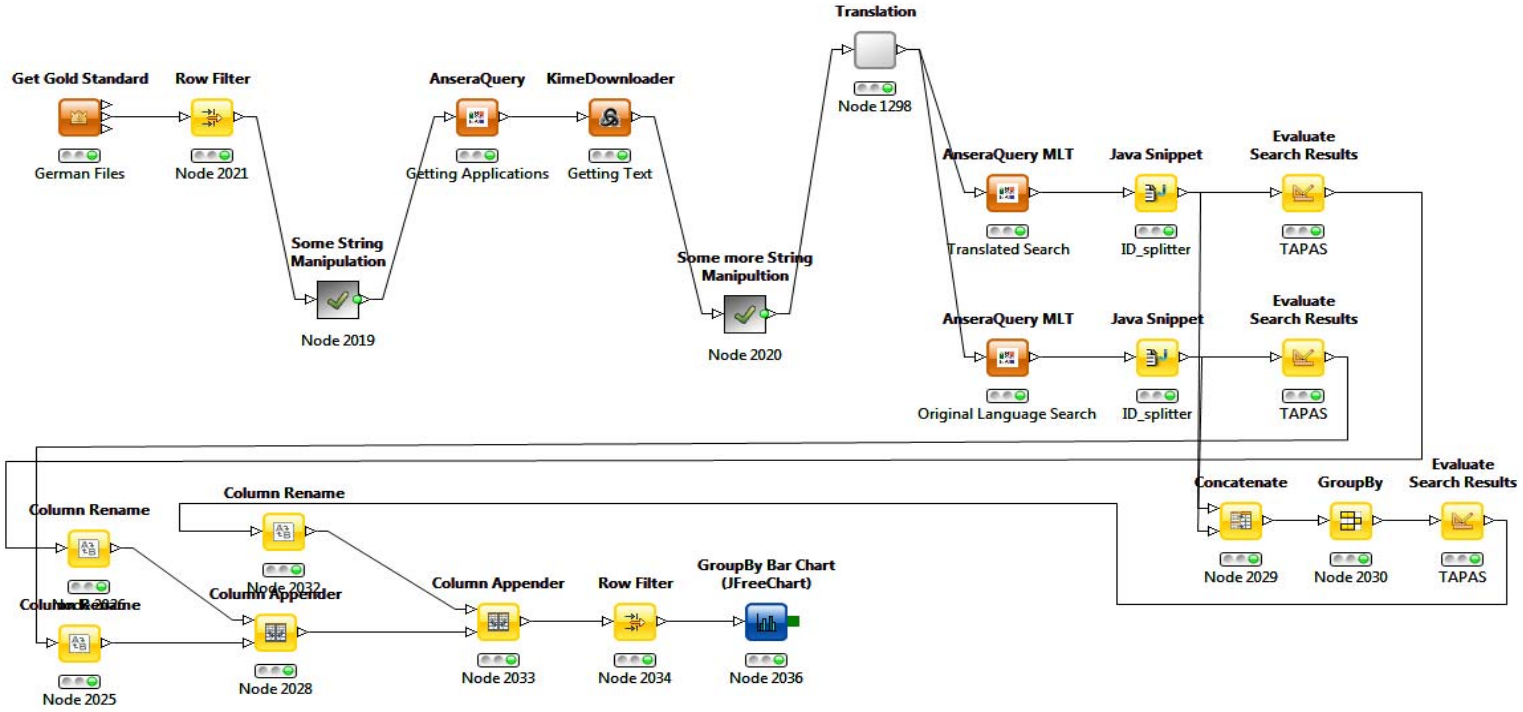


- Distribution shown as Box-plot



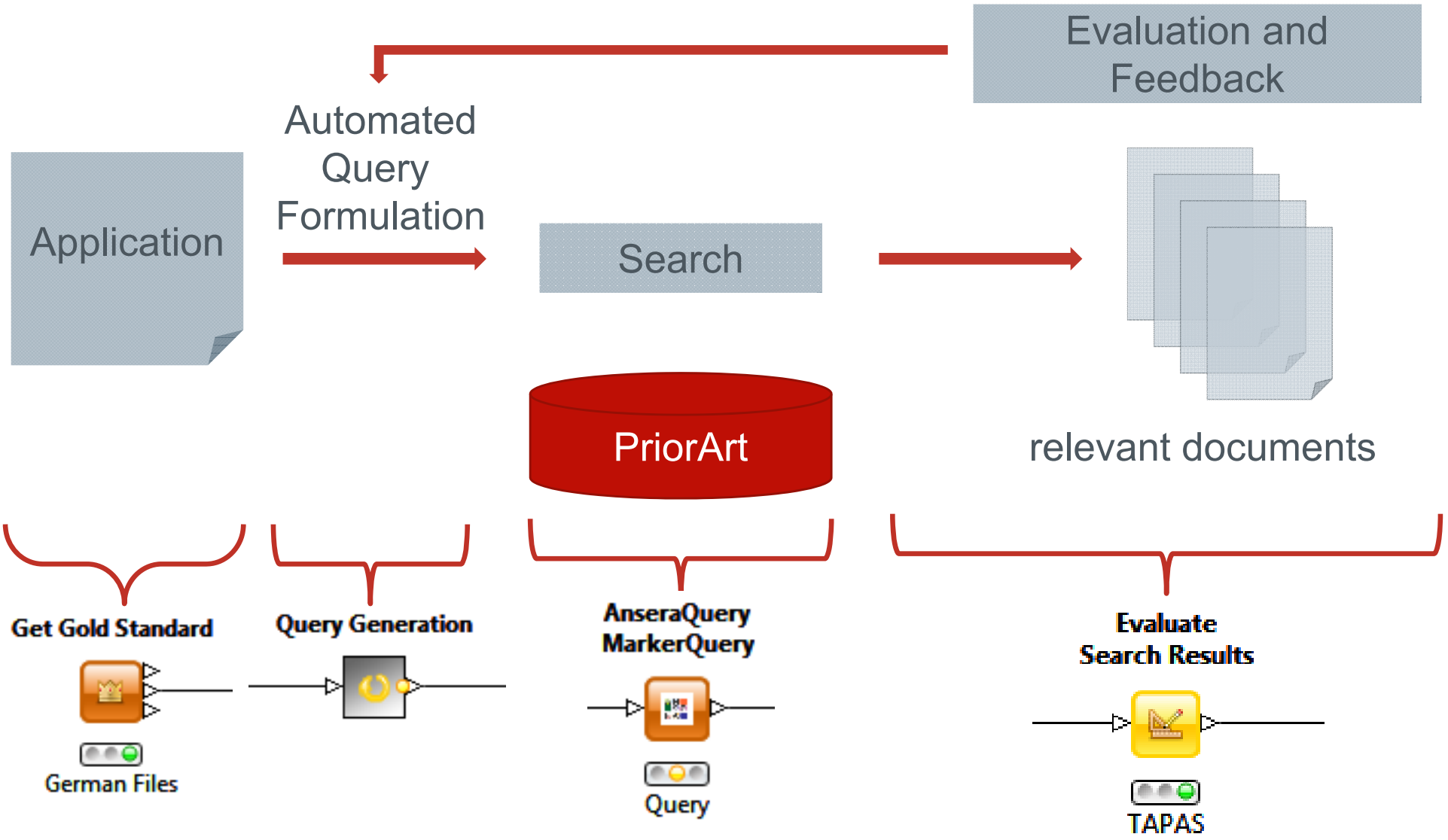
Search Benchmarking

- We implemented a holistic prototyping and benchmark system



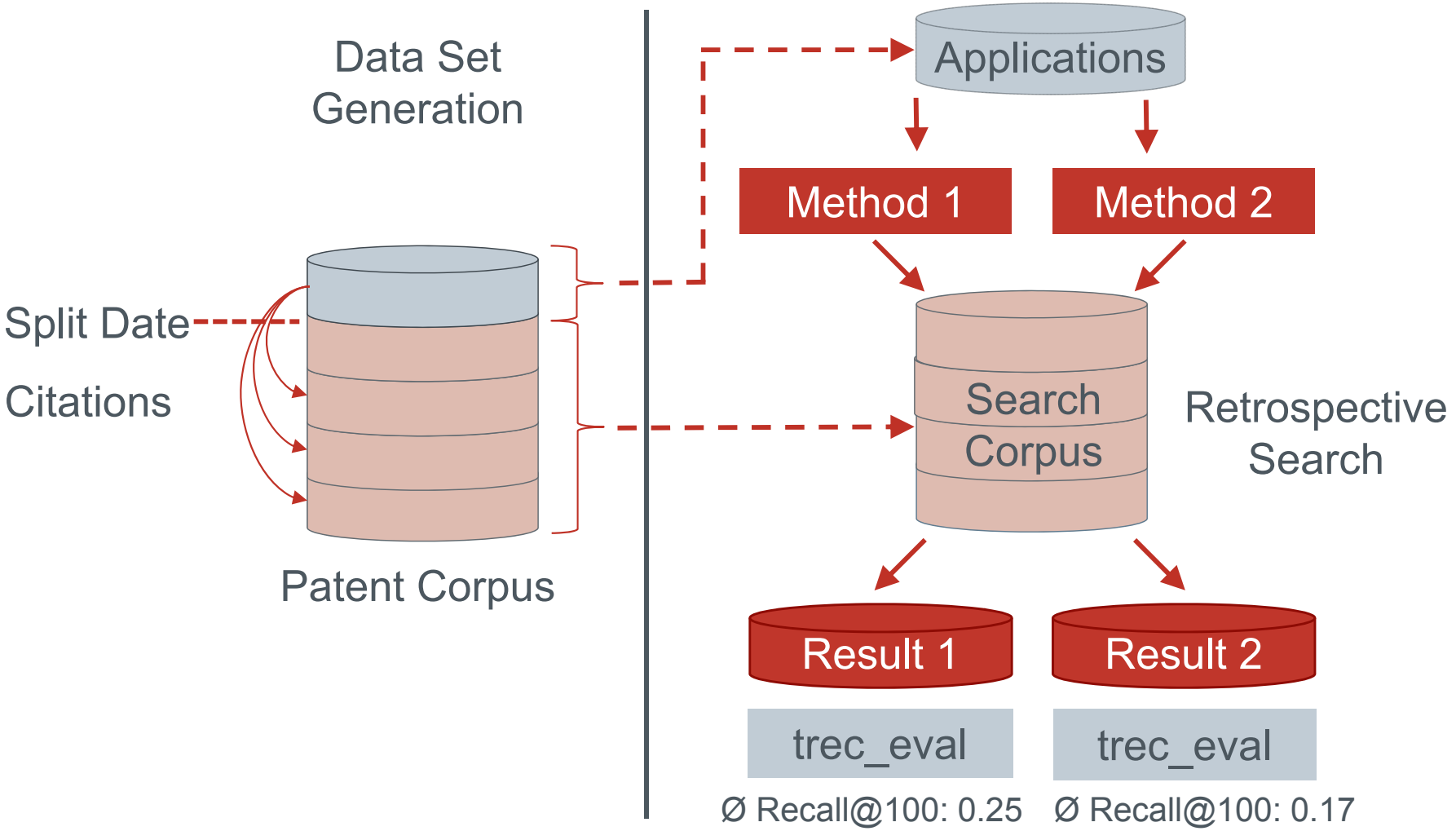
- We have access to all distributed system in the EPO IT landscape and can connect them in "visual" workflows

Automated Queries exploiting the shown tools



Benchmarking Environment

- From anecdotal to statistical evidence
- "Towards Automated Prior Art Search" (TAPAS)



Benchmarking Environment

- Retrieval success measured by different metrics
- Computed by the trec_eval tool

Top 50 Documents Returned

Cited Documents not Returned

Recall @ 50 $\frac{5}{10} = 0.5$

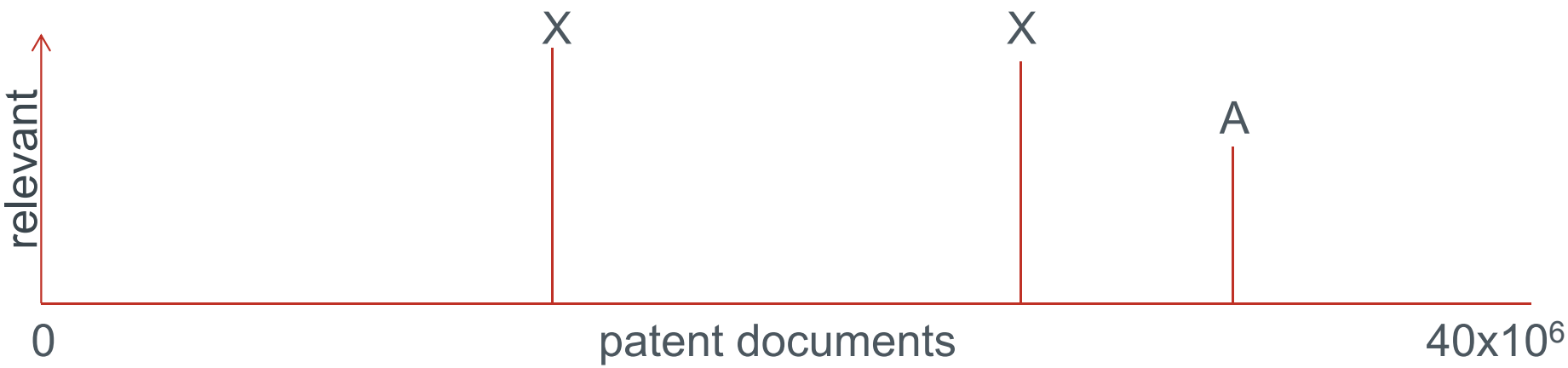
Hit Rate @ 50 $recall > 0? = 1$

- Average over all n simulated applications

$$\sum metric@X \Big/ n$$

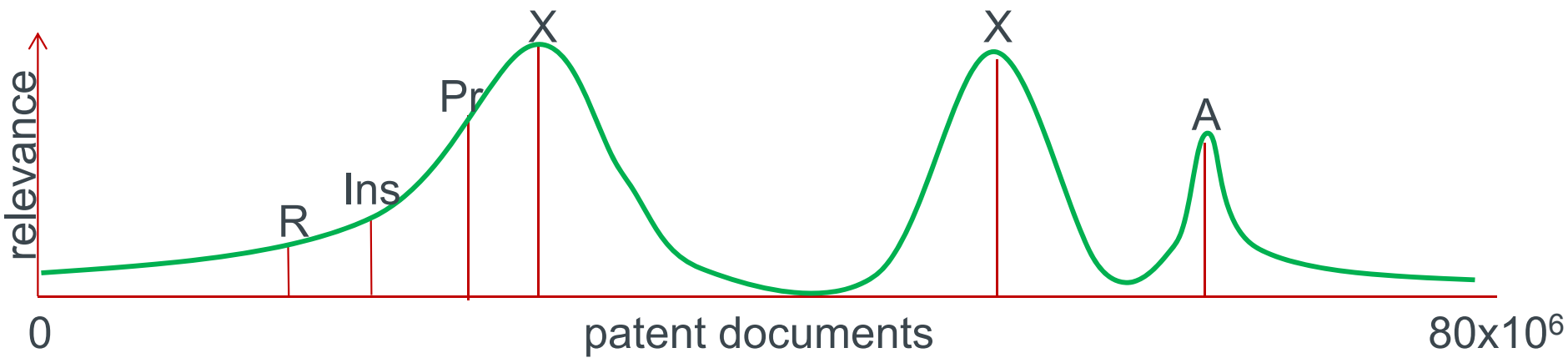
Benchmarking Environment

- Search reports for about 40 million simple patent families
- The relevant documents are mentioned in the search report as either
  - X or A citations
- Only few citations per document (2 X-category, 3 A-category)



Benchmarking Environment

- Search reports for about 40 million simple patent families
- The relevant documents are mentioned in the search report as either
  - X or A citations
- Only few citations per document (2 X-category, 3 A-category)
- But more "soft" information available: Returned during search (R), put aside for detailed inspection (Ins) , printed (Pr), stored for citation (Cs)



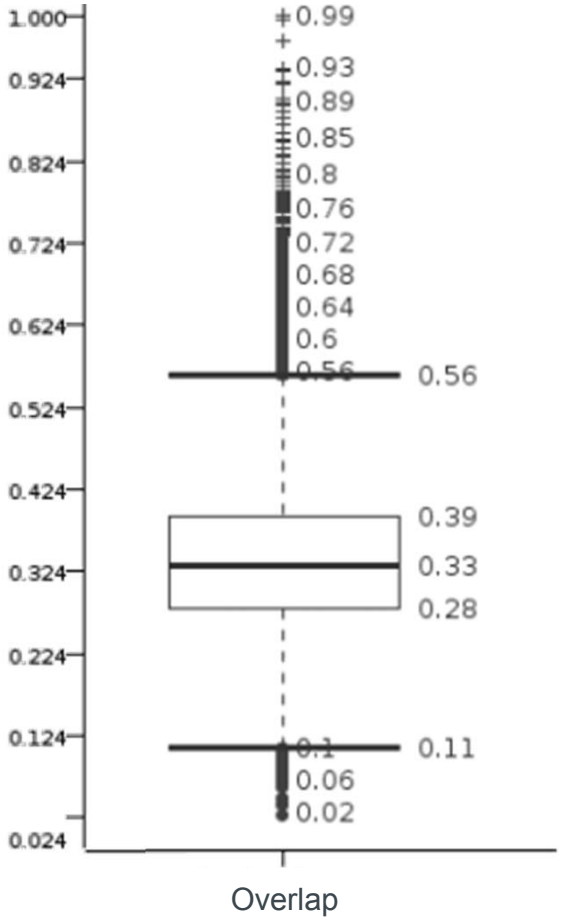
- Anecdotal Evidence and the need for a Benchmarking Environment
- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence

Term-Based Search

- Can we hope to find all citations with keywords?

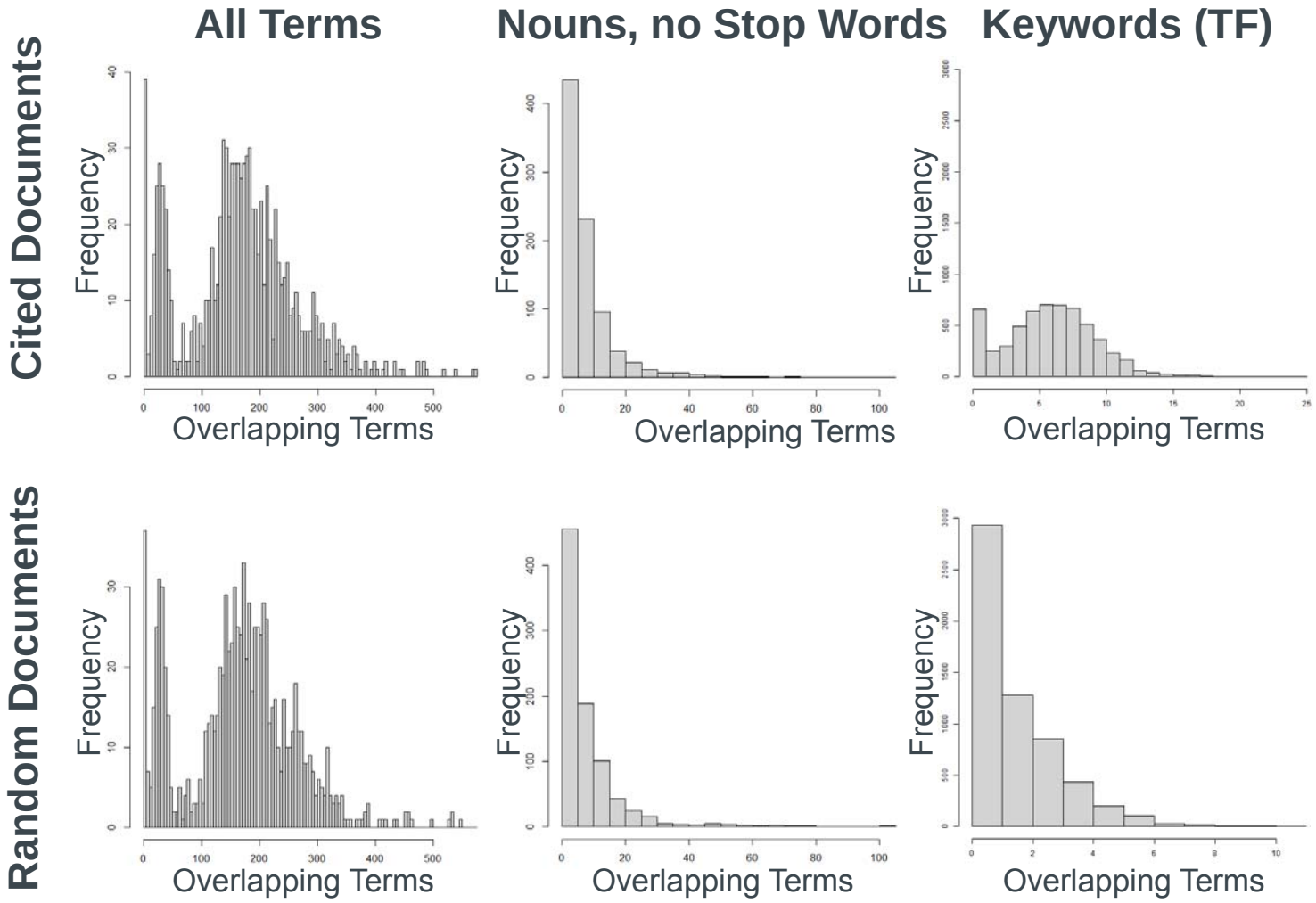
➔ overlap in vocabulary (non-stopwords)

Application **Citation**



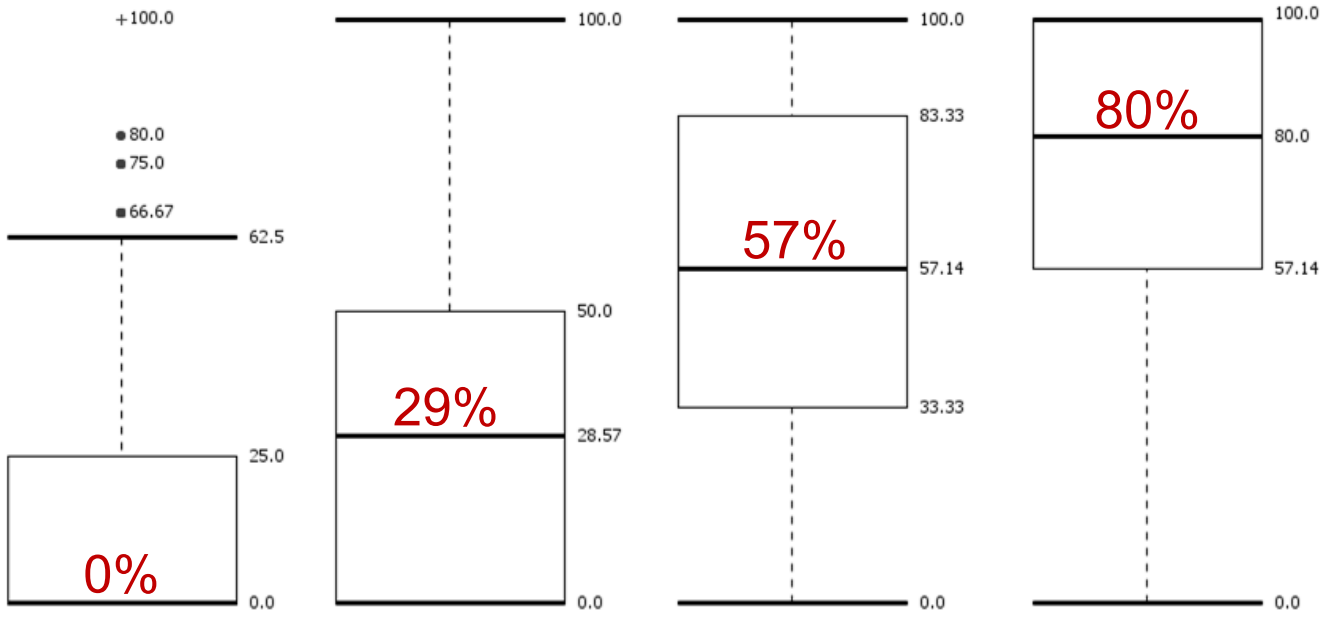
Term-Based Search

- However: also overlap with random documents
- Random: Y-Scrambling



Term-Based Search fully automated

- How good can we get?
- More Like This on Title / Abstract / Description / Claims
- Percentage of citations found in the top k

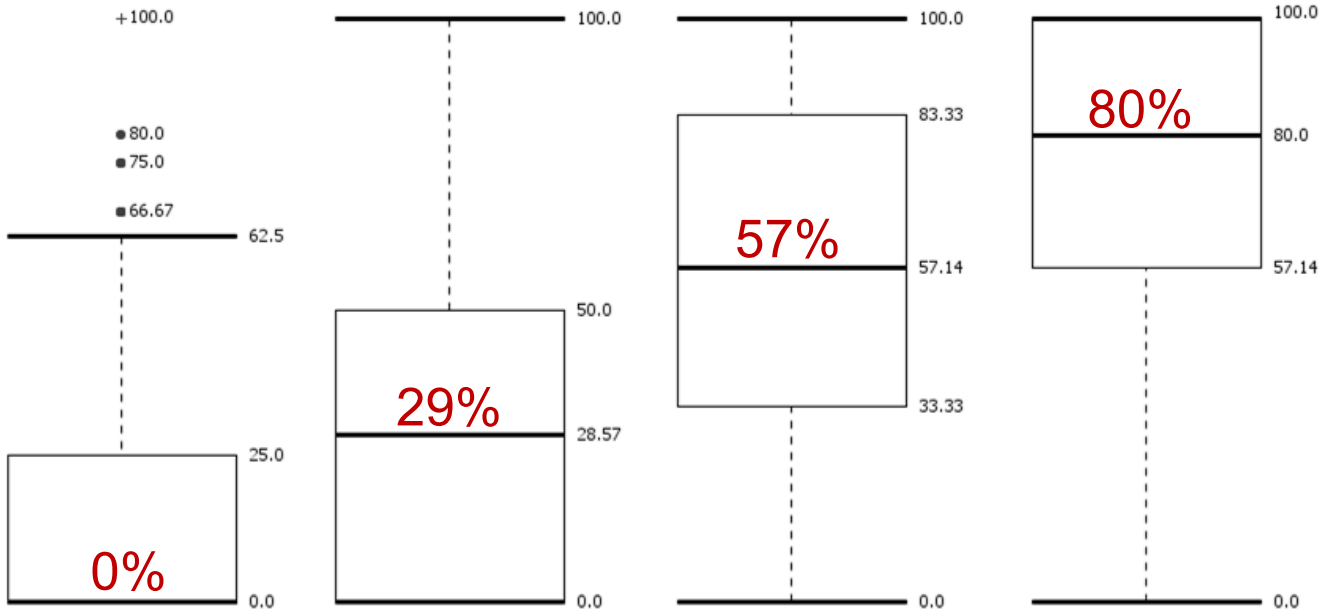


Top 100 Top 1 000 Top 10 000 Top 100 000

Term-Based Search fully automated

- **Can we close the gap?**

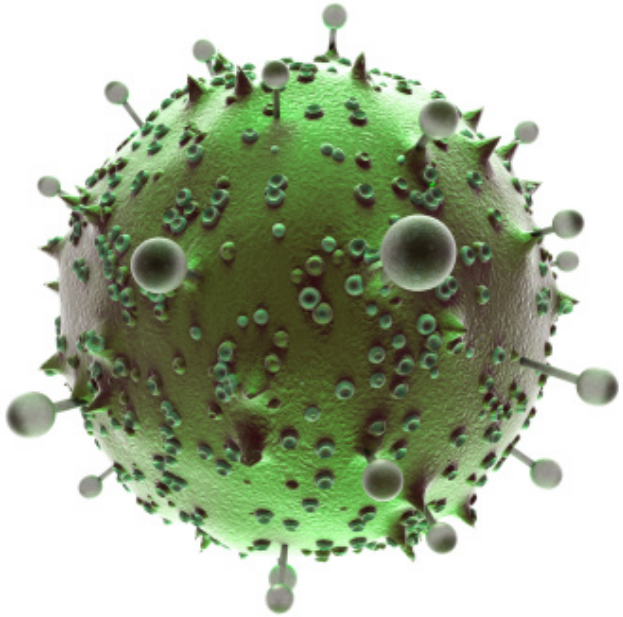
... using semantic search?



Top 100 Top 1 000 Top 10 000 Top 100 000

- Anecdotal Evidence and the need for a Benchmarking Environment
- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence

Introducing "Semantics"



Virus

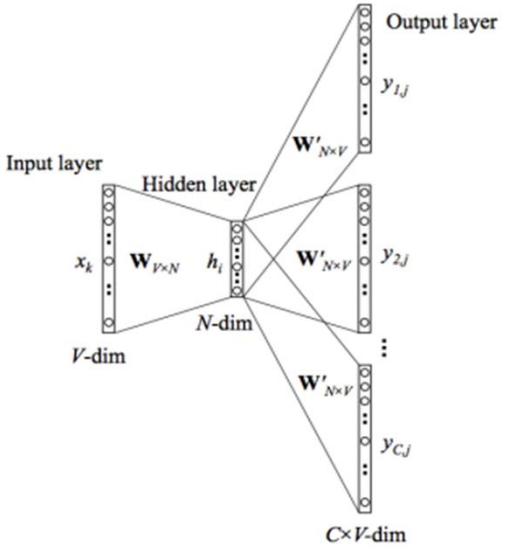
Automated Query Expansion

- Analyze word embeddings

Word

Context — eating **apples** is healthy

eating **fruits** is healthy

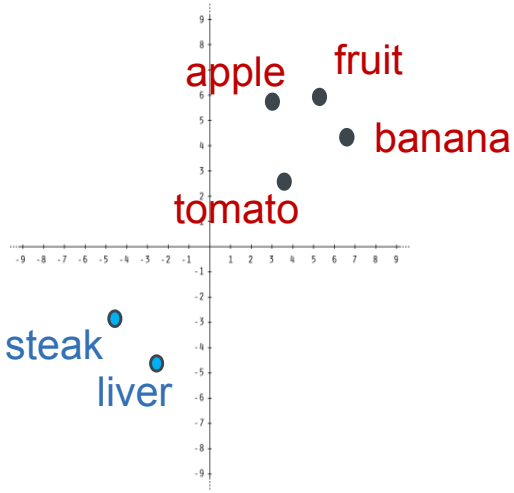


- Words with similar meaning have similar context

- Train neural network ➔ words represented by vectors

apple = [0.253, 1.793, ..., 5.555, 3,142]

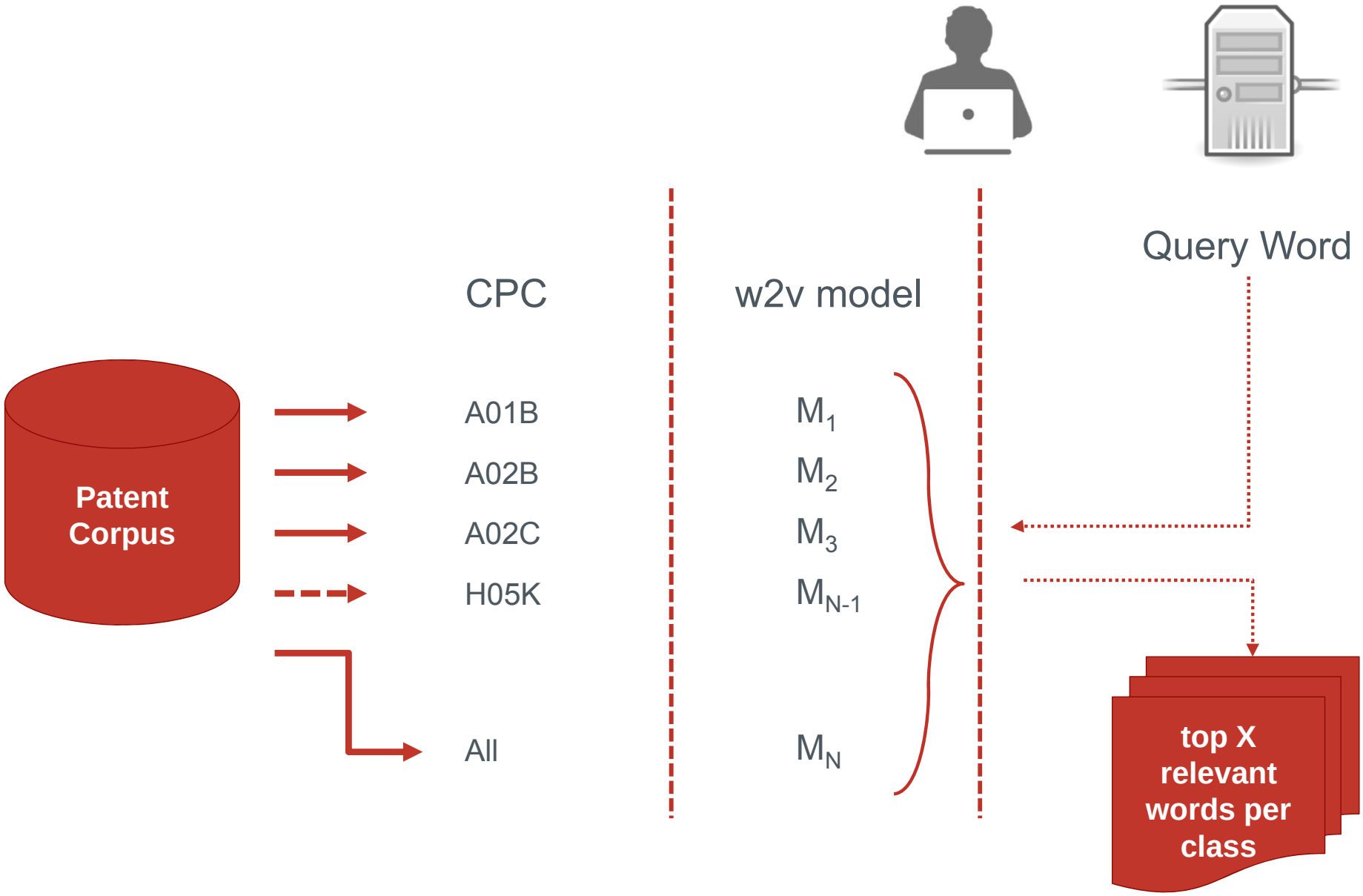
- Words with similar meaning have similar vectors



→ allows distance calculations between words

→ semantic similarity

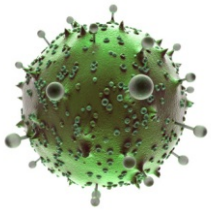
Automated Query Expansion



Automated Query Expansion

A61K

A virus is a small infectious agent that replicates only inside the living cells of other organisms.



hiv, viral, human immunodeficiency, viral genome, genome, replication, pathogen, adenovirus, replicating

G06F

A ~~computer~~ virus is a type of malicious software program ("malware") that, when executed, replicates by reproducing itself (copying its own source code) or infecting other computer programs by modifying them.

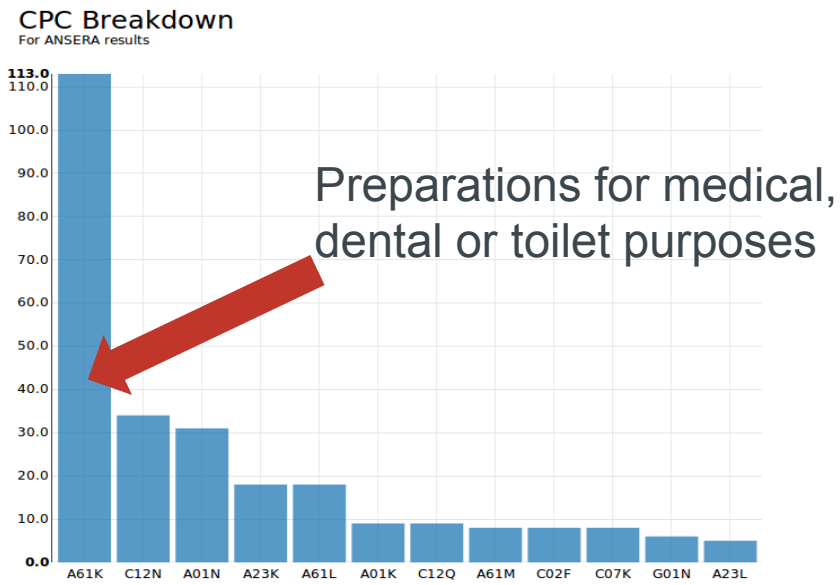
infected, virus infection, malicious, anti-virus, malware, virus-scanning, virus worm, macro virus, virus scanner

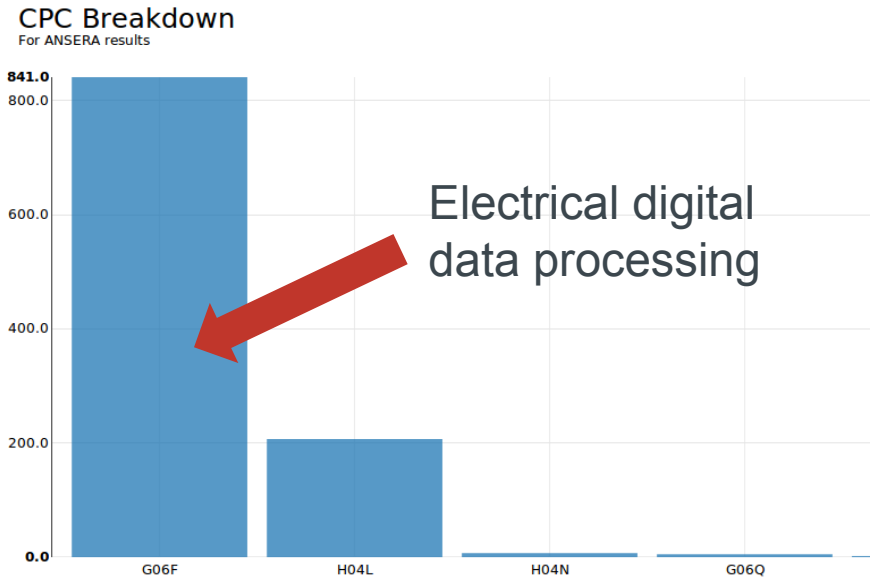
Semantics through Machine Learning

Example Queries: descriptions from Wikipedia

A **virus** is a small infectious agent that replicates only inside the living cells of other organisms.

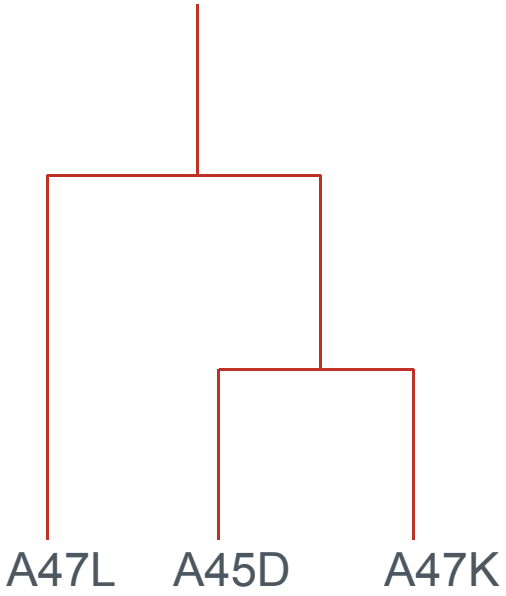
A ~~**computer**~~ **virus** is a type of malicious software program ("malware") that, when executed, replicates by reproducing itself (copying its own source code) or infecting other computer programs by modifying them.

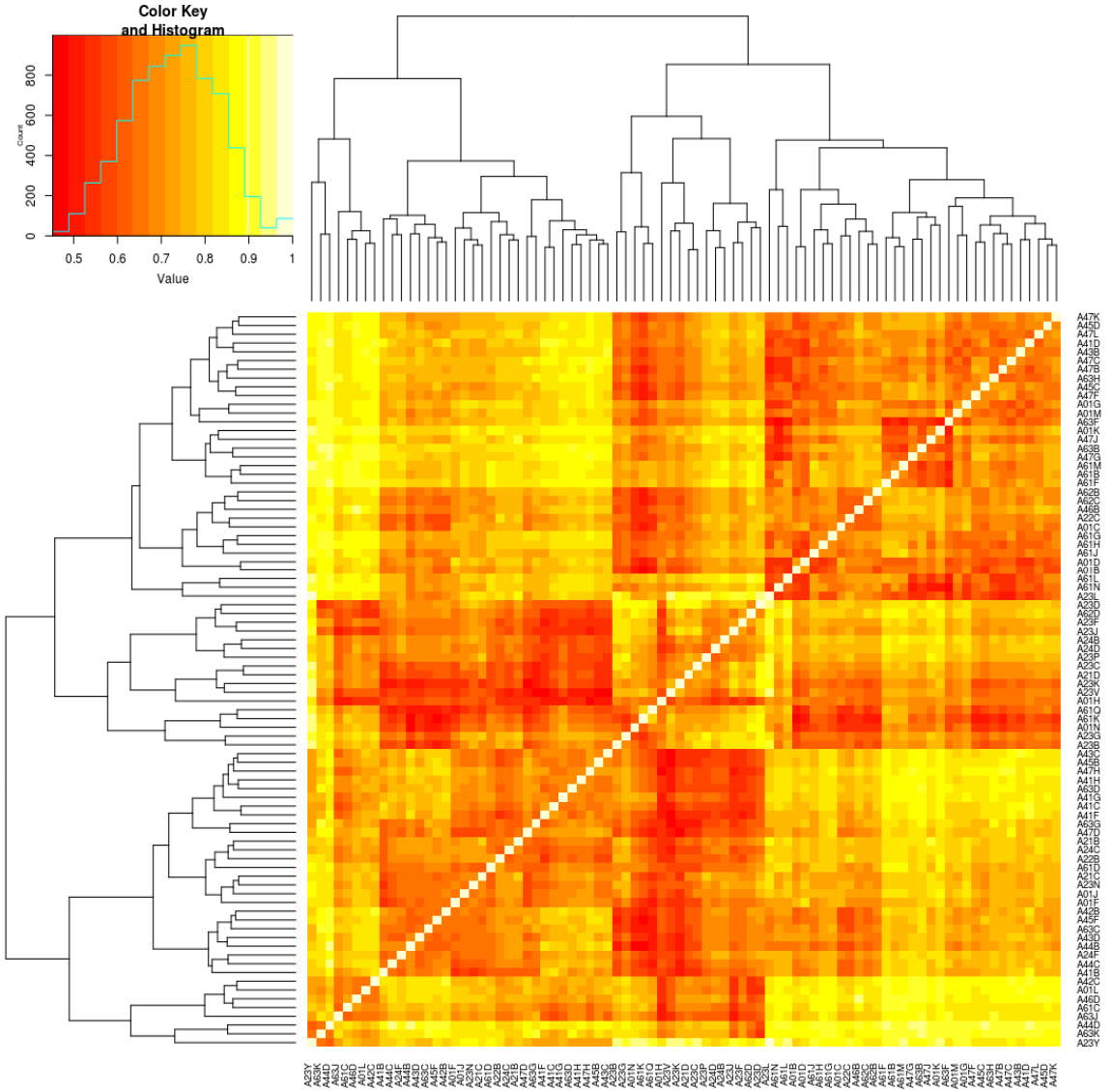




Automated Query Expansion

Inter model similarity in CPC A Human Necessities based on vocabularies





Automated Query Expansion

- Word2Vec Models applied to (100) extracted TF-IDF Keywords

"Touch sensing system and display apparatus"

EP2869168A120150506

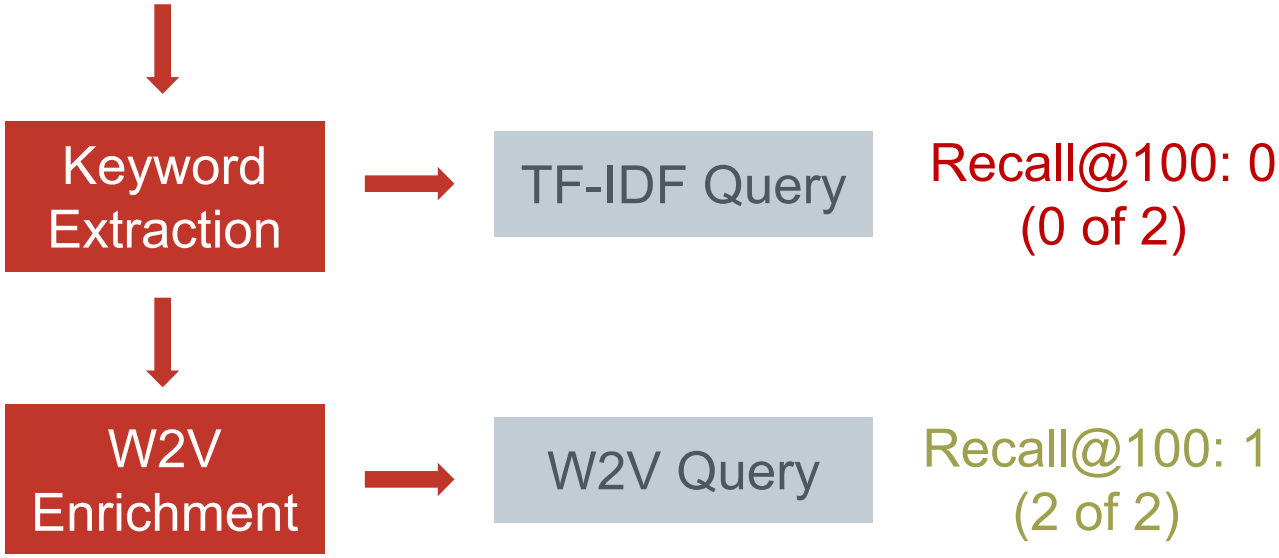
TF-IDF Keywords (10 of 100)	TF-IDF Scores	Word2Vec Enrichments (G06F) max 5 terms, at least Cosine 0.6
possessory	49.01	possessory
sensing	33.04	sensing OR sensed OR sensor OR touch
touch	31.36	touch OR touched OR touch screen OR touches OR touch panel OR touching
bridge	27.88	bridge OR bridges OR pci bus OR bus OR interfaces 22a-22b OR pci-pci bridge
nodes	27.59	nodes OR node OR nodes n2
shared	26.86	shared OR share OR sharing OR non-shared
electrodes	25.52	electrodes OR electrode OR electrodes y2
controller	23.18	controller OR control OR controllers OR bus OR unit OR processor
instructing	22.11	instructing OR instruct OR instructs
masters	21.23	masters
memory	20.39	memory OR ram OR non-volatile OR memories OR processor OR flash
sensor	20.38	sensor OR sensors OR sensing OR humidity sensor OR sensed OR sensor senses
algorithm	20.02	algorithm OR algorithms
senses	20.02	senses OR sensed
data	19.91	data OR stored OR stores OR store OR storing
...	...	...

Automated Query Expansion

- Query Enrichment brings increased performance

"Touch sensing system and display apparatus"

EP2869168A120150506



Automated Query Expansion

- Documents found with increased Term Overlap with Query

	TF-IDF Query	W2V Query
	Recall@100: 0 (0 of 2)	Recall@100: 1 (2 of 2)
39100958	39	80
48143418	43	80

Automated Query Expansion

- Word2Vec Expansion Terms in Overlap

New Overlap	Original Keyword	Word2Vec Enrichment
outputs	signal	OR signals OR circuit OR output OR clock OR outputs
apparatus	controlling	OR control OR controlled OR apparatus
screen	display	OR screen OR displays OR displaying OR displayed OR lcd
increasing	increases	OR increase OR decreases OR increased OR increasing OR decrease
	decreasing	OR increasing OR decrease OR increase OR decreases OR decreased
serially	parallel	OR serially

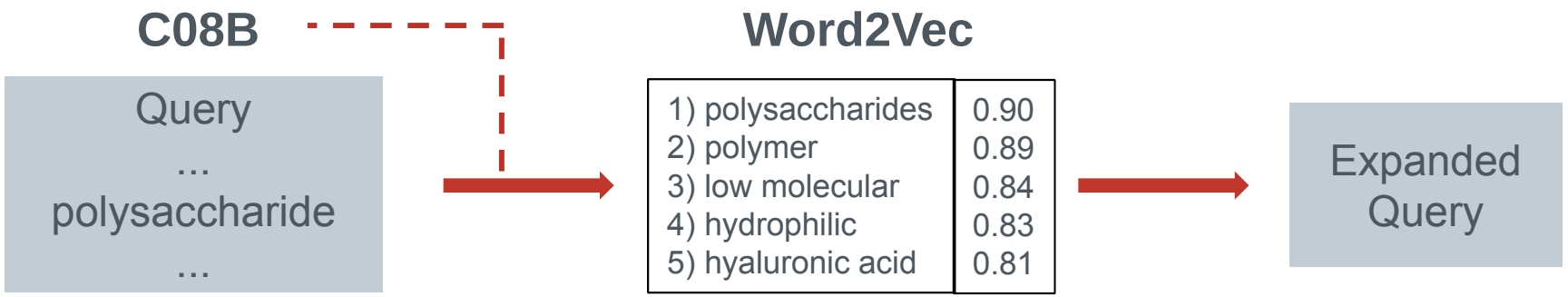
- Re-Suggestion of known Query Terms

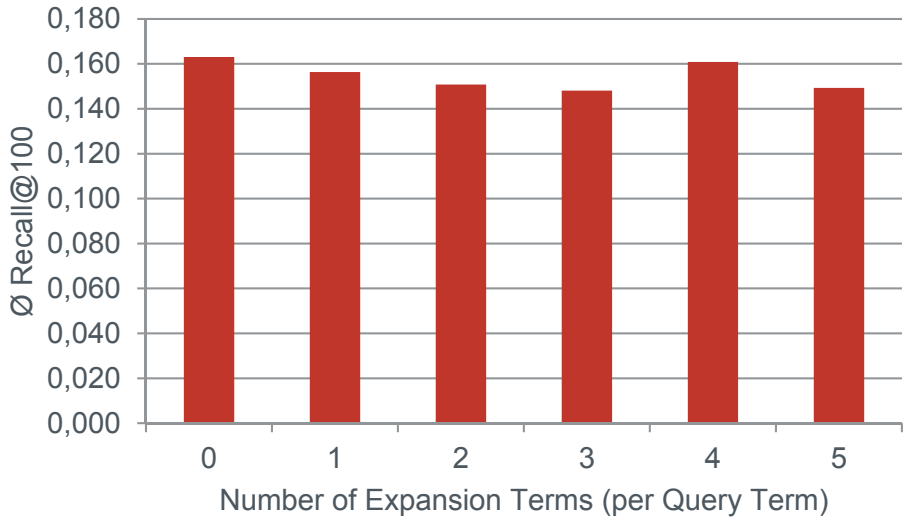
Term	Original Keyword	Word2Vec Enrichment
transmission	transmitted	OR transmits OR received OR transmit OR transmission OR transmitting
	reception	OR transmission OR transmitted OR transmits
	transmitting	OR receiving OR transmitted OR transmission OR sending OR transmit

- Should this result in higher weight?
- Models heavily suggest inflections

Automated Query Expansion

- Query Expansion (Word2Vec, 647 models at CPC subclass level)
  - Top 70 TFIDF keywords, CLEF-IP





Top 30 Documents

	Overlap	Improvement
1 Term	84 %	5 %
2 Terms	69 %	11 %
3 Terms	66 %	12 %
4 Terms	96 %	2%
5 Terms	78 %	6%

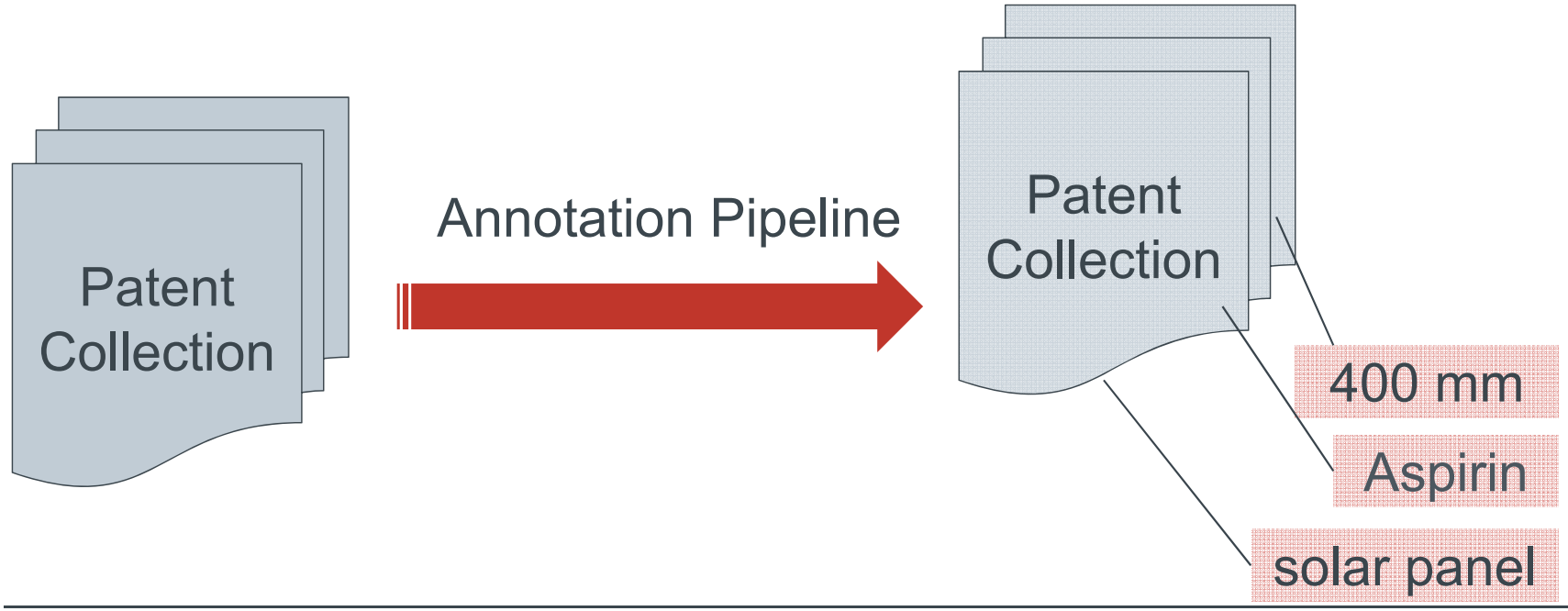
- Anecdotal Evidence and the need for a Benchmarking Environment
- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence

Introducing semantics through annotations

We are in the process of annotating the full prior art collection with normalised

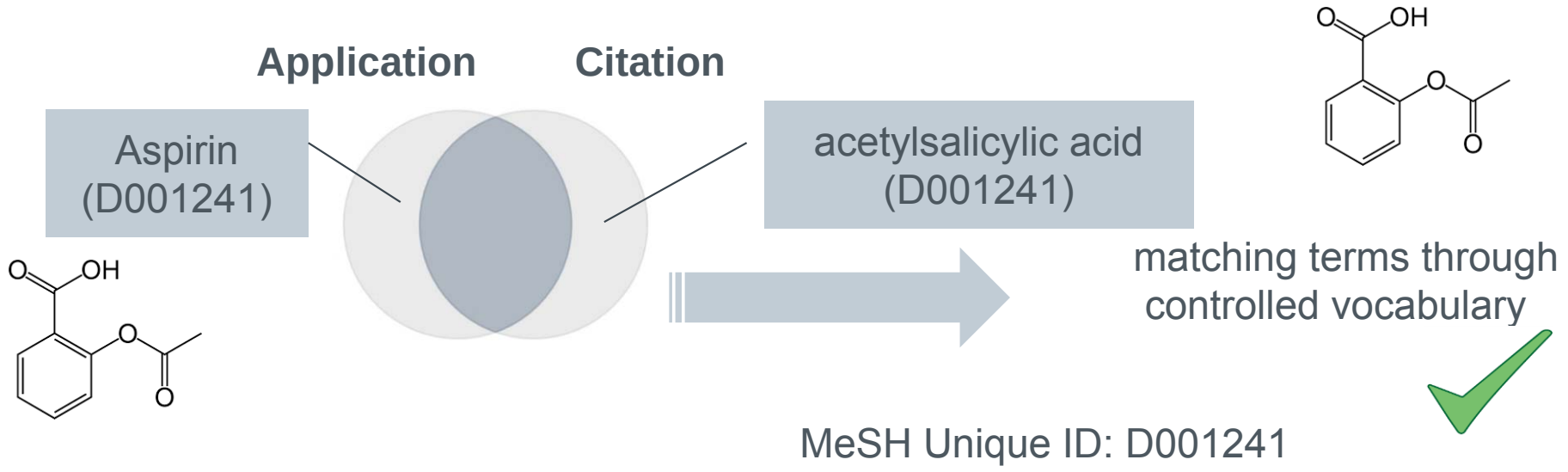
- Chemical Entities
- Physical Units
- Citations
- Controlled Terminologies (e.g. MeSH)

to enable a semantic search of those entities.



Introducing annotations and semantics

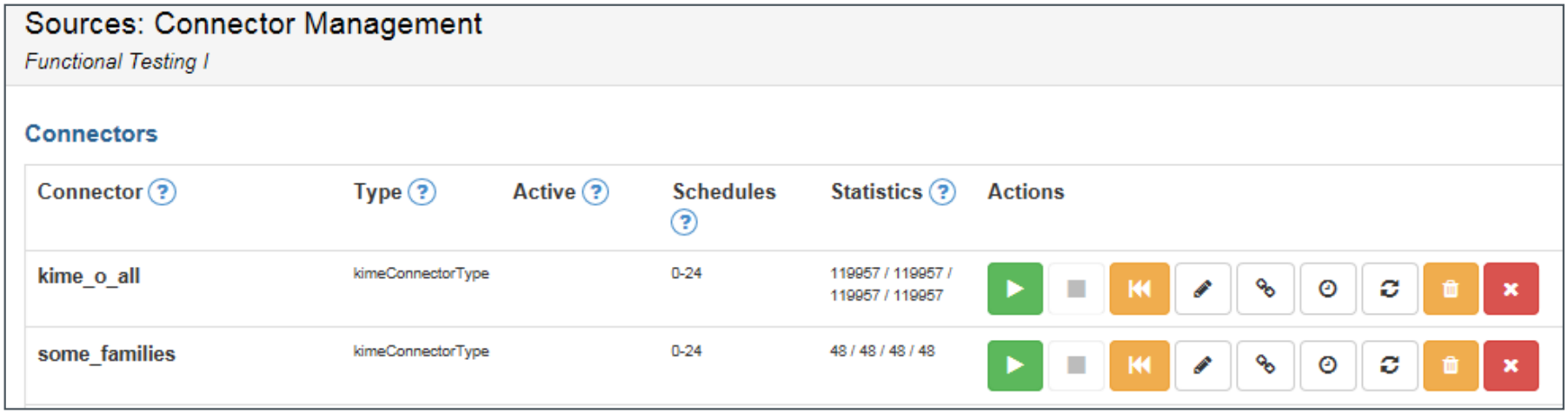
- Our annotation platform's key elements
  - Based on the open source framework UIMA
  - Plug and play of new components (analysis engines)
  - Quality Evaluation Workbench
  - Knowledge Base for controlled vocabulary
  - Scaling architecture
  - Data and annotations stored in noSQL databases



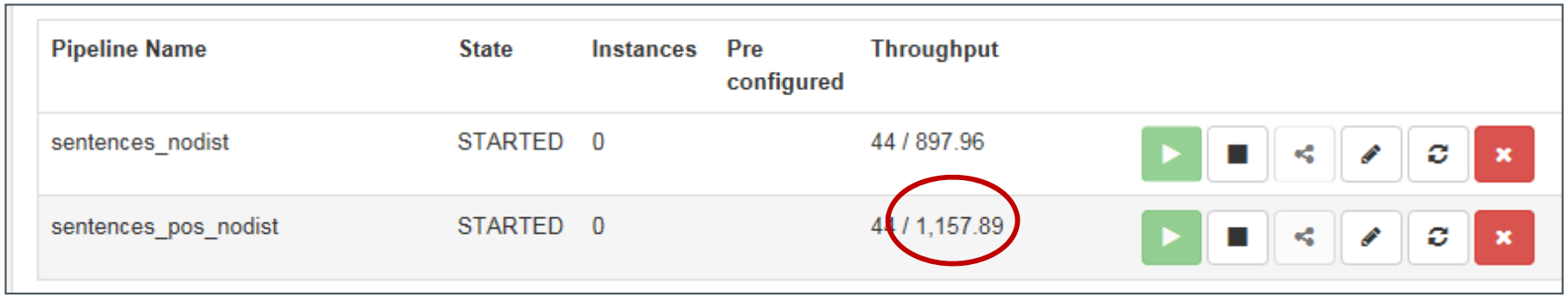
Introducing annotations and semantics

Project in Execution - Platform installed and tested at this moment

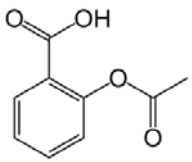
1. Define (sub)set of documents to work on



2. Define an annotation Pipeline

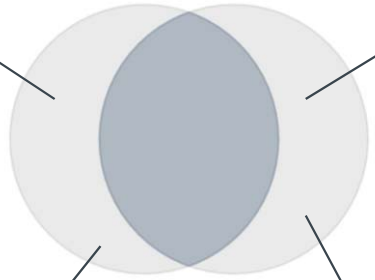


Word2Vec & APL



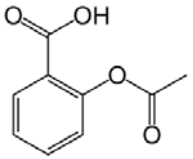
Application

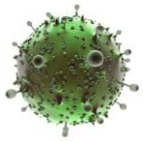
Aspirin



Citation

acetylsalicylic acid





Virus

Virus

matching terms through controlled vocabulary

Aspirin = acetylsalicylic acid (would have been False Negative)

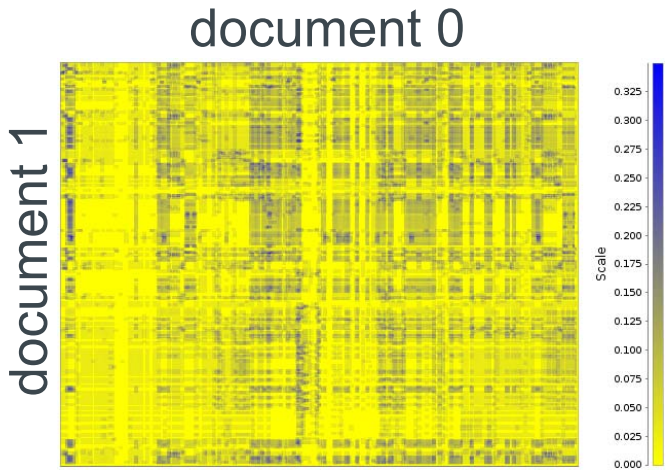
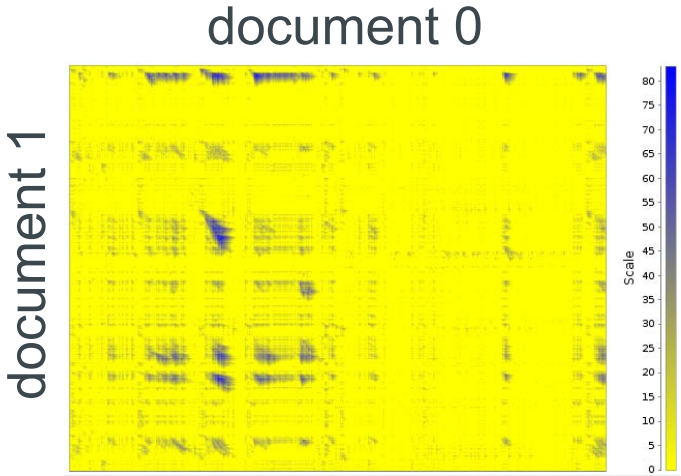
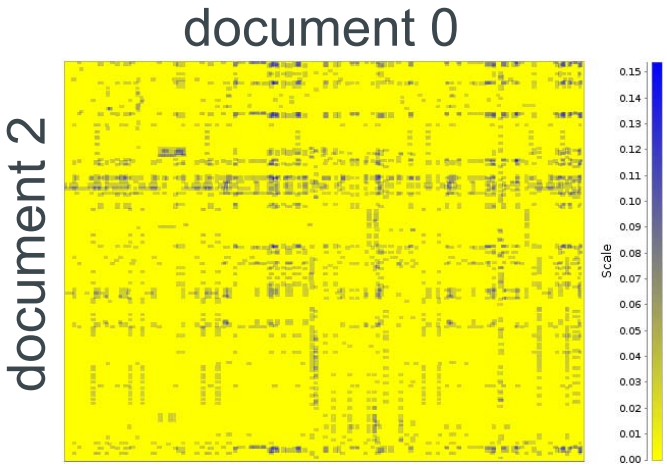
separating identical terms with different meaning

Virus = Virus (would have been False Positive)

- Anecdotal Evidence and the need for a Benchmarking Environment
- Keywords and query generation in (automatic) search scenarios (Project ERa)
- Introducing machine learned semantic search technologies and (automatic) query expansion
- Introducing terminology based semantics within the APL project
- (Automated) Search result confidence

Identify Successful (Automated) Searches

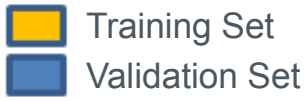
Machine-Learning: look at the documents

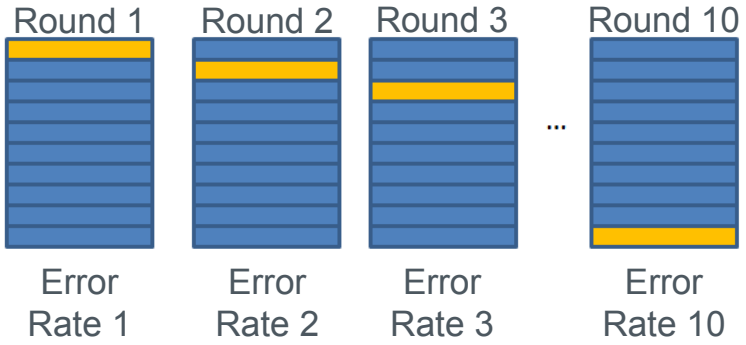
	Word2Vec Heatmap	**Sequence Alignment (Bioinformatics)**
Citation		
Random		

Can we identify successful searches? (Machine learned)

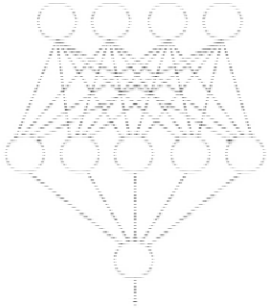
- Prediction Model for Citability based on Distribution Statistics of heat maps

  [mean, variance, skewness, kurtosis, min, q1, q2, q3, max]
- 10-fold Cross Validation (Y-Scrambling)





Multi-Layer Perceptron



~8%

Predicted Relevance

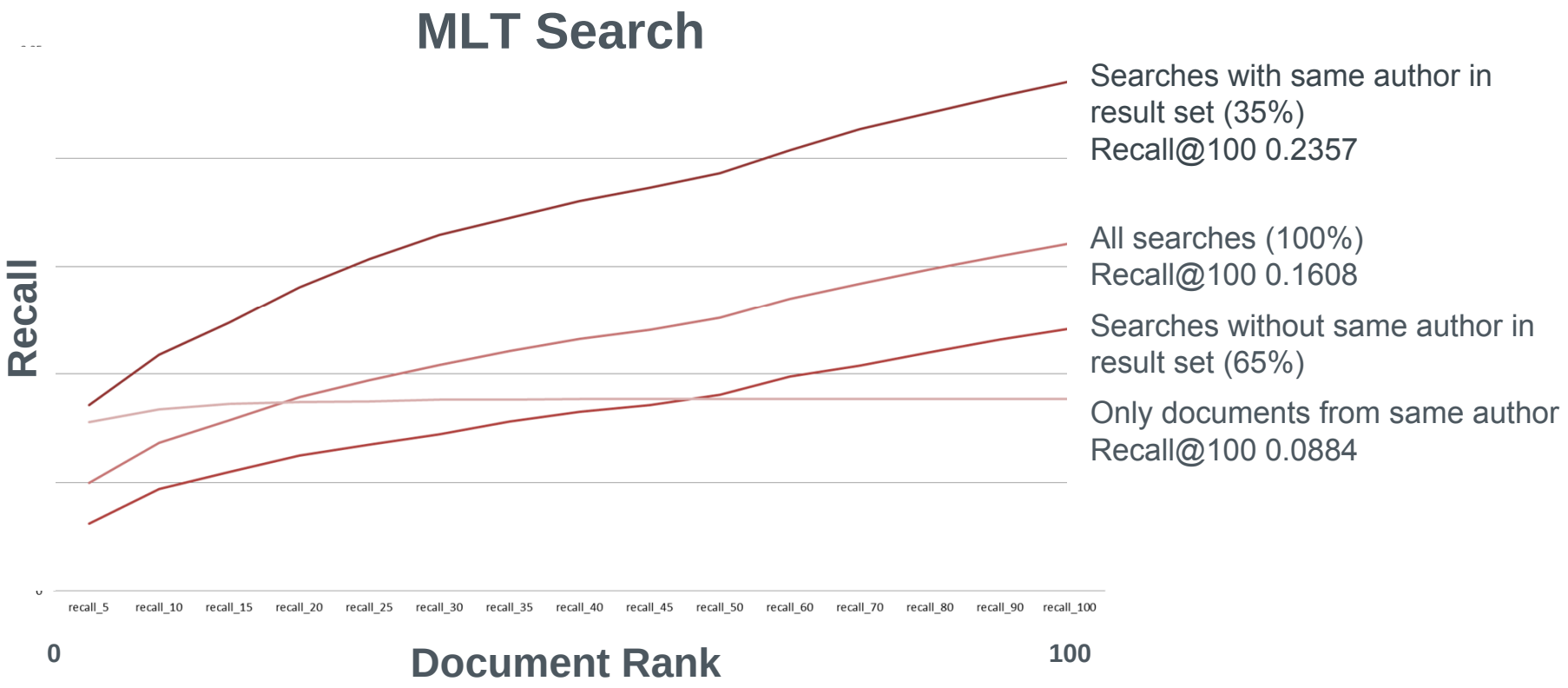
Random Forest

~5%

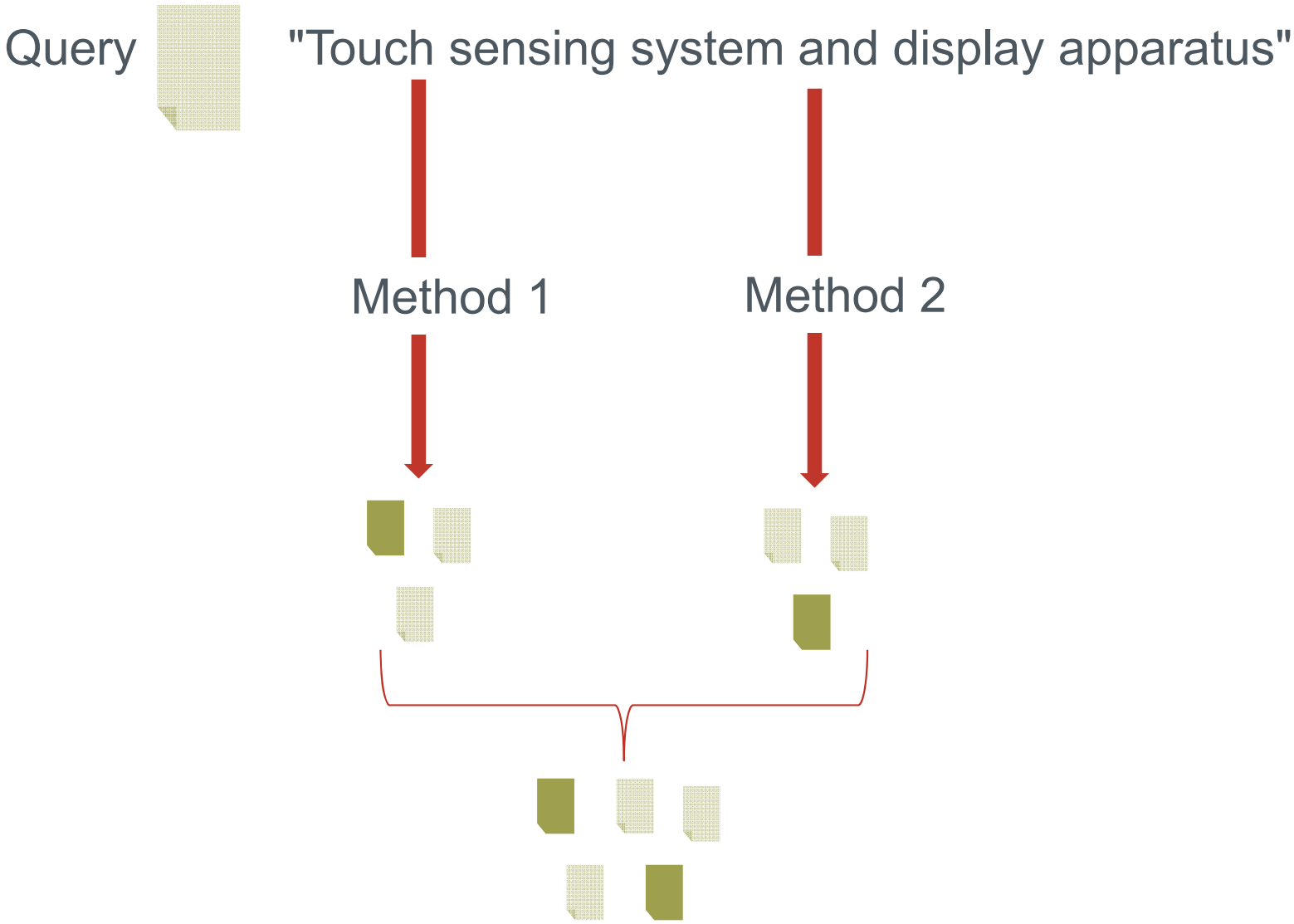
- ensemble of decision trees
- bootstrap aggregating
- random selection of features

Identify Successful (Automated) Searches

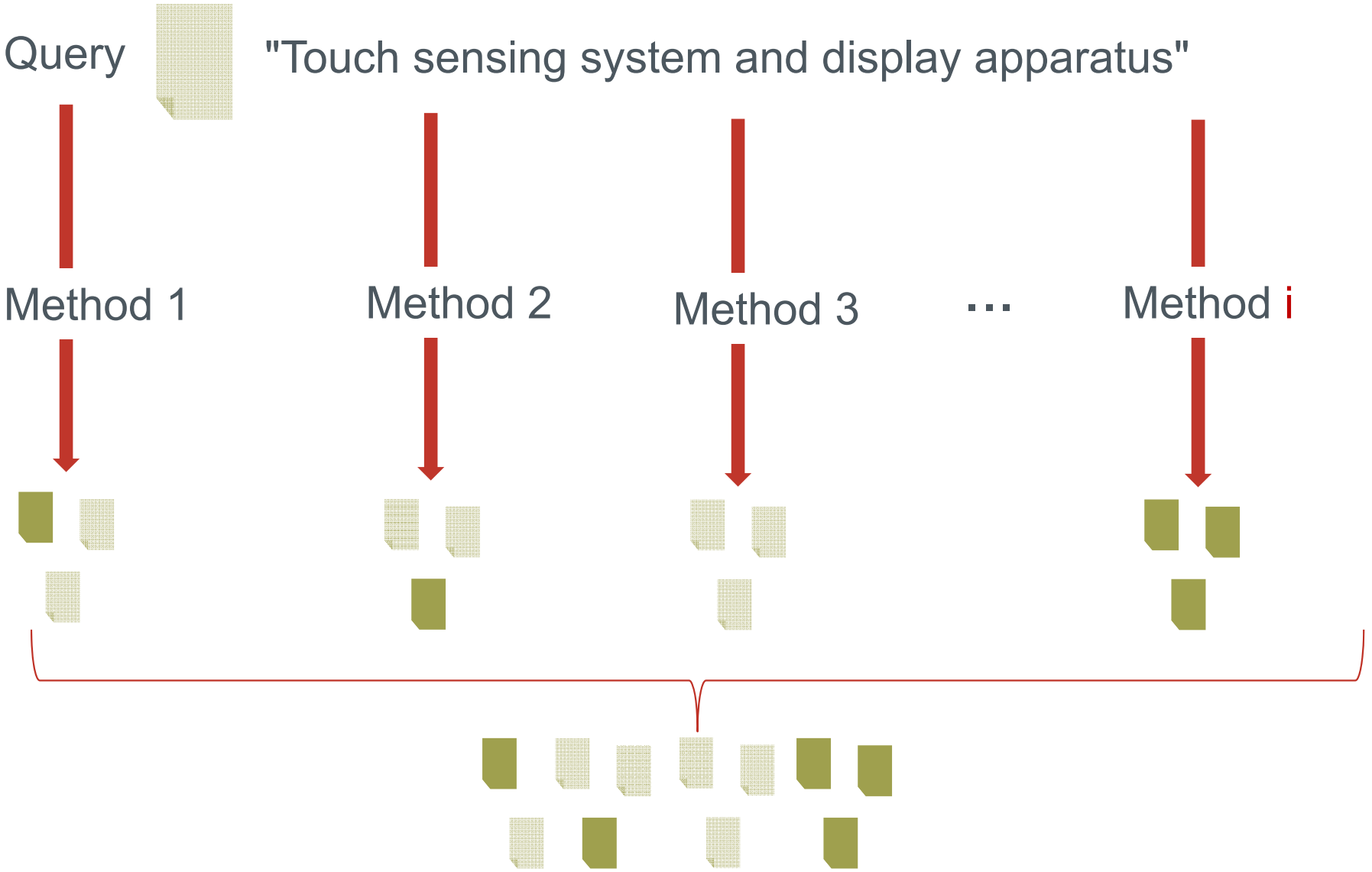
- Meta Data: same author as in application occurs in result set
- Interesting Item Set Mining & Subgroup Discovery



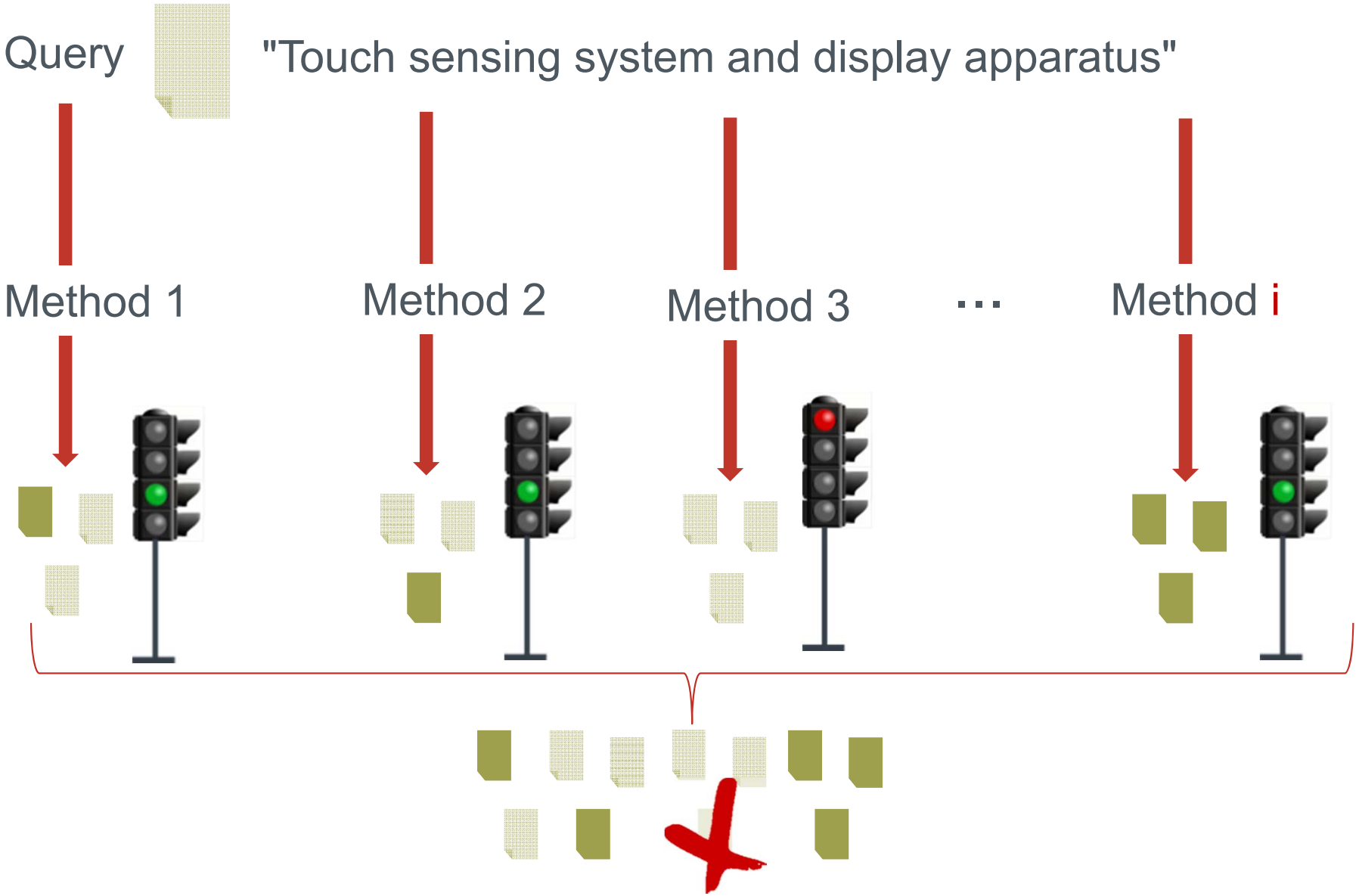
Complementariness of Results



Complementariness of Results



Complementariness of Results



Conclusion and outline

- We are exploiting state-of-the-art semantic- and other search technologies
- We have created a prototyping and benchmarking environment to evaluate the performance and quality of a new search algorithm learning from prior searches
- Bringing all these new technologies together in a productive search environment is the biggest challenge ahead
- Semantic (text) search is one small mosaic piece in the complex search environment at the EPO

Acknowledgments

- The Search & Knowledge Directorate
  - Domenico Golzio — Director
  - Stefan Klocke — Head of Department
  - Volker Hähnke — Ranking and Confidence
  - Matthias Wirth — Neo4J, CPC classification
  - Pavel Goncharik — Ansera, ranking, lucene
- Examiners
  - Robert Herman — ERa
  - Ype Kingma — ERa
  - Markus Arndt — ERa
  - Christos Stergio — APL