

Text Mining for Creative Cross-Domain Knowledge Discovery

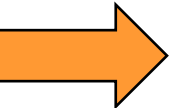
Nada Lavrač

Jožef Stefan Institute, Ljubljana, Slovenia

with contributors

Bojan Cestnik, Matjaž Juršič, Tanja Urbančič, Borut Sluban
et al.

Talk outline

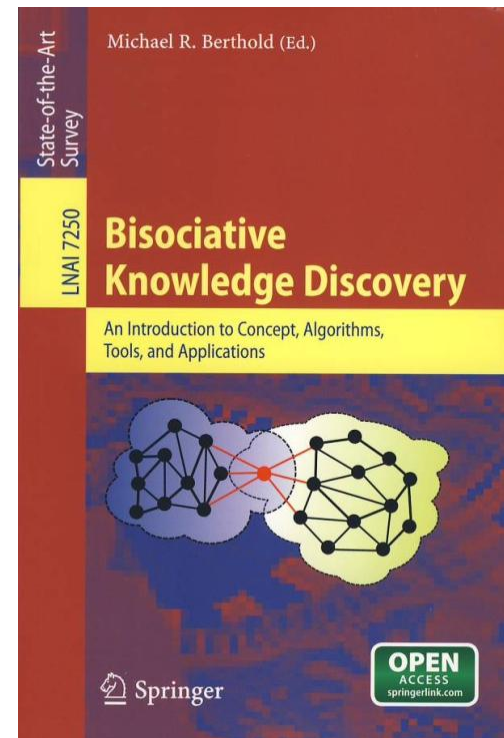
- 
- Background and motivation
 - Literature-based discovery
 - Cross-domain literature mining approaches
 - Outlier detection for cross-domain knowledge discovery
 - Cross-domain knowledge discovery with CrossBee
 - Summary and conclusions

The BISON project

- Explore the idea of bisociation (Arthur Koestler, The act of creation, 1964):
 - The mixture - in one human mind – of **different contexts** or **different categories of objects**, that are normally considered **separate categories** by the processes of the mind.
 - The **thinking process** that is the functional basis of **analogical or metaphoric thinking** as compared to logical or associative thinking.

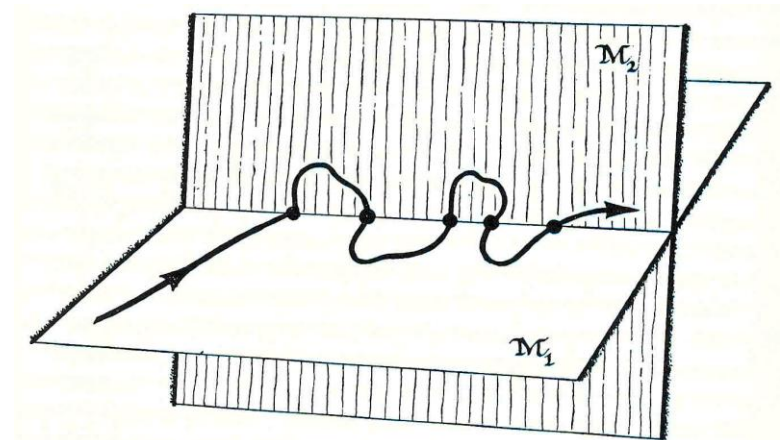
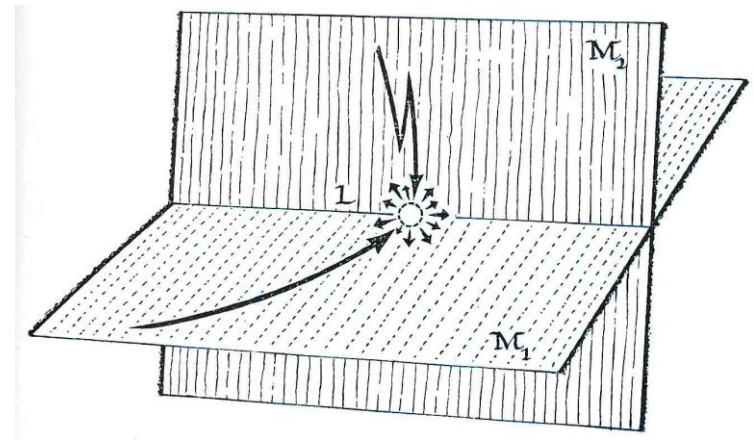
The BISON project

- BISON: Bisociation Networks for Creative Information Discovery, European 7FP project, www.bisonet.eu
- 12 partners (2008-2011)
- Open access book (Springer 2012):
Bisociative Knowledge Discovery
edited by M. Berthold



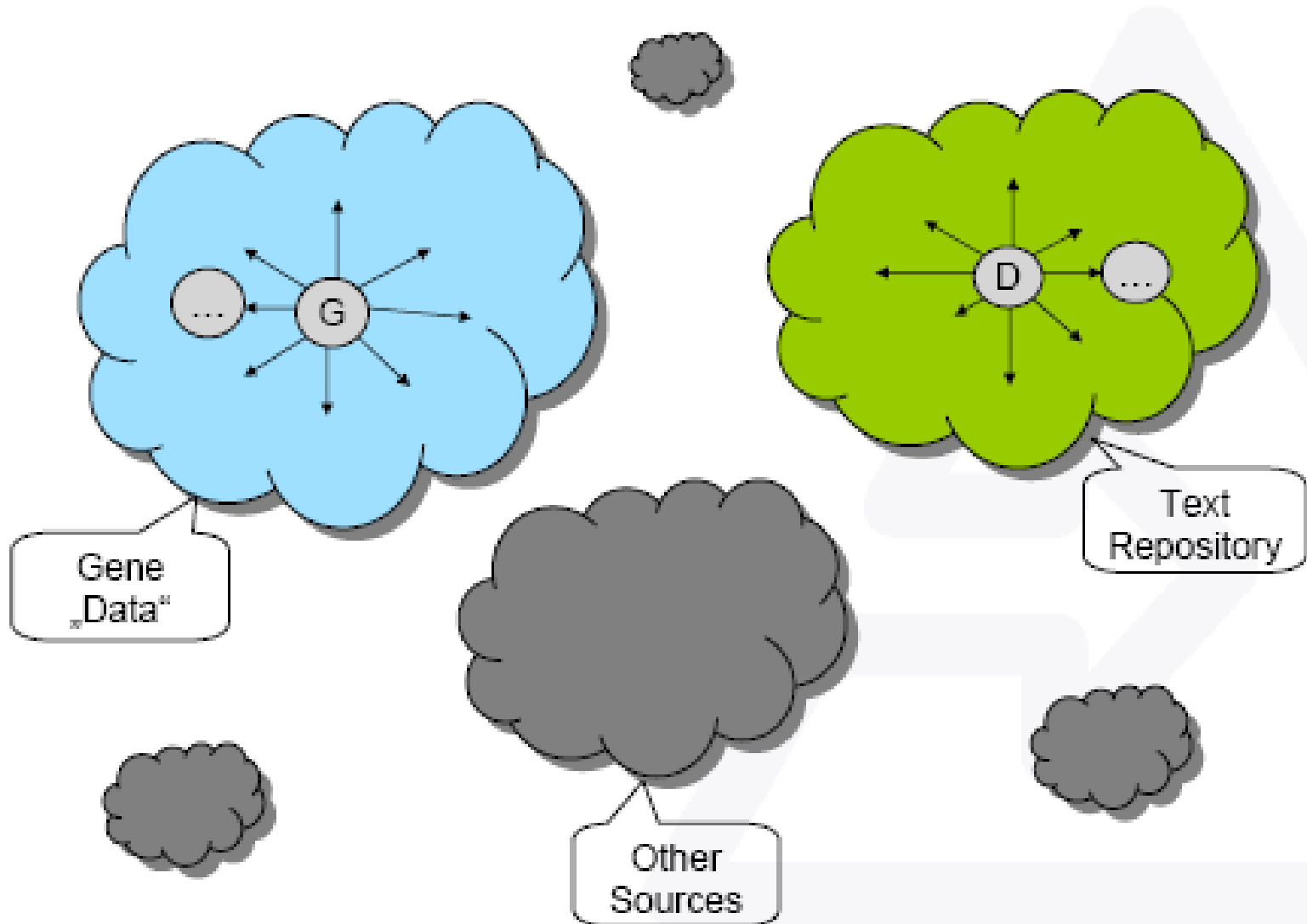
Bisociation discovery in BISON

- BISON challenge:
 - Find new insights: new **bisociations**, i.e., interesting new links **across domains**
- Two concepts are bisociated if and only if:
 - There is no direct, obvious evidence linking them
 - One has to cross contexts to find the link
 - This new link provides some novel insight



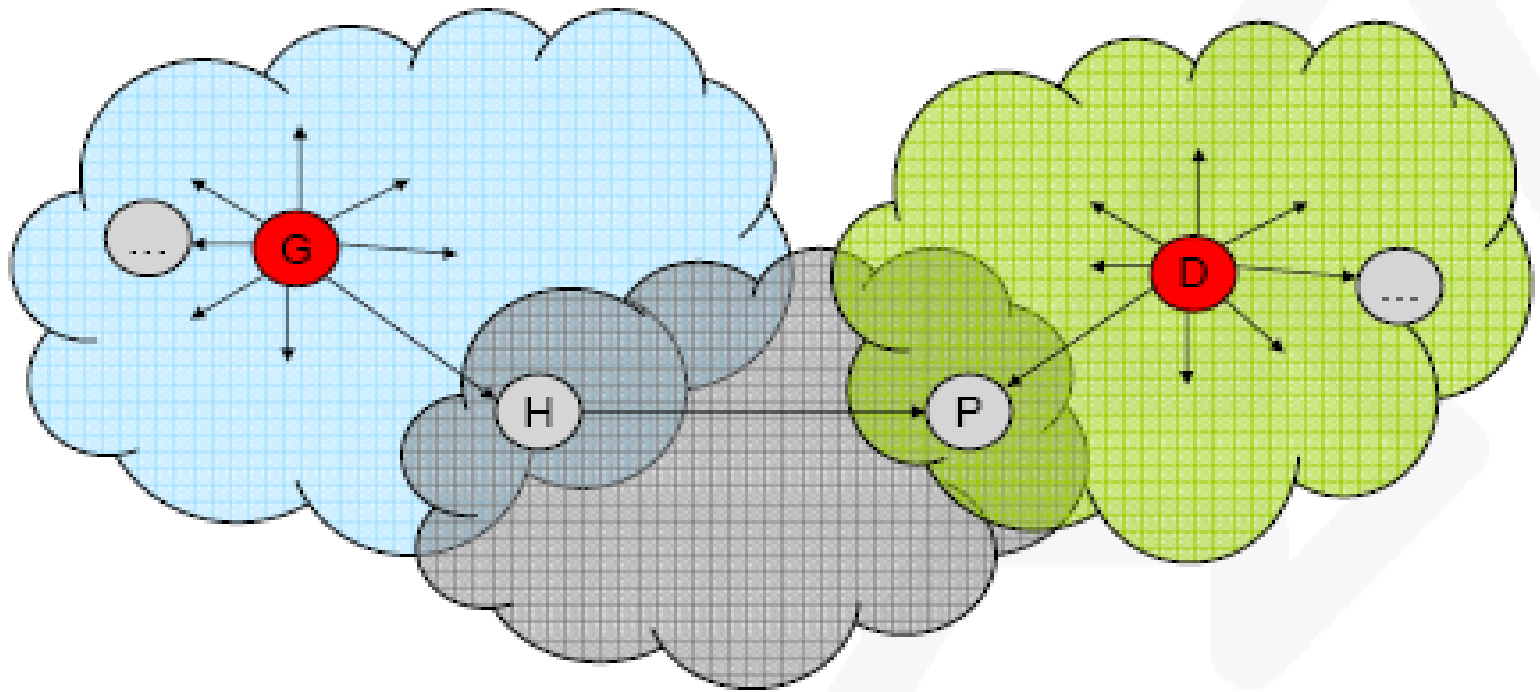
Heterogeneous data sources

(BISON, M. Berthold, 2008)

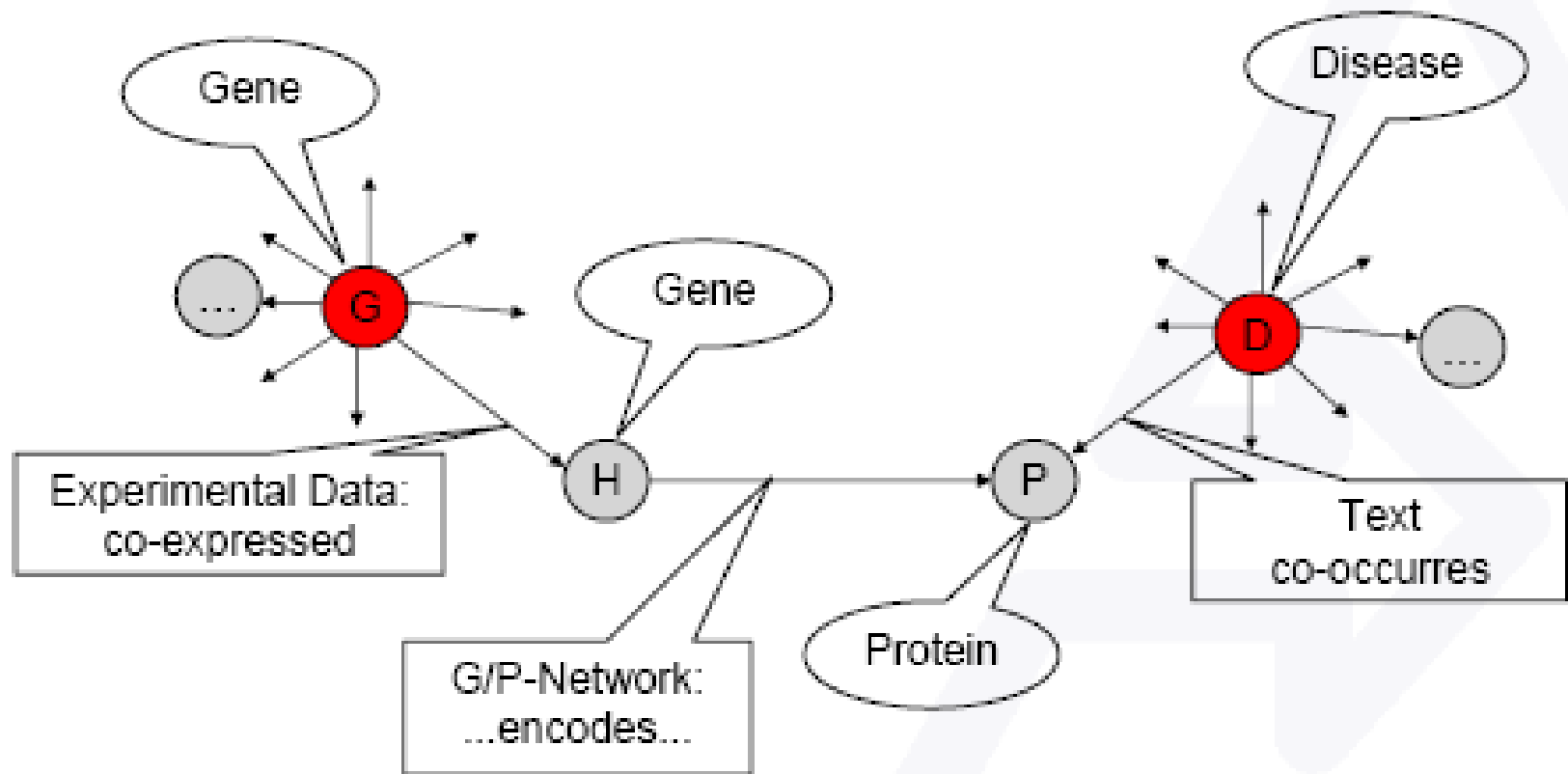


Bridging concepts

(BISON, M. Berthold, 2008)



Chains of associations across domains (BISON, M. Berthold, 2008)



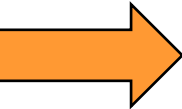
Main BISON approach

- Main approach: graph exploration
 - Find yet unknown links in a graph, crossing different contexts (domains)
- Open problems:
 - Crossing different contexts (domains): Finding unexpected, previously unknown links between BisoNet nodes belonging to different contexts
 - Crossing different types of data and knowledge sources: Fusion of heterogeneous data/knowledge sources into a joint representation format - a large information network named BisoNet (consisting of nodes and relationships between nodes)

Complementary BISON approach

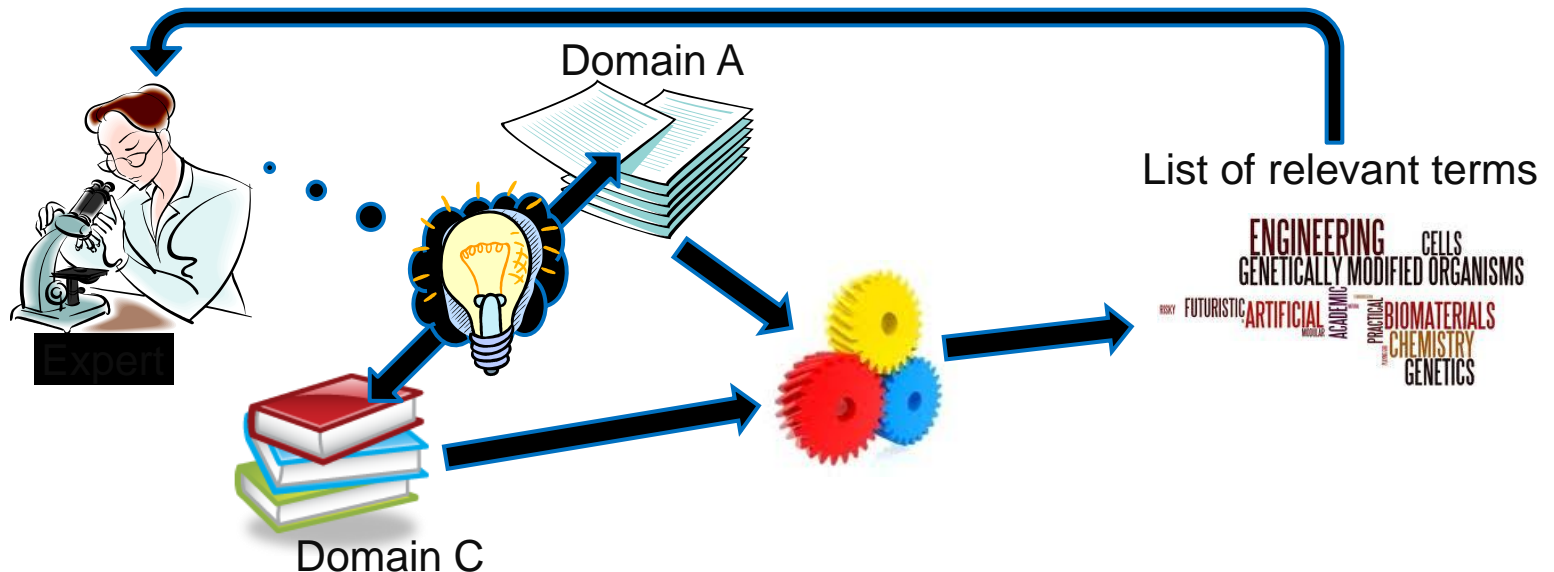
- Complementary approach: text mining
 - Find yet unknown terms in the intersection of documents, crossing different contexts (domains/literatures)
- Early related work: literature-based discovery (LBD)
 - Swanson (1988, 1990)
 - Smalheiser, Swanson (1998): ARROWSMITH
 - Weeber et al. (2001)
 - Hristovski et al. (2001): BITOLA
- Recent work: cross-domain literature mining
 - Petrič et al. (2007, 2009): RaJoLink
 - Juršič et al. (2012): CrossBee
 - ...

Talk outline

- Background and motivation
-  Literature-based discovery
- Cross-domain literature mining approaches
 - Outlier detection for cross-domain knowledge discovery
 - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions

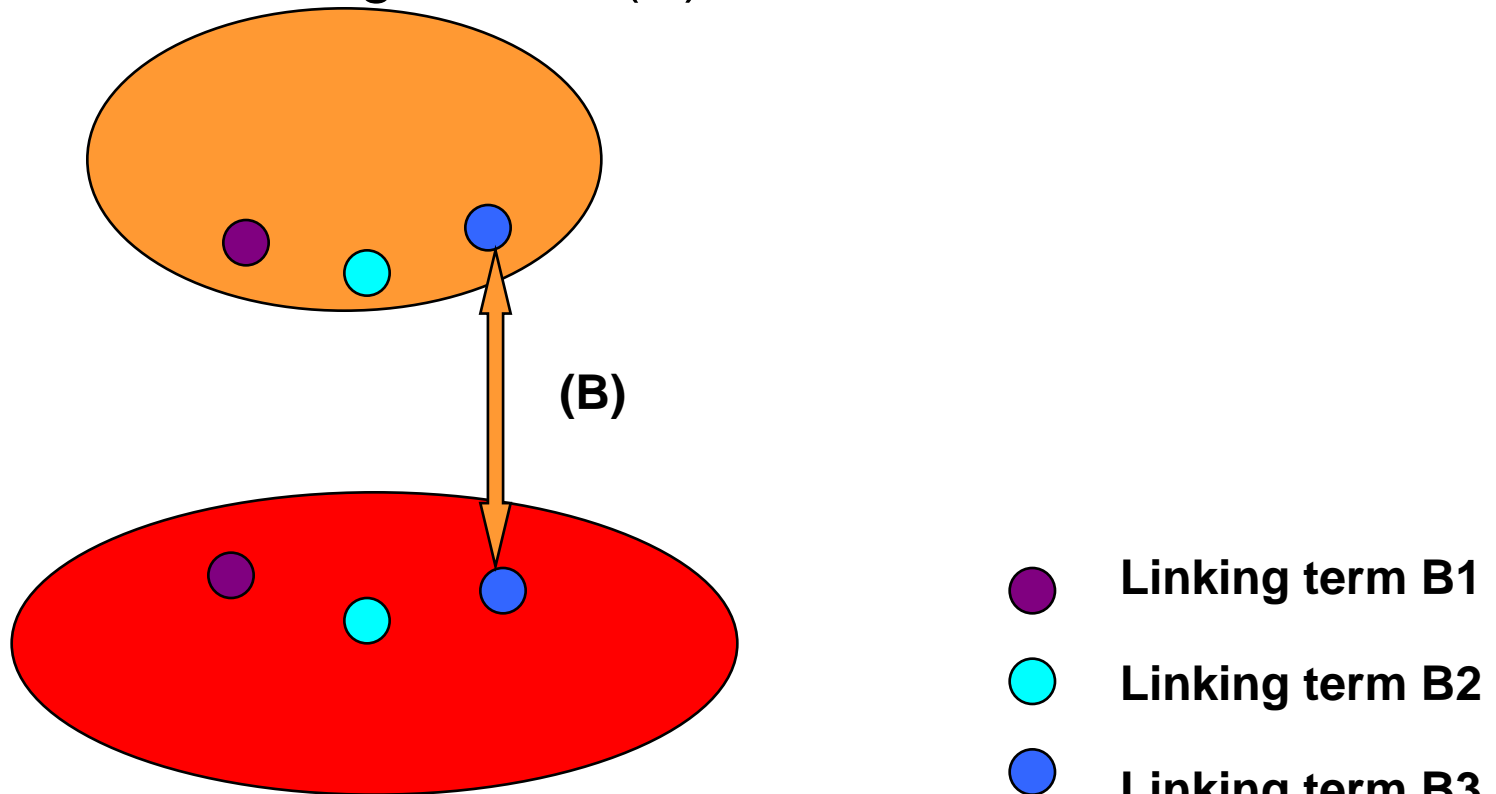
Literature-based discovery

- Help experts in cross-domain discovery for unknown facts/new findings
 - Closed discovery setting
 - Early work by Swanson: Medical literature as a potential source of new knowledge, 1990



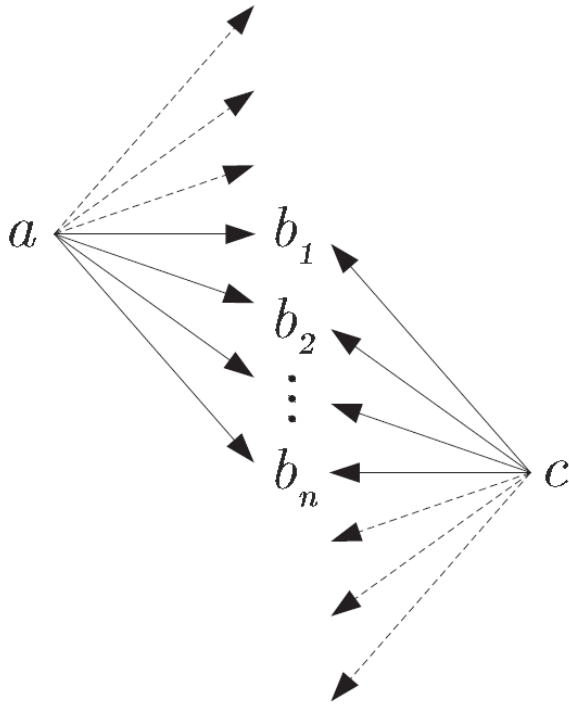
Closed discovery setting: Finding linking (bridging) terms

Literature about magnesium (A): 38,000 articles



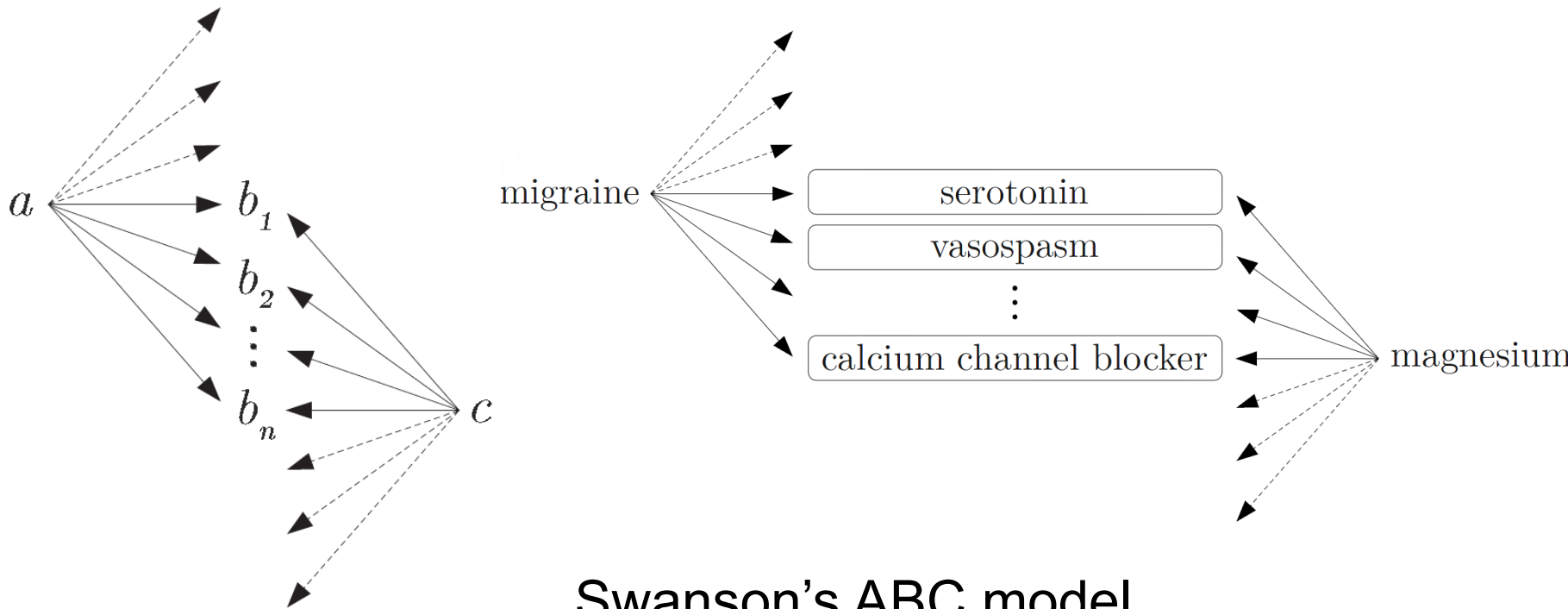
Literature about migraine (C): 4,600 articles

Closed discovery setting: Finding linking (bridging) terms



Swanson's ABC model

Closed discovery setting: Finding linking (bridging) terms



Swanson's ABC model

B-terms: calcium channel blocker, ...

Closed discovery setting: Finding linking (bridging) terms

Argument 1 (magnesium literature)

- Mg is a natural calcium channel blocker.
- Stress and Type A behavior can lead to body loss of Mg.
- Magnesium has anti-inflammatory properties.
- . . .

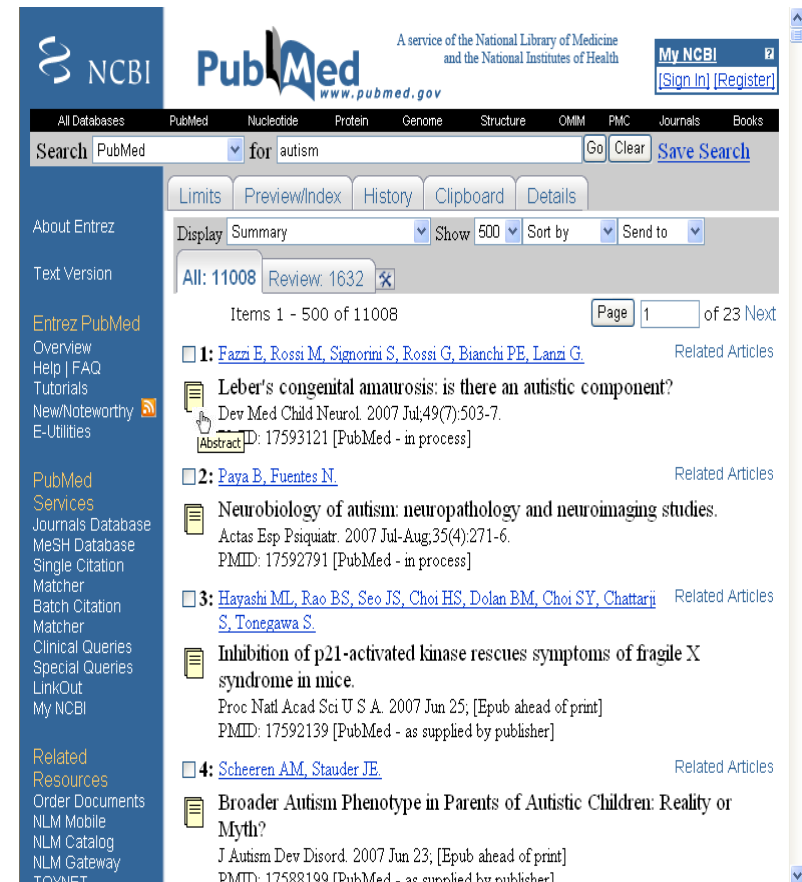
Argument 2 (migraine literature)

- Calcium channel blockers can prevent migraine attacks.
- Stress and Type A behavior are associated with migraine.
- Migraine may involve sterile inflammation of the cerebral blood vessels.
- . . .

Scientific literature as a source of knowledge

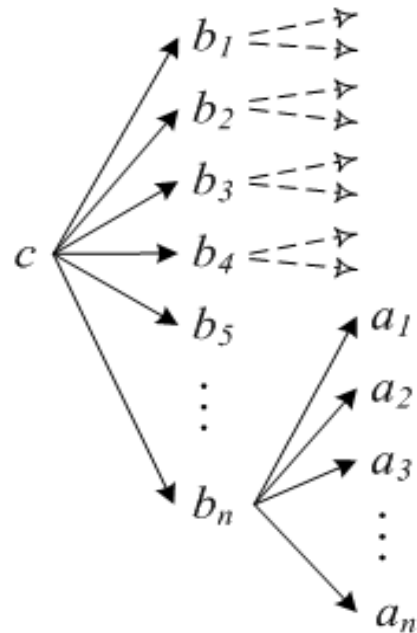
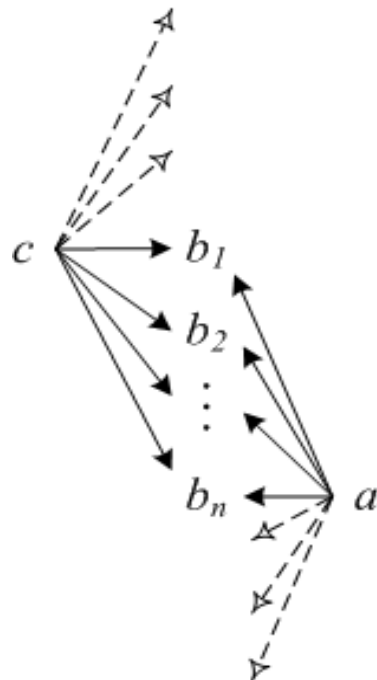
Example:

- Biomedical bibliographical database PubMed
- US National Library of Medicine
- More than 21M citations
- More than 5,600 journals
- 2,000 – 4,000 references added each working day!



Closed vs. open discovery (Weeber et al. 2001)

- **Closed discovery:**
 - A and C are known: Given two separate literatures A and C, find bridging terms B
- **Open discovery:**
 - Only C is known: Given literature C, how do we find A?



Closed vs. open discovery

(Weeber et al. 2001)

- **Closed discovery:**
 - A and C are known: Given two separate literatures A and C, find bridging terms B
- **Open discovery:**
 - Only C is known: Given literature C, how do we find A?
 - Swanson: “Search proceeds via some intermediate literature (B) toward an unknown destination A. ... Success depends entirely on the knowledge and ingenuity of the searcher.”
- **Text mining for cross-domain knowledge discovery:**
 - Can we provide systematic support to the closed and open discovery process ?

Text mining for cross-domain knowledge discovery

- **Situation:**

- Growing speed of knowledge growth, huge amounts of literature available on-line
- High specialization of researchers
- Potentially useful connections between “islands” of knowledge may remain hidden

- **Research objective:**

- To develop methods and text mining tools to support researchers in the discovery of new knowledge from literature

Talk outline

- Background and motivation
- Literature-based discovery
- Cross-domain literature mining approaches
 - Outlier detection for cross-domain knowledge discovery
 - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions

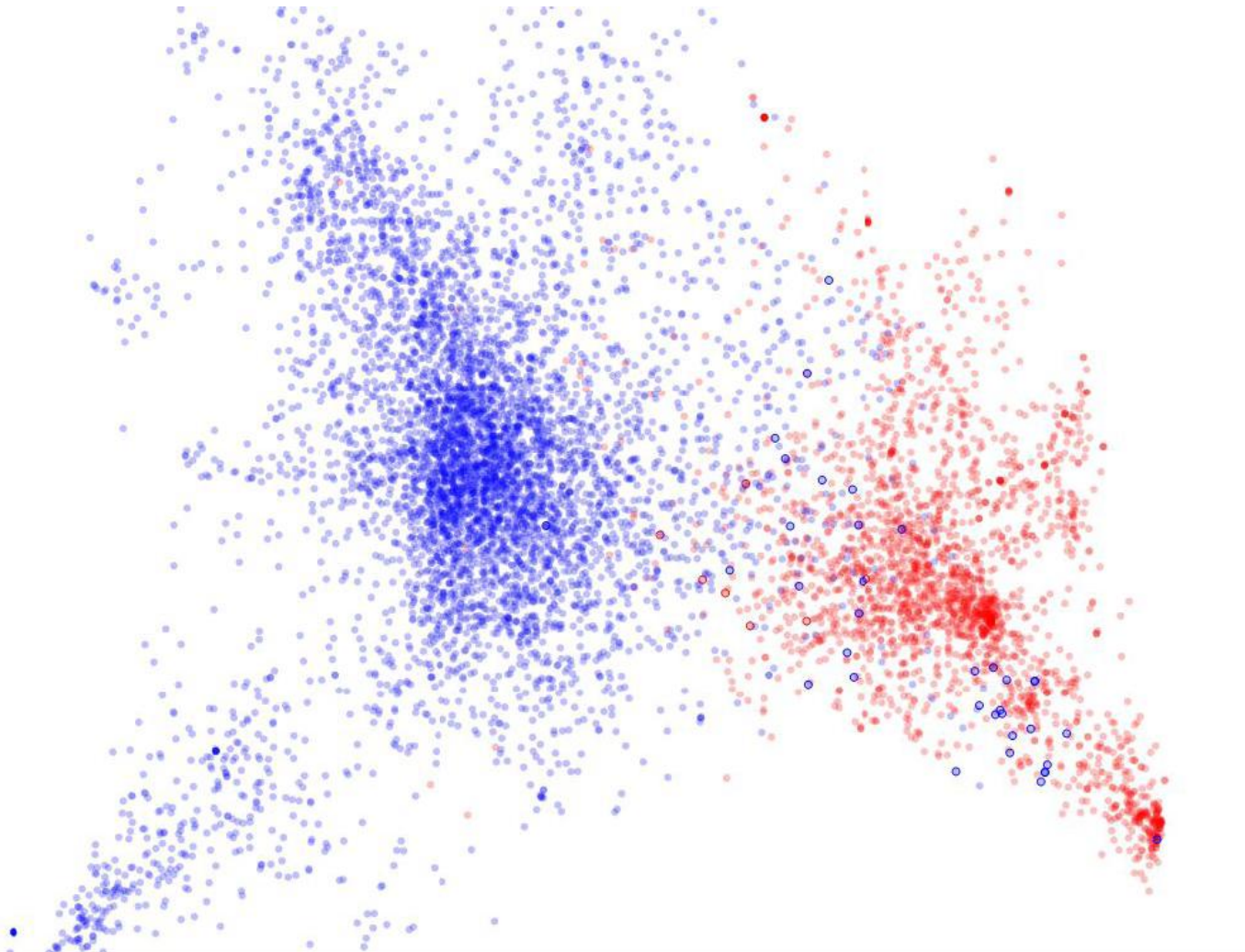
Outlier detection



Outlier detection for cross-domain knowledge discovery

- The goal is to identify interesting **terms** or **concepts** which relate or link separate domains.
⇒ *bridging terms (b-terms) / bridging concepts*
- We explore the utility of *outlier detection* in the task of *cross-domain bridging term discovery*

Outlier detection for cross-domain knowledge discovery



2-dimensional projection of documents (about autism (red) and calcineurin (blue). Outlier documents are bolded for the user to easily spot them.

Our research has shown that most domain bridging terms appear in outlier documents.

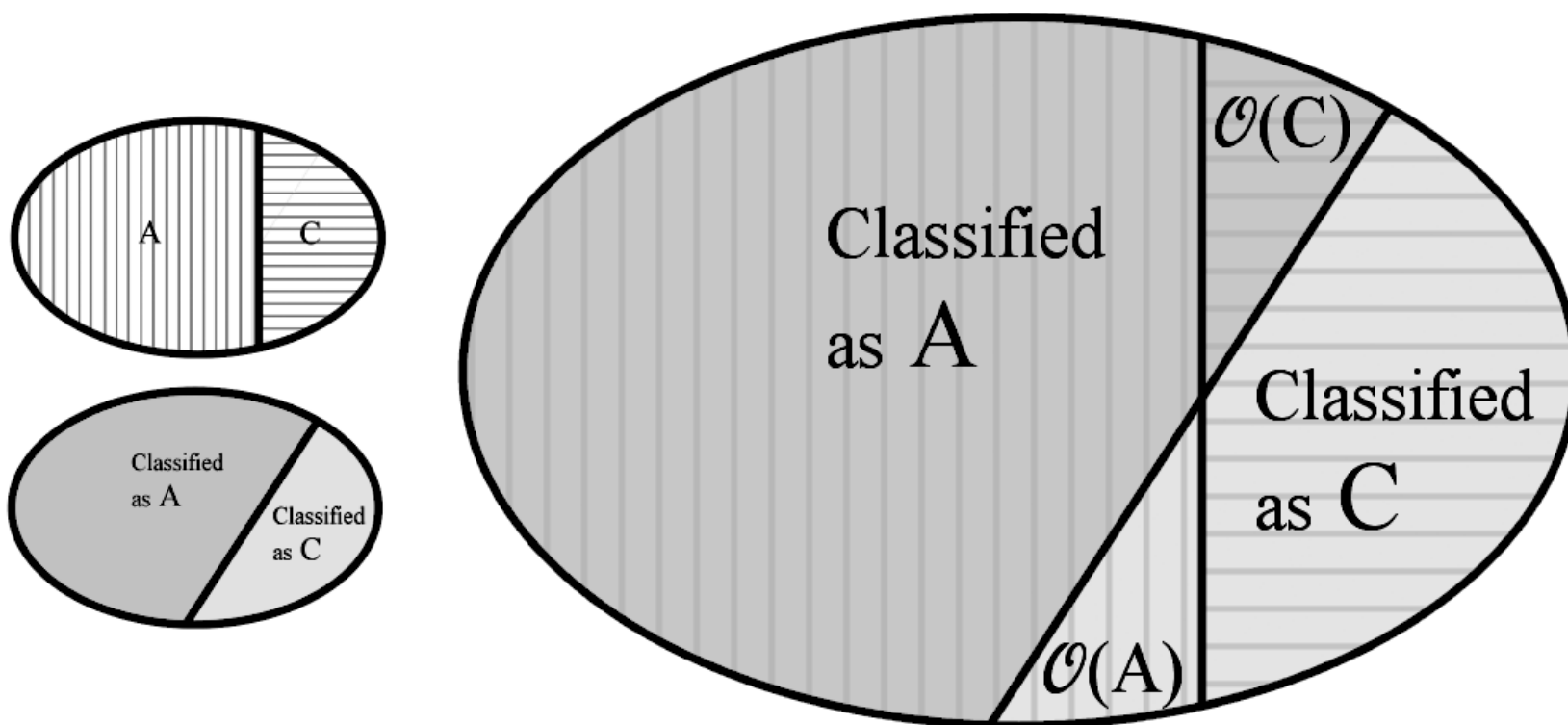
(Lavrač, Sluban, Grčar, Juršič 2010)

Outlier detection for cross-domain knowledge discovery

- Outlier document and bridging term detection
- Three approaches
 - Outlier detection through noise/outlier detection and ranking with NoiseRank
 - Outlier document detection through document clustering with OntoGen
 - Outlier document and outlier term detection using Banded matrices (current work, out of scope of this presentation)

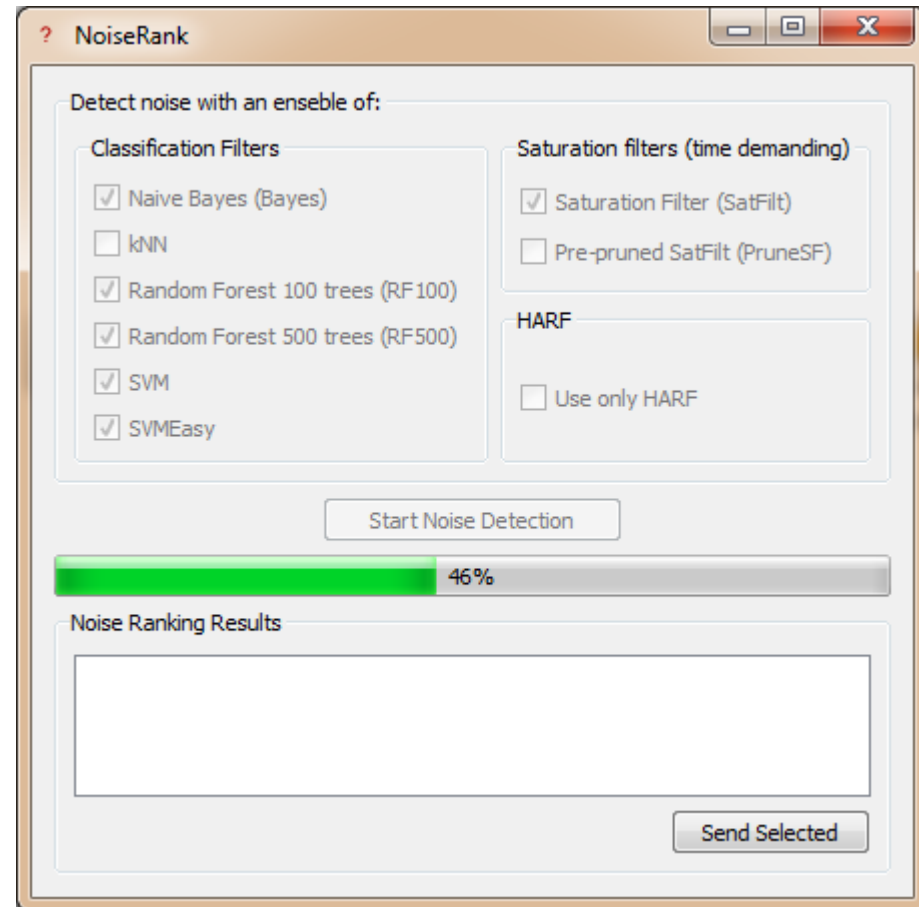
Detecting outlier documents

- By classification noise detection on a domain pair dataset, assuming two separate document corpora



NoiseRank: Ensemble-based noise and outlier detection

- Misclassified document detection by an ensemble of diverse classifiers (e.g., *Naive Bayes*, *Random Forest*, *SVM*, ... classifiers)
- Ranking of misclassified documents by “voting” of classifiers



NoiseRank on news articles

Articles on Kenyan elections: local vs. Western media

Rank | Class | ID | Detected by:

1.	WE	352	__Bayes__	RF100	RF500	SVM	SVMEasy	SatFilt	#HARF#
2.	LO	25	__Bayes	RF100	RF500	SVM	SVMEasy		#HARF#
3.	LO	101	__Bayes	RF100	RF500	SVM	SVMEasy		#HARF#
4.	LO	173	__Bayes	RF100	RF500	SVM	SVMEasy		#HARF#
5.	WE	348	__Bayes	RF100	RF500	SVM	SVMEasy		#HARF#
6.	WE	326	__Bayes	RF100	RF500	SVM	SVMEasy		
7.	WE	357	__Bayes	RF100	RF500	SVM	SatFilt		
8.	WE	410	__Bayes	RF100	RF500	SVM	SVMEasy		
9.	LO	21	RF100	RF500	SVM	SVMEasy			#HARF#
10.	LO	4	__Bayes	RF500	SVM	SVMEasy			
11.	LO	68	RF100	RF500	SVM	SVMEasy			
12.	LO	162	__Bayes	RF500	SVM	SVMEasy			
13.	WE	358	__Bayes	RF100	RF500	SVM			
14.	WE	464	RF100	RF500	SVM	SVMEasy			
15.	LO	153	__Bayes	SVM	SVMEasy				
16.	LO	201	RF100	RF500	SatFilt				
17.	WE	238	RF100	RF500	SVM				
18.	WE	364	__Bayes	RF500	SVM				
19.	WE	370	__Bayes	RF100	SVM				
20.	WE	379	RF100	RF500	SVMEasy				

NoiseRank on news articles

- **Article 352: Out of topic**

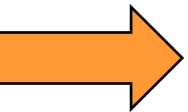
The article was later indeed removed from the corpus used for further linguistic analysis, since it is not about Kenya(ns) or the socio-political climate but about British tourists or expatriates' misfortune.

- **Article 173: Guest journalist**

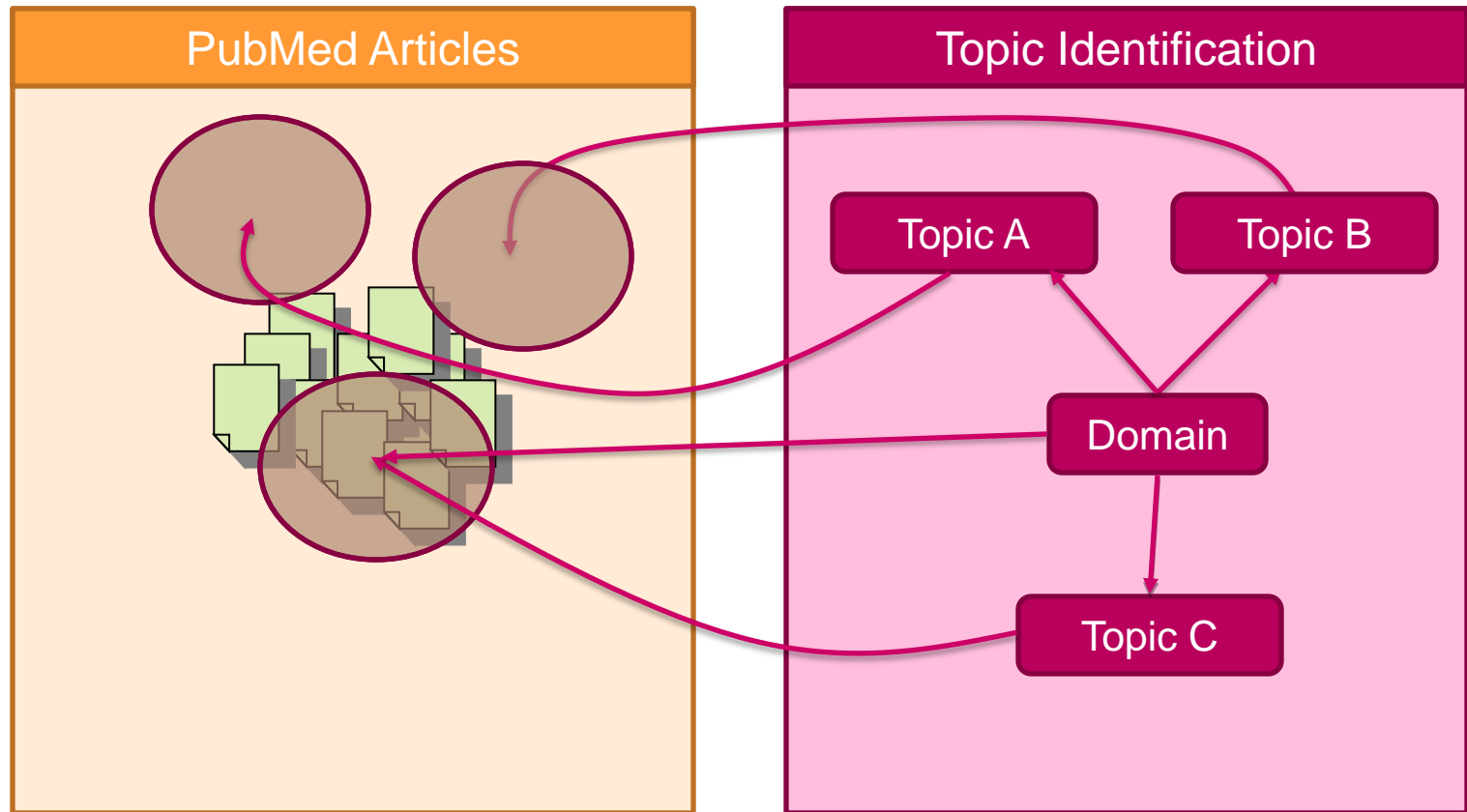
Wrongly classified because it could be regarded as a “Western article” among the local Kenyan press. The author does not have the cultural sensitivity or does not follow the editorial guidelines requiring to be careful when mentioning words like tribe in negative contexts. One could even say that he has a kind of “Western” writing style.

Talk outline

- Background and motivation
- Literature-based discovery
- Cross-domain literature mining approaches
 - Outlier detection for cross-domain knowledge discovery
 - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions

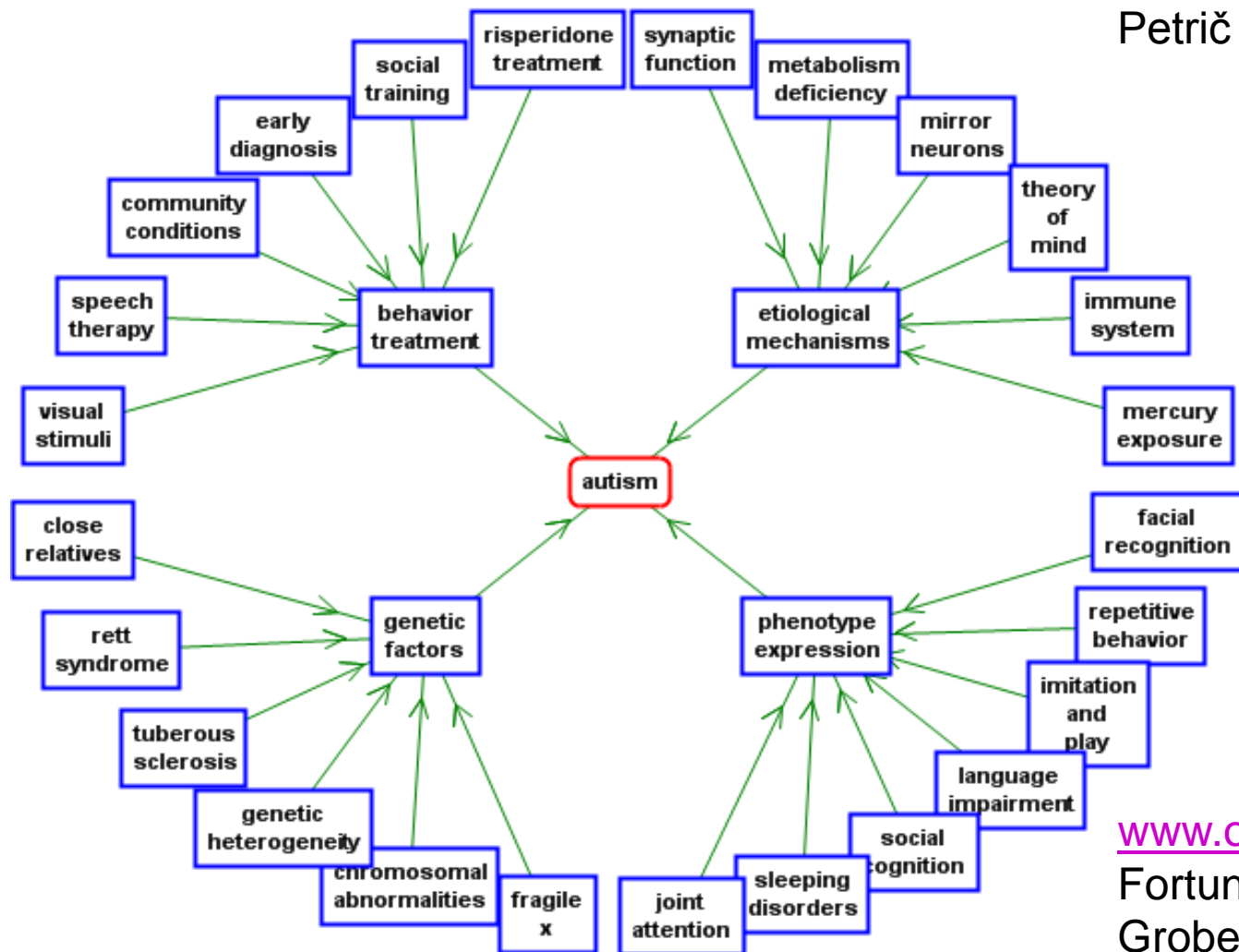


Outlier detection by clustering of PubMed articles



Using OntoGen for clustering PubMed articles on autism

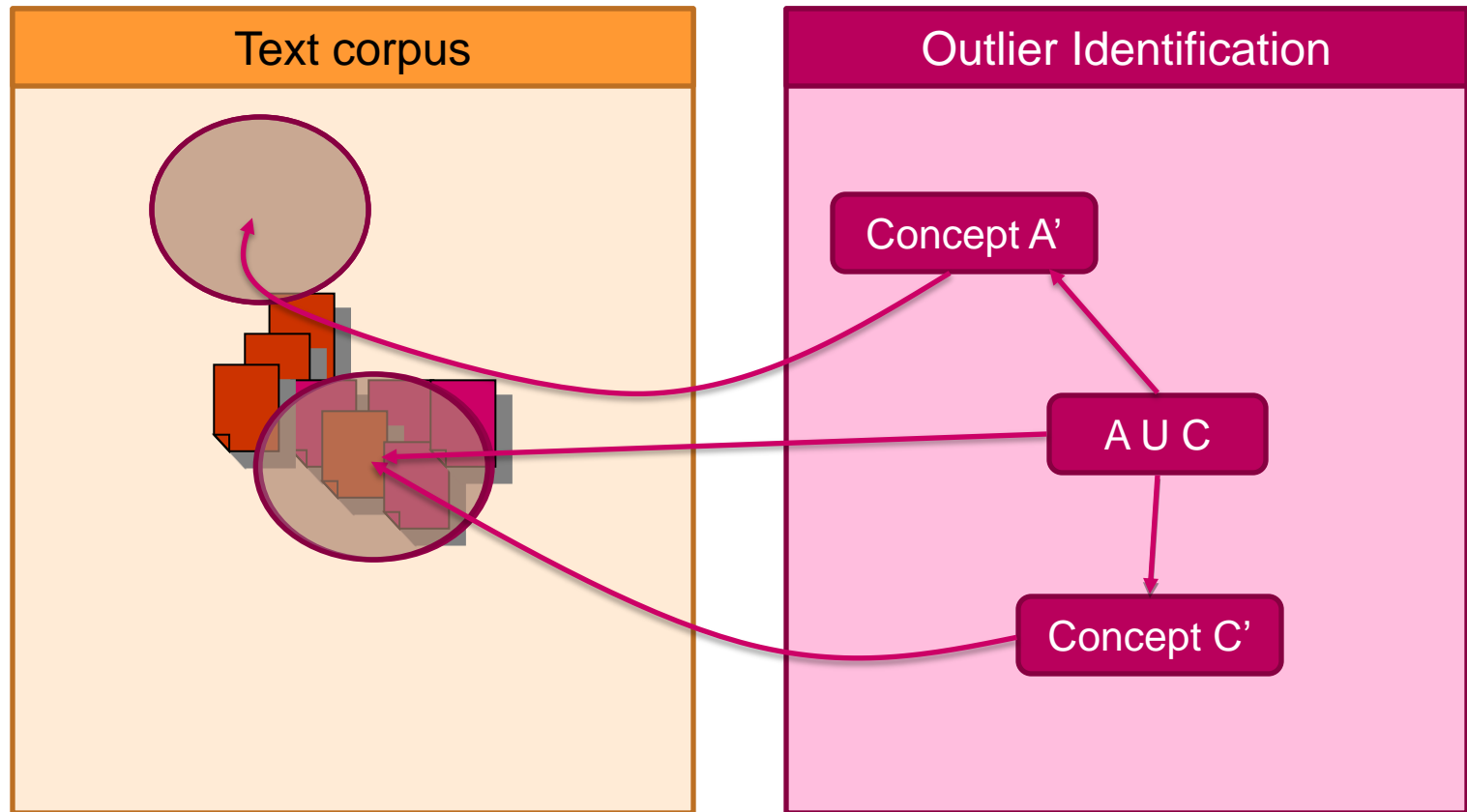
Work by
Petrič et al. 2009



www.ontogen.si

Fortuna, Mladenić,
Grobelnik 2006

Using OntoGen for outlier document identification



Results on autism-calcineurin: Outlier calcineurin document CN423

The screenshot displays the OntoGen Text Garden software interface. The main window is titled "OntoGen -- Text Garden" and contains several panels:

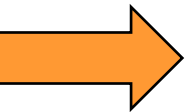
- Concepts:** A tree view showing a hierarchy starting from "root", branching into "A' autism" and "C' calcineurin".
- Concept properties:** A section with tabs for "Details", "Suggestions", and "Relations". The "Details" tab is active, showing fields for "Name" (set to "A' autism"), "Keywords" (including "children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation"), "SVM Keywords", and document counts (All documents: 10285, Unused documents: 10285).
- Ontology details:** A section with tabs for "Ontology visualization", "Concept's documents", and "Concept Visualization". The "Concept's documents" tab is active, showing a list of documents sorted by similarity. The list includes documents 3874, 8939, CN1065, 6372, 2402, CN3661, 220, 7163, 6864, 7686, CN3207, CN423, 5168, CN2549, and 4072, all with a similarity score of 0.146. Document CN423 is highlighted.
- Document details:** A panel on the right showing the full text of document CN423. The text describes calcineurin as a neuron-enriched phosphatase that regulates synaptic plasticity and neuronal adaptation. It mentions that activation of calcineurin, overall, antagonizes the effects of the cyclic AMP activated protein/kinase A. The text also discusses the critical role of the kinase/phosphatase dynamic balance in neurons and the potential behavioral impairments in animal models.
- Keywords for selected documents:** A section below the document list showing keywords for the selected documents: "children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation".

At the bottom of the interface, there is a "Document name:" field and a "Calc" button. The overall layout is typical of a web-based application from the early 2000s.


Work by
Petrič et al. 2010

Talk outline



- Background and motivation
- Literature-based discovery
- Cross-domain literature mining approaches
 - Outlier detection for cross-domain knowledge discovery
 - Cross-domain knowledge discovery with CrossBee
- Summary and conclusions



CrossBee: Cross Context Bisociation Explorer



CROSS BEE
CROSS CONTEXT BISOCIATION EXPLORER

Supported by

SEVENTH FRAMEWORK PROGRAMME

StartDownloadsTerm ViewDocument ViewBTerms

SEARCH

MAIN MENU

ITEM BASKET

B-Term Identify (Term "paroxysmal" Analysis)

<< Start < Previous | 1 - 10 of 10 | Next > End >> << Start < Previous | 1 - 3 of 3 | Next > End >>

2270. **Paroxysmal** and other **features** of th...

1012. **Paroxysmal** dysequilibrium in the **mi...**

2164. **Paroxysmal** supraventricular tachyc...

1152. **Migraine** as a **cause** of benign **parox...**

1393. The distinction between **paroxysmal** ...

1868. [Benign **paroxysmal** vertigo of childh...

1605. Benign **paroxysmal** vertigo in childh...

2241. Benign **paroxysmal** vertigo of childh...

503. **Chronic paroxysmal migraine**. A rev...

1104. **Paroxysmal** arrhythmias and **migraine...**

3456. [A **case** of **paroxysmal** tachycardia o...

3263. **Spontaneous paroxysmal activity** ind...

4678. **Paroxysmal nocturnal** hemoglobinuria...

Document: #2270
Go in depth, Add to basket
Domain: MIG

Paroxysmal and other **features** of the electroencephalogram in **migraine**.

Document's Important Terms (ordered by importance):

1. **paroxysmal** (0,999)
2. **migraine** (0,855)
3. **feature** (0,564)
4. **electroencephalogram migraine** (0,053)
5. **electroencephalogram** (0,029)

Document's Important Terms (ordered by alphabet):

1. **electroencephalogram** (0,029)
2. **electroencephalogram migraine** (0,053)
3. **feature** (0,564)
4. **migraine** (0,855)
5. **paroxysmal** (0,999)

Document: #3456
Go in depth, Add to basket
Domain: MAG

[A **case** of **paroxysmal** tachycardia of the torsade de pointes **type**: the role of **magnesium** in the **etiology** and **treatment**]

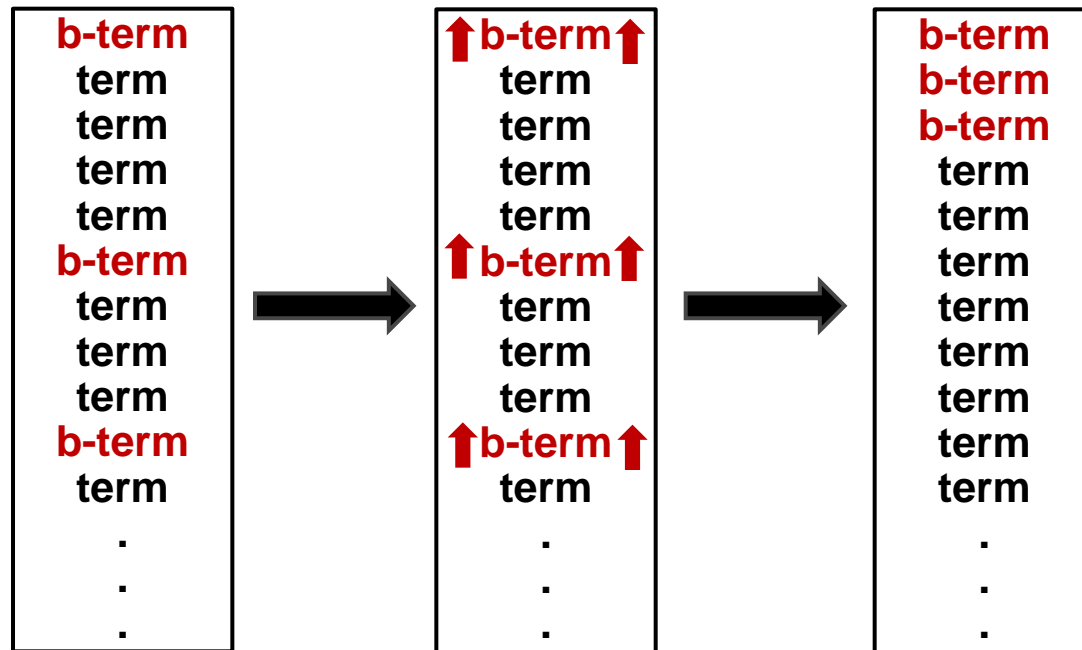
Document's Important Terms (ordered by importance):

1. **paroxysmal** (0,999)
2. **case** (0,855)
3. **treatment** (0,712)
4. **type** (0,711)
5. **etiology** (0,711)
6. **magnesium** (0,568)
7. **role** (0,424)
8. **tachycardia** (0,421)
9. **etiology treatment** (0,277)
10. **de** (0,086)
11. **role magnesium** (0,077)

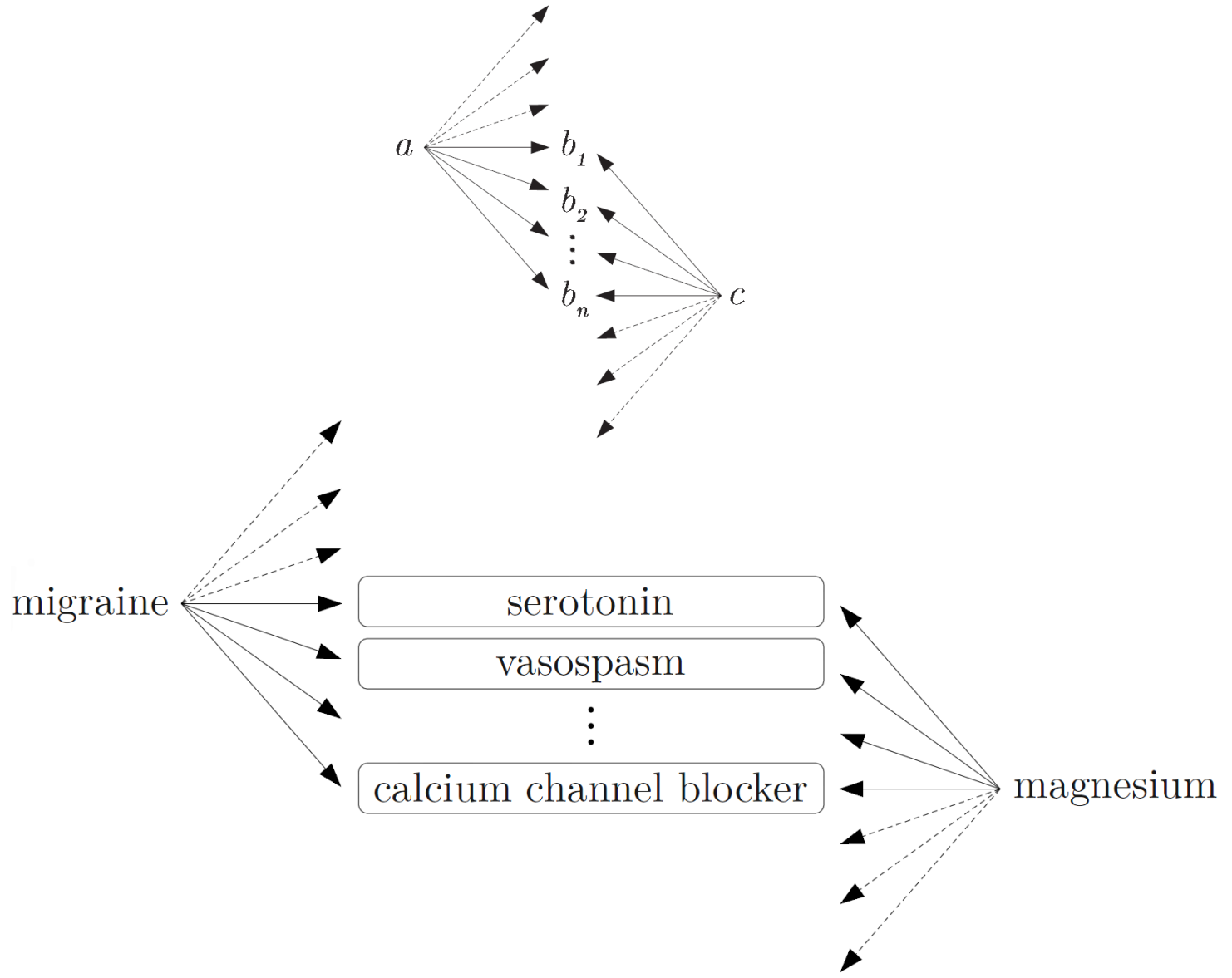
The research was supported by the European Commission under the 7th Framework Programme FP7 ICT 2007 C FET Open project BISON 211898.

Problem definition

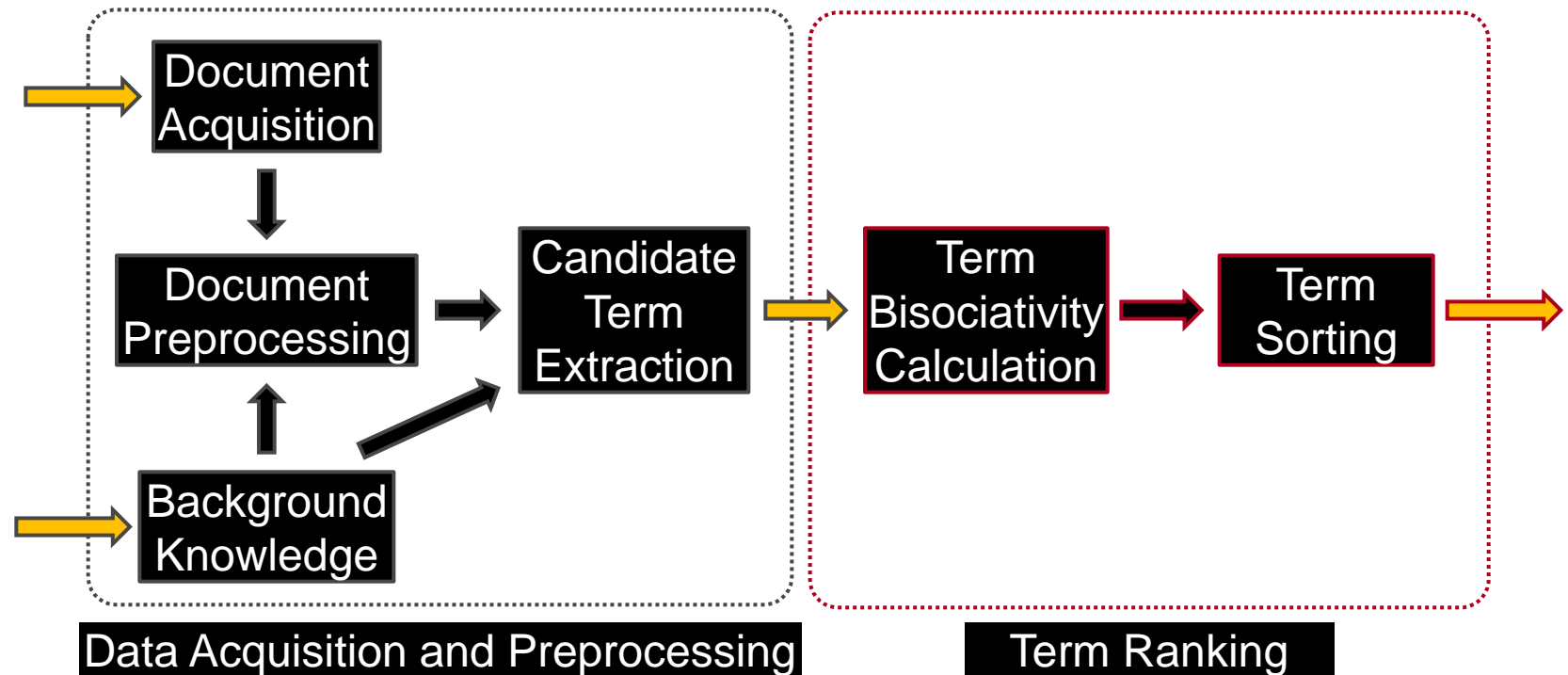
Goal: Develop a term ranking methodology that ranks high all the terms which have high bisociation potential (denoted as *bridging* terms or *b-terms*)



Closed discovery setting



CrossBee: Methodology overview

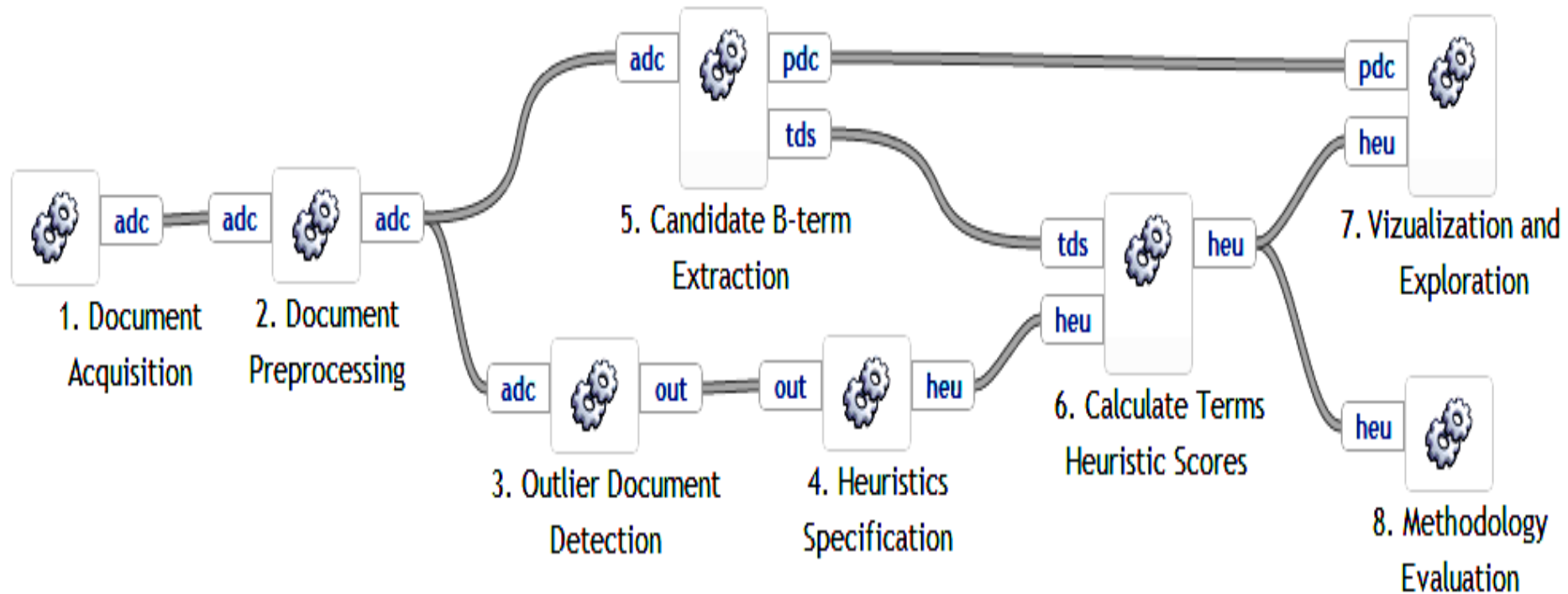


Incorporating available background knowledge

Vocabularies: e.g. for word/term filtering

Ontologies: e.g. for enriching documents term sets

Methodology implementation



Data acquisition and preprocessing

- Document acquisition from the Web
 - Acquiring documents from PubMed
 - Snippets returned from web search engines
 - Crawling the Internet and gathering documents from web pages
- Document preprocessing
 - Tokenization
 - Stopwords removal
 - Stemming or lemmatization: LemmaGen
 - Part of speech tagging or syntactic parsing
- Candidate term extraction
 - Frequent n-grams in preprocessed documents

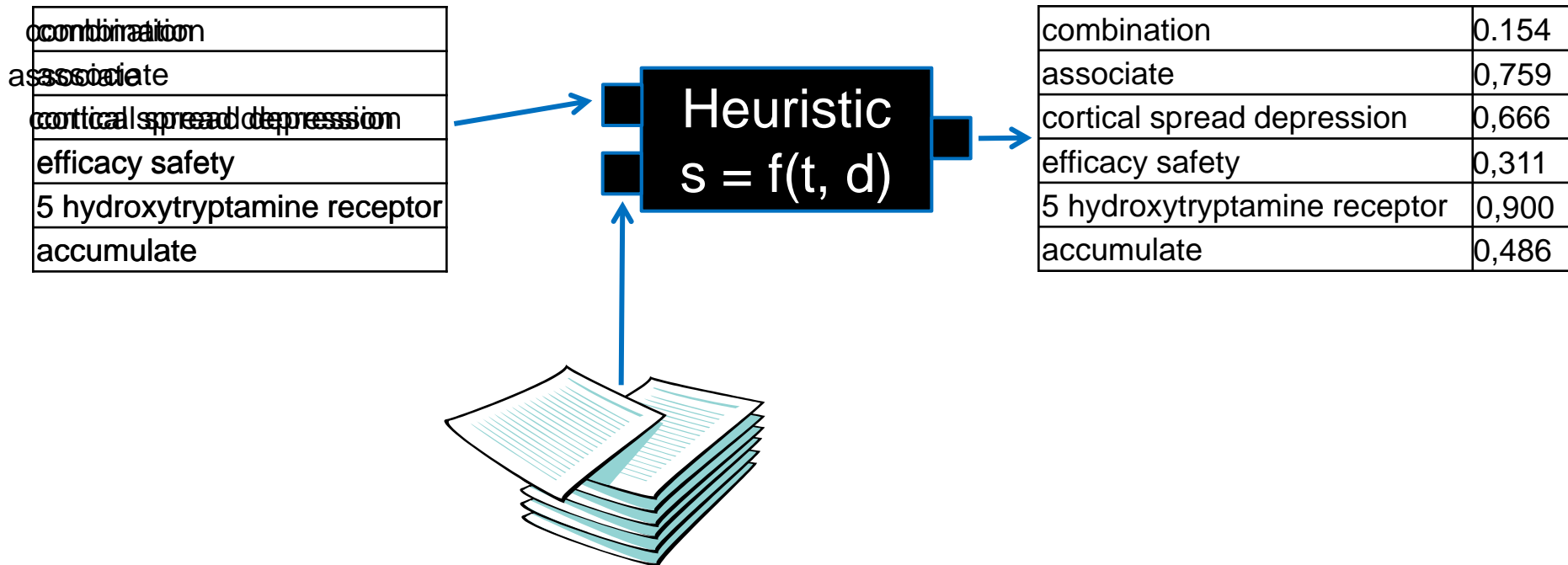
Term ranking

- Term ranking:
 - Assign scores to all the terms
 - Sort the terms according to the assigned scores
- How to assign scores to terms?
 - Using a heuristic function that estimates the probability that a term is b-term
- How to construct the “optimal” heuristic using training data?
 1. Create several promising heuristics
 2. Evaluate the constructed heuristics on a training dataset
 3. Construct the ensemble heuristic using the best individual heuristics
 4. Use the ensemble heuristic for scoring the terms

Heuristic function

- Input: a term with its statistic properties calculated from texts
- Output: a number [0,1] which ranks the term (its probability of being a b-term)

Ideal heuristic: such that ranks all true b-terms very high and all the others lower



Bisociation potential heuristics

- Heuristics can be grouped based on:
 - frequency (variations of the term occurrences)
 - $freqTerm(t) = countTerm_{D_u}(t)$: term frequency across both domains
 - tf-idf (combinations of tf-idf weights of a term)
 - $tfidfDomnProd(t) = tfidf_{D_1}(t) \cdot tfidf_{D_2}(t)$: product of a term's importance in both domains
 - similarity (similarity of a term to the average terms)
 - outliers (frequency of a term in documents at the border of the two domains)
 - $outFreqRelRF(t) = \frac{countTerm_{D_{RF}}(t)}{countTerm_{D_u}(t)}$: relative frequency in RF outlier set

Ensemble heuristic

heuristic 1
heuristic 2
heuristic 3

ensemble heuristic

heuristic 1

term 1	0,149
term 2	0,759
term 3	0,900
term 4	0,666
term 5	0,311
term 6	0,071
term 7	0,175
term 8	0,637
term 9	0,429
.	.
.	.
.	.

heuristic 2

term 1	0,429
term 2	0,149
term 3	0,071
term 4	0,175
term 5	0,637
term 6	0,759
term 7	0,970
term 8	0,636
term 9	0,311
.	.
.	.
.	.

heuristic 3

term 1	0,680
term 2	0,311
term 3	0,071
term 4	0,175
term 5	0,637
term 6	0,429
term 7	0,149
term 8	0,759
term 9	0,980
.	.
.	.
.	.

Ensemble heuristic

heuristic 1

term 3
term 2
term 1
term 8
term 9
term 5
term 7
term 4
term 6
.
.
.

heuristic 2

term 7
term 6
term 5
term 8
term 1
term 9
term 4
term 2
term 3
.
.
.

heuristic 3

term 7
term 8
term 1
term 5
term 6
term 2
term 4
term 7
term 9
.
.
.

ensemble heuristic

term 1	2
term 2	1
term 3	1
term 4	0
term 5	2
term 6	1
term 7	2
term 8	3
term 9	0
.	.
.	.
.	.

Ensemble heuristic

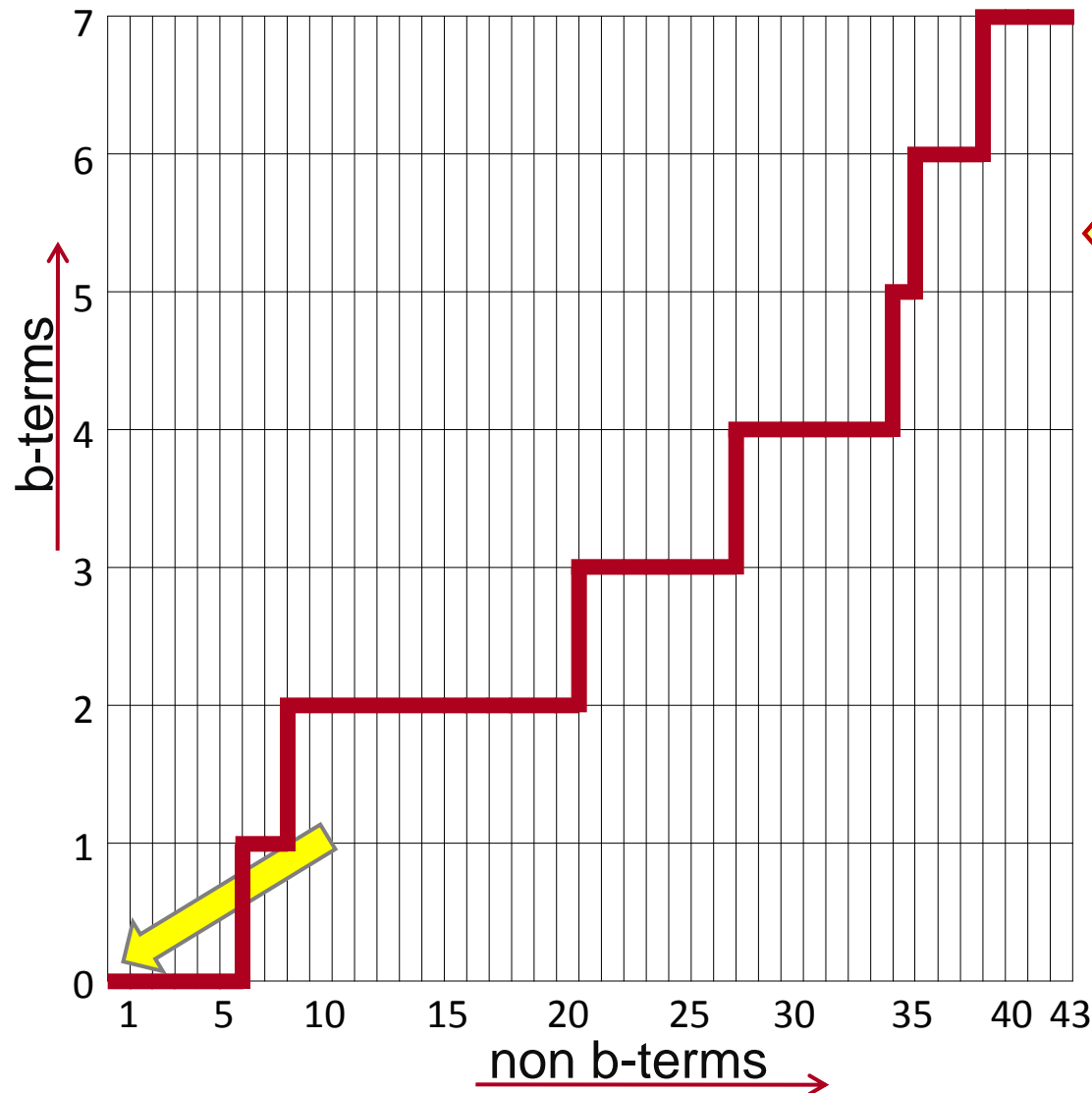
final ensemble heuristic

term 8	heuristic 1, heuristic 2, heuristic 3	term 8
term 1	heuristic 1, heuristic 3	term 1
term 5	heuristic 2, heuristic 3	term 5
term 7	heuristic 2, heuristic 3	term 7
term 2	heuristic 1	term 2
term 3	heuristic 1	term 3
term 6	heuristic 2	term 6
term 7	-	term 7
term 9	-	term 9
.	.	.
.	.	.
.	.	.

Domains and datasets

- Training dataset: migraine-magnesium
 - 8,058 documents (2,425- 5,633), 13,433 distinct terms
 - 43 expert identified b-terms (work by Swanson, D. R., Smalheiser, N. R., Torvik, V. I.: Ranking indirect connections in literature-based discovery : The role of Medical Subject Headings (MeSH))
- Test dataset: autism-calcineurin
 - 22,262 documents (14,890-7,372), 17,514 distinct terms
 - 12 expert identified b-terms (work by Petric, I., Urbancic, T., Cestnik, B., Macedoni-Luksic, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts)

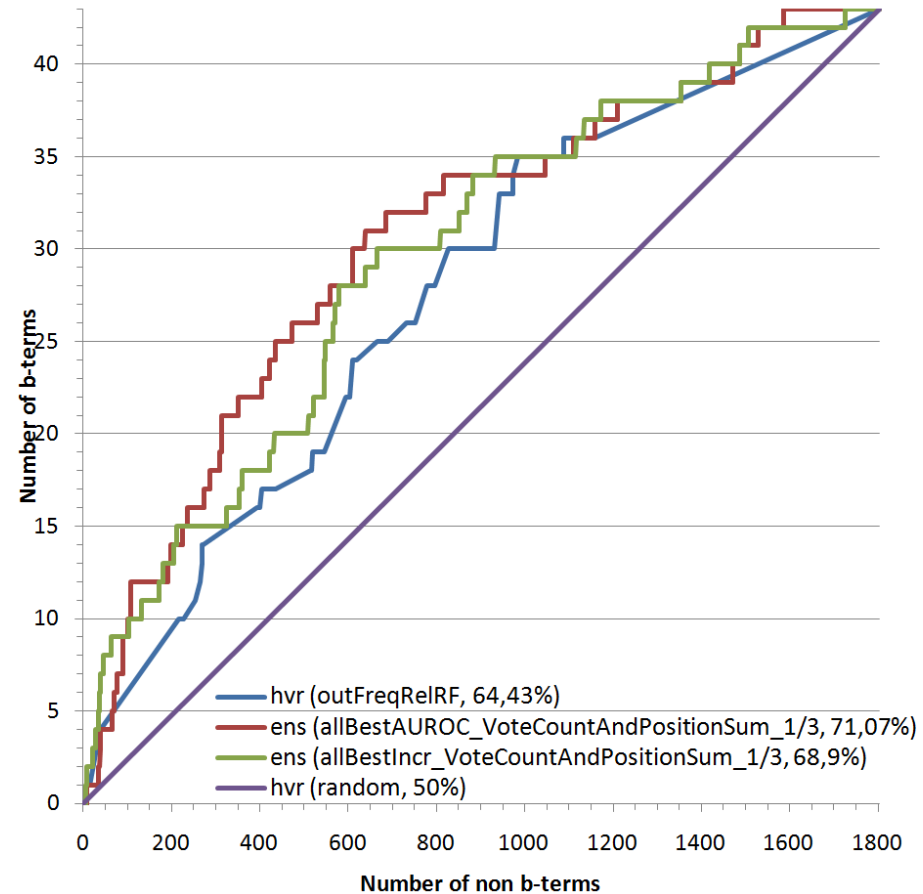
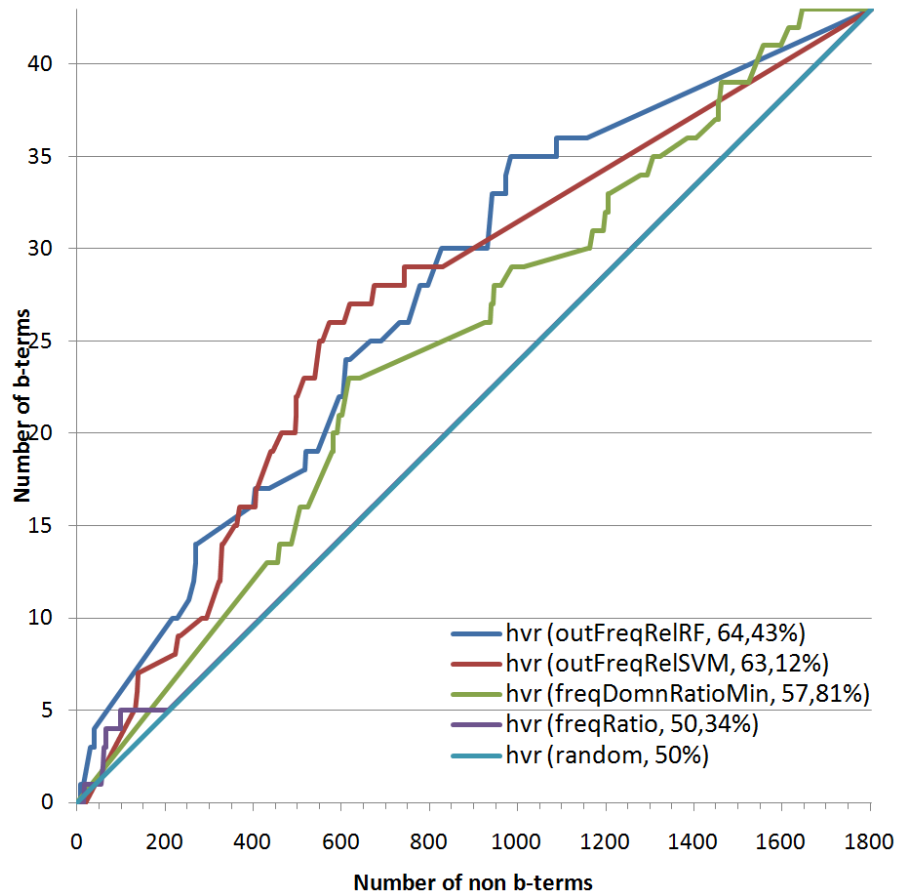
Evaluation ROC curve construction



Ranked term list:
50 terms = 7 b-terms +
43 non b-terms

400
animal human
anti inflammatory agent
basal
bruxism
biochemical aspect
brain serotonin
arteriopathy
cerebral artery
cerebral vasospasm
child treatment
clinical comparative
clinical form
clinical statistical
combination treatment
comparative double
comparative double blind

Results on training data set

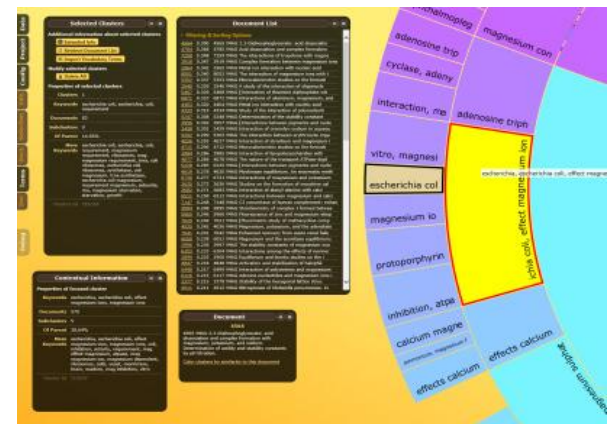
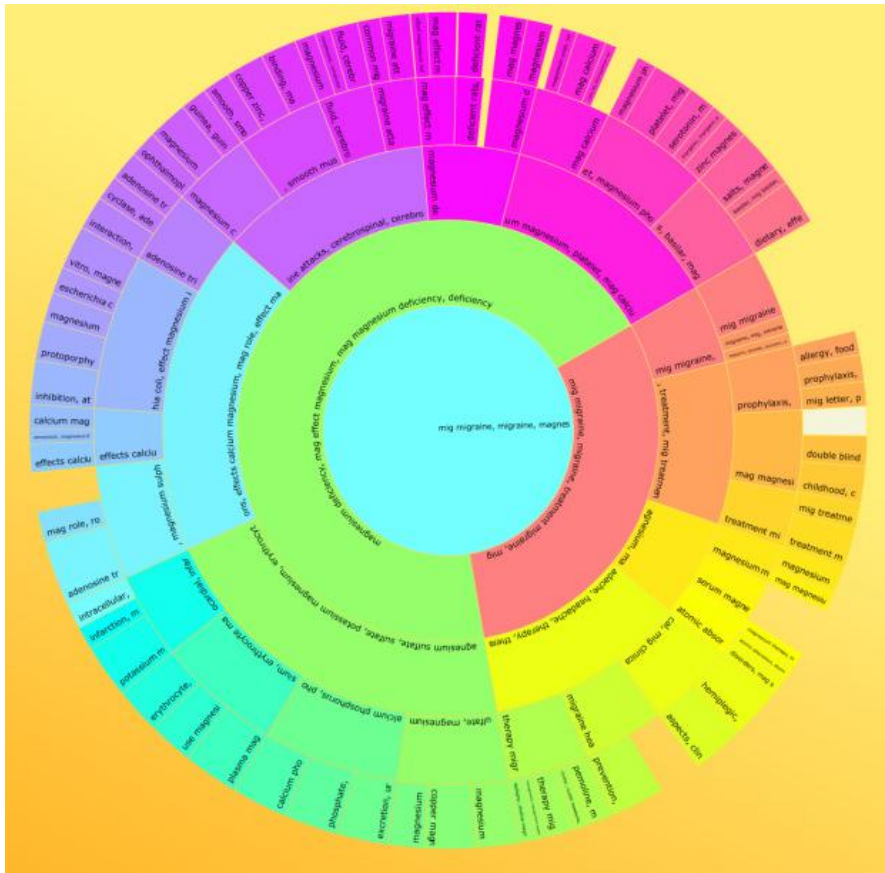


CrossBee system

- Cross Context Bisociation Explorer
- What is CrossBee?
- Web user interface which fuses multiple approaches developed for discovering bisociations in text
- Why CrossBee?
- Collaborating with domain experts on their data in real time on user friendly system (and thus evaluating their and our hypotheses)

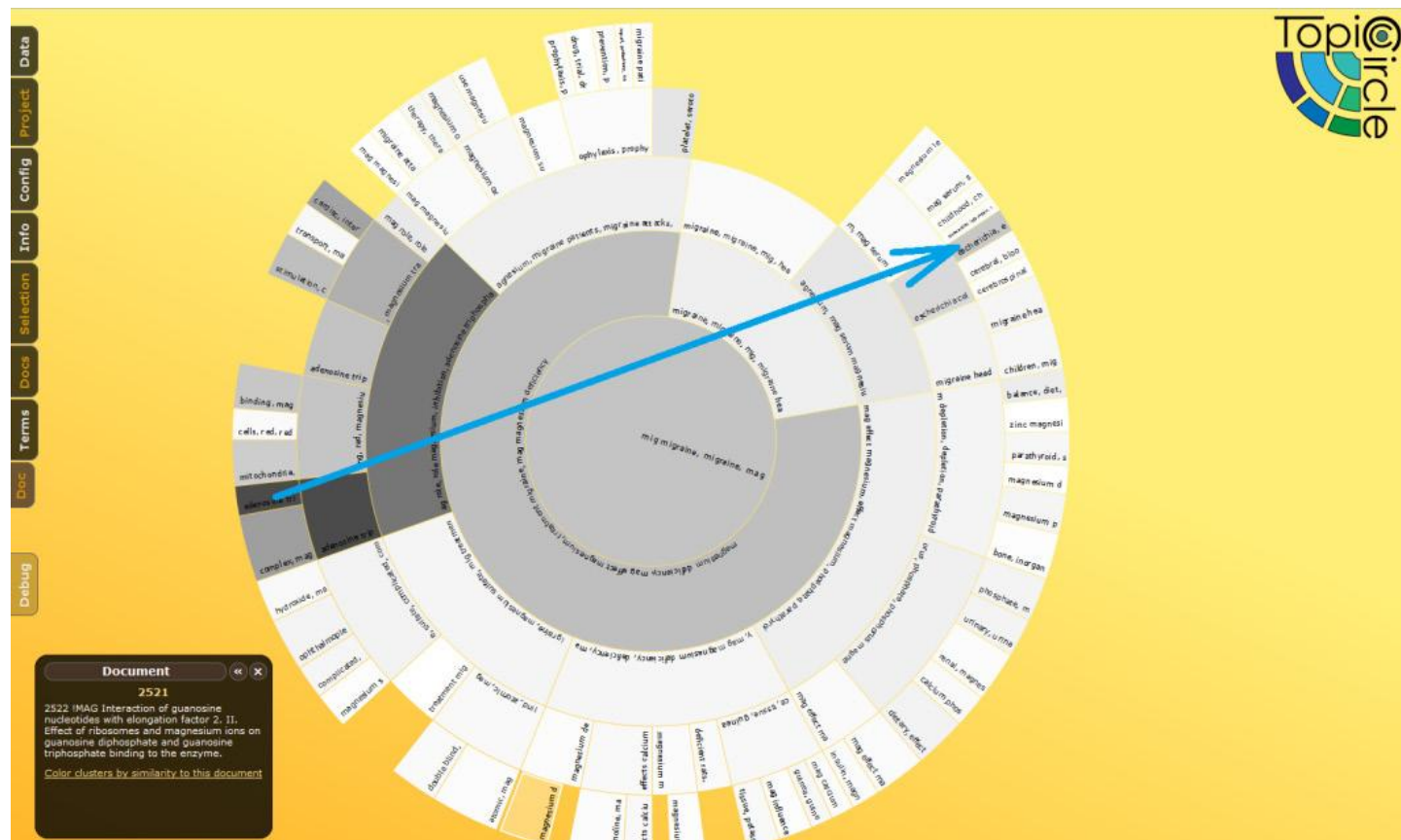
Additional CrossBee functionality

CrossBee Topic Circle for top-down document clustering



Additional CrossBee functionality

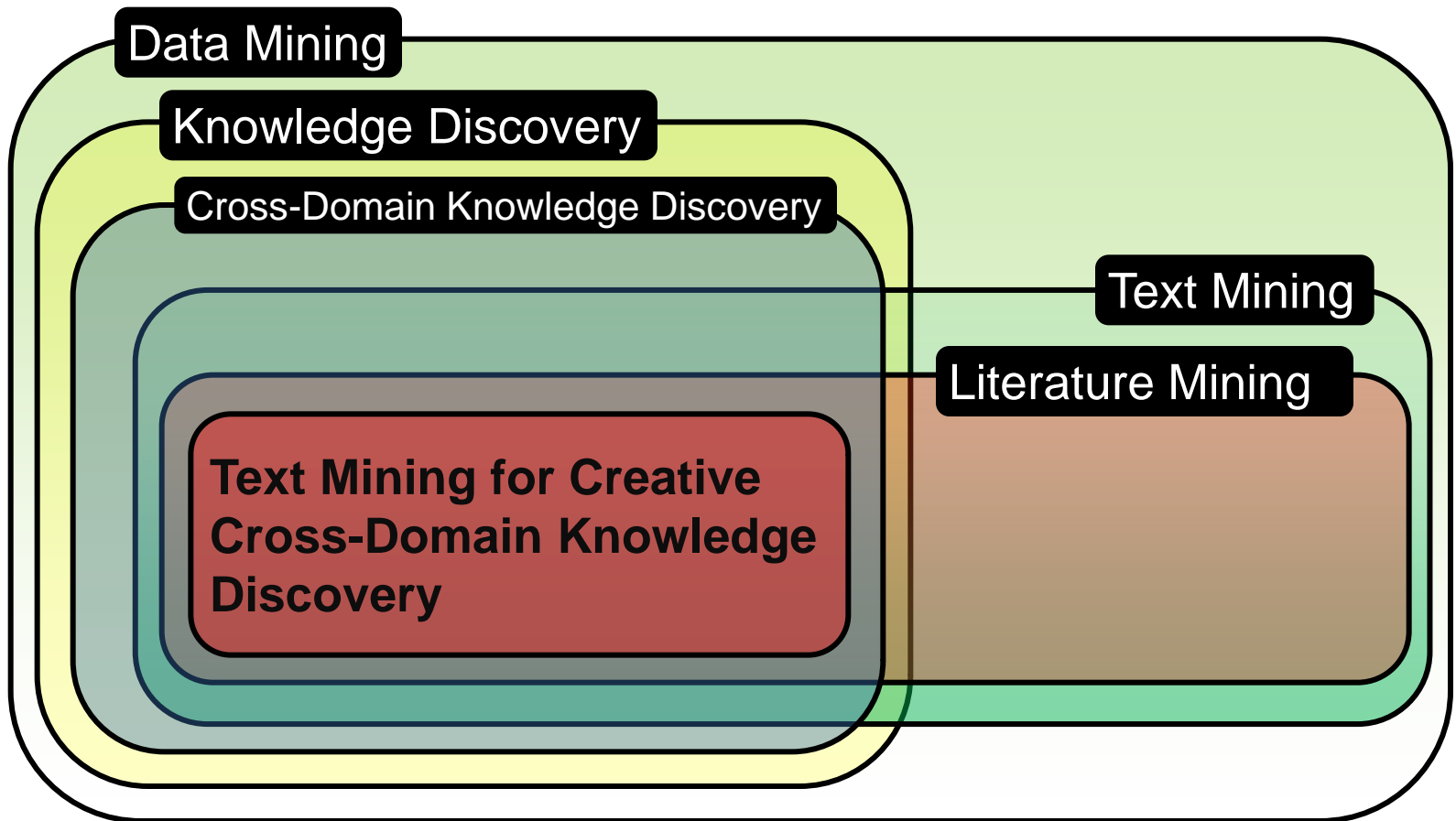
Cluster colors can show e.g., cluster's similarity to a single selected document. The arrow shows similar clusters in two different domains, potentially indicate to a novel bisociative link between the two domains.



Summary and conclusions

- Current literature-based approaches mostly depend on simple associative information search
- Potential of outlier detection for b-term discovery
 - Document outlier detection and ranking by NoiseRank
 - Document outlier detection by OntoGen
- CrossBee: improving computational creativity by supporting the expert in the task of cross-domain literature mining (novelty: ensemble-based bridging term ranking)

Summary and conclusions



Selected readings

- M. Berthold (2012): Bisociative Knowledge Discovery, Springer (open access)
- Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Cross-domain literature mining: Finding bridging concepts with CrossBee. In: Proc. 3rd International Conference on Computational Creativity (2012)
- Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: HCI empowered literature mining for cross-domain knowledge discovery. In: Proc. HCI-KDD, pp. 124-135, Springer (2013)
- Petrič, I., Urbančič, T., Cestnik, B., Macedoni-Lukšič, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics. vol. 42/2, pp. 219–227 (2009)

Selected readings

- Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier Detection in Cross-Context Link Discovery for Creative Literature Mining. *Computer Journal* 55/1, pp. 47–61 (2012)
- Sluban, B., Gamberger, D., Lavrač, N. Ensemble-based noise detection : noise ranking and visual performance evaluation. *Data mining and knowledge discovery* (2013)
- Swanson, D. R.: Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc.* vol. 78/1, pp. 29–37 (1990)
- Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech.* vol. 52/7, pp. 548–57 (2001)