

# **Text Mining for Creative Cross-Domain Knowledge Discovery**

**Nada Lavrač**

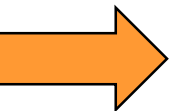
Jožef Stefan Institute, Ljubljana, Slovenia

with contributors

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

JOTA, FRI, 29.3.2016

# Talk outline

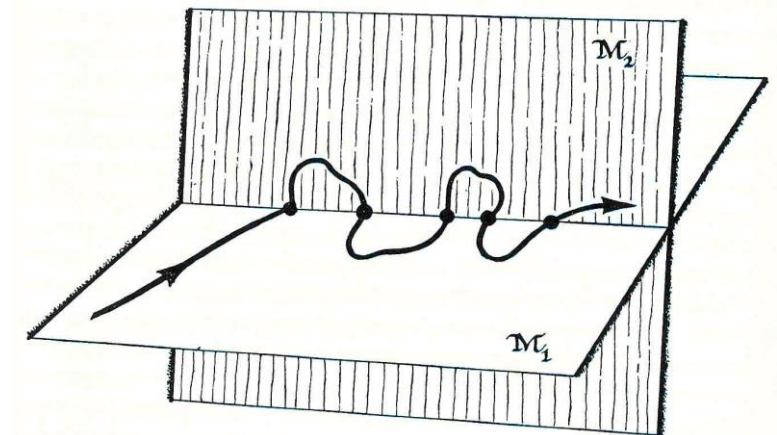
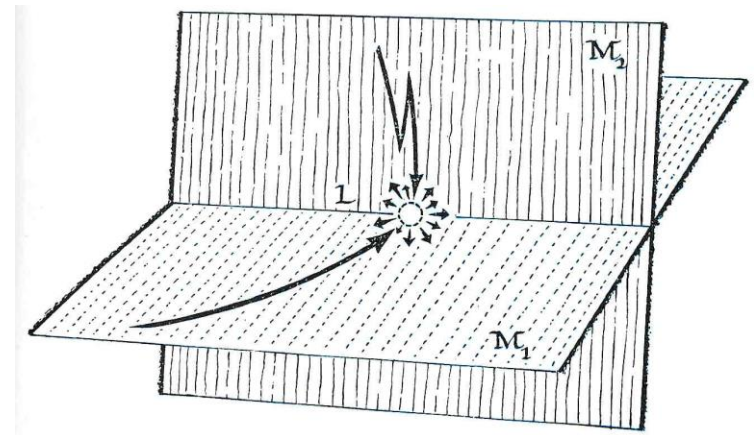
- 
- Background and motivation
    - Literature-based discovery
    - Cross-domain literature mining approaches
      - Outlier detection for cross-domain knowledge discovery
      - Cross-domain knowledge discovery with CrossBee
    - TextFlows text mining platform
    - Summary and conclusions

# 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.

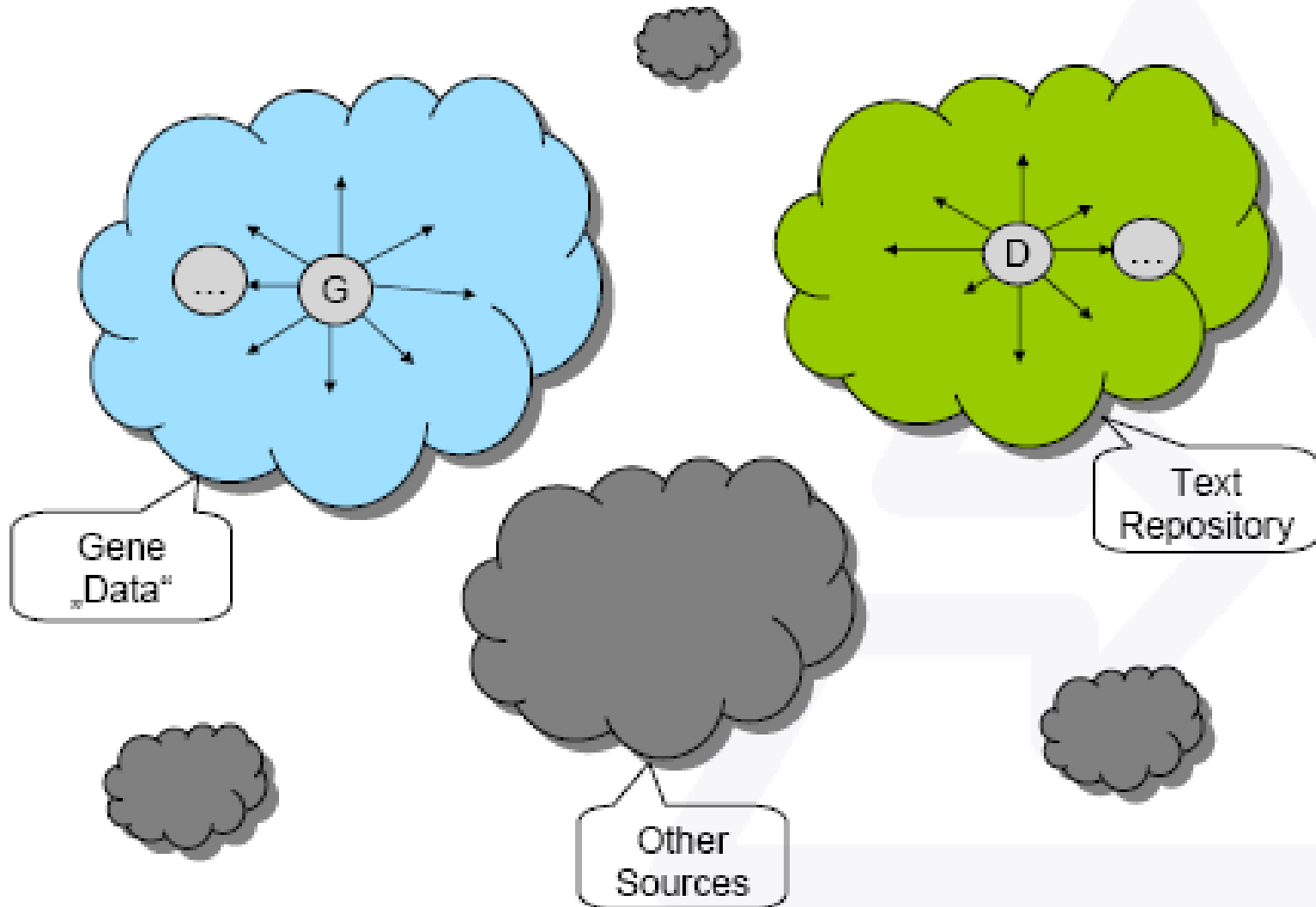
# Bisociation discovery in BISON

- 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



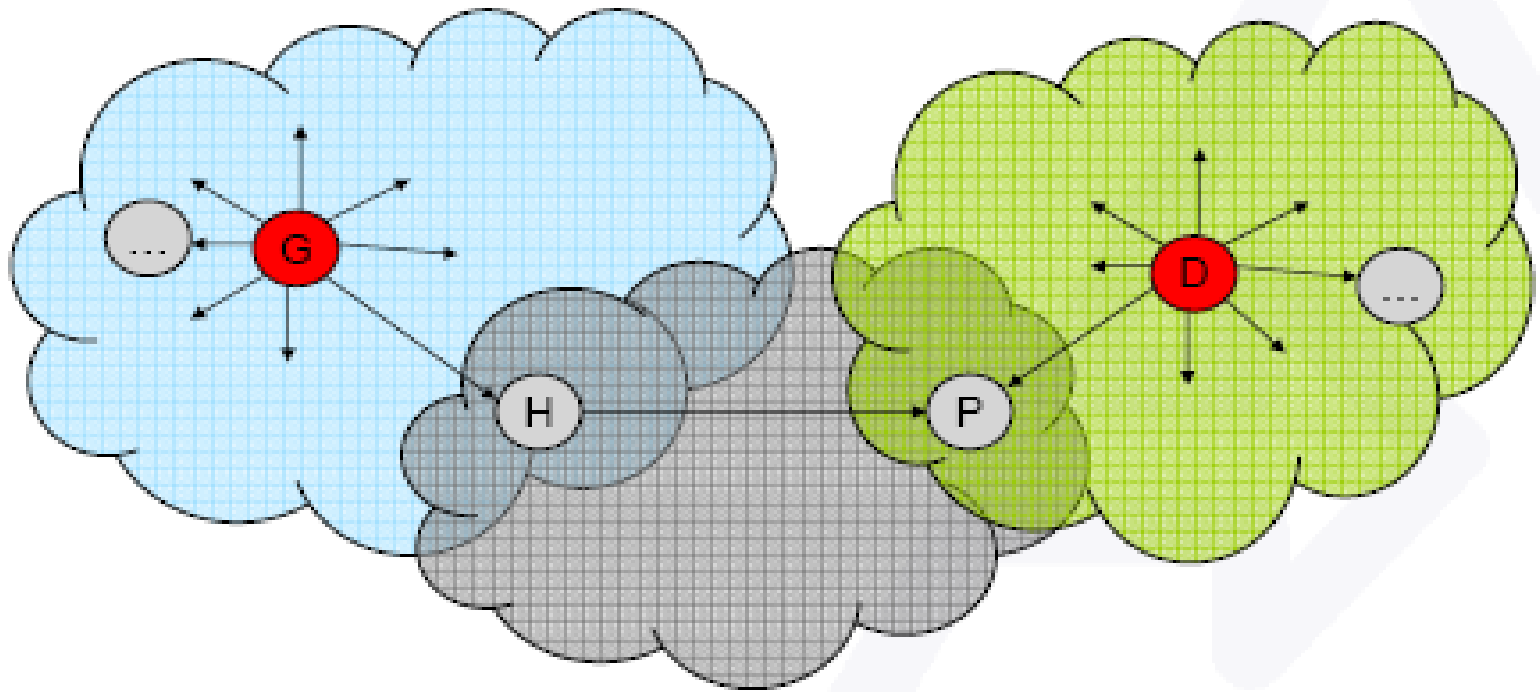
# Heterogeneous data sources

(BISON, M. Berthold, 2008)

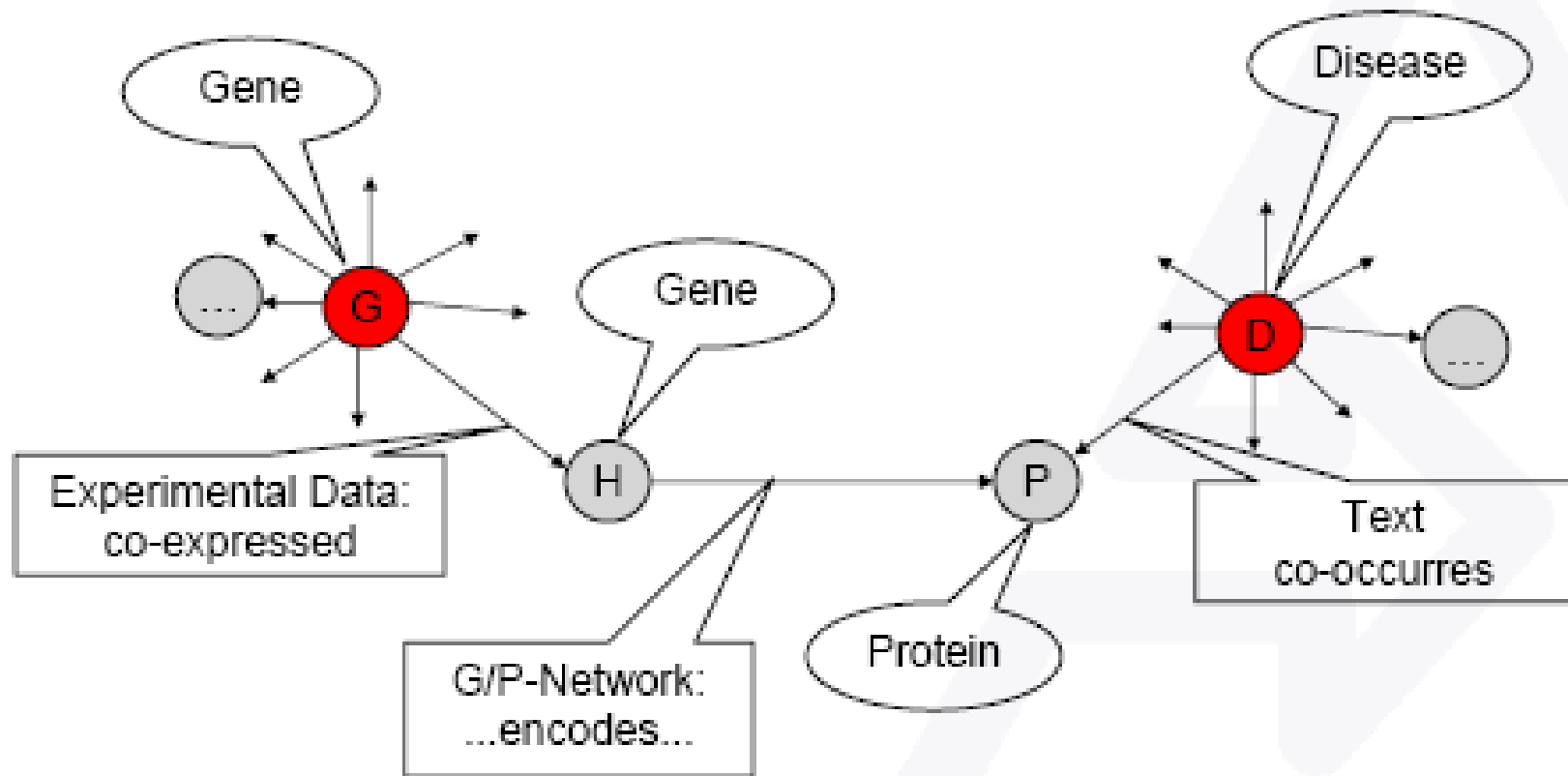


# Bridging concepts

(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)

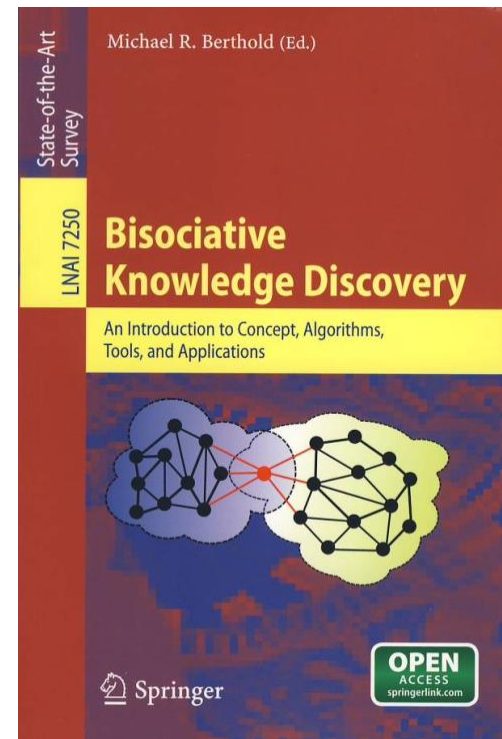


# Complementary BISON approach

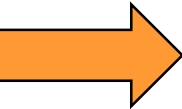
- Complementary approach: text mining
  - Find yet unknown terms in the intersection of documents, crossing different contexts (domains/literatures)
- Early related work: literature-based discovery (LBD)
  - Swanson (1988, 1990)
  - Smalheiser, Swanson (1998): ARROWSMITH
  - Weeber et al. (2001)
  -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

# The BISON project

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

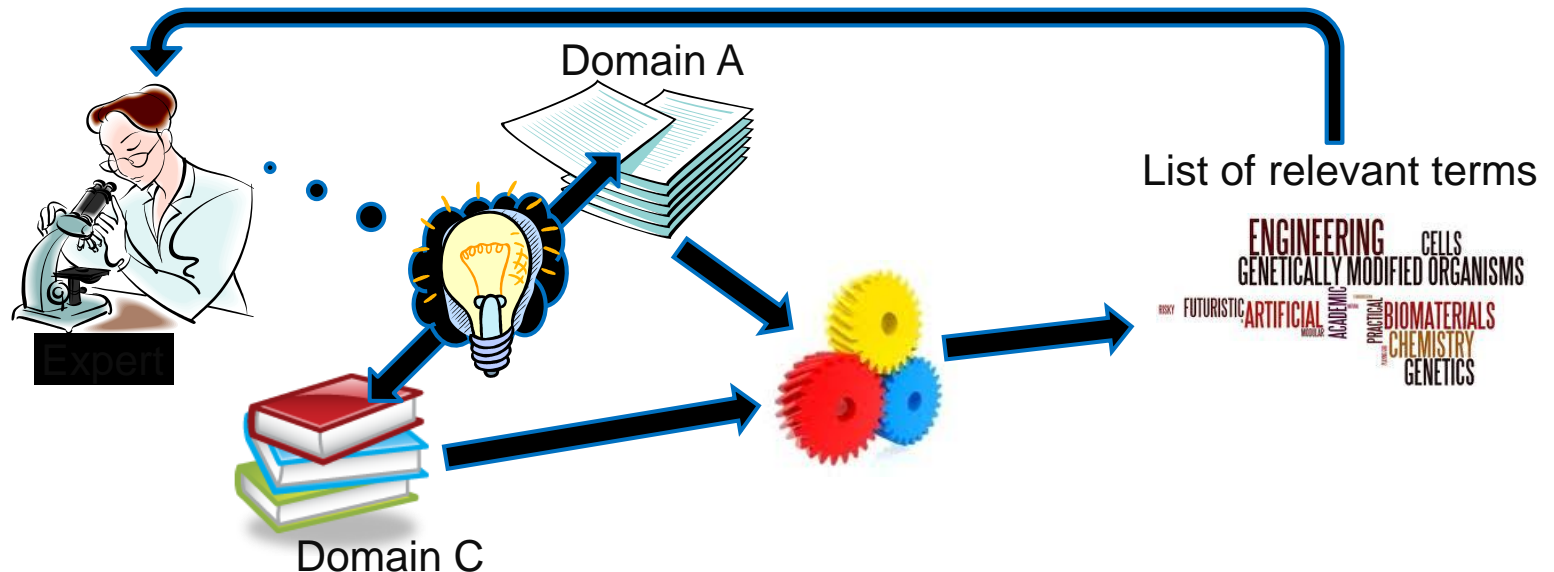


# Talk outline

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

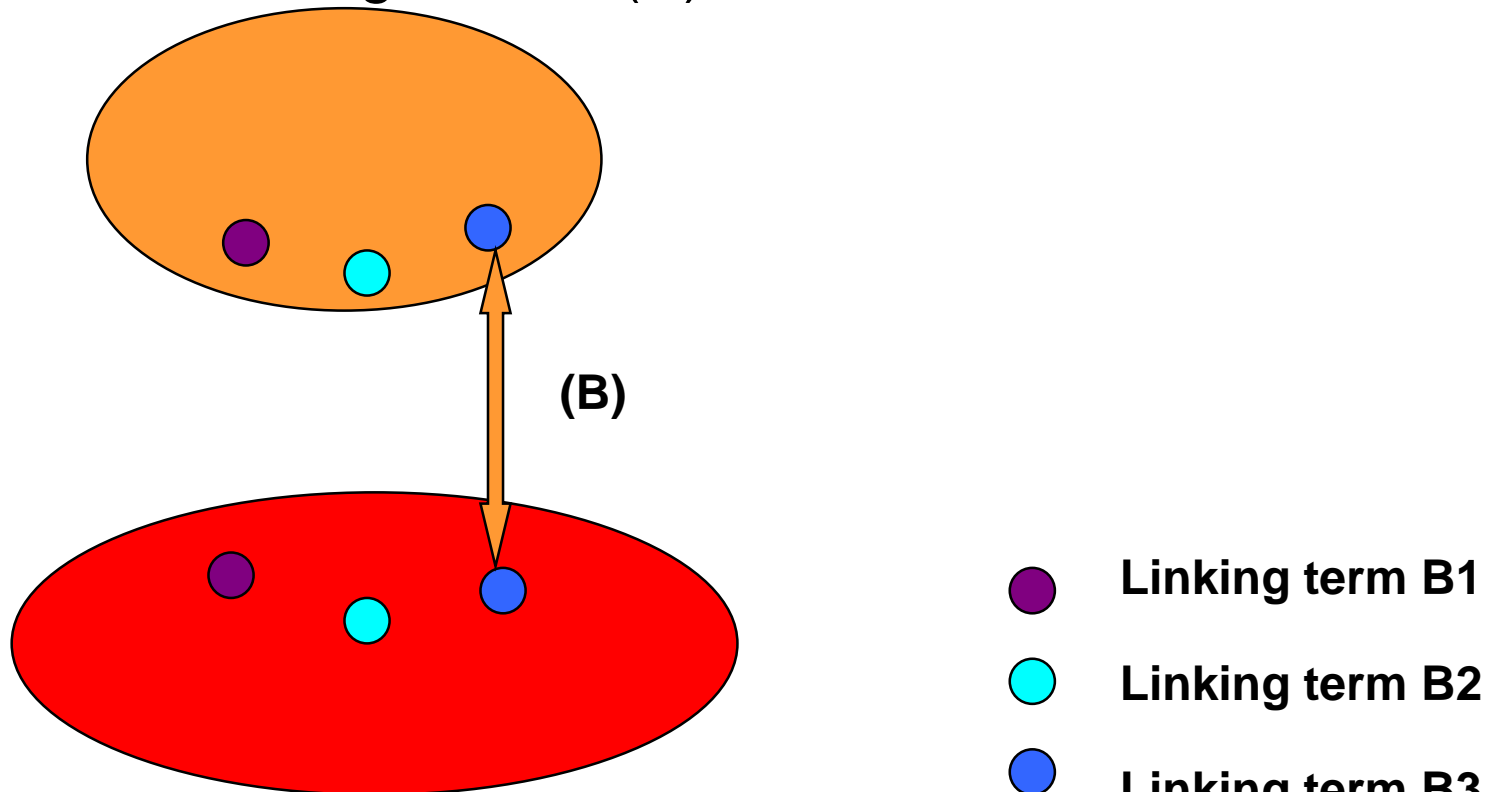
# Literature-based discovery

- 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



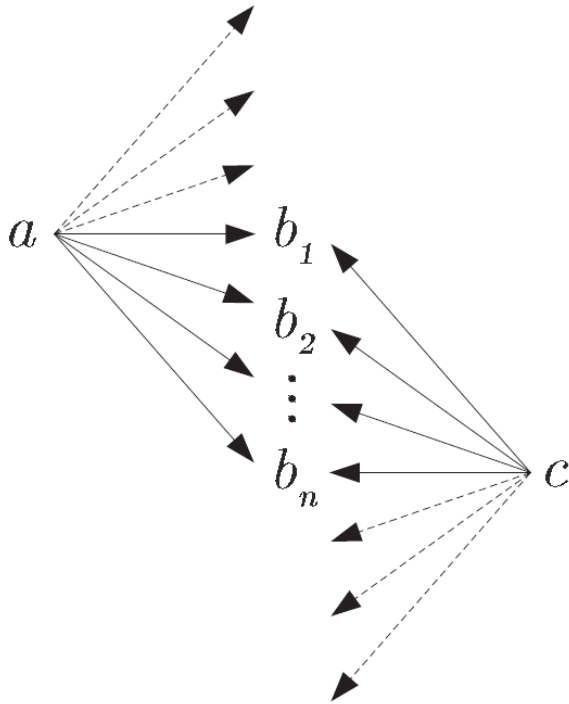
# Closed discovery setting: Finding linking (bridging) terms

Literature about magnesium (A): 38,000 articles



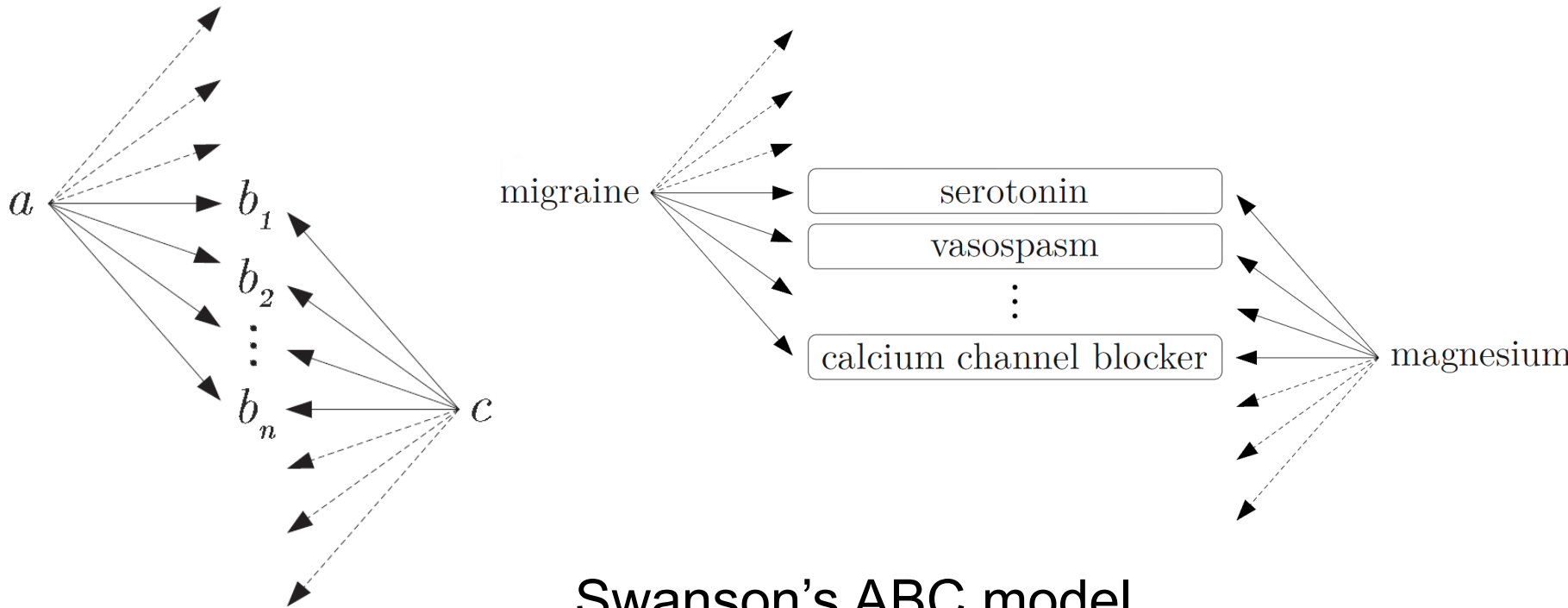
Literature about migraine (C): 4,600 articles

# 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:

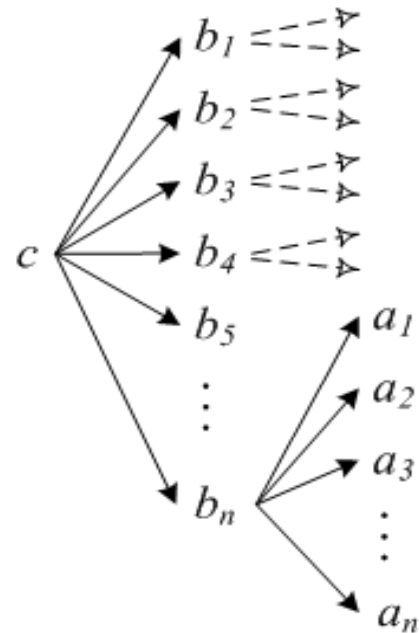
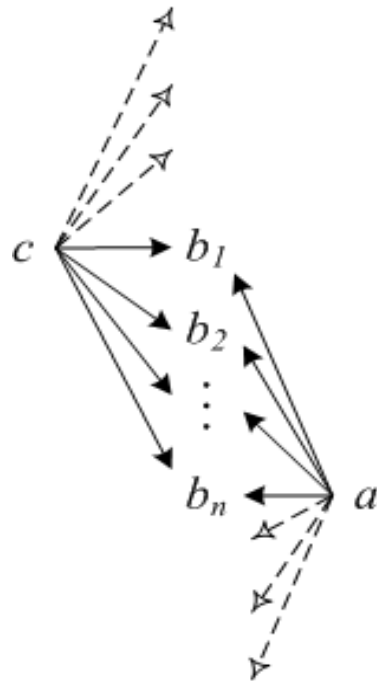
- Biomedical bibliographical database PubMed
- US National Library of Medicine
- More than 21M citations
- More than 5,600 journals
- 2,000 – 4,000 references added each working day!

The screenshot displays the PubMed website interface. At the top, the NCBI logo and 'PubMed' branding are visible, along with the text 'A service of the National Library of Medicine and the National Institutes of Health' and the URL 'www.pubmed.gov'. A search bar contains the query 'autism', with 'PubMed' selected as the database. The search results page shows a total of 11,008 items and 1,632 reviews. The first four results are listed:

- 1:** [Fazzi E, Rossi M, Signorini S, Rossi G, Bianchi PE, Lanzi G.](#) *Leber's congenital amaurosis: is there an autistic component?* *Dev Med Child Neurol.* 2007 Jul;49(7):503-7. PMID: 17593121 [PubMed - in process]
- 2:** [Paya B, Fuentes N.](#) *Neurobiology of autism: neuropathology and neuroimaging studies.* *Actas Esp Psiquiatr.* 2007 Jul-Aug;35(4):271-6. PMID: 17592791 [PubMed - in process]
- 3:** [Hayashi ML, Rao BS, Seo JS, Choi HS, Dolan BM, Choi SY, Chattarji S, Tonegawa S.](#) *Inhibition of p21-activated kinase rescues symptoms of fragile X syndrome in mice.* *Proc Natl Acad Sci U S A.* 2007 Jun 25; [Epub ahead of print] PMID: 17592139 [PubMed - as supplied by publisher]
- 4:** [Scheeren AM, Stauder JE.](#) *Broader Autism Phenotype in Parents of Autistic Children: Reality or Myth?* *J Autism Dev Disord.* 2007 Jun 23; [Epub ahead of print] PMID: 17588199 [PubMed - as supplied by publisher]

# 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
- TextFlows text mining platform
- Summary and conclusions

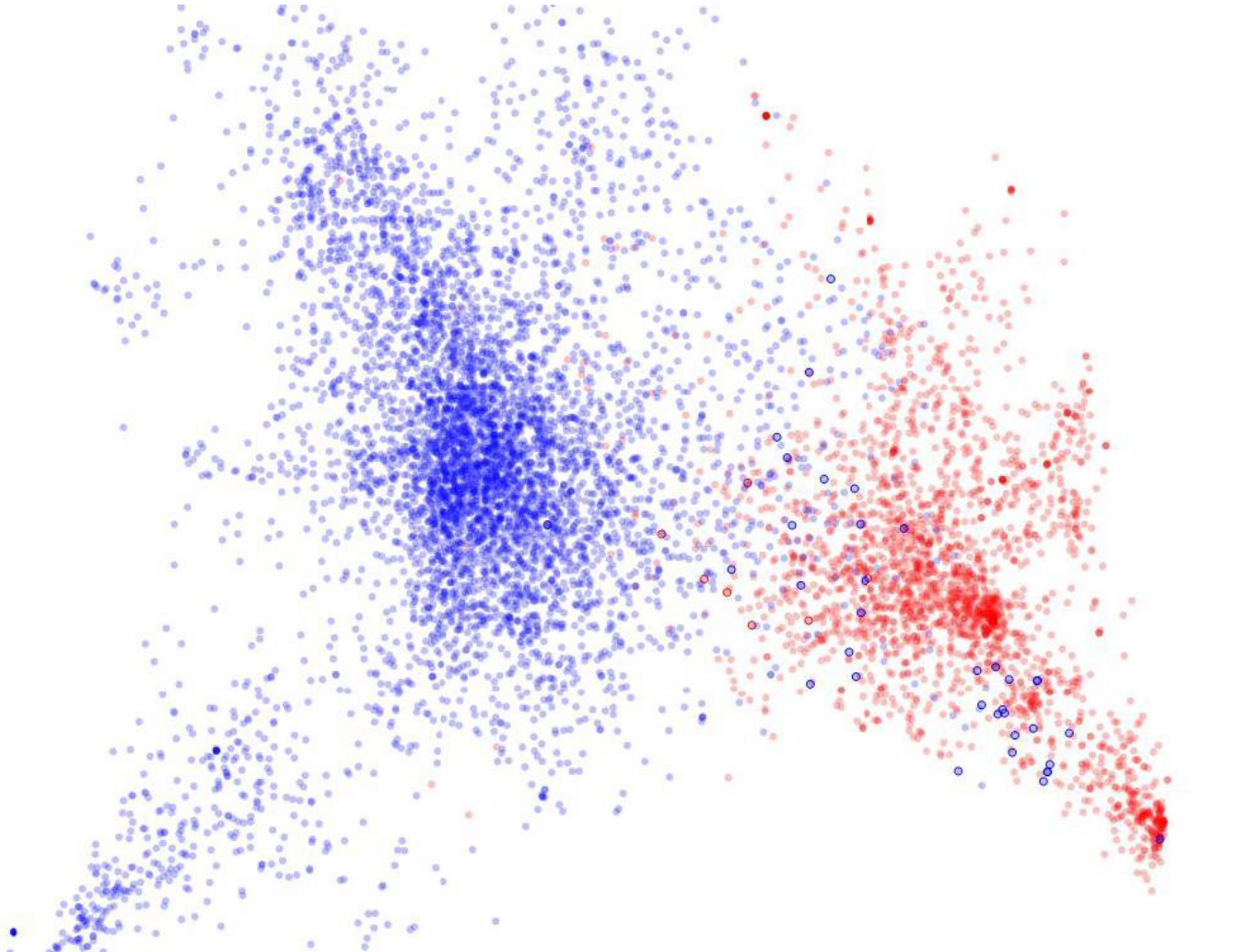
# Outlier detection



# 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



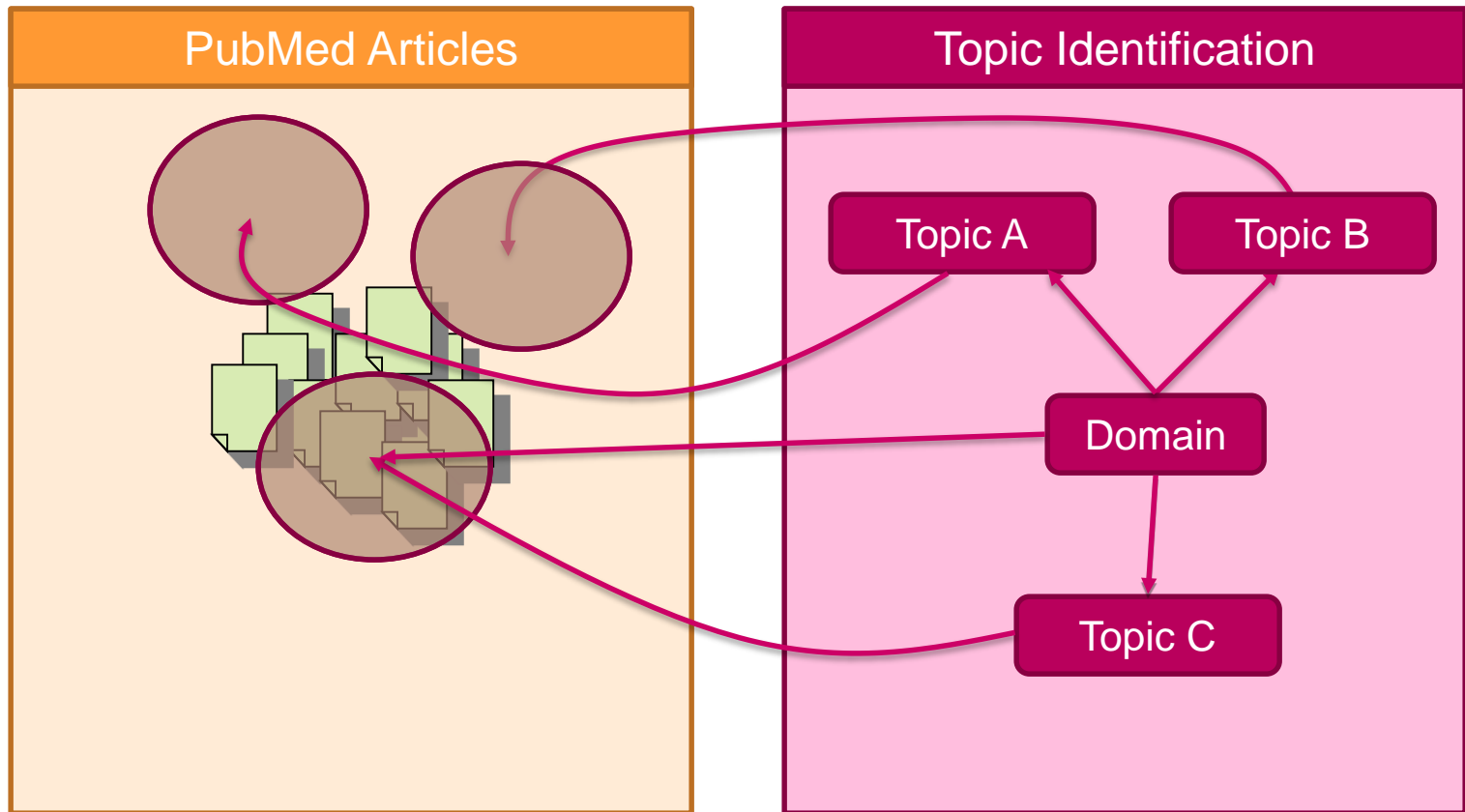
*2-dimensional projection of documents (about autism (red) and calcineurin (blue)). Outlier documents are bolded for the user to easily spot them.*

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

(Lavrač, Sluban, Grčar, Juršič 2010)

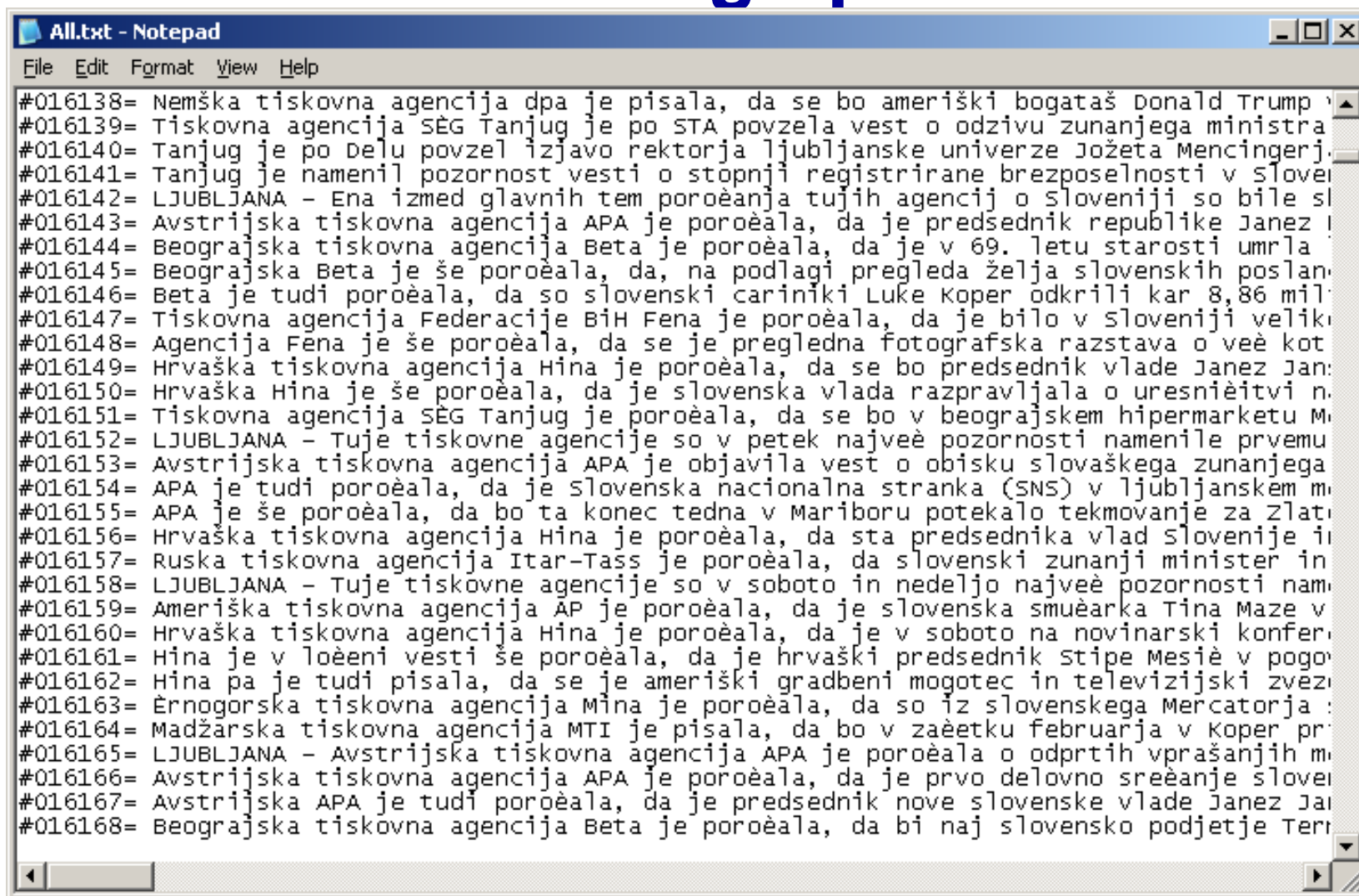


# Outlier detection by clustering of PubMed articles



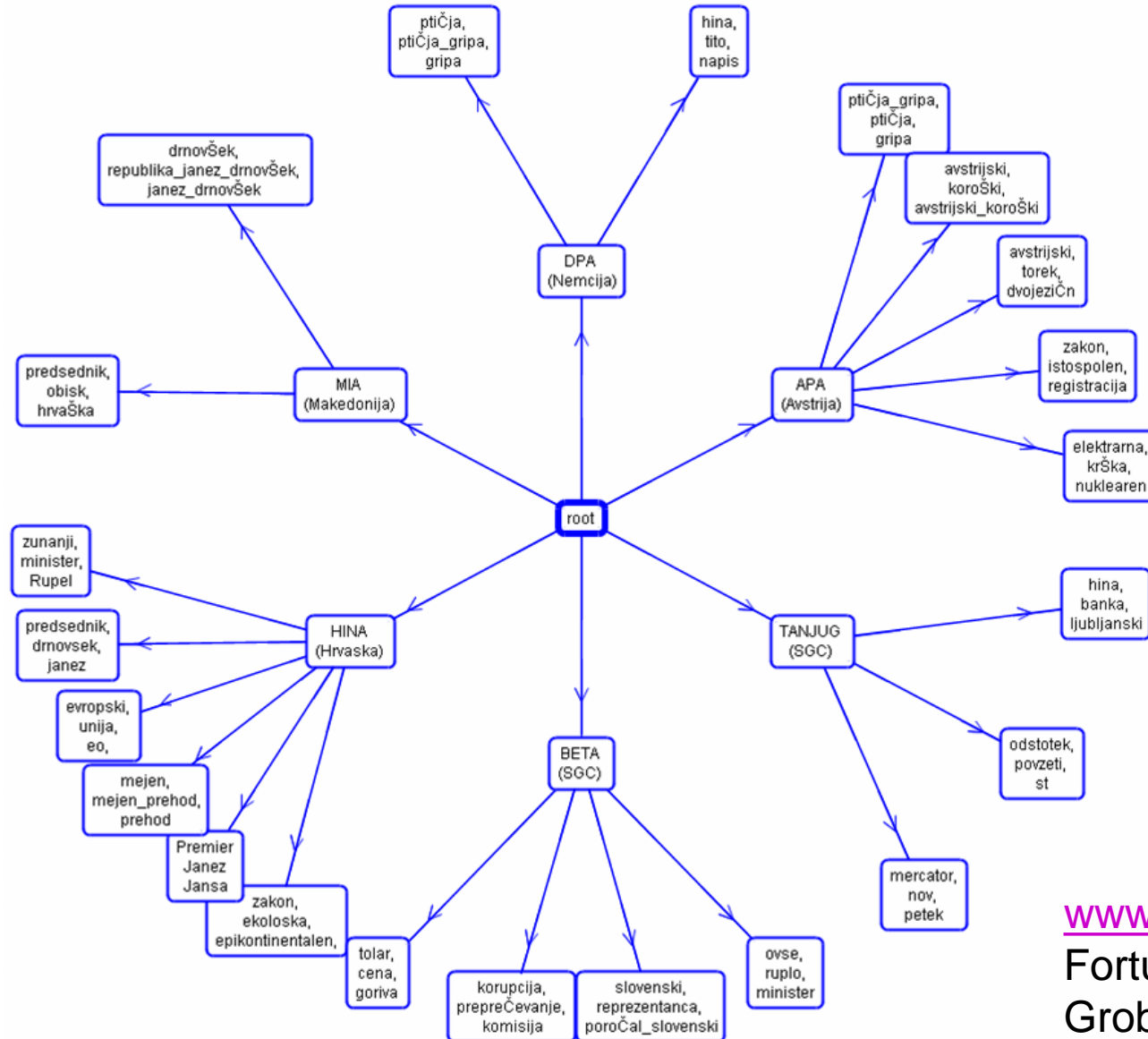
Slide adapted from D. Mladenić, JSI

# Illustrative example: Input to OntoGen clustering - STA news about Slovenia from foreign press



```
All.txt - Notepad
File Edit Format View Help
#016138= Nemška tiskovna agencija dpa je pisala, da se bo ameriški bogataš Donald Trump
#016139= Tiskovna agencija S&E Tanjug je po STA povzela vest o odzivu zunanjega ministra
#016140= Tanjug je po Delu povzel izjavo rektorja ljubljanske univerze Jožeta Mencingerj.
#016141= Tanjug je namenil pozornost vesti o stopnji registrirane brezposelnosti v Sloveni
#016142= LJUBLJANA - Ena izmed glavnih tem poro&anja tujih agencij o Sloveniji so bile sl
#016143= Avstrijska tiskovna agencija APA je poro&ala, da je predsednik republike Janez I
#016144= Beograjska tiskovna agencija Beta je poro&ala, da je v 69. letu starosti umrla
#016145= Beograjska Beta je še poro&ala, da, na podlagi pregleda želja slovenskih poslan
#016146= Beta je tudi poro&ala, da so slovenski cariniki Luke Koper odkrili kar 8,86 mil
#016147= Tiskovna agencija Federacije BiH Fena je poro&ala, da je bilo v Sloveniji veliki
#016148= Agencija Fena je še poro&ala, da se je pregledna fotografska razstava o ve& kot
#016149= Hrvaška tiskovna agencija Hina je poro&ala, da se bo predsednik vlade Janez Jan
#016150= Hrvaška Hina je še poro&ala, da je slovenska vlada razpravljala o uresni&itvi n.
#016151= Tiskovna agencija S&E Tanjug je poro&ala, da se bo v beograjskem hipermarketu M
#016152= LJUBLJANA - Tuje tiskovne agencije so v petek najve& pozornosti namenile prvemu
#016153= Avstrijska tiskovna agencija APA je objavila vest o obisku slovaškega zunanjega
#016154= APA je tudi poro&ala, da je slovenska nacionalna stranka (SNS) v ljubljanskem m
#016155= APA je še poro&ala, da bo ta konec tedna v Mariboru potekalo tekmovanje za Zlati
#016156= Hrvaška tiskovna agencija Hina je poro&ala, da sta predsednika vlad Slovenije in
#016157= Ruska tiskovna agencija Itar-Tass je poro&ala, da slovenski zunanji minister in
#016158= LJUBLJANA - Tuje tiskovne agencije so v soboto in nedeljo najve& pozornosti nam
#016159= Ameriška tiskovna agencija AP je poro&ala, da je slovenska smu&arka Tina Maze v
#016160= Hrvaška tiskovna agencija Hina je poro&ala, da je v soboto na novinarski konfer
#016161= Hina je v lo&eni vesti še poro&ala, da je hrvaški predsednik stipe Mesi& v pogo
#016162= Hina pa je tudi pisala, da se je ameriški gradbeni mogotec in televizijski zvez
#016163= &rnogorska tiskovna agencija Mina je poro&ala, da so iz slovenskega Mercatorja :
#016164= Madžarska tiskovna agencija MTI je pisala, da bo v za&etku februarja v Koper pr
#016165= LJUBLJANA - Avstrijska tiskovna agencija APA je poro&ala o odprtih vpra&anjih m
#016166= Avstrijska tiskovna agencija APA je poro&ala, da je prvo delovno sre&anje slove
#016167= Avstrijska APA je tudi poro&ala, da je predsednik nove slovenske vlade Janez Jan
#016168= Beograjska tiskovna agencija Beta je poro&ala, da bi naj slovensko podjetje Terr
```

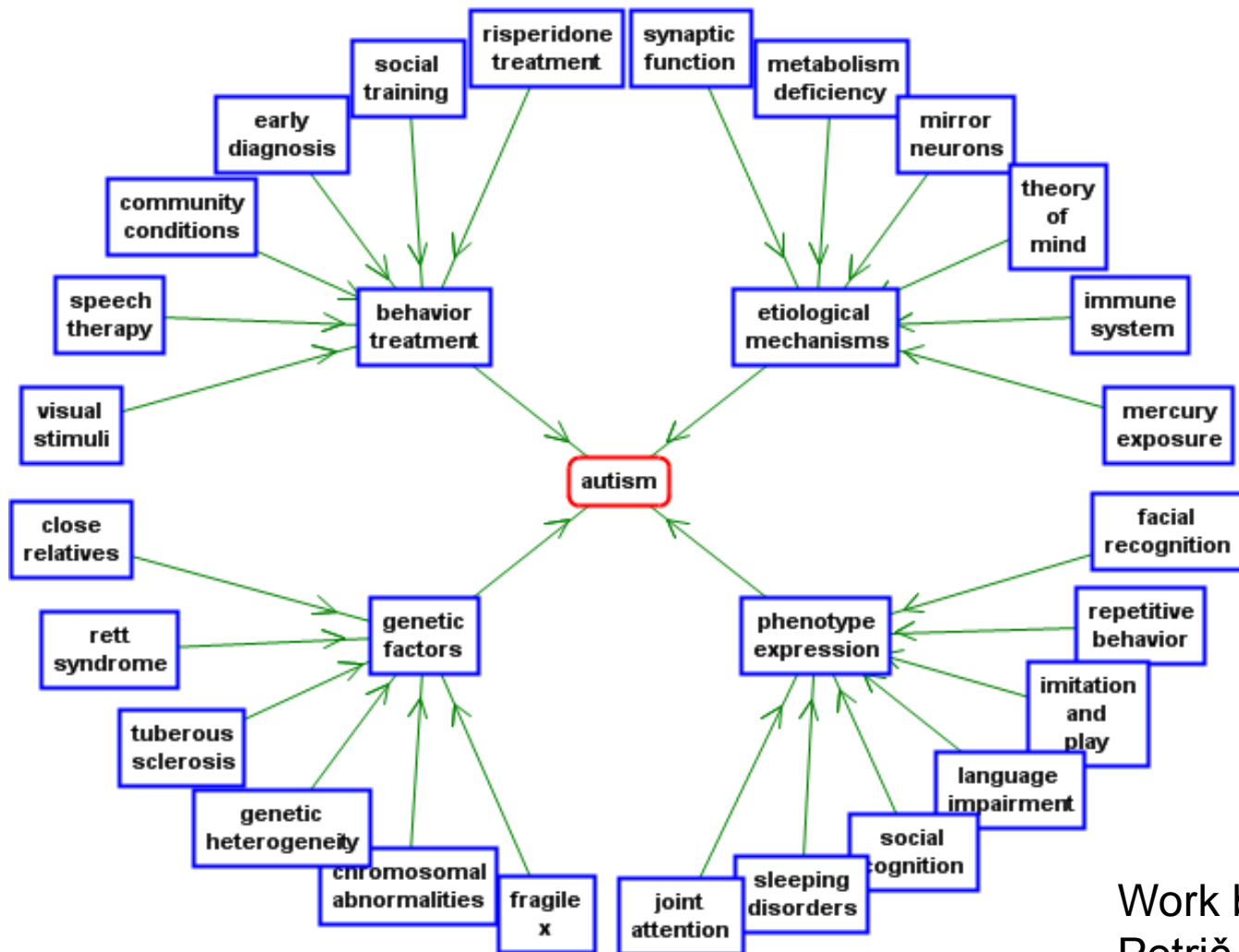
# OntoGen clustering for STA news analysis



[www.ontogen.si](http://www.ontogen.si)

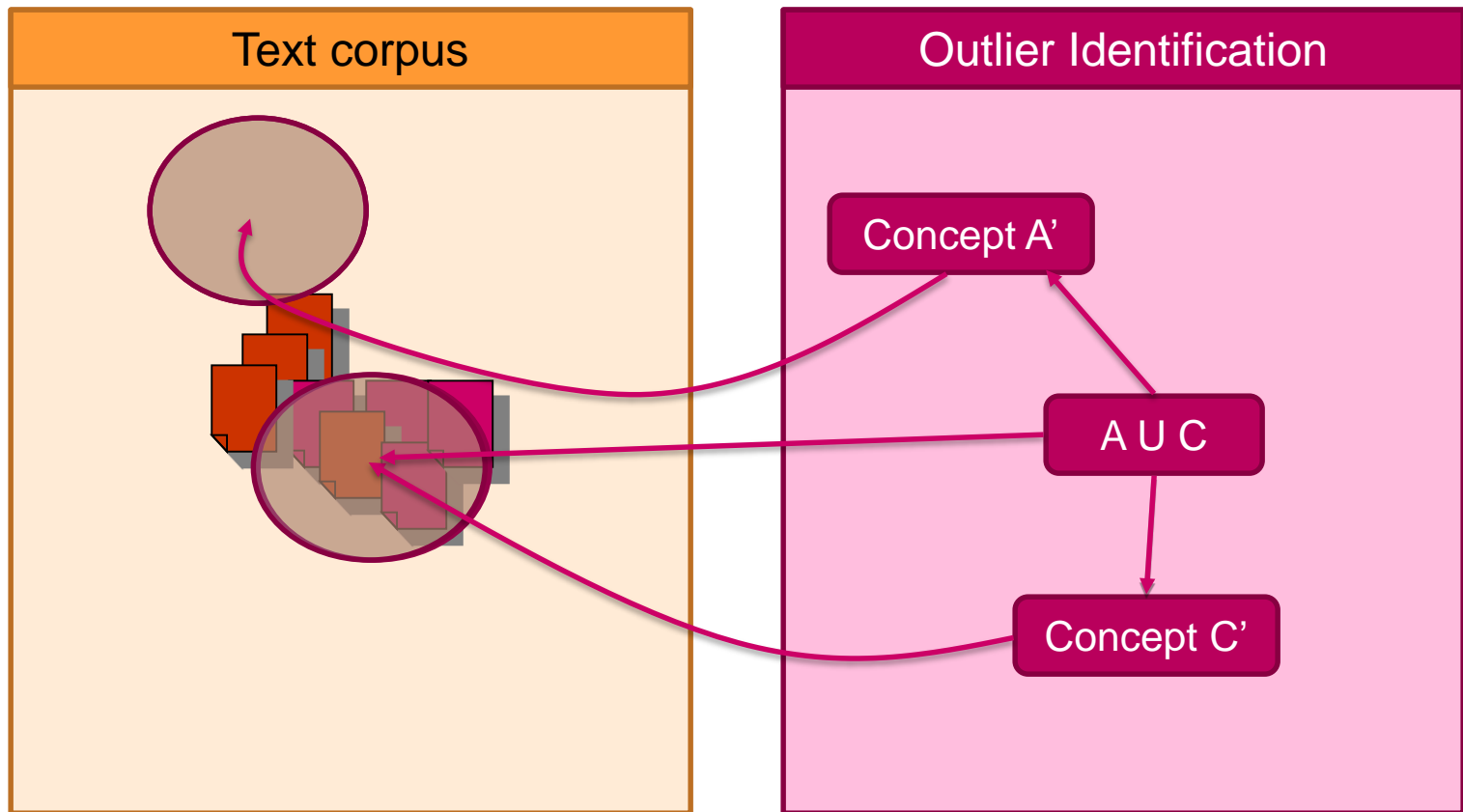
Fortuna, Mladenić,  
Grobelnik 2006

# Using OntoGen for clustering PubMed articles on autism



Work by  
Petrič et al. 2009

# Using OntoGen for outlier document identification



Slide adapted from D. Mladenić, JSI

# Results on autism-calcineurin: Outlier calcineurin document CN423

The screenshot displays the OntoGen software interface, titled "OntoGen -- Text Garden". The interface is divided into several panels:

- Concepts:** A tree view showing a hierarchy starting with "root", followed by "A' autism", and then "C' calcineurin".
- Concept properties:** A section with tabs for "Details", "Suggestions", and "Relations". The "Details" tab is active, showing the name "A' autism" and a list of keywords: "children, autism, patient, autistic, disorders, group, behaviors, asd, social, transplantation".
- 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 with their similarity scores. Document CN423 is highlighted as an outlier.
- Document preview:** A text box showing the content of document CN423, which discusses calcineurin's role in synaptic plasticity and neuronal adaptation.
- Document name:** A field at the bottom of the interface.

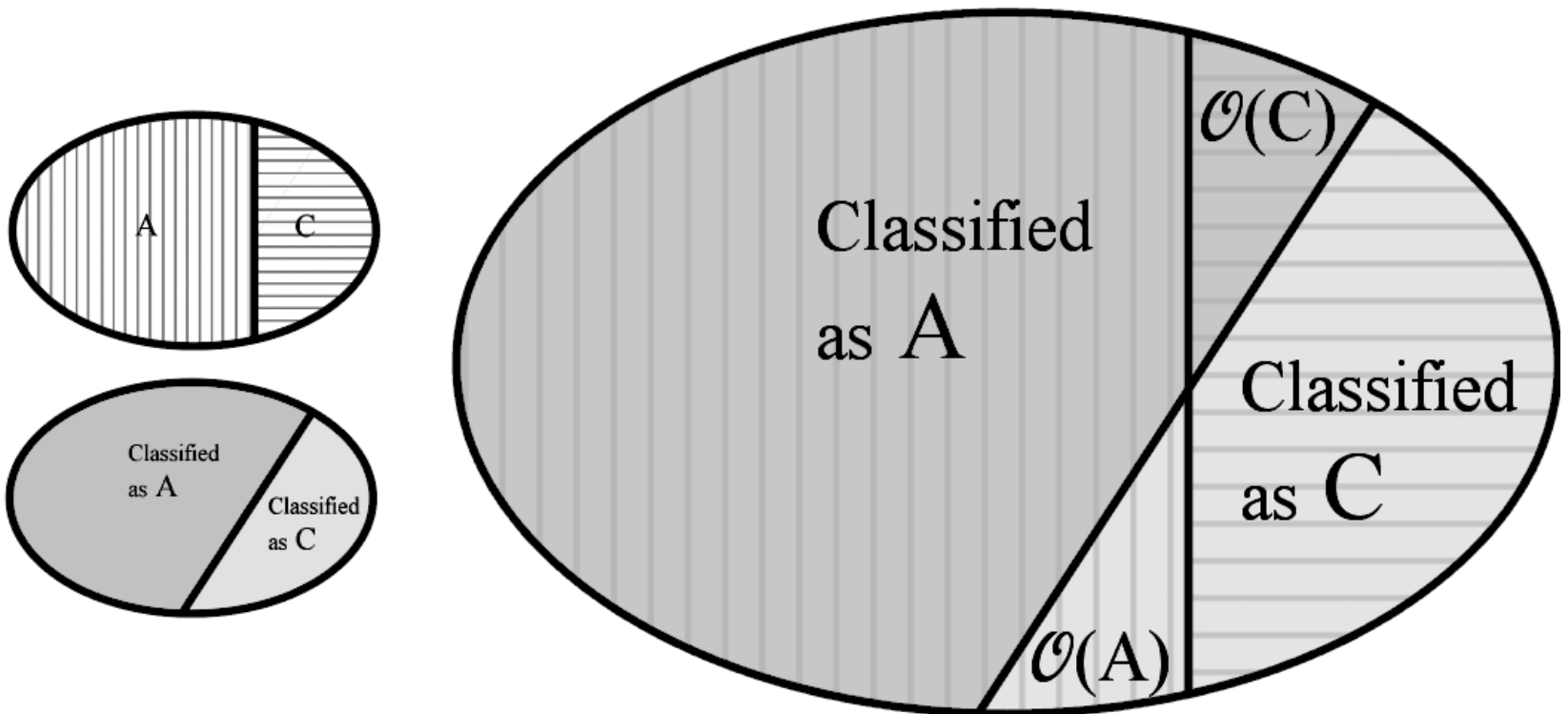
Document	Similarity
<input checked="" type="checkbox"/> 3874 -- This meta-analysis of 12 dependent...	0,146
<input checked="" type="checkbox"/> 8939 -- Administered the Stanford-Binet an...	0,146
<input checked="" type="checkbox"/> CN1065 -- Sirolimus-associated interstitial p...	0,146
<input checked="" type="checkbox"/> 6372 -- The last 40 years has seen a virtual...	0,146
<input checked="" type="checkbox"/> 2402 -- Early experiences affect brain funct...	0,146
<input type="checkbox"/> CN3661 -- Allograft rejection is a leading ca...	0,146
<input checked="" type="checkbox"/> 220 -- Kraepelin's dichotomy, manic-depres...	0,146
<input checked="" type="checkbox"/> 7163 -- A neurochemical assessment of nor...	0,146
<input checked="" type="checkbox"/> 6864 -- This paper reports findings from an ...	0,146
<input checked="" type="checkbox"/> 7686 -- A group of high-functioning autistic ...	0,146
<input checked="" type="checkbox"/> CN3207 -- Recent advances in immunosup...	0,146
<input type="checkbox"/> CN423 -- Calcineurin is a neuron-enriched ...	0,146
<input checked="" type="checkbox"/> 5168 -- Conventional antipsychotic medicat...	0,146
<input checked="" type="checkbox"/> CN2549 -- Steroids have accompanied oth...	0,146
<input checked="" type="checkbox"/> 4072 -- Autism is a complex genetic neurod...	0,146

Document name:

Work by  
Petrič et al. 2010

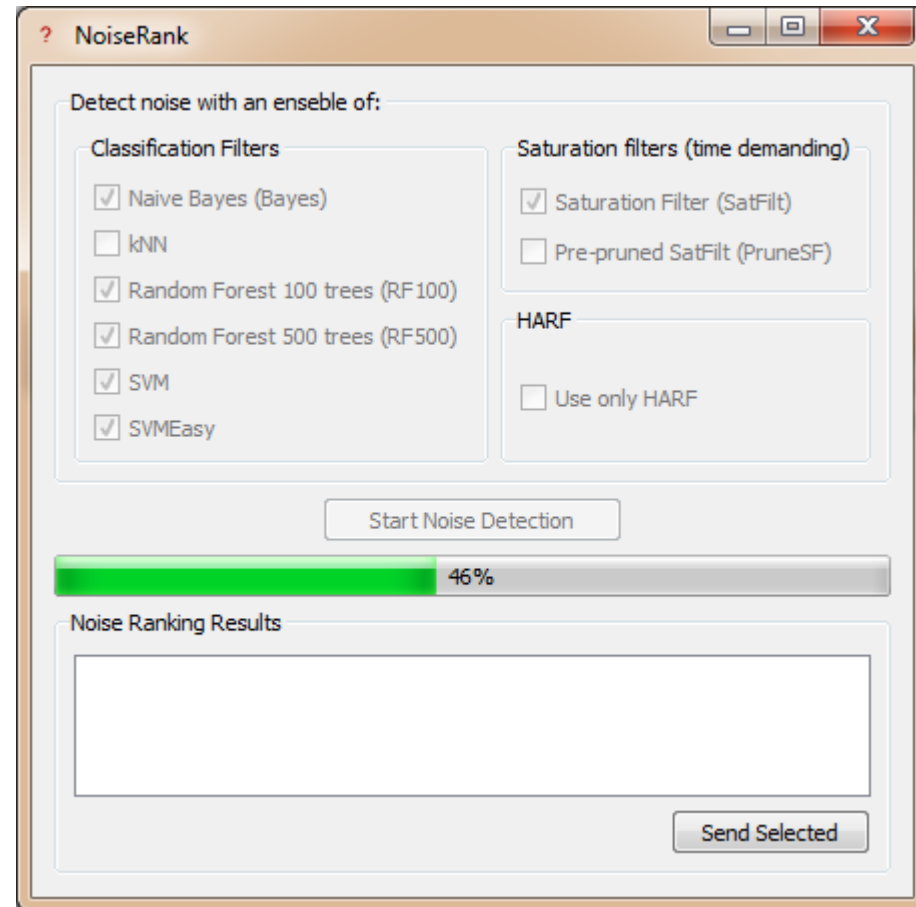
# Detecting outlier documents

- 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
- TextFlows text mining platform
- Summary and conclusions

# CrossBee: Cross Context Bisociation Explorer



Supported by   

Start Downloads Term View Document View BTerms

SEARCH

MAIN MENU

- Start
- Downloads
- Term View
- Document View
- BTerms
- Display Settings

ITEM BASKET

Empty - drag items (terms, documents or views to this basket to save them)

## 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 child...

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):

- paroxysmal** (0,999)
- migraine** (0,855)
- feature** (0,564)
- electroencephalogram migraine (0,053)
- electroencephalogram (0,029)

Document's Important Terms (ordered by alphabet):

- electroencephalogram (0,029)
- electroencephalogram migraine (0,053)
- feature** (0,564)
- migraine** (0,855)
- 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):

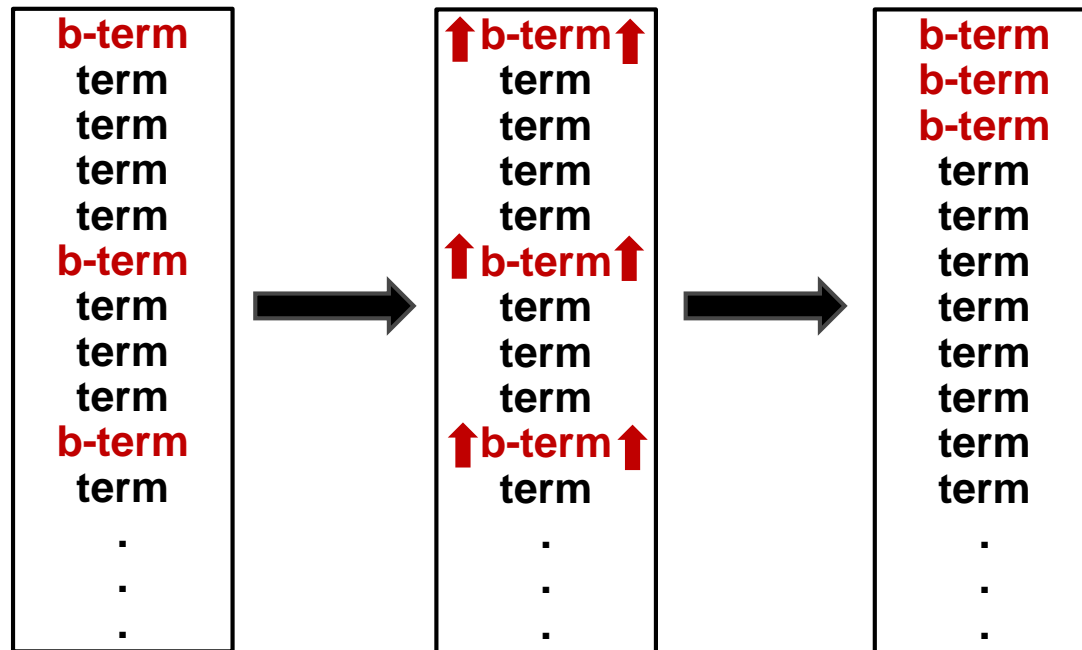
- paroxysmal** (0,999)
- case** (0,855)
- treatment** (0,712)
- type** (0,711)
- etiology** (0,711)
- magnesium** (0,568)
- role (0,424)
- tachycardia (0,421)
- etiology treatment (0,277)
- de (0,086)
- role magnesium (0,077)

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

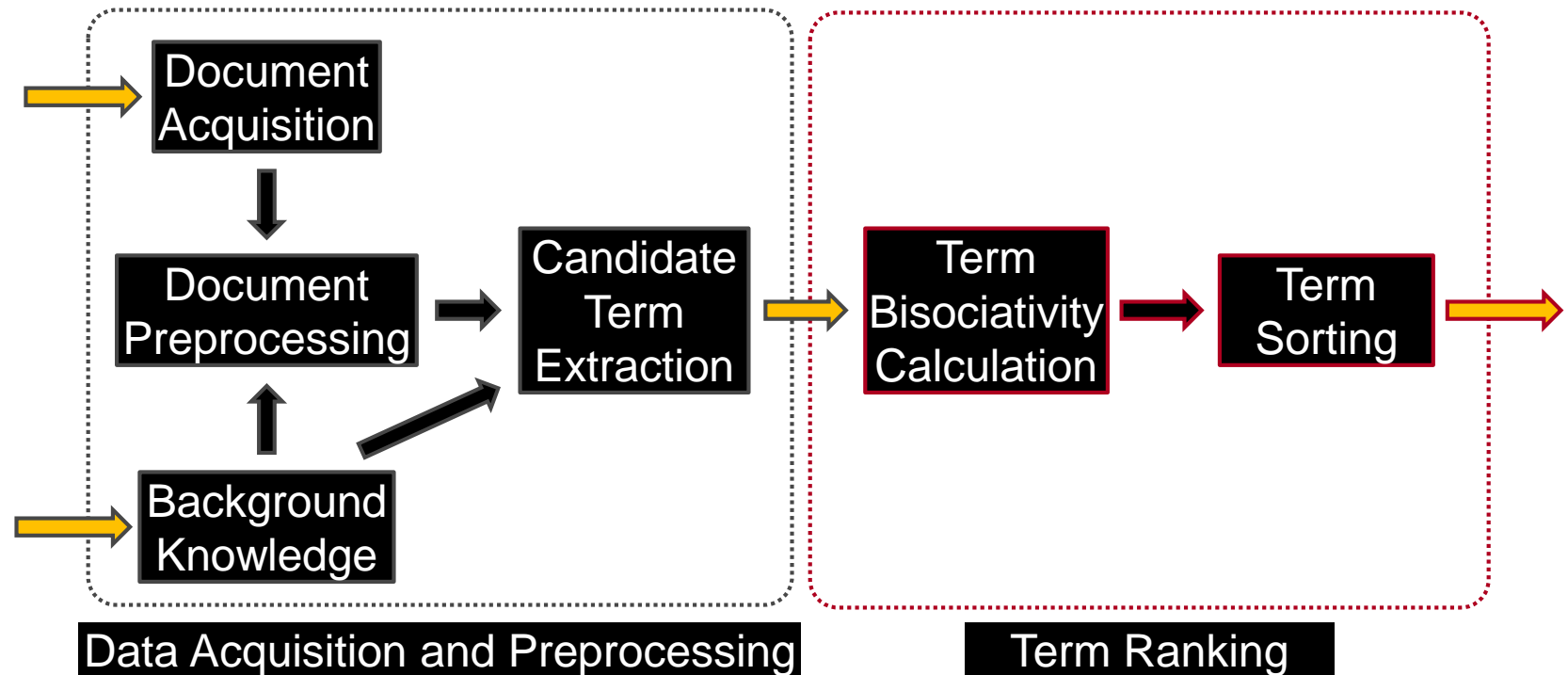
CrossBee: Application version: 3.0, built on: 17.1.2012  
In synch with the results published in the Bison book.  
Copyright © 2010 Jozef Stefan Institute. Style designed by Free CSS Templates. SiteMap.

# 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*)



# CrossBee: Methodology overview

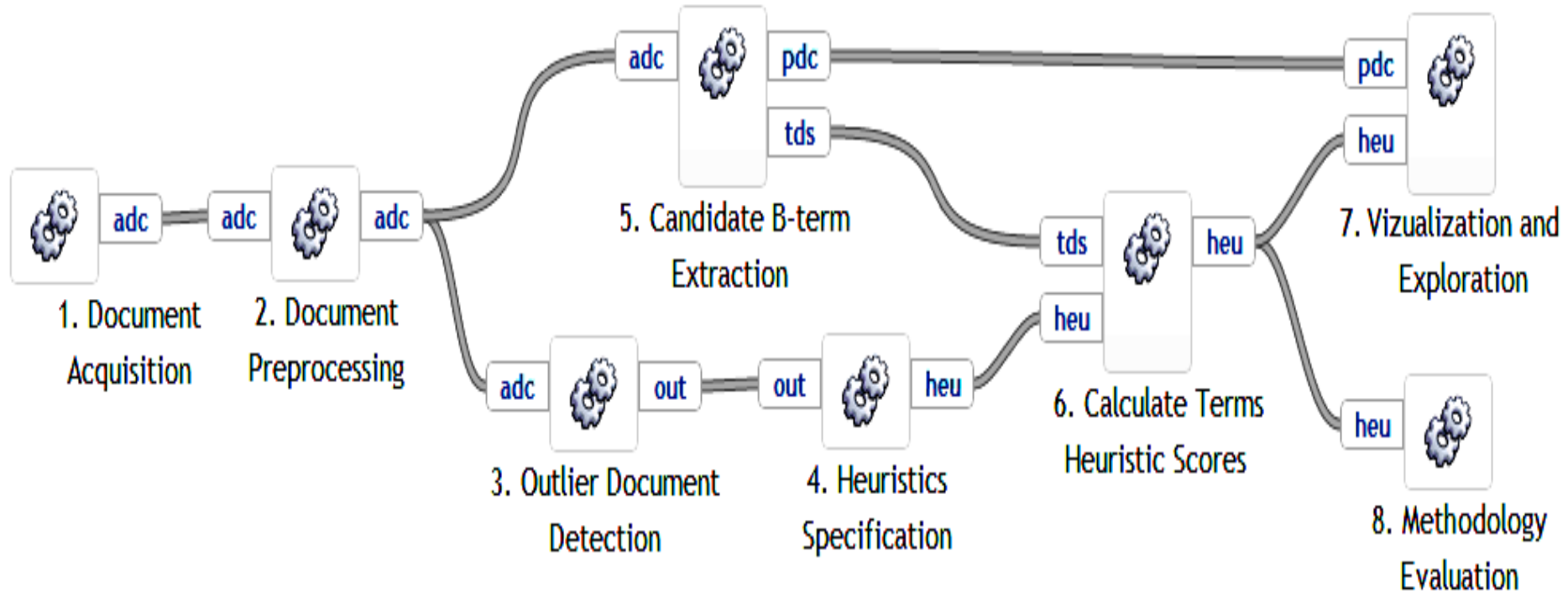


Incorporating available background knowledge

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



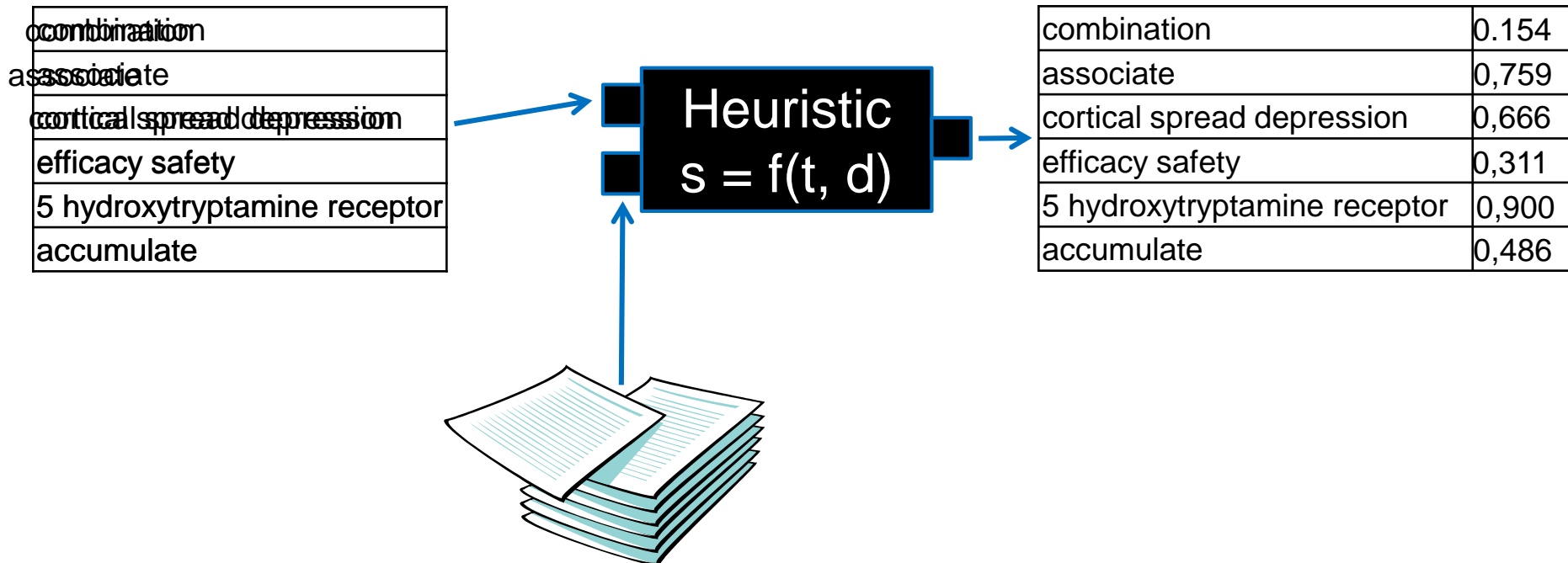
# Term ranking

- 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

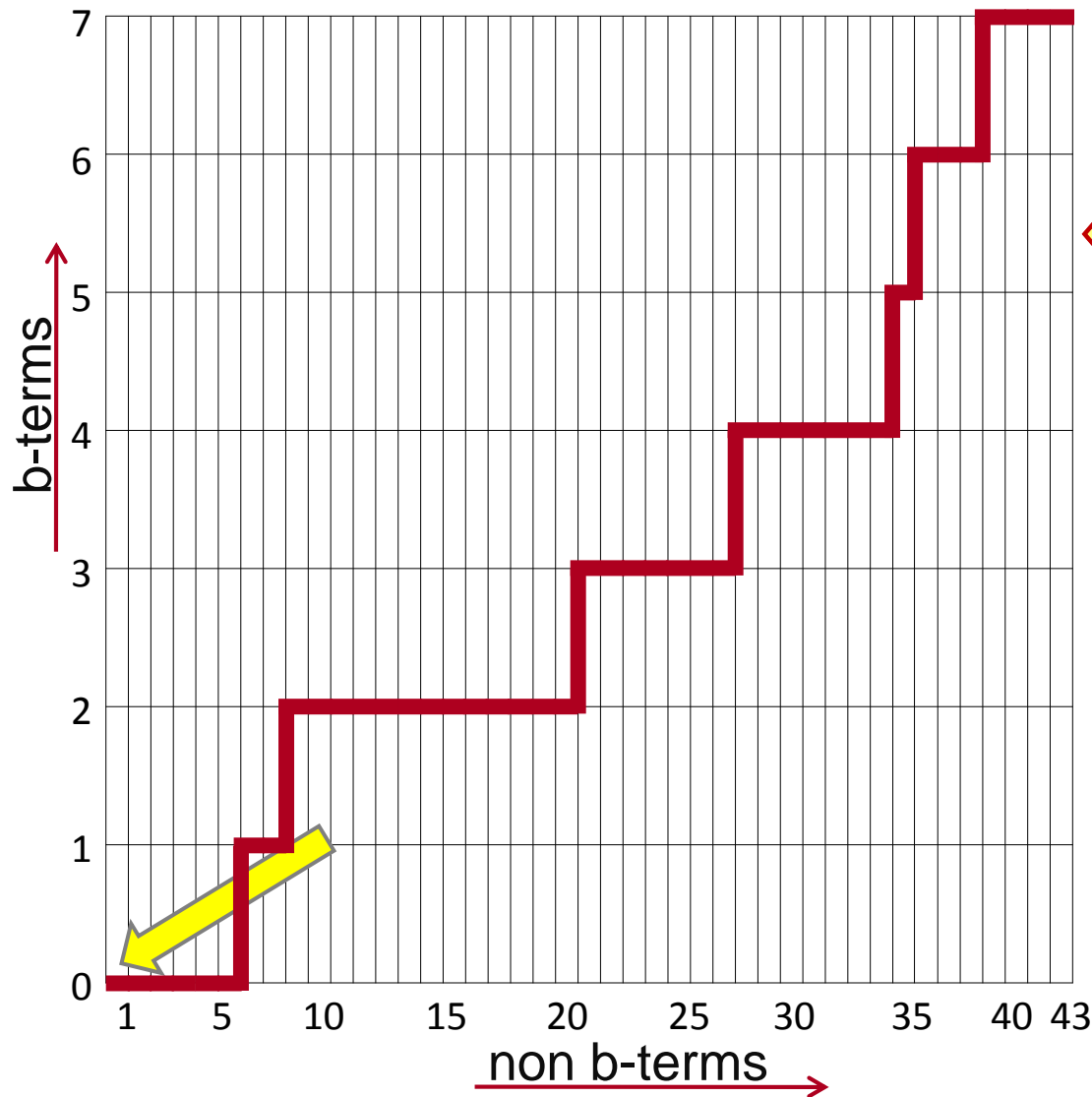
## final ensemble heuristic

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

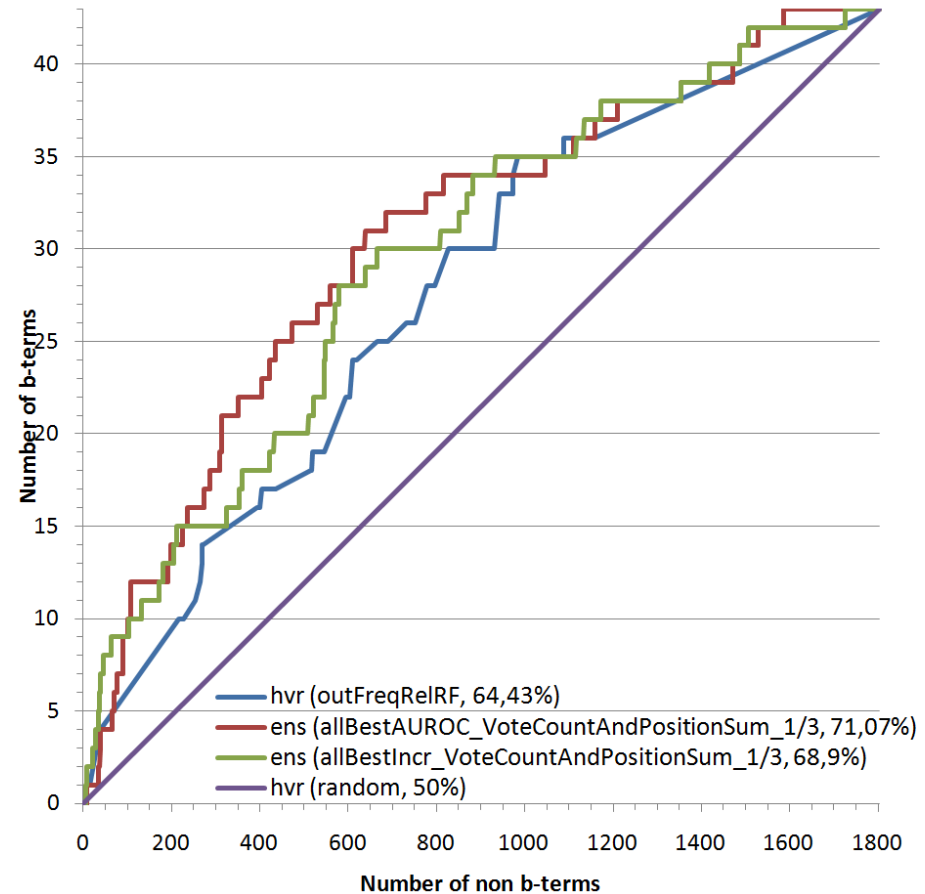
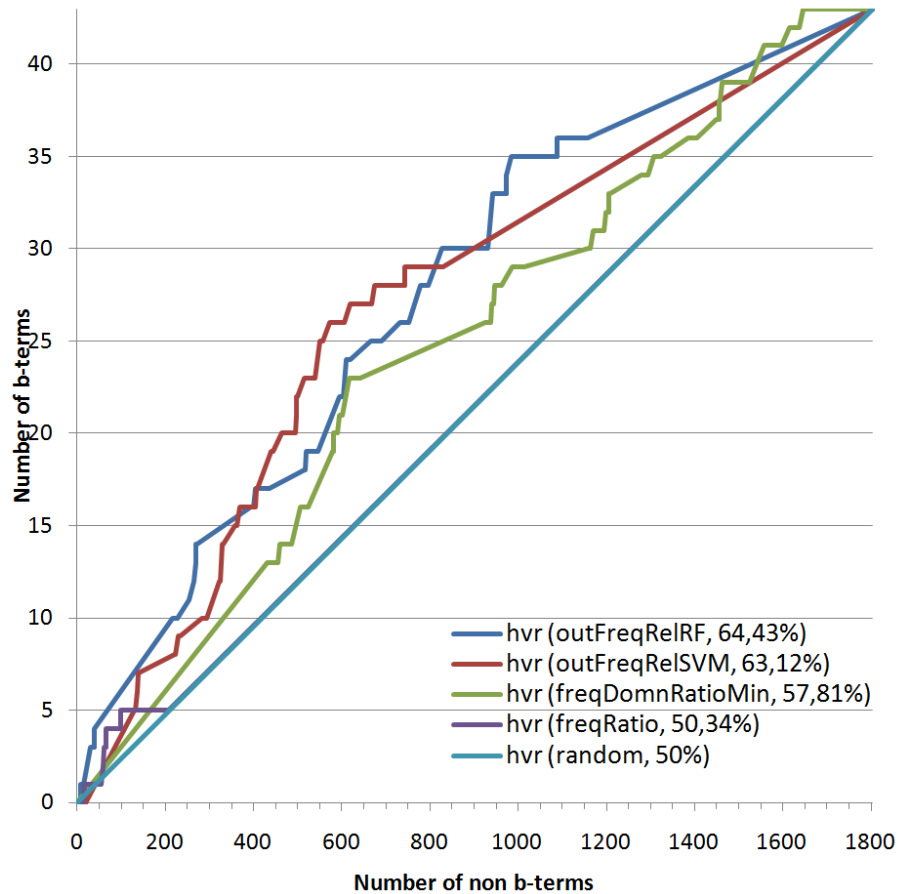


Ranked term list:  
 50 terms = 7 b-terms +  
 43 non b-terms

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

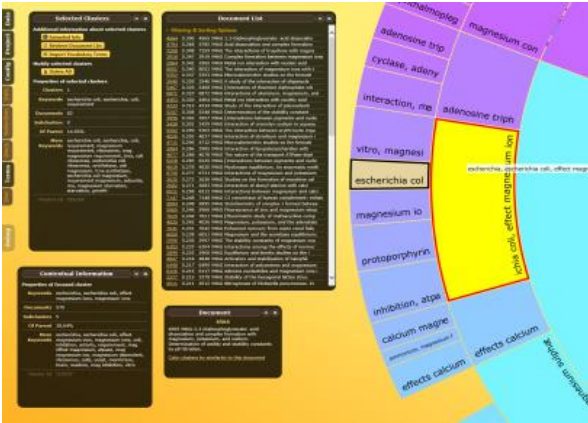
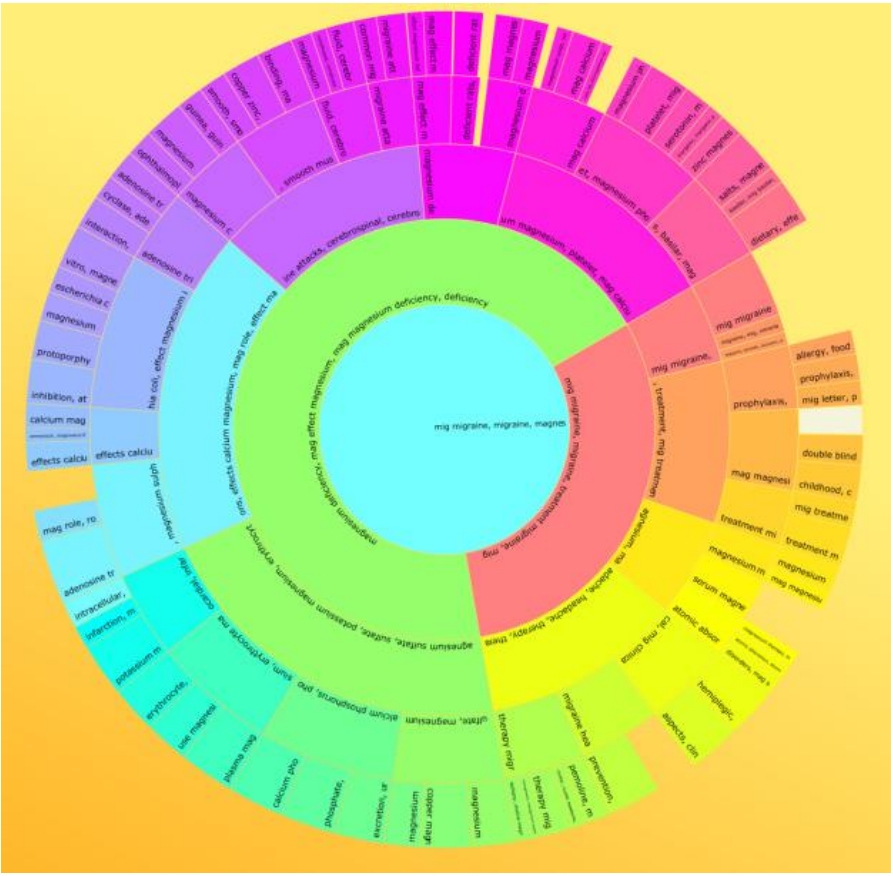


# CrossBee system

- 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)

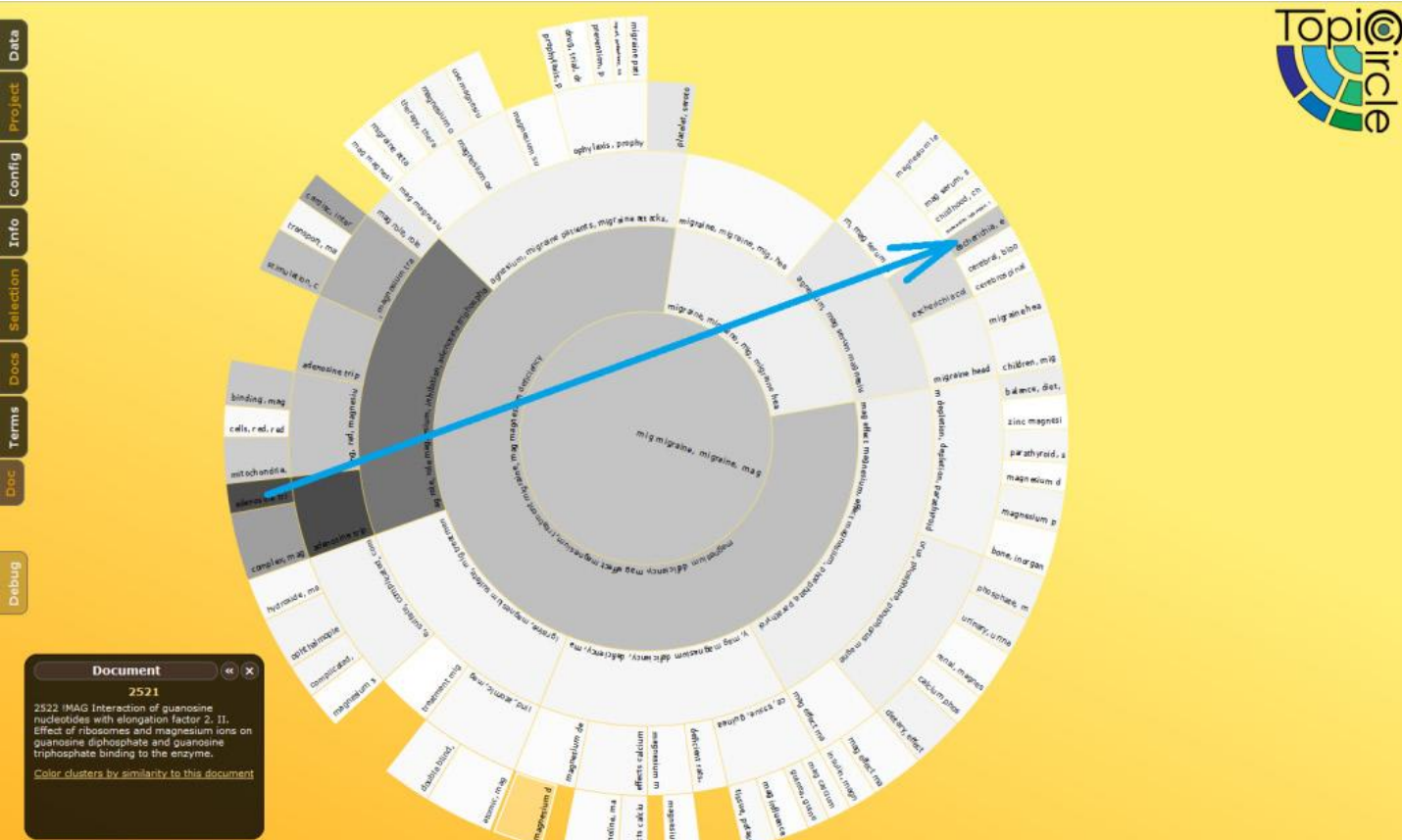
# Additional CrossBee functionality

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.

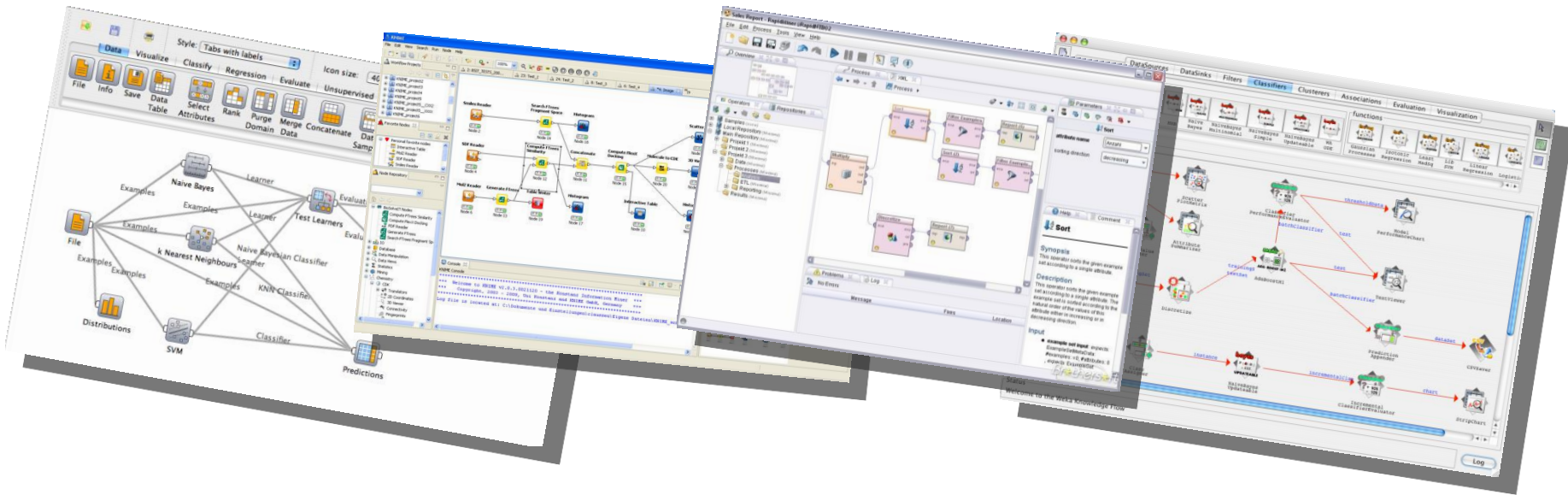


# Talk outline

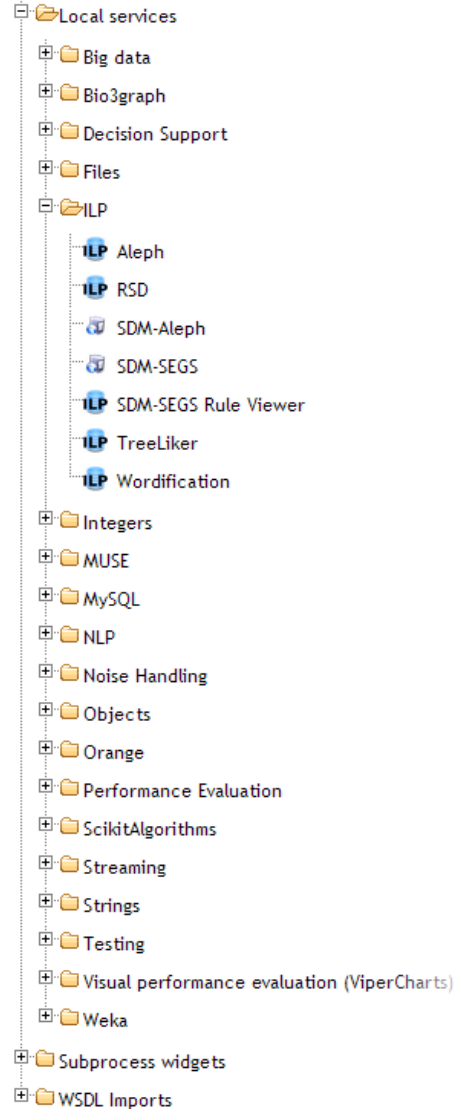
- Background and motivation
- Literature-based discovery
- Cross-domain literature mining approaches
  - Outlier detection for cross-domain knowledge discovery
  - Cross-domain knowledge discovery with CrossBee
- TextFlows text mining platform
- Summary and conclusions

# Data mining platforms

WEKA, KNIME, RapidMiner, Orange (FRI), Orange4WS (IJS)



- Incorporate numerous data mining algorithms
- Enable data analysis and visualization
- Enable workflow construction

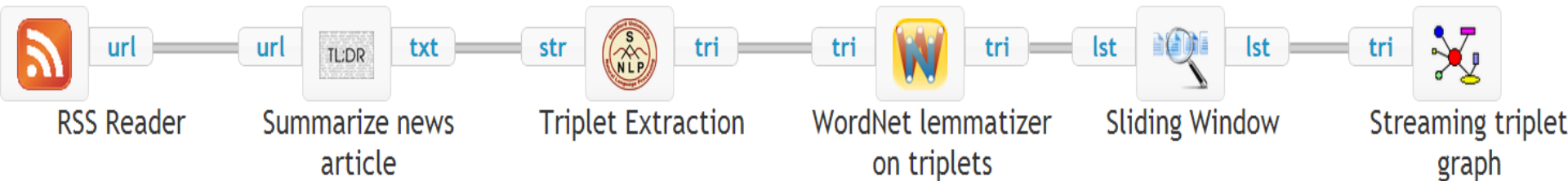


# CloudFlows platform

- **Large algorithm repository**
  - Relational data mining
  - All Orange algorithms
  - WEKA algorithms as web services
  - Data and results visualization
  - Text analysis
  - Social network analysis
  - Analysis of big data streams
- **Large workflow repository**
  - Enables access to our technology heritage

# “Big Data” Use Case

- Real-time analysis of big data streams
- Example: semantic graph construction from news streams. <http://clowdflows.org/workflow/1729/>.



- Example: news monitoring by graph visualization (graph of CNN RSS feeds)

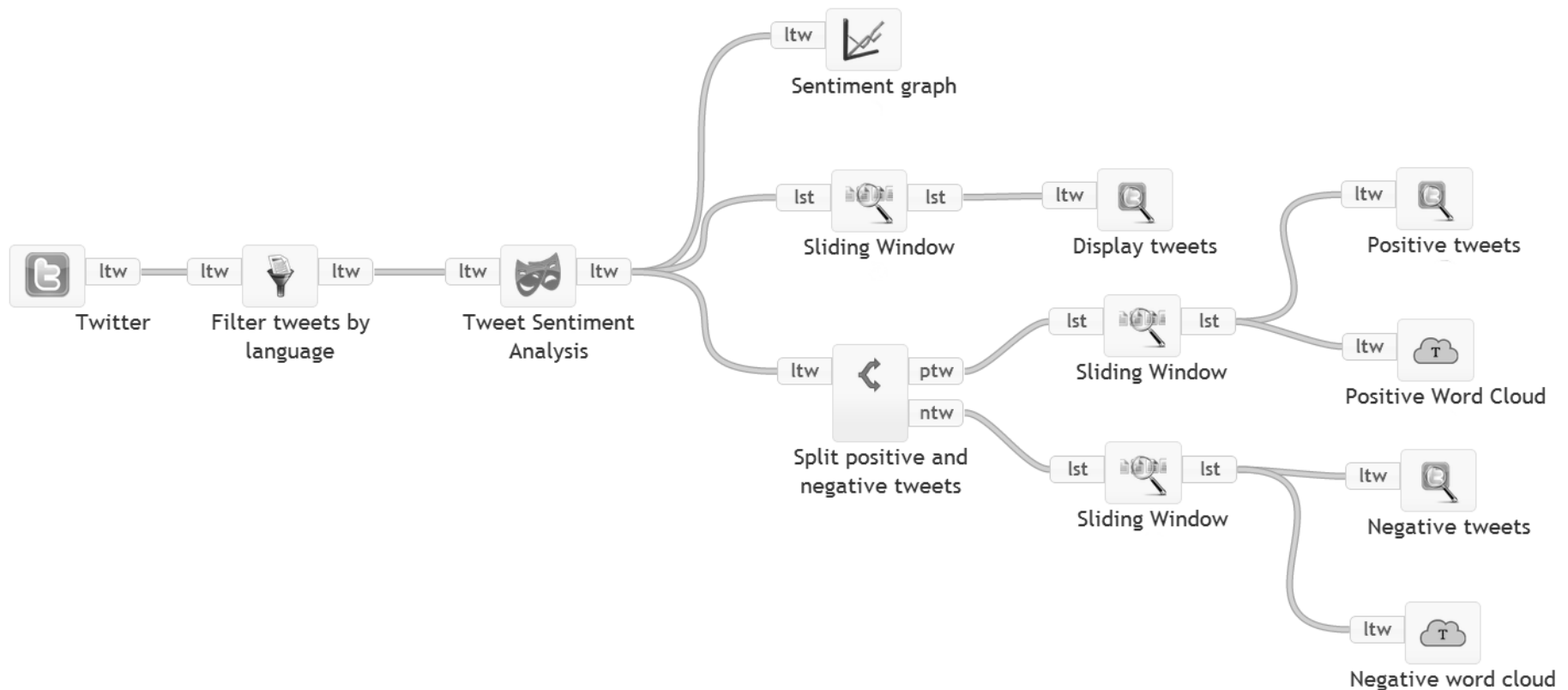
<http://clowdflows.org/streams/data/31/1>





# “Big Data” Use Case

- Analysis of positive/negative sentiment of tweets in real time: <http://clowdflows.org/workflow/1041/>.



# TextFlows

- Motivation:
  - Develop an online text mining platform for composition, execution and sharing of text mining workflows
- TextFlows platform – fork of ClowdFlows.org:
  - Web-based user interface
  - Visual programming
  - Big roster of existing workflow (mostly data mining) components
  - Cloud-based service-oriented architecture

# Comparison with ClowdFlows

- ClowdFlows:
  - Roster of not fully compatible widgets, developed separately by each workflow developer, non-systematic approach
  - Missing components for text mining and natural language processing
- TextFlows:
  - Includes numerous text mining and NLP widgets
  - Widgets grouped by their functionality
  - New common text representation structure

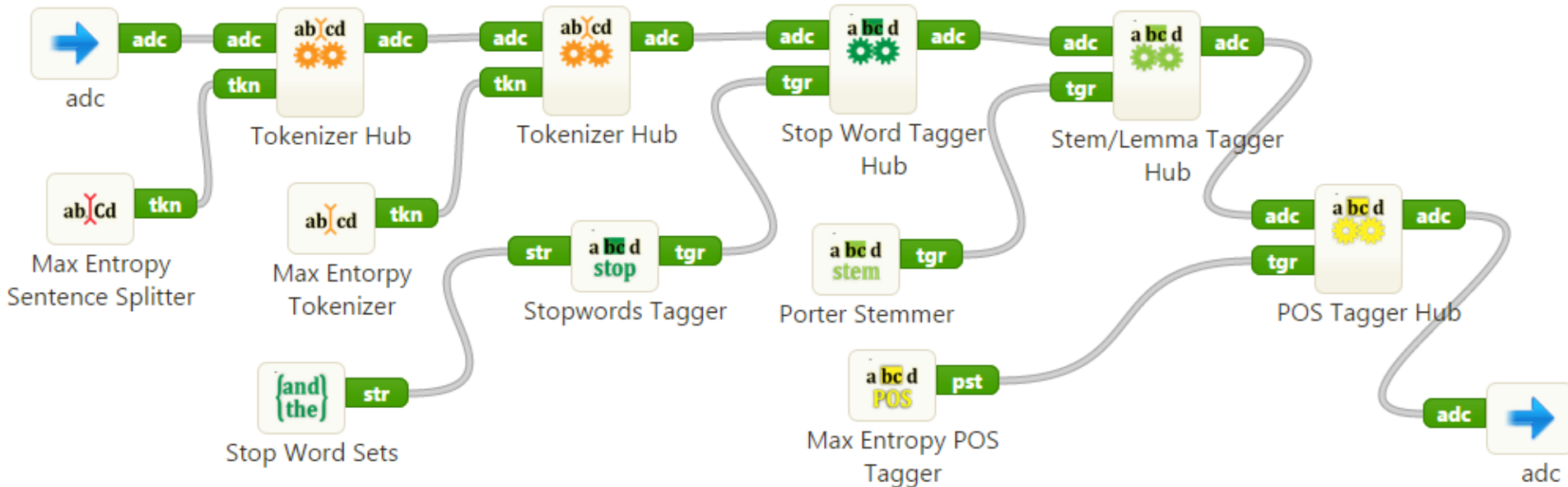
# The TextFlows Modules

- Implemented packages:
  - Text preprocessing
    - Latino (Grčar, 2015)
    - NLTK (Bird et. al., 2006)
    - Scikit-learn (Pedregosa et. al., 2011)
  - Text Categorisation
  - Literature based discovery (Juršič et. al., 2012)
  - Noise handling (Sluban et. al., 2012)
  - Visual performance evaluation (Sluban et. al., 2012)
  - Relational data mining through wordification (Perovšek et al., 2015)
- Advanced text processing workflows and use cases developed in this thesis

# Advanced workflows in TextFlows

- NLP scenarios
  1. Document preprocessing
  2. Classifier evaluation
  3. POS tagging classification evaluation
  4. Stemming classification evaluation
  5. Outlier document detection
- Complex data/text analysis scenarios
  1. Relational data mining, propositionalization
  2. Literature based cross-context knowledge discovery

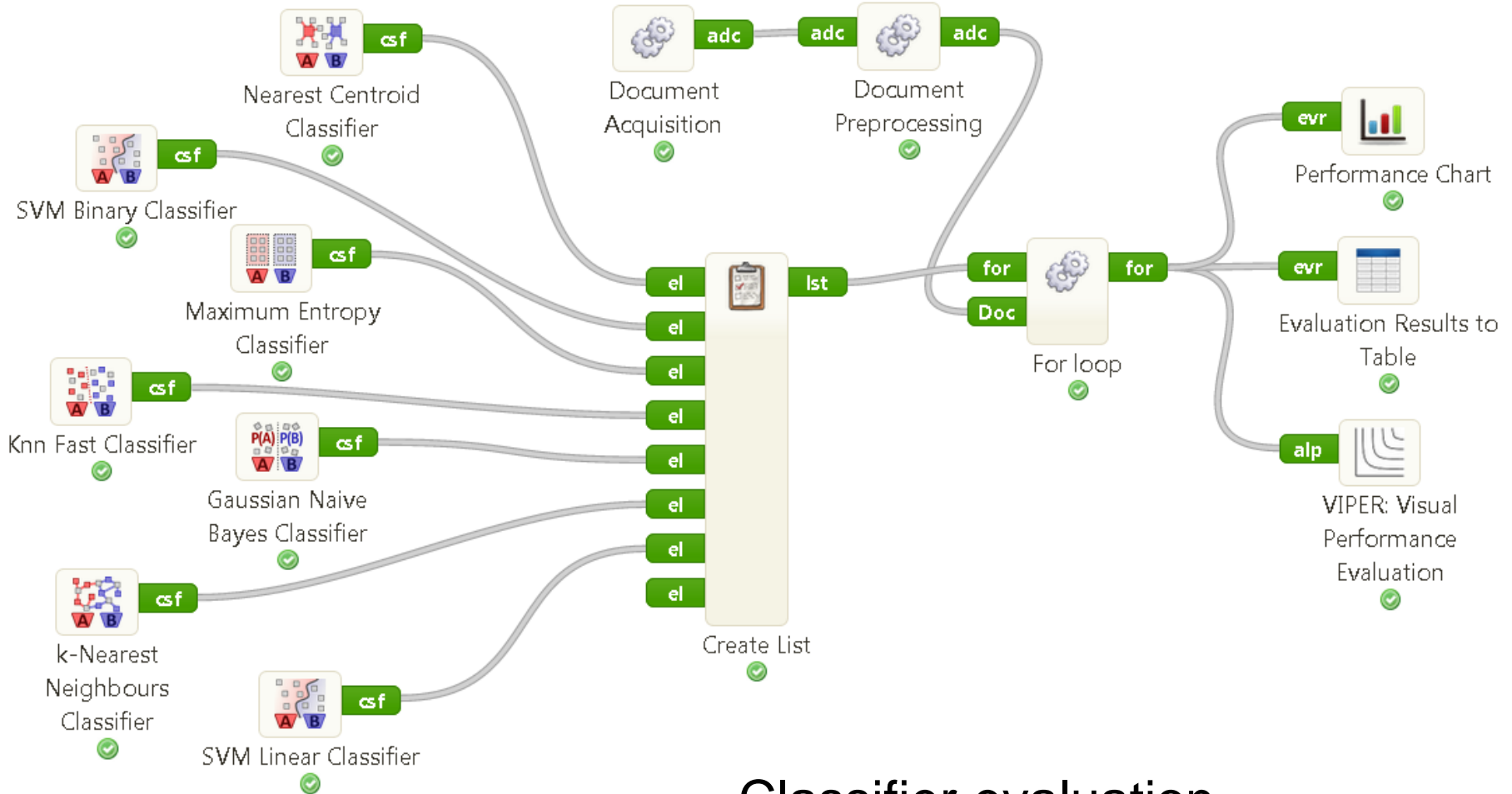
# 1. Document preprocessing



Simple Document Preprocessing

<http://textflows.org/workflow/604/>

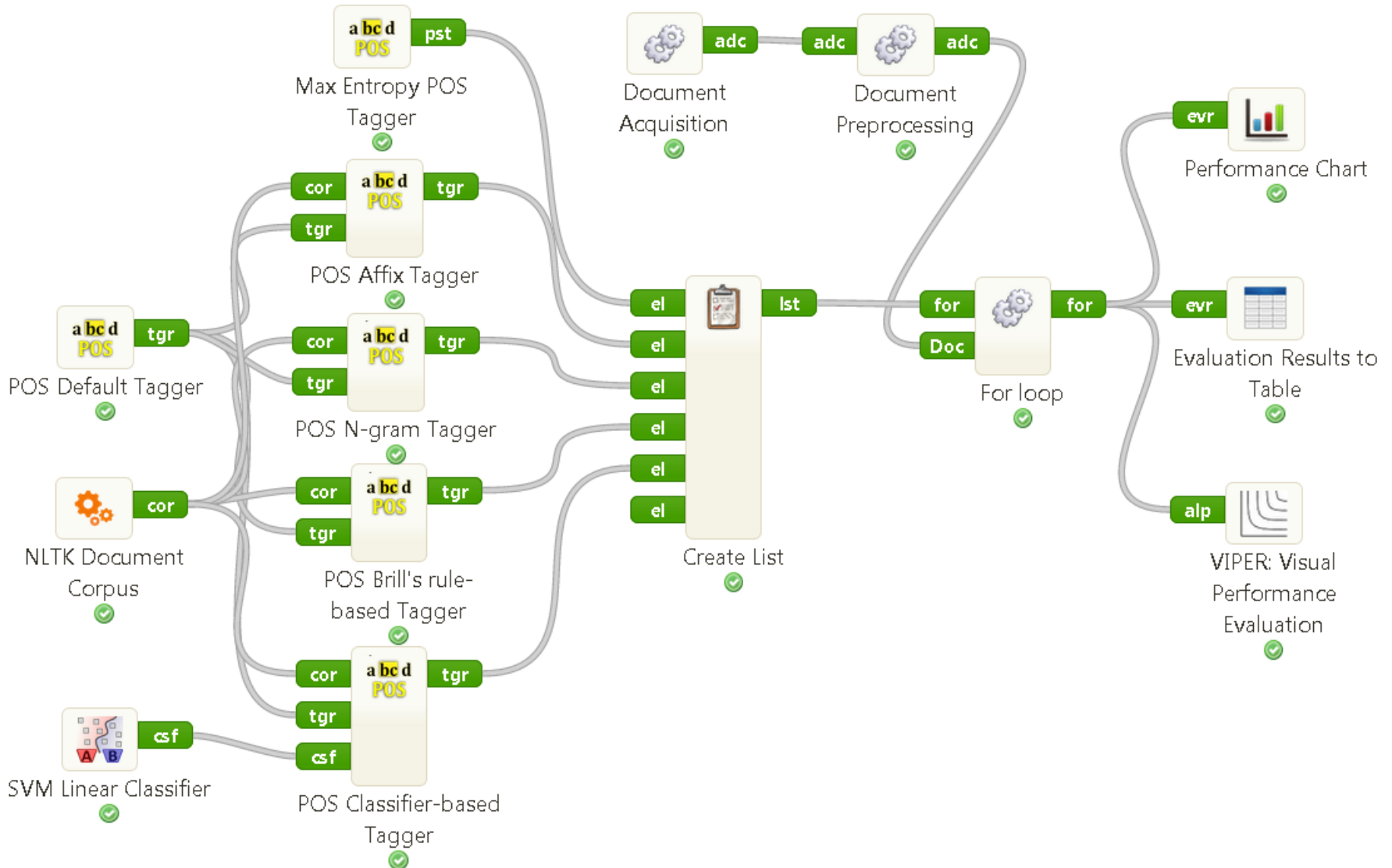
# 2. Classifier Evaluation



Classifier evaluation

<http://textflows.org/workflow/350/>

# 3. POS Tagger Evaluation



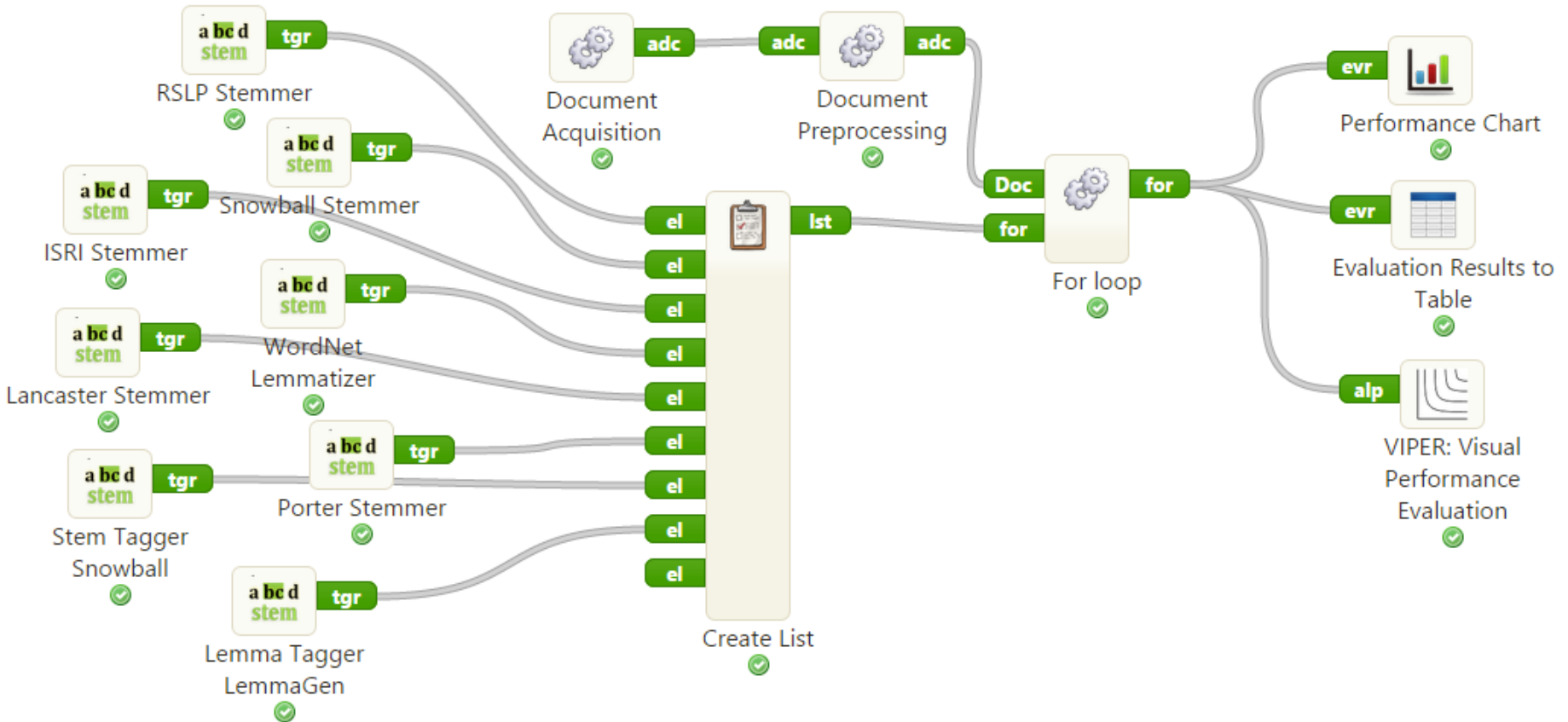


# 3. POS Tagger Evaluation

Table 2: POS tagger evaluation on the Kenyan elections database.

Library	Tagger	Recall	Precision	F1 score	Classification accuracy	AUC
	no POS tagger	0.98	0.93	0.95	95.24%	0.95
LATINO	Maximum Entropy POS Tagger	0.98	0.94	0.96	96.10%	0.96
NLTK	POS Affix Tagger	0.98	0.94	0.96	95.67%	0.96
NLTK	POS Ngram Tagger	0.98	0.95	0.96	96.10%	0.96
NLTK	POS Brill Tagger	0.97	0.93	0.95	95.24%	0.95
NLTK+ scikit-learn	POS Classifier Based Tagger (using SVM Linear Classifier)	0.98	0.95	0.96	96.32%	0.96

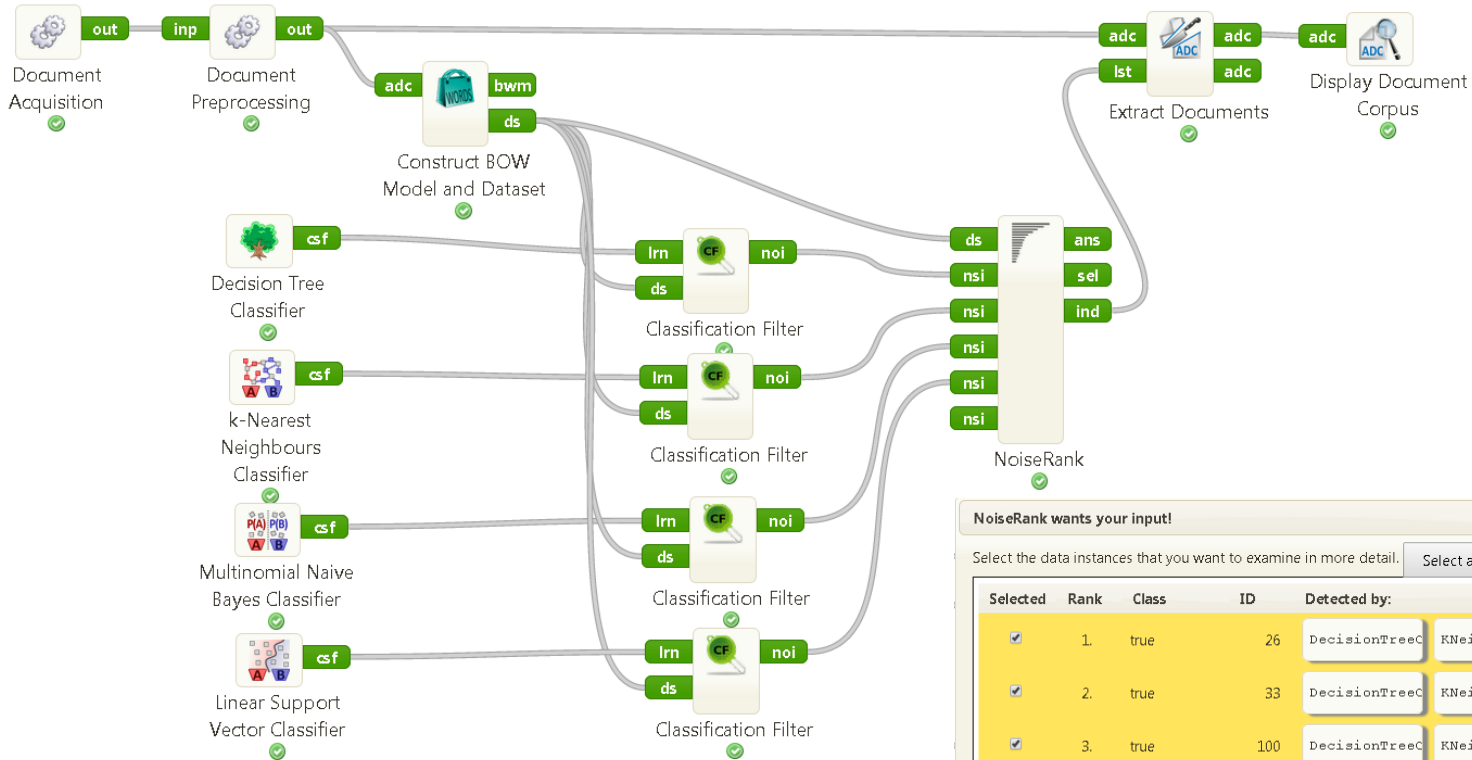
# 4. Stemmer Evaluation



# 4. Stemmer Evaluation

Library	Stemmer/Lemmatizer	F1 score	Class. accuracy	AUC
	no stemmer	0.94	94.16%	0.94
NLTK	RSLP Stemmer	0.95	95.24%	0.95
NLTK	Snowball Stemmer	0.96	96.10%	0.96
NLTK	ISRI Stemmer	0.96	96.10%	0.96
NLTK	WordNet Lemmatizer	0.96	96.32%	0.96
NLTK	Lancaster Stemmer	0.95	94.81%	0.95
NLTK	Porter Stemmer	0.95	95.24%	0.95
Latino	Stem Tagger Snowball	0.95	95.24%	0.95
Latino	Lemma Tagger LemmaGen	0.95	95.24%	0.95

# 5. Outlier Document Detection



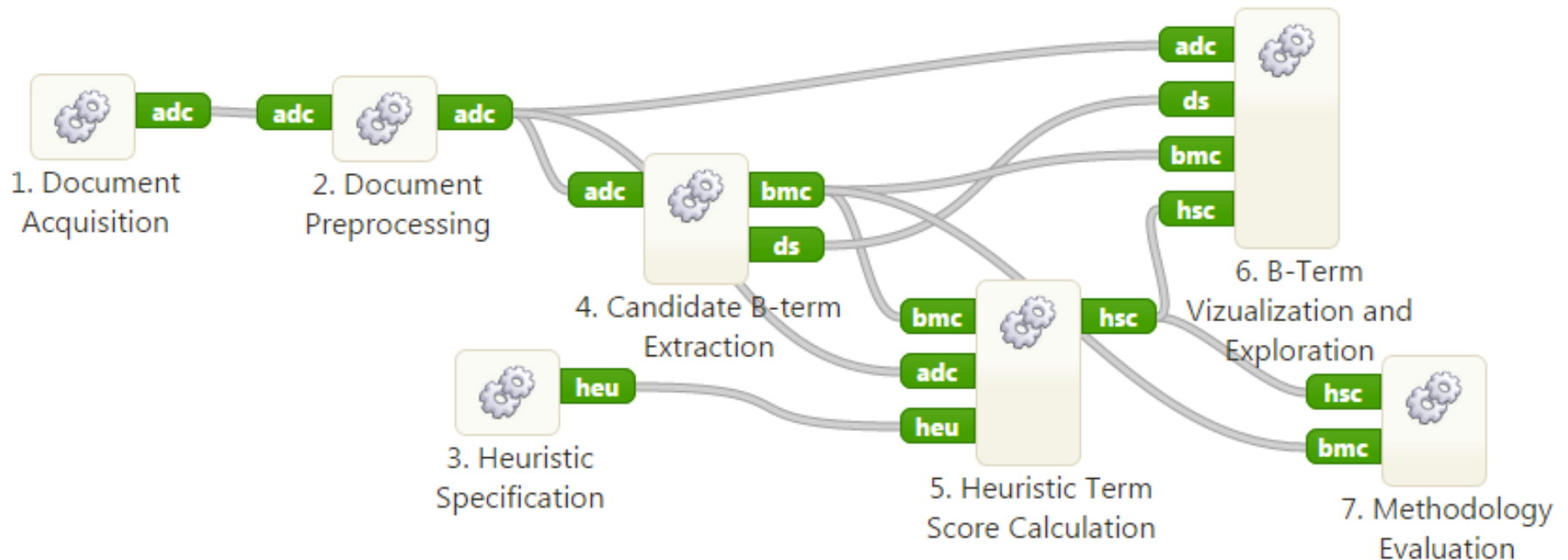
NoiseRank wants your input!

Select the data instances that you want to examine in more detail.

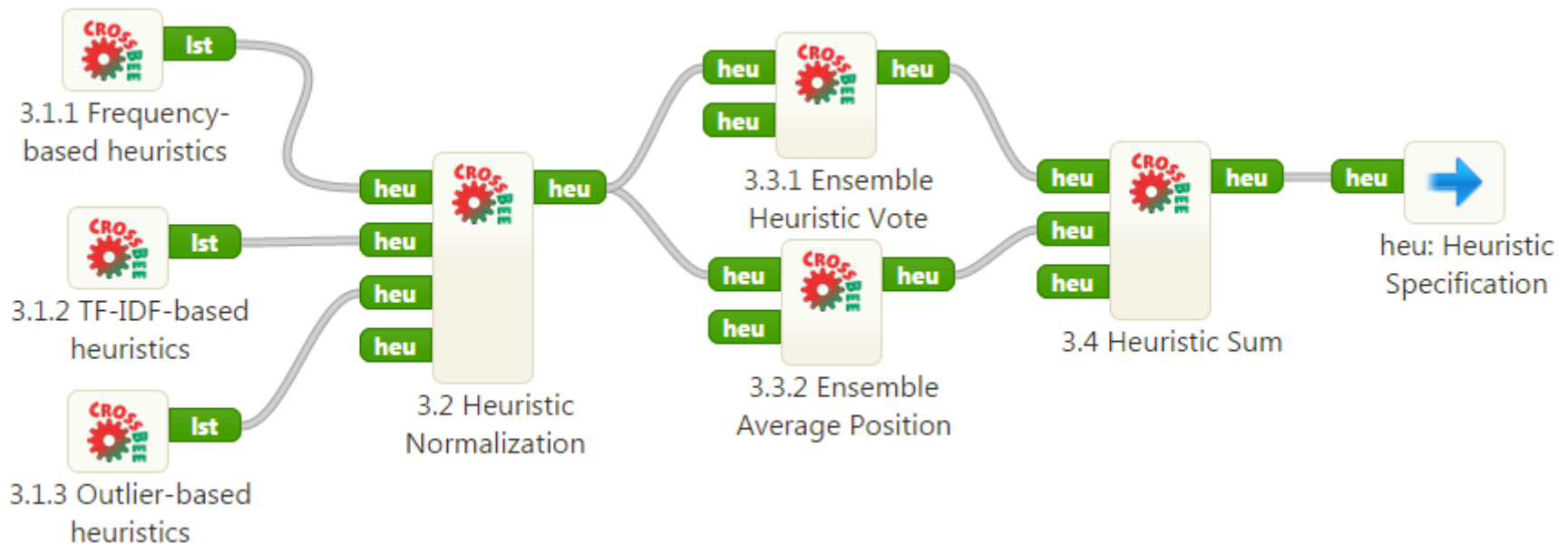
Selected	Rank	Class	ID	Detected by:			
<input checked="" type="checkbox"/>	1.	true	26	DecisionTreeC	KNeighborsCle	MultinomialNE	LinearSVC
<input checked="" type="checkbox"/>	2.	true	33	DecisionTreeC	KNeighborsCle	MultinomialNE	LinearSVC
<input checked="" type="checkbox"/>	3.	true	100	DecisionTreeC	KNeighborsCle	MultinomialNE	LinearSVC
<input type="checkbox"/>	4.	true	6	KNeighborsCle	MultinomialNE	LinearSVC	
<input type="checkbox"/>	5.	true	10	DecisionTreeC	MultinomialNE	LinearSVC	
<input type="checkbox"/>	6.	true	18	DecisionTreeC	MultinomialNE	LinearSVC	
<input type="checkbox"/>	7.	true	44	KNeighborsCle	MultinomialNE	LinearSVC	
<input type="checkbox"/>	8.	true	57	KNeighborsCle	MultinomialNE	LinearSVC	

Apply

# The Methodology Workflow



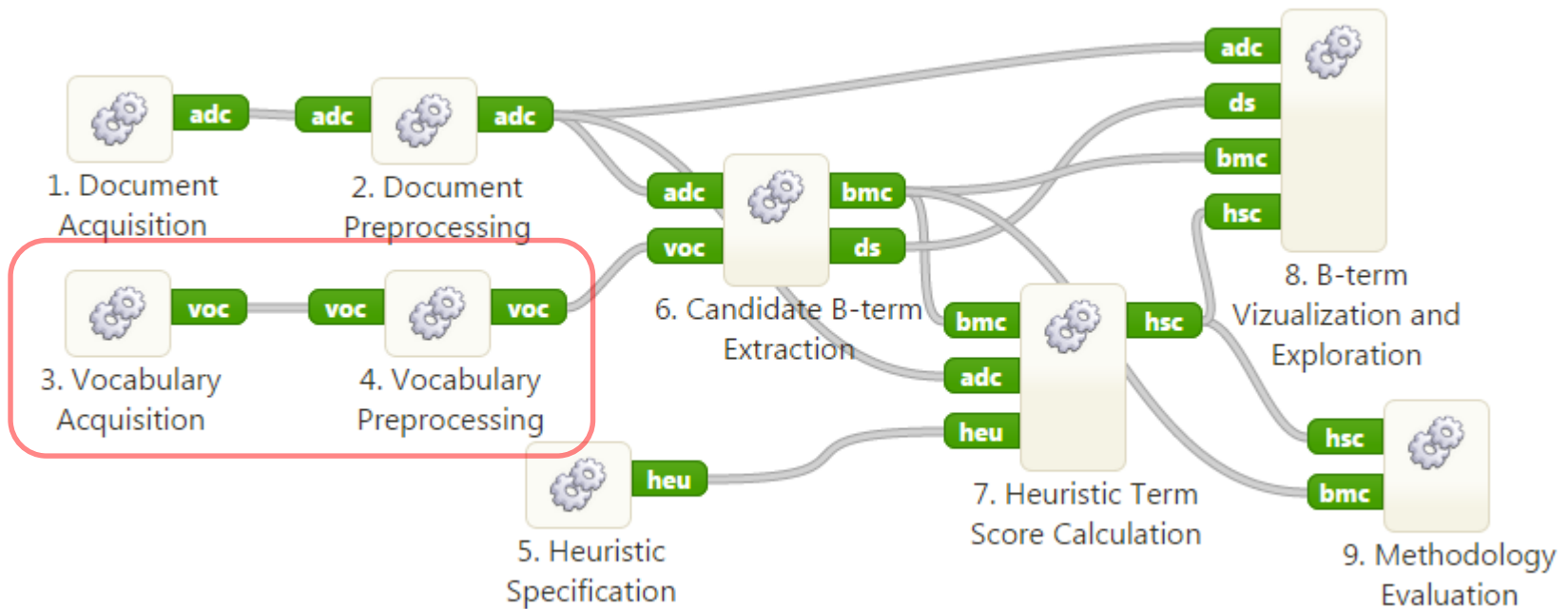
# Heuristics Specification Subprocess



# Controlled Vocabularies

- TextFlows BoW construction widget accepts synonyms and term whitelist as an additional input
- New lexicology package in TextFlows contains controlled vocabularies:
  - MeSH term filter
  - GO term filter
  - HGNC synonyms

# Extended Methodology

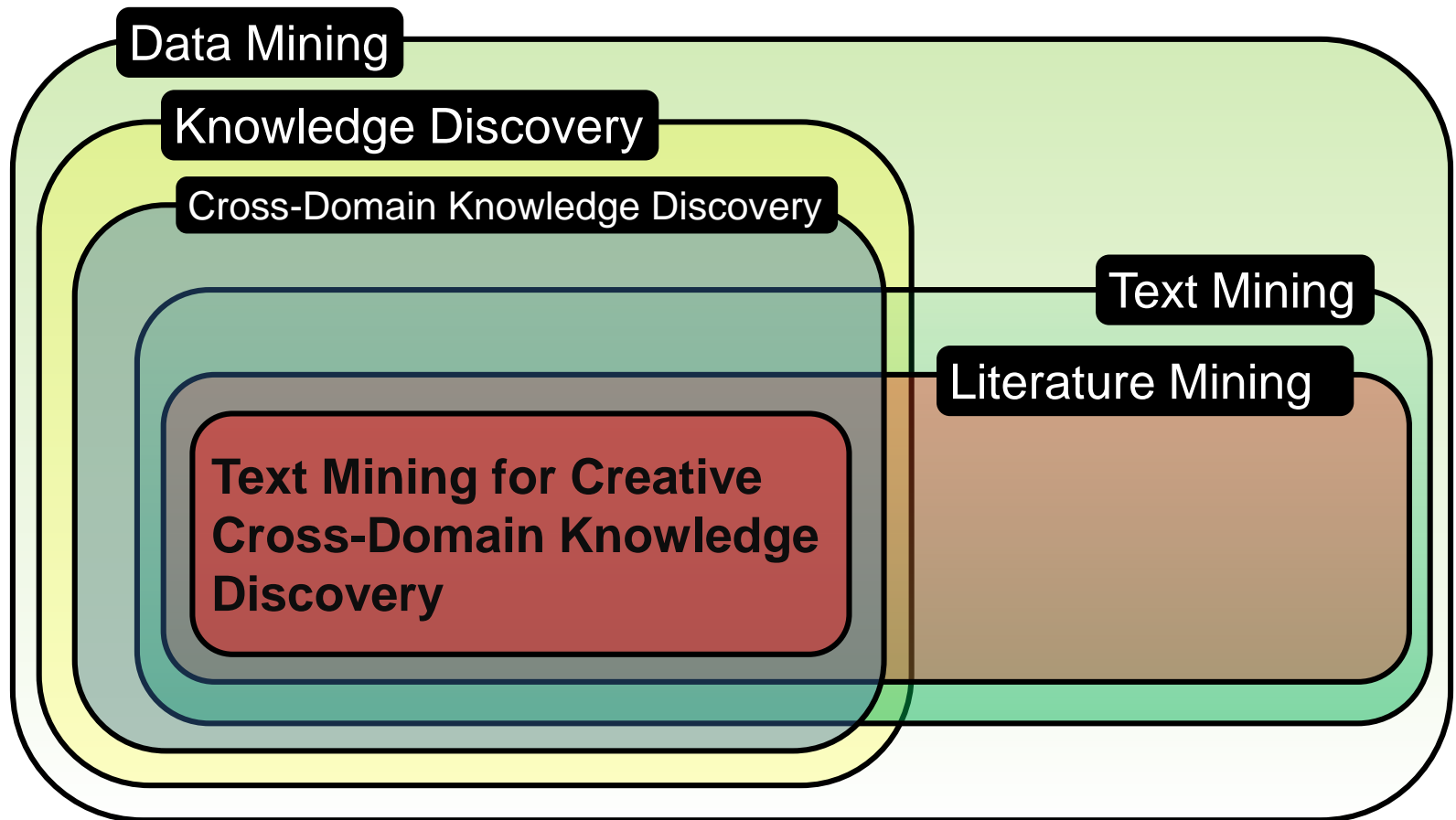




# Summary and conclusions

- 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)
- TextFlows text processing and text mining platform

# 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)