MANCHESTER
1824
The University of Manchester
Manchester Institute of Biotechnology

NaCTeM
The National Centre for Text Mining

Text mining:
The view from NaCTeM

John McNaught
Deputy Director
National Centre for Text Mining
www.nactem.ac.uk
John.McNaught@manchester.ac.uk

Overview

- Brief background
- Semantic search
- There’s more to TM than search
- Supporting systematic reviewing
- Interoperability aspects
- Gaps and issues, research data

How do we (humans) discover?

- Find, read, learn, analyse a lot
- Ask “What if…?”
- Construct hypotheses, test them
  - Explore many avenues, associations
- Work collaboratively
- Share results and data with others
  - Reproducibility → validation
- Integrate heterogeneous data/information/ knowledge
- (vs. serendipity)

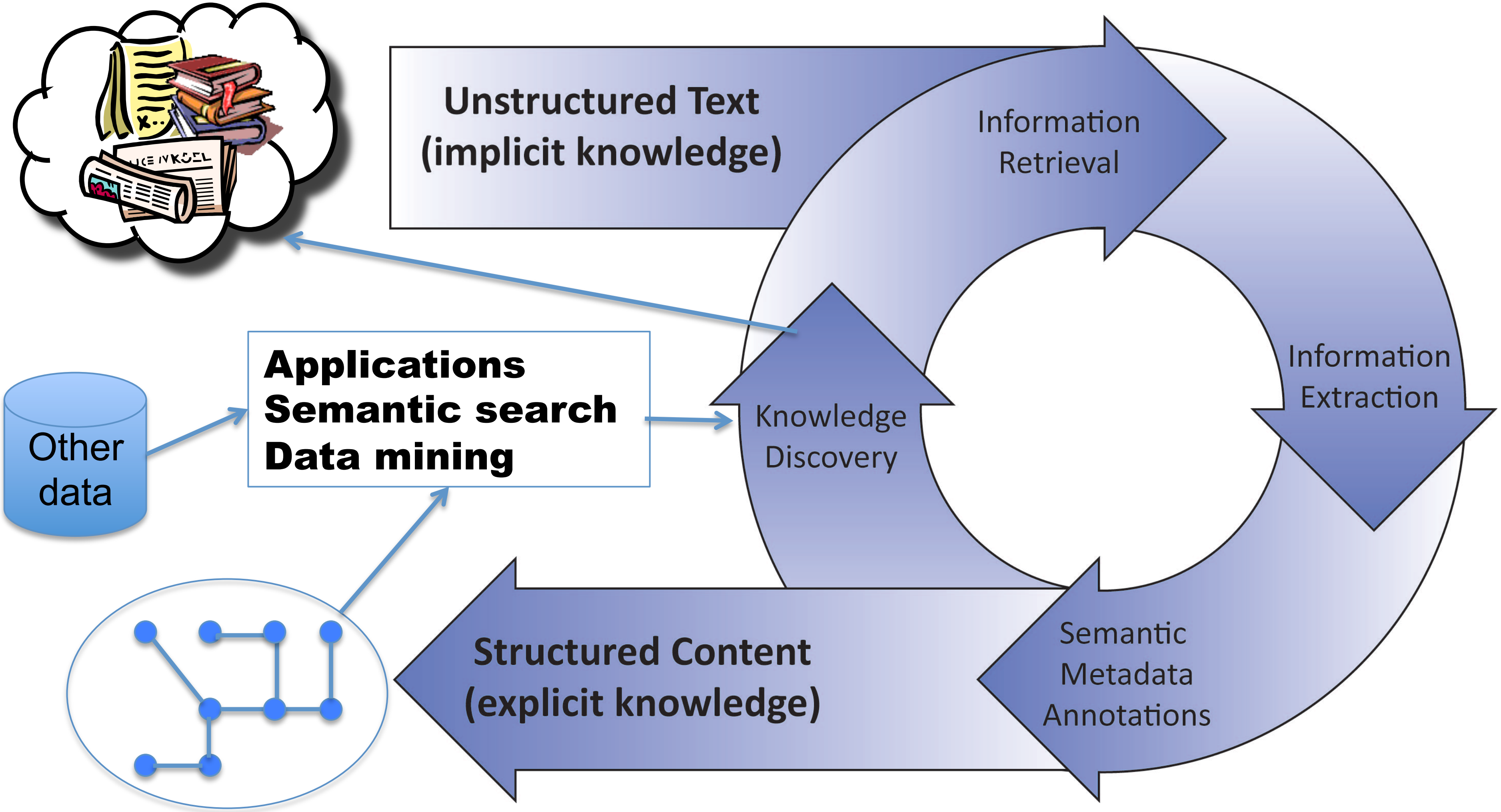
Barriers to discovery

- Find: *document* oriented, too many hits
- Read: too much to read, even if we find relevant hits
- Learn: too fast growth to keep up, to know most things
- Analyse: duplication of efforts, many new results to *document*
- Construct hypotheses: hard, can't tell which are most promising, or if have missed any
- Share: primary vehicles are *documents* and curated databases (massive curation backlog)
- Integrate: *document* often the key, hard to link in to different worlds of data, information, knowledge

How does TM aid discovery?

- **Find**: more precise, relevant information, *within* and *across* documents
- **Read**: “machine reading” much faster than human
- **Learn**: extracts, packages, links, synthesises, summarises, reduces burden
- **Analyse**: recognises duplication; clusters, classifies, drives semantic author aids
- **Construct hypotheses**: rapidly finds and *ranks* unknown associations for testing
- **Share**: reduces curation effort, complements and validates data bases
- **Integrate**: links documents deeply into worlds of data, information and knowledge

Text mining in a nutshell



Unstructured Text
(implicit knowledge)

Information Retrieval

Information Extraction

Semantic Metadata Annotations

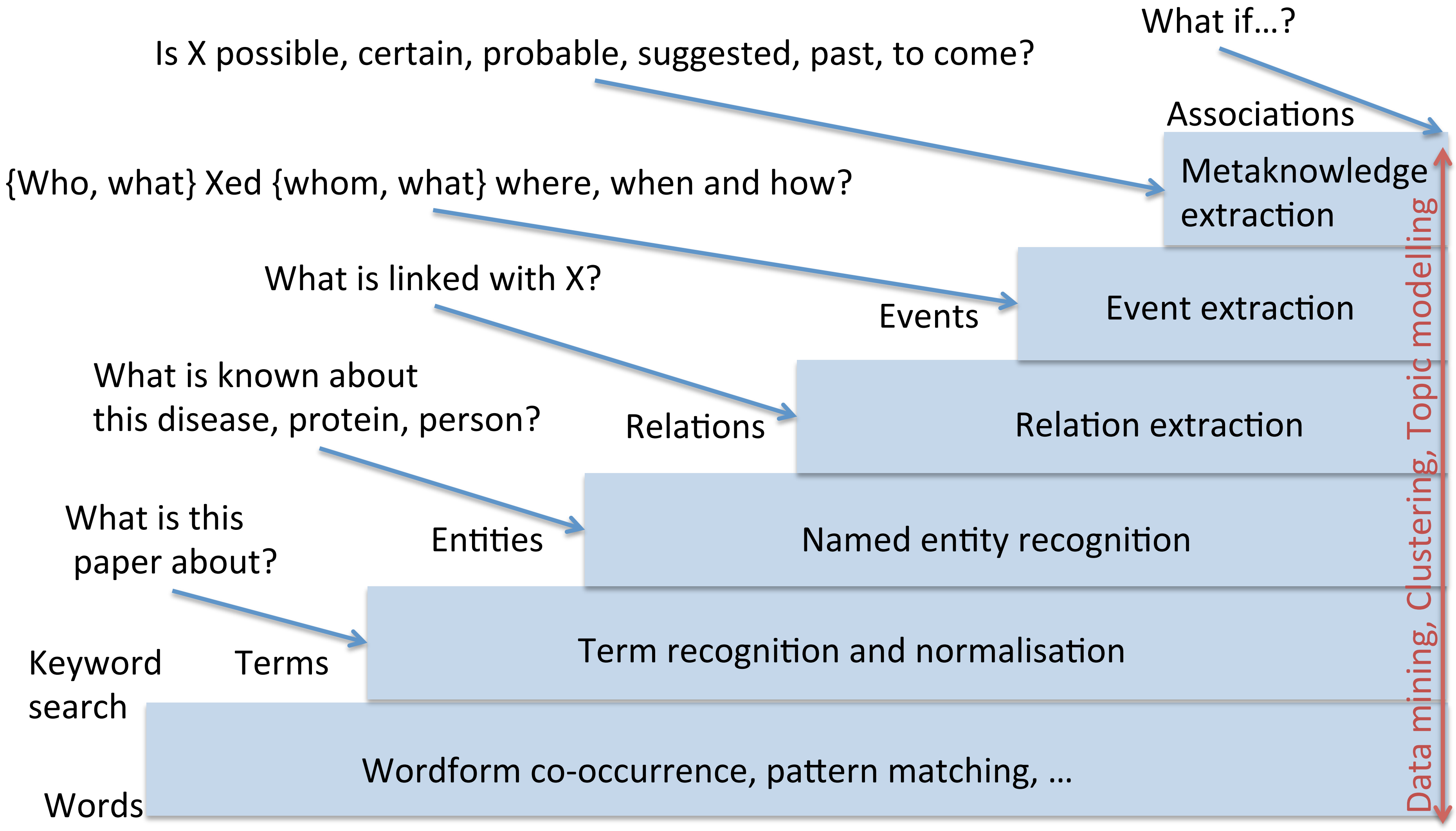
Structured Content
(explicit knowledge)

Knowledge Discovery

Applications
Semantic search
Data mining

Other data

Increased sophistication? Increased customisation!



What if…?

Is X possible, certain, probable, suggested, past, to come?

Associations

Metaknowledge extraction

{Who, what} Xed {whom, what} where, when and how?

What is linked with X?

Events

Event extraction

What is known about this disease, protein, person?

Relations

Relation extraction

What is this paper about?

Entities

Named entity recognition

Keyword search

Terms

Term recognition and normalisation

Words

Wordform co-occurrence, pattern matching, …

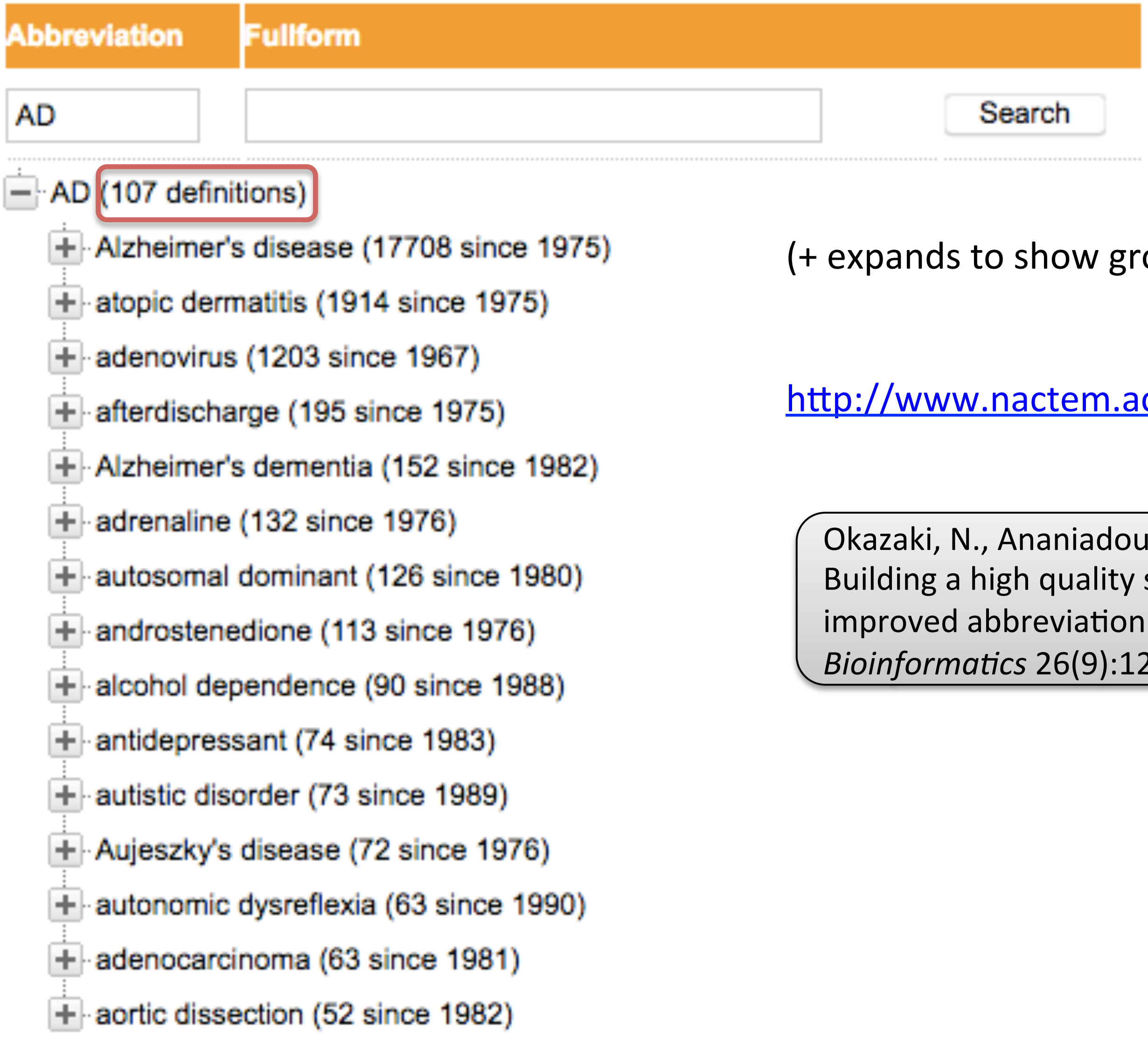
Data mining, Clustering, Topic modelling

Why is NL hard?
Ambiguity, ambiguity, ambiguity

- Picking up your litter puts road-workers at risk

  (sign on motorway)
- Women bitten by rabid bat found in crates outside pub
- "Last night I shot an elephant in my pyjamas. How he got into my pyjamas I'll never know." (Groucho Marx)
- FOOT HEADS ARMS BODY (The Times)
- Donald Trump, Mr Trump, Trump, the President of the United States, he, that man, …
- Replace cover and lock

Ambiguous Acronyms
(just in MEDLINE)



AD

- AD (107 definitions)
  - Alzheimer's disease (17708 since 1975)
  - atopic dermatitis (1914 since 1975)
  - adenovirus (1203 since 1967)
  - afterdischarge (195 since 1975)
  - Alzheimer's dementia (152 since 1982)
  - adrenaline (132 since 1976)
  - autosomal dominant (126 since 1980)
  - androstenedione (113 since 1976)
  - alcohol dependence (90 since 1988)
  - antidepressant (74 since 1983)
  - autistic disorder (73 since 1989)
  - Aujeszky's disease (72 since 1976)
  - autonomic dysreflexia (63 since 1990)
  - adenocarcinoma (63 since 1981)
  - aortic dissection (52 since 1982)

(+ expands to show grouped variants of a form)

http://www.nactem.ac.uk/software/acromine/

Okazaki, N., Ananiadou, S. and Tsujii, J. (2010). Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics* 26(9):1246–53

Nuances of language, hedging

S1 = We found that Y activates the expression of X

S2 = We examined the effect of Y on expression of X

S3 = These results suggest that Y has no effect on expression of X

S4 = Y is known to increase expression of X

S5 = Addition of Y slightly increased the expression of X

S6 = These results suggest that Y might affect the expression of X

Same ‘activation’ and ‘expression’ events, different meanings

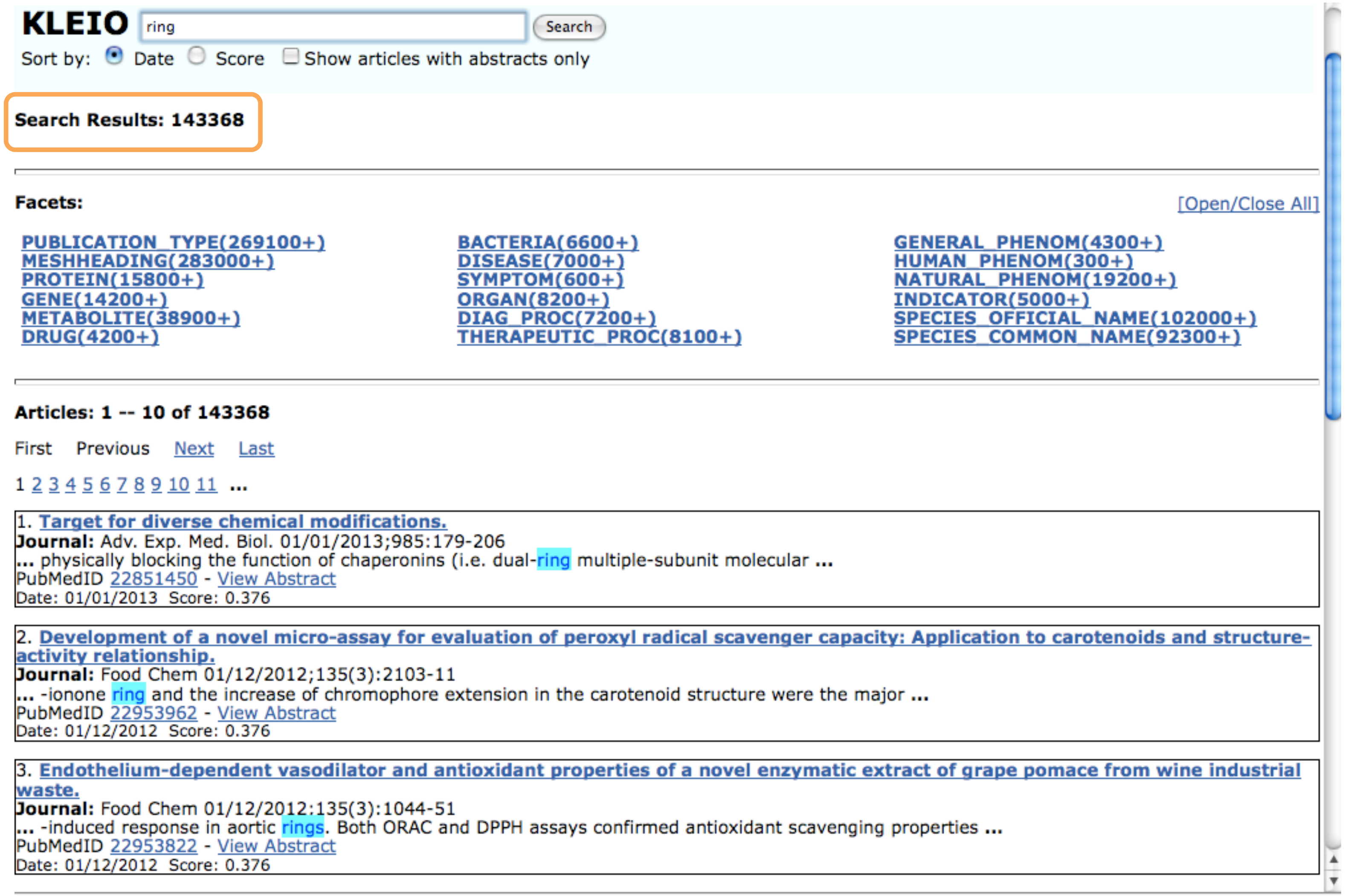
Semantic search using named entities

- Process collection to identify names of entities of domain interest
  - Gene, protein, metabolite, species, …
  - Person, organisation, location, date, …
  - Disease, symptom, drug, treatment, …
- Index entities for semantic search
- http://www.nactem.ac.uk/Kleio/

  (Currently being re-engineered and updated, watch this space)

Trying to find documents on a protein zinc finger domain called RING

Classic keyword search for “ring” returns 143368 hits



KLEIO ring Search

Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 143368

Facets: [Open/Close All]

PUBLICATION_TYPE(269100+)
MESHHEADING(283000+)
PROTEIN(15800+)
GENE(14200+)
METABOLITE(38900+)
DRUG(4200+)

BACTERIA(6600+)
DISEASE(7000+)
SYMPTOM(600+)
ORGAN(8200+)
DIAG_PROC(7200+)
THERAPEUTIC_PROC(8100+)

GENERAL_PHENOM(4300+)
HUMAN_PHENOM(300+)
NATURAL_PHENOM(19200+)
INDICATOR(5000+)
SPECIES_OFFICIAL_NAME(102000+)
SPECIES_COMMON_NAME(92300+)

Articles: 1 -- 10 of 143368

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

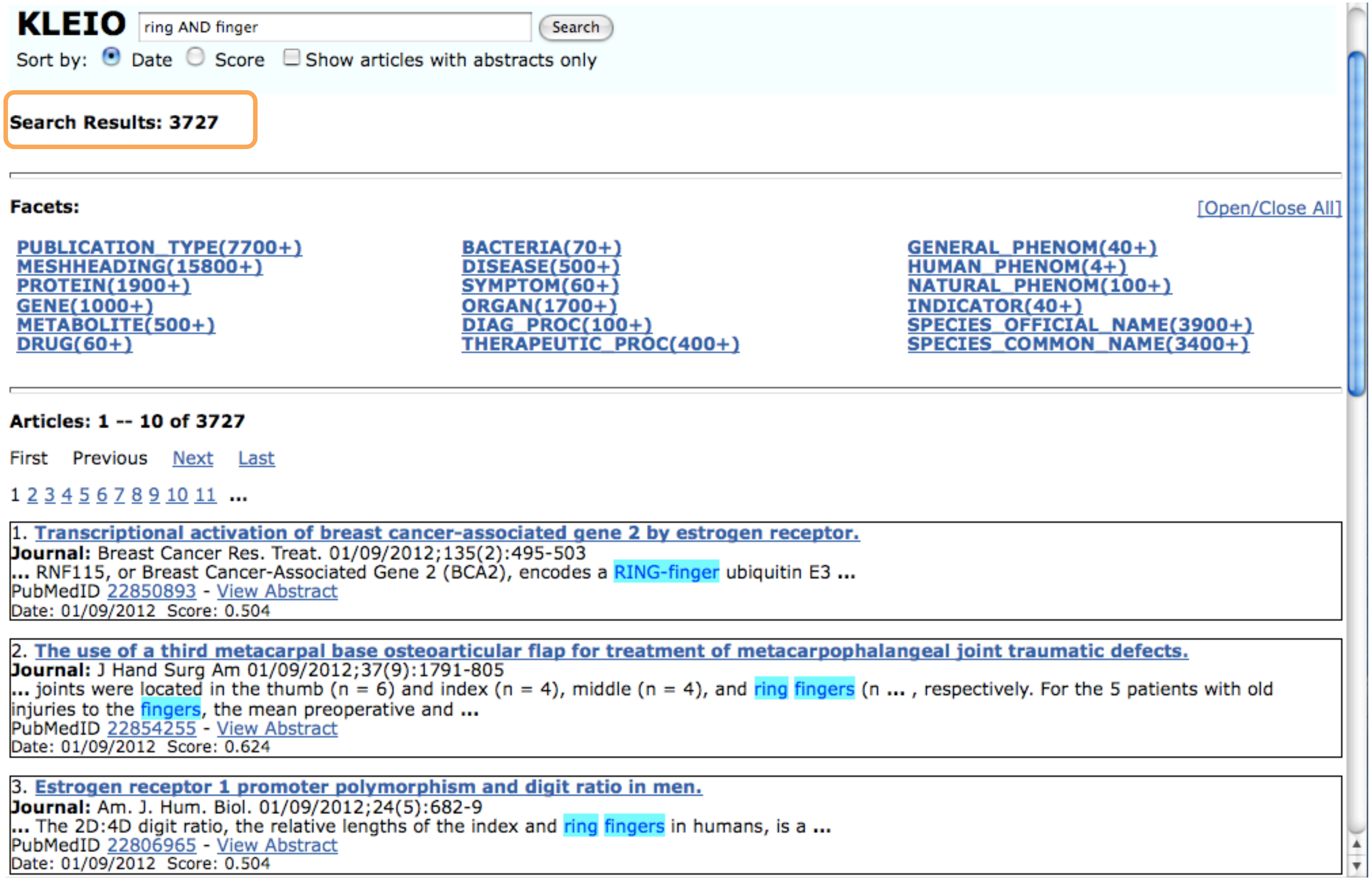
1. **Target for diverse chemical modifications.**
Journal: Adv. Exp. Med. Biol. 01/01/2013;985:179-206
... physically blocking the function of chaperonins (i.e. dual-ring multiple-subunit molecular ...
PubMedID 22851450 - View Abstract
Date: 01/01/2013 Score: 0.376

2. **Development of a novel micro-assay for evaluation of peroxyl radical scavenger capacity: Application to carotenoids and structure-activity relationship.**
Journal: Food Chem 01/12/2012;135(3):2103-11
... -ionone ring and the increase of chromophore extension in the carotenoid structure were the major ...
PubMedID 22953962 - View Abstract
Date: 01/12/2012 Score: 0.376

3. **Endothelium-dependent vasodilator and antioxidant properties of a novel enzymatic extract of grape pomace from wine industrial waste.**
Journal: Food Chem 01/12/2012;135(3):1044-51
... -induced response in aortic rings. Both ORAC and DPPH assays confirmed antioxidant scavenging properties ...
PubMedID 22953822 - View Abstract
Date: 01/12/2012 Score: 0.376

“Power user” tries Boolean search

Classic Boolean keyword search for “ring AND finger” does not help much



KLEIO ring AND finger Search

Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 3727

Facets: [Open/Close All]

PUBLICATION_TYPE(7700+)
MESHHEADING(15800+)
PROTEIN(1900+)
GENE(1000+)
METABOLITE(500+)
DRUG(60+)

BACTERIA(70+)
DISEASE(500+)
SYMPTOM(60+)
ORGAN(1700+)
DIAG_PROC(100+)
THERAPEUTIC_PROC(400+)

GENERAL_PHENOM(40+)
HUMAN_PHENOM(4+)
NATURAL_PHENOM(100+)
INDICATOR(40+)
SPECIES_OFFICIAL_NAME(3900+)
SPECIES_COMMON_NAME(3400+)

Articles: 1 -- 10 of 3727

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

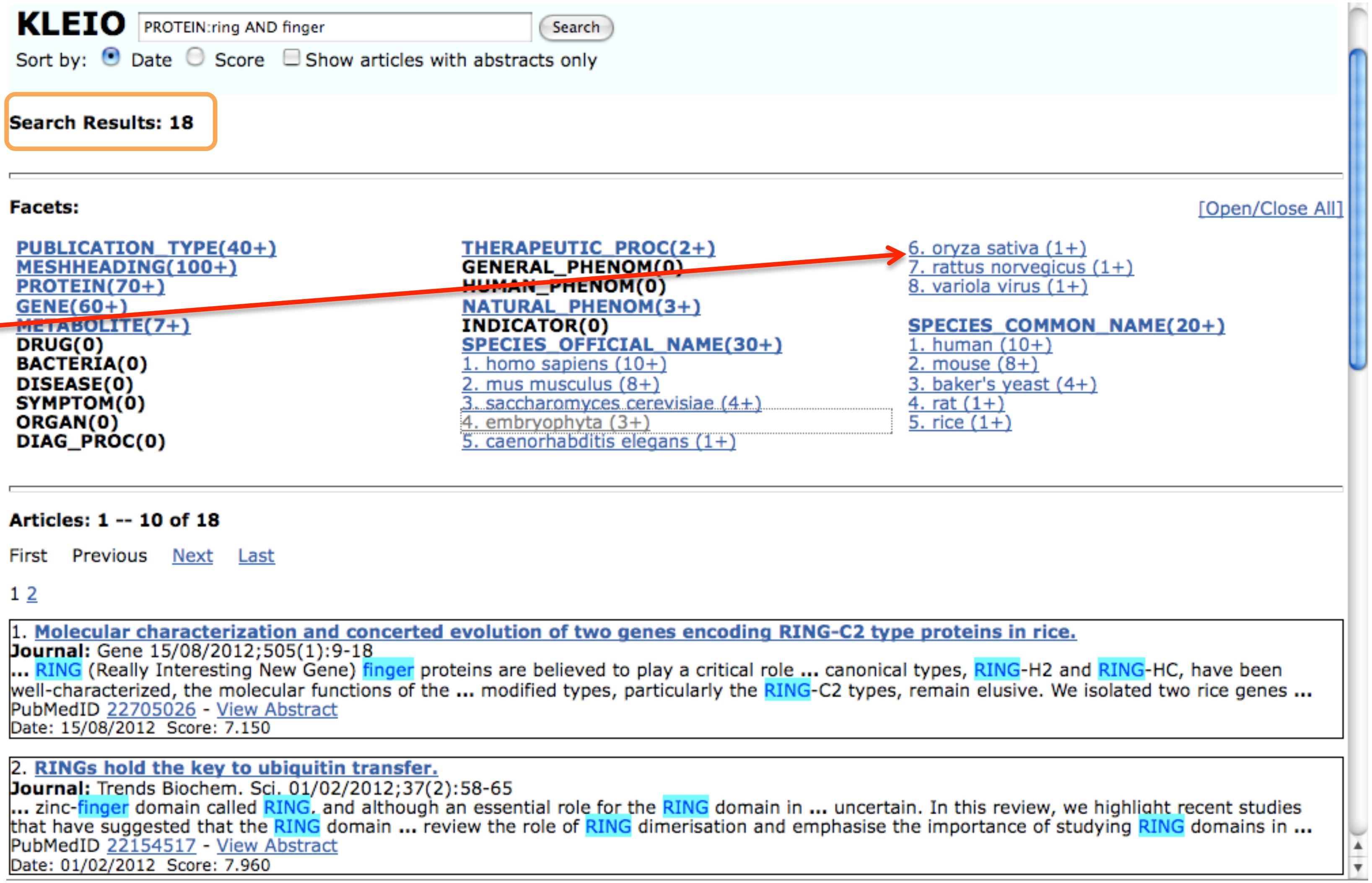
1. **Transcriptional activation of breast cancer-associated gene 2 by estrogen receptor.**
Journal: Breast Cancer Res. Treat. 01/09/2012;135(2):495-503
... RNF115, or Breast Cancer-Associated Gene 2 (BCA2), encodes a RING-finger ubiquitin E3 ...
PubMedID 22850893 - View Abstract
Date: 01/09/2012 Score: 0.504

2. **The use of a third metacarpal base osteoarticular flap for treatment of metacarpophalangeal joint traumatic defects.**
Journal: J Hand Surg Am 01/09/2012;37(9):1791-805
... joints were located in the thumb (n = 6) and index (n = 4), middle (n = 4), and ring fingers (n ... , respectively. For the 5 patients with old injuries to the fingers, the mean preoperative and ...
PubMedID 22854255 - View Abstract
Date: 01/09/2012 Score: 0.624

3. **Estrogen receptor 1 promoter polymorphism and digit ratio in men.**
Journal: Am. J. Hum. Biol. 01/09/2012;24(5):682-9
... The 2D:4D digit ratio, the relative lengths of the index and ring fingers in humans, is a ...
PubMedID 22806965 - View Abstract
Date: 01/09/2012 Score: 0.504

Use semantic search to greater effect: named entity search for "PROTEIN:ring AND finger"

Choosing entity facets narrows search further…



KLEIO PROTEIN:ring AND finger [Search]

Sort by: ● Date ○ Score ☐ Show articles with abstracts only

Search Results: 18

Facets: [Open/Close All]

PUBLICATION_TYPE(40+)
MESHHEADING(100+)
PROTEIN(70+)
GENE(60+)
METABOLITE(7+)
DRUG(0)
BACTERIA(0)
DISEASE(0)
SYMPTOM(0)
ORGAN(0)
DIAG_PROC(0)

THERAPEUTIC_PROC(2+)
GENERAL_PHENOM(0)
HUMAN_PHENOM(0)
NATURAL_PHENOM(3+)
INDICATOR(0)
SPECIES_OFFICIAL_NAME(30+)
1. homo sapiens (10+)
2. mus musculus (8+)
3. saccharomyces cerevisiae (4+)
4. embryophyta (3+)
5. caenorhabditis elegans (1+)
6. oryza sativa (1+)
7. rattus norvegicus (1+)
8. variola virus (1+)

SPECIES_COMMON_NAME(20+)
1. human (10+)
2. mouse (8+)
3. baker's yeast (4+)
4. rat (1+)
5. rice (1+)

Articles: 1 -- 10 of 18

First Previous Next Last

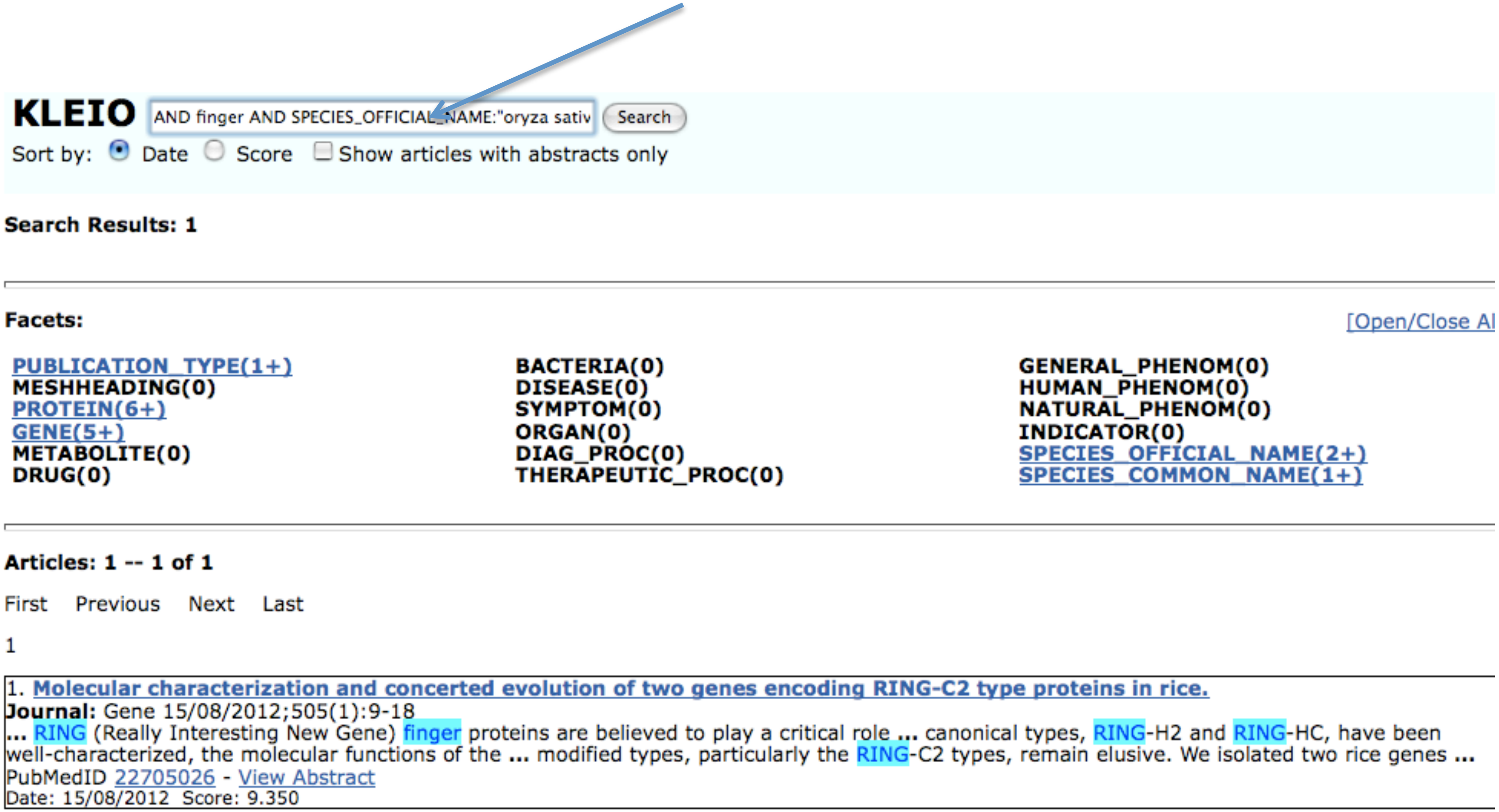
1 2

1. **Molecular characterization and concerted evolution of two genes encoding RING-C2 type proteins in rice.**
Journal: Gene 15/08/2012;505(1):9-18
... RING (Really Interesting New Gene) finger proteins are believed to play a critical role **...** canonical types, RING-H2 and RING-HC, have been well-characterized, the molecular functions of the **...** modified types, particularly the RING-C2 types, remain elusive. We isolated two rice genes **...**
PubMedID 22705026 - View Abstract
Date: 15/08/2012 Score: 7.150

2. **RINGs hold the key to ubiquitin transfer.**
Journal: Trends Biochem. Sci. 01/02/2012;37(2):58-65
... zinc-finger domain called RING. and although an essential role for the RING domain in **...** uncertain. In this review, we highlight recent studies that have suggested that the RING domain **...** review the role of RING dimerisation and emphasise the importance of studying RING domains in **...**
PubMedID 22154517 - View Abstract
Date: 01/02/2012 Score: 7.960

A couple of steps thus leads to the 1 document on RING related to rice

Choosing a facet value automatically expands the query for the user



KLEIO AND finger AND SPECIES_OFFICIAL_NAME:"oryza sativ Search

Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 1

Facets: [Open/Close Al

PUBLICATION_TYPE(1+)
MESHHEADING(0)
PROTEIN(6+)
GENE(5+)
METABOLITE(0)
DRUG(0)
BACTERIA(0)
DISEASE(0)
SYMPTOM(0)
ORGAN(0)
DIAG_PROC(0)
THERAPEUTIC_PROC(0)
GENERAL_PHENOM(0)
HUMAN_PHENOM(0)
NATURAL_PHENOM(0)
INDICATOR(0)
SPECIES_OFFICIAL_NAME(2+)
SPECIES_COMMON_NAME(1+)

Articles: 1 -- 1 of 1

First Previous Next Last

1

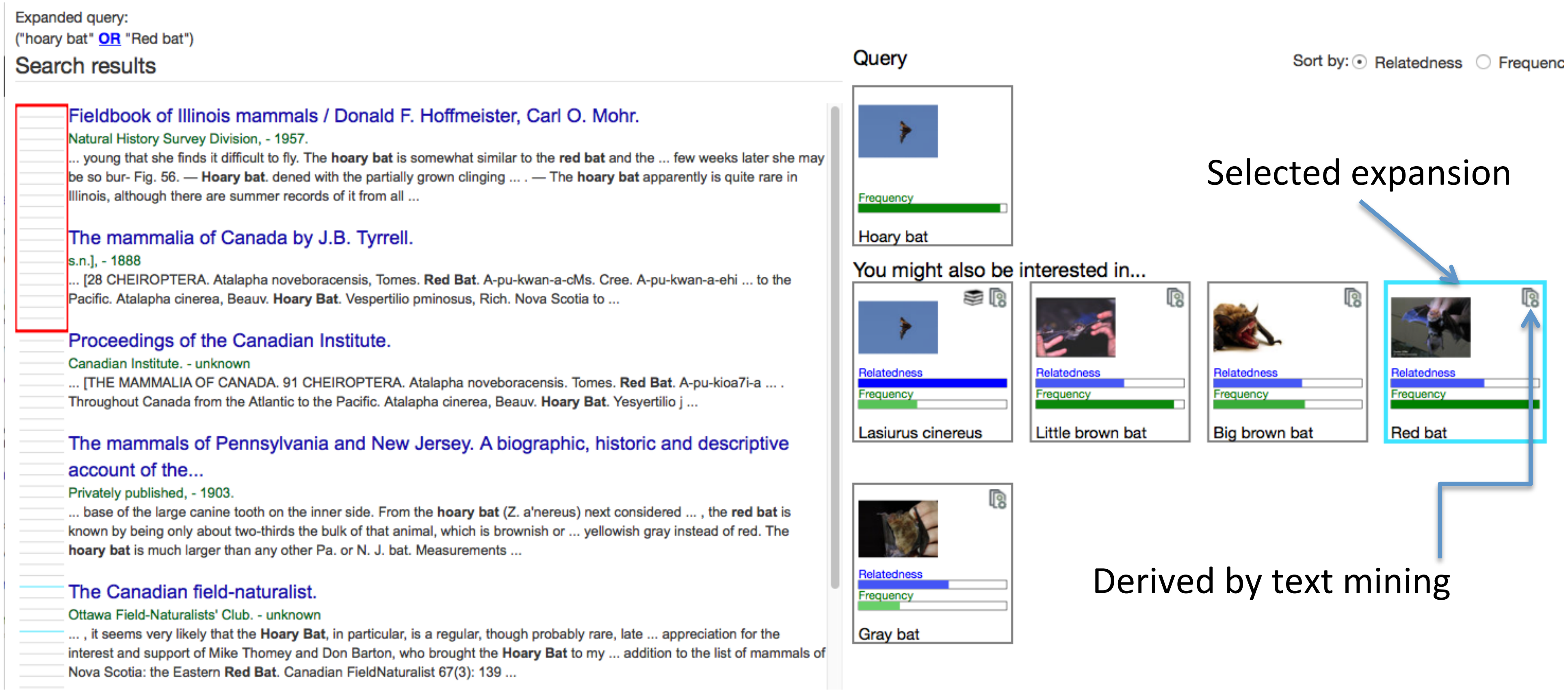
1. **Molecular characterization and concerted evolution of two genes encoding RING-C2 type proteins in rice.**
Journal: Gene 15/08/2012;505(1):9-18
... RING (Really Interesting New Gene) finger proteins are believed to play a critical role ... canonical types, RING-H2 and RING-HC, have been well-characterized, the molecular functions of the ... modified types, particularly the RING-C2 types, remain elusive. We isolated two rice genes ...
PubMedID 22705026 - View Abstract
Date: 15/08/2012 Score: 9.350

Semantic query expansion for Biodiversity Heritage Library

- Query expansion via *terminological inventory*
  - Derived from existing resources
    - Catologue of Life, Encylopedia of Life, Global Biodiversity Information Facility (millions of species)
  - ***PLUS*** extraction of semantically related terms from 24M pages (English subset of BHL)
  - Taxonomically related scientific and vernacular names, *species sharing same family or habitat*
    - Use of distributional semantics methods (words in context)

Nguyen et al. (2017) Constructing a biodiversity terminological inventory. *PLOS ONE* 12(4): e0175277

BHL query expansion from “hoary bat”



Expanded query:
("hoary bat" OR "Red bat")

Search results

Fieldbook of Illinois mammals / Donald F. Hoffmeister, Carl O. Mohr.
Natural History Survey Division, - 1957.
... young that she finds it difficult to fly. The **hoary bat** is somewhat similar to the **red bat** and the ... few weeks later she may be so bur- Fig. 56. — **Hoary bat**. dened with the partially grown clinging ... . — The **hoary bat** apparently is quite rare in Illinois, although there are summer records of it from all ...

The mammalia of Canada by J.B. Tyrrell.
s.n.], - 1888
... [28 CHEIROPTERA. Atalapha noveboracensis, Tomes. **Red Bat**. A-pu-kwan-a-cMs. Cree. A-pu-kwan-a-ehi ... to the Pacific. Atalapha cinerea, Beauv. **Hoary Bat**. Vespertilio pminosus, Rich. Nova Scotia to ...

Proceedings of the Canadian Institute.
Canadian Institute. - unknown
... [THE MAMMALIA OF CANADA. 91 CHEIROPTERA. Atalapha noveboracensis. Tomes. **Red Bat**. A-pu-kioa7i-a ... . Throughout Canada from the Atlantic to the Pacific. Atalapha cinerea, Beauv. **Hoary Bat**. Yesyertilio j ...

The mammals of Pennsylvania and New Jersey. A biographic, historic and descriptive account of the...
Privately published, - 1903.
... base of the large canine tooth on the inner side. From the **hoary bat** (Z. a'nereus) next considered ... , the **red bat** is known by being only about two-thirds the bulk of that animal, which is brownish or ... yellowish gray instead of red. The **hoary bat** is much larger than any other Pa. or N. J. bat. Measurements ...

The Canadian field-naturalist.
Ottawa Field-Naturalists' Club. - unknown
... , it seems very likely that the **Hoary Bat**, in particular, is a regular, though probably rare, late ... appreciation for the interest and support of Mike Thomey and Don Barton, who brought the **Hoary Bat** to my ... addition to the list of mammals of Nova Scotia: the Eastern **Red Bat**. Canadian FieldNaturalist 67(3): 139 ...

Query

Sort by: ● Relatedness ○ Frequenc

Frequency
Hoary bat

You might also be interested in...

Relatedness
Frequency
Lasiurus cinereus

Relatedness
Frequency
Little brown bat

Relatedness
Frequency
Big brown bat

Relatedness
Frequency
Red bat

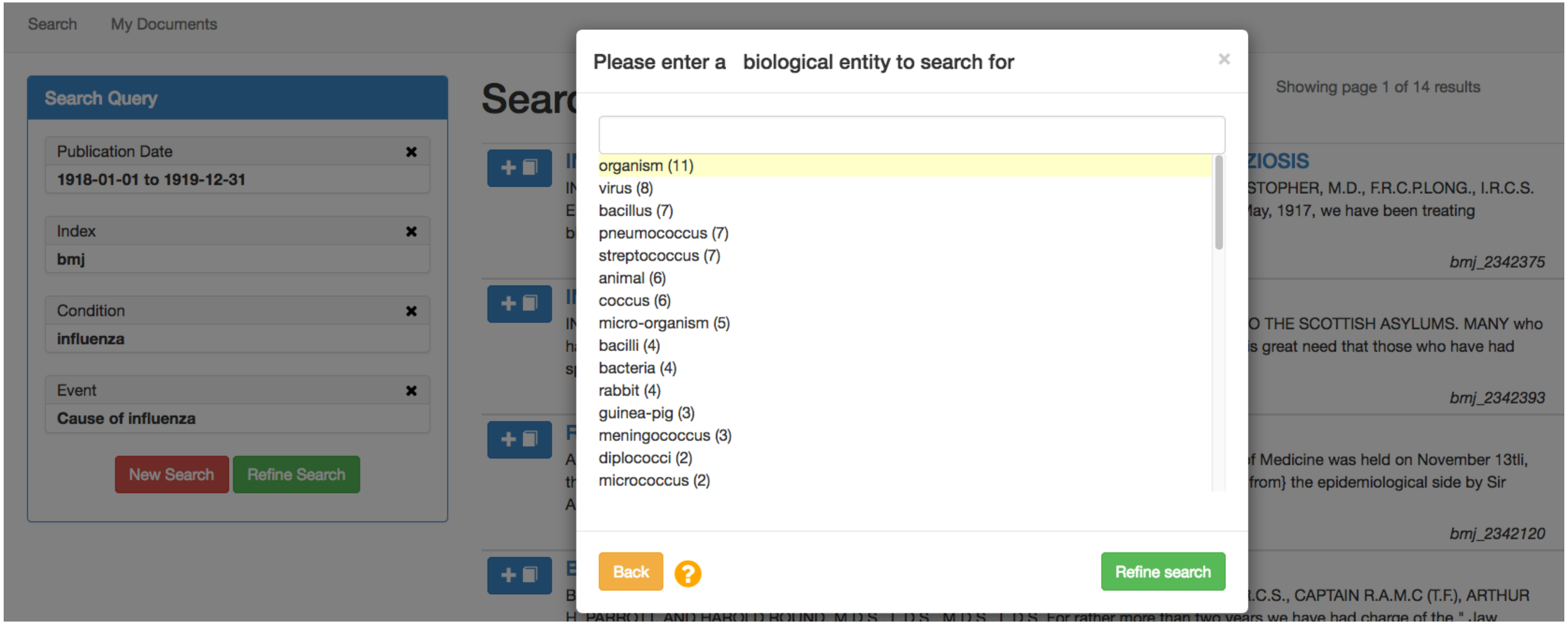
Relatedness
Frequency
Gray bat

Selected expansion

Derived by text mining

http://nactem.ac.uk/BHLQueryExpansion/

Semantic facets for History of Medicine



Search My Documents

Search Query

Publication Date ✖
1918-01-01 to 1919-12-31

Index ✖
bmj

Condition ✖
influenza

Event ✖
Cause of influenza

Please enter a biological entity to search for ×

- organism (11)
- virus (8)
- bacillus (7)
- pneumococcus (7)
- streptococcus (7)
- animal (6)
- coccus (6)
- micro-organism (5)
- bacilli (4)
- bacteria (4)
- rabbit (4)
- guinea-pig (3)
- meningococcus (3)
- diplococci (2)
- micrococcus (2)

Showing page 1 of 14 results

ZIOSIS
STOPHER, M.D., F.R.C.P.LONG., I.R.C.S.
lay, 1917, we have been treating
bmj_2342375

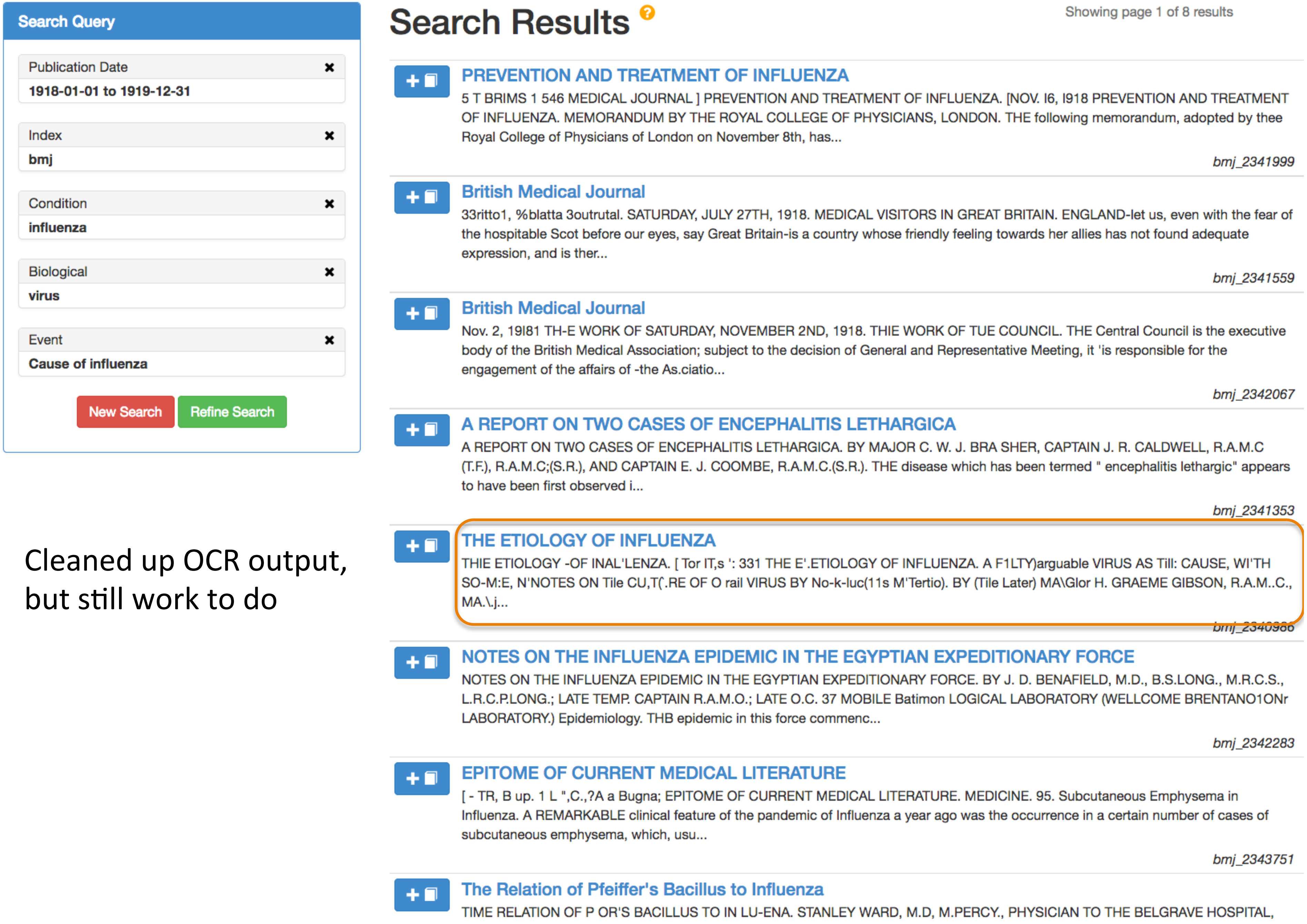
O THE SCOTTISH ASYLUMS. MANY who
is great need that those who have had
bmj_2342393

f Medicine was held on November 13tli,
from} the epidemiological side by Sir
bmj_2342120

.C.S., CAPTAIN R.A.M.C (T.F.), ARTHUR
H. PARROTT, AND HAROLD ROUND, M.D.S., L.D.S., M.D.S., L.D.S. For rather more than two years we have had charge of the " Jaw

What did doctors think was the cause of influenza at the time of the 1918 epidemic? Just beginning to see evidence of a “filterable virus” as opposed to until then accepted bacterium (-bacillus, -coccus), as seen from “virus (8)” in these 14 results

Thompson et al. (2016) Text mining the History of Medicine. *PLOS ONE*, 2016, 11(1): e0144717



Search Query

Publication Date ✖
1918-01-01 to 1919-12-31

Index ✖
bmj

Condition ✖
influenza

Biological ✖
virus

Event ✖
Cause of influenza

Search Results

Showing page 1 of 8 results

PREVENTION AND TREATMENT OF INFLUENZA

5 T BRIMS 1 546 MEDICAL JOURNAL] PREVENTION AND TREATMENT OF INFLUENZA. [NOV. I6, I918 PREVENTION AND TREATMENT OF INFLUENZA. MEMORANDUM BY THE ROYAL COLLEGE OF PHYSICIANS, LONDON. THE following memorandum, adopted by thee Royal College of Physicians of London on November 8th, has...

bmj_2341999

British Medical Journal

33ritto1, %blatta 3outrutal. SATURDAY, JULY 27TH, 1918. MEDICAL VISITORS IN GREAT BRITAIN. ENGLAND-let us, even with the fear of the hospitable Scot before our eyes, say Great Britain-is a country whose friendly feeling towards her allies has not found adequate expression, and is ther...

bmj_2341559

British Medical Journal

Nov. 2, 19I81 TH-E WORK OF SATURDAY, NOVEMBER 2ND, 1918. THIE WORK OF TUE COUNCIL. THE Central Council is the executive body of the British Medical Association; subject to the decision of General and Representative Meeting, it 'is responsible for the engagement of the affairs of -the As.ciatio...

bmj_2342067

A REPORT ON TWO CASES OF ENCEPHALITIS LETHARGICA

A REPORT ON TWO CASES OF ENCEPHALITIS LETHARGICA. BY MAJOR C. W. J. BRA SHER, CAPTAIN J. R. CALDWELL, R.A.M.C (T.F.), R.A.M.C;(S.R.), AND CAPTAIN E. J. COOMBE, R.A.M.C.(S.R.). THE disease which has been termed " encephalitis lethargic" appears to have been first observed i...

bmj_2341353

THE ETIOLOGY OF INFLUENZA

THIE ETIOLOGY -OF INAL'LENZA. [Tor IT,s ': 331 THE E'.ETIOLOGY OF INFLUENZA. A F1LTY)arguable VIRUS AS Till: CAUSE, WI'TH SO-M:E, N'NOTES ON Tile CU,T(.RE OF O rail VIRUS BY No-k-luc(11s M'Tertio). BY (Tile Later) MA\Glor H. GRAEME GIBSON, R.A.M..C., MA.\.j...

bmj_2340986

NOTES ON THE INFLUENZA EPIDEMIC IN THE EGYPTIAN EXPEDITIONARY FORCE

NOTES ON THE INFLUENZA EPIDEMIC IN THE EGYPTIAN EXPEDITIONARY FORCE. BY J. D. BENAFIELD, M.D., B.S.LONG., M.R.C.S., L.R.C.P.LONG.; LATE TEMP. CAPTAIN R.A.M.O.; LATE O.C. 37 MOBILE Batimon LOGICAL LABORATORY (WELLCOME BRENTANO1ONr LABORATORY.) Epidemiology. THB epidemic in this force commenc...

bmj_2342283

EPITOME OF CURRENT MEDICAL LITERATURE

[- TR, B up. 1 L ",C.,?A a Bugna; EPITOME OF CURRENT MEDICAL LITERATURE. MEDICINE. 95. Subcutaneous Emphysema in Influenza. A REMARKABLE clinical feature of the pandemic of Influenza a year ago was the occurrence in a certain number of cases of subcutaneous emphysema, which, usu...

bmj_2343751

The Relation of Pfeiffer's Bacillus to Influenza

TIME RELATION OF P OR'S BACILLUS TO IN LU-ENA. STANLEY WARD, M.D, M.PERCY., PHYSICIAN TO THE BELGRAVE HOSPITAL,

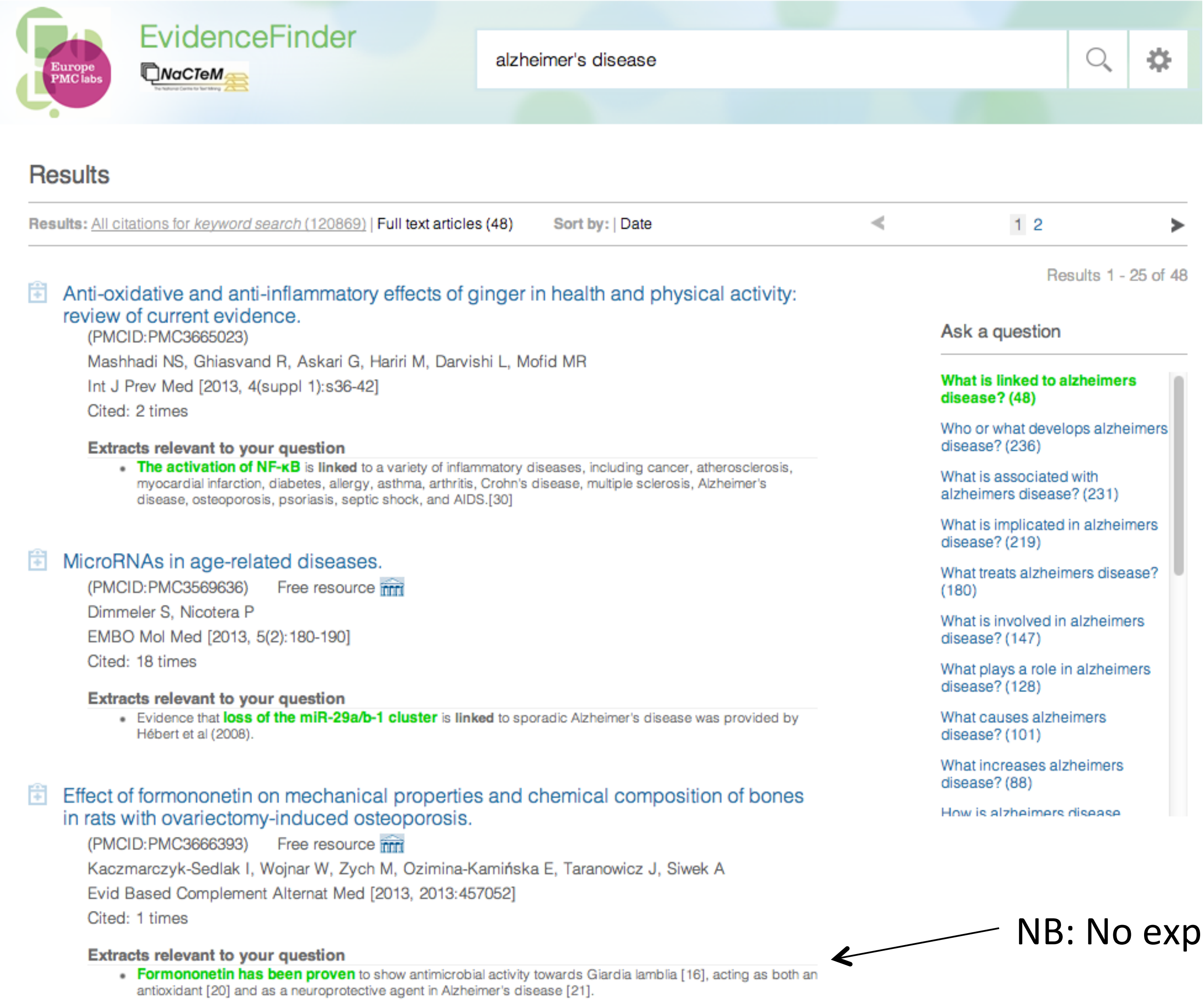
Cleaned up OCR output, but still work to do

Semantic search over events

- Entities participate in events
  - Identify and extract events
  - Index events for semantic search
- On user input, generate questions derived from stored events
  - Questions that have *known answers*
- EvidenceFinder for Europe PubMed Central Labs (>3M full text documents, >40M events)

Black et al. (2016) Text mining for semantic search in Europe PubMed Central Labs. In: Tonkin & Tourte (eds) *Working with Text: Tools, Techniques and Approaches for Text Mining*. Chandos. Pages 111−132

Semantic search: *known answers* for "what is linked to AD"



alzheimer's disease

Results

Results: All citations for *keyword search* (120869) | Full text articles (48) **Sort by:** | Date 1 2

Results 1 - 25 of 48

Anti-oxidative and anti-inflammatory effects of ginger in health and physical activity: review of current evidence.

(PMCID:PMC3665023)

Mashhadi NS, Ghiasvand R, Askari G, Hariri M, Darvishi L, Mofid MR

Int J Prev Med [2013, 4(suppl 1):s36-42]

Cited: 2 times

Extracts relevant to your question

- **The activation of NF-κB** is **linked** to a variety of inflammatory diseases, including cancer, atherosclerosis, myocardial infarction, diabetes, allergy, asthma, arthritis, Crohn's disease, multiple sclerosis, Alzheimer's disease, osteoporosis, psoriasis, septic shock, and AIDS.[30]

MicroRNAs in age-related diseases.

(PMCID:PMC3569636) Free resource

Dimmeler S, Nicotera P

EMBO Mol Med [2013, 5(2):180-190]

Cited: 18 times

Extracts relevant to your question

- Evidence that **loss of the miR-29a/b-1 cluster** is **linked** to sporadic Alzheimer's disease was provided by Hébert et al (2008).

Effect of formononetin on mechanical properties and chemical composition of bones in rats with ovariectomy-induced osteoporosis.

(PMCID:PMC3666393) Free resource

Kaczmarczyk-Sedlak I, Wojnar W, Zych M, Ozimina-Kamińska E, Taranowicz J, Siwek A

Evid Based Complement Alternat Med [2013, 2013:457052]

Cited: 1 times

Extracts relevant to your question

- **Formononetin has been proven** to show antimicrobial activity towards Giardia lamblia [16], acting as both an antioxidant [20] and as a neuroprotective agent in Alzheimer's disease [21].

Ask a question

What is linked to alzheimers disease? (48)

What or who develops alzheimers disease? (236)

What is associated with alzheimers disease? (231)

What is implicated in alzheimers disease? (219)

What treats alzheimers disease? (180)

What is involved in alzheimers disease? (147)

What plays a role in alzheimers disease? (128)

What causes alzheimers disease? (101)

What increases alzheimers disease? (88)

How is alzheimers disease

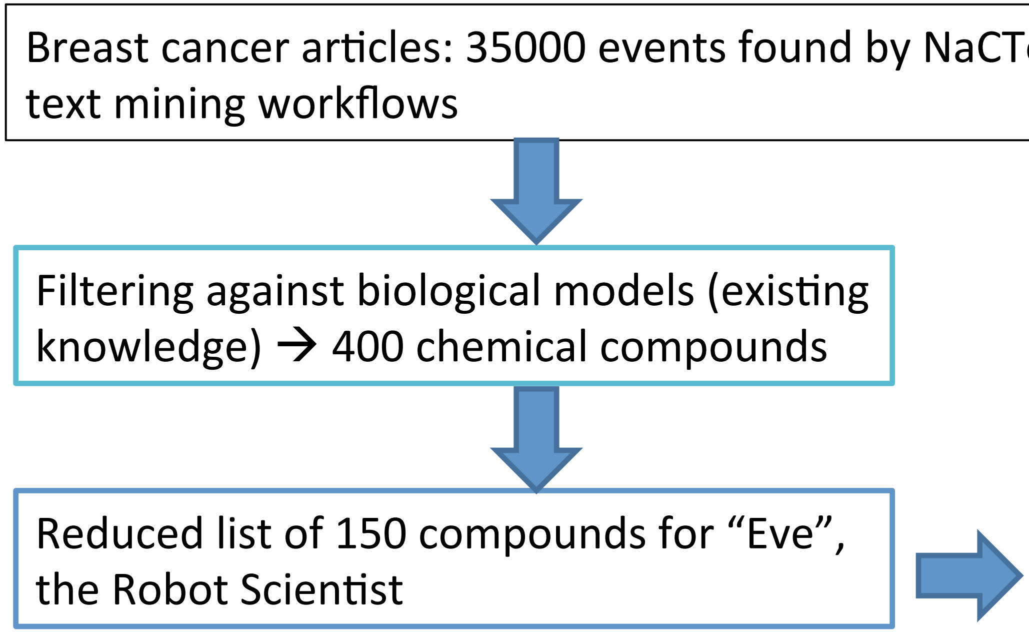
Nucleic Acids Research, 2015, 43

NB: No explicit 'link'

There’s more to (TM) life than search

- Much text mining not directly connected to user-facing search
- Curated databases: consistency checking, complementing human efforts
- Linking documents in to world of data
- ‘Feeding’ other processes
- Example: Automatic execution of experiments
  - Hypothesis generation via text mining
  - Validation by “robot scientist”
  - A sub-project of DARPA’s Big Mechanism (Cancer)

TM → filtering → Robot Scientist



Breast cancer articles: 35000 events found by NaCTeM's text mining workflows

Filtering against biological models (existing knowledge) → 400 chemical compounds

Reduced list of 150 compounds for "Eve", the Robot Scientist

"Eve"

Courtesy of R. King

Experiments to find
"Which compounds control expression of ESR1?"
(key gene for breast cancer)

Rzhetsky, A. (2016) The Big Mechanism program: Changing how science is done. Procs. XVIII Int. Conf. Data Analytics and Management in Data Intensive Domains. Pp. 1–2. http://ceur-ws.org/Vol-1752/paper01.pdf

Results (in part)

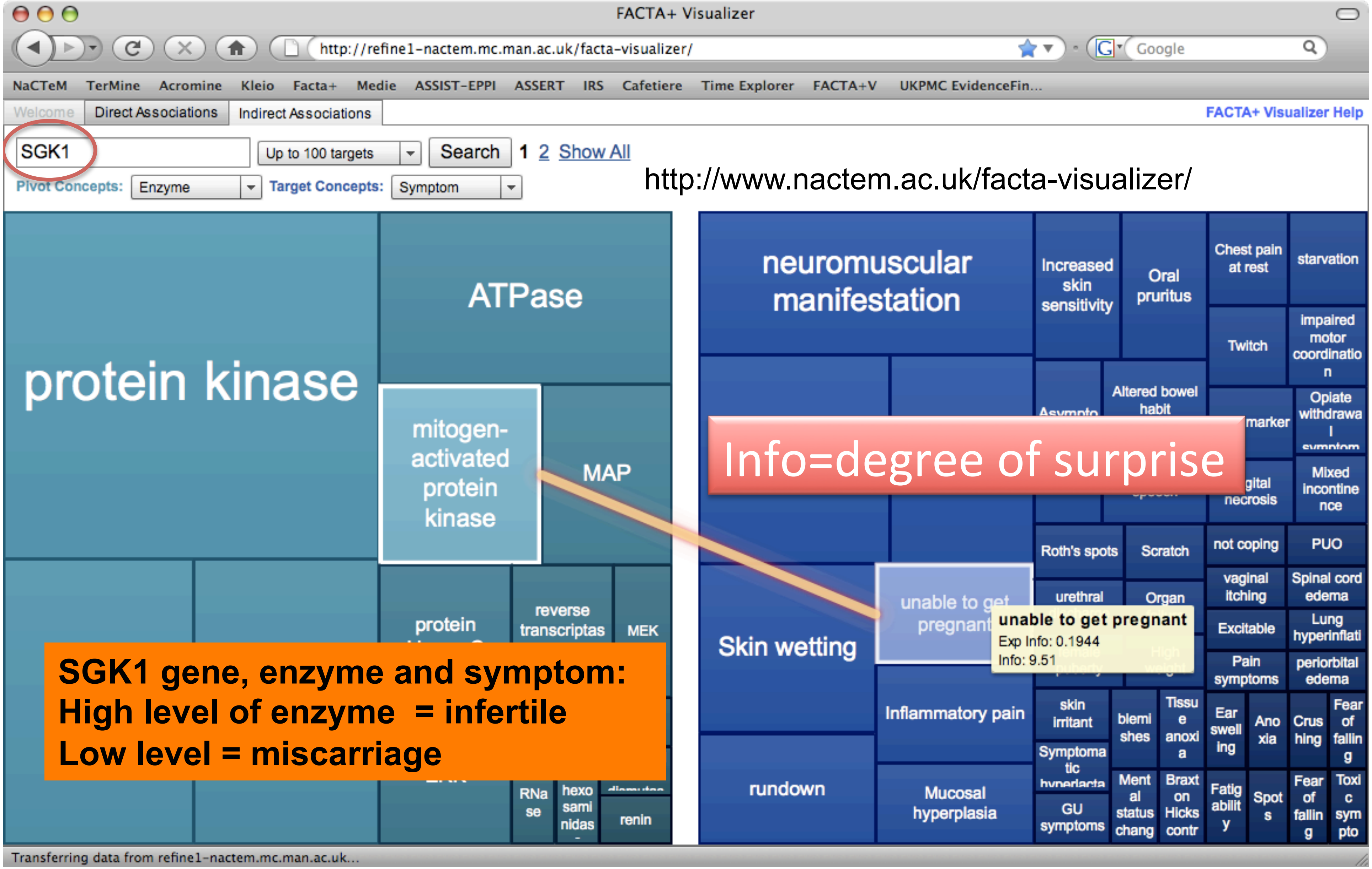
Compound	Experiment	TM prediction	Corresponds?
curcumin	Decrease	Negative Regulation	Yes
EGCG	Lethal at 10uM (?)	Positive Regulation (upregulated)	?
EGCG	Lethal at 10uM (?)	Positive Regulation(reactivate)	?
EGCG	Lethal at 10uM (?)	Regulation	?
fulvestrant	Decrease	Negative_regulation (supressed)	Yes
fulvestrant	Decrease	Negative_regulation (supressed)	Yes
oh-tam	Lethal at 10uM (?)	Negative_regulation (repressed)	?
oh-tam	Lethal at 10uM (?)	Positive_regulation (Activated)	?
PEITC	Lethal at 10uM (?)	Regulation	?
pterostilbene	Decrease	Negative_regulation (inhibits)	Yes
quercetin	Increase	Negative_regulation (decrease)	No
quercetin	Increase	Positive__regulation (increase)	Yes
resveratrol	Decrease	Positive_regulation (increasing)	No

(cellular death)

FACTA+: Association finding

- Based on named entities and events
- Rank associations
- Finds direct associations (known)
- Finds indirect associations
    - Not known
    - *Not explicitly written down*
- Much knowledge lies unsuspected in literature for many years before discovered via experiment

Reproducing a discovery – 11/**2011** in *Nature Medicine* – with FACTA+ running over MEDLINE **2009**



SGK1 | Up to 100 targets

Pivot Concepts: Enzyme | Target Concepts: Symptom

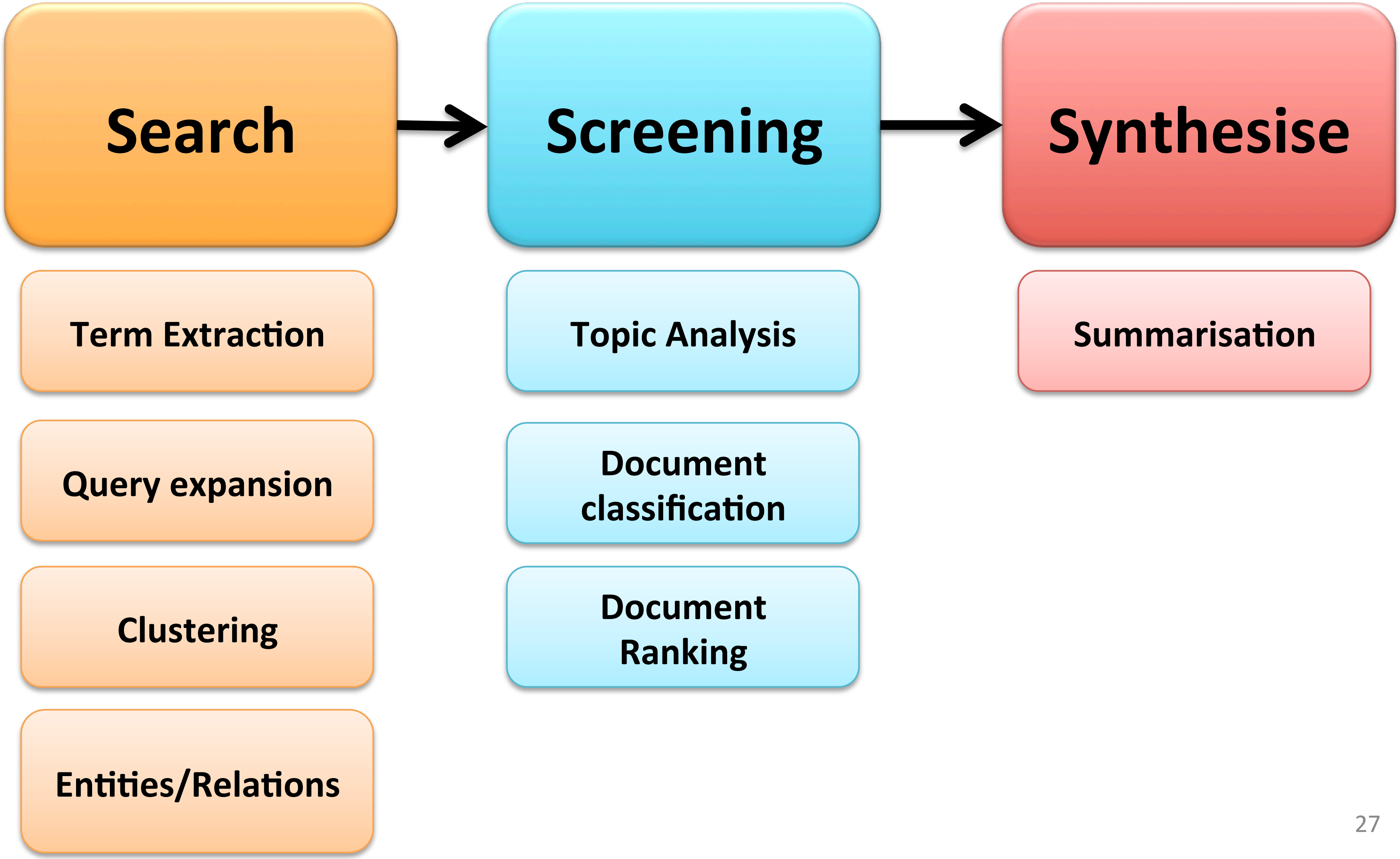
unable to get pregnant
Exp Info: 0.1944
Info: 9.51

http://www.nactem.ac.uk/facta-visualizer/

Info=degree of surprise

SGK1 gene, enzyme and symptom:
High level of enzyme = infertile
Low level = miscarriage

Text Mining Methods for Systematic Reviews



Search → Screening → Synthesise

Search
- Term Extraction
- Query expansion
- Clustering
- Entities/Relations

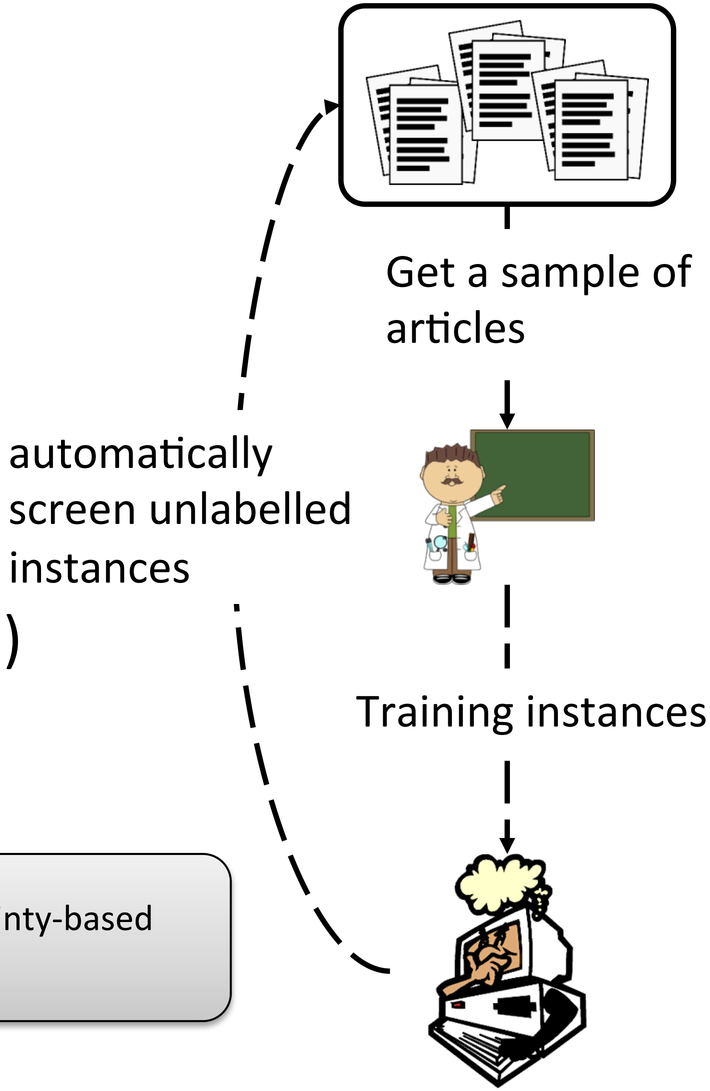
Screening
- Topic Analysis
- Document classification
- Document Ranking

Synthesise
- Summarisation

Supporting screening via document classification

Active learning

- Manually annotated samples used to train machine
- Algorithm learns from corrections, re-trains models, incrementally improves performance
- Analyst screens a subset of citations (direct reduction of screening workload)



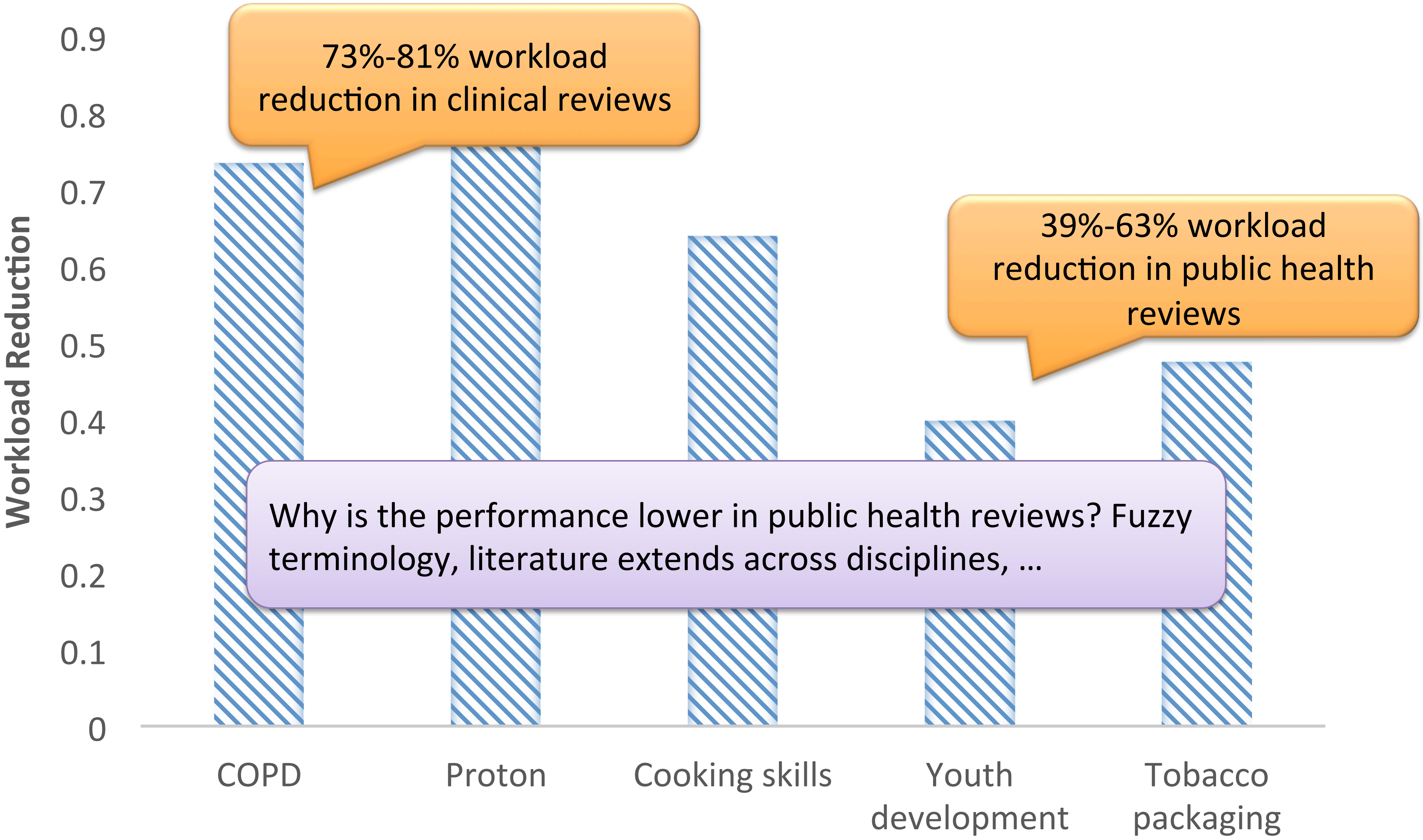
Get a sample of articles

automatically screen unlabelled instances

Training instances

Miwa et al. (2014) Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 51, 242–253.

Screening workload reduction (clinical and public health reviews)



73%-81% workload reduction in clinical reviews

39%-63% workload reduction in public health reviews

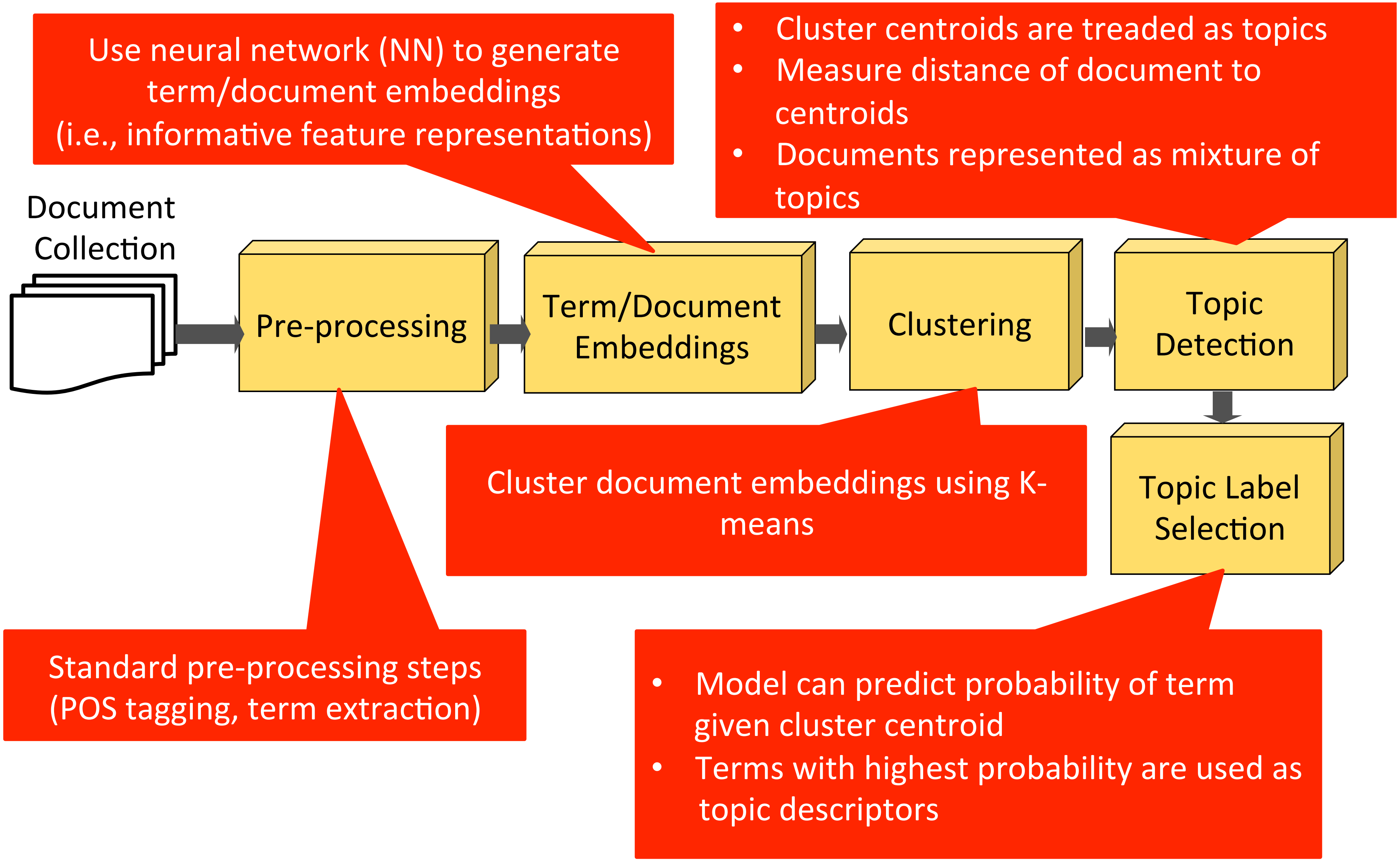
Why is the performance lower in public health reviews? Fuzzy terminology, literature extends across disciplines, …

Improving screening via topic modelling

- A new topic detection method that uses **deep learning** representation techniques
- Documents are represented as mixtures of topics
- Method takes into consideration semantics of words and documents
- Better captures terminological variation in public health reviews

Hashimoto et al. (2016) Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of Biomedical Informatics* 62, 59–65.

Architecture of a new topic detection model



Document Collection

Pre-processing

Term/Document Embeddings

Clustering

Topic Detection

Topic Label Selection

Use neural network (NN) to generate term/document embeddings (i.e., informative feature representations)

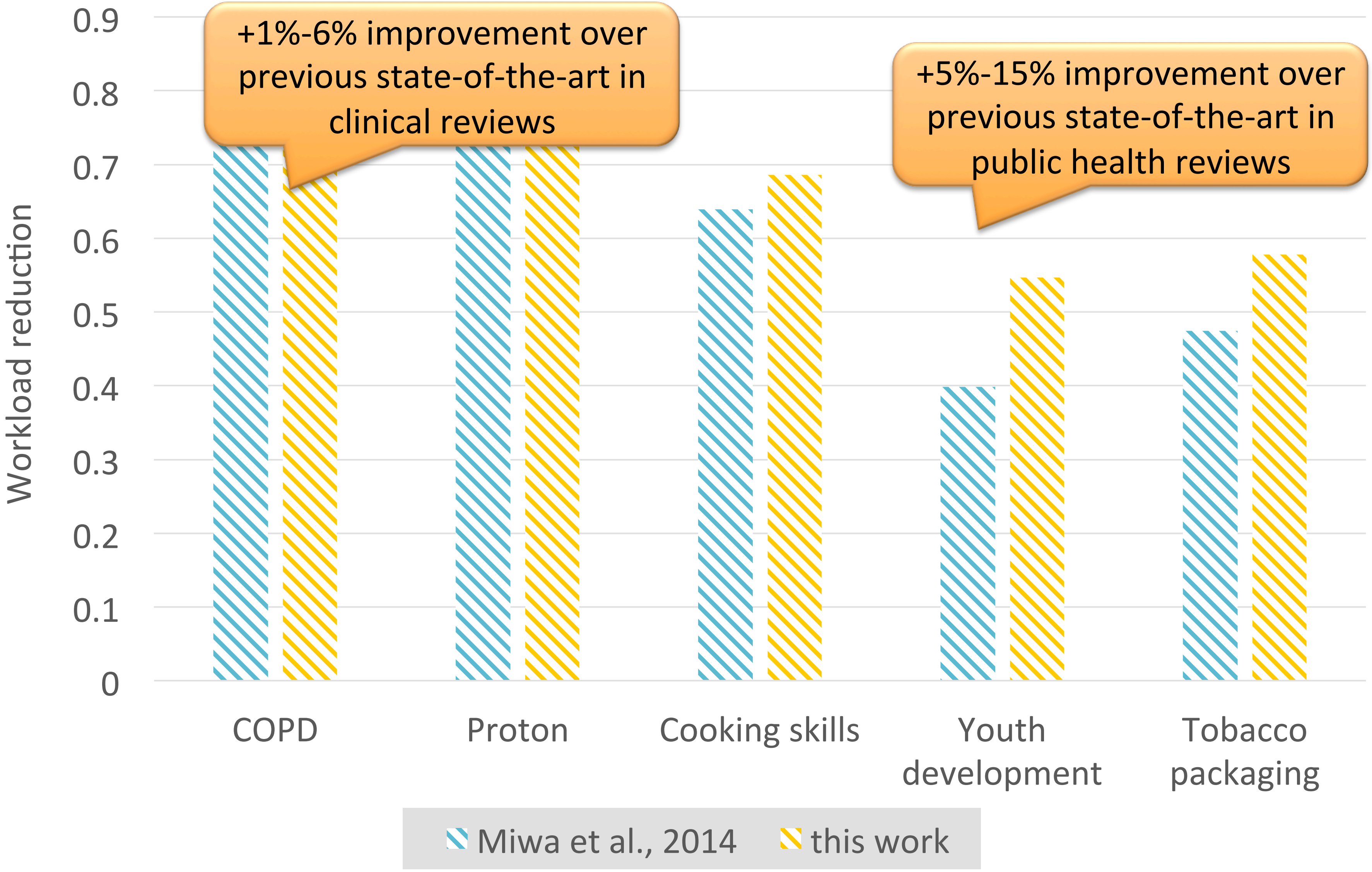
- Cluster centroids are treaded as topics
- Measure distance of document to centroids
- Documents represented as mixture of topics

Standard pre-processing steps (POS tagging, term extraction)

Cluster document embeddings using K-means

- Model can predict probability of term given cluster centroid
- Terms with highest probability are used as topic descriptors

Improving screening via topic modelling



+1%-6% improvement over previous state-of-the-art in clinical reviews

+5%-15% improvement over previous state-of-the-art in public health reviews

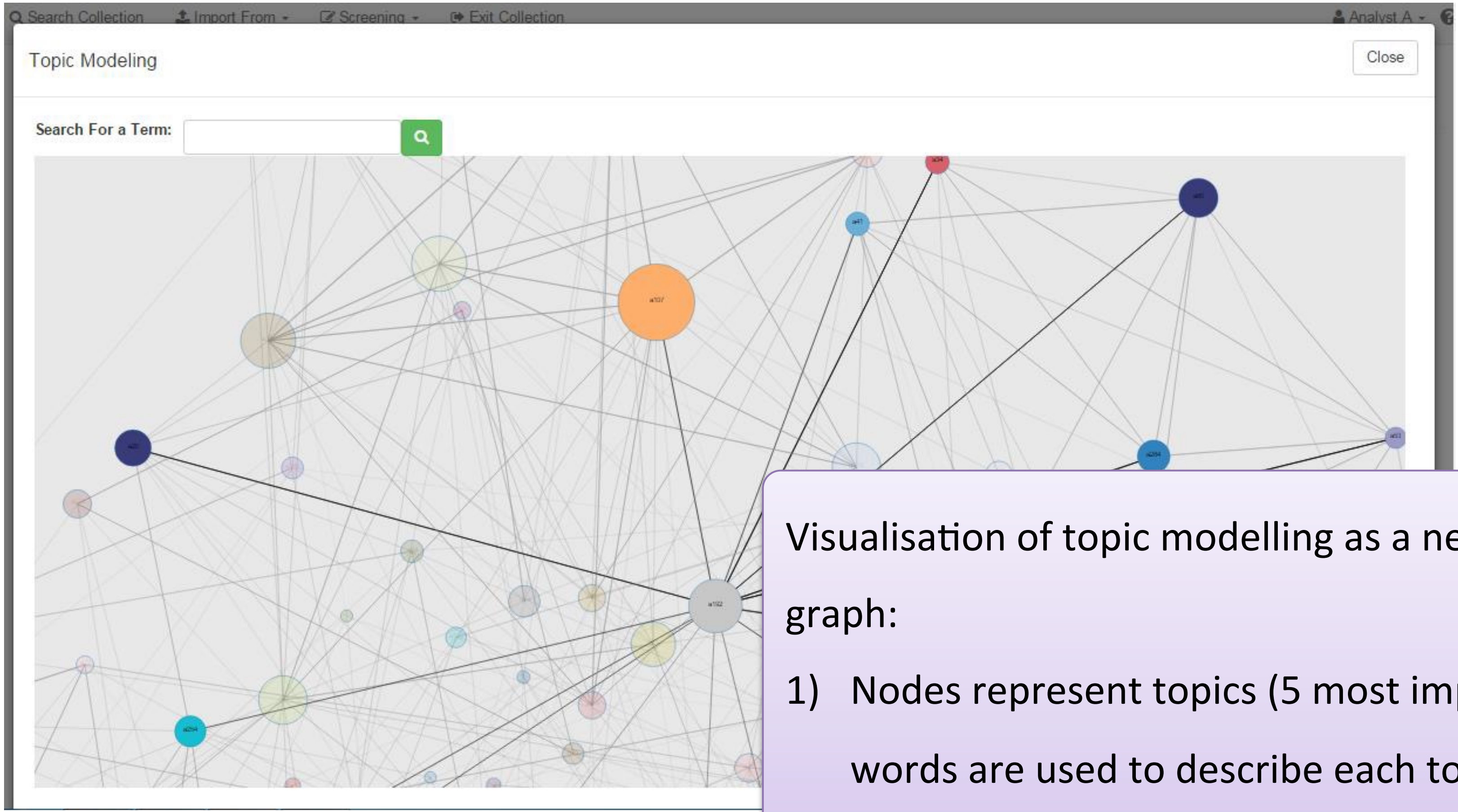
RobotAnalyst

- End-to-end reference management system enhanced with TM functionalities
  - Facetted search
  - Topic-based search
  - Semi-automatic citation screening
- Initially developed as evaluation workbench of TM methods for systematic reviews
- Currently used by public health analysts at NICE for screening

Latest project publication:
Kontonatsios et al. (2017). A semi-supervised approach using label propagation to support citation screening. *Journal of Biomedical Informatics* 72, 67−76.

http://www.nactem.ac.uk/robotanalyst/

Visualisation of topics

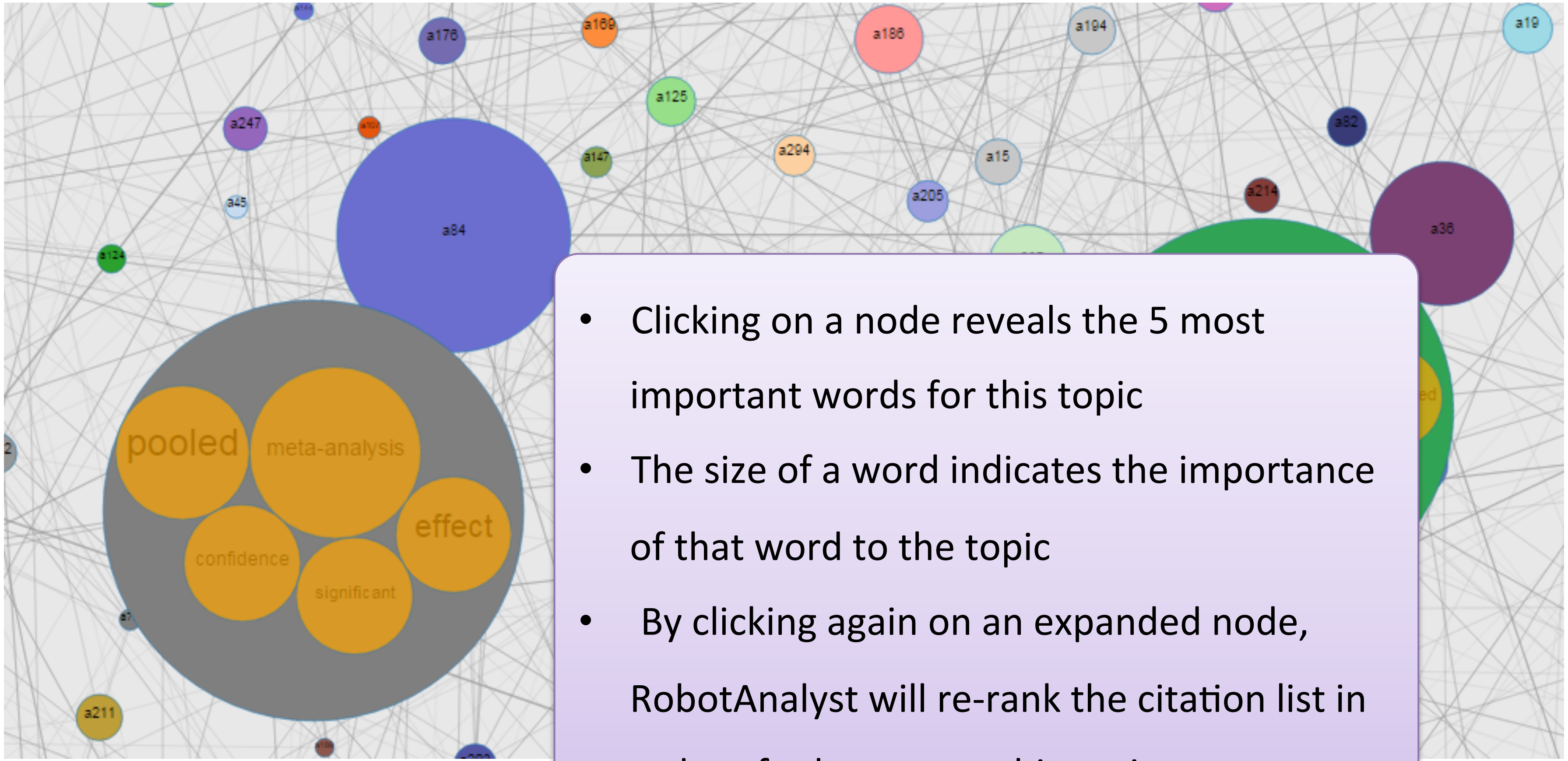


Visualisation of topic modelling as a network graph:

1) Nodes represent topics (5 most important words are used to describe each topic)

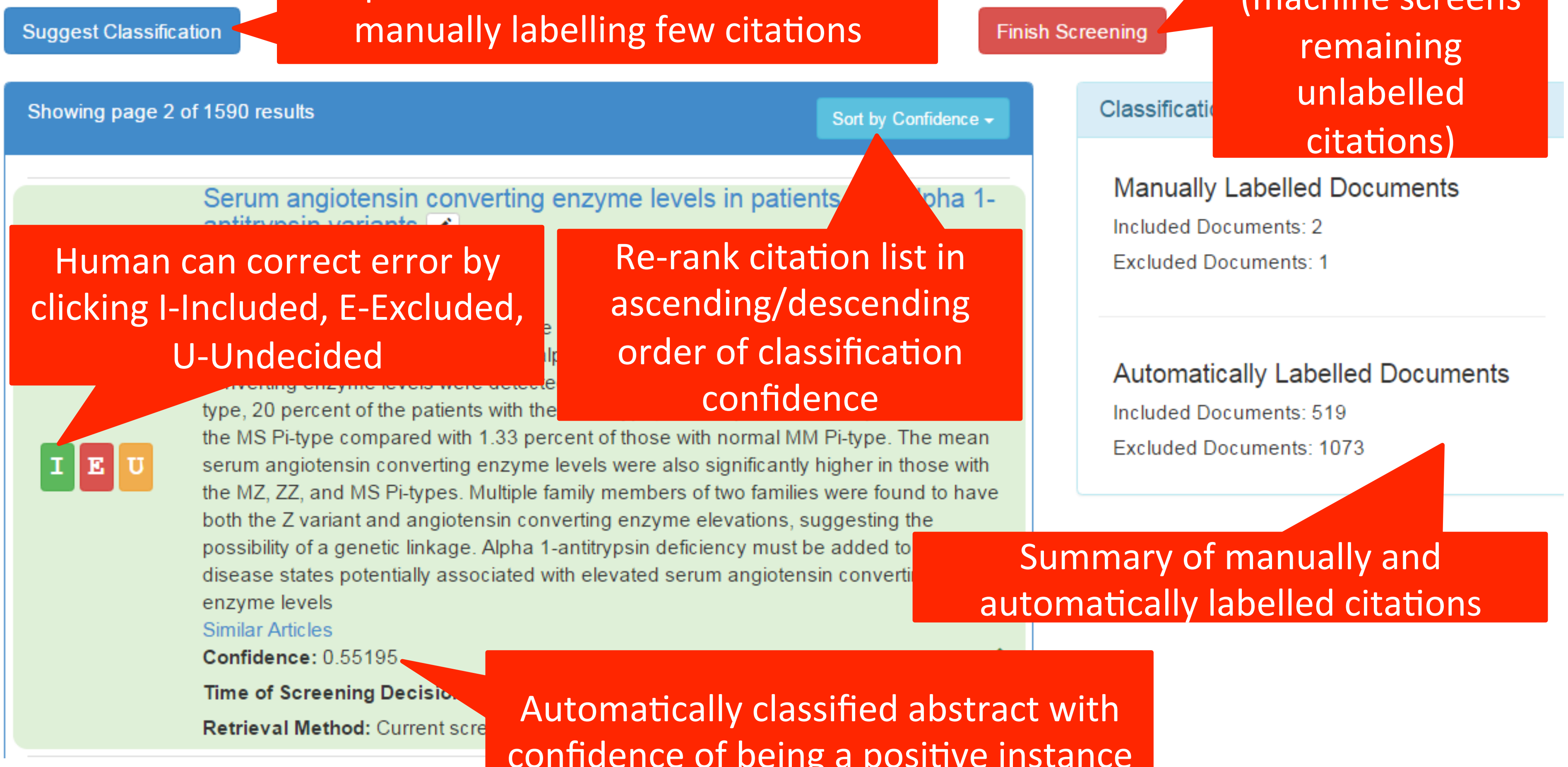
2) Edges show topic-to-topic correlation

Visualisation of topics



- Clicking on a node reveals the 5 most important words for this topic
- The size of a word indicates the importance of that word to the topic
- By clicking again on an expanded node, RobotAnalyst will re-rank the citation list in order of relevance to this topic

Semi-automatic citation screening



Suggest Classification

Finish Screening

Showing page 2 of 1590 results

Sort by Confidence

Serum angiotensin converting enzyme levels in patients … lpha 1-antitrypsin variants

type, 20 percent of the patients with the … the MS Pi-type compared with 1.33 percent of those with normal MM Pi-type. The mean serum angiotensin converting enzyme levels were also significantly higher in those with the MZ, ZZ, and MS Pi-types. Multiple family members of two families were found to have both the Z variant and angiotensin converting enzyme elevations, suggesting the possibility of a genetic linkage. Alpha 1-antitrypsin deficiency must be added to … disease states potentially associated with elevated serum angiotensin converti… enzyme levels

Similar Articles

Confidence: 0.55195

Time of Screening Decisio…

Retrieval Method: Current scre…

I E U

Classificati…

Manually Labelled Documents

Included Documents: 2

Excluded Documents: 1

Automatically Labelled Documents

Included Documents: 519

Excluded Documents: 1073

Update classification models after manually labelling few citations

Finish screening (machine screens remaining unlabelled citations)

Human can correct error by clicking I-Included, E-Excluded, U-Undecided

Re-rank citation list in ascending/descending order of classification confidence

Summary of manually and automatically labelled citations

Automatically classified abstract with confidence of being a positive instance

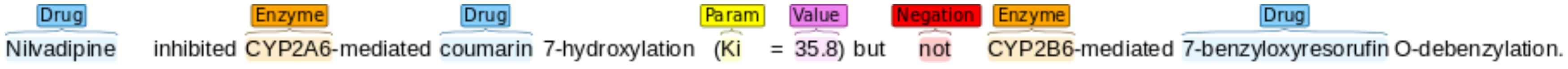
Infrastructure and interoperability

- TM is not monolithic
- Many levels of analysis
- Many tools and resources
  - *At each level*
- Adaptation according to requirements, domain, research question, text type, …
- How do we ensure interoperability of individual components and of infrastructures?

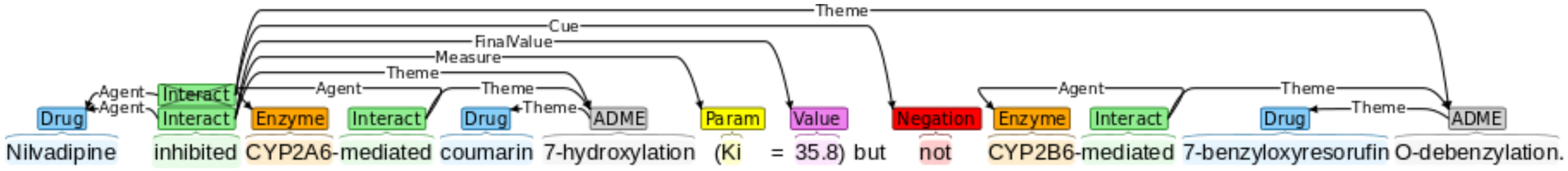
Example: 3 levels and processes

Nilvadipine inhibited CYP2A6-mediated coumarin 7-hydroxylation (Ki = 35.8) but not CYP2B6-mediated 7-benzyloxyresorufin O-debenzylation.

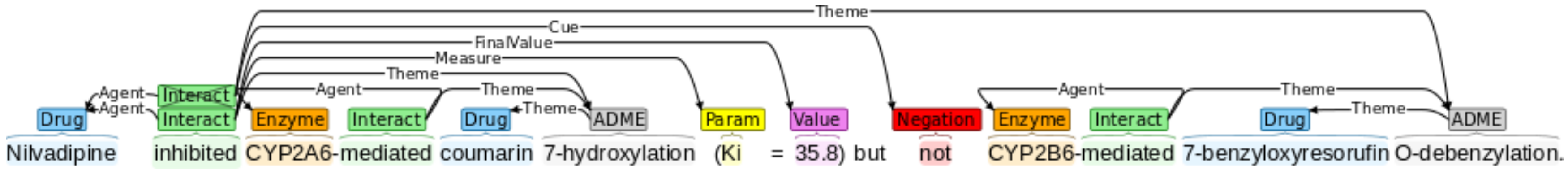
1 Named entity recognition


Drug
Enzyme
Drug
Param
Value
Negation
Enzyme
Drug
Nilvadipine inhibited CYP2A6-mediated coumarin 7-hydroxylation (Ki = 35.8) but not CYP2B6-mediated 7-benzyloxyresorufin O-debenzylation.

2 Event extraction



3 Identification of concepts and their interactions



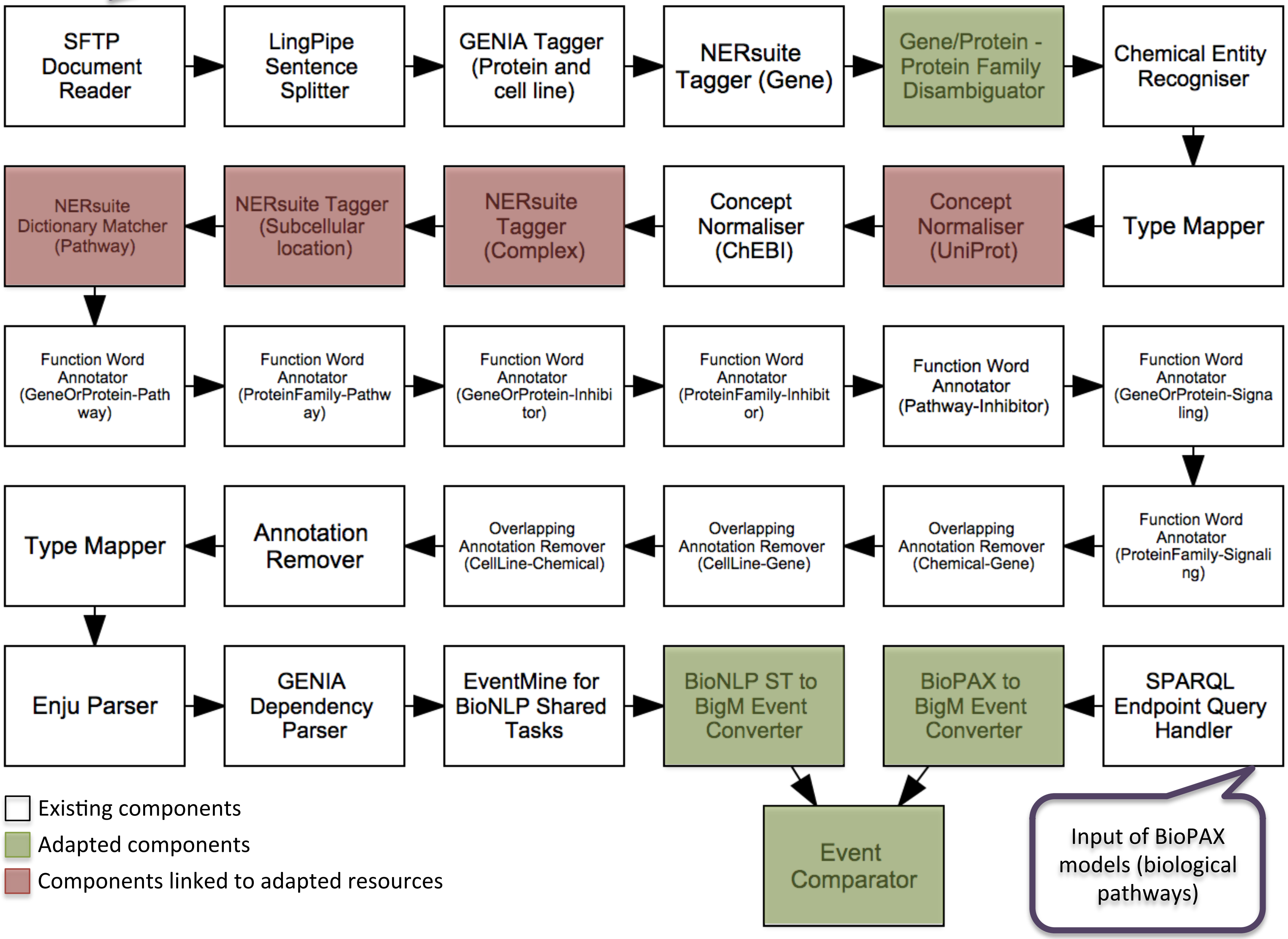
DrugBank:DB06712

DrugBank:DB00682

DrugBank:DB04610

One workflow for *one* sub-project within the DARPA Big Mechanism (Cancer) project

Input of documents



SFTP Document Reader

LingPipe Sentence Splitter

GENIA Tagger (Protein and cell line)

NERsuite Tagger (Gene)

Gene/Protein - Protein Family Disambiguator

Chemical Entity Recogniser

Type Mapper

Concept Normaliser (UniProt)

Concept Normaliser (ChEBI)

NERsuite Tagger (Complex)

NERsuite Tagger (Subcellular location)

NERsuite Dictionary Matcher (Pathway)

Function Word Annotator (GeneOrProtein-Pathway)

Function Word Annotator (ProteinFamily-Pathway)

Function Word Annotator (GeneOrProtein-Inhibitor)

Function Word Annotator (ProteinFamily-Inhibitor)

Function Word Annotator (Pathway-Inhibitor)

Function Word Annotator (GeneOrProtein-Signaling)

Function Word Annotator (ProteinFamily-Signaling)

Overlapping Annotation Remover (Chemical-Gene)

Overlapping Annotation Remover (CellLine-Gene)

Overlapping Annotation Remover (CellLine-Chemical)

Annotation Remover

Type Mapper

Enju Parser

GENIA Dependency Parser

EventMine for BioNLP Shared Tasks

BioNLP ST to BigM Event Converter

SPARQL Endpoint Query Handler

BioPAX to BigM Event Converter

Event Comparator

Input of BioPAX models (biological pathways)

□ Existing components

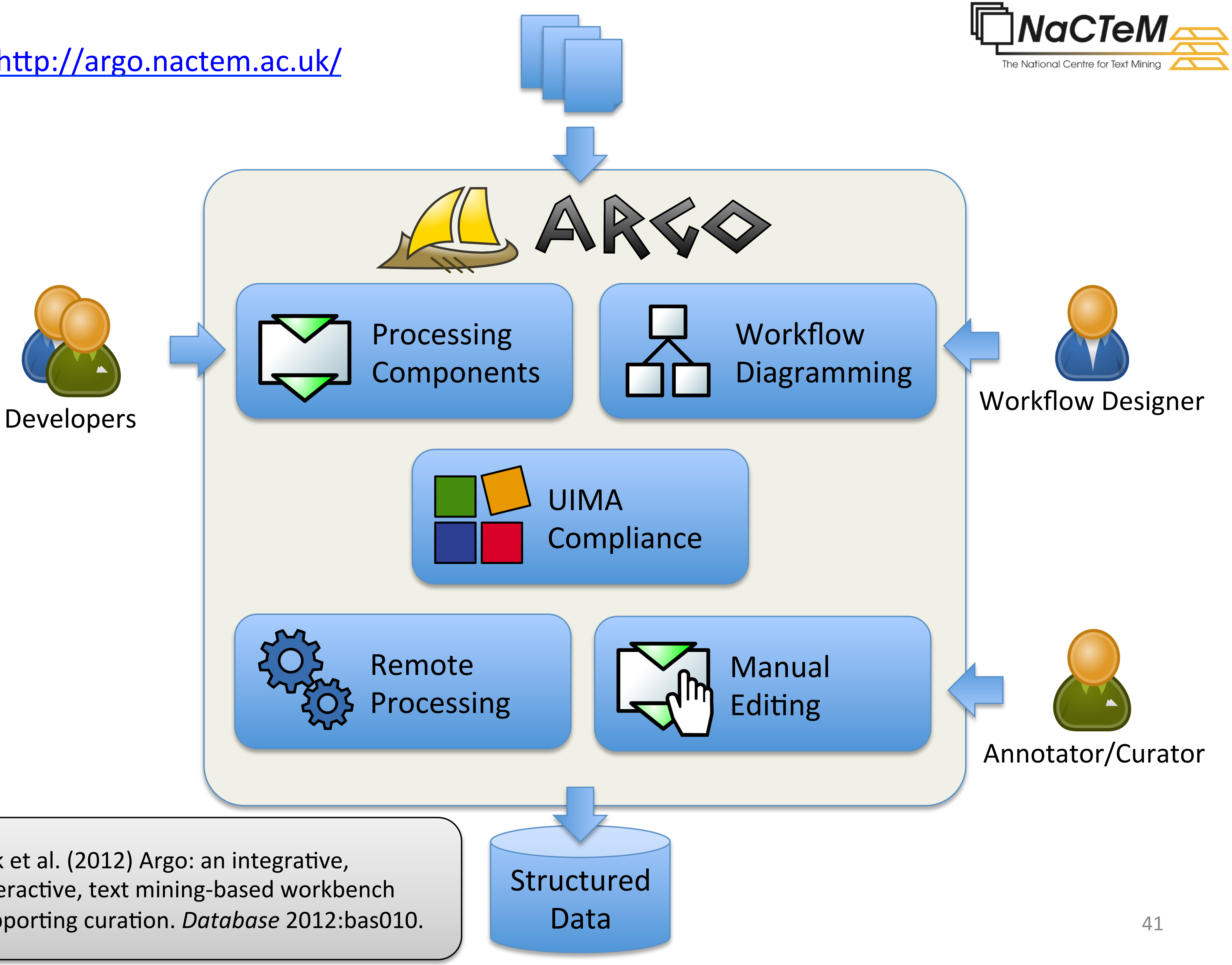
🟩 Adapted components

🟥 Components linked to adapted resources

Interoperability problems

- Needs of users in many domains
- Need to avoid dispersed, duplicated efforts
- Need to increase collaboration
- Need to evaluate components and workflows
- Need to assemble, network TM results
  - Links to public repositories
  - Persistence of results
- Need to ensure reproducible research
- Need to link several/many TM infrastructures

http://argo.nactem.ac.uk/



ARGO

Processing Components

Workflow Diagramming

UIMA Compliance

Remote Processing

Manual Editing

Developers

Workflow Designer

Annotator/Curator

Structured Data

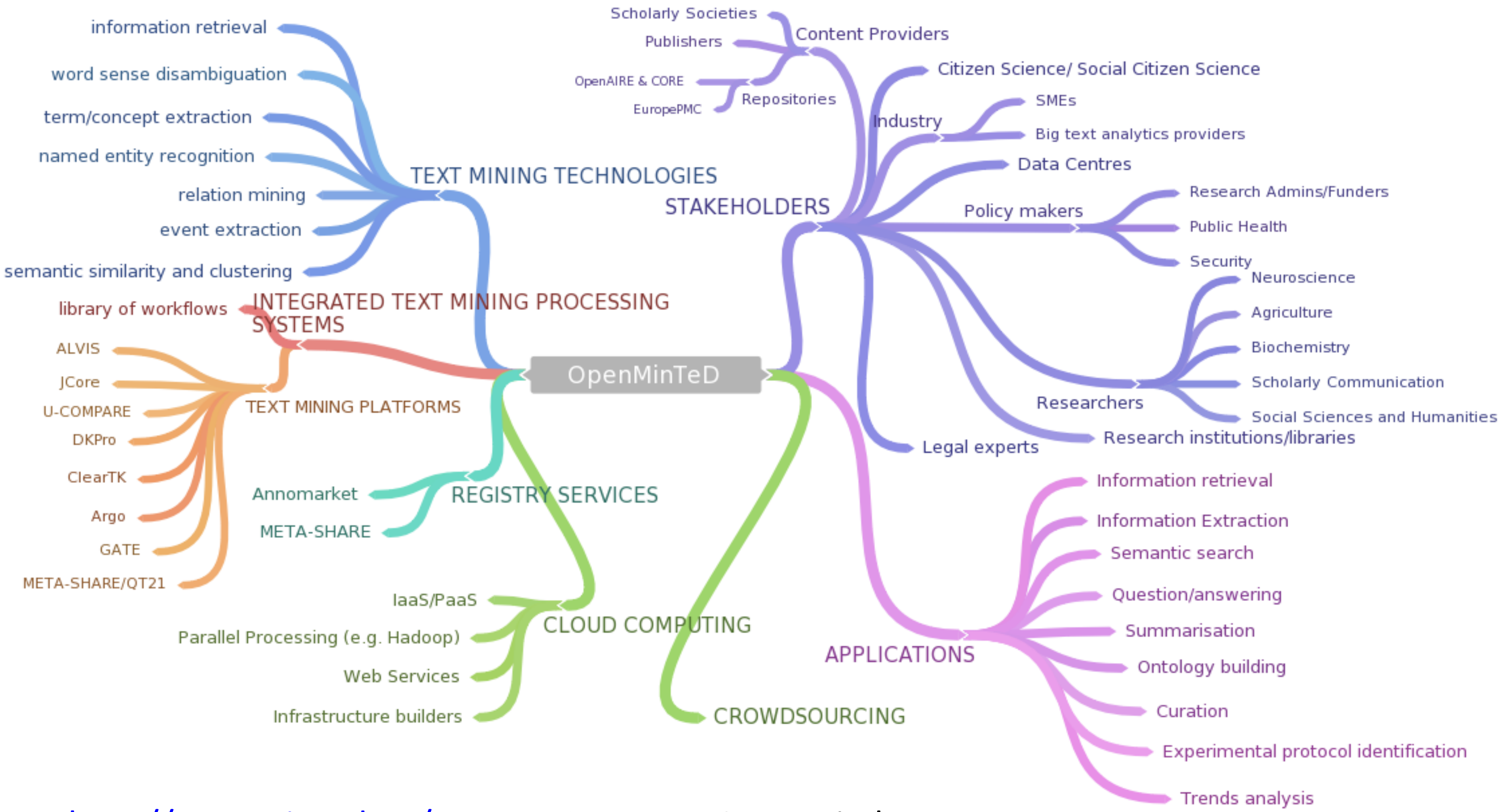
Rak et al. (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database* 2012:bas010.

Progress towards interoperability

- OpenMinTeD (H2020 project 2015–2018)
  - Open infrastructure to facilitate, encourage TM
- Leverages existing efforts including Argo
- Tool/resource discovery from registry
- Levels of interoperability
  - Services, platform, components, licences
  - Language resources
  - Sharing of storage and computational resources

NaCTeM
The National Centre for Text Mining

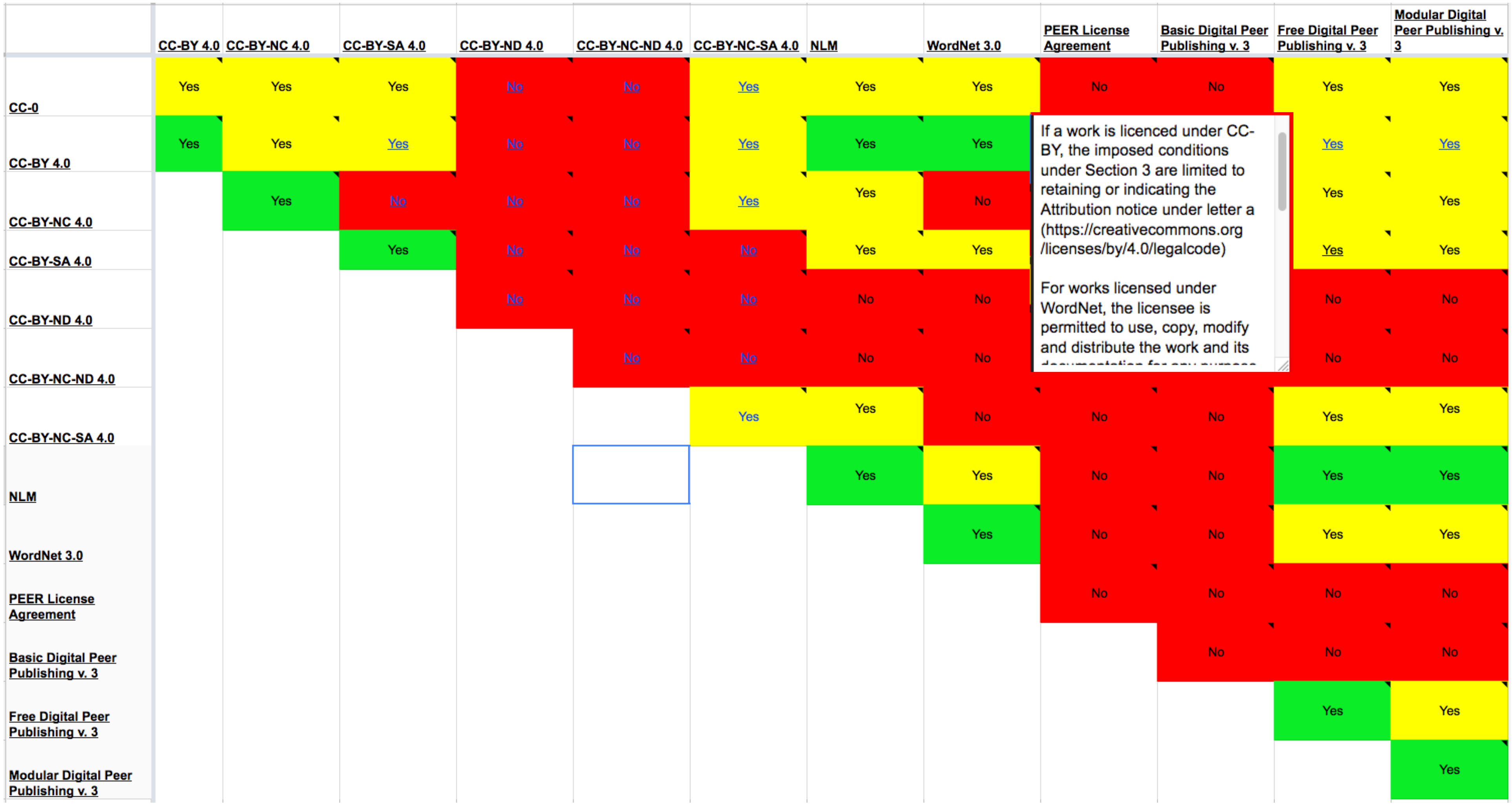
openMINTED
Open Mining Infrastructure for Text & Data



http://openminted.eu/ Courtesy: S. Ananiadou

Interoperability of licences
(here content, also software)

openMINTED
Open Mining Infrastructure for Text & Data



	CC-BY 4.0	CC-BY-NC 4.0	CC-BY-SA 4.0	CC-BY-ND 4.0	CC-BY-NC-ND 4.0	CC-BY-NC-SA 4.0	NLM	WordNet 3.0	PEER License Agreement	Basic Digital Peer Publishing v. 3	Free Digital Peer Publishing v. 3	Modular Digital Peer Publishing v. 3
CC-0	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	Yes	Yes
CC-BY 4.0	Yes	Yes	Yes	No	No	Yes	Yes	Yes			Yes	Yes
CC-BY-NC 4.0		Yes	No	No	No	Yes	Yes	No			Yes	Yes
CC-BY-SA 4.0			Yes	No	No	No	Yes	Yes			Yes	Yes
CC-BY-ND 4.0				No	No	No	No	No			No	No
CC-BY-NC-ND 4.0					No	No	No	No			No	No
CC-BY-NC-SA 4.0						Yes	Yes	No	No	No	Yes	Yes
NLM							Yes	Yes	No	No	Yes	Yes
WordNet 3.0								Yes	No	No	Yes	Yes
PEER License Agreement									No	No	No	No
Basic Digital Peer Publishing v. 3										No	No	No
Free Digital Peer Publishing v. 3											Yes	Yes
Modular Digital Peer Publishing v. 3												Yes

If a work is licenced under CC-BY, the imposed conditions under Section 3 are limited to retaining or indicating the Attribution notice under letter a (https://creativecommons.org/licenses/by/4.0/legalcode)

For works licensed under WordNet, the licensee is permitted to use, copy, modify and distribute the work and its documentation for any purpose

Gaps, issues

- Lack of people with training in TM
- Lack of tools and resources
- Copyright
  - Although somewhat improved due to UK copyright exception for non-commercial TDM
- Role of research libraries
  - Who is your TM champion?

Research data management

- Advanced TM produces formal (linguistic, semantic) annotations at numerous levels
- Massive amounts of rich research data
- Traditionally, individual researcher managed own TM data (and input content)
- Library could have key role in managing annotation store for Institution, linked to (subscribed) content, institutional repository
  - Foster inter/cross/trans-disciplinary research

Implications

- Open Annotation Store software for Europe PubMed Central, developed by NaCTeM
  - Test population of analysis of 150000 documents
  - Produced **30M** linguistic/semantic annotations
  - Represented in **10^9** RDF triples
- Consider! Many different analyses at same or different levels by different contributing researchers

Institutional risk?

- *N* researchers and students downloading (subscribed) content under TM copyright exception, multiple copies proliferating
  - Unacceptable risk of content leaks for Institution?
  - Central control? Researcher point of view?
  - Library manages access to (subscribed) content for TM purposes?

Take home points

- Research libraries are well-placed to support text mining in their institutions: readiness?
- Every domain can benefit from TM
  - Some more provided for than others
  - Differing levels of awareness of TM and of its possibilities among researchers
- Research data (annotations) produced during TM are a valuable resource for the institution
- The scope for new TM applications is enormous
  - Talk to us about joint research

Licence and credits

- Except where otherwise noted, this presentation is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence

  https://creativecommons.org/licenses/by/4.0/
- The University of Manchester and NaCTeM logos are copyright the University of Manchester
- The image of "Eve" is copyright R. King
- The OpenMinTeD logo is copyright the OpenMinTeD consortium
- The OpenMinTeD mind map is copyright S. Ananiadou