

Text Mining – Techniques & Limitations

View of an Industry Chemist

14.9.2013, EC-L4E-WG4



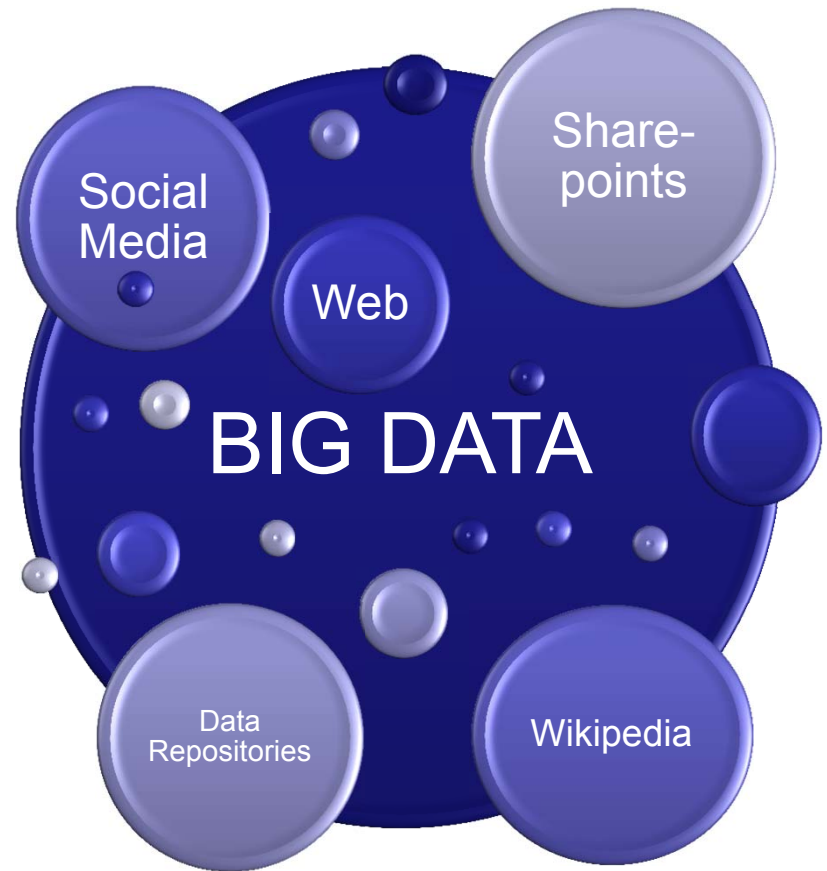
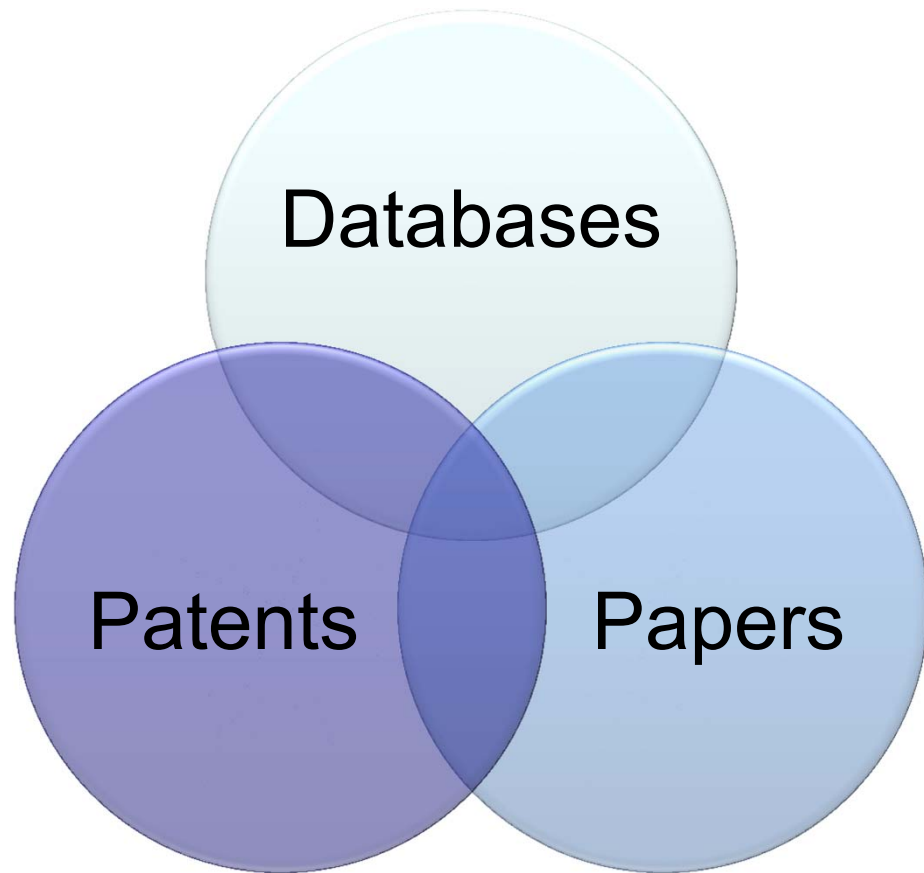
Frank Oellien
Secretary of EuCheMS DCC
Industrial Chemists/TDM Scientist



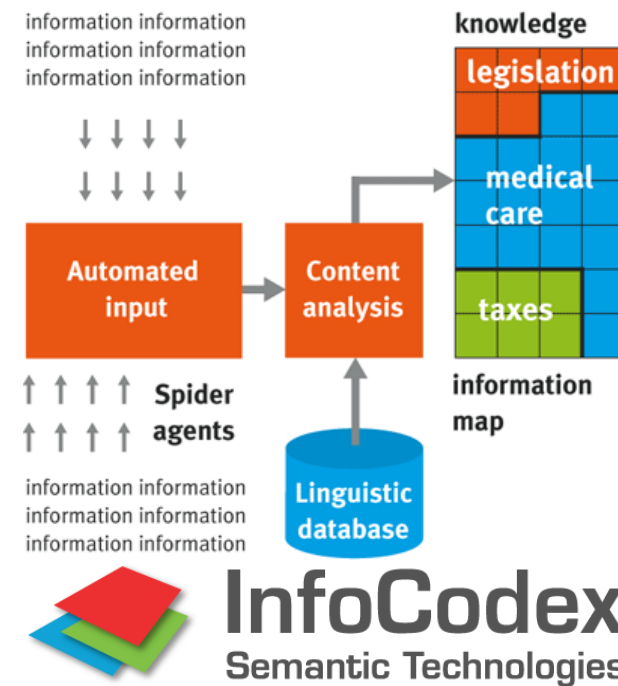
Overview

- Text Mining Sources
- Text Mining Techniques
 - Software Overview
 - Identification and Retrieval
 - Content Processing (Information to Knowledge)
 - Classification
 - Semantics/Ontology
 - Data Extraction (e.g. Chemical Objects)
 - Delivery and Presentation
- Current Limitations
- Workarounds
- (Ideal) Future Perspective

Text Mining Sources



Text Mining Techniques - Software

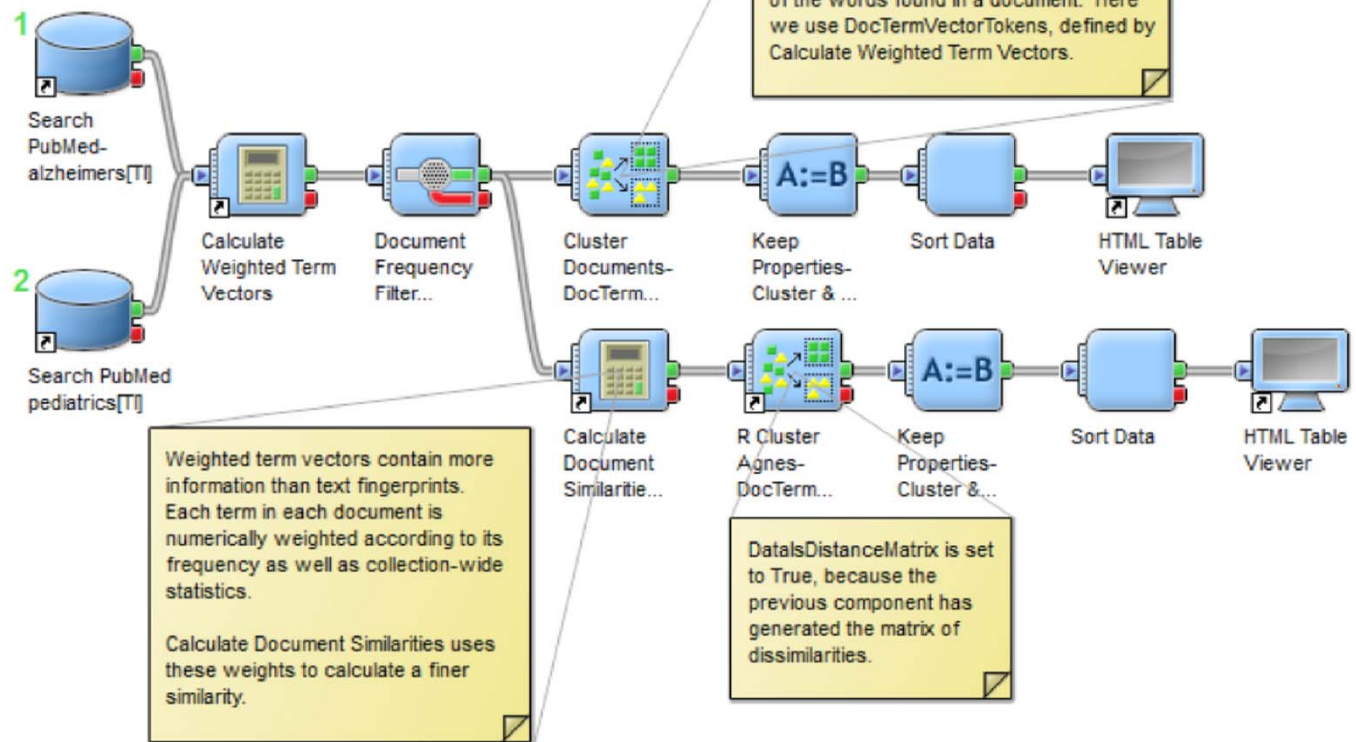


Text Mining Techniques - Software

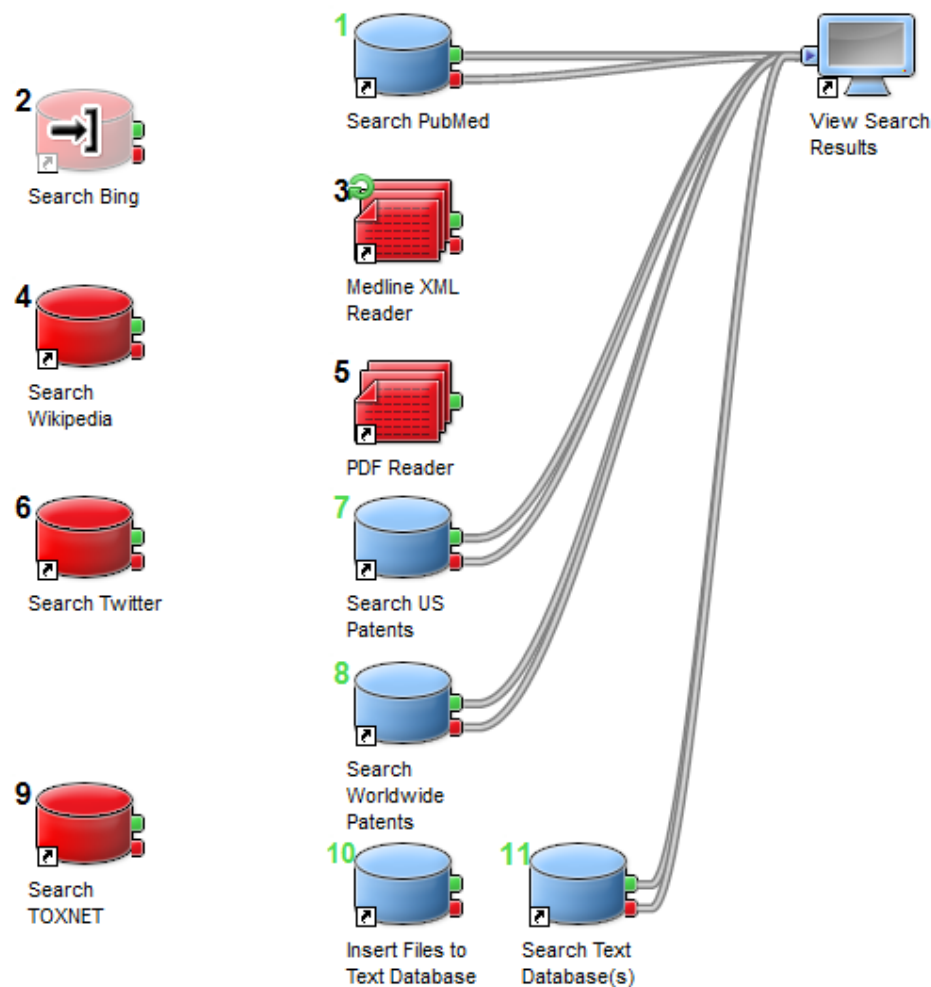


Use Accelrys Pipeline Pilot (Text Mining Collection) to introduce some basic concepts

Cluster Comparison Example -
Agnes Versus "Cluster Documents"



Techniques – Identify and Retrieve

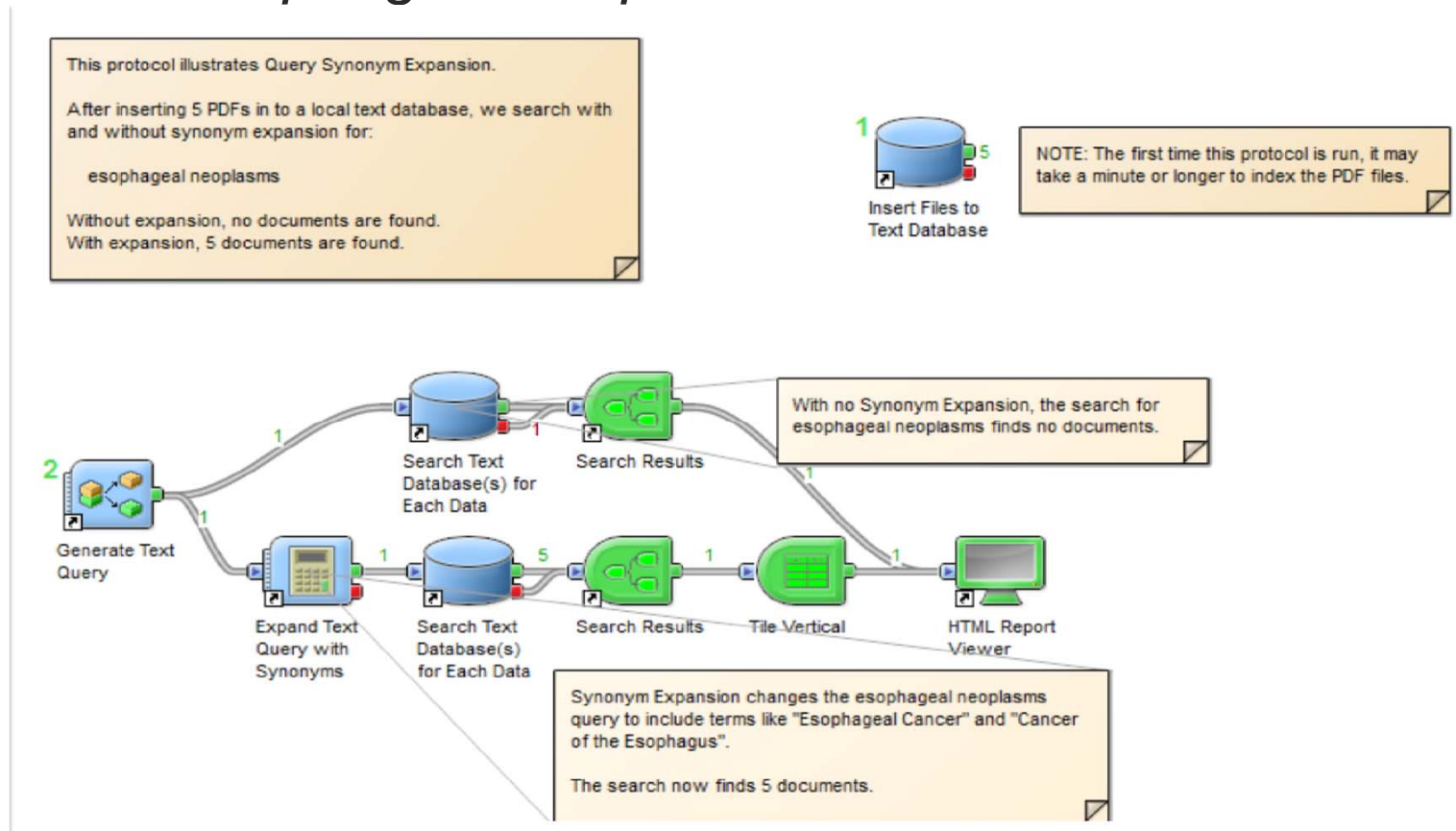


- A lot of readers and search components are available
- Simple text queries
- But also specific search capabilities like “*TAC Universal Query Language*” or “*PubMed Fields syntax*”
- But more sophisticated approaches are sometimes necessary
→ Semantics/Ontology

Techniques – Identify and Retrieve II

Semantics: Expands text query terms by using *Concept Dictionaries* (MeSH)

Search: “*esophageal neoplasms*”



Techniques – Identify and Retrieve II

Semantics: Expands text query terms by using *Concept Dictionaries* (MeSH)

Search: “*esophageal neoplasms*”

cancerIndex: No results for DocAllText:Esophageal AND DocAllText:Neoplasms

Warning: No results were found.

cancerIndex: Results 1 to 5 of 5 for DocAllText:"Esophageal Neoplasms" OR DocAllText:"Esophageal Neoplasm" OR DocAllText:"Esophagus Neoplasm" OR DocAllText:"Esophagus Neoplasms" OR DocAllText:"Cancer of Esophagus" OR DocAllText:"Cancer of the Esophagus" OR DocAllText:"Esophagus Cancer" OR DocAllText:"Esophagus Cancers" OR DocAllText:"Esophageal Cancer" OR DocAllText:"Esophageal Cancers" OR DocAllText:"Head and Neck Neoplasms" OR DocAllText:"Gastrointestinal Neoplasms"

Warning: Use uppercase AND, OR, or NOT for boolean operators

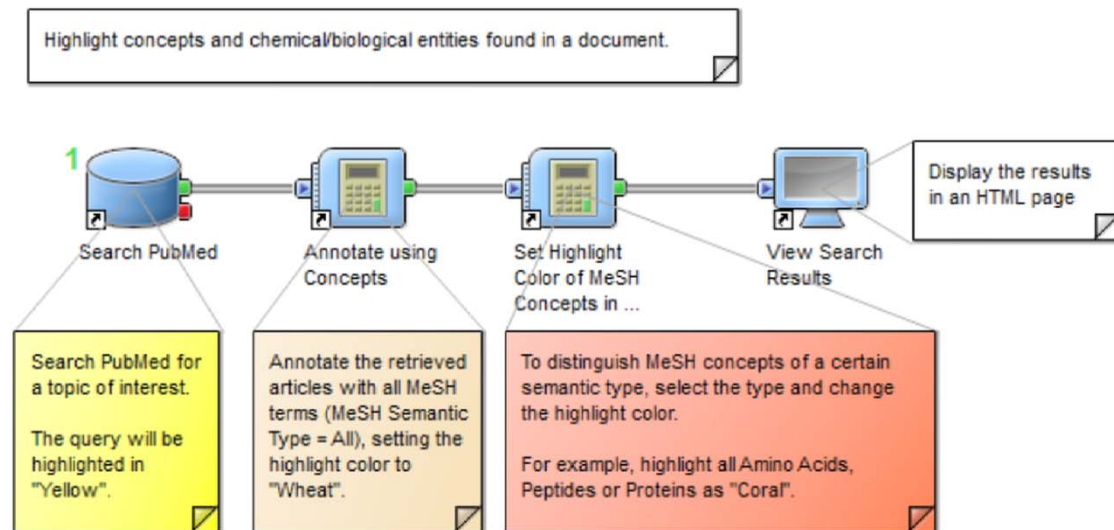
- 1 <http://deshlnx11-9944/scitegic-bin/DownloadFile/filename/opt/accelrys/AEP/apps/scitegic/textanalytics/public/data/ExampleDocuments/cancerPDFs/135.pdf>
... **Esophageal Cancer**, in: Comprehensive Registry of **Esophageal Cancer** in Ja- pan (1988–1994). Tokyo: Japanese Society for Esoph- ageal Diseases, 1998. (in Japanese) 5. Registration Committee for **Esophageal Cancer**. In: Comprehensive Registry of **Esophageal** ... can be effective for superficial **esophageal cancer**, while it is inapplicable for advanced **esophageal cancers** because the sentinel nodes are not clear in advanced **esophageal cancers** 14) The overall efficacy of sentinel node naviga- tion surgery for ... the esophagus: a summary of responses to a ques- tionnaire on superficial **cancer of the esophagus** in Ja- pan. Surgery 1998; 123: 432-9. 11. Makuuchi H, Shimada H, Chino O, et al. Endoscopic mucosal resection for m3.sm1 **esophageal cancer**. Clin ...
- 2 **Esophageal Cancer - Middle East Cancer Consortium (MECC) Cancer Incidence**
http://deshlnx11-9944/scitegic-bin/DownloadFile/filename/opt/accelrys/AEP/apps/scitegic/textanalytics/public/data/ExampleDocuments/cancerPDFs/mecc_esophageal.pdf
... Chapter 2 GUL ERGOR BACKGROUND **Cancer of the esophagus** is the eighth most common cancer worldwide [1], with ... **esophageal cancers** in all the countries in the ... rates were 2 to 3 times higher in males (11.5) than in females (4.7). Sex ratios for Middle Eastern populations can be judged from data taken from ... populations ... squamous cell carcinoma to adenocarcinoma (the other main type of **esophageal cancer**) was far higher in females than in males. Overall, approximately half of the **esophageal cancers** are squamous cell ...
- 3 <http://deshlnx11-9944/scitegic-bin/DownloadFile/filename/opt/accelrys/AEP/apps/scitegic/textanalytics/public/data/ExampleDocuments/cancerPDFs/980.pdf>
... CT) is the most important method of stag ing **cancer of the esophagus**. ... CT not only can identify the presence of disease but can also delineate the extent of tumor more accurately than endoscopy. Relying ... DOI: 10.3322/canjclin.34.2.127 1984;34;127 CA Cancer J Clin James W ... April 30, 2009 http://caonline.amcancersoc.org the World Wide Web at: The online version of this article, along ... November 1950, is published six times per year ... for Clinicians by on April 30, 2009 ((c)American Cancer Society, Inc.) ca o nline.am cancersoc.org D ow nloaded from **Esophageal Cancer** To the ...
- 4 <http://deshlnx11-9944/scitegic-bin/DownloadFile/filename/opt/accelrys/AEP/apps/scitegic/textanalytics/public/data/ExampleDocuments/cancerPDFs/980.pdf>
... currently applied staging of **esophageal cancer** ... Key Words: **esophageal cancer**; 18F-FDG PET; upstaging J Nucl Med 2004; 45:980-987 Adequate staging of **esophageal cancer** ... FDG PET is sensitive and accurate in the preoperative staging of distant metastases in patients with **cancer of the esophagus** and GEJ and leads to upstaging ... **esophageal cancer**: incremental value over computed tomography. Clin Positron Imaging. 1999;2:255-260. 8. Fukunaga T, Okazumi S, Koide Y, Isono K, Imazeki K. Evaluation of 18F-fluorodeoxyglucose PET. J Nucl Med. 1998;39: ...
- 5 <http://deshlnx11-9944/scitegic-bin/DownloadFile/filename/opt/accelrys/AEP/apps/scitegic/textanalytics/public/data/ExampleDocuments/cancerPDFs/1112.pdf>
... **cancer of the esophagus**, gastric cardia, or body of the stomach (8)). The deletion pat- terns on all chromosomes were similar between the family history-positive and family history-negative groups, except for chromosome 13, where a suggestion of a ... Short Communication Evidence for a Familial **Esophageal Cancer** Susceptibility Gene on Chromosome 13 Nan Hu, Alisa M. Goldstein, Paul S. Albert, Carol Giffen, Ze-Zhong Tang, Ti Ding, Philip R. Taylor1, and Michael R. Emmert-Buck2 Center for Cancer ... studies support a role for genetic sus- ceptibility to **esophageal cancer**, although the exact mechanism is unclear. To search for genes involved in the development or progression of ESCC in patients from Shanxi Province, we previously performed a genome- ...

More relevant papers are retrieved by semantic expansion

Techniques – Content Processing

How to retrieve Knowledge from Information?

- Search: “*RNAi*”
- Concept: MeSH
- Specific MeSH Concept



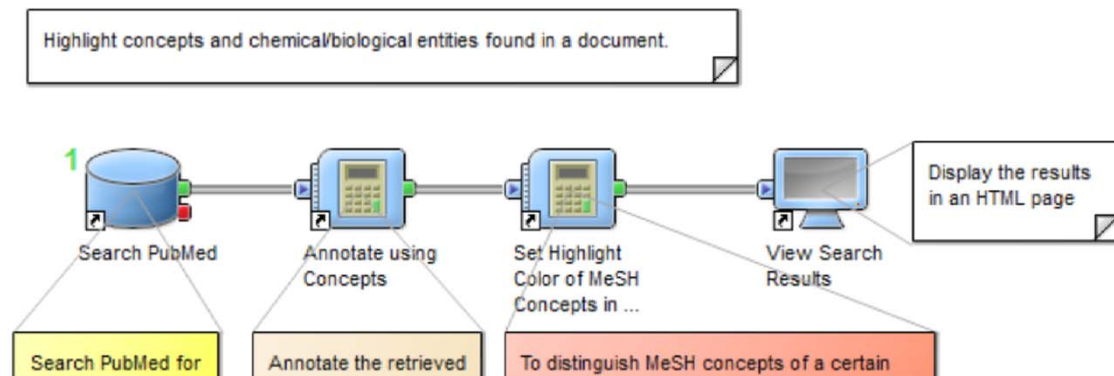
Parameters	
Properties To Process	DocAbstract
Concept Dictionary	MeSH - Medical Subject Headings (NLM)
Custom Dictionary Path	
Custom Annotation Key	dict
Highlight Color	Wheat
Selection Options	
Additional Options	

Parameters	
SemanticType	Amino Acid, Peptide, or Protein (mh_T116)
Background Color	Coral

Techniques – Content Processing

How to retrieve Knowledge from Information?

- Search: “*RNAi*”
- Concept: MeSH
- Specific MeSH Concept



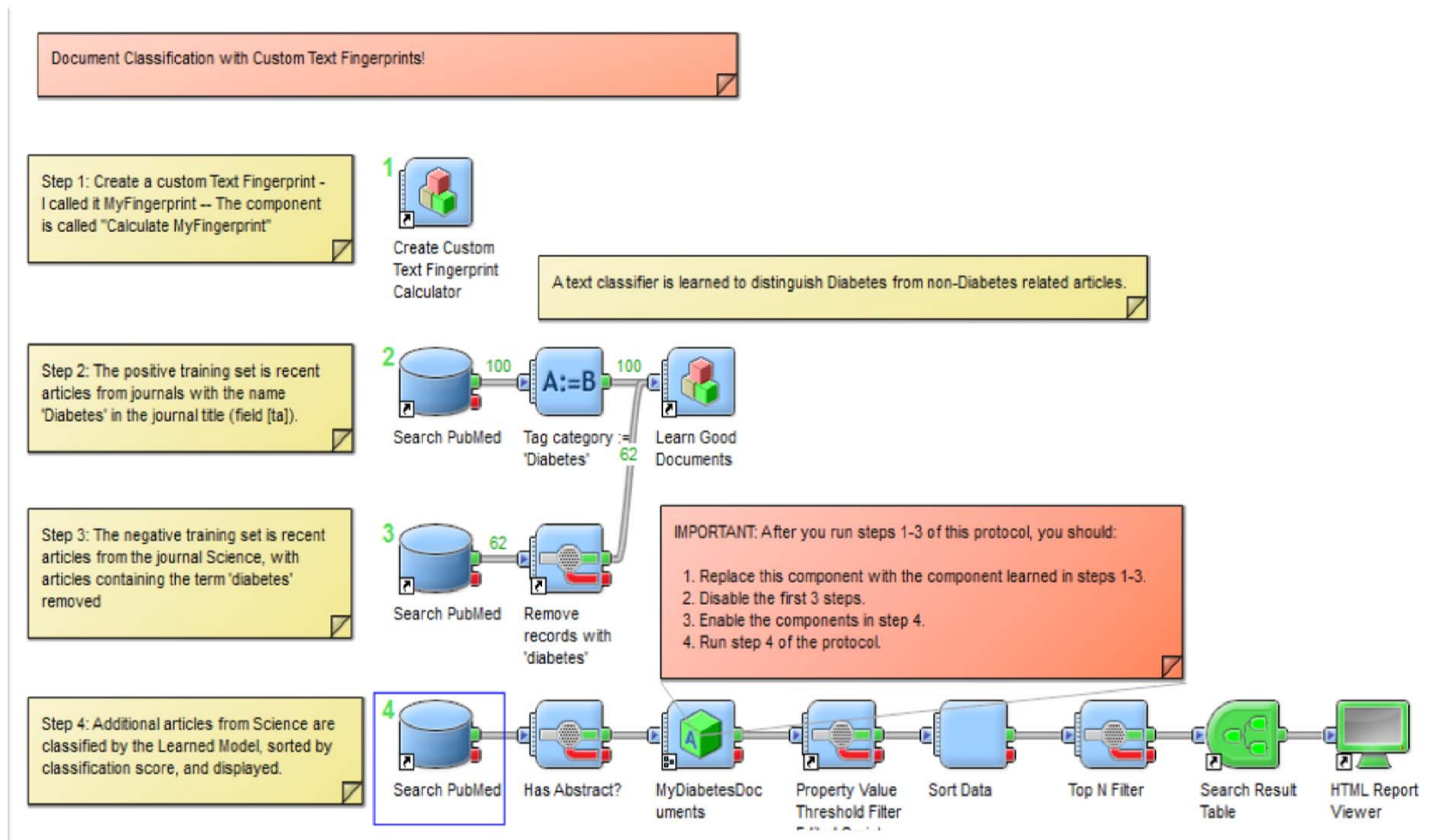
PubMed: Results 1 to 3 of 14488 for RNAi

- 1 The Importance of the 45S Ribosomal Small Subunit-related Complex for Mitochondrial Translation in Trypanosoma brucei.**
J Biol Chem 2013/10/02: Ridlon, Lucie; Skodova, Ingrid; Pan, Songqin; Lukes, Julius; Maslov, Dmitri A
UC Riverside, United States;
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=24089529
The mitochondrial 45S SSU* complex in Trypanosoma brucei contains the 9S SSU ribosomal RNA, a set of SSU ribosomal proteins, several pentatricopeptide repeat (PPR) proteins, and proteins not typically found in ribosomes, including rhodanese-domain protein (Rhod) and a 200 kDa coiled-coil protein. To investigate the function of this complex, PPR29, Rhod, 200 kDa proteins and mitochondrial ribosomal protein S17 were knocked-down by RNAi in procyclic T. brucei. A growth retardation phenotype, a reduction in the amount of the 45S SSU* complexes, and the preferential inhibition of synthesis of the cytochrome c oxidase subunit I (COI) over cytochrome b (Cyb) were observed as early as day 2 post induction of RNAi. On the contrary, the down- ...
- 2 Identification of Transcription Factors Involved in Rice Secondary Cell Wall Formation.**
Plant Cell Physiol 2013/10/01: Hirano, Ko; Kondo, Mari; Aye, Koichiro; Miyao, Akio; Sato, Yutaka; Antonio, Baltazar A; Namiki, Nobukazu; Nagamura, Yoshiaki; Matsuoka, Makoto
Bioscience and Biotechnology Center, Nagoya University, Nagoya, Aichi, Japan.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=24089432
Through co-expression network analysis, we identified 123 rice transcription factors (TFs) as candidate rice secondary cell wall regulators (Hirano et al., the accompanying paper). To validate whether these TFs are associated with secondary cell wall formation, six TF genes belonging to either the MYB, NAC or homeodomain containing TF families were over-expressed or down-regulated in rice. With the exception of OsMYB58/63-RNAi plants, all transgenic plants showed phenotypes possibly related to secondary cell wall alteration, such as dwarfism, narrow and dark green leaves, and also altered rice cinnamyl alcohol dehydrogenase 2 (OsCAD2) gene expression and lignin content. These results suggest that many of the 123 candidate secondary cell ...
- 3 Knockdown of beta-catenin with Dicer-Substrate siRNAs Reduces Liver Tumor Burden In Vivo.**
Mol Ther 2013/10/03: Dudek, Henryk; Wong, Darren H; Arvan, Rokhand; Shah, Anee; Wortham, Kathleen; Ying, Bo; Diwanji, Rohan; Zhou, Wei; Holmes, Benjamin; Yang, Hailin; Cyr, Wendy A; Zhou, Yi; Shah, Aalok; Farkiwala, Ruchir; Lee, Michael; Li, Yiting; Rettig, Garrett R; Collingwood, Michael A; Basu, Sujit K; Behlke, Mark A; Brown, Bob D
Dicerna Pharmaceuticals, Watertown, MA 02472 USA.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=24089139
... appropriate genes to target, and achieving sufficient knockdown in tumors. We have developed high potency Dicer-substrate siRNAs (DsiRNAs) targeting beta-catenin and delivered these in vivo using lipid nanoparticles, resulting in significant reduction of beta-catenin expression in liver cancer models. Reduction of beta-catenin strongly reduced tumor burden, alone or in combination with sorafenib and as effectively as DsiRNAs that target mitotic genes such as PLK1 and KIF11. beta-catenin knockdown also strongly reduced expression of beta-catenin-regulated genes including MYC, providing a potential mechanism for tumor inhibition. These results validate beta-catenin as a target for liver cancer therapy, and demonstrate the promise of RNAi in ...

This search at PubMed: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=search&db=pubmed&doptcmdl=DocSum&term=RNAi&dispmx...>

Techniques – Content Processing II

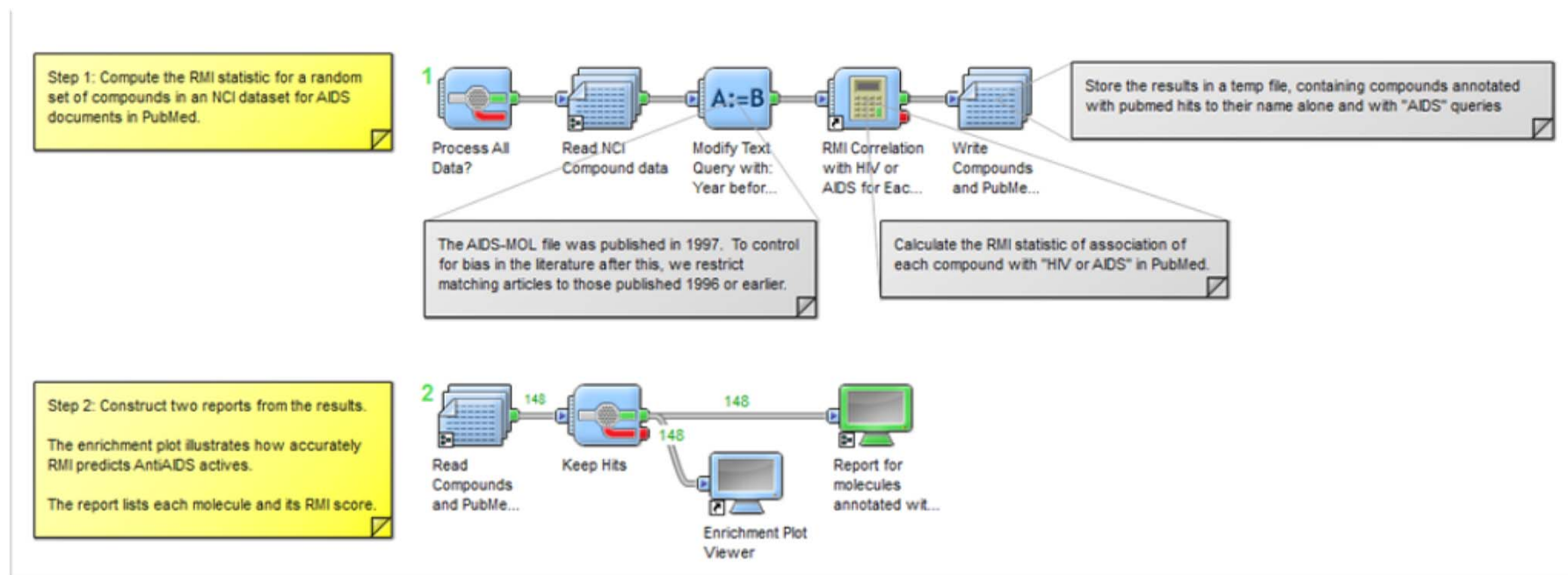
Advanced Classifications, Text Fingerprints



Techniques – Content Processing III

Relative Mutual Information (RMI)

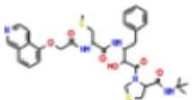
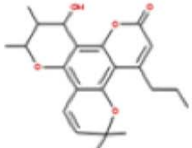
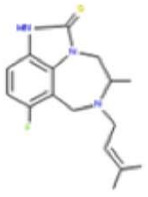
→ Use Text to Find Anti-AIDS Actives



Techniques – Content Processing III

Relative Mutual Information (RMI)

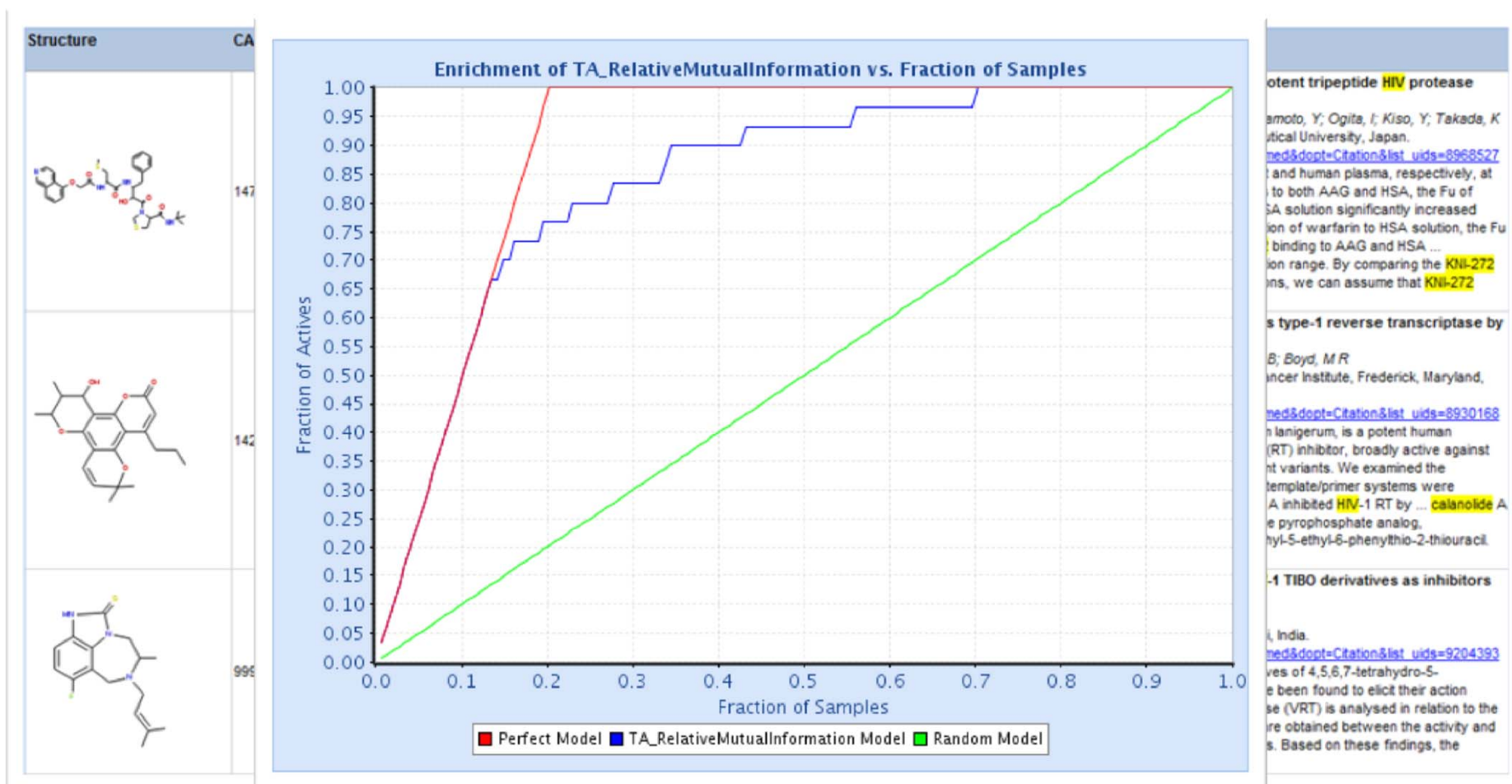
→ Use Text to Find Anti-AIDS Actives

Structure	CAS_RN	Name (1st listed)	Activity	PubMed Hits	RMI	Recent Citation
	147318-81-8	4-Thiazolidinecarboxamide, N-(1,1-dimethylethyl)-3-[2-hydroxy-3-[[2-[(5-isoquinolinyloxy)acetyl]amino]-3-(methylthio)-1-oxopropyl]amino]-1-oxo-4-phenylbutyl]-[4R-[3[2S*, 3S*(R*)],4R*]]	CA	23	-12.53	Binding characteristics of KNI-272 to plasma proteins, a new potent tripeptide HIV protease inhibitor. Biopharm Drug Dispos 1996: Kiriya, A; Nishiura, T; Ishino, M; Yamamoto, Y; Ogita, I; Kiso, Y; Takada, K Department of Pharmaceutics and Pharmacokinetics, Kyoto Pharmaceutical University, Japan. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=8968527 ... The unbound fractions (Fu) of KNI-272 were 12.13 and 2.24% in rat and human plasma, respectively, at the drug concentration of 1.0 microgram mL ⁻¹ . Although KNI-272 binds to both AAG and HSA, the Fu of KNI-272 in AAG solution was 1.83%, and only one- ... of KNI-272 in HSA solution significantly increased when warfarin and diazepam were added. In particular, with the addition of warfarin to HSA solution, the Fu of KNI-272 increased to 16%. The modified Scatchard plots of KNI-272 binding to AAG and HSA ... disopyramide on AAG and site I on HSA in the low KNI-272 concentration range. By comparing the KNI-272 binding parameters obtained in human plasma and these protein solutions, we can assume that KNI-272 binding at low concentration in human plasma is ...
	142632-32-4	2H,6H,10H-Benzo[1,2-b:3,4-b':5,6-b'']tripyran-2-one, 11,12-dihydro-12-hydroxy-6,6,10,11-tetramethyl-4-propyl-, [10R-[10.alpha.,11.beta.,12.alpha.]]- (9CI)	CA	10	-12.53	Kinetic analysis of inhibition of human immunodeficiency virus type-1 reverse transcriptase by calanolide A . J Pharmacol Exp Ther 1996: Currens, M J; Mariner, J M; McMahon, J B; Boyd, M R Laboratory of Drug Discovery Research and Development, National Cancer Institute, Frederick, Maryland, USA. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=8930168 Calanolide A , first isolated from the tropical rain forest tree Calophyllum lanigerum, is a potent human immunodeficiency virus type-1 (HIV -1) specific reverse transcriptase (RT) inhibitor, broadly active against diverse HIV -1 strains, including nucleoside and nonnucleoside-resistant variants. We examined the biochemical mechanism of inhibition of HIV -1 RT by calanolide A . Two template/primer systems were examined: ribosomal RNA and homopolymeric rA-dT 12-18. Calanolide A inhibited HIV -1 RT by ... calanolide A bound HIV -1 RT in a mutually exclusive fashion with respect to both the pyrophosphate analog, phosphonoformic acid and the acyclic nucleoside analog 1-ethoxymethyl-5-ethyl-6-phenylthio-2-thiouracil. This indicates that calanolide A ...
	999-99-9	TBO CACVII-23	CA	2	-12.53	Quantitative structure-activity relationship studies on anti- HIV -1 TBO derivatives as inhibitors of viral reverse transcriptase. J Enzyme Inhib 1996: Gupta, S P; Garg, R Department of Chemistry, Birla Institute of Technology & Science, Pilani, India. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=9204393 The anti-human-immunodeficiency-virus (HIV -1) activity of the derivatives of 4,5,6,7-tetrahydro-5-methylimidazo [4,5,1-jk] [1,4] benzodiazepin-2(1H)-one (TBO) that have been found to elicit their action through the allosteric inhibition of the enzyme viral reverse transcriptase (VRT) is analysed in relation to the physicochemical properties of the molecules. Significant correlations are obtained between the activity and the hydrophobic constant and some dummy parameters of substituents. Based on these findings, the mechanism of action of these anti- HIV drugs is discussed.

Techniques – Content Processing III

Relative Mutual Information (RMI)

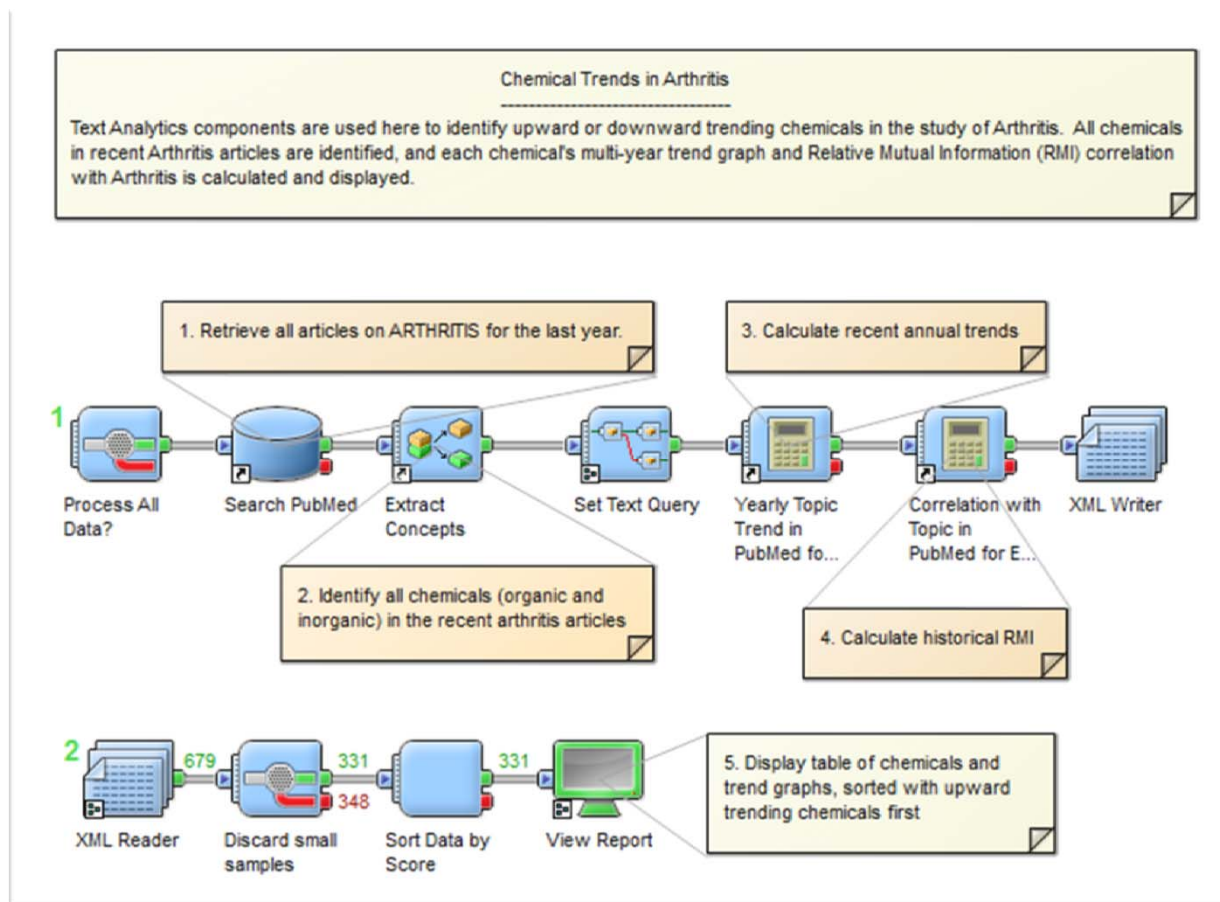
→ Use Text to Find Anti-AIDS Actives



Techniques – Content Processing IV

RMI Correlation and Trends

→ Chemical Trends in Arthritis



Techniques – Content Processing IV

RMI Correlation and Trends

→ Chemical Trends in Arthritis

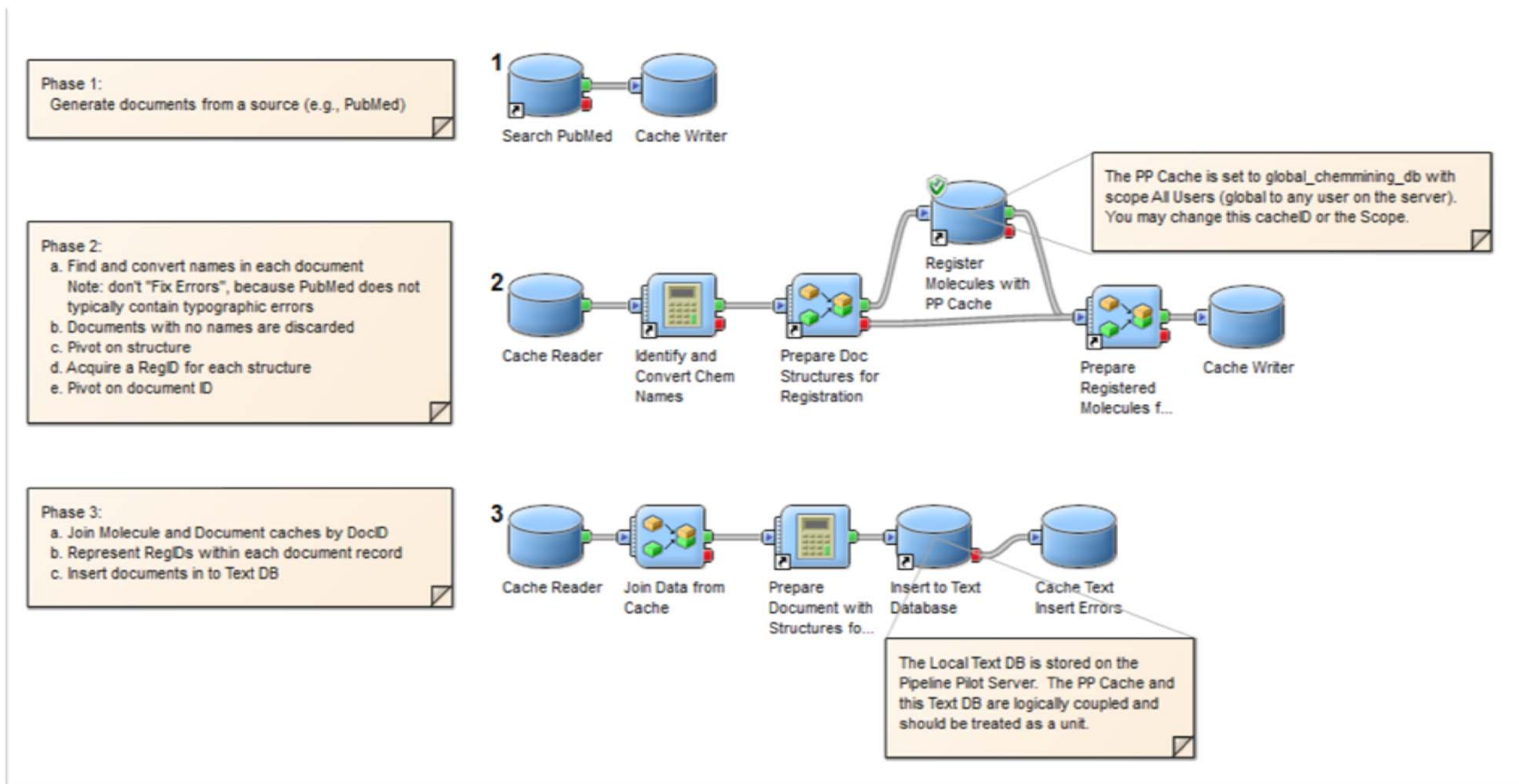
Chemicals trending with: arthritis

Name	2010 Count (in arthritis)	Mean (all years)	Change Ratio (2010-3yrWtdAvg) / (2010+3yrWtdAvg)	Trend (yearly entity AND arthritis citation count normalized by year count)
INCB018424	3	-13.606	1	
trizol	3	-16.4205	1	
Methylene Chloride	3	-18.4705	0.94	
beta Carotene	3	-17.679	0.94	
imiquimod	3	-17.5265	0.7885	
tanshinone	3	-16.6425	0.7125	
tanshinone II A	3	-16.6425	0.7125	

Techniques – Content Processing V

Data Mining – Extract Chemical Knowledge/Objects

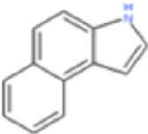
- Identify and extract chemicals by name, convert them to chemical structures and store them into a chemical database
- Store related literature in text database



Techniques – Content Processing V

Data Mining – Extract Chemical Knowledge/Objects

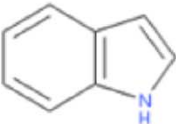
- Identify and extract chemicals by name, convert them to chemical structures and store them into a chemical database
- Store related literature in text database

Query Structure	Query Constraints	Result Summary
 2,3-BENZINDOLE	Database(s): /SATA/USR/foellien/TextDBs Search Type: Search by Similarity Threshold: Low (0.5) Text Query: oxidation Search for Each Molecule Separately: True	Molecules Matching Search by Similarity: 2 Molecules with Documents Matching Text Query: 1

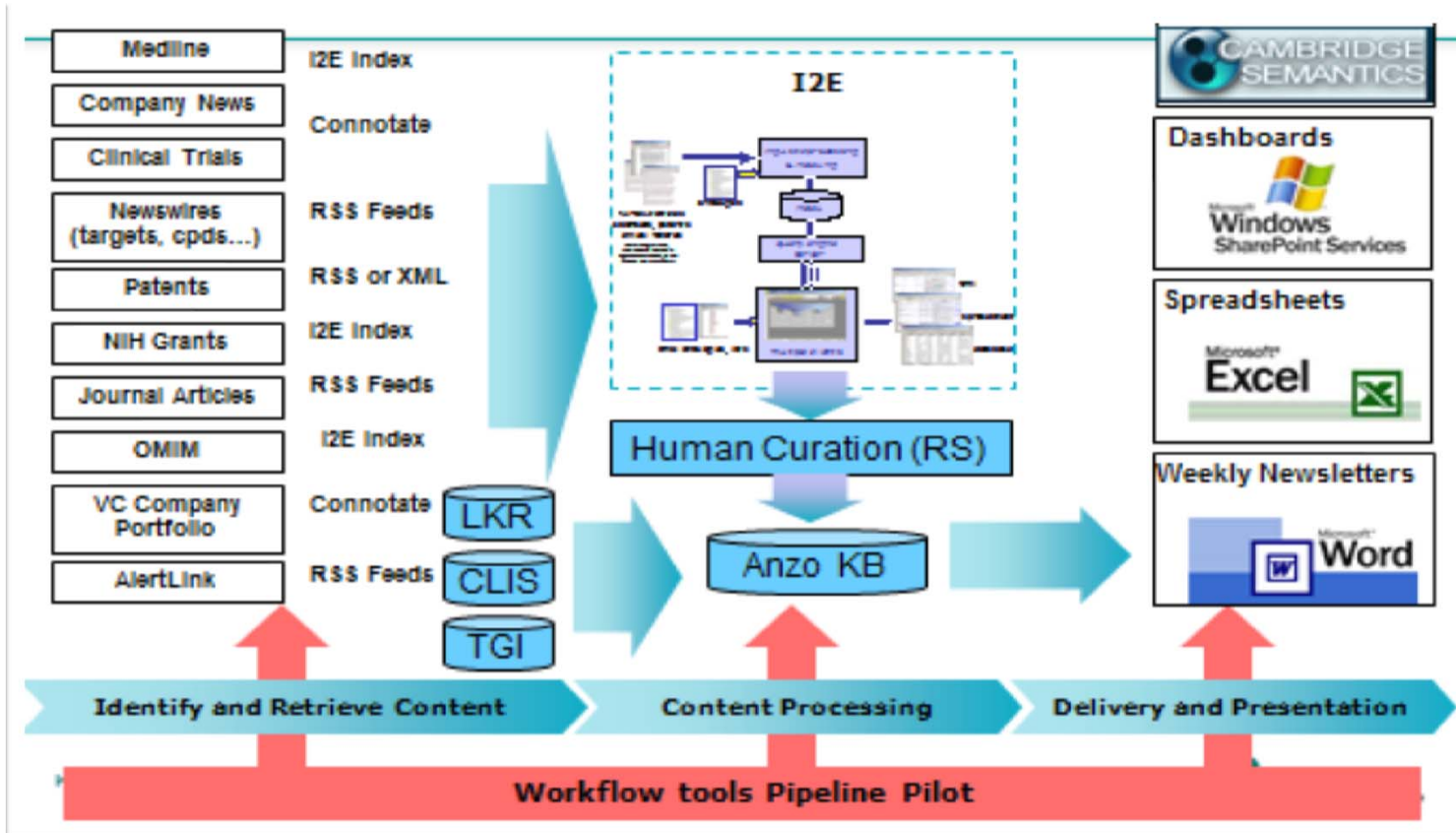
Color Key: Matching Molecules Other Molecules Unconverted Names Text Query

Molecules **1** to **1** of **1** (Molecules with Documents Matching Text Query)

Export Molecules: 1-2

indole	Search Results
 indole	<p>Result 1 of 3 for indole (all variants) AND (oxidation)</p> <p><input type="button" value="More Documents..."/> <input type="button" value="Export All 3 Documents"/></p> <p>1 Synthesis of indoles via 6pi-electrocyclic ring closures of trienecarbamates. J Am Chem Soc 2006/04/19: Greshock, Thomas J; Funk, Raymond L Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16802, USA. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=16608316 A new method for the preparation of indoles from readily available alpha-haloenones and alpha-(trialkylstanny)enecarbamates is described. Following a Stille coupling, trienecarbamate 2 is electronically activated to undergo a facile 6pi-electrocyclic ring closure and subsequent oxidation to afford protected aniline 4. Upon deprotection and reductive amination, acid 5 underwent clean cyclization to N-acetylindole 6 (Ac₂O, NEt₃, 130 degrees C). This method has been used to construct a variety of substituted indoles that are not easily prepared by conventional indole annelation methods.</p>

Techniques – Delivery and Presentation





Current Limitations

Nature News, 21.03.2013

IN FOCUS NEWS

PUBLISHING

Text-mining spat heats up

“Fearful that their content might be freely redistributed, publishers tend to block programs that they find crawling the full text of articles, making no exceptions for users who have paid for access. They give permission only on a case-by-case basis to those who negotiate agreements on access and use.”

complained bitterly over the past year that publishers do not allow them to use computer programs to download and crawl across the text of research articles, a methodology known as text mining that can reveal large-scale patterns in the studies (see ‘Uses of text mining’).

Fearful that their content might be freely

Linking genes to research papers.

The text2genome project pulls out DNA sequences from around 3 million research papers to produce an online genome map in which each region is linked to relevant articles.

go.nature.com/lupijx

might be permitted by ‘fair use’ rights, which allow snippets of text to be freely copied. But no one knows for sure, and many researchers are wary of testing the bounds of this legal grey area.

Some publishers say that unrestricted text mining could strain their servers, and so agreements will always be needed to specify when



Full Paper Accessibility Limitations

General Aspects

Old Full Papers (<1990)

- Not available as full text PDF, but as scanned images saved as PDFs
- OCR software needed to get text

Subscriptions

- Customers do not subscribe to each and every journal
- Only Full Papers that are covered by subscriptions or OA papers are (could be) available for text mining



Full Paper Accessibility Limitations II

Licenses and Copyright

- Almost all publishers have clauses in licenses prohibiting any robotic, systematic mining of their websites
- **This also affects customers with valid subscriptions**
- Copyright issues prevent bulk downloading and local storage of full papers
- Some progress has already been made by publishers
- Intermediary workarounds like CCC will be used
- Additional solutions might become available by the end of this year
- But currently, we are still facing some limitations
- Even if there are no copyright and license issues like in the case of Open Access publishers, restrictions still exist that handle if and how frequently you scrape or make calls to the publisher websites



Full Paper Accessibility Limitations III

Hardware Infrastructure

- Publishers common websites and web interfaces are generally not equipped to handle extreme traffic
- Massive impact on Server infrastructure
- Text Mining activities are interfering with their normal web traffic
 - Bulk download of full papers
 - On the fly text mining of full papers

Clauses protect publishers also for text mining caused breakdowns of their common websites



Full Paper Accessibility Limitations IV

Improper Distribution and Search Capabilities

- Full text papers are improperly distributed for text mining purposes
 - Many publishers and sources
 - Specific readers for each source needed
- No standardized APIs to perform searches
 - In some cases no API at all
 - Different APIs force the development of specific readers
 - No standardized feeds available
 - Additional post-processing needed
- Limited Search Capabilities
 - Only simple text queries are available or simple advanced queries
 - Semantic/Concept Dictionary-driven searches are not possible
 - More sophisticated search
 - Covers more of the relevant information
 - Limits searches to the important full paper → limitation of the number of papers that have to be downloaded)



Full Paper Accessibility Limitations V

Financial Aspects

- Many, sophisticated text mining tools are available and will be used by customers. However these tools are costly and sometimes require specialists to operate
- Downloading tools or additional services as CCC and QUOSA are also costly and even increase the financial expenses
- Text Mining agreements are often related to additional costs, even if subscription contracts are already in place
- A text mining-driven full paper download has **not** the same value as a common, manual full paper download! Most/many papers will be rejected during the text mining approach. Only few papers contain the desired information
- Text Mining cannot be performed on a financial negotiated case-by-case basis (long-term planning) → **text mining tasks are too diverse and occur as part of the daily work**
- In the current practice lies another obstacle in times of budgetary constraints



Workarounds – Abstracts, PM Central

Use Public Available Information - Abstracts

- The greatest success has been made in mining publicly available information such as citations and abstracts from PubMed/Medline
- This has been/can be licensed in as a flat file for text mining purpose
- Limited to Life Science field
- Can be used as starting point to identify promising and valuable full papers

Use PubMed Central (OA Repository)

- Access to full papers without license restrictions
- This has been/can be licensed in as downloaded copy for text mining
- Only 2-3 million papers available, 10% of PubMed/Medline
- Only Papers after 1990, old articles are not included

Other OA Full Papers from OA Publishers

- No copyright issue
- But, all other restrictions (incl. restricting clauses)



Workarounds – Customized Solution

Publisher-specific, customized Solutions

- Evaluate customized text mining opportunities in private projects/contracts between customer and publisher
- Pharma customers have experimented with having access to a publisher's website to mine at a specified rate during off-hours, and in having feeds or APIs from publishers
 - Try to prevent interfering with publishers normal web traffic
 - Try to spare publishers infrastructure
- These options come at additional cost and the coverage of published literature has not been comprehensive

→ therefore most of these efforts have been short-lived

Workarounds – CCC and QUOSA

QUOSA: Scientific Literature Management Software (Elsevier)



“QUOSA has been used in a limited fashion by pharma industry at least 7 years as a downloading tool. It works by linking from specified search results to download pdf files for mining. The product has been very problematic in our environment, both in terms of speed and accuracy. PubGet was a competitor, primarily focused on PubMed search results and had limited practicality. PubGet was purchased by the Copyright Clearance Center which plans to develop a new text mining tool, but has nothing to offer at this time.”

Copyright Clearance Center (CCC)



*“The Copyright Clearance Center (CCC) in Danvers, Massachusetts, which works with publishers on rights licensing, is pursuing a more ambitious effort. It would act as an intermediary, collecting publishers’ terms and content and storing them on a website for researchers. It is working with six publishers (including Nature Publishing Group) and with **drug and chemical firms** eager to mine the literature.”*

Help to handle the copy rights and other issues (e.g. bulk downloading) with the publishers



(Ideal) Future Perspective

- Use of standardized APIs and feeds (over all publishers incl. OA)
- Use mirror sites where tools can operate without interfering with the primary publisher's website
- Proper distribution / platforms that cover several publishers (CrossRef)
- Availability of semantic searches (concept dictionaries) to allow advanced searches → limit number of full paper to retrieve
- Text Mining should be possible directly between publishers and customers based on existing subscriptions
- Reasonable costs / Adapted pricing schemes
- Copyright/License harmonization
- **No case-by-case basis for access, agreements and licenses but a general agreement that will allow text mining approaches in general**

It's about having Access and not about costs

Thank You

