Jul 6th, 12:00 AM

Innovative usage of unstructured information sources: From text- and data-mining to model-driven decision-support

Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany

“Innovative usage of unstructured information sources: From text- and data-mining to model-driven decision-support”

Prof. Dr. Martin Hofmann-Apitius

Head of the Department of Bioinformatics

Fraunhofer Institute for Algorithms and Scientific Computing

36th Annual IATUL Conference 2015

July 6 - 9, Hannover, Germany

Fraunhofer SCAI

Where I come from: Fraunhofer Society

*Joseph von Fraunhofer (1787 – 1826)
Scientist, Inventor and Entrepreneur

- Founded 1949
- Europe´s largest applied research organisation
- 60 Research Institutes (7 Institutes in the US)
- > 23.000 Employees
- Annual Budget > 2 Billion Euro
- Financial model: 2/4 industry collaborations
  1/4 public funding
  1/4 institutional funding

The Fraunhofer Institute Center Schloss Birlinghoven

- Largest research centre for informatics and applied mathematics in Germany
- Around 700 employees, thereof 500 scientists, approx. 200 students and trainees
- University links:
  - Bonn
  - Aachen

Fraunhofer SCAI

Expertise at the Department of Bioinformatics at SCAI

Fraunhofer SCAI Department of Bioinformatics currently comprises:

- 10 scientists
- 3 scientific software developers
- 7 PhD students
- ~ 5 Master students
- ~ 10 student workers

- predominantly computer scientists & biologists
- some PhD students via University of Bonn (Bonn-Aachen International Center for Information Technology)

SCAI Department of Bioinformatics: R&D in a nutshell

Fraunhofer SCAI Department of Bioinformatics R&D activities:

1. Information extraction in the **life sciences**:

    - Recognition of named entities and relationships in text
    - Large-scale, automated Information Extraction

2. Integrative biology; disease modelling

    - Focus on neurodegenerative diseases

3. eScience, Grid- & Cloud- Computing / HPC (Cluster)

    - Focus on scaling of information extraction workflows

Making Scientific Content available for Computing

Imagine …

- **Cancer Patient, final stage, metastatic pancreas carcinoma**
  - Surgery, chemotherapy without success
  - Distant metastasis in bone marrow, lung and liver
  - Remaining life span: 2 – 4 weeks
- **Last chance: sequencing the cancer genome (< 10k€)**
  - Getting insight into mutations underlying cancer dysregulation
  - Understanding of mechanisms triggering uncontrolled growth
  - Identification of (experimental) compounds that inhibit tumour growth
- **This is not fiction – this is reality**

The challenge …

- **Cancer genome sequencing delivers vast amount of information**
  - Tens to hundreds of mutations
  - Functional relevance of a significant number of mutations unclear
  - Contribution of mutation to tumour growth and metastasis ?
- **How do we assess the biological impact of genetic variation information?**
  - Putting genetic variation information into a functional context
  - Reasoning over genetic variation information and inference of consequences
  - From inference to personalised recommendation …. within 2 weeks of time
- **Let us see where we stand …**

Semantic Search and Knowledge Discovery in Scientific Literature

ProMiner

Recognition and Normalization of Named Entities in Scientific Text

SCAIVIEW

Identification and normalisation of the relevant Life Science terminology is key for information retrieval, information extraction and inferring of knowledge

Fraunhofer SCAI

Named Entity Recognition Technologies @SCAI

ProMiner
Recognition and Normalization of Named Entities in Scientific Text

Machine learning based approaches
(CRF based)
SNP, IUPAC, epigenetic modifications

Dictionary based approaches
with normalisation and
embedding in hierarchies

Genes/Proteins, Disease, Drugs,
Cells...

Regular expression
(partly to normalise)
Chrom. Locations, rs numbers...

Combined entities
Combined normalisation
SNP, Histonmodification

Fraunhofer SCAI

Different input output formats

ProMiner

Recognition and Normalization of Named Entities in Scientific Text

MEDLINE abstracts

ASCII text

XML text, e.g. patents

Defined format

HTML text e.g. Journal articles, Web pages

PDF text e.g. Journal articles

1. Association of breast cancer resistance protein/ABCG2 phenotypes and novel promoter and intron 1 single nucleotide polymorphisms.

PubMed 18180275 **Authors:** Balasubramanian Poonkuzhali, Jatinder Lamba, Stephen Strom, Alex Sparreboom, Kenneth Thummel, Paul Watkins, Erin Schuetz, **Date:** 2008-04- **Journal:** Drug metabolism and disposition: the biological fate of chemicals **SciMago:** 0.43

Statistics

The hypothesis was tested that sequence diversity in breast cancer resistance protein (BCRP)'s cis-regulatory region is a significant determinant of BCRP expression. The BCRP promoter and intron 1 were resequenced in lymphoblast DNA from the polymorphism discovery resource (PDR) 44 subset. BCRP single nucleotide polymorphisms (SNPs) were genotyped in donor human livers

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Matrix Biology 27 (2008) 3–11

MATRIX BIOLOGY

www.elsevier.com/locate/matbio

Ucma — A novel secreted factor represents a highly specific marker for distal chondrocytes

Andreas Tagariello [a], Julia Luther [a], Melanie Streiter [a], Lydia Didt-Koziel [b], Manuela Wuelling [b], Cordula Surmann-Schmitt [c], Michael Stock [c], Nadia Adam [c], Andrea Vortkamp [b], Andreas Winterpacht [a,*]

[a] Institute of Human Genetics, University Hospital Erlangen, Schwabachanlage 10, D-91054 Erlangen, Germany
[b] Center for Medical Biotechnology, University of Duisburg-Essen, Germany
[c] Department of Experimental Medicine I, Nikolaus-Fiebiger Center of Molecular Medicine, University of Erlangen-Nuremberg, Erlangen, Germany

Received 4 April 2007; received in revised form 3 July 2007; accepted 6 July 2007

Abstract

Growth and development of most parts of the vertebrate skeleton takes place by endochondral ossification, a process during which chondrocytes undergo distinct stages of differentiation resulting in a successive replacement of the cartilage anlagen by bone. In the context of an EST project we isolated a novel transcript from a human fetal growth plate cartilage cDNA library. The transcript which we called Ucma (unique cartilage matrix-associated protein) encodes a short protein of 138 amino acids. The protein sequence is evolutionary conserved throughout vertebrates and comprises a signal peptide, a coiled-coil domain, and a putative dibasic cleavage site for proprotein convertases. Using RNA *in situ* hybridization and immunohistochemistry with a polyclonal anti-Ucma antibody we found high expression of Ucma uniquely in distal (resting) chondrocytes in developing long bones of wildtype mice. This restricted expression could also be observed in *Ihh*$^{-/-}$, *Ihh*$^{-/-}$; *Gli3*$^{-/-}$, *Gli3*$^{-/-}$ mice, and in mice that overexpress *Ihh* under the control of the *Col2a1* promoter indicating that expression of Ucma is regulated independent of hedgehog signaling. During insulin-induced differentiation of ATDC5 cells we found gradual increase of Ucma expression at day 21 with a maximum at day 24 and a decrease correlating with a simultaneous increase in the expression of cartilage link protein (Crtl1), a protein with maximum expression in column-forming proliferating chondrocytes. The present data strongly suggest an important function of Ucma in the

Named Entity Recognition and Normalisation

☐ Chromosomal Locations ☑ Drug Names ☑ Protein/Gene ☐ iupac ☐ OMIM Reference
☐ @neurlST ☑ non Normalized SNP ☑ Normalized SNP ☑ MeSH Disease ☐ Relations
☐ Cell Lines ☐ Genetic Association ☐ CRF-laura ☐ Brenda

1. Association of breast cancer resistance protein/ABCG2 phenotypes and novel promoter and intron 1 single nucleotide polymorphisms.

PubMed 18180275 **Authors:** Balasubramanian Poonkuzhali, Jatinder Lamba, Stephen Strom, Alex Sparreboom, Kenneth Thummel, Paul Watkins, Erin Schuetz, **Date:** 2008-04- **Journal:** Drug metabolism and disposition: the biological fate of chemicals **SciMago:** 0.43

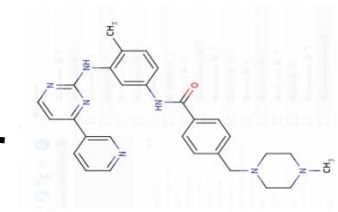
☐ *Statistics*

The hypothesis was tested that sequence diversity in breast cancer resistance protein (BCRP)'s cis-regulatory region is a significant determinant of BCRP expression. The BCRP promoter and intron 1 were resequenced in lymphoblast DNA from the polymorphism discovery resource (PDR) 44 subset. BCRP single nucleotide polymorphisms (SNPs) were genotyped in donor human livers, intestines, and lymphoblasts quantitatively phenotyped for BCRP mRNA expression. Carriers of the -15622C>T SNP had lower BCRP expression in multiple tissues. The intron 1 SNP 16702C>T was associated with high expression in livers; 1143G>A was associated with low expression in intestine; 12283T>C was associated with higher expression in the PDR44 and White livers. The -15994C>T promoter SNP was significantly associated with higher BCRP expression in multiple tissues. Patients with the -15994C>T genotype had substantially higher clearance of p.o. imatinib. We next determined whether BCRP expression was related to polymorphic alternative splicing or alternative promoter use. Liver polymorphically expressed an alternatively spliced mRNA [splice variant (SV) 1] skipping exon 2. Although SV1+ livers did not uniformly carry the exon 2 G34A allele, 90% of G34A livers expressed SV1 (versus 4% of 34GG livers). BCRP mRNA was significantly lower among Hispanic livers with the G34A variant genotype and may be due, in part, to polymorphic exon 2 splicing. Analysis of allele expression imbalance (AEI) showed that PDR44 samples with AEI had lower BCRP mRNA expression; however, no linked cis-polymorphisms were identified. BCRP used multiple promoters, and livers differentially using alternative exon 1b had lower BCRP. In conclusion, BCRP expression in lymphoblasts, liver, and intestine is associated with novel promoter and intron 1 SNPs.

GeneID: 9429: ABCG2
GO:0016887: ATP hydrolase
KEGG:02010: ATP transporter

Neoplasms by Site
Breast Neoplasms (D001943)
Breast Neoplasms, Male
Carcinoma, Ductal, Breast
Phyllodes Tumor

DB00619: imatinib



refSNP ID: rs2231137
ABCG2-Position:18897
Avg Het: 0.203+/-0.246

Fraunhofer SCAI

SCAIVIEW

Entity Tree View, select Entity Class to view and search

multiple sclerosis

Search | Documents | Entity | Analysis | Help | About

Subcorpus Statistics | Server Statistics

The following entities relating to 'multiple sclerosis' were found in 1244 documents.

Select Columns | Export Table | Export PMIDs | Export Entities | Sentence

- Human Genes / Proteins
- Chromosomal Location
- STS Marker
- non Normalized SNP
- Normalized SNP
- Normalized CRF SNP
- Drug Names
  - ALIMENTARY TR
  - ANTIINFECTIVES
  - ANTINEOPLAST
  - ANTIPARASITIC
  - BLOOD AND BL
  - CARDIOVASCUL
  - DERMATOLOGIC
  - GENITO URINAR
  - MUSCULO-SKEL
  - + NERVOUS SYS
  - RESPIRATORY S
  - SENSORY ORGA
  - SYSTEMIC HOR
  - VARIOUS
- IUPAC-like
- Corpora
- Human miRNA
- Mouse Genes
- MeSH Disease
- @neurIST Ontology
- GO Component
- GO Function
- GO Process

Select Confidence:

1 2 3 4

From MEDLINE®, Pub
PubMed Central®,datab
U.S. National Library o
PubMed PubM

SCAIVIEW

Filter next search in the current entity class exclusively to this entity

multiple sclerosis

Search | Documents | Entity | Analysis | Help | About

Result for 'multiple sclerosis', NER run 'DrugBank' for entity Carbamazepine, Page 0 > with 10 documents per page, totals to 81 and took 281 ms.

Toggle Abstracts

Chromosomal Locations
IUPAC | Epigenetics

- Human Genes / Proteins
- Chromosomal Location
- STS Marker
- non Normalized SNP
- Normalized SNP
- Normalized CRF SNP
- Drug Names
  - ALIMENTARY TRACT AND METAB
  - ANTIINFECTIVES FOR SYSTEMIC
  - ANTINEOPLASTIC AND IMMUNO
  - ANTIPARASITIC PRODUCTS, INS
  - BLOOD AND BLOOD FORMING C
  - CARDIOVASCULAR SYSTEM
  - DERMATOLOGICALS
  - GENITO URINARY SYSTEM AND
  - MUSCULO-SKELETAL SYSTEM
  - + NERVOUS SYSTEM
  - RESPIRATORY SYSTEM
  - SENSORY ORGANS
  - SYSTEMIC HORMONAL PREPAR
  - VARIOUS
- IUPAC-like
- Corpora
- Human miRNA
- Mouse Genes
- MeSH Disease
- @neurIST Ontology
- GO Component
- GO Function
- GO Process

Select Confidence:

1 2 3 4 5

From MEDLINE®, PubMed®, and PubMed Central®,databases of the U.S. National Library of Medicine.
PubMed PubMedCentral

1. Pharmacologic
PubMed 20568832
Statistics | Select ID
Multiple sclerosis (MS) is ar
directly relieve MS symptom
of patients. Nociceptive pair
direct consequence of a les
other symptoms. Central ne
tricyclic antidepressants and
randomized, placebo-contro
treatment perceived an impr
studied pain syndrome. The
antispasticity medications s
often unrecognized by clinic
highlight shortcomings in cli
MeSH: Analgesics therapeu

2. Tuberous scler
PubMed 20845641
and Paediatrics, University of
Statistics | Select ID
BACKGROUND: Tuberous
mental retardation. The aim
primary two pupil of Ijaw trib
for six years. He also had a
has been on Carbamazepin
on the malar area of the fac
tomography showed multipl
in the cortical white matter c
clinic. CONCLUSION: Tube
MeSH: Analgesics; Non-Na
Tuberous Sclerosis complic

3. Voltage-gated
PubMed 20298965
Besta, Milano, Italy.
Statistics | Select ID

SCAIVIEW

Fulltext Query for search, or select saved Searches under blue down arrow

multiple sclerosis

Search | Documents | Entity | Analysis | Help | About

Subcorpus Statistics | Server Statistics

The following entities relating to 'multiple sclerosis' were found in 1244 documents.

Select Columns | Export Table | Export PMIDs | Export Entities | Sentence

131 entities found, displaying 1 to 10.[First/Prev] 1, 2, 3, 4, 5, 6, 7, 8[Next/Last] 10 Entities per Page

Select	Structure Image	Entity	Relative Entropy	Ref. Doc Count	Doc Count	Date Reported	Links
		Choline	0.2579	17936	108	2010-11	SNP D sp
	Fe	Iron	0.2267	90528	161	2010-11-	D
		Carbamazepine	0.2144	9757	81	2010-07-	SNP D sp
	No Structure Available	Calcium	0.2141	206066	217	2010-10-	sp SNP D

- Human Genes / Proteins
- Chromosomal Location
- STS Marker
- non Normalized SNP
- Normalized SNP
- Normalized CRF SNP
- Drug Names
  - ALIMENTARY TRACT AND METAB
  - ANTIINFECTIVES FOR SYSTEMIC
  - ANTINEOPLASTIC AND IMMUNO
  - ANTIPARASITIC PRODUCTS, INS
  - BLOOD AND BLOOD FORMING C
  - CARDIOVASCULAR SYSTEM
  - DERMATOLOGICALS
  - GENITO URINARY SYSTEM AND
  - MUSCULO-SKELETAL SYSTEM
  - + NERVOUS SYSTEM
  - RESPIRATORY SYSTEM
  - SENSORY ORGANS
  - SYSTEMIC HORMONAL PREPAR
  - VARIOUS
- IUPAC-like
- Corpora
- Human miRNA
- Mouse Genes
- MeSH Disease
- @neurIST Ontology
- GO Component
- GO Function
- GO Process

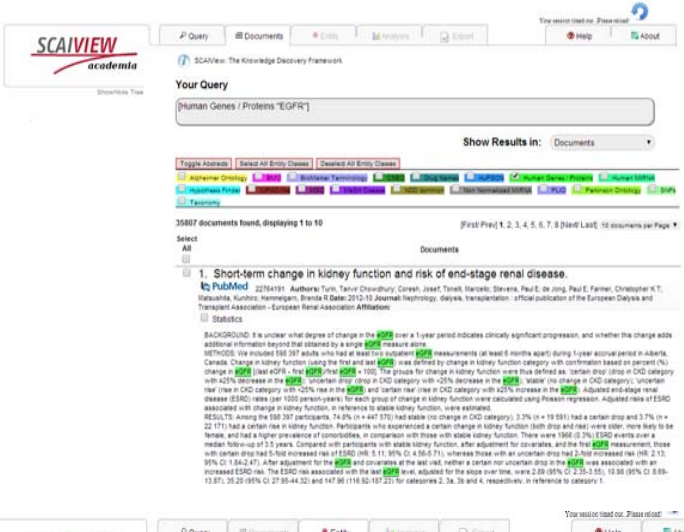
Select Confidence:

1 2 3 4 5

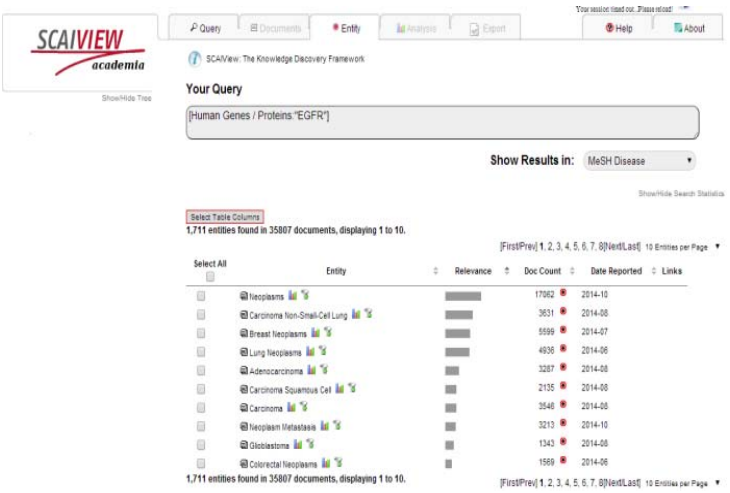
From MEDLINE®, PubMed®, and PubMed Central®,databases of the U.S. National Library of Medicine.
PubMed PubMedCentral

SCAIView functionalities

- **Document View -** Displays all the documents retrieved based on the search query. Entity classes can be selected that you want to highlight. By default, Documents are displayed according the date (newest on the top).

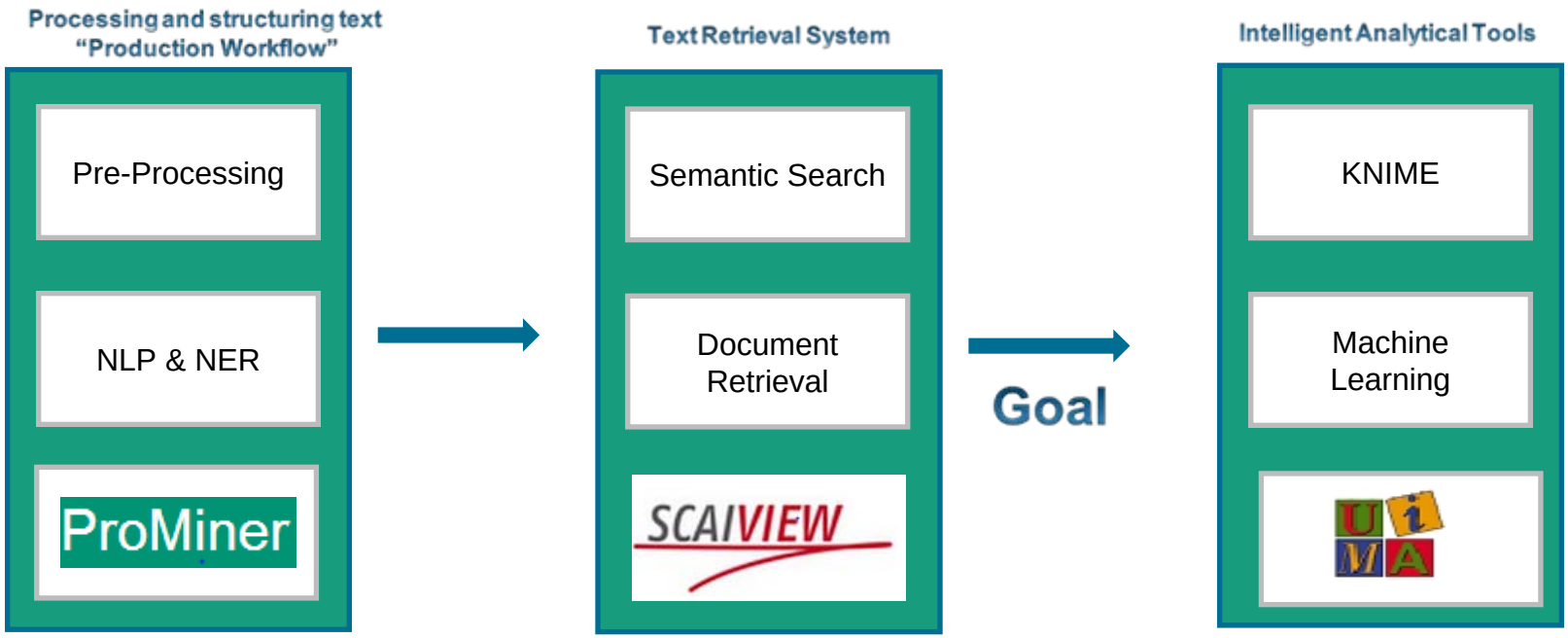


- **Entity View -** Displays named entities under the column entities and are linked to corresponding abstracts



- **Export -** PMID and Entity tables can be exported to text files or excel sheets

Motivation


Processing and structuring text
"Production Workflow"
Pre-Processing
NLP & NER
ProMiner
Text Retrieval System
Semantic Search
Document Retrieval
SCAIVIEW
Goal
Intelligent Analytical Tools
KNIME
Machine Learning
UIMA

UIMA workflow and UI applications

- UIMA (Unstructured Information Management Architecture) is a software architecture for deploying and developing unstructured information management application.
- Originally developed by IBM, now open source.
- Unstructured information application may be defined as a software system designed to analyses large volume of unstructured information in order to discover, organise, and deliver knowledge to the end user
- Thus this architecture provides analytical platform by converting unstructured text to structured information .

UIMA based analysis at SCAI

- BEL like Statement Extraction
- Co-occurrence and Tri-occurrence based relationship extraction
- Machine Learning based relationship extraction
- Topic Modelling
- Term Frequency based Analysis
- …

Capturing Knowledge on Causes and Effects: OpenBEL

Term Expression	Relationship	Term Expression
Subject	Predicate	Object

decreases

```
a(CHEBI:corticosteroid) -| bp(NCI:"Tissue Damage")
```

The abundance of molecules designated by the name "corticosteroid" in the CHEBI namespace.

The biological process designated by the name "Tissue Damage" in the NCI namespace.

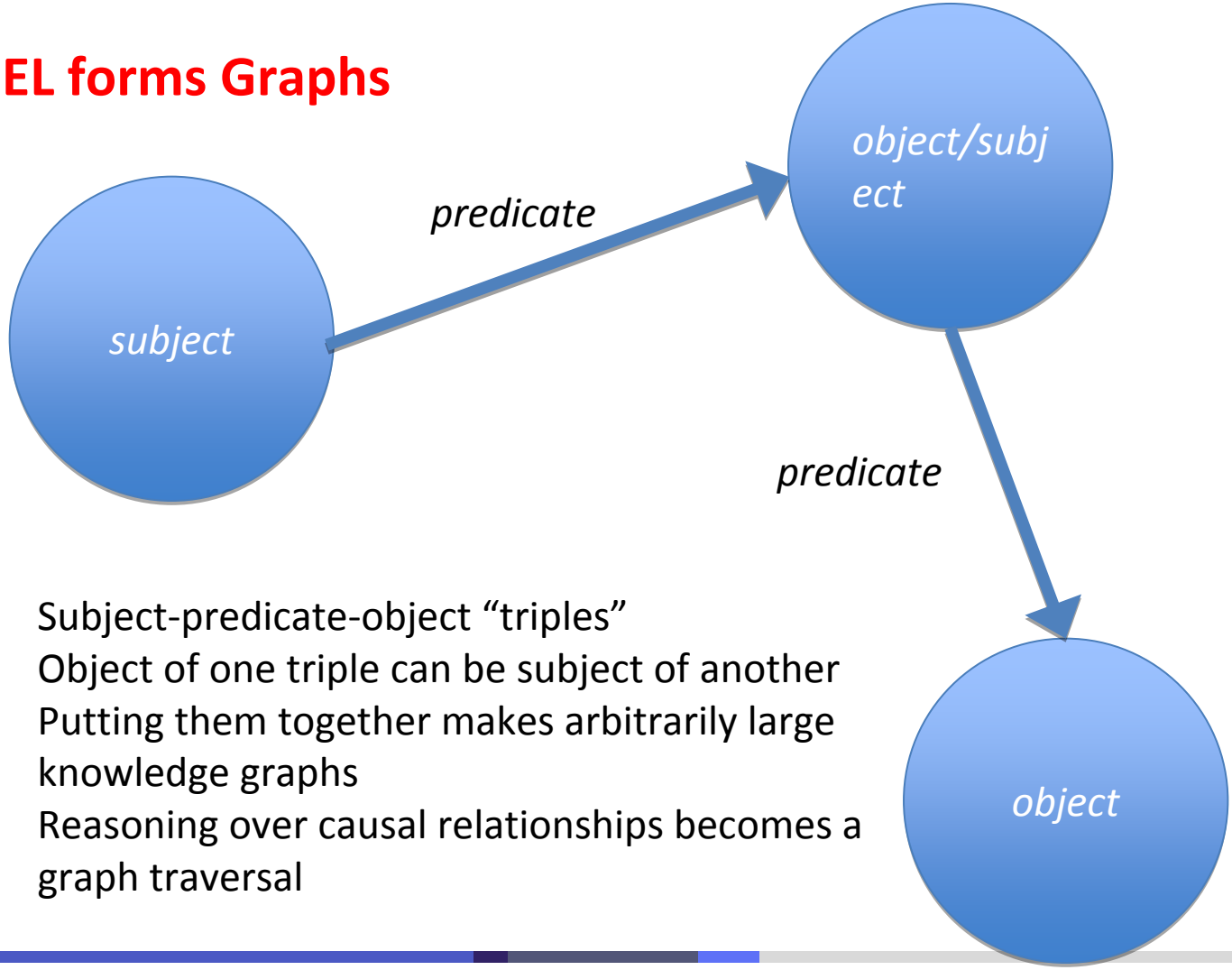
OpenBEL: Capturing of Knowledge and "encoding" of data

Phosphorylation of **glycogen synthase kinase 3beta** at **Threonine, 668** **increases** the **degradation** of **Amyloid precursor protein**.

p (HGNC:GSK3B, pmod (P,T,668)) -> deg (p (HGNC:APP))

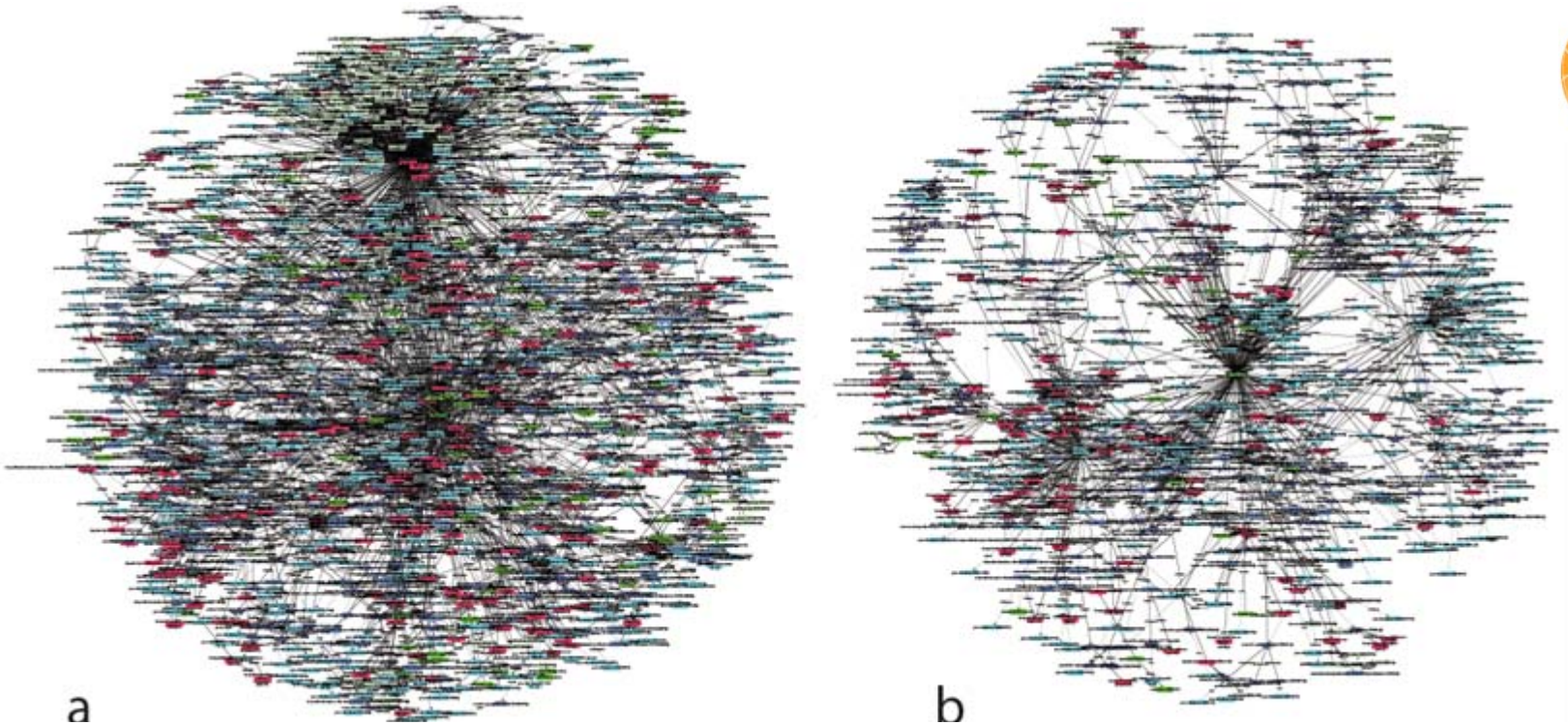
BEL Functions **Namespace Identifires** **Entity Definitions**

BEL forms Graphs



- Subject-predicate-object “triples”
- Object of one triple can be subject of another
- Putting them together makes arbitrarily large knowledge graphs
- Reasoning over causal relationships becomes a graph traversal

The World´s largest Computable Model for Alzheimer´s Disease

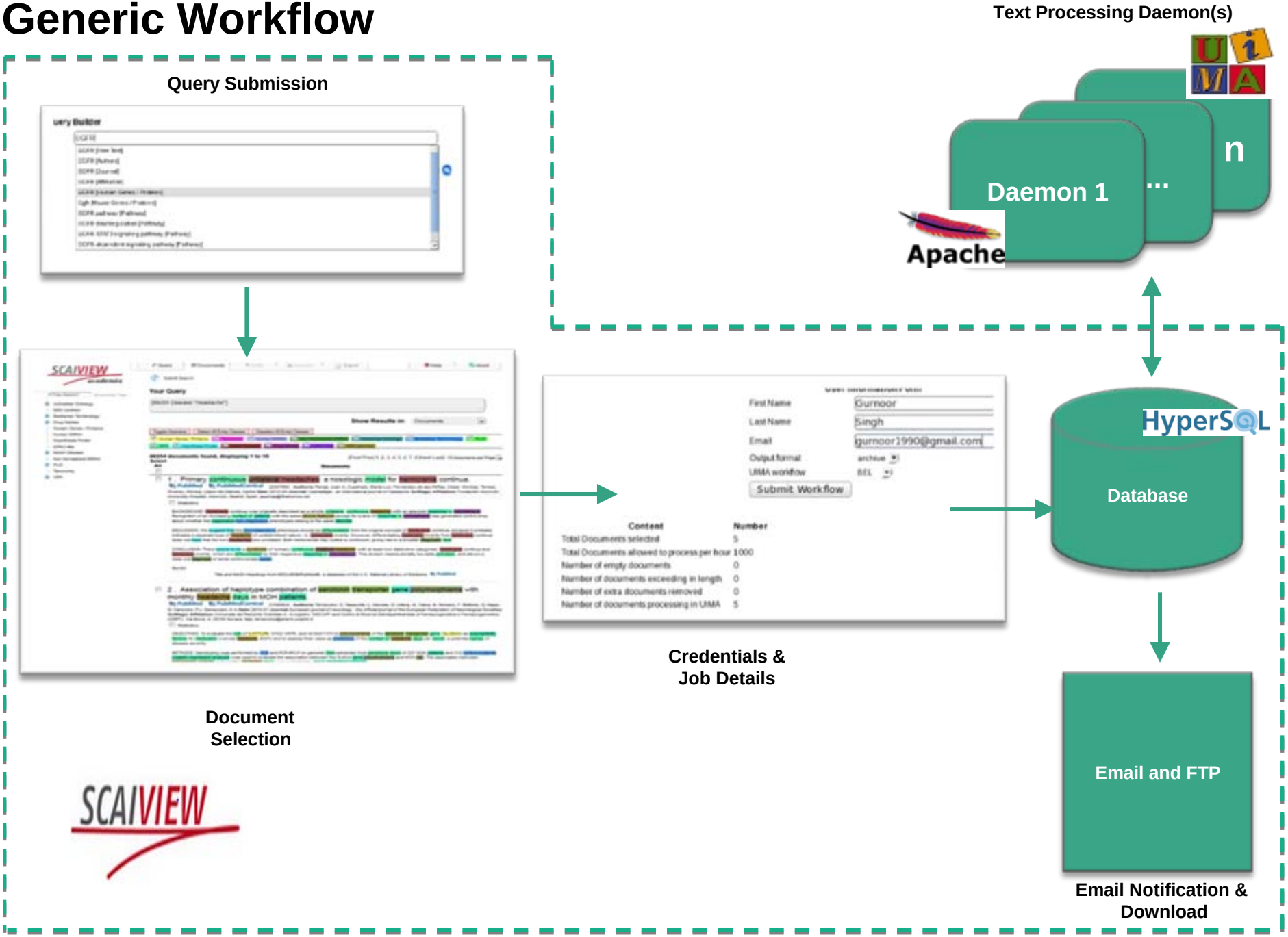


Kodamullil, Alpha Tom, et al. "Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis." Alzheimer's & Dementia (2015).

Project Goals of the Work of Gurnoor Singh

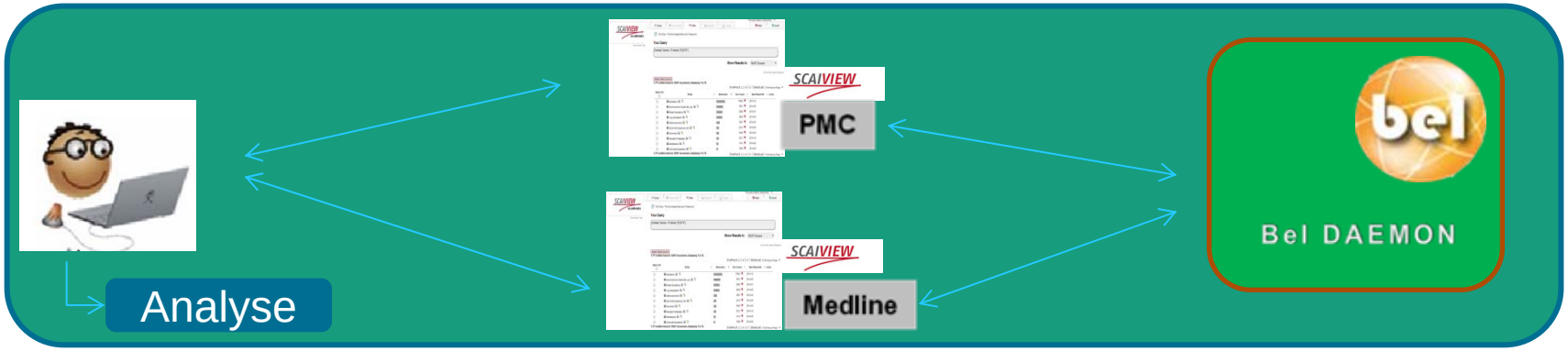
- Implement a solution integrating document retrieval via SCAIView and analytical tools which extracts information/knowledge based on a UIMA workflow
- A generic solution which works well on any analysis workflow(‘wrapper’)
- A well distributed, flexible, and efficient solution for multitasking
- Show application by, performing an exemplary analysis which measures the difference of information gain between abstract representation and full text representation of Biomedical journals

Generic Workflow



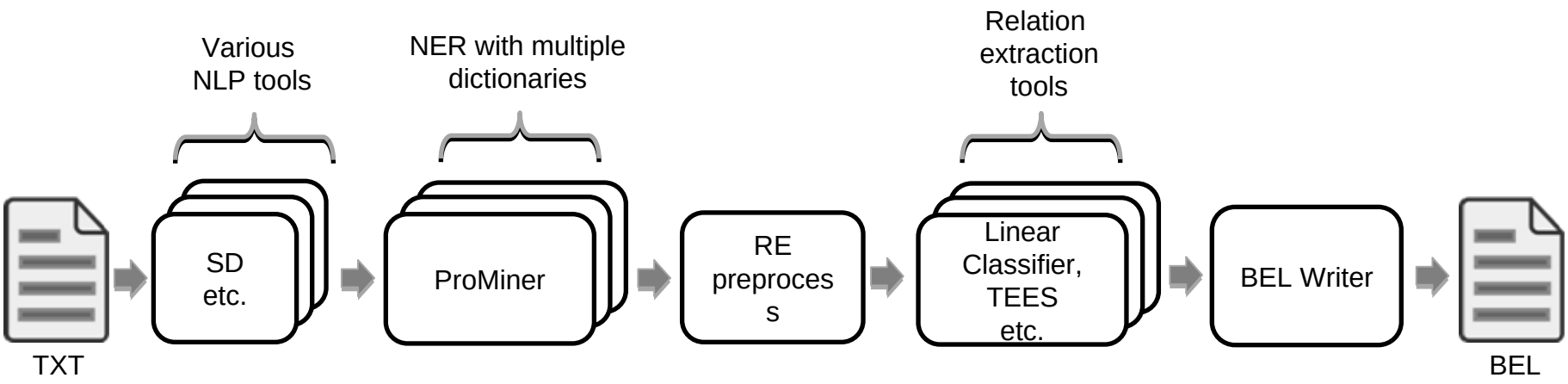
Exemplary Application

- Biological Research Question : “Is there a difference in Information Gain between abstract and full text of a document?”
- Information Gain can be measured as number of unique BEL like statements.
- Install daemon for analysing BEL Like Statement.



- Corpus : Collection of PMID as defined by user query in SCAIView

BELIEF Workflow



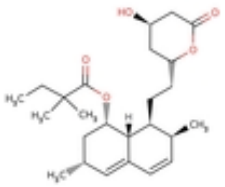
Unstructured
Information Management
Architecture
An Apache Project

Fraunhofer
SCAI

What text mining can deliver – NER & Normalisation

3 . Simvastatin inhibits induction of matrix metalloproteinase-9 in rat alveolar macrophages
moke extract.

Simvastatin:
Simvastatin - A derivative of lovastatin and potent competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase (hydroxymethylglutaryl COA reductases), which is the rate-limiting enzyme in cholesterol biosynthesis



D

Central 19299917 **Authors:** Kim, Sang Eun; Thanh Thuy, Tran Thi; Lee, Ji Hyun; Ro, Jai Youl; Bae, Young An; Kong, ; Oh, Yeon Mock; Lee, Sang Do; Lee, Yun Song **Date:** 2009-0 **Journal:** Experimental & molecular medicine **Affiliation:** y, Sungkyunkwan University School of Medicine, Samsung Biomedical Research Institute, Suwon 440-746, Korea.

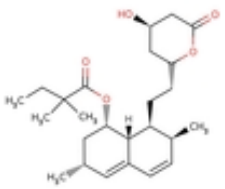
9) may play an important role in emphysematous change in chronic obstructive pulmonary disease (COPD), one of the orbidity worldwide. We previously reported that simvastatin, an inhibitor of HMG-CoA reductase, attenuates emphysematous e lungs of rats exposed to cigarette smoke. However, it remained uncertain how cigarette smoke induced MMP-9 and how ke-induced MMP-9 expression in alveolar macrophages (AMs), a major source of MMP-9 in the lungs of COPD patients. d signaling for MMP-9 induction and the inhibitory mechanism of simvastatin on MMP-9 induction in AMs exposed to solated rat AMs, CSE induced MMP-9 expression and phosphorylation of ERK and Akt. A chemical inhibitor of MEK1/2 or ERK or Akt, respectively, and also inhibited CSE-mediated MMP-9 induction. Simvastatin reduced CSE-mediated MMP-9 d inhibition was reversed by farnesyl pyrophosphate (FPP) or geranylgeranyl pyrophosphate (GGPP). Similar to sferase or GGPP transferase suppressed CSE-mediated MMP-9 induction. Simvastatin attenuated CSE-mediated activation of RAS and phosphorylation of ERK, Akt, p65, IkappaB, and nuclear AP-1 or NF-kappaB activity. Taken together, these results suggest that simvastatin may inhibit CSE-mediated MMP-9 induction, primarily by blocking prenylation of RAS in the signaling pathways, in which Raf-MEK-ERK, PI3K/Akt, AP-1, and IkappaB-NF-kappaB are involved.

Fraunhofer SCAI

What text mining can deliver – NER & Normalisation

3 . Simvastatin inhibits induction of matrix metalloproteinase-9 in rat alveolar macrophages
moke extract.

Simvastatin:
Simvastatin - A derivative of lovastatin and potent competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase (hydroxymethylglutaryl COA reductases), which is the rate-limiting enzyme in cholesterol biosynthesis



Central 19299917 **Authors:** Kim, Sang Eun; Thanh Thuy, Tran Thi; Lee, Ji Hyun; Ro, Jai Youl; Bae, Young An; Kong, ; Oh, Yeon Mock; Lee, Sang Do; Lee, Yun Song **Date:** 2009-0 **Journal:** Experimental & molecular medicine **Affiliation:** y, Sungkyunkwan University School of Medicine, Samsung Biomedical Research Institute, Suwon 440-746, Korea.

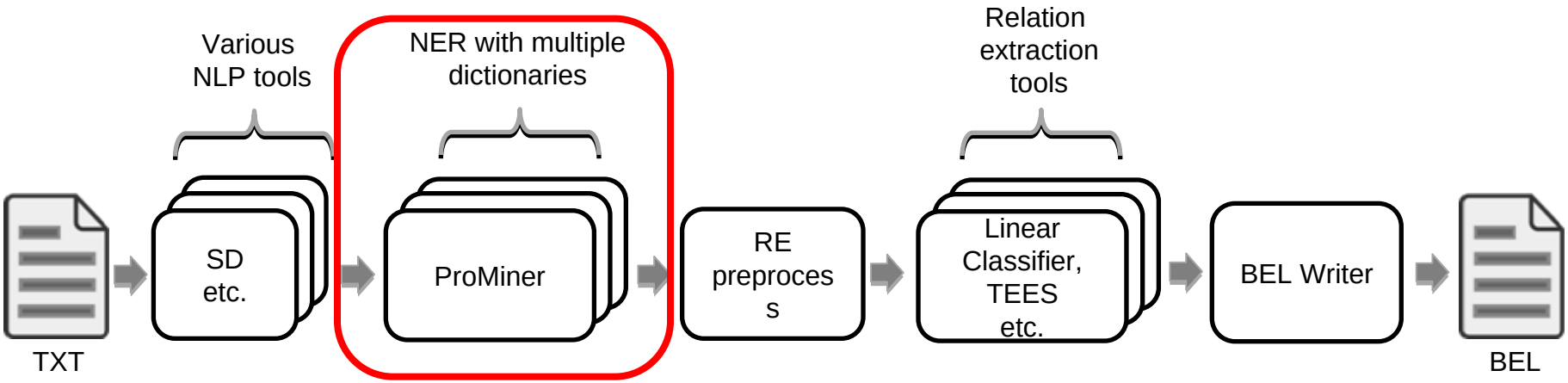
9) may play an important role in emphysematous change in chronic obstructive pulmonary disease (COPD), one of the orbidity worldwide. We previously reported that simvastatin, an inhibitor of HMG-CoA reductase, attenuates emphysematous e lungs of rats exposed to cigarette smoke. However, it remained uncertain how cigarette smoke induced MMP-9 and how ke-induced MMP-9 expression in alveolar macrophages (AMs), a major source of MMP-9 in the lungs of COPD patients. d signaling for MMP-9 induction and the inhibitory mechanism of simvastatin on MMP-9 induction in AMs exposed to solated rat AMs, CSE induced MMP-9 expression and phosphorylation of ERK and Akt. A chemical inhibitor of MEK1/2 or ERK or Akt, respectively, and also inhibited CSE-mediated MMP-9 induction. Simvastatin reduced CSE-mediated MMP-9 d inhibition was reversed by farnesyl pyrophosphate (FPP) or geranylgeranyl pyrophosphate (GGPP). Similar to sferase or GGPP transferase suppressed CSE-mediated MMP-9 induction. Simvastatin attenuated CSE-mediated activation of RAS and phosphorylation of ERK, Akt, p65, IkappaB, and nuclear AP-1 or NF-kappaB activity. Taken together, these results suggest that simvastatin may inhibit CSE-mediated MMP-9 induction, primarily by blocking prenylation of RAS in the signaling pathways, in which Raf-MEK-ERK, PI3K/Akt, AP-1, and IkappaB-NF-kappaB are involved.

- Recall and Precision rates are between 70% and 90% for biomedical NER

- Recall: How many of the existing names does the system detect
- Precision: How many of the detected names are correct

Fraunhofer SCAI

BELIEF Workflow



Fraunhofer
SCAI

Current dictionaries included

Entity class	Resources	BEL namespace
Human Genes/Proteins	EntrezGene/ Uniprot	HGNC
Mouse Genes/Proteins	EntrezGene/ Uniprot	MGI
Rat Genes/Proteins	EntrezGene/ Uniprot	RGD
Protein family names	OpenBEL	PFH
Protein complex names	OpenBEL	NCH
Protein complex names	Gene Ontology	GOCCTERM
Chemical names	OpenBEL	SCHEM
Chemical names	ChEBI	CHEBI
Chemical names	ChEMBL	SCHEM
Disease names	MeSH	MESHD
Anatomy names	MeSH	MESHA

Use case: relation between small molecules (mainly protein inhibitors) and their targets

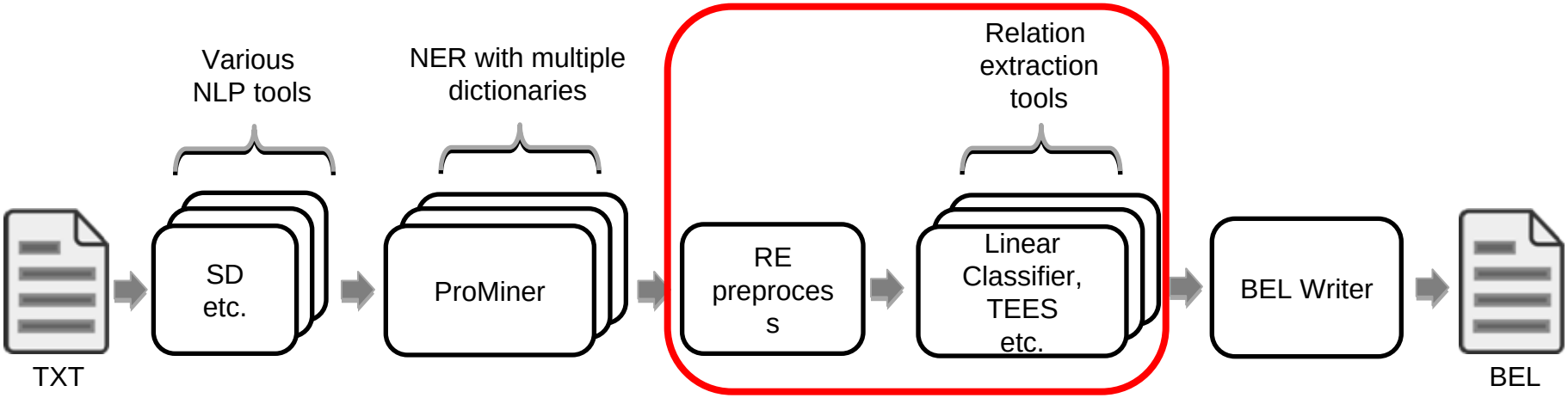
Dictionary	Recall rate initial version	Recall rate application adapted
Genes/Protein: (HGNC)	80 %	93 %
Chemical compounds: ChEBI	15 %	66 %
Chemical compounds: SCHEM	30 %	75 %
Chemical compounds: ChEBI + SCHEM+ ChEMBL	not determined	91 %
Selventa-human-complex	40 %	46 %
GO-Complex	not determined	64 %
Selventa-human-complex + Complex	not determined	82 %
GO-Function	22 %	not determined
Selventa-human-families	8 %	77 %

Recognition and normalization of terminology

- Normalisation is needed!
- Use external and internal (OpenBEL) resources for named entity recognition (Mapping!)
- Combine various resources
- Adapt terminology to use cases (OpenBel namespaces provide no synonyms)
- Offer curators the annotation of different concepts

For relation extraction high recall is a precondition!!!

BELIEF Workflow



Fraunhofer
SCAI

Relation Extraction

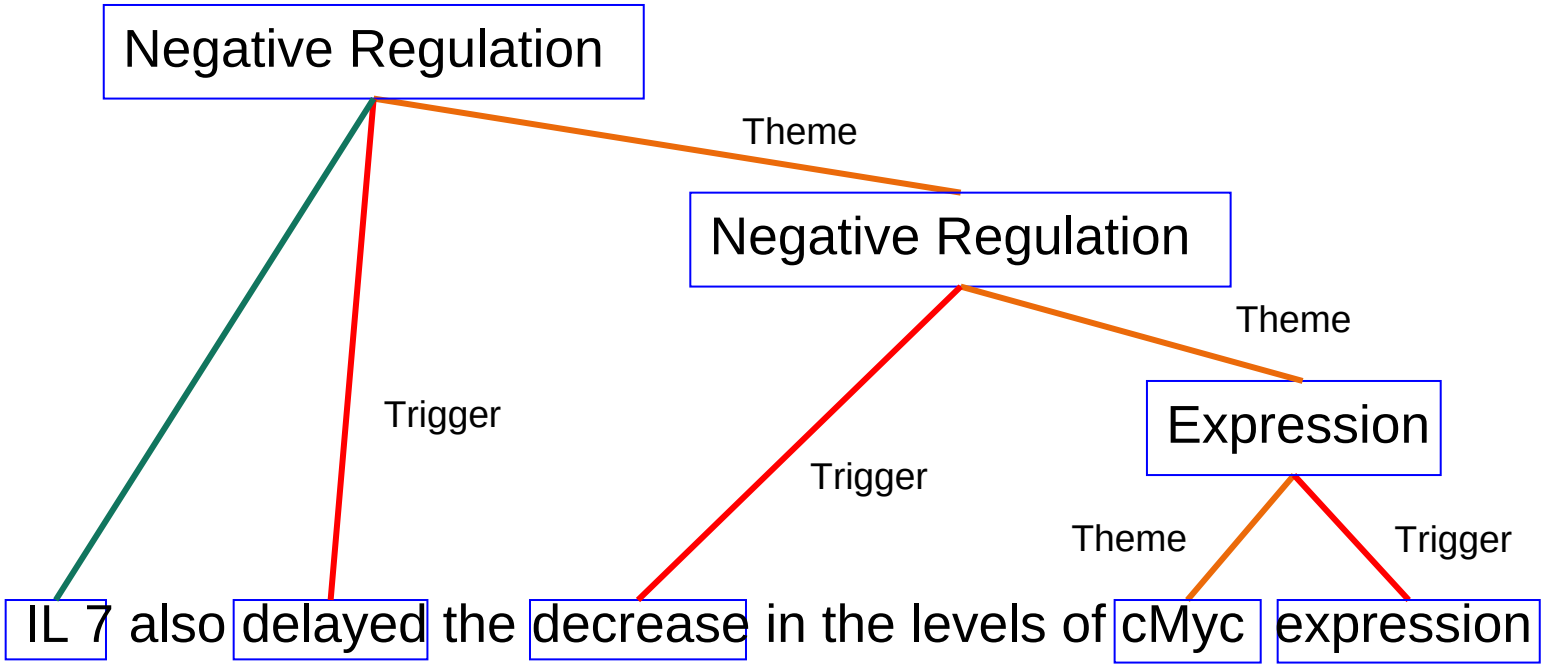
Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL

Relation Extraction

Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL



Relation Extraction

Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL

- Simpler binary classification:

  IL 7 also delayed the decrease in the levels of cMyc ….

  IL7 – cMyc Relation: Yes

Classifies if a relation between 2 entities is existing but gives no information about the direction or type

Technology – Performance

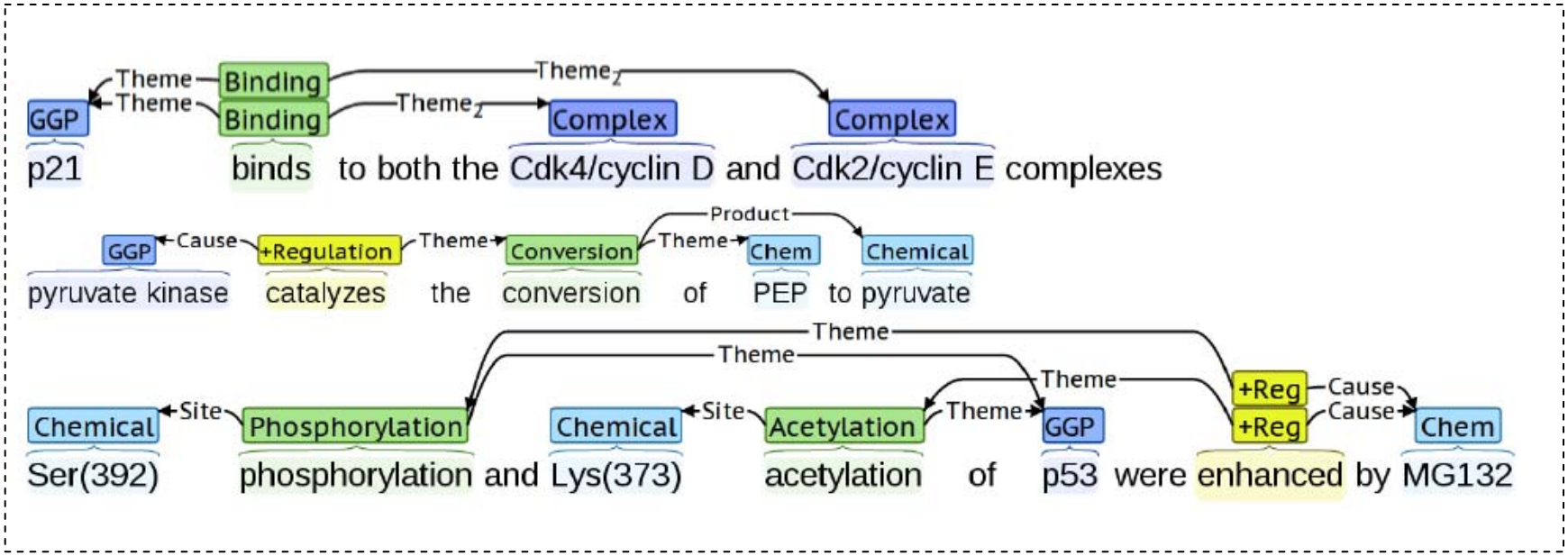
- NLP (Sentence Detection ~6% error) 94
- NLP (Tokenization ~8% error) 86
- NER (Different Classes ~15% error) 73
- RelationExtraction (Multi-step ~25% error) 54

complexity

Propagated error!

Fraunhofer SCAI

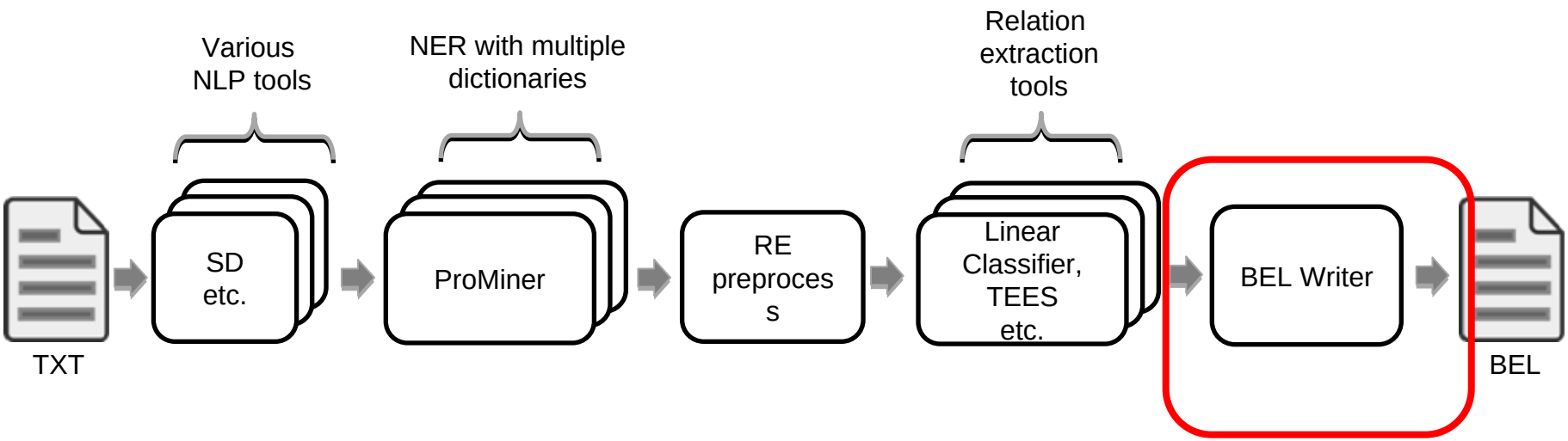
What text mining can deliver: Relation Extraction – Example from BioNLP shared Task



- Recall ~30% and Precision ~50% for regulation events – binary classification has higher recall and precision rates

That seems not very promising but many relations might be redundant!

BELIEF Workflow



Fraunhofer
SCAI

BioNLP Shared Task to openBEL conversion[1]

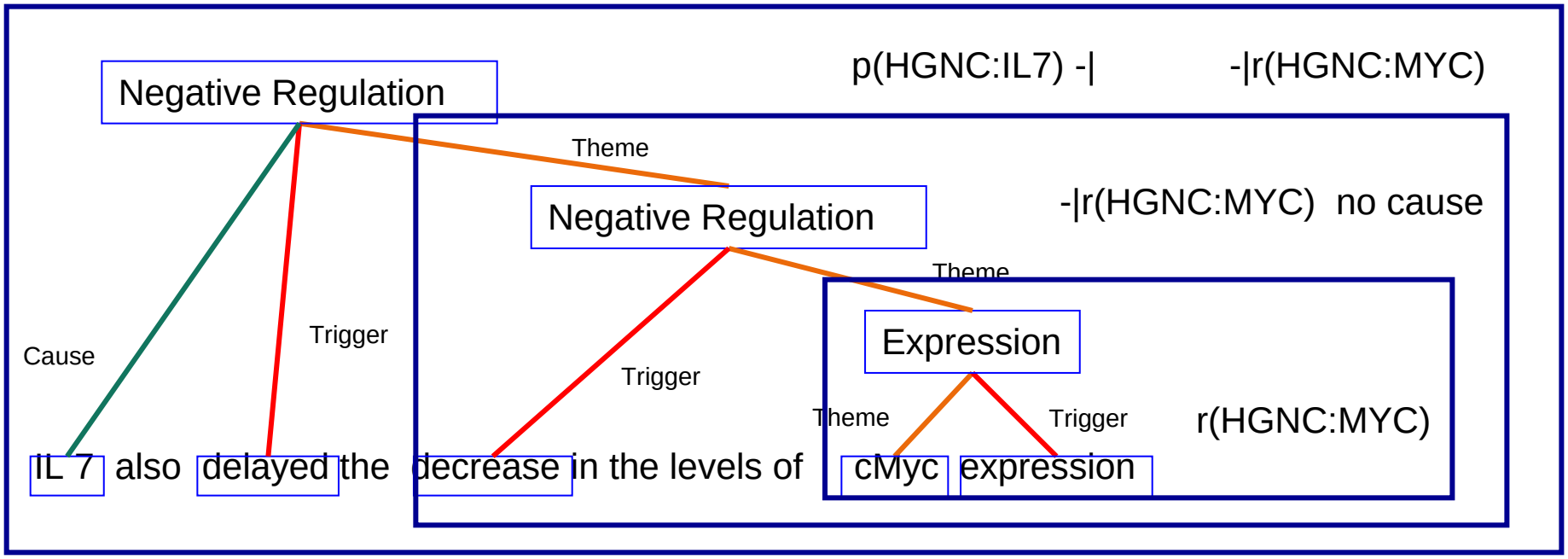
We implemented a rule set translating BioNLP SharedTask to BEL

Abundances	p(),..,
Modifications	pmod()
Catabolism	deg()
Location	tloc()
Binding/Complex	complex()

Positive Regulation	->
Negative Regulation	-|
Regulation/Association	- -

[1]**BEL networks derived from qualitative translations of BioNLP Shared Task annotations.**
The Association for Computational Linguistics (ACL) Sofia 2013
Fluck, J.; Klenner, A.; Madan, S.; Ansari, S.; Bobic, T.; Hoeng, J.; Hofmann-Apitius, M.; Peitsch, M.;

Relation extraction example result



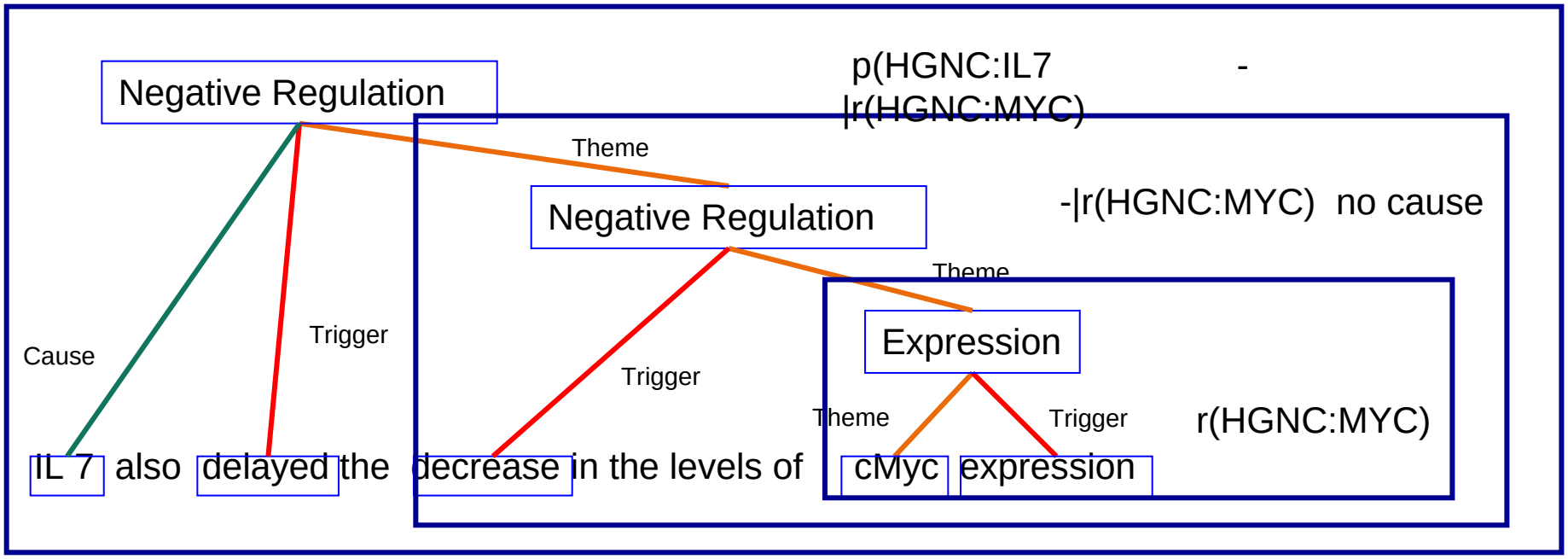
Fraunhofer SCAI

Relation extraction example result

Automatic extension to full statements in workflow:

Fixme -|r(HGNC:MYC

p(HGNC:IL7) -| Fixme -|r(HGNC:MYC)



Fraunhofer SCAI

Relation extraction example result

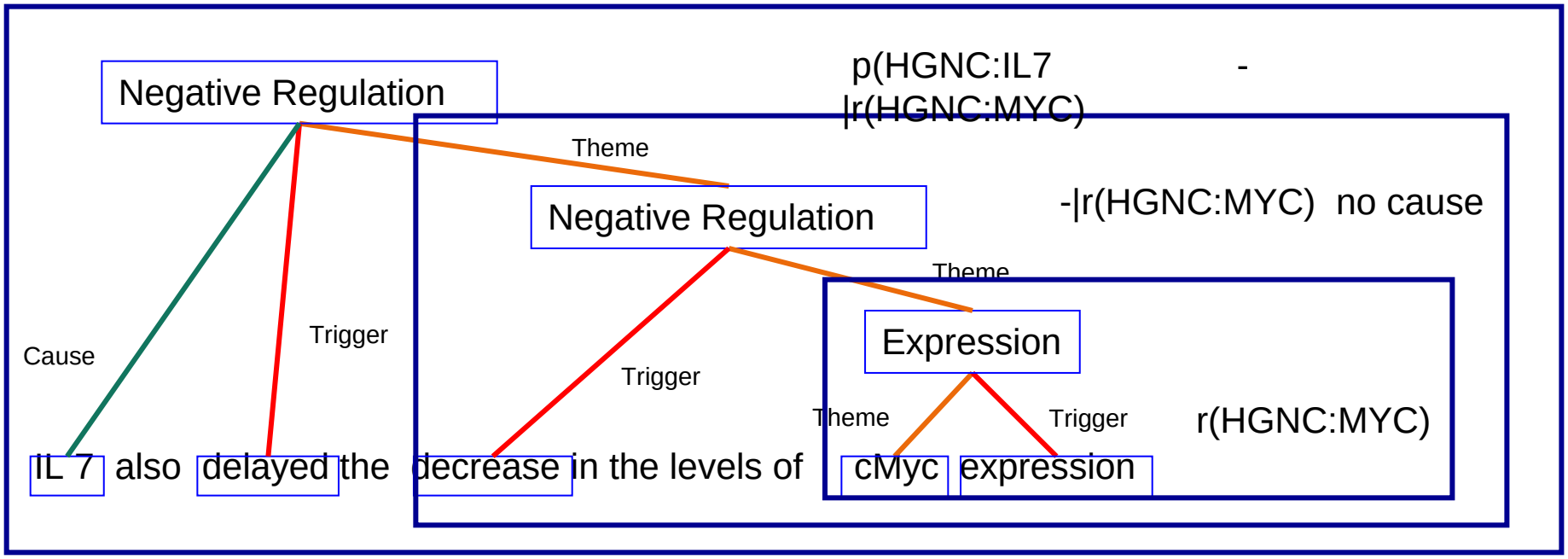
Automatic extension to full statements in workflow:

Fixme -|r(HGNC:MYC)

p(HGNC:IL7) -| Fixme -|r(HGNC:MYC)

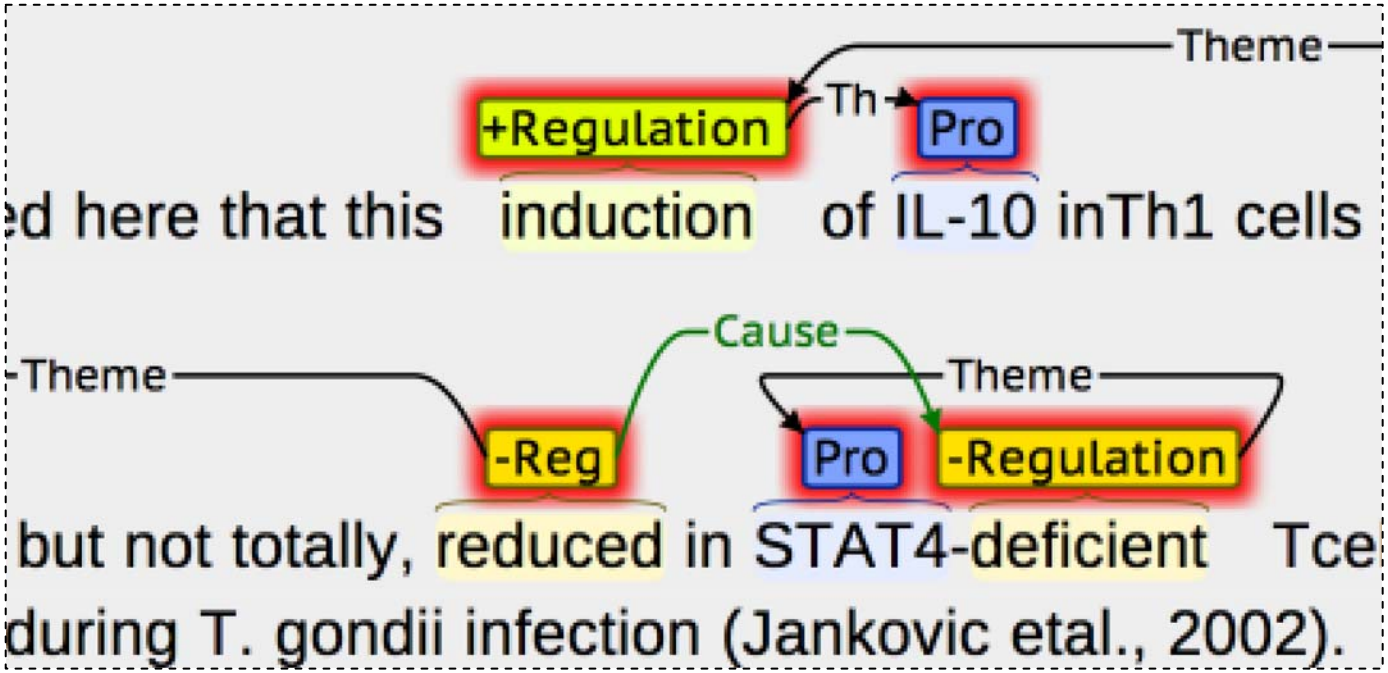
Binary classification:

p(HGNC:IL7) -- r(HGNC:MYC)



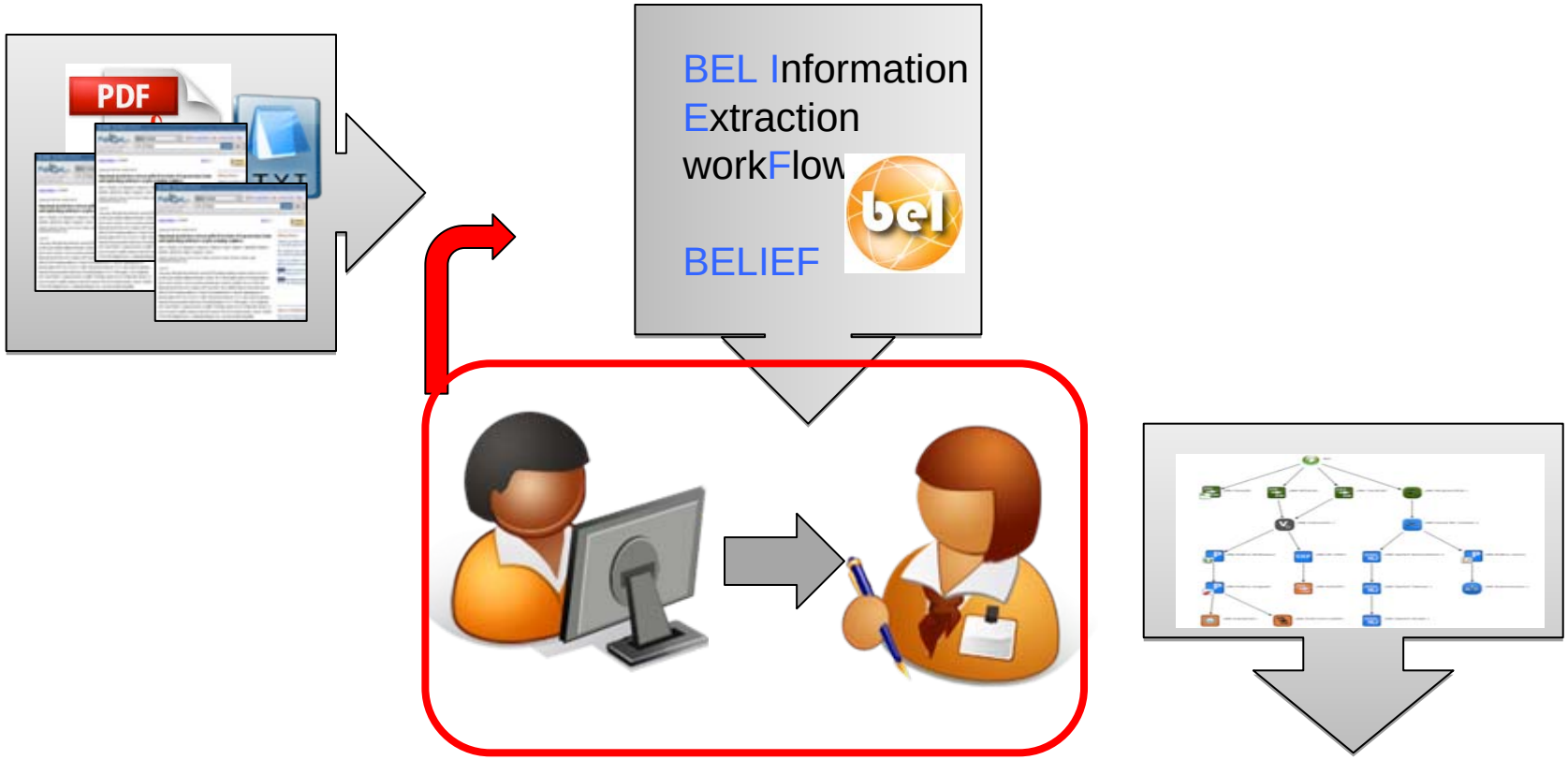
Fraunhofer SCAI

What text mining currently does not deliver: Interpretations



- Generated Statement:
  - p(FIXME) -| p(HGNC:STAT4) -| p(FIXME) -> p(HGNC:IL10)
- Manual Statement:
  - p(HGNC:STAT4) -> p(HGNC:IL10)

Semi automatic BEL Knowledge Extraction Pipeline

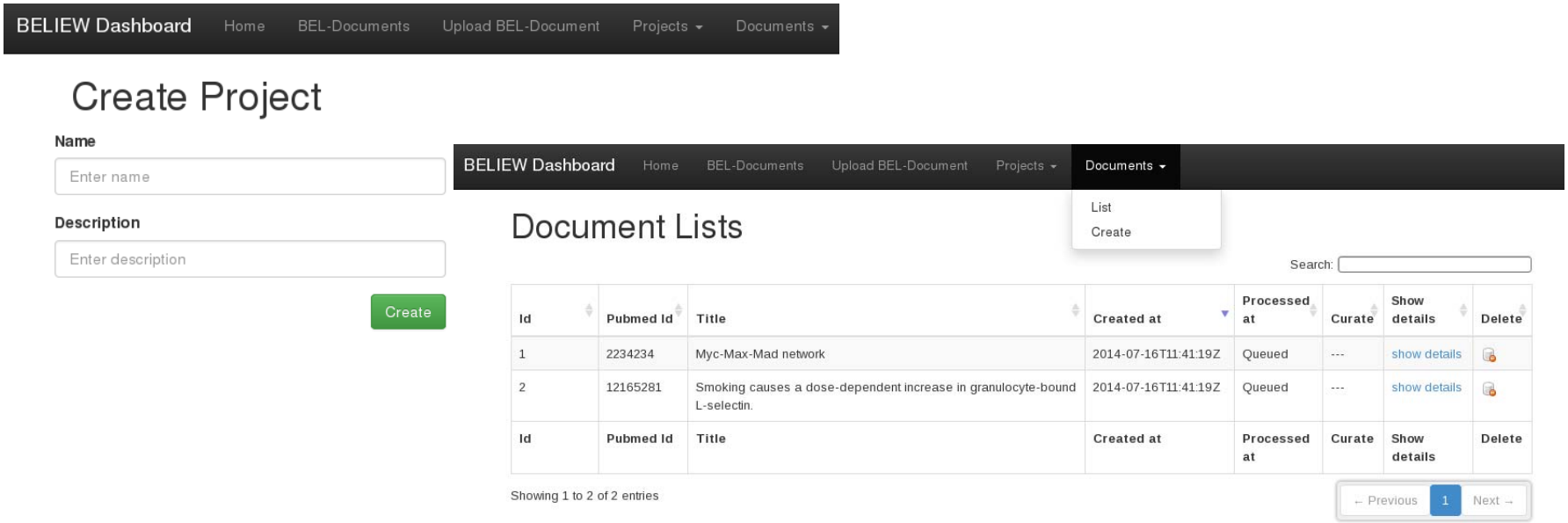


Application

BeliefDashboard

Projects management and multiple document upload

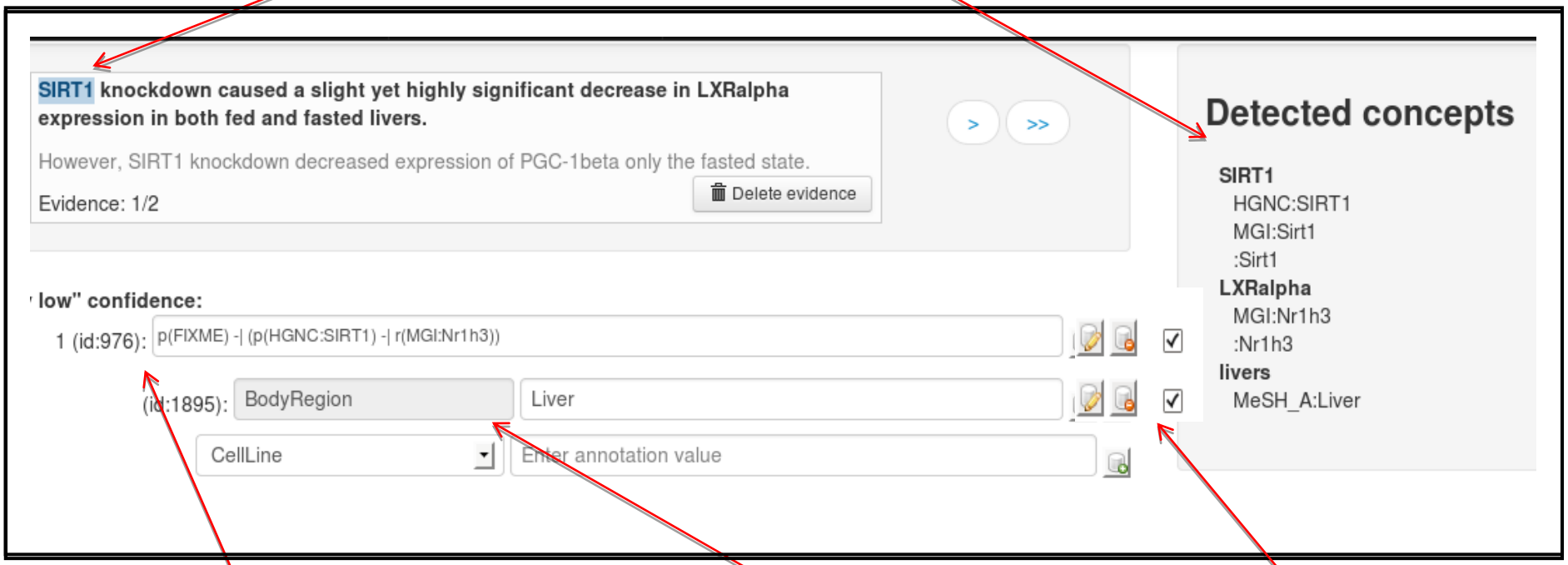
- Create, delete and list functionality



Fraunhofer SCAI

BELIEF Dashboard Curation Interface

Detected concepts:
-Highlighted text entity in mouse over
-Provide a fast overview of all entities in the evidence



BEL statement

Context annotation

Edit/Delete/Export

Fraunhofer SCAI

BEL Editor – Update document information

Document information

Pubmed Id: 11350768 Update

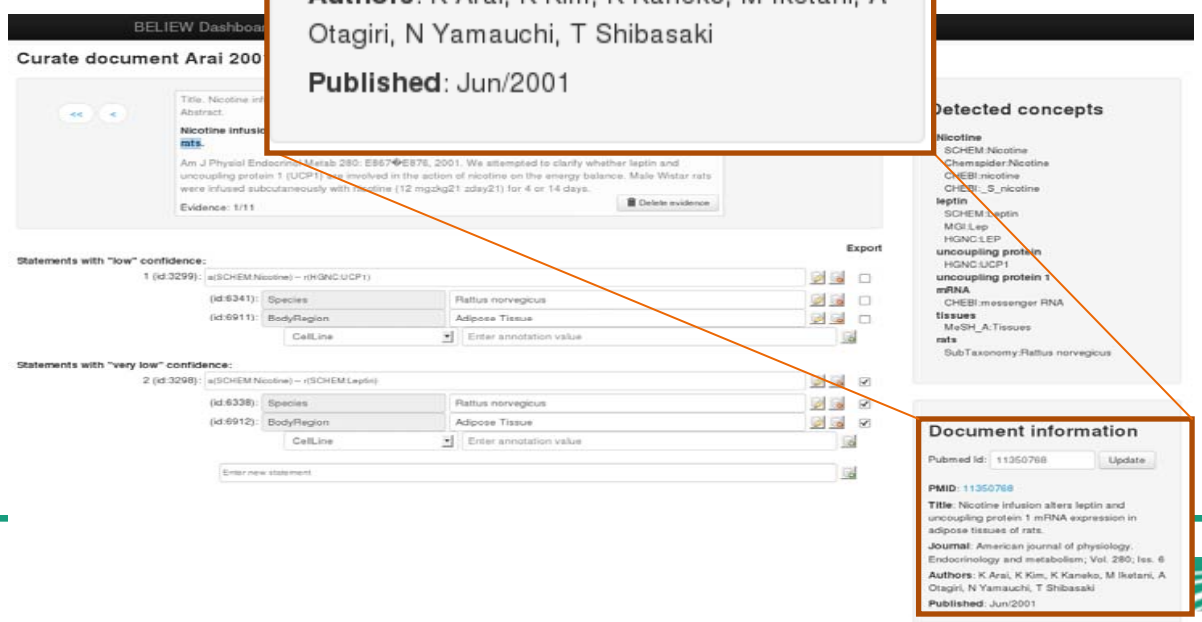
PMID: 11350768

Title: Nicotine infusion alters leptin and uncoupling protein 1 mRNA expression in adipose tissues of rats.

Journal: American journal of physiology. Endocrinology and metabolism; Vol. 280; Iss. 6

Authors: K Arai, K Kim, K Kaneko, M Iketani, A Otagiri, N Yamauchi, T Shibasaki

Published: Jun/2001

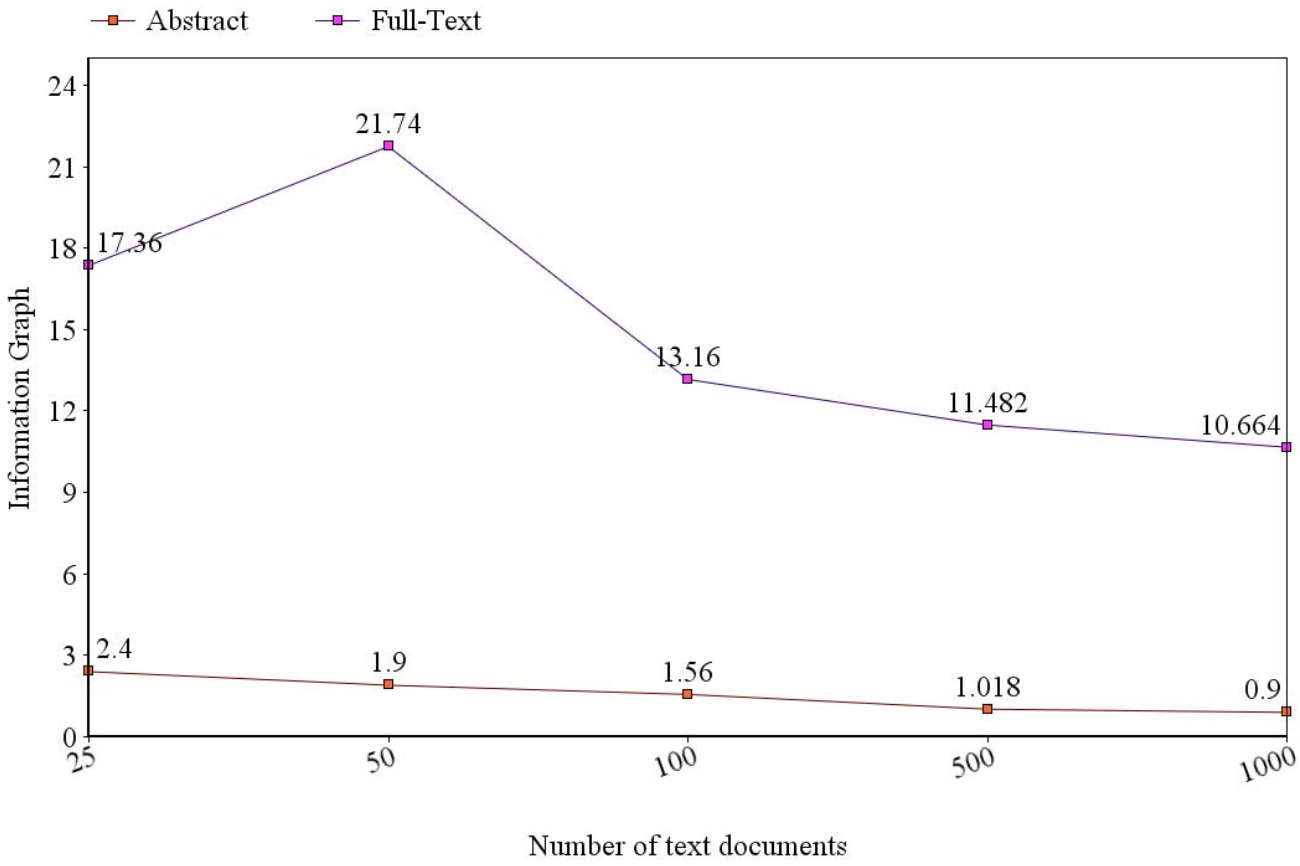


Fraunhofer SCAI

Back to Gurnoor Singh: Experimental Setup

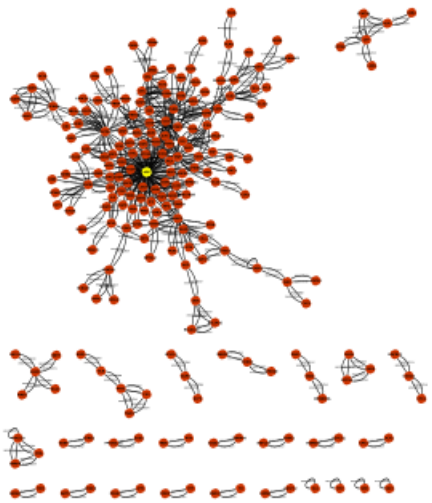
- Query disease under study: Alzheimer's Disease (AD)
- A total 10 jobs selecting top 25, 50, 100, 500, and 1000 text documents from both the SCAIView systems were exported to BEL processing daemon.
- BEL documents retrieved via SCAIview were analyzed using KAM navigator and Cytoscape
- Biological networks were further narrowed down to Protein-Protein Interactions networks

Information Graphs



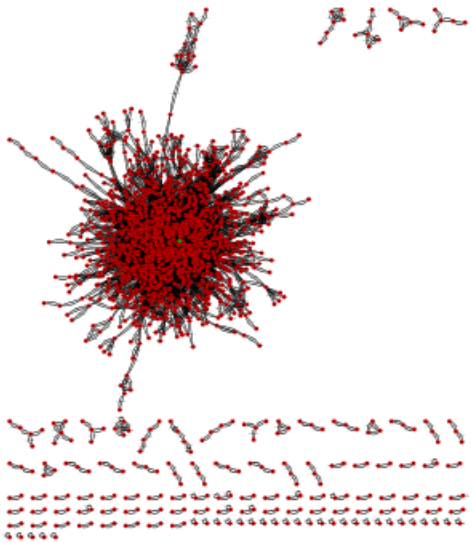
Graph Topology

500- Abstract



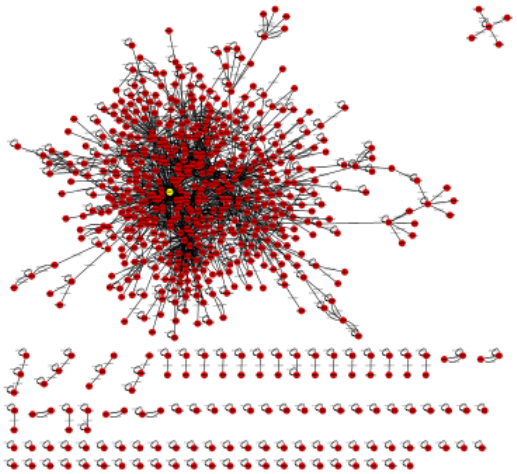
Documents: 500
Nodes: 215
Edges: 509

500- Full Text



Documents : 500
Nodes : 1490
Edges : 5741

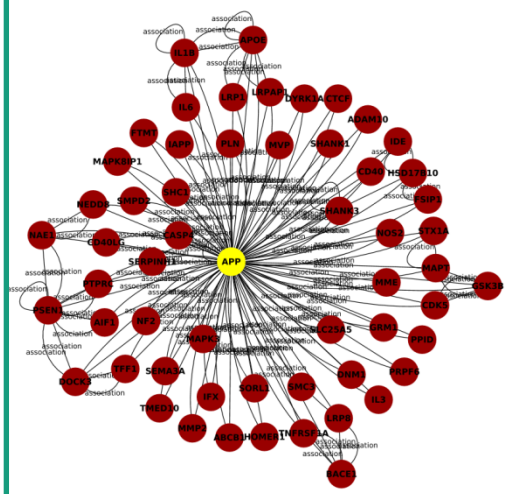
Manually Curated



Documents : 500
Nodes : 553
Edges : 3525

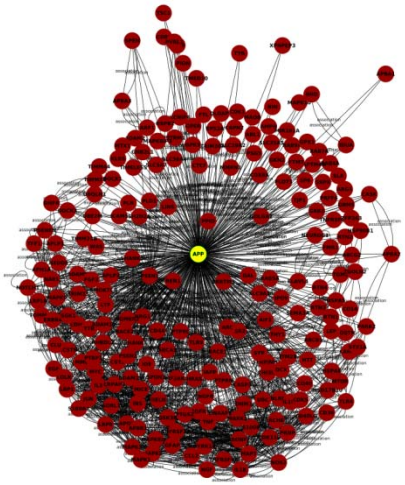
Plotting APP (yellow) and its First Neighbors

500- Abstract



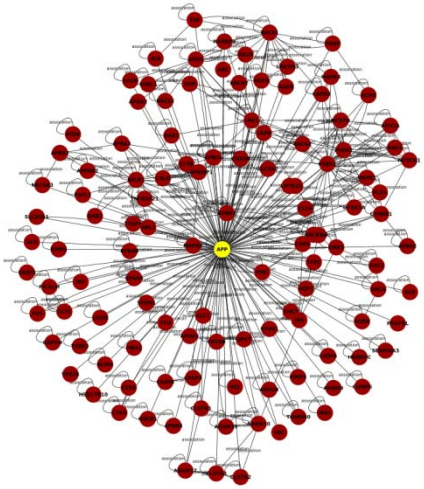
Nodes: 58

500- Full Text



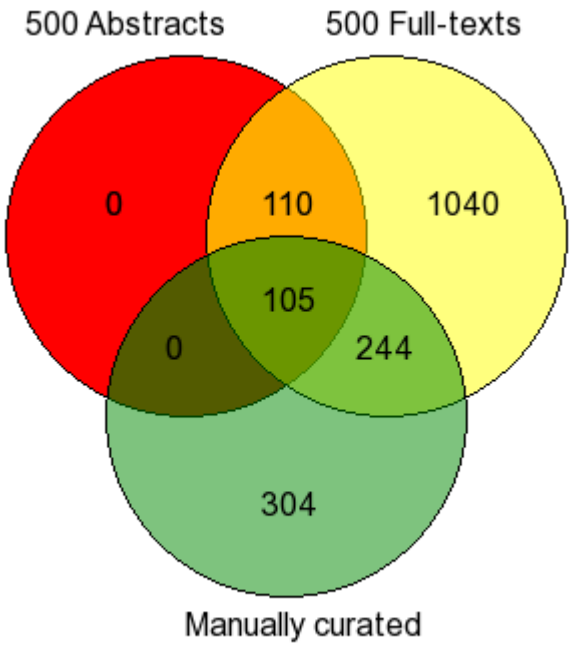
Nodes : 230

Manually Curated

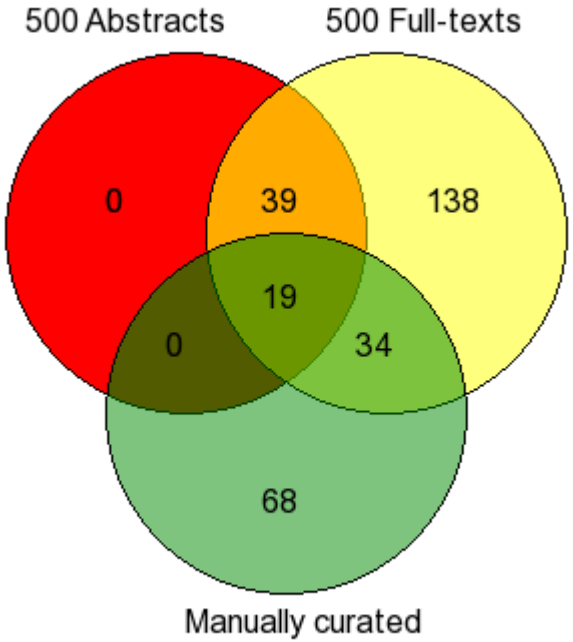


Nodes : 121

Venn Diagrams



All triples



APP triples

Text Mining & Decision Support

- **Our cancer patient may die, because:**
  - Nobody can read all the papers that contain relevant information in only two weeks
  - The publishing industry does not permit machines to do the job
- **How will we reason over genetic variation information in a functional context?**
  - See Naz et al., Briefings in Bioinformatics, in press
- **Time for a "GRAND CHALLENGE"**
  - Let us work together to organise a Grand Challenge that demonstrates how our cancer patient could be saved if automated text mining would be supported by the publishing industry

Take home message …..

Innovative Text Mining and Decision Support

- We can use unstructured text like any database
- We can extract useful and interesting facts, such as triples that represent causal relationships in biomedicine
- We can use these semi-automated information extraction processes to generate a knowledge base in languages such as BEL
- In the future, such knowledge bases will enable decision support in life saving, time-critical scenarios