Text Mining

Sprach- und Wissenstechnologien am Beispiel der Lebenswissenschaften

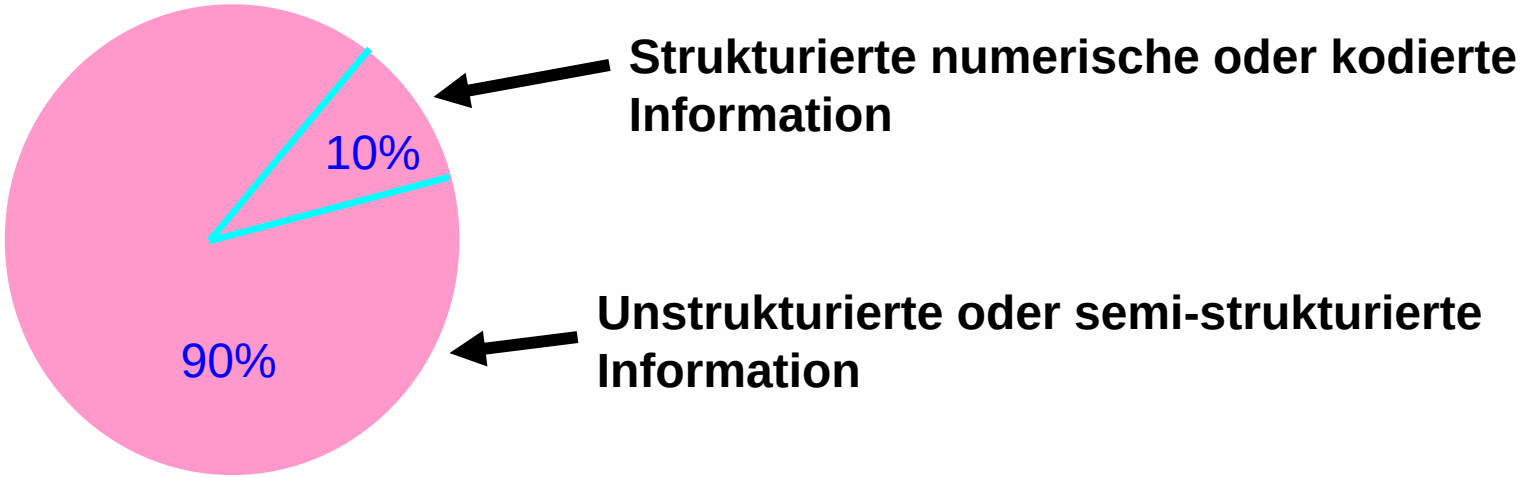
Udo Hahn Joachim Wermter

Jena University
Language and Information Engineering (JULIE) Lab

Jena University Language & Information Engineering Lab

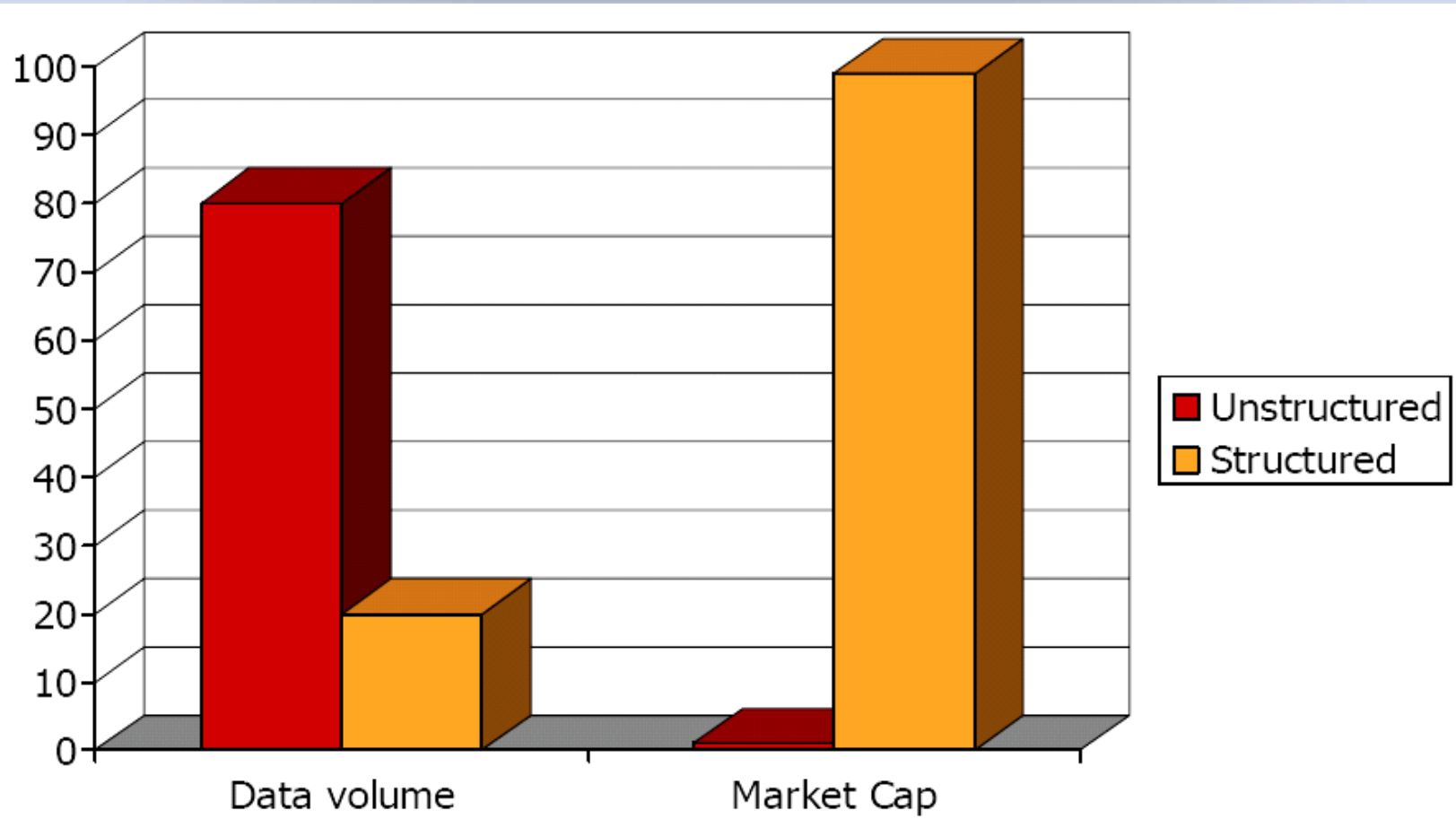
Warum ist Text Mining Relevant? (1/4)

90:10 Argument



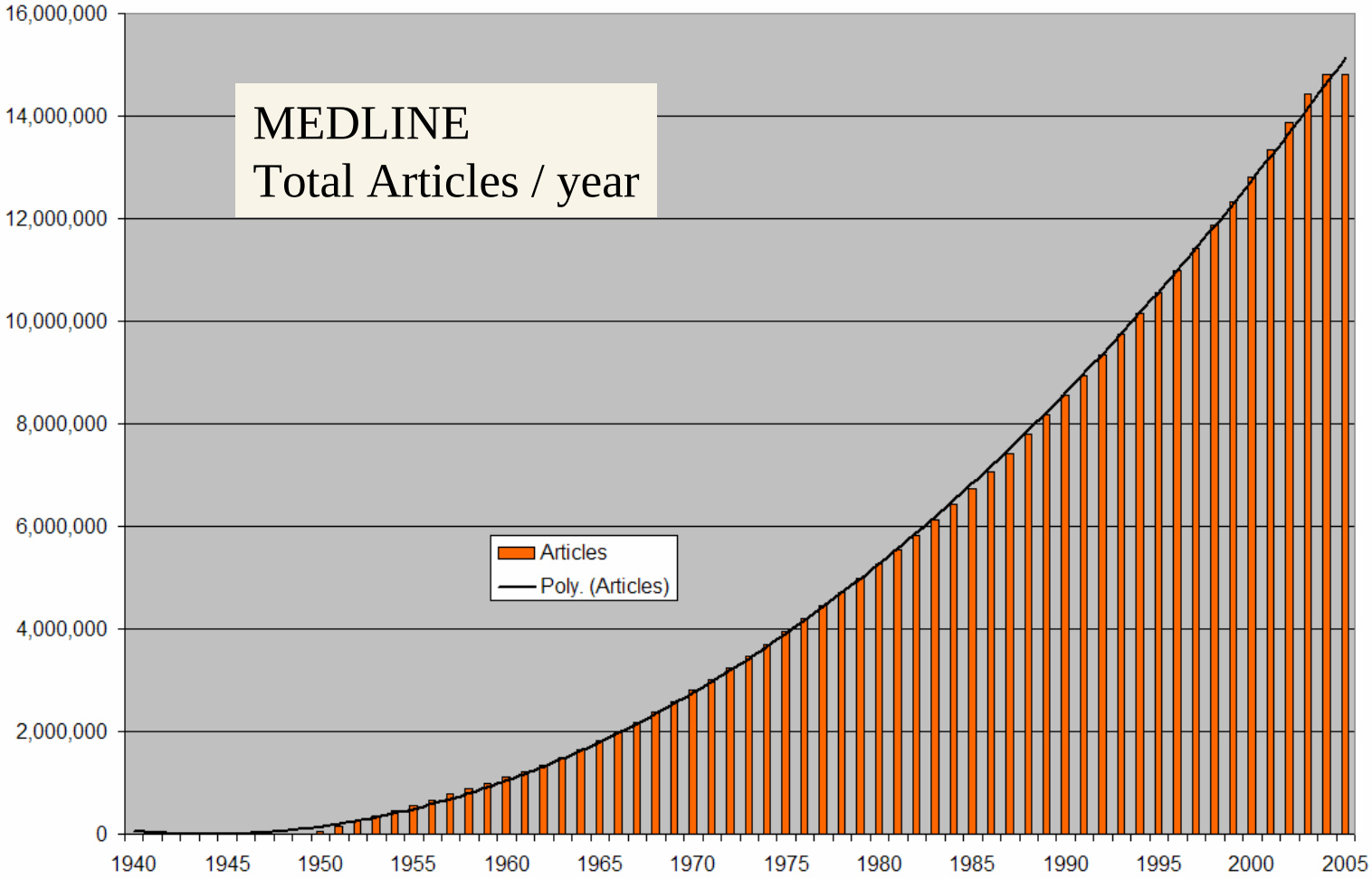
Quelle: Oracle Corp.

Warum ist Text Mining Relevant? (2/4)



Quelle: Prabhakar & Raghavan, Verity (2002)

Warum ist Text Mining Relevant? (3/4)



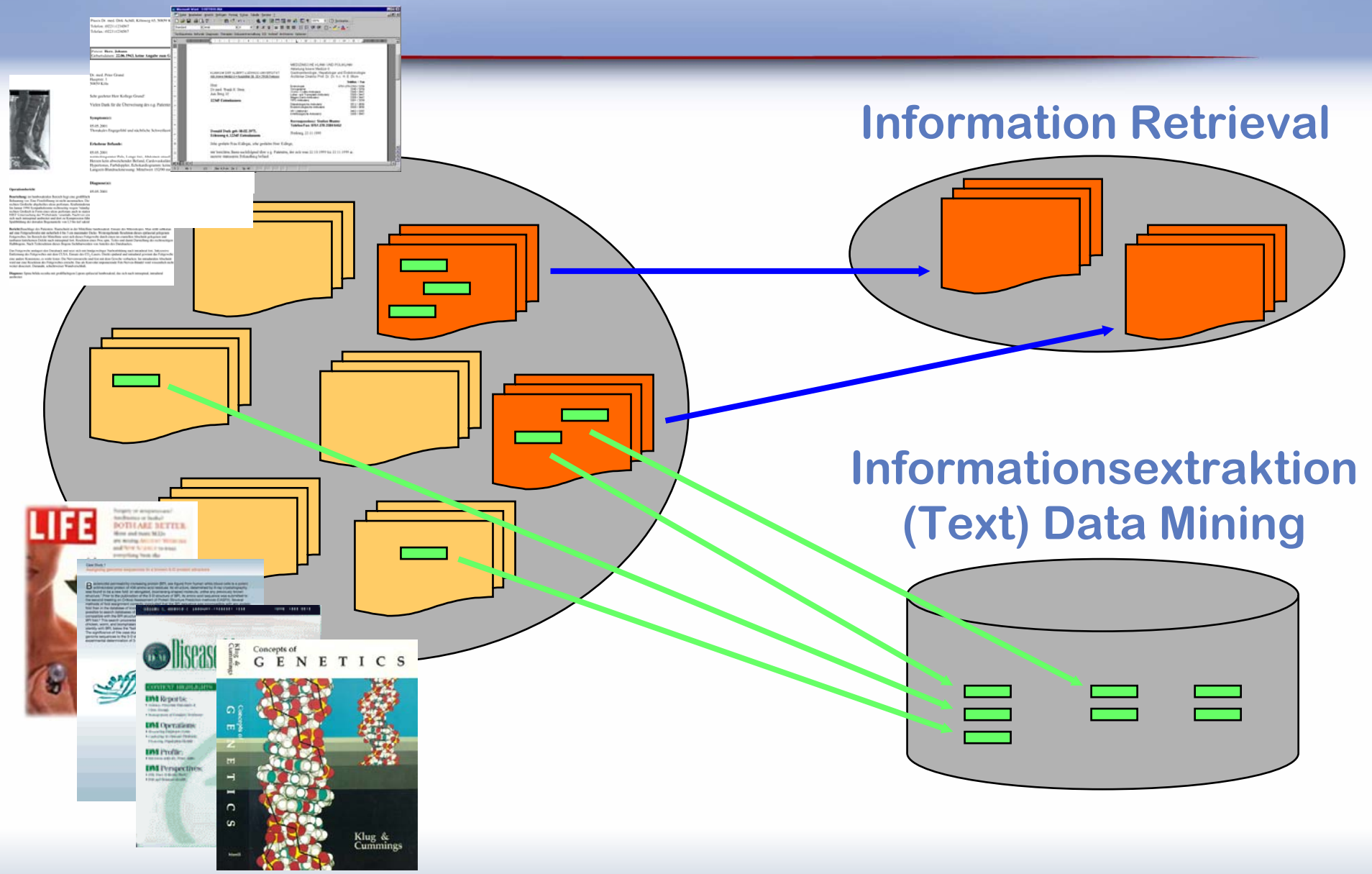
Warum ist Text Mining Relevant? (4/4)

"About a quarter of late stage failures we surveyed could have been eliminated two years earlier by making all external and internal information in the form of text documents more widely available."

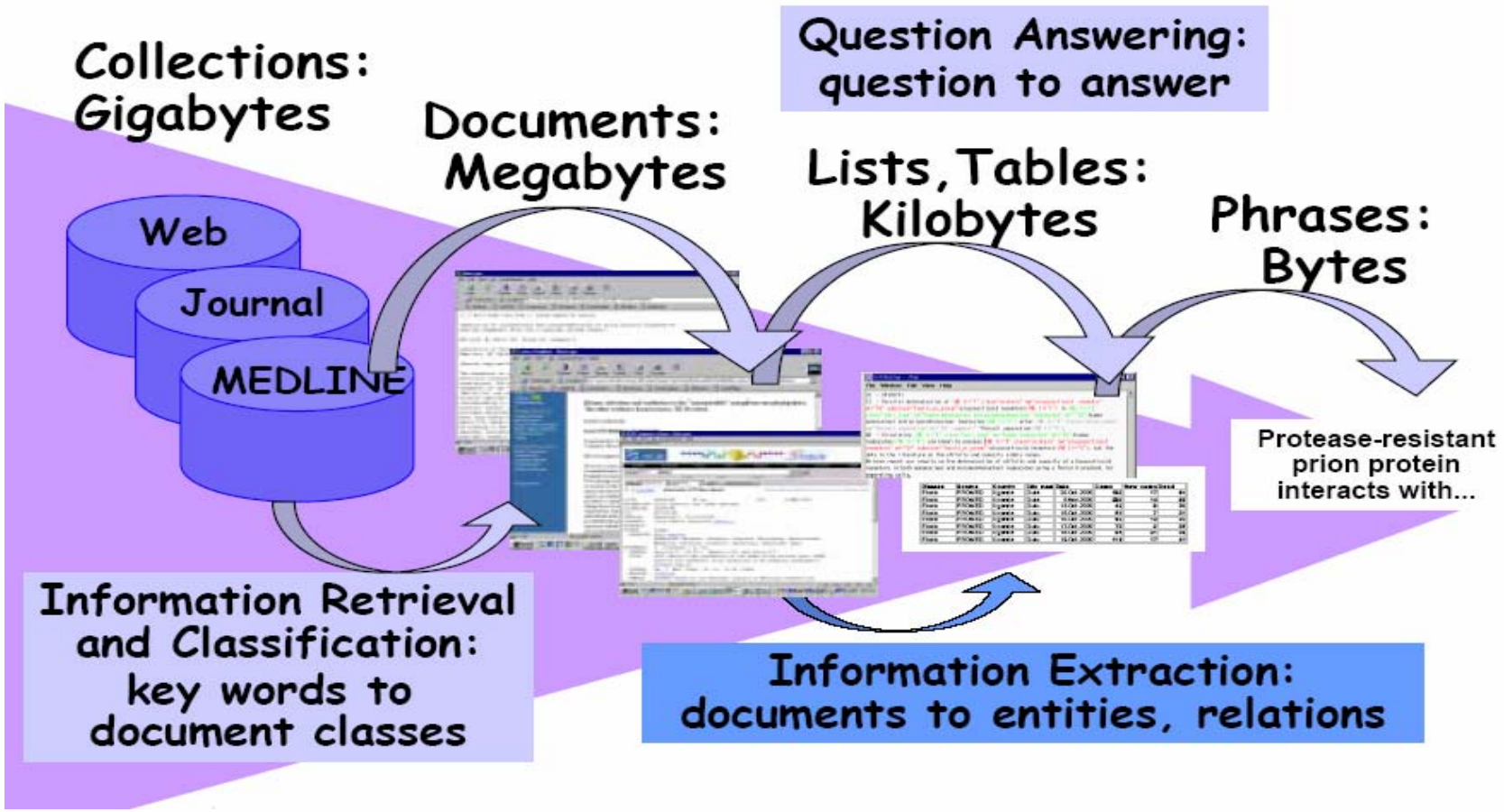
Top 5 Pharmaceutical Senior VP

MBC Informatics Committee, July 31, 2003

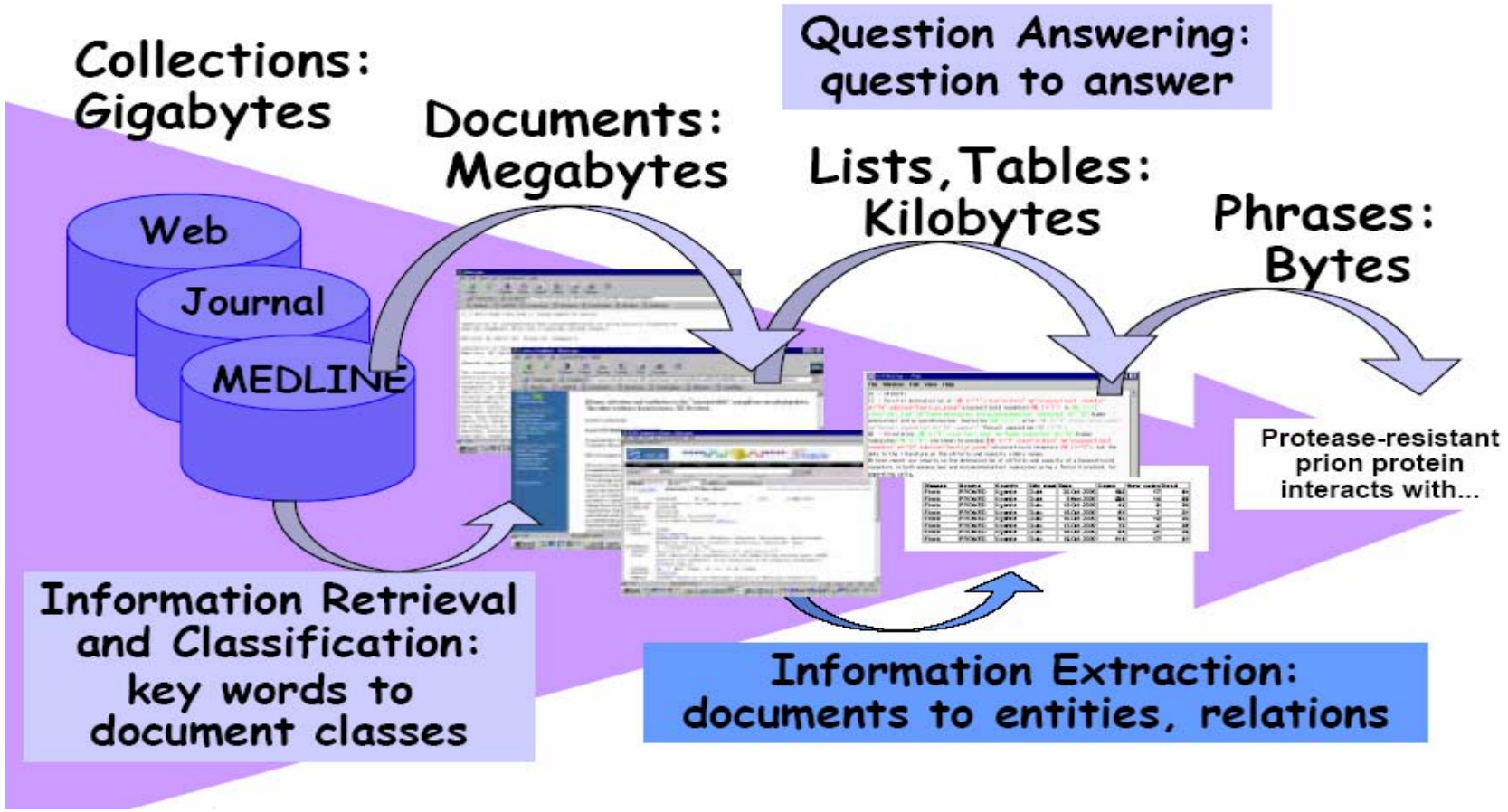
Zwei Text Mining Paradigmen



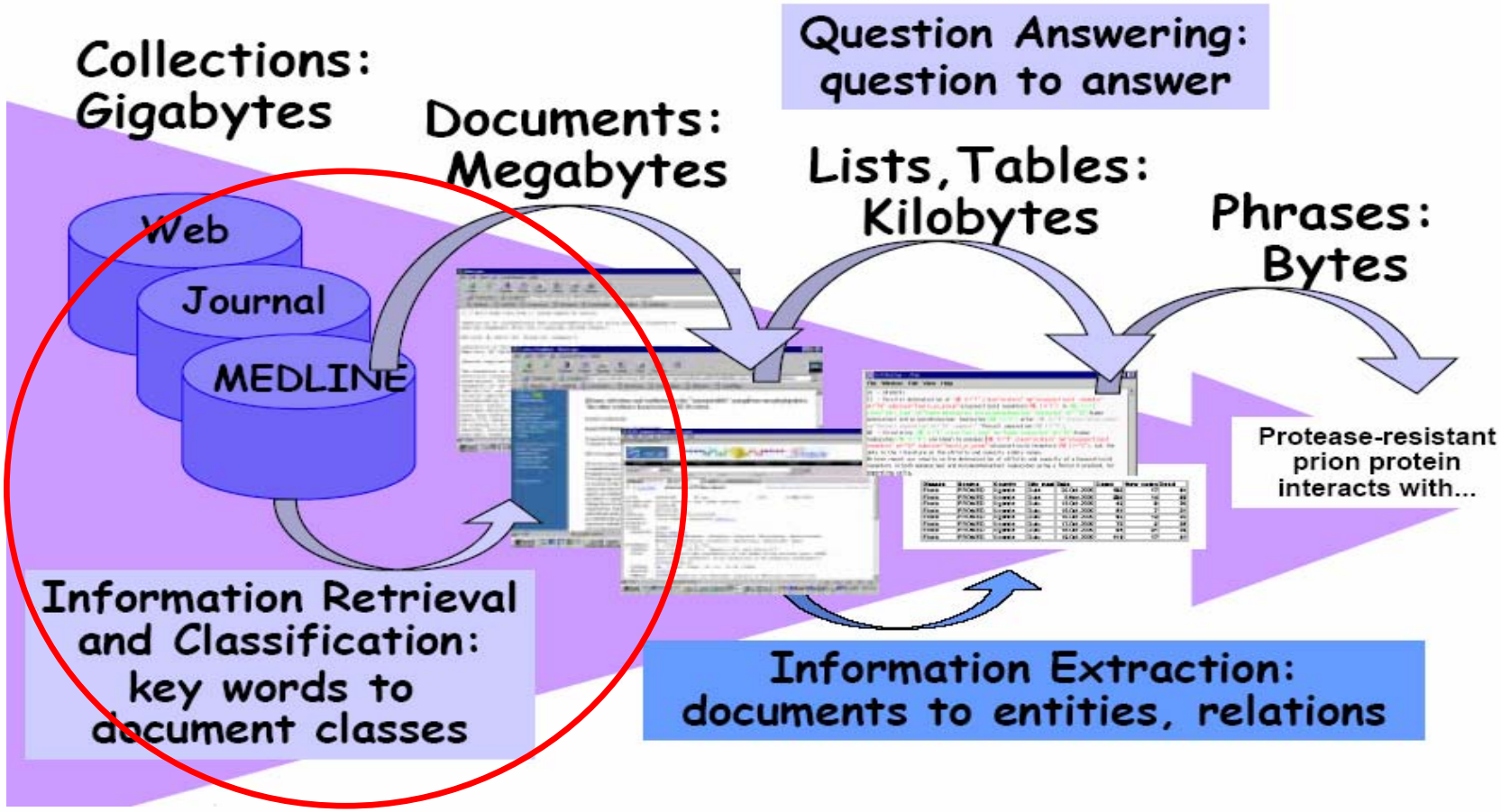
Text Mining Filter



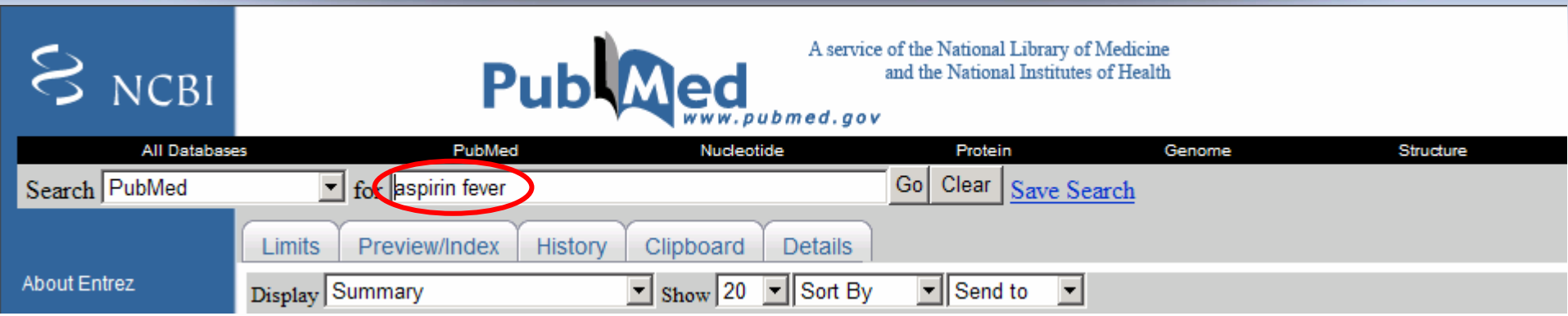
Informationskristallisationsfilter



Text Mining Filter

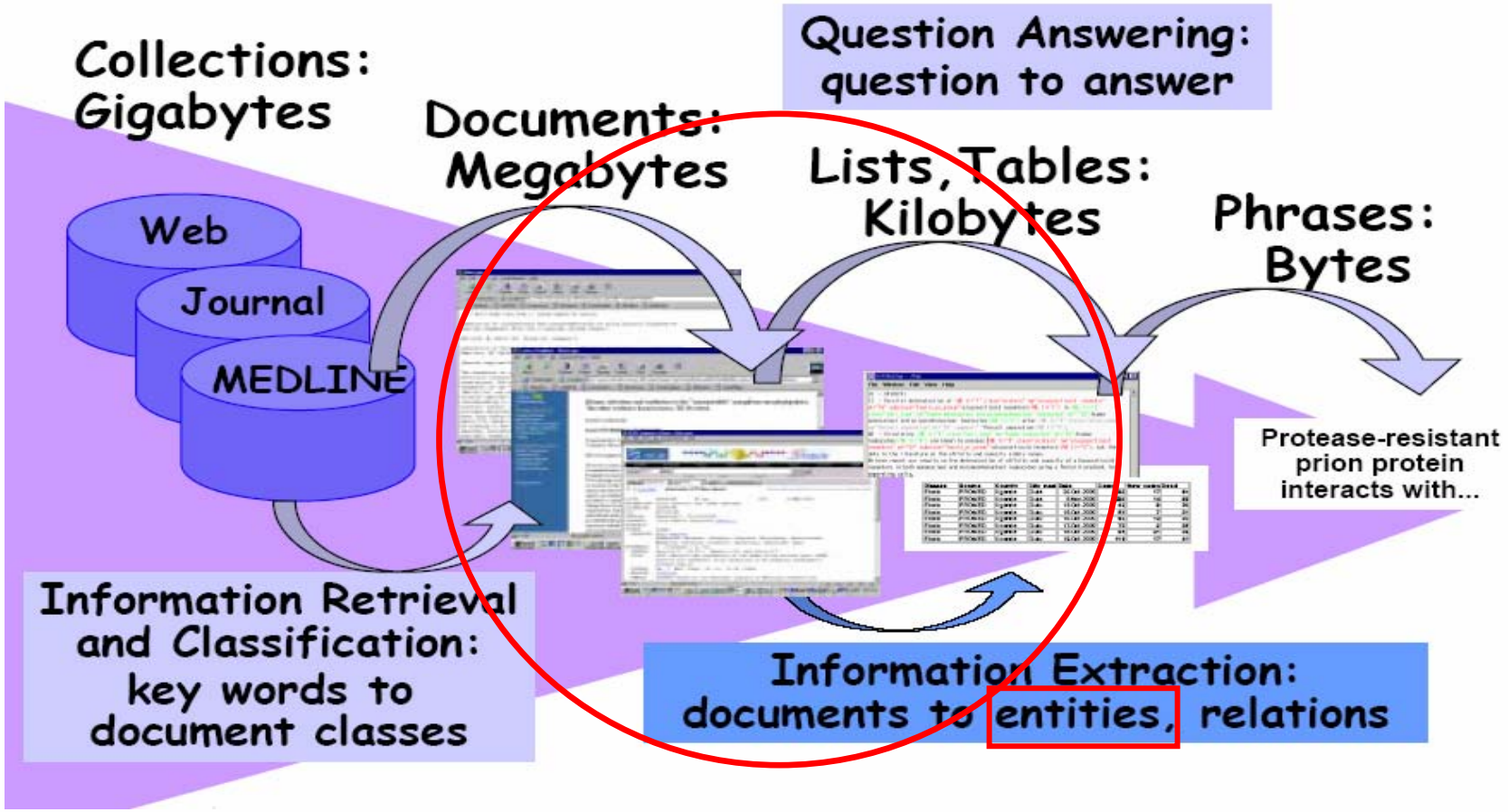


Information Retrieval: Keyword Suche



Thalidomide was found to be highly effective in managing the cutaneous manifestations of erythema nodosum leprosum (ENL) and even to be superior to **aspirin** (acetylsalicyclic acid) in controlling ENL-associated **fever**.

Text Mining Filter



Informationsextraktion: Entitätenerkennung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Entitätenerkennung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

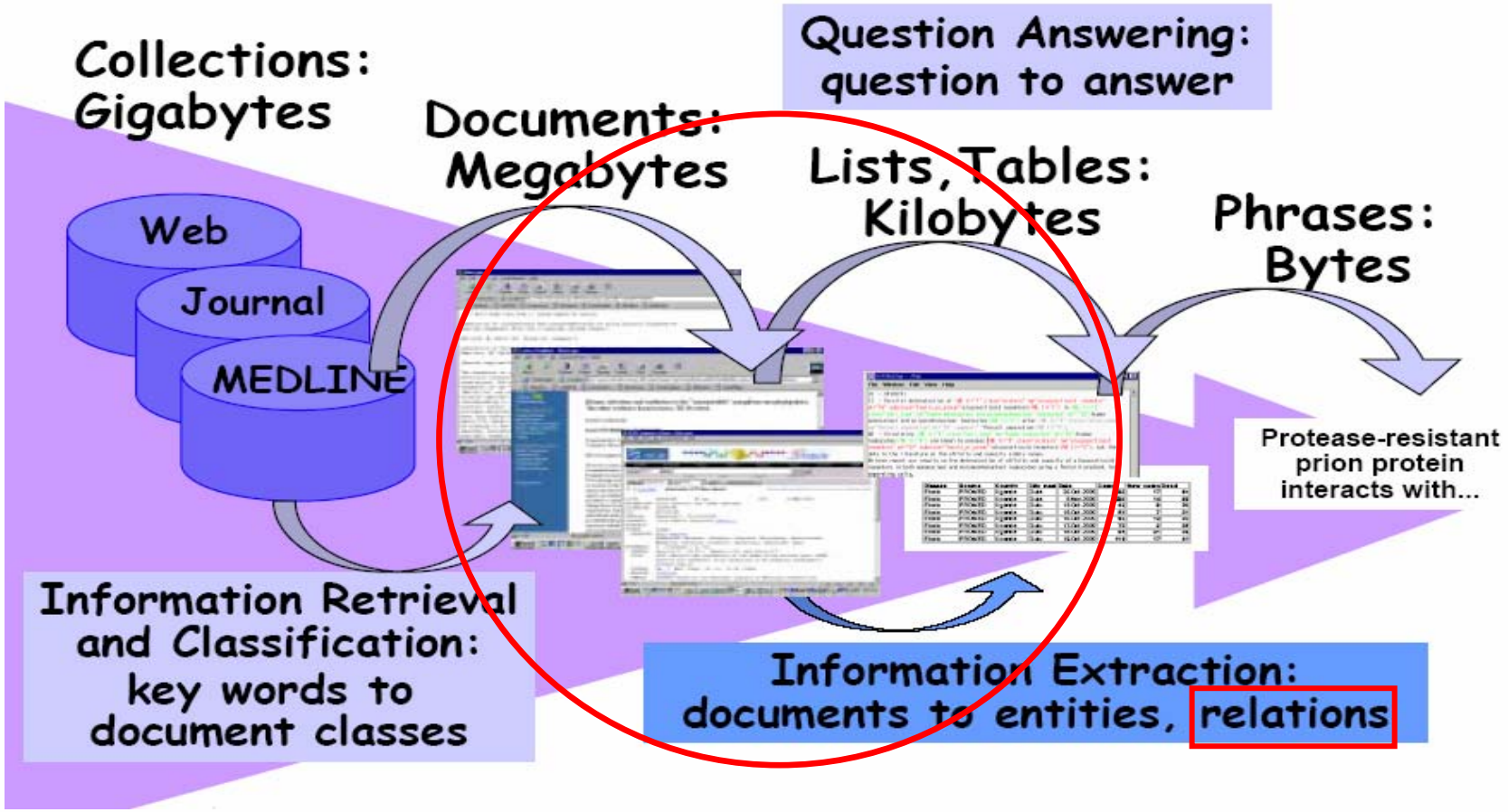
Informationsextraktion: Entitätennormalisierung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Entitätennormalisierung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> leprosy </DISEASE> (<DISEASE> leprosy </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> leprosy-associated fever </DISEASE>.

Text Mining Filter



Informationsextraktion: Relationserkennung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Relationserkennung

thalidomide

RELATION
EFFECTIVE-FOR

leprosy

<DRUG> Thalidomide </DRUG> was found to be highly **effective** in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Relationserkennung

thalidomide

RELATION
EFFECTIVE-FOR

leprosy-associated fever

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in **controlling** <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Relationserkennung

aspirin

RELATION
EFFECTIVE-FOR

leprosy-associated fever

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in **controlling** <DISEASE> ENL-associated fever </DISEASE>.

Informationsextraktion: Relationserkennung

thalidomide **>** aspirin

RELATION
EFFECTIVE-FOR

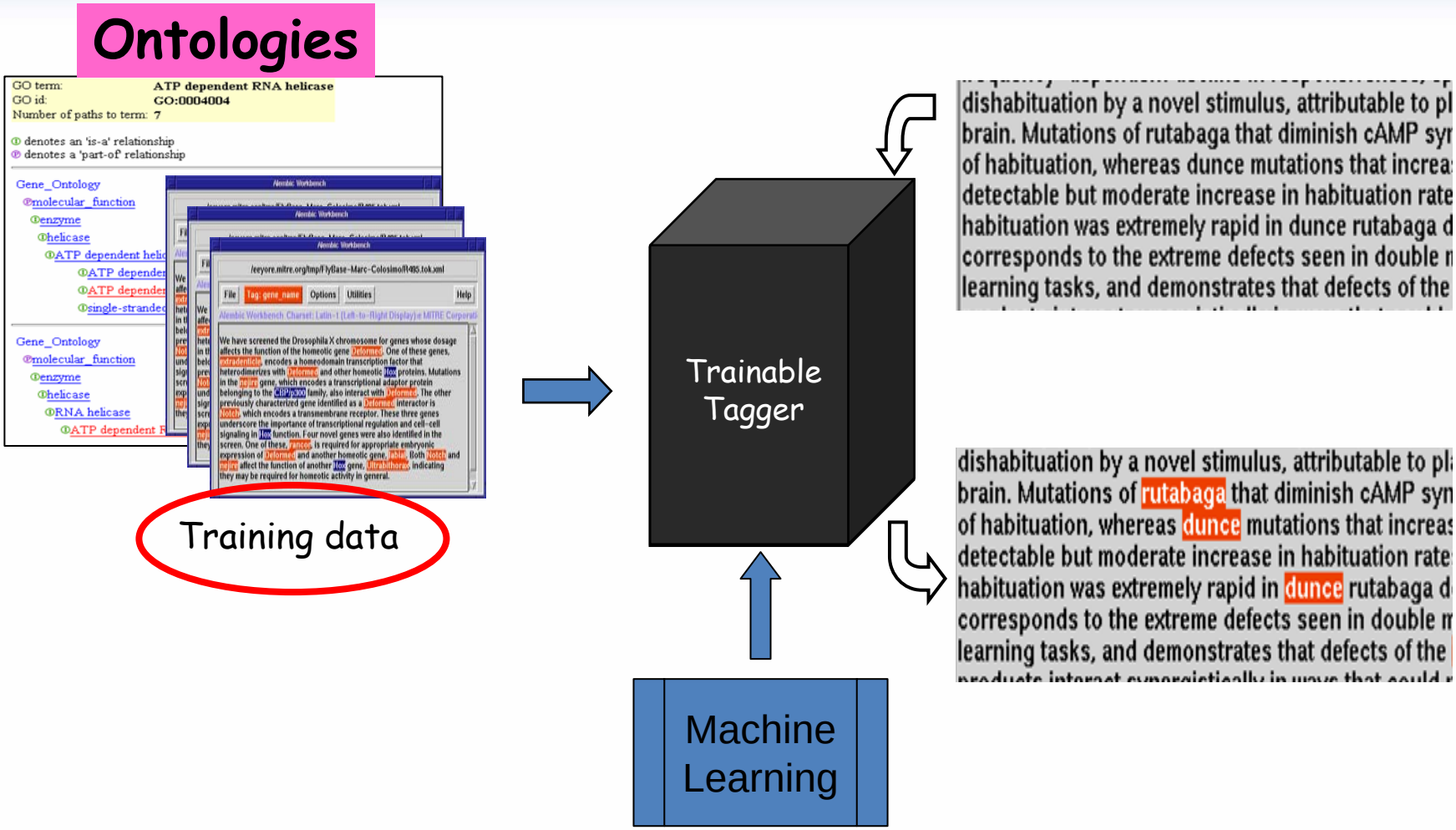
leprosy-associated fever

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be **superior** to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in **controlling** <DISEASE> ENL-associated fever </DISEASE>.

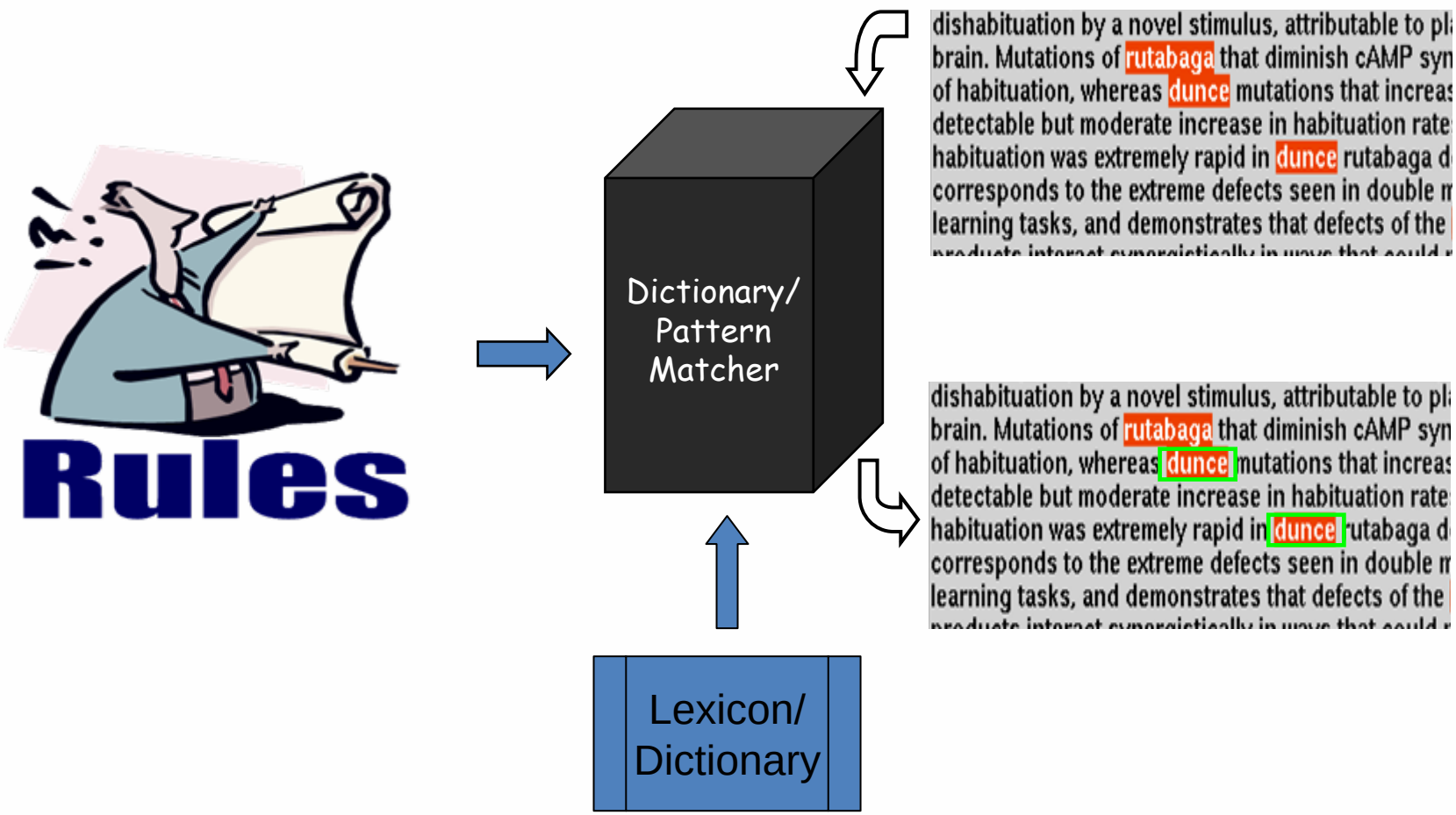
Informationsextraktion: Relationserkennung

<DRUG> Thalidomide </DRUG> was found to be highly effective in managing the <TISSUE> cutaneous </TISSUE> manifestations of <DISEASE> erythema nodosum leprosum </DISEASE> (<DISEASE> ENL </DISEASE>) and even to be superior to <DRUG> aspirin </DRUG> (<SUBSTANCE> acetylsalicyclic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

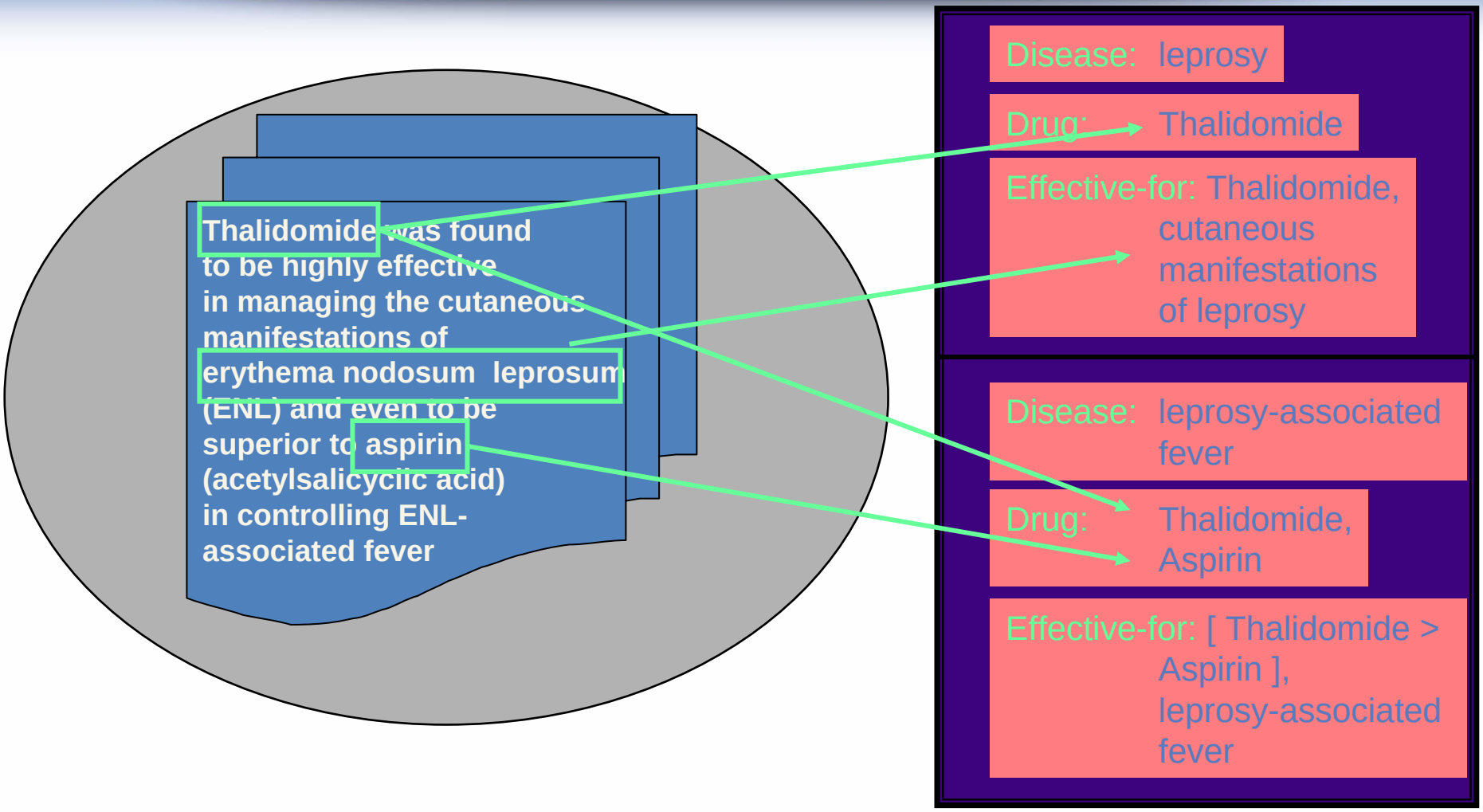
Maschinelles Lernen für Entitäten- und Relationserkennung



Dictionary-/Regelbasierte Entitätennormalisierung

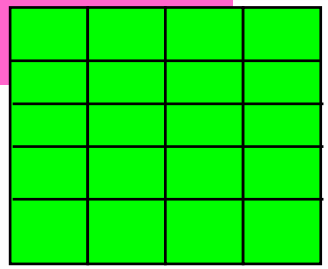


Textwissensbereitstellung (1)
Datenbanken und Templates



Textwissensbereitstellung (2)
Biomedizinisches Wissensmanagement

Experimental Data



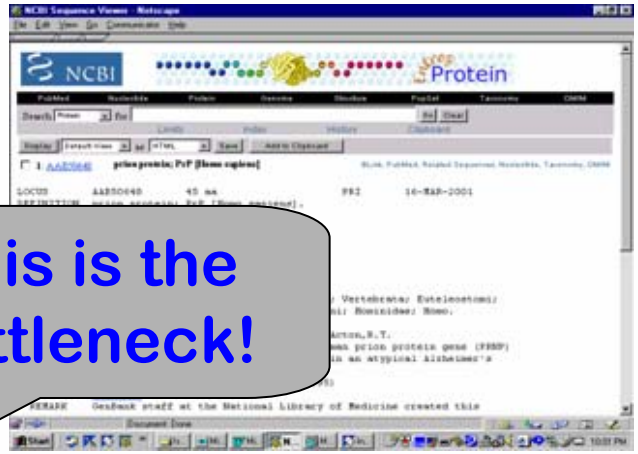
Ontologies

GO term: ATP dependent RNA helicase
GO id: GO:0004004
Number of paths to term: 7

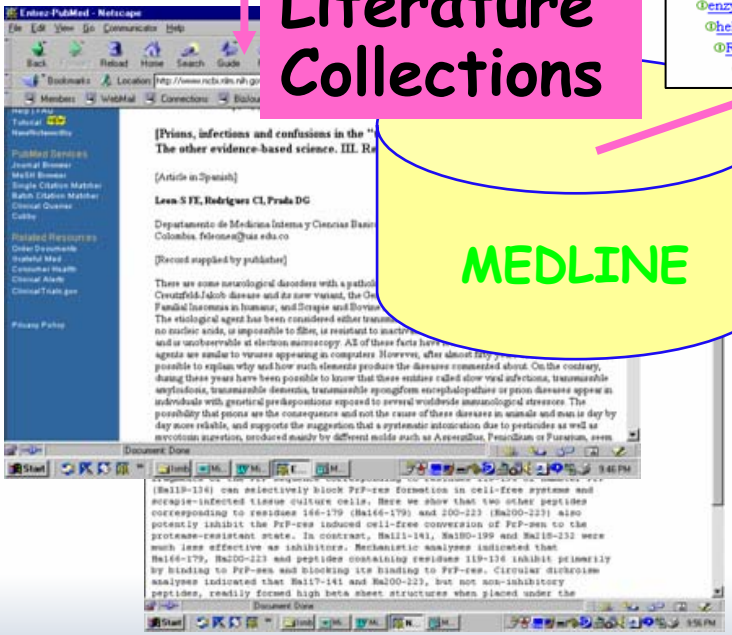
denotes an 'is-a' relationship
denotes a 'part-of' relationship

Gene_Ontology
molecular_function
enzyme
helicase
ATP dependent helicase
ATP dependent DNA helicase
ATP dependent RNA helicase [GO:0004004]
single-stranded DNA dependent ATP dependent

Gene_Ontology
molecular_function
enzyme
helicase
RNA helicase
ATP dependent RNA helicase [GO:0004004]



Literature Collections



MEDLINE

This is the bottleneck!

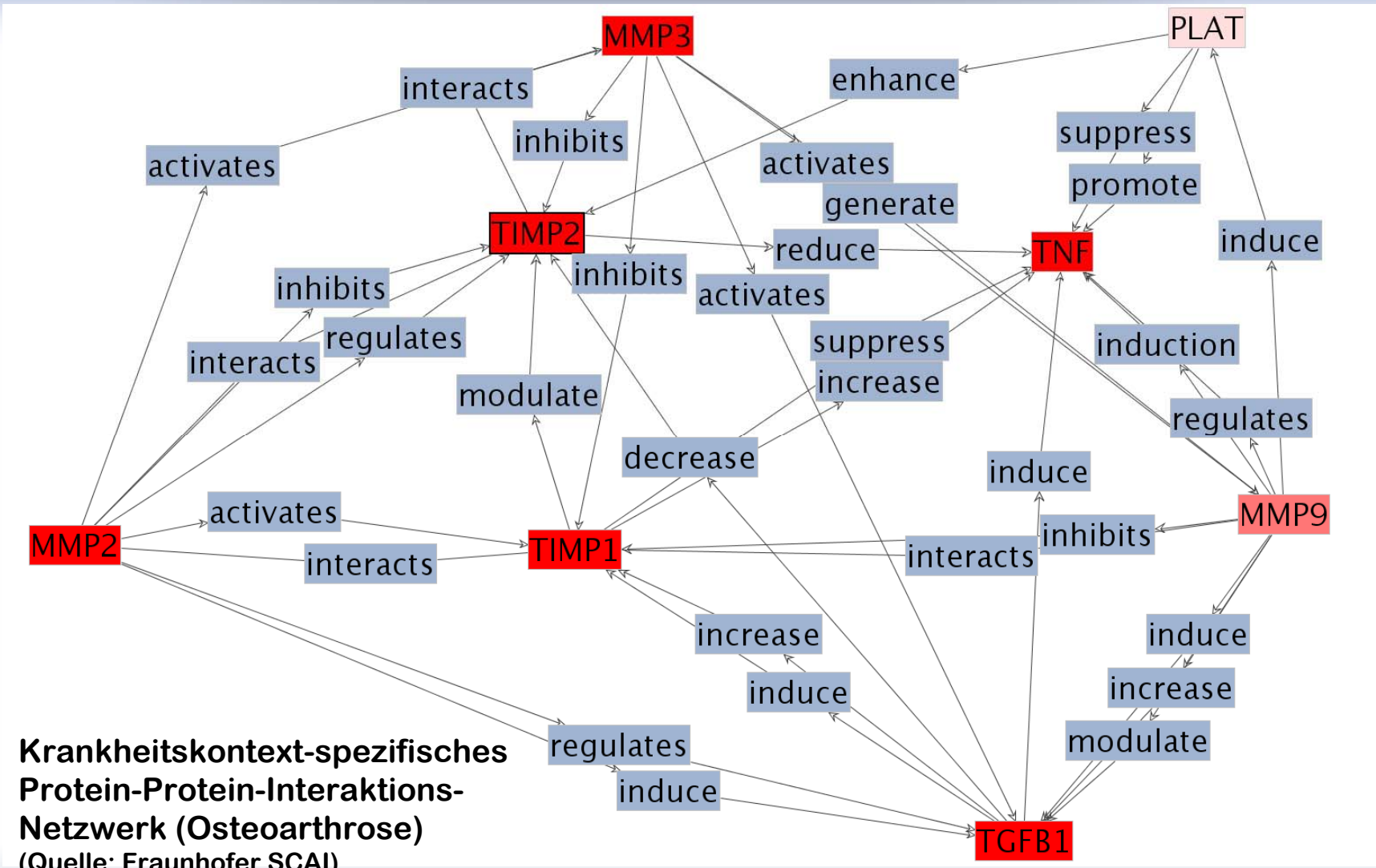
Expert Annotation

Automatic text mining

Databases

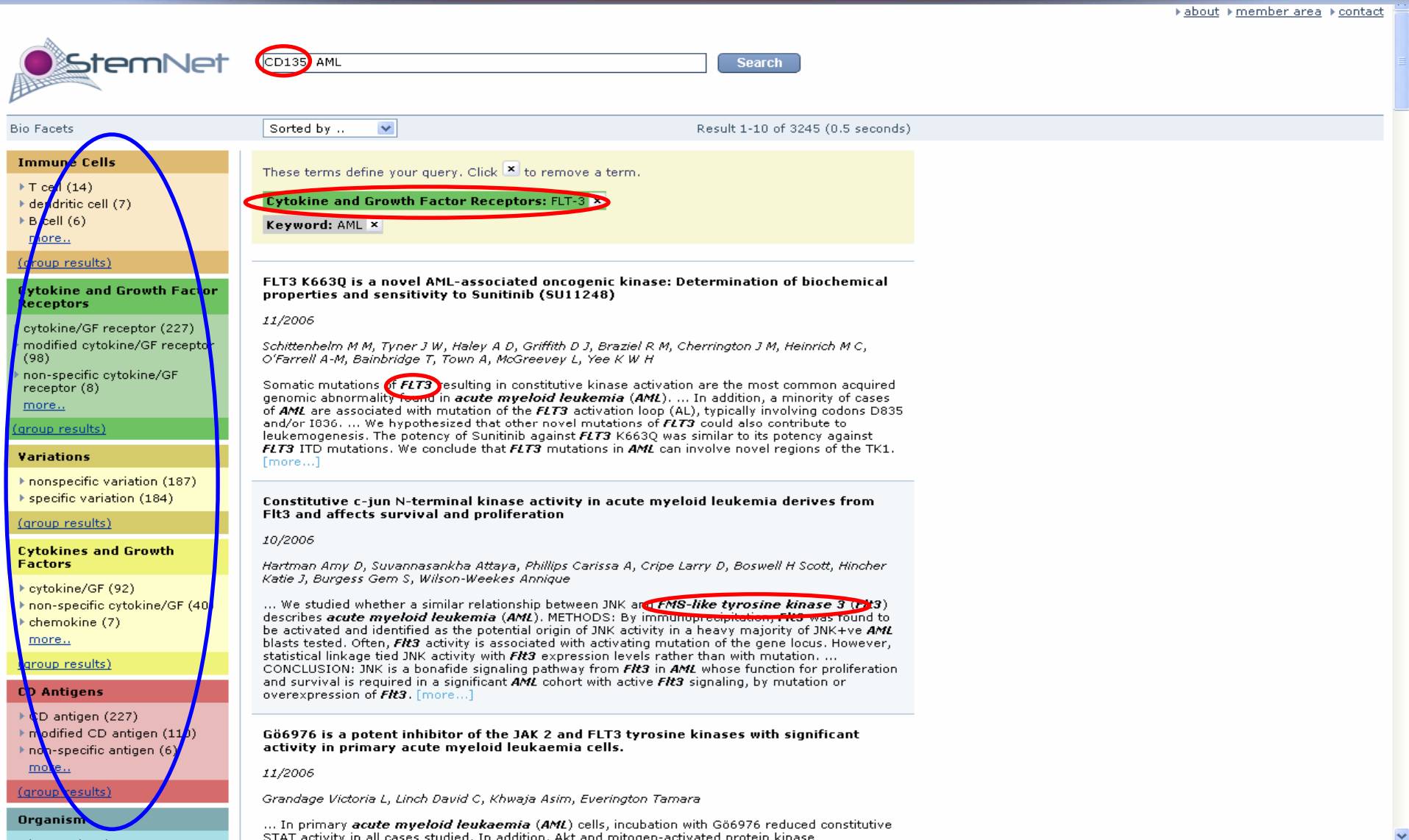
Genbank
UniProt
OMIM

Textwissensbereitstellung (3) Biomedizinisches Interaktionsnetzwerk



Krankheitskontext-spezifisches Protein-Protein-Interaktions-Netzwerk (Osteoarthrose)
(Quelle: Fraunhofer SCAI)

Textwissensbereitstellung (4)
Semantische Suche



Text Mining

Sprach- und Wissenstechnologien am Beispiel der Lebenswissenschaften

joachim.wermter@uni-jena.de

Jena University
Language and Information Engineering (JULIE) Lab

Jena University Language & Information Engineering Lab