

Text Mining

Sprach- und Wissenstechnologien am Beispiel der
Lebenswissenschaften

Udo Hahn

Joachim Wermter

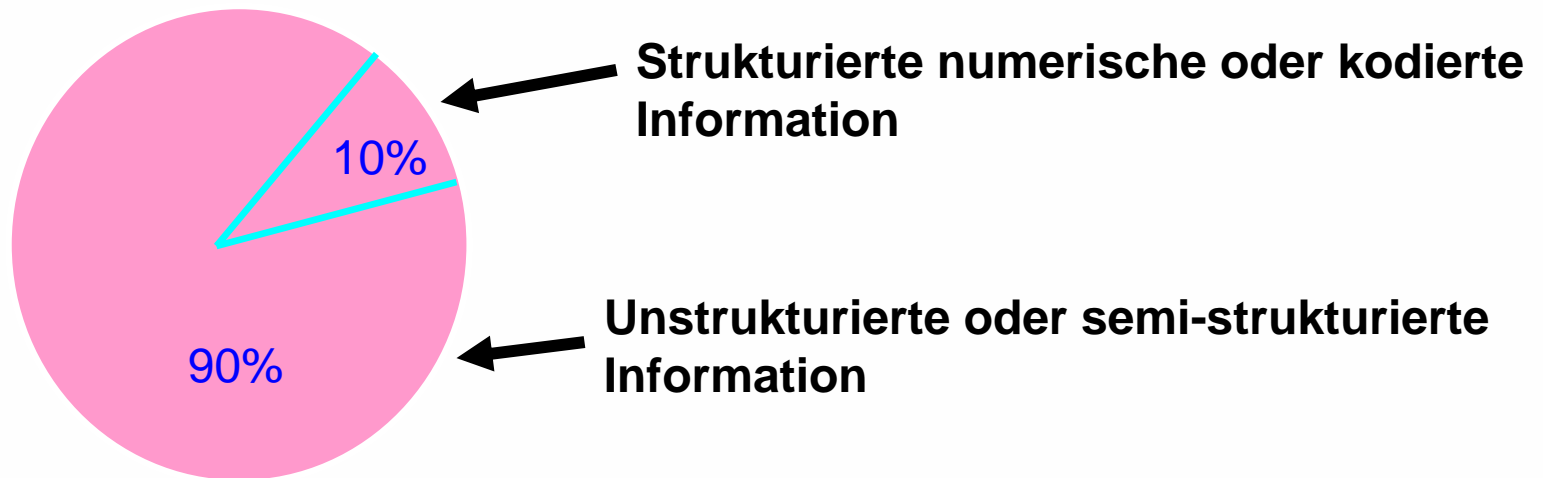
Jena University

Language and Information Engineering (JULIE) Lab



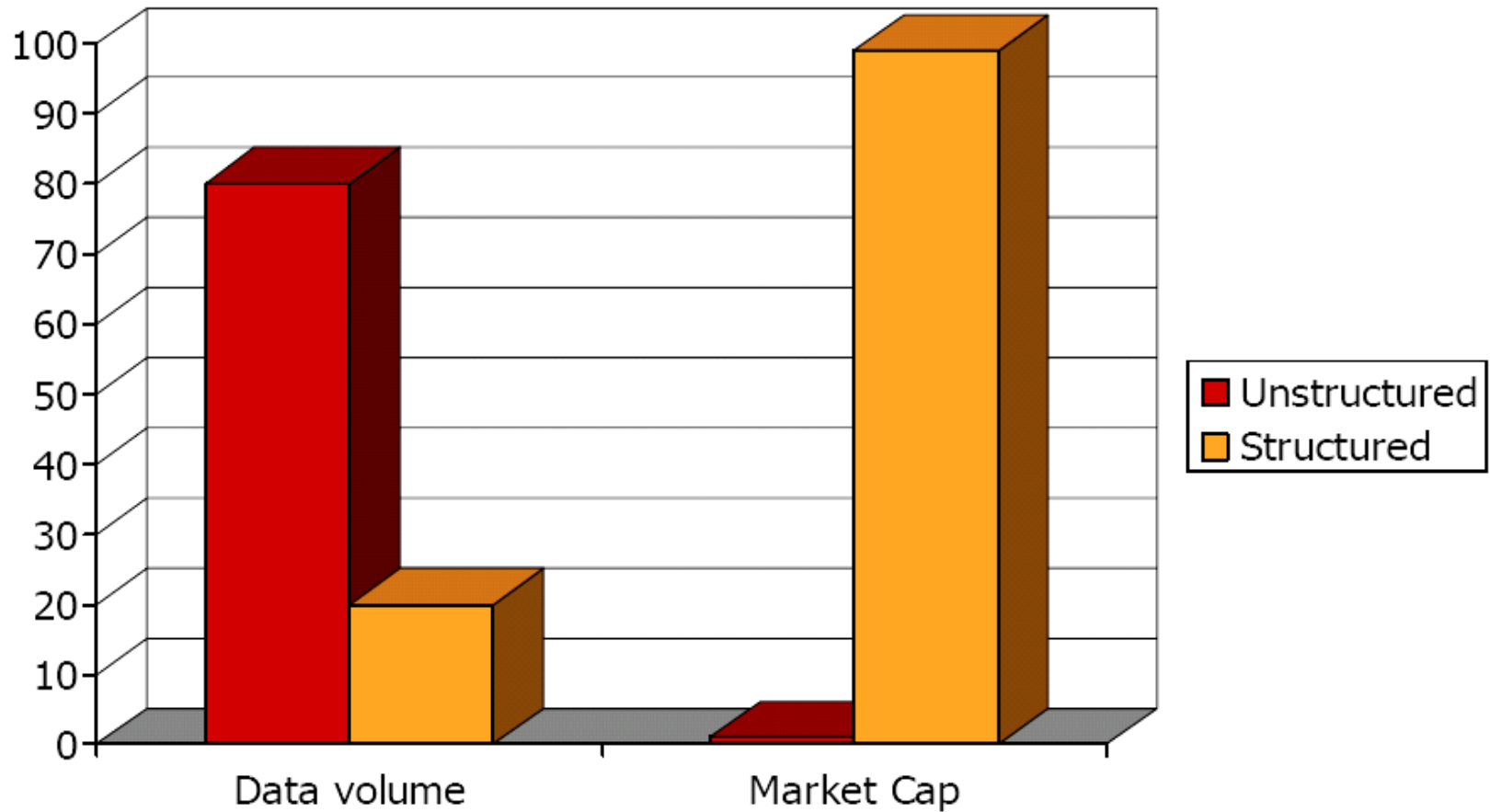
Warum ist Text Mining Relevant? (1/4)

90:10 Argument



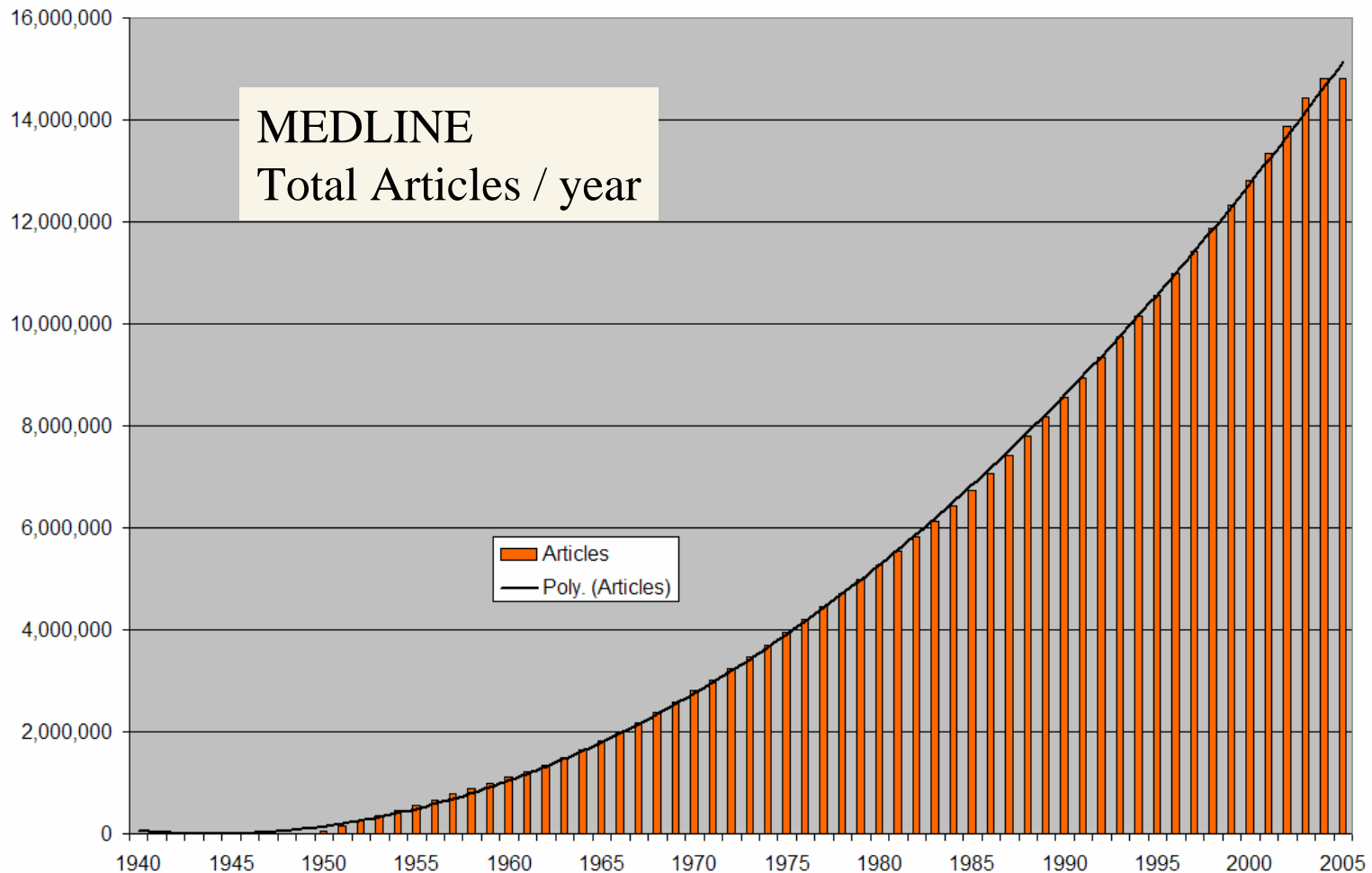
Quelle: Oracle Corp.

Warum ist Text Mining Relevant? (2/4)



Quelle: Prabhakar & Raghavan, Verity (2002)

Warum ist Text Mining Relevant? (3/4)



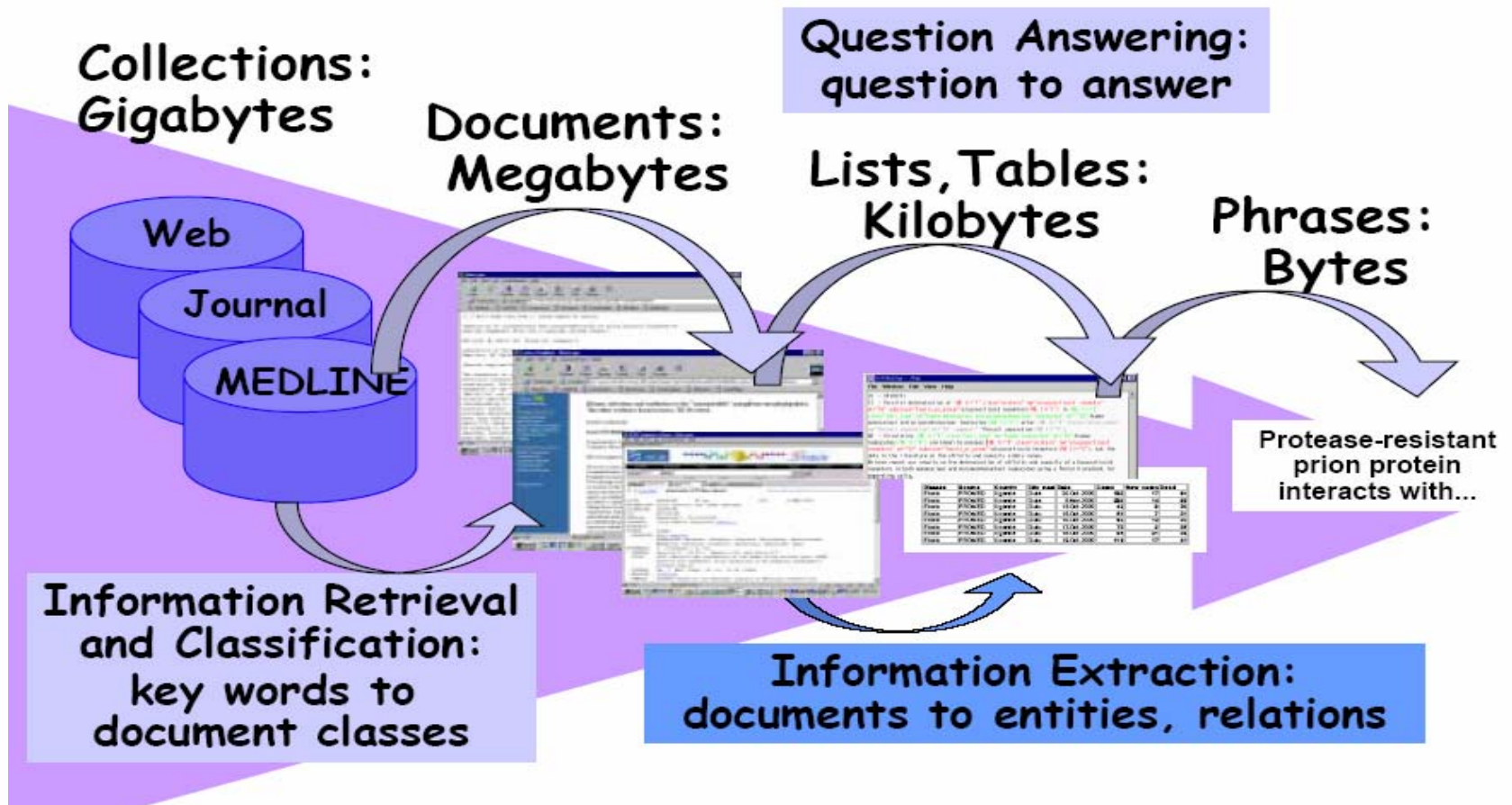
Warum ist Text Mining Relevant? (4/4)

“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.”

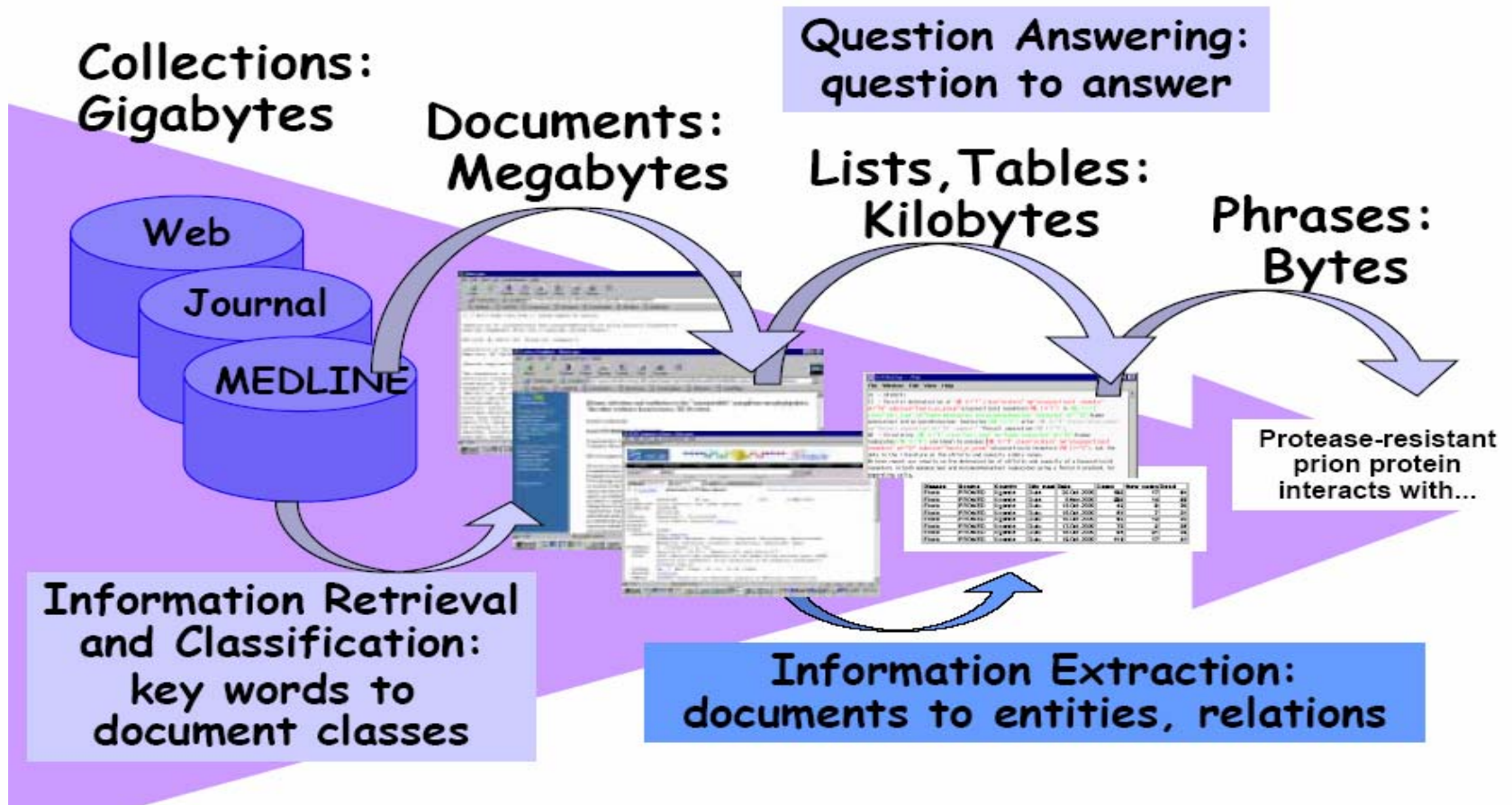
Top 5 Pharmaceutical Senior VP

MBC Informatics Committee, July 31, 2003

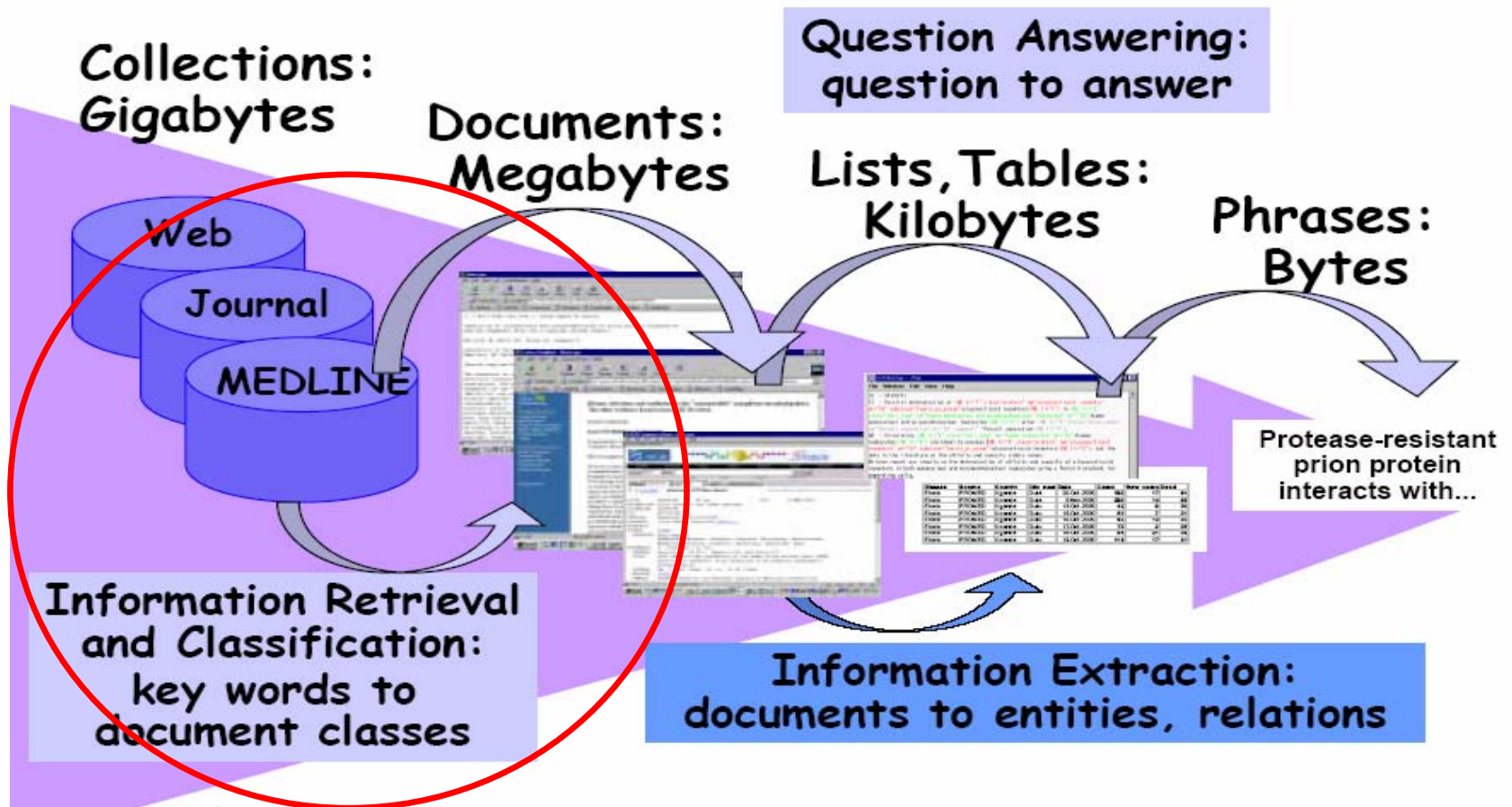
Text Mining Filter



Informationskristallisationsfilter



Text Mining Filter



Information Retrieval: Keyword Suche

NCBI

PubMed
www.pubmed.gov

A service of the National Library of Medicine
and the National Institutes of Health

All Databases PubMed Nucleotide Protein Genome Structure

Search PubMed for **aspirin fever** Go Clear [Save Search](#)

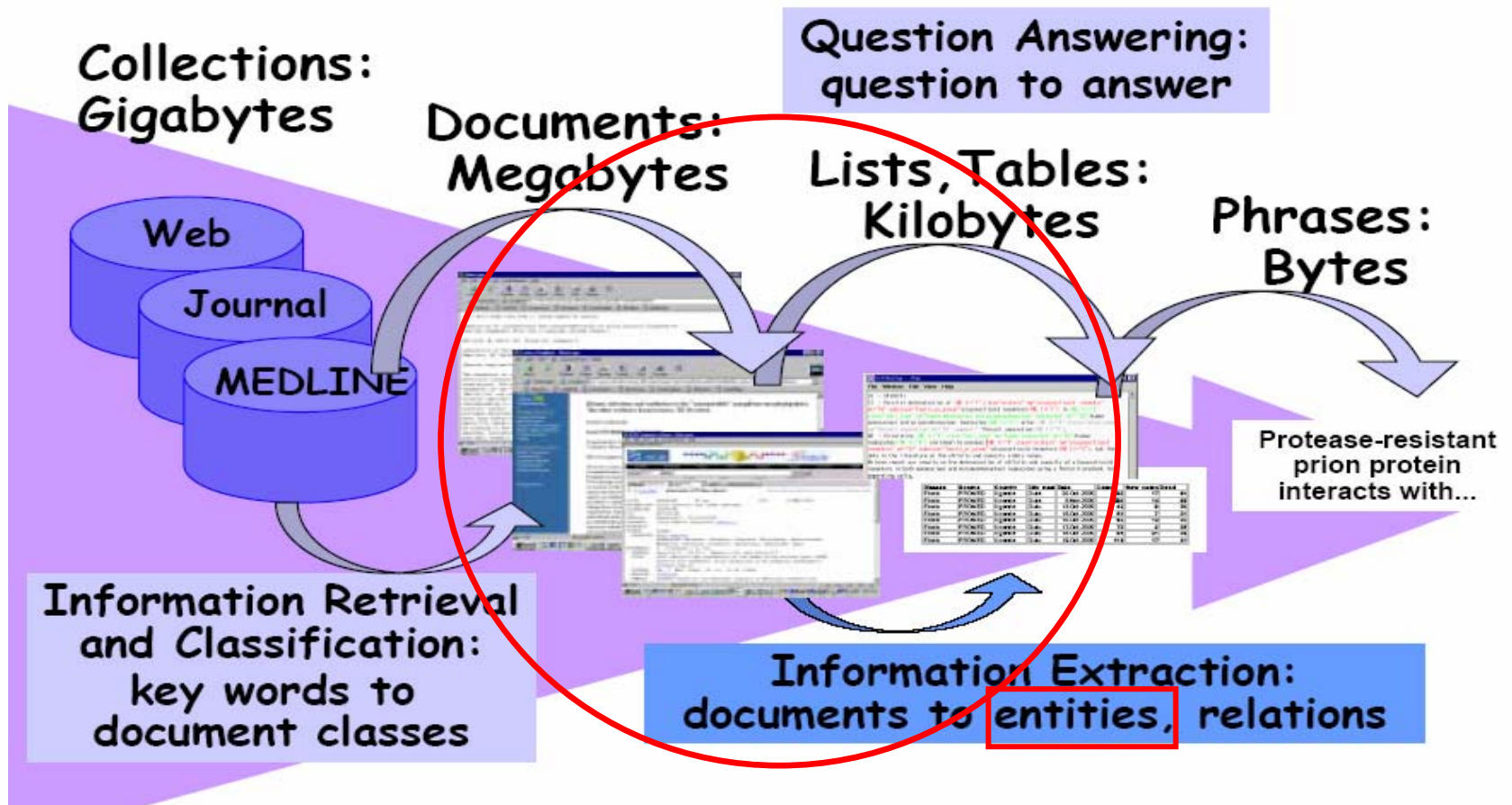
Limits Preview/Index History Clipboard Details

About Entrez

Display Summary Show 20 Sort By Send to

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> acetylsalicylic 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> acetylsalicylic acid </SUBSTANCE>) in controlling <DISEASE> ENL-associated fever </DISEASE>.

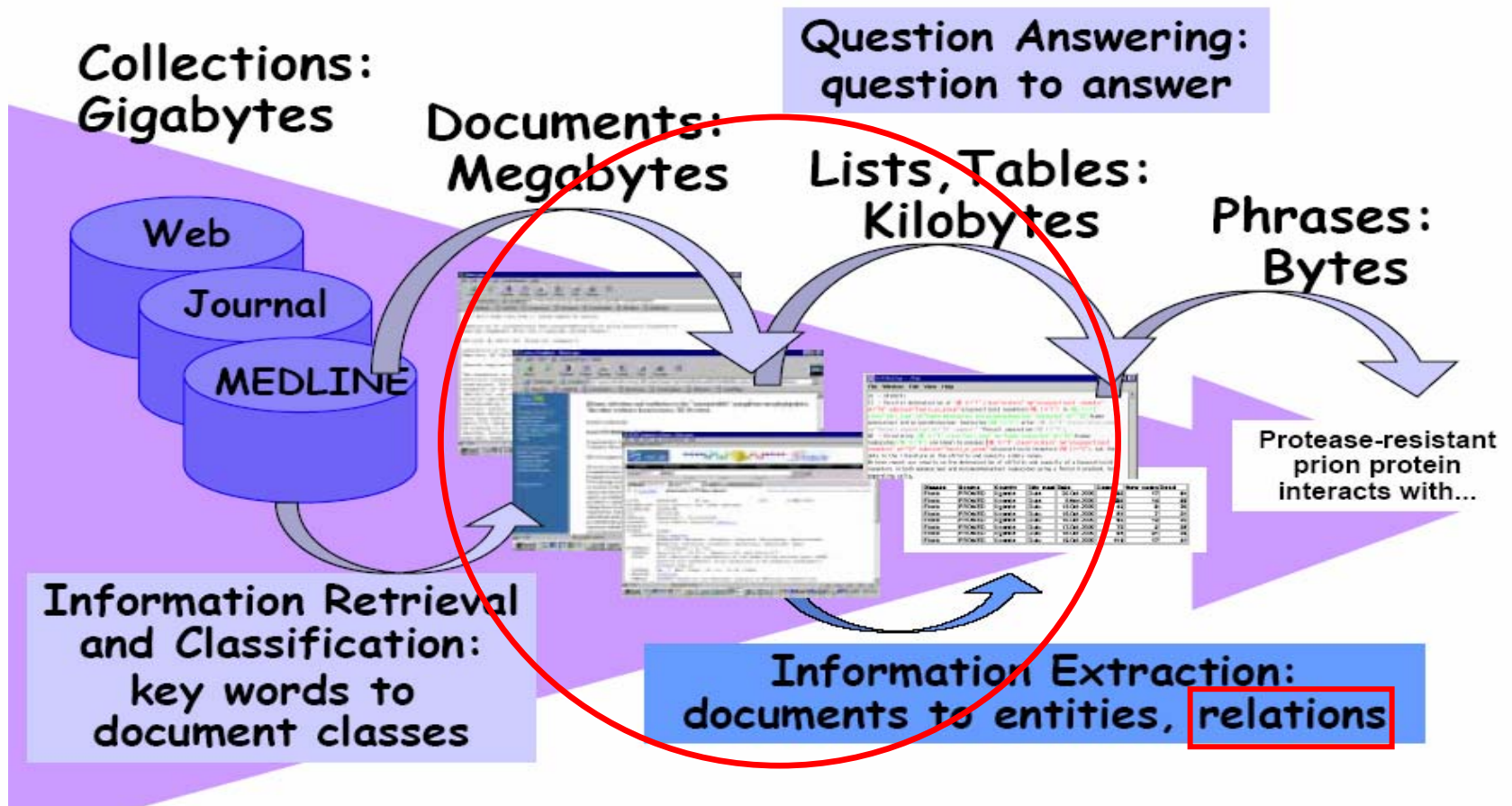
Informationsextraktion: Entitätennormalisierung

<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>.

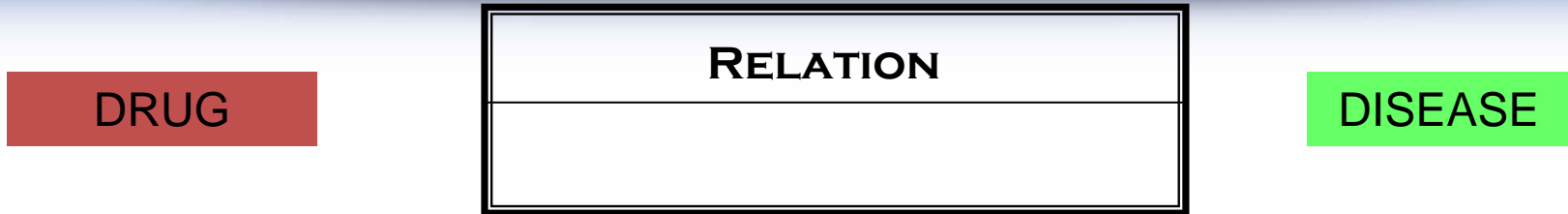
Informationsextraktion: Entitätennormalisierung

<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



<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> acetylsalicylic 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

Ontologies

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

① denotes an 'is-a' relationship
② denotes a 'part-of' relationship

Gene_Ontology
@molecular_function
①enzyme
①helicase
②ATP dependent helicase
②ATP dependent helicase
②ATP dependent helicase
②single-stranded DNA helicase

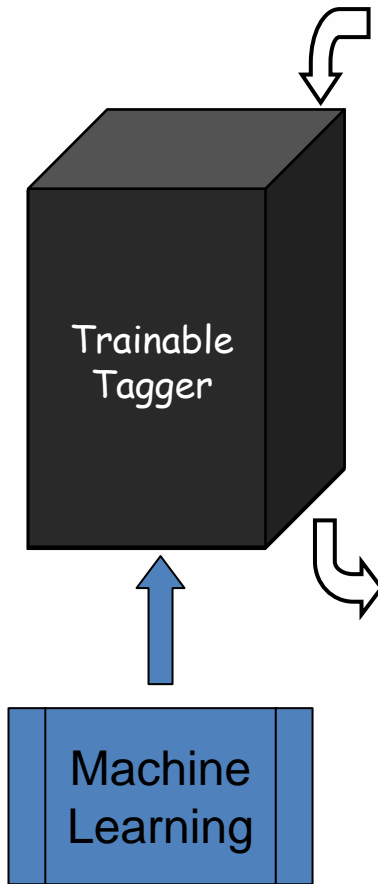
Gene_Ontology
@molecular_function
①enzyme
①helicase
②RNA helicase
②ATP dependent helicase

File: tag_gene_name Options Utilities Help

Alameda Workbench Charset: Latin-1 (Left-to-Right Display) MITRE Corporation

We have screened the Drosophila X chromosome for genes whose dosage affects the function of the homeotic gene **Ultrabithorax**. One of these genes, **extralimbic**, encodes a homeodomain transcription factor that heterodimerizes with **Dfd** and other homeotic **TCF** proteins. Mutations in the **Ubx** gene, which encodes a transcriptional adaptor protein belonging to the **POU** family, also interact with **Ubx**. The other previously characterized gene identified as a **Ubx** interactor is **Notch**, which encodes a transmembrane receptor. These three genes underscore the importance of transcriptional regulation and cell-cell signaling in **TCF** function. Four novel genes were also identified in the screen. One of these, **rutabaga**, is required for appropriate embryonic expression of **Ubx**, and another homeotic gene, **Notch**, both **rutabaga** and **Notch** affect the function of another **TCF** gene, **Ultrabithorax**, indicating they may be required for homeotic activity in general.

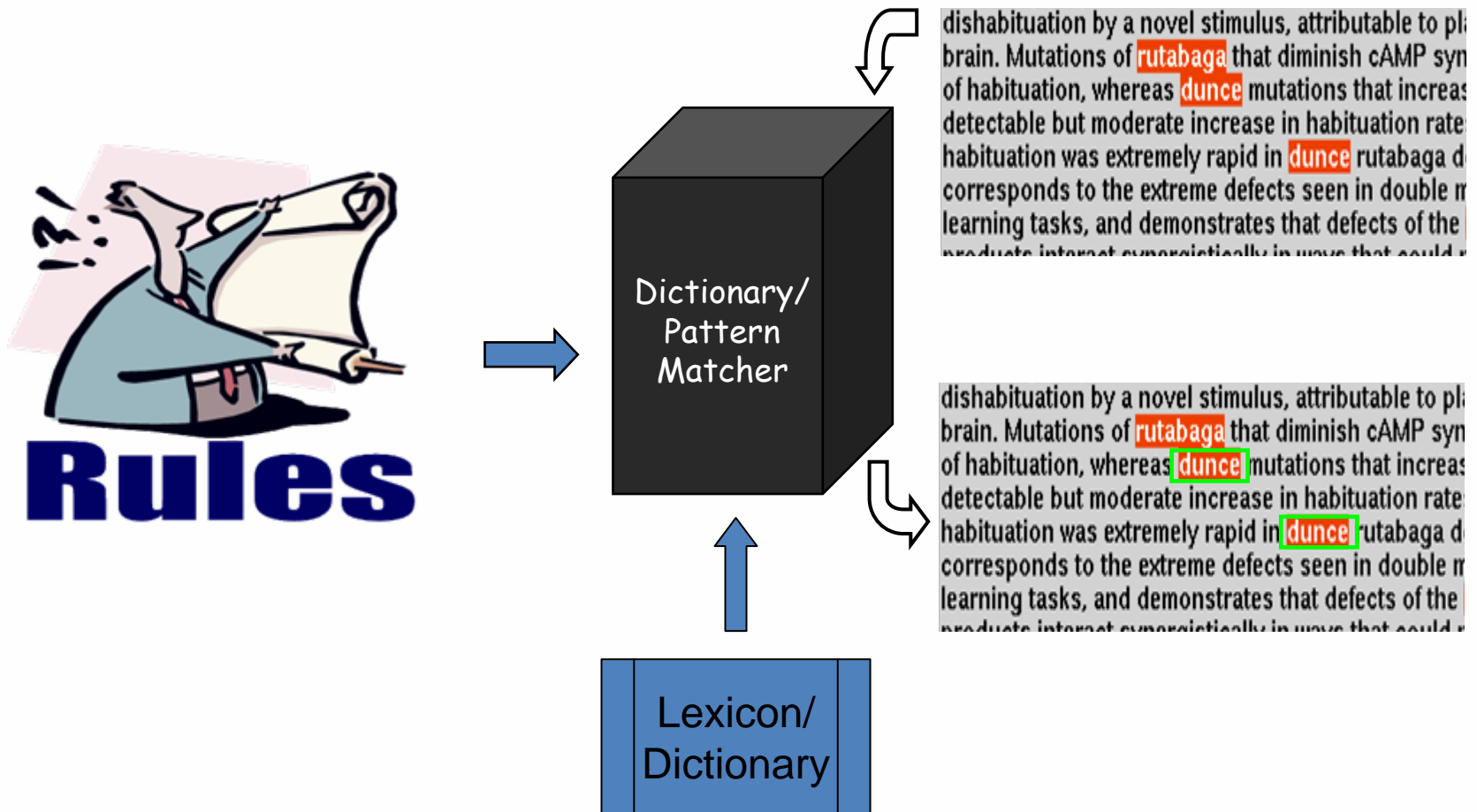
Training data



dishabitu...
ation by a novel stimulus, attributable to pl...
brain. Mutations of **rutabaga** that diminish cAMP syn...
of habituation, whereas **dunce** mutations that increa...
detectable but moderate increase in habituation rate...
habituation was extremely rapid in **dunce** rutabaga d...
corresponds to the extreme defects seen in double m...
learning tasks, and demonstrates that defects of the

dishabitu...
ation by a novel stimulus, attributable to pl...
brain. Mutations of **rutabaga** that diminish cAMP syn...
of habituation, whereas **dunce** mutations that increas...
detectable but moderate increase in habituation rate...
habituation was extremely rapid in **dunce** rutabaga d...
corresponds to the extreme defects seen in double m...
learning tasks, and demonstrates that defects of the
products interact synergistically in ways that could

Dictionary-/Regelbasierte Entitätennormalisierung



Textwissensbereitstellung (1)

Datenbanken und Templates

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

Disease: leprosy

Drug: Thalidomide

Effective-for: Thalidomide, cutaneous manifestations of leprosy

Disease: leprosy-associated fever

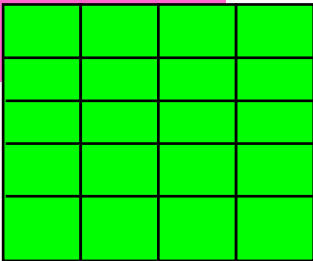
Drug: Thalidomide, Aspirin

Effective-for: [Thalidomide > Aspirin], leprosy-associated fever

Textwissensbereitstellung (2)

Biomedizinisches Wissensmanagement

Experimental
Data



Ontologies

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

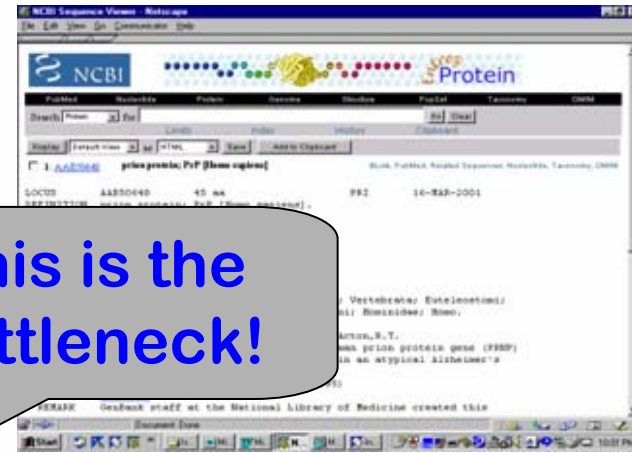
Ⓞ denotes an 'is-a' relationship
Ⓢ denotes a 'part-of' relationship

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 RNA helicase

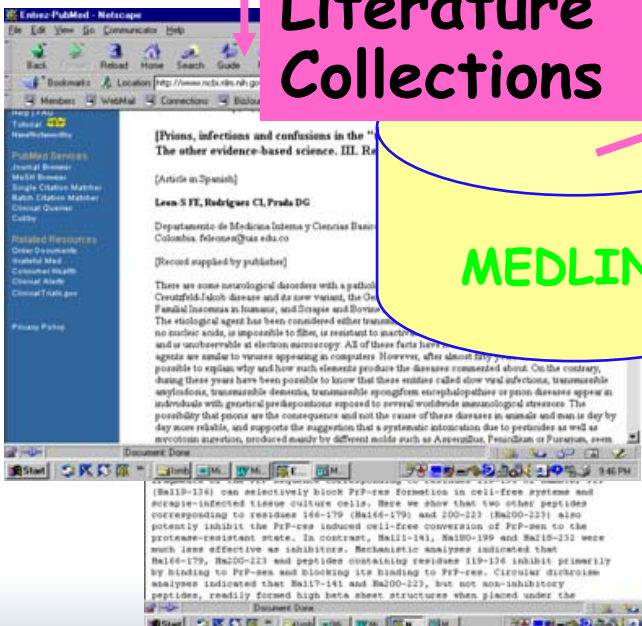
Gene_Ontology

- Ⓢmolecular_function
- Ⓞenzyme
- Ⓞhelicase
 - ⓄRNA helicase
 - ⓄATP dependent RNA helicase [GO:0004004]



This is the bottleneck!

Literature
Collections

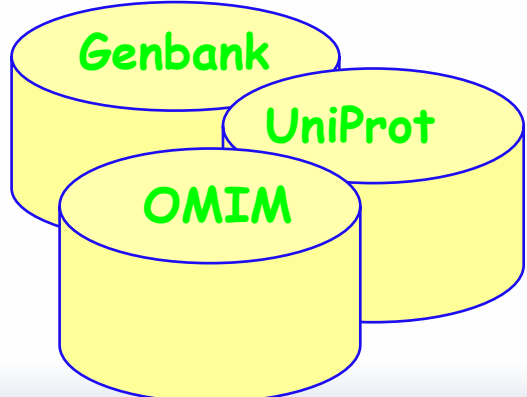


MEDLINE

Expert
Annotation

Automatic
text mining

Databases



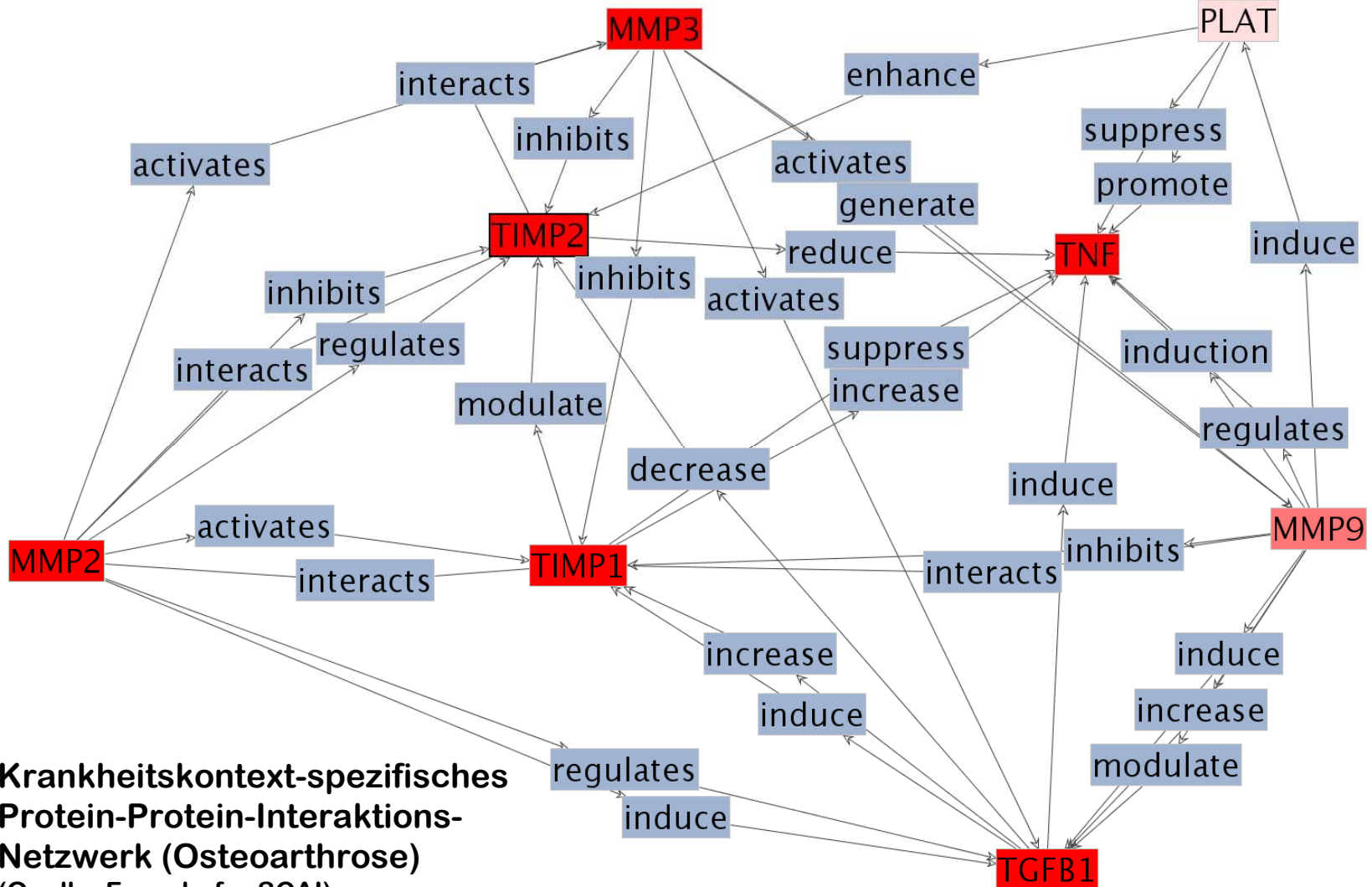
Genbank

UniProt

OMIM

Textwissensbereitstellung (3)

Biomedizinisches Interaktionsnetzwerk



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

Textwissensbereitstellung (4)

Semantische Suche

[about](#) [member area](#) [contact](#)



ICD135 AML

Search

Bio Facets

Sorted by ..

Result 1-10 of 3245 (0.5 seconds)

Immune Cells

- ▶ T cell (14)
- ▶ dendritic cell (7)
- ▶ B cell (6)
- [more..](#)

[\(group results\)](#)

Cytokine and Growth Factor receptors

- ▶ cytokine/GF receptor (227)
- ▶ modified cytokine/GF receptor (98)
- ▶ non-specific cytokine/GF receptor (8)
- [more..](#)

[\(group results\)](#)

Variations

- ▶ nonspecific variation (187)
- ▶ specific variation (184)

[\(group results\)](#)

Cytokines and Growth Factors

- ▶ cytokine/GF (92)
- ▶ non-specific cytokine/GF (40)
- ▶ chemokine (7)
- [more..](#)

[\(group results\)](#)

CD Antigens

- ▶ CD antigen (227)
- ▶ modified CD antigen (110)
- ▶ non-specific antigen (6)
- [more..](#)

[\(group results\)](#)

Organism

These terms define your query. Click to remove a term.

Cytokine and Growth Factor Receptors: FLT-3

Keyword: AML

FLT3 K663Q is a novel AML-associated oncogenic kinase: Determination of biochemical properties and sensitivity to Sunitinib (SU11248)

11/2006

Schittenhelm M M, Tyner J W, Haley A D, Griffith D J, Brazier R M, Cherrington J M, Heinrich M C, O'Farrell A-M, Bainbridge T, Town A, McGreevey L, Yee K W H

Somatic mutations of **FLT3** resulting in constitutive kinase activation are the most common acquired genomic abnormality found in **acute myeloid leukemia (AML)**. ... In addition, a minority of cases of **AML** are associated with mutation of the **FLT3** activation loop (AL), typically involving codons D835 and/or I636. ... We hypothesized that other novel mutations of **FLT3** could also contribute to leukemogenesis. The potency of Sunitinib against **FLT3** K663Q was similar to its potency against **FLT3** ITD mutations. We conclude that **FLT3** mutations in **AML** can involve novel regions of the TK1. [\[more...\]](#)

Constitutive c-jun N-terminal kinase activity in acute myeloid leukemia derives from Flt3 and affects survival and proliferation

10/2006

Hartman Amy D, Suvannasankha Attaya, Phillips Carissa A, Cripe Larry D, Boswell H Scott, Hinchey Katie J, Burgess Gem S, Wilson-Weekes Annique

... We studied whether a similar relationship between JNK and **FMS-like tyrosine kinase 3 (FLT3)** describes **acute myeloid leukemia (AML)**. METHODS: By immunoprecipitation, **Flt3** was found to be activated and identified as the potential origin of JNK activity in a heavy majority of JNK+ve **AML** blasts tested. Often, **Flt3** activity is associated with activating mutation of the gene locus. However, statistical linkage tied JNK activity with **Flt3** expression levels rather than with mutation. ... CONCLUSION: JNK is a bonafide signaling pathway from **Flt3** in **AML** whose function for proliferation and survival is required in a significant **AML** cohort with active **Flt3** signaling, by mutation or overexpression of **Flt3**. [\[more...\]](#)

Gö6976 is a potent inhibitor of the JAK 2 and FLT3 tyrosine kinases with significant activity in primary acute myeloid leukaemia cells.

11/2006

Grandage Victoria L, Lynch David C, Khwaja Asim, Everington Tamara

... In primary **acute myeloid leukaemia (AML)** cells, incubation with Gö6976 reduced constitutive STAT activity in all cases studied. In addition, Akt and mitogen-activated protein kinase

Text Mining

Sprach- und Wissenstechnologien am Beispiel der
Lebenswissenschaften

joachim.wermter@uni-jena.de

Jena University

Language and Information Engineering (JULIE) Lab



14.09.2007