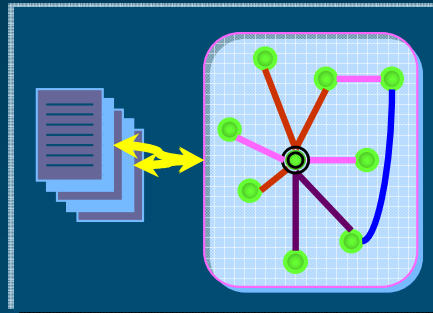


Forum on Informatics Solutions
November 3, 2006

Advanced Library Services

*Developing a Biomedical Knowledge Repository
to Support Advanced Information Management Applications*

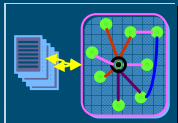


Olivier Bodenreider, M.D., Ph.D.
Thomas C. Rindflesch, Ph.D.

Context

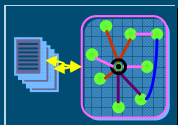
- ◆ Provide biomedical information to health care professionals and consumers
 - Exploit NLM resources
 - Maintain NLM's cutting edge

- ◆ Proposal overview
 - *Advanced Library Services*
 - *Biomedical Knowledge Repository*
- ◆ Pilot projects



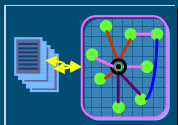
Why additional services?

- ◆ Biomedical information is growing at an increasingly faster pace
 - High-throughput approach to knowledge processing
- ◆ Information retrieval is the starting point, not the end of the journey for the researcher
 - Towards “computable” knowledge
- ◆ Integration between literature and other resources is insufficient
 - Adequate for navigation purposes
 - Insufficient for knowledge processing



What additional services?

- ◆ Refined information retrieval
 - Indexing on relations in addition to concepts
 - *Find articles asserting that **IL-13** inhibits **COX-2***
- ◆ Multi-document summarization
 - Extract and visualize facts from the literature
 - *Summarize the top 300 papers on **panic disorder***
- ◆ Question answering
 - Clinical and biological questions
 - *What drugs **interact** with **imipramine**?*
- ◆ Knowledge discovery
 - Reasoning with facts from heterogeneous resources
 - *From MEDLINE and UMLS together*



Normalized and integrated knowledge

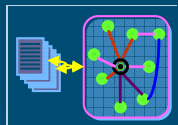
◆ Normalized knowledge

- Common format
- Common identification mechanism

◆ Integrated knowledge

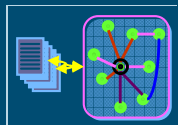
- Single repository
- Seamless environment
- *Phenotype and genotype information together*

Biomedical Knowledge Repository

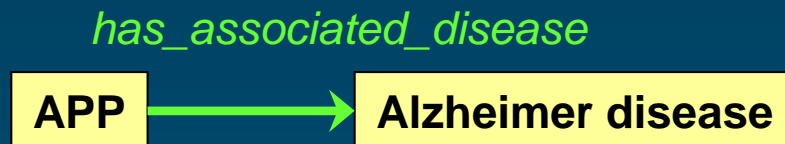


Sources of knowledge

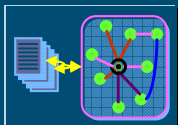
- ◆ Biomedical literature
 - Predications extracted from **MEDLINE** abstracts and full-text publicly available articles using text mining techniques
 - Other corpora (e.g., **ClinicalTrials.gov**)
- ◆ Terminological knowledge
 - **UMLS**
- ◆ Structured knowledge bases
 - NCBI resources (e.g., **Entrez Gene**)
 - Functional annotations from model organism databases
 - ...
- ◆ Contributed knowledge
 - The repository is open to collaborators outside NLM



Formalism Triples



- ◆ Facts
- ◆ Assertions
- ◆ Relations
- ◆ Semantic predications
- ◆ RDF triples



Annotated knowledge

◆ Provenance information

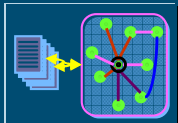
- Source (e.g., PMID)
- Extraction mechanism
- Timestamp

◆ Frequency information

- Redundancy

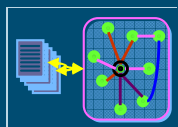
◆ Collaborative annotation

- “Was this information useful?”
- Context of use/usefulness

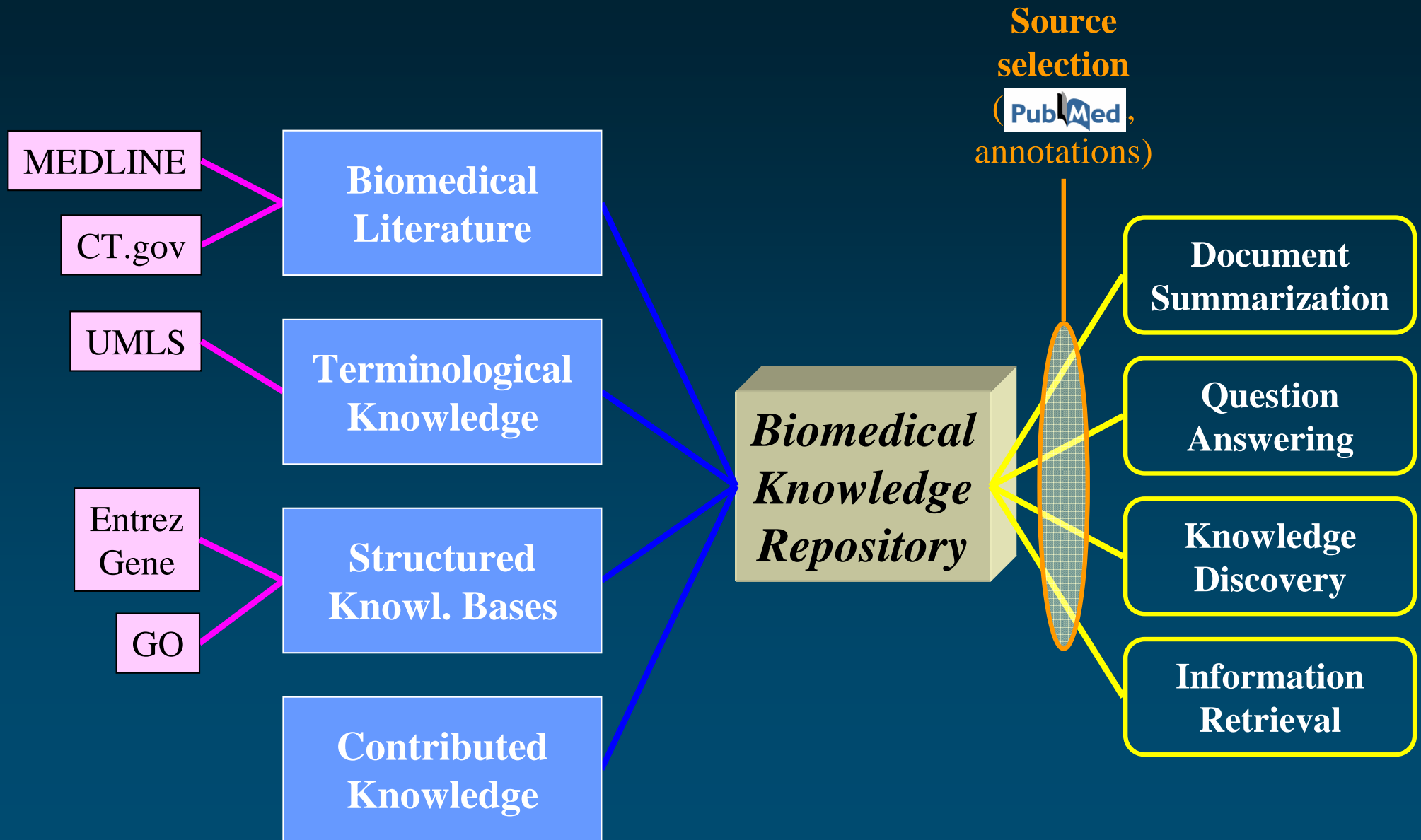


Semantic Web perspective

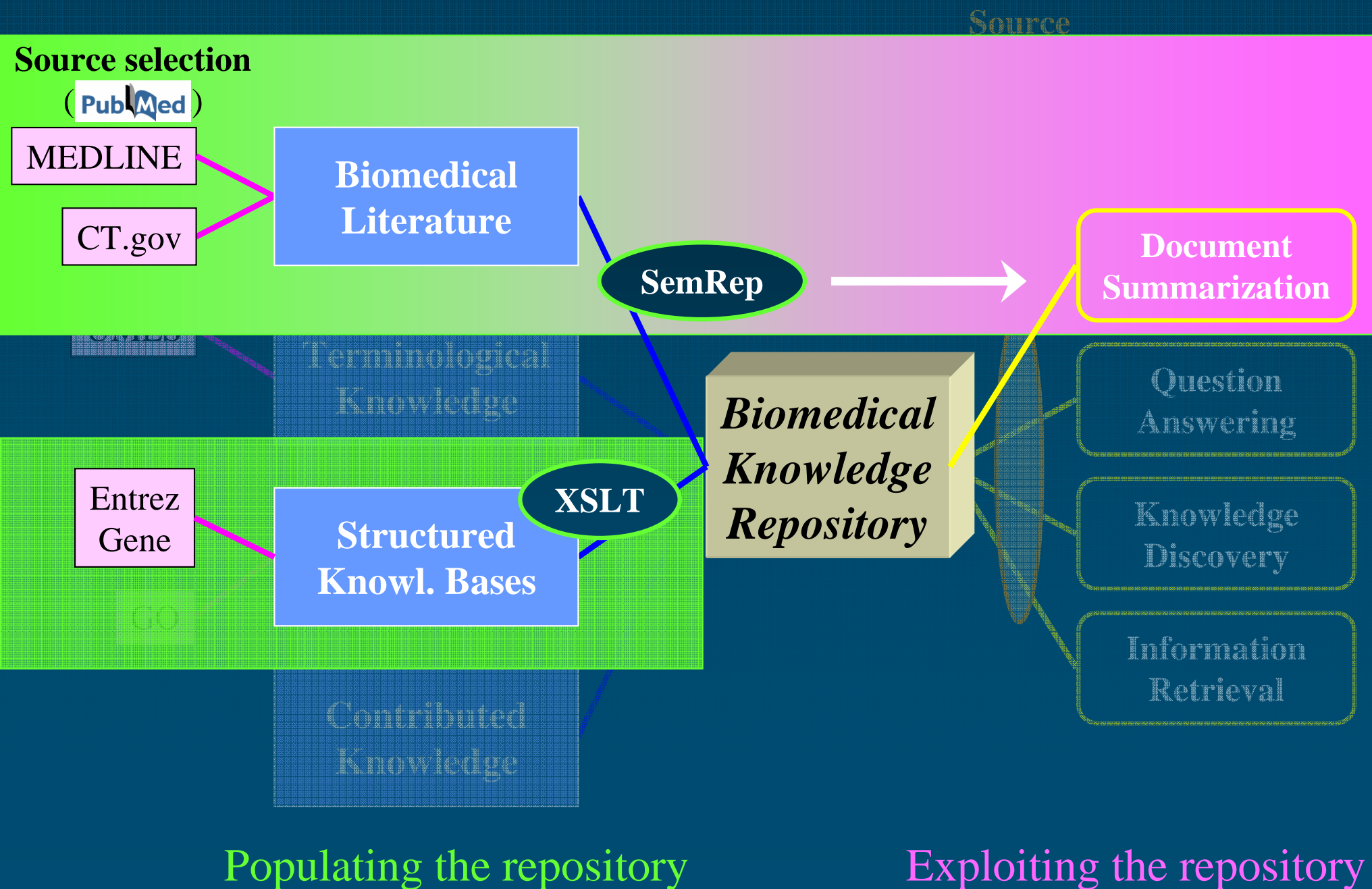
- ◆ Common format for knowledge
 - Resource Description Format (RDF)
- ◆ Common identification scheme
 - Unified Resource Identifier (URI)
- ◆ Standard tools
 - RDF browsers
 - RDF “reasoners”
- ◆ High level of interest for biomedicine in the SW community
 - Health Care and Life Sciences Interest Group



Advanced Library Services Summary



Advanced Library Services Pilot projects



Pilot #1

Populating and exploiting the
Biomedical Knowledge Repository

Converting Entrez Gene into RDF

With Satya Sahoo (U. Georgia)
and Kelly Zeng (LHC)

Search for

Display Show Send to

All: 1

☐ 1: **APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)** [*Homo sapiens*]
GeneID: 351 Primary source: [HGNC:620](#) updated 26-Jul-2006

[Entrez Gene Home](#)

[Table Of Contents](#)

Summary

Official Symbol: APP **and Name:** amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) **provided by** [HUGO Gene Nomenclature Committee](#)

See related: [HPRD:00100](#), [MIM:104760](#)

Gene type: protein coding

Gene name: APP

Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- HIV-1 protein interactions
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

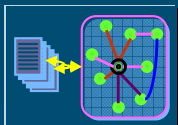
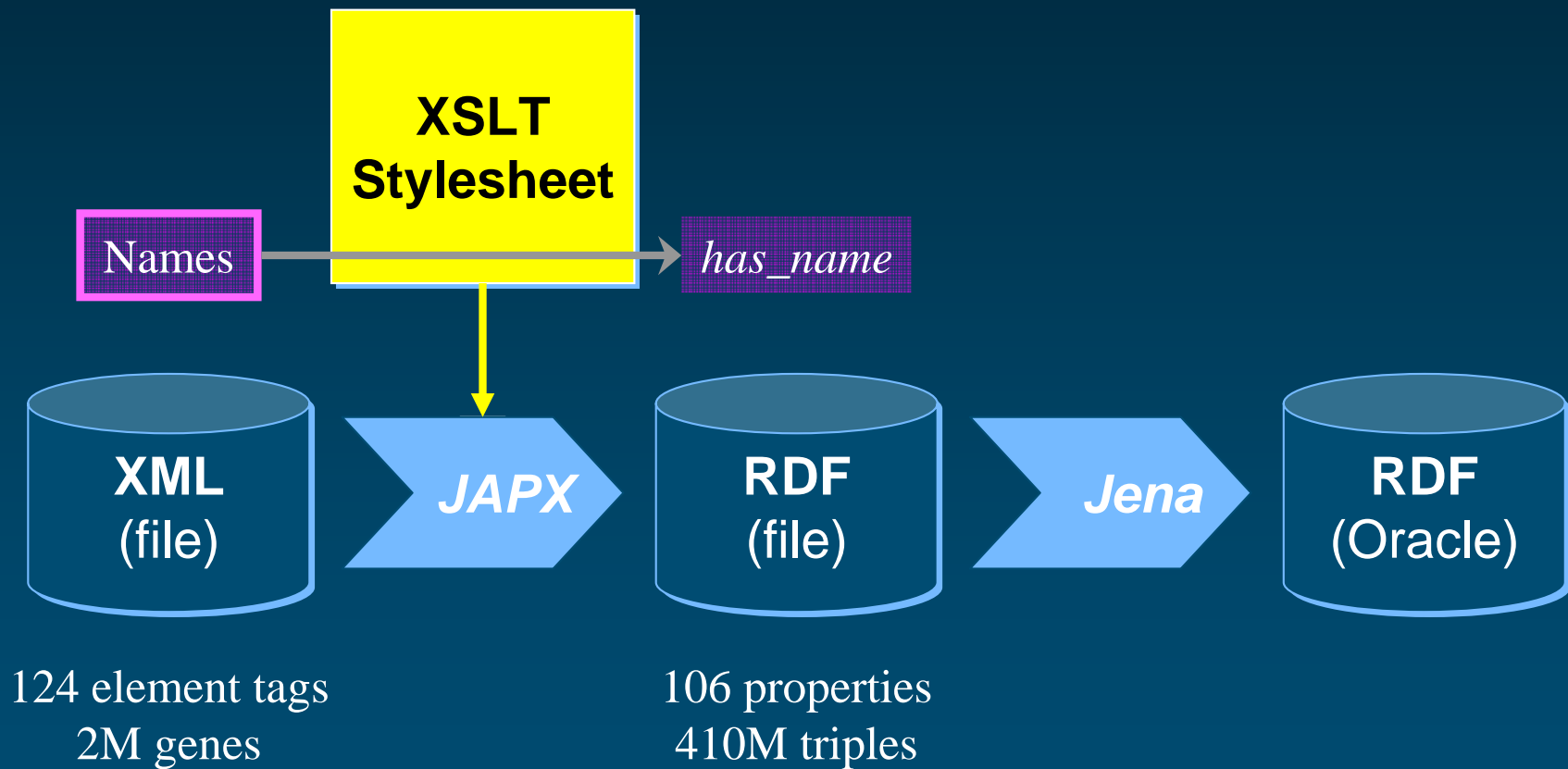
[Links](#)

General protein information

Names: amyloid beta A4 protein

protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

Overview



Search Gene for APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Show 5 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

1: APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [Homo sapiens]
GeneID: 351 Primary source: HGNC:620 updated 26-Jul-2006

Summary

Official Symbol: APP and Name: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) provided by HUGO Gene Nomenclature Committee

See related: HPRD:00100, MIM:104760

Gene type: protein coding

Gene name: APP

Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

has_protein_name

amyloid beta A4 protein

General protein information

Names: amyloid beta A4 protein

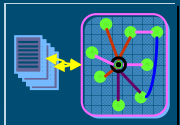
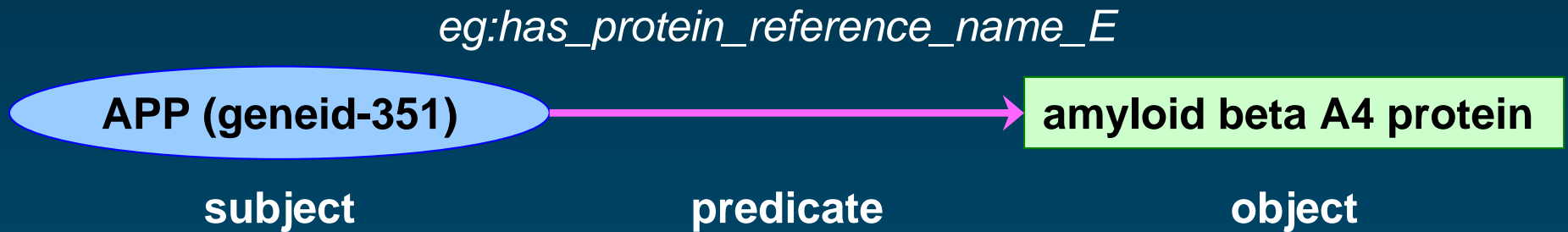
protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

Entrez Gene Home

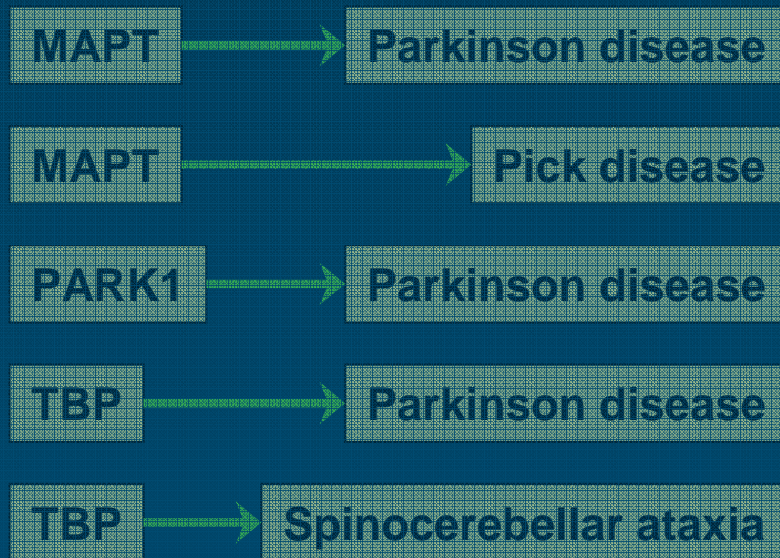
Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- HIV-1 protein interactions
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links
- Links

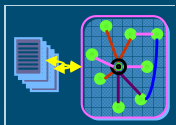
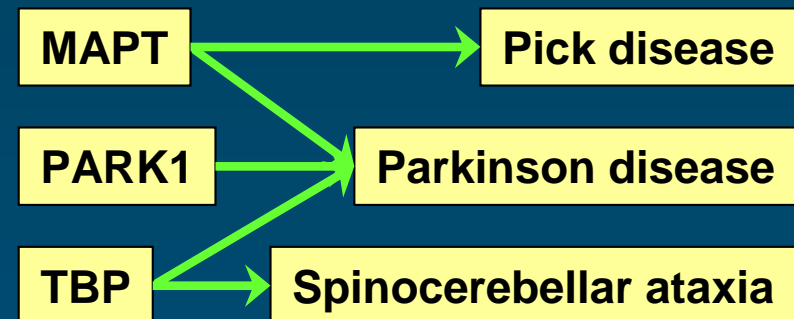
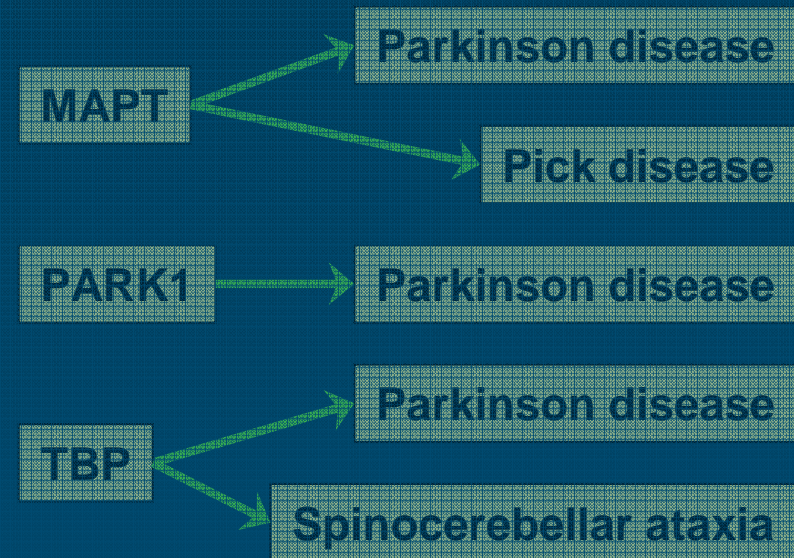
RDF triple Gene property



RDF graph Connecting several genes



has_associated_disease



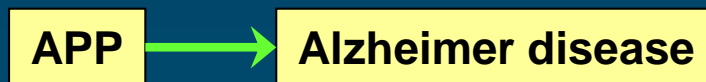
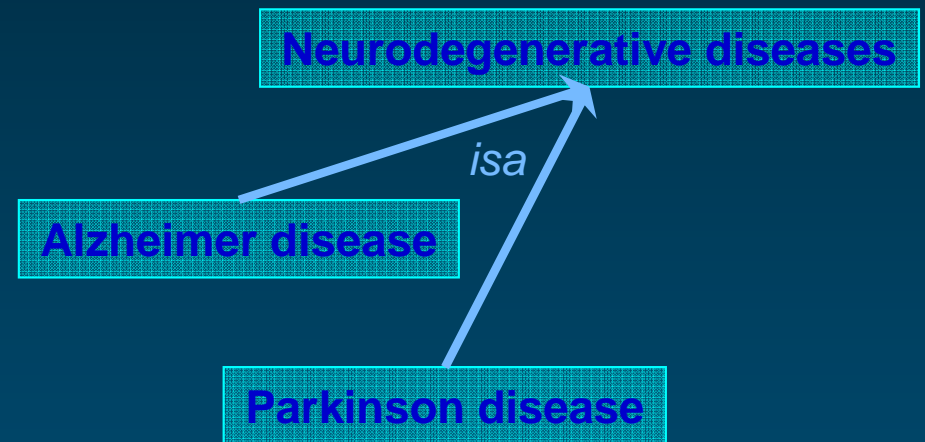
Future work

◆ Transform additional resources into RDF

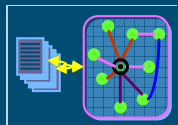
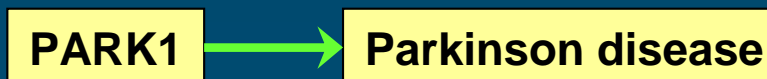
- UMLS Metathesaurus
- Other NCBI databases
- Drug knowledge bases
- ...

◆ Integrate resources

- Query across resources



has_associated_disease



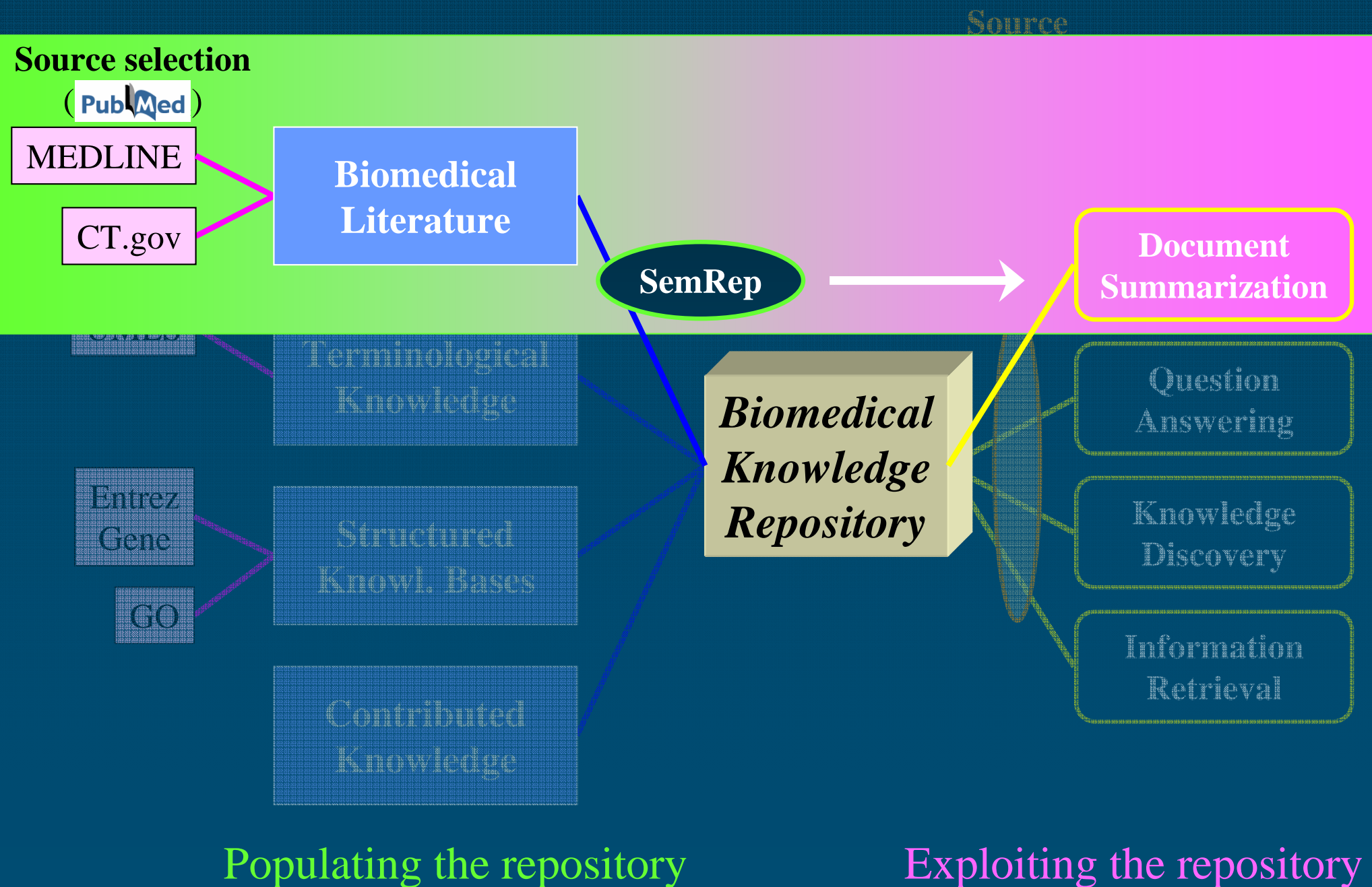
Pilot #2

Populating and exploiting the Biomedical Knowledge Repository

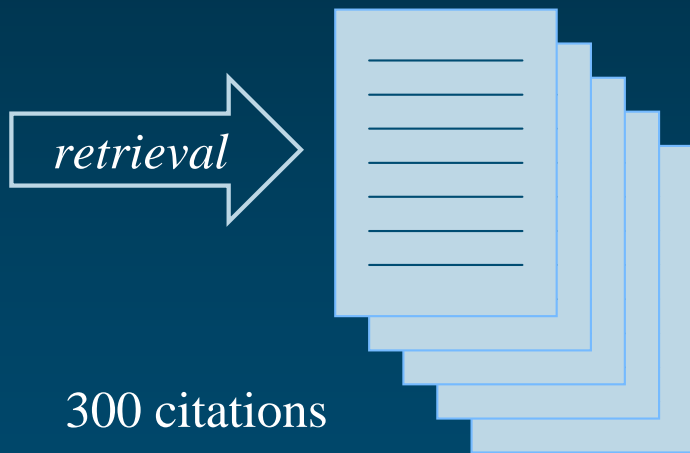
*Semantic Medline:
Multi-document summarization
and visualization*

With Marcelo Fiszman, M.D., Ph.D.
and Halil Kilicoglu, M.S.

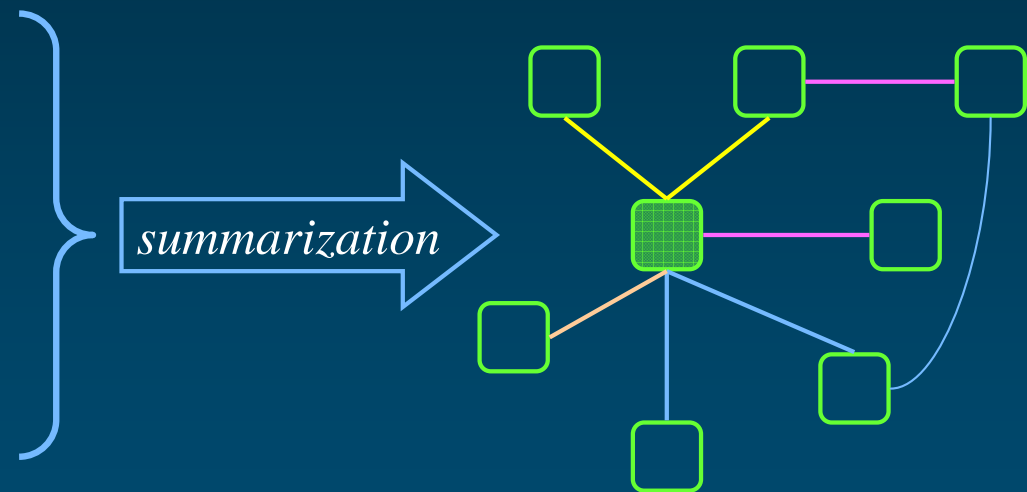
Advanced Library Services Pilot projects



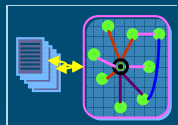
Managing retrieval results



Information retrieval



Semantic Medline

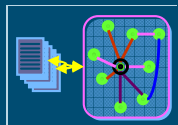
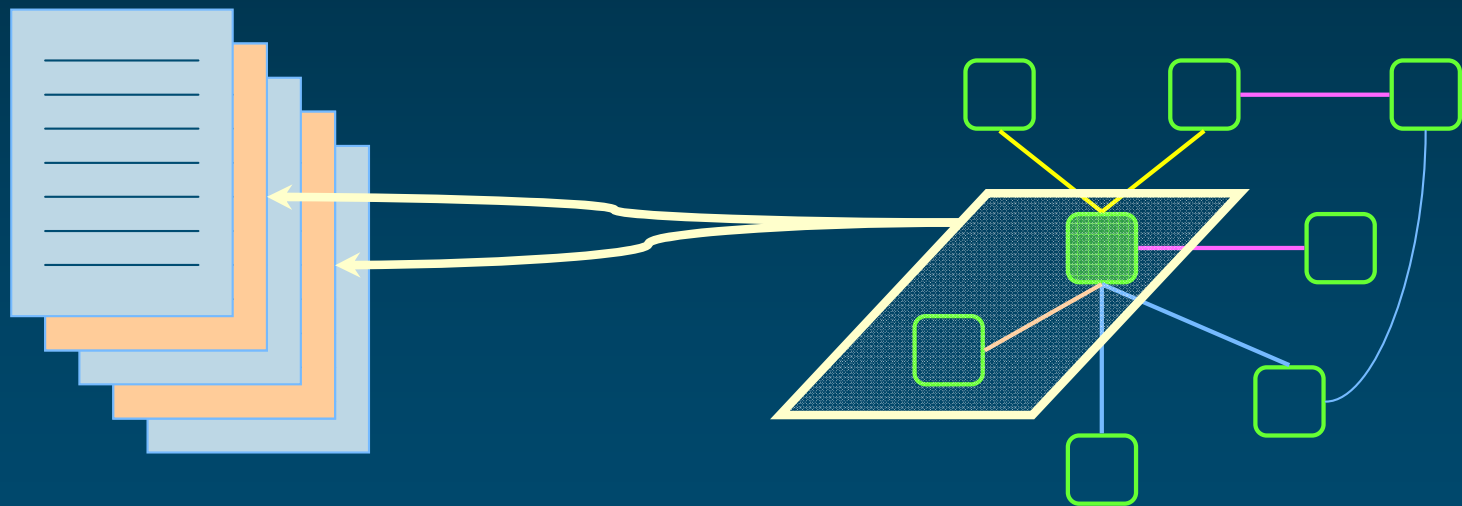


Managing retrieval results

NCBI PubMed National Library of Medicine NLM

All Databases PubMed Nucleotide Protein Genome Structure

Search PubMed for panic disorder Go Clear



Guiding principles

◆ Visualization

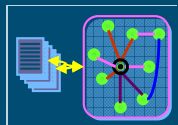
- Overview first
- Details on demand

[Shneiderman 1996]

◆ Integration of knowledge content [BoSC, April, 2006]

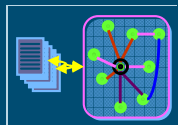
◆ Automated management of knowledge from text

◆ Seamless application interfaces

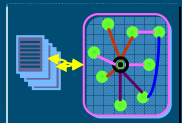
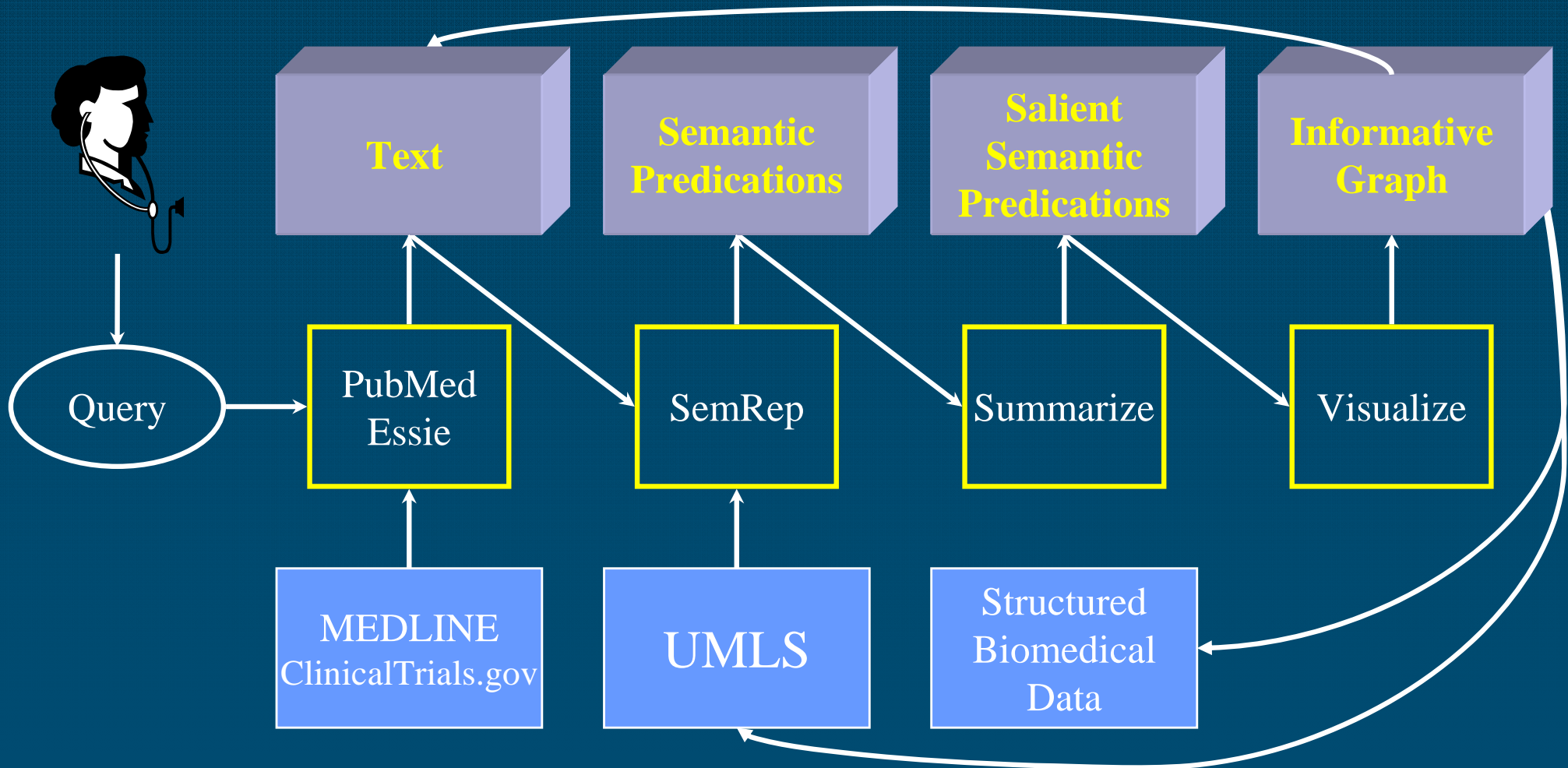


Seamless integration of technologies

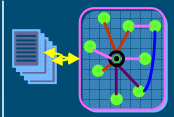
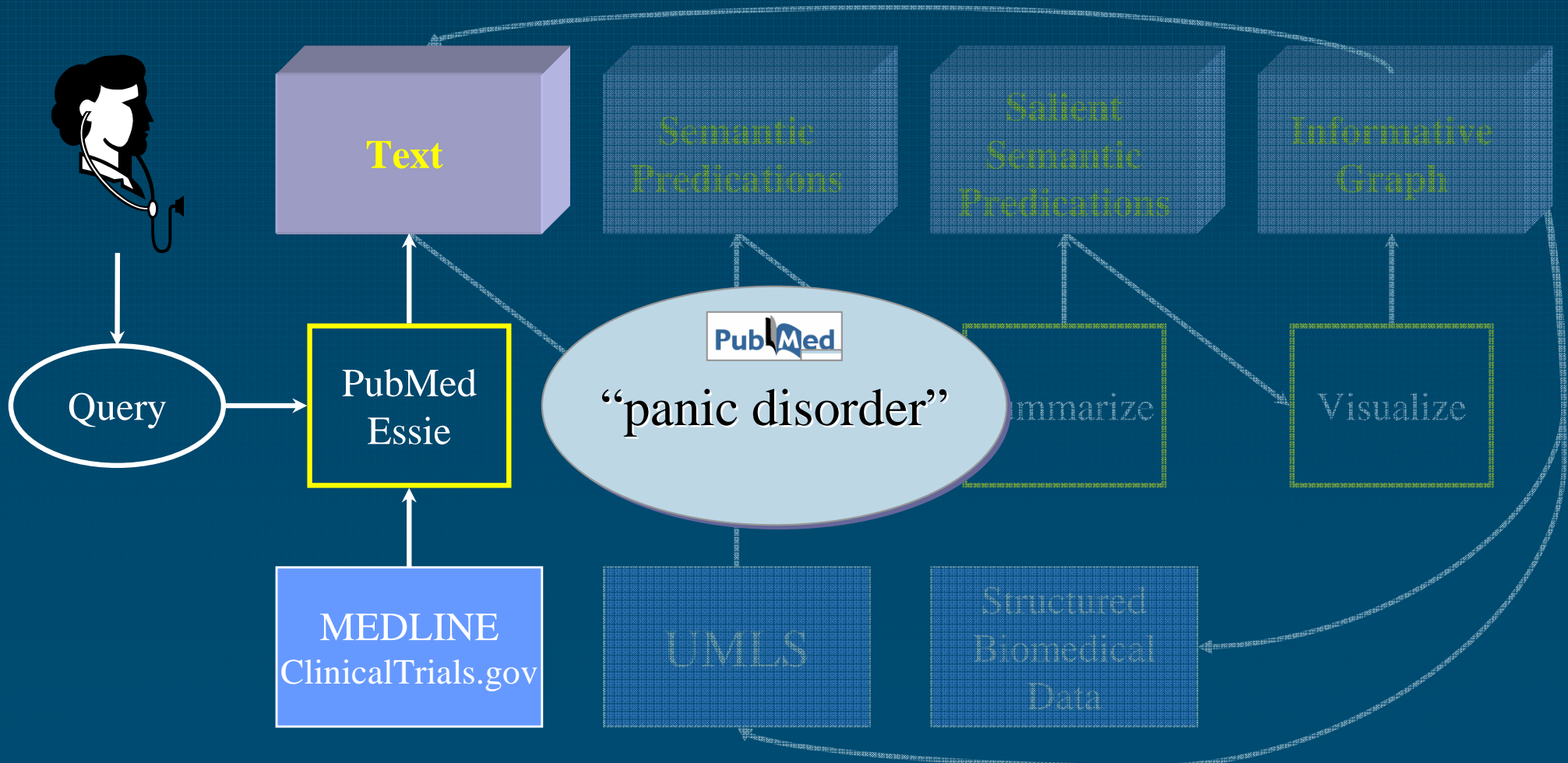
- ◆ Information retrieval
 - PubMed - MEDLINE
 - Essie - ClinicalTrials.gov
- ◆ Natural language processing: [SemRep](#)
 - Represent content of text with semantic predications
- ◆ Abstraction summarization
 - Informative: Overview of most salient information
- ◆ Visualization
 - Indicative: Links to source text and additional information



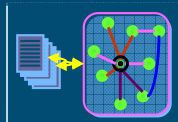
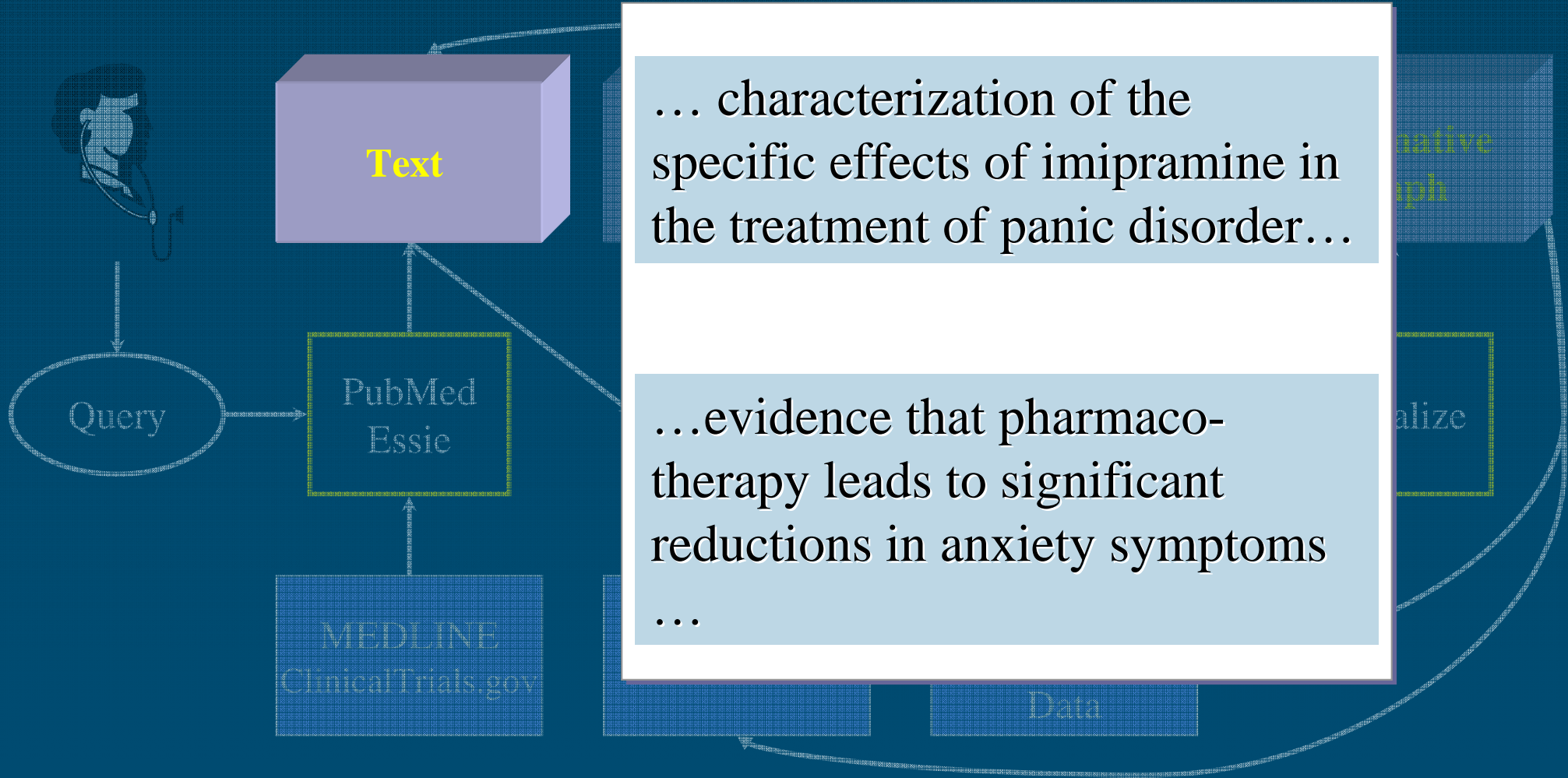
Semantic Medline Overview



Document selection



MEDLINE citations



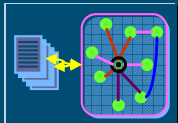
The diagram illustrates the SemRep system architecture. At the bottom, two input sources, "MEDLINE ClinicalTrials.gov" and "Structured Biomedical Data", feed into a central blue box labeled "UMLS". From "UMLS", an arrow points up to a yellow-outlined box labeled "SemRep". "SemRep" has two outgoing arrows: one to a purple box labeled "Text" and another to a purple box labeled "Semantic Predications". The "Text" box has an arrow pointing to a "Query" oval on the left, which is associated with a doctor icon. The "Semantic Predications" box has an arrow pointing to a "Summarize" box, which in turn points to a "Salient Semantic Predications" box. The "Summarize" box also has an arrow pointing to a "Visualize" box, which points to an "Informative Graph" box. Finally, the "Informative Graph" box has an arrow pointing back to the "Text" box, completing a cycle.

Semantic interpretation

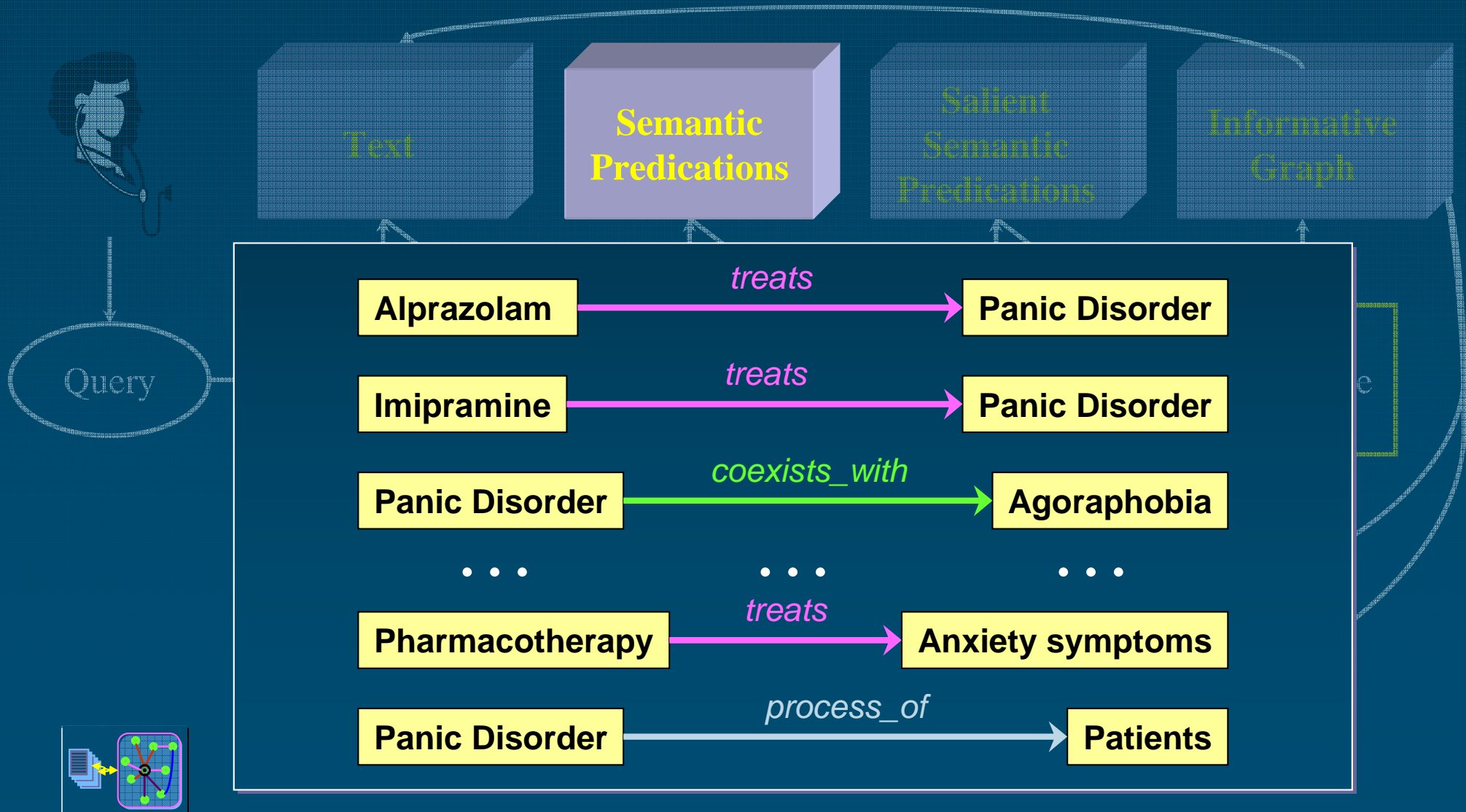
... characterization of the specific effects of
imipramine in the treatment of panic disorder ...



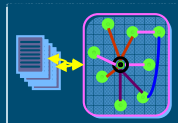
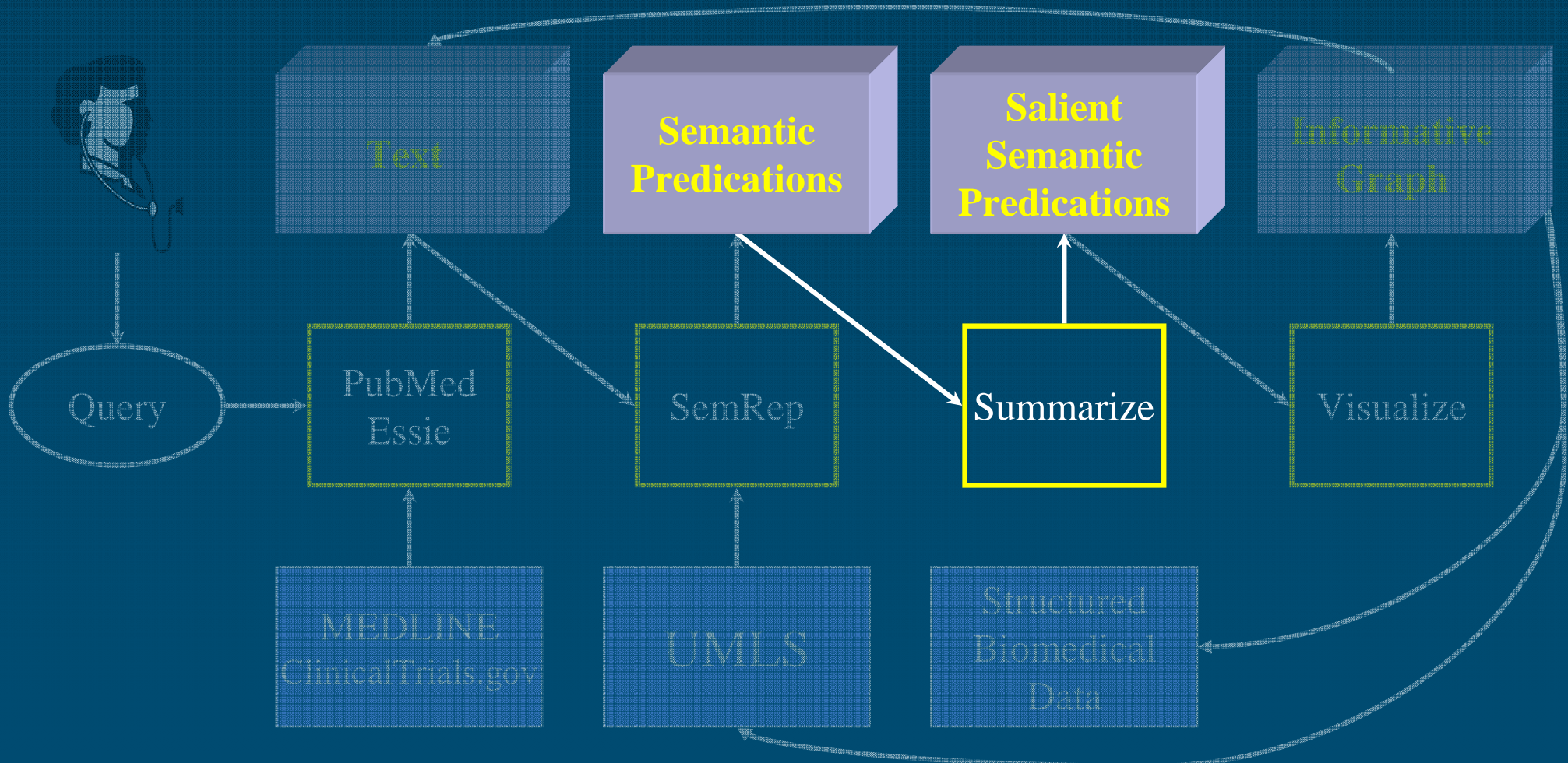
...evidence that pharmacotherapy leads to significant
reductions in anxiety symptoms ...



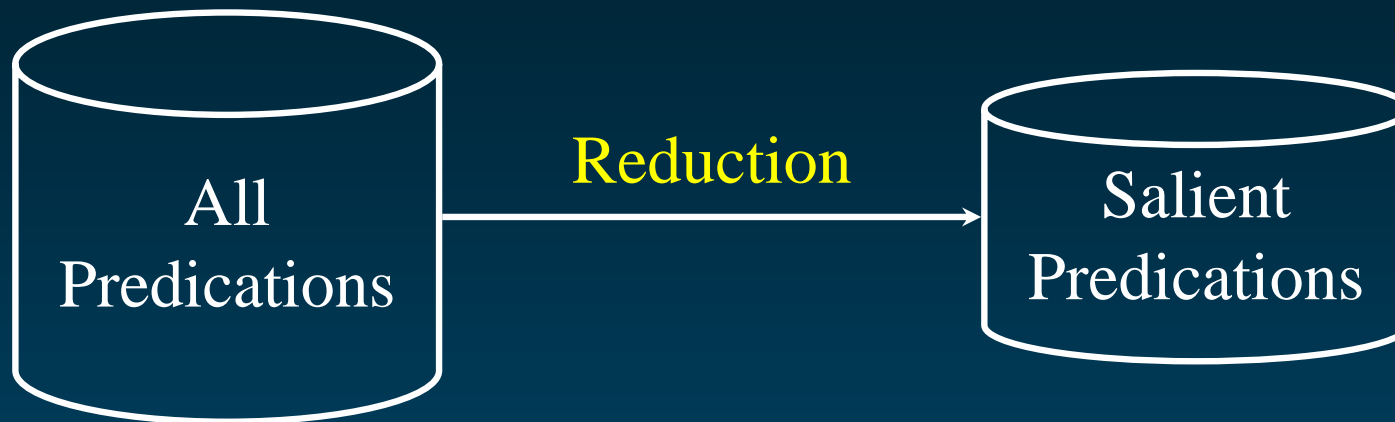
Semantic predications



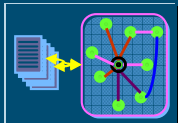
Summarization



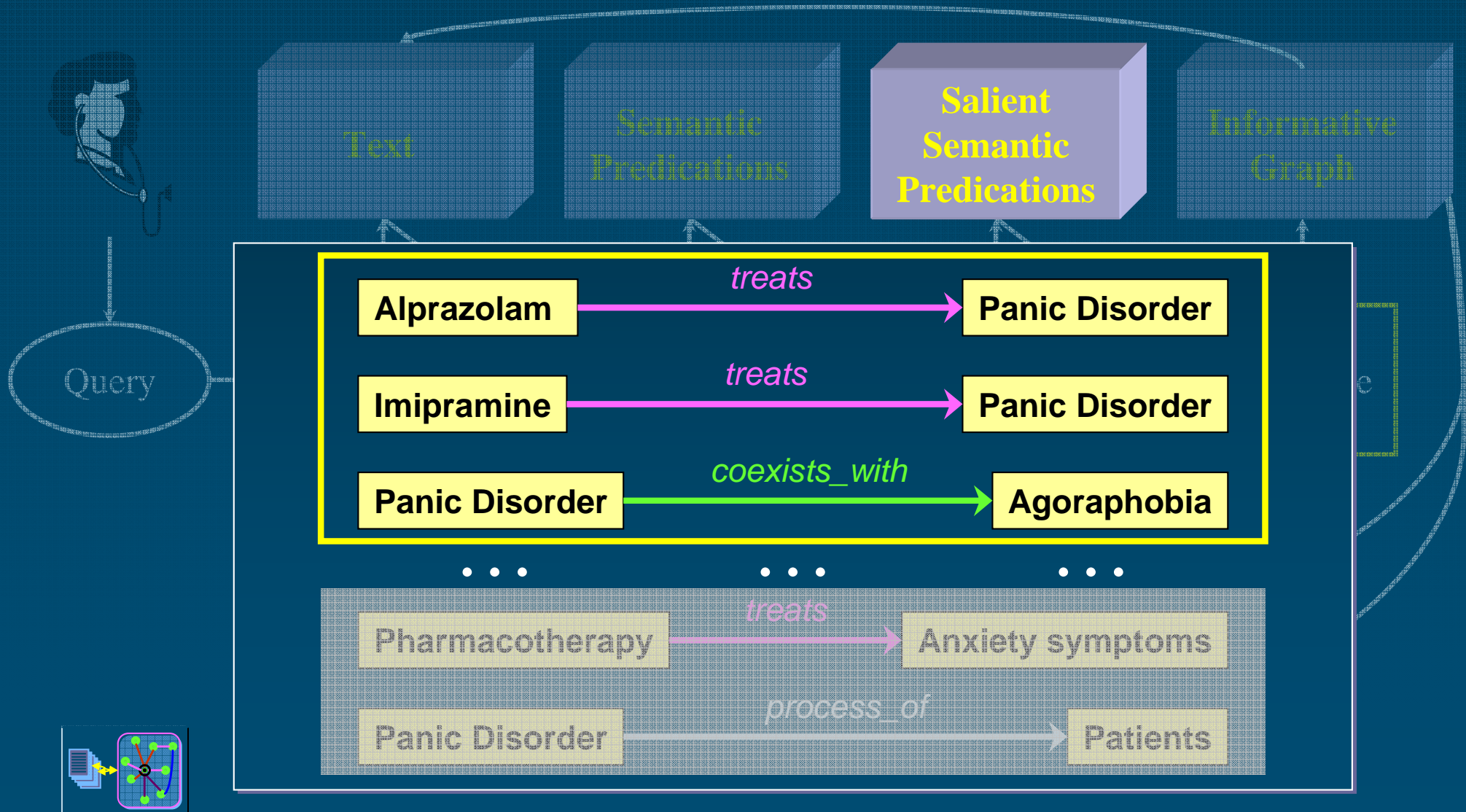
Abstraction summarization



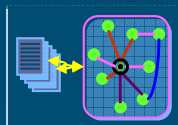
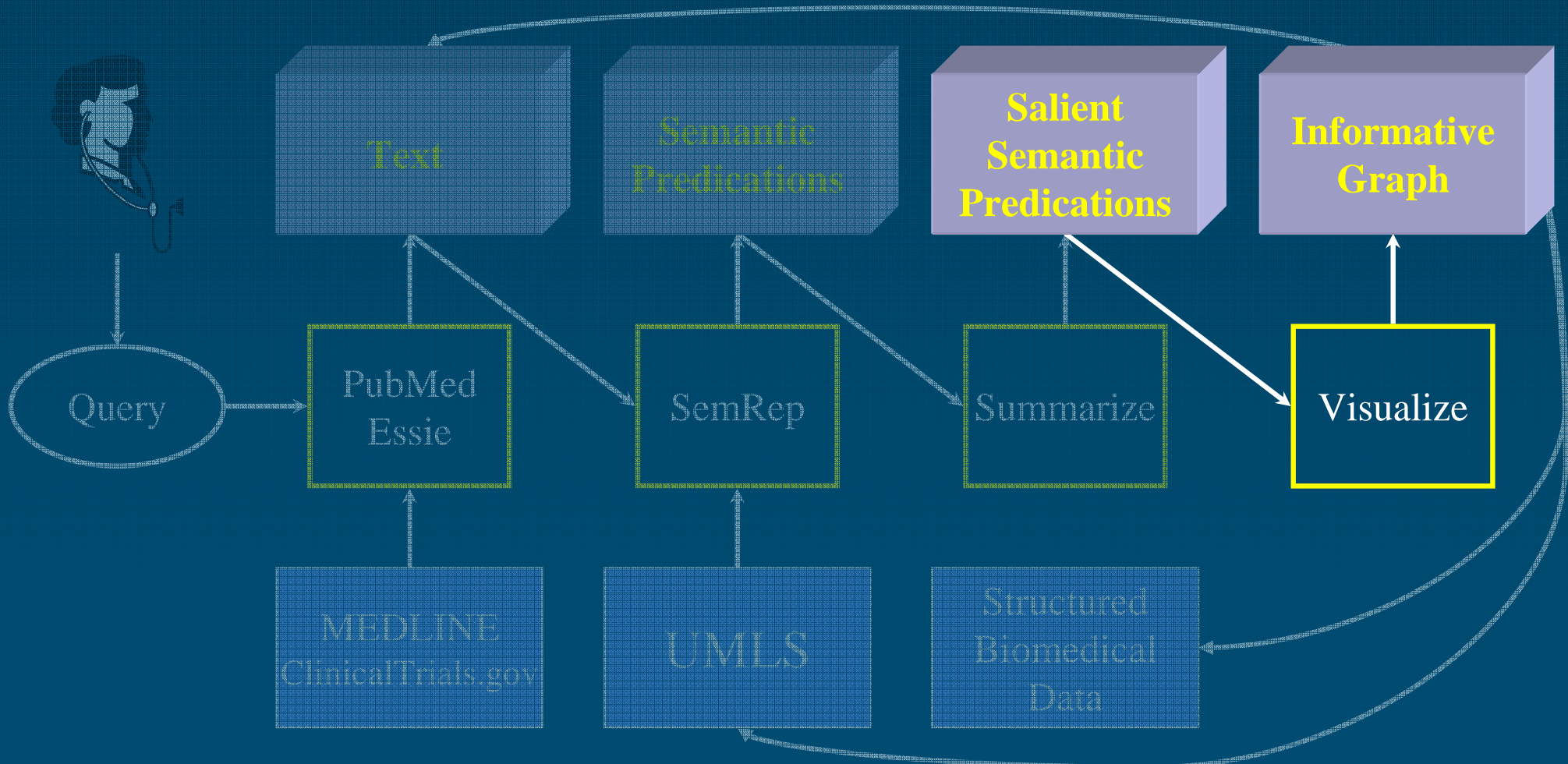
- ◆ Specify a topic
- ◆ Retain predications on the topic
- ◆ Eliminate uninformative predications
- ◆ Retain most frequent predications



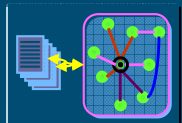
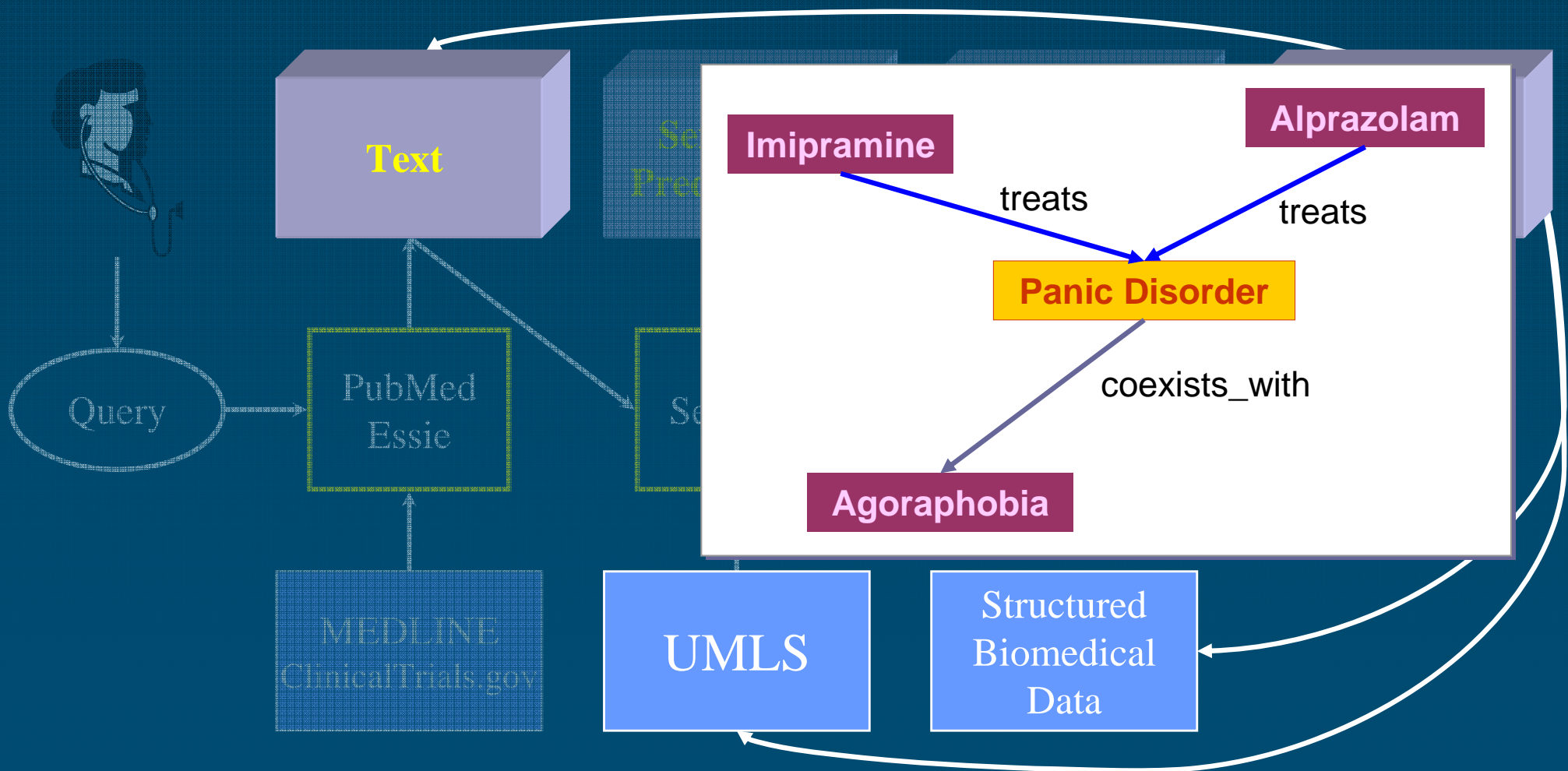
Salient semantic predications



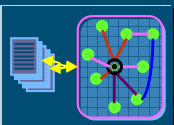
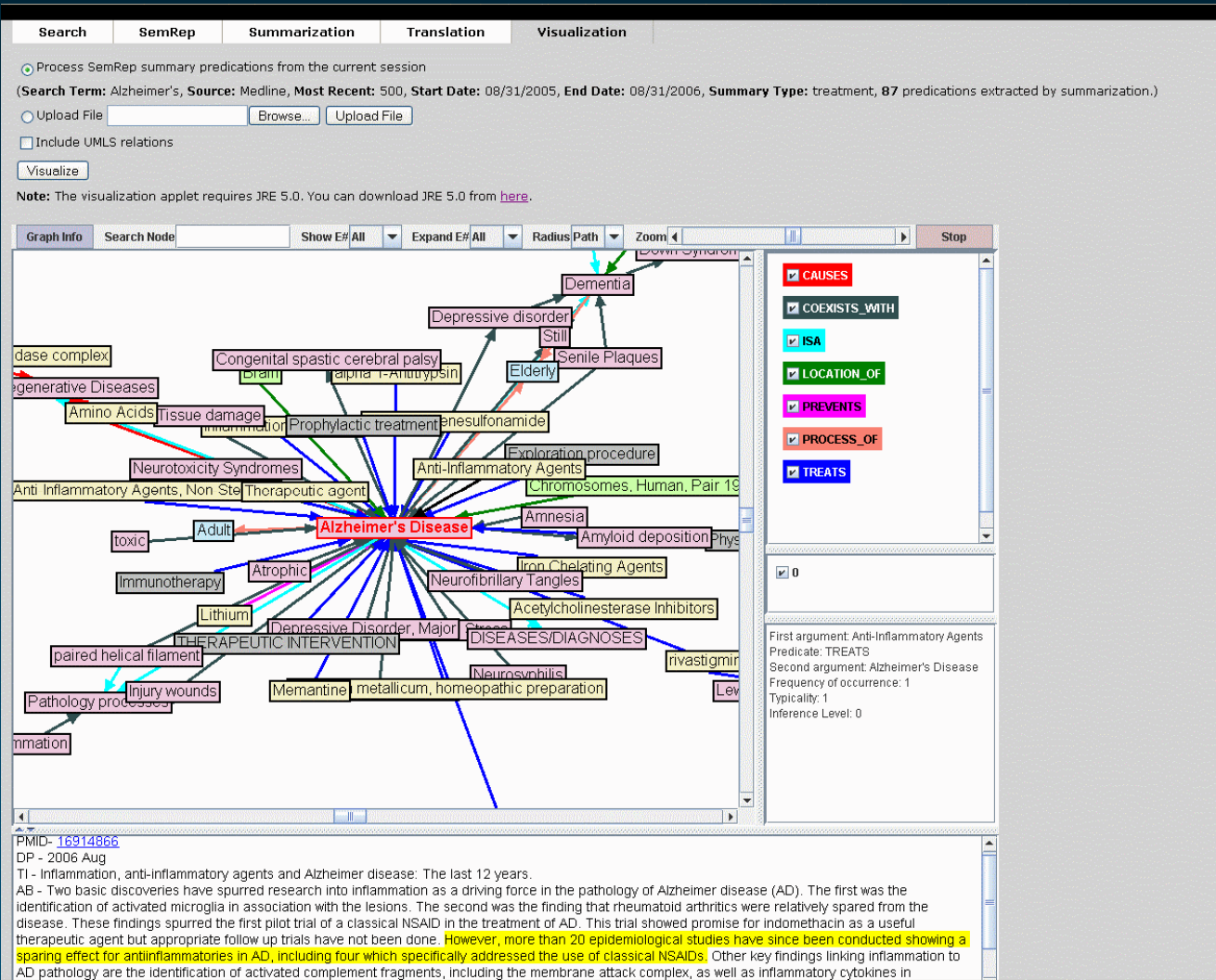
Visualization



Informative graph

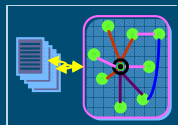


Live



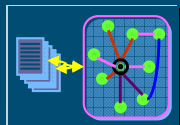
Related research Visualizing relations

- ◆ Maps of linked concepts among document
[Fuller et al. 2004]
- ◆ Literature network of co-occurring genes
[Jensen et al. 2001]
- ◆ Associative concept space for discovery
[van der Eijk et al. 2004]
- ◆ Genomic information across structured and textual databases
[Tao et al. 2005]



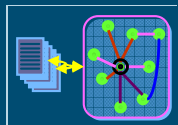
Future work

- ◆ Process all of MEDLINE/PubMed
 - With SemRep
- ◆ Incrementally integrate structured knowledge sources
 - Entrez databases
 - UMLS
 - Genetics Home Reference
- ◆ Implementation
 - Efficiency
 - Large amount of data



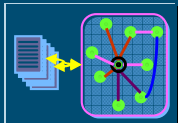
Summary

- ◆ Deliver health information
 - Biomedical Knowledge Repository
 - Advanced Library Services
- ◆ Exploit
 - Current Library resources
 - Advanced information technology
- ◆ Support timely translation
 - Of biomedical research
 - Into improvements in patient care and public health



Acknowledgments

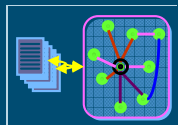
- ◆ Caroline Ahlers
- ◆ Mariana Dimitrov
- ◆ Marcelo Fiszman
- ◆ Halil Kilicoglu
- ◆ François-Michel Lang
- ◆ Lee Peters
- ◆ Anna Ripple
- ◆ Graciela Rosemblat
- ◆ Satya Sahoo
- ◆ Kelly Zeng

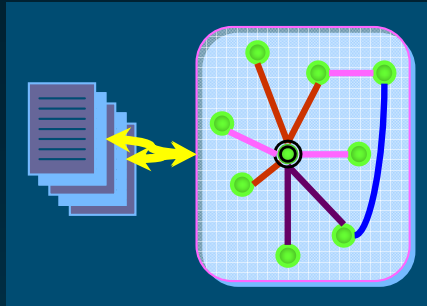


References

- ◆ Bodenreider O, Rindflesch TC. *Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications*. Technical report. Bethesda, Maryland: Lister Hill National Center for Biomedical Communications, National Library of Medicine; September 14, 2006.

<http://lhncbc.nlm.nih.gov/lhc/docs/reports/2006/tr2006001.pdf>





Advanced Library Services

Olivier Bodenreider

olivier@nlm.nih.gov

Thomas C. Rindflesch

tcr@nlm.nih.gov



Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA