



RESEARCH, CONDITION, AND DISEASE CATEGORIZATION

July 17, 2007

NLM Resources for Mining Biomedical Text



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

Overview

- ◆ An example
- ◆ Three types of resources for mining biomedical text
 - Lexical resources
 - SPECIALIST Lexicon
 - Lexical Tools
 - Terminological resources
 - UMLS Metathesaurus
 - MetaMap, MTI
 - Ontological resources
 - UMLS Semantic Network
 - SemRep
- ◆ Application: Semantic Medline



An example

Neurofibromatosis 2

Neurofibromatosis 2

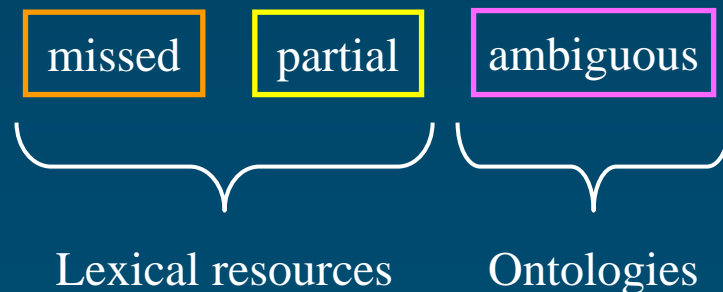
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

[Uppal, S., and A. P. Coatesworth. "Neurofibromatosis Type 2." *Int J Clin Pract*, 57, no. 8, 2003, pp. 698-703.]



Entity recognition

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.



Relation extraction

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

- vestibular schwannomas *manifestation of* neurofibromatosis 2
- neurofibromatosis 2 *associated with* mutation of NF2 gene
- NF2 gene *located on* chromosome 22

NLM resources for mining biomedical text

Types of resources

◆ Lexical resources

- Collections of lexical items
- Additional information
 - Part of speech
 - Spelling variants
- Useful for entity recognition
- UMLS SPECIALIST Lexicon, WordNet

◆ Ontological resources

- Collections of
 - kinds of entities (substances, qualities, processes)
 - relations among them
- Useful for **relation extraction**
- UMLS Semantic Network, SNOMED CT



Unified Medical Language System

◆ SPECIALIST Lexicon

- 200,000 lexical items
- Part of speech and variant information

◆ Metathesaurus

- 5M names from over 100 terminologies
- 1.5M concepts
- 16M relations

◆ Semantic Network

- 135 high-level categories
- 7000 relations among them

Lexical
resources

Terminological
resources

Ontological
resources



Lexical resources

*SPECIALIST
Lexicon*

- *Lexical tools*



Lexical Systems Group

<http://umlslex.nlm.nih.gov/>

SPECIALIST Lexicon

◆ Content

- English lexicon
- Many words from the biomedical domain

◆ 200,000+ lexical items

◆ Word properties

- morphology
- orthography
- syntax

```
{  
  base=hemoglobin           (base form)  
  spelling_variant=haemoglobin  
  entry=E0031208            (identifier)  
  cat=noun                  (part of speech)  
  variants=uncount          (no plural)  
  variants=reg              (plural: hemoglobins , haemoglobins)  
}
```

◆ Used by the lexical tools



Lexical tools

- ◆ To manage lexical variation in biomedical terminologies
- ◆ Major tools
 - Normalization
 - Indexes
 - Lexical Variant Generation program (lvg)
- ◆ Based on the SPECIALIST Lexicon
- ◆ Used by noun phrase extractors, search engines



Terminological resources

*UMLS
Metathesaurus*

- *MetaMap*
- *Medical Text Indexer (MTI)*


[http://www.nlm.nih.gov/
research/umls/](http://www.nlm.nih.gov/research/umls/)

INDEXING INITIATIVE

<http://ii.nlm.nih.gov/>



Source Vocabularies

(2007AB)

- ◆ 143 source vocabularies
 - 17 languages
- ◆ Broad coverage of biomedicine
 - 5.9M names
 - 1.4M concepts
 - 16M relations
- ◆ Common presentation

Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

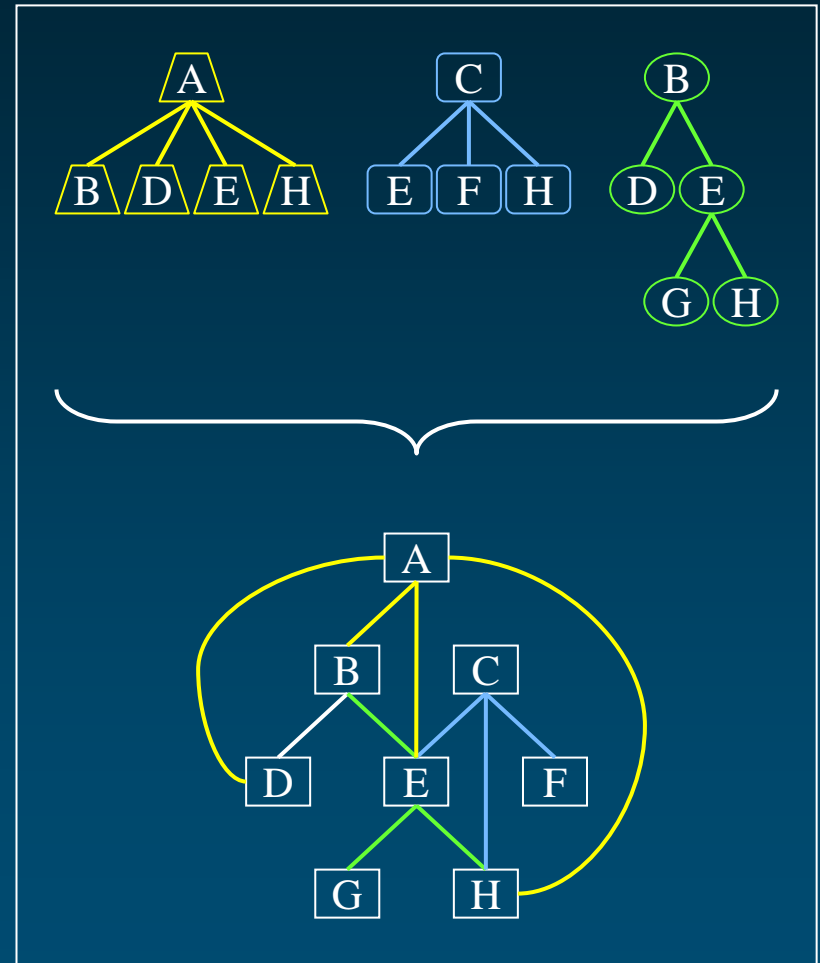
C0001403

Addison's disease

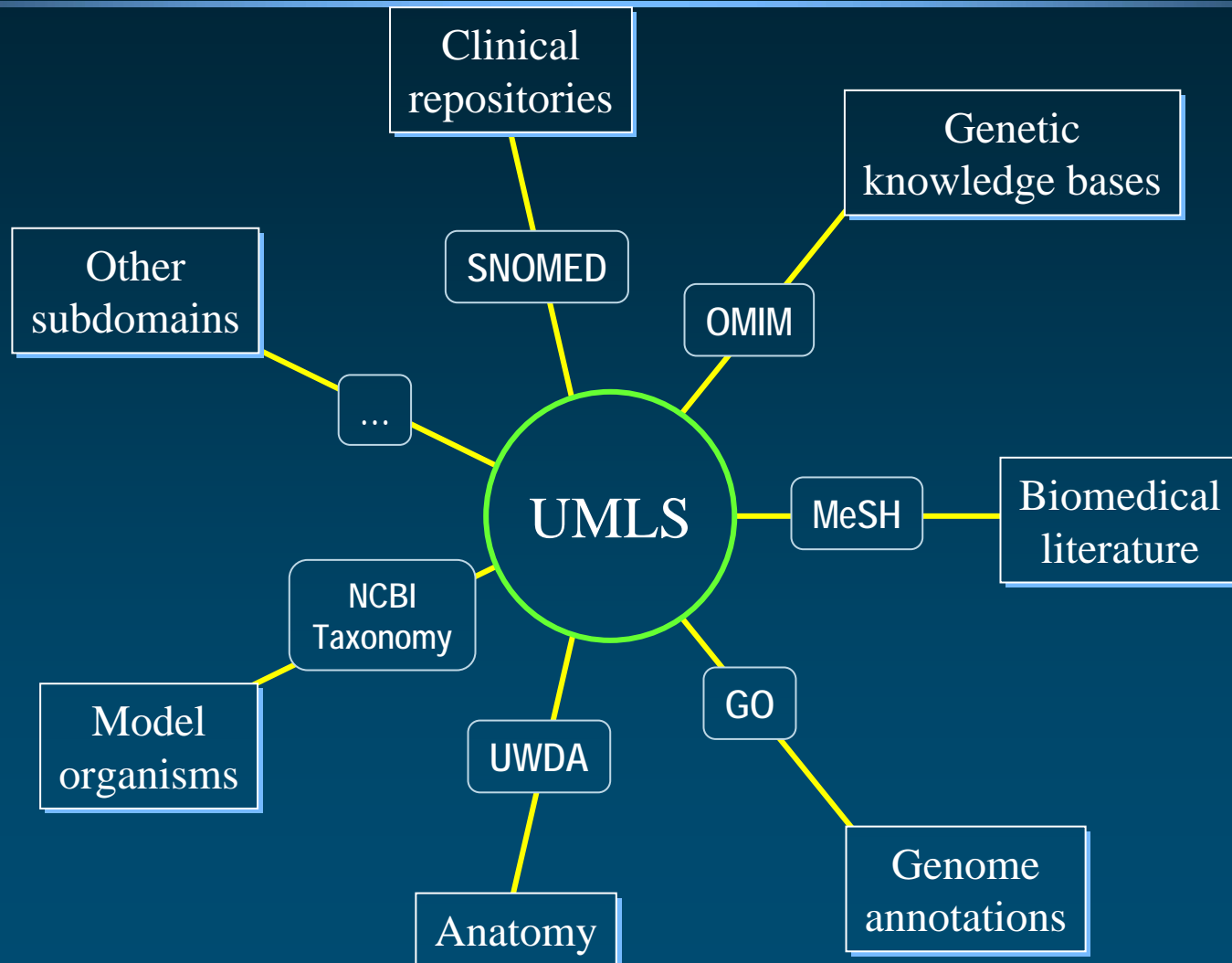


Organize concepts

- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)



Integrating subdomains



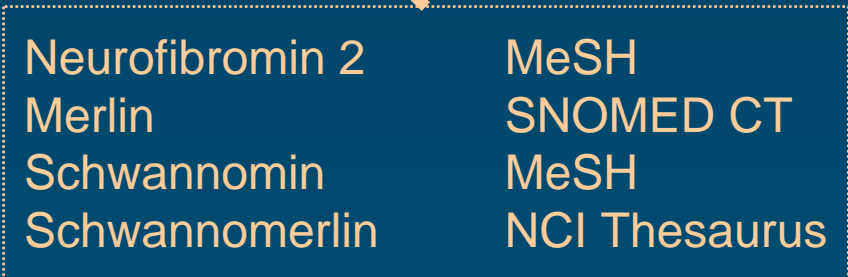
MetaMap

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

INDEXING INITIATIVE

<http://ii.nlm.nih.gov/>

C0254123



Neurofibromin 2	MeSH
Merlin	SNOMED CT
Schwannomin	MeSH
Schwannomerlin	NCI Thesaurus



Medical Text Indexer

- ◆ Semi-automatic indexing of MEDLINE citations
 - Suggest MeSH main headings
 - Complement to manual indexing
 - Integrated into the DCMS indexing environment
- ◆ Automatic indexing of collections not indexed previously

INDEXING INITIATIVE

<http://ii.nlm.nih.gov/>



Ontological resources

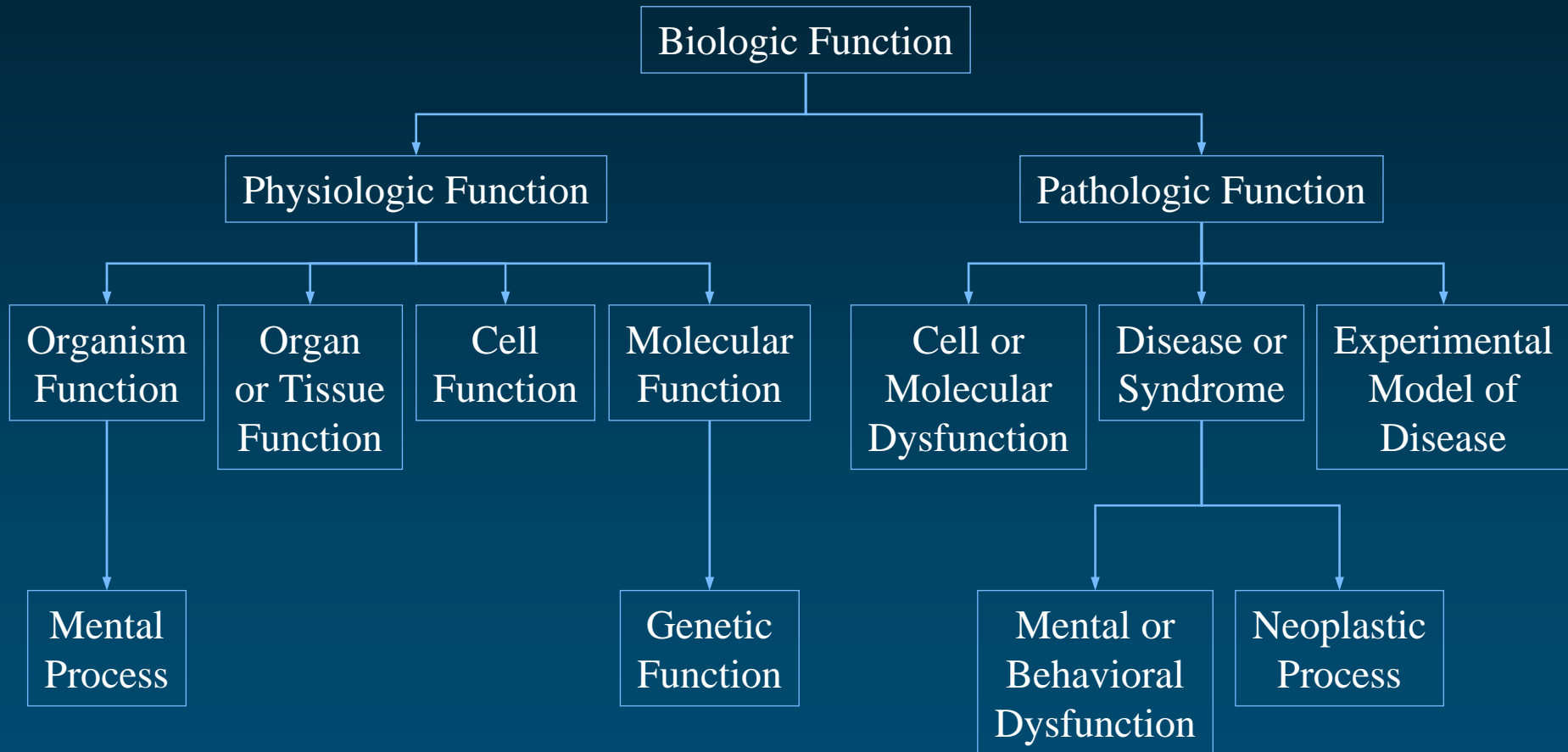
*UMLS
Semantic
Network*

<http://semanticnetwork.nlm.nih.gov/>

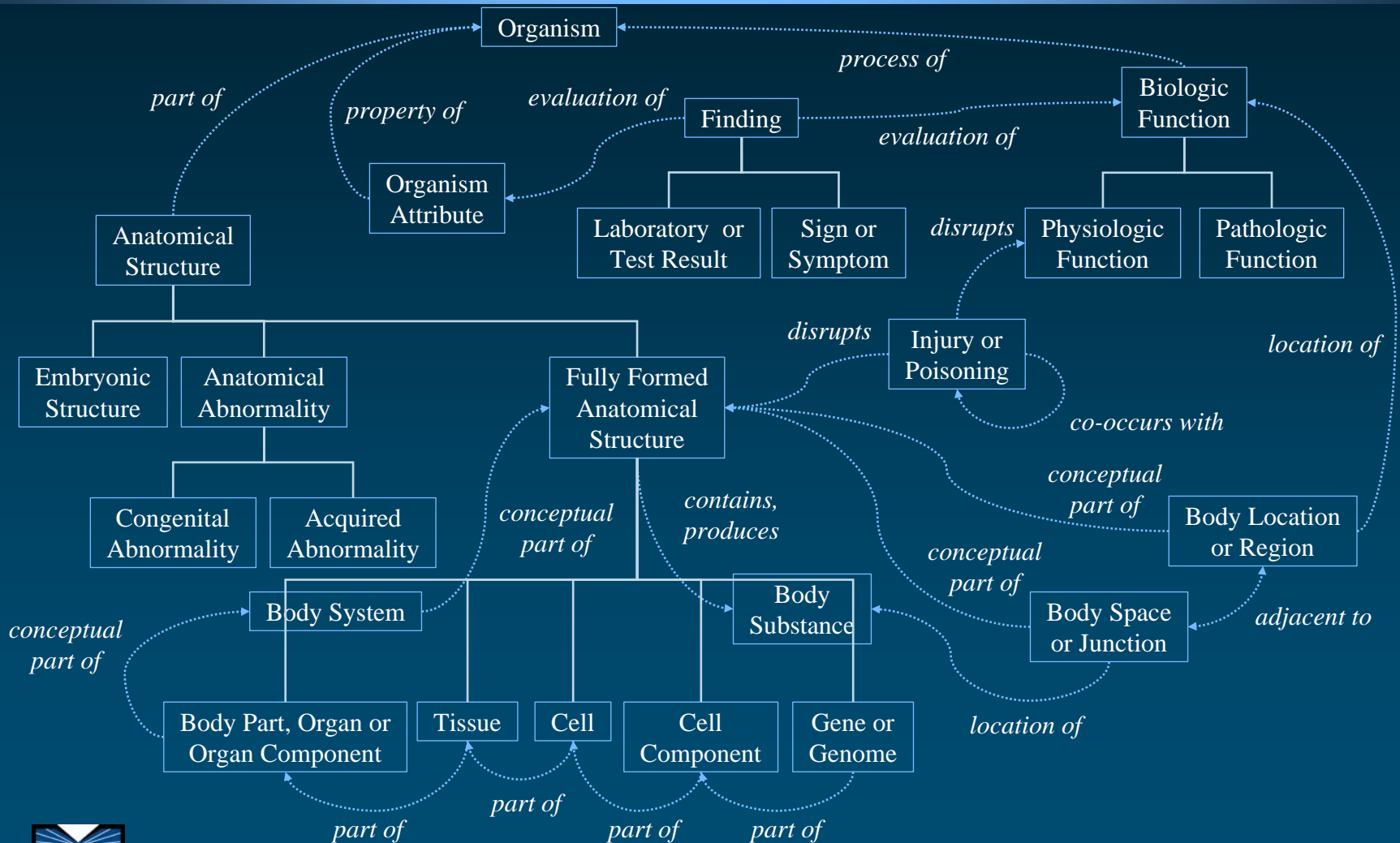
• *SemRep*

<http://skr.nlm.nih.gov/>

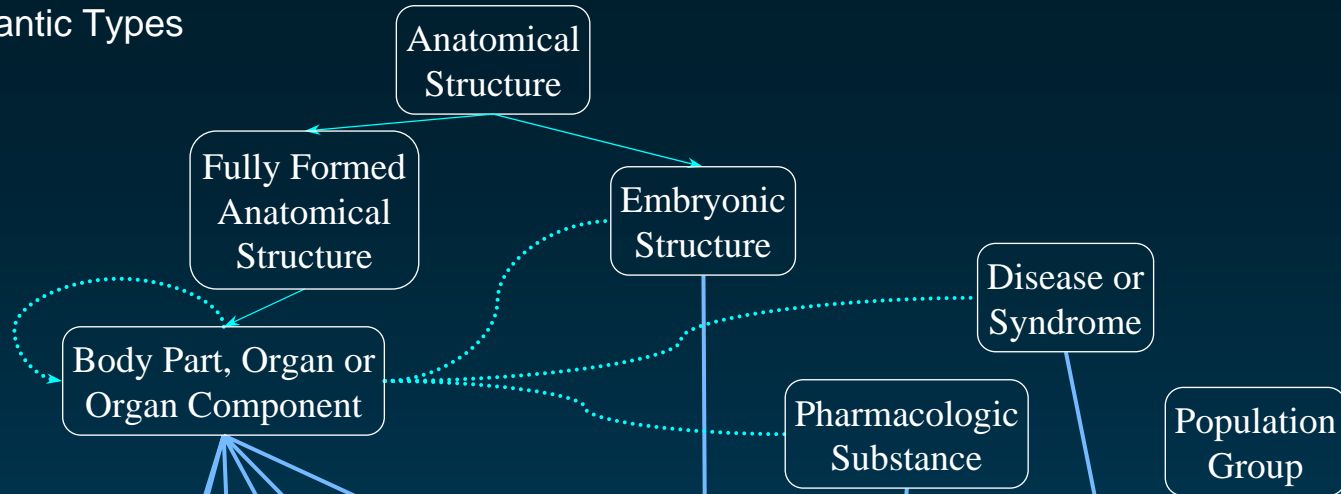
“Biologic Function” hierarchy (isa)



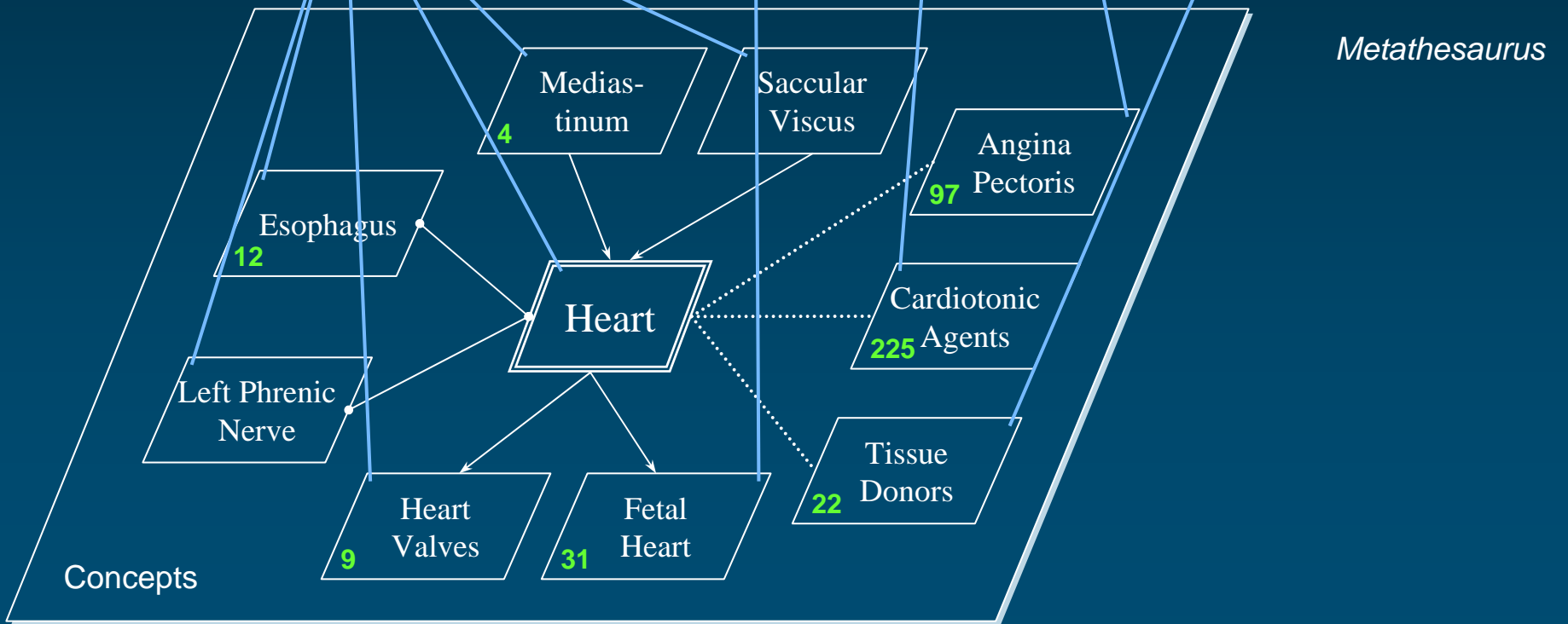
Associative (non-isa) relationships



Semantic Types



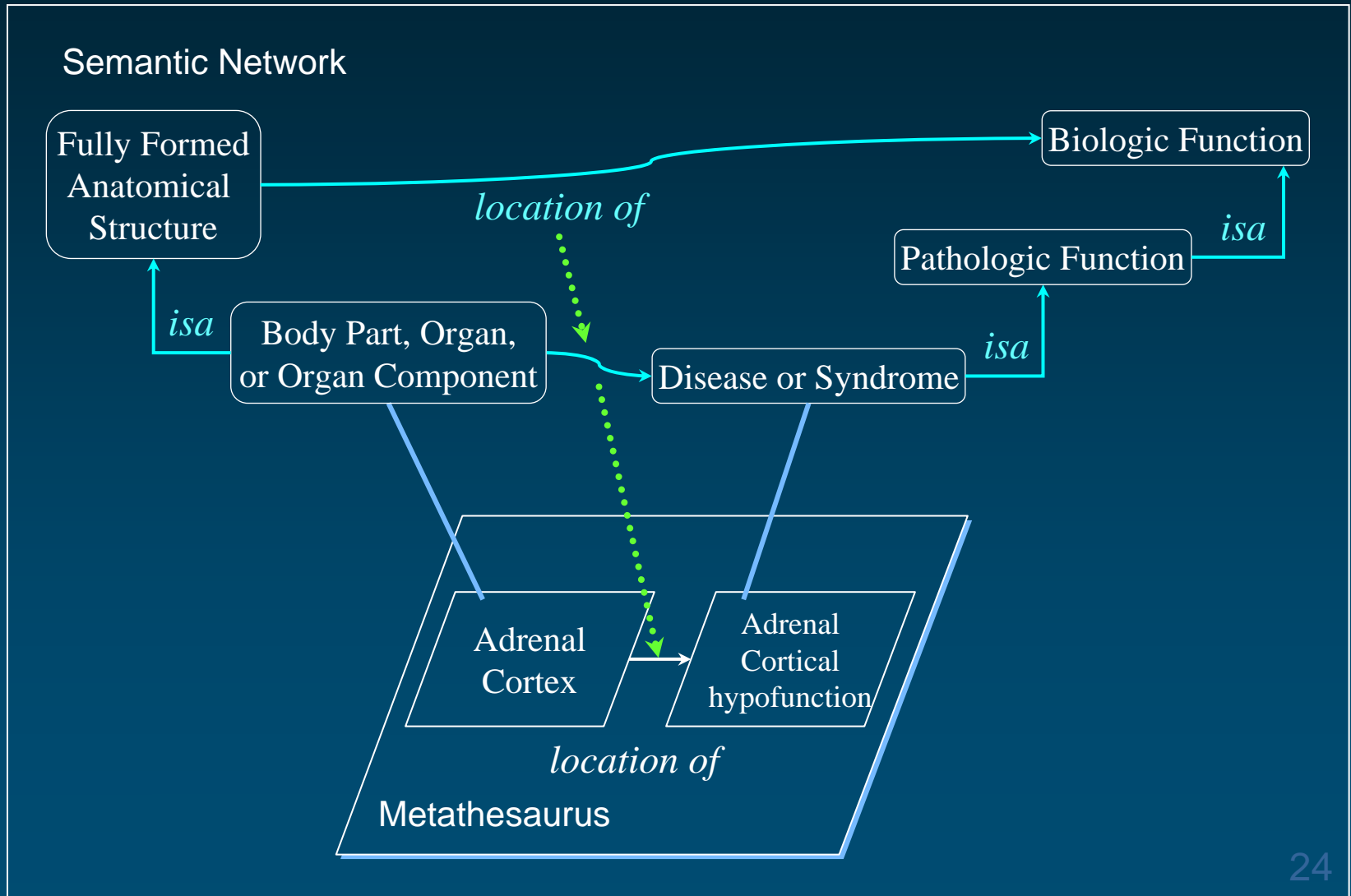
Semantic Network



Metathesaurus

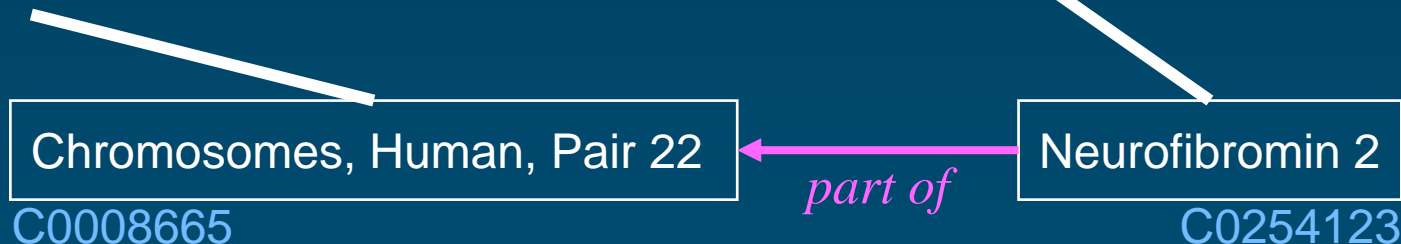
Concepts

Relationships can inherit semantics



SemRep Relation extraction

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.



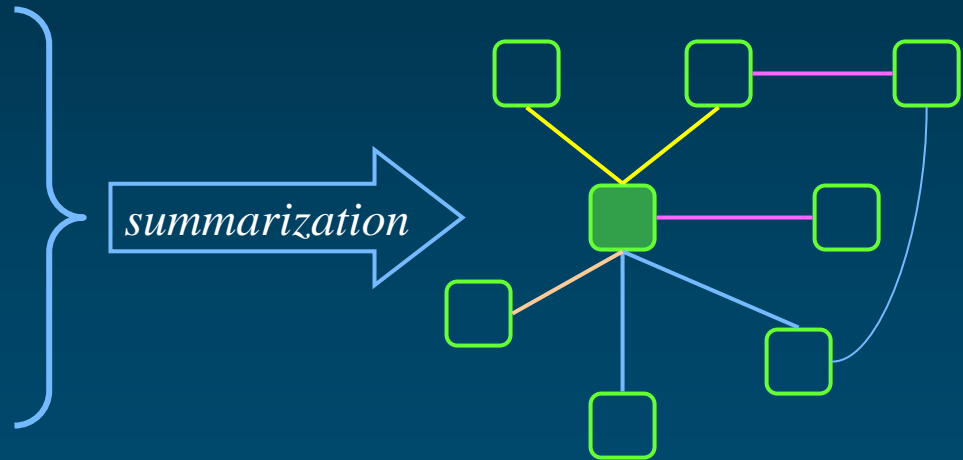
NLM resources for mining
biomedical text in action
Semantic Medline

Managing retrieval results



500 citations

Information retrieval

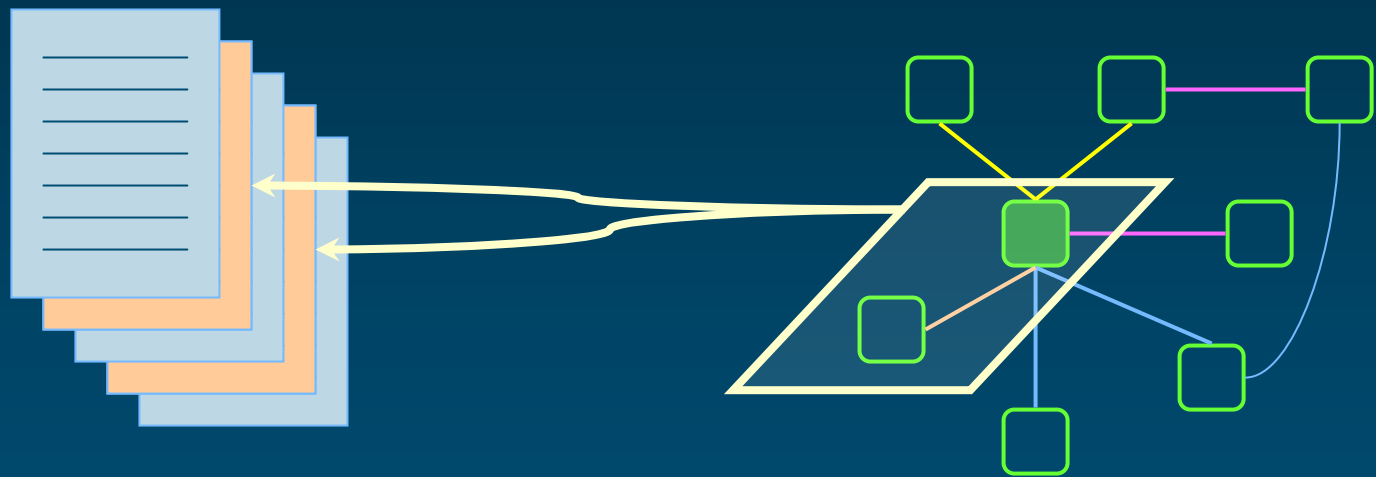


Network of relations

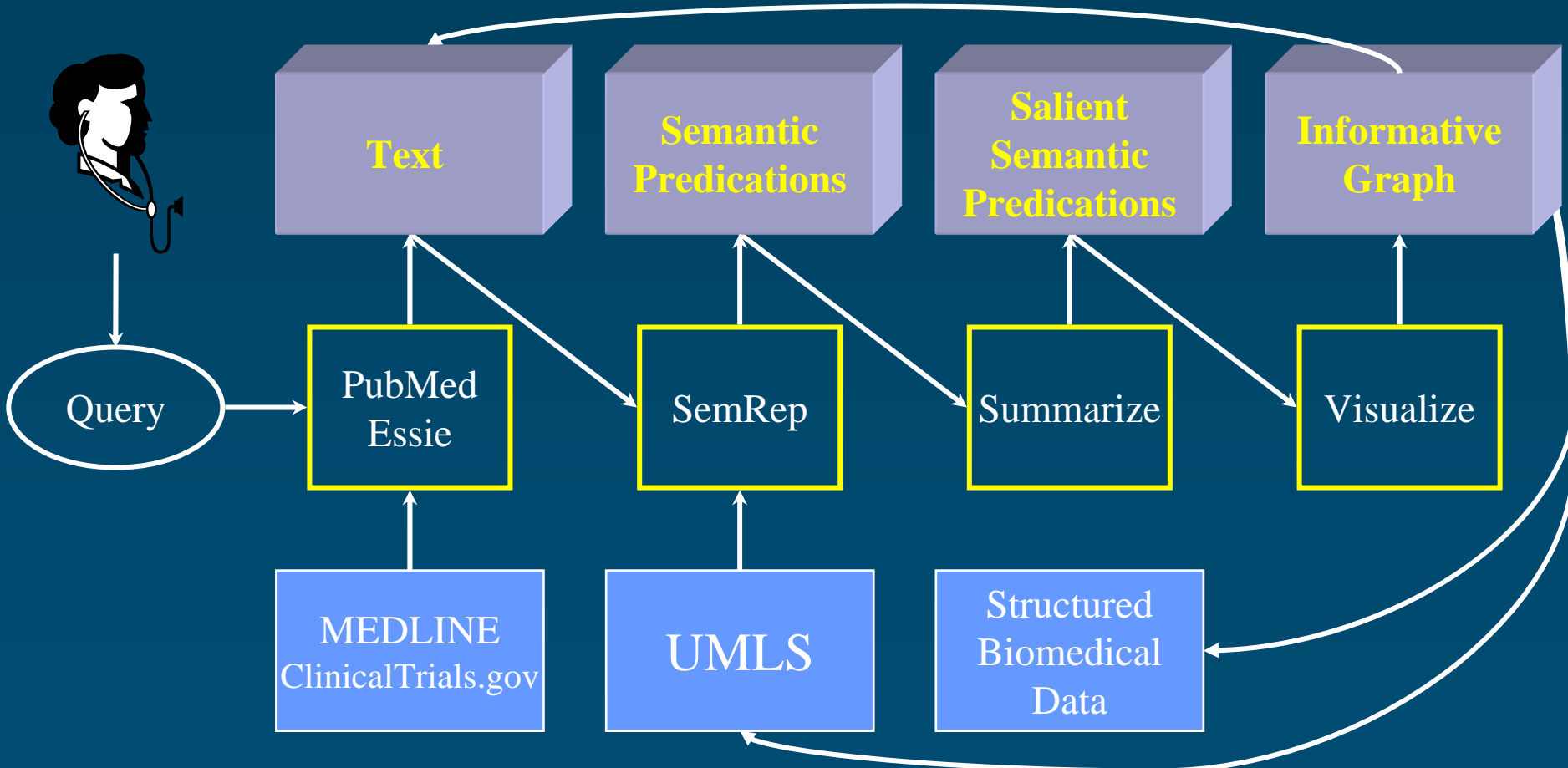
Semantic Medline



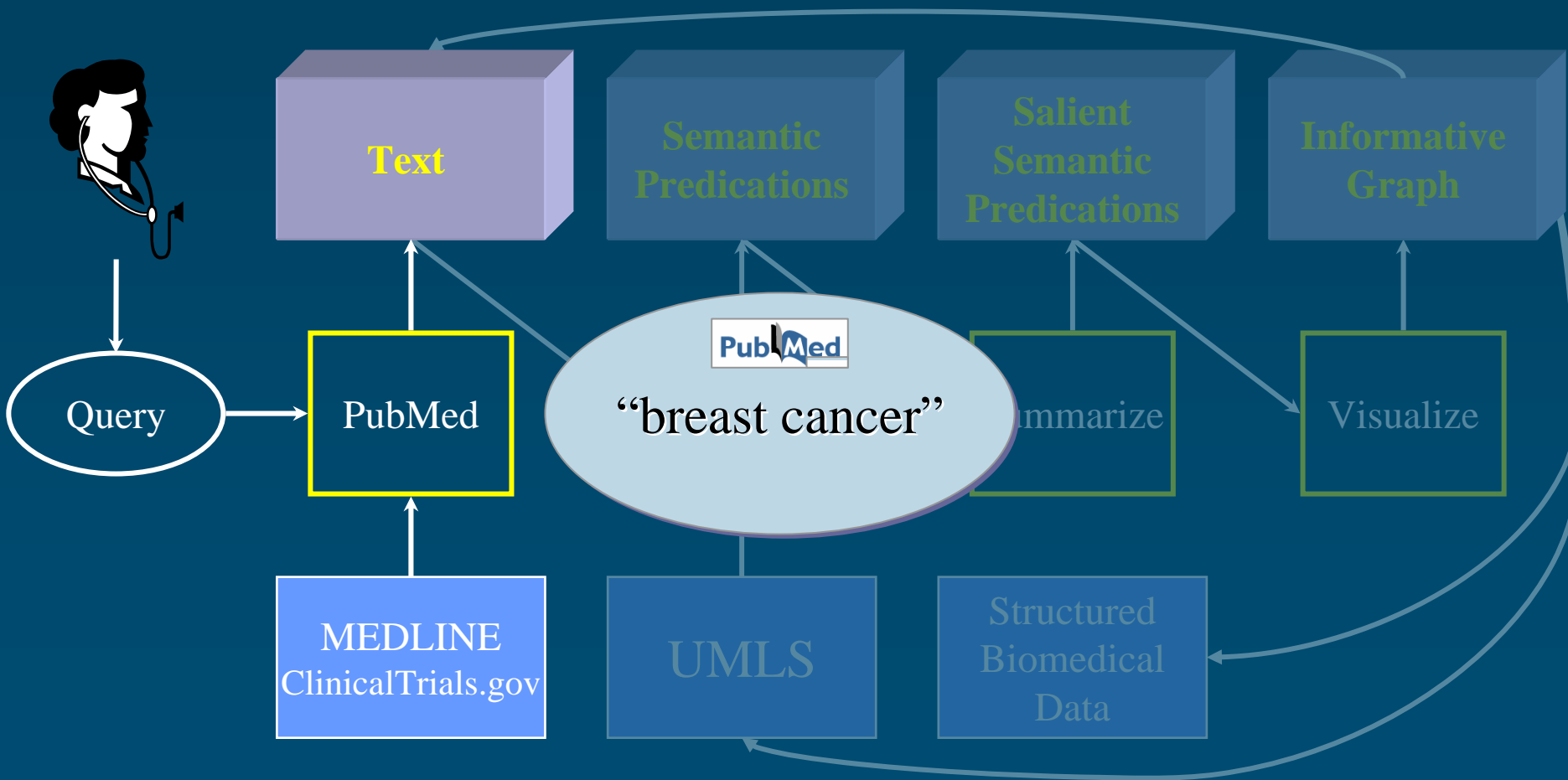
Managing retrieval results



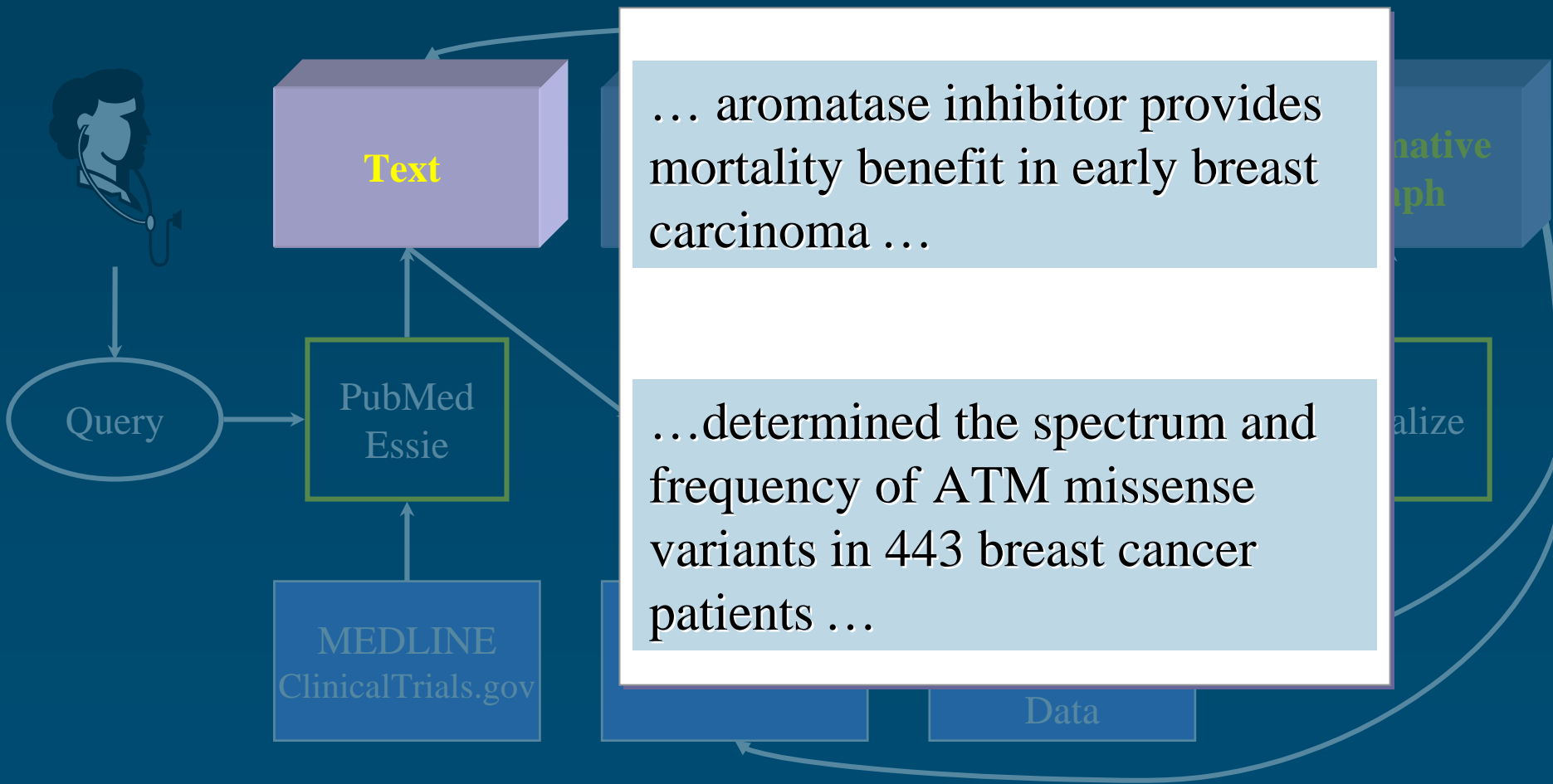
Semantic Medline Overview



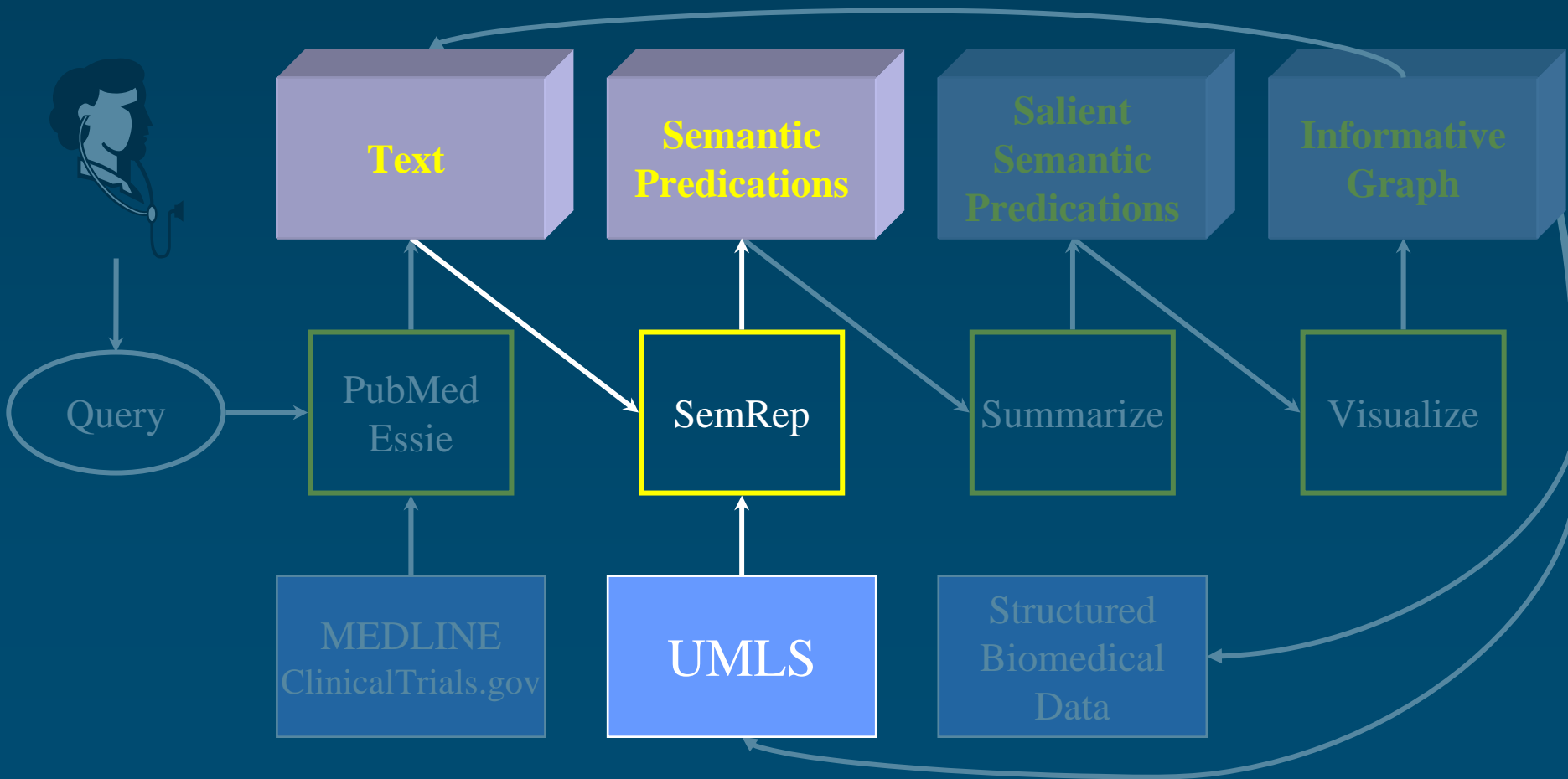
Document selection



MEDLINE citations



Semantic interpretation



Semantic interpretation

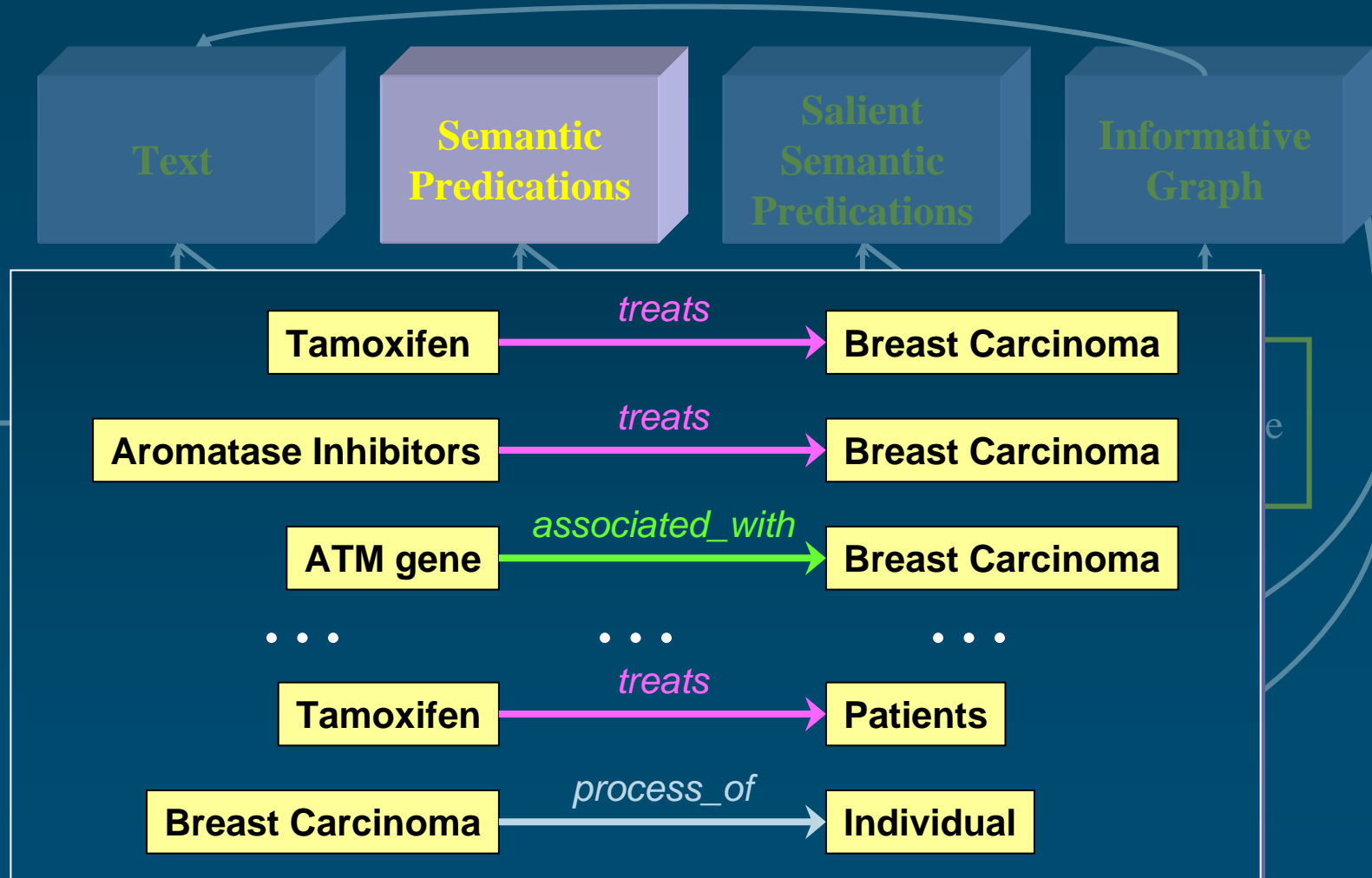
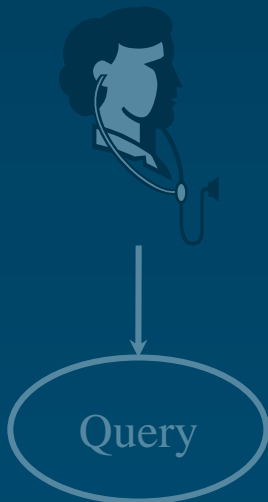
... aromatase inhibitor provides mortality benefit in early breast carcinoma ...



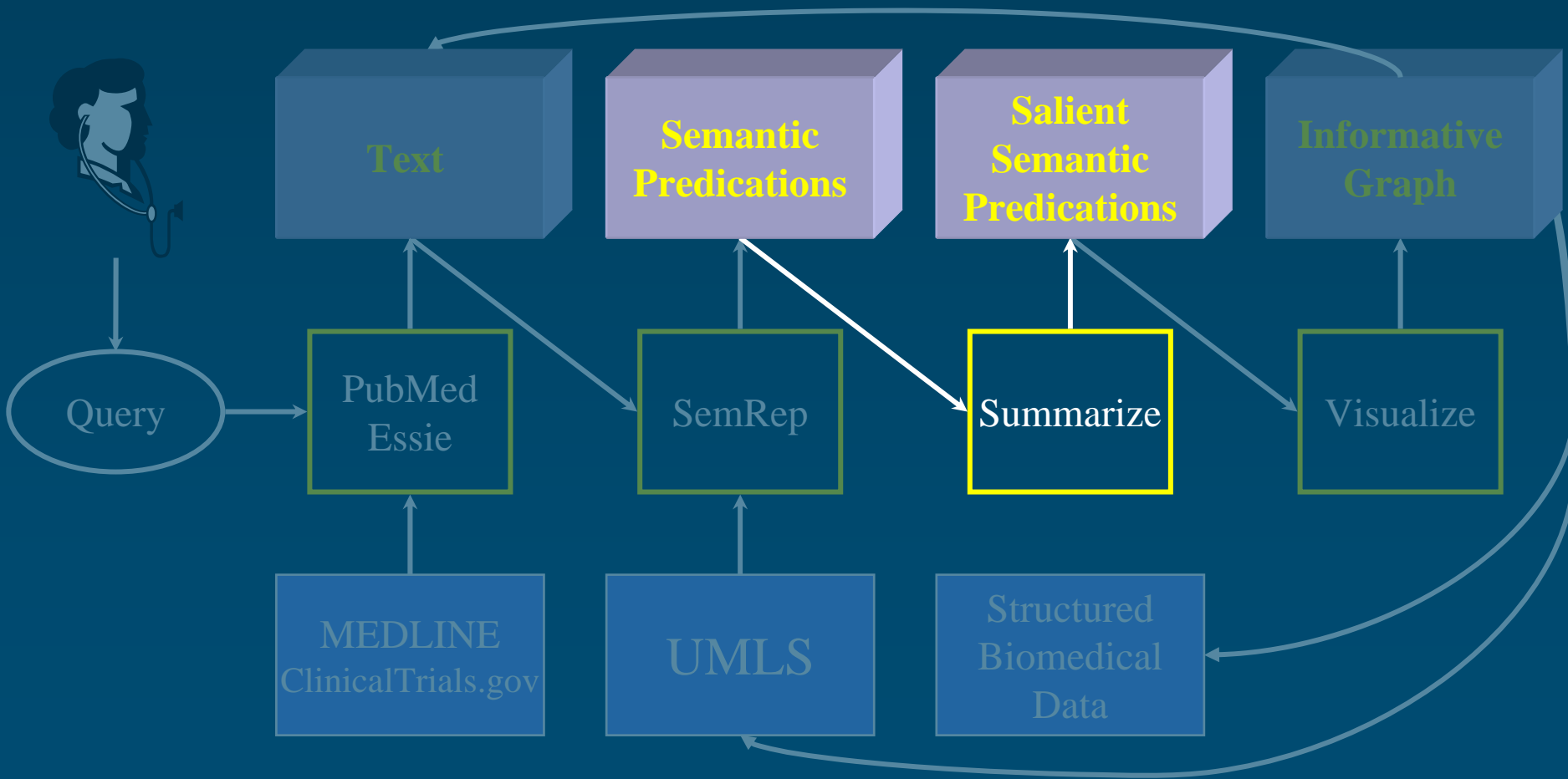
... determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients ...



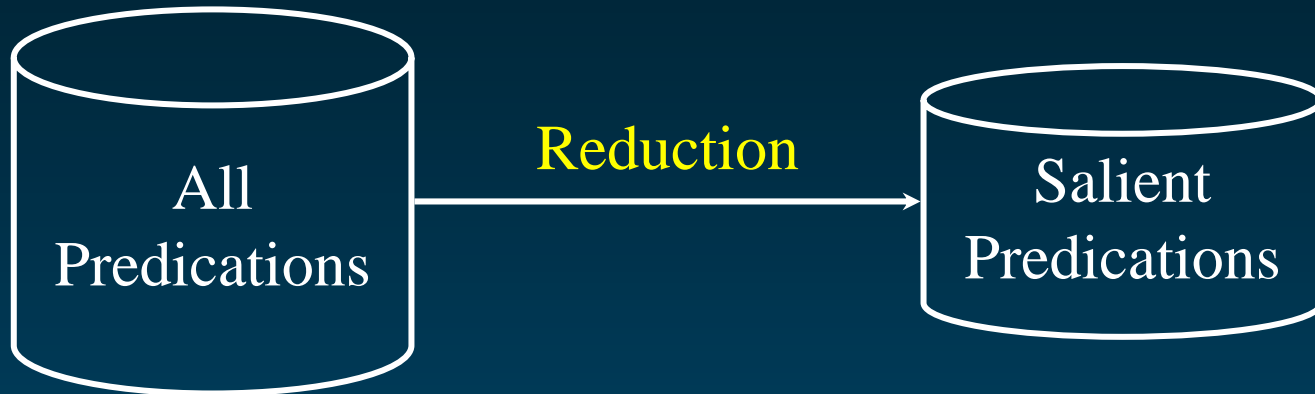
Semantic predications



Summarization

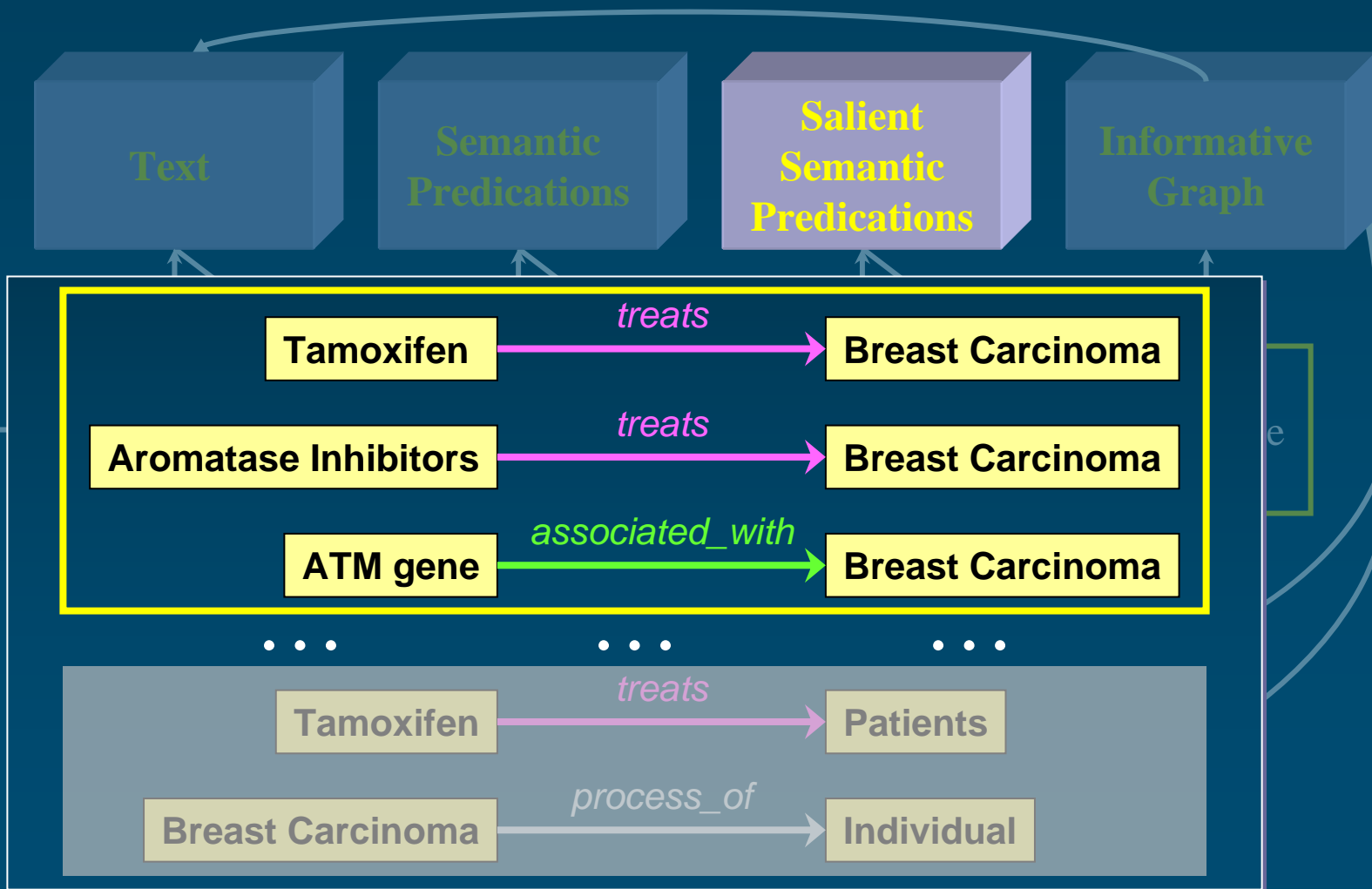
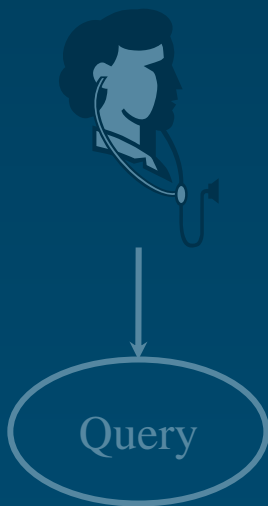


Abstraction summarization

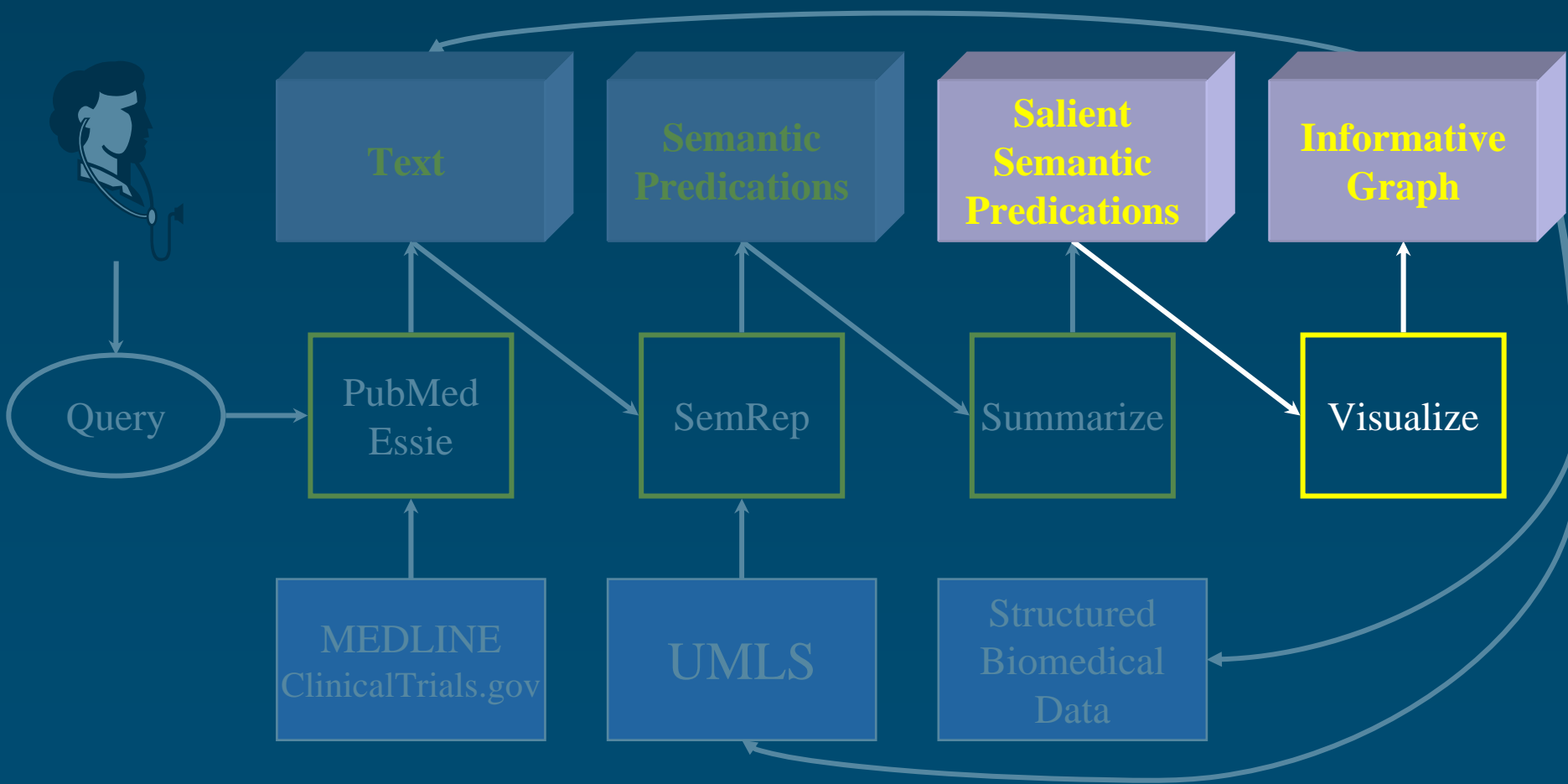


- ◆ Specify a topic
- ◆ Retain predications on the topic
- ◆ Eliminate uninformative predications
- ◆ Retain most frequent predications

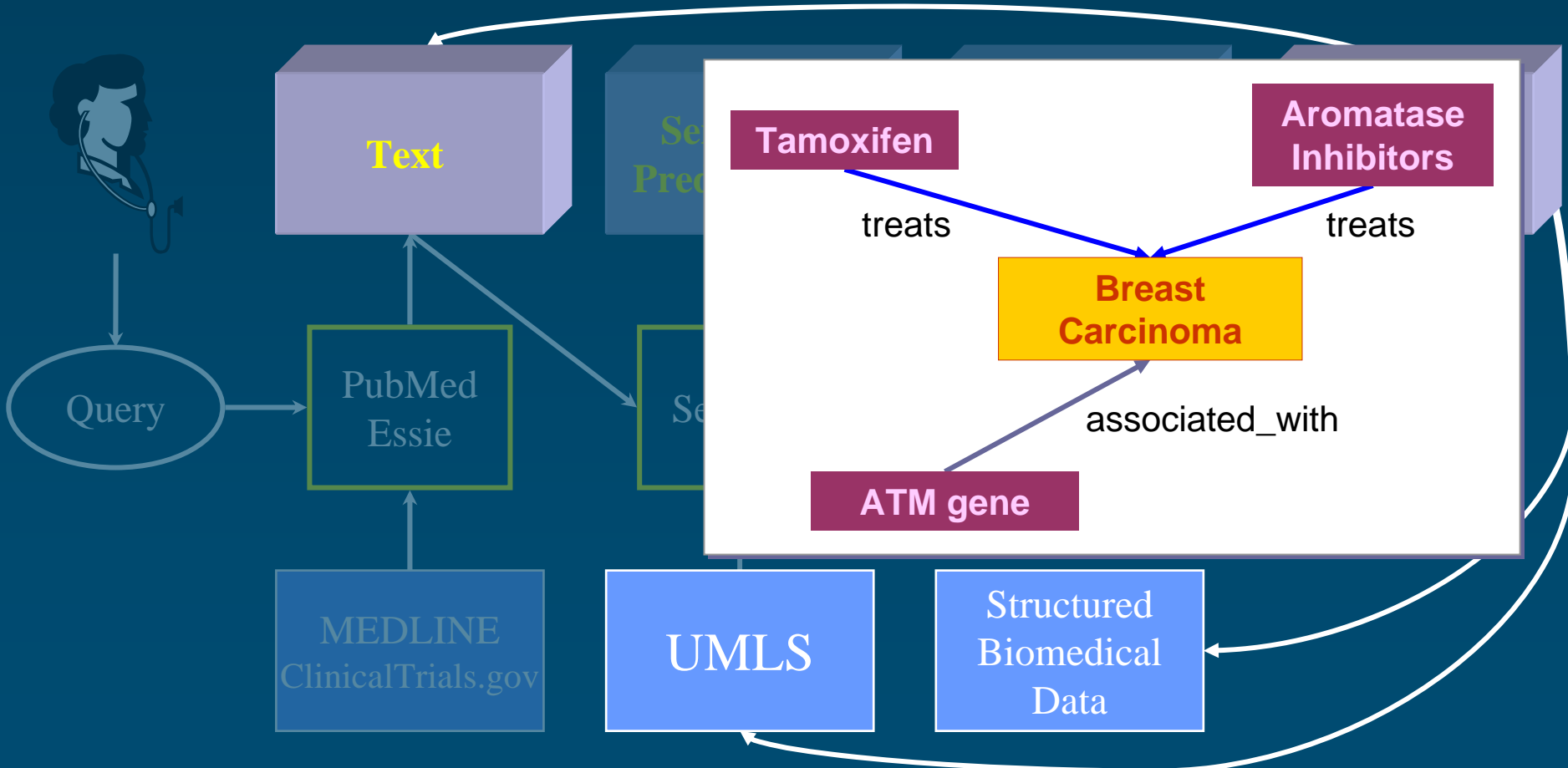
Salient semantic predications



Visualization



Informative graph



Semantic Medline Live

The screenshot displays the Semantic MEDLINE Prototype interface within a Mozilla Firefox browser window. The main area features a network graph with 'Breast Carcinoma' at the center, connected to various entities such as 'ERBB2 gene', 'Herceptin', 'Estrogens', 'Rastuzumab', 'Immunologic Adjuvants', 'BRCA1 Protein', 'BRCA2 Protein', 'Aromatase Inhibitors', 'BCL-2 Protein', 'Lignans', 'Aromatase Inhibitors', 'cyclooxygenase 2', 'Estradiol', 'Tamoxifen', 'Protein p53', 'Oncologist - Health', 'Eld', 'Anthracycline Antibiotics', 'Untranslated Regions', 'Carbon', 'Mannose Binding Lectin', 'ESR1 protein, human', 'DUSP22 gene', 'Immunologic Adjuvants', 'Allyl sulfide', 'Cells', 'NA Strand Break', 'nylimidazo(4,5-b)pyridine', 'Melatonin', 'Progestins', 'Mice, Nude', 'Mammary Neoplasms', 'Neoplasms', 'cohort', 'Breast cancer invasive NOS', 'BRCA1 Mutation', 'BRCA1 Mutation', 'Adiponectin', 'Hormones', 'Breast cancer regression', 'woman', 'cancer metastatic', 'Neoplasm Metastasis', '17q23', and 'Neoplasm progres'. A sidebar on the right lists semantic relations: ASSOCIATED_WITH, CAUSES, COEXISTS_WITH, DISRUPTS, INHIBITS, INTERACTS_WITH, ISA, LOCATION_OF, and PART_OF. Below the graph, a text box provides details for PMID- 17393301, dated 2007 Mar 28, with a title about ATM missense variants and breast cancer. The text includes a highlighted sentence: 'We have determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients diagnosed before age 50, including 247 patients who subsequently developed CBC.' The footer of the browser window shows the URL 'http://skr3.nlm.nih.gov/SemMedDemo/Summary.do' and the applet name 'Applet.com.touchgraph.semrepbrowser.SemRepVisualization started'.

PMID- 17393301
DP - 2007 Mar 28
T1 - The spectrum of ATM missense variants and their contribution to contralateral breast cancer.
AB - Heterozygous carriers of ATM mutations are at increased risk of breast cancer. In this case-control study, we evaluated the significance of germline ATM missense variants to the risk of contralateral breast cancer (CBC). We have determined the spectrum and frequency of ATM missense variants in 443 breast cancer patients diagnosed before age 50, including 247 patients who subsequently developed CBC. Twenty-one per cent of the women with

