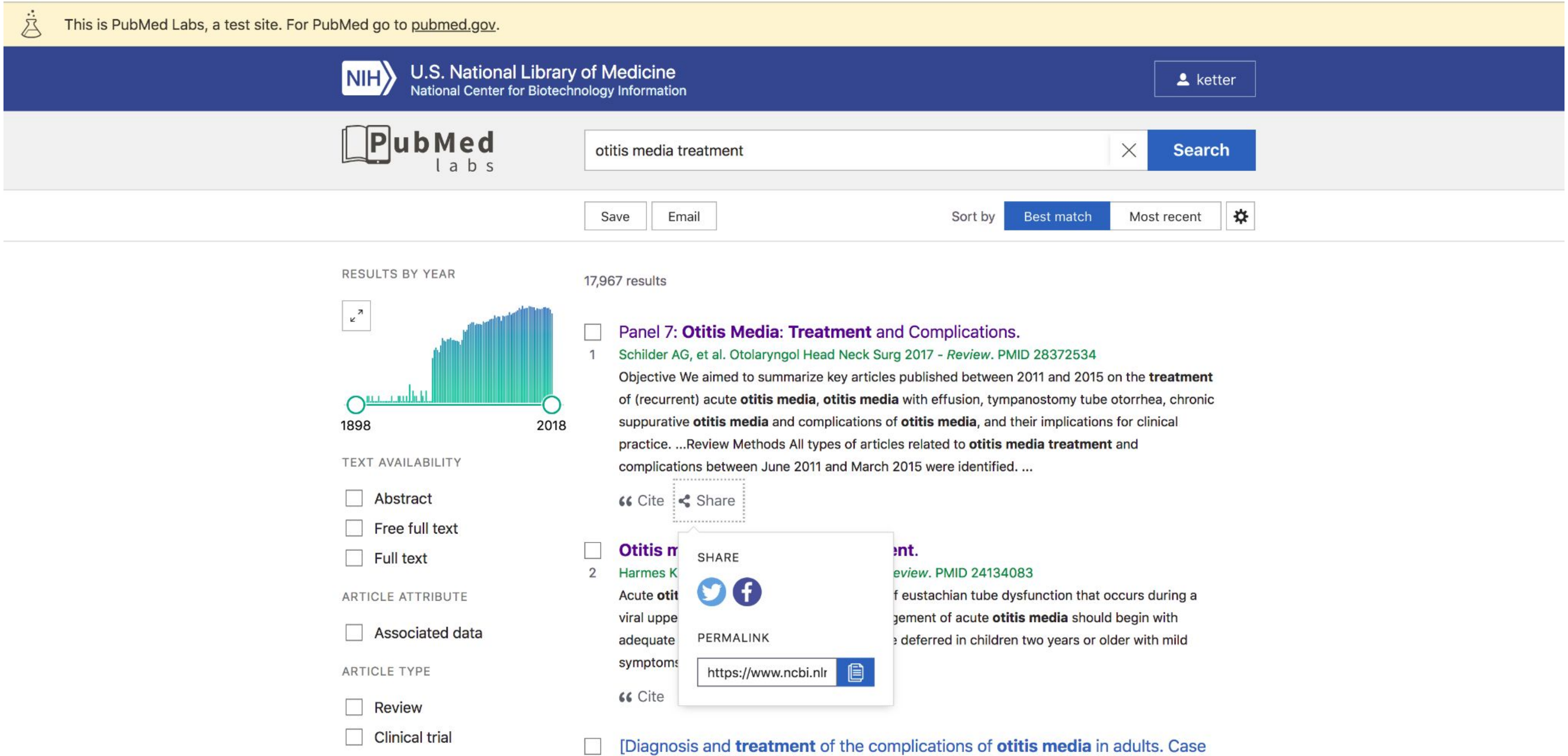
# NLP and Text Mining in Bioinformatics

Manu Madhavan

# Bioinformatics Literature

- Much literature generated quickly.
  - 11 million citations in MEDLINE.
  - 400,000 added yearly.
- Need methods to deal with data.
  - Query
  - Summarize
  - Organize
  - Understand
- For mankind to benefit from bioinformatics research, the sequence and structure of proteins and other molecules **must be linked to functional genomics and proteomics.**

# PubMed



This is PubMed Labs, a test site. For PubMed go to pubmed.gov.

U.S. National Library of Medicine
National Center for Biotechnology Information

ketter

PubMed labs

otitis media treatment

Search

Save | Email | Sort by | Best match | Most recent

RESULTS BY YEAR

1898 2018

TEXT AVAILABILITY

□ Abstract
□ Free full text
□ Full text

ARTICLE ATTRIBUTE

□ Associated data

ARTICLE TYPE

□ Review
□ Clinical trial

17,967 results

□ 1 Panel 7: **Otitis Media: Treatment** and Complications.
Schilder AG, et al. Otolaryngol Head Neck Surg 2017 - *Review*. PMID 28372534
Objective We aimed to summarize key articles published between 2011 and 2015 on the **treatment** of (recurrent) acute **otitis media**, **otitis media** with effusion, tympanostomy tube otorrhea, chronic suppurative **otitis media** and complications of **otitis media**, and their implications for clinical practice. ...Review Methods All types of articles related to **otitis media treatment** and complications between June 2011 and March 2015 were identified. ...
Cite Share

SHARE

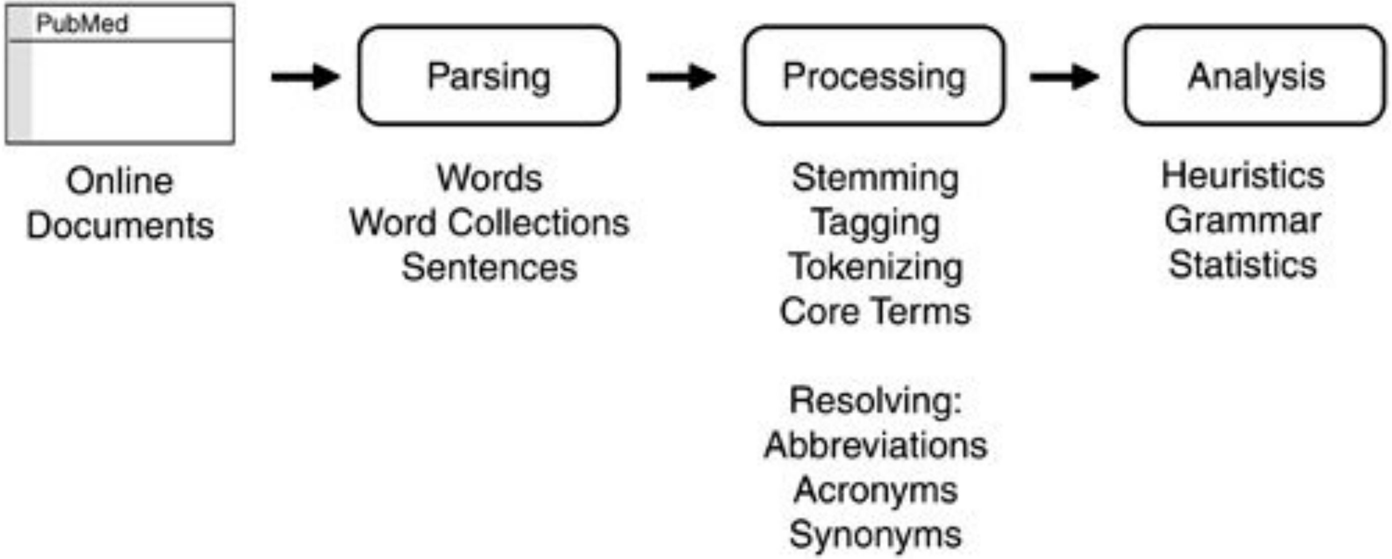
PERMALINK
https://www.ncbi.nlr

□ 2 **Otitis m** **ent.**
Harmes K *eview*. PMID 24134083
Acute **otit** f eustachian tube dysfunction that occurs during a viral uppe gement of acute **otitis media** should begin with adequate deferred in children two years or older with mild symptoms
Cite

□ [Diagnosis and **treatment** of the complications of **otitis media** in adults. Case

- The primary store of functional data that links clinical medicine, pharmacology, sequence data, and structure data is in the form of biomedicine documents in online bibliographic databases such as PubMed
- Mining these databases is expected to reveal the relationships between structure and function at the molecular level and their relationship to pharmacology and clinical medicine.
- **Text mining**—automatically extracting this data from documents, which is published in the form of unstructured free text, often in several languages

# The process

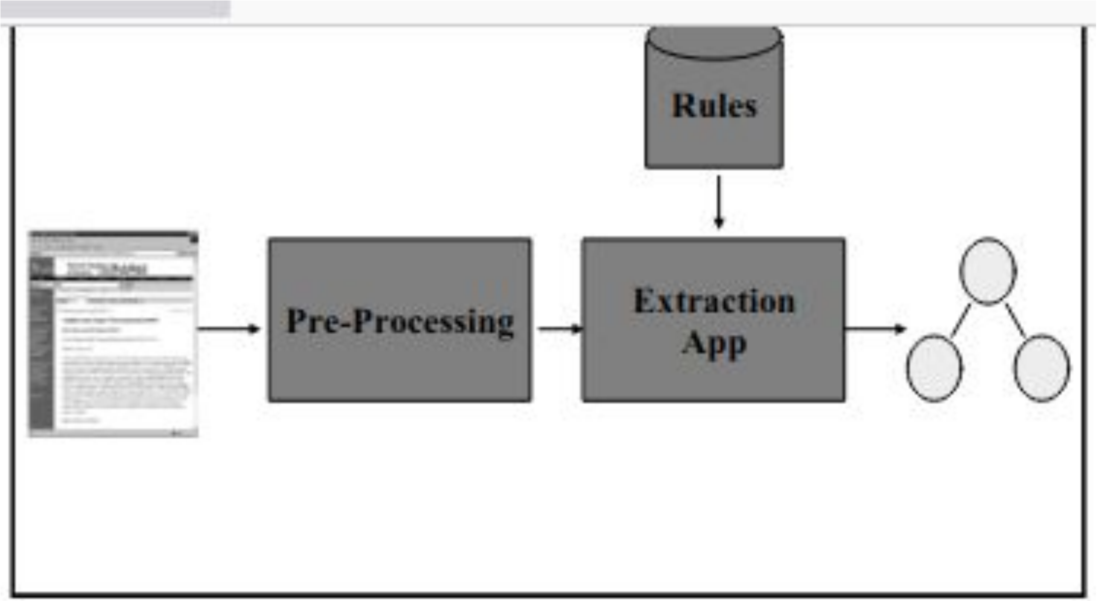


IF <protein name>
AND <experimental method name> are in the same sentence
THEN the <experimental method name> refers to the <protein name>

# NLP Research areas

- Information Retrieval
- Information Extraction
- Q&A
- Named Entity Recognition
- Entity Relation Extraction
- Curated Databases
- Text summarization
- Ontology/Knowledge graph

# Information Extraction



# Gene Ontology and Mesh Terms

- Gene Ontology:The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects: Molecular Function. Molecular-level activities performed by gene products.

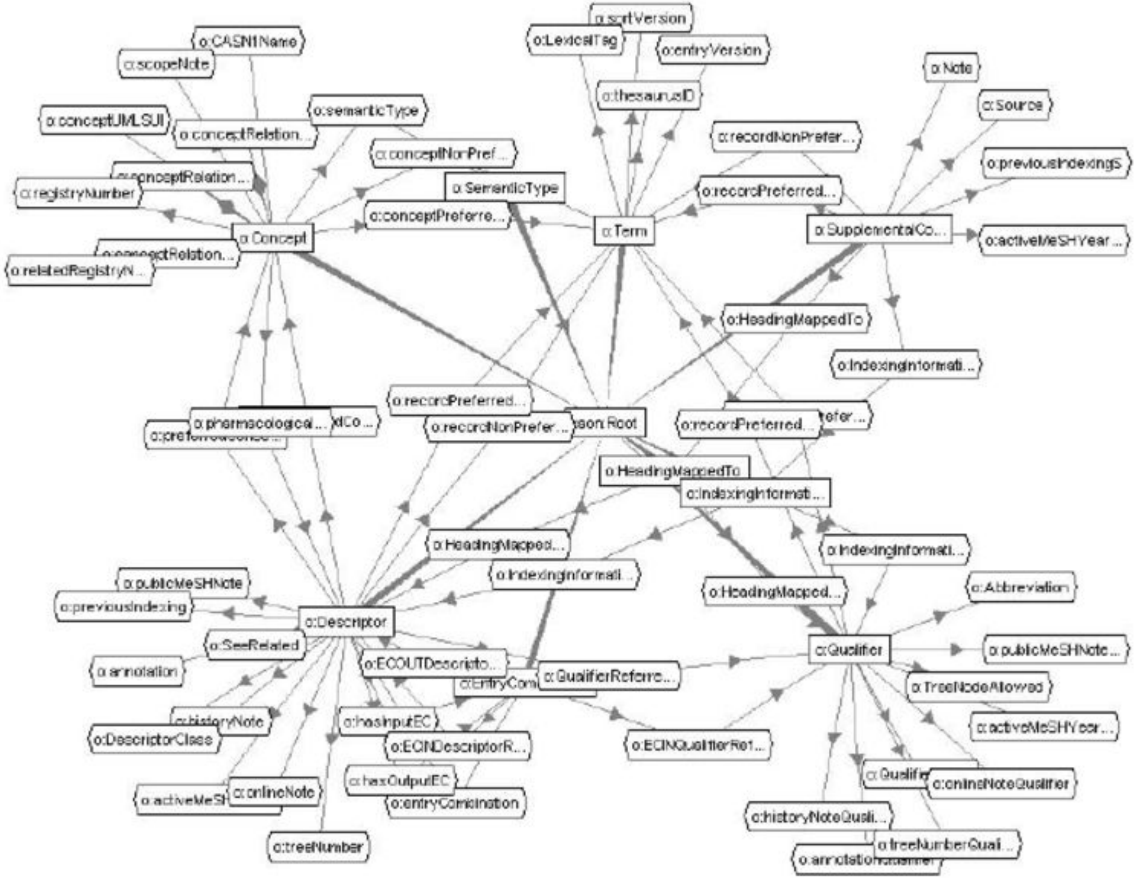
- Mesh Terms: The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine.

# Gene Ontology

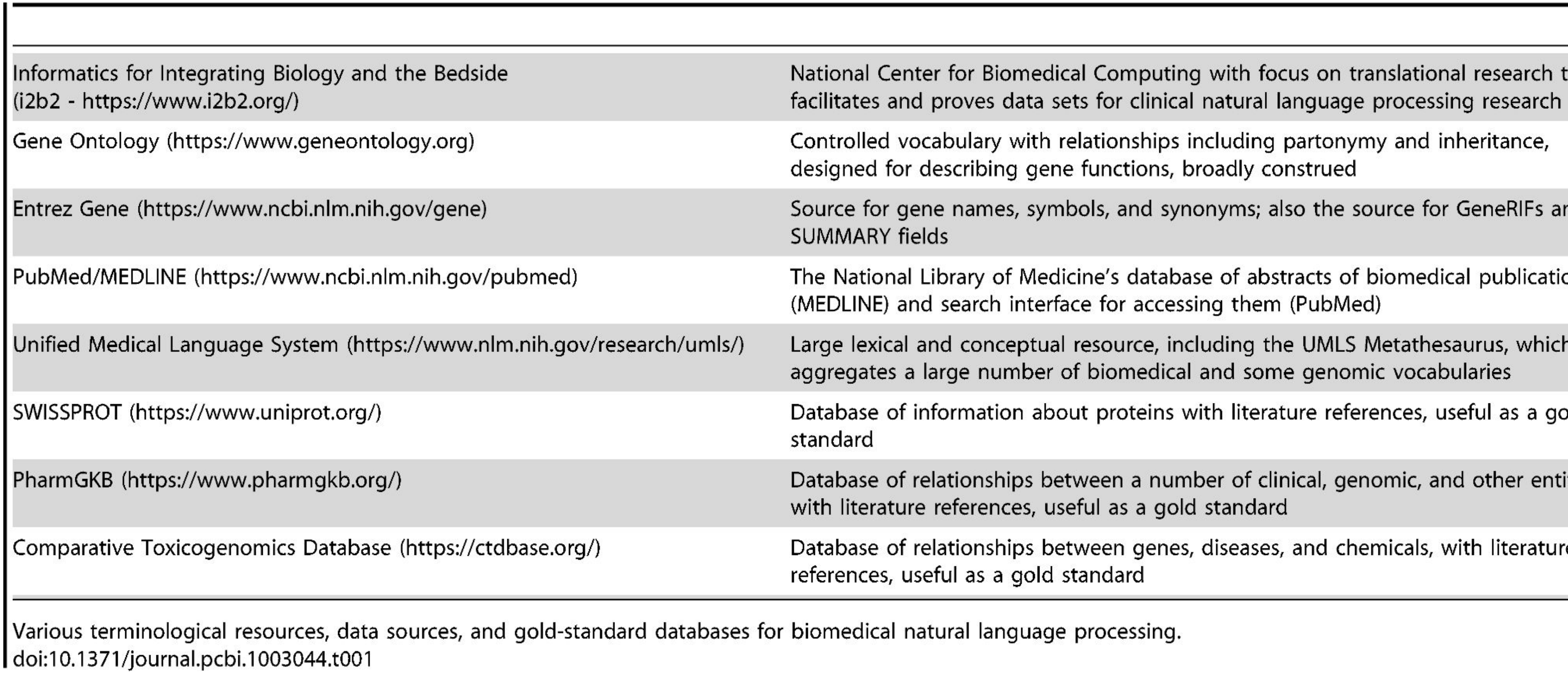
http://geneontology.org/docs/ontology-documentation/

Gene Enrichment Analysis: http://geneontology.org/

# Mesh


o:CASN1Name
o:scopeNote
o:conceptUMLSUI
o:conceptRelation...
o:semanticType
o:conceptNonPref...
o:conceptRelation...
o:registryNumber
o:Concept
o:conceptPreferre...
o:conceptRelation...
o:relatedRegistryN...
o:SemanticType
o:sortVersion
o:LexicalTag
o:entryVersion
o:thesaurusID
o:Term
o:recordNonPrefer...
o:recordPreferred...
o:Note
o:Source
o:previousIndexingS
o:SupplementalCo...
o:activeMeSHYear...
o:HeadingMappedTo
o:IndexingInformati...
o:recordPreferred...
o:pharmacological...
o:preferredConce...
o:recordNonPrefer...
son Root
o:recordPreferred...
o:HeadingMappedTo
o:IndexingInformati...
o:HeadingMapped...
o:IndexingInformati...
o:IndexingInformati...
o:HeadingMapped...
o:Abbreviation
o:publicMeSHNote
o:previousIndexing
o:Descriptor
o:SeeRelated
o:annotation
o:ECOUTDescripto...
o:EntryCom
o:QualifierReferre...
o:Qualifier
o:publicMeSHNote...
o:TreeNodeAllowed
o:activeMeSHYear...
o:historyNote
o:DescriptorClass
o:hasInputEC
o:ECINDescriptorR...
o:ECINQualifierRef...
o:activeMeSH
o:onlineNote
o:hasOutputEC
o:entryCombination
o:Qualifier
o:onlineNoteQualifier
o:historyNoteQuali...
o:treeNumber
o:treeNumberQuali...
o:annotationQualifier

# Resources

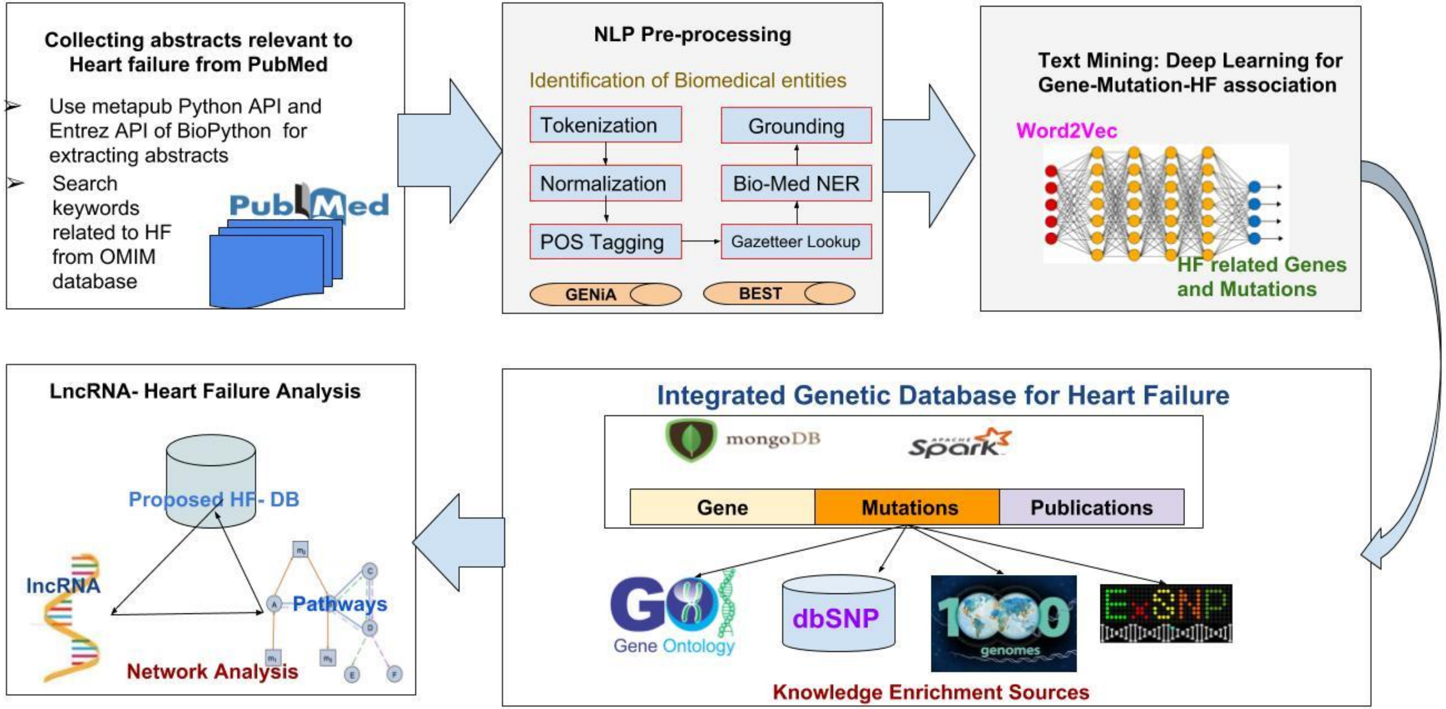


| | |
|---|---|
| Informatics for Integrating Biology and the Bedside (i2b2 - https://www.i2b2.org/) | National Center for Biomedical Computing with focus on translational research t facilitates and proves data sets for clinical natural language processing research |
| Gene Ontology (https://www.geneontology.org) | Controlled vocabulary with relationships including partonymy and inheritance, designed for describing gene functions, broadly construed |
| Entrez Gene (https://www.ncbi.nlm.nih.gov/gene) | Source for gene names, symbols, and synonyms; also the source for GeneRIFs ar SUMMARY fields |
| PubMed/MEDLINE (https://www.ncbi.nlm.nih.gov/pubmed) | The National Library of Medicine's database of abstracts of biomedical publicatio (MEDLINE) and search interface for accessing them (PubMed) |
| Unified Medical Language System (https://www.nlm.nih.gov/research/umls/) | Large lexical and conceptual resource, including the UMLS Metathesaurus, which aggregates a large number of biomedical and some genomic vocabularies |
| SWISSPROT (https://www.uniprot.org/) | Database of information about proteins with literature references, useful as a go standard |
| PharmGKB (https://www.pharmgkb.org/) | Database of relationships between a number of clinical, genomic, and other enti with literature references, useful as a gold standard |
| Comparative Toxicogenomics Database (https://ctdbase.org/) | Database of relationships between genes, diseases, and chemicals, with literature references, useful as a gold standard |

Various terminological resources, data sources, and gold-standard databases for biomedical natural language processing.
doi:10.1371/journal.pcbi.1003044.t001

# Case Studies

# Text Minded database of Mutations in HF



**Collecting abstracts relevant to Heart failure from PubMed**

- Use metapub Python API and Entrez API of BioPython for extracting abstracts
- Search keywords related to HF from OMIM database

**NLP Pre-processing**

Identification of Biomedical entities

**Text Mining: Deep Learning for Gene-Mutation-HF association**

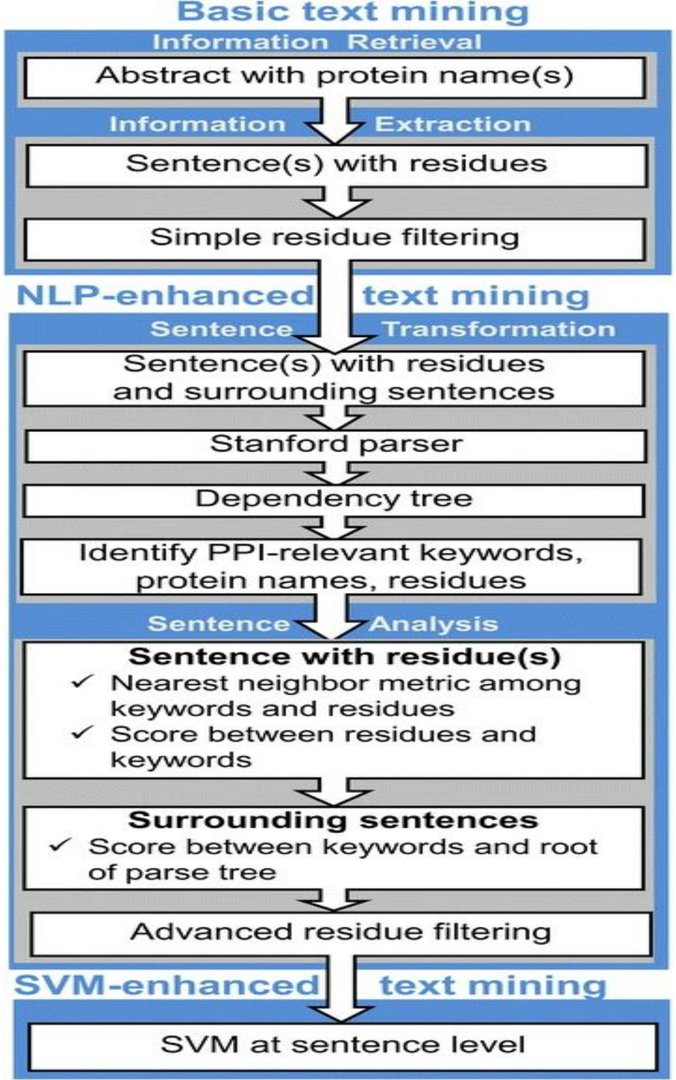
**Integrated Genetic Database for Heart Failure**

Knowledge Enrichment Sources

**LncRNA- Heart Failure Analysis**

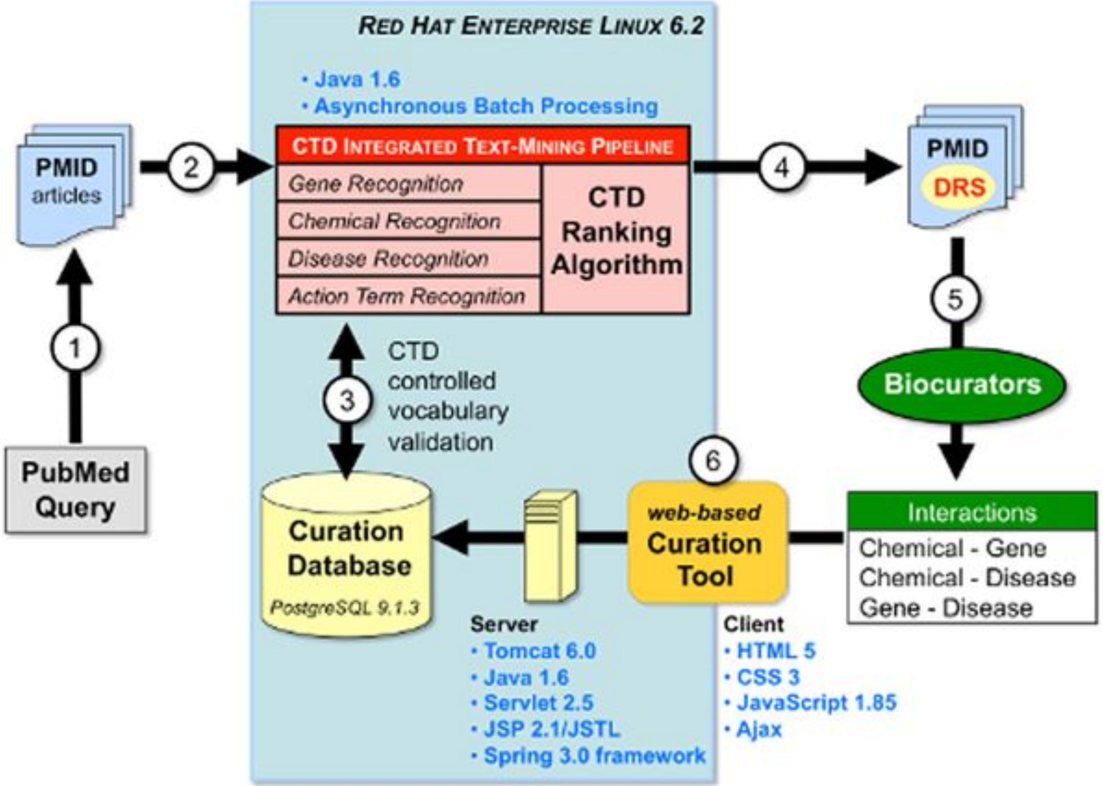
# Text mining for structural modeling of protein complexes

- TM tool, which utilizes natural language processing (NLP) for analyzing the context of the residue occurrence
- TM procedure for extracting protein-protein binding site residues from the PubMed abstracts was significantly advanced by the deep parsing (NLP techniques for contextual analysis) in purging of the initial pool of the extracted residues.

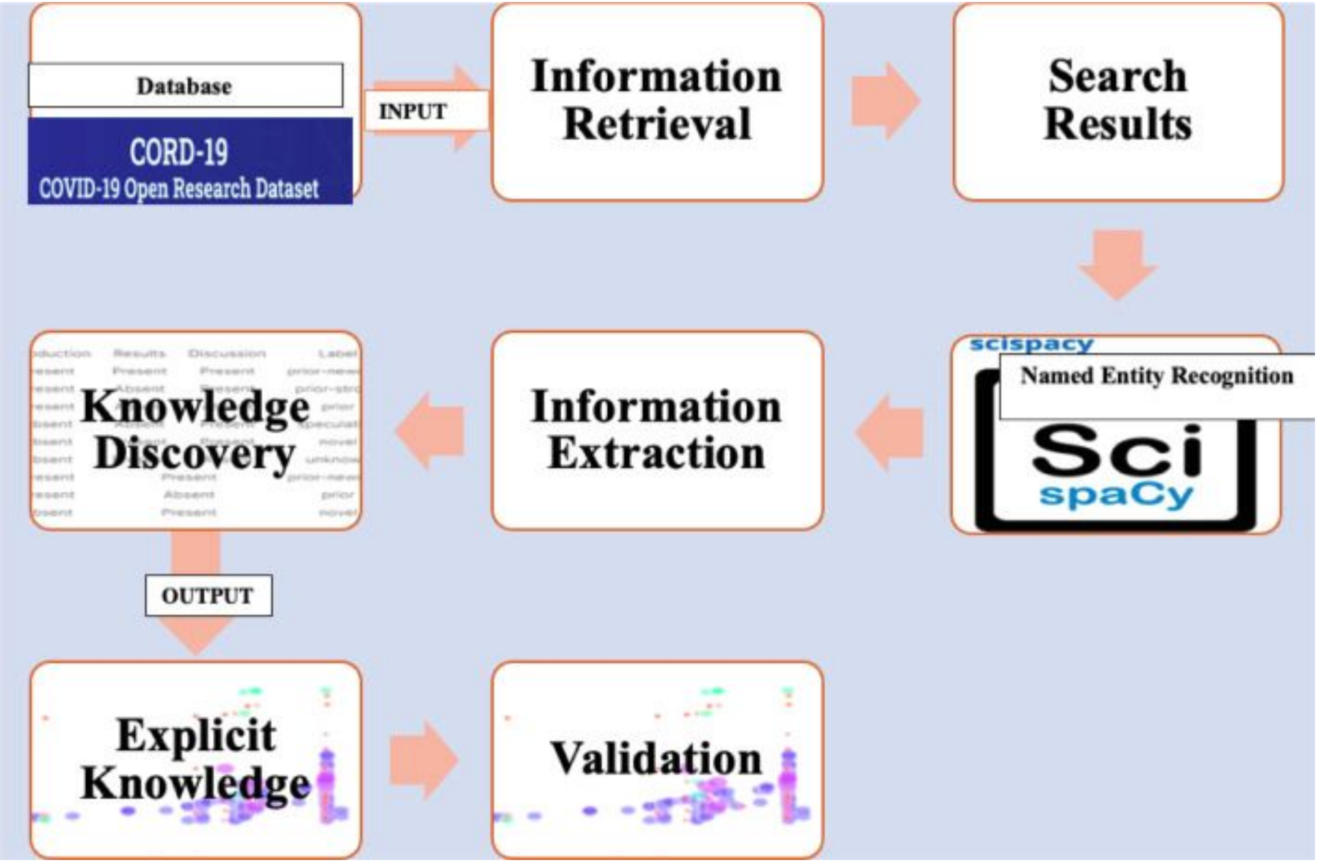


Basic text mining
Information Retrieval
Abstract with protein name(s)
Information Extraction
Sentence(s) with residues
Simple residue filtering
NLP-enhanced text mining
Sentence Transformation
Sentence(s) with residues and surrounding sentences
Stanford parser
Dependency tree
Identify PPI-relevant keywords, protein names, residues
Sentence Analysis
Sentence with residue(s)
✓ Nearest neighbor metric among keywords and residues
✓ Score between residues and keywords
Surrounding sentences
✓ Score between keywords and root of parse tree
Advanced residue filtering
SVM-enhanced text mining
SVM at sentence level

# Biocuration



# Covid Dataset



# Pubtator Demo

https://www.ncbi.nlm.nih.gov/research/pubtator/

# NCBI text mining tools

https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/

# More Examples

1. Pathway extraction and reasoning
2. Gene prioritization and gene function prediction
3. Pharmacology
4. Drug repurposing