Biomedical Text Mining

Fabio Rinaldi
Dalle Molle Institute for Artificial Intelligence (IDSIA) Lugano, Switzerland, USI/SUPSI
nlp.idsia.ch

November 16, 2021

Outline

Topic

What is text mining?

The ability to process text written in some human language (unstructured data), typically a large set of documents, interpret the meaning, and automatically extract concepts, as well as the relationships among those concepts, to directly answer questions of interest.

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

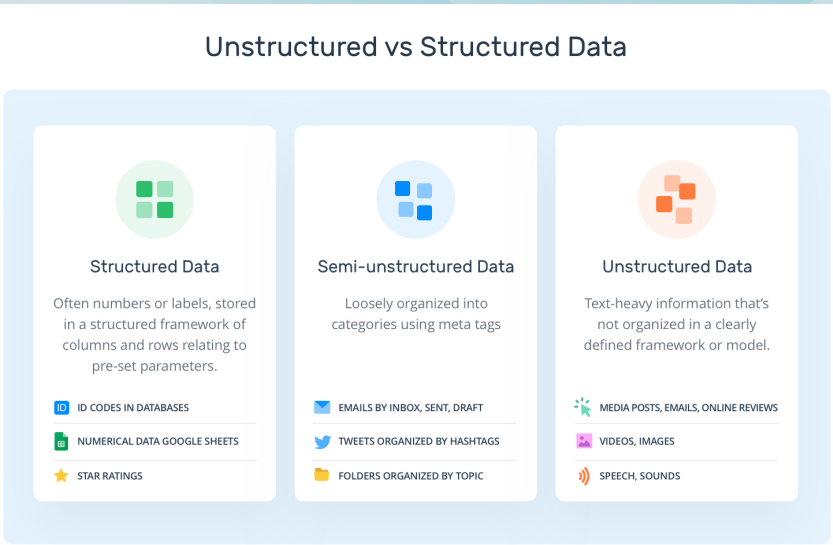
Structured data

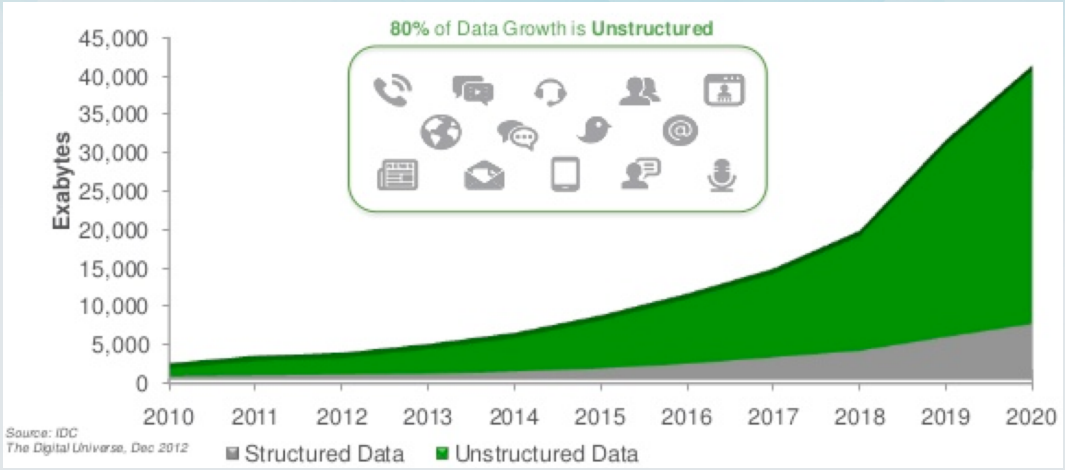
ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

What is unstructured data?

- Information that does not have a pre-defined data model (structure)
- Cannot be organized in a relational DB
- Typically text-heavy and usually contains free text in form of natural language
  - types: text, picture, videos (multimedia)
  - scientific literature, electronic health records, web pages, tweets and other social media text.
- These data sources/files may have internal structure but it is challenging to store information in row-column basis
- Estimates 80-90% of usable information in unstructured form
- Considerable growth and accumulation of unstructured data
- Often used as synonym of Big data (which is however both structured and unstructured)

Unstructured vs Structured Data

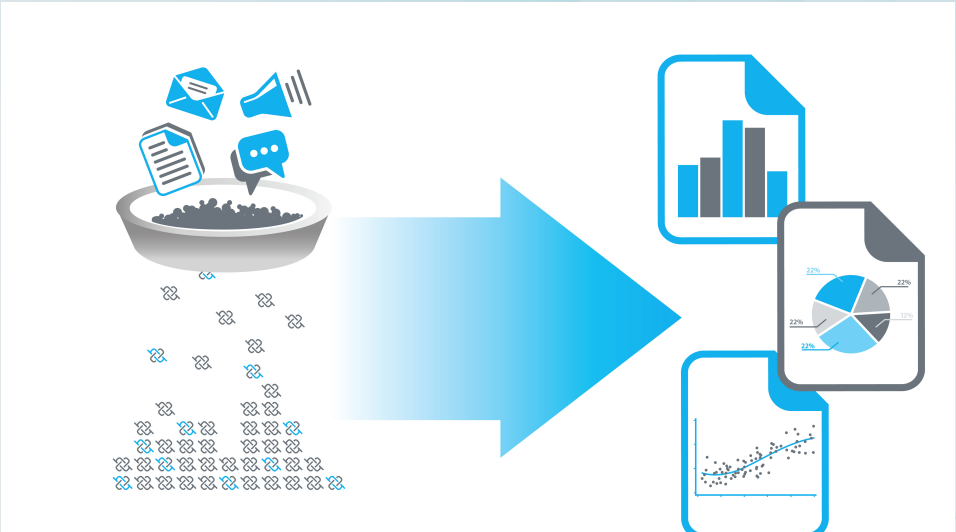



80% of Data Growth is Unstructured
Exabytes
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Source: IDC
The Digital Universe, Dec 2012
Structured Data
Unstructured Data

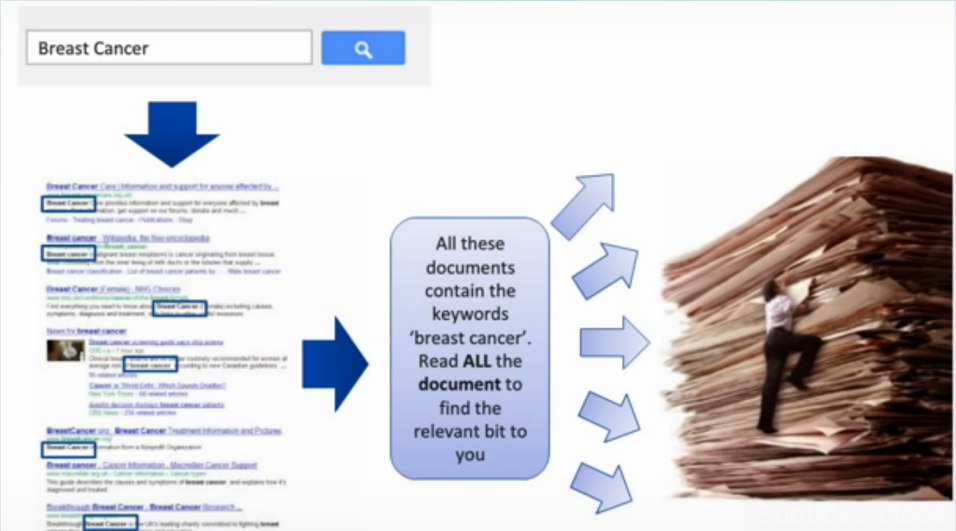
What is text mining?

- Data Mining needs structured data, usually in numerical form
- Text mining deals with unstructured data in form of natural language data (free text)
  - Goal: Retrieve what is hidden in text, present distilled knowledge in concise form
  - Process of discovery and extraction of knowledge from unstructured data
  - Extraction of non-trivial information or new information from natural language data collection
  - Applications integrating natural language processing components (such as document retrieval, classification and recognition of biological entities)

Text Mining process



Conventional document search

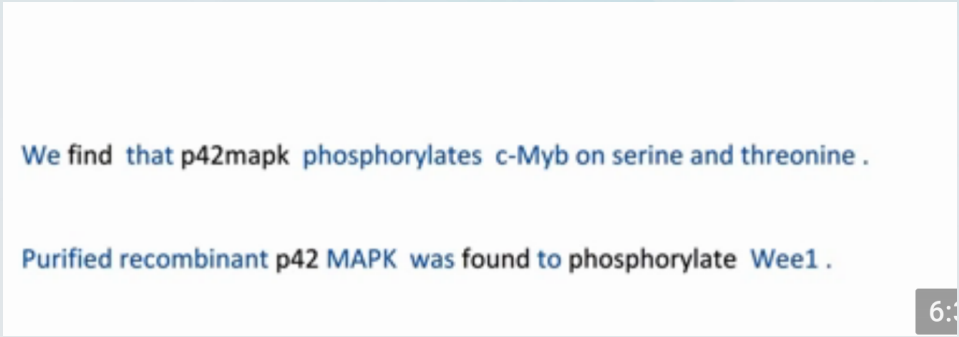


Pros and cons

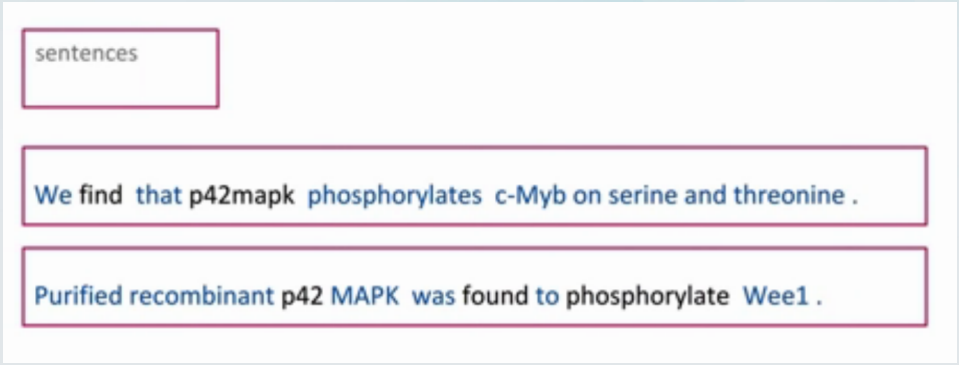
- Pros
  - speculative browsing for information
  - finding general information, e.g. address of the post office
- Cons
  - large result sets, normally not enough time to read them all
  - noisy and irrelevant bits
  - need to consider variants of terms
    - e.g. *cancer*, *tumor*, *neoplasm*

What is Text Mining

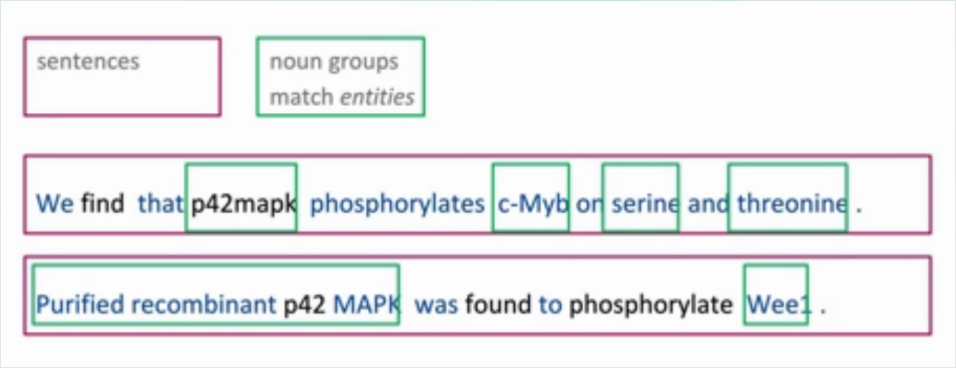
Example


We find that p42mapk phosphorylates c-Myb on serine and threonine .
Purified recombinant p42 MAPK was found to phosphorylate Wee1 .
6:

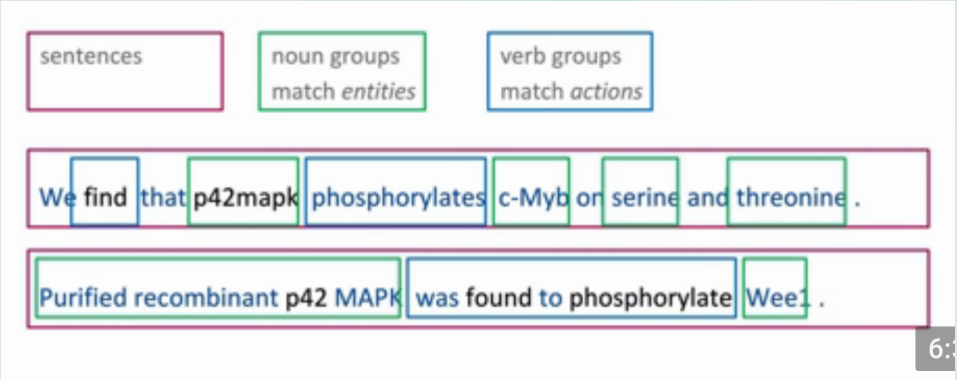
Example


sentences
We find that p42mapk phosphorylates c-Myb on serine and threonine .
Purified recombinant p42 MAPK was found to phosphorylate Wee1 .

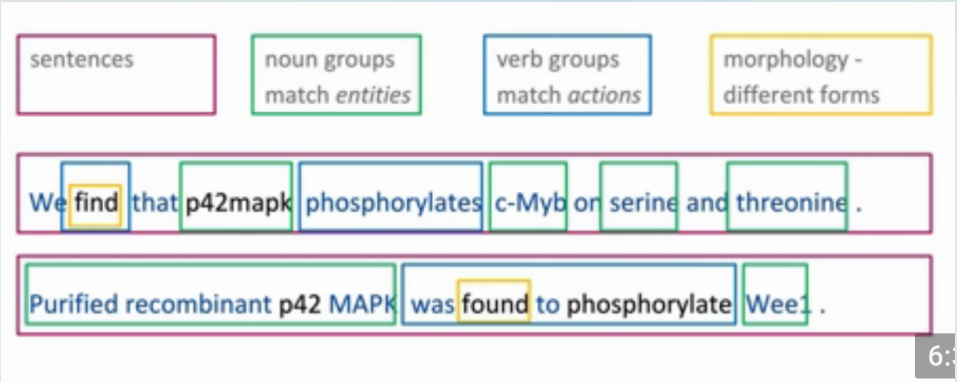
Example


sentences
noun groups
match *entities*
We find that p42mapk phosphorylates c-Myb on serine and threonine .
Purified recombinant p42 MAPK was found to phosphorylate Wee1 .

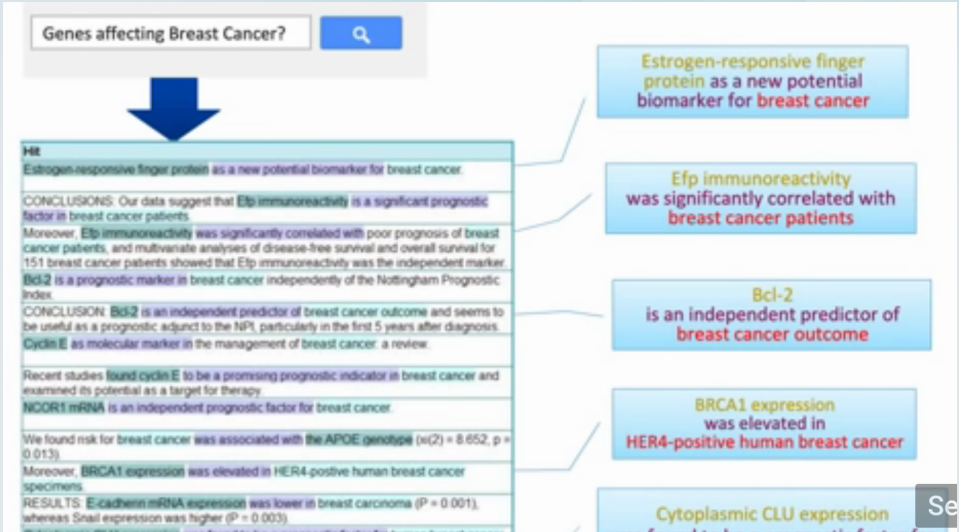
Example


sentences
noun groups
match *entities*
verb groups
match *actions*
We find that p42mapk phosphorylates c-Myb on serine and threonine .
Purified recombinant p42 MAPK was found to phosphorylate Wee1 .

Example


sentences
noun groups match *entities*
verb groups match *actions*
morphology - different forms
We find that p42mapk phosphorylates c-Myb on serine and threonine .
Purified recombinant p42 MAPK was found to phosphorylate Wee1 .

Example


Genes affecting Breast Cancer?
Hit
Estrogen-responsive finger protein as a new potential biomarker for breast cancer
CONCLUSIONS: Our data suggest that Efp immunoreactivity is a significant prognostic factor in breast cancer patients.
Moreover, Efp immunoreactivity was significantly correlated with poor prognosis of breast cancer patients, and multivariate analyses of disease-free survival and overall survival for 151 breast cancer patients showed that Efp immunoreactivity was the independent marker.
Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index.
CONCLUSION: Bcl-2 is an independent predictor of breast cancer outcome and seems to be useful as a prognostic adjunct to the NPI, particularly in the first 5 years after diagnosis.
Cyclin E as molecular marker in the management of breast cancer: a review.
Recent studies found cyclin E to be a promising prognostic indicator in breast cancer and examined its potential as a target for therapy.
NCOR1 mRNA is an independent prognostic factor for breast cancer.
We found risk for breast cancer was associated with the APOE genotype (x(2) = 8.652, p = 0.013).
Moreover, BRCA1 expression was elevated in HER4-positive human breast cancer specimens.
RESULTS: E-cadherin mRNA expression was lower in breast carcinoma (P = 0.001), whereas Snail expression was higher (P = 0.003).
Estrogen-responsive finger protein as a new potential biomarker for breast cancer
Efp immunoreactivity was significantly correlated with breast cancer patients
Bcl-2 is an independent predictor of breast cancer outcome
BRCA1 expression was elevated in HER4-positive human breast cancer
Cytoplasmic CLU expression

Topic

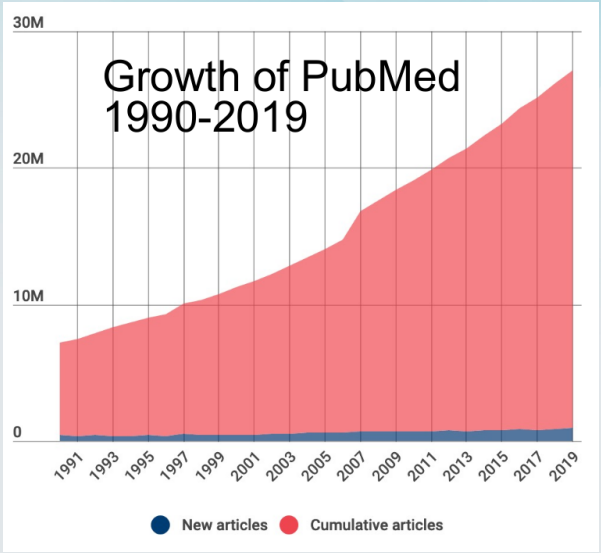
Biomedical literature

- Huge number of documents available
  - ~28 million references (Medline)
    - more than half of them have an abstract in English
  - two publications every minute
  - about 1.5 million full text (PMC)
- Aims of literature mining
  - locate and extract relevant biological information
  - integrate and cross-link it with other biological data and knowledge bases

Secondary usage of the literature

Major resource: Medline/PubMed

- One of the resources maintained by the National Library of Medicine
- Origin: Index Medicus, 1879-2004
- Going digital: MEDLARS, 1964-
  MEDLARS (Medical Literature Analysis and Retrieval System) is a computerised biomedical bibliographic retrieval system.
- Going online: MEDLINE, 1971-
  In late 1971, an online version called MEDLINE ("MEDLARS Online") became available as a way to do online searching of MEDLARS from remote medical libraries. This early system covered 239 journals and boasted that it could support as many as 25 simultaneous online users.
- Going on the web: PubMed, 1997-
  more than 28 million records from 5,639 selected publications covering biomedicine and health from 1950 to the present.


Growth of PubMed
1990-2019
30M
20M
10M
0
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
New articles
Cumulative articles

Open Access

- 2003, Bethesda Statement on Open Access Publishing (http://www.earlham.edu/~peters/fos/bethesda.htm)
- 2003, Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities https://openaccess.mpg.de/Berlin-Declaration
- 2004, PubMedCentral (PMC) https://www.ncbi.nlm.nih.gov/pmc/
  - 4.5 million articles
  - 2027 "Full participation" journals
  - ~4700 Partial participation journals

Open Access

- NIH requirement
  - "Policy on Enhancing Public Access to Archived Publications Resulting From NIH-Funded Research,"
  - all publications generated by NIH-funded research (in whole or part) must be submited to PMC
  - from May 2, 2005
  - NLM is digitizing earlier print issues of many of the journals already in PMC, extending the availability of full texts back to before the implementation of the 2005 policy.

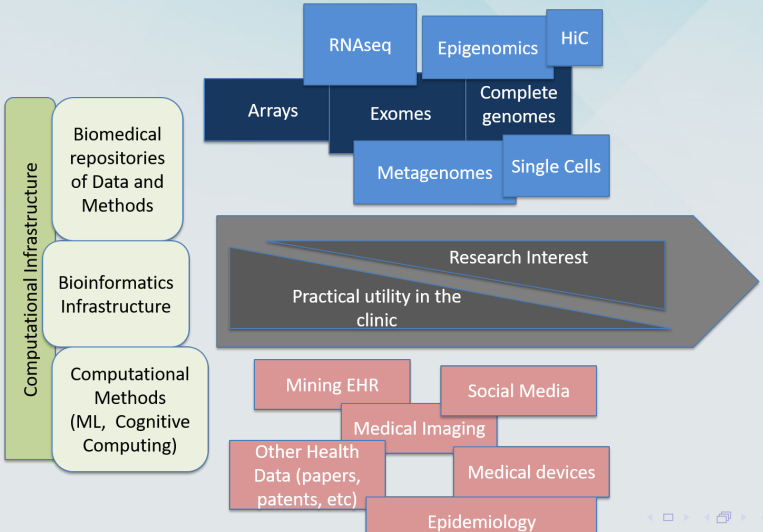
NOT Open Access

Unfortunately, free full-text access remains impossible for a large portion of scientific journals. In some fields, such as chemistry, even article abstracts are inaccessible for a large-scale analysis. The obvious outcome is that articles published in open-access journals have a better chance of being identified as relevant hits than others appearing in traditional "closed-access" journals. Electronic access to text obviously impacts all stages of text mining.
[Rzhetsky et al., 2009]

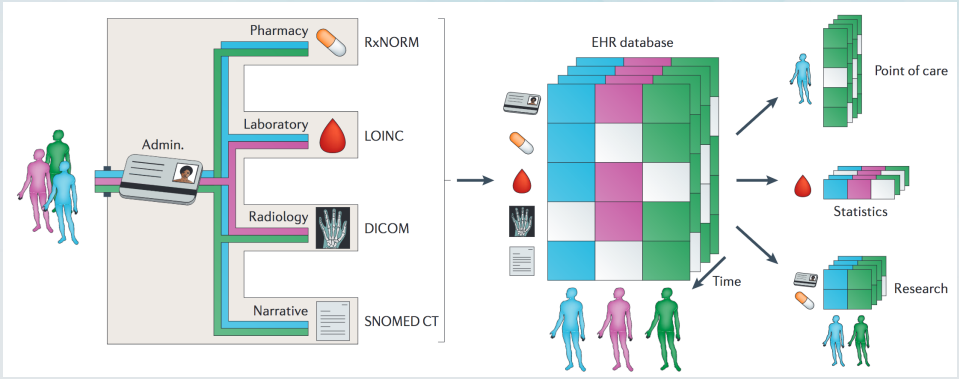
Why PMC?

PMC "makes it possible to integrate the literature with a variety of other information resources such as sequence databases and other factual databases that are available to scientists, clinicians and everyone else interested in the life sciences. The intentional and serendipitous discoveries that such links might foster excite us and stimulate us to move forward."

Not only PubMed: Precision Medicine



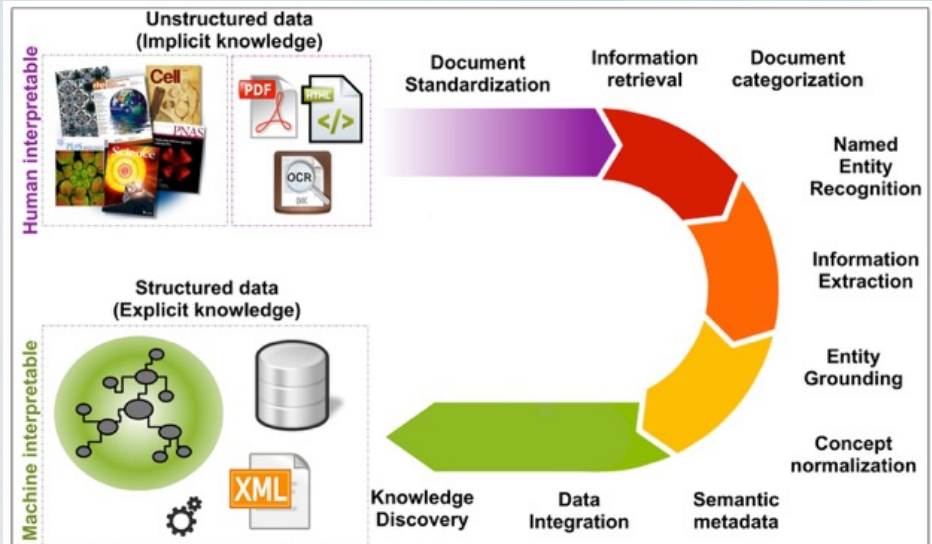
Precision Medicine



Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature reviews. Genetics, 13(6), 395.

Topic

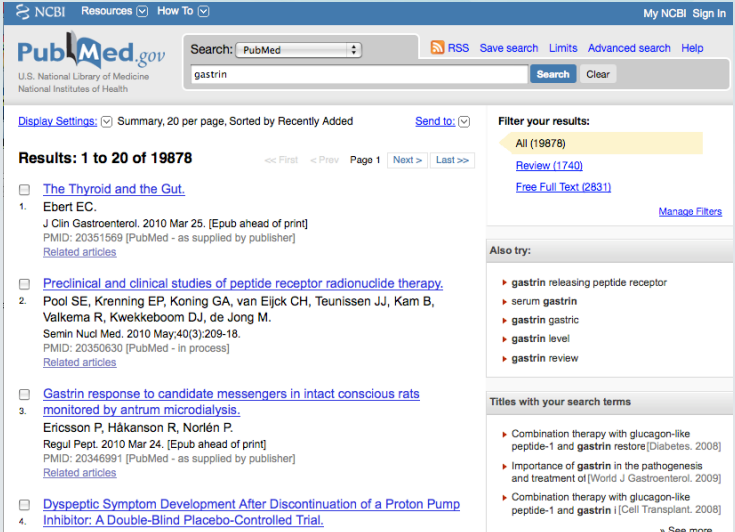
Canonical Text Mining flowchart



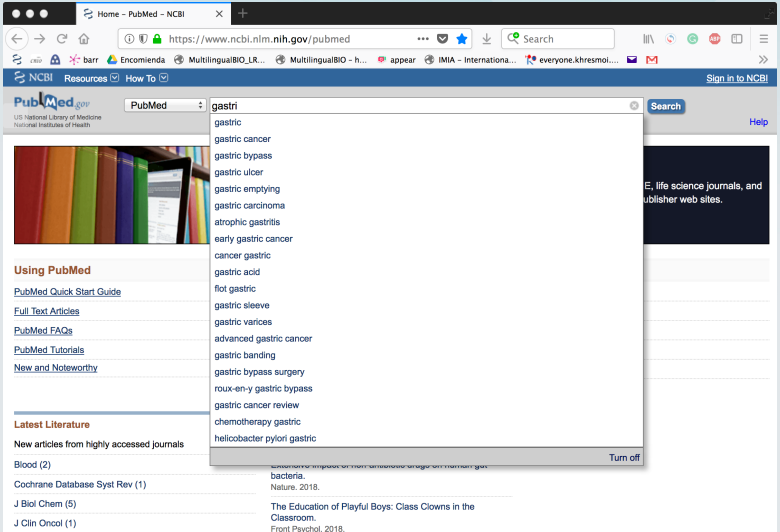
IR, IE, DM

- Information Retrieval (IR)
  - find relevant documents (unstructured)
  - example: PubMed search
- Information Extraction (IE)
  - find detailed information (semi-structured)
  - NER: Named Entity Recognition
  - Concept Recognition
  - Relation Extraction
- Data Mining (DM)
  - find associations, build networks, make predictions

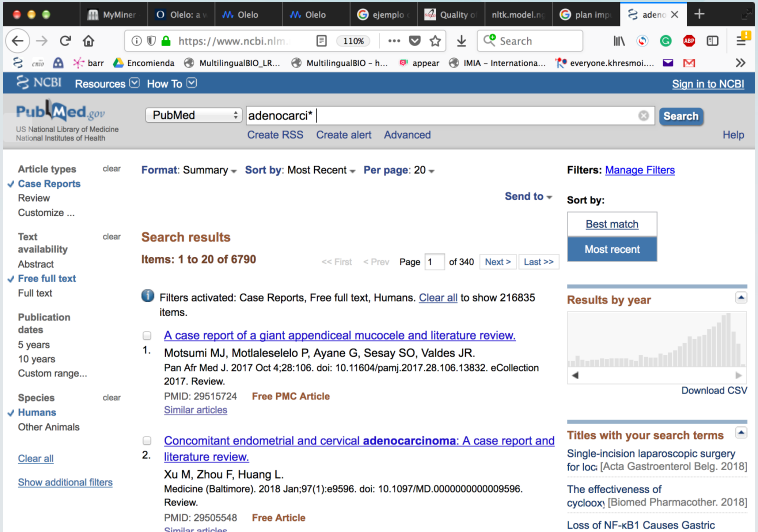
PubMed as an IR engine



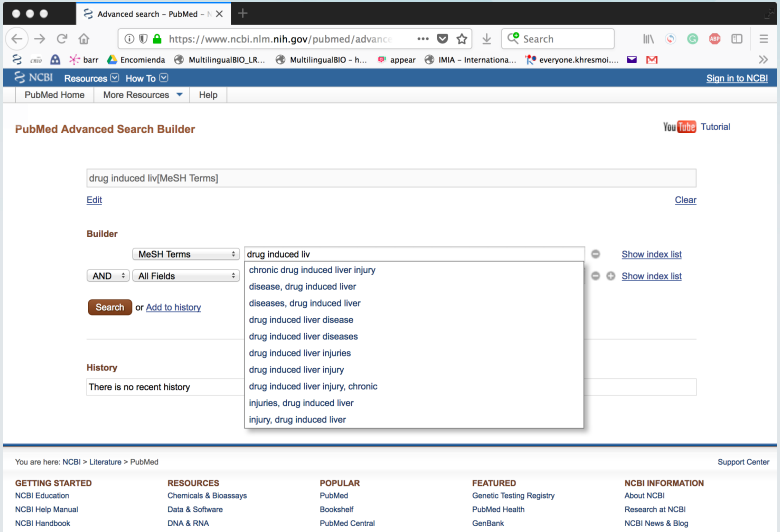
PubMed as an IR engine



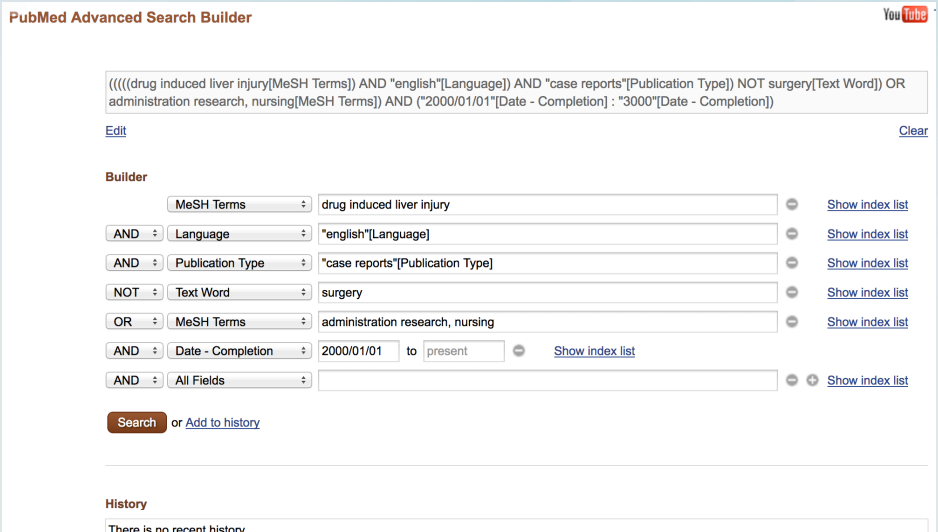
PubMed as an IR engine



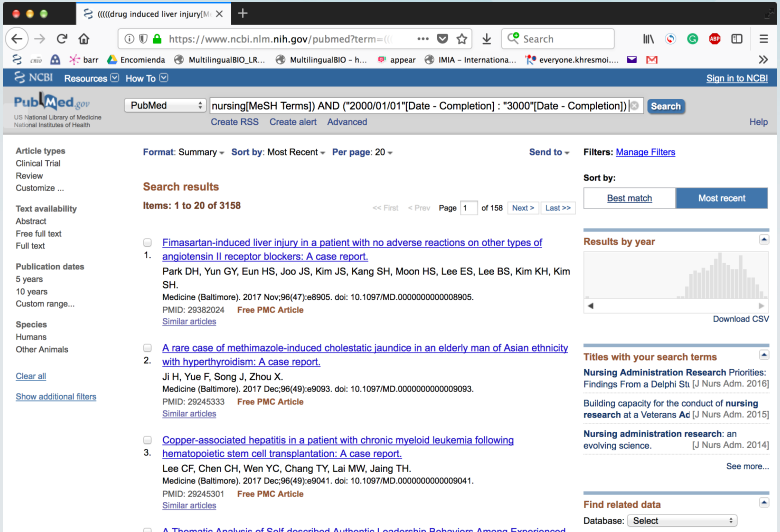
PubMed as an IR engine



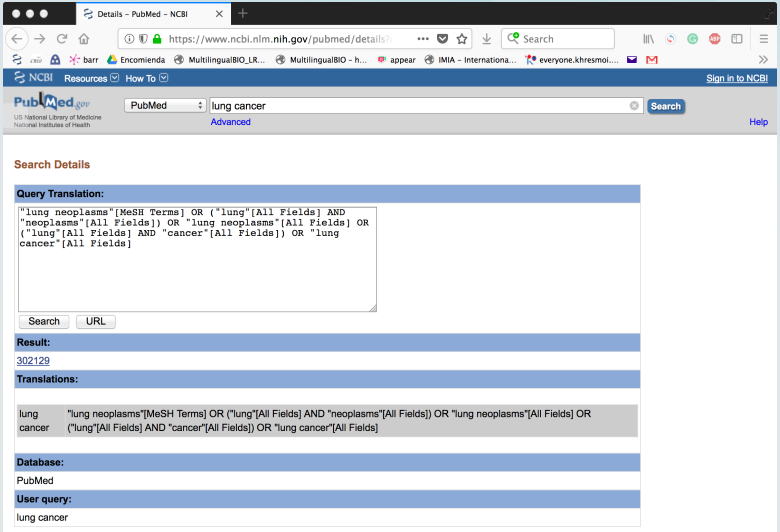
PubMed as an IR engine



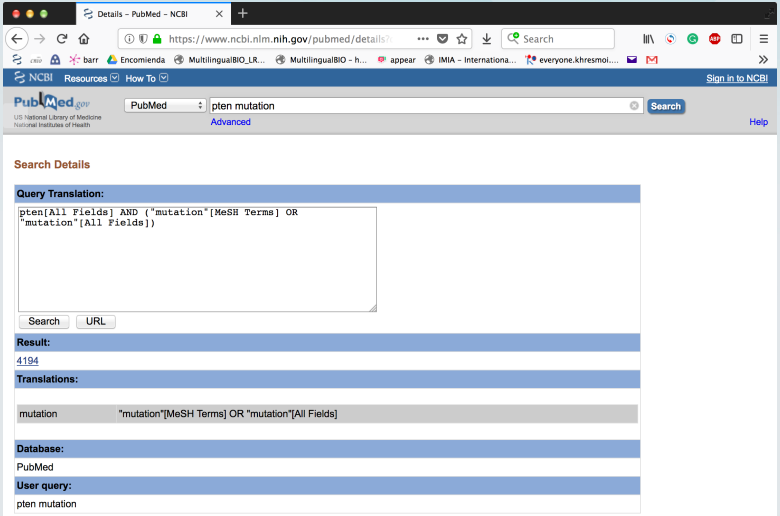
PubMed as an IR engine

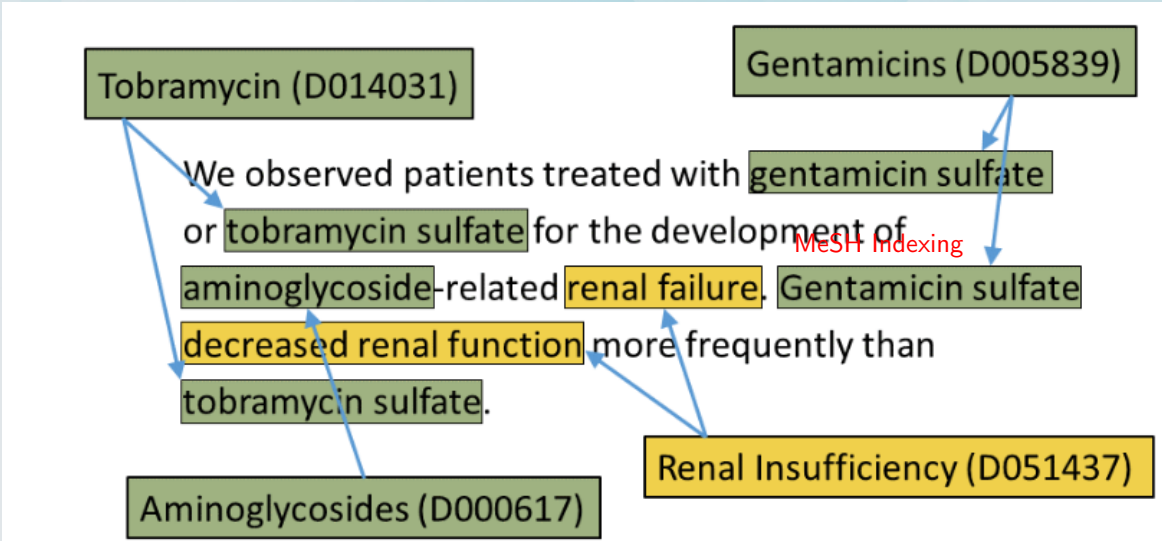


PubMed as an IR engine



PubMed as an IR engine




Tobramycin (D014031)
Gentamicins (D005839)
We observed patients treated with gentamicin sulfate
or tobramycin sulfate for the development of
aminoglycoside-related renal failure. Gentamicin sulfate
decreased renal function more frequently than
tobramycin sulfate.
MeSH Indexing
Renal Insufficiency (D051437)
Aminoglycosides (D000617)

J Alzheimers Dis. 2012;32(2 Disease Gene Variant

Highly pathogenic Alzheimer's disease presenilin 1 P117R mutation causes a specific increase in p53 and p21 protein levels and cell cycle dysregulation in human lymphocytes.

Bialopiotrowicz E[1], Szybinska A, Kuzniewska B, Buizza L, Uberti D, Kuznicki J, Wojda U.

Rank	Gene or Protein ID	Gene SYM	WTAA	MTAA	POS	Disease	PMIDs
1	Q13131	PRKAA1	Q	R	16	Breast cancer	16959974
2	P31749	AKT1	E	K	17	Breast cancer	17611497 18954143 19713527 21793738
3	P10275	AR	H	Y	874	Prostate cancer	17591767
	.....	....	....	....	....	....	.....

NLP steps

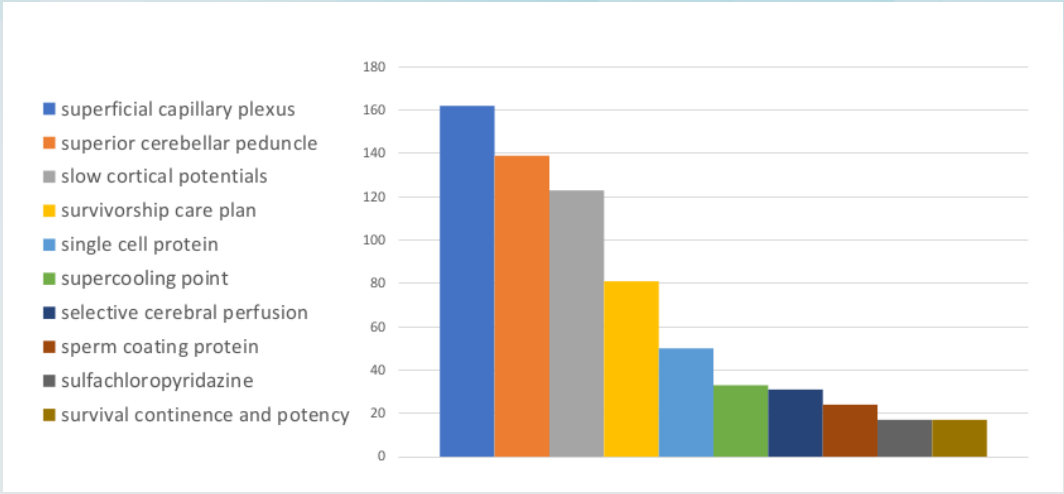
- sentence segmentation splitting a multi-sentence input into its individual sentences
  in biomedical texts, sentences can even begin with lower-case letters, e.g.,
  lush mutants are also defective for pheromone-evoked behavior (PMID 15664171)

NLP steps

- Tokenization is the task of splitting an input into individual words (as well as other units, such as punctuation marks).
  - apparently trivial, but in fact difficult even in general text
  - plus a number of domain-specific challenges in biomedical text Example: hyphen.
    - *-fever* absence of a fever
    - *Cl-* negative charge on a chlorine ion
    - *BRCA1-* non-functional gene
    - *BRCA1-IRIS* joins two synonymous forms of the gene name
    - *BRCA1-mediated* syntactic, not semantic
    - *BRCA1-CtIP* the BRCA1 protein forms a complex with the CtIP substrate of ATM protein kinase

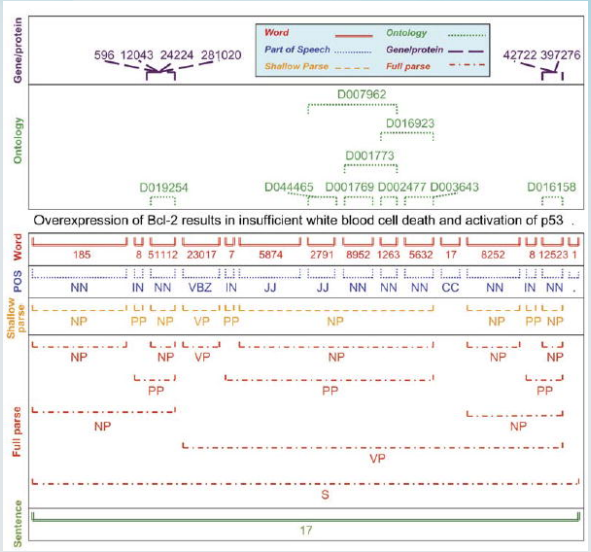
NLP steps

- Abbreviations E.g. does "PDA" mean *patent ductus arteriosus* or *posterior descending artery* or *phorbol 12,13 diacetate*?
  Almost 22% of abbreviations in one sample of biomedical text having more than one possible expansion and an average of 4.61 possible definitions for abbreviations six or fewer characters long [Chang et al., 2002]
  Possible approach: Schwartz and Hearst's algorithm (Schwartz and Hearst 2003); easily implementable; returns a single output with a high likelihood of being correct, and of requiring no connections with external applications


superficial capillary plexus
superior cerebellar peduncle
slow cortical potentials
survivorship care plan
single cell protein
supercooling point
selective cerebral perfusion
sperm coating protein
sulfachloropyridazine
survival continence and potency
180
160
140
120
100
80
60
40
20
0

NLP steps

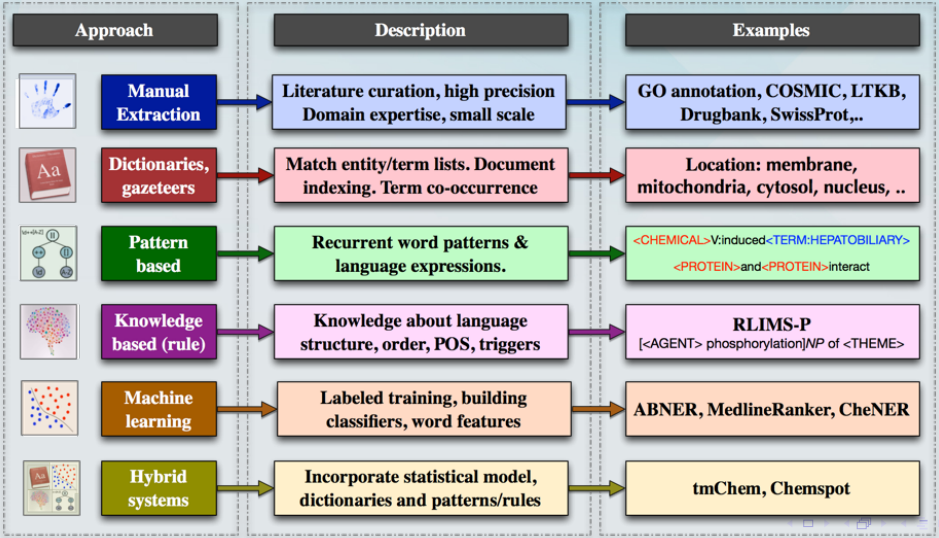
- Named Entity Recognition
  - Entity recognition is a required component for almost any biomedical text mining application.
  - Examples
    - ABNER system (Settles 2005). Its advantages include robustness (it is distributed as a .jar file and works well on multiple platforms) and a variety of input and output format options. In addition to gene/protein names, it also locates cell types and cell lines.
    - LingPipe's gene-mention-finding module
    - MetaMap tool (Aronson 2001) recognition of a wider range of biomedically relevant categories–diseases, drugs, chemicals, treatment modalities, and the like, available as a Java API or via a server. Recognizes variant forms of domain-relevant concepts
    - Banner http://banner.sourceforge.net/



BioNLP analysis levels

[Hunter and Cohen, 2006]

TM approaches



Problems

- Synonyms and Term variability: one concept, multiple names For example, BRCA1 could be referred to by any of its alternate symbols *IRIS*, *PSCP*, *BRCAI*, *BRCC1*, or *RNF53* (or by any of their many spelling variants, which include *BRCA1*, *BRCA-1*, and *BRCA 1*) or by any of the variants of its full name, viz. *breast cancer 1, early onset* (its official name per Entrez Gene and the Human Gene Nomenclature Committee), as *breast cancer susceptibility gene 1*, or as the latter's variant *breast cancer susceptibility gene-1*. Similarly, *breast cancer* could be referred to alternatively as *carcinoma of the breast*, or *mammary neoplasm*.

- pneumonia of unknown aetiology
- 2019-nCov infection
- novel coronavirus pneumonia
- SARS-CoV-2 infection
- COVID-19
- coronavirus disease 2019
- 2019 novel coronavirus infection disease
- nCOVID-19
- severe acute respiratory syndrome coronavirus 2 infection
- Coronavirus disease of 2019
- Wuhan coronavirus pneumonia
- CoV 19 infection
- COIVD-19 disease
- COVID-19 ARDS
- nCov-19 infection
- SARS-CoV-2 infectious disease
- CV-19
- coronavirus 2 syndrome
- SARS-CoV-2 associated ARDS
- C19

Problems

- Ambiguity: one name, multiple concepts/meanings A primary problem that either type of system must deal with is the issue of ambiguity: the existence of multiple relationships between language and meanings or categories. Ambiguity exists at every level of linguistic structure, from the part of speech of words to subtle issues in pragmatics.
  - with general language
  - across entity types
    - metonimy
  - within the same entity type
    - across species
    - within a species
  - grammatical ambiguity

Ambiguity

- across species fat also turns out to be the name or symbol of a number of different genes. Humans, mice, rats, Drosophila, zebrafish, chickens, M. mulatta, and two Lactobacilli have at least one gene whose name, official symbol, or alias is fat.
- within a species in humans, fat is the official symbol of Entrez Gene entry 2195 and an alternate symbol for Entrez Gene entry 948. The distinction is not trivial. The former is a cadhedrin, and is associated with tumor suppression and with bipolar disorder, while the latter is a thrombospondin receptor associated with atherosclerosis, platelet glycoprotein deficiency, hyperlipidemia, and insulin resistance, to name just a few phenotypes. These ambiguities are not trivial: if your analysis is wrong, you miss or erroneously extract information on relations between molecular biology and human disease.

- "… suggests a role for BRCA1 in HIF-1alpha regulation."
  - Human = NCBI Gene 672
  - M. musculus = NCBI Gene 12189
  - R. norvegicus = NCBI Gene 497672

- "… to lower blood cholesterol levels in comparison to sugar cane policosanol (SCP) in rabbits."

Source: PMID 18030615

Ambiguity

- grammatical ambiguity *fat*: noun or verb? PubMed returns almost 112 K hits for that single-word query (and more than 13 K even if we try to restrict the query to genomics by including the disjunction (gene OR genetic OR genetics).
- using a common term many Drosophila symbols and names are the same as common English words: *a*, *to*, and *And* are all symbols of Drosophila genes (Entrez Gene IDs 43852, 43036, and 44913, respectively) [Hunter and Cohen, 2006]
- metonimy "referring to something by an entity that is related to it": it is often not clear whether a string like p53 refers to the gene of that name, to the protein that it codes for, or to its mRNA

Met28 binds to DNA

…binding of Met28 to DNA…

…Met28 and DNA bind…

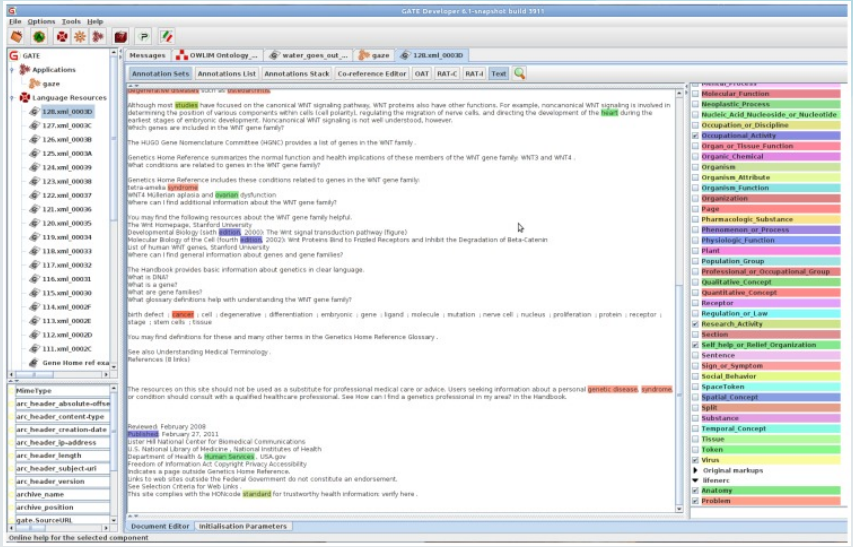
…binding between Met28 and DNA…

…Met28 is sufficient to bind DNA…

…DNA bound by Met28…

Generic Environments for TM pipelines

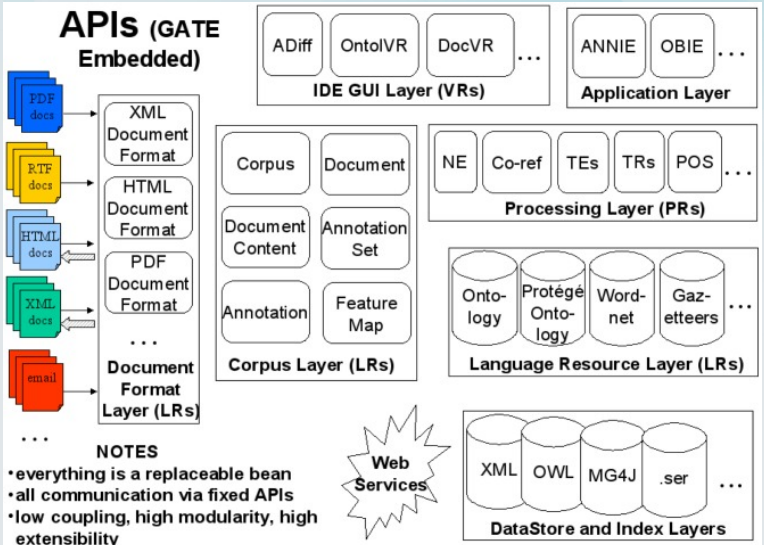
- GATE (General Architecture for Text Engineering, Cunningham 2002) system. It has been available since the mid 1990s, is used by many groups, and has been heavily field-tested over the years.
- UIMA (Unstructured Information Management Architecture, Ferrucci and Lally 2004, Mack et al. 2004), IBM, java
- LingPipe, produced by Alias-i, offers a mid-point of sorts between unintegrated pipelines and comprehensive architectures.



GATE

[Cunningham et al., 2013]

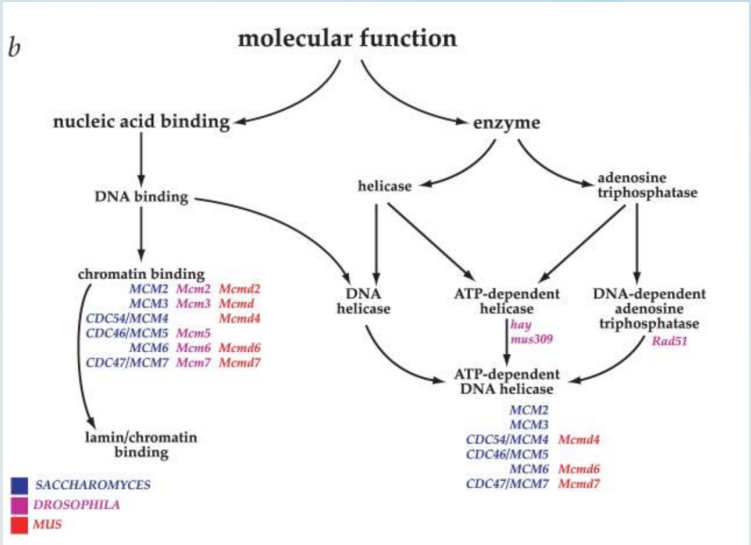
GATE


APIs (GATE Embedded)
ADiff
OntoIVR
DocVR
...
IDE GUI Layer (VRs)
ANNIE
OBIE
...
Application Layer
PDF docs
RTF docs
HTML docs
XML docs
email
...
XML Document Format
HTML Document Format
PDF Document Format
...
Document Format Layer (LRs)
Corpus
Document
Document Content
Annotation Set
Annotation
Feature Map
Corpus Layer (LRs)
NE
Co-ref
TEs
TRs
POS
...
Processing Layer (PRs)
Onto-logy
Protégé Onto-logy
Word-net
Gaz-etteers
...
Language Resource Layer (LRs)
NOTES
•everything is a replaceable bean
•all communication via fixed APIs
•low coupling, high modularity, high extensibility
Web Services
XML
OWL
MG4J
.ser
...
DataStore and Index Layers

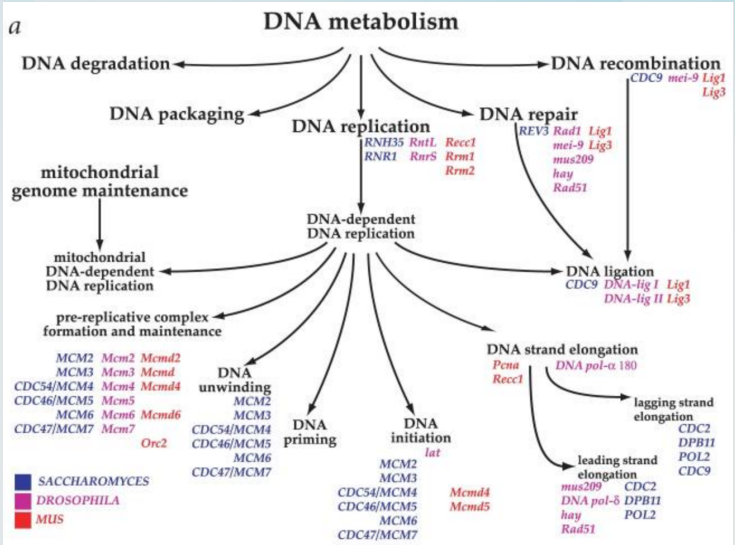
Ontologies

- are the most effective representational mechanism
- verbs and relations might benefit from richer, frame-like or event-related structure
- Gene Ontology [Ashburner et al., 2000] The ontologies are built from a structured, controlled vocabulary.

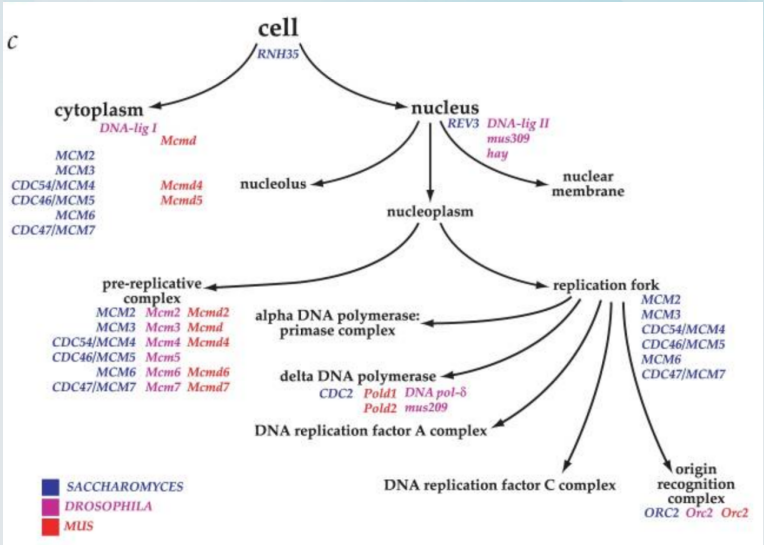
GO


b
molecular function
nucleic acid binding
enzyme
DNA binding
helicase
adenosine triphosphatase
chromatin binding
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7 Mcmd7
DNA helicase
ATP-dependent helicase
hay
mus309
DNA-dependent adenosine triphosphatase
Rad51
ATP-dependent DNA helicase
MCM2
MCM3
CDC54/MCM4 Mcmd4
CDC46/MCM5
MCM6 Mcmd6
CDC47/MCM7 Mcmd7
lamin/chromatin binding
SACCHAROMYCES
DROSOPHILA
MUS

GO

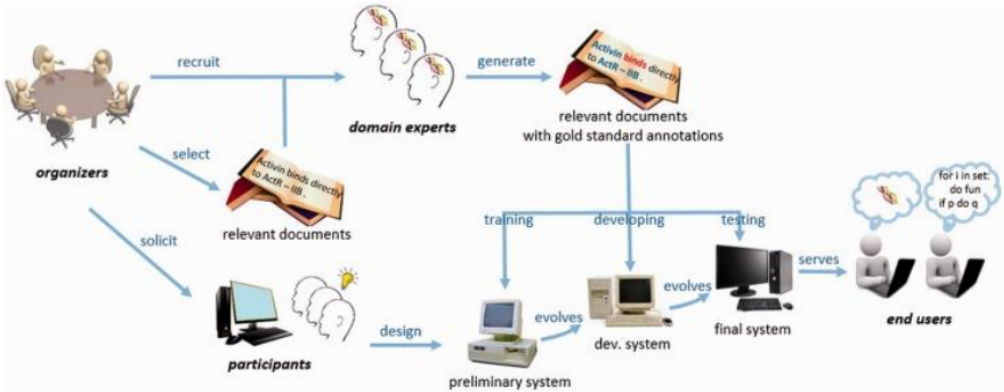

a
DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNH35 RntL Recc1
RNR1 RnrS Rrm1
Rrm2
DNA repair
REV3 Rad1 Lig1
mei-9 Lig3
mus209
hay
Rad51
DNA recombination
CDC9 mei-9 Lig1
Lig3
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
pre-replicative complex formation and maintenance
DNA ligation
CDC9 DNA-lig I Lig1
DNA-lig II Lig3
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
DNA initiation
lat
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
DNA strand elongation
Pcna
Recc1
DNA pol-α 180
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
mus209
DNA pol-δ
hay
Rad51
CDC2
DPB11
POL2
SACCHAROMYCES
DROSOPHILA
MUS

GO


C
cell
RNH35
cytoplasm
DNA-lig I
Mcmd
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
nucleus
REV3
DNA-lig II
mus309
hay
nucleolus
nuclear membrane
nucleoplasm
pre-replicative complex
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7 Mcmd7
alpha DNA polymerase: primase complex
delta DNA polymerase
CDC2 Pold1 DNA pol-δ
Pold2 mus209
DNA replication factor A complex
replication fork
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA replication factor C complex
origin recognition complex
ORC2 Orc2 Orc2
SACCHAROMYCES
DROSOPHILA
MUS

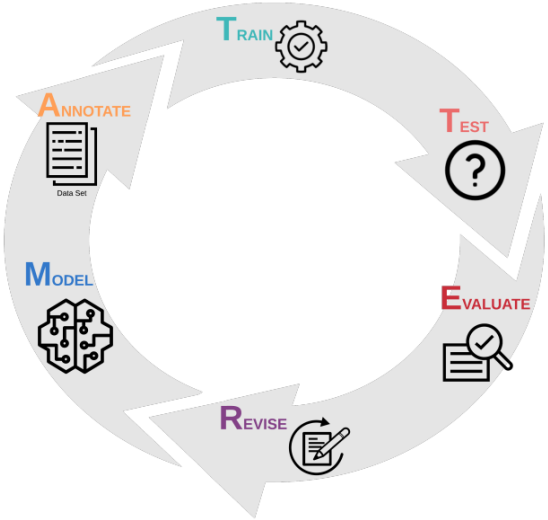
Competitive Evaluations

- also called "Shared Tasks", "Competitive scientific challenges"

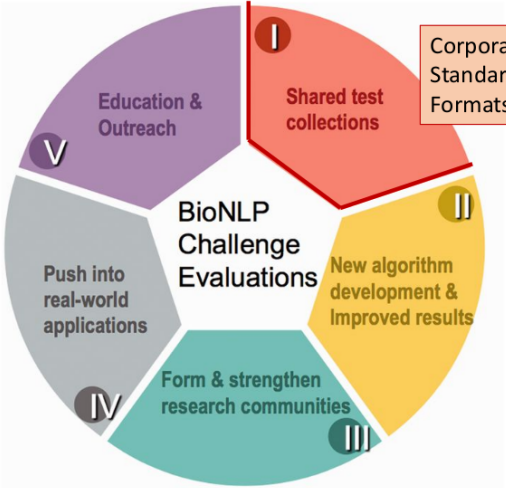


Competitive Evaluations

- TREC Genomics Track (2003-2007) http://trec.nist.gov/data/genomics.html
  Supported by the National Institute of Standards and Technology, TREC provides a forum for evaluation of information retrieval systems.
- BioNLP competition
- BioCreative
- IE Task @ KDD Cup earliest biomedical text data mining competition, an information extraction task sponsored by the Knowledge Discovery and Data mining (KDD) Cup in which participants built systems to aid in the FlyBase curation process (Yeh et al., 2003) [Yeh et al., 2003]



MATTER Cycle concept: Pustejovsky J. & Stubbs A. (2013) Natural Language Annotation for Machine Learning. Edited by O'Reilly

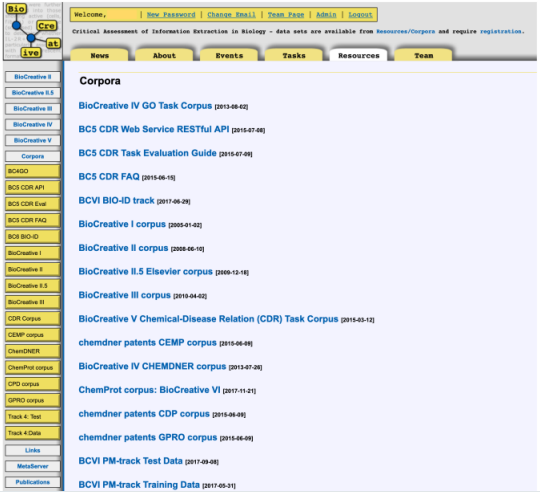


Corpora
Standards
Formats for NLP interoperability

Modified from Chung-Chi Huang, and Zhiyong Lu Brief Bioinform 2015;bib.bbv024

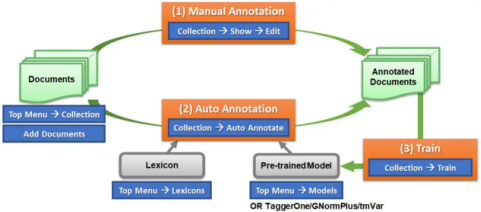
Briefings in Bioinformatics

http://www.biocreative.org

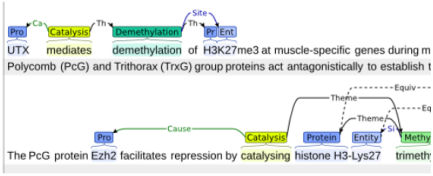


- High quality corpora
- Covers abstracts, full text and patents
- Annotation guidelines aligned with domain experts
- Inter-annotator agreement

ezTag https://eztag.bioqrator.org/



brat https://brat.nlplab.org/



TeamTat https://www.teamtat.org/

TextAE http://textae.pubannotation.org/

Literature-Based Discovery

Literature-based discovery is the attempt to automatically induce novel hypotheses by processing existing publications. [Swanson, 1988]

Early example: Arrowsmith system by Don R. Swanson (1924-2012).

"Imagine that the pieces of a puzzle are independently designed and created, and that, when retrieved and assembled, they then reveal a pattern undesigned, unintended, and never before seen, yet a pattern that commands interest and invites interpretation. So it is, I claim, that independently created pieces of knowledge can harbor an unseen, unknown, and unintended pattern. And so it is that the world of recorded knowledge can yield genuinely new discoveries"

https://en.wikipedia.org/wiki/Arrowsmith_System

Chillibot

http://www.chilibot.net [Chen and Sharp, 2004]

Chilibot is a freely available, web-based application that takes sets of gene names, and (optionally) additional keywords as input, and finds information about the relationships among them. In an initial information retrieval step, it constructs queries to send to the PubMed search engine. It uses basic language processing techniques to identify sentences that describe stimulatory, inhibitory, and other relationships between pairs of genes. For pairs of genes for which large sets of sentences are found, it then uses techniques from the field of automatic summarization (the area of natural language processing concerned with constructing shortened versions of input texts) to select the best sentences to display to the user. A graph displays the entire set of interactions amongst all of the genes that were input. By clicking on links between pairs of genes, the user can see the full set of sentences that describe relations between those genes, and clicking on the sentences themselves displays the original PubMed abstract.

Textpresso

http://www.textpresso.org [MIler et al., 2004]

The Textpresso system is an example of an information extraction project that strives to create an up-to-date summary of current knowledge related to a specific model organism. Users can search for information involving any of the 33 semantic classes of things and relationships between them that the system extracts information about (e.g., genes, clones, pathways, or mutations). The original system, devoted to C. elegans, has scanned thousands of full-text articles, and the resulting database is accessed more than 1500 times daily by nematode biologists worldwide. Work is in progress to extend this method to a variety of other organisms, including N. crassa, D. melanogaster, and S. cerevisiae, and preliminary systems are now available on several model organism database sites.

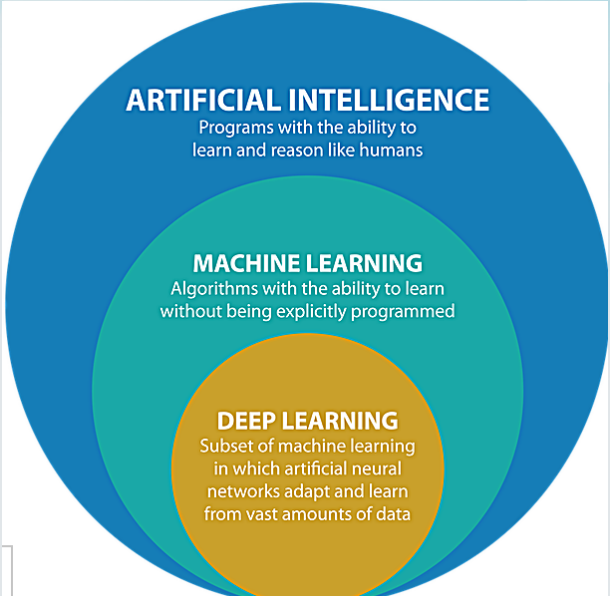
iHOP

http://www.ihop-net.org/

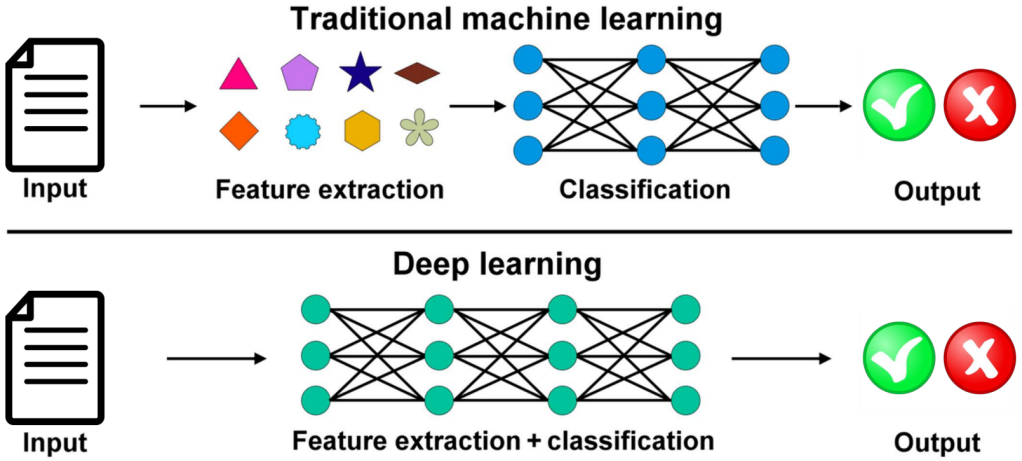
Information Hyperlinked over Proteins (or iHOP) is an online text-mining service that provides a gene-guided network to access PubMed abstracts. The service was established by Robert Hoffmann and Alfonso Valencia in 2004.

The concept underlying iHOP is that by using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource. Navigating across interrelated sentences within this network rather than the use of conventional keyword searches allows for stepwise and controlled acquisition of information. Moreover, this literature network can be superimposed upon experimental interaction data to facilitate the simultaneous analysis of novel and existing knowledge

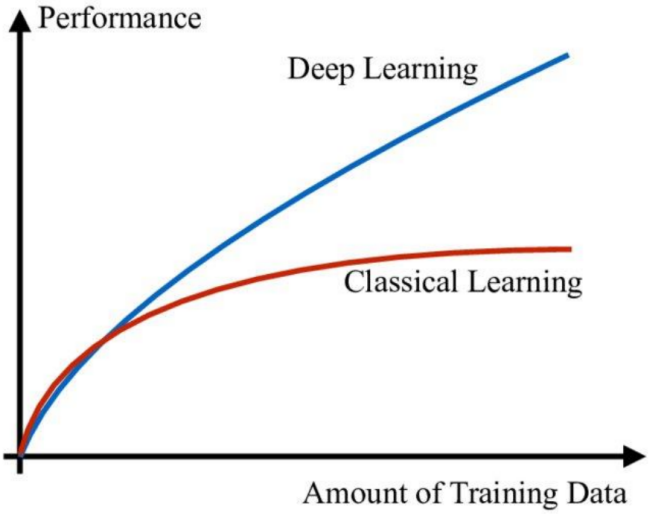
https://en.wikipedia.org/wiki/Information_Hyperlinked_over_Proteins



What is deep learning?

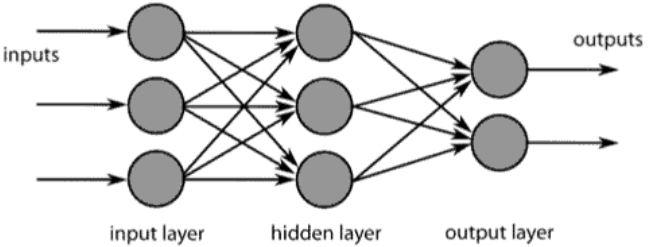


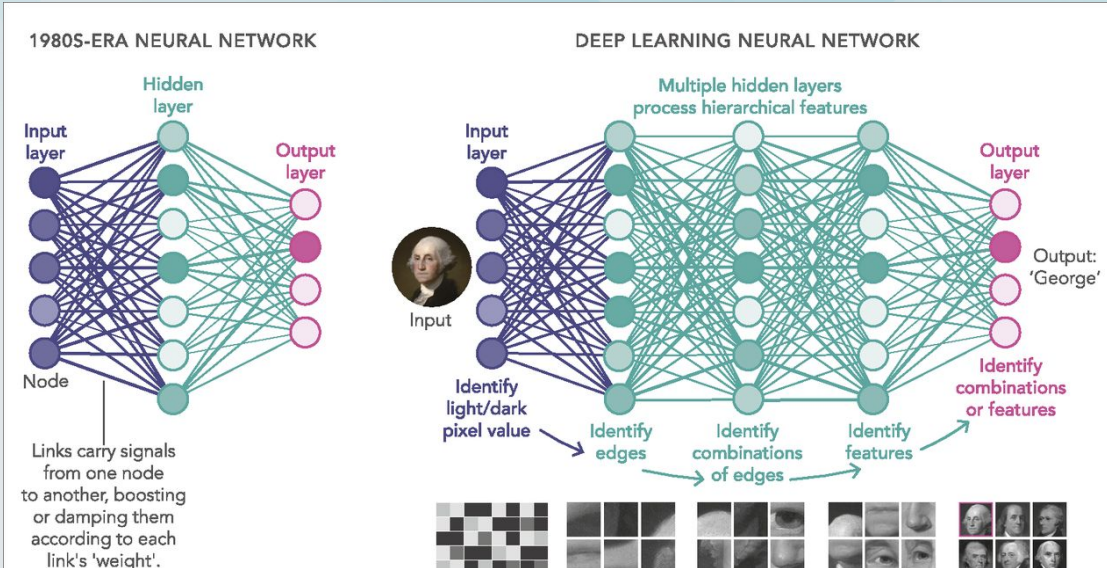
Renn, Alex, et al. "Advances in the prediction of mouse liver microsomal studies: From machine learning to deep learning." *Wiley Interdisciplinary Reviews: Computational Molecular Science* (2020): e1479.

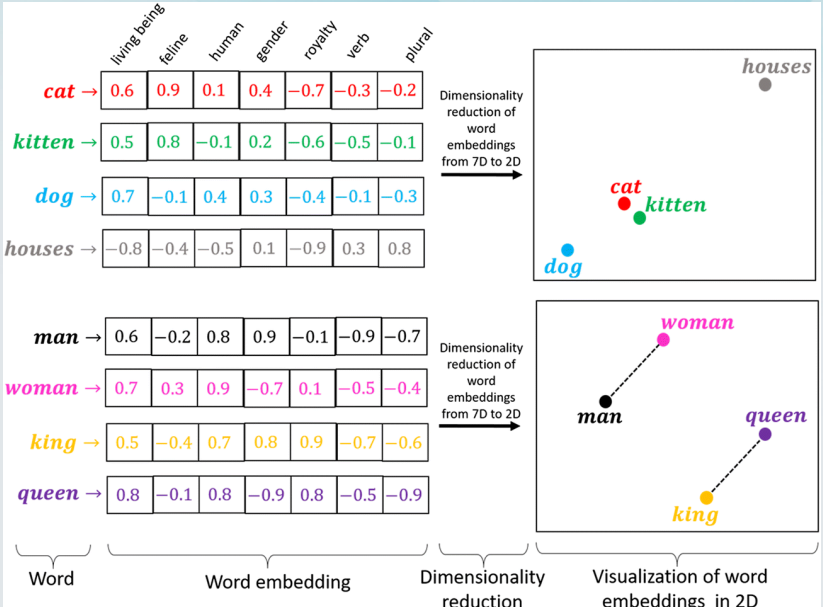


Zappone, Alessio, Marco Di Renzo, and Mérouane Debbah. "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?." *IEEE Transactions on Communications* 67.10 (2019): 7331-7376.

Was ist Deep Learning?




1980S-ERA NEURAL NETWORK
Hidden layer
Input layer
Output layer
Node
Links carry signals from one node to another, boosting or damping them according to each link's 'weight'.
DEEP LEARNING NEURAL NETWORK
Multiple hidden layers process hierarchical features
Input layer
Output layer
Input
Output: 'George'
Identify light/dark pixel value
Identify edges
Identify combinations of edges
Identify features
Identify combinations or features


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

word2vec

- Focus on local context
- Implicitly learn embeddings based on local co-occurrences
- Has two versions: cbow and skip-gram

glove

- Focus on global context
- Explicitly learn embeddings based on global co-occurrences

fastText

- Extends from word2vec by using word subgrams
- Predict an out-of-vocabulary word based on subgram statistics

All of these models are Transformer models

ELMo
Oct 2017
Training:
800M words
42 GPU days

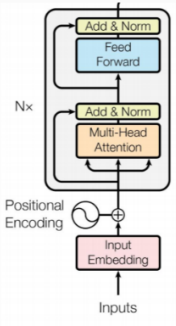
GPT
June 2018
Training
800M words
240 GPU days

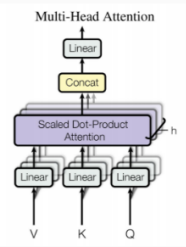
BERT
Oct 2018
Training
3.3B words
256 TPU days
~320–560
GPU days

GPT-2
Feb 2019
Training
40B words
~2048 TPU v3
days according to
a reddit thread

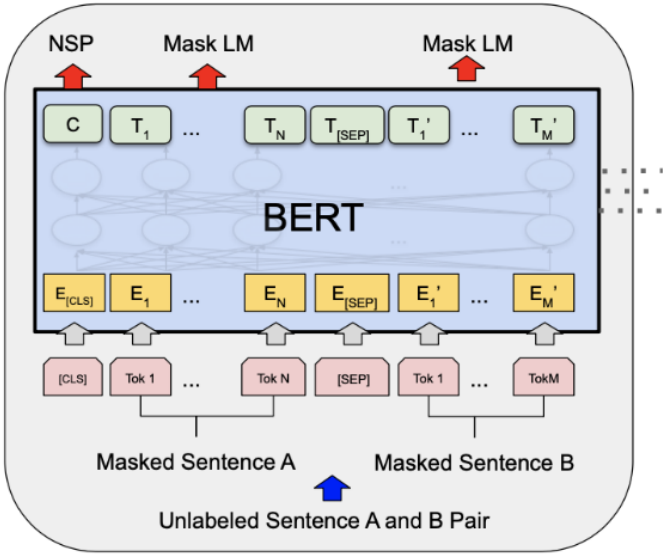
XL-Net,
ERNIE,
Grover
RoBERTa, T5
July 2019—

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning

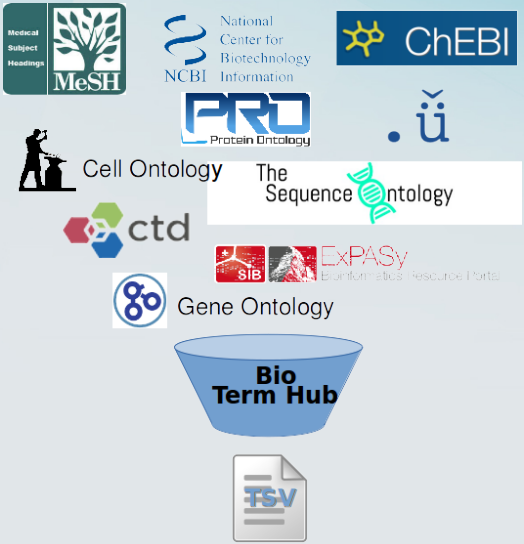




http://jalammar.github.io/illustrated-transformer/


NSP
Mask LM
Mask LM
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E1
...
EN
E[SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair

Topic


Medical Subject Headings
MeSH
National Center for Biotechnology Information
NCBI
ChEBI
PRO
Protein Ontology
Cell Ontology
The Sequence Ontology
ctd
SIB
ExPASy
Bioinformatics Resource Portal
Gene Ontology
Bio Term Hub
TSV

Bio Term Hub

Resource Selection

Please select the resources to be included:

☐ *select all*

☐ Cell Ontology	(→ source)	Update available.	Update
☐ Cellosaurus	(→ source)	Up-to-date.	...
☐ ChEBI	(→ source)	Up-to-date.	...
☐ CTD chemicals	(→ source)	Up-to-date.	...
☐ CTD diseases	(→ source)	Up-to-date.	...
☐ EntrezGene	(→ source)	Update available.	Update
☐ Gene Ontology	(→ source)	Update available.	Update
☐ MeSH	(→ source)	Up-to-date.	...
☐ NCBI Taxonomy	(→ source)	Update available.	Update
☐ Protein Ontology	(→ source)	Up-to-date.	...
☐ Sequence Ontology	(→ source)	Up-to-date.	...
☐ Swiss-Prot	(→ source)	Up-to-date.	...
☐ Uberon	(→ source)	Up-to-date.	...

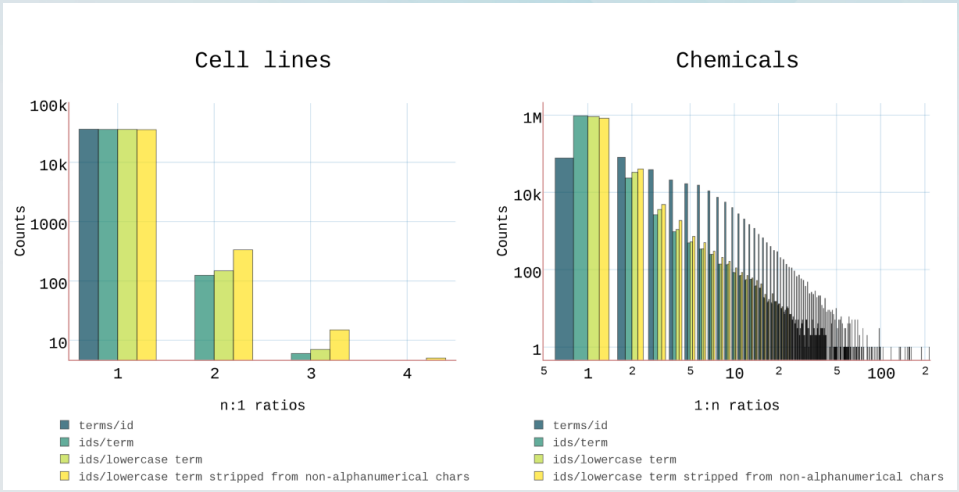
The Bio Term Hub provides access to biomedical terminology resources in a unified format. We envision its main use as a basis for text mining systems.

Select any of the external resources on the left to obtain a custom terminology list. The list is a tab-separated table including terms (names, synonyms), preferred name, concept identifier, entity type, and original resource. The contents of the last two fields can be modified using replacement patterns (uncover the **Renaming** section below using the triangle button). Submit your request with the **Create resource** button at the bottom of the page. Afterwards you will be offered the opportunity to annotate a text with the selected resources.

Aggregating, filtering, reformatting, and optionally renaming all this information takes time. Creation time mainly depends on the size of each resource. Compiling a list of all terminologies except for EntrezGene takes approximately 90 seconds; however, EntrezGene is very large, and processing it requires 3 to 4 minutes.

The requested terminology list is compiled on the fly, based on local copies of the external terminology resources. Whenever one of the local copies becomes out-of-date with respect to its remote source, an **Update** button is shown next to the corresponding resource name. This button triggers downloading the latest version to our server. Please note that, due to data volume and bandwidth restrictions, updating may take several minutes for some of the resources.

Bio Term Hub


Cell lines
Counts
100k
10k
1000
100
10
1
2
3
4
n:1 ratios
terms/id
ids/term
ids/lowercase term
ids/lowercase term stripped from non-alphanumerical chars
Chemicals
Counts
1M
10k
100
1
5
1
2
5
10
2
5
100
2
1:n ratios
terms/id
ids/term
ids/lowercase term
ids/lowercase term stripped from non-alphanumerical chars

BTH: Term Statistics

Tabelle: Overview of Termfile Statistics

	genes/proteins	chemicals	diseases	species	cell lines	all
Terms in Resource	**10,429,162**	979,418	67,614	1,333,903	36,249	12,846,346
Avg. term length (number letters)	11.73	**37.49**	26.98	22.87	**7.611**	14.92
Avg. terms/ID	1.1455	**3.545**	**6.018**	1.326	**1.000**	1.328
Avg. IDs/term	**1.371**	1.049	**1.000**	1.003	1.004	1.306
Avg. IDs/term (case insensitive)	1.383	1.062	1.000	1.003	1.004	1.317
Avg. IDs/term (case insensitive, stripped)	1.387	1.086	1.000	1.006	1.010	1.324

BTH: Term confusion matrix

Overlap among entity types

	chemical	sequence	gene/ protein	cellular component	cell	cell line	molecular function	biological process	organism	disease
chemical		7.46%	27.19%	4.51%	0.26%	2.56%	3.70%	0.08%	0.76%	1.61%
sequence	23.68%		23.72%	15.41%	0.00%	1.69%	4.18%	0.29%	1.10%	0.36%
gene/protein	8.89%	2.44%		2.77%	1.12%	3.21%	2.22%	0.12%	2.29%	1.77%
cellular component	29.63%	31.84%	55.70%		16.74%	1.60%	4.43%	0.60%	0.22%	1.27%
cell	3.97%	0.00%	51.83%	38.57%		4.38%	0.82%	0.45%	0.06%	3.49%
cell line	17.75%	3.69%	67.86%	1.69%	2.00%		2.10%	0.19%	3.07%	5.77%
molecular function	34.50%	12.27%	63.34%	6.29%	0.51%	2.82%		0.68%	1.09%	2.30%
biological process	0.41%	0.49%	2.07%	0.50%	0.16%	0.15%	0.40%		0.03%	3.79%
organism	3.14%	1.42%	28.79%	0.14%	0.02%	1.83%	0.48%	0.02%		0.35%
disease	4.09%	0.29%	13.75%	0.49%	0.59%	2.12%	0.63%	1.78%	0.21%	

OGER

Human prostate cancer metastases target the hematopoietic stem cell niche to establish footholds in mouse bone marrow.
HSC homing, quiescence, and self-renewal depend on the bone marrow HSC niche.
A large proportion of solid tumor metastases are bone metastases, known to usurp HSC homing pathways to establish footholds in the bone marrow.
However, it is not clear whether tumors target the HSC niche during metastasis.

Legend

disease
chemical
sequence
gene/protein
biological_process
organism
cell

https://covid19.nlp.idsia.ch/oger-rest.html

https://oger.nlp.idsia.ch/

OGER: annotation service

The OntoGenes Biomedical Entity Recogniser (OGER)

- RESTful web service, using BTH terminologies
- Allows annotation of a collection of documents.
- Evaluated in the Bio Text Mining services challenge BioCreative/TIPS
  - best results according to several of the evaluation metrics.

https://covid19.nlp.idsia.ch/oger-rest.html

OGER: annotation service

- Annotates input text with entities from the BTH
  - Except EntrezGene
- Can be used as a web demo (for annotation of single articles) or as a web service (batch).
- Input: PubMed, PubMed Central, Free Text
- Formats: text(I), BioC (I/O), pxml (I), tsv (O), brat (O), odin-xml (O)

https://oger.nlp.idsia.ch/

Note: user-provided terminologies can be used, but this is not yet supported by the interface and web service.

BioCreative V.5 / TIPS

 seconds/document	1.06943 s	**1st**
 seconds/Byte	0.00086 s	**1st**
 seconds/annotation	0.07227 s	
 annotations/document	14.7923	
 time between failures	[no failure]	**1st** (shared)
 time to repair	[no failure]	**1st** (shared)

Previous history...

472 IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 7, NO. 3, JULY-SEPTEMBER 2010

OntoGene in BioCreative II.5

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Thérèse Vachon, and Martin Romacker

Abstract—We describe a system for the detection of mentions of protein-protein interactions in the biomedical scientific literature. The original system was developed as a part of the OntoGene project, which focuses on using advanced computational linguistic techniques for text mining applications in the biomedical domain. In this paper, we focus in particular on the participation to the BioCreative II.5 challenge, where the OntoGene system achieved best-ranked results. Additionally, we describe a feature-analysis experiment performed after the challenge, which shows the unexpected result that one single feature alone performs better than the combination of features used in the challenge.

Index Terms—Biomedical text mining, Natural Language Processing (NLP), protein interactions, BioCreative.

1 Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt [1] collects information on all known proteins. MINT [2] and IntAct [3] are the databases collecting protein interactions. Most of the information in these databases is derived from the primary literature by a process of manual revision known as "literature curation."

Typically, a text mining system is designed to locate relevant information within a given text collection (e.g., the PubMed database of biomedical literature), not only by finding relevant articles (which is the task of information retrieval systems), but also by extracting very specific information of interest to the user, such as, for example, protein-protein interactions (PPI).[1] Text mining systems are based on a variety of different approaches, ranging from simple bag-of-words techniques to more complex natural language approaches, with various degrees of success.

The work presented here is part of a larger effort undertaken in the OntoGene project [6] aimed at improving biomedical text mining through the usage of advanced natural language processing techniques. The results of the entity detection feed directly into the process of identification of protein interactions. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [7], [8]. The syntactic parser [9] takes into account constituent boundaries defined by previously identified multiword entities. Therefore, the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus, leads to better recognition of interactions [10].

There have recently been numerous results showing the potential of dependency-based language analysis for text mining. Pyysalo et al. [11] describe a manually annotated corpus, which includes a dependency-based analysis of each sentence. Clegg and Shepherd [12] use dependency graphs in order to benchmark four publicly available natural language parsers. Fundel et al. [13] describe a large-scale relation mining application based upon the Stanford Lexicalized Parser.

In this paper, we first describe in Section 2 the overall architecture of our text mining system. We then present in Section 3 the results obtained in the recent BioCreative II.5 competition. Additionally, in Section 4, we briefly sketch how the system could be used in the curation process.

2 The OntoGene Text Mining System

In this section, we provide an overall description of the OntoGene text mining environment, with a specific focus on its application to the detection of PPI. In Section 2.1, we explain the process used to automatically annotate different types of entities, and ground them to reference identifiers (IDs). In Section 2.2, we illustrate how we collect information about the focus organisms mentioned in the articles and how we use it to disambiguate protein mentions. In Section 2.3, we describe our approach to the detection of interactions among entities (proteins in particular).

2.1 Detection and Grounding of Domain Entities

In this section, we describe our approach to the problem of detecting names of relevant domain entities in biomedical literature (we consider, in particular, proteins, genes, species, experimental methods, and cell lines) and grounding them to widely accepted IDs assigned by four different knowledge bases: UniProt Knowledgebase [14], National

1. Thorough surveys of the role of Text Mining systems in biology can be found in [4] and [5].

• F. Rinaldi, G. Schneider, K. Kaljurand, and S. Clematide are with the Institute of Computational Linguistics, University of Zurich, Binzmühlestr. 14, CH-8050 Zurich, Switzerland. E-mail: {rinaldi, gschneid, siclemat}@ifi.uzh.ch, kalju@ifi.uzh.ch.
• T. Vachon and M. Romacker are with Novartis Pharma AG, NIBR/NITA/Text Mining Services, CH-4002 Basel, Switzerland. E-mail: {therese.vachon, martin.romacker}@novartis.com.

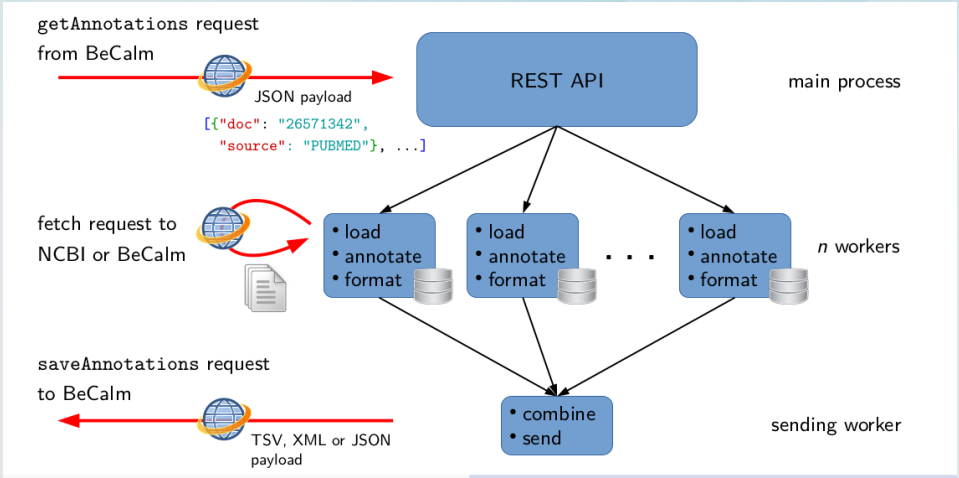
Manuscript received 11 Jan. 2010; revised 2 Apr. 2010; accepted 3 May 2010; published online 17 May 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2010-01-2013.
Digital Object Identifier no. 10.1109/TCBB.2010.50.

1545-5963/10/$26.00 © 2010 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

- [2006] BioCreative II: PPI (3rd), IMT (best)
- [2009] BioCreative II.5 PPI (best results); BioNLP
- [2010] BioCreative III: ACT, IMT, IAT
- [2011] CALBC (large scale entity extraction), BioNLP
- [2012] CTD task at BioCreative 2012
- [2013] BioCreative IV: BioC, CTD, IAT

http://www.biomext.org/

OGER in TIPS



NOTE: Just won an OpenMinTeD tender! (17'500 EUR)

Disambiguation challenges

Human prostate cancer metastases target the hematopoietic stem cell niche to establish footholds in mouse bone marrow.
HSC homing, quiescence, and self-renewal depend on the bone marrow HSC niche.
A large proportion of solid tumor metastases are bone metastases, known to usurp HSC homing pathways to establish footholds in the bone marrow.
However, it is not clear whether tumors target the HSC niche during metastasis.

Legend
disease
chemical
sequence
gene/protein
biological_process
organism
cell

gene/protein	galactoside 2-alpha-L-fucosyltransferase 1	PR:000007702	Protein Ontology
cellular component	Hedgehog signaling complex	GO:0035301	Gene Ontology
gene/protein	chaperone protein HscA	PR:000022925	Protein Ontology
chemical	N-(3-carboxypropanoyl)-N-hydroxycadaverine	CHEBI:50443	ChEBI

Term ambiguity

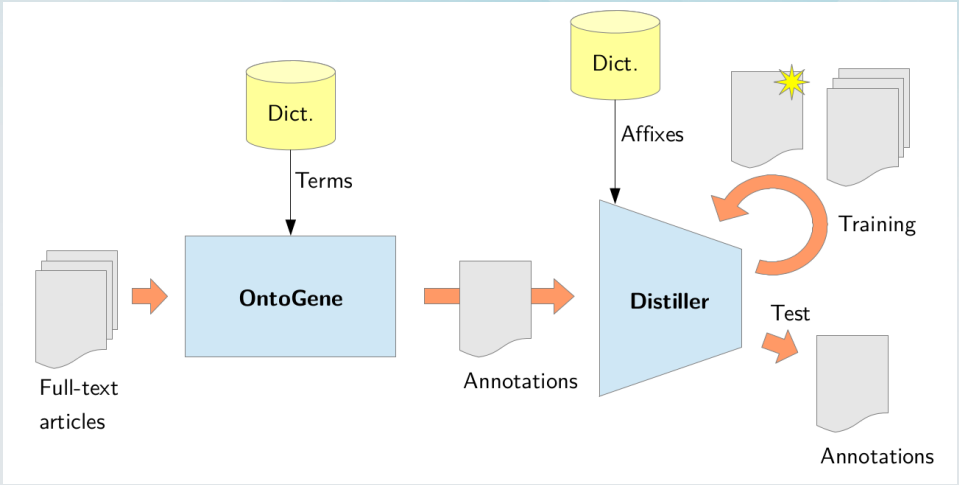
Name/Gene ID	Description	Location	Aliases
thiS ID: 2847702	immediate sulfur donor in thiazole formation [*Escherichia coli str. K-12 substr. MG1655*]	NC_000913.3 (4192637..4192837, complement)	b4407, ECK3983, JW3955
thiS ID: 939811	sulfur carrier protein ThiS [*Bacillus subtilis subsp. subtilis str. 168*]	NC_000964.3 (1244844..1245044)	BSU11680
thiS ID: 4524717	ThiS [*Thalassiosira pseudonana*]	NC_008589.1 (84740..84952)	ThpsCp091, ycf40
thiS ID: 4524549	ThiS [*Phaeodactylum tricornutum*]	NC_008588.1 (81332..81544)	PhtrCp091
ThiS ID: 33350719	ThiS [*Sheathia arcuata*]	NC_035231.1 (47491..47706, complement)	CGW65_pgp144

Term ambiguity

Name/Gene ID	Description	Location	Aliases	MIM
WAS ID: 7454	Wiskott-Aldrich syndrome [*Homo sapiens* (human)]	Chromosome X, NC_000023.11 (48683753..48691427)	IMD2, SCNX, THC, THC1P, WASPA, WAS	300392
Was ID: 22376	Wiskott-Aldrich syndrome [*Mus musculus* (house mouse)]	Chromosome X, NC_000086.7 (8081466..8090491, complement)	U42471p, Was	
Was ID: 317371	Wiskott-Aldrich syndrome [*Rattus norvegicus* (Norway rat)]	Chromosome X, NC_005120.4 (15155246..15164099)	WASP	
WAS ID: 107131173	Wiskott-Aldrich syndrome [*Bos taurus* (cattle)]	Chromosome X, AC_000187.1 (91842853..91848367)		

Term ambiguity

Name/Gene ID	Description	Location	Aliases	MIM
CAT ID: 847	catalase [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (34438925..34472060)		115500
Cat ID: 40048	Catalase [*Drosophila melanogaster* (fruit fly)]	Chromosome 3L, NT_037436.4 (18822604..18828188)	Dmel_CG6871, CAT, CATA, CG6871, CT21282A, DMCATHPO, DROCATHPO, Dmel\CG6871, U00145, bs36h11.y1, cat, catalase, Cat	
CAT ID: 101093891	catalase [*Felis catus* (domestic cat)]	Chromosome D1, NC_018732.2 (88501679..88538456)		
CAT ID: 531682	catalase [*Bos taurus* (cattle)]	Chromosome 15, AC_000172.1		


Dict.
Terms
Dict.
Affixes
Training
OntoGene
Distiller
Test
Full-text articles
Annotations
Annotations

CRAFT corpus

Corpus

- 67 full-text articles (PubMed Central)
- 100 k manually verified concept annotations, completed in 2012
- refers to 7 ontologies/terminologies:
  ChEBI, CL, Entrez Gene, GO, NCBI taxonomy, PRO, SO

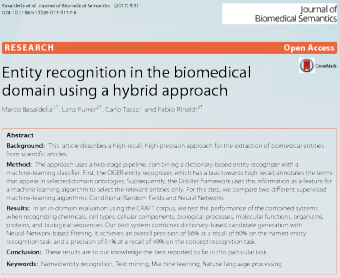
Usage

CRAFT is being used to date[1,2] for evaluating concept recognition systems.

---

[1] Christopher Funk et al. (2014). "Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters". In: *BMC Bioinformatics* 15.1, pp. 1–29

[2] Eugene Tseytlin et al. (2016). "NOBLE – Flexible concept recognition for large-scale



Rinaldi et al. Journal of Biomedical Semantics 2010, 1:5
http://www.jbiomedsem.com/content/1/1/5

Journal of Biomedical Semantics

RESEARCH Open Access

Entity recognition in the biomedical domain using a hybrid approach

Marco Basaldella[1†], Lenz Furrer[2†], Carlo Tasso[1] and Fabio Rinaldi[2*]

Abstract

Background: This article describes a high-recall, high-precision approach for the extraction of biomedical entities from scientific articles.

Method: The approach uses a two-stage pipeline, combining a dictionary-based entity recognizer with a machine-learning classifier. First, the OGER entity recognizer, which has a bias towards high recall, annotates the terms that appear in selected domain ontologies. Subsequently, the Distiller framework uses this information as a feature for a machine learning algorithm to select the relevant entities only. For this step, we compare two different supervised machine-learning algorithms: Conditional Random Fields and Neural Networks.

Results: In an in-domain evaluation using the CRAFT corpus, we test the performance of the combined systems when recognizing chemicals, cell types, cellular components, biological processes, molecular functions, organisms, proteins, and biological sequences. Our best system combines dictionary-based candidate generation with Neural Network-based filtering. It achieves an overall precision of 86% at a recall of 60% on the named entity recognition task, and a precision of 51% at a recall of 49% on the concept recognition task.

Conclusion: These results are to our knowledge the best reported so far in this particular task.

Keywords: Named entity recognition, Text mining, Machine learning, Natural language processing

Background

The scientific community in the biomedical domain is a vibrant community, producing a large amount of scientific findings in the form of data, publications, reports, and so on, each year, making it difficult for scholars to find the right information in the large sea of knowledge.

To tackle that problem, researchers have developed different text mining techniques with the goal of detecting the relevant information for the intended purpose. This paper's focus is the technique called Named Entity Recognition (herein NER), which solves the problem of detecting terms belonging to a limited set of predefined entity types.

NER can be performed on both "generic" documents, to recognize concepts like *person*, *date* or *location*, or on technical documents, to recognize concepts like *cells*, *diseases* or *proteins*. NER can be used by itself, with the goal of recognizing the main *processes* of a text in an extract portion of the text, or as a preliminary stage for Concept Recognition (CR), also known as Entity Linking or Normalization, where a term is not only recognized but also linked to a terminological resource, such as an ontology, through the use of a unique identifier [1].

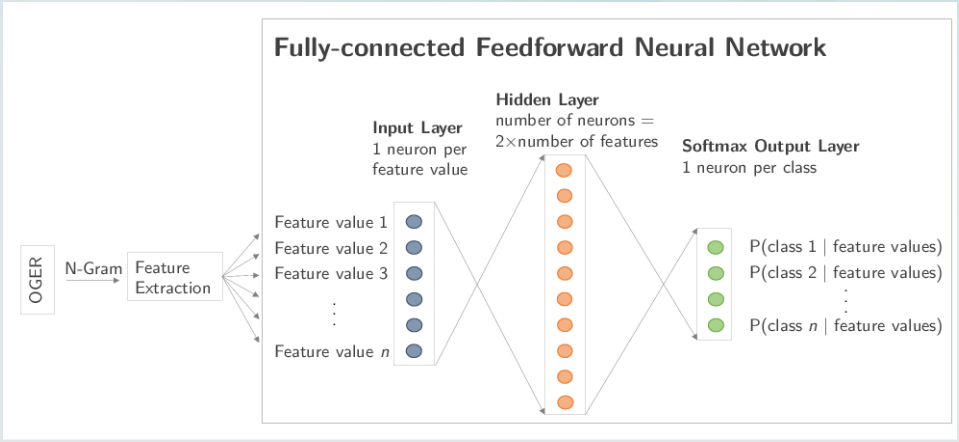
NER can be solved using several techniques:

- Using manual, hand-written rules. A group of experts develops these rules using domain knowledge. The rules typically rely on orthographic patterns, such as particular use of capitalization or punctuation. Even though rule-based systems can perform well if sufficient expert time is available for creating the rules, their maintenance requires repeated manual efforts, since the rules need to be revised or even entirely replaced whenever the system is adapted to new data (different entity types, another text collection). In the biomedical domain, plain rule-based approaches like [2] have become rare, however

*Correspondence: rinaldi@ifi.uzh.ch
†Equal contributors
Full list of author information is available at the end of the article

	Neural Network	Conditional Random Fields
Implementation		
Software	R [47], `nnet` library	CRFSuite [48]
Model parameters	1 hidden layer of size $2 \times (number\ of\ input\ features)$, softmax output layer	Training algorithm: averaged perceptron, default epsilon, 2 words window
Input	n-grams selected by OGER	single tokens
Features		
Candidate character count	count	—
Candidate is all uppercase	label yes/no	label yes/no
Candidate is all lowercase	label yes/no	label yes/no
Candidate contains Greek (i.e. "alpha", α)	label yes/no	label yes/no
Candidate contains dashes ('-')	count	label yes/no
Candidate contains numbers	count	label yes/no
Candidate ends with a number	label yes/no	label yes/no
Candidate contains capital letter not in first position	label yes/no	label yes/no
Candidate contains lowercase characters	count	label yes/no
Candidate contains uppercase characters	count	label yes/no
Candidate contains spaces	count	label yes/no
Candidate contains symbols	count	label yes/no
2-3 character affixes appearing in an ontology in [36]	normalized frequency	label yes/no
Candidate is symbol	—	label yes/no
Candidate's part-of-speech	—	yes, using [49]
Candidate's stem	—	yes, using [50]
Candidate pre-selected by OGER	—	yes (see Section 2.5)
Total features	36	about 2.8 million
Tagging speed	1286 tokens/sec	632 tokens/sec

New JointNN filter



Results

Table 4: F1-scores

Entity Type	OGER	OGER+DistillerNN	OGER+JointNN
All	0.51	0.70	**0.71**
Chemicals	0.55	0.77	**0.79**
Cells	0.80	0.76	0.72
Biological processes/molecular functions	0.30	0.35	**0.45**
Cellular components	0.55	0.70	**0.73**
Organisms	0.44	0.94	**0.95**
Proteins	0.62	0.80	0.79
Sequences	0.54	0.75	0.73

[Anna Jancso, Lenz Furrer, Fabio Rinaldi, in preparation]

Conclusions

- Biomedical Text Mining is indispensable to make sense of huge amounts of text: e.g. scientific literature, clinical records, social media
- The complexity of human language makes it difficult for machines to extract meaning out of text
- The biomedical field has been at the forefront of research in text mining, with numerous competitions and shared resources
- Recent advances in NLP, based on deep learning, have generated a quantum leap in our ability to extract information from text

Topic

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000).
Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
Nat. Genet., 25:25–29.

Chang, J. T., Schtze, H., and Altman, R. B. (2002).
Creating an online dictionary of abbreviations from MEDLINE.
J Am Med Inform Assoc, 9(6):612–20.
PMID:12386112.

Chen, H. and Sharp, B. M. (2004).
Content-rich biological network constructed by mining PubMed abstracts.
BMC Bioinformatics, 5:147.
PMID:15473905.