

Biomedical Text Mining

Tianxiao Li

cbb752

Structured vs unstructured data (free text)

[▲] node1	node2	node1 accession	node2 accession	score
ATM	BARD1	ENSP00000278616	ENSP00000260947	0.975
ATM	BRCA1	ENSP00000278616	ENSP00000418960	0.995
ATM	BRCA2	ENSP00000278616	ENSP00000369497	0.994
ATM	FAM175A	ENSP00000278616	ENSP00000369857	0.950
ATM	FANCD2	ENSP00000278616	ENSP00000287647	0.998
ATM	PALB2	ENSP00000278616	ENSP00000261584	0.979
ATM	RAD51	ENSP00000278616	ENSP00000372088	0.995
ATM	TOPBP1	ENSP00000278616	ENSP00000260810	0.996
ATM	TP53	ENSP00000278616	ENSP00000269305	0.999
ATM	UIMC1	ENSP00000278616	ENSP00000366434	0.978
BARD1	ATM	ENSP00000260947	ENSP00000278616	0.975
BARD1	BRCA1	ENSP00000260947	ENSP00000418960	0.998
BARD1	BRCA2	ENSP00000260947	ENSP00000369497	0.994
BARD1	FAM175A	ENSP00000260947	ENSP00000369857	0.999
BARD1	FANCD2	ENSP00000260947	ENSP00000287647	0.965
BARD1	PALB2	ENSP00000260947	ENSP00000261584	0.973
BARD1	RAD51	ENSP00000260947	ENSP00000372088	0.996
BARD1	TOPBP1	ENSP00000260947	ENSP00000260810	0.993
BARD1	TP53	ENSP00000260947	ENSP00000269305	0.987
BARD1	UIMC1	ENSP00000260947	ENSP00000366434	0.982

Although further work is required to fully understand each steps of the FA pathway, some of the downstream players involved are: FANCD1 (BRIP1), DNA-dependent ATPase and 5'-3' DNA helicase able to interact with **BRCA1**; FANCD1 (BRCA2), able to bind ssDNA and dsDNA and to stimulate RAD51 action; FANCD1 (partner and localizer of BRCA2, PALB2), required for FANCD1 stabilization and for the recruitment of BRCA2 and RAD51; and FANCO (RAD51C) involved in HJ resolution ([Kottemann and Smogorzewska, 2013](#)).

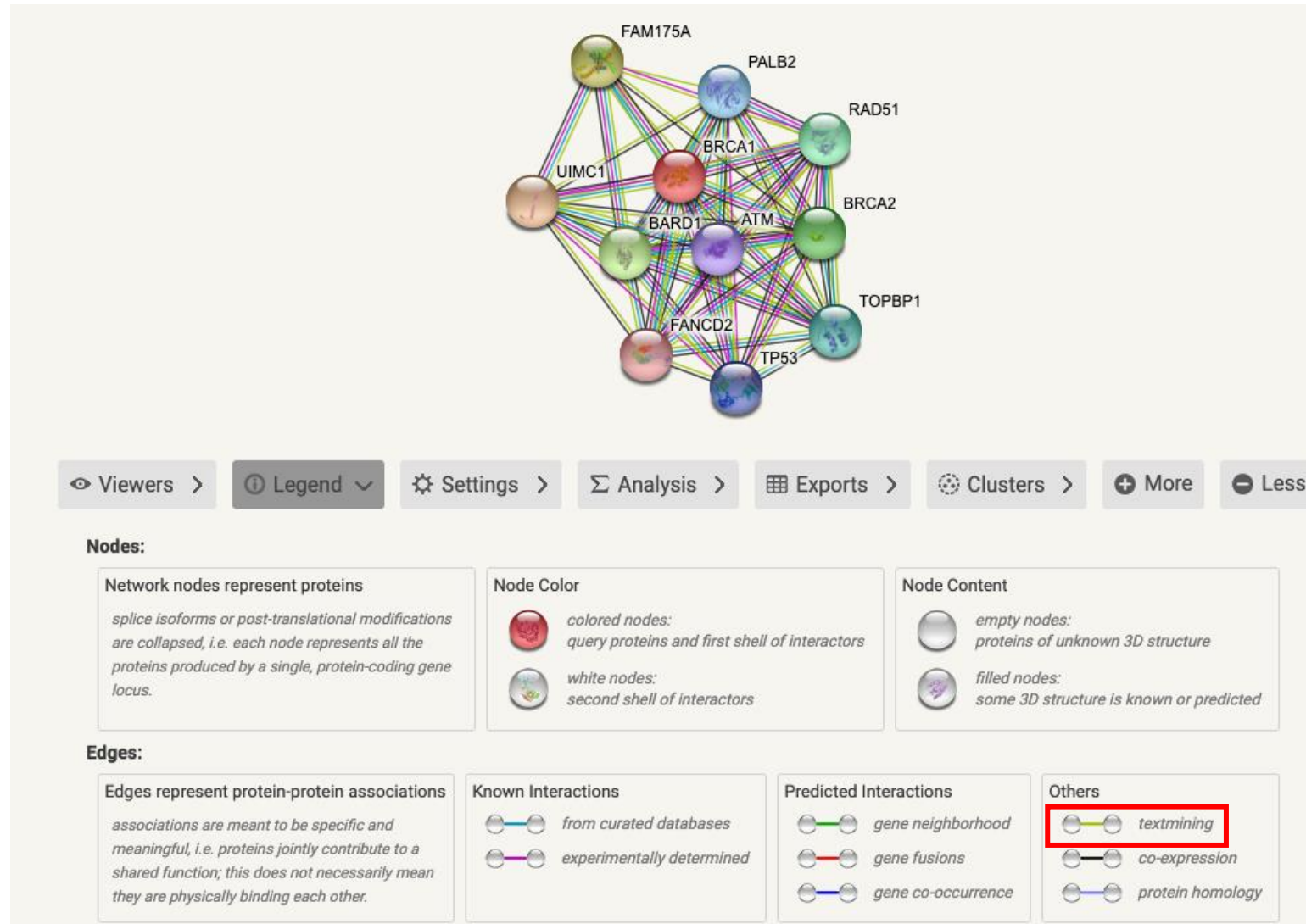
Breast cancer susceptibility gene 1 is a component of three different multiprotein complexes involved in all cell cycle checkpoints: the **BRCA1A** complex (composed of Abraxas, BARD1, RAP80, BRCC36, BRCC45, and MERIT40), responsible to recruit **BRCA1** to damaged sites; the **BRCA1B** complex (formed with BRIP1 and TOPBP1), mainly associated with replication-coupled DNA repair and the **BRCA1C** complex (formed together with CtIP and the MRN complex), which promotes HR despite NHEJ ([Huen et al., 2010](#)).

<https://string-db.org/>

Torgovnick, Alessandro, and Björn Schumacher. "DNA repair mechanisms in cancer development and therapy." *Frontiers in genetics* 6 (2015): 157.

PART I: Supervised Mining

Mining of Biomedical Literature



Relevant publications mentioning your query species (Homo sapiens):

PMID:25954303: DNA repair mechanisms in cancer development and therapy.

▼ *Torgovnick A, Schumacher B*
Front Genet. 6:157 2015.



Abstract:

DNA damage has been long recognized as causal factor for cancer development. When erroneous DNA repair leads to mutations or chromosomal aberrations affecting oncogenes and tumor suppressor genes, cells undergo malignant transformation resulting in cancerous growth. Genetic defects can predispose to cancer: mutations in distinct DNA repair systems elevate the susceptibility to various cancer types. However, DNA damage not only comprises a root cause for cancer development but also continues to provide an important avenue for chemo- and radiotherapy. Since the beginning of cancer therapy, genotoxic agents that trigger DNA damage checkpoints have been applied to halt the growth and trigger the apoptotic demise of cancer cells. We provide an overview about the involvement of DNA repair systems in cancer prevention and the classes of genotoxins that are commonly used for the treatment of cancer. A better understanding of the roles and interactions of the highly complex DNA repair machineries will lead to important improvements in cancer therapy.

Excerpts from full text:

... FAN1 (FA-associated nuclease 1) is another nuclease recruited to the damaged site by ubiquitinated **FANCD2** (●). FAN1 abrogation does not affect ICLs-induced DSBs formation most likely [...] involved are: FANCD1 (BRIP1), DNA-dependent ATPase and 5'-3' DNA helicase able to interact with **BRCA1** (●); FANCD1 (**BRCA2** (●)), able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; FANCD1 [...] able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; FANCD1 (**partner and localizer of BRCA2** (●), **PALB2** (●)), required for FANCD1 stabilization and for the recruitment of **BRCA2** (●) and **RAD51** (●); and [...] components (Adamo et al., 2010; Pace et al., 2010). Breast Cancer Susceptibility Gene 1 and 2 (**BRCA1** (●) and 2) Double strand breaks are the most threatening forms of DNA damage, if left [...] complexes involved in all cell cycle checkpoints: the BRCA1A complex (composed of Abraxas, **BARD1** (●), **RAP80** (●), BRCC36, BRCC45, and MERIT40), responsible to recruit **BRCA1** (●) to damaged [...] MERIT40), responsible to recruit **BRCA1** (●) to damaged sites; the BRCA1B complex (formed with BRIP1 and **TOPBP1** (●)), mainly associated with replication-coupled DNA repair and the BRCA1C complex [...] with CtIP and the MRN complex), which promotes HR despite NHEJ (Huen et al., 2010). Interestingly, **BRCA1** (●)/BARD mutations cannot only fuel genome instability due to impaired HR activity, but [...] et al., 2014). The genome instability in smc-5/6 mutants could be reversed upon inactivation of the **BRCA1** (●)/BARD complex. It is tempting to speculate that mutations in **BRCA1** (●) might be [...] conditions of replication fork breakdown prevention of HR could benefit genome stability. The **BRCA2** (●) protein was recently purified and functionally validated by three independent [...] repair pathways. If the damage is too severe, programmed cell death will be induced through the **ATM** (●)/ATR/**TP53** (●) pathway (Damia et al., 2000; Pabla et al., 2008). Although cisplatin is a ...

Relevant publications mentioning your query species (Homo sapiens):

PMID:25954303: DNA repair mechanisms in cancer development and therapy.

▼ Torgovnick A, Schumacher B
Front Genet. 6:157 2015.

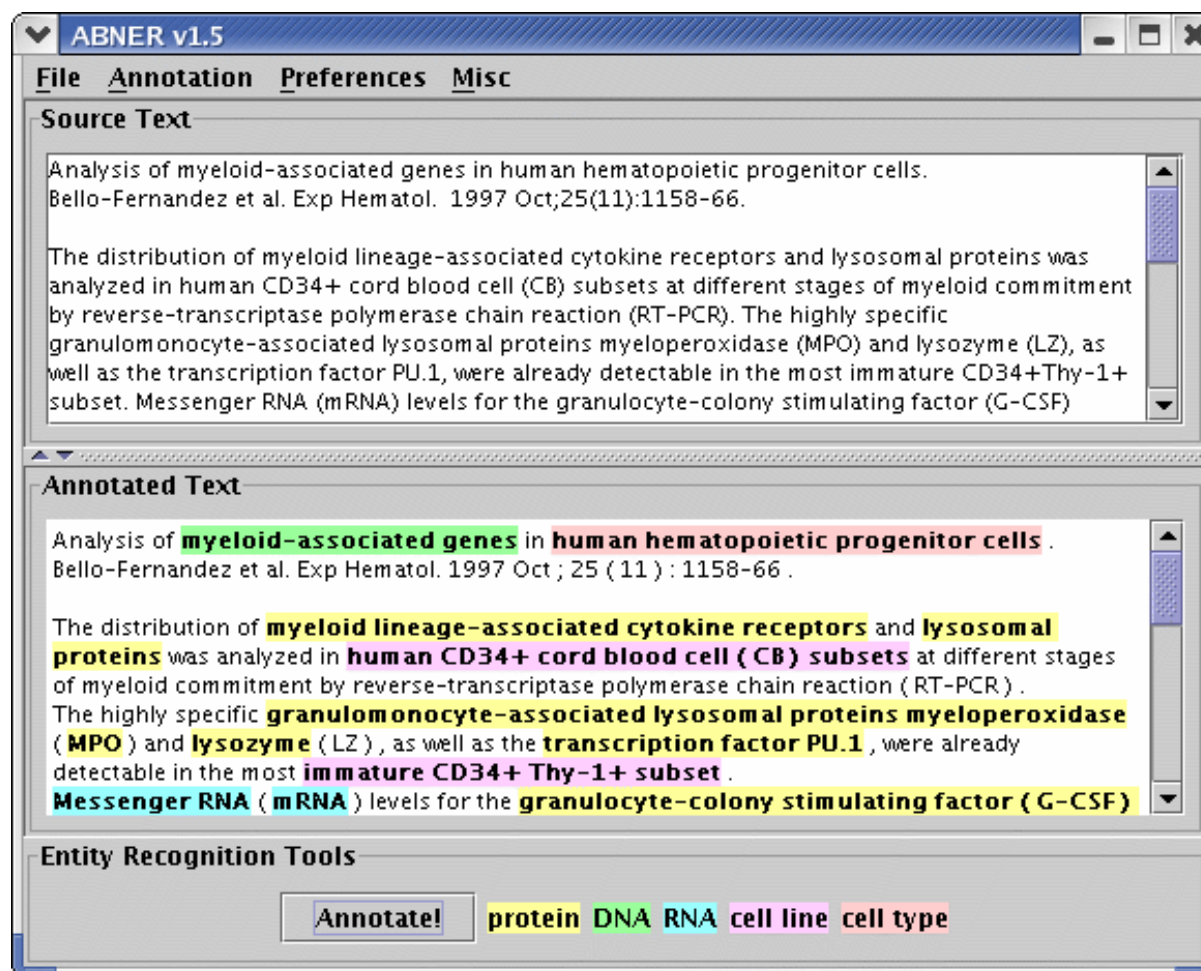


Abstract:

DNA damage has been long recognized as causal factor for cancer development. When erroneous DNA repair leads to mutations or chromosomal aberrations affecting oncogenes and tumor suppressor genes, cells undergo malignant transformation resulting in cancerous growth. Genetic defects can predispose to cancer: mutations in distinct DNA repair systems elevate the susceptibility to various cancer types. However, DNA damage not only comprises a root cause for cancer development but also continues to provide an important avenue for chemo- and radiotherapy. Since the beginning of cancer therapy, genotoxic agents that trigger DNA damage checkpoints have been applied to halt the growth and trigger the apoptotic demise of cancer cells. We provide an overview about the involvement of DNA repair systems in cancer prevention and the classes of genotoxins that are commonly used for the treatment of cancer. A better understanding of the roles and interactions of the highly complex DNA repair machineries will lead to important improvements in cancer therapy.

Excerpts from full text:

... FAN1 (FA-associated nuclease 1) is another nuclease recruited to the damaged site by ubiquitinated **FANCD2** (●). FAN1 abrogation does not affect ICLs-induced DSBs formation most likely [...] involved are: FANCD1 (BRIP1), DNA-dependent ATPase and 5'-3' DNA helicase able to interact with **BRCA1** (●); **FANCD1** (**BRCA2** (●)), able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; FANCD1 [...] able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; **FANCD1** (**partner and localizer of BRCA2** (●), **PALB2** (●)), required for FANCD1 stabilization and for the recruitment of **BRCA2** (●) and **RAD51** (●); and [...] components (Adamo et al., 2010; Pace et al., 2010). Breast Cancer Susceptibility Gene 1 and 2 (**BRCA1** (●) and 2) Double strand breaks are the most threatening forms of DNA damage, if left [...] complexes involved in all cell cycle checkpoints: the BRCA1A complex (composed of Abraxas, **BARD1** (●), **RAP80** (●), BRCC36, BRCC45, and MERIT40), responsible to recruit **BRCA1** (●) to damaged [...] MERIT40), responsible to recruit **BRCA1** (●) to damaged sites; the BRCA1B complex (formed with BRIP1 and **TOPBP1** (●)), mainly associated with replication-coupled DNA repair and the BRCA1C complex [...] with CtIP and the MRN complex), which promotes HR despite NHEJ (Huen et al., 2010). Interestingly, **BRCA1** (●)/BARD mutations cannot only fuel genome instability due to impaired HR activity, but [...] et al., 2014). The genome instability in smc-5/6 mutants could be reversed upon inactivation of the **BRCA1** (●)/BARD complex. It is tempting to speculate that mutations in **BRCA1** (●) might be [...] conditions of replication fork breakdown prevention of HR could benefit genome stability. The **BRCA2** (●) protein was recently purified and functionally validated by three independent [...] repair pathways. If the damage is too severe, programmed cell death will be induced through the **ATM** (●)/ATR/**TP53** (●) pathway (Damia et al., 2000; Pabla et al., 2008). Although cisplatin is a ...



<http://pages.cs.wisc.edu/~bsettles/abner/>

Named Entity Recognition

- Is it (part of) a named entity?
- Boundary?
- Class?

1601297|t|Electrocardiographic evidence of **myocardial injury** in psychiatrically hospitalized **cocaine** abusers.

1601297|a|The electrocardiograms (ECG) of 99 **cocaine**-abusing patients were compared with the ECGs of 50 **schizophrenic** controls. Eleven of the **cocaine** abusers and none of the controls had ECG evidence of significant **myocardial injury** defined as **myocardial infarction**, **ischemia**, and **bundle branch block**.

1601297	33	50	myocardial injury	Disease	D009202
1601297	83	90	cocaine	Chemical	D003042
1601297	135	142	cocaine	Chemical	D003042
1601297	194	207	schizophrenic	Disease	D012559
1601297	232	239	cocaine	Chemical	D003042
1601297	305	322	myocardial injury	Disease	D009202
1601297	334	355	myocardial infarction	Disease	D009203
1601297	357	365	ischemia	Disease	D007511
1601297	371	390	bundle branch block	Disease	D002037

Nguyen, Dat Quoc, and Karin Verspoor. "Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings." *arXiv preprint arXiv:1805.10586* (2018).

Wei, Chih-Hsuan, et al. "Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task." *Database* 2016 (2016).

Named Entity Recognition

- Challenges in biomedical texts
 - Large, and increasing, numbers of new technical terms
 - Names are long (“12-o-tetradecanoylphorbol 13-acetate”)
 - Abbreviations (“TCF: “T cell factor” or “Tissue Culture Fluid”)
 - Combinations

WT1 missense mutations, associated with **male pseudohermaphroditism** in **Denys-Drash syndrome**, fail to ...

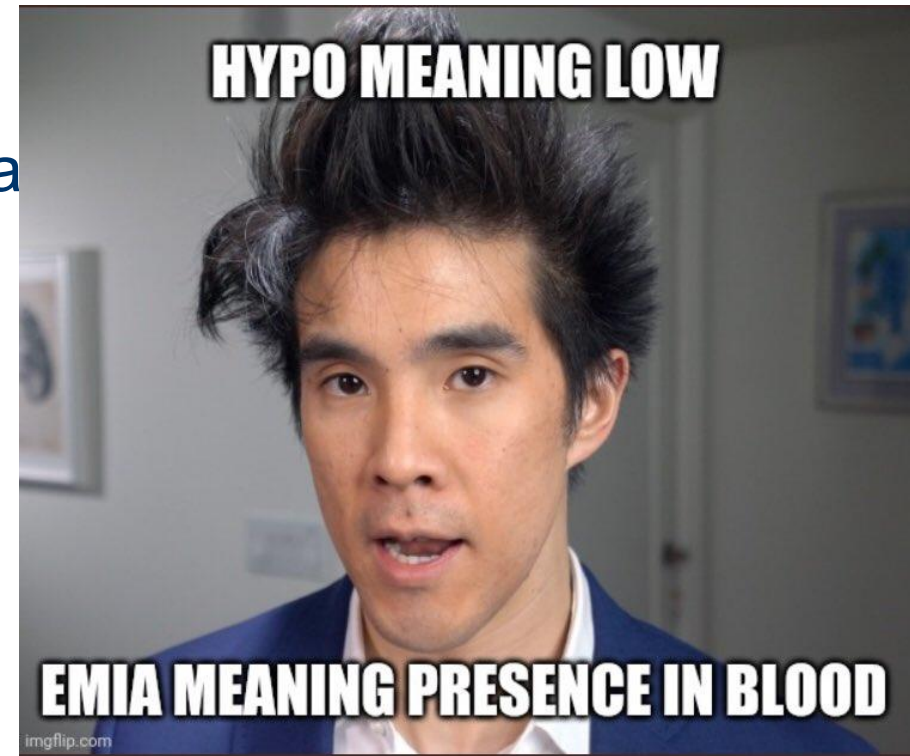
... a case of oral **penicillin anaphylaxis** is described, and the terminology ...

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

Song, Hye-Jeong, et al. "Comparison of named entity recognition methodologies in biomedical documents." *Biomedical engineering online* 17.2 (2018): 1-14.

Named Entity Recognition

- Dictionary-based
- Rule-based
 - “Mr. [X]” -> X is a named entity (a person’s name)
 - Upper cases, hyphens and numbers, etc.
 - “**Methylphenidate**” -> likely a chemical
 - “**Hypoxemia**” -> likely a disease



Named Entity Recognition

- Supervised learning
 - Need hand-curated training data

1601297|t|Electrocardiographic evidence of myocardial injury in psychiatrically

Features

The word itself

- Capitalized?
- Contains non-alphabetic characters?
- ...

Parse features

- Part-of-speech tag?
- Part of a lexical/grammatical rule?
- Which part?
- ...

Contextual features

...



Model

Label

Beginning/End/Inside/Outside
of a [CLASS] named entity?

Finkel, Jenny Rose, Trond Grenager, and Christopher D. Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*. 2005.

Finkel, Jenny Rose, Alex Kleeman, and Christopher D. Manning. "Efficient, feature-based, conditional random field parsing." *Proceedings of ACL-08: HLT*. 2008.

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

Relevant publications mentioning your query species (Homo sapiens):

PMID:25954303: DNA repair mechanisms in cancer development and therapy.

▼ Torgovnick A, Schumacher B
Front Genet. 6:157 2015.





Abstract:

DNA damage has been long recognized as causal factor for cancer development. When erroneous DNA repair leads to mutations or chromosomal aberrations affecting oncogenes and tumor suppressor genes, cells undergo malignant transformation resulting in cancerous growth. Genetic defects can predispose to cancer: mutations in distinct DNA repair systems elevate the susceptibility to various cancer types. However, DNA damage not only comprises a root cause for cancer development but also continues to provide an important avenue for chemo- and radiotherapy. Since the beginning of cancer therapy, genotoxic agents that trigger DNA damage checkpoints have been applied to halt the growth and trigger the apoptotic demise of cancer cells. We provide an overview about the involvement of DNA repair systems in cancer prevention and the classes of genotoxins that are commonly used for the treatment of cancer. A better understanding of the roles and interactions of the highly complex DNA repair machineries will lead to important improvements in cancer therapy.

Excerpts from full text:

... FAN1 (FA-associated nuclease 1) is another nuclease recruited to the damaged site by ubiquitinated **FANCD2** (●). FAN1 abrogation does not affect ICLs-induced DSBs formation most likely [...] involved are: FANCD1 (BRIP1), DNA-dependent ATPase and 5'-3' DNA helicase able to interact with **BRCA1** (●); FANCD1 (**BRCA2** (●)), able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; FANCD1 [...] able to bind ssDNA and dsDNA and to stimulate **RAD51** (●) action; FANCD1 (**partner and localizer of BRCA2** (●), **PALB2** (●)), required for FANCD1 stabilization and for the recruitment of **BRCA2** (●) and **RAD51** (●); and [...] components (Adamo et al., 2010; Pace et al., 2010). Breast Cancer Susceptibility Gene 1 and 2 (**BRCA1** (●) and 2) Double strand breaks are the most threatening forms of DNA damage, if left [...] complexes involved in all cell cycle checkpoints: the BRCA1A complex (composed of Abraxas, **BARD1** (●), **RAP80** (●), BRCC36, BRCC45, and MERIT40), responsible to recruit **BRCA1** (●) to damaged [...] MERIT40), responsible to recruit **BRCA1** (●) to damaged sites; the BRCA1B complex (formed with BRIP1 and **TOPBP1** (●)), mainly associated with replication-coupled DNA repair and the BRCA1C complex [...] with CtIP and the MRN complex), which promotes HR despite NHEJ (Huen et al., 2010). Interestingly, **BRCA1** (●)/BARD mutations cannot only fuel genome instability due to impaired HR activity, but [...] et al., 2014). The genome instability in smc-5/6 mutants could be reversed upon inactivation of the **BRCA1** (●)/BARD complex. It is tempting to speculate that mutations in **BRCA1** (●) might be [...] conditions of replication fork breakdown prevention of HR could benefit genome stability. The **BRCA2** (●) protein was recently purified and functionally validated by three independent [...] repair pathways. If the damage is too severe, programmed cell death will be induced through the **ATM** (●)/ATR/**TP53** (●) pathway (Damia et al., 2000; Pabla et al., 2008). Although cisplatin is a ...

Relation Extraction

- Co-occurrence
 - “[A] interacts with [B]” 
 - “[A] does not interact with [B]” 
- Rule-based
 - “[A] **[relational verb phrase]** [B]”
 - “[A] **does not [relational verb phrase]** [B]”

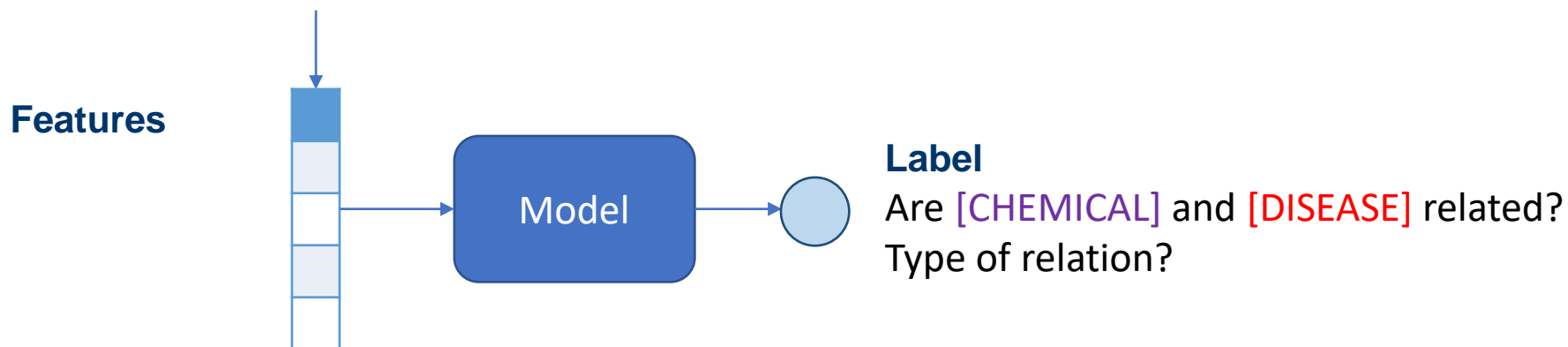
Relation Extraction

- Supervised learning

Eleven of the cocaine abusers and none of the controls had ECG evidence of significant myocardial injury defined as ...

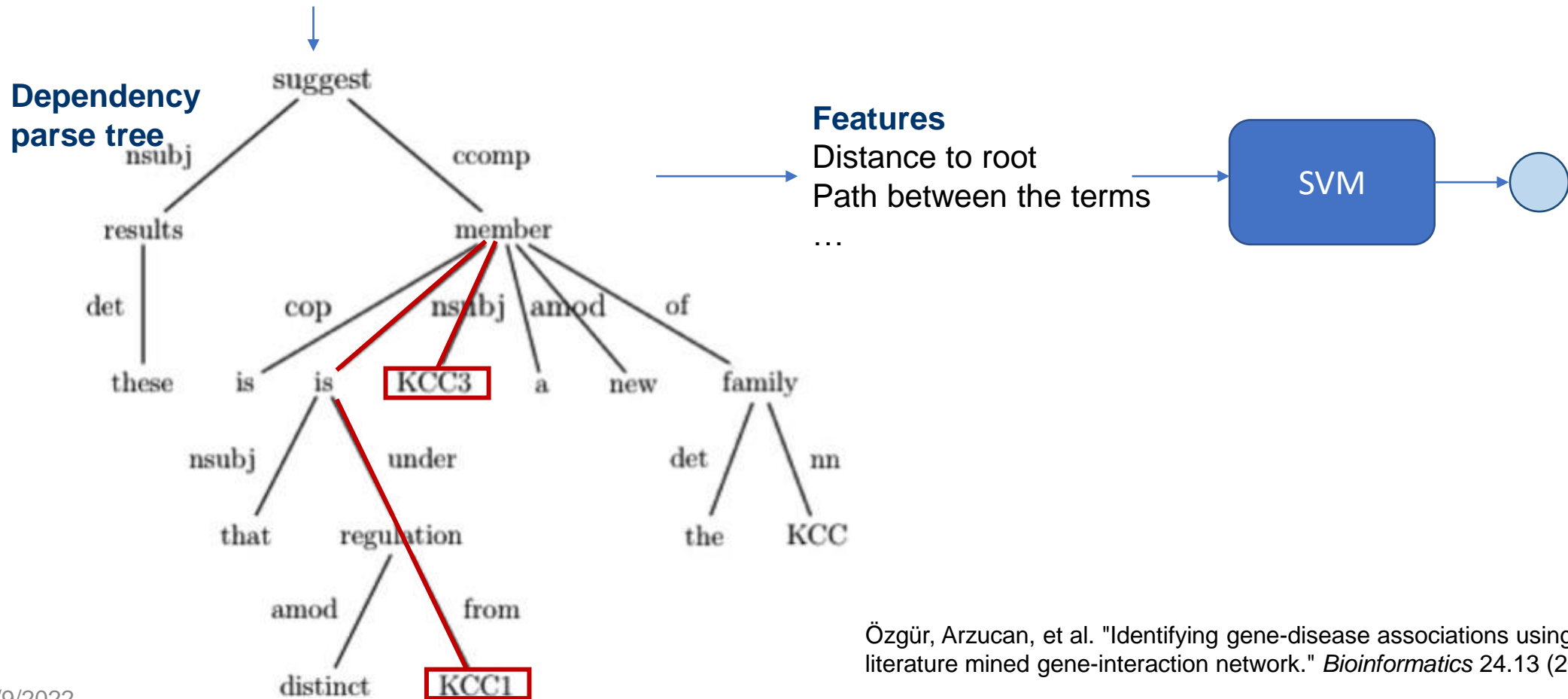


Eleven of the [CHEMICAL] abusers and none of the controls had ECG evidence of significant [DISEASE] defined as ...

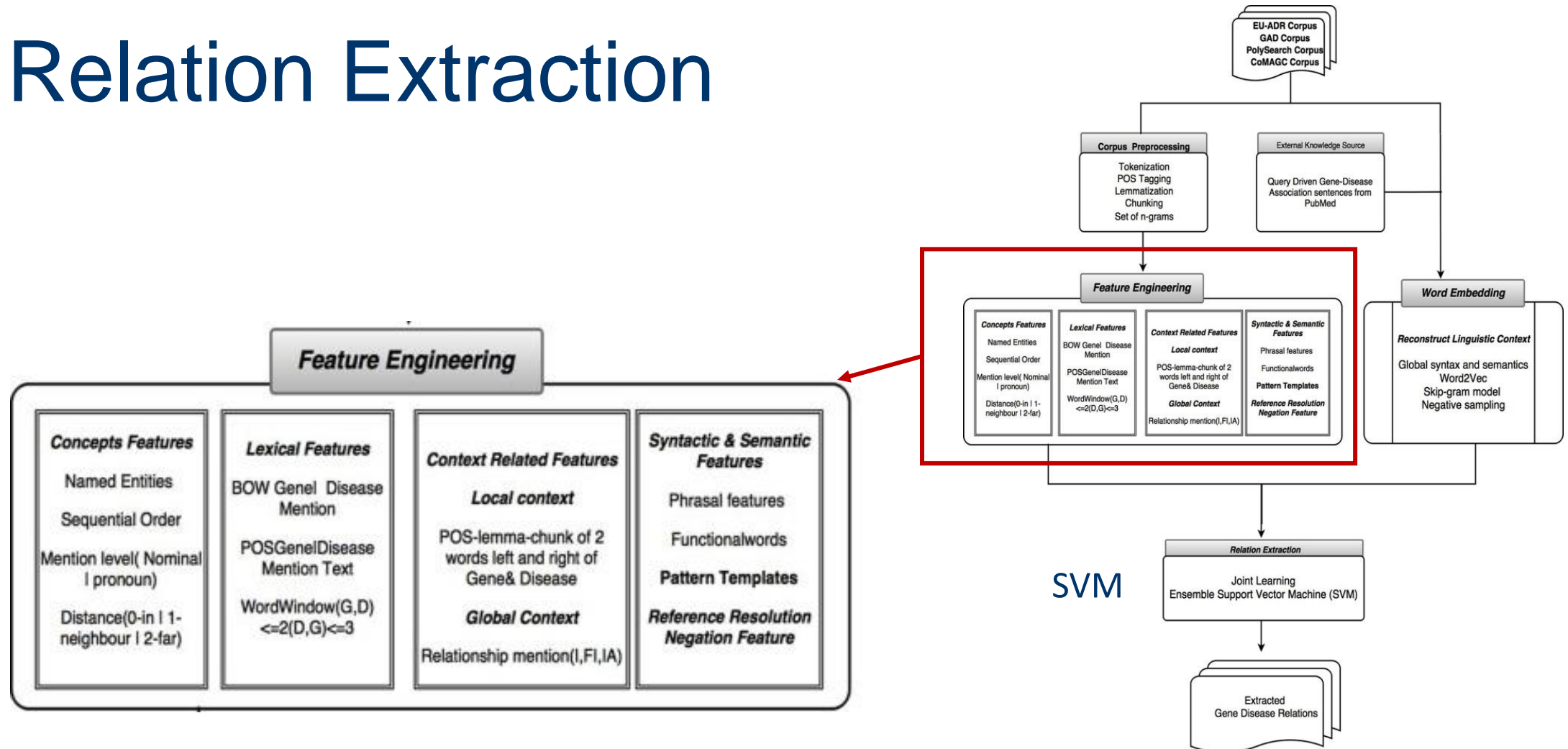


Relation Extraction

*These results suggest **KCC3** is a new member of the KCC family that is under distinct regulation from **KCC1***



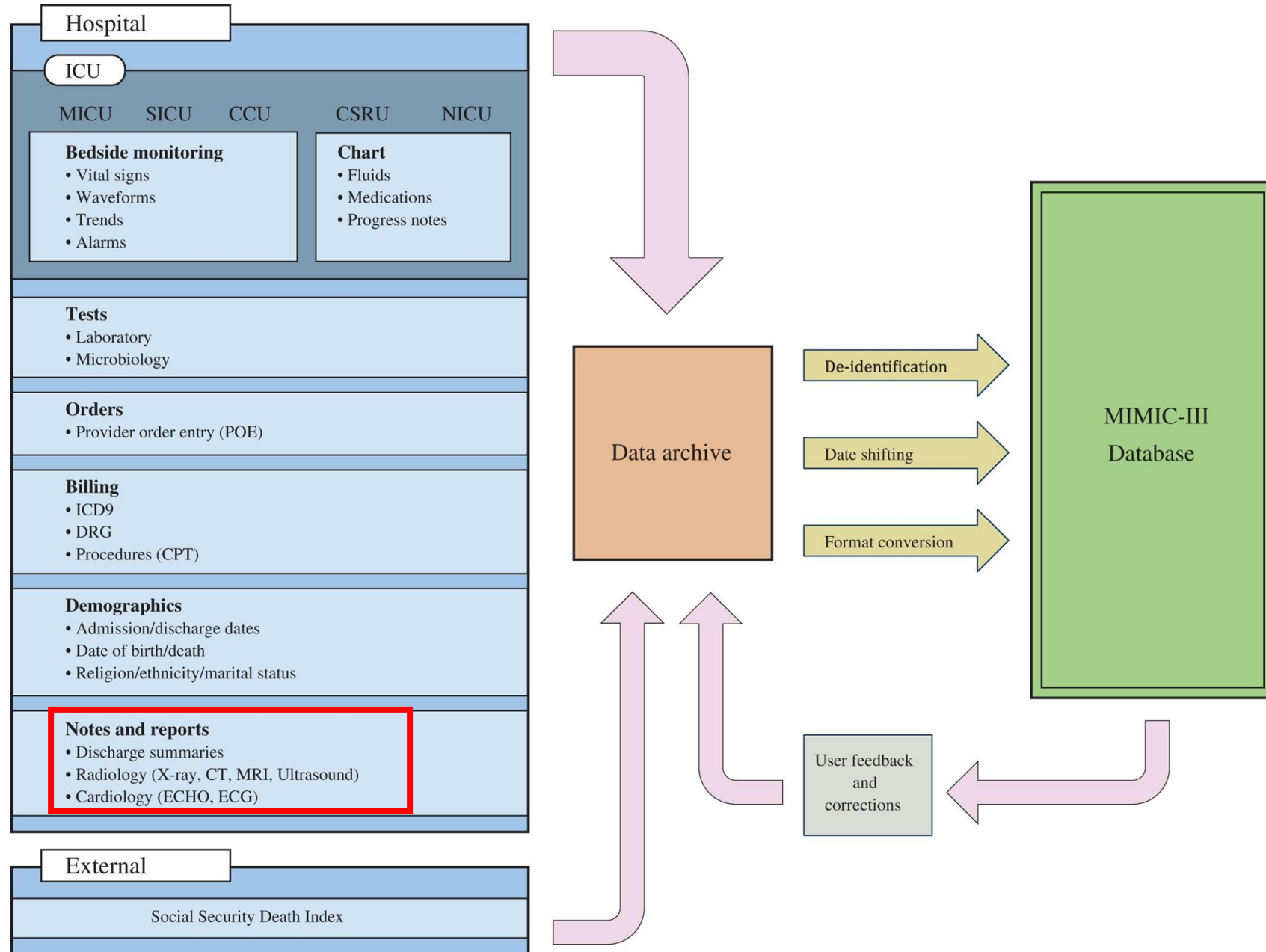
Relation Extraction



Bhasuran, Balu, and Jeyakumar Natarajan. "Automatic extraction of gene-disease associations from literature using joint ensemble learning." *PloS one* 13.7 (2018): e0200699.

Mining of Clinical Records

MIMIC-III



PRINCIPAL DIAGNOSIS: 1. OSTEOARTHRITIS OF THE LEFT HIP
SECONDARY DIAGNOSIS: 2. WOLFF-PARKINSON-WHITE SYNDROME

PROCEDURES: Left total hip replacement (uncemented), 2-2-93.

HISTORY OF PRESENT ILLNESS: The patient is a 54 year old white male with a 9 month history of left hip pain. He has noted a severe limitation of ambulation over this period of time and presently is limited to non reciprocal stairs and short distances. He has trouble getting out of a chair as well as a car. The examination and radiographs ... confirmed bilateral hip osteoarthritis with left greater than right. He is admitted for an elective left total hip replacement. He has donated three units of autologous blood.

PAST MEDICAL HISTORY: Notable for osteoarthritis as noted above and WPW syndrome.
PAST SURGICAL HISTORY: Notable for tonsillectomy at age 3 and bilateral hammer toe corrections. MEDICATIONS ON ADMISSION: At the time of admission, the patient was on Ferrous Sulfate 325 mg po t.i.d. ALLERGIES: NK DA.

PHYSICAL EXAMINATION: HEENT examination was within normal limits. The lungs were clear. The cardiac examination revealed no murmurs. The abdomen was benign. The extremity examination revealed a left antalgic gait with no lurch. There was negative bilateral Trendelenburg sign. His range of motion of both hips are as follows: flexion is 90 bilaterally and extension was -10 degrees bilaterally. He had abduction to only 5 degrees bilaterally and adduction of 30 degrees bilaterally. His external rotation was 5 degrees and internal rotation was 0 degrees bilaterally. His knees and ankles had full range of motion. Distal sensory motor examination was intact. Distal pulses were intact.

LABORATORY DATA: The patient's admission hematocrit was 38.1. Electrolytes were within normal limits. Coagulation factors were normal. Sed rate was 11.

HOSPITAL COURSE: The patient underwent a left total hip replacement on 2-2-93. Post-operatively, he was transferred to the floor in stable condition. His hematocrit immediately postoperative was 38 and trended down to a hematocrit of 34. His postoperative course was notable for quick progression in physical therapy and he was discharged on 2-9-93. He was anticoagulated in routine fashion postoperatively and at discharge his PT was 13.8 with iron of 1.6. Vascular ultrasound and x-rays were taken prior to discharge and the results were not available at the time of this dictation. He was to continue on 6 weeks of coumadinization and follow up with Dr. ... at that time.

MEDICATIONS ON DISCHARGE: At the time of discharge, the patient was on Percocet 1-2 q 3 prn, Coumadin 5 mg po q d until directed otherwise. DISPOSITION: To home.

ICD9 Codes for this discharge summary:

D715.95 Osteoarthritis, unspecified whether generalized or localized, involving pelvic region and thigh

D426.7 Anomalous atrioventricular excitation

Principle diagnosis code (ICD9)

Free text (discharge summary)

Larkey, Leah S., and W. Bruce Croft. *Automatic assignment of icd9 codes to discharge summaries*. Technical report, University of Massachusetts at Amherst, Amherst, MA, 1995.

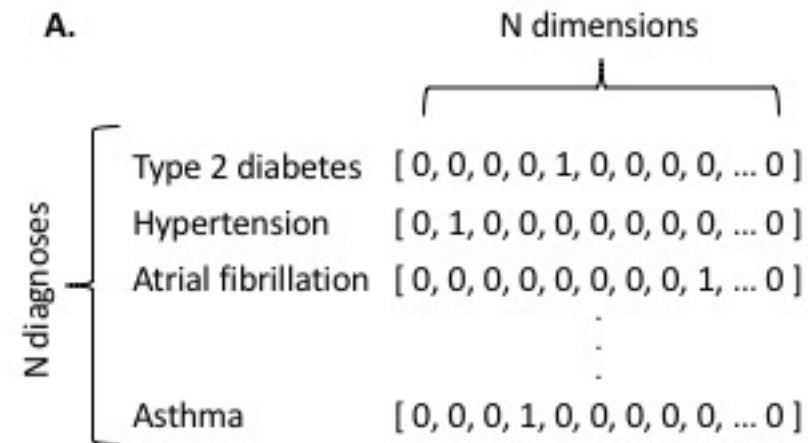
PART II:

Unsupervised Mining

Word Embedding

Representation of Words

- Hand-curated features
- One-hot encoding



Word Embedding

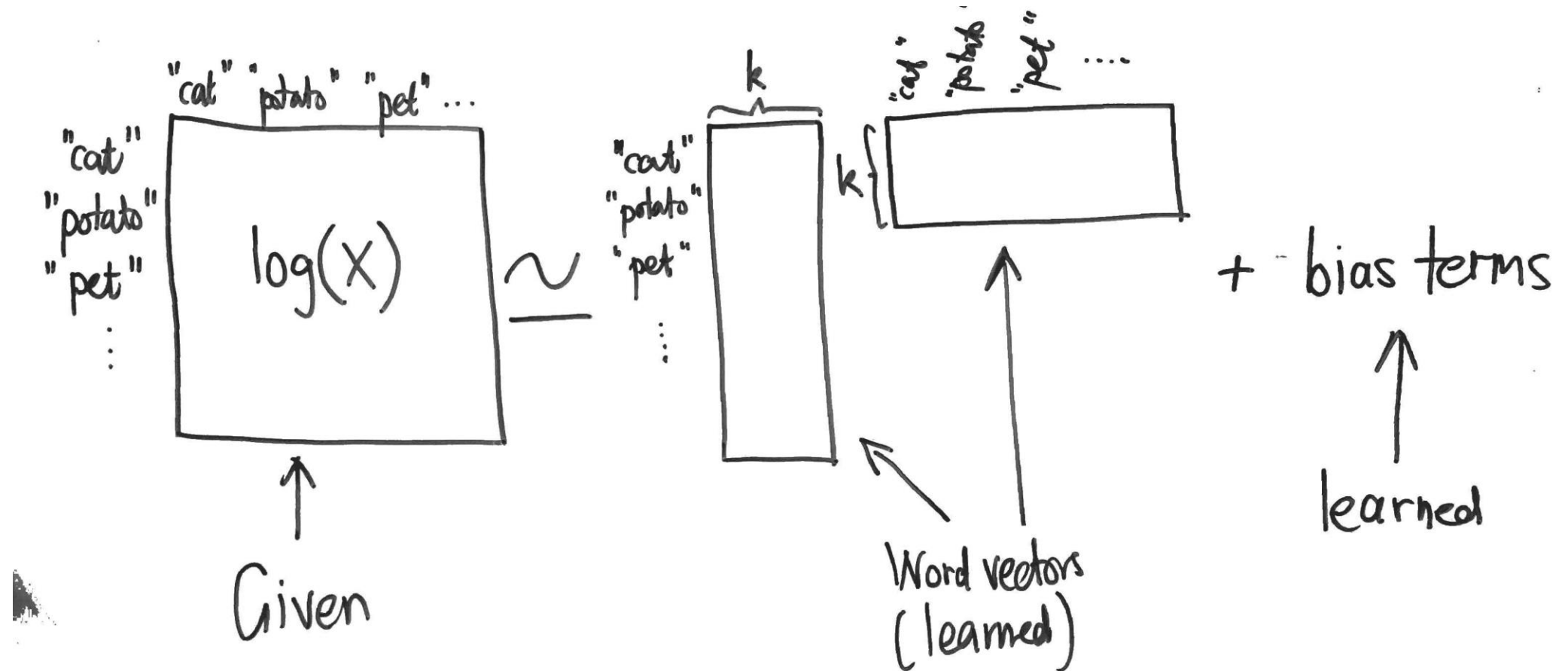
- Reduce dimensionality
- Information-rich
- **Unsupervised**

A.		N dimensions	B.		X dimensions
N diagnoses					
	Type 2 diabetes	[0, 0, 0, 0, 1, 0, 0, 0, 0, ... 0]	Type 2 diabetes	[0.2, -1.2, -0.2, 1.1 ... 0.8]	
	Hypertension	[0, 1, 0, 0, 0, 0, 0, 0, 0, ... 0]	Hypertension	[0.1, 1.1, 0.3, 0.6, ... 0.5]	
	Atrial fibrillation	[0, 0, 0, 0, 0, 0, 0, 0, 1, ... 0]	Atrial fibrillation	[1.2, 1.1, 0.3, -0.88, ... 0.2]	
	⋮	⋮	⋮	⋮	
Asthma	[0, 0, 0, 1, 0, 0, 0, 0, 0, ... 0]	Asthma	[-0.3, -0.9, -0.1, 1.9... 1.2]		

Word Embedding

- A straightforward idea: can we learn an embedding vector for each word from which we can...
 - Retrieve their similarities/co-occurrences/dependencies, etc.?
 - Infer their relations? (e.g., king -> queen, man -> woman)
- Pre-training
 - Training data are typically small
 - But many rules, dependencies, etc. in the language are universal
 - First learn the word embeddings from a large corpus of text, then apply them to the specific supervised task

GloVe



<http://building-babylon.net/2015/07/29/glove-global-vectors-for-word-representations/>

word2vec

The quick brown fox jumps over the lazy dog.

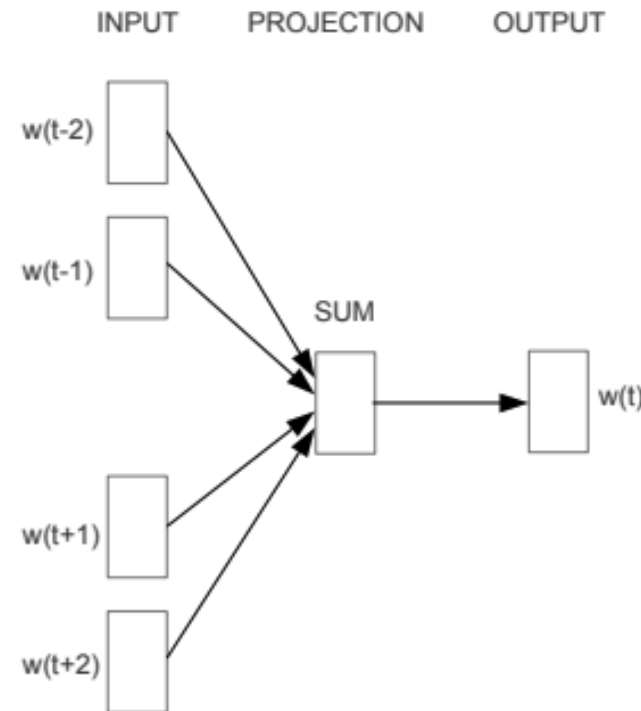
The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog.

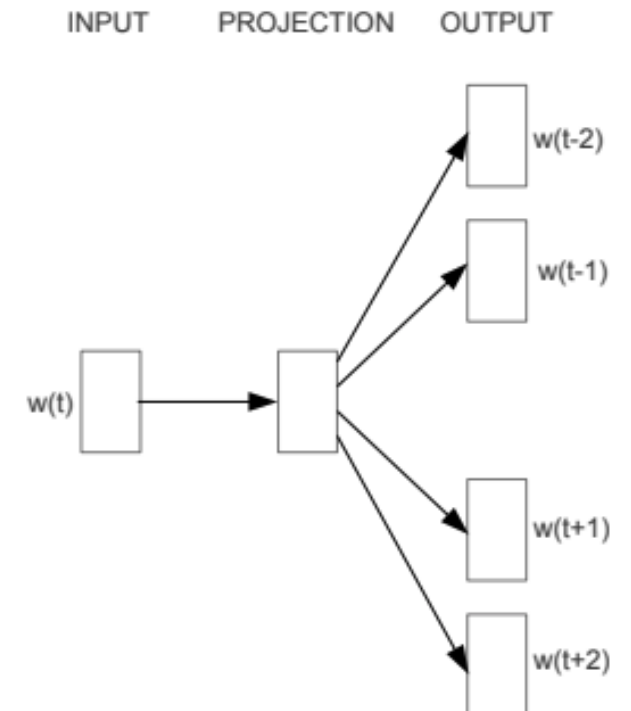
The quick brown fox jumps over the lazy dog.



Patient has a history of coronary artery diseases and hypertension. He.....



CBOW



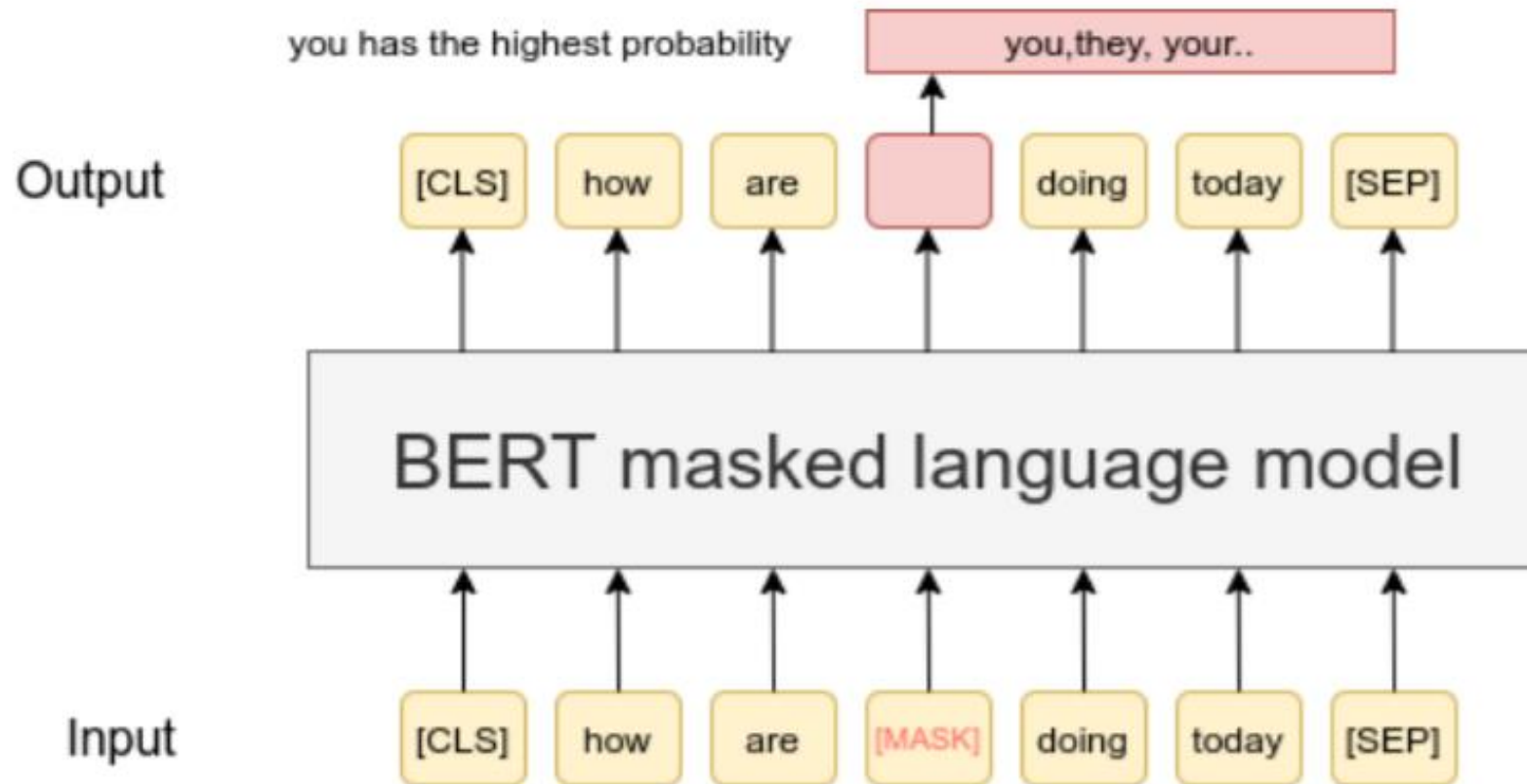
Skip-gram

<https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1>

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Bai, Tian, et al. "EHR phenotyping via jointly embedding medical concepts and words into a unified vector space." *BMC medical informatics and decision making* 18.4 (2018): 15-25.

BERT



https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

BioBERT

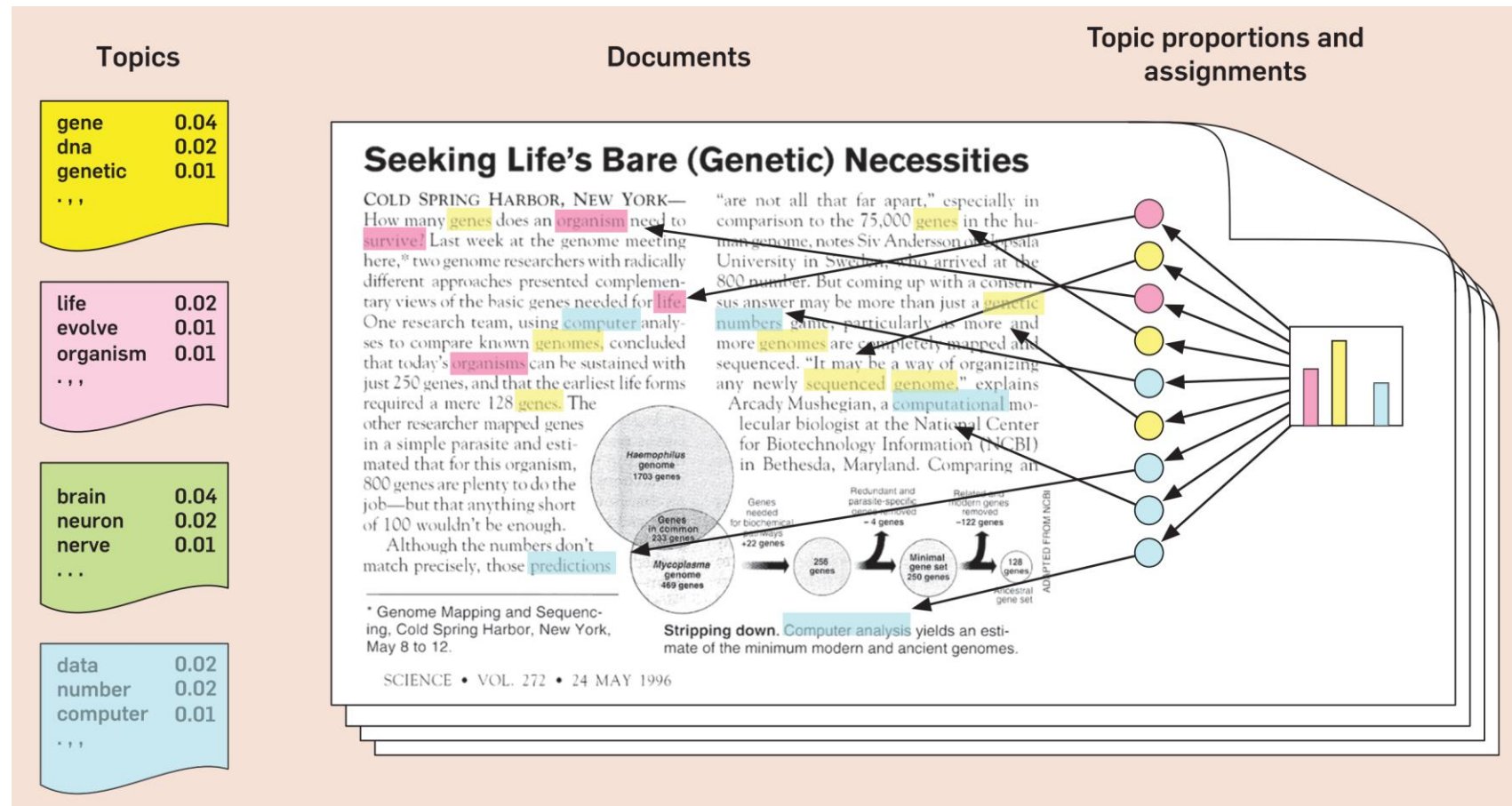
Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical

Model	Corpus combination
BERT (Devlin <i>et al.</i> , 2019)	Wiki + Books
BioBERT (+PubMed)	Wiki + Books + PubMed
BioBERT (+PMC)	Wiki + Books + PMC
BioBERT (+PubMed + PMC)	Wiki + Books + PubMed + PMC

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

Topic Modelling

Topic Modelling



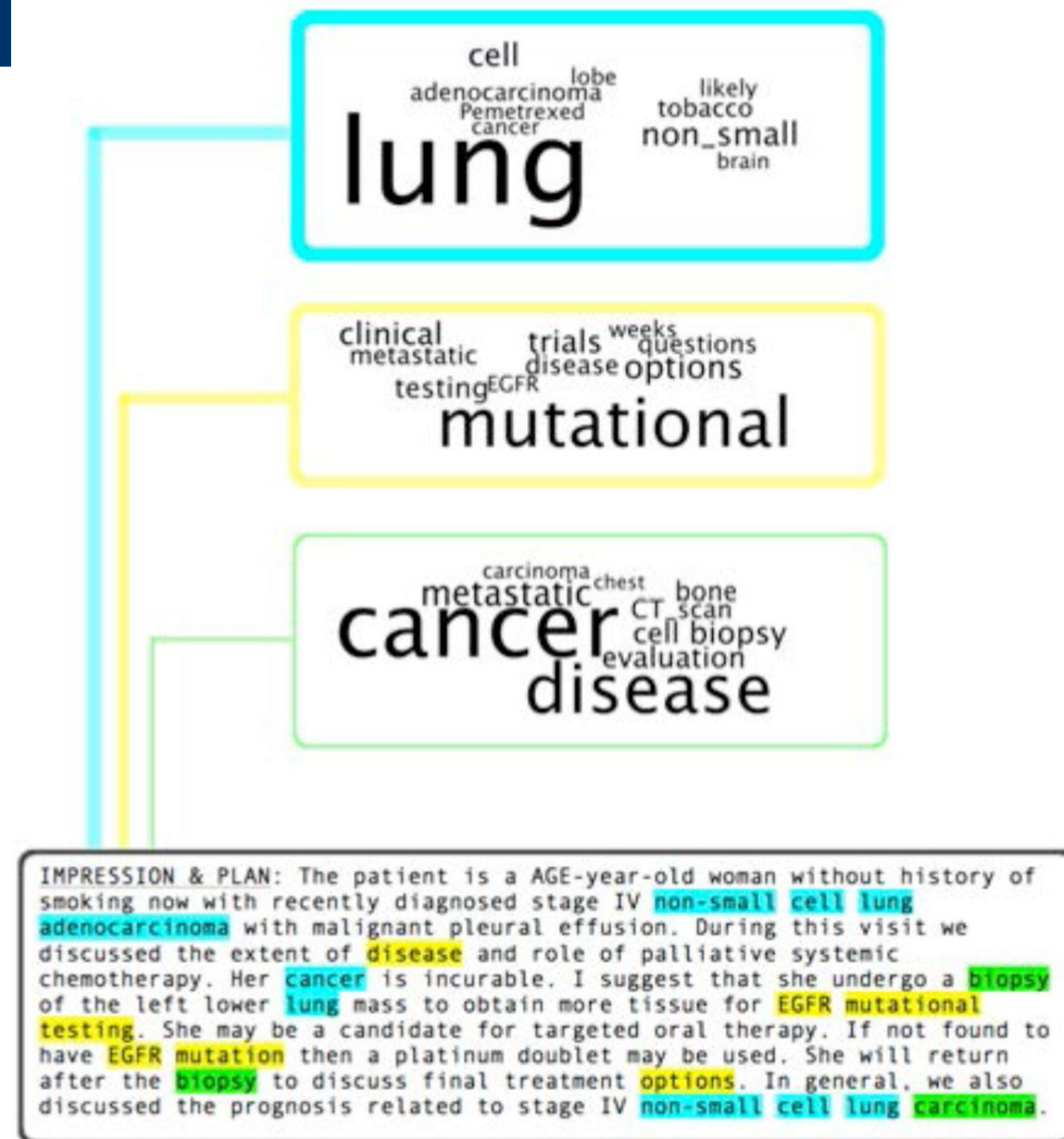
Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.

Topic Modelling

- Some basic assumptions:
 - Each document contains certain latent topics
 - Each topic generates certain words (with a certain probability)~component
- In other words:
 - If a document has high proportion of a certain topic, it should also contain many words related to that topic.

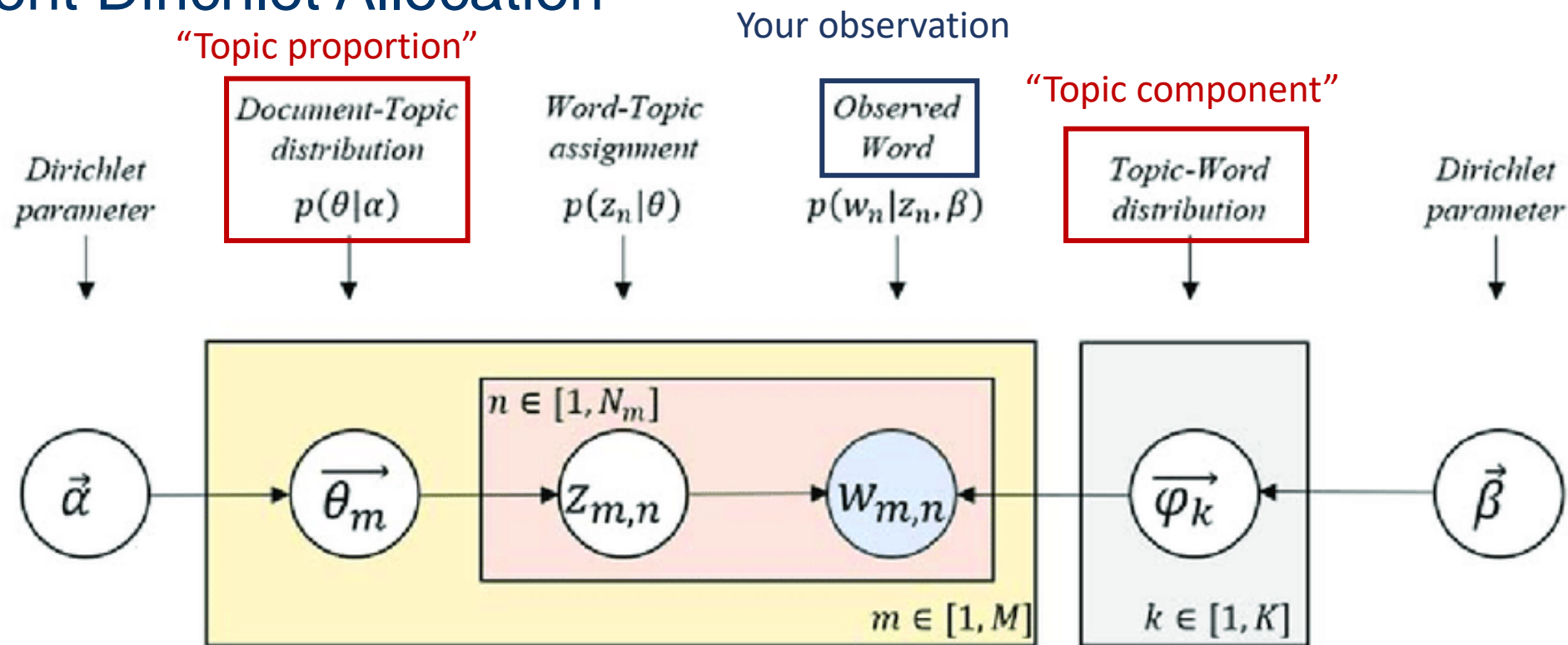
Chang, Katherine Redfield, et al. "An Empirical Analysis of Topic Modeling for Mining Cancer Clinical Notes." *bioRxiv* (2016): 062307.

3/9/2022



Topic Modelling

- Latent Dirichlet Allocation



<https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/>

Topic Modelling

- Latent Dirichlet Allocation

- Gibbs sampling:

- Assign words to topics based on topic proportion & topic component
 - Update topic proportion & topic component based on topic assignments

Topic proportion \sim # of words are assigned to the topic (for each doc)

Topic component \sim # of times a certain word is assigned to the topic

Thank you!