Overview of Natural Language Processing (NLP) in biomedical and cancer research

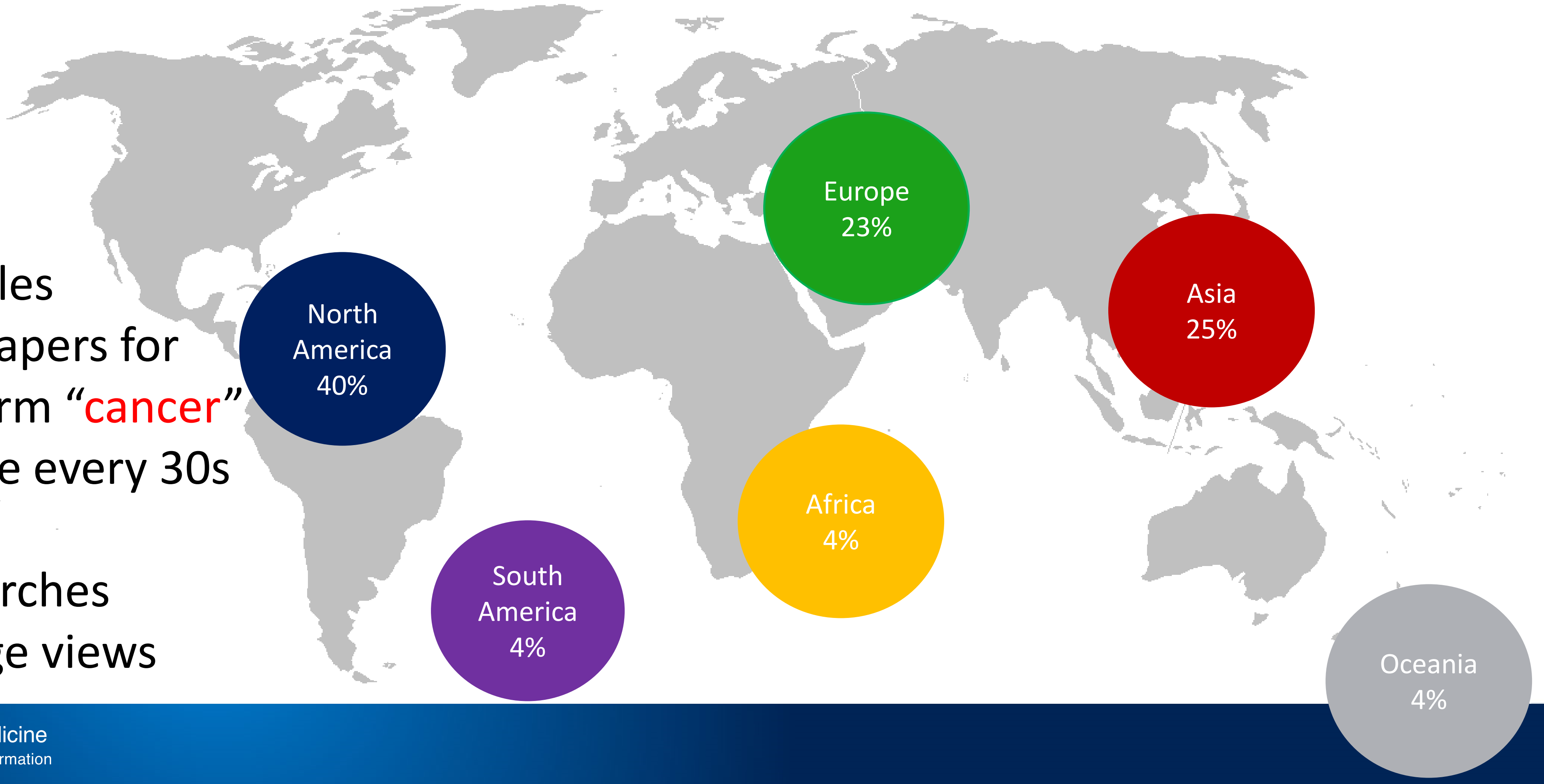
Yifan Peng, Qingyu Chen

NCBI/NLM/NIH

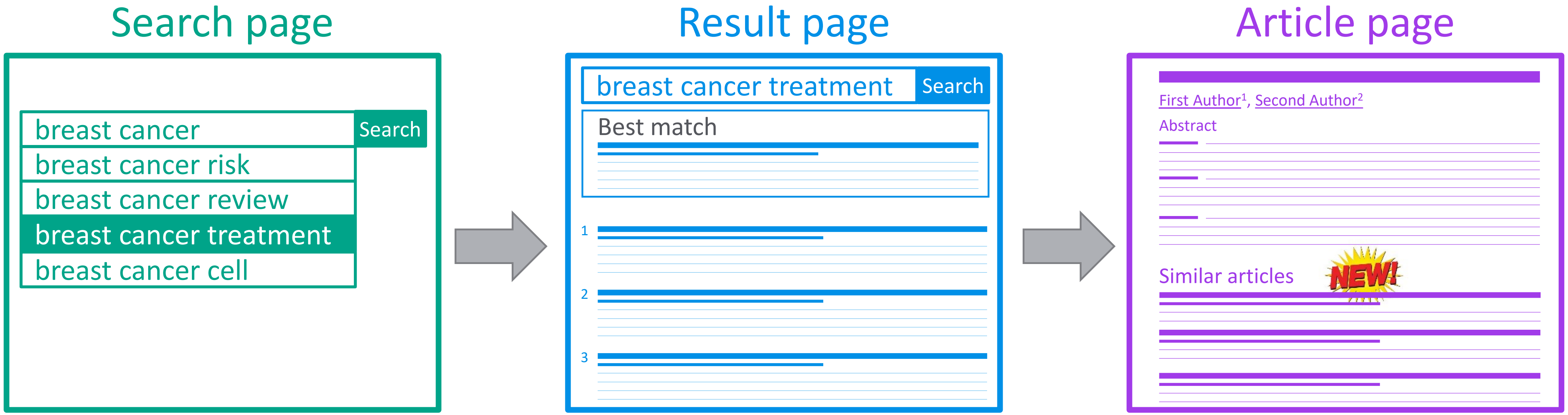
U.S. National Library of Medicine
National Center for Biotechnology Information

PubMed: biomedical literature search engine

- 28+ million articles
  - 3.8 million papers for the query term “cancer”
  - A new article every 30s
- Daily usage
  - 3 million searches
  - 9 million page views



What happens when you click the "Search" button on PubMed



- Spelling correction
- Query expansion
- Query suggestion

- Navigational searches
- Relevance match

- Related articles suggestions
- Author name disambiguation
- Citation sensor

https://www.ncbi.nlm.nih.gov/labs/pubmed/

Fiorini, et al., Best Match: New relevance search for PubMed, PLoS Biology, 2018

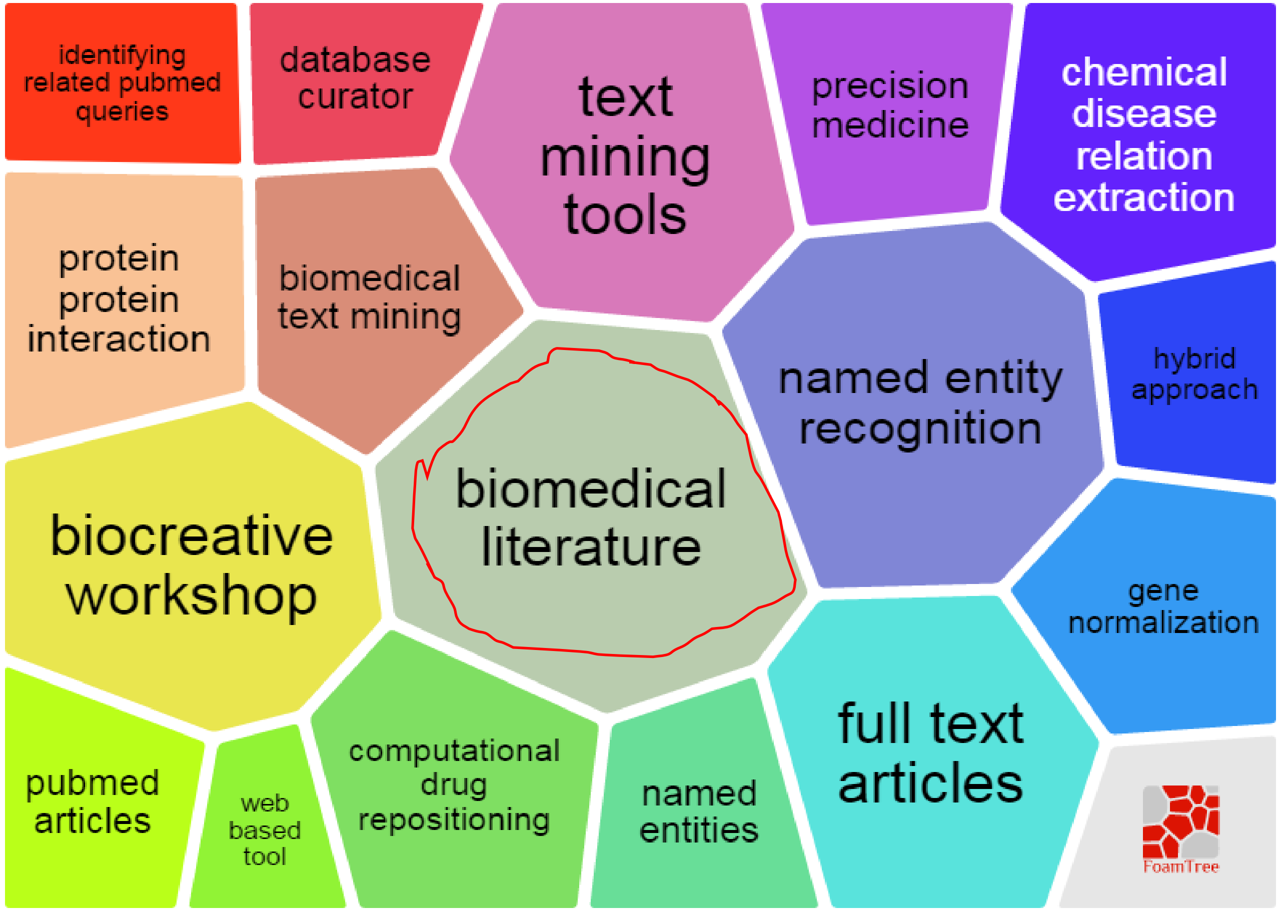
What is Natural Language Processing (NLP)?

- **Natural language processing** is a field at the intersection of
  - Computer science
  - Artificial intelligence
  - Linguistics

- **Goal**: for computers to “understand” natural language in order to perform tasks that are useful

NLP goes beyond the biomedical literature

- Biomedical Literature
- Clinical notes, EMRs
  - Chest X-ray & retinal images

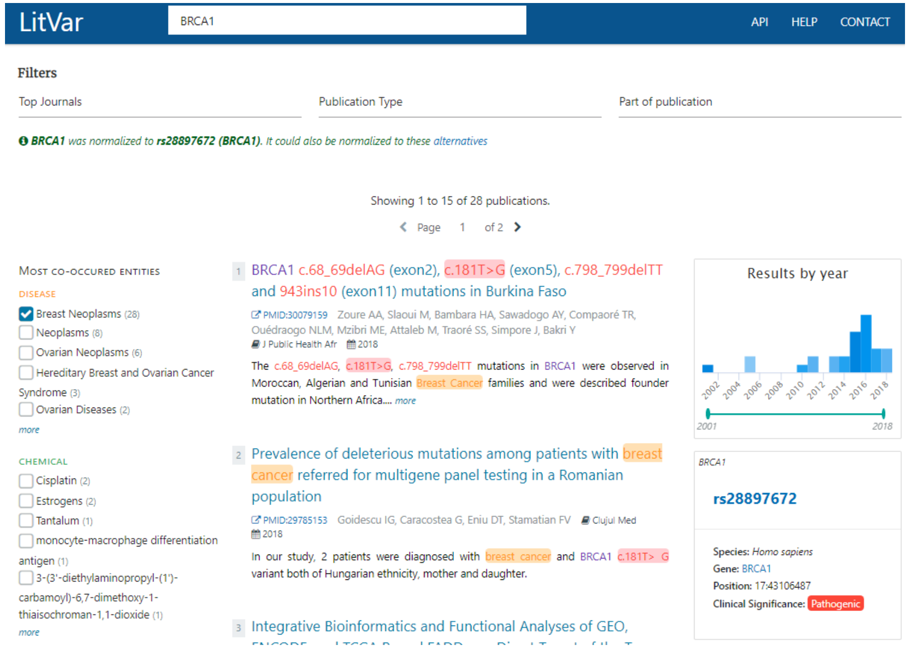


NLP is important for cancer research

- Finding relevant literature
- Extracting important entities such as cancers and treatment mentioned in literature
- Understanding the semantics of language
- Classifying related documents for manual curation
- Helping image analysis

NLP helps extract information

LitVar: Extracting mutation information from articles



Filters

Top Journals | Publication Type | Part of publication

*BRCA1 was normalized to **rs28897672 (BRCA1)**. It could also be normalized to these alternatives*

Showing 1 to 15 of 28 publications.

Page 1 of 2

MOST CO-OCCURED ENTITIES

DISEASE

- ☑ Breast Neoplasms (28)
- □ Neoplasms (8)
- □ Ovarian Neoplasms (6)
- □ Hereditary Breast and Ovarian Cancer Syndrome (3)
- □ Ovarian Diseases (2)

more

CHEMICAL

- □ Cisplatin (2)
- □ Estrogens (2)
- □ Tantalum (1)
- □ monocyte-macrophage differentiation antigen (1)
- □ 3-(3'-diethylaminopropyl-(1')-carbamoyl)-6,7-dimethoxy-1-thiaisochroman-1,1-dioxide (1)

more

1 BRCA1 c.68_69delAG (exon2), c.181T>G (exon5), c.798_799delTT and 943ins10 (exon11) mutations in Burkina Faso

PMID:30079159 Zoure AA, Slaoui M, Bambara HA, Sawadogo AY, Compaoré TR, Ouédraogo NLM, Mzibri ME, Attaleb M, Traoré SS, Simpore J, Bakri Y
J Public Health Afr 2018

The c.68_69delAG, c.181T>G, c.798_799delTT mutations in BRCA1 were observed in Moroccan, Algerian and Tunisian Breast Cancer families and were described founder mutation in Northern Africa.... *more*

2 Prevalence of deleterious mutations among patients with breast cancer referred for multigene panel testing in a Romanian population

PMID:29785153 Goidescu IG, Caracostea G, Eniu DT, Stamatian FV Clujul Med
2018

In our study, 2 patients were diagnosed with breast cancer and BRCA1 c.181T> G variant both of Hungarian ethnicity, mother and daughter.

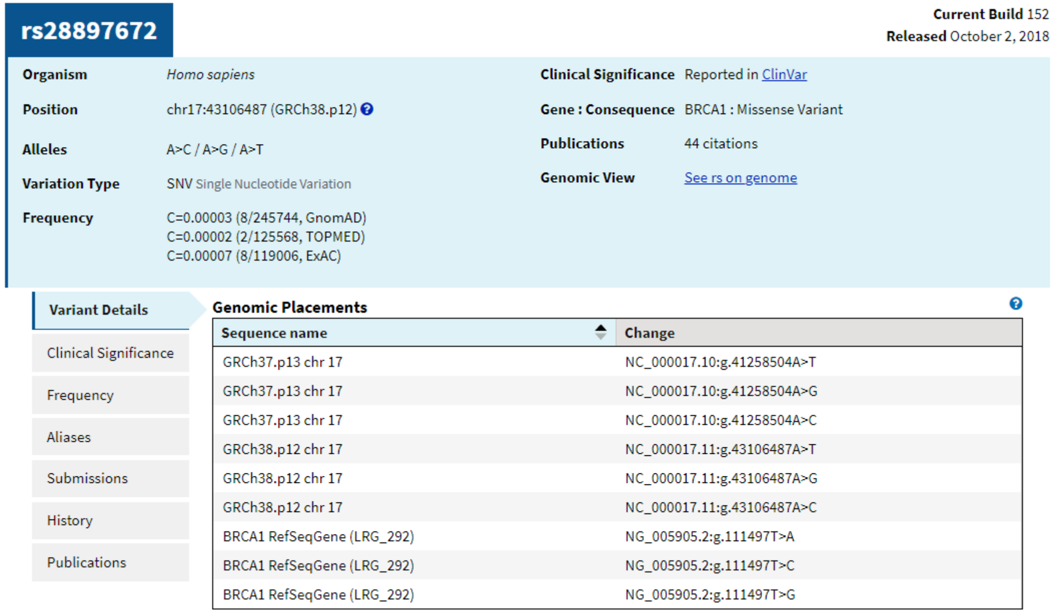
3 Integrative Bioinformatics and Functional Analyses of GEO,

Results by year

BRCA1

rs28897672

Species: *Homo sapiens*
Gene: BRCA1
Position: 17:43106487
Clinical Significance: Pathogenic



rs28897672

Current Build 152
Released October 2, 2018

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr17:43106487 (GRCh38.p12)	**Gene : Consequence**	BRCA1 : Missense Variant
Alleles	A>C / A>G / A>T	**Publications**	44 citations
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome

Frequency	C=0.00003 (8/245744, GnomAD)
	C=0.00002 (2/125568, TOPMED)
	C=0.00007 (8/119006, ExAC)		

Genomic Placements

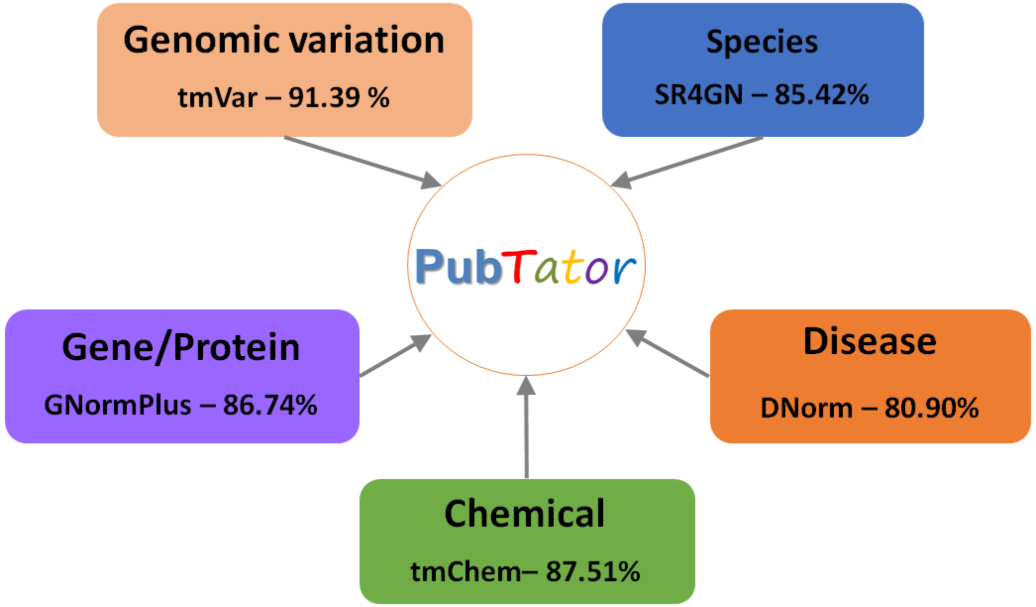
Sequence name	Change
GRCh37.p13 chr 17	NC_000017.10:g.41258504A>T
GRCh37.p13 chr 17	NC_000017.10:g.41258504A>G
GRCh37.p13 chr 17	NC_000017.10:g.41258504A>C
GRCh38.p12 chr 17	NC_000017.11:g.43106487A>T
GRCh38.p12 chr 17	NC_000017.11:g.43106487A>G
GRCh38.p12 chr 17	NC_000017.11:g.43106487A>C
BRCA1 RefSeqGene (LRG_292)	NG_005905.2:g.111497T>A
BRCA1 RefSeqGene (LRG_292)	NG_005905.2:g.111497T>C
BRCA1 RefSeqGene (LRG_292)	NG_005905.2:g.111497T>G

https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/

Allot et al., LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Research. 2018

LitVar is supported by PubTator

Named entity recognition tool



Genomic variation

tmVar – 91.39 %

Species

SR4GN – 85.42%

PubTator

Gene/Protein

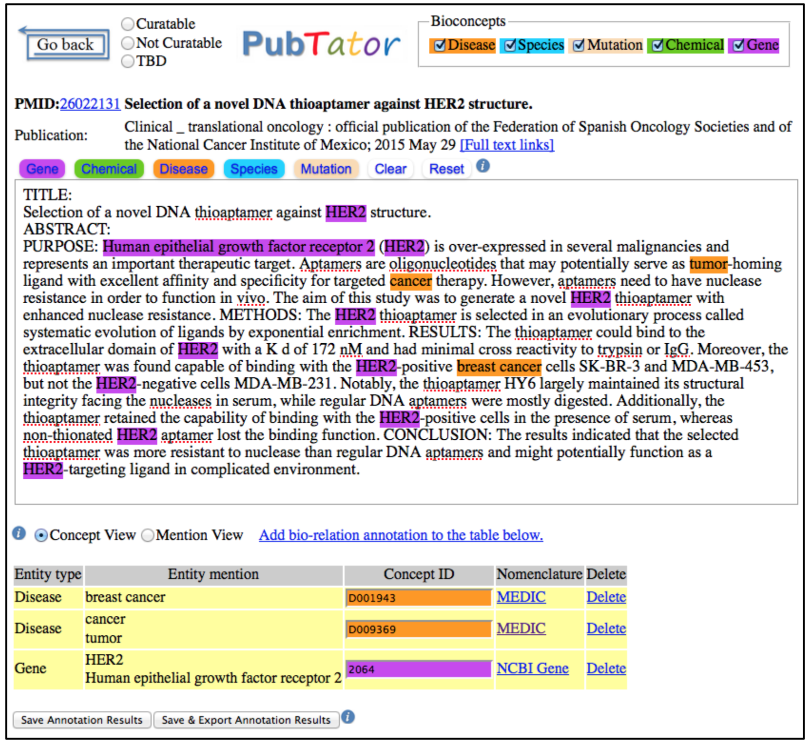
GNormPlus – 86.74%

Disease

DNorm – 80.90%

Chemical

tmChem– 87.51%


Go back
○Curatable
○Not Curatable
○TBD
PubTator
Bioconcepts
☑Disease ☑Species ☑Mutation ☑Chemical ☑Gene

PMID:26022131 Selection of a novel DNA thioaptamer against HER2 structure.

Publication: Clinical _ translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico; 2015 May 29 [Full text links]

Gene Chemical Disease Species Mutation Clear Reset

TITLE:
Selection of a novel DNA thioaptamer against HER2 structure.
ABSTRACT:
PURPOSE: Human epithelial growth factor receptor 2 (HER2) is over-expressed in several malignancies and represents an important therapeutic target. Aptamers are oligonucleotides that may potentially serve as tumor-homing ligand with excellent affinity and specificity for targeted cancer therapy. However, aptamers need to have nuclease resistance in order to function in vivo. The aim of this study was to generate a novel HER2 thioaptamer with enhanced nuclease resistance. METHODS: The HER2 thioaptamer is selected in an evolutionary process called systematic evolution of ligands by exponential enrichment. RESULTS: The thioaptamer could bind to the extracellular domain of HER2 with a K d of 172 nM and had minimal cross reactivity to trypsin or IgG. Moreover, the thioaptamer was found capable of binding with the HER2-positive breast cancer cells SK-BR-3 and MDA-MB-453, but not the HER2-negative cells MDA-MB-231. Notably, the thioaptamer HY6 largely maintained its structural integrity facing the nucleases in serum, while regular DNA aptamers were mostly digested. Additionally, the thioaptamer retained the capability of binding with the HER2-positive cells in the presence of serum, whereas non-thionated HER2 aptamer lost the binding function. CONCLUSION: The results indicated that the selected thioaptamer was more resistant to nuclease than regular DNA aptamers and might potentially function as a HER2-targeting ligand in complicated environment.

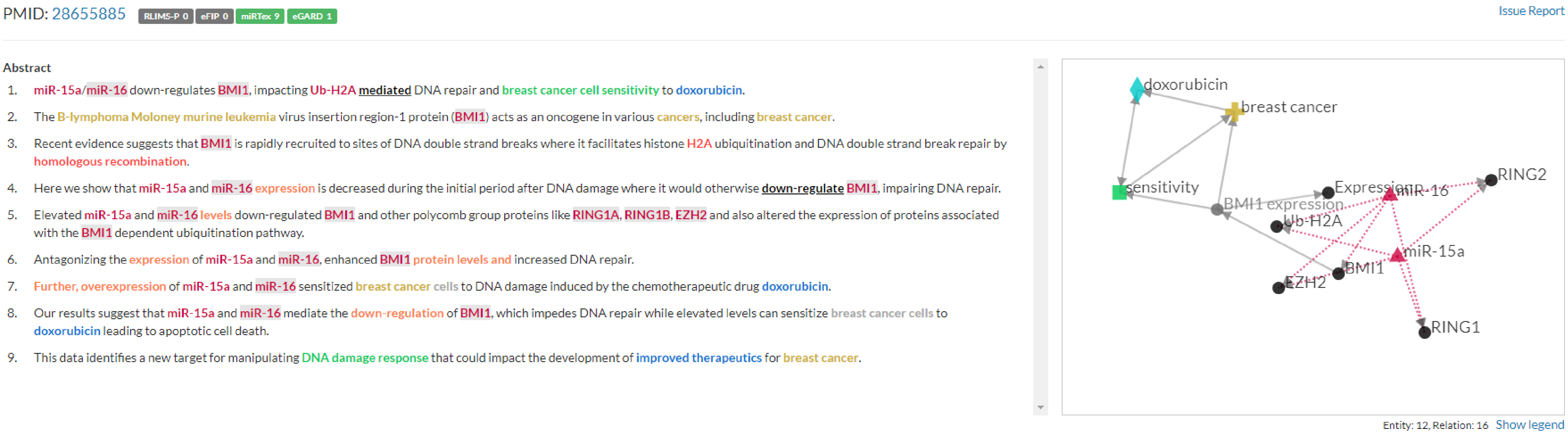
● Concept View ○ Mention View Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	Delete
Disease	breast cancer	D001943	MEDIC	Delete
Disease	cancer tumor	D009369	MEDIC	Delete
Gene	HER2 Human epithelial growth factor receptor 2	2064	NCBI Gene	Delete

https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

Wei et. al., PubTator: a Web-based text mining tool for assisting Biocuration, Nucleic acids research, 2013.

From named entities to relations



PMID: 28655885 RLIMS-P 0 eFIP 0 miRTex 9 eGARD 1

Issue Report

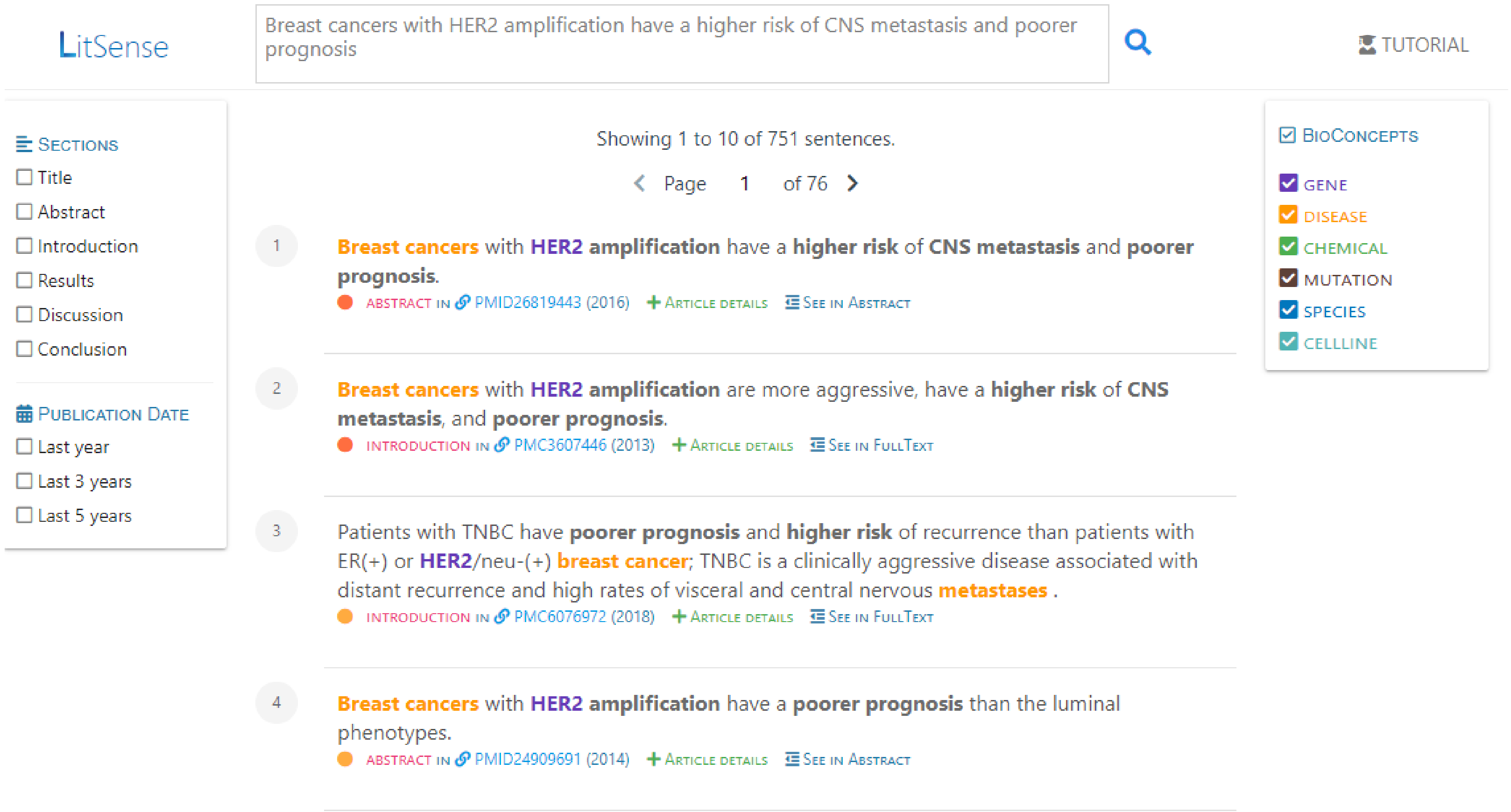
Abstract

1. miR-15a/miR-16 down-regulates BMI1, impacting Ub-H2A mediated DNA repair and breast cancer cell sensitivity to doxorubicin.
2. The B-lymphoma Moloney murine leukemia virus insertion region-1 protein (BMI1) acts as an oncogene in various cancers, including breast cancer.
3. Recent evidence suggests that BMI1 is rapidly recruited to sites of DNA double strand breaks where it facilitates histone H2A ubiquitination and DNA double strand break repair by homologous recombination.
4. Here we show that miR-15a and miR-16 expression is decreased during the initial period after DNA damage where it would otherwise down-regulate BMI1, impairing DNA repair.
5. Elevated miR-15a and miR-16 levels down-regulated BMI1 and other polycomb group proteins like RING1A, RING1B, EZH2 and also altered the expression of proteins associated with the BMI1 dependent ubiquitination pathway.
6. Antagonizing the expression of miR-15a and miR-16, enhanced BMI1 protein levels and increased DNA repair.
7. Further, overexpression of miR-15a and miR-16 sensitized breast cancer cells to DNA damage induced by the chemotherapeutic drug doxorubicin.
8. Our results suggest that miR-15a and miR-16 mediate the down-regulation of BMI1, which impedes DNA repair while elevated levels can sensitize breast cancer cells to doxorubicin leading to apoptotic cell death.
9. This data identifies a new target for manipulating DNA damage response that could impact the development of improved therapeutics for breast cancer.

Entity: 12, Relation: 16 Show legend

https://research.bioinformatics.udel.edu/itextmine/

LitSense: sentence-level retrieval



Breast cancers with HER2 amplification have a higher risk of CNS metastasis and poorer prognosis

Showing 1 to 10 of 751 sentences.

Page 1 of 76

1 **Breast cancers** with **HER2 amplification** have a **higher risk** of **CNS metastasis** and **poorer prognosis**.
● ABSTRACT IN PMID26819443 (2016) + ARTICLE DETAILS SEE IN ABSTRACT

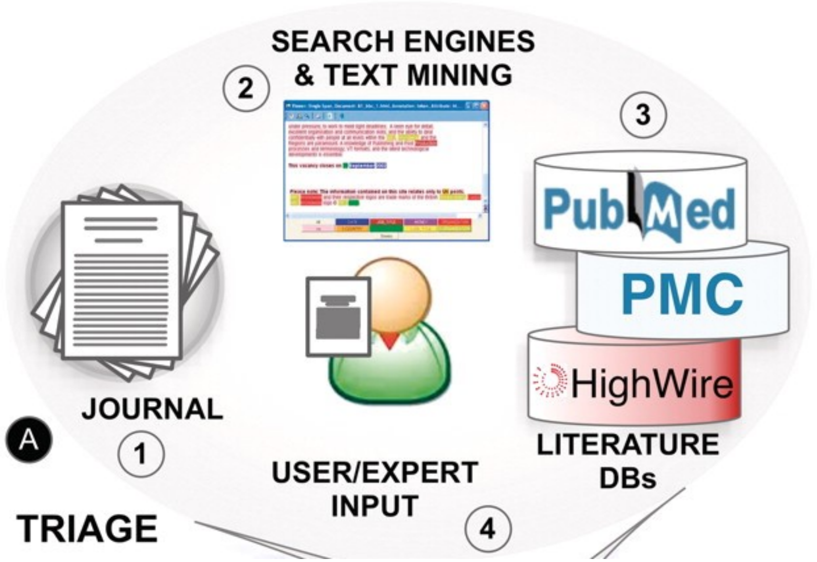
2 **Breast cancers** with **HER2 amplification** are more aggressive, have a **higher risk** of **CNS metastasis**, and **poorer prognosis**.
● INTRODUCTION IN PMC3607446 (2013) + ARTICLE DETAILS SEE IN FULLTEXT

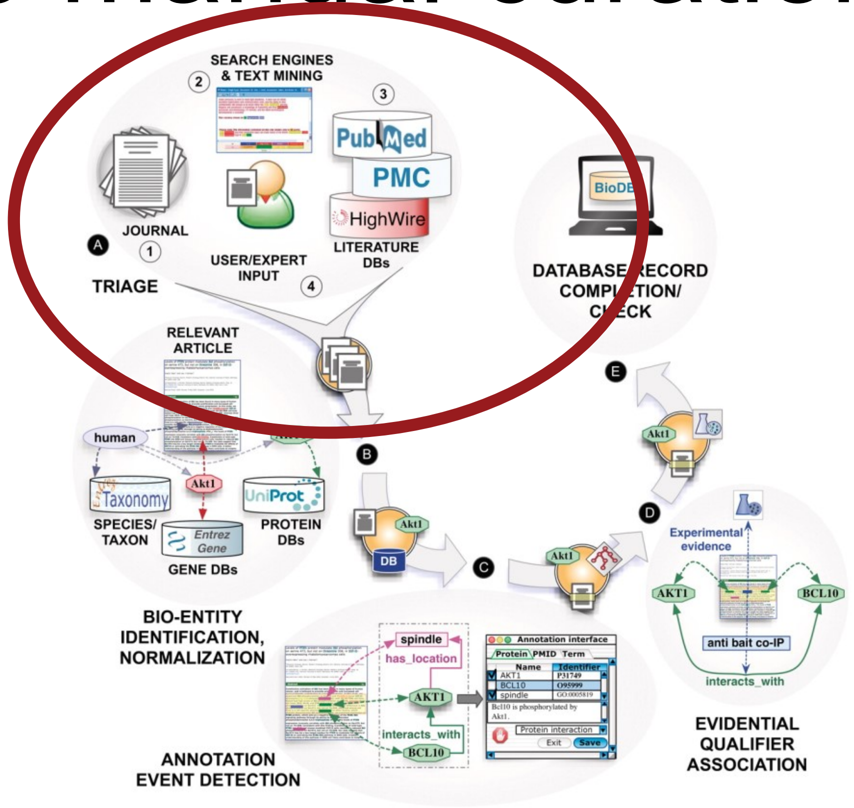
3 Patients with TNBC have **poorer prognosis** and **higher risk** of recurrence than patients with ER(+) or **HER2**/neu-(+) **breast cancer**; TNBC is a clinically aggressive disease associated with distant recurrence and high rates of visceral and central nervous **metastases** .
● INTRODUCTION IN PMC6076972 (2018) + ARTICLE DETAILS SEE IN FULLTEXT

4 **Breast cancers** with **HER2 amplification** have a **poorer prognosis** than the luminal phenotypes.
● ABSTRACT IN PMID24909691 (2014) + ARTICLE DETAILS SEE IN ABSTRACT

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

NLP helps scale up manual curation





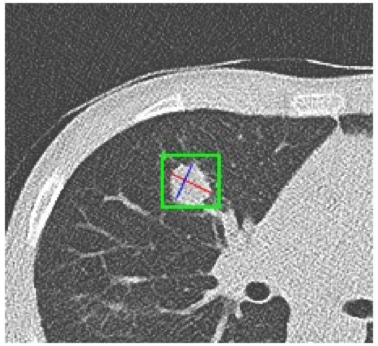
GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies

SIB swissprot EBI

Lee et al., Scaling up data curation using deep learning: An application to literature triage in genomic variation resources, PLoS Comp Biol. 2018.

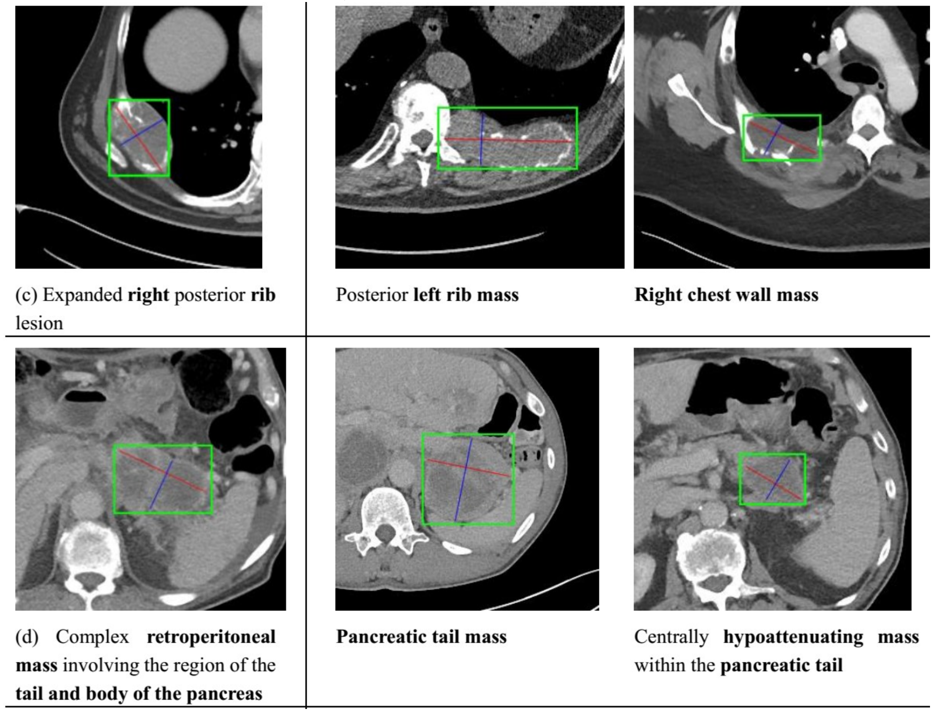
NLP helps biomedical image analysis

DeepLesion: Lesion annotation, detection, and retrieval



Unchanged **large** **nodule** bilaterally for example **right lower lobe** [OTHER BOOKMARK] and **right middle lobe** [BOOKMARK]

Size:	large
Type:	nodule
Body part:	right mid lobe
Unrelated:	right lower lobe



(c) Expanded **right** posterior **rib** lesion

Posterior **left rib mass**

Right chest wall mass

(d) Complex **retroperitoneal mass** involving the region of the **tail and body of the pancreas**

Pancreatic tail mass

Centrally **hypoattenuating mass** within the **pancreatic tail**

Yan et al., Fine-grained lesion annotation in CT images with knowledge mined from radiology reports., ISBI, 2019

Common Thorax Disease Classification and Reporting in Chest X-rays

Chest X-ray: the largest public X-ray image dataset generated by NLP tools

- Over 100,000 frontal-view X-ray images
- 30k unique patients
- 14 common thorax diseases (e.g. pneumonia)

Wang et al., TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. CVPR 2018

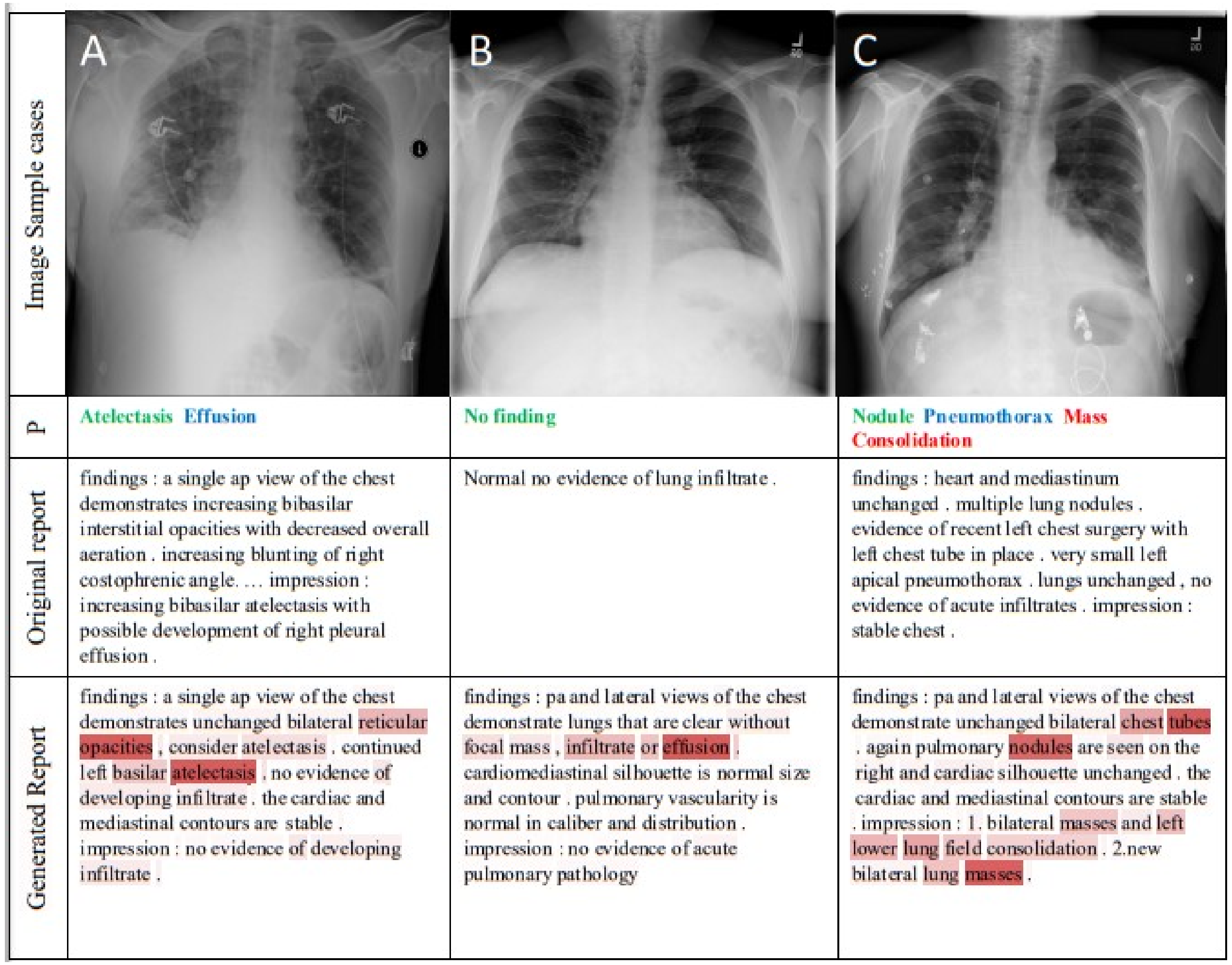


Image Sample cases	A	B	C
P	Atelectasis Effusion	No finding	Nodule Pneumothorax Mass Consolidation
Original report	findings : a single ap view of the chest demonstrates increasing bibasilar interstitial opacities with decreased overall aeration . increasing blunting of right costophrenic angle. … impression : increasing bibasilar atelectasis with possible development of right pleural effusion .	Normal no evidence of lung infiltrate .	findings : heart and mediastinum unchanged . multiple lung nodules . evidence of recent left chest surgery with left chest tube in place . very small left apical pneumothorax . lungs unchanged , no evidence of acute infiltrates . impression : stable chest .
Generated Report	findings : a single ap view of the chest demonstrates unchanged bilateral reticular opacities , consider atelectasis . continued left basilar atelectasis . no evidence of developing infiltrate . the cardiac and mediastinal contours are stable . impression : no evidence of developing infiltrate .	findings : pa and lateral views of the chest demonstrate lungs that are clear without focal mass , infiltrate or effusion . cardiomediastinal silhouette is normal size and contour . pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology	findings : pa and lateral views of the chest demonstrate unchanged bilateral chest tubes . again pulmonary nodules are seen on the right and cardiac silhouette unchanged . the cardiac and mediastinal contours are stable . impression : 1. bilateral masses and left lower lung field consolidation . 2.new bilateral lung masses .

A summary of NLP methods

- Rule-based
  - For instance, if two genes co-occur in literature, it will be considered as interacted
  - Simple and efficient, but cannot tackle complex scenarios and not generalizable

- Machine learning
  - Manually derive features as inputs
  - Have been used over decades, but are limited to domain knowledge

- Recent methods: **deep learning**
  - Automatically derive features and representations
  - Have outperformed traditional machine learning methods since 2010
  - Have been widely applied in NLP applications: question answering, translation…
  - Open issues: privacy and interpretability

Summary

- What is NLP?
- Why is NLP important?
  - NLP helps find relevant papers
  - NLP helps extract information (named entity, relation, …)
  - NLP helps image analysis
- NLP methods overview

Text Mining Group @ NCBI/NLM

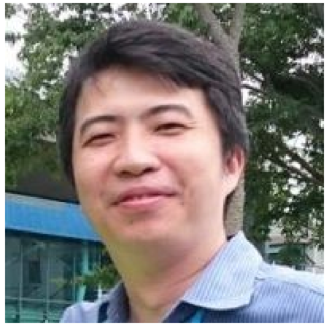
Zhiyong Lu
(Principal Investigator)

Alexis Allot

Qingyu Chen

Donald Comeau

Rezarta Dogan

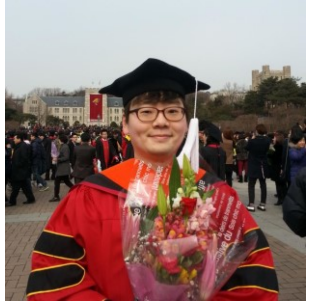


Alan Hsu

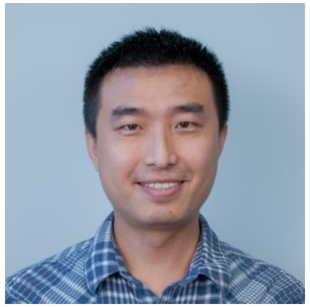
Sun Kim

Won Kim

Robert Leaman



Kyubum Lee



Yifan Peng

Chih-Hsuan Wei

Lana Yeganova

Resources

- FAES course
  - BIOF395 “Introduction to Text Mining”, Fall 2019 (instructed by us)

- NIH.AI workshop on NLP
  - **Who**: Entry-level to advanced NIH researchers working with NLP
  - **When**: April or May, 2019
  - **Mail list**: BIOINFORMATICS-SIG-L@LIST.NIH.GOV

- Reviews
  - https://www.ncbi.nlm.nih.gov/labs/pubmed/27807747
  - https://www.ncbi.nlm.nih.gov/labs/pubmed/22549152
  - https://www.ncbi.nlm.nih.gov/labs/pubmed/19649304

Q & A

yifan.peng@nih.gov

qingyu.chen@nih.gov