

Natural Language Processing

Sonja Aits

Lund University, 20250924



Natural Language Processing (NLP)

= Computational analysis and generation of natural language (text or speech)

BioNLP

= NLP related to medicine and life sciences



Google Search

I'm Feeling Lucky

Google offered in: [svenska](#)

NLP tasks

Examples of text and speech used
in medicine and life sciences?

Language datasets

- Patient-doctor conversations
- Emergency calls
- Electronic health records
- Literature – abstracts or fulltext
- Reports
- Patents
- Social media posts
- ...

Summarization

Risk factors for breast cancer

🌐 3 languages ▾

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

(Redirected from [Risk factors of breast cancer](#))



This article **needs more [reliable medical references](#) for [verification](#)** or **relies too heavily on [primary sources](#)**. Please review the contents of the article and [add the appropriate references](#) if you can. Unsourced or poorly sourced material may be challenged and [removed](#).

Find sources: ["Risk factors for breast cancer"](#) – [news](#) · [newspapers](#) · [books](#) · [scholar](#) · [JSTOR](#) (November 2016)



Risk factors for breast cancer may be divided into preventable and non-preventable. Their study belongs in the field of [epidemiology](#). Breast cancer, like other forms of cancer, can result from multiple environmental and hereditary risk factors. The term "environmental", as used by cancer researchers, means any risk factor that is not genetically inherited.

For breast cancer, the list of environmental risk factors includes the individual person's development, exposure to [microbes](#), "medical interventions, dietary exposures to nutrients, energy and toxicants, ionizing radiation, and chemicals from industrial and agricultural processes and from consumer products...reproductive choices, energy balance, adult weight gain, body fatness, voluntary and involuntary physical activity, medical care, exposure to tobacco smoke and alcohol, and occupational exposures, including shift work" as well as "metabolic and physiologic processes that modify the body's internal environment."^[1] Some of these environmental factors are part of the physical environment, while others (such as diet and number of pregnancies) are primarily part of the social, cultural, or economic environment.^[1]

Although many epidemiological risk factors have been identified, the cause of any individual breast cancer is most often unknowable. Epidemiological research informs the patterns of breast cancer incidence across certain populations, but not in a given individual. Approximately 5% of new breast cancers are attributable to hereditary syndromes, and well-established risk factors accounts for approximately 30% of cases.^[2]



Common risk factors for breast cancer are age, female sex, hereditary factors (e.g. BRCA1/2 status) and life style factors (e.g. obesity and alcohol).

Translation

≡ Google Translate

Text

Documents

DETECT LANGUAGE

ENGLISH

SPANISH

FRENCH



GERMAN

SWEDISH

ENGLISH



Oxidative stress causes lysosomal damage and subsequent cell death in many types of cells.



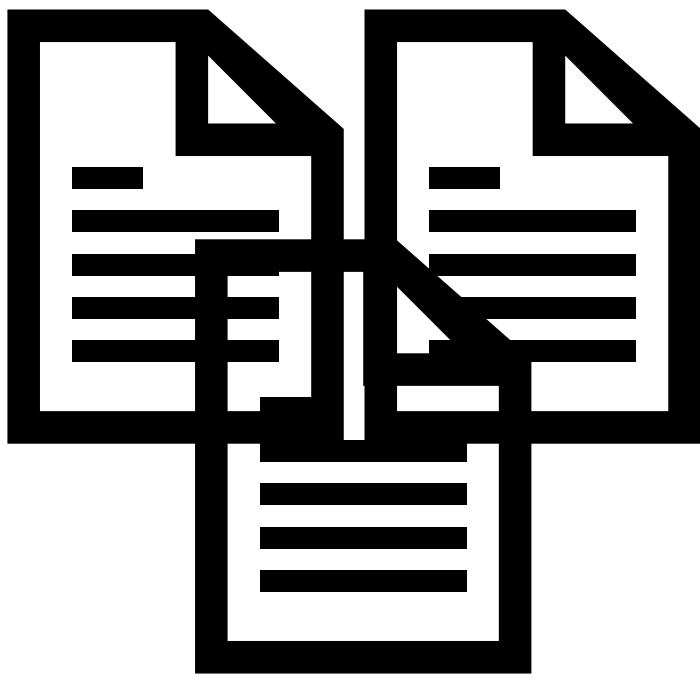
Oxidativer Stress führt bei vielen Zelltypen zu lysosomalen Schäden und anschließendem Zelltod.



90/5000

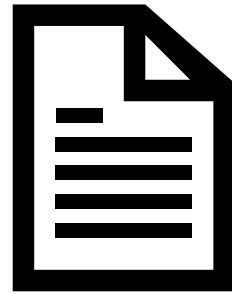


Text clustering



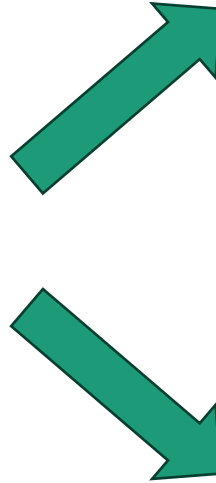
Text classification

Electronic
health record

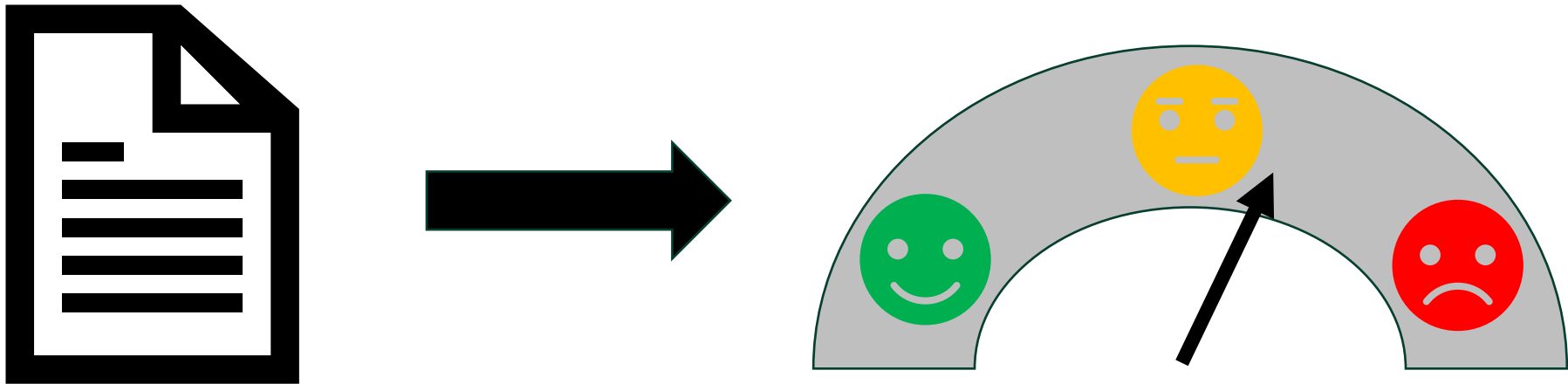


High risk of heart attack

Low risk of heart attack



Sentiment analysis



Named entity recognition (token classification)

Disease

Drug/treatment

Gene/protein

Process/location

Relation

[PLoS One](#). 2012;7(10):e45381. doi: 10.1371/journal.pone.0045381. Epub 2012 Oct 11.

Identification of cytoskeleton-associated proteins essential for lysosomal stability and survival of human cancer cells.

[Groth-Pedersen L](#)¹, [Aits S](#), [Corcelle-Termeau E](#), [Petersen NH](#), [Nylandsted J](#), [Jäättelä M](#).

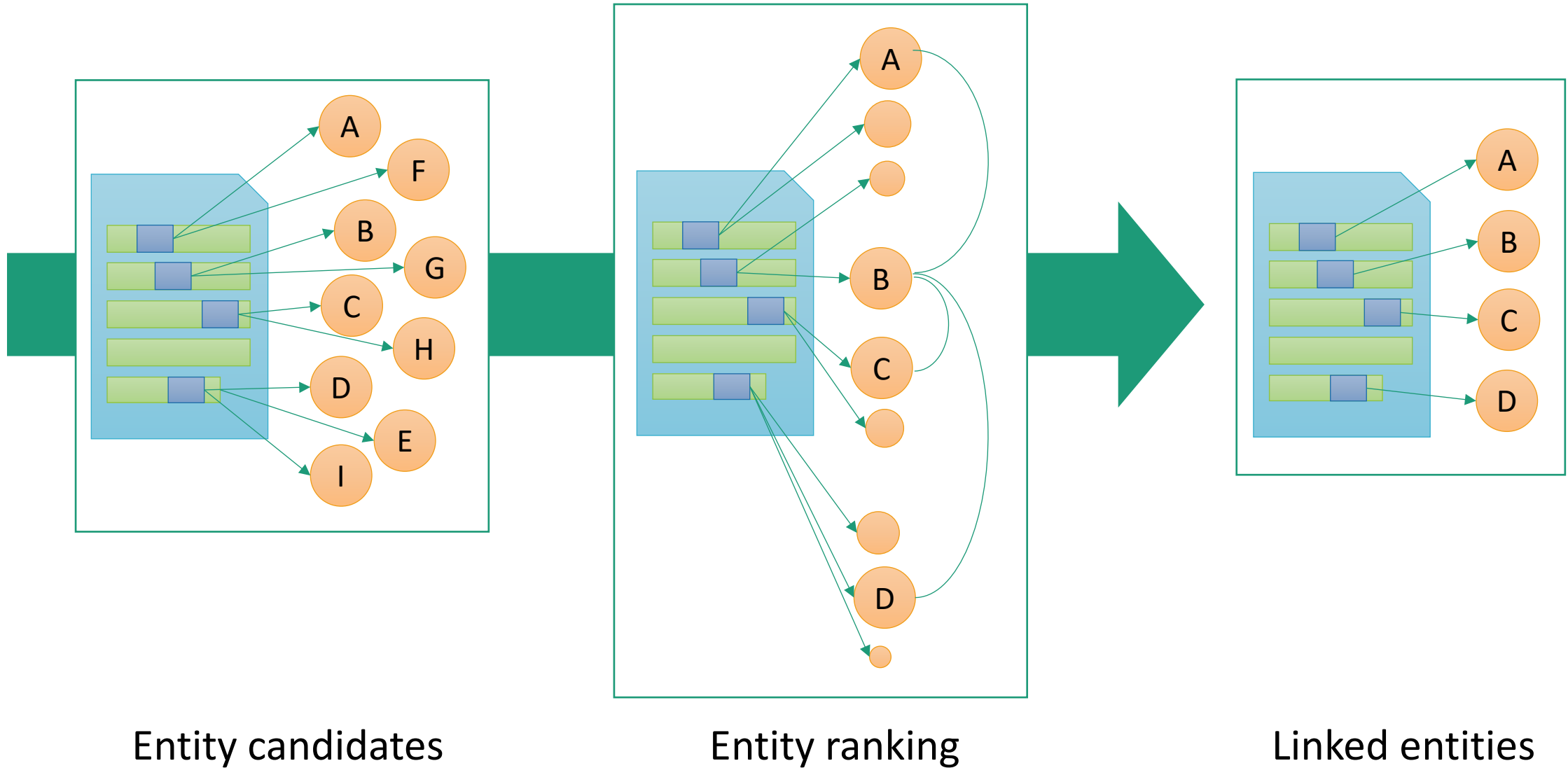
[Author information](#)

Abstract

Microtubule-disturbing drugs inhibit lysosomal trafficking and induce lysosomal membrane permeabilization followed by cathepsin-dependent cell death. To identify specific trafficking-related proteins that control cell survival and lysosomal stability, we screened a molecular motor siRNA library in human MCF7 breast cancer cells. SiRNAs targeting four kinesins (KIF11/Eg5, KIF20A, KIF21A, KIF25), myosin 1G (MYO1G), myosin heavy chain 1 (MYH1) and tropomyosin 2 (TPM2) were identified as effective inducers of non-apoptotic cell death. The cell death induced by KIF11, KIF21A, KIF25, MYH1 or TPM2 siRNAs was preceded by lysosomal membrane permeabilization, and all identified siRNAs induced several changes in the endo-lysosomal compartment, i.e. increased lysosomal volume (KIF11, KIF20A, KIF25, MYO1G, MYH1), increased cysteine cathepsin activity (KIF20A, KIF25), altered lysosomal localization (KIF25, MYH1, TPM2), increased dextran accumulation (KIF20A), or reduced autophagic flux (MYO1G, MYH1). Importantly, all seven siRNAs also killed human cervix cancer (HeLa) and osteosarcoma (U-2-OS) cells and sensitized cancer cells to other lysosome-destabilizing treatments, i.e. photo-oxidation, siramesine, etoposide or cisplatin.



Named entity linking



Relation extraction

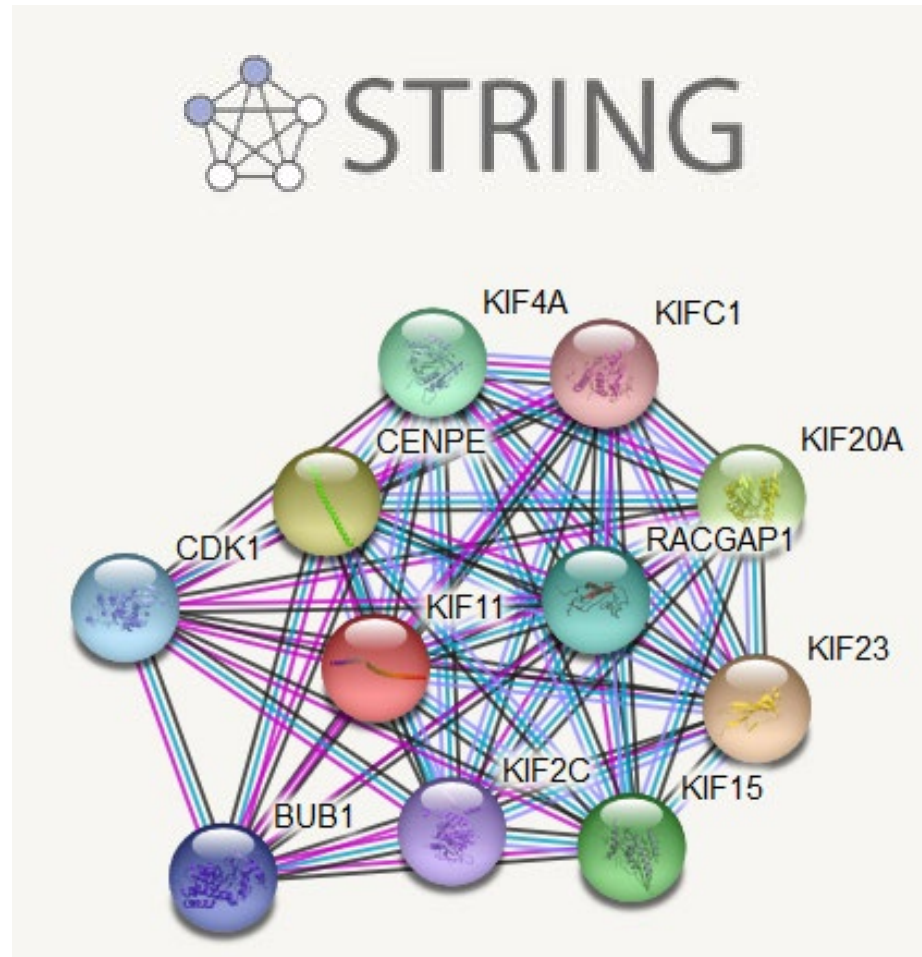
**Rapamycin inhibits mTOR, and
thereby affects autophagy.**

Entity 1: Rapamycin

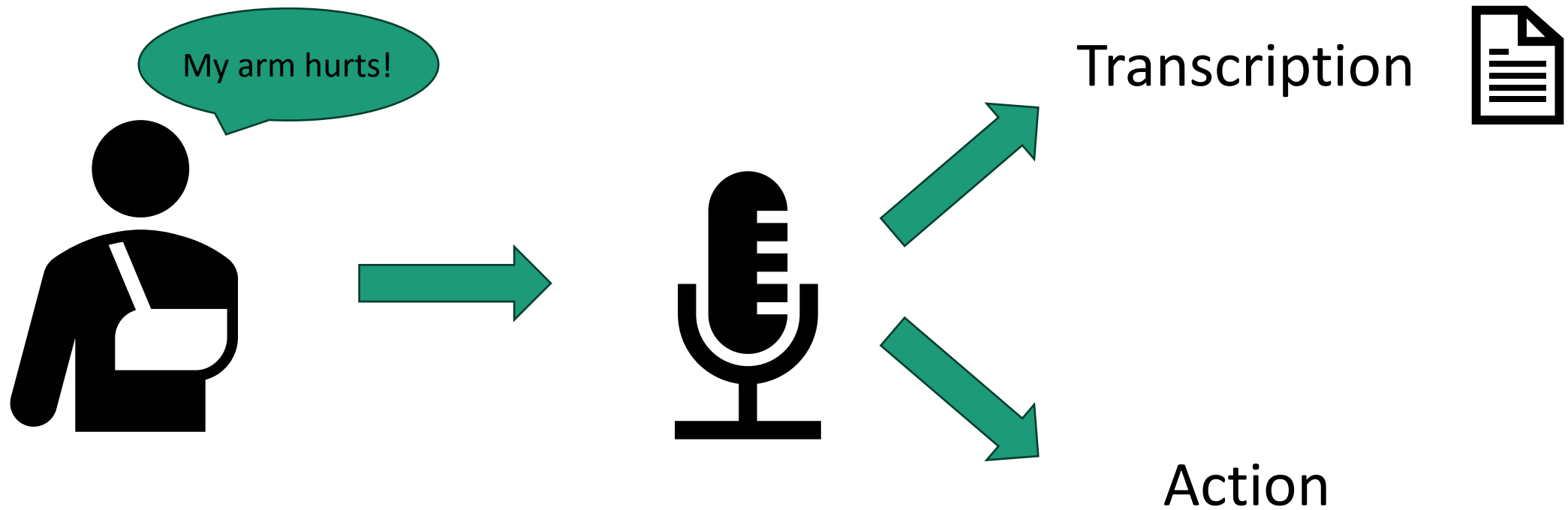
Entity 2: mTOR

Relation: inhibition

Relation extraction is used to populate protein-protein interaction databases



Speech recognition



Question answering

Question: Which disease conditions are treated with aspirin?

Context

The screenshot shows the DrugBank Online interface. At the top, the logo 'DRUGBANK Online' is on the left, and navigation links 'Explore', 'Data Library', 'Academic Downloads', and 'Interaction Checker' are on the right. A 'NEW' tag is next to 'Data Library'. Below the navigation bar is a dark sidebar with a search input field and three menu items: 'Identification', 'Pharmacology', and 'Interactions'. The main content area has a dark header with the word 'Aspirin' and a 'Drugs' tab. Below this, the 'Summary' section contains the text: 'Aspirin is a salicylate used to treat pain, fever, inflammation, migraines, and reducing the risk of major adverse cardiovascular events.'

DRUGBANK Online

Explore Data Library NEW Academic Downloads Interaction Checker

Drugs

Aspirin

Identification
Pharmacology
Interactions

Summary

Aspirin is a salicylate used to treat pain, fever, inflammation, migraines, and reducing the risk of major adverse cardiovascular events.

Answer: Pain, fever, inflammation and migraines are treated with aspirin.

Conversational AI: Chat bots

I cannot smile or
raise my right arm.



You probably have
a stroke. I am calling 112.



Conversational AI: Chat bots



You

what are the most common nlp tasks for deep neural networks?



ChatGPT

Natural Language Processing (NLP) involves a variety of tasks that can be tackled using deep neural networks. Some of the most common NLP tasks for deep neural networks include:

1. **Text Classification:**

- **Task:** Assigning predefined categories or labels to a given text.
- **Example:** Spam detection, sentiment analysis, topic categorization.

2. **Named Entity Recognition (NER):**

- **Task:** Identifying and classifying entities (such as names of people, organizations, locations) in text.
- **Example:** Extracting names of people and locations from news articles.

Mini quiz: NLP tasks

- Matching electronic health records to ICD-10 disease codes
- Identifying the major types of adverse effects of a new vaccine from reports submitted by doctors
- Finding all links between mutations and diseases reported in the literature
- Flagging high priority patients in emergency calls
- Building a treatment recommender that suggests drugs based on mutation status in sequencing reports

The technical side...



NLP is a challenging task!

She killed the man with the tie.

Why is NLP challenging?

- **Ambiguity**

- Homographs: We **ran** a Western blot to measure **RAN** levels.
- Homophones: to, two, too

- **Co-reference**

Cancer is one of the most common causes of death. **This disease...**

- **Synonymous expressions**

- This caused **cell death**./This led to **cellular demise**./This **killed the cells**./The **viability was greatly reduced**./The **cells were eradicated**.

- **Abbreviations**

State-of-the-art NLP models encode words based on their context

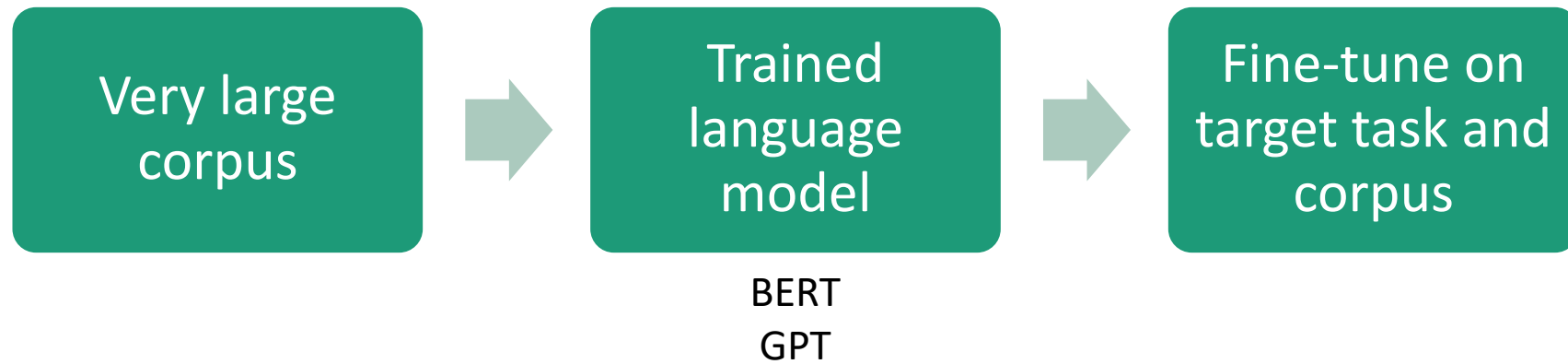
Occupation groups who spend many hours doing computer works often suffer from wrist pain related to **mouse** use.

mouse → vector a

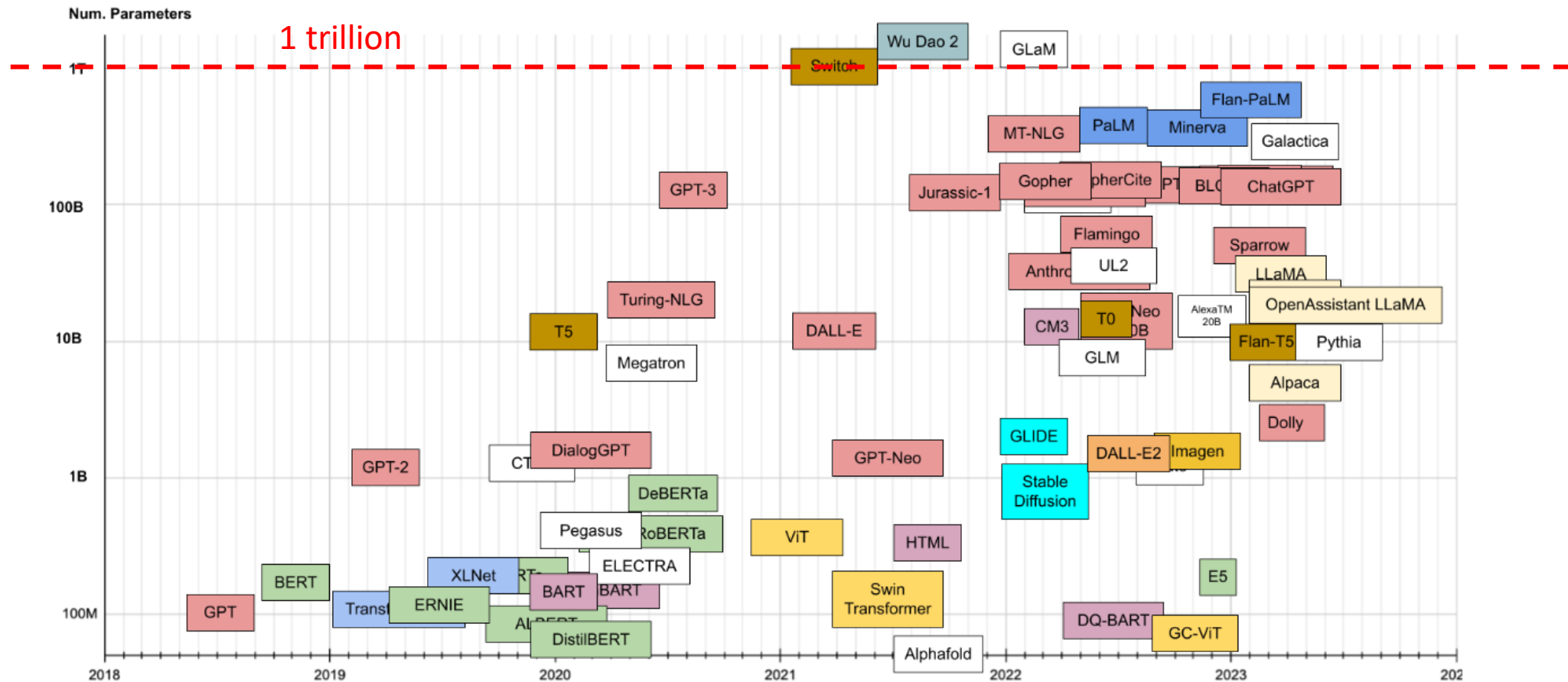
This study used a transgenic **mouse** model to evaluate the impact of TP53 mutations on lung cancer survival.

mouse → vector b

Transfer learning with large language models (LLMs)

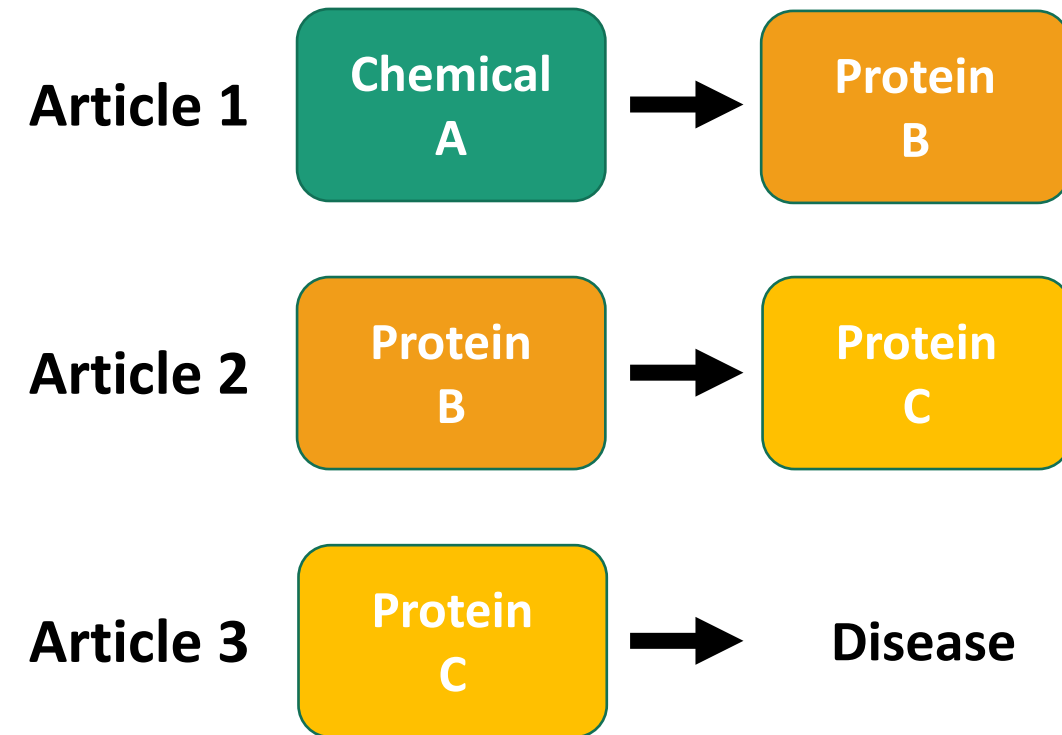


LLMs are becoming really large

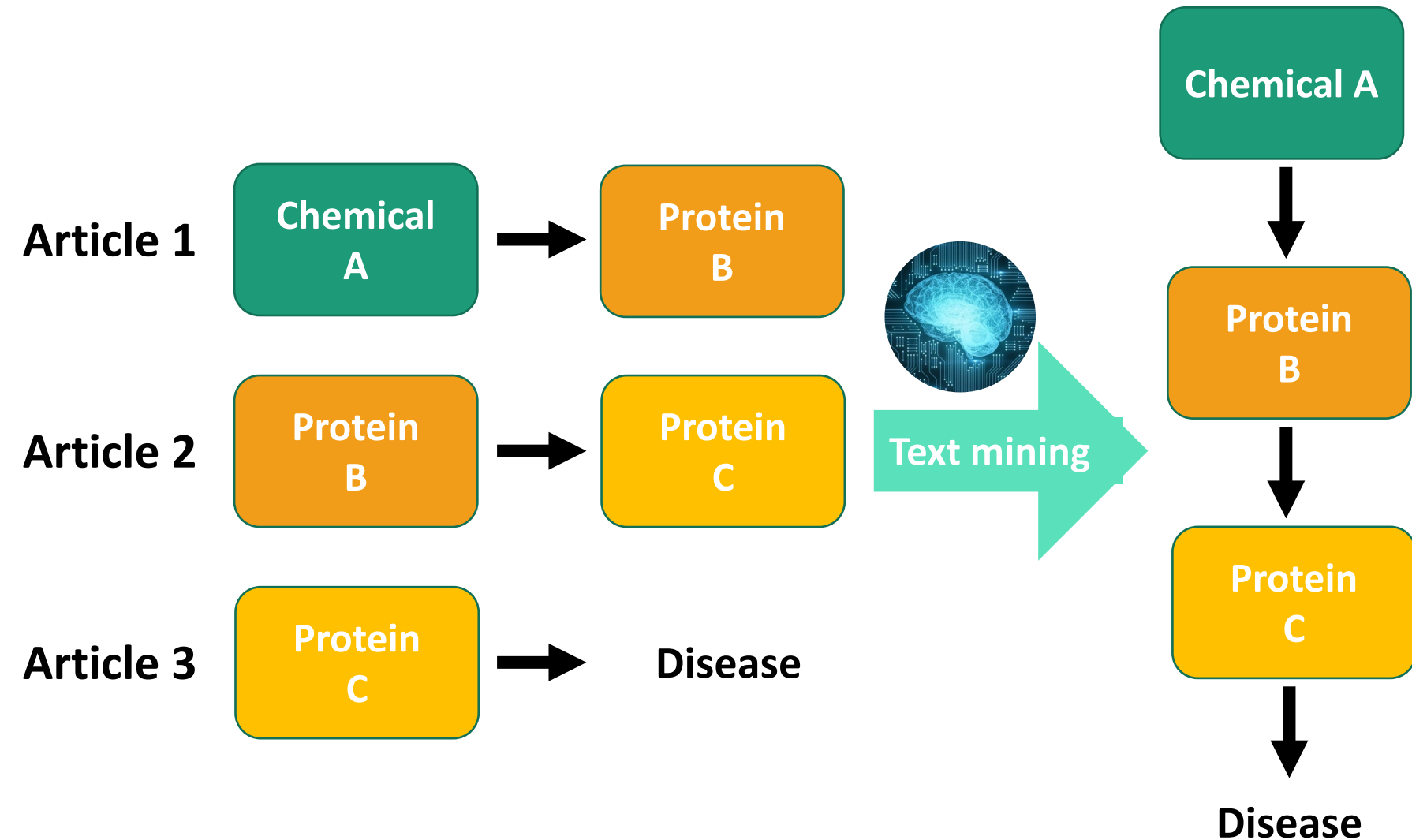


Research examples:
Information extraction

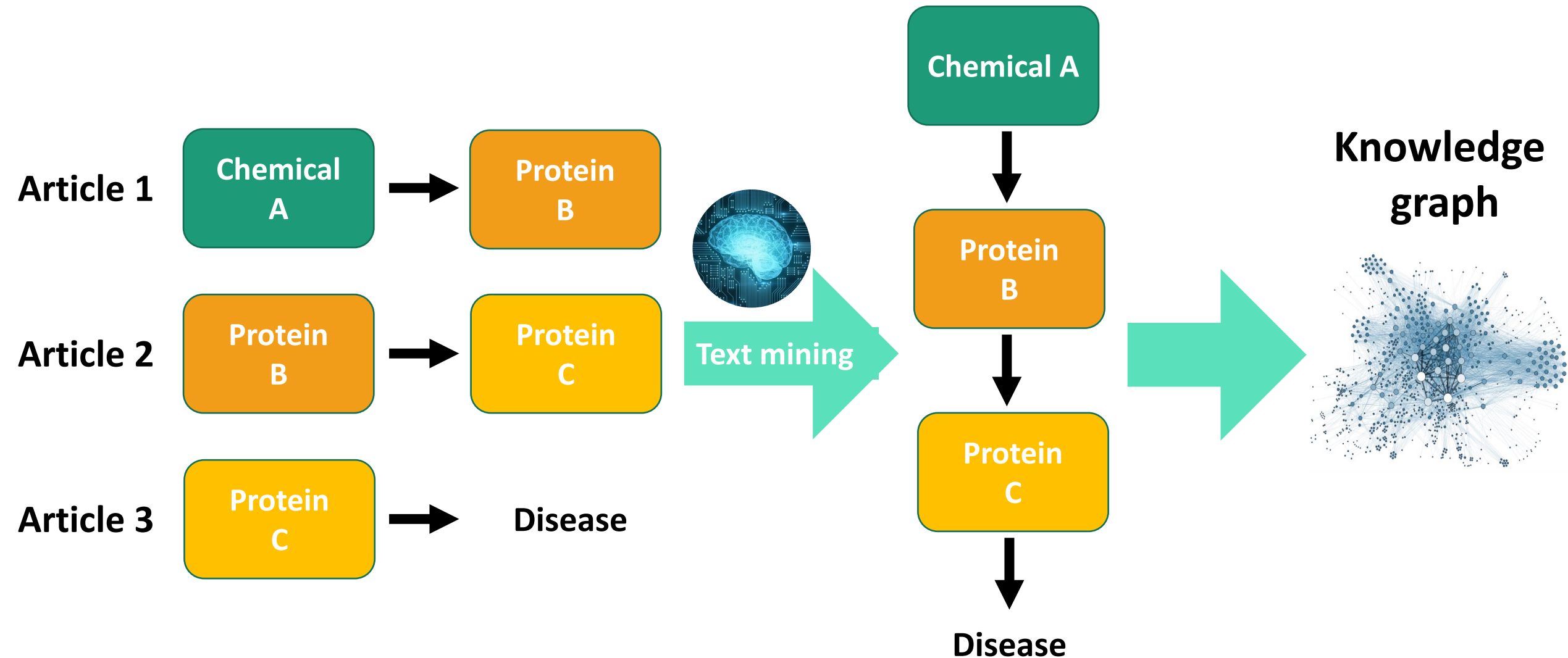
LLM- and dictionary-based text mining connects scattered pieces of information



LLM- and dictionary-based text mining connects scattered pieces of information



LLM- and dictionary-based text mining connects scattered pieces of information



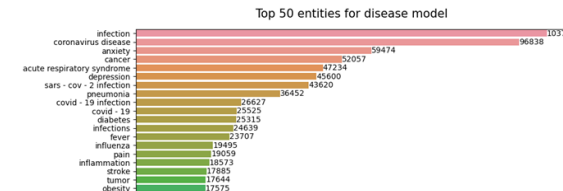
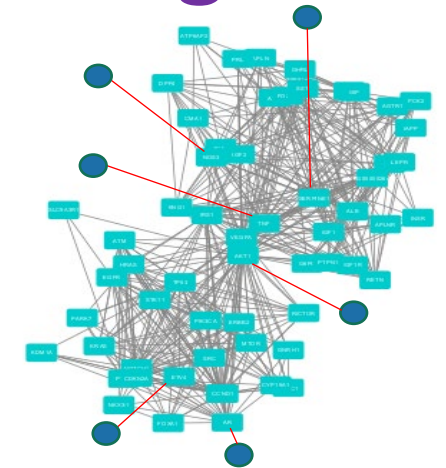
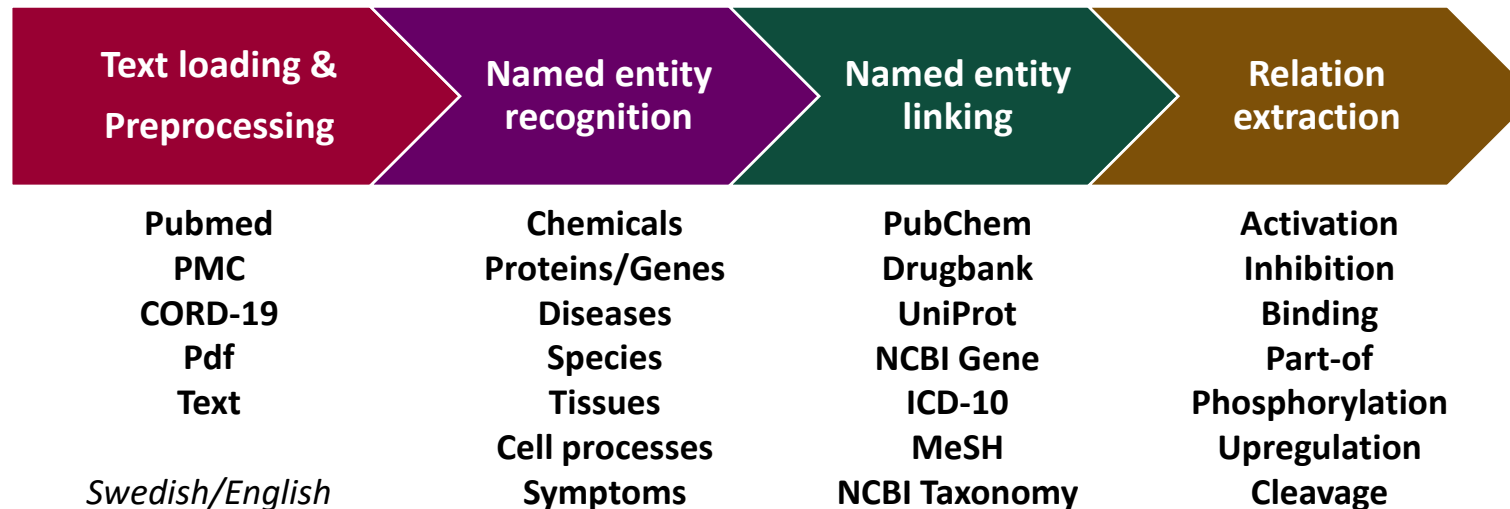
EasyNER: easy-to-use, customizable text mining with full traceability

EasyNER

<https://github.com/Aitslab/EasyNER>

SIRNAs targeting four kinesins (KIF11/Eg5, KIF20A, KIF21A, KIF25), myosin 1G (MYO1G), myosin heavy chain 1 (MYH1) and tropomyosin 2 (TPM2) were identified as effective inducers of non-apoptotic cell death. All seven killed human cervix cancer (HeLa) and osteosarcoma (U-2-OS) cells and sensitized cancer cells to other lysosome-destabilizing treatments, i.e. siramesine, etoposide or cisplatin.

PMID
34088762
43463464
57357822
54728368
14634568
24624647
25475427
45716478
26462477
24746274



SIRNAs targeting four kinesins (KIF11/Eg5, KIF20A, KIF21A, KIF25), myosin 1G (MYO1G), myosin heavy chain 1 (MYH1) and tropomyosin 2 (TPM2) were identified as effective inducers of non-apoptotic cell death. All seven killed human cervix cancer (HeLa) and osteosarcoma (U-2-OS) cells and sensitized cancer cells to other lysosome-destabilizing treatments, i.e. siramesine, etoposide or cisplatin.

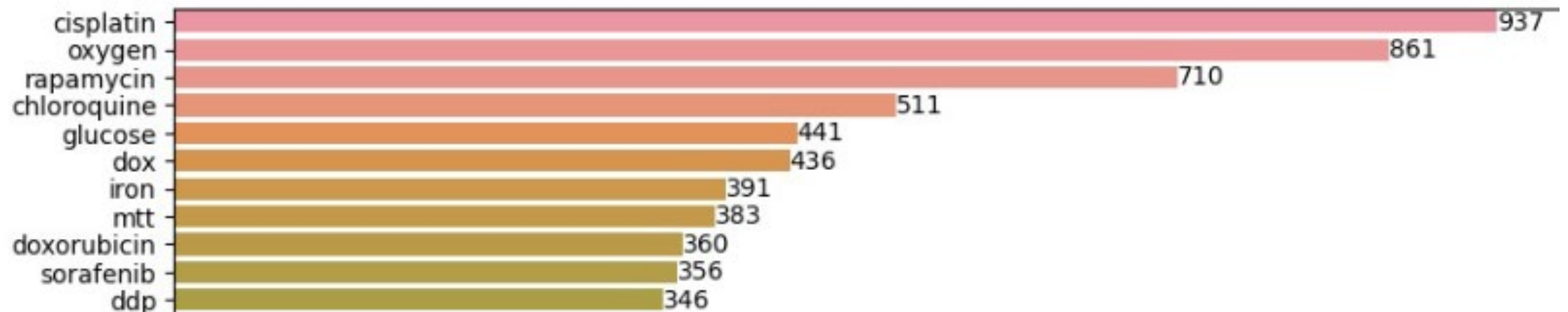
Example:

Chemicals linked to autophagy in cancer

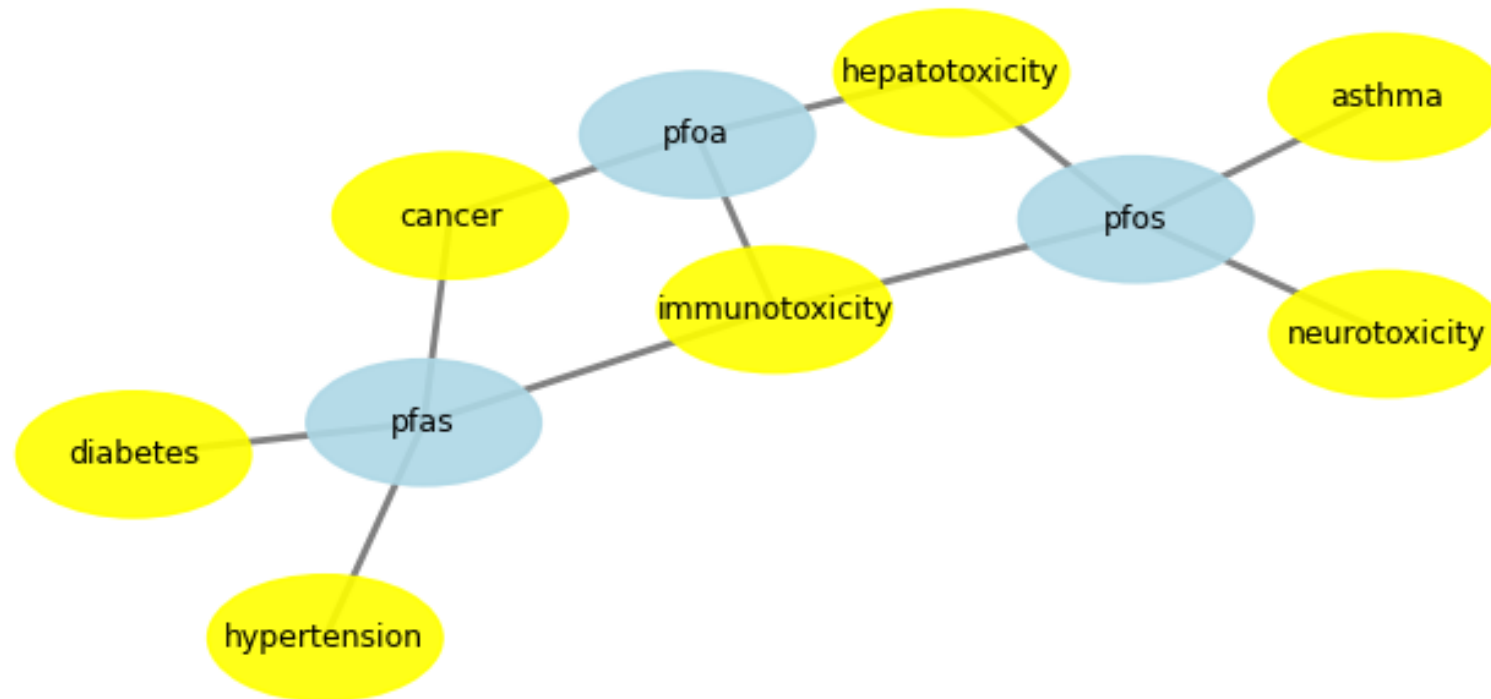
PubMed search: autophagy AND cancer



EasyNER: chemical detection



Information linkage in knowledge graphs



Analysis of electronic health records

Tidigare sjukdommar: DM2 insulinbehandlad. Hösnuva (tar cetrizin ibland).
Aktuellt: Insjuknat för tre dagar sedan med andnöd vid gång. Tillkomst av hosta under gårdagen. Kräkt en gång idag. Ej noterat feber. Ingen buksmärta eller ÖNH symptom.

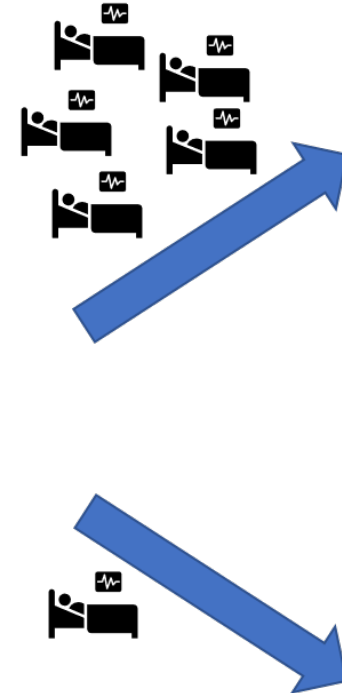
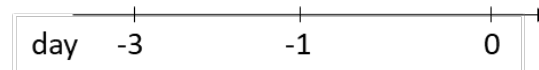


Tidigare sjukdommar: DM2 DIS insulinbehandlad DRUG. Hösnuva DIS (tar cetrizin DRUG ibland).
Aktuellt: Insjuknat för tre dagar sedan TIME med andnöd SYMP vid gång. Tillkomst av hosta SYMP och yrsel SYMP under gårdagen TIME. Kräkt SYMP en gång idag TIME. Ej NEG noterat feber SYMP. Ingen NEG buksmärta SYMP eller ÖNH symptom SYMP.

Disease: E11.2, J30

Drugs: DB00030, DB00341

Timeline: dyspnea cough, vertigo vomiting



Epidemiology/
clinical
research



Personalized
treatment
decision

Summary: common tasks for NLP

- Summarization
- Translation
- Clustering
- Text classification
- Sentiment analysis
- Named entity recognition
- Named entity linking
- Speech recognition
- Question answering
- Chat bots