

EMPOWERING BIOMEDICAL LITERATURE UNDERSTANDING THROUGH LLMS AND HPC

Presented By: Enrique Noriega

Data Science Institute

Dept of Computer Science

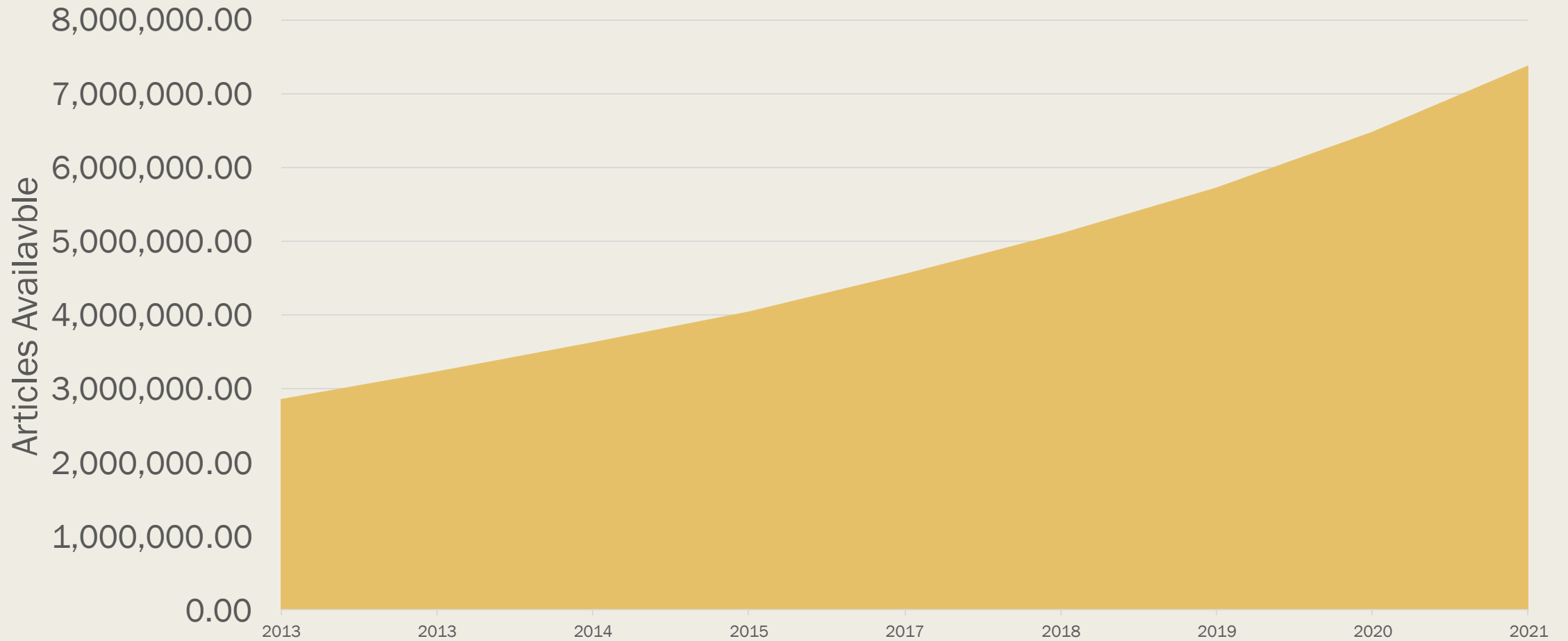


- Researcher @ Data Science Institute, Dept of Computer Science
- Specialized in Natural Language Processing (NLP) and building applications with Large Language Models (LLMs)
 - Specially in the bio medical domain

MOTIVATION

The information overload

PubMed Size per Year



Keeping Up with Literature

- Research productivity increases overtime
- Use automated tools to keep up
 - *Search for relevant articles*
 - *Obtain insights automatically*
 - *Draw hypotheses based on the literature*

High-level Retrieval Augmented Generation



NATURAL LANGUAGE PROCESSING

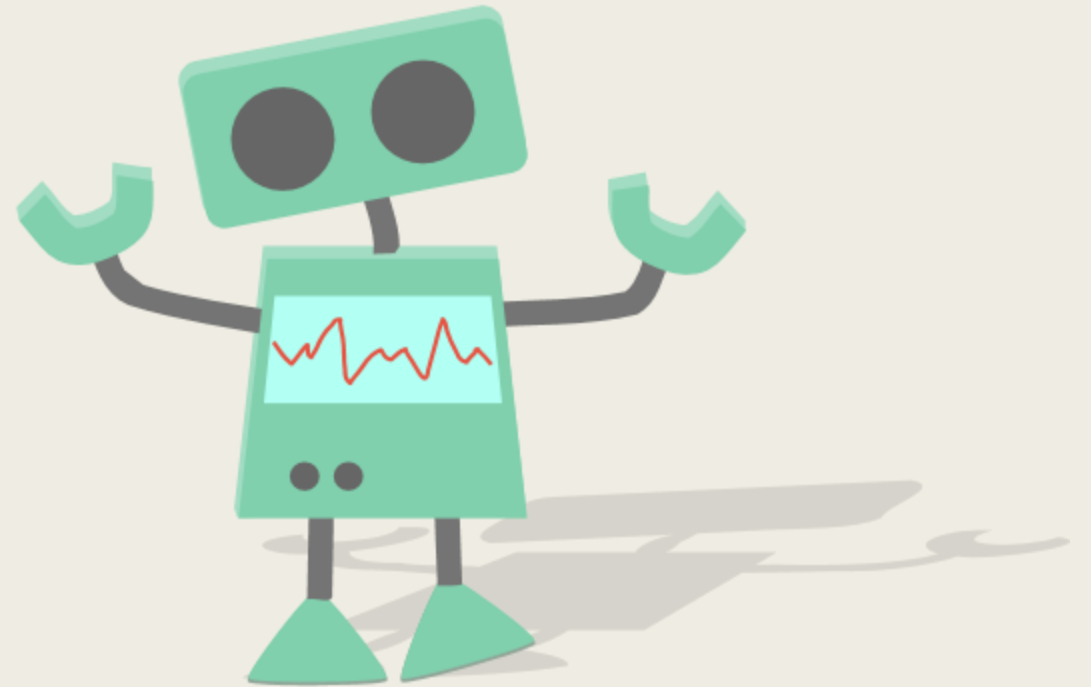
The AI tools used to ingest human-written text

What is Natural Language Processing?

NLP, in part, is the application of computational techniques to understand texts written by people

Examples

- Sentiment classification
- Spam detection
- Machine translation
- Question answering



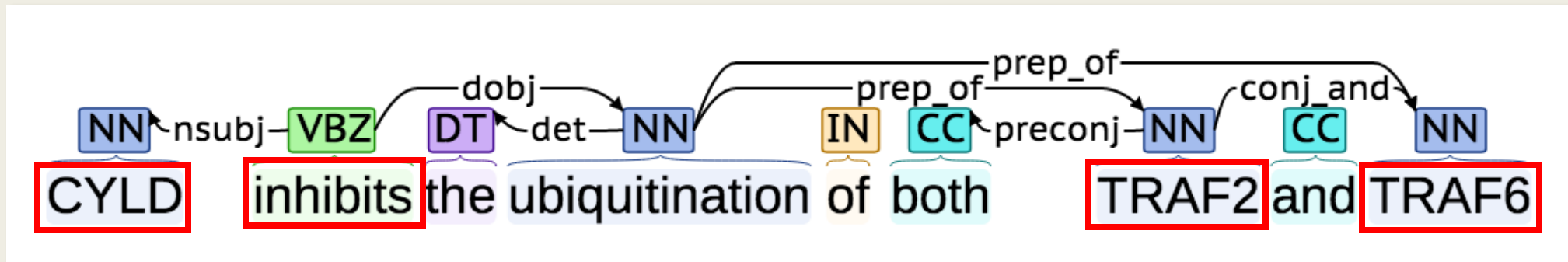
INFORMATION EXTRACTION

Drawing computer data from human language

*Enrique is talking about AI tools to the
biomedical community of
The University of Arizona*

Enrique is talking about AI tools to the
biomedical community of
The University of Arizona

Rule-based Information Extraction

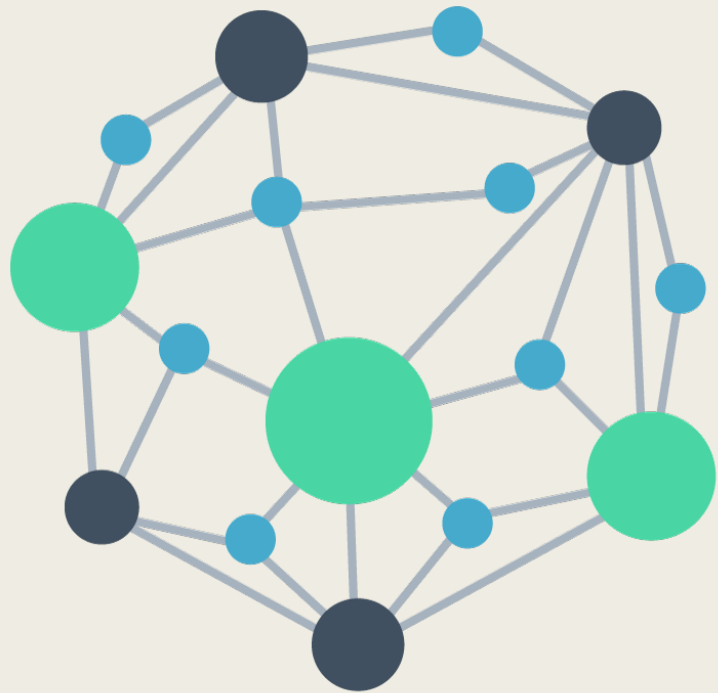


1. Match a Trigger: *inhibits*
2. Find biochemical entity: *CYLD*
3. Find a second biochemical entity: *TRAF2* or *TRAF6*
4. Constructing the biochemical event(s)

Information Extraction

- Commonly trained using supervised machine learning
- Multi-step process:
 - *Concept recognition*
 - *Normalization to database/ontology*
 - *Relation detection*

Information Extraction



INFORMATION RETRIEVAL

Finding a needle in a haystack

“Finding a needle in a haystack”

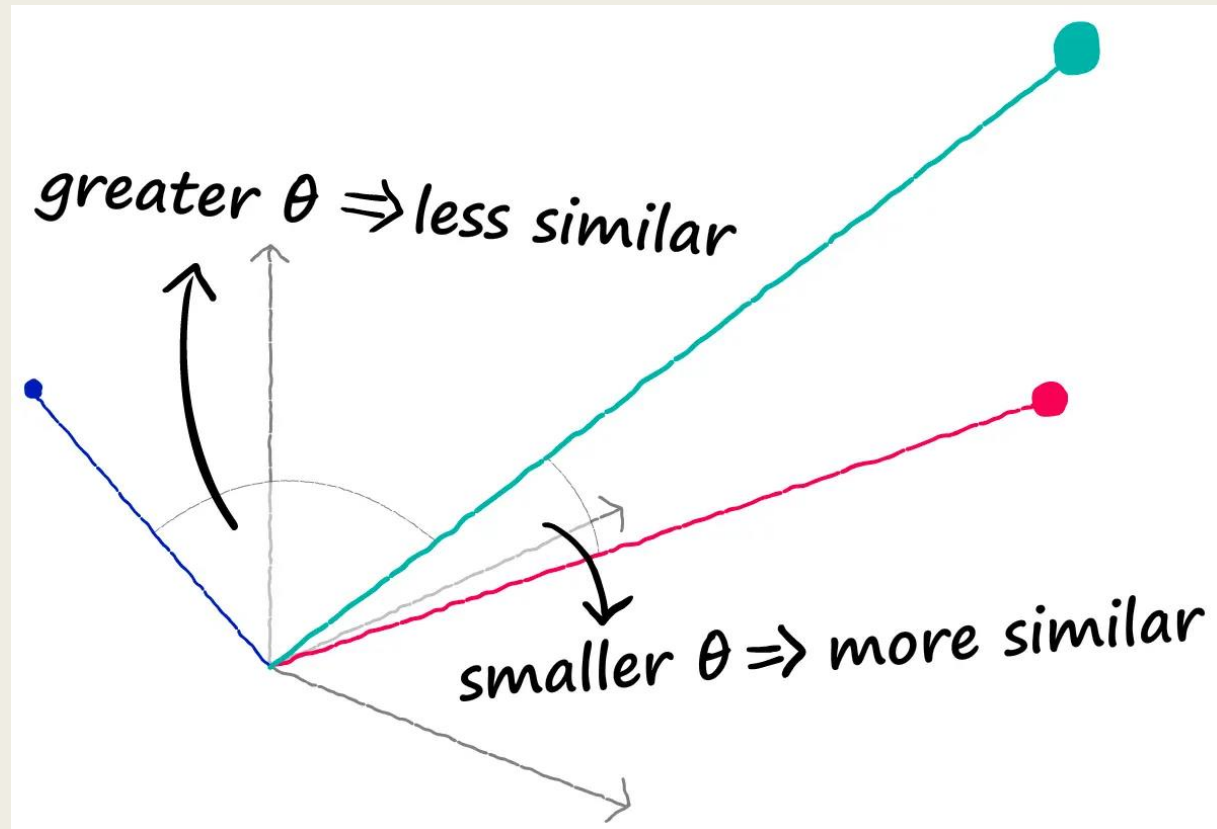
- Think about search engines
- Use *queries* to search for relevant documents
 - *Google*
 - *Bing*
 - *For the life sciences: PubMed*
- Traditionally: Keyword/phrase search approach
- Contemporary: *Word embedding* search approach
- Hybrid search: Combine keyword + embedding approaches

Information Retrieval

Token	Document Id
Harry	1, 2
Potter	1, 2
And	1, 2
The	1, 2
Half	1
Blood	1
Prince	1
Deathly	2
Hallows	2

Inverted index

Keyword search



Word embedding search

LARGE LANGUAGE MODELS

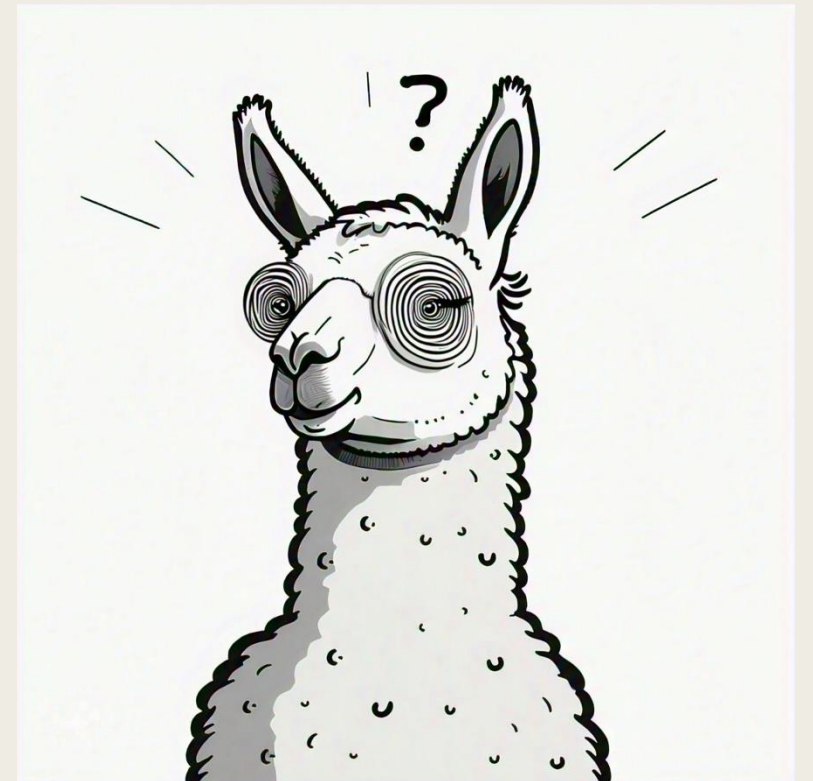
Eloquently eliciting insights based on data

What is a Large Language Model?

- “Giant” computational program that generates language in a coherent, human-like fashion.
- Fundamentally: Predicts the next word based on its prior context
- Requires significant amount of computing resources
- With enough *parameters* and *training data* exhibit “emergent” capabilities: Reasoning over, understand and generate human language.

LLMs continued

- They have limitations:
 - *Hallucinations!*
 - *Stubbornness*
 - *Expensive to maintain*
- Mitigations:
 - *Retrieval Augmented Generation*
 - *Guardrails*
 - *Agentic behavior*



INTEGRATING AI

How everything fits together

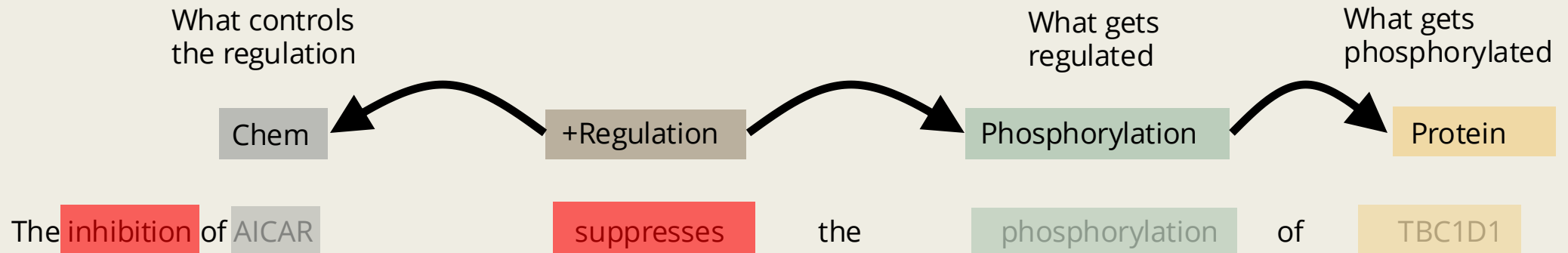
High-level Retrieval Augmented Generation



Information Extraction in the Biomed Domain

- Symbolic systems:
 - *Use rules*
 - *Interpretable*
 - *Higher precision – lower recall*
 - *High maintenance*
- Statistical systems:
 - *Base off Machine Learning (i.e. LLMs)*
 - *Black box*
 - *Higher recall – lower precision*
 - *Low maintenance*

IE in the Biomedical Domain



Biomedical IE Systems

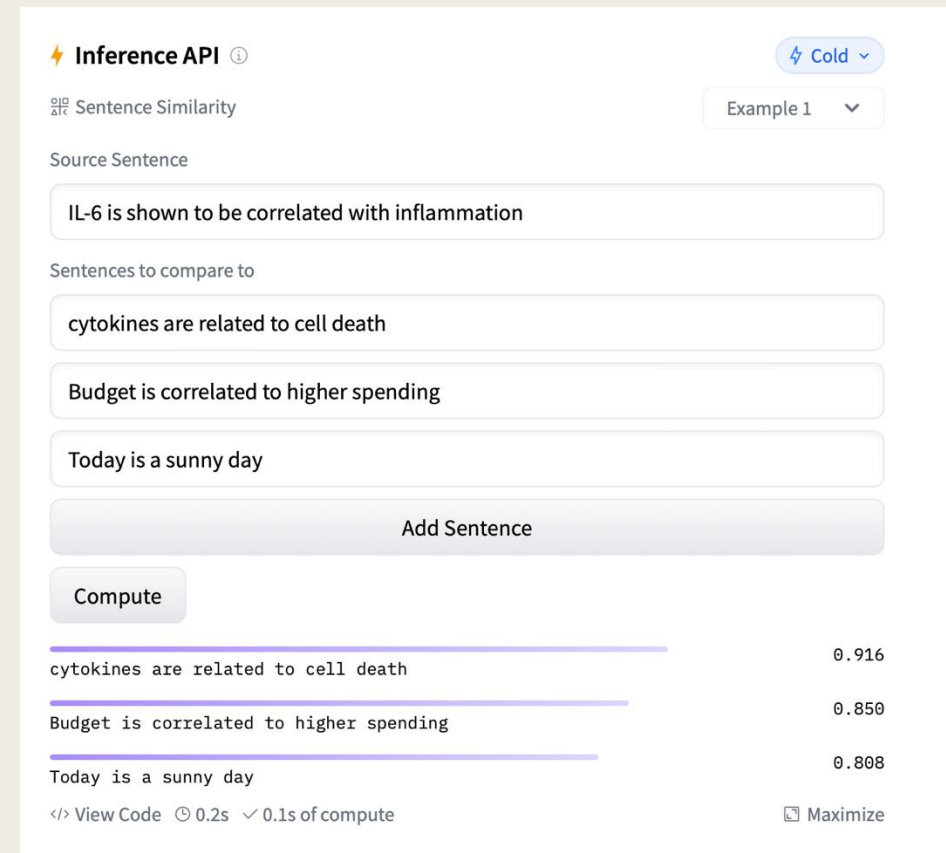
- Rule based systems:
 - *REACH* – Cancer biology P-2-P annotator
 - *MetaMap* – UMLS annotator
- Statistical based systems:
 - *SciSpacy* – NLP pipeline trained on MedMentions dataset
 - *GENIA Tagger* – Similar but trained on MEDLINE abstracts

Information Retrieval Systems

- Elastic Search
 - *Keyword based retrieval*
 - *Scalable and highly optimized*
- Weaviate, Pinecone
 - *Embedding based retrieval*
 - *Works with any encoder-based language model*

Encoder Language Models

- Similar to LLMs in principle
- Create a numerical representation of phrases
- Represent semantic similarity
- *Sentence Transformers*
 - *Specialized in biomedical domain: S-PubMedBert-MS-MARCO*

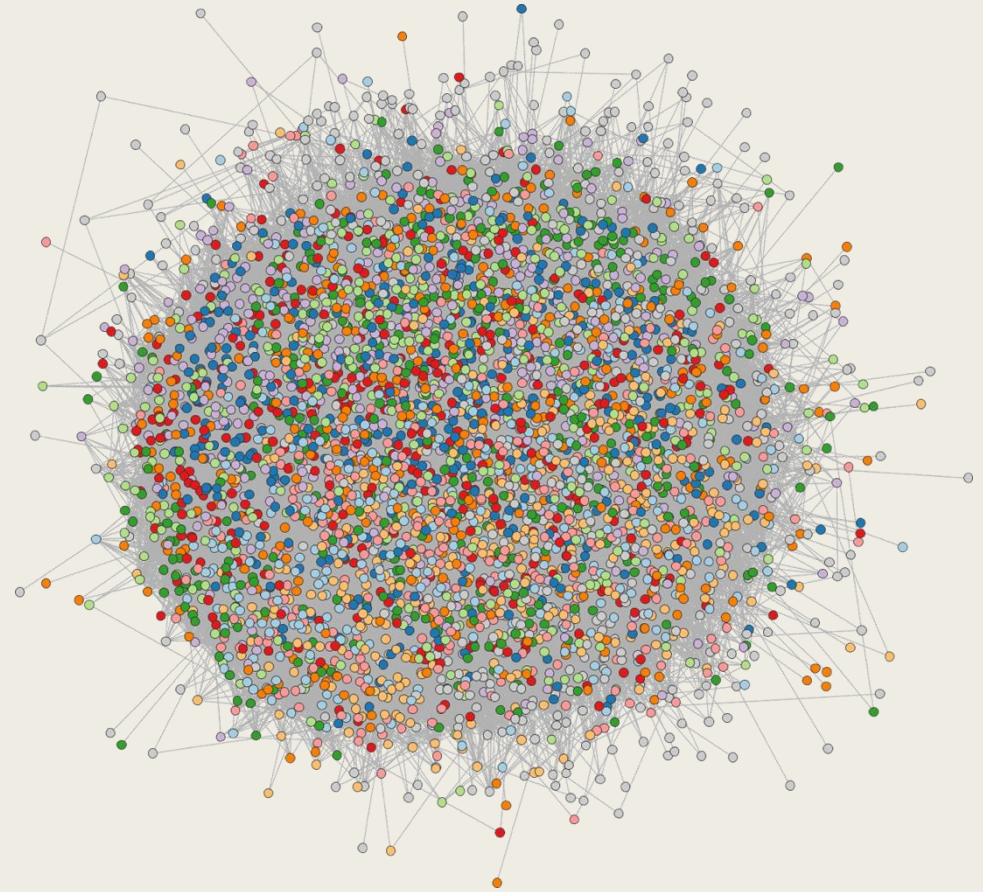


The screenshot shows the Inference API interface for Sentence Similarity. The source sentence is "IL-6 is shown to be correlated with inflammation". Three comparison sentences are listed: "cytokines are related to cell death" (0.916), "Budget is correlated to higher spending" (0.850), and "Today is a sunny day" (0.808). The interface includes a "Compute" button and a "Maximize" option.

Sentence	Similarity Score
IL-6 is shown to be correlated with inflammation	-
cytokines are related to cell death	0.916
Budget is correlated to higher spending	0.850
Today is a sunny day	0.808

Example: Biomedical Interactions Visualization

- Efficiently search and locate:
 - *Mechanistic interactions*
 - *Underlying textual evidence*
 - *The pointer to the source of the information*
- Reduce the *hairball* effect
- Search paradigm:
 - *Narrow down search space*
 - *Iteratively search*



Structural Search

Search, navigate, and visualize exploiting underlying network structure

Explore Biomedical Literature

[Interactions Overview](#) | [Graphic Overview \(IL-6\)](#) | [Graphic Overview \(TNF-FAT\)](#) | [Search Evidence](#) | [Structured Evidence Search](#)

Overview of *Interleukin-6* - uniprot:P05231

Type an entity name or id...

Sort by: Filter by:

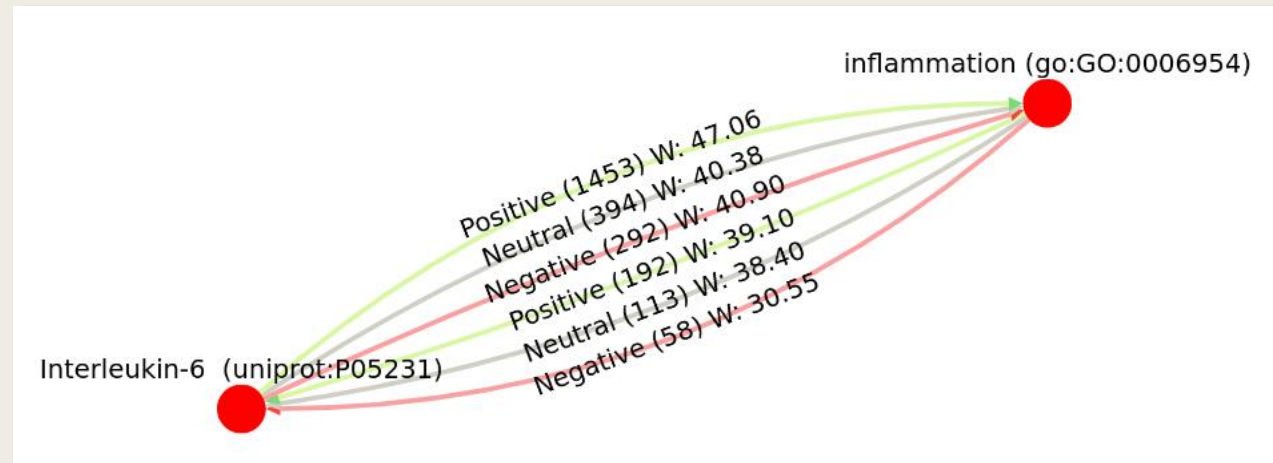
Weighting

Columns to display:

Influenced Reciprocal Influence

Influenced By:	Reciprocal With:	Influence:
Biological Process - (422)	Biological Process - (175)	Biological Process - (32)
Cells, Organs and Tissues - (24)	Cells, Organs and Tissues - (5)	pre-frailty (frailty:FR00018) - F: 3 - W: 9.29 - D: 3
Chemicals - (109)	Chemicals - (354)	ltd (go:GO:0060292) - F: 2 - W: 1.10 - D: 1
Diseases - (256)	ros (chebi:CHEBI:26523) - F: 305 - W: 17.09 - D: 113	skeletal muscle hypertrophy (go:GO:0014734) - F: 2 - W: 4.56 - D: 2
Proteins or Gene Products - (739)	lipopolysaccharide (chebi:CHEBI:16412) - F: 289 - W: 13.53 - D: 12	activation of creb (go:GO:0032793) - F: 1 - W: 0.69 - D: 1
Aromatase (uniprot:P11511) - F: 34 - W: 14.07 - D: 22	curcumin (chebi:CHEBI:3962) - F: 267 - W: 13.15 - D: 3	ccl3 production (go:GO:0071608) - F: 1 - W: 0.69 - D: 1
Cadherin-2 (uniprot:P19022) - F: 17 - W: 16.40 - D: 13	glucose (chebi:CHEBI:17234) - F: 263 - W: 15.51 - D: 97	cellular localization (go:GO:0051641) - F: 1 - W: 0.69 - D: 1
Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial (uniprot:P08559) - F: 16 - W: 17.90 - D: 3	no (pubchem:24822) - F: 173 - W: 16.13 - D: 85	chondrocyte hypertrophy (go:GO:0003415) - F: 1 - W: 8.30 - D: 1
	ligand (chebi:CHEBI:52214) - F: 154 - W: 14.92 - D: 52	cit (go:GO:0106106) - F: 1 - W: 0.69 - D: 1
		dna demethylation (go:GO:0080111) - F: 1 - W: 0.69 - D: 1

Node-Link Visualization



Evidence Panel

Evidence: Close

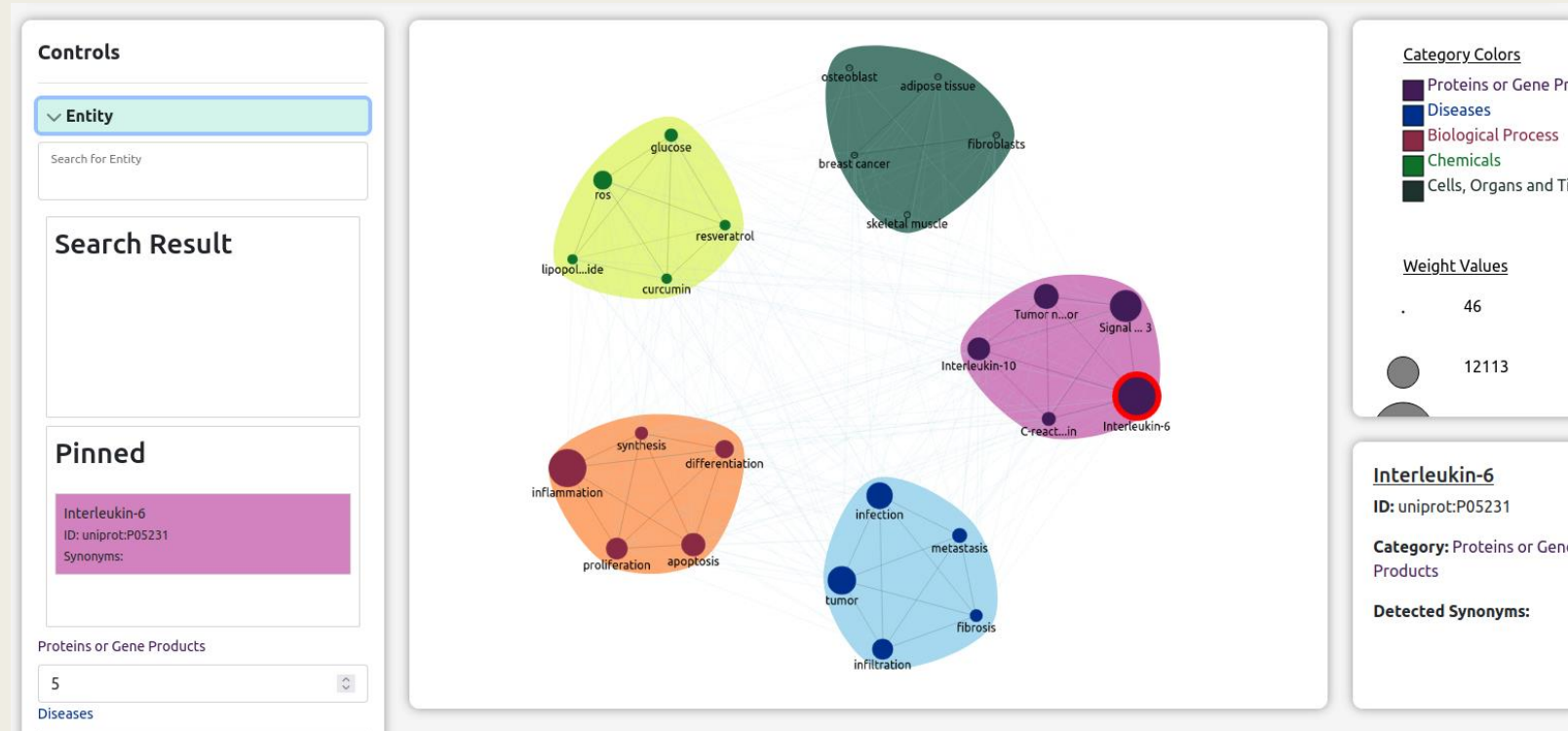
- (14.29) [PMC3919204](#): **Interleukin-6 Signaling Drives Fibrosis in Unresolved Inflammation** .
- (12.56) [PMC7402632](#): Overall , these results support a framework in which an ongoing inflammation is driven by PAMPs and DAMPs. In this study , we identified an impaired type I IFN response in severe and critical COVID-19 patients , accompanied by high blood viral load and an excessive NF-kappa B - driven **inflammatory response associated with increased TNF-alpha and IL-6** .
- (6.26) [PMC4391624](#): Because acute **inflammation is associated with increased levels of other cytokines** , we examined the effects of these cytokines on the tonic current .
- (3.60) [PMC3436816](#): This gene set association corroborates previously identified **links between PD and inflammation** (Knott et al . , 2000) and reports of elevated **levels of interleukin-6** in the cerebrospinal fluid of PD patients (Blum-Degen et al . , 1995) .
- (3.60) [PMC3940382](#): Although IL-6 is considered necessary for initiation of the acute-phase response , IL-6 and its receptor have pleiotropic effects with both proinflammatory and anti-inflammatory activity , with **IL-6 having been shown to act in an anti-inflammatory manner in previous models of LPS induced lung inflammation** .

Tag evidence as: Close

Relevant

Add new

Graphical Overview



Textual Search

Enter your query
causes inflammation

Max results
500

Search

- [PMC5976511](#): *Noxa* deficiency **causes** increased *inflammation*.
- [PMC5342343](#): Binge *ethanol* exposure **causes** pancreatic *inflammation*.
- [PMC4195784](#): *Rab11a* ablation in intestinal epithelia **causes** *inflammation*.
- [PMC7820820](#): In sepsis , inflammation and coagulation are cross linked , and *inflammation* **causes** *coagulation* activation.
- [Source](#): *IL-6* **causes** cardiac *inflammation* and suppresses AMPK.
- [Source](#): *IL-6* causes cardiac inflammation and **suppresses** *AMPK*.

FUTURE DEVELOPMENTS

Where is the research moving toward?

Revisiting RAG

Consider the following context passage. Put together the signaling pathway from IL-6 to inflammation as a JSON array, where each element contains the endpoints and the text that backs the extraction. Write your output as a markdown block of JSON code.

Interleukin-6 (IL-6) regulates inflammation through two primary signaling pathways: classical signaling and trans-signaling.

In classical signaling, IL-6 binds to its membrane-bound receptor (mIL-6R), forming a complex with glycoprotein 130 (gp130), which triggers downstream pathways like JAK/STAT, MAPK, and PI3K.

These pathways promote the expression of genes involved in inflammation, such as acute-phase proteins and anti-inflammatory mediators like SOCS3, which also provide negative feedback. Classical signaling is restricted to cells expressing mIL-6R, such as hepatocytes and immune cells, and is generally associated with protective responses in acute inflammation. In trans-signaling, IL-6 binds to soluble IL-6 receptors (sIL-6R), enabling it to act on cells lacking mIL-6R, such as endothelial cells and fibroblasts, thereby amplifying inflammatory responses.

This pathway often drives pathological inflammation, contributing to chronic diseases like rheumatoid arthritis and inflammatory bowel disease.

Regulatory mechanisms, including SOCS3-mediated inhibition and receptor internalization, maintain signaling balance. Targeting IL-6 signaling, particularly trans-signaling, with therapeutic agents like tocilizumab has proven effective in mitigating chronic inflammatory conditions and cytokine release syndrome.

Revisiting RAG



IL-6 Signaling Pathway to Inflammation

json

Copy Code

```
[
  {
    "endpoints": "IL-6 -> mIL-6R",
    "text": "IL-6 binds to its membrane-bound receptor (mIL-6R)"
  },
  {
    "endpoints": "IL-6/mIL-6R -> gp130",
    "text": "forming a complex with glycoprotein 130 (gp130)"
  },
  {
    "endpoints": "IL-6/mIL-6R/gp130 -> JAK/STAT, MAPK, PI3K",
    "text": "which triggers downstream pathways like JAK/STAT, MAPK, and
```

```

    "endpoints": "JAK/STAT, MAPK, PI3K -> Gene Expression",
    "text": "These pathways promote the expression of genes involved in i
  },
  {
    "endpoints": "IL-6 -> sIL-6R",
    "text": "IL-6 binds to soluble IL-6 receptors (sIL-6R)"
  },
  {
    "endpoints": "IL-6/sIL-6R -> Endothelial cells, Fibroblasts",
    "text": "enabling it to act on cells lacking mIL-6R, such as endothel
  },
  {
    "endpoints": "IL-6/sIL-6R/Endothelial cells, Fibroblasts -> Inflammat
    "text": "thereby amplifying inflammatory responses"
  }
]
```

Revisiting RAG

Make a succinct summary of the interactions described in the JSON array



IL-6 Signaling Pathway Summary IL-6 interacts with its membrane-bound receptor (mIL-6R) and soluble receptor (sIL-6R) to trigger two primary signaling pathways:

- **Classical Signaling:** IL-6/mIL-6R binds to gp130, activating JAK/STAT, MAPK, and PI3K pathways, which promote inflammation in cells expressing mIL-6R (e.g., hepatocytes and immune cells).
- **Trans-Signaling:** IL-6/sIL-6R acts on cells lacking mIL-6R (e.g., endothelial cells and fibroblasts), amplifying inflammatory responses.



Large Language Models

- LLMs exhibit understanding and reasoning-like capabilities
 - *Result of training them to follow instructions*
- Two avenues:
 - *Fine-tuning*
 - Requires significant data and engineering skills
 - Achieves the best results
 - *In Context Learning*
 - Allows to set *an example* for an LLM to imitate
 - Improves performance but has limitations

WHERE TO DO IT?

Computing resources for the university's community

High Performance Computing

- Ideal for highly parallelizable jobs
 - *Scale to hundreds of CPUs*
 - *(Some) GPU access*

<https://hpcdocs.hpc.arizona.edu>

Welcome to the UArizona HPC Documentation Site

Introduction

The University of Arizona offers High Performance Computing (HPC) resources in the Research Data Center (RDC), a state-of-the-art facility that hosts our large computer clusters. HPC services are available at no cost to researchers. Each faculty member is eligible for a free standard allocation of [CPU time](#) and [storage space](#).

This documentation site provides technical details relevant to using our HPC system. Whether you are just starting your journey into computational sciences or are a seasoned programmer, we hope you will find something useful in these pages. This site is managed by the HPC Consult team. Please [contact us](#) if you have questions or comments about the content of this site.



CyVerse

- Ideal for data storage and sharing
 - *iRods*
- Cloud-based workflows
- Sharing *applications*
 - *Discovery Environment*
 - *VICE applications*

<http://cyverse.org>

The screenshot displays the CyVerse Discovery Environment dashboard. At the top, there is a navigation bar with the 'DISCOVERY ENVIRONMENT' logo, a search bar, and user profile information. A sidebar on the left contains various navigation icons. The main content area features a 'Tour the Discovery Environment' section with a 'Take a Tour' button. Below this is the 'Resource Usage' section, which indicates the current subscription tier is 'Pro' and shows the effective period from 2024-01-19 to 2025-01-19. Two bar charts are present: 'Data Storage' showing 2.52% usage of a 3.0 TiB limit, and 'CPU Consumption' showing 46.67% usage of a 20,000 Core Hours limit. At the bottom, 'Analyses Stats' shows a total of 2 analyses, with 2 completed and 0 canceled.

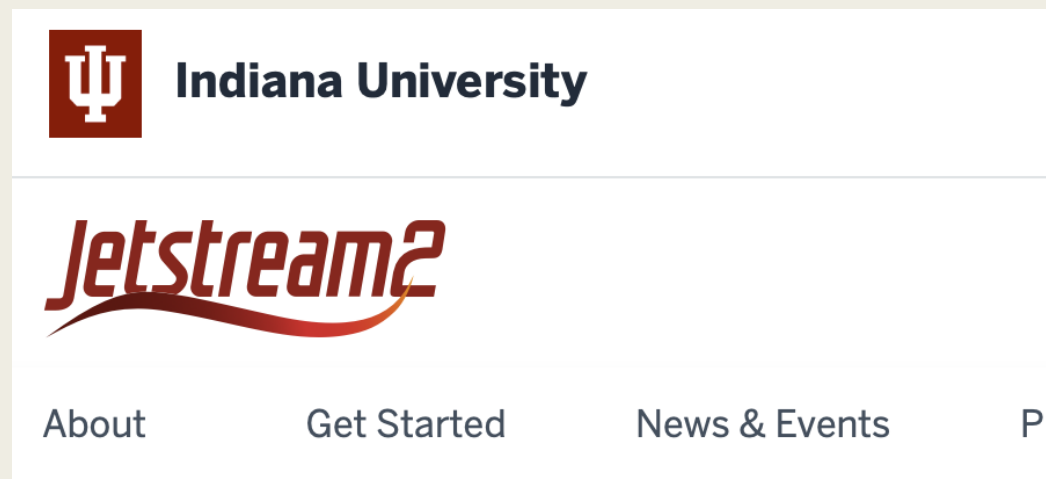
Resource	Usage	Limit
Data Storage	2.52%	3.0 TiB
CPU Consumption	46.67%	20,000 Core Hours

Analyses Stats	Count
Completed	2
Canceled	0

Jetstream 2

- Free access to GPUs for training
 - *Competitively allocated through a proposal*
- Cloud based workflows
- Compatible with Cacao (Cyverse's orchestration software)

<https://jetstream-cloud.org>



THANK YOU