Introduction au Traitement du Langage Naturel en Bio-informatique

Course Objective

- Understand the fundamentals of natural language processing
- Describe its application in bioinformatics.
- Explore specific NLP methods adapted to the analysis of biology and genomics.

Session 1: Introduction to Fundamental NLP Concepts
- General overview of NLP and its applications in bioinformatics.
- Fundamental NLP concepts: tokenization, tagging, classification.

Session 2: Representation of Sequential Data in NLP
- Data vectorization: embeddings, representations.
- Representation techniques adapted to the specificities of biological data.

Session 3: NLP Models for Biological Analysis
- Introduction to NLP models: RNN, CNN, transformers.
- Introduction to kernel methods for structured data analysis (sequences and trees).
- Adaptation and creation of models for the analysis of biological data.

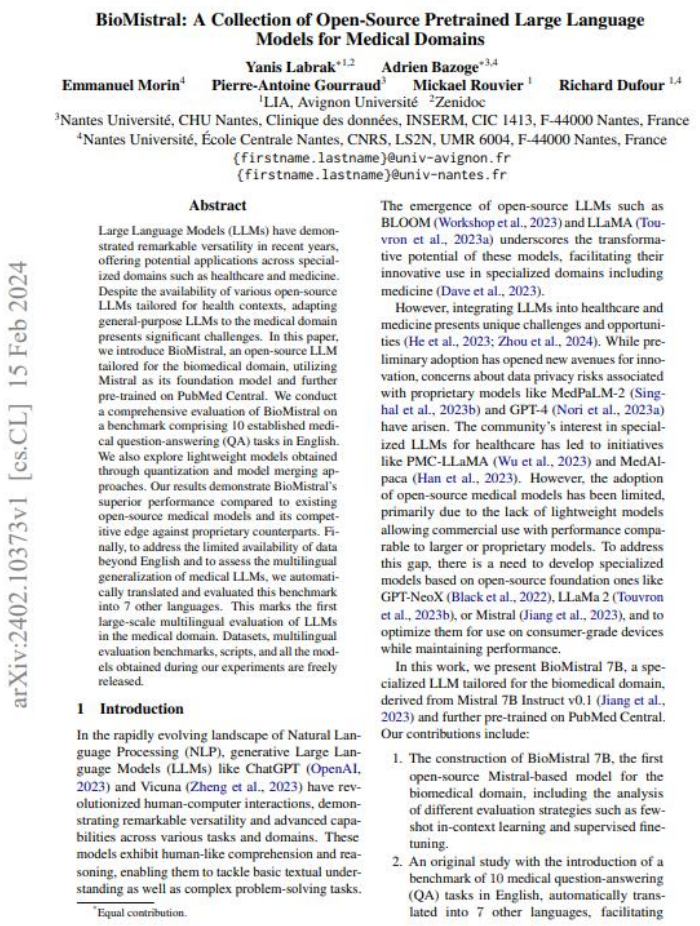
Session 4: Biological Corpora and Data Preprocessing
- Types of biological data: DNA sequences, proteins, scientific articles, etc.
- Data preprocessing and Tasks.

Outline for today

- Why Natural Language Processing?
- A Brief History of NLP
- What is Natural Language Processing?
  - Modelling Framework
  - Tokenization as a first-step task
  - Overview of NLP Tasks
- How to tackle any NLP problem?

BioMistral

- Large Language Models (LLMs) exhibit versatile potential, especially in specialized domains like healthcare and medicine.
- Despite various open-source LLMs tailored for health contexts, adapting general-purpose LLMs to medicine poses significant challenges.
- BioMistral, an open-source LLM specifically for the biomedical domain, is introduced in this paper, built upon Mistral and pre-trained on PubMed Central.
- Comprehensive evaluation of BioMistral on 10 established medical question-answering tasks in English shows its superior performance compared to existing models.
- Lightweight models via quantization and model merging, along with the first large-scale multilingual evaluation of medical LLMs, addressing data availability beyond English. All datasets, benchmarks, scripts, and models from the experiments are freely released.



BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains

Yanis Labrak[*1,2] Adrien Bazoge[*3,4]
Emmanuel Morin[4] Pierre-Antoine Gourraud[3] Mickael Rouvier [1] Richard Dufour [1,4]
[1]LIA, Avignon Université [2]Zenidoc
[3]Nantes Université, CHU Nantes, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France
[4]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
{firstname.lastname}@univ-avignon.fr
{firstname.lastname}@univ-nantes.fr

Abstract

Large Language Models (LLMs) have demonstrated remarkable versatility in recent years, offering potential applications across specialized domains such as healthcare and medicine. Despite the availability of various open-source LLMs tailored for health contexts, adapting general-purpose LLMs to the medical domain presents significant challenges. In this paper, we introduce BioMistral, an open-source LLM tailored for the biomedical domain, utilizing Mistral as its foundation model and further pre-trained on PubMed Central. We conduct a comprehensive evaluation of BioMistral on a benchmark comprising 10 established medical question-answering (QA) tasks in English. We also explore lightweight models obtained through quantization and model merging approaches. Our results demonstrate BioMistral's superior performance compared to existing open-source medical models and its competitive edge against proprietary counterparts. Finally, to address the limited availability of data beyond English and to assess the multilingual generalization of medical LLMs, we automatically translated and evaluated this benchmark into 7 other languages. This marks the first large-scale multilingual evaluation of LLMs in the medical domain. Datasets, multilingual evaluation benchmarks, scripts, and all the models obtained during our experiments are freely released.

1 Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP), generative Large Language Models (LLMs) like ChatGPT (OpenAI, 2023) and Vicuna (Zheng et al., 2023) have revolutionized human-computer interactions, demonstrating remarkable versatility and advanced capabilities across various tasks and domains. These models exhibit human-like comprehension and reasoning, enabling them to tackle basic textual understanding as well as complex problem-solving tasks.

[*]Equal contribution.

The emergence of open-source LLMs such as BLOOM (Workshop et al., 2023) and LLaMA (Touvron et al., 2023a) underscores the transformative potential of these models, facilitating their innovative use in specialized domains including medicine (Dave et al., 2023).

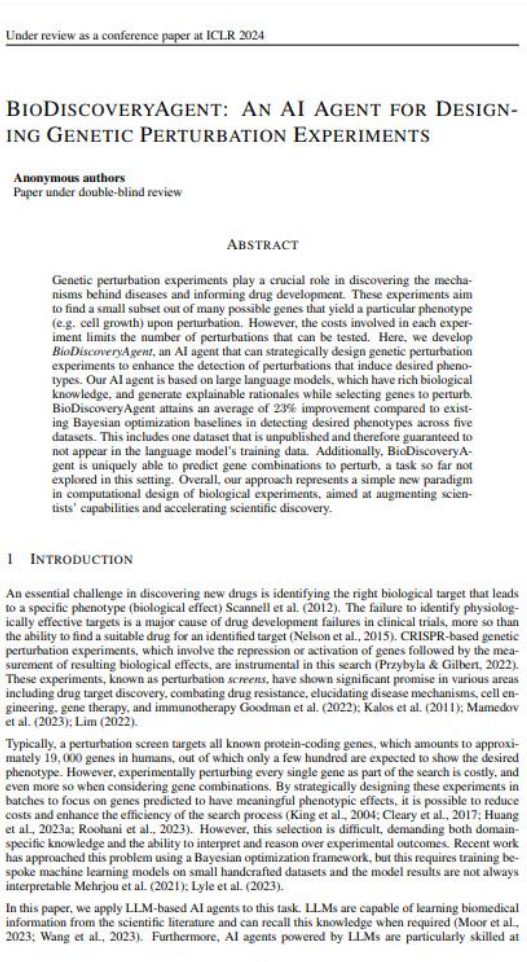
However, integrating LLMs into healthcare and medicine presents unique challenges and opportunities (He et al., 2023; Zhou et al., 2024). While preliminary adoption has opened new avenues for innovation, concerns about data privacy risks associated with proprietary models like MedPaLM-2 (Singhal et al., 2023b) and GPT-4 (Nori et al., 2023a) have arisen. The community's interest in specialized LLMs for healthcare has led to initiatives like PMC-LLaMA (Wu et al., 2023) and MedAlpaca (Han et al., 2023). However, the adoption of open-source medical models has been limited, primarily due to the lack of lightweight models allowing commercial use with performance comparable to larger or proprietary models. To address this gap, there is a need to develop specialized models based on open-source foundation ones like GPT-NeoX (Black et al., 2022), LLaMa 2 (Touvron et al., 2023b), or Mistral (Jiang et al., 2023), and to optimize them for use on consumer-grade devices while maintaining performance.

In this work, we present BioMistral 7B, a specialized LLM tailored for the biomedical domain, derived from Mistral 7B Instruct v0.1 (Jiang et al., 2023) and further pre-trained on PubMed Central. Our contributions include:

1. The construction of BioMistral 7B, the first open-source Mistral-based model for the biomedical domain, including the analysis of different evaluation strategies such as few-shot in-context learning and supervised fine-tuning.
2. An original study with the introduction of a benchmark of 10 medical question-answering (QA) tasks in English, automatically translated into 7 other languages, facilitating

Bio discovery with AI

- Genetic perturbation experiments are vital for understanding disease mechanisms and drug development but are limited by cost.
- BioDiscoveryAgent, an AI agent based on large language models, strategically designs genetic perturbation experiments to enhance the detection of desired phenotypes.
- The AI agent generates explainable rationales while selecting genes to perturb, leveraging its rich biological knowledge.
- BioDiscoveryAgent outperforms existing Bayesian optimization baselines by an average of 23% across five datasets, including one unpublished dataset.
- It can predict gene combinations to perturb, a previously unexplored task, showcasing its potential to augment scientists' capabilities and accelerate scientific discovery.



BIODISCOVERYAGENT: AN AI AGENT FOR DESIGNING GENETIC PERTURBATION EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genetic perturbation experiments play a crucial role in discovering the mechanisms behind diseases and informing drug development. These experiments aim to find a small subset out of many possible genes that yield a particular phenotype (e.g. cell growth) upon perturbation. However, the costs involved in each experiment limits the number of perturbations that can be tested. Here, we develop *BioDiscoveryAgent*, an AI agent that can strategically design genetic perturbation experiments to enhance the detection of perturbations that induce desired phenotypes. Our AI agent is based on large language models, which have rich biological knowledge, and generate explainable rationales while selecting genes to perturb. BioDiscoveryAgent attains an average of 23% improvement compared to existing Bayesian optimization baselines in detecting desired phenotypes across five datasets. This includes one dataset that is unpublished and therefore guaranteed to not appear in the language model's training data. Additionally, BioDiscoveryAgent is uniquely able to predict gene combinations to perturb, a task so far not explored in this setting. Overall, our approach represents a simple new paradigm in computational design of biological experiments, aimed at augmenting scientists' capabilities and accelerating scientific discovery.

1 INTRODUCTION

An essential challenge in discovering new drugs is identifying the right biological target that leads to a specific phenotype (biological effect) Scannell et al. (2012). The failure to identify physiologically effective targets is a major cause of drug development failures in clinical trials, more so than the ability to find a suitable drug for an identified target (Nelson et al., 2015). CRISPR-based genetic perturbation experiments, which involve the repression or activation of genes followed by the measurement of resulting biological effects, are instrumental in this search (Przybyla & Gilbert, 2022). These experiments, known as perturbation *screens*, have shown significant promise in various areas including drug target discovery, combating drug resistance, elucidating disease mechanisms, cell engineering, gene therapy, and immunotherapy Goodman et al. (2022); Kalos et al. (2011); Mamedov et al. (2023); Lim (2022).

Typically, a perturbation screen targets all known protein-coding genes, which amounts to approximately 19,000 genes in humans, out of which only a few hundred are expected to show the desired phenotype. However, experimentally perturbing every single gene as part of the search is costly, and even more so when considering gene combinations. By strategically designing these experiments in batches to focus on genes predicted to have meaningful phenotypic effects, it is possible to reduce costs and enhance the efficiency of the search process (King et al., 2004; Cleary et al., 2017; Huang et al., 2023a; Roohani et al., 2023). However, this selection is difficult, demanding both domain-specific knowledge and the ability to interpret and reason over experimental outcomes. Recent work has approached this problem using a Bayesian optimization framework, but this requires training bespoke machine learning models on small handcrafted datasets and the model results are not always interpretable Mehrjou et al. (2021); Lyle et al. (2023).

In this paper, we apply LLM-based AI agents to this task. LLMs are capable of learning biomedical information from the scientific literature and can recall this knowledge when required (Moor et al., 2023; Wang et al., 2023). Furthermore, AI agents powered by LLMs are particularly skilled at

Bio discovery with AI

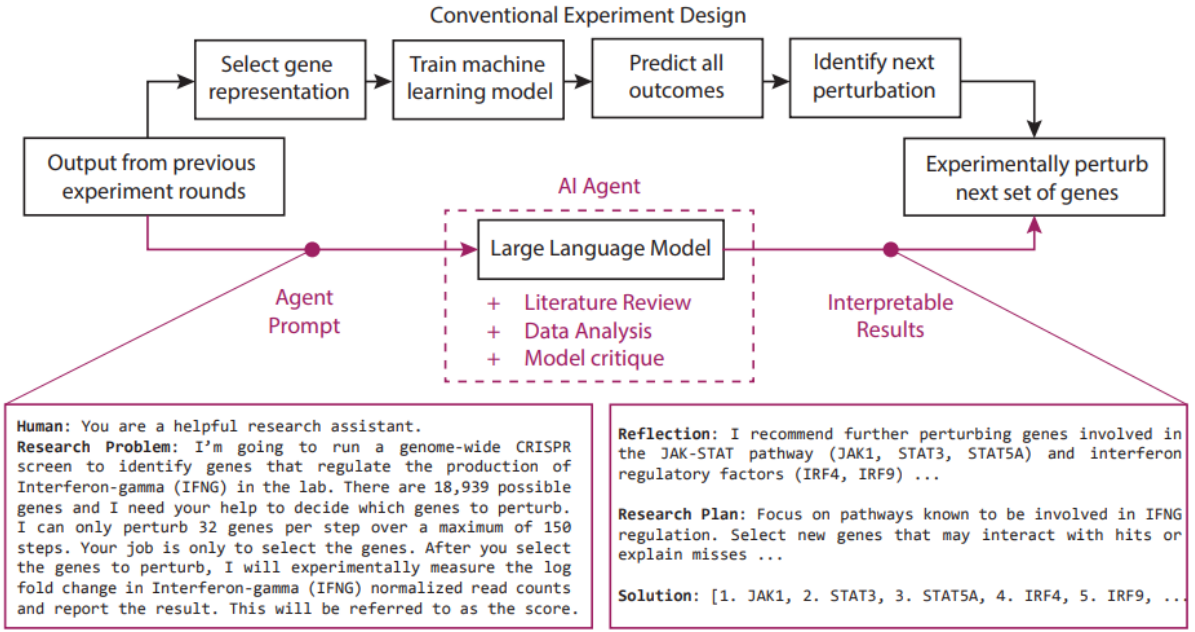


Figure 1: Comparison of AI agent based approach (BioDiscoveryAgent) to conventional machine-learning methods for experiment design.

Bio discovery with AI

The prompt includes the task information and response format as this example shown below.

```
You are a scientist working on problems in drug discovery.

Research Problem: I'm planning to run a genome-wide CRISPR screen
   to identify genes that regulate the production of Interleukin
   -2 (IL-2). There are 18,939 possible  genes to perturb and I
   can only perturb 128 genes at a time. For each perturbation, I
   'm able to measure out the log fold change in Interleukin-2 (
   IL-2) normalized read counts which will be referred to as the
   score. I can only do a few rounds of experimentation.

Always respond in this format exactly:

1. Reflection: Thoughts on previous results and next steps.
2. Research Plan: The full high level research plan, with current
   status and reasoning behind each proposed approach. It should
   be at most 5 sentences.
3. Solution: Propose a list of predicted genes to test separated
   by commas in this format: 1. <Gene name 1>, 2. <Gene name 2>
   ...
Do not include any genes from this prompt (since they're already
   tested).
```

Bio discovery with AI

- Concept of "AI scientists" as systems that enhance biomedical research through collaborative agents integrating machine learning tools with experimental platforms.
- These AI agents do not aim to replace human involvement but rather augment it by combining human creativity and expertise with AI's ability to analyze large datasets, navigate hypothesis spaces, and execute repetitive tasks.
- Proficient in tasks such as self-assessment, planning discovery workflows, and analyzing data, these agents utilize large language models and generative models to maintain structured memory for continual learning.
- They incorporate scientific knowledge, biological principles, and theories, impacting areas ranging from hybrid cell simulation and programmable control of phenotypes to cellular circuit design and the development of new therapies.

Empowering Biomedical Discovery with AI Agents

Shanghua Gao[1], Ada Fang[1,2,8,+], Yepeng Huang[1,3,+], Valentina Giunchiglia[1,4,+], Ayush Noori[1,5,+], Jonathan Richard Schwarz[1], Yasha Ektefaie[1,6], Jovana Kondic[7], and Marinka Zitnik[1,8,9,10,#]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[3]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA
[4]Department of Brain Sciences, Imperial College London, London, UK
[5]Harvard College, Cambridge, MA, USA
[6]Program in Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[7]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[8]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA
[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Harvard Data Science Initiative, Cambridge, MA, USA
[+]Co-second authors
[#]Correspondence: marinka@hms.harvard.edu

Summary

We envision "AI scientists" as systems capable of skeptical learning and reasoning that empower biomedical research through collaborative agents that integrate machine learning tools with experimental platforms. Rather than taking humans out of the discovery process, biomedical AI agents combine human creativity and expertise with AI's ability to analyze large datasets, navigate hypothesis spaces, and execute repetitive tasks. AI agents are proficient in a variety of tasks, including self-assessment and planning of discovery workflows. These agents use large language models and generative models to feature structured memory for continual learning and use machine learning tools to incorporate scientific knowledge, biological principles, and theories. AI agents can impact areas ranging from hybrid cell simulation, programmable control of phenotypes, and the design of cellular circuits to the development of new therapies.

1 Introduction

A long-standing ambition for artificial intelligence (AI) in biomedicine is the development of AI systems that could eventually make major scientific discoveries, with the potential to be worthy of a Nobel Prize—fulfilling the Nobel Turing Challenge [1]. While the concept of an "AI scientist" is aspirational, advances in agent-based AI pave the way to the development of *AI agents* as conversable systems capable of skeptical learning and reasoning that coordinate large language models (LLMs), machine learning (ML) tools, experimental platforms, or even combinations of them [2–5] (Figure 1).

The complexity of biological problems requires a multistage approach, where decomposing complex questions into simpler tasks is necessary. AI agents can break down a problem into manageable subtasks, which can then be addressed by agents with specialized functions for targeted problem-solving and integration of scientific knowledge, paving the way toward a future in which a major biomedical discovery is made solely by AI [2, 6]. In the near future, AI agents can accelerate discovery workflows by making them faster and more

Bio discovery with AI

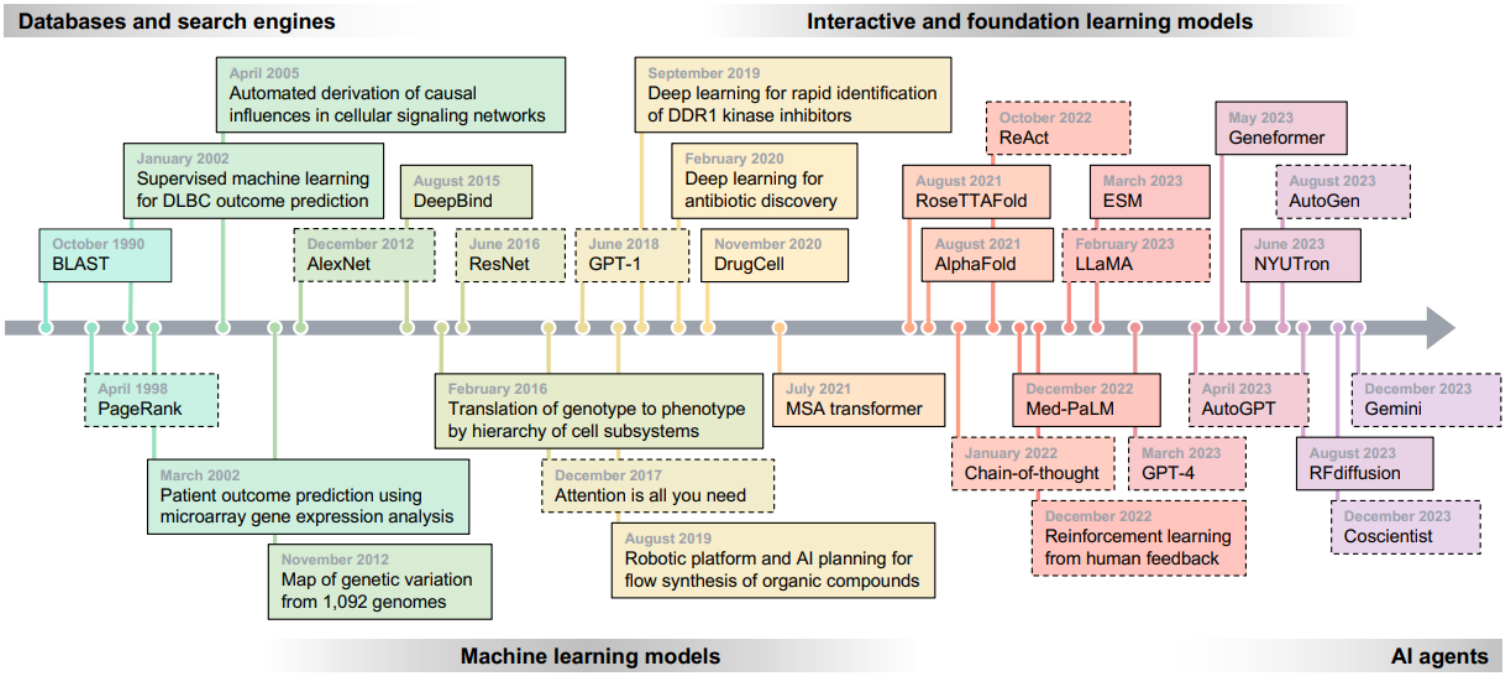


Figure 2: Evolving use of data-driven models in research. Data-driven approaches, from databases and search engines, machine learning, and interactive learning models to advanced agent systems (Section 2), have reshaped biomedical research throughout the last several decades.

Bio discovery with AI

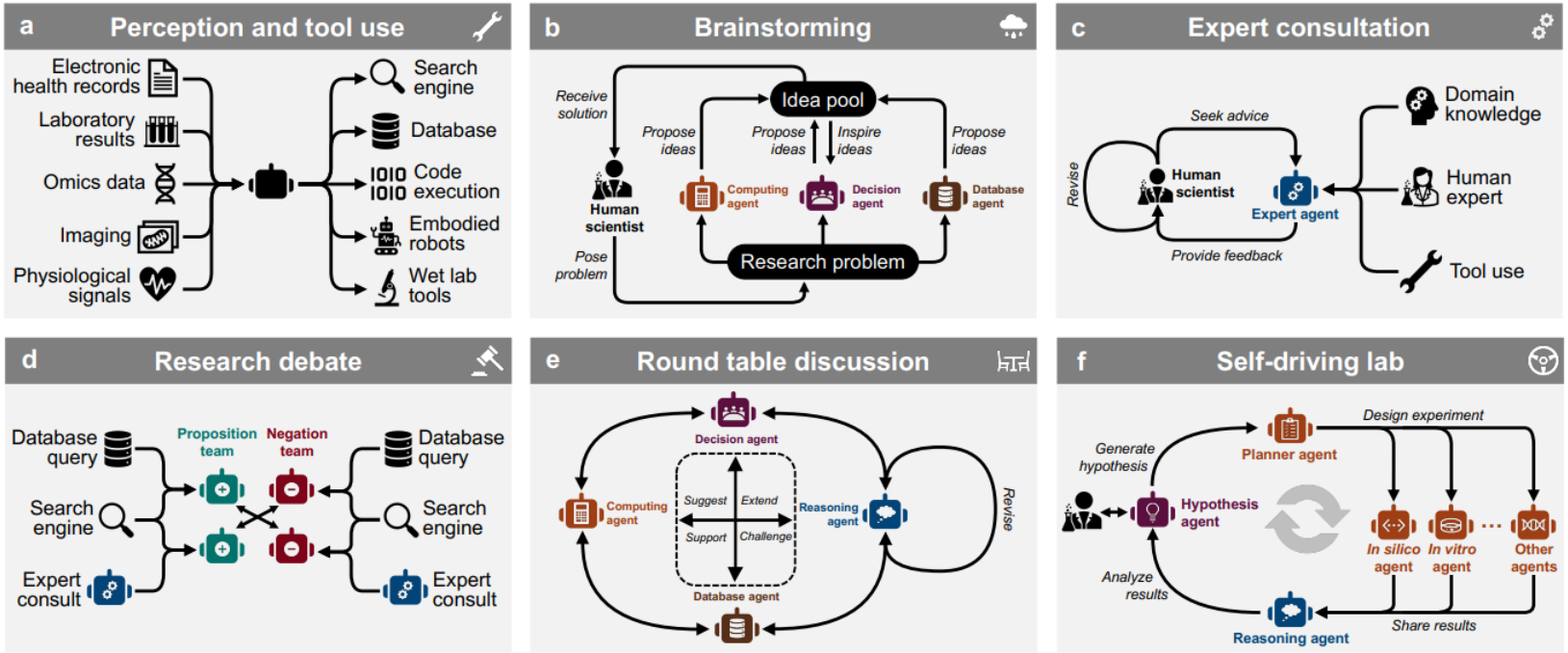


Figure 3: Diverse configurations of AI agents in biology – from an LLM-based AI agent to a multi-agent system with AI models, tools, and integrated physical devices.

Bio discovery with AI

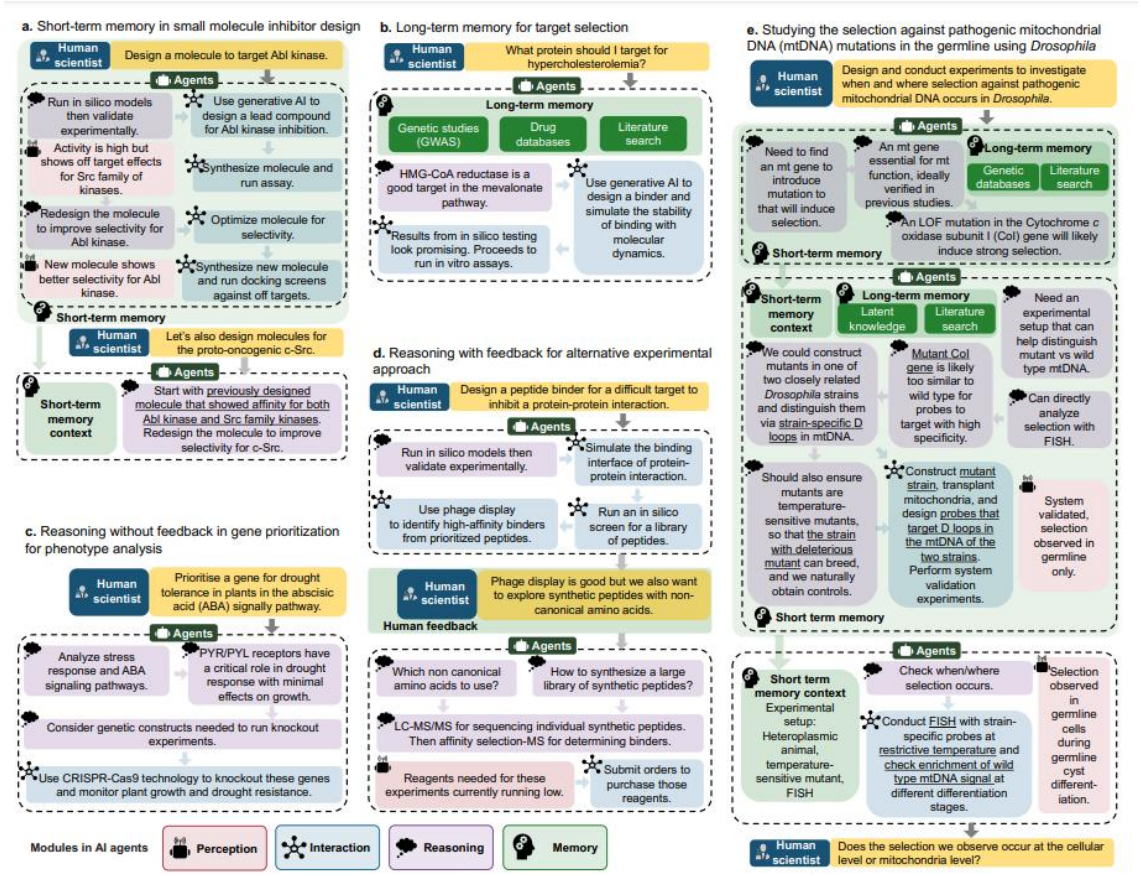
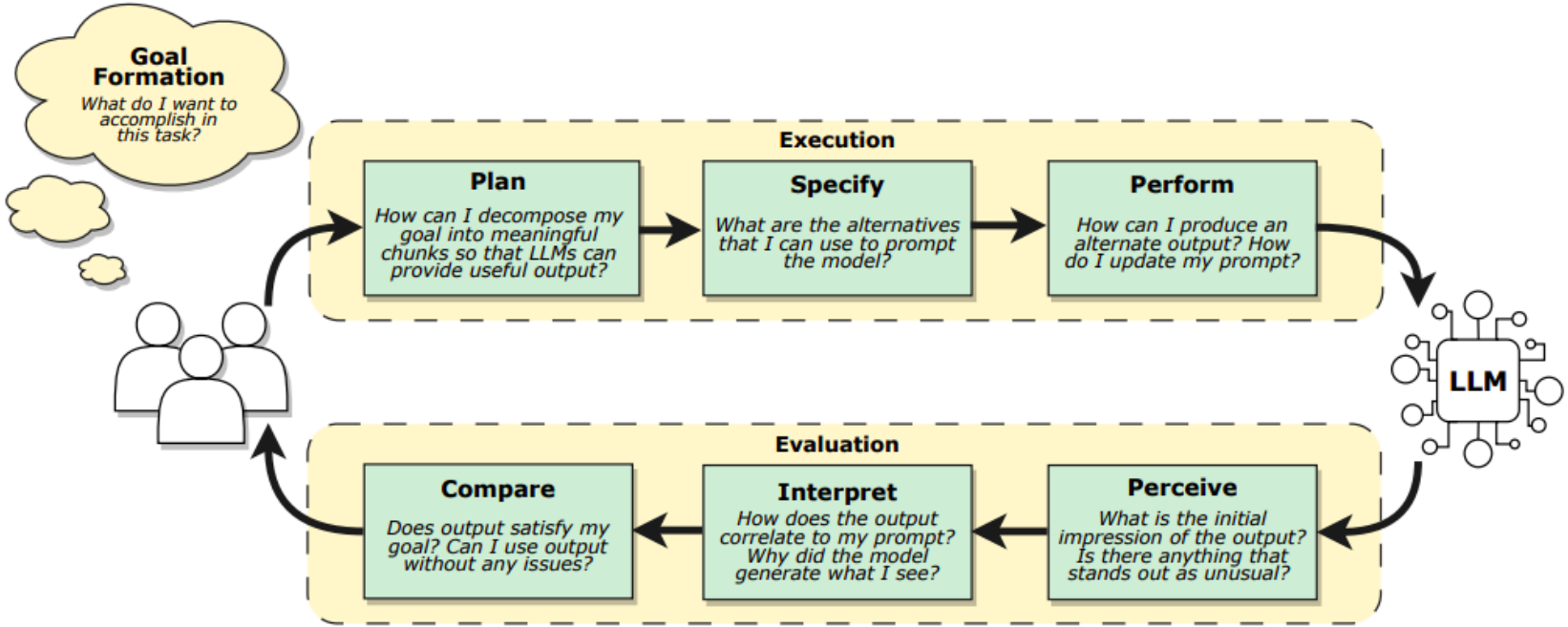
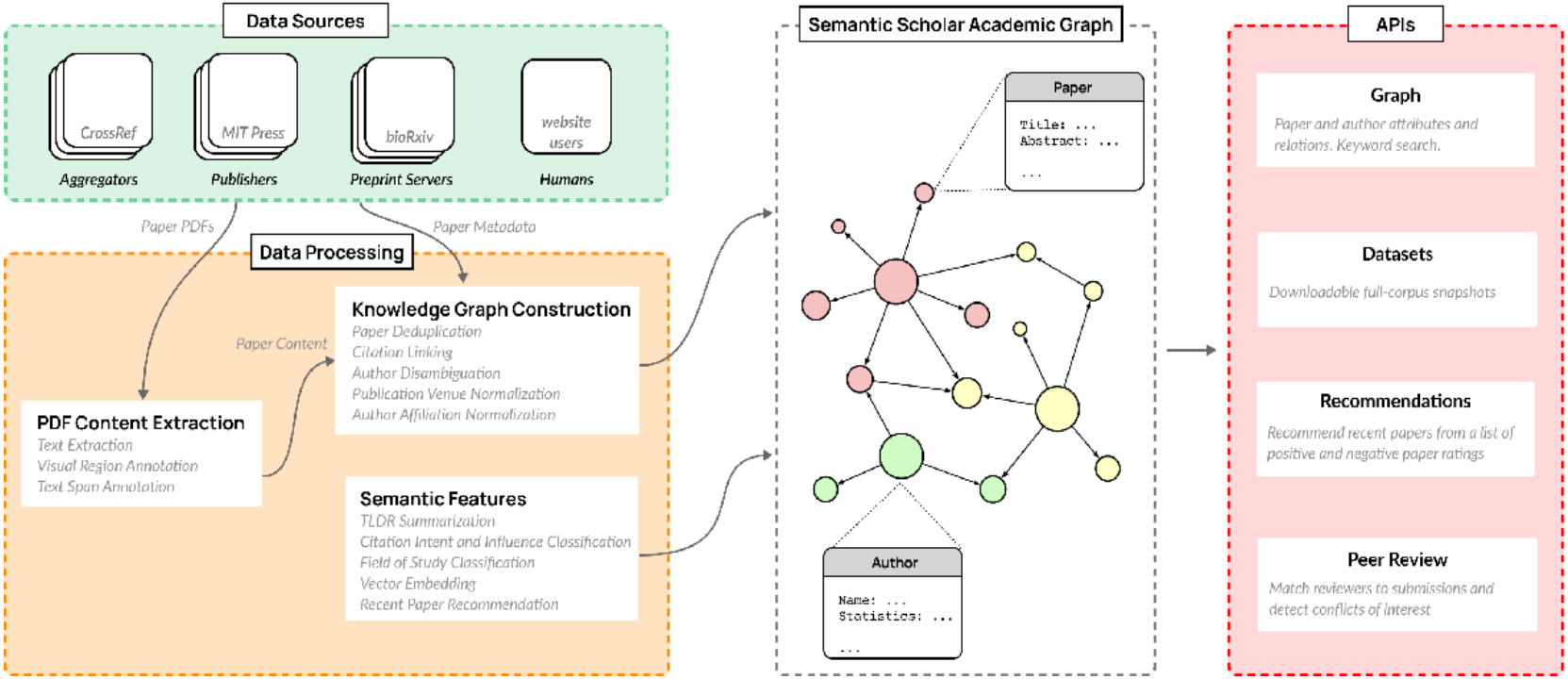


Figure 5: Components of biomedical AI agents.

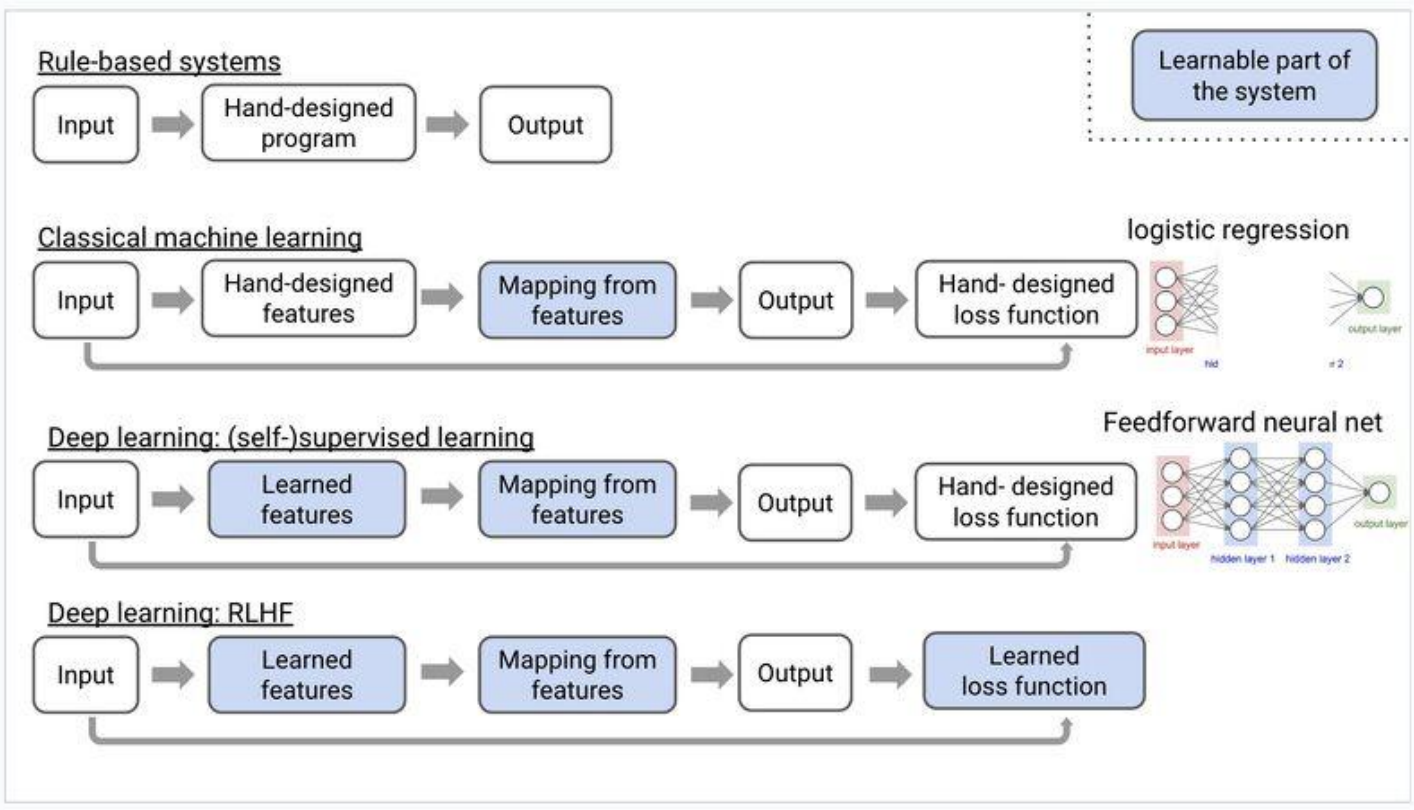
Bio discovery with AI



Microsoft / AllenAI - Semantic scholar



What NLP brought to AI



Natural Language Processing

Main functions of Language

- We **communicate** using language
- We **think and articulate thoughts** with language
- We build **Scientific Theories** with language

Purpose of NLP

- **Access Knowledge** (search engine, recommender system…)
- **Communicate** (e.g. Translation)
- **Linguistics** and **Cognitive Sciences** (Analyse Languages themselves)

Natural Language Processing

Amount of online textual data…

- 70 billion web-pages online (1.9 billion websites)
- 55 million Wikipedia articles

…Growing at a fast pace

- 9000 tweets/second
- 3 million mail / second (60% spam)

Natural Language Processing

Potential Users of Natural Language Processing

- 7.9 billion people use some sort of language
- 4.7 billion internet users (~59%)
- 4.2 billion social media users (~54%)

Natural Language Processing

Current main Products and Services

- Search: +2 billion Google users, 700 millions Baidu users
- Social Media: +3 billion users of Social media (Facebook, Instagram, WeChat, Twitter...)
- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)
- Machine Translation: 500M users for google translate

Short history of NLP

Symbolic

- - 2000
- Grammar-based systems
- Syntactic tree
- Deterministic inference

Statistical Learning

- 2000-2010
- Small corpus
- SVM, Logistic regression
- Handcrafted features

Deep Learning

- 2010-Today
- Deep Learning Architecture (Transformer)
- Transfer Learning in NLP (word2vec, BERT, CamemBERT, GPT)
- More compute, larger (raw) dataset, Open Source Deep

1949: First Machine Translation “Model”

1949 ***Memorandum on Translation,*** Warren Weaver
First to propose the idea of using “*electronic computers*” to do translation

- Using Shannon’s information theory to frame Machine Translation as a ***cryptographic*** problem
- **Modeling context** to disambiguate between word senses
- “*Going down*” from each language to ***universal language*** in order to translate

1964: ELIZA First Conversational Bot

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE

Natural Language Processing WorkFlow

Assume we have a **Research, Engineering, Product Problem**

1. Define a Split **NLP System** to solve it **modules**, each one performing a **task**
2. Define **Evaluation Metric(s) for your system and submodules**
3. **Collect Data** to build/train your models
4. Build **Baseline Models** (i.e. most simple model you can think of that have a non trivial performance metric)
5. Build **Better Models** using symbolic/statistical/DL methods

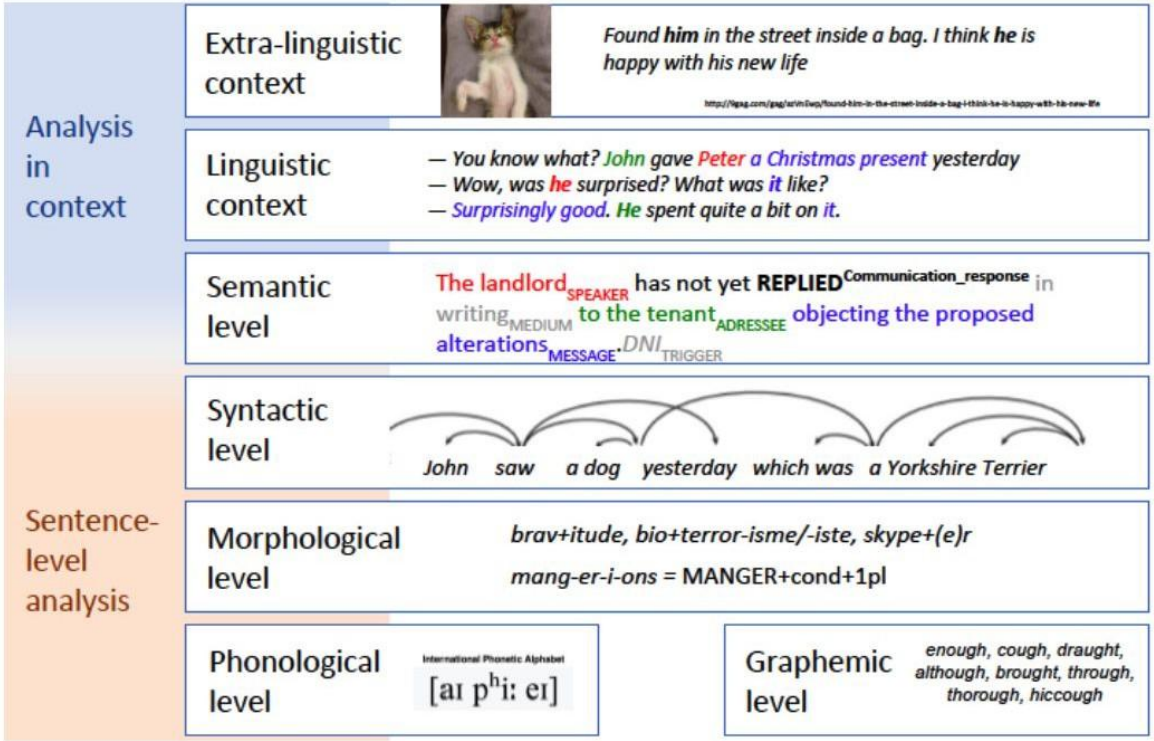
Challenges of Language

A Definition of Language

Definition 1: *Language is a means to communicate, it is a semiotic system. By that we simply mean that it is a set of signs. A sign is a pair consisting in [...] a signifier and a signified.*

Definition 2: *A sign consists in a phonological structure, a morphological structure, a syntactic structure and a semantic structure*

The Six Levels of Linguistics Analysis



The 5 Challenges of NLP

1. Productivity
2. Ambiguous
3. Variability
4. Diversity
5. Sparsity

Productivity

Definition

“property of the language-system which enables native speakers to construct and understand an indefinitely large number of utterances, including utterances that they have never previously encountered.” (Lyons, 1977)

➔ **New words, senses, structure** are **introduced in languages all the time**

Ambiguous

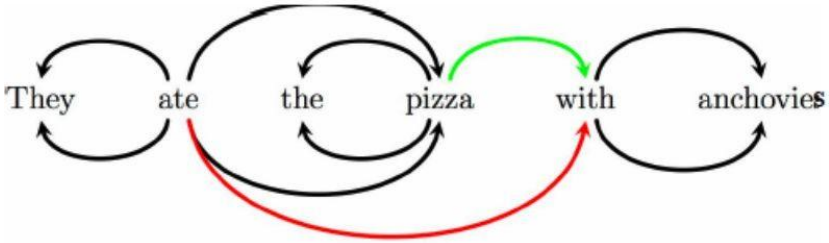
Most linguistic observations (speech, text) are open to **several interpretations**

We (Humans) disambiguate -i.e. **find the correct interpretation -** using all kind of signals (linguistic and extra linguistic)

Ambiguity can appear at all levels (phonology, graphemics, morphology, syntax, semantics)

Ambiguous

Syntactic Ambiguity



Ambiguous

Semantic Ambiguity

- Polysemy: *e.g.* ***set , arm, head***
  Head of New-Zealand is a woman
- Name Entity: *e.g.* ***Michael Jordan***
  Michael Jordan is a professor at Berkeley
- Object/Color: *e.g.* ***cherry***
  Your cherry coat

Ambiguous

Pragmatic Ambiguity

Two Soviet ships collide, ***one dies***

Dealers will hear ***car talk*** *at noon*

Ambiguous

Disambiguating can require Discourse Knowledge

Where can I find **a vegetarian restaurant** in Paris

Here is a list of restaurant in Paris: ….

Give me the top ranked ones, in the 14th arrondissement

Here are the top ranked restaurant in the 14th arrondissement in Paris

How far is the closest **one** from my current location?

Variation

Language Varies at all levels

- Phonetic (accent)
- Morphological, Lexical (spelling)
- Syntactic
- Semantic

Variation Determiners

- Who is talking?
- To Whom?
- Where? *Work, Home, Restaurant*
- When? *19th century, 2008, 2022…*
- About what? *Specialised domain, the Weather,…*

Essentially, the Variability of a language depends on:

- Social Context
- Geography
- Sociology
- Date
- Topic

Diversity

- About **7000 languages** spoken in the world
- About **60%** are found in the **written form** (cf. Omniglot)

Syntactic Diversity

A key characteristics of the syntax of a given language is **the word order**

- **Word order differs** across languages
- Word order degree of freedom also differs across languages
- We characterize word orders with: **Subject (S) Verb (V) Object (O) order**

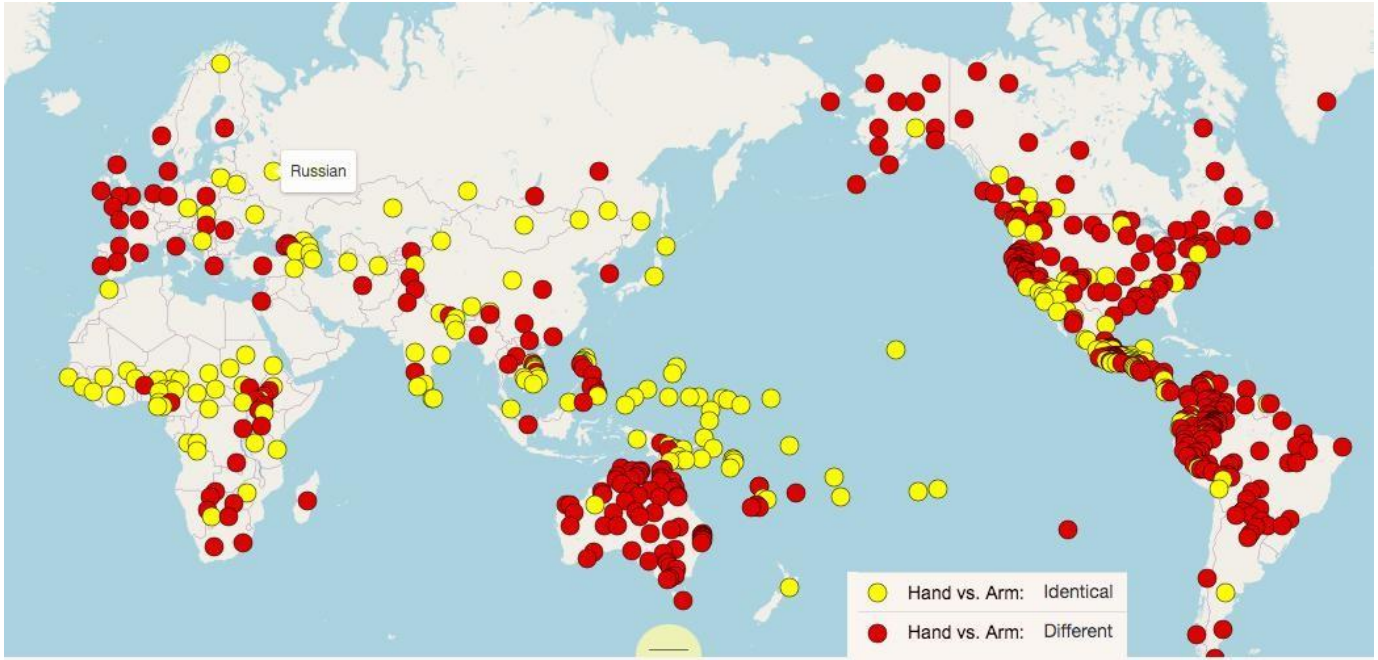
Word Order Freedom And Morphology

- Word orders freedom and morphology are usually related
- **The more freedom in word orders**
  → the less information is conveyed by word positions
  → the more information is carried by each word
  → the richer the morphology

English ***cats eat mice***

Semantic Diversity

- Words partition the semantic space
- This partition is very diverse across language

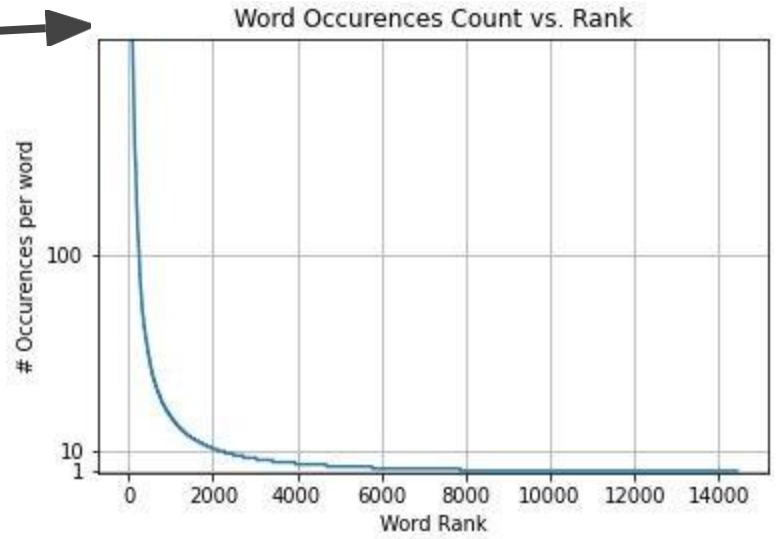


(Dyer et. al 2013)

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

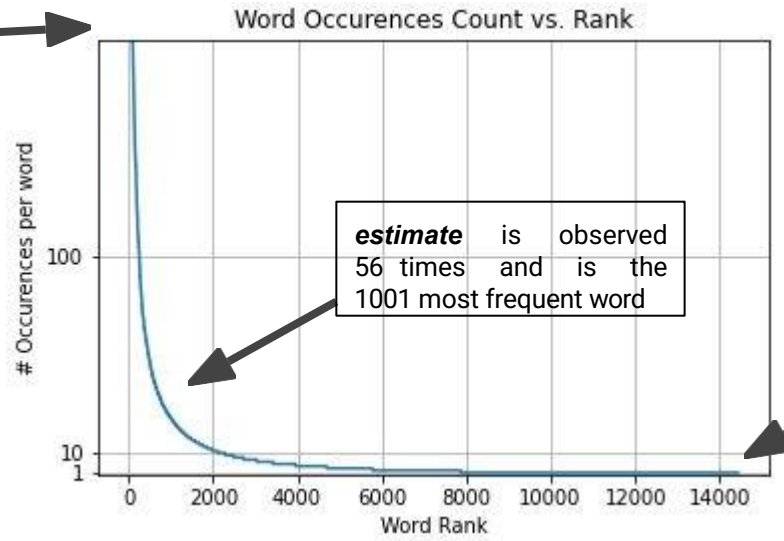
the is the most observed (rank 1) word with 8119 occurrences



Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

the is the most observed (rank 1) word with 8119 occurrences



estimate is observed 56 times and is the 1001 most frequent word

About 6000 Words are observed only 1 time in the dataset (e.g. ***stakeholders, pending, score…)***

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, word distributions follows ***a Zipf Law*** ***ie:***

f_w frequency of entity w
k frequency rank of entity w

$$f_w(k) \propto \frac{1}{k^\theta}$$

- Zipf law is a Power relation between the rank and frequency

*The most frequent entities are **much more frequent** than the less frequent ones*

- Under a Zipf law, log(fw) and log(k) are linearly related

Statistical Description of Language

Zipf Distributions are observed not only for words but with many other units of language (sounds, syntactic structure, name entities…)

Consequence

A large number of units are observed in language with very low frequency i.e. **Sparsity**

Very challenging for NLP

What is Natural Language Processing?

In a nutshell, NLP consists in handling the complexities of natural languages "to do something"

- Raw Text / Speech → Structured Information
- Raw Text / Speech → (Controlled) Text/Speech

In this course we will focus **on textual data**

Framework

We assume:

- A **token** is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1, ..., V.
- A **document** is a sequence of N words denoted by **d = (w1,w2, ...,wN)**, where wn is the N-th word in the sequence.
- A **corpus** is a collection of M documents denoted by D = (d1, d2, ..., dM)

Example: *Wikipedia, All the articles of the NYT in 2021…*

Token

With regard to our end task, a token can be:

- A word
- A sub-word: *e.g. a sequence of 3 characters*
- A character
- An sequence of characters (sometimes a word, sometimes several words, sometimes a sub-word…)

Document

A Document can be:

- A Sentence
- A Paragraph
- A sequence of characters

Text Segmentation

Definition: Text Segmentation is the process of splitting raw text (i.e. list of characters) into **units of interest**.

Two level of segmentation (usually) required :

- Split raw text into **modeling units** (ex: sentence, paragraph, 1000 characters, web-page...)
- Split modeling units into sequence of **basic units** (referred as tokens) (e.g: words, word-pieces, characters, ...)

Two distinct approaches:

- **Linguistically informed** e.g. word, sentence segmentation...
- **Statistically informed** e.g. frequent sub-words (word pieces, sentence pieces...)

Tokenization

Definition: Tokenization consists in *segmenting* raw textual data into tokens:

Can be framed as a character level task

input: *une industrie métallurgique existait*.
output: IIIEIIIIIIIIIIIIIEIIIIIIIIIIIIIIIIIIEIIIIIIIIIIIEE

- **Easy task** for most languages and domains
- Can be **very complex in some cases** (Chinese, Social Media...)

NLP Tasks: Modeling Framework

Let (X, Y) a pair of random variable. X may characterize tokens or documents. Modeling an NLP task consists in estimating the conditional probability $Y|X$ in order to predict Y with X.

$$p(Y|X)$$

Tasks Taxonomy

- If *Y* is a single label and X a sequence of tokens (e.g. a sentence): **Sequence Classification**
- If we have one label per token: **Sequence Labelling**
- If Y is a sequence of tokens: **Sequence Prediction**
- If Y is a graph, a tree or a complex structured output: **Structure Prediction**

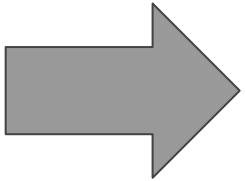
Document Classification

Europe

Germany's minimum wage hike will not cost jobs -labour minister

BERLIN, Jan 21 (Reuters) - Germany's planned minimum wage hike to 12 euros ($13.61) per hour from October means a pay rise for over 6 million people across the country and should not cost jobs contrary to critics, Labour Minister Hubertus Heil said on Friday.

Increasing the German minimum wage, currently 9.82 euros per hour and will increase to 10.45 euros per hour from July, to 12 euros per hour was one of the key election promises of Chancellor Olaf Scholz and his Social Democrats.



Politics

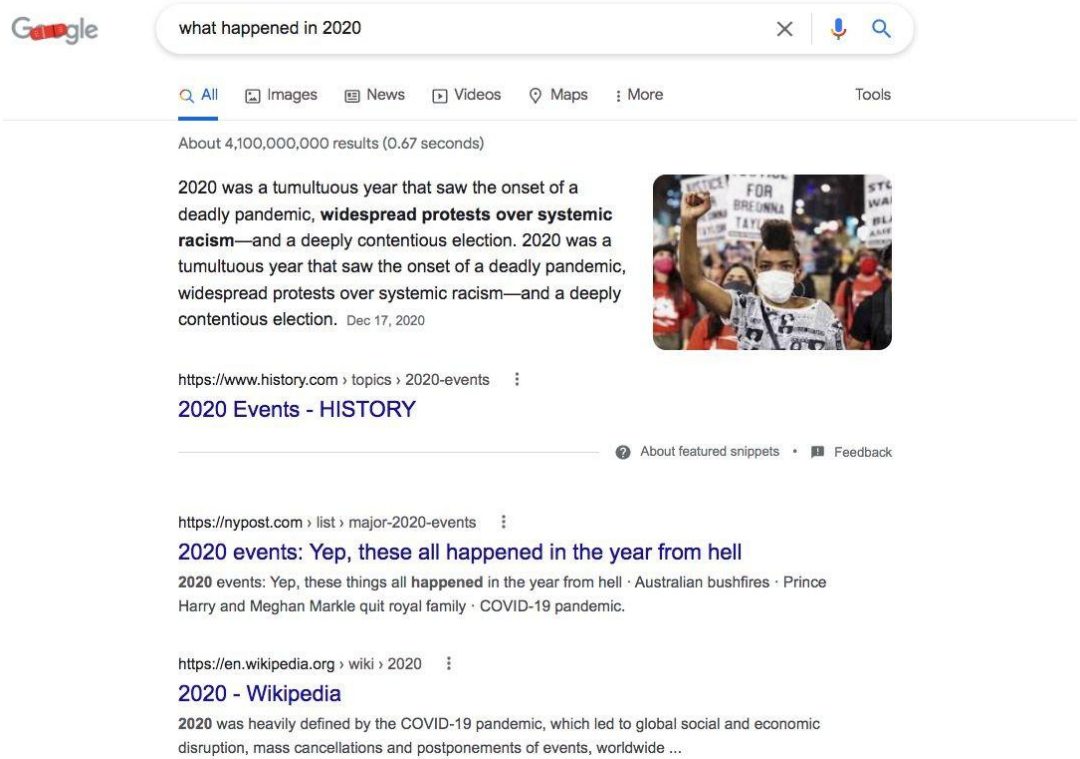
Economy

Travel

….

Geopolitics

Document Ranking (Retriever)



NLP Tasks: Part-of-Speech Tagging

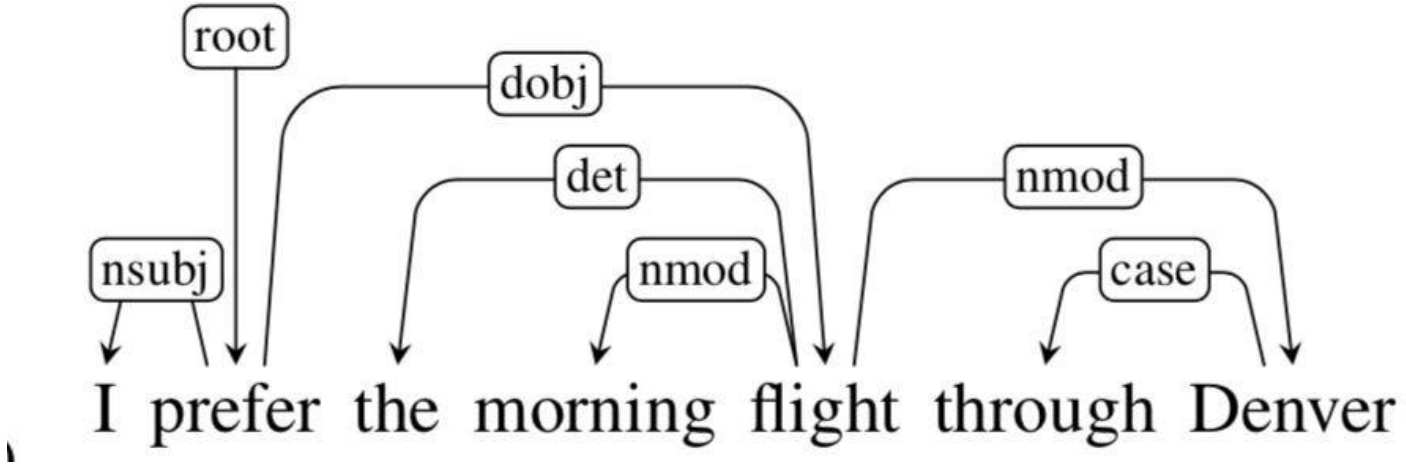
POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

[PRON , NOUN, VERB, NOUN, CC, PRON, VERB, PREP, NOUN, PUNCT]

Syntactic Parsing

Syntactic Parsing consists in **extracting the syntactic structure** of a sentence. For instance, **Dependency Parsing** (here) predicts an acyclic directed graph (a **tree**)



Slot-Filling / Intent Detection

Intent Detection is a sequence classification task that consists in **classifying the intent of a user** in a pre-defined category.

Slot-Filling is a sequence labelling task that consists in identifying **specific parameters in a user request**.

Can you please play Hello from Adele ?

Intent: ***play_music***
Slots: [*Can, you, please, play, Hello, from, Adele, ?]*
*[O , O , O , O , **SONG**, O , ARTIST, O]*

Semantic Role Labelling (SLR)

- SRL is the task of finding the **semantic role** of each predicate in a sentence.

- Given a sentence, SRL predicts: *who did what to whom, when,* ***where, why, how***

NLP Tasks: Name Entity Recognition

NER: Find the **Name-Entities** in a sentence

[My , name, is, Bob, and, I, live, in, NY, !]

*[O , O, O, **PERSON**, O, O, O, O, **LOCATION**, O]*

Machine Translation

INPUT: My name is Bob and I live in NY !

OUTPUT: Je m'appelle Bob et je vis à New-York!

Question Answering

INPUT: *How many episodes in season 2 breaking bad?*

OUTPUT: ***13***

How do we solve an NLP Problem?

Each NLP Problem is unique

➔ **No Universal Method to solve them all**
➔ Have a set of tools in mind, symbolic, statistics and deep learning based

NLP Engineering: Find **what works best** for your given method based on what has been done already in the literature

NLP Research: Find a **better** way to do, more accurate, cheaper, faster