

# Knowledge Graphs: Extracting and Leveraging Meaning from Unstructured Text

Bang Liu, Haochen Shi  
April 3rd, 2023



## 2 Introduction about the teacher and teaching assistant

- ◎ **Bang Liu (刘邦)** is an Assistant Professor in the Department of Computer Science and Operations Research (**DIRO**) at the **University of Montreal**. He is a core member of the **RALI** laboratory (Applied Research in Computer Linguistics) of DIRO, an associate member of **Mila – Quebec Artificial Intelligence Institute**, and a **Canada CIFAR AI (CCAI) Chair**. His research interests primarily lie in the areas of **natural language processing, multimodal & embodied learning, theory and techniques for AGI (e.g., understanding and improving large language models), and AI for science (e.g., health, material science, XR)**.



- ◎ **Haochen Shi (史皓琛)** is a Ph.D. candidate in the Department of Computer Science and Operations Research (**DIRO**) at the **University of Montreal**, supervised by Prof. Bang Liu. He received his master's degree in 2021 from Zhejiang University, as well as his bachelor's degree from Huazhong University of Science and Technology. His research interests primarily lie in the areas of **natural language processing (NLP), multimodal & embodied learning, and resource-economical deep learning**.





# Tutorial outline

## ◎ Part I: Introduction to knowledge graph

- What is a knowledge graph?
- Why are knowledge graphs important?
- Success stories of knowledge graphs

## ◎ Part II: Semantic Web Knowledge Graph

- W3C Semantic Web
- RDF, SPARQL, etc.

## ◎ Part III: Knowledge graph construction

- A pipeline: KG source -> KG construction -> KG maintenance -> KG application

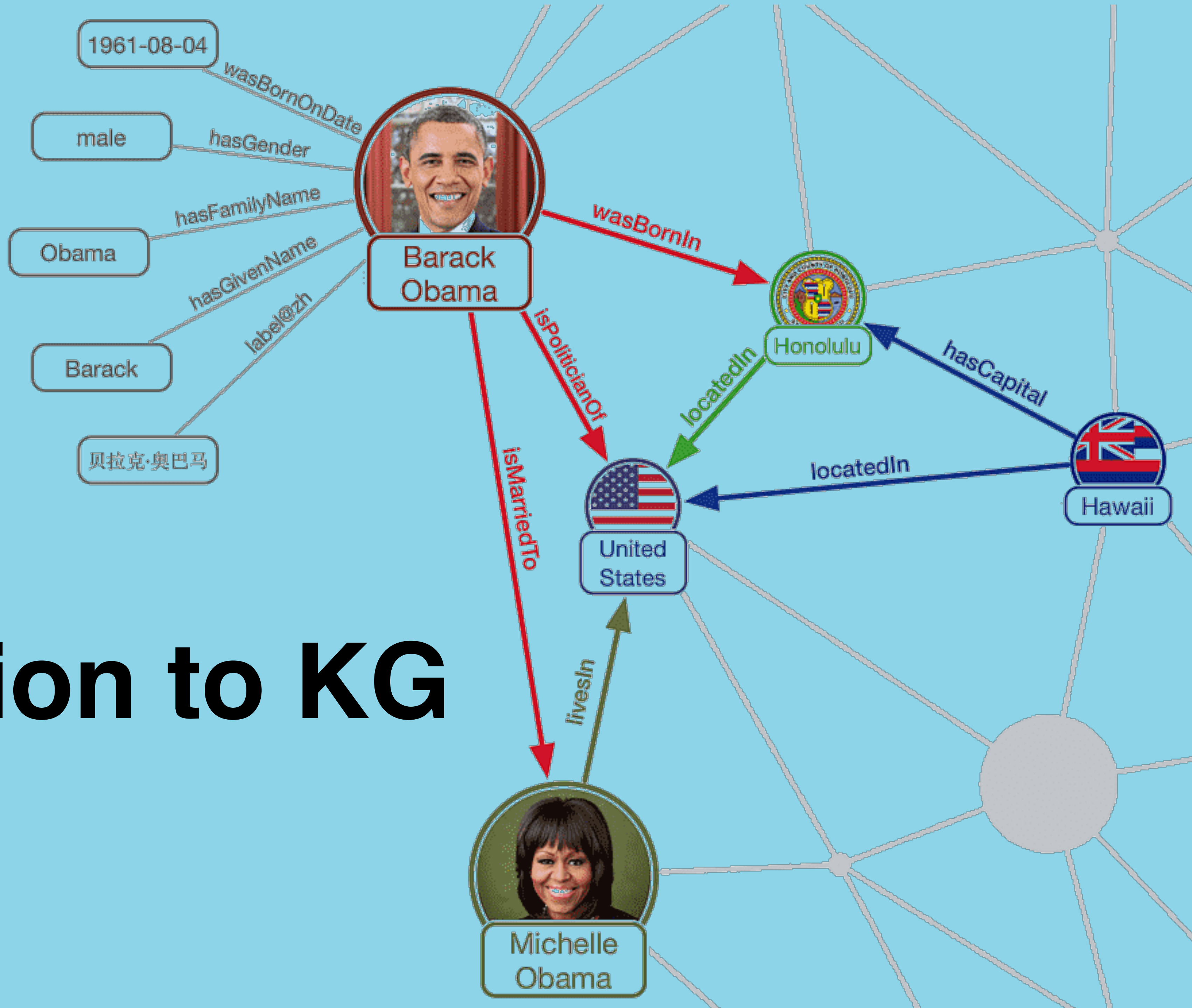
## ◎ Part IV: Knowledge graph applications

- Emphasize on KG-based question answering

## ◎ Part V: Active Research about KG

## ◎ Part VI: ChatGPT and KG

# Introduction to KG





5 Suppose you are searching...



**What he is interested in?**

**Query: "Honda Civic, Hyundai Elantra"**

## 6 Is-A relationship between things

Honda Civic



Hyundai Elantra



◆ **Fuel-efficient cars**

◆ **Economy cars**



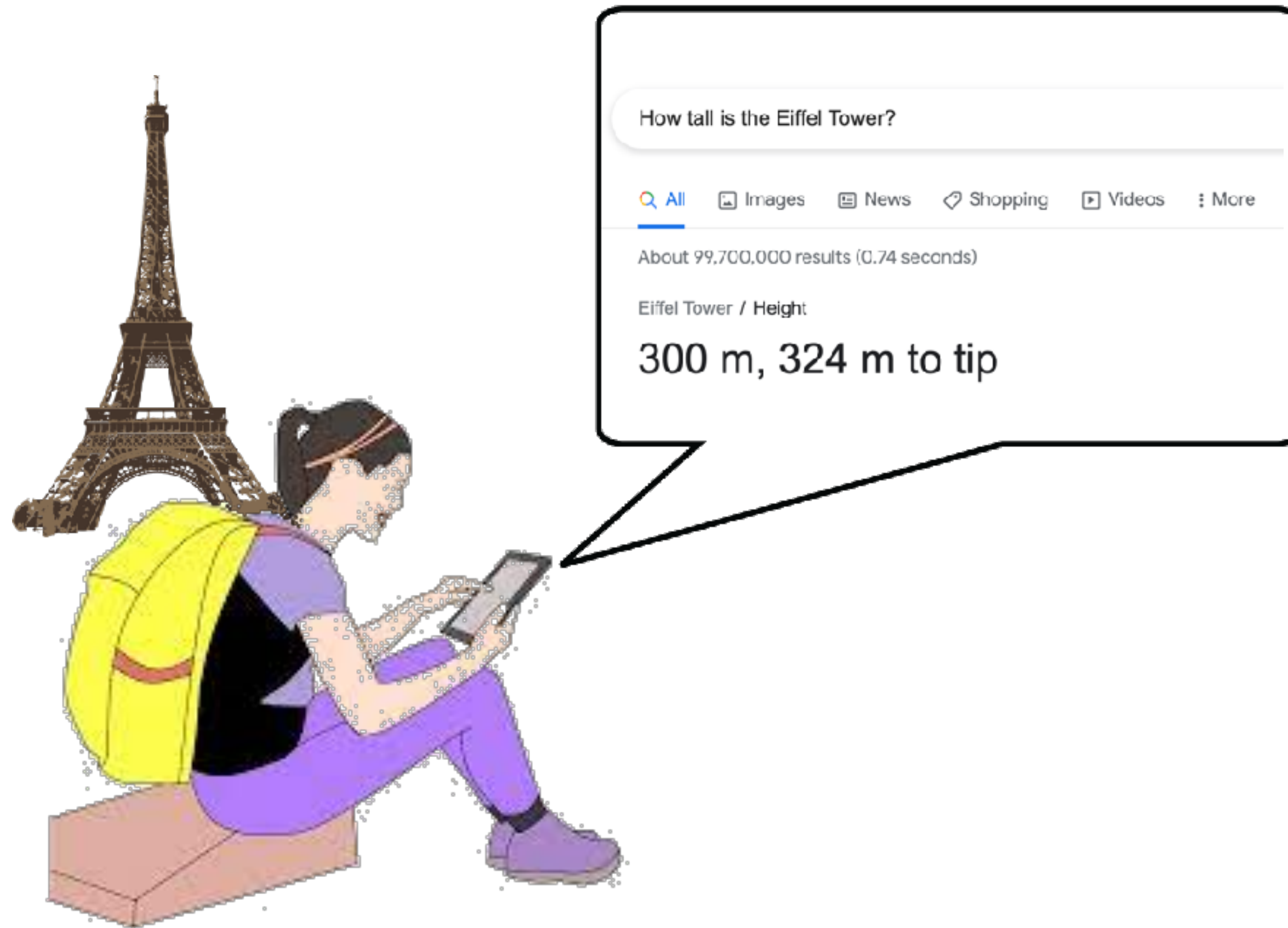
◆ **Marvel heroes**

◆ **Revenagers**

**Concept:** a collection of things that share some common attributes



## 7 Suppose you are asking...



**How does Google know?**

**Question: "How tall is the Eiffel Tower?"**

## 8 Properties of things

# The Eiffel Tower at a glance

An object of discord, desire and fascination, the Eiffel Tower never fails to impress. Enriched by a history full of new developments, here you can discover all of its key information.

### *Main figures*

Current height	1063 feet
Original height without the antennas	1024 feet
Total width	410 feet (on the ground)
Width of a pillar	82 feet (on the ground)
First floor	187 feet, 14,485 square feet
Second floor	377 feet, 4,692 square feet
Third floor	906 feet, 820 square feet
Lifts	5 lifts from the esplanade to second floor, 2 x 2 duolifts from second floor to the top
Weight of the metal frame	7,300 tons



## 9 Suppose you are shopping...



Roll over image to zoom in

<b>Brand</b>	Cetaphil
<b>Ingredients</b>	Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.
<b>Scent</b>	Fragrance free
<b>Additional Item Information</b>	Non-Comedogenic, Fragrance-free, Natural
<b>Skin Type</b>	Sensitive

### About this item


- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean
- Mild & non irritating: Soap free formulation won't strip skin of its natural protective oils and emollients

**How does Amazon provide the information?**

# 10 Suppose you are shopping...

Visit the Bigmuscles Nutrition Store ★★★★☆ 6,270  
Bigmuscles Nutrition Crude Whey 1kg, Whey Protein Concentrate 80%, 24g Protein, 5.5g BCAA, 4 g Glutamine

40% off



Style name: **Crude Whey [Unflavored]**


Flavor: **Unflavored**

Size: **1Kg[2.2lbs]**

₹ 779 (₹ 82.50 / 100 g) **Lightning Deal**  
MRP: ₹-1,299.00 **Save ₹520**  
prime

Visit the Bigmuscles Nutrition Store ★★★★☆ 6,270  
Bigmuscles Nutrition Crude Whey 1kg, Whey Protein Concentrate 80%, 24g Protein, 5.5g BCAA, 4 g Glutamine

40% off



Style name: **Crude Whey [Unflavored]**

Flavor: **Unflavored**

<b>Unflavored</b>	Caffe Latte	Isola
₹ 779.00 prime (₹ 77.90 / 100 g) Deal is 34% Claimed	See available options	See a optio

Size:

How does Amazon provide choices?



# Suppose you are shopping...



**KitchenAid KSM150PSE Artisan Tilt-Head Stand Mixer with Pouring Shield, 5-Quart, Empire Red**  
by KitchenAid  
★★★★★ 6,331 customer reviews | 972 answered questions

List Price: \$429.99  
Price: **\$249.99 & FREE Shipping**  
You Save: **\$180.00 (42%)**

Item is eligible for **6 Month Special Financing** with your Amazon.com Store Card. [Learn more](#)

**Note:** Not eligible for Amazon Prime. Available with free Prime shipping from [other sellers on Amazon](#).

**Only 14 left in stock.**  
**Estimated Delivery Date:** July 28 - Aug. 2 when you choose Standard at checkout.  
Ships from and sold by [GE Premier](#) in [easy-to-open packaging](#).

Color: **Empire Red**



Share   

Qty:

**\$249.99** + Frt  
**Only 14 left in sto**  
**Prem**

Include 2-Year I  
**\$14.99**




 **Add**

[Turn on 1-Click orderi](#)

**Ship to:**  
KEVIN DAVENPOF

## How does Amazon improves searching and recommendation?

### Similar But Different

More Capacity	More Attachments	Different Brand
		
<input type="button" value="See more choices"/>	<input type="button" value="See Color Options"/>	<input type="button" value="See Color Options"/>
<a href="#">KitchenAid® MSE Certified®</a>	<a href="#">KitchenAid KP26M1XER 6</a>	<a href="#">Hamilton Beach 63232 Electric</a>

**Other Sellers**

**\$264.99**  
  
Sold by: Amazon.com

**\$264.99**  
+ Free Shipping  
Sold by: Marcus AV

**\$289.00**  
  
Sold by: goldentech



## 12 Suppose you are diagnosing...



### Patients

SUBJ_ID	NAME	DOB	Gender	...
12	"John"	2089-11-12	"M"	...
13	"Jane"	2117-12-25	"F"	...

### Admissions

ADM_ID	SUBJ_ID	Age	Reason	...
231	12	81	"arrhythmia"	...
232	13	56	"cancer"	...

### Diagnoses

DX_ID	ADM_ID	ICD9	...
75	231	"401.9"	...
76	232	"162.9"	...

### Medications

MED_ID	ADM_ID	Drug	Dosage	...
6	231	"amiodarone"	20	...
7	231	"lisinopril"	10	...

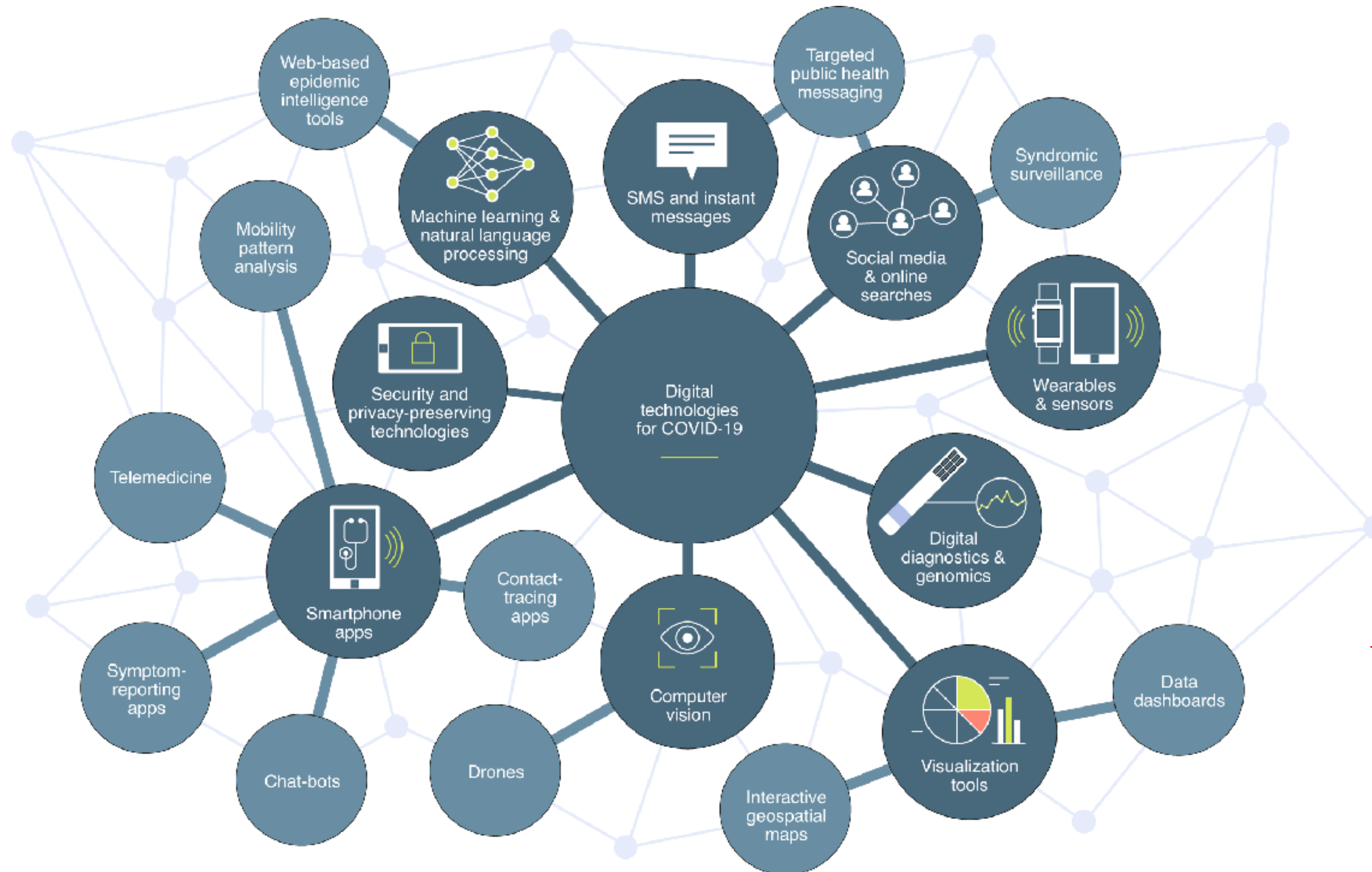
### ICD9\_Dx\_Codes

ICD_ID	ICD9	Short	Long
872	"401.1"	"Ben.Hyp."	"Benign Hypertension"
873	"401.9"	"Uns.Hyp."	"Unspecified Hypertension"

**How to better diagnose and answer questions with EHR?**

# 13 Suppose you would like to analyze COVID-19...

- You would like to build an application to **visualize and analyze COVID-19**



**What would you primarily need?**



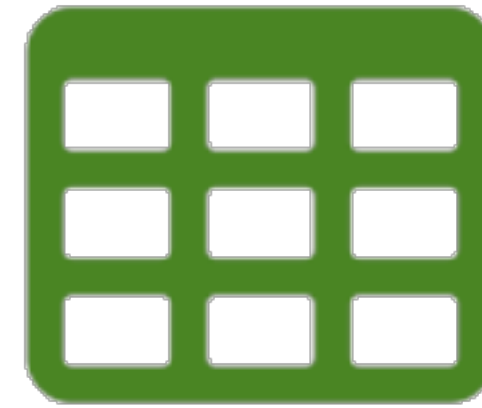
14 Data



VS.



*Closed vs. Open Data*



Tabular

VS.



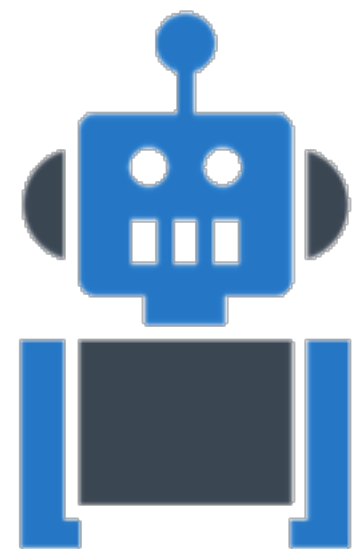
Graph

*Tabular vs. Graph*



JPG, PNG, PDF

VS.



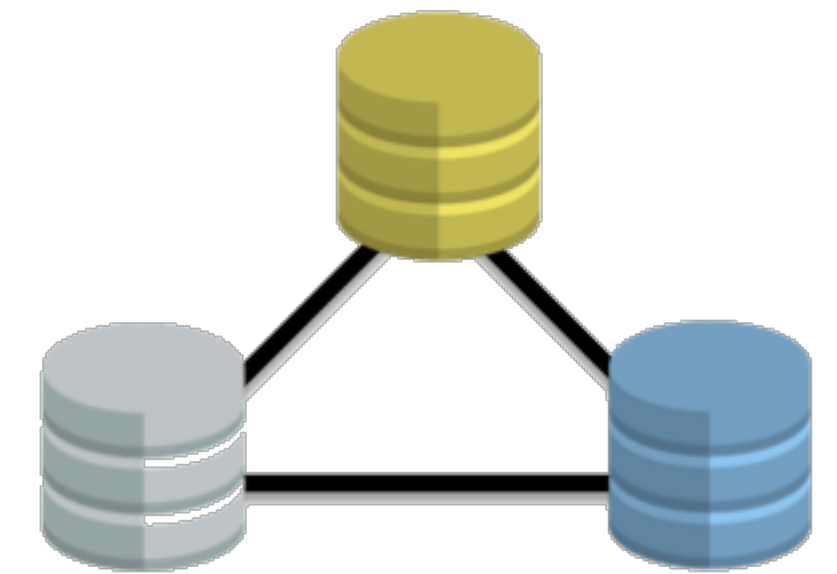
XLS, CSV

*Human Readable vs. Machine Readable*



Separated

VS.

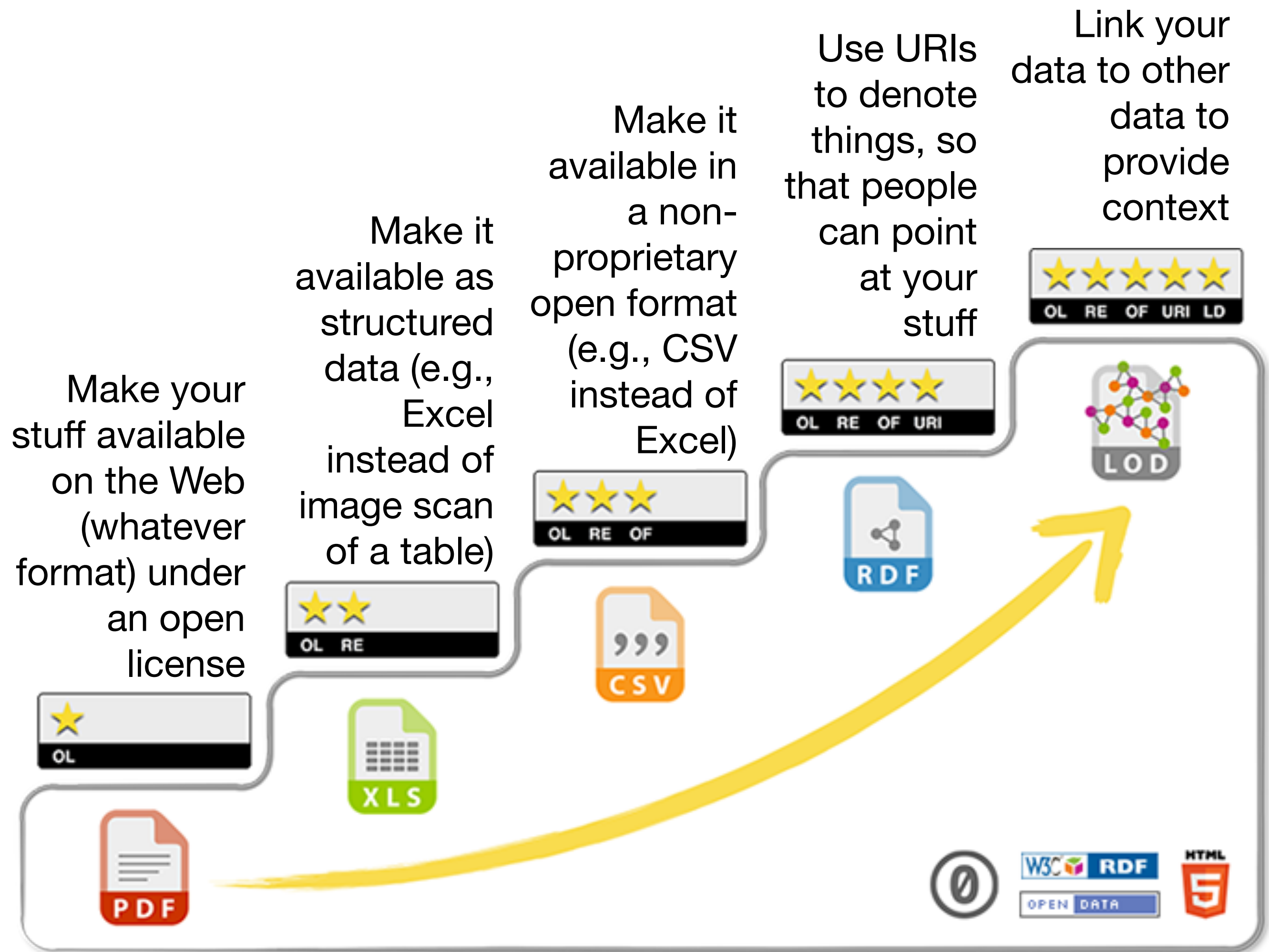


Integrated/Linked

*Separated vs. Integrated/Linked*

# 5-Star Open Data


- **Tim Berners-Lee**, the inventor of the Web and Linked Data initiator, suggested a **5-star deployment scheme** for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.





# What is Knowledge

- Plato's definition: Justified true belief

 SINCE 1828 knowledge

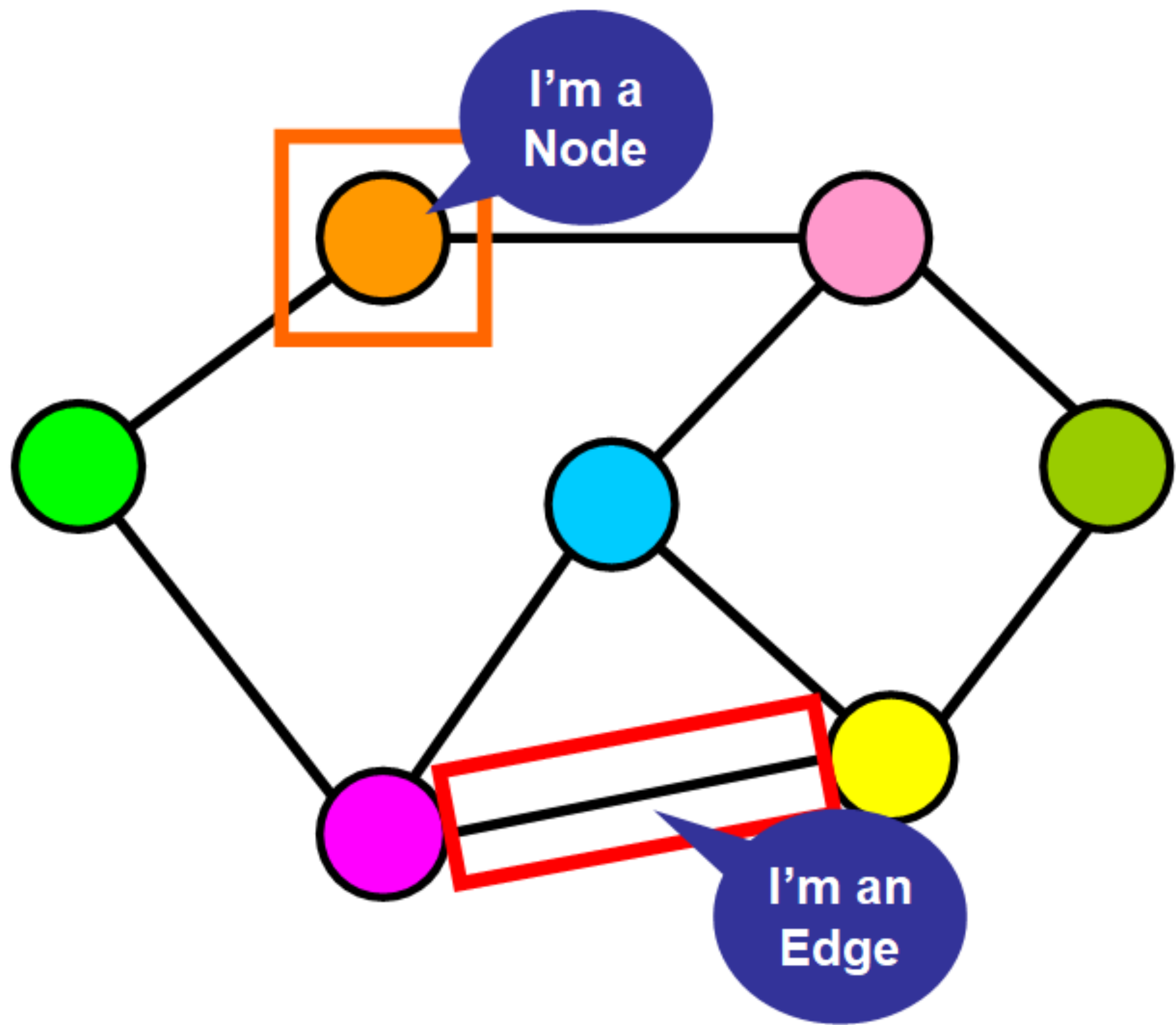
[Dictionary](#) [Thesaurus](#)

### Essential Meaning of *knowledge*

**1** : information, understanding, or skill that you get from experience or education  
*// She has little/no/some **knowledge** of fashion/history/cooking.*  
*// He has devoted himself to the pursuit of **knowledge**.*  
[See More Examples](#)

**2** : awareness of something : the state of being aware of something  
*// He claimed to have no **knowledge** of the plan. = He denied all/any **knowledge** of the plan.*  
*// Did you have any **knowledge** of her intentions?*  
*// The decision was made **without my knowledge**. [=I did not know about the decision]*

# 17 What is Graph



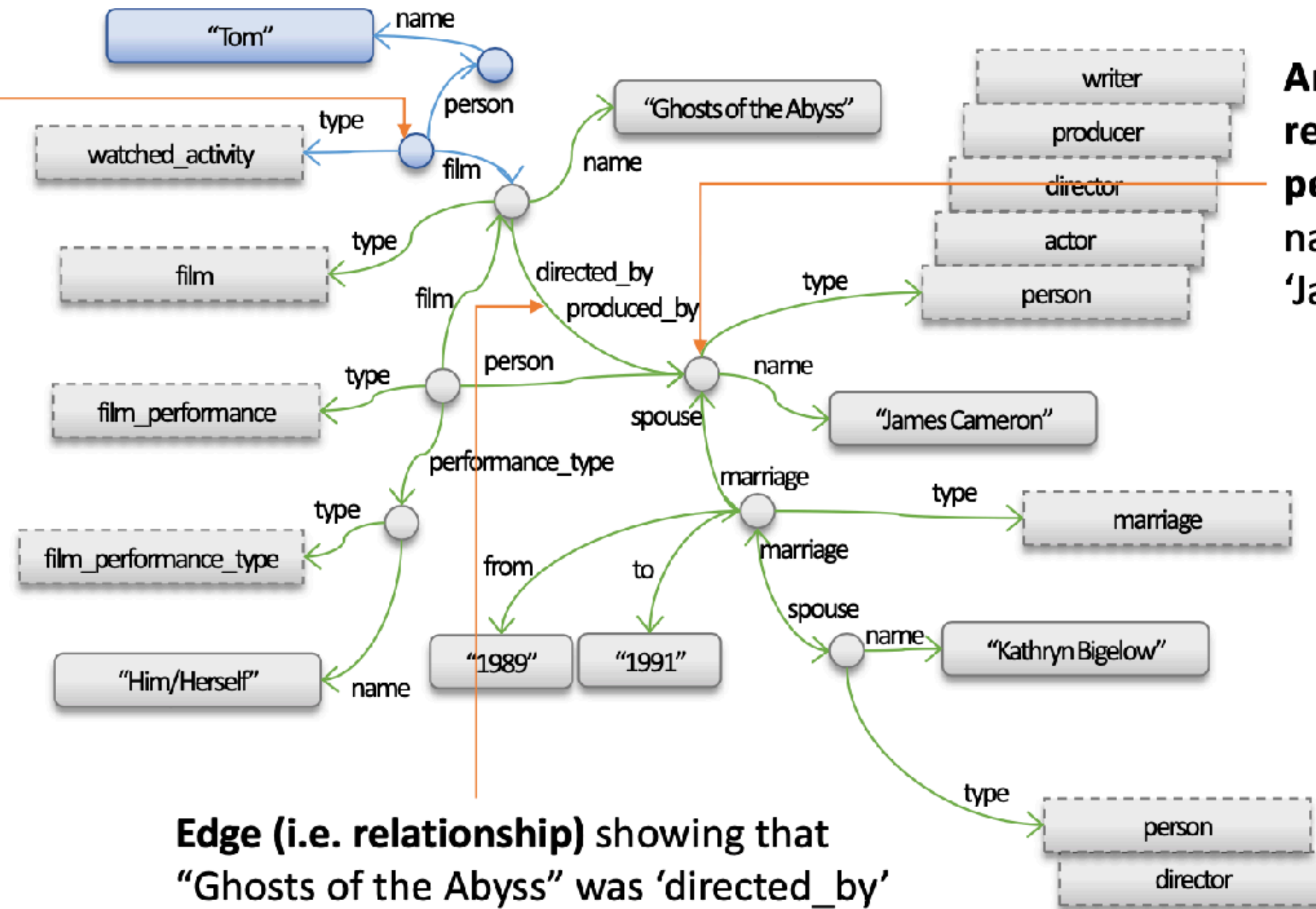


# 18 So a knowledge graph is...

- Knowledge represented as entities, edges and attributes

**Personal entity** showing that Tom watched Ghosts of the Abyss

**An entity** representing a person with name attribute 'James Cameron'



## Key concepts

<b>Entity</b>	Represent something in the real world
<b>Edge</b>	Represent relationship
<b>Attribute</b>	Represent something about an entity
<b>Ontology</b>	Definition of possible types of entities, relationships and attributes

**Edge (i.e. relationship)** showing that "Ghosts of the Abyss" was 'directed\_by' and 'produced\_by' James Cameron

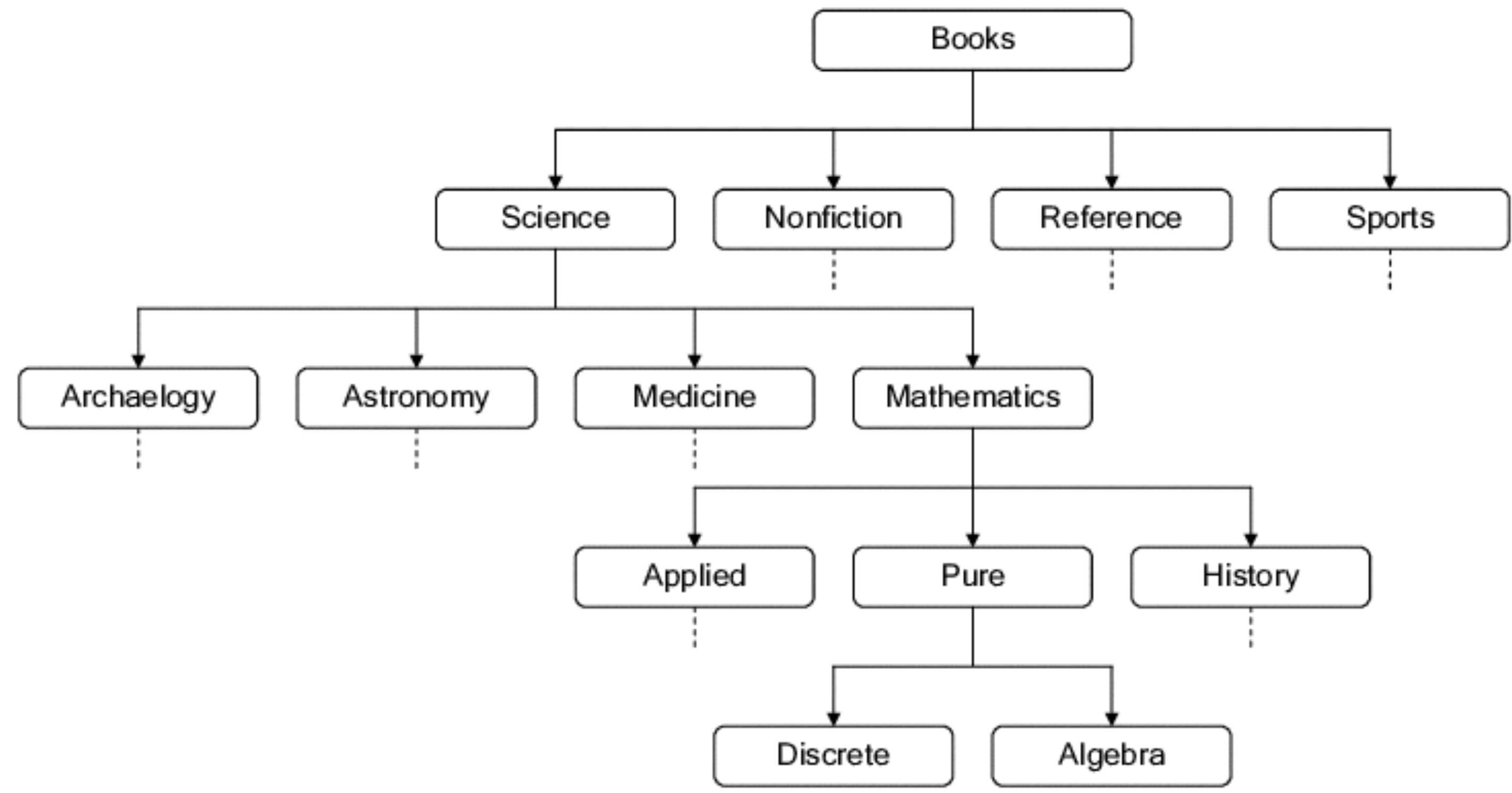
The background features a dark blue gradient with various colorful, semi-transparent geometric shapes (squares and rectangles) in shades of purple, teal, and light blue, creating a layered, architectural effect.

# Understanding the Role of Taxonomies, Ontologies, Schemas and Knowledge Graphs



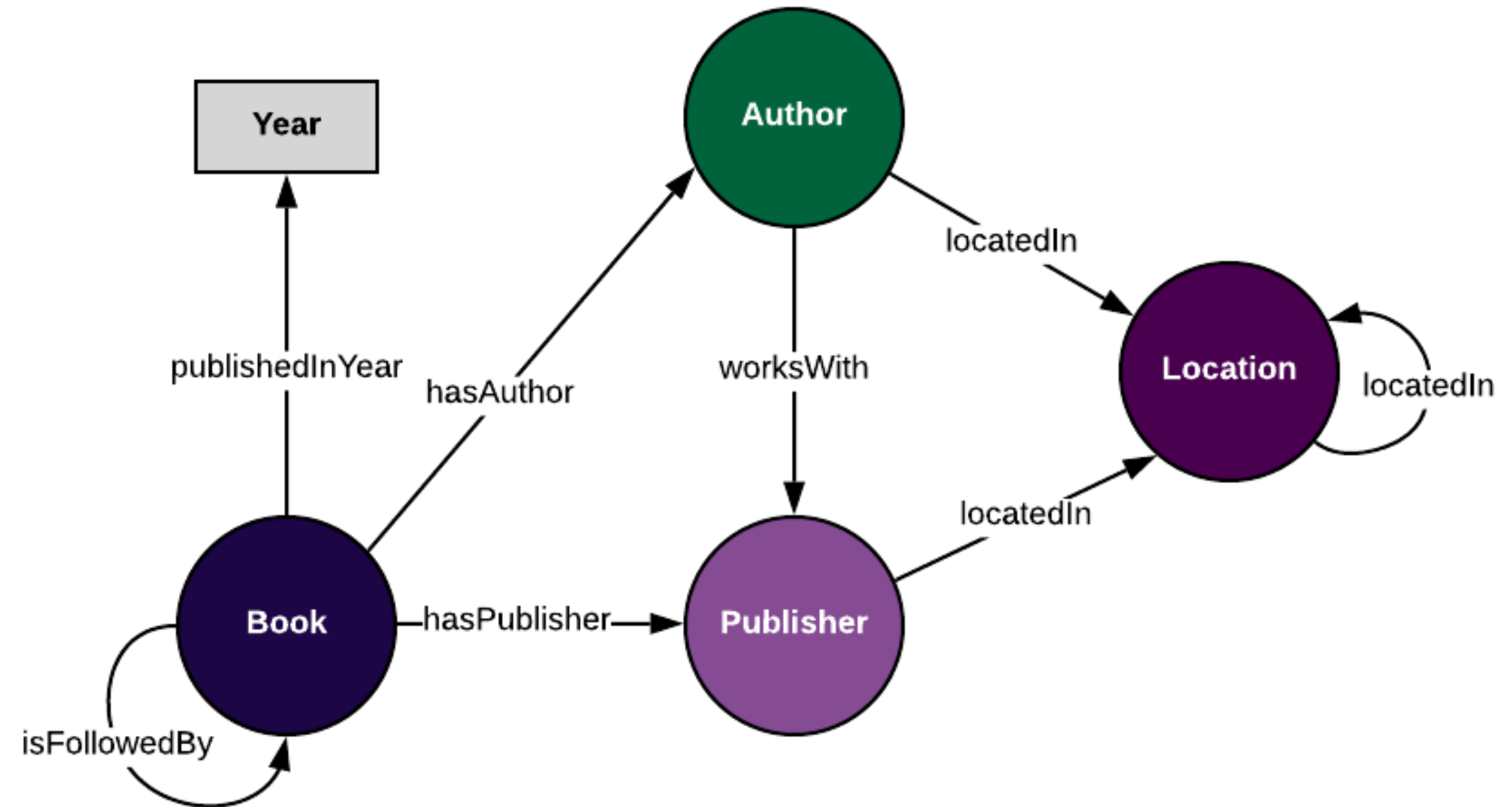
# What is a Taxonomy?

- A data taxonomy is the classification of data into categories and sub-categories.
- Taxonomy identifies hierarchical relationships within a category.



# What is an Ontology?

- ◎ Ontologies are semantic data models that define the types of things that exist in our domain and the properties that can be used to describe them. There are three main components to an ontology, which are usually described as follows:
  - **Classes:** the distinct types of things that exist in our data.
  - **Relationships:** properties that connect two classes.
  - **Attributes:** properties that describe an individual class.
  
- ◎ Ontologies are generalized data models, meaning that they only model general types of things that share certain properties, but don't include information about specific individuals in our domain.



**The main difference between Taxonomy and Ontology is that taxonomy is simpler in nature than ontology. Taxonomy takes into consideration one type of relationship, whereas ontology takes into account many different complex relationships between the concepts.**



# What is a Schema?

- A schema is the organization or structure for a database.

**Ontologies focus more on the semantic relationships whereas schemas focus more on the data structure (e.g., the data class Elephant that has the data properties: name and livesIn).**

**Books**

Title	Author	Publisher	Year Published	Followed By
To Kill a Mockingbird	Harper Lee	J. B. Lippincott Company	1960	Go Set a Watchman
Go Set a Watchman	Harper Lee	HarperCollins, LLC; Heinemann	2015	
The Picture of Dorian Gray	Oscar Wilde	J. B. Lippincott & Co.	1890	
2001: A Space Odyssey	Arthur C. Clarke	New American Library, Hutchinson	1968	

**Publishers**

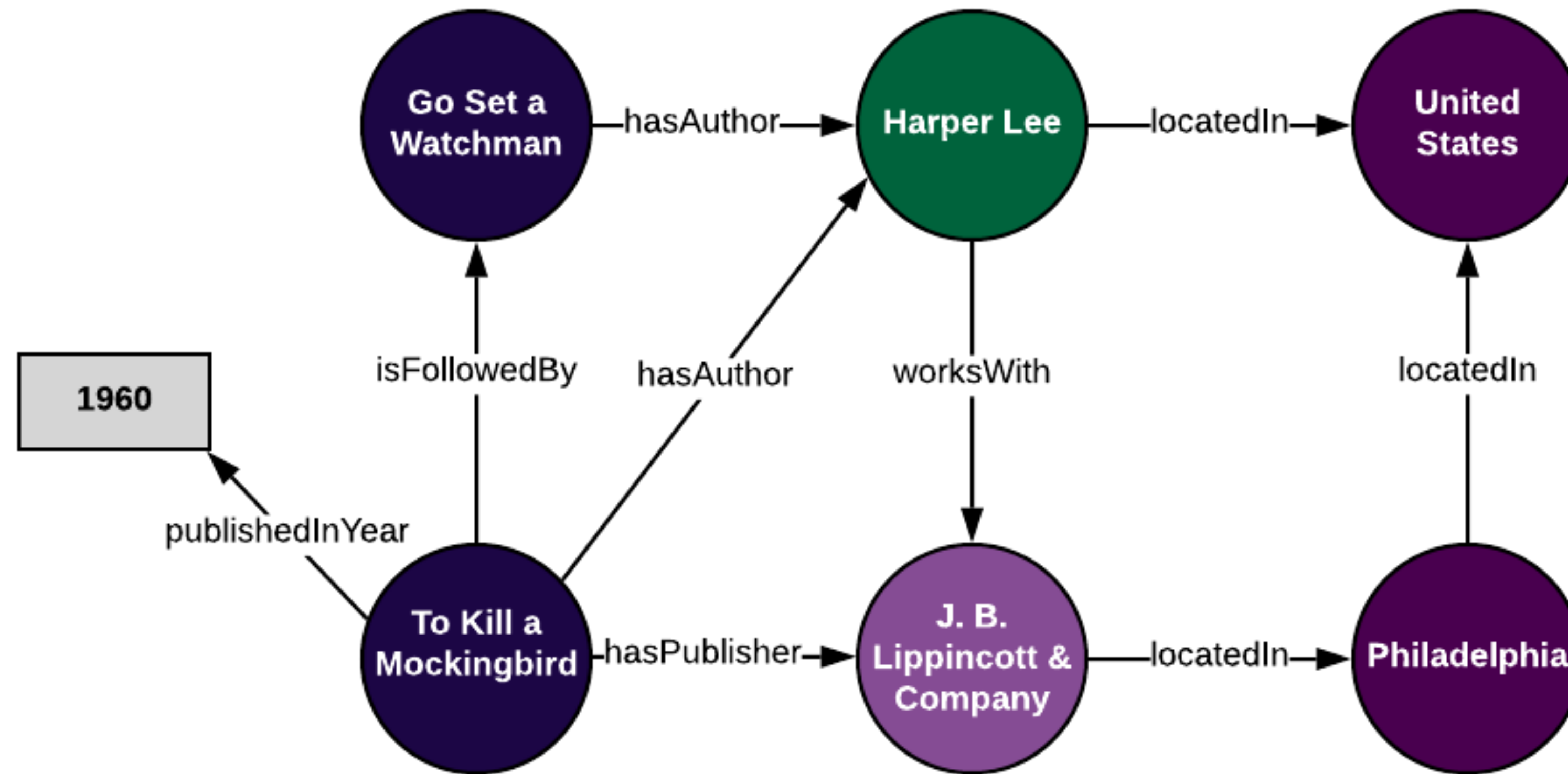
Name	City	Country
J. B. Lippincott & Company	Philadelphia	United States
HarperCollins, LLC	New York City	United States
Heinemann	Portsmouth	United States
New American Library	New York City	United States
Hutchinson	London	United Kingdom

**Authors**

Name	Country of Birth
Harper Lee	United States
Oscar Wilde	Ireland
Arthur C. Clarke	United Kingdom

# What is a Knowledge Graph?

- Using our ontology as a framework, we can add in real data about individual books, authors, publishers, and locations to create a knowledge graph.



ontology + data = knowledge graph



# Why Knowledge Graph

## Why Knowledge Graphs?

- **Graphs are a natural way** to represent entities and their relationships
- Graphs can **capture a broad spectrum of data** (structured / unstructured)
- Graphs can be managed **efficiently**

**Game-changing data integration**

**Robust data quality assurance**

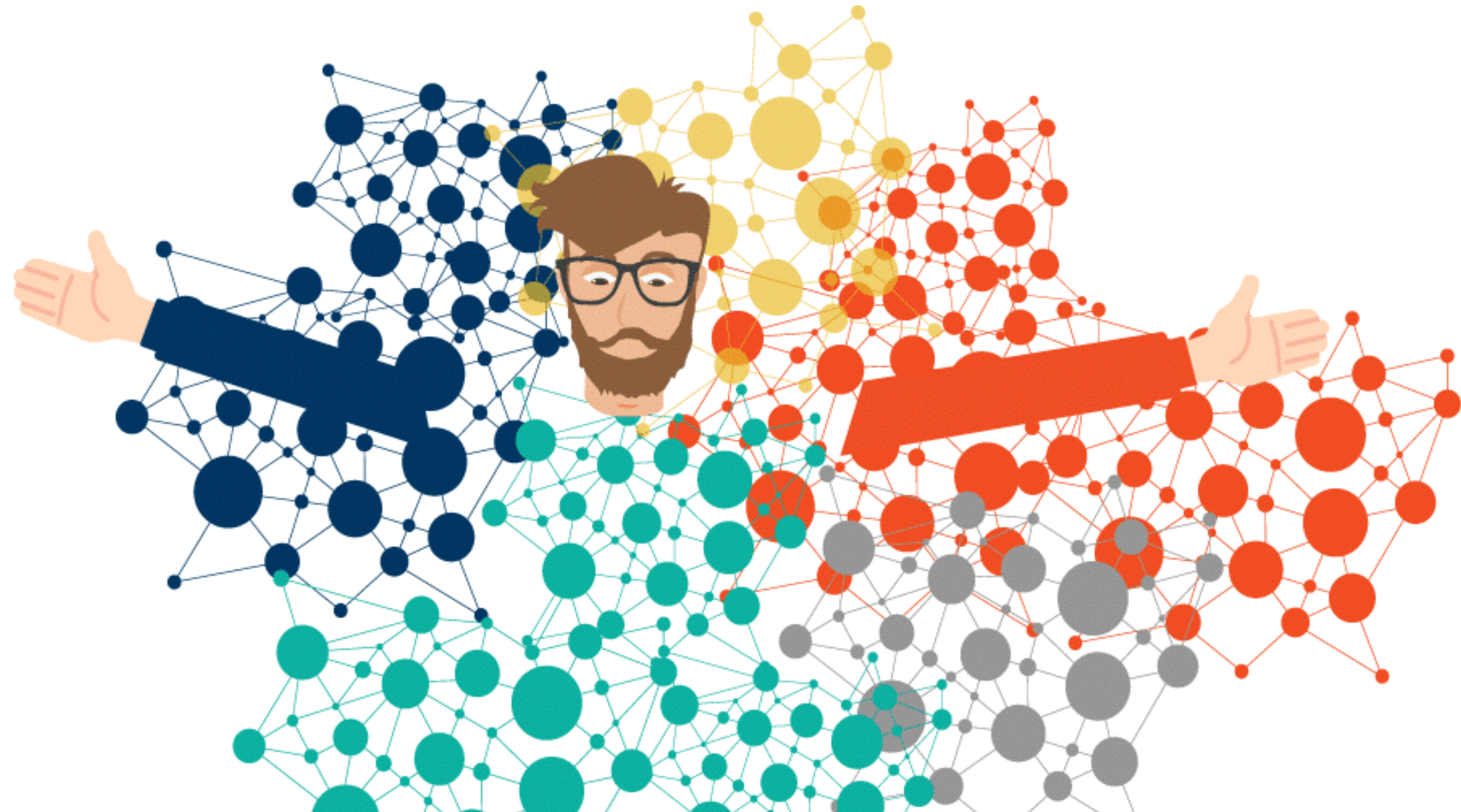
**Intuitive domain modelling**

**Flexibility & performance**

**Low up-front investment**

## 25 Minimum set of characteristics of knowledge graphs

- Mainly describes real world entities and their interrelations, organized in a graph.
- Defines possible classes and relations of entities in a schema.
- Allows for potentially interrelating arbitrary entities with each other.
- Covers various topical domains.





# Open Knowledge Graphs

Open knowledge graphs are published online, making their content accessible for the public good.

- DBpedia
- Freebase
- Wikidata
- YAGO
- .....



Cover many domains and are either extracted from **Wikipedia**, or built by **communities of volunteers**



WIKIPEDIA  
The Free Encyclopedia



# State of the art knowledge graphs



***Cyc is a long-term artificial intelligence project that aims to assemble a comprehensive ontology and knowledge base that spans the basic concepts and rules about how the world works. Hoping to capture common sense knowledge, Cyc focuses on implicit knowledge that other AI platforms may take for granted.***



***Freebase was a large collaborative knowledge base consisting of data composed mainly by its community members. It was an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions. [3][2] Freebase aimed to create a global resource that allowed people (and machines) to access common information more effectively.***



***Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.***

# State of the art knowledge graphs



*DBpedia is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web.*



*YAGO is an open source knowledge base developed at the Max Planck Institute for Computer Science in Saarbrücken. It is automatically extracted from Wikipedia and other sources.*

## Read the Web

Research Project at Carnegie Mellon University

*Never-Ending Language Learning system (NELL) is a semantic machine learning system developed by a research team at Carnegie Mellon University, and supported by grants from DARPA, Google, NSF, and CNPq with portions of the system running on a supercomputing cluster provided by Yahoo!*



# 29 NELL: Never-Ending Language Learning

## Recently-Learned Facts

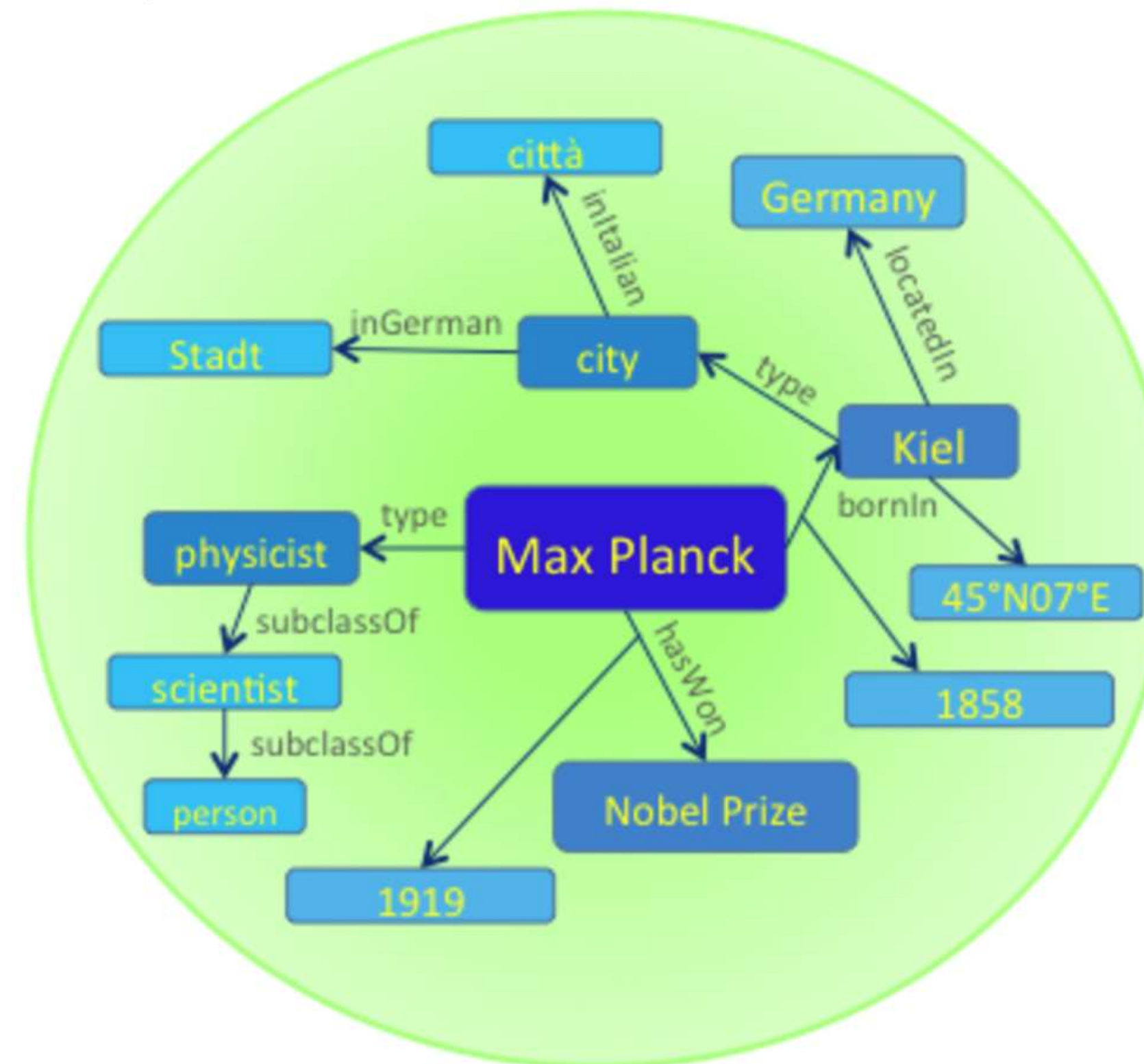


Refresh

instance	iteration	date learned	confidence		
<u>high_end_clock</u> is an <u>item found on a wall</u>	1111	06-jul-2018	99.6		
<u>movie edward scissorhands</u> is a <u>movie</u>	1111	06-jul-2018	100.0		
<u>n1999 toyota camry</u> is a <u>model of automobile</u>	1111	06-jul-2018	94.7		
<u>michael i</u> is a <u>monarch</u>	1111	06-jul-2018	96.7		
<u>gonzales music festival 1968</u> is a <u>music festival</u>	1111	06-jul-2018	100.0		
<u>roof</u> is a building feature <u>made from</u> the building material <u>steel</u>	1115	03-sep-2018	92.7		
<u>tyco healthcare</u> <u>hired</u> <u>dennis kozlowski</u>	1112	24-jul-2018	100.0		
<u>air pacific</u> is a company that <u>has an office in</u> the city <u>los angeles ca</u>	1112	24-jul-2018	99.2		
the sports team <u>yankees</u> was the <u>winner of</u> <u>n1962 world series</u>	1116	12-sep-2018	100.0		
<u>silvestre herrera</u> <u>died in</u> the city <u>glendale</u>	1116	12-sep-2018	100.0		

# 30 YAGO

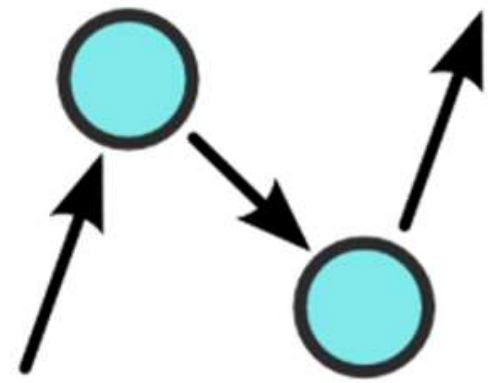
- **Input:** Wikipedia infoboxes, WordNet and GeoNames
- **Output:** KG with 350K entity types, 10M entities, 120M facts
- Temporal and spatial information



<https://yago-knowledge.org/>



# 31 ConceptNet



## ConceptNet

An open, multilingual knowledge graph

[Link](#)

## en beatles

An English term in ConceptNet 5.5

### Derived terms

- [en](#) [beatle](#) →
- [en](#) [beatledom](#) →
- [en](#) [beatlemania](#) →
- [en](#) [beatlesque](#) →
- [en](#) [fourth beatle](#) →

### beatles is a type of...

- [en](#) [a British band](#) →
- [en](#) [man](#) <sup>(n)</sup> →
- [en](#) [band](#) <sup>(n)</sup> →
- [en](#) [musician](#) <sup>(n)</sup> →
- [en](#) [album](#) <sup>(n)</sup> →

### Links to other sites

- [dbpedia.org](#) [The Beatles](#) →
- [sw.opencyc.org](#) [Beatle](#) →
- [umbel.org](#) [Beatle](#) →
- [wordnet-rdf.princeton.edu](#) [400520405-N](#) →
- [wordnet-rdf.princeton.edu](#) [108386847-n](#) →
- [wikidata.dbpedia.org](#) [Q1299](#) →
- [en.wiktionary.org](#) [Beatles](#) →
- [dbpedia.org](#) [The Beatles \(No. 1\)](#) →
- [wikidata.dbpedia.org](#) [Q738260](#) →
- [fr.wiktionary.org](#) [Beatles](#) →
- [dbpedia.org](#) [The Beatles \(The Original Studio Recordings\)](#) →
- [wikidata.dbpedia.org](#) [Q603122](#) →

<https://conceptnet.io/>



# Enterprise Knowledge Graphs

- ◎ Enterprise knowledge graphs are typically **internal to a company** and applied for **commercial use-cases**
  - **Web search:** Bing, Google, Airbnb, Amazon, eBay, Uber, ...
  - **Social network:** Facebook, LinkedIn, ...
  - **Finance:** Accenture, Bloomberg, Capital One, Wells Fargo, ...
- ◎ Applications include search, recommendations, personal agents, advertising, business analytics, risk assessment, automation, and more besides.

33 **Recap: suppose you are searching...**



**What he is interested in?**

**Query: "Honda Civic, Hyundai Elantra"**

# 34 Recap: Is-A relationship between things

Honda Civic



Hyundai Elantra



◆ **Fuel-efficient cars**

◆ **Economy cars**

◆ **Marvel heroes**

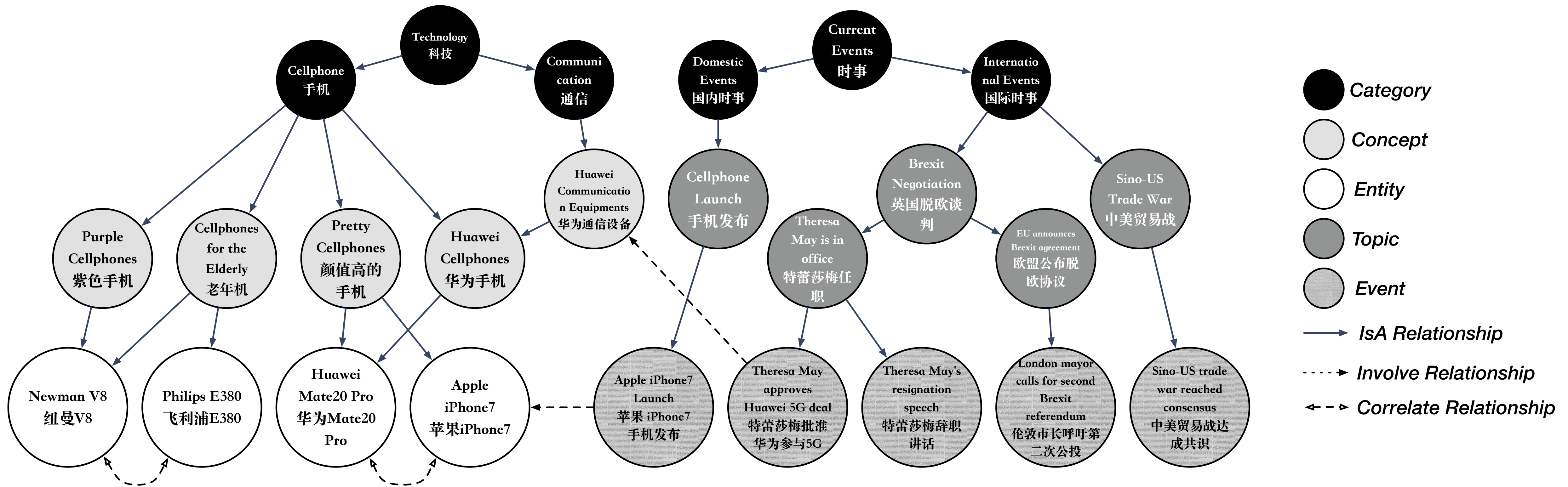
◆ **Revenagers**

**Concept:** a collection of things that share some common attributes



# 35 Tencent GIANT Attention Ontology

**Create a web-scale ontology to represent user interests and document topics.**

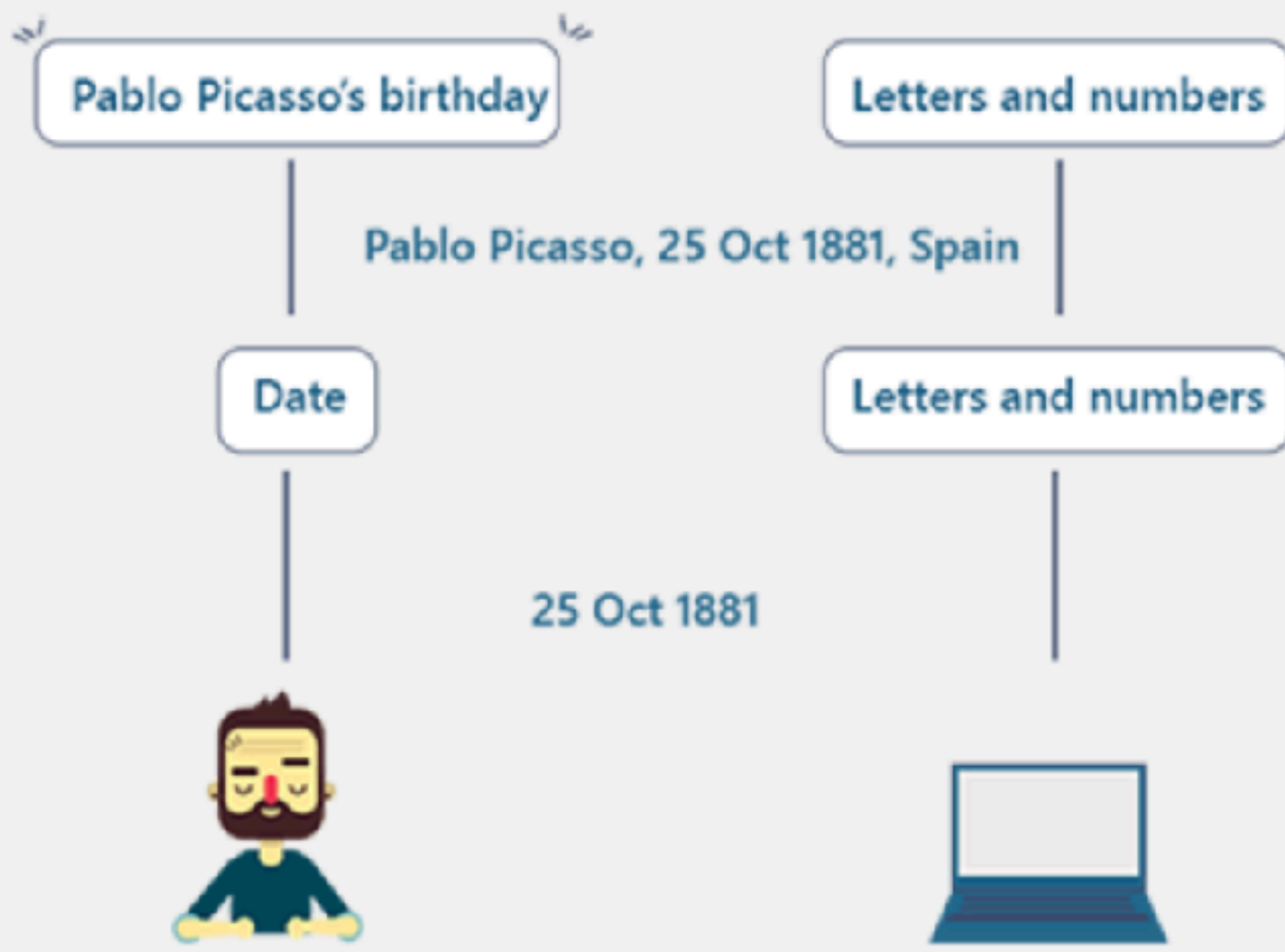


# Microsoft Concept Graph

## Microsoft Concept Graph

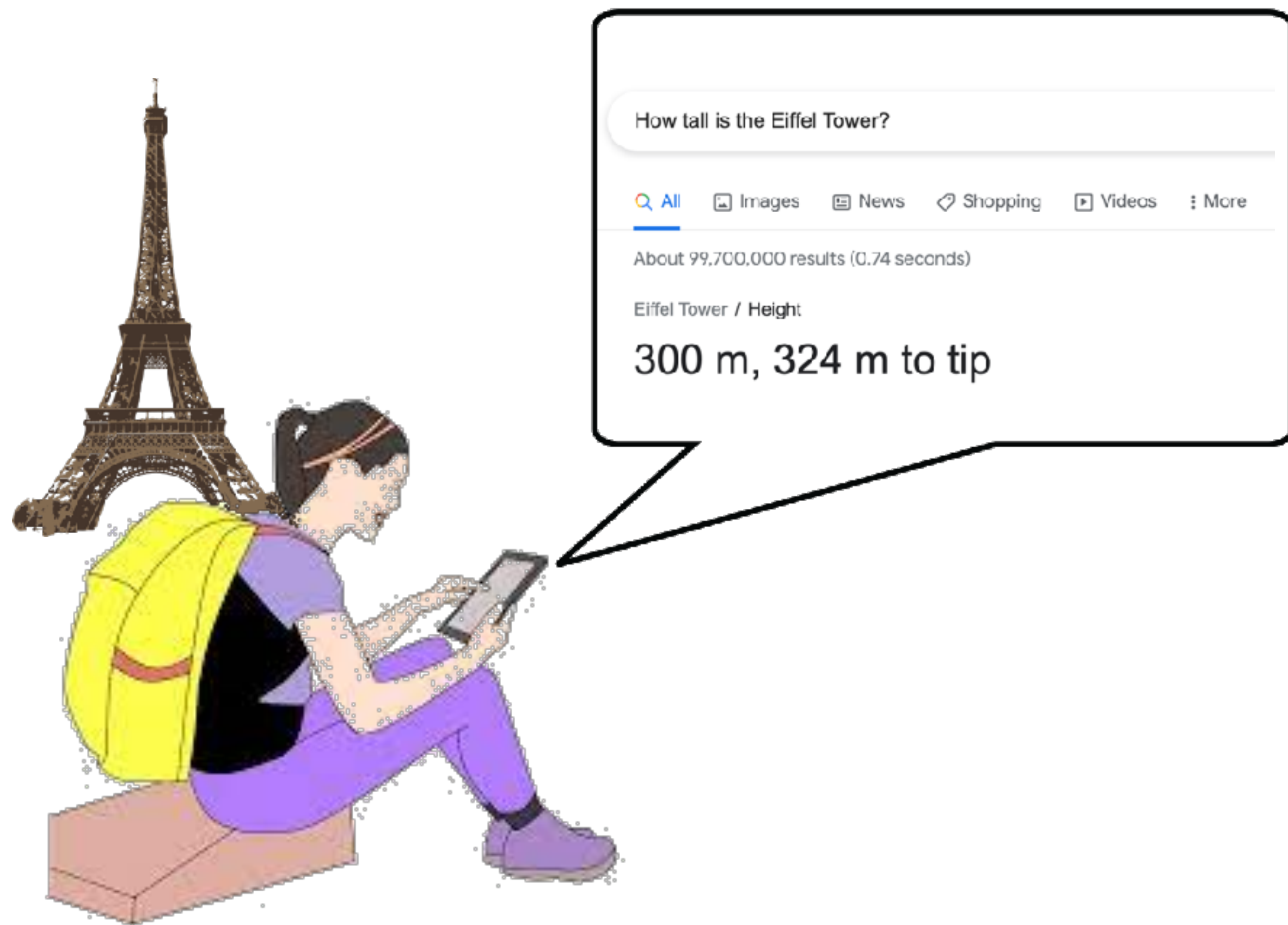
Our goal is to enable machines to better understand human communication. An important question is, what does the word “understand” mean here? Consider the following example. For human beings, when we see “25 Oct 1881”, we recognize it as a date, although most of us do not know what it is about. However, if we are given a little more context, say the date is embedded in the following piece of short text “Pablo Picasso, 25 Oct 1881, Spain”, most of us would have guessed (correctly) that the date represents Pablo Picasso’s birthday. We are able to do this because we possess certain knowledge, and in this case, “one of the most important dates associated with a person is his birthday.”

As another example, consider a problem in natural language processing. Humans do not find sentences such as “animals other than dogs such as cats” ambiguous, but machine parsing can lead to two possible understandings: “cats are animals” or “cats are dogs.” Common sense tells us that cats cannot be dogs, which renders the second parsing improbable.





# 37 Recap: suppose you are asking...

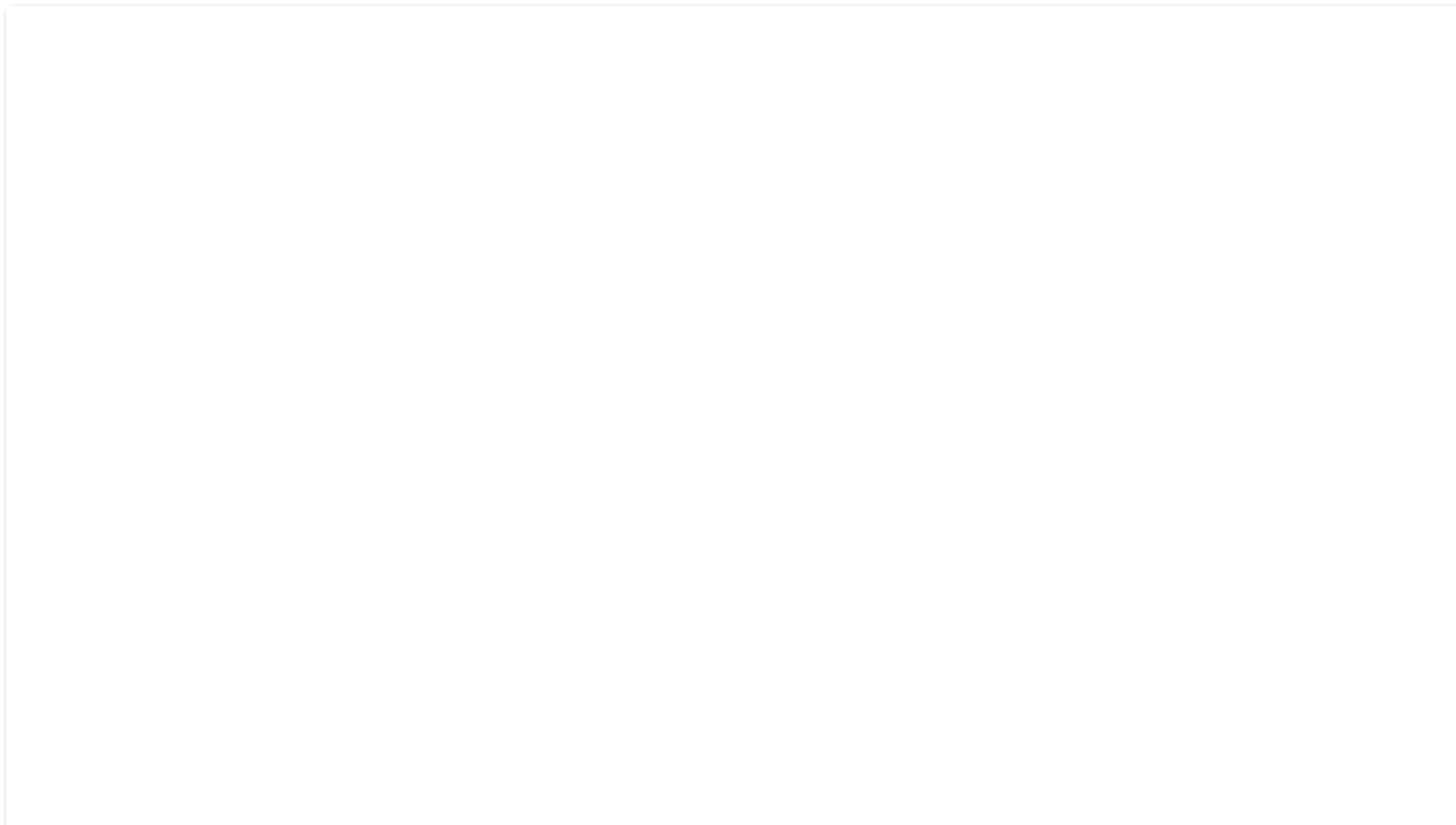


**How does Google know?**

**Question: “How tall is the Eiffel Tower?”**



# 38 Google Knowledge Graph



<https://blog.google/products/search/introducing-knowledge-graph-things-not/>

# 39 Recap: suppose you are shopping...



Roll over image to zoom in

<b>Brand</b>	Cetaphil
<b>Ingredients</b>	Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.
<b>Scent</b>	Fragrance free
<b>Additional Item Information</b>	Non-Comedogenic, Fragrance-free, Natural
<b>Skin Type</b>	Sensitive

## About this item

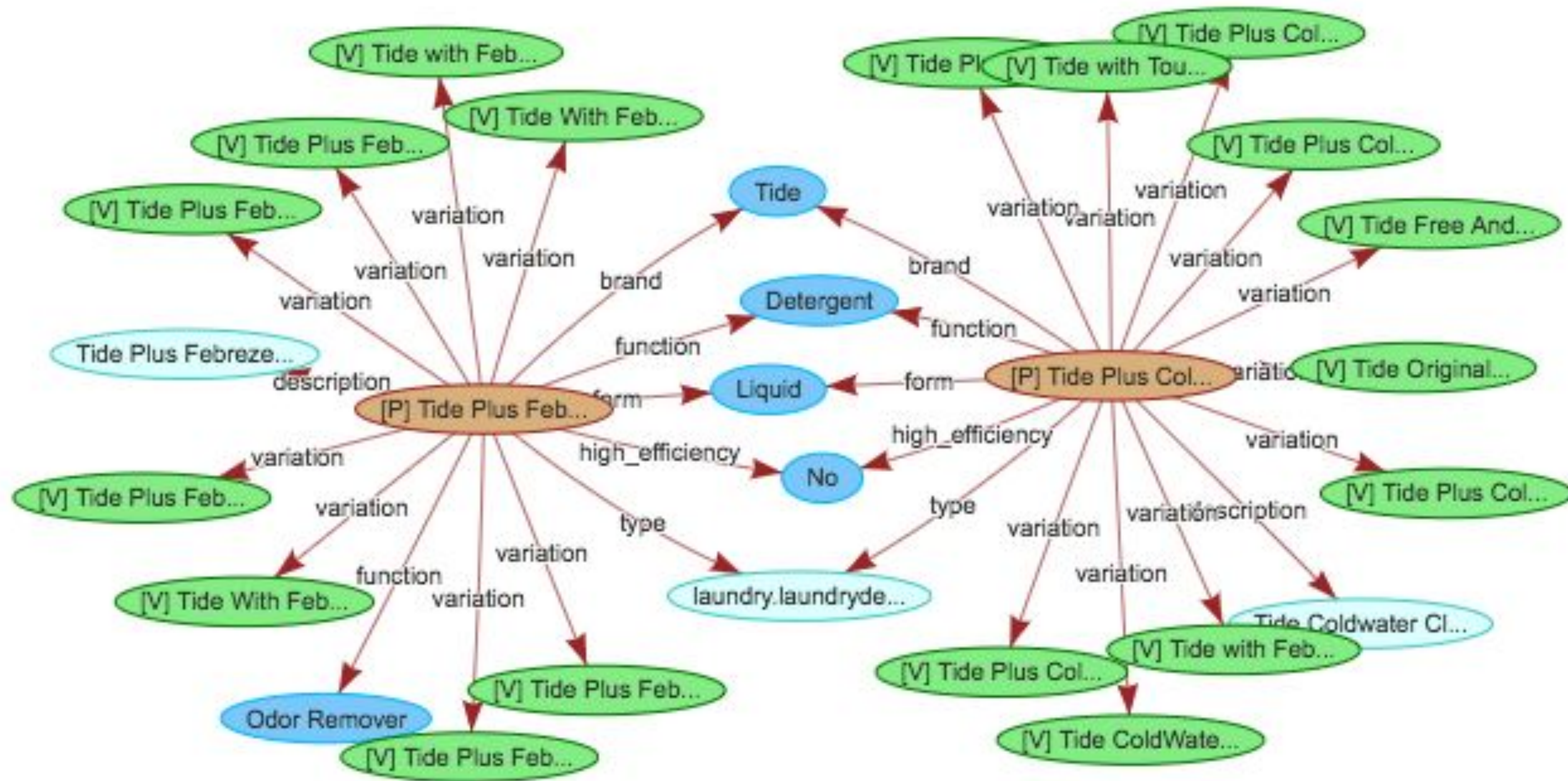
- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean
- Mild & non irritating: Soap free formulation won't strip skin of its natural protective oils and emollients

**How does Amazon provide the information?**

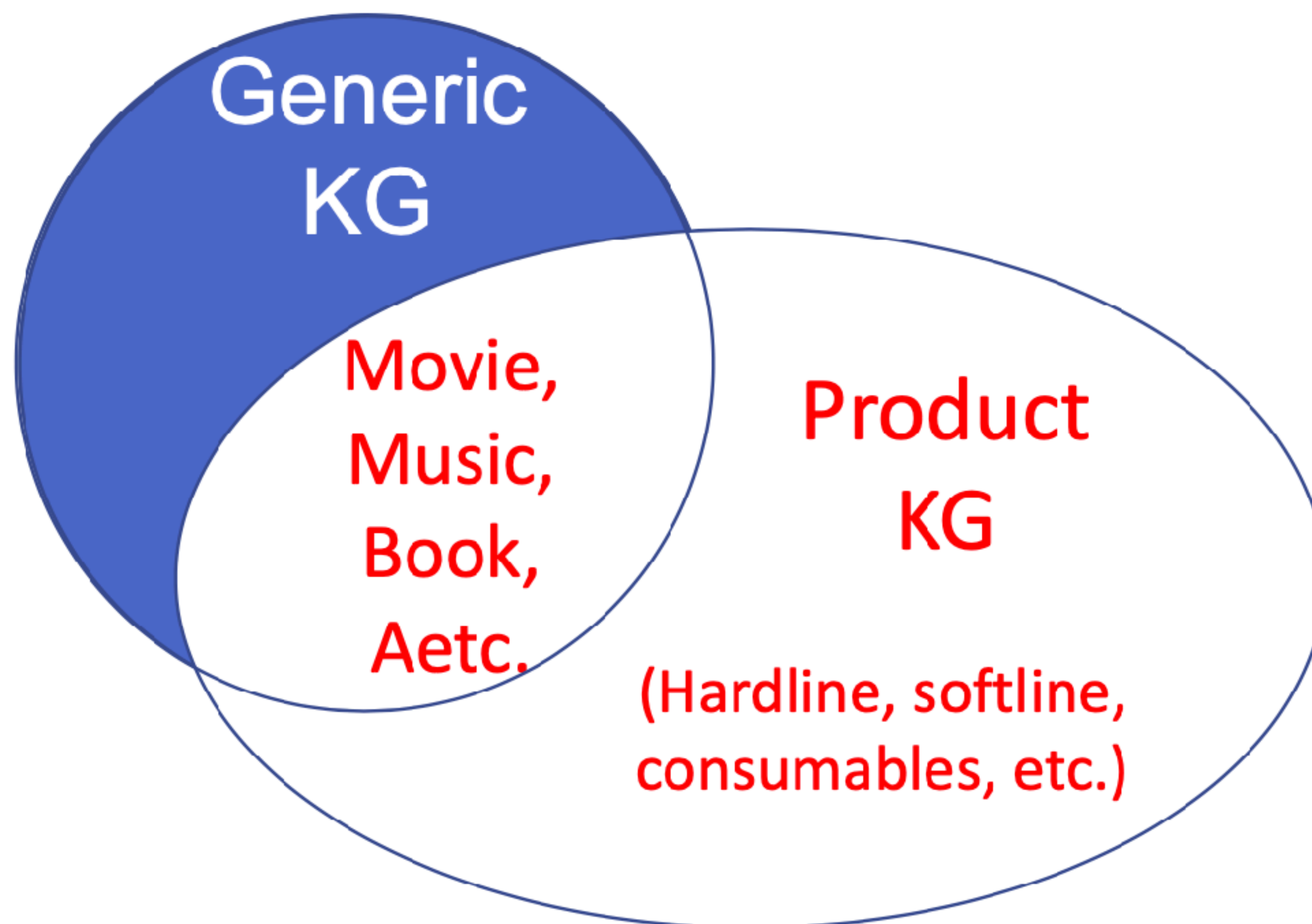


# Amazon Product Knowledge Graph

## Product Graph Example for 2 Products

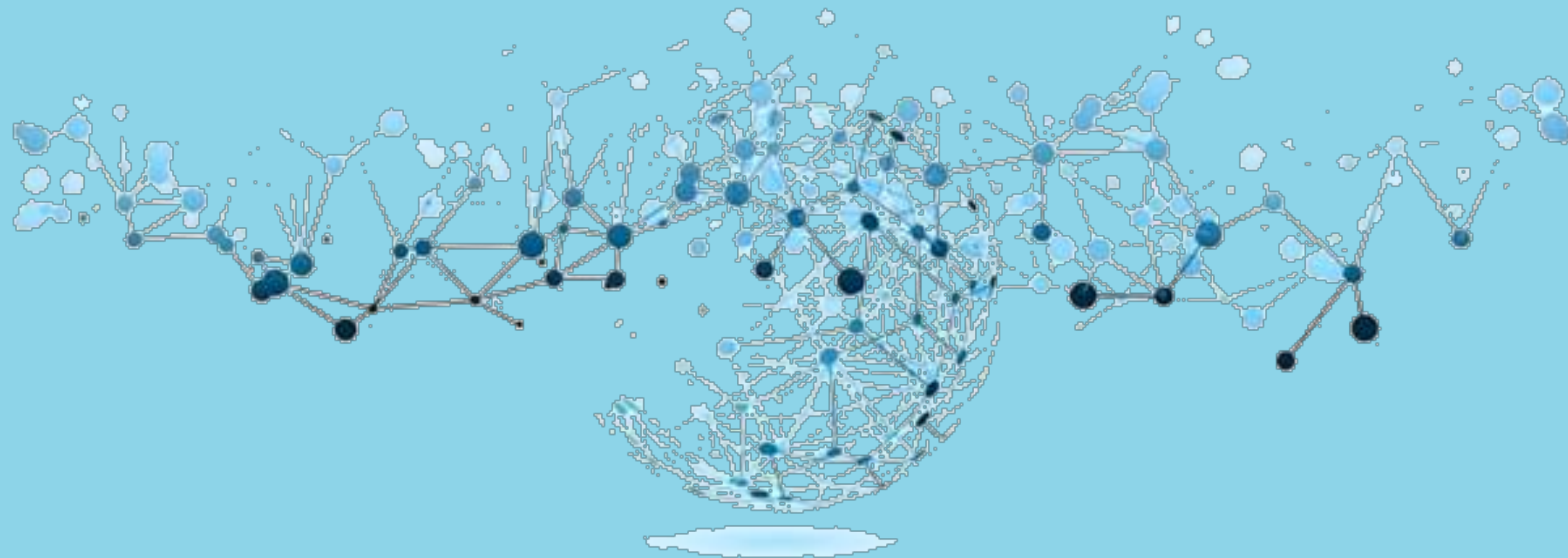


# 41 Product Graph vs. Knowledge Graph





# Semantic Web Knowledge Graph







- W3C semantic web or linked data standards are widely used in the development of knowledge graphs. JSON-LD, in fact, is one of those standards.
- “**Semantic web**” is a term that originated with web pioneer and World Wide Web Consortium founder Tim Berners-Lee (TBL) in an article in *Scientific American* in 1999, one that described a vision of a relationship-rich, scalable, contextual data web, a web of graph-oriented and well-described data that could live side by side with existing web content.
- Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data.
- Linked data are empowered by technologies such as **RDF, SPARQL, JSON-LD, OWL, SHACL and SKOS**.

## Semantic Web vs. KG

- Semantic Web is based on RDF (Resource Description Framework) and OWL (Web Ontology Language).
- Google Knowledge Graph is based on its API , which in turn is based on Schema in JSON-LD format and which in turn is also based on RDF or RDF Schema. JSON-LD is a lightweight Linked Data format. It is easy for humans to read and write. It is based on the already successful JSON format and provides a way to help JSON data interoperate at Web-scale.
- There are different technologies to build and operate a knowledge graph. **A knowledge graph that employs Semantic Web standards and technologies is called a Semantic Web Knowledge Graph.**



the data part



the schema part

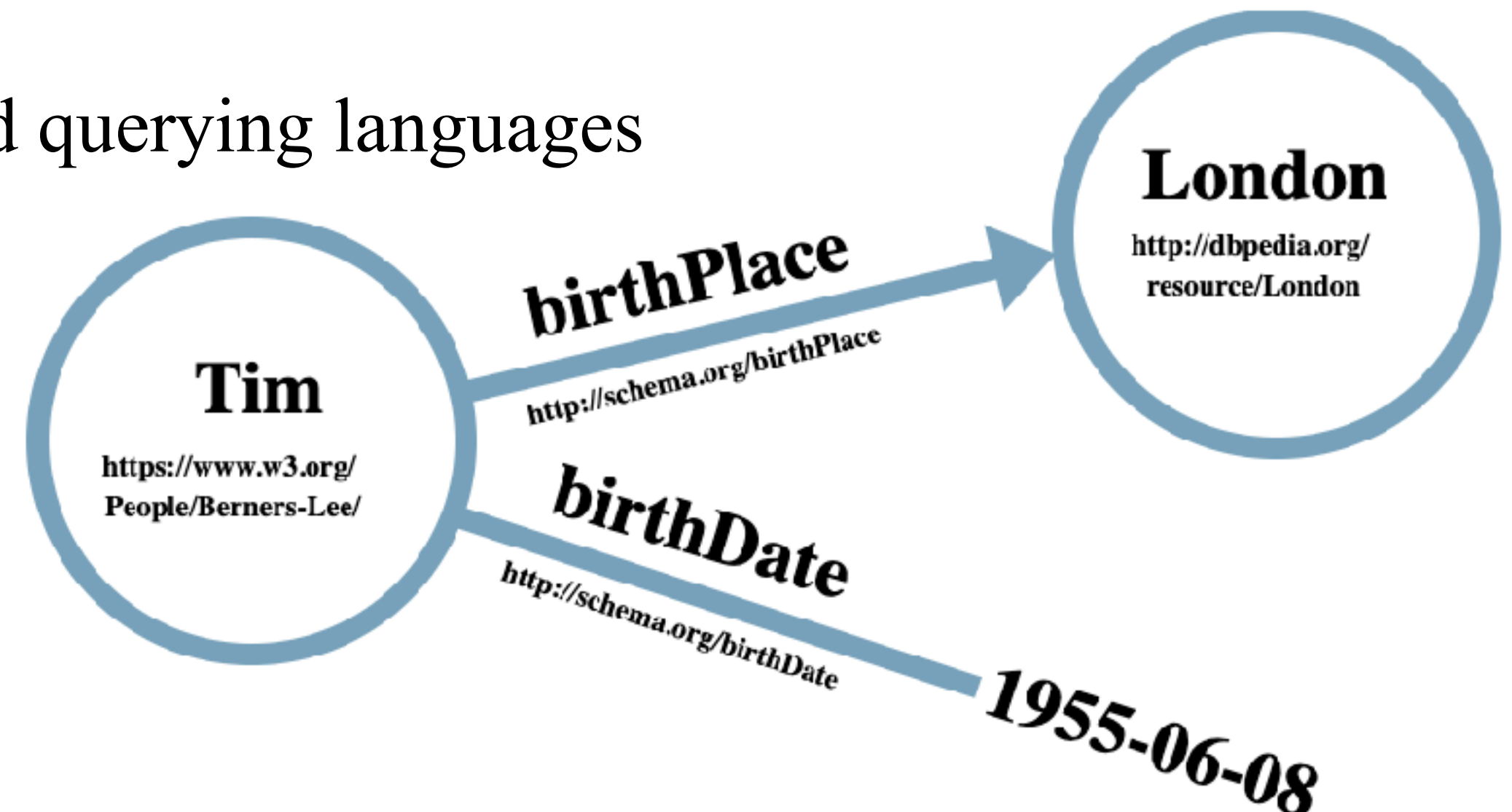


the query part



# Resource Description Framework (RDF)

- Data model, based on S-P-O triple structure (Subject, Predicate, Object)
- Used for describing things, yes, every, single, thing And anyways, **RDF = Resource Description Framework**
- Four Key Principles for RDF Graphs
  - Uniform Resource Identifiers (URIs) as identifiers
  - HTTP URI as information access protocol
  - RDF and SPARQL as knowledge representation and querying languages
  - Link information from one source to other



# RDF Schema (RDFS)

- RDFS is a vocabulary for modeling schemas used in RDF data.
- RDFS provides mechanisms for defining classes and properties.
- Via RDFS, e.g., one can provide the domains and ranges of properties.
- The official document of RDFS is available: <https://www.w3.org/TR/rdf-schema/>



## RDF Schema 1.1

[W3C Recommendation 25 February 2014](#)

**This version:**

<http://www.w3.org/TR/2014/REC-rdf-schema-20140225/>

**Latest published version:**

<http://www.w3.org/TR/rdf-schema/>

**Previous version:**

<http://www.w3.org/TR/2014/PER-rdf-schema-20140109/>

**Editors:**

[Dan Brickley](#), Google

R.V. Guha, Google

**Previous Editors:**

Brian McBride

## 48 Simple Knowledge Organization System (SKOS)

- SKOS is short for **Simple Knowledge Organization System**.
- It is a vocabulary for expressing the basic structure and content of concept schemes such as thesauri and taxonomies.
- SKOS allows concepts and relationships among concepts to be composed and published on the World Wide Web.
- Official documentation links of SKOS:  
<https://www.w3.org/TR/skos-primer/>  
<https://www.w3.org/TR/skos-reference/>

### SKOS & RDF

SKOS provides a standard way to represent knowledge organization systems using the [Resource Description Framework \(RDF\)](#). Encoding this information in RDF allows it to be passed between computer applications in an interoperable way.

Using RDF also allows knowledge organization systems to be used in distributed, decentralised metadata applications. Decentralised metadata is becoming a typical scenario, where service providers want to add value to metadata harvested from multiple sources.



# 49 SPARQL

- If RDF captures knowledge, then SPARQL retrieves knowledge, querying knowledge captured by RDF!
- Short for: SPARQL Protocol and RDF Query Language  
SPARQL documentation: <https://www.w3.org/TR/rdf-sparql-query/>

● Example:

```
@prefix ex:    <https://example.org/resource/> .
@prefix prop: <https://example.org/property/> .

ex:matt prop:knows ex:scott .
ex:matt prop:twitterAccount <https://twitter.com/mattt7812> .
```

```
SELECT ?twitter WHERE {
  ?person prop:twitterAccount ?twitter .
  ?person prop:knows ex:scott }
```

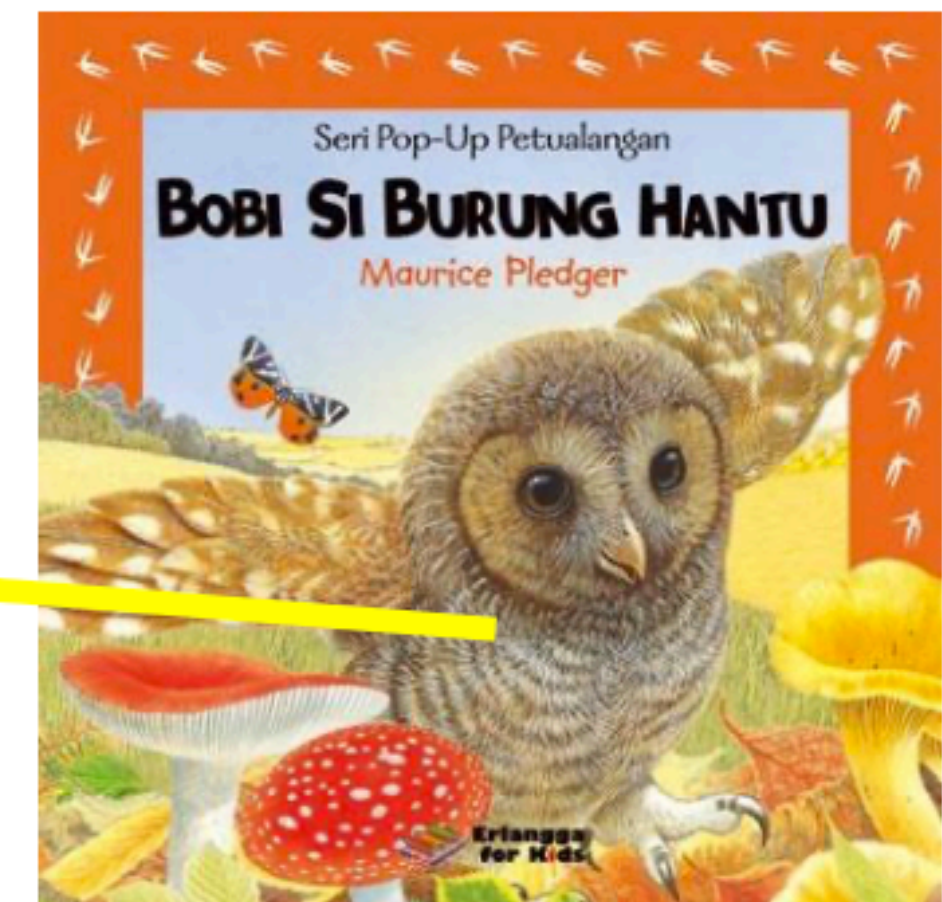
*What are the Twitter accounts of people knowing Scott?*



# 50 Web Ontology Language (WOL? No, it is OWL!)

- Short for: Web Ontology Language. OWL is a language to create ontologies
- OWL is more expressive/heavyweight compared to RDFS
- Key features:
  - Vocabulary description
  - Reasoning: You can conclude new things based on existing facts!  
For example: owl subClassOf bird + bird subClassOf animal Now, if Bobi is an owl, do you think Bobi is an animal?
- Documentation:  
<https://www.w3.org/TR/owl-primer/>

Yes!





# More on Semantic Web

- <https://www.w3.org/standards/semanticweb/>

**SEMANTIC WEB**

On this page → [technology topics](#) • [news](#) • [upcoming events and talks](#)

In addition to the classic “Web of documents” W3C is helping to build a technology stack to support a “Web of data,” the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term “Semantic Web” refers to W3C’s vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

### Linked Data

The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files.

### Vocabularies

At times it may be important or valuable to organize data. Using OWL (to build vocabularies, or “ontologies”) and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

### Query

Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data. SPARQL is the query language for the Semantic Web.

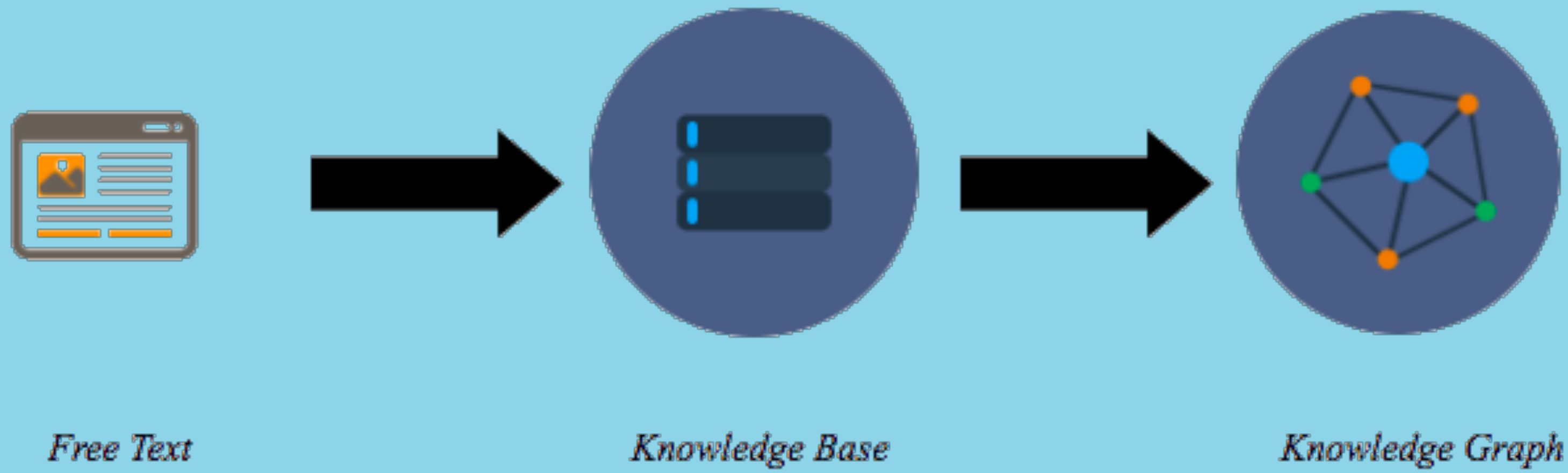
### Inference

Near the top of the Semantic Web stack one

### Vertical Applications

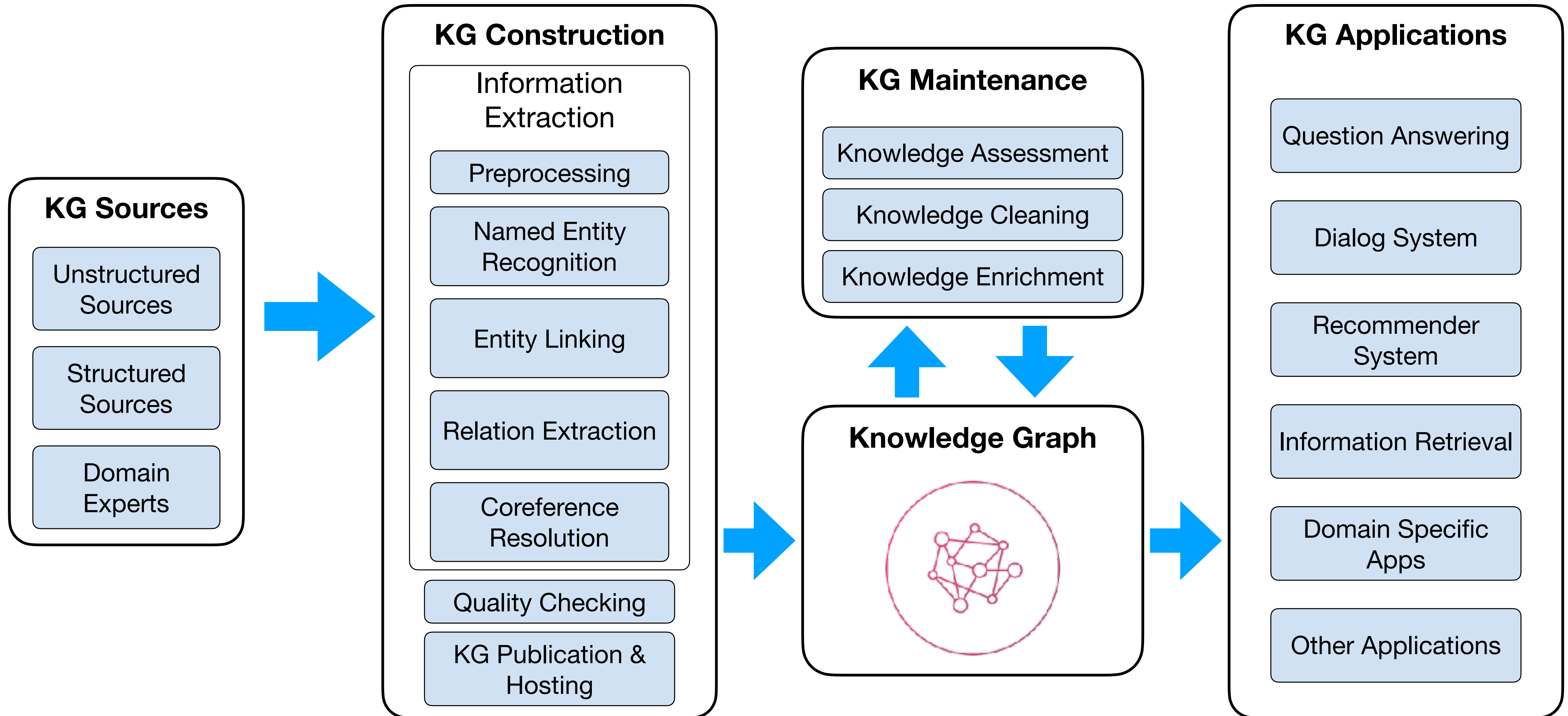
W3C is working with different industries —



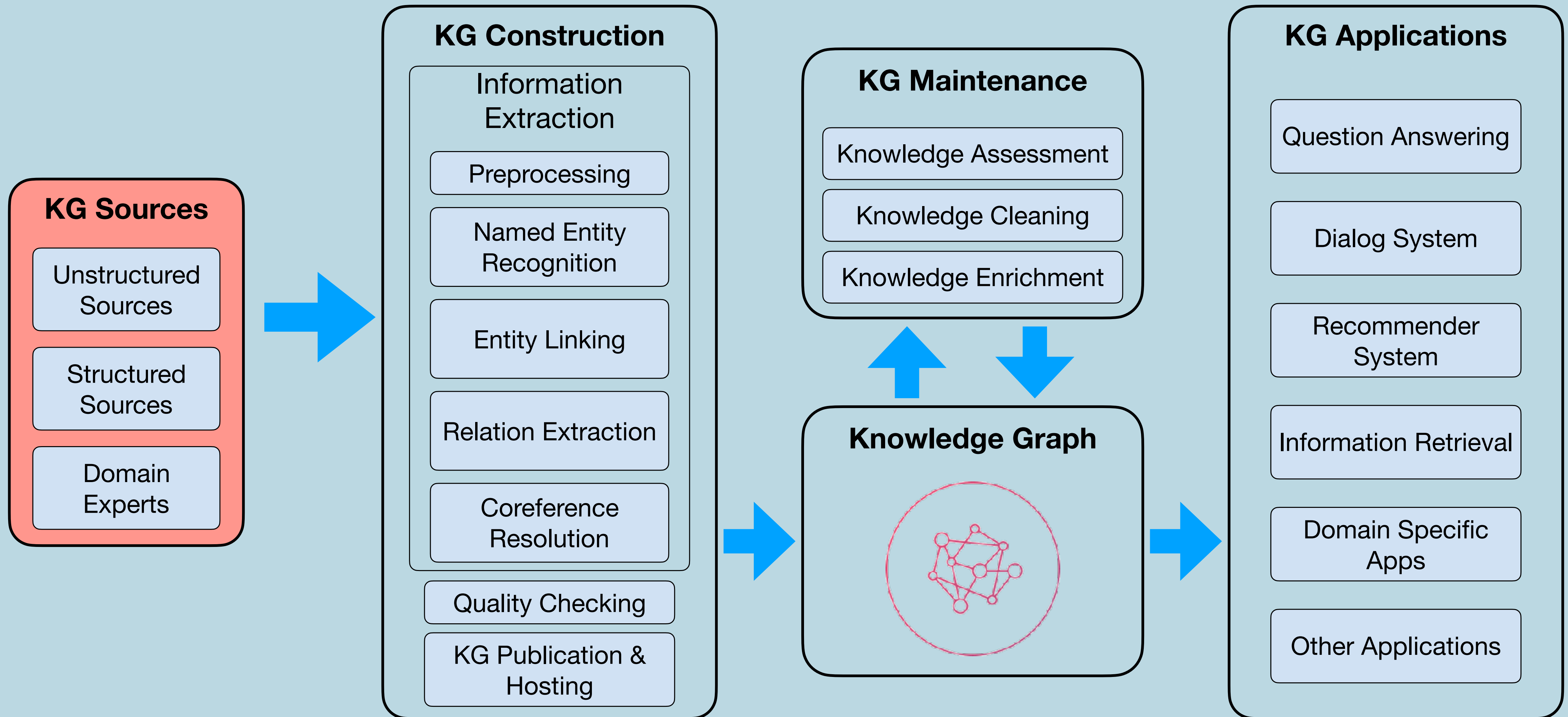


# KG Construction

# 53 KG Workflow: How to build a KG from Scratch



# 54 KG Workflow: How to build a KG from Scratch





## 55 KG Construction

- © We know that a knowledge graph is when an **ontology** is **applied to** a set of individual **data**.

**All of which are based on RDF triples.**

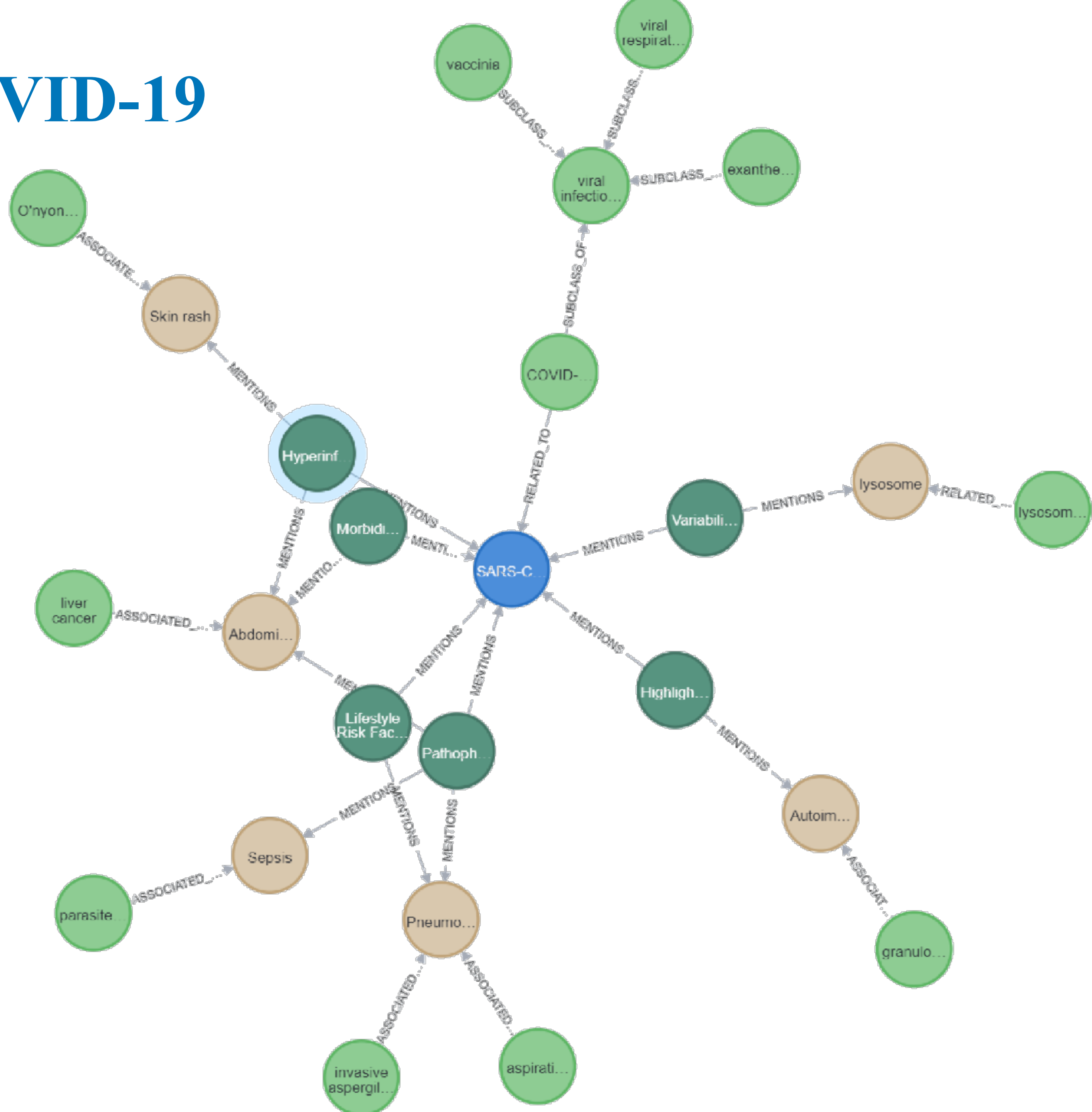


The construction of KG will involve how do we get triples from data of different sources.

# Example: KG-COVID-19

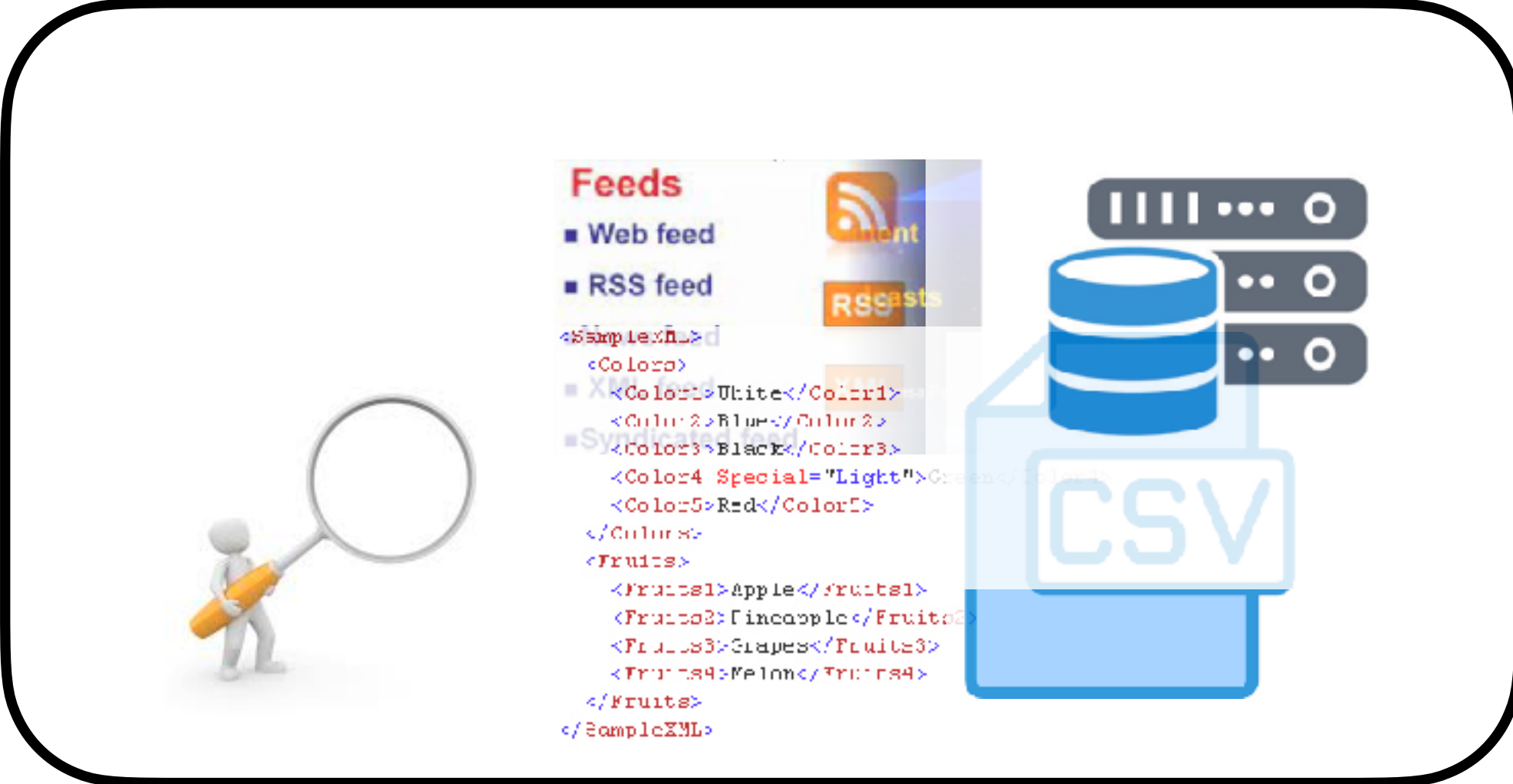
The right figure shows the sub-KG that covers the 3-hop neighbors of SARS-COV-2 with disease type nodes. We sampled 10 paths from the complete KG-Covid-19.

**Data source:** <https://github.com/Knowledge-Graph-Hub/kg-covid-19>



# Knowledge Sources

- Structured Sources
  - Relational Databases
  - Feeds
  - XML
  - CSV
  - ...



- Unstructured Sources
  - Web Crawls
  - Scientific Articles
  - Social Media
  - Emails
  - ...





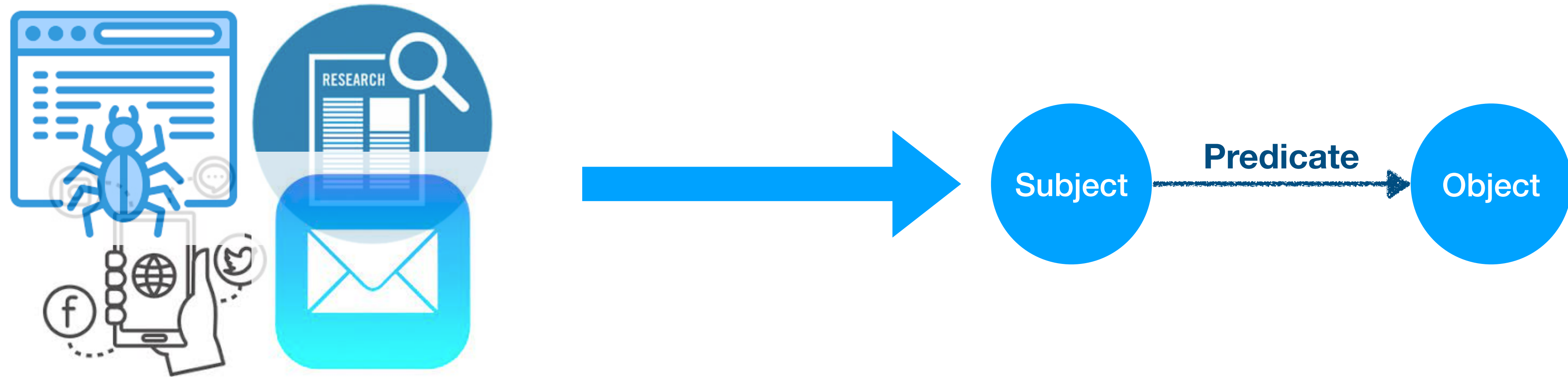
## 58 Structured Knowledge Sources to KGs

- Structured information can be mapped to KGs through a **semantic integration** process.
- Common strategy:
  - First adopting **reference ontologies** as global schemas.
  - Then construct **mappings** between the global schema and the local schema of the target data source.



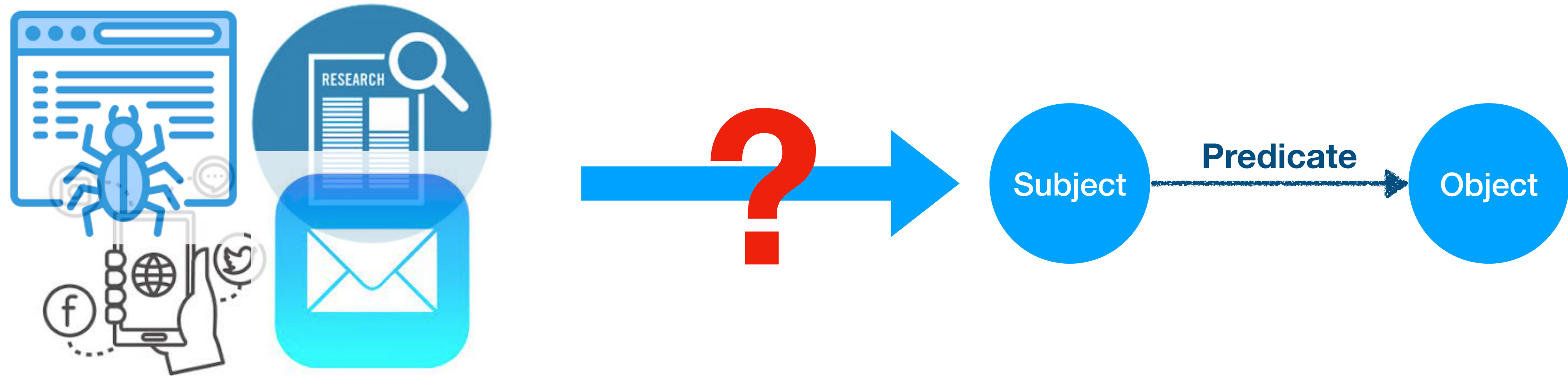
# 59 Unstructured Data to Structured Knowledge

- **How to excavate structured knowledge from unstructured data automatically?**



# 60 Unstructured Data to Structured Knowledge

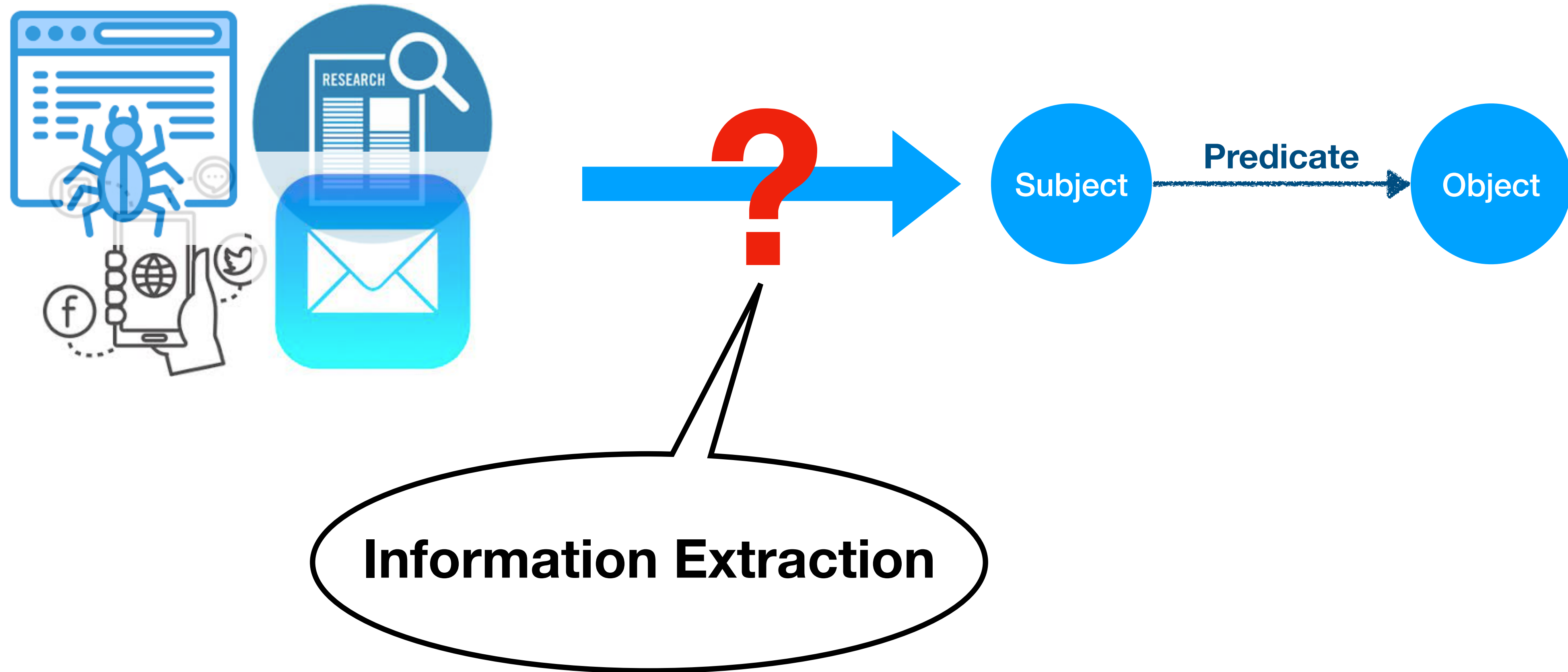
- **How to excavate structured knowledge from unstructured data automatically?**



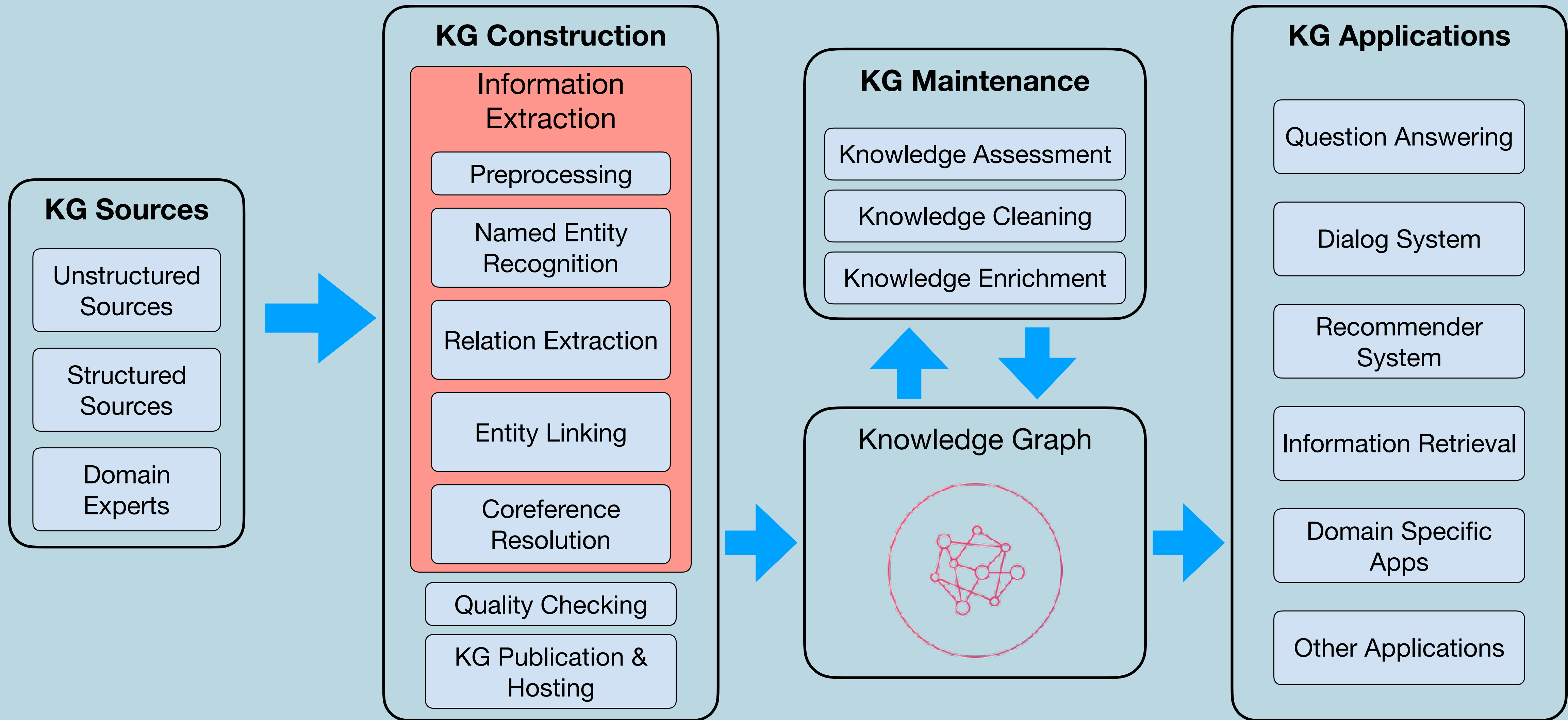


# 61 Unstructured Data to Structured Knowledge

- How to excavate structured knowledge from unstructured data automatically?

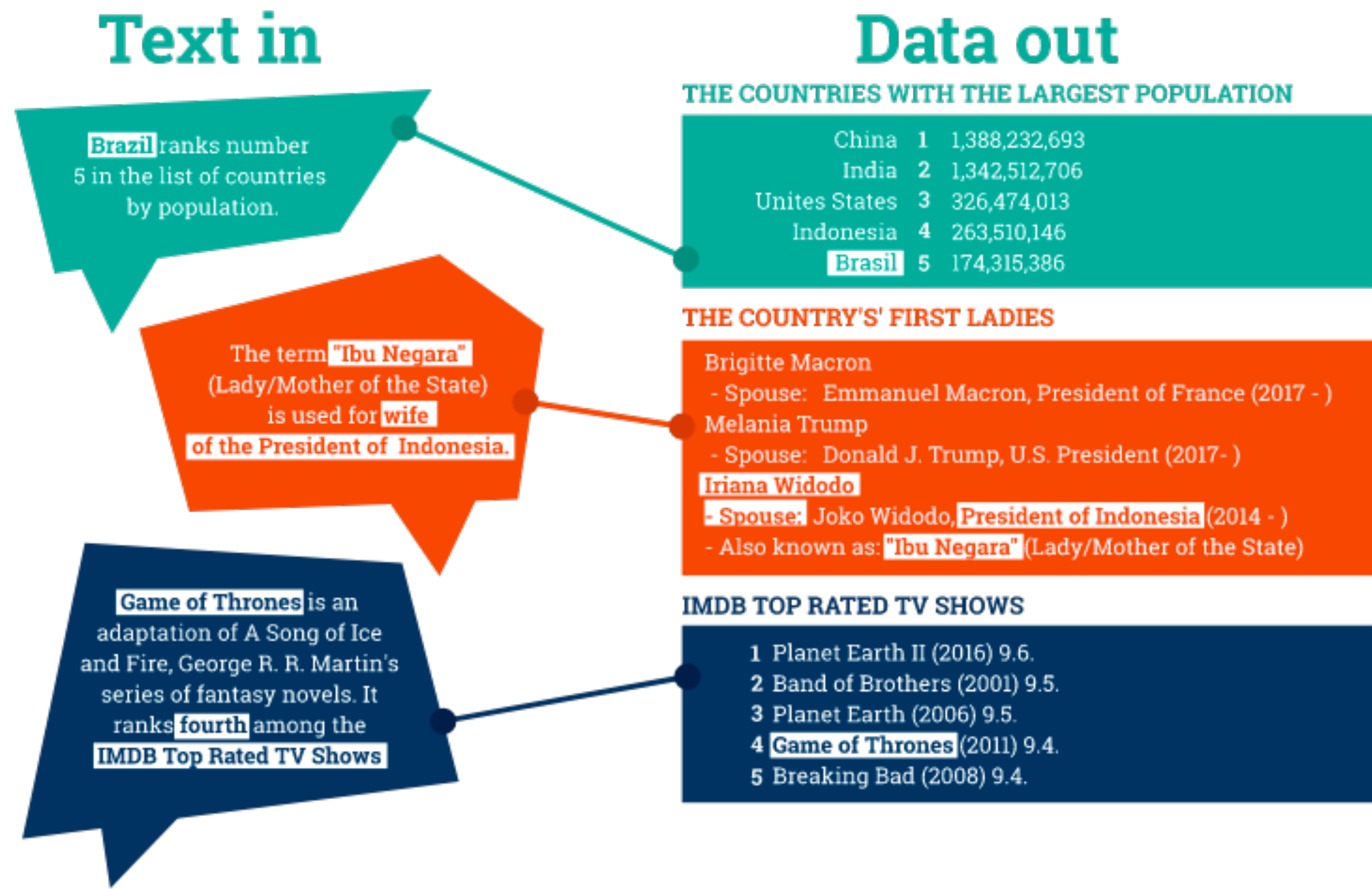


# Information Extraction



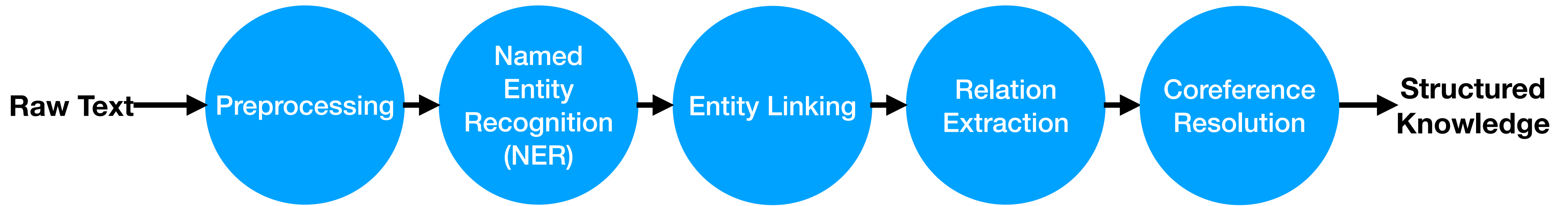
# What is information extraction?

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources. (From Wikipedia)





# 64 How does IE work?



Raw Text

Robert was stuck at the airport because of the snow storm. He missed the wedding of his daughter.

Robert PER was stuck at the airport FAC because of the snow storm. He PER missed the wedding of his PER daughter PER

Robert PER was stuck at the airport FAC because of the snow storm. He PER missed the wedding of his PER daughter PER  Robert\_De\_Niro

Robert SUBJ was stuck at the airport OBJ because of the snow storm. He missed the wedding of his SUBJ daughter OBJ .

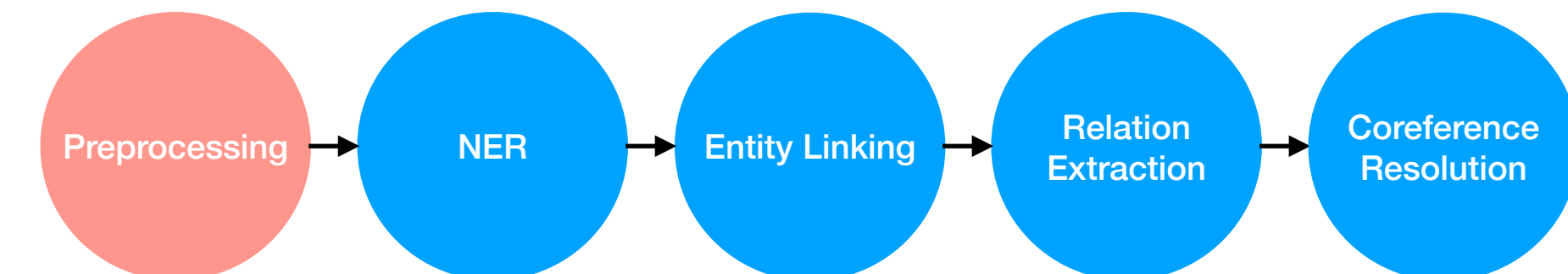
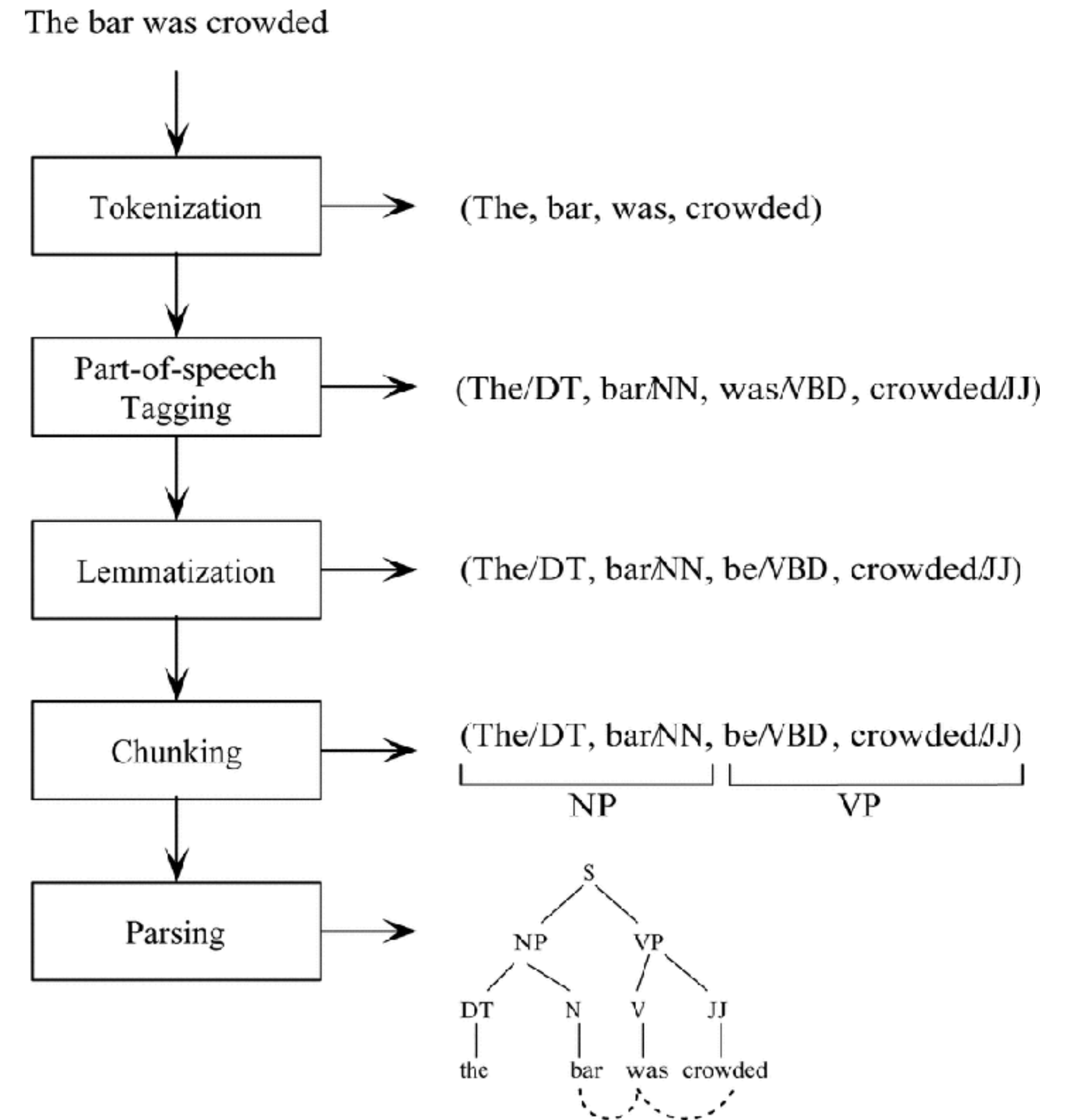
Robert was stuck at the airport because of the snow storm. He missed the wedding of his daughter.

Structured Knowledge

(Robert, PHYS, the airport), (Robert, PER-SOC, daughter)

# 65 Preprocessing

- The pre-processing usually involves varied procedures (depends on followed algorithms):
  - Language Detection
  - Sentence Segmentation
  - Tokenization
  - Part-of-Speech (POS) Tagging
  - Dependency Parsing
  - ...

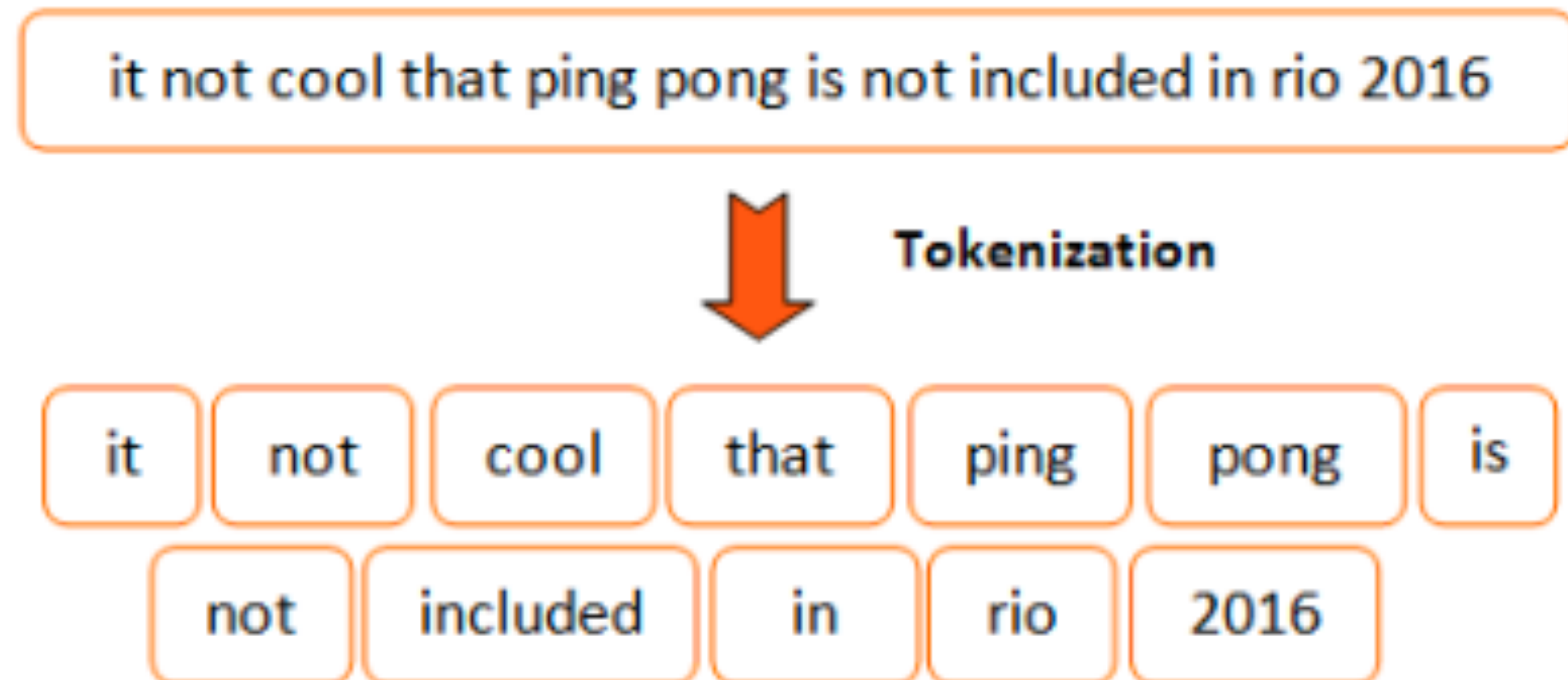


## 66 Tokenization

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters.

Tools:

- **NLTK Word Tokenize**
- **Stanford CoreNLP**
- **Gensim word tokenizer**
- **spaCy Tokenizer**
- TextBlob Word Tokenize
- Tokenization with Keras
- BERT Tokenizer
- ...





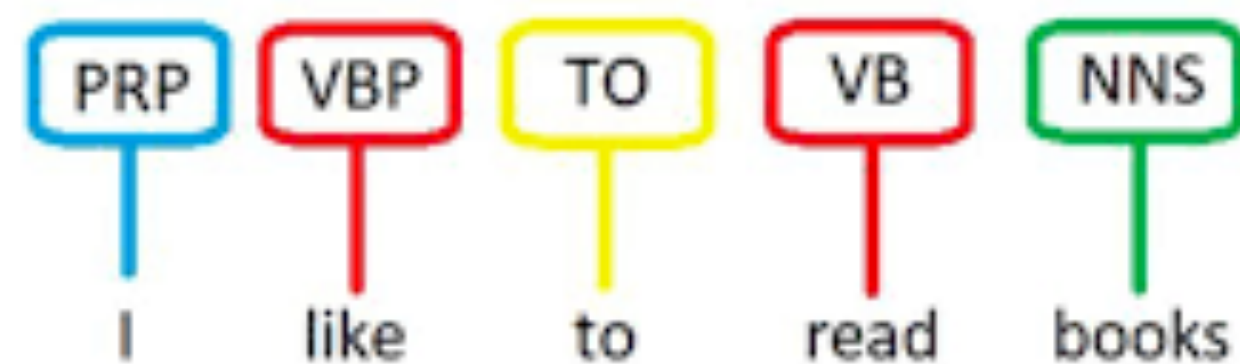
# 67 POS Tagging

● POS Tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

● **Tools:**

- spaCy Tagger
- NLTK pos\_tag
- ...

## POS Tagging



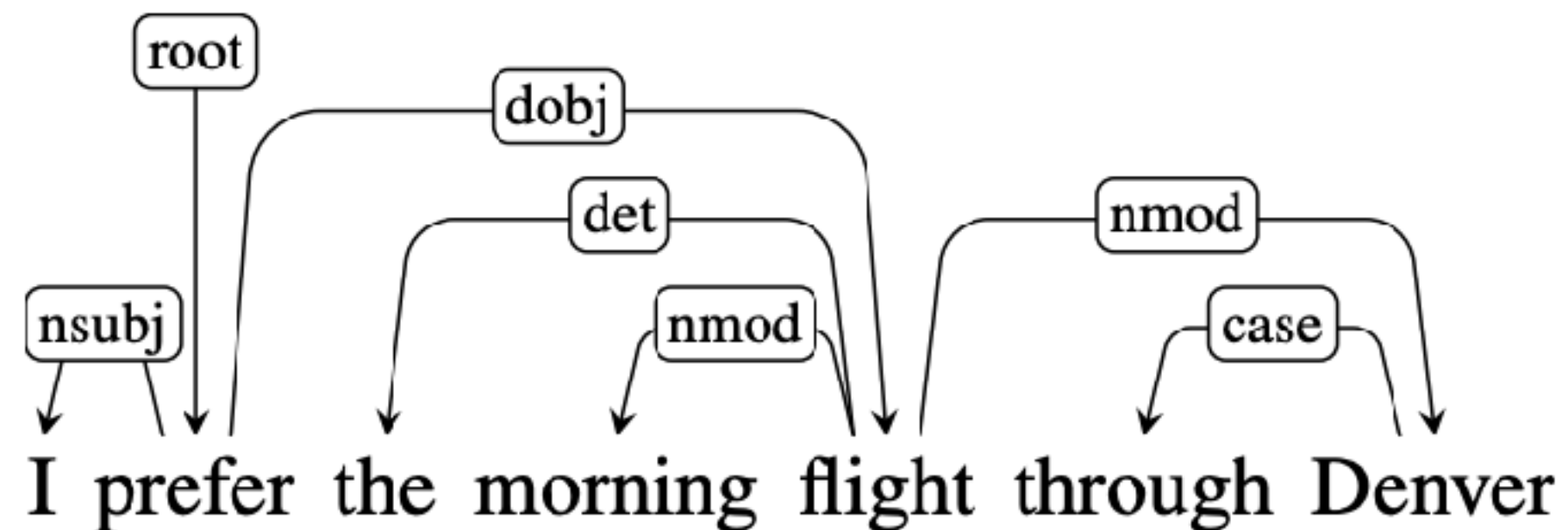
Tag	Meaning	English Examples
ADJ	adjective	<i>new, good, high, special, big, local</i>
ADP	adposition	<i>on, of, at, with, by, into, under</i>
ADV	adverb	<i>really, already, still, early, now</i>
CONJ	conjunction	<i>and, or, but, if, while, although</i>
DET	determiner, article	<i>the, a, some, most, every, no, which</i>
NOUN	noun	<i>year, home, costs, time, Africa</i>
NUM	numeral	<i>twenty-four, fourth, 1991, 14:24</i>
PRT	particle	<i>at, on, out, over per, that, up, with</i>
PRON	pronoun	<i>he, their, her, its, my, I, us</i>
VERB	verb	<i>is, say, told, given, playing, would</i>
.	punctuation marks	<i>. , ; !</i>
X	other	<i>ersatz, esprit, dunno, gr8, univeristy</i>

# 68 Dependency Parsing

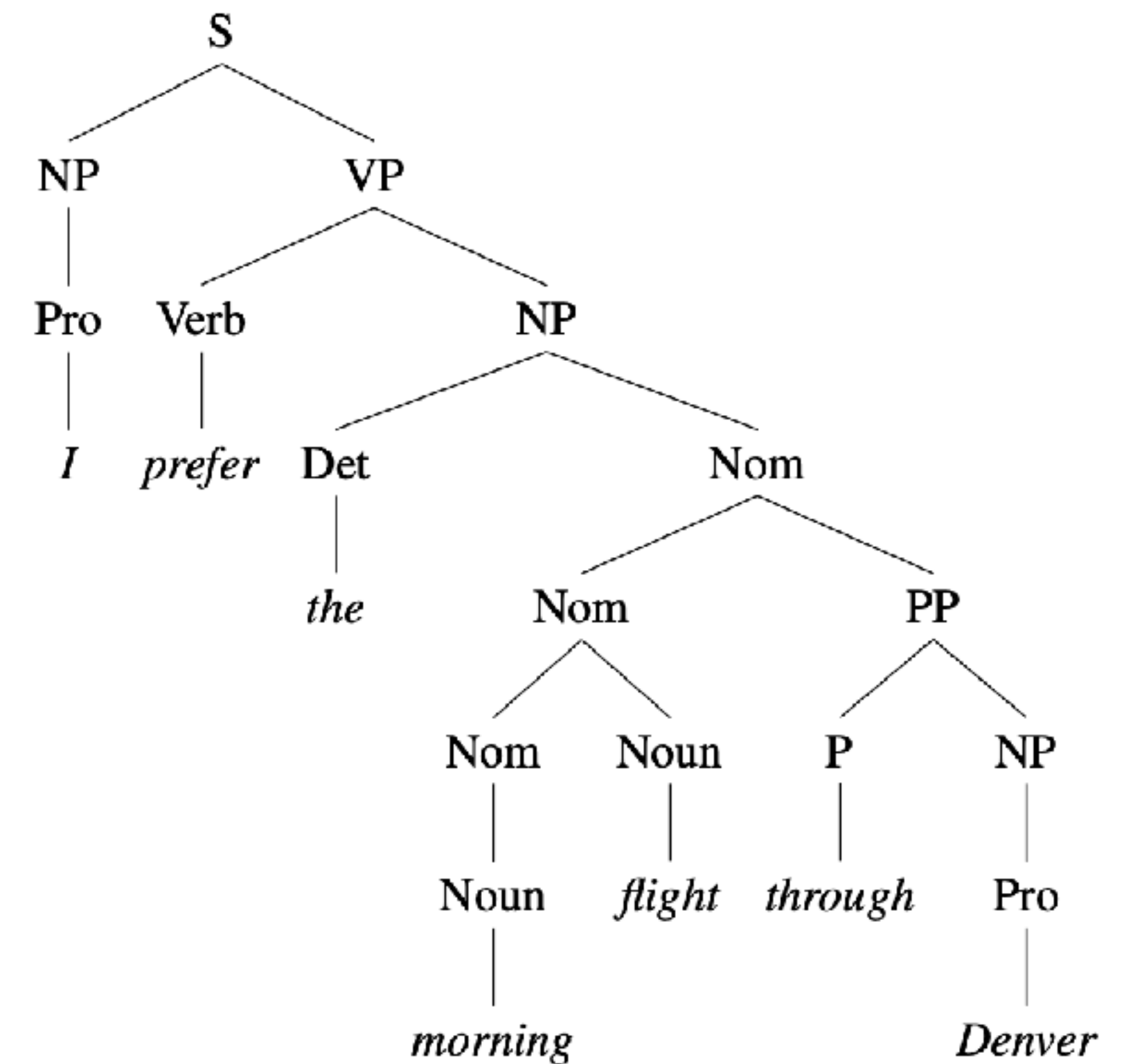
Dependency Parsing is the process to analyze the **grammatical structure** in a sentence and find out related words as well as the type of the relationship between them.

## Tools:

- spaCy Parser
- NLTK StanfordDependencyParser
- Stanza
- ...



Dependency Parsing



Constituency Parsing



- **NLTK** is a leading platform for building Python programs to work with human language data. (<https://www.nltk.org/>)

NLTK

Documentation

Search

Natural Language Toolkit

NLTK Documentation

**API Reference****Example Usage****Module Index****Wiki****FAQ****Open Issues****NLTK on GitHub**

Installation

**Installing NLTK****Installing NLTK Data**

More

**Release Notes****Contributing to NLTK****NLTK Team**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to **over 50 corpora and lexical resources** such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active **discussion forum**.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

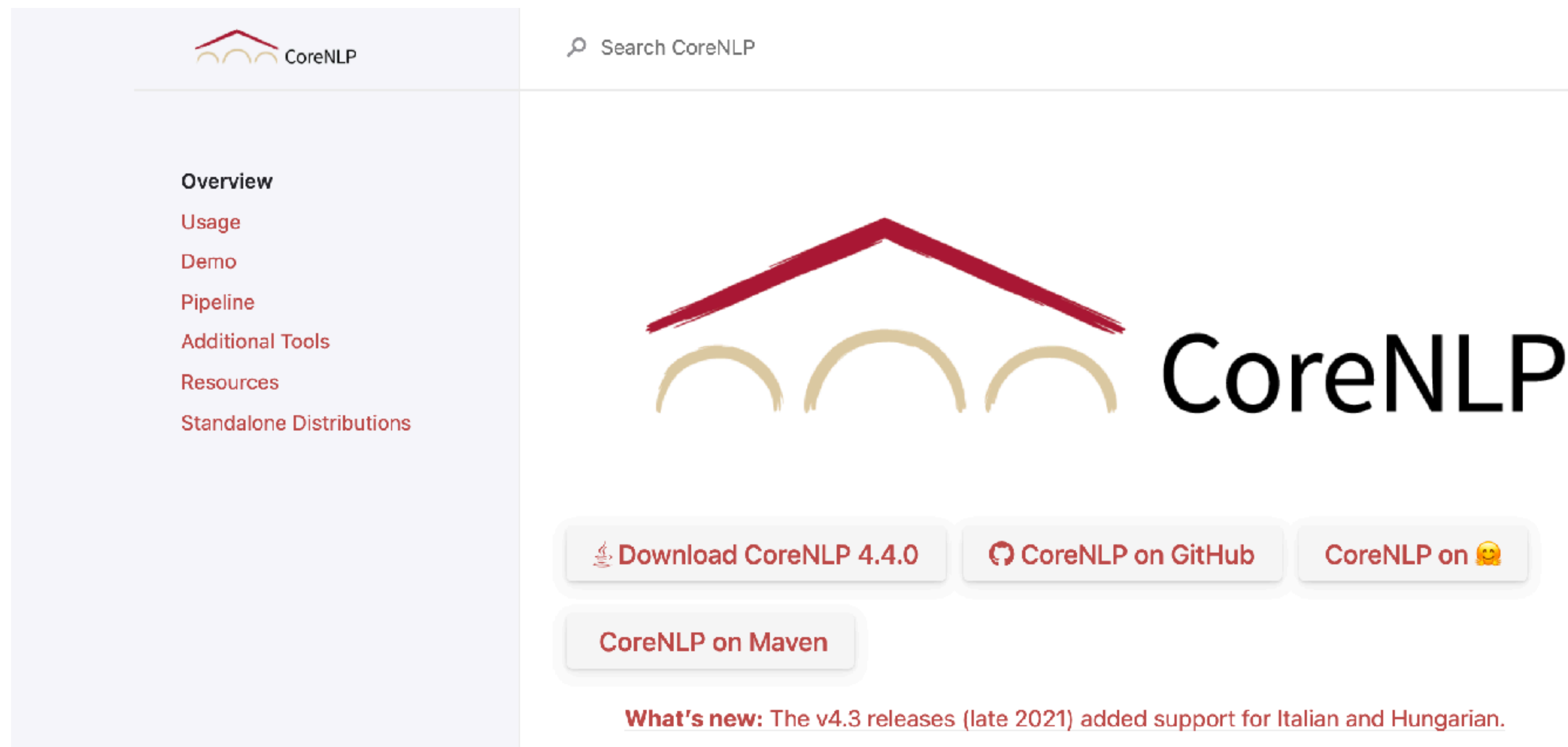
**Natural Language Processing with Python** provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at [https://www.nltk.org/book\\_1ed.](https://www.nltk.org/book_1ed.))

Some simple things you can do with NLTK



## 70 Stanford CoreNLP

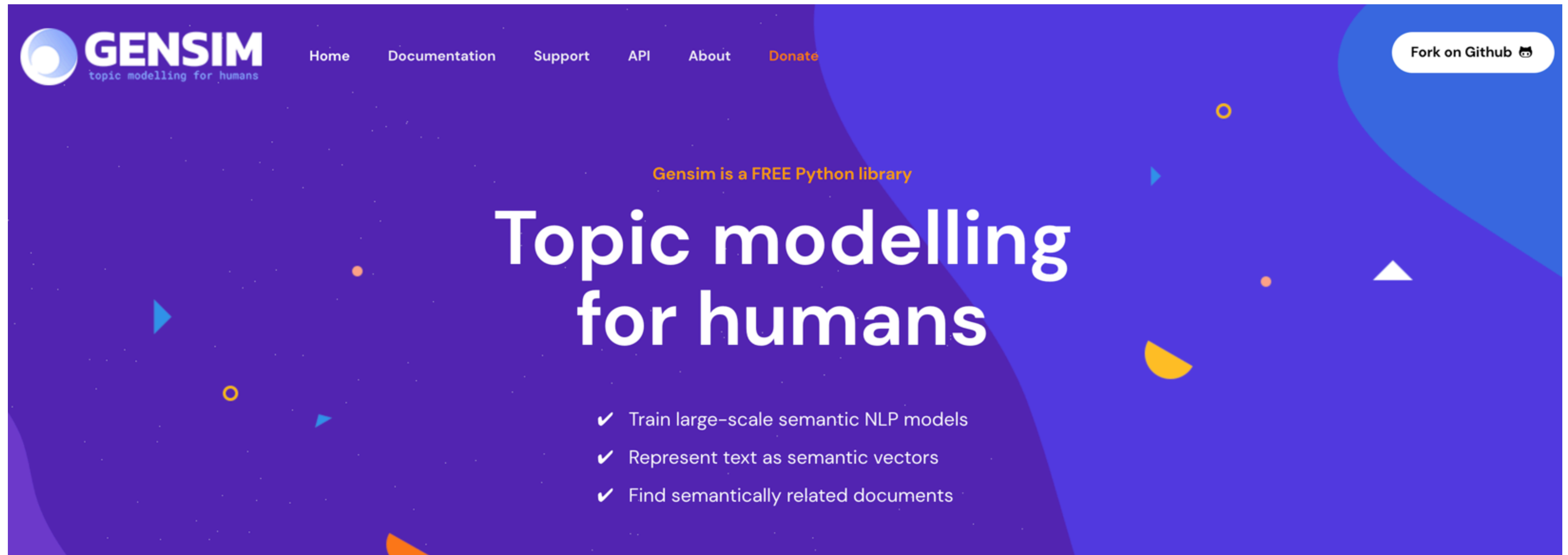
- **CoreNLP** is your one stop shop for natural language processing in Java and currently supports 8 languages: Arabic, Chinese, English, French, German, Hungarian, Italian, and Spanish. (<https://stanfordnlp.github.io/CoreNLP/>)



The screenshot shows the Stanford CoreNLP website homepage. On the left is a navigation menu with links for Overview, Usage, Demo, Pipeline, Additional Tools, Resources, and Standalone Distributions. The main content area features the CoreNLP logo (a red roof over three yellow arches) and the text "CoreNLP". Below the logo are four buttons: "Download CoreNLP 4.4.0", "CoreNLP on GitHub", "CoreNLP on [emoji]", and "CoreNLP on Maven". At the bottom, a red text block states: "What's new: The v4.3 releases (late 2021) added support for Italian and Hungarian."


# 71 GENSIM

- **Gensim** is a Python library for *topic modelling*, *document indexing* and *similarity retrieval* with large corpora. (<https://radimrehurek.com/gensim/>)



**GENSIM**  
topic modelling for humans

Home Documentation Support API About **Donate**

Fork on Github 

Gensim is a FREE Python library

# Topic modelling for humans

- ✓ Train large-scale semantic NLP models
- ✓ Represent text as semantic vectors
- ✓ Find semantically related documents

- spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. (<https://spacy.io/>)

# Industrial-Strength Natural Language Processing

IN PYTHON

**Get things done**

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

**GET STARTED**

**Blazing fast**

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

**FACTS & FIGURES**

**Awesome ecosystem**

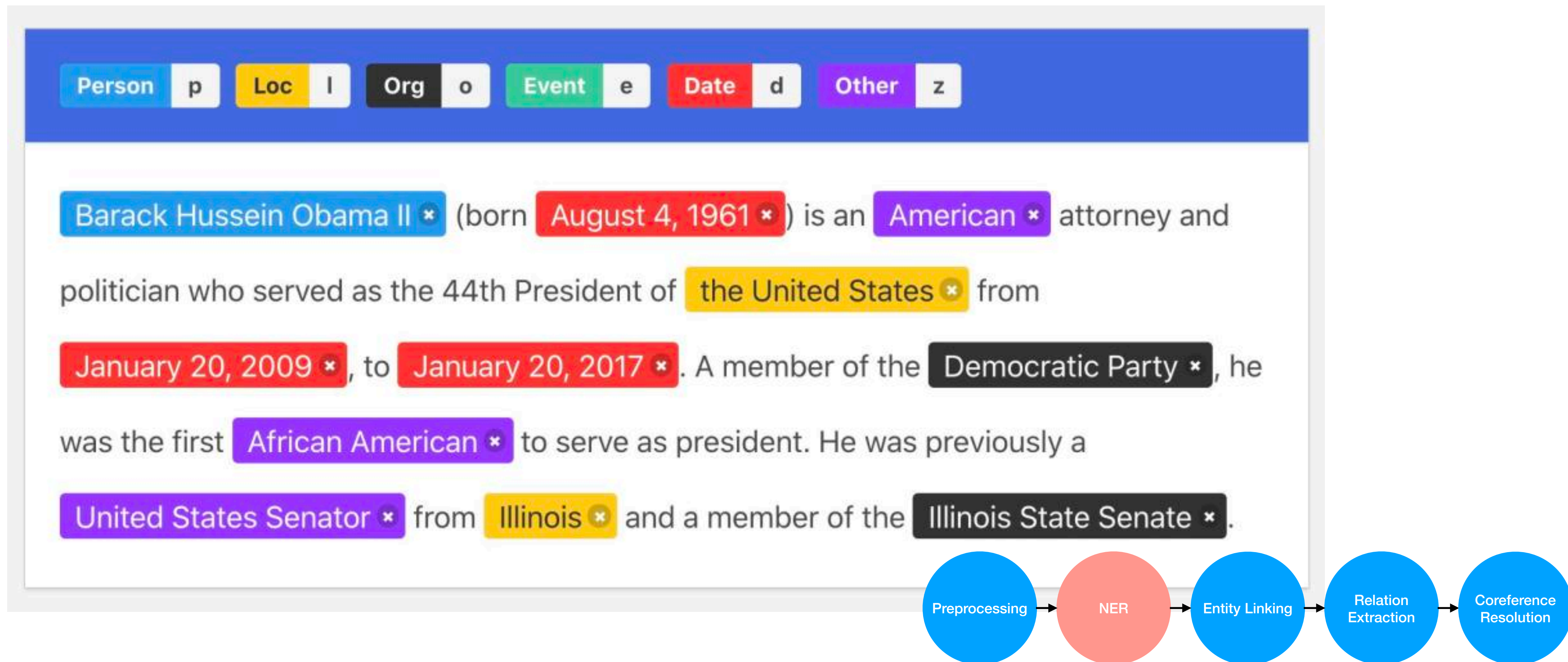
In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

**READ MORE**



# 73 Named Entity Recognition (NER)

- NER seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.



## 74 Named Entity Recognition (NER)

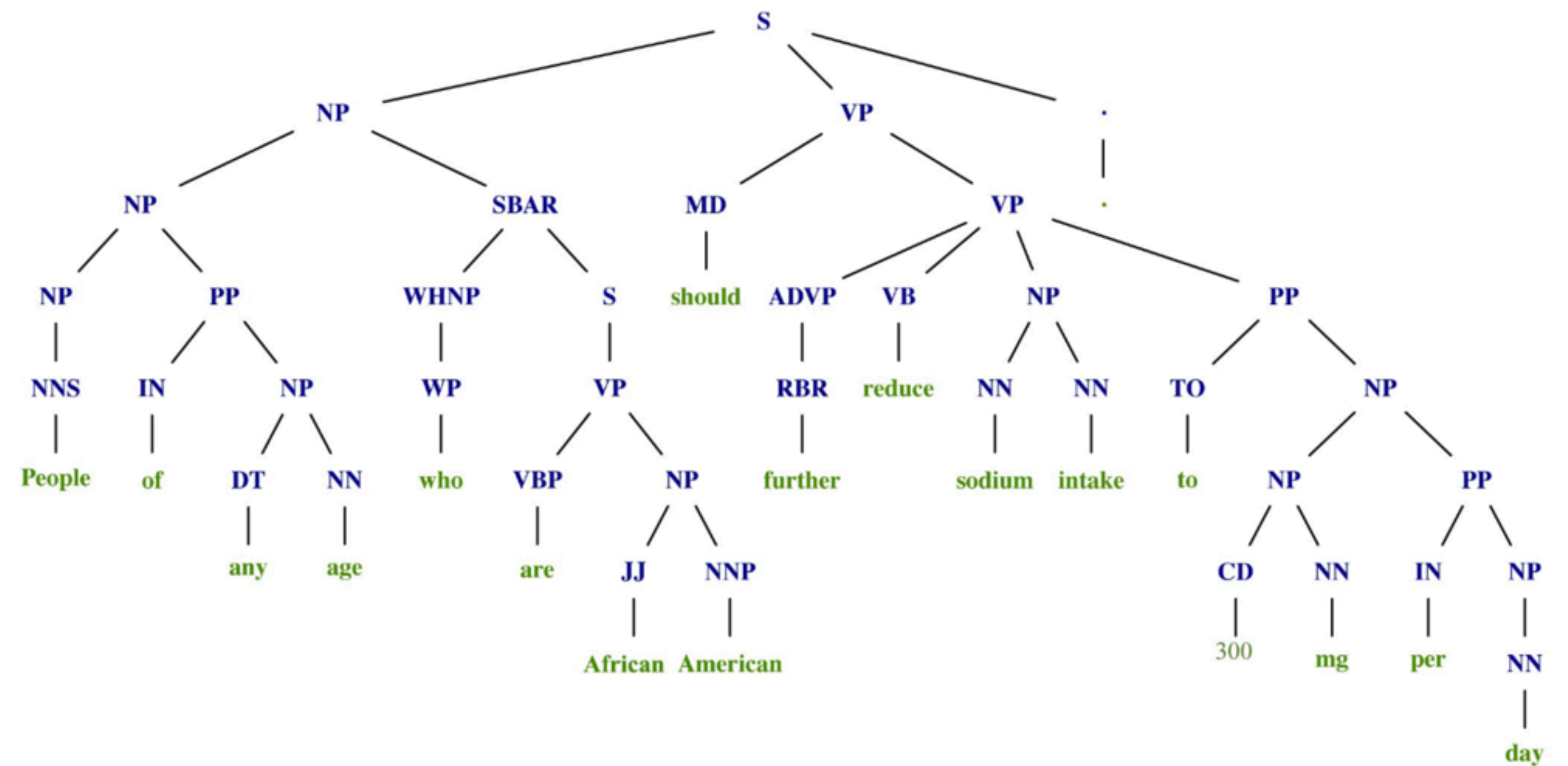
### ◎ NER Approaches:

- **Rule-based Approaches** rely on hand-crafted rules. The Rules can be designed based on domain-specific gazetteers and syntactic-lexical patterns.
- **Unsupervised Learning Approaches** typically leverage clustering based on lexical resources, patterns, and statistics on a large corpus to infer mentions of named entities.
- **Feature-based Supervised Learning Approaches** cast NER to a multi-class classification or sequence labeling task, with annotated data samples and carefully designed features.
- **Deep Learning-based Approaches** discovering hidden features automatically.

# 75 Named Entity Recognition (NER)

## ● Rule-based Approaches

- **Pros:**  
The domain-specific rules usually brings high precision.
- **Cons:**  
Low recall resulted by incomplete dictionaries.  
Construction of rules can be expensive.  
Cannot be transferred to other domains.
- **Rule-based systems:**  
ProMiner, LaSIE-II, NetOwl, Facile, SAR, LTG, ...

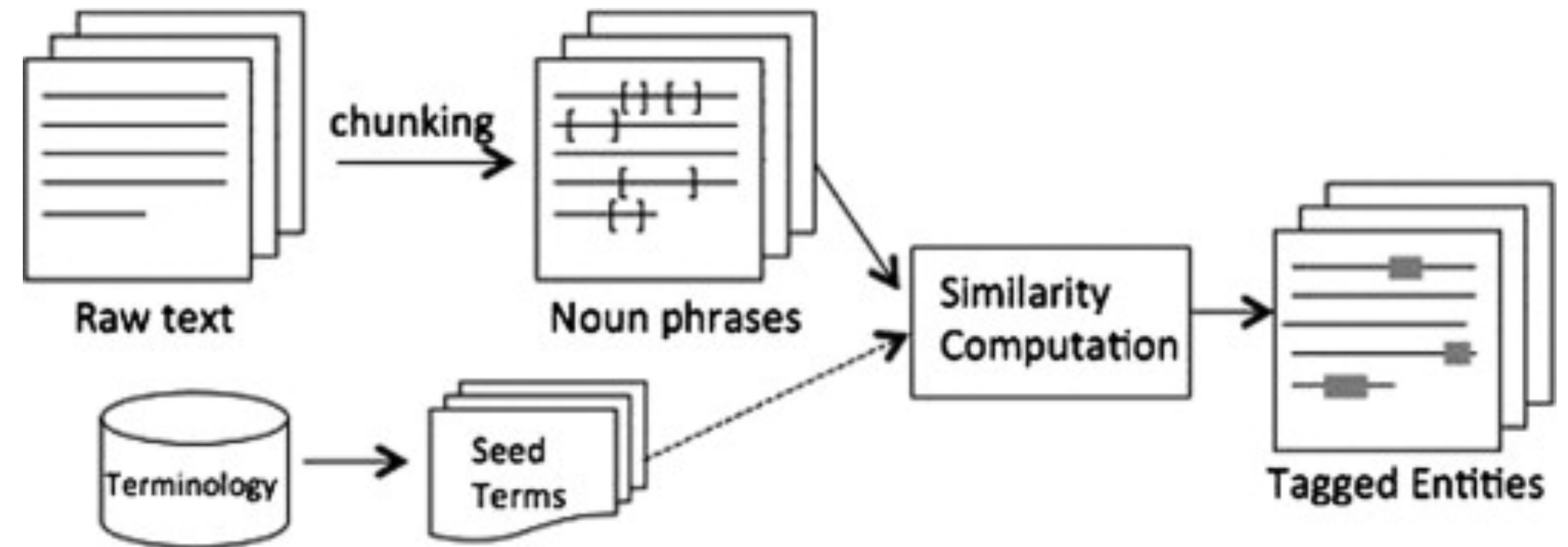




# 76 Named Entity Recognition (NER)

## ● Unsupervised Learning Approaches:

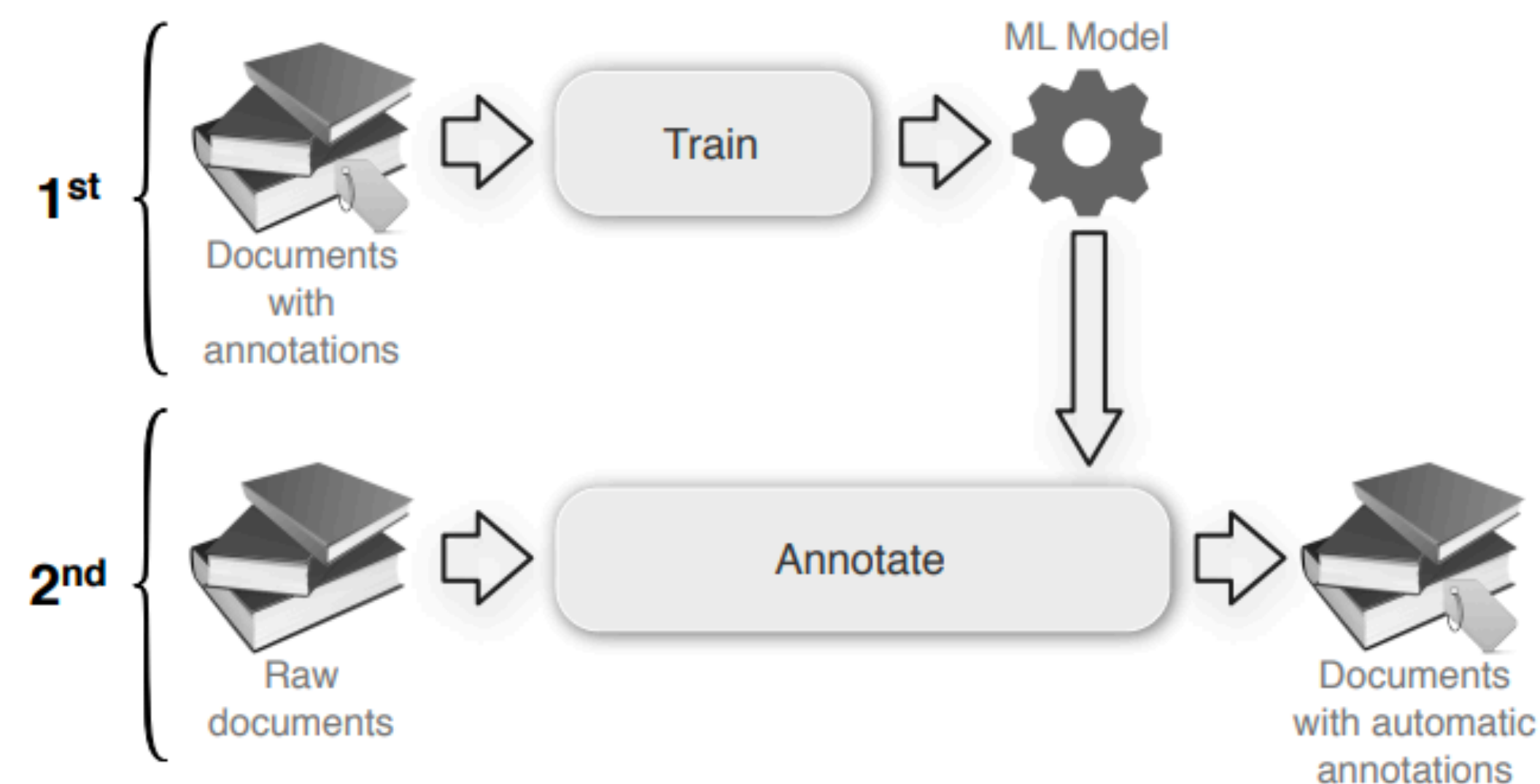
- **Pros:**  
Low required resources.
- **Cons:**  
Low performance.
- **Unsupervised NER systems:**  
KNOWITALL, Collins et al., Nadeau et al.,  
Zhang and Elhadad, ...



# 77 Named Entity Recognition (NER)

## ● Feature-based Supervised Learning Approaches:

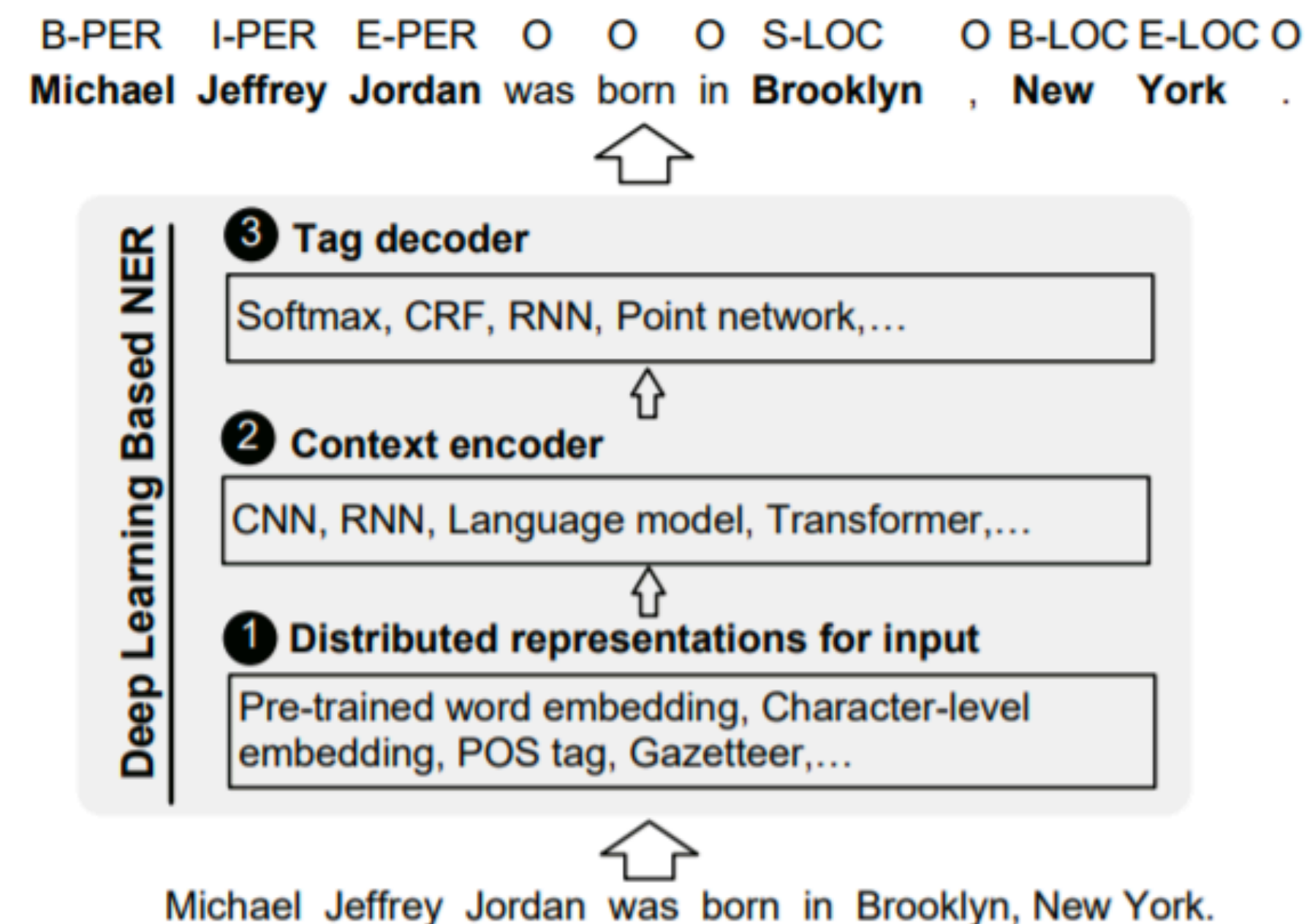
- **Pros:**  
High in-domain performance.
- **Cons:**  
The designing of features may involve experts.  
The labeling of data may be expensive.
- **Systems:**  
IdentiFinder, MENE, McNamee and Mayfield,  
...



# 78 Named Entity Recognition (NER)

## Deep Learning-based Approaches

- **Pros:**
  - End-to-end
  - High performance
  - Automatic feature discovery
- **Cons:**
  - Generally requires abundant labeled data
  - Requires higher computational resources
- **Systems:**
  - CNN-based, LSTM-based, ELMo-based, BERT-based, ...





Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	<a href="https://catalog ldc.upenn.edu/LDC2003T13">https://catalog ldc.upenn.edu/LDC2003T13</a>
MUC-6 Plus	1995	Additional news to MUC-6	7	<a href="https://catalog ldc.upenn.edu/LDC96T10">https://catalog ldc.upenn.edu/LDC96T10</a>
MUC-7	1997	New York Times news	7	<a href="https://catalog ldc.upenn.edu/LDC2001T02">https://catalog ldc.upenn.edu/LDC2001T02</a>
CoNLL03	2003	Reuters news	4	<a href="https://www.clips.uantwerpen.be/conll2003/ner/">https://www.clips.uantwerpen.be/conll2003/ner/</a>
ACE	2000 - 2008	Transcripts, news	7	<a href="https://www ldc.upenn.edu/collaborations/past-projects/ace">https://www ldc.upenn.edu/collaborations/past-projects/ace</a>
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	<a href="https://catalog ldc.upenn.edu/LDC2013T19">https://catalog ldc.upenn.edu/LDC2013T19</a>
W-NUT	2015 - 2018	User-generated text	6/10	<a href="http://noisy-text.github.io">http://noisy-text.github.io</a>
BBN	2005	Wall Street Journal	64	<a href="https://catalog ldc.upenn.edu/LDC2005T33">https://catalog ldc.upenn.edu/LDC2005T33</a>
WikiGold	2009	Wikipedia	4	<a href="https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500">https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500</a>
WiNER	2012	Wikipedia	4	<a href="http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner">http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner</a>
WikiFiger	2012	Wikipedia	112	<a href="https://github.com/xiaoling/figer">https://github.com/xiaoling/figer</a>
HYENA	2012	Wikipedia	505	<a href="https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena">https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena</a>
N <sup>3</sup>	2014	News	3	<a href="http://aksw.org/Projects/N3NER/NEDNIF.html">http://aksw.org/Projects/N3NER/NEDNIF.html</a>
Gillick	2016	Magazine, news, web, etc.	89	<a href="https://arxiv.org/e-print/1412.1820v2">https://arxiv.org/e-print/1412.1820v2</a>
FG-NER	2018	Various	200	<a href="https://fgner.alt.ai/">https://fgner.alt.ai/</a>
NNE	2019	Newswire	114	<a href="https://github.com/nickyringland/nested_named_entities">https://github.com/nickyringland/nested_named_entities</a>
GENIA	2004	Biology and clinical text	36	<a href="http://www.geniaproject.org/home">http://www.geniaproject.org/home</a>
GENETAG	2005	MEDLINE	2	<a href="https://sourceforge.net/projects/bioc/files/">https://sourceforge.net/projects/bioc/files/</a>
FSU-PRGE	2010	PubMed and MEDLINE	5	<a href="https://julielab.de/Resources/FSU_PRGE.html">https://julielab.de/Resources/FSU_PRGE.html</a>
NCBI-Disease	2014	PubMed	1	<a href="https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/">https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/</a>
BC5CDR	2015	PubMed	3	<a href="http://bioc.sourceforge.net/">http://bioc.sourceforge.net/</a>
DFKI	2018	Business news and social media	7	<a href="https://dfki-lt-re-group.bitbucket.io/product-corpus/">https://dfki-lt-re-group.bitbucket.io/product-corpus/</a>



# Recent works on neural NER

Work	Input representation			Context encoder	Tag decoder	Performance (F-score)
	Character	Word	Hybrid			
[94]	-	Trained on PubMed	POS	CNN	CRF	GENIA: 71.01%
[89]	-	Trained on Gigaword	-	GRU	GRU	ACE 2005: 80.00%
[95]	-	Random	-	LSTM	Pointer Network	ATIS: 96.86%
[90]	-	Trained on NYT	-	LSTM	LSTM	NYT: 49.50%
[91]	-	SENNA	Word shape	ID-CNN	CRF	CoNLL03: 90.65%; OntoNotes5.0: 86.84%
[96]	-	Google word2vec	-	LSTM	LSTM	CoNLL04: 75.0%
[100]	LSTM	-	-	LSTM	CRF	CoNLL03: 84.52%
[97]	CNN	GloVe	-	LSTM	CRF	CoNLL03: 91.21%
[105]	LSTM	Google word2vec	-	LSTM	CRF	CoNLL03: 84.09%
[19]	LSTM	SENNA	-	LSTM	CRF	CoNLL03: 90.94%
[106]	GRU	SENNA	-	GRU	CRF	CoNLL03: 90.94%
[98]	CNN	GloVe	POS	BRNN	Softmax	OntoNotes5.0: 87.21%
[107]	LSTM-LM	-	-	LSTM	CRF	CoNLL03: 93.09%; OntoNotes5.0: 89.71%
[103]	CNN-LSTM-LM	-	-	LSTM	CRF	CoNLL03: 92.22%
[17]	-	Random	POS	CNN	CRF	CoNLL03: 89.86%
[18]	-	SENNA	Spelling, n-gram, gazetteer	LSTM	CRF	CoNLL03: 90.10%
[20]	CNN	SENNA	capitalization, lexicons	LSTM	CRF	CoNLL03: 91.62%; OntoNotes5.0: 86.34%
[116]	-	-	FOFE	MLP	CRF	CoNLL03: 91.17%
[101]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.07%
[113]	LSTM	GloVe	Syntactic	LSTM	CRF	W-NUT17: 40.42%
[102]	CNN	SENNA	-	LSTM	Reranker	CoNLL03: 91.62%
[114]	CNN	Twitter Word2vec	POS	LSTM	CRF	W-NUT17: 41.86%
[115]	LSTM	GloVe	POS, topics	LSTM	CRF	W-NUT17: 41.81%
[118]	LSTM	GloVe	Images	LSTM	CRF	SnapCaptions: 52.4%
[109]	LSTM	SSKIP	Lexical	LSTM	CRF	CoNLL03: 91.73%; OntoNotes5.0: 87.95%
[119]	-	WordPiece	Segment, position	Transformer	Softmax	CoNLL03: 92.8%
[121]	LSTM	SENNA	-	LSTM	Softmax	CoNLL03: 91.48%
[124]	LSTM	Google Word2vec	-	LSTM	CRF	CoNLL03: 86.26%
[21]	GRU	SENNA	LM	GRU	CRF	CoNLL03: 91.93%
[126]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.71%
[142]	-	SENNA	POS, gazetteers	CNN	Semi-CRF	CoNLL03: 90.87%
[143]	LSTM	GloVe	-	LSTM	Semi-CRF	CoNLL03: 91.38%
[88]	CNN	Trained on Gigaword	-	LSTM	LSTM	CoNLL03: 90.69%; OntoNotes5.0: 86.15%
[110]	-	GloVe	ELMo, dependency	LSTM	CRF	CoNLL03: 92.4%; OntoNotes5.0: 89.88%
[108]	CNN	GloVe	ELMo, gazetteers	LSTM	Semi-CRF	CoNLL03: 92.75%; OntoNotes5.0: 89.94%
[133]	LSTM	GloVe	ELMo, POS	LSTM	Softmax	CoNLL03: 92.28%
[137]	-	-	BERT	-	Softmax	CoNLL03: 93.04%; OntoNotes5.0: 91.11%
[138]	-	-	BERT	-	Softmax +Dice Loss	CoNLL03: 93.33%; <b>OntoNotes5.0: 92.07%</b>
[134]	LSTM	GloVe	BERT, document-level embeddings	LSTM	CRF	CoNLL03: 93.37%; OntoNotes5.0: 90.3%
[135]	CNN	GloVe	BERT, global embeddings	GRU	GRU	CoNLL03: 93.47%
[132]	CNN	-	Cloze-style LM embeddings	LSTM	CRF	CoNLL03: <b>93.5%</b>
[136]	-	GloVe	Flooded contextual embeddings	RNN	CRF	CoNLL03: 93.47%

# Off-the-shelf NER tools

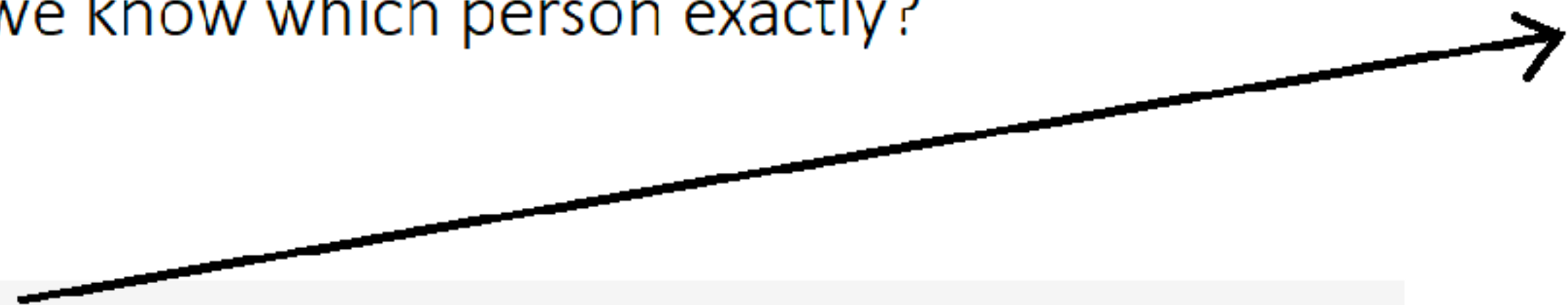
NER System	URL
StanfordCoreNLP	<a href="https://stanfordnlp.github.io/CoreNLP/">https://stanfordnlp.github.io/CoreNLP/</a>
OSU Twitter NLP	<a href="https://github.com/aritter/twitter_nlp">https://github.com/aritter/twitter_nlp</a>
Illinois NLP	<a href="http://cogcomp.org/page/software/">http://cogcomp.org/page/software/</a>
NeuroNER	<a href="http://neuroner.com/">http://neuroner.com/</a>
NERsuite	<a href="http://nersuite.nlplab.org/">http://nersuite.nlplab.org/</a>
Polyglot	<a href="https://polyglot.readthedocs.io">https://polyglot.readthedocs.io</a>
Gimli	<a href="http://bioinformatics.ua.pt/gimli">http://bioinformatics.ua.pt/gimli</a>
spaCy	<a href="https://spacy.io/api/entityrecognizer">https://spacy.io/api/entityrecognizer</a>
NLTK	<a href="https://www.nltk.org">https://www.nltk.org</a>
OpenNLP	<a href="https://opennlp.apache.org/">https://opennlp.apache.org/</a>
LingPipe	<a href="http://alias-i.com/lingpipe-3.9.3/">http://alias-i.com/lingpipe-3.9.3/</a>
AllenNLP	<a href="https://demo.allennlp.org/">https://demo.allennlp.org/</a>
IBM Watson	<a href="https://natural-language-understanding-demo.ng.bluemix.net">https://natural-language-understanding-demo.ng.bluemix.net</a>
FG-NER	<a href="https://fgner.alt.ai/extractor/">https://fgner.alt.ai/extractor/</a>
Intellexer	<a href="http://demo.intellexer.com/">http://demo.intellexer.com/</a>
Repustate	<a href="https://repustate.com/named-entity-recognition-api-demo">https://repustate.com/named-entity-recognition-api-demo</a>
AYLIEN	<a href="https://developer.aylien.com/text-api-demo">https://developer.aylien.com/text-api-demo</a>
Dandelion API	<a href="https://dandelion.eu/semantic-text/entity-extraction-demo">https://dandelion.eu/semantic-text/entity-extraction-demo</a>
displaCy	<a href="https://explosion.ai/demos/displacy-ent">https://explosion.ai/demos/displacy-ent</a>
ParallelDots	<a href="https://www.paralldots.com/named-entity-recognition">https://www.paralldots.com/named-entity-recognition</a>
TextRazor	<a href="https://www.textrazor.com/named_entity_recognition">https://www.textrazor.com/named_entity_recognition</a>



# Entity Linking (EL)

- Entity Linking (or Named-Entity Disambiguation) assigns a unique identity in target knowledge bases (KBs) to entity mentions in text.

We know 'Sebastian Thrun' is a person but do we know which person exactly?



When **Sebastian Thrun** **PERSON** started at **Google** **ORG** in **2007** **DATE**, few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** **NORP** car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** **PERSON**, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** **ORG** **earlier this week** **DATE**.

A little **less than a decade later** **DATE**, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

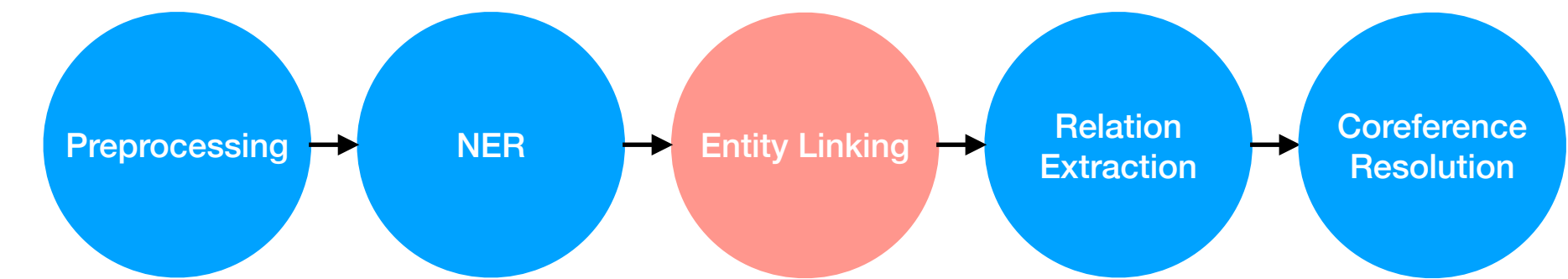
### About: Sebastian Thrun

An Entity of Type : scientist, from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](http://dbpedia.org)

Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech.

Property	Value
<a href="#">dbpedia:abstract</a>	<ul style="list-style-type: none"> <li>Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech. Thrun led development of the robotic vehicle Stanley which won the 2005 DARPA Grand Challenge, and which has since been placed on exhibit in the Smithsonian Institution's National Museum of American History. His team also developed a vehicle called Junior, which placed second at the DARPA Grand Challenge (2007). Thrun led the development of the Google self-driving car. Thrun is also known for his work on probabilistic algorithms for robotics with applications including robotic mapping. In recognition of his contributions, and at age 39, Thrun was elected into the National Academy of Engineering and also into the Academy of Sciences Leopoldina in 2007. In 2011, Thrun received the Max-Planck-Research Award, and the inaugural AAAI Ed Feigenbaum Prize. Fast Company selected Thrun as the fifth most creative person in the business world. The Guardian recognized Thrun as one of 20 "fighters for internet freedom". (en)</li> </ul>

[http://dbpedia.org/page/Sebastian\\_Thrun](http://dbpedia.org/page/Sebastian_Thrun)



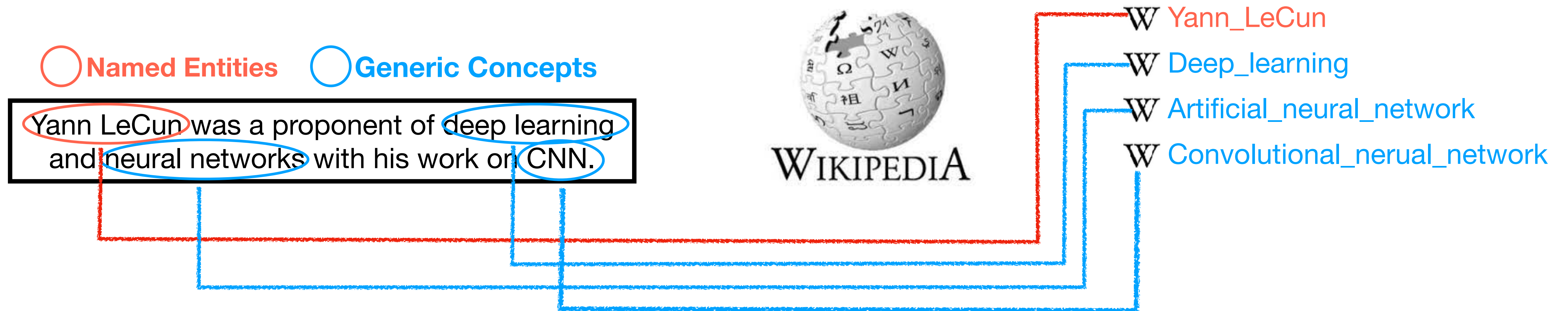
# Entity Linking (EL)

- A formal definition of EL consist of:
  - A definition of the mentions (concepts, entities)
  - Determining the target KB
  - Defining the unique identity in the target KBs

## 1. Define the mentions

## 2. Determine target KBs

## 3. Find the unique identity



# Entity Linking: Define the mentions

- A **mention**: a phrase used to refer to something in the world

**Named Entity**  **Object**  **Substance**  **Event**  **Philosophy**  **Mental State**  
 **Rule**  ...

- Task definitions of **EL** vary across the definition of **mentions**
  - All N-grams
  - Dictionary-based selection
  - Data-driven controlled vocabulary (e.g., all Wikipedia titles)
  - **Named entities** (by NER in our demo).
- Ideally, the definition of mentions should adapt to the application



# 85 Entity Linking: Define the mentions

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Some task definitions insist on dealing only with mentions that are **named entities**

How about: *Hosni Mubarak's wife?*

Both entities have a Wikipedia page

Blumenthal (D) is a candidate for the U.S. Senate seat now held by Christopher Dodd (D), and he has held a commanding lead in the race since he entered it. But the Times report has the potential to fundamentally reshape the contest in the Nutmeg State.

Richard Blumenthal  
From Wikipedia, the free encyclopedia

Christopher Dodd  
From Wikipedia, the free encyclopedia

Chris Dodd  
From Wikipedia, the free encyclopedia

The New York Times  
From Wikipedia, the free encyclopedia

Connecticut  
From Wikipedia, the free encyclopedia

Google search for "Hosni Mubarak's wife" results in a knowledge panel for Suzanne Mubarak. The panel includes her photo, title "Former First Lady of Egypt", and biographical information: born February 23, 1941 (age 80 years), Menia Governorate, Egypt; full name Suzanne Saleh Thabet; spouse Hosni Mubarak (m. 1959-2020); grandchildren Faride Gamal Mubarak, Omer Alea Mubarak, Mahmoud Gamal Mubarak, Mohammed Mubarak; parents Saleh Thabet, Lily May Palmer; children Gamal Mubarak, Alaa Mubarak; siblings Mounir Sabet. A "Claim this knowledge panel" button is visible at the bottom.

People also search for: Gamal Mubarak, Hosni Mubarak, Alaa Mubarak, Jehen Sadat, Khadija El Gamal, Mounir Sabet, Naglea Mah...

Wikipedia snippet: Suzanne Mubarak is the widow of Egyptian former president Hosni Mubarak and was the First Lady of Egypt during her husband's presidential tenure from 14 October 1981 to 11 February 2011.

# English Entity Linking Demo



## Demo

### English Entity Linking Demo

About This Demo

125 views

BLINK is one of the state-of-the-art Entity Detection and Linking (EDL) models, which is to disambiguate entity mentions and link them to the correct Wikipedia entries. This demo takes a paragraph as the input, calls the Spacy NER model to extract entities, and uses BLINK to link these entities to Wikipedia.

Here is the [original paper](#) introducing BLINK.

#### Annotators

EDL

#### Examples:

Sigmund Freud was an Austrian neurologist and the ...

#### Text:

Sigmund Freud was an Austrian neurologist and the founder of psychoanalysis, a clinical method for treating psychopathology through dialogue between a patient and a psychoanalyst. Freud was born to Galician Jewish parents in the Moravian town of Freiberg, in the Austrian Empire. He qualified as a doctor of medicine in 1881 at the University of Vienna. Freud lived and worked in Vienna, having set up his clinical practice there in 1886. In 1938, Freud left Austria to escape Nazi persecution. He died in exile in the United Kingdom in 1939.

Run >

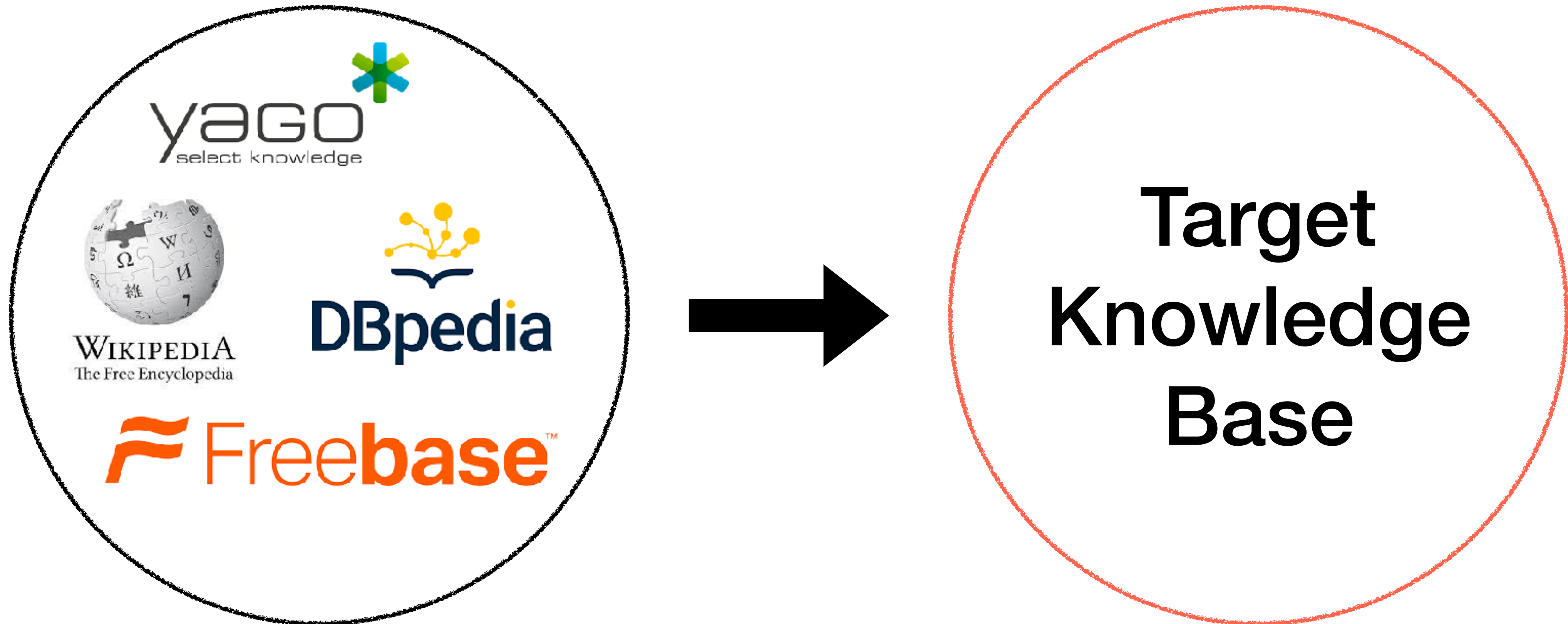
The definition of which **mentions** to highlight should depends on the **expertise** and **users' interests**.

**Sigmund Freud** was an **Austrian** neurologist and the founder of psychoanalysis , a clinical method for treating psychopathology through dialogue between a patient and a psychoanalyst . Freud was born to **Galician Jewish** parents in the **Moravian** town of **Freiberg** , in **the Austrian Empire** . He qualified as a doctor of medicine in **1881** at **the University of Vienna** . Freud lived and worked in **Vienna** , having set up his clinical practice there in **1886** . In **1938** , Freud left Austria to escape **Nazi** persecution . He died in exile in **the United Kingdom** in **1939** .

[https://cogcomp.seas.upenn.edu/page/demo\\_view/EDLEnglish](https://cogcomp.seas.upenn.edu/page/demo_view/EDLEnglish)

## 87 Entity Linking: Define the target KB

- In principle, **multiple KBs** can be used as the **target KB**.

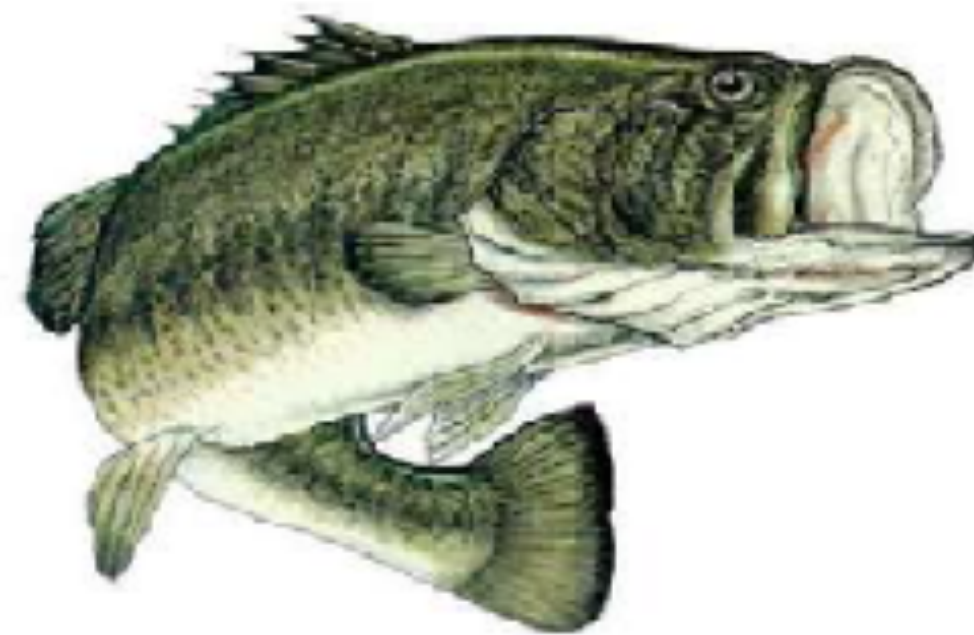




# Entity Linking: Find the Unique ID

Find the unique identity in the target KB (Disambiguation)

Which one to link?



This is a **bass**

This is also a **bass**

Google

[All](#)
[News](#)
[Images](#)
[Videos](#)
[Maps](#)
[More](#)
[Tools](#)

About 120,000 results (0.72 seconds)

[https://en.wikipedia.org/wiki/James\\_Craig\\_\(police\\_chief\)](https://en.wikipedia.org/wiki/James_Craig_(police_chief))

**James Craig (police chief) - Wikipedia**

James E. Craig (born 1956) is the former head of the Detroit Police Department. He served in this capacity from 2013 to 2021, after previously serving as ...

Political party: [Republican](#) (2011–present) Mayor: [Dave Bing](#); [Mike Duggan](#)

Other political affiliations: [Democratic](#) (before ... Education: [West Coast University \(BS\)](#); [Uni...](#)

[Policing career](#) · [2022 Michigan gubernatorial...](#)

[https://en.wikipedia.org/wiki/James\\_Craig,\\_1st\\_Viscount\\_Craigavon](https://en.wikipedia.org/wiki/James_Craig,_1st_Viscount_Craigavon)

**James Craig, 1st Viscount Craigavon - Wikipedia**

James Craig, 1st Viscount Craigavon PC PC (NI) DL (8 January 1871 – 24 November 1940), was a prominent Irish unionist politician, leader of the Ulster ...

Prime Minister: [David Lloyd George](#)

[https://en.wikipedia.org/wiki/James\\_Craig\\_\(actor\)](https://en.wikipedia.org/wiki/James_Craig_(actor))

**James Craig (actor) - Wikipedia**

James Craig (born James Henry Meador, February 4, 1912 – June 27, 1985) was an American actor. He is best known for appearances in films like Kitty Foyle ...

Children: 3 Born: [James Henry Meador](#); February 4, 19...

Spouse(s): [Sumie Jossi](#) (1969–1980); [Jil Jarm...](#) Died: [June 27, 1985](#) (aged 73); [Santa Ana](#), ...

[Biography](#) · [Paramount](#) · [MGM](#) · [Personal life](#)

# Entity Linking: Find the Unique ID

Find the unique identity in the target KB (Disambiguation)

What if multiple **mentions** correspond to the same concept is **outside KB**?

The image shows a Google search interface. The search bar contains the text "Marc-Alexandre Côté site:en.wikipedia.org", which is circled in red. A red arrow points from the text "Marc-Alexandre Côté" in the text below to the search bar. Below the search bar, the search results are displayed, including two Wikipedia entries: "2021 Montreal municipal election - Wikipedia" and "2021 Quebec municipal elections - Wikipedia".

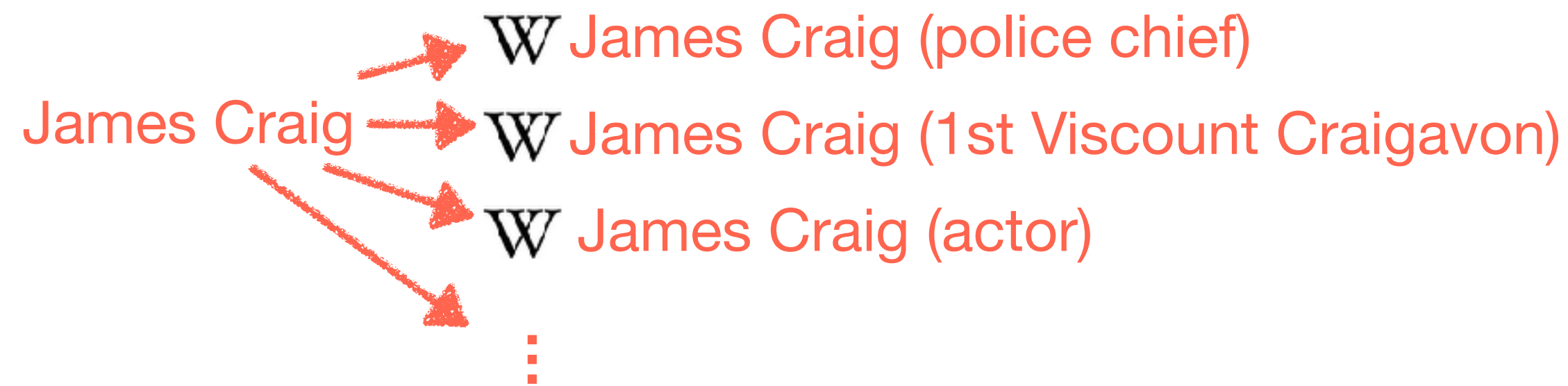
**Marc-Alexandre Côté** is an author of **Textworld**.

First **cluster relevant mentions** as representing a single concept  
Then map the **cluster** to **NIL**.

# Entity Linking (EL)

## What are the main issues with EL?

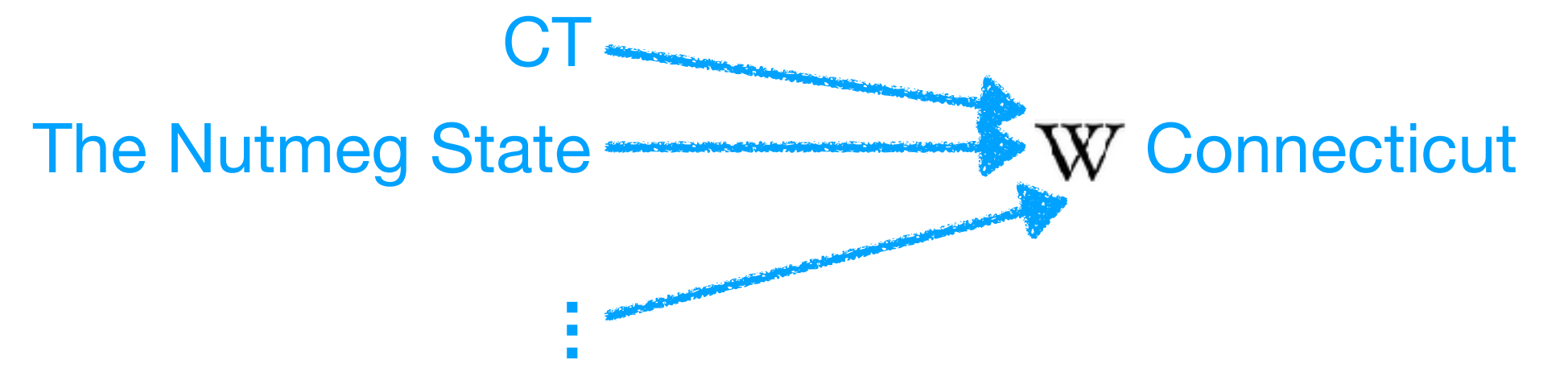
### Ambiguity



### Concepts outside of KB (NIL)

**Marc-Alexandre Côté** is an author of **Textworld**.

### Variability



### Scale

### Millions of labels



# Entity Linking (EL)

- Other challenges
  - Popularity Bias**

## Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

**Michael Jordan** (born 1963) is an American businessman and former professional basketball player.

Michael Jordan or Mike Jordan may also refer to:

### People [\[edit\]](#)

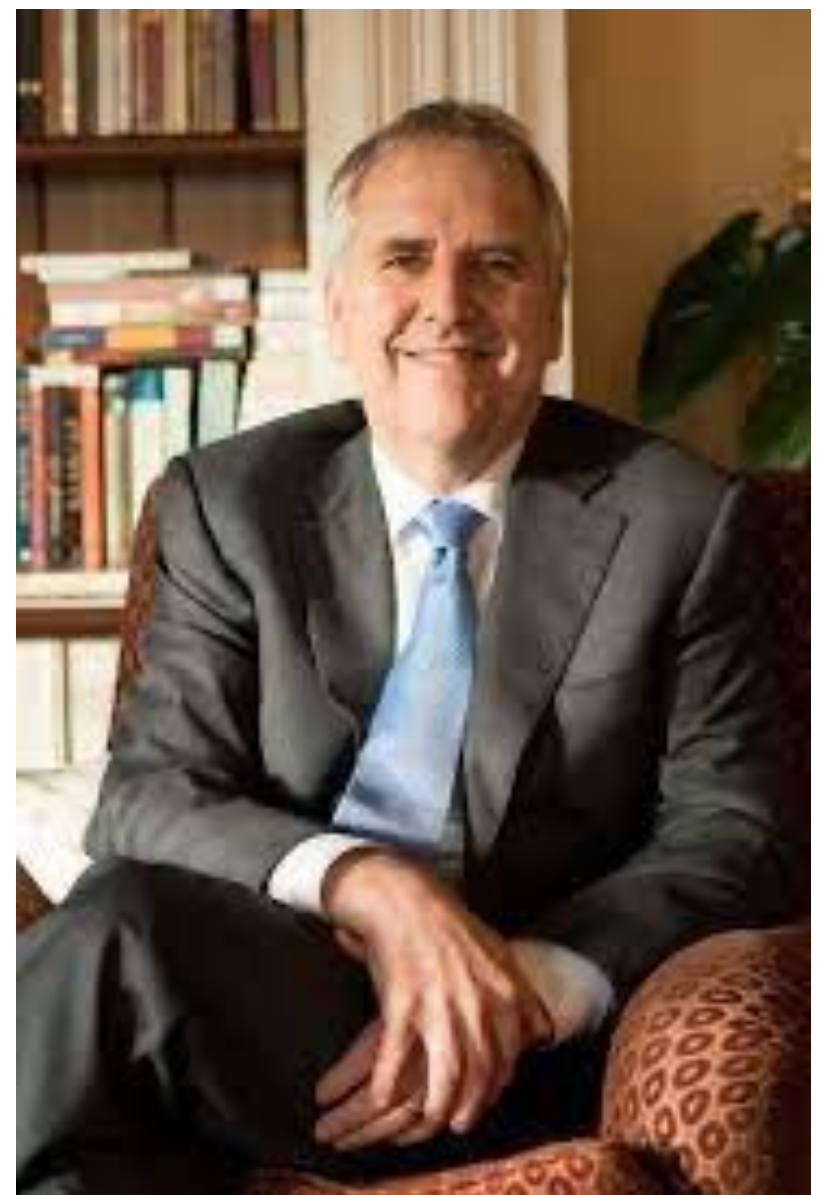
### Sports [\[edit\]](#)

- [Michael Jordan \(footballer\)](#) (born 1986), English goalkeeper
- [Mike Jordan \(racing driver\)](#) (born 1958), English racing driver
- [Mike Jordan \(baseball, born 1863\)](#) (1863–1940), American baseball player
- [Mike Jordan \(cornerback\)](#) (born 1992), American football cornerback
- [Michael Jordan \(offensive lineman\)](#) (born 1998), American football offensive lineman
- [Michael–Hakim Jordan](#) (born 1977), American professional basketball player
- [Michal Jordán](#) (born 1990), Czech ice hockey player

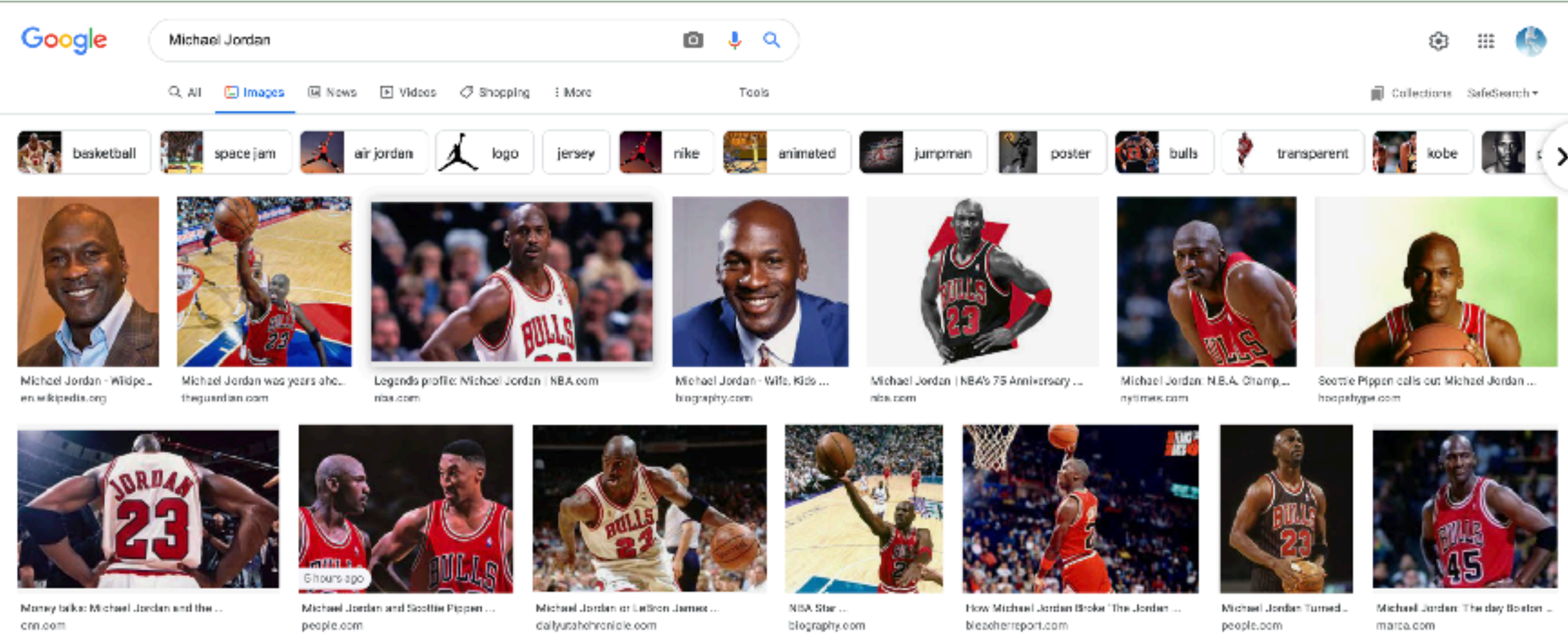
### Other people [\[edit\]](#)

- [Michael B. Jordan](#) (born 1987), American actor
- [Michael I. Jordan](#) (born 1956), American researcher in machine learning and artificial intelligence
- [Michael Jordan \(insolvency baron\)](#) (born 1931), English businessman
- [Michael Jordan \(Irish politician\)](#), Irish Farmers' Party TD from Wexford, 1927–1932
- [Michael H. Jordan](#) (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- [Michael Jordan \(mycologist\)](#), English mycologist

When you talk about **Michael Jordan**



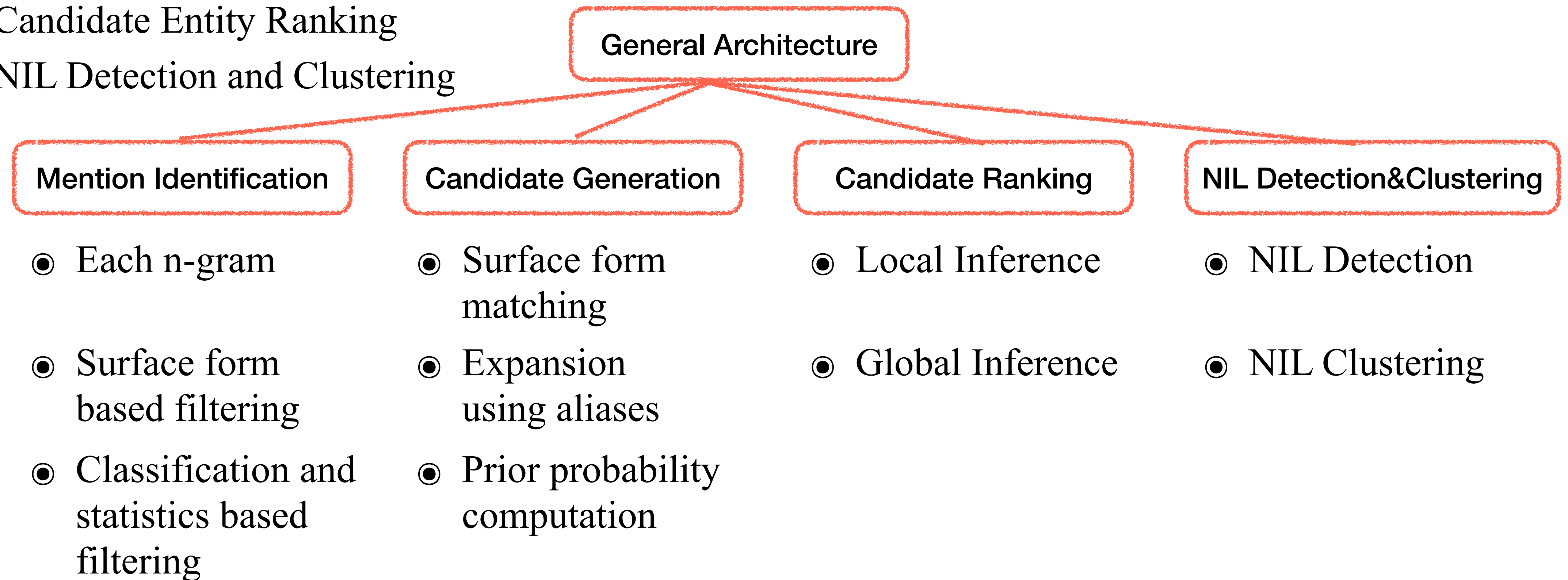
But if you search for **Michael Jordan**





# EL Pipeline

- Entity Linking requires addressing several **subtasks**:
  - Mention Identification
  - Candidate KB Entities Generation
  - Candidate Entity Ranking
  - NIL Detection and Clustering



# Mentions Identification

- ⊙ Highest recall: Each n-gram is a potential concept mention
  - Intractable for large documents
  
- ⊙ Surface form based filtering
  - Shallow parsing (especially NP chunks)
  - NP's augmented with surrounding tokens
  - Capitalized words
  - ...
  
- ⊙ Classification and statistics based filtering
  - Named entity recognition (our demo)
  - Key phrase extraction
  - ...



# Mentions Identification

- ◎ Methods used by some famous systems
  - CCG @ U. Penn. Entity Detection and Linking (2019)
    - spaCy NER model
  - Illinois Wikifier (Cheng and Roth, 2013)
    - NP chunks and substrings, NER (+nesting), prior anchor text
  - DBpedia Spotlight (Mendes et al., 2011)
    - Dictionary-based chunking with string matching (via DBpedia lexicalization dataset)
  - RPI Wikifier (Huang et al., 2014)
    - Mention Extraction (Li and Ji, 2014)

# Candidate Generation

- ◎ Surface form matching
  - A **candidate list** is composed of **entities** that **match various surface forms** of mentions in the text. (Our demo)
- ◎ Expansion using aliases
  - A dictionary of additional **aliases** is constructed using **KG metadata** or a **dictionary of aliases and synonyms**
  - e.g. **NYC** -> **New York City**
- ◎ Prior probability computation
  - Based on pre-calculated **prior probabilities** of correspondence between **certain mentions** and **entities**,  $p(e|m)$ .
  - e.g. Compute  $p(e|m)$  based on **Wikipedia entity hyperlinks (e)** and **anchor text (m)**. Based on **CrossWikis**, leveraging the **frequency of mention-entity links** in **web crawl data**.

## Mention

James Craig

## Candidates

James Craig



1st Prime Minister of Northern Ireland

	In office
	7 June 1921 – 24 November 1940
Monarch	George V Edward VIII George VI
Governor	The Duke of Abercorn
Preceded by	Position established
Succeeded by	J. M. Andrews

James Craig



Craig in *Boys' Ranch* (1946)

Born	James Henry Meador February 4, 1912 Nashville, Tennessee, U.S.
Died	June 27, 1985 (aged 73) Santa Ana, California, U.S.
Other names	James Mead
Occupation	Film actor
Years active	1937–1972
Spouse(s)	Sumie Jossi (1969–1980) Jill Jarmyn (1959–1962) (divorced) Mary June Ray (?–?)
Children	3

# Candidate Generation

- Example

Method	5 sample candidate entities for the example mention “Big Blue”
<b>surface form matched based</b> on DBpedia names	Big_Blue_Sky, Big_Blue_(Old_Dominion), Big_Blue_Crane_collapse, Dexter_Bexley_and_the_Big_Blue_Beastie, Big_Bluegrass
<b>expansion using aliases</b> on YAGO means	Big_Blue_River_(Indiana), Big_Blue_River_(Kansas), Big_Blue_(crane), Big_Red_(drink), <b>IBM</b>
<b>probability + expansion using aliases</b> on [44]:Anchor prob. + CrossWikis + YAGO	<b>IBM</b> , Big_Blue_River_(Kansas), The_Big_Blue Big_Blue_River_(Indiana), Big_Blue_(crane)



# Candidate Ranking

⦿ For each mention  $m_i$  in document  $d$ , rank the corresponding candidates  $e_j \in T(m_i)$ .

- Local Inference

- For each  $m_i$  in  $d$ :

- Generate corresponding candidate set  $T(m_i)$

- Rank entities  $e_i \in T(m_i)$ .

(i.e., consider local statistics of edges)

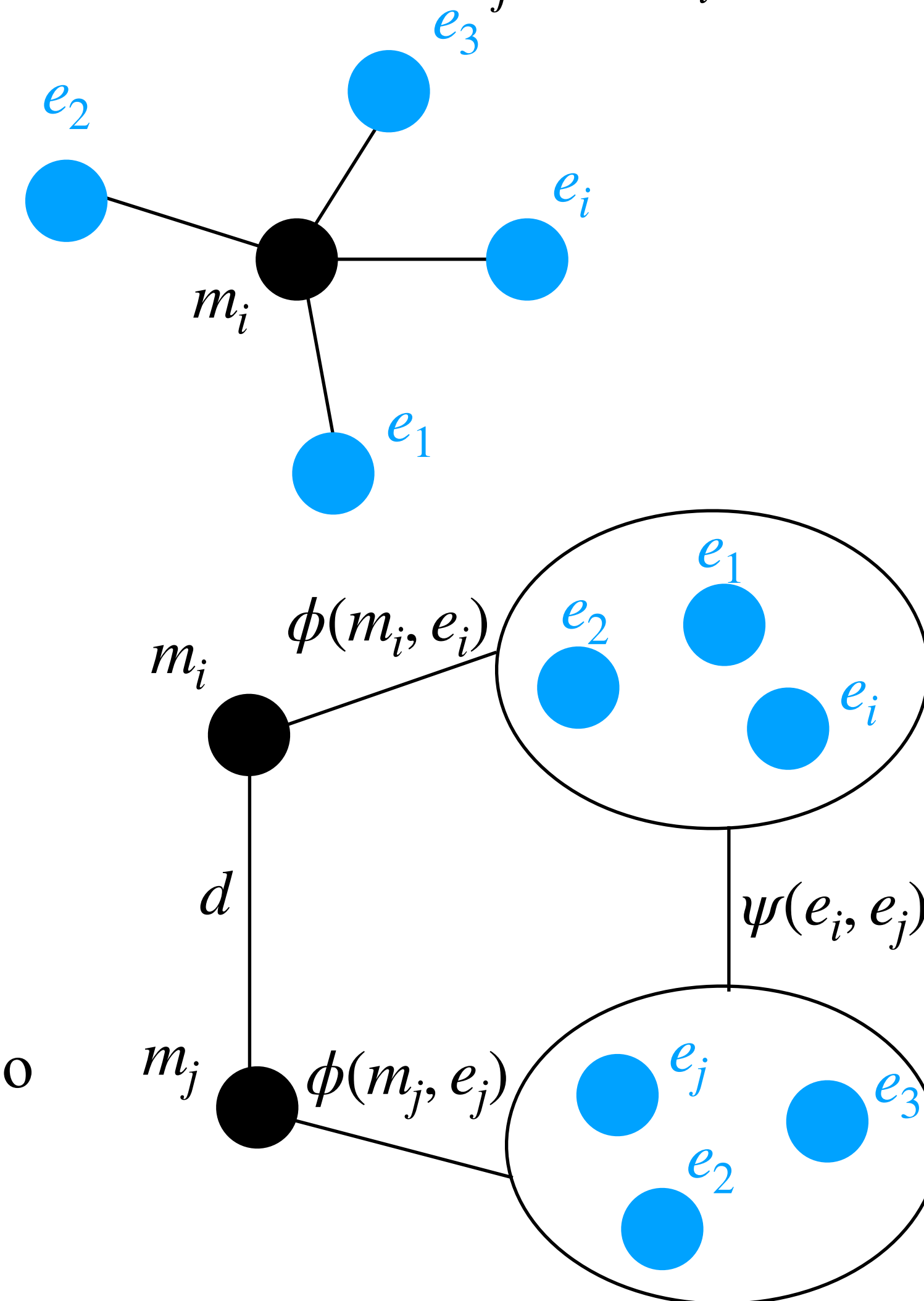
- Global Inference

- For each document  $d$ ,

- Consider all  $m_i \in d$ , and all  $e_i \in T(m_i)$

- Re-rank entities  $e_i \in T(m_i)$

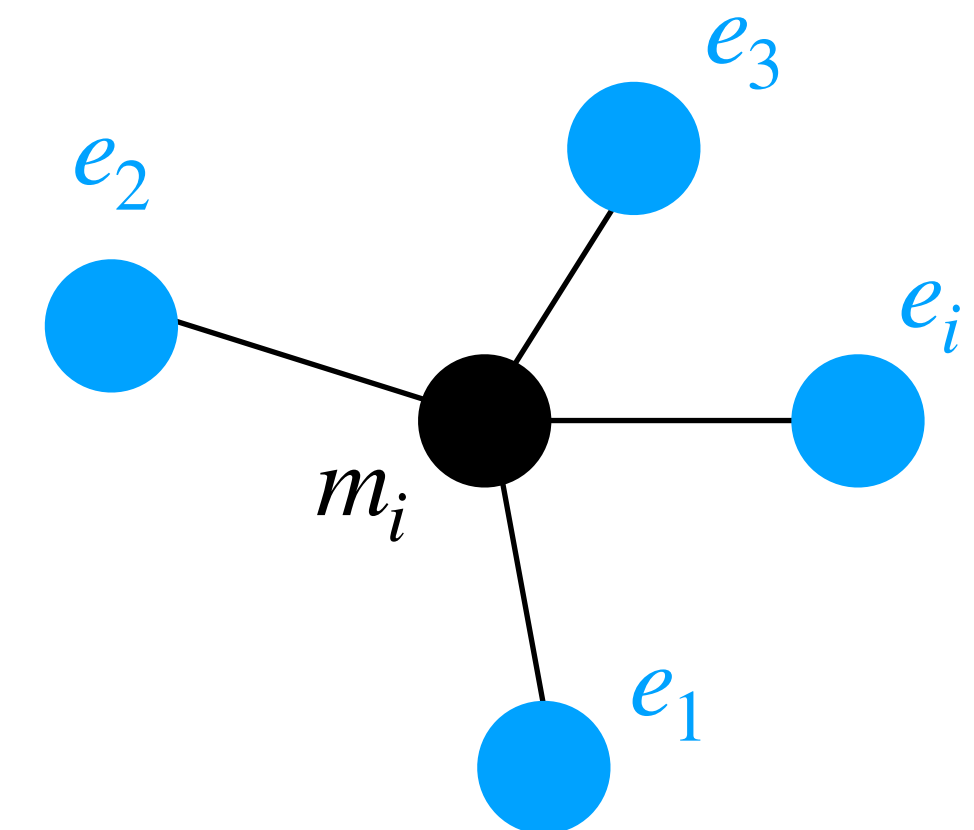
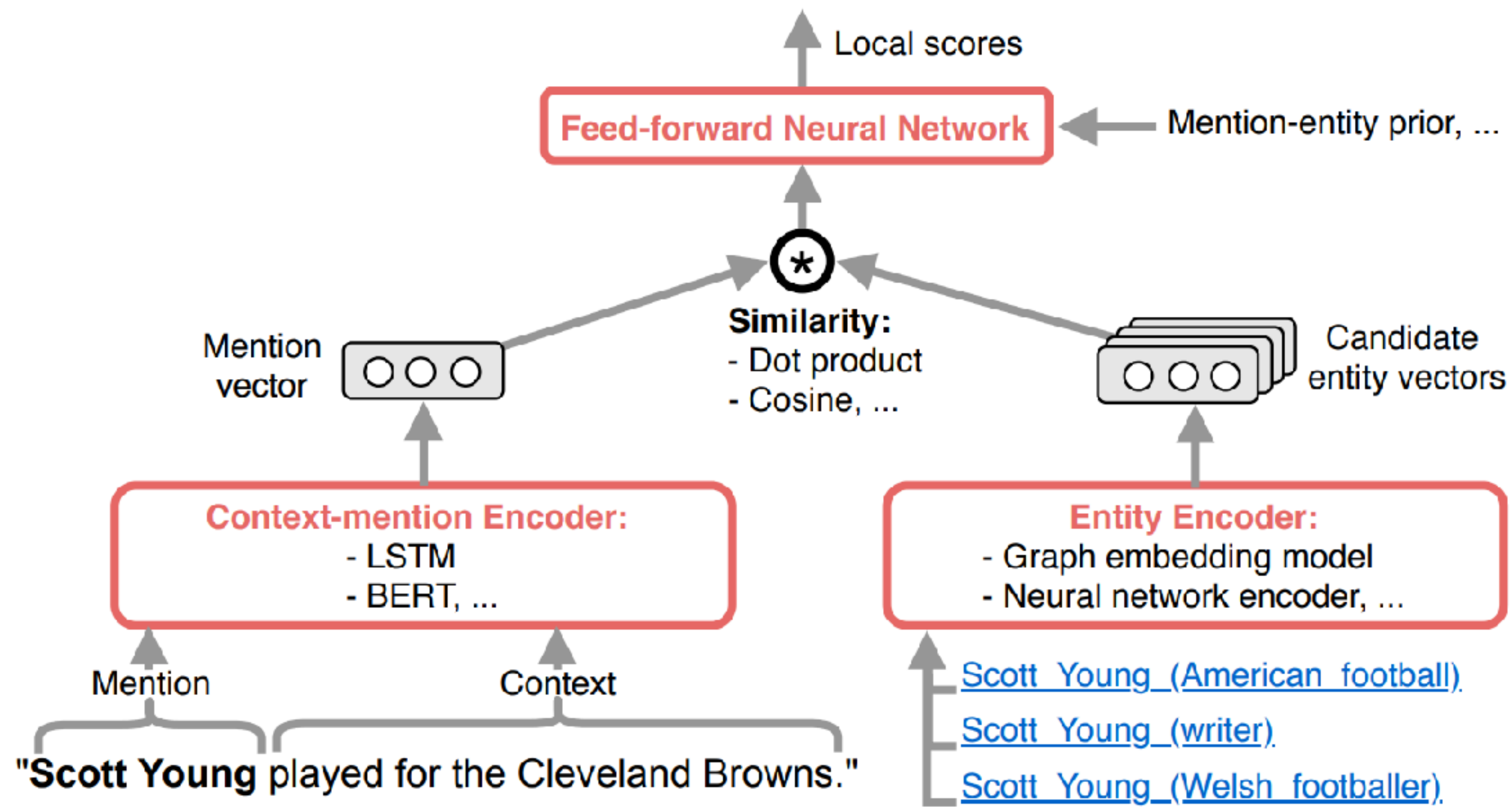
(i.e., the corresponding  $e_i$  of related  $m_i$  may also be related)



# Local Ranking

## Basic idea:

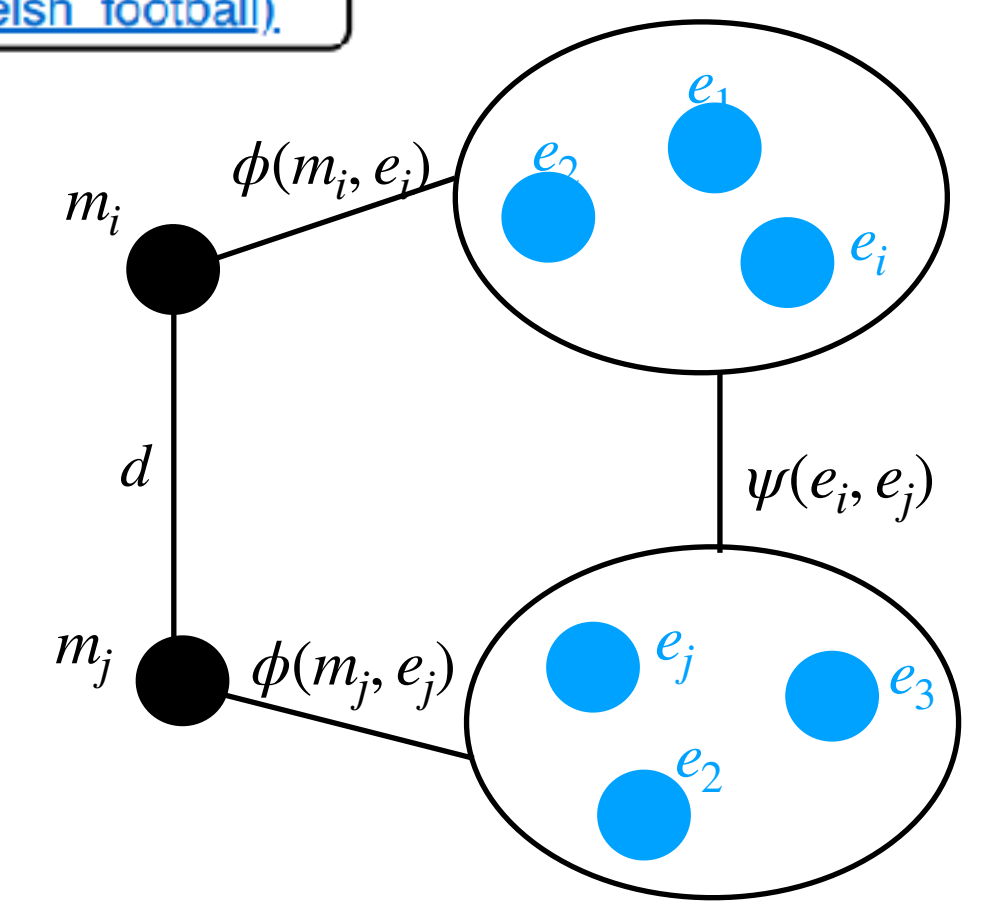
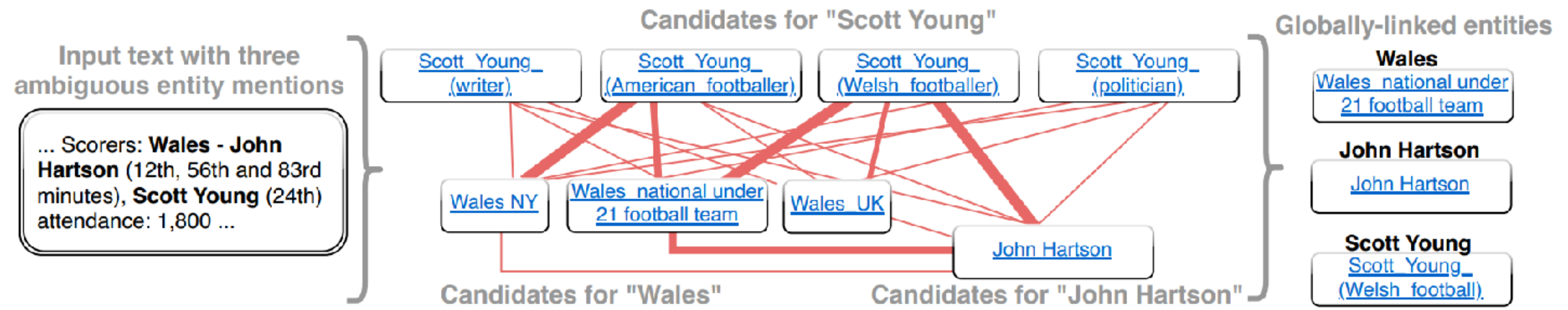
- Use **similarity measure** between **candidates**  $e_i \in T(m_i)$  and **mention**  $m_i$  with its **context**  $c_i$ .



# Global Ranking

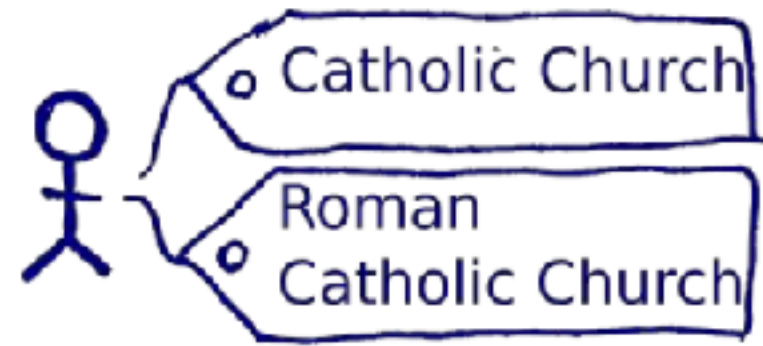
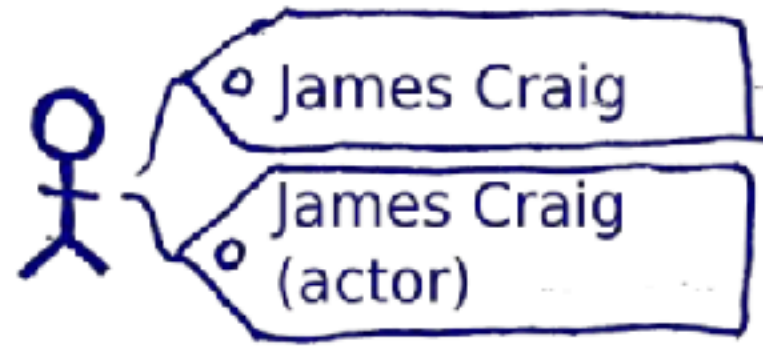
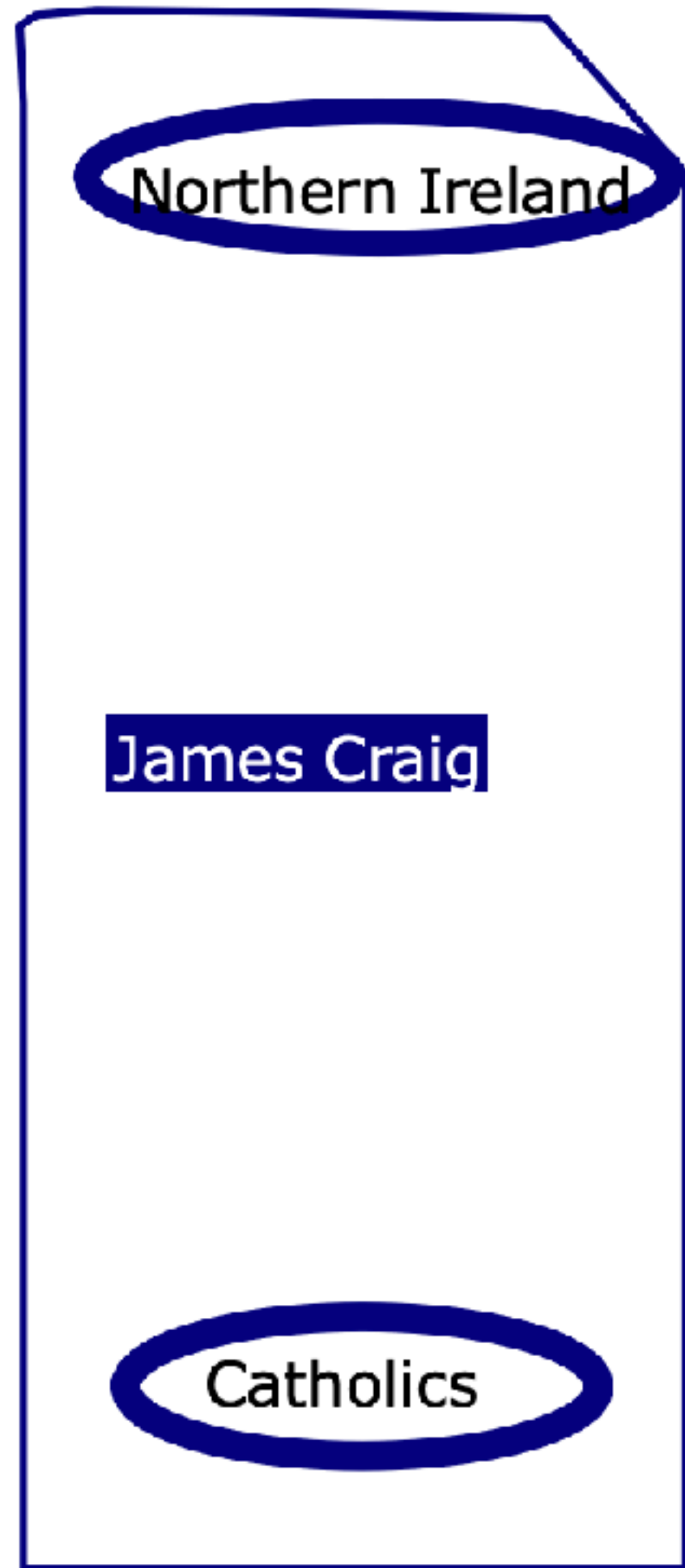
## Basic idea:

- Resolve all mentions simultaneously based on **entity coherence** and **relatedness**.

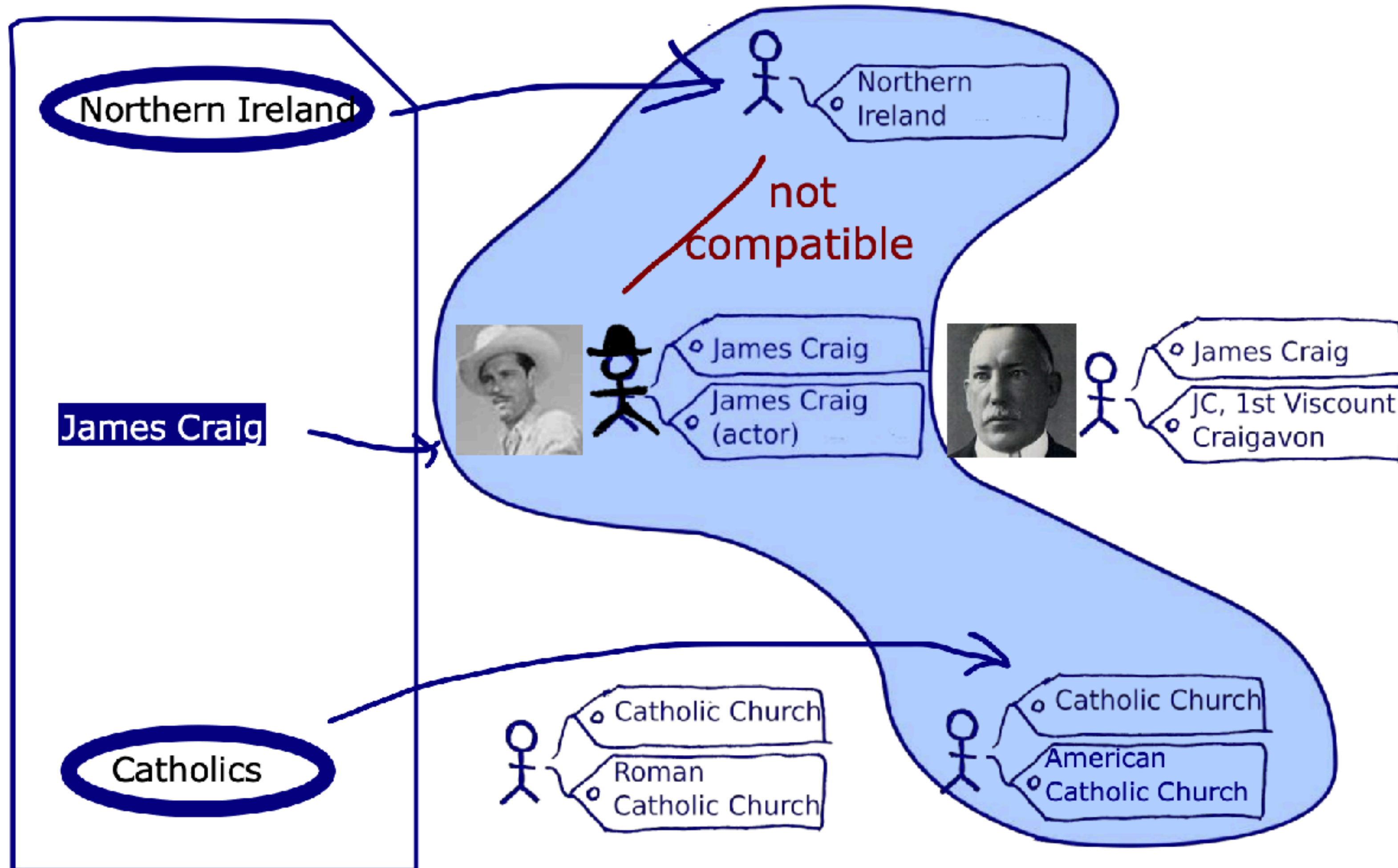




# Global Ranking: Illustration



# Global Ranking: Illustration



# Global Ranking: Formulation

- A Combinatorial Optimization Problem

- Problem Formulation

$$\Gamma^* = \arg \max_{\Gamma} \sum_{i=1}^N [\phi(m_i, e_i) + \sum_{e_i \in \Gamma_i, e_j \in \Gamma_j} \psi(e_i, e_j)]$$

- How to define the coherence and relatedness between two candidate entities?  
(What is  $\psi$ ?)



## Co-occurrence

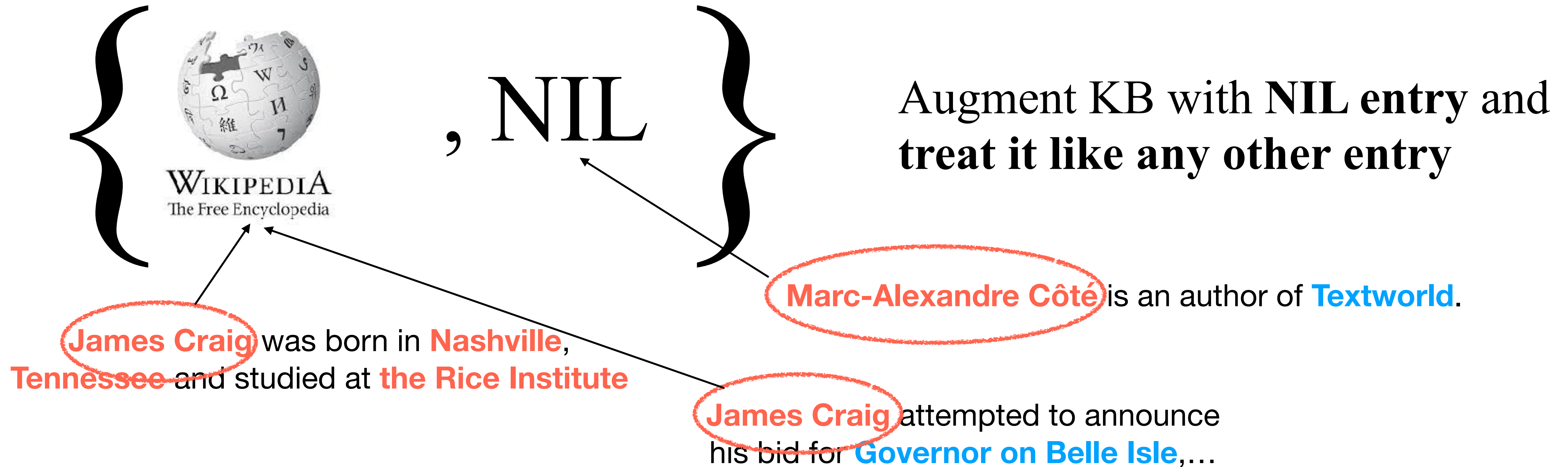


The screenshot shows the Wikipedia article for Chicago. A blue box highlights the text: "New York City and Los Angeles often appear". The article text includes: "Chicago (/ˈtʃɪkɑːɡoo/ (listen) *shih-KAH-goh*, locally also /ˈtʃɪkɑːɡoo/ *shih-KAV-goh*<sup>[5]</sup>), officially the City of Chicago, is the most populous city in the U.S. state of Illinois, and the third-most populous city in the United States, following New York City and Los Angeles. With a population of 2,746,388 in the 2020 census,<sup>[4]</sup> it is also the most populous city in the Midwestern United States and the fifth-most populous city in North America. Chicago is the county seat of Cook County, the second-most populous county in the U.S., while a small portion of the city's O'Hare Airport also extends into DuPage County. Chicago is the principal city of the Chicago metropolitan area, defined as either the U.S. Census Bureau's metropolitan statistical area (9.6 million people) or the combined statistical area (almost 10 million residents), often called Chicagoland. It is one of the 40 largest urban areas in the world. Located on the shores of freshwater Lake Michigan, Chicago was incorporated as a city in 1837 near a portage between the Great Lakes and the Mississippi River watershed and grew rapidly in the mid-19th century.<sup>[7]</sup> After the Great Chicago Fire of 1871, which destroyed several square miles and left more than 100,000 homeless, the city rebuilt.<sup>[8]</sup> The construction boom accelerated population growth throughout the following decades, and by 1900, less than 30 years after the great fire, Chicago was the fifth-largest city in the world.<sup>[9]</sup> Chicago made noted contributions to urban planning and zoning standards, including new construction styles (including the Chicago School of architecture), the development of the City Beautiful Movement, and the steel-framed skyscraper.<sup>[10][11]</sup>

- Typical approach: generate a graph containing **candidate entities**  $e$  of **mentions**  $m$  in  $d$  and perform **random walk algorithms** (e.g. PageRank) over it to select highly consistent entities.
- Calculate  $\psi(e_i, e_j)$  as Pointwise Mutual Information of  $e_i, e_j$  based on their incoming sets.
- ...

# 104 NIL Detection & Clustering

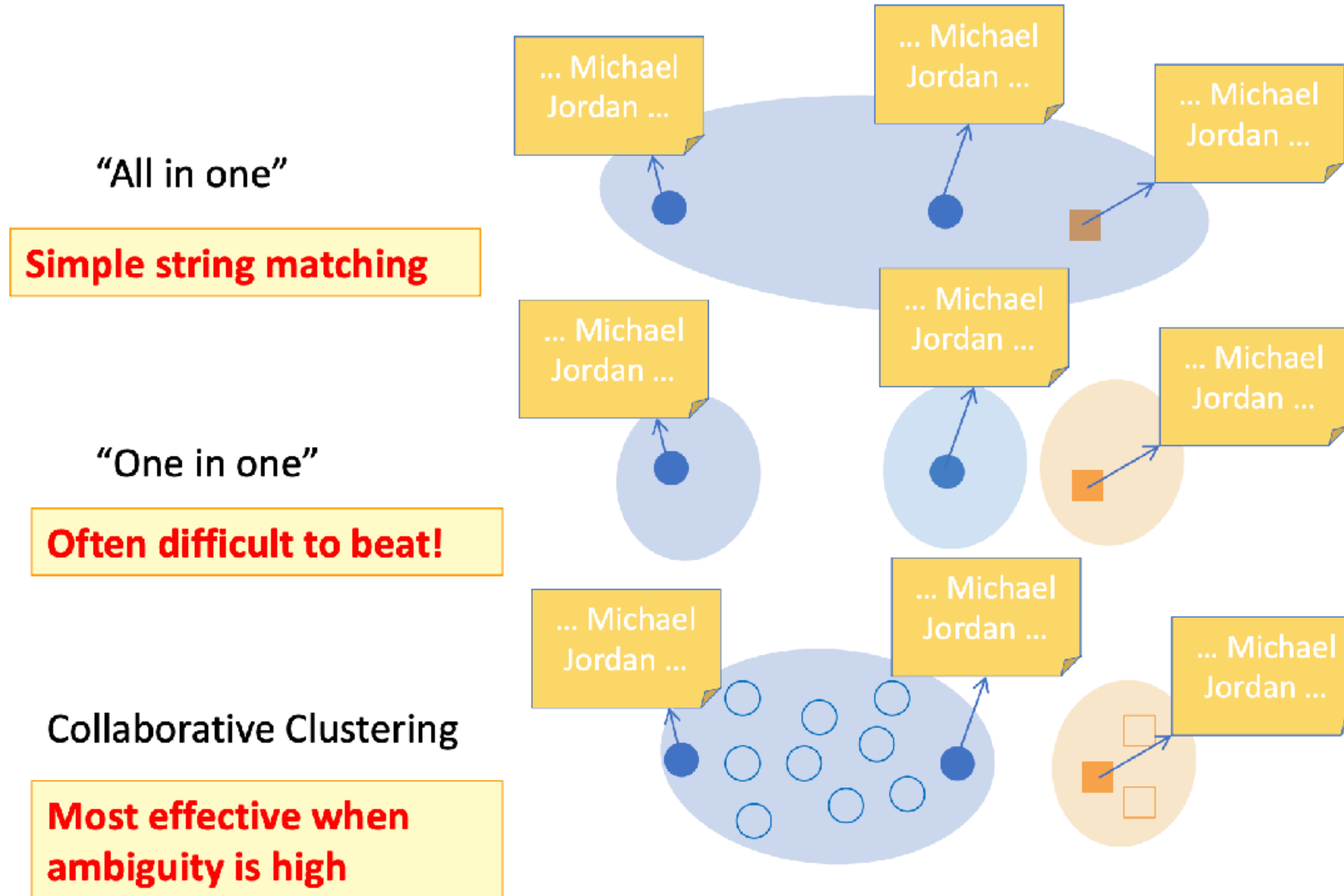
## © NIL Detection



- **Binary classification (Within KB vs. NIL)**
- **Select NIL cutoff by tuning confidence threshold**

# 105 NIL Detection & Clustering

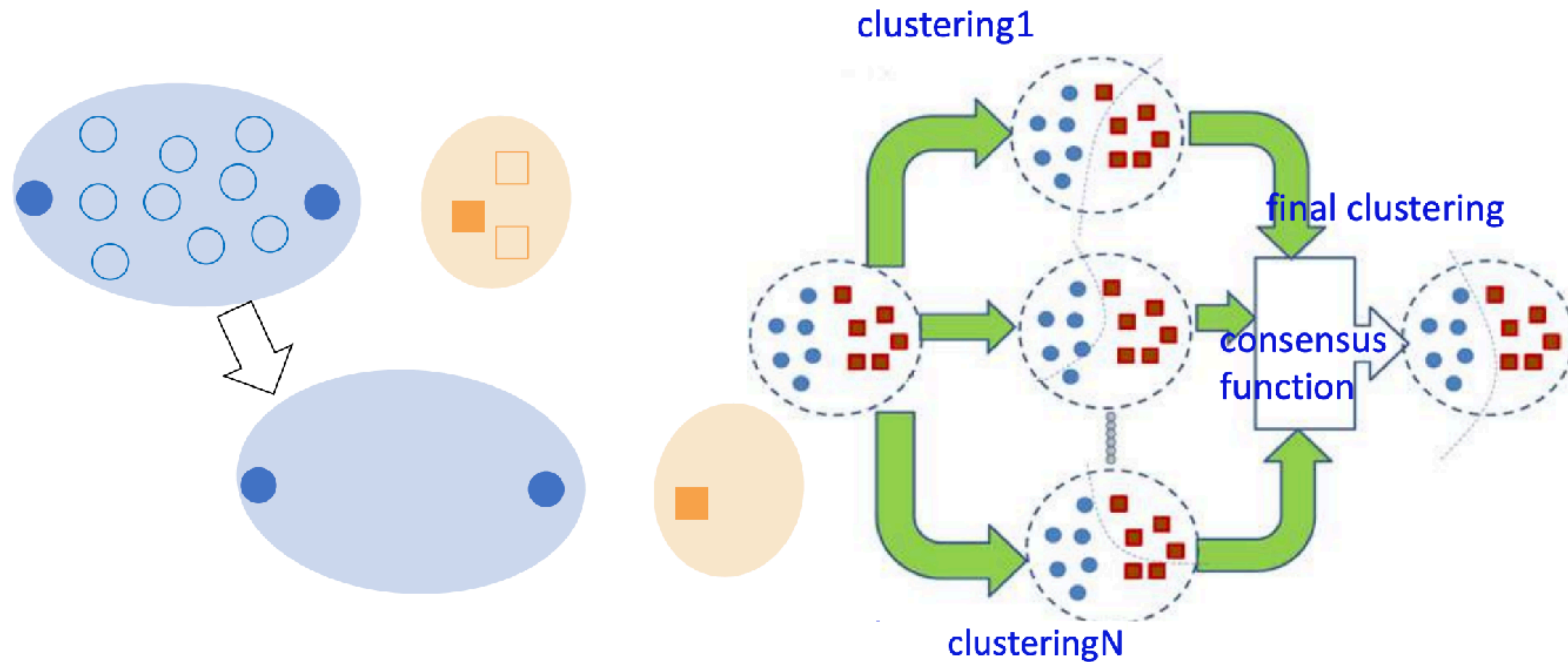
## ● NIL Clustering





# 106 NIL Detection & Clustering

## ● Collaborative Clustering



## 107 EL Tools

### Tools (Target KB: Wikipedia):

- FacebookResearch BLINK 
  - SOTA of zero-shot EL
  - Well-documented
- OpenAI DeepType  OpenAI
  - SOTA of EL
- Spacy Entity Linker 
  - Easy to use

# spaCy

```
import spacy # version 3.0.6'

# initialize language model
nlp = spacy.load("en_core_web_md")

# add pipeline (declared through entry_points in setup.py)
nlp.add_pipe("entityLinker", last=True)

doc = nlp("I watched the Pirates of the Caribbean last silvester")

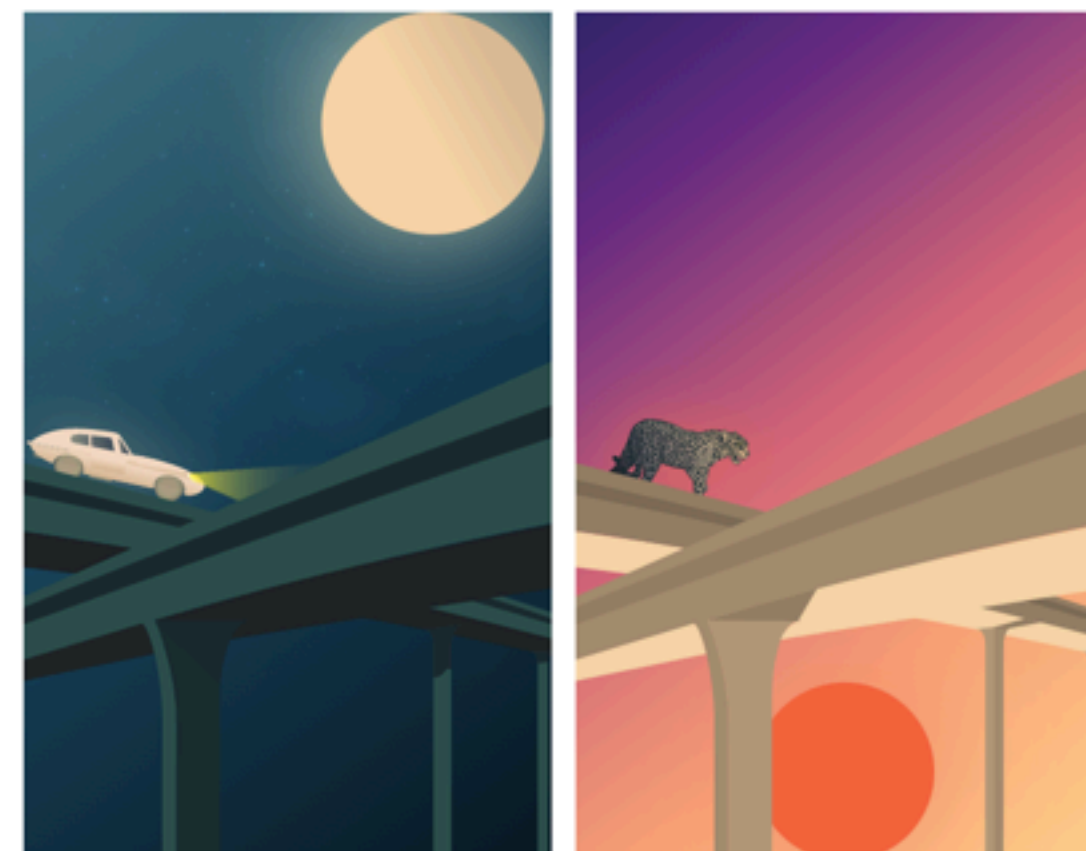
# returns all entities in the whole document
all_linked_entities = doc._.linkedEntities
# iterates over sentences and prints linked entities
for sent in doc.sents:
    sent._.linkedEntities.pretty_print()

# OUTPUT:
# https://www.wikidata.org/wiki/Q194318      Pirates of the Caribbean
# https://www.wikidata.org/wiki/Q12525597    Silvester
```

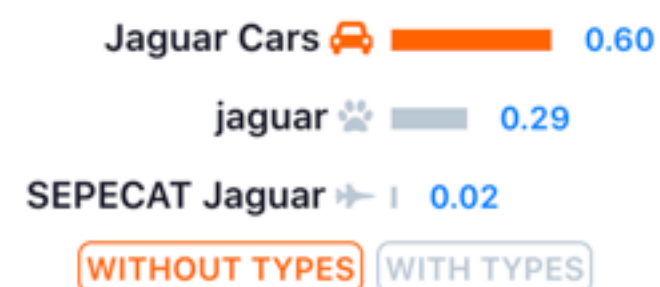
● OpenAI DeepType 

Figuring out which object is meant by a word by having a neural network decide if the word belongs to each of about 100 automatically-discovered “types”.

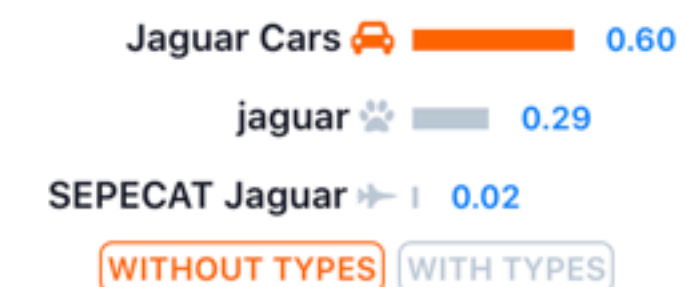
Rather than trying to reason directly whether **jaguar** means the **car**, **animal**, or **something else**, the system **plays “20 questions”** with a **pre-chosen set of categories**.



The man saw a Jaguar speed on the highway.

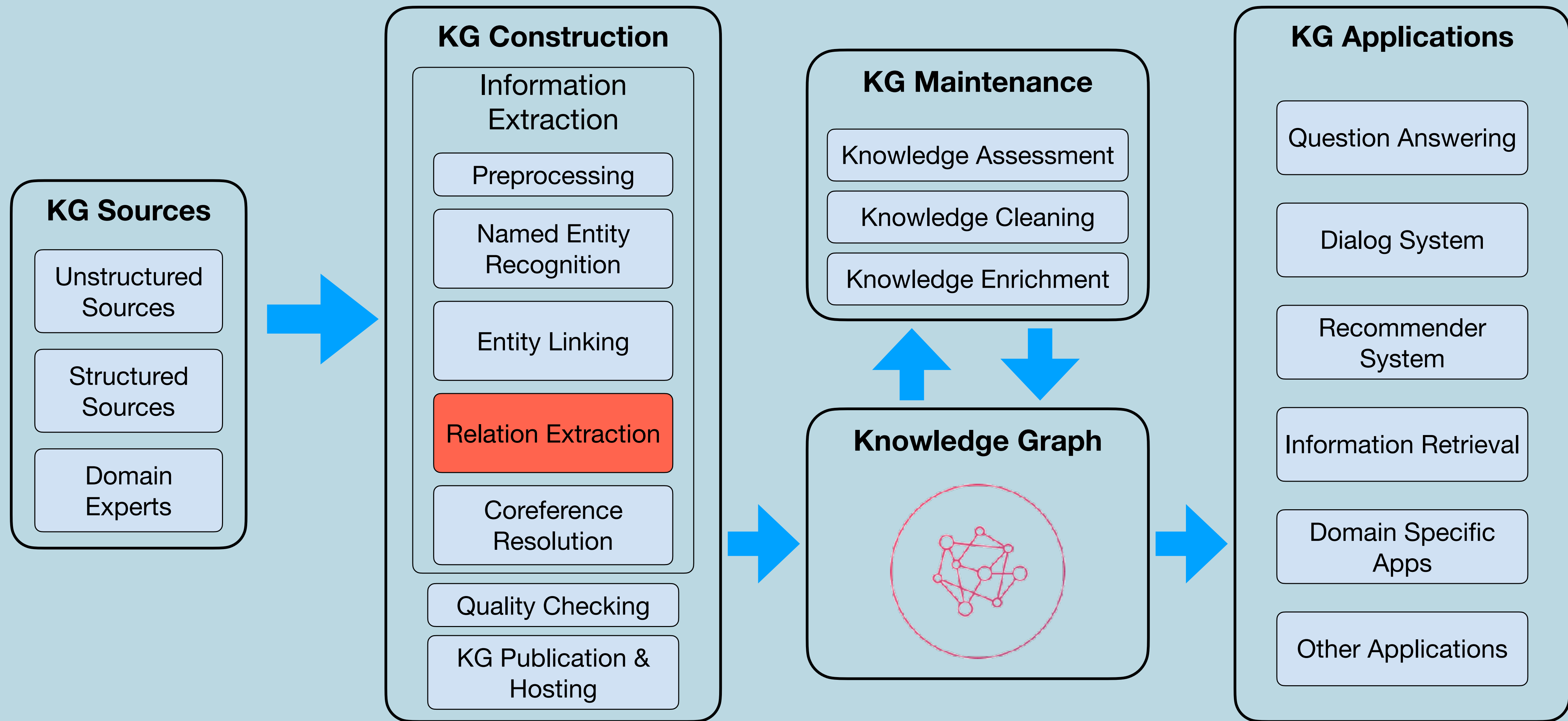


The prey saw the jaguar cross the jungle.





# Relation Extraction



# 110 What is relation extraction (RE)?

- Company report: “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co.(C-T-R) ...”
- Complex relation in the report:

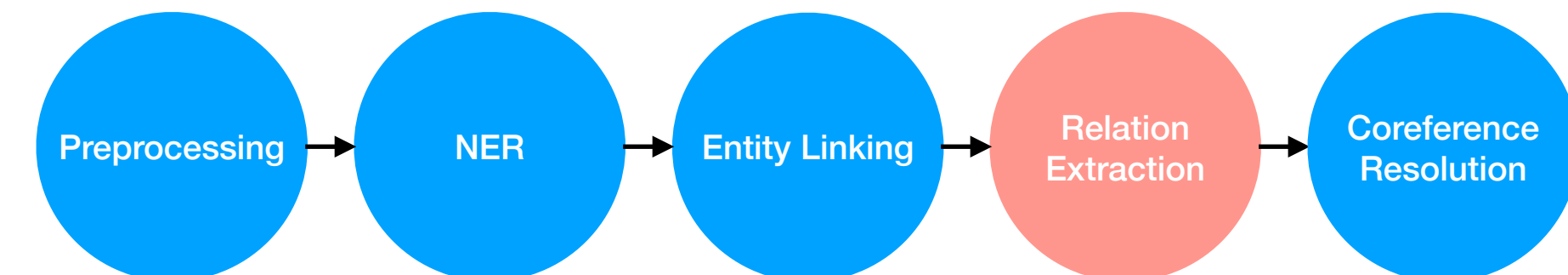
Company-Founding	
Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

- But we will focus on the simpler task of extracting relation **triples**:

Founding-year(IBM, 1991)

Founding-location(IBM, New York)

⋮



# 111 Extracting relation triples from text



**Textual abstract:  
Summary for human**



Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

**Structured knowledge extraction:  
Summary for machine**



# 112 Extracting relation triples from text: illustration

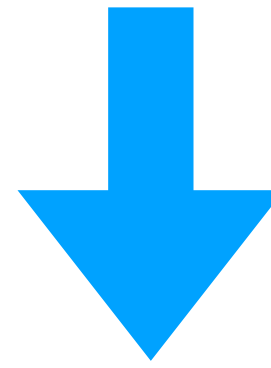


WIKIPEDIA  
The Free Encyclopedia

## Wiki Text

### Policing career [edit]

Craig first worked at the [Detroit Police Department](#) from 1977 until he was laid off in 1981. He went to [Los Angeles](#) to work for the [LAPD](#) for nearly three decades.<sup>[3]</sup> From 2009 to 2011, he was police chief for the [Portland Police Department](#) in [Portland, Maine](#). From 2011 to 2013, he was chief of police for the [Cincinnati Police Department](#) in [Cincinnati, Ohio](#). During his tenure with the Cincinnati police, the crime rate was its lowest since 2000, though it was part of an overall trend of decreasing crime.<sup>[4]</sup>



Craig first worked at the Detroit Police Department from 1977 until he was laid off in 1981. He went to Los Angeles to work for the LAPD for nearly three decades.



#### Named Entity Recognition

Craig **PERSON** first worked at **the Detroit Police Department** **ORG** from **1977** **DATE** until he was laid off in **1981** **DATE**. He went to **Los Angeles** **GPE** to work for the **LAPD** **ORG** for **nearly three decades** **DATE**.

#### Entity Mention Detection

Craig **PER** first worked at the **Detroit Police Department** **ORG** from 1977 until **he** **PER** was laid off in 1981. **He** **PER** went to **Los Angeles** **GPE** to work for **the LAPD** **ORG** for nearly three decades.

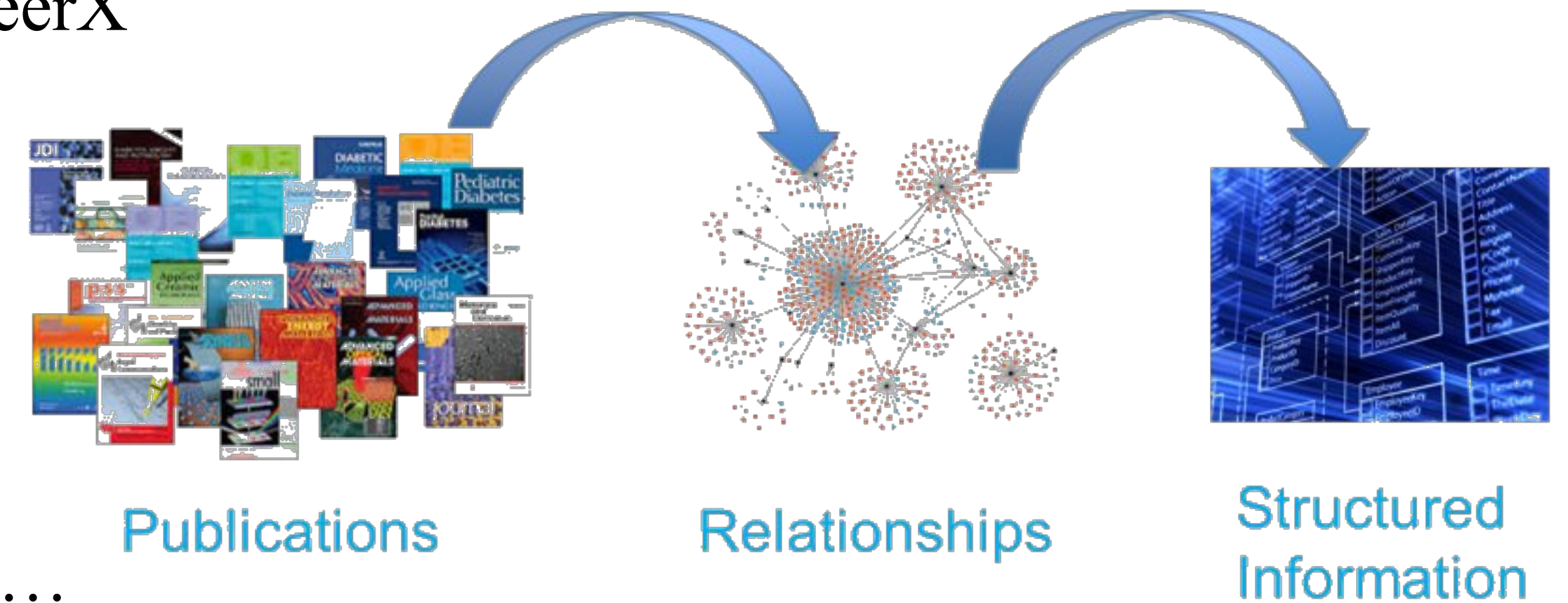
#### Relation Extraction

**Craig** **SUBJ** first worked at **the Detroit Police Department** **OBJ** from 1977 until he was laid off in 1981. **He** **SUBJ** **SUBJ** went to **Los Angeles** **OBJ** to work for **the LAPD** **OBJ** for nearly three decades.

Relations: **ORG-AFF** (Craig to the Detroit Police Department), **PHYS** (He to Los Angeles), **ORG-AFF** (He to the LAPD).

# Why Relation Extraction?

- Building & extending knowledge bases and ontologies
- Scholarly literature databases: Google Scholar, CiteSeerX
- People directories: Rapleaf, Spoke, Naymz
- Shopping engines & product search
- Bioinformatics: clinical outcomes, gene interactions, ...
- Stock analysis: deals, acquisitions, earnings, hirings & firings
- Intelligence analysis for business & government
- ...





# 114 But what relations should we extract?

⊙ Depends on the **domain of interest**.

- For generic news texts

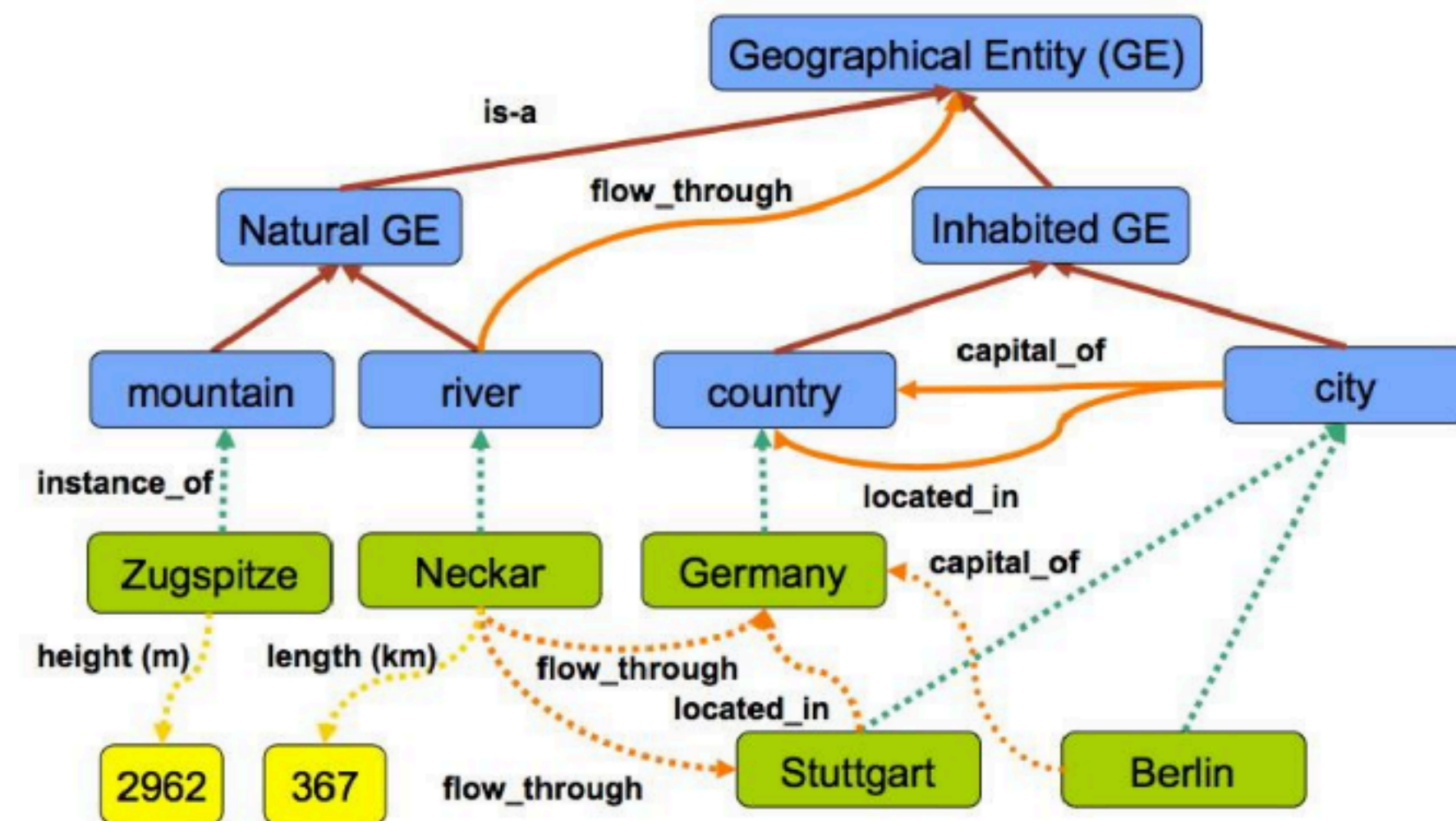
Relations	Examples	Types
<b>Affiliations</b>		
Personal	<i>married to, mother of</i>	PER → PER
Organizational	<i>spokesman for, president of</i>	PER → ORG
Artifactual	<i>owns, invented, produces</i>	(PER   ORG) → ART
<b>Geospatial</b>		
Proximity	<i>near, on outskirts</i>	LOC → LOC
Directional	<i>southeast of</i>	LOC → LOC
<b>Part-Of</b>		
Organizational	<i>a unit of, parent of</i>	ORG → ORG
Political	<i>annexed, acquired</i>	GPE → GPE

- Freebase relation types

- 23 million entities, thousands of relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States

- Geographical relation types



- Protein interactions

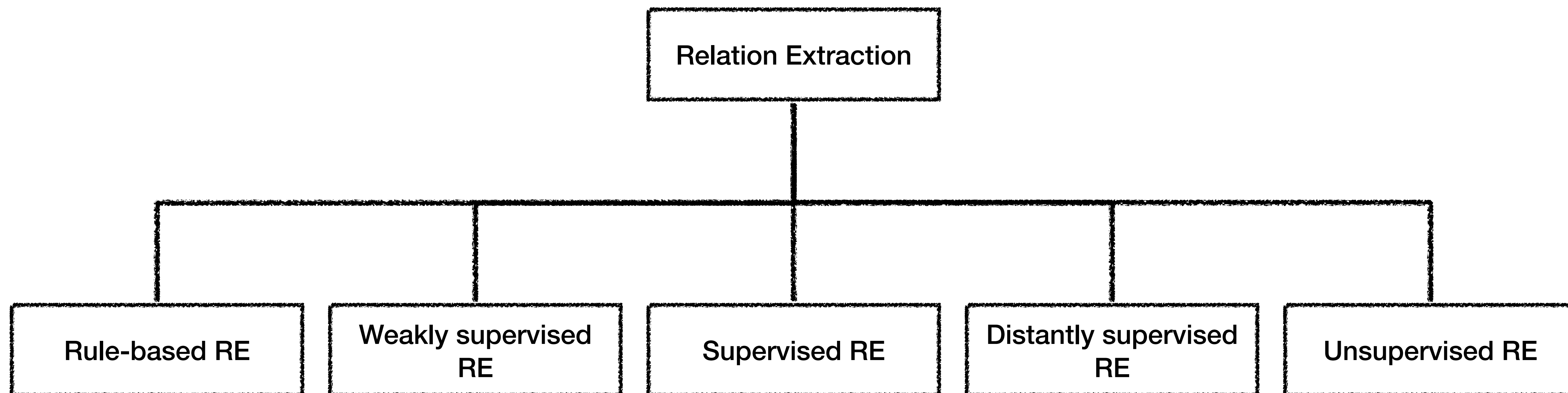
„We show that **CBF-A** and **CBF-C** interact with each other to form a **CBF-A-CBF-C complex** and that **CBF-B** does not interact with **CBF-A** or **CBF-C** individually but that it **associates** with the **CBF-A-CBF-C complex**.“

CBF-A  $\xleftrightarrow[\text{complex}]{\text{interact}}$  CBF-C

CBF-B  $\xrightarrow{\text{associates}}$  CBF-A-CBF-C complex



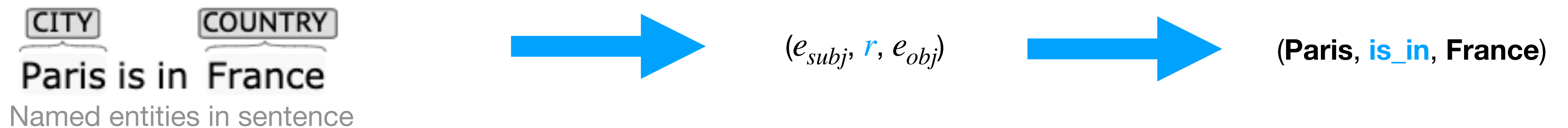
# Relation extraction approaches



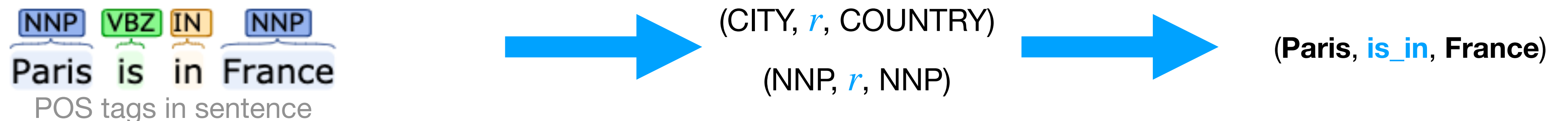
# Rule-based RE

- Many instances of **relations** can be identified through **hand-crafted patterns**, looking for triples  $(X, r, Y)$  where  $X, Y$  are entities and  $r$  are words in between.

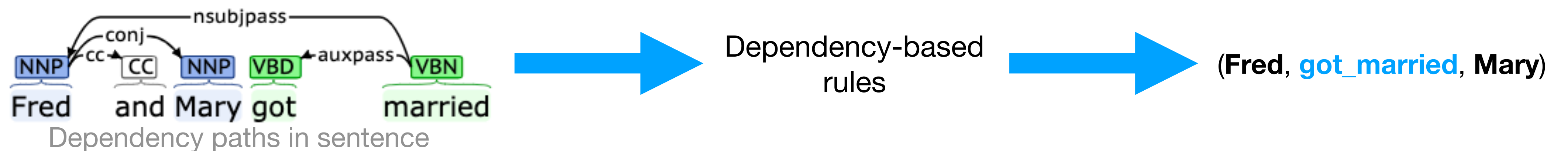
- A simple rule



- Consider entity type and part-of-speech (POS) tags to **remove additional false positive**



- Use dependency paths for **longer-range patterns** and **increase coverage**.



# 117 Rule-based RE

## ● Pros

- Humans created patterns tend to lead **high precision**
- Can be **tailored to specific domains**

## ● Cons

- Human patterns suffer from **low-recall**
- **Creation of patterns** can be **expensive** or even **intractable**
  - Hard to write and maintain the rules
  - Have to create rules for every relation type

Examples of the Hearst patterns

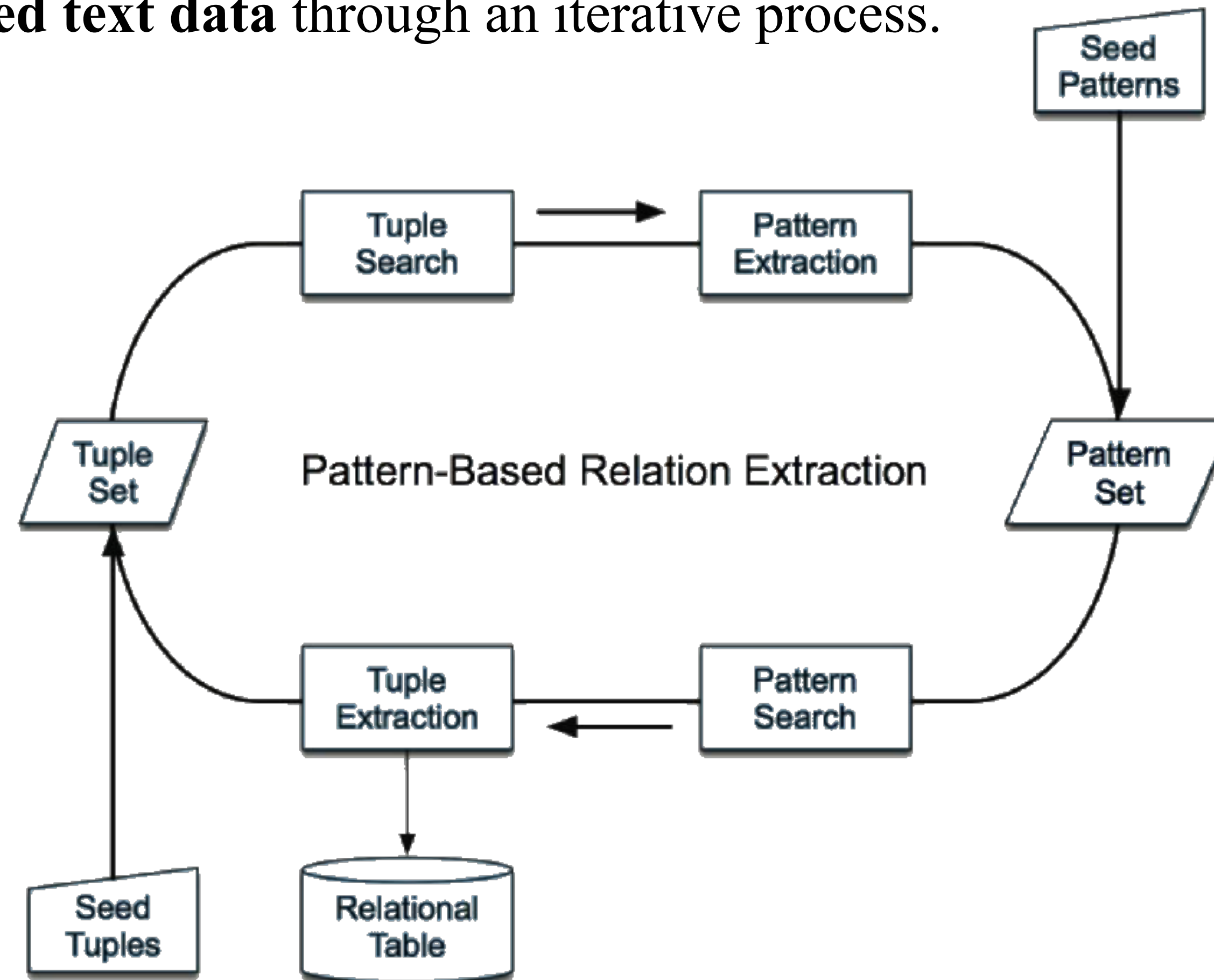
Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...



# Weakly Supervised RE

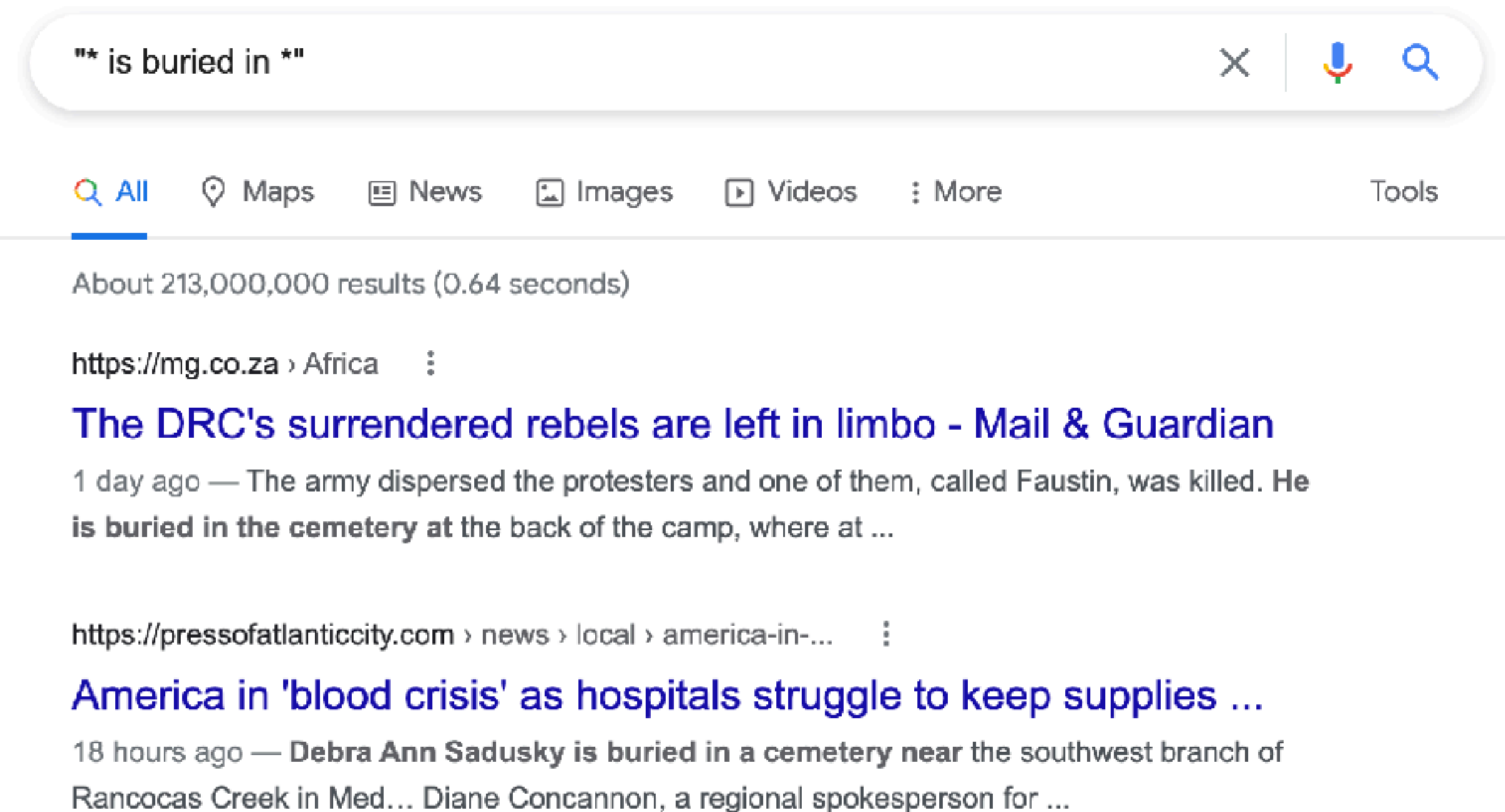
## ● Basic idea:

- **Bootstrapping:** start out with a set of hand-crafted rules and **automatically find new ones** from the **unlabeled text data** through an iterative process.



# Weakly Supervised RE: Example

- **Target relation:** burial place
- **Seed tuple:** [Mark Twain, Elmira]
- Grep/Google for “Mark Twain” and “Elmira”
  - “Mark Twain is buried in Elmira, NY”  
 → X is buried in Y
  - “The grave of Mark Twain is in Elmira”  
 → The grave of X is in Y
  - “Elmira is Mark Twain’s final resting place”  
 → Y is X’s final resting place
- Use those patterns to search for new tuples

Google search for “\*\* is buried in \*\*”

About 213,000,000 results (0.64 seconds)

<https://mg.co.za> > Africa

**The DRC's surrendered rebels are left in limbo - Mail & Guardian**  
 1 day ago — The army dispersed the protesters and one of them, called Faustin, was killed. He is buried in the cemetery at the back of the camp, where at ...

<https://pressofatlanticcity.com> > news > local > america-in-...

**America in 'blood crisis' as hospitals struggle to keep supplies ...**  
 18 hours ago — Debra Ann Sadusky is buried in a cemetery near the southwest branch of Rancocas Creek in Med... Diane Concannon, a regional spokesperson for ...

# Weakly Supervised RE

## Pros

- **Higher recall:** more relations can be discovered
- **Less cost:** only require a high quality seed

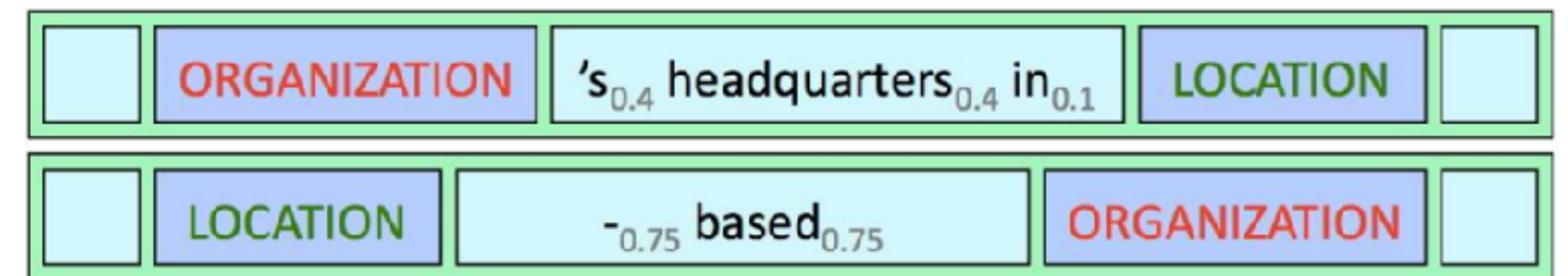
## Cons

- **Sensitive to original set of seeds**
- Problem of **semantic drift** at each iteration
- **Precision** tends to be **not that high**
- **Still hard for relation type extension:** require manually labeled new seeds for new relation types

A classical work about weakly supervised RE

**Snowball (Agichtein & Gravano 2000)**

Require that X and Y be named entities of particular types



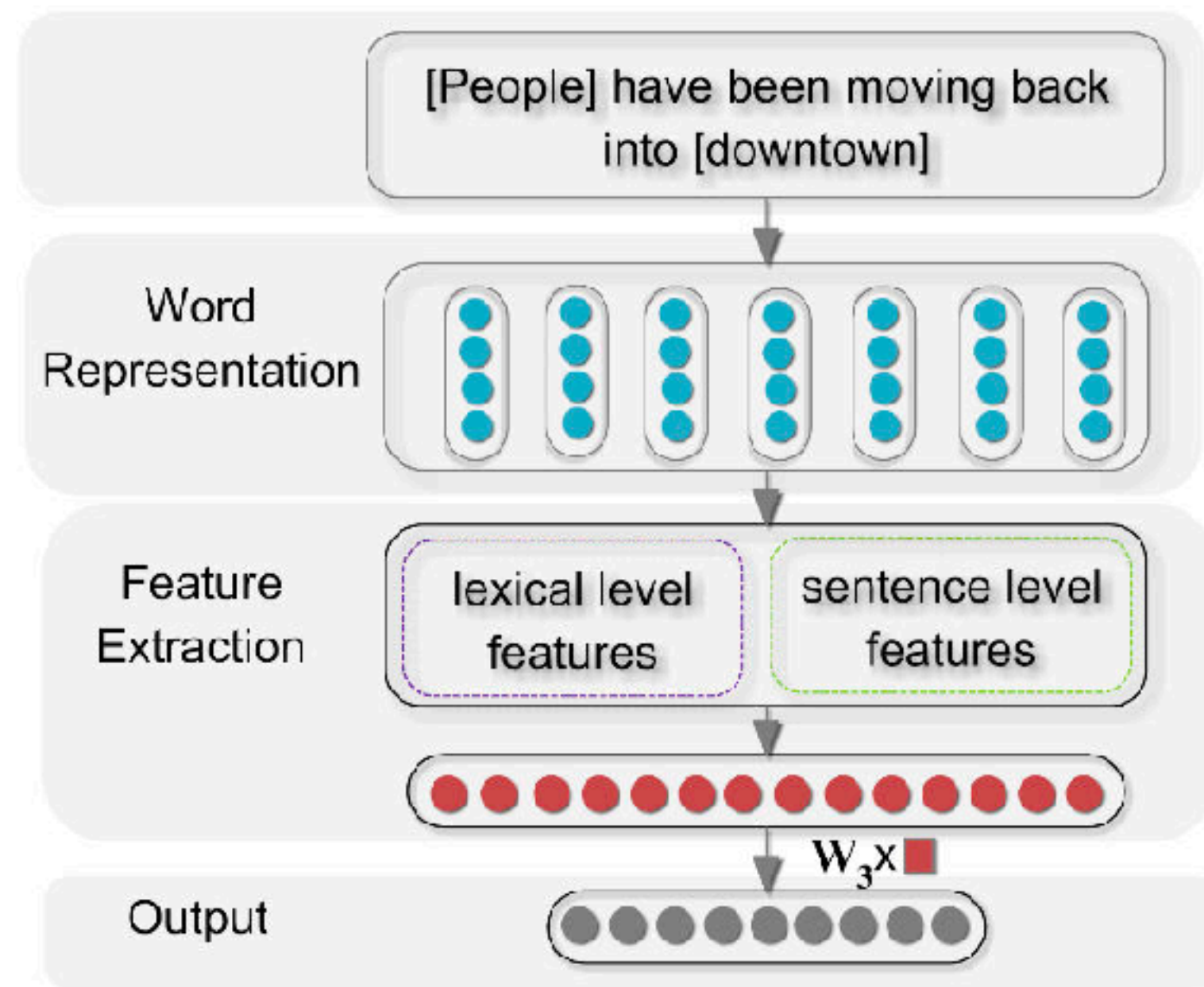


# Supervised RE

- ◎ The supervised approach requires:
  - Defining an inventory of output **labels**
    - Relation detection: **True/False**
    - Relation classification: **located\_in, employee\_of, be\_born\_in, ...**
  - Collection sufficient **labeled training data (Hard Part)**
    - Public benchmarks: MUC-7, ACE-2005, SemEval, NYT10, FewRel, ...
  - Designing architecture of neural networks for **DL-based methods**  
(or designing feature representation & choosing classifier for **ML-based methods**)
    - CNNs, RNNs, Transformers, ...
  - **Training and evaluation**

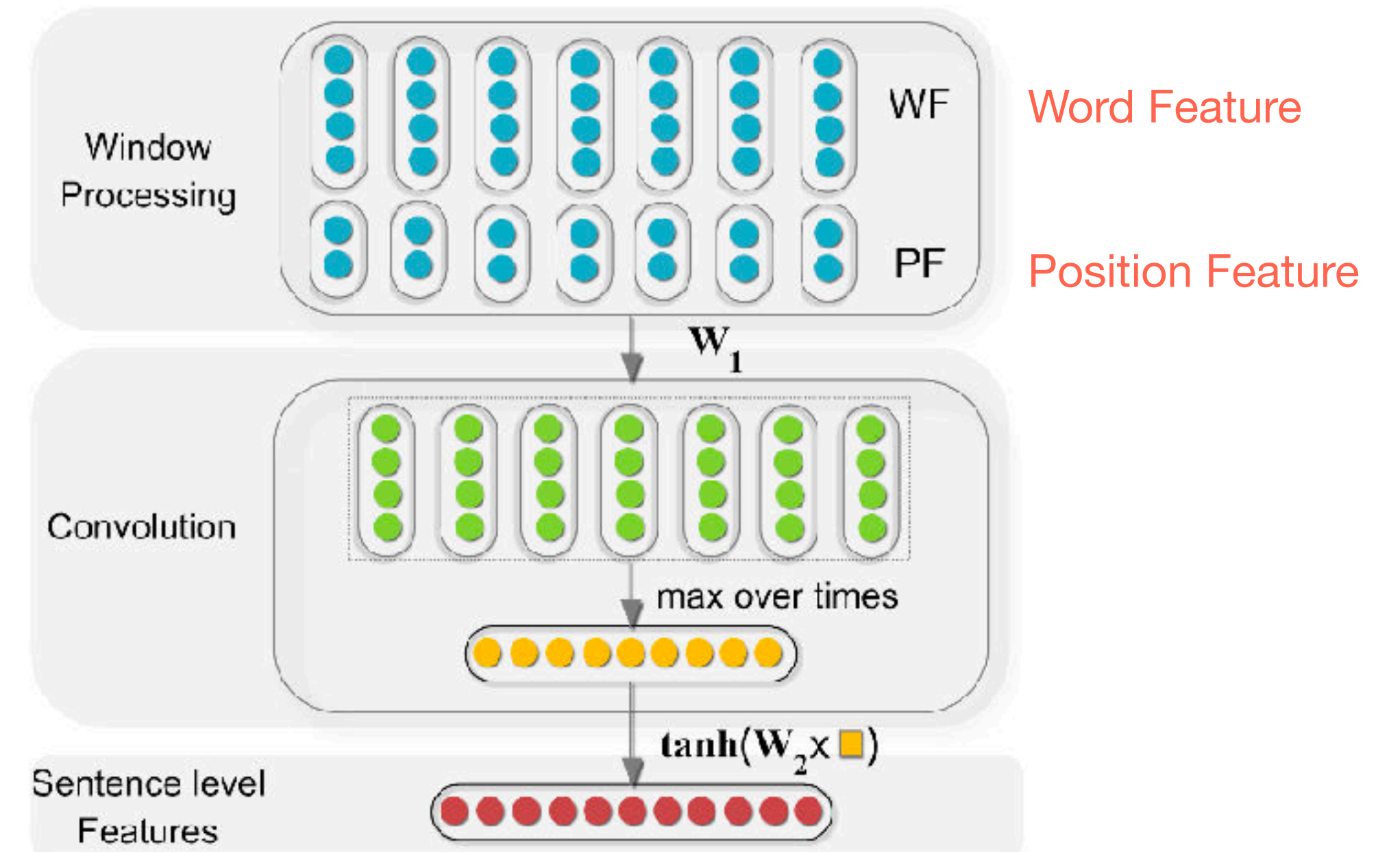
# 122 How are neural networks being used to do RE?

- [Zeng et al. 2014] apply **word embeddings** and **Convolutional Neural Network (CNN)** to relation classification.



lexical features:  
specific to entity pairs

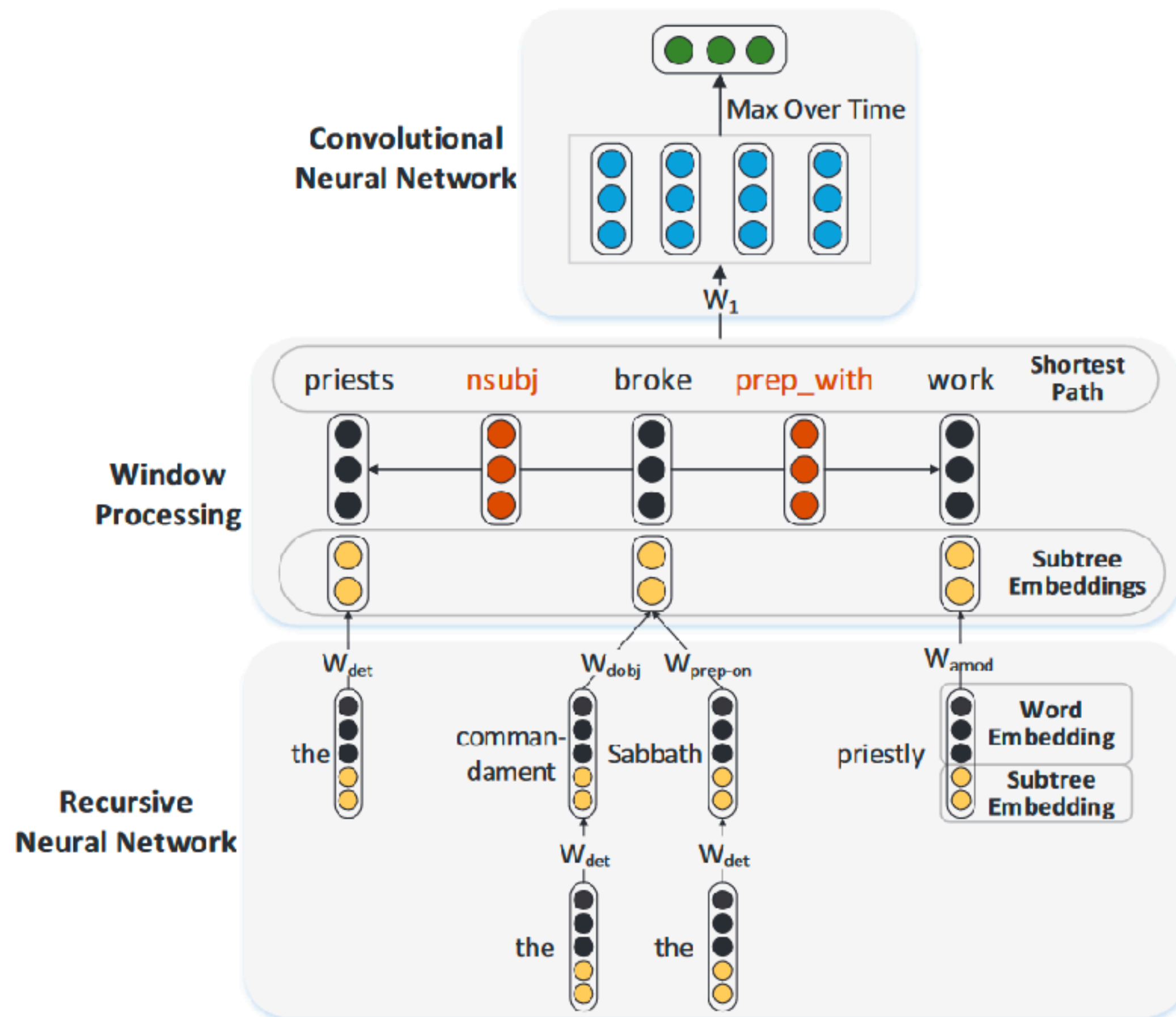
(a) Neural Network Architecture for Relation Classification.



(b) Extracting Sentence-level Features using CNN.

# 123 How are neural networks being used to do RE?

- [Liu et al. 2015] propose a **recursive neural network (RNN)** to model the **dependency subtrees**, and a CNN to capture the most important features on the shortest path.

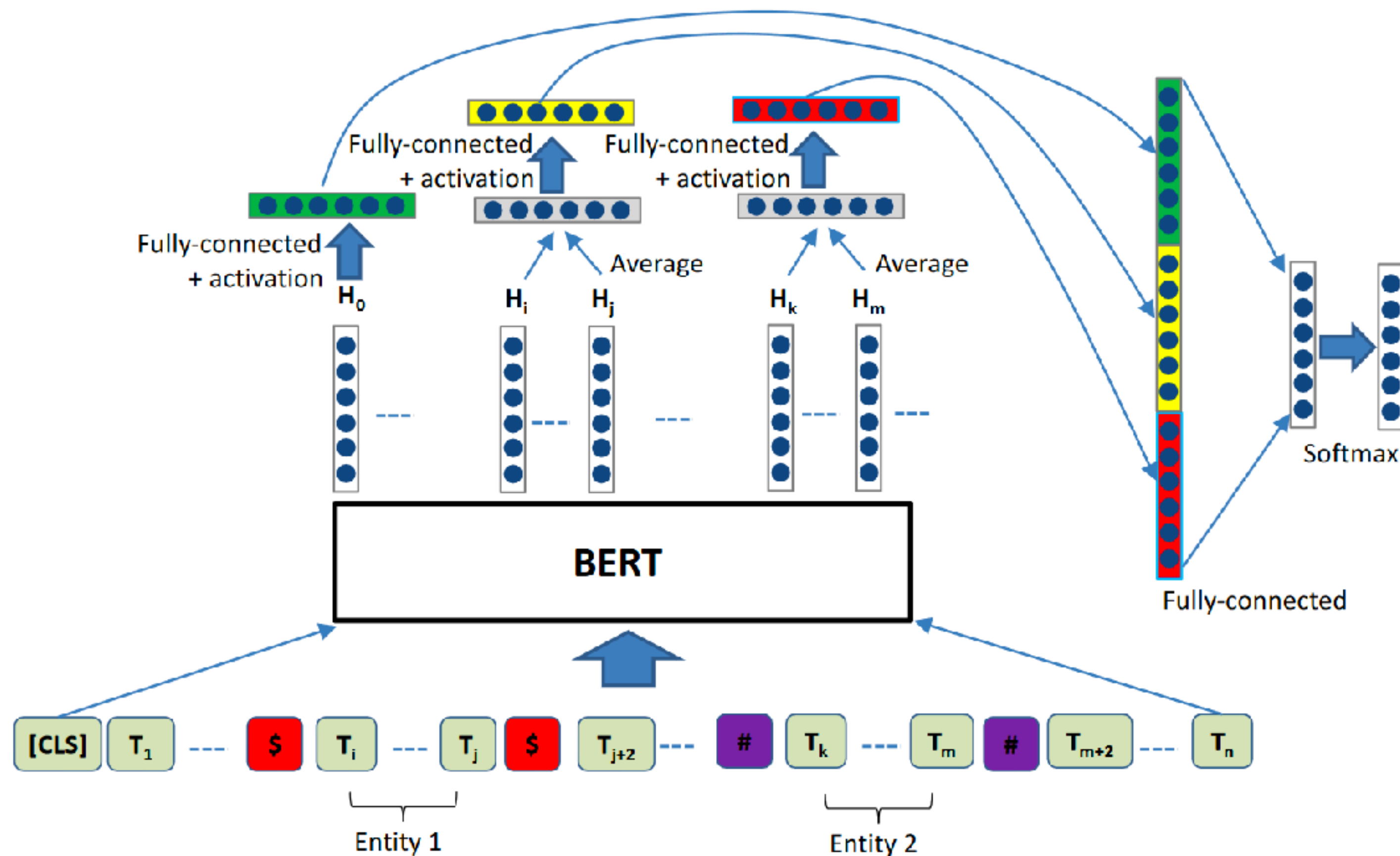


1. Get subtree representations by Recursive NN
2. Get sentence level representation by CNN through the shortest path on dependency tree from subject to object.

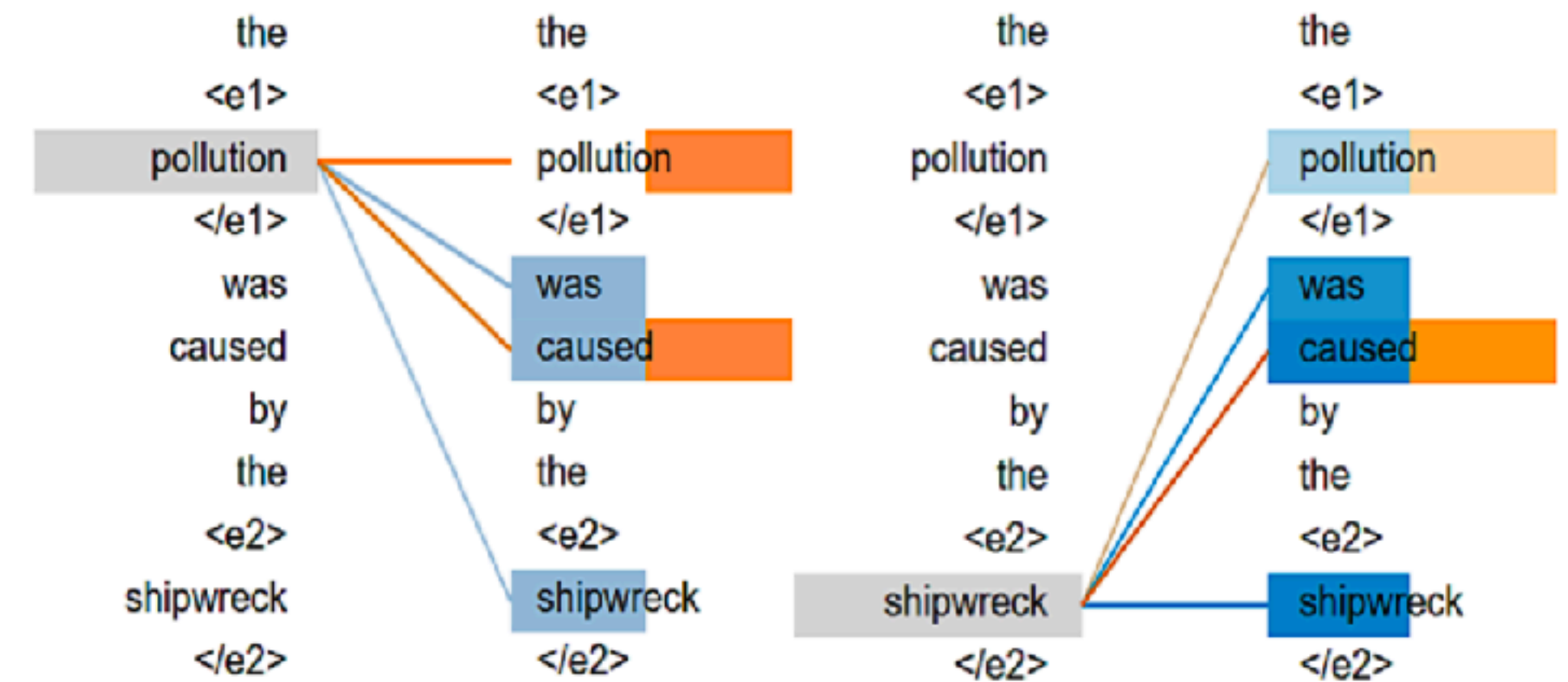


# 124 How are neural networks being used to do RE?

- © [Wu and He 2019] apply **BERT pre-trained language model** to relation extraction. (R-BERT)



## visualization of self-attention



# Supervised RE

## ⦿ Pros:

- **High accuracy**
  - At least for **some seen relations**
  - If we have **sufficient training data**

## ⦿ Cons:

- Performance is limited by the amount of **expensive** labeling samples
- Does **not generalize** well to **new domains**

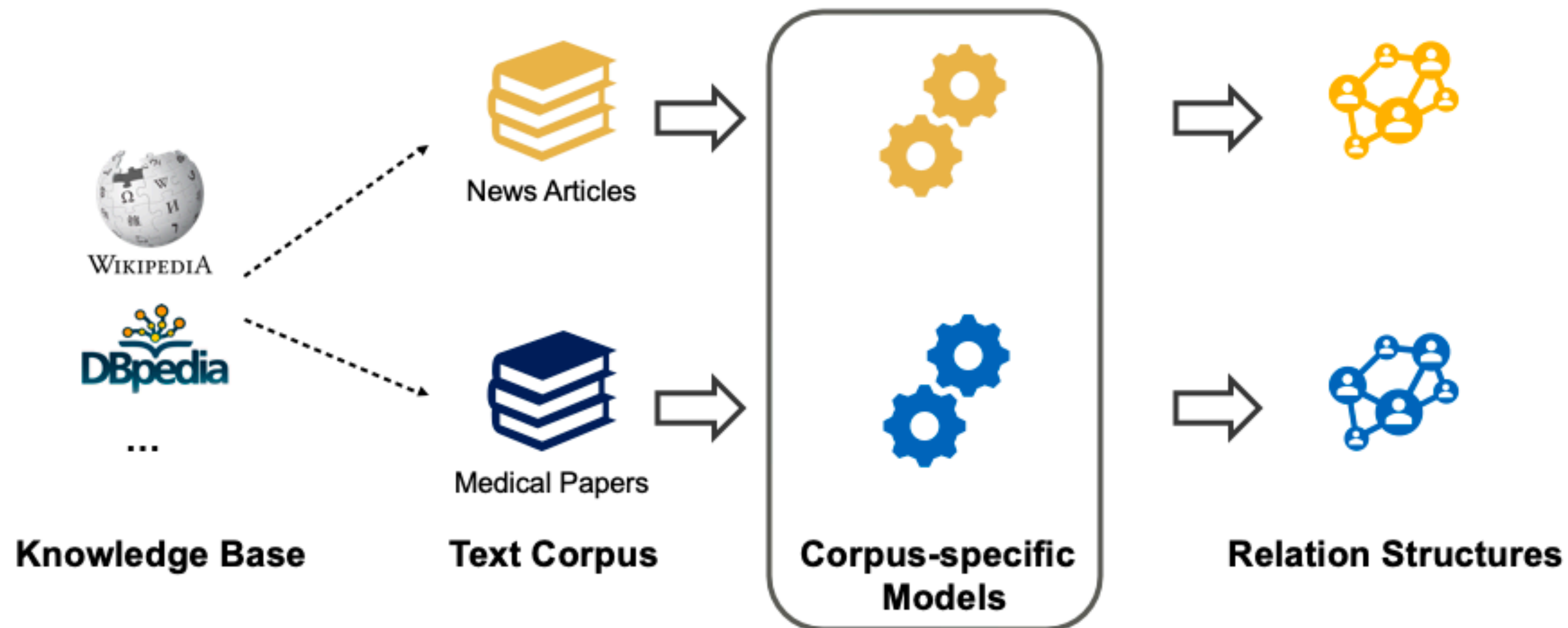
## Beyond supervised RE?

- Collect **sufficient** labeled data **at low cost**: **Distantly supervised RE**
- Get rid of the reliance on labeled data: **Unsupervised RE**

# Distantly Supervised RE

## ● Basic idea:

- Combine the idea of **using seed data** (as for Weakly Supervised RE) with **training a classifier** (as for Supervised RE).
- Instead of providing seed tuples ourselves, we can take it from KBs.





# 127 Distantly Supervised RE: Illustration

- ⦿ For each relation
  - For each tuple in KB
    - Find sentences in large corpus with both entities
    - Assume these sentences are positive examples for this relation type
  - Train a supervised classifier on this distantly-labeled dataset

## Born\_In

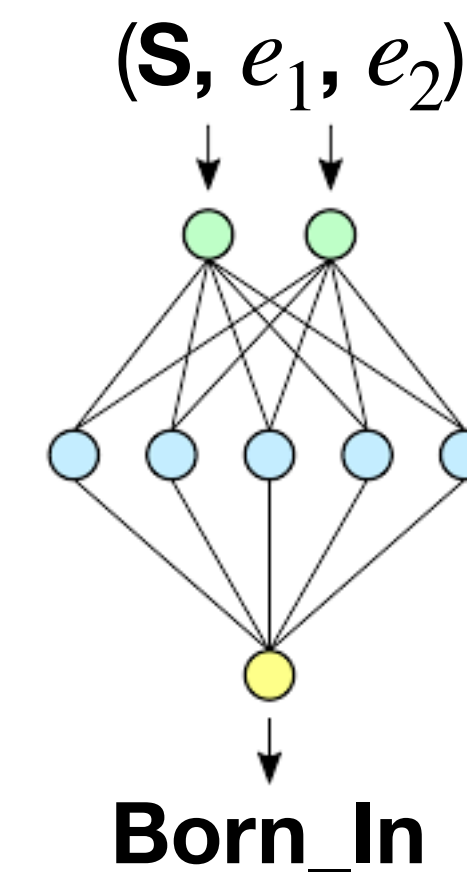
<Edwin Hubble, **Born\_In**, Marshfield>

<Albert Einstein, **Born\_In**, Ulm>

**S1: Hubble was born in Marshfield**

**S2: Einstein, born (1879), Ulm**

**S3: Hubble ... Steven Tyler ... Marshfield**



# Distantly Supervised RE

## ● Pros

- **Less manual effort**
- Can **scale** to large amount of corpus and many relations
- **No iterations required**  
(unlike Weakly Supervised RE)

## ● Cons

- Generated **labels** can be **noisy**
- Be **restricted** to the referred **KBs**

[Mintz et al. 2009] propose **distant supervision** to **reduce the cost of manually labeling corpus**.

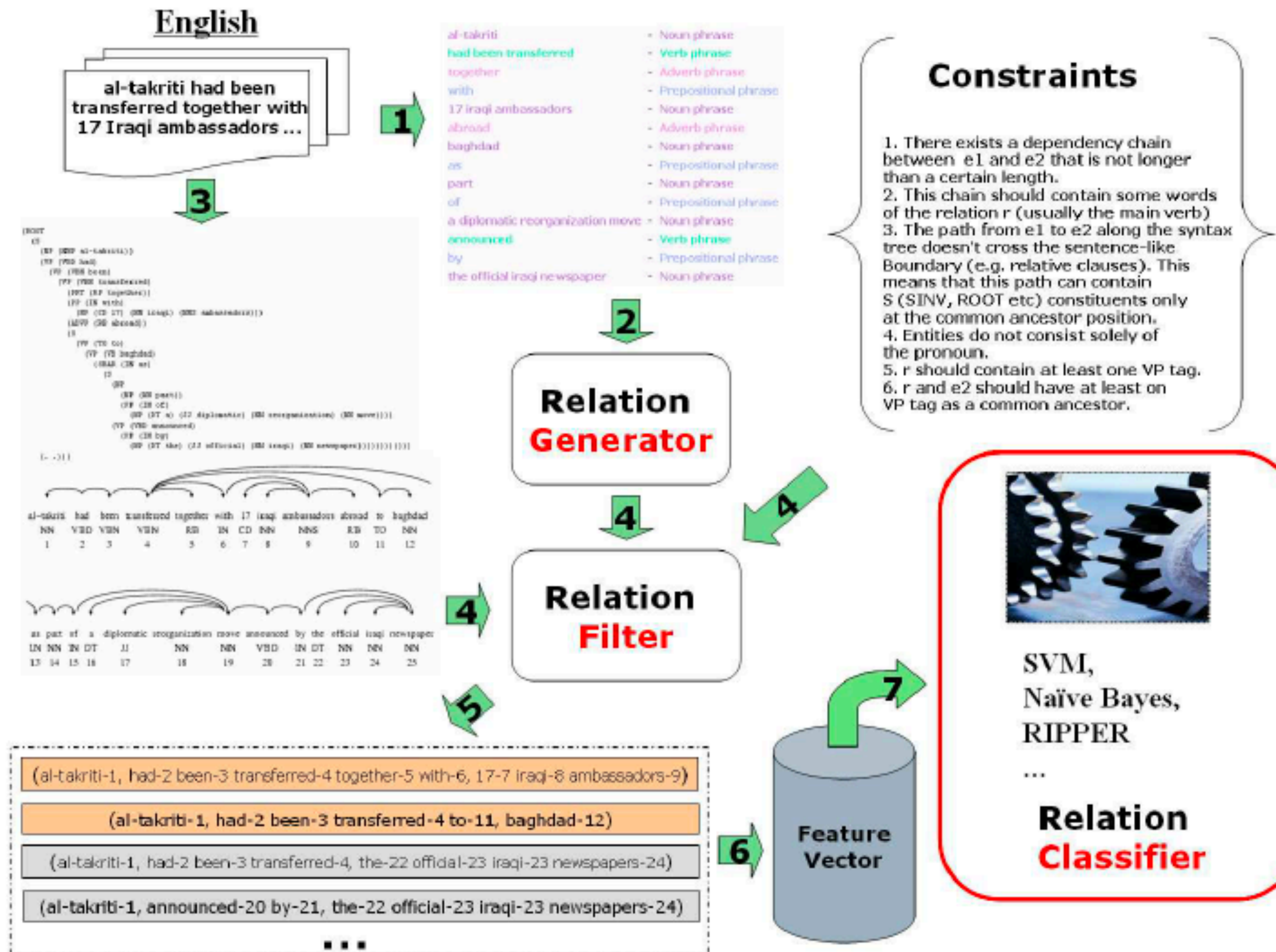
### **Distant supervision for relation extraction without labeled data**

**Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky**  
Stanford University / Stanford, CA 94305  
{mikemintz, sbills, rion, jurafsky}@cs.stanford.edu

# Unsupervised RE for Open Information Extraction

## Basic idea:

- Extract relations from text based on a set of very **general constraints and heuristics**, thus **get rid of the reliance on labeled data, seed tuples, and hand write rules.**



## TextRunner

[Bach, Nguyen, and Sameer Badaskar 2007]

Constraints are not relation-specific patterns.



# Unsupervised RE

## ◎ Influential work by Allen Institute for AI (Oren Etzioni's group)

- 2005: KnowItAll
  - Generalizes Hearst patterns to other relations
  - Requires zillions of search queries; very slow
- 2007: TextRunner
  - No predefined relations; highly scalable; imprecise
- 2011: ReVerb
  - Improves precision using simple heuristics
- 2012: Ollie
  - Operates on Stanford dependencies, not just tokens
- 2013: OpenIE 4.0
  - Use heuristics from semantic role labeling (SRL) frames
  - Extends the definition of Open IE extractions to include n-ary extractions
- 2018: OpenIE 5.1
  - Improves upon extractions from numerical sentences
  - Improves upon conjunctive sentences by breaking conjunctions in arguments to generate multiple extractions.
- 2020: OpenIE6
  - Iterative Grid Labeling and Coordination Analysis



# Unsupervised RE

## ● Pros:

- **No/Almost none labeled training data** required
- **Without** the require of manually **pre-specifying** all the **relation** of interest, it **considers all possible relation types**

## ● Cons:

- **Performance depends** a lot on how well constructed the **constraints** and **heuristics** are
- **Relations are not normalized** as pre-specified relation types



## Open Information Extraction



**Example Queries:**

- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

**Typed Example Queries:**

- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?

Argument 1:       Relation:       Argument 2:       Corpus:

To learn more about Open IE, watch our [YouTube video!](#)

Powered by [ReVerb](#), our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

[Open IE 4.0](#), the successor to [ReVerb](#) and [Ollie](#), has been released. [Download it from GitHub!](#)

### Publications:

- [Search Needs a Shake-up](#) (Nature 2011)
- [Open Information Extraction](#) (IJCAI 2011)
- [Ollie](#) (EMNLP 2012)
- [Reverb](#) (EMNLP 2011)
- [TextRunner](#) (IJCAI 2007)

### Public resources based on Open IE:

- [1B million question-paraphrases](#) (Fader et al. ACL, 2013)
- [94M Rel-grams](#) n-grams, but over relations in text (Balasubramanian et al. AKBC-WEKEX, 2012)
- [600K ReVerb relation phrases](#)
- [15 million high-precision ReVerb extractions](#) (826MB compressed)
- [3 million entity-linked ReVerb extractions](#) (Lin et al. AKBC-WEKEX, 2012)
- [Domain/range for 50K relations](#) (Ritter et al. ACL 2010)
- [10K Functional relations](#) (Lin et al. EMNLP 2010)
- [30K Horn clauses](#) learned from text (Schoenmackers et al. EMNLP 2010)
- [10M CLEAN entailment rules](#) (Berant et al. ACL 2012)
- [Code for extracting sentences from ClueWeb](#)

# Relation Extraction: Tools

## ◎ Domain-dependent:

- spaCy Kindred (Biomedical)
- OpenNRE (NYT10m and Wiki20m)
- ...

## ◎ Domain-independent:

- Stanford OpenIE
  - About 4x faster than AllenAI OpenIE, while brings lower performance.
- AllenAI OpenIE 5.1
  - Utilize pre-trained language model
  - Higher performance, with higher computational cost.
- ...





## Open Information Extraction

Hosted by



Created at



### Example Queries: ⓘ

What kills bacteria?

Who built the Pyramids?

What did Thomas Edison invent?

What contains antioxidants?

### Typed Example Queries: ⓘ

What countries are located in Africa?

What actors starred in which films?

What is the symbol of which country?

What foods are grown in which countries?

What drug ingredients has the FDA approved?

Argument 1:

Relation:

Argument 2:

Corpus:



To learn more about Open IE, watch our [YouTube video!](#)

Powered by [ReVerb](#), our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

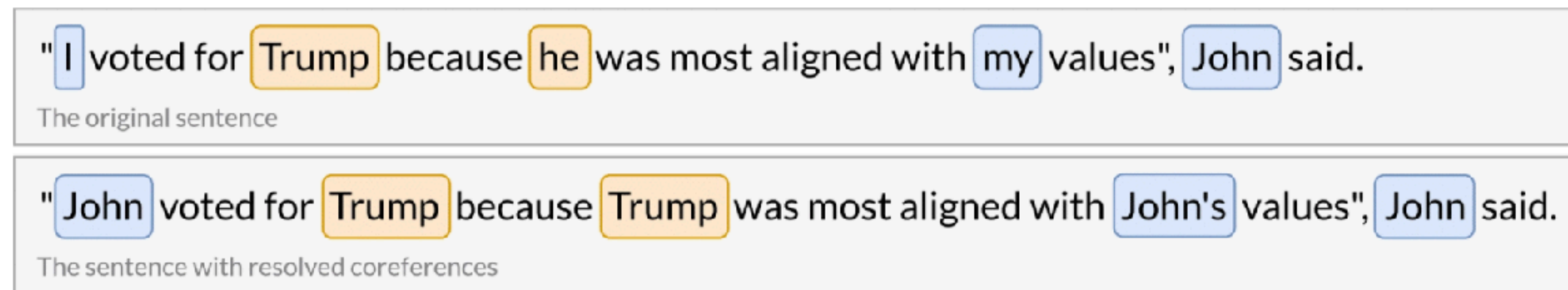
[Open IE 4.0](#), the successor to [ReVerb](#) and [Ollie](#), has been released. [Download it from GitHub!](#)

<https://openie.allenai.org/>

# Coreference Resolution

## What is Coreference Resolution?

- Coreference Resolution is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity.

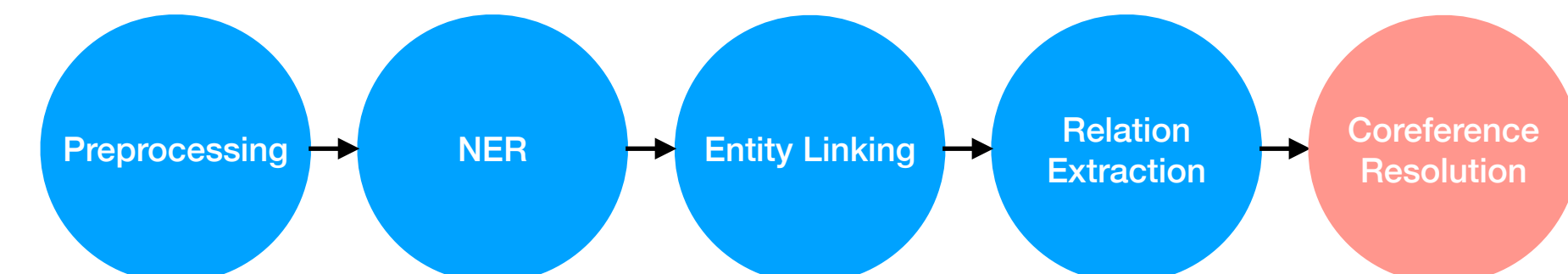


## Why Coreference Resolution?

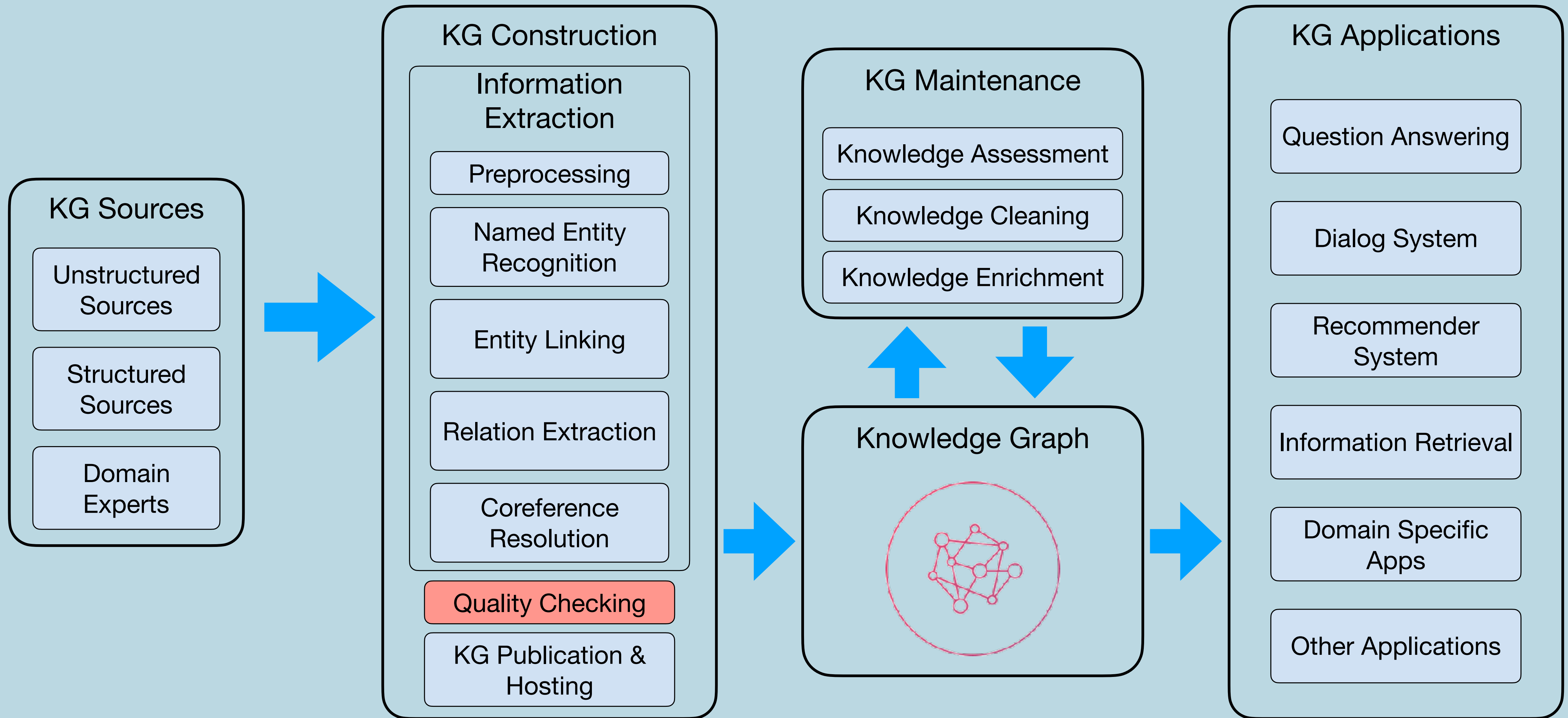
- Coreference resolution could increase the overall efficiency and the coverage of the information extraction.

## Tools:

- spaCy·neuralcoref, NLTK·coref, BART, ...



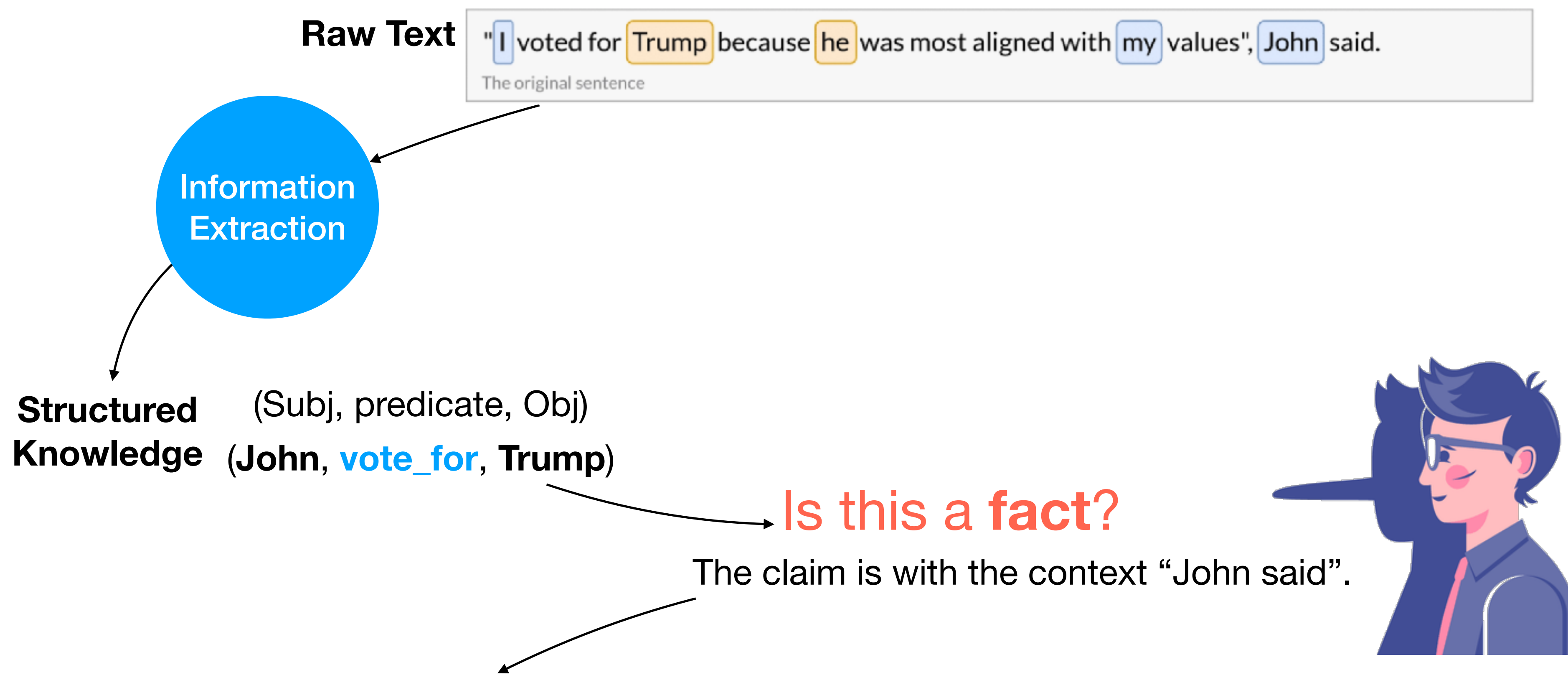
# Quality Checking





# Quality Checking

- Not everything accurately extracted is **fact**



**We need to further check the quality of the extracted knowledge.**

# Quality Checking

- ◎ Not everything accurately extracted is **fact**
- ◎ Many recent efforts on assessing truth and finding supports
  - Knowledge-based Trust [Dong et.al 2015]
  - Multilingual answer validation [Rodrigo et.al2009] [Kobayashi et.al 2017]
  - FactChecker [Nakashole & Mitchell 2014]
  - PolitiFact [Vlachos & Riedel 2014], [Wang 2017]
  - Fake News challenge [Pomerleau & Rao 2017]
  - Fake news detection via crowd signals [Tschiatschek et.al 2018]
  - Fact Verification competition [Thorne et.al 2018]

# Quality Checking

- Fact Verification competition (**FEVER**) [[Thorne et.al 2018](#)]
- **Goal:**
  - Given a claim, label claim **SUPPORTS**, **REFUTES**, or **NOT-ENOUGH-INFO**
  - For the first two classes, **select relevant sentences** from Wikipedia intro sections.
- **Largest annotated fact sets**
  - 185,445 annotated claims.
  - Claims generated by mutating Wikipedia sentences: paraphrasing, negation, substitution of entity/relation, generalize/specialize claims

**Claim:** The Rodney King riots took place in the most populous county in the USA.

**[wiki/Los Angeles Riots]**

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

**[wiki/Los Angeles County]**

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

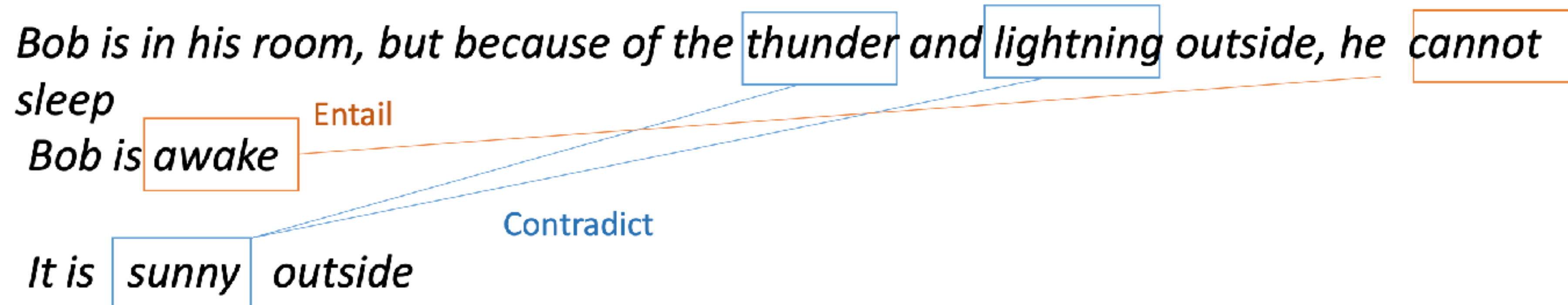
**Verdict:** Supported



# Quality Checking

◎ FEVER baseline – sentence classification [Thorne et.al 2018]

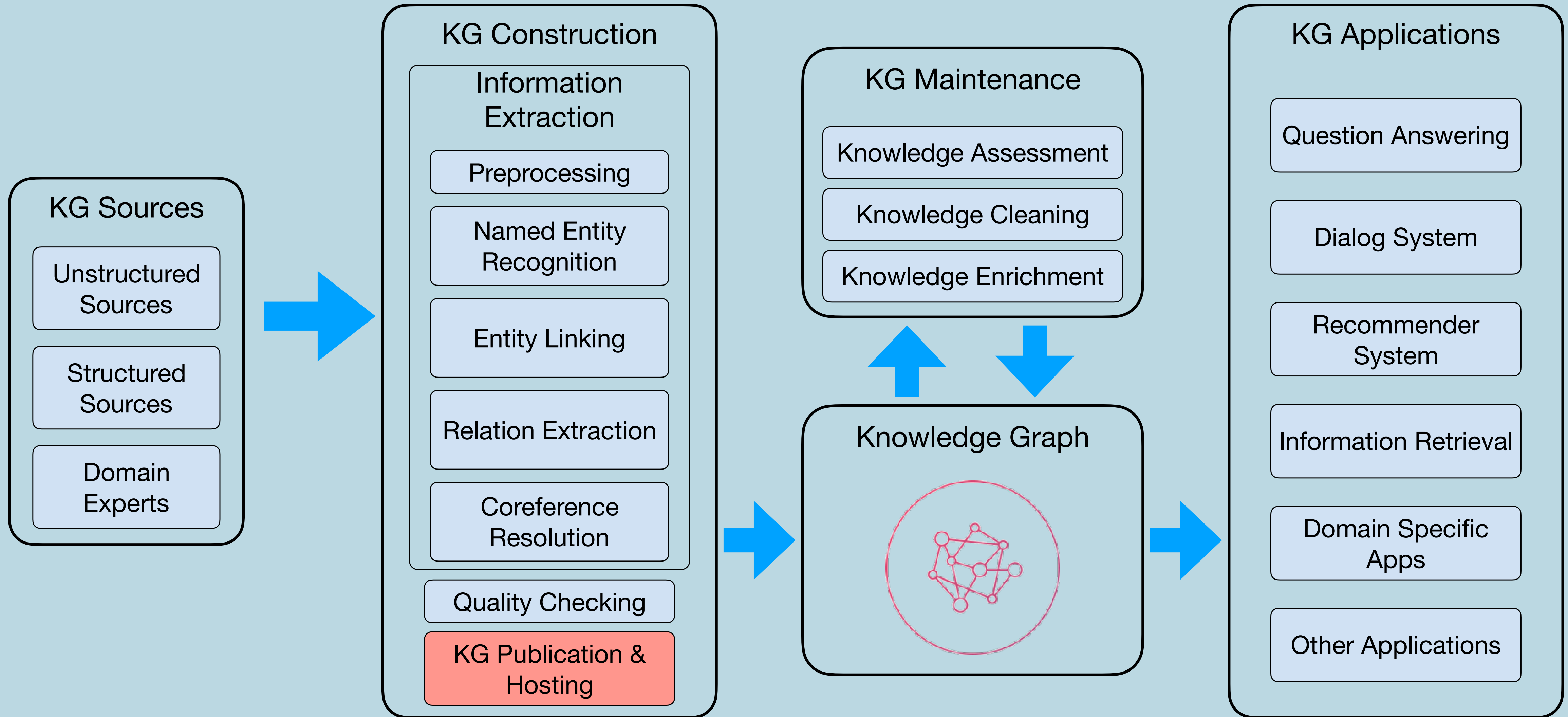
- Basic idea: **align parts** of the text in sentences **a** and **b** and then aggregate info to predict the label
- Example



## ● FEVER Progress

Final Leaderboard		2021 Shared Task				
Rank	Team	FEVEROUS score	Accuracy	Evidence F1	Evidence Precision	Evidence Recall
1	Bust a move!	0.2701	0.5607	0.1308	0.0773	0.4258
2	Papelo	0.2592	0.5757	0.1187	0.0716	0.3460
3	NCU	0.2514	0.5229	0.1581	0.0991	0.3907
4	Z team	0.2251	0.4901	0.1312	0.0776	0.4264
5	EURECOM_Fever	0.2001	0.4779	0.1952	0.1373	0.3373
6	FEVEROUS Baseline	0.1773	0.4548	0.1503	0.1017	0.2878
7	Saturday_Night_Fever	0.1763	0.4804	0.1618	0.1122	0.2900
8	Martin Funkquist	0.1261	0.4302	0.1045	0.0642	0.2789
9	Albatross	0.1159	0.4035	0.0963	0.0644	0.1902
10	METUIS	0.0636	0.3897	0.0634	0.0462	0.1011
11	ChaCha	0.0389	0.4194	0.0398	0.0251	0.0969
12	seda_kaist	0.0362	0.4140	0.0384	0.0242	0.0920
13	qmul_uou_iiith	0.0223	0.3999	0.0282	0.0245	0.0330

# 141 KG Hosting



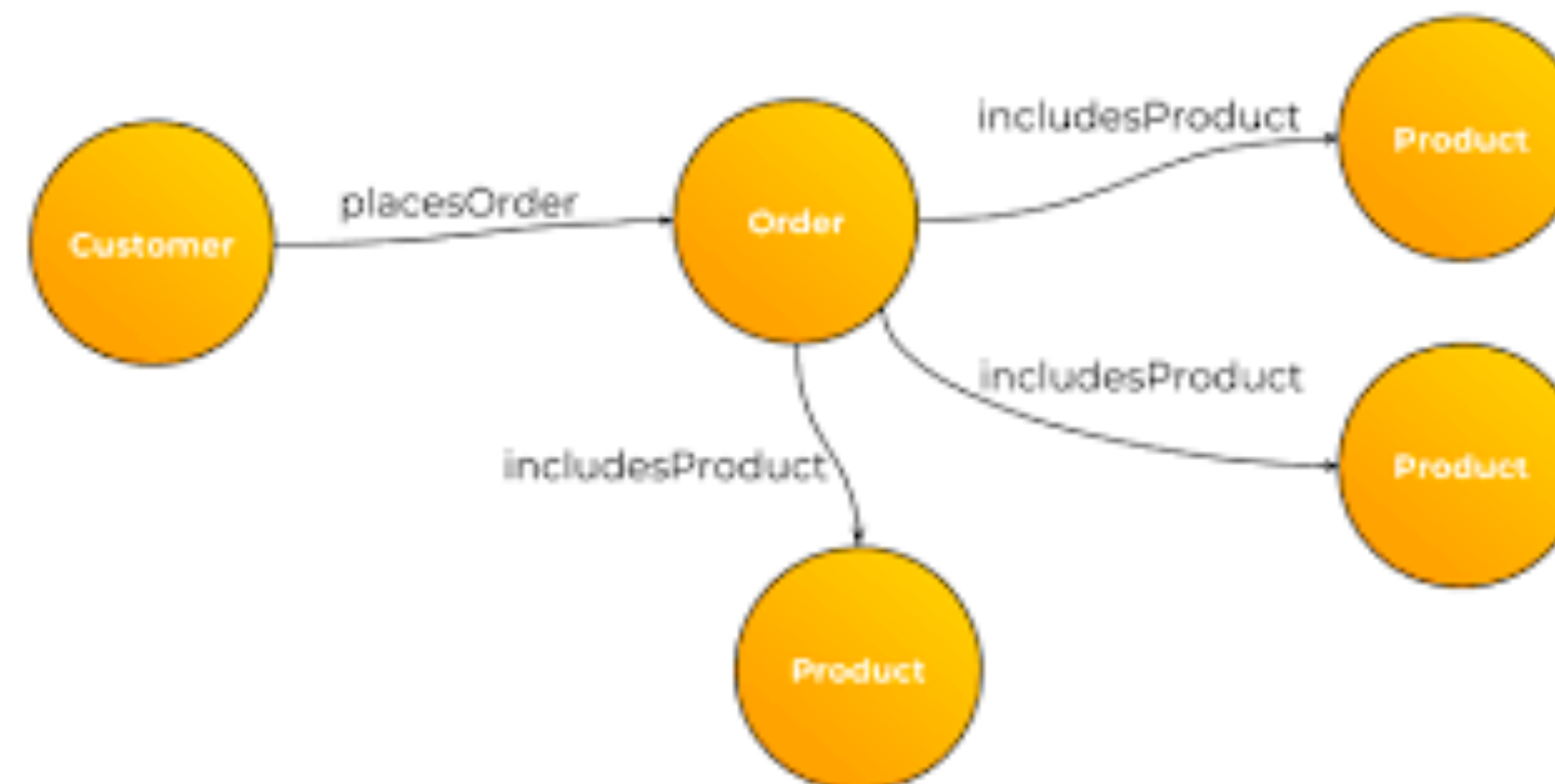


# 142 KG Hosting

- After checking the quality of extracted knowledge, we can store the **RDF triples** as **knowledge graph** by:
  - Storing them into **RDF Triple Stores**. (<https://www.w3.org/wiki/LargeTripleStores>)

#	Name	# triples tested with
1	Oracle Spatial and Graph with Oracle Database 12c	1.08 T
2	AnzoGraph DB by Cambridge Semantics	1.065 T
3	AllegroGraph	1+ T
4	Stardog	50 B
5	OpenLink Virtuoso v7+	39.8 B
6	GraphDB™ by Ontotext	17 B

- Or converting them to **nodes** and **edges** and then storing them into **Graph Databases**.



## ● Graph Databases vs. RDF Triple Stores

### ● How They Are Similar

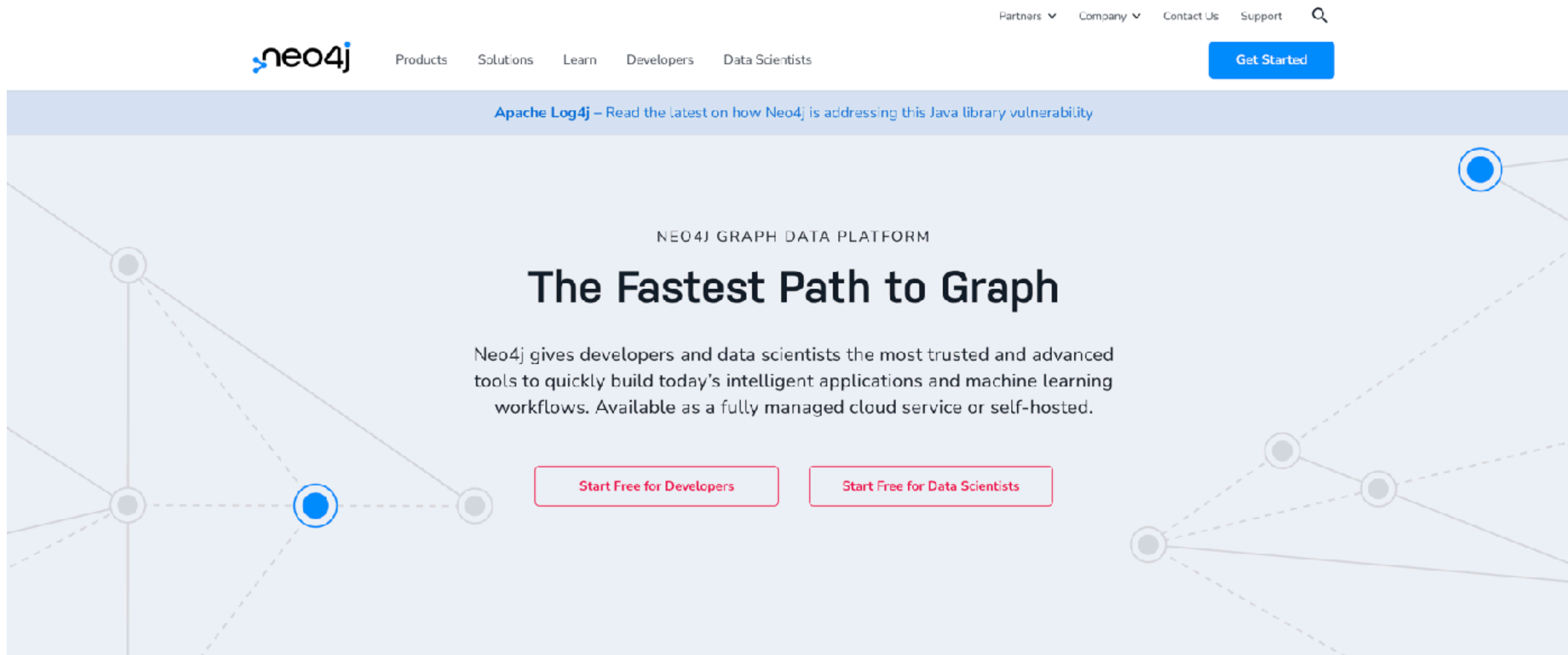
- Both two focus on the **relationships between the data**, often referred to as “**linked data**”.

### ● How They Are Different

- Graph databases are more **versatile with query languages**: **Neo4J** can run an RDF triple store and use SPARQL but generally focuses on its own proprietary language, **Cypher**; RDF triple stores **only use SPARQL** as the query language.
- Graph databases can store **various types** of graphs; RDF triple stores focus solely on storing rows of **RDF triples**.
- Graph databases are **node centric**, whereas RDF triple stores are **edge-centric**.
- Graph databases are **better optimized for graph traversals**, while the cost of traversing an edge tends to be **logarithmic** with RDF triple stores.
- Graph databases do **not provide inferences on data**, but RDF triple stores do.

# KG Hosting

- **Neo4j**, a graph database for high-performance graph queries on large datasets. More on <https://neo4j.com/>



The screenshot shows the Neo4j website homepage. At the top, there is a navigation bar with the Neo4j logo on the left and links for Partners, Company, Contact Us, and Support on the right. A search icon is also present. Below the navigation bar, there is a blue banner with the text "Apache Log4j – Read the latest on how Neo4j is addressing this Java library vulnerability". The main content area features the heading "NEO4J GRAPH DATA PLATFORM" and the sub-heading "The Fastest Path to Graph". Below this, a paragraph states: "Neo4j gives developers and data scientists the most trusted and advanced tools to quickly build today's intelligent applications and machine learning workflows. Available as a fully managed cloud service or self-hosted." At the bottom of the main content area, there are two red buttons: "Start Free for Developers" and "Start Free for Data Scientists". The background of the main content area is a light blue color with a faint graph structure consisting of nodes and edges.



## Storing as Knowledge Graph:

- Query:

```
1 MATCH p=shortestPath((n {name:"SARS-CoV-2"})-[*1..3]-(m:Disease))
2 RETURN p LIMIT 10;
```

- Cypher

<https://neo4j.com/docs/cypher-manual/current/>

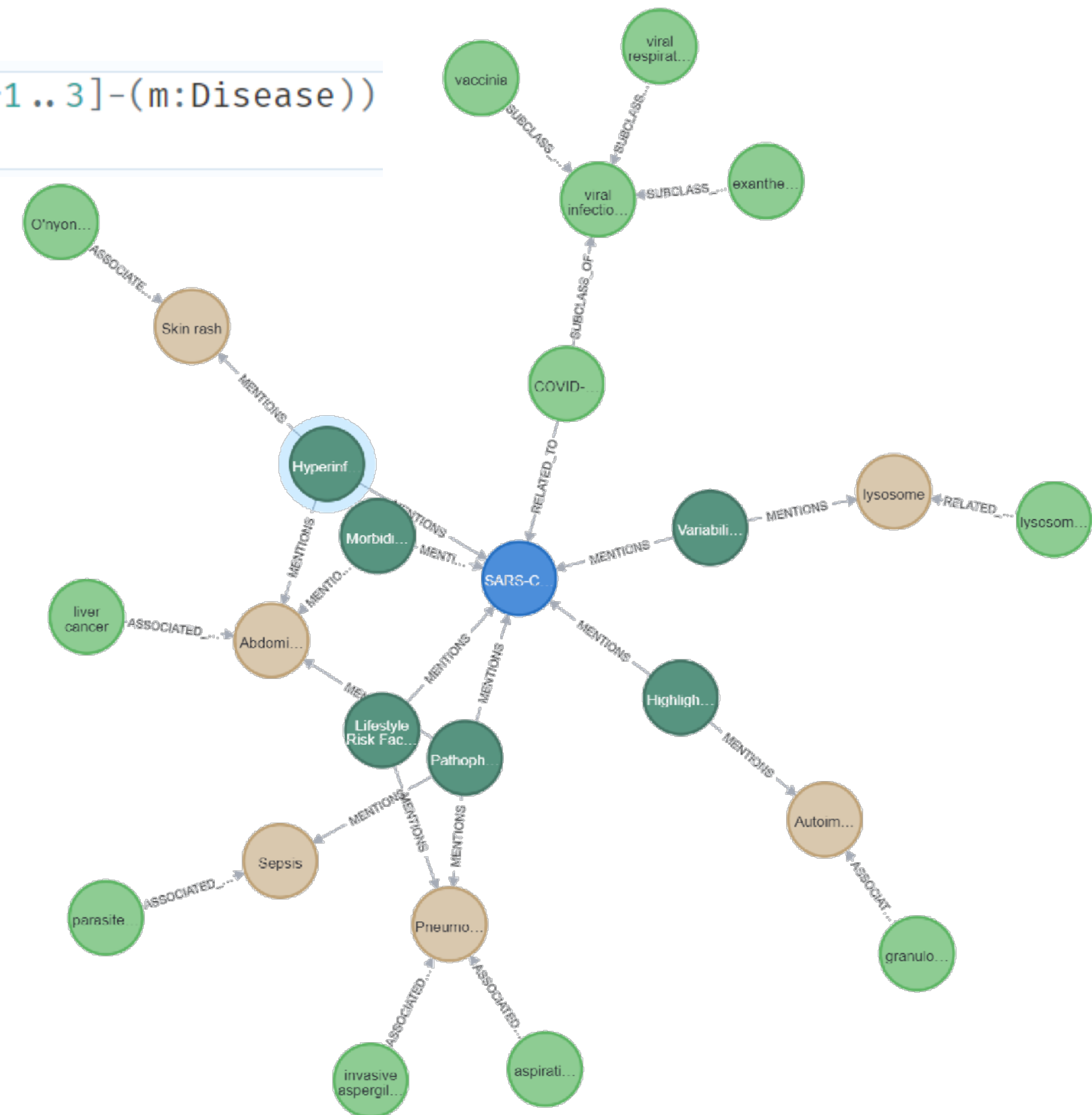
- Python Driver

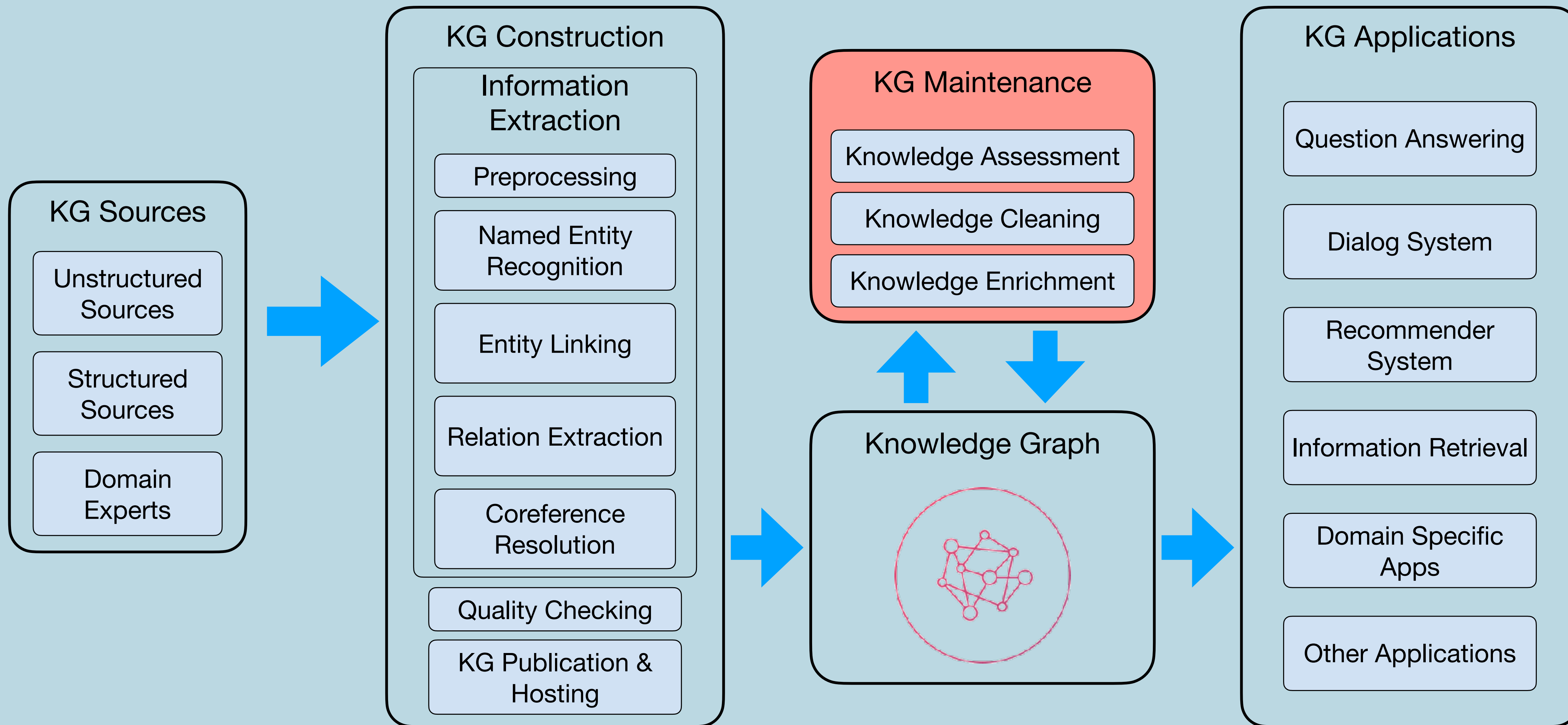
<https://neo4j.com/developer/python/>

- Others

- Summary:

high performance graph store,  
multiple programming language driver  
supported robust graph database





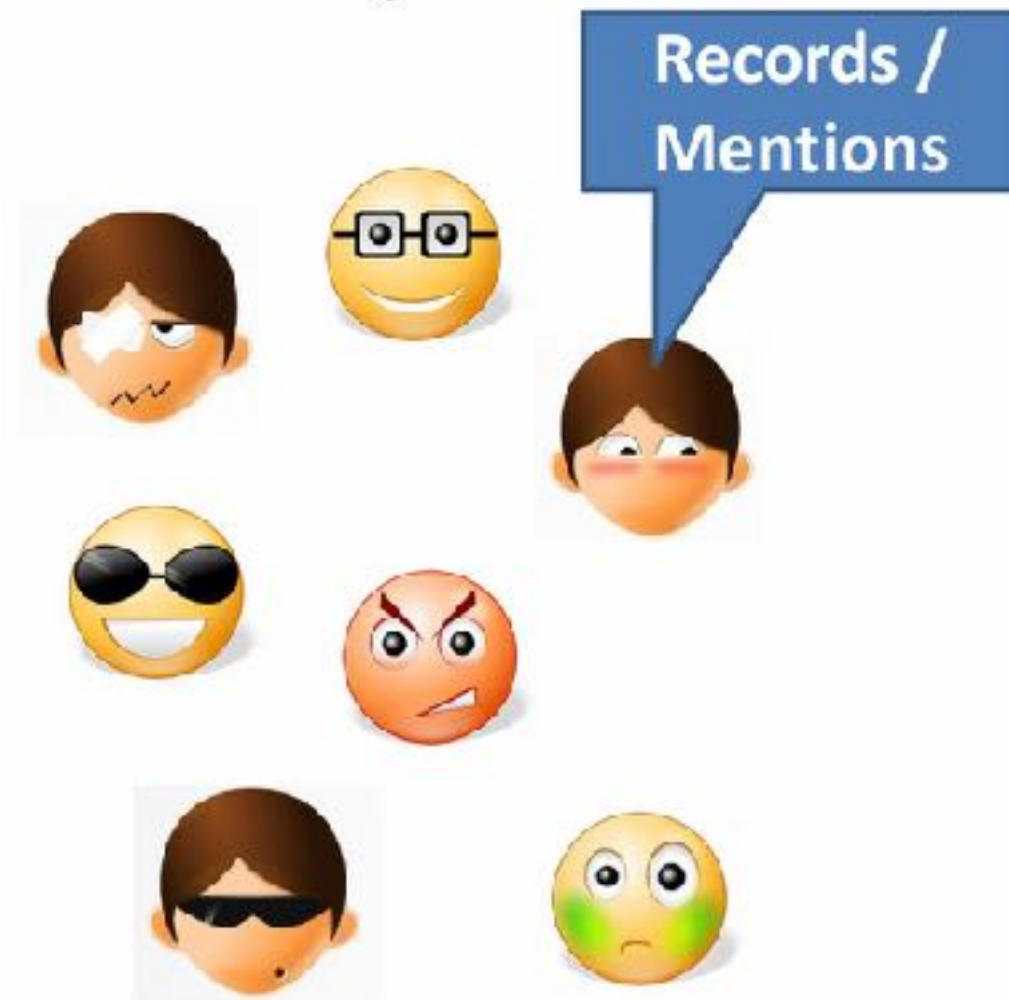
# KG Maintenance

- ◎ The extracted facts can be **incomplete**, **conflicting**, and **incorrect**:
  - Missing Data
  - Data error due to IE tech or human errors
  - Abbreviations and truncation
  - ...

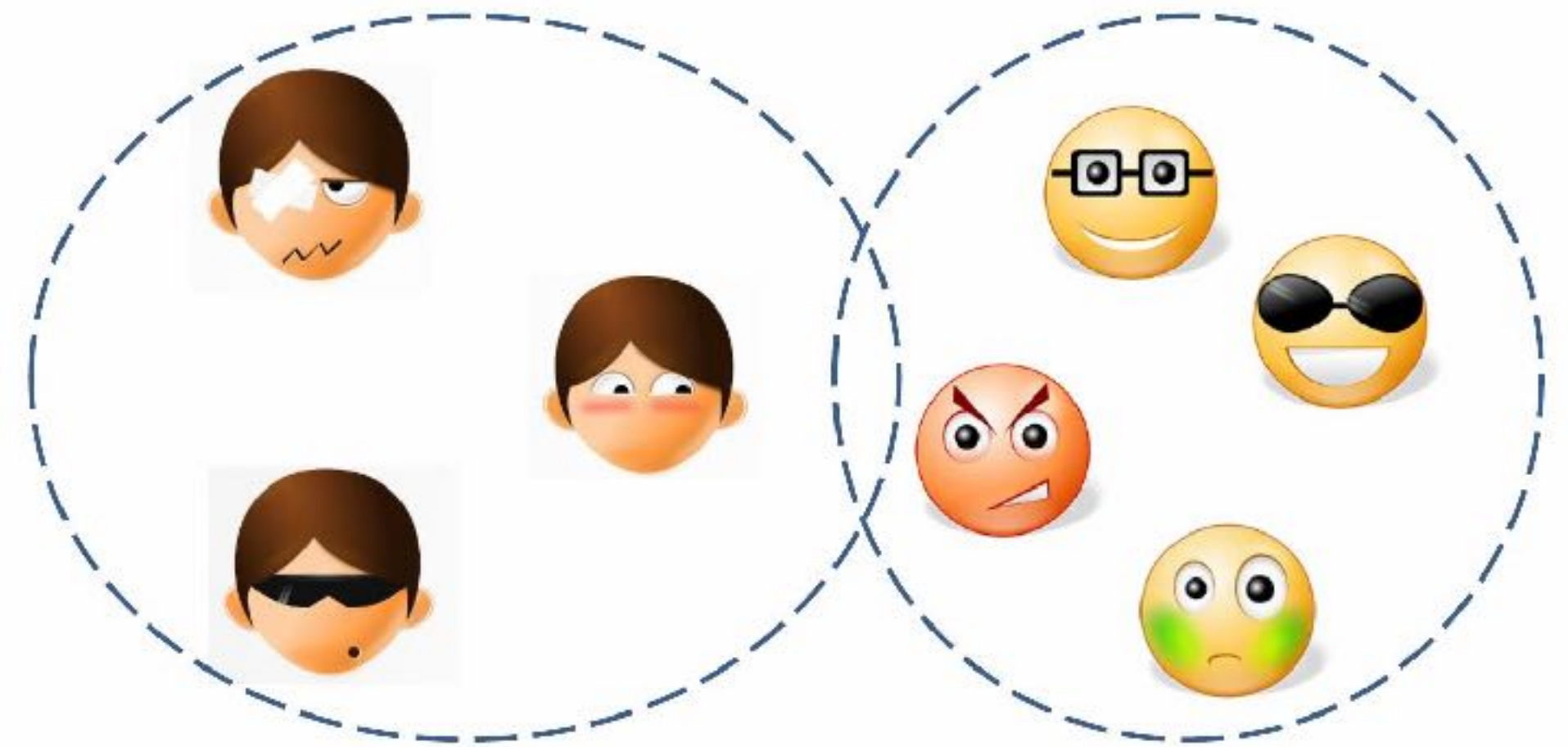
Real World



Digital World

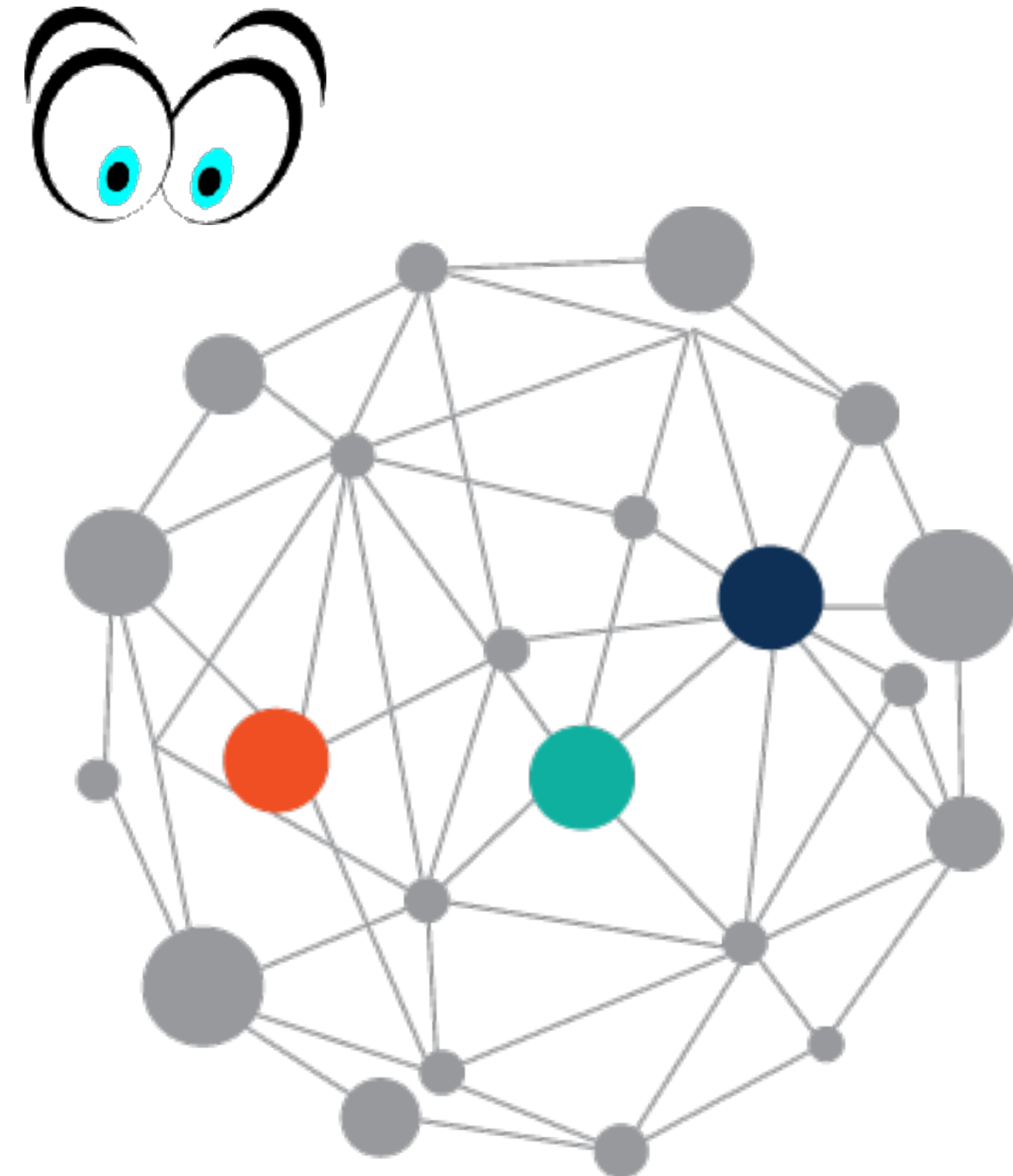


Mentions in digital world is biased and various.  
We need to de-bias and handle the diversity.



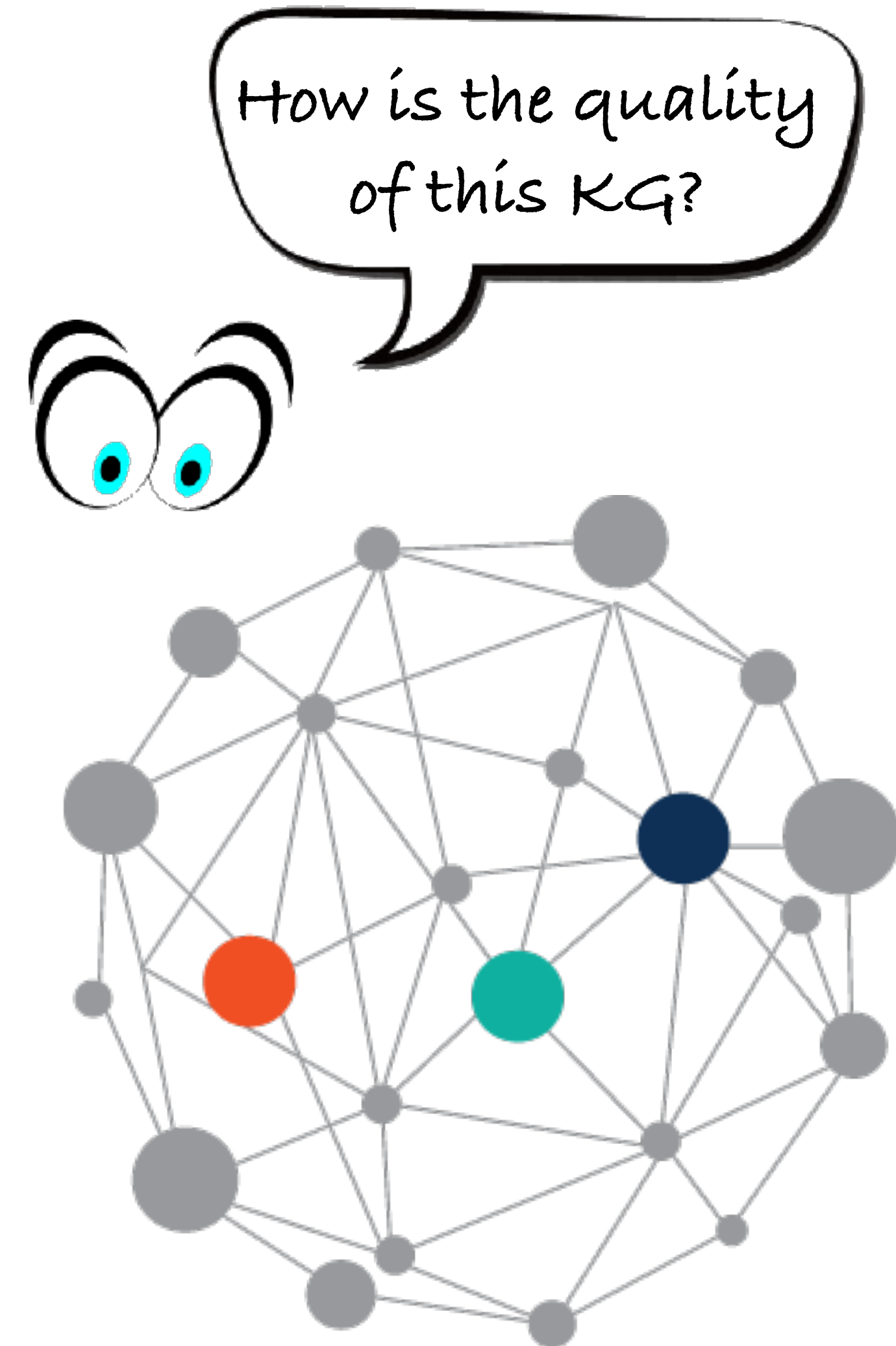


- Knowledge **Assessment**
- Knowledge **Cleaning**
  - Error detection
  - Error correction
- Knowledge **Enrichment**
  - Identifying new knowledge sources
  - Data integration



# KG Assessment

- First step to improve the quality of a KG: **assess the situation**
- Closely related to **data quality** literature
- Various dimensions for data quality assessment introduced [Batini & Scannapieco, 2006], [Färber et al., 2018], [Pipino et al., 2002], [Wang, 1998], [Wang & Strong, 1996], [Cao et al., 2001], [Zaveri et al., 2016])

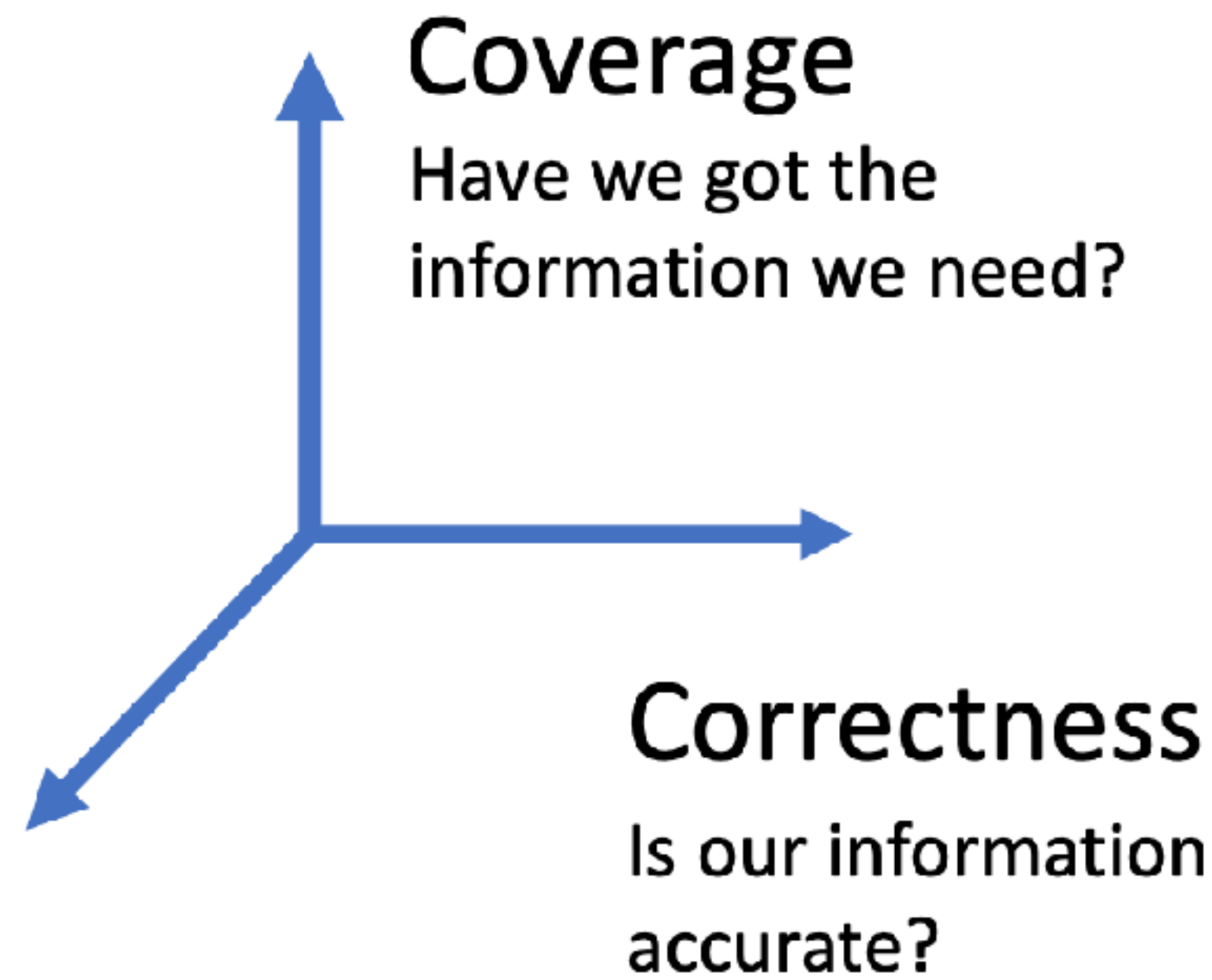


1. accessibility
  2. accuracy (veracity)
  3. appropriate amount
  4. believability
  5. completeness
  6. concise representation
  7. consistent representation
  8. cost-effectiveness
  9. easy of manipulating
  10. easy of operation
  11. easy of understanding
  12. flexibility
  13. free-of-error
  14. interoperability
  15. objectivity
  16. relevancy
  17. reputation
  18. security
  19. timeliness (velocity)
  20. traceability
  21. understandability
  22. value-added
  23. variety
  24. fitness for use
- .....

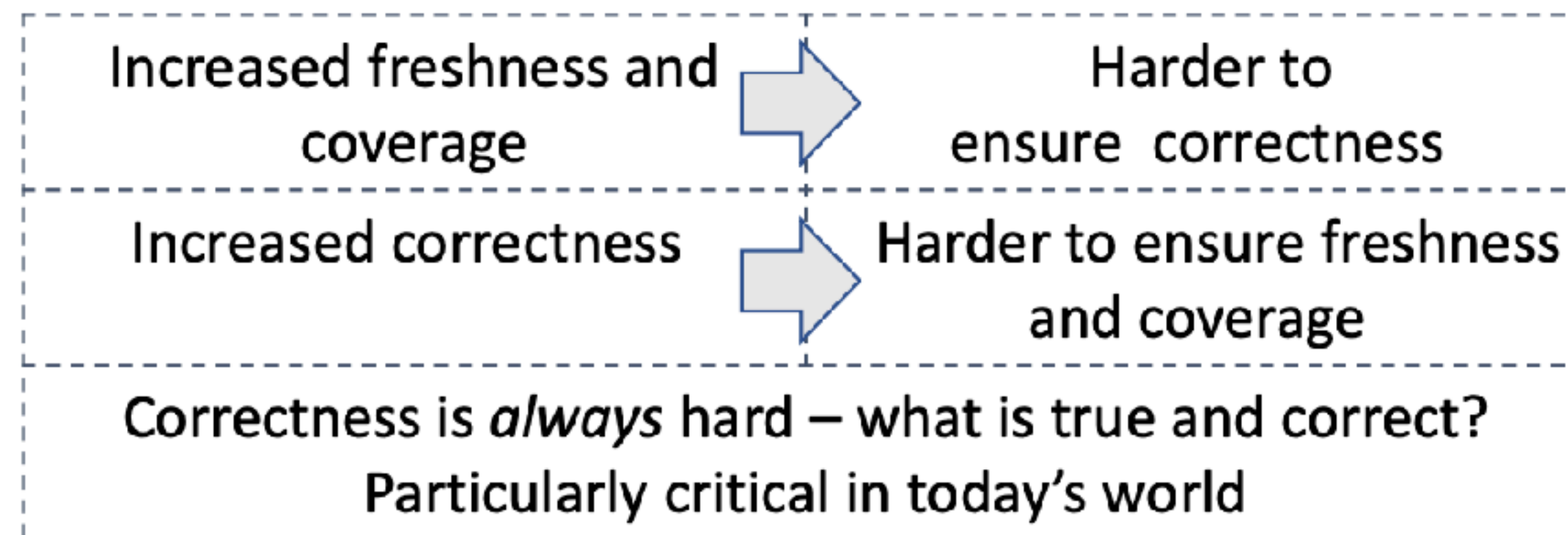




# 151 Several core assessment dimensions for KG



## Three forces in constant conflict:



**Will Smith: Single entity, 108K facts assembled from 41 web sites. There are 200 Will Smiths on Wikipedia alone.**

The collage shows multiple entries for 'Will Smith' from different sources, illustrating the fragmentation of information. It includes a Wikipedia entry for the actor, an IMDb entry for a baseball player, a Twitter profile, and a Wikipedia entry for a comedian. Each entry contains a small image and a brief bio, demonstrating how a single name can refer to multiple distinct entities.

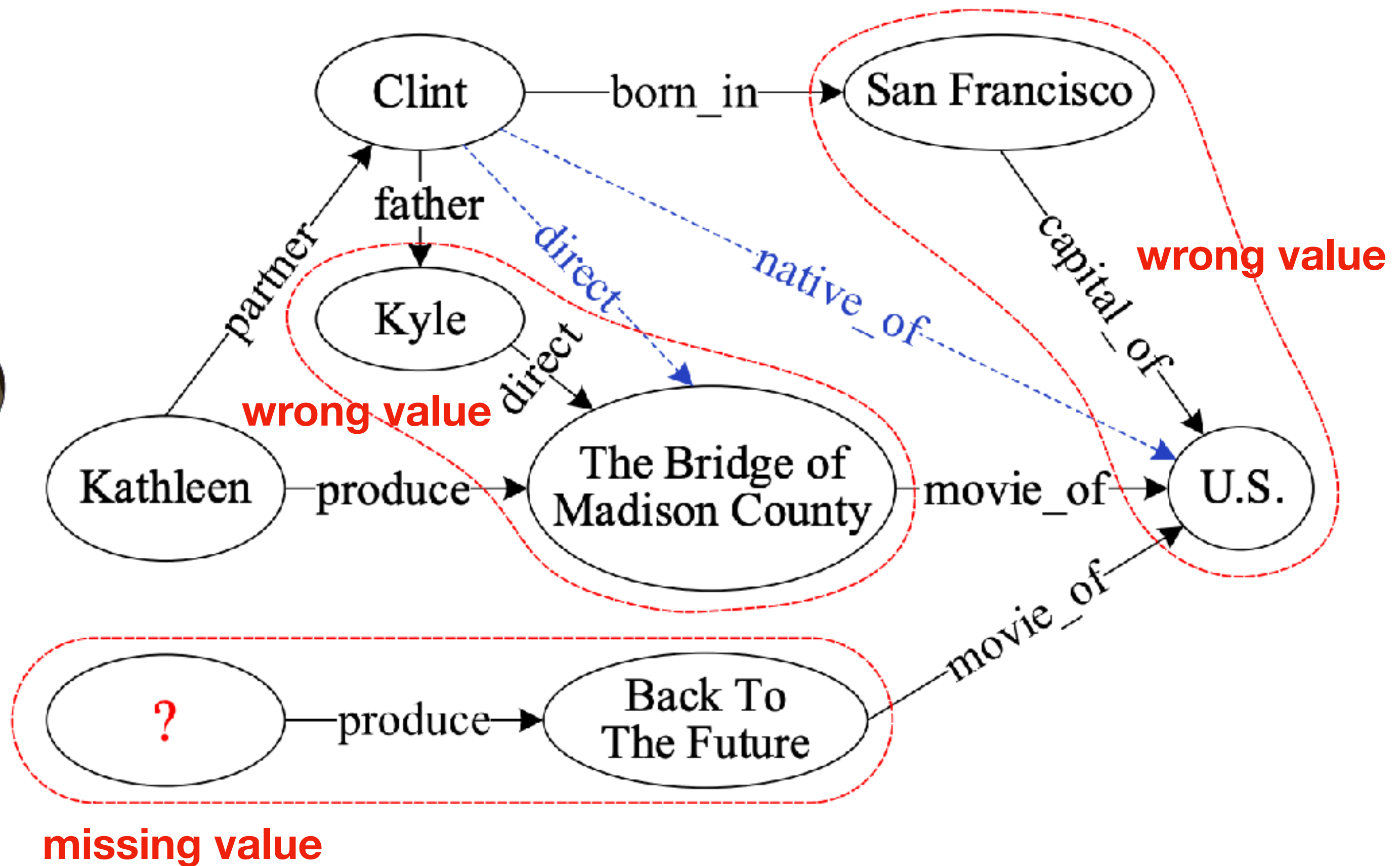
# Knowledge Assessment Tools

- LINK-QA [[Guéret et al., 2012](#)]
  - Benefits from network features to assess data quality (e.g. counting open chains to find wrongly asserted isSameAs relationships)
- Sieve [[Mendes et al., 2012](#)]
  - Uses data quality indicators, scoring functions and assessment metrics
- SDType [[Paulheim & Bizer, 2013](#)]
  - Uses statistical distributions to predict the types of instances. Incoming and outgoing properties are used as indicators for the types of resources.
- RDFUnit [[Kontokostas et al., 2014](#)]
  - A framework that assesses linked data quality based on test cases defined in various ways (e.g. RDFS/OWL axioms can be converted into constraints)
- Validata [[Hansen et al., 2015](#)]
  - An online tool check the conformance of RDF graphs against ShEx (Shape Expressions)
- Luzzu [[Debattista et al., 2016](#)]
  - Allows declarative definitions of quality metrics and produces machine-readable assessment reports based on Dataset Quality Vocabulary



# Knowledge Cleaning

- The actions taken to **improve the correctness** of a knowledge graph.
- Two major steps:
  - Error **detection**
  - Error **correction**



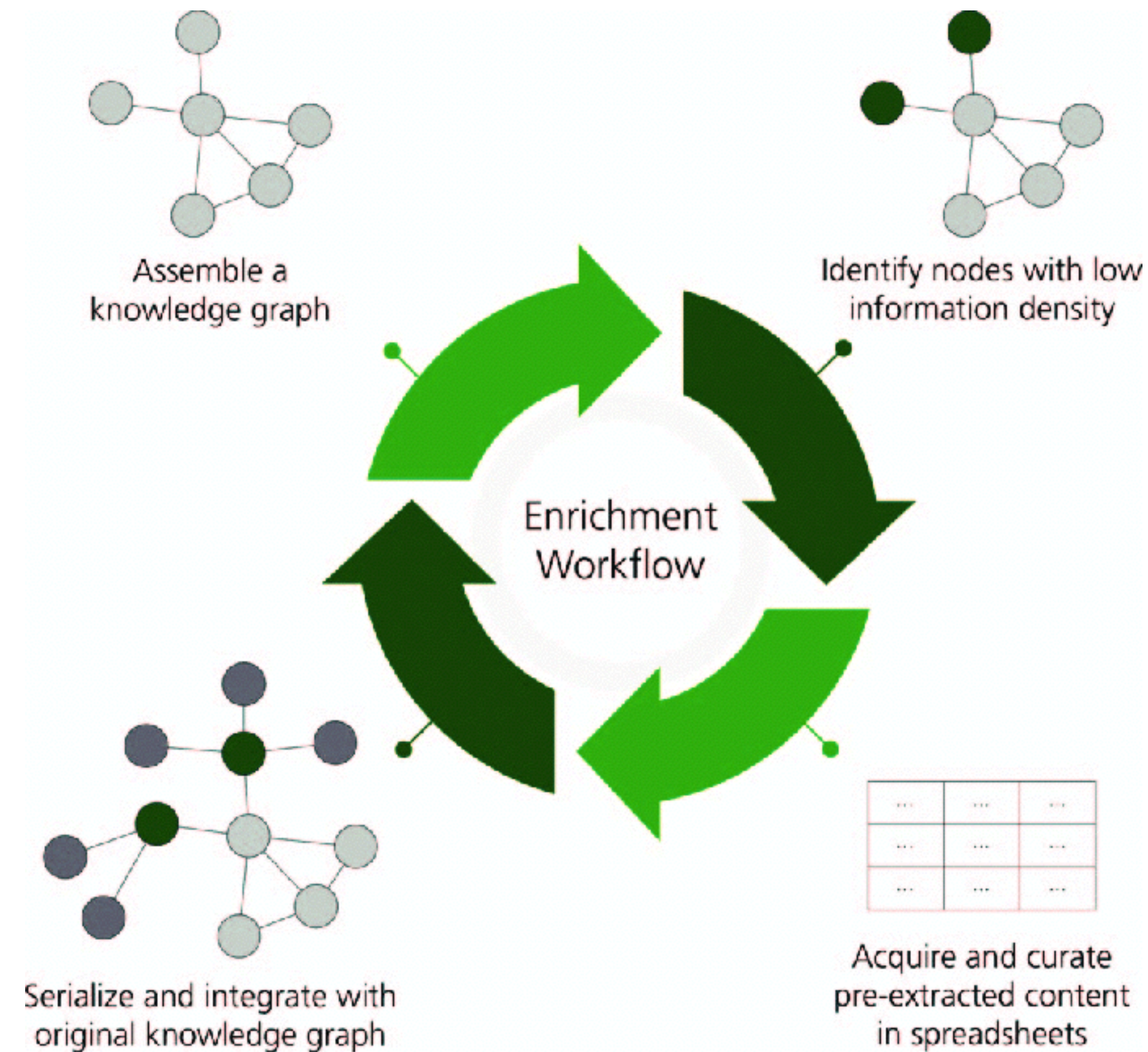


# Knowledge Cleaning Tools

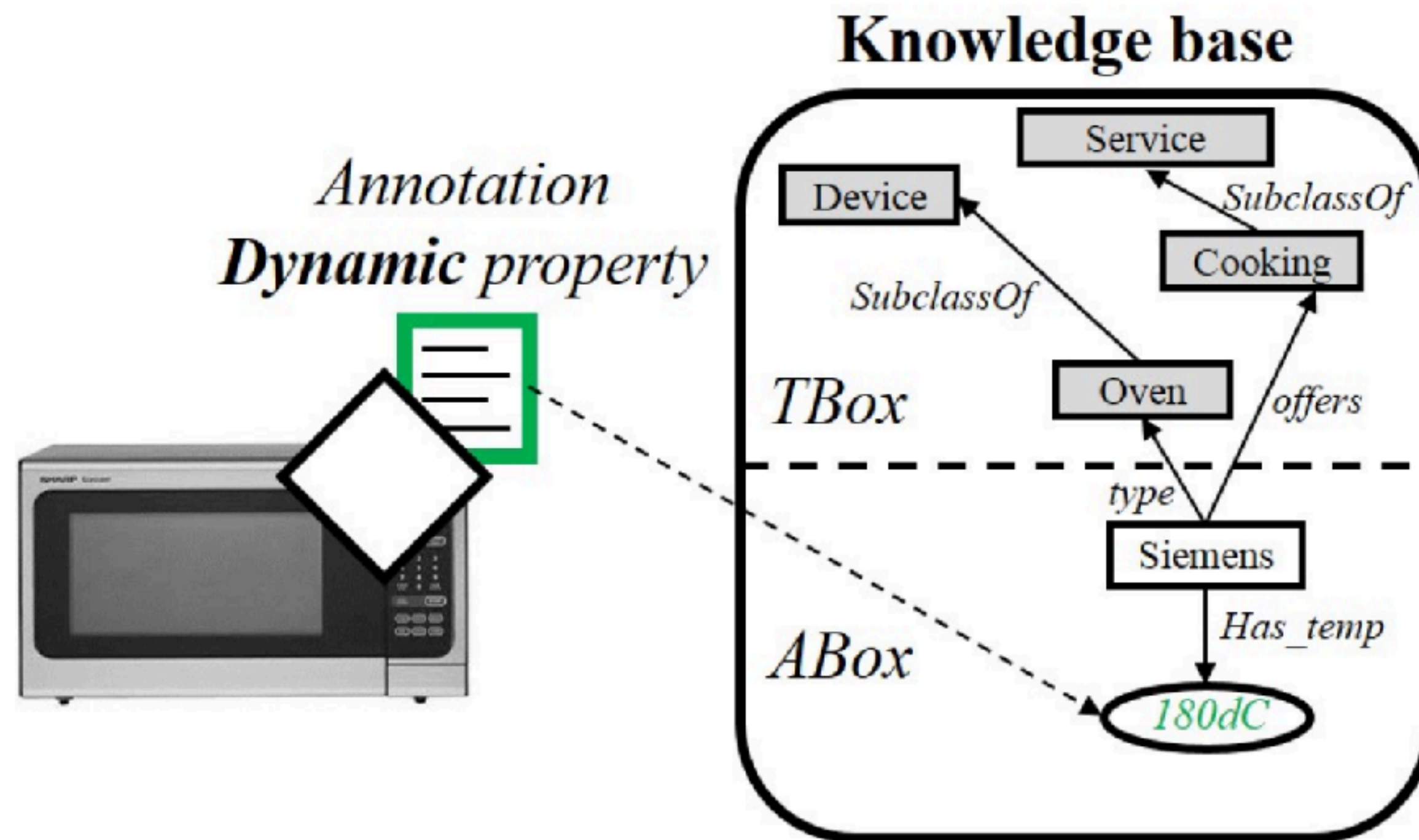
- ◎ DValidate [Paulheim & Bizer, 2014]
  - Uses statistical distribution to detect erroneous statements that connect two resources. The statements with less frequent predicate-object pairs are selected as candidates for being wrong.
  
- ◎ KATARA [Chu et al., 2015]
  - Learns the relationships between data columns and validate the learn patterns with the help of existing Knowledge Bases and crowd, in order to detect errors in the data. Afterwards it also suggests possible repairs.
  
- ◎ HoloClean [Rekatsinas et al., 2017]
  - An error detection and correction tool based on integrity constraints to identify conflicting and invalid values, external information to support the constraints, and quantitative statistics to detect outliers.
  
- ◎ TISCO [Rula et al., 2019]
  - A framework that tries to identify the time interval where a statement was correct. It uses external knowledge bases and the web content to extract evidence to assess the validity of a statement for a time interval.

# Knowledge Enrichment

- Improve the **completeness** of a knowledge graph by adding new statements.

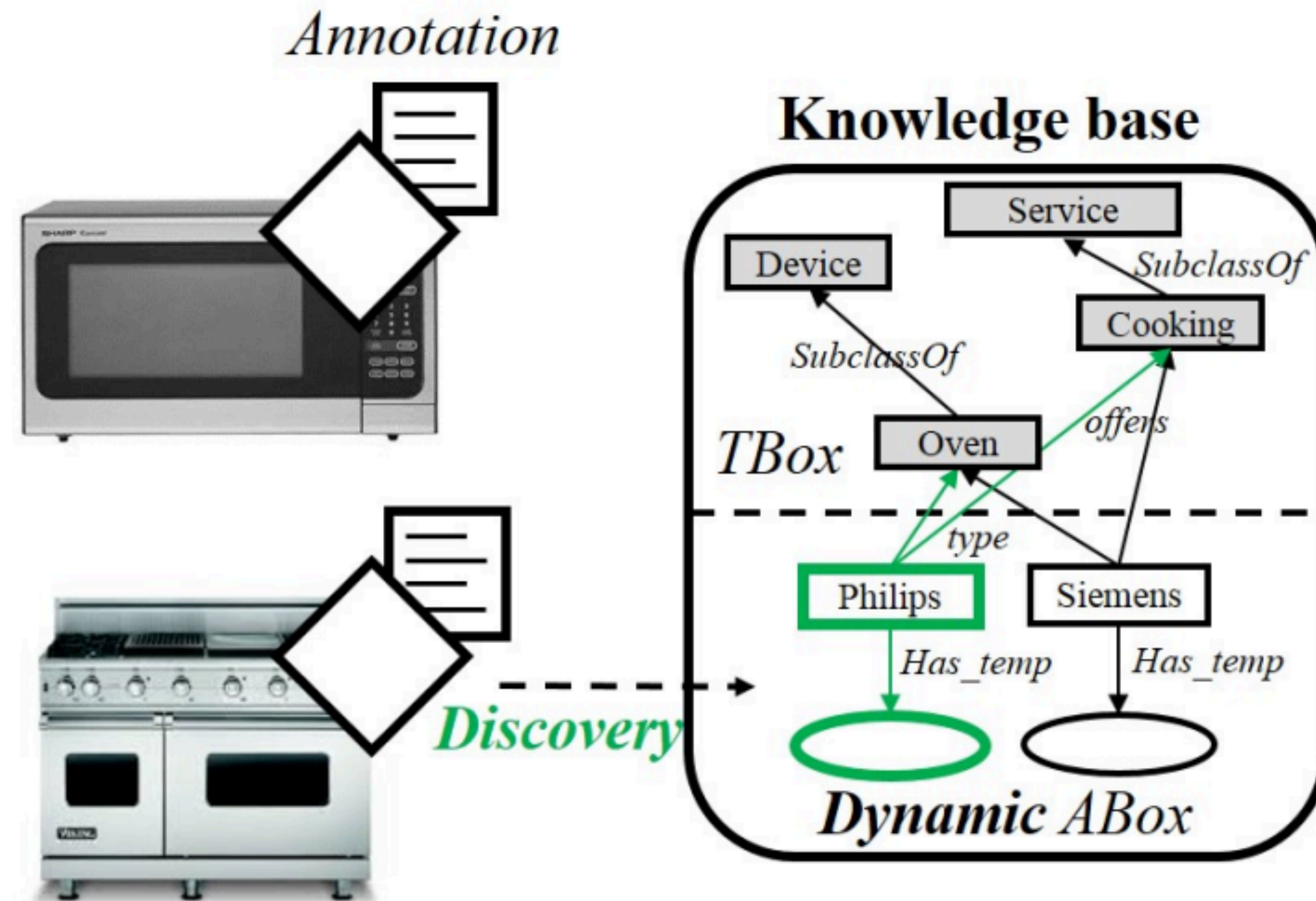


# 156 Knowledge enrichment: the property level



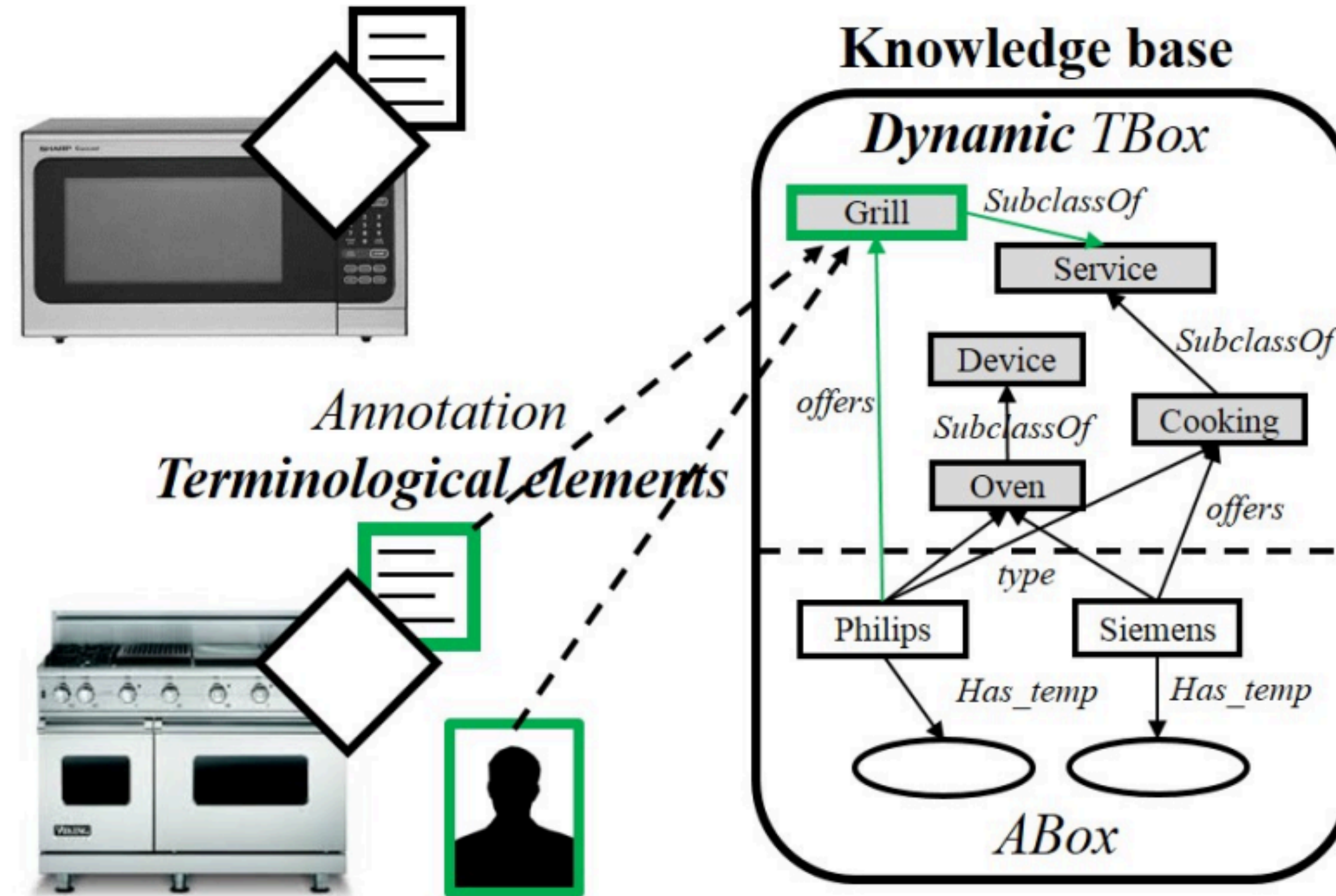


# 157 Knowledge enrichment: the instance level (ABox)



**ABox statements:**  
"assertion component"  
– facts associated with  
the TBox's conceptual  
model or ontologies.

# 158 Knowledge enrichment: the terminological level (TBox)



**TBox statements:**  
"terminology component"  
describe a domain of interest by defining classes and properties as a domain vocabulary.

# Knowledge Enrichment Tools

## Duplication detection and resolution tools

- ◎ Dedupe
  - A Python library that uses machine learning to find duplicates in a dataset and to link two datasets.
  
- ◎ LIMES [Ngomo & Auer, 2011]
  - A link discovery approach that benefits from the metric spaces (in particular triangle inequality) to reduce the amount of comparisons between source and target dataset.
  
- ◎ Duke [Garshol & Borge, 2013]
  - Uses various similarity metrics to detect duplicates in a dataset or link records between two datasets based on a given configuration.
  
- ◎ Legato [Achichi et al., 2017]
  - A recording linkage tool that utilizes Concise Bounded Description\* of resources for comparison.

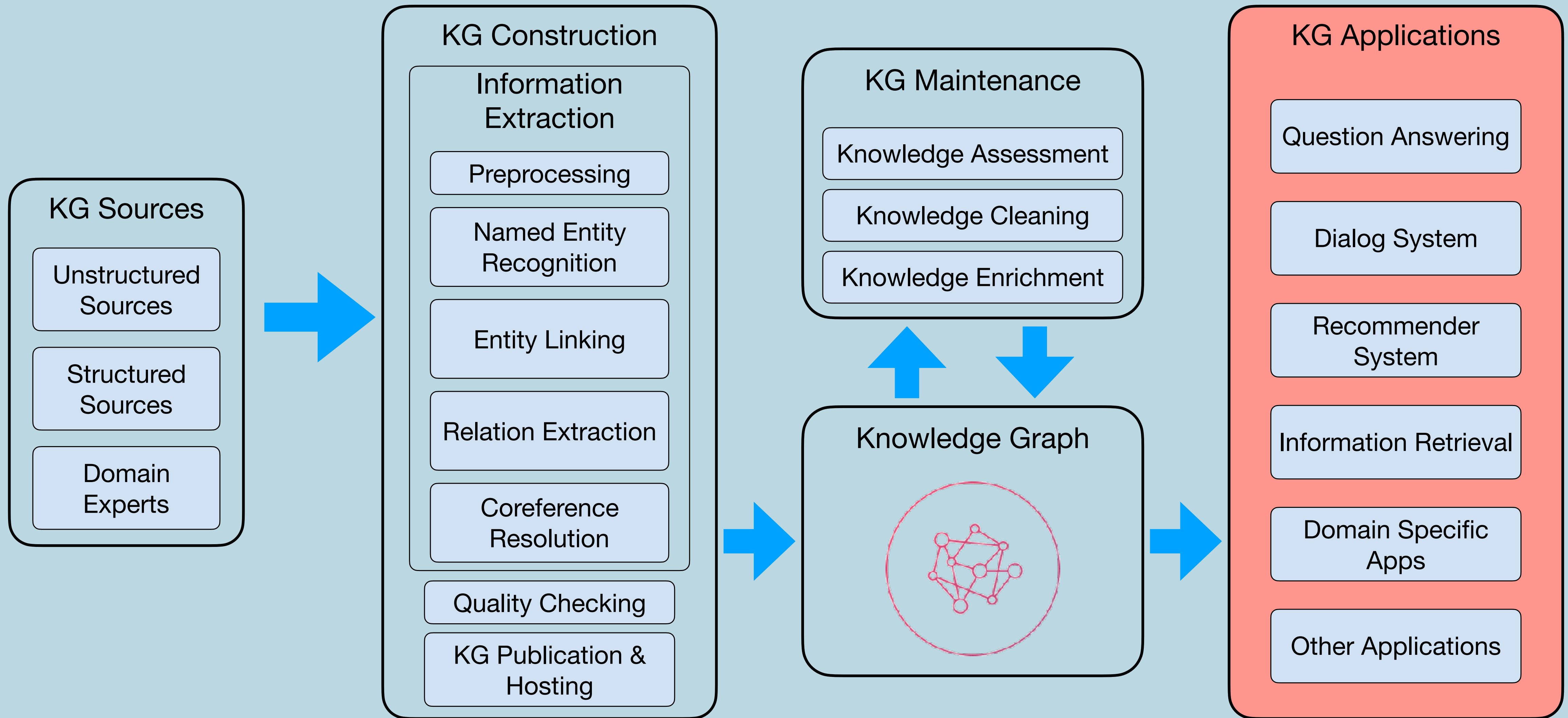


# Demo part 1

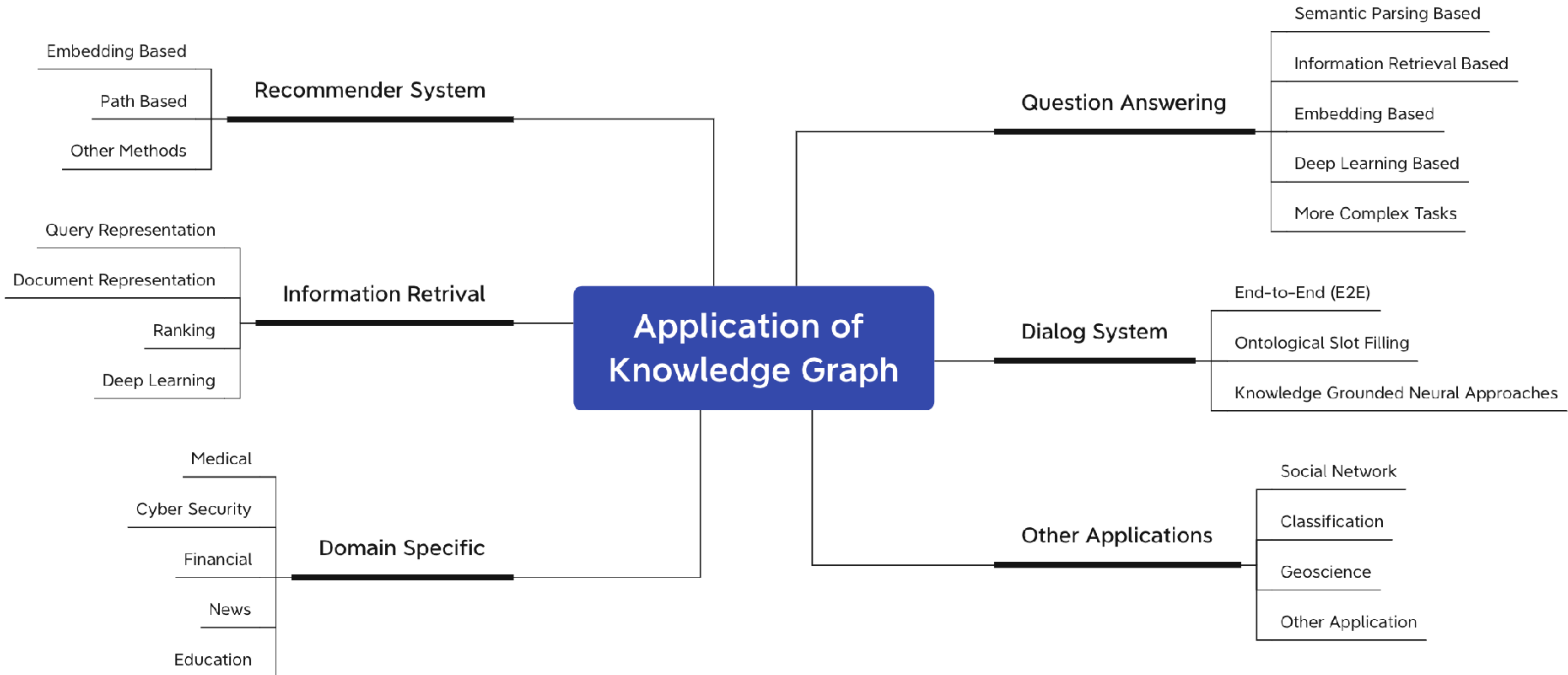
## Constructing KG

# KG Applications



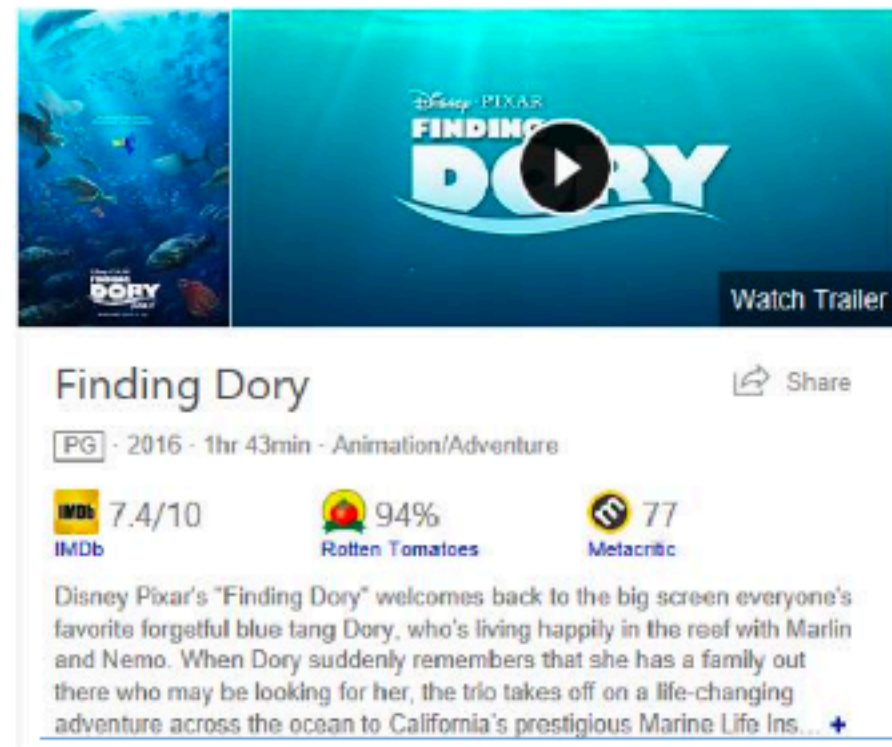




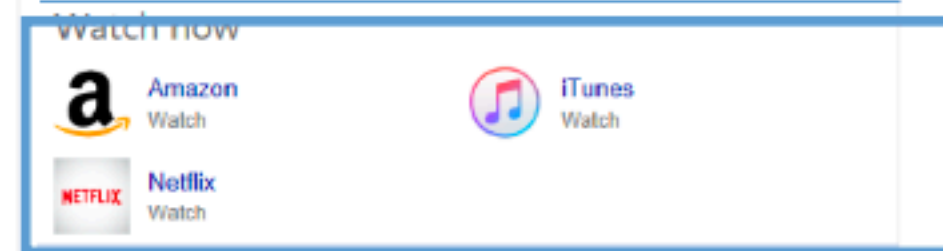


# KG Applications: Bing & Cortana

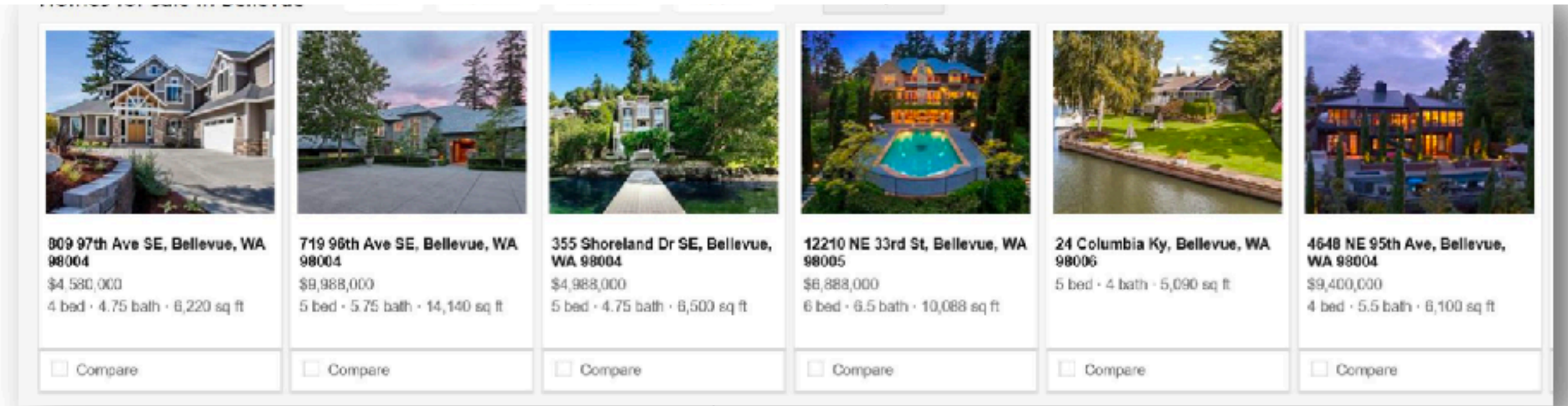
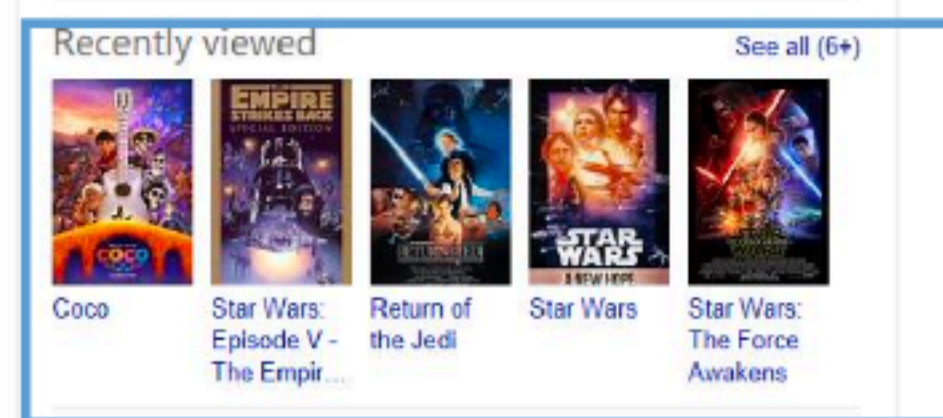
- Knowledge graph has a tremendous impact on Bing impressions



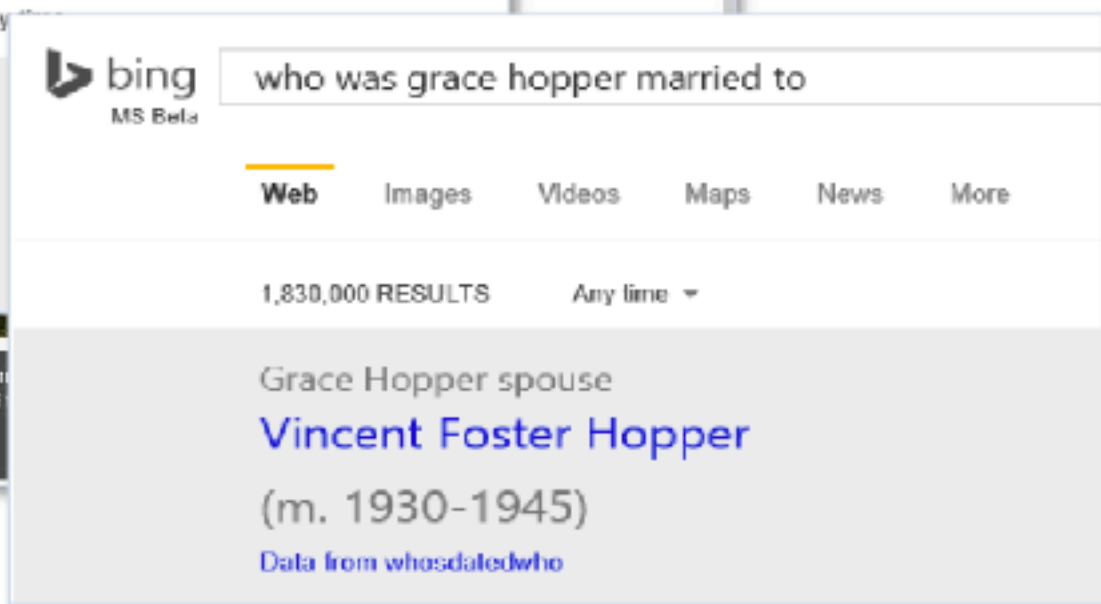
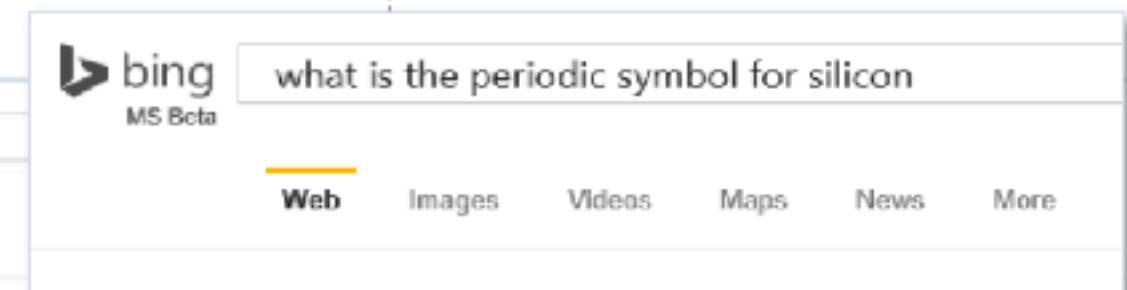
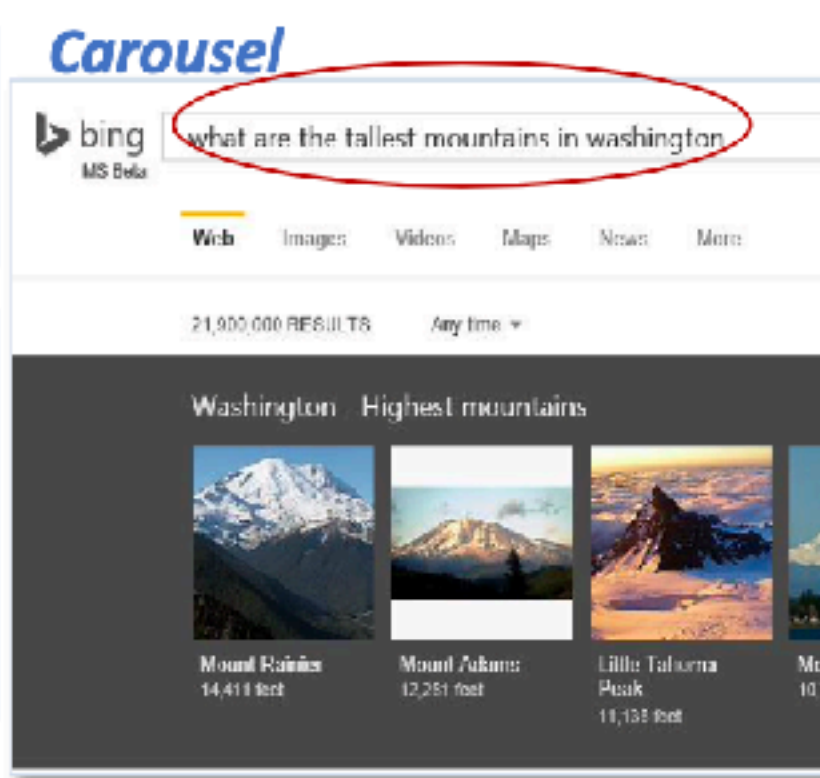
Watch now  
actions for  
movie  
entities



Recently  
viewed  
shows  
personal  
history



Carousel of information from Satori

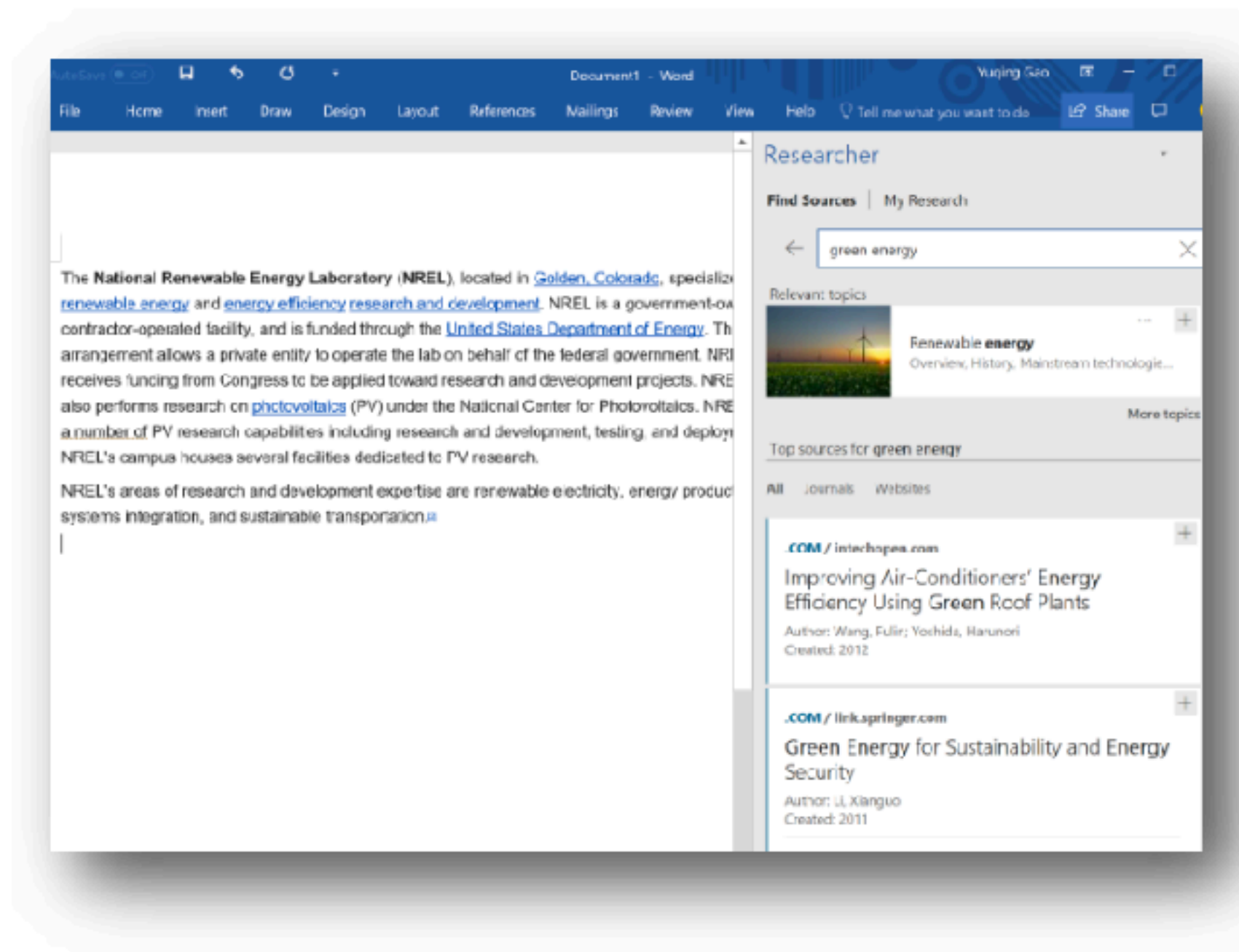


Natural language Fact Answers




# KG Applications: Office

- Enriching the Office experience with knowledge graph.



**Researcher in Word & OneNote**  
Get topic information straight into your documents

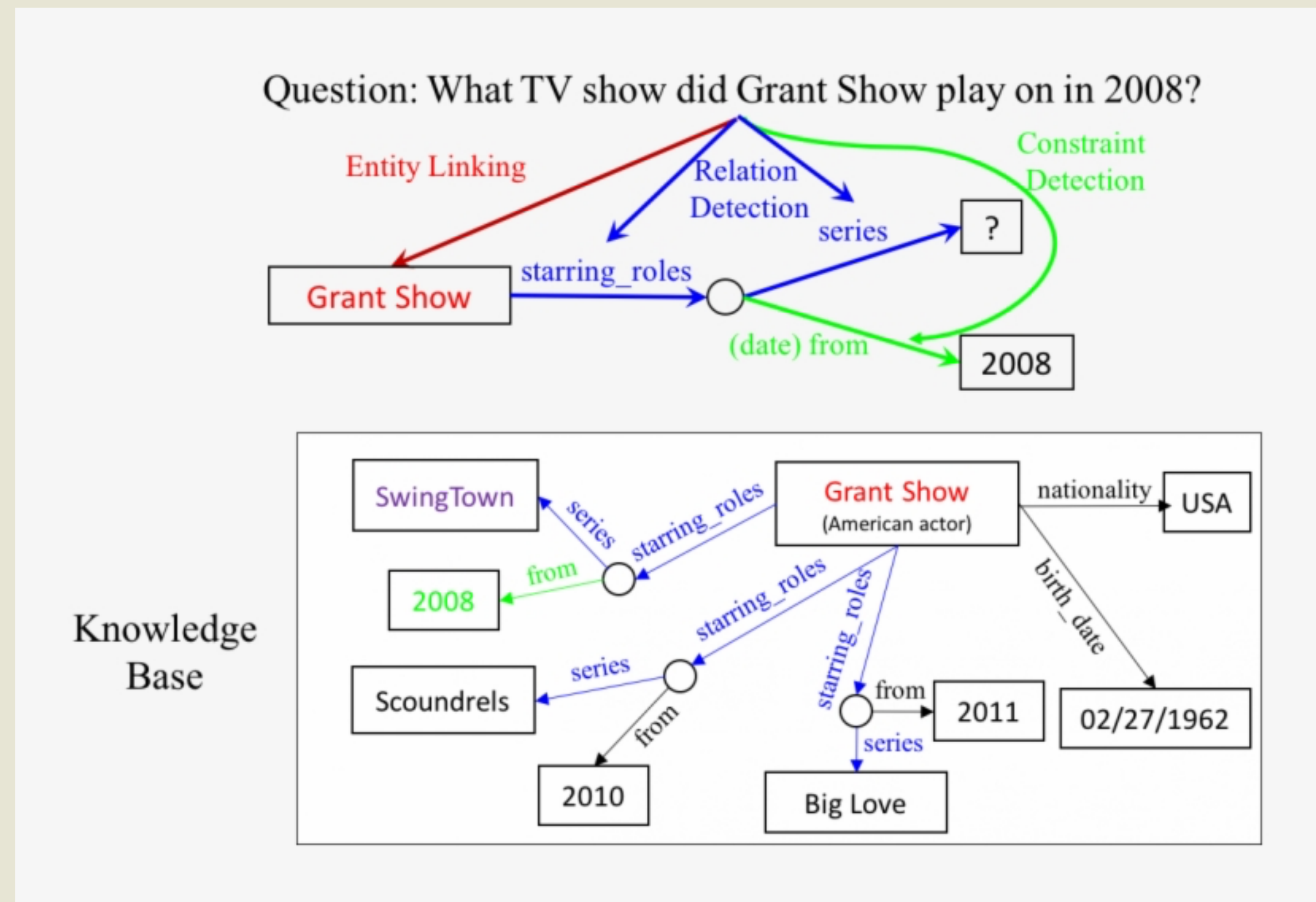
	A	B	C	D	E
2					
3		Stock	52 week high	52 week low	Exchange
4		Microsoft Corp	96.07	63.62	NASDAQ
5		Alphabet Inc	1,186.89	803.37	NASDAQ
6		Apple Inc			
7		Amazon.com Inc			
8		Walt Disney Co			
9		Starbucks Corp			
10		Nike Inc	Walt Disney Co		
11		AT&T Inc	Price	52 week high	
12		HP Inc	103.39	116.10	
13		International Busin	Price (after hours)	52 week low	
14			102.61	96.20	
15			Shares outstanding	Beta	
16			1,503,675,479	1.31	
17					
18					
19					

**Project Yellow (Excel)**  
Finance and demographic information available based on cell contents



# KG-based Question Answering

- Given:
  - Knowledge graph ingested from unstructured, structured, and semi-structured data sources
- Input:
  - Natural language query
- Output:
  - Answer in the form of knowledge



# KG-based Question Answering

✕ 🔊 🔍

[All](#) [Pictures](#) [News](#) [Shopping](#) [Videos](#) [More](#) [Tools](#)

Approximately 56,100,000 results (0.61 seconds)

Seattle Seahawks / Head Coach

## Pete Carroll

Seahawks **coach Pete Carroll** has no worries about job status: 'I'm in great shape' On a day in which some NFL head coaches might see their tenures end, Pete Carroll isn't fretting his future in Seattle following Sunday's 38-30 season -ending victory over the Arizona Cardinals. Jan 10, 2022

<https://www.nfl.com/news/seahawks-coach-pete-carroll...>

### Seahawks coach Pete Carroll has no worries about job status

associated researches

## Pete Carroll

Football coach

Peter Clay "Pete" Carroll is an American football coach. He is currently the head coach of the Seattle Seahawks of the National Football League. [Wikipedia](#)

**Date/Place of birth** : September 15, 1951 (Age: 70), [San Francisco, California, United States](#)

**End of career** : 1972

**Date of registration** : 1971 (Pacific Tigers football)

**Team** : [Seattle Seahawks](#)

**Teams Coached** : [USC Trojans American Football Team](#) (Head Coach, 2001–2009) , [MORE](#)

**Wife** : [Glenn Goranson](#) (d. 1976), [Wendy Pearl](#) (d. 1973–1975)

**Children** : [Brennan Carroll](#) , [Jaime Carroll](#) , [Nathan Carroll](#) , [Nate Carroll](#)

[to report a problem](#)

[to report a problem](#)

# KG-based Question Answering

## ● Challenges:

- Matching natural language relations to formal relations in KGs
- There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, ...etc
- Scalable entity linking
- Word sense disambiguation
- Semantic roles and relationships extraction

## ● Large search space

- Every entity can have hundreds of edges and every entity instance can have hundreds of millions of edges/facts

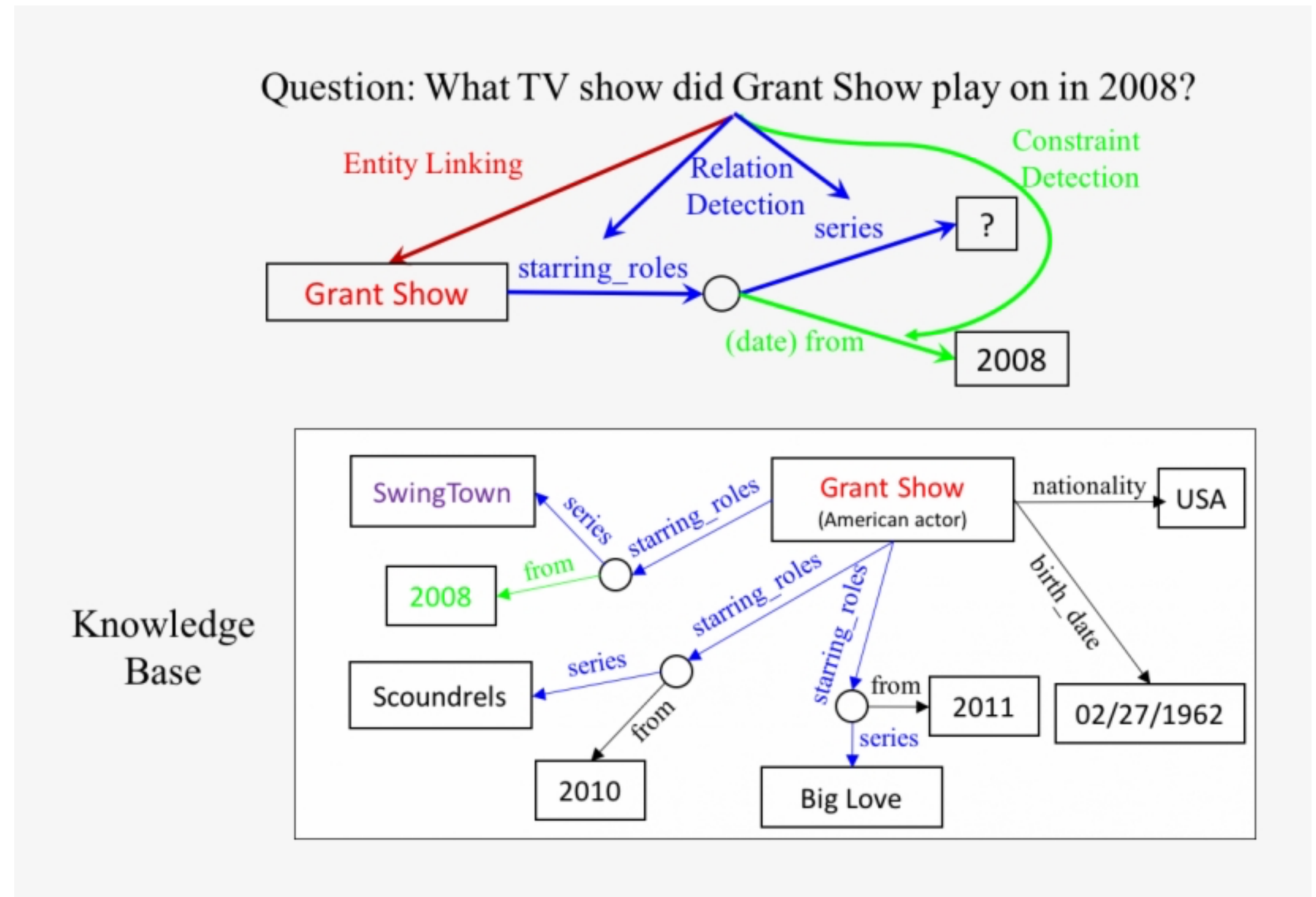
## ● Compositionality: multi-hop reasoning

- {Movies starring the first wife of tom hanks}, {movies directed by the director of titanic}



# 169 Approaches for KG-based Question Answering

- Semantic parsing approaches:
  - Generic semantic parsing followed by ontology grounding
- Knowledge embedding
- Information extraction approaches:
  - Information retrieval methods with semantic enrichment



# 170 Semantic Parsing

2017 movies starring the actor that played batman in batman

All Images Videos Maps News Shop | My saves

Microsoft Sign in to see results from Microsoft >

12,900,000 Results Any time ▾

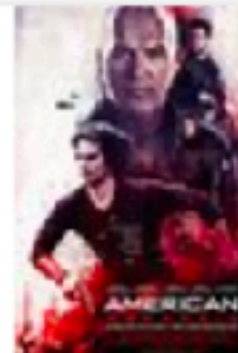
2017 Movies starring Michael Keaton who played Characters named Batman who acted in Batman (1989)

Spider-Man: Homecoming



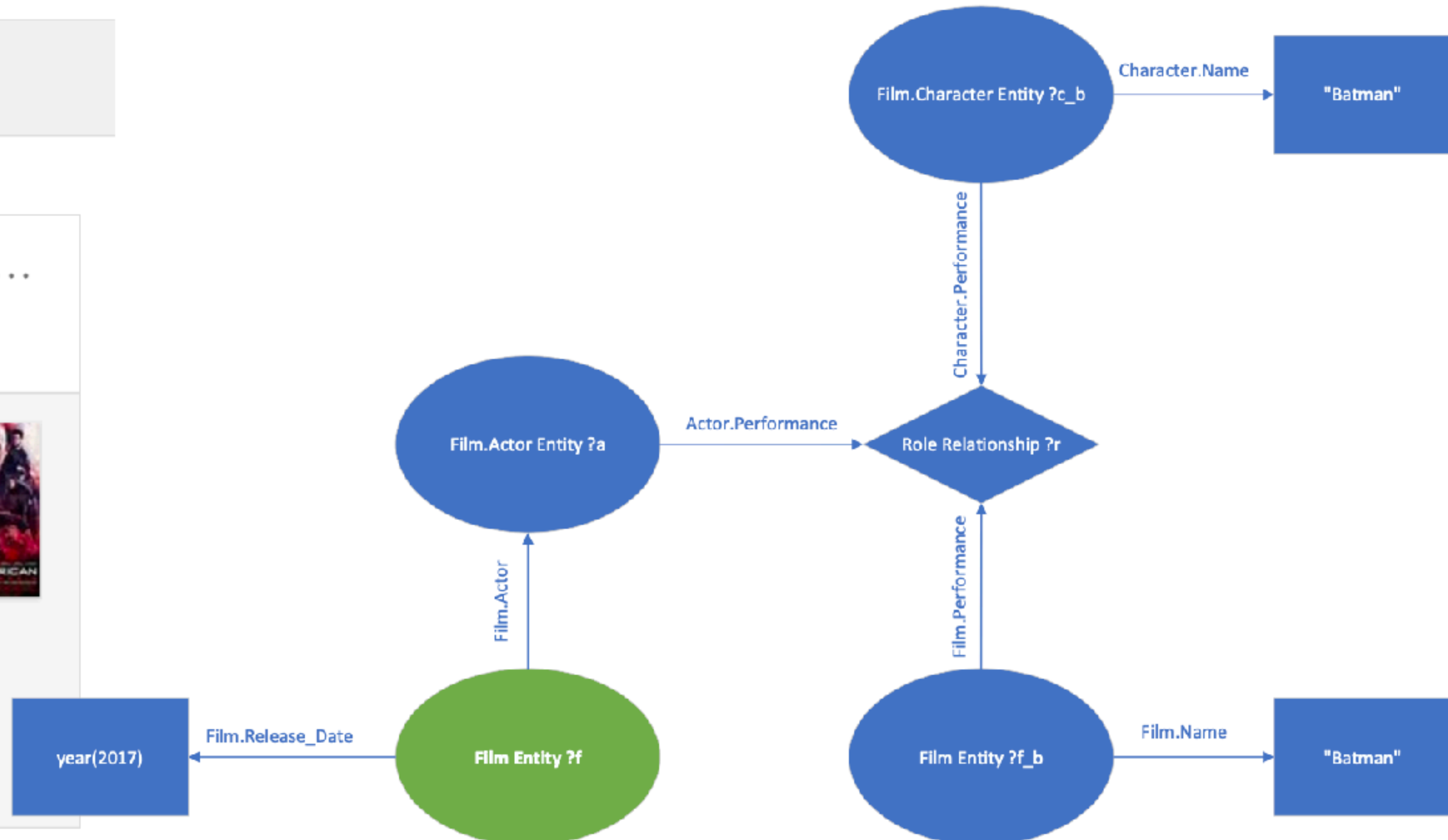
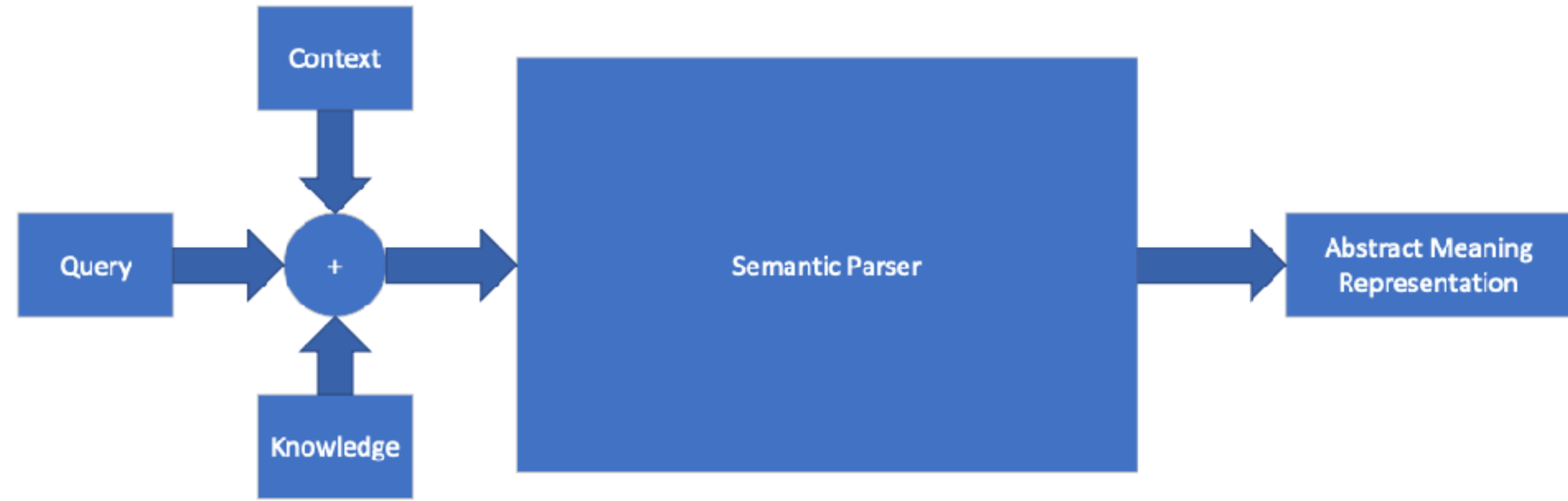
American Assassin

2017 · Suspense



The Founder

2017 · Biography



# Generic Semantic Parsing

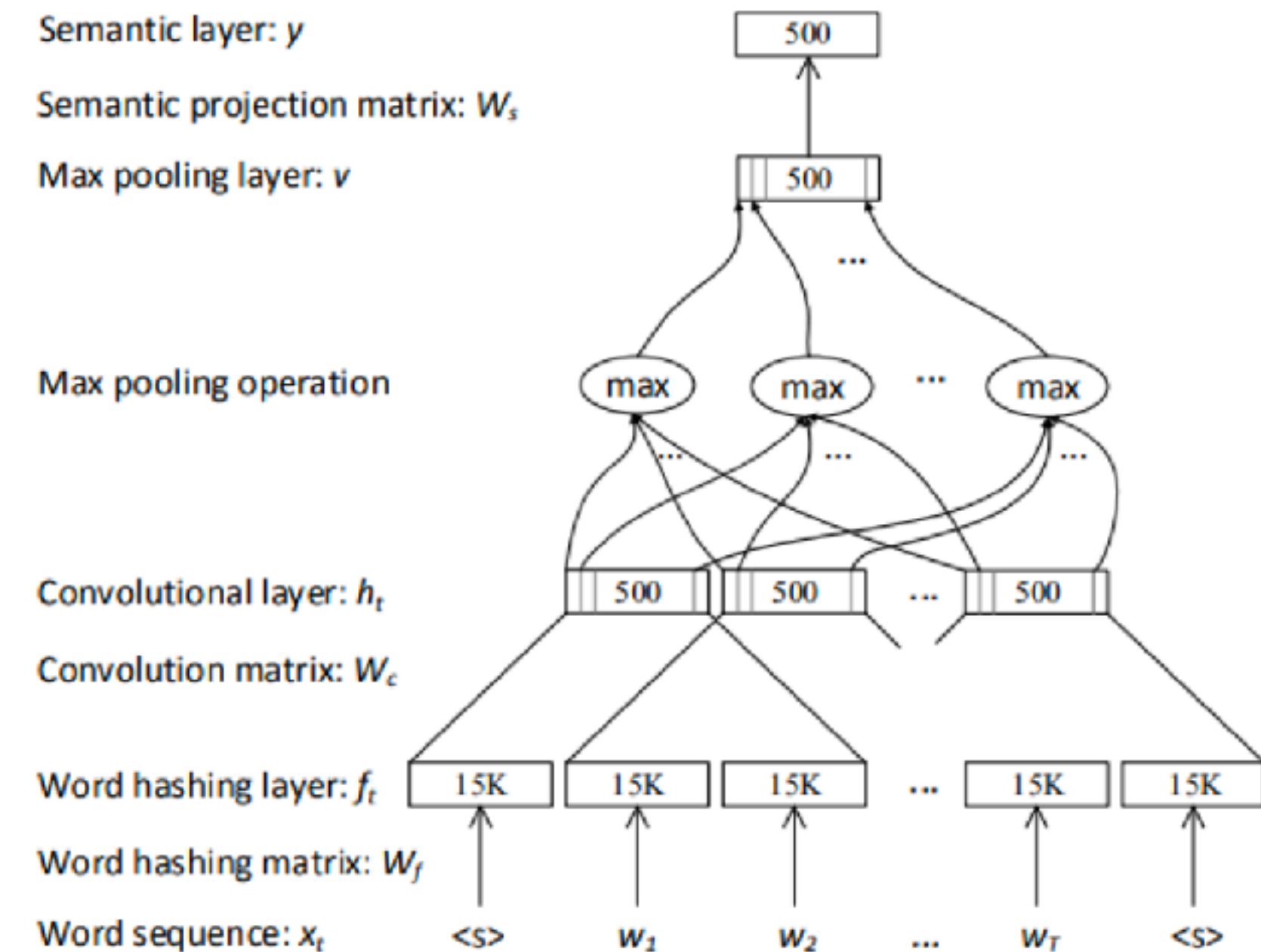
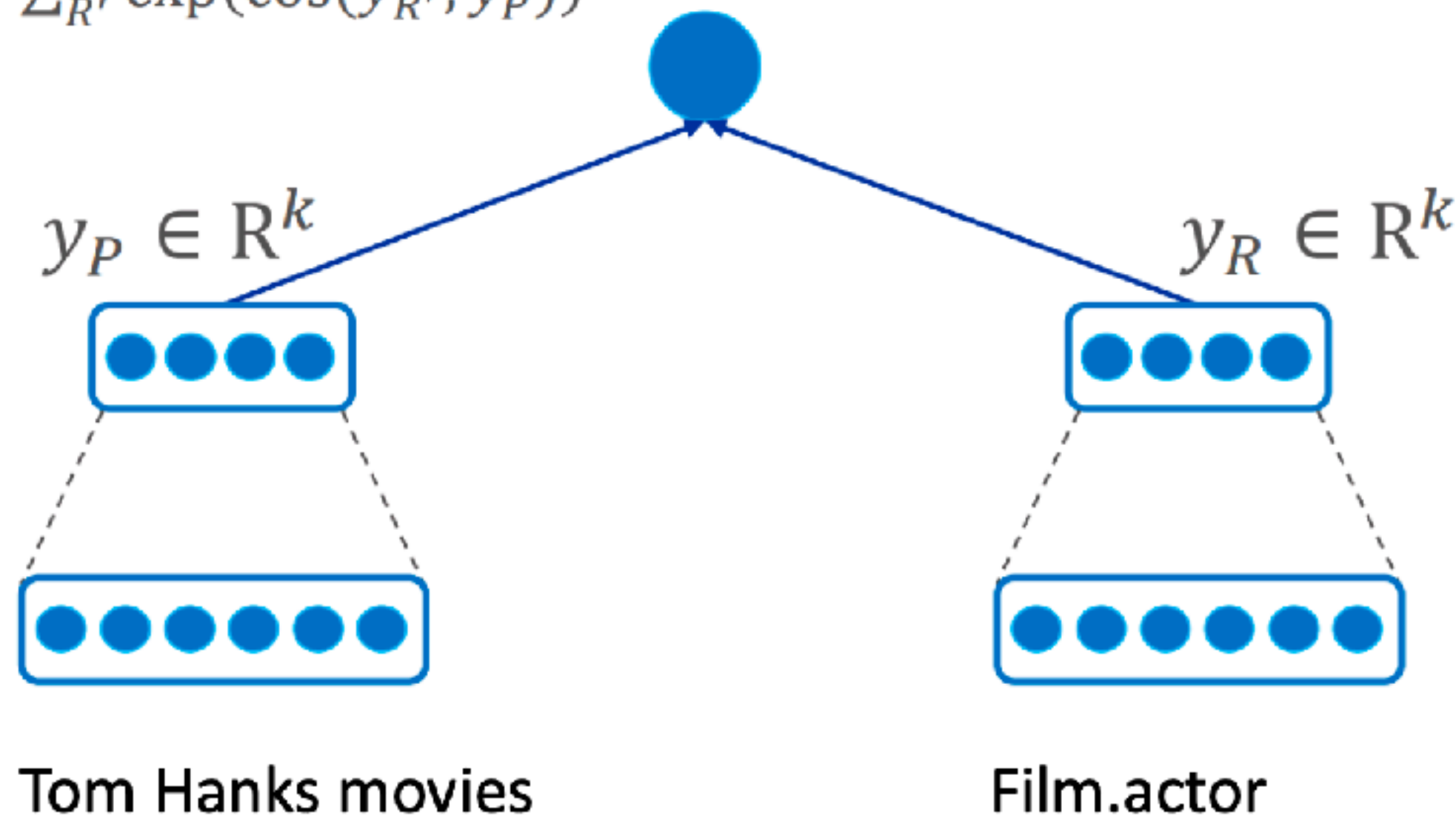
- In this approach as in the example provided by [Kwiatkoski 13], we:
  1. Perform a generic semantic parsing of the utterances
  2. Perform ontology matching on relationships
- For example: **{who is Donald Trump's Daughter}**
  - $\lambda x. \text{daughter\_of Donald Trump, } x$
  - $\lambda x. \text{child\_of Donald Trump, } x \wedge \text{gender}(x, \text{female})$
- This semantic expression can be then compiled into a knowledge graph database query e.g. Cypher/SPARQL and executed to return the results



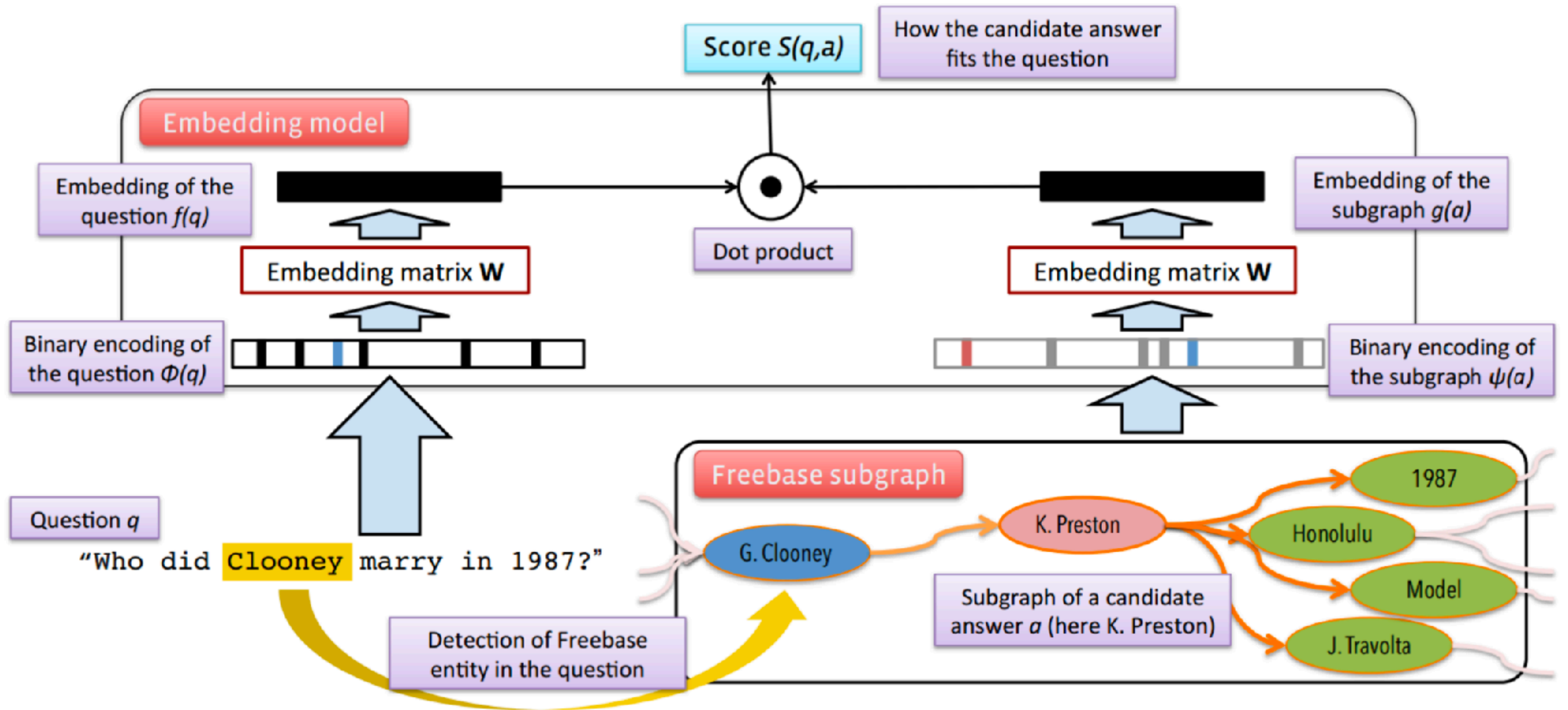
# 172 Ontology Matching on Relationships using DSSM

- Input is mapped into two  $k$  dimensional vectors
- Probability is determined by *Softmax* of their cosine similarity

$$P(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$$

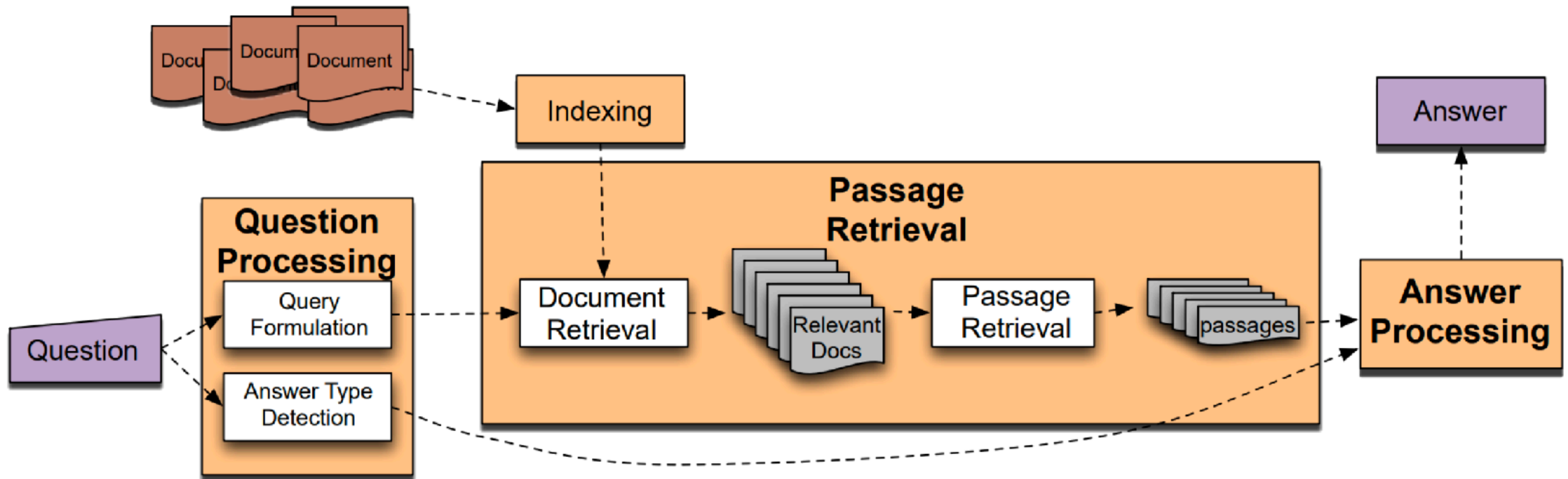


# Knowledge Embedding Approach



# 174 Information Extraction Approaches

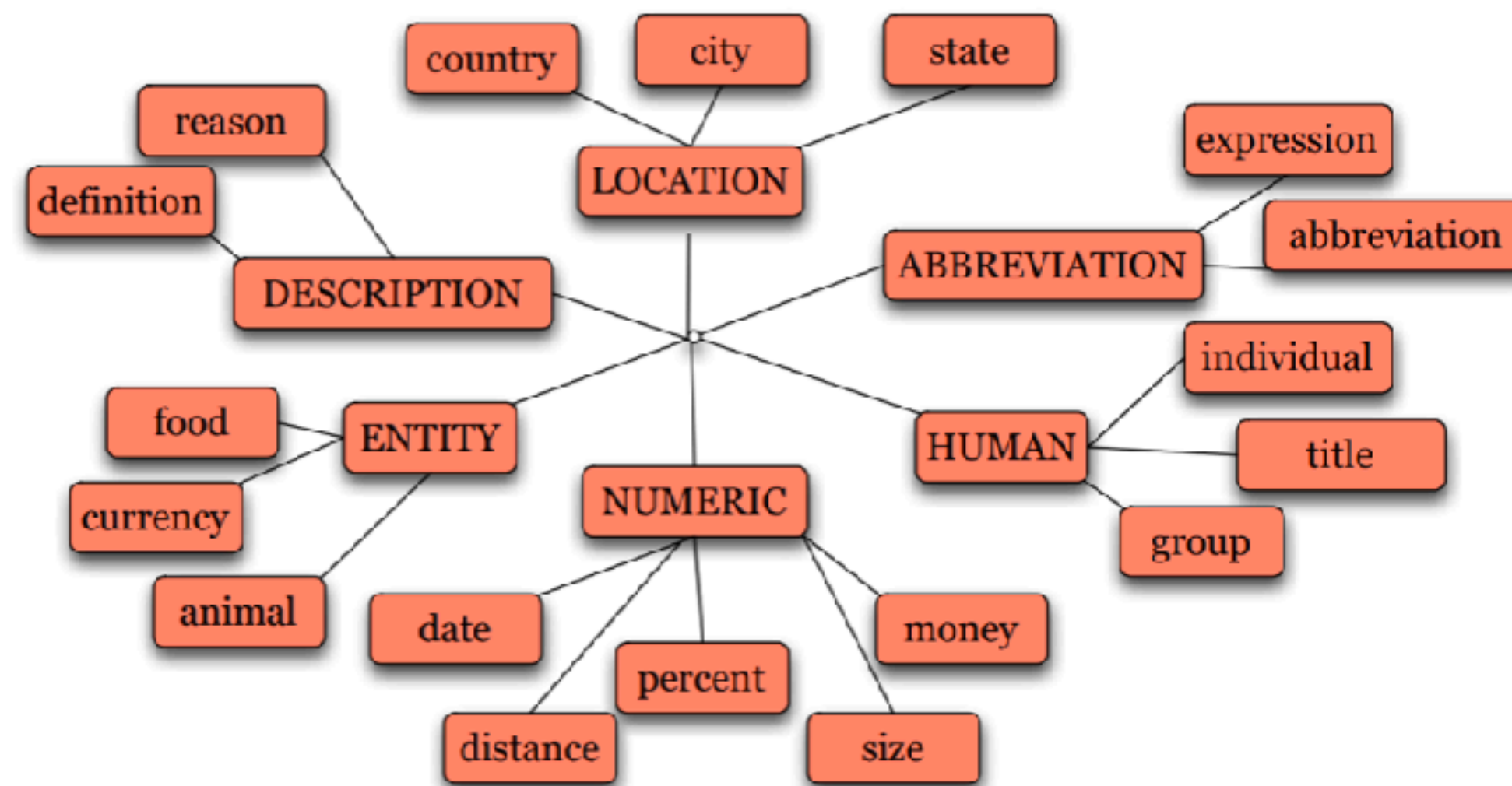
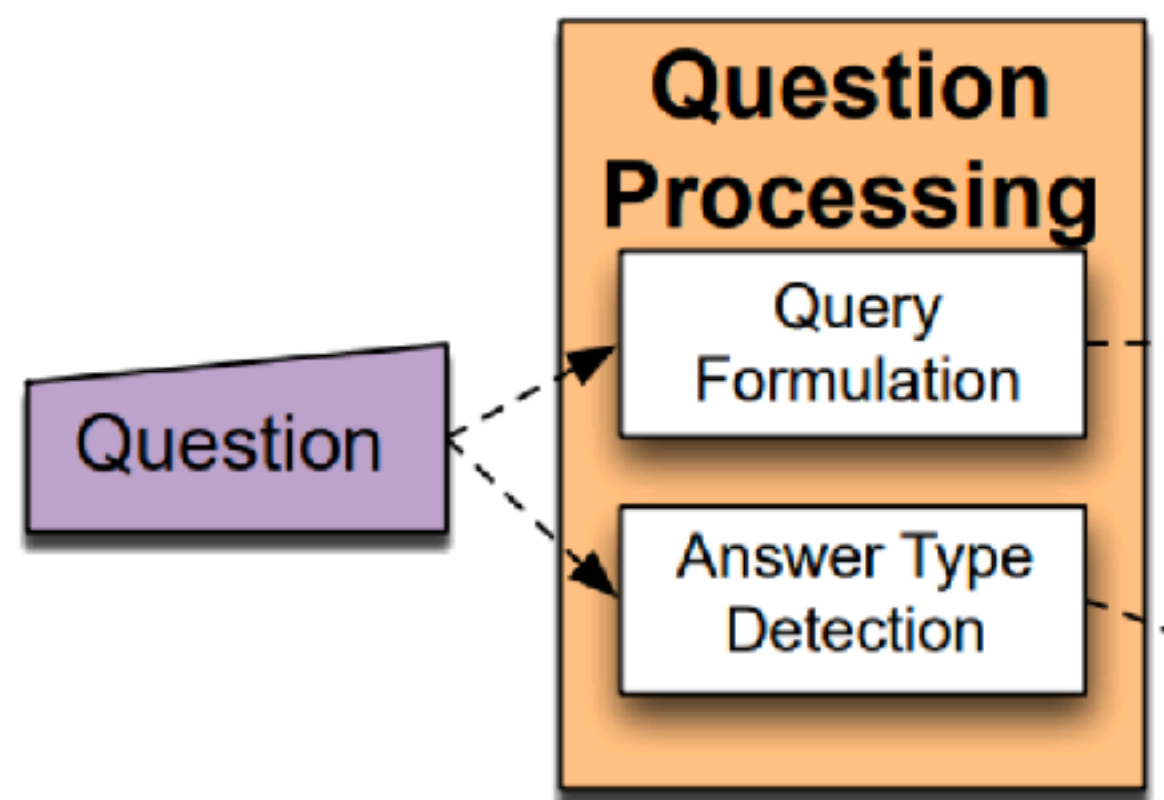
- Extracting and answers on the fly.
- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.





# Answer Type Detection

- Who first landed on the moon => Person
- Where is the headquarters of Microsoft => Location
- What is the largest country in population => Country
- Highest flying bird => Animal/Bird



Learning Question Classifiers [Xin Li & Dan Roth, COLING 2002]

Question Answering [Dan Jurafsky, Stanford]

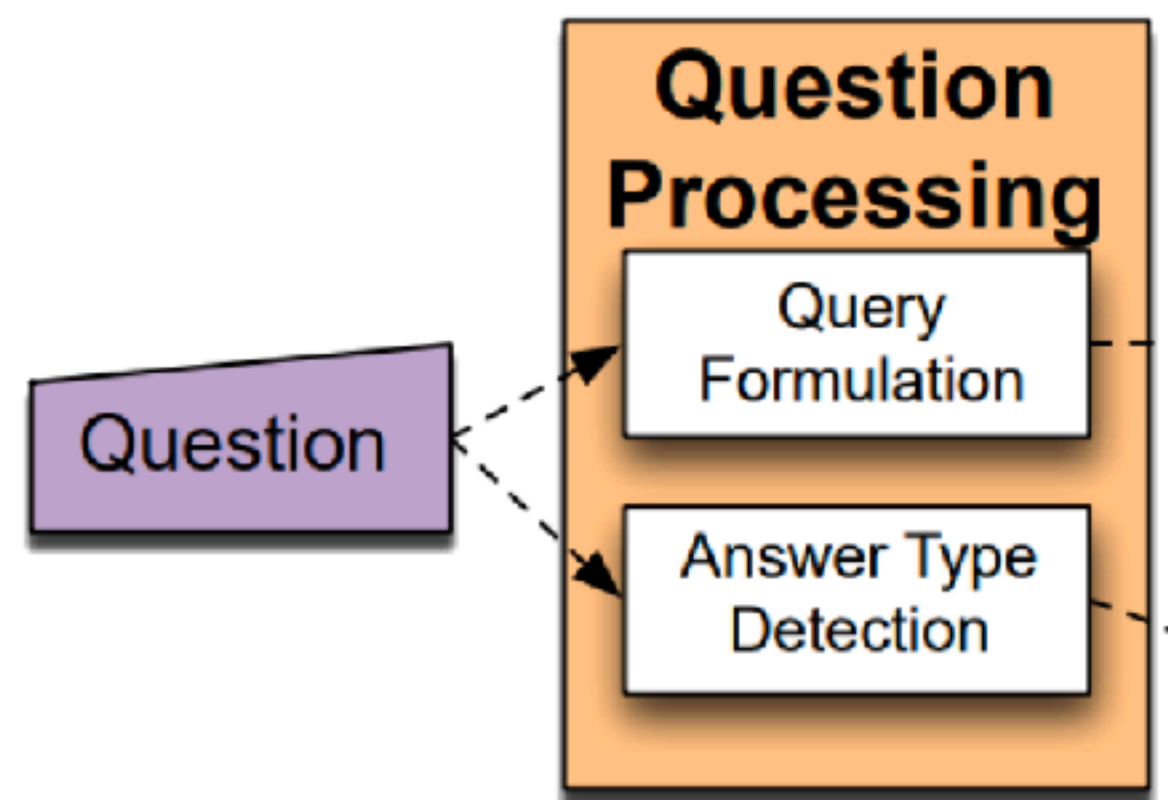
# Answer Type Detection

## ● Rules

- Grammar for e.g. who be/... => Person
- Head word for e.g. which city is the largest
- ...

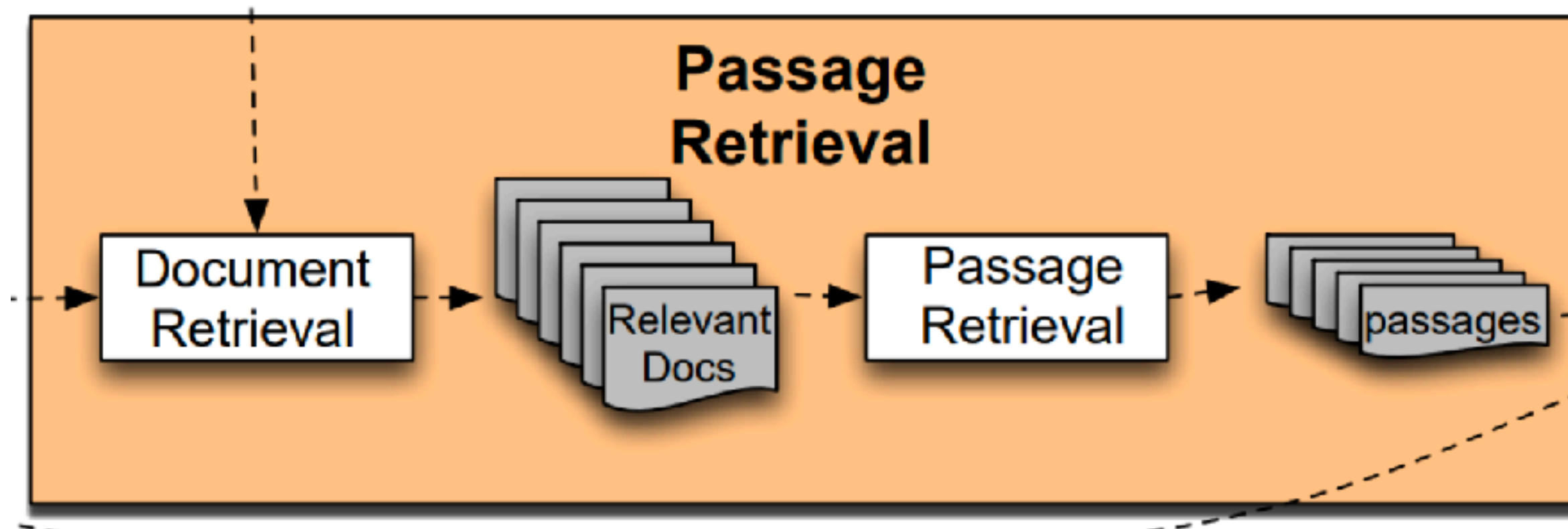
## ● Learned type classifier

e.g. SVM utilizing features like question words, phrases, POS tags, headwords, mentioned entities, ...etc [Dan Jurafsky]



# Passage Retrieval

- Retrieve documents using expanded query terms + search engine
- Segment the documents into smaller units e.g. passages/paragraphs
- Rank passages using learned model utilizing features like:
  - Number of named entities of the right type in the passage
  - Number of query words in the passage
  - Number of question n-grams in the passage
  - Proximity of query words in the passage
  - Longest sequence of question words
  - Rank of document containing passage,...etc

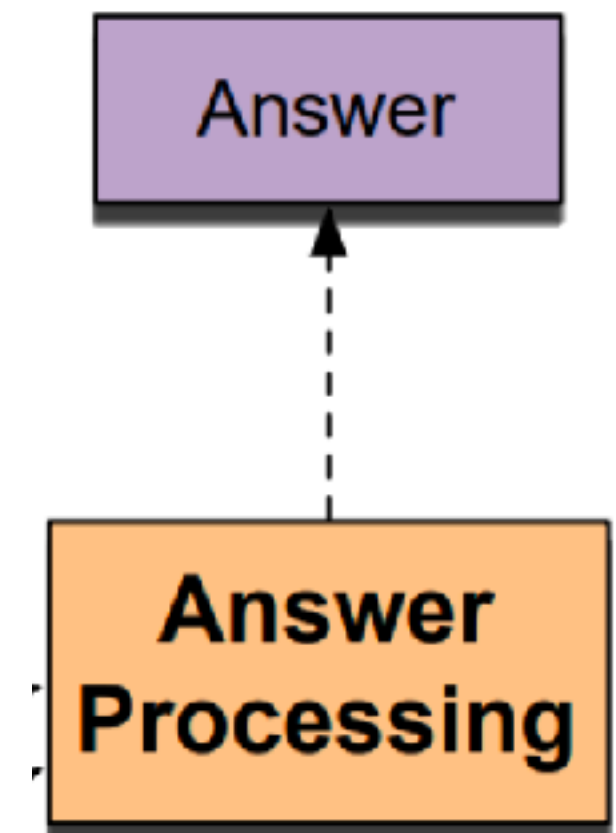




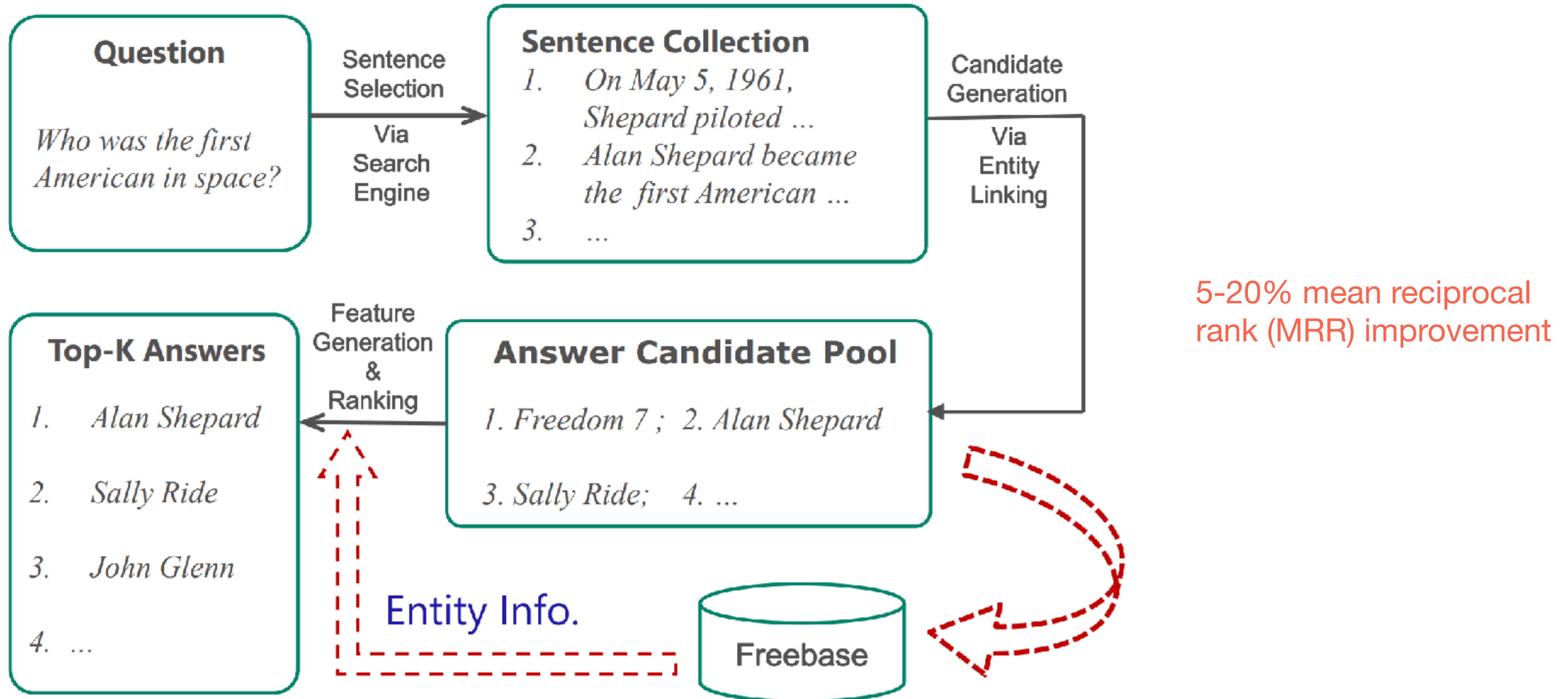
# Process Answer

- Detect answer entity by running NER on the passage
- Mark the answer entity in the passage
- How many bones in an adult human body? (**Number**)
  - The human skeleton is the internal framework of the body. It is composed of 270 bones at birth – this total decreases to **206 bones** by adulthood after some bones have fused together.

Question Answering [Dan Jurafsky, Stanford]



# 179 Answer Semantic Enrichment using KB



Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

◎ Approaches:

- E2E Seq2seq (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015)
- Knowledge based ontological slot filling (Dai+ 2017)
- Knowledge grounded neural approaches (Ghazvininejad+ 2018)
- ...



# E2E Dialog Systems

- Suitable for chitchat kind of bots.
- Predicted target sequences are usually free from facts

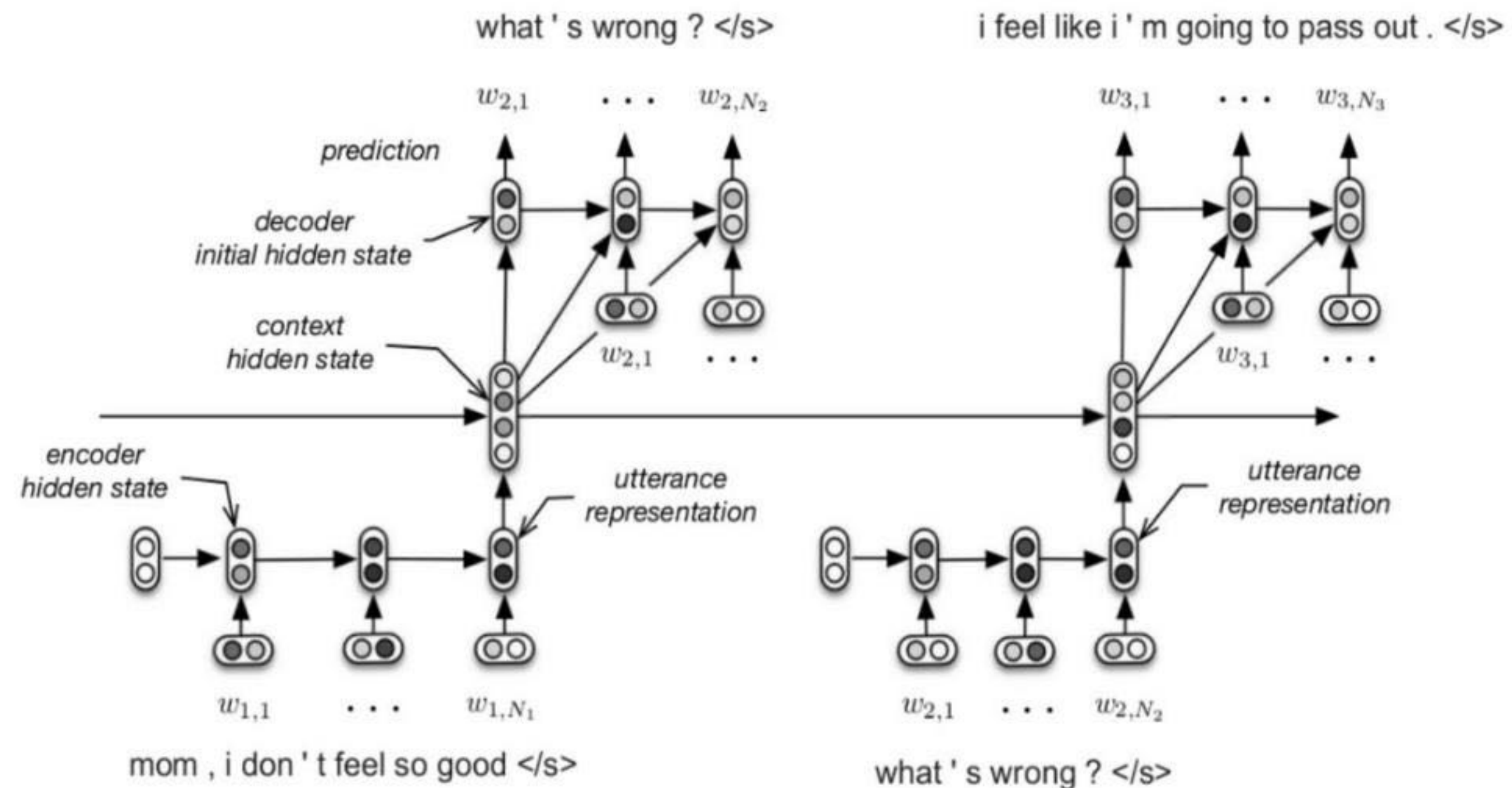
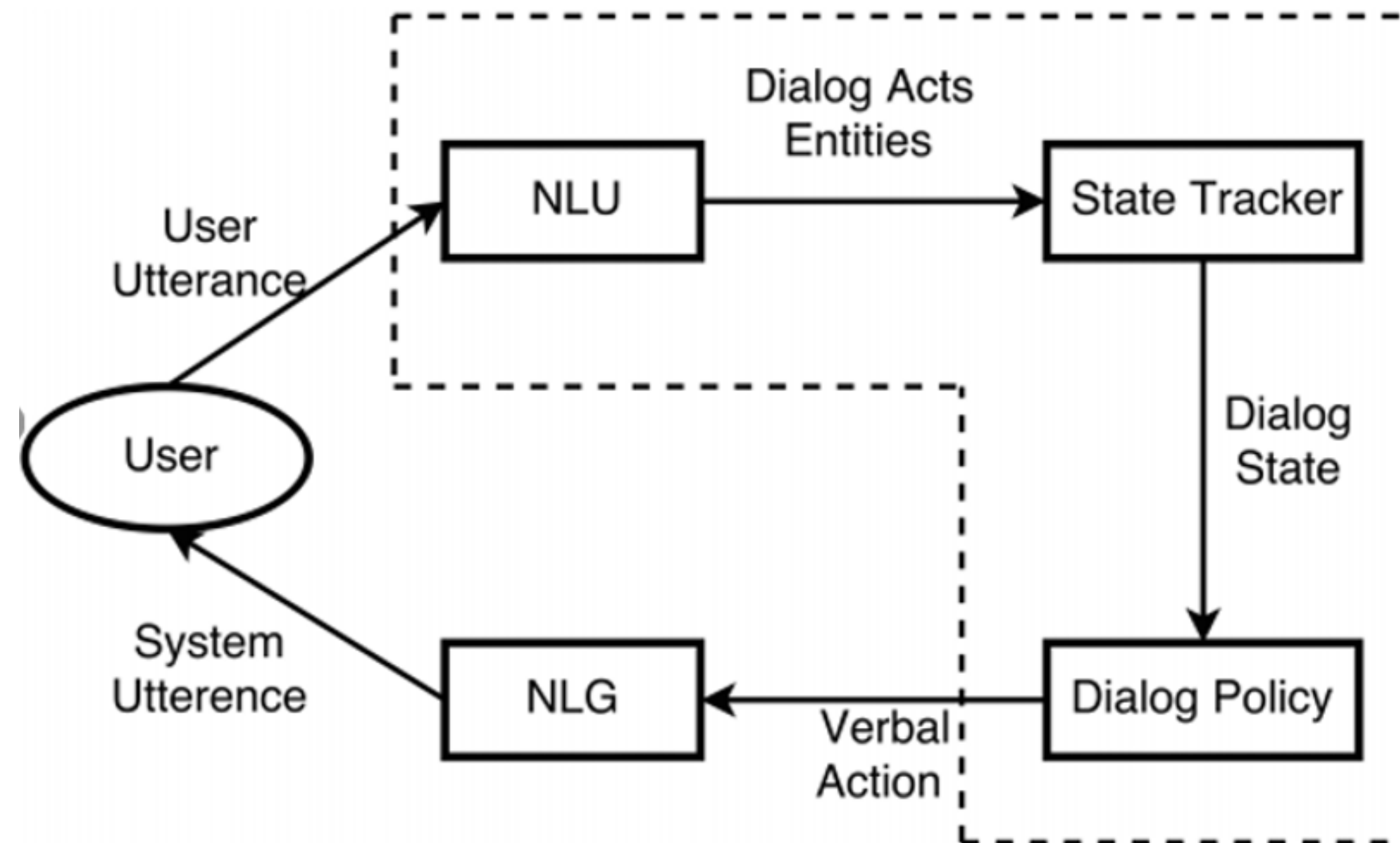


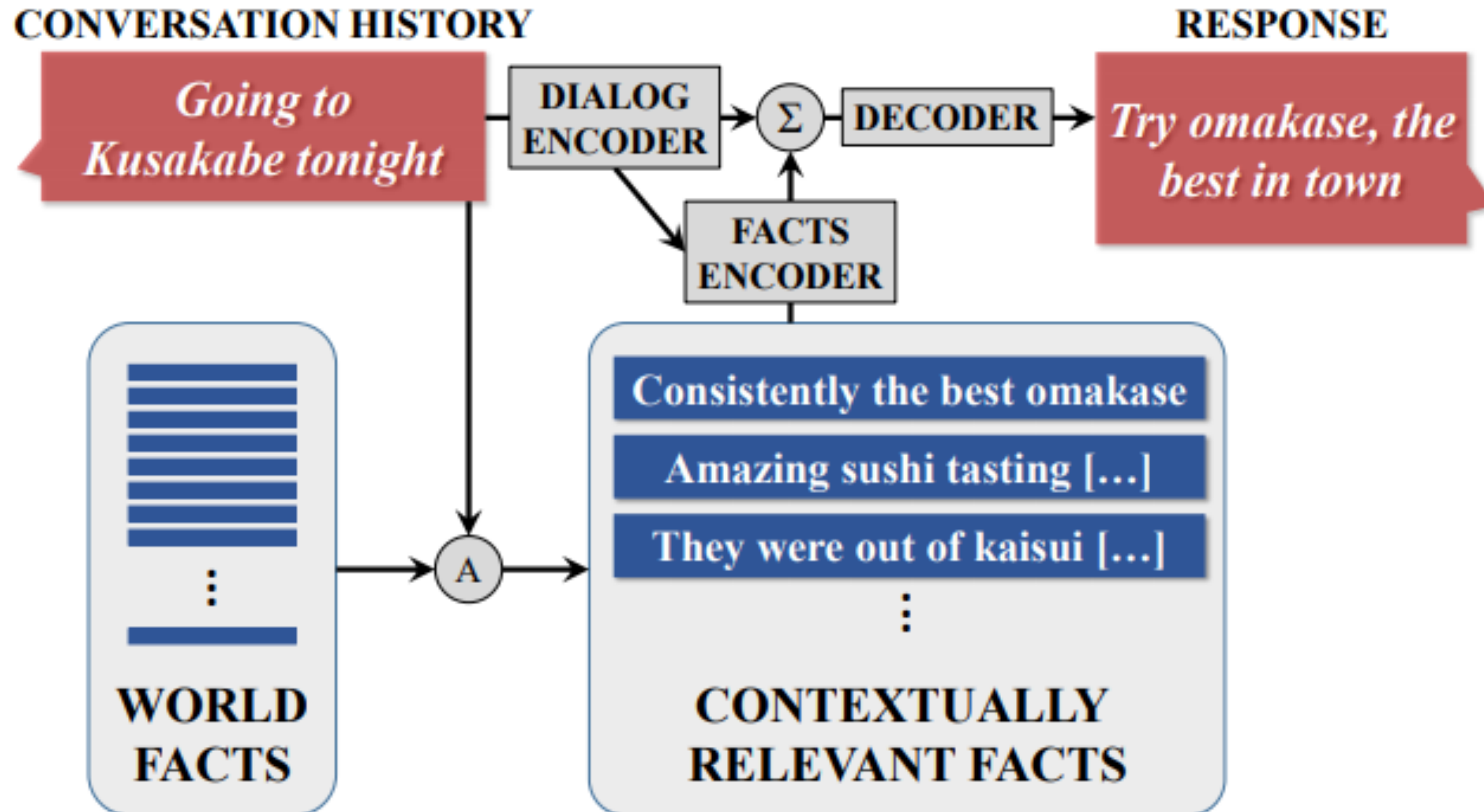
Figure 8: A computational graph representing the HRED architecture for dialogue over a span of three turns. The major addition to the architecture is a higher-level *context*-RNN keeping track of past utterances by progressively processing over time each utterance vector and conditioning the decoding on the last hidden state of the context vector (middle).

# 182 Knowledge Based Ontological Slot Filling

- Suitable for chitchat kind of bots.
- Predicted target sequences are usually free from facts



# 183 Knowledge Grounded Neural Approaches





# Enterprise Scenarios

◎ **All the challenges mentioned previously plus the following:**

- Compliance
- Different data formats: databases, emails, chat logs, discussion forums, web blogs, pdfs, PowerPoint/Word/Excel documents etc.
- Different schemas: schema mapping and merging, and new schema discovery.
- Consumption via dialog systems, search interface, mobile devices or other modalities, API.
- Highly domain-specific models required, bootstrapped by pre-trained models. Need on-prem domain-adaptation.

# Demo part 2

## KG-based QA



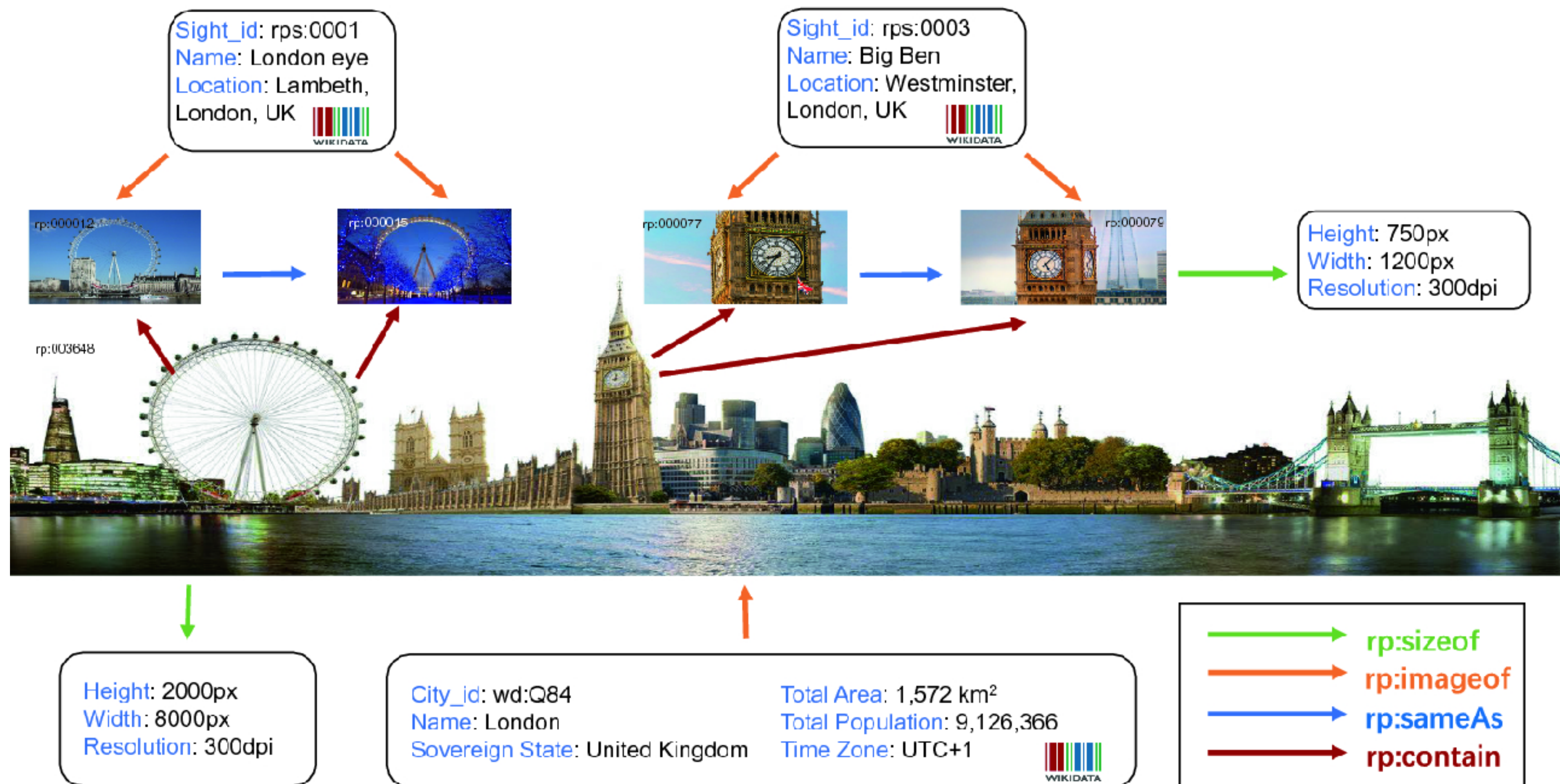
A woman's profile is shown in silhouette against a blue background. She is looking towards the left. The background is filled with various digital and data-related visualizations. On the left, there is a network graph with yellow nodes and lines. In the center and right, there are several floating rectangular panels containing different types of data visualizations: bar charts, line graphs, pie charts, and network diagrams. Some panels have text labels such as 'Administration', 'Human Resources', 'Legal', 'Accounting', 'Marketing', 'Publicity', 'Research', 'Business', 'Development', 'Engineering', 'Manufacturing', and 'Planning'. The overall aesthetic is futuristic and high-tech, with a strong emphasis on data and research.

# Exciting Active Research



# Multimodal Knowledge Graph

- Richpedia: provide a comprehensive multi-modal knowledge graph by distributing sufficient and diverse images to textual entities in Wikidata





# CommonSense Knowledge Graph

## ● **CSKG**: The CommonSense Knowledge Graph

### Choosing an intuitive follow-up scene

Q: On stage, a woman takes a seat at the piano. She:

1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
- > 4. nervously sets her fingers on the keys.



- Answering this question requires knowledge that humans possess and apply, but machines cannot distill directly in communication.
  - **ConceptNet**: pianos have keys and are used to perform music
  - **WordNet**: pianos are played by pressing keys
  - **ATOMIC**: before a person plays piano, they need to sit at it, be on stage, and reach for the keys.
  - **FrameNet**: frame of a performance contains two separate roles for the performer and the audience, meaning that these two are distinct entities, which can be seen as evidence against answer

# Causal Knowledge Graph

- **CausalKG**: Causal Knowledge Graph - Explainability using interventional and counterfactual reasoning

Representation of causality in AI systems using knowledge-graph based approach is needed for better explainability, and support for intervention and counterfactuals, leading to improved understanding of AI systems by humans.

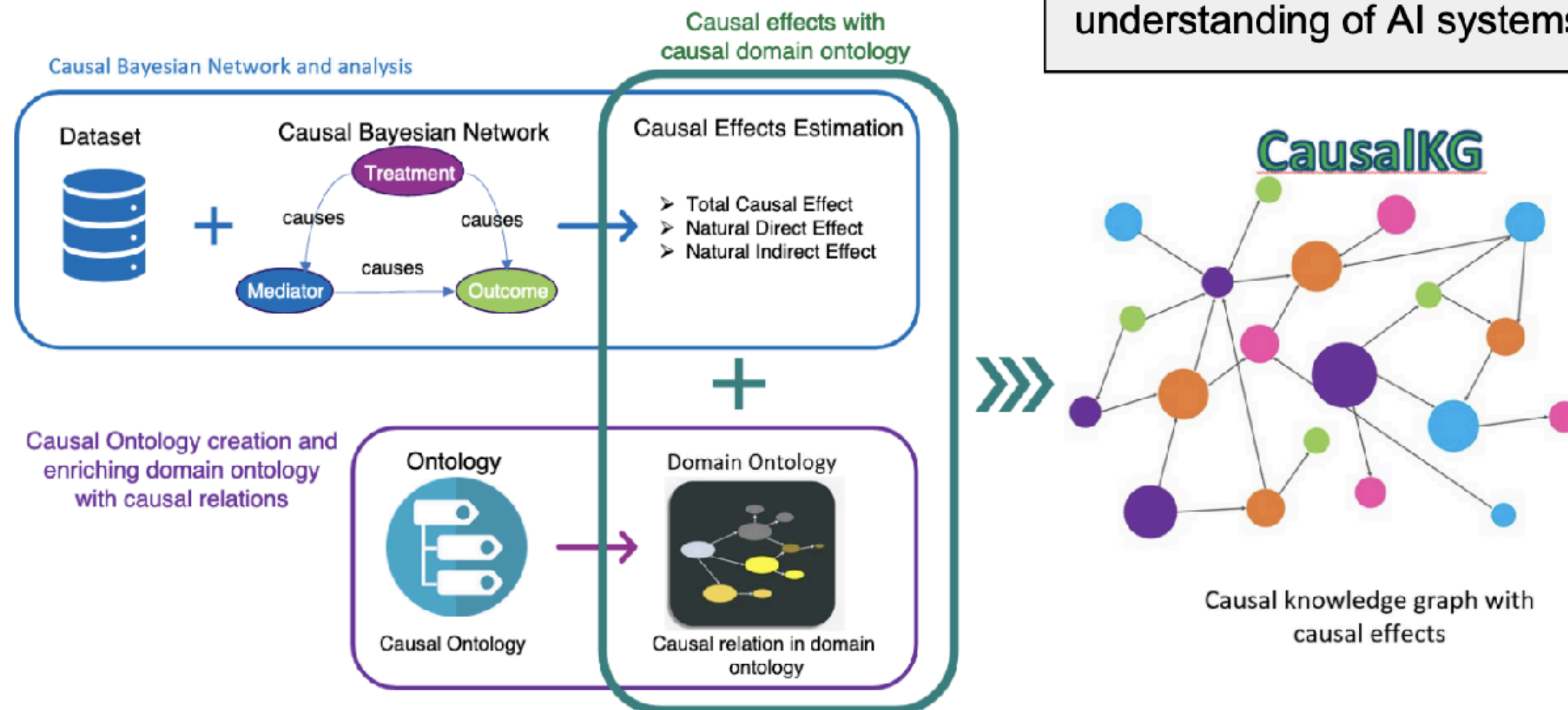
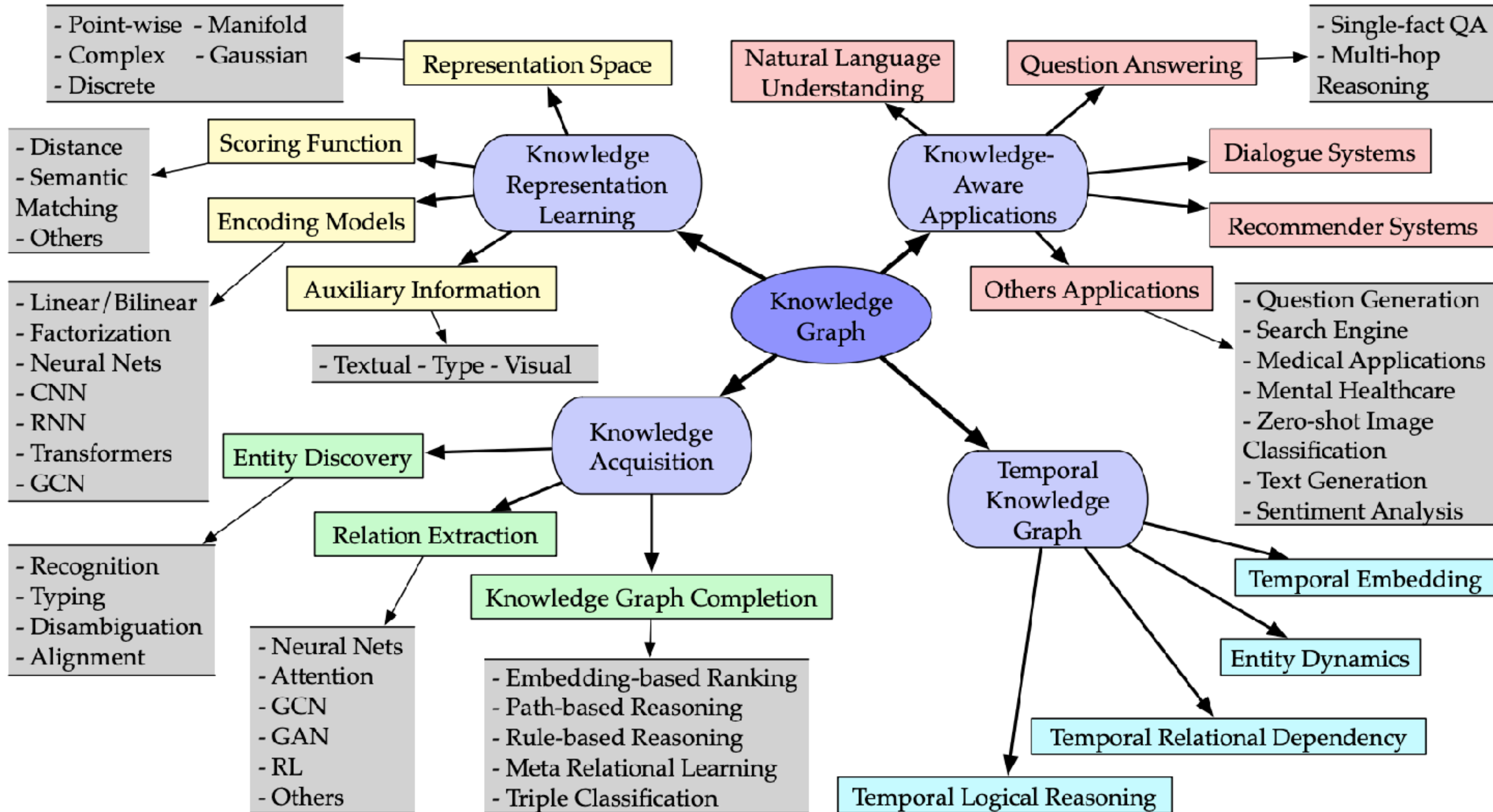


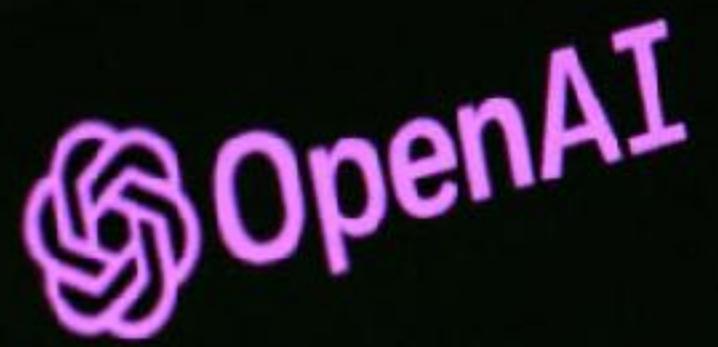
Figure 2. Causal Knowledge Graph Framework consists of three main steps, i) a Causal Bayesian Network and a domain-specific observational dataset, ii) Causal Ontology creation and enriching the domain ontology with causal relationships, and iii) Estimating the causal effects of the treatment, mediator, and outcome variable in the domain for a given context



# 190 More research on knowledge graphs



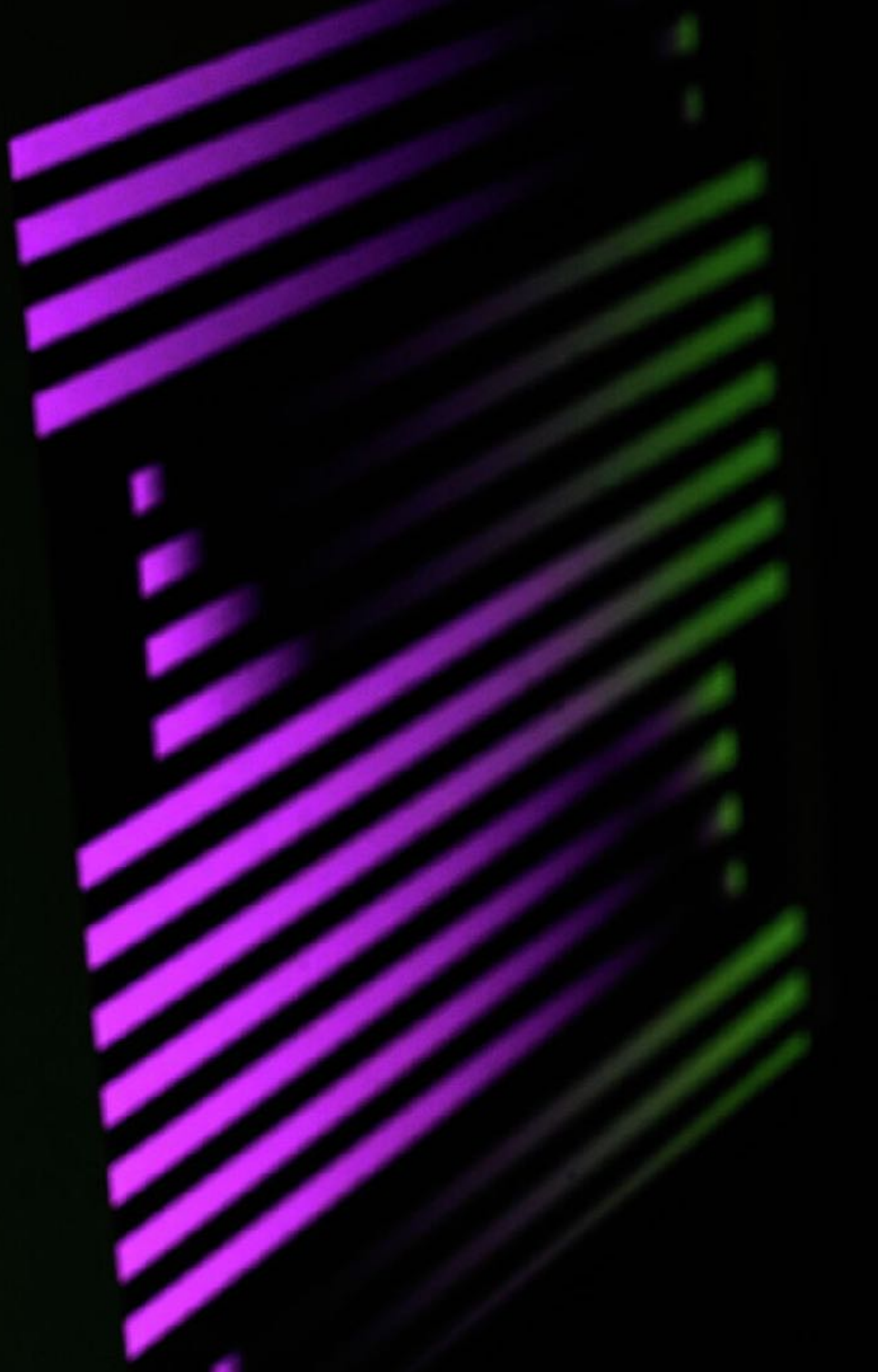




# ChatGPT and KG

ChatGPT: Optimizing  
Language Models  
for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for GPT to answer followup questions, admit its mistakes, correct premises, and reject inappropriate ones. InstructGPT is a sibling model to InstructGPT, which is trained to follow instructions in a prompt and provide a





## 192 Large Language Models: ChatGPT and GPT-4



Hi ChatGPT, can you briefly yourself and GPT-4?



Hello! I'm ChatGPT, a large language model designed by OpenAI. I was trained using the GPT-3.5 architecture, which is a variant of the GPT-3 architecture. As a language model, I'm capable of generating human-like responses to text-based inputs, thanks to my extensive training on a diverse range of text data.




As for GPT-4, I'm not sure about its specific details as it has not been officially announced or released yet. However, based on OpenAI's previous developments with GPT models, it's expected to be an even more powerful language model than its predecessors. It may have a larger training dataset, improved training algorithms, and potentially new features that enhance its language understanding and generation capabilities. Nonetheless, this is all speculation, and we'll have to wait for official announcements from OpenAI for more accurate information.

<https://chat.openai.com/chat>



# Large Language Models: ChatGPT and GPT-4

 **GPT-4** >

### ChatGPT

ChatGPT, or **Chatbot Generalized Pre-trained Transformer**, is an advanced **language model** developed by **OpenAI**. It builds on the success of the **GPT series of language models** and is specifically designed for generating human-like conversational responses. ChatGPT is capable of understanding context, answering questions, and providing relevant information in a coherent and fluent manner. It is widely used in various applications, including **customer support**, **virtual assistants**, **content creation**, and more. The model is trained on **diverse text data** from the internet, which allows it to generate **contextually appropriate responses** in a wide range of domains.

### GPT-4

GPT-4, the fourth iteration of the Generalized Pre-trained **Transformer series**, is a hypothetical and more advanced version of the GPT-3 model. As GPT-3 is the state-of-the-art language model at the time of this response's **knowledge cutoff** in September 2021, GPT-4 represents a potential future development in the field. Like its predecessors, GPT-4 would be expected to showcase even greater performance and capabilities in **natural language understanding**, generation, and other language-related tasks. With each iteration of the GPT series, the models typically become more powerful, showcasing better **context understanding**, improved coherence, and a wider range of applications.

👍 Like
👎 Dislike
🔗 Share

**Ask GPT-4**

Chat
Compose
Insights
⋮
✕

Can you briefly introduce yourself, ChatGPT, and GPT-4?

- ✓ Searching for: **introduction to ChatGPT and GPT-4**
- ✓ Searching for: **introduction to ChatGPT and GPT-4 wikipedia**
- ✓ Searching for: **introduction to ChatGPT and GPT-4 openai**
- ✓ Generating answers for you...

ChatGPT is a conversational AI model developed by OpenAI that can generate human-like responses to text-based inputs. It is a member of the generative pre-trained transformer (GPT) family of language models <sup>1</sup>. ChatGPT Plus is a version of ChatGPT based on GPT-4, which was released on March 14, 2023 <sup>2</sup>. GPT-4 is a large multimodal model that accepts image and text inputs and emits text outputs <sup>3</sup>. It exhibits human-level performance on various professional and academic benchmarks <sup>3</sup>. It was created by OpenAI and is the fourth in its GPT series <sup>2</sup>.

**Learn more:**

- 1. [en.wikipedia.org](https://en.wikipedia.org)
- 2. [en.wikipedia.org](https://en.wikipedia.org) +11 more

**Ask Bing**



# Language Models are Few-Shot Learners

Tom B. Brown\*    Benjamin  
Jared Kaplan†    Prafulla Dhariwal  
Amanda Askell    Sandhini Agarwal  
Rewon Child    Aditya Ramesh  
Christopher Hesse    Mark Ch  
Benjamin Chess  
Sam McCandlish    Alex

**GPT-3**

# Training language models to follow instructions with human feedback

Long Ouyang\*    Jeff Wu\*  
Pamela Mishkin\*    Chong Zh  
John Schulman    Jacob Hilt  
Amanda Askell†  
Jan Leike\*

**InstructGPT**

# GPT-4 Technical Report

We report the development of accept image and text inputs humans in many real-world s on various professional and ac bar exam with a score around based model pre-trained to pre alignment process results in in adherence to desired behavior infrastructure and optimization range of scales. This allowed performance based on models GPT-4.

**GPT-4**

# Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck    Varun Chandrasekaran    Ronen Eldan    Johannes Gehrke  
Eric Horvitz    Ece Kamar    Peter Lee    Yin Tat Lee    Yuanzhi Li    Scott Lundberg  
Harsha Nori    Hamid Palangi    Marco Tulio Ribeiro    Yi Zhang

Microsoft Research

## Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis

**GPT-4**



# Do we still need knowledge graph?



Do we still need Knowledge Graph as ChatGPT and GPT-4 are so powerful?



Yes, Knowledge Graphs can still be valuable even with the advancements of language models like ChatGPT and future models like GPT-4. While language models like ChatGPT and GPT-4 are capable of generating responses to a wide range of questions and inputs, they may still lack the depth and breadth of knowledge that a well-structured Knowledge Graph can provide.



Knowledge Graphs can organize and represent complex information in a structured and machine-readable format, which can be used to augment the language models' understanding of certain concepts and relationships between them. Knowledge Graphs can also help address issues with ambiguity and context that may arise in language-based interactions.

Moreover, Knowledge Graphs can be used for other applications beyond just question-answering, such as semantic search, recommendation systems, and data integration across different domains and sources.

In summary, while ChatGPT and GPT-4 are powerful language models, Knowledge Graphs can still play a valuable role in enhancing their understanding of complex information and improving the accuracy of their responses.





# Knowledge graph for large language models

## KG can help address a few of the major shortcomings of language models

- Memory
  - Knowledge graphs can provide accurate information about any specific subject, which ChatGPT can incorporate into its text generation, resulting in more informative and accurate output. It helps reducing the *hallucination* problem of language models.
- Logic and reasoning
  - Knowledge graphs provide a structured way of linking concepts together, allowing ChatGPT to make inferences and draw conclusions based on the information it has been provided, thus *improving the reasoning capabilities of the model*.
- Interpretability
  - By observing what memories are accessed, we can *understand sources of resulting statements* and information flow.
  - Alternatively, by observing how the language model processes and categorizes information, we can understand the inherent structure of the information learned by the raw language model.
- Computational resource use for training
  - Knowledge Graphs can store vast amount of data, which allows chatbot providers to *scale* their chatbot to handle large numbers of users and queries, without continuously retraining the model parameters.
- Personalization
  - By using knowledge graph, chatbot providers can create *personalized chatbot experiences* for their customers, as the chatbot can use the user's browsing history, interests, and demographic data to provide personalized recommendations and responses.



# 198 Large language models for knowledge graph

## **Zero-Shot Information Extraction via Chatting with ChatGPT**

**Xiang Wei<sup>1</sup>, Xingyu Cui<sup>1</sup>, Ning Cheng<sup>1</sup>, Xiaobin Wang<sup>2</sup>, Xin Zhang, Shen Huang<sup>2</sup>,  
Pengjun Xie<sup>2</sup>, Jinan Xu<sup>1</sup>, Yufeng Chen<sup>1</sup>, Meishan Zhang, Yong Jiang<sup>2</sup>, and Wenjuan Han<sup>1</sup>**

<sup>1</sup> Beijing Jiaotong University, Beijing, China

<sup>2</sup> DAMO Academy, Alibaba Group, China

<https://arxiv.org/pdf/2302.10205.pdf>

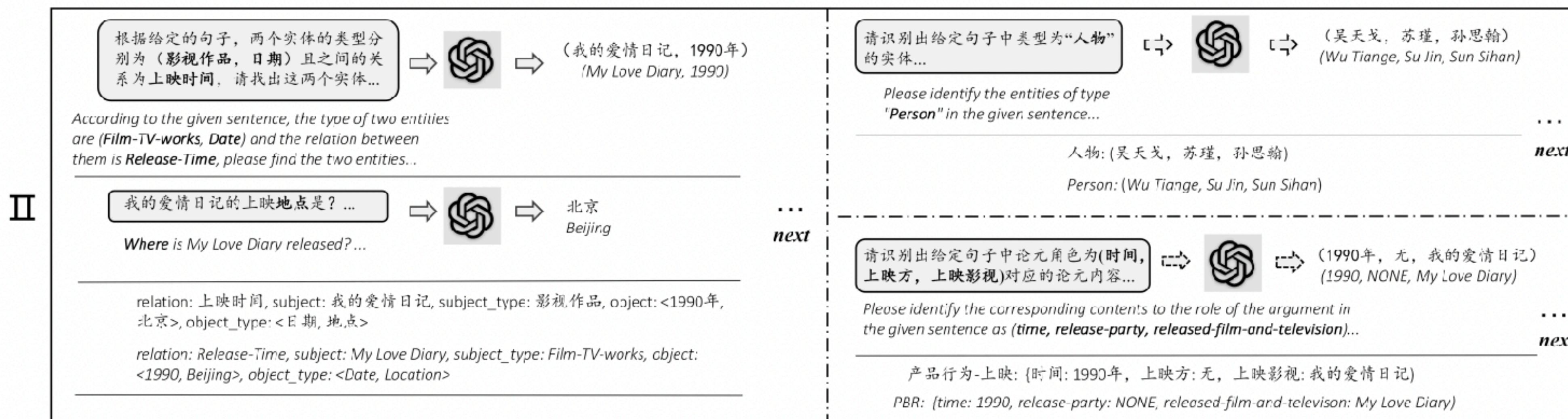
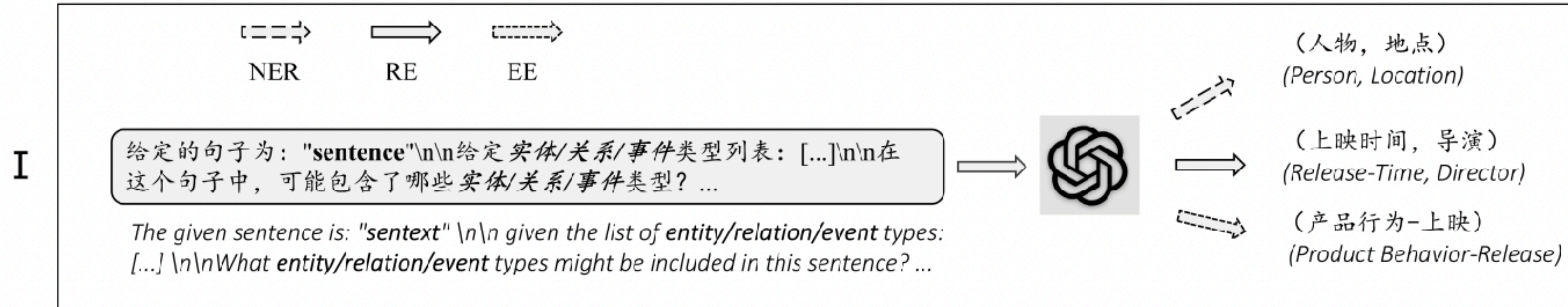


# Large language models for knowledge graph

## IE with prompt engineering

**sentence:** 《我的爱情日记》是1990年在北京上映的中国剧情电视剧，由吴天戈执导，苏瑾、孙思翰等主演

"My Love Diary" is a Chinese TV series released in Beijing in 1990, directed by Wu Tiange and starred by Su Jin and Sun Sihan.





# 200 Large language models for knowledge graph

## Distill knowledge from LLM to KG

### Symbolic Knowledge Distillation: from General Language Models to Commonsense Models

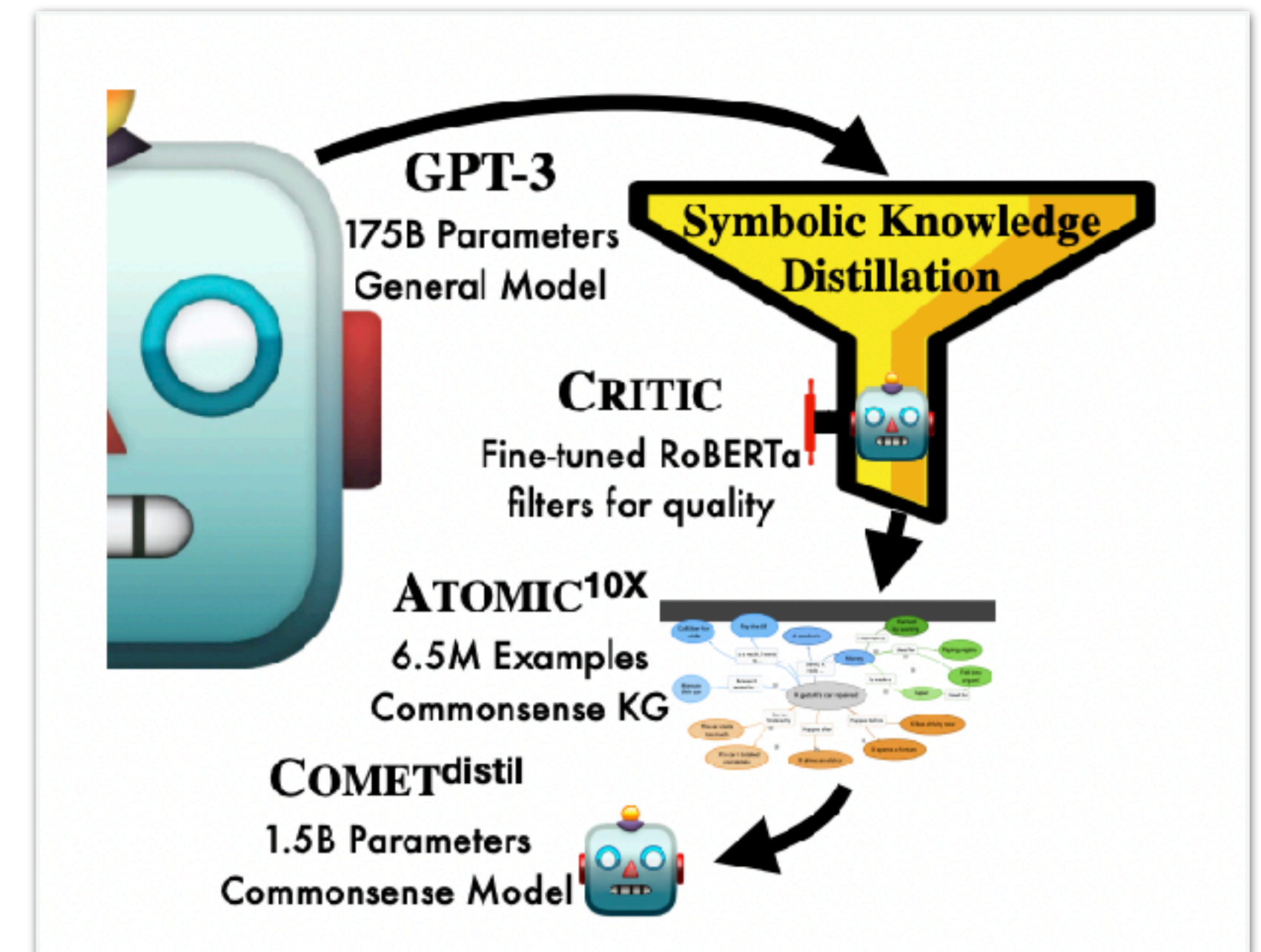
Peter West<sup>†‡\*</sup> Chandra Bhagavatula<sup>‡</sup> Jack Hessel<sup>‡</sup> Jena D. Hwang<sup>‡</sup>

Liwei Jiang<sup>†‡</sup> Ronan Le Bras<sup>‡</sup> Ximing Lu<sup>†‡</sup> Sean Welleck<sup>†‡</sup> Yejin Choi<sup>†‡\*</sup>

<sup>†</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington

<sup>‡</sup>Allen Institute for Artificial Intelligence

<https://aclanthology.org/2022.naacl-main.341.pdf>



# Demo part 3

## ChatGPT for KG



# Demo of our tutorial

## ◎ Demo part 1: Pipeline of Knowledge Graph Construction

- [https://colab.research.google.com/drive/1FEmL6hE9UK\\_Do7dOqrgbKowVaG4WYdUO?usp=sharing](https://colab.research.google.com/drive/1FEmL6hE9UK_Do7dOqrgbKowVaG4WYdUO?usp=sharing)

## ◎ Demo part 2: Knowledge-Graph for Question Answering

- <https://github.com/MrBearWithHisSword/KGQA-COVID>

## ◎ Demo part 3: ChatGPT for Knowledge Graph

- <https://colab.research.google.com/drive/1hSGcFQ2QCQSzhBYFfejJuyq0hue1GbYw?usp=sharing>

- [What is Information Extraction? \(n.d.\). Ontotext.](#)
- [Jurafsky, D., & Martin, J. H. \(2018\). Speech and language processing.](#)
- [Li, J., Sun, A., Han, J., & Li, C. \(2020\). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34\(1\), 50-70.](#)
- [Nadeau, D., Turney, P. D., & Matwin, S. \(2006, June\). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In Conference of the Canadian society for computational studies of intelligence \(pp. 266-277\). Springer, Berlin, Heidelberg](#)
- [Fariz Darari. Knowledge graphs mini handbook.](#)
- [BLOMQVIST, Eva, D'AMATO, Claudia, DE MELO, Gerard, et al. Knowledge Graphs.](#)
- [Pujara, J., & Singh, S. \(2018, February\). Mining knowledge graphs from text. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining \(pp. 789-790\).](#)
- [Gao, Y., Liang, J., Han, B., Yakout, M., & Mohamed, A. \(2018\). Building a large-scale, accurate and fresh knowledge graph. KDD-2018, Tutorial, 39, 1939-1374.](#)
- [Marta Maślankowska, Paweł Mielniczuk. Intro to coreference resolution in NLP](#)
- [Sundar V. Entity Linking: A primary NLP task for Information Extraction](#)
- [Raiman, J., & Raiman, O. \(2018, April\). Deeptype: multilingual entity linking by neural type system evolution. In Proceedings of the AAAI Conference on Artificial Intelligence \(Vol. 32, No. 1\).](#)
- [Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C. \(2020\). Neural entity linking: A survey of models based on deep learning. arXiv preprint arXiv:2006.00575.](#)
- [Xiang Ren. Entity Linking and Coreference Resolution](#)
- [Laura Dietz. Entity Linking](#)
- <https://github.com/tomhartke/knowledge-graph-from-GPT-3>
- <https://www.ontotext.com/blog/why-should-you-combine-chatgpt-with-knowledge-graphs/#:~:text=In%20conclusion%2C%20while%20ChatGPT%20is,and%20informative%20text%20generation%20system.>

## Citation about this slides

- Please cite with the following BibTex entry if these slides are helpful to your research works

```
@misc{Liu2022KGTutorial,  
  author={Bang Liu and Haochen Shi},  
  title = {Knowledge Graphs: Extracting and Leveraging Meaning from Unstructured Text},  
  howpublished = {\url{http://www-labs.iro.umontreal.ca/~liubang/files/KG_tutorial_ivado22.pdf}},  
  year = {2022}  
}
```



# Thanks! Q&A

**Bang Liu**

**Email:** [bang.liu@umontreal.ca](mailto:bang.liu@umontreal.ca)

**Homepage:** <http://www-labs.iro.umontreal.ca/~liubang/>

**Github:** <https://github.com/BangLiu/>



# IVADO



**CANADA  
FIRST**  
RESEARCH  
EXCELLENCE  
FUND

**APOGÉE  
CANADA**  
FONDS  
D'EXCELLENCE  
EN RECHERCHE

Québec 

Canada 