Knowledge Graphs: Extracting and Leveraging Meaning from Unstructured Text

Teacher and TA: Bang Liu, Haochen Shi
University of Montreal, DIRO & Mila

NLP Workshop Series, IVADO & CLIQ-AI

2 Introduction about the teacher and teaching assistant

- **Bang Liu** 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 (NLP), data mining, multimodal and multitask learning, AI + X (e.g., animation, VR, health)**.

- **Haochen Shi** 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 recevied 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 learning, and resource-economical deep learning**.

3 Tutorial outline

- **Part I: Introduction to knowledge graph**
  - What is a knowledge graph?
  - Why are knowledge graphs important?
  - Success stories of knowledge graphs

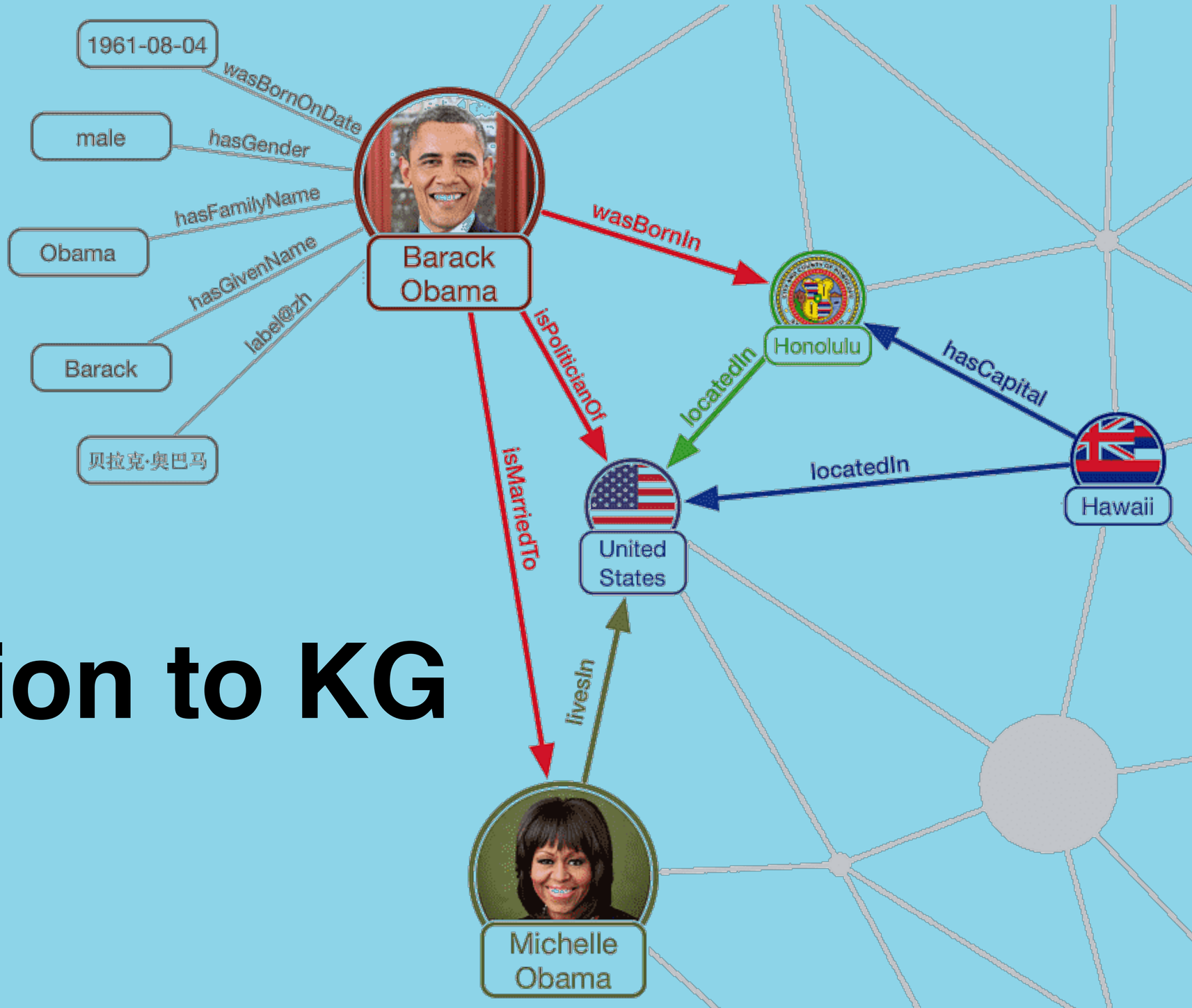
- **Part II: Semantic Web Knowledge Graph**
  - W3C Semantic Web
  - RDF, SPARQL, etc.

- **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**

Introduction to KG



5 Suppose you are searching…

Query: "Honda Civic, Hyundai Elantra"

What he is interested in?

6 Is-A relationship between things

Honda Civic

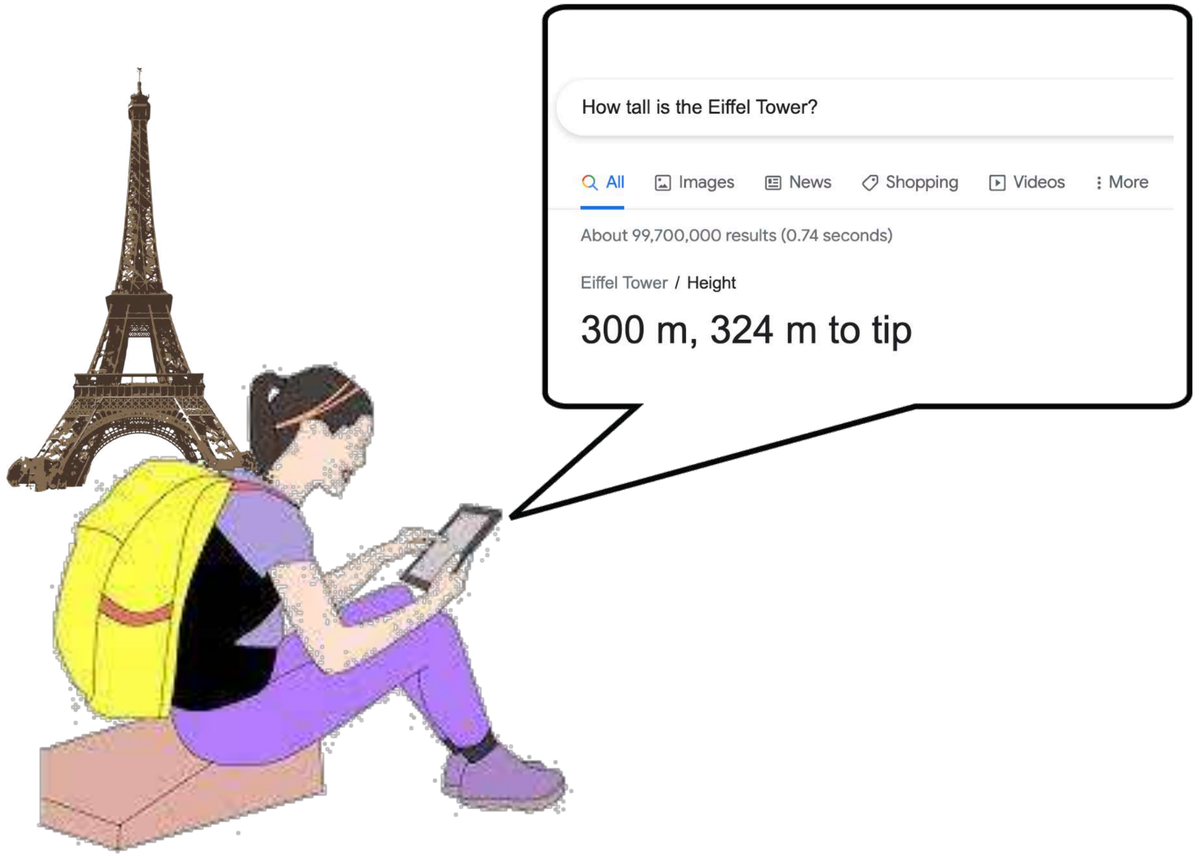
Hyundai Elantra

- **Fuel-efficient cars**
- **Economy cars**

- **Marvel heroes**
- **Revengers**

Concept: a collection of things that share some common attributes

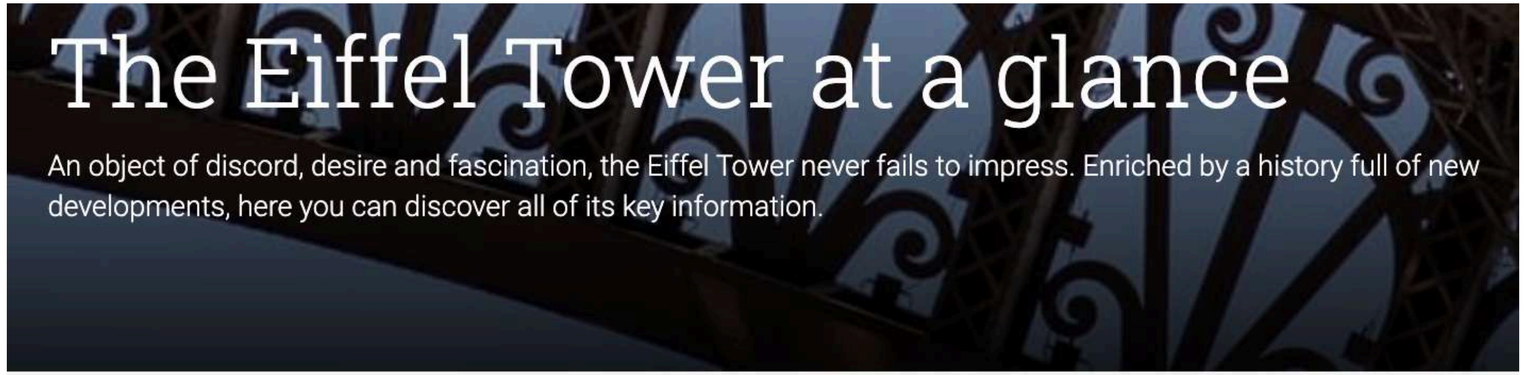
7 Suppose you are asking…



How does Google know?

Question: "How tall is the Eiffel Tower?"

8 Properties of things



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

https://www.toureiffel.paris/en/the-monument/key-figures

9 Suppose you are shopping…

Roll over image to zoom in

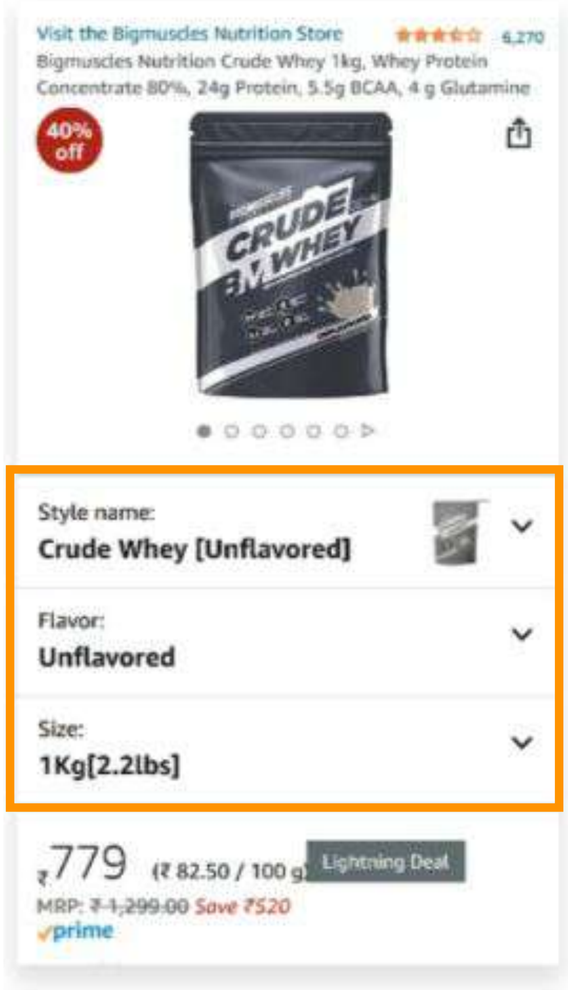
Brand	Cetaphil
Ingredients	Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.
Scent	Fragrance free
Additional Item Information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

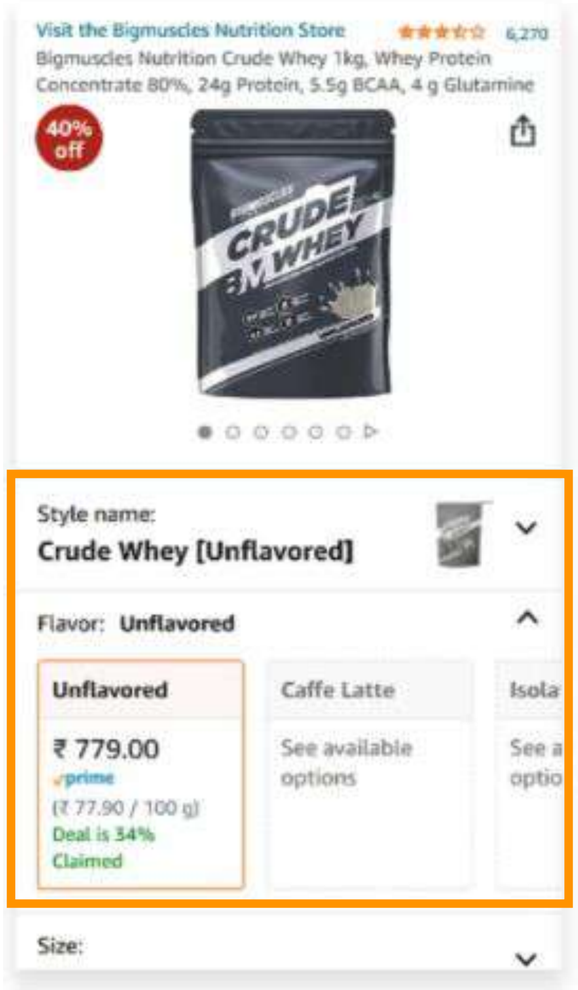
About this item

- 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?

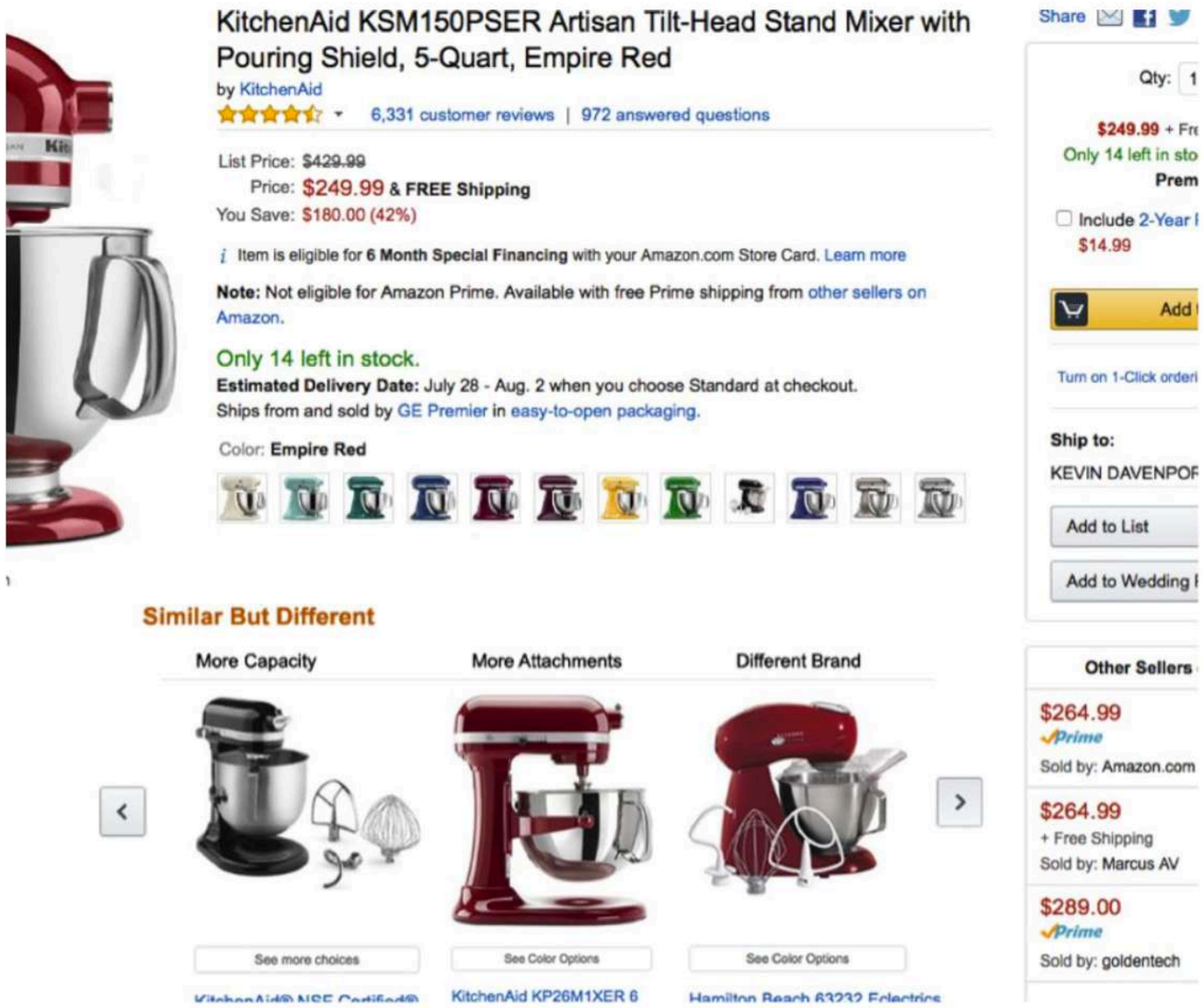
10 Suppose you are shopping…





How does Amazon provide choices?

11 Suppose you are shopping…



How does Amazon improves searching and recommendation?

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	...

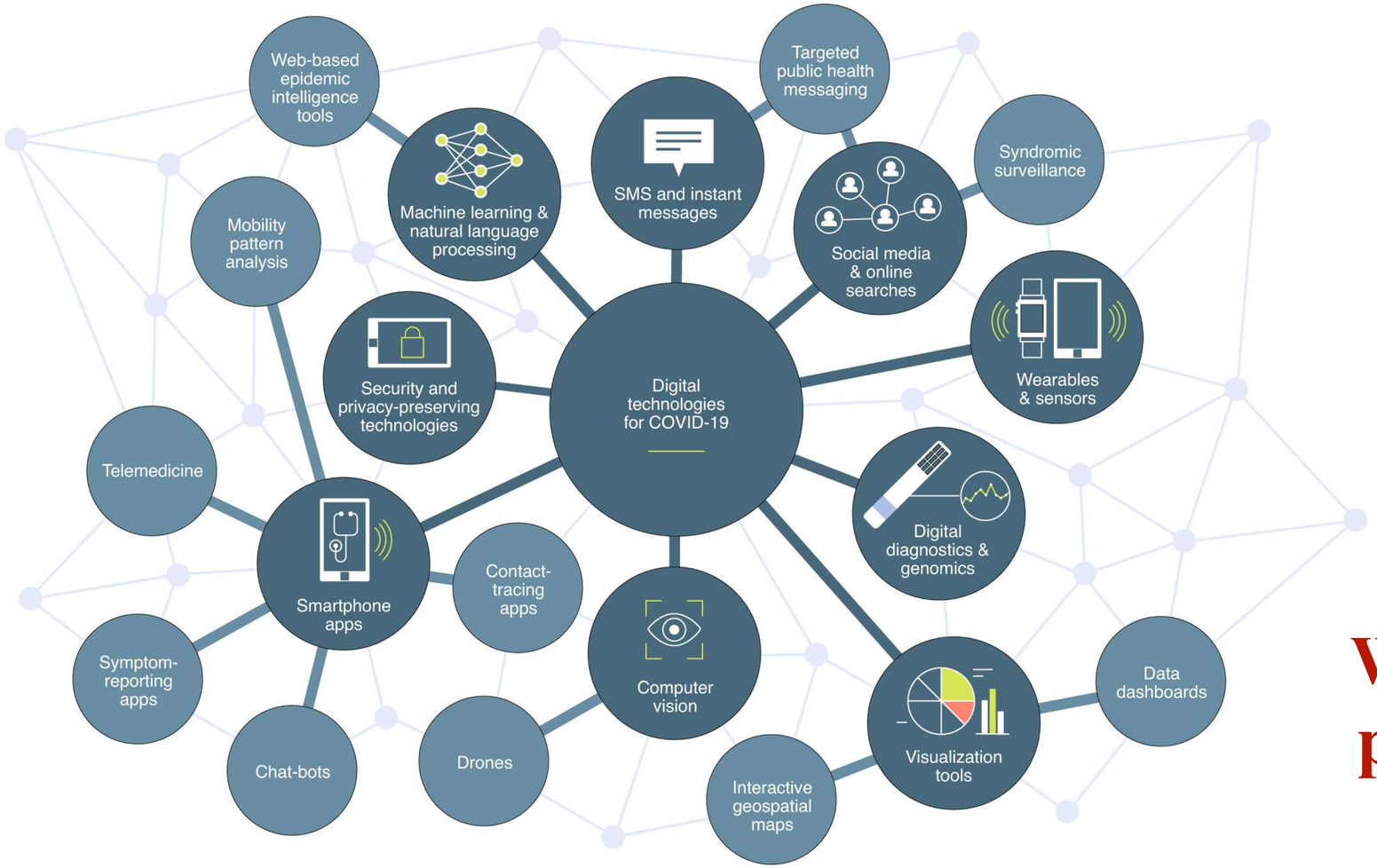
ICD9_Dx_Codes

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…

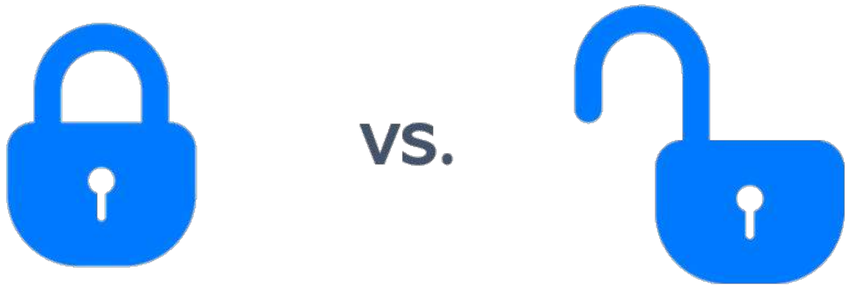
- You would like to build an application to **visualize and analyze COVID-19**



What would you primarily need?

Jobie Budd et al., Digital technologies in the public-health response to COVID-19, nature medicine

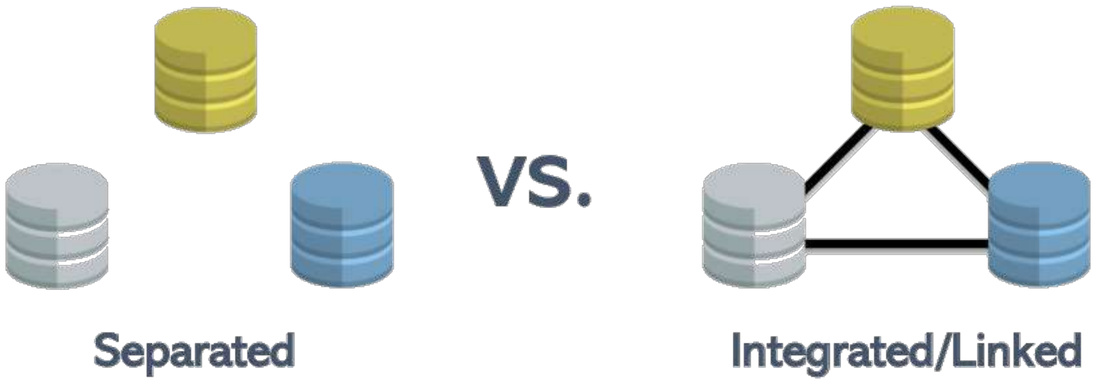
Data



Closed vs. Open Data

Tabular vs. Graph

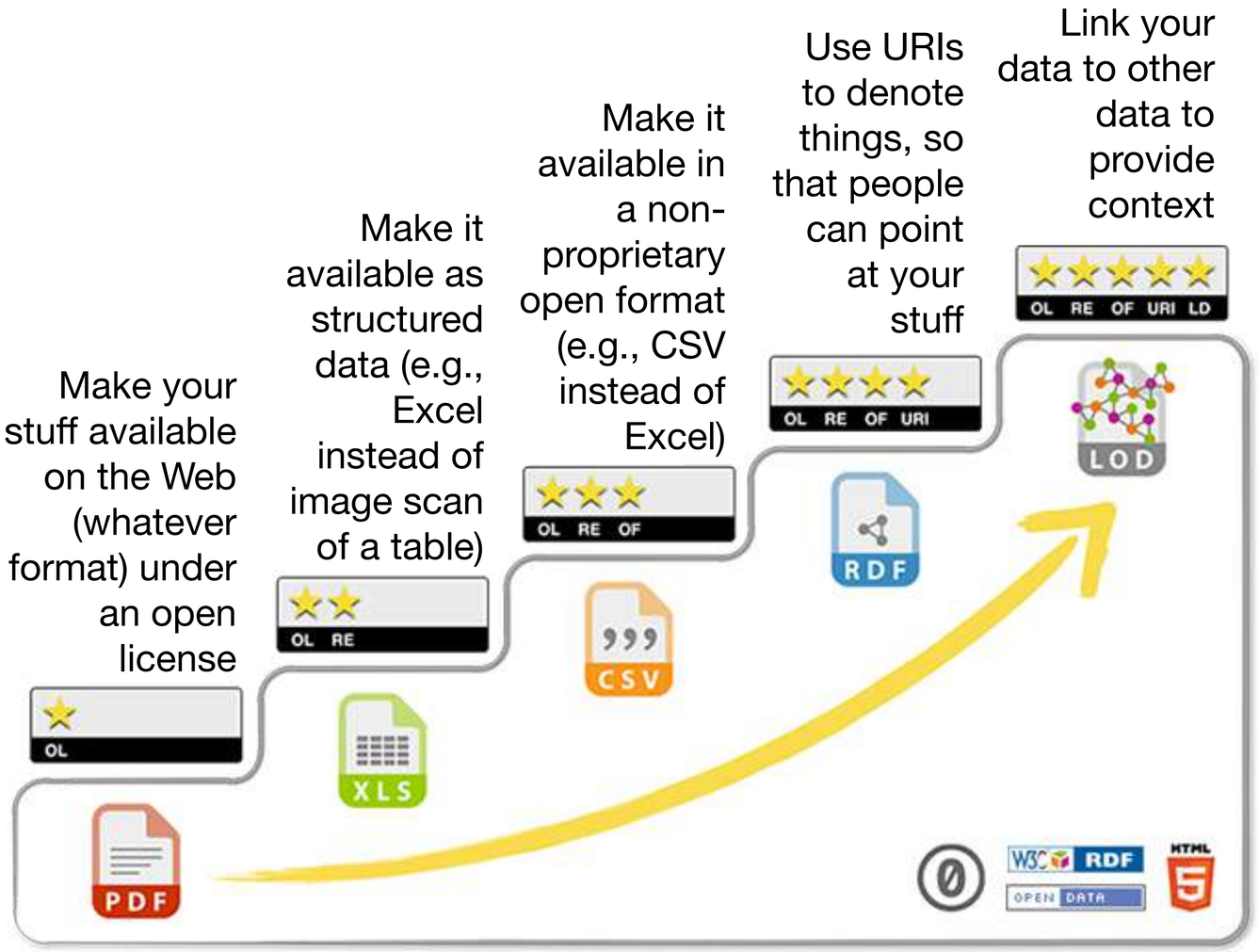
Human Readable vs. Machine Readable



Separated vs. Integrated/Linked

15 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.



https://5stardata.info/en/

16 What is Knowledge

- Plato's definition: Justified true belief

Merriam-Webster 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

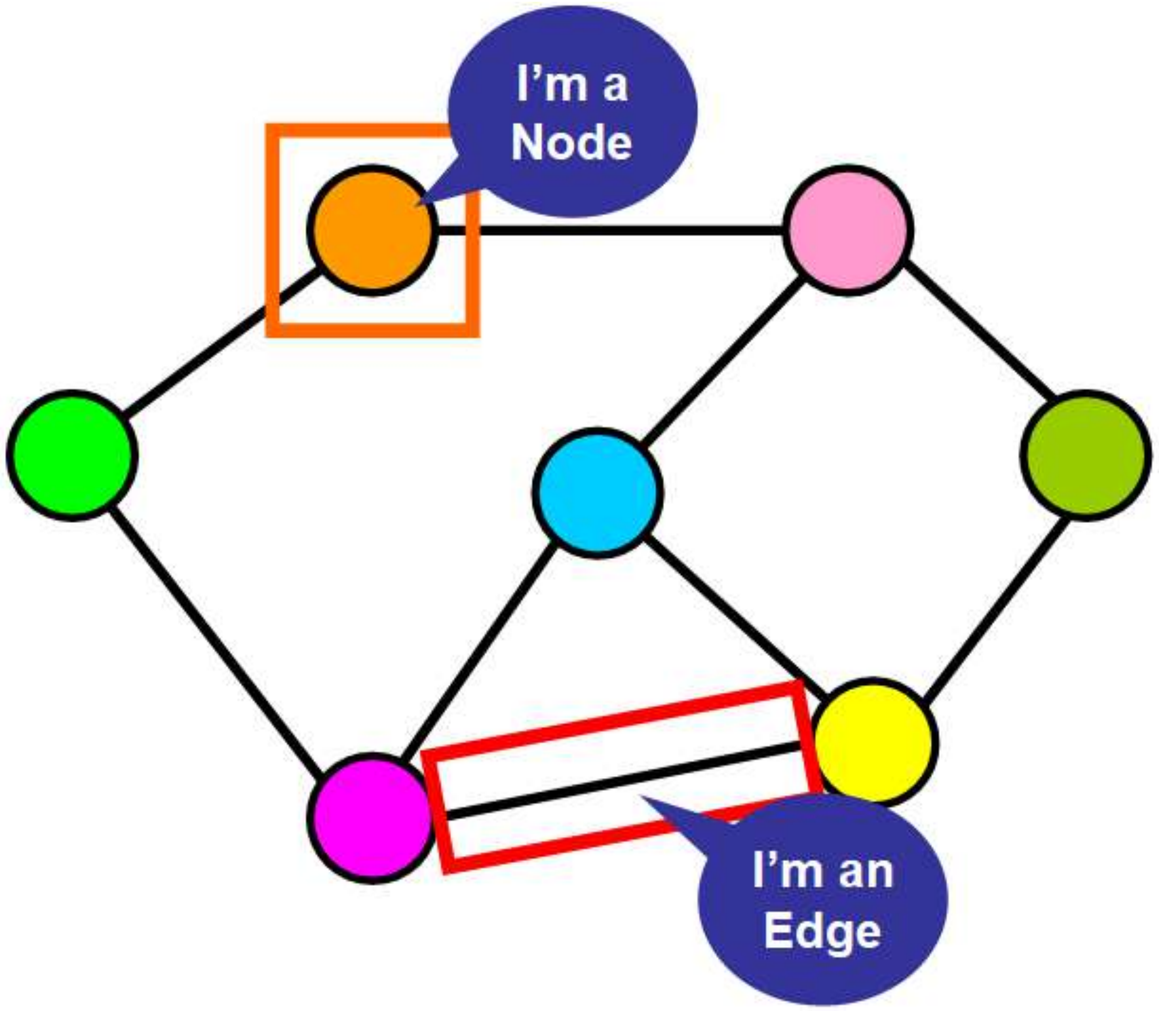
2 **:** awareness of something **:** the state of being aware of something

// 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]

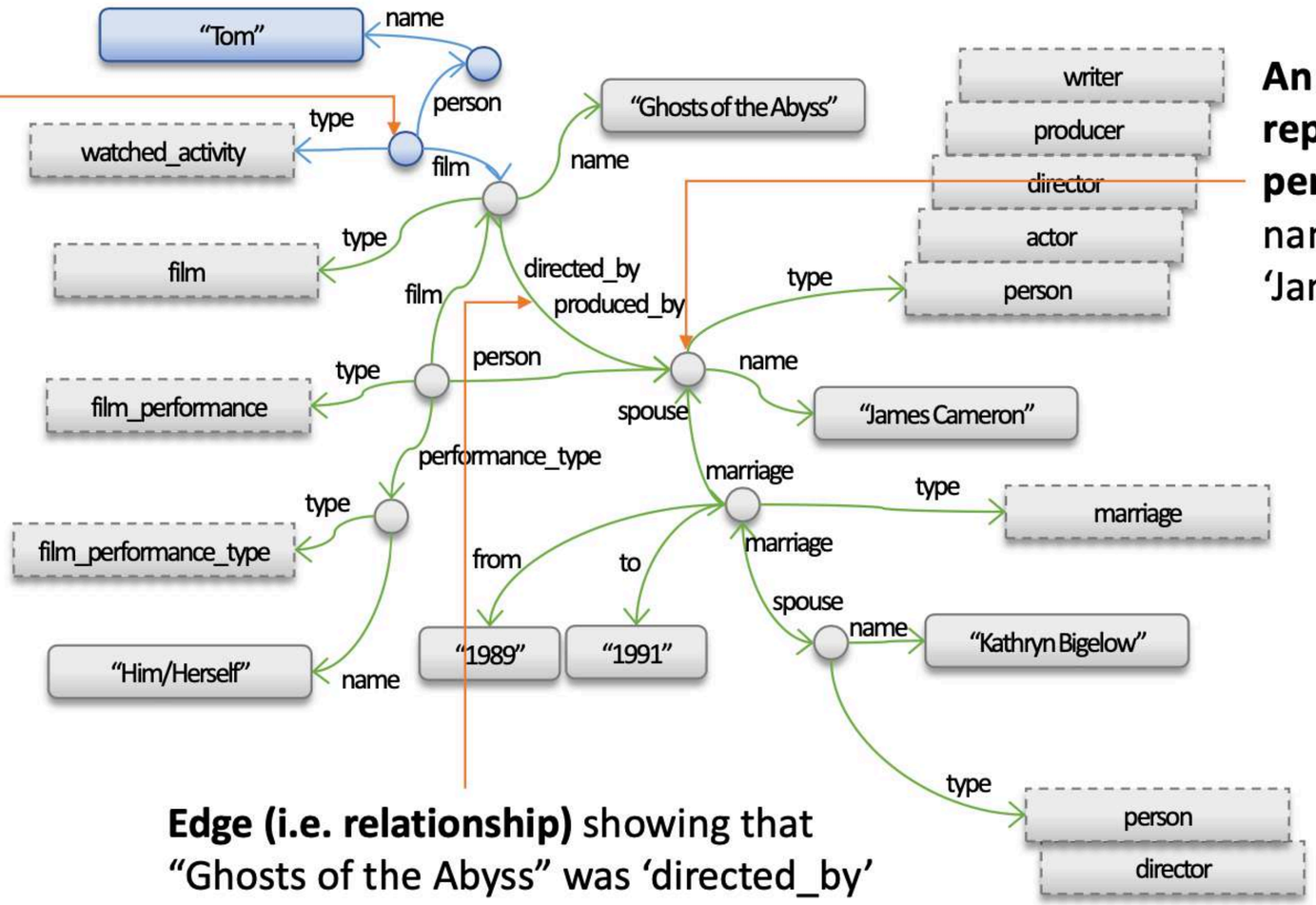
17 What is Graph



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'

Edge (i.e. relationship) showing that "Ghosts of the Abyss" was 'directed_by' and 'produced_by' James Cameron

Key concepts

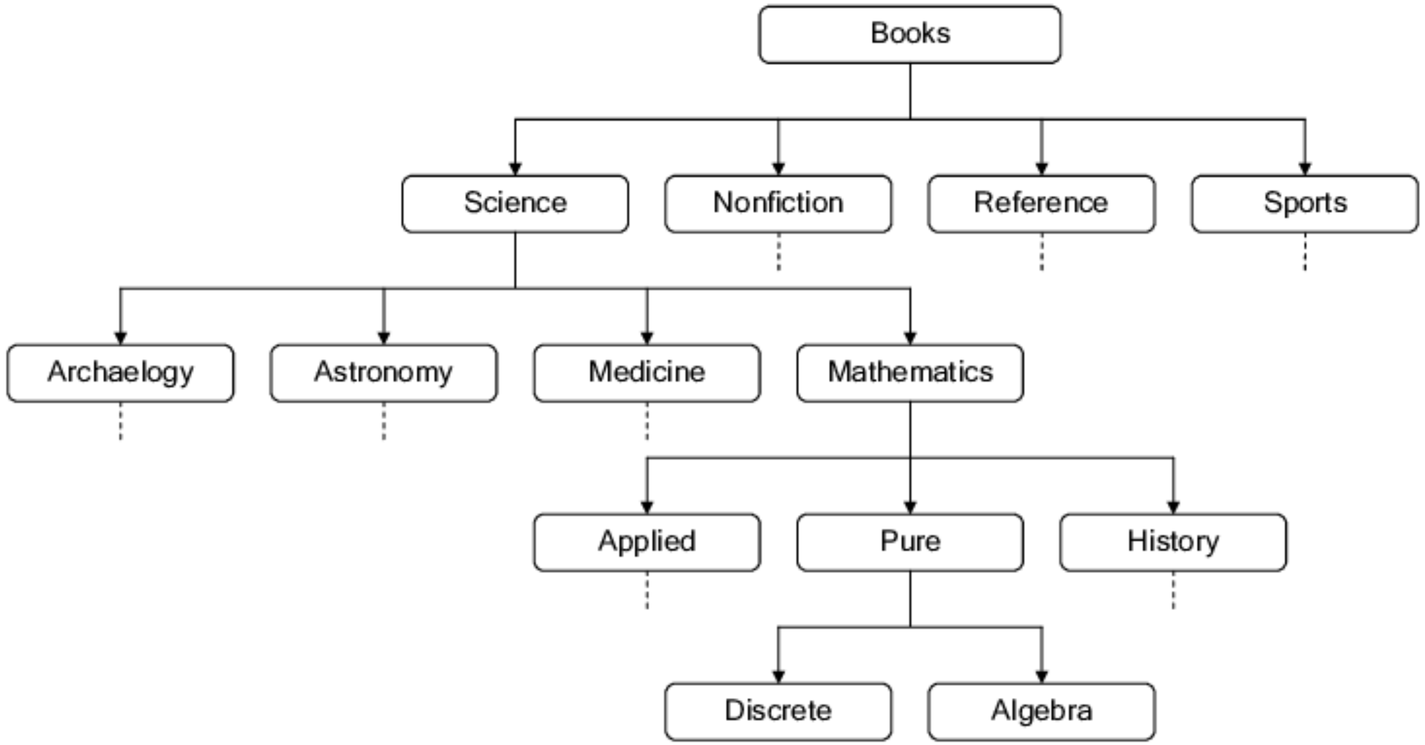
Entity	Represent something in the real world
Edge	Represent relationship
Attribute	Represent something about an entity
Ontology	Definition of possible types of entities, relationships and attributes

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

20 What is a Taxonomy?

- A data taxonomy is the classification of data into categories and sub-categories.

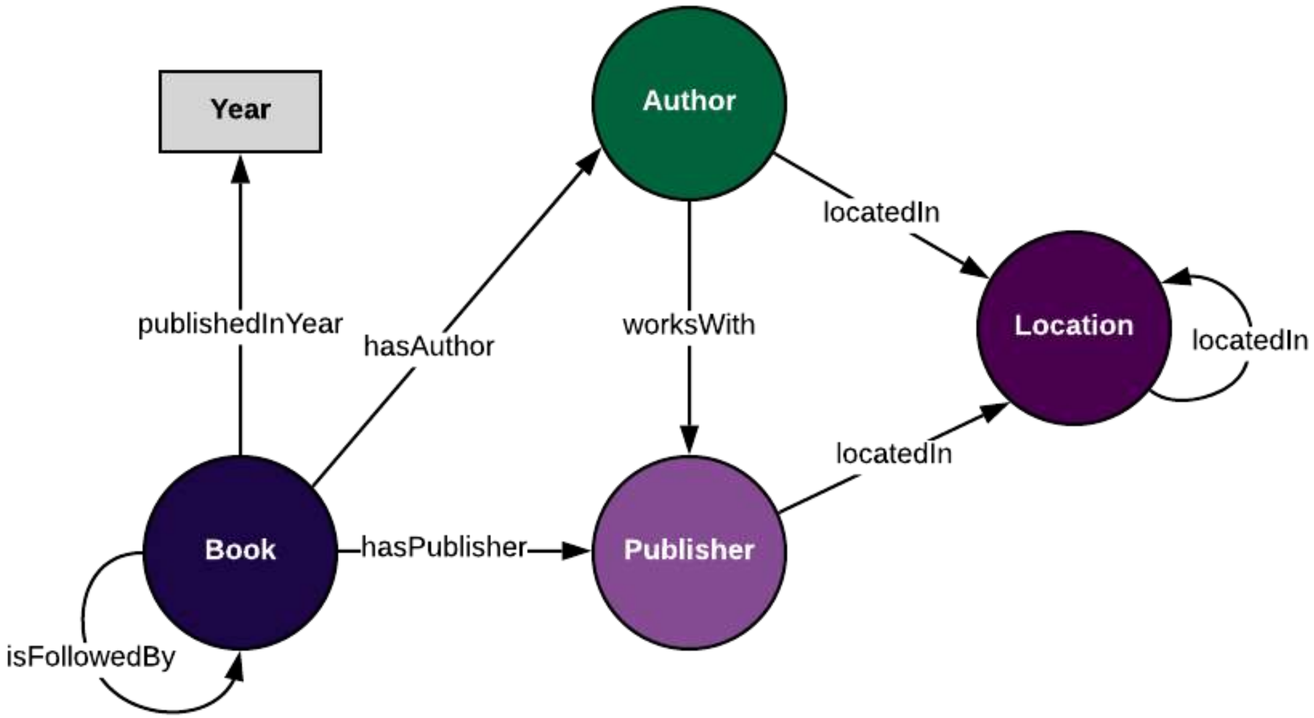
- Taxonomy identifies hierarchical relationships within a category.



21 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.

22 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

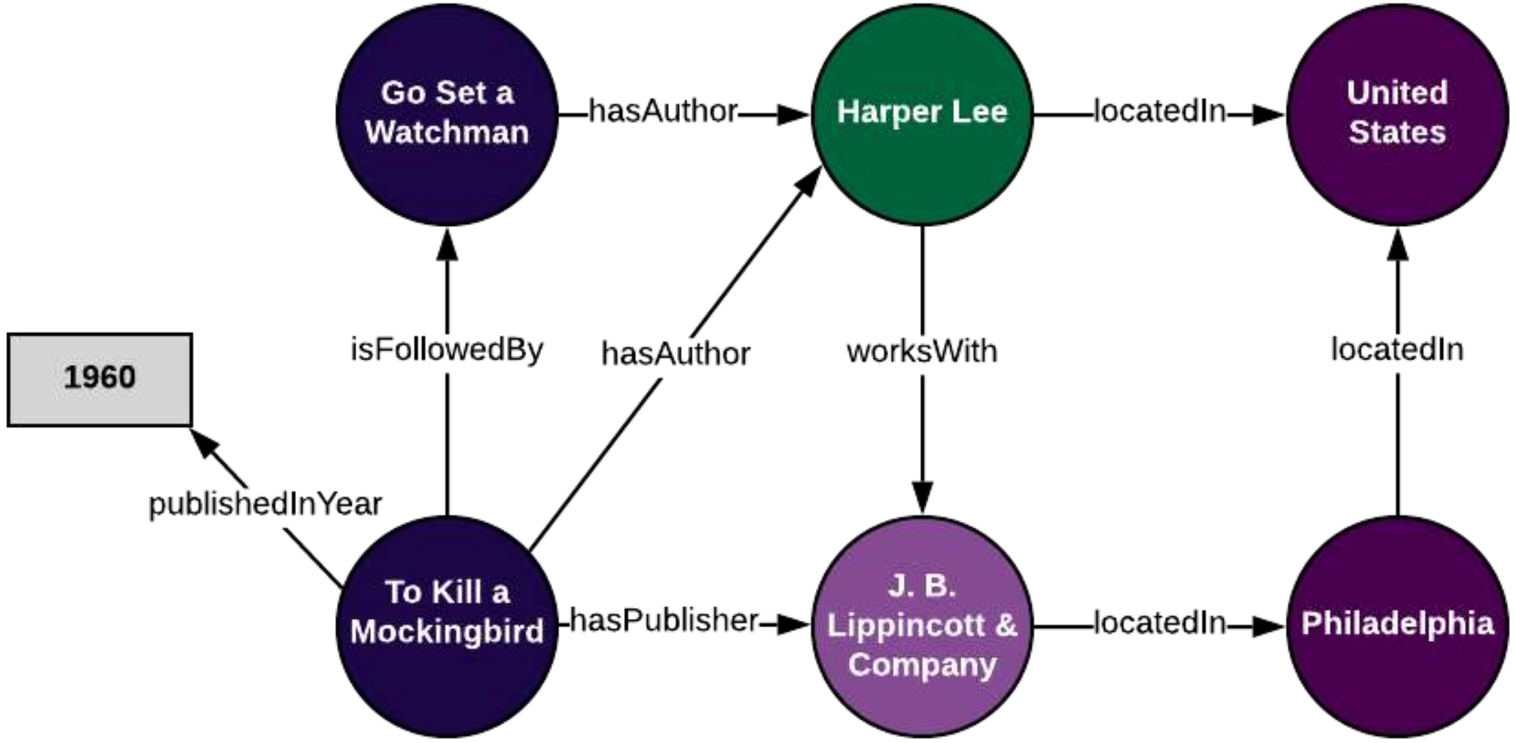
Name	City	Country
J. B. Lippincott & Company	Philadelphia	United States
HarperCollins, LLC	New York City	United States
Heinemann	Portsmouth	United States
New American Library	New York City	United States
Hutchinson	London	United Kingdom

Authors

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

23 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

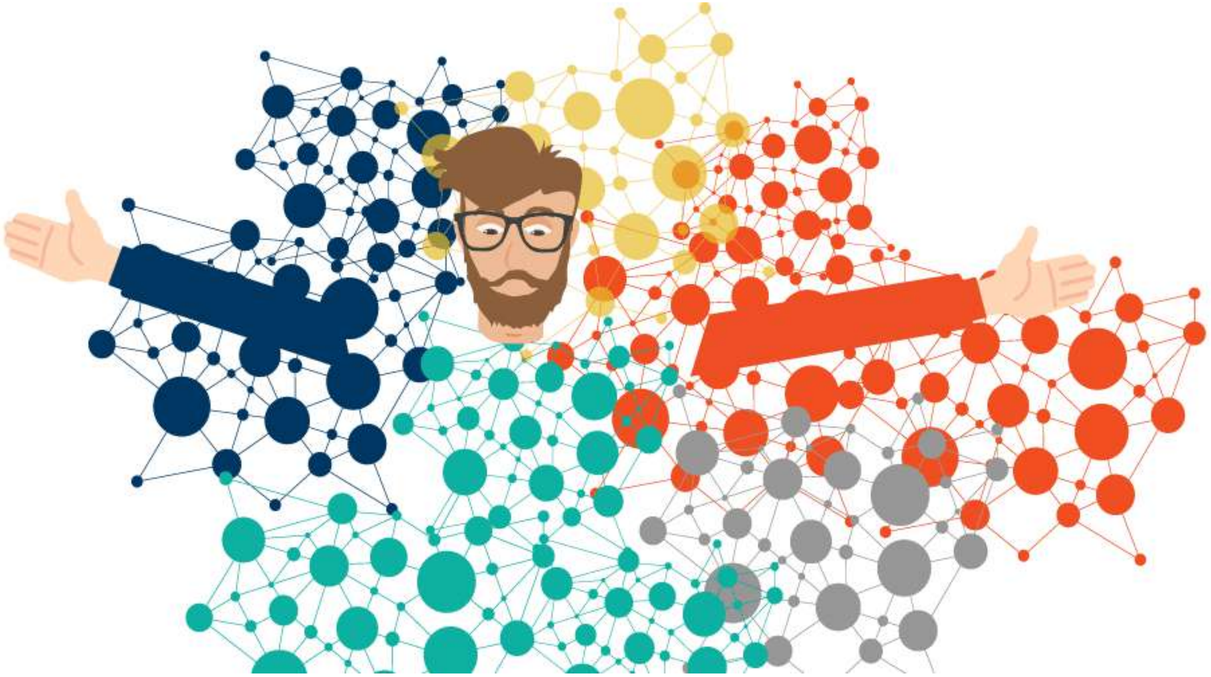
Intuitive domain modelling

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**

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

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
select knowledge

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!

NELL: Never-Ending Language Learning

Recently-Learned Facts

twitter

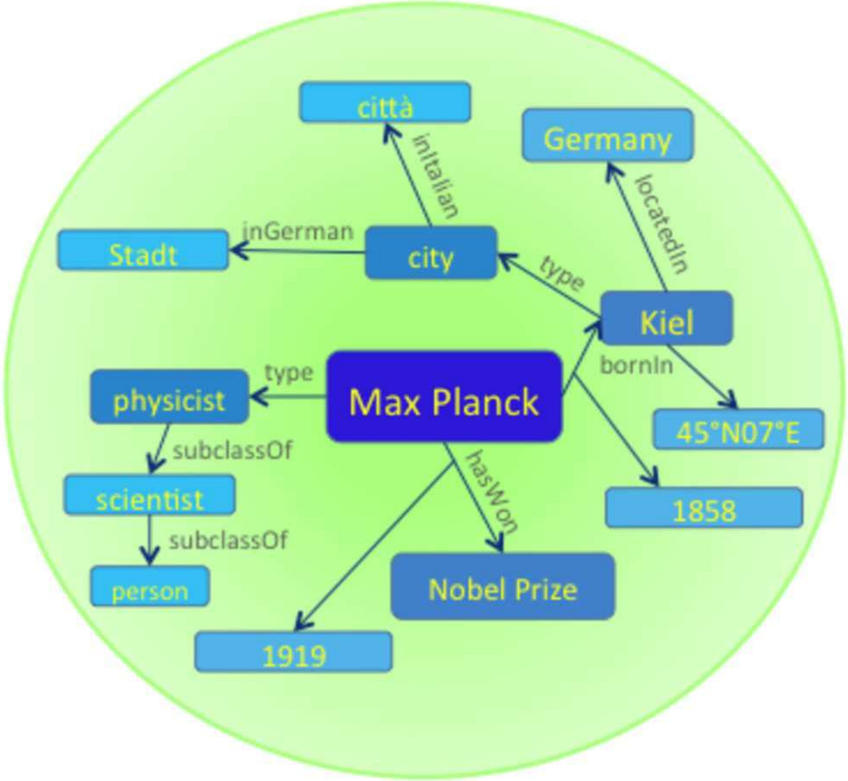
Refresh

instance	iteration	date learned	confidence
high_end_clock is an item found on a wall	1111	06-jul-2018	99.6
movie_edward_scissorhands is a movie	1111	06-jul-2018	100.0
n1999_toyota_camry is a model of automobile	1111	06-jul-2018	94.7
michael_i is a monarch	1111	06-jul-2018	96.7
gonzales_music_festival_1968 is a music festival	1111	06-jul-2018	100.0
roof is a building feature made from the building material steel	1115	03-sep-2018	92.7
tyco_healthcare hired dennis_kozlowski	1112	24-jul-2018	100.0
air_pacific is a company that has an office in the city los_angeles_ca	1112	24-jul-2018	99.2
the sports team yankees was the winner of n1962_world_series	1116	12-sep-2018	100.0
silvestre_herrera died in the city glendale	1116	12-sep-2018	100.0

http://rtw.ml.cmu.edu/rtw/

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 (n)
- en band (n)
- en musician (n)
- en album (n)

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.

Recap: suppose you are searching…

What he is interested in?

Query: "Honda Civic, Hyundai Elantra"

Recap: Is-A relationship between things

Honda Civic

Hyundai Elantra

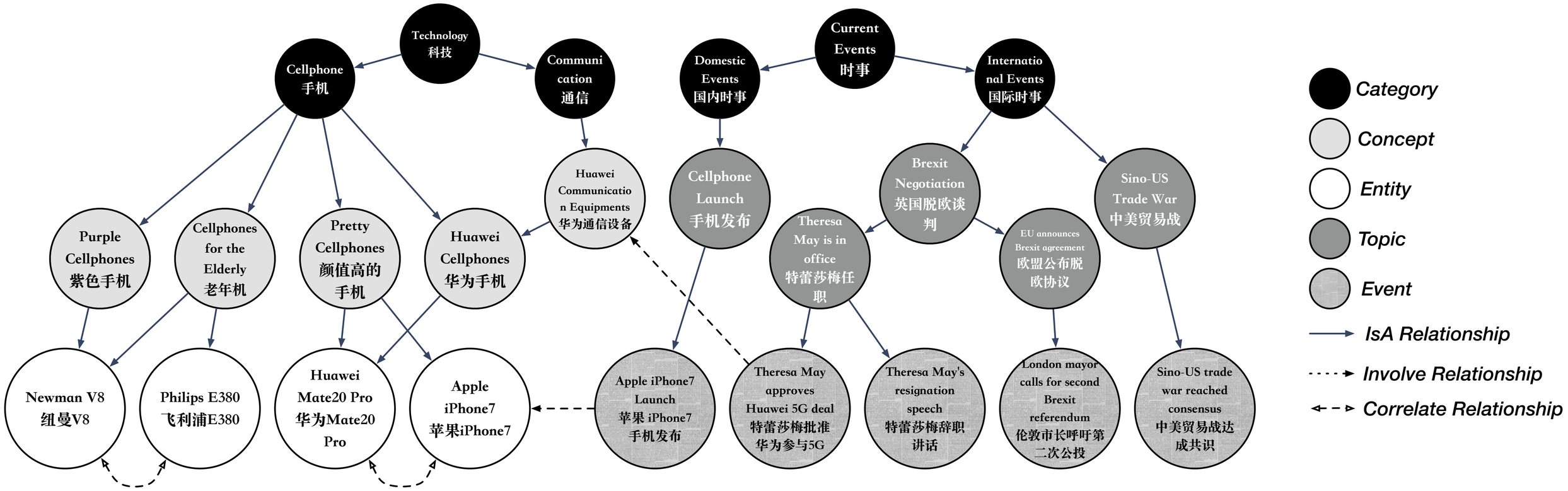
- **Fuel-efficient cars**
- **Economy cars**

- **Marvel heroes**
- **Revengers**

Concept: a collection of things that share some common attributes

Tencent GIANT Attention Ontology

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

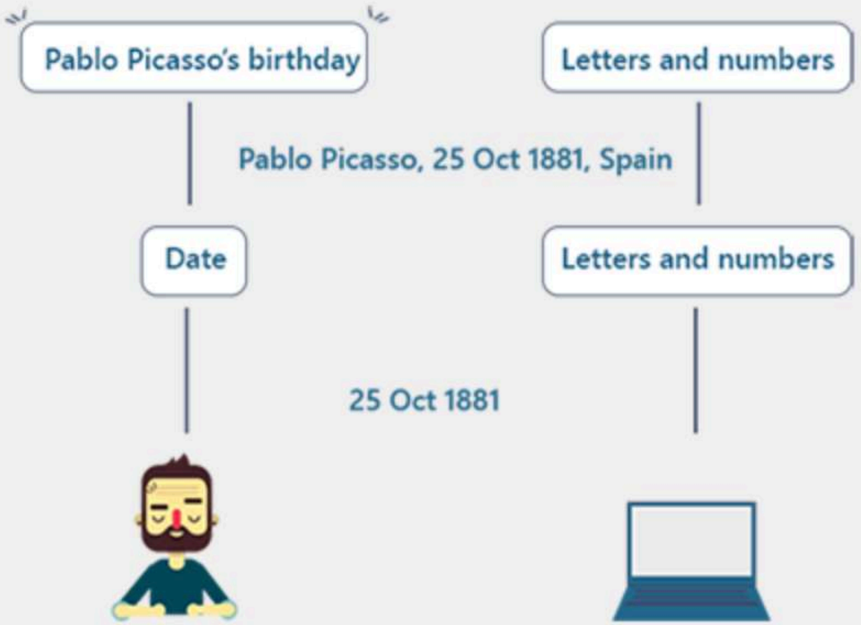


Liu et al., SIGMOD 2020, GIANT: Scalable Creation of a Web-scale Ontology

Microsoft Concept Graph

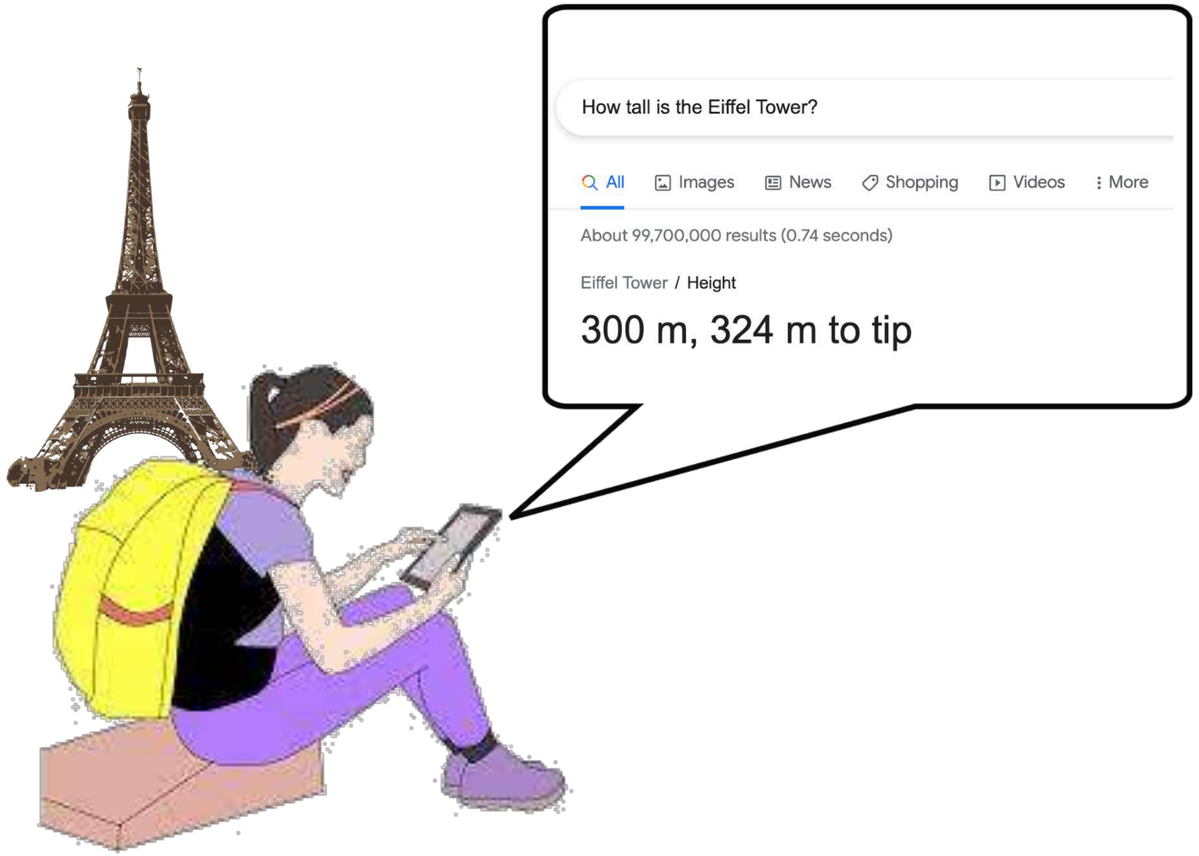
Microsoft Concept Graph

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.

Recap: suppose you are asking…



How does Google know?

Question: "How tall is the Eiffel Tower?"

Google Knowledge Graph

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

Recap: suppose you are shopping…

Roll over image to zoom in

Brand	Cetaphil
Ingredients	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.
Scent	Fragrance free
Additional Item Information	Non-Comedogenic, Fragrance-free, Natural
Skin Type	Sensitive

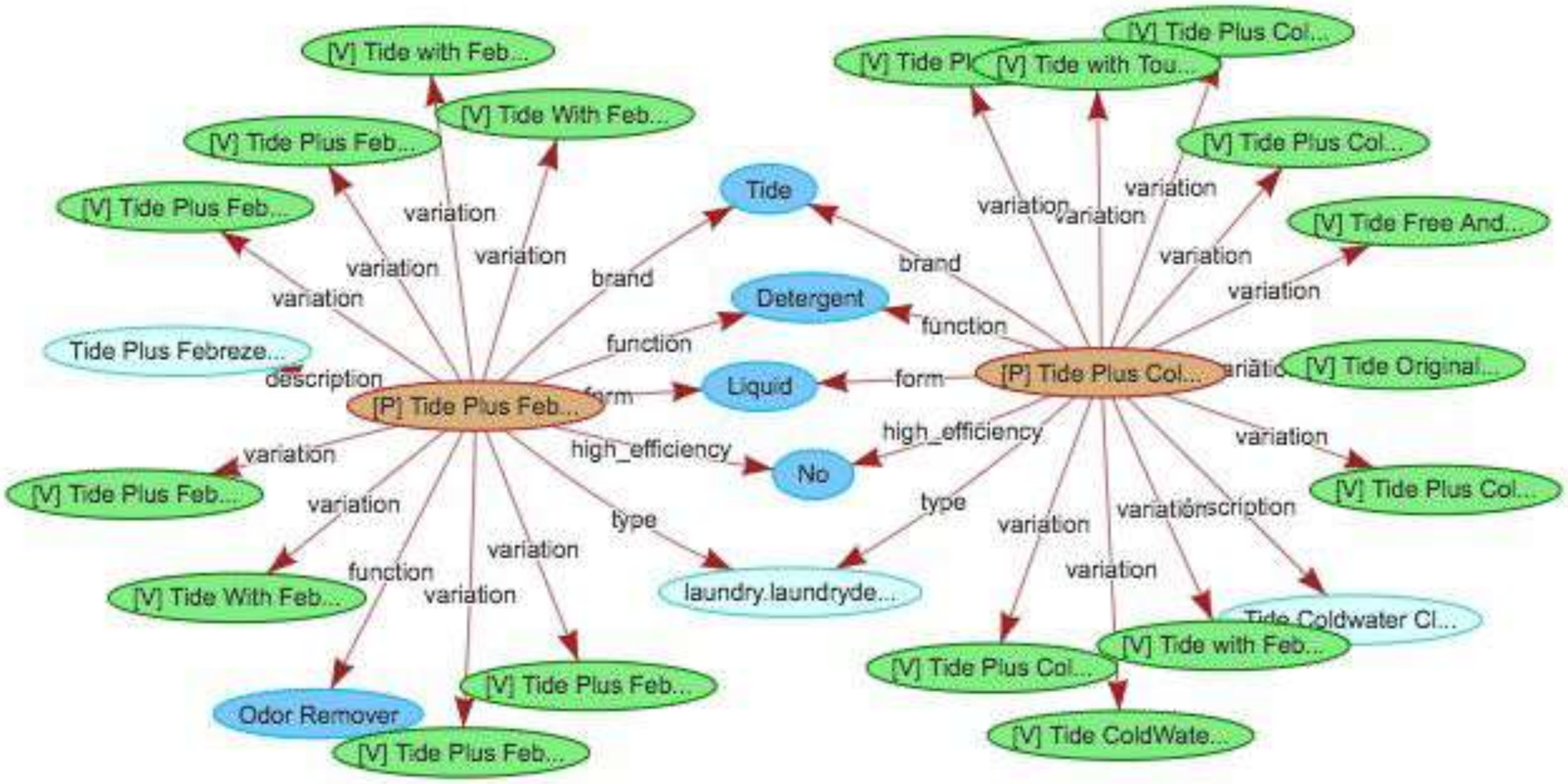
About this item

- 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?

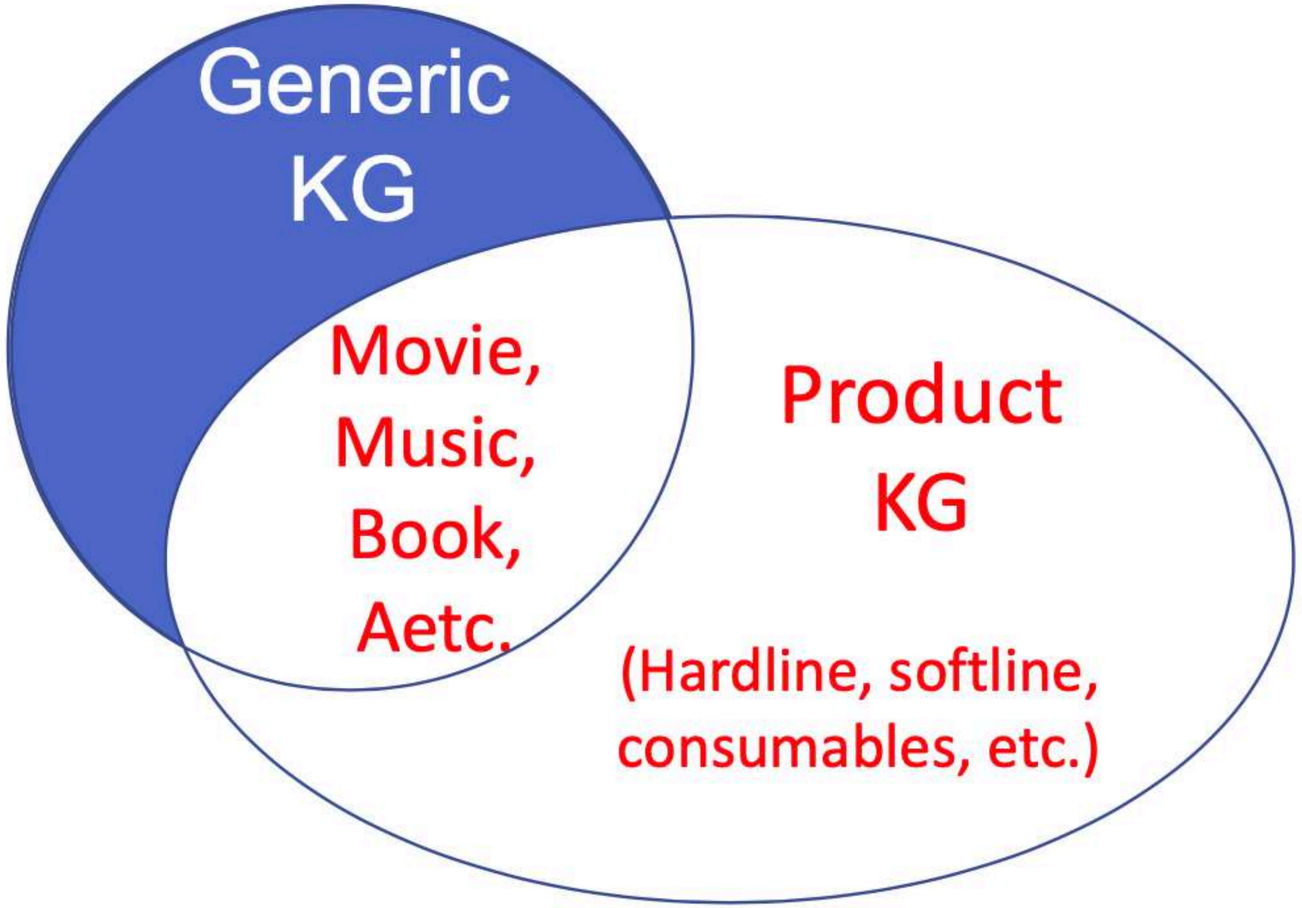
Amazon Product Knowledge Graph

Product Graph Example for 2 Products



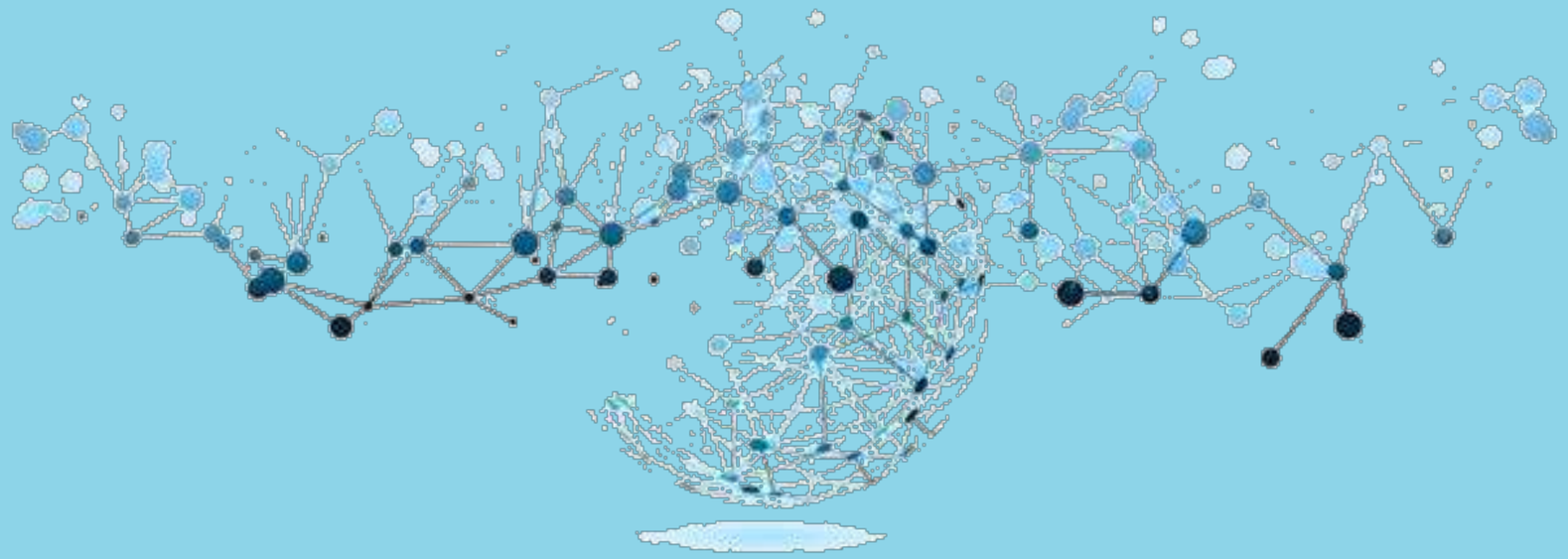
https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/

Product Graph vs. Knowledge Graph



https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/

Semantic Web Knowledge Graph



W3C Semantic Web

- 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**.

44 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 Scheme 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.**

KG Standards

W3C RDF — the **data** part

W3C OWL — the **schema** part

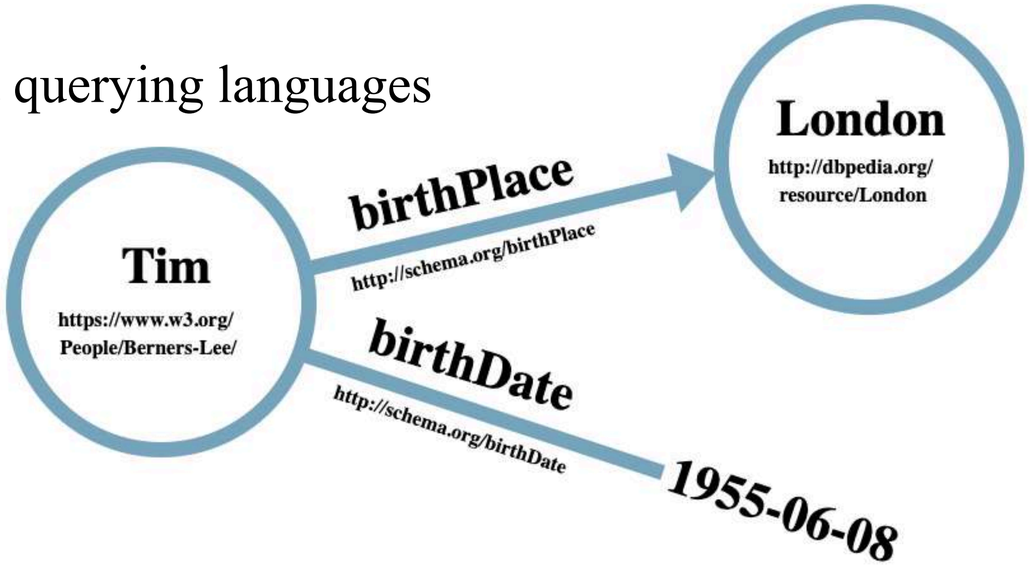
W3C SPARQL — the **query** part

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/

W3C

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.

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!)

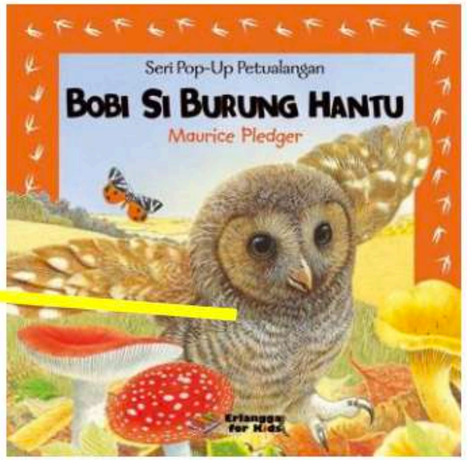
- Short for: Web Ontology Language. OWL is a language to create ontologies

- 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!



51 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

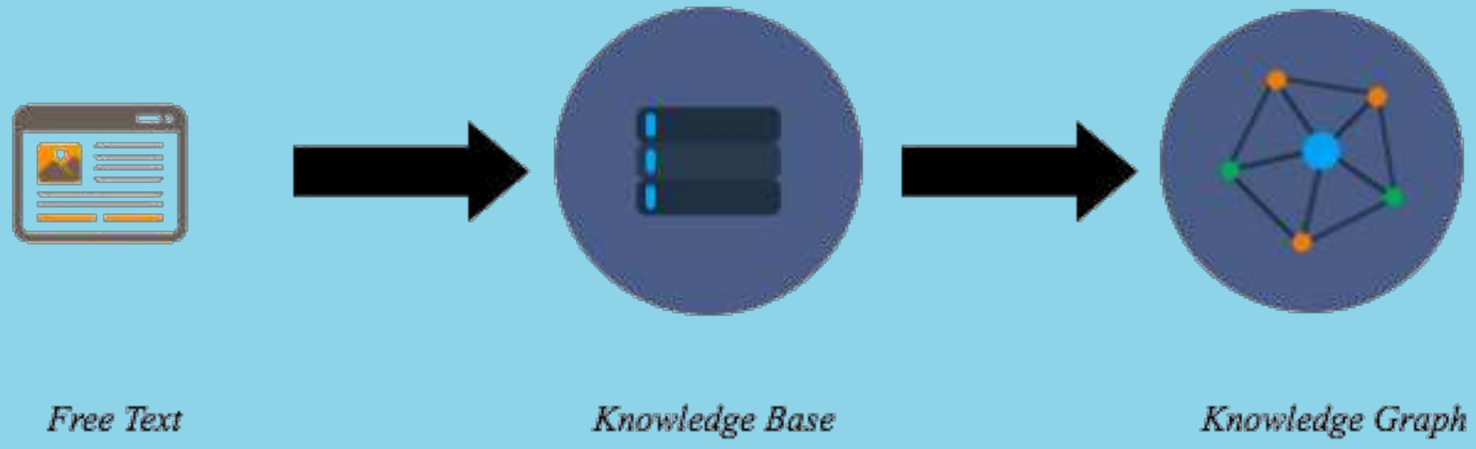
Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data. SPARQL is the query language for the Semantic Web.

Inference

Near the top of the Semantic Web stack one

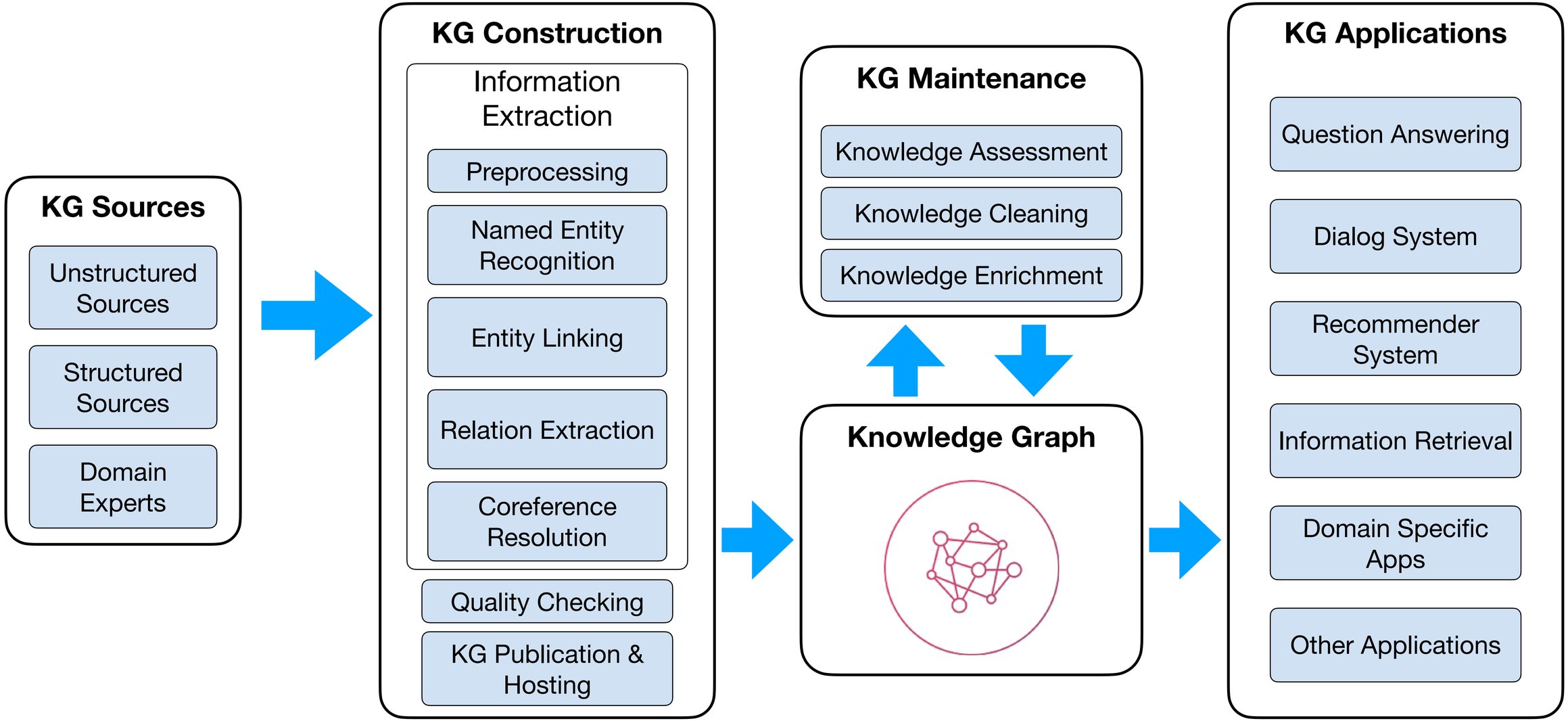
Vertical Applications

W3C is working with different industries —

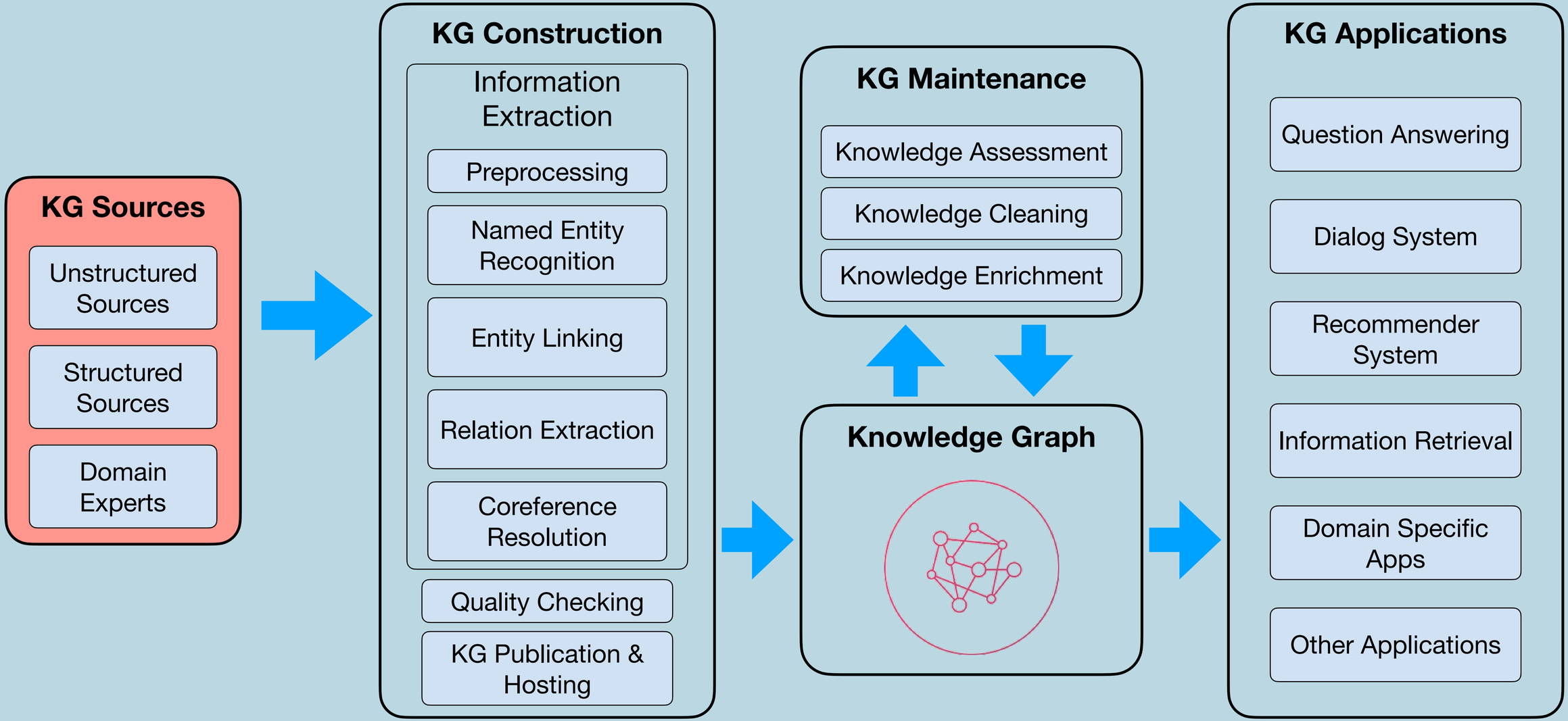


KG Construction

KG Workflow: How to build a KG from Scratch



KG Workflow: How to build a KG from Scratch



KG Construction

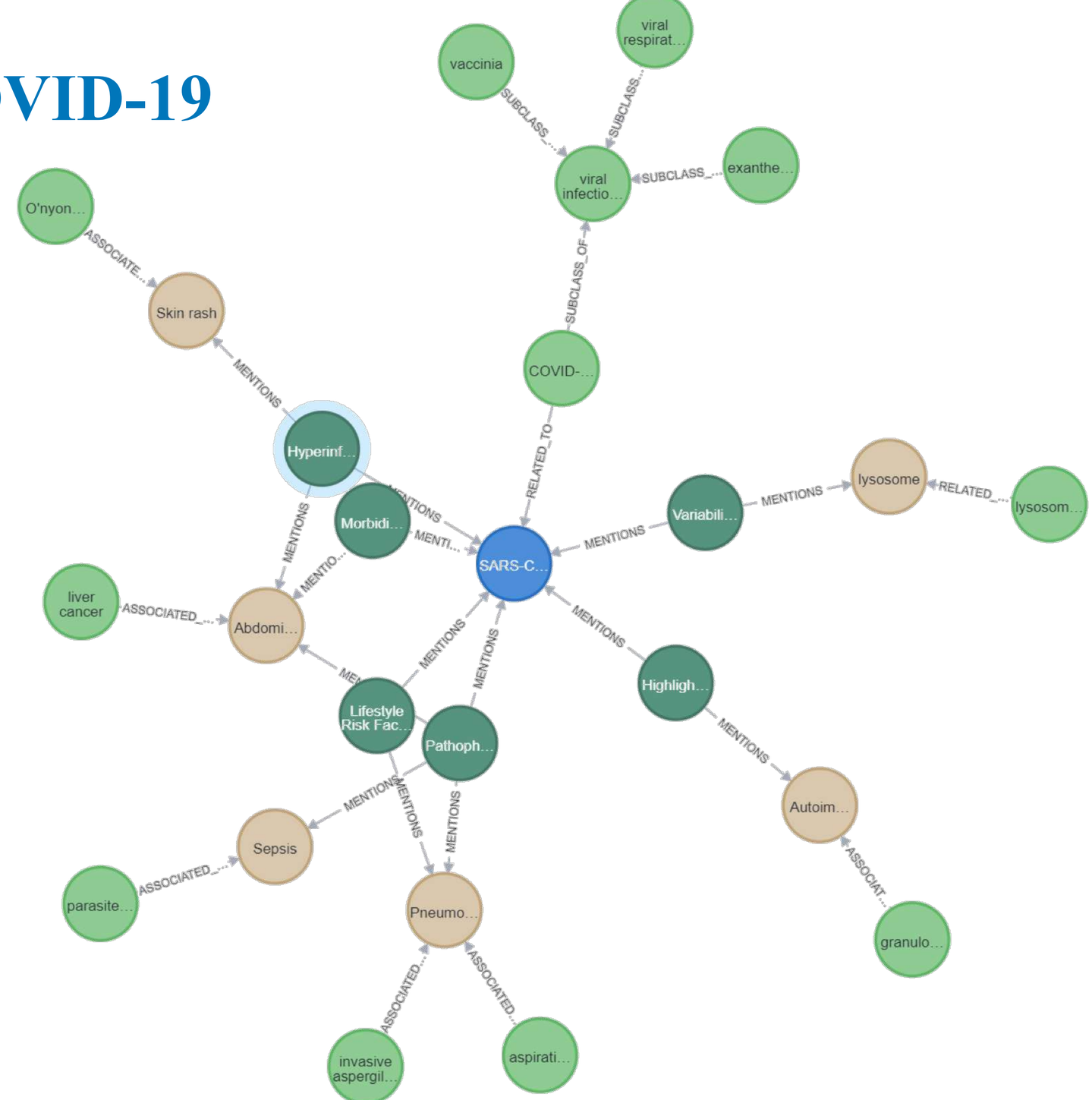
- We know that a knowledge graph is when an **ontology** is **applied to** a set of individual **data**.

All of which are based on RDF triples.

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

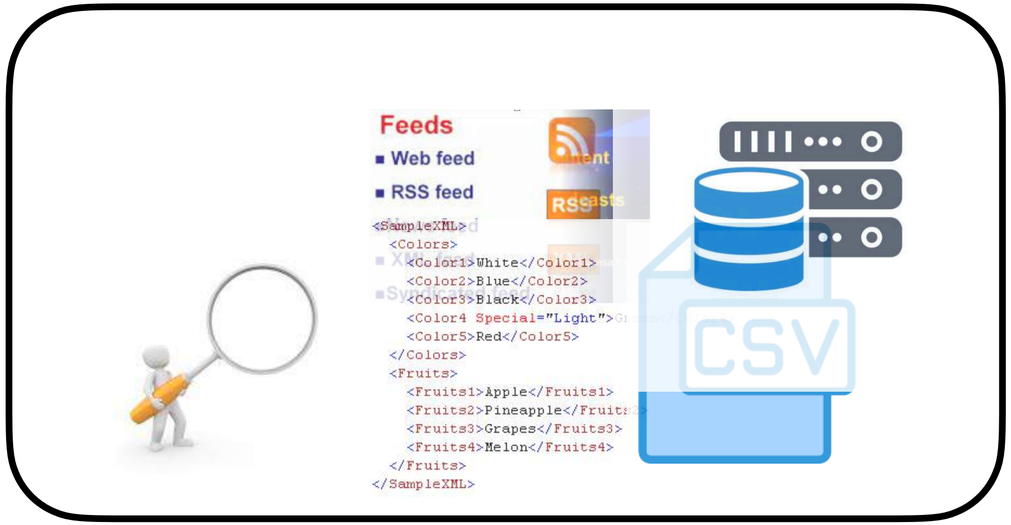
Example: KG-COVID-19

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

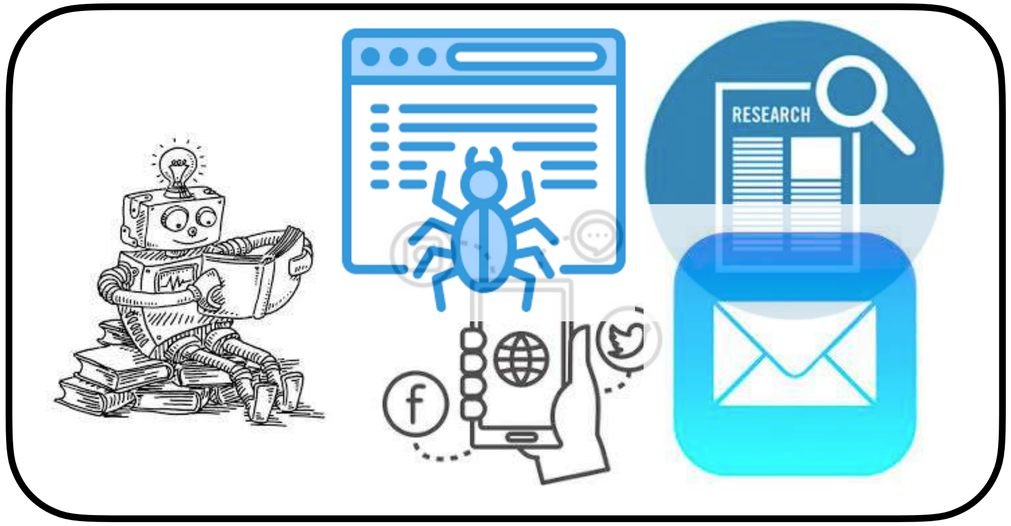


Knowledge Sources

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



- Unstructured Sources
  - Web Craws
  - Scientific Articles
  - Social Media
  - Emails
  - …



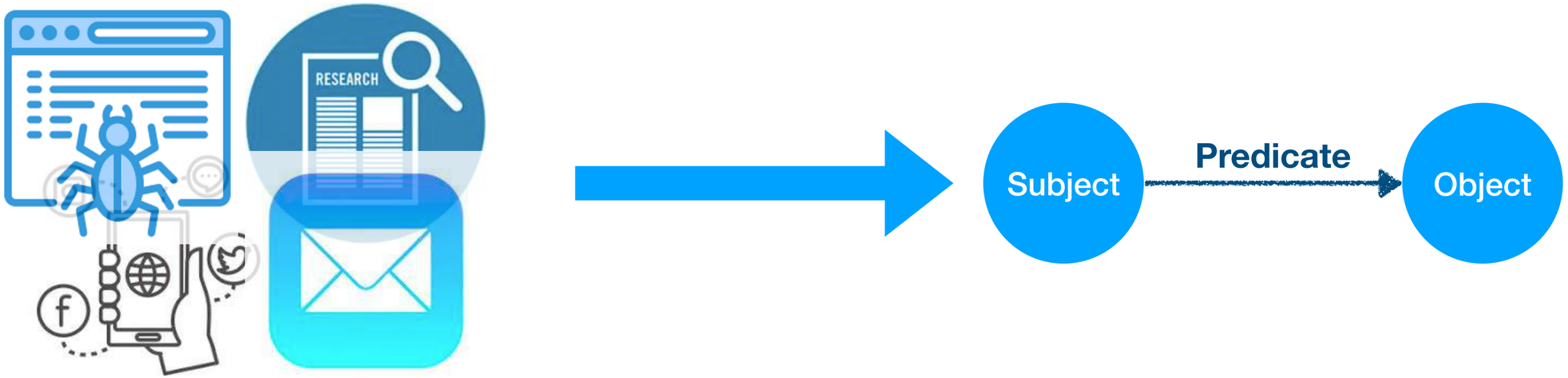
Structured Knowledge Sources to KGs

- 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.

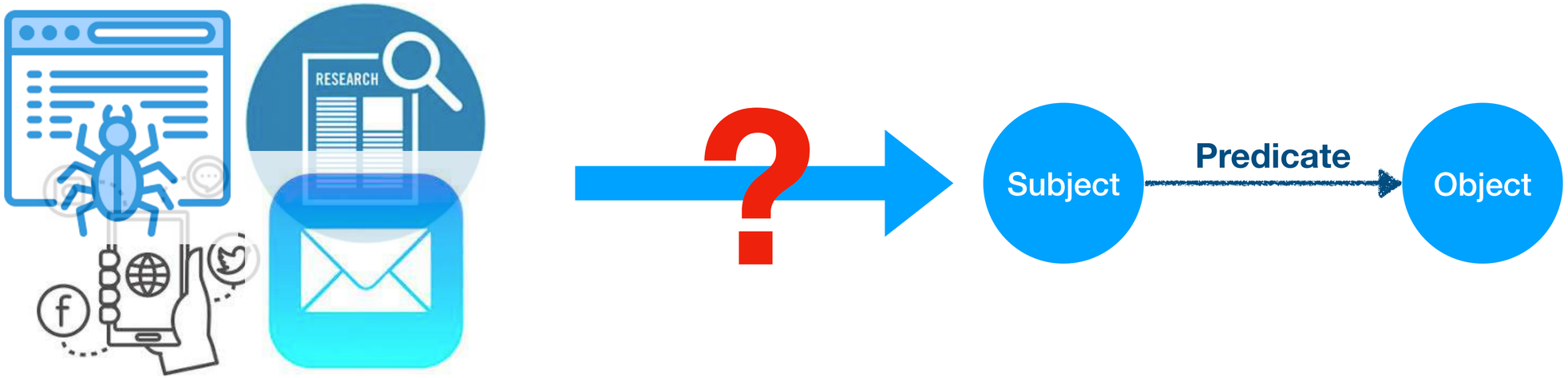
Unstructured Data to Structured Knowledge

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



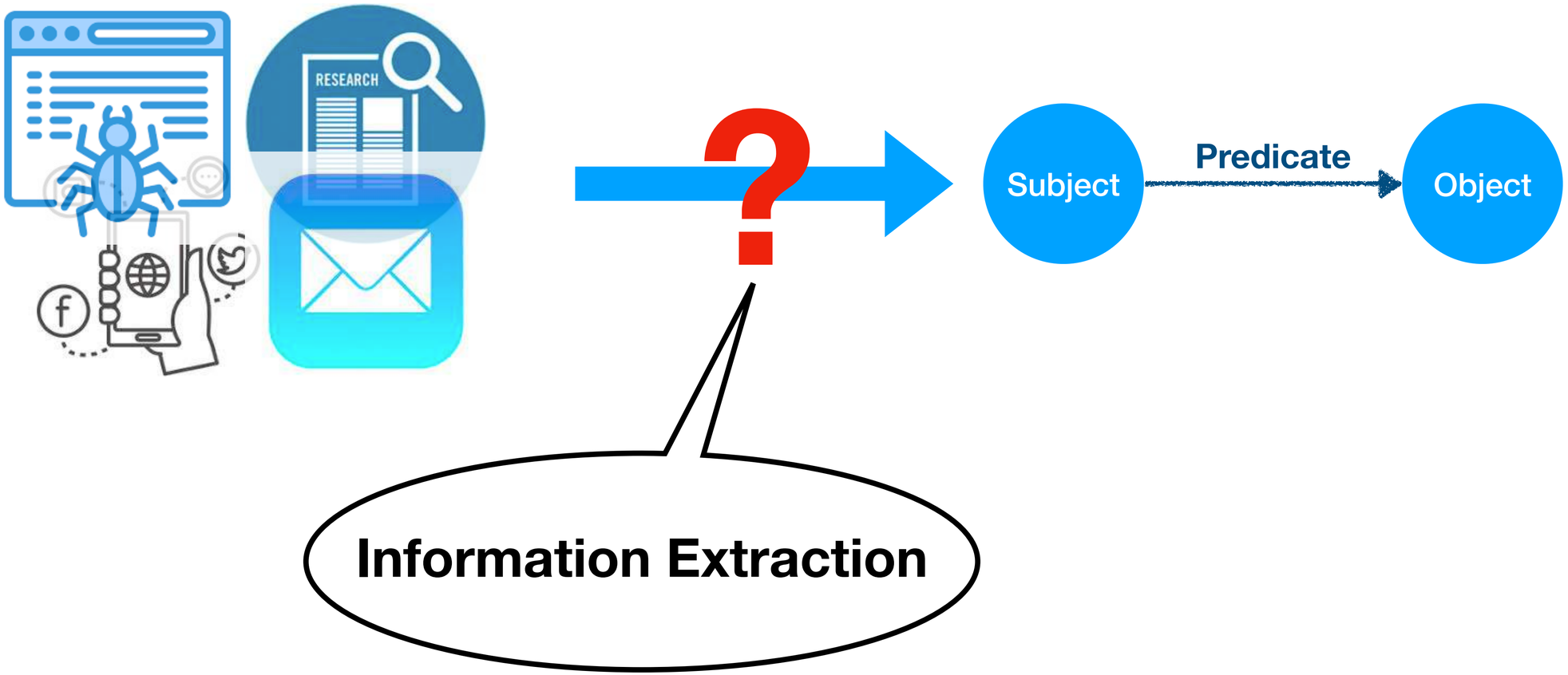
Unstructured Data to Structured Knowledge

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

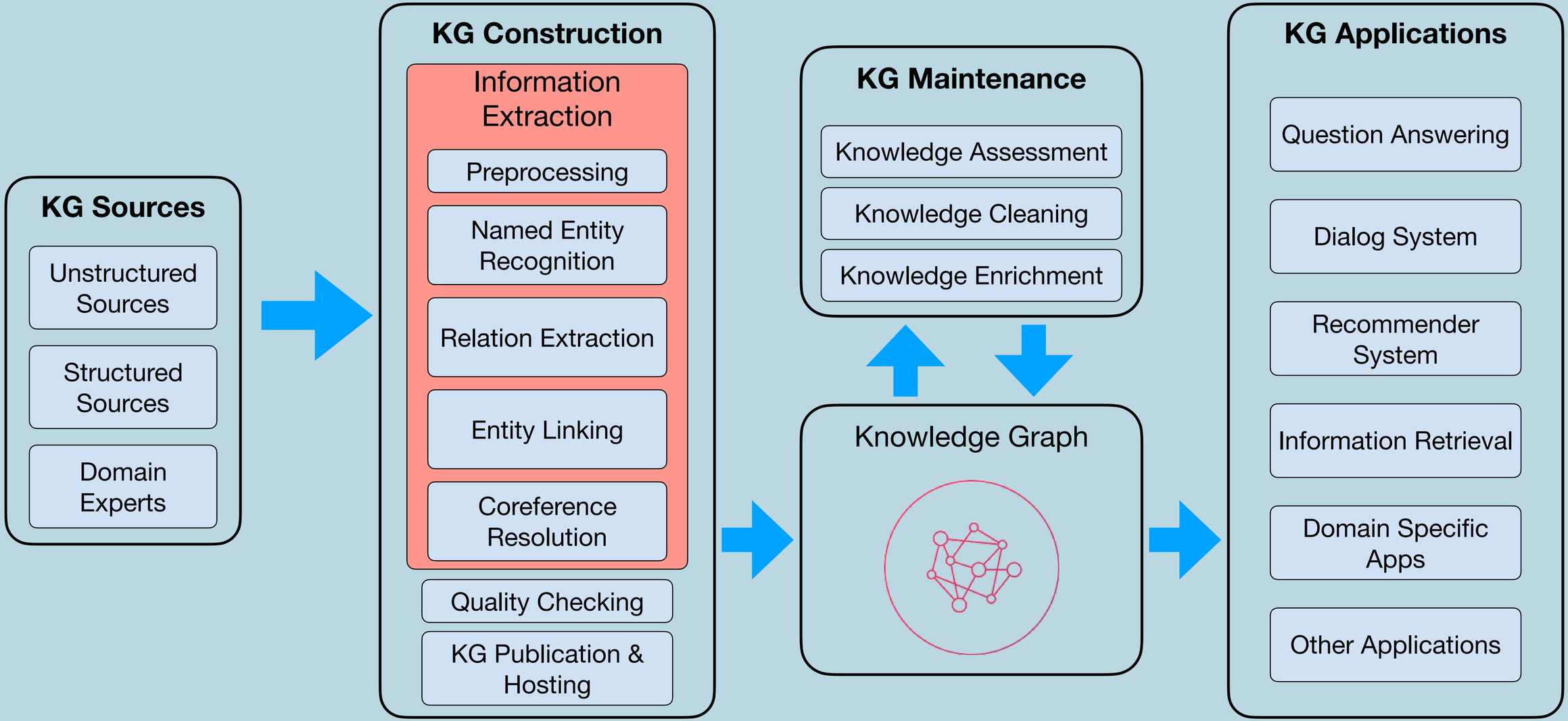


Unstructured Data to Structured Knowledge

- How to excavate structured knowledge from unstructured data automatically?

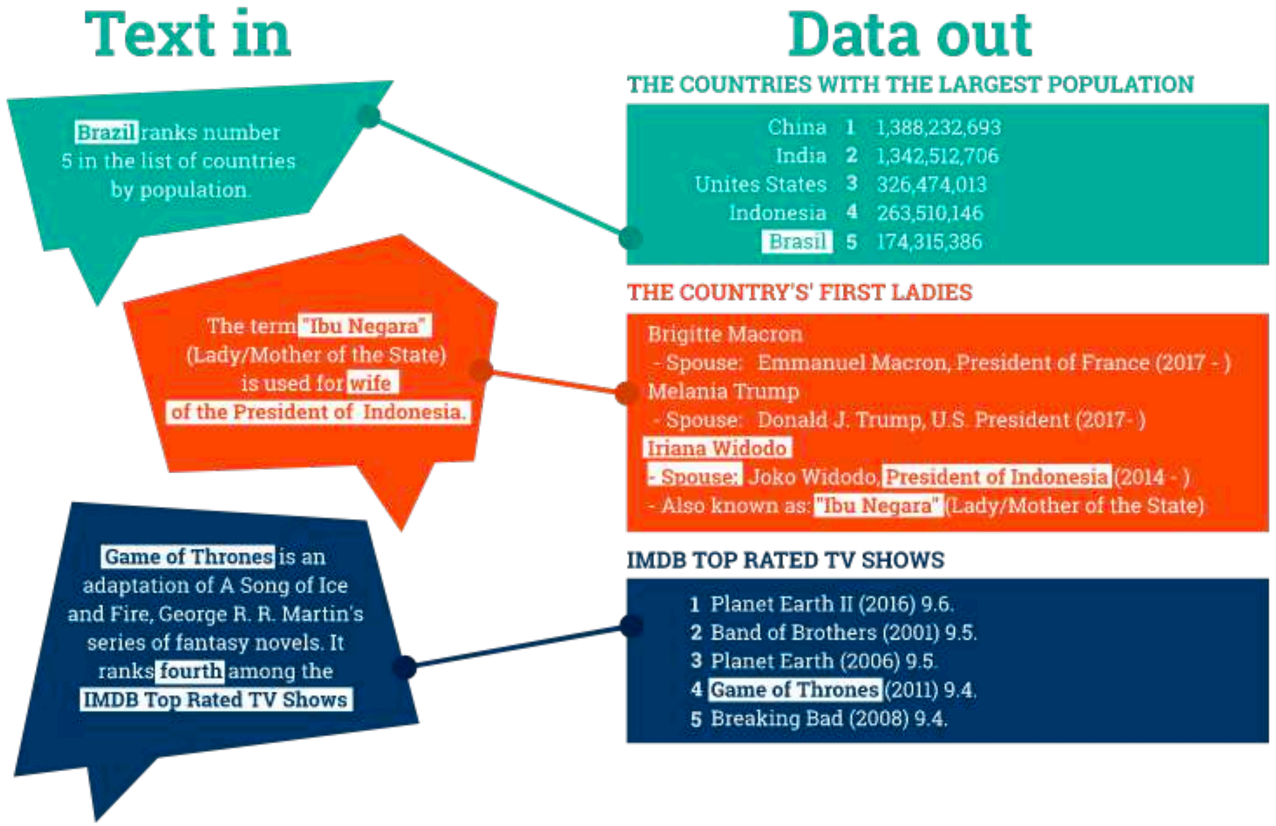


Information Extraction

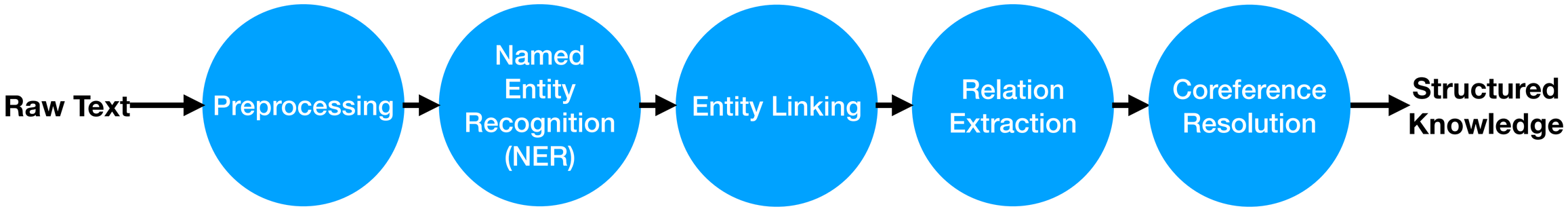


What is information extraction?

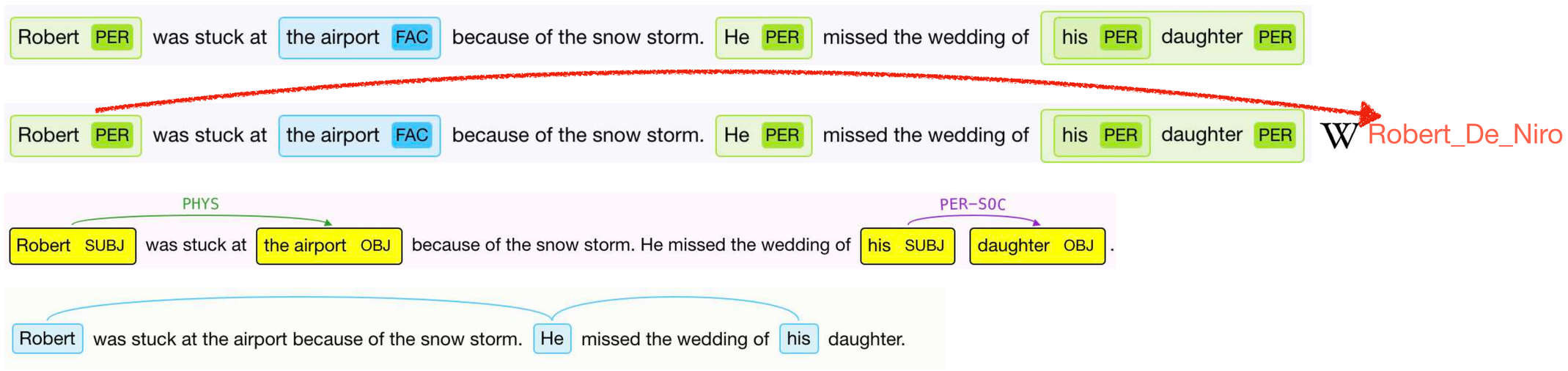
- **Information extraction (IE)** is the task of **automatically extracting structured information** from unstructured and/or semi-structured machine-readable documents and other electronically represented sources. (From Wikipedia)



How does IE work?



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

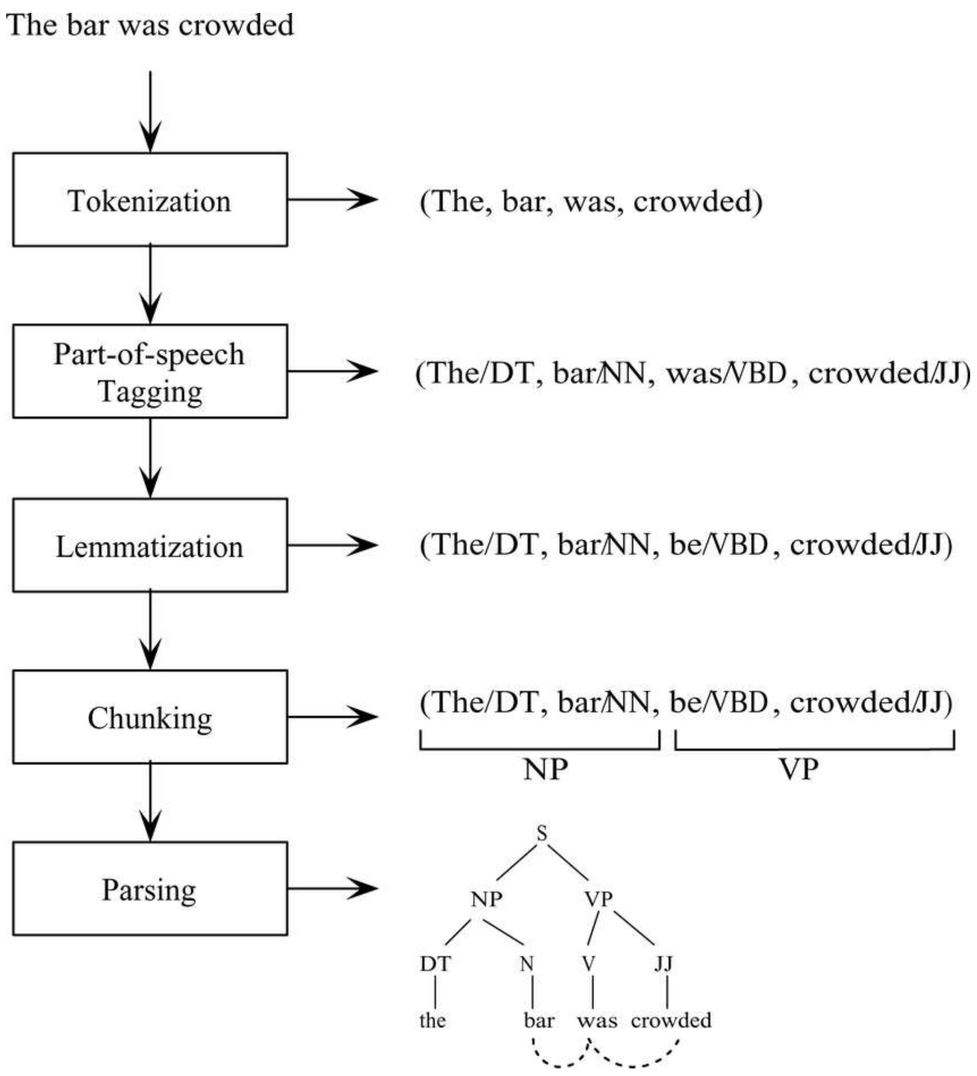


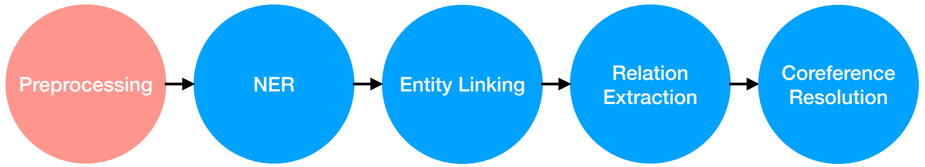
Structured Knowledge (**Robert**, **PHYS, the airport**), (**Robert**, **PER-SOC, daughter**)

https://huggingface.co/hmtl/#emd

Preprocessing

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

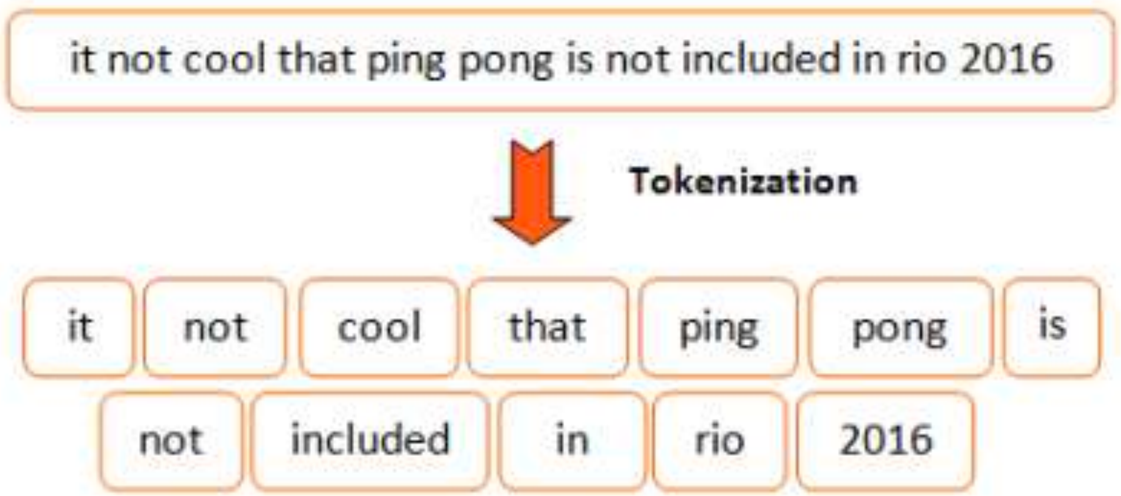




Tokenization

- 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
  - ...



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.

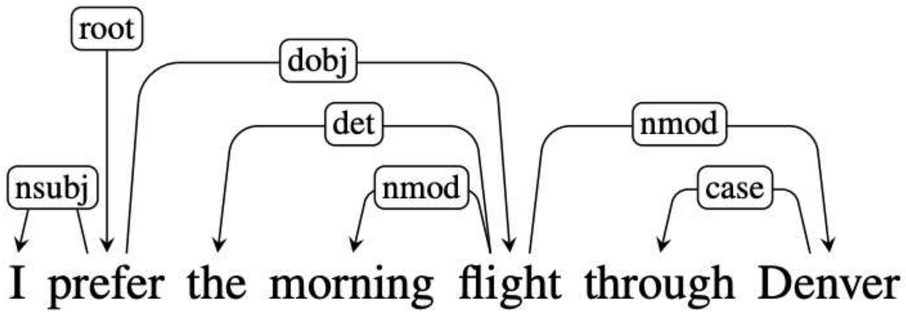
- **Tools:**
  - spaCy Tagger
  - NLTK pos_tag
  - …

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

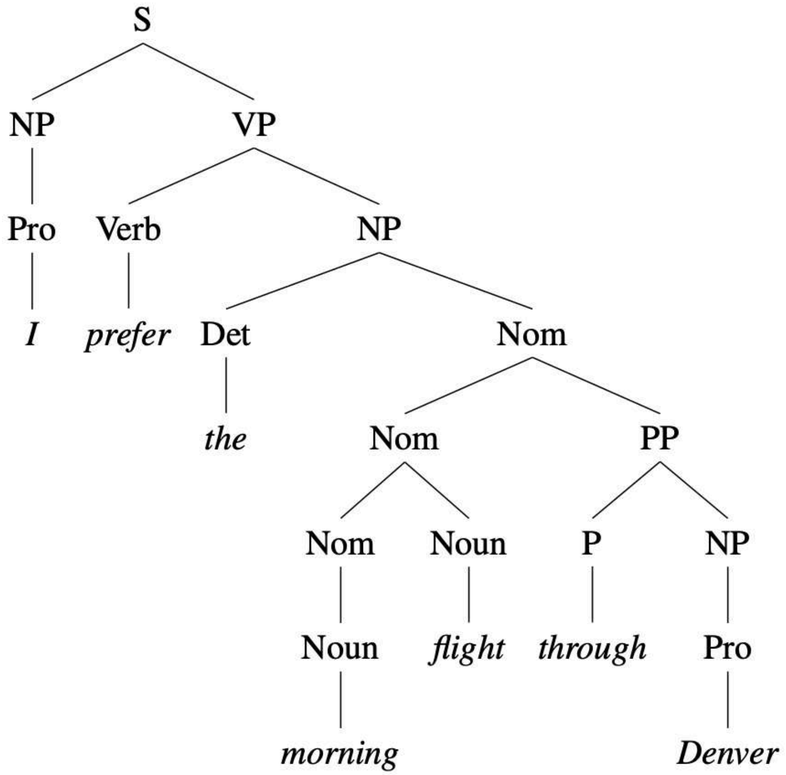
Dependency Parsing

- 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

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

NLTK

Search

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

Documentation

Natural Language Toolkit

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.

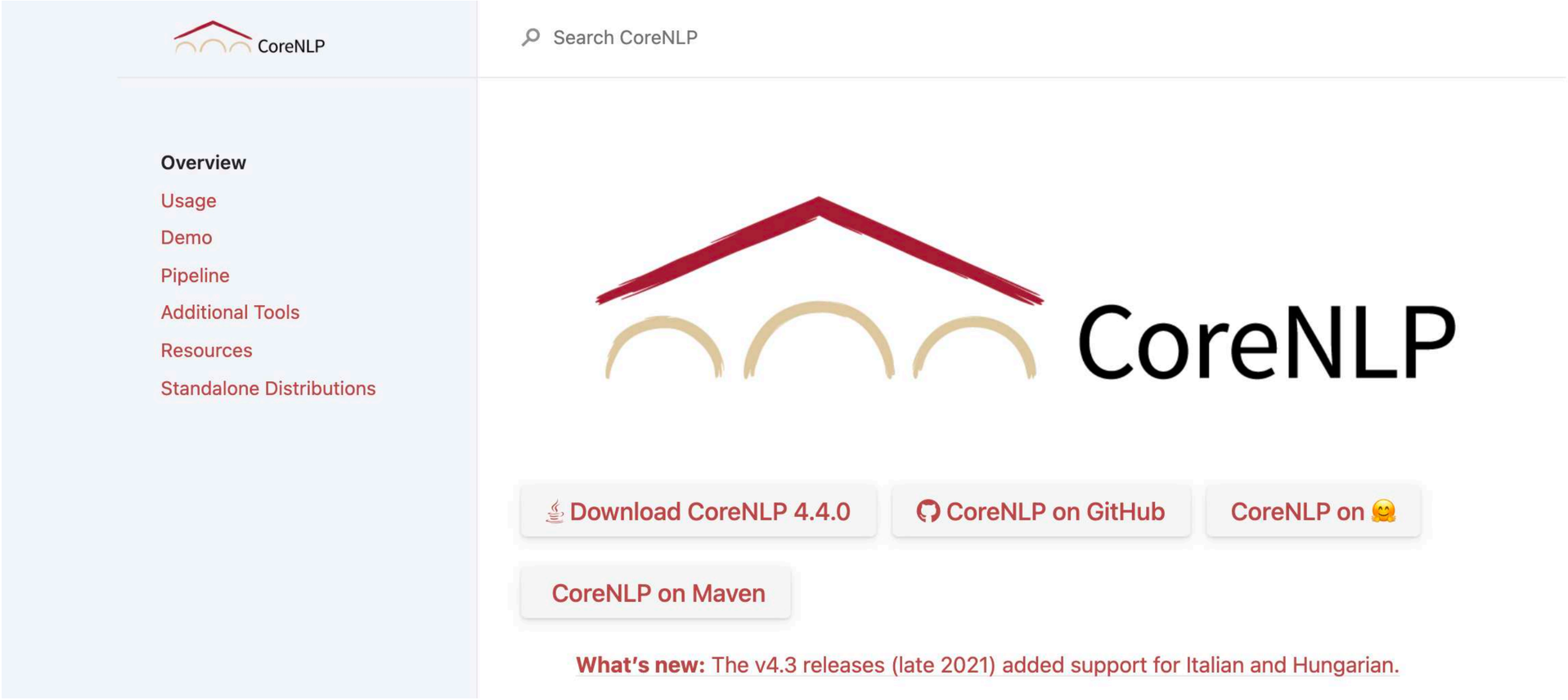
NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

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 been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at **https://www.nltk.org/book_1ed**.)

Some simple things you can do with NLTK

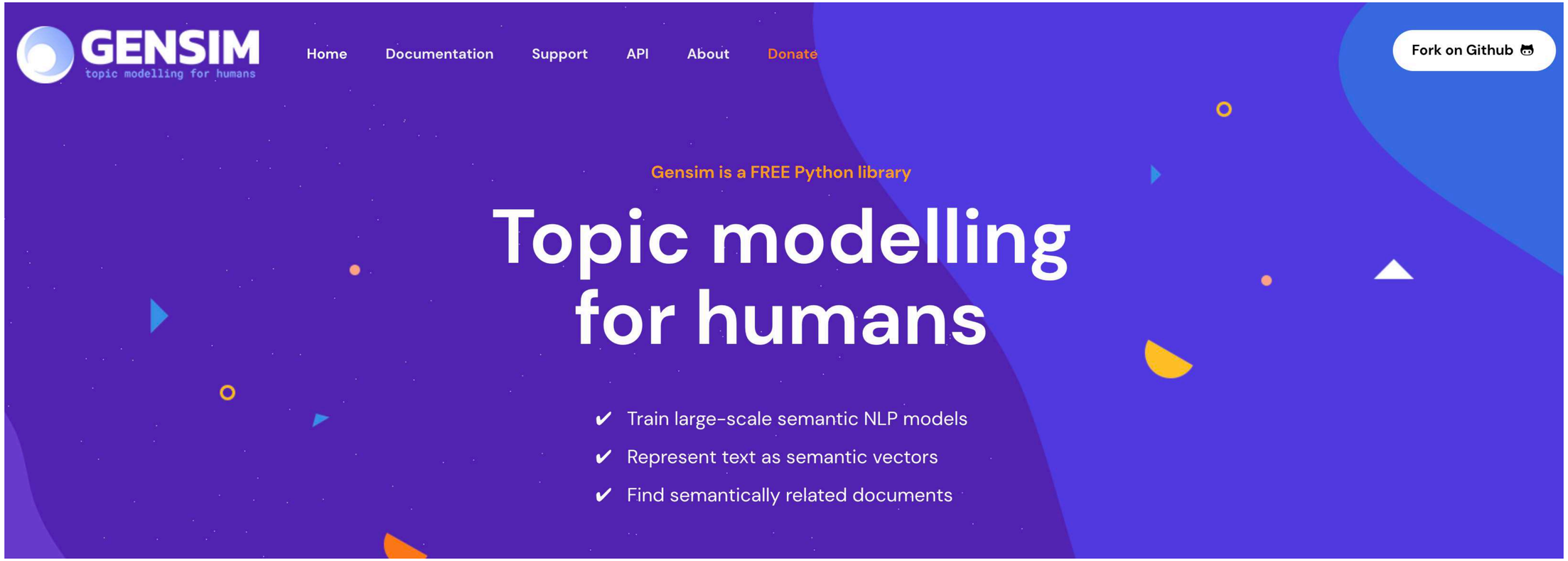
70 Stanford CoreNLP

- **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/)



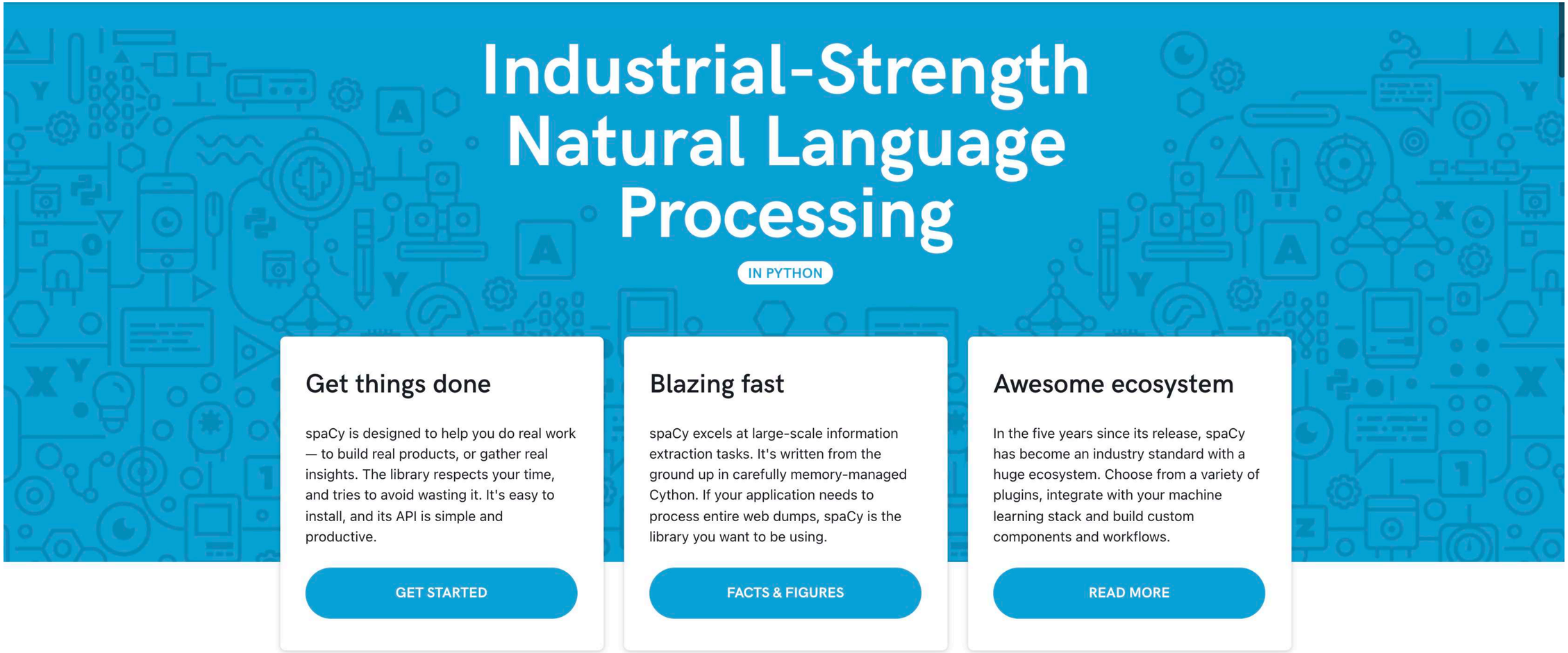
GENSIM

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



spaCy

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



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.

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.

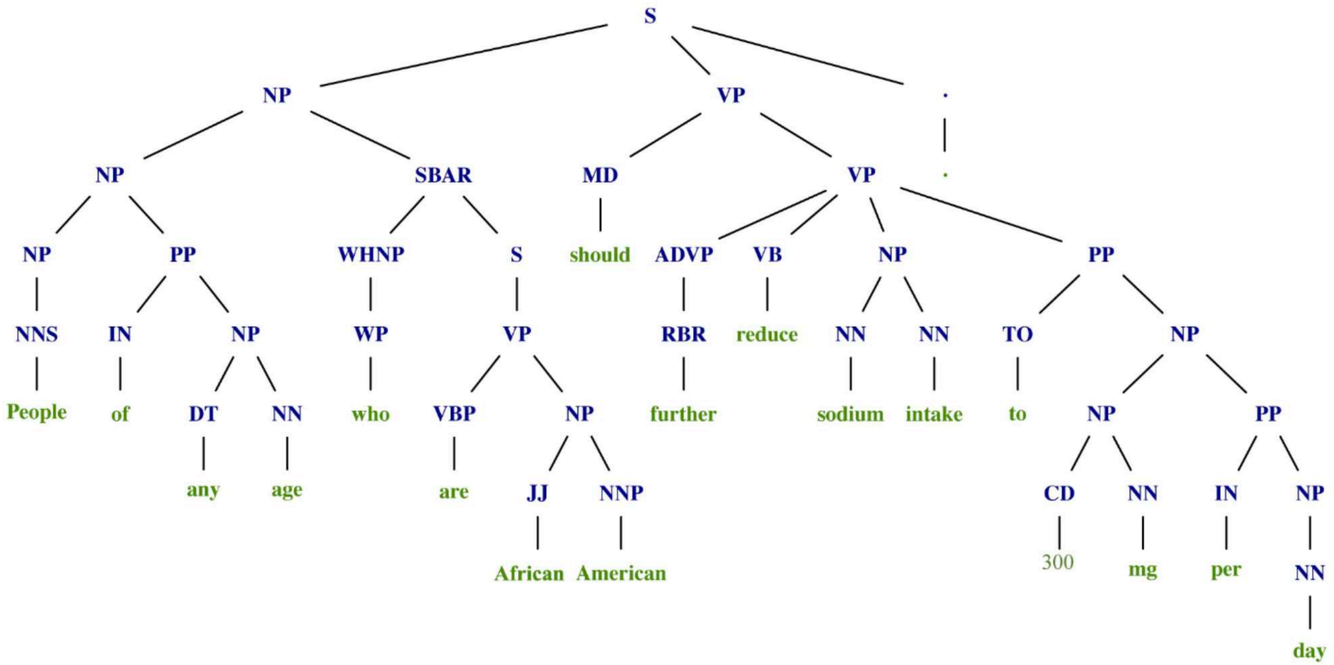
Named Entity Recognition (NER)

- **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, …



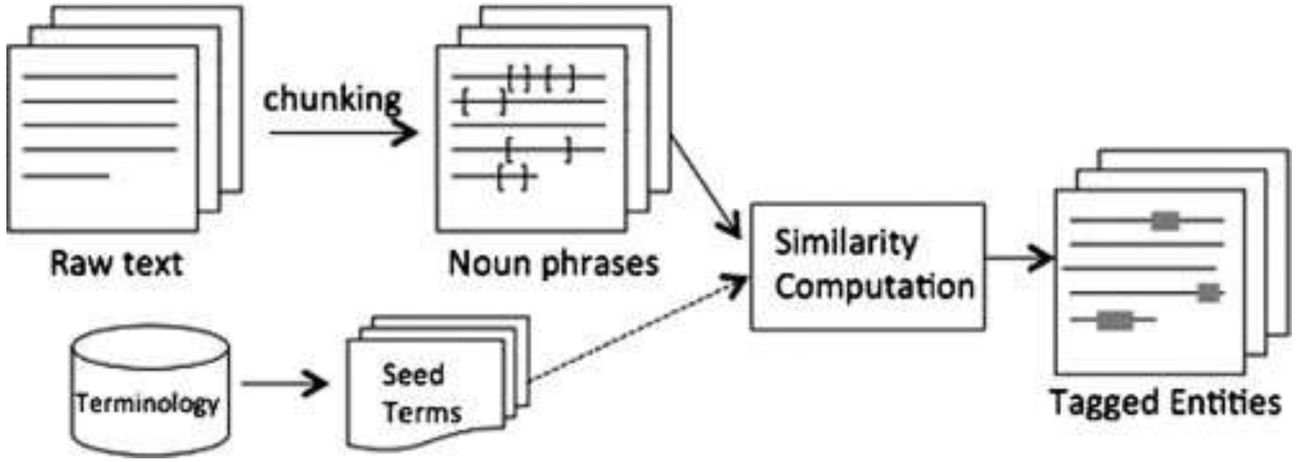
Named Entity Recognition (NER)

- **Unsupervised Learning Approaches:**

  - **Pros**:
    Low required resources.

  - **Cons**:
    Low performance.

  - **Unsupervised NER systems**:
    KNOWITALL, Collins et al., Nadeau et al., Zhang and Elhadad, …



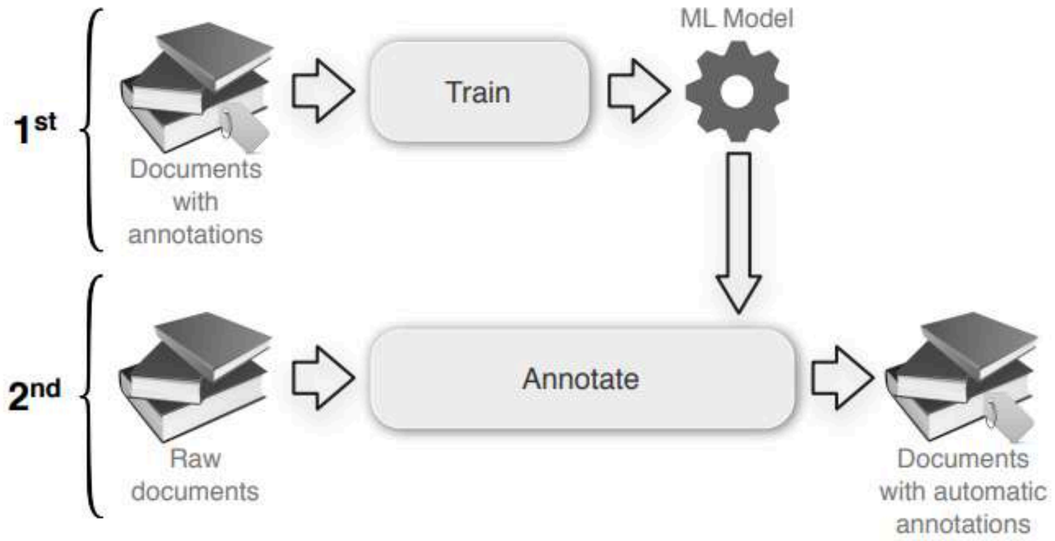
Named Entity Recognition (NER)

- **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,
    …



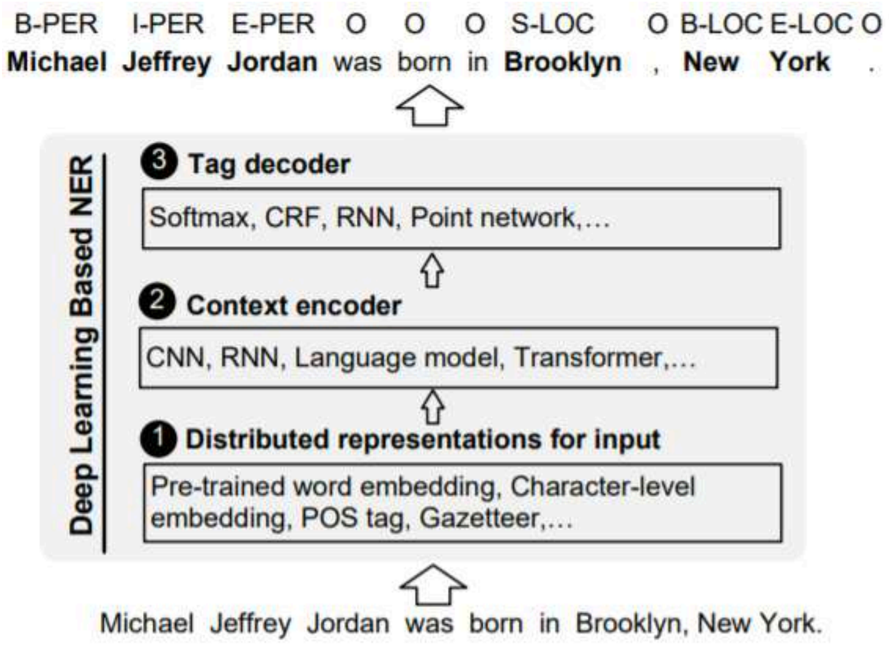
Named Entity Recognition (NER)

◉ 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, …



NER Datasets

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	6/10	http://noisy-text.github.io
BBN	2005	Wall Street Journal	64	https://catalog.ldc.upenn.edu/LDC2005T33
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	112	https://github.com/xiaoling/figer
HYENA	2012	Wikipedia	505	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
Gillick	2016	Magazine, news, web, etc.	89	https://arxiv.org/e-print/1412.1820v2
FG-NER	2018	Various	200	https://fgner.alt.ai/
NNE	2019	Newswire	114	https://github.com/nickyringland/nested_named_entities
GENIA	2004	Biology and clinical text	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	1	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

https://arxiv.org/pdf/1812.09449.pdf

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%; **OntoNotes5.0: 92.07%**
[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: 93.5%**
[136]	-	GloVe	Plooled contextual embeddings	RNN	CRF	CoNLL03: 93.47%

https://arxiv.org/pdf/1812.09449.pdf

Off-the-shelf NER tools

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

More in this survey: https://arxiv.org/pdf/1812.09449.pdf

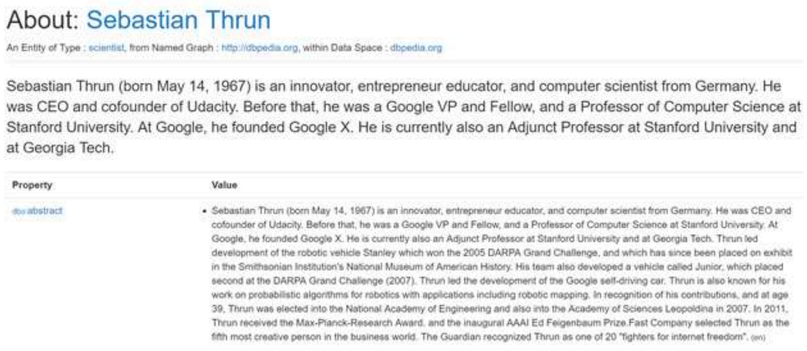
Entity Linking (EL)

- **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.

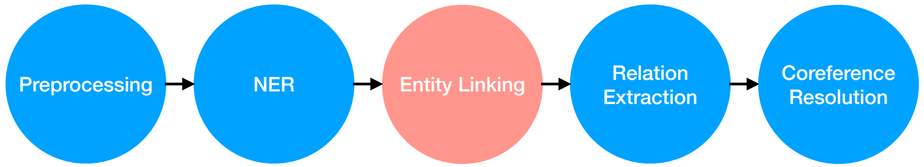

About: Sebastian Thrun

An Entity of Type : scientist, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

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
dbo:abstract	• 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)

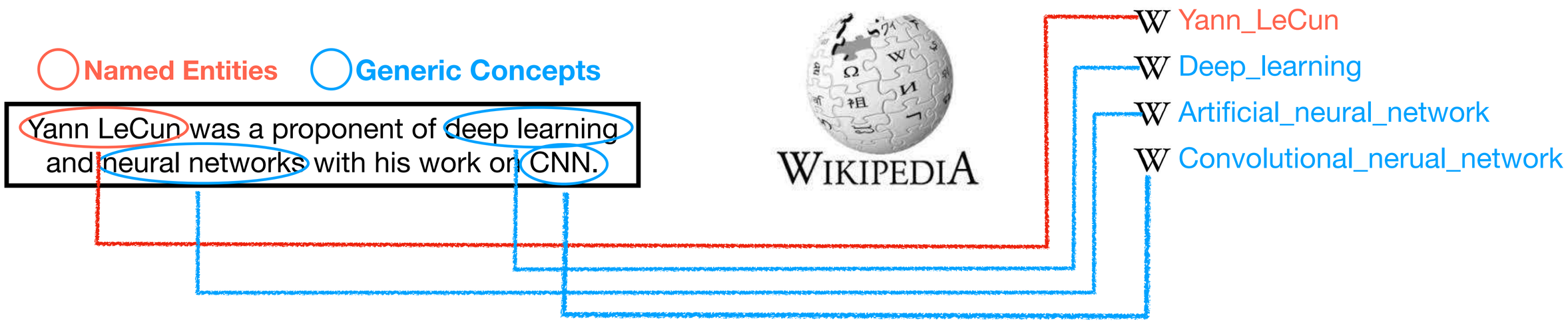
http://dbpedia.org/page/Sebastian_Thrun



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 adapts to the application

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 fundam

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

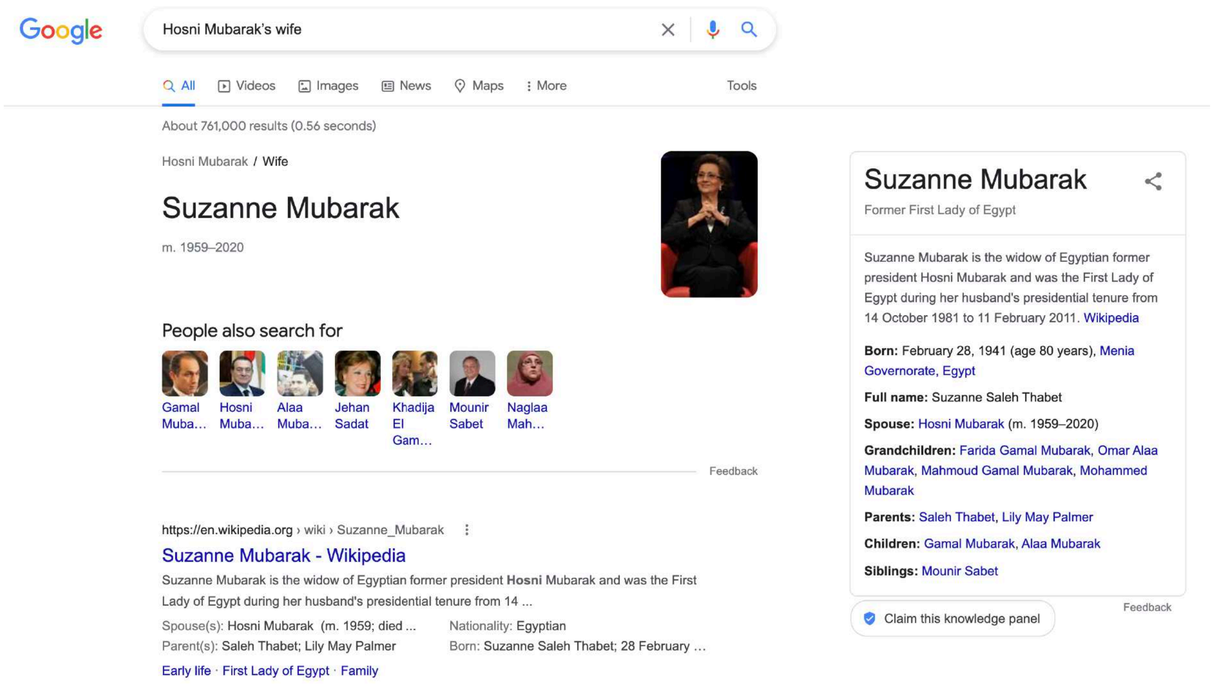
Richard Blu ... ate
From Wikipedia, the free ...

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.

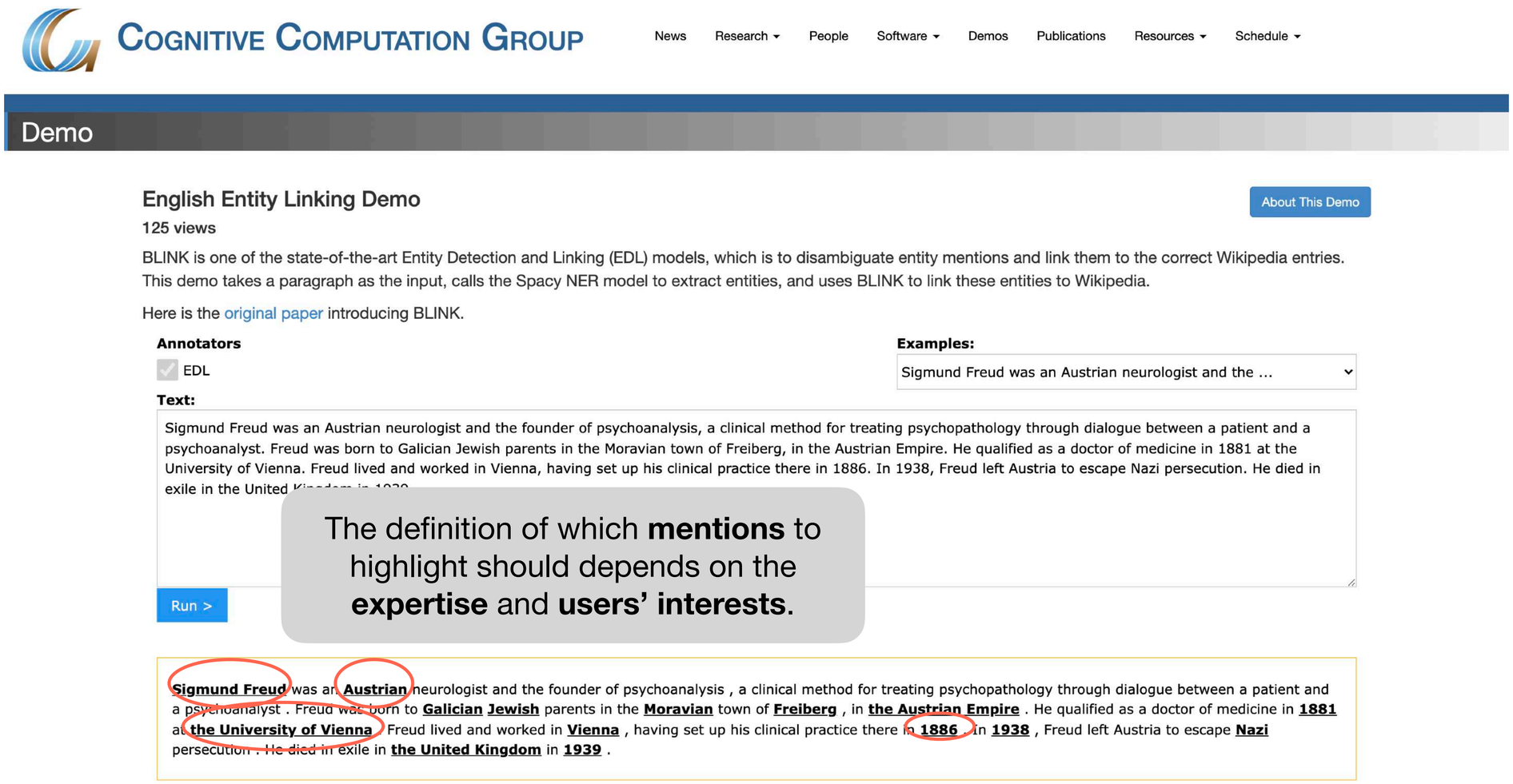
Chris Dodd
From Wikipedia, the free encyclopedia

The New York Times
From Wikipedia, the free encyclopedia

Connecticut
From Wikipedia, the free encyclopedia



English Entity Linking Demo

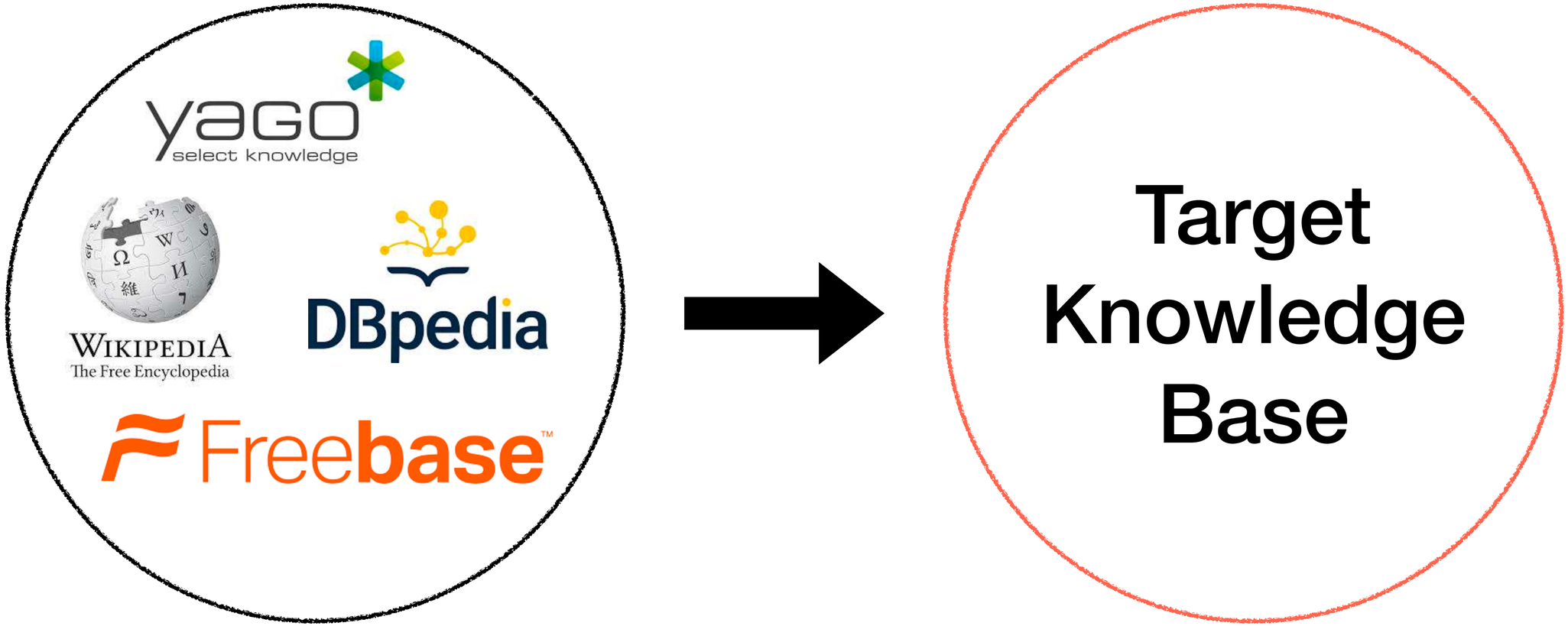


The definition of which **mentions** to highlight should depends on the **expertise** and **users' interests**.

https://cogcomp.seas.upenn.edu/page/demo_view/EDLEnglish

Entity Linking: Define the target KB

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

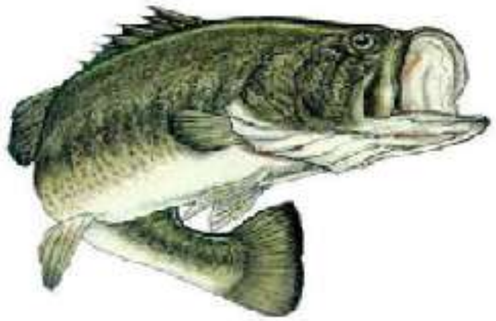


Entity Linking: Find the Unique ID

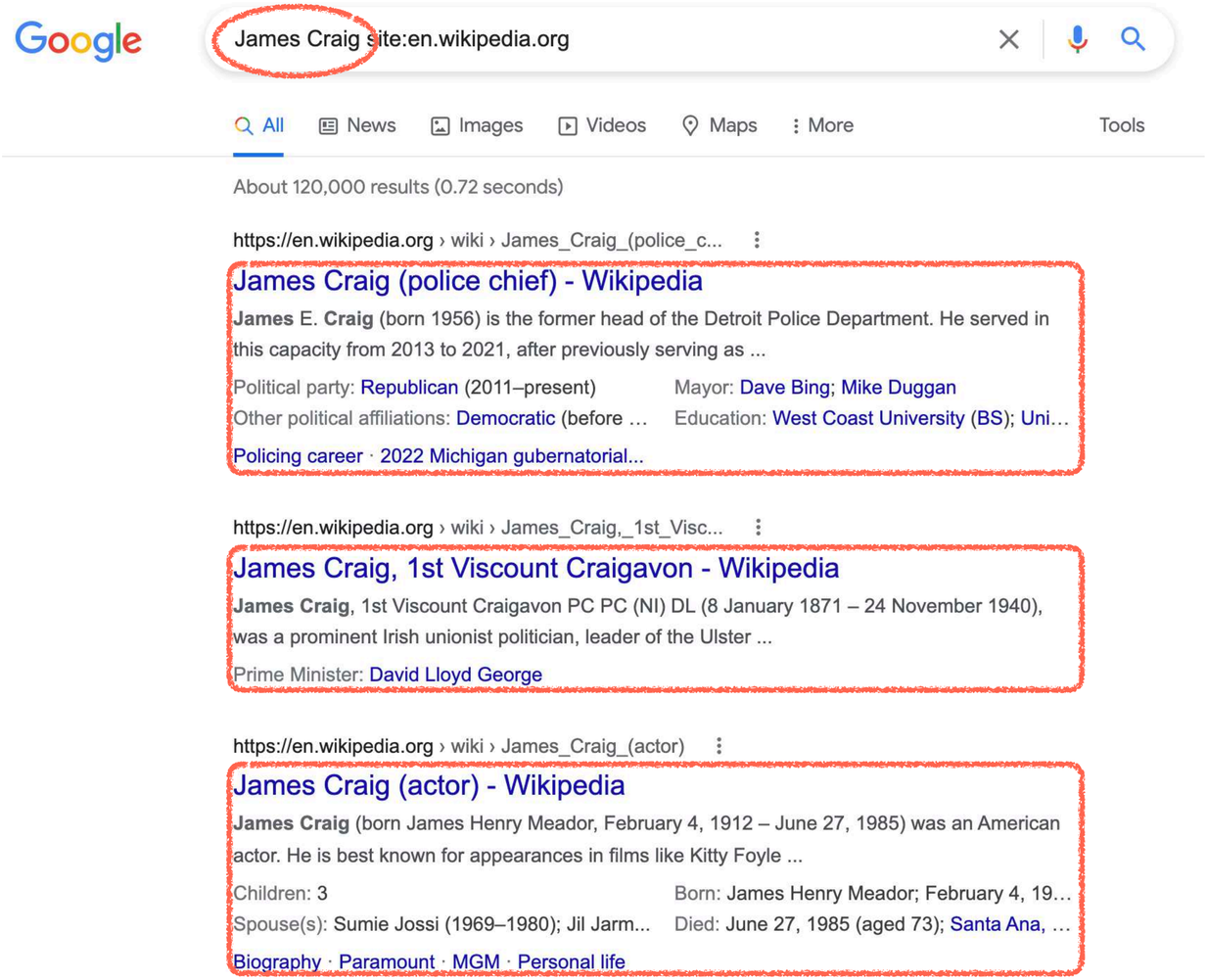
Find the unique identity in the target KB (Disambiguation)

Which one to link?

This is a bass



This is also a bass



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**?

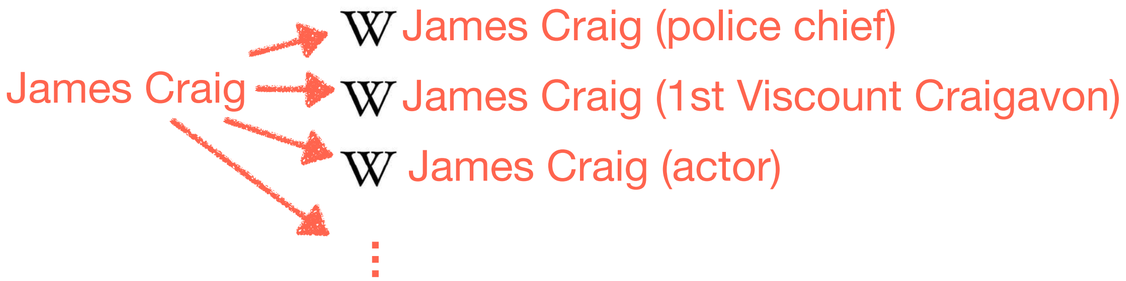
Marc-Alexandre Côté is an author of **Textworld**.

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

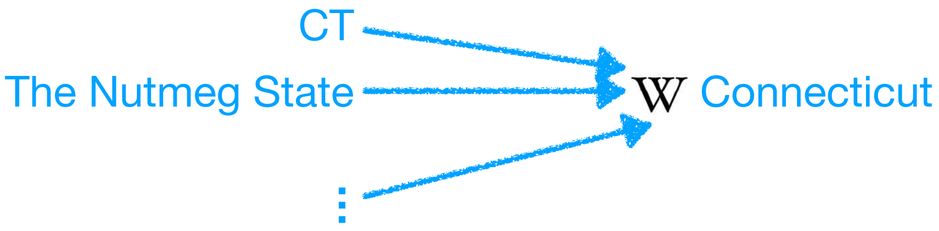
Entity Linking (EL)

- **What are the main issues with EL?**

- **Ambiguity**



- **Variability**



- **Concepts outside of KB (NIL)**

Marc-Alexandre Côté is an author of Textworld.

- **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]

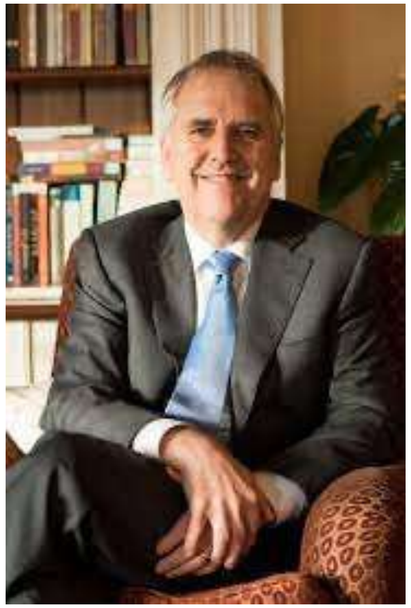
Sports [edit]

- 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

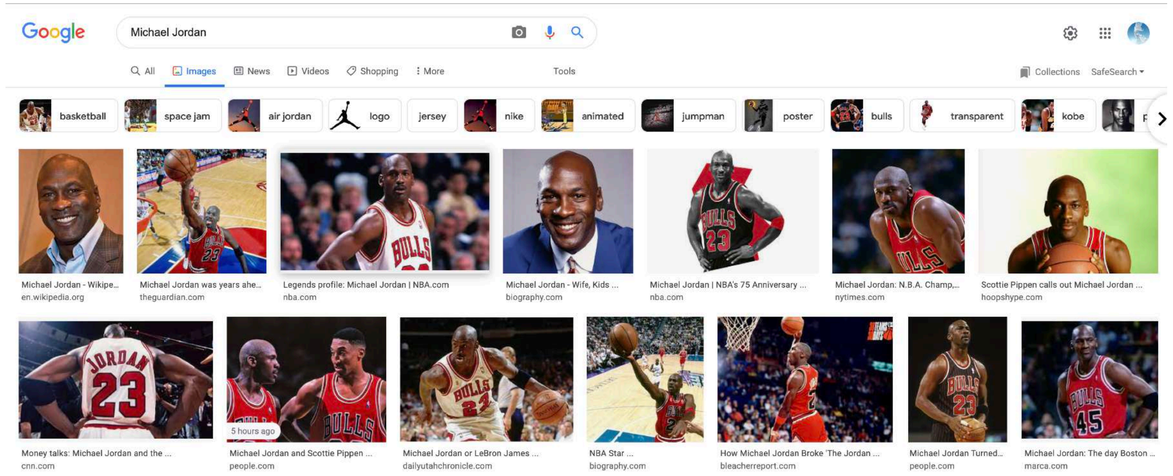
Other people [edit]

- 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

General Architecture

Mention Identification

- Each n-gram
- Surface form based filtering
- Classification and statistics based filtering

Candidate Generation

- Surface form matching
- Expansion using aliases
- Prior probability computation

Candidate Ranking

- Local Inference
- Global Inference

NIL Detection&Clustering

- NIL Detection
- NIL 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**

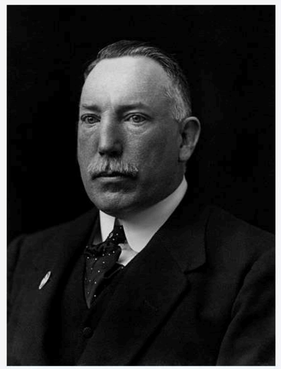
- Prior probability computation
  - Based on pre-calculated **prior probabilities** of correspondence between **certain mentions** and **entities**, p(e|m).
  - e.g. Compute p(e|m) based on **Wikipedia entity hyperlinks (e)** and **anchor text (m)**.
    Based on **CrossWikis,** leveraging the **frequency** 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) Jil Jarmyn (1959—1962) (divorced) Mary June Ray (?–?)
Children	3

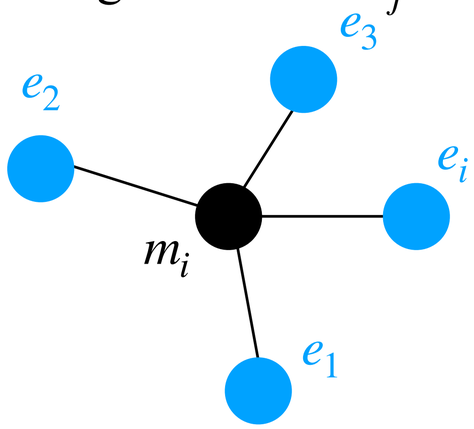
Candidate Generation

- Example

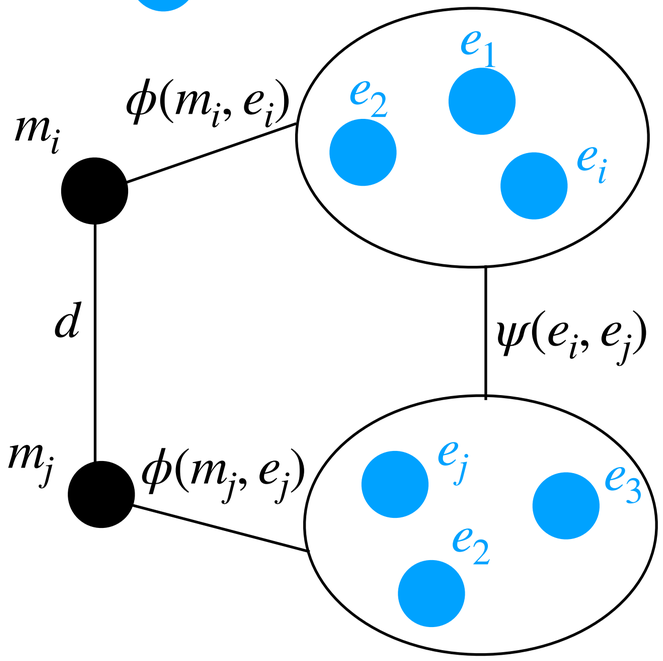
Method	5 sample candidate entities for the example mention "Big Blue"
surface form matched based on DBpedia names	Big_Blue_Sky, Big_Blue_(Old_Dominion), Big_Blue_Crane_collapse, Dexter_Bexley_and_the_Big_Blue_Beastie, Big_Bluegrass
expansion using aliases on YAGO means	Big_Blue_River_(Indiana), Big_Blue_River_(Kansas),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

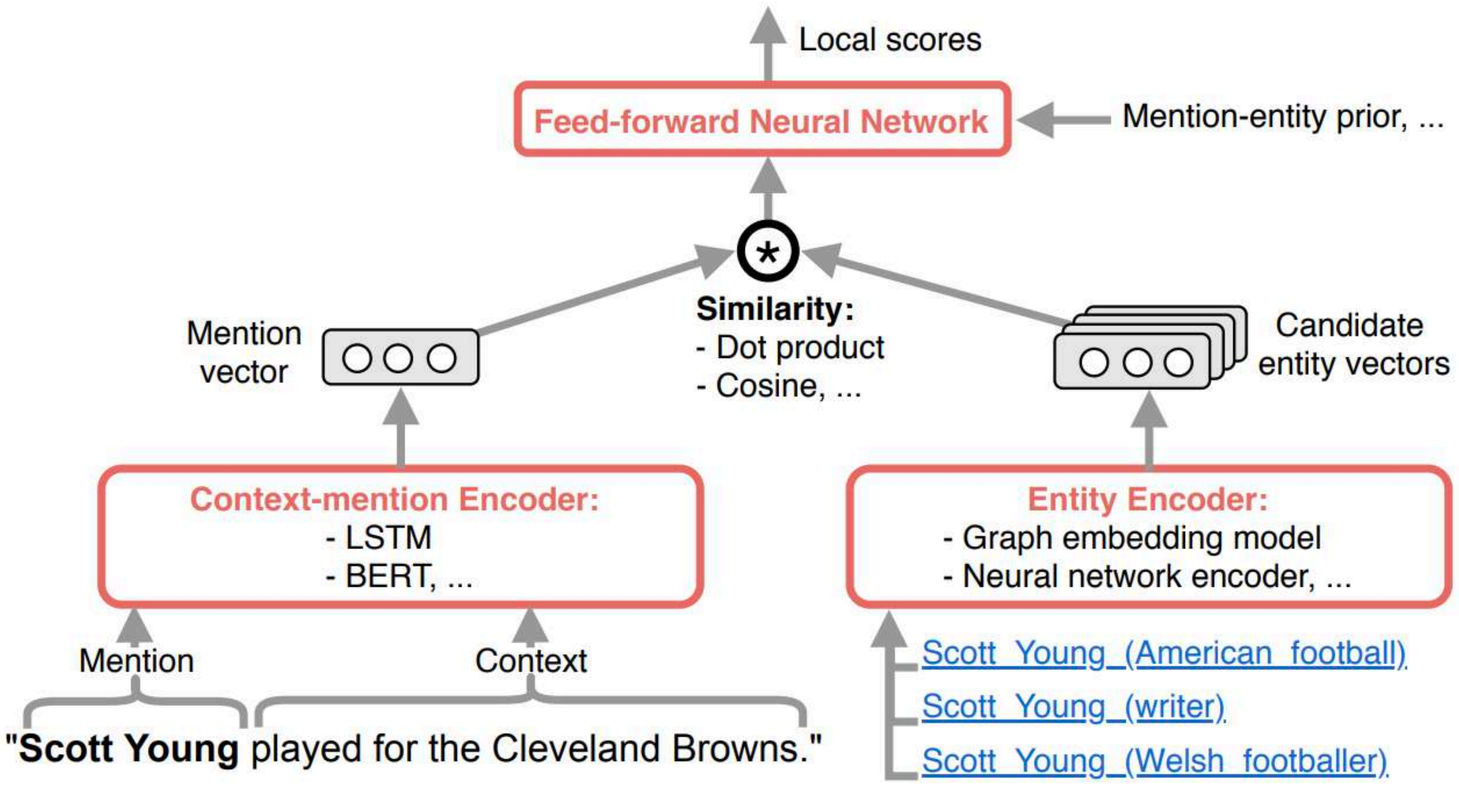


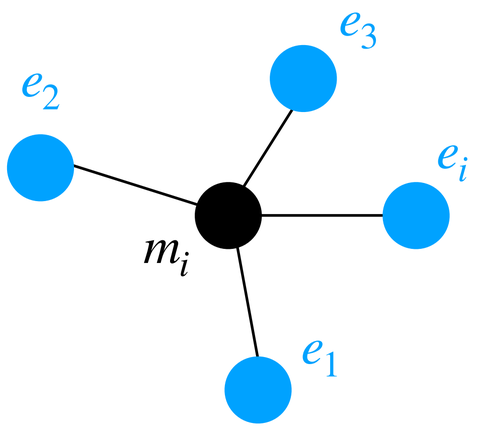
  - Global Inference
    - For each document d,
      - Consider all $m_i \in d$, and all $e_i \in T(m_i)$
      - Re-rank entities $e_i \in T(m_i)$
        (i.e., the corresponding e_i of related m_i may also be related)



Local Ranking

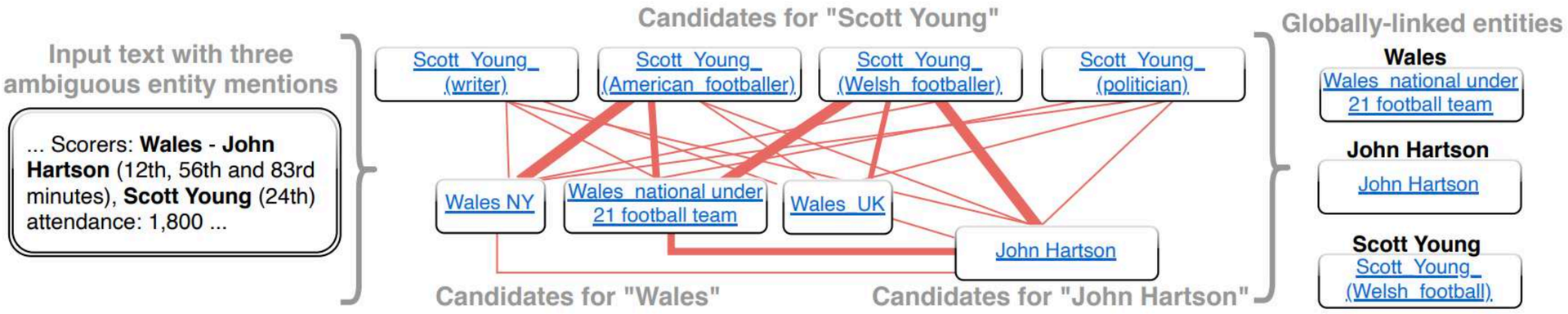
- **Basic idea:**
  - Use **similarity measure** between **candidates** $e_i \in T(m_i)$ and **mention** m_i **with its context** c_i.

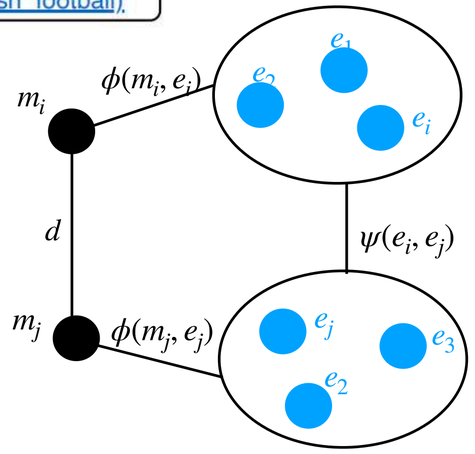




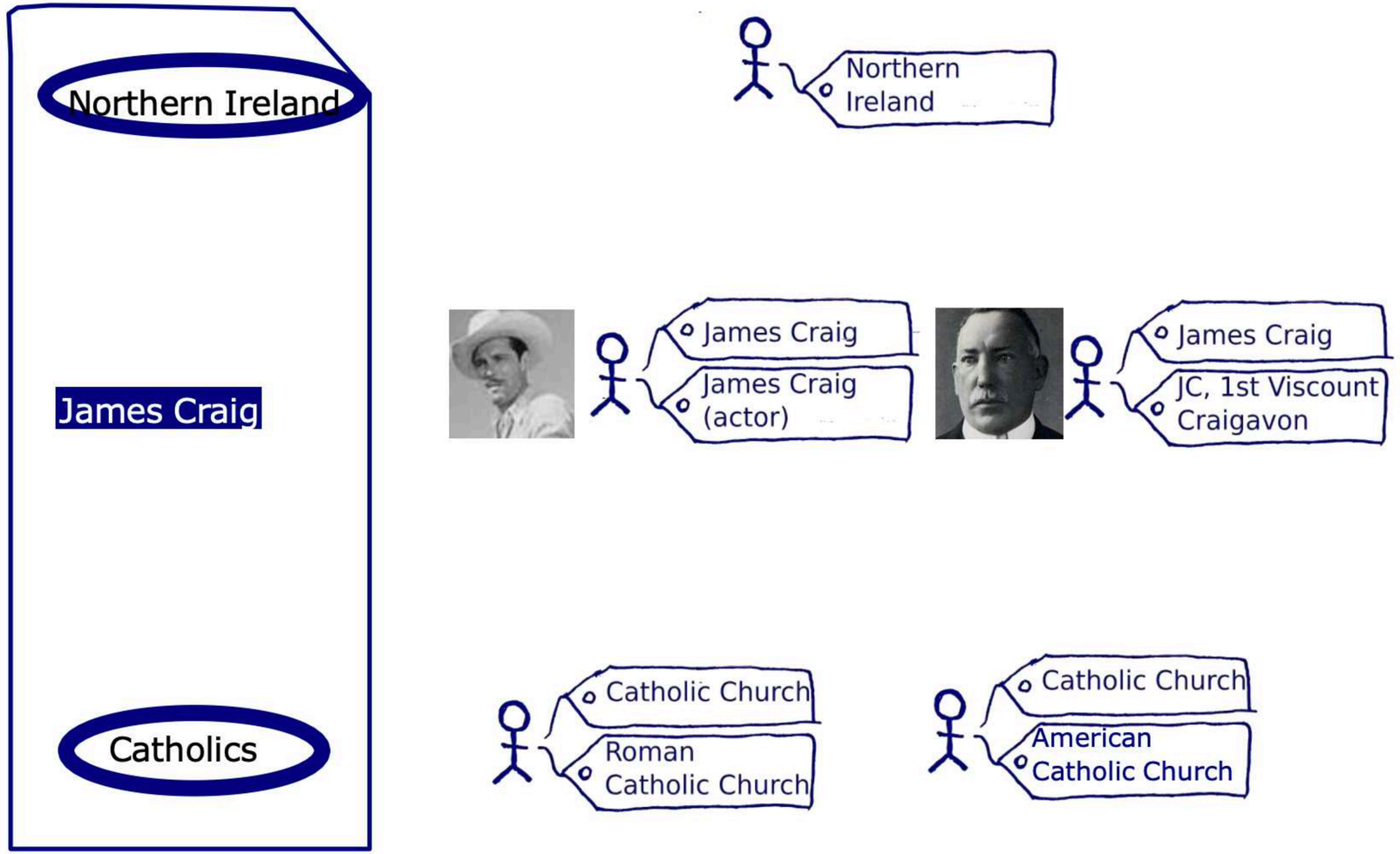
Global Ranking

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

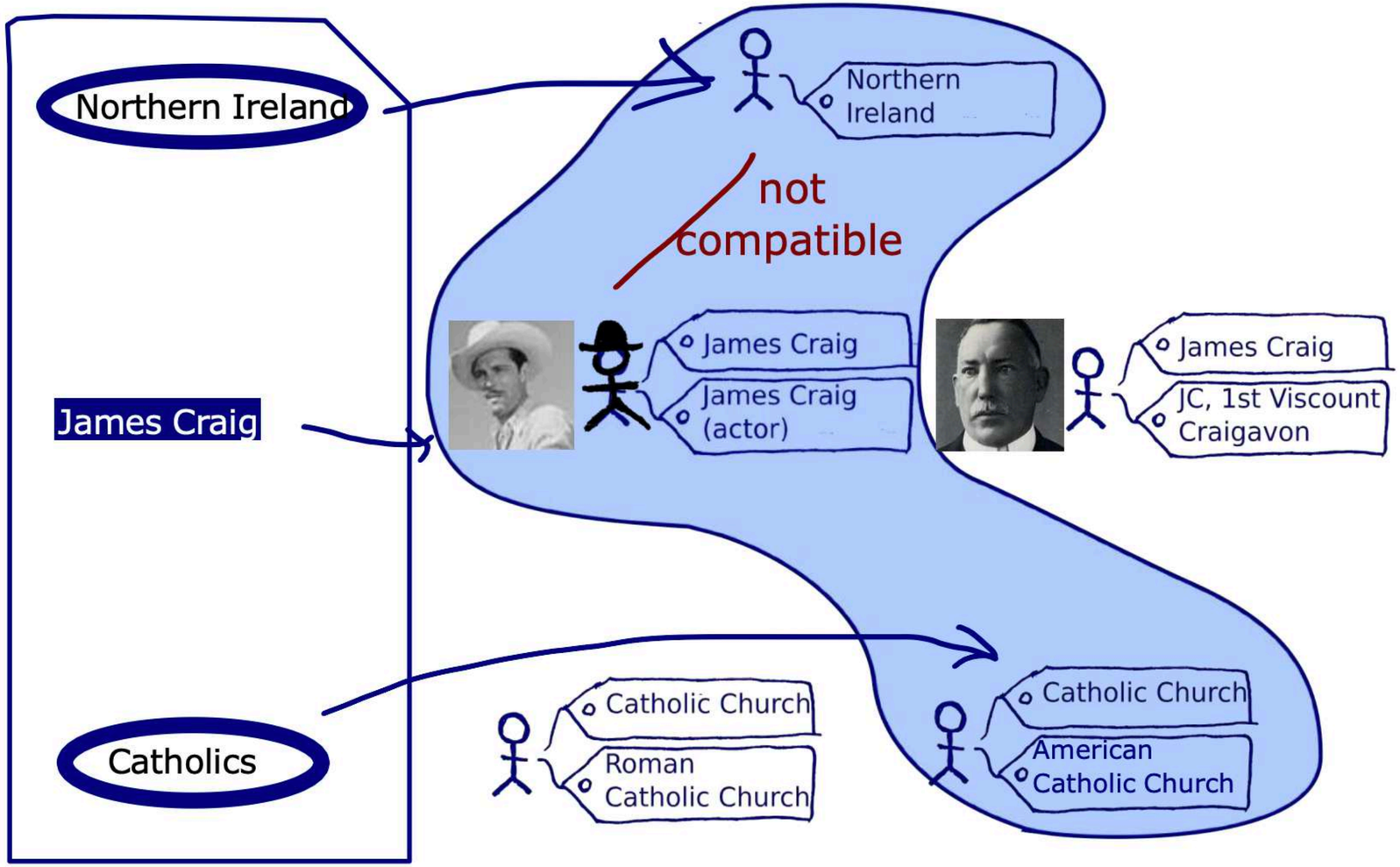




Global Ranking: Illustration



101 Global Ranking: Illustration



Global Ranking: Formulation

- A Combinatorial Optimization Problem

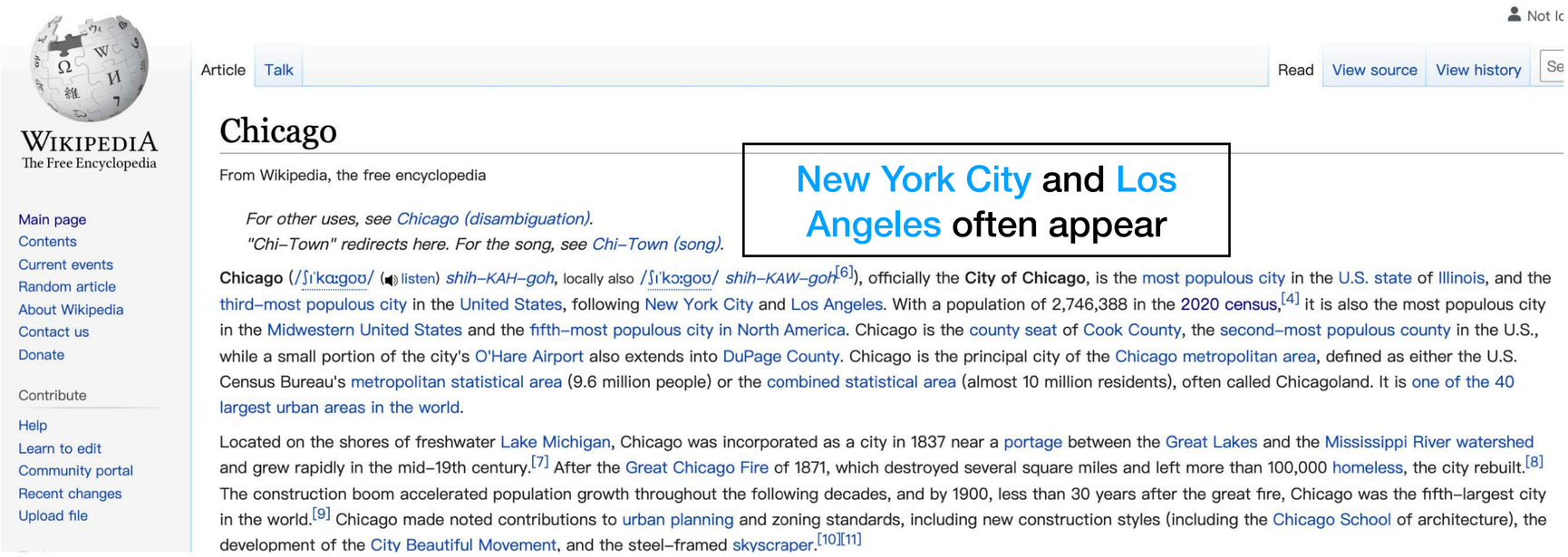
- Problem Formulation

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

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

Entity Coherence & Relatedness

- Co-occurrence



- 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.
- …

NIL Detection & Clustering

- NIL Detection

{ WIKIPEDIA The Free Encyclopedia , NIL }

Augment KB with **NIL entry** and **treat it like any other entry**

Marc-Alexandre Côté is an author of Textworld.

James Craig was born in Nashville, Tennessee and studied at the Rice Institute

James Craig attempted to announce his bid for Governor on Belle Isle,…

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

NIL Detection & Clustering

- NIL Clustering

"All in one"

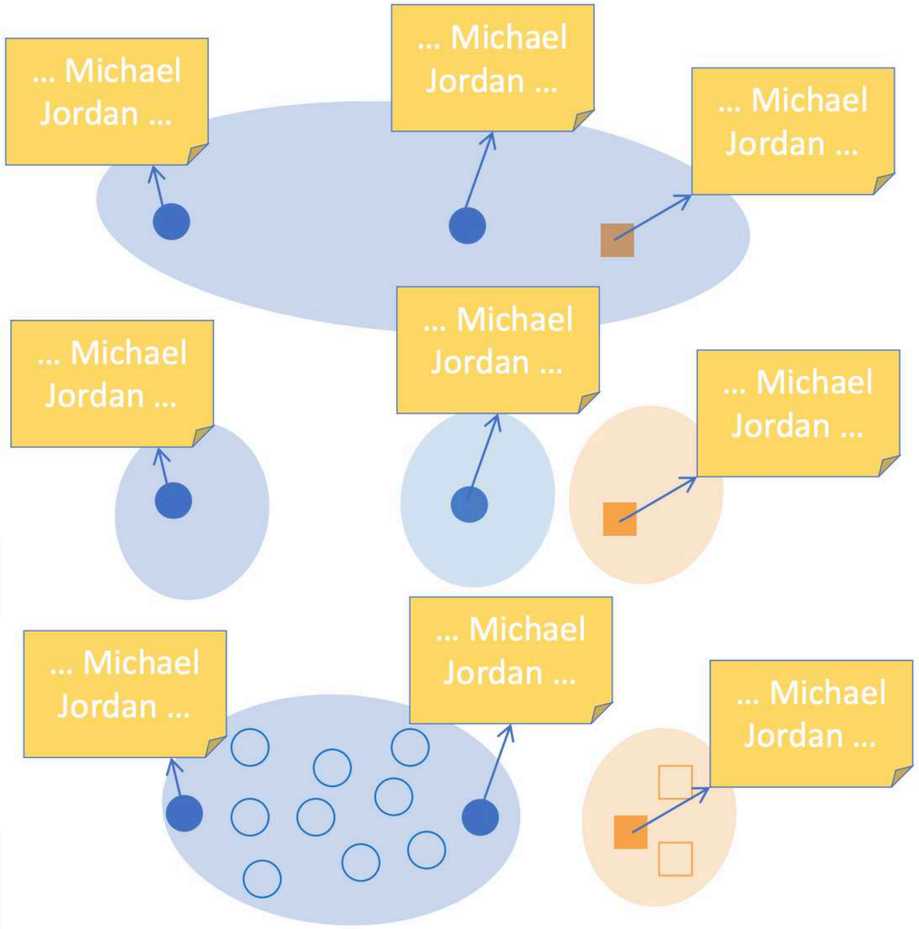
Simple string matching

"One in one"

Often difficult to beat!

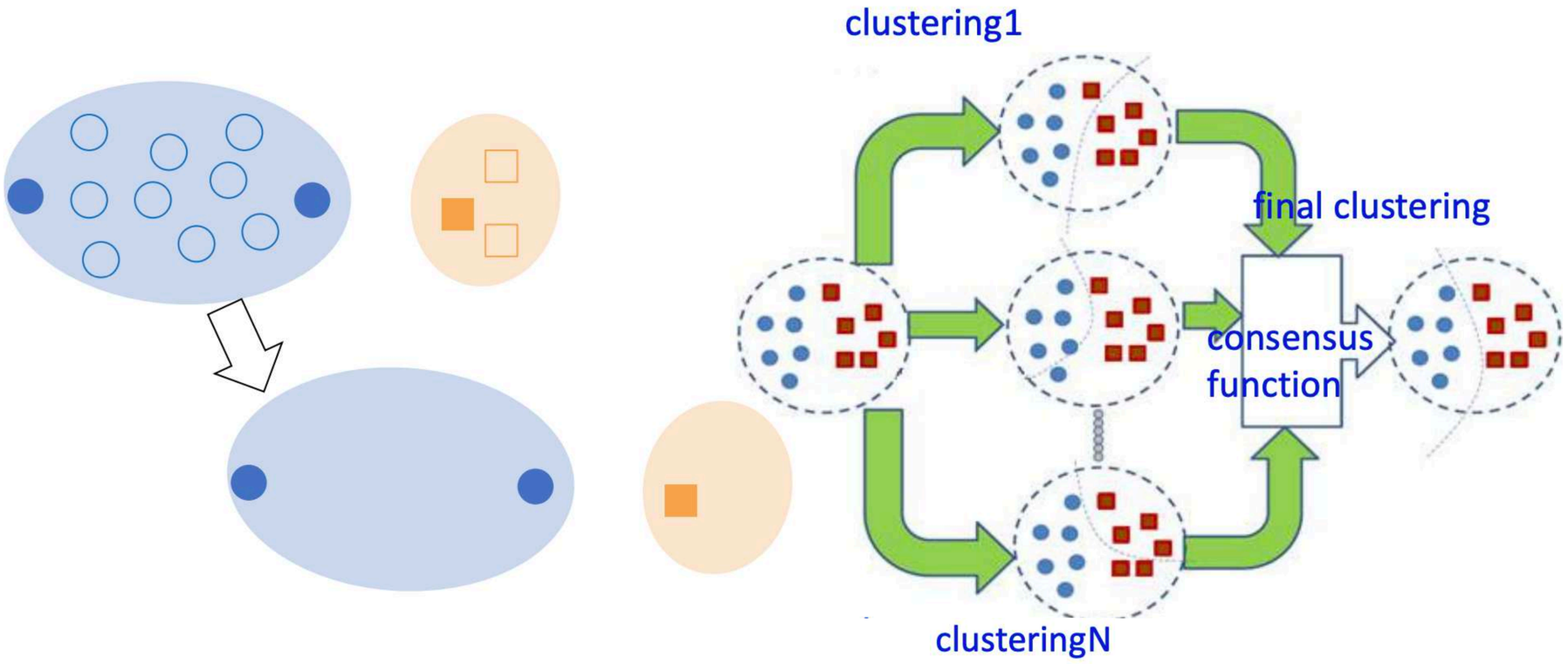
Collaborative Clustering

Most effective when ambiguity is high



NIL Detection & Clustering

- Collaborative Clustering



EL Tools

- Tools (Target KB: Wikipedia):
  - FacebookResearch BLINK
    - **SOTA** of **zero-shot** EL
    - Well-documented
  - OpenAI DeepType
    - SOTA of EL
  - Spacy Entity Linker
    - Easy to use

spaCy

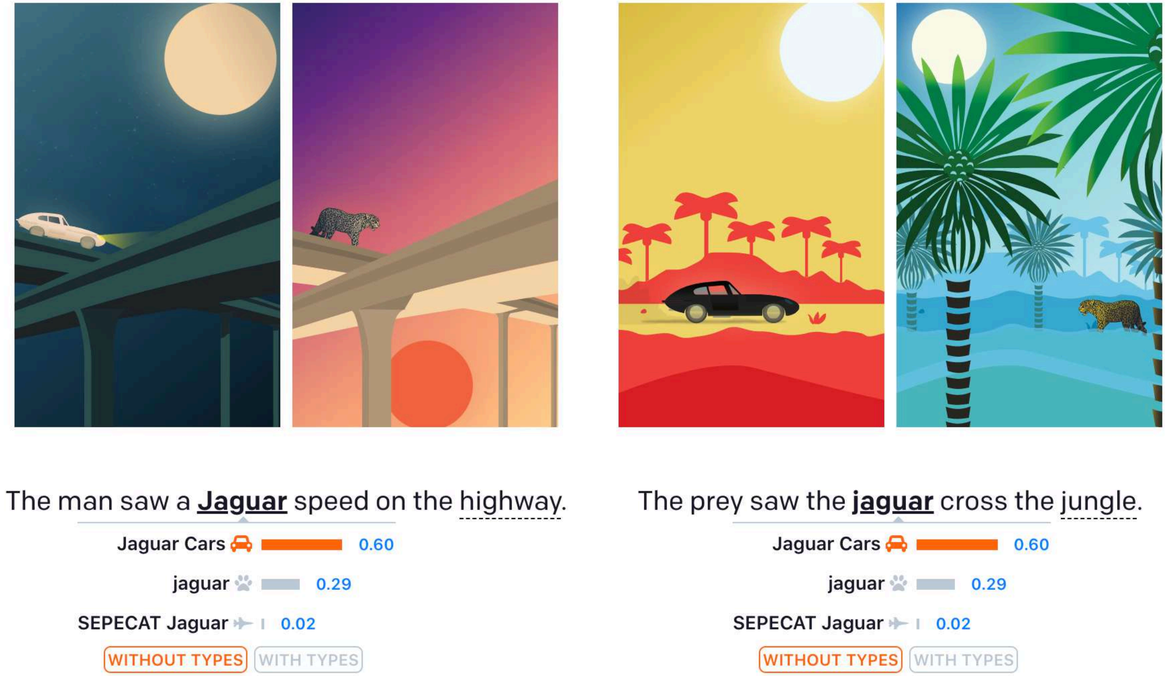
```
import spacy  # version 3.0.6'

# 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
```

https://github.com/egerber/spaCy-entity-linker

EL Tools

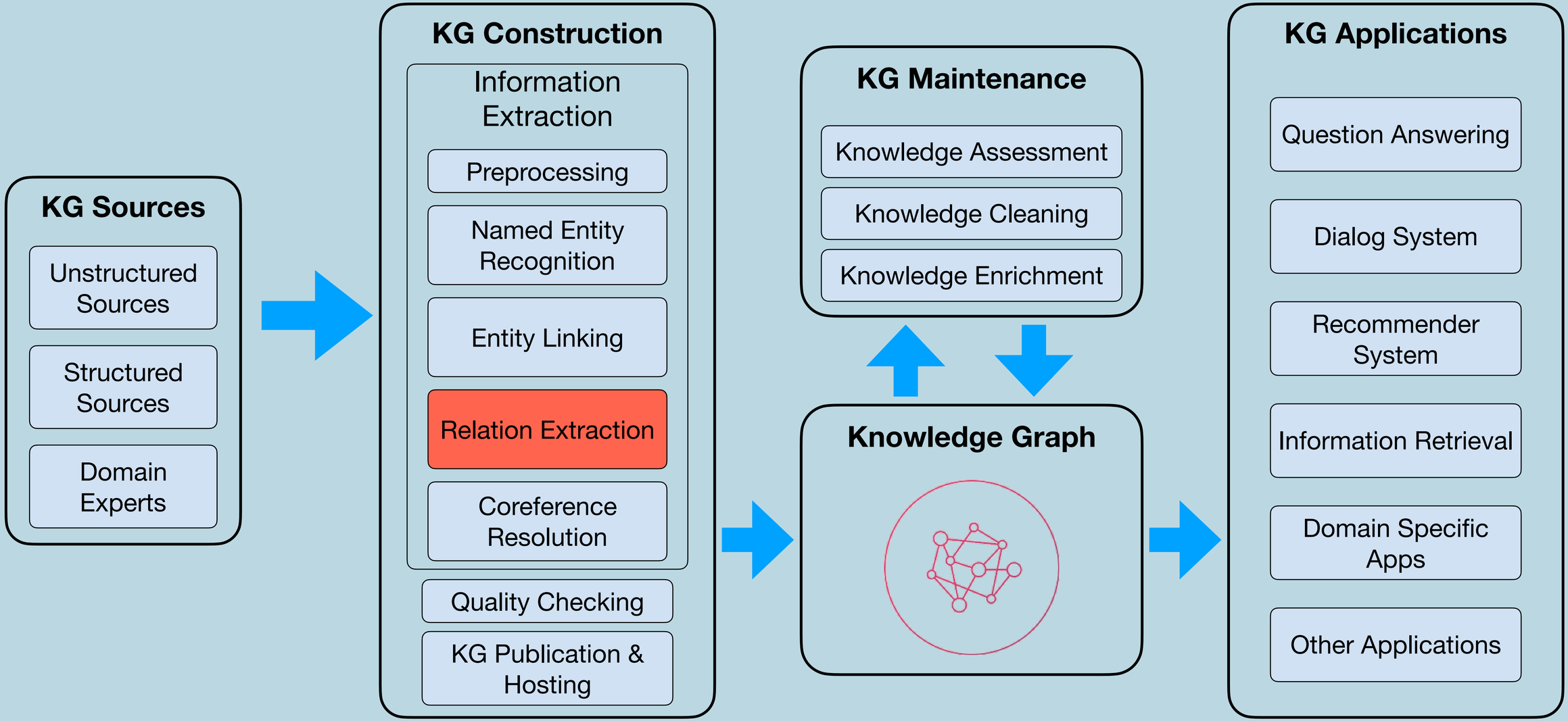
- **OpenAI DeepType**
  Figuring out which object is meant by a word by having a neural network decide if the word belongs to each of about 100 automatically-discovered “types”.

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.**



https://openai.com/blog/discovering-types-for-entity-disambiguation/

Relation Extraction



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

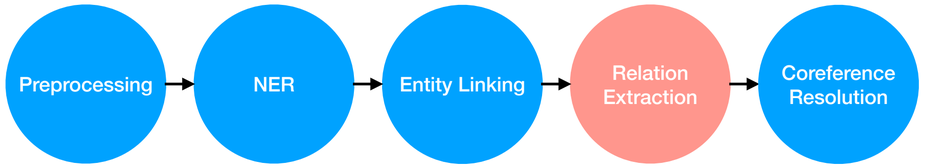
Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

- 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

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.[3] 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.[4]

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.

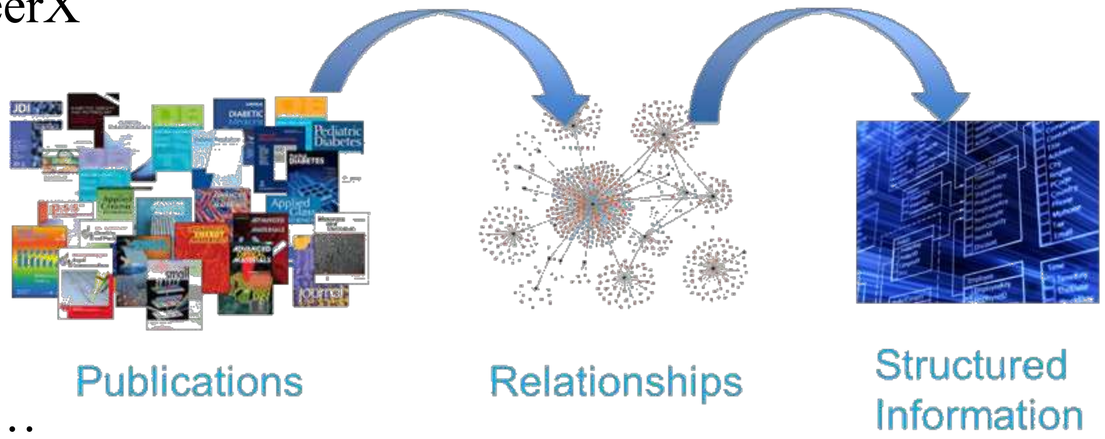
Relation Extraction

ORG-AFF PHYS ORG-AFF

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.

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
- …



But what relations should we extract?

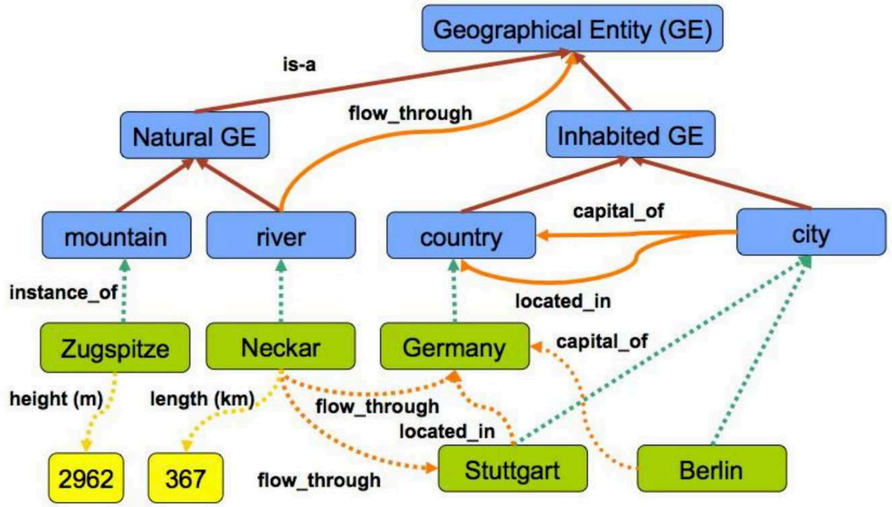
- Depends on the **domain of interest**.
  - For generic news texts

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



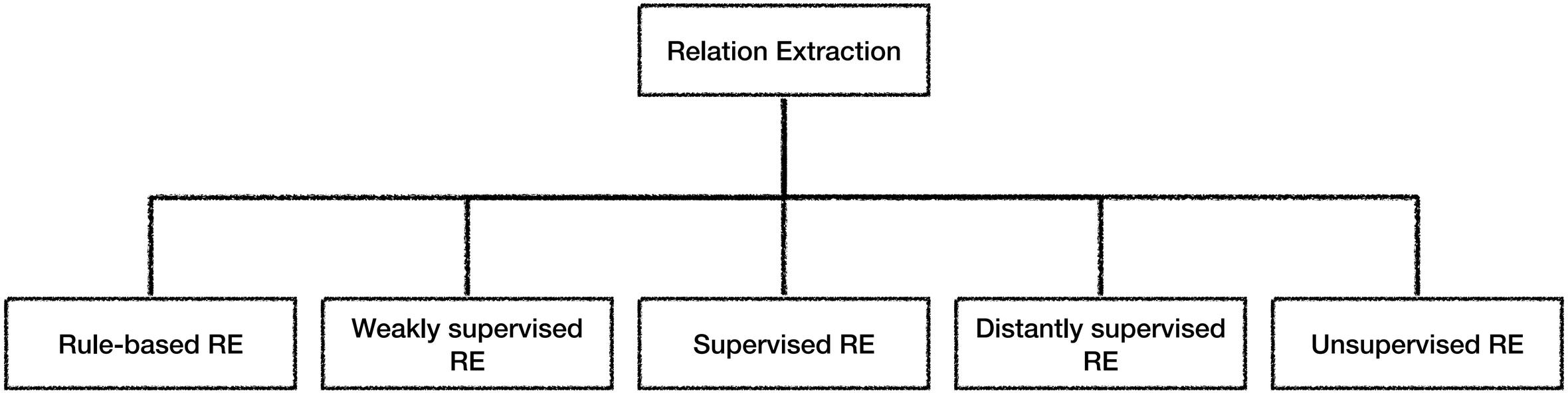
  - Protein interactions

„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 ← interact complex → CBF-C

CBF-B → associates → CBF-A-CBF-C complex

Relation extraction approaches



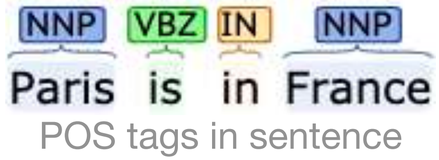
Rule-based RE

- 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

(e_{subj}, r, e_{obj})

(**Paris**, **is_in**, **France**)

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

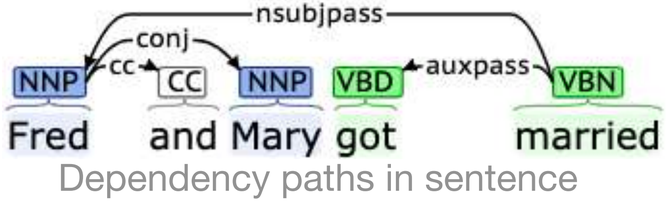


(CITY, *r*, COUNTRY)

(NNP, *r*, NNP)

(**Paris**, **is_in**, **France**)

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



Dependency-based rules

(**Fred**, **got_married**, **Mary**)

Rule-based RE

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

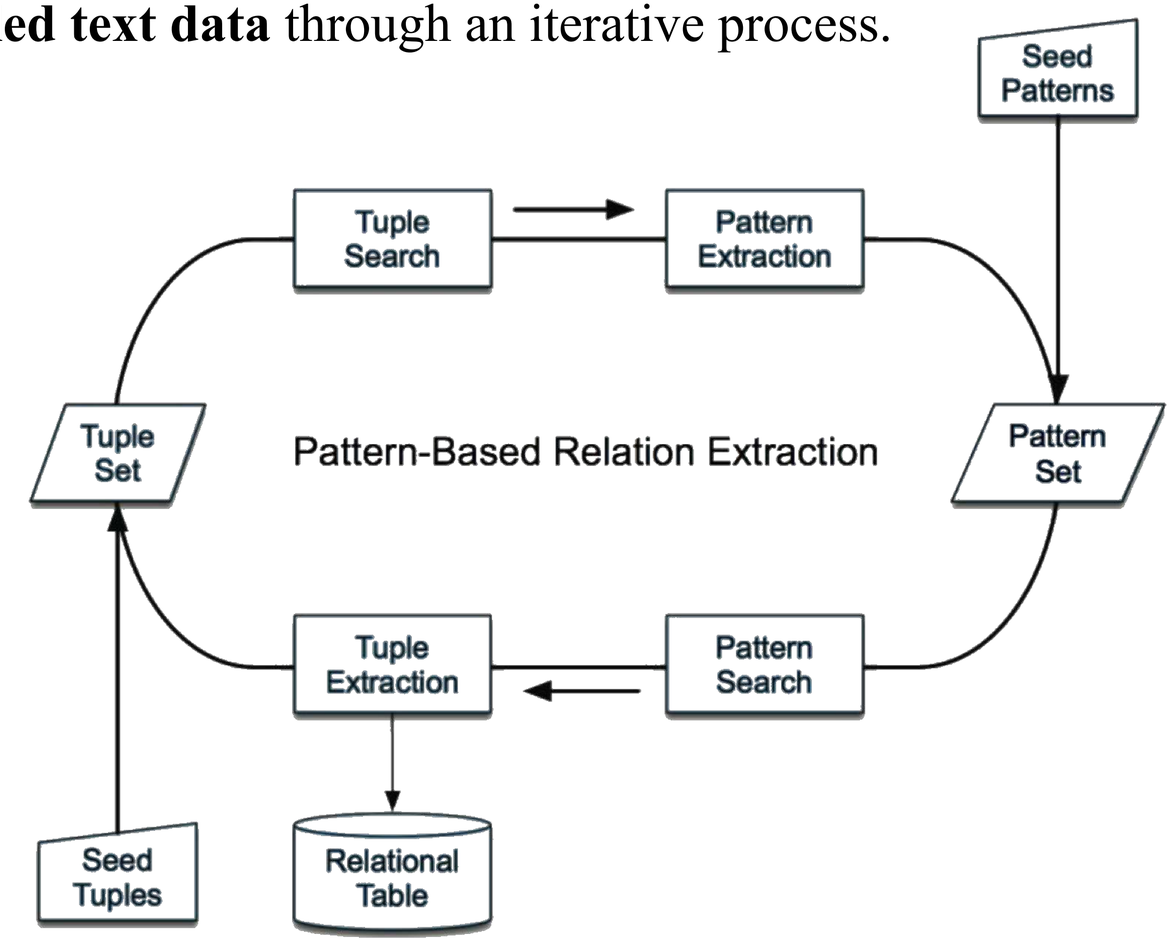
- **Cons**
  - Human patterns suffer from **low-recall**
  - **Creation of patterns** can be **expensive** or **even intractable**
    - Hard to write and maintain the rules
    - Have to create rules for every relation type

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.



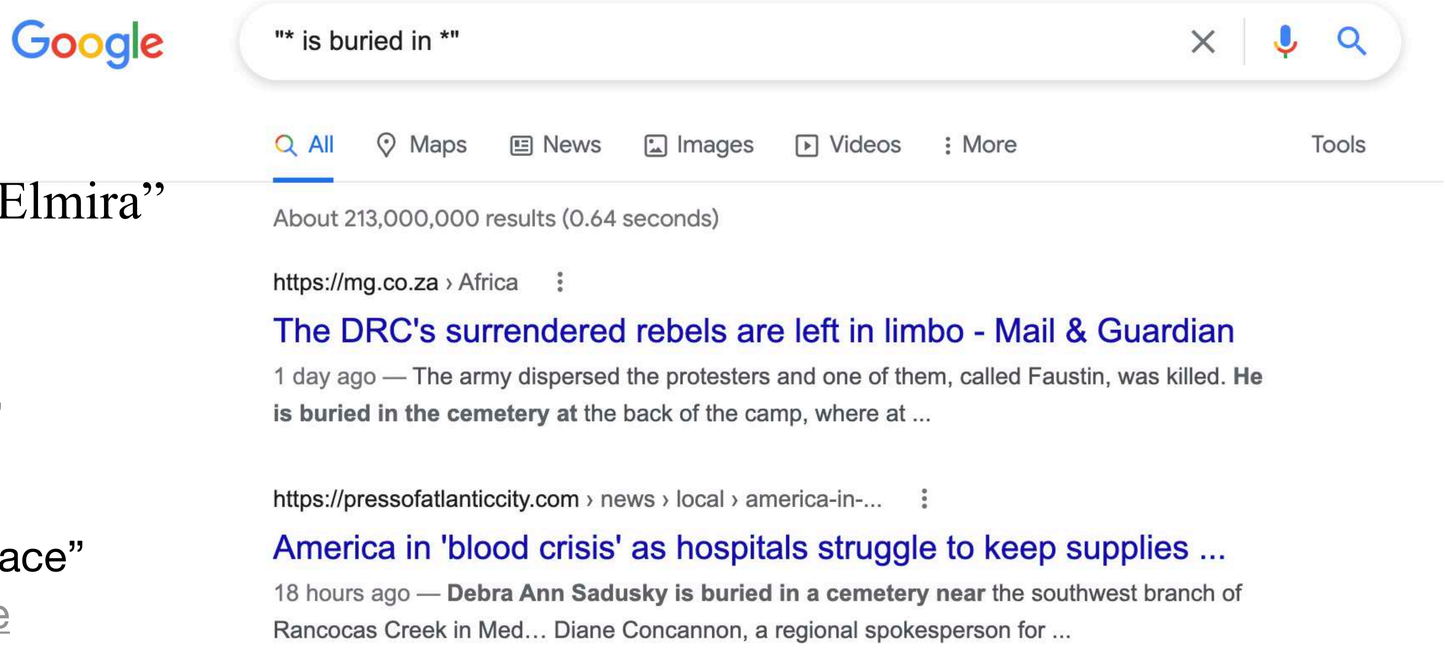
Weakly Supervised RE: Example

- **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



Weakly Supervised RE

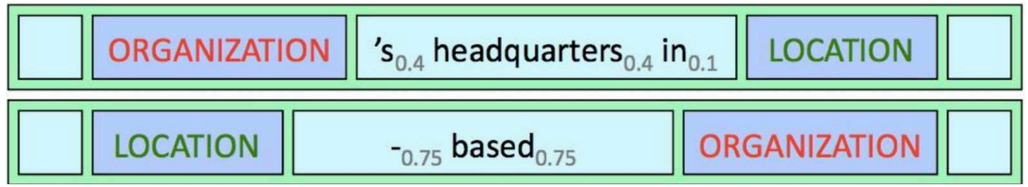
- **Pros**
  - **Higher recall**: more relations can be discovered
  - **Less cost**: only require a high quality seed

- **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

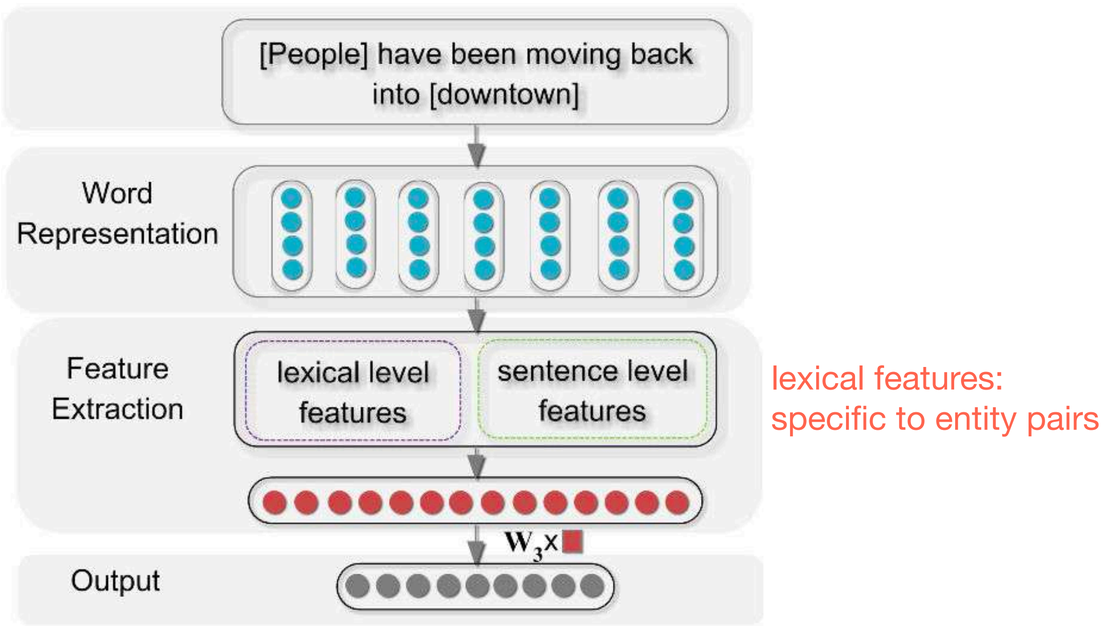


Supervised RE

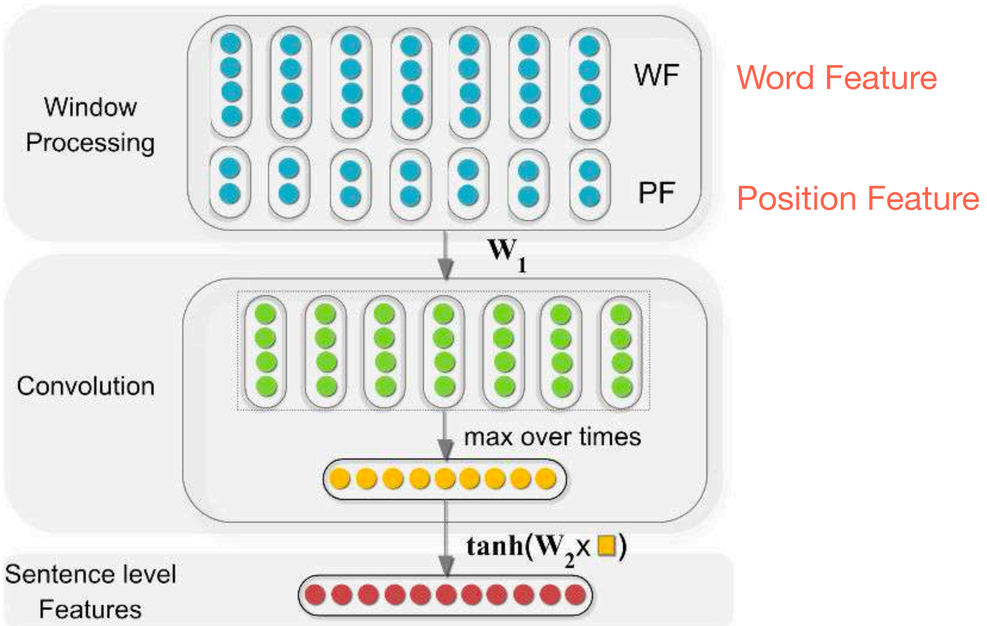
- The supervised approach requires:
  - Defining an inventory of output **labels**
    - Relation detection: True/False
    - Relation classification: located_in, employee_of, be_born_in, …
  - Collection sufficient **labeled training data** (Hard Part)
    - Public benchmarks: MUC-7, ACE-2005, SemEval, 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**

How are neural networks being used to do RE?

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



(a) Neural Network Architecture for Relation Classification.

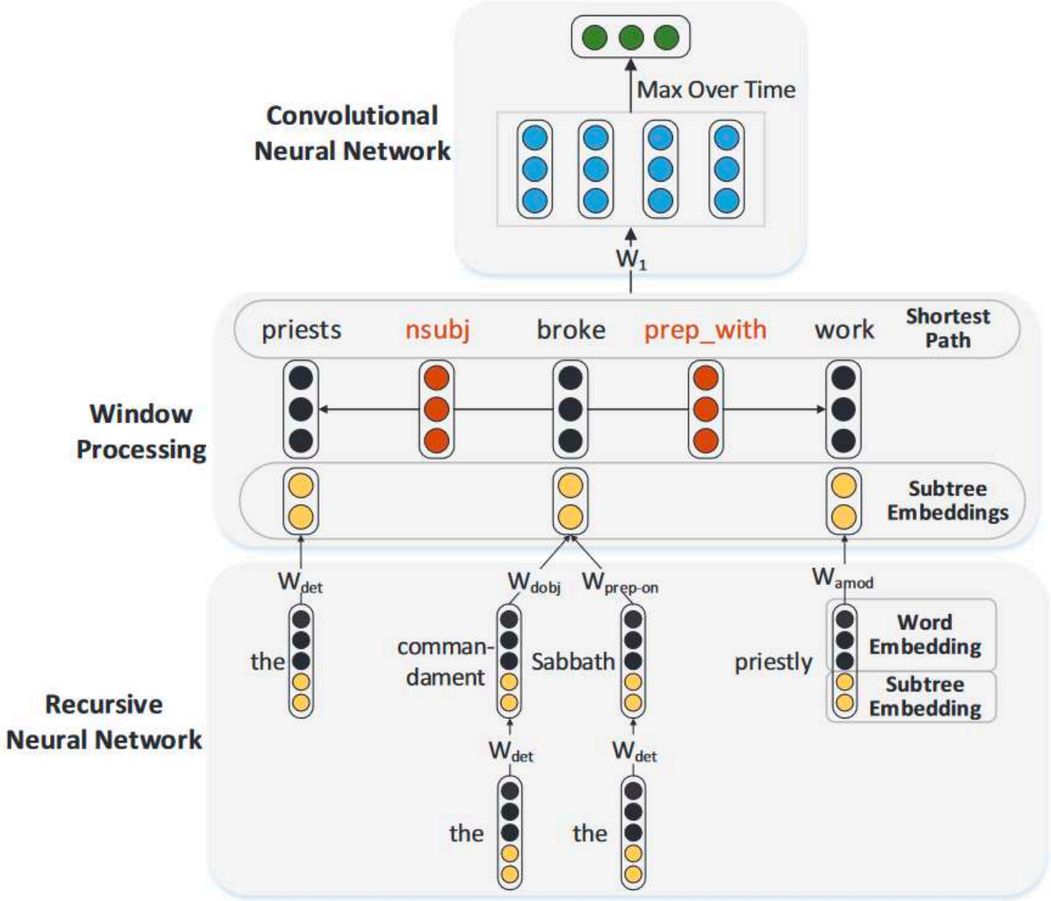


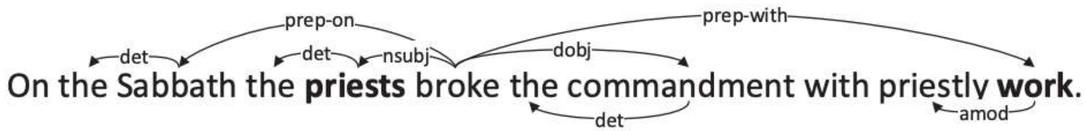
(b) Extracting Sentence-level Features using CNN.

https://aclanthology.org/C14-1220/

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.



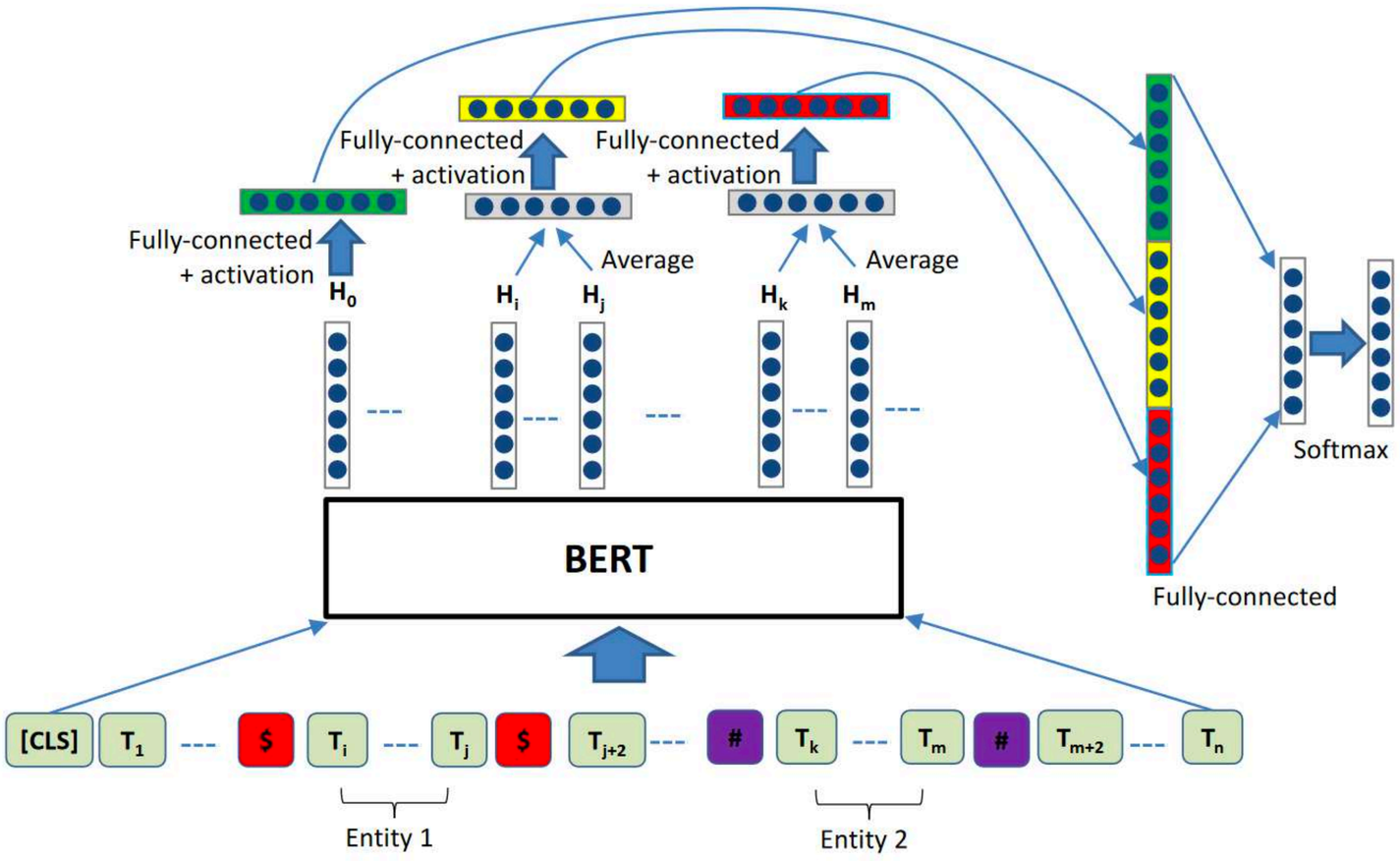


1. Get subtree representations by Recursive NN
2. Get sentence level representation by CNN through the shortest path on dependency tree from subject to object.

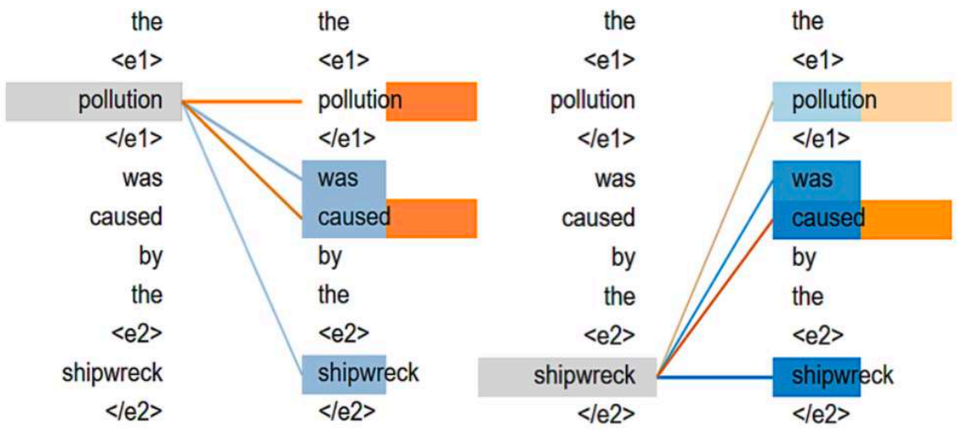
https://aclanthology.org/P15-2047/

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



https://arxiv.org/abs/1905.08284

Supervised RE

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

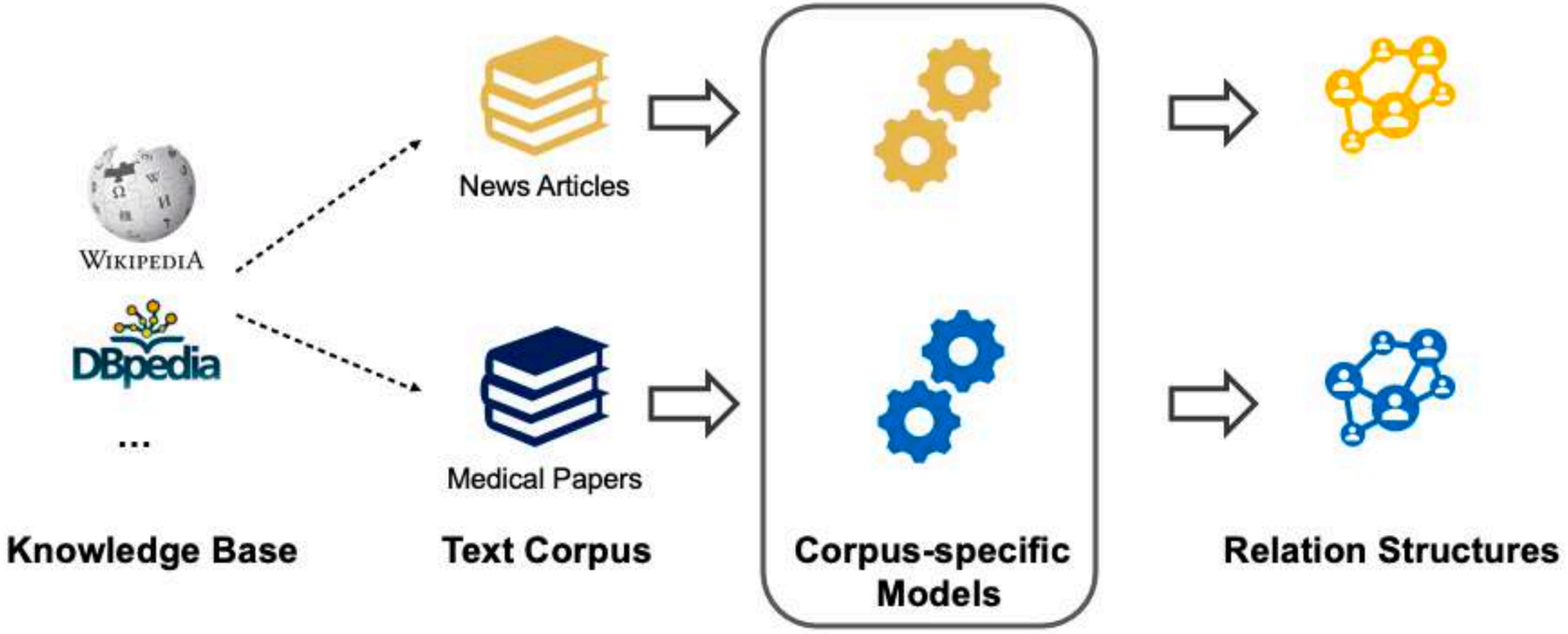
- **Cons:**
  - Performance is limited by the amount of **expensive** labeling samples
  - Does **not generalize** well to **new domains**

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.



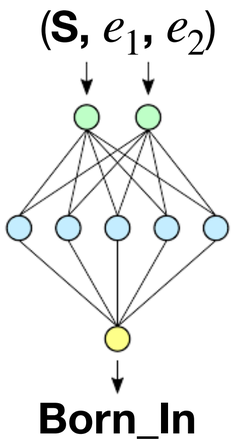
Distantly Supervised RE: Illustration

- 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



128 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**.

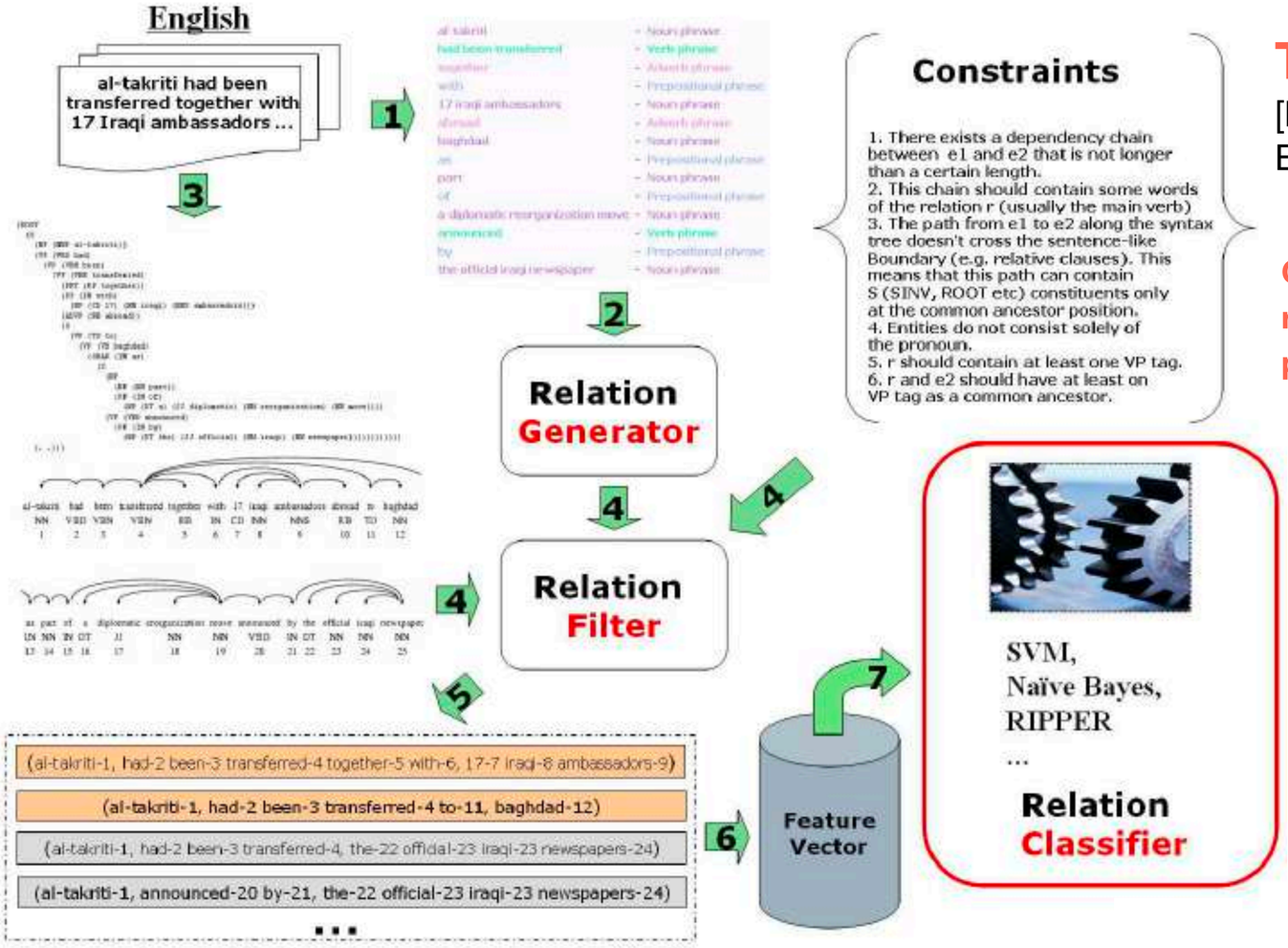
Distant supervision for relation extraction without labeled data

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

https://aclanthology.org/P09-1113/

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**.



Constraints

1. There exists a dependency chain between e1 and e2 that is not longer than a certain length.
2. This chain should contain some words of the relation r (usually the main verb)
3. The path from e1 to e2 along the syntax tree doesn't cross the sentence-like Boundary (e.g. relative clauses). This means that this path can contain S (SINV, ROOT etc) constituents only at the common ancestor position.
4. Entities do not consist solely of the pronoun.
5. r should contain at least one VP tag.
6. r and e2 should have at least on VP tag as a common ancestor.

TextRunner

[Bach, Nguyen, and Sameer Badaskar 2007]

Constraints are not relation-specific patterns.

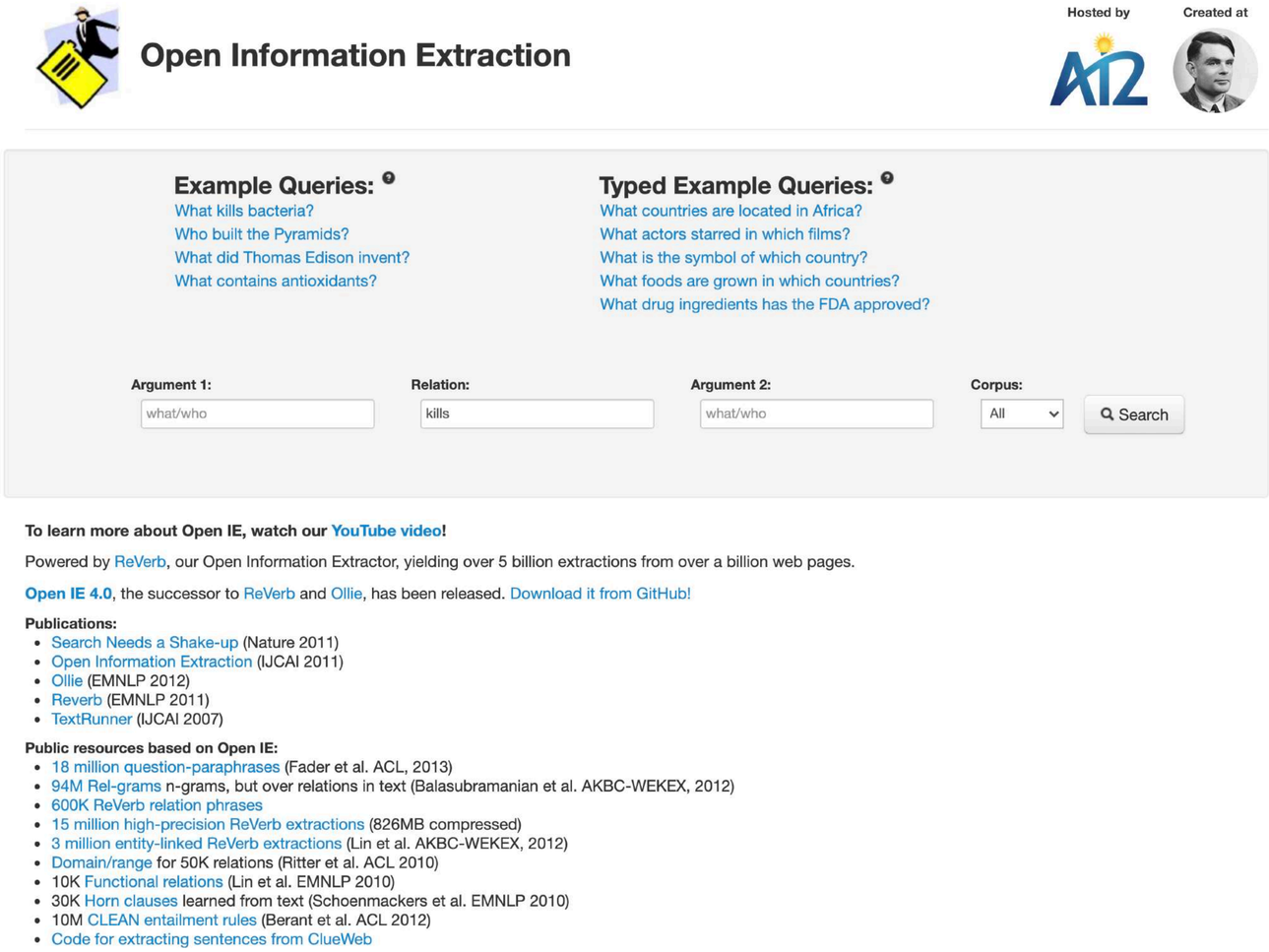
Unsupervised RE

- **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 defintion 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



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.
  - …

OpenIE

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: what/who

Relation: verb phrase

Argument 2: what/who

Corpus: All

Search

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.

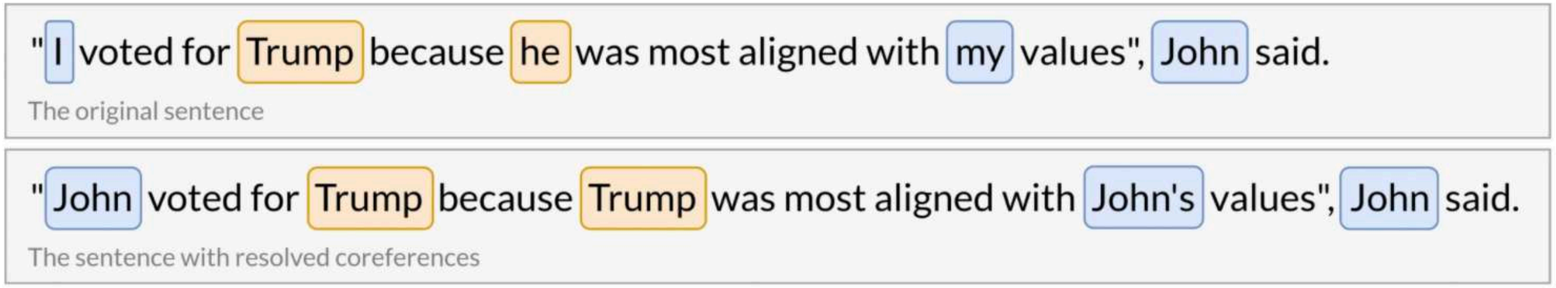
Open IE 4.0, the successor to ReVerb and Ollie, has been released. Download it from GitHub!

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**.

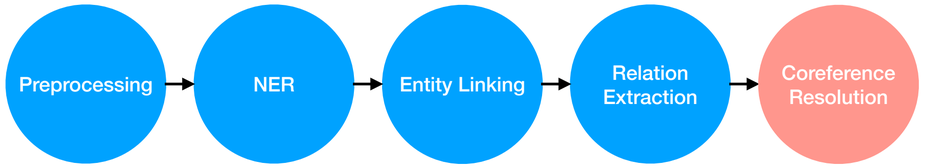


Why Coreference Resolution?

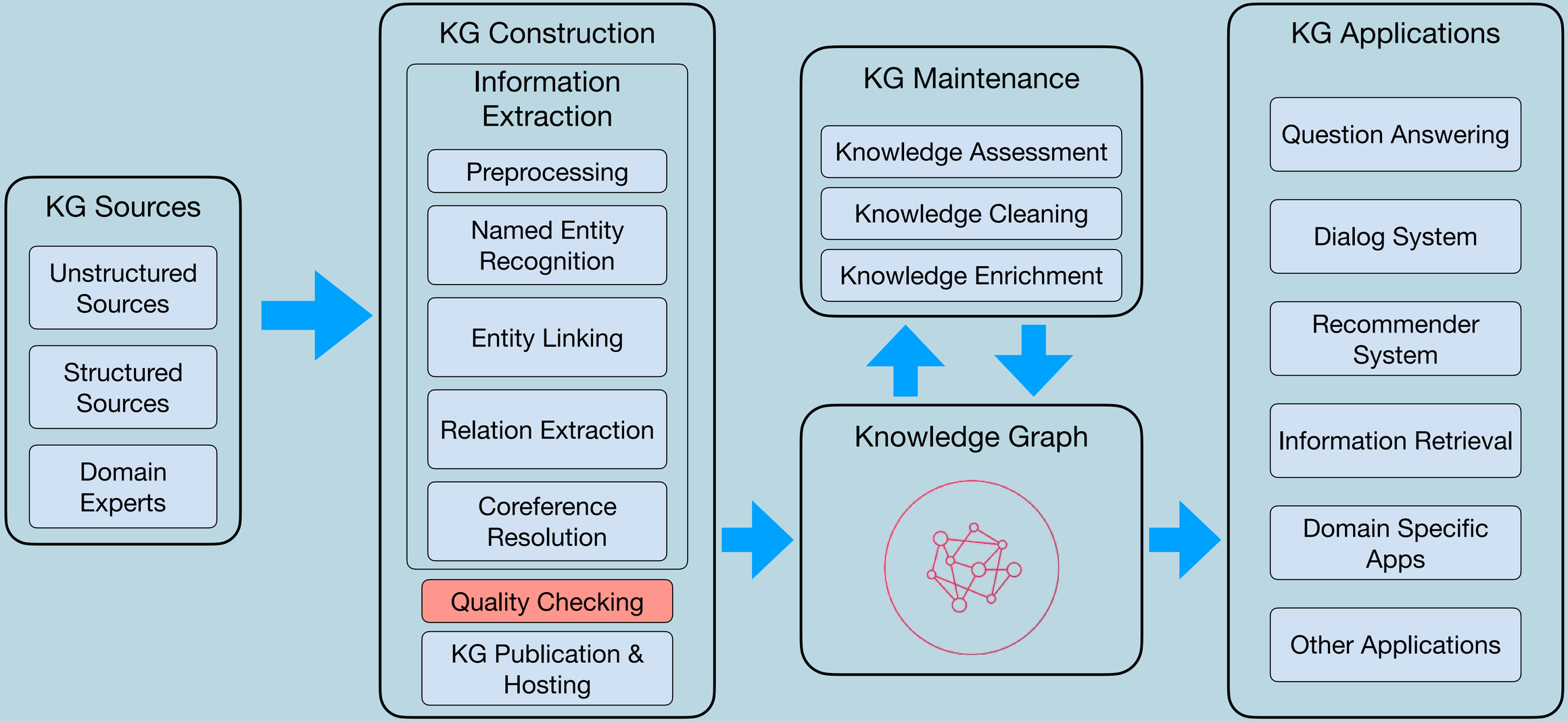
- Coreference resolution could **increase** the overall **efficiency** and the **coverage** of the information extraction.

Tools:

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



Quality Checking



Quality Checking

- Not everything accurately extracted is **fact**

Raw Text

"I voted for Trump because he was most aligned with my values", John said.

The original sentence

Information Extraction

Structured Knowledge

(Subj, predicate, Obj)

(**John**, vote_for, **Trump**)

Is this a **fact**?

The claim is with the context "John said".

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.

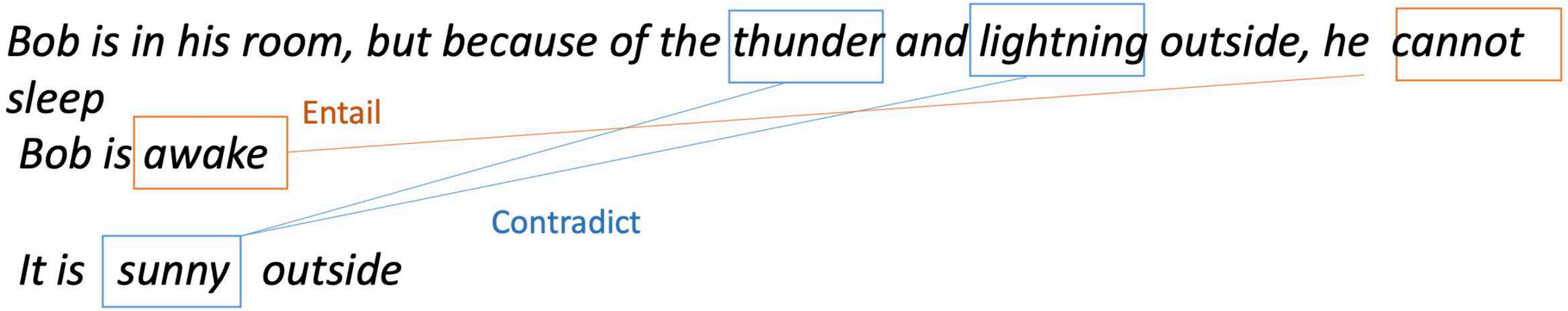
Verdict: Supported

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



Quality Checking

- FEVER Progress

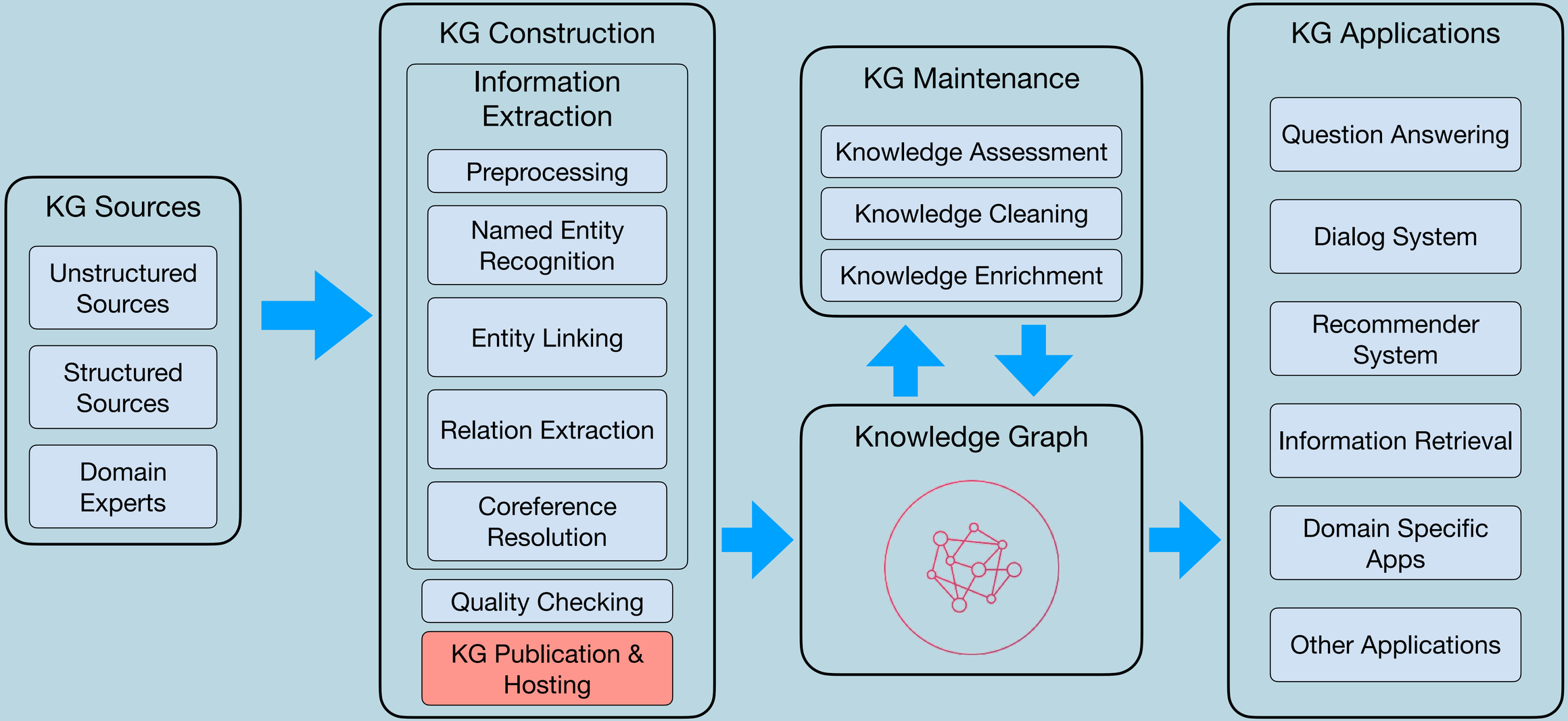
Final Leaderboard

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

https://fever.ai/task.html

KG Hosting

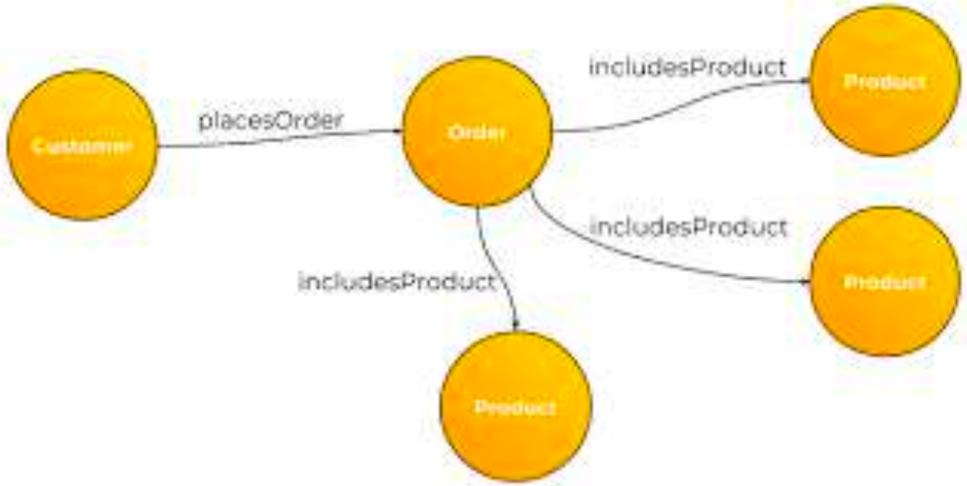

KG Sources
Unstructured Sources
Structured Sources
Domain Experts
KG Construction
Information Extraction
Preprocessing
Named Entity Recognition
Entity Linking
Relation Extraction
Coreference Resolution
Quality Checking
KG Publication & Hosting
KG Maintenance
Knowledge Assessment
Knowledge Cleaning
Knowledge Enrichment
Knowledge Graph
KG Applications
Question Answering
Dialog System
Recommender System
Information Retrieval
Domain Specific Apps
Other Applications

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**.

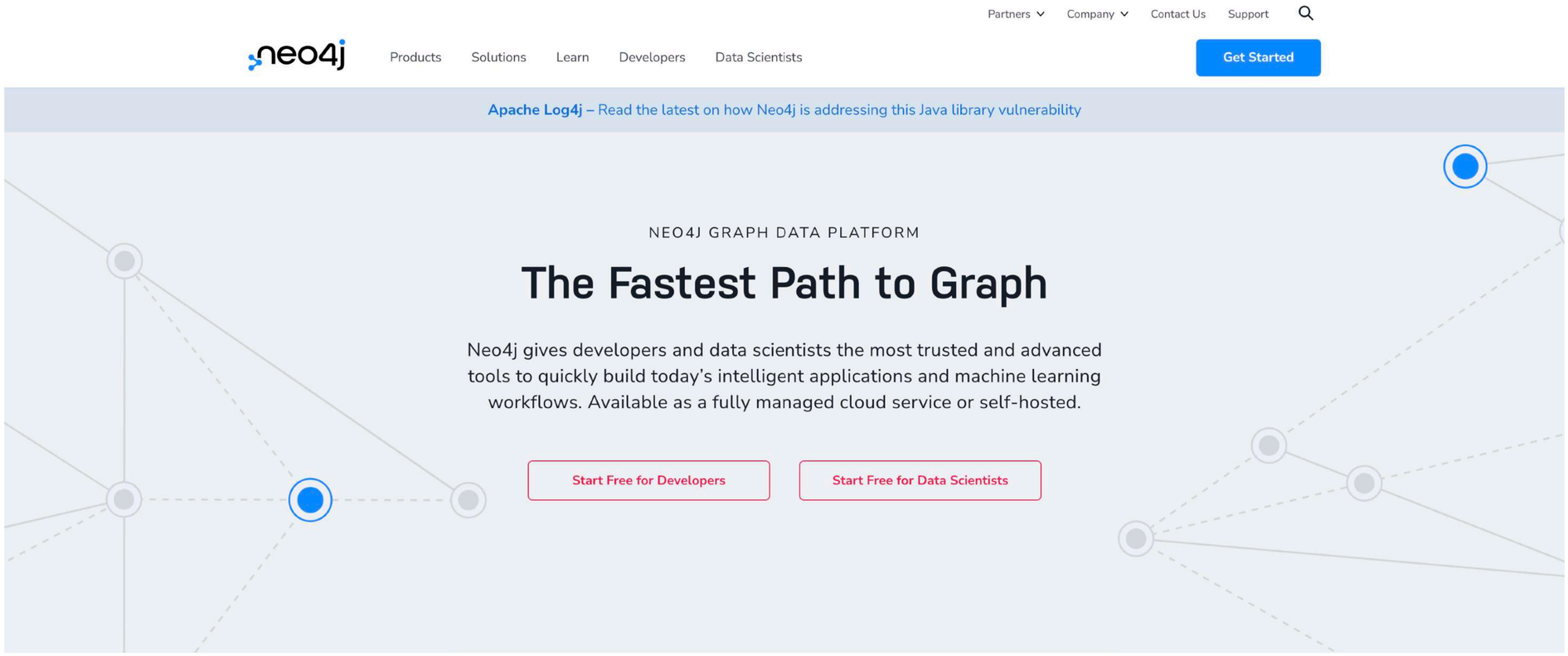


KG Hosting

- **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/



KG Hosting

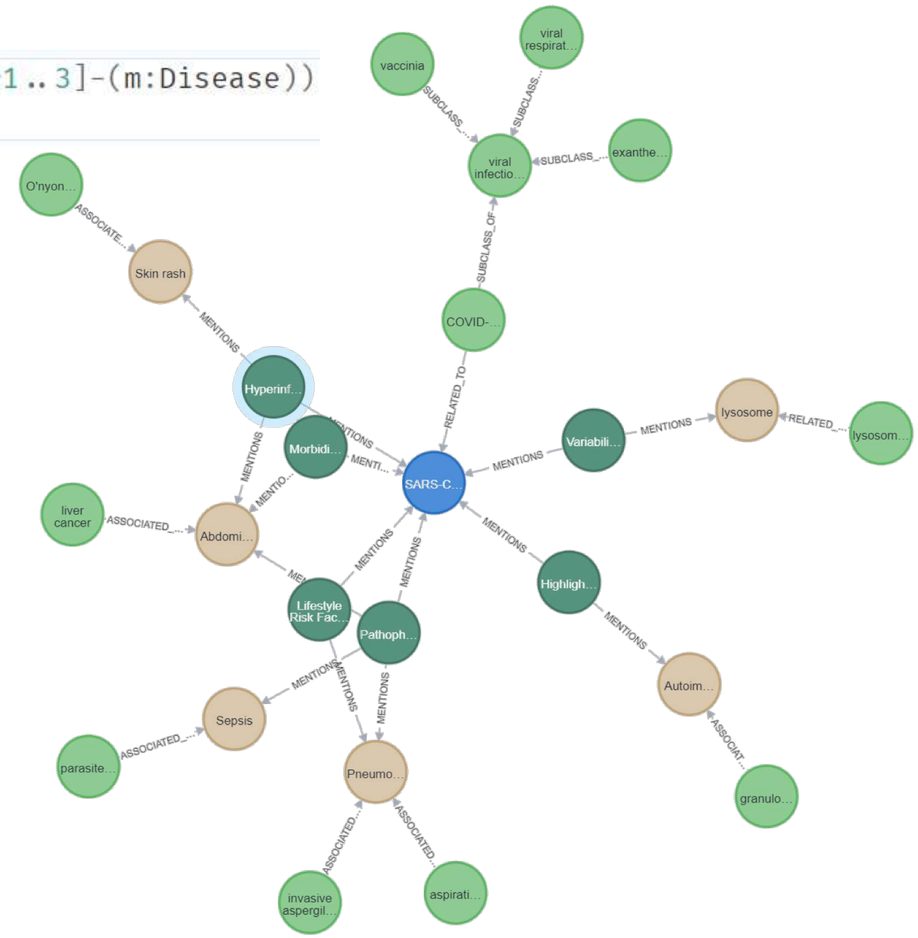
◉ Storing as **Knowledge Graph**:

- **Query:**

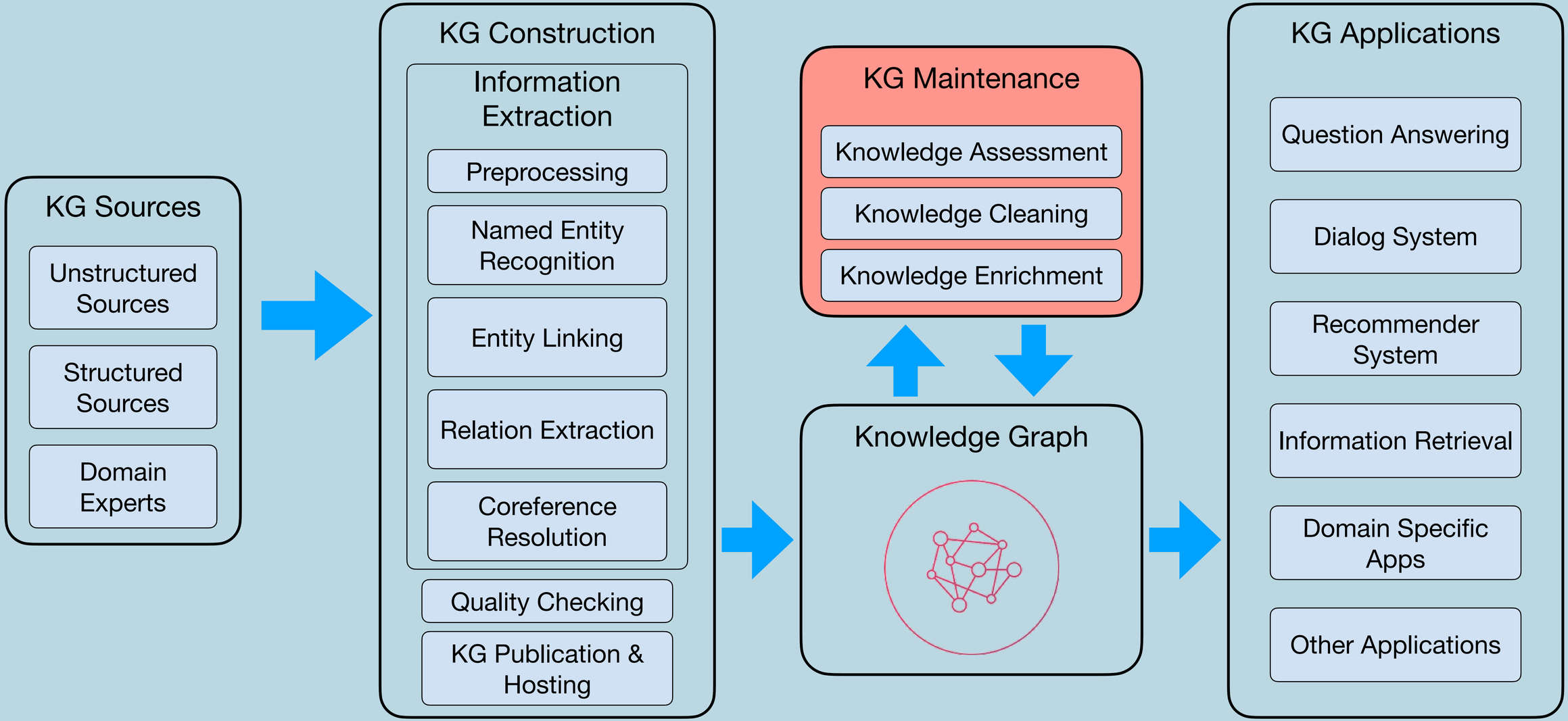
```
MATCH p=shortestPath((n {name:"SARS-CoV-2"})-[*1..3]-(m:Disease))
RETURN p LIMIT 10;
```

  - **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

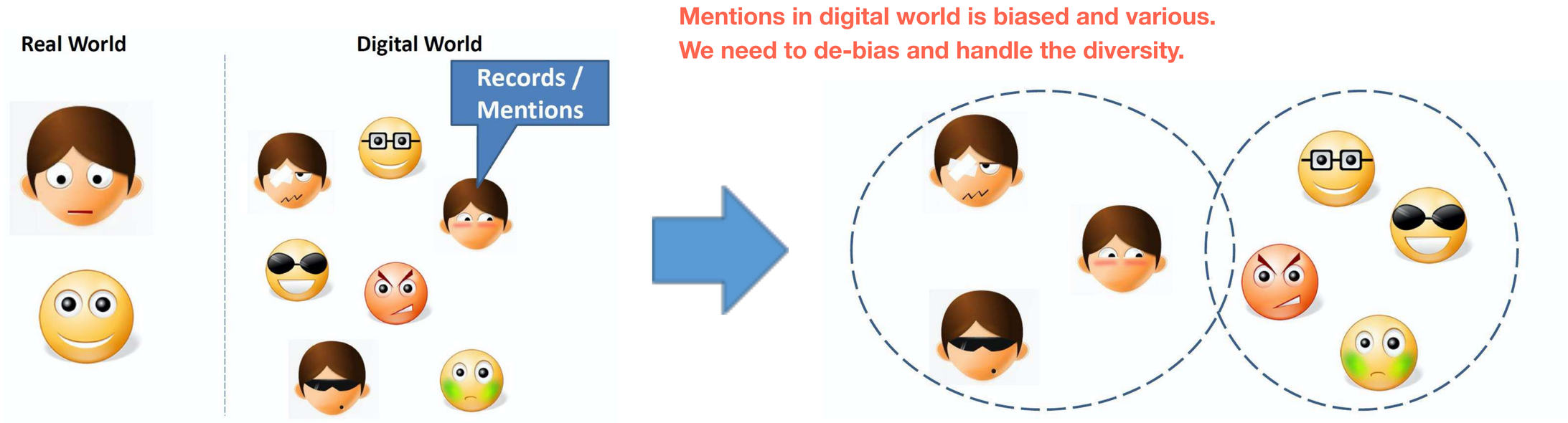


KG Maintenance



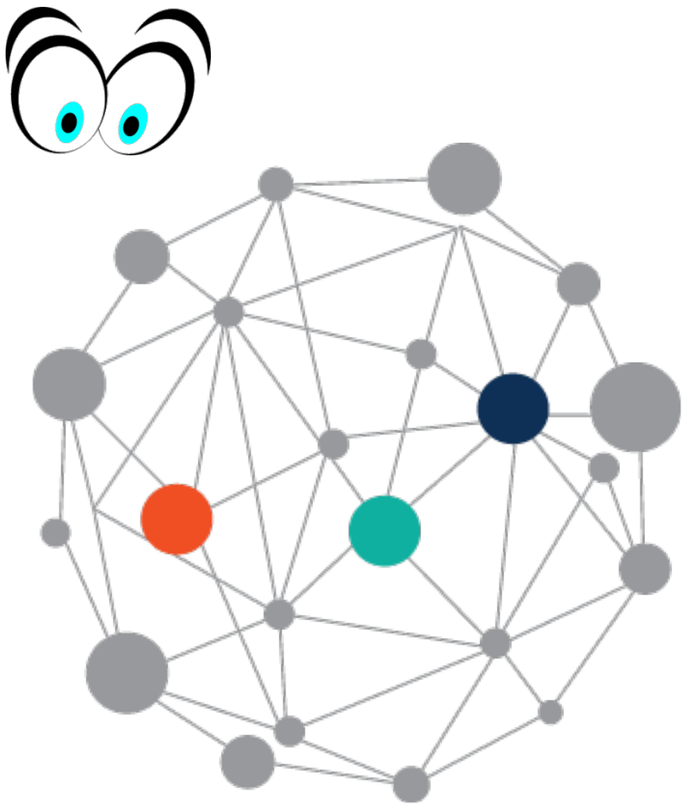
KG Maintenance

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



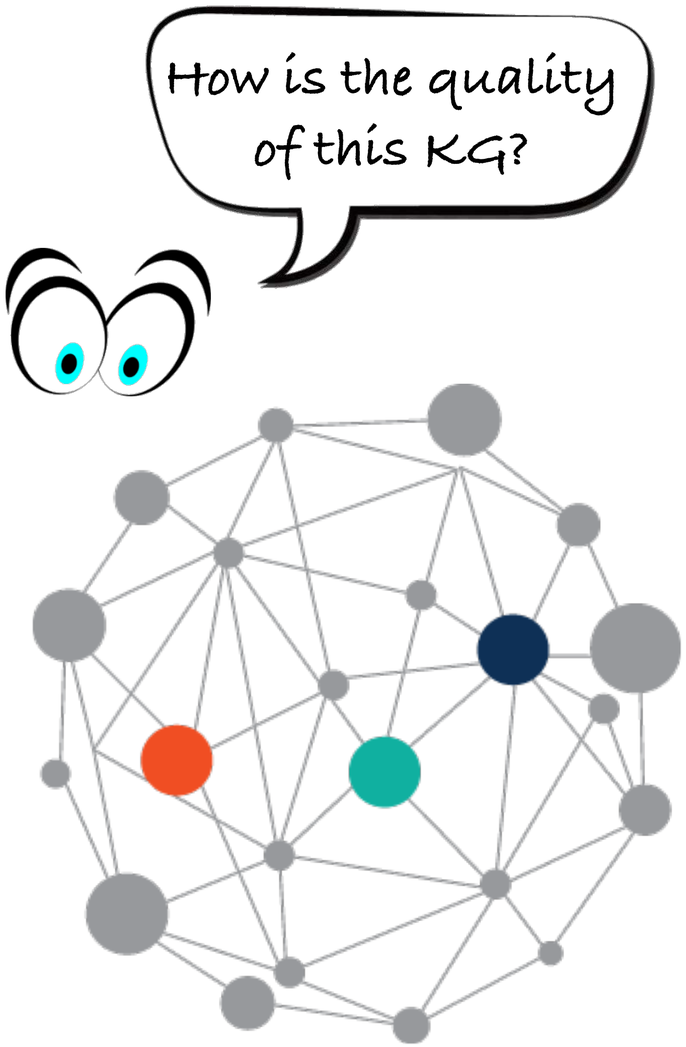
KG Maintenance

- 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])

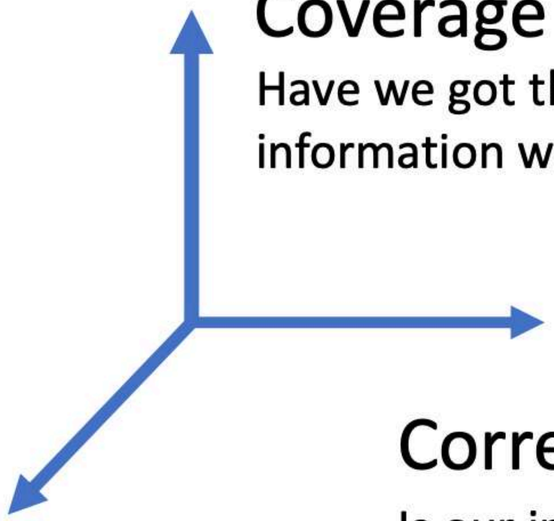


KG Assessment

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

……

Several core assessment dimensions for KG



Coverage

Have we got the information we need?

Freshness

Is information up to date?

Correctness

Is our information accurate?

Three forces in constant conflict:

Increased freshness and coverage	⇨	Harder to ensure correctness
Increased correctness	⇨	Harder to ensure freshness and coverage

Correctness is *always* hard – what is true and correct?
Particularly critical in today's world

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

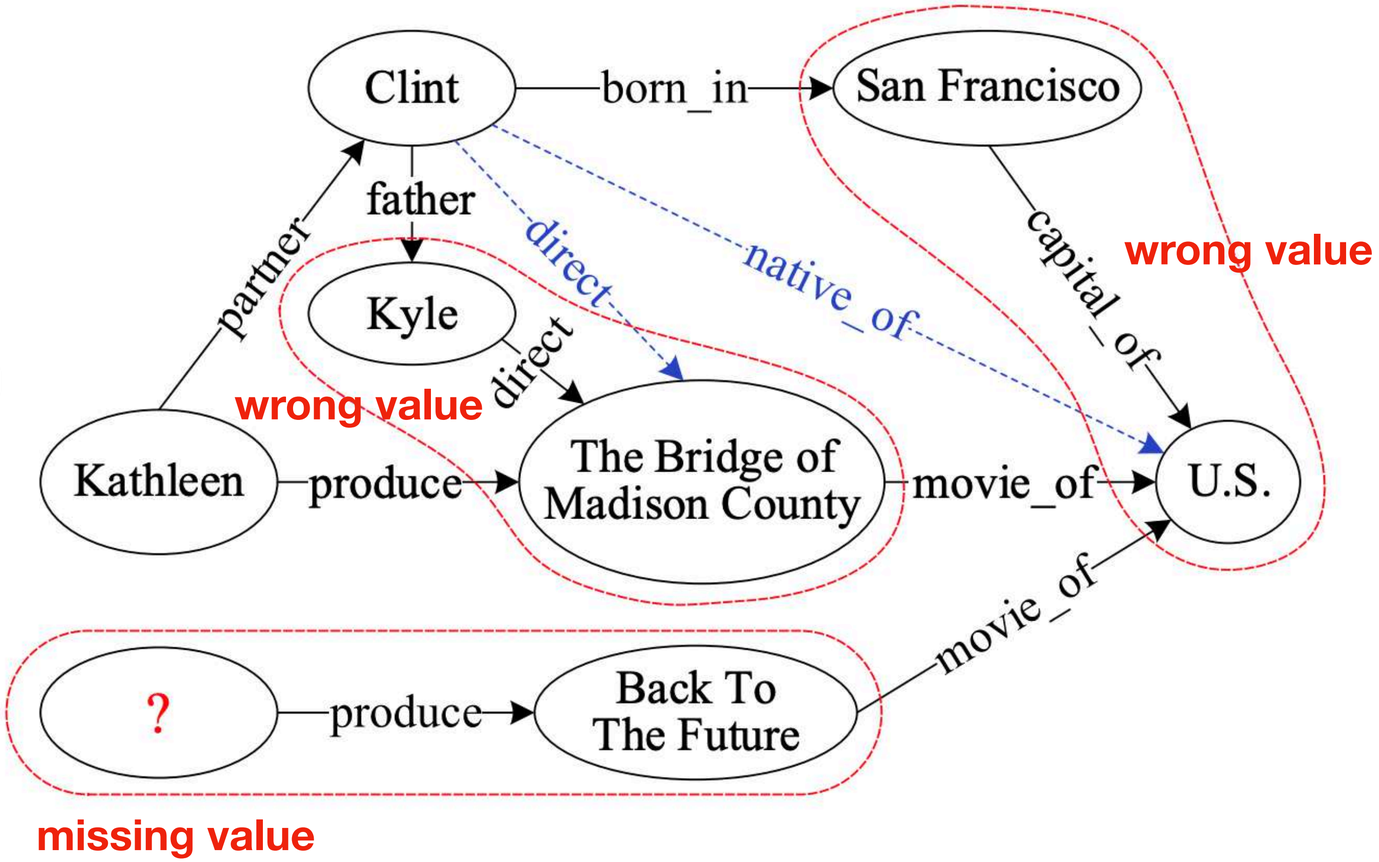
Knowledge Assessment Tools

- 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.

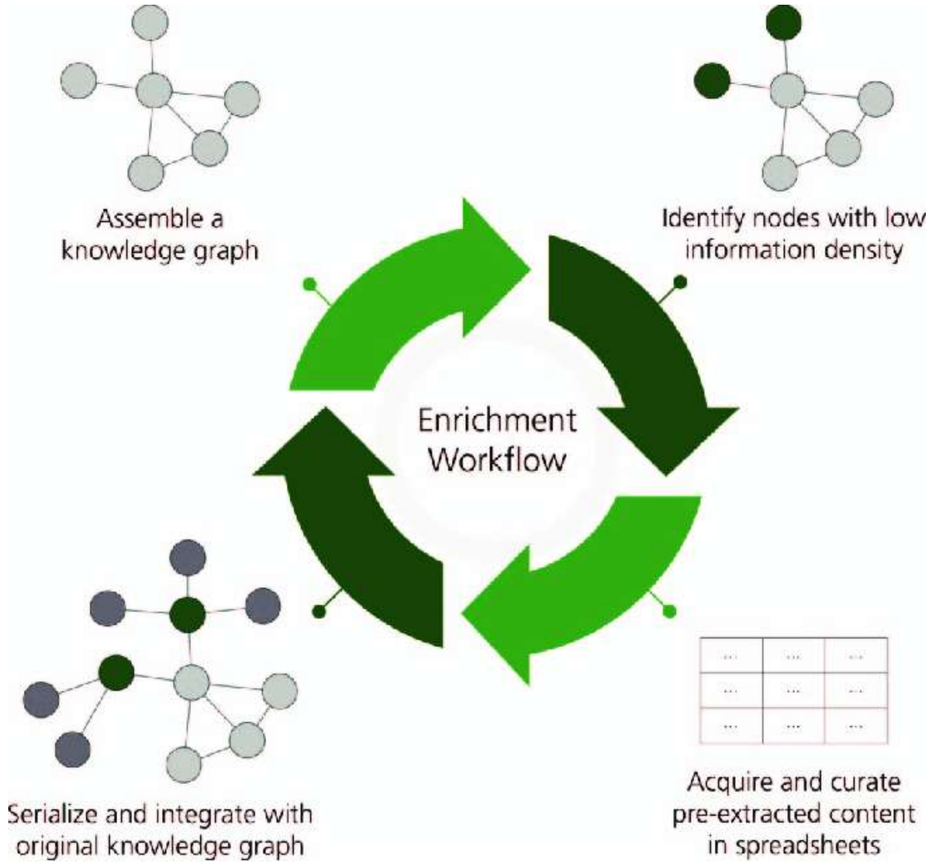
- KATARA [Chu et al., 2015]
  - Learns the relationships between data columns and validate the learn patterns with the help of existing Knowledge Bases and crowd, in order to detect errors in the data. Afterwards it also suggests possible repairs.

- 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.

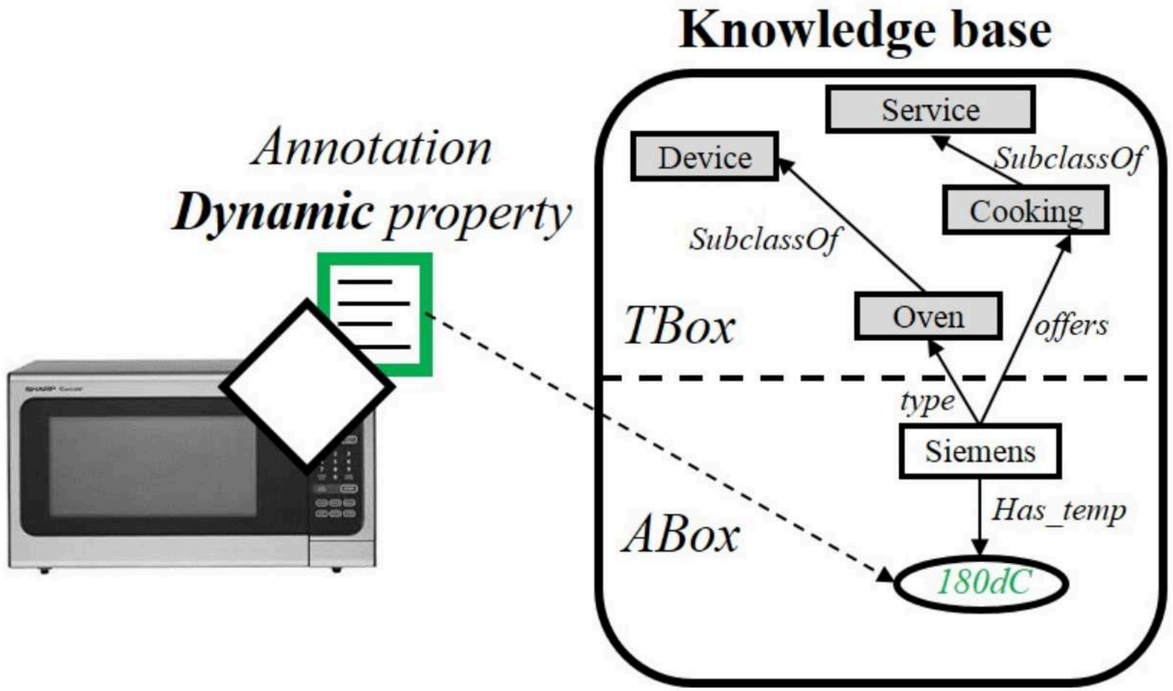
- TISCO [Rula et al., 2019]
  - A framework that tries to identify the time interval where a statement was correct. It uses external knowledge bases and the web content to extract evidence to assess the validity of a statement for a time interval.

Knowledge Enrichment

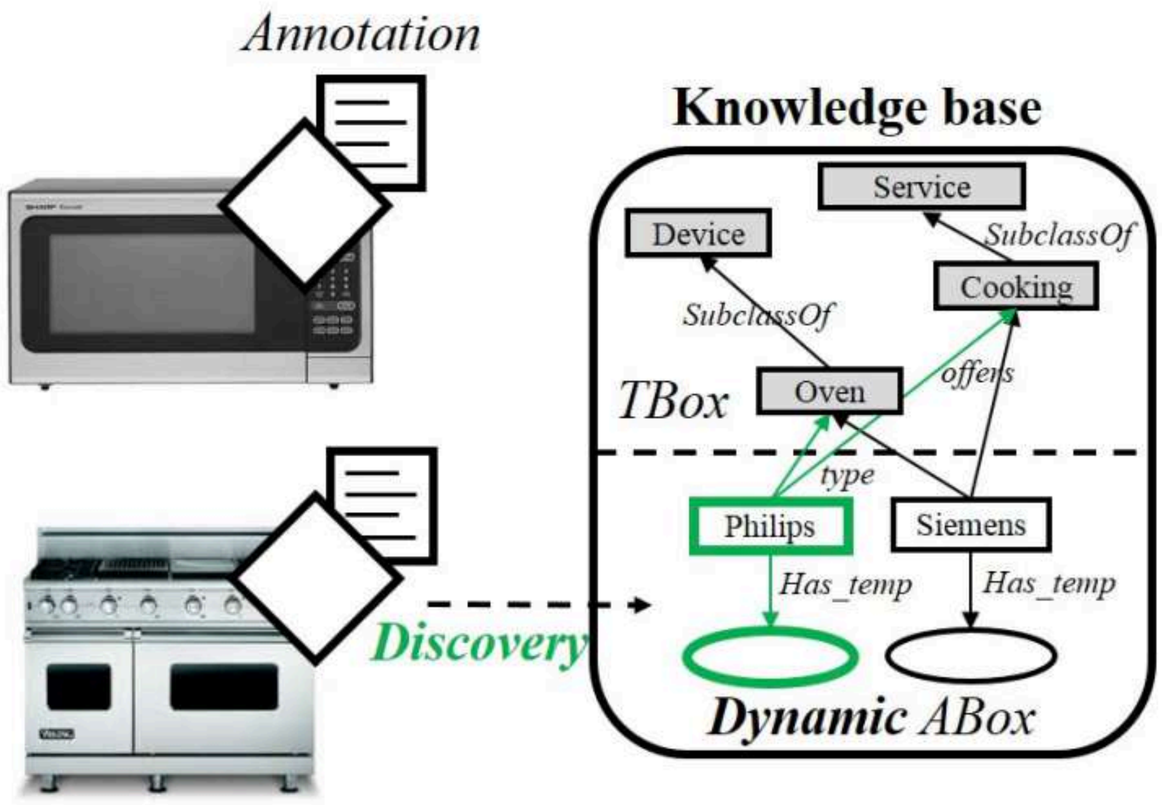
- Improve the **completeness** of a knowledge graph by adding new statements.



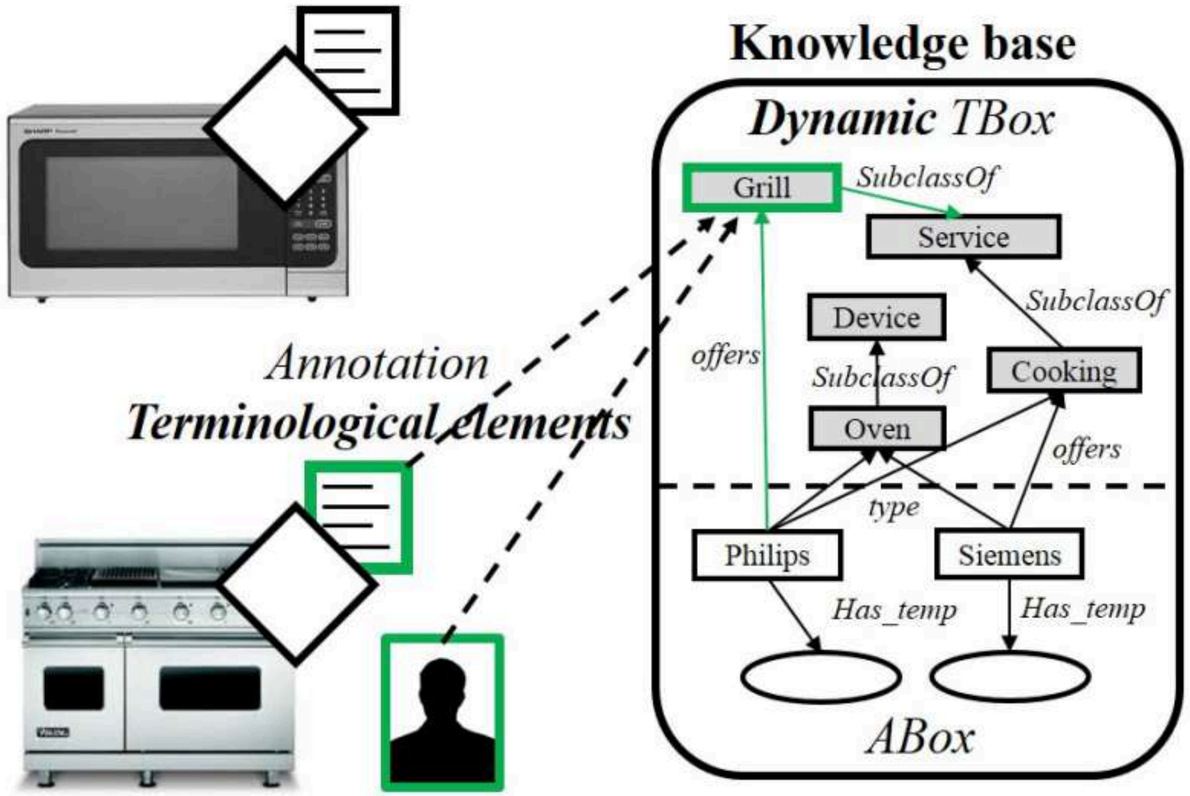
Knowledge enrichment: the property level



Knowledge enrichment: the instance level (Box)



Knowledge enrichment: the terminological level (TBox)



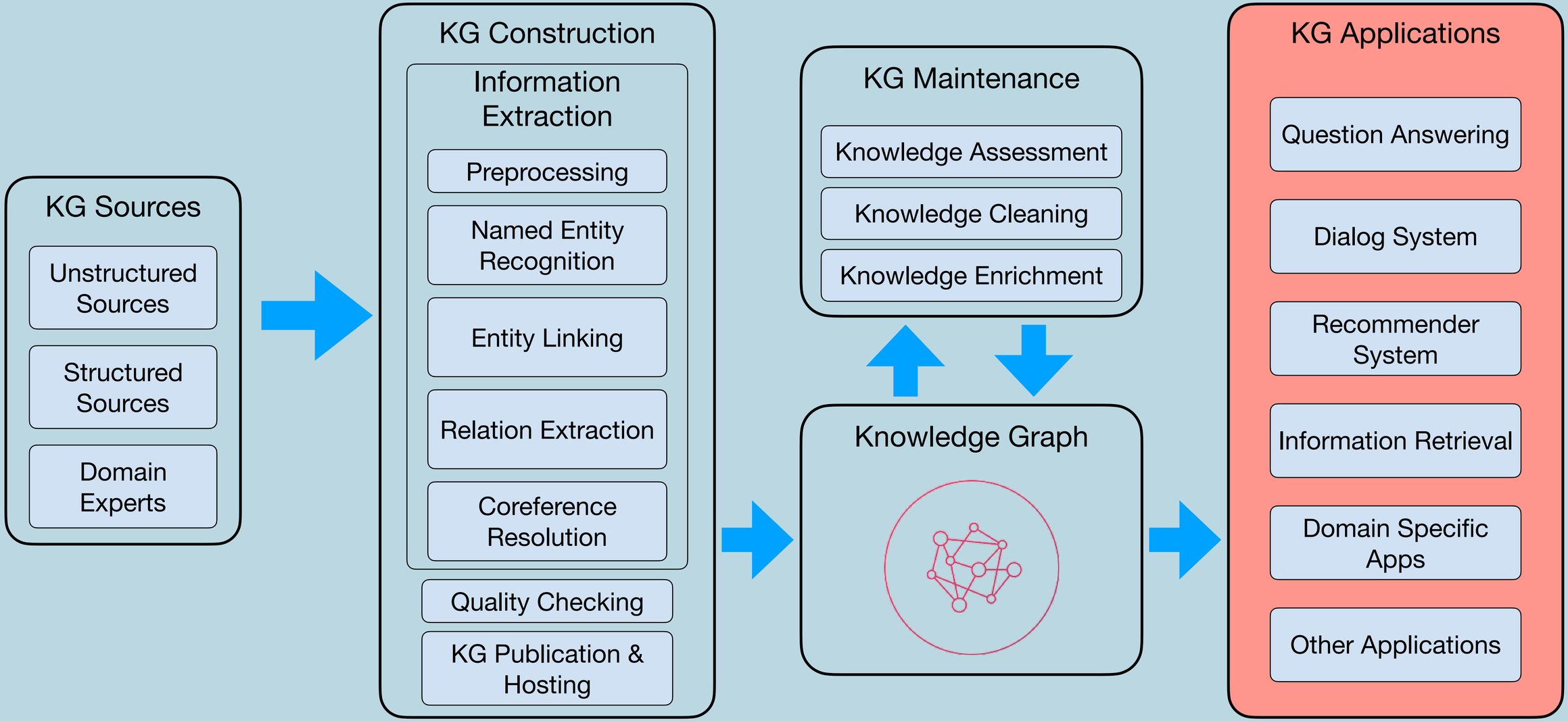
Knowledge Enrichment Tools

Duplication detection and resolution tools

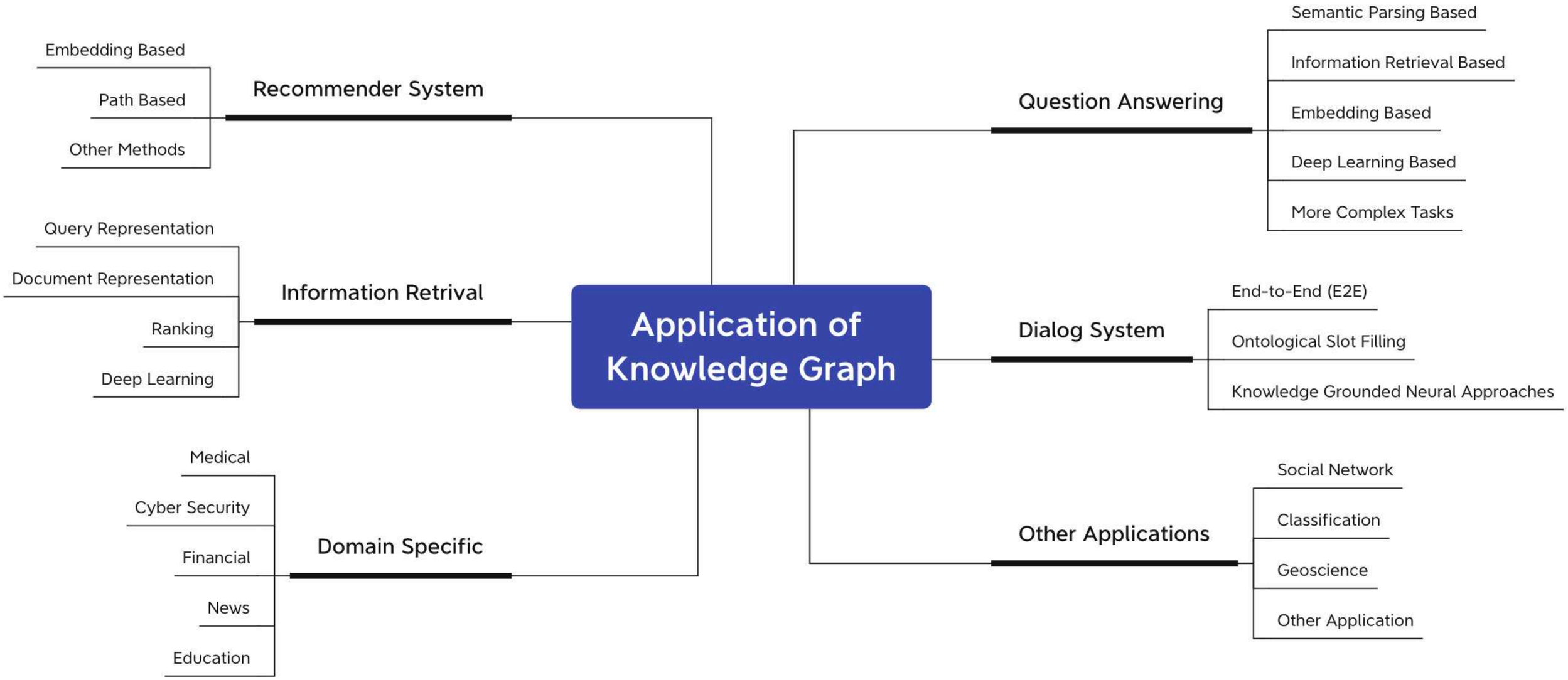
- Dedupe
  - A Python library that uses machine learning to find duplicates in a dataset and to link two datasets.
- LIMES [Ngomo & Auer, 2011]
  - A link discovery approach that benefits from the metric spaces (in particular triangle inequality) to reduce the amount of comparisons between source and 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.

KG Applications

KG Applications

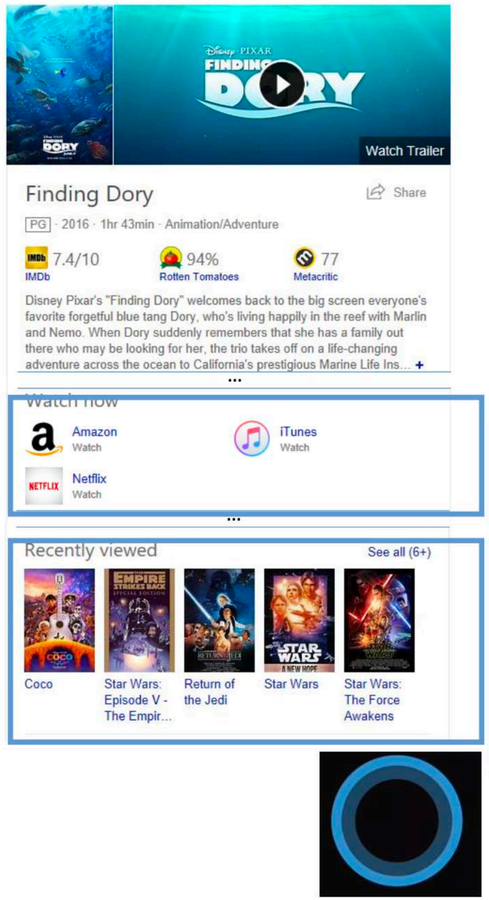


KG Applications



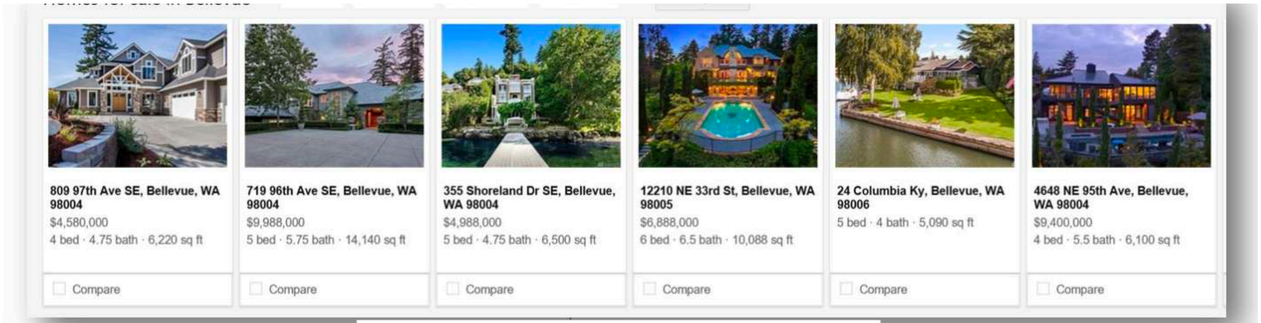
KG Applications: Bing & Cortana

- Knowledge graph has a tremendous impact on Bing impressions

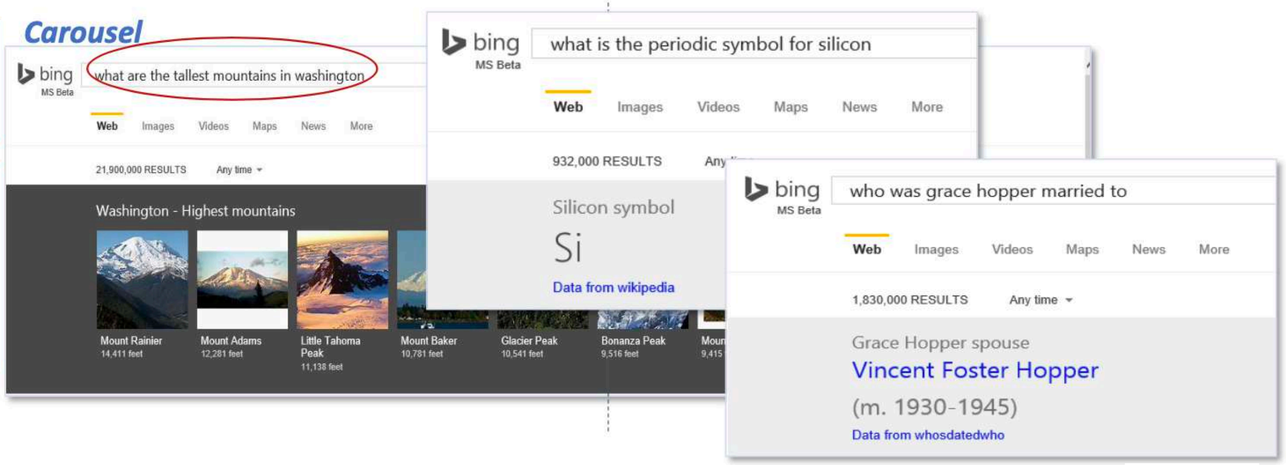


Watch now actions for movie entities

Recently viewed shows personal history



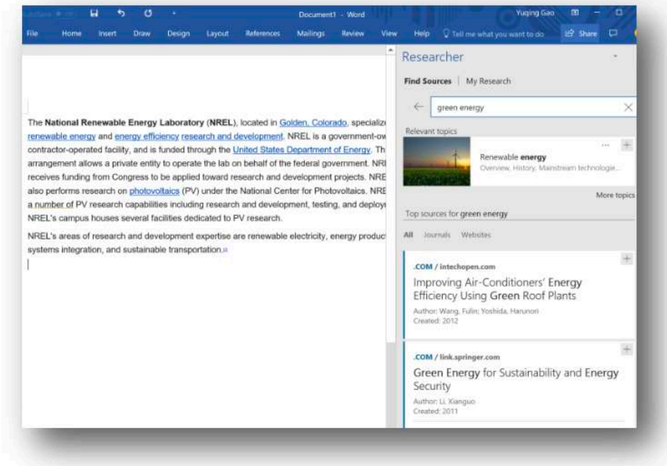
Carousel of information from Satori



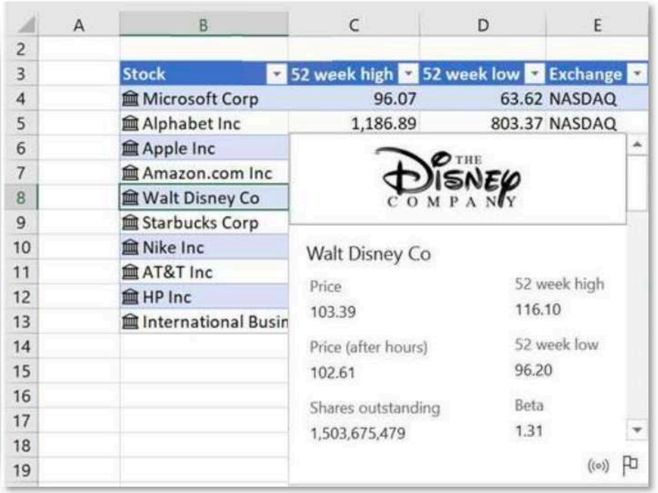
Natural language **Fact Answers**

KG Applications: Office

- Enriching the Office experience with knowledge graph.



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



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

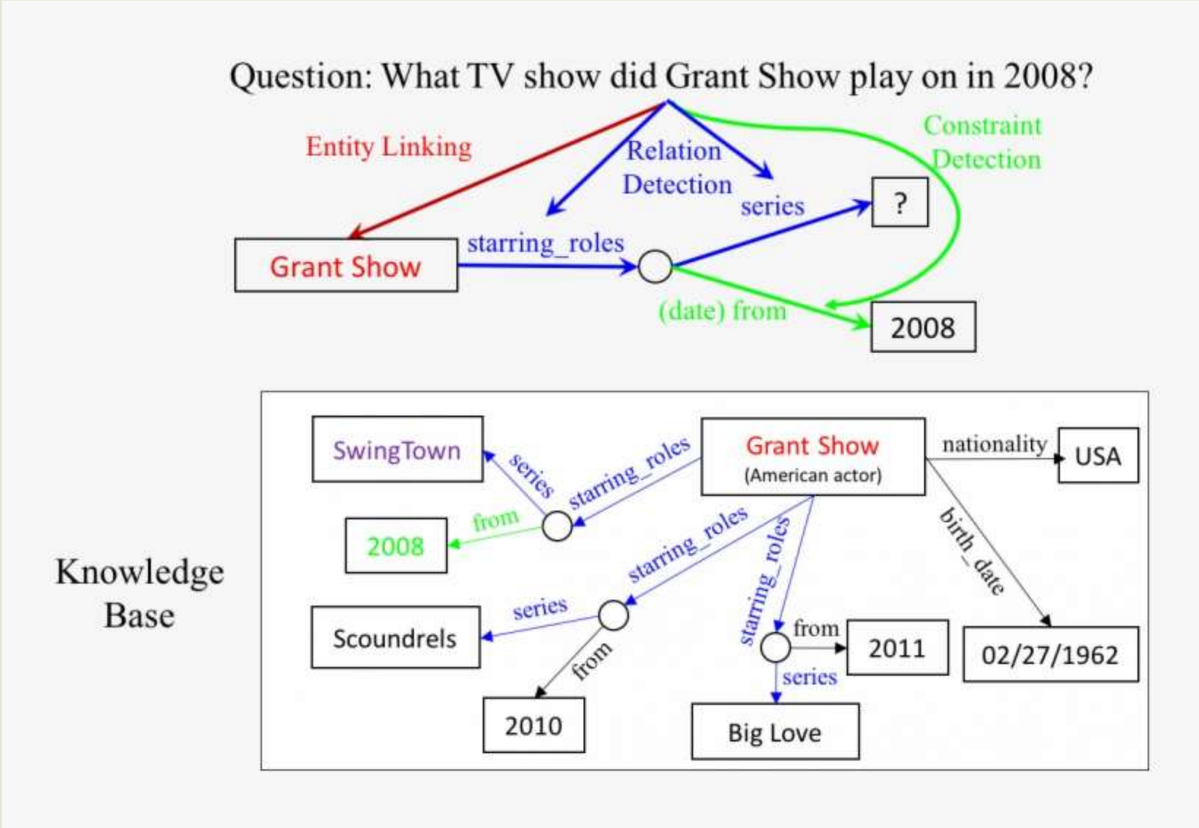
LinkedIn profile information visible in O365 People Card through Satori

Challenges

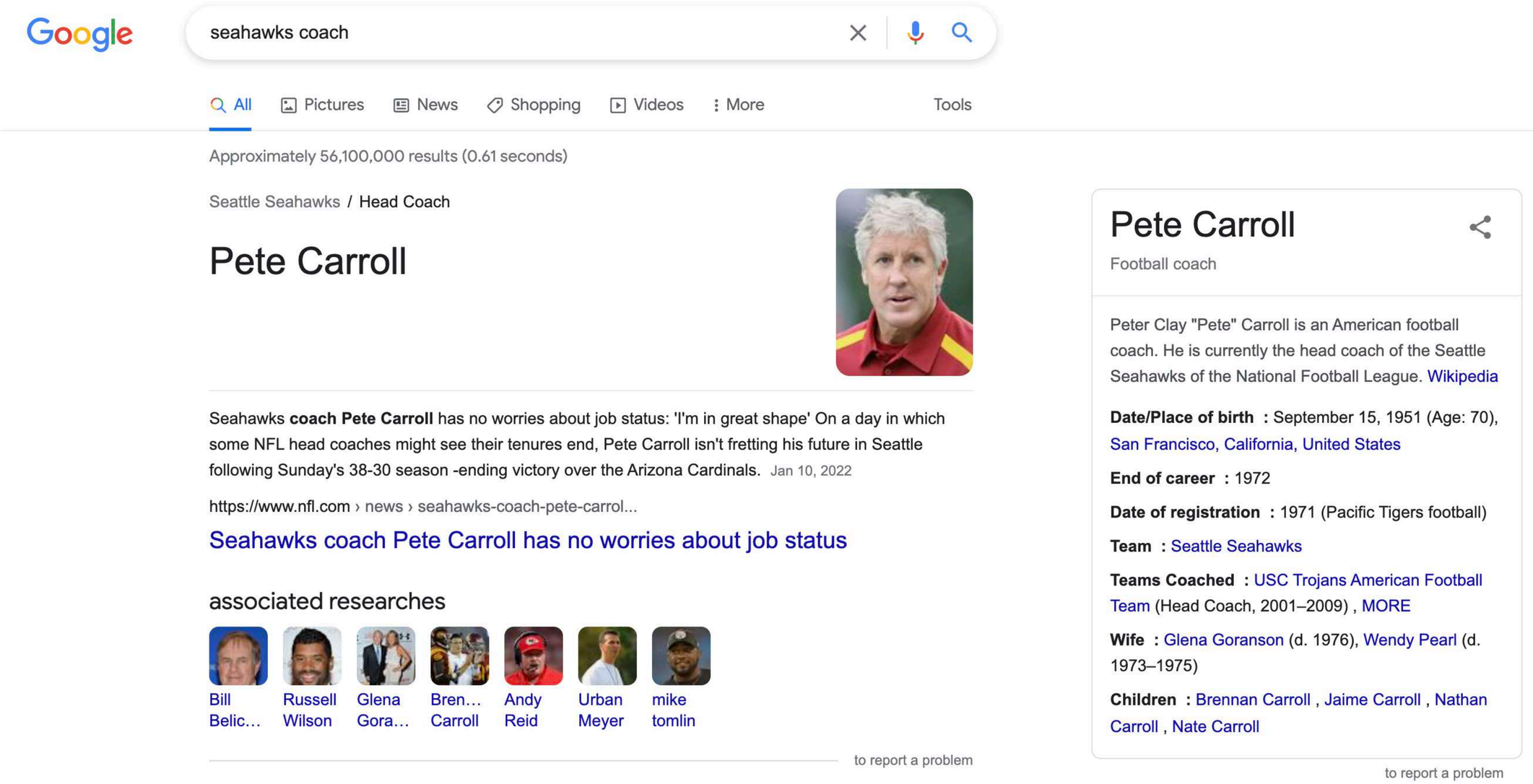
- Online conflation of people
- Compliance (training, code scanning & fixing, onboarding to new tools, etc.)
- Relevance with sparse profiles w/o access to the raw queries

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



KG-based Question Answering

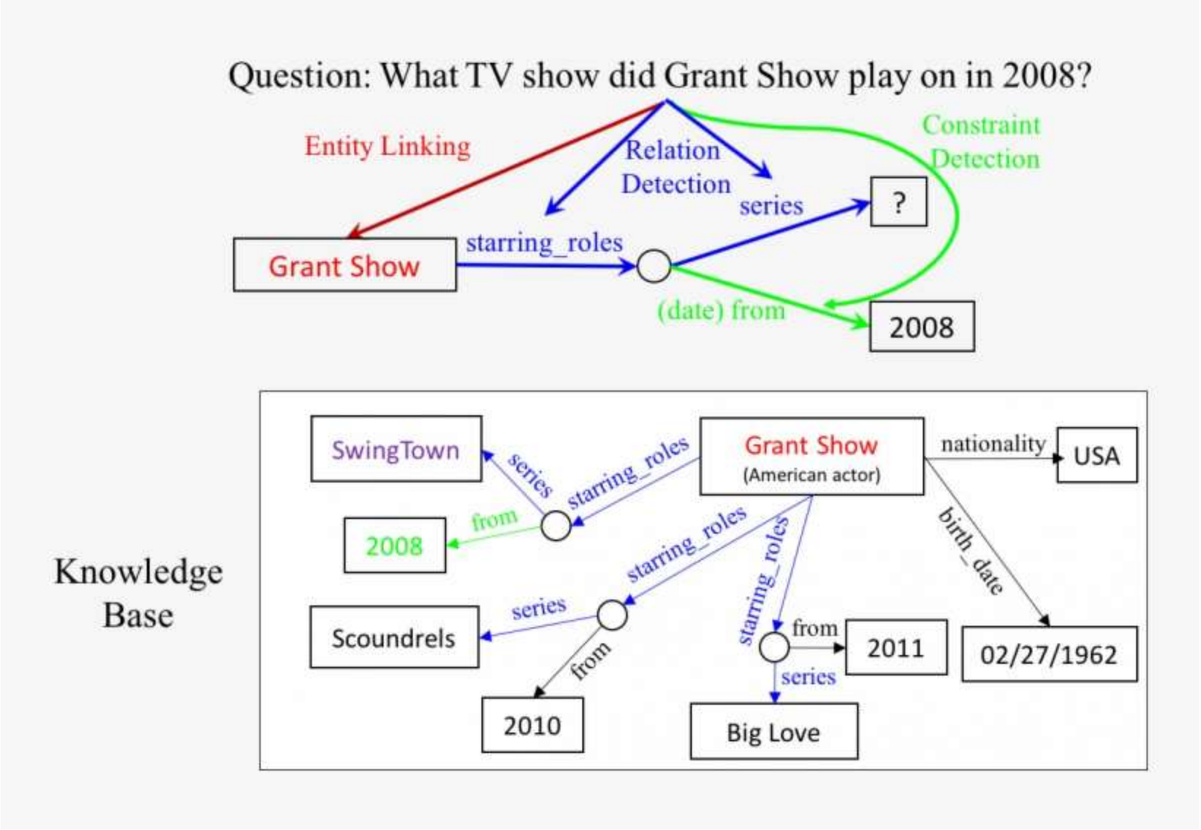
- **Challenges:**
  - Matching natural language relations to formal relations in KGs
  - There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, …etc
  - Scalable entity linking
  - Word 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

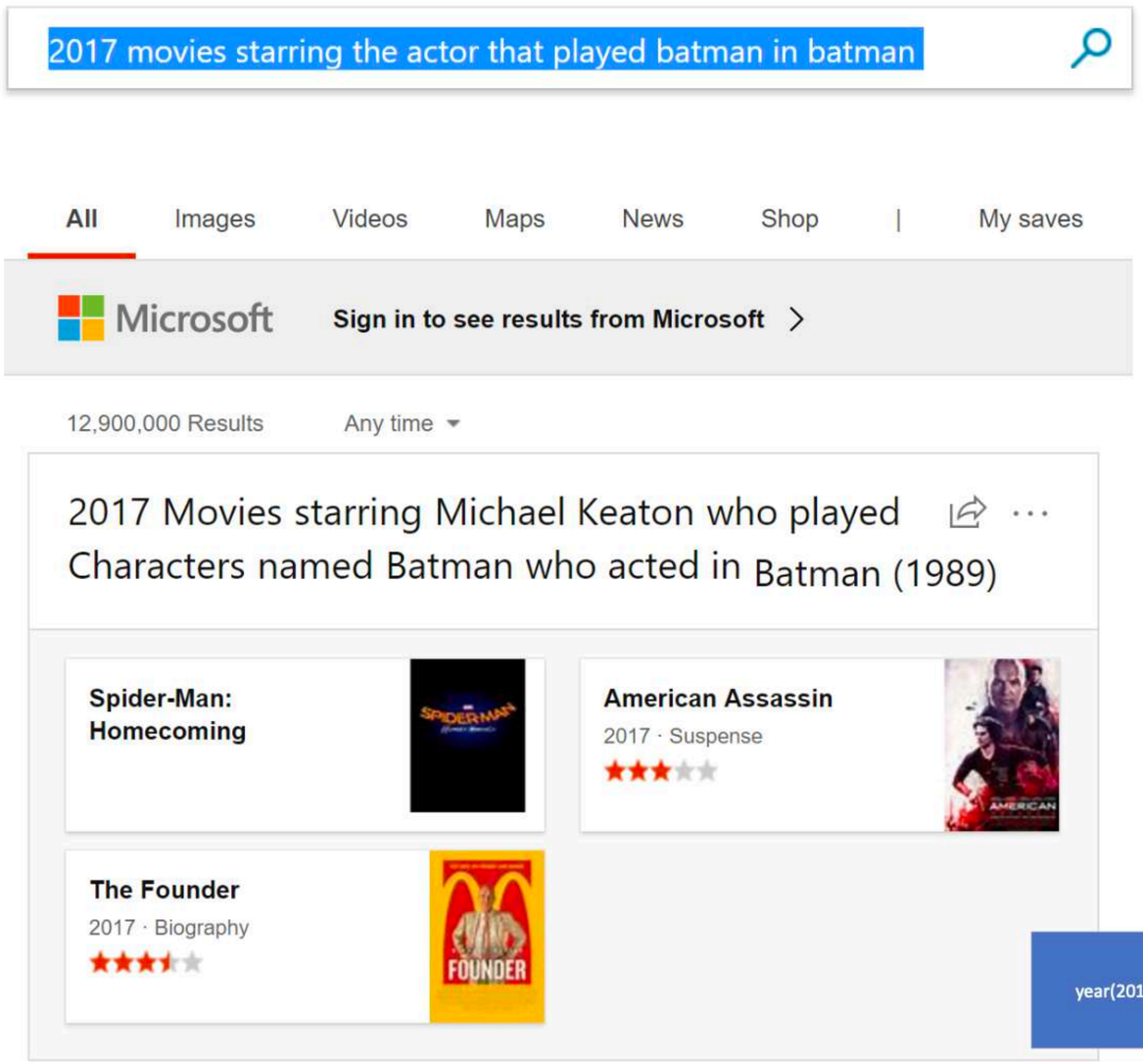
- Compositionality: multi-hop reasoning
  - {Movies starring the first wife of tom hanks}, {movies directed by the director of titanic}

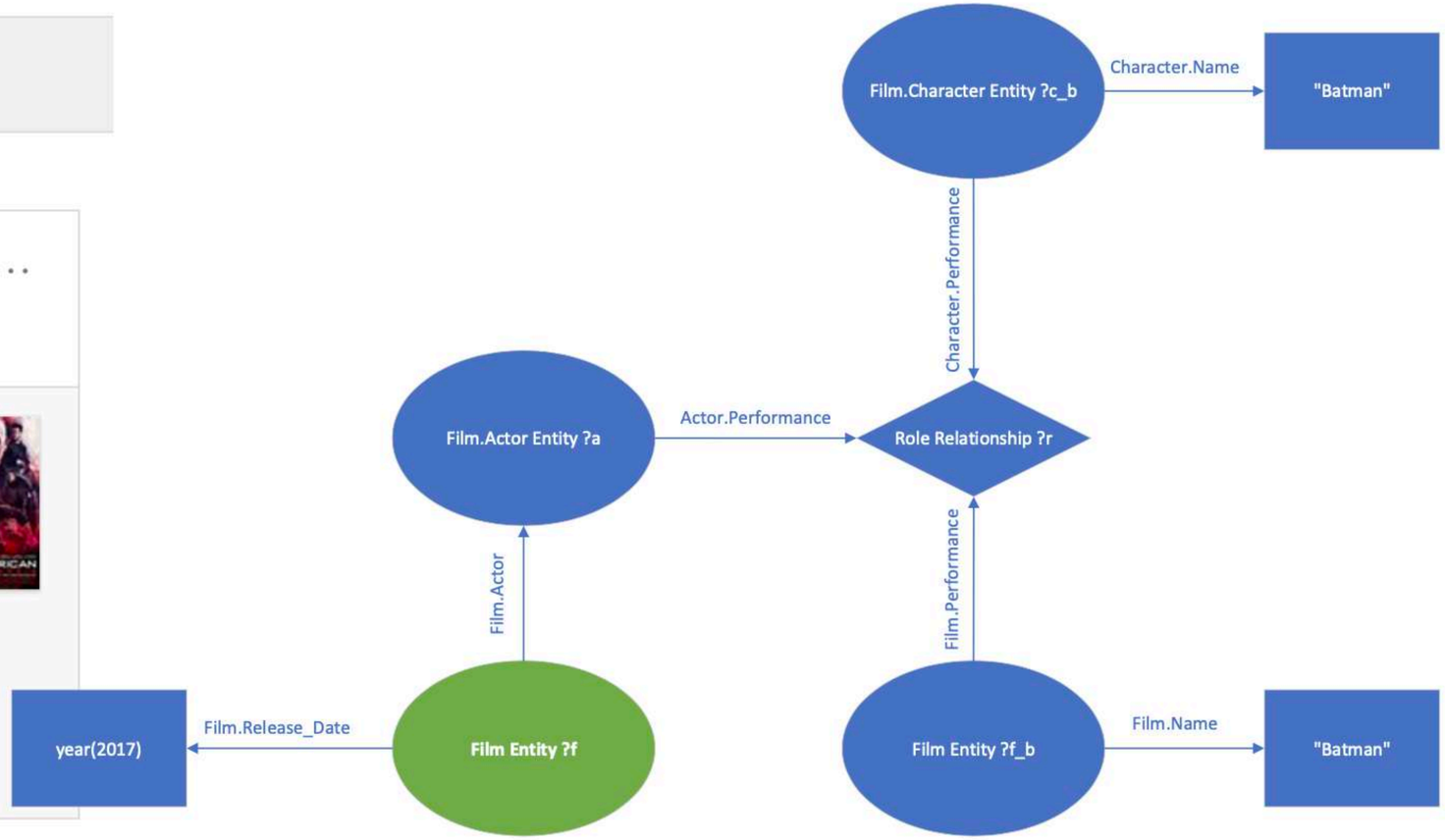
Approaches for KG-based Question Answering

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



Semantic Parsing





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}**

  $\lambda x.\ daughter\_of\ Donald\ Trump,\ x$

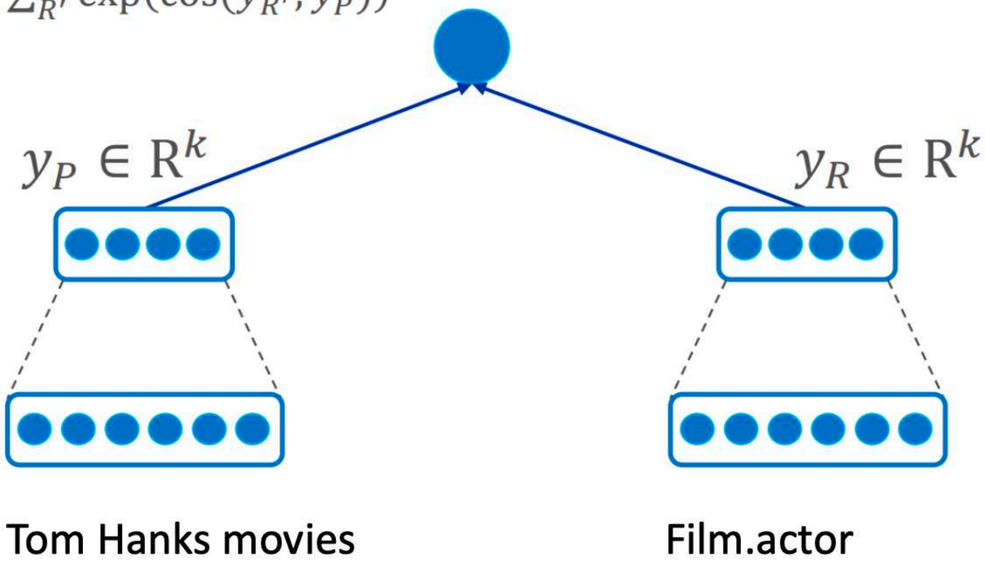
  $\lambda x.\ child\_of\ Donald\ Trump,\ x\ \wedge\ gender\ (x,\ female)$

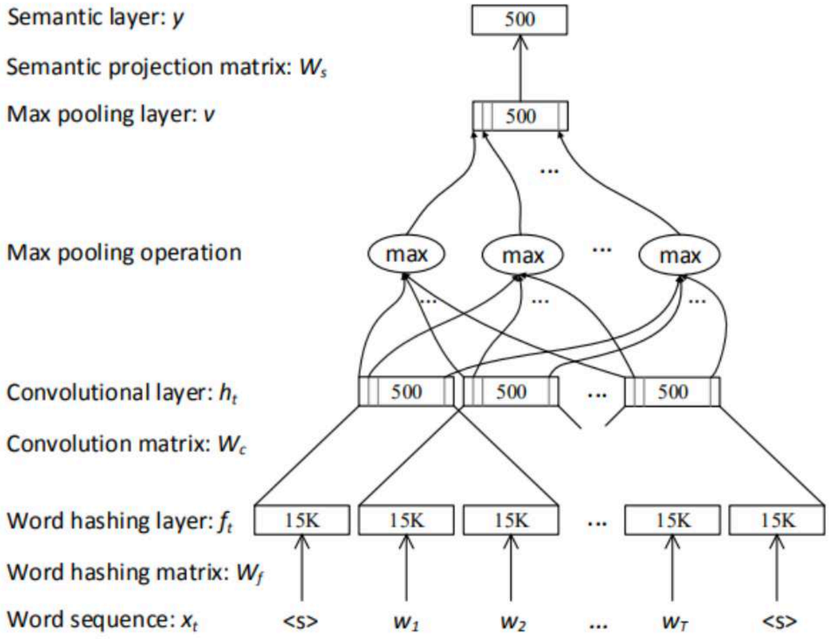
- This semantic expression can be then compiled into a knowledge graph database query e.g. Cypher/SPARQL and executed to return the results

Ontology Matching on Relationships using DSSM

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

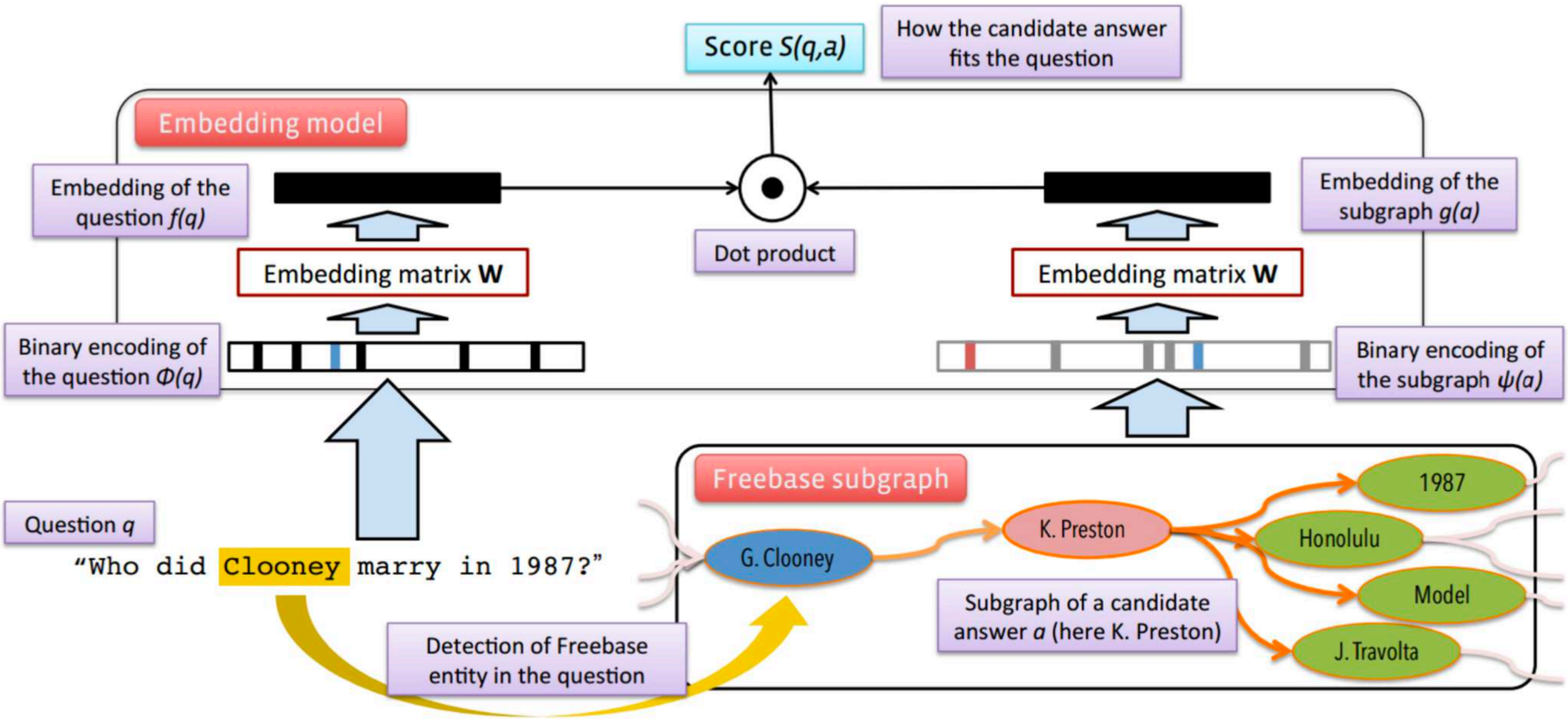
$$P(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$$





https://aclanthology.org/D14-1002.pdf

Knowledge Embedding Approach

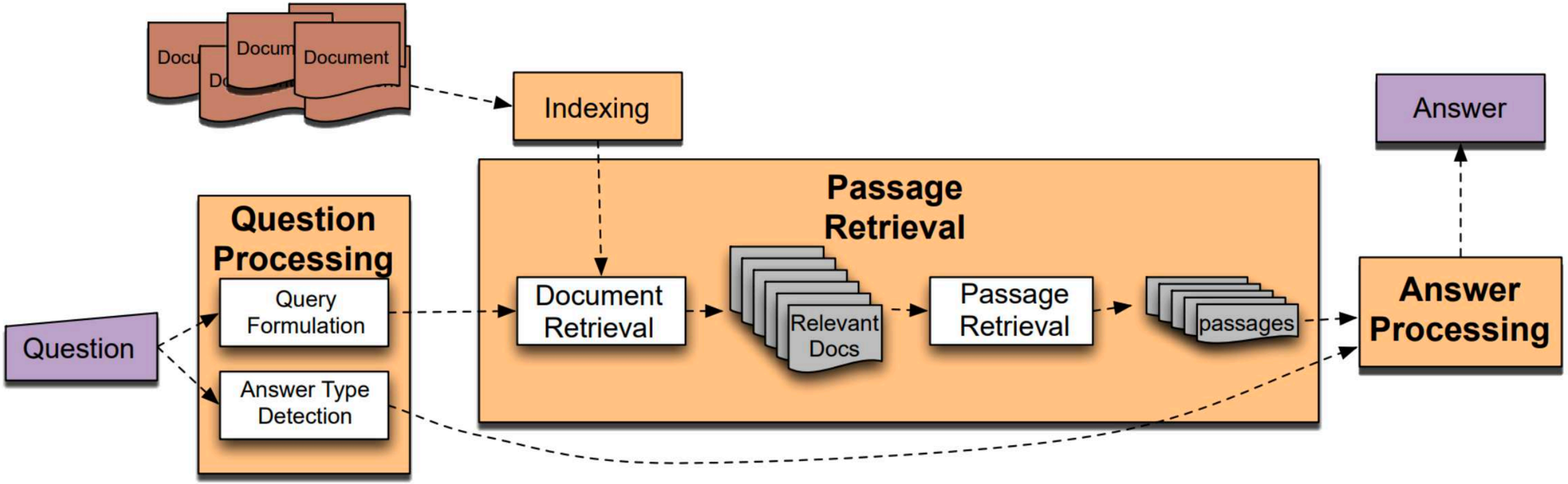


https://arxiv.org/pdf/1406.3676.pdf

Information Extraction Approaches

- Extracting and answers on the fly.

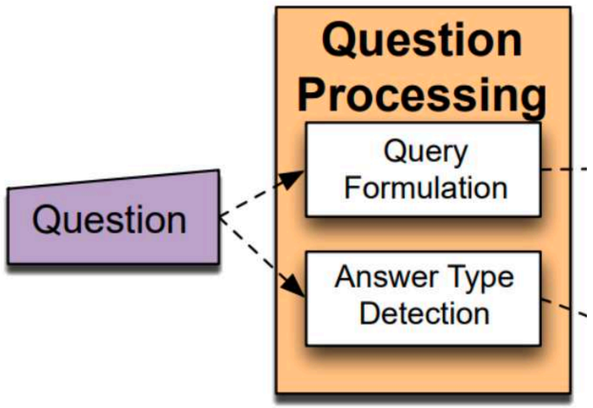
- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.

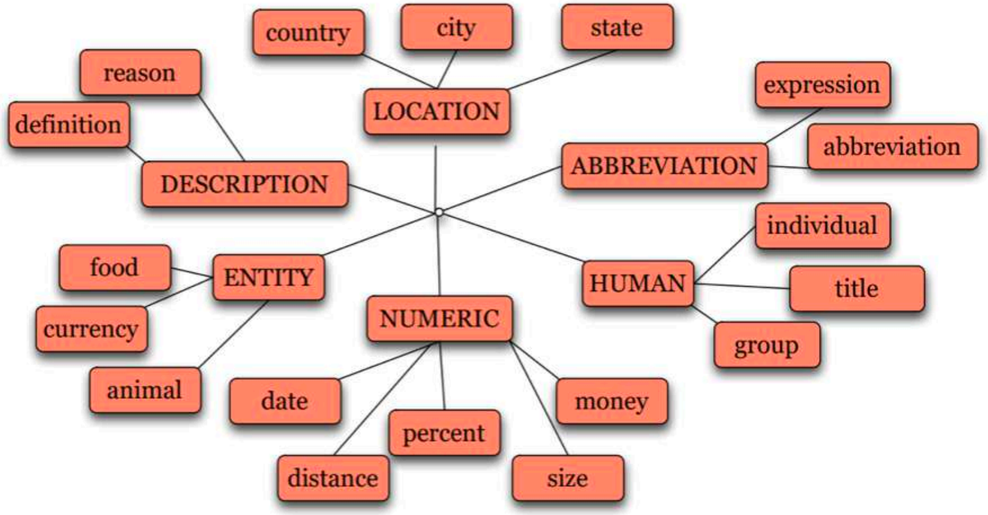


Question Answering [Dan Jurafsky, Stanford]
https://web.stanford.edu/~jurafsky/slp3/old_oct19/25.pdf

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





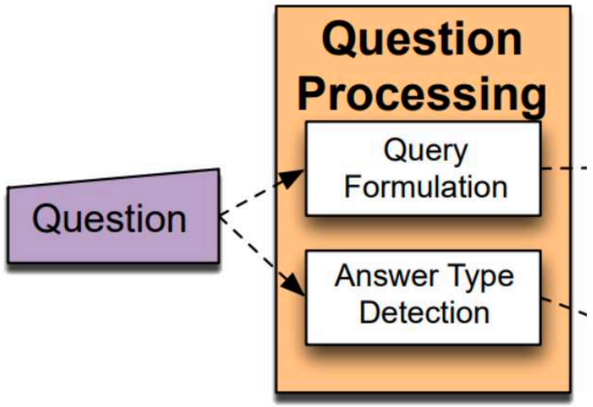
Learning Question Classifiers [Xin Li & Dan Roth, COLING 2002]

Question Answering [Dan Jurafsky, Stanford]

Answer Type Detection

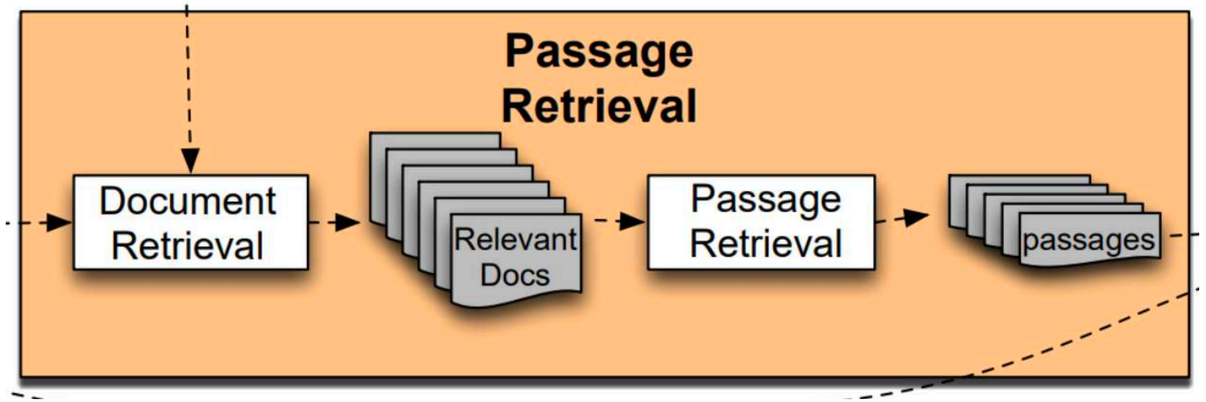
- **Rules**
  - Grammar for e.g. who be/... => Person
  - Head word for e.g. which city is the largest
  - …

- **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

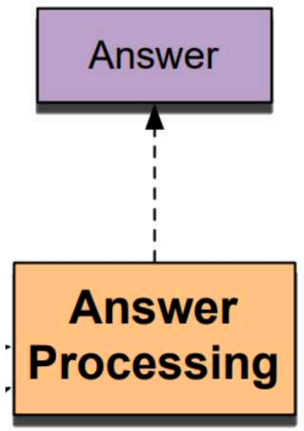


Process Answer

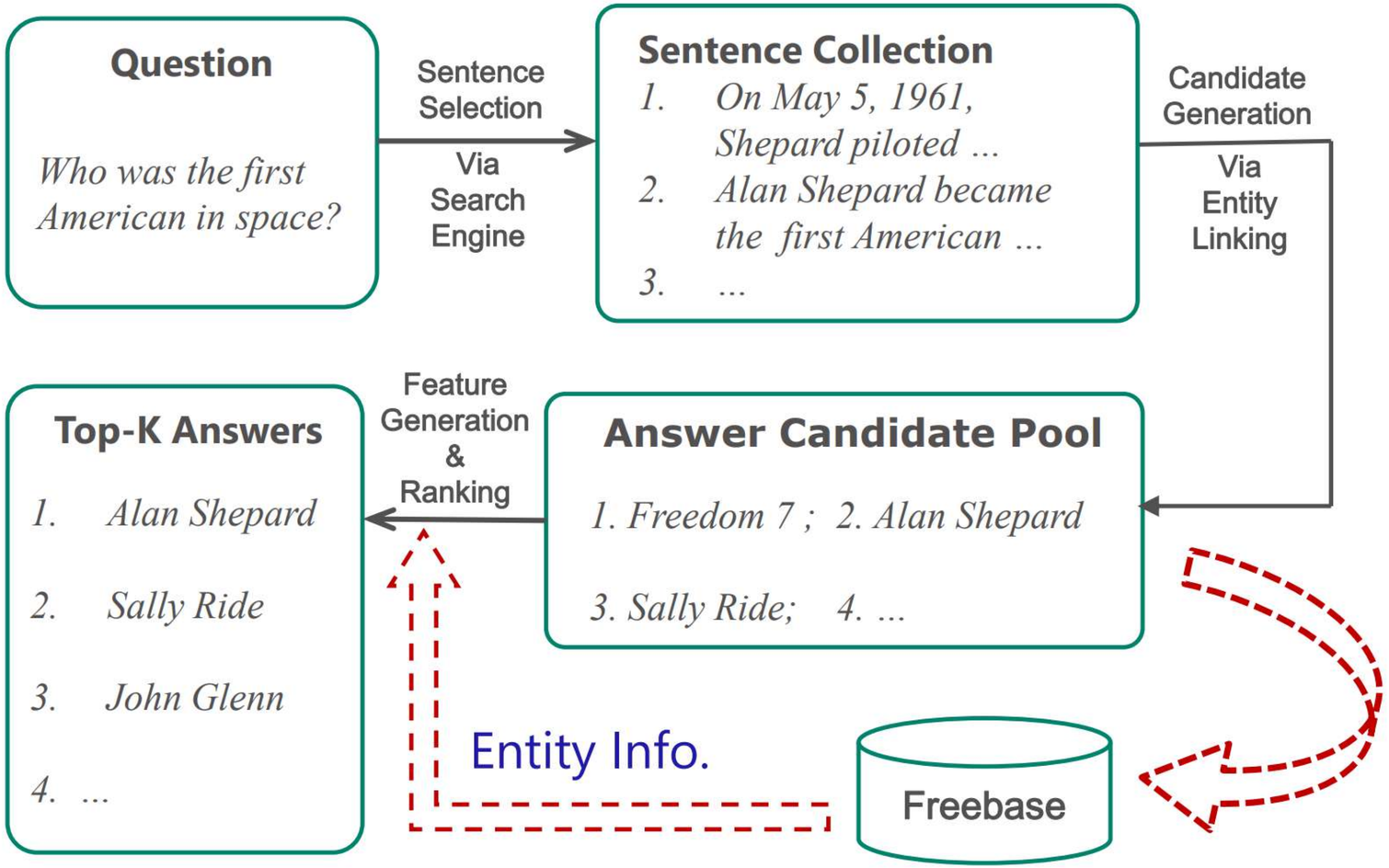
- Detect answer entity by running NER on the passage
- Mark the answer entity in the passage

- 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]



Answer Semantic Enrichment using KB



5-20% mean reciprocal rank (MRR) improvement

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

https://dl.acm.org/doi/10.1145/2736277.2741651

KG-based Dialogs

- 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

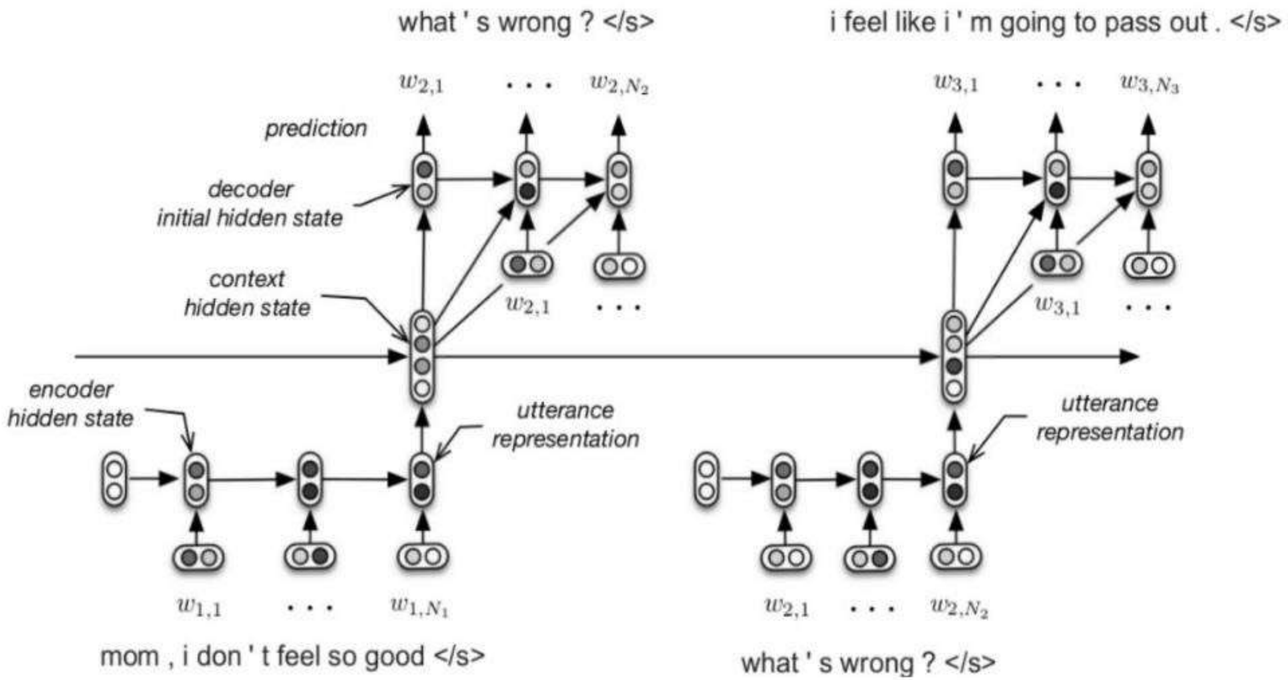
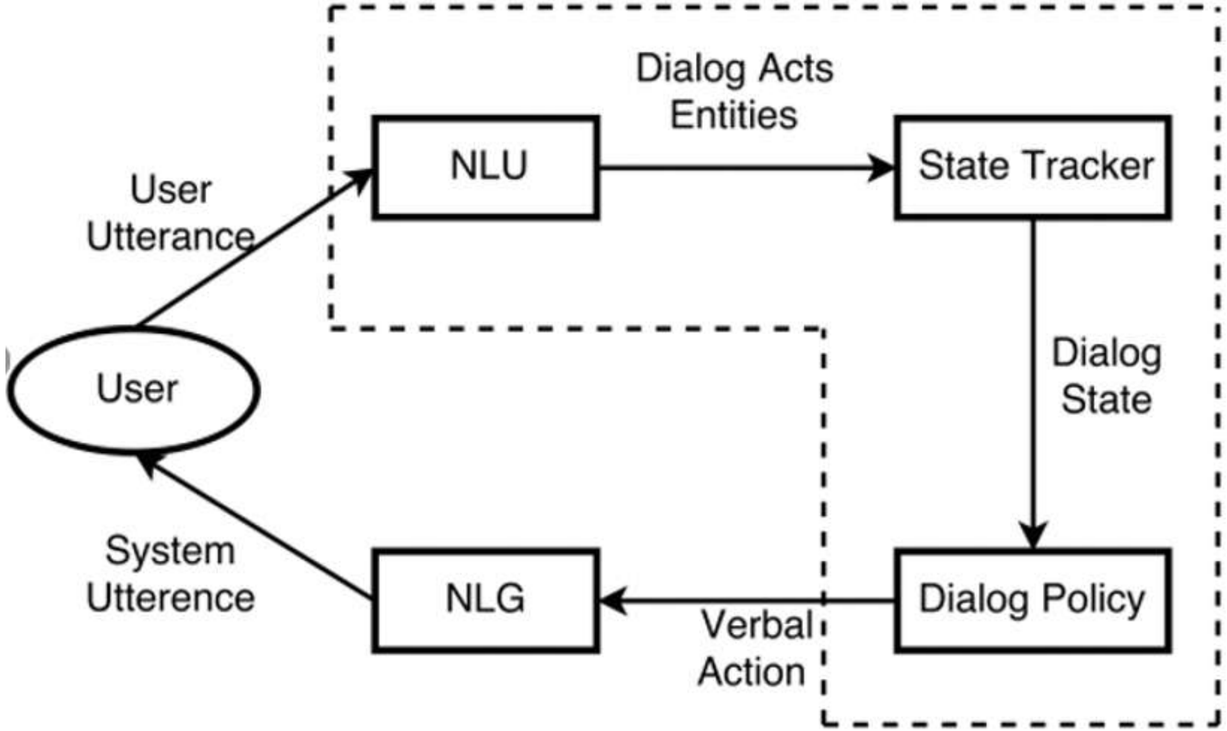


Figure 8: A computational graph representing the HRED architecture for dialogue over a span of three turns. The major addition to the architecture is a higher-level *context*-RNN keeping track of past utterances by progressively processing over time each utterance vector and conditioning the decoding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

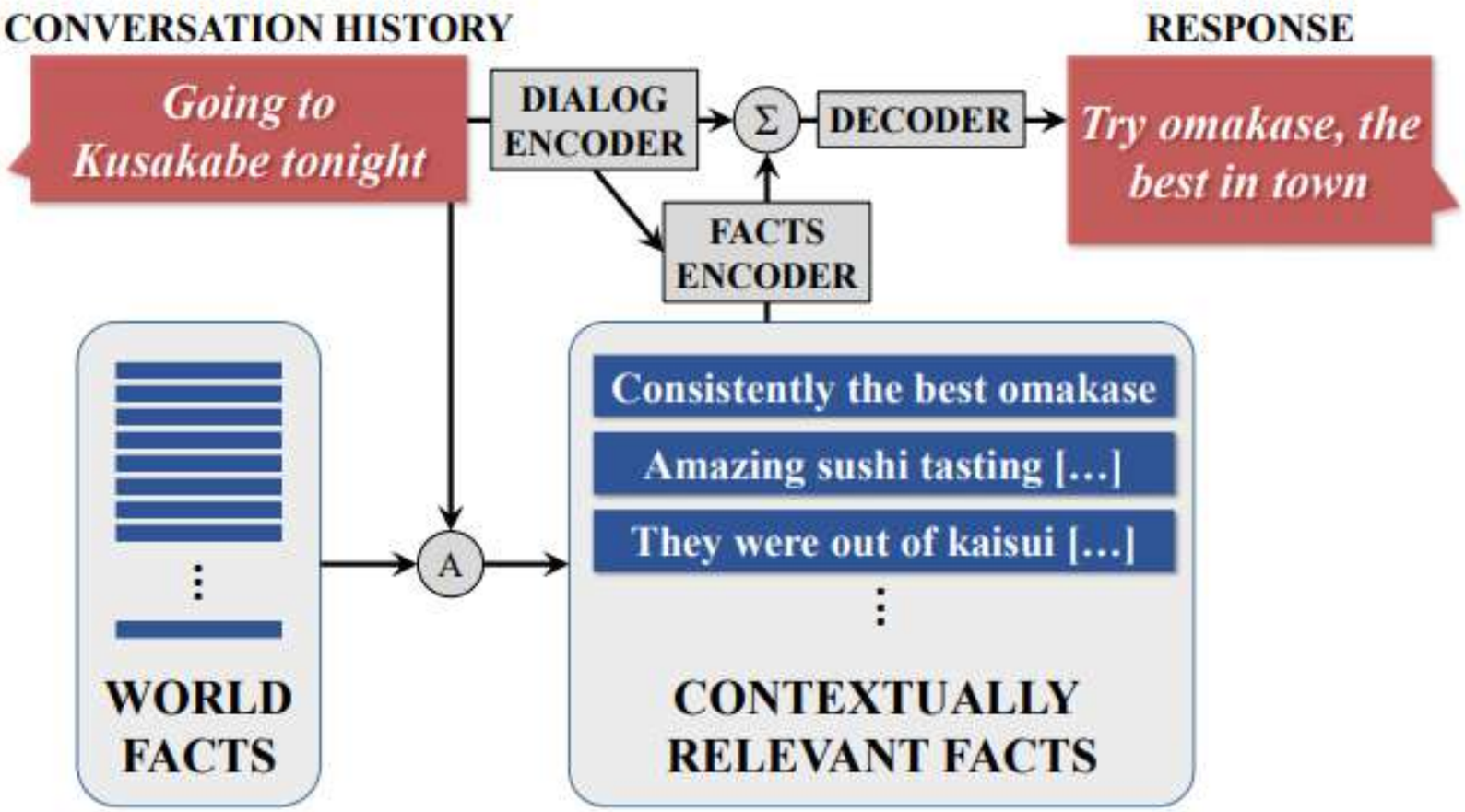
https://arxiv.org/pdf/1507.02221.pdf

Knowledge Based Ontological Slot Filling

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



Knowledge Grounded Neural Approaches



https://arxiv.org/pdf/1702.01932.pdf

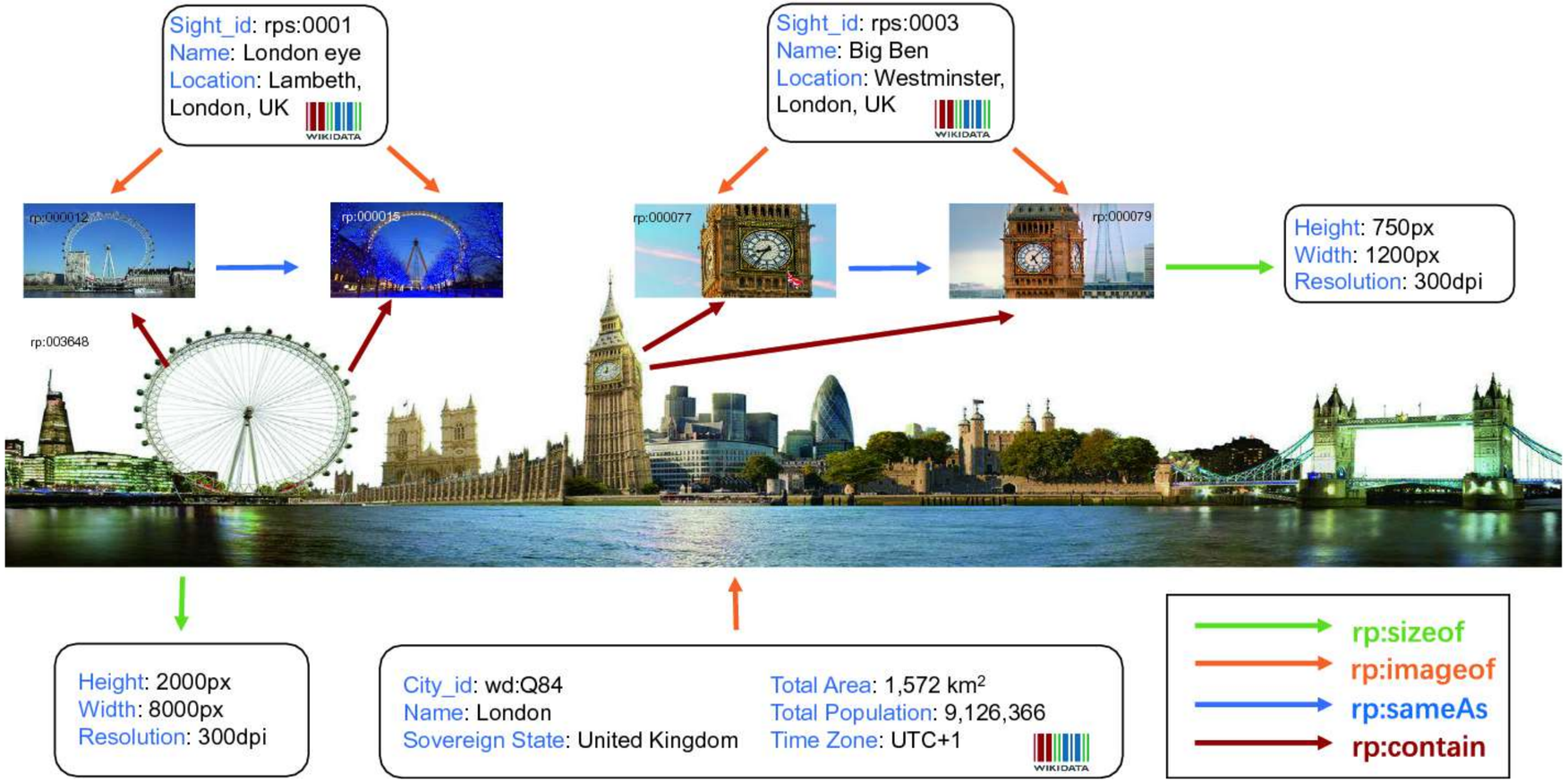
Enterprise Scenarios

- **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.

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.

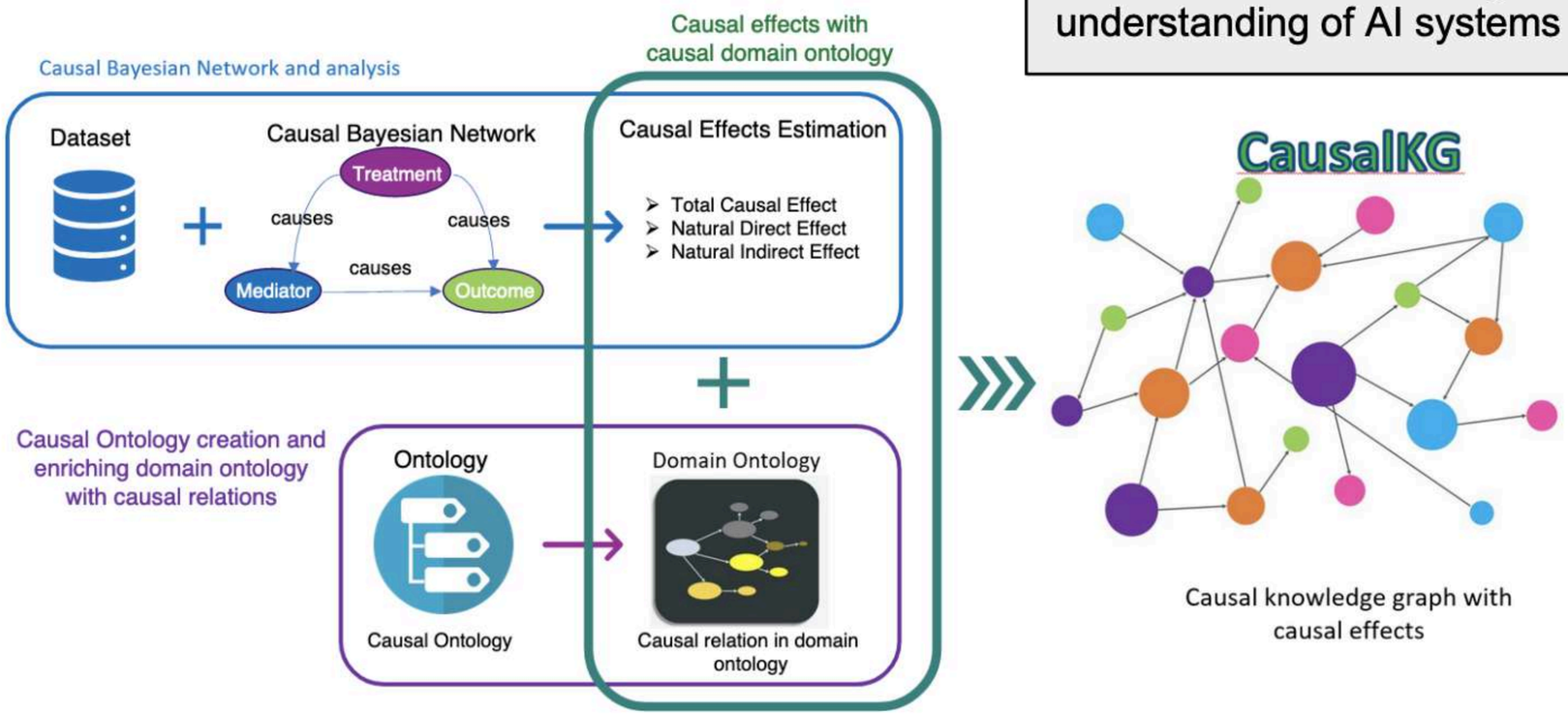
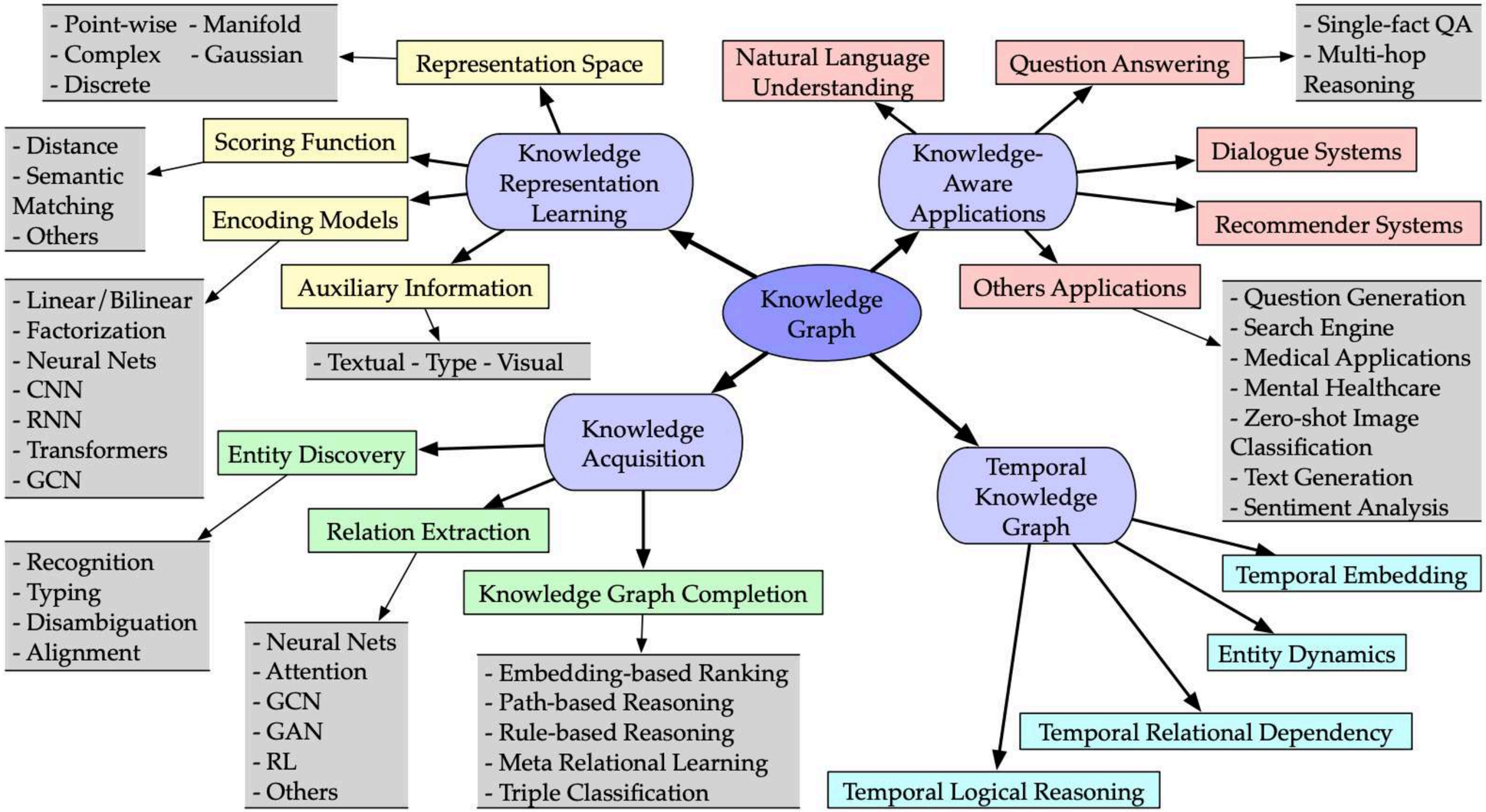


Figure 2. Causal Knowledge Graph Framework consists of three main steps, i) a Causal Bayesian Network and a domain-specific observational dataset, ii) Causal Ontology creation and enriching the domain ontology with causal relationships, and iii) Estimating the causal effects of the treatment, mediator,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

More research on knowledge graphs



https://arxiv.org/pdf/2002.00388v1.pdf

References

- 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. K nowledge G raphs.
- 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

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
Homepage: http://www-labs.iro.umontreal.ca/~liubang/
Github: https://github.com/BangLiu/

Université de Montréal

rali | recherche appliquée en linguistique informatique

Mila