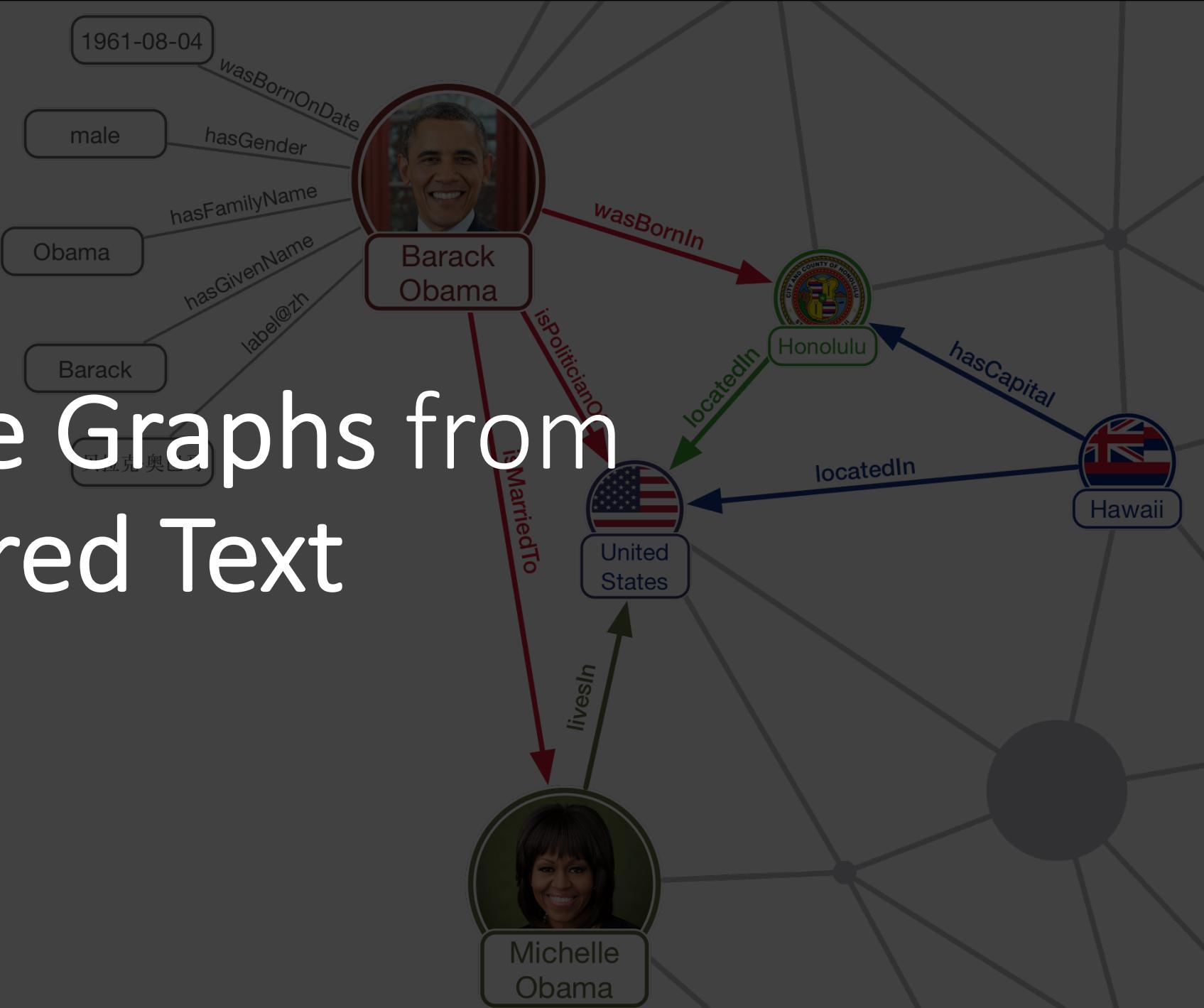


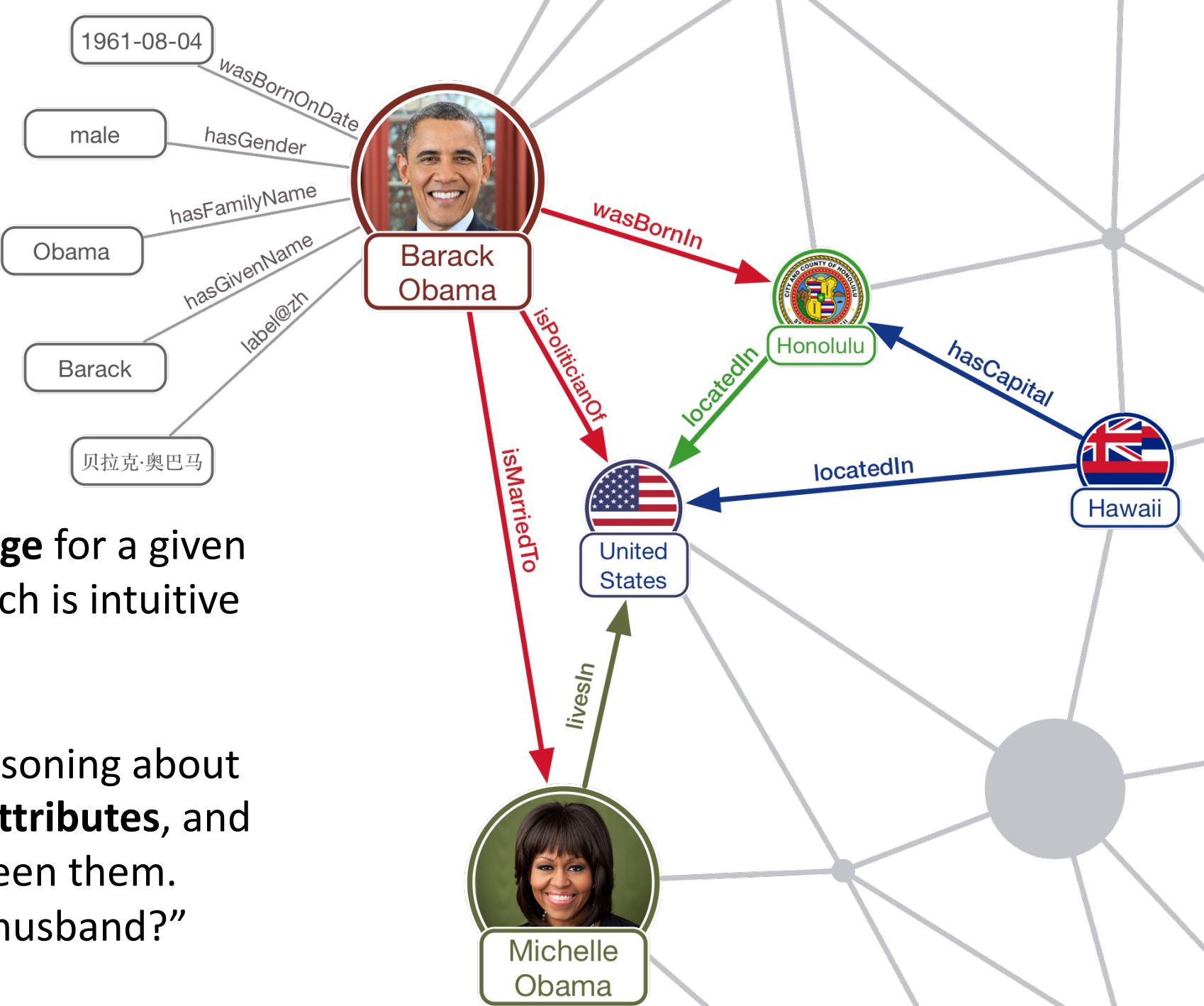
Building Knowledge Graphs from Unstructured Text

Chaitanya Joshi
chaitjo.github.io



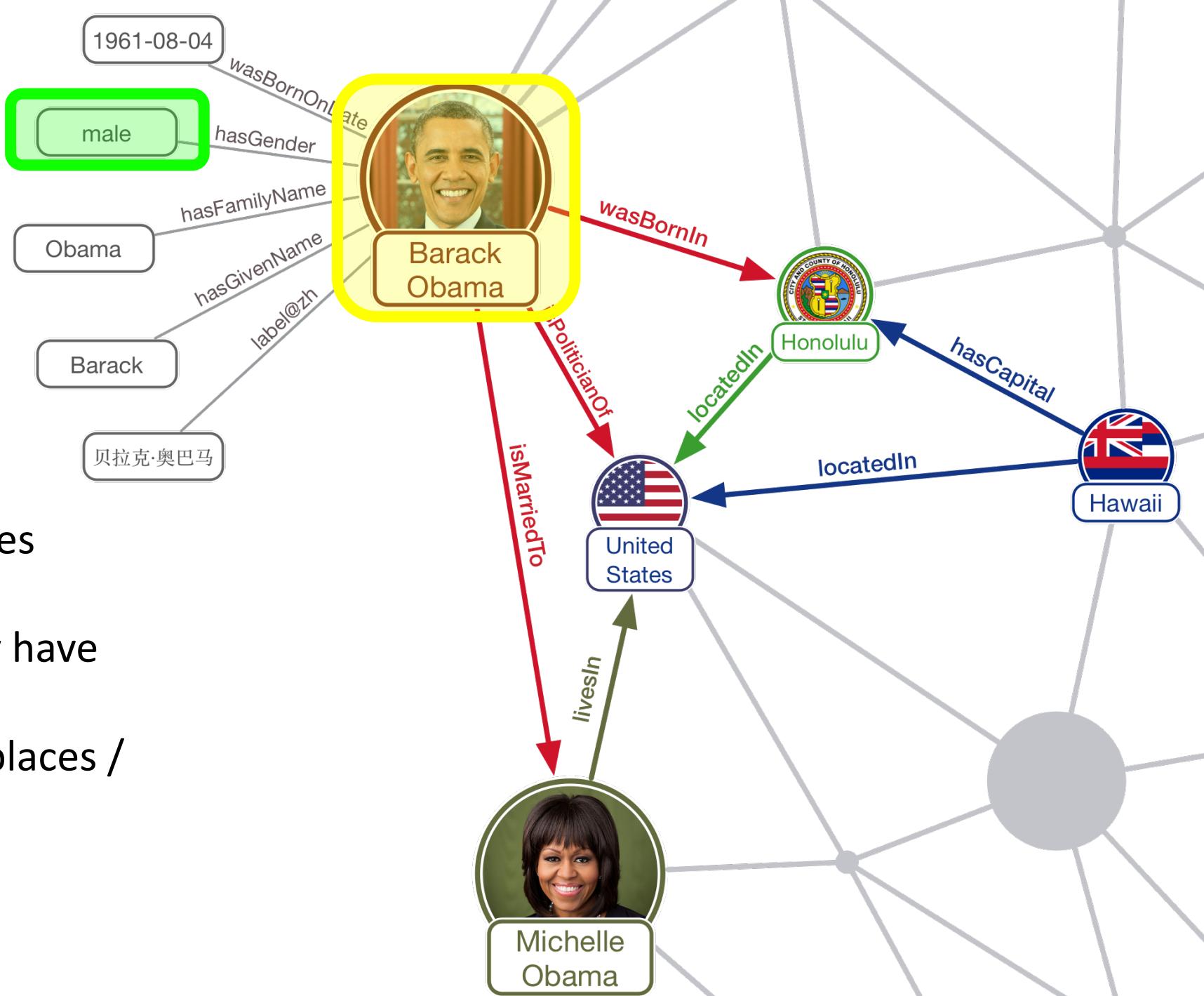
What are Knowledge Graphs?

- KGs represent **knowledge** for a given domain as a **graph**, which is intuitive to understand.
- KGs can be used for reasoning about various **entities**, their **attributes**, and the **relationships** between them.
 - “Who is Michelle’s husband?”



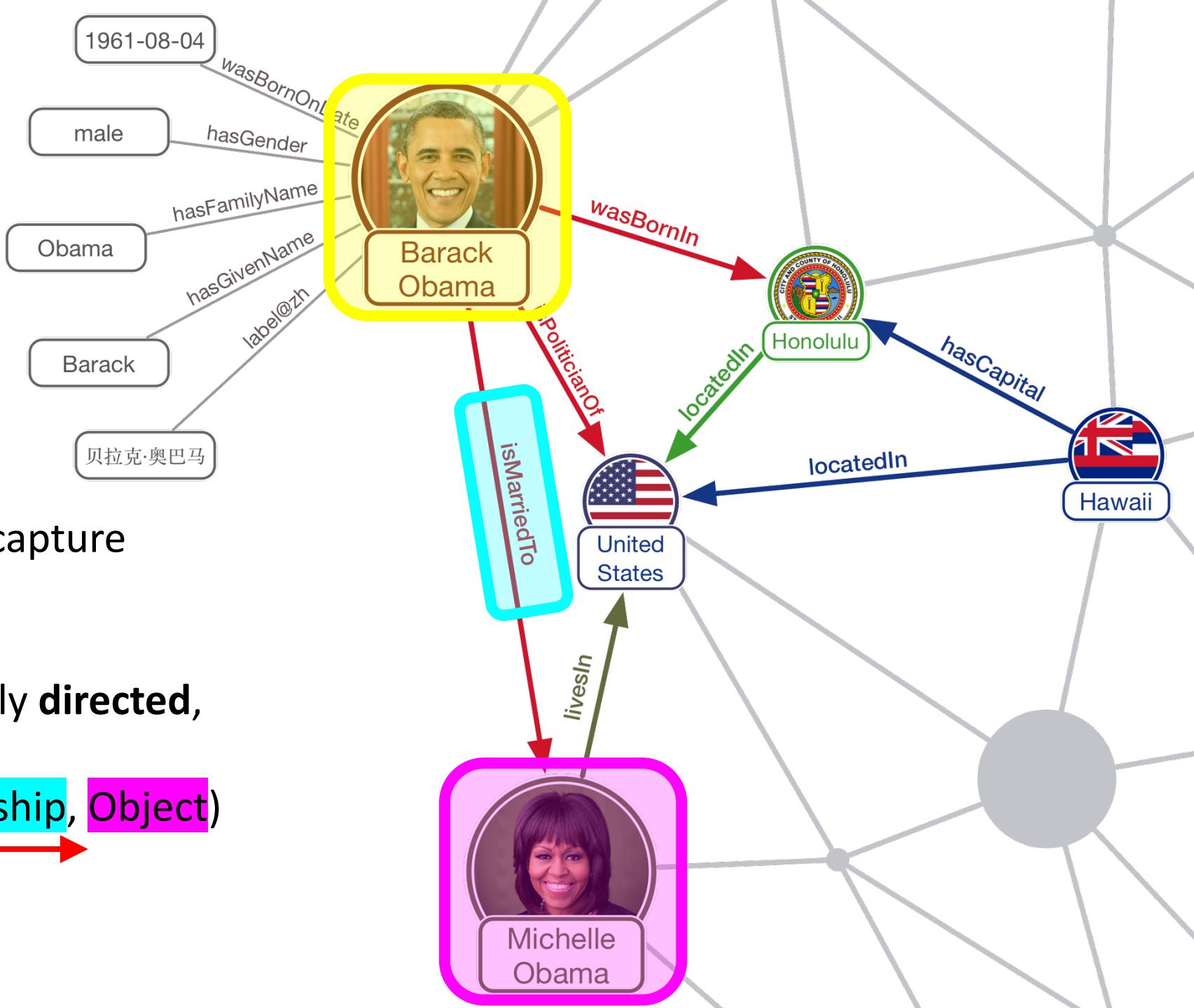
What are Knowledge Graphs?

- **Graph nodes** are entities
- Entities may optionally have **types/attributes**
 - Are they people / places / organizations?



What are Knowledge Graphs?

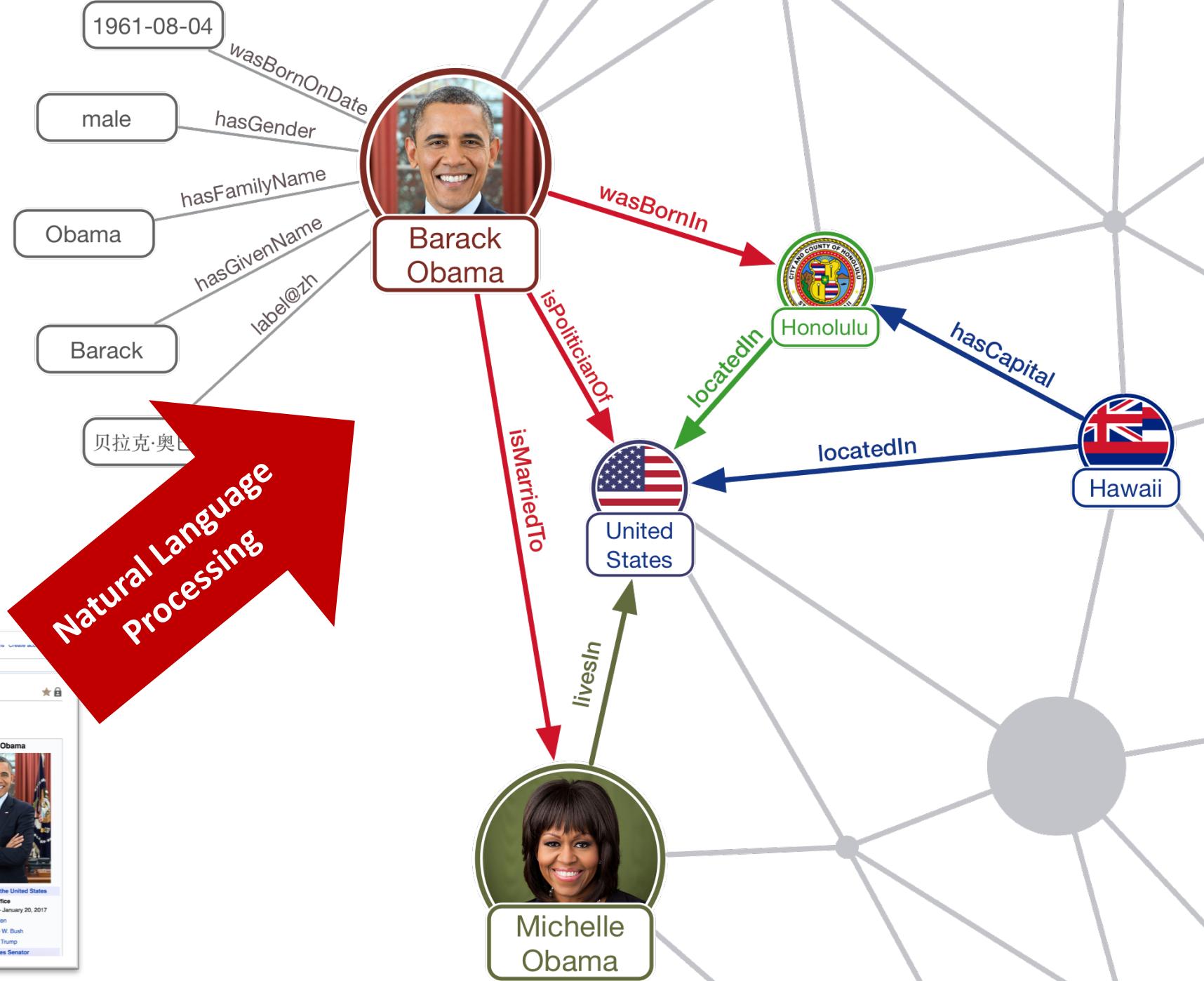
- Edges between nodes capture relationships
- Relationships are usually **directed**, forming triplets:
(Subject, Relationship, Object)



How to build Knowledge Graphs? → NLP

(Unstructured Text Document)

The screenshot shows the Wikipedia article for "Barack Obama". The page title is "Barack Obama". Below the title, it says "From Wikipedia, the free encyclopedia". The main content starts with a brief introduction: "Barack" and "Obama" redirect here. For other uses, see [Barack \(disambiguation\)](#), [Obama \(disambiguation\)](#), and [Barack Obama \(disambiguation\)](#). The article then provides a detailed biography of Barack Hussein Obama II, born August 4, 1961, and his political career.



And how to collect Unstructured Text?

→ Web Scraping

Main Document: Barack Obama

Not logged in | Talk | Contributions | Create account | Log in

Barack Obama

From Wikipedia, the free encyclopedia

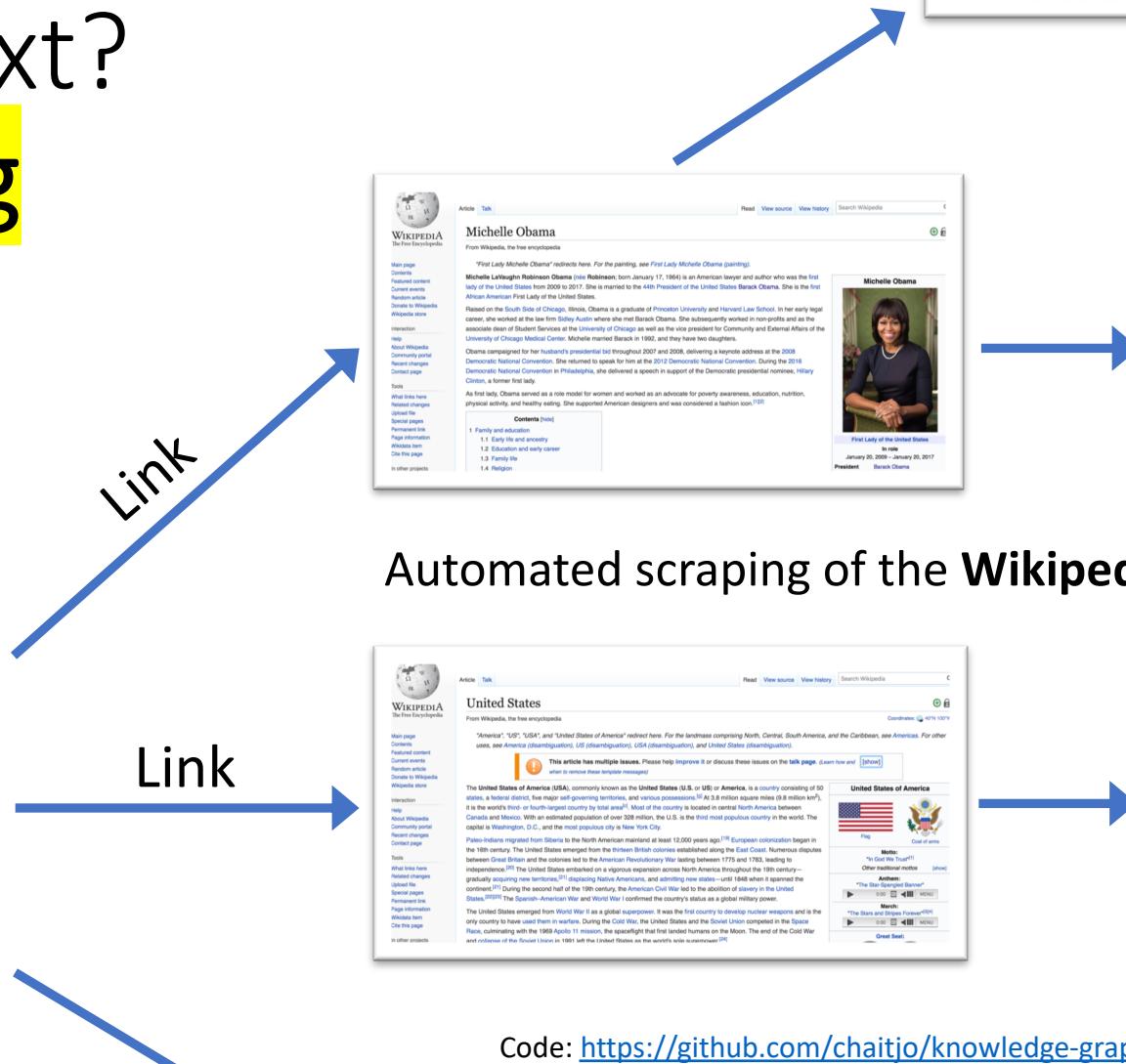
"Barack" and "Obama" redirect here. For other uses, see [Barack \(disambiguation\)](#), [Obama \(disambiguation\)](#), and [Barack Obama \(disambiguation\)](#).

Barack Hussein Obama II (*bḁrək hoō̥zəm əbəmə*) (*listen*)^[1] born August 4, 1961) is an American politician and attorney who served as the 44th [President of the United States](#) from 2009 to 2017. A member of the Democratic Party, he was the first African-American president of the United States. He previously served as a U.S. senator from Illinois from 2005 to 2008 and an Illinois state senator from 1997 to 2004.

Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black person to head the [Harvard Law Review](#). After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004. To turn to elective politics, he represented the 13th district from 1997 until 2004 in the Illinois Senate, while he ran for the U.S. Senate. Obama received national attention in 2004 with his March [Senate-primary](#) win, his well-received July Democratic National Convention keynote address, and his landslide November election to the Senate. In 2008, he was nominated for president a year after his presidential campaign began, and after [close primary](#) campaigns against Hillary Clinton. Obama was elected over Republican John McCain and was inaugurated on January 20, 2009. Nine months later, he was named the 2009 [Nobel Peace Prize laureate](#).

Obama signed many landmark bills into law during his first two years in office. The main reforms that were passed include the Patient Protection and Affordable Care Act (commonly referred to as the "Affordable Care Act" or "Obamacare"), the Dodd-Frank Wall Street Reform and Consumer Protection Act, and the Don't Ask, Don't Tell Repeal of 2010. The American Recovery and Reinvestment Act of 2009 and Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010 served as economic stimulus amidst the Great Recession. After a lengthy debate over the national debt limit, he signed the Budget Control and the American Taxpayer Relief Act. In foreign policy, he increased U.S. troop levels in [Afghanistan](#), reduced nuclear weapons with the United States-Russia New START treaty, and ended military involvement in the Iraq War. He ordered military involvement in Libya, contributing to the overthrow of Muammar Gaddafi. He also ordered the military operations that resulted in the death of Osama bin Laden, and supported

Automated scraping of the Wikipedia links network



Code: https://github.com/chaitjo/knowledge-graphs/blob/master/scrapers_utils.py

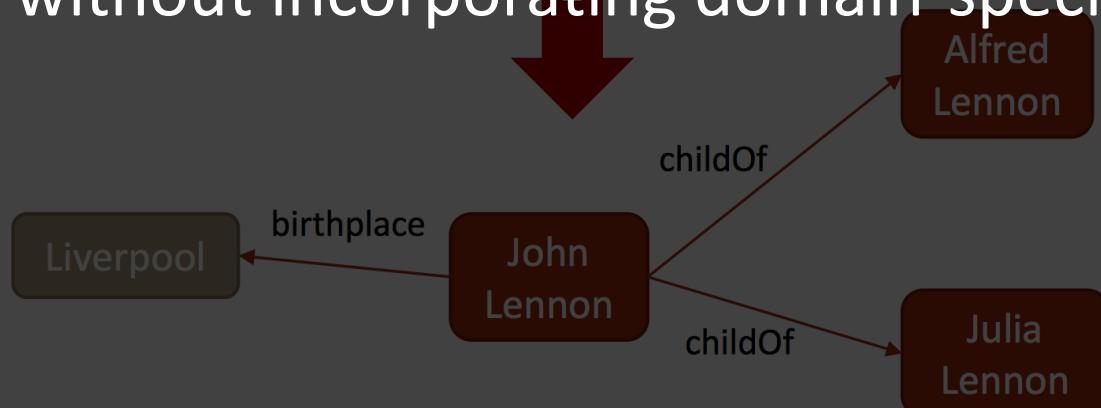
John was born in Liverpool, to Julia and Alfred Lennon.



Person	Location	Person	Person
John	Liverpool	Julia	Alfred Lennon
NNP	VBD	VBD	IN

NLP Pipeline Overview

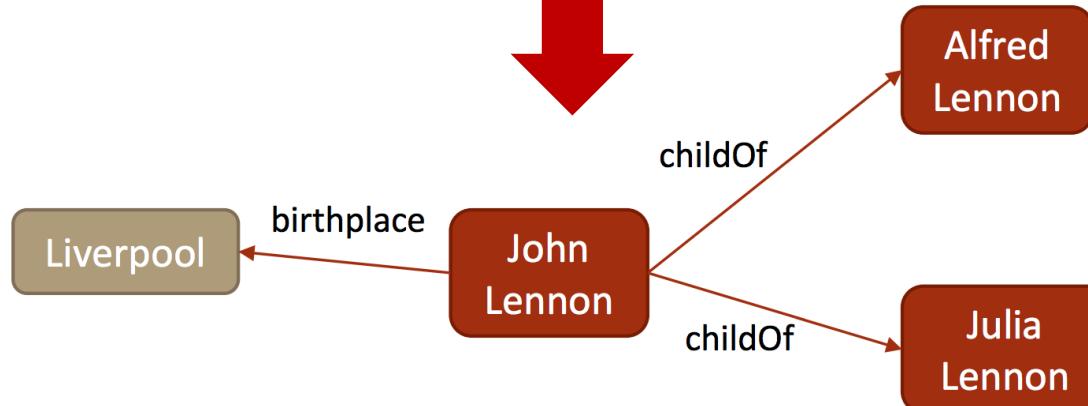
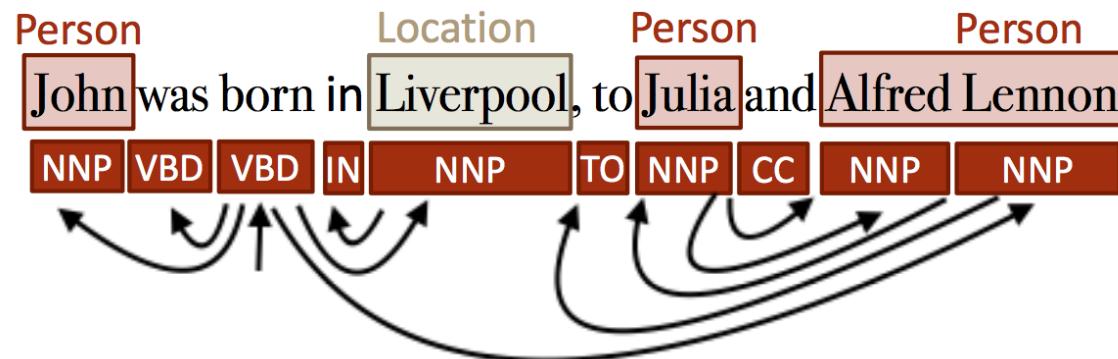
General overview of NLP approaches for KG construction, without incorporating domain-specific heuristics



- Text Cleaning
- Coreference Resolution
- Named Entity Recognition
- Part-of-speech Tagging
- Dependency Parsing

- Entity Extraction
- Relationship Extraction
- Entity-Relationship Linking
- Graph Refining/Pruning

John was born in Liverpool, to Julia and Alfred Lennon.



- Text Cleaning
- Coreference Resolution
- Named Entity Recognition
- Part-of-speech Tagging
- Dependency Parsing

- Relationship Identification
- Entity-Relationship Linking
- Graph Refining/Pruning

At a high level...

- We identify **Named Entities** and **noun chunks** in a given document. These are **candidates for nodes** in our KG.
- We extract the main **relationship** (i.e. **verb span**) for each sentence.
- For each relationship, we detect the **subject** and the **object** from the candidate set of Named Entities and noun chunks in the sentence.
- After performing some basic validation, the extracted triplet (**Subject**, **Relationship**, **Object**) is incorporated into our KG.
- Tools: **Spacy** (NER, PoS, Parsing), HuggingFace **NeuralCoref** (Coreference Resolution), HuggingFace **Transformers** (alternative to Spacy's NER).

Build a Knowledge Graph for the company **Bayer**, focused on their **Pharmacology business**

Domain-specific Challenges

How do we adapt a generic Web Scrapping + NLP pipeline for building domain-specific KGs?

Screenshot of the Wikipedia page for Bayer AG. The page title is "Bayer". The main content discusses Bayer AG's history as a German multinational pharmaceutical and life sciences company, its areas of business, and its status as a component of the Euro Stoxx 50 stock market index. A sidebar on the right provides detailed information about the company, including its logo, type (Aktiengesellschaft AG), traded as (FWB: BAYN, DAX Component), ISIN (DE000BAY0017), industry (Life sciences, Pharmaceuticals, Chemicals), founded (1 August 1863), founder (Friedrich Bayer), headquarters (Leverkusen, Germany), area served (Worldwide), and key people (Werner Baumann (CEO)).

Build a Knowledge Graph for the company **Bayer**, focused on their **Pharmacology business**

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia | 

Bayer

From Wikipedia, the free encyclopedia

This article is about the life science, chemical and pharmaceutical company. For other uses, see [Bayer \(disambiguation\)](#).
"Bayer Aspirin" redirects here. For the pharmaceutical product, see [aspirin](#).

Bayer AG (*/bər.ər, ˈbaɪ.ər/*; German: [\[bareɪ̯\]](#)) is a German multinational pharmaceutical and life sciences company and one of the largest pharmaceutical companies in the world. Headquartered in Leverkusen, Bayer's areas of business include human and veterinary pharmaceuticals; consumer healthcare products; agricultural chemicals, seeds and biotechnology products. The company is a component of the Euro Stoxx 50 stock market index.^[5] Werner Baumann has been CEO since 2016.^[6]

Founded in Barmen in 1863 as a dyestuffs factory, Bayer's first and best-known product was aspirin. In 1898 Bayer trademarked the name heroin for the drug diacetylmorphine and marketed it as a cough suppressant and non-addictive substitute for morphine until 1910. Bayer also introduced phenobarbital; prontosil, the first widely used antibiotic and the subject of the 1939 Nobel Prize in Medicine; the antibiotic Cipro (ciprofloxacin); and Yaz (drospirenone) birth control pills.

In 1925 Bayer was one of six chemical companies that merged to form IG Farben,^[7] the world's largest chemical and pharmaceutical company. The Allied Control Council seized IG Farben after World War II,^{[8][9]} because of its role in the Nazi war effort and involvement in the Holocaust, which included using slave labour from concentration camps and the purchase of humans for dangerous medical testing. It was split into its six constituent companies in 1951, then split again into three: BASF, Bayer and Hoechst.^{[9][10]}

Bayer played a key role in the Wirtschaftswunder in post-war West Germany, quickly regaining its position as one of the world's largest chemical and pharmaceutical corporations. In 2006 the company acquired Schering, in 2014 it acquired Merck & Co.'s consumer business, with brands such as Claritin, Coppertone and Dr. Scholl's, and in 2018 it acquired Monsanto, a leading producer of genetically engineered crops, for \$63 billion.^[11] Bayer CropScience develops genetically modified crops and pesticides.^[citation needed]

Bayer AG

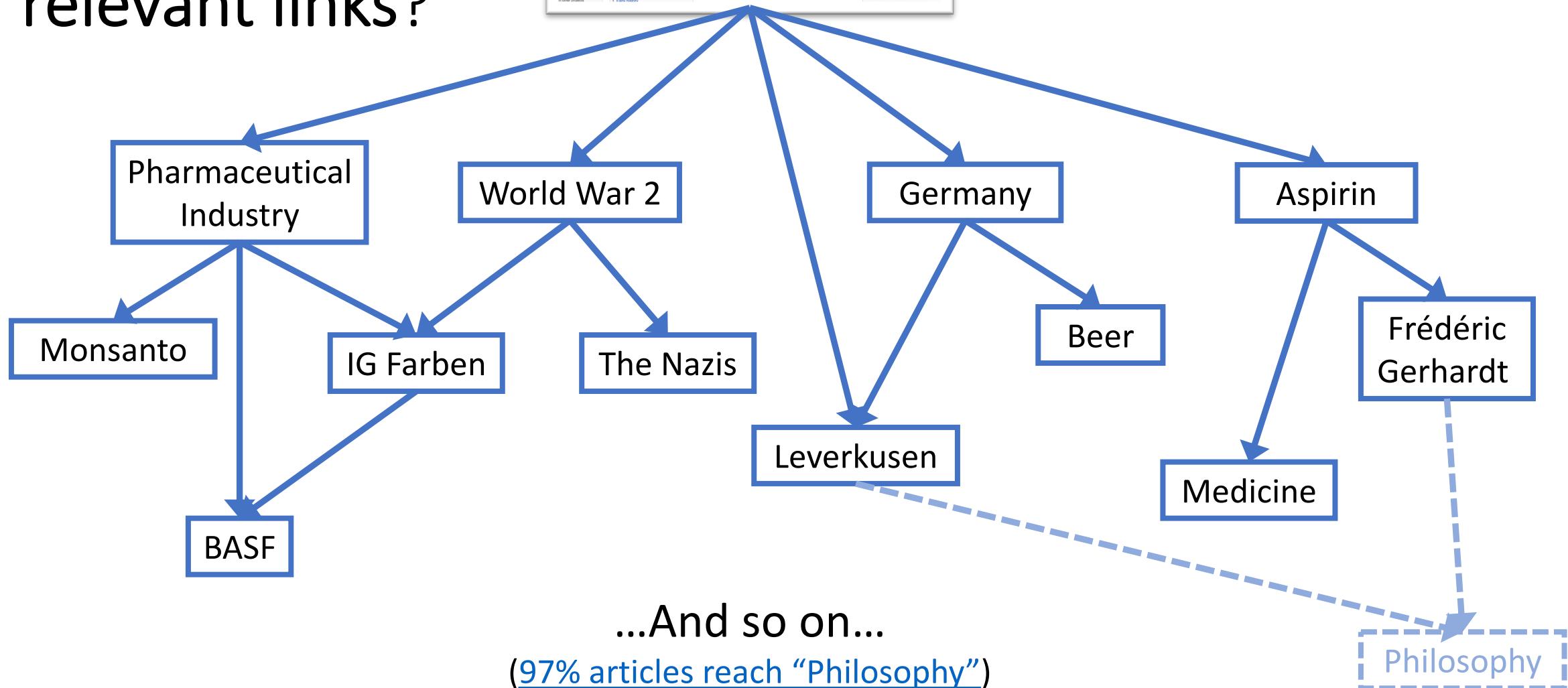
Type	Aktiengesellschaft (AG)
Traded as	FWB: BAYN  DAX Component
ISIN	DE000BAY0017 
Industry	Life sciences Pharmaceuticals Chemicals
Founded	1 August 1863; 156 years ago ^[1]
Founder	Friedrich Bayer
Headquarters	Leverkusen, Germany
Area served	Worldwide
Key people	Werner Baumann (CEO)

Contents [hide]

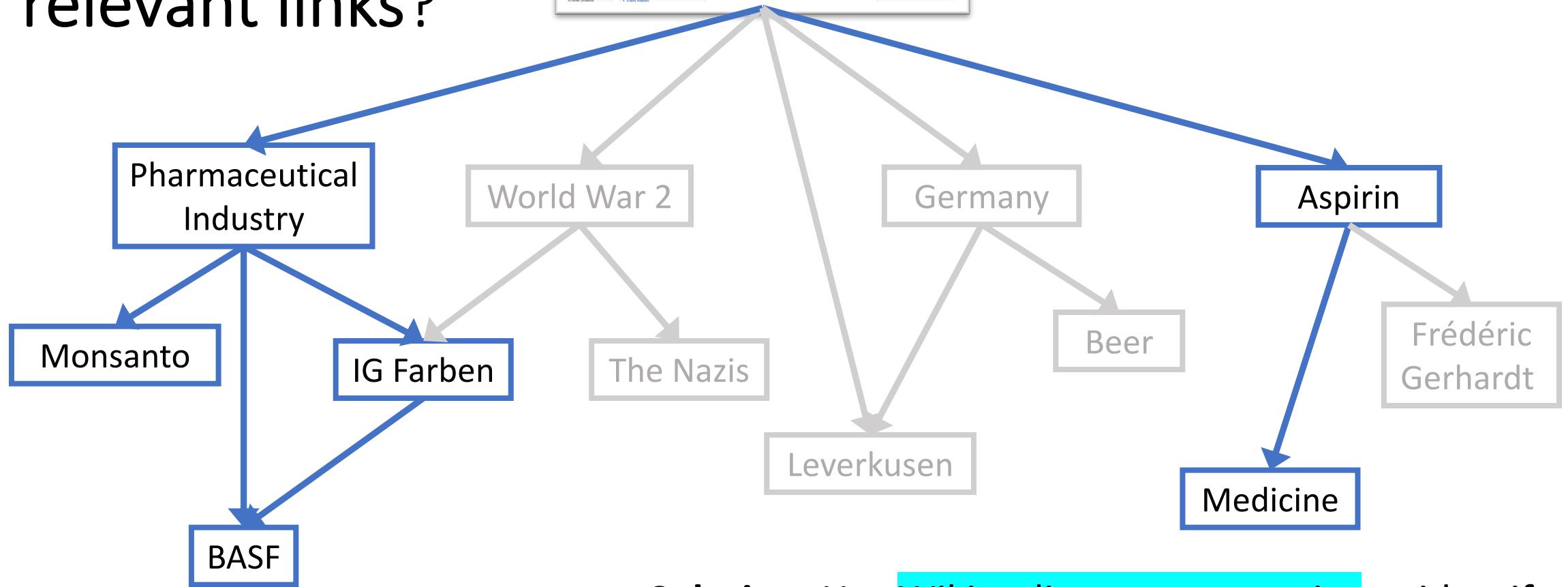
1 Early history

1.1 Foundation

Challenge 1: How to scrape relevant links?



Challenge 1: How to scrape relevant links?



Solution: Use Wikipedia page categories to identify and follow only relevant and domain-specific links

Domain-specific Wikipedia Scraping

Steps	Examples
1) Identify a small subset of important and relevant pages	Bayer , Pharmaceutical Industry, Aspirin
2) Extract categories from the relevant pages subset and identify domain-specific categories	Genetic engineering, Life sciences industry , Nanotechnology companies, Pharmaceutical companies
3) Follow links from the relevant pages subset and only construct KG from pages who's categories overlap with domain-specific categories	Biotechnology, Monsanto , BASF, Roche Pharmaceuticals , Merck & Co, Life sciences, Aspirin

Potential improvements:

- **Network science**: Analyzing how pages link and back-link to each other.
- **Topic modelling**: Using the content of pages in addition to Wiki categories for identifying relevance.

Challenge 2: How to identify domain-specific entities?

Off-the-shelf NER models (e.g. the generic NLP pipeline) will either **miss** or **not assign importance** to domain-specific entities:

- names of chemicals and drugs
- names of less well known organizations

“Bayer invented prontosil, which was the subject of the 1939 Nobel Prize in medicine”

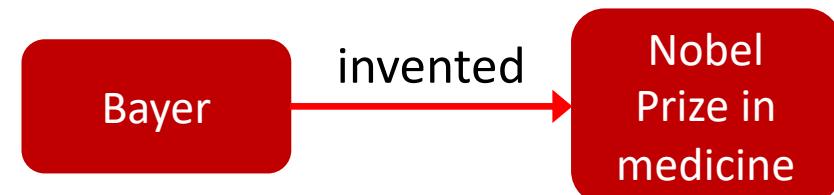


Generic NLP
Tools (Spacy)

Entities: {Bayer, Nobel Prize in medicine}
Verbs: {invented, was subject of}



Domain-specific entities (prontosil) are missed!

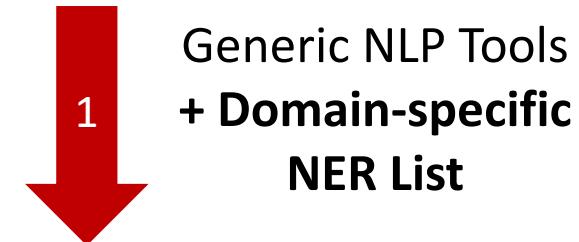


KG does not make sense...

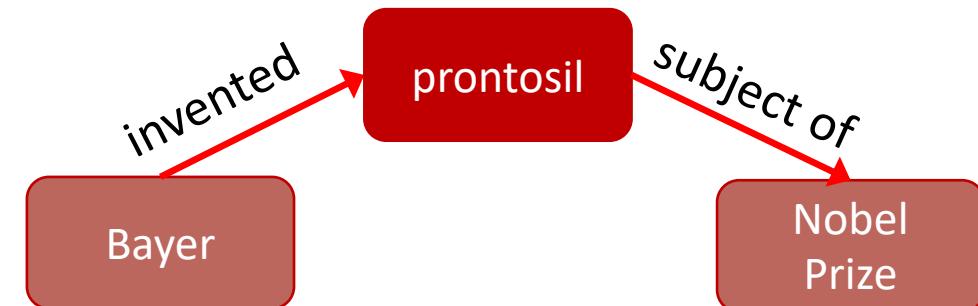
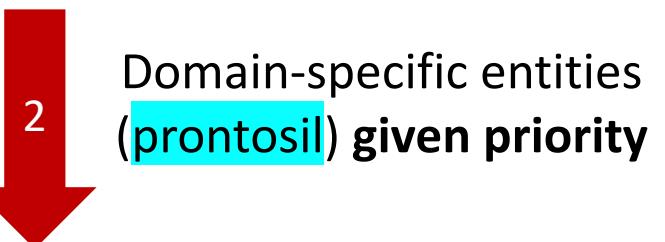
Challenge 2: How to identify domain-specific entities?

- **Manually curate** a list of domain-specific entities as the **titles of Wikipedia pages linked from Bayer**.
- Overcomes Spacy/BERT missing names of chemicals, drugs and companies which are **potentially key for useful KGs**.

“Bayer invented prontosil, which was the subject of the 1939 Nobel Prize in medicine”



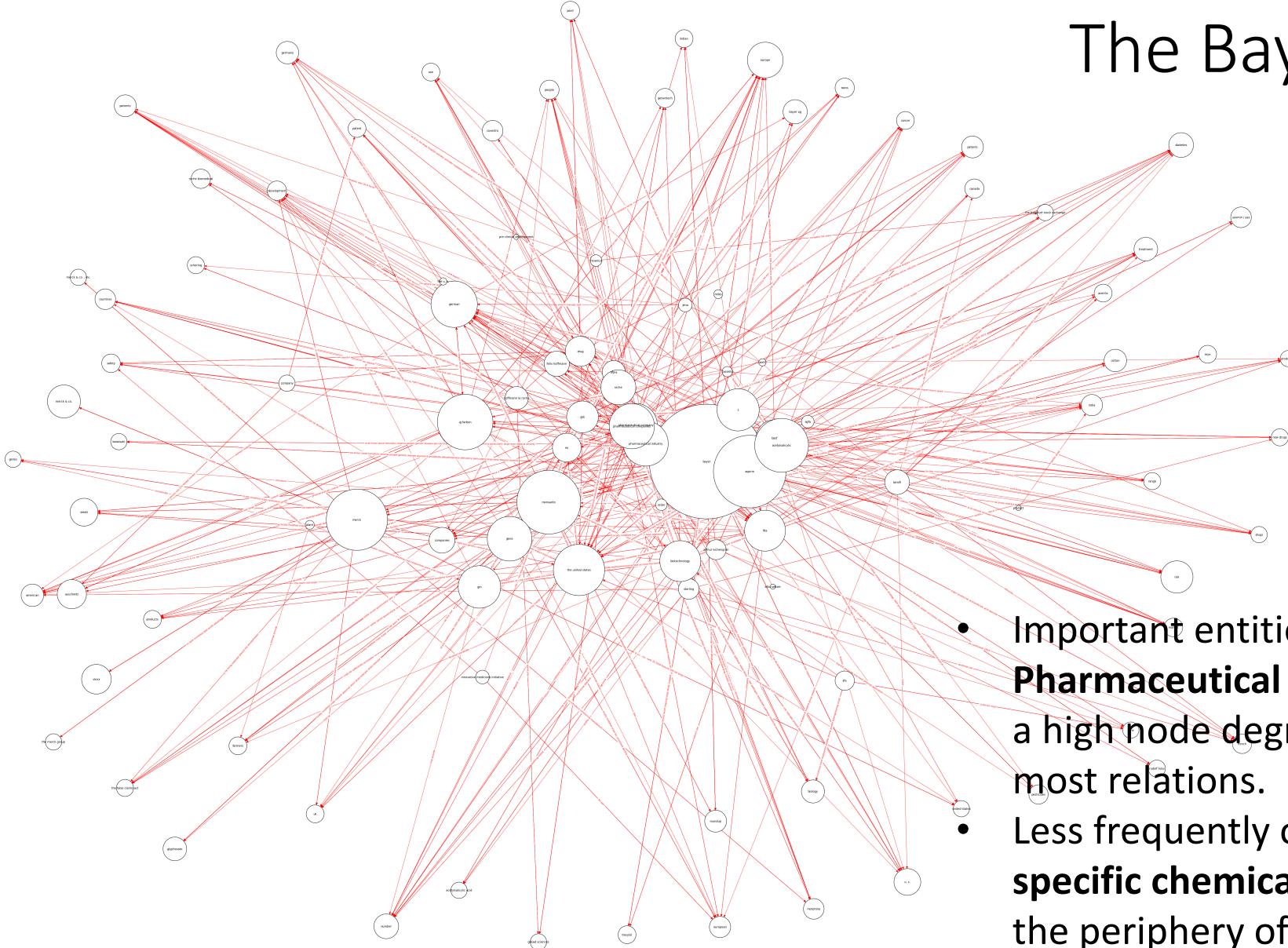
Entities: {Bayer, prontosil, Nobel Prize}
Verbs: {invented, was subject of}



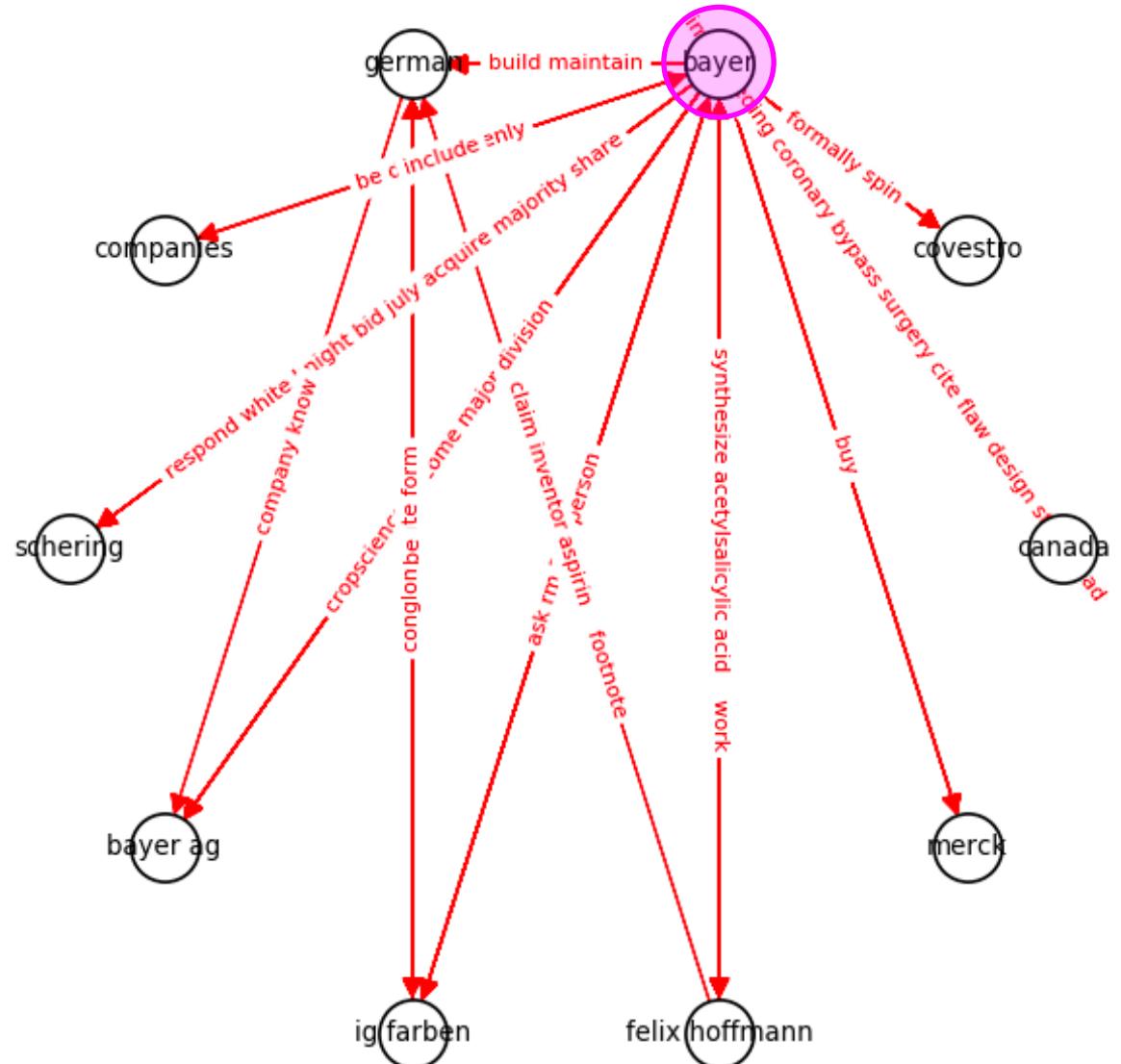
Qualitative Results

Visually inspecting the Bayer-Pharma industry Knowledge Graph for interesting relationships

The Bayer-Pharma KG

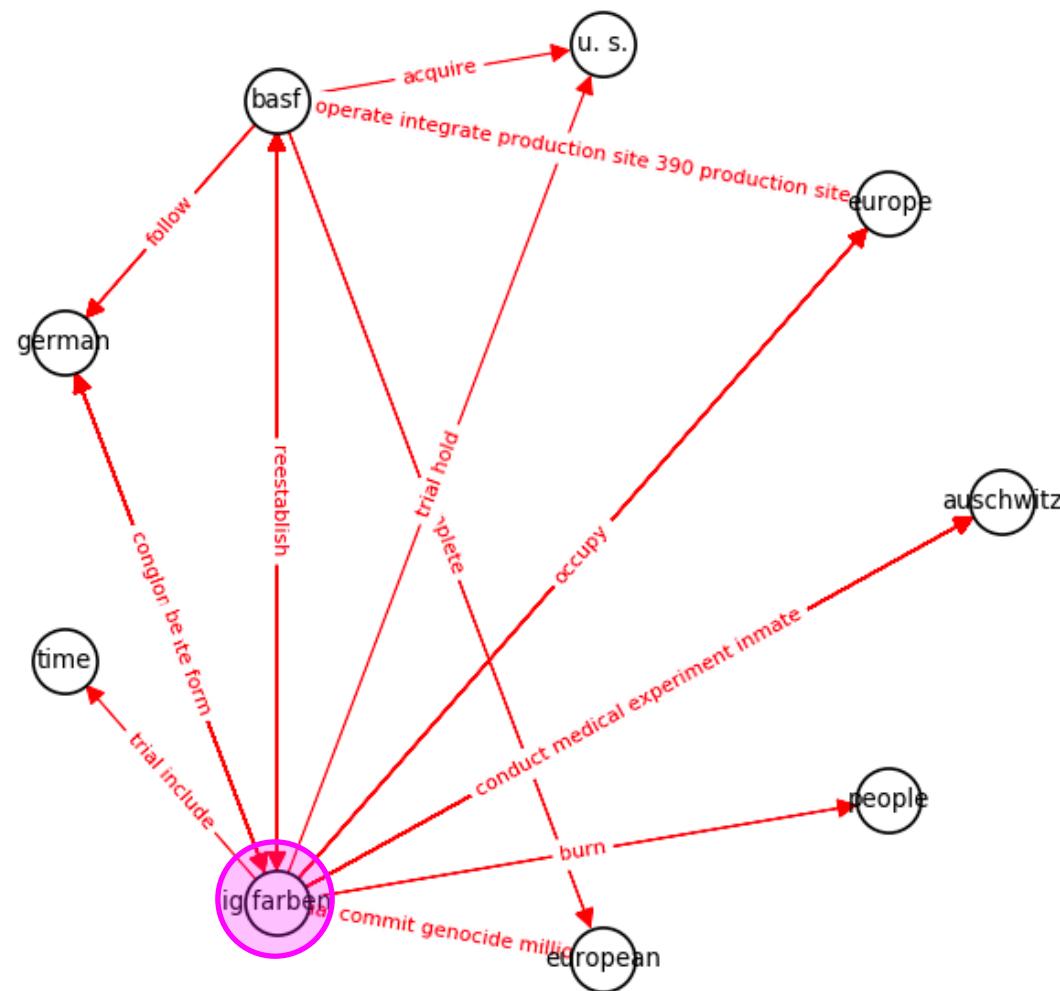


- Important entities such as **Bayer**, **Aspirin**, **Pharmaceutical Companies** are larger, i.e. have a high node degree in the KG, i.e. are part of the most relations.
- Less frequently occurring entities such as **specific chemicals** or **smaller companies** are on the periphery of the graph and are part of fewer relations.

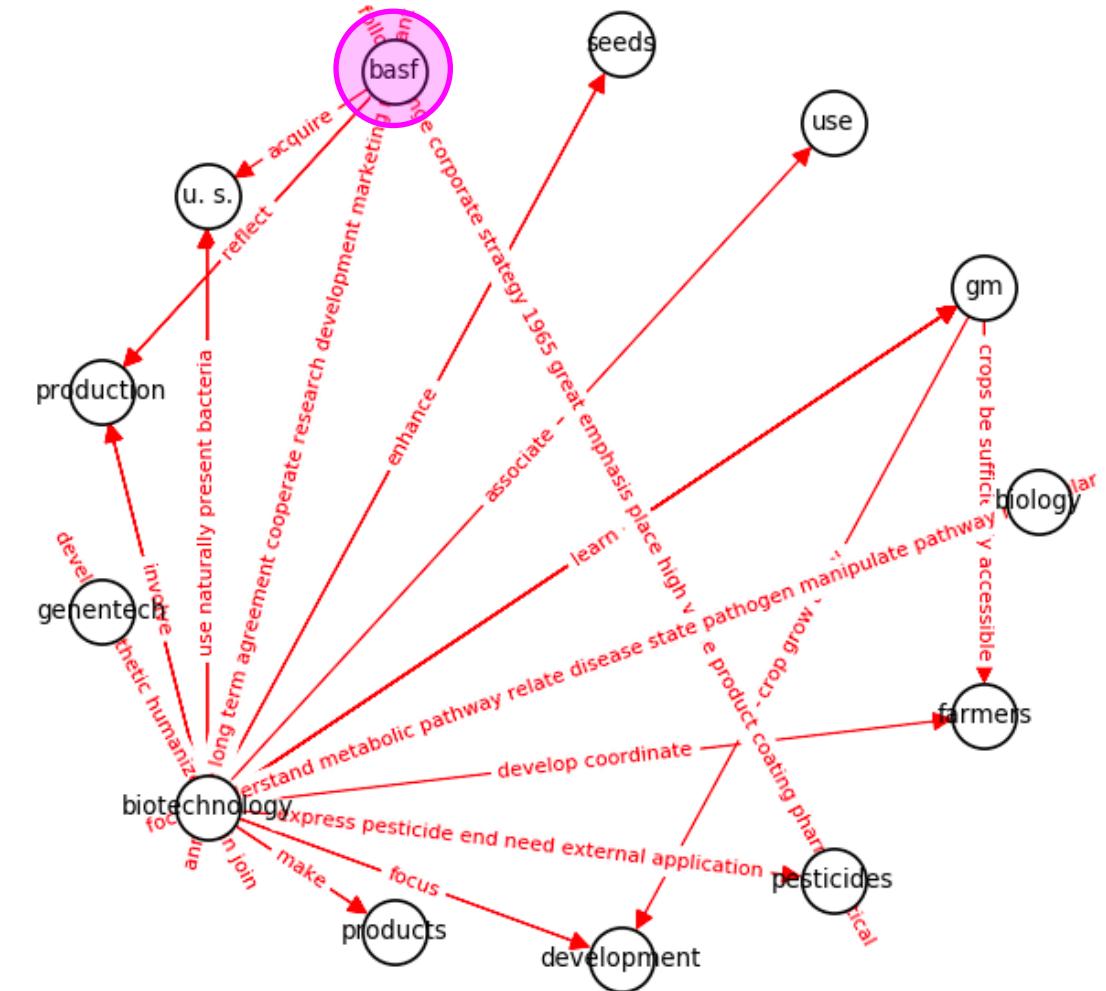


Sub-graphs around entities using 2-hop neighborhoods

- To understand specific entities better, we can plot 2-hop neighborhood sub-graphs around these entities, e.g. **Bayer** and all entities which are reachable via at most 2 ‘hops’/relationships from Bayer.
 - **Bayer** bought **Merck**
 - **Felix Hoffman** worked at **Bayer**
 - **Bayer** was part of **German** conglomerate **IG Farben**
 - **Bayer Cropscience** is a major division of **Bayer AG**
 - **Bayer** acquired a majority stake in **Schering**
- To-do: We need to improve and make more concise the content of relationship arrows!
- To-do: We can add color to each graph node according to entity type: Company, Person, Place, Noun, etc.



IG Farben



BASF

Next Steps

- How do we build powerful NLP pipelines **without handcrafting heuristics** and manual domain knowledge?
 - Supervised learning using manually annotated data, e.g. finetuning BERT-NER on domain-specific entities.
- How to refine or **improve the KG** after initial construction?
 - Semantic embeddings for KG completion, e.g. via Graph NNs.
- How do we **evaluate domain-specific KGs** extracted from free text? How do we know we're improving the system?
 - Establish **performance metrics on downstream tasks**, e.g. Question-Answering, Information Retrieval, etc.

BASF