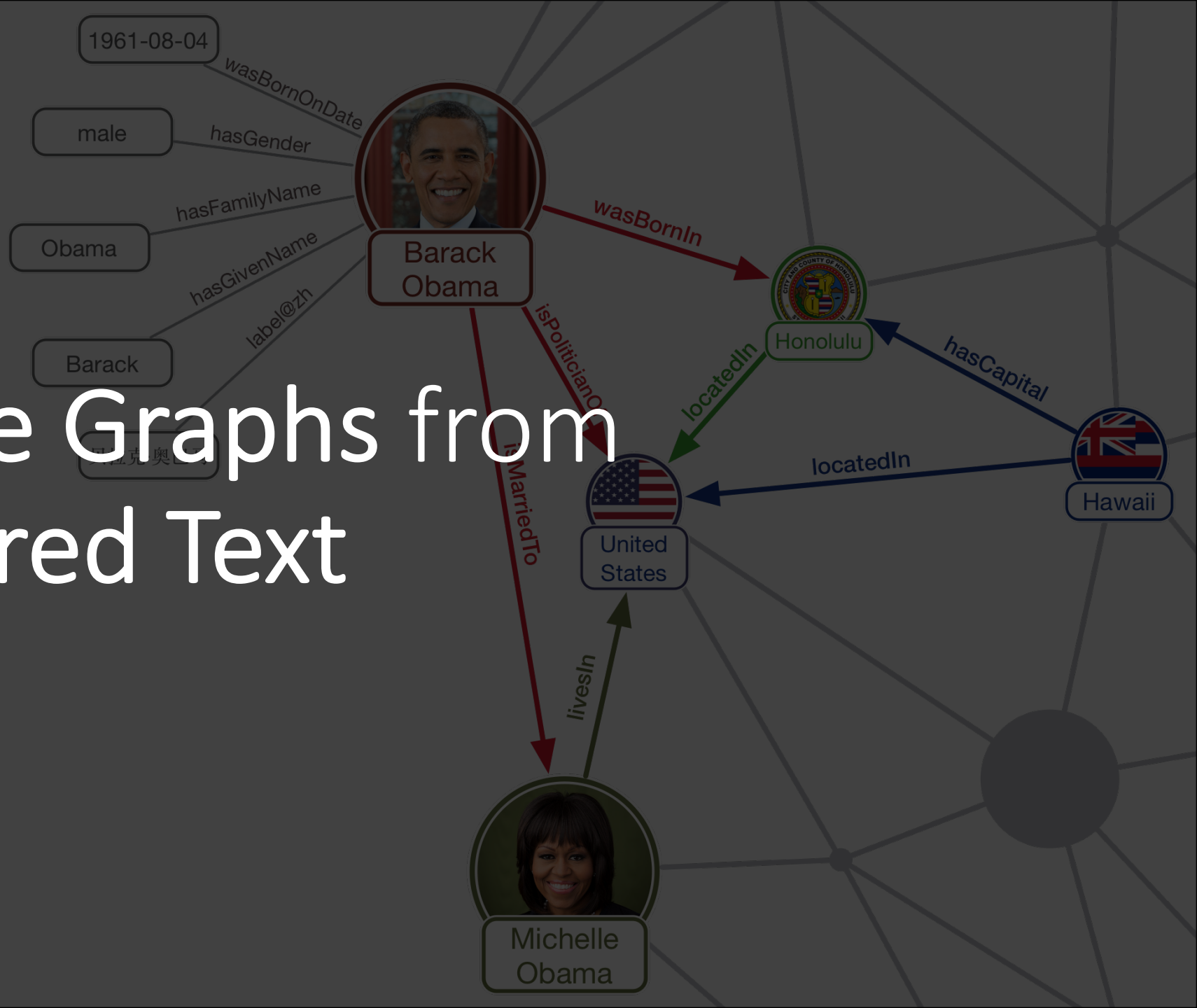
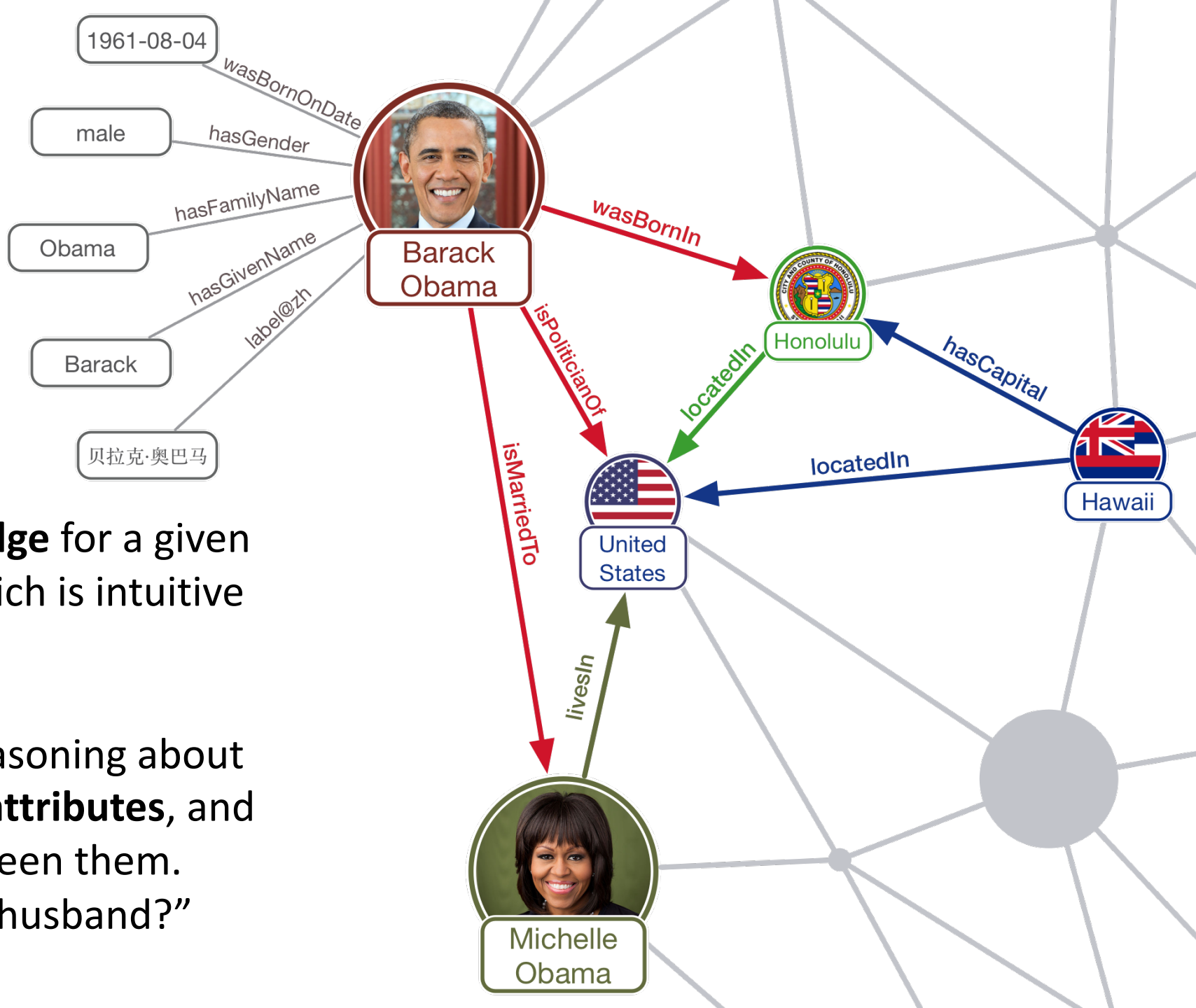
Building Knowledge Graphs from Unstructured Text

Chaitanya Joshi

(chaitjo.github.io)



What are Knowledge Graphs?

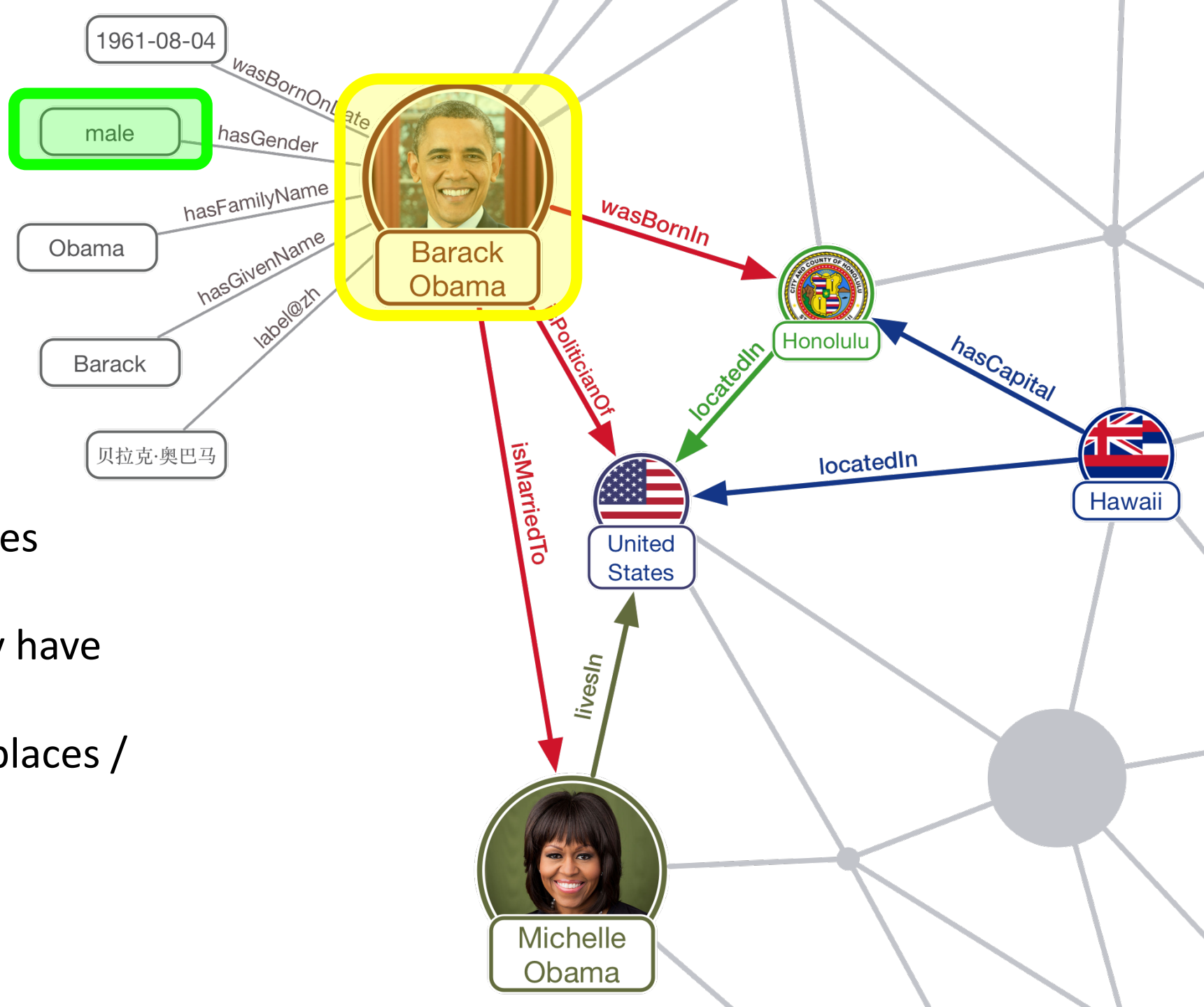


- KGs represent **knowledge** for a given domain as a **graph**, which is intuitive to understand.

- 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?

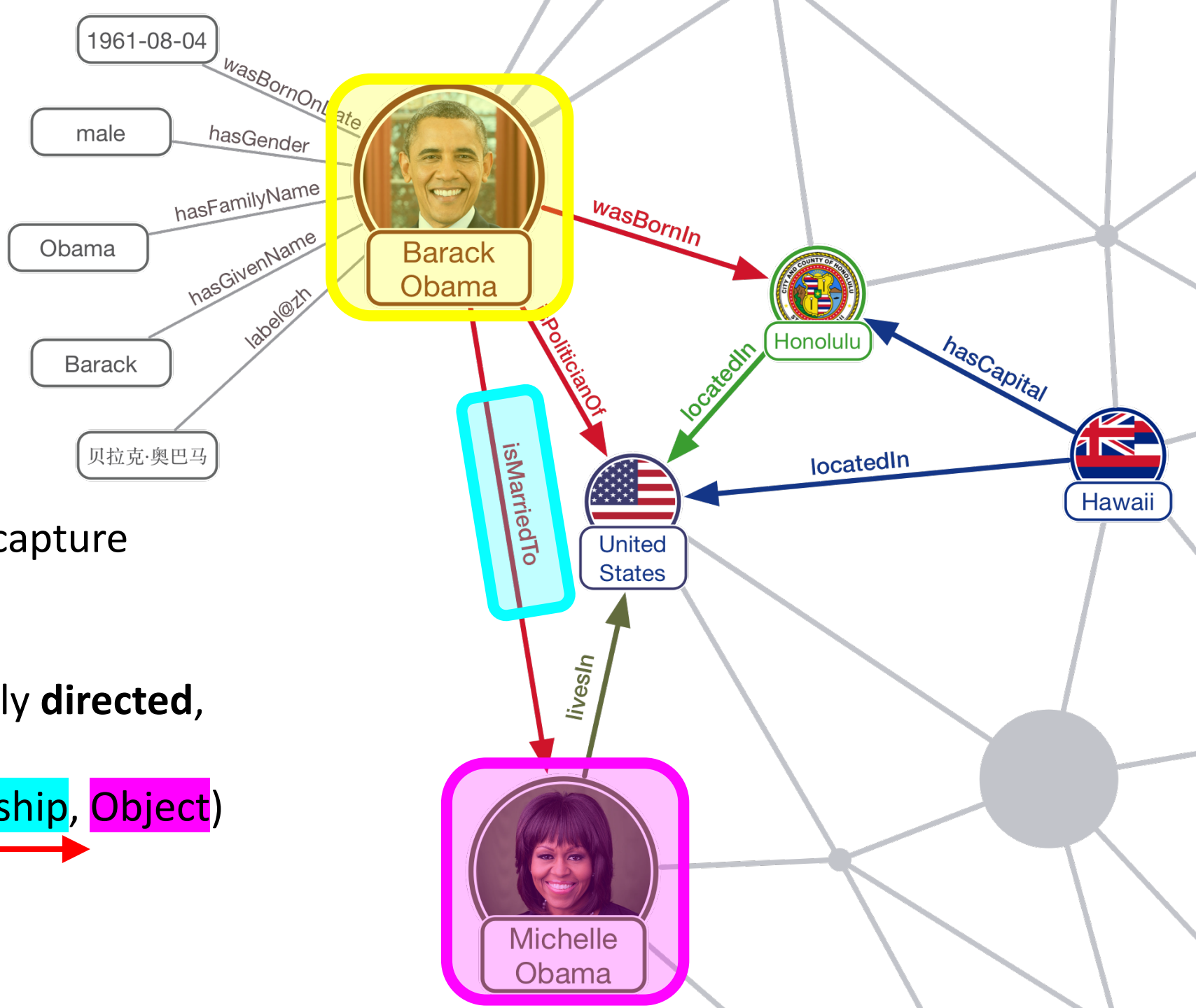
- Graph nodes are entities
- Entities may optionally have types/attributes
  - Are they people / places / organizations?



What are Knowledge Graphs?

- Edges between nodes capture relationships
- Relationships are usually **directed**, forming triplets:

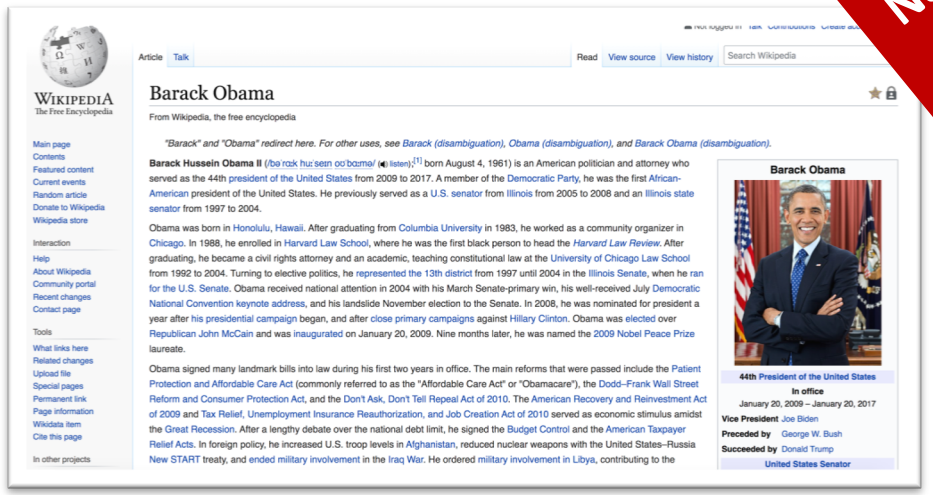
  (Subject, Relationship, Object)



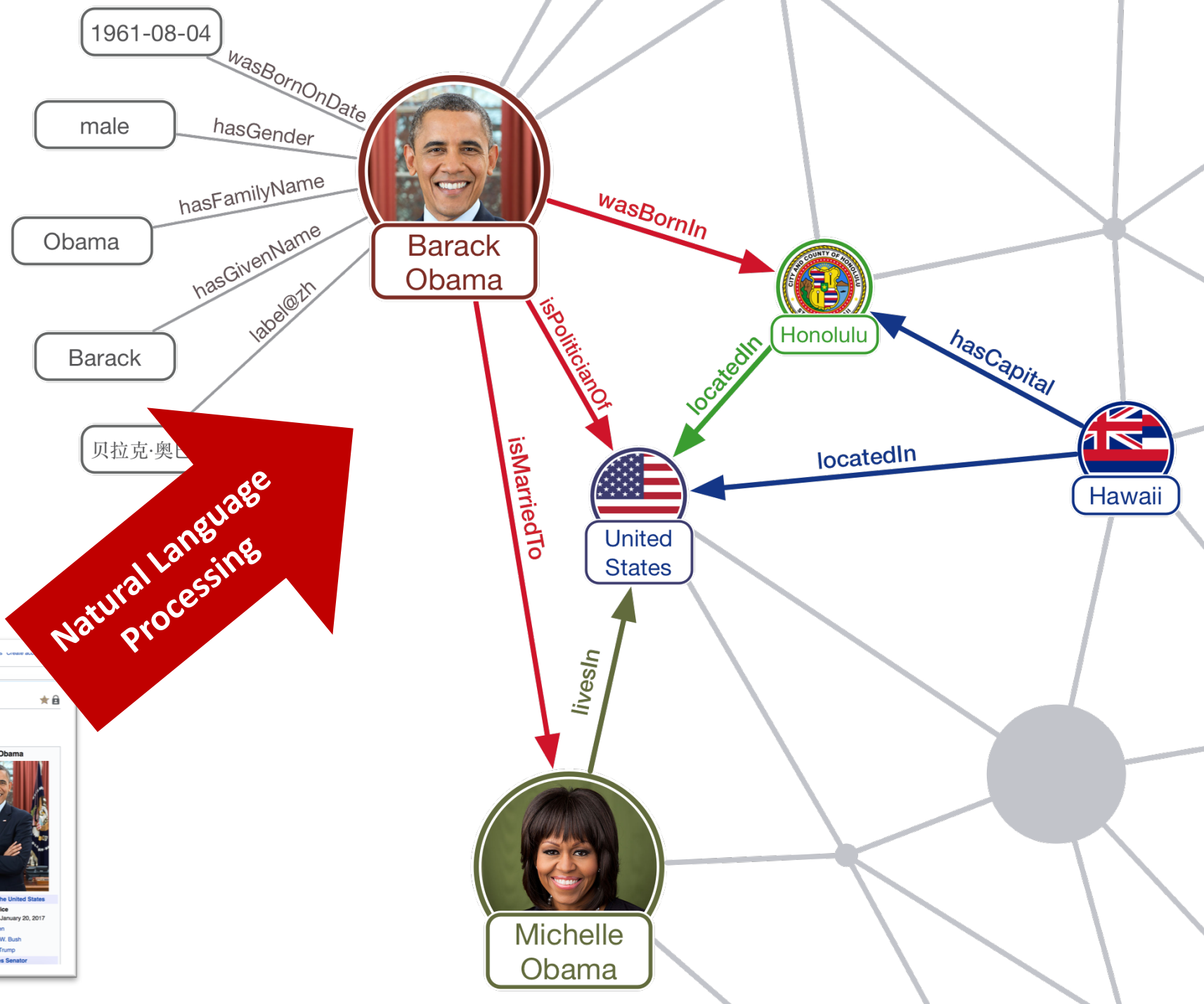
How to build Knowledge Graphs?

→ NLP

(Unstructured Text Document)



Natural Language Processing

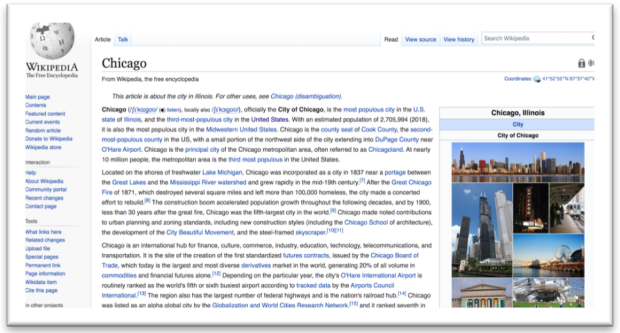


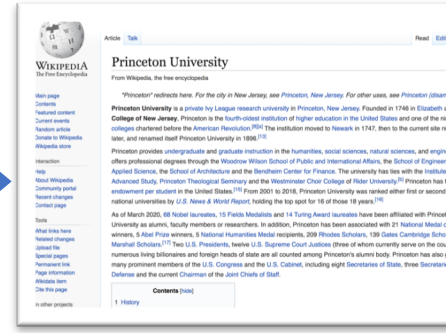
And how to collect Unstructured Text?

→ Web Scraping

Main Document: **Barack Obama**

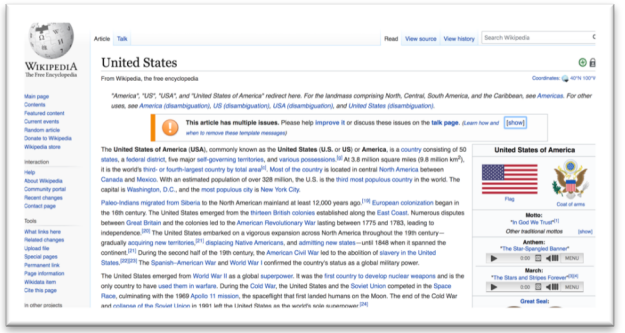
Link

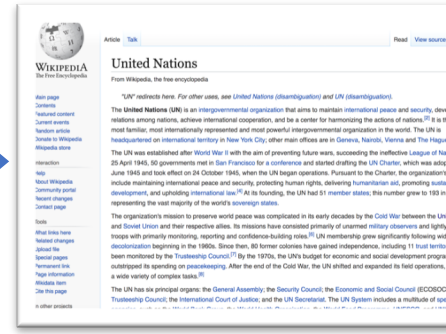




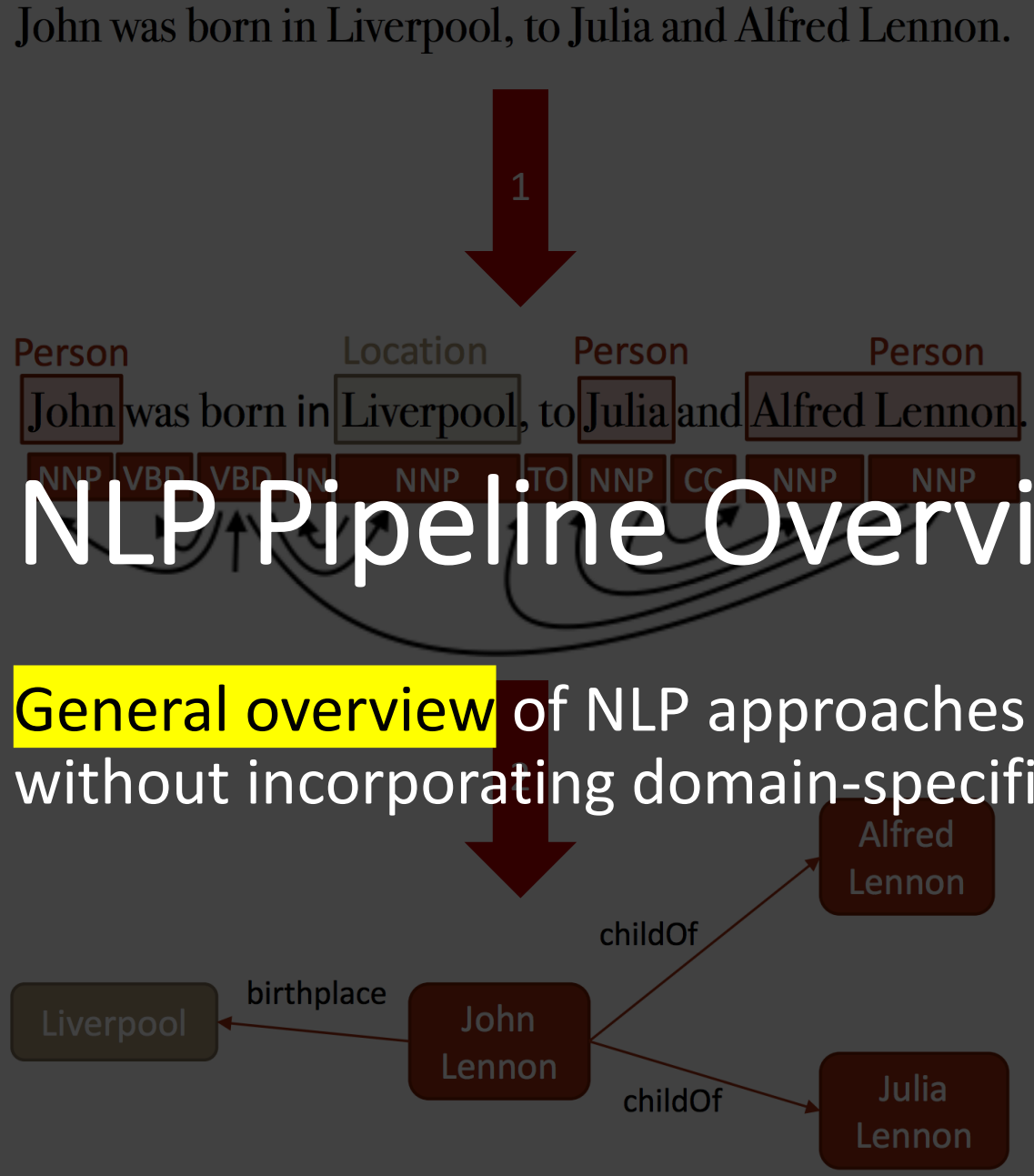
Automated scraping of the **Wikipedia links network**

Link





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



NLP Pipeline Overview

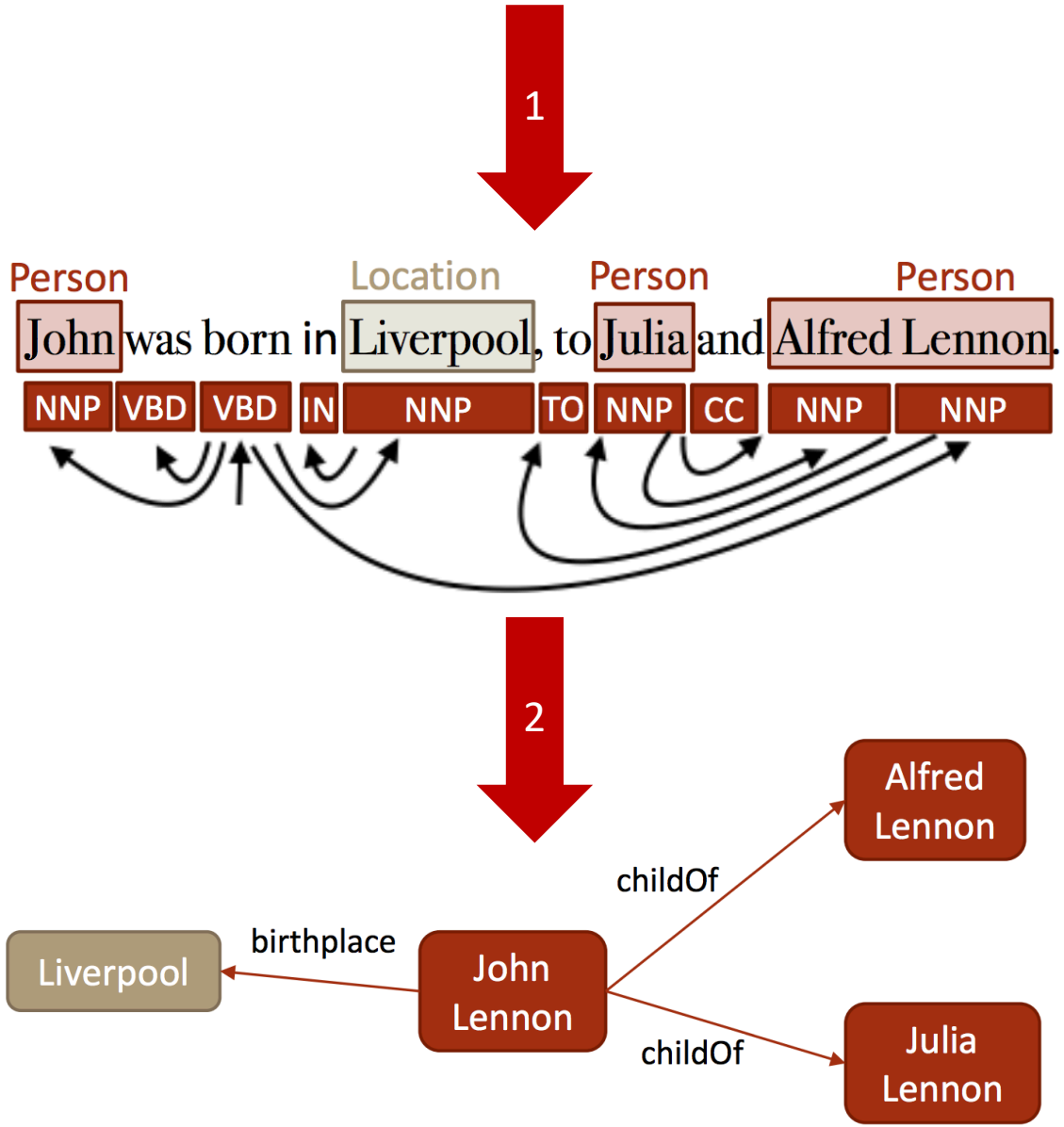
- Text Cleaning
- Coreference Resolution
- Named Entity Recognition
- Part-of-speech Tagging
- Dependency Parsing

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

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

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

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

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

At a high level…

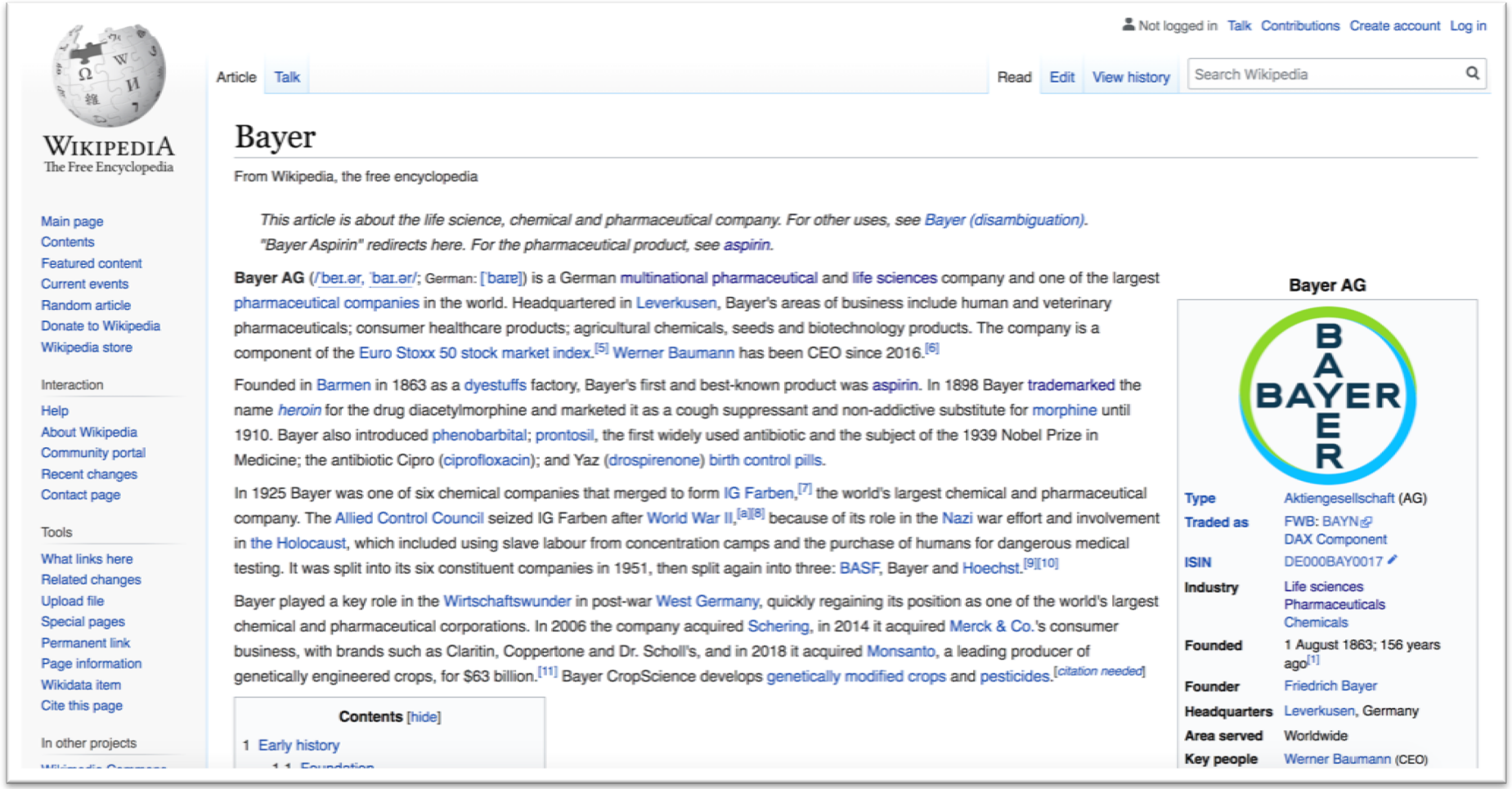
- We identify **Named Entities** and **noun chunks** in a given document. These are candidates for nodes in our KG.
- We extract the main **relationship** (i.e. verb span) for each sentence.
- For each relationship, we detect the **subject** and the **object** from the candidate set of Named Entities 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**?

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

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

In other projects

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 (/ˈbeɪ.ər, ˈbaɪ.ər/; German: [ˈbaɪɐ]) 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,[a][8] 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]

Contents [hide]
1 Early history

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)

Challenge 1: How to scrape **relevant links**?

Pharmaceutical Industry

World War 2

Germany

Aspirin

Monsanto

IG Farben

The Nazis

Beer

Frédéric Gerhardt

Leverkusen

Medicine

BASF

Philosophy

...And so on...

(97% articles reach “Philosophy”)

Challenge 1: How to scrape **relevant links**?

Pharmaceutical Industry

World War 2

Germany

Aspirin

Monsanto

IG Farben

The Nazis

Beer

Frédéric Gerhardt

Leverkusen

Medicine

BASF

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:

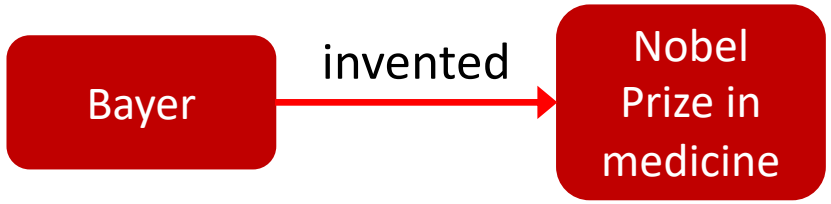
- names of chemicals and drugs
- names of less well known organizations

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

1 **Generic NLP Tools (Spacy)**

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

2 Domain-specific entities (prontosil) are missed!



KG does not make sense…

Challenge 2: How to identify **domain-specific entities**?

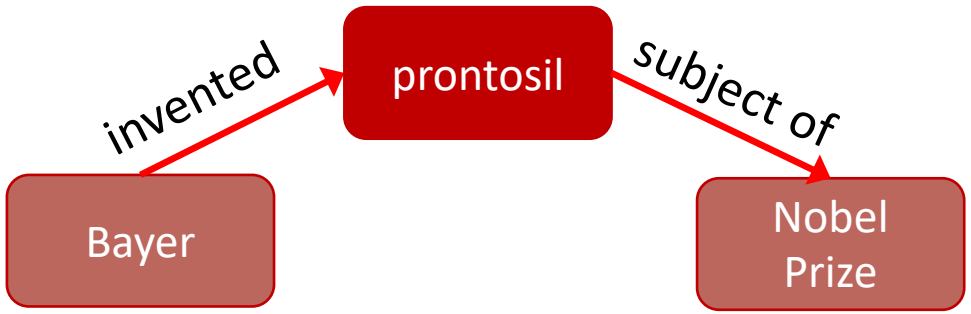
- **Manually curate** a list of domain-specific entities as the **titles of Wikipedia pages linked** from Bayer.
- Overcomes Spacy/BERT missing names of chemicals, drugs and companies which are **potentially key for useful KGs.**

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

1 → Generic NLP Tools **+ Domain-specific NER List**

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

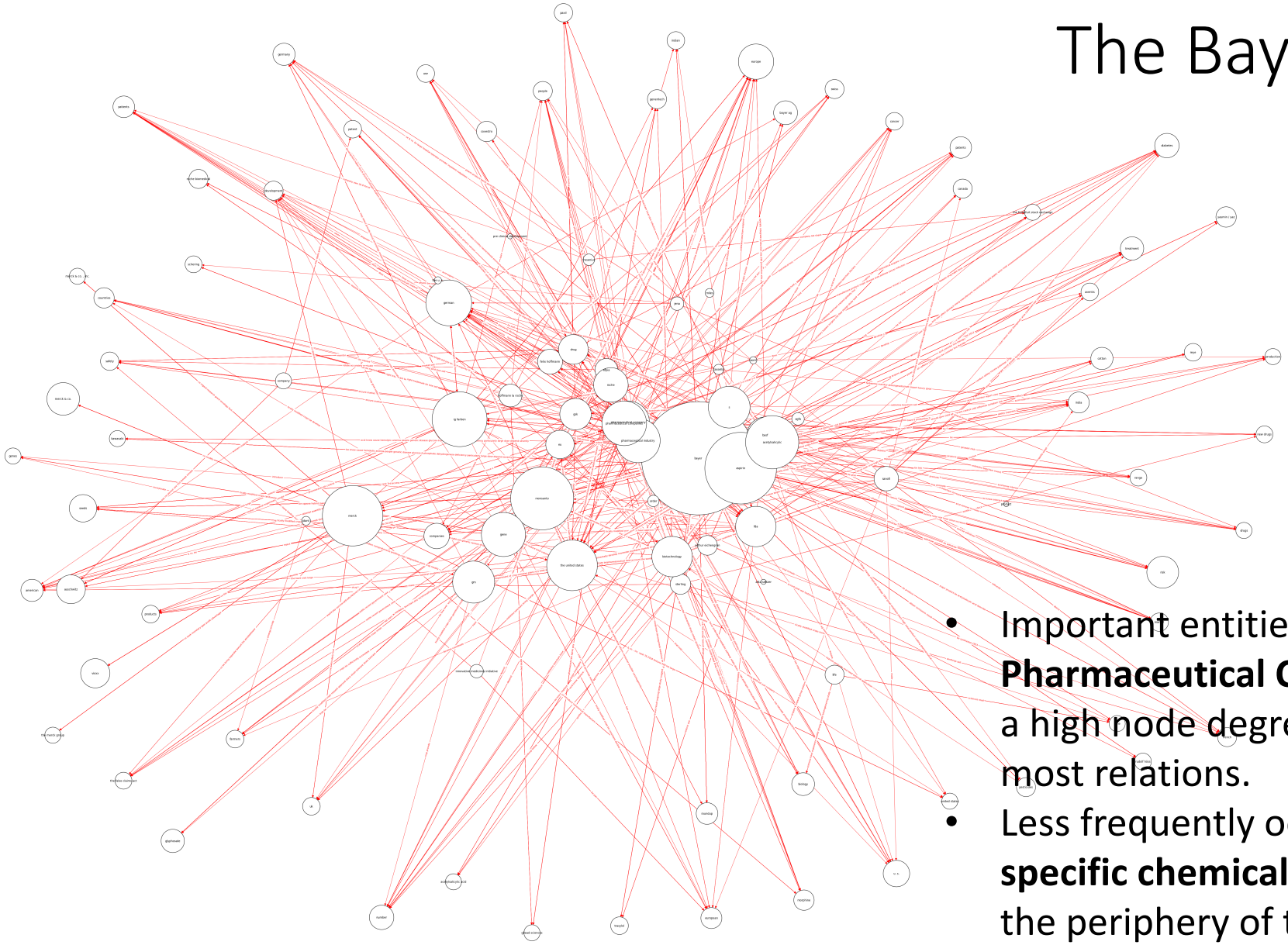
2 → Domain-specific entities (prontosil) **given priority**



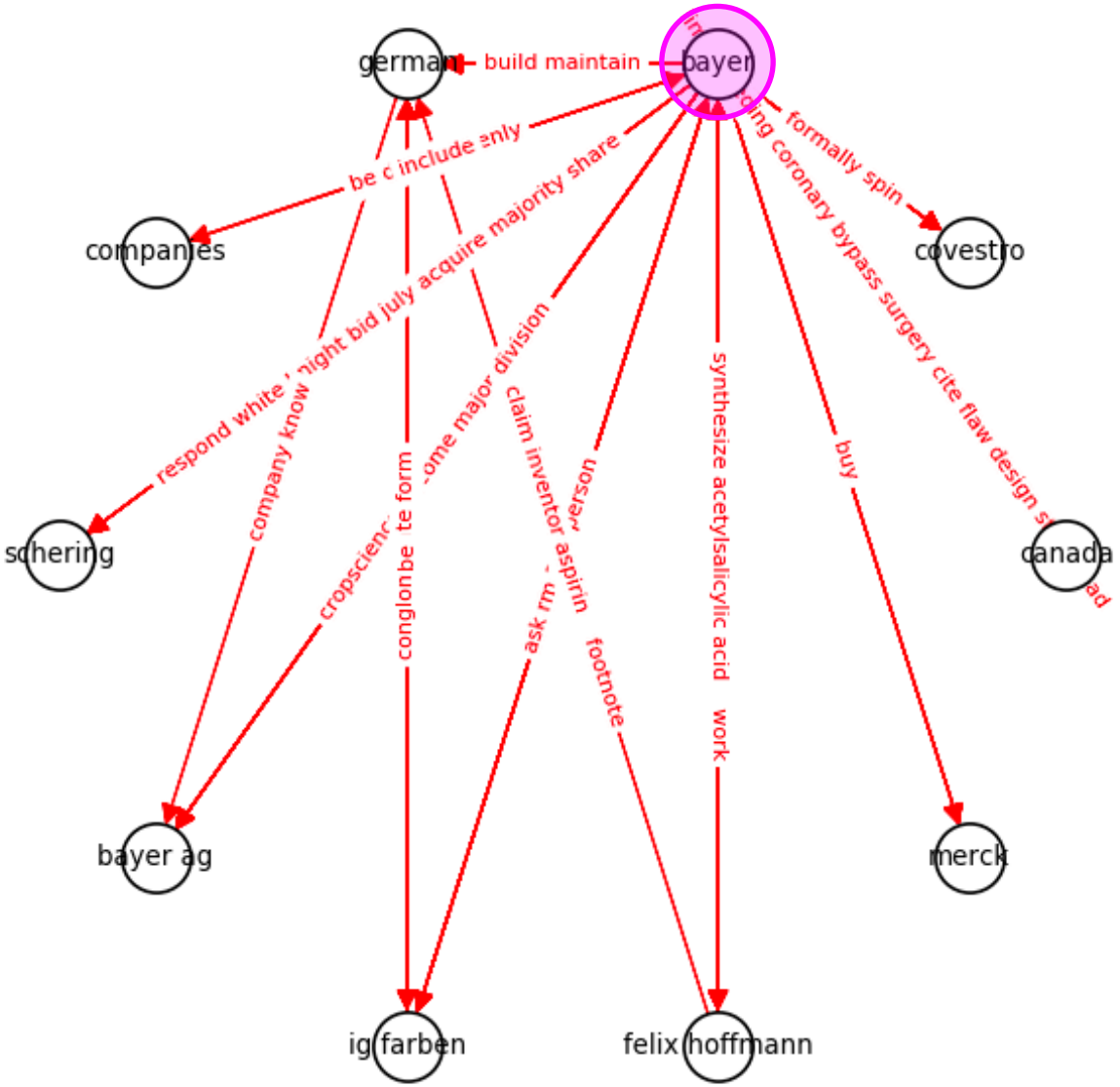
Qualitative Results

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

The Bayer-Pharma KG



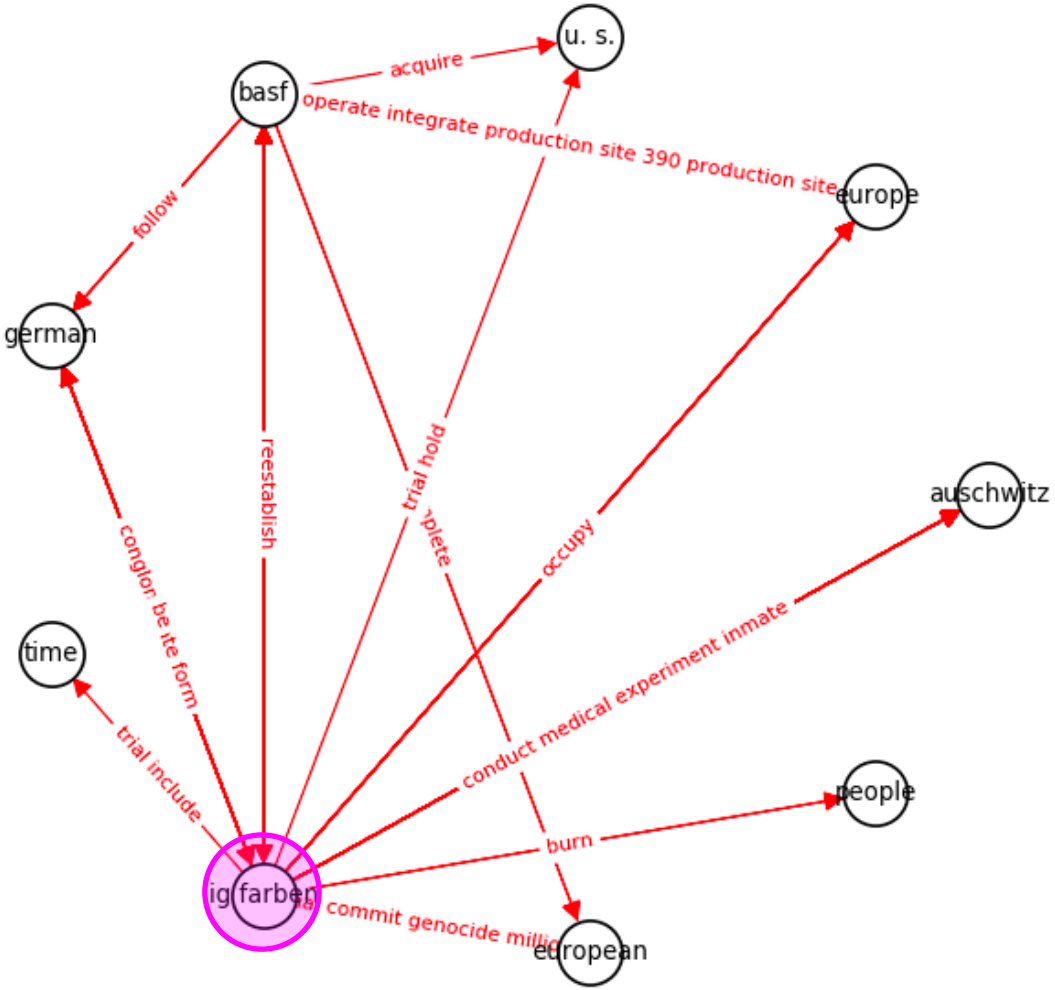
- Important entities such as **Bayer**, **Aspirin**, **Pharmaceutical Companies** are larger, i.e. have a high node degree in the KG, i.e. are part of the most relations.
- Less frequently occurring entities such as **specific chemicals** or **smaller companies** are on the periphery of the graph and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



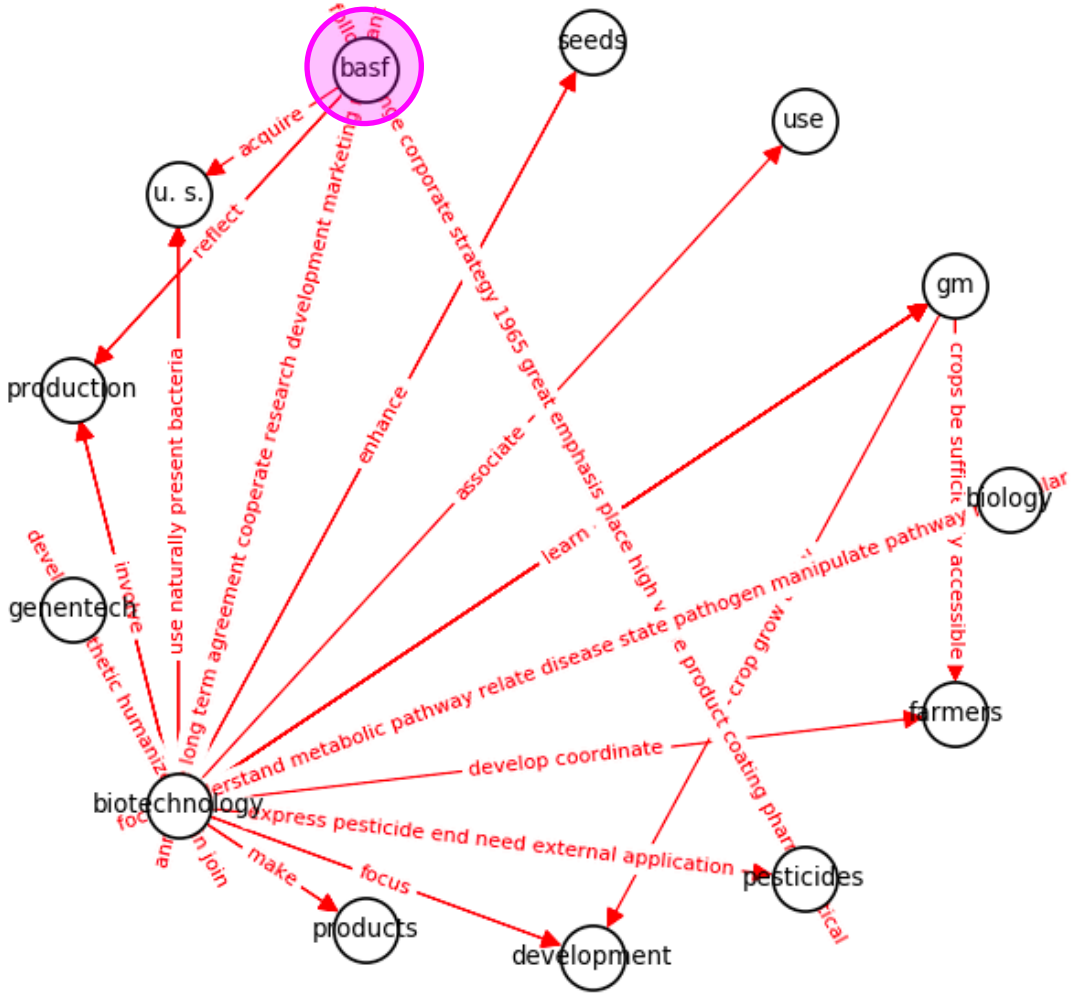
Sub-graphs around entities using 2-hop neighborhoods

- 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.