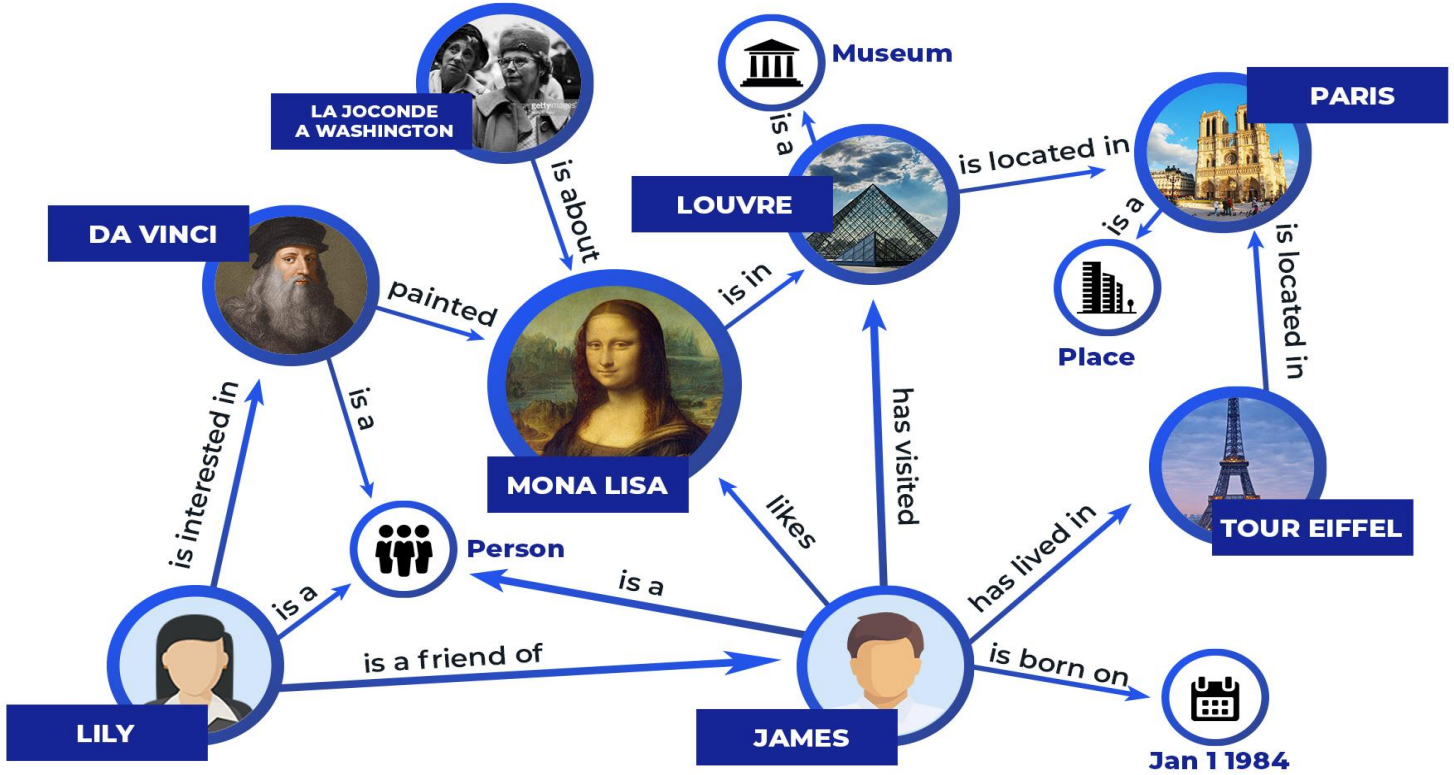
COVID-19 Knowledge Graph Analysis

CSE 261: Omkar Patil

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

What is a Knowledge Graph (KG)?

A way to represent heterogeneous data points and their relationships using nodes and edges of a graph structure.



Current Contents of KG-COVID-19

- Drug and chemical compound data: DrugCentral, PharmGKB, TTD, ChEMBL
- Functional annotations and synonyms of coronavirus genes and proteins: Gene Ontology, UniProtKB, HGNC and SciBite-CORD-19
- Ontology-based annotation: HPO and Mondo
- Biological systems like protein pathways: GO-CAM
- In total **14** different data sources comprising of **447,766** nodes and **21,611,628** edges.
- Apart from KG-COVID-19 the tool also references externals websites like Wikipedia, NBCI,

https://knowledge-graph-hub.github.io/kg-covid-19-dashboard/

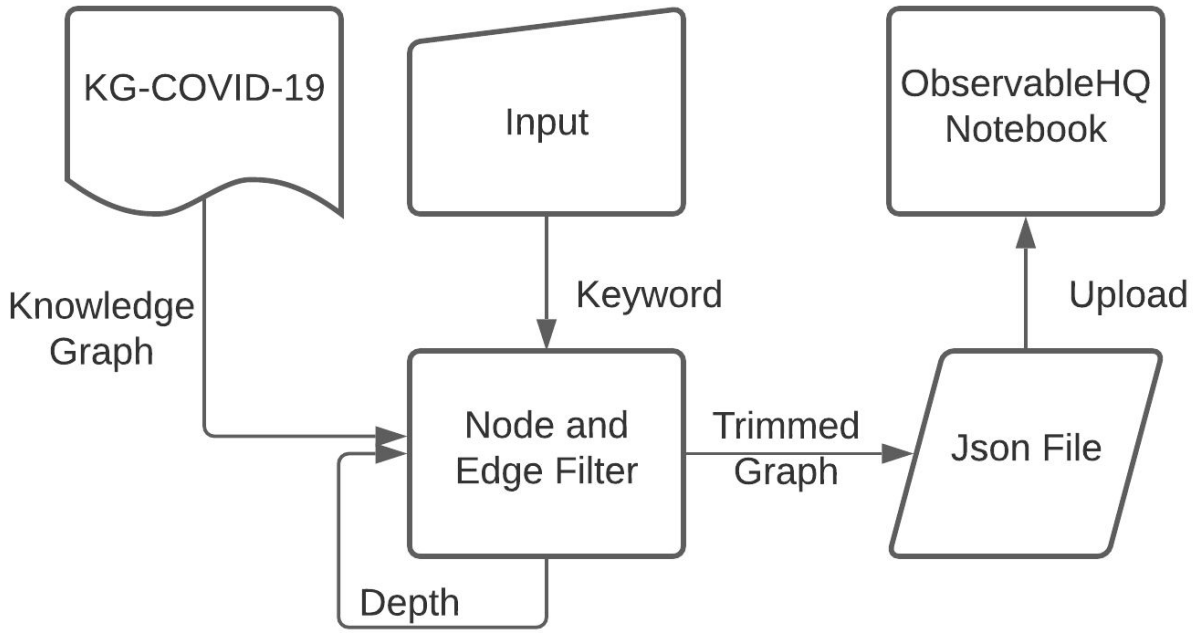
Project Idea

Not everyone can formulate proper machine learning problems to develop ML models on this graph data.
Solution: Develop an **Interactive graph exploration tool** for KG-COVID-19 to visually analyze it.

- Use attributes from the KG to **filter** out certain edges to better identify relationships between nodes beyond frequent pattern mining as proposed in the paper.
- Provide ability to perform a **macroscopic** as well as **microscopic** analysis of the graph.
- **Embed links** to original/external sources for additional reading.
- **Target Users**: Novices and Domain experts (to an extent)

Process Flow

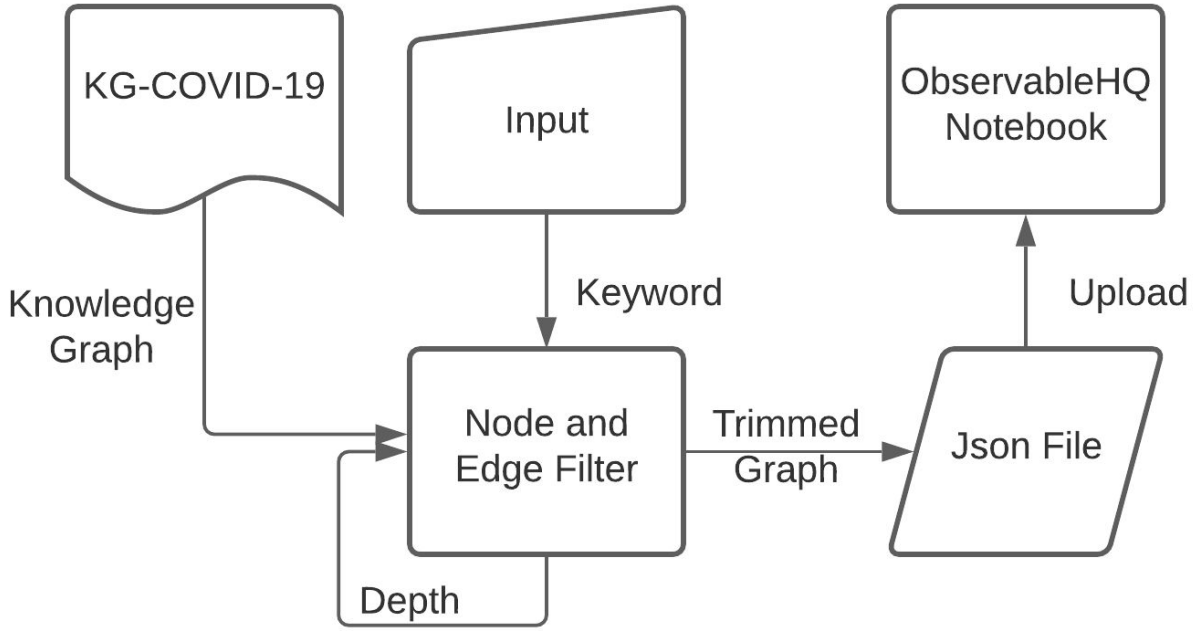
1. KG in the form of two CSV files (nodes & edges) is loaded into memory using Pandas.
2. **Filter** function identifies all nodes in the KG containing the keyword.
3. **Match** all the edges containing the nodes either as source or target.
4. **Update** the nodes list with corresponding nodes in the edges.
5. Repeat steps 3 & 4 for the given number of depth each time adding more nodes and edges to the graph.
6. Create a json file from this trimmed down graph and upload it to Observable HQ Notebook.



Process Flow

Why is **Update** and **Depth** required after **Filter** and **Match** operations?

- Due to the connected nature of the graph, once we have identified all the edges originating from a particular node, It is also important to add the destination nodes to our list otherwise the edge doesn't make any sense.
- This raises another problem though as we add additional nodes they may be connected to more edges thus it is important to limit the degree of separation from our original nodes. This is done use **Depth**.



Challenges

- Learning D3.js
  No prior experience working in frontend or pure Javascript. Observable is a very helpful tool for beginners.
- Understanding data
  All of the data sources are extremely technical so significant time was spent doing background reading and understanding the data especially the scientific nomenclature.
- Data loading and processing (High RAM machines required)
  Raw Nodes and Edges files are of 83M and 3.8G. When loaded into pandas they use 78M and 5G after dropping a significant number of columns.
- Missing nodes
  Edges file contain 65K nodes ids which are not present in the nodes file.

Current Viz

1. Categorical Bubble Charts
   - Gives the counts of different entities for a particular categorical attribute.
   - Used for a macroscopic analysis of the dataset to understand size of the contents.
2. Force Directed Graph based on subsets of KG-COVID-19
   - Plots the trimmed graph as an interactive force directed graph.
   - Clicking on any of the nodes provides additional information about the element along with any available further links.
3. Node Exploration Graph based on Wikipedia
   - Clicked node is searched on Wikipedia to identify connected pages based on hyperlinks present in the given page.

Demo

Publically available notebook present at:

https://observablehq.com/@omkarpat/kg-covid-19-visualization

Future Expansions

- Creating a document view for each of the nodes in the KG. The DocView will use NLP techniques to help people understand technical terms using *named entity recognition* and it's mappings in KG-COVID-19.
- Automatic annotations of technical document like papers to original reference materials.
- Identifying similar or conflicting views/approaches and opinions between the original documents.

Thank you!