Foundations of Semantic Knowledge Graphs

Prof. Dr. Stefan Linus Zander

Introduction to Knowledge Graphs

What are the Goals ◎ of this Course ?

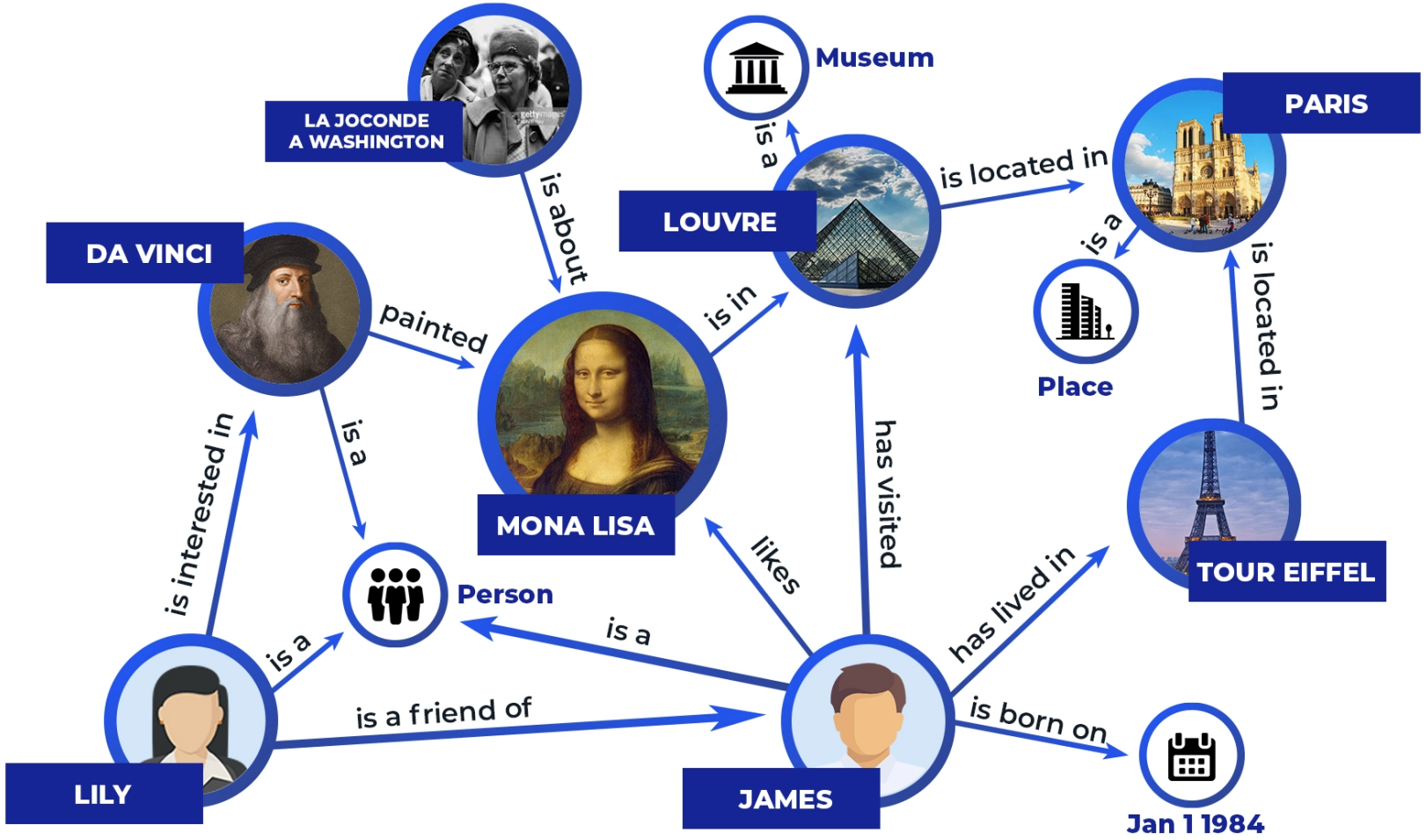
- to motivate and give a comprehensive introduction to knowledge graphs
- describe their foundational data models and how they can be queried
- discuss representations relating to schema, identity, and context
- discuss deductive and inductive ways to make knowledge explicit
- present a variety of techniques that can be used for the creation and enrichment of graph-structured data
- describe how the quality of knowledge graphs can be discerned and how they can be refined
- discuss standards and best practices by which knowledge graphs can be published
- provide an overview of existing knowledge graphs found in practice.

This course is for students and people who are new to knowledge graphs. As such, we do **not** assume that participants have specific expertise or background knoweldge on knowledge graphs.

What are Knowledge Graphs and why should we talk about them ?

A Brief History of Knowledge Graphs

The modern incarnation of the term stems from the 2012 announcement of the Google Knowledge Graph



Now, every big company has their own knowledge graph

The industrial uptake makes it hard to ignore KGs

- Airbnb ("Homes you may like"),
- Amazon (product recommendations engine; Alexa),
- eBay (eBay Shopper Experience Graph),
- Facebook (Facebook Knowledge Graph),
- IBM (Watson),
- LinkedIn (People You May Know),
- Microsoft (Cortana),
- Uber (Uber Technology Platform)

The core idea of using graphs is to represent data, often enhanced with means to explicitly represent knowledge

Sources:
https://blog.google/products/search/introducing-knowledge-graph-things-not/
https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/

Hype Cycle for Artificial Intelligence, 2020

gartner.com/SmarterWithGartner

Gartner

Excursus: How eBay uses Knowlegde Graphs

The eBay Shopper Experience Graph

eBay also uses knowledge graphs to improve its platform's functionality and provide a better user experience. eBay's knowledge graph is known as the "eBay Shopper Experience Graph".

The eBay Shopper Experience Graph uses machine learning and artificial intelligence to analyze and understand the relationships between different types of data on the platform, including product listings, user profiles, search queries, and more. This allows eBay to provide more relevant search results and personalized recommendations to its users.

The eBay Shopper Experience Graph is used to power a number of features on the platform, such as the "Shop Similar Items" and "You May Also Like" recommendations, which provide users with personalized product recommendations based on their browsing and purchase history. The knowledge graph is also used to improve eBay's search functionality, making it easier for users to find the products they are looking for.

In addition, eBay also uses its knowledge graph to identify fraudulent activity on the platform, such as fake product listings or seller accounts. By using machine learning algorithms to analyze and understand patterns of suspicious behavior, eBay is able to identify and prevent fraudulent activity before it can harm users or the platform itself.

Overall, the eBay Shopper Experience Graph plays a crucial role in helping eBay provide a more personalized and efficient platform to its users, while also improving its security and preventing fraud.

Excursus: How Airbnb uses Knowledge Graphs

The Airbnb Knowledge Graph

Airbnb also uses knowledge graphs to improve its platform and provide a better user experience. Airbnb's knowledge graph is known as the "Airbnb Knowledge Graph".

The Airbnb Knowledge Graph uses machine learning and artificial intelligence to analyze and understand the relationships between different types of data on the platform, including user profiles, listings, bookings, and more. This allows Airbnb to provide more relevant search results and personalized recommendations to its users.

The Airbnb Knowledge Graph is used to power a number of features on the platform, such as the "More Places to Stay" and "Homes You May Like" recommendations, which provide users with personalized accommodation recommendations based on their search history and preferences. The knowledge graph is also used to improve Airbnb's search functionality, making it easier for users to find the right accommodation for their needs.

In addition, Airbnb uses its knowledge graph to optimize its pricing and revenue management strategy. By analyzing patterns and trends in user behavior and accommodation demand, Airbnb is able to adjust its pricing algorithm to maximize revenue for hosts while still offering competitive prices for users.

Overall, the Airbnb Knowledge Graph plays a crucial role in helping Airbnb provide a more personalized and efficient platform to its users, while also optimizing its pricing strategy and revenue management.

Source: https://medium.com/airbnb-engineering/contextualizing-airbnb-by-building-knowledge-graph-b7077e268d5a

Knowledge Graphs are Everywhere

Search Engine Optimization and Web Commerce

- Schema.org used by >20% of Web sites (cf. https://queue.acm.org/detail.cfm?id=2857276)
- Major search engines exploit semantic descriptions (cf. schema.org; GoodRelations-ontology; RDFa; JSON-LD)

Pharma and Lifesciences

- Mature and comprehensive vocabularies and ontologies (eg. Gene Ontology describing genes and their functions)
- Billions of disease, drug, clinical trial descriptions

Digital Libraries

- Many established vocabularies (DublinCore, FRBR, EDM)
- Millions of aggregated from thousands of memory institutions on Europeana, German Digital Library

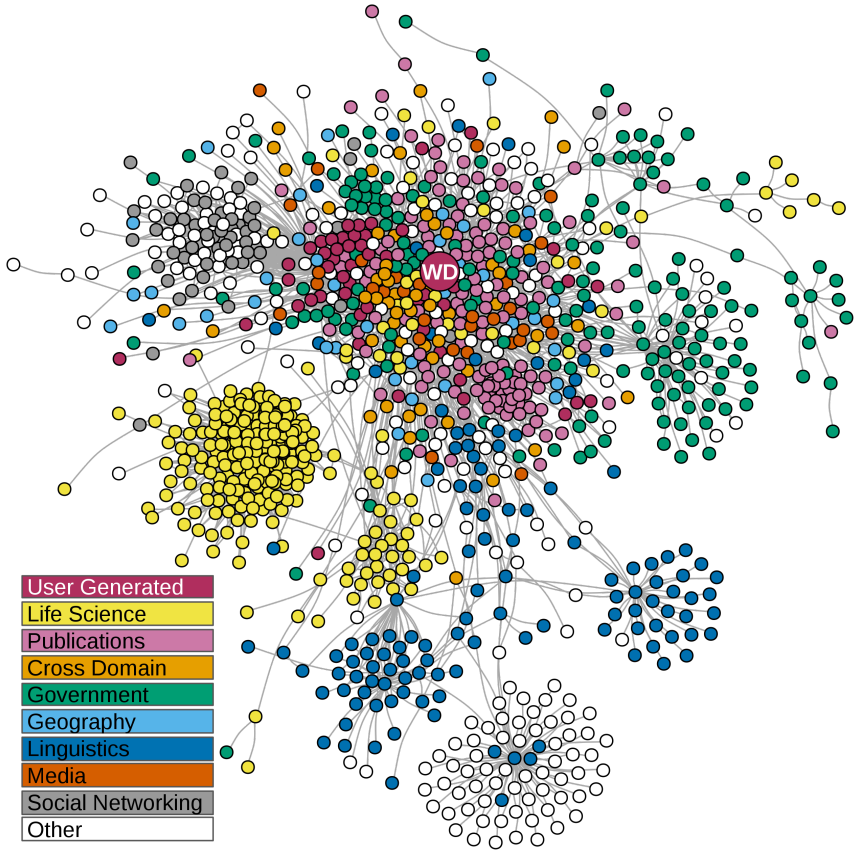
Source: https://www.slideshare.net/soeren1611/knowledge-graph-introduction (Slide 34)

The Impact of Knowledge Graphs is already visible

Key Application Areas for Knowledge Graphs

Three key applications accellerated the popularity of KGs:

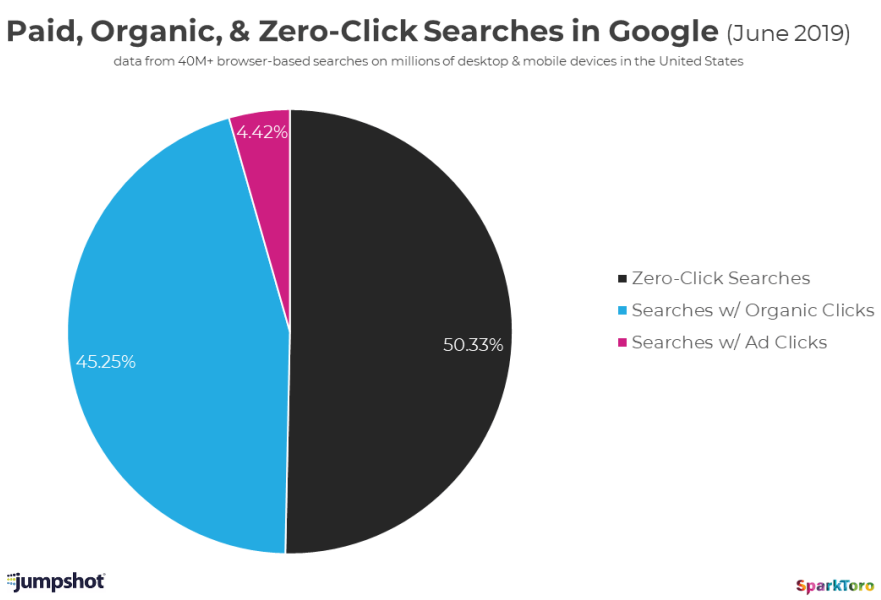
1. The integration and organization of information about known "entities", either as an openly accessible resource on the Web
2. The representation of input and output information for AI/ML algorithms.
3. The deployment in organisations or enterprises in order to
    i. improve **search** and **discovery** by explicit relationships
    ii. enable better **decision-making** through holistic views on data
    iii. enhance **customer experience** through personalized and connected information
    iv. facilitate **data integration** by generic, domain-independent KRFs
    v. support **knowledge management** by centralized KBs

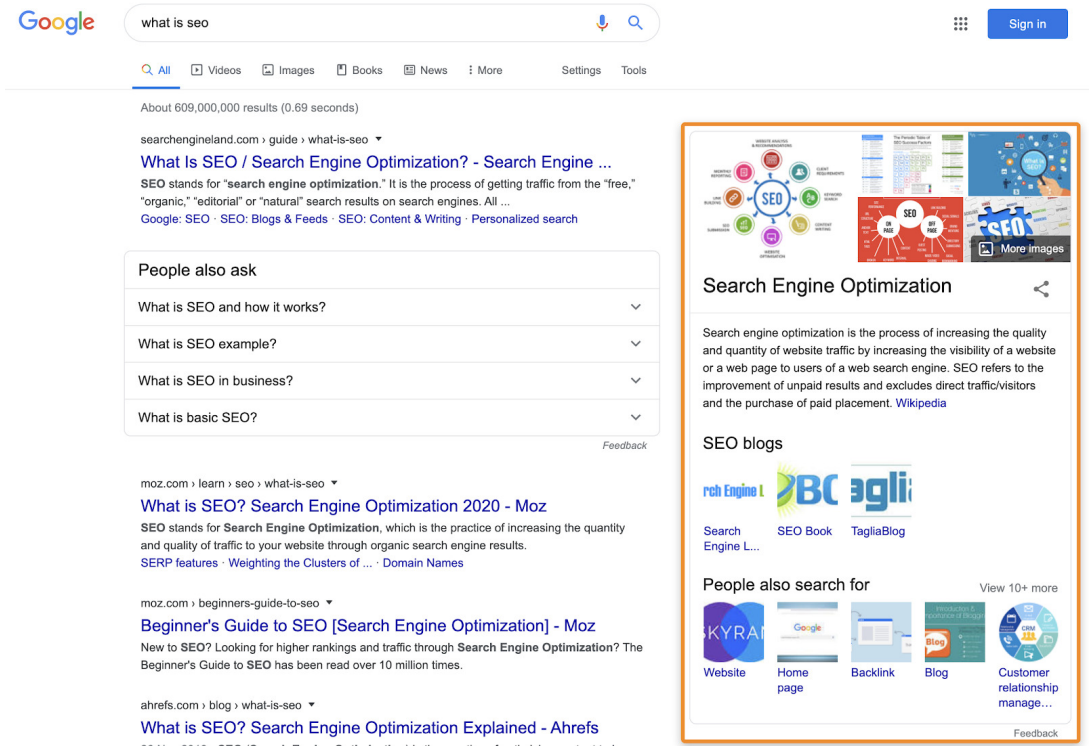


Source: Chaudhri, V. K., Baru, C., Chittar, N., Dong, X. L., Genesereth, M., Hendler, J., Kalyanpur, A., Lenat, D., Sequeda, J., Vrandečić, D., and Wang, K. 2022. "Knowledge graphs: Introduction, history, and perspectives.". AI Magazine 43: 17–29. https://doi.org/10.1002/aaai.12033

Knowledge Graphs influence Search Behaviour

Based on a recent study (see sources), more than **50% of Web searches** result in **no clicks**.





The Knowledge Graph helps Google in answering queries directly using a **knowldge panel**; ⇝ the query for "what is seo" on the right picture

Sources: https://ahrefs.com/blog/google-knowledge-graph/ and https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/

Knowledge Graphs allow for organizing open information – Wikidata

- **Wikidata** is a collaboratively edited open KG that provides data for Wikipedia and for other uses on the web.
- The Wikidata KG enhances/improves the quality of information in Wikipedia
- Semantic KGs might employ terms whose semantics are defined on the basis of logical theories
- The entailment regimes computed by inference engines allow for deriving relationships that are not inherently present
- When other open KGs use well-known vocabularies with commonly defined semantics, entities can be linked together (e.g. using schema.org)
- Multi-level annotations can be created and published

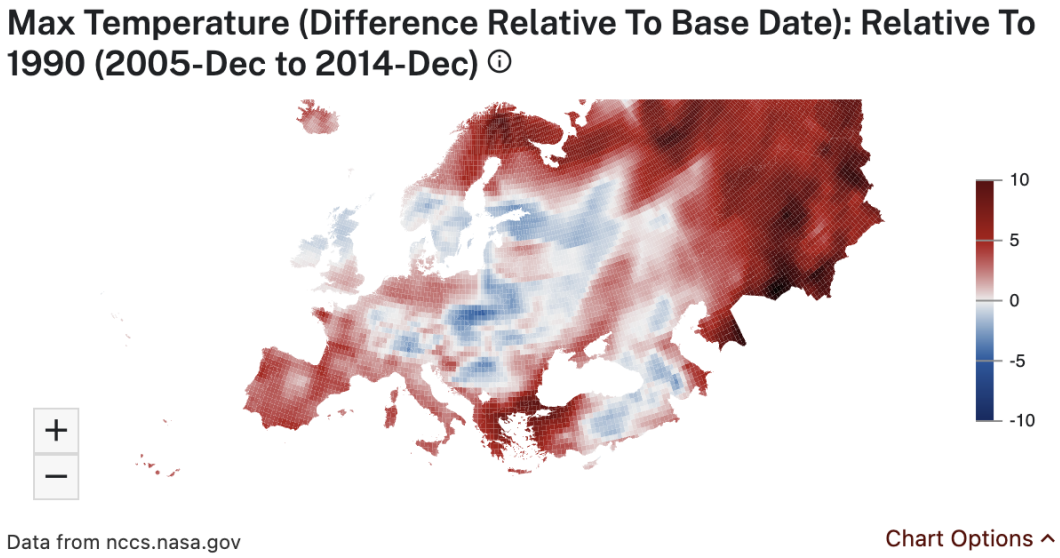
Example

Consider the Wikipedia page for the town "Winterthur", which includes a list of all of Winterthur's twin towns: two are in Switzerland, one in the Czech Republic, and one in Austria. Wikipedia also has an entry for the city, Ontario, in California, which lists Winterthur as its sister city. The "sister city" and "twin city" relationships are meant to be identical as well as reciprocal. Thus, if a city A is a sister (twin) of another city B, then B must be a sister (twin) of A. In Wikipedia, "Sister cities" and "Twin towns" are simply section headings without any relationship/linkage specified between the two. Therefore, it is difficult to detect this discrepancy automatically. In contrast, the Wikidata representation of Winterthur includes a relationship called **twinned administrative body**, which includes the city of Ontario, CA. As this relationship is defined to be a **symmetrical** relationship in the KG, a SPARQL query engine can infer that the Wikidata page for the city of Ontario, CA is to be linked to the Wikidata page of Winterthur.

Datacommons interlink many open and freely accessible data sources

https://datacommons.org/ is another publicly available large KG that incorporates data from many different, mostly governmental sources and authorities such as

- demographics (US Census, Eurostat)
- economics (World Bank, Bureau of Labor Statistics, Bureau of Economic Analysis)
- health (World Health Organization, Center for Disease Control)
- climate (Intergovernmental Panel on Climate Change, National Oceanic and Atmospheric Administration)
- sustainability (U.S. National Renewable Energy Laboratory (NREL), RE Atlas, etc.)



Data analysis about COVID-19 affection indicators

COVID-19

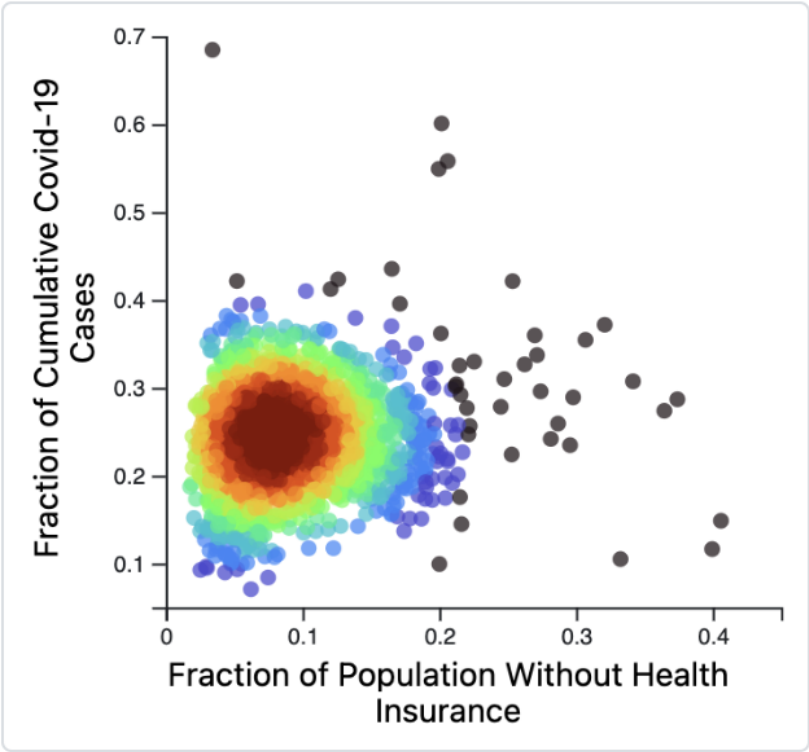
As many insightful articles from the New York Times and others pointed out, Covid-19 affected African American communities much more. Unfortunately, Covid-19 prevalence is correlated with many other indicators.

For example, we see that Covid-19 infection rates are highly correlated

- with the fraction of the population that is uninsured,
- with the fraction of the population in poverty,
- with the fraction of the population on food stamps, etc.

Of course, these are just correlations.

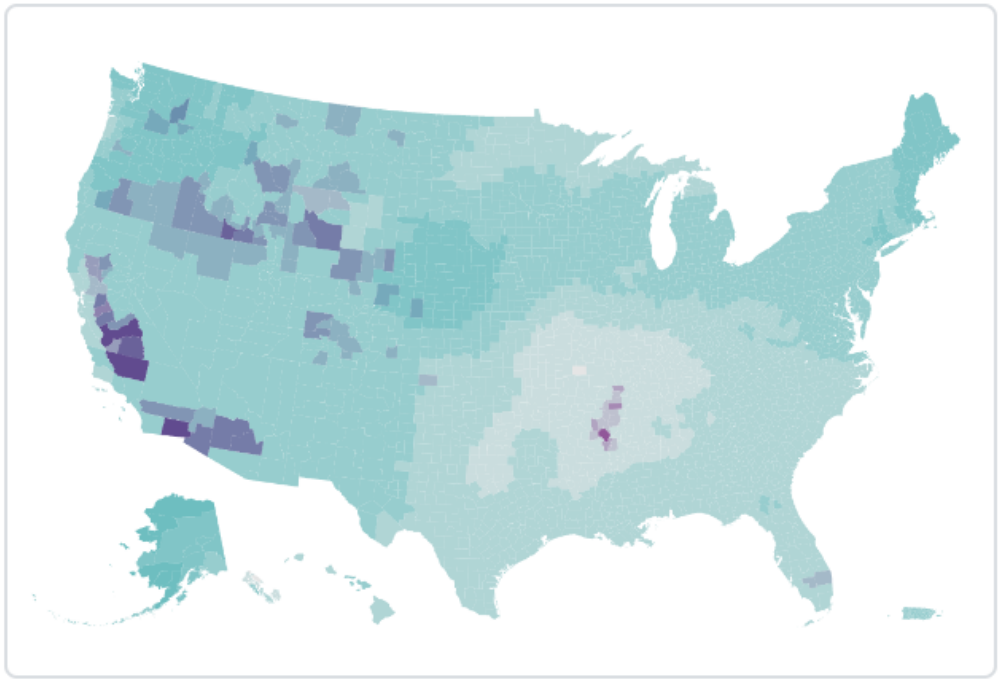
By performing a causal analysis, it is possible to discover the most variables most causally predictive of Covid-19 occurrence and morbidity.



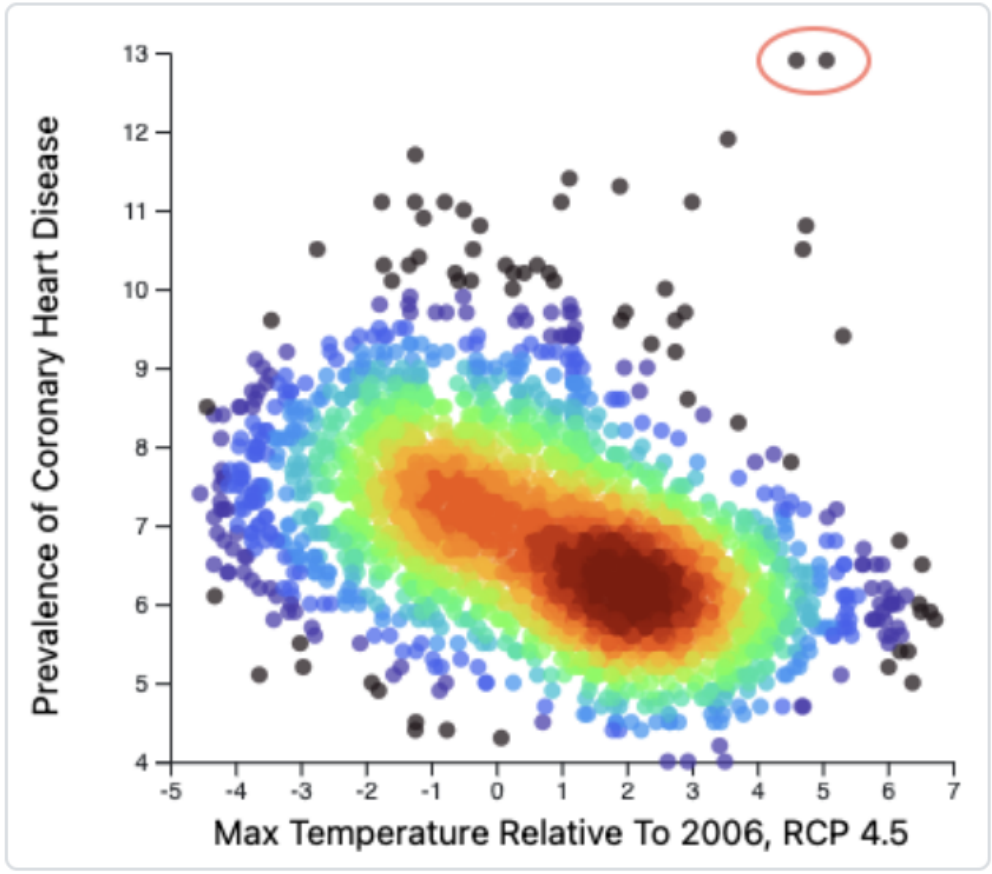
Fraction of positive Covid-19 cases vs. Fraction of uninsured across US counties (source: US Census, New York Times)

Source: Datacommons webpage

Visualisations about environmental issues from the datacommons KG



Water withdrawal for irrigation vs. projected temperature rise across US counties (source: USGS, NASA)



Heart condition vs. max projected summer temperature for US counties (RCP 4.5) (source: CDC, NASA)

Knowledge Graphs in Organisations

Data Silos – many valuable Information is still locked in proprietary data formats and systems

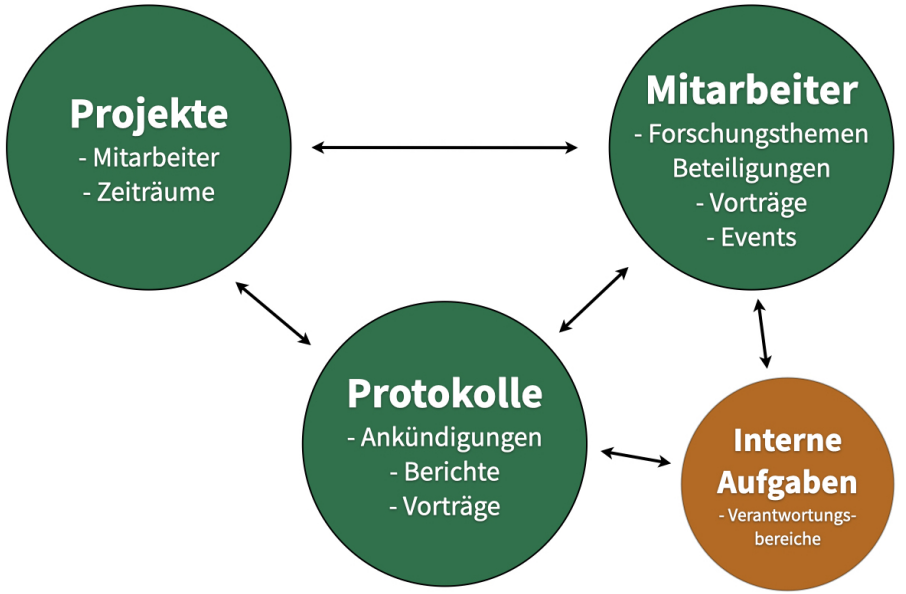
Picture source: ...

Knowledge Graphs support the Organization of Enterprise Information

Data integration is essential to the functioning of modern enterprises where corporate data typically reside across many distinct databases and unstructured sources.

Examples:

- Companies can create a "360-degree-view" of its customers by aggregating data from different external sources
- Companies can track user behaviour and relate it to internal data for more effective business operations



The data integration process for creating the 360-degree view of a customer might begin with knowledge engineers working with business analysts to sketch out a schema of the key entities, events, and the relationships that they are interested in tracking.

The **meaning** of the data stored in enterprise databases is hidden in logic embedded in queries, data models, application code, written documentation, or simply in the minds of subject matter experts requiring both human and machine effort in the mapping process.

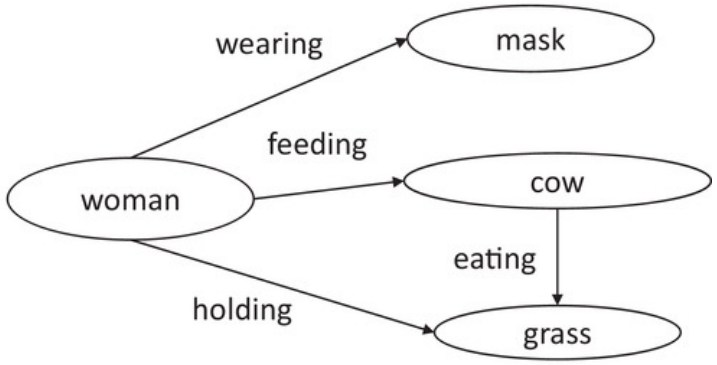
New and exciting aspects of using KGs for data integration

1. First, the integrated information may come from text and other unstructured sources (for example, news, social media, and others) as well as structured data sources (for example, relational databases). As many information extraction systems already output information in triples, using a generic schema of triples substantially **reduces the cost** of starting data integration projects.
2. Second, it can be easier to adapt a triple-based schema in response to **changes** than the comparable effort required to adapt a traditional relational database. This is because a relational system is typically modeled to support the application, and thus, schema changes often require database reorganization.
3. Lastly, modern KG engines are highly optimized for **answering questions** that require traversing the graph relationships in the data.

Due to the relative ease of creating and visualizing the **schema** and the availability of **built-in analytics operations**, KGs are becoming a popular solution for turning **data** into **intelligence** in organisations and enterprises.

Source: https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12033

Knowledge Graphs in Artificial Intelligence and Computer Vision



In CV, an image is represented as a set of objects with a set of properties, where each object corresponds to a bounding box, identified by an object detector.

- Objects are interconnected by a set of named relationships predicted by a model trained for identifying visual relationships.

A CV algorithm produces the KG (aka scene graph) shown to the left with objects such as a `woman`, a `cow`, and a `mask`, and relationships such as `holding`, `feeding` etc.

A scene graph became a central tool for achieving compositional behavior in CV algorithms.

- Once a CV algorithm is able to recognize certain objects, by leveraging scene graphs, it can be trained to **recognize any combination** of those objects with fewer examples.
- Scene graphs also provide the foundation for tasks such as **visual question answering**.

Source: https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12033

What are the Benefits of Knowledge Graphs

A **graph-based abstraction** of knowledge has numerous benefits in professional scenarios when compared with traditional data organisation systems such as the relational model or NoSQL alternatives.

- Graphs provide a concise and intuitive abstraction for a variety of domains, where edges capture the relations between the entities inherent in the domain.
- Graphs allow maintainers to postpone the definition of a schema, allowing the data and its scope to evolve in a more flexible manner than typically possible in a relational setting, particularly for capturing incomplete knowledge.
- Unlike (other) NoSQL models, specialised graph query languages support not only standard relational operators (joins, unions, projections, etc.), but also navigational operators for recursively finding entities connected through arbitrary-length paths.
- Standard knowledge representation formalisms – such as ontologies and rules – can be employed to define and reason about the semantics of the terms used to label and describe the nodes and edges in the graph.
- Scalable frameworks for graph analytics can be leveraged for computing centrality, clustering, summarisation, etc., in order to gain insights about the domain being described.
- Various representations have also been developed that support applying machine learning techniques directly over graphs.

✨ In summary, the decision to build and use a knowledge graph opens up a range of techniques that can be brought to bear for integrating and extracting value from diverse sources of data.