Knowledge Graphs – And why we need them

CS520 Guest Lecture

Chaitan Baru, UC San Diego / National Science Foundation

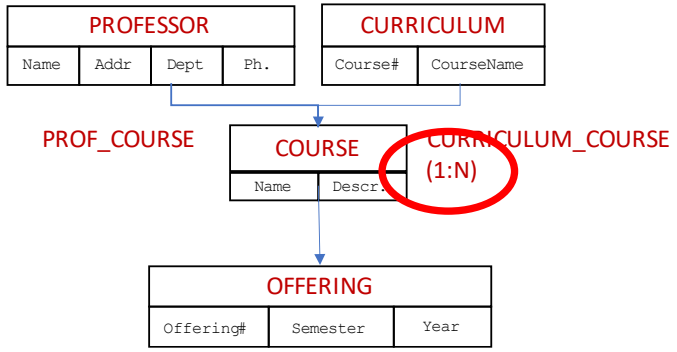
Let’s begin with Graphical Representations of data…for data management and processing

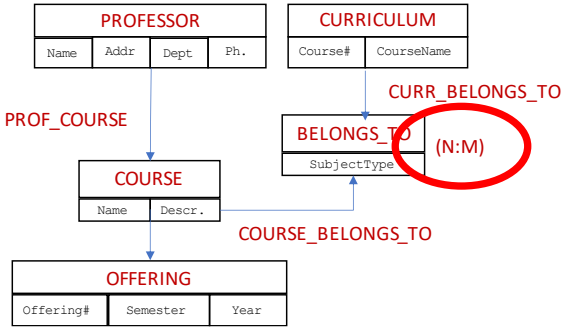
- Early 1970’s: The Network Model, CODASYL (COnference on DAta SYstems Languages DataBase Task Group)
  - *Bachman Diagrams*
  - Introduced the ideas of a Schema, Subschema, Data Definition Language, Data Manipulation Langu
  - “Record types” and one-many a ong Record Types
  - Mapping of structure to magnetic disk (2D to 1D)
  - Had to traverse data structures using operations like Next, Previous, FirstChild, Parent, etc.

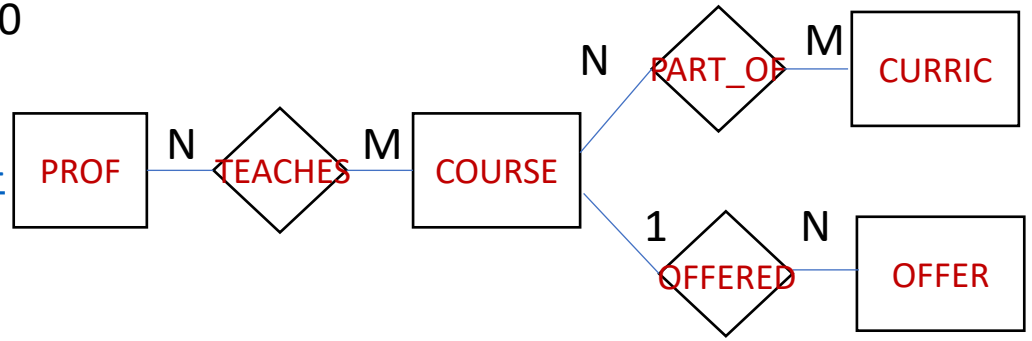
Focus on physical data structures and operations on access data on disk

- Mid-1970’s Entity-Relationship Diagrams
  - *The Entity-Relationship Model –* Peter P. Chen, ACM’s Transactio No 1, March 1976, https://dl.acm.org/d 20440
  - **A Short History of the ER Diagram and Information Modeling,** Shannon Kempe on September 25, 2012 https://www.dataversity.net/a-short-history-of-the-er-diagram-and-information-modeling/#

Focus on conceptual design

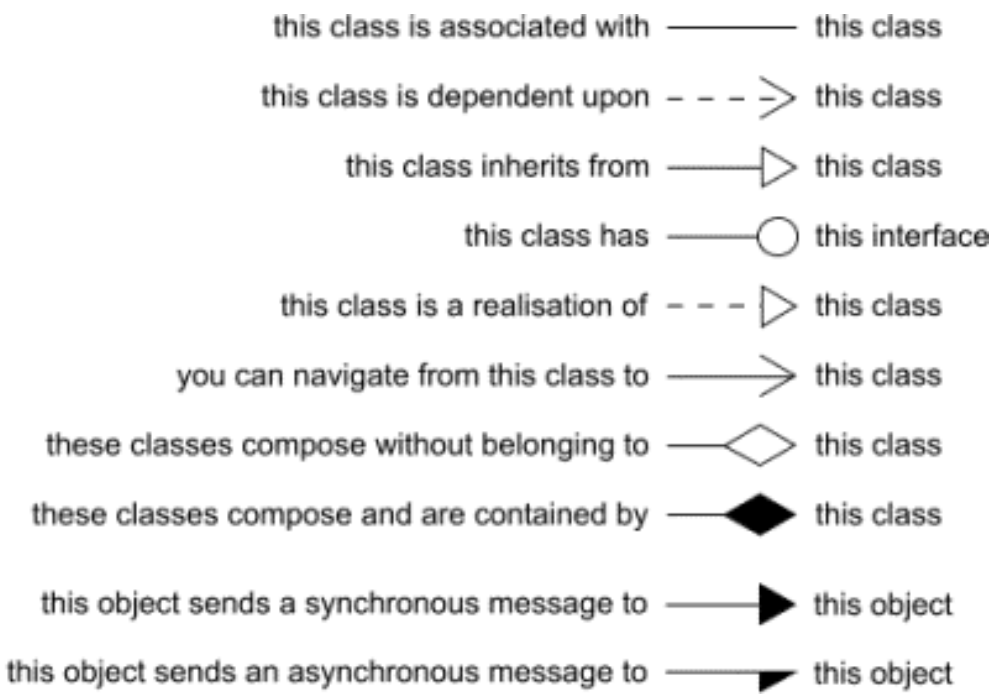






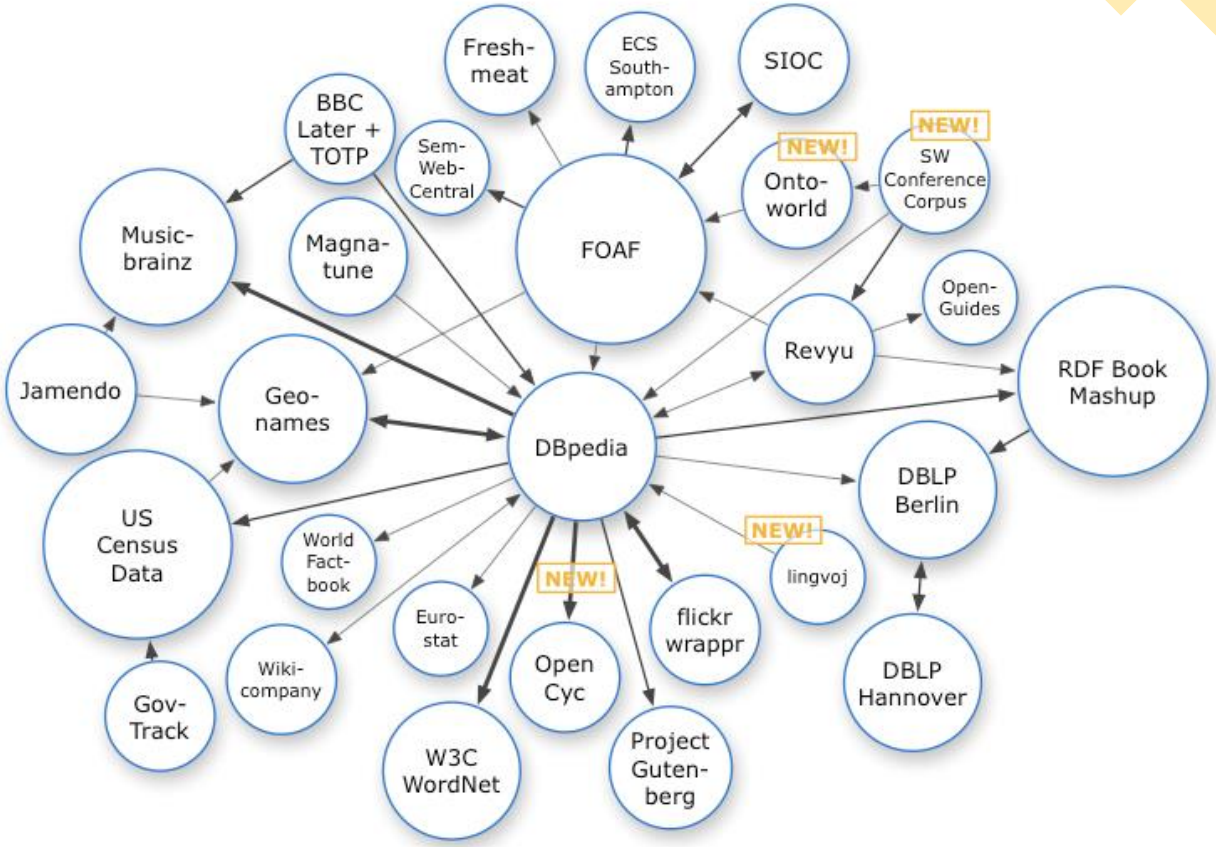
Object-oriented modeling

- Mid-1990s: Advent of Object-oriented design methodologies
  - *Unified Modeling Language* User Guide, Addison-Wesley 2005, ISBN 0321267974
- Rich variety of relationships, including inheritance, part-of, etc.

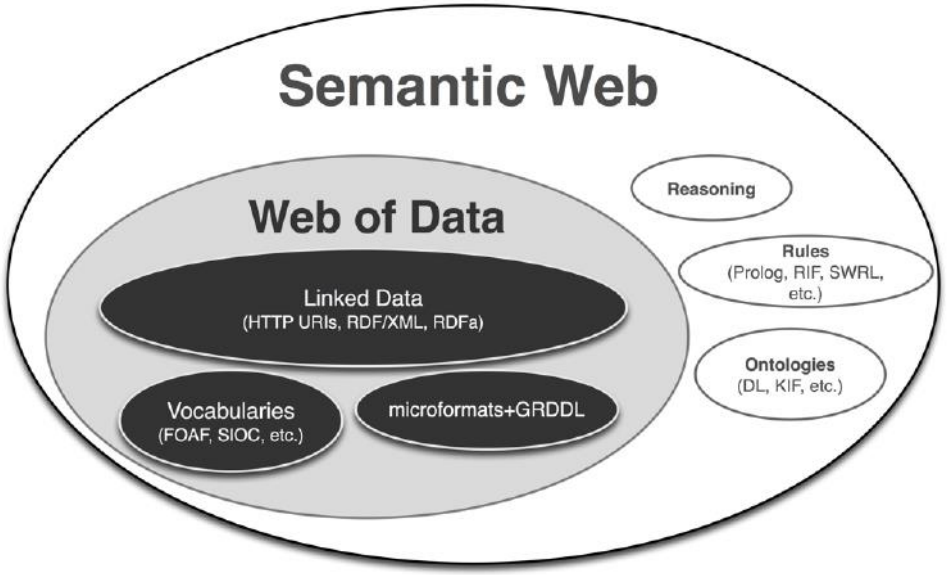


Fast Forward 30 years…to Now!

- Huge amounts of open, unstructured data on the Web (and structured data in the "Deep Web")
- Heterogeneous data – Real-world entities associated with a wide variety of information
    - **You** – as an individual
    - A **disease** and all of its relationships
    - A **geographic location**, e.g., your home
    - Ecological information for a **region**
    - …
- The need to search all of this data and "integrate" data (e.g., Google/Bing search)
- Knowledge Representation and Querying Systems
    - OWL, SPARQL
- Computational power
    - Big data – BigQuery, BigTable
    - Graph databases

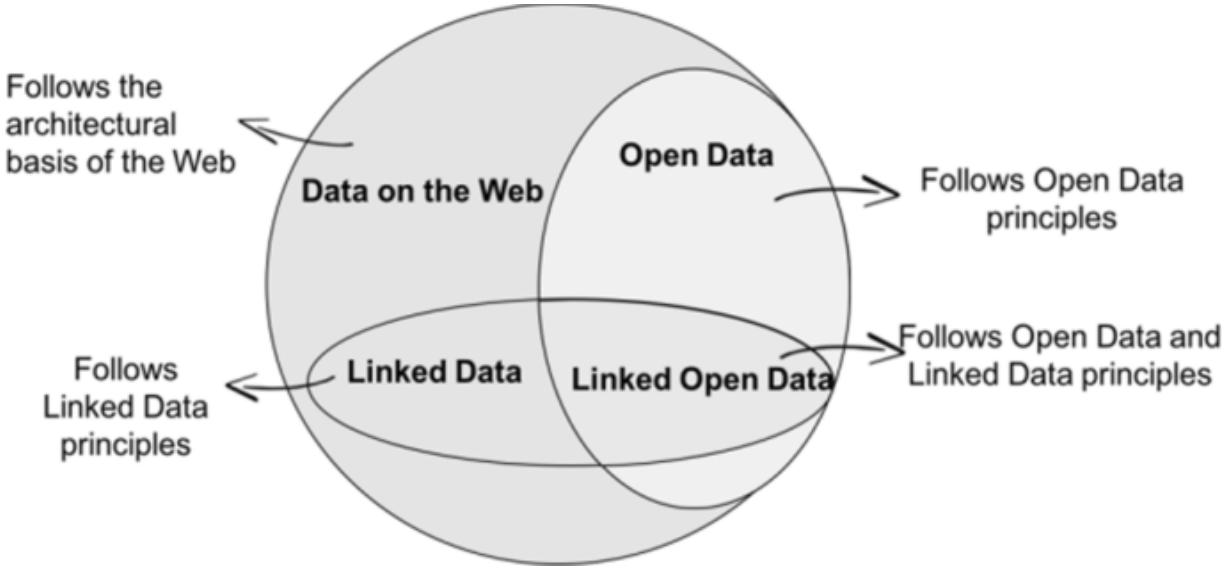


Semantic Web, Linked Open Data



- Semantic Web: Web data + Ontologies + Reasoning using the Web Ontology Language (OWL) (web of data, 2018)

- Linked Open Data: Web-based structured data, interlinked with other data so it becomes more useful through *semantic queries*, using RDF etc.



Emerging services: E.g., data.world for structured data

Can your data catalog do *this*?

All your knowledge

data.world gives you complete context, so you actually *understand* the data, whether it's in the cloud or on-prem. This includes metadata, dashboards, analysis, code, docs, project management, and social collaboration features.

Born in the cloud

We've been cloud-native from day one. Our multi-tenant offerings are highly available, scale bigger, perform better, and evolve faster. And as a SaaS company, we provide open and transparent pricing.

Agile innovation

data.world is the only enterprise data catalog with a continuous release cycle. That means our platform is always getting better, and your data tools are never out of date. To access the latest features and capabilities, just refresh your browser.

Powered by a knowledge graph

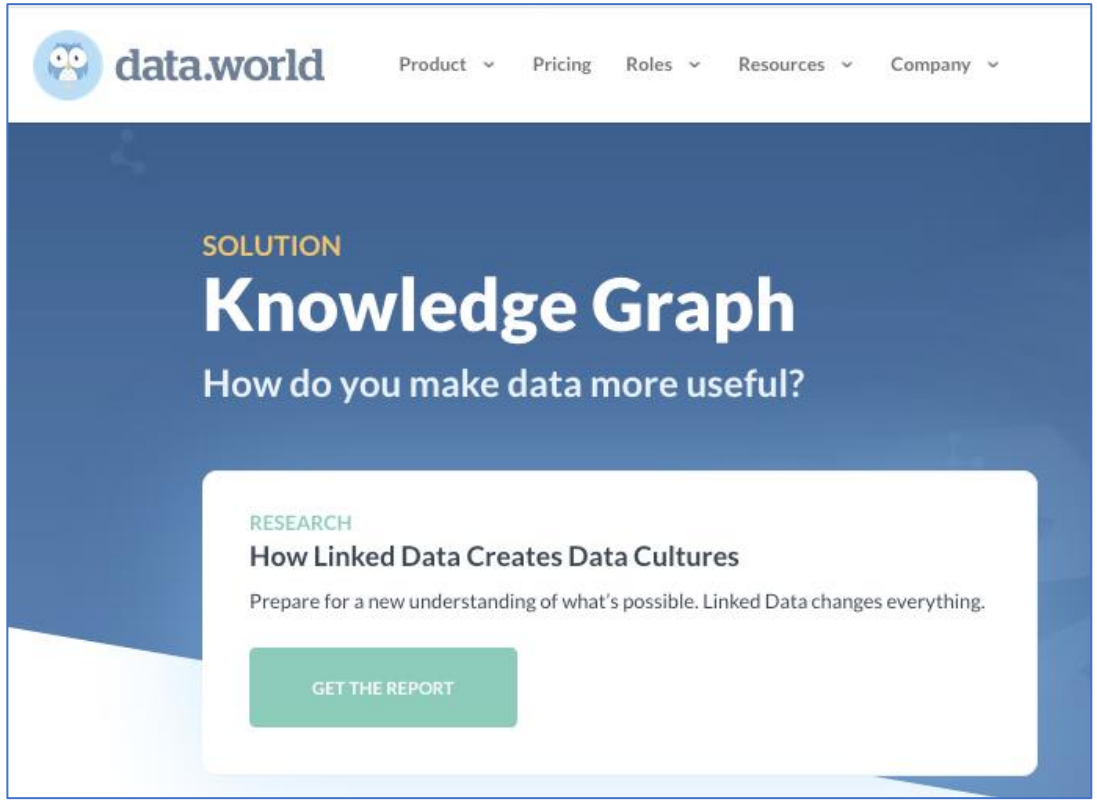
data.world automatically builds a connected web of data and insights so you can explore the relationships within. Get recommendations on related assets to enrich your analysis. The more you use data.world's patented knowledge graph technology, the smarter you and your data get.

Real-time integration

Deploy faster and extend your capabilities farther with data.world's growing array of one-click, pre-built integrations, connectors, and APIs.

One-of-a-kind expertise

data.world built the world's largest open data community. Imagine what we've learned from hundreds of thousands of users, datasets, and interactions. This deep and unrivaled knowledge informs everything we do for enterprises, too.



Knowledge

a. Wikipedia (and Webster) definitions of “knowledge”
   - A familiarity, awareness, or understanding of someone or something, such as facts, skills, or objects. … knowledge can be acquired in many different ways and from many sources … perception, reason, memory, testimony, scientific inquiry, education, and practice.

b. Scientific Knowledge
   - To be termed scientific, a method of inquiry must be based on gathering observable and measurable evidence subject to specific principles of reasoning and experimentation. The scientific method consists of the collection of data through observation and experimentation, and the formulation and testing of hypotheses.

c. (Scientific) Knowledge must be “usefully available”
   - The system should apparently be dynamic and self-organizing (unlike a mere book on its own).
   - The knowledge must constitute some sort of representation of "the outside world", or ways of dealing with it (directly or indirectly).
   - Some way must exist for the system to access this information quickly enough for it to be useful.

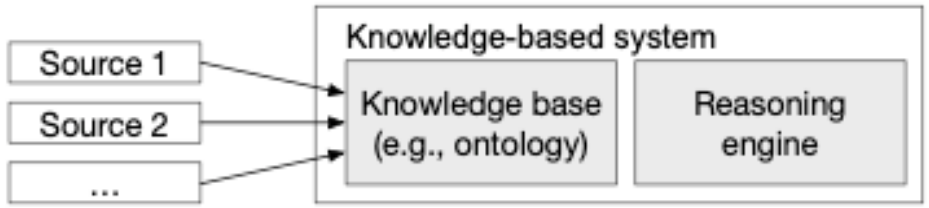
What is the “knowledge” in Knowledge Graphs?

- Focus (for now…) is not on knowledge in the epistemological sense—i.e., what do we *know?* And, representing what we know.
- Instead, on a data-driven approach. Represent the data I have—about “entities” (in the real world) and their relationship to other entities
- This graph of entities and their relationships is the start of the knowledge graph

Towards a Definition of Knowledge Graphs

- Lisa Ehrlinger and Wolfram, Poster, Semantics 2016, September 2016, Leipzig, https://2016.semantics.cc/posters-and-demos-madness.

Definition	Source
"A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for pot don	Paulheim [16]
"Kn rela	[12]
"Kn to a rela	3]
"W trip subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$."	
"[...] systems exist, [...], which use a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph."	Pujara et al. [17]

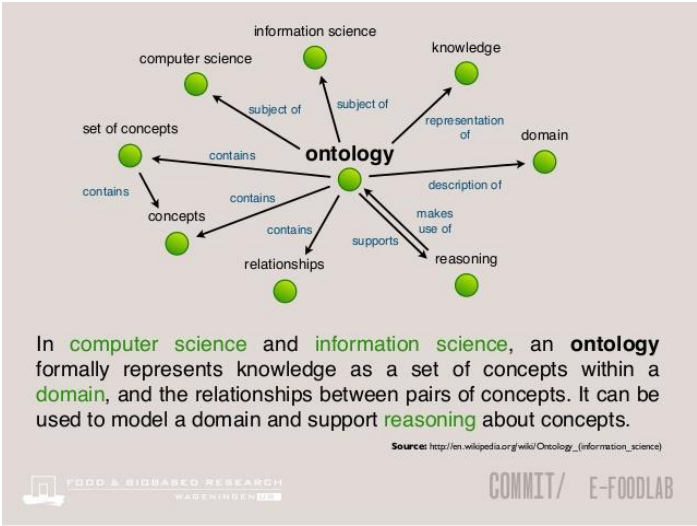


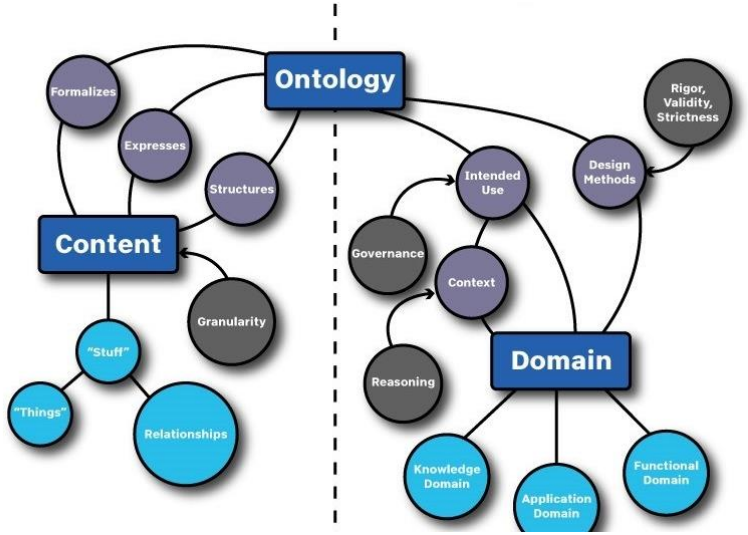
A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.

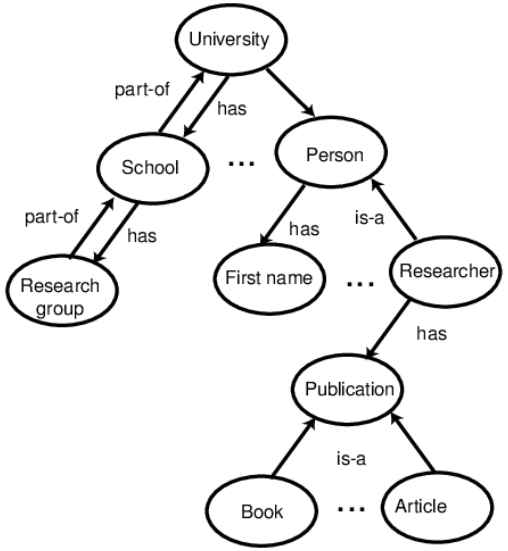
Figure 1: Architecture of a knowledge graph

Table 1: Selected definitions of knowledge graph

Ontologies…

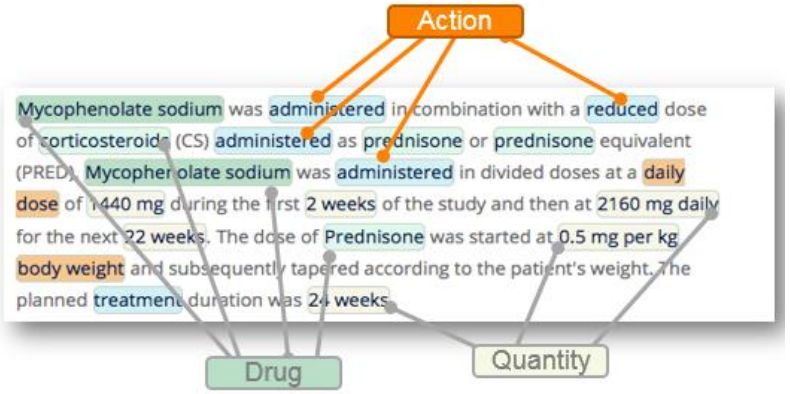






From natural language to a formal logic system
--The Resource Description Framework (RDF) triple model: A set of three entities codifying a statement in the form subject–predicate–object expressions, e.g., "Bob is 35", or "Bob knows John")

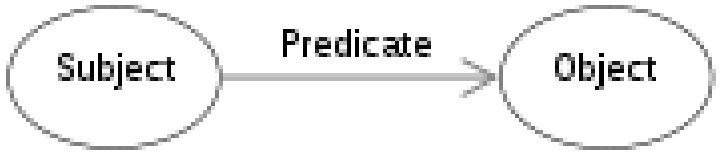
- Where do ontologies come from? Designing versus *deriving* ontologies



Where do the entities come from? Structured and Unstructured data…

- Structured databases
  - Corporate data, many scientific databases
- Vast majority of data on the web and in science domains are unstructured
    - Images, text, video, signals (waveforms)
- Representation using the RDF model

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery



Prof. Chaudhri Teaches KnowledgeGraphs

RDF schema: Schema.org

Welcome to Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts. It is in this spirit that the founders, together with the larger community have come together - to provide a shared collection of schemas.

We invite you to get started!

View our blog at blog.schema.org or see release history for version 12.0.

Types:

Close hierarchy / Open hierarchy

- Thing -
  - Action +
  - CreativeWork +
  - Event +
  - Intangible +
  - MedicalEntity +
  - Organization +
  - Person +
  - Place +
  - Product +

DataTypes:

Close hierarchy / Open hierarchy

- DataType -
  - Boolean +
  - Date
  - DateTime
  - Number +
  - Text +
  - Time

Schema.org

SpecialAnnouncement

A Schema.org Type

This term is proposed for full integration into Schema.org, pe

Thing > CreativeWork > SpecialAnnouncement

Reacting to COVID-19 data on the Web

schema blog

Official blog for schema.org

MONDAY, MARCH 16, 2020

Schema for Coronavirus special a
Testing Facilities and more

The COVID-19 pandemic is causing a large number of
changes in schedules and other aspects of everyday li
rescheduling of events but also new availability of med

We have today published Schema.org 7.0, which inclu
global response to the Coronavirus outbreak.

It includes a "SpecialAnnouncement" type that provides
well as markup to associate the announcement with a
and to indicate URLs for various kinds of update such
quarantine guidelines, travel bans, and information abo

Many new testing facilities are being rapidly establishe
now has a CovidTestingFacility type to represent these
established medical facilities or temporary adaptations

We are also making improvements to other areas of Sc
to working online and working from home, for example
event has moved from having a physical location to bei
whether the event's "eventAttendanceMode" is online, c

7.0 2020-03-17	Version 7.0. See planning pages.	Vocabulary

The basic content of SpecialAnnouncement is similar to that of an RSS or Atom feed. For publishers without such feeds, basic feed-li
shared by posting SpecialAnnouncement updates in a page, e.g. using JSON-LD. For sites with Atom/RSS functionality, you can poi
webFeed property. This can be a simple URL, or an inline DataFeed object, with encodingFormat providing media type information e
"application/rss+xml" or "application/atom+xml".

Property	Expected Type	Description
Properties from SpecialAnnouncement		
announcementLocation	CivicStructure or LocalBusiness	Indicates a specific CivicStructure or LocalBusiness associated w SpecialAnnouncement. For example, a specific testing facility or t opening hours. For a larger geographic region like a quarantine o use spatialCoverage.
category	PhysicalActivityCategory or Text or Thing or URL	A category for the item. Greater signs or slashes can be used to i category hierarchy.
datePosted	Date or DateTime	Publication date of an online listing.
diseasePreventionInfo	URL or WebContent	Information about disease prevention.
diseaseSpreadStatistics	Dataset or Observation or URL or WebContent	Statistical information about the spread of a disease, either as We described directly as a Dataset, or the specific Observations in th WebContent URL is provided, the page indicated might also conta markup.
gettingTestedInfo	URL or WebContent	Information about getting tested (for a MedicalCondition), e.g. in t pandemic.
governmentBenefitsInfo	GovernmentService	governmentBenefitsInfo provides information about government b with a SpecialAnnouncement.
newsUpdatesAndGuidelines	URL or WebContent	Indicates a page with news updates and guidelines. This could of required to be) the main page containing SpecialAnnouncement r
publicTransportClosuresInfo	URL or WebContent	Information about public transport closures.
quarantineGuidelines	URL or WebContent	Guidelines about quarantine rules, e.g. in the context of a panden
schoolClosuresInfo	URL or WebContent	Information about school closures.
travelBans	URL or WebContent	Information about travel bans, e.g. in the context of a pandemic.
webFeed	DataFeed or URL	The URL for a feed, e.g. associated with a podcast series, blog, o stamped updates. This is usually RSS or Atom.

emoved several largely un-used medical health properties whose
appropriately general: *action, background, cause, cost, function,*
n, outcome, overview, phase, population, purpose, source, subtype.
not remove terms casually, but in the current case the usability
f keeping them in the system outweighed the benefits of retaining
agged as archived/superseded.
ntroduced a VirtualLocation type, to support description of Events
is online. Added eventAttendanceMode to clarify the current
de of an event (online, offline or a mix). Also added a new
lue for eventStatus for Event: EventMovedOnline.

section:

dded (fast track) a SpecialAnnouncement type with several
supporting documentation, as a contribution to the global response
9 Coronavirus pandemic. We expect to iterate on this design after
and publisher feedback.
dded properties to EducationalOccupationalProgram to support cases
clear the distinction between classroom-based and work-based

dded petsAllowed for use with ApartmentComplex, and petsAllowed,
and floorSize to FloorPlan. Created new properties for
oms and a new tourBookingPage.
dded *draft* type MediaReview, and associated
ionRatingEnumeration with two example values AuthenticContent
text. Note that this is Editorial work in progress and not a complete
o NiemanLab background article on this work.)
dded usageInfo, and acquireLicensePage to CreativeWork.

redirect to https for http requests.
mplemented in sdopythonapp submodule Issue #6: Enhancements
rt, providing appropriate responses to HTTP OPTIONS request.

SPARQL: Query language for RDF

- A simple query example using the FOAF, Friends Of A Friend, ontology definition.
- Returns *names* and *emails* of every *person* in the dataset

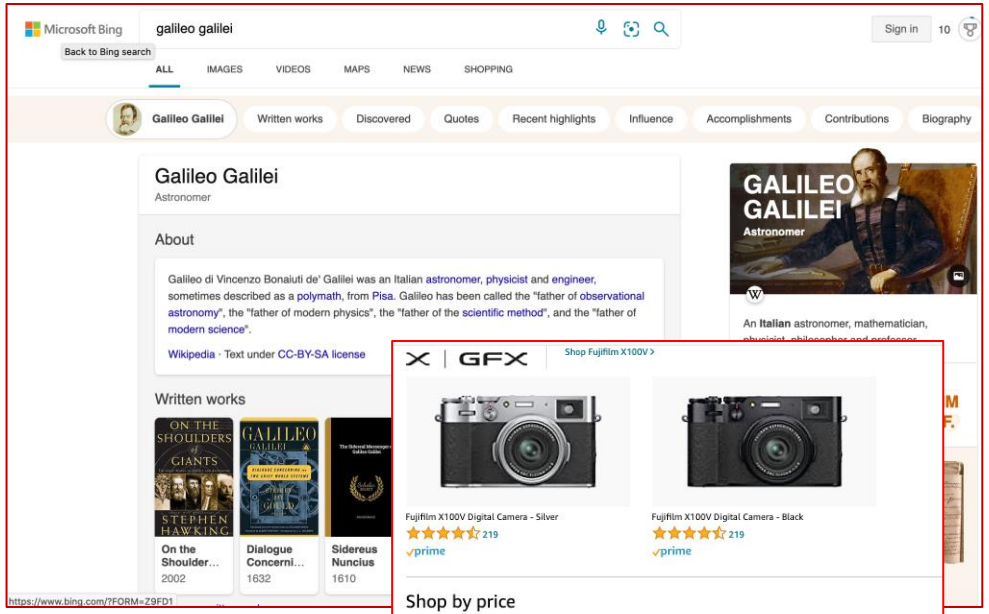
```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
       ?email
WHERE
  {
    ?person  a          foaf:Person .
    ?person  foaf:name  ?name .
    ?person  foaf:mbox  ?email .
  }
```

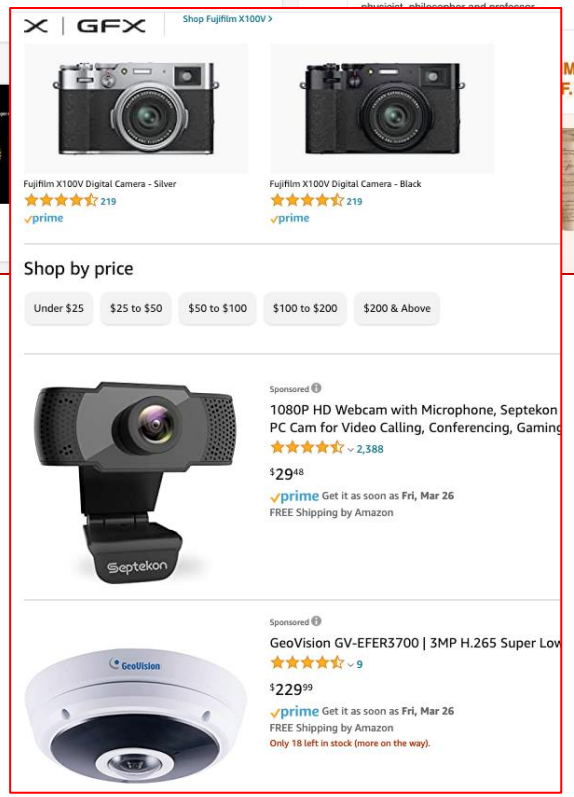
- An example query showing *country capitals* in Africa (using a fictional ontology)

```
PREFIX ex: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x ex:cityname ?capital ;
     ex:isCapitalOf ?y .
  ?y ex:countryname ?country ;
     ex:isInContinent ex:Africa .
}
```

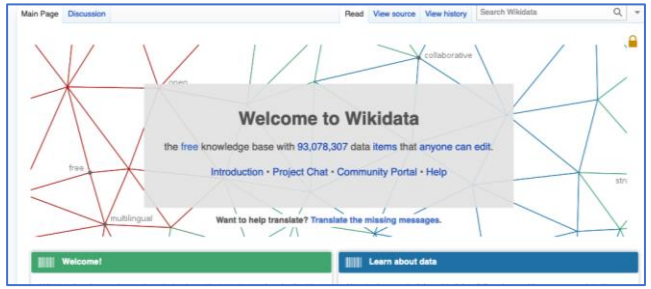
Why Knowledge Graphs? The use cases

- Searching the Web – Google, Bing, Maps
- Product information for online shopping – Amazon, …all retailers
- Smart Assistants – Siri, Cortana, Echo, …
- In science applications:
  - Search, discovery, data exploration
  - Finding connections (“integrating”) among heterogeneous data
  - Enabling analysis
  - Connecting data with (data-driven) models, e.g., ML models





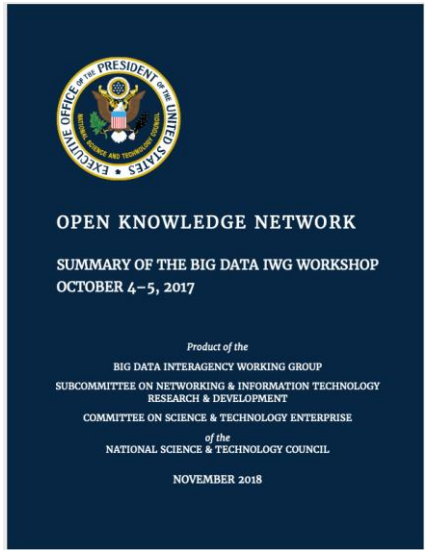
Harvesting entities



- 2007: Freebase—a large collaborative knowledge base
  - Knowledge graph containing data harvested from variety of sources, e.g., Wikipedia, Notable Names Database, Fashion Model Directory and MusicBrainz
  - Entity-Relationship, triple model of representation
- 2010: Google acquires Freebase
- 2012: Wikidata – a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation. A common source of open data used by Wikimedia projects such as Wikipedia and others, under the CC0 public domain license.
- 2015: Google announces Knowledge Graph API, using standard schema.org types, compliant with the JSON-LD (JavaScript Object Notation for Linked Data)
- 2018: Challenges and Innovations in Building a Product Knowledge Graph, Luna Dong, Amazon, Applied Data Science Talk KDD 2018, http://www.kdd.org/kdd2018/applied-data-science-invited-talks/view/luna-dong
- Similarly, Apple, Microsoft Cortana, …

Open Knowledge Network

- Interagency Federal Government Workshop, Oct 2017
- “An open and broad community effort to develop a national-scale data infrastructure—an Open Knowledge Network—would distribute the development expense, be accessible to a broad group of stakeholders, and be domain-agnostic. This infrastructure has the potential to drive innovation across medicine, science, engineering, and finance, and achieve a new round of explosive scientific and economic growth not seen since the adoption of the Internet.”



https://www.nitrd.gov/nitrdgroups/index.php?title=Open_Knowledge_Network

NSF Convergence Accelerator
Track A: Open Knowledge Network

- Creating OKN requires collaboration among KG technical experts and science domain experts
- Data may come from highly-curated scientific databases, as well as crowdsourcing
- Need to educate the next generation of scientists in tools to author and query and use KG/KNs.

Urba
(urban r

ntent
on EventS
ds)

Knowledge Network Programming System:
An IDE for Knowledge Graphs

Services for enriching data with geographic context

NSF Convergence Accelerator Model

Ideation Process

- Begins with a released RFI to solicit ideas
- Selected ideas will be invited to submit a workshop proposal
- Workshops are created to further the ideas
- Final workshop results/reports assist NSF in selecting future track topic(s)

Convergence Research Focus

- Chosen track topics are released in a program solicitation
- Selected awards focus on accelerating basic research into practice
- Leverage multidisciplinary approach, partnerships

PHASE I

- Funding up to $750K
- Nine months planning phase
- Teams participate in innovation curriculum
- Pitch comp
and proposal
evaluation

PHASE II

- Funding up to $5M
- 24 months solution development phase
- Focused on impact deliverables (e.g., otype, sustainability plan)

Innovation curriculum

Societal Impact

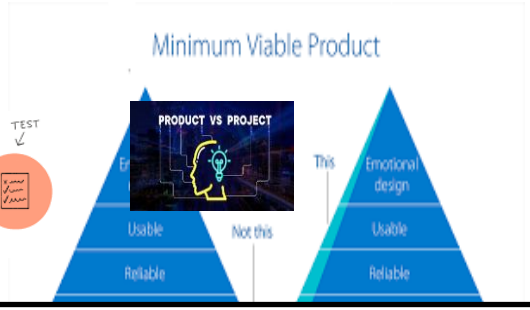
- National-scale
- Sustainability beyond NSF support may include:
  - Follow-on funding and investment
  - Integration into existing systems
  - Tools/knowledge accessible thru open source
  - Solutions expanded into new markets

19

NSF Convergence Accelerator Curriculum

- User Discovery
- Human-centered Design

- Prototyping



- Coaching/Mentorship
- Financials
- Communication Skills
- Pitching
- Public Expo

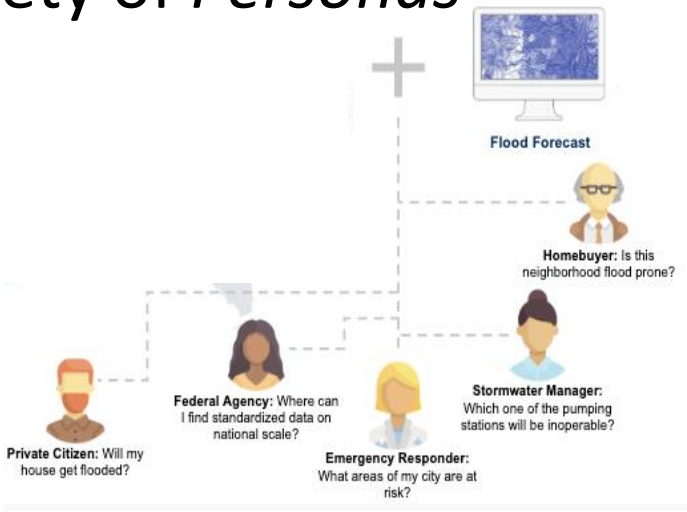
Essential for developing Open Knowledge Networksl

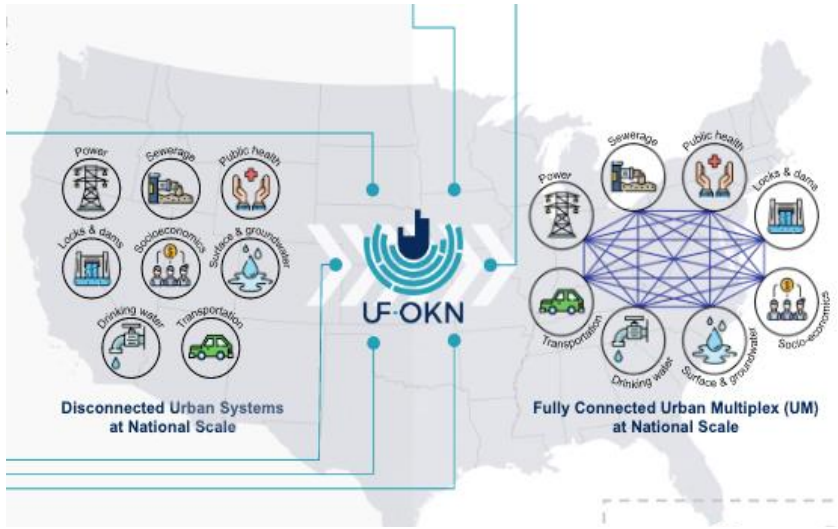
- Team Science

TOOLBOX
DIALOGUE INITIATIVE ™

Urban Flooding Open Knowledge Network
UF-OKN, PI: Prof. Lilit Yeghizarian, Univ of Cincinnati

- Link together all information in an urban multiplex
- Connect flood-related data and models
- Serve a variety of *Personas*

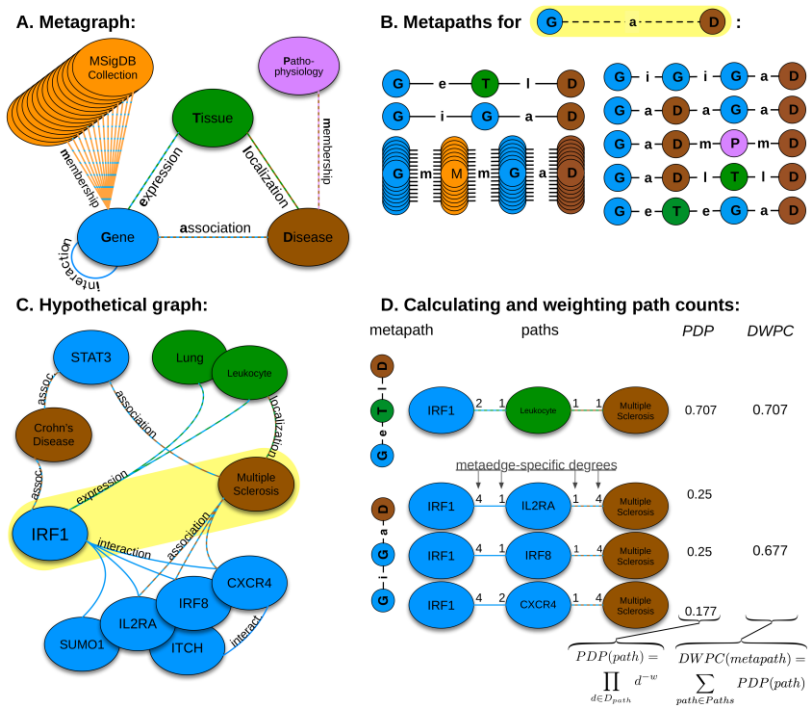


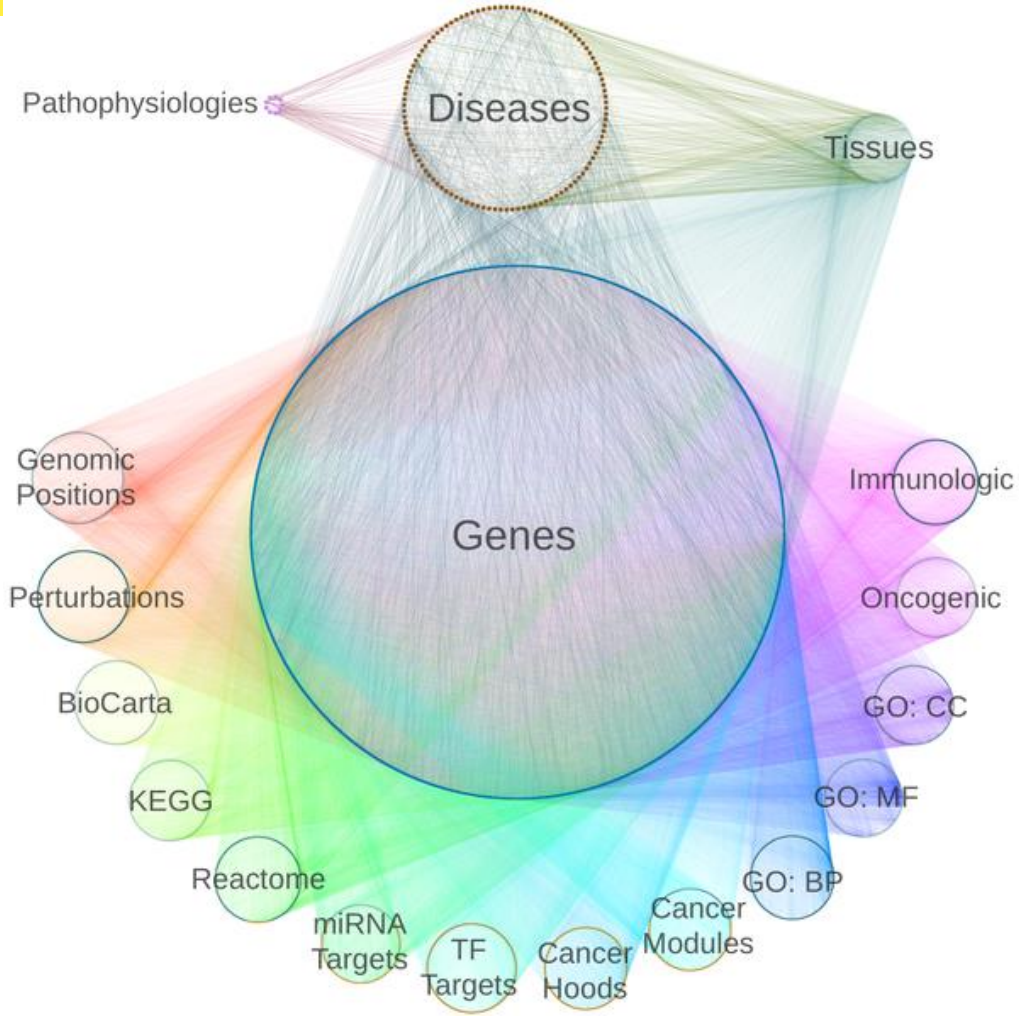


- Dealing with uncertainty in flood models
  - 1000's of local/regional flood models across the country

SPOKE: Scalable Precision Medicine Oriented Knowledge Engine, PI: Prof. Sergio Baranzini, UCSF

- Heterogeneous network integrating diverse information domains
- Metanodes and metaedges
- Graphical query interfaces





Himmelstein DS, Baranzini SE (2015) Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLOS Computational Biology 11(7): e1004259. https://doi.org/10.1371/journal.pcbi.1004259
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004259

SPOKE references

- Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes, Daniel S. Himmelstein, Sergio E. Baranzini, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004259.
- Knowledge Network Embedding of Transcriptomic Data from Space-flown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases, Charlotte A. Nelson, Ana Uriarte Acuna,, Amber M. Paul, Ryan T. Scott, Atul J. Butte, Egle Cekanaviciute, Sergio E. Baranzini, and Sylvain V. Costes, Life2021,11, https://doi.org/10.3390/life11010042.
- Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings, Charlotte A. Nelson, Atul J. Butte & Sergio E. Baranzini, Nature Communications, (2019) 10:3045, https://doi.org/10.1038/s41467-019-11069-0.
- Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini, Life. 2017, https://doi.org/10.7554/eLife.26726

SCALES–OKN: Systematic Content Analysis of Litigation EventS OKN, PI: Prof. Luis Amaral, Northwestern

- Extracting knowledge graphs from pdf text
- All court records are currently locked in pdf's behind a paywall (PACER)
- Impossible to perform simple queries, e.g.,
  - How many cases in a jurisdiction have self-representation; and how many of those are successfully litigated for the defendant?
  - In what types of cases are court fees waived? And what is the ethnicity of the defendant?
- Inferring document type from document structure (physical layout of pdfs)
- Developing ontologies for:
  - Legal docs, events – order, notice, report; agreement, brief, motion, notice;
  - Stages of a case—lifecycle; and how cases end

KNPS: Knowledge Graph Programming System, PI: Prof. Michael Cafarella, Michigan / MIT

- A programming language and toolset for:
  - Building KG-driven software;
  - Debugging data quality problems quickly and efficiently
- "Social" Knowledge Networks
  - How KNs like W
  - Data sharing a
- Supporting prov
- Tools using ML
  - Ingest web dat
  - Quickly create
  - Extract knowle ayout

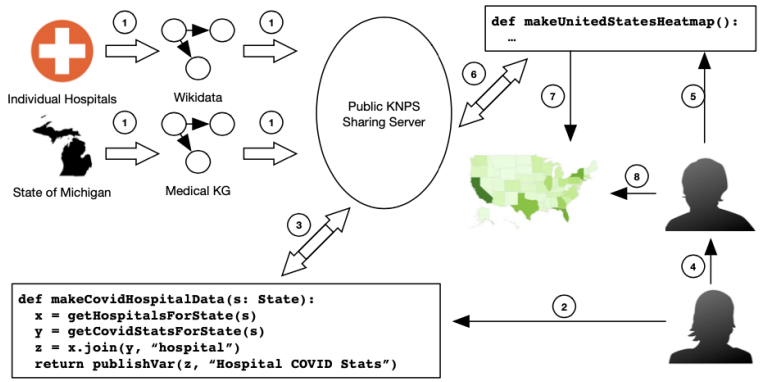
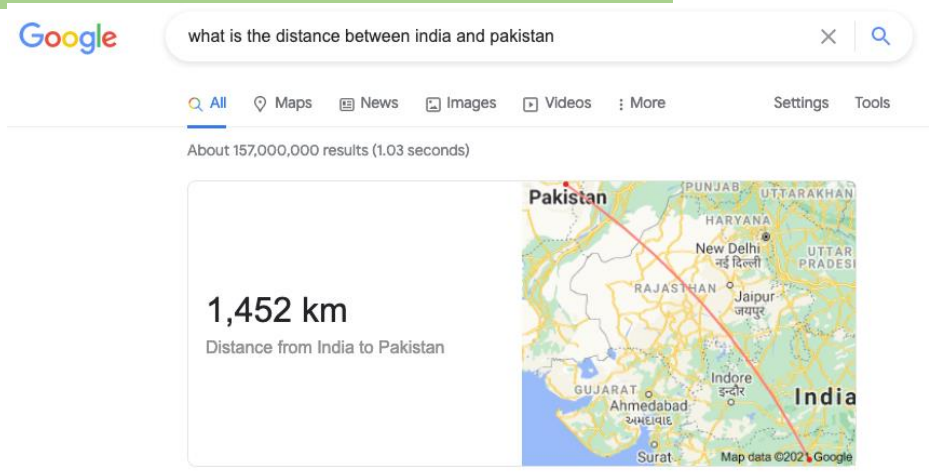


Figure 6: An example of how the **Knowledge Graph Programming System** enables efficient and high-quality knowledge applications in the case of COVID-19 analysis.

Technical Report on Data Integration and Preparation, El Kindi Rezig, Michael Cafarella, and Vijay Gadepally, MIT, arXiv:2103.01986v1, Mar 2, 2021.

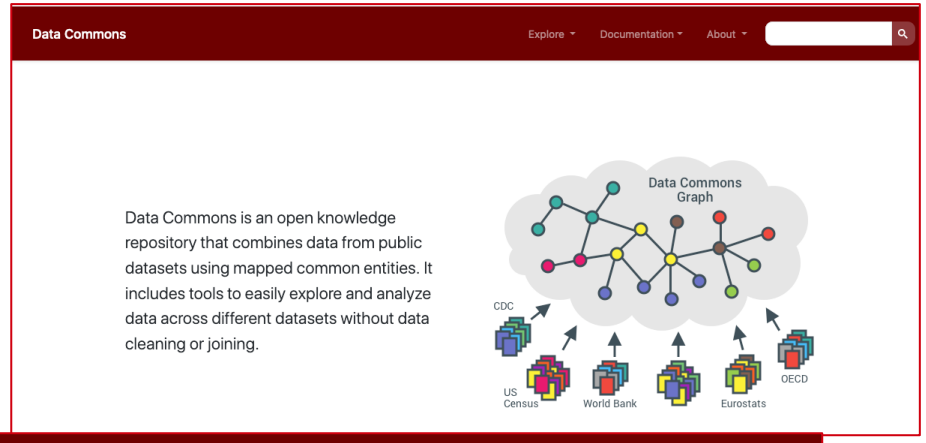
KnowWhereGraph: Geospatial information

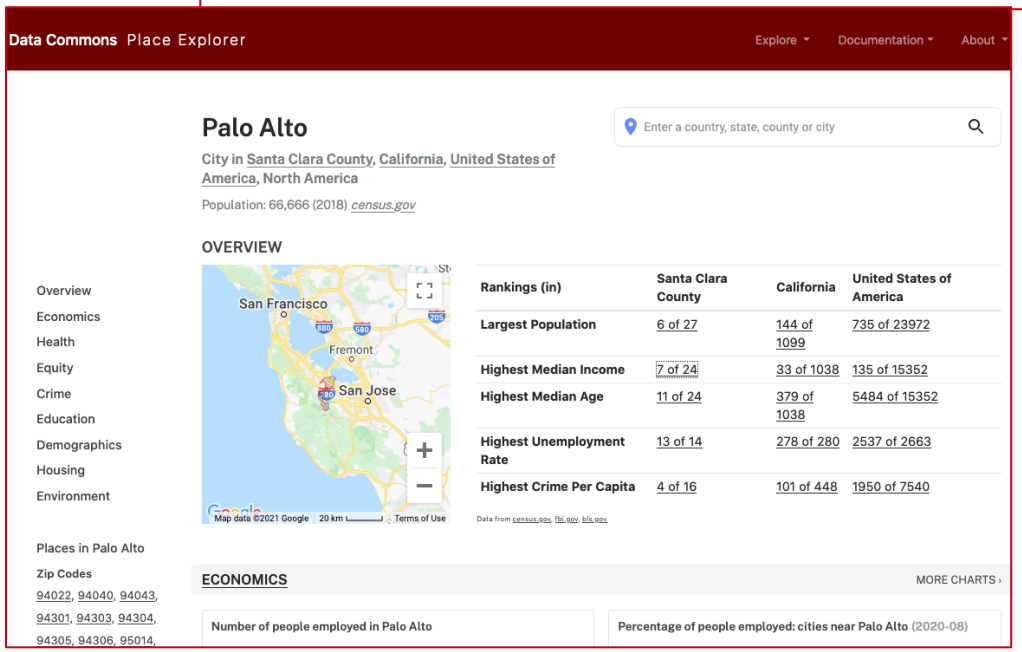
- The need for “geoenrichment” of knowledge graphs:
  - Ask Google: What is the distance between India and Pakistan?
- GeoEnrichment
  - Adding spatial “intelligence” to knowledge graphs
  - E.g., knowledge network of experts
    - Where are they located?
    - Which geospatial regions, or types of regions, does their expertise apply to?
- Tools to ingest RDF data into mapping software, e.g. ESRI ArcMap



Other efforts: datacommons.org

- Effort at Google
- Currently, mostly socioeconomic data
- Uses schema.org
- Becoming incorporated into Google search





Some issues…

- Implementation / scaling considerations
  - Underlying representation: Tables vs graphs
  - Graph databases, e.g. Neo4j, Stardog, OrientDB, InfiniteGraph, AllegroGraph, …
  - Tabular representation, e.g., Relational databases, Apache Spark, BigTable, BigQuery
- Dynamic data
  - Changing data and changing relationships; evolving knowledge graph; versioning
- User views
  - E.g., error and data quality viewed differently by different users / applications
  - *Comprehensive Modular Ontology IDE,* Modular Ontology Modeling, Cogan Shimizu, Karl Hammar, Pascal Hitzler, Semantic Web, http://www.semantic-web-journal.net/content/modular-ontology-modeling.
    - "More than one domain expert with overlapping expertise, and more than one ontology engineer on the team.
    - Based on our experiences, three types of participants are needed in order to have a team that can establish a good modular ontology: domain experts, ontology engineers, and data scientists."