Knowledge Graphs for AI in the Enterprise

François Scharffe

Tech Innovations Forum

Columbia University Libraries

11/25/2019

About me

- François Scharffe

- @lechatpito on Twitter

- PhD from University of Innsbruck, in EU Semantic Web lab

- Senior lecturer at the School of Professional Studies in the Applied Analytics Master in Sciences,Columbia University

- Consulting through my company The Data Chefs
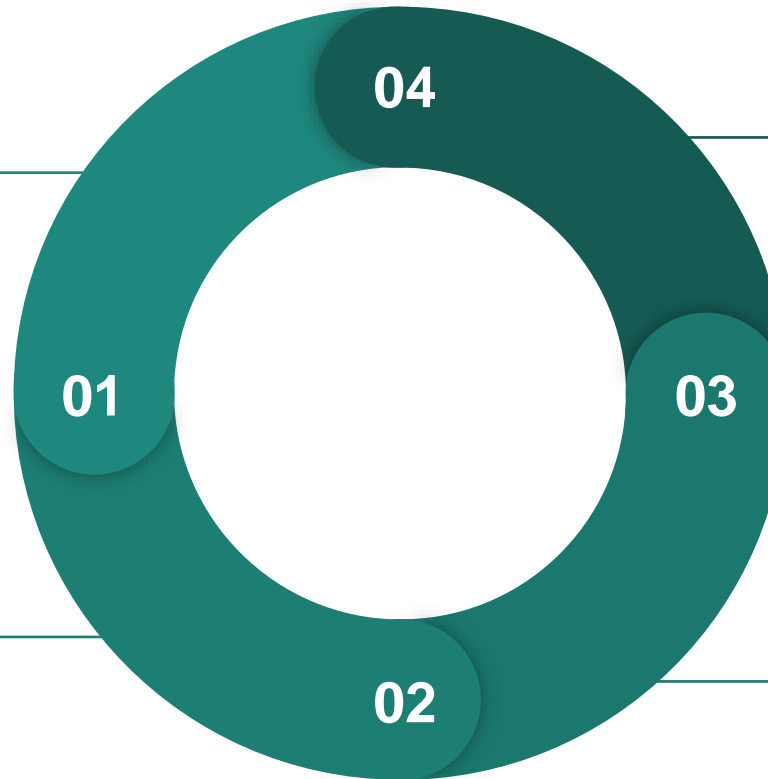
- Organizer of the Knowledge Graph Conference

The AI/ML/data science process

Define the business problem

What is the problem at hand? Is there a question to answer? Is it an open problem?

Build the dataset

What are the data sources needed to solve the problem? Which ones are available? How to integrate the data? How often will the analysis performed on the data?

Evaluate results

Run the model over the data. Evaluate the results. Should the problem be reformulated? Is the dataset sufficient to answer the questions?

Design the model

Will a set of queries able to answer our business problem? Do we need a statistical model? Which model is appropriate? What are the features? The parameters of the models?
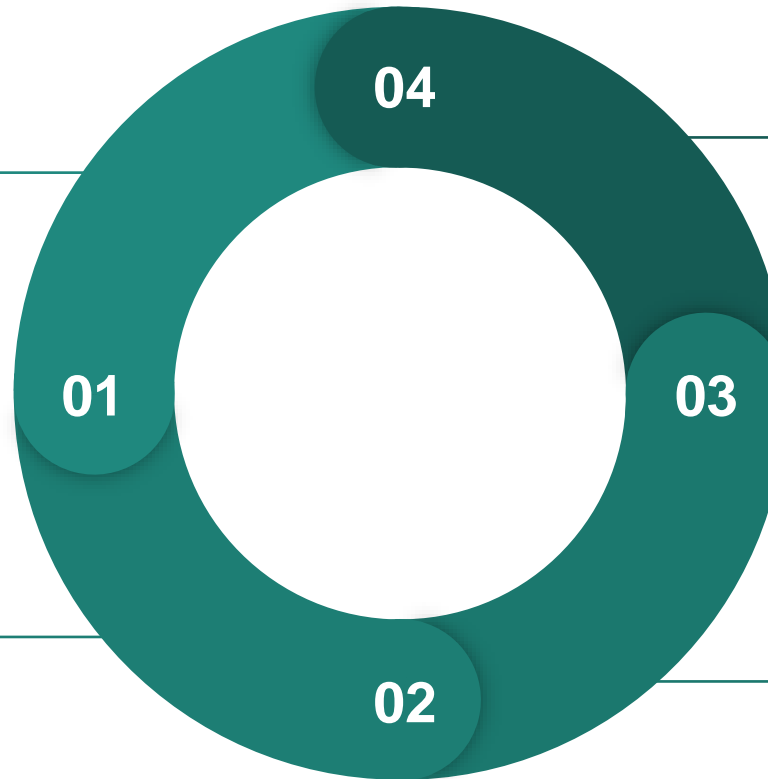
01 02 03 04

The The AI/ML/data science process

Define the business problem

What is the problem at hand? Is there a question to answer? Is it an open problem?

Build the dataset

What are the data sources needed to solve the problem? Which ones are available? How to integrate the data? How often will the analysis performed on the data?

Evaluate results

Run the model over the data. Evaluate the results. Should the problem be reformulated? Is the dataset sufficient to answer the questions?

Design the analytical model

Will a set of queries able to answer our business problem? Do we need a statistical model? Which model is appropriate? What are the features? The parameters of the models?

01 02 03 04

Building a data set is time consuming
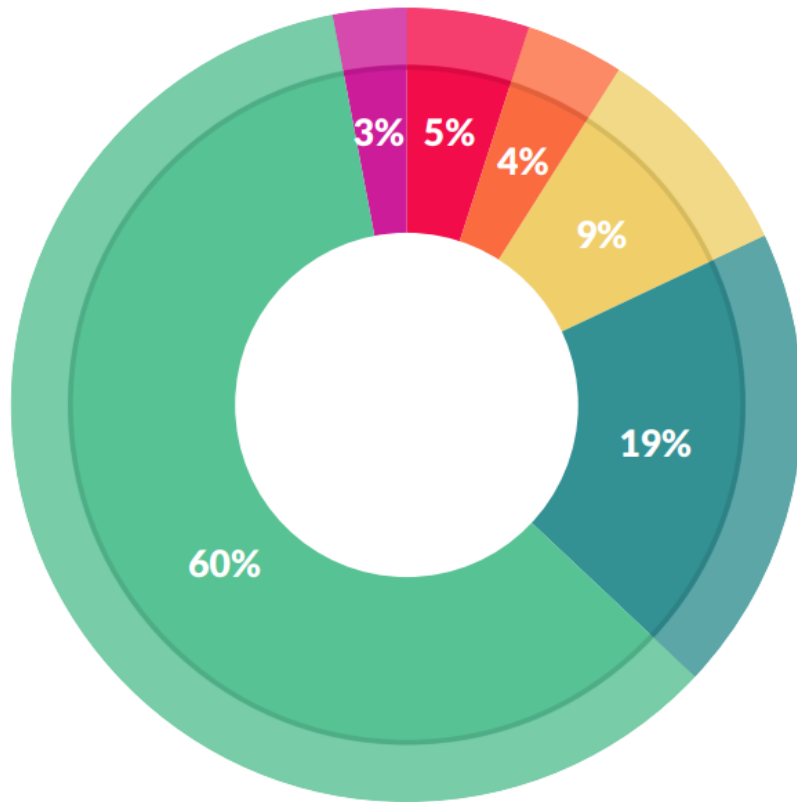
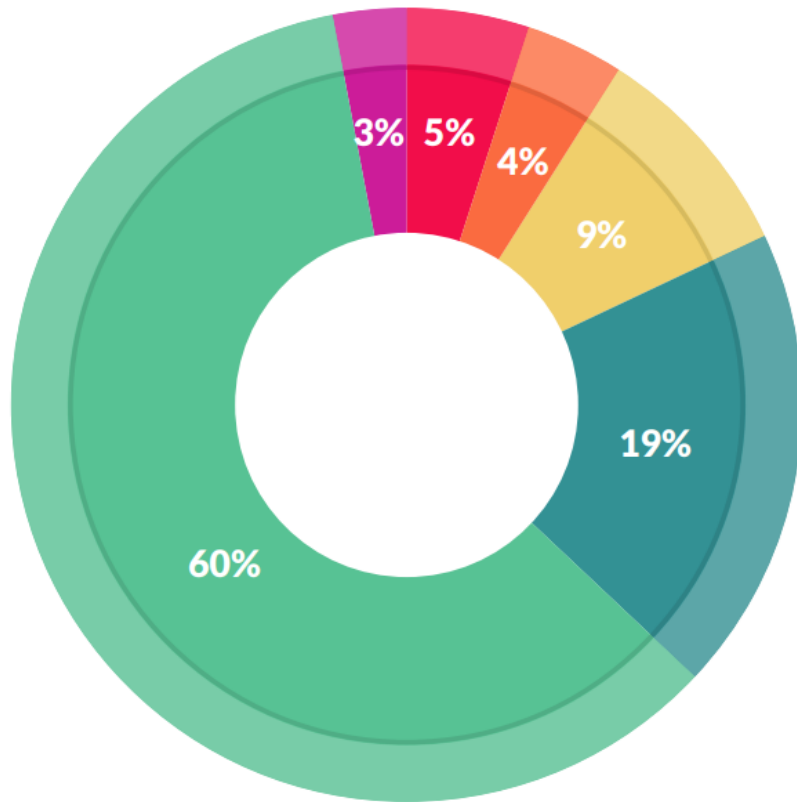What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Source Crowdflower http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

See also

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#54d5863f6f63

Building a data set is time consuming

What data scientists spend the most time doing

- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Credits Crowdflower http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

See also
https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#54d5863f6f63
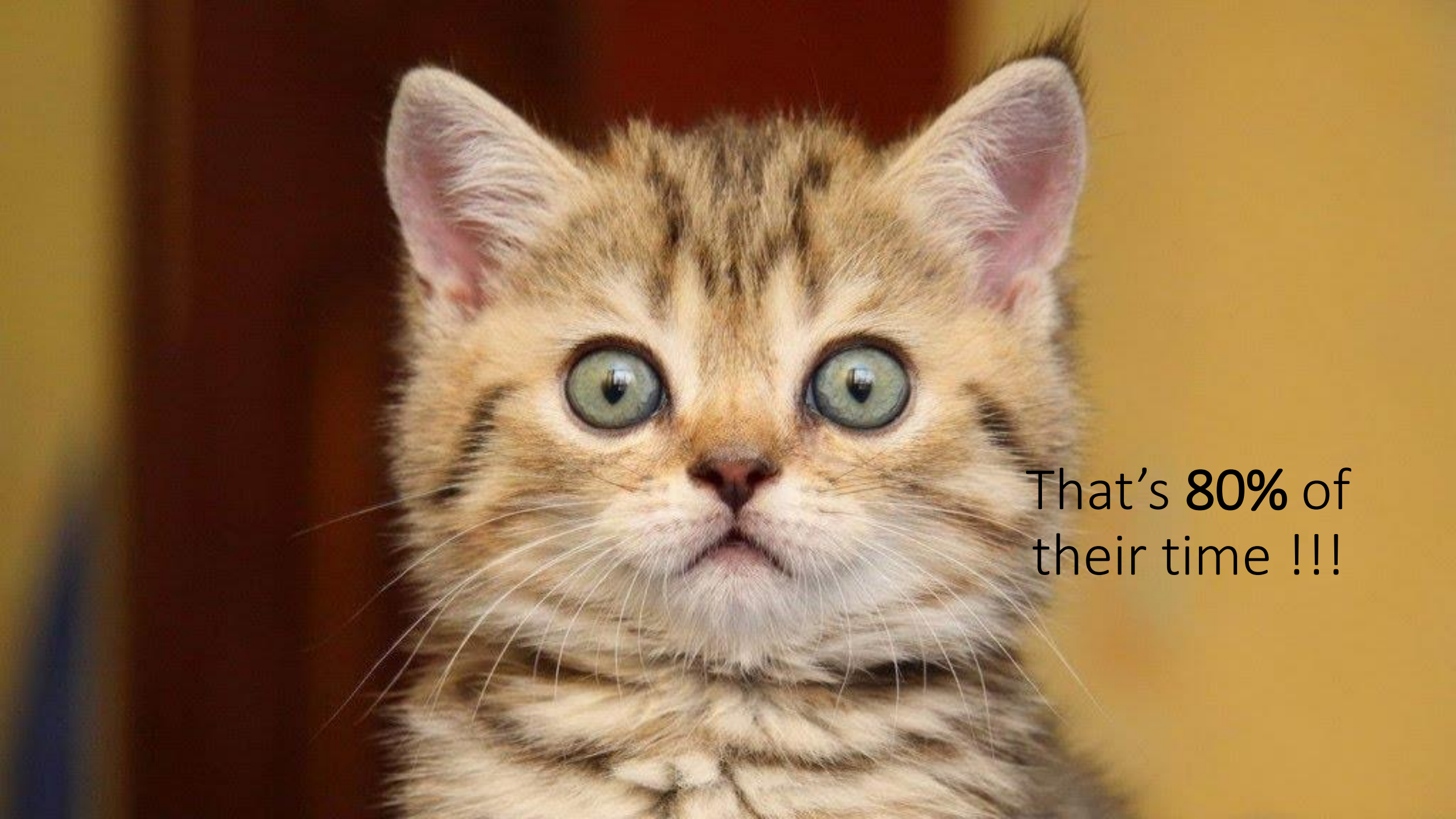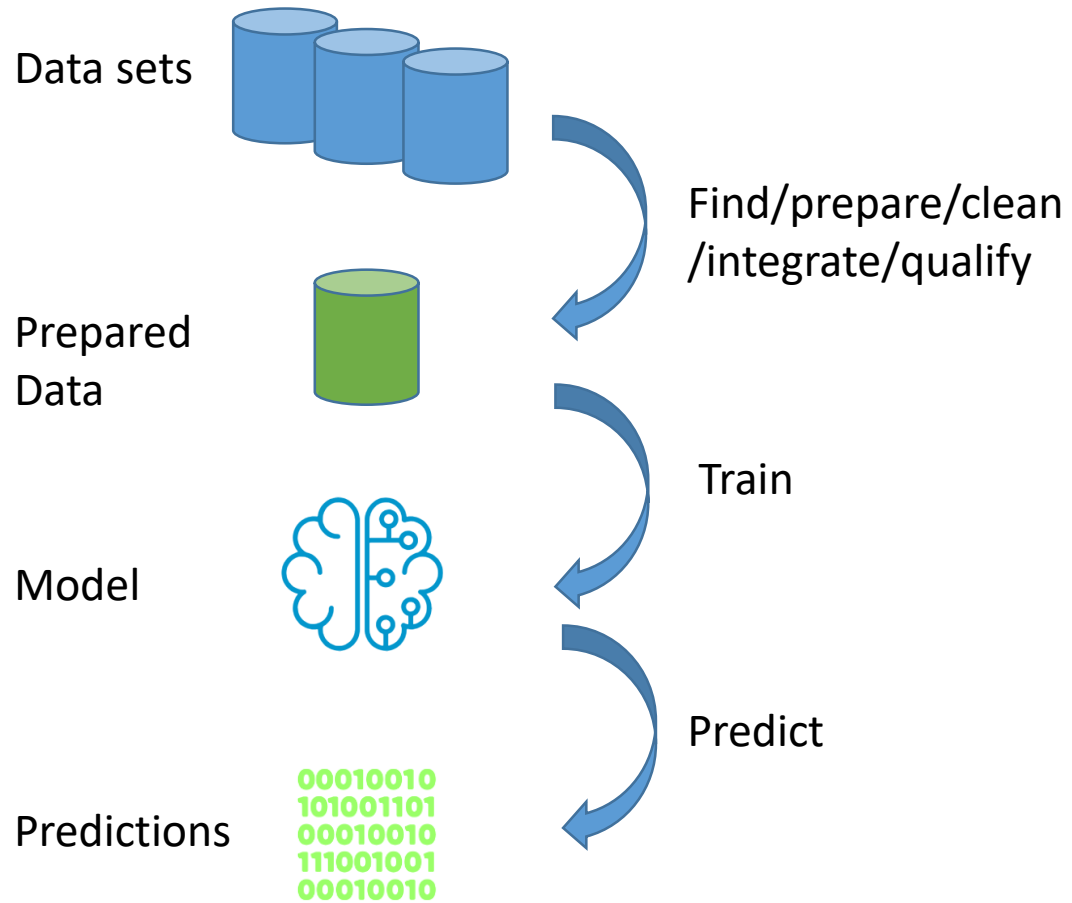
That's **80%** of their time !!!

Knowledge Graphs come to the rescue

Many data sources are **integrated** so that they can be **reused** across applications and models
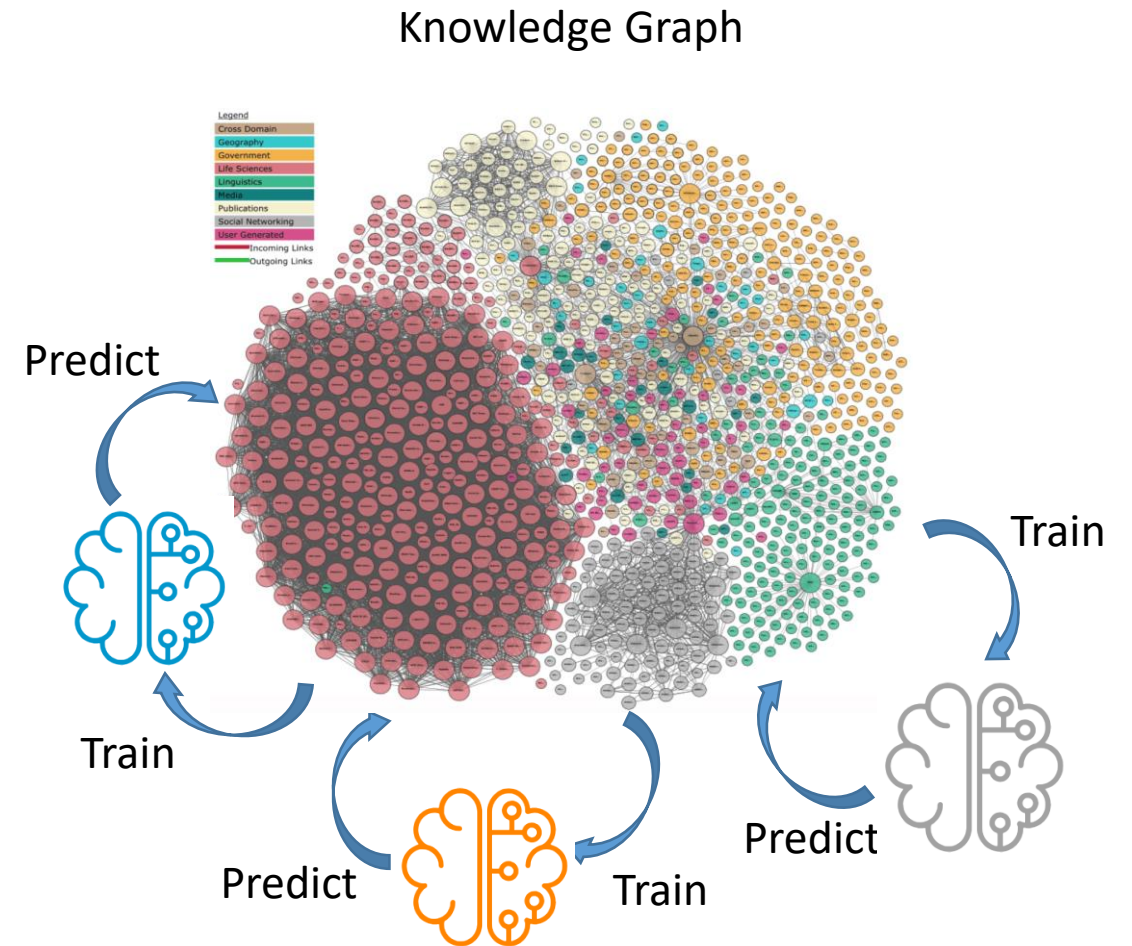
- The hard work is performed upstream
  - Data quality
  - Entity matching
  - Schema mapping
  - ETL
  - Ontology definition

KGs for smarter AI

- AI today

- AI tomorrow

Data sets

Find/prepare/clean
/integrate/qualify

Prepared
Data

Train

Model

Predict

Predictions

```
00010010
101001101
00010010
111001001
00010010
```

Knowledge Graph

Predict

Train

Train

Train

Predict

Predict
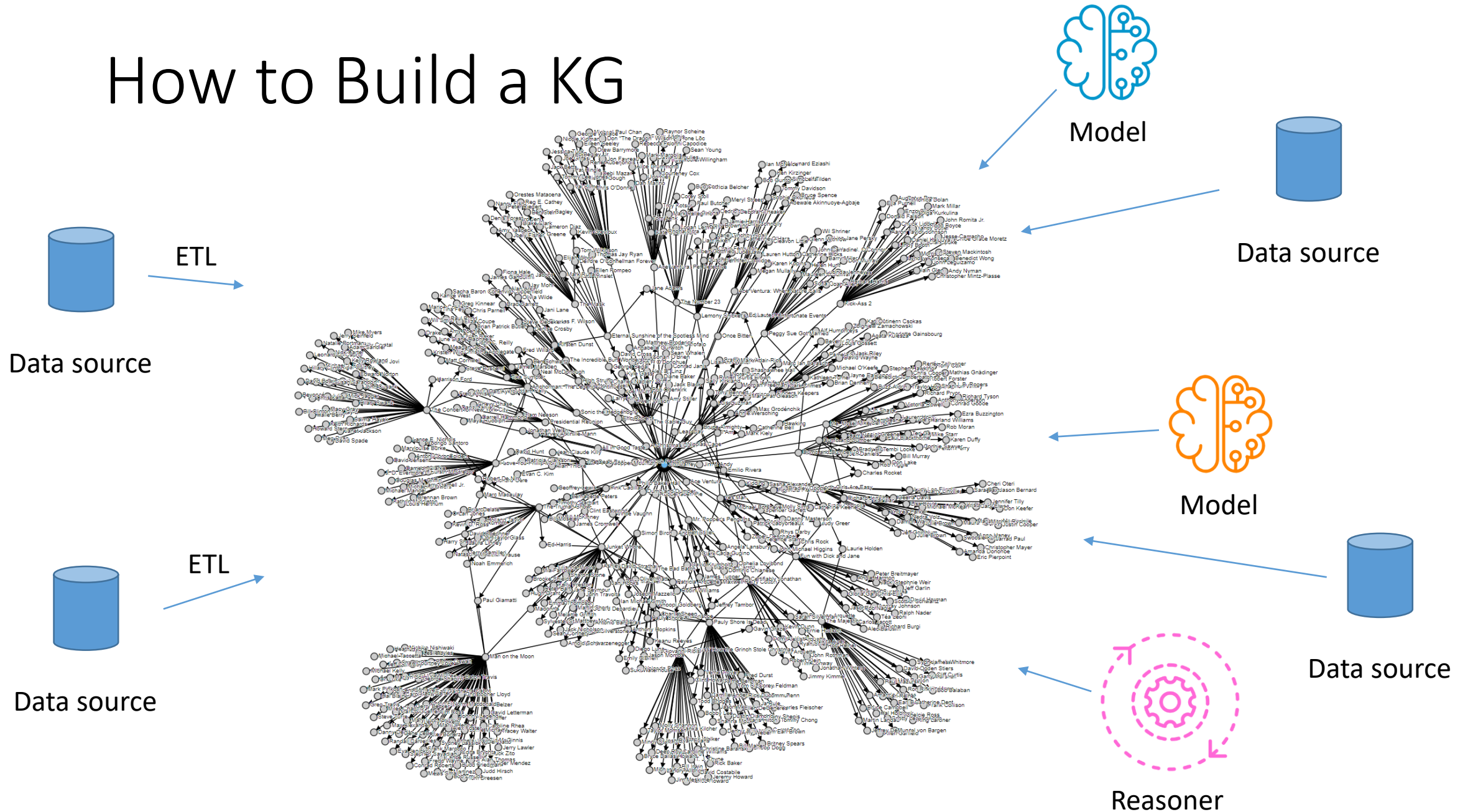
Train

Knowledge Graphs for AI

Intelligence needs a memory that
can be used to **reason logically**
about facts in the world

Knowledge graphs allows data to be
represented as facts
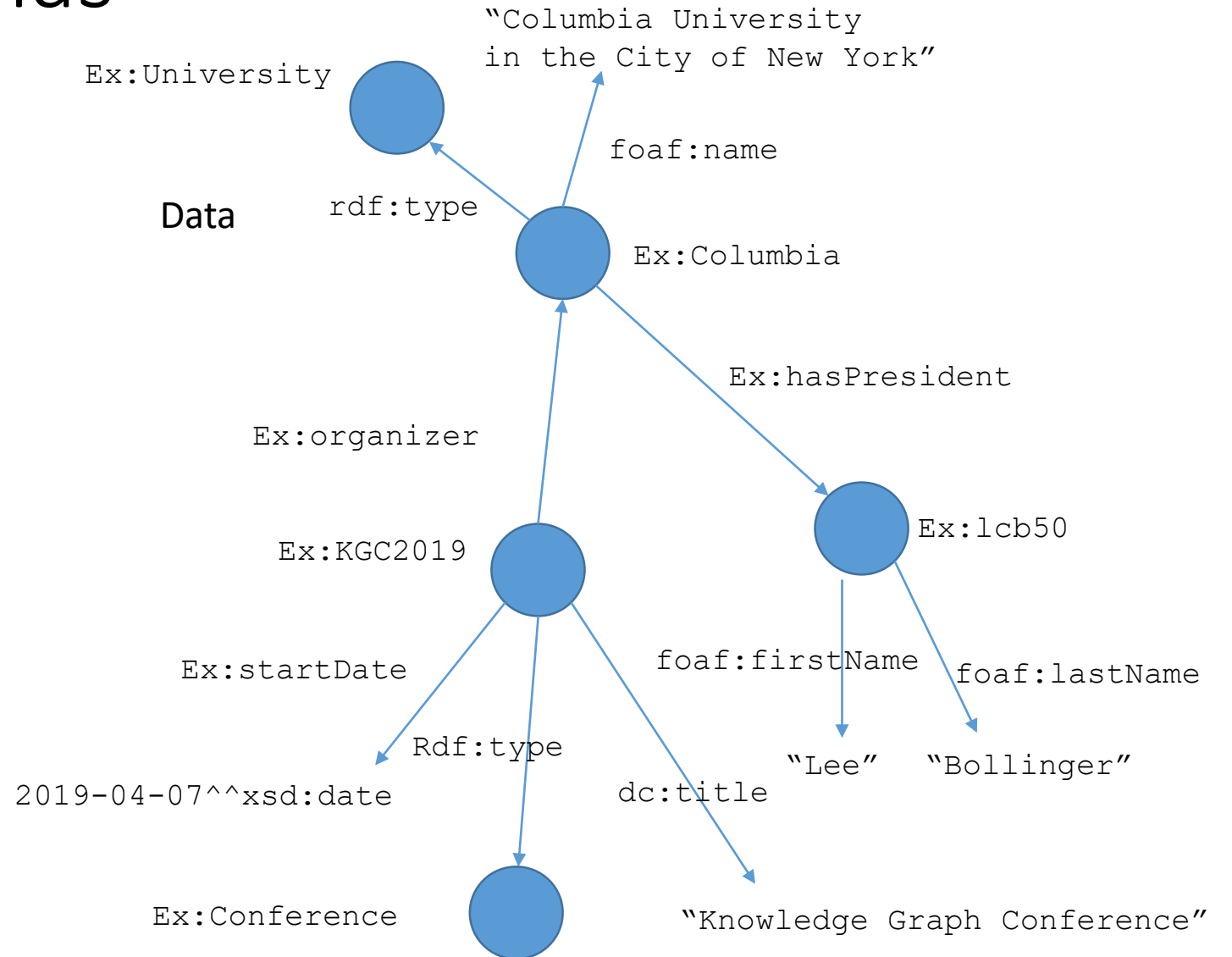
A brief history of knowledge graphs:
http://knowledgegraph.today/paper.html

How to Build a KG

Graphs and Schemas

Ontology

```
Ex:University a rdfs:Class.
Ex:organizer a rdfs:Property.
Ex:organizer rdfs:domain Ex:Event.
Ex:organizer rdfs:range foaf:Agent.
Ex:Conference a rdfs:Class.
Ex:Conference rdfs:subClassOf event:Event.
rdfs:subClassOf  foaf:Person.
…
```

Data

Ex:University

"Columbia University in the City of New York"

foaf:name

rdf:type

Ex:Columbia

Ex:hasPresident

Ex:organizer

Ex:KGC2019

Ex:lcb50

foaf:firstName   foaf:lastName

Ex:startDate

Rdf:type

"Lee"   "Bollinger"

2019-04-07^^xsd:date

dc:title

Ex:Conference

"Knowledge Graph Conference"

Graph Databases and Query Languages

- One open standard from the W3C
  - RDF the graph modeling language
  - SPARQL the graph query language
  - OWL the ontology language
- A mix of other more or less open languages
  - GSQL
  - GQL
  - Cypher
  - Gremlin
- Issues
  - fragmentation on one side
  - Complexity on the other. See RDF spec
- Ongoing effort to harmonize it all

Ontolowhat?

An ontology is a **formal specification** of a **shared conceptualization** of a **domain**

- a conceptualization: a description of the way we think about a domain

- a specification: there is a formal way of writing down the conceptualization

- formal: it is defined by axioms in a formal language

- Shared: it is meant to represent a community, it should be reusable across applications

Example: The Org Ontology
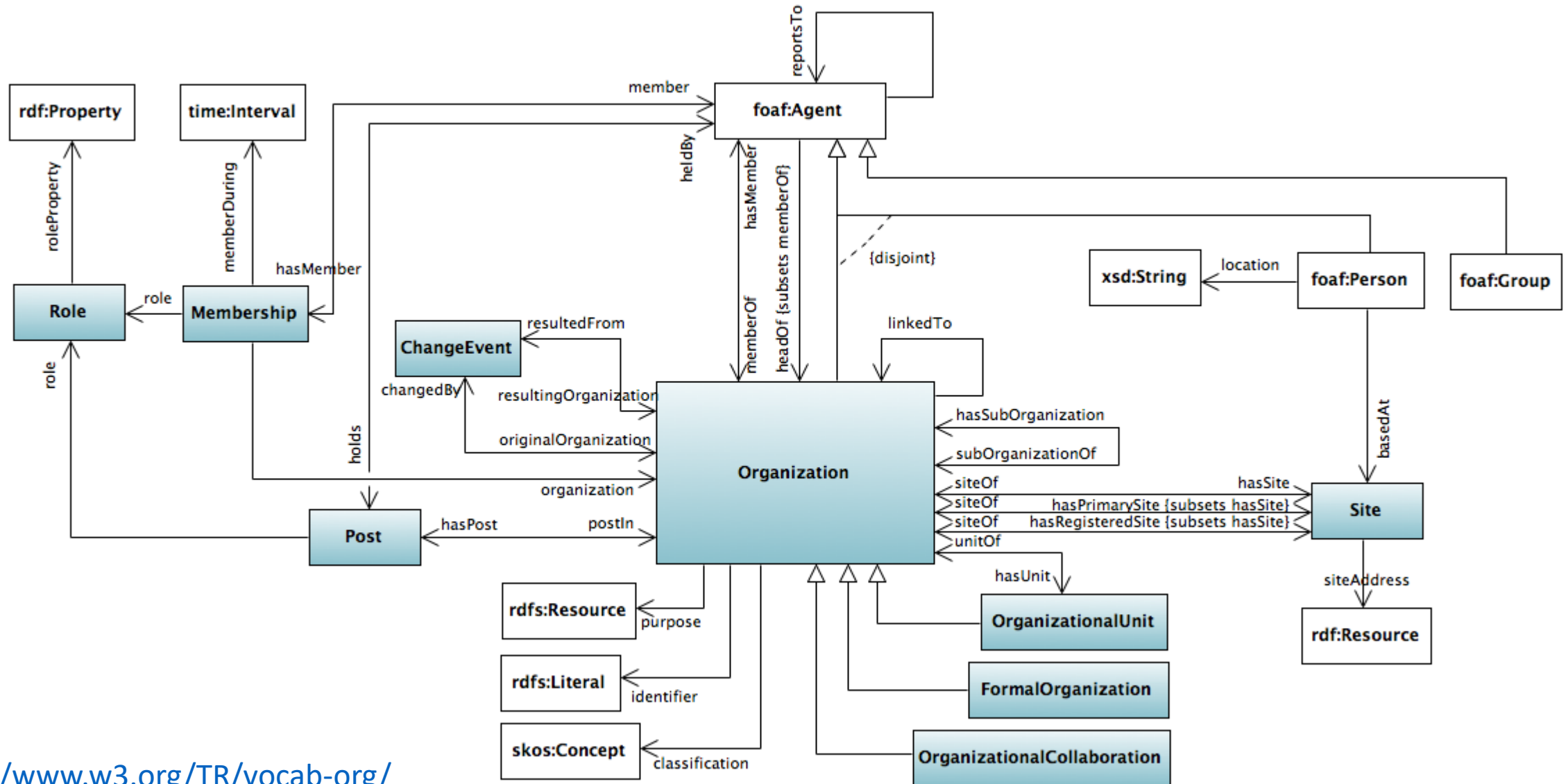
Ontologies

- http://schema.org
  - Open and collaborative development of an ontology for the annotation of web pages.

- FIBO: "the semantic standard for the financial industry"

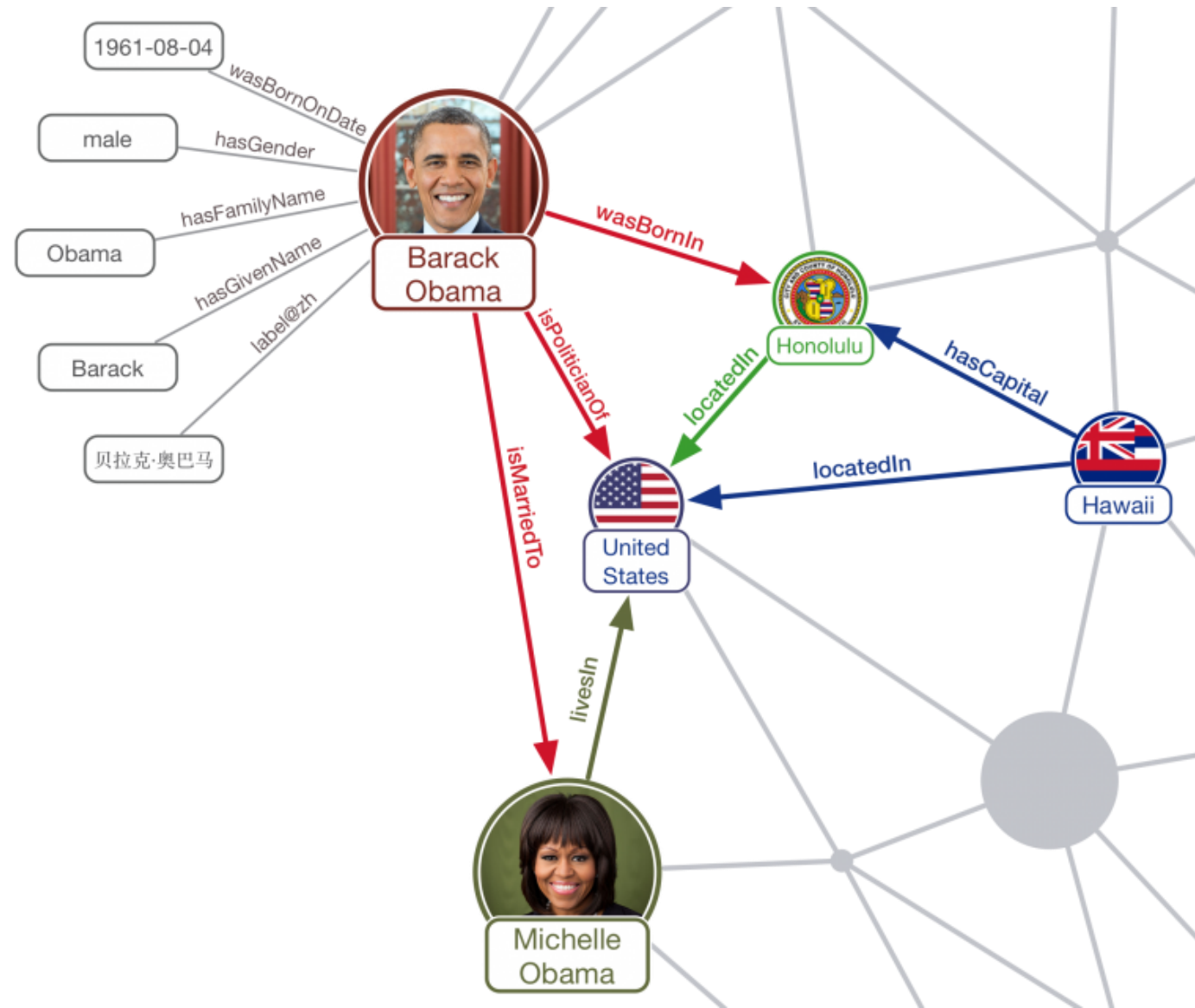- Lots of small ontologies https://lov.linkeddata.es/dataset/lov/
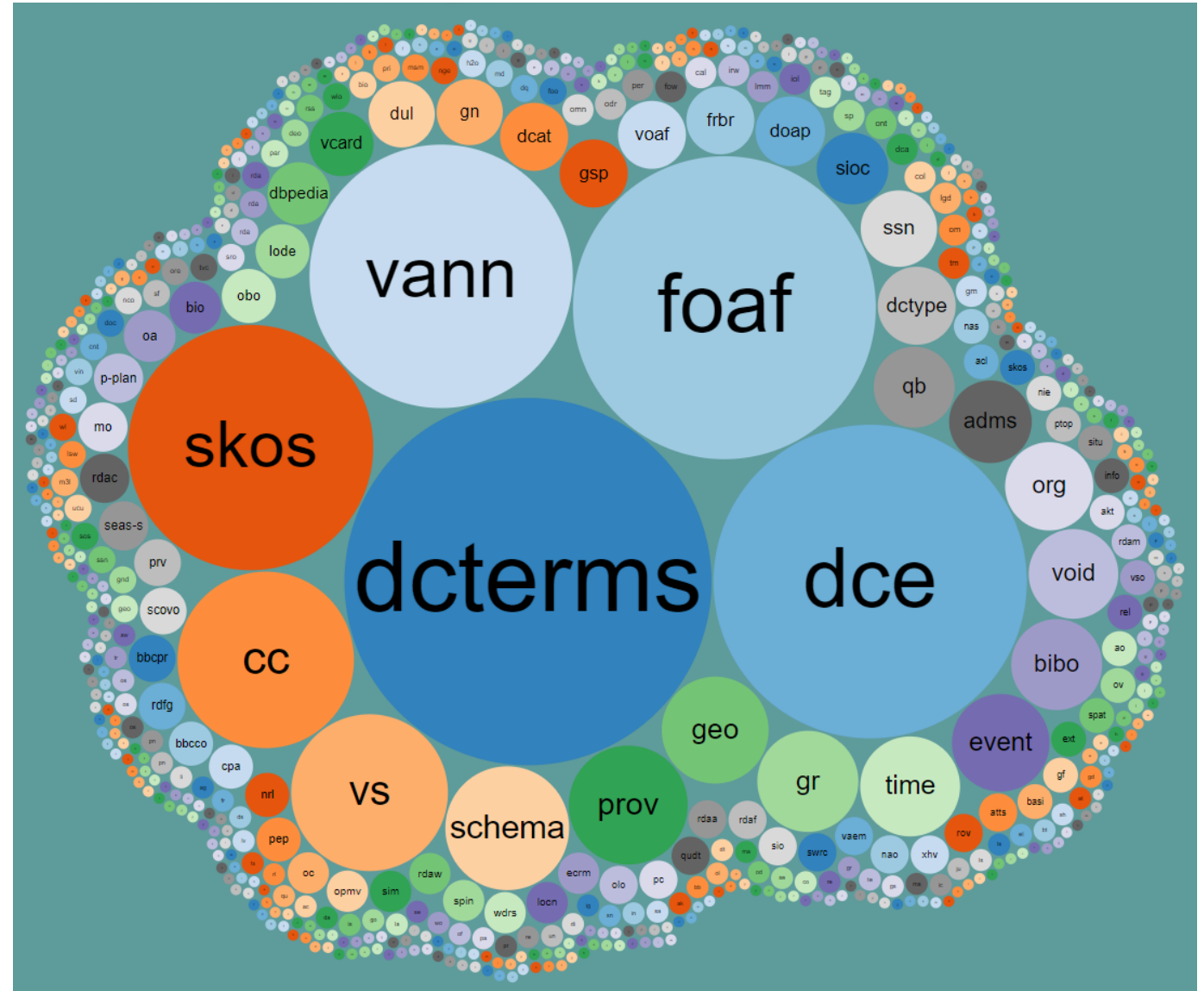
Graph vs Relational

- The inevitable question: why don't you use a SQL DB?

- There are many differences, let's cite the most relevant:
  - Graph DBs are more flexible
    - The data defines the graph structure. No a priori schema. Attributes (new links in the graph) can be added as you go and each resource has its own set of links.
  - Rich semantics are defined in ontologies
    - Ontologies define domains and are collaboratively created
    - This provide better integration with external datasets
    - Semantics can be tailored to the use case and the domain from very flexible to very crisp and constrained.
  - No need to join, path queries rule
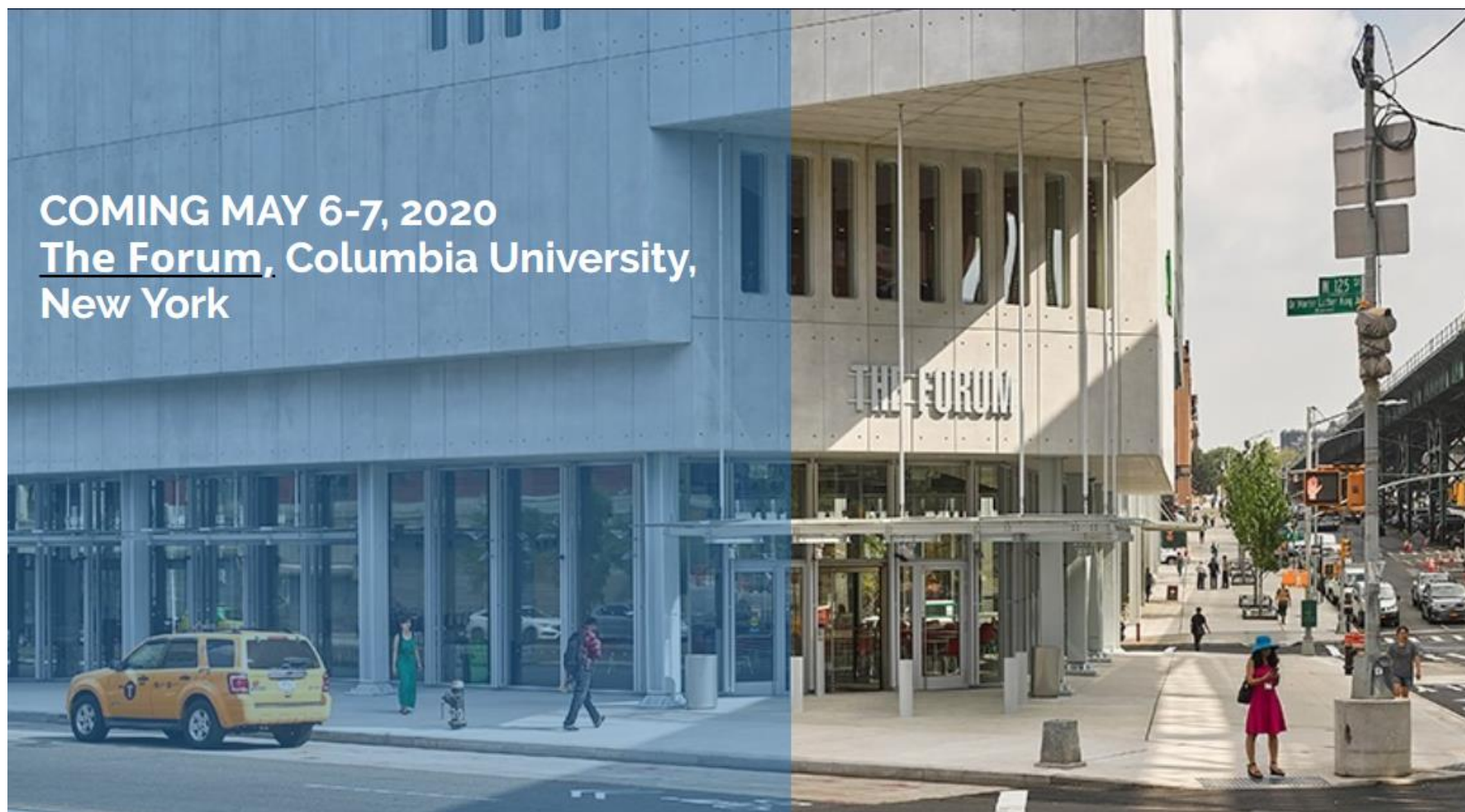
Knowledge Graphs in the Wild

- Google knowledge graph
  - https://developers.google.com/knowledge-graph/
  - Powers Google search result cards
  - Open ontology available at http://schema.org

- Facebook OpenGraph
  - Now not open anymore!
  - Let you navigate and query for uses, likes, entities through an API
  - Powers Facebook AI for recommendations

- Microsoft Concept Graph
  - Concepts extracted from the Web

- Bing Satori
  - https://azure.microsoft.com/en-us/services/cognitive-services/bing-entity-search-api/

- Wikidata
  - Source of Wikipedia structured data
  - http://www.Wikidata.org
  - https://angryloki.github.io/wikidata-graph-builder/

- Refinitiv
  - https://www.refinitiv.com/en/products/knowledge-graph-feed

- And many more

The Knowledge Graphs Conference 2019

The Knowledge Graphs Conference 2020

COMING MAY 6-7, 2020
The Forum, Columbia University, New York

THE FORUM

https://www.knowledgegraph.tech/

https://knowledgegraphconference.vhx.tv/
Groupe Slack