Becoming a Knowledge Scientist

Ernesto Jiménez-Ruiz

Lecturer in Artificial Intelligence

About the Module

Module Leader and Lecturer

- **Ernesto Jiménez-Ruiz**
- Research Centre for **Adaptive Computing Systems and Machine Learning**

Module Leader and Lecturer

- **Ernesto Jiménez-Ruiz**
- Research Centre for **Adaptive Computing Systems and Machine Learning**
- Contact:
  - Moodle forum / Teams
  - `ernesto.jimenez-ruiz@city.ac.uk`
  - `https://www.city.ac.uk/people/academics/ernesto-jimenez-ruiz`

Lecture and Lab: IN3067/INM713

Lecture:

– Lecture Capture: Recorded and live-streamed
– **Thursday, 09:00-10:50**,
– **C312 Tait Building**

Laboratory:

– **R201 (Franklin Building)**. 48 seats/PCs. 10 min walk.
– **Session 1: Thursday, 11:00-11:50**.
– Session 2: Thursday, 12:00-12:50 (Weeks: 1-4, weeks 7, 9 and 11). C301 on weeks 5, 8 and 10.
– Programming languages: Python and/or Java

Drop-in hours

- **Term 2**
  - Tuesday 2-3pm (online)
  - Thursdays 2-4pm (on-campus or online)
- Additional (online) drop-ins can be arranged via email.

Coursework

(Automatic) transformation of tabular data into semantic data.

– Part 1 (20%): **creation of an ontology** that covers the knowledge of a given domain. **Deadline**: Sunday, 3 March 2024, 5:00 PM

– Part 2 (80%): Design and development of a **software component**. **Deadline**: Sunday, 12 May 2024, 5:00 PM

– Allowed to **work in pairs**, or individually

– **To deliver:** codes or ontology, short report and video.

– There may be a short interview/presentation per student.

Reading list

- General resources list in moodle and Reading List Online.

- Recommended reading for each week.

- Suggested Books:
  - Semantic Web for the Working Ontologist. Morgan Kaufmann 2020
  - Knowledge Graphs. Morgan & Claypool 2021
  - Enterprise Knowledge Graphs. Morgan & Claypool 2021
  - Foundations of Semantic Web Technologies. CRC Press 2009

Students background questionnaire (2024)

- Knowledge of...
  - Semantic Web: 27%
  - Knowledge Graphs: 18%
  - Ontologies: 18%

Students background questionnaire (2024)

- Knowledge of...
  - Semantic Web: 27%
  - Knowledge Graphs: 18%
  - Ontologies: 18%

- Python preference: 64%, Java preference: 27%, Both: 9%

Students background questionnaire (2024)

- Knowledge of...
  - Semantic Web: 27%
  - Knowledge Graphs: 18%
  - Ontologies: 18%

- Python preference: 64%, Java preference: 27%, Both: 9%

- Topics of interest:
  - ML/AI + KGs

Module topics (tentative) - (i)

- **Week 1**: Introduction: becoming a knowledge scientist.
  - Release of coursework part 1 and part 2.
- **Week 2**: RDF-based knowledge graphs.
- **Week 3**: OWL ontology language. Focus on modelling.
- **Week 4**: SPARQL 1.0.
- **Week 5**: From tabular data to KG.
  - Assignment coursework part 1 (March 3).

Module topics (tentative) - (ii)

- *Week 6: Reading week*
  - Release model OWL ontology for coursework part 2.
- **Week 7**: RDFS Semantics and OWL 2 profiles.
- **Week 8**: Ontology (KG) Alignment.
- **Week 9**: Ontology (KG) Embeddings and Machine Learning.
- **Week 10**: SPARQL 1.1, Rules and Graph Database solutions
- **Week 11**: To be defined (LLMs and KGs?)
  - Seminar and/or invited talk(s)

Brief introduction

About Semantic Web technologies

- Originally designed as an extension of the World Wide Web to **make Internet data machine-readable**.

- Nowadays also covers the encoding of the semantics of arbitrary data (*e.g.*, enterprise data): **Knowledge Graphs**

About Semantic Web technologies

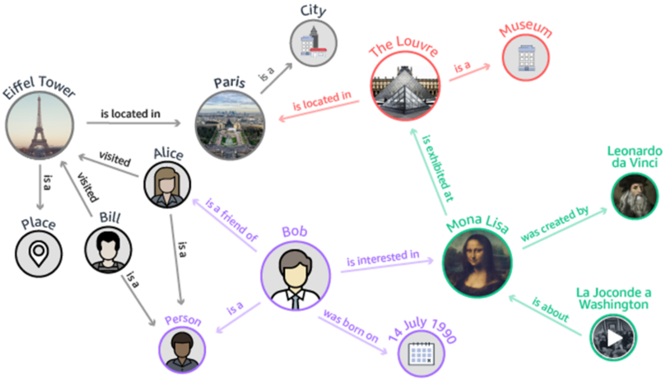
- Originally designed as an extension of the World Wide Web to **make Internet data machine-readable**.

- Nowadays also covers the encoding of the semantics of arbitrary data (*e.g.*, enterprise data): **Knowledge Graphs**

(*) How LLMs fit within the Semantic Web / Knowledge Graphs vision?

About Knowledge Graphs

- **(data graph)** capture information about entities of interest (like people, places or events), and
- **(knowledge)** with an underlying formal model



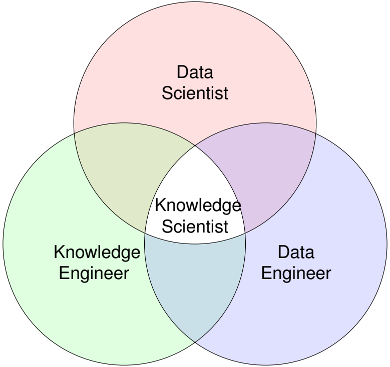
Knowledge Graphs: `https://arxiv.org/abs/2003.02320`.

About Knowledge Graphs

- In data science and AI, knowledge graphs are commonly used to:
  - Serve as bridges between humans and systems;
  - Facilitate access to and integration of data sources; and
  - Add context and depth to other, more data-driven AI techniques such as machine learning (*e.g.*, **Neurosymbolic AI**)

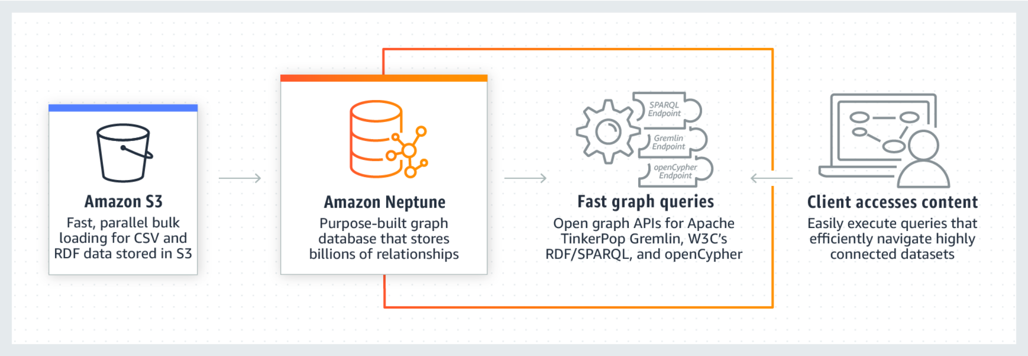
Interest Group @ The Alan Turing Institute: `https://www.turing.ac.uk/research/interest-groups/knowledge-graphs`

Relevance to BSc and MSc programmes



Knowledge Scientist: `https://www.knowledgescientist.org/`

Relevance to BSc and MSc programmes



KGs on AWS: `https://aws.amazon.com/neptune/knowledge-graphs-on-aws/`, `https://aws.amazon.com/neptune/`

Industry interest in KGs

There is a growing demand of computer scientists with KG experience

Semantic Web

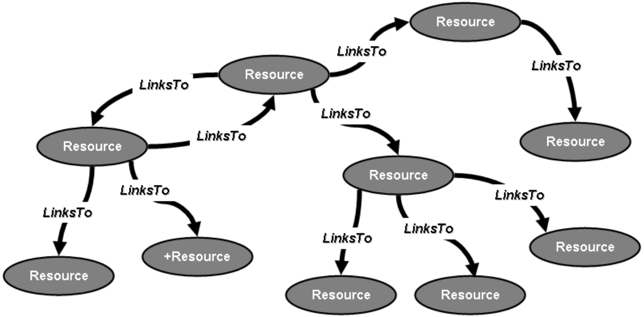
The Vision of a Semantic Web: Tim Berners-Lee

«I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize.»

Quoted from: *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web.* Tim Berners-Lee with Mark Fischetti. Harper San Francisco, 1999.

What is the Web?

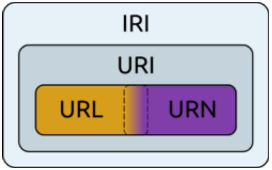
– The Web can be seen as a distributed network of hypertext pages that can refer to each other via URLs.



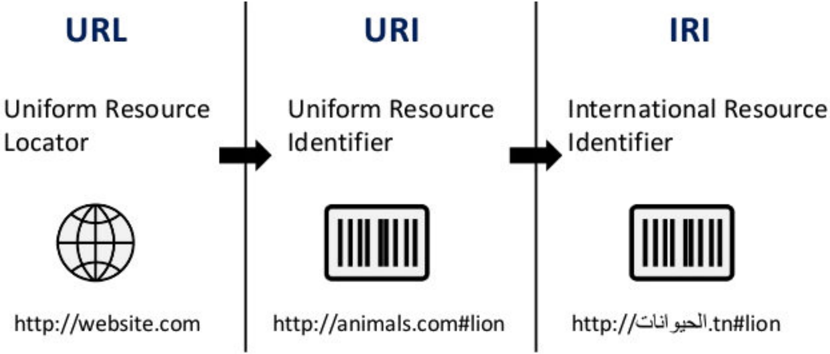
Semantic Web, and Other Technologies to Watch, 2007. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/`.

Resource identifiers

- **IRI** (Internationalized Resource Identifier): allows Unicode characters.
- **URI** (Uniform Resource Identifier): ASCII characters, includes URL and URN
- **URL** (Uniform Resource Locator): usually locates resources within the World Wide Web.
- **URN** (Uniform Resource Name): uses the *urn:* scheme and it is adopted by the ISBN system.



Resource identifiers (examples)



(*) Prefix definitions (or namespaces). `an: = http://animals.com#`
`http://animals.com#lion` → `an:lion`

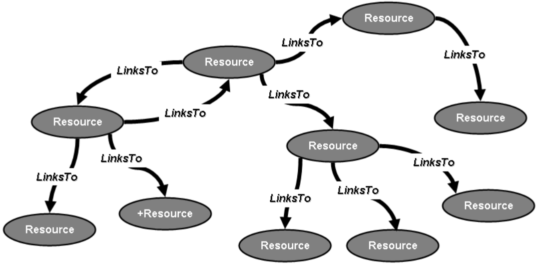
What is the Semantic Web? (i)

- The Semantic Web aims at going beyond the Web, towards a **Web of Data** where each individual data element has its own URI.

Semantic Web, and Other Technologies to Watch, 2007. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/`.

What is the Semantic Web? (i)

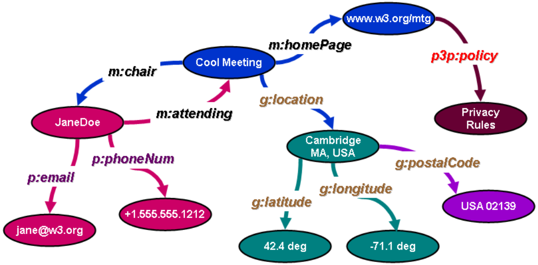
– The Semantic Web aims at going beyond the Web, towards a **Web of Data** where each individual data element has its own URI.



Semantic Web, and Other Technologies to Watch, 2007. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/`.

What is the Semantic Web? (i)

– The Semantic Web aims at going beyond the Web, towards a **Web of Data** where each individual data element has its own URI.



Semantic Web, and Other Technologies to Watch, 2007. `https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/`.

What is the Semantic Web? (ii)

The Semantic Web also aims for a Web of Data with **clear semantics**:

- Data is published in a **machine-readable format**
- Different information **sources can be linked**.
- Data is enriched with machine-interpretable meaning (domain knowledge).
- Smart agents can **draw conclusions** from the available information.

What is the Semantic Web? (ii)

The Semantic Web also aims for a Web of Data with **clear semantics**:

- Data is published in a **machine-readable format** (*)
- Different information **sources can be linked**.
- Data is enriched with machine-interpretable meaning (domain knowledge).
- Smart agents can **draw conclusions** from the available information.

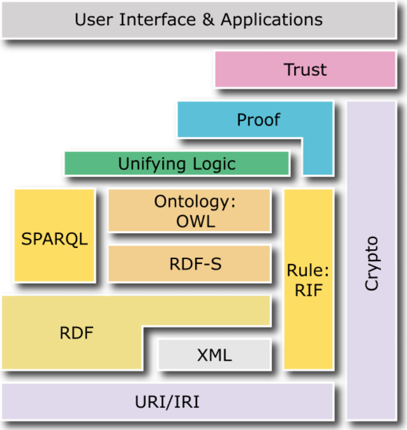
(*) How LLMs fit within the Semantic Web vision?

Semantic Web technology stack (i)

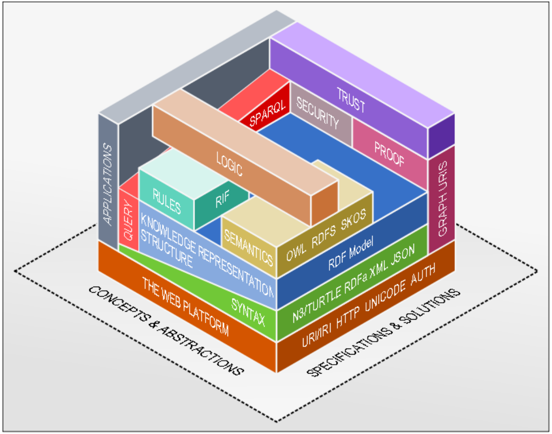
- The **World Wide Web Consortium (W3C)** is an international community that develops open standards to ensure the long-term growth of the Web: `https://www.w3.org/`
- **Why standards?**
  - broader industry (and academic) agreement,
  - interoperability across organizations and applications,
  - avoids vendor lock-in of a particular (exchange) format.
- The Semantic Web (as the Web) is built around W3C standards: the **Semantic Web stack**.

Semantic Web technology stack (ii)

- Identification: **URI/IRI**
- Data Representation: **RDF**
- Knowledge and Reasoning: **RDFS and OWL**
- Query language: **SPARQL**
- Exchange format: **XML**



Semantic Web technology stack (iii)



Overview standards. `https://csiro-enviro-informatics.github.io/info-engineering/standards.html`.
Image created by Benjamin Nowack (CC BY 3.0 Deed).

Motivating example

How to get to the cinemas in London that show a comedy?
Current situation:

- search to first find a possibly incomplete list of cinemas
- check if they are screening comedies
- locate the cinemas in the map
- find the best way to get there

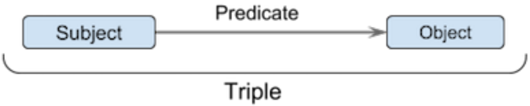
Motivating example: necessary ingredients

How to get to the cinemas in London that show a comedy?

1. Semantic language to express this query
2. Information with shared syntax and semantics across providers (*e.g.*, interoperability among ODEON, Cineworld and Transport of London)
3. A smart agent that can orchestrate all the information sources

(*) Google already facilitates a number of searches with its own mash-ups (now also in the form of a KG), but it does not contain mash-ups for all possible and desired Web searches. It may also lack updated information.
(**) LLMs can facilitate some of the steps.

Motivating example: RDF triples (i)



Cinema 1:

subject	predicate	object
http://cinemas/1	rdf:type	http://cinemas/Cinema
http://cinemas/1	http://cinemas/screening	http://movies/1
http://cinemas/1	http://cinemas/location	http://places/london
http://cinemas/1	http://cinemas/address	‘‘London W1T 1BX’’
http://movies/1	rdf:type	http://movies/Movie
http://movies/1	http://movies/has_genre	http://genres/parody

Motivating example: RDF triples (ii)

Cinema 2:

subject	predicate	object
http://cinemas/2	http://cinemas/screening	http://movies/2
http://movies/2	http://movies/has_genre	http://genres/zombies

Additional background knowledge:

http://places/london	rdf:type	http://places/City
http://genres/parody	rdf:type	http://genres/Comedy
http://genres/Comedy	rdfs:subClassOf	http://genres/Genre
http://genres/zombies	rdf:type	http://genres/Horror
http://genres/Comedy	owl:disjointWith	http://genres/Horror

Motivating example: SPARQL query

Example of (partial) query to extract relevant cinema addresses:

```
PREFIX cine:  <http://cinemas/>
PREFIX place:  <http://places/>
PREFIX movie:  <http://movies/>
PREFIX gen:  <http://genres/>
SELECT DISTINCT ?address WHERE {
  ?x rdf:type cine:Cinema .
  ?x cine:location place:london .
  ?x cine:address ?address .
  ?x cine:screening [ movie:has_genre [rdf:type gen:Comedy] ]
}
```

Challenges and Future of the Semantic Web (i)

- Distributed **Web of Data** applied to the whole Web is **challenging**
  - Data should be provided in RDF
  - Data should be linked to knowledge (*e.g.*, about cinemas, movies, genres, places, etc.)
  - Data providers should agree about the knowledge in intersecting domains
  - Consistency and trust of the data and knowledge
  - Smart agents in the Web

Challenges and Future of the Semantic Web (ii)

- **We are moving forward**:
  - There was also scepticism about the Web
  - **Examples**: Wikidata, DBPedia, the Linked Open Data Cloud, Bio2RDF, Google's Knowledge Graph.
  - Semantic Web within an organisation: **Graph(s) of Knowledge**
  - Combination of LLMs and KGs.

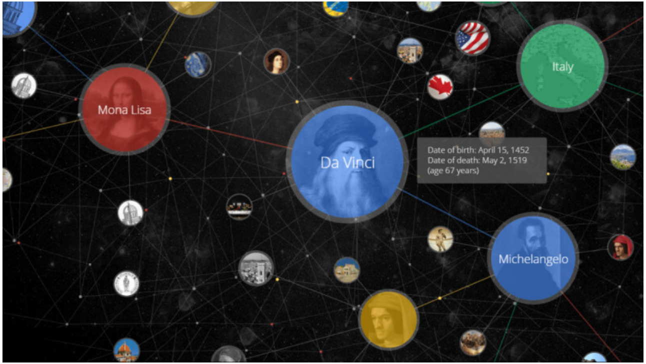
Graph(s) of Knowledge (i)

- Semantic Web in more **controlled scenarios**
- Not new but fresh notion of **knowledge graph in industry**
- Availability of (mature) **Semantic Web technology**
- Enterprise data as a knowledge graph to **drive products** and make them more **"intelligent"**

Graph(s) of Knowledge (ii)

- Identify and integrate disparate resources in the Web (*e.g.*, transport, cinemas, films)
- Integrate data within an organisation (*e.g.*, multiple data sources and departments)
- Combine life science data from genetic, pharmaceutical, patient database, etc.
- Cross-reference disparate digital libraries

Graph(s) of Knowledge: new search experience



Google’s Knowledge Graph

Challenges to create Graph(s) of Knowledge

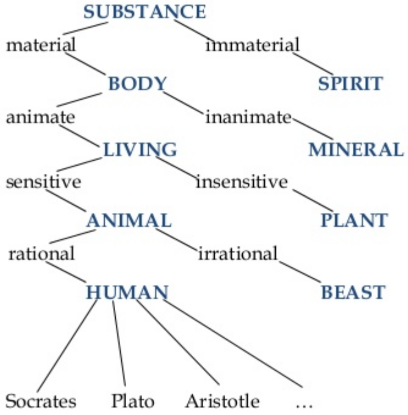
- How to **expose** data (*e.g.*, databases, csv files) as knowledge graphs?
- How to **create** (or reuse) and use (abstract) **knowledge** (*i.e.*, *Ontologies*)?
- How to **align** different knowledge graphs? (*)
- How to check **consistency and trust** of the data and knowledge? (*)

(*) Better with things than with strings

Knowledge Representation and Reasoning

Ontologies (in philosophy)

- Ontology is a discipline that deals with how to represent and categorise entities
- Aristotle (384-322BC): first systematic taxonomy of biology
- Porphyrian tree or Aristotle's categories (right)



Ontologies (information sciences)

- Play a key role in the development of the Semantic Web
- “**Formal** specifications of a shared domain conceptualization”
- “Abstract symbolic representations of a domain expressed in a **formal** language”

Thomas R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. 1993
Pim Borst, Hans Akkermans, and Jan Top. Engineering ontologies. 1999.

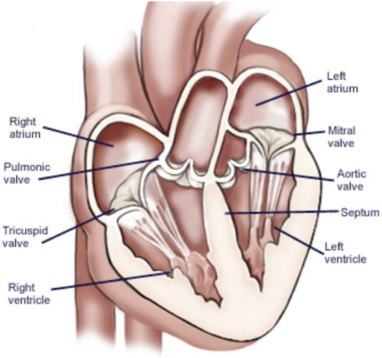
Ontologies as domain models (i)

- A model is a simplified (abstract) representation of certain aspects of the real world.
- Models help people communicate.
- Models explain and make predictions.
- Models mediate among multiple viewpoints.

Dean Allemang, James Hendler. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL.

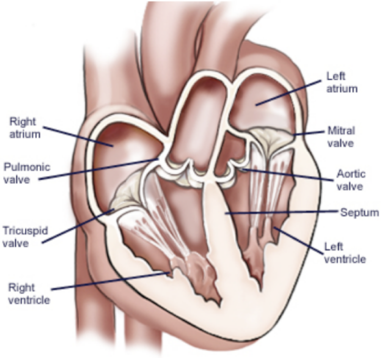
Ontologies as domain models (ii)

- include **vocabulary** relevant to a domain
- specify meaning (**semantics**) of terms
  - Heart is a muscular organ that is part of the circulatory system



Ontologies as domain models (ii)

- include **vocabulary** relevant to a domain
- specify meaning (**semantics**) of terms
  - Heart is a muscular organ that is part of the circulatory system
- are **formalised** using a suitable logic language
  - $\forall x.[Heart(x) \rightarrow MuscularOrgan(x) \wedge \exists y.[isPartOf(x,y) \wedge CirculatorySystem(y)]]$



Logic-based languages

- Symbolic logic is a subset of the formal logic
- **Propositional logic**:
  - simple facts or declarative propositions: a=*‘Ernesto is a Lecturer’*
  - Connectives: $\vee$, $\wedge$, $\neg$, $\rightarrow$, $\leftrightarrow$.

Logic-based languages

- Symbolic logic is a subset of the formal logic

- **Propositional logic**:
  - simple facts or declarative propositions: a=*‘Ernesto is a Lecturer’*
  - Connectives: $\vee$, $\wedge$, $\neg$, $\rightarrow$, $\leftrightarrow$.

- **Predicate logic** (*e.g.*, First Order Logic)
  - use of predicates (relationships)
  - unary predicates: $Lecturer(ernesto)$, $University(city)$
  - binary predicates: $TeachesIn(ernesto, city)$,
  - use of quantifiers ($\forall$, $\exists$) over variables in formulas:
    $\forall x.(Lecturer(x) \rightarrow Academic(x))$

Description Logics and OWL

- **Origin:** semantic networks and other graph-based models and the attempt to formalise them with First Order Logic (FOL).

Description Logics and OWL

- **Origin:** semantic networks and other graph-based models and the attempt to formalise them with First Order Logic (FOL).
- Core reasoning problems in **FOL** are **undecidable** (*e.g.*, is $ernesto$ an $Academic$?)
- **Trade-off** between **expressiveness** and computational properties

Description Logics and OWL

- **Origin:** semantic networks and other graph-based models and the attempt to formalise them with First Order Logic (FOL).
- Core reasoning problems in **FOL** are **undecidable** (*e.g.*, is $ernesto$ an $Academic$?)
- **Trade-off** between **expressiveness** and computational properties
- **Description Logics** (DL):
  - Family of knowledge representation languages
  - Decidable subset of FOL
  - Original called: *Terminological language* or *concept language*
  - OWL is based on DL

Calculating with Knowledge (i)

– Syllogisms (*i.e.*, inference) can be traced back to Aristotle

– Example:

$$\frac{\begin{array}{c}\text{All humans are mortal}\\ \text{Socrates is a human}\end{array}}{\text{Therefore, Socrates is mortal}}$$

Calculating with Knowledge (i)

- Syllogisms (*i.e.*, inference) can be traced back to Aristotle
- Example:

All humans are mortal
Socrates is a human

Therefore, Socrates is mortal

- Algorithmic manipulation of *knowledge*. . .
- . . . where the *meaning* of the words is not needed, e.g.,

All A are B
a is a A

Therefore, a is a B

Calculating with Knowledge (ii)

– Is Ernesto and Academic?

$$\frac{\begin{array}{c} Lecturer(ernesto) \\ \forall x.(Lecturer(x) \rightarrow Academic(x)) \end{array}}{Academic(ernesto)}$$

Ontologies and Knowledge Graphs

– Google has relaunched the interest on KGs

– Graph data models extensively studied in AI

– Core idea of knowledge graphs is the enhancement of the graph data model with knowledge.

– In this module: **OWL-layered RDF-based knowledge graphs**

Aidan Hogan and others. Knowledge Graphs. CoRR abs/2003.02320 (2020)

Ontologies VS relational databases

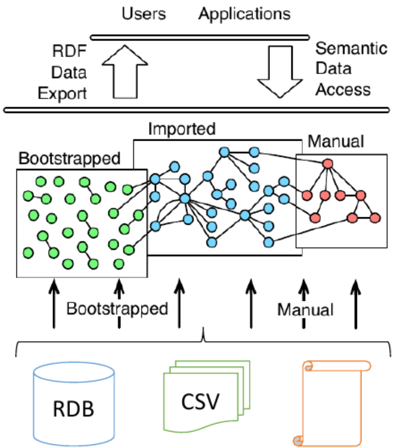
- Independence of logical/physical schema: **domain model**
- Vocabulary closer to domain experts: **more user-friendly**
- Incomplete and semi-structured data: **flexibility**
- Integration of heterogeneous sources: **unified view**

(*) They can complement each other.

Exposing data sources as RDF data

– **Virtual** (Knowledge Graph) exposure of the data (*OBDA: Ontology Based Data Access*)

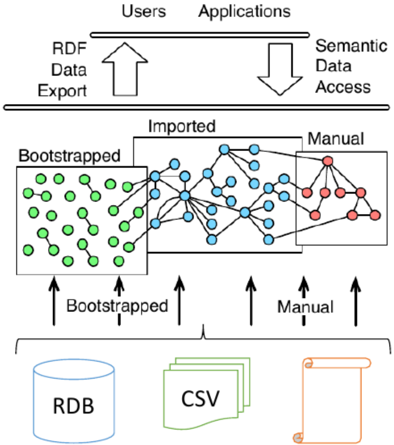
– **Materialised** Data Export. Useful to exchange data.



Exposing data sources as RDF data: ingredients

– **Ontology vocabulary**. Custom and/or given by a public KG.

– **Mappings**
  – RDB to RDF: relate ontology terms to queries over the RDB (W3C standard).
  – CSV to RDF: transformation functions. (W3C standard)

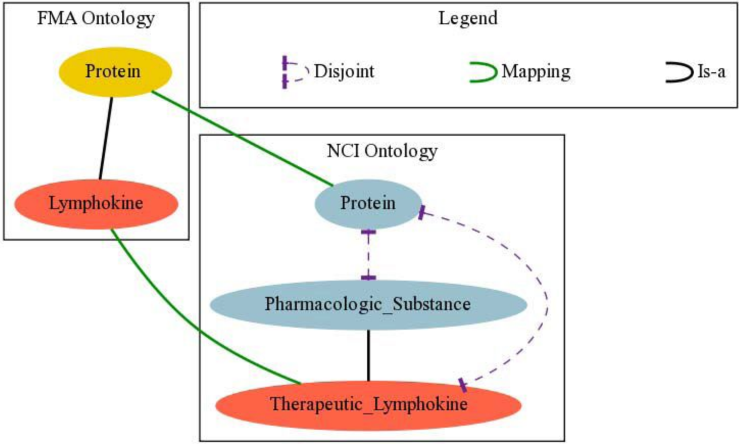


Knowledge Graph Alignment

- The same domain can be modelled from **multiple points of views**
- Same entity may come from **different data sources**
- Key to enable **interoperability**
- Easier to find **agreement among things** than strings

Ian Harrow and others. Ontology mapping for semantically enabled applications. Drug Discovery Today, 2019.
Jérôme Euzenat, Pavel Shvaiko: Ontology Matching, Second Edition. Springer 2013, ISBN

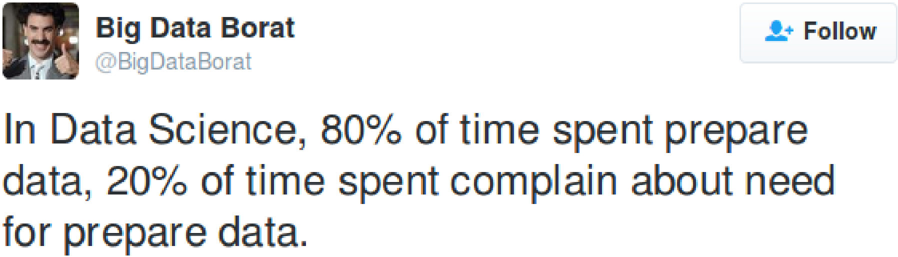
Knowledge Graph Alignment (example)



KGs for AI and Data Science

Motivation

Data understanding and data preparation involves the 80% of work on a data mining project.



Adding Semantics to Tabular Data

- **Tabular data** in the form of **CSV files** is the common input format in a data analytics pipeline.
- The **lack of semantics and context in datasets** hinders their usability.
- Gaining **semantic understanding** will be very valuable for KG creation, data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.

Contribution of Semantics in Data Wrangling Challenges

- *Data parsing*, e.g. converting csv's or tables.
- (+++)*Data dictionary*: basic types and semantic types.
- (++)*Data integration* from multiple sources (foreign key discovery).
- (++)*Entity resolution*: duplication and record linkage.
- (+)*Format variability*: e.g. for dates and names.
- (+)*Structural variability* in the data.
- (++)Identifying and repairing *missing data*.
- (+)*Anomaly detection* and repair.
- (+++)**Metadata/contextual information**. (Semantic) data governance.

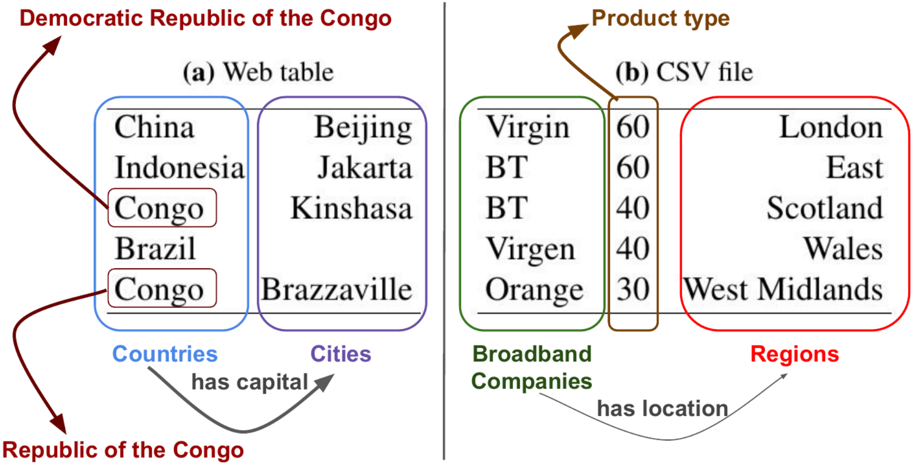
Adding Semantics to Tabular Data: Basic Tasks

- Matching a cell to a KG entity (**CEA task** - Cell-Entity Annotation)
- Assigning a semantic type (*e.g.*, a KG class) to an (entity) column (**CTA task** - Column-Type Annotation)
- Assigning a KG property to the relationship between two columns (**CPA task** - Columns-Property Annotation)

(*) We assume the existence of a (possibly incomplete) ***Knowledge Graph (KG)*** *relevant to the domain.*
(**) We may or not transform the (tabular) data into semantic data (i.e., extend the KG).

Ernesto Jiménez-Ruiz and others. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. ESWC 2020

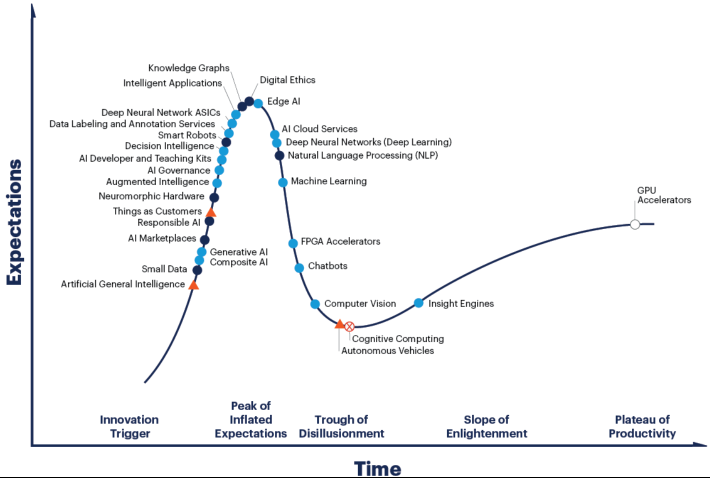
Adding Semantics to Tabular Data: Basic Tasks



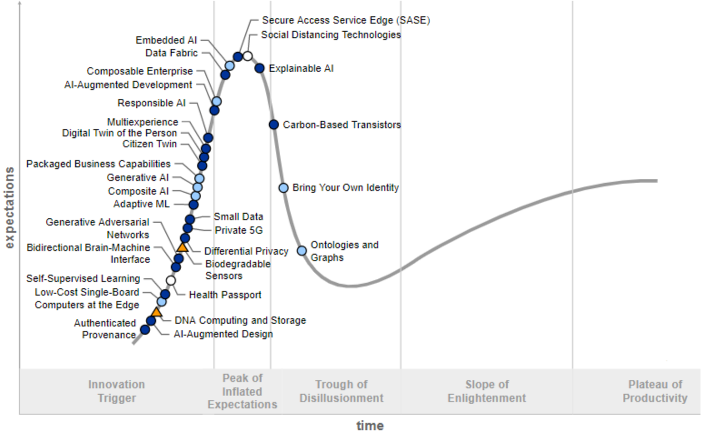
Next Decade in AI

- Gary Marcus has recently highlighted the need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.
- Impressive results in Deep Learning (including LLMs) but require (very) large datasets and lack explanation.
- Limitations of KG systems: maintenance and flexibility in the inference.
  - *e.g.*, $\forall x.\forall y.(A(x) \wedge R(x,y) \wedge B(y) \rightarrow C(x))$, $A(a)$, $B'(b)$, and $R(a,b)$. Does $C(a)$ hold?
- **Solution?** Combinations of connectionist or sub-symbolic systems with symbolic system (*e.g.*, **neuro-symbolic integration**, **KG-augmented LLMs**. **LLMs-enhanced KGs**)

Gartner's 2020 Hype Cycles: AI



Gartner’s 2020 Hype Cycles: Emerging Technologies



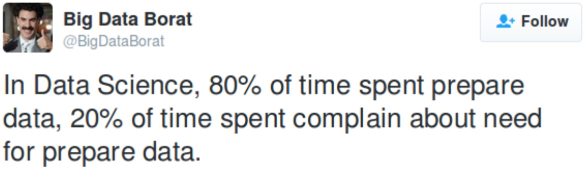
The Knowledge Scientist

Tasks of a Data Scientist

- Understand the data and its context
- Reliability of the data (shared with Data Engineers)
- Data wrangling
- Data analytics

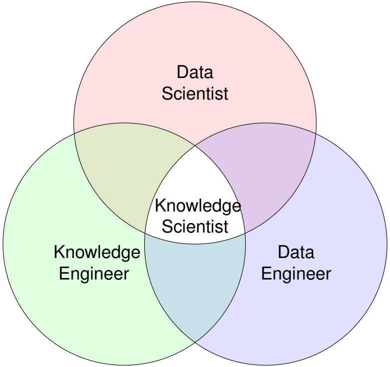
Tasks of a Data Scientist

- Understand the data and its context
- Reliability of the data (shared with Data Engineers)
- Data wrangling
- Data analytics



The Knowledge Scientist (i)

- **Data Engineer:** harnesses and collects data.
- **Data Scientist:** draws value from data.
- **Knowledge Engineer:** encodes domain expertise.
- **Knowledge Scientist:** adds context to the data to make it more useful, clean, reliable and ready to be used.

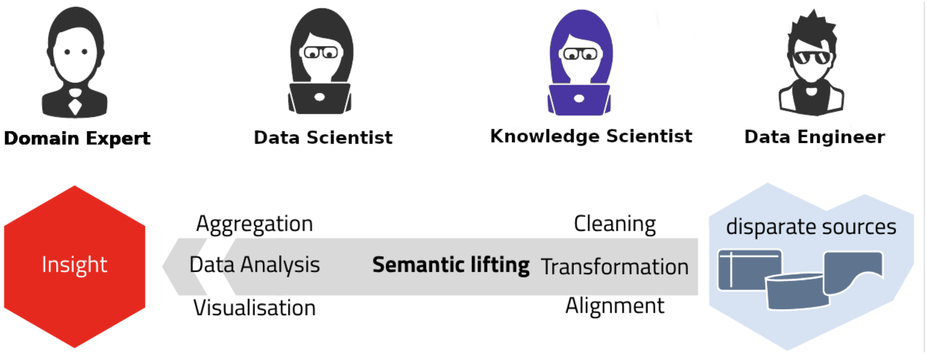


The Knowledge Scientist (ii)

- Bridges the data and the business requirements/questions.
- Outputs a data model (*e.g.*, a knowledge graph): how business users see the world.
- Drives a semantic-lifting of the data (from Data Engineers to Data Scientists)
- Relies on the **technology and skills** we cover in this module

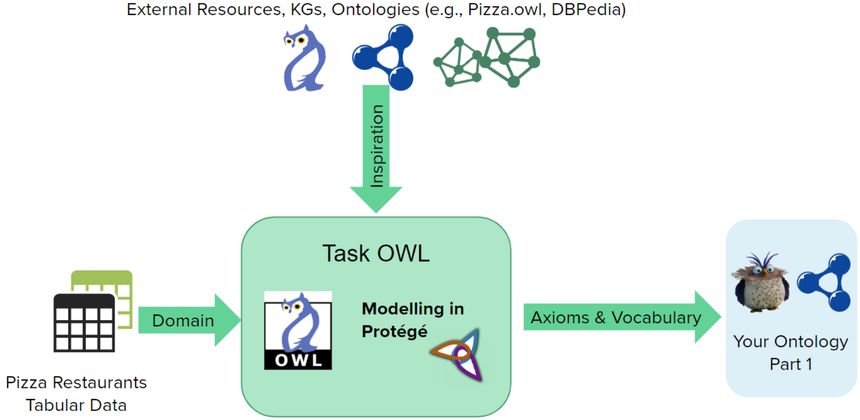
George Fletcher and others. Knowledge Scientists: Unlocking the data-driven organization. 2020

The Knowledge Scientist (iii)



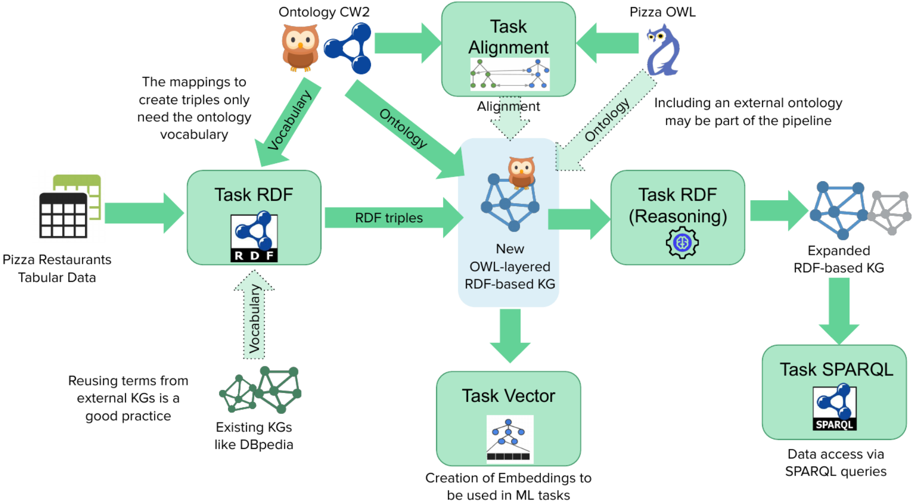
Coursework

Part 1: Ontology modelling



(*) Ontology CW2 = Model solution ≠ Your Ontology Part 1

Part 2: The global picture



Laboratory Session

Laboratory Session

- R201 (Franklin building) - 10 min. walk.
- This session is about infrastructure set-up.
- `https://staffhub.city.ac.uk/timetabling/rooms-by-building/franklin-building/r201`