



Becoming a Knowledge Scientist

Ernesto Jiménez-Ruiz

Lecturer in Artificial Intelligence

About the Module

Module Leader and Lecturer

- Ernesto Jiménez-Ruiz
- Research Centres for **Machine Learning and Adaptive Computing Systems** and *Artificial Intelligence*

Module Leader and Lecturer

- Ernesto Jiménez-Ruiz
- Research Centres for **Machine Learning and Adaptive Computing Systems** and *Artificial Intelligence*
- (Online) drop-in hours: **Wednesday from 1pm to 3pm**
 - Better to arrange meeting via e-mail.
- Contact:
 - Moodle forum
 - `ernesto.jimenez-ruiz@city.ac.uk`
 - `ernesto.jimenez.ruiz@gmail.com`
 - <https://www.city.ac.uk/people/academics/ernesto-jimenez-ruiz>

Lecture and Lab

- **Lecture:** Wednesdays 9am-10:50am (approx.)
 - Recording will be added to moodle.
 - Questions allowed: in chat or unmute.

Lecture and Lab

- **Lecture:** Wednesdays 9am-10:50am (approx.)
 - Recording will be added to moodle.
 - Questions allowed: in chat or unmute.
- **Lab:** Wednesdays 11am-12pm
 - Exercises related to the lecturer topic.
 - Programming languages: Python and/or Java
 - Optional break-out rooms

Lecture and Lab

- **Lecture:** Wednesdays 9am-10:50am (approx.)
 - Recording will be added to moodle.
 - Questions allowed: in chat or unmute.
- **Lab:** Wednesdays 11am-12pm
 - Exercises related to the lecturer topic.
 - Programming languages: Python and/or Java
 - Optional break-out rooms
- **Observers** (students in other modules):
 - Can be added to moodle

Course work

- Design and development of a **software component** applied to the Data Science pipeline
- Codes and **short report**.
- **Deadline**: April 21, 5pm (Wednesday)

Reading list

- General resources list in moodle.
- Recommended reading for each week.
- Books:
 - Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann. 2nd Edition:
<http://www.kevenlw.name/downloads/Ontologist.pdf>
 - Foundations of Semantic Web Technologies. CRC Press 2009:
<http://people.mpi-inf.mpg.de/~dstepano/KRSW/literature/SWTechnologies.pdf>

Initial module feedback

- Knowledge of...
 - Semantic Web: 25%
 - Knowledge Graphs: 50%
 - Ontologies: 42%

Initial module feedback

- Knowledge of...
 - Semantic Web: 25%
 - Knowledge Graphs: 50%
 - Ontologies: 42%
- Python preference: 92%, Java preference: 8%

Initial module feedback

- Knowledge of...
 - Semantic Web: 25%
 - Knowledge Graphs: 50%
 - Ontologies: 42%
- Python preference: 92%, Java preference: 8%
- Topics of interest:
 - KG and ML (*e.g.*, GNNs)
 - Practical implementations
 - (Semantic) Web development
 - Artificial General Intelligence

Module topics

- Semantic Web.
 - Knowledge graphs (KG).
 - Query language for KGs.
 - Ontologies and their semantics
 - Inference and Reasoning Engines.
-
- Knowledge graph or ontology alignment.
 - Applications to Data Science.
 - Machine learning and knowledge graphs.
 - Additional relevant topics

Semantic Web

The Vision of a Semantic Web: Tim Berners-Lee

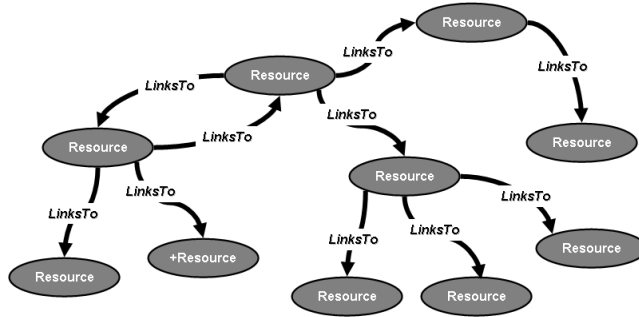
«I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy 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?

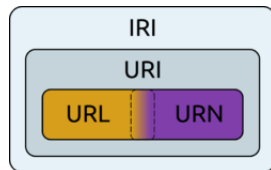
- The Web can be seen as a distributed network of hypertext pages that can refer to each other via URLs.



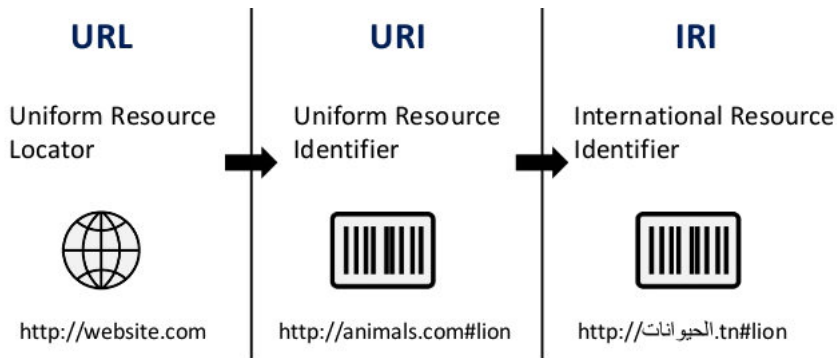
Semantic Web, and Other Technologies to Watch, 2007. <https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/>.

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`

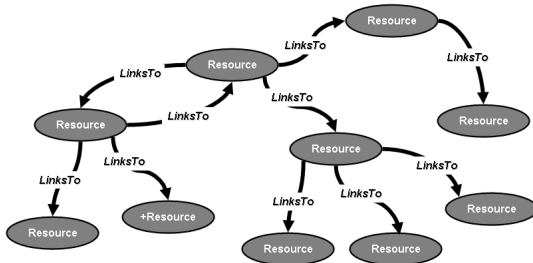
What is the Semantic Web? (i)

- 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)

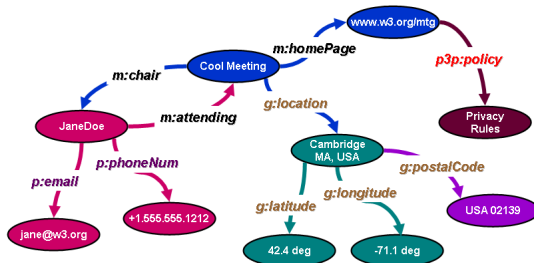
- The Semantic Web aims at going beyond the Web, towards a **Web of Data** where each individual data element has its own URI.



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**:

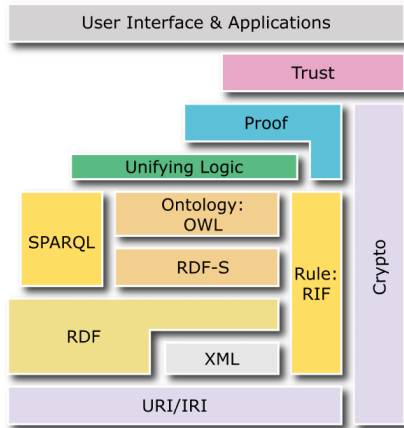
- Data is published in a **machine-readable format**.
- Different information **sources can be linked**.
- Data is enriched with machine-interpretable meaning (domain knowledge).
- Smart agents can **draw conclusions** from the available information.

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**



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

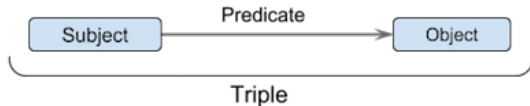
Motivating example: necessary ingredients

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 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.

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
<code>http://cinemas/2</code>	<code>http://cinemas/screening</code>	<code>http://movies/2</code>
<code>http://movies/2</code>	<code>http://movies/has_genre</code>	<code>http://genres/zombies</code>

Additional background knowledge:

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

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 were also sceptical 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**

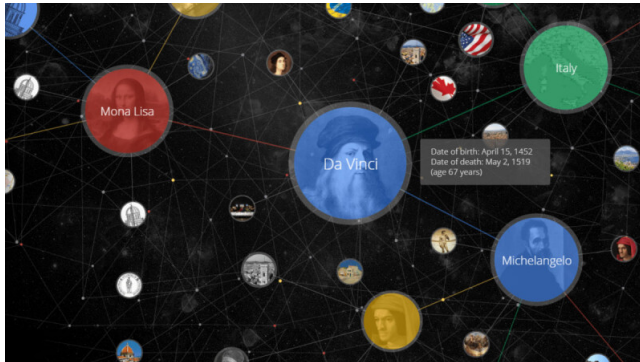
Graph(s) of Knowledge (i)

- 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



Leonardo da Vinci

Polymath

Leonardo da Vinci was an Italian polymath of the High Renaissance who is widely considered one of the most diversely talented individuals ever to have lived.

[Wikipedia](#)

Born: 15 April 1452, Anchiano, Italy

Died: 2 May 1519, Château du Clos Lucé, Amboise, France

On view: Ambrosian Library, Louvre Museum, [MORE](#)


Periods: High Renaissance, Early renaissance, Renaissance, Italian Renaissance, Florentine painting

Height: 1.75 m

Full name: Leonardo di ser Piero da Vinci

Artworks

[View 15+ more](#)




Mona Lisa
1503



The Last Supper



Vitruvian Man



Salvator Mundi

Google's Knowledge Graph

Challenges to create Graph(s) of Knowledge

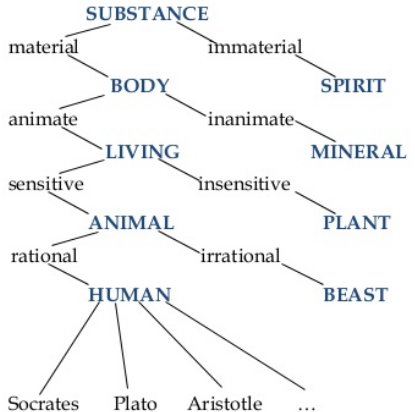
- How to **expose** data (*e.g.*, databases, csv files) as knowledge graphs?
- How to **create** (or reuse) and use (abstract) **knowledge** (*i.e.*, *Ontologies*)?
- How to **align** different knowledge graphs? (*)
- How to check **consistency and trust** of the data and knowledge? (*)

(*) 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.

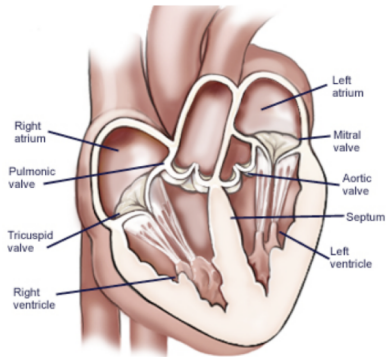
Ontologies as domain models (i)

- 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.

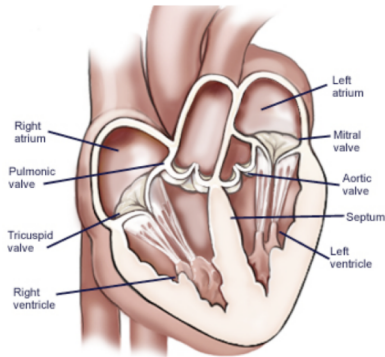
Ontologies as domain models (ii)

- 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: $a = \text{'Ernesto is a Lecturer'}$
 - 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: $a = \text{'Ernesto is a Lecturer'}$
 - 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:

All	men	are	mortal
Socrates	is a	man	
<hr/>			
Therefore,	Socrates	is	mortal

Calculating with Knowledge (i)

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

$$\begin{array}{l} \text{All men are mortal} \\ \text{Socrates is a man} \\ \hline \text{Therefore, Socrates is mortal} \end{array}$$

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

$$\begin{array}{l} \text{All A are B} \\ \text{a is a A} \\ \hline \text{Therefore, a is a B} \end{array}$$

Calculating with Knowledge (ii)

- Is Ernesto and Academic?

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

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)

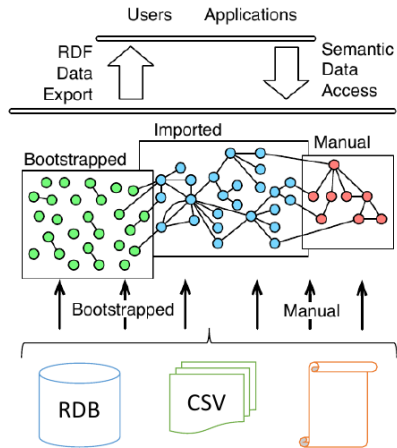
Ontologies VS relational databases

- 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.

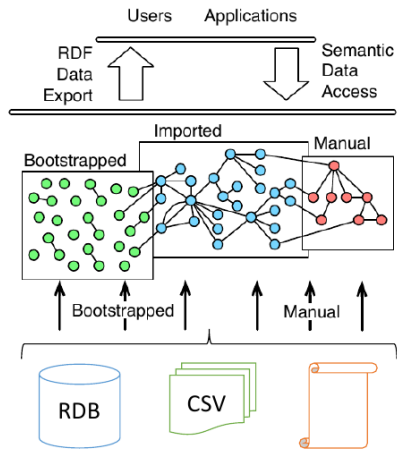
Exposing data sources as RDF data

- **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)

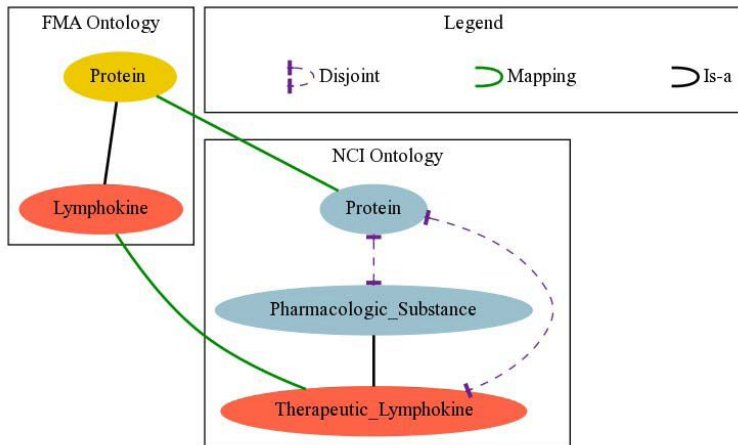


Knowledge Graph Alignment

- 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.



Big Data Borat
@BigDataBorat



Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

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 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.

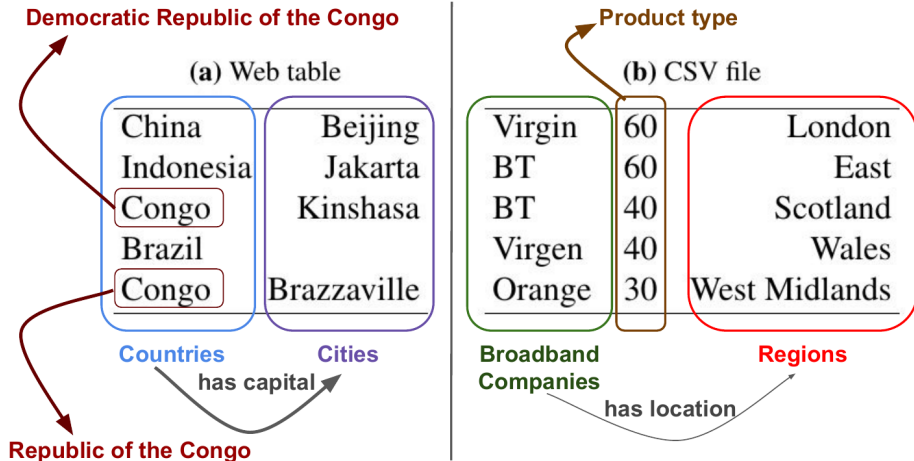
Adding Semantics to Tabular Data: Basic Tasks

- 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.*

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

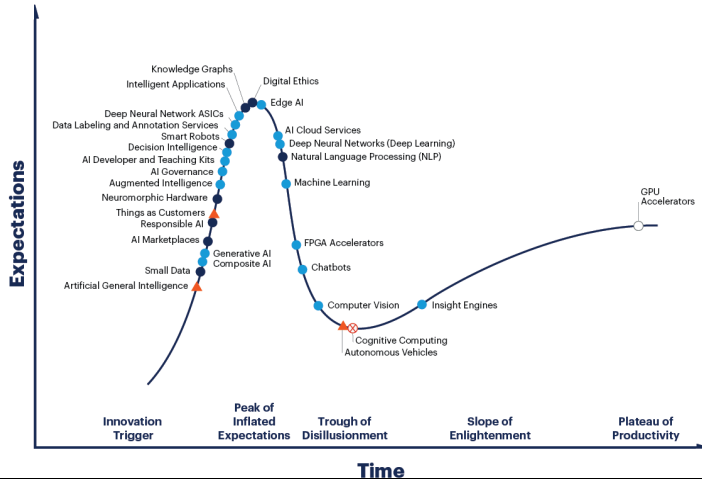
Adding Semantics to Tabular Data: Basic Tasks



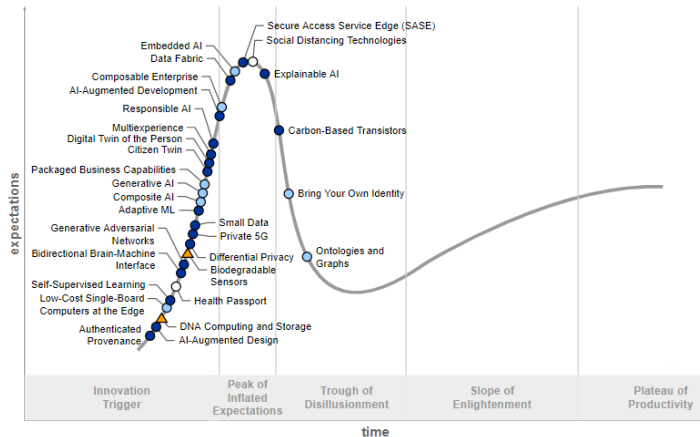
Next Decade in AI

- Gary Marcus has recently highlighted the need of **richer AI** systems, *i.e.*, **semantically sound**, **explainable**, and **reliable**.
- Impressive results in Deep Learning but require large datasets and lack explanation.
- Limitations of Knowledge representation 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 (**neuro-symbolic integration**)

Gartner's 2020 Hype Cycles: AI



Gartner's 2020 Hype Cycles: Emerging Technologies



Plateau will be reached:

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



Big Data Borat

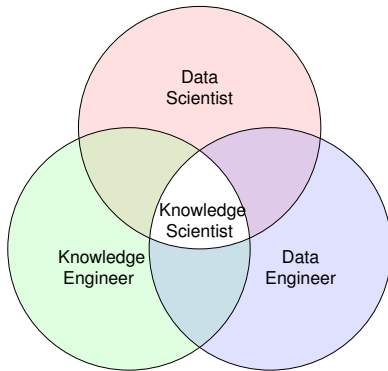
@BigDataBorat

 Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

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.

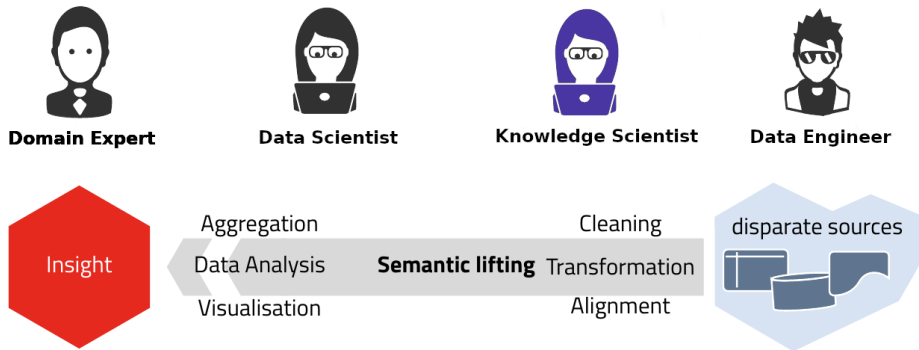


The Knowledge Scientist (ii)

- 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)



Laboratory Session

Laboratory Session

- Optional break-out rooms
- Questions in “main room”
- This session is about environment and library set-up