Engaging Content
Engaging People

Statement-Level Metadata in Knowledge Graphs

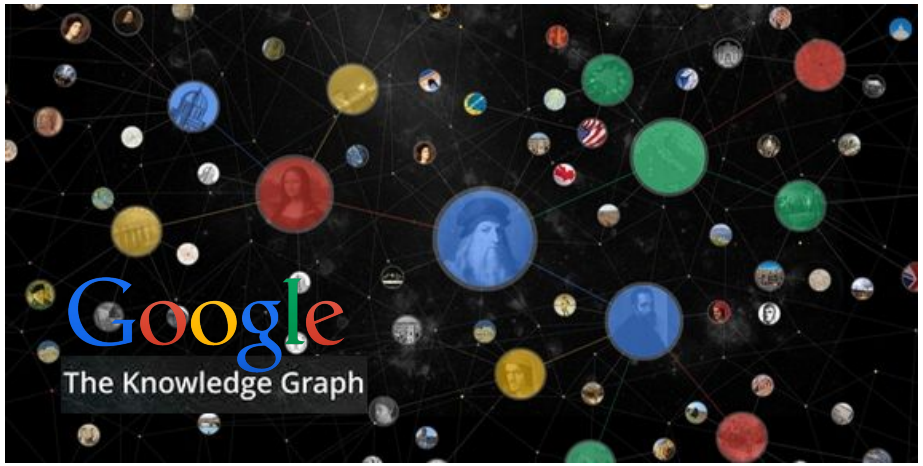
Dr. Fabrizio Orlandi

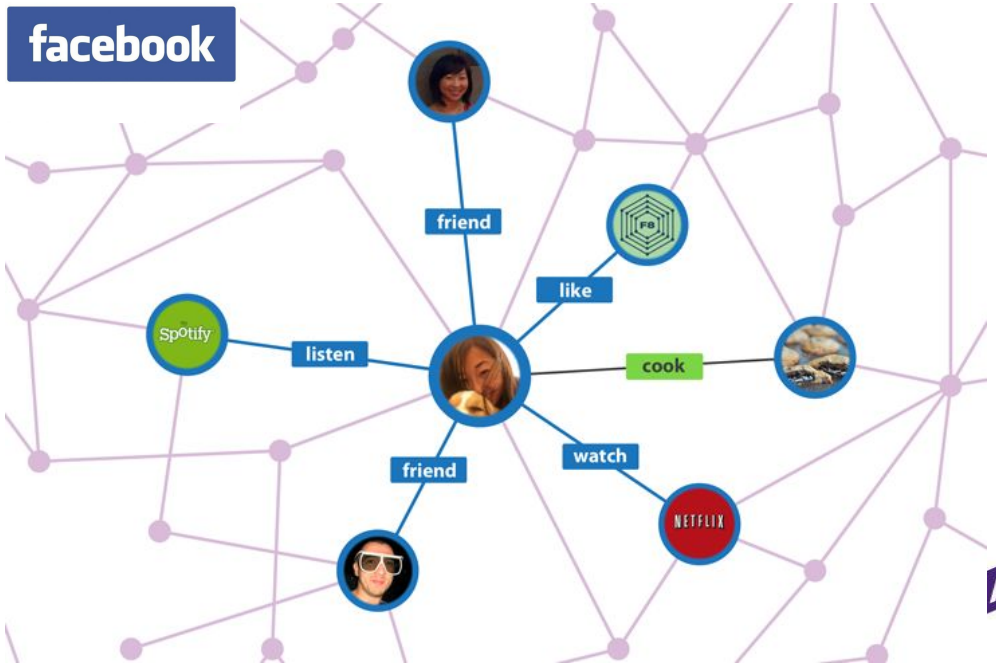
ADAPT Research Centre (TCD)

Knowledge Graphs

THOMSON REUTERS™

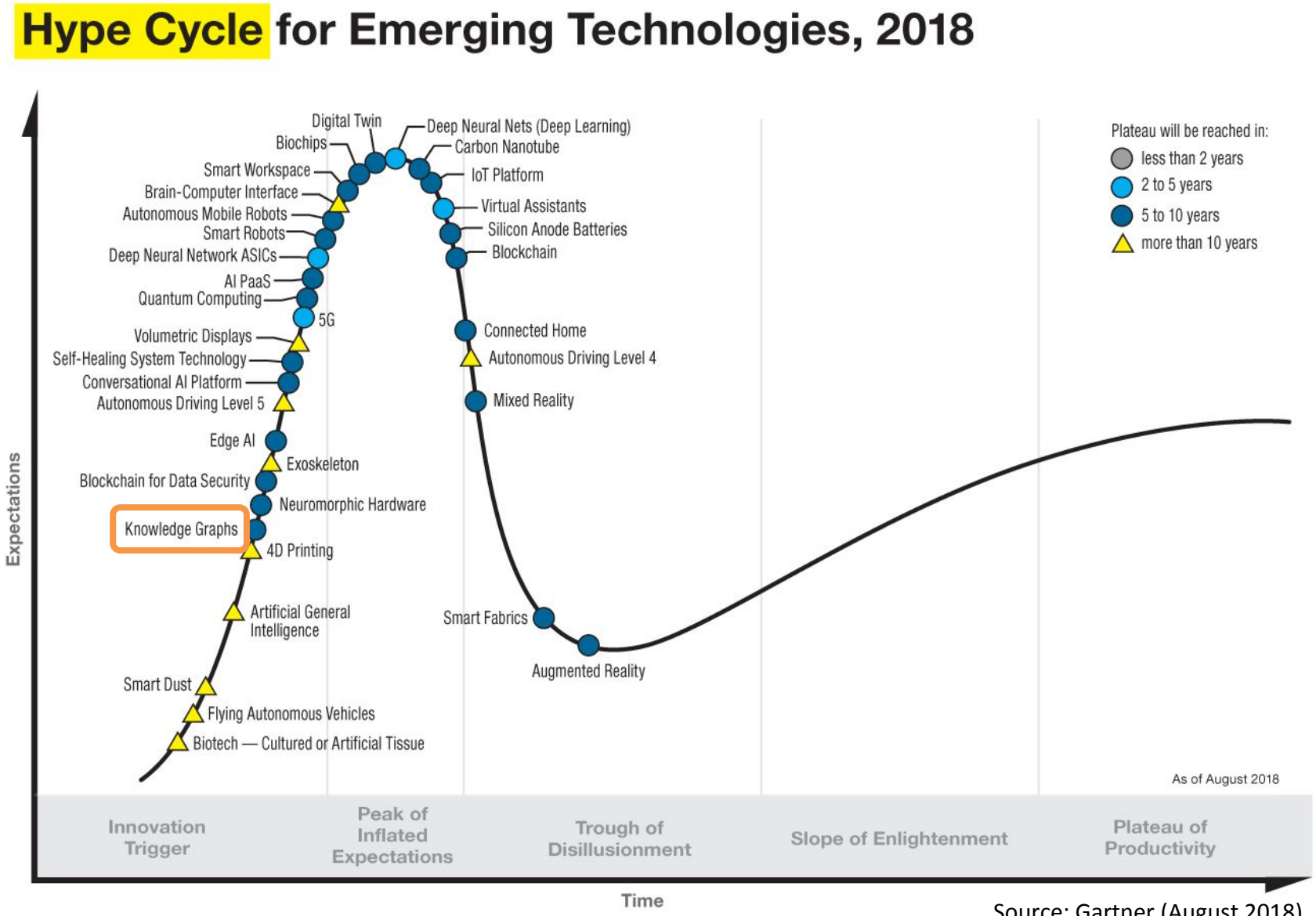
WIKIDATA





KGs in the Enterprise

- **Multi-domain applications:** Life Sciences, Economy, Sociology, Security, Libraries
- **Commercial uptake:** Since the launch of the Google KG in 2012, several companies such as Amazon, Airbnb, eBay, Elsevier, Facebook, Microsoft announced their own KGs



Source: Gartner (August 2018)

Knowledge Graphs - Example

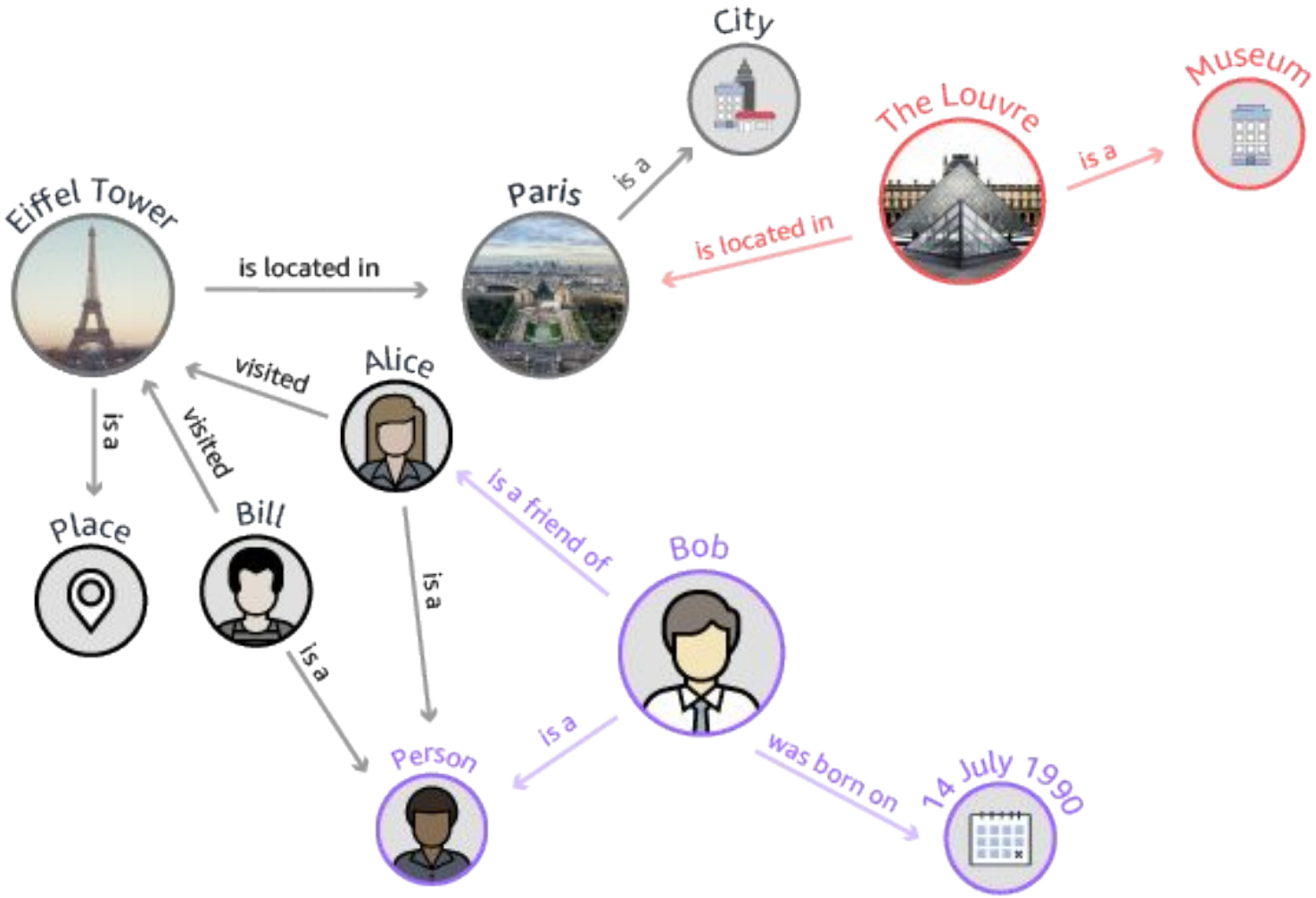
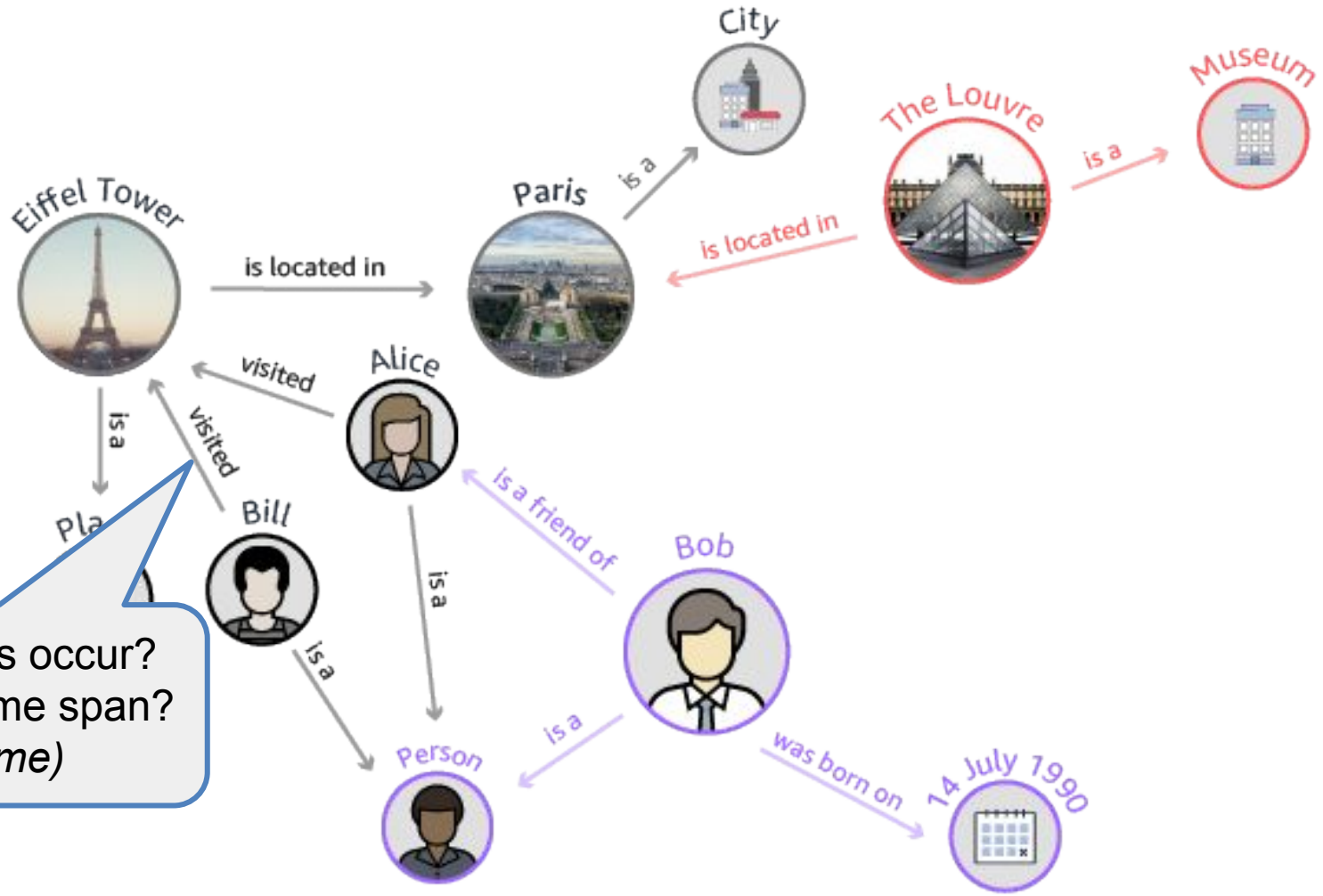


Image source: https://aws.amazon.com/neptune/

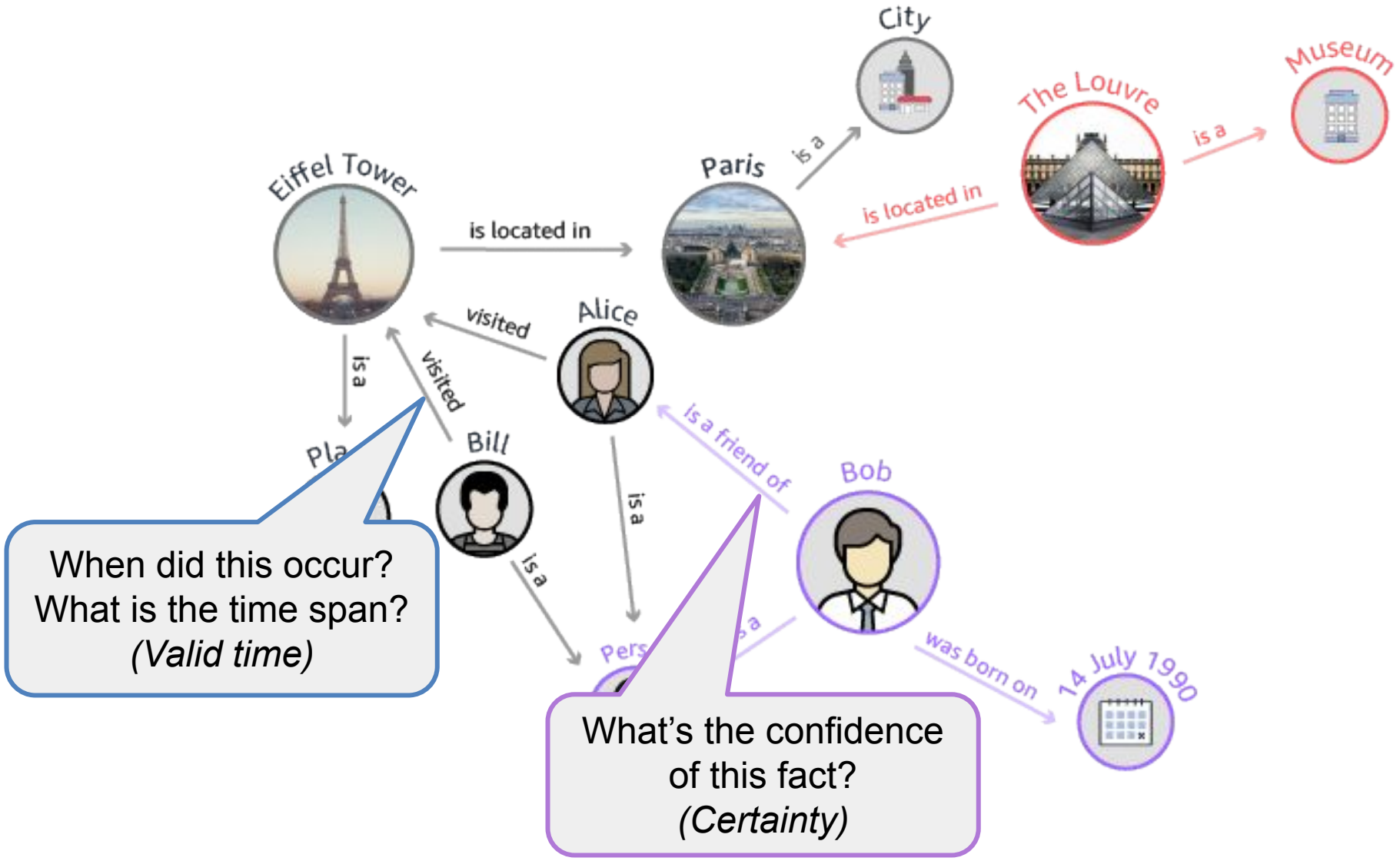
Knowledge Graphs - Example



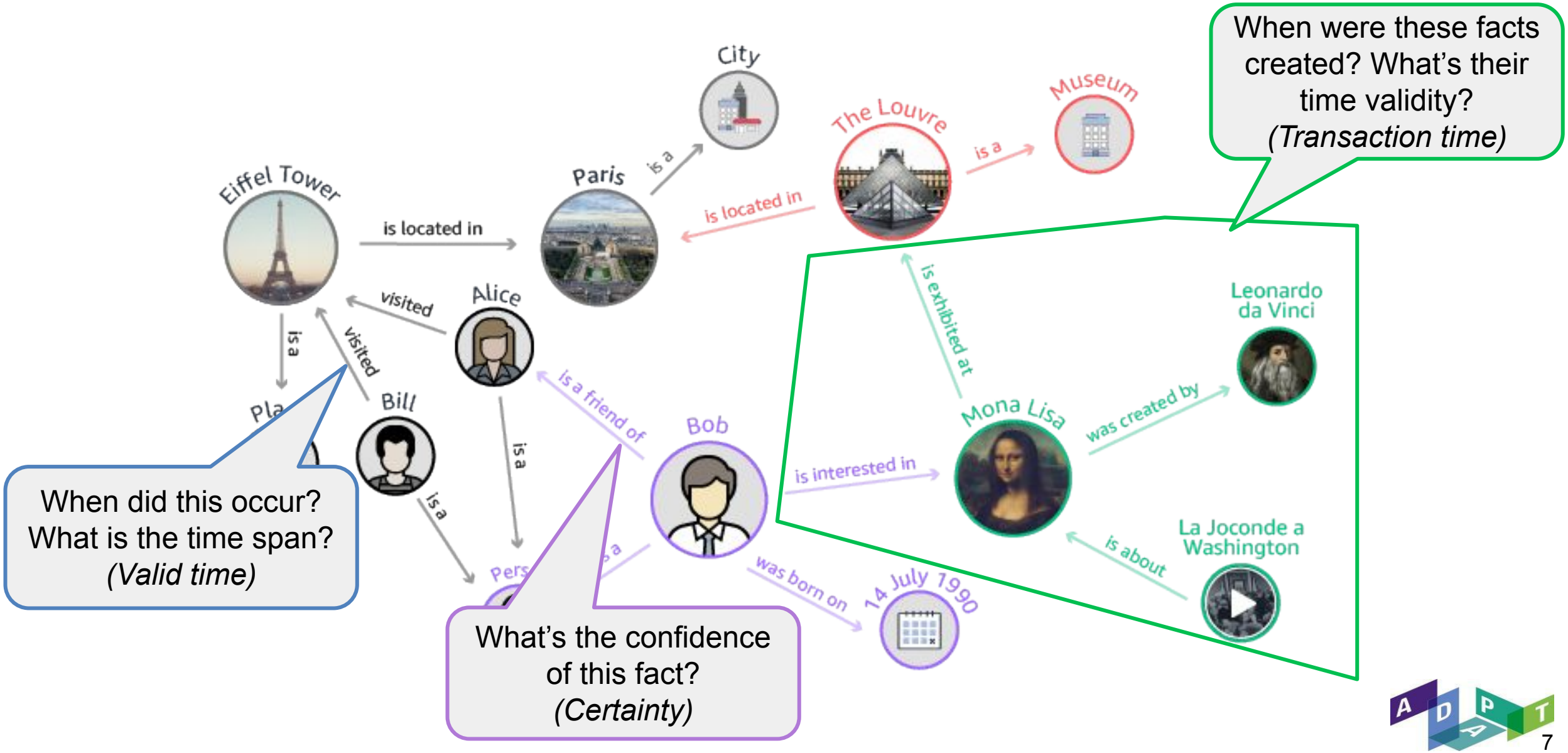
When did this occur?
What is the time span?
(Valid time)

Image source: https://aws.amazon.com/neptune/

Knowledge Graphs - Example



Knowledge Graphs - Example



Popular Use Cases for Statement-Level Metadata

- Temporal aspects of facts are usually not reflected in KGs
  (When are specific statements - triples - valid?)

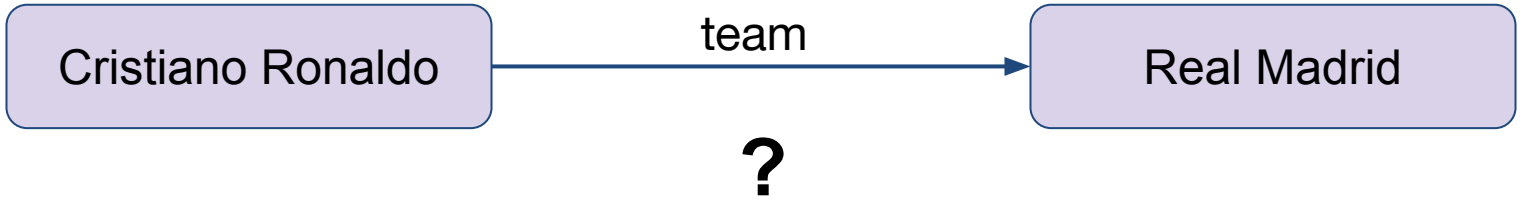
- Facts extracted from heterogeneous data sources hold different degrees of certainty, depending on the source or the extraction/generation process
  (Provenance of data)

- Missing efficient solutions for managing the dynamics (the evolution) of KGs
  (When were specific statements added/updated?)

Example of Statement-Level Metadata

Subject	Predicate	Object	Starts	Ends
Cristiano Ronaldo	team	Real Madrid	1 July 2009	10 July 2018

How to represent this in a graph?



RDF graphs vs. Property graphs

RDF Graphs

- Formally defined data model
- Various well-defined serialization formats
- Well-defined query language with a formal semantics
- Natural support for globally unique identifiers
- Semantics of data can be made explicit in the data itself
- W3C recommendations (standards!)
- High usage complexity

Labeled-Property Graphs (e.g. neo4j)

- Easy to manage statement-level metadata
- Efficient graph traversals
- Fast and scalable implementations
- No open standards defined
- Different proprietary implementations and query languages
- Good adoption in enterprise

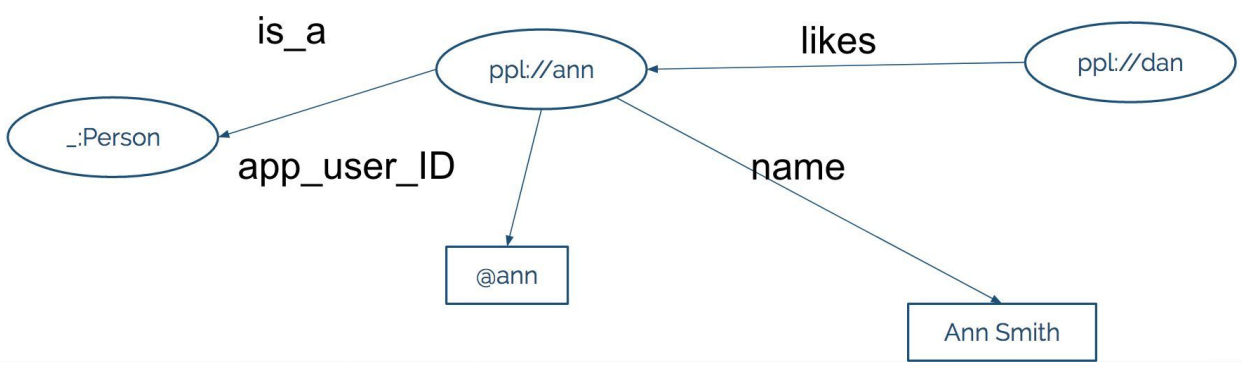
RDF graphs vs. Property graphs

RDF Graphs

Vertices
Every statement produces two vertices in the graph.
Some are uniquely identified by URIs: Resources
Some are property values: e.g. Literals

Edges
Every statement produces an edge.
Uniquely identified by URIs

Vertices or Edges have NO internal structure

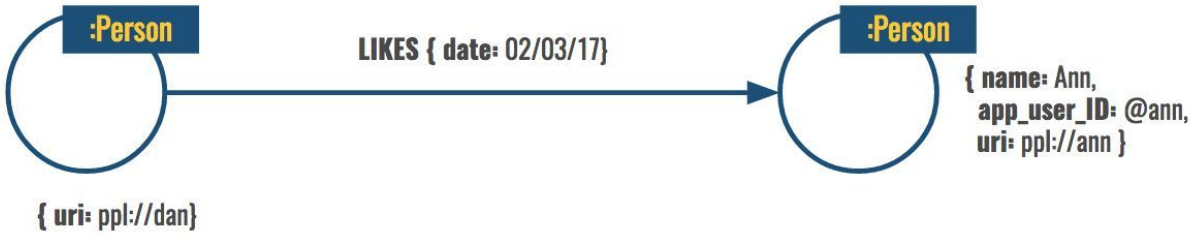


Labeled-Property Graphs (e.g. neo4j)

Vertices
Unique Id + set of key-value pairs

Edges
Unique Id + set of key-value pairs

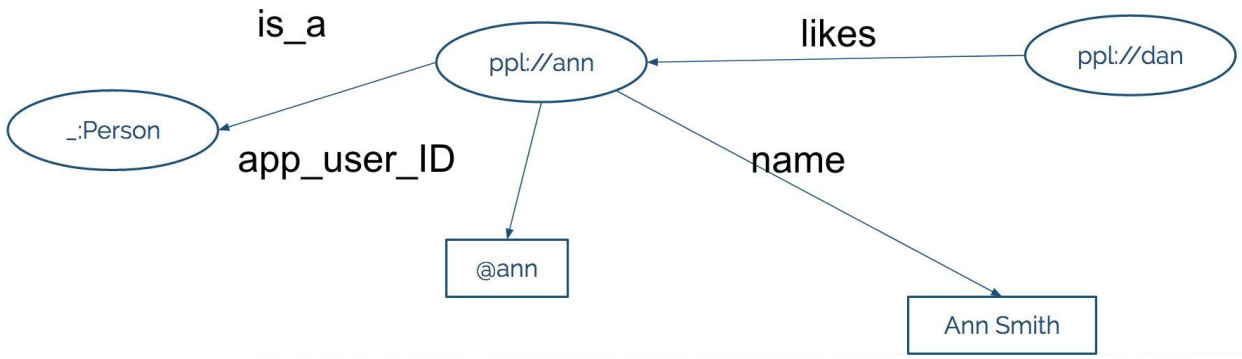
Vertices and Edges have internal structure



Query: Who likes a person named “Ann”?

SPARQL

```
SELECT ?who
WHERE
{
    ?who ms:likes ?a .
    ?a rdf:type ms:Person .
    ?a ms:name ?asName .
    FILTER regex(?asName,'Ann')
}
```

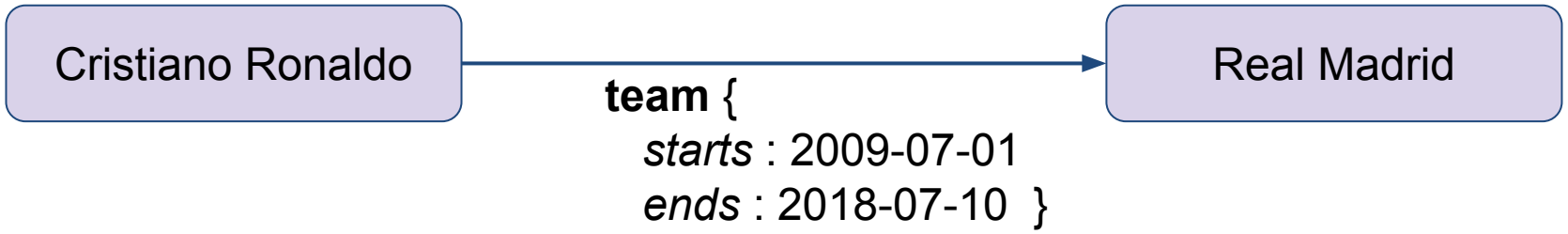


Cypher (neo4j)

```
MATCH
    (who)-[:LIKES]->(a:Person)
WHERE
    a.name CONTAINS ‘Ann’
RETURN who
```

Statement-Level Metadata with Property Graphs

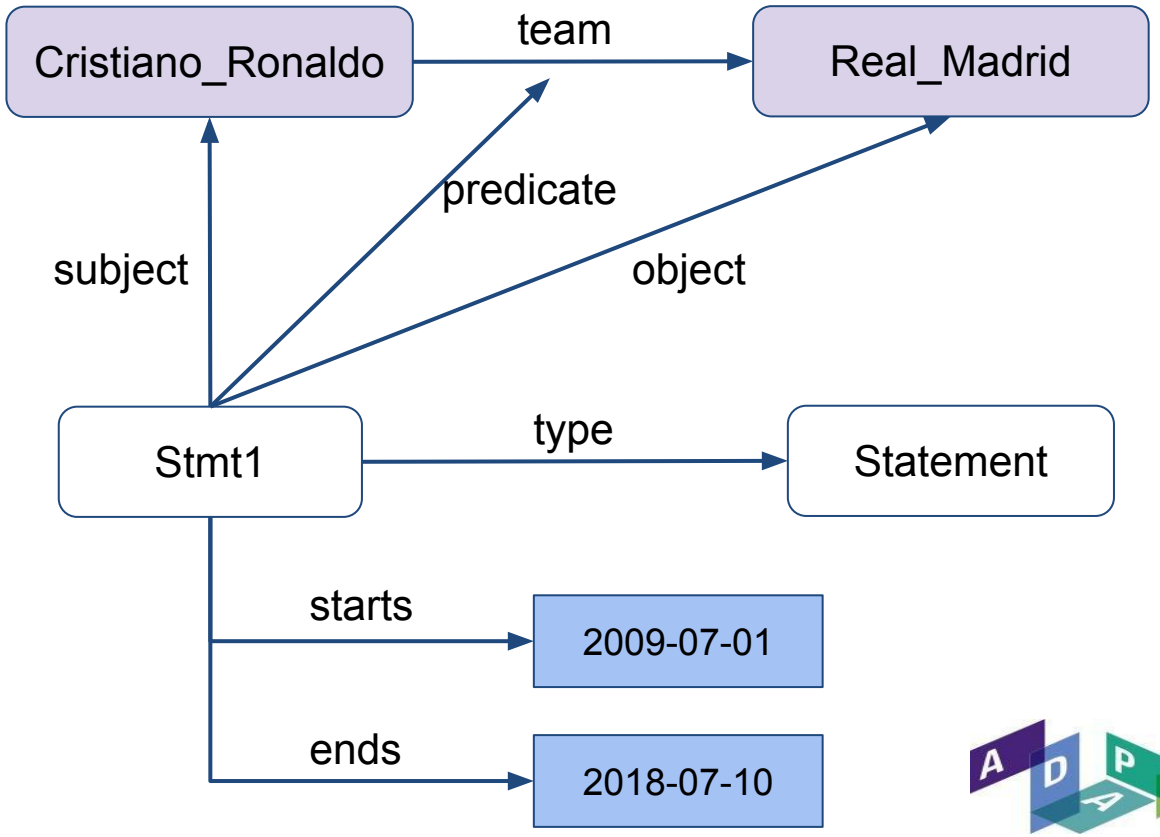
Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018



Modelling - RDF Reification

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	**Predicate**	**Object**
Cristiano_Ronaldo	team	Real_Madrid
Stmt1	type	Statement
Stmt1	subject	Cristiano_Ronaldo
Stmt1	predicate	team
Stmt1	object	Real_Madrid
Stmt1	starts	2009-07-01
Stmt1	ends	2018-07-10



Modelling - RDF Reification

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	**Predicate**	**Object**
Cristiano_Ronaldo	team	Real_Madrid
Stmt1	type	Statement
Stmt1	subject	Cristiano_Ronaldo
Stmt1	predicate	team
Stmt1	object	Real_Madrid
Stmt1	starts	2009-07-01
Stmt1	ends	2018-07-10

4N

Pros:

1. Easy to understand

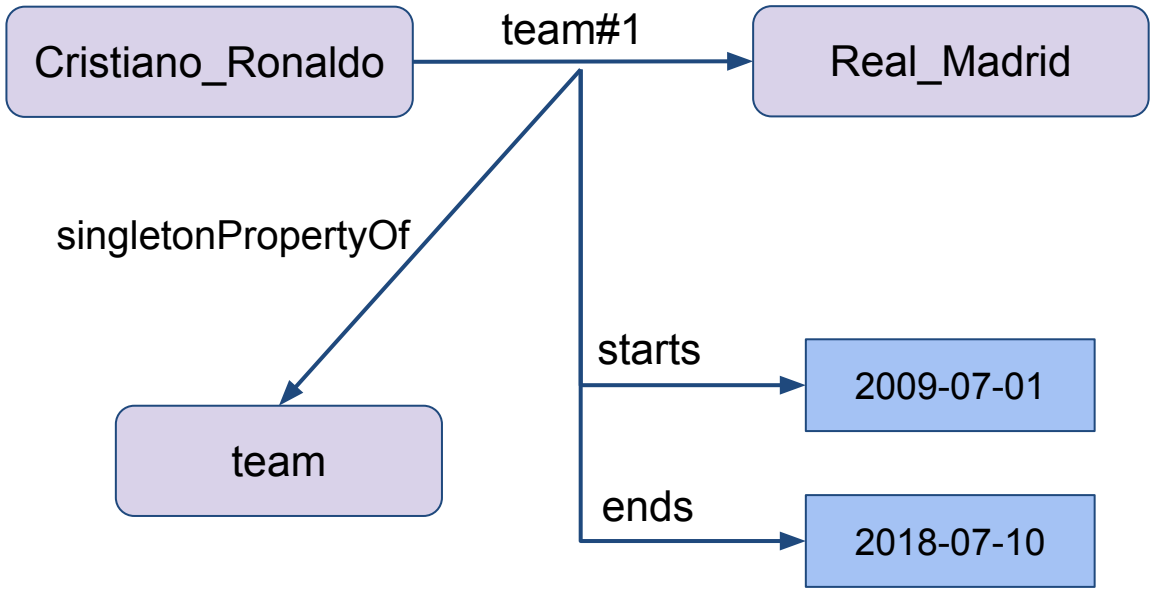
Cons:

1. Not Scalable => Takes 4N to represent a statement
2. No formal semantics defined
3. Discouraged in LOD!

Modelling - Singleton Property

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	**Predicate**	**Object**
Cristiano_Ronaldo	team#1	Real_Madrid
team#1	singletonPropertyOf	team
team#1	starts	2009-07-01
team#1	ends	2018-07-10



Vinh Nguyen, Olivier Bodenreider, and Amit Sheth. "Don't like RDF reification?: making statements about statements using singleton property." In Proceedings of the 23rd international conference on World wide web, ACM, 2014.

Modelling - Singleton Property

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	**Predicate**	**Object**
Cristiano_Ronaldo	team#1	Real_Madrid
team#1	singletonPropertyOf	team
team#1	starts	2009-07-01
team#1	ends	2018-07-10

Pros:

1. More scalable => only 1 extra triple

Cons:

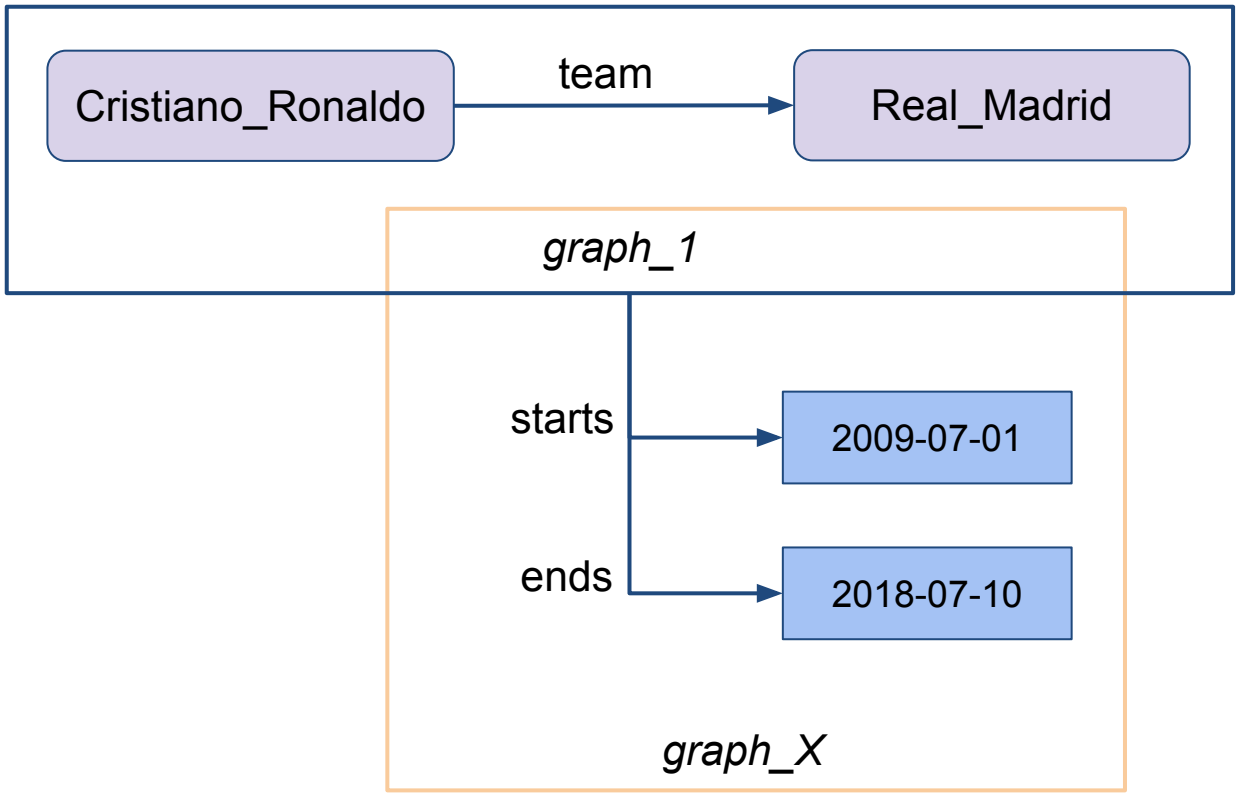
1. Less intuitive
2. Large number of unique predicates
3. Requires verbose constructs in queries

Vinh Nguyen, Olivier Bodenreider, and Amit Sheth. "Don't like RDF reification?: making statements about statements using singleton property." In Proceedings of the 23rd international conference on World wide web, ACM, 2014.

Modelling - Named Graphs (Quads)

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	Predicate	Object	NG
Cristiano_Ronaldo	team	Real_Madrid	graph_1
graph_1	starts	2009-07-01	graph_X
graph_1	ends	2018-07-10	graph_X



Carroll, Jeremy J., et al. "Named graphs, provenance and trust." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

Modelling - Named Graphs (Quads)

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

Subject	**Predicate**	**Object**	**NG**
Cristiano_Ronaldo	team	Real_Madrid	graph_1
graph_1	starts	2009-07-01	graph_X
graph_1	ends	2018-07-10	graph_X

A possible specification is N-Quads that extends N-Triples with an optional context value at the fourth position http://www.w3.org/TR/n-quads/ (W3C Recommendation)

Pros:

1. Intuitive - creates N named graphs for N sources
2. Attach metadata for a **set** of triples
3. SPARQL support

Cons:

1. Restricts usage of named graphs to provenance only
2. Requires verbose constructs in queries

Carroll, Jeremy J., et al. "Named graphs, provenance and trust." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

Modelling - RDF* and SPARQL*

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

RDF extension for nested triples:

```
<< :Cristiano_Ronaldo :team :Real_Madrid >>
        :starts “2009-07-01” ;
        :ends “2018-07-10”.
```

SPARQL extension with nested triple patterns:

```
SELECT ?player WHERE {
        << ?player :team :Real_Madrid >> :starts ?date .
        FILTER (?date >= “2009-07-01”)  }
```

Modelling - RDF* and SPARQL*

Subject	Predicate	Object	Starts	Ends
Cristiano_Ronaldo	team	Real_Madrid	1 July 2009	10 July 2018

1. **Purely syntactic "sugar" on top of standard RDF and SPARQL**
   a. Can be parsed directly into standard RDF and SPARQL
   b. Can be implemented easily by a small wrapper on top of any existing RDF DBMS
2. **A logical model in its own right, with the possibility of a dedicated physical schema**
   a. Extension of the RDF data model and of SPARQL to capture the notion of nested triples
   b. Supported by some triplestores (e.g. Blazegraph)

O Hartig: "Foundations of RDF* and SPARQL* - An Alternative Approach to Statement-Level Metadata in RDF." In Proc. of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW), 2017.

Concrete Examples...

Wikidata

https://www.wikidata.org/wiki/Q11571

English | Not logged in | Talk | Contribu

WIKIDATA

Item | Discussion | Read | View history | Search Wikidata

Subject

Cristiano Ronaldo (Q11571)

Portuguese association football player

Cristiano Ronaldo dos Santos Aveiro | CR7 | Ronaldo

In more languages

Configure

Language	Label	Description	Also known as
English	Cristiano Ronaldo	Portuguese association football player	Cristiano Ronaldo dos Santos A… CR7 Ronaldo
Italian	Cristiano Ronaldo	calciatore portoghese	Cristiano Ronaldo dos Santos A… CR7 Ronaldo
Irish	Cristiano Ronaldo	Portaingéilis peileadóir	Cristiano Ronaldo dos Santos A…
French	Cristiano Ronaldo	footballeur portugais	Cristiano Ronaldo dos Santos A… CR7

All entered languages

Predicate

Statements

instance of	human 2 references

Object

Main page
Community portal
Project chat
Create a new Item
Create a new Lexeme
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Wikipedia (129 entries)

af Cristiano Ronaldo
ak Cristiano Ronaldo
am ክርስቲያኖ ሮናልዶ
an Cristiano Ronaldo
ar كريستيانو رونالدو
arz كريستيانو رونالدو
ast Cristiano Ronaldo
azb كريستيانو رونالدو
az Kriştiano Ronaldo
bar Cristiano Ronaldo
be_x_old Крыштыяну Раналду
be Крыштыяну Раналду
bg Кристиано Роналдо
bn ক্রিস্তিয়ানো রোনাল্দো
br Cristiano Ronaldo
bs Cristiano Ronaldo
ca Cristiano Ronaldo dos Santos Aveiro
ckb کریستیانۆ ڕۆناڵدۆ
co Cristiano Ronaldo
cs Cristiano Ronaldo
cv Криштиану Роналду

Wikidata

Predicate

member of sports team

Manchester United F.C.

start time	2003
end time	1 July 2009
number of matches played/races/starts	196
number of points/goals/set scored	84
acquisition transaction	transfer
departure transaction	transfer

2 references

Object

Real Madrid CF

start time	1 July 2009
number of matches played/races/starts	292
number of points/goals/set scored	311
end time	10 July 2018
acquisition transaction	transfer
departure transaction	transfer

1 reference

Statement-level Metadata (Wikidata "Qualifiers")

Wikidata

Wikidata Query Service | Examples | Help | More tools | English

```
SELECT ?team ?teamLabel ?starttime ?endtime
WHERE
{
     wd:Q11571 p:P54 ?statement.
      ?statement ps:P54 ?team.
      ?statement pq:P580 ?starttime.
      ?statement pq:P582 ?endtime.
     ?team rdfs:label ?teamLabel.
     FILTER(LANG(?teamLabel) = "en")
}
ORDER BY ?starttime
```

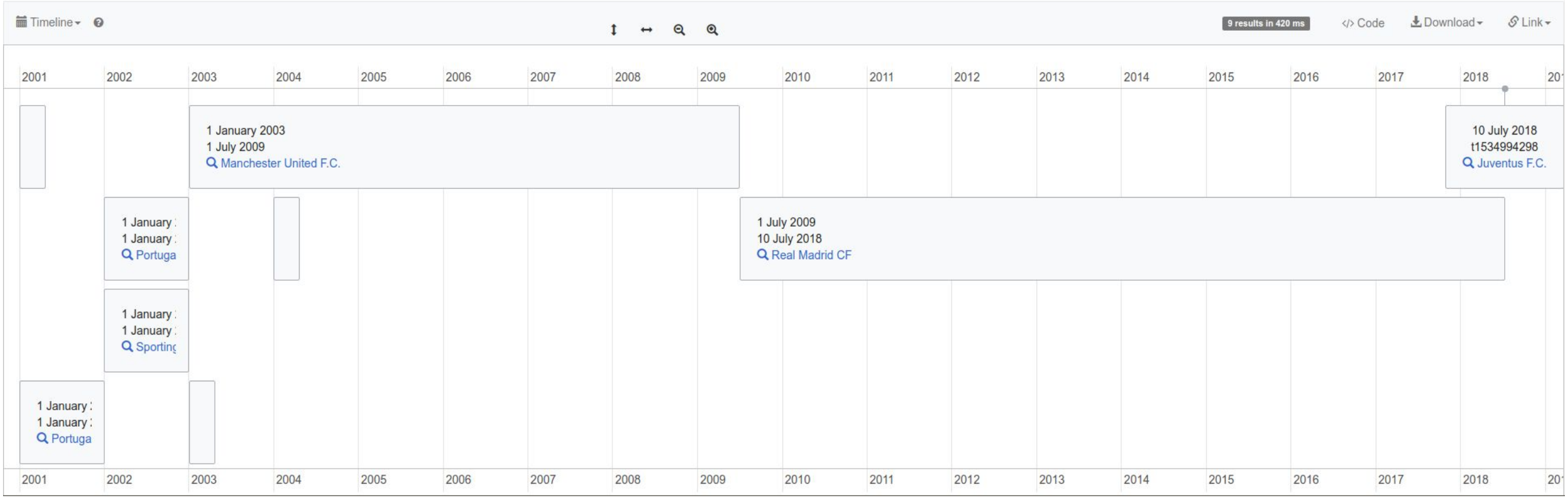
Table | 9 results in 420 ms | Code | Download | Link

team	teamLabel	starttime	endtime
wd:Q3590754	Portugal national under-17 football team	1 January 2001	1 January 2002
wd:Q21079208	Portugal national under-15 football team	1 January 2001	1 January 2001
wd:Q75729	Sporting CP	1 January 2002	1 January 2003
wd:Q1630430	Portugal national under-21 football team	1 January 2002	1 January 2003
wd:Q18656	Manchester United F.C.	1 January 2003	1 July 2009
wd:Q1772776	Portugal national under-20 football team	1 January 2003	1 January 2003
wd:Q3590758	Portugal Olympic football team	1 January 2004	1 January 2004
wd:Q8682	Real Madrid CF	1 July 2009	10 July 2018
wd:Q1422	Juventus F.C.	10 July 2018	t1534994298

```
SELECT ?team ?teamLabel ?starttime ?endtime
WHERE
{
     wd:Q11571 p:P54 ?statement.
      ?statement ps:P54 ?team.
      ?statement pq:P580 ?starttime.
      ?statement pq:P582 ?endtime.
     ?team rdfs:label ?teamLabel.
     FILTER(LANG(?teamLabel) = "en")
}
ORDER BY ?starttime
```

Try it out at https://query.wikidata.org/

(or directly at: https://w.wiki/BWZ)



Additional examples...

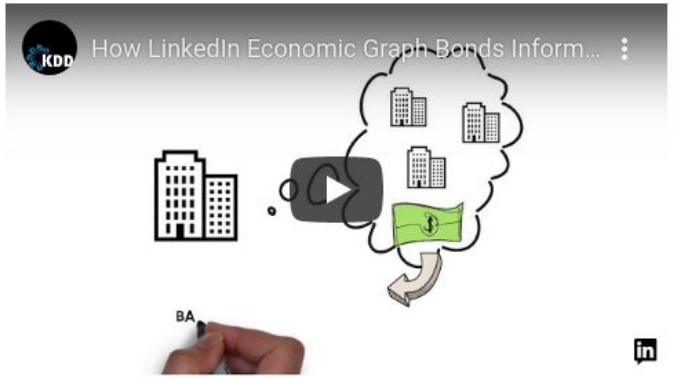
https://engineering.linkedin.com/blog/2018/12/using-economic-graph-data-to-power-the-linkedin-salary-product

LinkedIn Engineering Home Blog Data Open Source Jobs Women in Tech

Using Economic Graph Data to Power the LinkedIn Salary Product

Xi Chen December 14, 2018 Share Tweet Share

Co-authors: Xi Chen, Yiqun Liu, Liang Zhang, and Krishnaram Kenthapadi



Online professional social networks and job platforms, such as LinkedIn, play a key role in ensuring an efficient labor marketplace by connecting talent (job seekers) with opportunities (jobs). Studies show that salary is an important factor when looking for new opportunities, but salary information isn't always as readily apparent as, say, the job location. Products such as LinkedIn Salary have the potential to reduce asymmetry of compensation knowledge, and to serve as market-perfecting tools for job seekers and job providers.

from their ACM KDD 2018 paper

Additional examples...

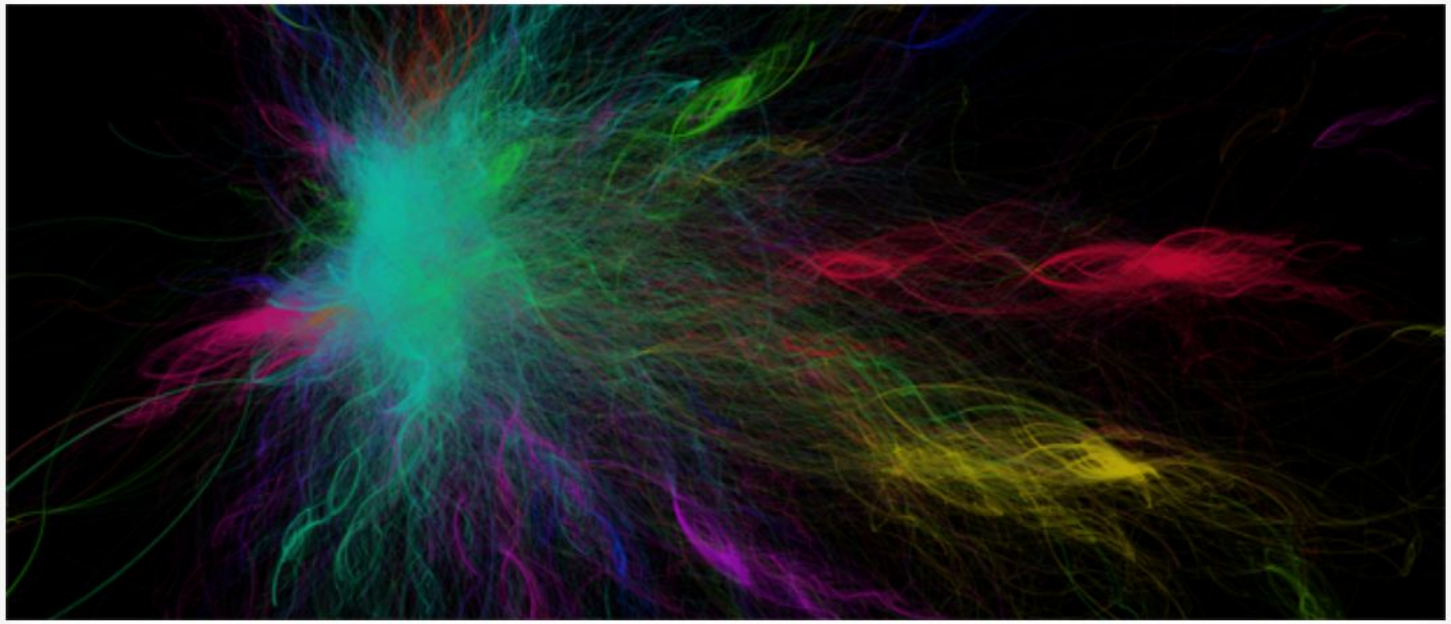
https://blog.gdeltproject.org/gdelt-global-knowledge-graph/

The GDELT Project

THE GDELT PROJECT BLOG | WEBSITE

GDELT Global Knowledge Graph

JANUARY 25, 2014



We are tremendously excited to announce the debut of the GDELT Global Knowledge Graph (GKG), which expands GDELT's ability to quantify global human society beyond cataloging physical occurrences towards actually representing all of the latent dimensions, geography, and network structure of the global news. To sum up the Global Knowledge Graph in a single sentence, it attempts to connect every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day.

The Global Knowledge Graph actually consists of two parallel data streams. The first is the daily Counts File, which records mentions of counts of things with respect to a set of predefined categories such as a number of protesters, a number killed, or a number displaced or sickened. Such counts may occur independently

Research in our group…

How can we effectively represent and manage temporal dynamics and uncertainty of facts in knowledge graphs?

Current activities:

- Model and characterise facts in KGs according to temporal and uncertainty aspects
- Develop solutions for real-time processing, update and propagation of changes in KGs
- Evaluate the developed solutions, applying them to different use cases

Some Industrial Use-Cases

1) Finance (temporal aspects)

Data about companies, their shares & market is complex, available and very time-dependent

2) Law / Court Cases (uncertainty)

Legal search and Q&A systems on large corpora of court cases need the uncertainty dimension for their different information extraction systems

3) News & Social Media (dynamics)

Very time-dependent & uncertain data which needs an efficient management solution for its dynamics

Engaging Content
Engaging People

Questions?