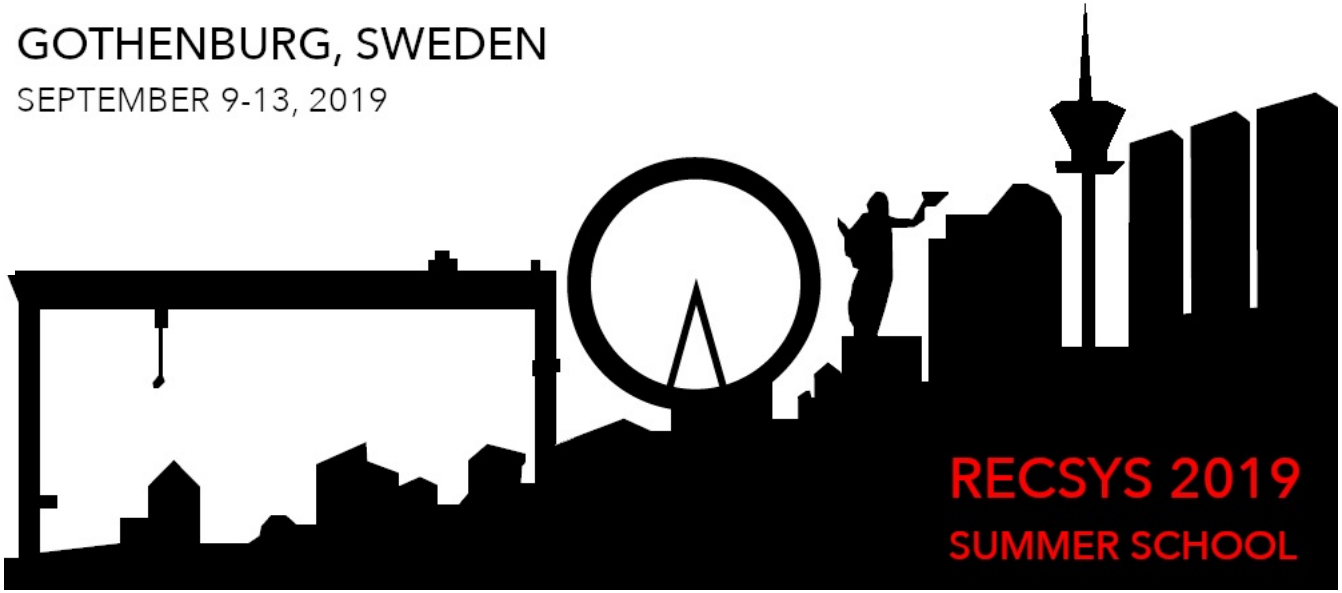
GOTHENBURG, SWEDEN
SEPTEMBER 9-13, 2019



Advances in Content-based RecSys

Semantic Representations

Pasquale Lops

Department of Computer Science
University of Bari Aldo Moro, Italy

Semantic representations

Endogenous
(implicit semantics)

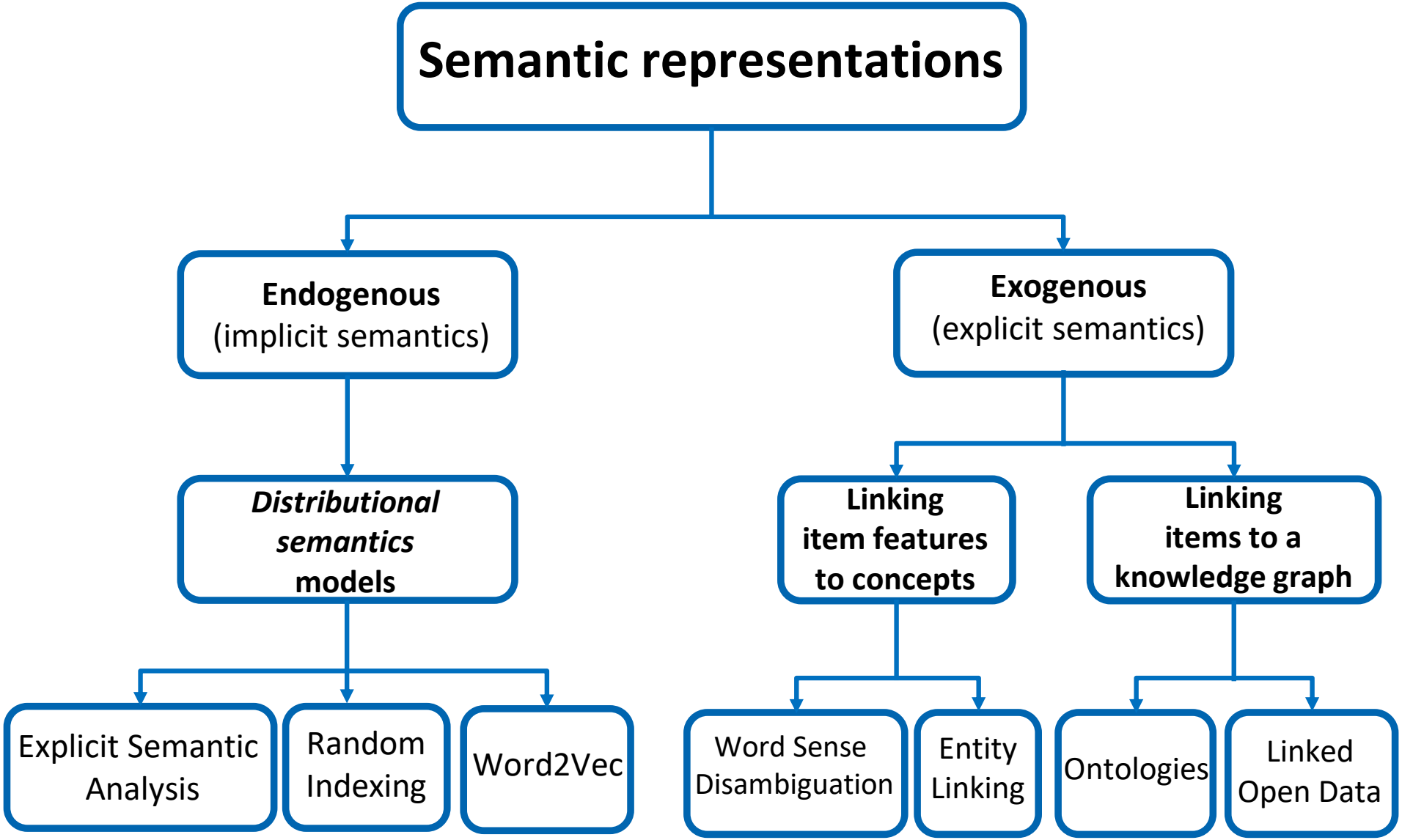
Exogenous
(explicit semantics)

bottom-up

approaches that determine the **meaning of a word** by analyzing the **rules of its usage** in **large corpora of textual content**

top-down

approaches based on the integration of **external knowledge** for representing content.
Able to provide the **linguistic, cultural** and **backgroud knowledge** in the **content representation**


Semantic representations
Endogenous
(implicit semantics)
Exogenous
(explicit semantics)
Distributional semantics models
Linking item features to concepts
Linking items to a knowledge graph
Explicit Semantic Analysis
Random Indexing
Word2Vec
Word Sense Disambiguation
Entity Linking
Ontologies
Linked Open Data

Agenda

Why?

Why do we need **content**?
Why do we need **semantics**?

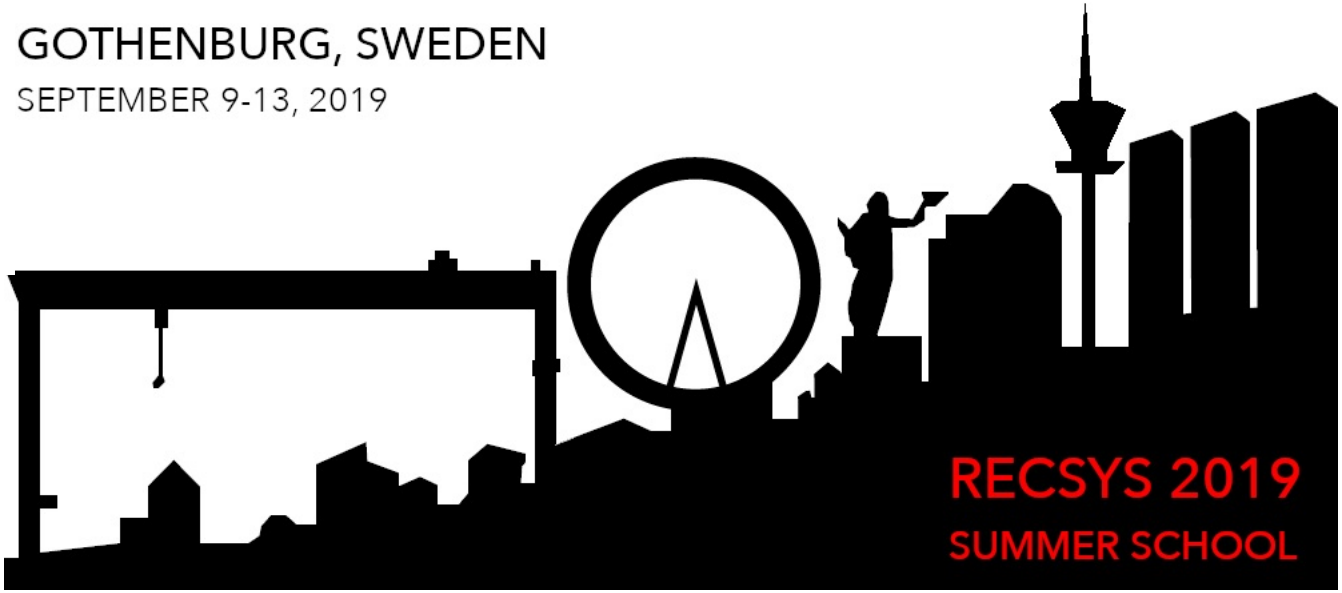
How?

How to **introduce semantics**?
Basics of **Natural Language Processing**
Encoding **exogenous semantics**,i.e. *explicit* semantics
Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

Explanation of Recommendations
Conversational Recommender Systems (hands-on)

GOTHENBURG, SWEDEN
SEPTEMBER 9-13, 2019



Advances in Content-based RecSys

Encoding Exogenous Semantics

Pasquale Lops

Department of Computer Science
University of Bari Aldo Moro, Italy

"Knowledge is power" hypothesis

The power of an intelligent program
to perform its task well depends primarily
on the **quantity** and **quality** of
knowledge it has
about that task

crucial ingredient for intelligent information
access is the availability of **machine-readable knowledge**

Buchanan BG, Feigenbaum EA (1982) Forward. Knowl Based Syst Artif Intell

Knowledge sources

unstructured

dmoz
open directory project

WIKIPEDIA
The Free Encyclopedia

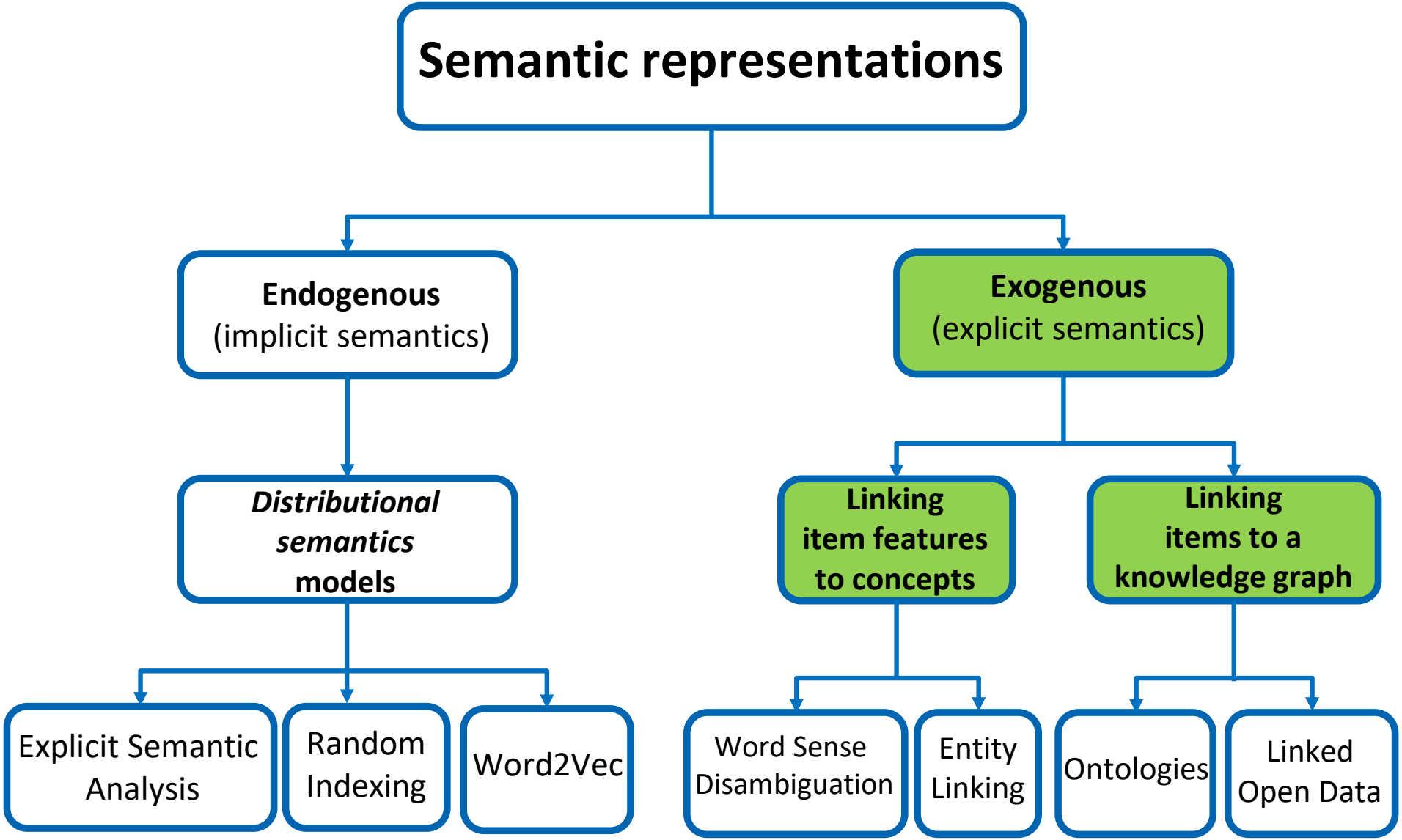
structured

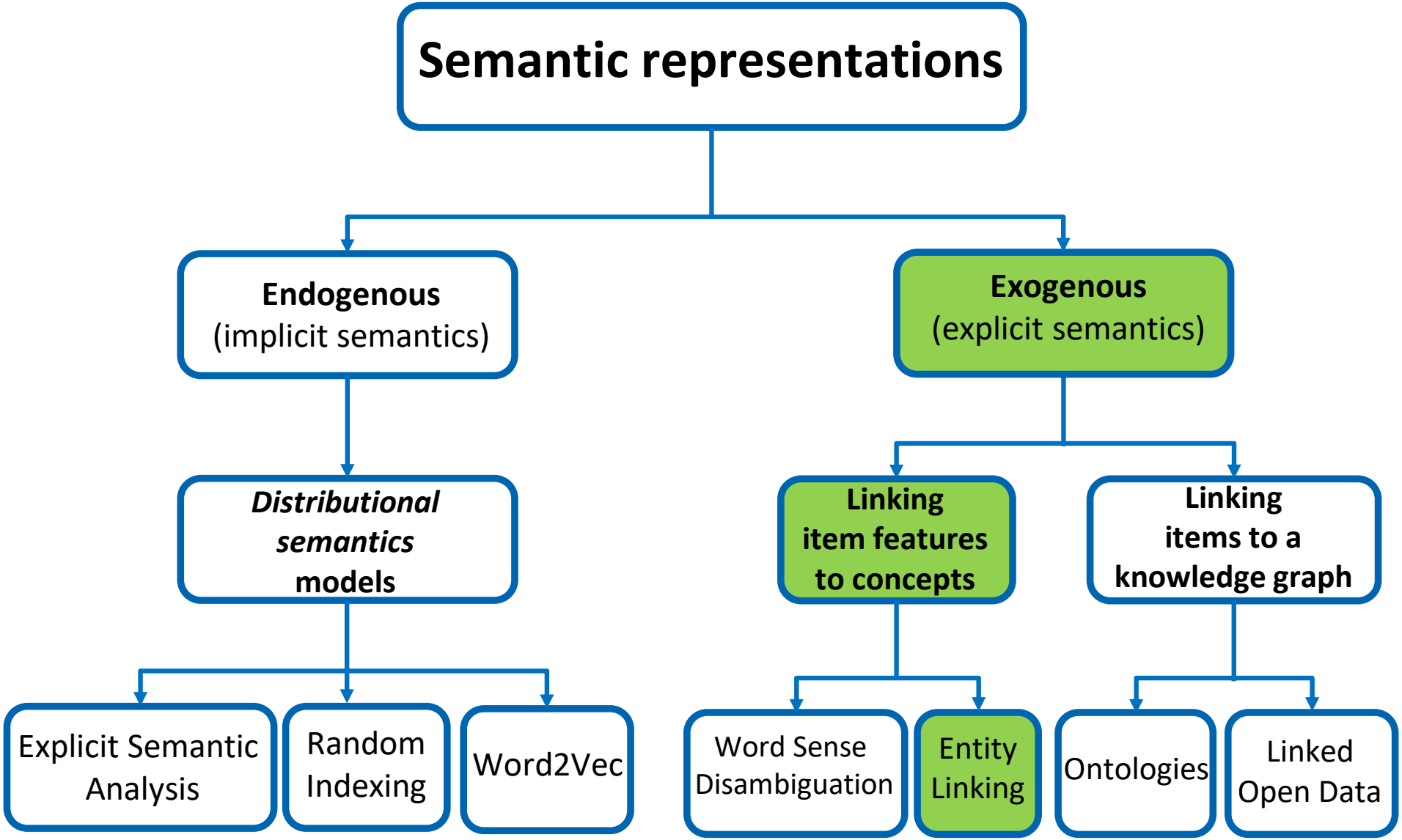
DBpedia

Freebase

WIKIDATA

BabelNet


Semantic representations
Endogenous
(implicit semantics)
Exogenous
(explicit semantics)
Distributional semantics models
Linking item features to concepts
Linking items to a knowledge graph
Explicit Semantic Analysis
Random Indexing
Word2Vec
Word Sense Disambiguation
Entity Linking
Ontologies
Linked Open Data


Semantic representations
Endogenous
(implicit semantics)
Exogenous
(explicit semantics)
Distributional semantics models
Linking item features to concepts
Linking items to a knowledge graph
Explicit Semantic Analysis
Random Indexing
Word2Vec
Word Sense Disambiguation
Entity Linking
Ontologies
Linked Open Data

Entity Linking

Associating the **mention** of an entity in a text to an entity of the real world stored in a **knowledge base**

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American-Australian science fiction action film written and directed by The Wachowski Brothers, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[4] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[5] Jean Baudrillard's *Simulacra and Simulation*[6] and

The Matrix

Theatrical release poster

Input: free text

e.g. Wikipedia abstract

Output: **entities** mentioned in the text.

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

Why Entity Linking?

because **we need to identify the entities**
mentioned in the textual description
to better catch user preferences
and information needs.

Several state-of-the-art implementations are already available

TAG·ME

Babelfy

… and many more

Entity Linking methodology

- entity **recognition**, aka **mention detection**
- identification of **candidate entities** in the **knowledge base**
- entity **disambiguation**, if there is more than one candidate entity for a given mention

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

Output

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

very **transparent** and **human-readable** content representation

non-trivial NLP tasks automatically performed

(stopwords removal, n-grams identification,
named entities recognition and disambiguation)

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

Output

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

Advantage #1: several common sense concepts are now identified

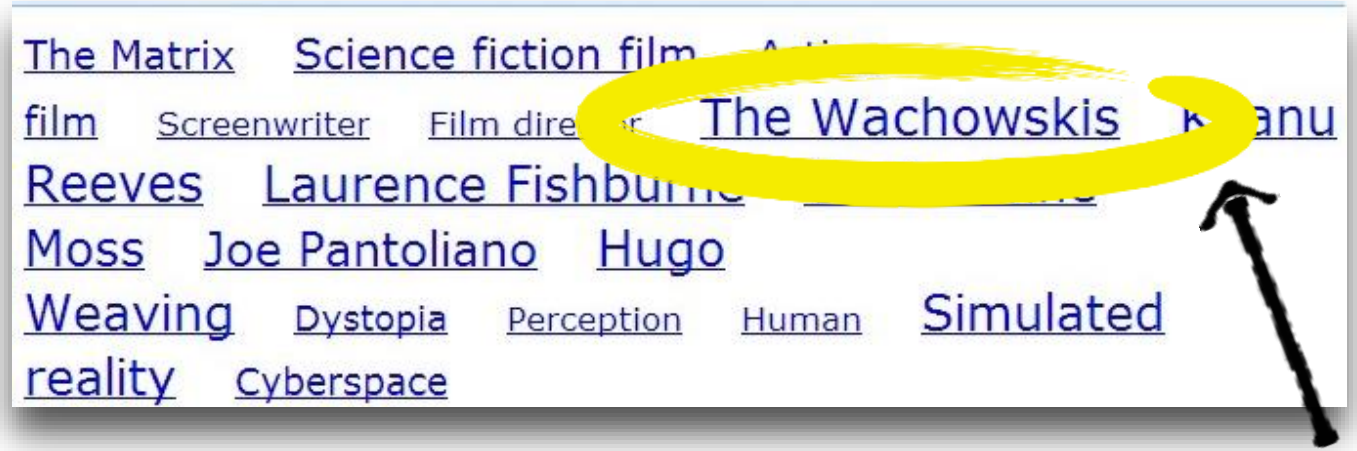
Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

Output



not a simple textual feature!

Advantage #2: each **entity** is a reference to a **Wikipedia page**

http://en.wikipedia.org/wiki/The_Wachowskis

Entity Linking Algorithms

Tag.me + Wikipedia Categories

https://tagme.d4science.org/tagme/

The Wachowskis

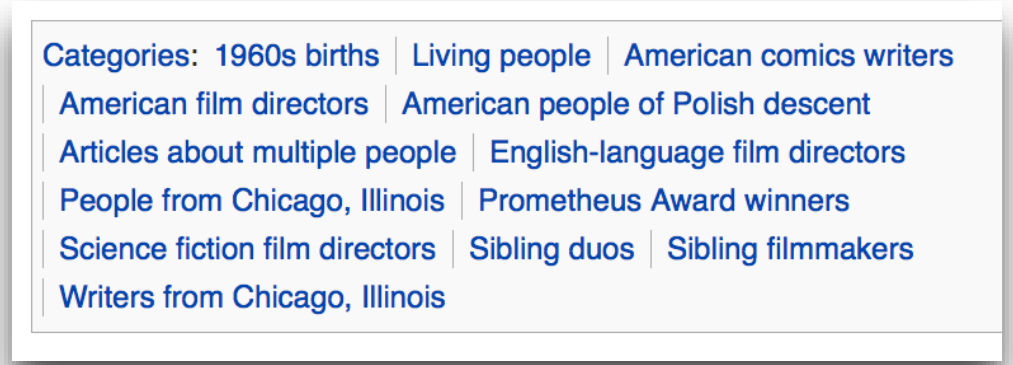
From Wikipedia, the free encyclopedia

Lana Wachowski (born **Laurence "Larry" Wachowski**; June 21, 1965) and **Andrew Paul "Andy" Wachowski** (born December 29, 1967), known together professionally as **the Wachowskis** and formerly as the **Wachowski Brothers**, are American film directors, screenwriters and producers. [5]

They made their directing debut in 1996 with *Bound*, and reached fame with their second film *The Matrix* (1999), for which they won the Saturn Award for Best Director. They wrote and directed its two sequels *The Matrix Reloaded* and *The Matrix Revolutions* (both in 2003), and were heavily involved in the writing and production of other works in the franchise.

Lana Wachowski

Andy (left) and Lana Wachowski in September 2012, at the Fantastic Fest screening of *Cloud Atlas*.



Categories: 1960s births | Living people | American comics writers | American film directors | American people of Polish descent | Articles about multiple people | English-language film directors | People from Chicago, Illinois | Prometheus Award winners | Science fiction film directors | Sibling duos | Sibling filmmakers | Writers from Chicago, Illinois

We **can enrich** this **entity-based representation** by exploiting the **Wikipedia categories' tree**

Entity Linking Algorithms

Tag.me + Wikipedia Categories

https://tagme.d4science.org/tagme/

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

features = entities + wikipedia categories

Categories: 1960s births | Living people | American comics writers | American film directors | American people of Polish descent | Articles about multiple people | English-language film directors | People from Chicago, Illinois | Prometheus Award winners | Science fiction film directors | Sibling duos | Sibling filmmakers | Writers from Chicago, Illinois

final representation of items obtained by merging **entities** identified in the text with the **(most relevant) Wikipedia categories** each entity is linked to

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

Output

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

entities and Wikipedia categories identified in the content can be features of a **semantics-aware content representation based on entity linking**

The Matrix representation

Matrix

1999

American

Australian

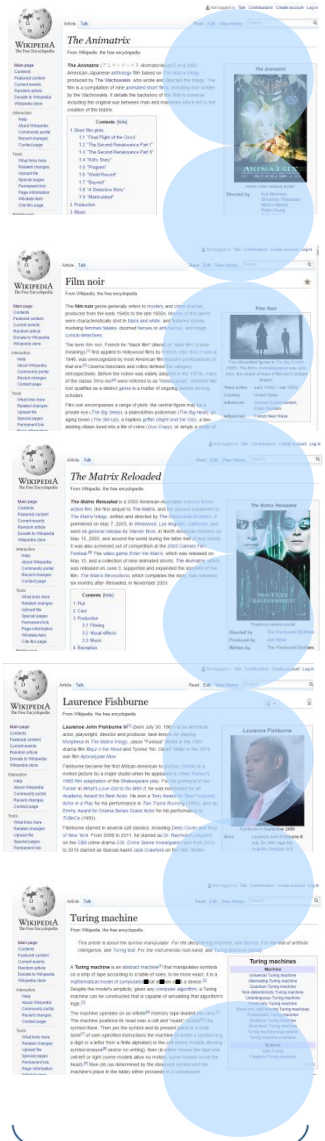
neo

science

fiction

…

world



keywords

Wikipedia pages

The Matrix representation

Matrix

1999

American

Australian

neo

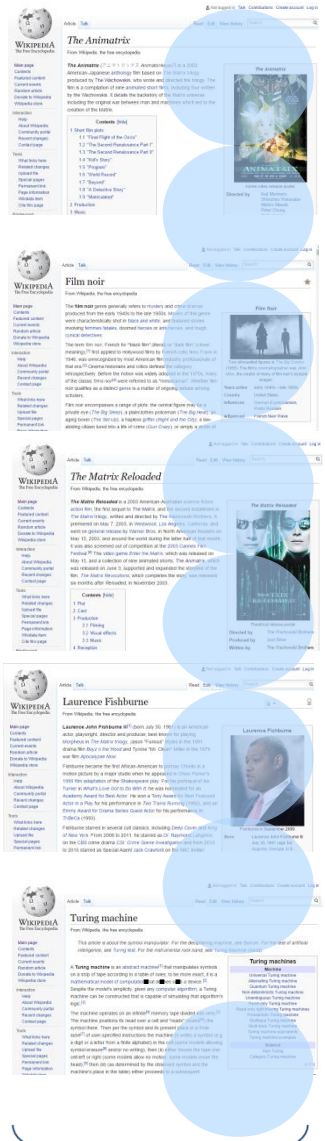
science

fiction

…

world

keywords



Wikipedia pages

entities recognized and modeled

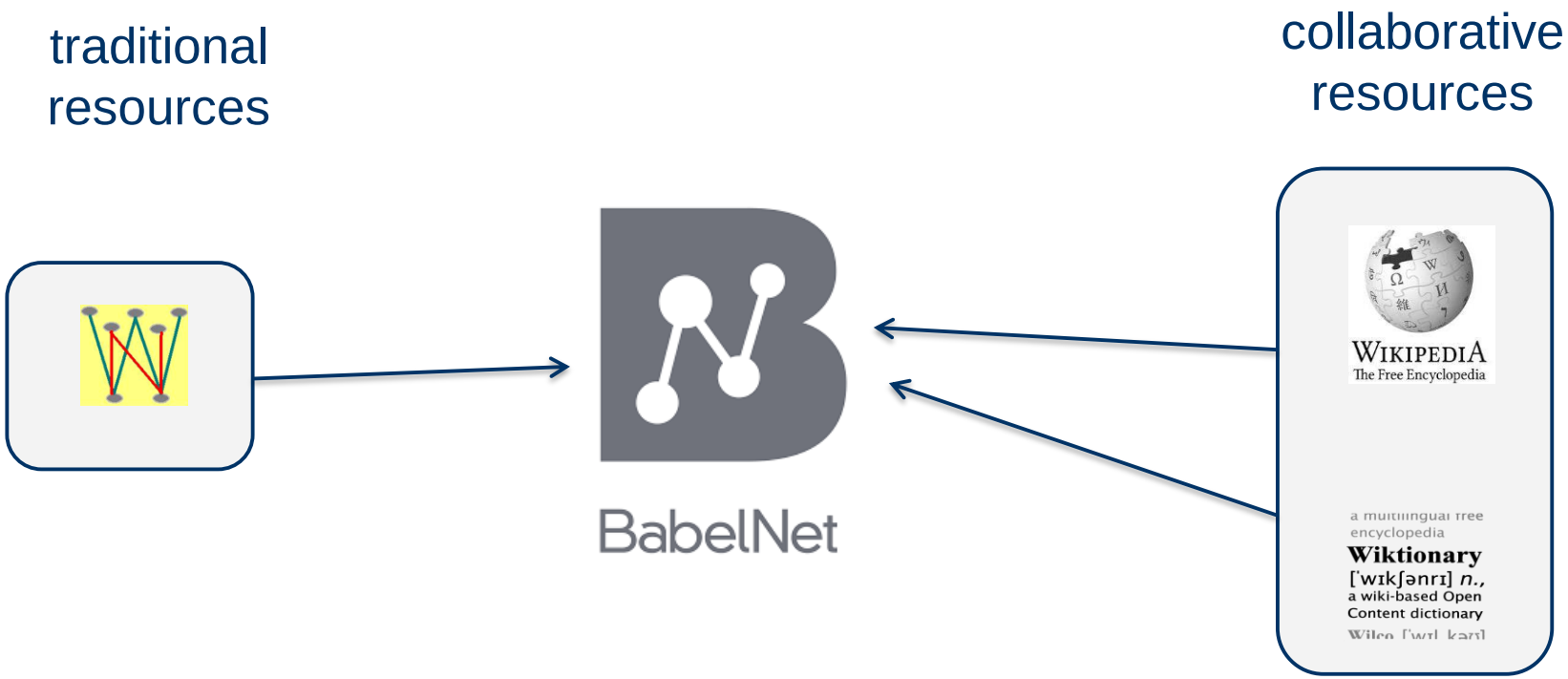
Wikipedia-based representation: some common sense terms included, and new interesting features (e.g. science fiction film, film director») can be generated

terms without a Wikipedia mapping are ignored

Entity Linking Algorithms

Babelfy

http://babelfy.org/



- manually curated by experts
- available for a few languages
- difficult to maintain and update

- collaboratively built by the crowd
- highly multilingual
- up-to-date

BabelNet

CERCA, TRADUCI, IMPARA!

Scrivi un termine o un testo... | ITALIAN | TRADUCI IN... | CERCA

BABELNET IN TIME MAGAZINE!

PREFERENZE

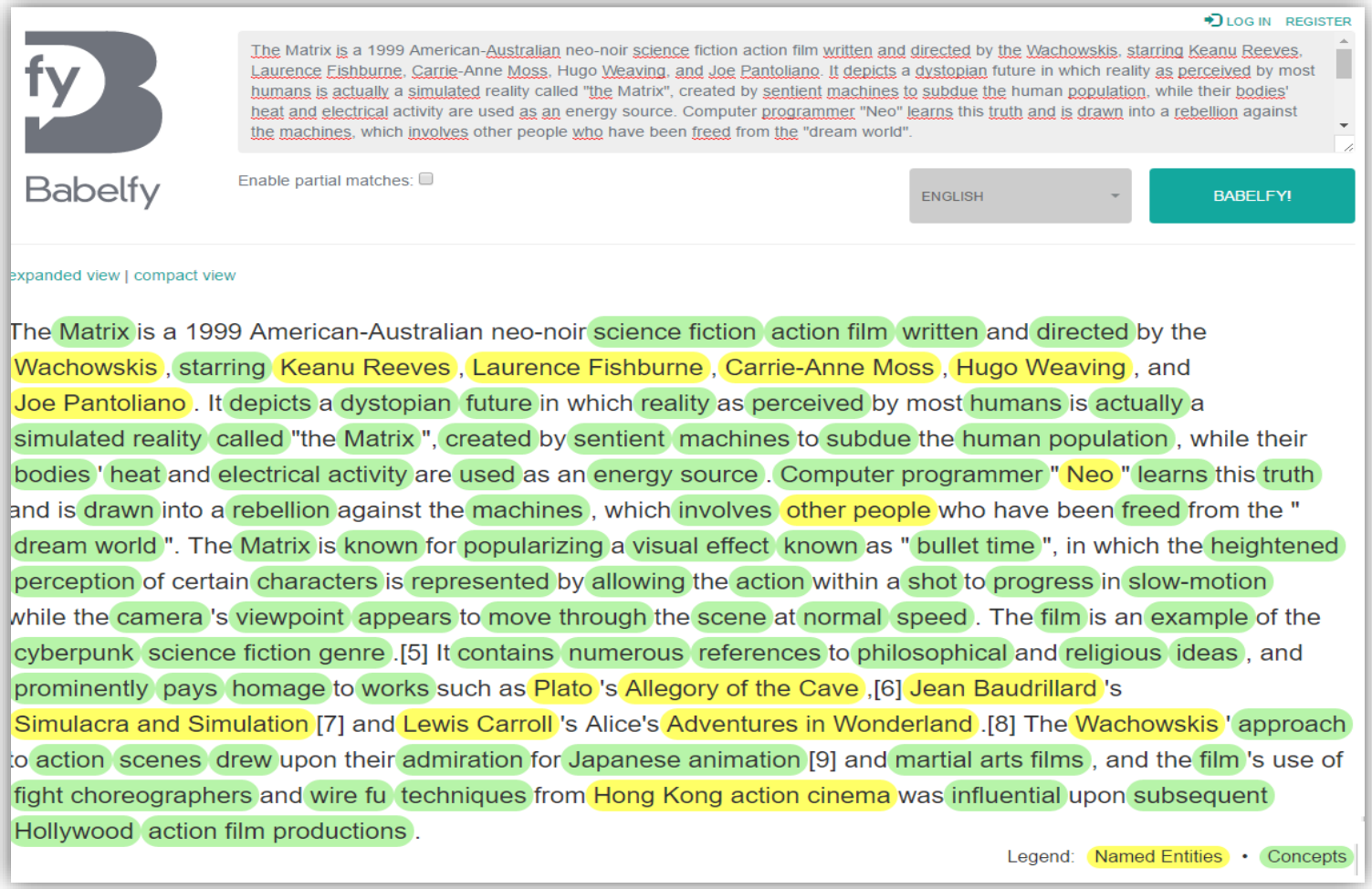
BabelNet 4.0: General statistics

Number of languages:	284
Total number of Babel synsets:	15,780,364
Total number of Babel senses:	808,974,108
Total number of concepts:	6,113,467
Total number of Named Entities:	9,666,897
Total number of lexico-semantic relations:	277,036,611
Total number of glosses (textual definitions):	91,218,220
Total number of images:	54,229,458
Total number of Babel synsets with at least one domain:	2,637,407
Total number of Babel synsets with at least one picture:	10,522,922
Total number of sources:	47

Entity Linking Algorithms

Babelfy

http://babelfy.org/

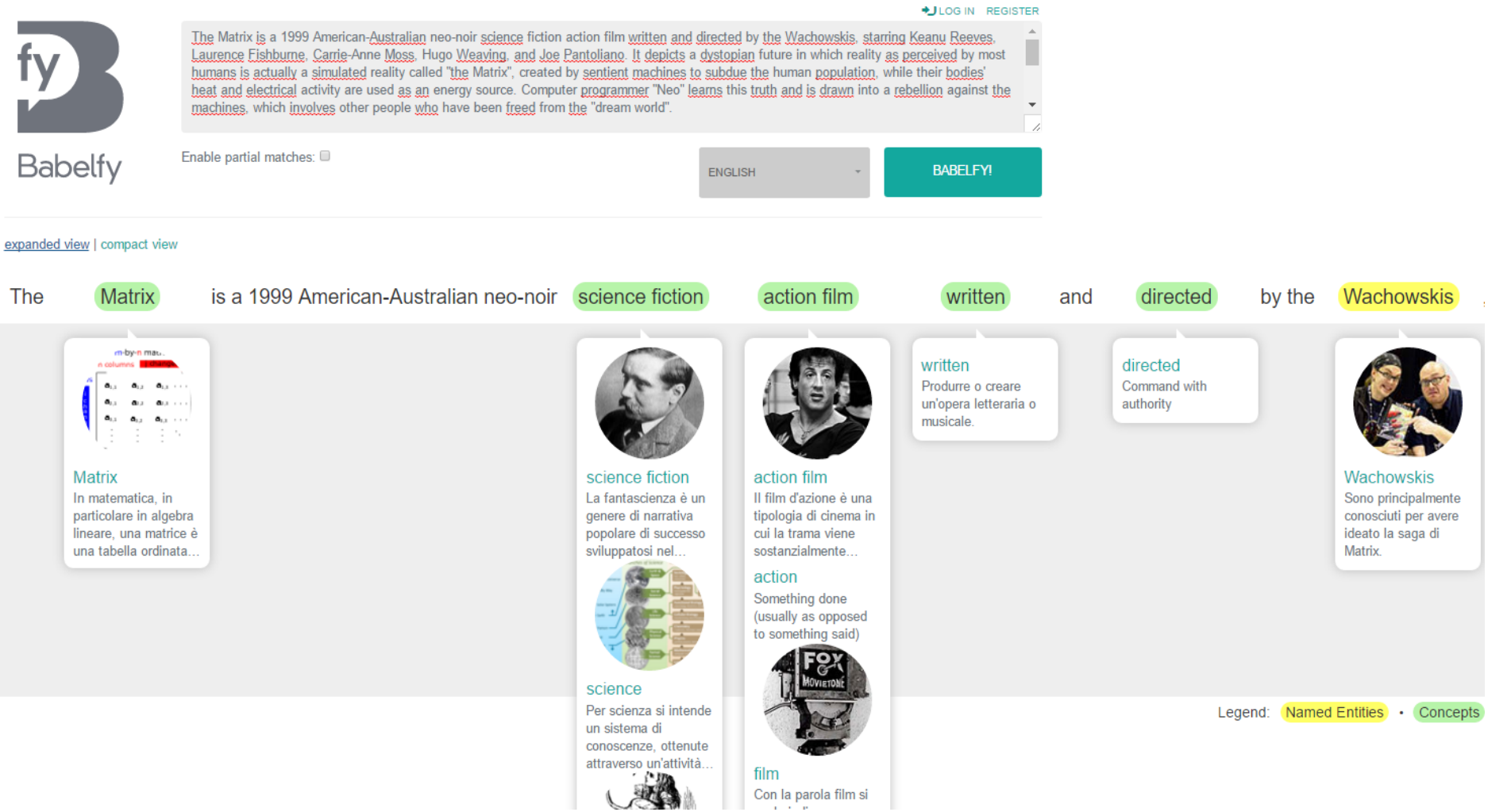


we have both **Named Entities** and **Concepts**!

Entity Linking Algorithms

Babelfy

http://babelfy.org/



The Matrix representation

The Matrix representation

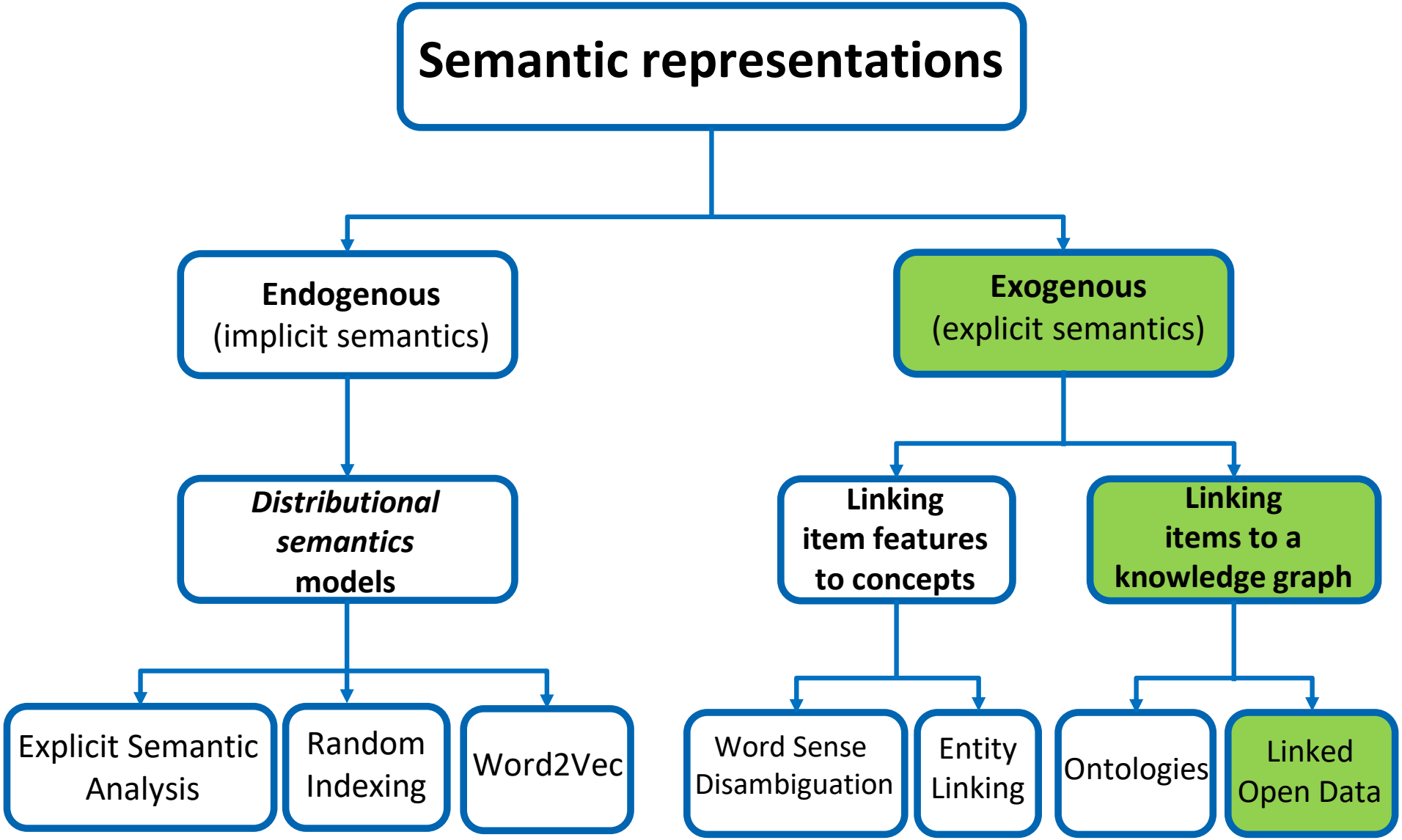
BabelNet

entities recognized and modeled (as in Tag.me)

Wikipedia-based representation: some common sense terms included, and new interesting features (e.g. science fiction, action film) can be generated

includes linguistic knowledge and is able to disambiguate terms

also multilingual!


Semantic representations
Endogenous
(implicit semantics)
Exogenous
(explicit semantics)
Distributional semantics models
Linking item features to concepts
Linking items to a knowledge graph
Explicit Semantic Analysis
Random Indexing
Word2Vec
Word Sense Disambiguation
Entity Linking
Ontologies
Linked Open Data

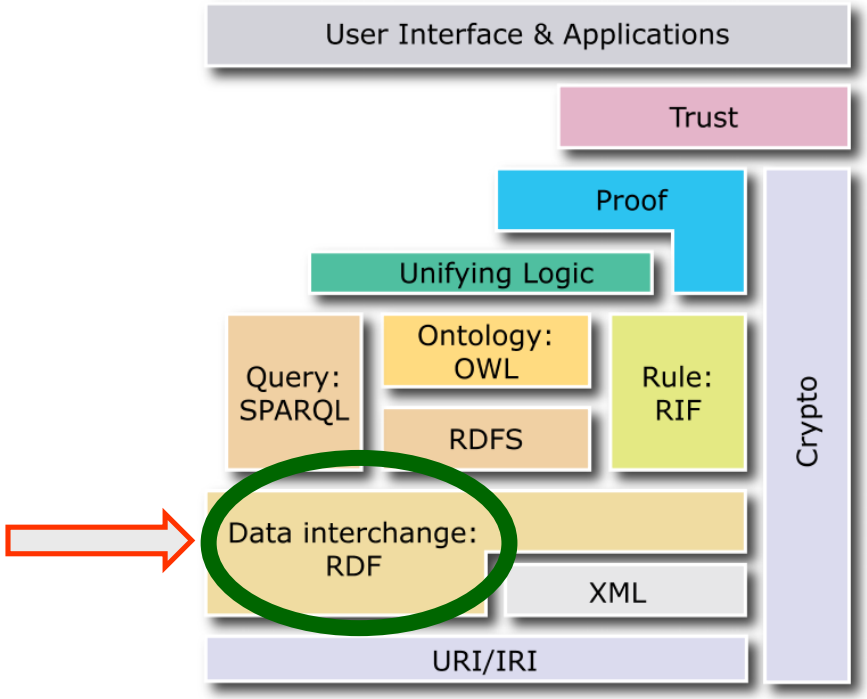
Linked Open Data

What is it?

Linked Open Data is a **methodology** to publish, share and link **structured data** on the Web

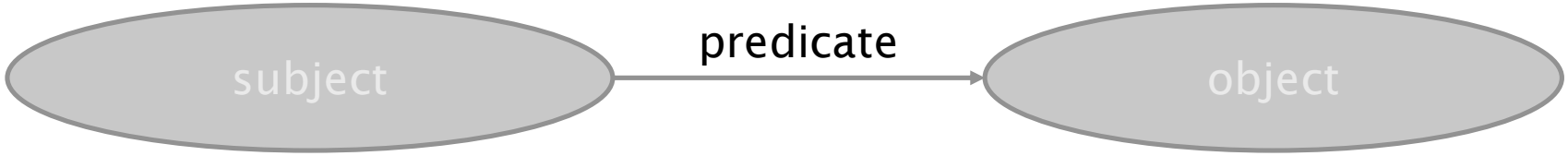
Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available



Linked Open Data - Cornerstones

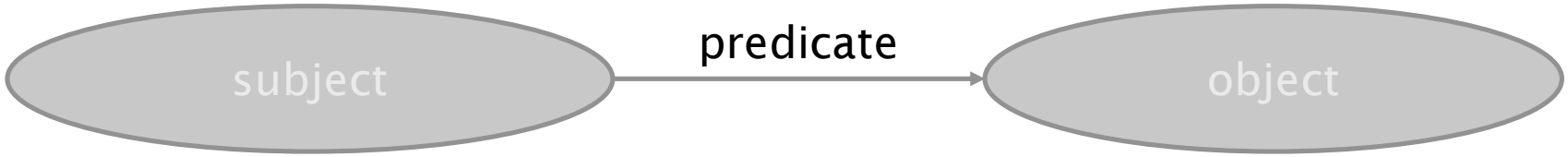
1. Use of RDF to model the information and make data publicly available



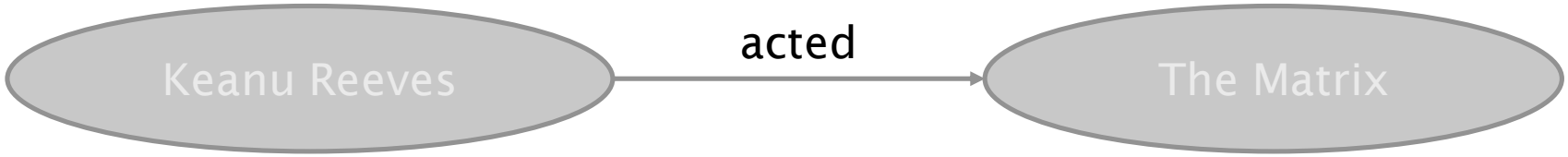
(this is called *RDF triple*)

Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available

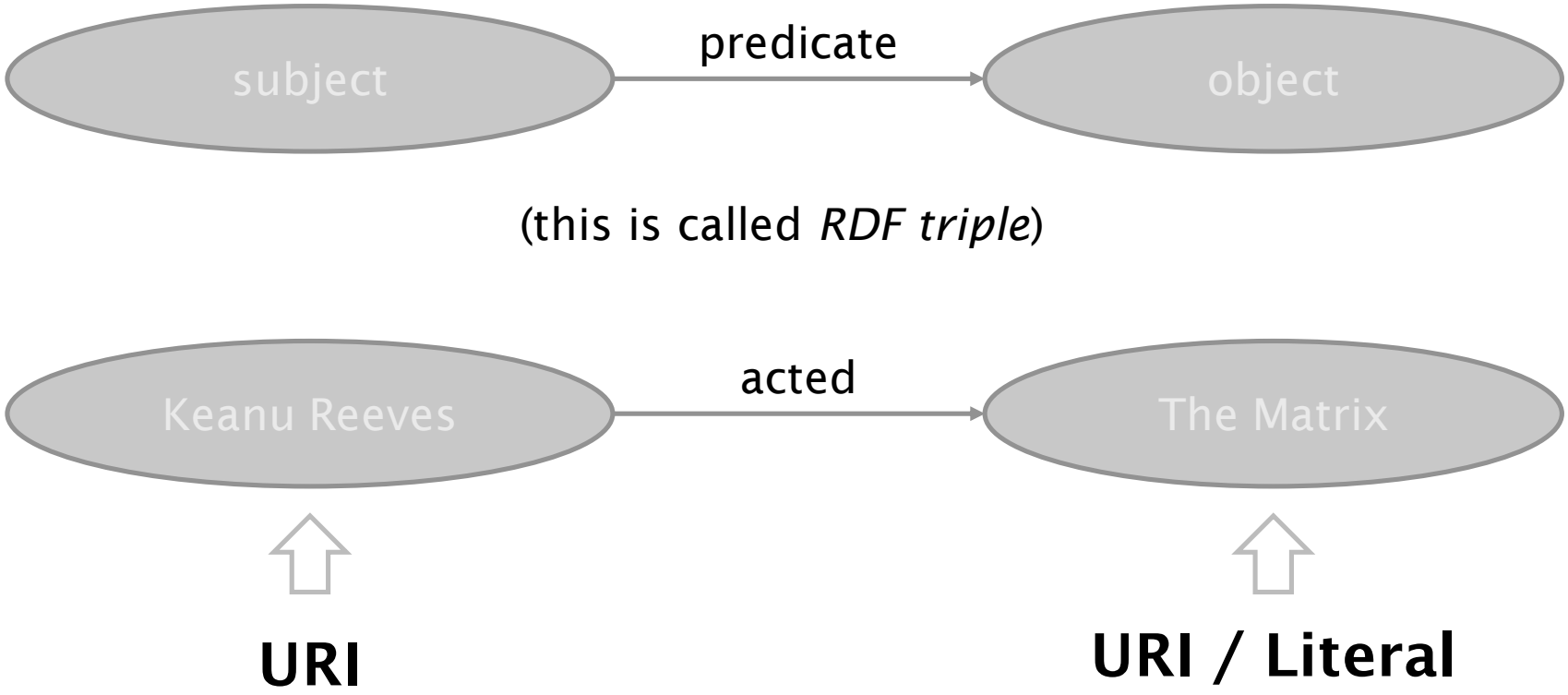


(this is called *RDF triple*)



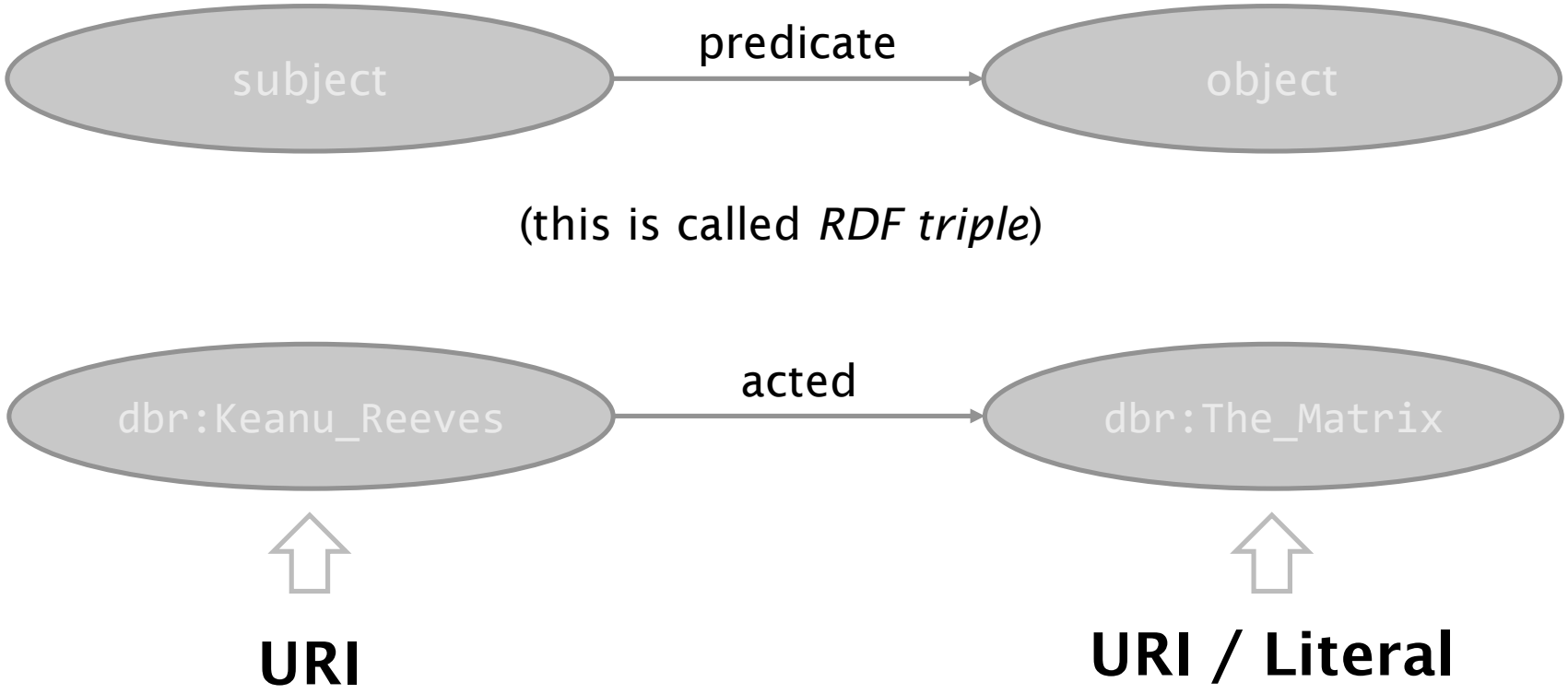
Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available



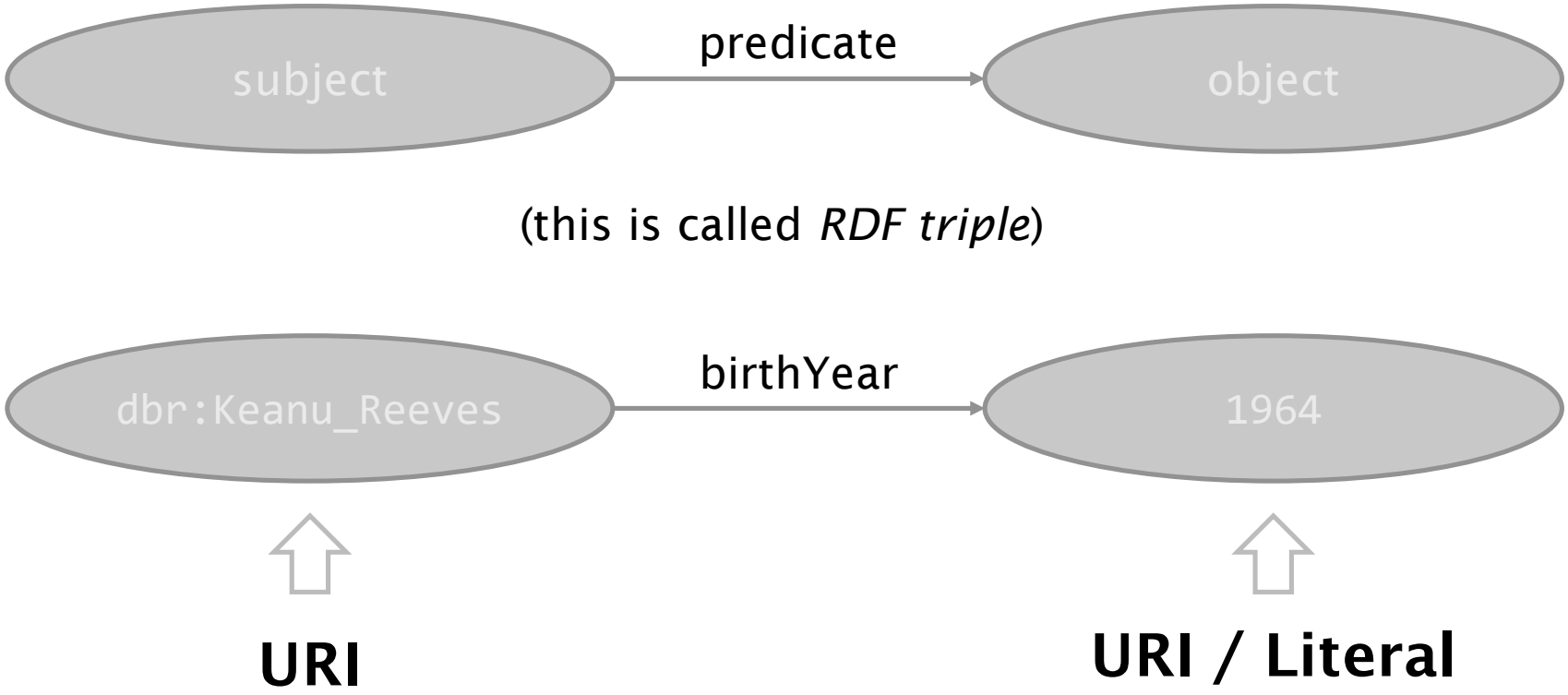
Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available



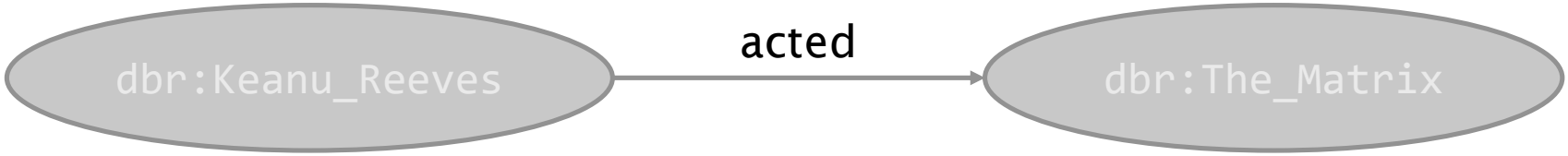
Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available



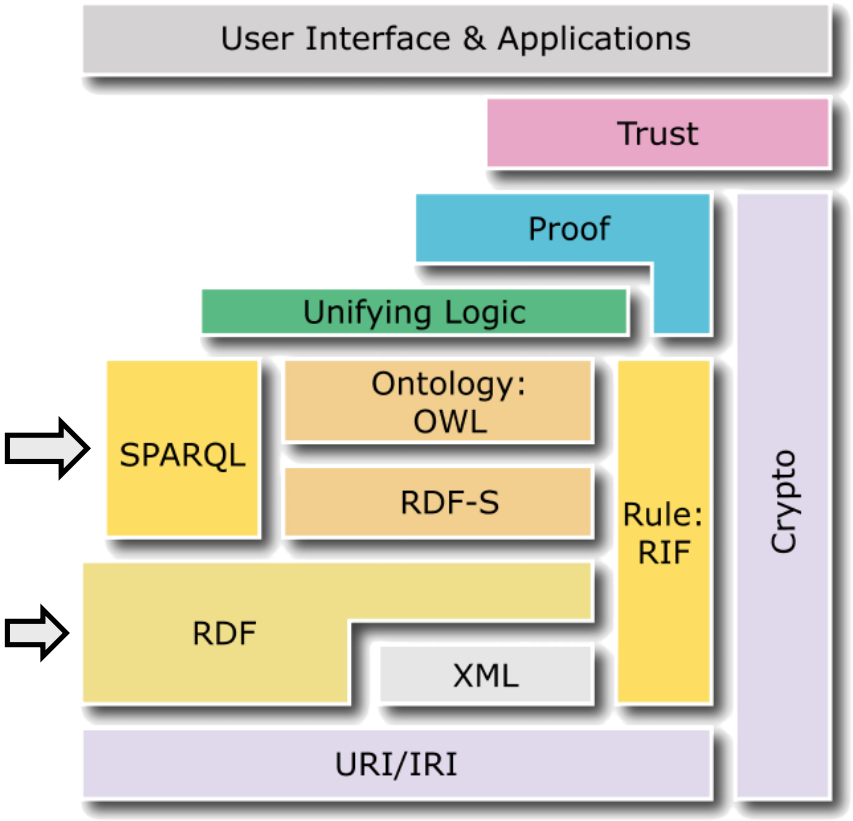
Linked Open Data - Cornerstones

1. Use of RDF to model the information and make data publicly available



2. Re-Use existing resources and properties in order to make the data inter-connected

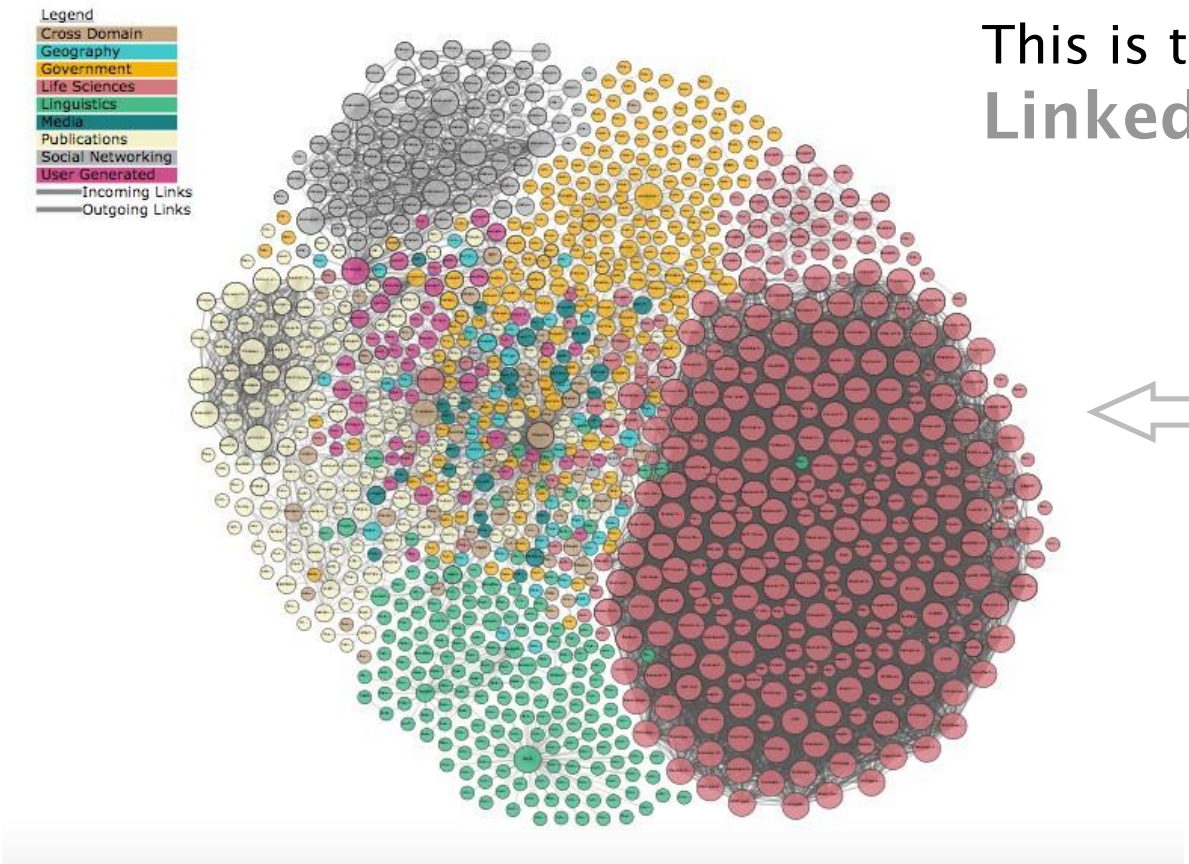
Linked Open Data



We only use a small subset of the 'Semantic web cake'

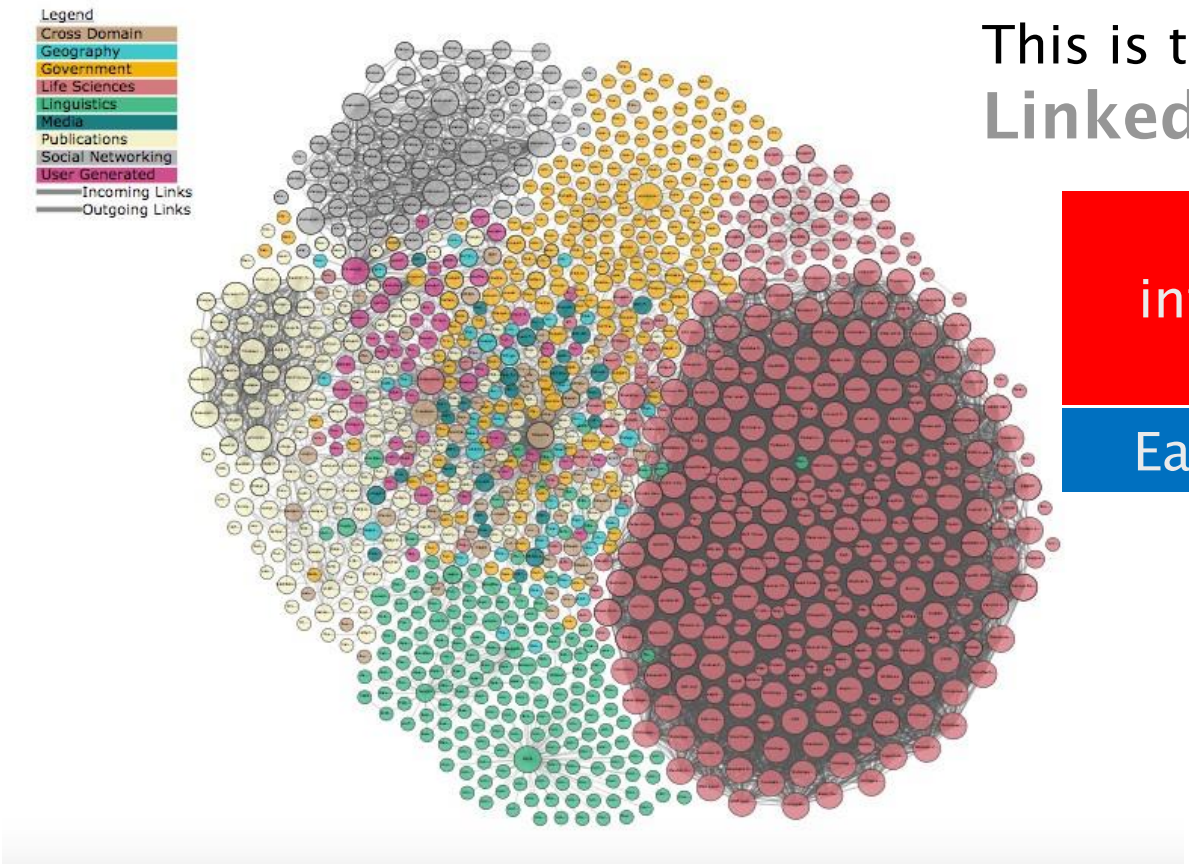
We use RDF to model our data and we use SPARQL as query language to gather data

Linked Open Data



This is the
Linked Open Data cloud

Linked Open Data

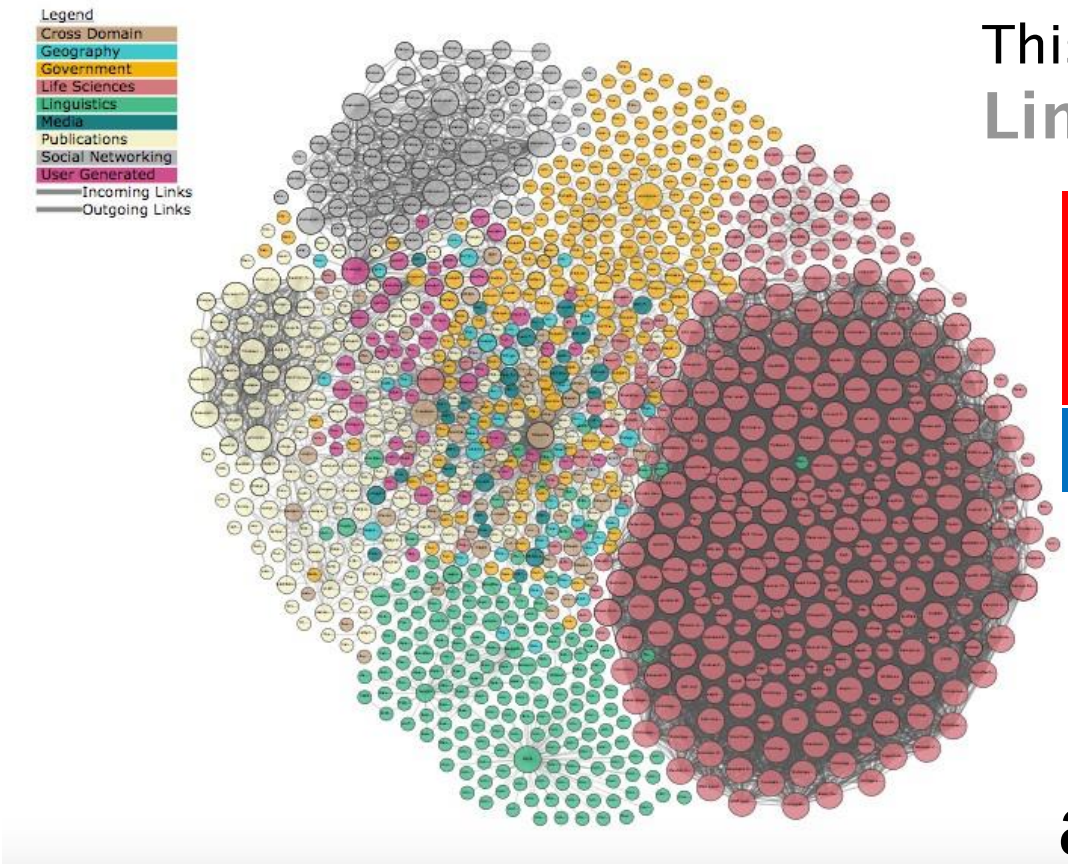


This is the
Linked Open Data cloud

It is a (huge) set of interconnected semantic datasets

Each bubble is a dataset!

Linked Open Data



This is the
Linked Open Data cloud

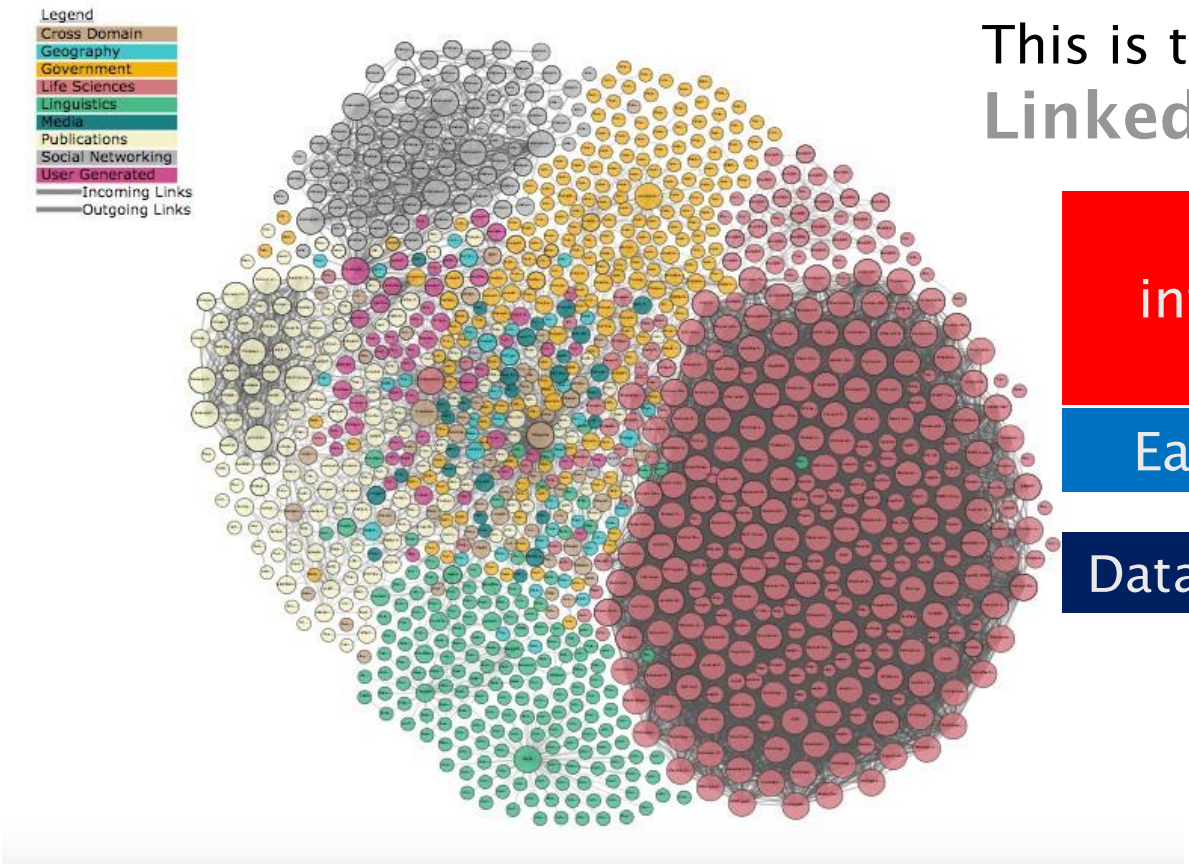
It is a (huge) set of interconnected semantic datasets

Each bubble is a dataset!

How many datasets do we have?

149 billions triples
and **9,960** datasets

Linked Open Data



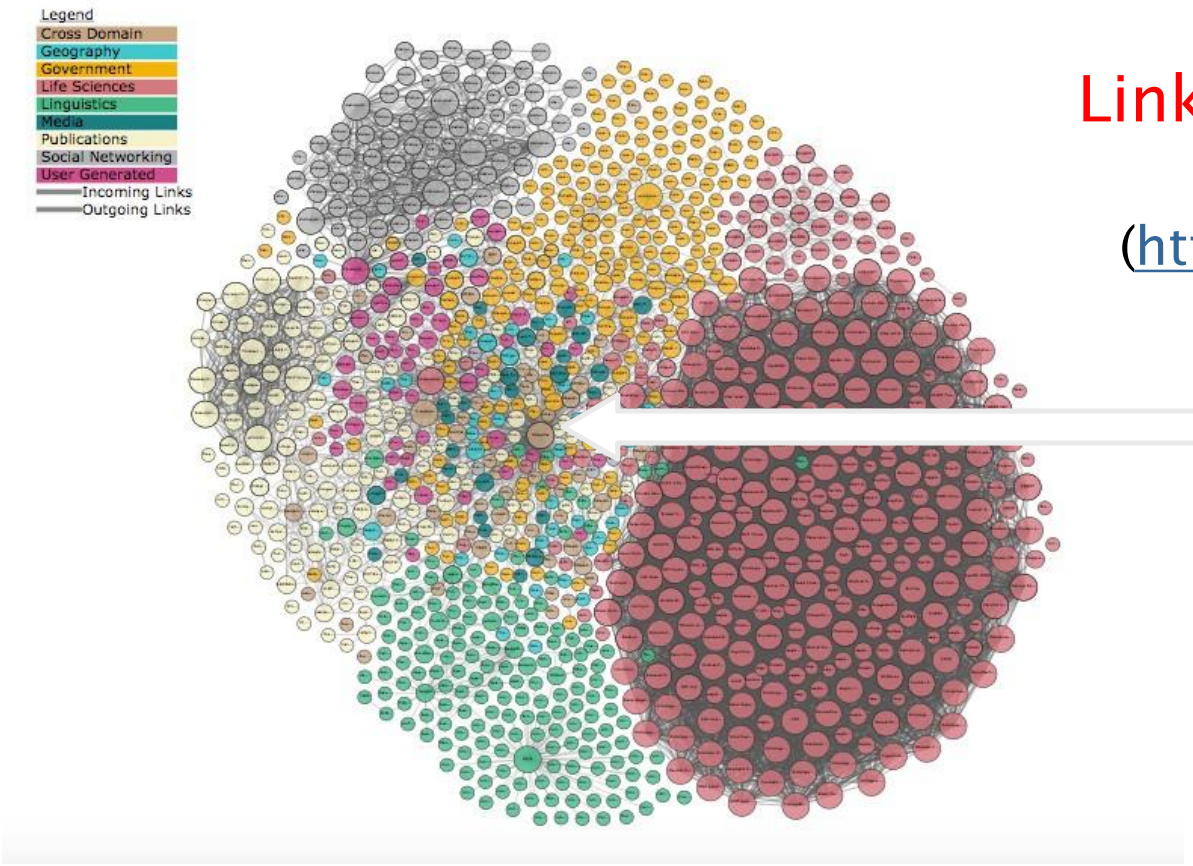
This is the
Linked Open Data cloud

It is a (huge) set of interconnected semantic datasets

Each bubble is a dataset!

Datasets cover many domains

Linked Open Data



The core of the
Linked Open Data cloud
is **DBpedia**
(http://www.dbpedia.org)

RDF mapping
of Wikipedia

DBpedia

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix was known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well-received by critics.[10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix

Theatrical release poster

Wikipedia
Unstructured Content

DBpedia

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

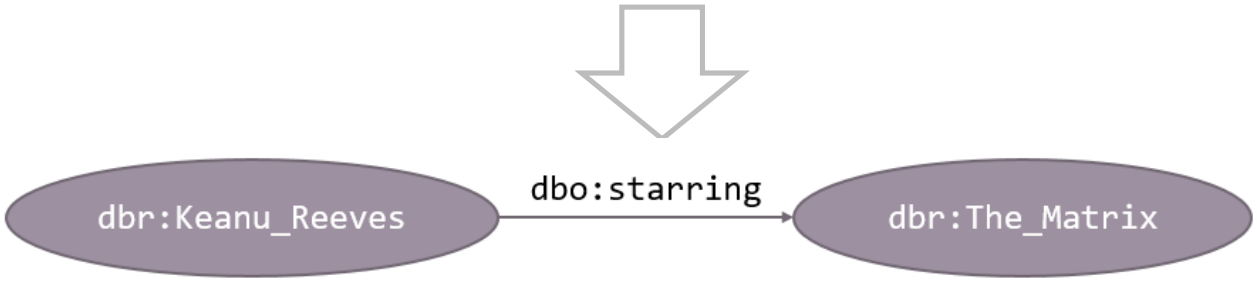
The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well-received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix

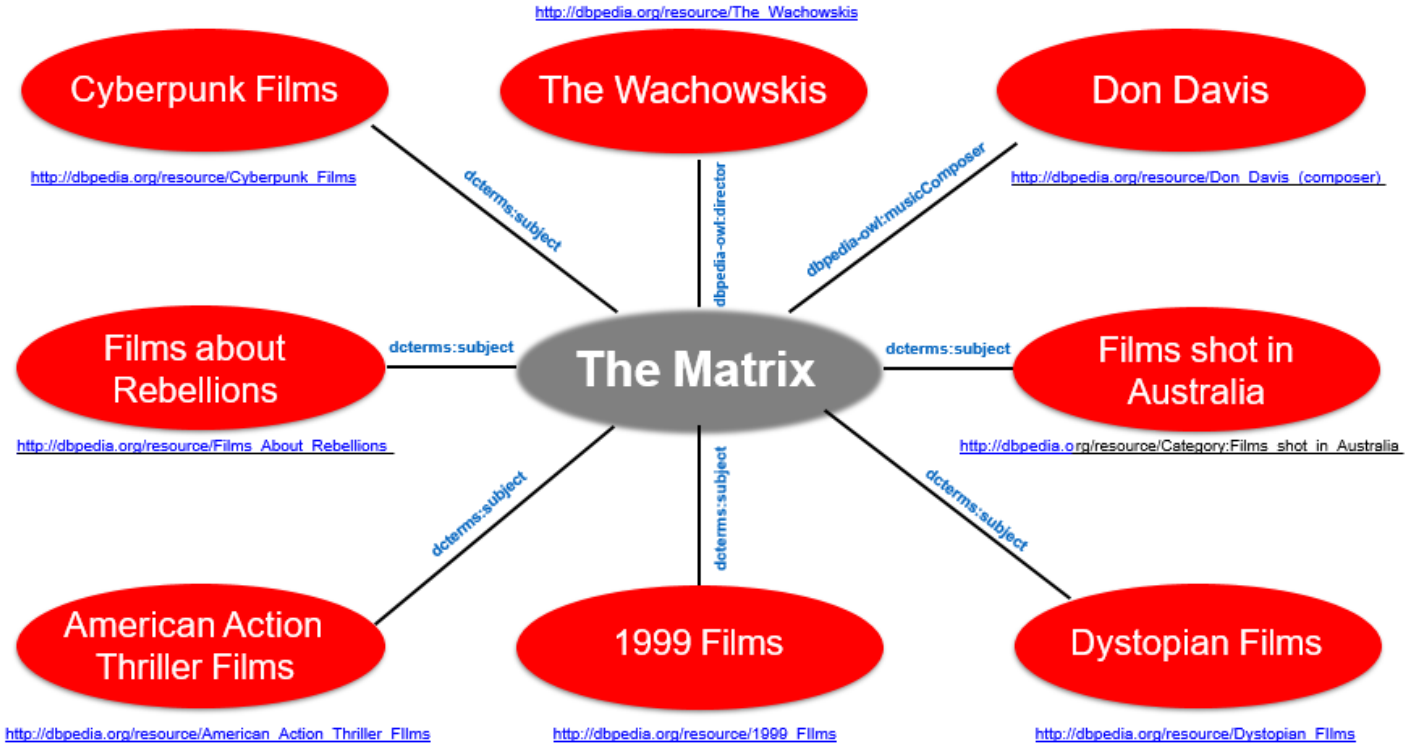
Theatrical release poster

Wikipedia
Unstructured Content



DBpedia
Structured Data

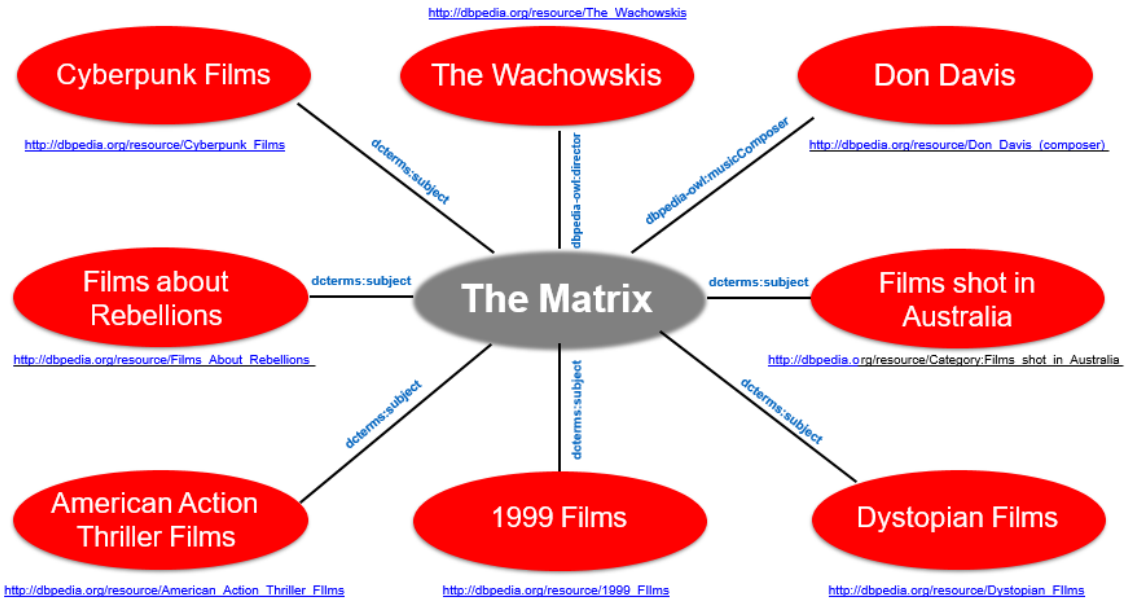
DBpedia



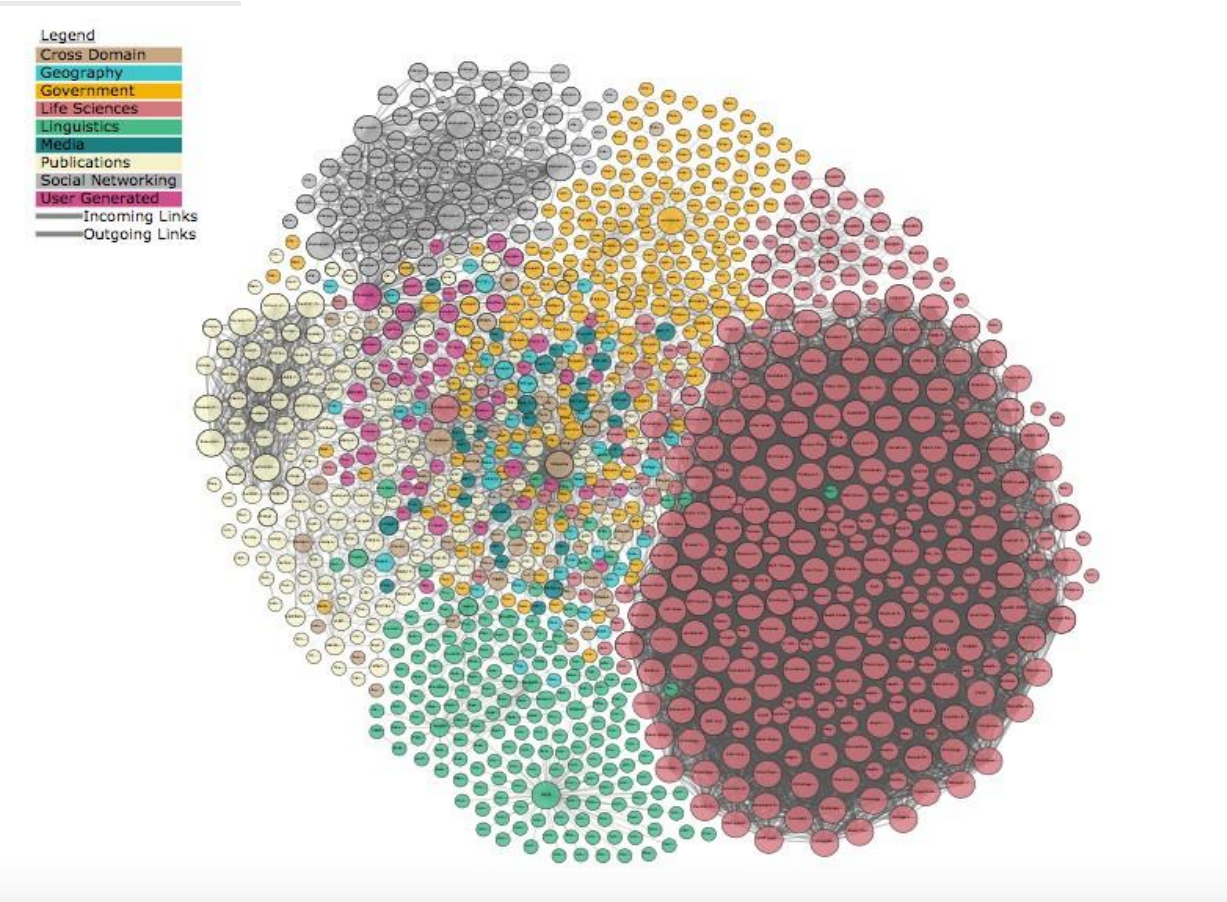
All the information available in Wikipedia

is modeled in RDF

...One step back



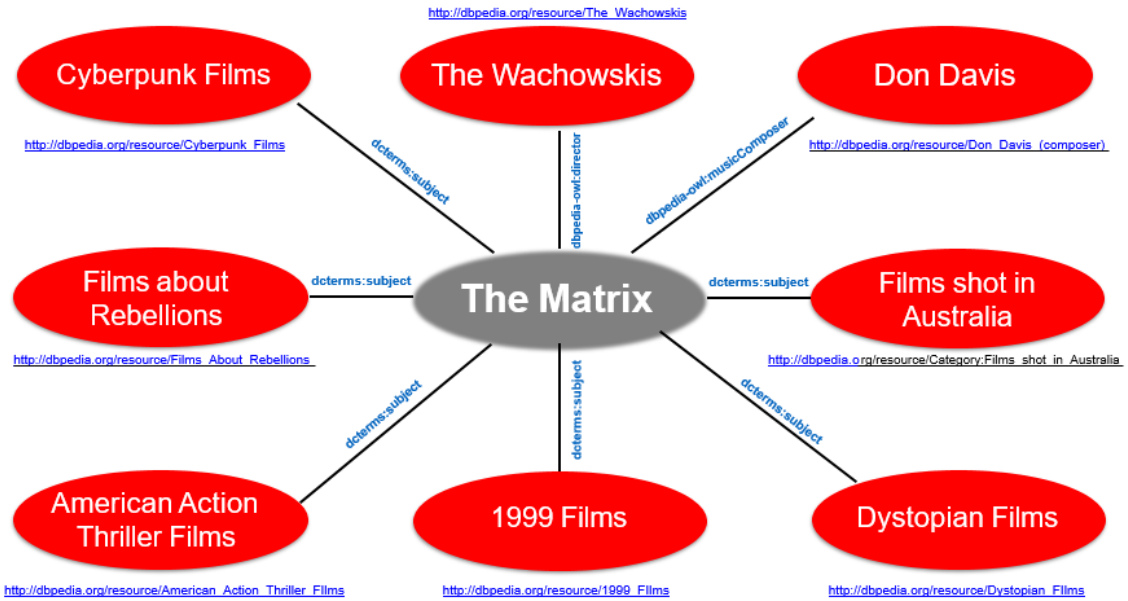
DBpedia – In a Nutshell



We know that **somewhere** in the Linked Open Data cloud **we have an entity referring the the movie 'The Matrix'**, and we have a lot of available properties!

How can we find it?

...One step back



SPARQL comes into play!

SPARQL

```
[…]
SELECT DISTINCT ?city ?name
WHERE {
  ?city dct:subject dbc:Cities_in_Italy .
  ?city rdfs:label ?name .
  ?city dbo:populationTotal ?population .
  FILTER (?population > 100000) .
  FILTER (lang(?name) = 'en')
}
```

An example of **SPARQL** query

SPARQL

```
[…]
SELECT DISTINCT ?city ?name
WHERE {
  ?city dct:subject dbc:Cities_in_Italy .
  ?city rdfs:label ?name .
  ?city dbo:populationTotal ?population .
  FILTER (?population > 100000) .
  FILTER (lang(?name) = 'en')
}
```

Returns **big cities in Italy** (more than 100,000 people)

An example of **SPARQL** query

SPARQL

```
[…]
SELECT DISTINCT ?city ?name
WHERE {
  ?city dct:subject dbc:Cities_in_Italy .
  ?city rdfs:label ?name .
  ?city dbo:populationTotal ?population .
  FILTER (?population > 100000) .
  FILTER (lang(?name) = 'en')
}
```

Returns **big cities in Italy** (more than 100,000 people)

How do we exploit SPARQL?

SPARQL

```
[…]
SELECT DISTINCT ?city ?name
WHERE {
  ?city dct:subject dbc:Cities_in_Italy .
  ?city rdfs:label ?name .
  ?city dbo:populationTotal ?population .
  FILTER (?population > 100000) .
  FILTER (lang(?name) = 'en')
}
```

Returns **big cities in Italy** (more than 100,000 people)

Key concept: mapping

SPARQL


```
SELECT DISTINCT ?uri, ?title
WHERE {
   ?uri rdf:type dbpedia-owl:Film.
   ?uri rdfs:label ?title.
   FILTER langMatches(lang(?title), "EN").
   FILTER regex(?title, "matrix", "i")
}
```

We can run a **SPARQL query** to find **the corresponding URI for the resource**

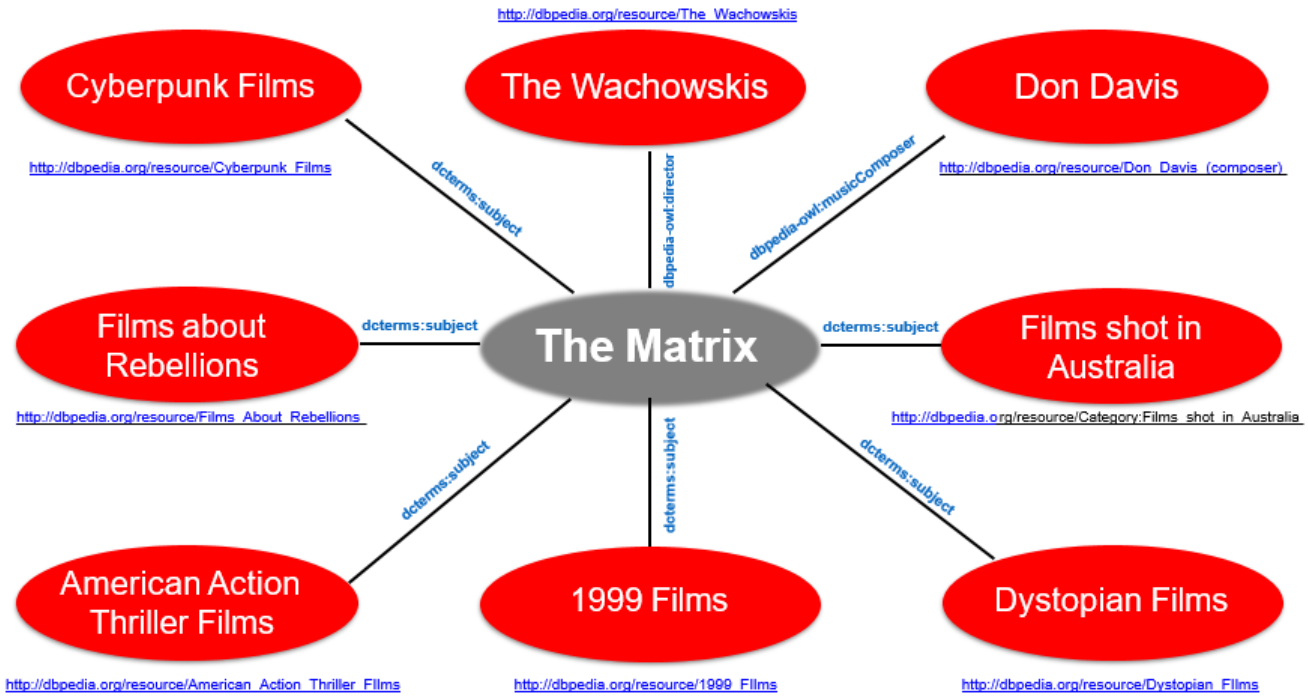
SPARQL

dbr:The_Matrix

```
SELECT DISTINCT ?uri, ?title
WHERE {
   ?uri rdf:type dbpedia-owl:Film.
   ?uri rdfs:label ?title.
   FILTER langMatches(lang(?title), "EN").
   FILTER regex(?title, "matrix", "i")
}
```

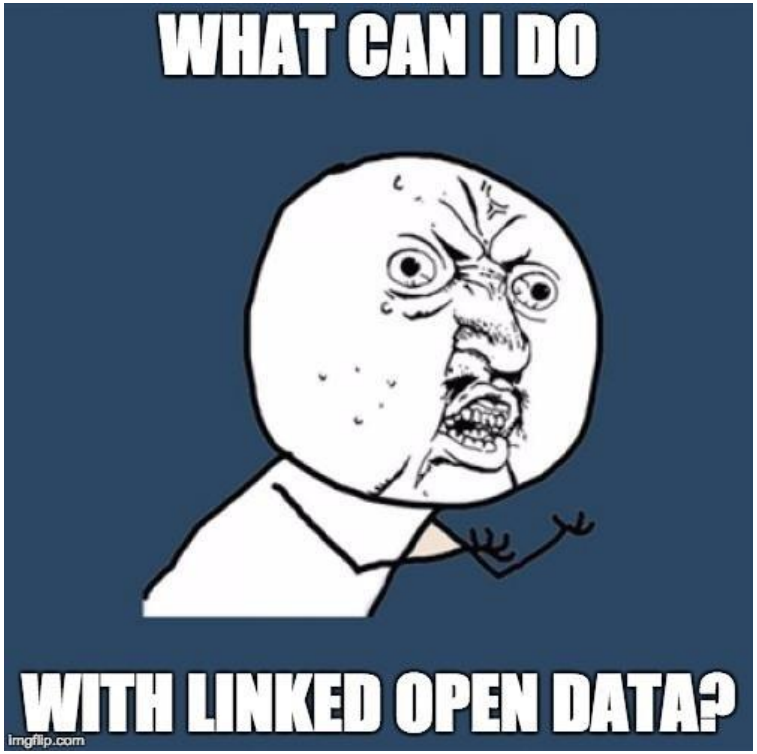
We want to link «logical» entities occurring in our data with «physical» entities occurring in the LOD cloud

LOD-aware data model



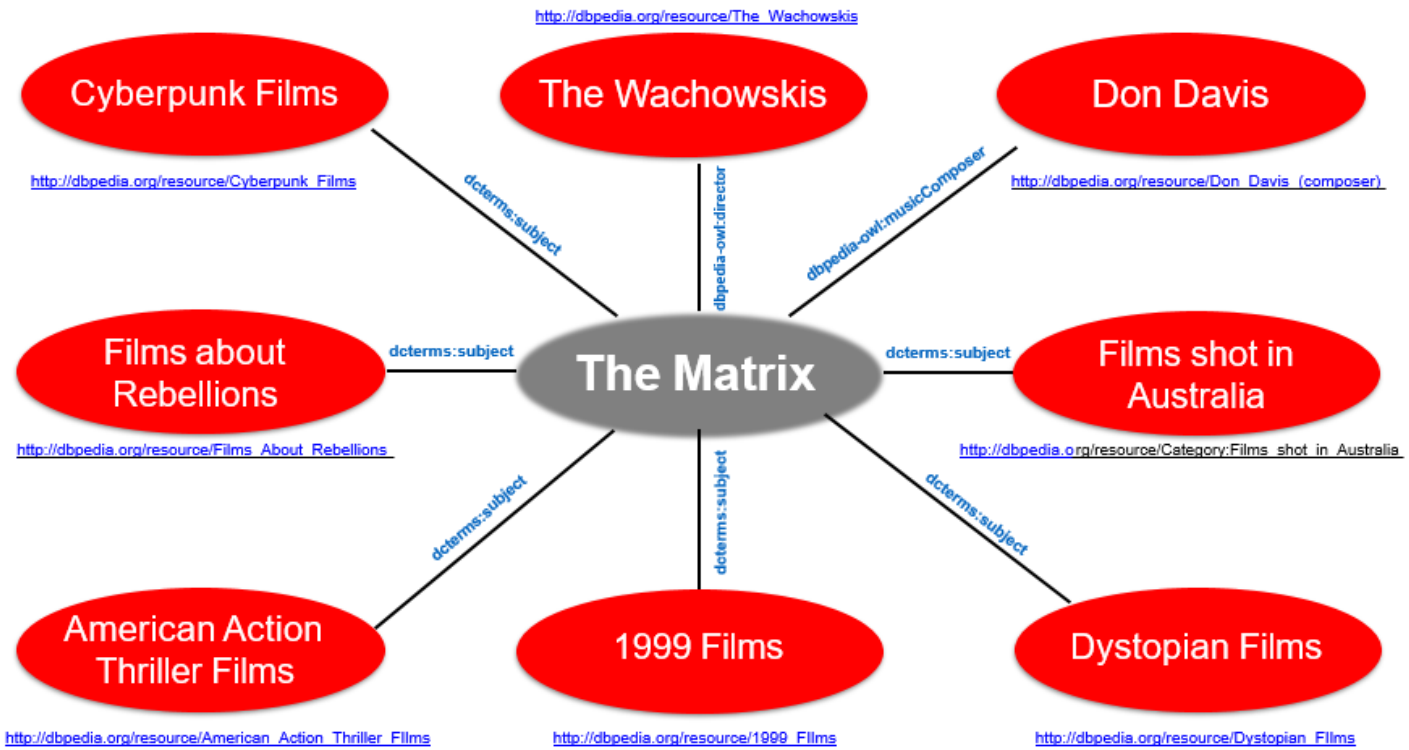
Once we have a mapping, properties can be extracted

LOD-aware RecSys



How can we use Linked Open Data for **Recommender Systems?**

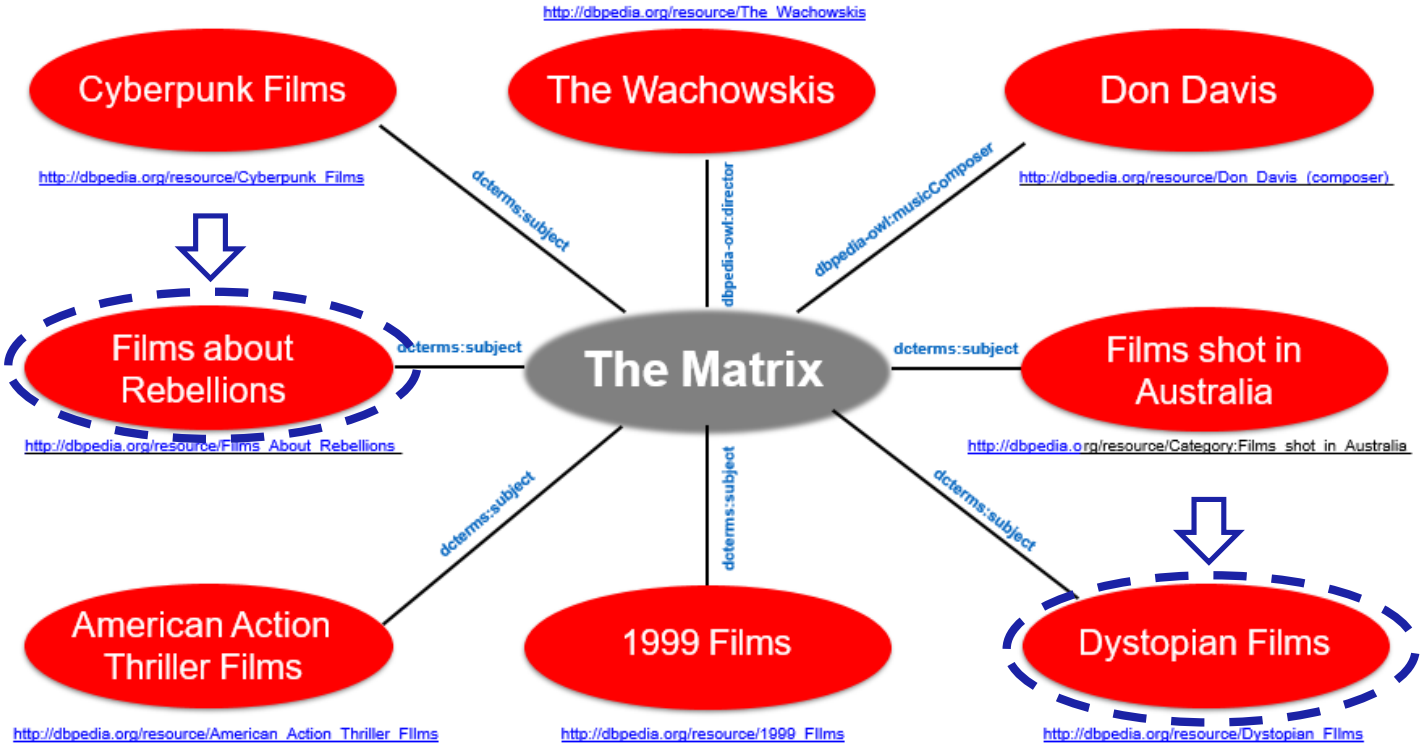
Motivations: Limited Content Analysis



In some scenarios, **we don't have enough features** to feed our recommendation models.

LOD cloud **can be helpful**

Motivations: Limited Content Analysis



Several **very fine-grained**
and interesting features can be
easily injected by querying DBpedia

LOD-aware RecSys

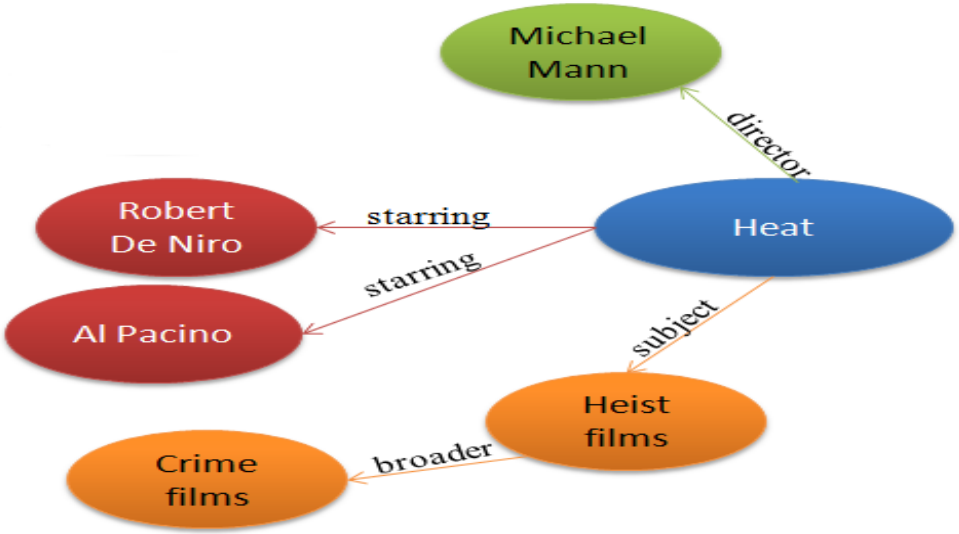
1. Approaches based on **Vector Space Models**

2. Approaches based on **Graph-based Models**

3. Approaches based on **Machine Learning techniques**

LOD-based Recommender Systems

approaches based on VSM



LOD-based Recommender Systems

Thanks to the LOD we can obtain a richer vector-space representation

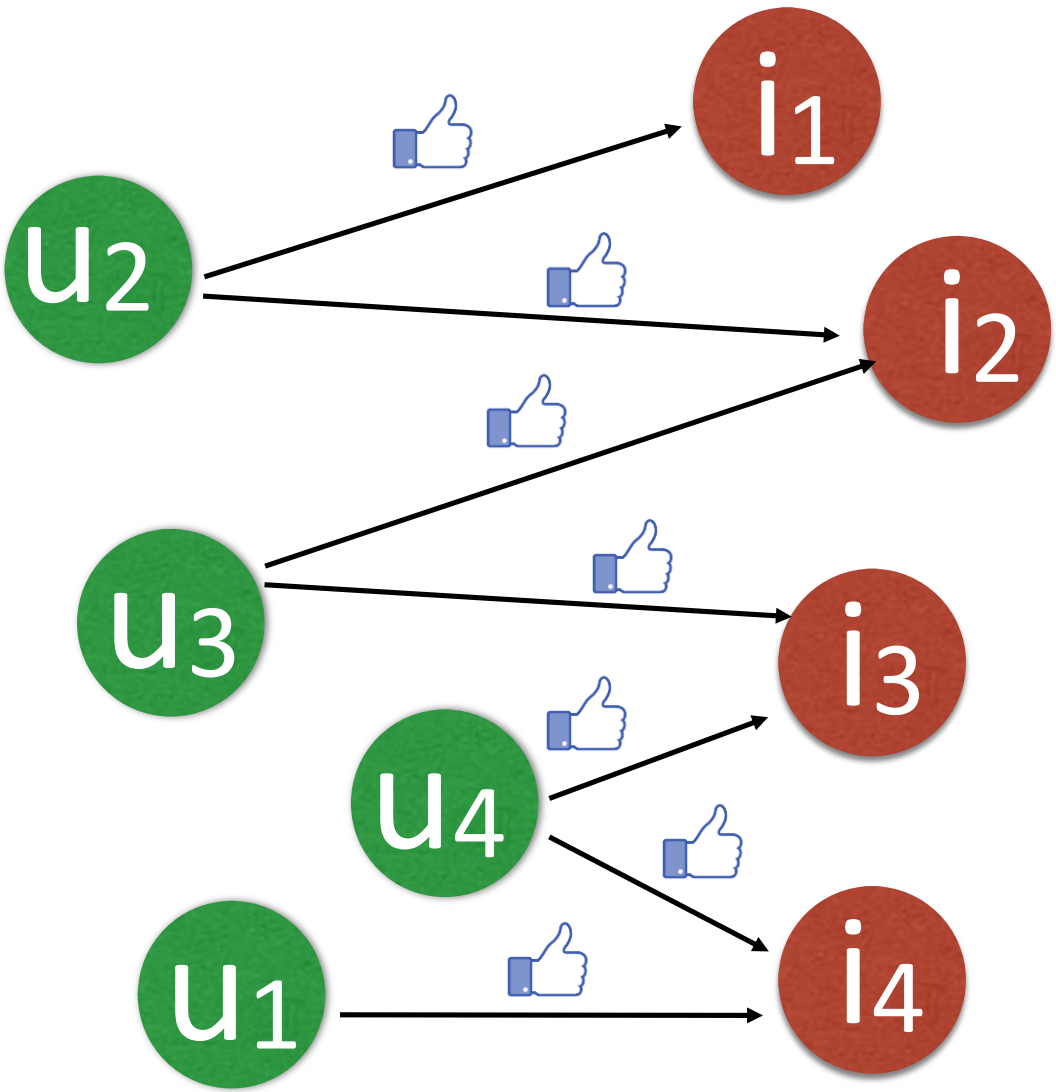
	STARRING		DIRECTOR	SUBJECT+BROADER	
Heat	Robert DeNiro	Al Pacino	Michael Mann	Heist films	Crime films

similarity between items based on the Vector Space representation, e.g. **Jaccard**

LOD-aware RecSys

1. Approaches based on **Vector Space Models**
2. Approaches based on **Graph-based Models**
3. Approaches based on **Machine Learning techniques**

Graph-based Data Model

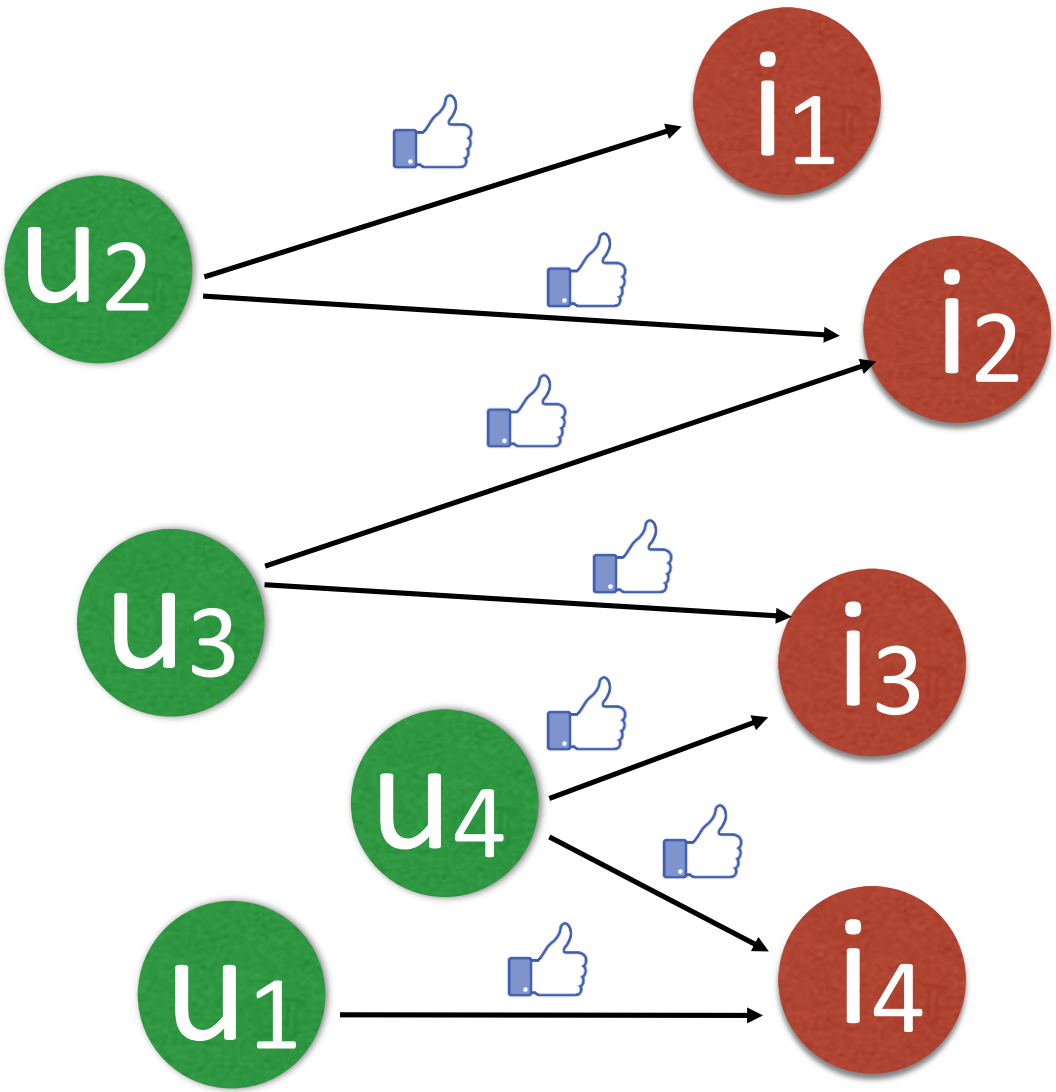


users = **nodes**
items = **nodes**
preferences = **edges**

(bipartite graph)

Very intuitive representation!

Graph-based Data Model



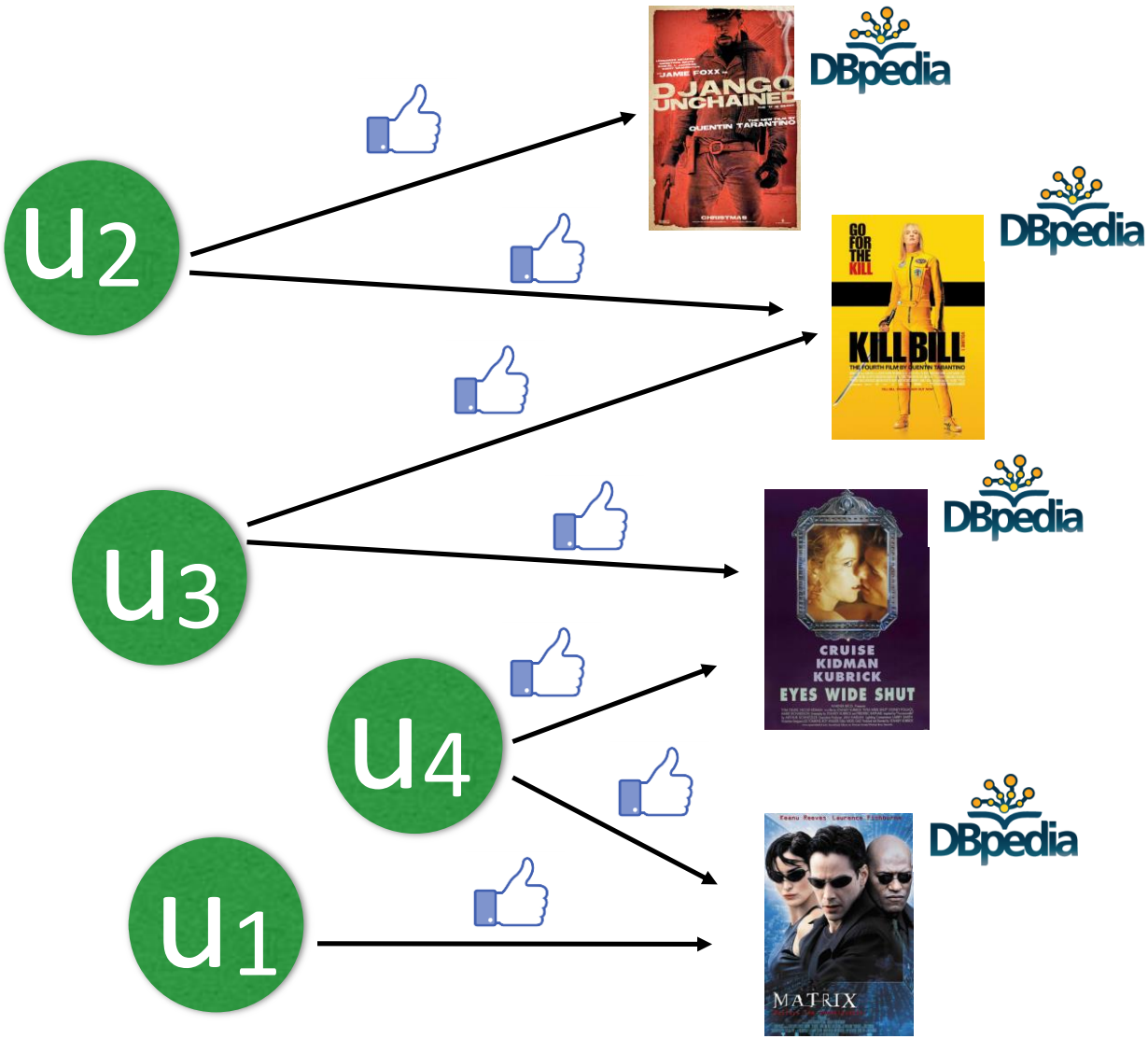
users = **nodes**
items = **nodes**
preferences = **edges**

(bipartite graph)

Basic graph-based data models only encode **collaborative data points**

We can extend such data model by introducing features gathered from **the LOD cloud**

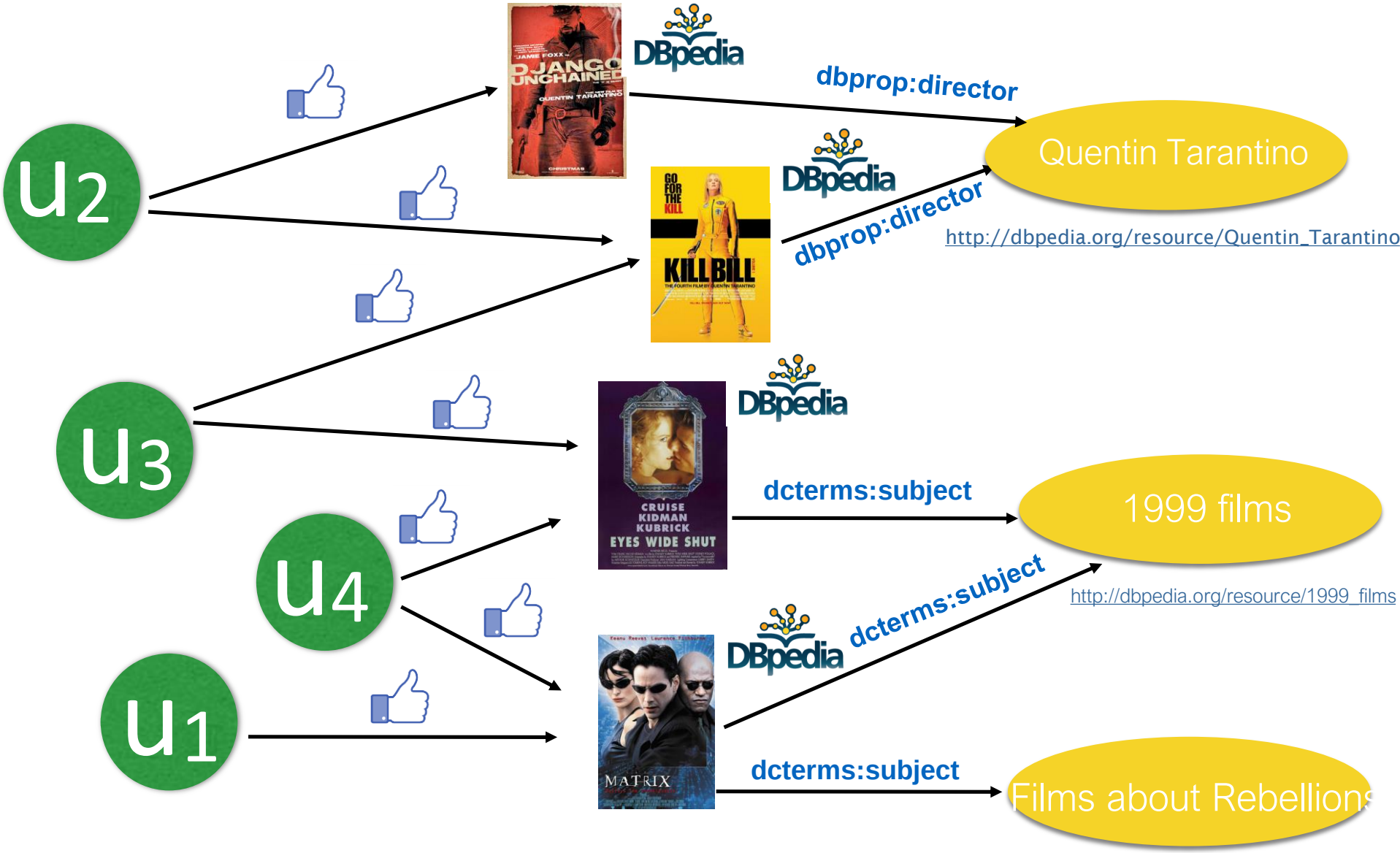
Semantic Graph-based Data Model



DBpedia mapping

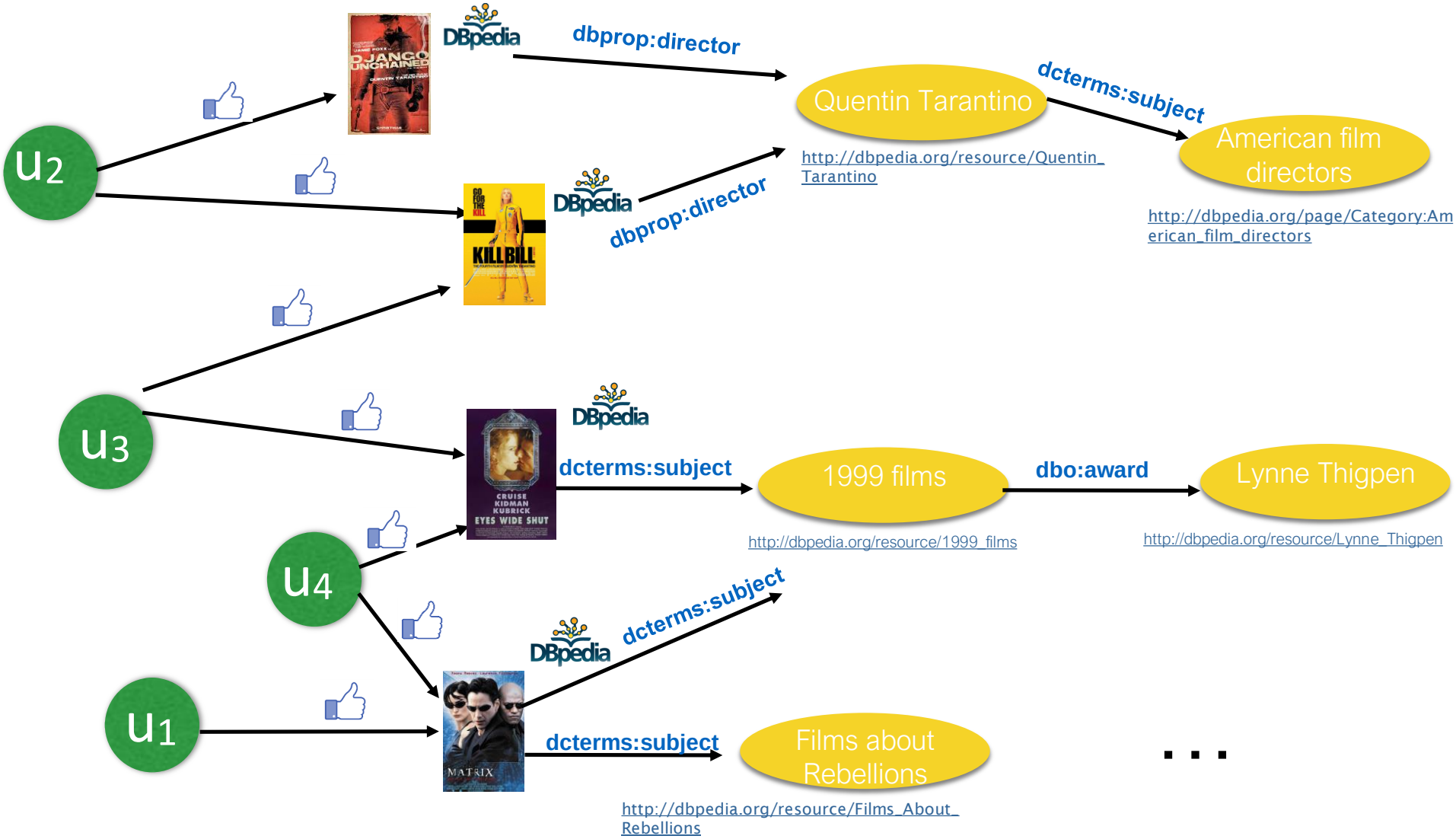
Semantic Graph-based Data Model

(1-hop)



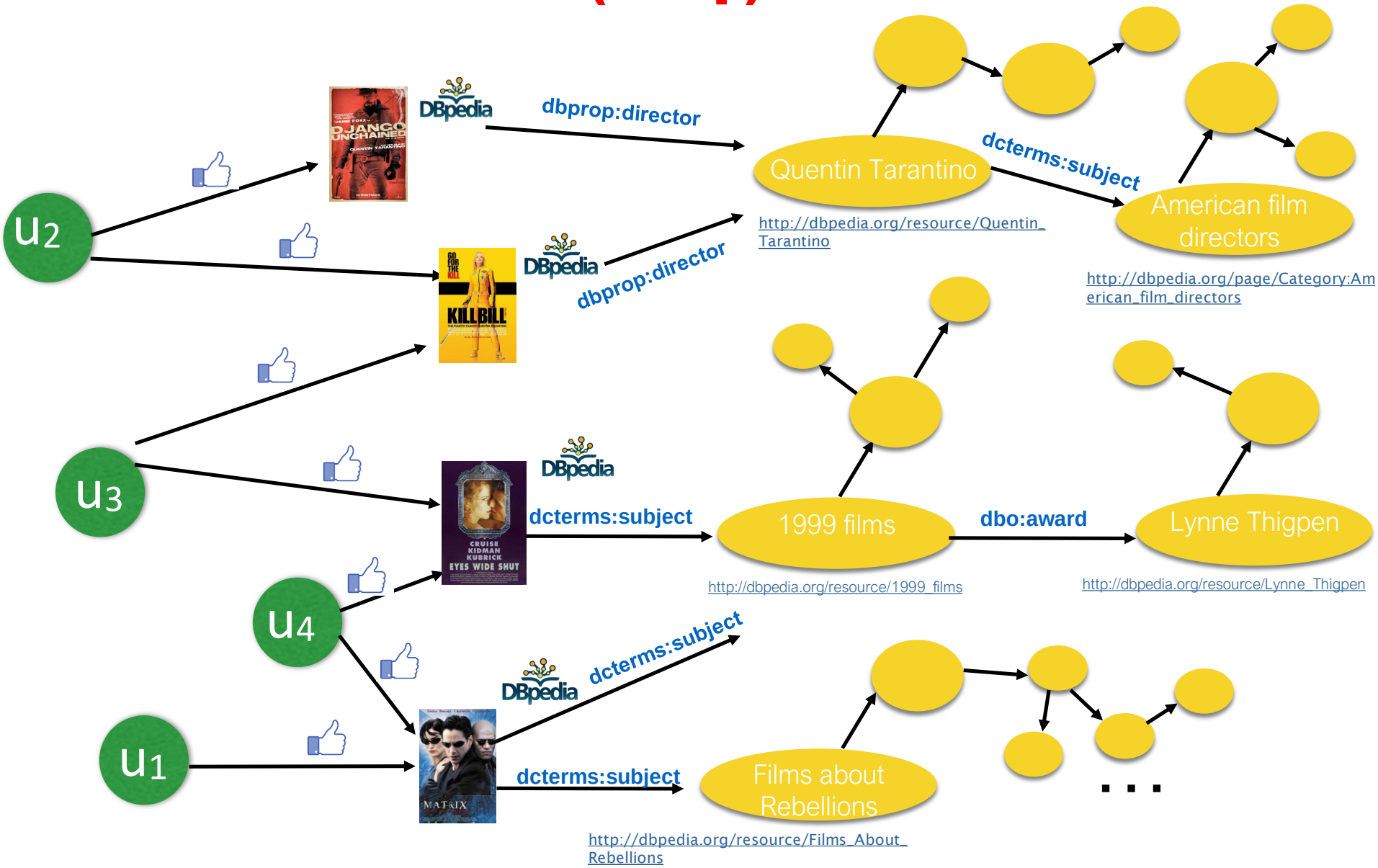
Semantic Graph-based Data Model

(2-hop)



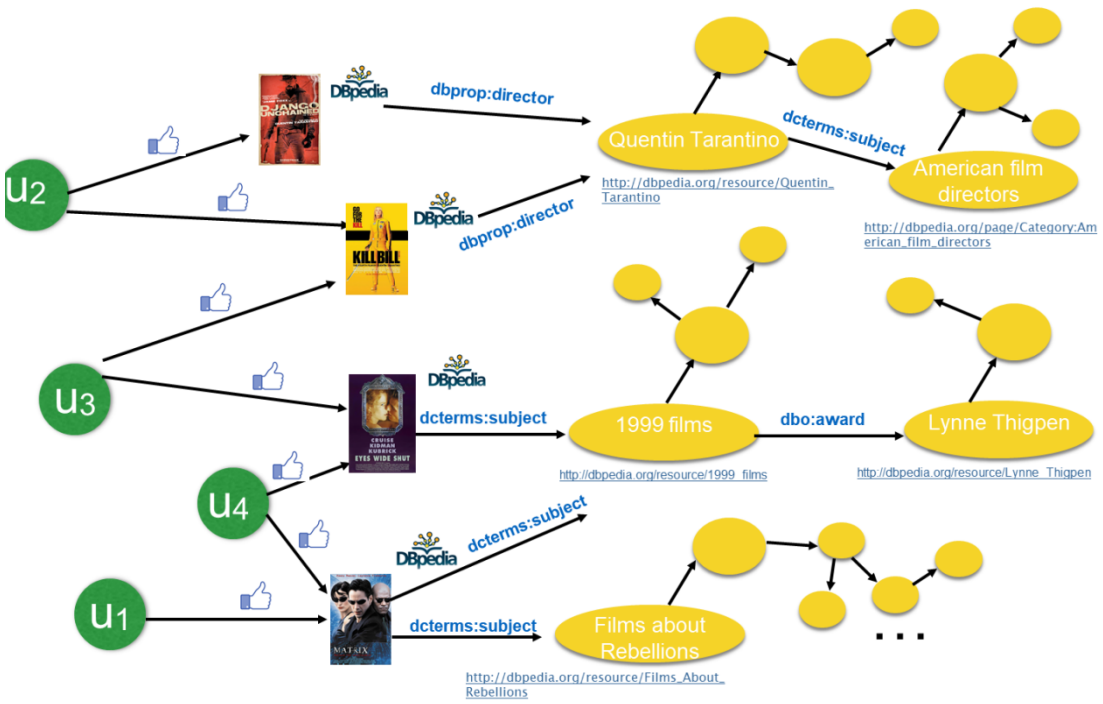
Semantic Graph-based Data Model

(n-hop)



Graph-based RecSys

How to get the recommendations?



Recommendations obtained by mining the graph

Identification of the most relevant (target) nodes, according to the recommendation scenario

PageRank
Spreading Activation
Personalized PageRank
...

Cataldo Musto, Pierpaolo Basile, Pasquale Lops, Marco De Gemmis, Giovanni Semeraro. **Introducing linked open data in graph-based recommender systems**. Information Processing & Management. 53(2): 405-435 (2017)

LOD-aware RecSys

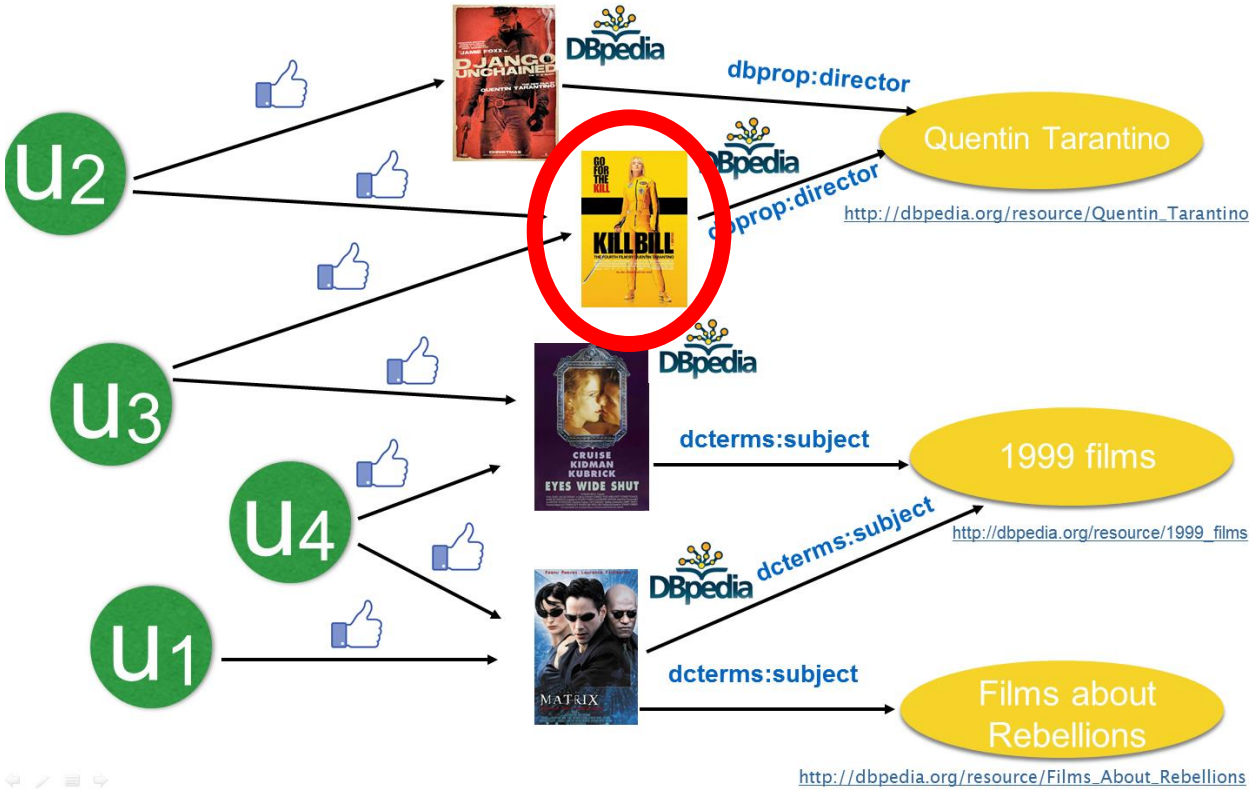
1. Approaches based on **Vector Space Models**

2. Approaches based on **Graph-based Models**

3. Approaches based on **Machine Learning techniques**

Semantic Graph-based Data Model

(Recap)



new features describing the item can be **inferred** by **mining** the **structure** of the **tripartite graph**

Average Neighbor degree
Degree Centrality
Node redundancy
Clustering coefficient

LOD-based Recommender Systems

Research Question: what **is the impact of such features** on the overall performance of the recommendation framework?

Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems. 136 : 1-14 (2017).*

LOD-based Recommender Systems

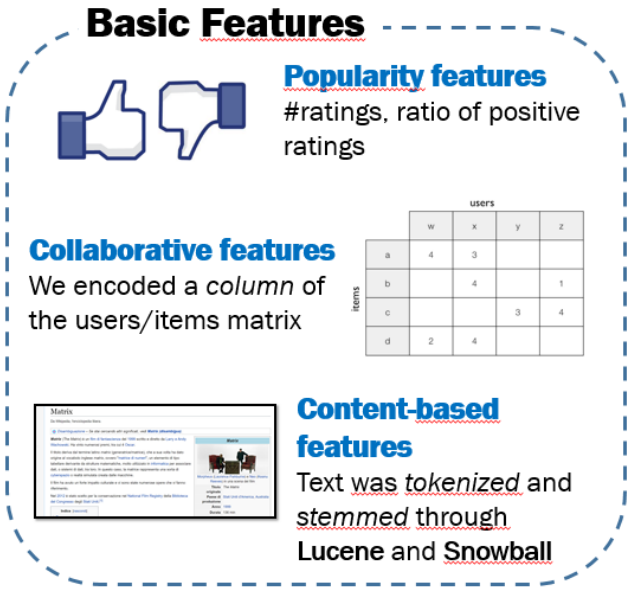
Research Question: what **is the impact of such features** on the overall performance of the recommendation framework?

Insight: to build a hybrid classification framework exploiting **LOD-based** and **graph-based features**

Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems. 136 : 1-14 (2017).*

LOD-based Recommender Systems

Methodology



We first model basic features

Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems. 136 : 1-14 (2017).*

LOD-based Recommender Systems

Methodology

Basic Features

Popularity features
#ratings, ratio of positive ratings

Collaborative features
We encoded a *column* of the users/items matrix

users

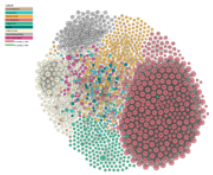
items	w	x	y	z
a	4	3		
b		4		1
c			3	4
d	2	4		

Content-based features
Text was *tokenized* and *stemmed* through **Lucene** and **Snowball**

+

Extended Features

LOD-based features
The most relevant features are extracted from **Dbpedia** by mapping **item names** to URIs



Graph-based features
Degree Centrality, Average Neighbor Degree, PageRank score, Node Redundancy and Cluster Coefficient, calculated with **Jung library**

Then we introduce extended features based on the Linked Open Data cloud

Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems. 136 : 1-14 (2017).*

LOD-based Recommender Systems

Methodology

Basic Features

Popularity features
#ratings, ratio of positive ratings

Collaborative features
We encoded a *column* of the users/items matrix

users

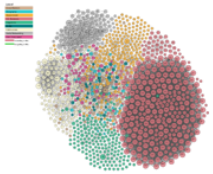
items	w	x	y	z
a	4	3		
b		4		1
c			3	4
d	2	4		

Content-based features
Text was *tokenized* and *stemmed* through **Lucene** and **Snowball**

+

Extended Features

LOD-based features
The most relevant features are extracted from **Dbpedia** by mapping **item names** to URIs



Graph-based features
Degree Centrality, Average Neighbor Degree, PageRank score, Node Redundancy and Cluster Coefficient, calculated with **Jung library**

→

Recommendation Framework

Item Representation
Items represented through different combinations of **basic and extended features**

Recommendation
Unseen items are labeled as **relevant** or **not relevant** by a classification algorithm

We used them to feed a hybrid classification framework

Best performance using both LOD and graph-based features

Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems. 136 : 1-14 (2017).*

Recap

encoding **exogenous semantics**

- **Exogenous** techniques use external knowledge sources to inject **semantics**
- **Entity Linking algorithms** focus on the identification of the **entities**
  Recent approaches also identify common sense terms
- **Knowledge graphs** used to feed recommendation models with **relevant features** and can help to overcome the **limited content analysis** problem