Recent Developments of Content-Based RecSys

Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro

Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

SWAP researchgroup

Recent Developments of Content-Based RecSys

Introduction

Giovanni Semeraro

Department of Computer Science
University of Bari Aldo Moro, Italy

About us

marco.degemmis@uniba.it

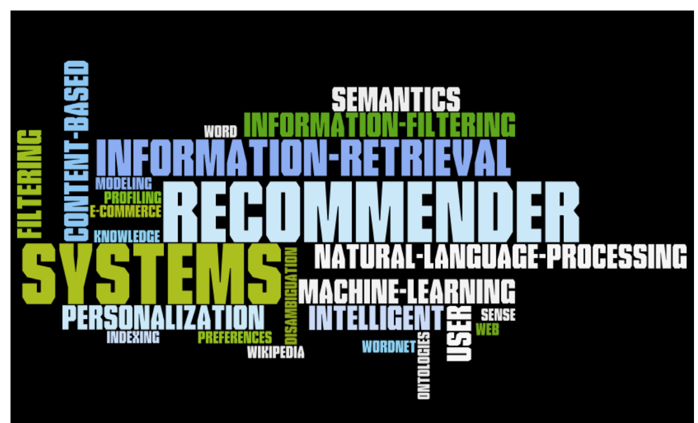
pasquale.lops@uniba.it

cataldo.musto@uniba.it

fedelucio.narducci@uniba.it

giovanni.semeraro@uniba.it

Semantic
Web
Access and
Personalization
"Antonio Bello" research group
http://www.di.uniba.it/~swap



in this tutorial…

how to represent content

to improve **information access** and build a new generation of services for **user modeling** and **recommender systems**?

Agenda

Why?

Why do we need **intelligent information access**?
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
Serendipity in Recommender Systems

Agenda

Why?

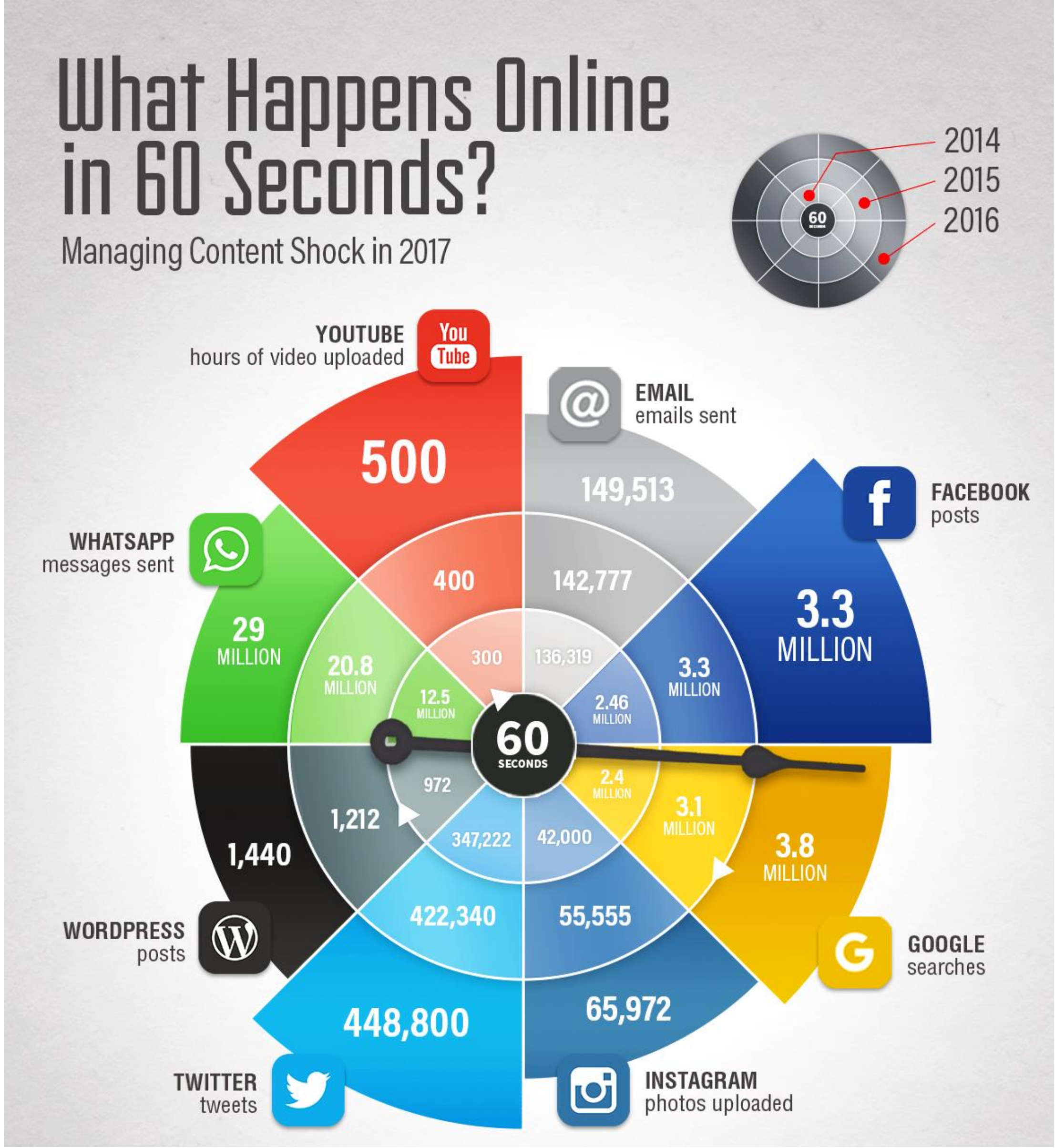
Why do we need **intelligent information access**?
Why do we need **content**?
Why do we need **semantics**?

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
Serendipity in Recommender Systems



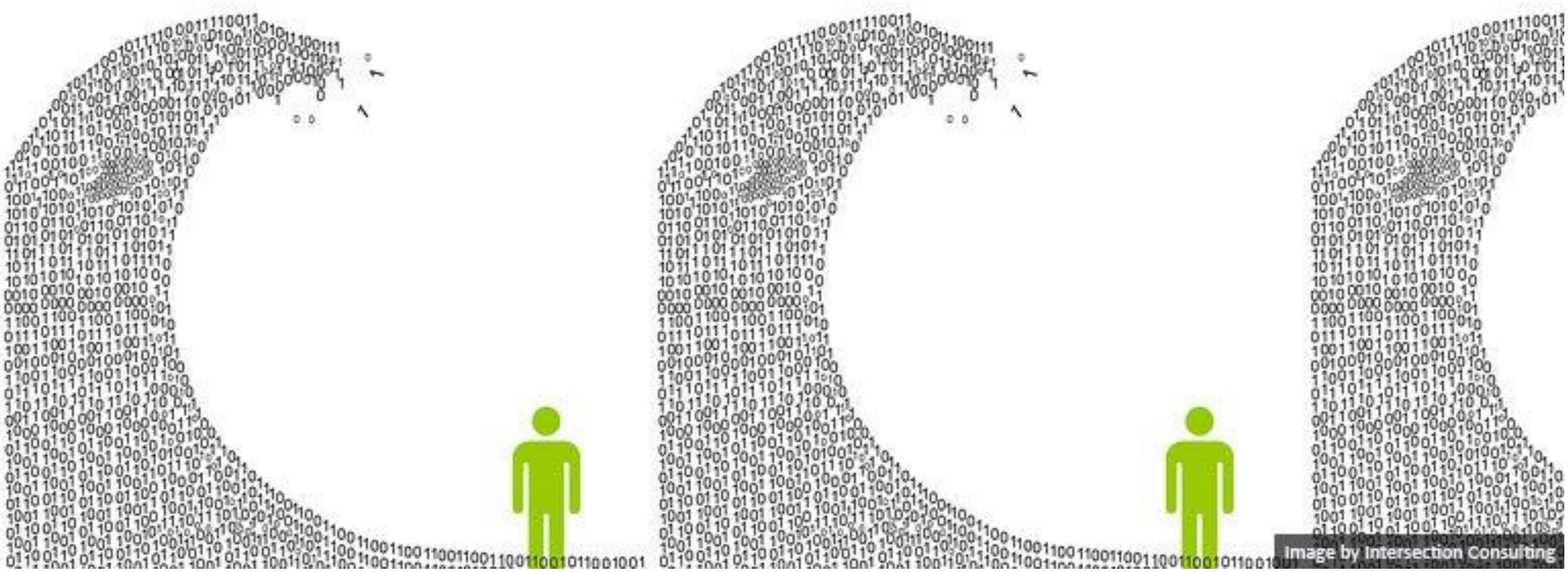
Sources: Email: 2014-2016: Radicati; Facebook: 2015 Qmee; 2016 Wishpond; Google: 2014 Statista; 2015 AdWeek; 2016 Internet Live Stats; Instagram: 2014 Tech Crunch; 2015 Nuke Suite; 2016 Instagram; Twitter: 2014 Internet Live Stats; 2015 Internet Live Stats; 2016 Tech Insider; WordPress: 2014 WordPress; 2015 WordPress; 2016 Internet Live Stats; WhatsApp: 2014 Fierce Mobile IT; 2015 Slash Gear; 2016 Expanded Ramblings; *YouTube:* 2014 Youtube Global Blog; 2015 Reel SEO.

physiologically

impossible

to follow the information flow

in **real time**

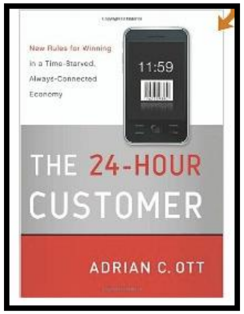


we can handle
126 bits of
information/second

we deal with
393 bits of
information/second

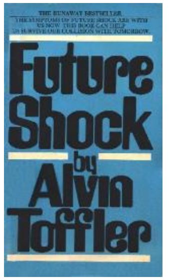
ratio: more than **3x**

source: Adrian C.Ott,
The 24-hour customer,
HarperCollins, 2010



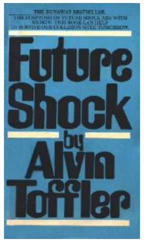
Information overload

Appeared for the first time in 1964 in «The Managing of Organizations» by Bertram Gross, popularized by Alvin Toffler in his best-seller «Future Shock» (1970)



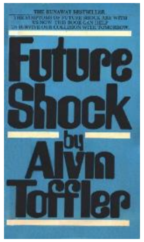
Information overload

Appeared for the first time in 1964 in «The Managing of Organizations» by Bertram Gross, popularized by Alvin Toffler in his best-seller «Future Shock» (1970)

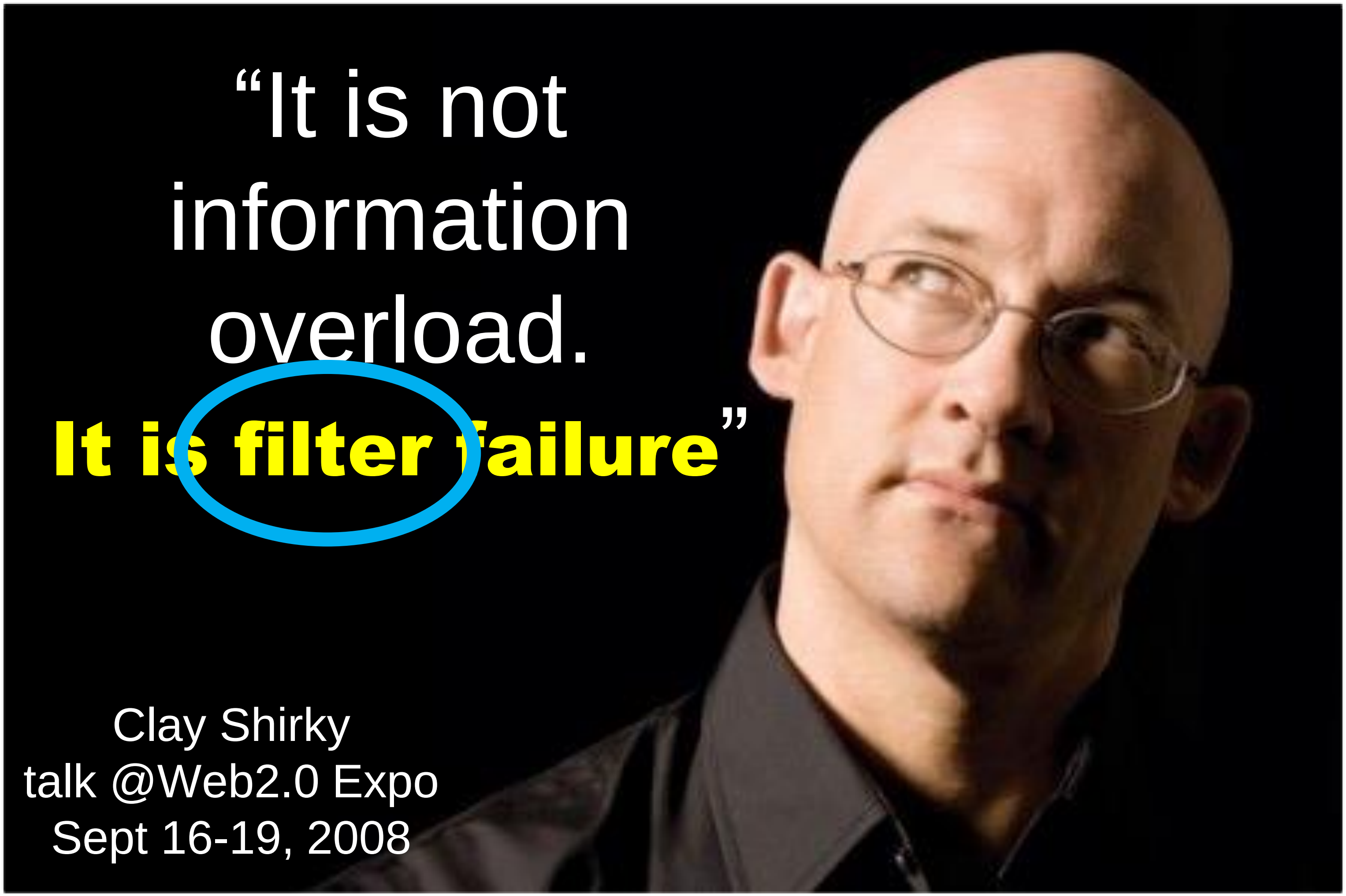


Information overload

Appeared for the first time in 1964 in «The Managing of Organizations» by Bertram Gross, popularized by Alvin Toffler in his best-seller «Future Shock» (1970)



Information overload



Challenge

To effectively cope with

information overload

& **bounded rationality**

we need to **filter** the information flow

We need technologies and algorithms for

intelligent information access

… and we already have some evidence!

Intelligent Information Access

success stories

Information Retrieval (Search Engines)

Intelligent Information Access

success stories

Information Filtering (Recommender Systems)

Agenda

Why?

Why do we need **intelligent information access**?
Why do we need **content**?
Why do we need **semantics**?

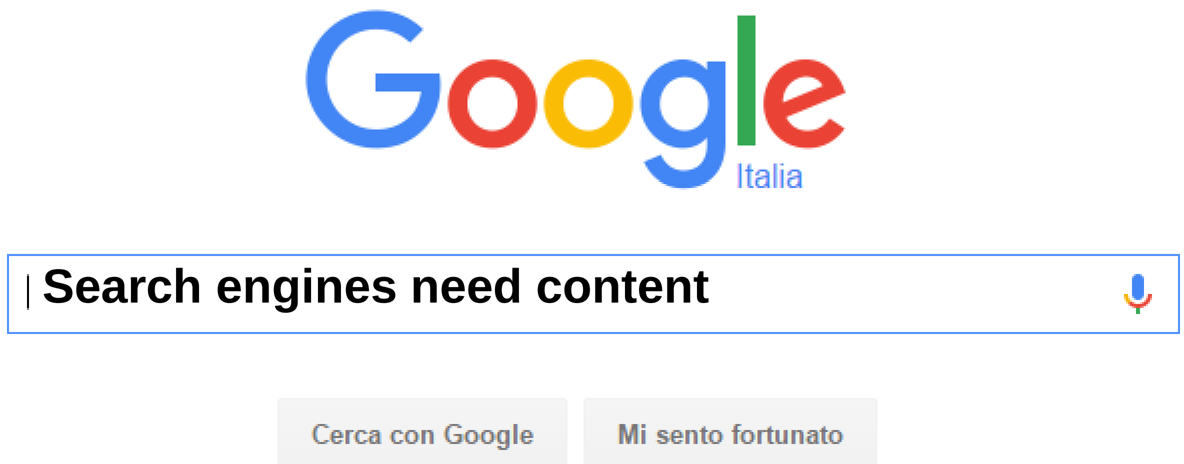
How?

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

What?

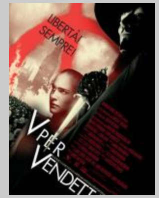
Explanation of Recommendations
Serendipity in Recommender Systems

Why do we need content?



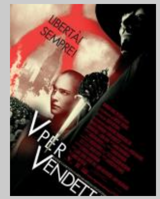
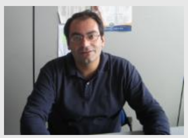
Trivial: search engines can’t work without content

Why do we need content?

					
	✓	✓		✓	
		✓			
	✓		✓		
				✓	✓
	✓	✓			

Recommender Systems: <u>not trivial!</u>

Why do we need content?

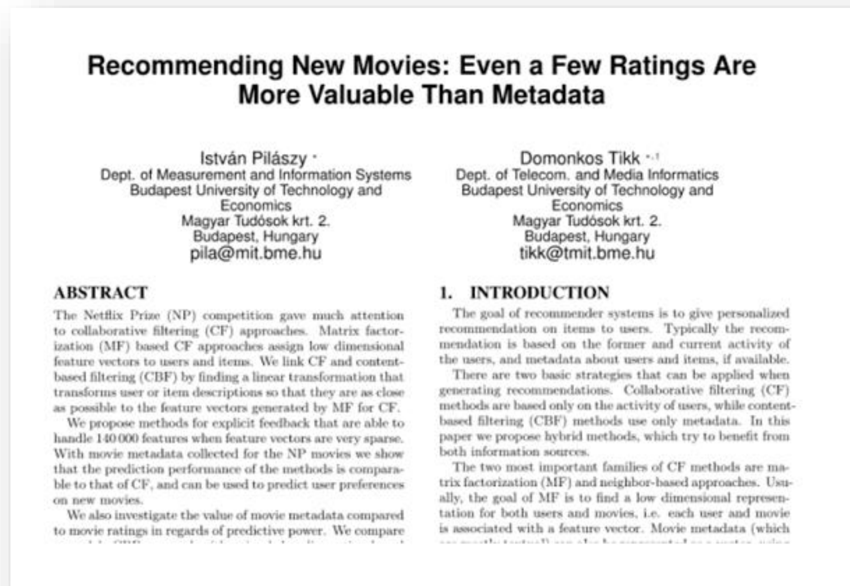
					
	✓	✓		✓	
		✓			
	✓		✓		
				✓	✓
	✓	✓			

Recommender Systems <u>can work</u> without content

Why do we need content?

Several Recommender Systems **perfectly work** **using no content!**

Collaborative Filtering (CF), **Matrix Factorization (MF)** and **Tensor Factorization (TF)** are state-of-the-art techniques for implementing Recommender Systems



Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy *
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk *,†
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that transforms user or item descriptions so that they are as close as possible to the feature vectors generated by MF for CF.

We propose methods for explicit feedback that are able to handle 140 000 features when feature vectors are very sparse. With movie metadata collected for the NP movies we show that the prediction performance of the methods is comparable to that of CF, and can be used to predict user preferences on new movies.

We also investigate the value of movie metadata compared to movie ratings in regards of predictive power. We compare

1. INTRODUCTION

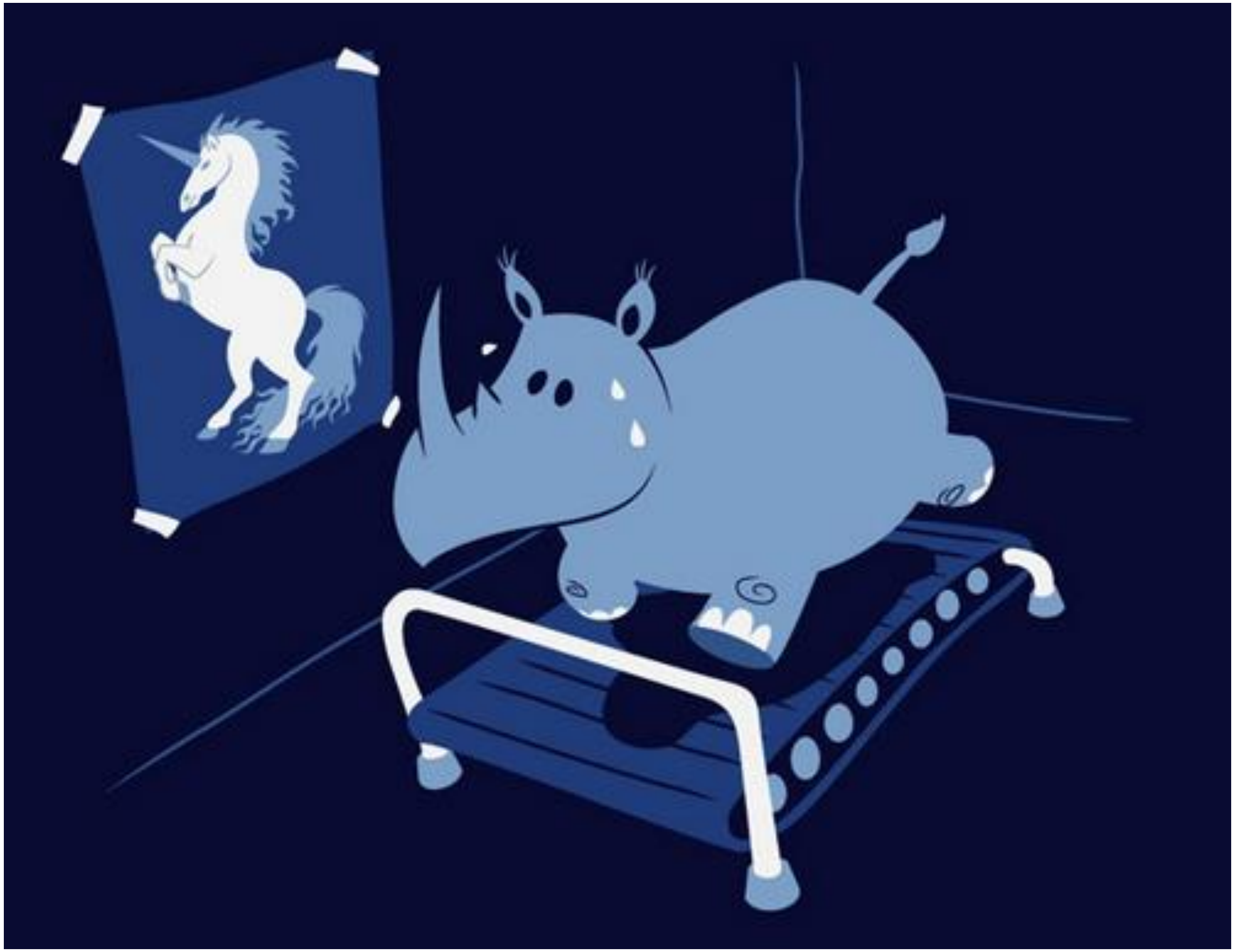
The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available.

There are two basic strategies that can be applied when generating recommendations. Collaborative filtering (CF) methods are based only on the activity of users, while content-based filtering (CBF) methods use only metadata. In this paper we propose hybrid methods, which try to benefit from both information sources.

The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which

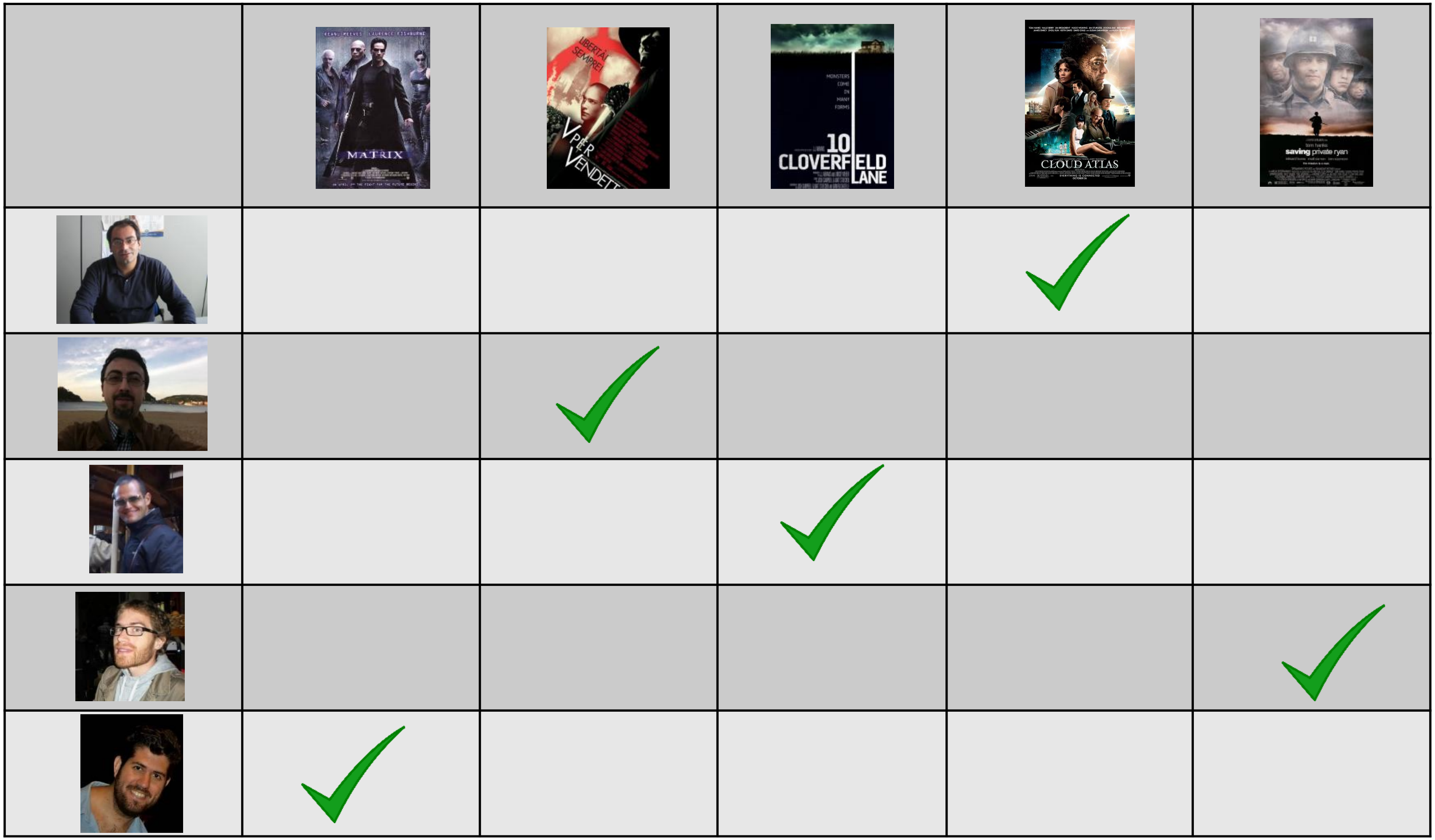
ACM RecSys 2009 paper
by Netflix Challenge winners

Why do we need content?



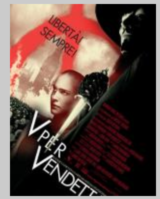
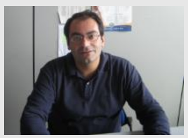
Content **can tackle some issues** of **Collaborative Filtering**

Why do we need content?



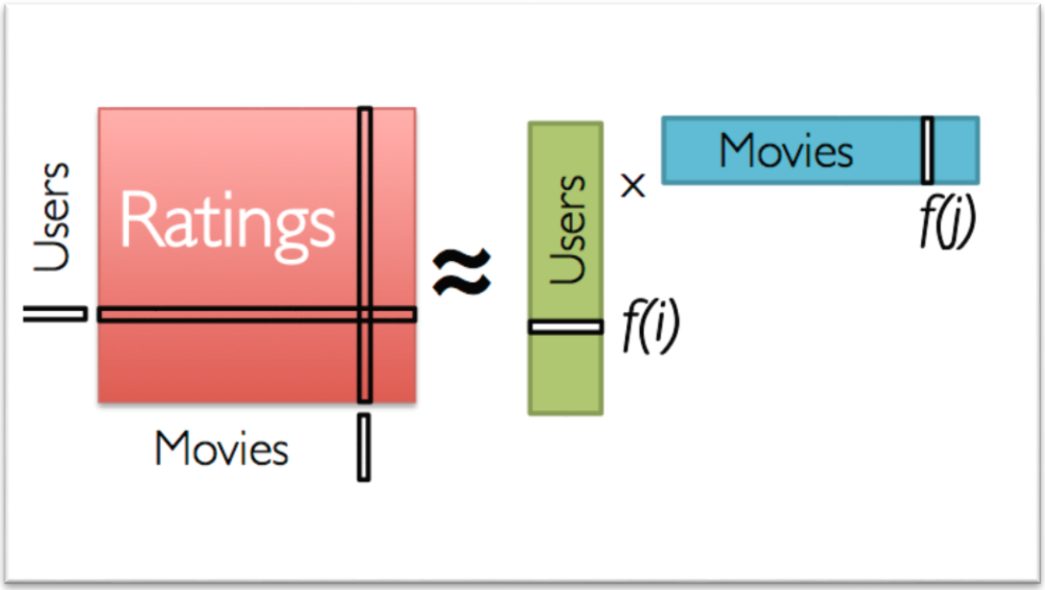
Collaborative Filtering issues: <u>sparsity</u>

Why do we need content?

					
	✓	✓		✓	
		✓			?
			✓		
				✓	
	✓	✓			

Collaborative Filtering issues: **<u>new item problem</u>**

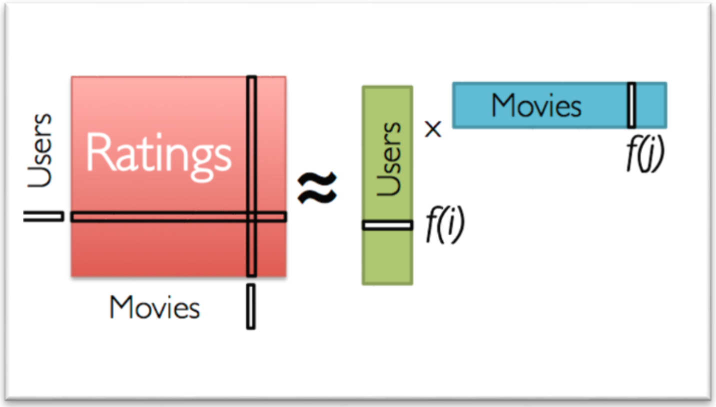
Why do we need content?



Why?

Collaborative Filtering issues: lack of transparency!

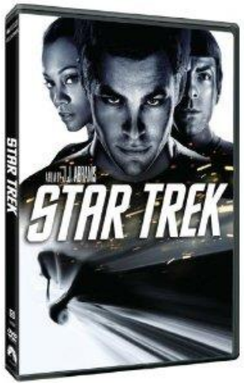
Why do we need content?



Who knows the «Customers Who Bought This Item …»?
Information Asymmetry

Collaborative Filtering issues: <u>poor explanations!</u>

Why do we need content?



accurate but obvious

⇩

not useful

- Content-based RecSys suggest items whose scores are high when matched against the user profile
  - ✓ the user is recommended items similar to those already liked in the past
  - ✓ No straight method for finding something unexpected → Overspecialization

Obviousness of recommendations!

[McNee06] S.M. McNee, J. Riedl, and J. Konstan. Accurate is not always good: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems*, pages 1-5, Canada, 2006.

Recap #1

Why do we need content?

- **In general:** to extend and improve user modeling
- To exploit the information **spread on social media**
- To overcome **typical issues of collaborative filtering** and matrix factorization
- Because **search engines can't simply work** without content ☺

Agenda

Why?

Why do we need **intelligent information access**?
Why do we need **content**?
Why do we need **semantics**?

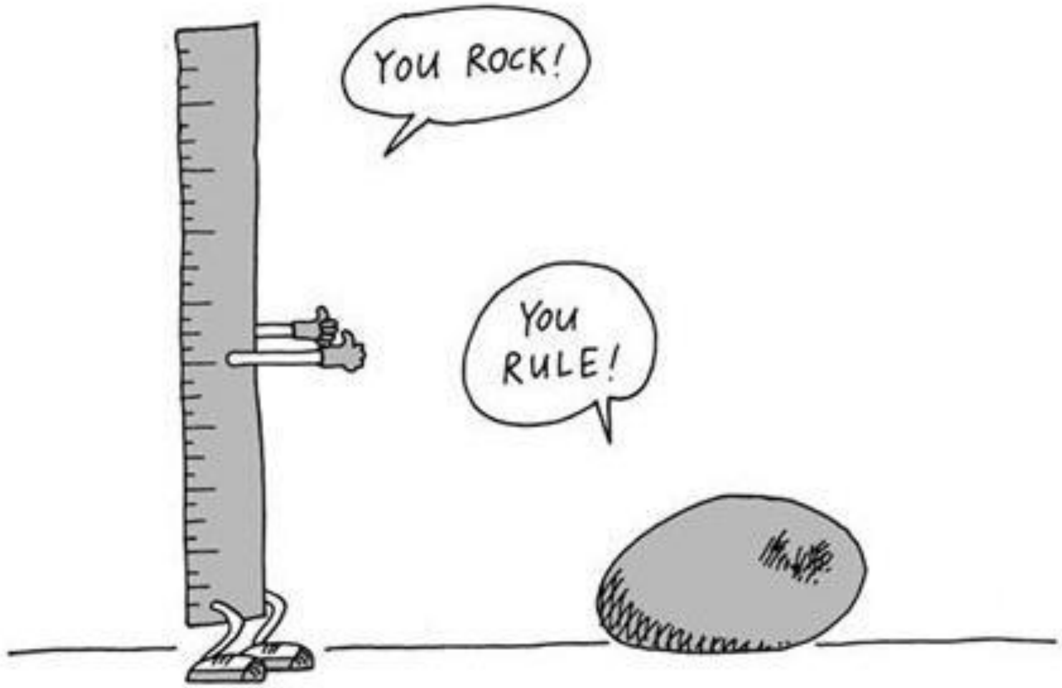
How?

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
Serendipity in Recommender Systems

Why do we need semantics?



Deep Rationality requires a **deep comprehension** of the information conveyed by textual content. To achieve that goal it is crucial to **improve the quality of user profiles** and the **effectiveness of intelligent information access platforms.**

Basics: Content-based RecSys (CBRS)

Suggest items similar to those the user liked in the past

Recommendations generated by matching the **description of items** with the **profile of the user's interests**

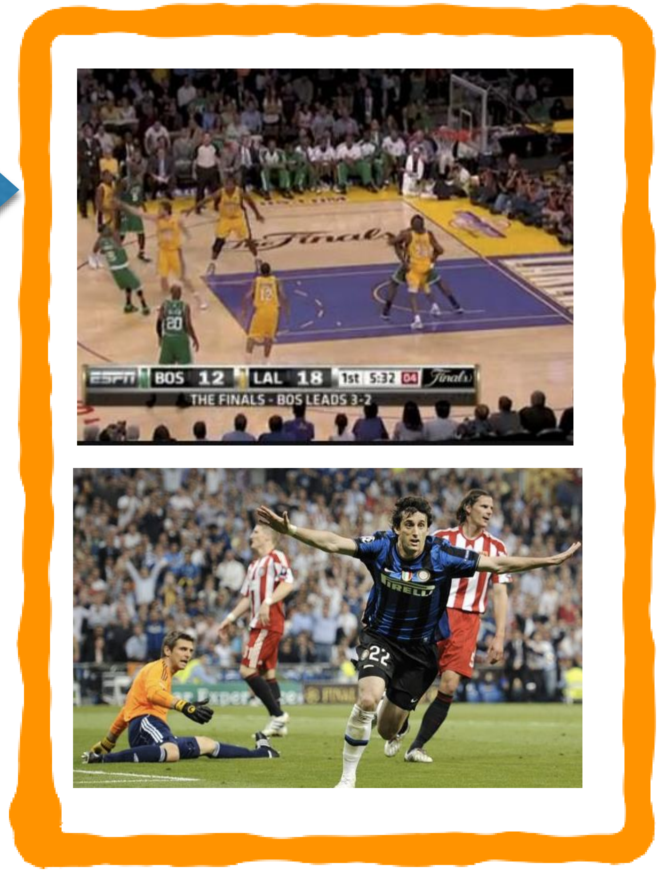
use of specific **features**

[Lops11] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.), *Recommender Systems Handbook,* Springer, 73–105, 2011.

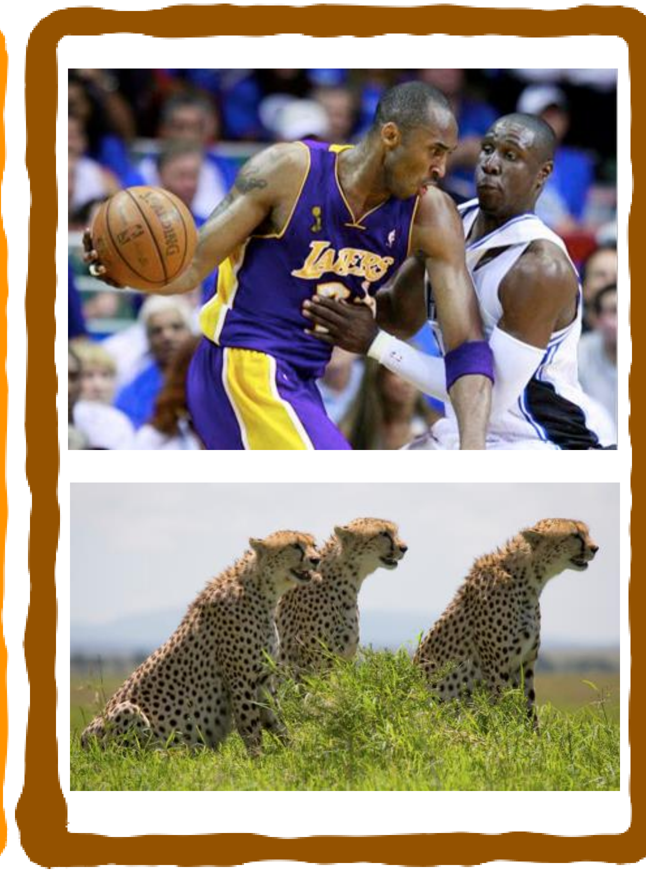
[Pazzani07] Pazzani, M. J., & Billsus, D. Content-Based Recommendation Systems. *The Adaptive Web*. Lecture Notes in Computer Science vol. 4321, 325-341, 2007.

Basics: Content-based RecSys (CBRS)

Recommendations are generated by matching the features stored in the user profile with those describing the items to be recommended.



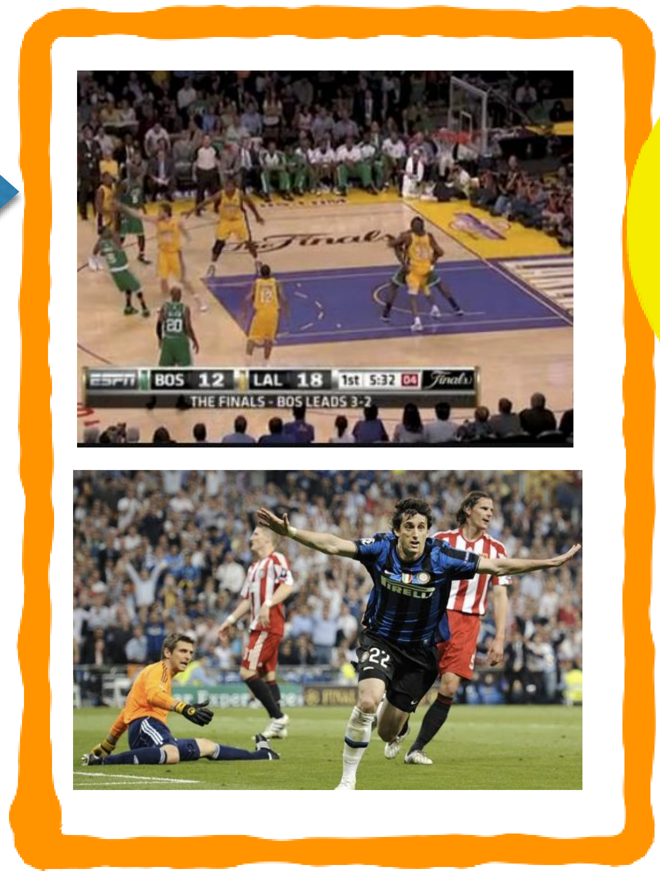
user profile



items

Basics: Content-based RecSys (CBRS)

Recommendations are generated by matching the features stored in the user profile with those describing the items to be recommended.

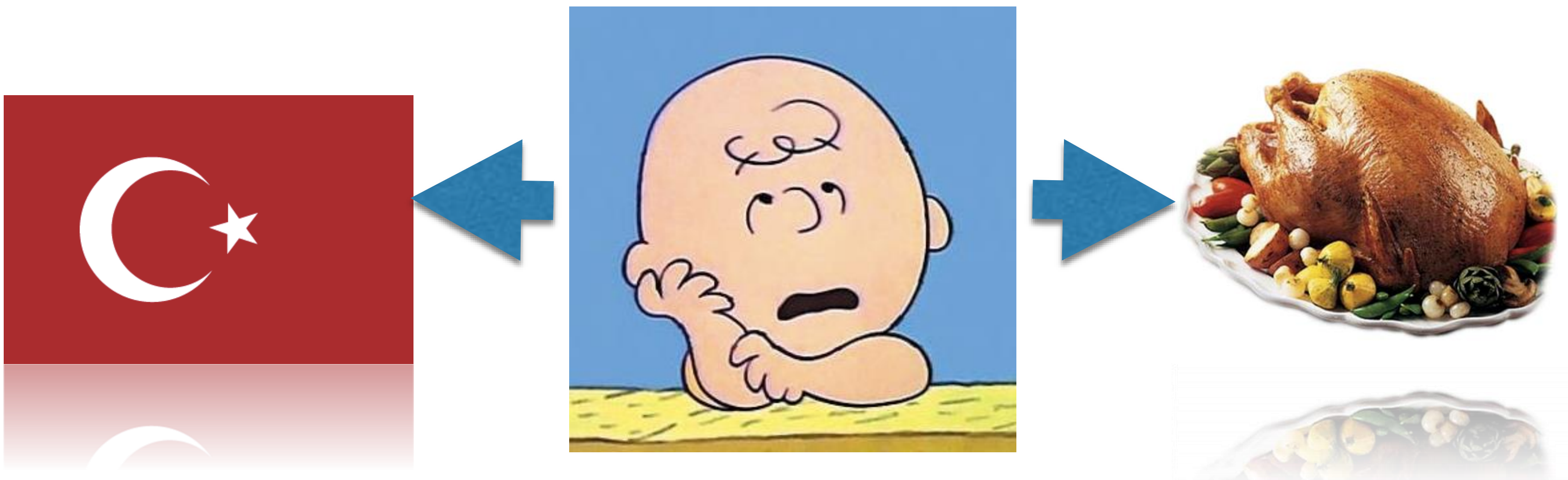


user profile

items

Lack of Semantics in User Models

"I love turkey. It's my choice for these #holidays!



Social Media can be helpful to avoid cold start

Lack of Semantics in User Models

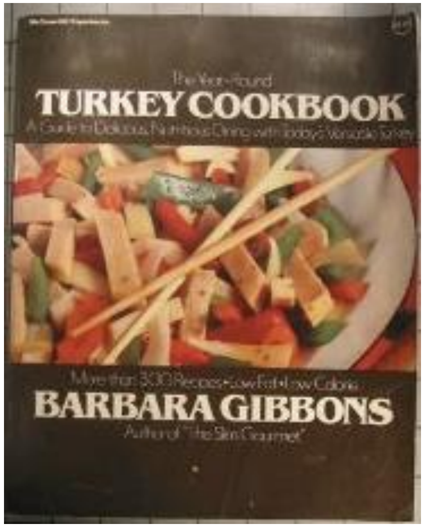
"I love turkey. It's my choice for these #holidays!

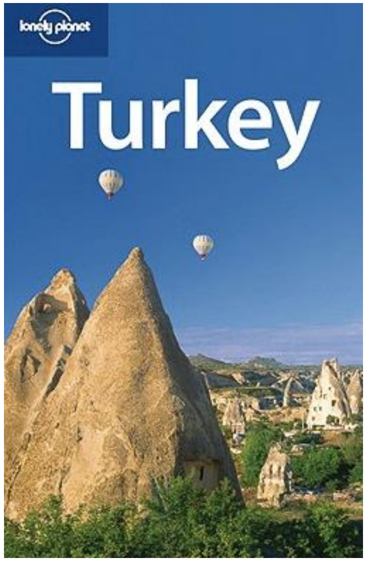
..but pure content-based representations

can't handle **polysemy**

Lack of Semantics in User Models

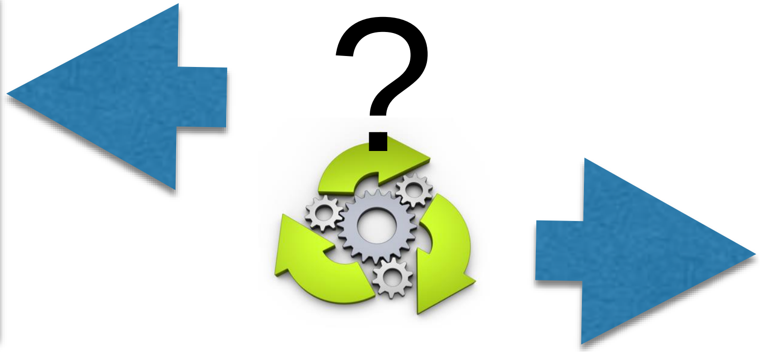
"I love turkey. It's my choice for these #holidays!





Pure content-based representations can easily drive a recommender system towards failures!

Lack of Semantics in Social Media Analysis



What are people worried about?

Are they worried about the eagle

or about the city of L'Aquila?

Lack of Semantics in User Models

…is not only about **polysemy**

doc1
AI is a branch of computer science

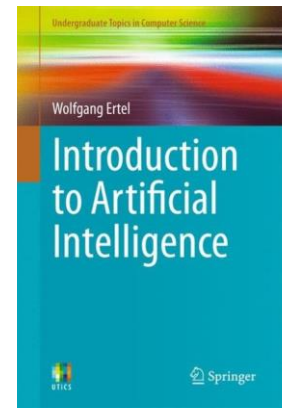
doc2
the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

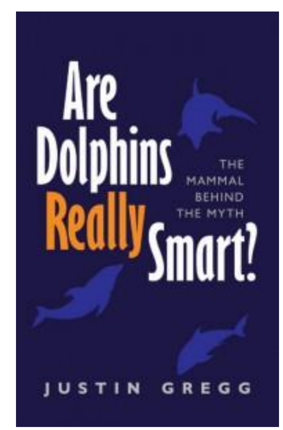
doc3
apple launches a new product…

USER PROFILE

artificial	0.11
intelligence	0.12
apple	0.20
AI	0.18
…	

?





Book recommendation

multi-word concepts

Lack of Semantics in User Models

…is not only about **polysemy**

doc1
AI is a branch of computer science

doc2
the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3
apple launches a new product…

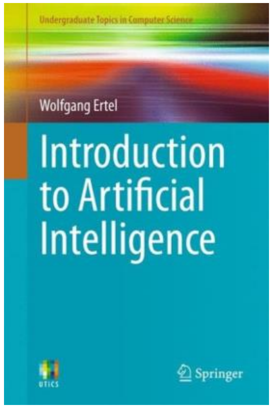
USER PROFILE

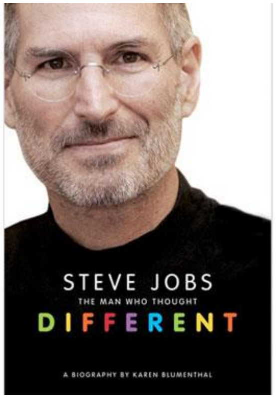
artificial	0.11
intelligence	0.12
apple	0.20
AI	0.18
…	

Most of the preferences regard AI, but «apple» is the most relevant feature in the profile due to synonymy

synonymy

?





Book recommendation

Lack of Semantics in CBRS

L'Italia prepara la sfida impossibile Con la Francia un miracolo non basta

SIAULIAI (Lituania), 3 settembre 2011

Gli azzurri devono battere Parker e soci, ancora senza sconfitte, con almeno 13 punti di scarto per continuare a sperare. Il c.t. Pianigiani ammette: "Non valiamo le prime 10 d'Europa"

87 | 10
Mi piace | Tweet
19 | 1
DI LA TUA | +1
Share

Invia articolo
Versione stampabile
Ascolta

PER SAPERNE DI PIÙ
Risultati e classifiche

Gli azzurri festeggiano la vittoria sulla Lettonia. Ansa

PIÙ LETTI | PIÙ COMMENTATI

Juve, la notte dell'orgoglio, Il nuovo stadio emoziona
Moratti: "Vicini a Gasperini, Il caso Forlan è grave"
Milan, la macchina da gol , contro Klose e Cisse
Portieri: sarà l'anno di Mirante?, Mutu, il riscatto

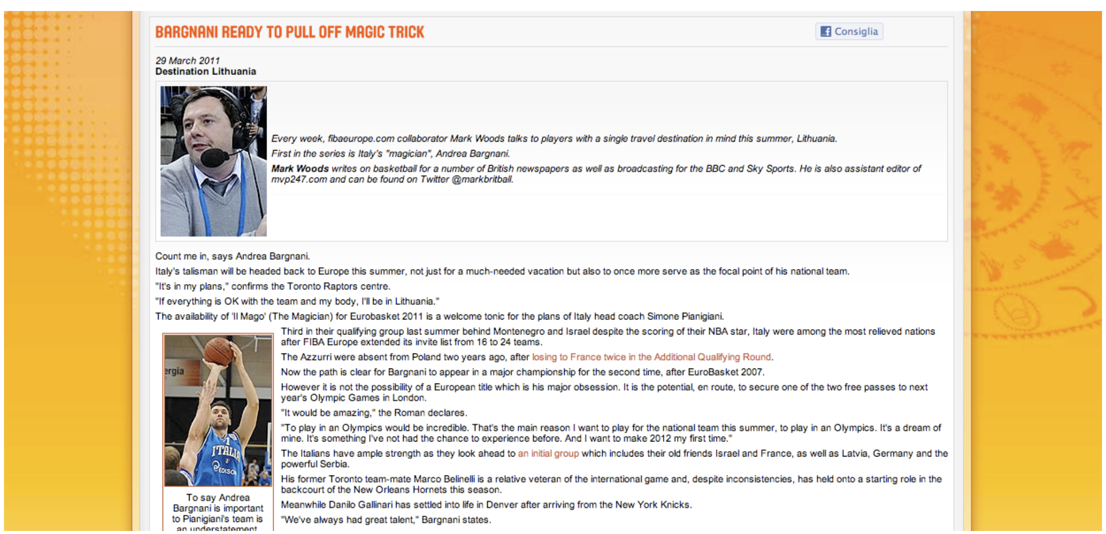
MAGIC LIBRO 2011
Non perdere nemmeno un colpo all'asta d'inizio anno! A soli 7,99 € in edicola

EXTRA MAGIC CHAMPIONS
Una squadra più forte di Barça e Red Devils? Tu puoi crearla. Gioca gratis e vinci!

LE GRANDI STORIE DELL'AUTO
Porta le grandi storie dell'auto sempre con te!Scarica subito per iPad e iPhone a SOLI 2.99€.

LINOMANIA

italian



BARGNANI READY TO PULL OFF MAGIC TRICK

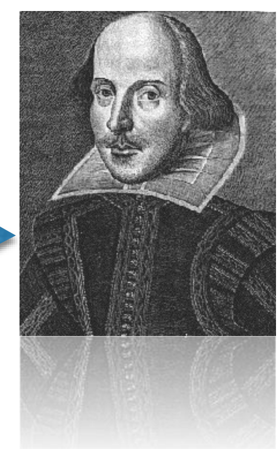
Consiglia

29 March 2011
Destination Lithuania

Every week, fibaeurope.com collaborator Mark Woods talks to players with a single travel destination in mind this summer, Lithuania.
First in the series is Italy's "magician", Andrea Bargnani.
***Mark Woods** writes on basketball for a number of British newspapers as well as broadcasting for the BBC and Sky Sports. He is also assistant editor of mvp247.com and can be found on Twitter @markbritball.*

Count me in, says Andrea Bargnani.
Italy's talisman will be headed back to Europe this summer, not just for a much-needed vacation but also to once more serve as the focal point of his national team.
"It's in my plans," confirms the Toronto Raptors centre.
"If everything is OK with the team and my body, I'll be in Lithuania."
The availability of 'Il Mago' (The Magician) for Eurobasket 2011 is a welcome tonic for the plans of Italy head coach Simone Pianigiani.
Third in their qualifying group last summer behind Montenegro and Israel despite the scoring of their NBA star, Italy were among the most relieved nations after FIBA Europe extended its invite list from 16 to 24 teams.
The Azzurri were absent from Poland two years ago, after losing to France twice in the Additional Qualifying Round.
Now the path is clear for Bargnani to appear in a major championship for the second time, after EuroBasket 2007.
However it is not the possibility of a European title which is his major obsession. It is the potential, en route, to secure one of the two free passes to next year's Olympic Games in London.
"It would be amazing," the Roman declares.
"To play in an Olympics would be incredible. That's the main reason I want to play for the national team this summer, to play in an Olympics. It's a dream of mine. It's something I've not had the chance to experience before. And I want to make 2012 my first time."
The Italians have ample strength as they look ahead to an initial group which includes their old friends Israel and France, as well as Latvia, Germany and the powerful Serbia.
His former Toronto team-mate Marco Belinelli is a relative veteran of the international game and, despite inconsistencies, has held onto a starting role in the backcourt of the New Orleans Hornets this season.
Meanwhile Danilo Gallinari has settled into life in Denver after arriving from the New York Knicks.
"We've always had great talent," Bargnani states.

To say Andrea Bargnani is important to Pianigiani's team is an understatement



english

Lack of Semantics in CBRS


L'Italia prepara la sfida impossibile
Con la Francia un miracolo non basta

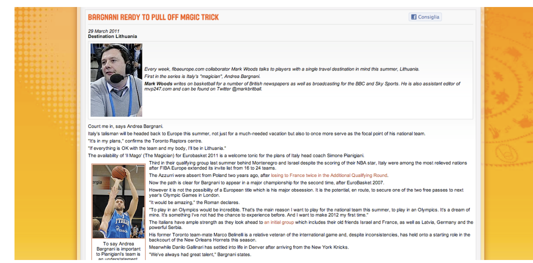
SIAULIAI (Lituania), 3 settembre 2011

Gli azzurri devono battere Parker e soci, ancora senza sconfitte, con almeno 13 punti di scarto per continuare a sperare. Il c.t. Pianigiani ammette: "Non valiamo le prime 10 d'Europa"

PIÙ LETTI | PIÙ COMMENTATI

Gli azzurri festeggiano la vittoria sulla Lettonia. Ansa

Italian-language news about basketball


BARGNANI READY TO PULL OFF MAGIC TRICK

English-language news about basketball

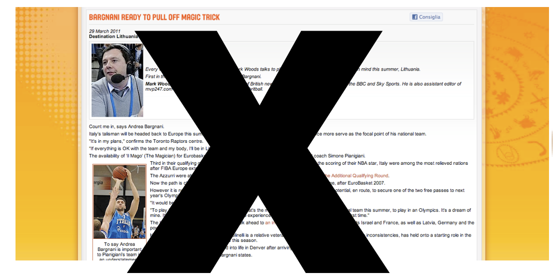
user profile **items**

Lack of Semantics in CBRS

It is likely that the algorithm **is not able to suggest a (relevant) English news** since there exist **no overlaps between the features!**

Italian-language news about basketball

user profile



English-language news about basketball

items

Lack of Semantics in CBRS

Content-based recommendations are language-dependent !

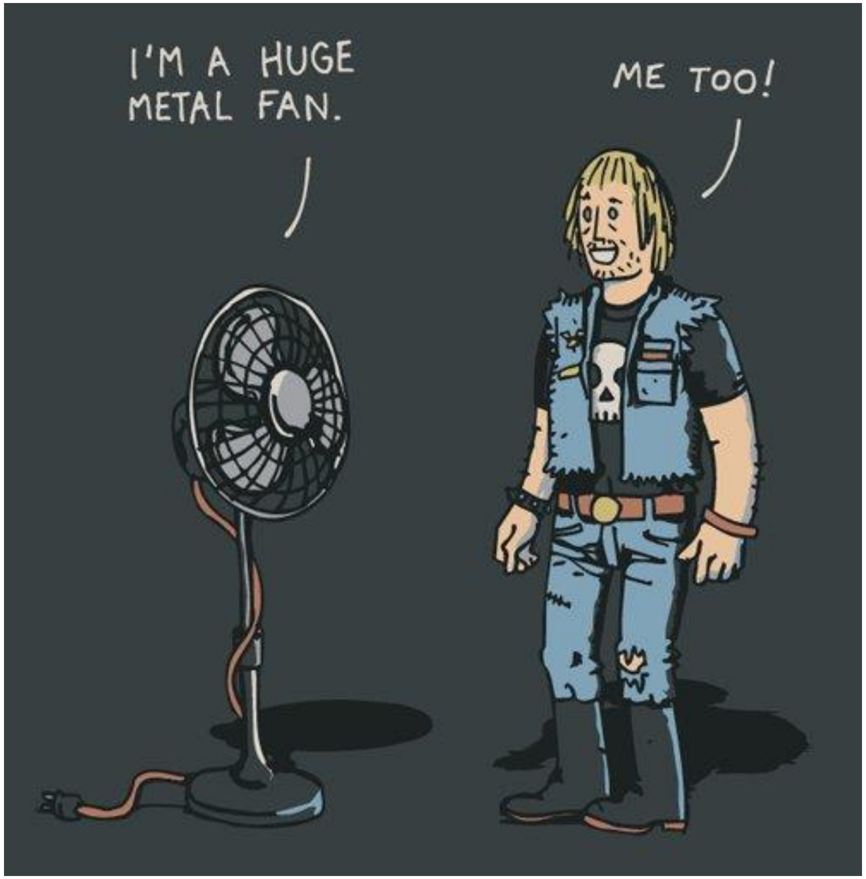
It is
algor
t
(rel
news since there exist
no overlaps between the features!

user profile

items

Recap #2

Why do we need semantics?



Because language is inherently ambiguous

[deG15] M. de Gemmis, P. Lops, C. Musto, F. Narducci and G. Semeraro. Semantics-aware content-based recommender systems. In F. Ricci, L. Rokach, and B. Shapira (Eds.), *Recommender Systems Handbook*, *2nd Ed.*, Springer, 119-159, 2015.

- **In general: to improve content representation** in intelligent information access platforms
- To avoid typical **issues of natural language representations** (polysemy, synonymy, multi-word concepts, etc.)
- To **model user preferences in an effective way**
- To **better understand** the information spread on social media
- To provide **multilingual recommendations**

Recent Developments of Content-Based RecSys

Basics of NLP and Exogenous Techniques

Pasquale Lops

Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

SWAP researchgroup

Agenda

Why?

Why do we need **intelligent information access**?
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
Serendipity in Recommender Systems

Information Retrieval and Filtering

Two sides of the same coin (Belkin&Croft,1992)

Information Retrieval

information need expressed through a **query**

goal: retrieve information which might be **relevant** to a user

Information Filtering

information need expressed through a **user profile**

goal: expose users to only the information that is **relevant** to them, rding to personal profiles

It's all about searching!

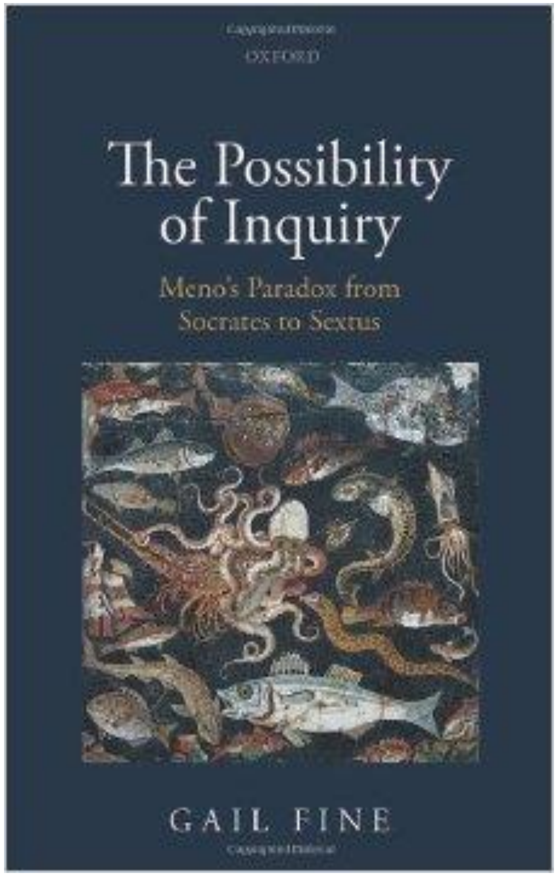
[Belkin&Croft, 1992] Belkin, Nicholas J., and W. Bruce Croft. "Information filtering and information retrieval: Two sides of the same coin?." *Communications of the ACM* 35.12 (1992): 29-38.

Search (and Content-based Recommendation) is not so simple as it might seem

Meno's Paradox of Inquiry:

Meno: and how will you enquire, Socrates, into that which you do not know? **What will you put forth as the subject of enquiry**? And if you find what you want, **how will you know that this is the thing you did not know**?

Socrates: I know, Meno, what you mean; but just see what a tiresome dispute you are introducing. **You argue that a man cannot search either for what he knows or for what he does not know**; if he knows it, there is no need to search; and if not, he cannot; he does not know the very subject about which he is to search.



Plato Meno 80d-81a

http://www.gutenberg.org/etext/1643

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support** **(as intelligent information access technologies)**

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support** **(as intelligent information access technologies)**

We need to **better understand and represent the content**

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

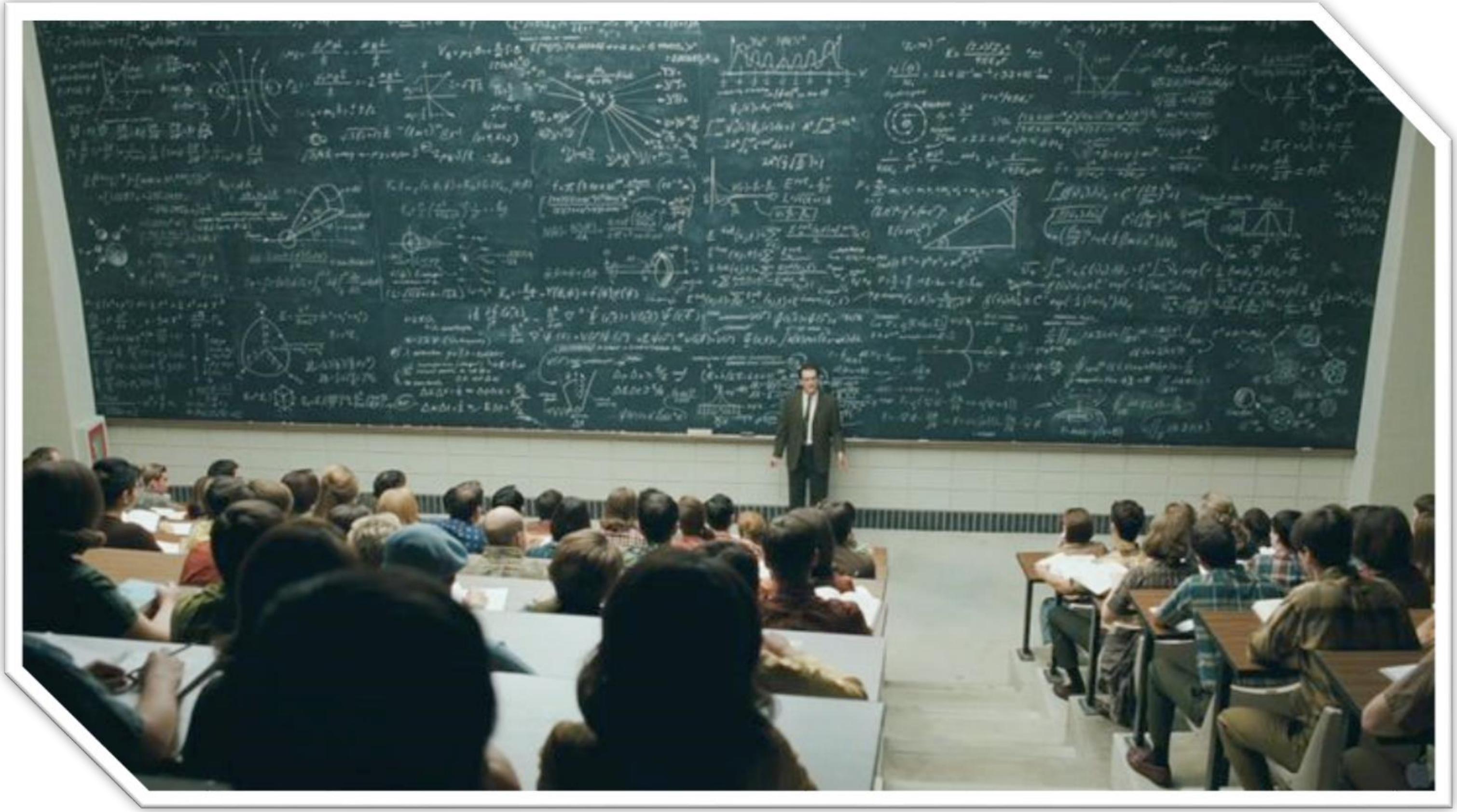
How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support (as intelligent information access technologies)**

We need to **better understand and represent the content**

...before semantics

some **basics**



of **Natural Language Processing (NLP)**

Agenda

Why?

Why do we need **intelligent information access**?
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?

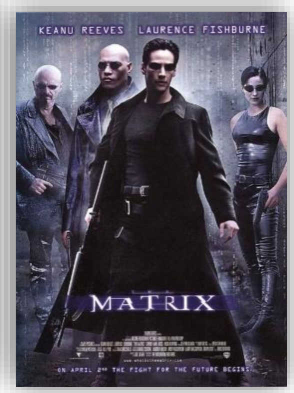
Explanation of Recommendations
Serendipity in Recommender Systems

Scenario

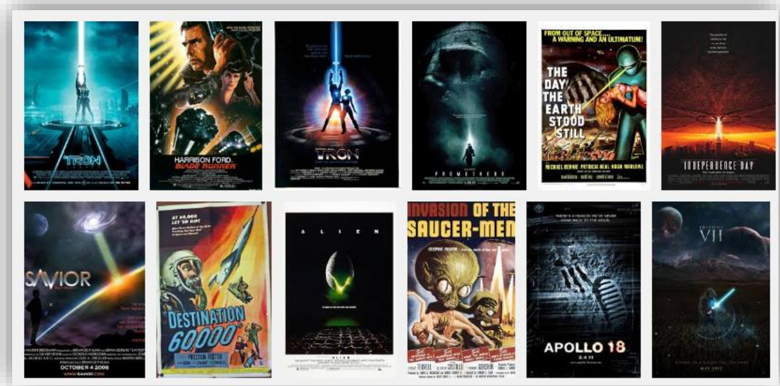
Pasquale really loves the movie «The Matrix», and he asks a content-based recommender system for some suggestions.

Question

How can we **feed the algorithm with some textual features** related to the **movie to build a (content-based) profile** and provide recommendations?

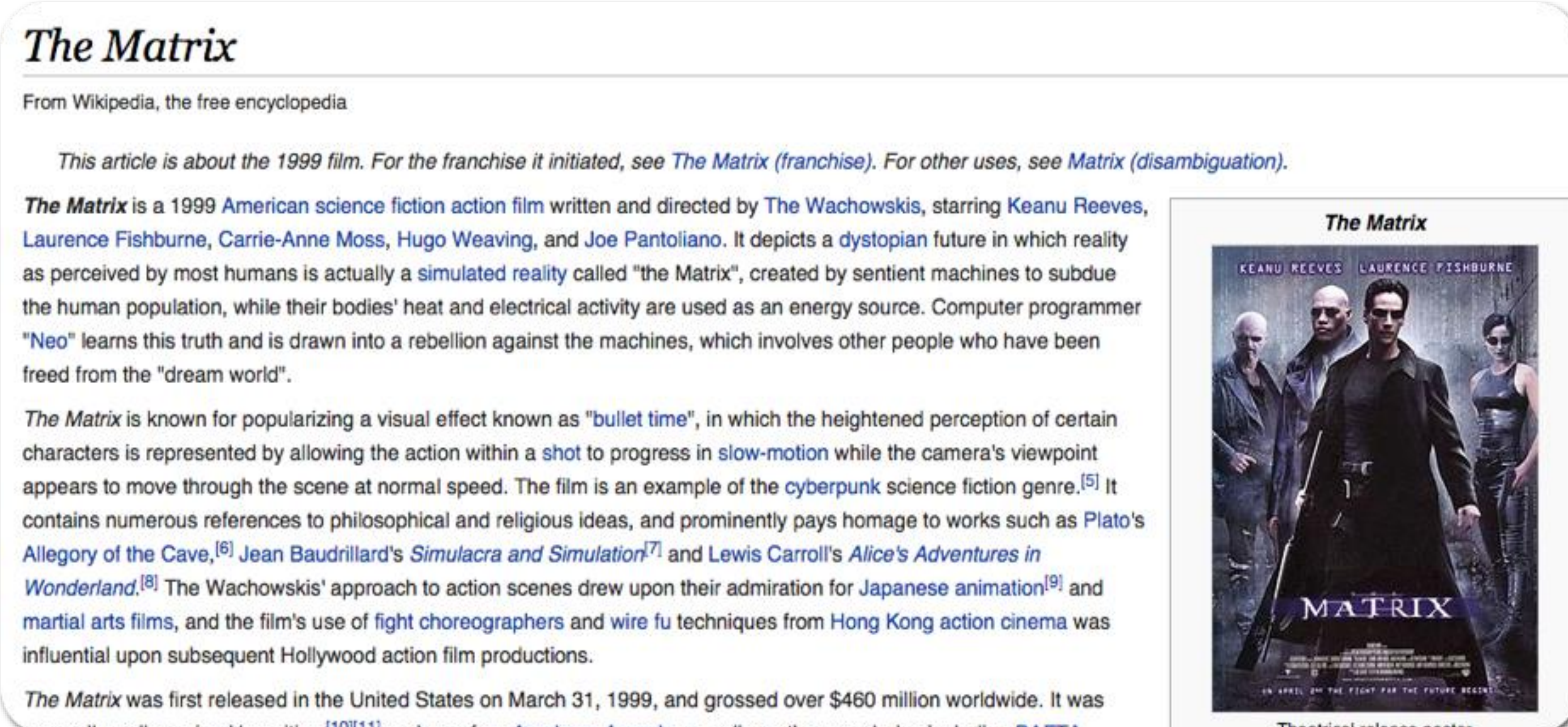


Recommendation Engine



?

Scenario



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

the plot can be a **rich source** of **content-based features**

Scenario

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

the plot can be a **rich source** of **content-based features**

…but we need to **properly process it** through a pipeline of **Natural Language Processing** techniques

Basic NLP operations

- ✓ **normalization** strip unwanted characters/markup (e.g. HTML/XML tags, punctuation, numbers, etc.)
- ✓ **tokenization** break text into tokens
- ✓ **stopword removal** exclude common words having little semantic content
- ✓ **lemmatization** reduce inflectional/variant forms to base form (lemma in the dictionary), *e.g. am, are, is → be*
- ✓ **stemming** reduce terms to their "roots", *e.g. automate(s), automatic, automation all reduced to* ***automat***

vocabulary

Example

The Matrix is a 1999 American-Australian neo-noir 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".

Example

The Matrix is a 1999 American-Australian neo-noir 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".

normalization

Example

The Matrix is a 1999 American Australian neo noir 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

tokenization

Tokenization issues

compound words

- science-fiction: break up **hyphenated** sequence?
- Keanu Reeves: **one token or two**? How do you decide it is one token?

numbers and dates

- 3/20/91 Mar. 20, 1991 20/3/91
- 55 B.C.
- (800) 234-2333

Tokenization issues

language issues

- German noun compounds not segmented
  Lebensversicherungsgesellschaftsangestellter means
  life insurance company employee
- Chinese and Japanese have no spaces between words (not always guaranteed a unique tokenization)

  莎拉波娃现在居住在美国东南部的佛罗里达

- Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right

  استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

  Algeria achieved its independence in 1962 after 132 years of French occupation

Example

The Matrix is a 1999 American Australian neo noir 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

stopword removal

Example

The Matrix is a 1999 American Australian neo noir 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

stopword removal

Example

The Matrix is a 1999 American Australian neo noir 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 bodyies 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

lemmatization

Example

Matrix 1999 American Australian neo noir science fiction action film write direct Wachowskis star Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving Joe Pantoliano depict dystopian future reality perceived human simulate reality call Matrix create sentient machine subdue human population body heat electrical activity use energy source Computer programmer Neo learn truth draw rebellion against machine involve people free dream world

next step: to give a weight to each feature

(e.g. through TF-IDF)

Weighting features: TF-IDF

terms frequency – inverse document frequency best known weighting scheme in information retrieval.

Weight of a term as product of **tf weight** and **idf weight**

$$\mathrm{w}_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log(N / \mathrm{df}_t)$$

tf **number of times** the term occurs in the document

idf depends on **rarity** of a term in a collection

tf-idf increases with the number of occurrences within a document, and with the rarity of the term in the collection.

Example

Matrix 1999 American Australian neo noir science fiction action **film** write direct Wachowskis star Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving Joe Pantoliano depict **dystopian** future reality perceived human simulate reality call Matrix create sentient machine subdue human population body heat electrical activity **use** energy source Computer programmer Neo learn truth draw **rebellion** against machine involve people free dream world

green=high IDF

red=low IDF

The Matrix representation

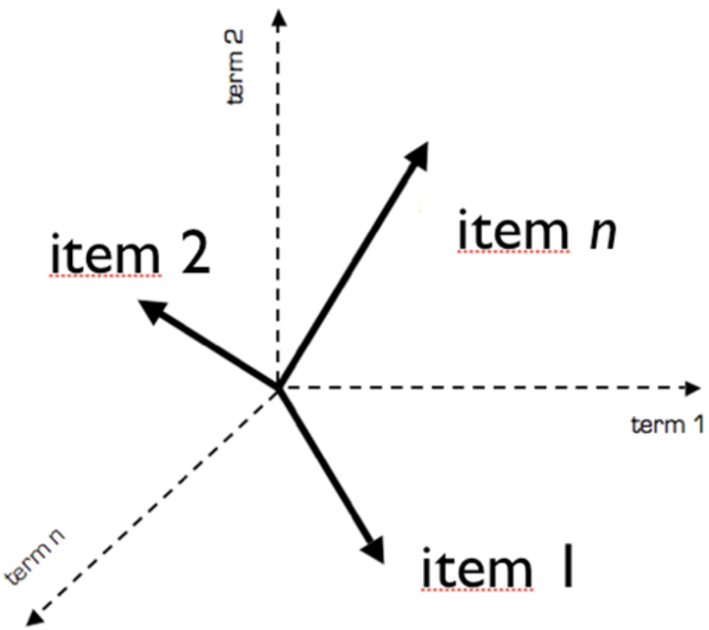
Matrix
1999
American
Australian
science
fiction
Hugo
…
world

keywords

a portion of Pasquale's **content-based profile**

given a content-based profile, we can **easily build a basic recommender system** through **Vector Space Model** and **similarity measures**

Vector Space Model (VSM)



given a set of *n* **features** (vocabulary)

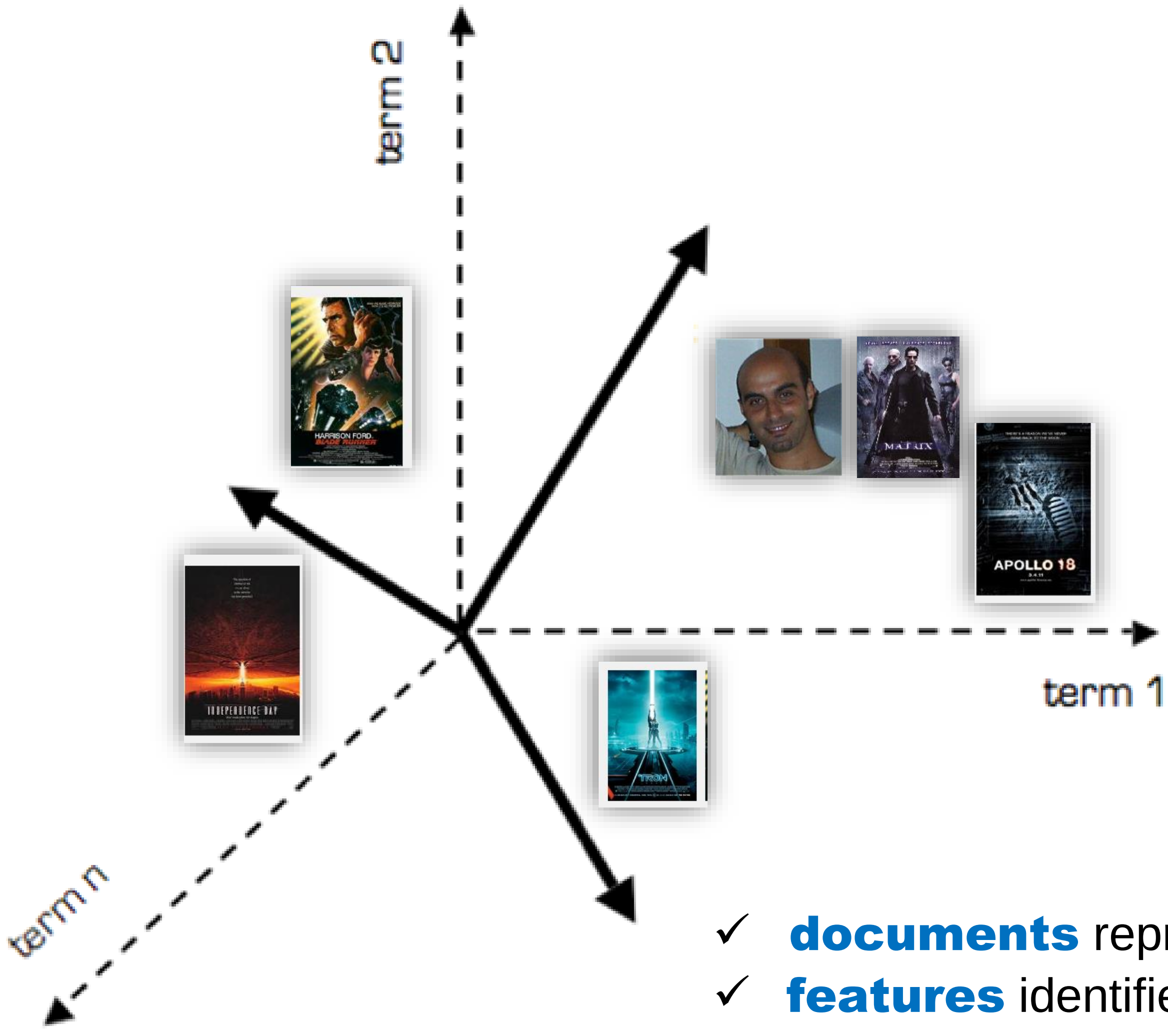
$$f = \{f_1, f_2, \ldots, f_n\}$$

given a set of *M* items, each item *I* represented as a point in a ***n*-dimensional vector space**

$$I = (w_{f1}, \ldots\ldots w_{fn})$$

w_{fi} is the **weight** of feature *i* in the item

Basic Content-based Recommendations



- ✓ **documents** represented as **vectors**
- ✓ **features** identified through **NLP operations**
- ✓ **features** weigthed using **tf-idf**
- ✓ **cosine measure** for computing similarity between vectors

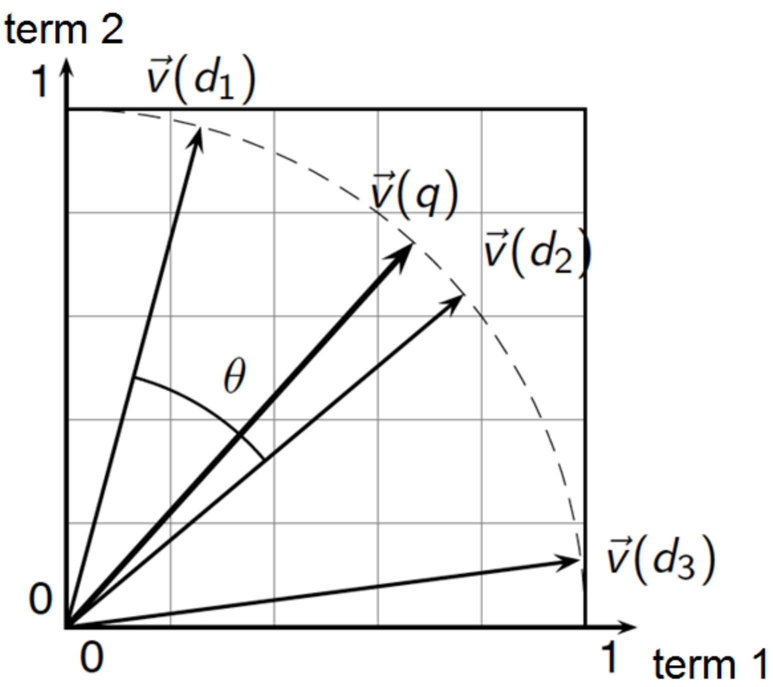
Similarity between vectors

cosine similarity

dot product

unit vectors

$$\cos(\vec{I},\vec{J}) = \frac{\vec{I} \bullet \vec{J}}{\left|\vec{I}\right|\left|\vec{J}\right|} = \frac{\vec{I}}{\left|\vec{I}\right|} \bullet \frac{\vec{J}}{\left|\vec{J}\right|} = \frac{\sum_{i=1}^{|V|} I_i J_i}{\sqrt{\sum_{i=1}^{|V|} I_i^2}\sqrt{\sum_{i=1}^{|V|} J_i^2}}$$



Basic Content-based Recommendations
Drawbacks

- Matrix
- 1999
- American
- Australian
- science
- fiction
- **Hugo**
- …
- world

a portion of Pasquale's
content-based profile

Recommendation:
Notre Dame de Paris,
by Victor Hugo

Why?
Entities as «Hugo Weaving» were not modeled

Basic Content-based Recommendations
Drawbacks

Matrix

1999

American

Australian

science

fiction

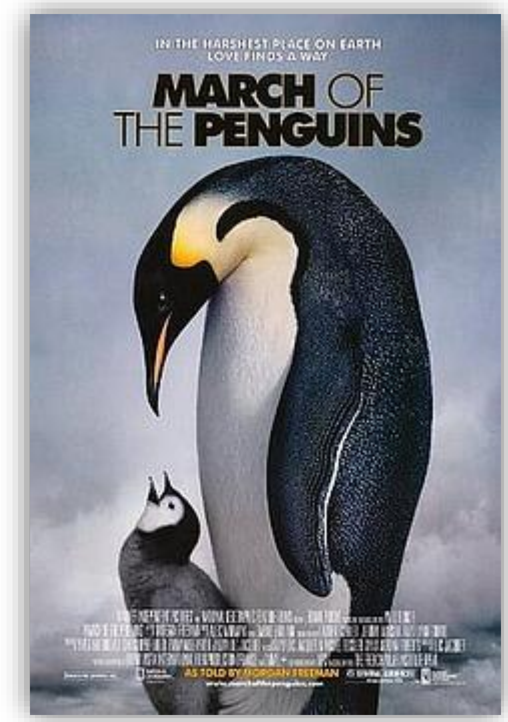
Hugo

…

world

a portion of Pasquale's
content-based profile

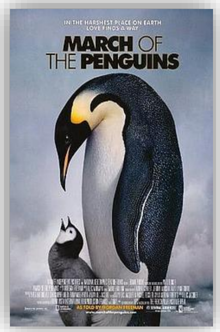
Recommendation:
The March of Penguins



Why?

More complex concepts as «science fiction» were not modeled as single features

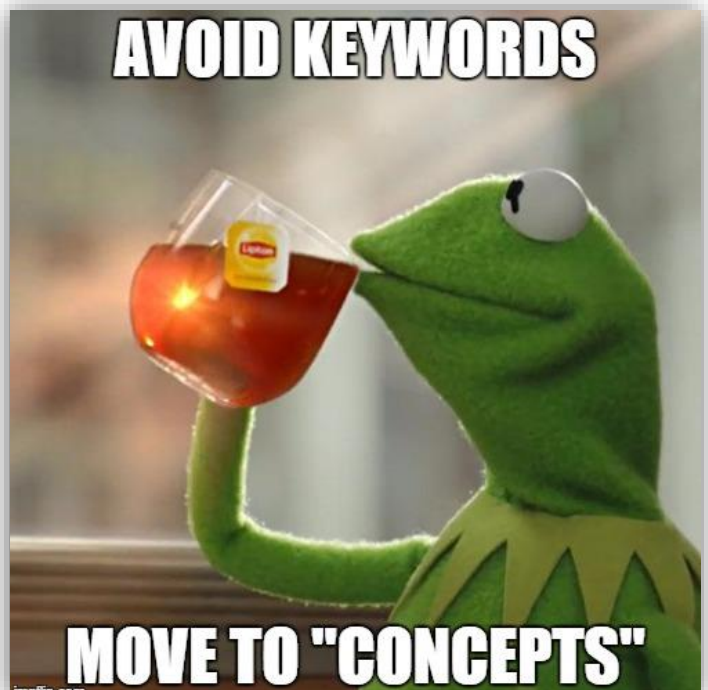
Basic Content-based Recommendations
Vision



Basic Content-based Recommendations
Vision




Bad recommendations



Recap #3

basics of **NLP** and **keyword**-based representation

- **Natural Language Processing techniques necessary** to build a content-based profile
- **basic content-based recommender systems can be easily built** through VSM and TF-IDF
- **keyword-based representation too poor** and can drive to bad modeling of preferences (and bad recommendations)
- **we need to shift from keywords to concepts**

Agenda

Why?

Why do we need **intelligent information access**?
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
Serendipity in Recommender Systems

Semantic representations

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

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

Explicit (Exogenous) Semantics

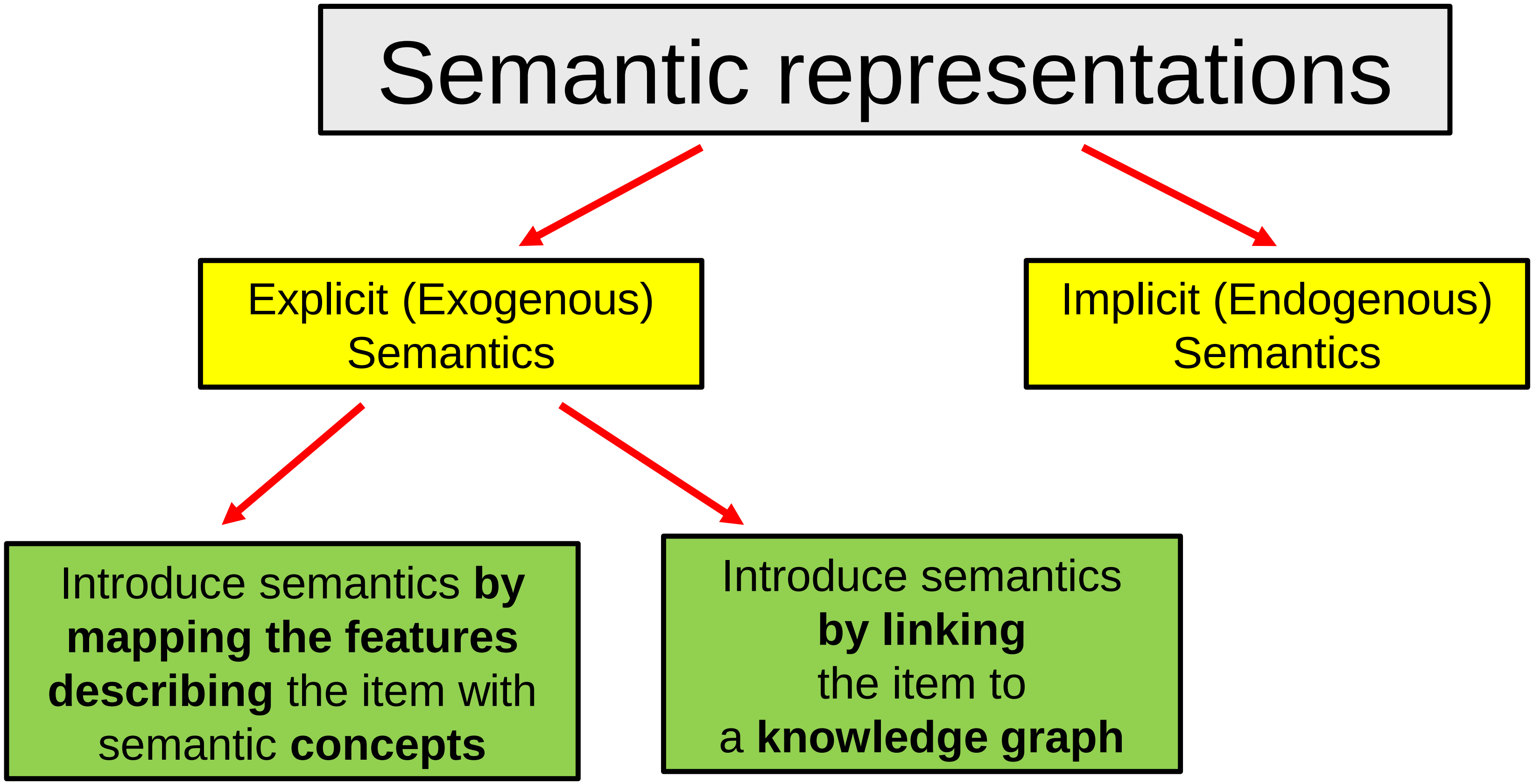
top-down

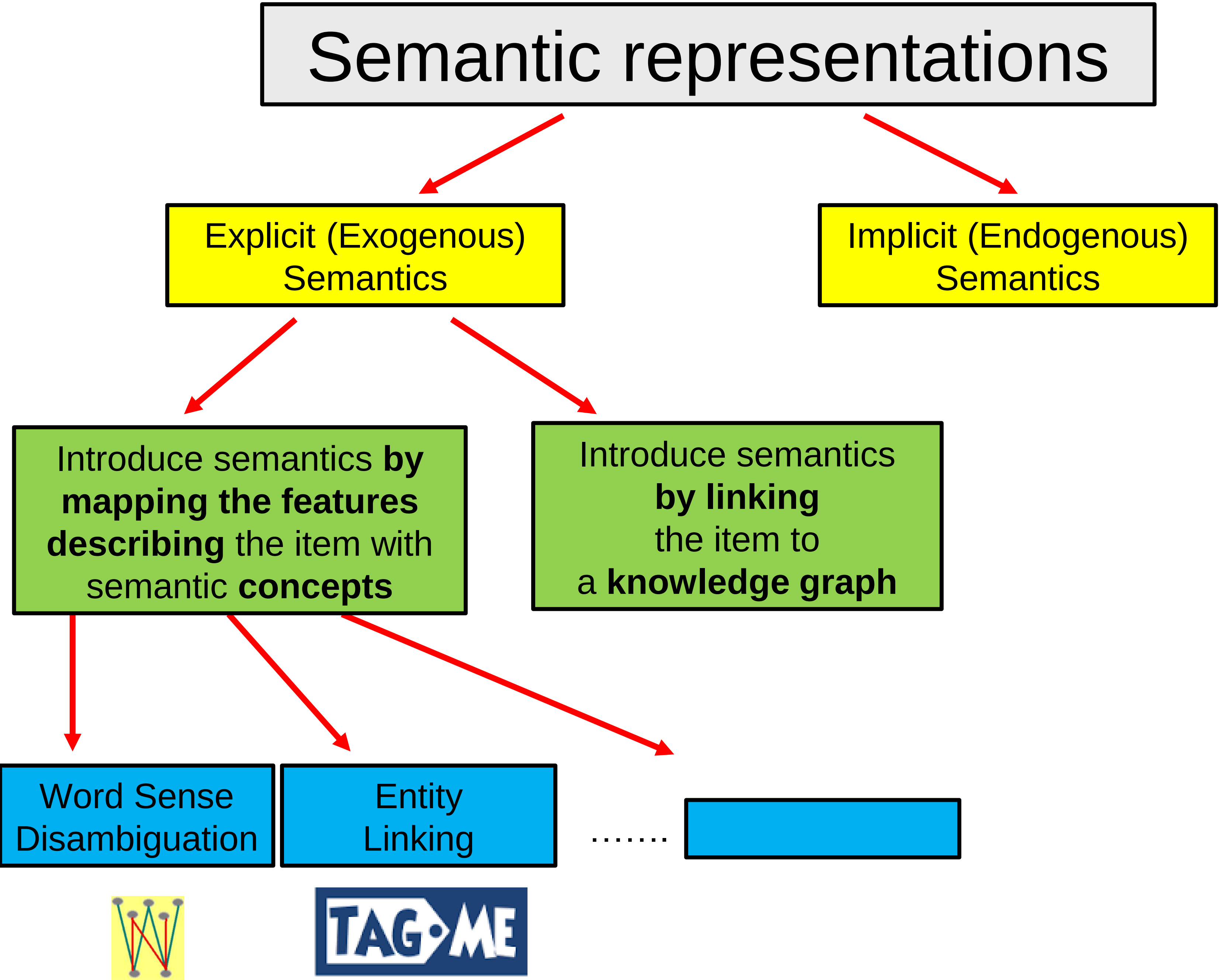
approaches based on the integration of **external knowledge** for representing content. Able to provide the **linguistic, cultural** and **backgroud knowledge** in the **content representation**

Implicit (Endogenous) Semantics

bottom-up

approaches that determine the **meaning** of a word by analyzing the rules of its **usage** in the context of **ordinary and concrete language behavior**


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking** the item to a **knowledge graph**


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking** the item to a **knowledge graph**
Word Sense Disambiguation
Entity Linking
…….
TAG ME

Semantic representations

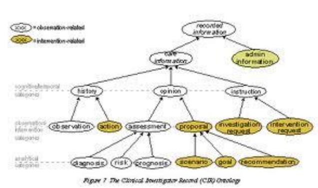
Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking** the item to a **knowledge graph**

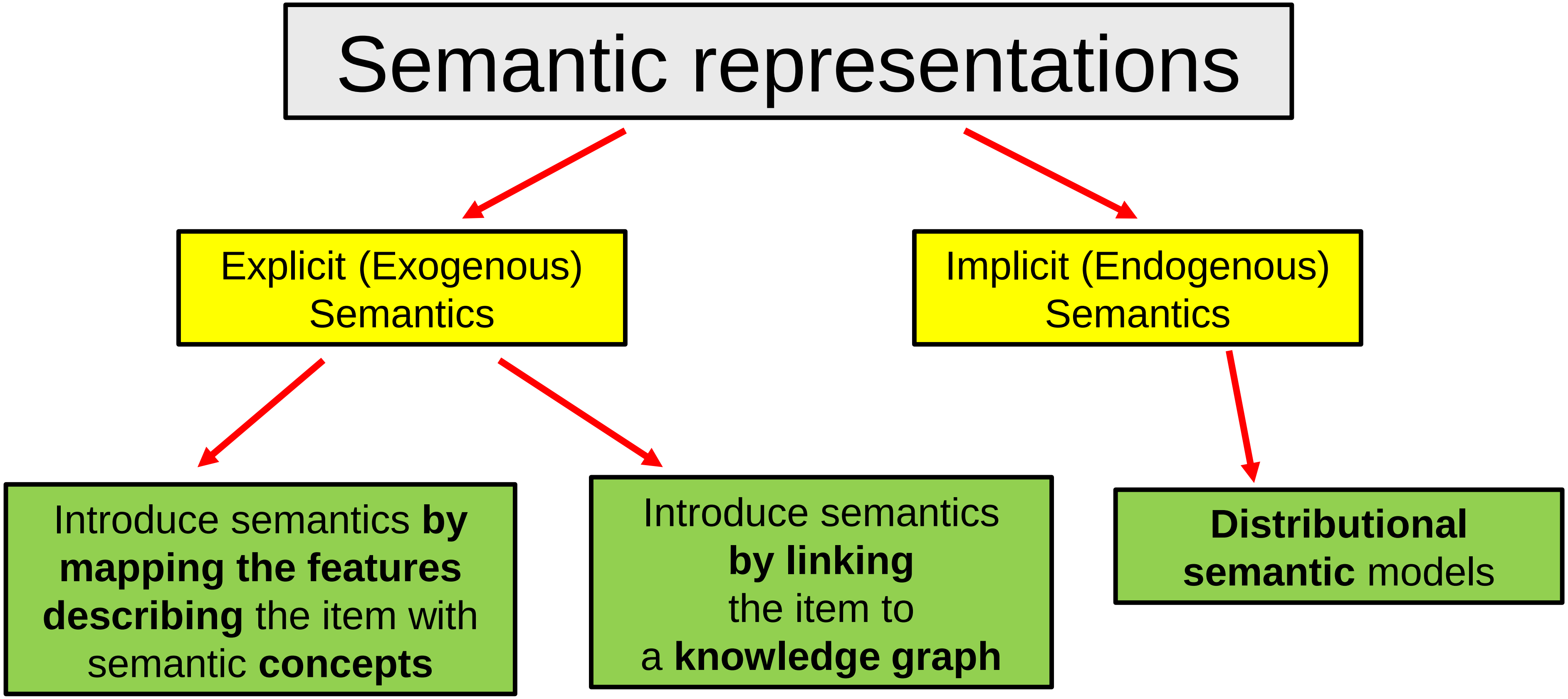
Ontologies

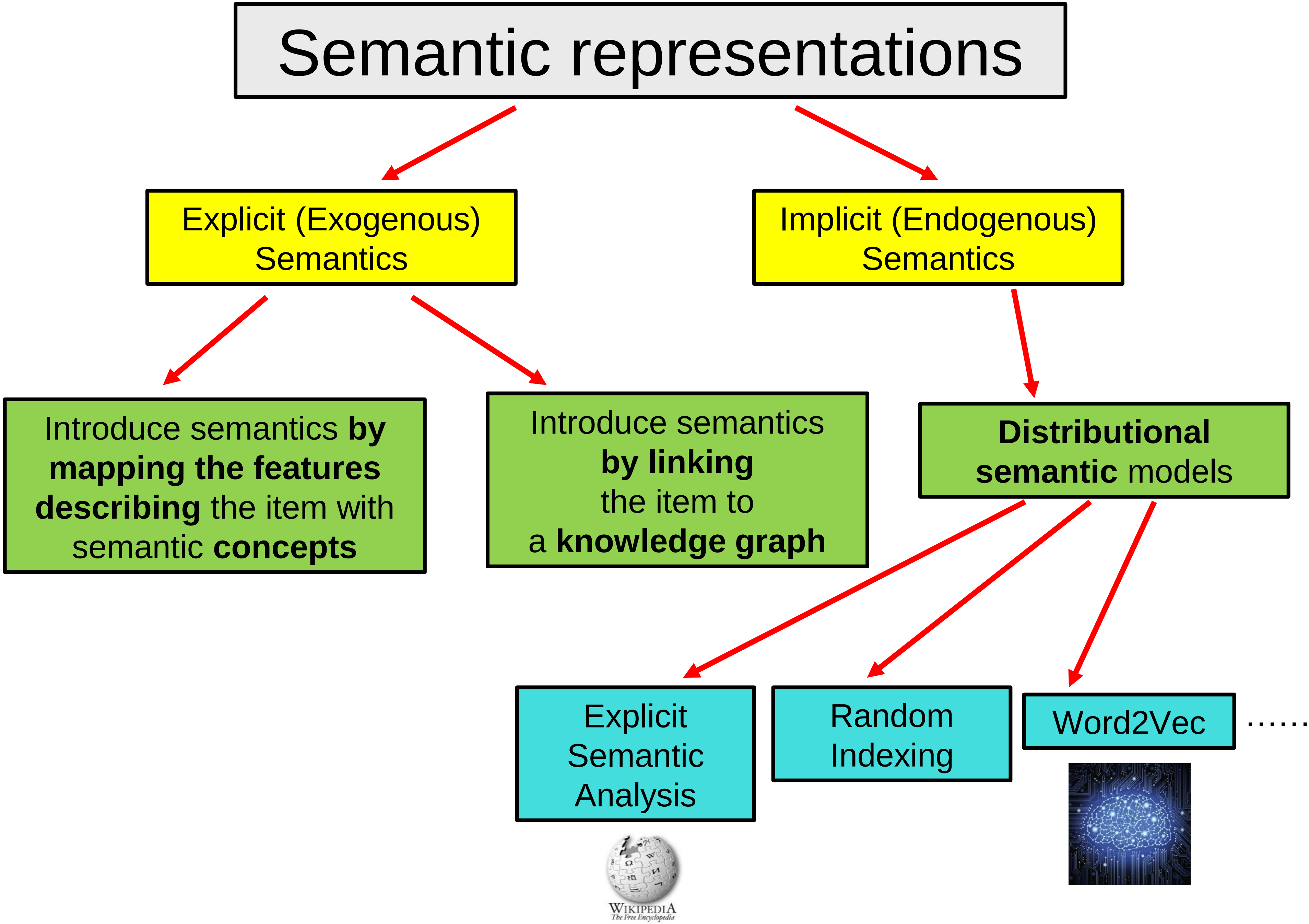


Linked Open Data

DBpedia

....


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking** the item to a **knowledge graph**
Distributional semantic models


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking** the item to a **knowledge graph**
Distributional semantic models
Explicit Semantic Analysis
Random Indexing
Word2Vec
……
WIKIPEDIA
The Free Encyclopedia

Agenda

Why?

Why do we need **intelligent information access**?
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?

Semantics-aware Recommender Systems
Cross-lingual Content-based Recommender Systems
Explanation of Recommendations
Real-time Semantic Analysis of Social Streams

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking** the item to a **knowledge graph**

Word Sense Disambiguation

Entity Linking

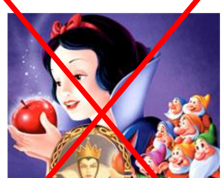
....

Word Sense Disambiguation (WSD)
using linguistic ontologies

WSD selects the proper meaning, i.e. **sense**, for a word in a text by taking into account the **context** in which it occurs

context

Apple Computer iPhone



#12567: computer brand

~~#22999: fruit~~

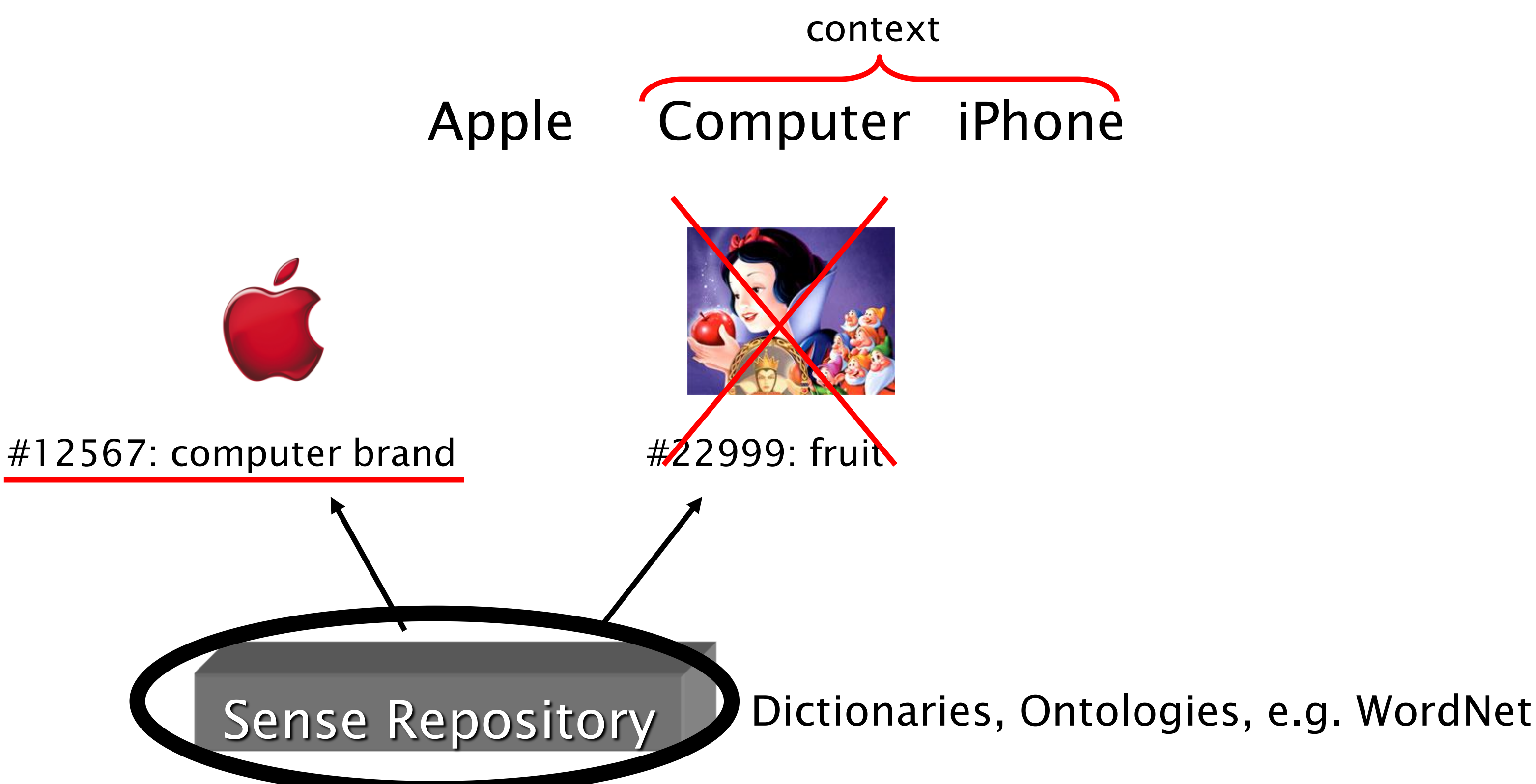
Sense Repository

Dictionaries, Ontologies, e.g. WordNet

Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile: Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. IJCAI 2007: 2856-2861

Word Sense Disambiguation (WSD)
using linguistic ontologies

WSD selects the proper meaning, i.e. **sense**, for a word in a text by taking into account the **context** in which it occurs



Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile: Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. IJCAI 2007: 2856-2861

Sense Repository

WordNet linguistic ontology [*]

https://wordnet.princeton.edu

WordNet groups words into sets of synonyms called **synsets**

It contains **nouns**, **verbs**, **adjectives**, **adverbs**

a lexical database for the English language

cognitive science laboratory | princeton university | 221 nassau st. | princeton, nj 08542

Word Meanings	***Word Forms***					
	F_1	F_2	F_3	…	…	F_n
M_1	V(1,1)	V(2,1)				
M_2		V(2,2)	V(3,2)			
M_3						
$M_{…}$						
M_m						V(m,n)

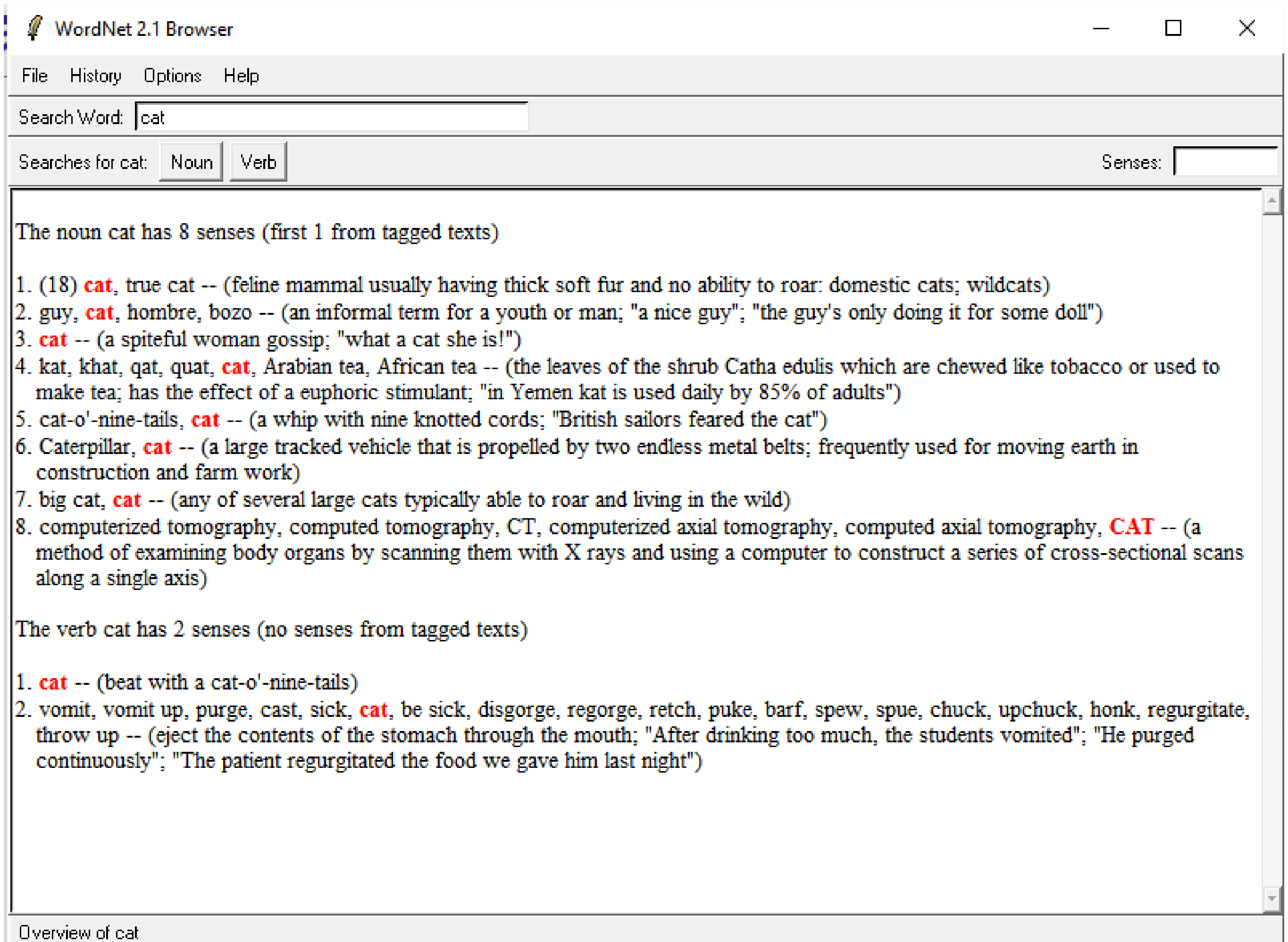
Synonym word forms (synset)

polysemous word: disambiguation needed

[*] *Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.*

Sense Repository
WordNet linguistic ontology

https://wordnet.princeton.edu

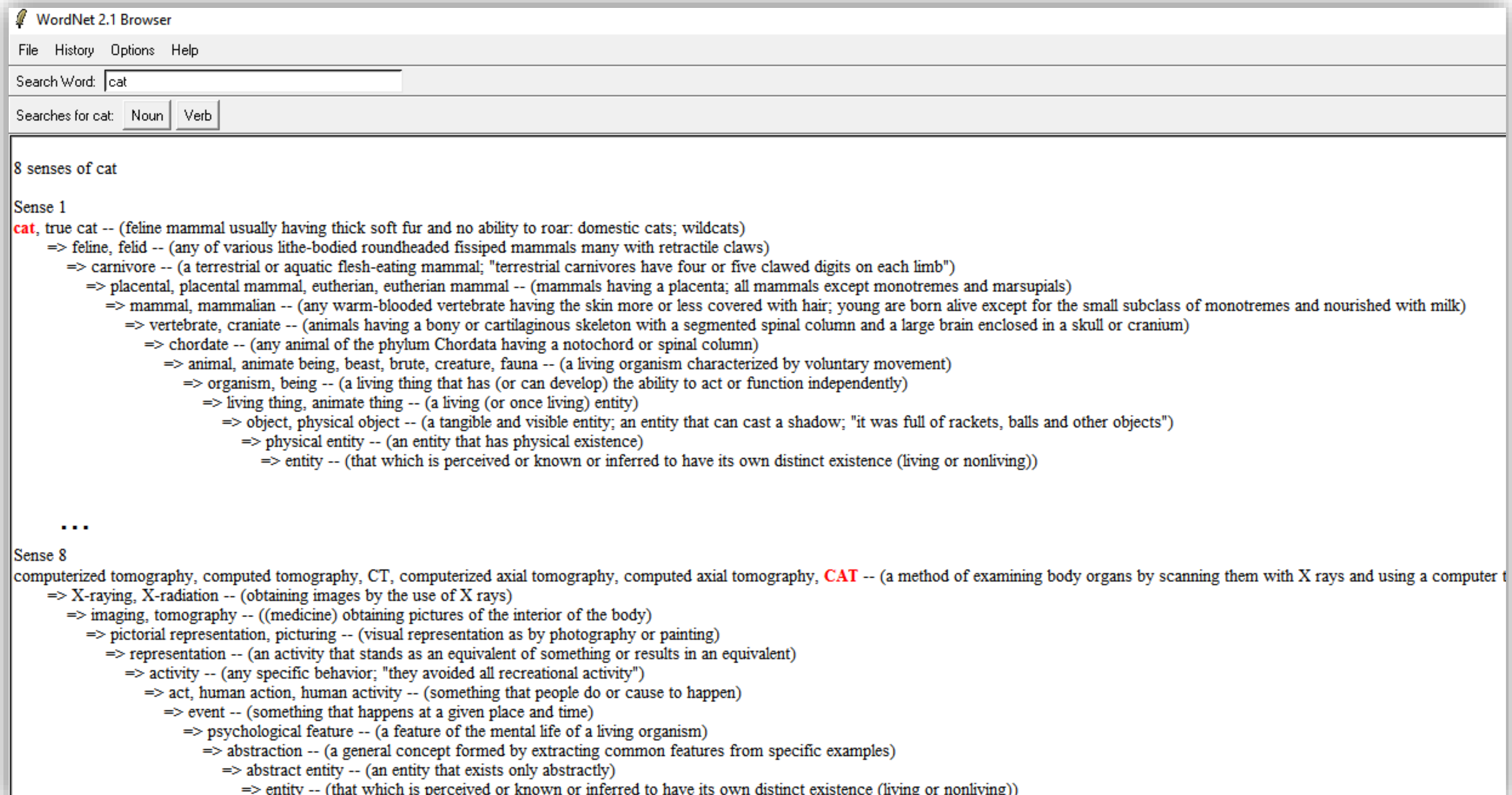


an example of **synset**

Sense Repository

WordNet linguistic ontology

https://wordnet.princeton.edu



WordNet Hierarchies

Word Sense Disambiguation

State of the art: JIGSAW algorithm [*]

Input

- $D = \{w_1, w_2, \ldots, w_h\}$ document

Output

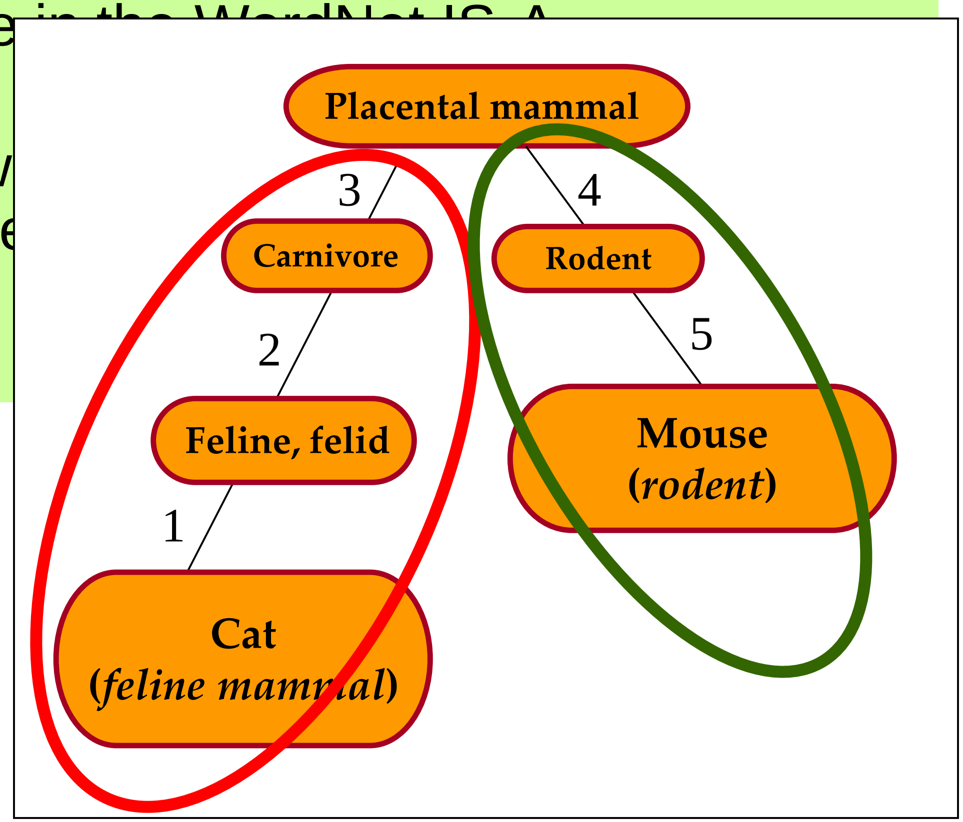
- $X = \{s_1, s_2, \ldots, s_k\}$ $(k \leq h)$
  - Each s_i obtained by disambiguating w_i based on the context of each word
  - Some words not recognized by WordNet
  - Groups of words recognized as a single concept

[*] Basile, P., de Gemmis, M., Gentile, A. L., Lops, P., & Semeraro, G. (2007, June). UNIBA: JIGSAW algorithm for word sense disambiguation. InProceedings of the 4th International Workshop on Semantic Evaluations (pp. 398-401). Association for Computational Linguistics.

JIGSAW WSD algorithm

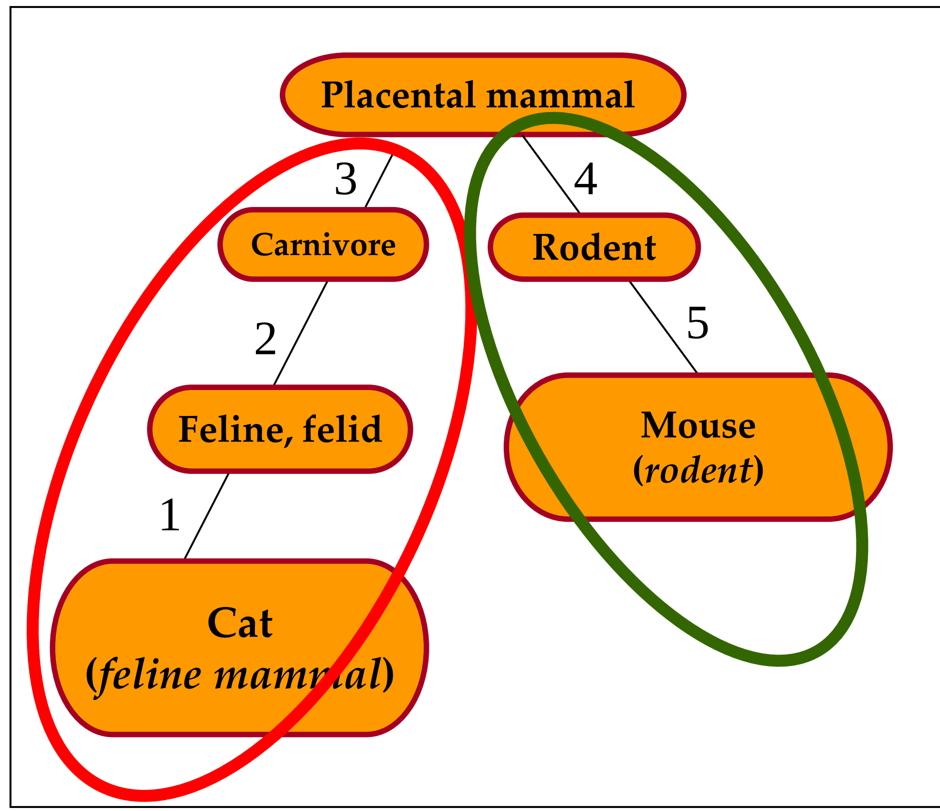
How to use WordNet for WSD?

- **Semantic similarity** between synsets inversely proportional to their distance in the WordNet IS-A hierarchy
- **Path length similarity** betw scores to synsets of a polyse choose the correct sense



Synset semantic similarity

function SINSIM(a, b) ▷ The similarity of the synsets a and b
    $N_p \leftarrow$ the number of nodes in path p from a to b
    $D \leftarrow$ maximum depth of the taxonomy ▷ In WordNet 1.7.1 $D = 16$
    $r \leftarrow -log(N_p/2D)$
    return r
end function



SINSIM(cat,mouse) =
-log(5/32)=0.806

Leacock-Chodorow similarity

JIGSAW WSD algorithm

"The white cat is hunting the mouse"

w = cat

C = {mouse}

W_{cat}={02037721,00847815}

T={02244530,03651364}

cat

02037721: feline mammal…

00847815: computerized axial tomography…

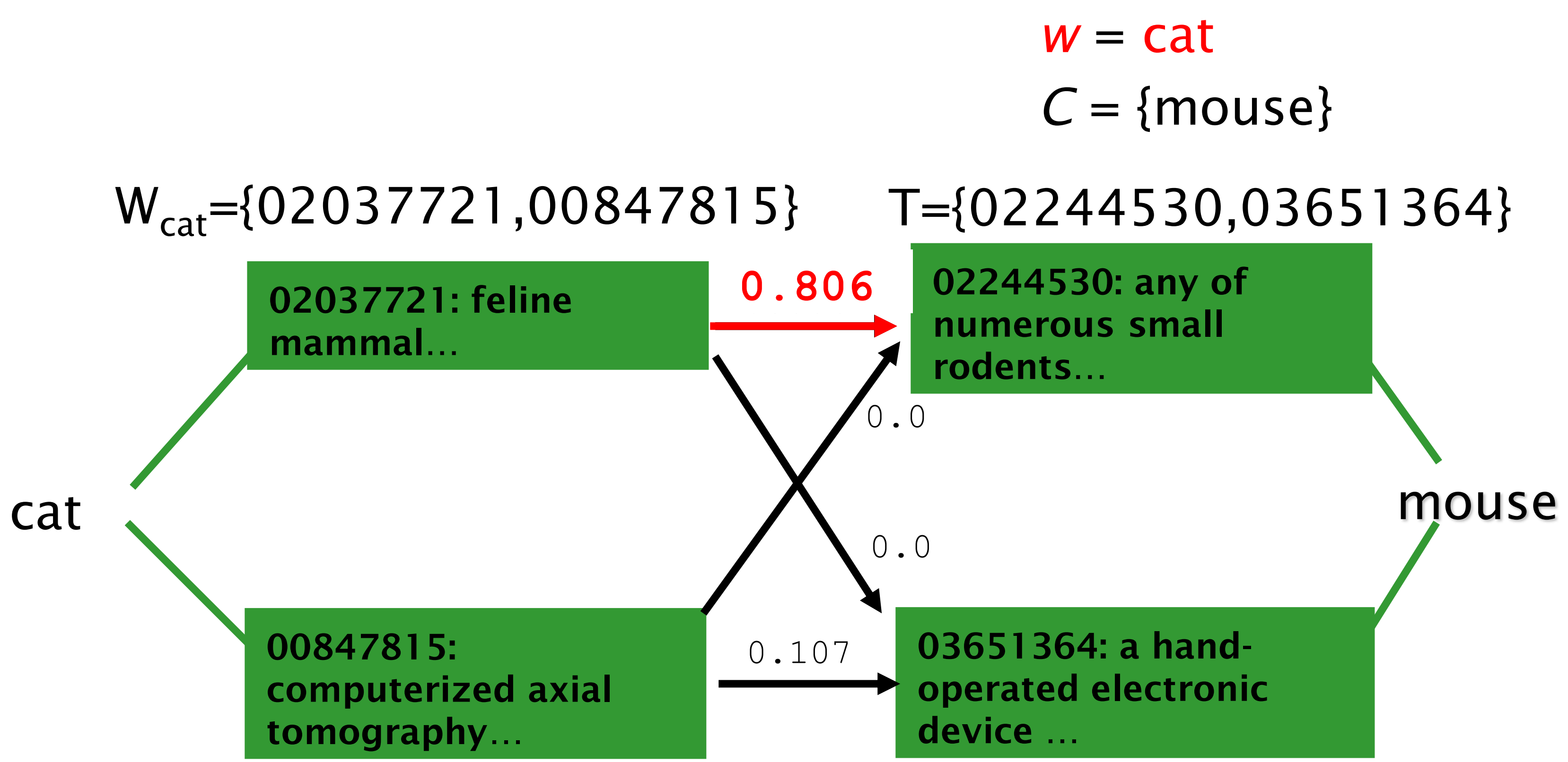
mouse

02244530: any of numerous small rodents…

03651364: a hand-operated electronic device …

JIGSAW WSD algorithm

"The white cat is hunting the mouse"

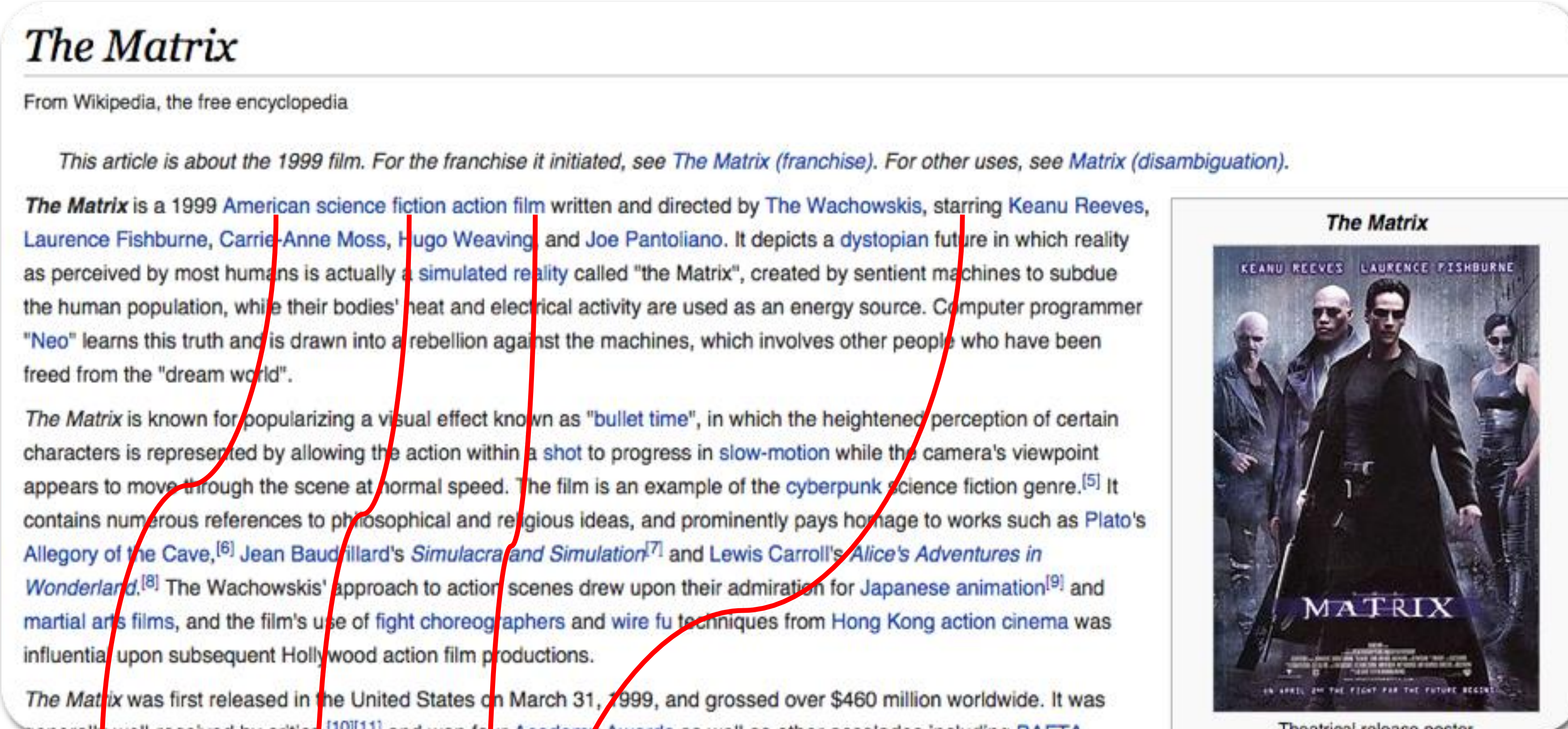


through WSD can we obtain a

semantics-aware representation

of textual content

Synset-based representation


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

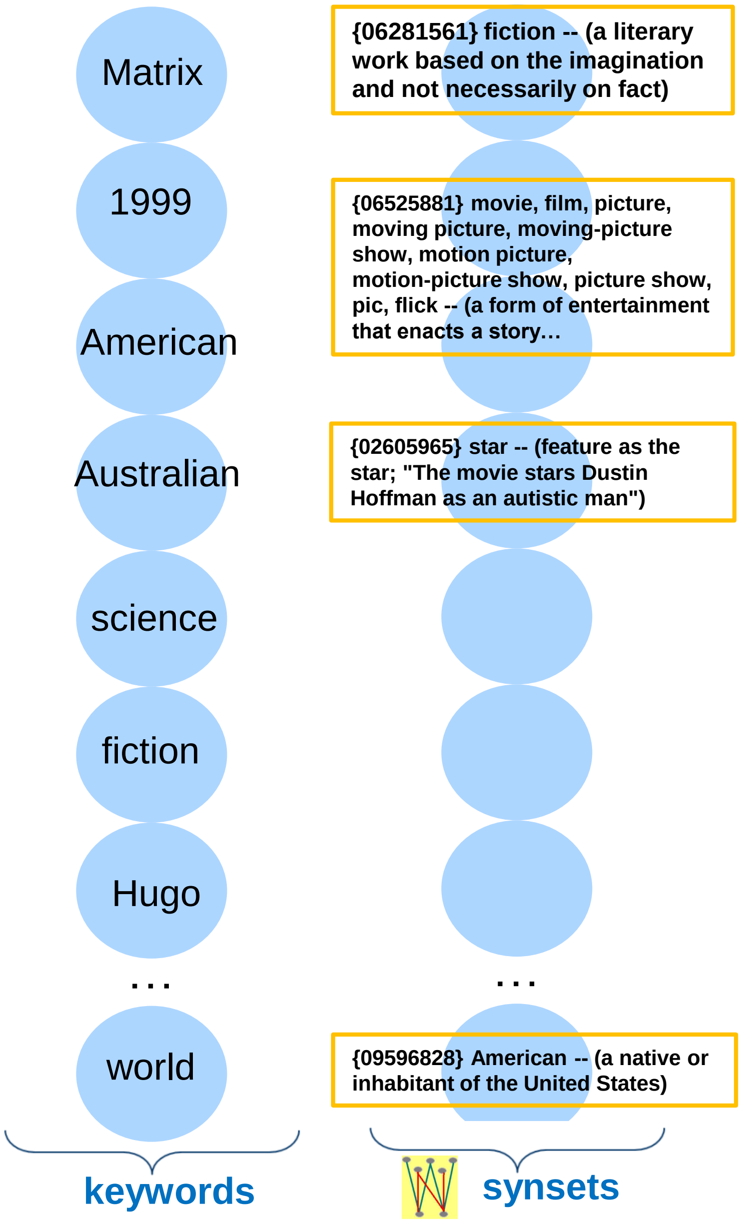
{09596828} American -- (a native or inhabitant of the United States)

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

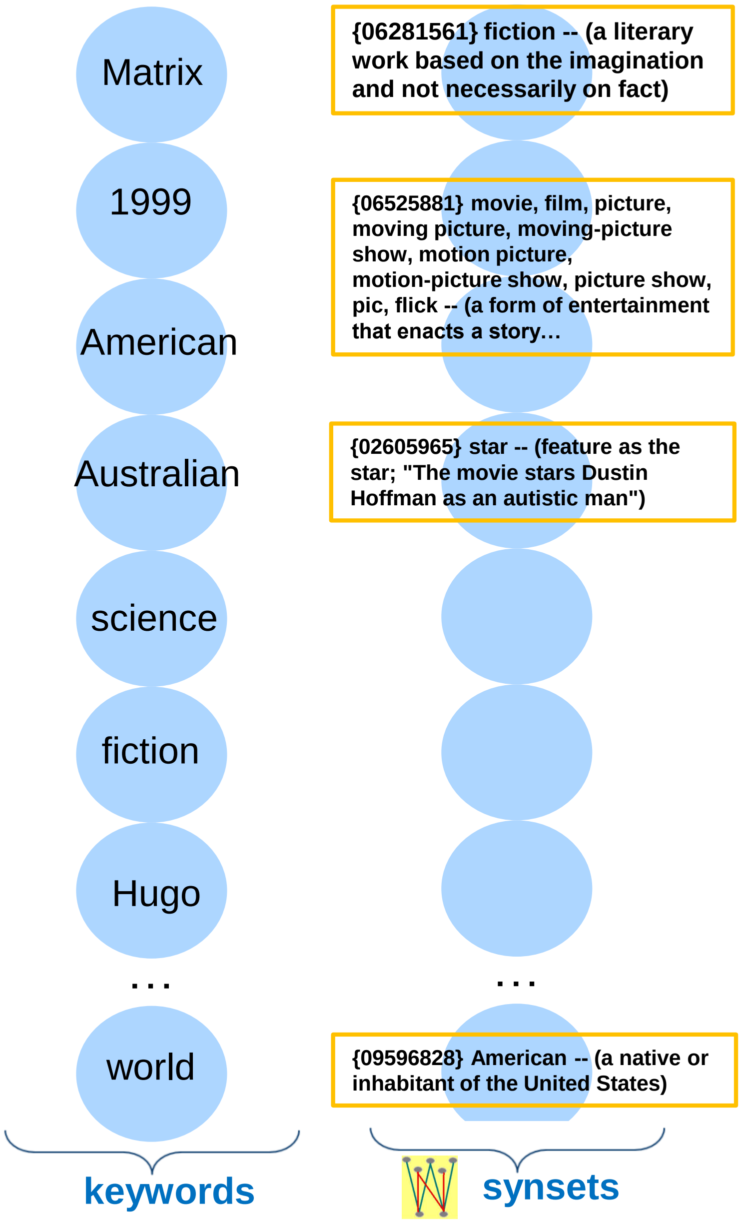
The Matrix representation



through WSD we process the textual description of the item and we obtain a **semantics-aware representation** of the item as output

keyword-based features **replaced with the concepts** (in this case WordNet synsets) they refer to

The Matrix representation



Word Sense Disambiguation

recap

polysemy and **synonymy** effectively handled

classical NLP techniques helpful to remove further noise (e.g. stopwords)

potentially language-independent (later)

entities (e.g. Hugo Weaving) still not recognized

Semantic representations

- Explicit (Exogenous) Semantics
  - Introduce semantics **by mapping the features describing** the item with semantic **concepts**
    - Word Sense Disambiguation
    - Entity Linking
    - ....
  - Introduce semantics **by linking** the item to a **knowledge graph**
- Implicit (Endogenous) Semantics

TAG ME

Entity Linking Algorithms

- **Basic Idea**
  - **Input:** free text
    - *e.g. Wikipedia abstract*
  - **Output:** identification of the **entities** mentioned in the text.

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

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

OPEN CALAIS
THOMSON REUTERS

TAG ME

Babelfy

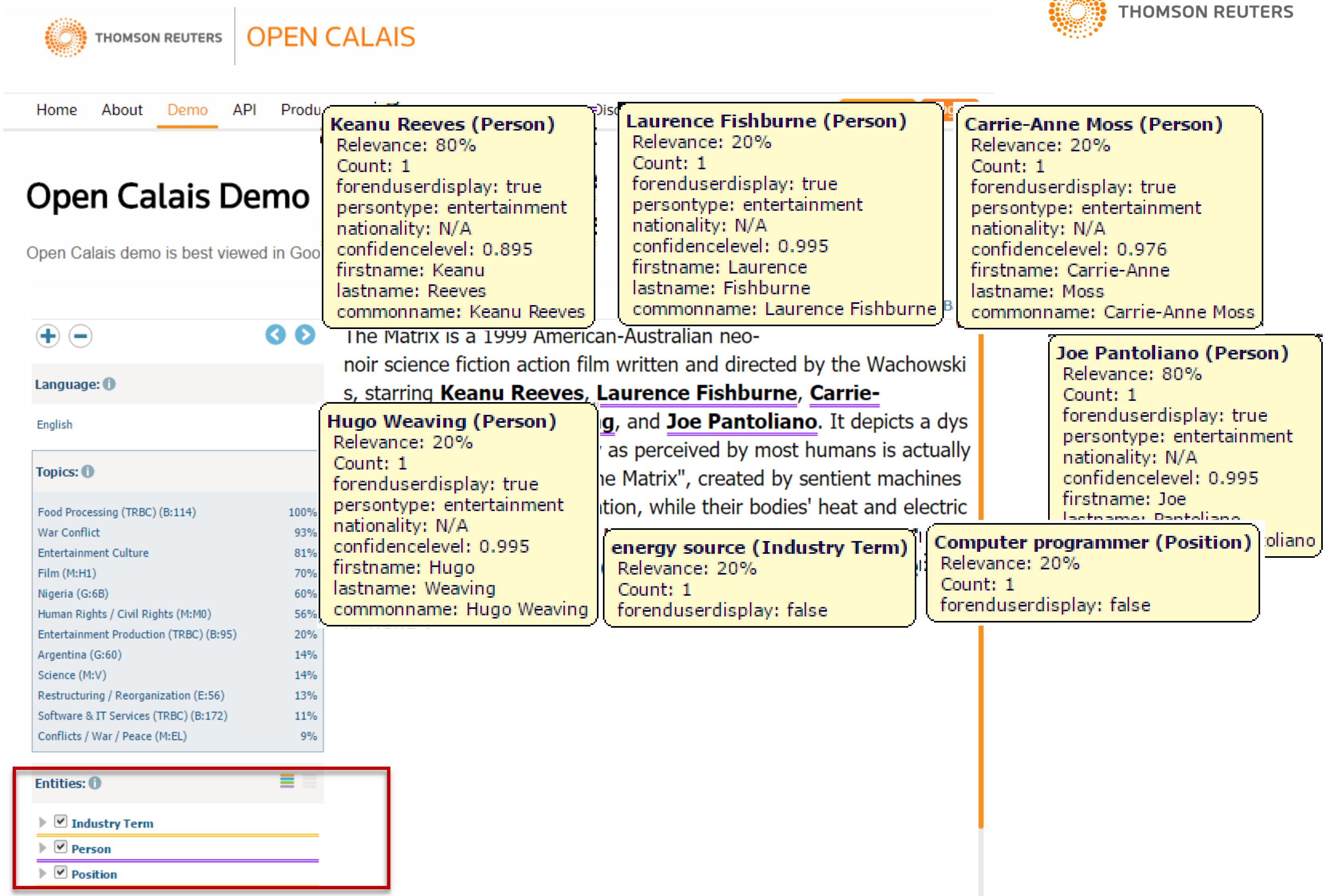
… and many more

Entity Linking Algorithms
OpenCalais

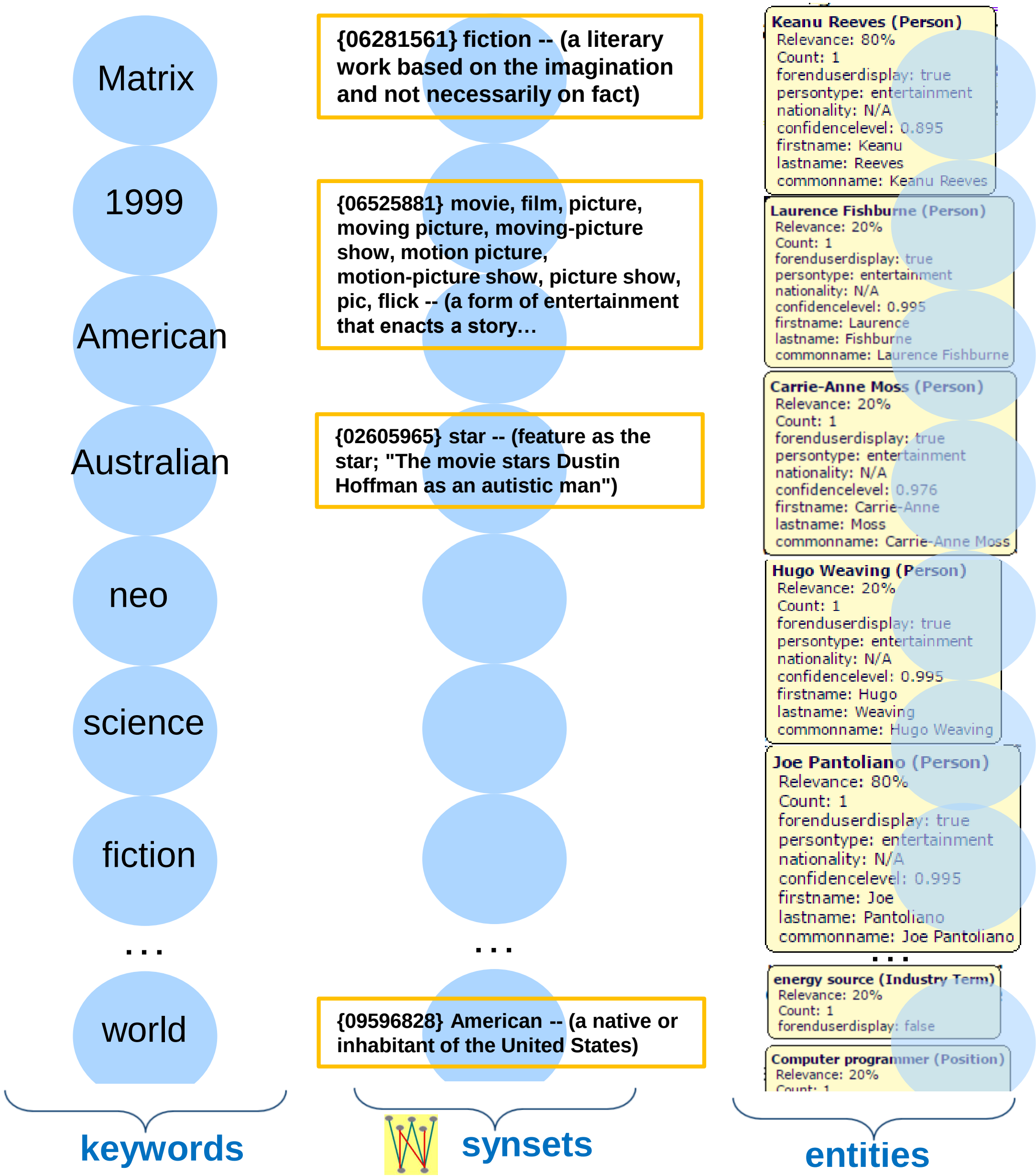
OPEN CALAIS

THOMSON REUTERS

http://www.opencalais.com/opencalais-api/



The Matrix representation



The Matrix representation

keywords	synsets
Matrix	{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)
1999	
American	{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…
Australian	{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")
neo	
science	
fiction	
…	…
world	{09596828} American -- (a native or inhabitant of the United States)

entities

Keanu Reeves (Person)
Relevance: 80%
Count: 1
forenduserdisplay: true
persontype: entertainment
nationality: N/A
confidencelevel: 0.895
firstname: Keanu
lastname: Reeves
commonname: Keanu Reeves

Laurence Fishburne (Person)
Relevance: 20%
Count: 1
forenduserdisplay: true
persontype: entertainment
nationality: N/A
confidencelevel: 0.995
firstname: Laurence
lastname: Fishburne
commonname: Laurence Fishburne

Carrie-Anne Moss (Person)
Relevance: 20%
Count: 1
forenduserdisplay: true
persontype: entertainment
nationality: N/A
confidencelevel: 0.976
firstname: Carrie-Anne
lastname: Moss
commonname: Carrie-Anne Moss

Hugo Weaving (Person)
Relevance: 20%
Count: 1
forenduserdisplay: true
persontype: entertainment
nationality: N/A
confidencelevel: 0.995
firstname: Hugo
lastname: Weaving
commonname: Hugo Weaving

Joe Pantoliano (Person)
Relevance: 80%
Count: 1
forenduserdisplay: true
persontype: entertainment
nationality: N/A
confidencelevel: 0.995
firstname: Joe
lastname: Pantoliano
commonname: Joe Pantoliano

…

energy source (Industry Term)
Relevance: 20%
Count: 1
forenduserdisplay: false

Computer programmer (Position)
Relevance: 20%
Count: 1

OPEN CALAIS

THOMSON REUTERS

entities are correctly recognized and modeled

partially multilingual
(entities are inherently multilingual, but other concepts aren't)

common sense and abstract concepts now ignored.

Entity Linking Algorithms

Tag.me

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

TAG-ME Output

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

each entity identified in the content can be a feature of a **semantics-aware content representation based on entity linking**

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

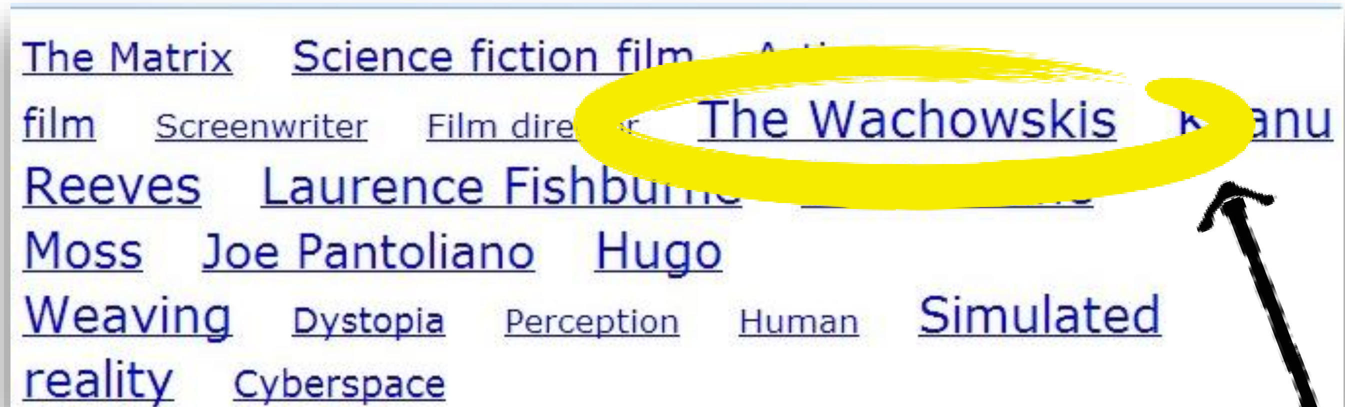
Advantage #1: several common sense concepts are now identified

Entity Linking Algorithms

Tag.me

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

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

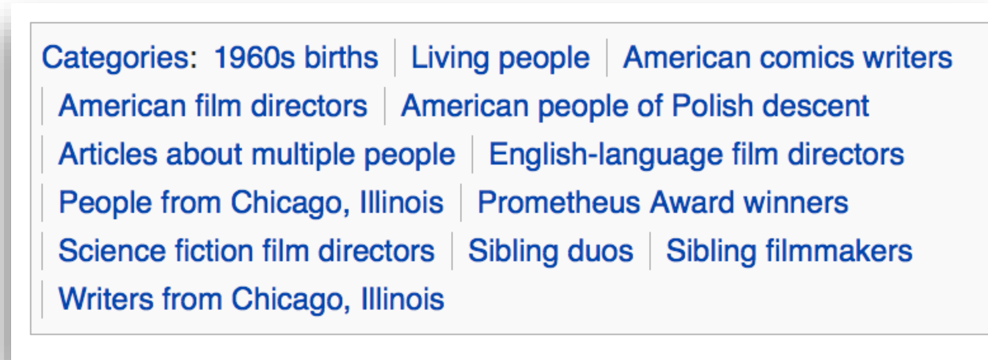
From Wikipedia, the free encyclopedia

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*.



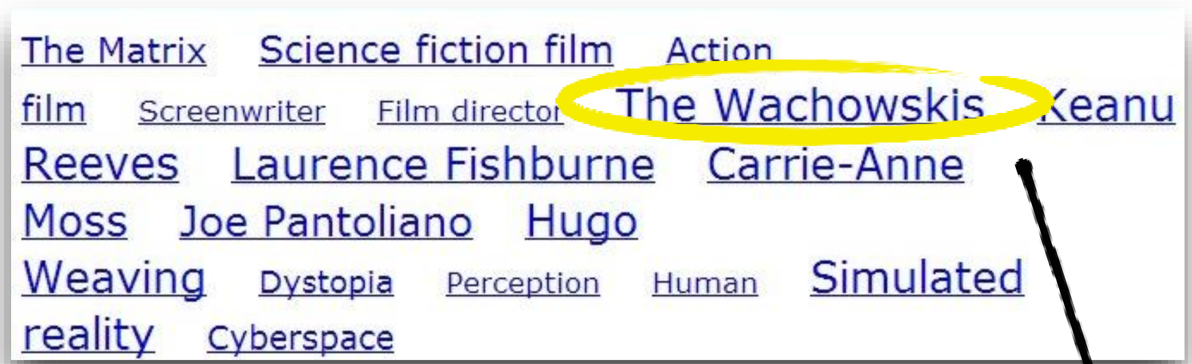
Categories: 1960s births | Living people | American comics writers | American film directors | American people of Polish descent | Articles about multiple people | English-language film directors | People from Chicago, Illinois | Prometheus Award winners | Science fiction film directors | Sibling duos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

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

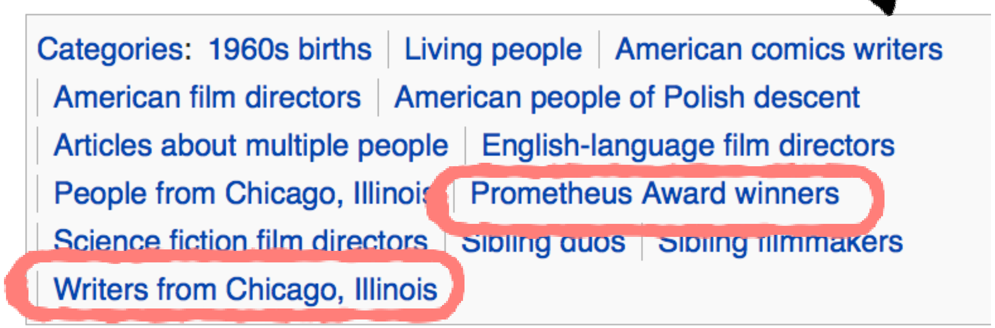
Entity Linking Algorithms

Tag.me + Wikipedia Categories

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



features = entities + wikipedia categories



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

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

...

world

keywords

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

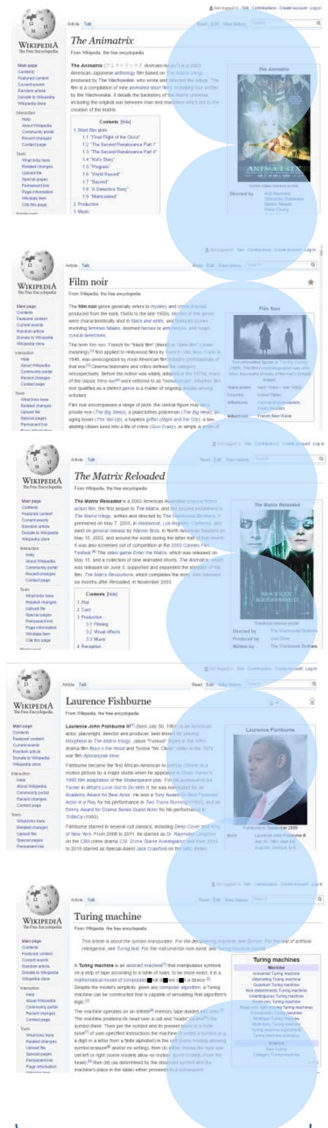
{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story...

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

...

{09596828} American -- (a native or inhabitant of the United States)

synsets



Wikipedia pages

The Matrix representation

Matrix
1999
American
Australian
neo
science
fiction
…
world

keywords

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

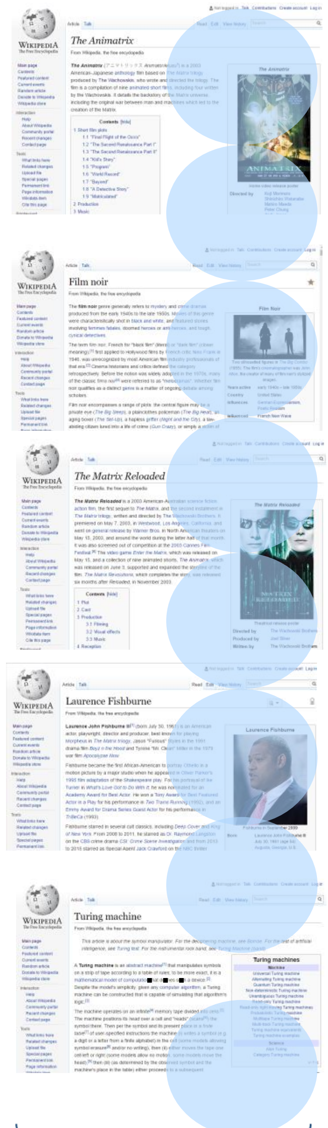
{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

…

{09596828} American -- (a native or inhabitant of the United States)

synsets



Wikipedia pages

TAG•ME

entities recognized and modeled (as in OpenCalais)

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

terms without a Wikipedia mapping are ignored

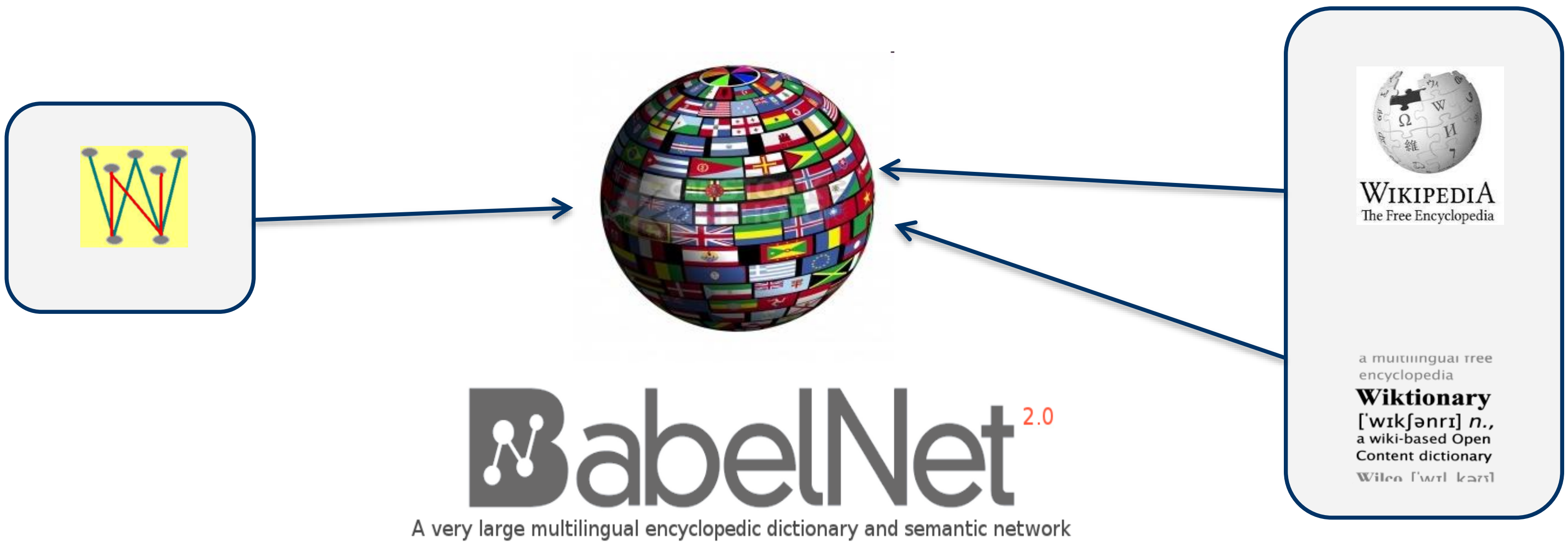
Entity Linking Algorithms

Babelfy

http://babelfy.org/

traditional resources

collaborative resources



- 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

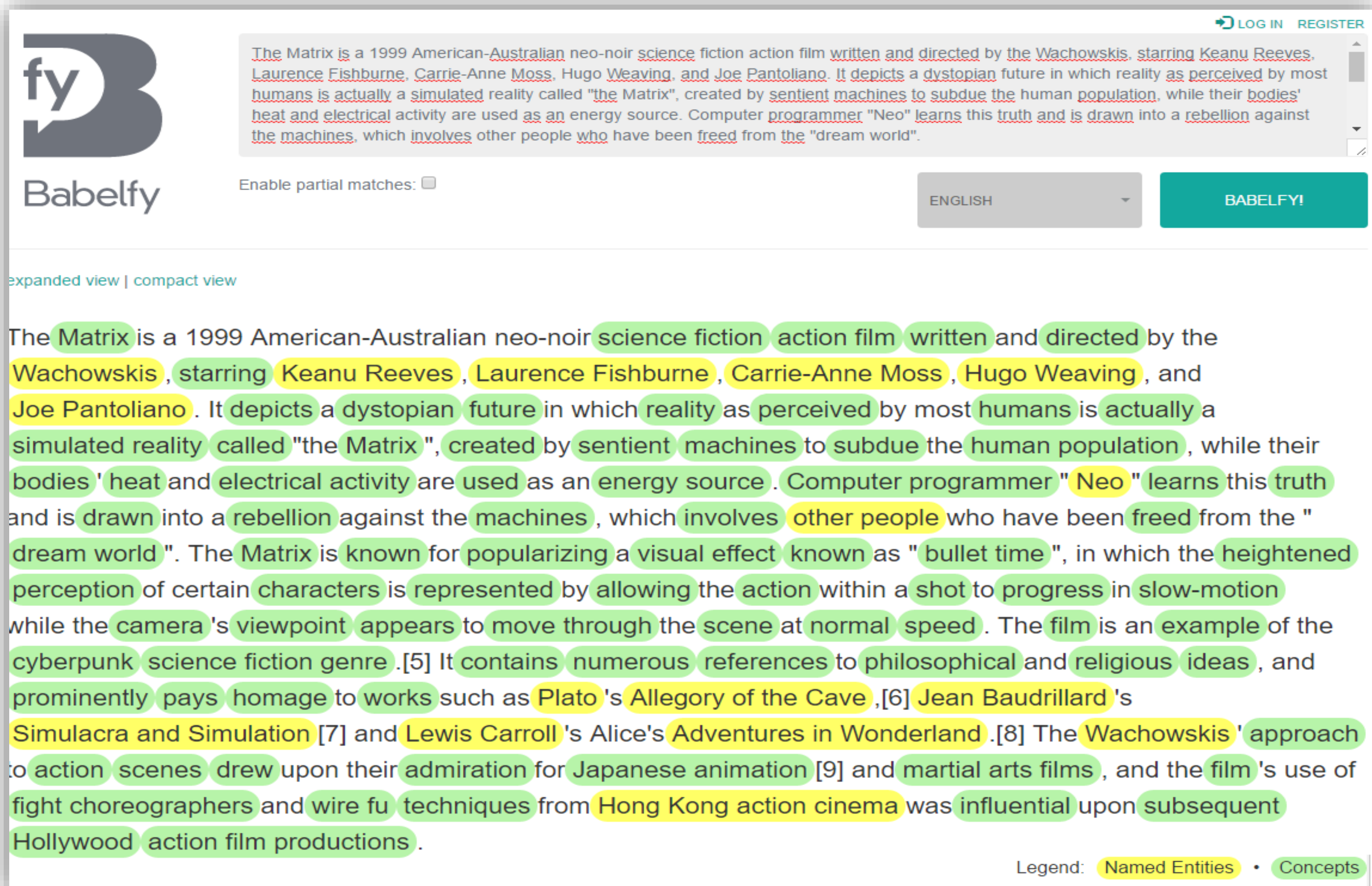
BabelNet 3.6: General statistics

Number of languages:	271
Total number of Babel synsets:	13,801,844
Total number of Babel senses:	745,856,326
Total number of concepts:	6,066,396
Total number of Named Entities:	7,735,448
Total number of lexico-semantic relations:	380,239,084
Total number of glosses (textual definitions):	40,705,588
Total number of images:	10,767,833
Total number of Babel synsets with at least one domain:	1,558,806
Total number of compounds:	743,296
Total number of other forms:	6,393,568
Total number of Babel synsets with at least one picture:	2,948,668
Total number of RDF triples:	1,971,744,856

Entity Linking Algorithms

Babelfy

http://babelfy.org/

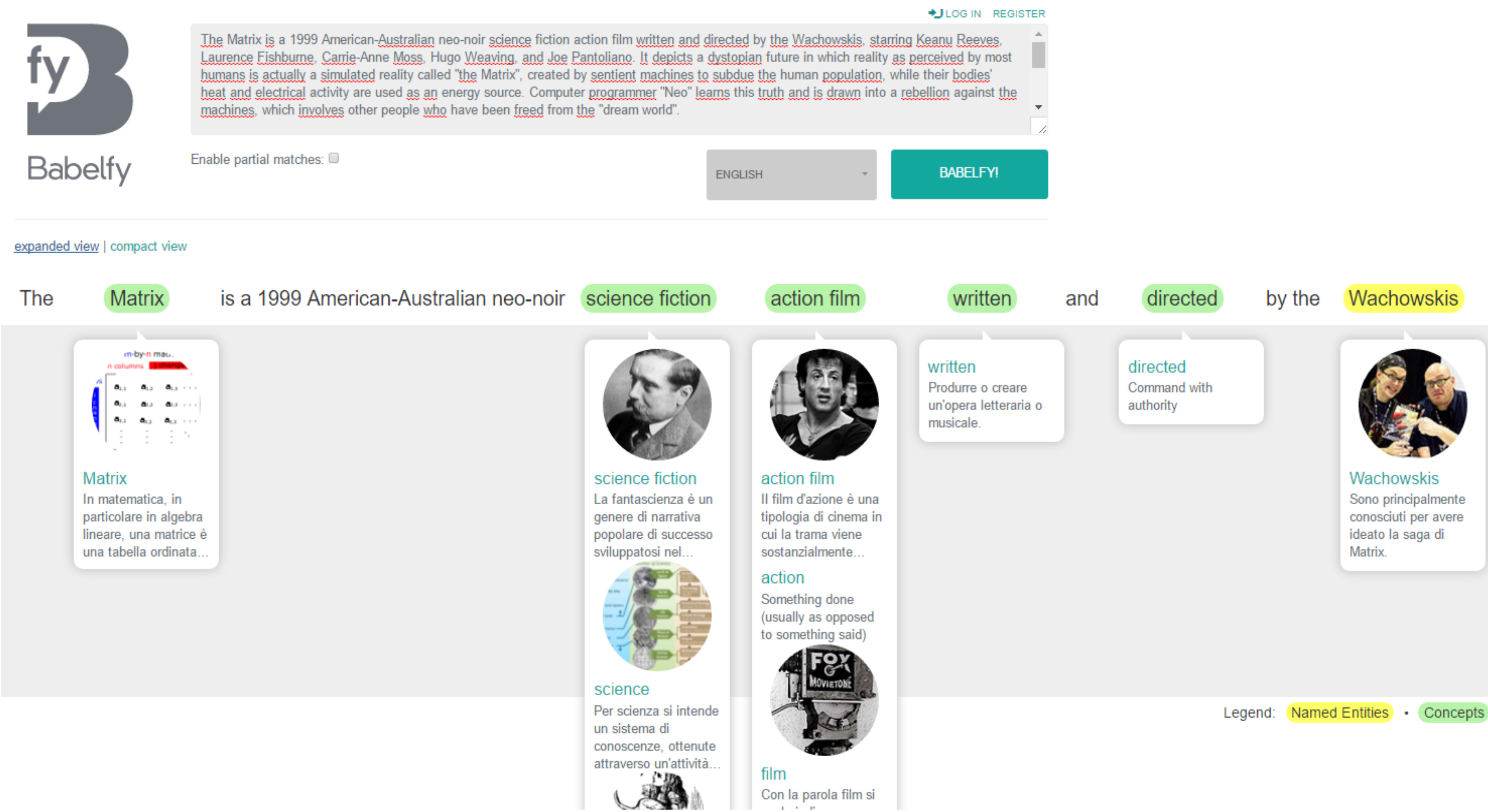


we have both **Named Entities** and **Concepts**!

Entity Linking Algorithms

Babelfy

http://babelfy.org/



Entity Linking Algorithms

Babelfy

http://babelfy.org/

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

...

world

keywords

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

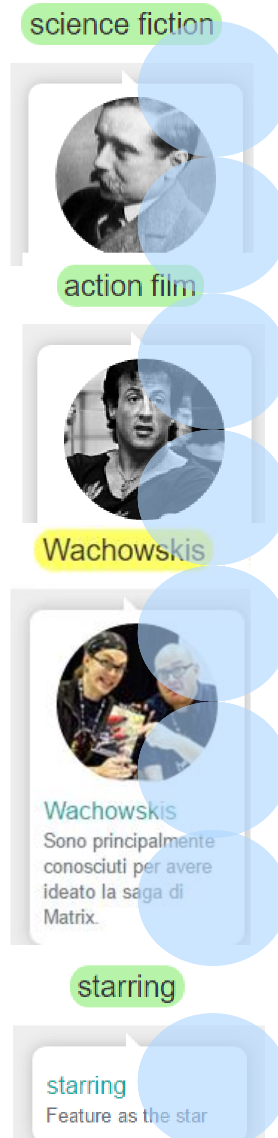
{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

...

{09596828} American -- (a native or inhabitant of the United States)

synsets



Babel synsets

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

...

world

keywords

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

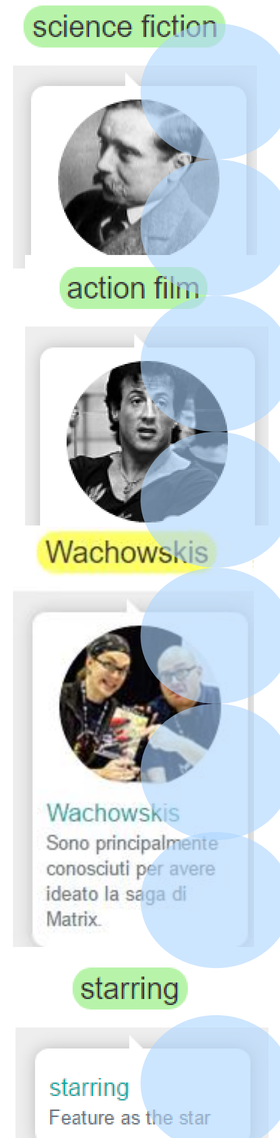
{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

...

{09596828} American -- (a native or inhabitant of the United States)

synsets



Babel synsets

BabelNet 2.5

A very large multilingual encyclopedic dictionary and semantic network

entities recognized and modeled (as in OpenCalais and Tag.me)

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

includes linguistic knowledge and is able to disambiguate terms

also multilingual!

Recap #4

encoding **exogenous semantics** by processing textual descriptions

- **«Exogenous» techniques use external knowledge sources to inject semantics**
- **Word Sense Disambiguation** algorithms process the textual description and replace keywords with semantic concepts (as synsets)
- **Entity Linking algorithms** focus on the identification of the entities. Some recent approaches also able to identify common sense terms
- **Combination of both approaches is potentially the best strategy**

Results in a cultural heritage scenario

CHAT

Cultural ***H***eritage fruition & e-learning applications of new ***A***dvanced (multimodal) ***T***echnologies

ENGINEERING

Dipartimento di Informatica
Università degli Studi di Bari

In the context of cultural heritage personalization, does the integration of UGC and textual description of artwork collections cause an increase of the prediction accuracy in the process of recommending artifacts to users?

M. de Gemmis, P. Lops, G. Semeraro, and P. Basile. **Integrating Tags in a Semantic Content-based Recommender**. *In RecSys '08, Proceed. of the 2nd ACM Conference on Recommender Systems, pages 163–170, October 23-25, 2008, Lausanne, Switzerland, ACM, 2008.*

Results in a cultural heritage scenario

27) Caravaggio - Deposition from the Cross

Textual description of items (static content)

Descrizione dell'opera

The Deposition, considered one of Caravaggio's greatest masterpieces, was commissioned by Girolamo Vittrice for his famil chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris i execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII's Pinacoteca. Caravaggio did not real portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in t tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is t stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arriv in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the us of colour and light, and was certainly the most important personage of the "realist" trend of seventeenth century painting.

Social Tags

Social Tags (from other users): caravaggio, deposition, christ, cross, suffering, religion

Inserisci il tuo voto e dei tag descrittivi (separati da una VIRGOLA, senza spazi)

5-point rating scale

1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Personal Tags

passion

Inserisci i voti e prosegui

Results in a cultural heritage scenario

27) Caravaggio - Deposition from the Cross

Descrizione dell'opera

The Deposition, considered one of Caravaggio's greatest masterpieces, was commissioned by Girolamo Vittrice for his famil
chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris i
execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII's Pinacoteca. Caravaggio did not real
portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in t
tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is t
stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus
and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arriv
in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the us
of colour and light, and was certainly the most important personage of the "realist" trend of seventeenth century painting.

USER PROFILE

caravaggio, deposition,

cross, christ, rome, …

- - - - - - - - - - - - - - - - - -

passion

- - - - - - - - - - - - - - - - - -

caravaggio, deposition,

christ, cross, suffering,

religion, ..

Static Content

Personal Tags

Social Tags

collaborative part of
the user profile

Results in a cultural heritage scenario

- Artwork representation
  - Artist
  - Title
  - Description
  - Tags
- change of text representation from vectors of words (BOW) into vectors of WordNet synsets (BOS)
  - From tags to semantic tags
- supervised Learning
  - Bayesian Classifier learned from artworks labeled with user ratings and tags

Results in a cultural heritage scenario

	Type of Content	Precision*	Recall*	F1*
Content-based Profiles	EXP#1: Static Content	75.86	94.27	84.07
Tag-based Profiles	EXP#2: Personal Tags	75.96	92.65	83.48
	EXP#3: Social Tags	75.59	90.50	82.37
Augmented Profiles	EXP#4: Static Content + Personal Tags	78.04	93.60	**85.11**
	EXP#5: Static Content + Social Tags	78.01	93.19	**84.93**

* Results averaged over the 30 study subjects

Overall accuracy F1 ≈ 85%

Recent Developments of Content-Based RecSys

Distributional Semantics

Fedelucio Narducci

Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

SWAP researchgroup

Agenda

Why?

Why do we need **intelligent information access**?
Why do we need **content**?
Why do we need **semantics**?

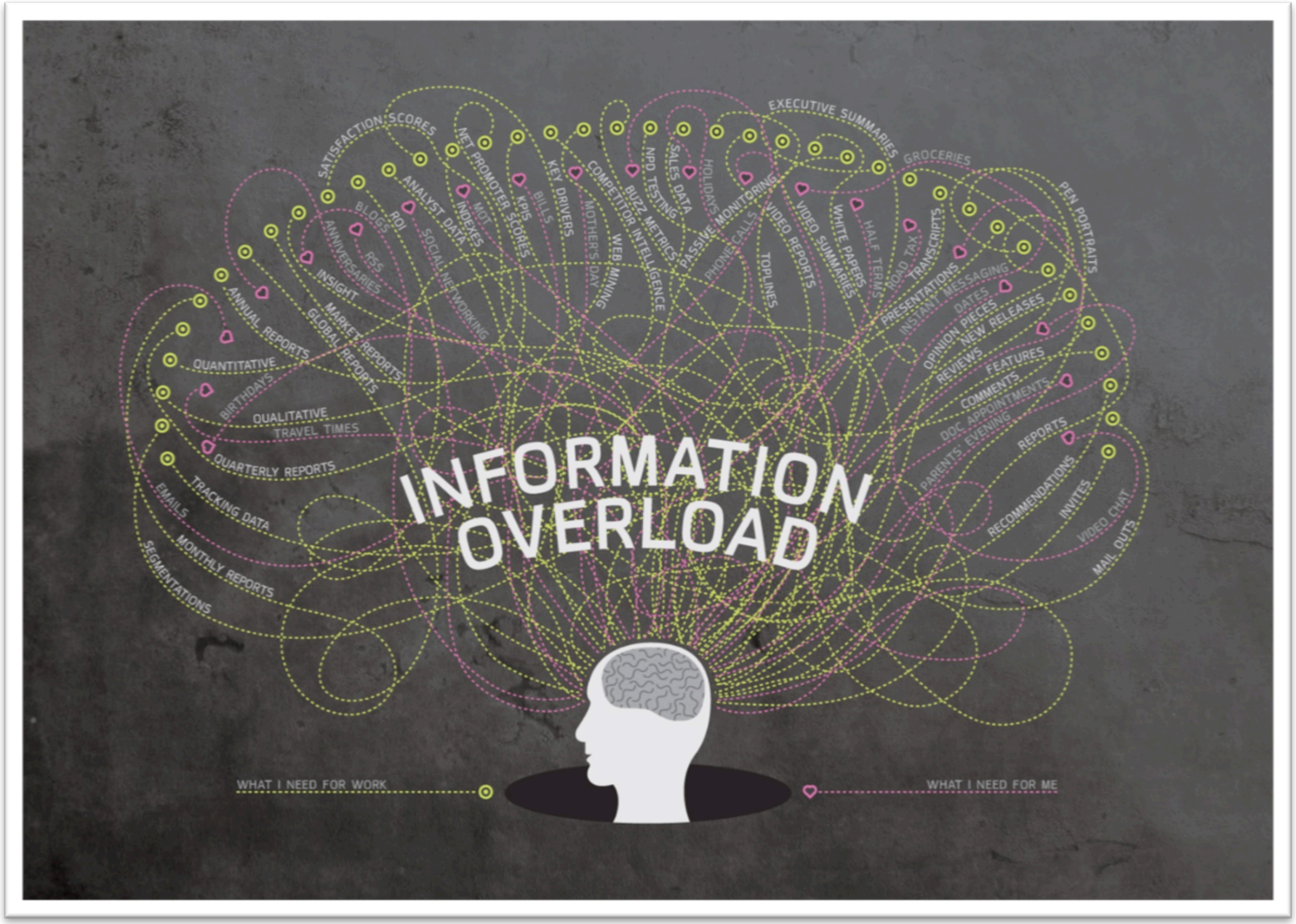
How?

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

What?

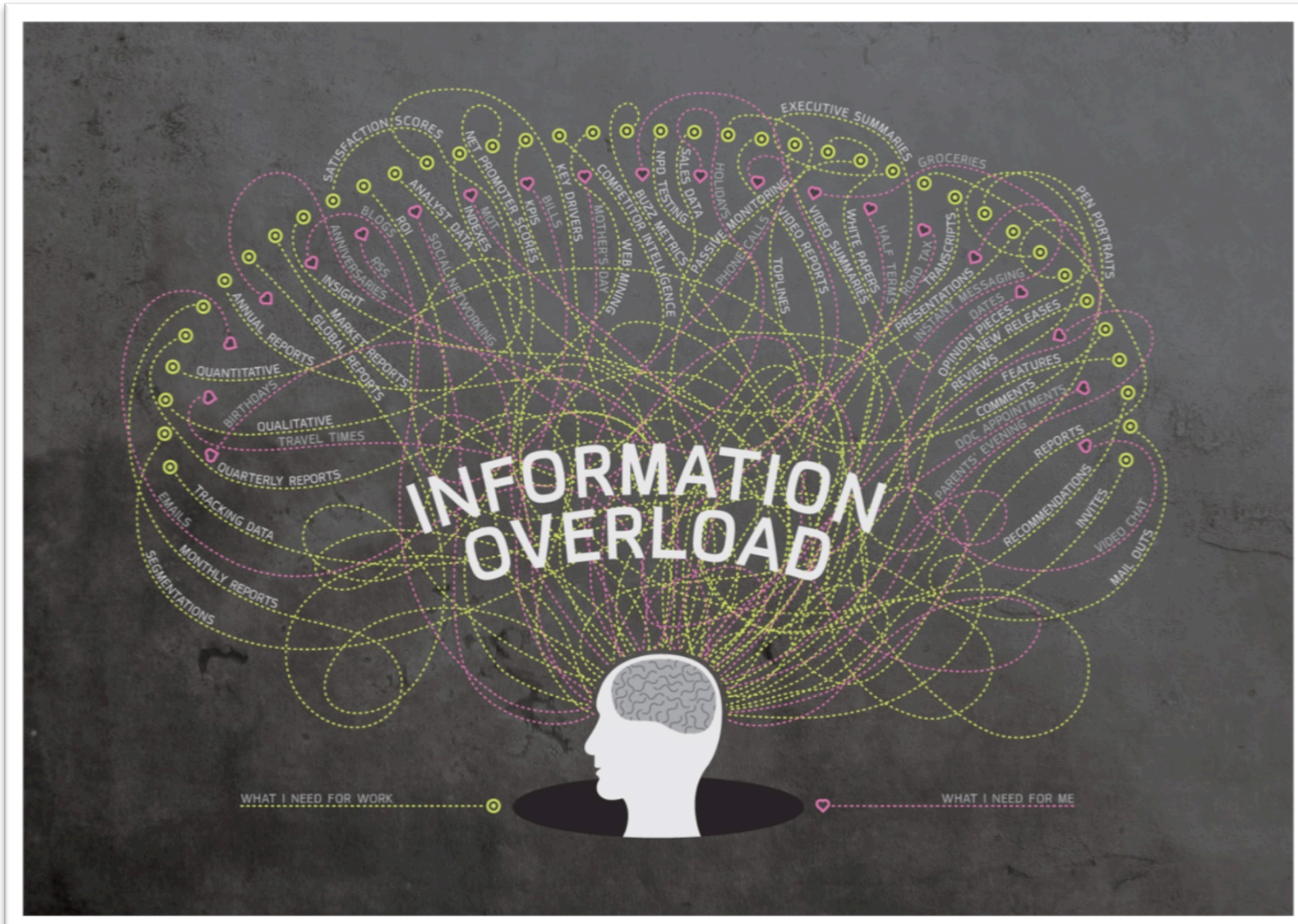
Explanation of Recommendations
Serendipity in Recommender Systems

Insight



Very huge availability of textual content

Insight



We can use this huge amount of content to directly learn a representation of words

Insight

Pass me a **Peroni!**

I like **Peroni**

Football and **Peroni**, what a perfect Saturday!

What is «Peroni» ?

Insight

Pass me a **Budweiser!**

I like **Budweiser**

Football and **Budweiser**, what a perfect Saturday!

What is «Budweiser» ?

Insight

Pass me a **Budweiser!**
I like **Budweiser**
Football and **Budweiser**, what a perfect Saturday!

What is «Budweiser» ?

Insight

Pass me a **Peroni!**

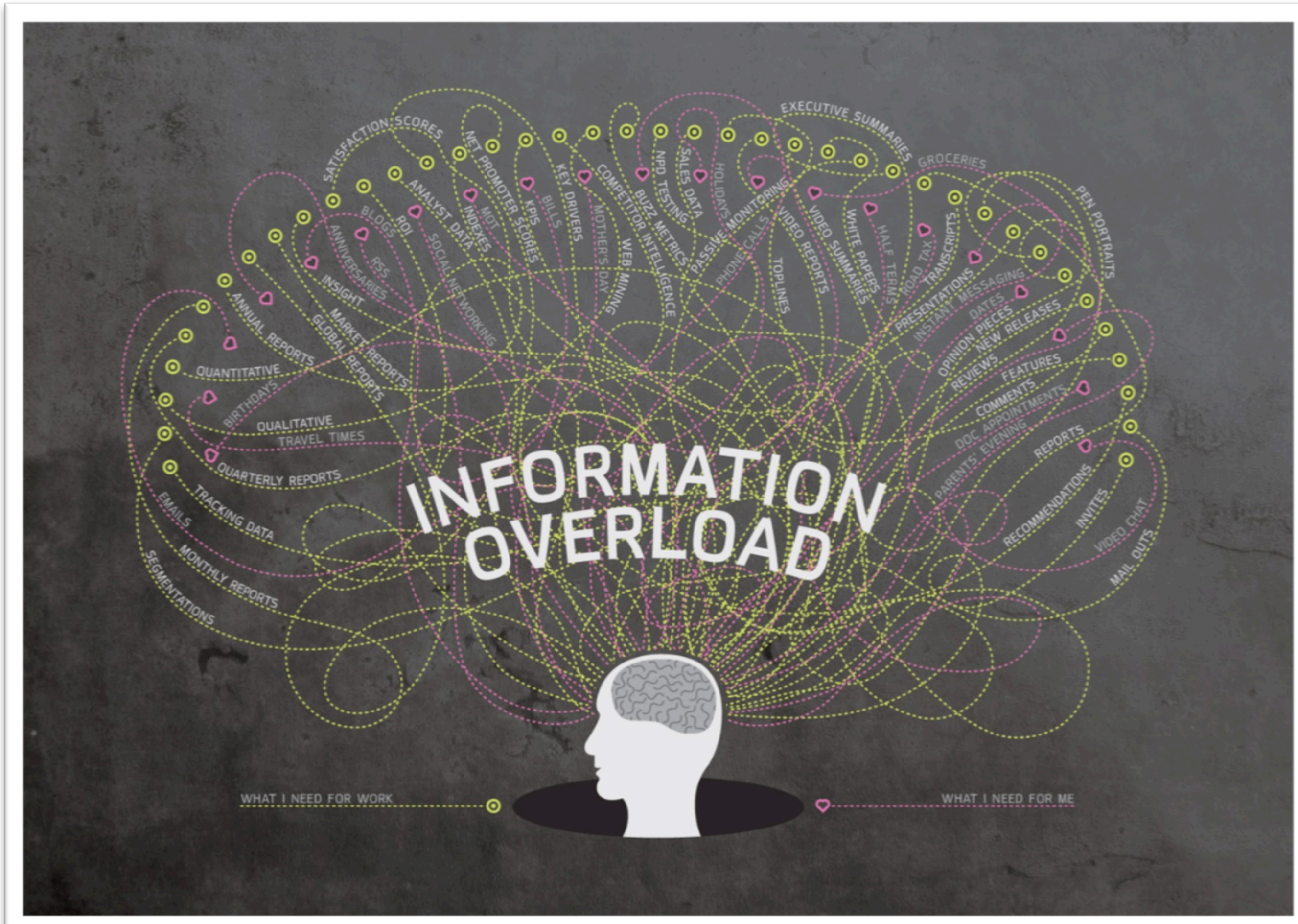
I like **Peroni**

Football and **Peroni**, what a perfect Saturday!

What is «Peroni» ?

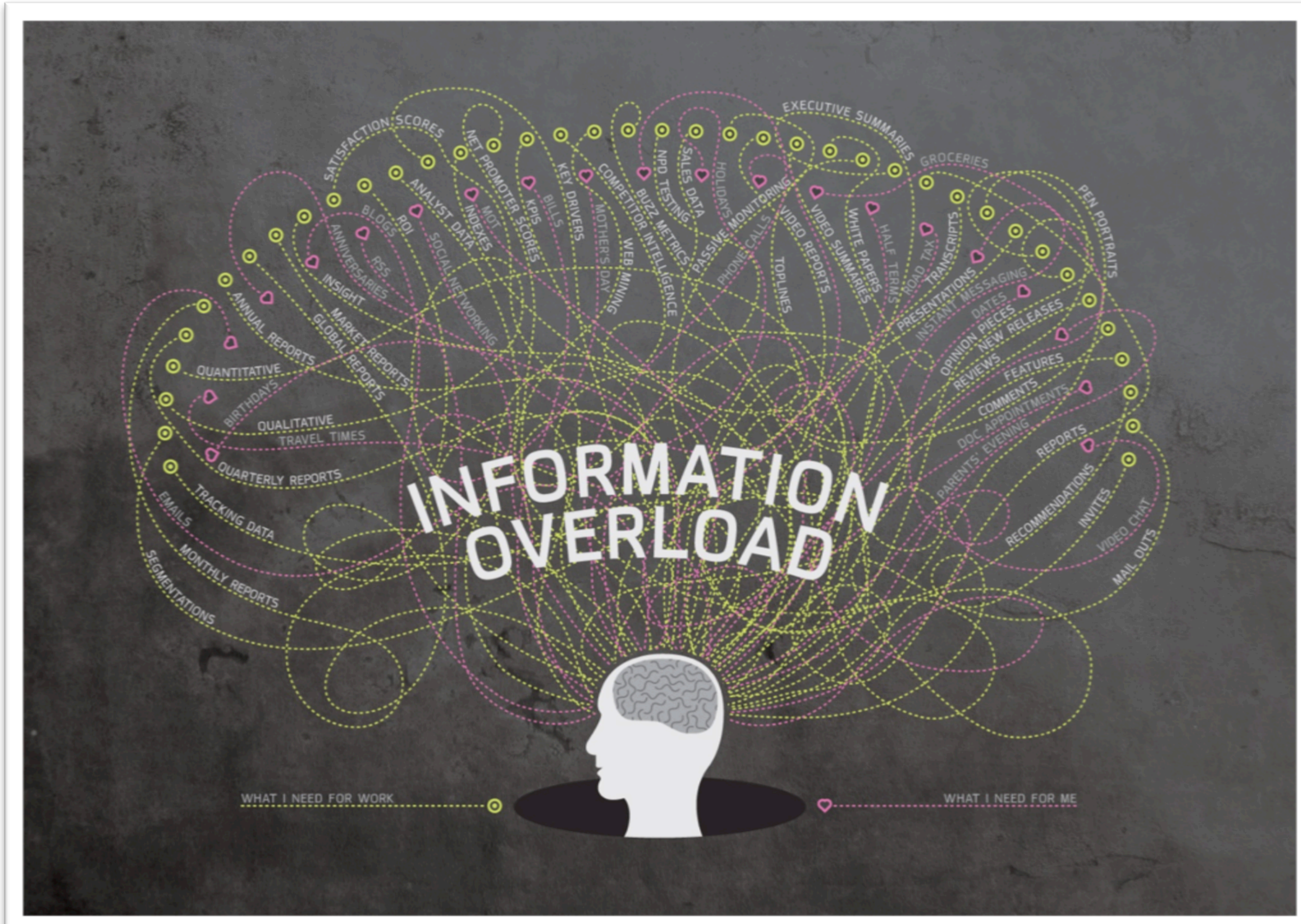
The most famous beer in Bari !

Insight



The semantics learnt according to term usage is called «distributional»

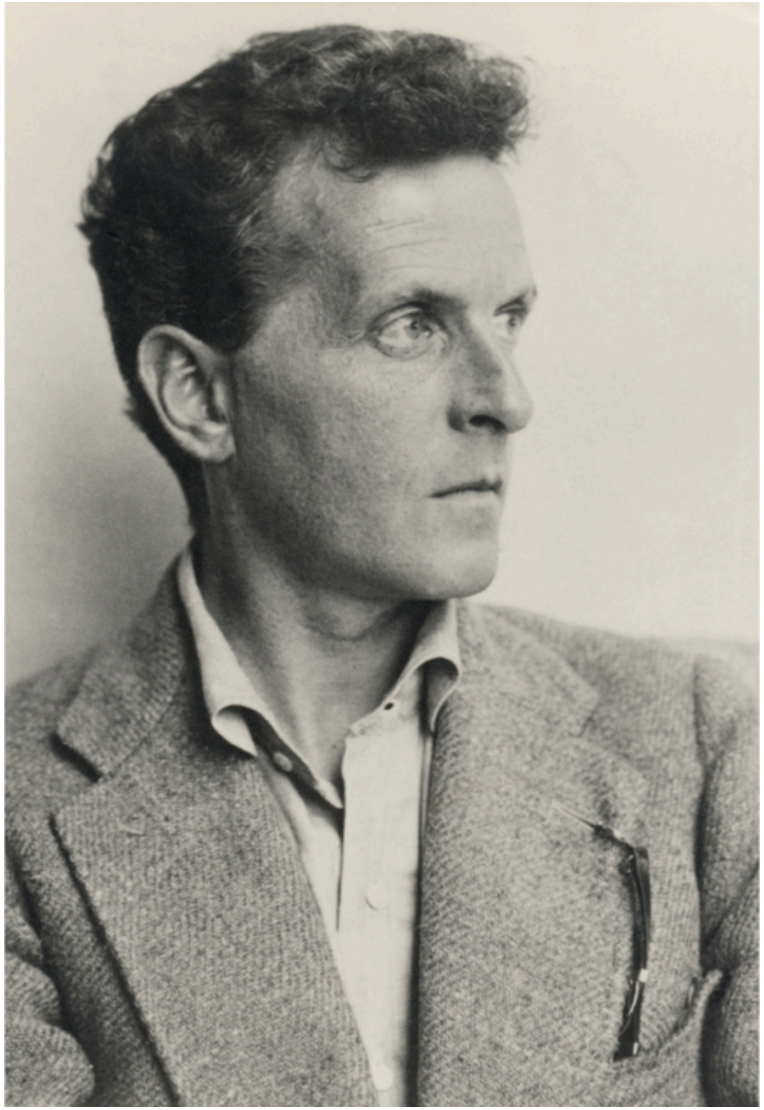
Insight



Distributional Hypothesis

«Terms used in similar contexts share a similar meaning»

Distributional Semantics



Meaning of a word is determined by its usage

Ludwig Wittgenstein
(Austrian philosopher)

Distributional Semantics

Definition

by analyzing large corpora of textual data it is possible to infer information about the usage (about the meaning) of the terms

(*) Firth, J.R. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32, 1957.

wine ≈ beer

similar meanings

dog ≈ cat

Distributional Semantics

wine ≈ beer

Definition

by analyzing large corpora of textual data it is possible to infer information about the usage (about the meaning) of the terms

similar meanings

(*) Firth, J.R. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32, 1957.

dog ≈ cat

Beer and wine, dog and cat share a similar meaning since they are often used in similar contexts

Distributional Semantics

Term-Context Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

A **vector-space representation** is learnt
by encoding in which context each term is used

<u>Each row of the matrix is a vector</u>

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

beer vs wine: good overlap

Similar!

Distributional Semantics

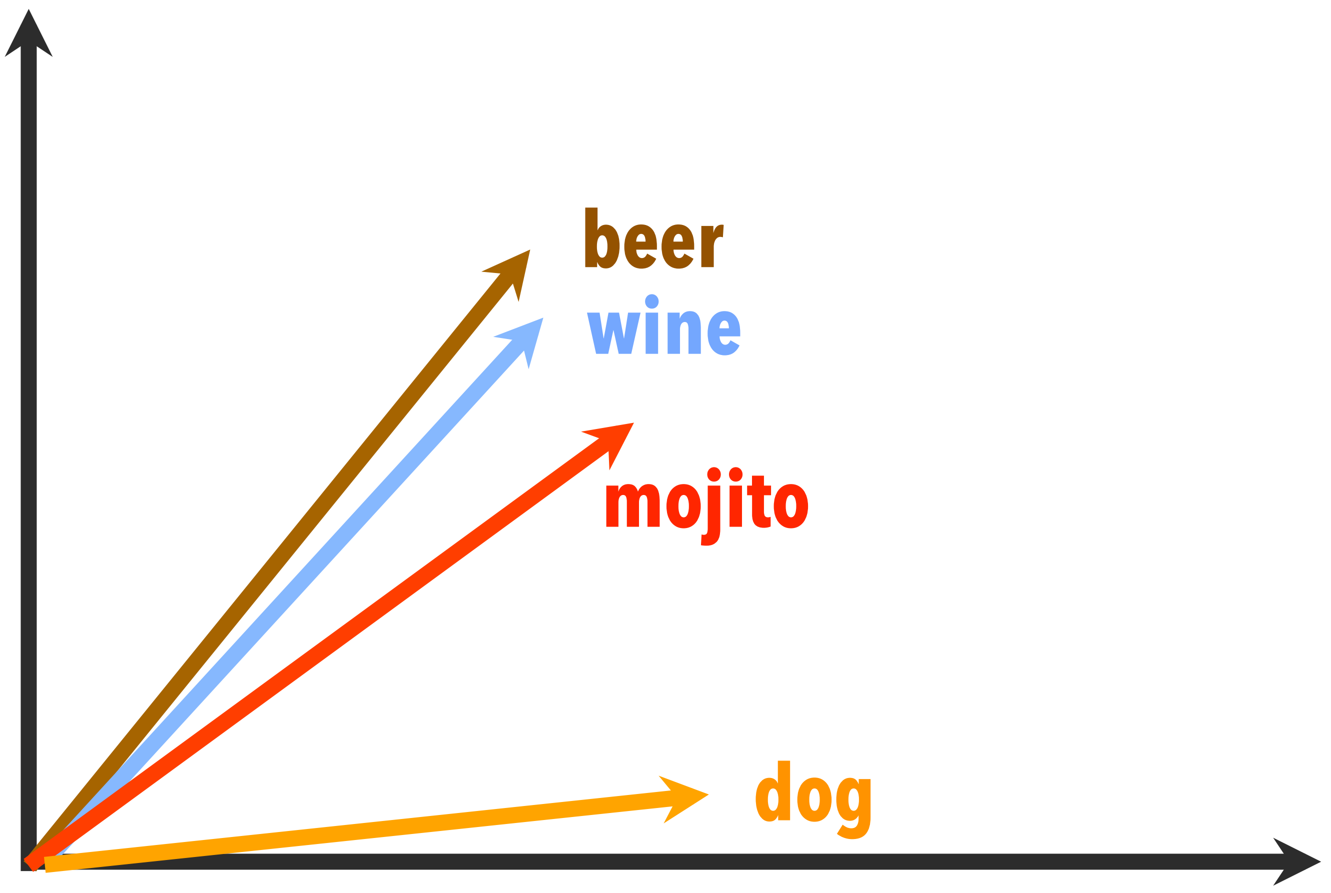
Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

beer vs spoon: no overlap

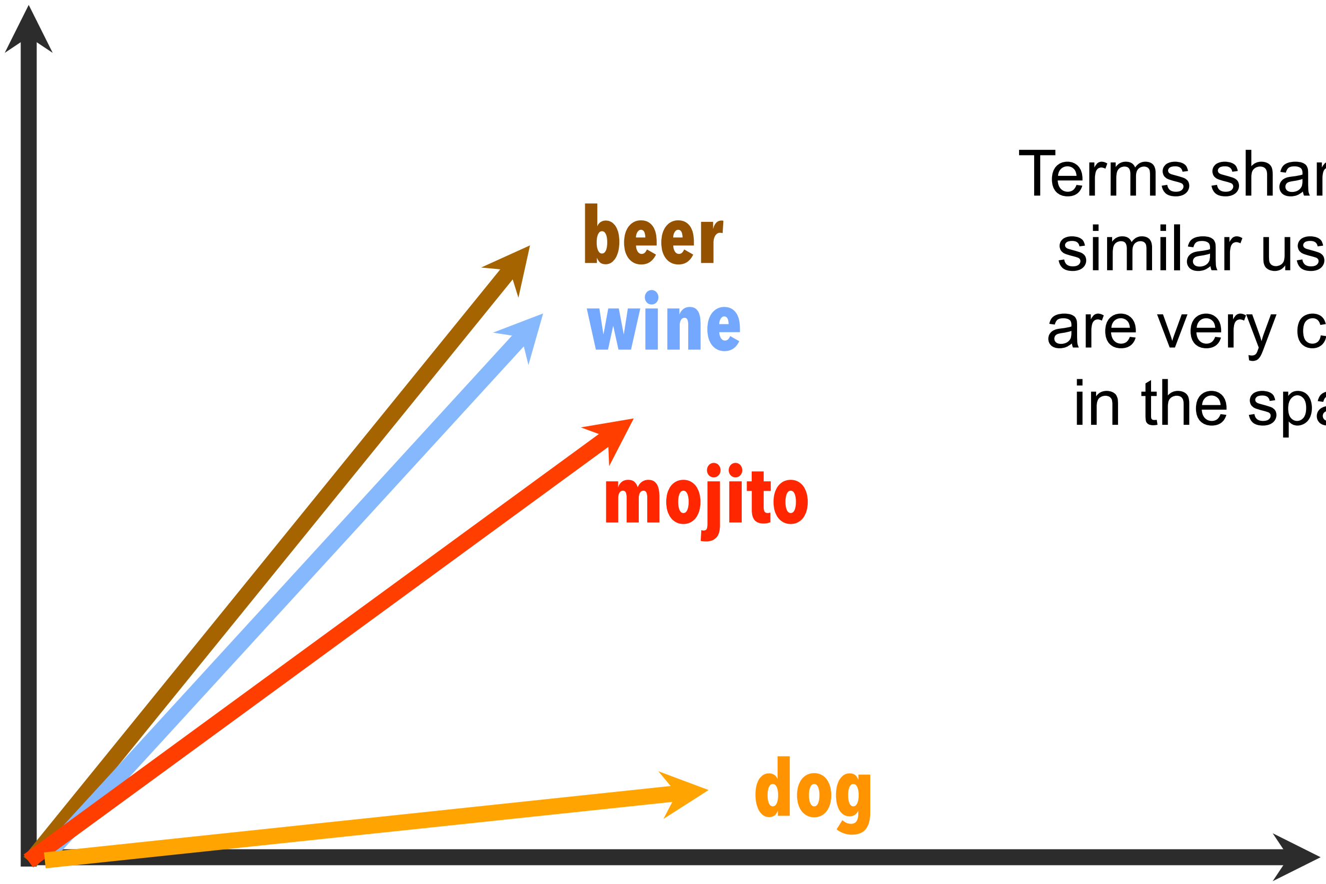
Not Similar!

WordSpace



A vector space representation (called WordSpace) is learnt according to terms usage in contexts

WordSpace



Terms sharing a similar usage are very close in the space

A vector space representation (called WordSpace) is learnt according to terms usage in contexts

Distributional Semantics

Term-Context Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Key question: **what is the context?**

Distributional Semantics

Term-Context Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Key question: **what is the context?**

These approaches are very flexible since the «context» can be set **according to the granularity required by the representation**

Distributional Semantics

Term-Context Matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Key question: **what is the context?**

Coarse-grained granularity:

context=whole document

Distributional Semantics

Term-Context Matrix = Term-Document Matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Key question: **what is the context?**

(This is Vector Space Model!)

Vector Space Model is a Distributional Model

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Key question: **what is the context?**

Fine-grained granularities:

context=paragraph, sentence, window of words

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

Fine-grained granularities:

PROs: the more fine-grained the representation, **more precise the vectors**
CONs: the more fine-grained the representation, **the bigger the matrix**

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✔	✔			✔	✔		
wine		✔	✔			✔	✔	✔	
spoon	✔			✔				✔	✔
glass	✔	✔	✔		✔				✔

The flexibility of distributional semantics models
also regards the rows of the matrix

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✔	✔			✔	✔		
concept2		✔	✔			✔	✔	✔	
concept3	✔			✔				✔	✔
concept4	✔	✔	✔		✔				✔

The flexibility of distributional semantics models
also regards the rows of the matrix

Keywords can be replaced with concepts
(as synsets or entities!)

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✔		✔		✔	✔		✔
Al Pacino			✔			✔			
American Writers	✔			✔				✔	✔
Laurence Fishburne	✔		✔		✔				✔

The flexibility of distributional semantics models
also regards the rows of the matrix

Keywords can be replaced with concepts
(as synsets or entities!)

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✔		✔		✔	✔		✔
Al Pacino			✔			✔			
American Writers	✔			✔					
Laurence Fishburne	✔		✔		✔				


Drama film

From Wikipedia, the free encyclopedia

This article **possibly contains original research**. Please improve it citations. Statements consisting only of original research should be r

A **drama film** is a film genre that depends mostly on in-depth development of realistic characters alcoholism, drug addiction, infidelity, moral dilemmas, racist prejudice, religious intolerance, sexua corruption put the characters in conflict with themselves, others, society and even natural phenom subgenres such as romantic drama, sport films, period drama, courtroom drama and crime.[1]

At the center of a drama is usually a character or characters who are in conflict at a crucial mome *Ordinary People* dig under the skin of everyday life to ask big questions and touch on the deepest tragic or at least painful resolutions and concern the survival of some tragic crisis, like the death o *Kramer*). Some of the greatest screen performances come from dramas, as there is ample opport afford.[2]

Drama films have been nominated frequently for the Academy Award (particularly Best Picture) - r

Contents [hide]

1 Sub-genres
2 Early film-1950s
3 1960s-1970s
4 1980s-1990s

Keanu Reeves and Al Pacino are «connected» because they both acted in Drama Films

Distributional Semantics

Representing Documents

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✔		✔		✔	✔		✔
Al Pacino			✔			✔			
American Writers	✔			✔				✔	✔
Laurence Fishburne	✔		✔		✔				✔

Given a WordSpace, a vector space representation of documents (called DocSpace) is typically built as the centroid vector of word representations

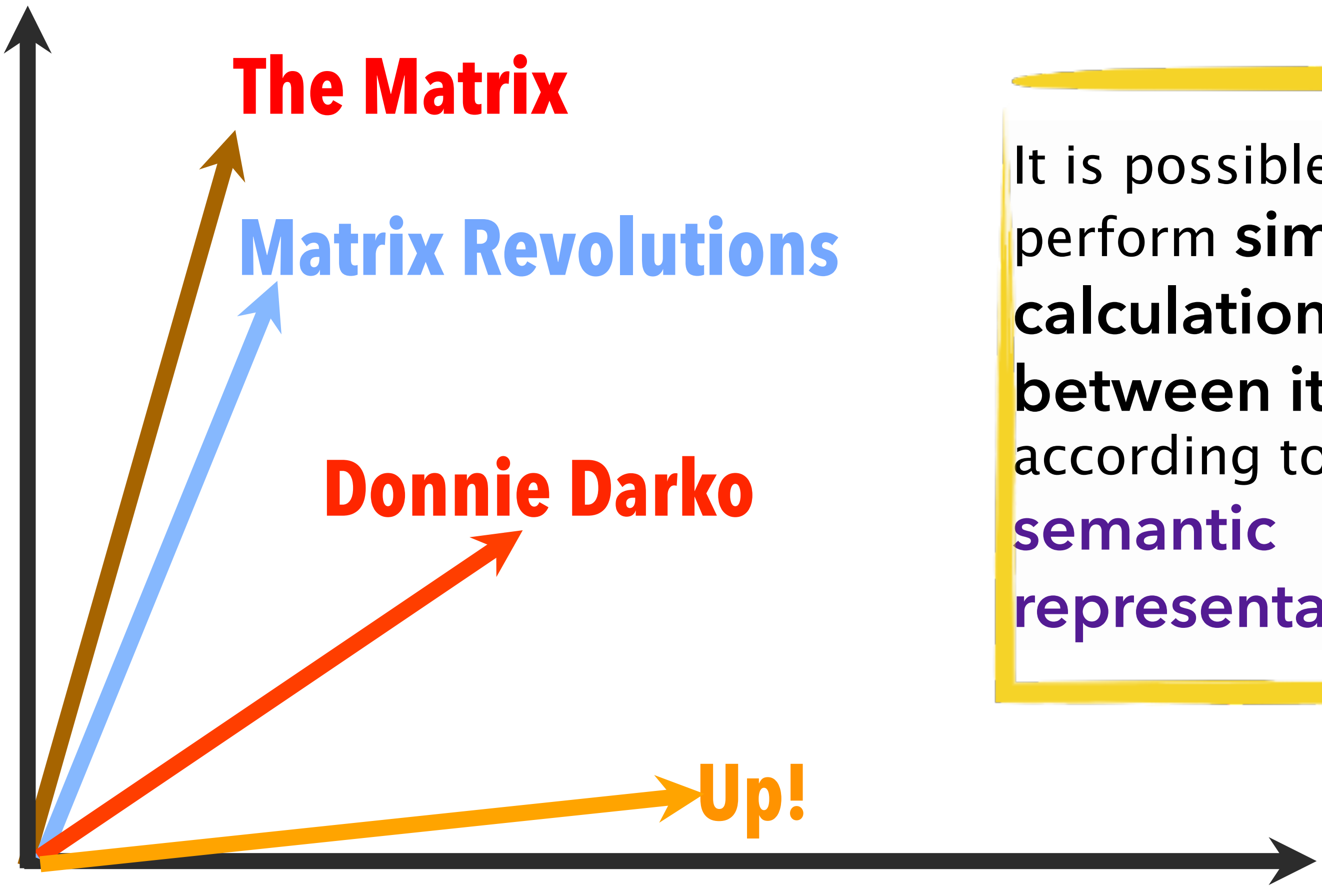
Distributional Semantics

Representing Documents

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✔		✔		✔	✔		✔
Al Pacino			✔			✔			
American Writers	✔			✔				✔	✔
Laurence Fishburne	✔		✔		✔				✔

The Matrix	✔	✔	✔	✔	✔	✔		✔	✔

DocSpace



It is possible to perform **similarity calculations between items** according to their **semantic representation**

Given a WordSpace, a vector space representation of documents (called DocSpace) is typically built as the centroid vector of word representations

Distributional Semantics

- We can exploit the (big) corpora of data to **directly learn a semantic vector-space representation of the terms** of a language
- **Lightweight semantics,** not formally defined
- **High flexibility:** everything is a vector: term/term similarity, doc/term, term/doc, etc..
- **Context can have different granularities**
- **Huge amount of content** is needed
- **Matrices are particularly huge** and difficult to build
    - **Too many features: need for dimensionality reduction**

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

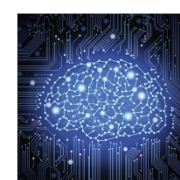
Introduce semantics **by linking** the item to a **knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

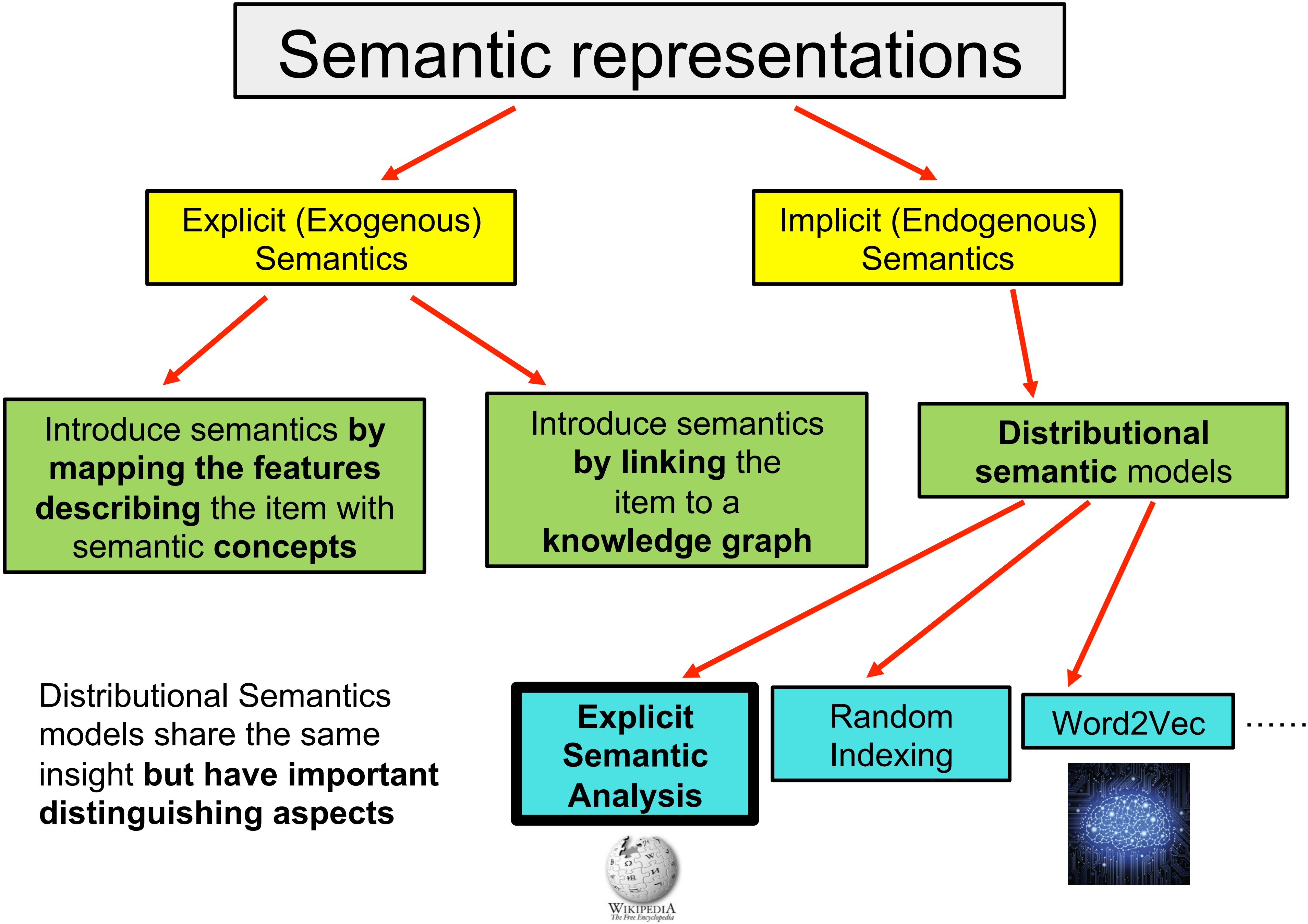
Random Indexing

Word2Vec



……

Distributional Semantics models share the same insight **but have important distinguishing aspects**


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics by mapping the features describing the item with semantic concepts
Introduce semantics by linking the item to a knowledge graph
Distributional semantic models
Explicit Semantic Analysis
Random Indexing
Word2Vec
WIKIPEDIA
The Free Encyclopedia
Distributional Semantics models share the same insight but have important distinguishing aspects

Explicit Semantic Analysis (ESA)

ESA matrix

ESA is a Distributional Semantic model which uses **Wikipedia articles** **as context**

Explicit Semantic Analysis (ESA)

ESA matrix

ESA is a Distributional Semantic model which uses **Wikipedia articles** **as context**

Wikipedia articles

Terms

ESA	Context 1	...	Context n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Explicit Semantic Analysis (ESA)

ESA matrix

semantic relatedness
between a word and a context
TF-IDF score

Wikipedia articles

ESA	Context 1	...	Context n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Terms

Explicit Semantic Analysis (ESA)

ESA matrix

semantic relatedness
between a word and a context
TF-IDF score

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Wikipedia articles

Terms

ESA	Context 1	...	Context n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Explicit Semantic Analysis (ESA)

Every Wikipedia article represents a **concept**

Panthera

From Wikipedia, the free encyclopedia

Panthera is a genus of the family Felidae (the cats), which contains four well-known living species: the lion, tiger, jaguar, and leopard. The genus comprises about half of the big cats. One meaning of the word ***panther*** is to designate cats of this family. Only these four cat species have the anatomical changes enabling them to roar. The primary reason for this was assumed to be the incomplete ossification of the hyoid bone. However, new studies show that the ability to roar is due to other morphological features, especially of the larynx. The snow leopard, *Uncia uncia*, which is sometimes included within *Panthera*, does not roar. Although it has an incomplete ossification of the hyoid bone, it lacks the special morphology of the larynx, which is typical for lions, tigers, jaguars and leopards.[1]

Panthera

Tiger

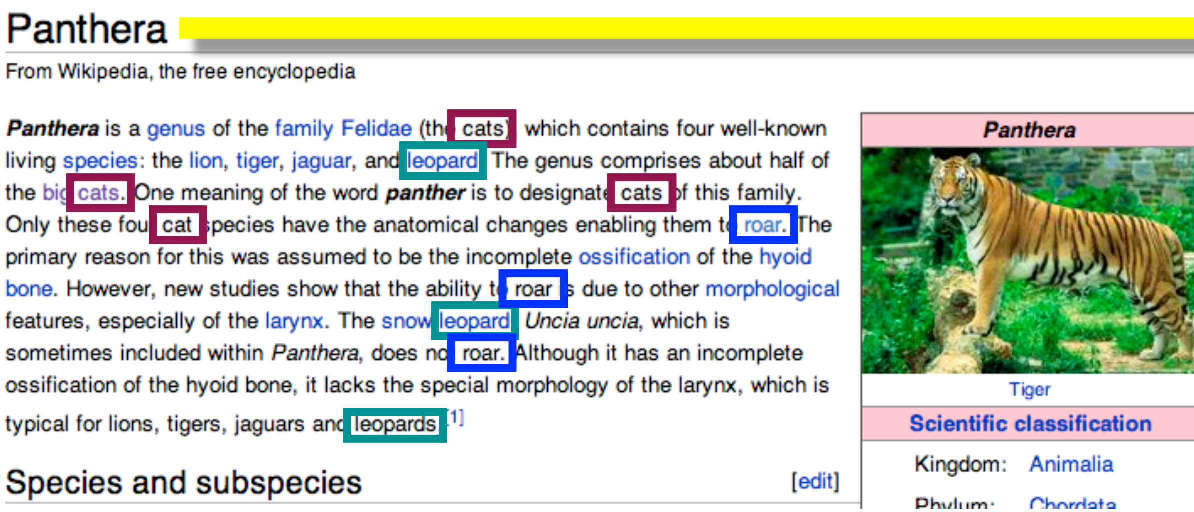
Scientific classification

Kingdom:	Animalia
Phylum:	Chordata

Species and subspecies [edit]

Explicit Semantic Analysis (ESA)

Every Wikipedia article represents a **concept**


Panthera

From Wikipedia, the free encyclopedia

Panthera is a genus of the family Felidae (the cats) which contains four well-known living species: the lion, tiger, jaguar, and leopard. The genus comprises about half of the big cats. One meaning of the word ***panther*** is to designate cats of this family. Only these four cat species have the anatomical changes enabling them to roar. The primary reason for this was assumed to be the incomplete ossification of the hyoid bone. However, new studies show that the ability to roar is due to other morphological features, especially of the larynx. The snow leopard, *Uncia uncia*, which is sometimes included within *Panthera*, does not roar. Although it has an incomplete ossification of the hyoid bone, it lacks the special morphology of the larynx, which is typical for lions, tigers, jaguars and leopards.[1]

Species and subspecies [edit]

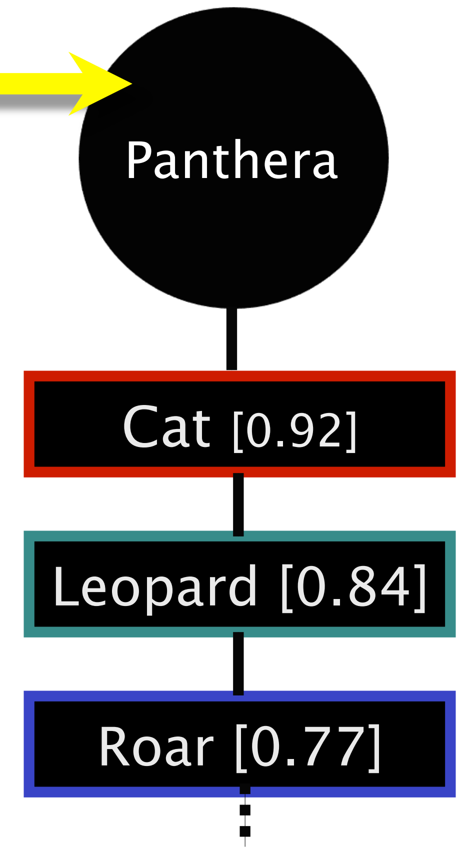
Panthera

Tiger

Scientific classification

Kingdom: Animalia

Phylum: Chordata



Article words are associated with the **concept** (TF-IDF)

(this is a ***column*** *of ESA matrix)*

Each Wikipedia page can be described in terms of the words with the highest TF-IDF score

Explicit Semantic Analysis (ESA)

ESA	Panthera	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

We iterate the process over (almost) all the Wikipedia pages and we obtain

the so-called ESA matrix

Explicit Semantic Analysis (ESA)

ESA matrix

ESA	Panthera	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Each row of the ESA matrix is called

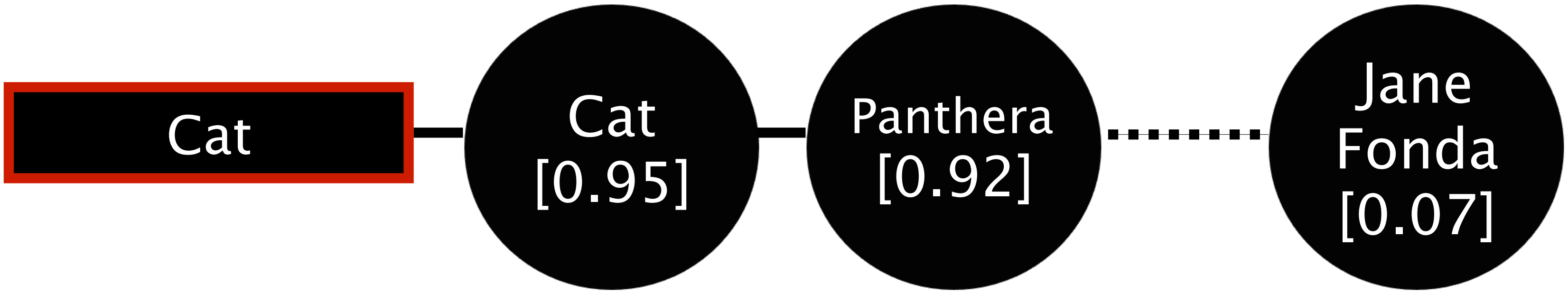
semantic interpretation vector

(of a term t)

Explicit Semantic Analysis (ESA)

Semantic interpretation vector of the term ‘cat’

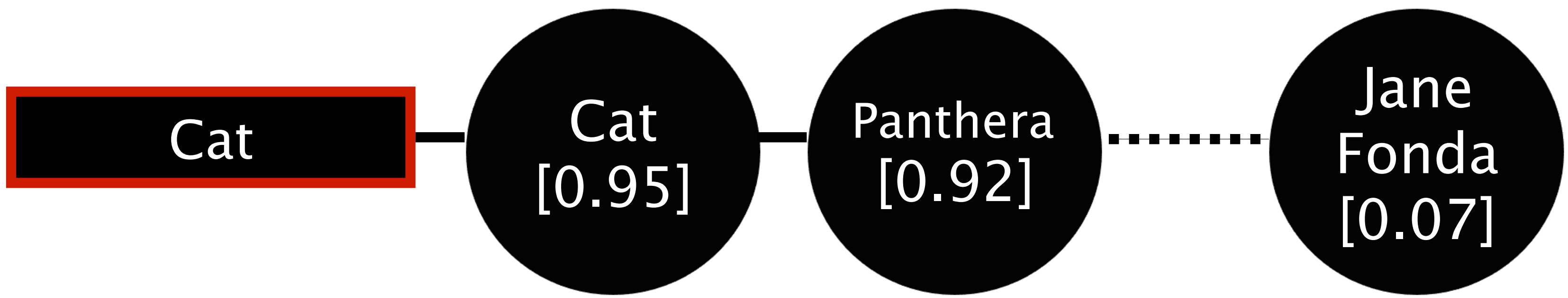
(Wikipedia articles are ranked in a descending order)



Explicit Semantic Analysis (ESA)

Semantic interpretation vector of the term ‘cat’

(Wikipedia articles are ranked in a descending order)

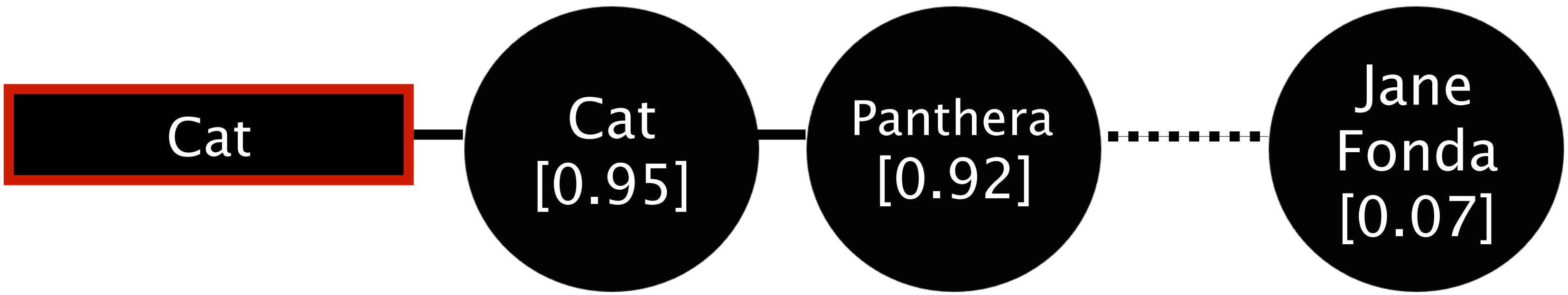


The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts.

Explicit Semantic Analysis (ESA)

Semantic interpretation vector of the term ‘cat’

(Wikipedia articles are ranked in a descending order)

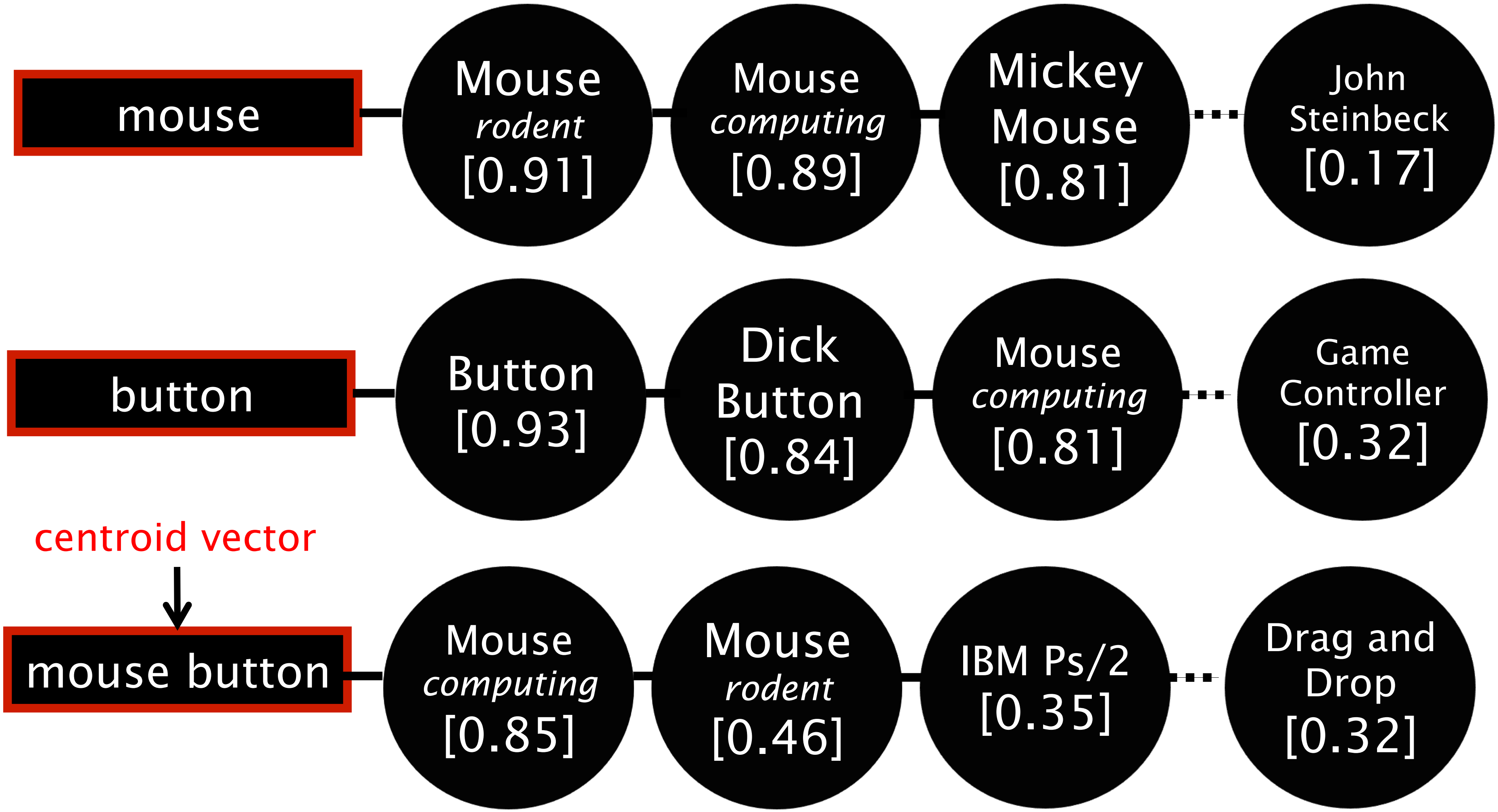


The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts.

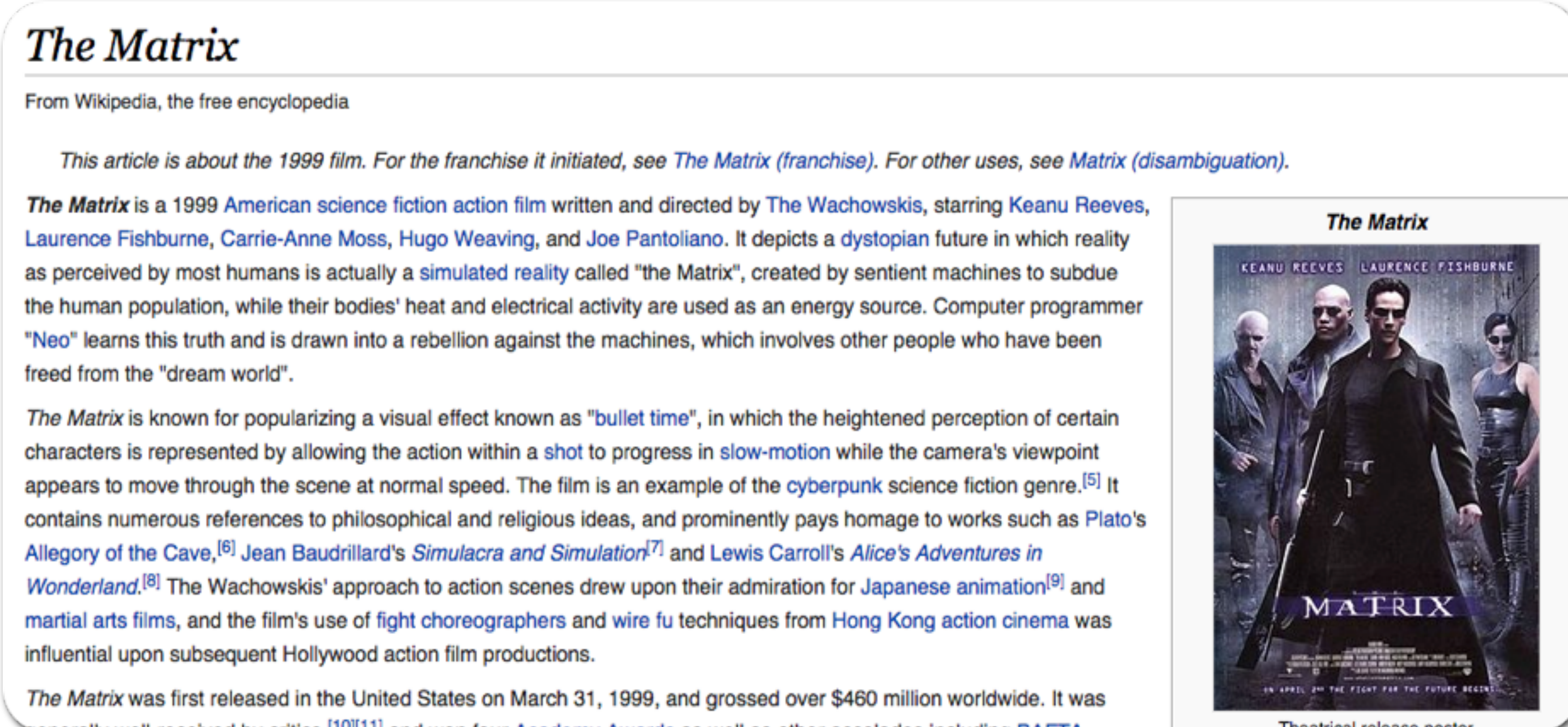
The **highest** the score, the **more** the strength of its ‘**semantic connection’** with a Wikipedia concept

Explicit Semantic Analysis (ESA)

Semantic interpretation vector of **text fragment**
(e.g. '**mouse button**')
is the centroid vector of the terms in the fragment



Explicit Semantic Analysis (ESA)



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

A semantic representation of an item can be built as the **centroid vector** of the **semantic interpretation vectors of the terms** in the item description

Explicit Semantic Analysis (ESA)

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.

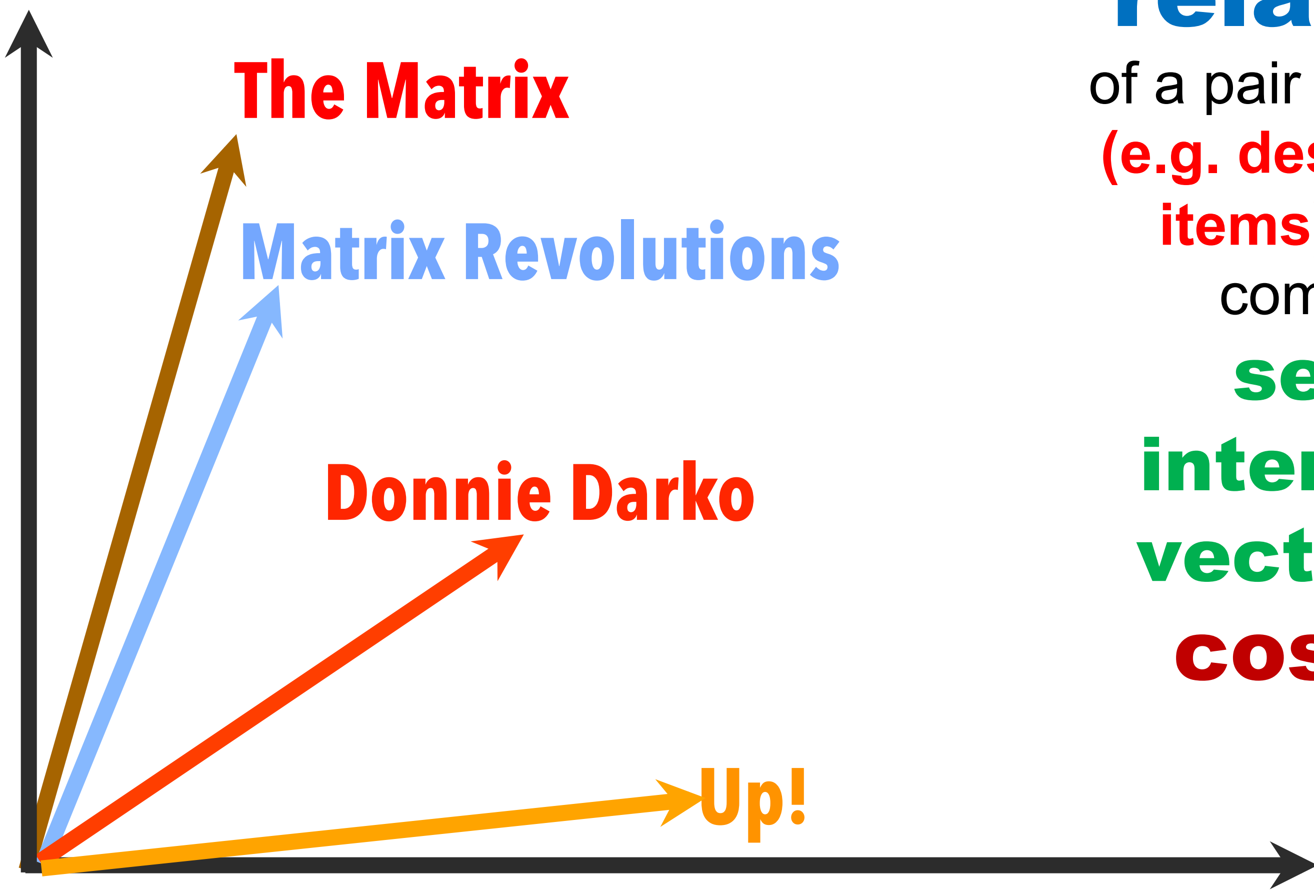
The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well received by critics,[10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix

Theatrical release poster

A semantic representation of an item can be built as the centroid vector of the **semantic interpretation vectors of the terms** in the item description

Explicit Semantic Analysis (ESA)



semantic relatedness of a pair of text fragments **(e.g. description of two items)** computed by comparing their **semantic interpretation vectors** using the **cosine** metric

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique

How can we generate a set of relevant extra concepts describing the items?

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique

How can we generate a set of relevant extra concepts describing the items?

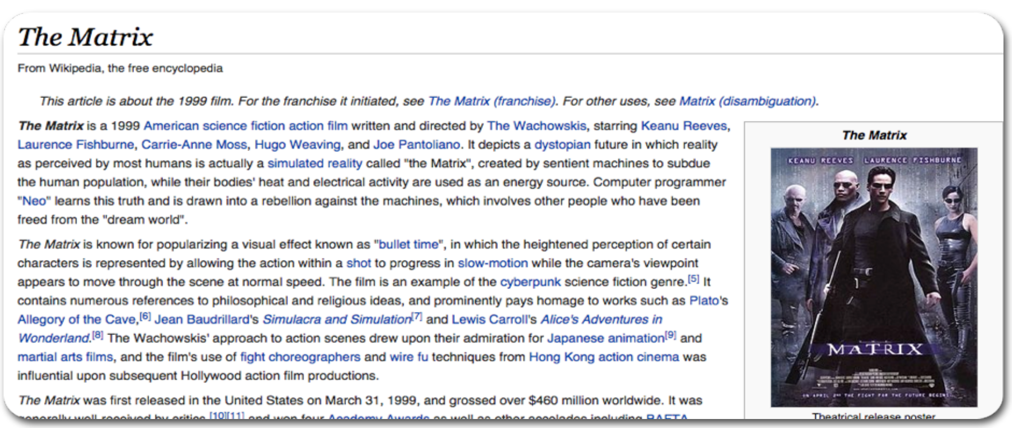
Given an item, we first generate its semantic interpretation vector

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique.

How can we generate a set of relevant extra concepts describing the items?

Given an item, we first generate its semantic interpretation vector



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

The Matrix

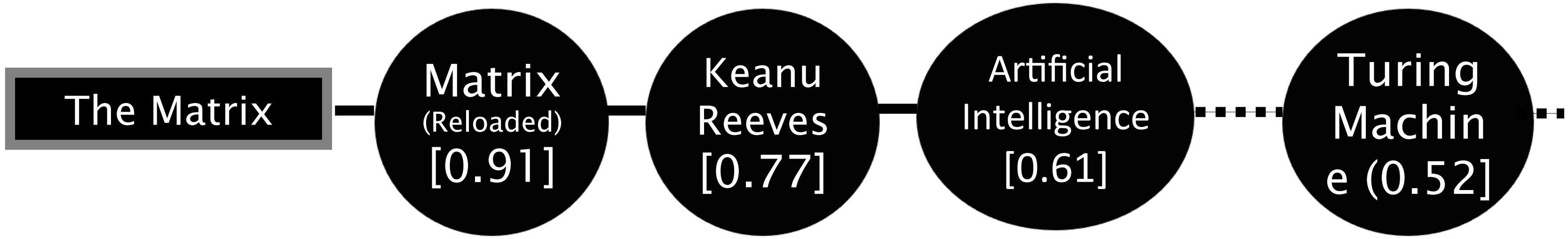
Theatrical release poster

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique.

How can we generate a set of relevant extra concepts describing the items?

Given an item, we first generate its semantic interpretation vector

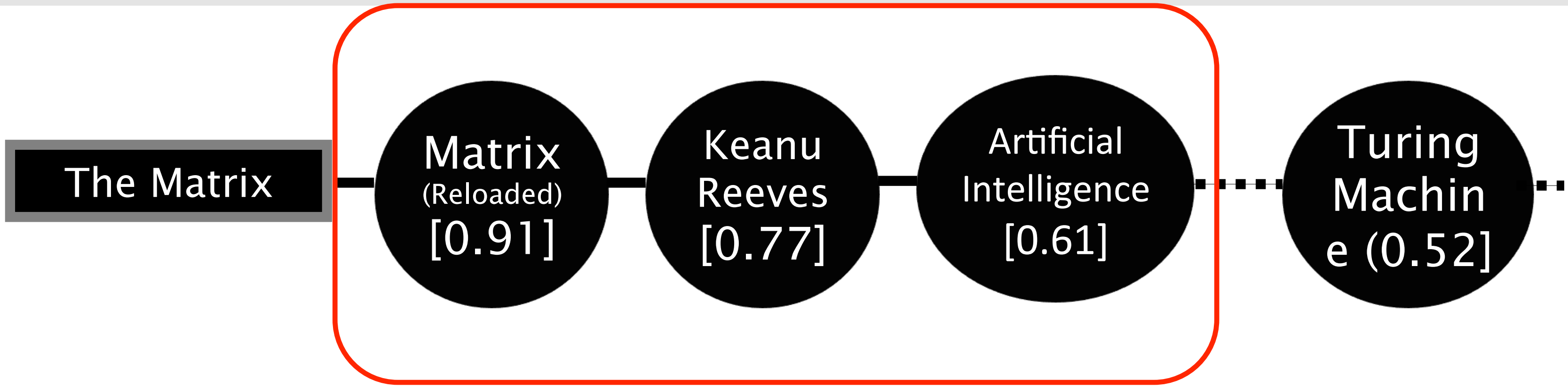


Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique.

How can we generate a set of relevant extra concepts describing the items?

The pages with the highest TF/IDF score in the semantic interpretation vector are the **most related concepts**



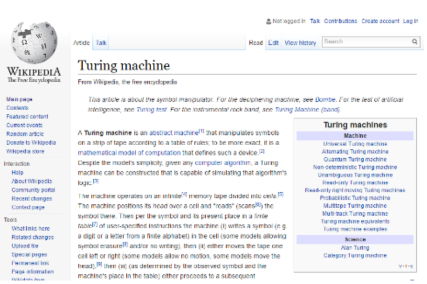
Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used as a **feature generation** technique.

How can we generate a set of relevant extra concepts describing the items?

The pages with the highest TF/IDF score in the semantic interpretation vector are the **most related concepts**

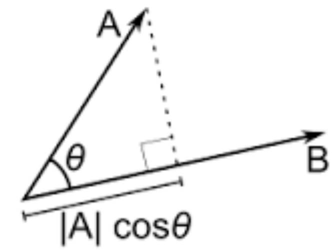
Extra features related to the item



ESA effectively used for

Text Categorization [Gabri09]
experiments on diverse datasets

Semantic relatedness of words and texts [Gabri09]
cosine similarity between vectors of ESA concepts



Information Retrieval [Egozi08, Egozi11]
ESA-based IR algorithm enriching documents and queries

what about **ESA** for **Information Filtering**?

[Gabri09] E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 34:443-498, 2009.

[Egozi08] Ofer Egozi, Evgeniy Gabrilovich, Shaul Markovitch: Concept-Based Feature Generation and Selection for Information Retrieval. *AAAI 2008*, 1132-1137, 2008.

[Egozi11] Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems* 29(2), April 2011.

Information Filtering using ESA

TV-domain

German Electronic Program Guides (EPG)

problem

description of TV shows **too short** or **poorly meaningful** to feed a **content-based recommendation algorithm**

solution

Explicit Semantic Analysis exploited to obtain an **enhanced representation**

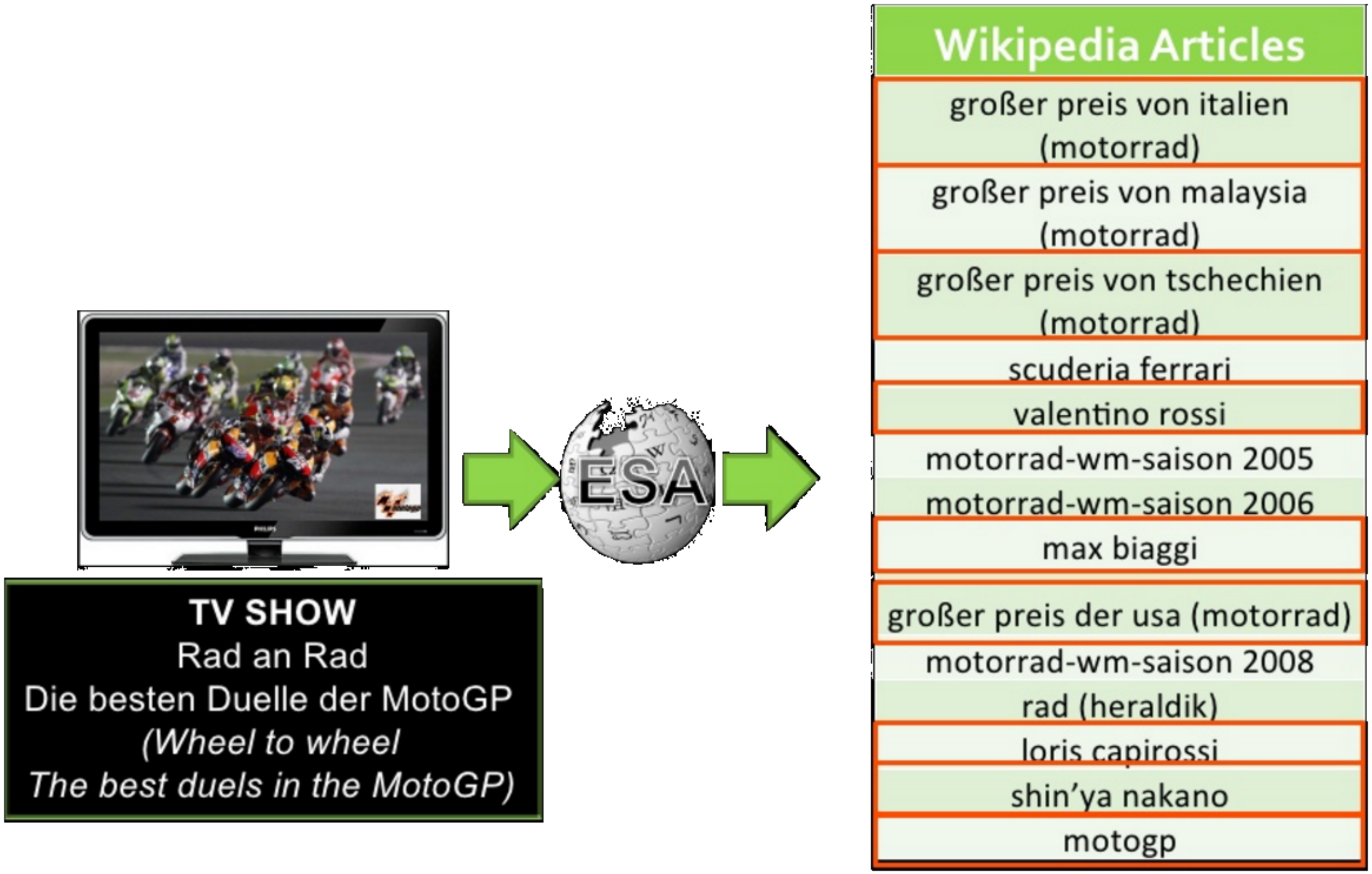
[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

TV SHOW
Rad an Rad
Die besten Duelle der MotoGP
(Wheel to wheel
The best duels in the MotoGP)

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides



[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

Wikipedia Articles related to the TV show **are added to the description**

TV SHOW
Rad an Rad
Die besten Duelle der MotoGP
(Wheel to wheel
The best duels in the MotoGP)

Wikipedia Articles
großer preis von italien (motorrad)
großer preis von malaysia (motorrad)
großer preis von tschechien (motorrad)
scuderia ferrari
valentino rossi
motorrad-wm-saison 2005
motorrad-wm-saison 2006
max biaggi
großer preis der usa (motorrad)
motorrad-wm-saison 2008
rad (heraldik)
loris capirossi
shin'ya nakano
motogp

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
sports
motorbike
...
competition

tv show

2012 Superbike Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
sports
motorbike
...
competition

No matching!

tv show

2012 Superbike Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
superbike
sports
motorbike
formula 1
…
competition

Through ESA we can add new features to the profile **and we can improve the overlap between textual description**

tv show

2012 Superbike Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
superbike
sports
motorbike
formula 1
…
competition

Matching!

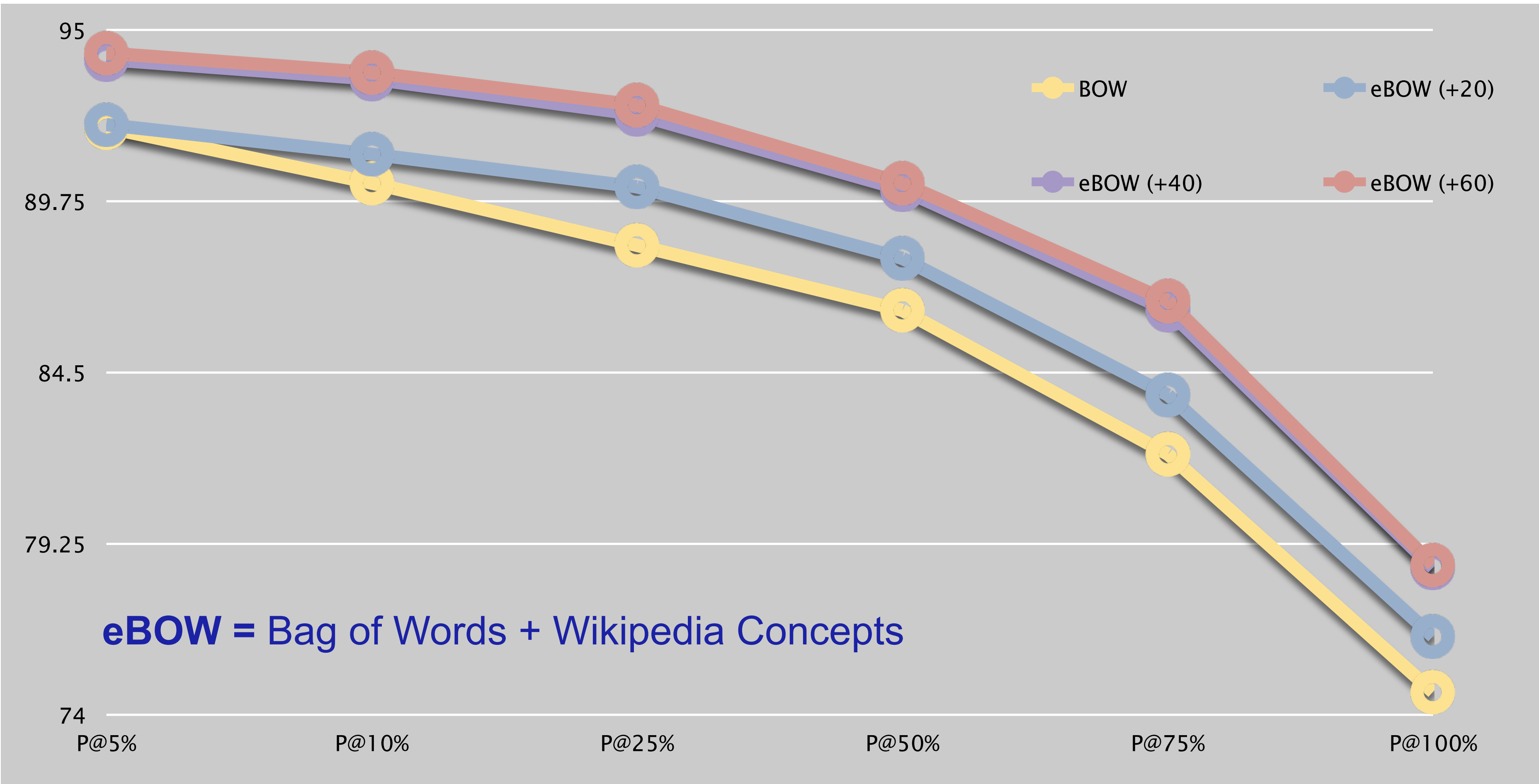
tv show

2012 Superbike Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

results on Aprico.tv data



The more Wikipedia Concepts are added to the textual description of the items (eBOW+60), **the best the precision of the algorithm**

Explicit Semantic Analysis (ESA)

Distributional Model which uses
Wikipedia Article as context

Very Transparent representation
(columns have an explicit meaning)

Can be used as a feature generation tool

The whole matrix is very huge

«Empirical» tuning of the parameters:
how many articles? How many terms?
What is the thresholding?

Recent Developments of Content-Based RecSys

Endogenous Approaches: Random Indexing & Word2Vec

Cataldo Musto

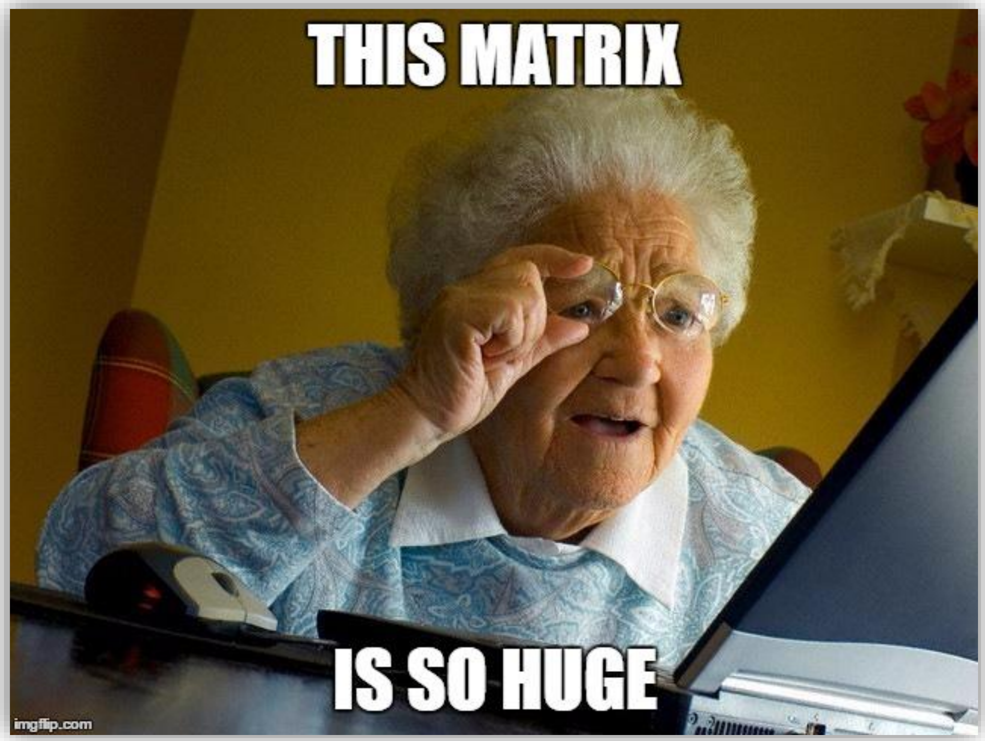
Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

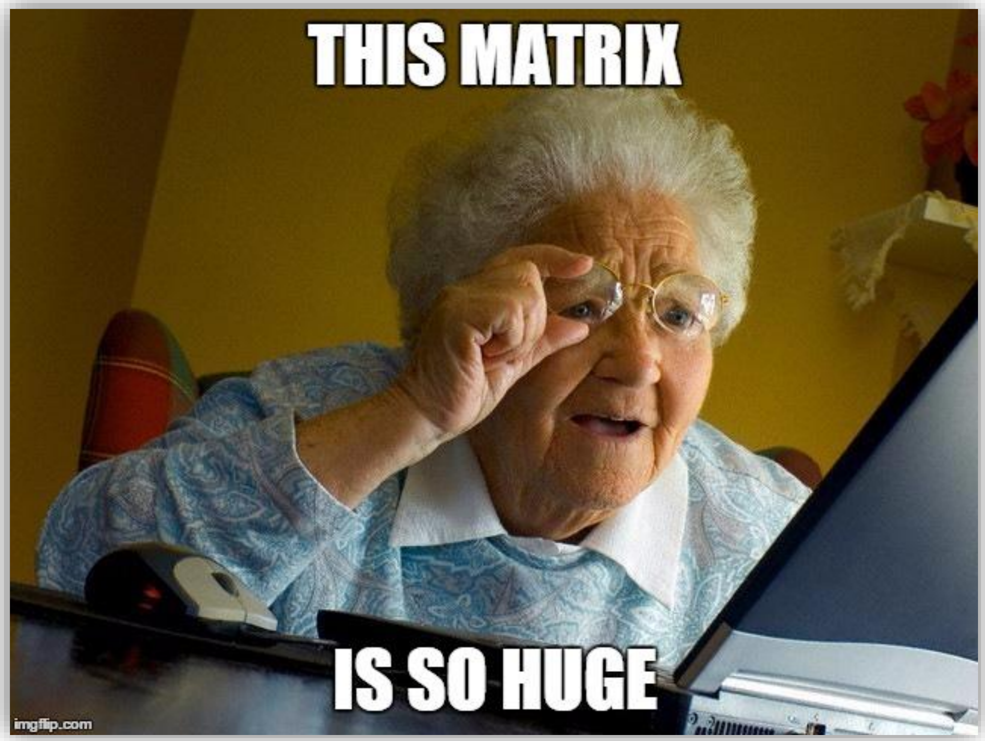
SWAP researchgroup

Dimensions

are important.

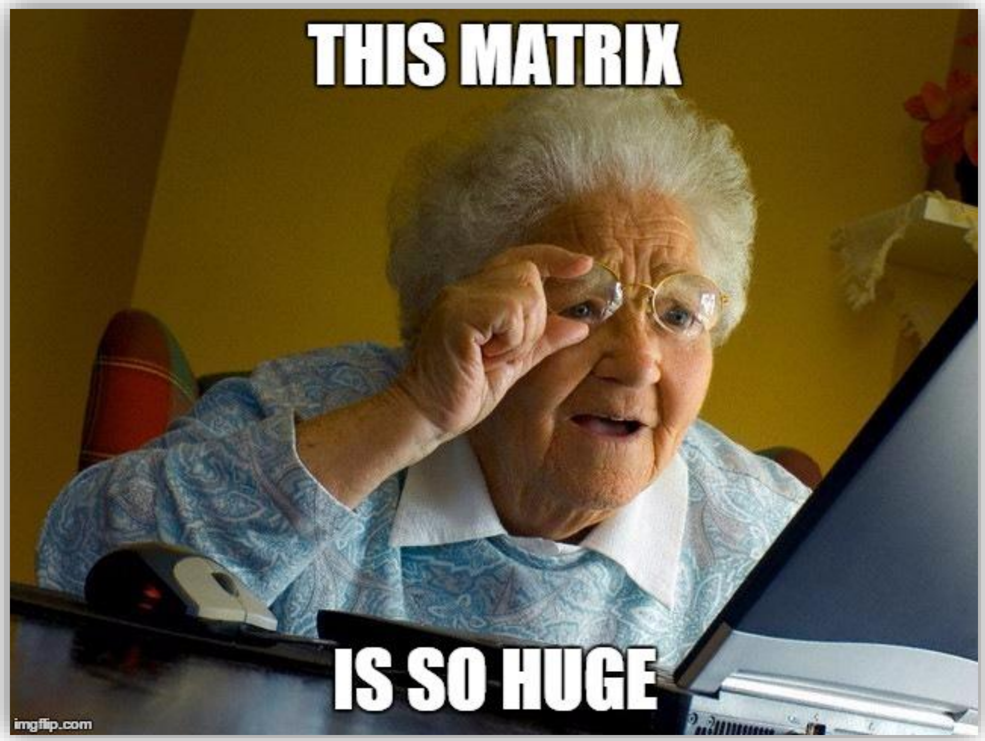


When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items



When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items

Dimensionality Reduction techniques



When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items

a.k.a. Word embedding techniques

Embedding = a smaller representation of words

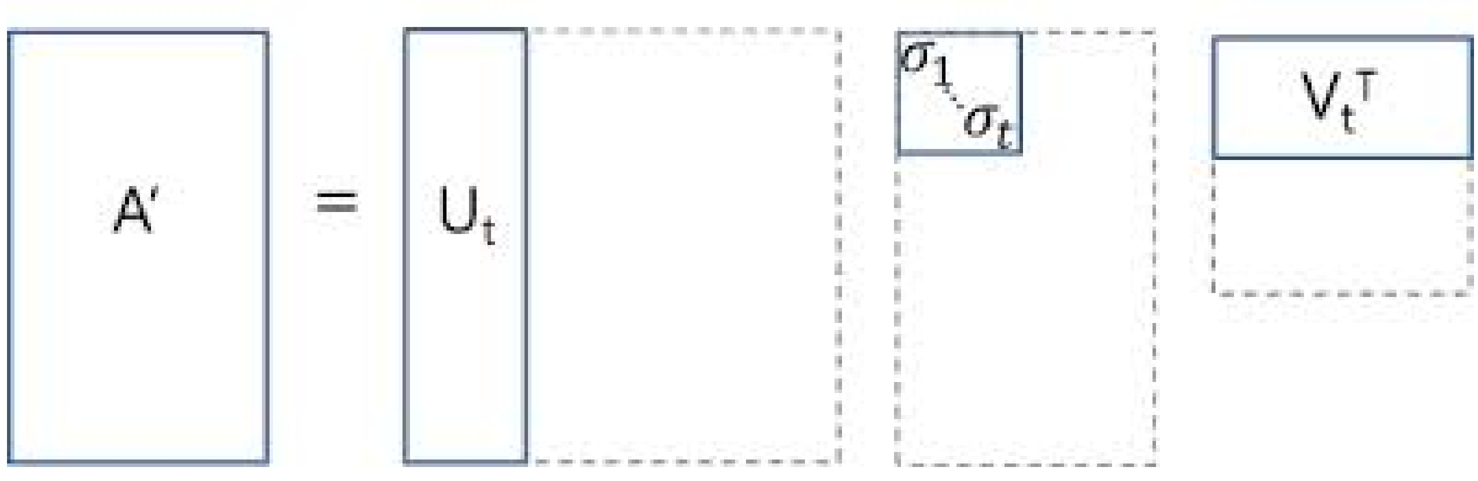
Is this new?

Dimensionality reduction techniques

Latent Semantic Analysis (LSA) **is a widespread** distributional semantics model **which builds a term/context matrix** and calculates SVD over that matrix.

Dumais, Susan T. "Latent semantic analysis." *Annual review of information science and technology* 38.1 (2004): 188-230.

Truncated Singular Value Decomposition



induces **higher-order (paradigmatic)** relations through the truncated SVD

Dimensionality reduction techniques

Singular Value Decomposition

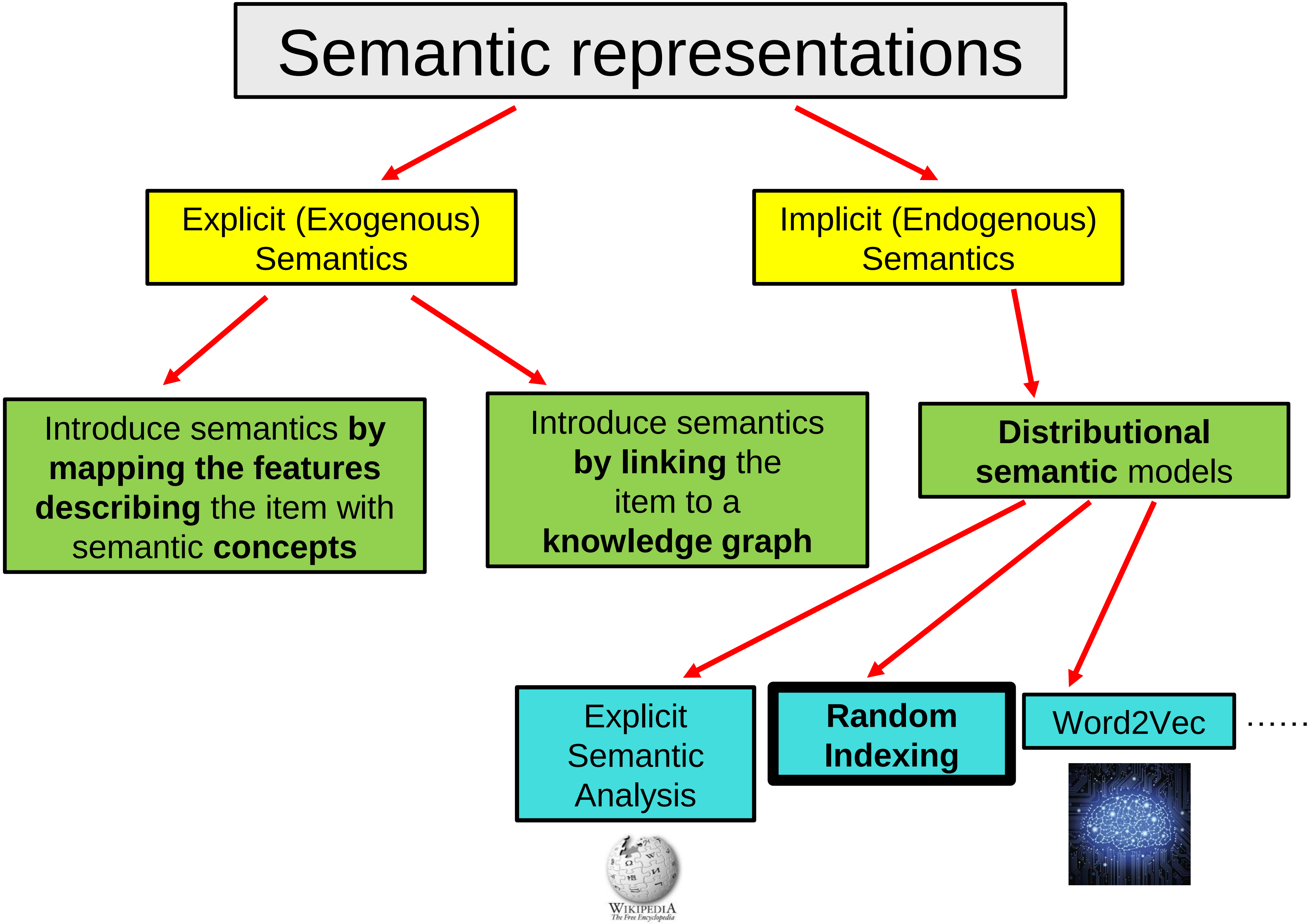
PROBLEM

the huge co-occurrence matrix

SOLUTION

don’t build the huge co-occurrence matrix!

Use incremental and scalable techniques


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking** the item to a **knowledge graph**
Distributional semantic models
Explicit Semantic Analysis
Random Indexing
Word2Vec
……
WIKIPEDIA
The Free Encyclopedia

Dimensionality reduction

Random Indexing

It is an incremental and scalable technique for dimensionality reduction.

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Dimensionality reduction

Random Indexing

It is an incremental and scalable technique for dimensionality reduction.

Insight

- Assign a vector to each context (word, documents, etc.). The vector can be as big as you want.
- Fill the vector with (almost) randomly assigned values.
- Given a word, collect the contexts where that word appears.
- Sum the context and obtain the final representation of the word
- The resulting representation is a smaller but (almost) equivalent to the original one

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Dimensionality reduction

Random Indexing

It is a … hnique

- Assign …). The vector …
- Fill the …
- Given … ppears.
- Sum the context and obtain the final representation of the word
- The resulting representation is a smaller but (almost) equivalent to the original one

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Random Indexing

Algorithm

Step 1 - definition of the context granularity:

Document? Paragraph? Sentence? Word?

Step 2 – building the random matrix R

each **'context'** (e.g. sentence) is assigned a **context vector**

- ✓ dimension = **k**
- ✓ allowed values = **{-1, 0, +1}**
  - values distributed in a **random way**
- ✓ small # of non-zero elements, i.e. **sparse vectors**

Random Indexing

Context vectors of dimension k = 8

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
r_n	…							

Each row is a «context»

Random Indexing

Algorithm

Step 3 –the **vector space representation** of a **term t is** obtained by **combining** the **random vectors** of the **context in which it occurs in**

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
…								
r_n	…							

Random Indexing

Algorithm

Step 3 – building the representation for t1

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							

$t1 \in \{c1, c2, c5\}$

Random Indexing

Algorithm

Step 3 – building the representation for t1

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							

$t1 \in \{c1, c2, c5\}$

r_1	0,	0,	-1,	1,	0,	0,	0,	0	+
r_2	1,	0,	0,	0,	0,	0,	0,	-1	+
r_5	1,	0,	0,	-1,	1,	0,	0,	0	+
t1	2,	0,	-1,	0,	1,	0,	0,	-1	

Random Indexing

Algorithm

Step 3 – building the representation for t1

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							

$t1 \in \{c1, c2, c5\}$

r_1	0,	0,	-1,	1,	0,	0,	0,	0	+
r_2	1,	0,	0,	0,	0,	0,	0,	-1	+
r_5	1,	0,	0,	-1,	1,	0,	0,	0	+
t1	2,	0,	-1,	0,	1,	0,	0,	-1	

Output: **WordSpace**

Random Indexing

Algorithm

Step 4 – building the document space

the **vector space representation** of a **document d** obtained by **combining** the **vector space representation** of the **terms that occur in the document**

Output: **DocSpace**

WordSpace and DocSpace

WordSpace

	c_1	c_2	c_3	c_4	...	c_k
t_1						
t_2						
t_3						
t_4						
...						
t_m						

DocSpace

	c_1	c_2	c_3	c_4	...	c_k
d_1						
d_2						
d_3						
d_4						
...						
d_n						

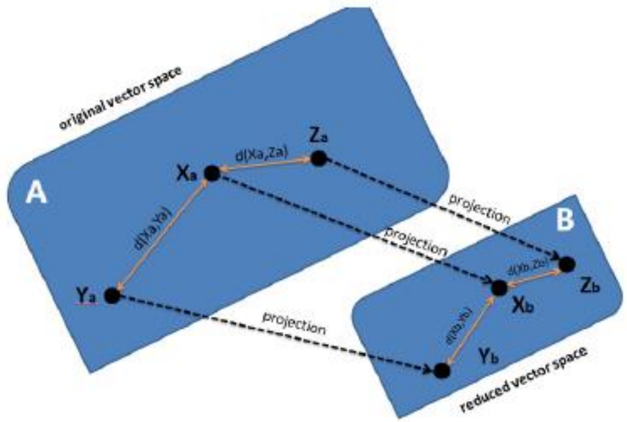
k is a simple parameter of the model

Uniform representation

Dimensionality reduction

..even if it sounds weird

theory: Johnson-Lindenstrauss' lemma [*]



$$B^{m,k} \approx A^{m,n} R^{n,k} \quad k << n$$

distances between the points in the reduced space
approximately preserved if

context vectors are nearly orthogonal

(and they are)

[*] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26(189-206), 1.

Dimensionality reduction

..even if it sounds weird

theory: Johnson-Lindenstrauss' lemma [*]

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
…								
r_n	…							

distanc… space

context vectors are nearly orthogonal

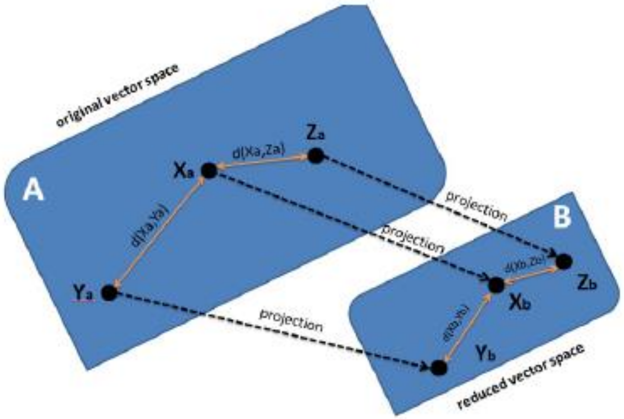
(and they are)

[*] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26(189-206), 1.

Dimensionality reduction

..even if it sounds weird

theory: Johnson-Lindenstrauss' lemma [*]



$B^{m,k} \approx A^{m,n} R^{n,k} \quad k << n$

distances between the points in the reduced space **approximately preserved if**

context vectors are nearly orthogonal

(and they are)

[*] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26(189-206), 1.

Random Indexing

Incremental and Scalable technique for learning word embeddings

Smaller vector space representation

Dimension of the space can be arbitrarily set

Incremental and Scalable

Not transparent anymore

Proper tuning to find the optimal size of the embeddings

Random Indexing @Work: eVSM

- **Enhanced Vector Space Model [*]**
  - **Content-based Recommendation Framework**

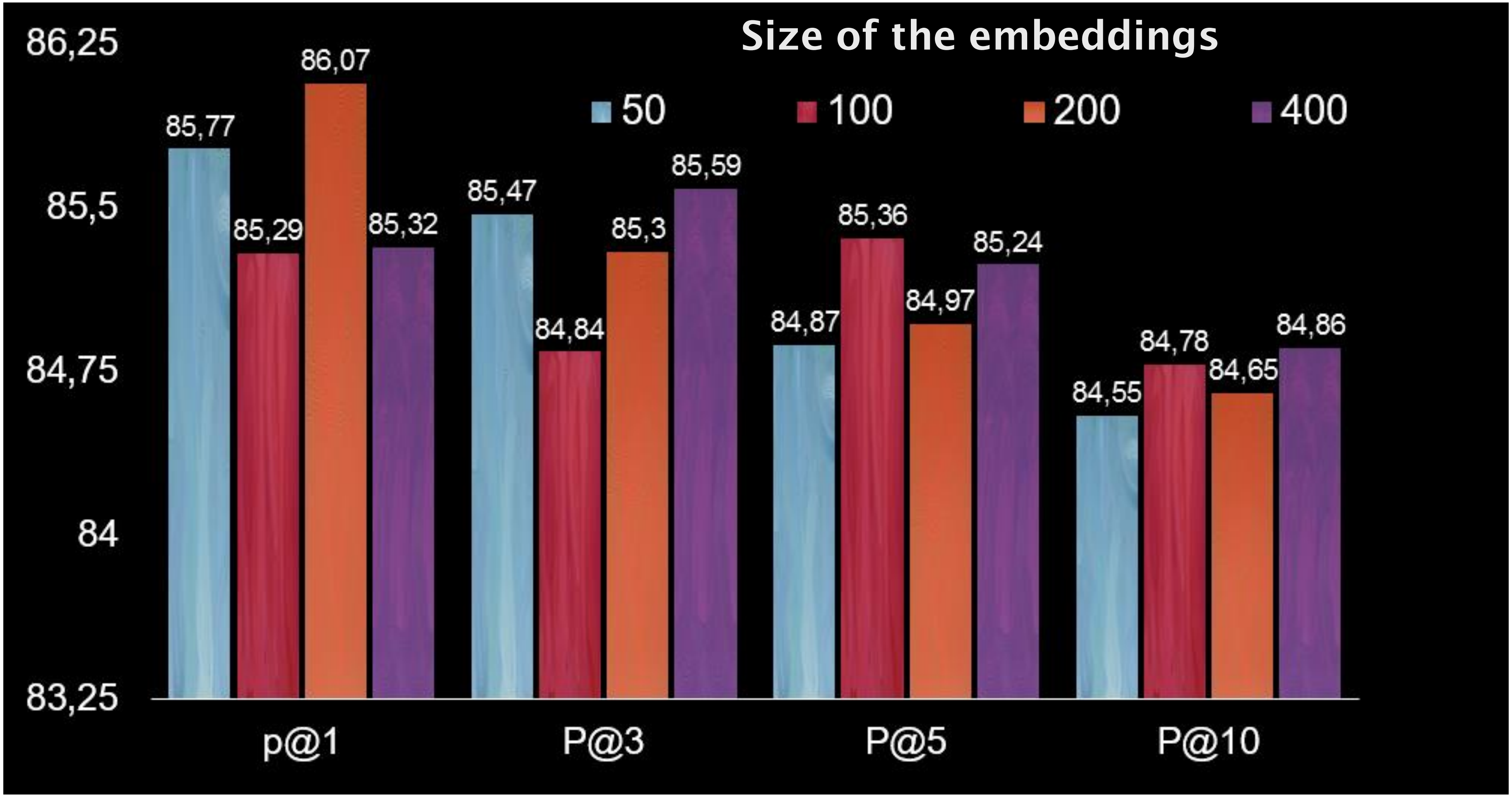
- **Cornerstones**
  - Semantics modeled through **Distributional Models**
  - ⇨ **Random Indexing** for Dimensionality Reduction
  - Negative Preferences modeled through **Quantum Negation** **[^]**
  - **User Profiles** as centroid vectors of items representation
  - Recommendations **through Cosine Similarity**

[*] Musto, Cataldo. **"Enhanced vector space models for content-based recommender systems."** *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

[^] Widdows, Dominic, and Stanley Peters. **"Word vectors and quantum logic: Experiments with negation and disjunction."** *Mathematics of language* 8.141-154 (2003).

eVSM

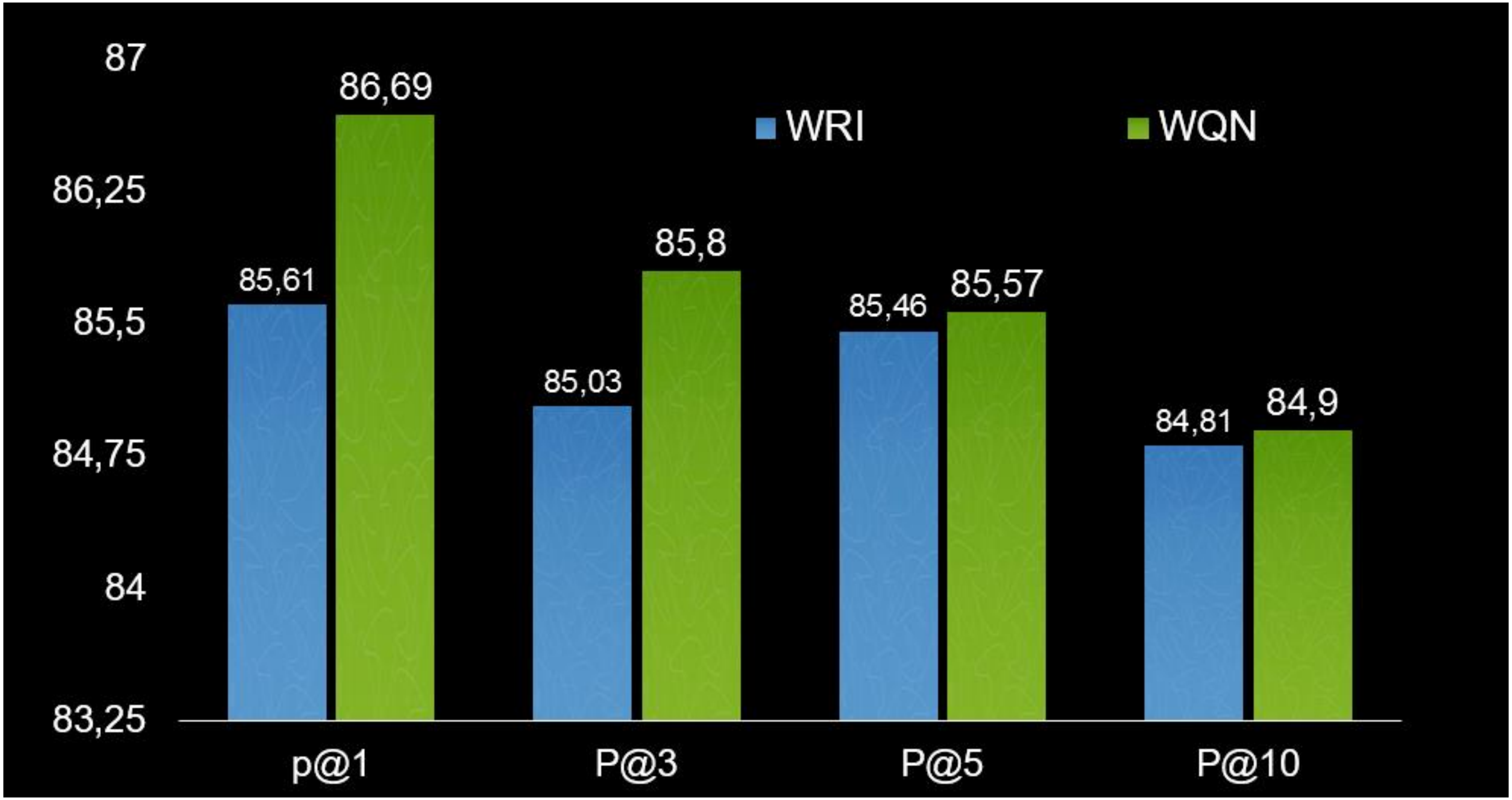
Experiments



The **size** of the embeddings **does not significantly affect** the overall accuracy of eVsm **(MovieLens data)**

eVSM

Experiments

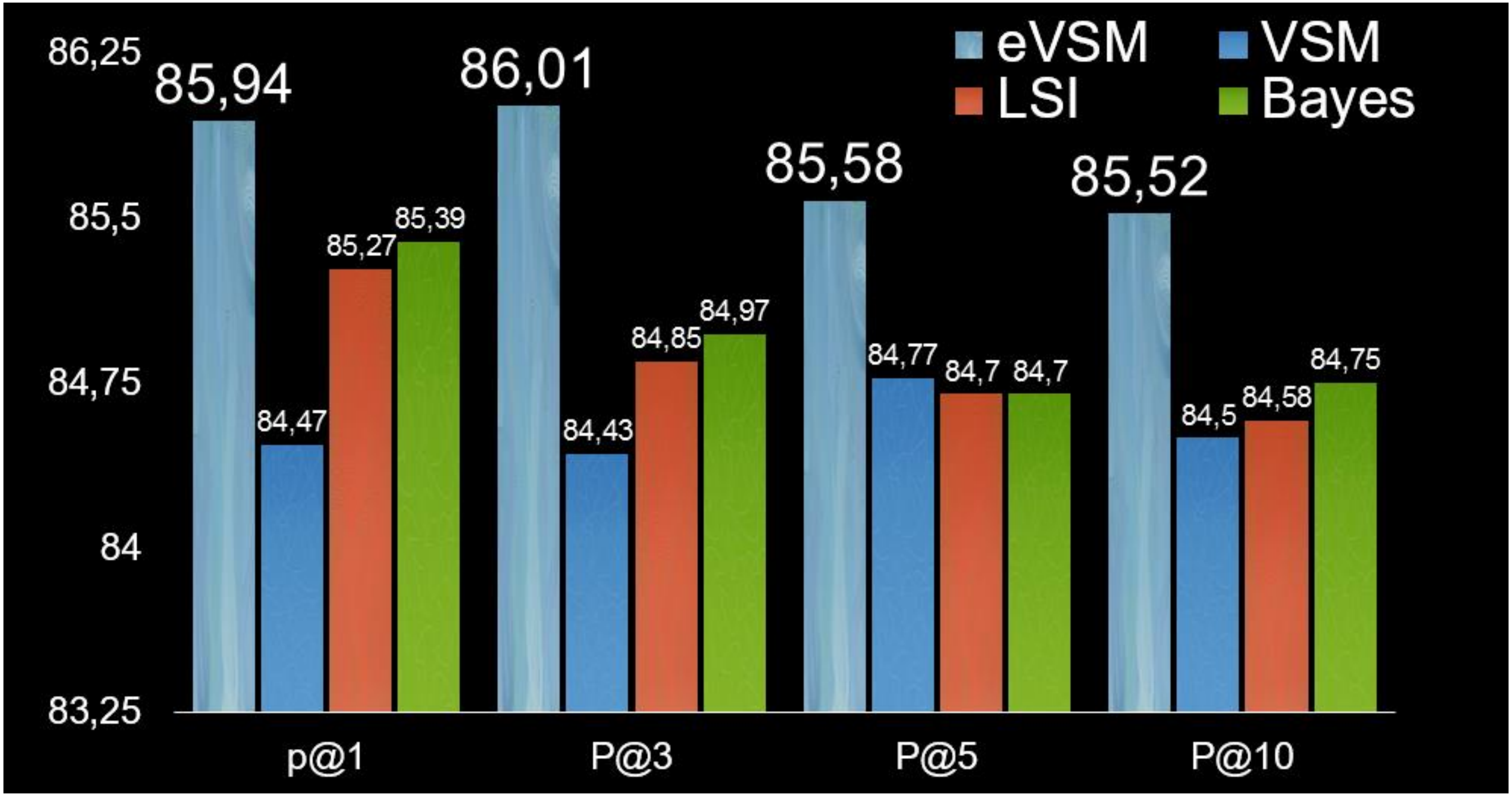


Quantum Negation improves the accuracy of the model

(MovieLens data, embedding size=100)

eVSM

Experiments



eVSM significantly overcame all the baselines.

(MovieLens data, embedding size=400)

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking** the item to a **knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

WIKIPEDIA
The Free Encyclopedia

Random Indexing

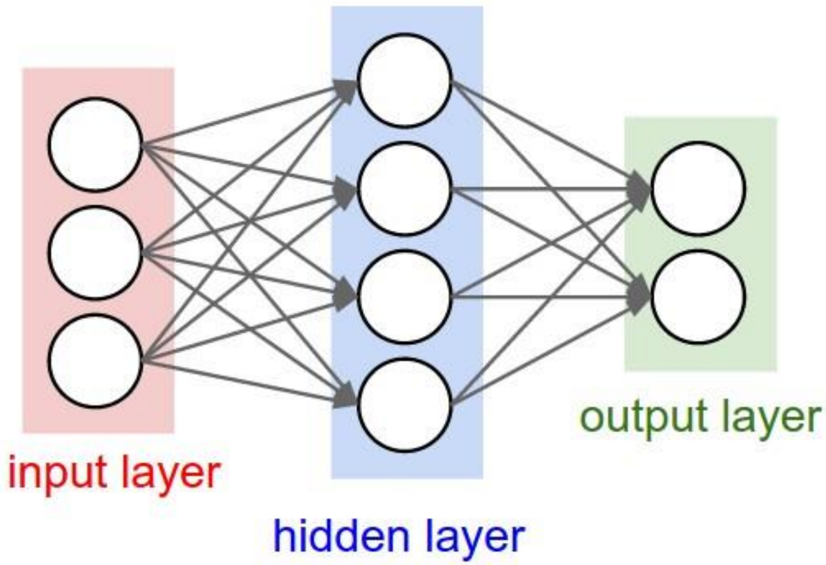
Word2Vec

……

Word2Vec

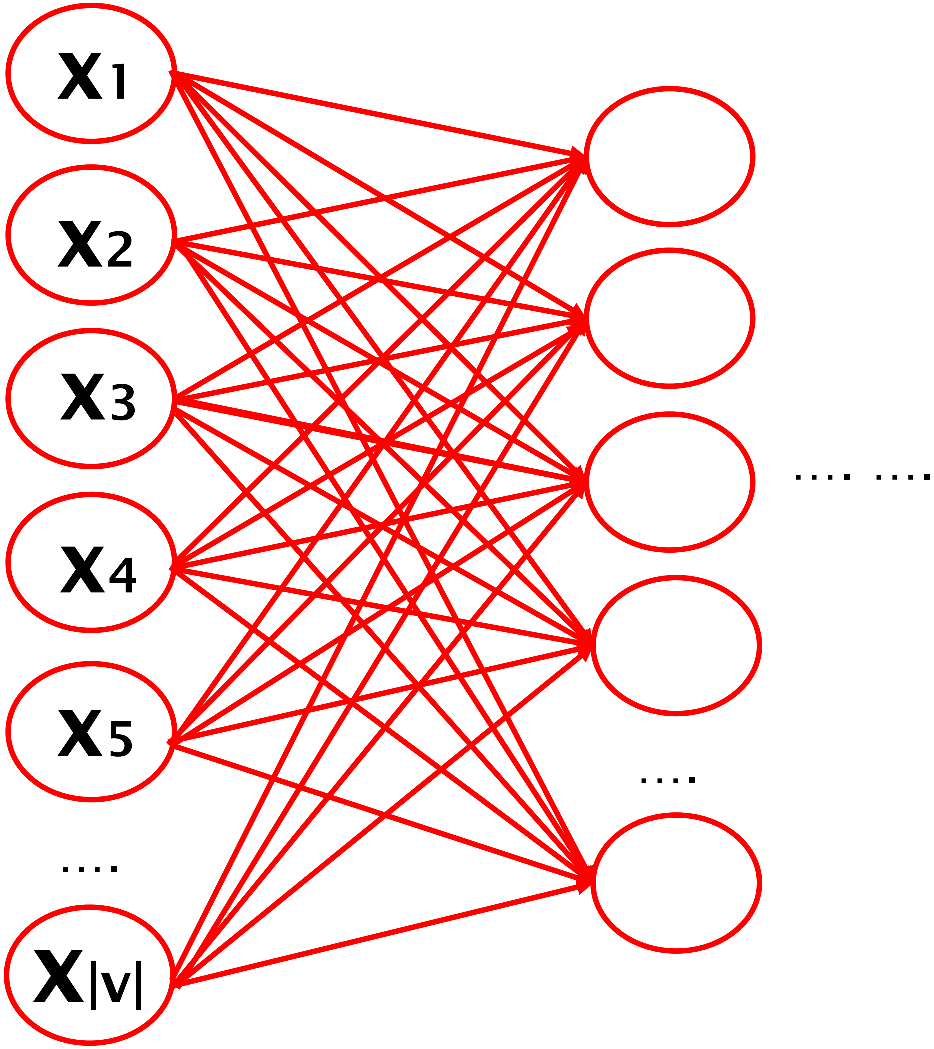
In a nutshell

- **Distributional Model** to learn Word Embeddings.
- Uses a **two-layers neural network**
- Training based on the **Skip-Gram methodology**
- **Update of the network** through Mini-batch and Stochastic Gradient Descent



Word2Vec

(Partial) Structure of the network

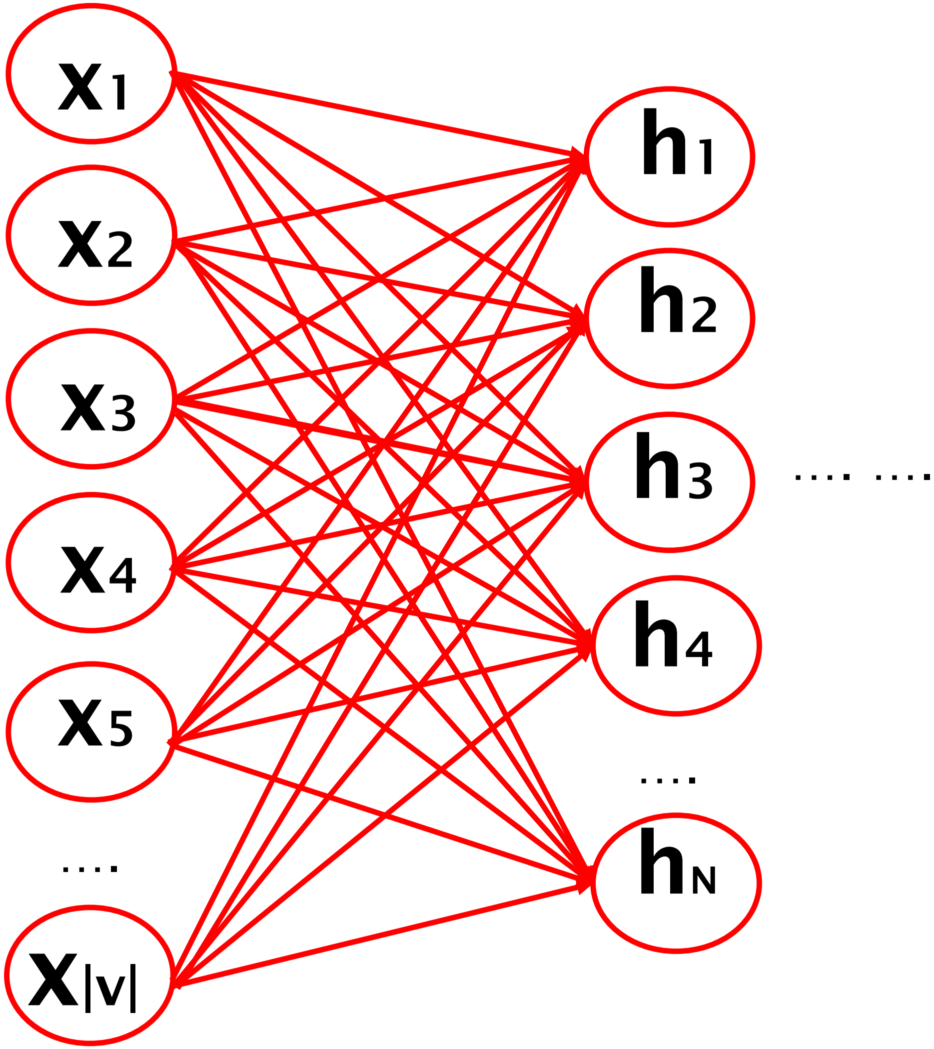


Input Layer:

- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **Each term is represented through a «one hot representation»**

Word2Vec

(Partial) Structure of the network



Input Layer:

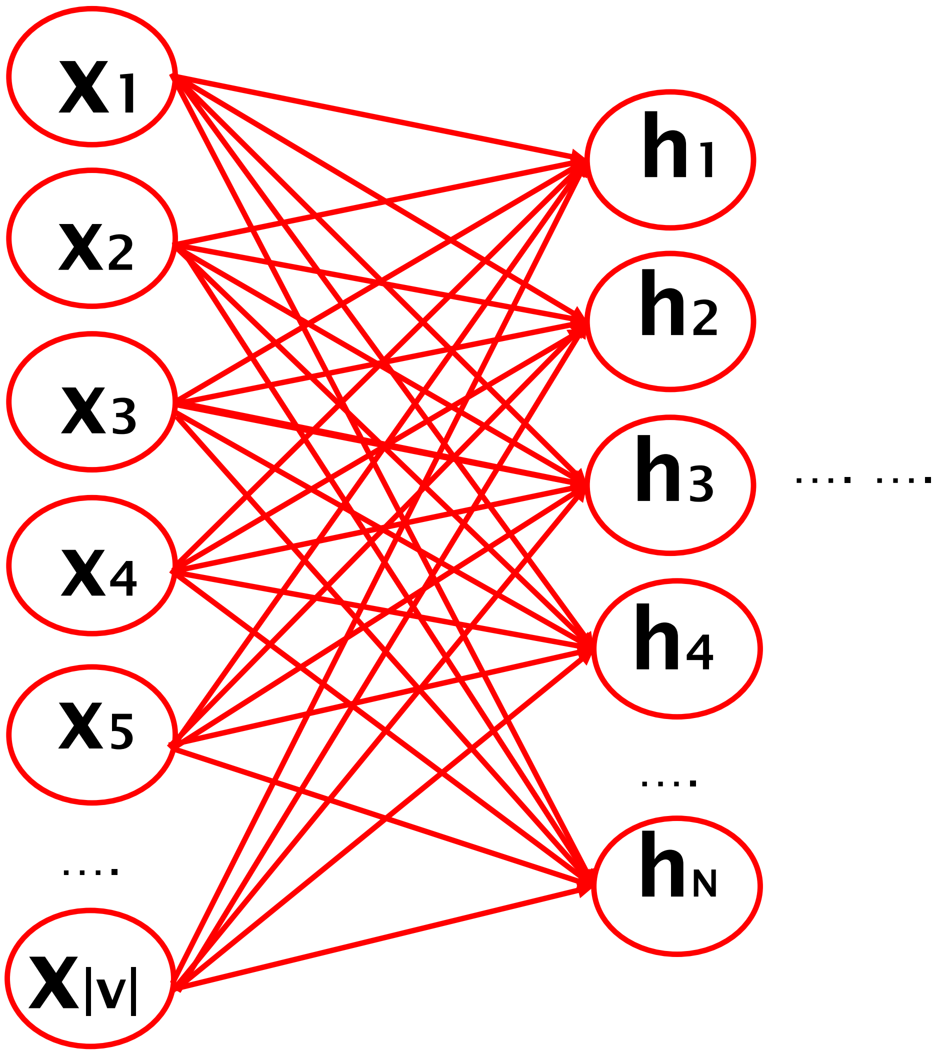
- Vocabulary V
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:

- N nodes
  - N = size of the embeddings
  - Parameter of the model

Word2Vec

(Partial) Structure of the network



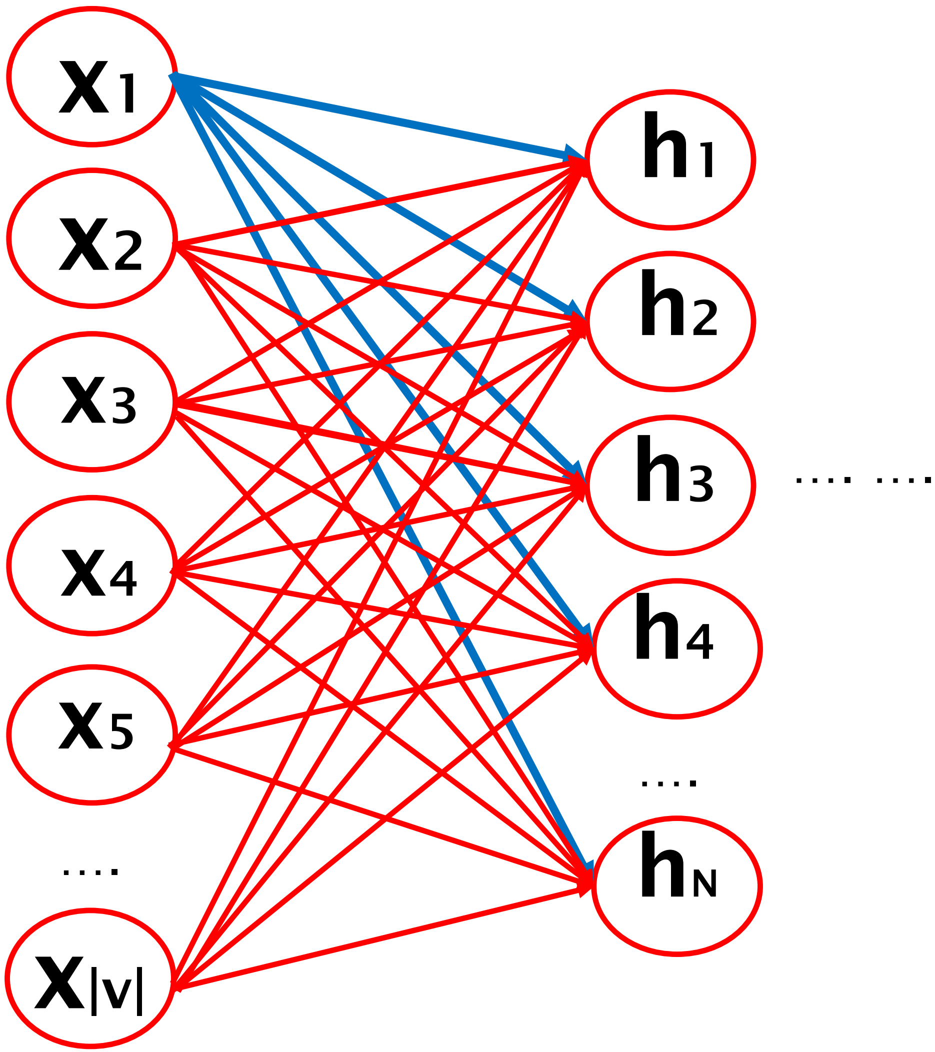
Input Layer:
- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:
- **N nodes**
  - N = size of the embeddings
  - Parameter of the model

Weight of the network:
- **Randomly set (initially)**
- **Updated through the training**

Word2Vec

(Partial) Structure of the network



Input Layer:

- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:

- **N nodes**
  - N = size of the embeddings
  - Parameter of the model

Weight of the network:

- **Randomly set (initially)**
- **Updated through the training**

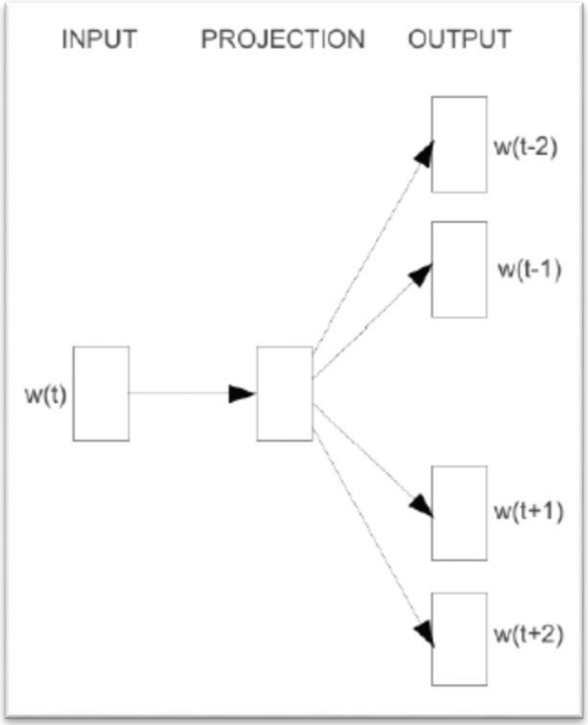
Final Representation for term t_k

- Weights Extracted from the network
- $t_k = [wt_{kv_1}, wt_{kv_2} \ldots wt_{kv_n}]$

Word2Vec

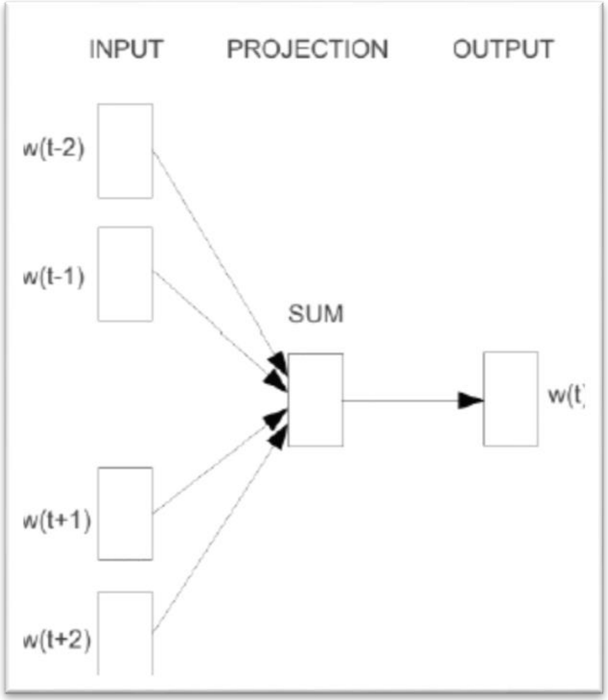
Training Procedure: how to create training examples?

Skip-Gram Methodology



Given **a word w(t),** predict its **context w(t-2), t(t-1).. w(t+1), w(t+2)**

Continuous Bag-of-Words Methodology

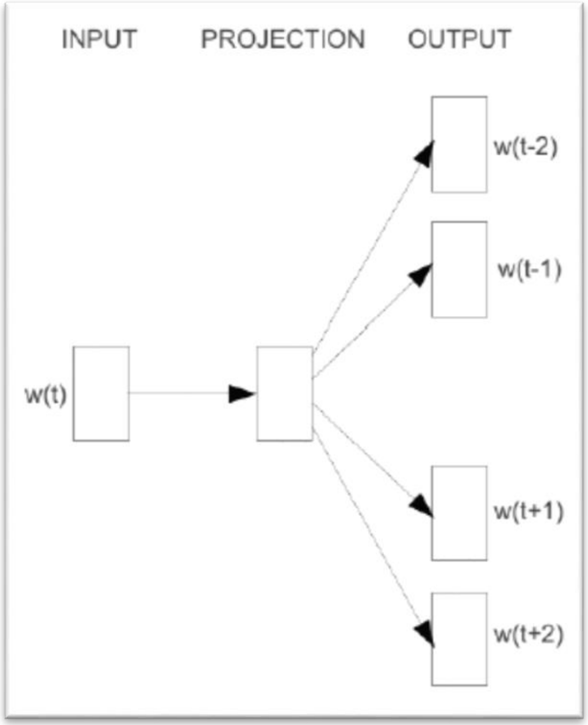


Given **a context w(t-2), t(t-1).. w(t+1), w(t+2)** predict **word w(t)**

Word2Vec

Training Procedure: how to create training examples?

Skip-Gram Methodology



Given **a word w(t),** predict its **context w(t-2), t(t-1).. w(t+1), w(t+2)**

Example

Input: ”the quick brown fox jumped over the lazy dog”

Window Size: 1

Contexts:

- ([the, brown], quick)
- ([quick, fox], brown)
- ([brown, jumped], fox) ...

Training Examples:

- (quick, the)
- (quick, brown)
- (brown, quick)
- (brown, fox) ...

Word2Vec

Training Procedure: how to optimize the model?

Given a corpus, we create **of training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context 'c' given a word 'w'

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

Word2Vec

Training Procedure: how to optimize the model?

Given a corpus, we create **of training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context 'c' given a word 'w'

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Word2Vec

Training Procedure: how to optimize the model?

Given a corpus, we create a **training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context C given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Intuitively, probability is high when scalar product is close to 1 → when **vectors are similar!**

Word2Vec

Training Procedure: how to optimize the model?

Given a corpus, we create a **training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context C given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Intuitively, probability is high when scalar product is close to 1 → when **vectors are similar!**

Word2Vec is a distributional model since it learns a representation such that couples (word,context) appearing together have similar vectors

Word2Vec

Training Procedure: how to optimize the model?

Given a corpus, we create a **training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context C given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Intuitively, probability is high when scalar product is close to 1 → when **vectors are similar!**

The error is collected and weights in the network are updated accordingly. **Typically is used Stochastic Gradient Descent or Mini-Batch (every 128 or 512 training examples)**

Word2Vec

Learning Word Embeddings through Neural Networks: it is not based on «counting» co-occurrences. It relies on **«predicting» the distribution**

Representation can be really really small (size≤100, typically)

Trending ☺ - Recent and Very Hot technique

Not transparent anymore

Needs more computational resources

Word2Vec

- ✓ Empirical Comparison of **Word Embedding Techniques** for **Content-based Recommender Systems [*]**

- ✓ **Methodology**
  - ✓ **Build a WordSpace using different Word Embedding techniques (and different sizes)**
  - ✓ Build a DocSpace as the centroid vectors of term vectors
  - ✓ **Build User Profiles as centroid** of the items they liked
  - ✓ Provide **Users with Recommendations**
  - ✓ **Compare the approaches**

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval.* ECIR 2016.

Word2Vec

Results on DBbook and MovieLens data

MovieLens	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,4645	0,4715	0,4921	0,4910	**0,5056**	0,5054	**0,5217**	*0,5022*	0,5141
F1@10	0,5393	0,5469	0,5622	0,5613	**0,5757**	0,5751	**0,5969**	0,5836	0,5928
F1@15	0,5187	0,5254	0,5349	0,5352	0,5672	**0,5674**	**0,5911**	0,5814	0,5876

Word Embedding overcomes **I2I-CF only on F1@5**. Needs to further process content **on less sparse datasets.**

DBbook	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,5056	0,5076	0,5064	0,5039	0,5183	**0,5186**	0,5193	*0,5111*	**0,5290**
F1@10	0,6256	**0,6260**	0,6239	0,6244	0,6207	0,6209	0,6229	*0,6194*	**0,6263**
F1@15	0,5908	**0,5909**	0,5892	0,5887	0,5829	0,5828	*0,5777*	*0,5776*	*0,5778*

Results comparable to CF and MF on more sparse datasets. LSI is the best-performing approach on F1@15

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval.* ECIR 2016.

Word2Vec

Results on DBbook and MovieLens data

MovieLens	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,4645	0,4715	0,4921	0,4910	**0,5056**	0,5054	**0,5217**	*0,5022*	0,5141
F1@10	0,5393	0,5469	0,5622	0,5613	**0,5757**	0,5751	**0,5969**	0,5836	0,5928
F1@15	0,5187	0,5254	0,5349	0,5352	0,5672	**0,5674**	**0,5911**	0,5814	0,5876

Word Embedding overcomes **I2I-CF only on F1@5**. Needs to further process content **on less sparse datasets.**

DBbook	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,5056	0,5076	0,5064	0,5039	0,5183	**0,5186**	0,5193	*0,5111*	**0,5290**
F1@10	0,6256	**0,6260**	0,6239	0,6244	0,6207	0,6209	0,6229	*0,6194*	**0,6263**
F1@15	0,5908	**0,5909**	0,5892	0,5887	0,5829	0,5828	*0,5777*	*0,5776*	*0,5778*

Results comparable to CF and MF on more sparse datasets. LSI is the best-performing approach on F1@10

Word Embedding techniques

Baselines

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval.* ECIR 2016.

Word2Vec

Results on DBbook and MovieLens data

MovieLens	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,4645	0,4715	0,4921	0,4910	**0,5056**	0,5054	**0,5217**	*0,5022*	0,5141
F1@10	0,5393	0,5469	0,5622	0,5613	**0,5757**	0,5751	**0,5969**	0,5836	0,5928
F1@15	0,5187	0,5254	0,5349	0,5352	0,5672	**0,5674**	**0,5911**	0,5814	0,5876

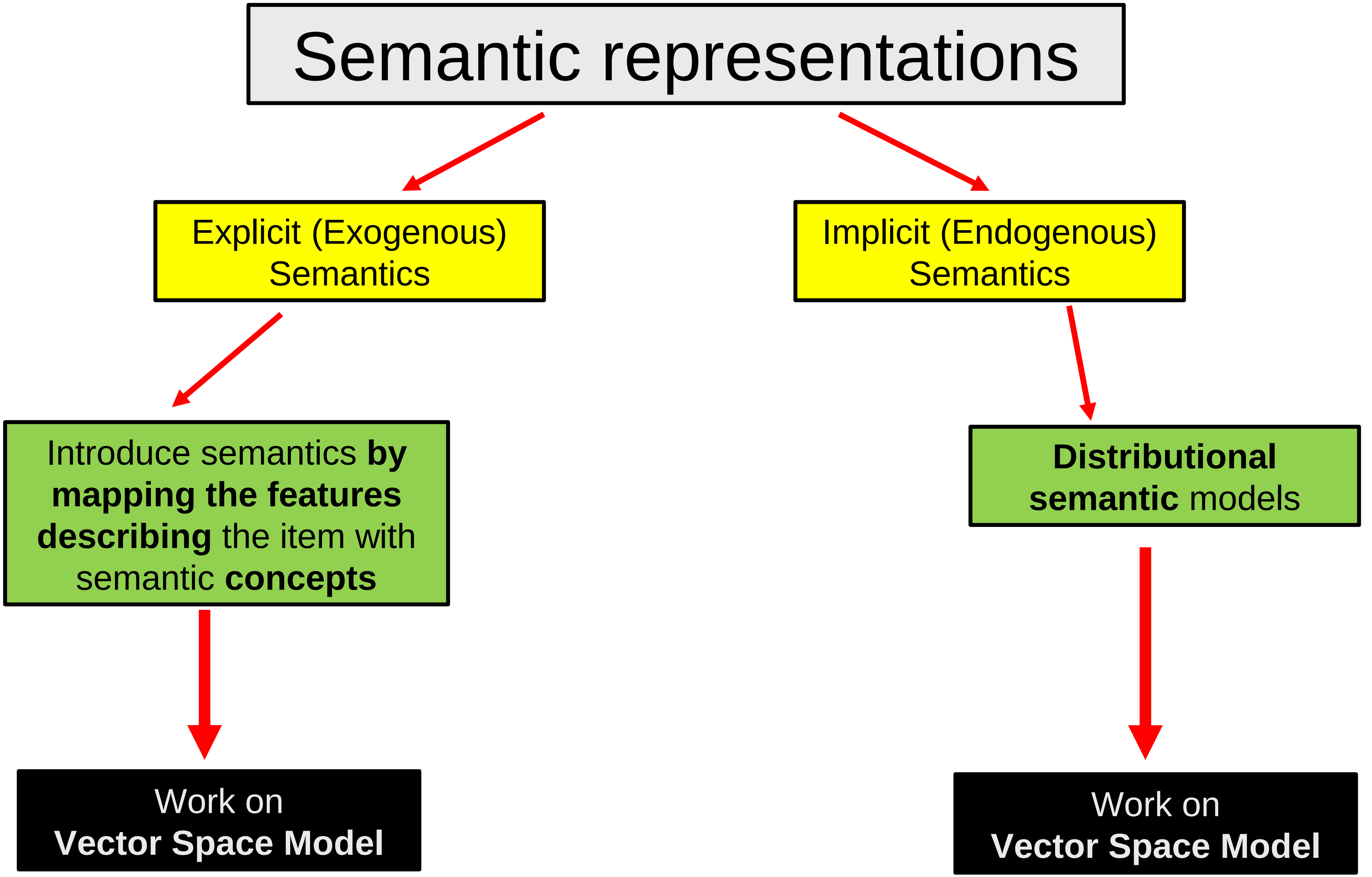
Word Embedding overcomes **I2I-CF only on F1@5**. Needs to further process content **on less sparse datasets.**

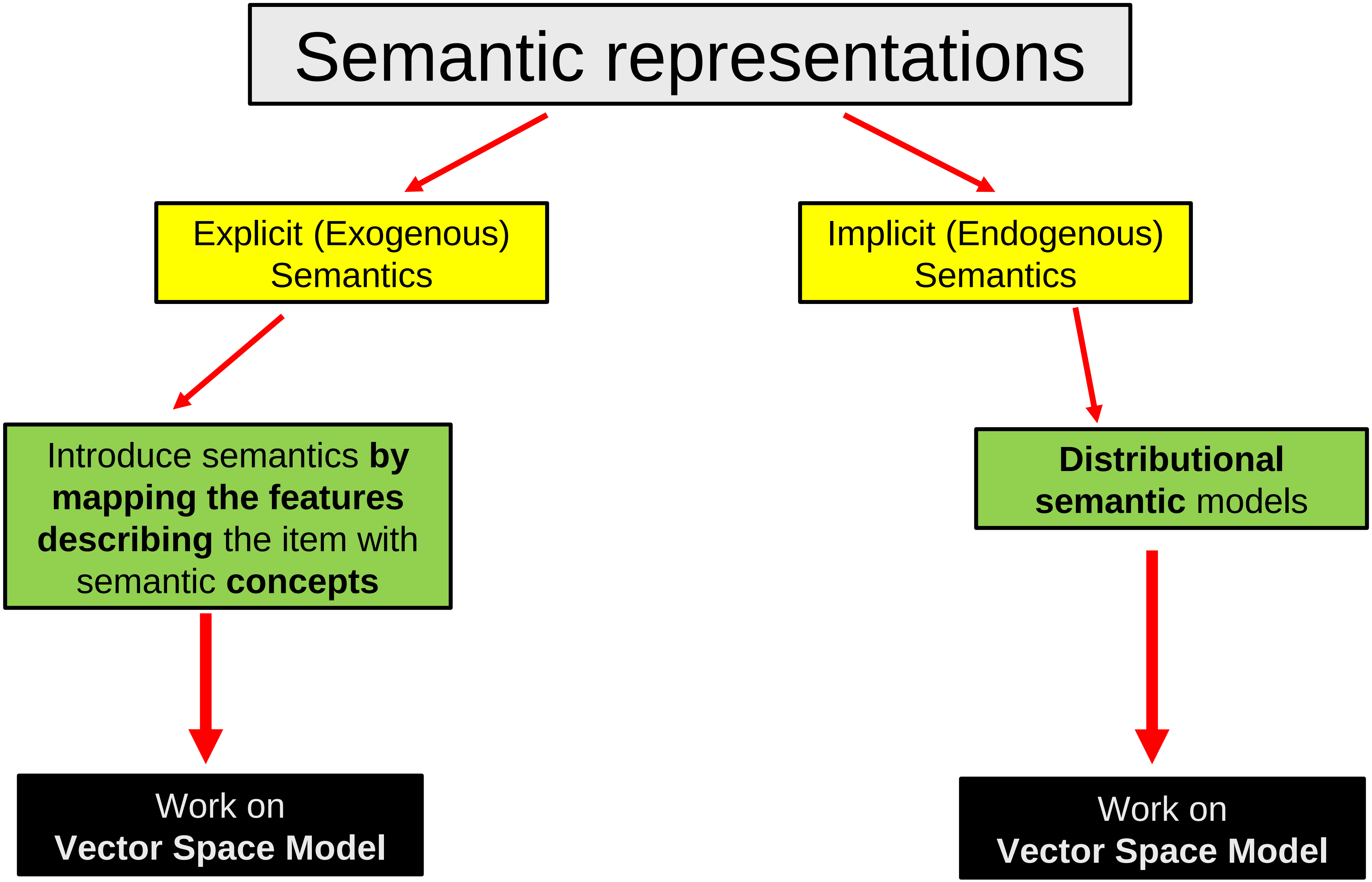
DBbook	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,5056	0,5076	0,5064	0,5039	0,5183	**0,5186**	0,5193	*0,5111*	**0,5290**
F1@10	0,6256	**0,6260**	0,6239	0,6244	0,6207	0,6209	0,6229	*0,6194*	**0,6263**
F1@15	0,5908	**0,5909**	0,5892	0,5887	0,5829	0,5828	*0,5777*	*0,5776*	*0,5778*

Results comparable to CF and MF on more sparse datasets. LSI is the best-performing approach on F1@15

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval.* ECIR 2016.

...Let’s put everything together


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics by mapping the features describing the item with semantic concepts
Distributional semantic models
Work on Vector Space Model
Work on Vector Space Model



Can Exogenous and Endogenous approaches be combined?

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

Endogenous Approaches as ESA or Word2Vec
work on the columns of the matrix

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

Endogenous Approaches as ESA or Word2Vec
work on the columns of the matrix

Both approaches can be combined to obtain richer and more precise semantic representations
(e.g. Word2Vec over textual description processed with WSD)

Recent Developments of Content-Based RecSys

Exogenous techniques:

RecSys based on Linked Open Data

Cataldo Musto

Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

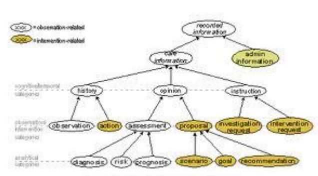
SWAP researchgroup

Semantic representations

Explicit (Exogenous) Semantics

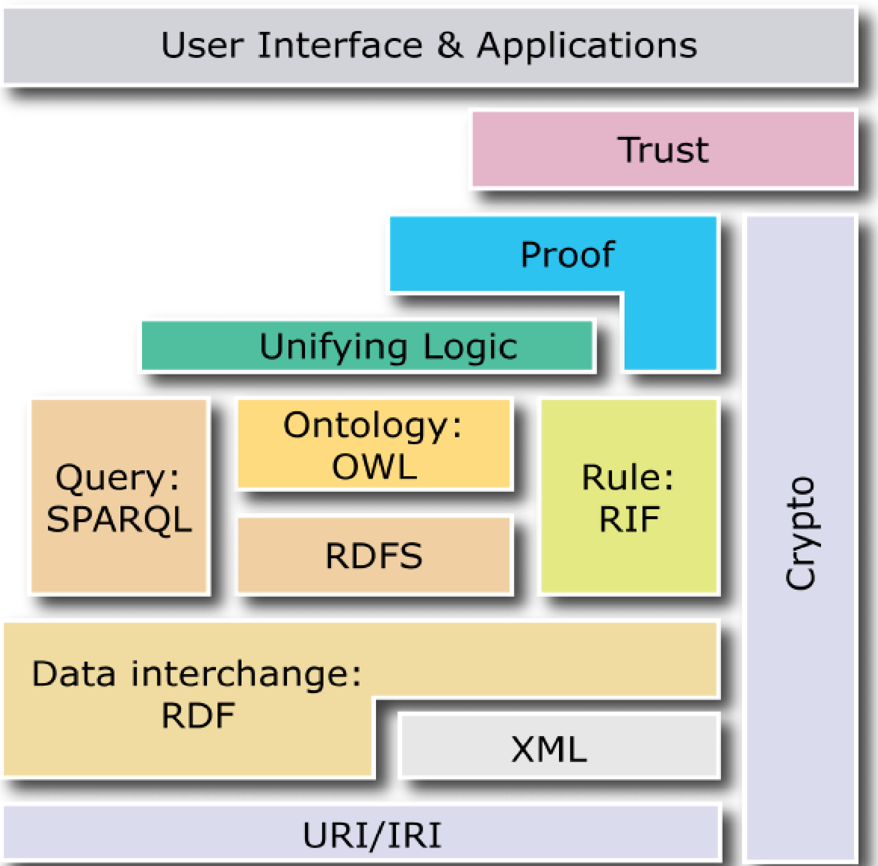
- Introduce semantics **by mapping the features describing** the item with semantic **concepts**
- Introduce semantics **by linking** the item to a **knowledge graph**
  - Ontologies
  - Linked Open Data
  - ....

Implicit (Endogenous) Semantics



DBpedia

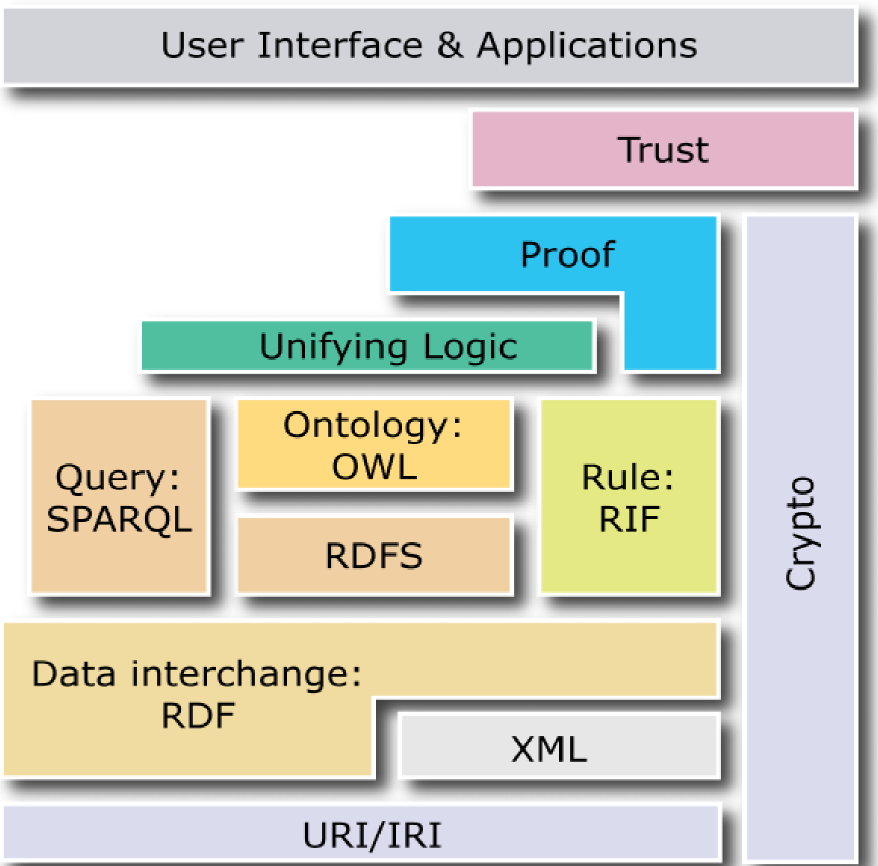
Semantic Web



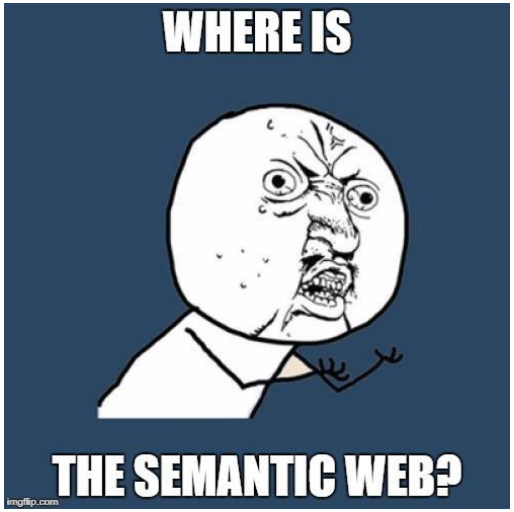
"The Semantic Web provides a common framework that allows data to be shared and reused across application enterprise, and community boundaries" ***[*]***

[*] Berners-Lee, Tim; James Hendler; Ora Lassila "The Semantic Web". Scientific American Magazine, 2001

Semantic Web



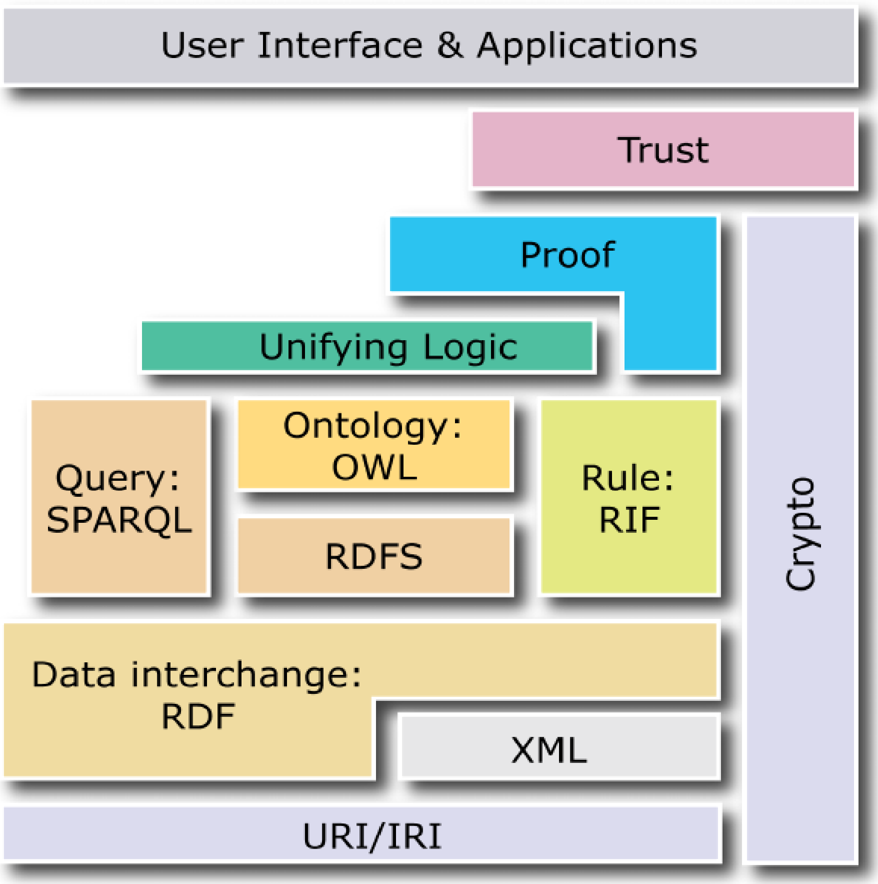
"The Semantic Web provides a common framework that allows data to be shared and reused across application enterprise, and community boundaries" ***[*]***



(Do we succed?)

[*] Berners-Lee, Tim; James Hendler; Ora Lassila "The Semantic Web". Scientific American Magazine, 2001

From Semantic Web to Linked Open Data



"The Semantic Web provides a common framework that allows data to be shared and reused across application enterprise, and community boundaries" ***[*]***

Linked Open Data Project

Goal: *to make* ***structured*** *and* ***interconnected*** *the* ***whole DATA*** *available on the Web* ***[^]***.

[*] Berners-Lee, Tim; James Hendler; Ora Lassila "The Semantic Web". Scientific American Magazine, 2001

Linked Open Data

What is it?

Linked Open Data

What is it?

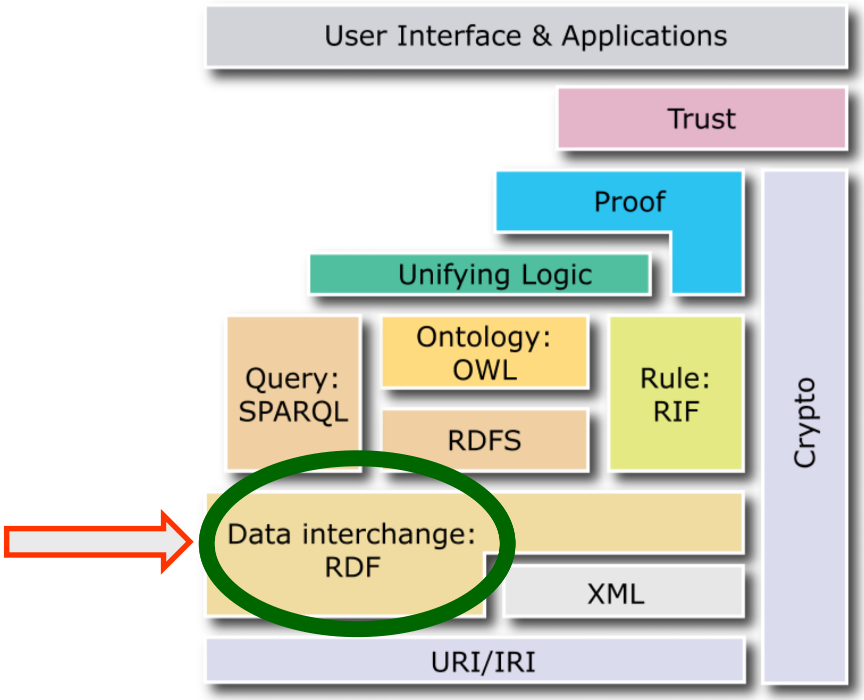
Linked Open Data is a **methodology** to publish, share and link **structured data** on the Web

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



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

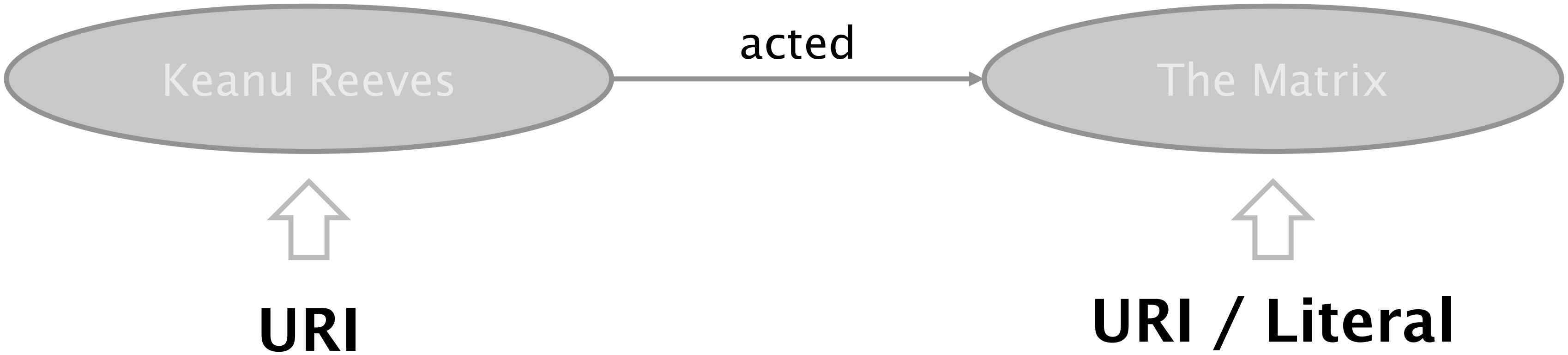
1. Use of RDF to model the information and make data publicly available

(this is called *RDF triple*)

Linked Open Data - Cornerstones

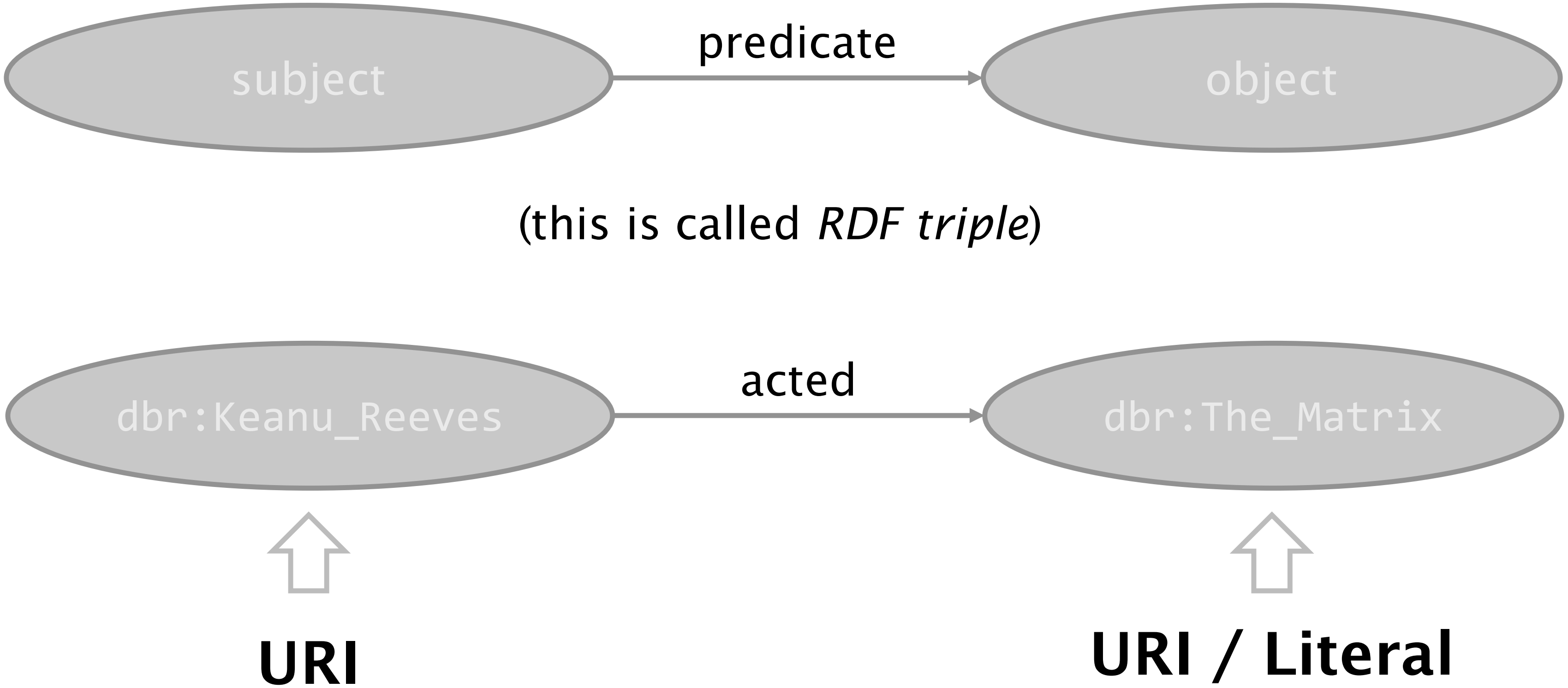
1. Use of RDF to model the information and make data publicly available

(this is called *RDF triple*)



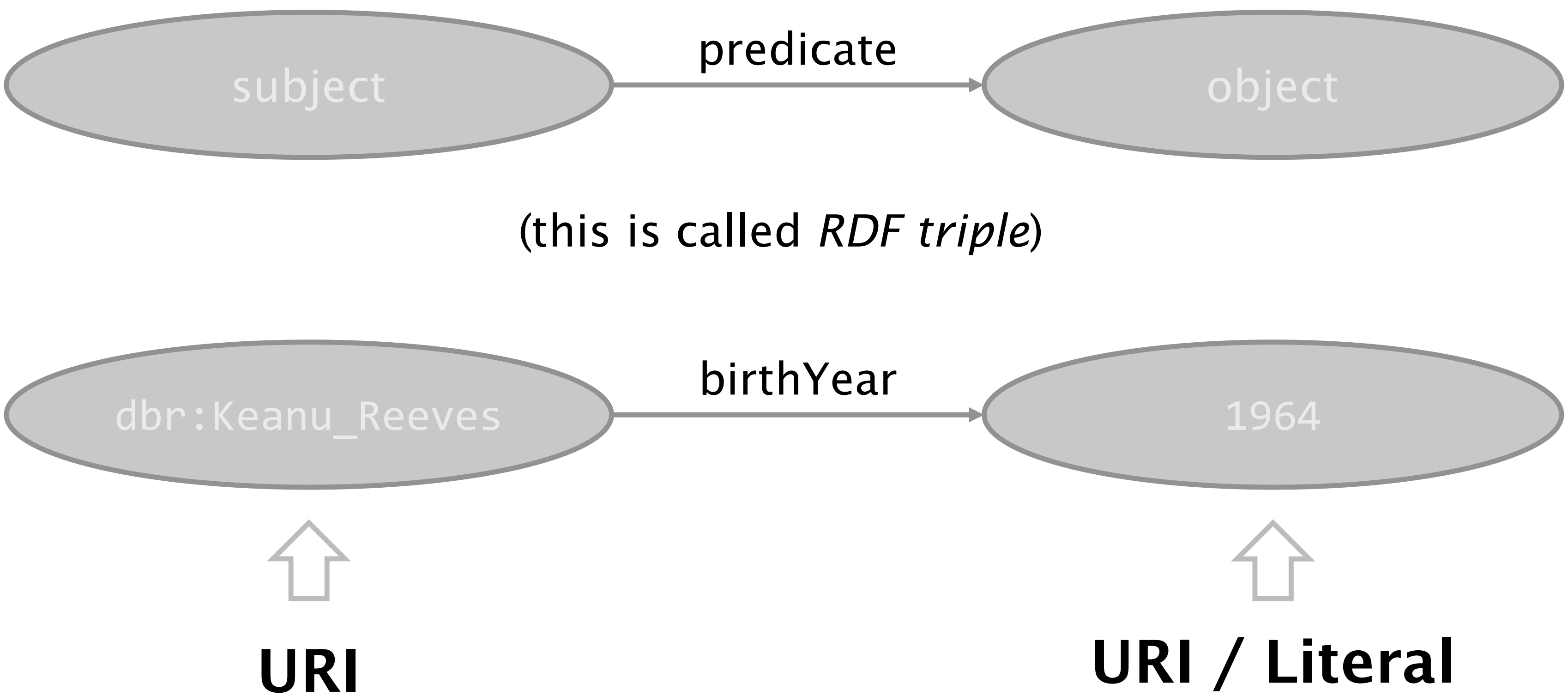
Linked Open Data - Cornerstones

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

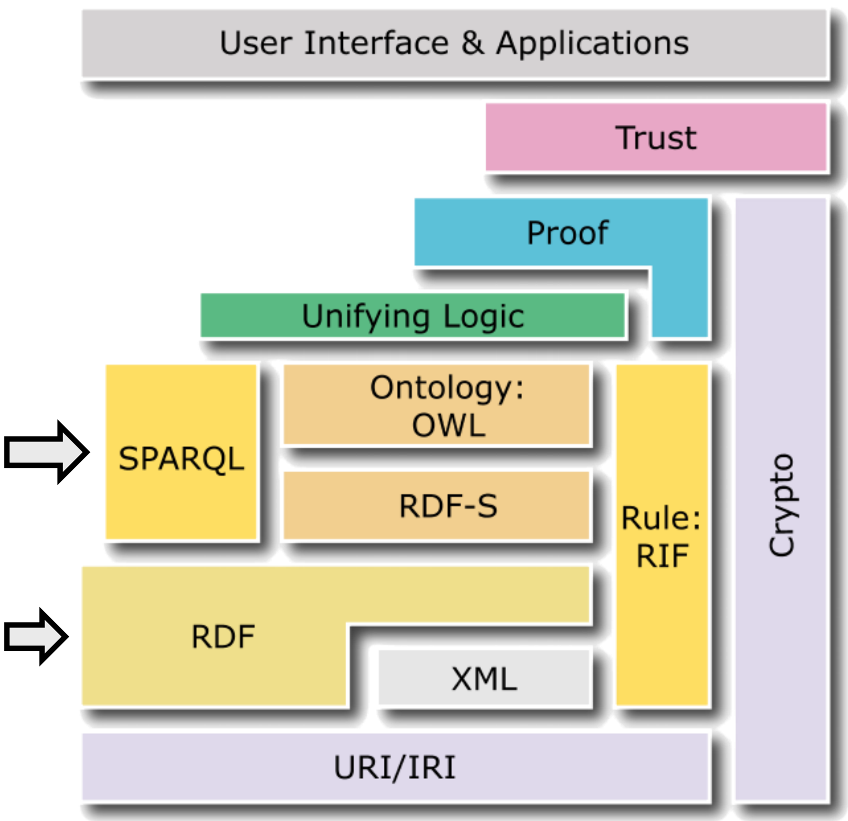


Linked Open Data - Cornerstones

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

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

Linked Open Data



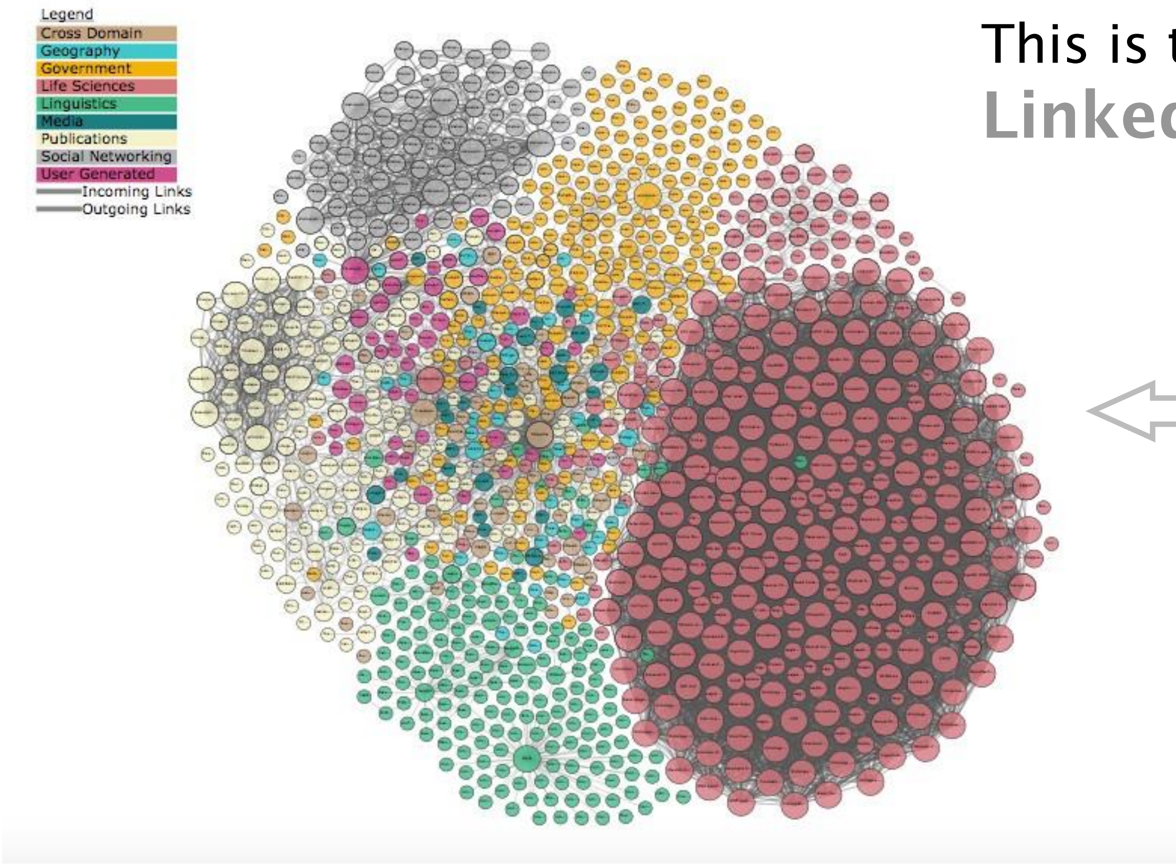
We only use a small subset of the 'Semantic web cake'

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

Linked Open Data

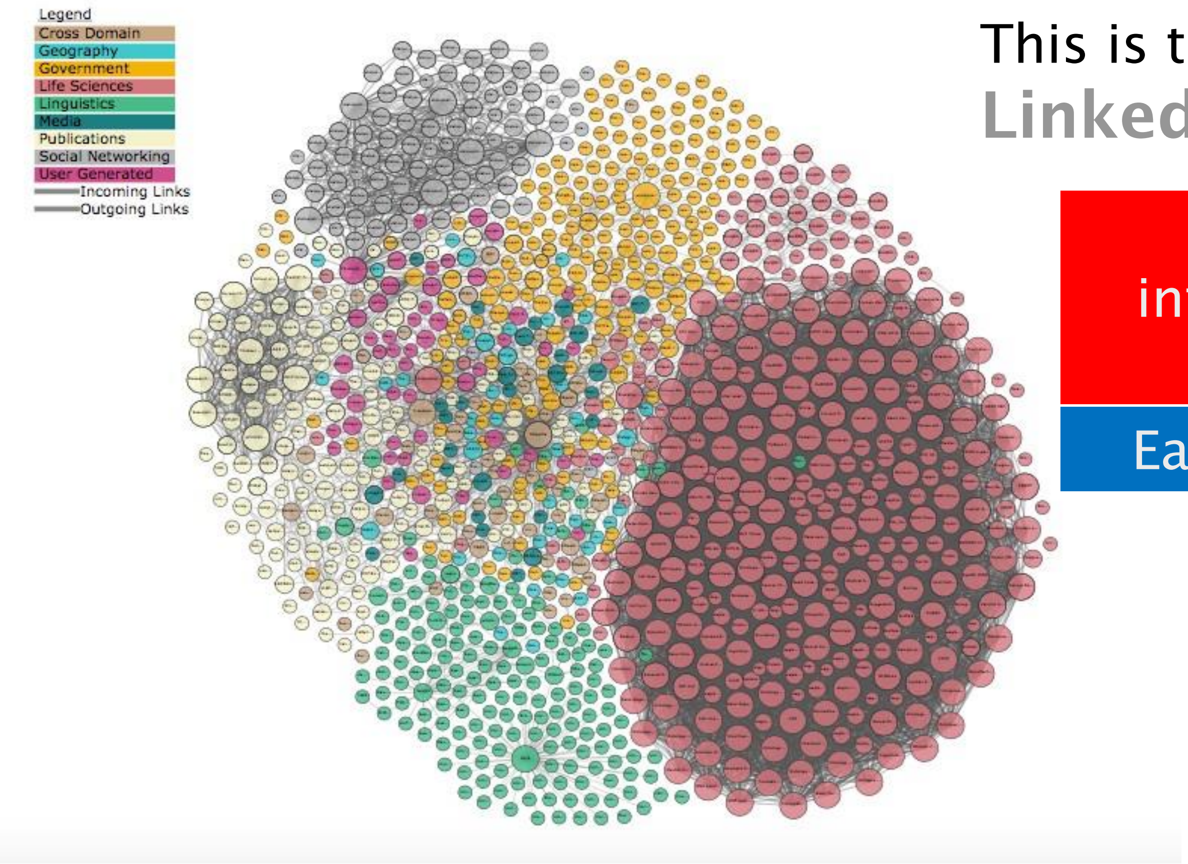
Do we succeed?

Linked Open Data



This is the
Linked Open Data cloud

Linked Open Data



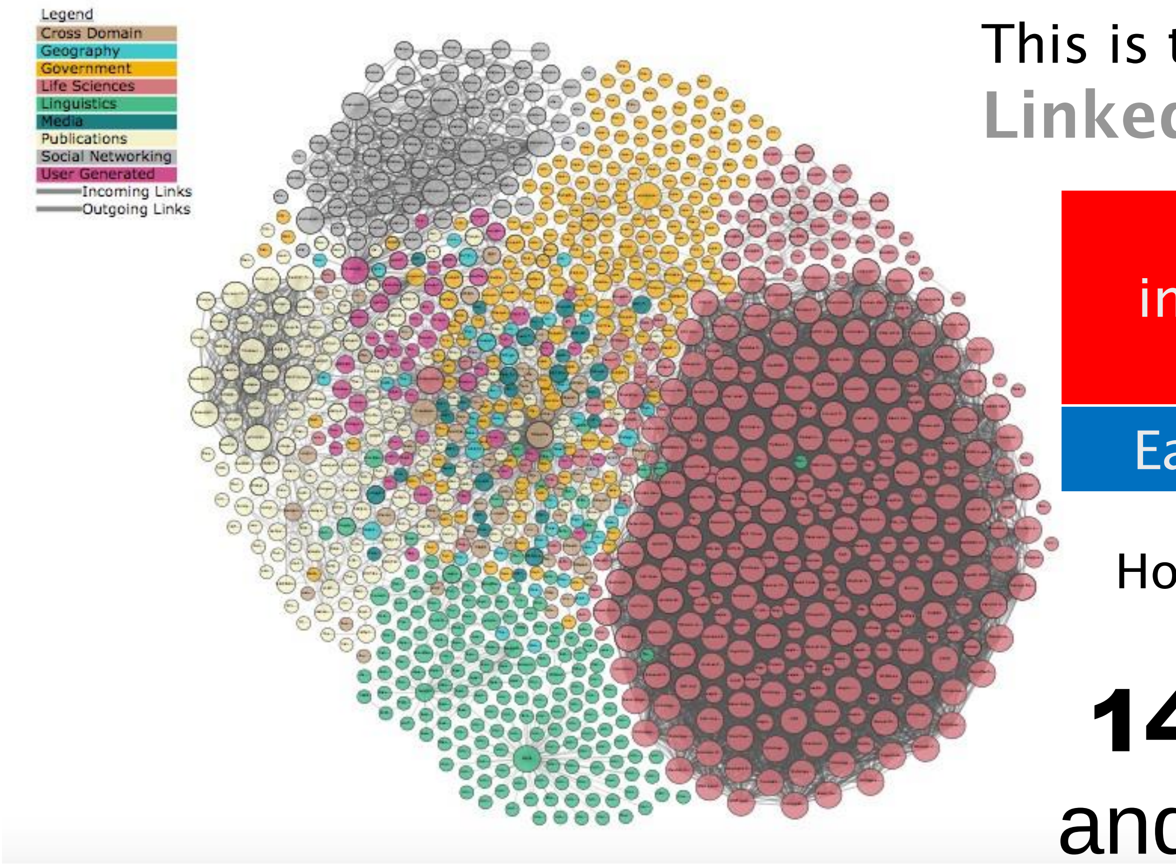
This is the

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

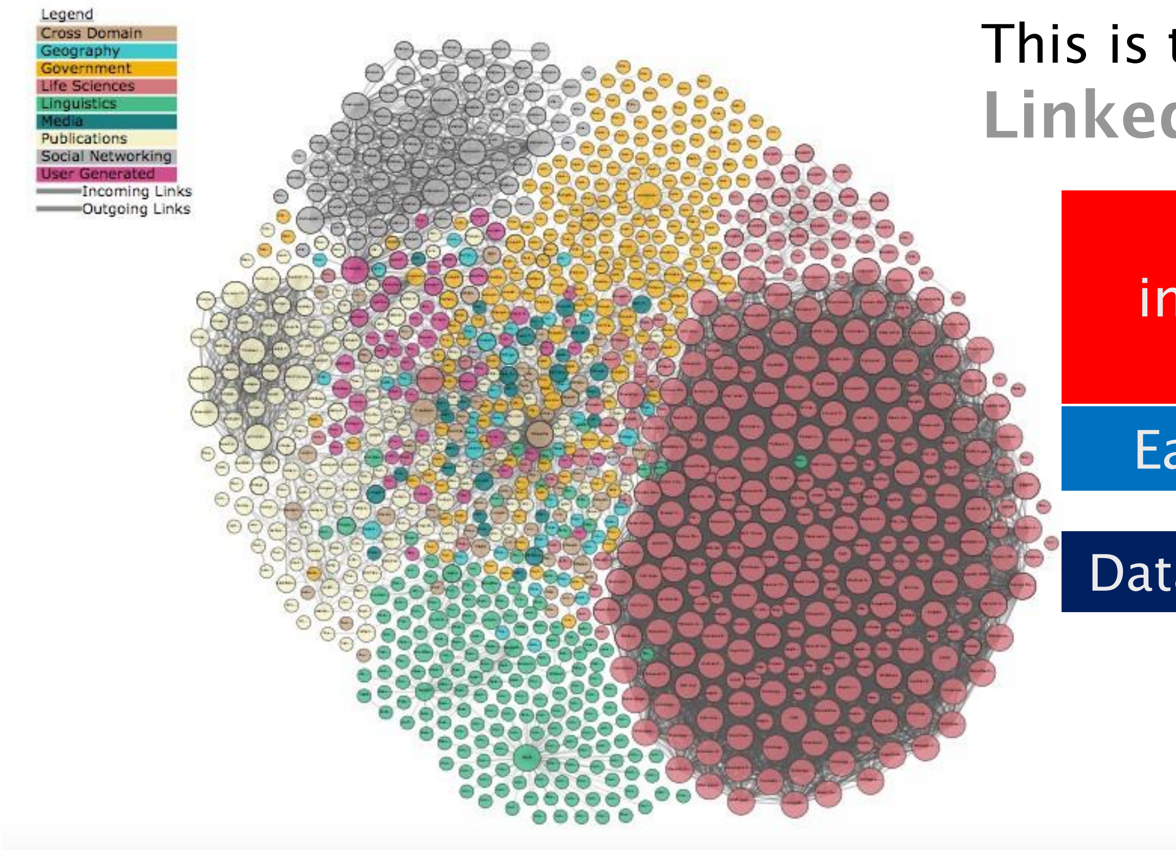
Each bubble is a dataset!

How many datasets do we have?

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

(source: http://stats.lod2.eu)

Linked Open Data



This is the

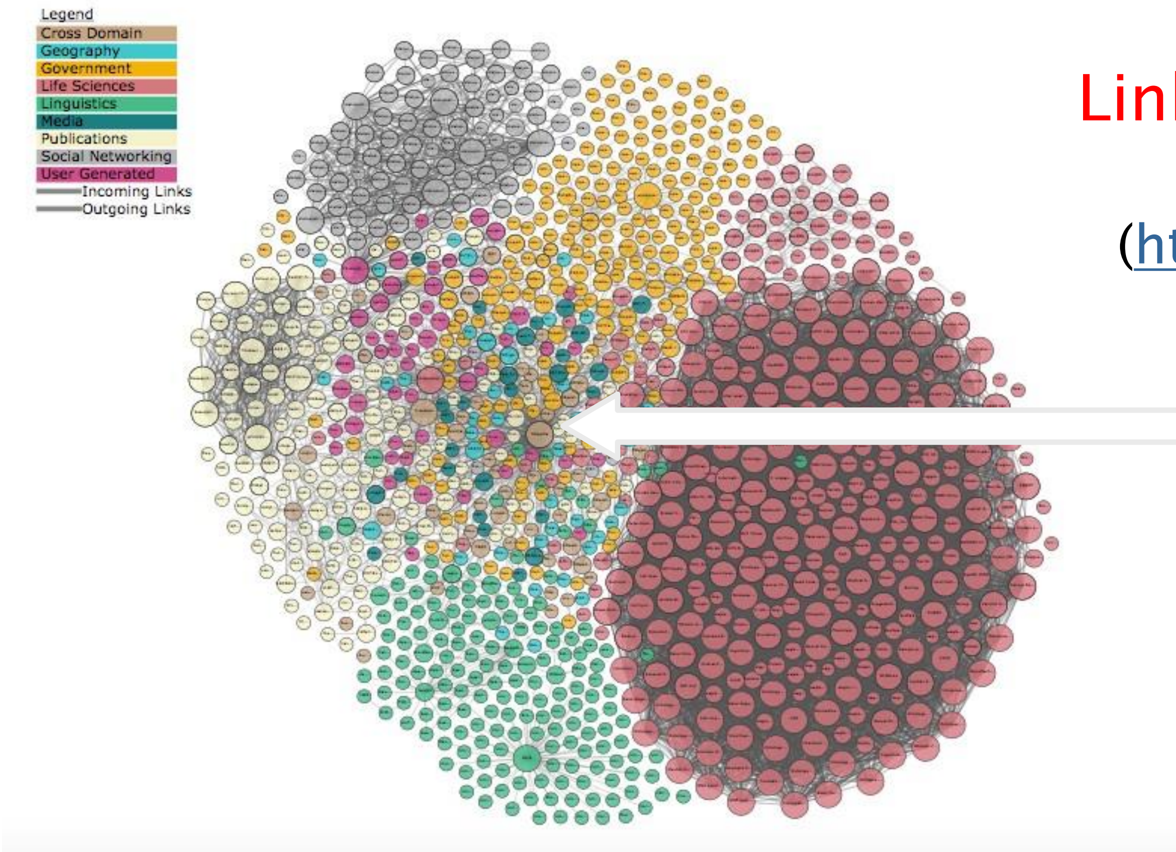
Linked Open Data cloud

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)

DBpedia

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 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

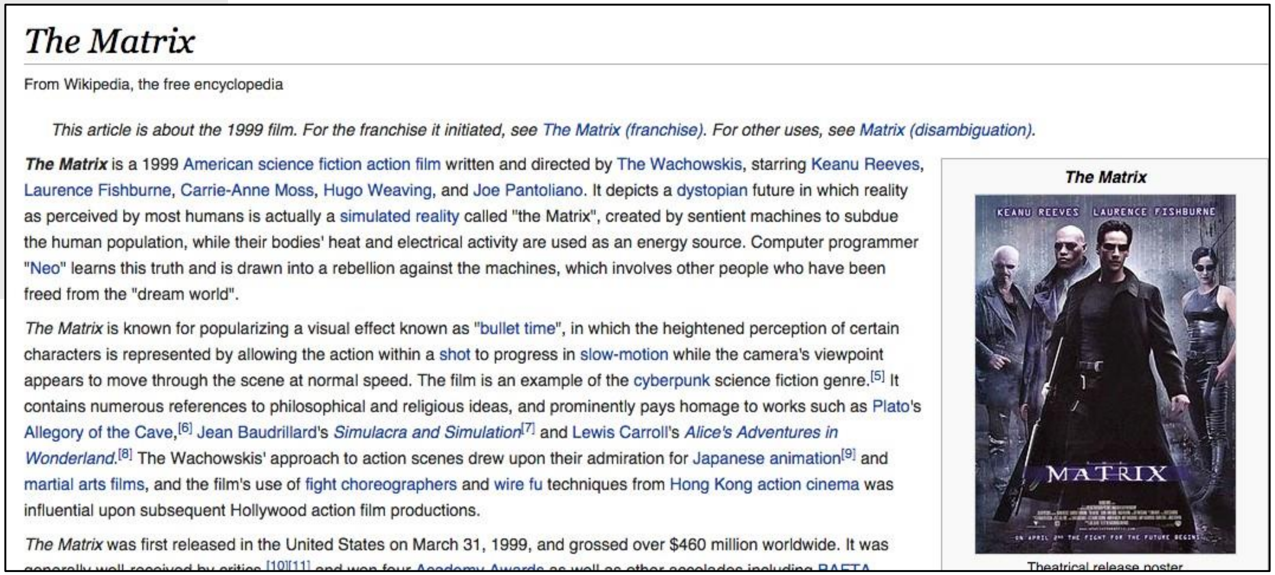
The Matrix

Theatrical release poster

Wikipedia

Unstructured Content

DBpedia



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

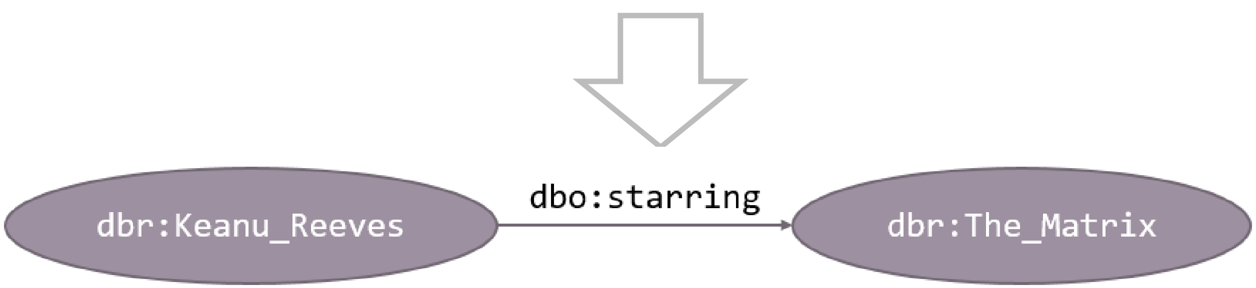
The Matrix

KEANU REEVES LAURENCE FISHBURNE

MATRIX

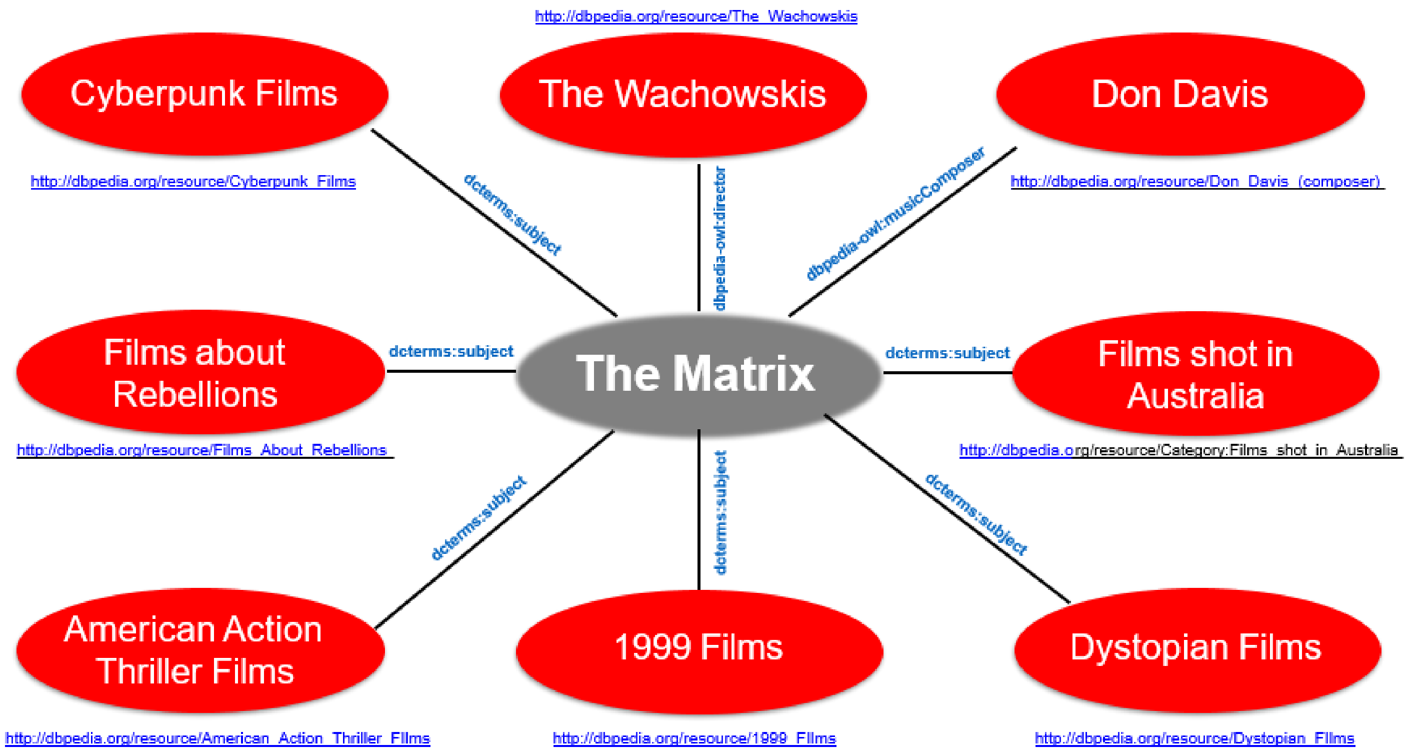
Theatrical release poster

Wikipedia
Unstructured Content



DBpedia
Structured Data

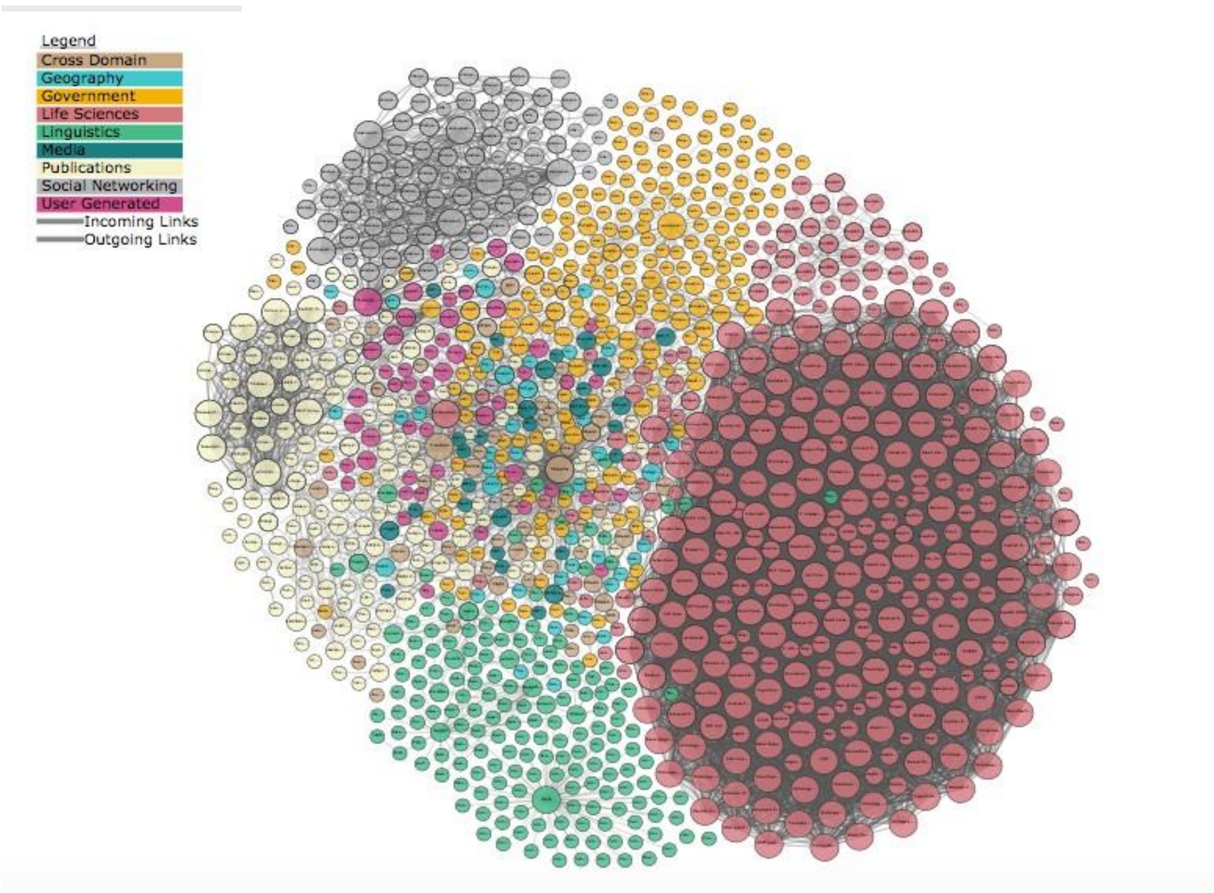
DBpedia



All the information available in Wikipedia

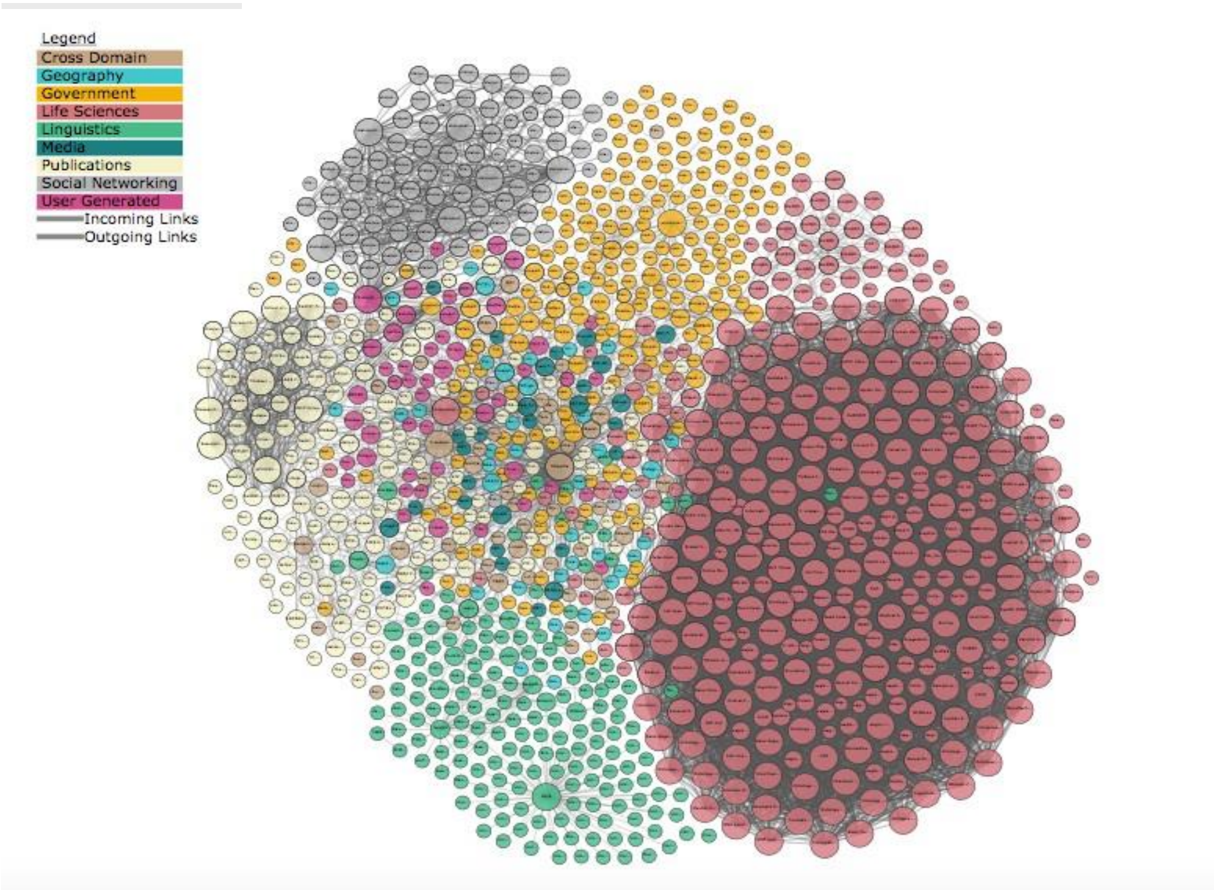
is modeled in RDF

Dbpedia – In a Nutshell



We have **interesting features coming from Wikipedia (and other sources)** and the advantage of **formal semantics defined in RDF**

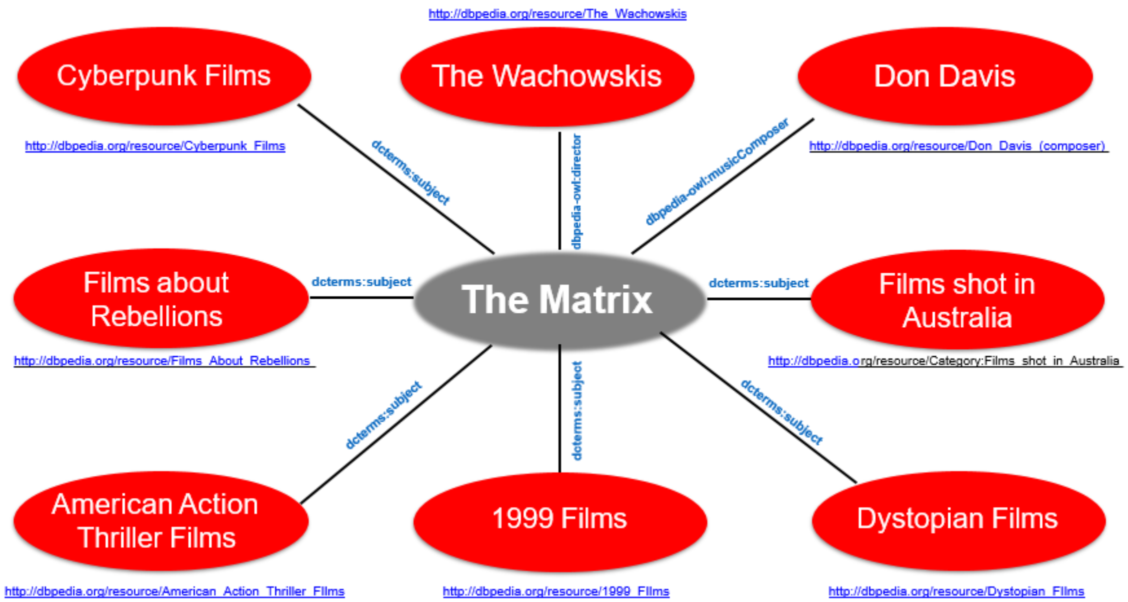
Dbpedia – In a Nutshell



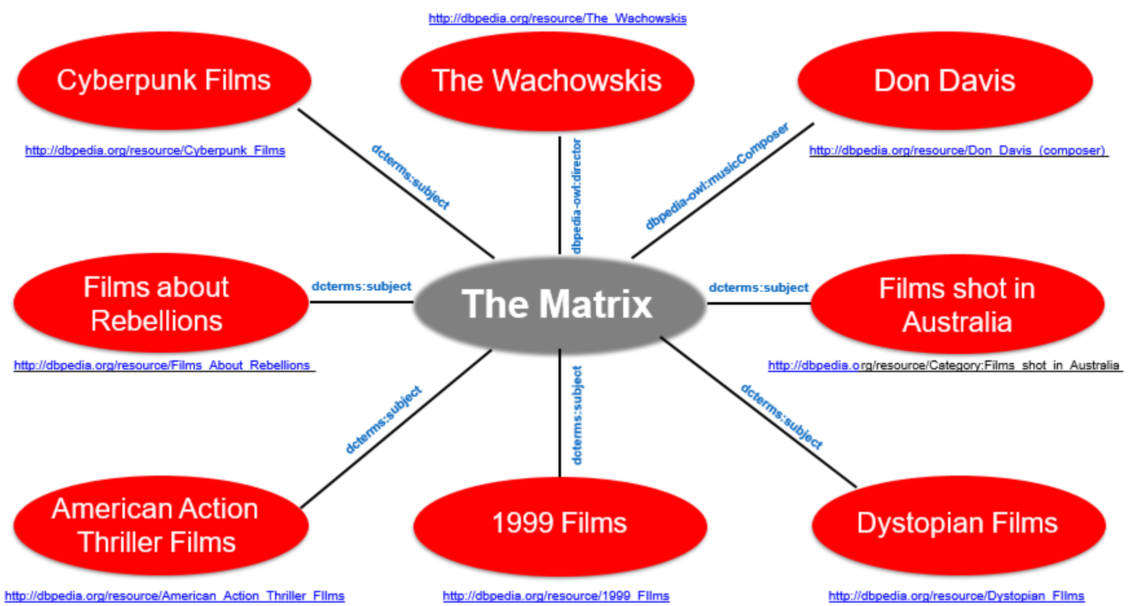
We have **interesting features coming from Wikipedia (and other sources)** and the advantage of **formal semantics defined in RDF**

We have semantics without the need of building and **manually populating an ontology**

...One step back



...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**

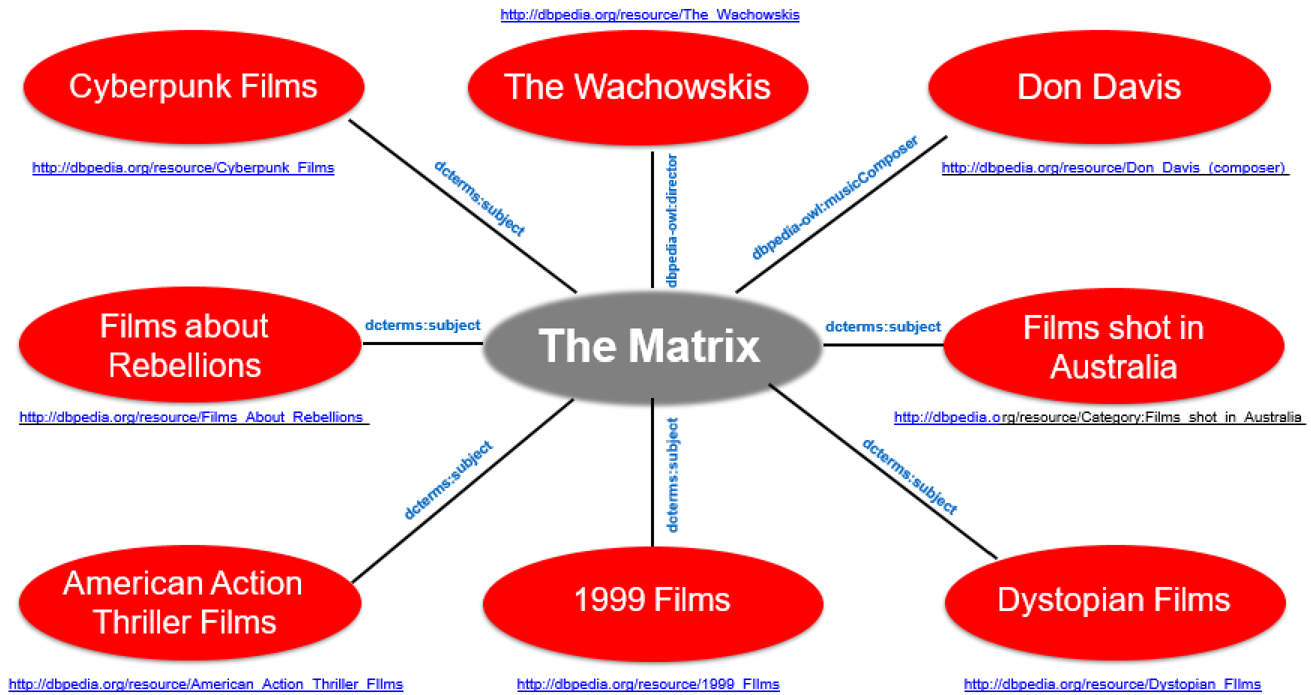
SPARQL

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")
}
```

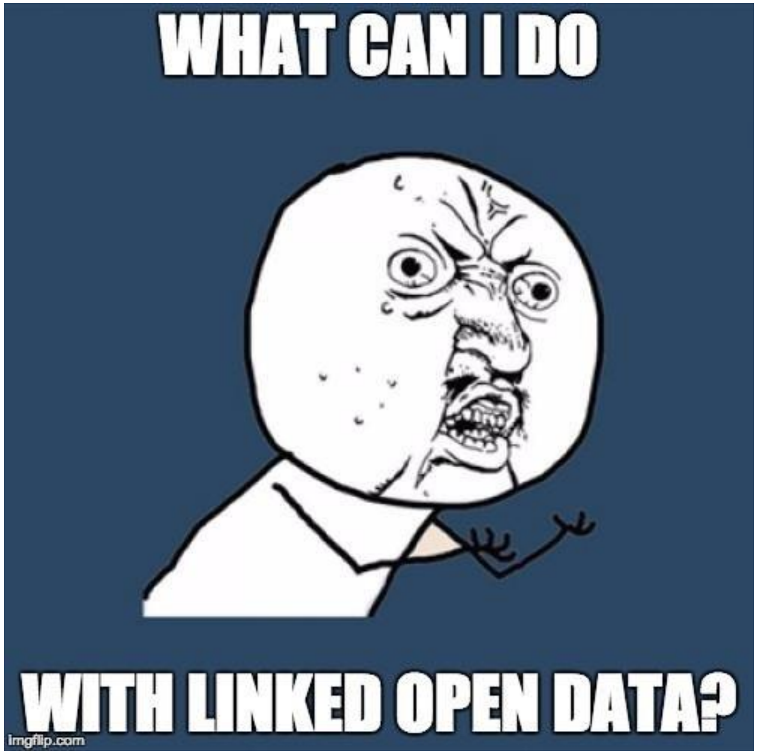
We want to link «logical» entities occurring in our data with «physical» entities occurring in the LOD cloud

LOD-aware data model



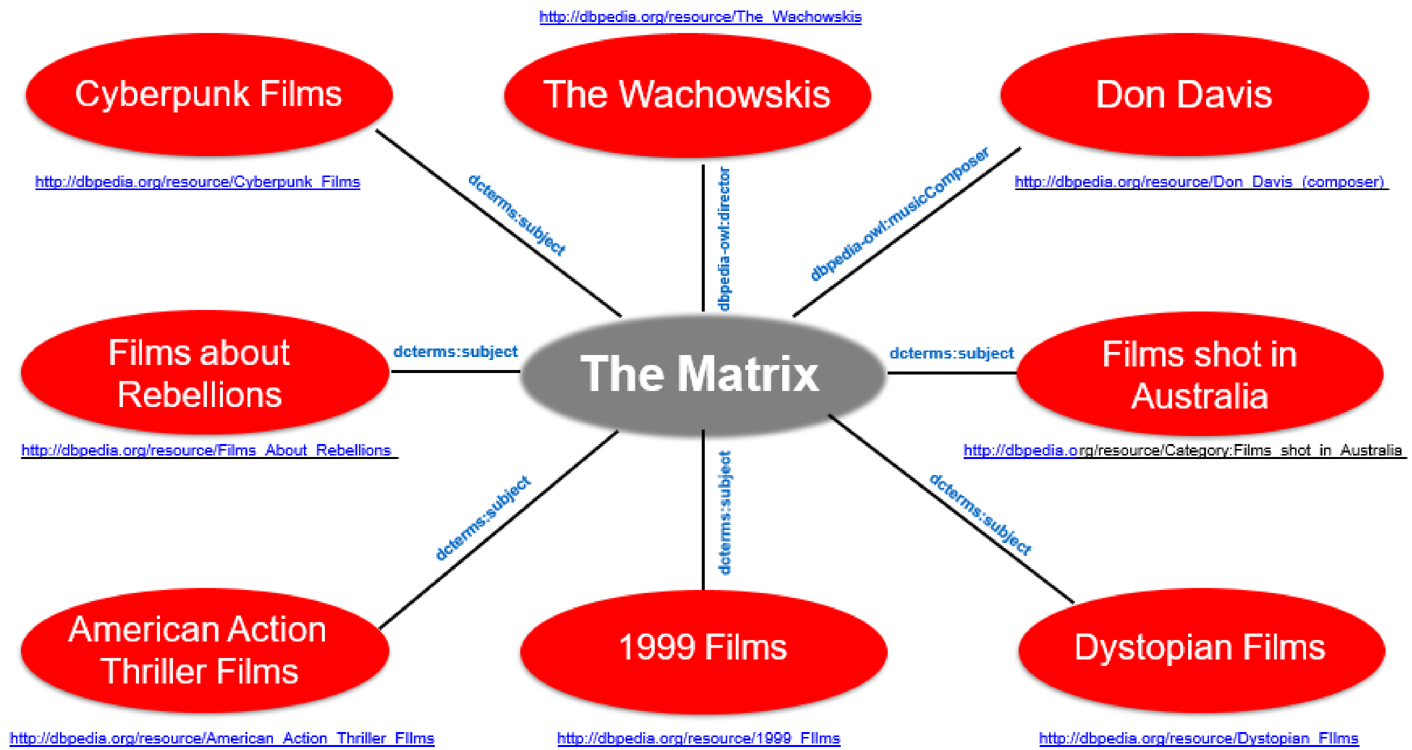
Once we have a mapping, properties can be extracted

LOD-aware RecSys



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

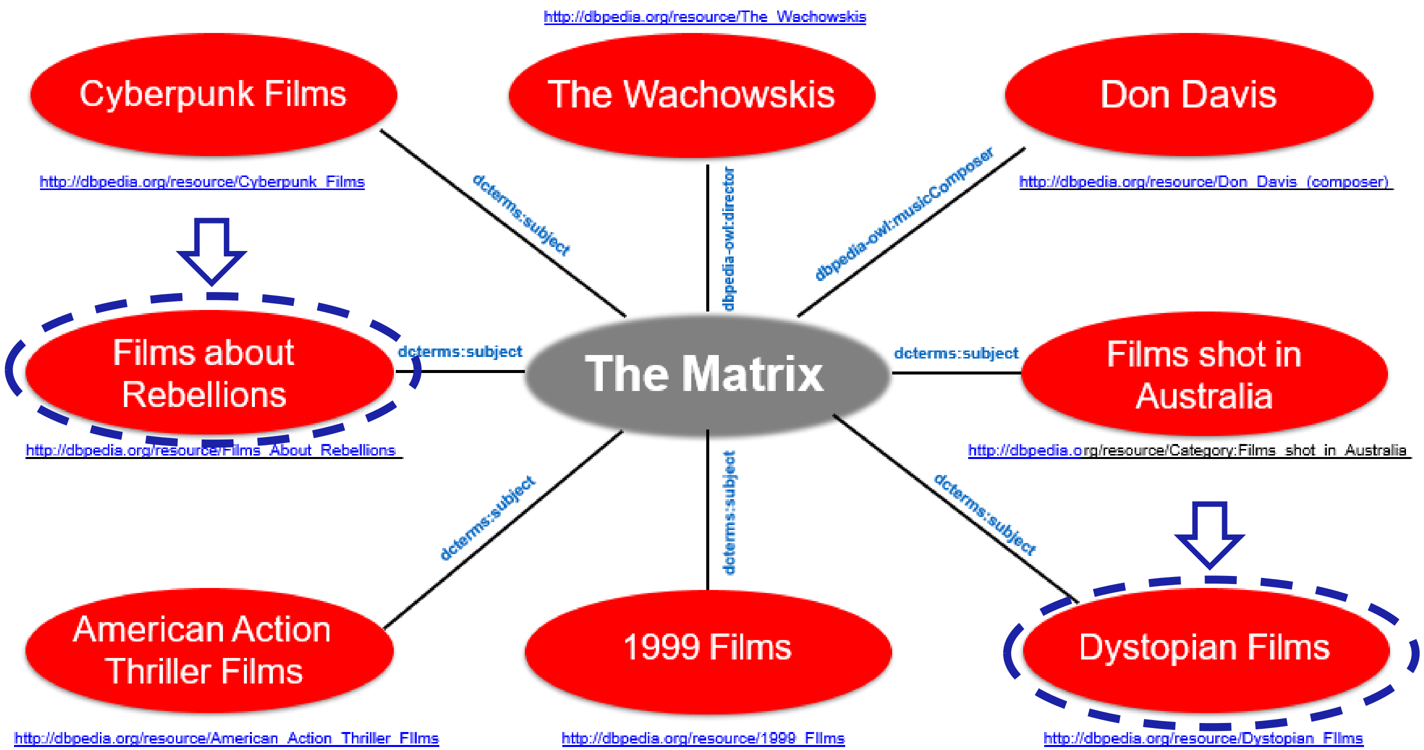
Motivations: Limited Content Analysis



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

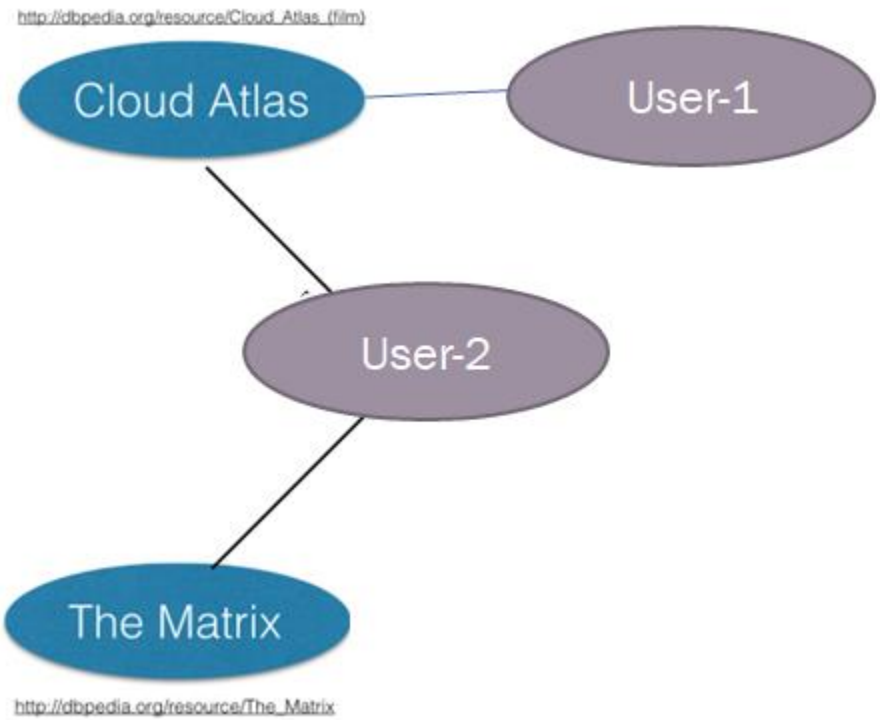
LOD cloud **can be helpful**

Motivations: Limited Content Analysis



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

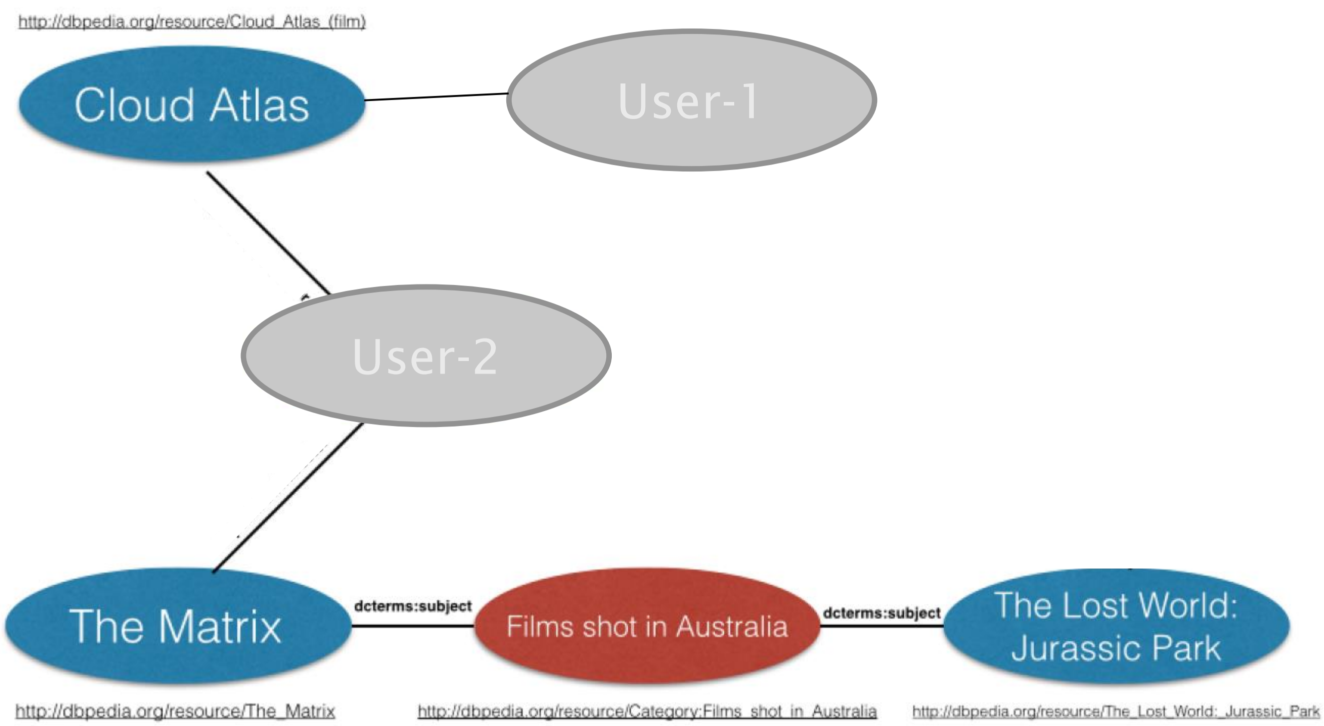
Motivations: Graph-based Data Model



Basic Graph-based Data Model

Only collaborative connections are modeled

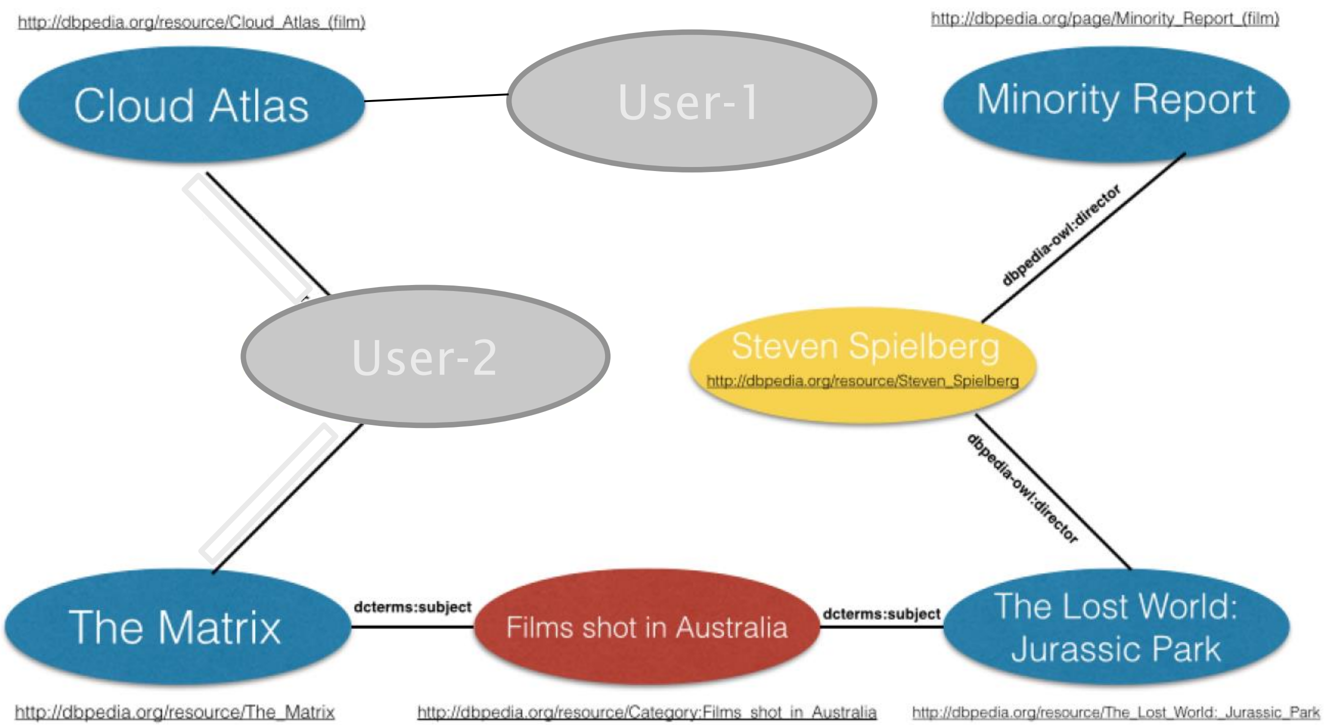
Motivations: Graph-based Data Model



Extended Graph-based Data Model

Richer representation based on properties gathered from the LOD cloud

Motivations: Graph-based Data Model



Extended Graph-based Data Model

New and unexpected connections may lead to more surprising recommendations

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 are typically used to cope with limited content analysis problem

LOD-based Recommender Systems

approaches based on VSM

LOD-based Recommender Systems

approaches based on VSM

DBpedia

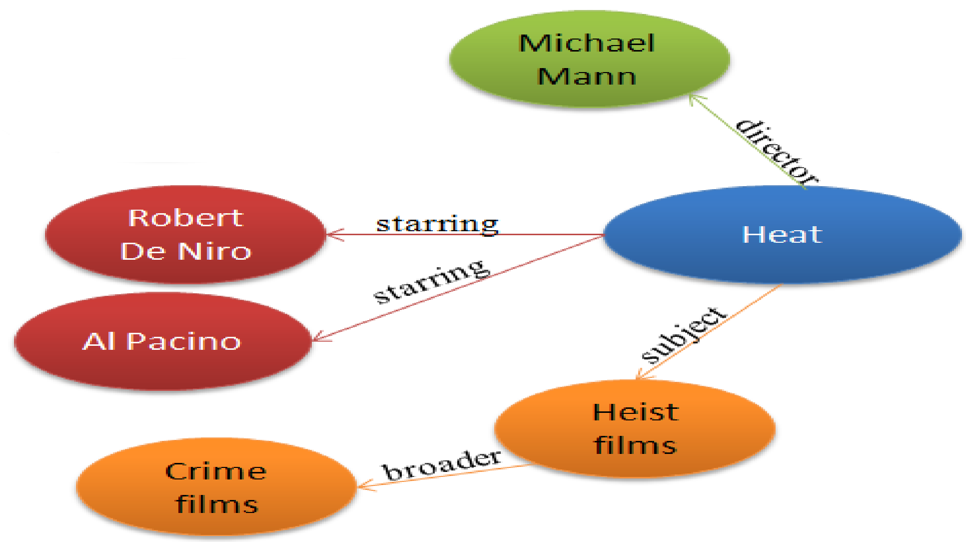
Linked Movie DataBase

WIKIDATA

SPARQL

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

$$sim_{jaccard}(x_i, x_j) = \frac{|N_d(x_i) \cap N_d(x_j)|}{|N_d(x_i) \cup N_d(x_j)|}$$

similarity between items

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

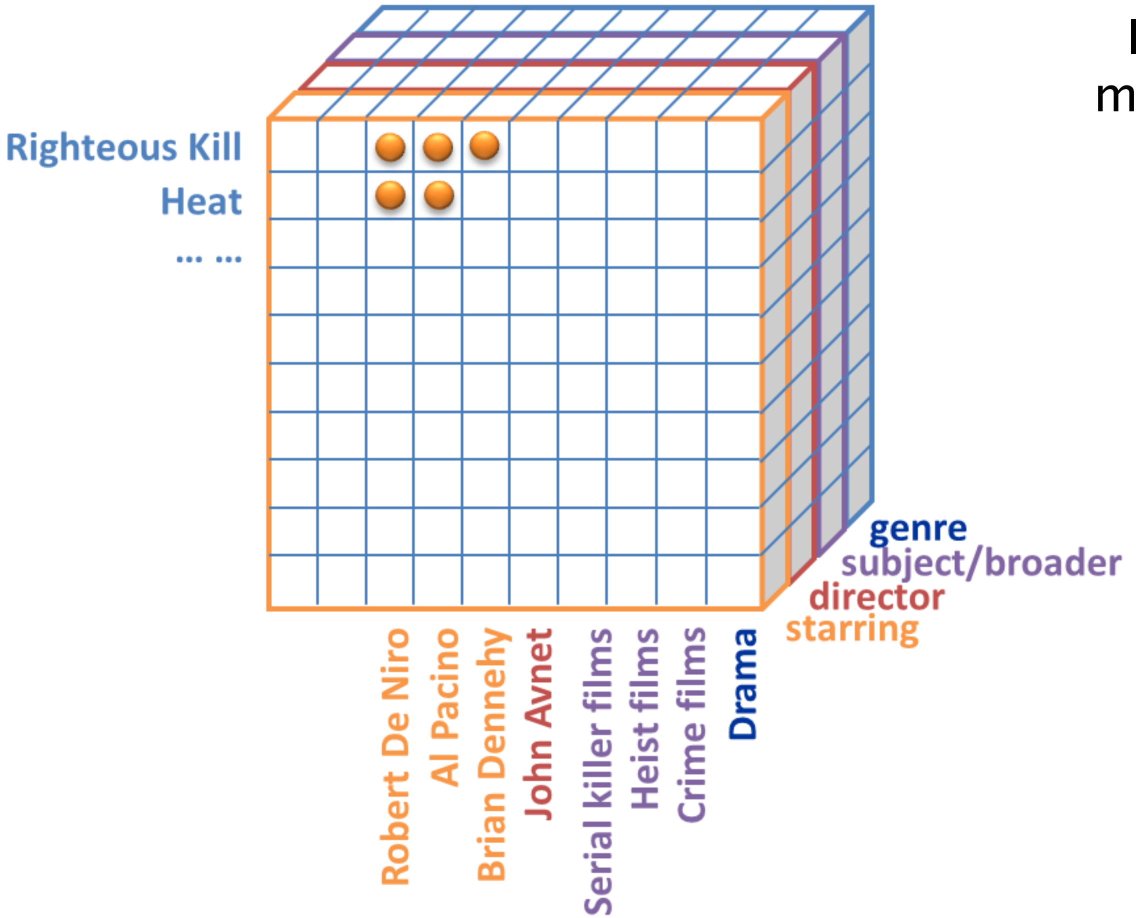
$$sim_{jaccard}(x_i, x_j) = \frac{|N_d(x_i) \cap N_d(x_j)|}{|N_d(x_i) \cup N_d(x_j)|}$$

similarity between items

Can we think about more complex models?

LOD-based Recommender Systems

Vector Space Model for LOD

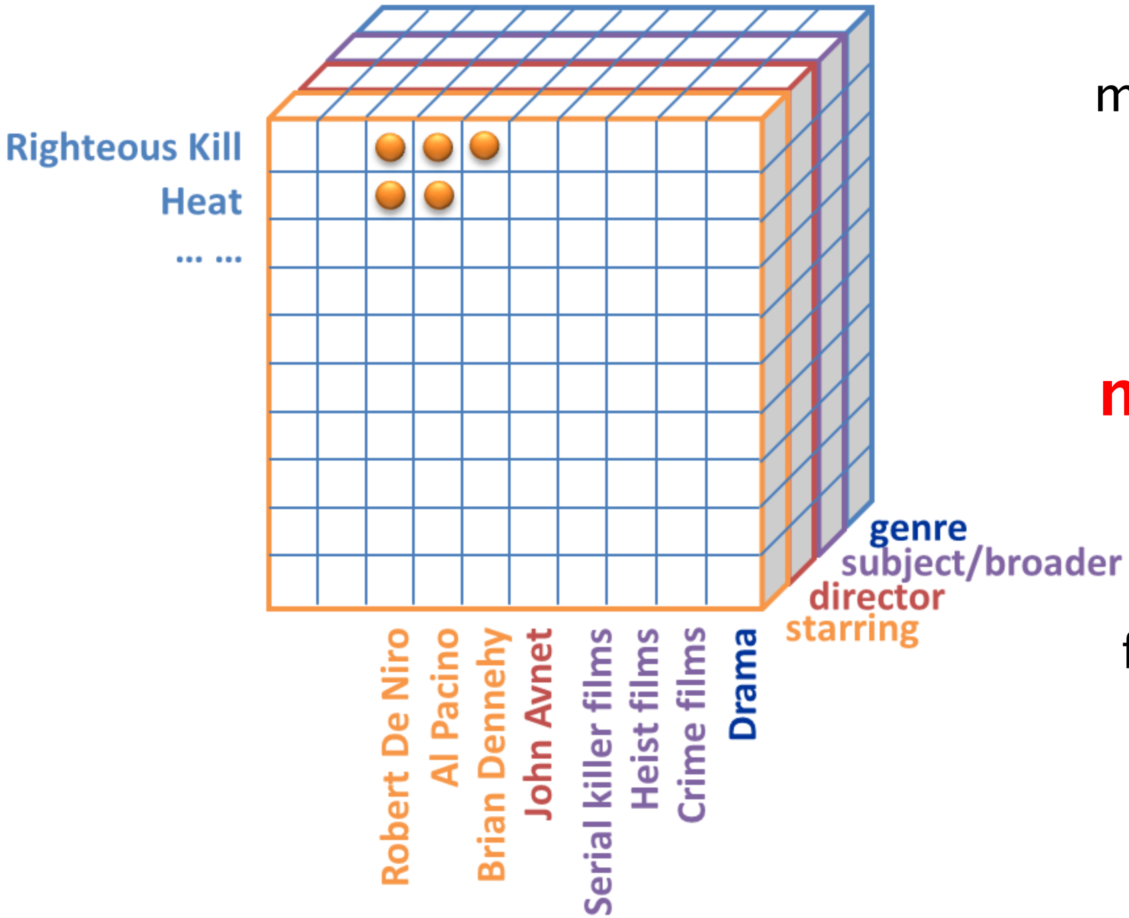


In DBpedia each item is modeled on the ground of **several facets**

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD-based Recommender Systems

Vector Space Model for LOD



In DBpedia each item is modeled on the ground of **several facets**

Each facet is modeled as a slice of a tensor.

Each slice encodes the features describing that particular facet.

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD-based Recommender Systems

Vector Space Model for LOD

$$\alpha_{starring} * sim_{starring}(\overrightarrow{x_i}, \overrightarrow{x_j})$$ +

$$\alpha_{director} * sim_{director}(\overrightarrow{x_i}, \overrightarrow{x_j})$$ +

$$\alpha_{subject} * sim_{subject}(\overrightarrow{x_i}, \overrightarrow{x_j})$$ +

...

=

$$sim\,(\overrightarrow{x_i}, \overrightarrow{x_j})$$

Similarity between items as linear combination of the similarity **among Dbpedia facets** *(starring, directors, subject, etc.)*

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD-based Recommender Systems

VSM content-based recommender

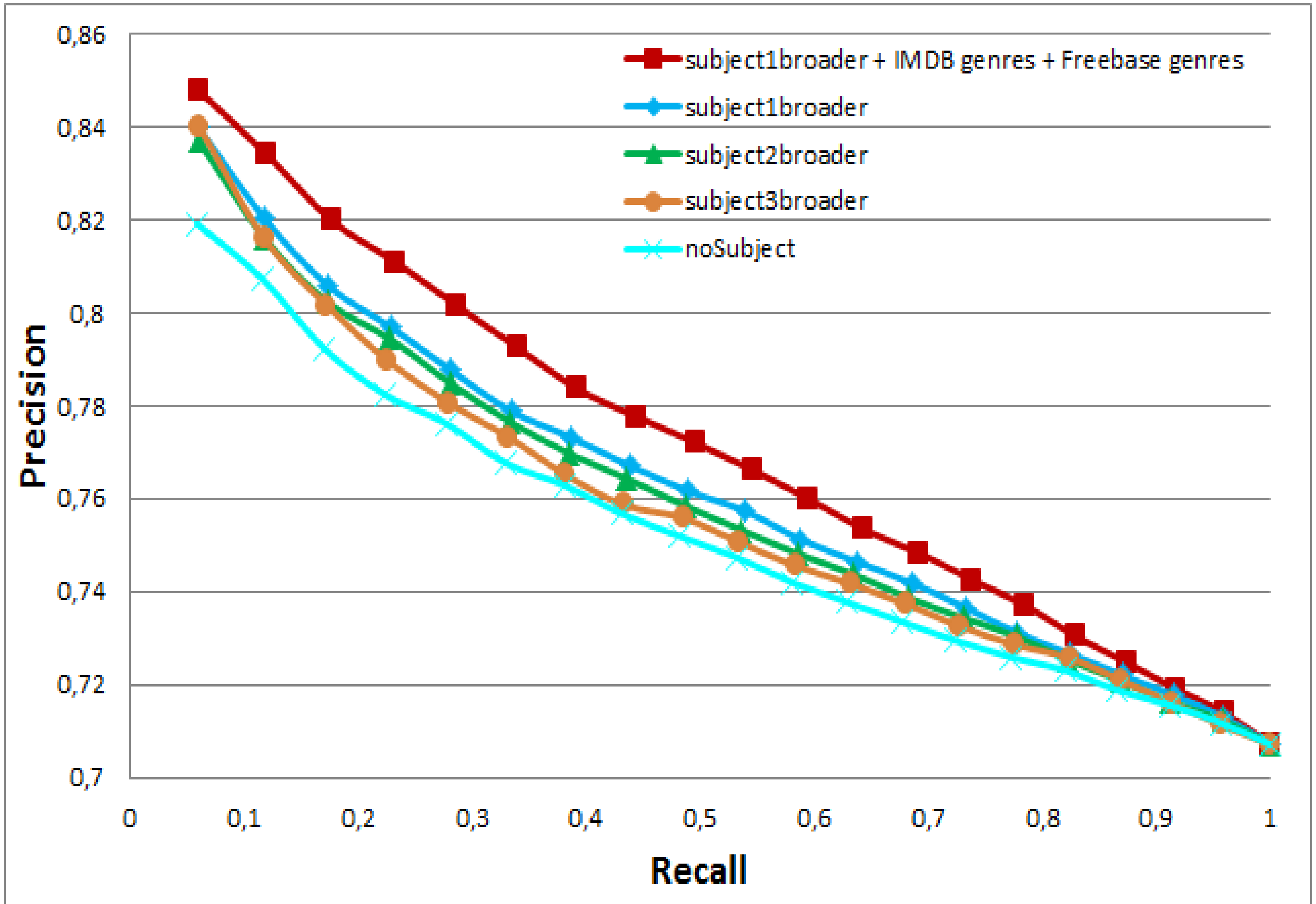
$$\tilde{r}(u, x_j) = \frac{\sum_{x_i \in Profile(u)} r(u, x_i) \cdot \frac{\sum_{p \in P} \alpha_p \cdot sim_p(x_i, x_j)}{|P|}}{|profile(u)|}$$

Predict the rating using a **Nearest Neighbor Classifier** wherein the similarity measure is a linear combination of **local property similarities**

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** 8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)

LOD-based Recommender Systems

Property subset evaluation



subject+broader solution better than only subject or subject+more broaders

too many broaders introduce noise

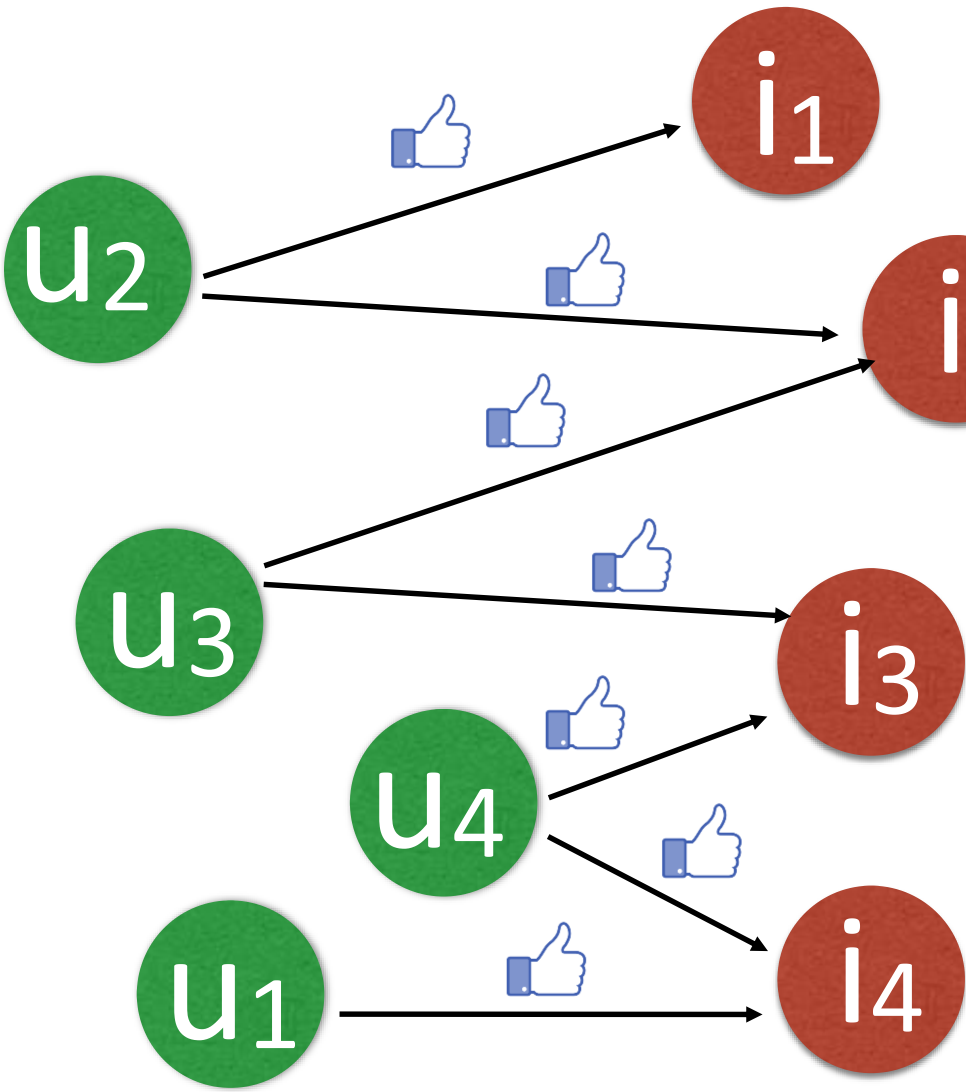
best solution achieved with **subject+broader+genres**

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

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

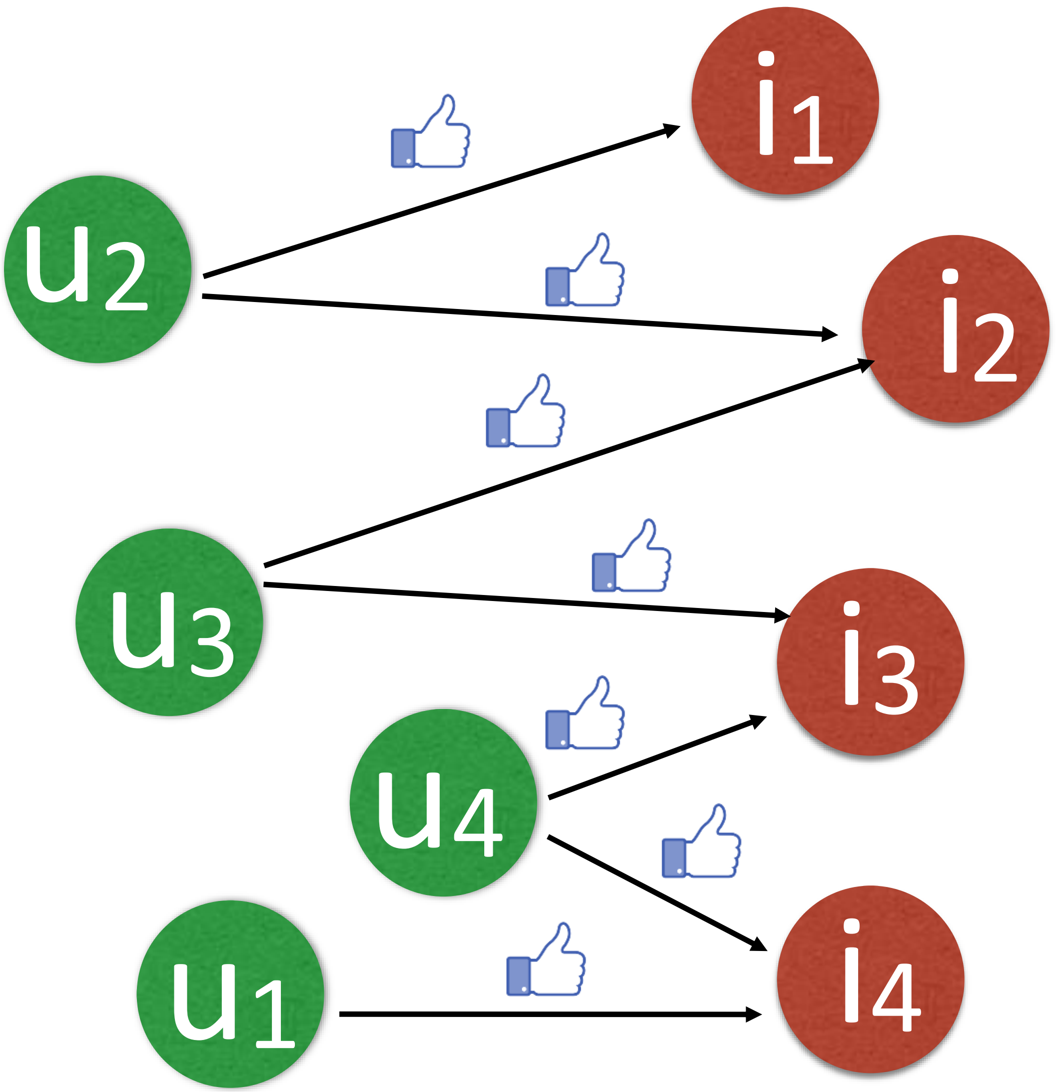


users = **nodes**
items = **nodes**
preferences = **edges**

(bipartite graph)

Very intuitive representation!

Graph-based Data Model



users = **nodes**

items = **nodes**

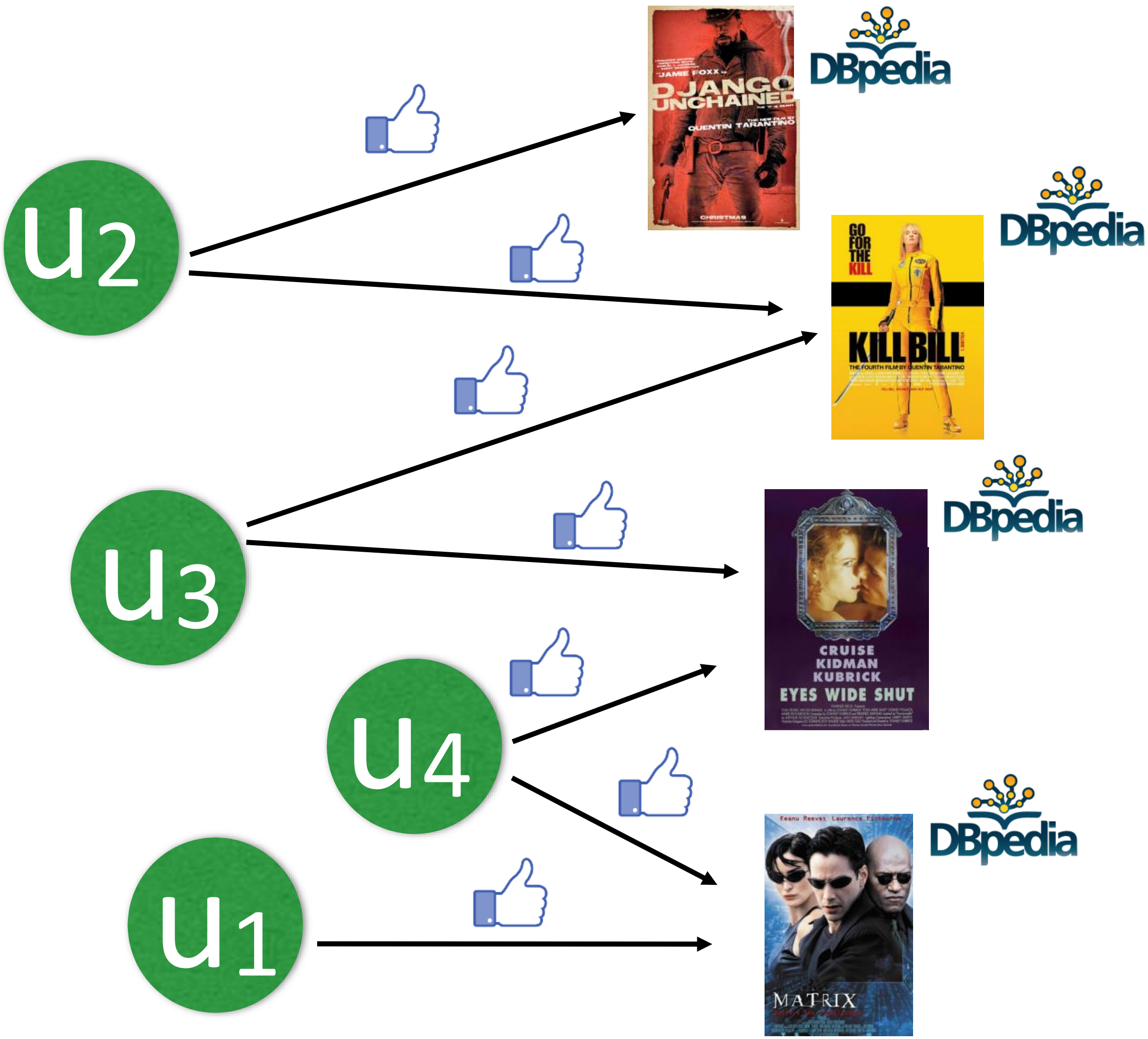
preferences = **edges**

(bipartite graph)

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

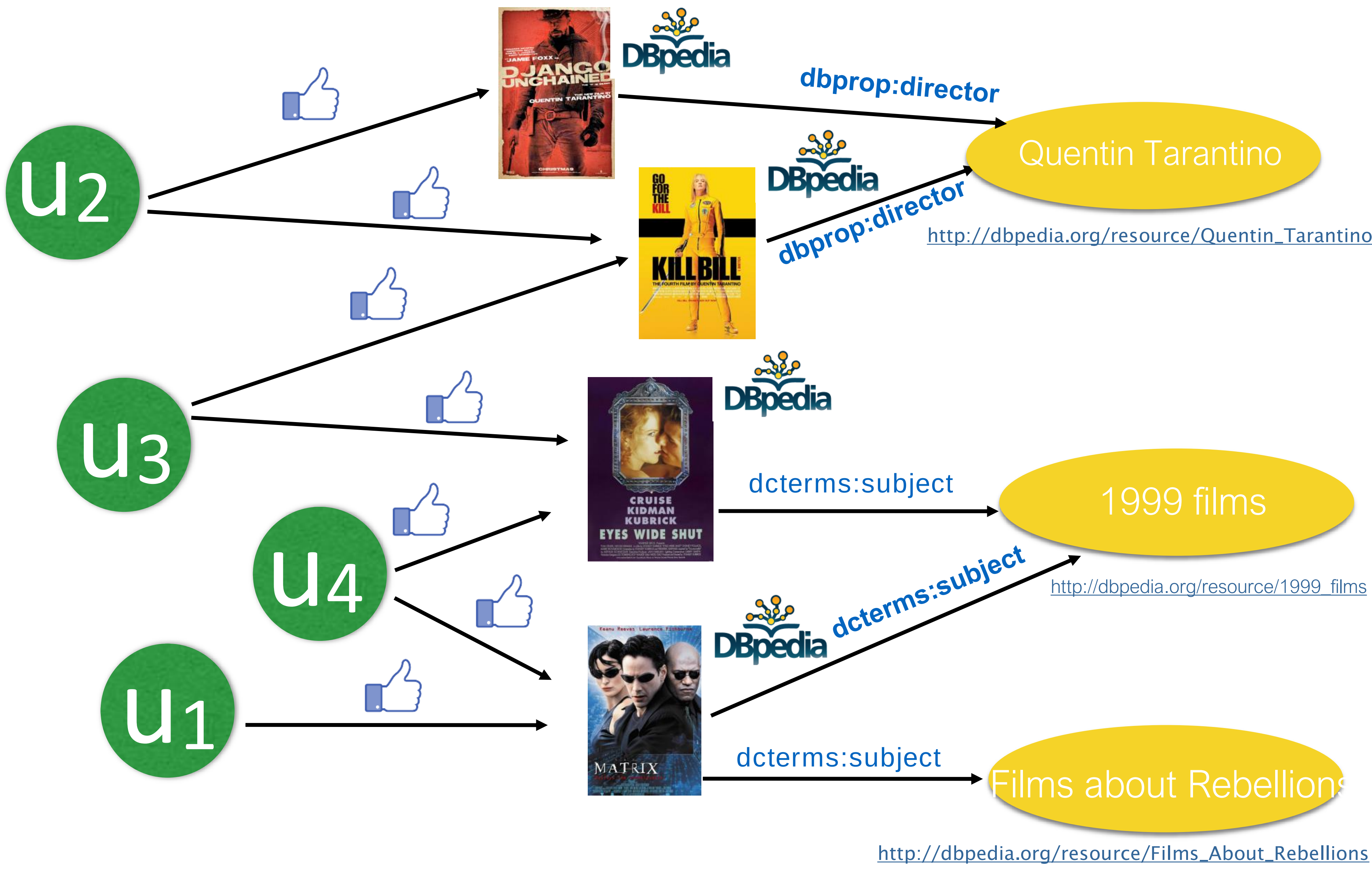
We can extend such data model by introducing features gathered from **the LOD cloud**

Semantic Graph-based Data Model

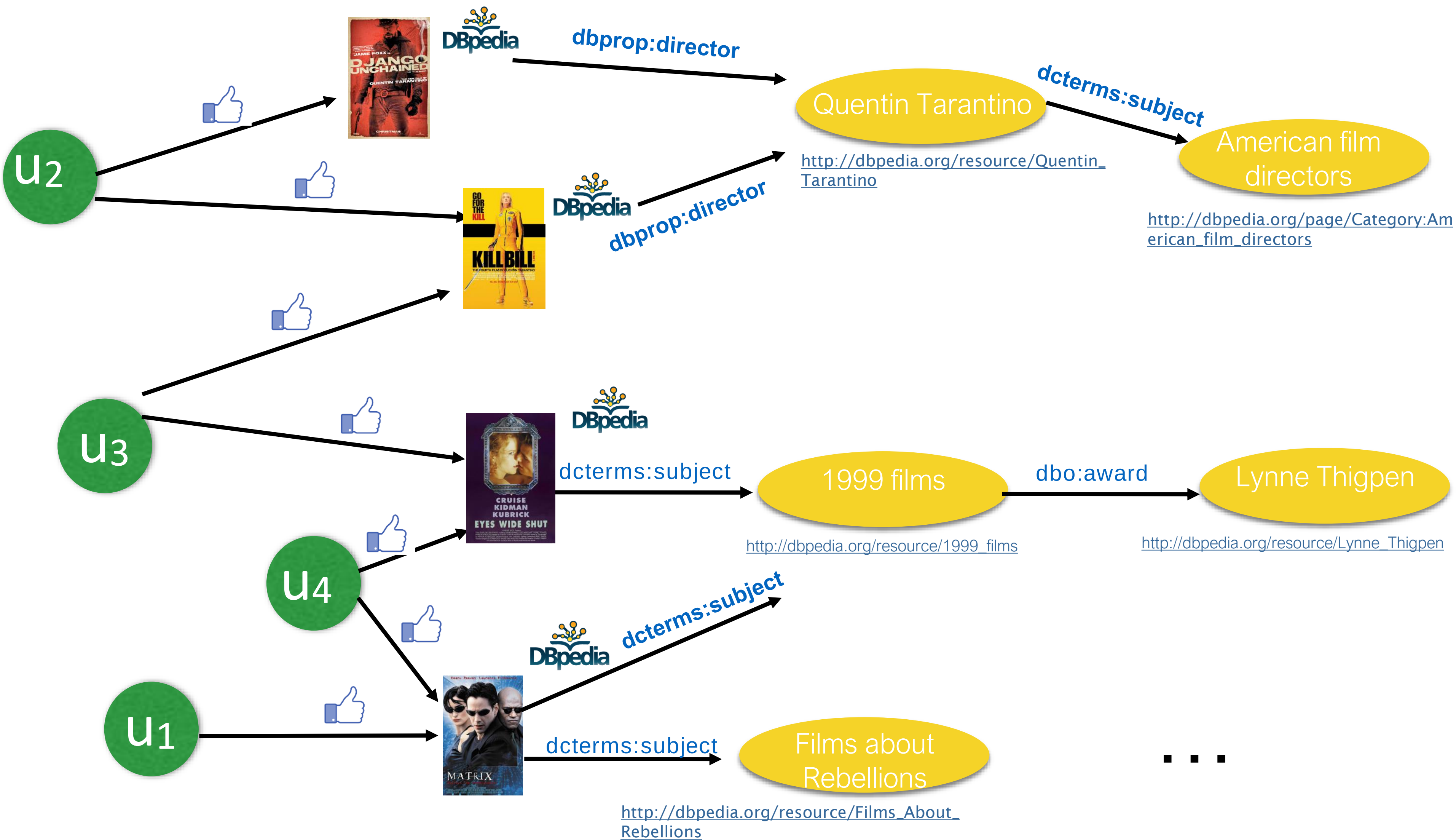


DBpedia
mapping

Semantic Graph-based Data Model
(1-hop)

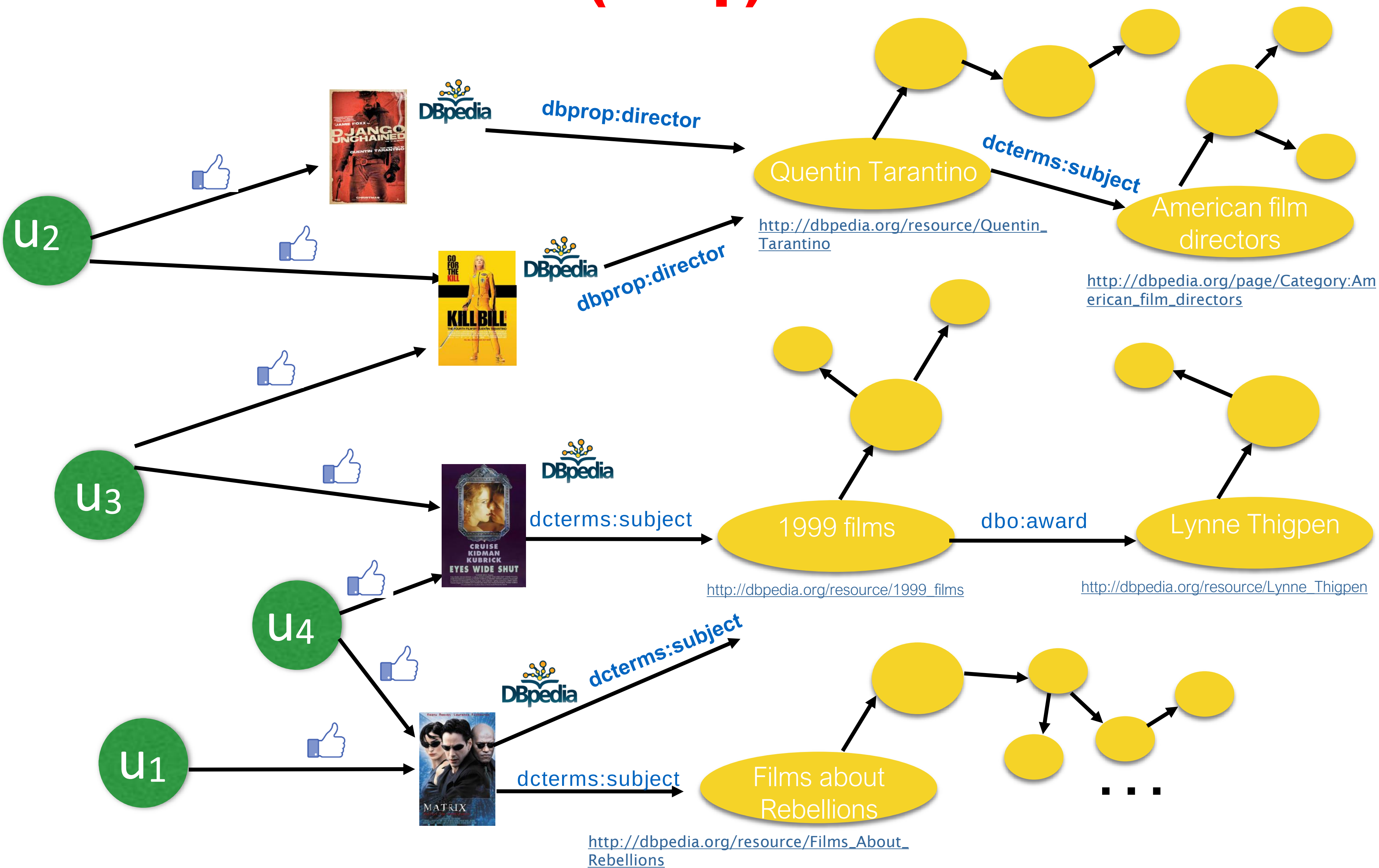


Semantic Graph-based Data Model
(2-hop)



Semantic Graph-based Data Model

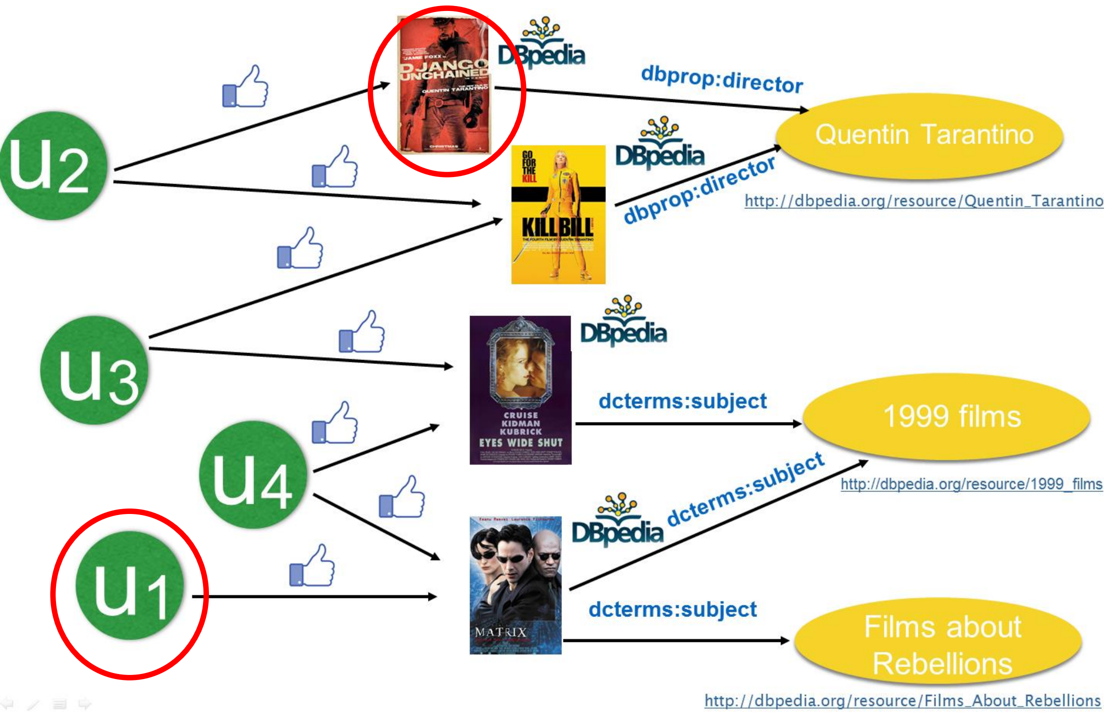
(n-hop)



Graph-based RecSys

How to get the recommendations?

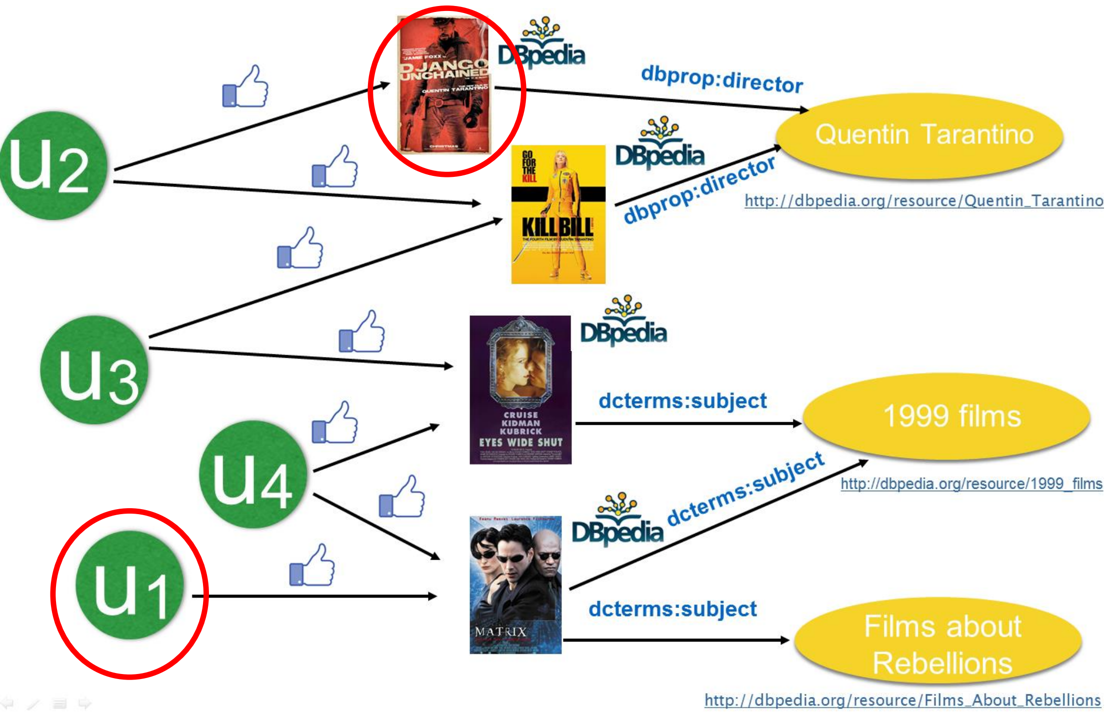
Recommendations obtained by mining the graph



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)

Graph-based RecSys

How to get the recommendations?



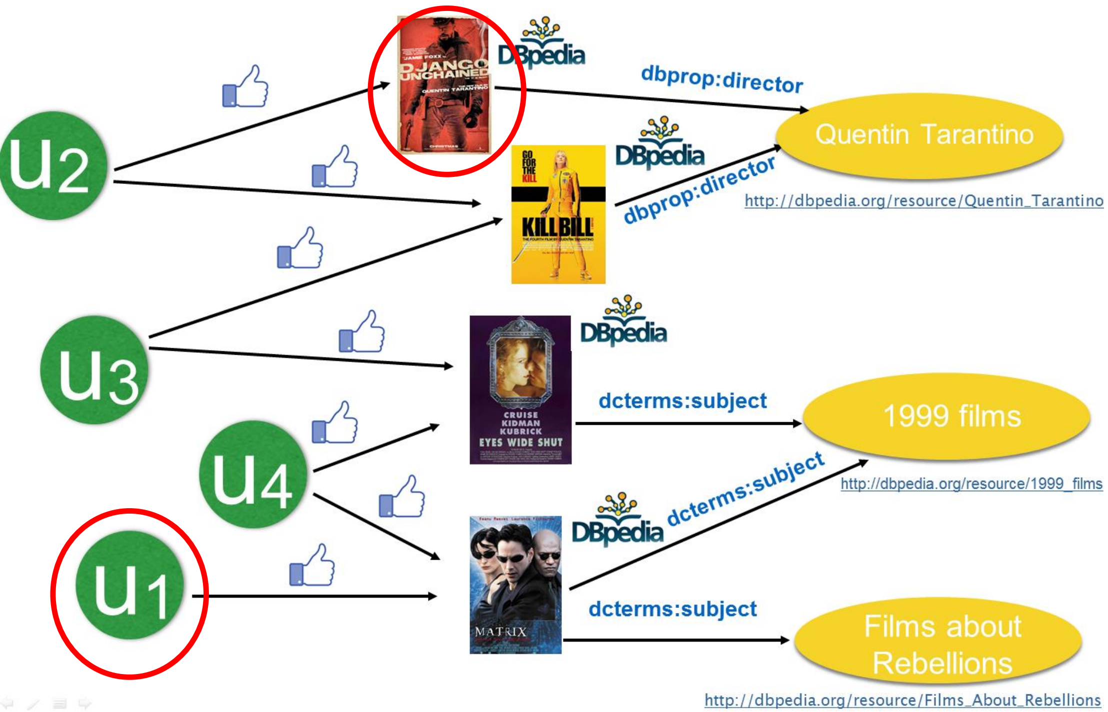
Recommendations obtained by mining the graph

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

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)

Graph-based RecSys

How to get the recommendations?



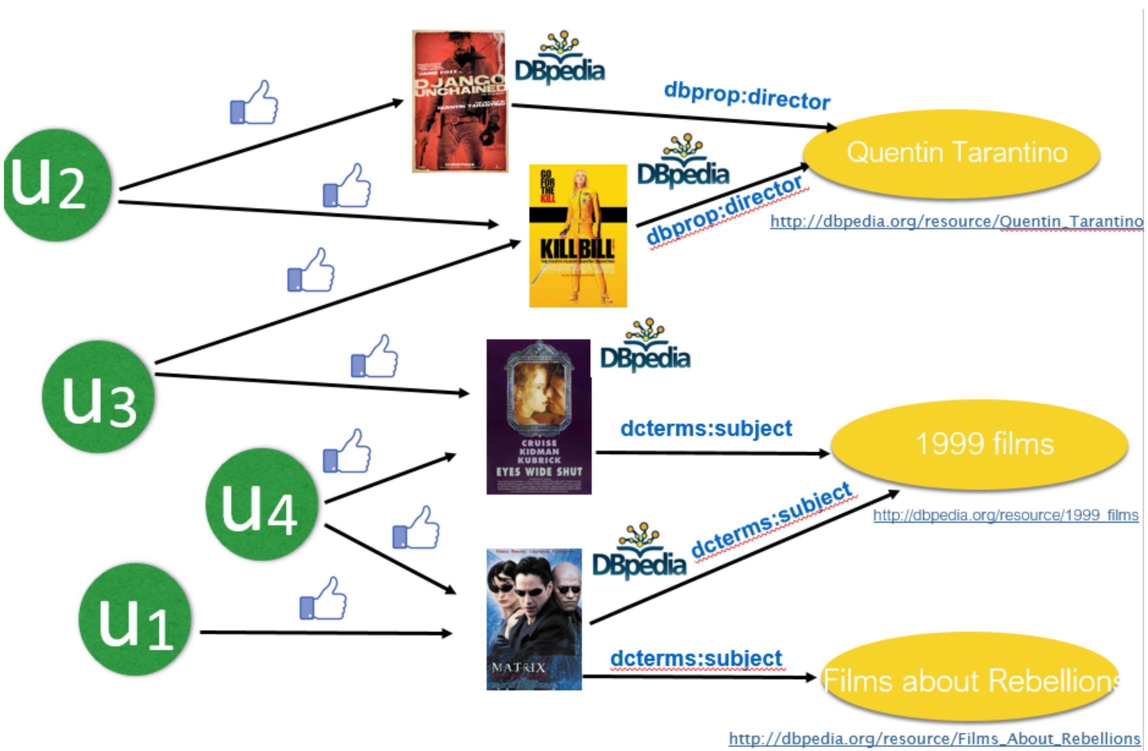
Recommendations obtained by mining the graph

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)

Graph-based RecSys



Recent work [*]

Task: top-N recommendation

Expansion: 1-hop, all the properties were injected

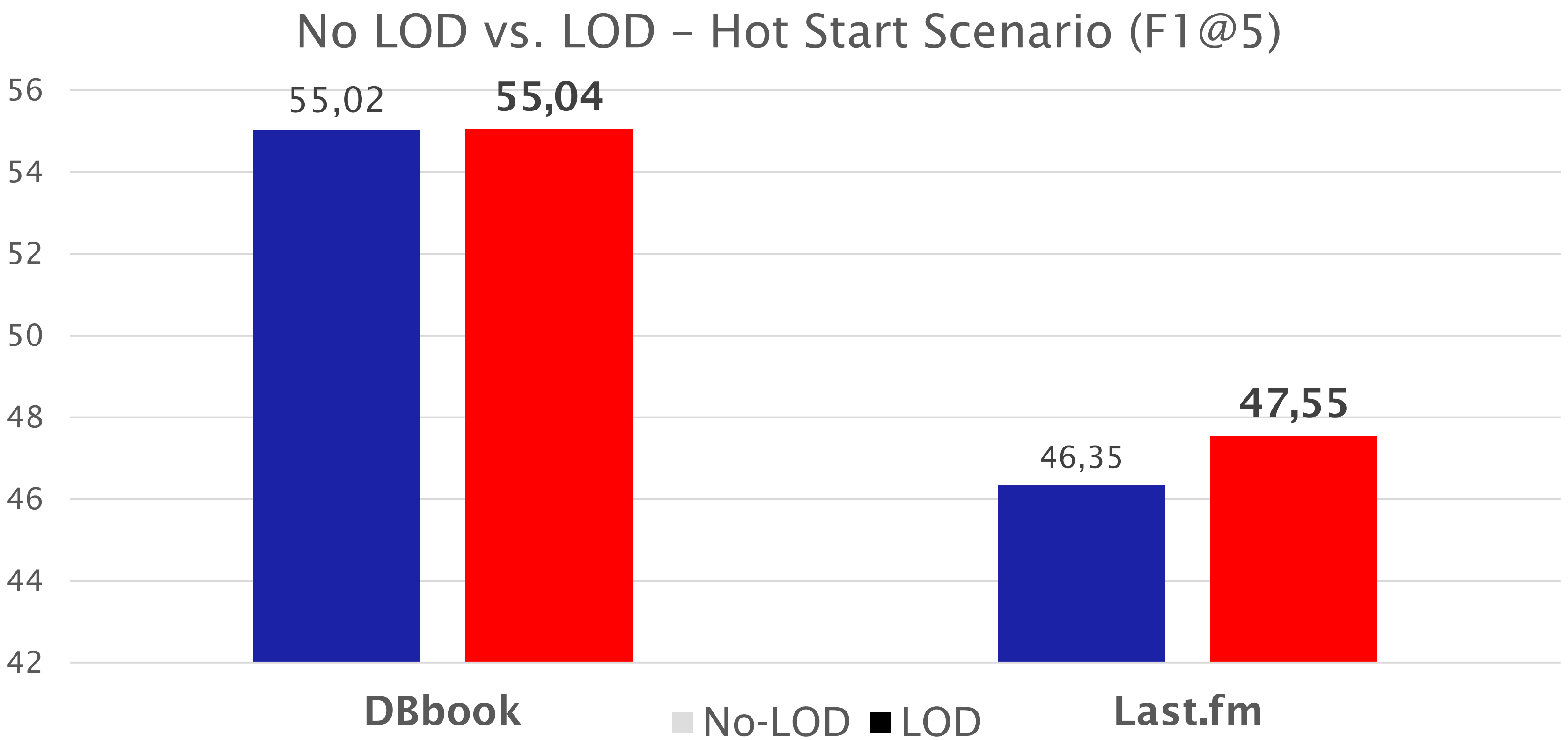
Recommendation algorithm: PageRank with Priors

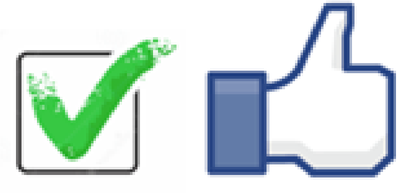
Settings: Hot Start, Cold Start

Topologies: NoLOD , LOD

[*] C. Musto, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: **Introducing linked open data in graph-based recommender systems.** Inf. Process. Manage. 53(2): 405-435 (2017)

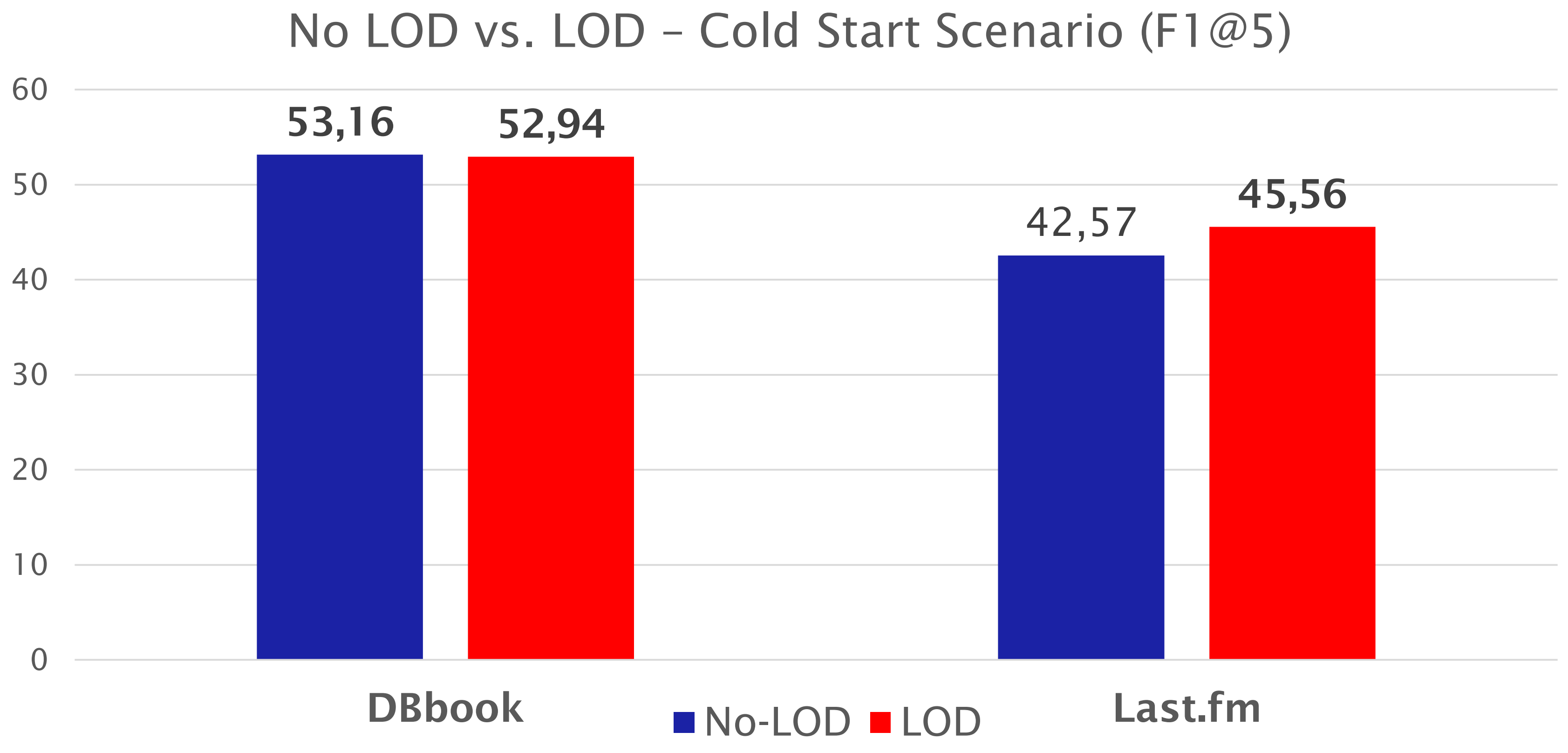
Graph-based RecSys





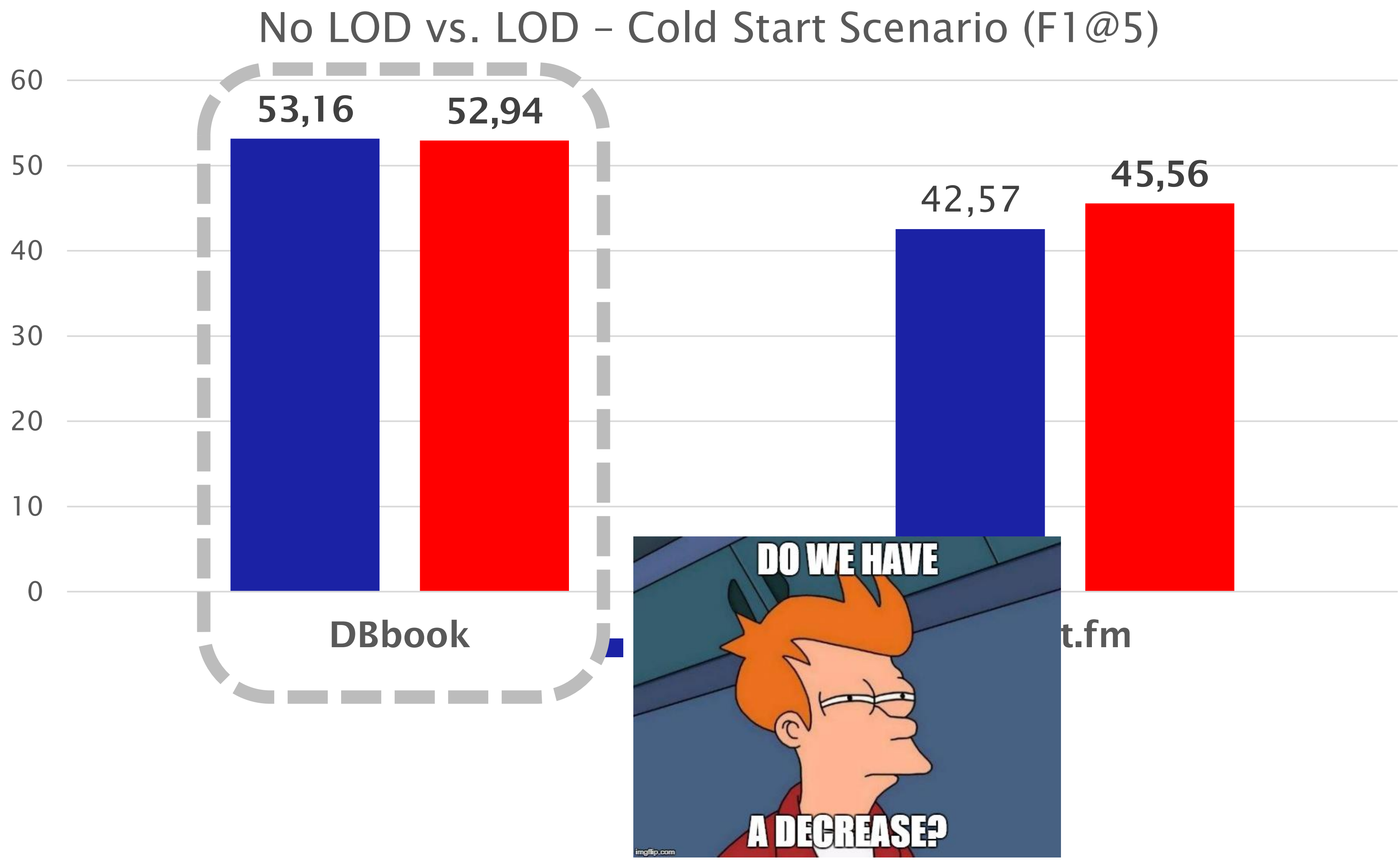
No-LOD = Bipartite User-Item graph
LOD= Tripartite graph also including LOD properties for items

Graph-based RecSys



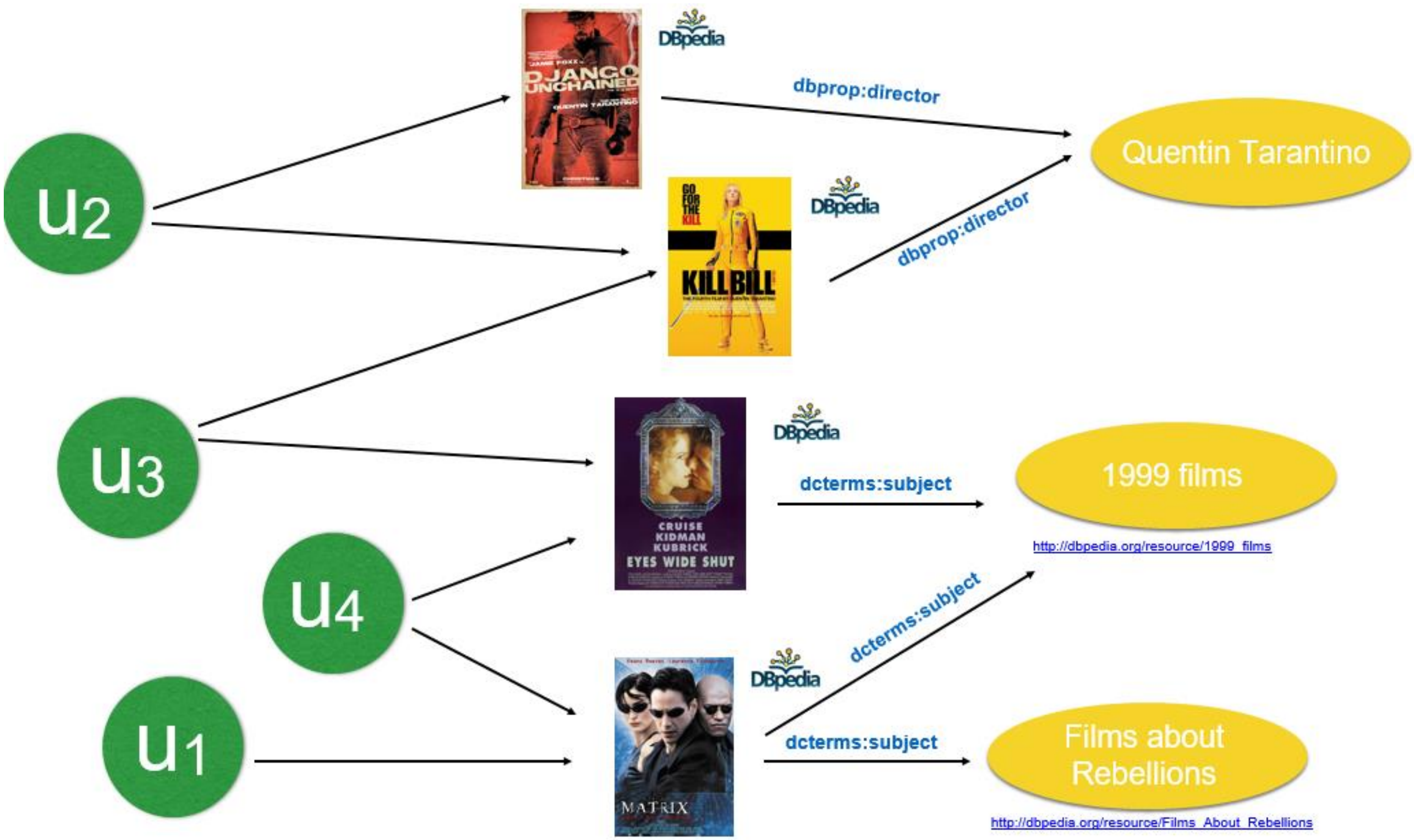
No-LOD = Bipartite User-Item graph
LOD= Tripartite graph also including LOD properties for items

Graph-based RecSys



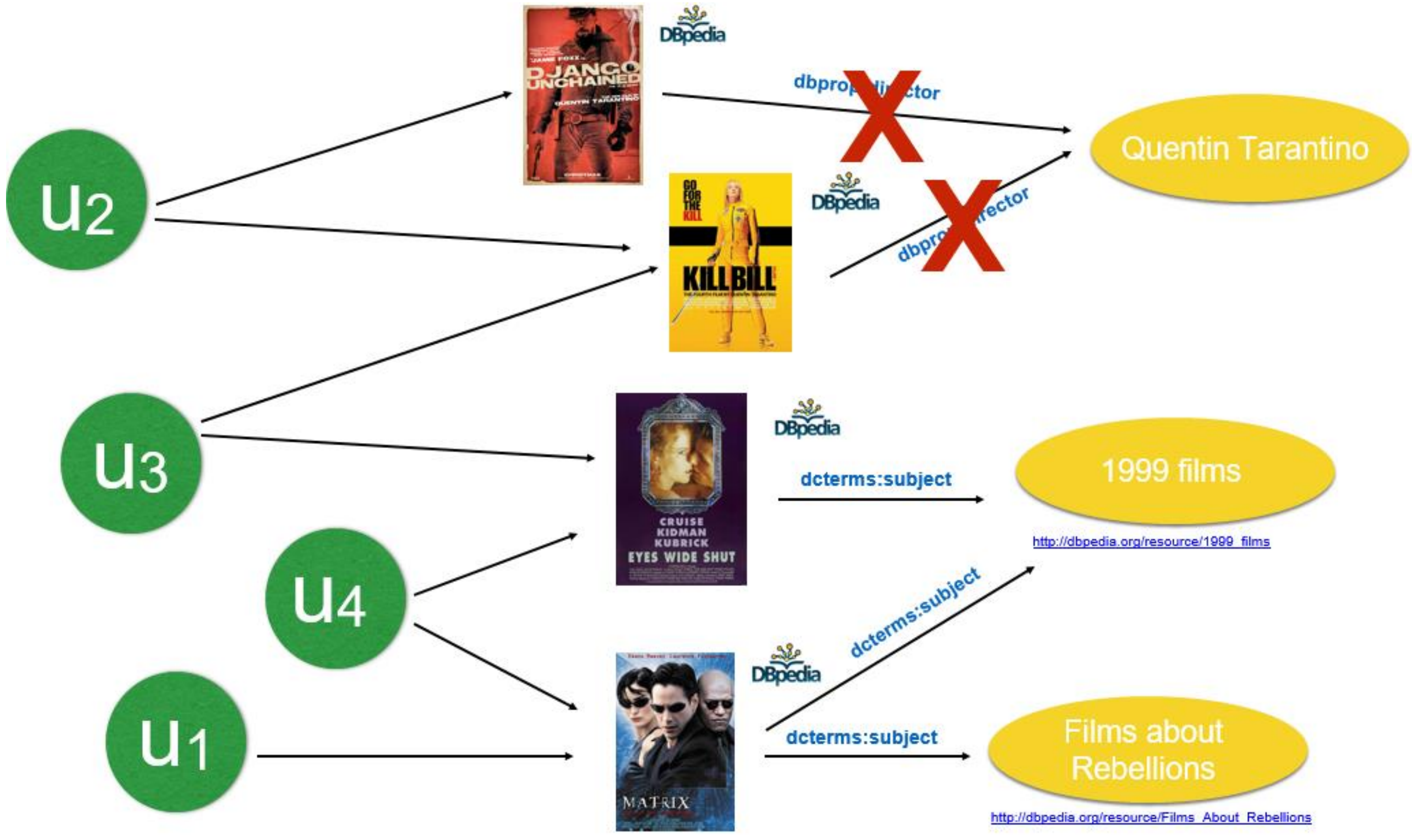
No-LOD = Bipartite User-Item graph
LOD= Tripartite graph also including LOD properties for items

Graph-based RecSys



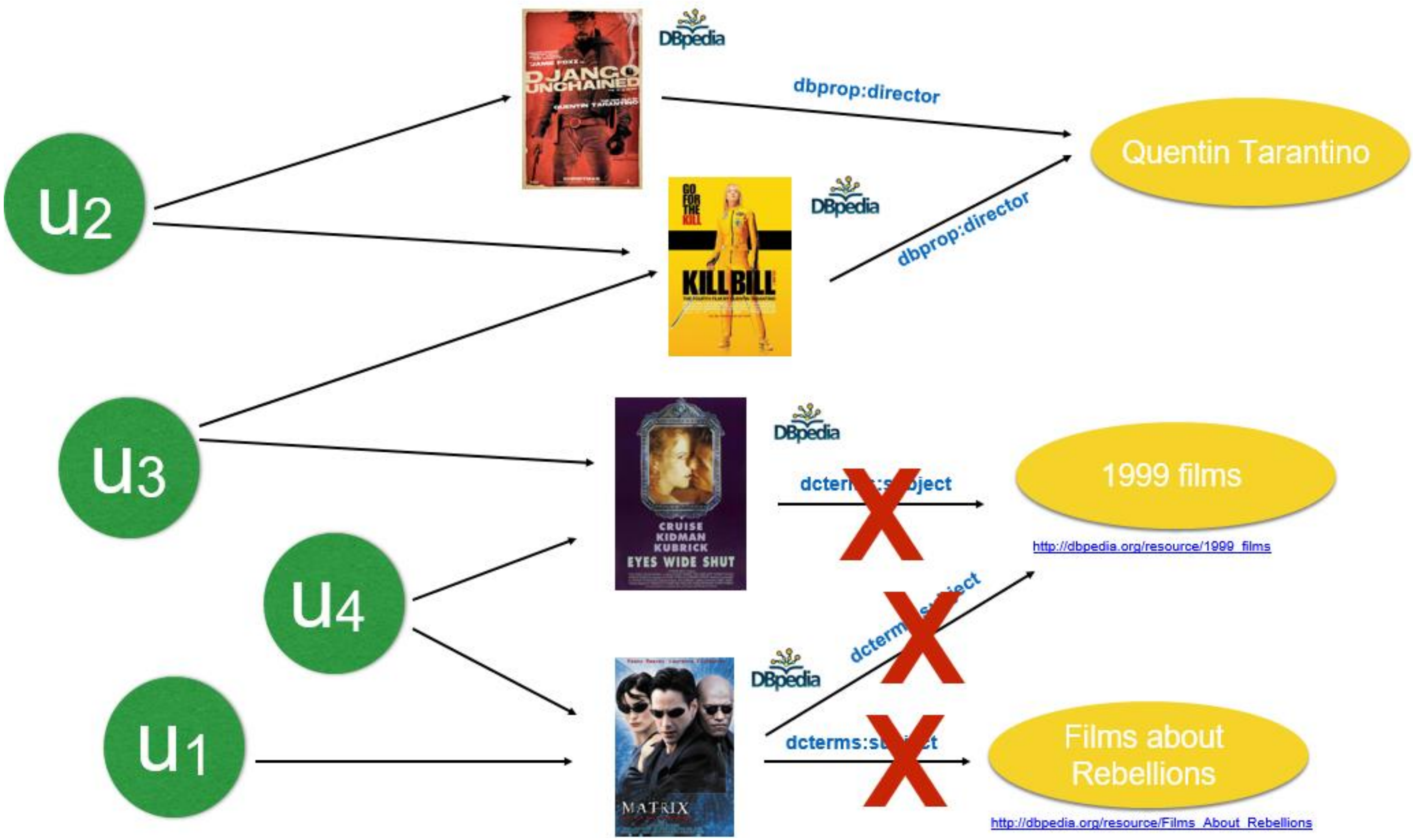
is it necessary to **inject all the properties** available in LOD cloud?

Graph-based RecSys



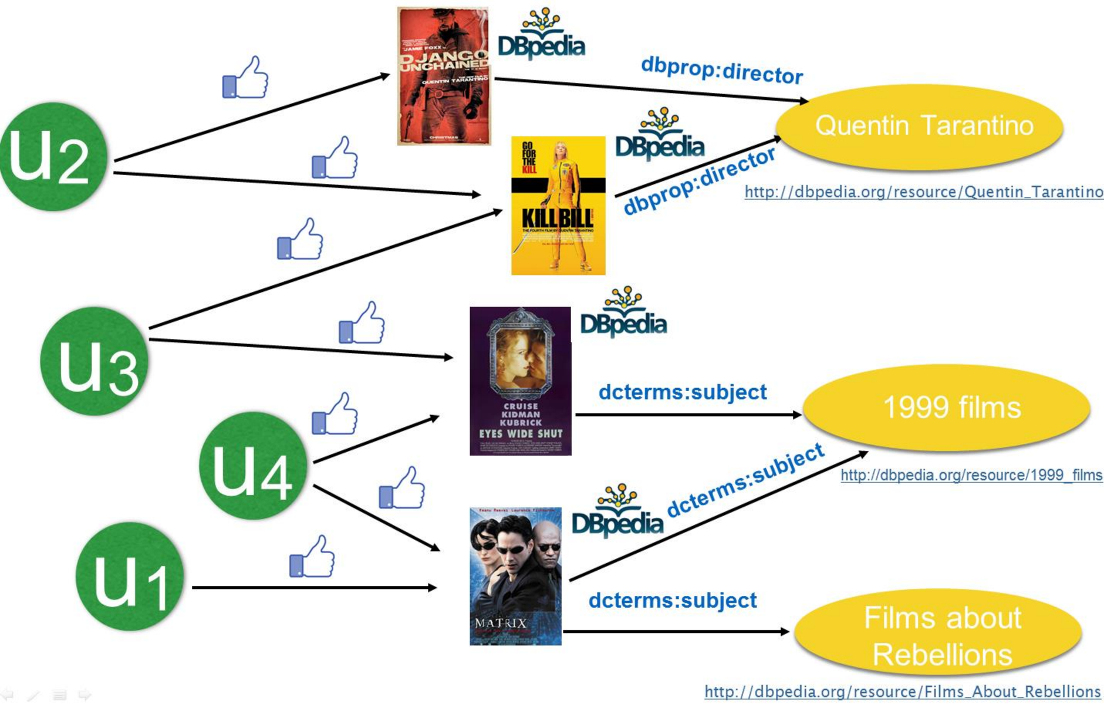
is it necessary to **inject all the properties** available in LOD cloud?

Graph-based RecSys



is it necessary to **inject all the properties** available in LOD cloud?

Graph-based RecSys



what are the most **promising properties** to include?

manual selection

- domain-specific properties
- most frequent properties
- …

automatic selection

- **more difficult** to implement

Feature selection

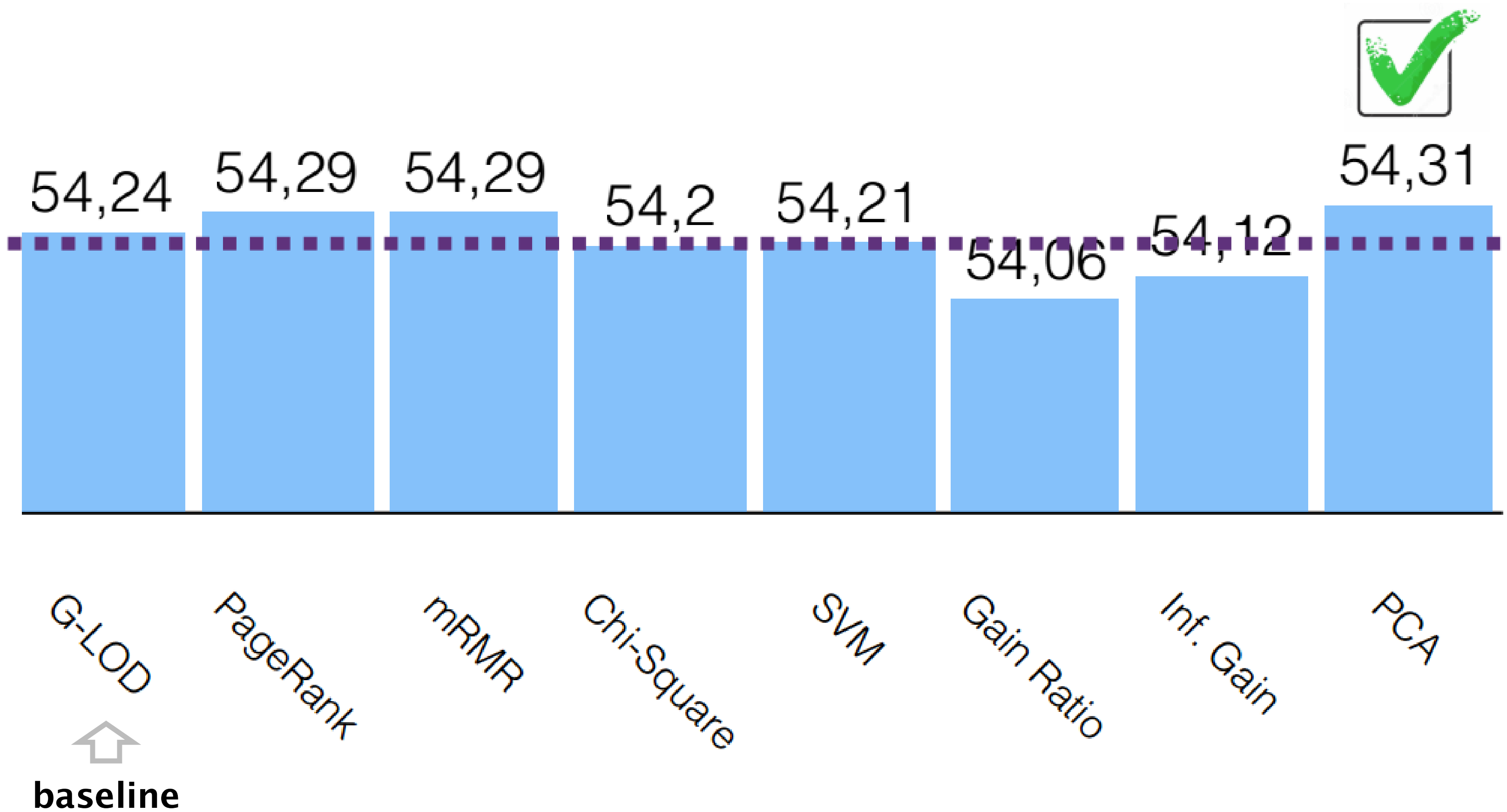
selecting the **most promising subset** of LOD-based properties

PageRank
Principal Component Analysis
Chi-square
Information Gain
Information Gain Ratio
Mininum Redundancy Maximum Relevance

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)

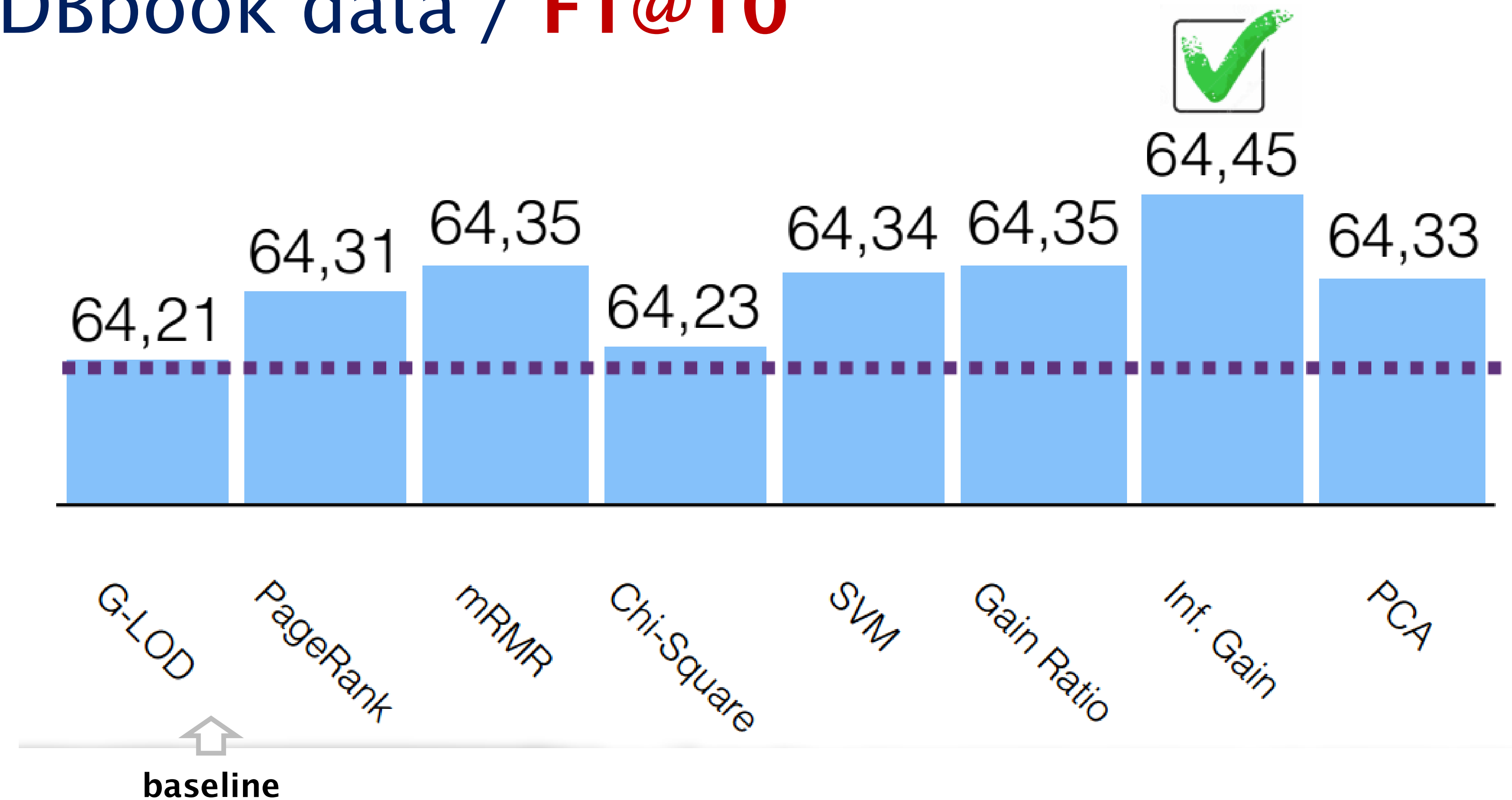
Graph-based RecSys

MovieLens data / F1@10



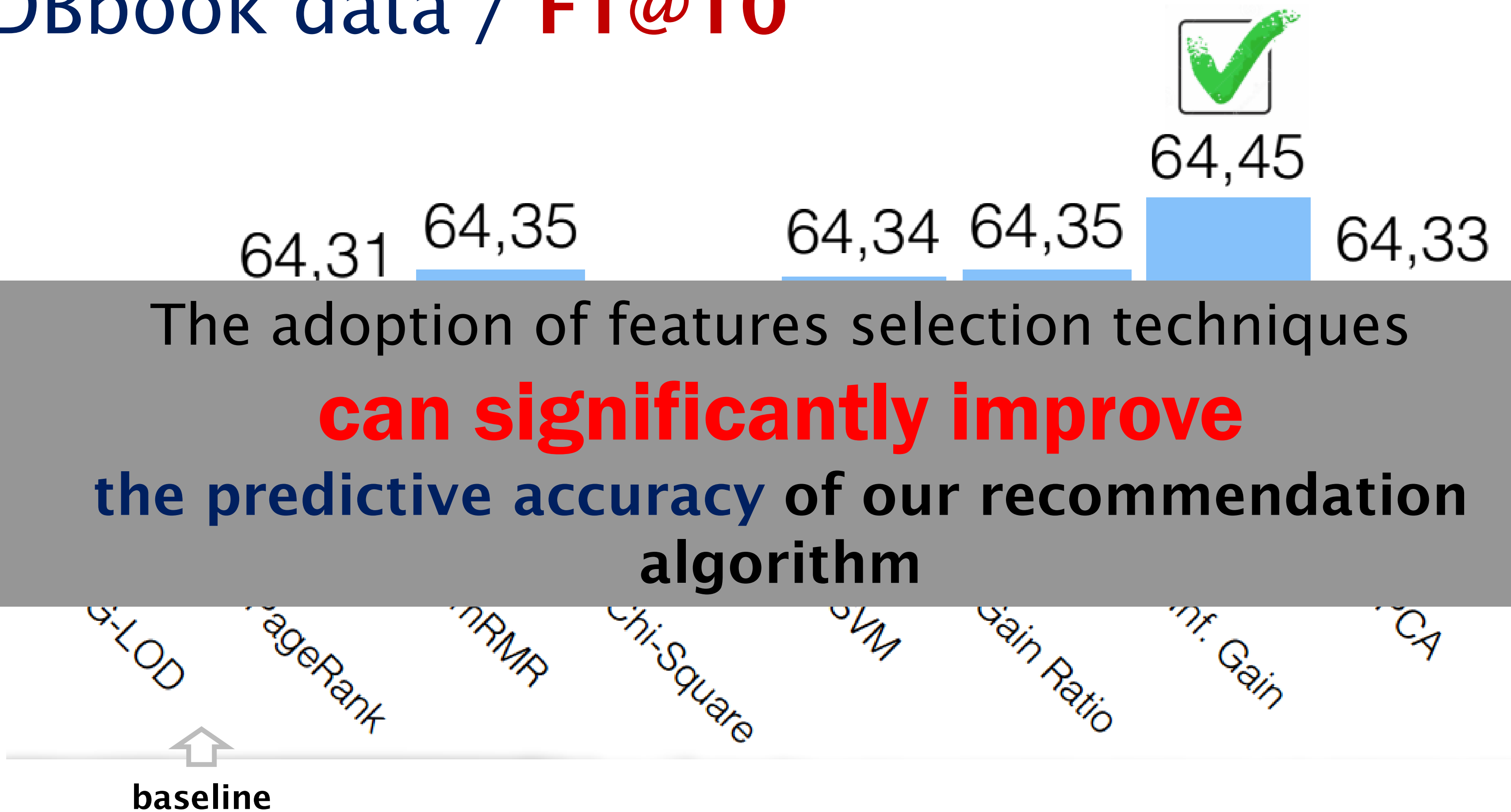
Graph-based RecSys

DBbook data / F1@10



Graph-based RecSys

DBbook data / **F1@10**



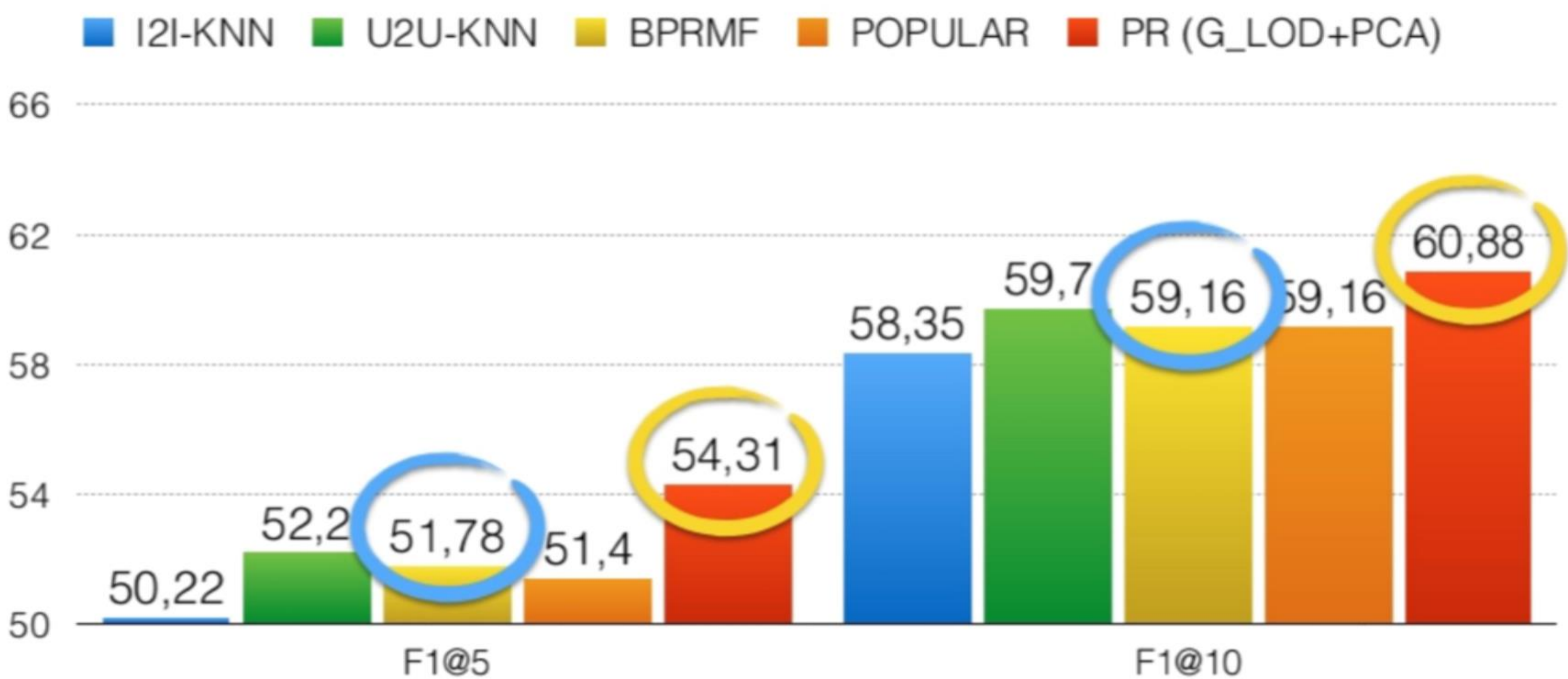
The adoption of features selection techniques

can significantly improve

the predictive accuracy of our recommendation algorithm

Graph-based RecSys

Comparison to state of the art
MovieLens 100K dataset

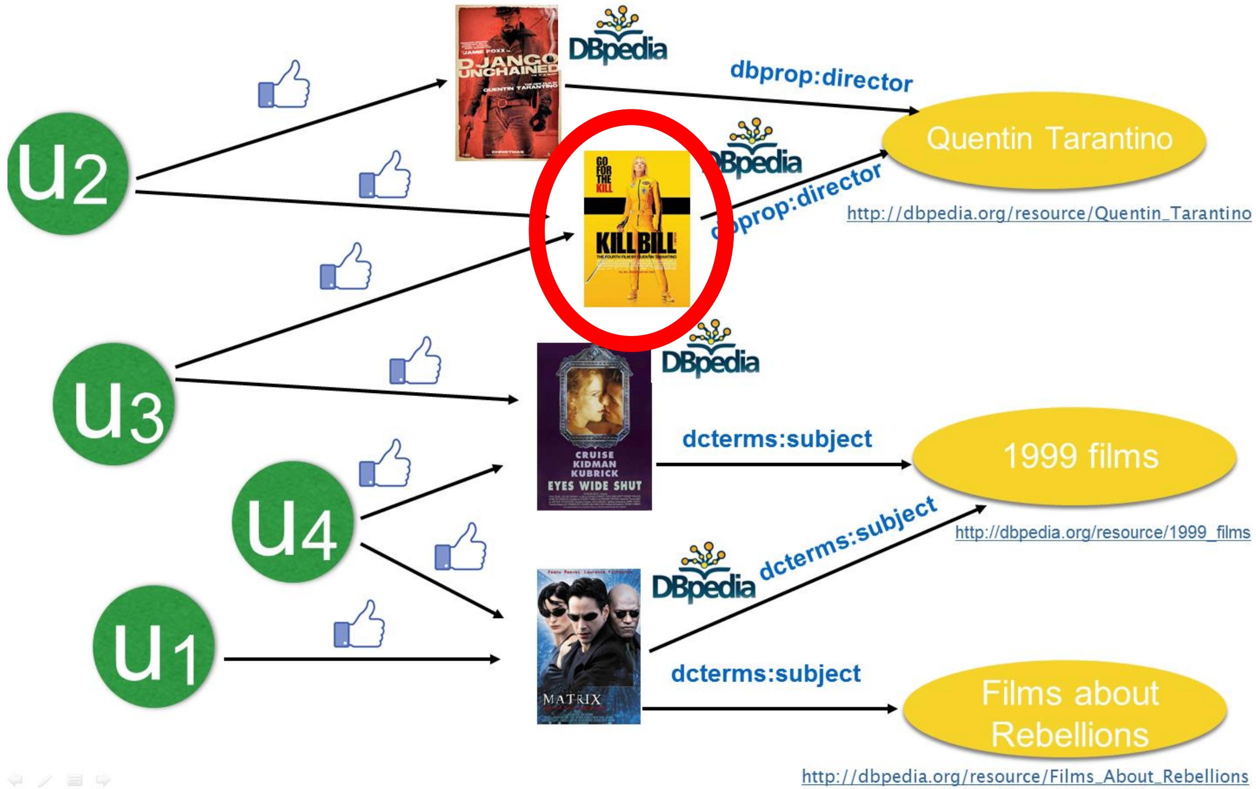


Even state-of-the-art approaches based **on Matrix Factorization are overcame** by our methodology

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 (accepted for publication), 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 (accepted for publication), 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**

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 (accepted for publication), 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

items \ users	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 (accepted for publication), 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

items \ users	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

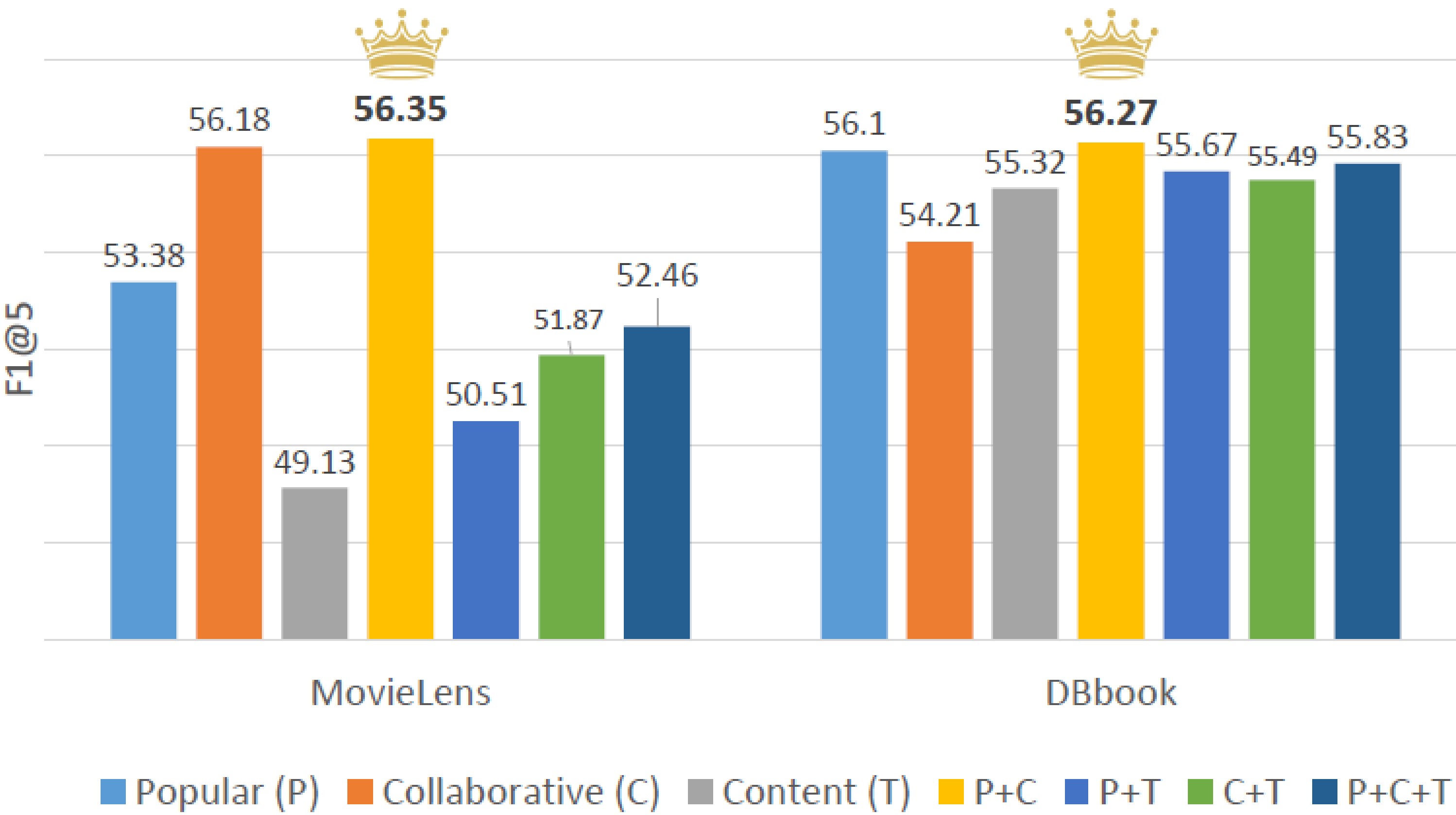
We used them to feed a hybrid classification framework

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

LOD-based Recommender Systems

Results

Experiment 1 – Performance of **basic features**

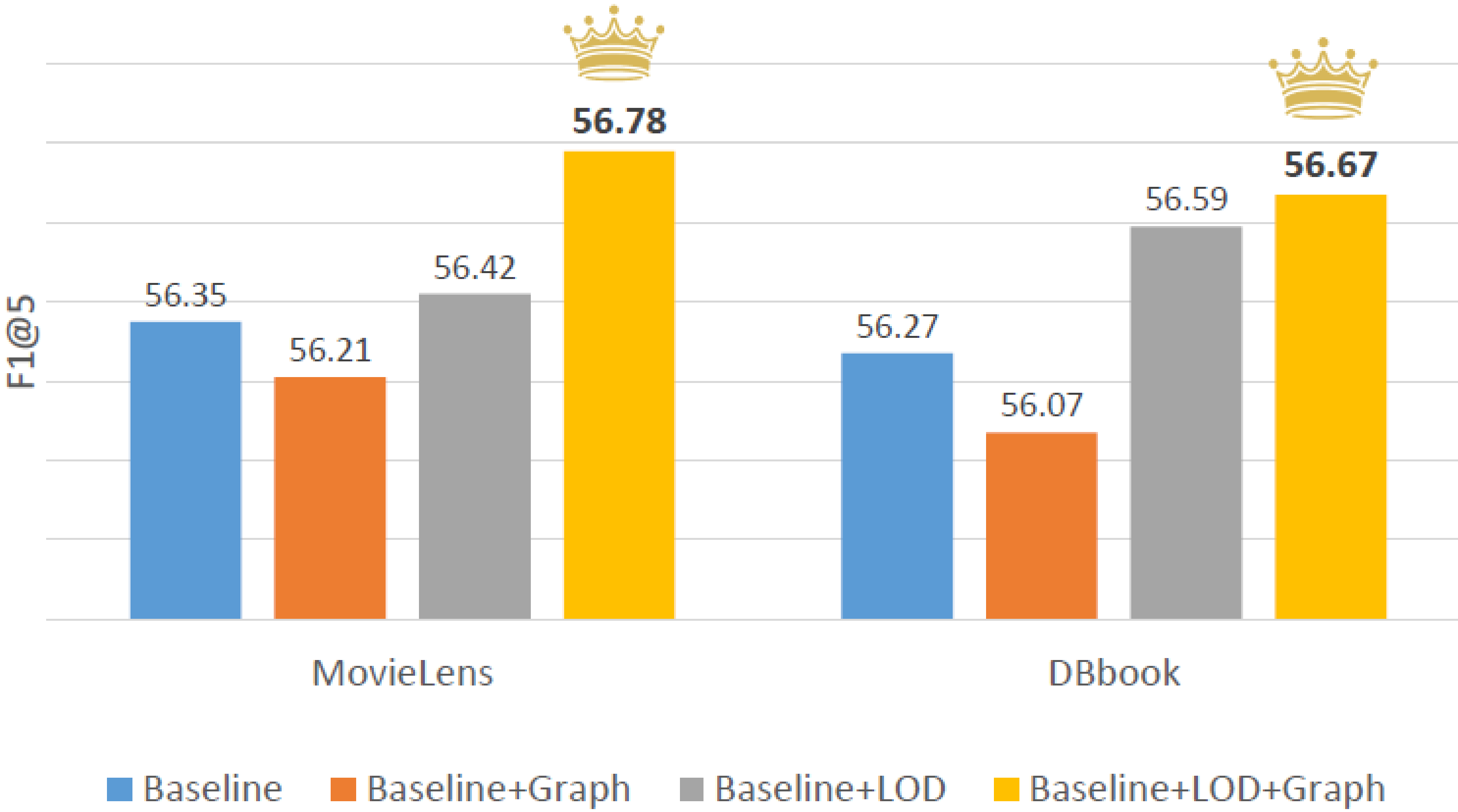


Cataldo Musto, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis. **Semantics-aware Recommender Systems Exploiting Linked Open Data and Graph-based Features.** *Knowledge-Based Systems (accepted for publication), 2017.*

LOD-based Recommender Systems

Results

Experiment 2 - Impact of **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 (accepted for publication), 2017.*

Recap #5

encoding exogenous semantics through **Knowledge Graphs**

1. Linked Open Data represent a huge data silos, which is freely available
2. They can easily let overcome the limited content analysis problem
3. They can enrich graph-based data model with interesting data points
4. They can feed machine learning models with new and relevant features
5. They improve the accuracy of recommender systems

Recent Developments of Content-Based RecSys

Applications: explanations, obviousness of recommendations

Marco de Gemmis

Department of Computer Science
University of Bari Aldo Moro, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

SWAP researchgroup

Agenda

Why?

Why do we need intelligent information access?
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
Serendipity in Recommender Systems

Explanatory aims

Aim	Description
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Persuasiveness	Convince users to try or buy
Trust	Increase users' confidence in the system
Effectiveness	Help users make good decisions
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of use or enjoyment

N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. UMUAI, 22(4-5):399{439, 2012.

Some Examples

"People who liked this movie also liked... "

TRANSPARENCY

"You might like this item because it won the Oscar"

"It is a funny comedy"

EFFECTIVENESS

EFFICIENCY

Explanation Strategies

❶ Preferences of similar users

- ✓ «customers who bought this item also bought...»

❷ Items similar to those in the user profile

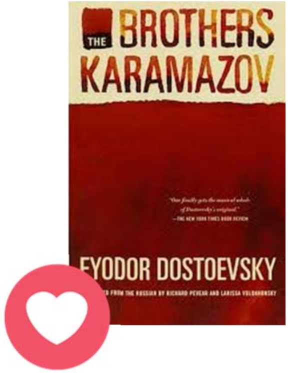
- ✓ «I recommend Star Trek because you liked Star Wars»

❸ Attributes of interest

- ✓ «You will like Forrest Gump because Tom Hanks is in the cast»

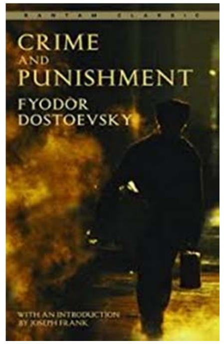
A detailed explanation for a book recommendation

User Profile



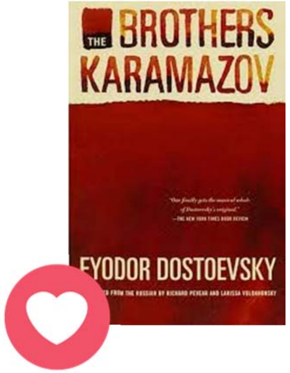
I suggest Crime and Punishment because you like books written by Fyodor Dostoevskij such as The Brothers Karamazov. Furthermore, you often like Psychological Russian Novels such as Anna Karenina and A hero of our time.

Recommendation



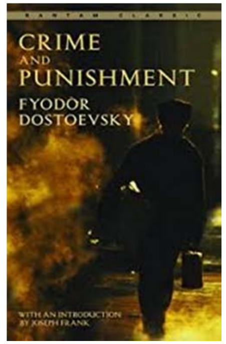
A detailed explanation for a book recommendation

User Profile



I suggest Crime and Punishment because you like books written by Fyodor Dostoevskij such as The Brothers Karamazov. Furthermore, you often like Psychological Russian Novels such as Anna Karenina and A hero of our time.

Recommendation

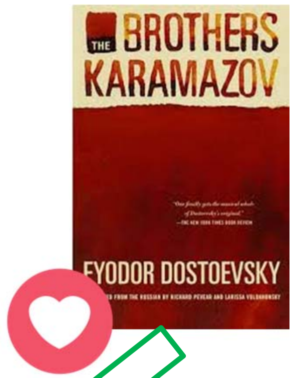


A Hybrid Explanation Strategy

❶ Preferences of similar users

- ✓ «customers who bought this item also bought...»

❷ Items similar to those in the user profile



I suggest Crime and Punishment because you like books written by Fyodor Dostoevskij such as The Brothers Karamazov.

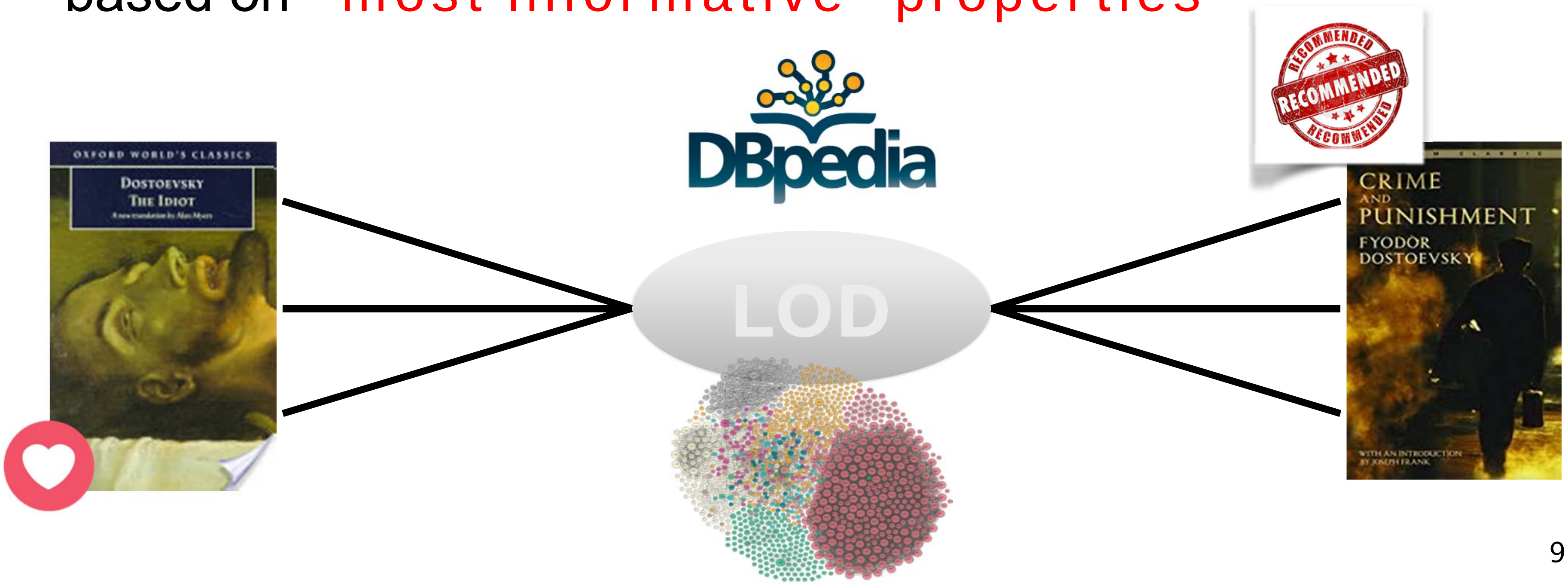
❸ Attributes of interest

I suggest Crime and Punishment because you like books written by Fyodor Dostoevskij such as The Brothers Karamazov.

Personalized explanation approach based on user preferences on items and their properties

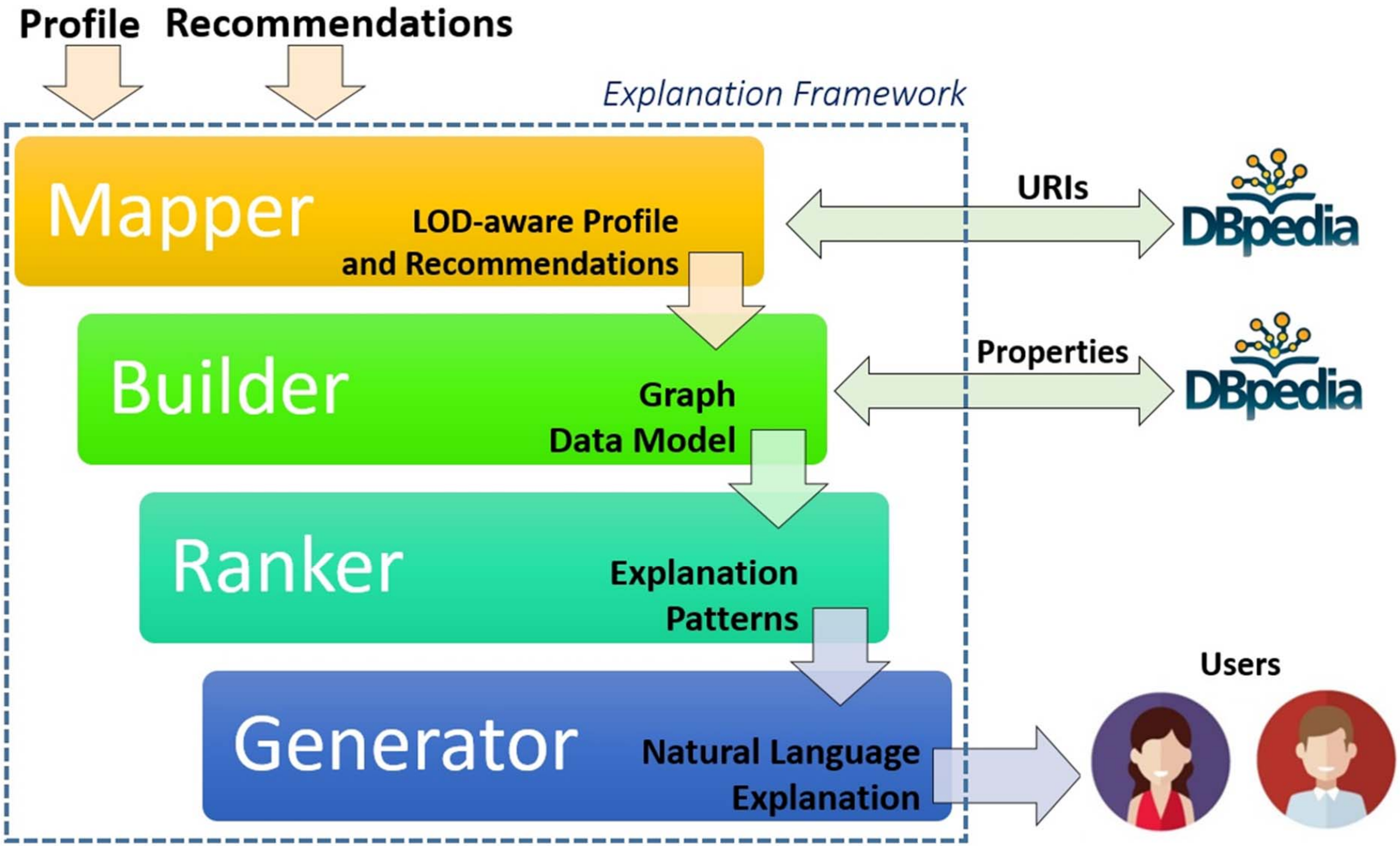
Explaining recommendations based on the Linked Open Data cloud

- connecting the items the user liked to the recommendations through properties in the LOD cloud
- generation of natural language explanations based on “most informative” properties



LOD-aware Representation

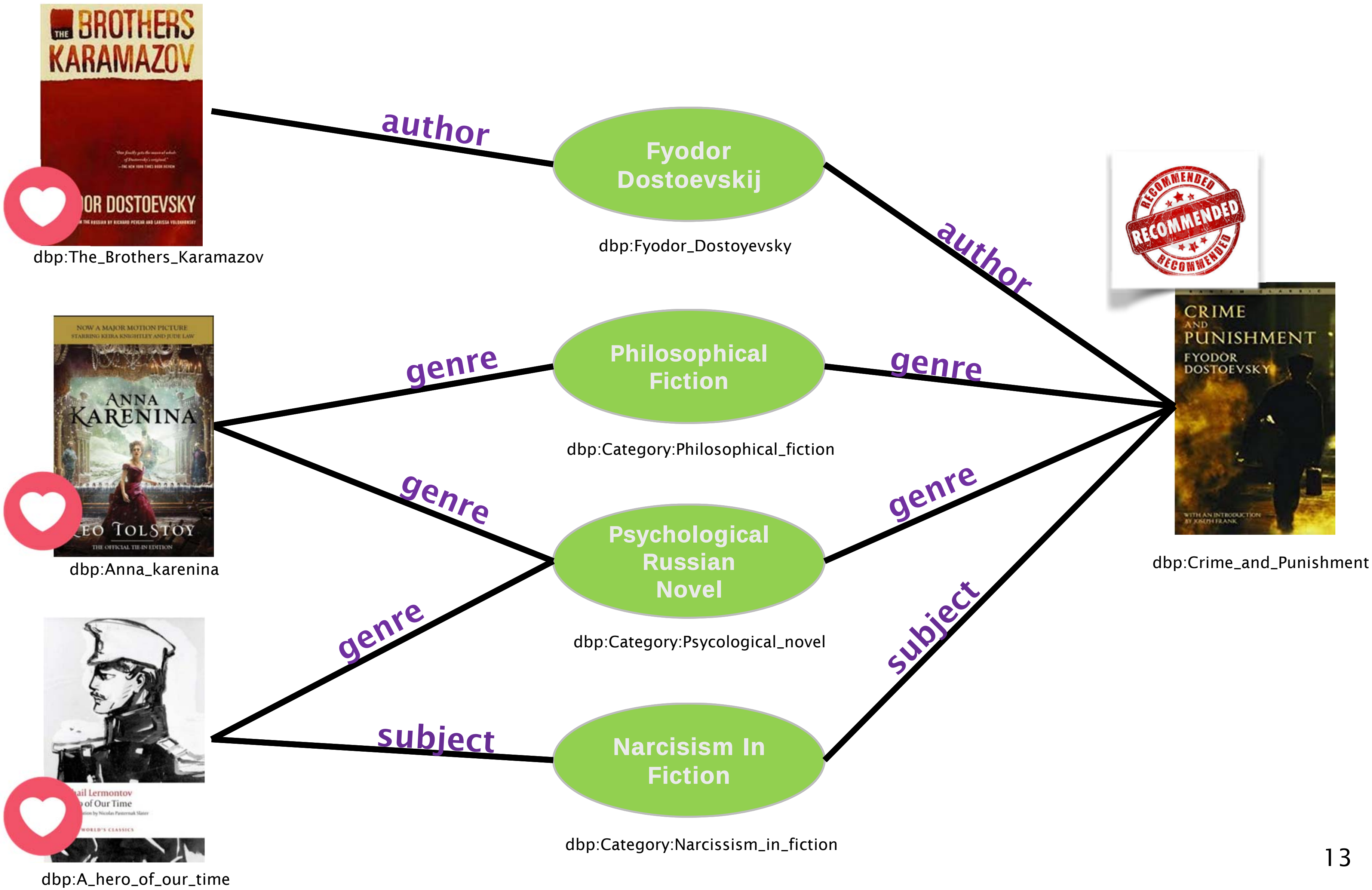
ExpLOD: Framework



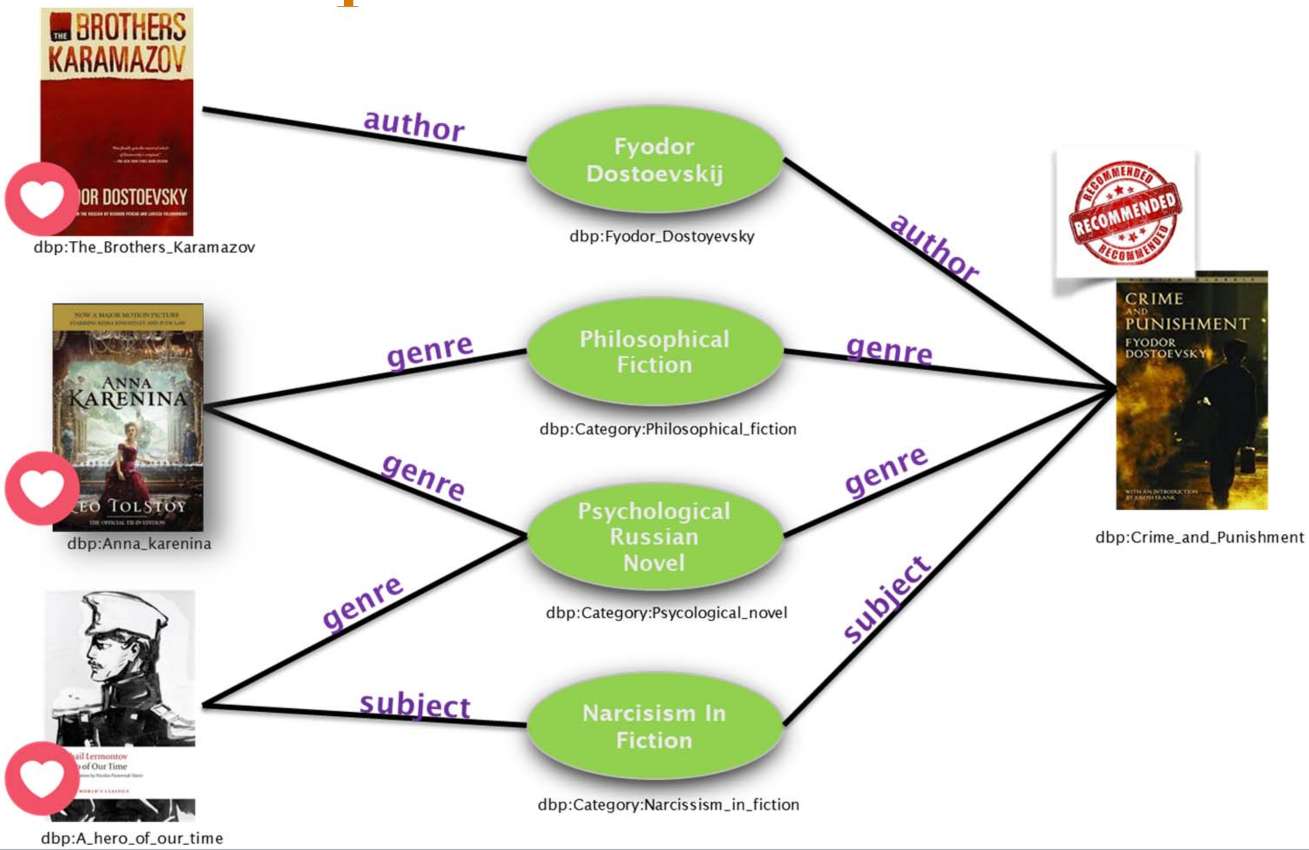
Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro: **ExpLOD: A Framework for Explaining Recommendations based on the Linked Open Data Cloud.** RecSys 2016: 151-154

ExpLOD: Mapper

ExpLOD: Builder



ExpLOD: Ranker



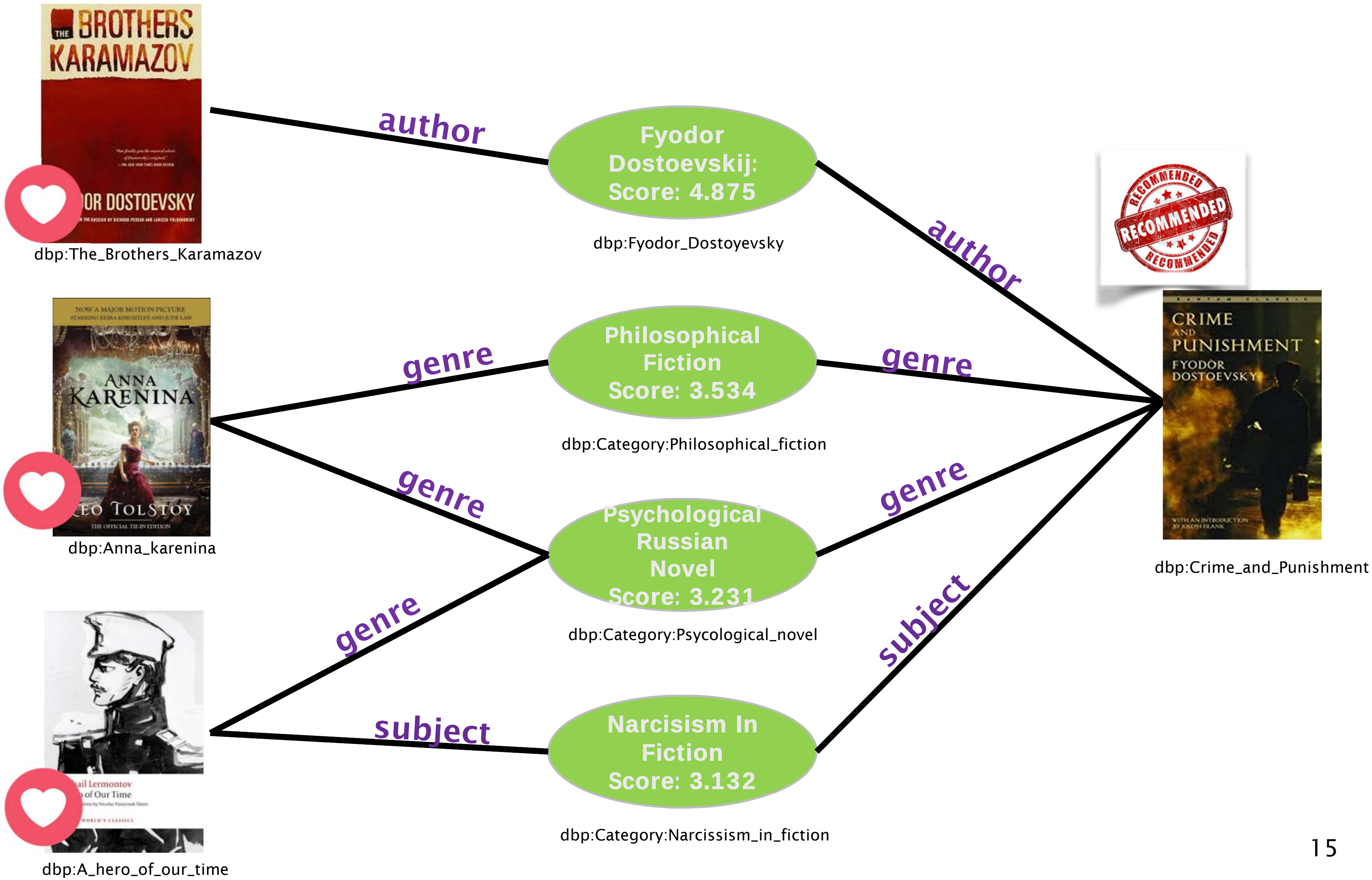
Scoring properties in ExpLOD

$$score(c, I_p, I_r) = (\alpha \frac{n_{c,I_p}}{|I_p|} + \beta \frac{n_{c,I_r}}{|I_r|}) * IDF_c$$

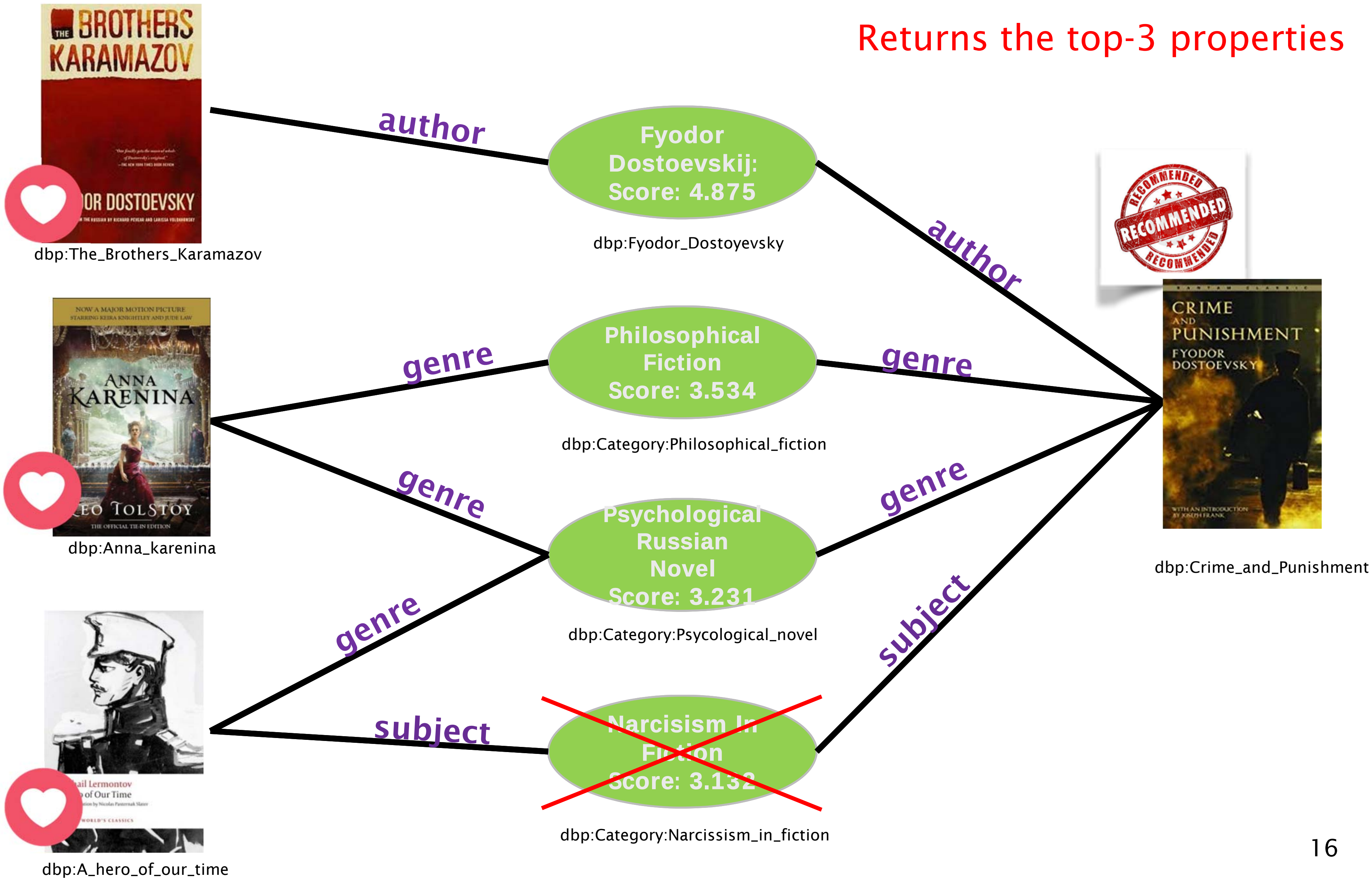
number of edges number of edges

higher score to uncommon properties highly connected to the items in both the **user profile** and the **recommendation list**

ExpLOD: Ranker



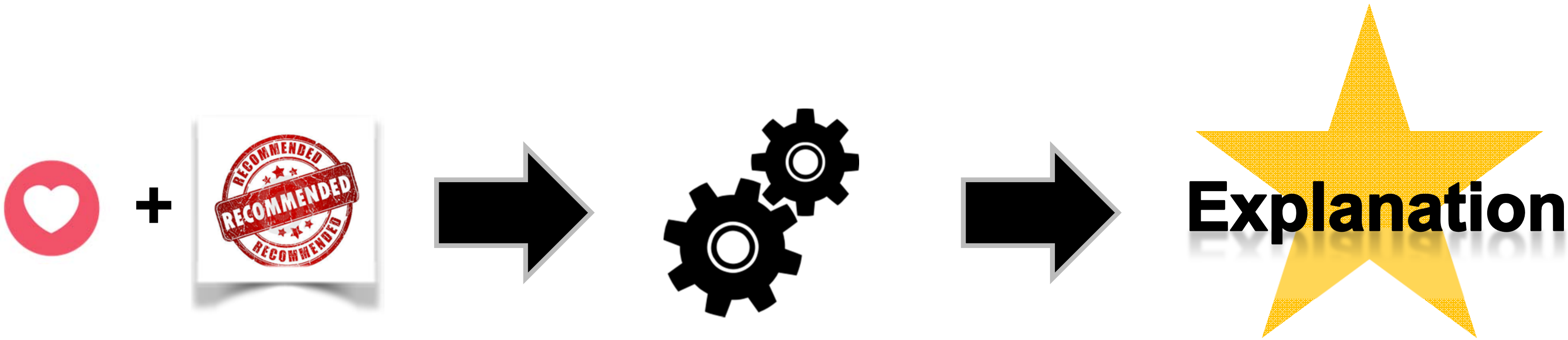
ExpLOD: Ranker



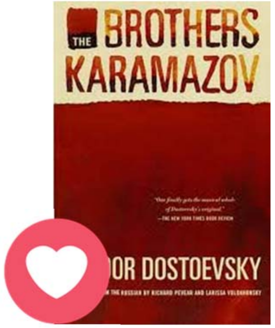
ExpLOD: Generator

Input:
- ✓ User Profile
- ✓ Recommended Items
- ✓ Top-k properties

Output:
- ✓ Natural Language Explanation



ExpLOD: Generator

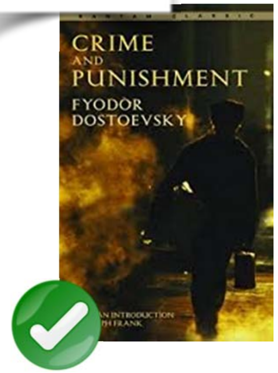


dbp:The_Brothers_Karamazov

dbp:Anna_karenina

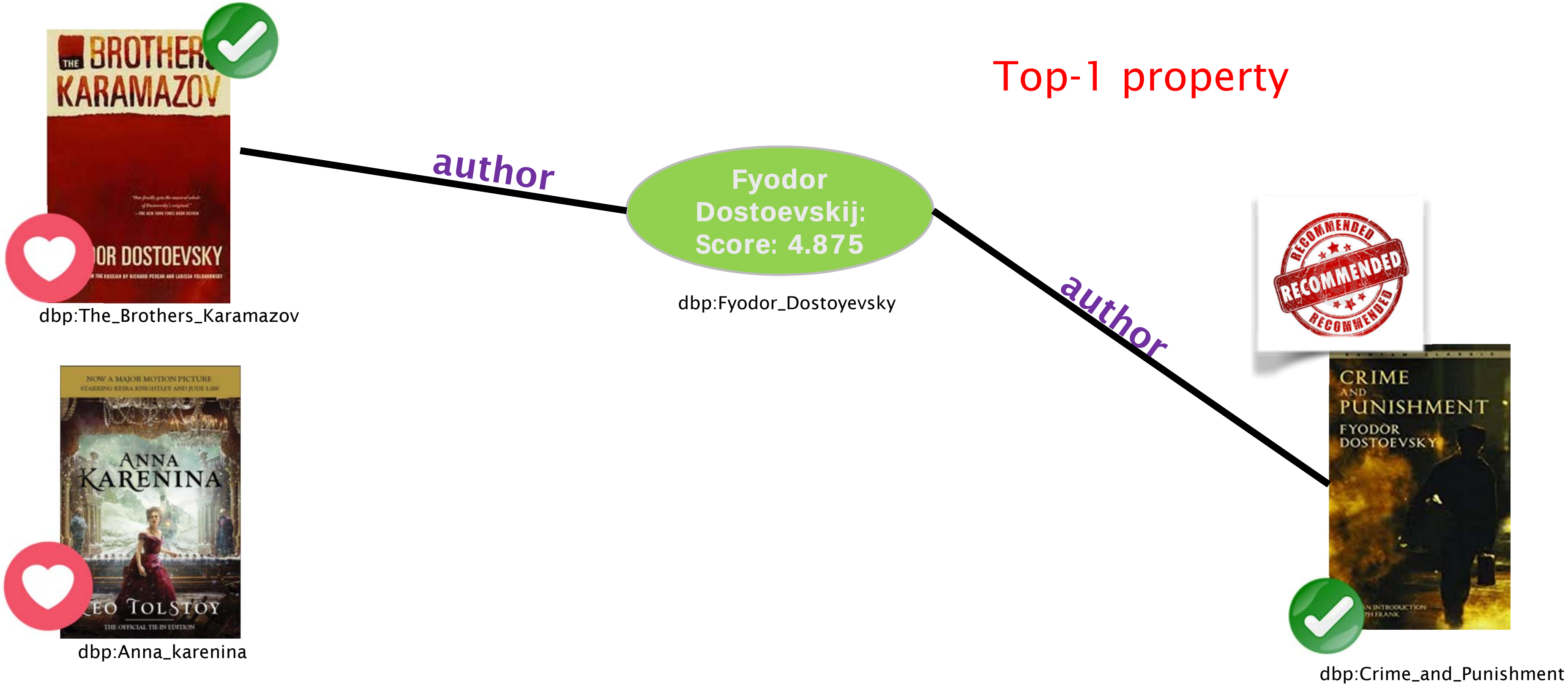
dbp:A_hero_of_our_time

I suggest **Crime and Punishment**...



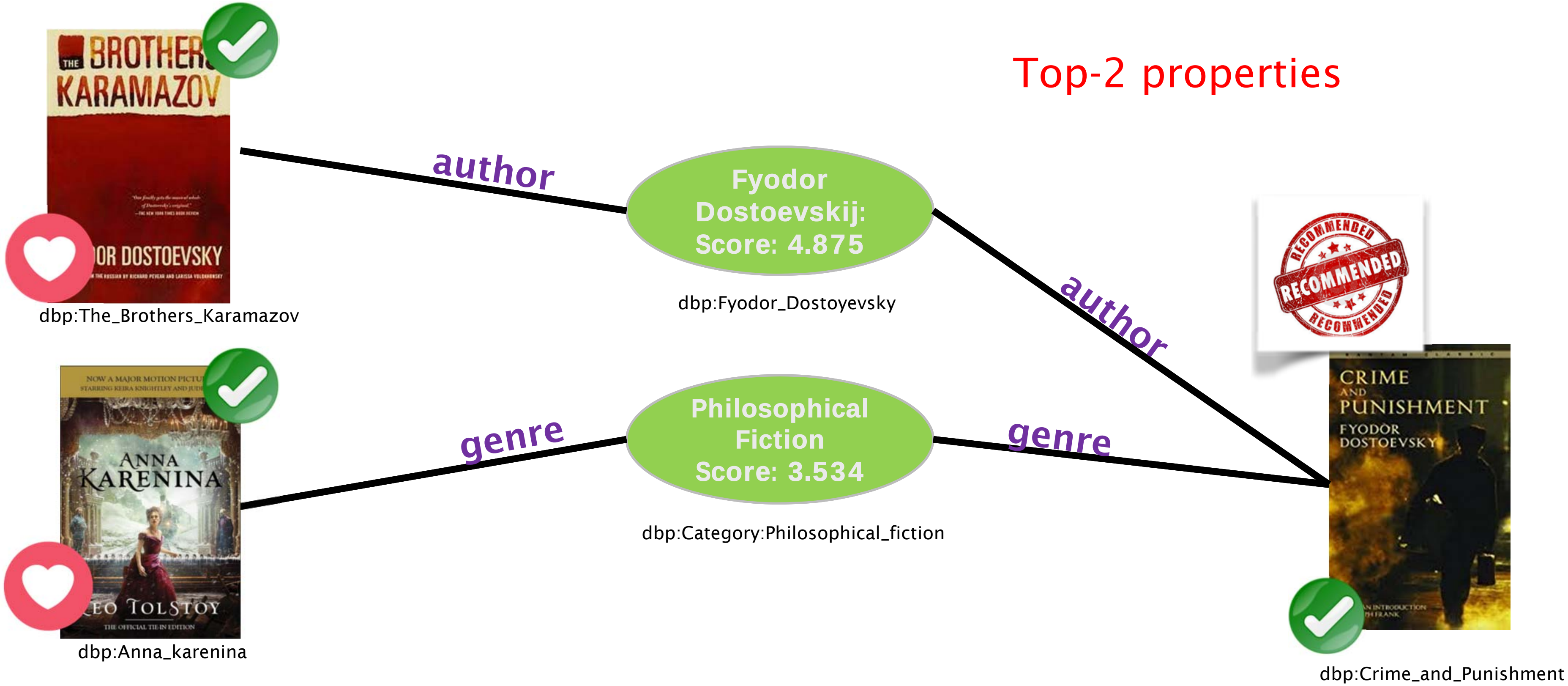
dbp:Crime_and_Punishment

ExpLOD: Generator



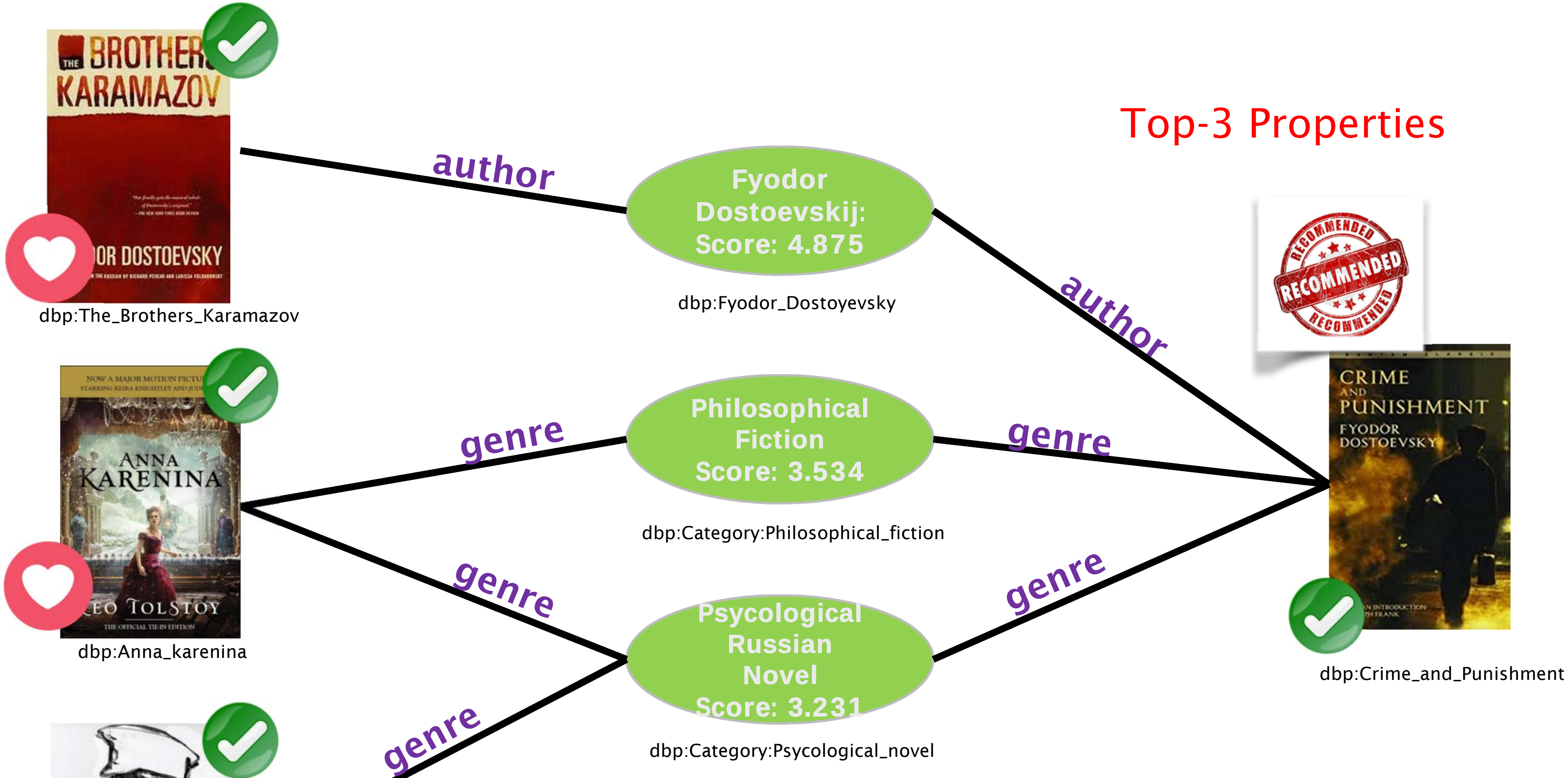
I suggest **Crime and Punishment** because you like books **written by Fyodor Dostoevskij** such as **The Brothers Karamazov**.

ExpLOD: Generator



I suggest **Crime and Punishment** because you like books **written by Fyodor Dostoevskij** such as **The Brothers Karamazov**. Furthermore, you like **Philosophical Fiction**, such as **Anna Karenina**.

ExpLOD: Generator



I suggest **Crime and Punishment** because you like books **written by Fyodor Dostoevskij** such as **The Brothers Karamazov**. Furthermore, you like **Philosophical Fiction**, such as **Anna Karenina**. Finally, you **often** like **Psychological Russian Novel**, such as **Anna Karenina** and **A hero of our time**.

Experimental Evaluation

user study

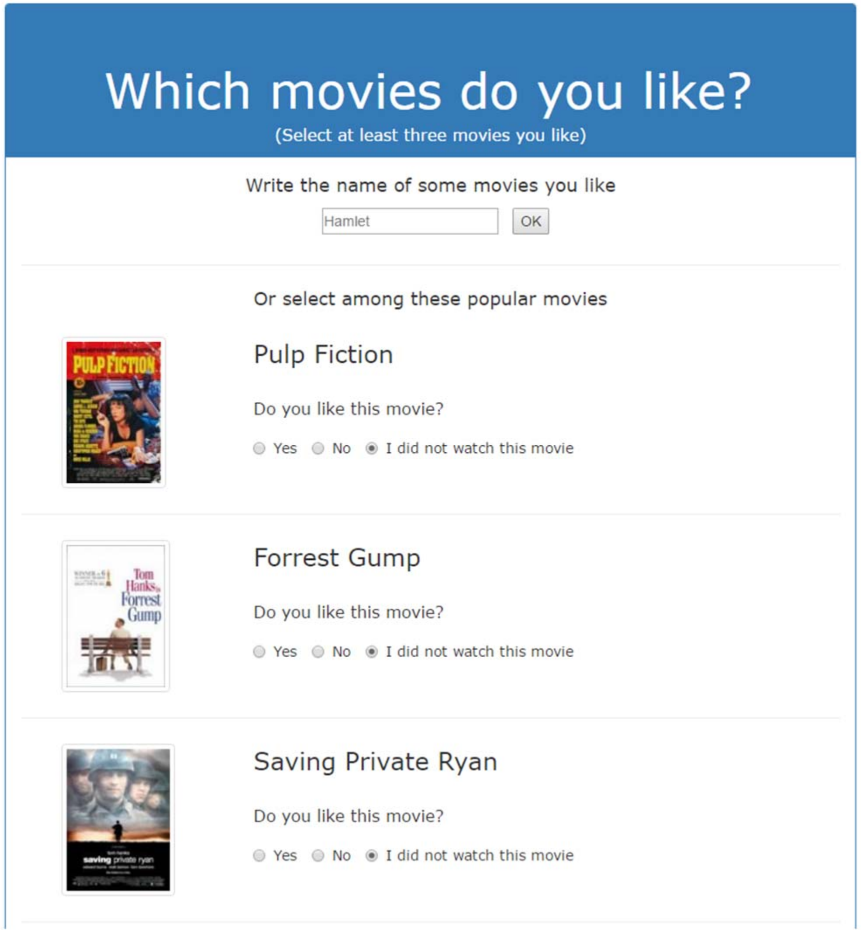
- ✓ movie domain, 308 users involved
- ✓ **protocol:**
  - Web Application → Building User Profiles → Recommendations + Explanations → Questionnaire + Ex-post Evaluation
- ✓ **explanation aims**
  - Transparency, Engagement, Persuasion, Trust, Effectiveness

three configurations compared

- ✓ popularity-based explanation (baseline)
- ✓ non-personalized explanation based on LOD
- ✓ EXPLOD

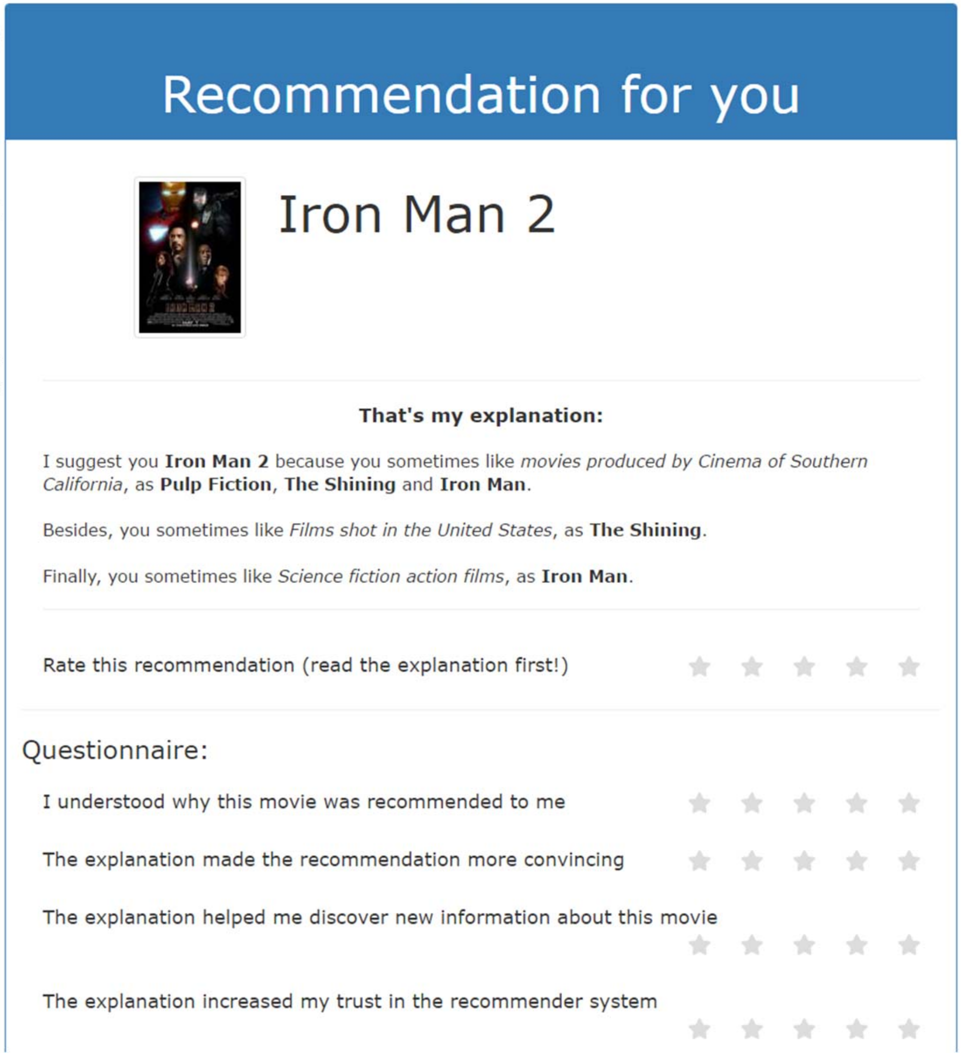
Experimental Evaluation: A User Study

- Gathering movie preferences
  - ✓ 308 users rated 20 movies randomly chosen from the most popular movies in IMDB



Experimental Evaluation: A User Study

- Recommendation and e
  - ✓ 1 recommendation per PageRank + explanatio
  - ✓ the user read the expla 5 statements to evaluat engagement, trust, effe



T. H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Trans. Knowl. Data Eng., 15(4):784-796, 2003.

Experimental Evaluation: A User Study

That's my explanation:

I suggest you **Iron Man 2** because you sometimes like *movies produced by Cinema of Southern California*, as **Pulp Fiction**, **The Shining** and **Iron Man**.

Besides, you sometimes like *Films shot in the United States*, as **The Shining**.

Finally, you sometimes like *Science fiction action films*, as **Iron Man**.

Rate this recommendation (read the explanation first!)

Questionnaire:

I understood why this movie was recommended to me

The explanation made the recommendation more convincing

The explanation helped me discover new information about this movie

The explanation increased my trust in the recommender system

Experimental Evaluation: A User Study

That's my explanation:

I suggest you **Iron Man 2** because you sometimes like *movies produced by Cinema of Southern California*, as **Pulp Fiction**, **The Shining** and **Iron Man**.

Besides, you sometimes like *Films shot in the United States*, as **The Shining**.

Finally, you sometimes like *Science fiction action films*, as **Iron Man**.

Rate this recommendation (read the explanation first!)

effectiveness

Questionnaire:

I understood why this movie was recommended to me

transparency

The explanation made the recommendation more convincing

persuasiveness

The explanation helped me discover new information about this movie

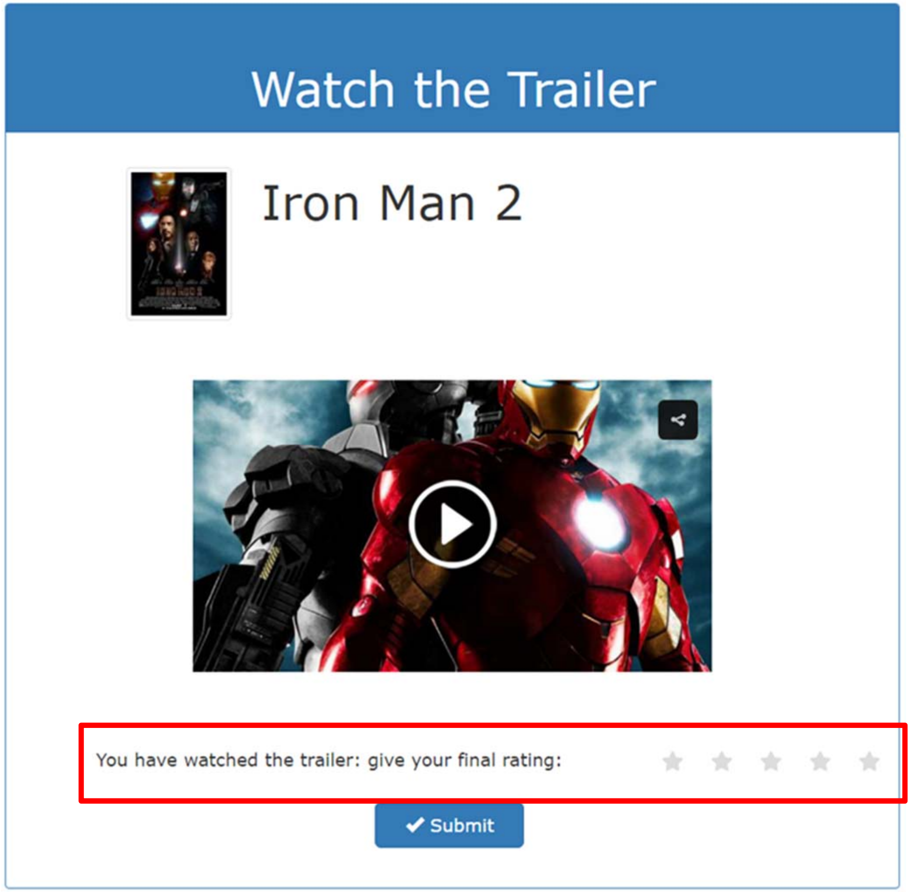
engagement

The explanation increased my trust in the recommender system

trust

Experimental Evaluation: A User Study

- Ex-post evaluation
  - ✓ the user watched the trailer of the recommended movie and provided again a 5-star rating



Experimental Evaluation: A User Study

- Three explanation strategies compared
  - ✓ ExpLOD
  - ✓ popularity-based explanation (baseline) → "We suggest this item since it is very popular among people who like the same movies as you"
  - ✓ non-personalized explanation based on LOD → movie properties extracted from Dbpedia (without any filtering or ranking of properties)

Experimental Evaluation: A User Study

- Gathering movie preferences
  - ✓ 308 users rated 20 movies randomly chosen from the most popular movies in IMDB
  - ✓ 1 recommendation per user computed by Personalized PageRank + explanation
- Evaluation of Explanations
  - ✓ the user read the explanation and provided a 5-star rating on 5 statements to evaluate transparency, persuasion, engagement, trust, effectiveness
  - ✓ Ex-post evaluation: the user watched the trailer of the recommended movie and provided again a 5-star rating

T. H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Trans. Knowl. Data Eng., 15(4):784{796, 2003.

Results

Aim		Statement
Transparency	✓	I understood why this movie was recommended to me
Persuasiveness	✓	The explanation made the recommendation more convincing
Engagement	✓	The explanation helped me discover new information about this movie
Trust	✓	The explanation increased my trust in the recommender system
Effectiveness	✓	I like this recommendation

EXPLOD compared to:

- ✓ popularity-based explanation → "We suggest this item since it is very popular among people who like the same movies as you"
- ✓ non-personalized explanation style → movie properties extracted from DBpedia

C. Musto, F. Narducci, P. Lops, M. de Gemmis, G. Semeraro: ExpLOD: a framework for Explaining Recommendations based on the Linked Open Data cloud. Proc. ACM RecSys 2016.

Explanations - Results

	ExpLOD	non-personalized	baseline (pop)
transparency	**4.18**	3.04	3.01
persuasion	**3.41**	2.84	2.59
engagement	**3.48**	3.28	2.31
trust	**3.39**	2.81	2.67
effectiveness	0.72	**0.66**	0.93

Results – Main findings

	ExpLOD	non-personalized	baseline (pop)
Transparency*	**4.18**	3.04	3.01
Persuasion*	**3.41**	2.84	2.59
Engagement*	**3.48**	3.28	2.31
Trust*	**3.39**	2.81	2.67
Effectiveness**	0.72	**0.66**	0.93

* average score collected through the user questionnaires
** difference between the pre- and post-trailer ratings

significant improvement in 4 out of **5 metrics**

non-significant gaps in terms of **effectiveness**

C. Musto, F. Narducci, P. Lops, M. de Gemmis, G. Semeraro: ExpLOD: a framework for Explaining Recommendations based on the Linked Open Data cloud. In Proc. of the 10th ACM Conference on Recommender Systems (RecSys '16). ACM, New York, NY, USA, 151-154.

Explanations - Results

Aim	Question		
transparency	I understood why this movie was recommended to me	topic director	distributor composer
persuasion	The explanation made the recommendation more convincing	awards director	location producer
engagement	The explanation helped me discover new information	writer director	producer distributor
trust	The explanation increased my trust in the recommender system	awards composer	producer topic
effectiveness	I like this recommendation	director writer	location composer

Agenda

Why?

Why do we need intelligent information access?
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

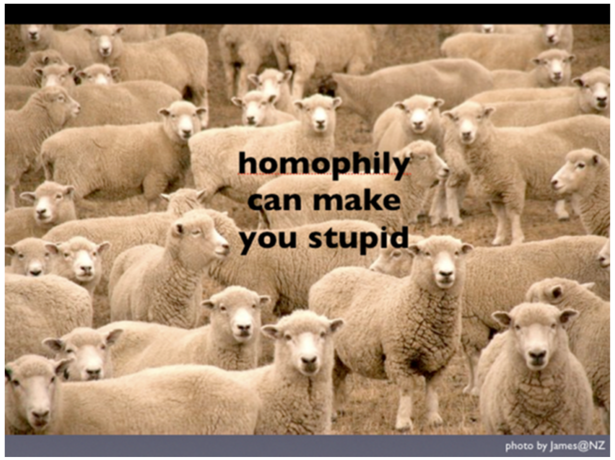
What?

Explanation of Recommendations
Serendipity in Recommender Systems

Obviousness of Recommendations: Homophily

- The tendency to surround ourselves by like-minded people [E. Zuckerman 2008]
- The *filter bubble* [Pariser 2011]
  - ✓ the user is provided with items within her existing range of interests
  - ✓ cultural impoverishment: “it’s possible to miss huge trends, changes and opportunities by talking solely to people who agree with you”

You are here!



[E. Zuckerman 2008] E. Zuckerman. *Homophily, serendipity, xenophilia*. April 25, 2008. www.ethanzuckerman.com/blog/2008/04/25/homophily-serendipity-xenophilia/]

[Pariser 2011] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, May 2011

Homophily in the digital world

- In the physical world, one of the strongest sources of homophily is **locality**, due to geographic proximity, family ties, and organizational factors (school, work, etc.)
- In the digital world, physical locality is less important. Other factors, such as **common interests**, might play a central role

2 main questions

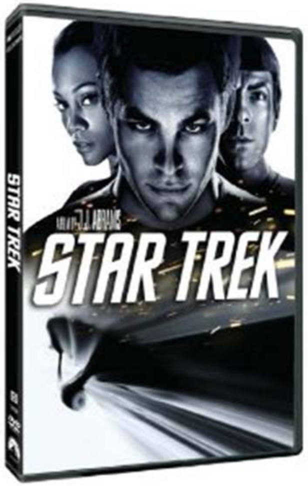
1. Are two users more likely to be friends if they share common interests?
2. Are two users more likely to share common interests if they are friends?

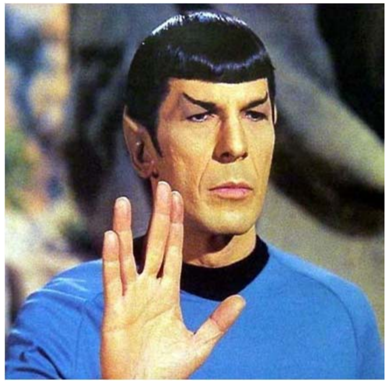
The answer to both questions is **YES** [Lauw et al. 2010]

[Lauw et al. 2010] Lauw, H.W., Schafer, J.C., Agrawal, R., & A. Ntoulas. Homophily in the Digital World: A LiveJournal Case Study. *IEEE Internet Computing* 14(2):15-23, March-April 2010.

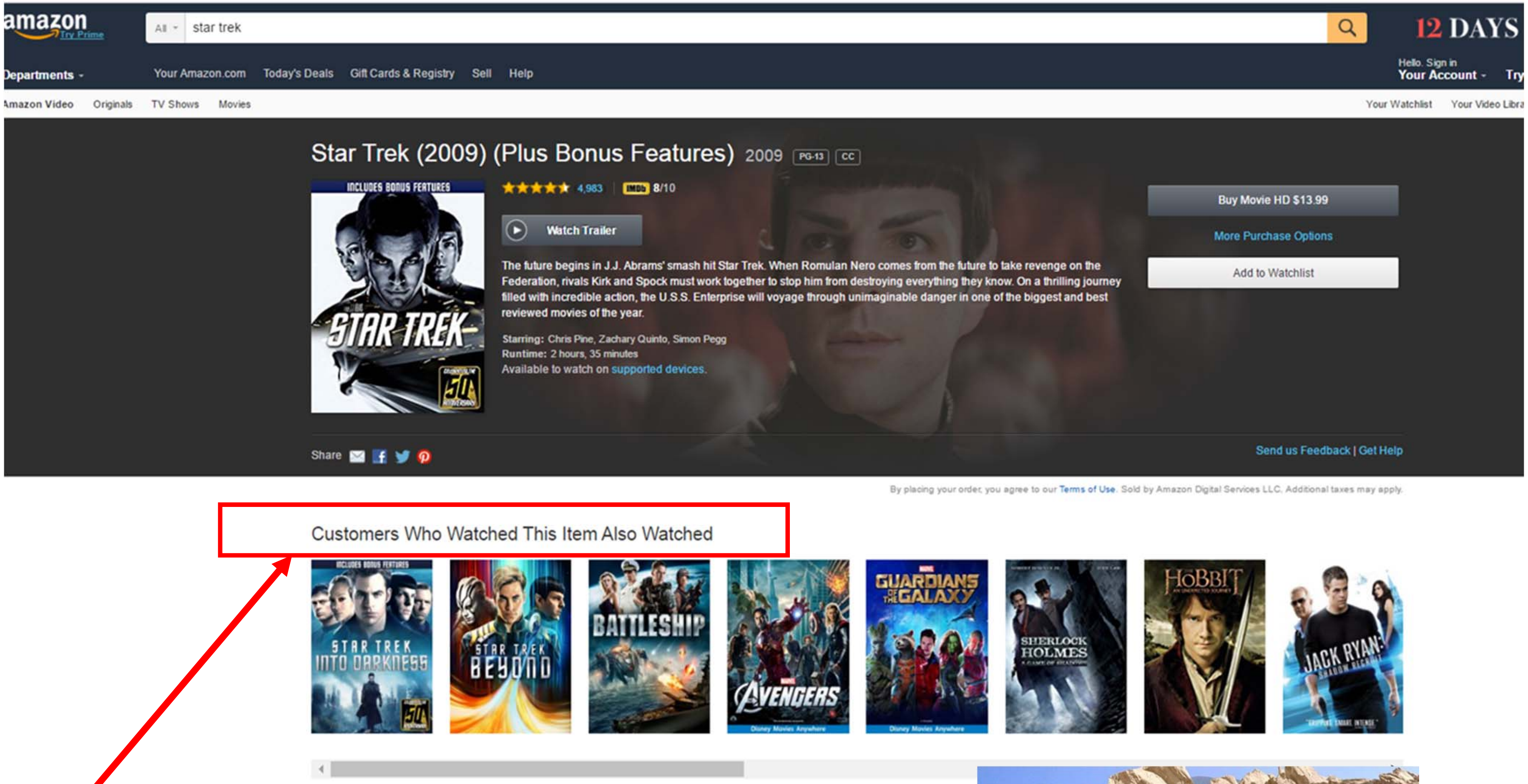
The homophily trap

- Does homophily hurt RecSys?
  - ✓ try to tell Amazon that you liked “Star Trek”…





The Homophily Trap: User-User



Recommendations by similar customers

Serendipitous recommendations

- *"Suggestions which help the user to find* ***surprisingly interesting*** *items she might not have discovered by herself"* [Herlocker et al. 2004]
  - ✓ Both *attractive* and *unexpected*

- *"The experience of receiving an unexpected and fortuitous item recommendation"* [McNee et al. 2006]

- Surprise or unexpectedness defined with respect to a benchmark model that generates expected recommendations [Ge10]

[Herlocker et al. 2004] Herlocker, L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22(1): 5-53, 2004.

[McNee et al. 2006] S.M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, 1097-1101, ACM, New York, NY, USA, 2006.

[Ge10] Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. Proc. of the ACM Conference on Recommender Systems, pp. 257–260. ACM (2010)

Operationally induced serendipity

- How to introduce **serendipity** in the recommendation process?
- Semantic matching is not a solution
  - ✓ Semantic profiles might provide more accurate recommendations than keyword-based profiles but they could be obvious too

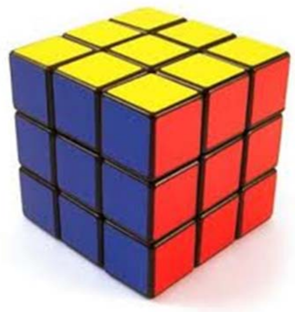
Serendipity in Information Seeking

- Information seeking metaphor investigated in literature (Toms 2000, André et al 2009, Bordino et al. 2013)
- Toms suggests 4 strategies
  - ✓ *Blind luck* or *"role of chance"* → random
  - ✓ *Pasteur Principle* or *"chance favors only the prepared mind"* → flashes of insight don't just happen, but they are the products of a "prepared mind"
  - ✓ *Anomalies and exceptions* or *"searching for dissimilarities"* → identification of items dissimilar to those the user liked in the past
  - ✓ *Reasoning by analogy* → abstraction mechanism allowing the system to discover the applicability of an existing schema to a new situation

(Toms 2000) E. Toms. Serendipitous Information Retrieval. *Proc.1st DELOS NoE Workshop on Information Seeking, Searching and Querying in Digital Libraries*, Zurich, Switzerland: ERCIM, 2000.
(André 2009) P. André, J. Teevan, S.T. Dumais. From x-rays to silly putty via Uranus: serendipity and its role in web search. Proc. ACM CHI 2009, ACM, New York, NY, USA, 2009,
(Bordino et al. 2013) I. Bordino, Y. Mejova, M. Lalmas, Penguins in sweaters, or serendipitous entity search on user-generated content. *Proc.22nd ACM CIKM 2013*, ACM, New York, NY, USA, 2013, pp. 109-118.

Operationally induced serendipity

- How to introduce serendipity in the recommendation process?

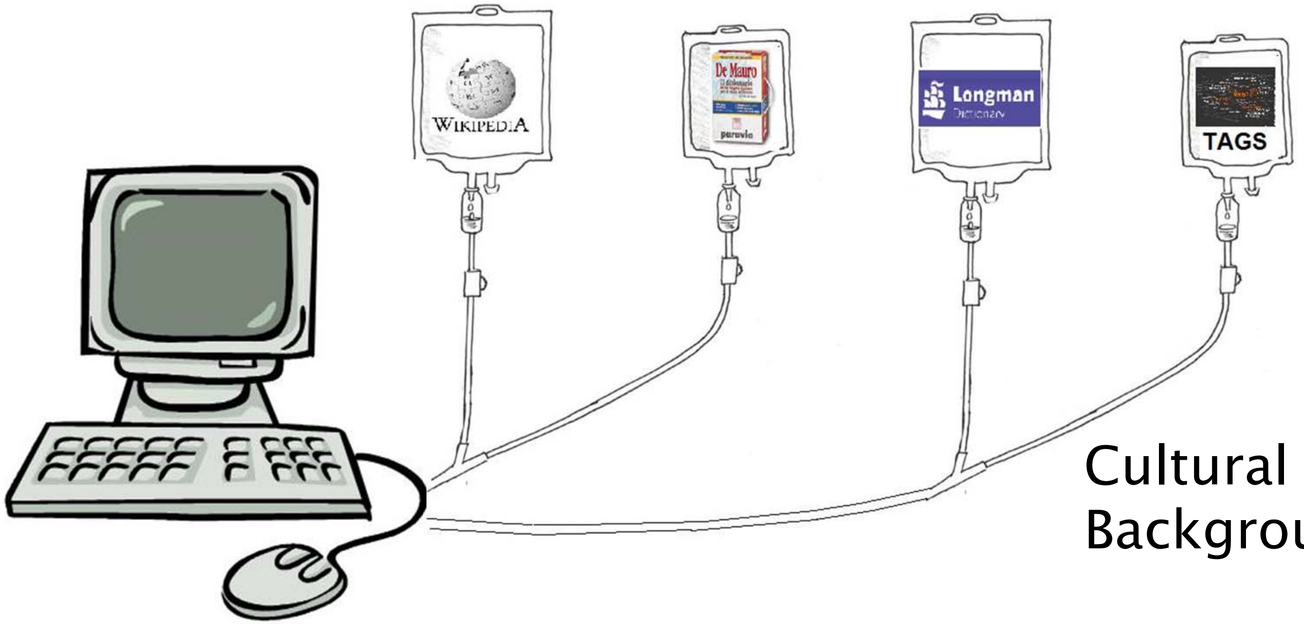


- Build a "prepared mind"!
  - ✓ Need some background knowledge → deep understanding of item descriptions
  - ✓ Need some reasoning capabilities → discovering non-obvious associations among items

Deep Content Analytics ⟺ **Semantics + Reasoning**

Knowledge Infusion: NLP+AI

- NLP techniques process the unstructured information stored in several (open) knowledge sources
  - ✓ The memory of the system
- Spreading Activation [And83] as the reasoning mechanism
  - ✓ The brain of the system



Cultural and Linguistic
Background Knowledge

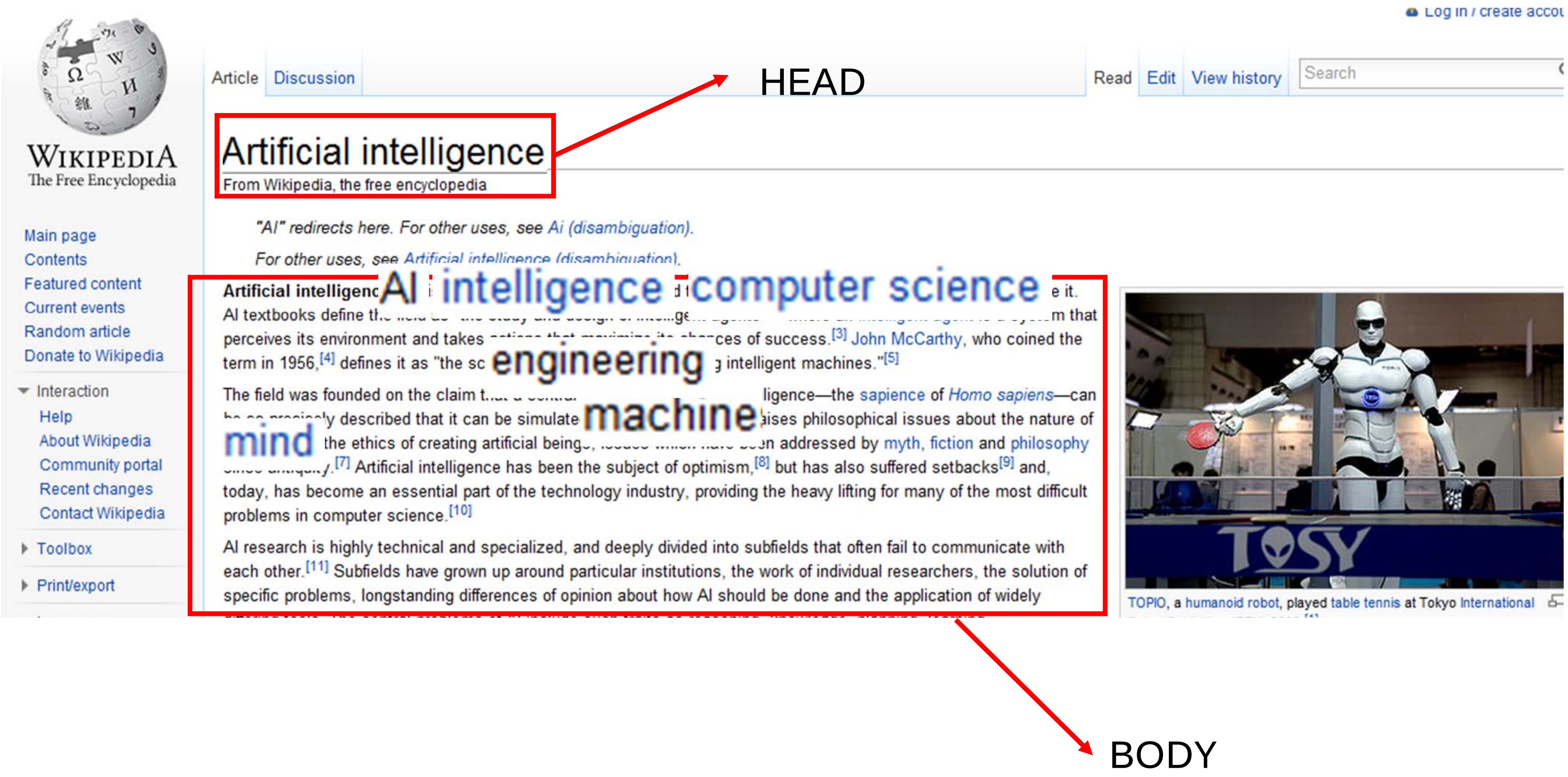
[And83] J. R. Anderson. A Spreading Activation Theory of Memory. Journal of Verbal Learning and Verbal Behavior, 22:261-295, 1983.

The Memory: Encoding Knowledge Sources as a CU Repository

- Information in long term memory of human beings encoded as Cognitive Units – ACT theory [And83]
- Cognitive Unit (CU) = textual description of a concept
  - ✓ HEAD = words identifying the concept represented by the CU
  - ✓ BODY = words describing the concept
  - ✓ [HEAD | BODY]

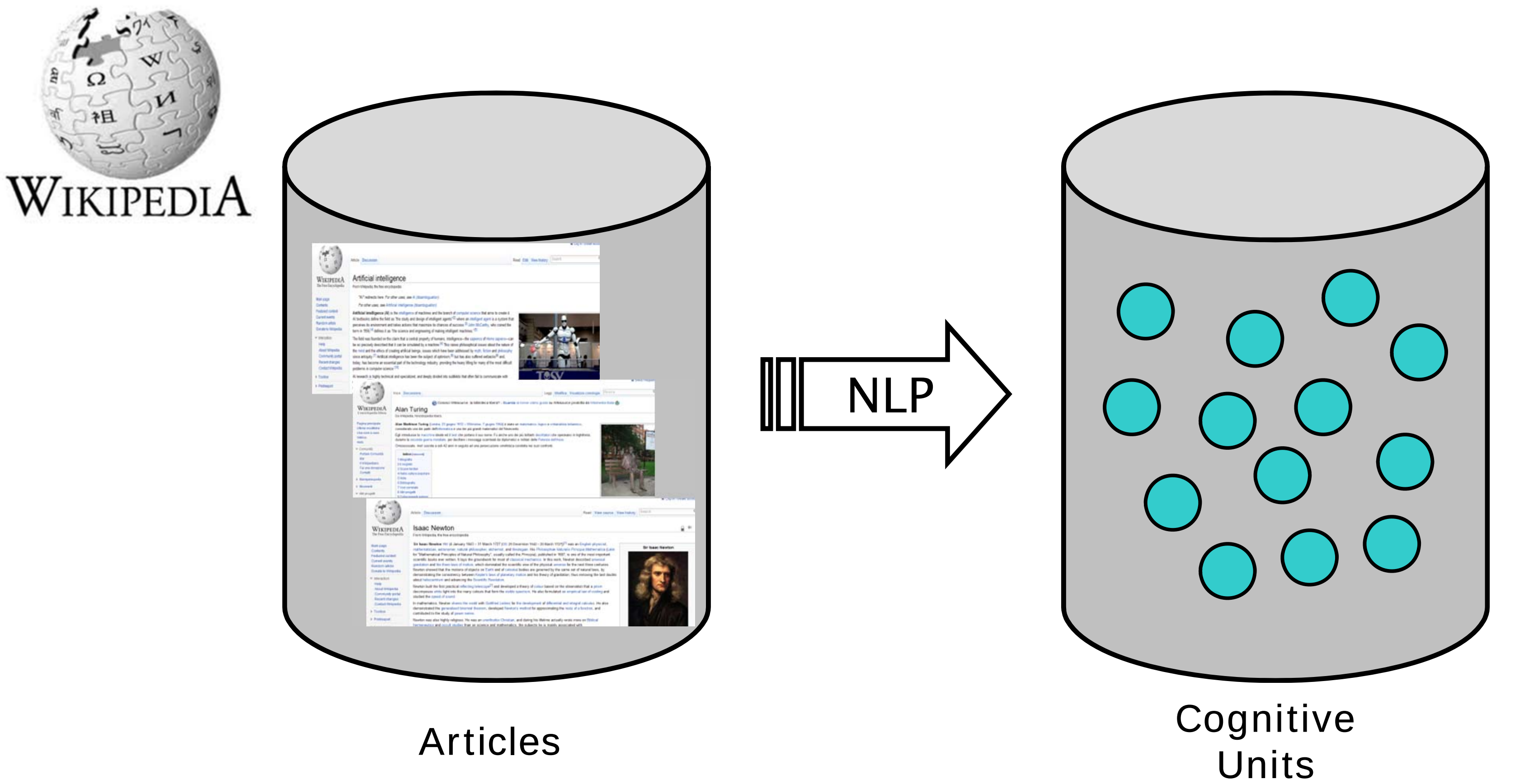
[And83] J. R. Anderson. A Spreading Activation Theory of Memory. Journal of Verbal Learning and Verbal Behavior, 22:261–295, 1983.

Encoding a Knowledge Source as Cognitive Unit Repository



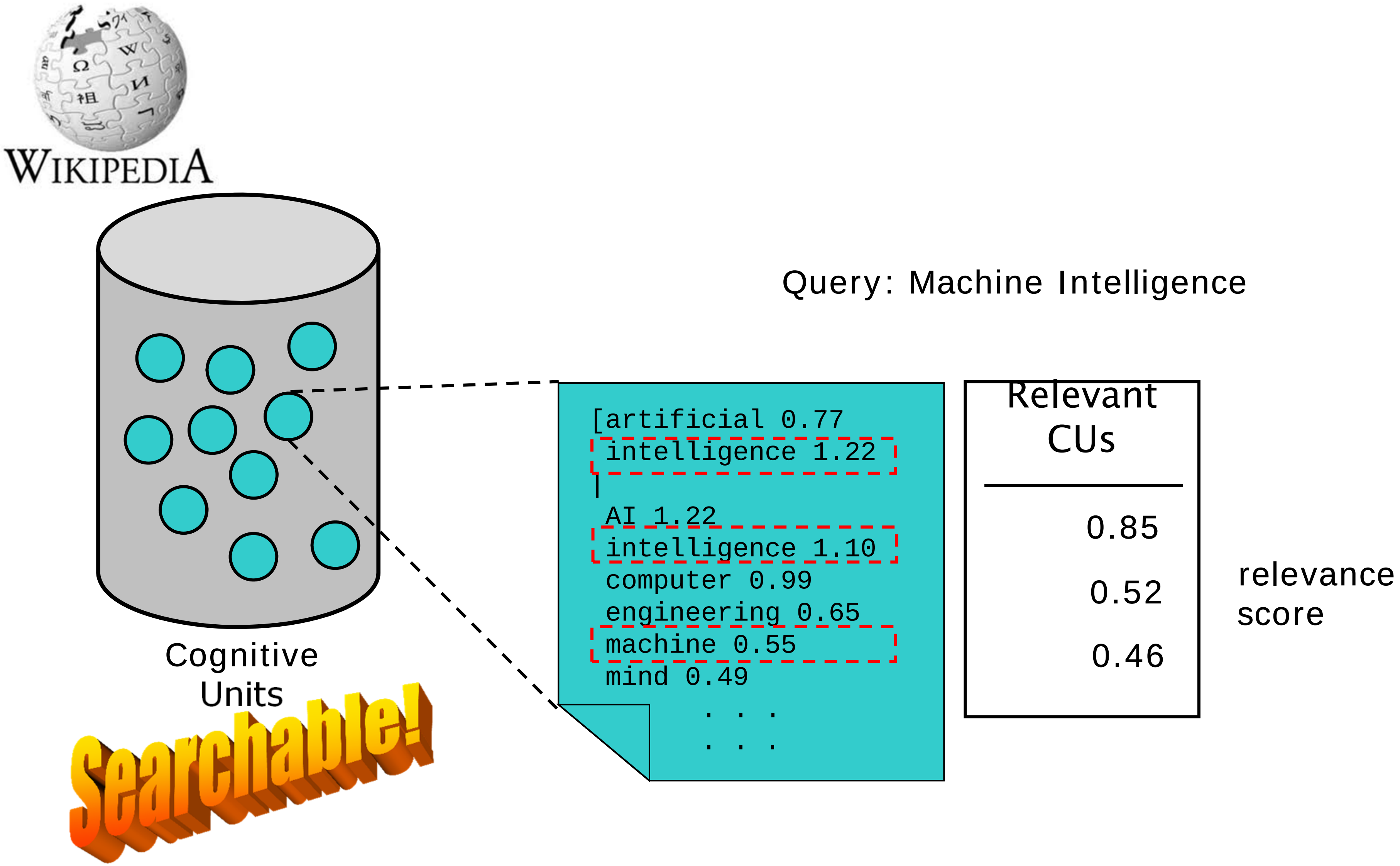

```
Artificial 0.77       |  AI 1.22      intelligence 1.10  computer  0.99
Intelligence 1.22     |  engineering 0.65   machine 0.55      mind  0.49
                      |       … … … …
```

From Wikipedia articles to CU

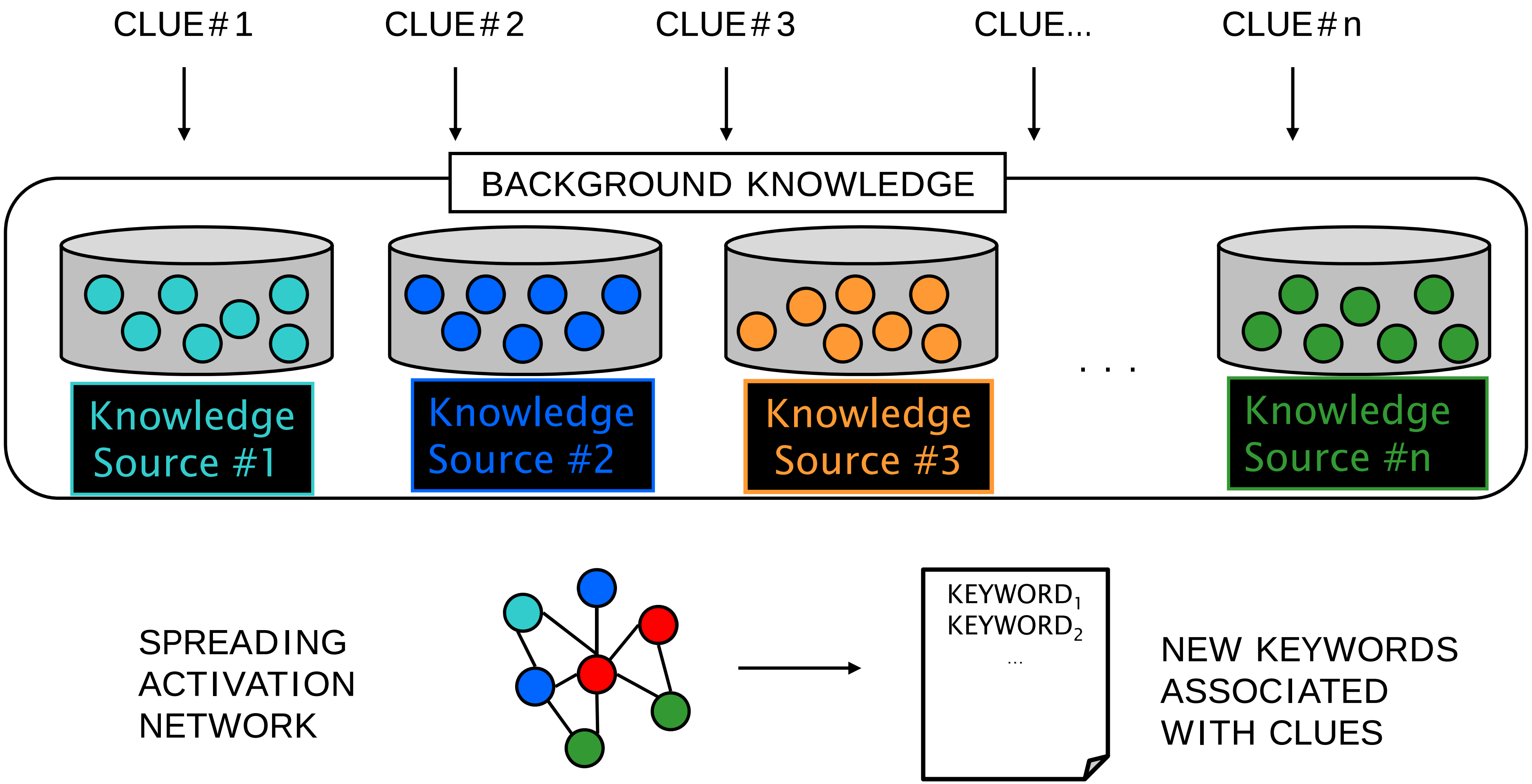


CU repositories can be queried

Query: Machine Intelligence



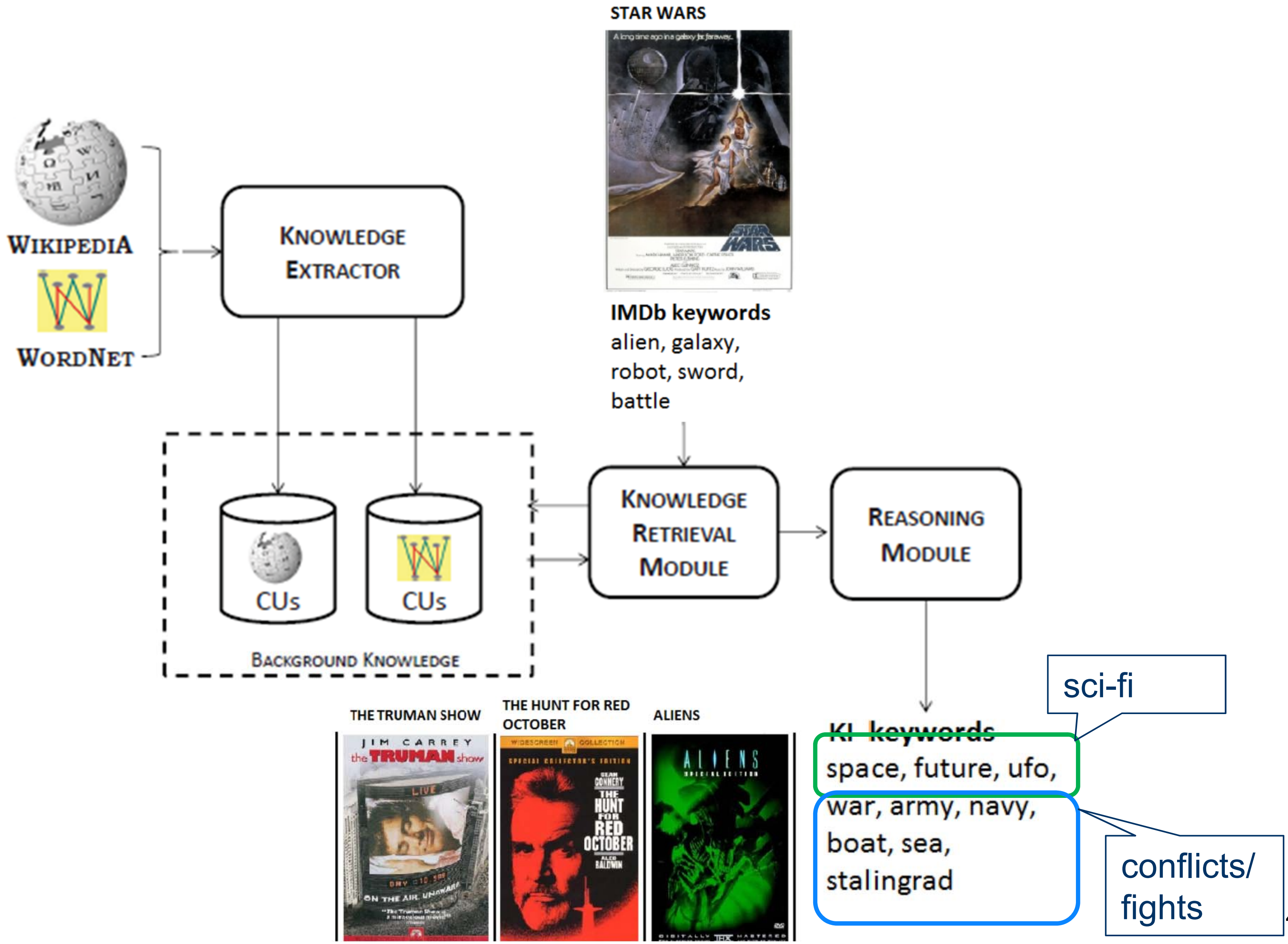
KI@Work



G. Semeraro, M. de Gemmis, P. Lops, P. Basile. An Artificial Player for a Language Game. *IEEE Intelligent Systems* 27(5): 36-43, 2012.

P. Basile, M. de Gemmis, P. Lops, G. Semeraro. Solving a Complex Language Game by using Knowledge-based Word Associations Discovery. *IEEE Transactions on Computational Intelligence and AI in Games*, 2016 DOI: 10.1109/TCIAIG.2014.2355859.

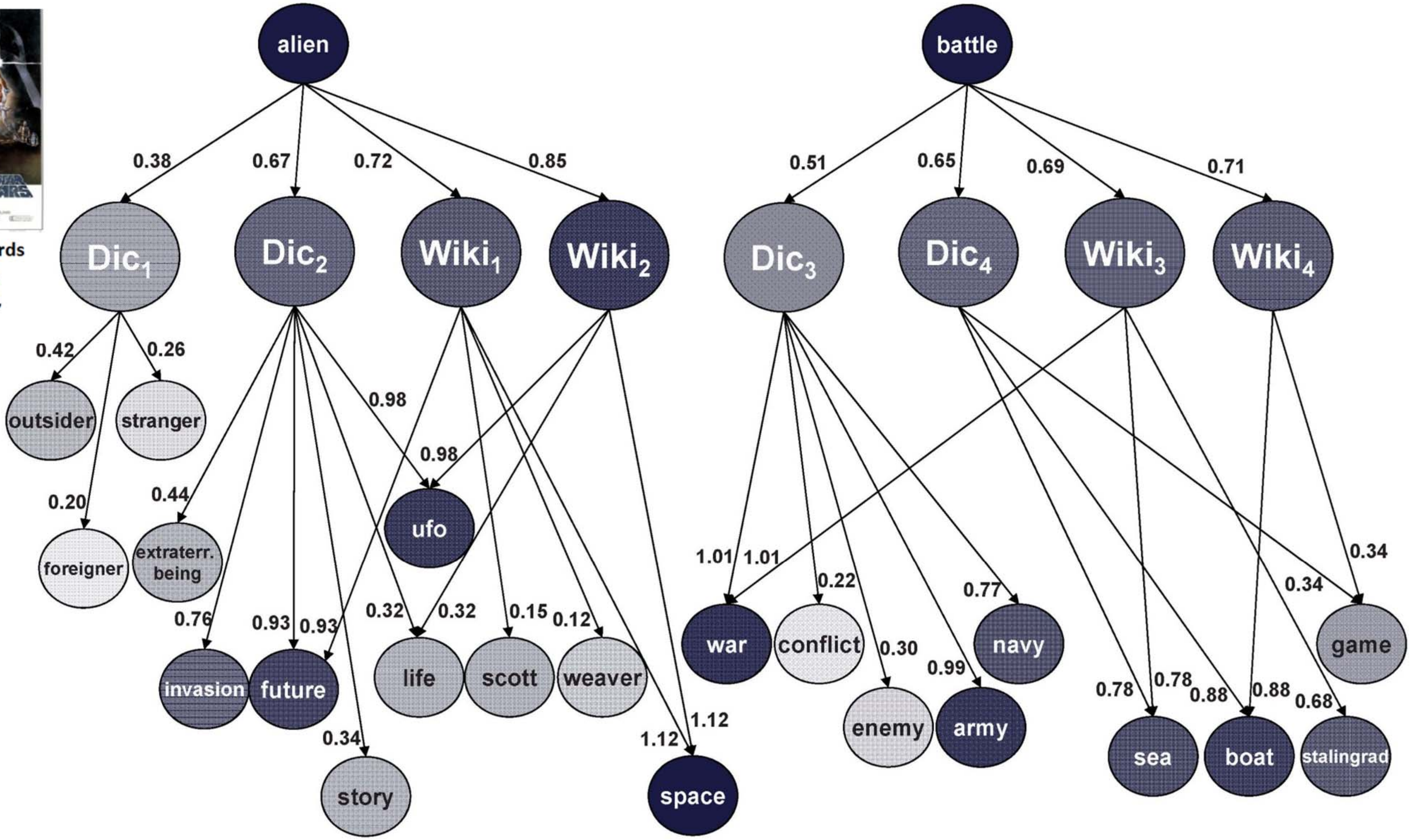
KI@work on Movies



The Spreading Activation Net

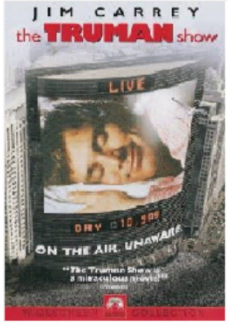
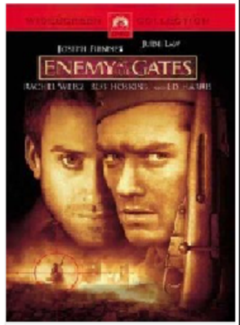
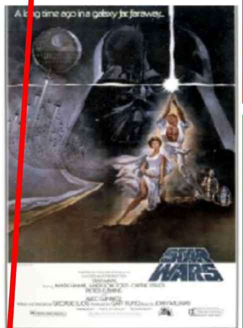
STAR WARS

IMDb keywords
alien, galaxy,
robot, sword,
battle



KI as a novel method for computing associations between items

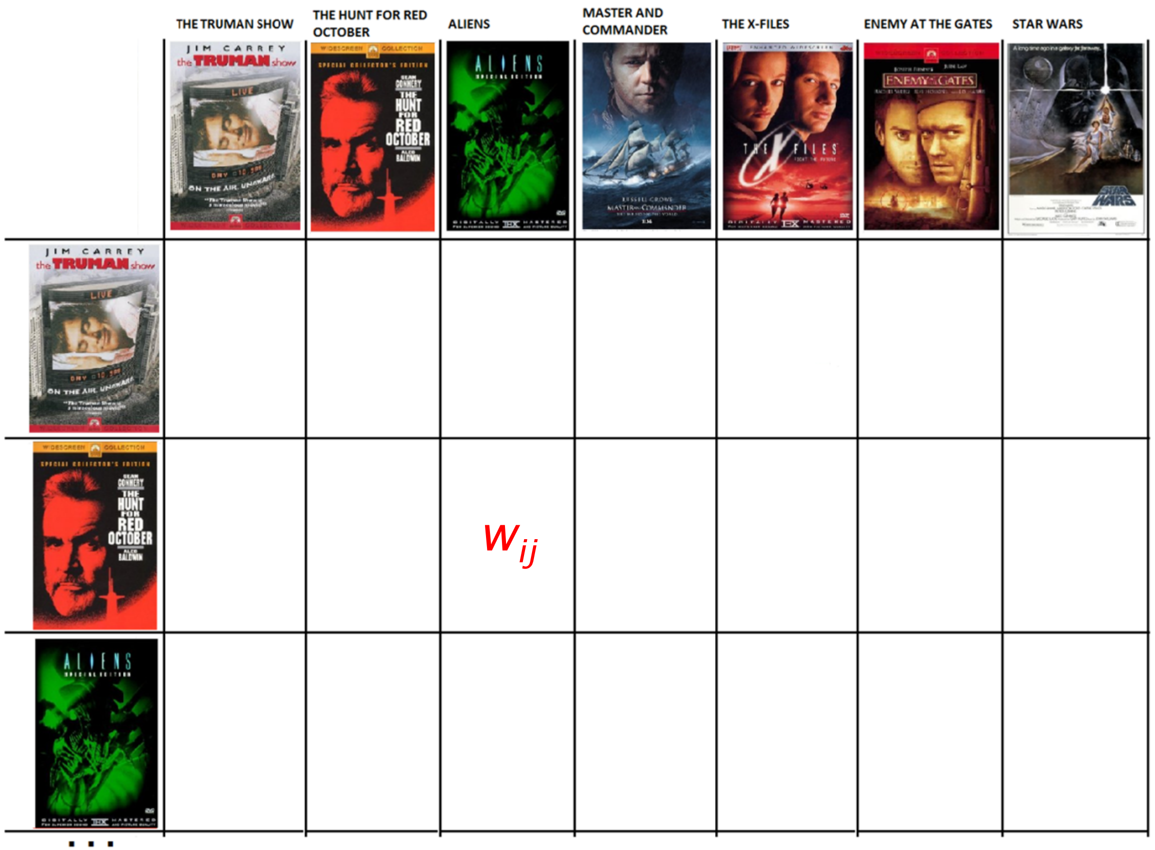
clues

	THE TRUMAN SHOW	THE HUNT FOR RED OCTOBER	ALIENS	MASTER AND COMMANDER	THE X-FILES	ENEMY AT THE GATES
						
STAR WARS  **IMDb keywords** alien, galaxy, robot, sword, battle **KI keywords** space, future, ufo war, army, navy, boat, sea, stalingrad	**Correlation index** 0.43 **Keywords matched** boat, future, storm at sea	**Correlation index** 0.67 **Keywords matched** navy, US navy, soviet navy, sea, cold war	**Correlation index** 0.55 **Keywords matched** outer space, space colony, space travel, future	**Correlation index** 0.72 **Keywords matched** sea, sea battle, navy, royal navy, ship, war	**Correlation index** 0.14 **Keywords matched** ufo	**Correlation index** 0.51 **Keywords matched** world war II, stalingrad, german army, boat

...

BM25 retrieval score

KI as a Serendipity Engine: Item-to-Item similarity matrix → Item-to-Item correlation matrix



w_{ij} computed in different ways

- ✓ #users co-rated items I_i and I_j
- ✓ cosine similarity between descriptions of items I_i and I_j
- ✓ KI Correlation index

Recommendation list computed by **Random Walk with Restart** (Lovasz 1996) working on KI matrix (RWR-KI)

(Lovasz 1996) L. Lovasz. Random Walks on Graphs: a Survey. *Combinatronics* 2:1–46, 1996.

Experimental Evaluation

- Validation of the hypothesis that recommendations produced by RWR-KI are *serendipitous* (*relevant*/*attractive* & *unexpected*/*surprising*)

- Difficulty of assessing *unexpectedness*
  - ✓ In-vitro experiments: *unexpectedness* measured as deviation from a *standard prediction criterion* such as *popularity* [Murakami et al. 2008]
  - ✓ User studies: how to measure the *pleasant surprise* that serendipity should convey?
- User study
  - ✓ Emotions observed in facial expressions are used as implicit feedback for serendipity → Analysis performed using Noldus FaceReader™

[Murakami et al. 2008] T. Murakami, K. Mori, R. Orihara, Metrics for Evaluating the Serendipity of Recommendation Lists, in K. Satoh, A. Inokuchi, K. Nagao, T. Kawamura (Eds.), New Frontiers in Artificial Intelligence, *Lecture Notes in Computer Science* 4914, pp. 40-46, Springer, 2008.

Noldus FaceReader™

- Recognize *basic emotions*: 6 categories of emotions, proposed by Ekman (1999)

✓ happiness

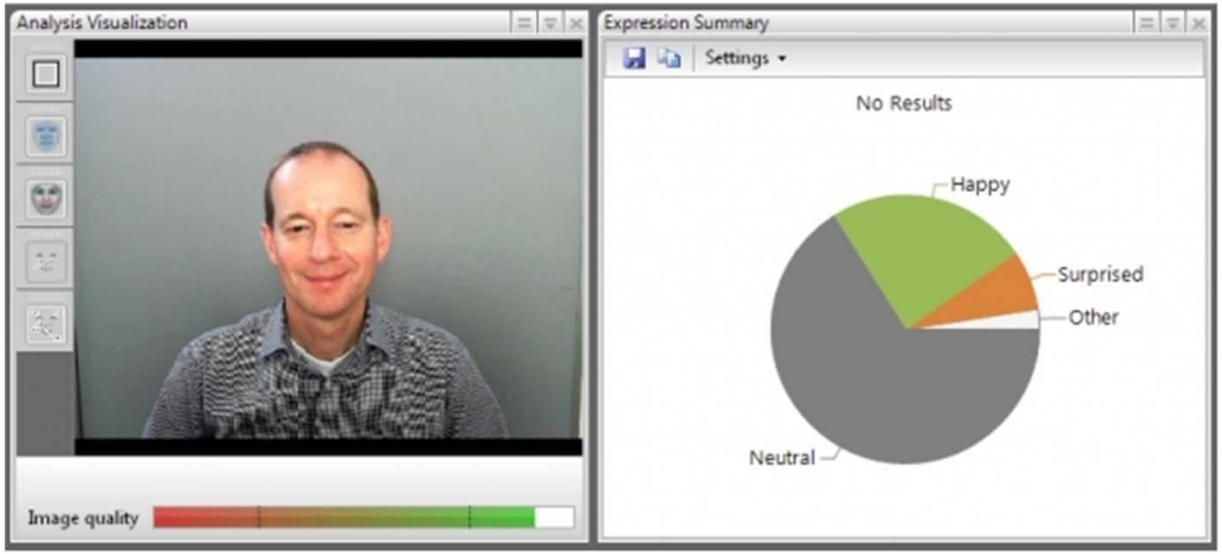
✓ anger

✓ sadness

✓ fear

✓ disgust

✓ surprise



(Ekman 1999) P. Ekman, Basic Emotions, in T. Dalgleish, M.J. Power (Eds.), *Handbook of Cognition and Emotion*, 45-60, John Wiley & Sons, 1999.

Dataset

- Experimental units: 40 master students (engineering, architecture, economy, computer science and humanities)
  - ✓ 26 male (65%), 14 female (35%)
  - ✓ Age distribution: from 20 to 35
- Dataset
  - ✓ 2,135 movies released between 2006 and 2011
  - ✓ Movie content – title, poster, plot keywords, cast, director, summary – crawled from the Internet Movie Database (IMDb)
  - ✓ Vocabulary of 32,583 plot keywords
  - ✓ Average: 12.33 keywords/item

Experimental Design (I)

- *Between-subjects* controlled experiment
  - ✓ 20 users randomly assigned to test RWR-KI
  - ✓ 20 users randomly assigned to test RANDOM (control group), a baseline inspired by the *blind luck* principle which produces random suggestions
- Procedure
  - ✓ Users interact with a web application
    - shows details of movies
    - 20 ratings collected (used only by RWR-KI)
    - displays 5 recommendations (movie poster & title) per user
  - ✓ Recommended items displayed 1 at a time

Experimental Design (II)

- Procedure
  - ✓ 2 binary questions to assess *user acceptance*
    - *"Have you ever heard about this movie?"* → *unexpectedness*
    - *"Do you like this movie?"* → *relevance*
    - (NO,YES) answers → *serendipitous* recommendation
  - ✓ Video started when a movie is recommended to the user and stopped when the answers to the 2 questions were provided
  - ✓ 5 videos per user→ 200 videos recorded to assess *user emotional response* when exposed to recommendations

Metrics

Relevance@N = #relevant_items/*N* *“Do you like this movie?” → YES!*

Unexpectedness@N = #unexpected_items/*N*

“Have you ever heard about this movie?” → NO!

Serendipity@N = #serendipitous_items/*N*
= #(relevant_items ∩ unexpected_items)/*N*

N = size of the recommendation list = 5

Results: Questionnaire Analysis

Metric	RWR-KI	RANDOM
Relevance	0.69	0.46
Unexpectedness	0.72	0.85
Serendipity	0.46	0.35

- ✓ Serendipity: RWR-KI outperforms RANDOM
- ✓ Statistically significant differences (Mann-Whitney U test, $p<0.05$)
- ✓ ~ Half of the recommendations are deemed serendipitous!
- ✓ RWR-KI: a better Relevance-Unexpectedness trade-off
- ✓ RANDOM: more unbalanced towards Unexpectedness

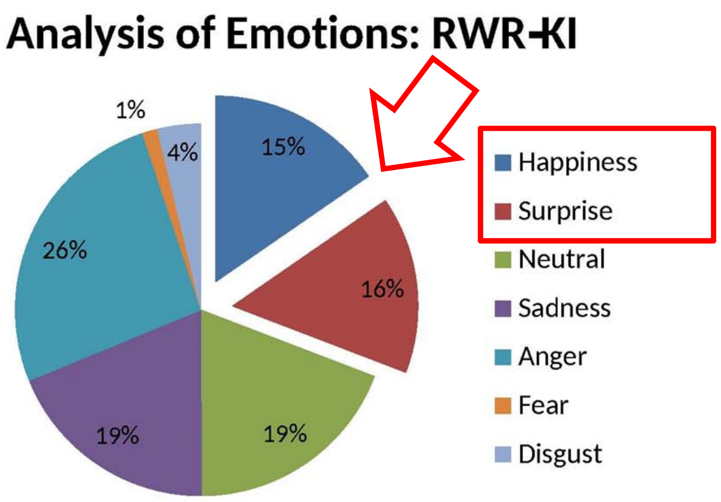
Results: Analysis of User Emotions

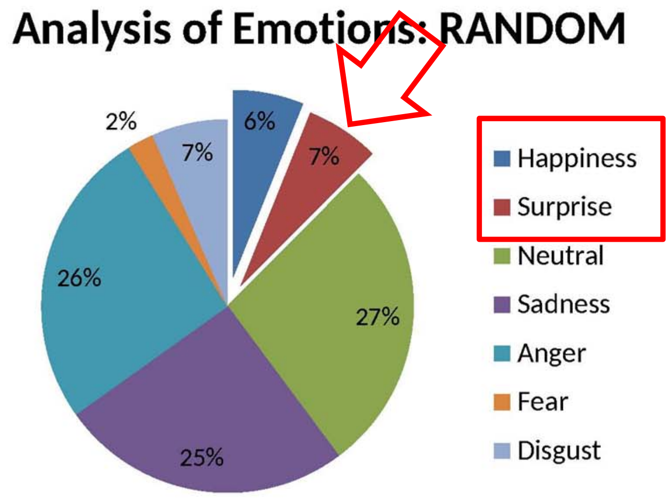
- Hypothesis: *users' facial expressions convey a mixture of emotions that helps to measure the perception of serendipity of recommendations*
- Serendipity associated to *surprise* and *happiness*
- 200 videos (40 users x 5 recommendations)
  - ✓ 41 videos filtered out (< 5 seconds)
- ∀ 159 videos, FaceReader™ computed the distribution of detected emotions + duration (emotions lasting < 1 sec. filtered out)

Algorithm	Serend. Recomm.	Non-Serend. Recomm.
RWR-KI	39	39
RANDOM	30	51

Results: analysis of user emotions associated to serendipitous suggestions

Algorithm	Serend. Recomm.	Non-Serend. Recomm.
RWR-KI	39	39
RANDOM	30	51

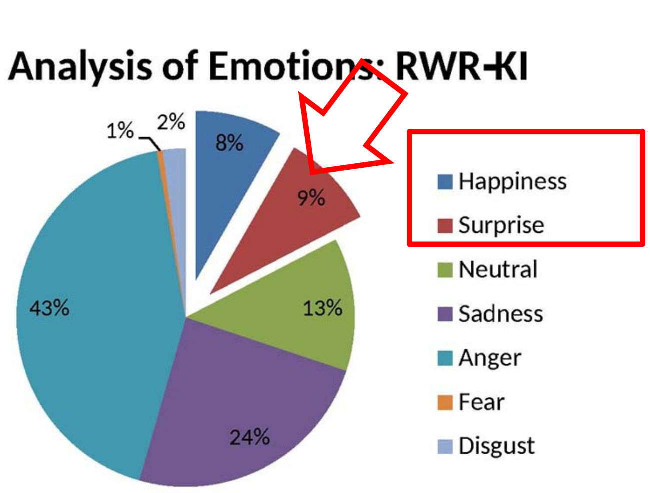


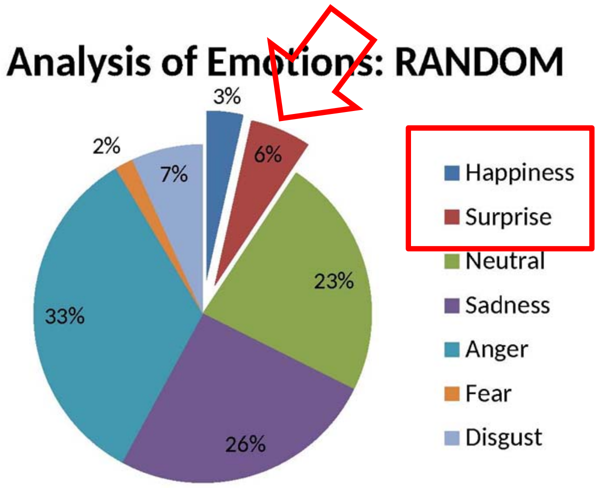


- ✓ Evidence of Happiness and Surprise for both algorithms
- ✓ RWR-KI > RANDOM (in line with the questionnaire results)
- ✓ High values of negative emotions (*sadness* and *anger*) → ?????

Results: analysis of user emotions associated to non-serendipitous suggestions

Algorithm	Serend. Recomm.	Non-Serend. Recomm.
RWR-KI	39	39
RANDOM	30	51





- ✓ General decrease of *surprise* and *happiness*
- ✓ High values of negative emotions (*sadness* and *anger*), also in this case → due to the fact that users assumed troubled expressions since they were very concentrated on the task

Main findings

- Positive emotions: Happyness, Surprise
  - ✓ marked difference between RWR-KI and RANDOM
  - ✓ marked difference between serendipitous and non-serendipitous recommendations
- Moderate agreement between *explicit feedback* (questionnaires) & *implicit feedback* (facial expressions/emotions)
  - ✓ Cohen's kappa coefficient
- Emotions can help to assess the actual perception of serendipity

M. de Gemmis, P. Lops, G. Semeraro, C. Musto. An Investigation on the Serendipity Problem in Recommender Systems. *Information Processing and Management*, 2015 DOI: 10.1016/j.ipm.2015.06.008

Readings

Semantics-aware Recommender Systems

- C. Musto, G.Semeraro, M.de Gemmis, P. Lops. **A Hybrid Recommendation Framework Exploiting Linked Open Data and Graph-based Features.** UMAP 2017
- C. Musto, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: **Introducing linked open data in graph-based recommender systems**. Inf. Process. Manage. 53(2): 405-435 (2017)
- C. Musto, G. Semeraro, M. de Gemmis, P. Lops: **Tuning Personalized PageRank for Semantics-Aware Recommendations Based on Linked Open Data**. ESWC (1) 2017: 169-183
- V. W. Anelli, V. Bellini, T. Di Noia, W. La Bruna, P. Tomeo, E. Di Sciascio: ***An analysis on Time- and Session-aware diversification in recommender systems.*** UMAP 2017
- A. Ragone, P. Tomeo, C. Magarelli, T. Di Noia, M. Palmonari, A. Maurino, E. Di Sciascio: ***Schema-summarization in Linked-Data-based feature selection for recommender systems.*** 32nd ACM SIGAPP Symposium On Applied Computing - 2017
- T. Di Noia, J. Rosati, P. Tomeo, E. Di Sciascio: ***Adaptive multi-attribute diversity for recommender systems.*** Inf. Sci. 382-383: 234-253(2017)
- S. Oramas, V. C. Ostuni, T. Di Noia, X. Serra, E. Di Sciascio: ***Sound and Music Recommendation with Knowledge Graphs.*** ACM TIST 8(2): 21:1-21:21 (2017)
- C. Musto, G. Semeraro, M. de Gemmis, P. Lops: ***Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems***. ECIR 2016: 729-734
- T. Di Noia, V. C. Ostuni, P. Tomeo, E. Di Sciascio: ***SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data.*** ACM TIST 8(1): 9:1-9:34 (2016)
- P. Tomeo, I. Fernández-Tobías, T. Di Noia, I. Cantador: ***Exploiting Linked Open Data in Cold-start Recommendations with Positive-only Feedback.*** CERI 2016: 11
- I. Fernández-Tobías, P. Tomeo, I. Cantador, T. Di Noia, E. Di Sciascio***: Accuracy and Diversity in Cross-domain Recommendations for Cold-start Users with Positive-only Feedback.*** RecSys 2016: 119-122
- C. Musto, G. Semeraro, M. de Gemmis, P. Lops: ***Word Embedding Techniques for Content-based Recommender Systems: An Empirical Evaluation***. RecSys Posters 2015

Readings

Semantics-aware Recommender Systems

- C. Musto, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, S. Rutigliano: ***Automatic Selection of Linked Open Data Features in Graph-based Recommender Systems***. CBRecSys@RecSys 2015: 10-13
- M. de Gemmis, P. Lops, C. Musto, F.Narducci, G. Semeraro: ***Semantics-Aware Content-Based Recommender Systems***. Recommender Systems Handbook 2015: 119-159
- P. Tomeo, T. Di Noia, M. de Gemmis, P. Lops, G. Semeraro, E. Di Sciascio: ***Exploiting Regression Trees as User Models for Intent-Aware Multi-attribute Diversity.*** CBRecSys@RecSys 2015: 2-9
- T. Di Noia, V. C. Ostuni: ***Recommender Systems and Linked Open Data.*** Reasoning Web 2015: 88-113
- P. Basile, C. Musto, M. de Gemmis, P. Lops, F. Narducci, G. Semeraro: ***Content-Based Recommender Systems + DBpedia Knowledge = Semantics-Aware Recommender Systems***. SemWebEval@ESWC 2014: 163-169
- C. Musto, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: ***Linked Open Data-enabled Strategies for Top-N Recommendations***. CBRecSys@RecSys 2014: 49-56
- C. Musto, G. Semeraro, P. Lops, M. de Gemmis: ***Combining Distributional Semantics and Entity Linking for Context-Aware Content-Based Recommendation***. UMAP 2014: 381-392
- V. C. Ostuni, T. Di Noia, R. Mirizzi, E. Di Sciascio: ***A Linked Data Recommender System Using a Neighborhood-Based Graph Kernel.*** EC-Web 2014: 89-100
- V. C. Ostuni, T. Di Noia, E. Di Sciascio, R. Mirizzi: ***Top-N recommendations from implicit feedback leveraging linked open data.*** RecSys 2013: 85-92
- C. Musto, G. Semeraro, P. Lops, M. de Gemmis: ***Contextual eVSM: A Content-Based Context-Aware Recommendation Framework Based on Distributional Semantics***. EC-Web 2013: 125-136
- T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, M. Zanker: ***Linked open data to support content-based recommender systems.*** I-SEMANTICS 2012: 1-8
- C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, R. Clout: ***Enhanced Semantic TV-Show Representation for Personalized Electronic Program Guides***. UMAP 2012: 188-199

Readings

Semantics-aware Recommender Systems

- M. Degemmis, P. Lops, G. Semeraro: ***A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation***. User Model. User-Adapt. Interact. 17(3): 217-255 (2007)
- G. Semeraro, M. Degemmis, P. Lops, P. Basile: ***Combining Learning and Word Sense Disambiguation for Intelligent User Profiling***. IJCAI 2007: 2856-2861

Cross-language Recommender Systems

- F. Narducci, P. Basile, C. Musto, P. Lops, A. Caputo, M. de Gemmis, L. Iaquinta, G. Semeraro: **Concept-based item representations for a cross-lingual content-based recommendation process**. Inf. Sci. 374: 15-31 (2016)
- C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: ***Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models***. AI*IA 2011: 250-261
- P. Lops, C. Musto, F. Narducci, M. de Gemmis, P. Basile, G. Semeraro: ***Cross-Language Personalization through a Semantic Content-Based Recommender System***. AIMSA 2010: 52-60

Explanations

- C. Musto, F. Narducci, P. Lops, M. de Gemmis, G. Semeraro: **ExpLOD: A Framework for Explaining Recommendations based on the Linked Open Data Cloud**. In Proc. of the 10th ACM Conference on Recommender Systems (RecSys '16). ACM, New York, NY, USA, 151-154.

Serendipity

- M. de Gemmis, P. Lops, G. Semeraro, C. Musto. **An Investigation on the Serendipity Problem in Recommender Systems.** Information Processing and Management, 2015 DOI: 10.1016/j.ipm.2015.06.008