

xkcd.com/1592/



# Retrieving Knowledge from the Web

**Laura Dietz**

dietz@cs.unh.edu



**University of New Hampshire**

# What is Knowledge? (Pragmatic Definition)

## United Kingdom



Coordinates: 55°N 3°W

The **United Kingdom of Great Britain and Northern Ireland**, commonly known as the **United Kingdom (UK)** or **Britain**, is a **sovereign country** in western Europe.

Lying off the north-western coast of the **European mainland**, the United Kingdom includes the island of **Great Britain**, the north-eastern part of the island of **Ireland** and many smaller

## Theresa May

**Theresa Mary May** (*née* **Brasier**;<sup>[1]</sup> born 1 October 1956) is the **Prime Minister of the United Kingdom** and **Leader of the Conservative Party**, having served as both since July 2016. She has been the **Member of Parliament (MP)** for **Maidenhead** since 1997. May identifies as a **one-nation conservative** and has been characterised as a **liberal conservative**. She is the second female Prime Minister and Conservative Party leader after **Margaret Thatcher**.

## Brexit

**Brexit** is a commonly used term for the **United Kingdom's planned withdrawal from the European Union**.<sup>[1]</sup> Following the **2016 referendum vote** to leave, the UK government started **the withdrawal process** on 29 March 2017, putting the UK on course to leave by April 2019.<sup>[2]</sup>



YAY go Brexit :-)

RETWEETS

2

LIKES

4

11:05 AM - 6 Apr 2017



2

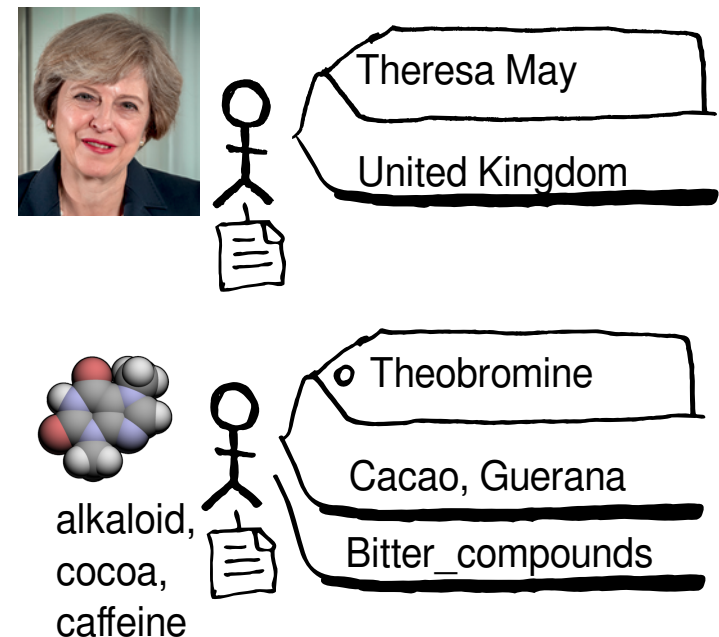
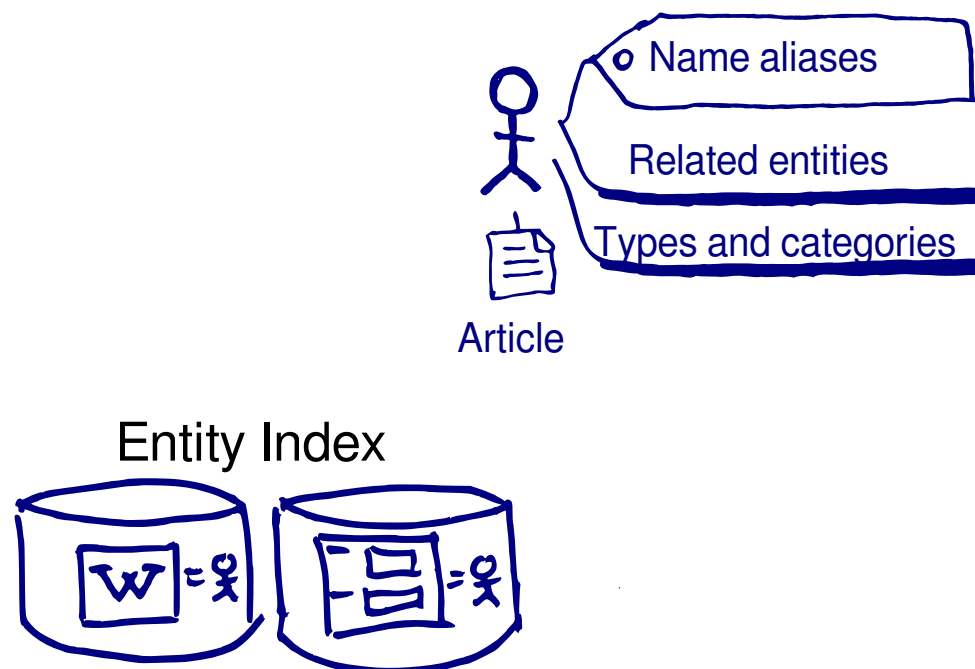


4

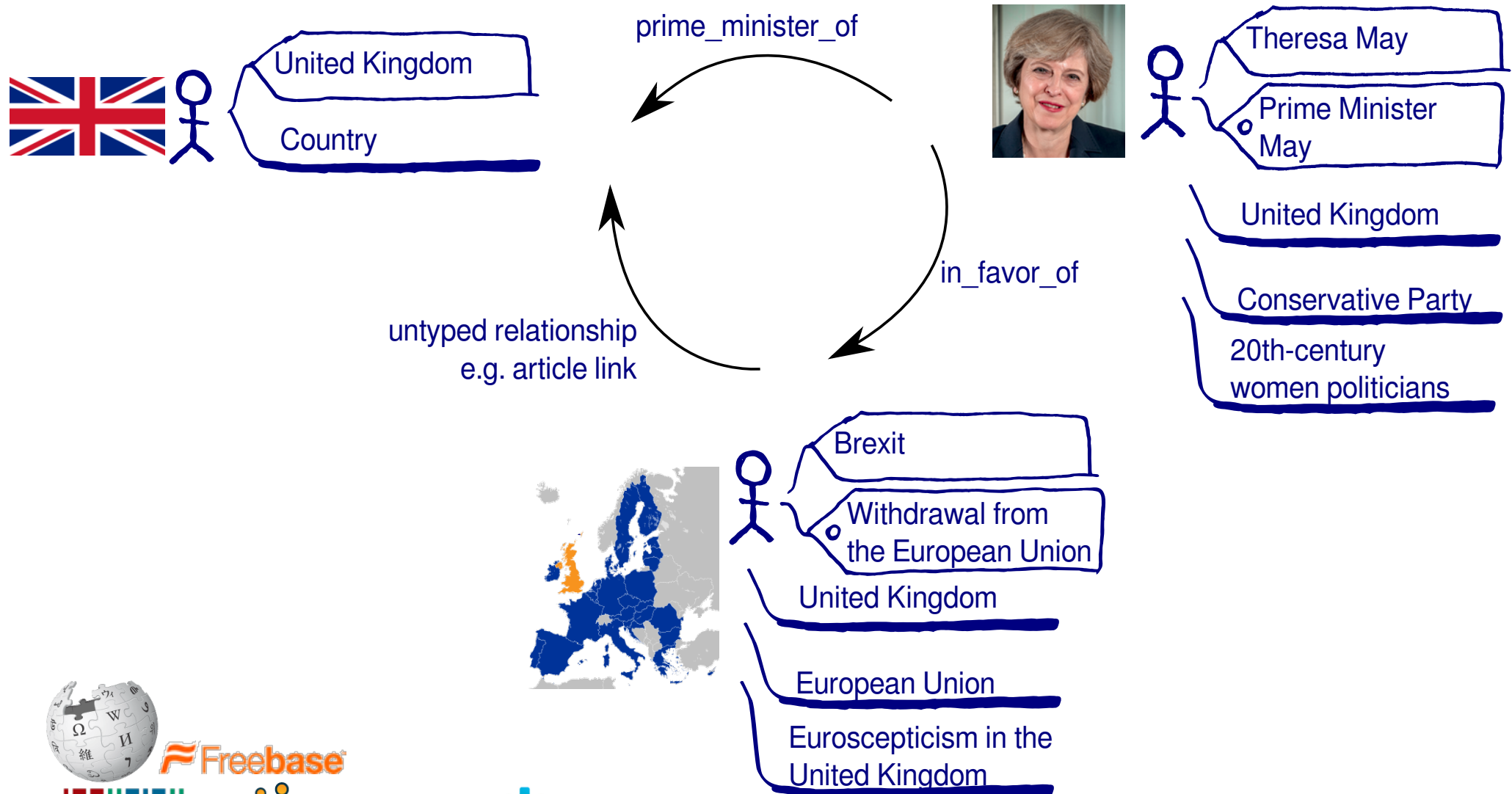
# What is an Entity?

Person, Place, Gene, Protein, Event, Thing

Anything with an entry in a knowledge base  
(here: anything with a Wikipedia article)



# What is a Knowledge Graph?



Freebase



# Open Information Needs

Requiring long, complex answers

Intended queries:

- drink water good
- dark chocolate health benefits
- causes conflict Middle East
- UK leaving Europe
- Spent nuclear fuel

If yes, why? If not, why not?

Causes? Involvements? Controversy? Backstory?

What do I need to know to understand the answer?

[xkcd.com/1592/](http://xkcd.com/1592/)



# Open Information Needs

Requiring long, complex answers

Intended queries:

- drink water good
- dark chocolate health benefits
- causes conflict Middle East
- UK leaving Europe
- Spent nuclear fuel

[xkcd.com/1592/](http://xkcd.com/1592/)



If yes, why? If not, why not?

Causes? Involvements? Controversy? Backstory?

What do I need to know to understand the answer?

# Health effects of chocolate

From Wikipedia, the free encyclopedia

The **health effects of chocolate** refer to the possible positive and negative effects on health of eating [chocolate](#).

Unconstrained consumption of large quantities of any energy-rich food, such as chocolate, without a corresponding increase in activity, increases the risk of [obesity](#). Raw chocolate is high in [cocoa butter](#), a fat removed during chocolate refining, then added back in varying proportions during manufacturing. Manufacturers may add other fats, [sugars](#), and [powdered milk](#) as well.

Although considerable research has been conducted to evaluate the potential health benefits of consuming chocolate, there are insufficient [studies](#) to confirm any effect and no medical or regulatory authority has approved any health claim.



Desperation , Pacification ,  
Expectation, Acclamation , Realization ,  
It's Fry's. Advertisement of Fry's 'Five  
Boys' [milk chocolate](#)

## Contents [\[hide\]](#)

### 1 Research

- 1.1 [Acne](#)
- 1.2 [Addiction](#)
- 1.3
- 1.4 [Heart and blood vessels](#)
- 1.5 [Stimulant](#)
- 1.6 [Weight gain](#)

### 2 Lead content

### 3 Polyphenol content

Provide More!



# dark chocolate health benefits

## 7 Proven Health Benefits of Dark Chocolate (No. 5 is Best)

[authoritynutrition.com/7-health-benefits-dark-chocolate](http://authoritynutrition.com/7-health-benefits-dark-chocolate)

**Dark chocolate** is loaded with nutrients that can positively affect your **health**. Made from the seed of the cocoa tree, it is one of the best sources of antioxidants on ...

Synthesize!

## Six Health Benefits of Dark Chocolate / Nutrition ...

[www.fitday.com/.../6-health-benefits-of-dark-chocolate.html](http://www.fitday.com/.../6-health-benefits-of-dark-chocolate.html)

**Dark chocolate** has recently been discovered to have a number of healthy **benefits**. While eating **dark chocolate** can lead to the **health benefits** described below ...

## Pick Dark Chocolate for Health Benefits - WebMD - Better ...

[www.webmd.com/diet/20120424/pick-dark-chocolate-health-benefits](http://www.webmd.com/diet/20120424/pick-dark-chocolate-health-benefits)

24/04/2012 · **Chocolate** and **Health Benefits**: Study Details. Hong compared white **chocolate**, which has no cocoa solids, to regular **dark chocolate** containing 70% ...

## Dark Chocolate Is Healthy Chocolate - WebMD - Better ...

[www.webmd.com/diet/20030827/dark-chocolate-is-healthy-chocolate](http://www.webmd.com/diet/20030827/dark-chocolate-is-healthy-chocolate)

27/08/2003 · **Dark Chocolate Is Healthy Chocolate**. By Daniel J. DeNoon on August 27, 2003. WebMD News Archive **Dark Chocolate** Has **Health Benefits** Not Seen in ...

## Health Benefits of Dark Chocolates - Mercola.com

[articles.mercola.com/.../03/31/dark-chocolate-health-benefits.aspx](http://articles.mercola.com/.../03/31/dark-chocolate-health-benefits.aspx)

31/03/2014 · Video embedded · By Dr. **Mercola**. The **health benefits of dark chocolate** are all the rage right now, with increasing numbers of studies pointing to its ...



# Bing News Search Results



## Synthesize!



### The **Benefits** of **Dark Chocolate**

Studies have shown that **dark chocolate** has numerous science-backed **health benefits**. Without a doubt, **dark chocolate** has remained one of

worldhealth.net 13d



### 5 Evidence-Based **Health Benefits** Of **Chocolate**

Aren't you truly ecstatic reading that something you so dearly love, helps you have better **health** too? Don't think twice before having that one bar

curejoy.com 10d

Chocolate

### Choose **Dark Chocolate** for **Health Benefits**

April 24, 2012 -- If you're eating **chocolate** for the **health benefits** -- and aren't we all? -- you must pick wisely, new research suggests. **Dark chocolate** was the clear winner, she says. She is due to present the

WebMD 4y

### More from Bing News



NC Defeats Oregon  
In Final 4



Dylan Collects Nobel  
Prize



Galaxy S8 Available  
For Preorder



Cargo Ship Vanishes  
In Atlantic

### Realtime news

#### **Chocolate** and the Elite Athlete

hpsnz.org.nz 9h

#### Vegan **Chocolate** Creme Eggs

itdoesnttastelikechicken.com 1d

#### The Best Vegan **Chocolate** Alternatives For Easter 2017

Forbes 2d

#### Science that Matters: **Benefits** of **Chocolate**

WABI-TV | WABI 4d

#### 7 Probiotic-Filled Foods You Definitely Need to Have in Your Diet

cheatsheet.com 7d

#### 20 Reasons for Swimmers to Eat **Chocolate**

Swimming World 7d

# Query 234 dark chocolate health benefits

## Chocolate

[More explanations...](#)

facts about chocolate and health. how much chocolate is good for your health?

[View Web Source](#)

## C-reactive protein

[More explanations...](#)

cocoa and chocolate can modulate platelet function through a multitude of pathways. chocolate and c-reactive protein levels dark chocolate effect on platelet activity c-reactive protein and lipid profile

[View Web Source](#)

## Theobromine

[More explanations...](#)

chocolate could alleviate some blood circulation problems in the body also increasing blood flow to the brain which could have benefits for memory and dementia theobromine is the main alkaloid in cocoa and dark chocolate some people say that the theobromine in dark chocolate works better for them

[View Web Source](#)

## Circulatory system

[More explanations...](#)

cocoa flavanols have also been shown to have potential anti inflammatory activities that are relevant to cardiovascular health with inflammation substances are formed which can produce adverse cardiovascular effects now dr shock will never let a change to promote chocolate consumption slip

[View Web Source](#)

## Yale University

[More explanations...](#)

a research study in 2008 at yale university suggests that consumption by pregnant women of chocolate rich in the chemical could help prevent pre eclampsia

[View Web Source](#)

## Italy

[More explanations...](#)

eating dark chocolate could help control diabetes and blood pressure, italian experts say

[View Web Source](#)

Demo  
available:



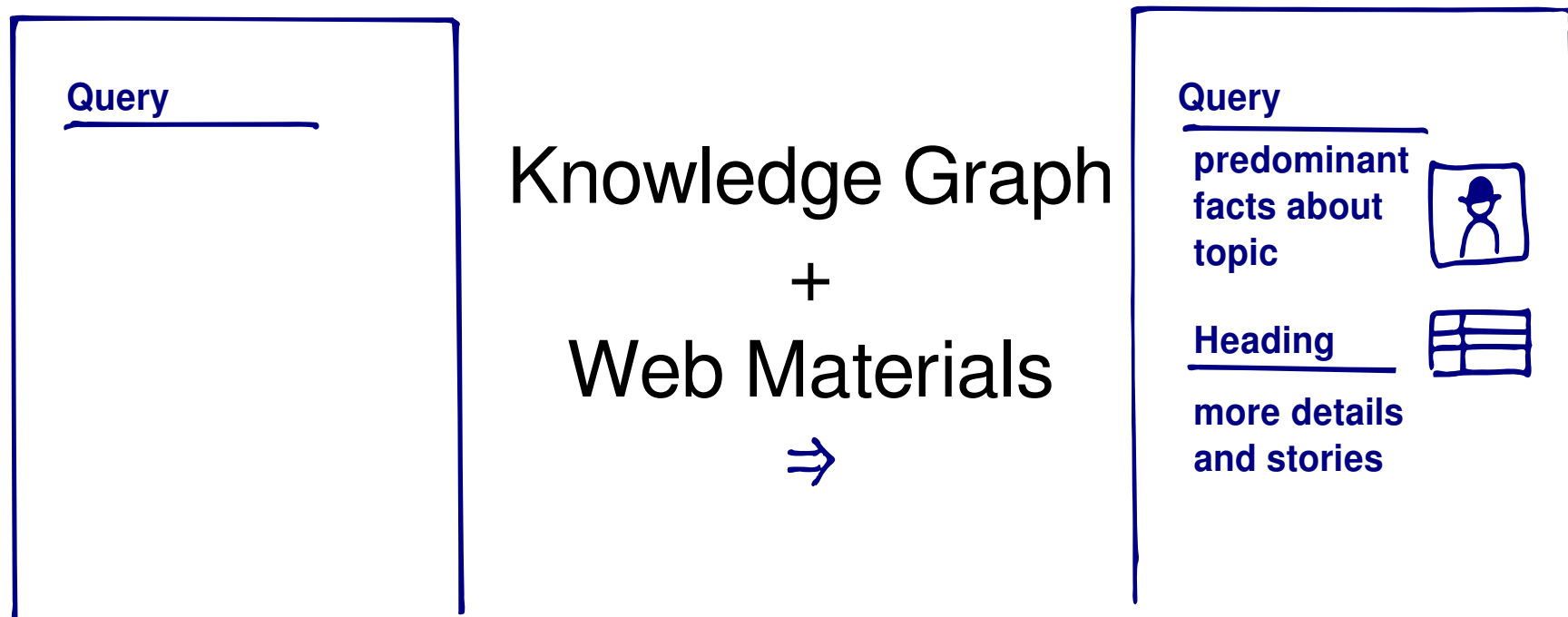
# Vision: Retrieving Knowledge from the Web

---

1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion

# Vision: Query-specific Wikipedia Construction

Given query  $Q$  (= open domain topic),  
automatically compose an encyclopedic article.

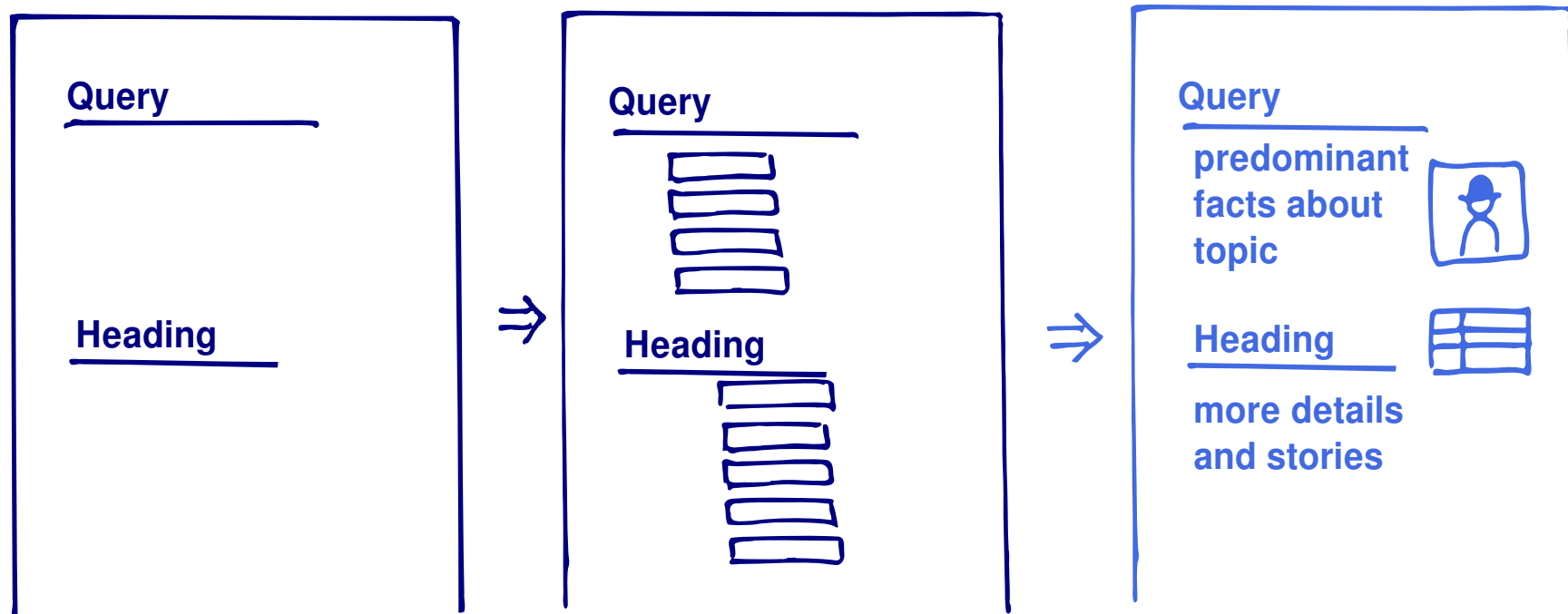


Advantage:

Access to near-infinite material on the Web

# Task: TREC Complex Answer Retrieval

Given: query  $Q$  (= open domain topic)  
and outline of headings ( $H1$ , ..  $Hn$ ).  
For every heading, return ranking of passages.



# TREC Complex Answer Retrieval Data Set

Task:

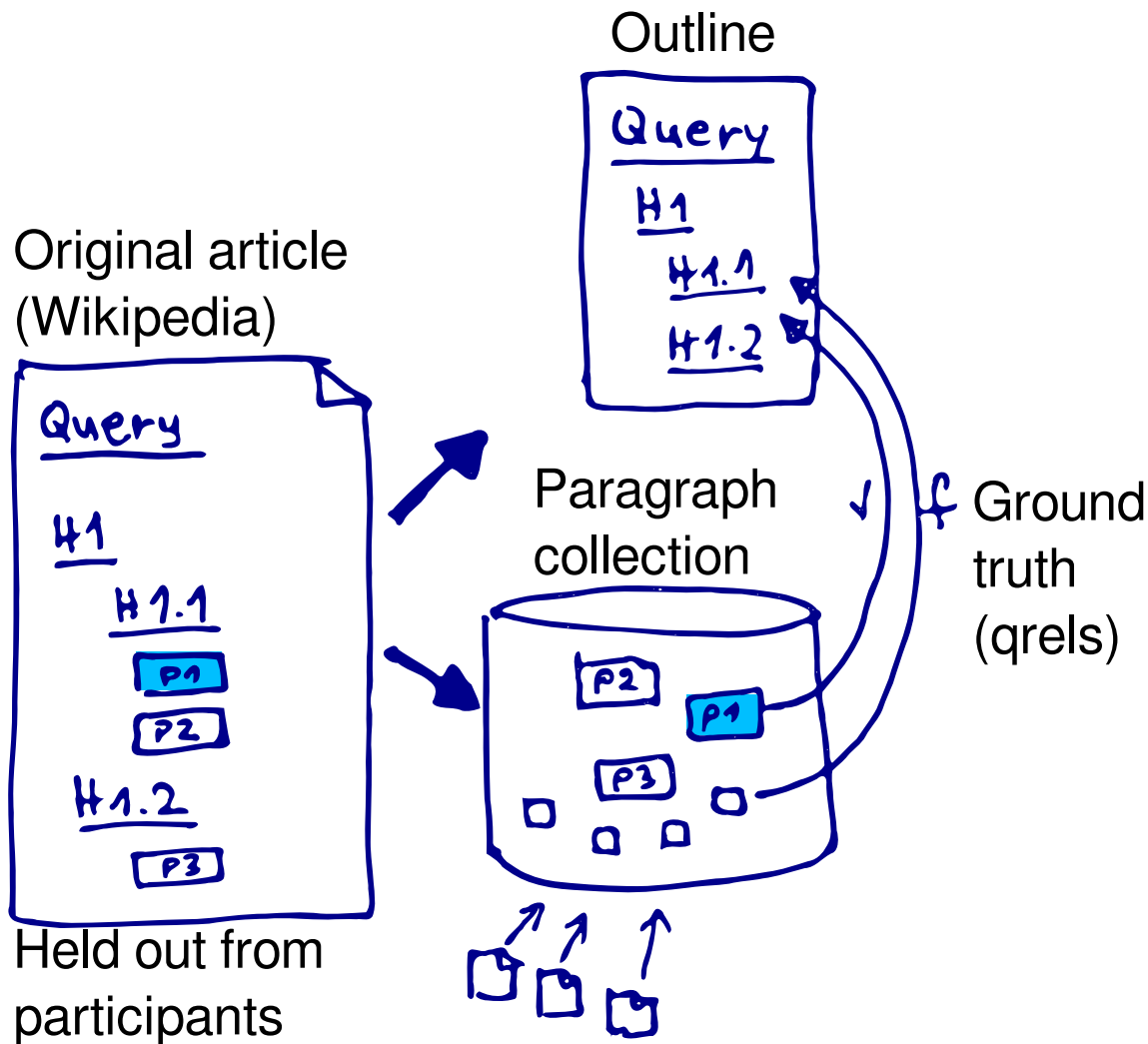
For each heading,  
rank paragraphs

Eval 1:

Article reconstruction

Eval 2:

Relevance judgments  
(by NIST)



Data online: <http://trec-car.cs.unh.edu>

# Great Problem, but: How to solve it?

Simple baseline:

$Q' = \text{Query} + \text{Heading 1} + \text{Heading 1.1} + \dots$

Rank passages with BM25.

Issue:


Many relevant passages do not contain query terms.

Reminder: we want long answers!

For complex answers, helpful:

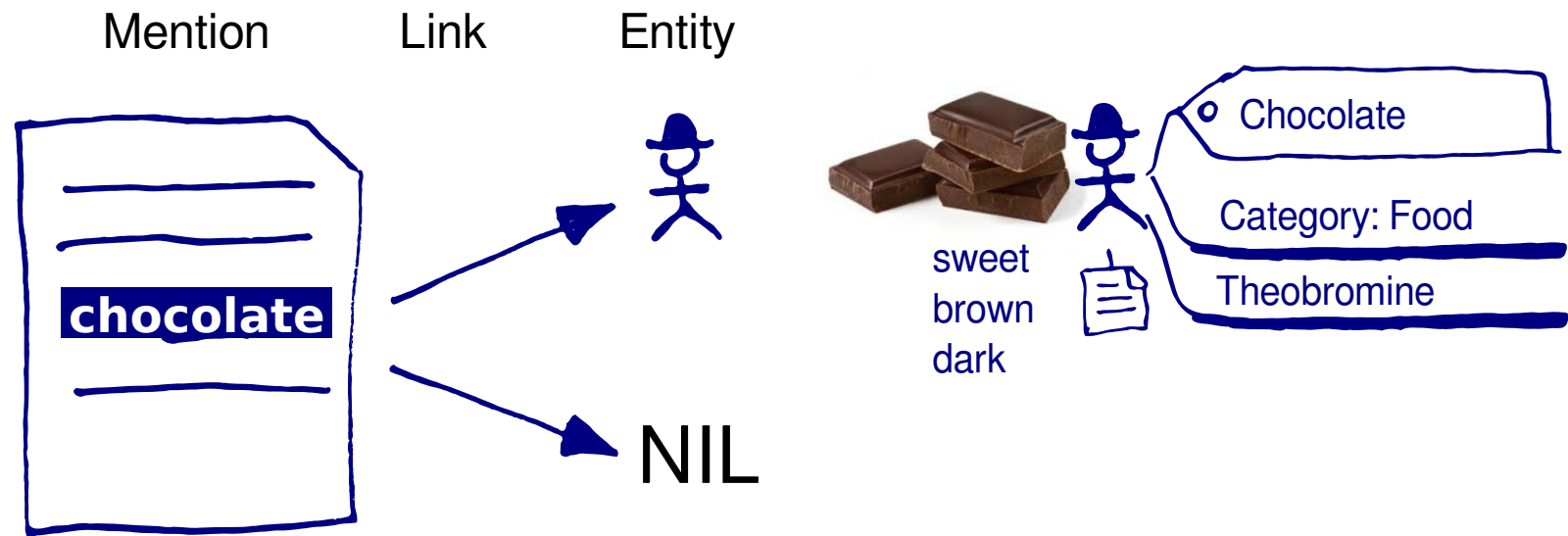
- deeper understanding of text
- relevant concepts, entities, and connections.

You disagree?  
Please prove  
me wrong!





# Task: Entity Linking (aka Wikification)

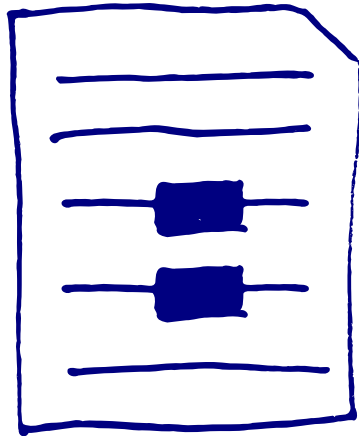


Entity linking algorithms detect entity mentions in text and align them to their knowledge base entry.

# Bag-of-words -> Bag-of-entities

Query: dark chocolate health benefits

$Q' =$    (Query Entities)



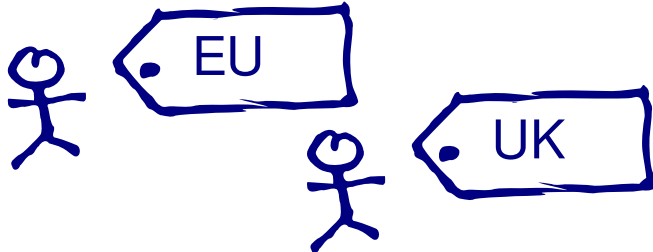
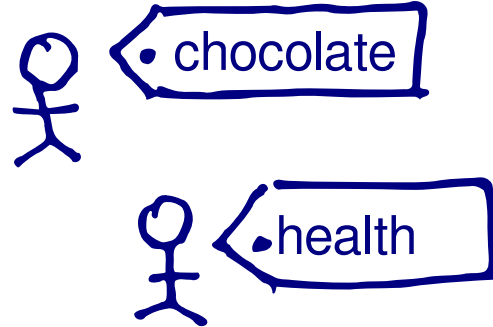
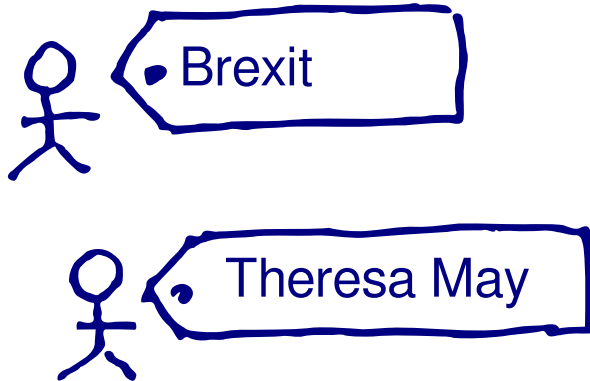
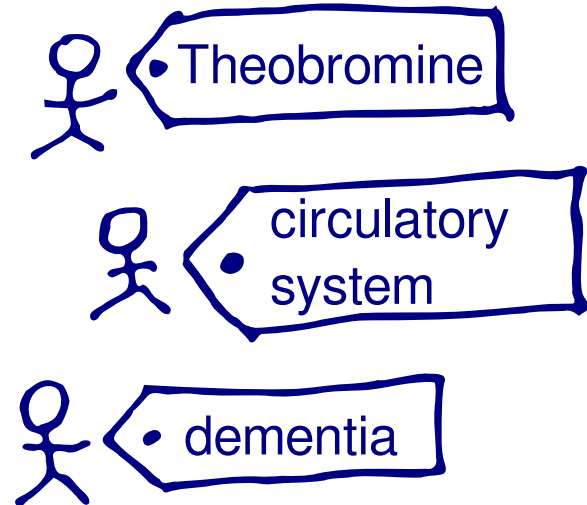
Rank passages with  
entity-BM25.

Issues:

- Entity linkers make mistakes
- Limited entity types
- Many relevant passages without query entities

Advantage:  
+ Disambiguation

# Different Queries - Different Entities

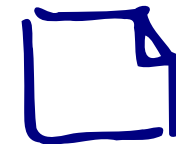
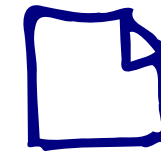
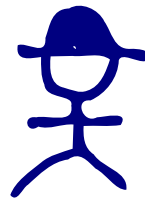
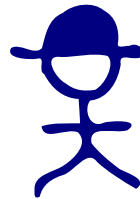
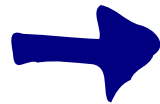
Query	EU UK relations	dark chocolate health benefits
Query entities		
Relevant entities		
	<b>Named Entities</b>	<b>Concept Entities</b>

# Document Retrieval with Entities

Query

Entities

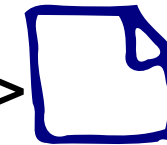
Documents



Entities believed  
to be relevant ->



Docs we  
want to rank ->



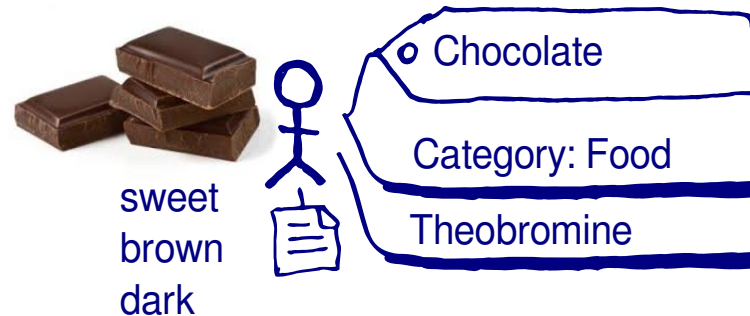
# Approaches: Utilizing KGs for Text IR

---

1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion

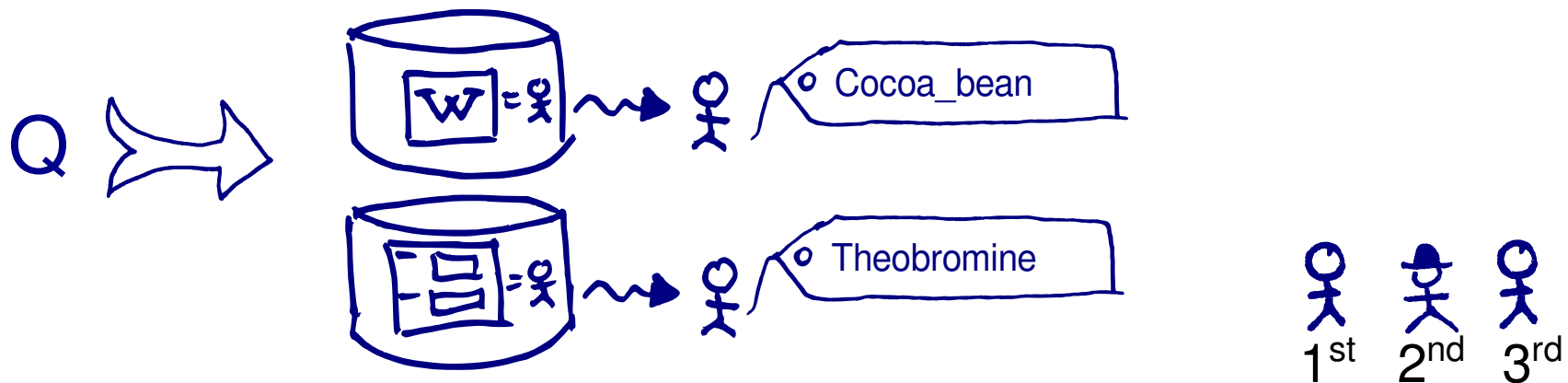
# Query Entities through Entity Linking

Query: dark chocolate health benefits



# Relevant Entities through Object Retrieval

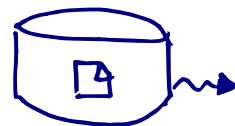
Retrieve entities from knowledge graph to obtain ranking of entities  (with score)



[Pound10, Balog11, Zhiltsov15, Dalton14, Xiong15]


Notation:

Search Index

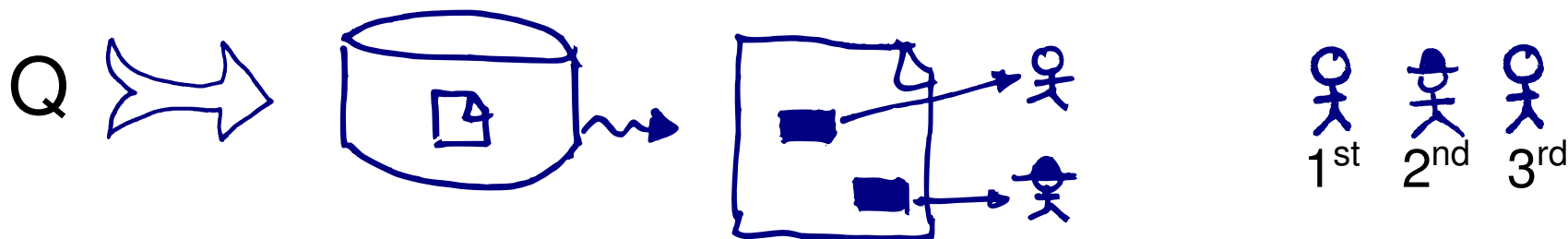




# Relevant Entities through Pseudo-Rel. Feedback


1. Retrieve document ranking
2. Entity link documents in top K
3. Derive distribution over  (bag of entities)  
(see Relevance Model / RM3)

[Lavrenko01; Dalton14, Liu15, Schuhmacher15]

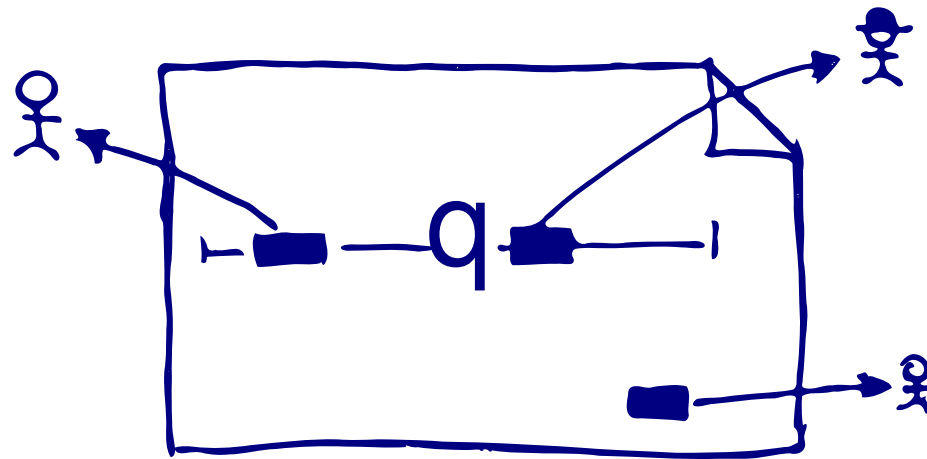


Issue: entities not necessarily near query terms.

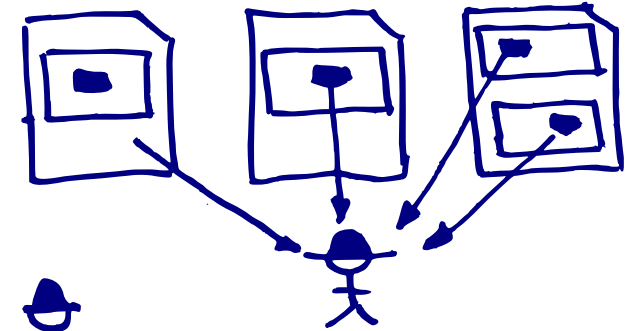
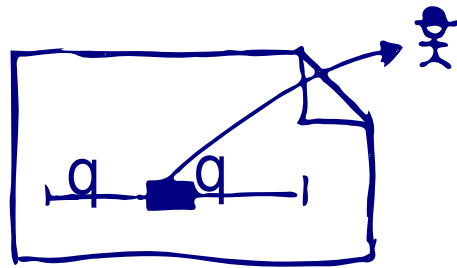
# Relev. Entities through Proximity to Query Words

Using distance between entity mentions and query words **q** as a measure for  relevance.

[Petkova & Croft 07; Liu & Fang 15]

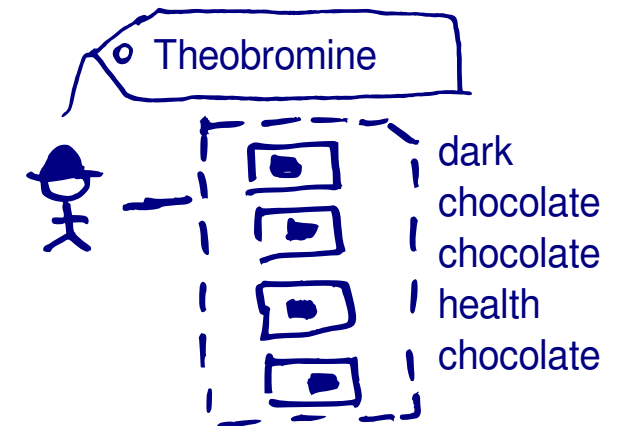


# Relevant Entities through Entity Context Model



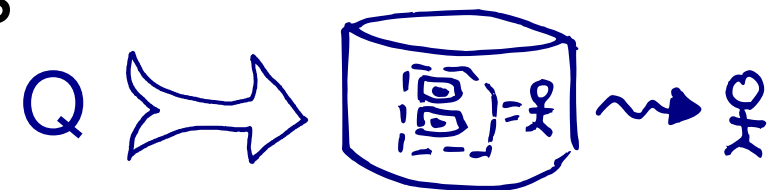
1. Collect contexts of entity links to 

2. Concat link context into one pseudodoc per entity

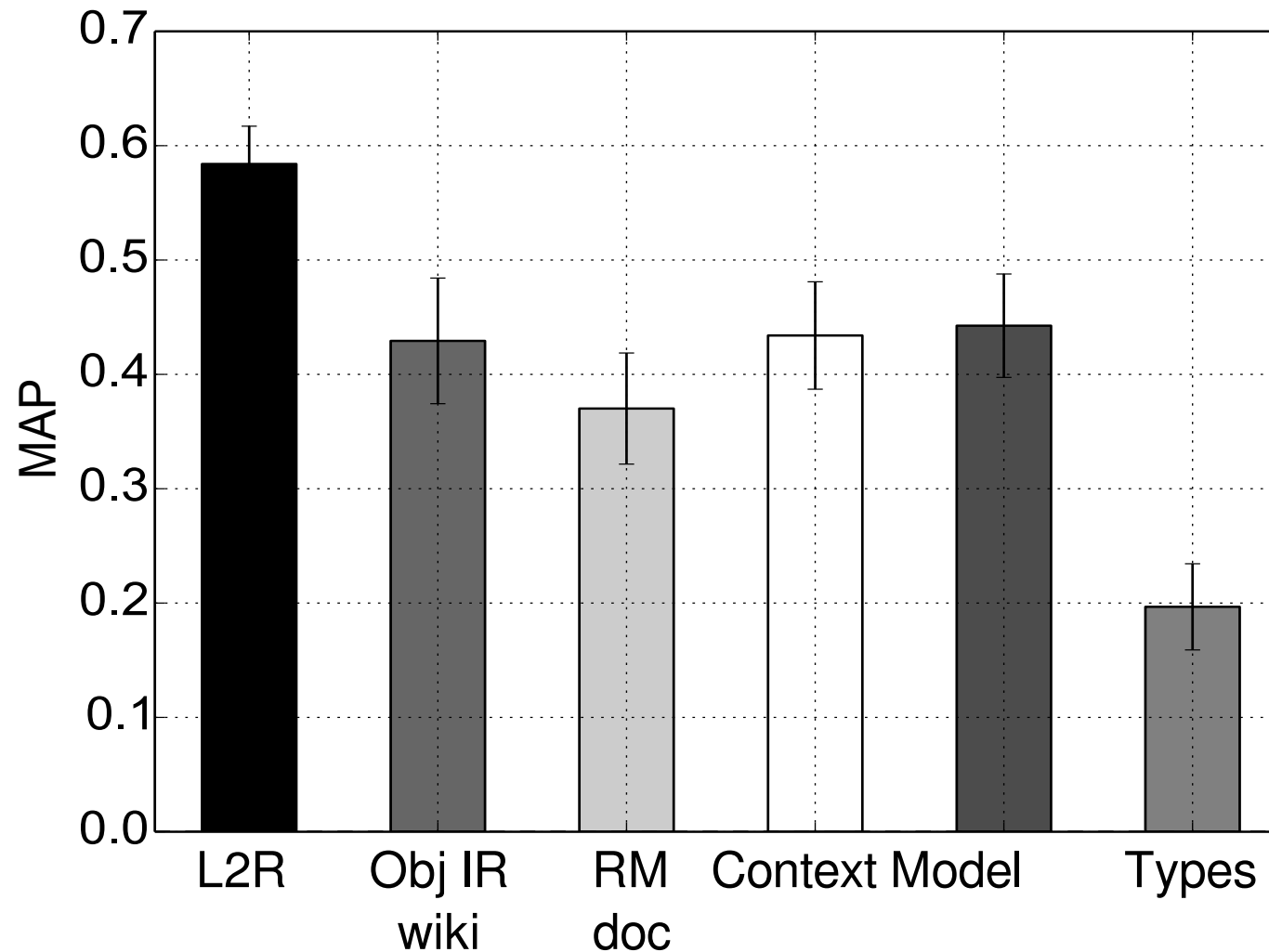


3. Given query  $Q$ , retrieve pseudodocs, thereby ranking entities

[Dalton 14, Liu 15]

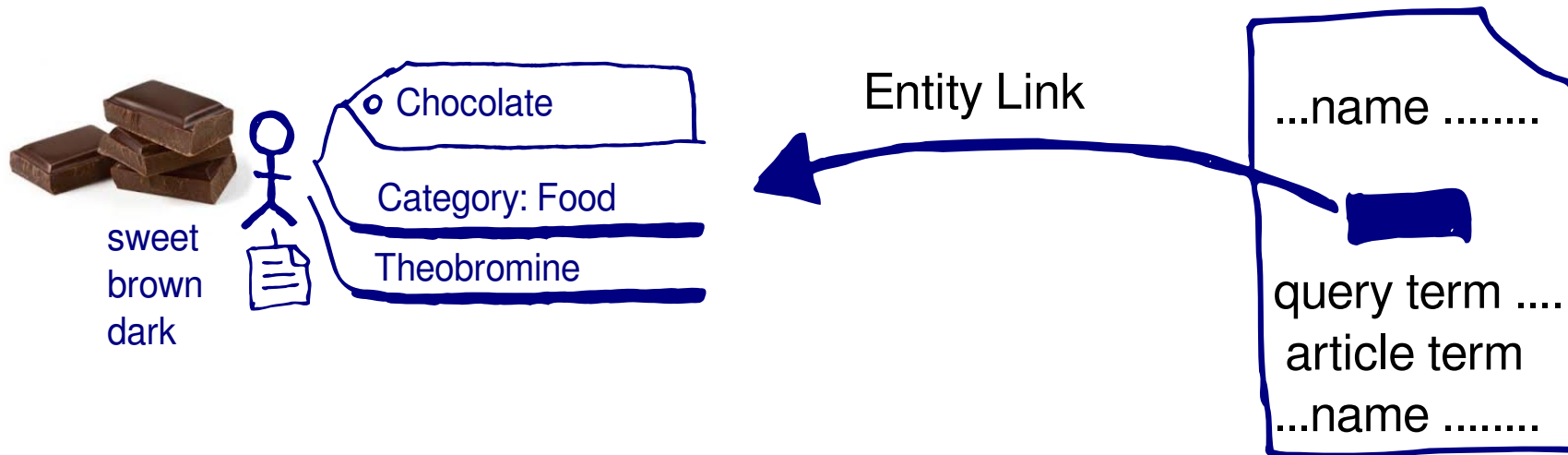


# Entity Ranking Evaluation on ClueWeb12



Evaluation Data: <http://rewq.dswlab.de/>

# Retrieval Models over Terms, Names, Links



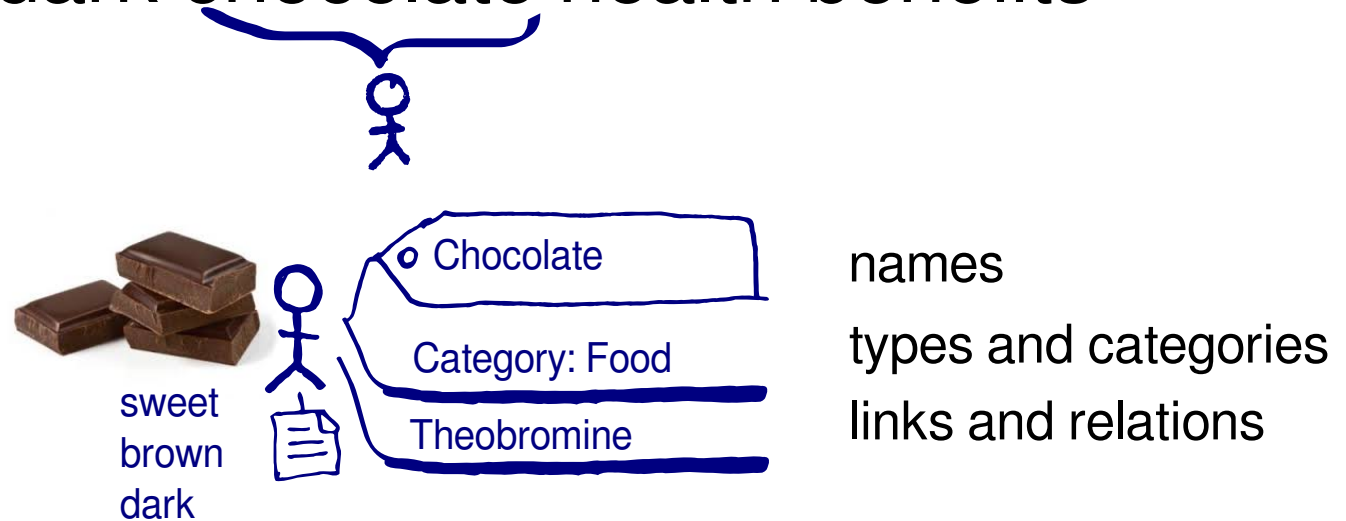
$$score(\text{document}) = \lambda_1 \text{query terms} + \lambda_2 \text{names} + \lambda_3 \text{entity links} + \lambda_4 \text{article terms} + \dots$$

use your favorite retrieval model here!

# So Far: Entities as Tags

But knowledge graphs contain so much more information!

Query: dark chocolate health benefits



How can we make use of the information?

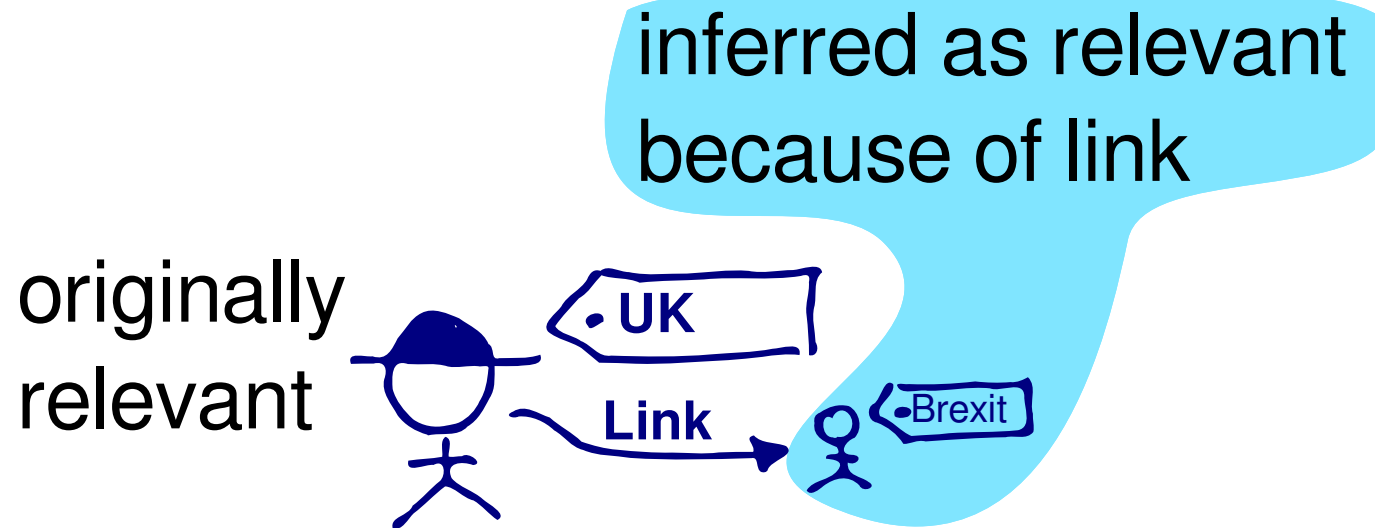
# Knowlegde Graph Expansion

---

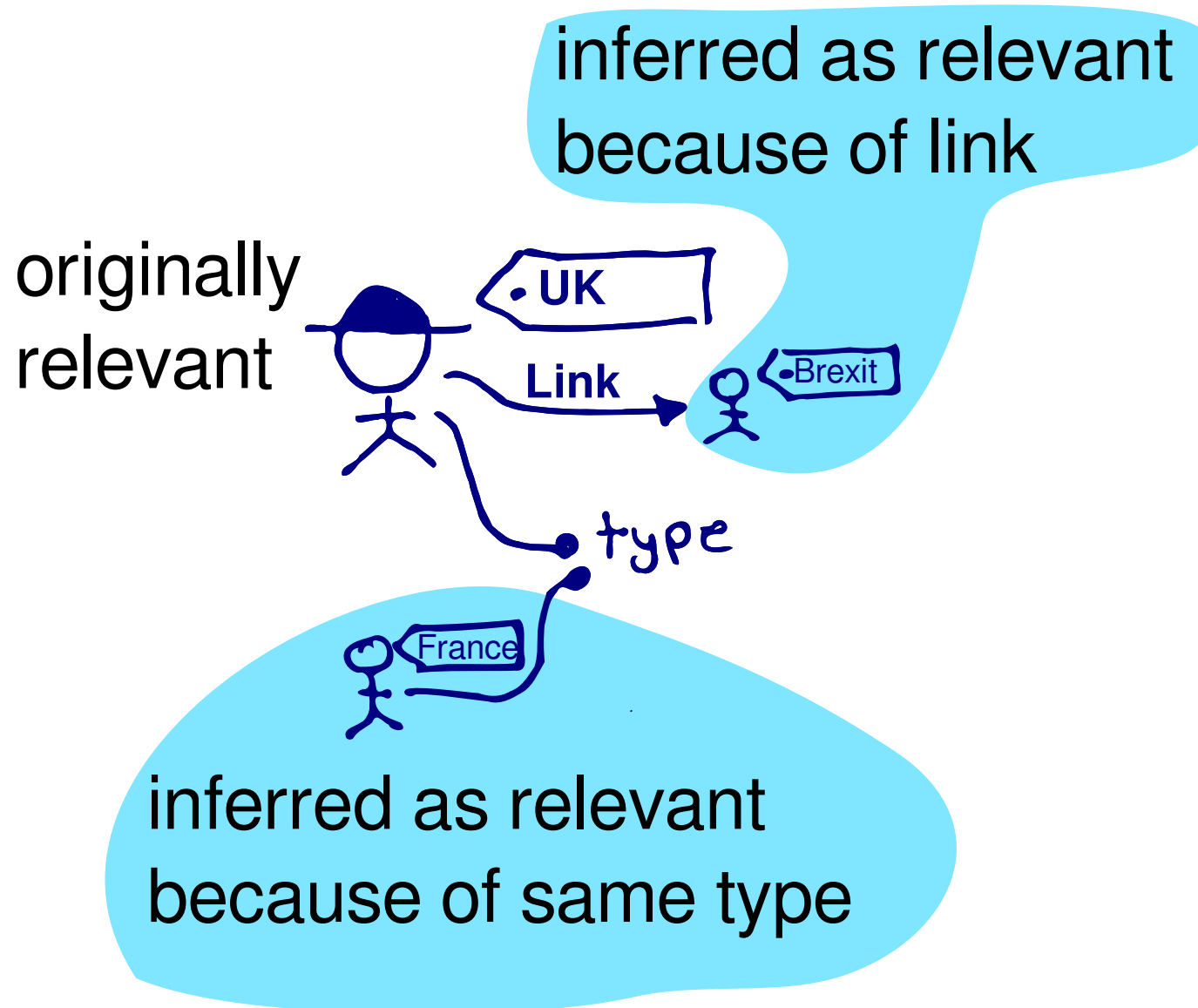
1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion



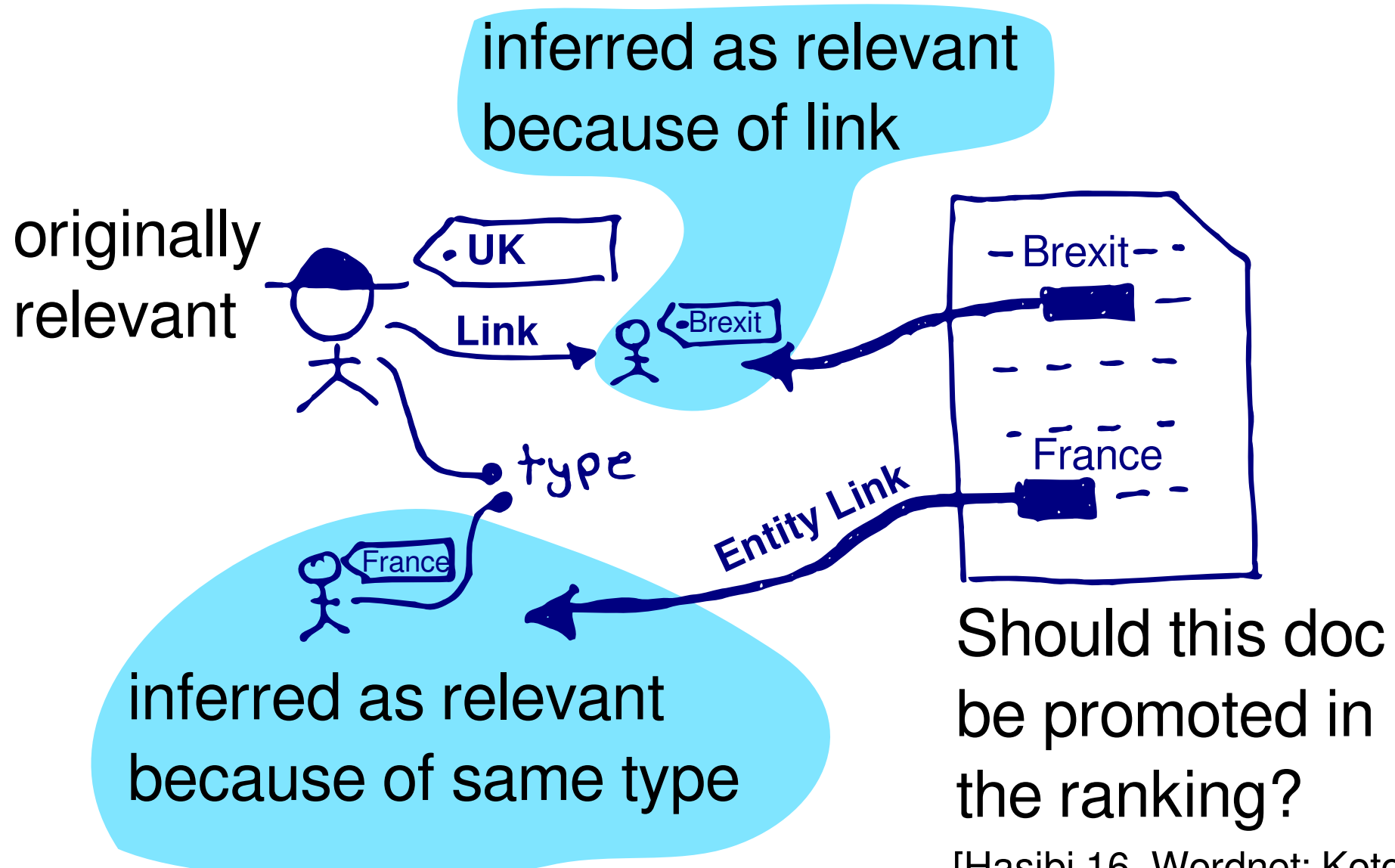
# Using Relations and Types with Entity Links



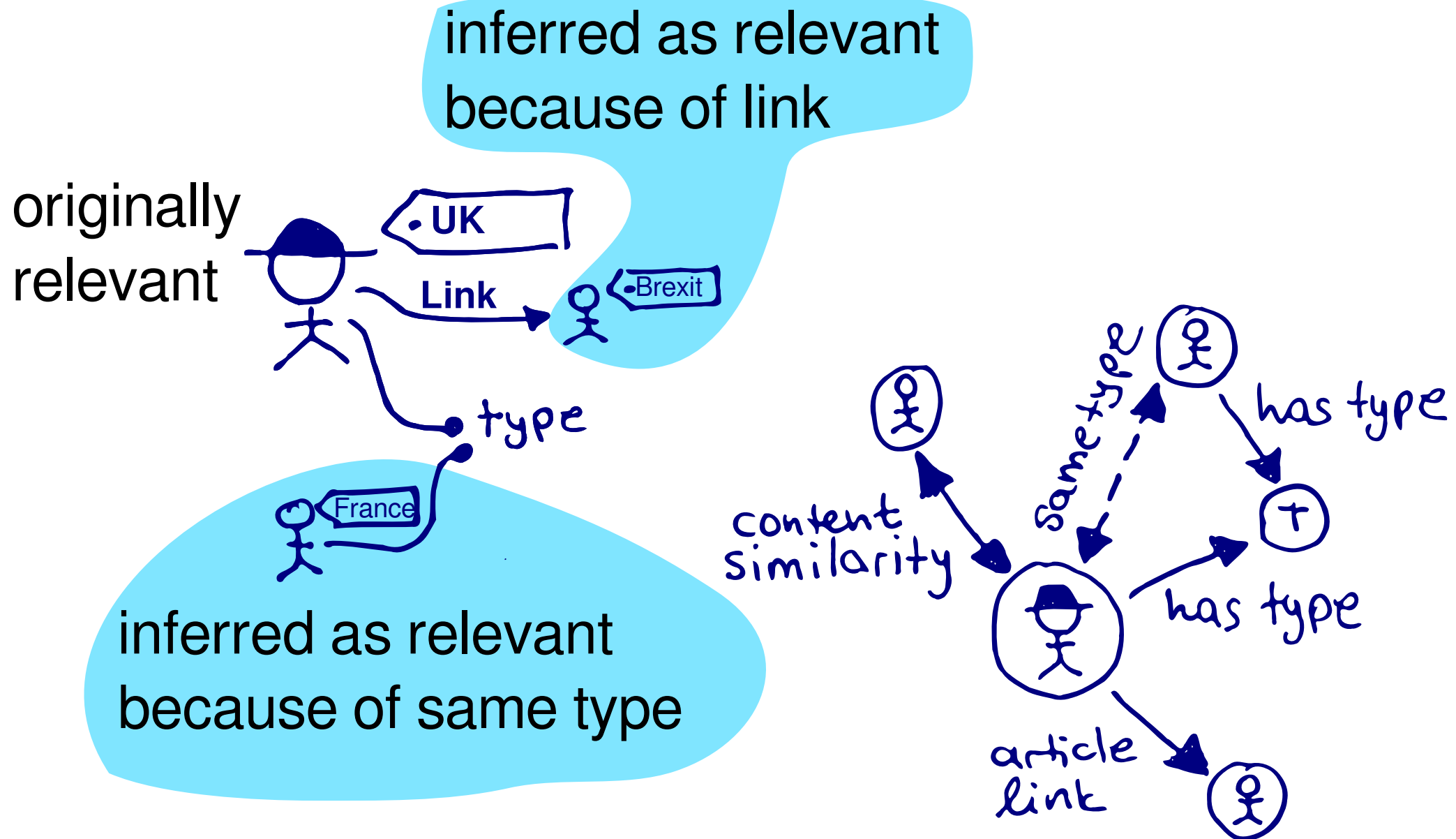
# Using Relations and Types with Entity Links



# Using Relations and Types with Entity Links



# Using Knowledge Graph Structure

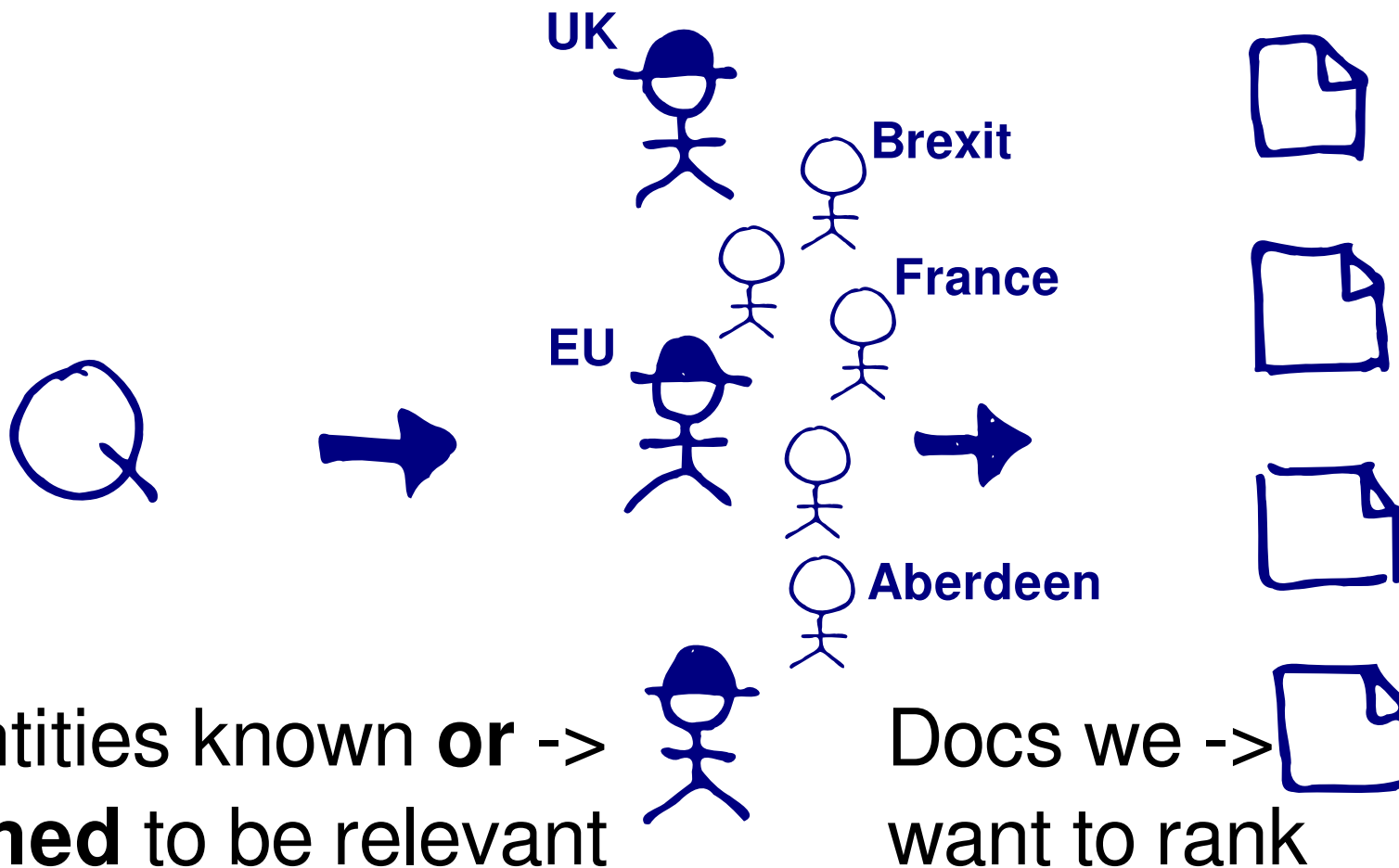


# Document Retrieval with (More) Entities

Query

Entities

Documents



# Boston et al 2013: Wikimantic: Toward effective ...

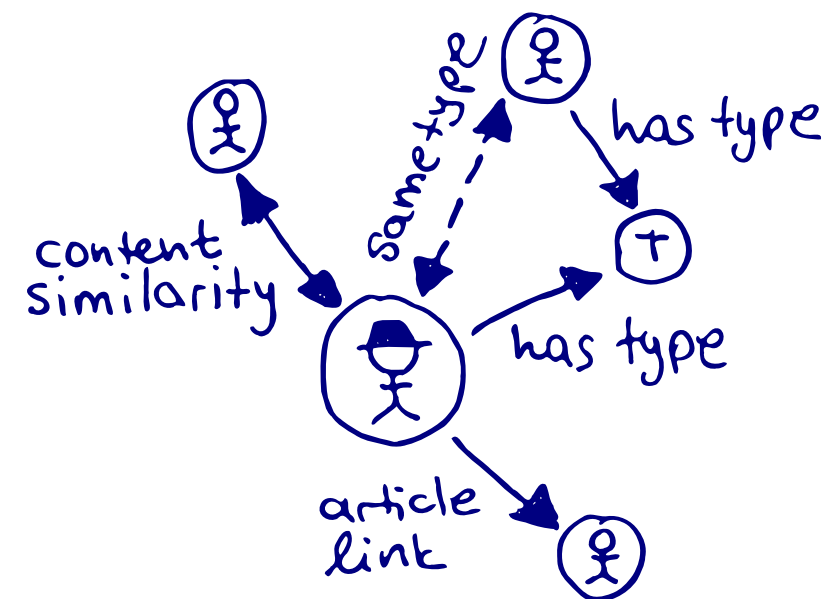
Weight entities by:

M: How well **E**s article content matches the query

MR: How often **E** is linked by others (PageRank)

Method	F1 on TREC QA
content	76.92
content +d*graph	79.47

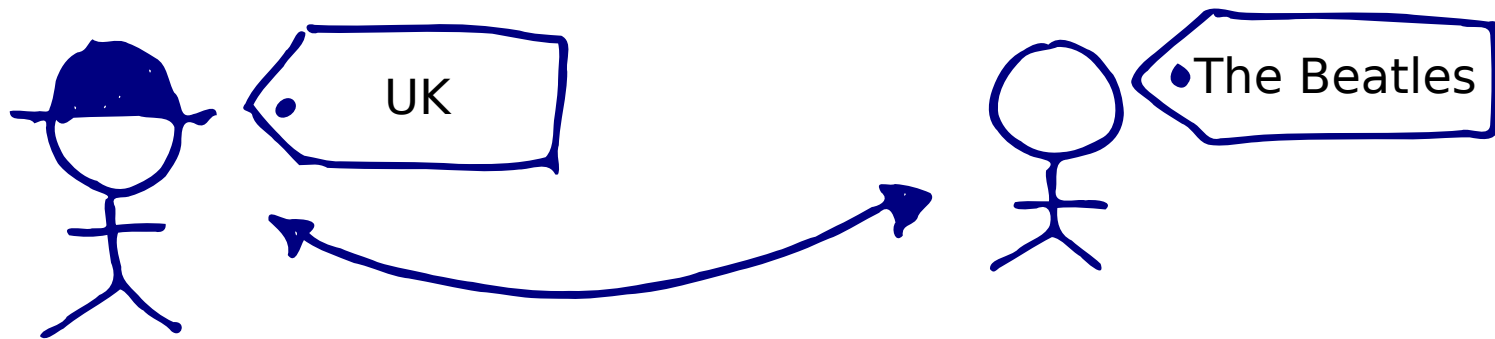
d=0.0001



# KG expansion: A Potential Issue

Example query: EU UK relations

Consider:



**Correct connection, but:**

The connection is not relevant in the context of "UK" as in "EU relations".

**If** we would promote docs because they talk about The Beatles, we are hurting the ranking quality.

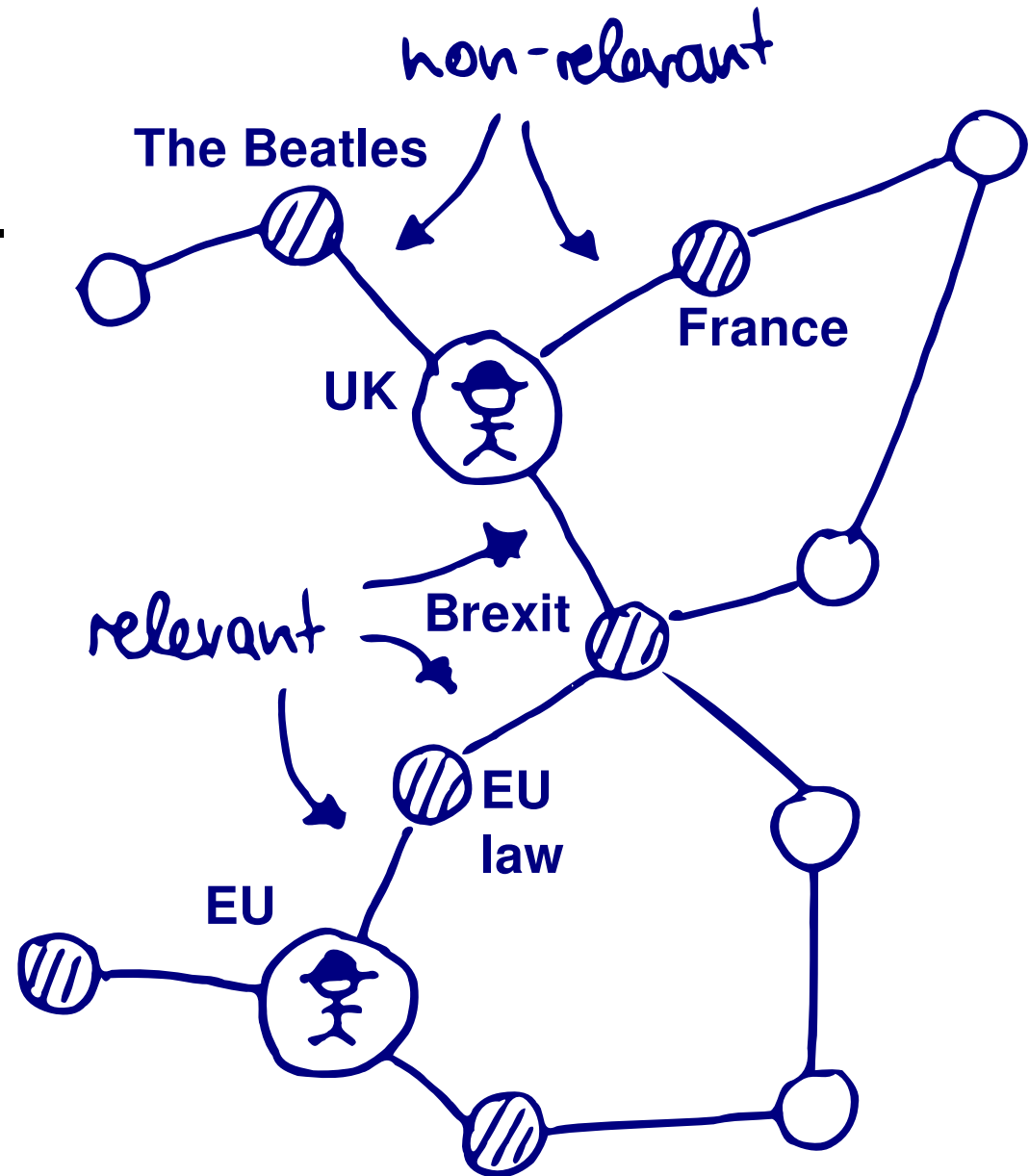


# General Approach: Graph Expansion

Many connections  
in a knowledge graph.

Only few are relevant!

Expanding with  
non-relevant entities  
leads to low precision  
rankings.



# Weight Edges / Nodes in the Knowledge Graph

Popularity measures:

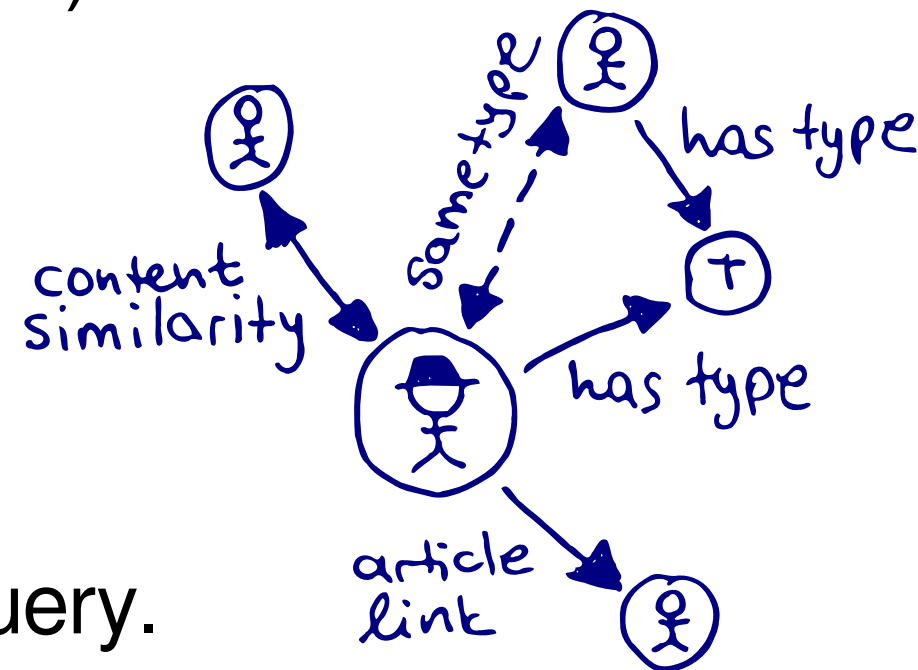
- Graph walks: PageRank / HITS
- Degree

Connectivity measures (seeds):

- Shortest paths
- Entity relatedness

Graph clustering

Issue: Do not consider the query.



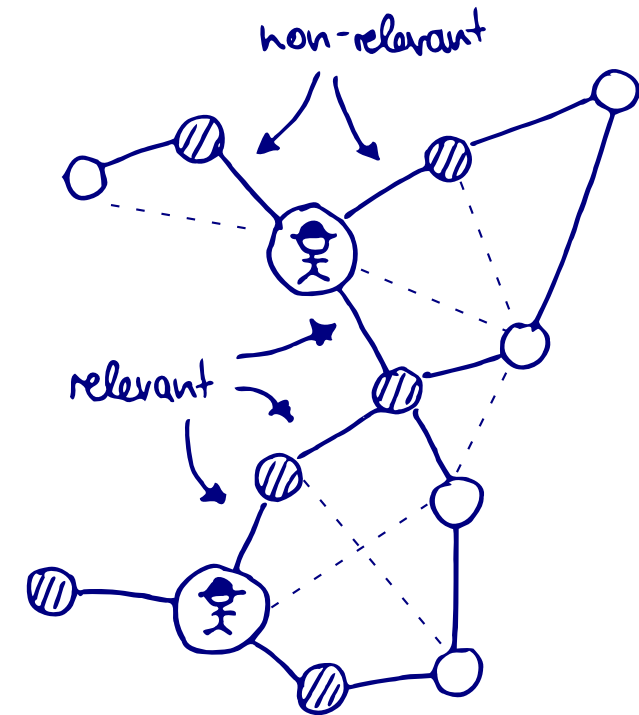
# Entity Aspects and the Graph Structure

An **open issue** remains:

- Entities have multiple aspects
- Graph = overlay of all aspects

Growth of KGs leads to

- better coverage of relevant facts
- many more spurious facts :-)



When relevant  $\neq$  popular:

How to tell which edges are relevant?

# Relation Extraction (for Relevant Relations)

---

1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion

# Fix: RM3 + Graph Walk

UK - The Beatles:

Can be solved with Pseudo-relevance Feedback.

1. Retrieve documents for Q
2. Delete edges to entities that are not mentioned

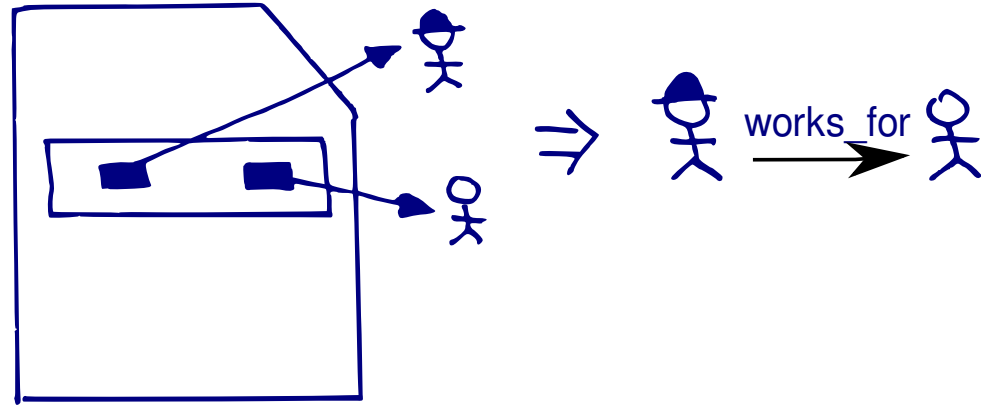
But:

non-relevant relations remain

lead to erroneous entity expansions

# Task: Extracting Relevant Relations

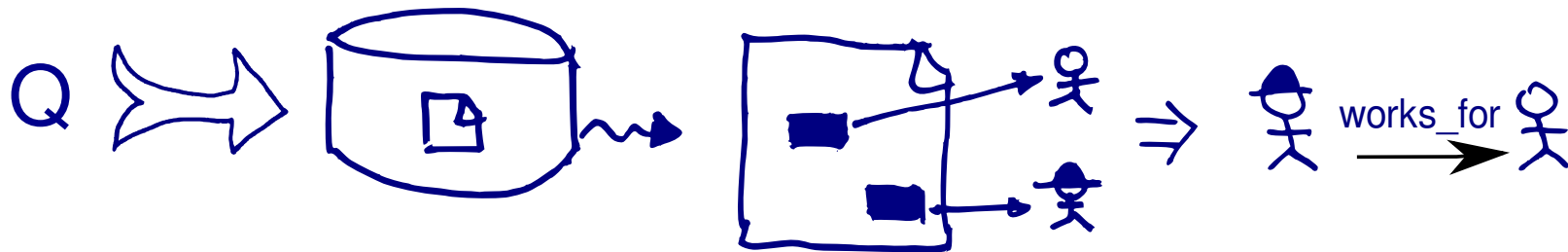
Relation Extraction:



Research question:

relevant documents + extraction = relevant relations?

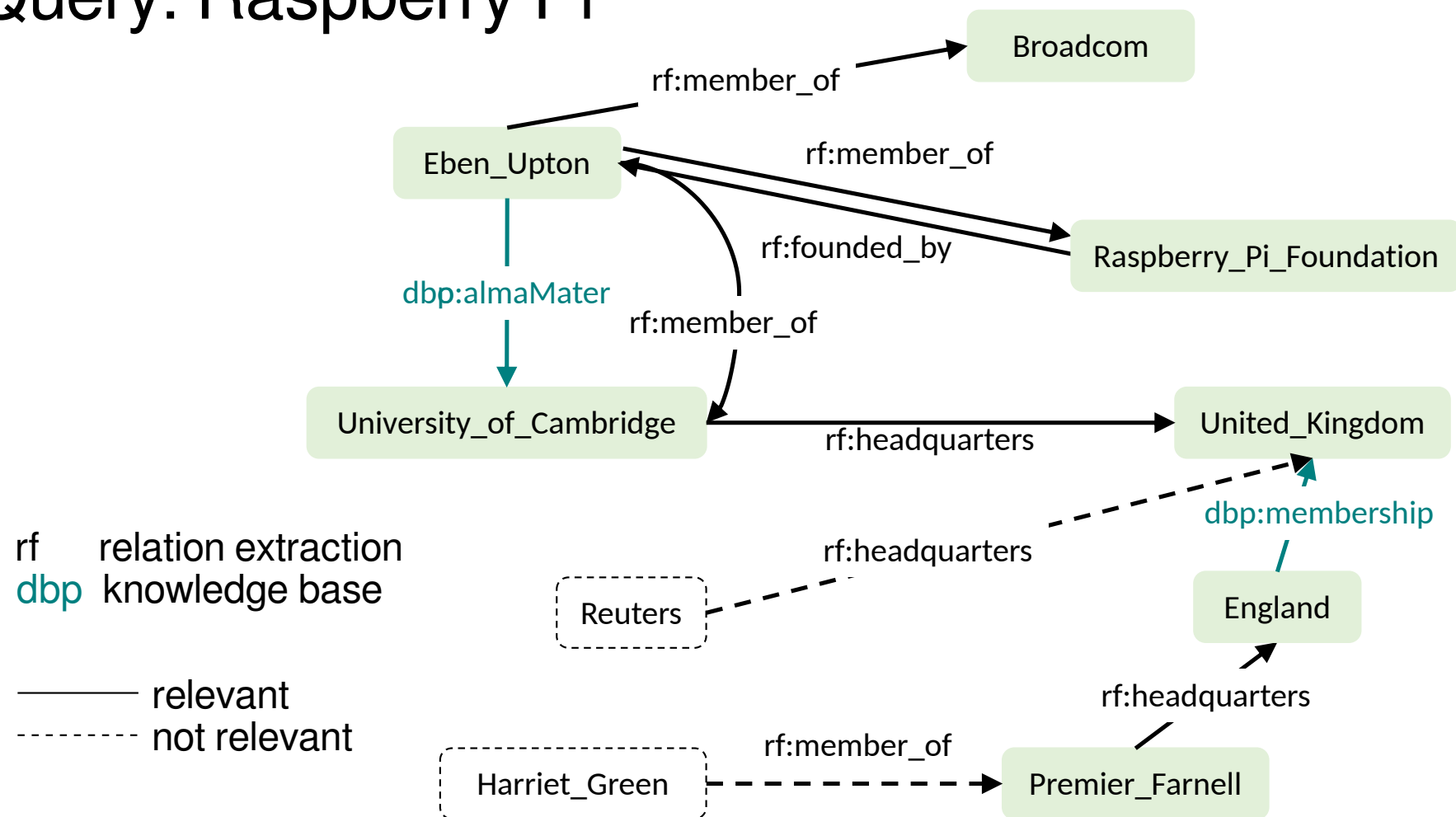
[Schuhmacher 16]



# Relevant Relations through Relevant Documents

Goal: Relations need to be relevant and correct

Query: Raspberry Pi



# Big Question: Edge Relevance



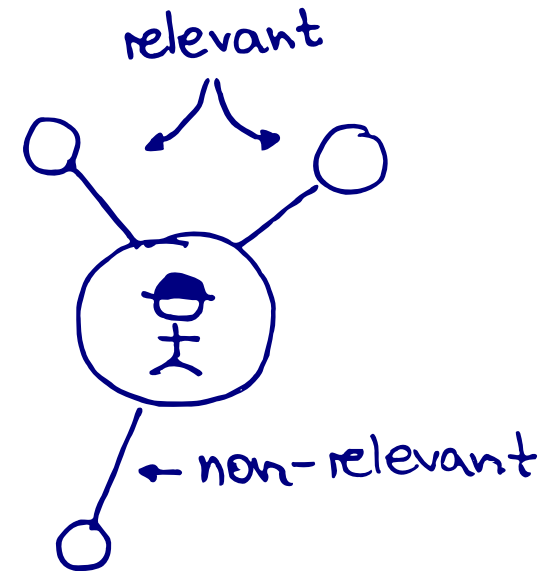
How to infer which other connected entities / nodes are relevant for the information need Q?

...and therefore safe for

- expansion
- and promotion in entity ranking?

Not just those with

- many connections (PageRank)
- mentioned in feedback docs
- extracted with relation extraction





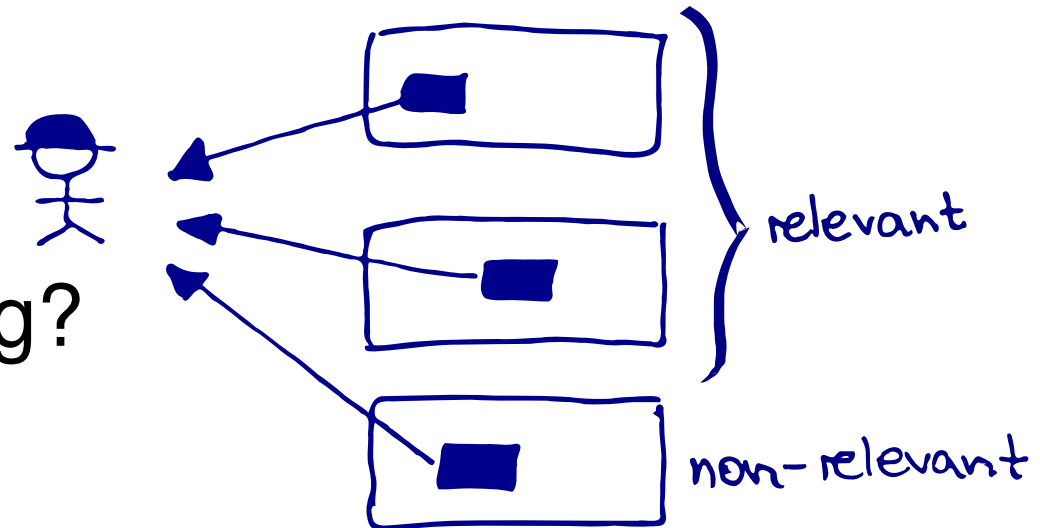
# Big Question: Context Relevance



How to infer which contexts of entity links are relevant for the information need Q?

...and therefore safe for

- expansion and
- promotion in psg ranking?



Not just those with

- popular words (RM3)
- frequent entity mentions

# Entity Aspects

---

1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion

# Entity Aspects

Danger: An entity is relevant, but:  
only because of one aspect  
=> many non-relevant aspects of relevant entities.

Example aspects about UK:

- still a member of the European Union
- is a constitutional monarchy
- the Raspberry Pi was invented in the UK
- there are many great UK bands

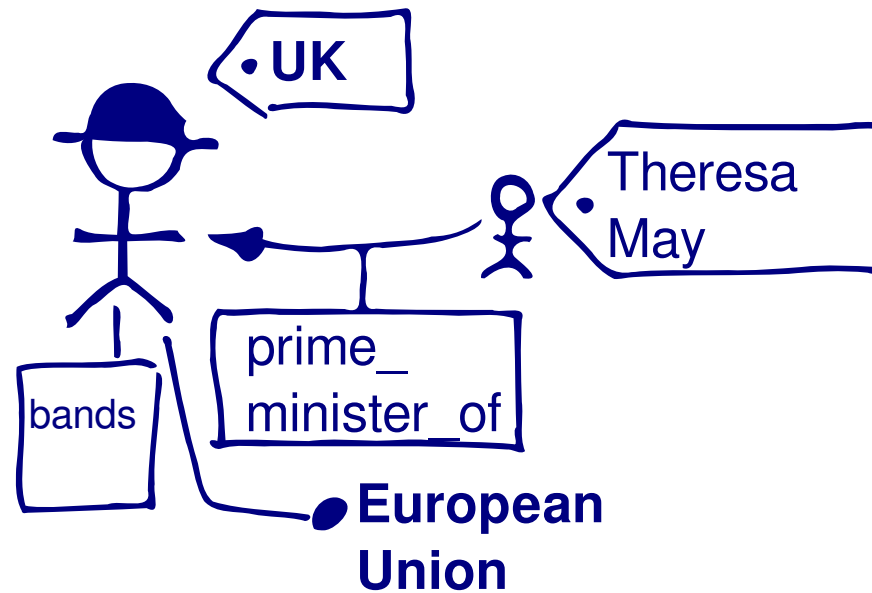
Depending on query, some are relevant, some not.

# How to Represent Entity Aspects?

As terms?	UK bands brexit
As types?	UK member of "European Union"
As is-a?	UK as a European country
Related entities?	UK Theresa_May
Relations?	Theresa_May prime_minister_of UK
Language Model	$p(\text{brexit})=0.4$ $p(\text{leave})=0.25$ $p(\text{immigration})=0.10$

[Reinanda SIGIR15, Liu IRJ15, Prasojo CIKM15]

# Entity Aspects: Using KG ...

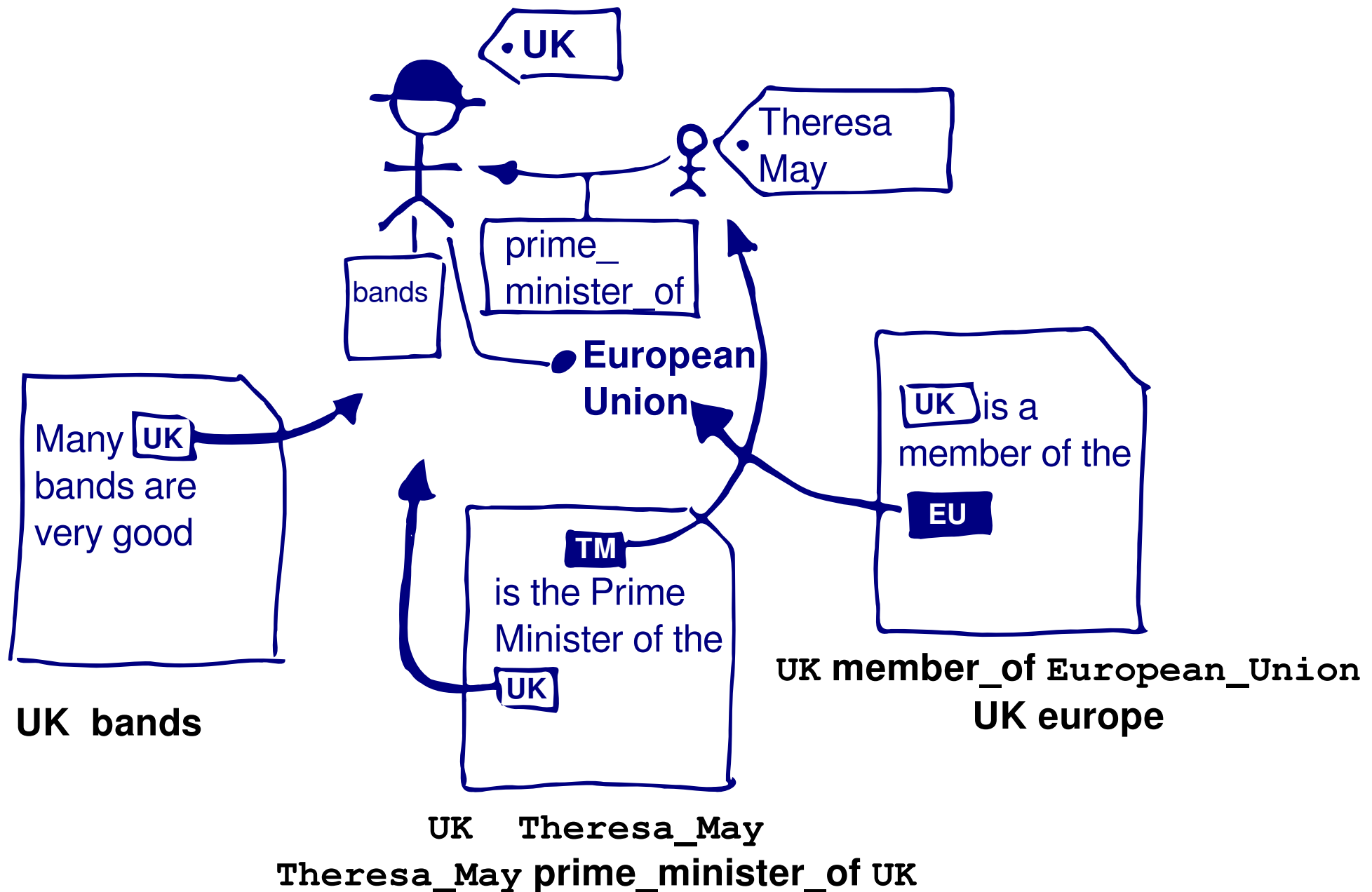


UK bands

UK member\_of European Union  
UK europe

UK Theresa\_May  
Theresa\_May prime\_minister\_of UK

# Entity Aspects: Using KG and Text



# Entity Aspects: Infer Relevance, Match, Extract

Use KG + text to model for each relevant entity:

- what are different aspects of the entity?
- which aspects are relevant?
- how are relevant aspects best represented?

Generic pattern:

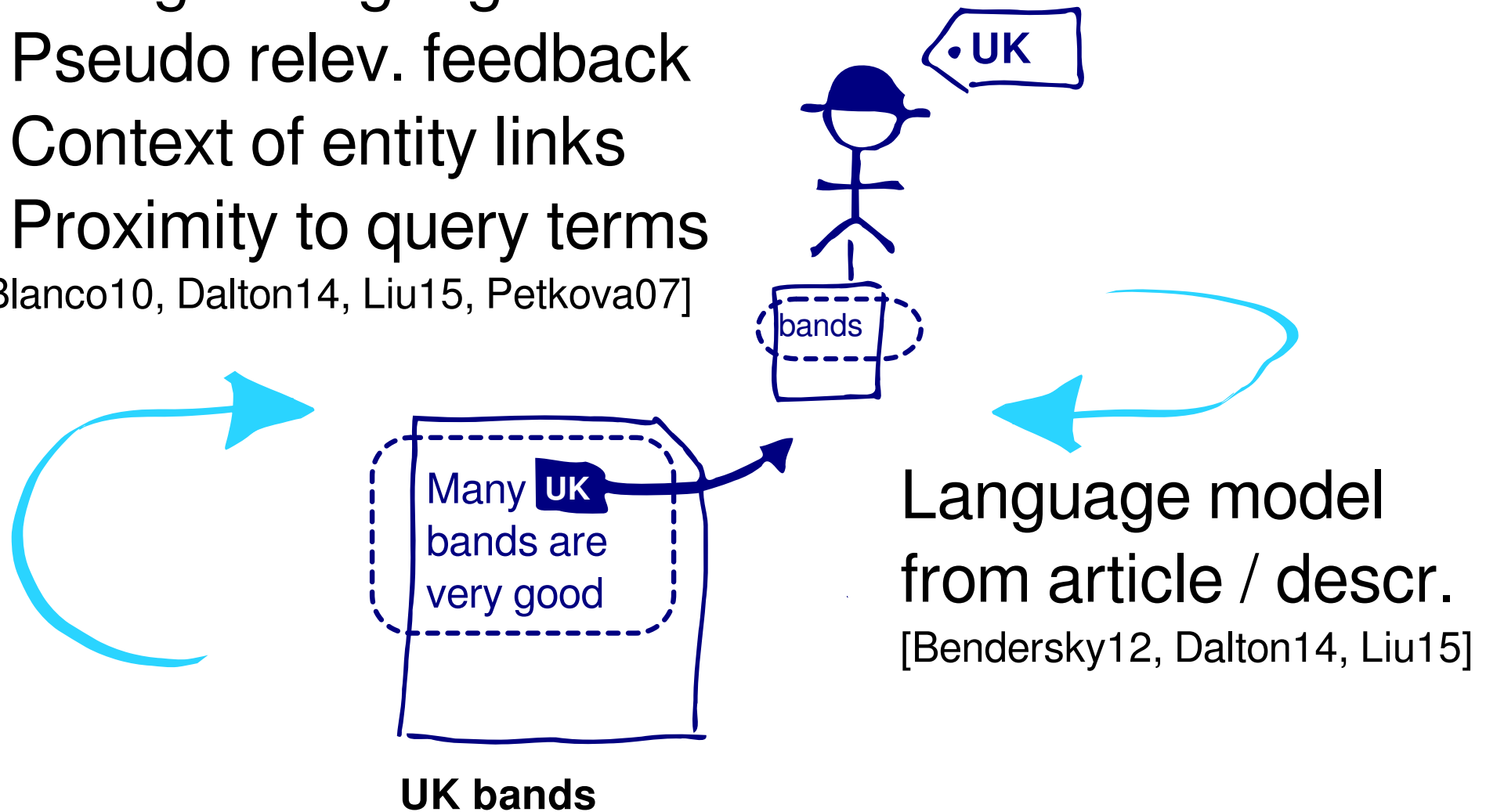
1. Information extraction
2. Relevance prediction
3. Matching (inverse extraction)

# Entity Aspects as Terms

## Passage-Language Model

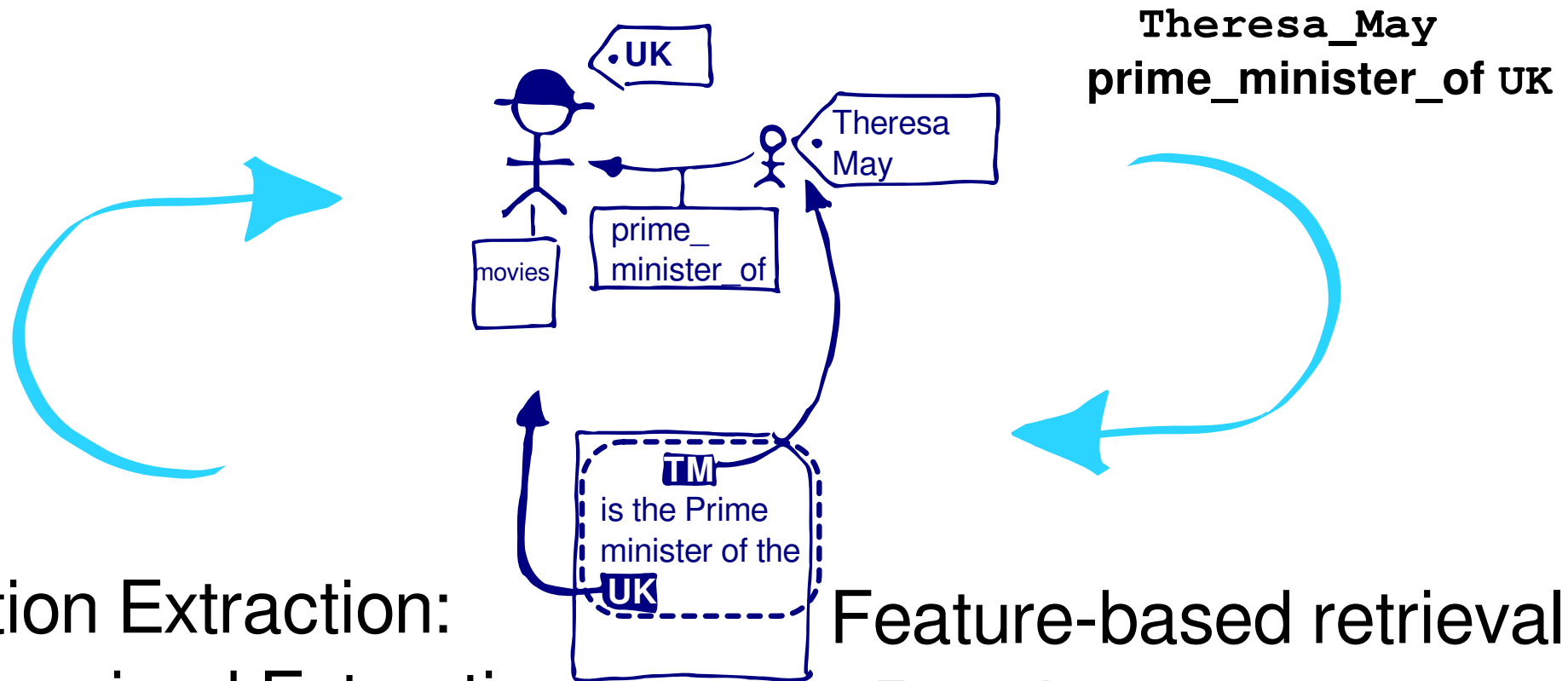
- Pseudo relev. feedback
- Context of entity links
- Proximity to query terms

[Blanco10, Dalton14, Liu15, Petkova07]





# Entity Aspects through Relations (Triples)



Relation Extraction:

- Supervised Extraction from Text

[Schuhmacher ECIR16]

Feature-based retrieval:

- Relation terms
- Cosine of word vectors

[Voskarides ACL15]

Infer & Extract Aspects

Match Aspects

# Conclusion

---

1. Introduction
2. Vision
3. Approaches: Utilizing KGs for Text IR
4. Knowledge Graph Expansion
5. Relation Extraction
6. Entity Aspects
7. Conclusion

# Conclusion: Retrieving Knowledge from the Web

Many "prob-pportunities" when retrieving detailed answers

- Relevant KG edges/elements?
- Relevant contexts of entities?
- Relevant entity aspects?

xkcd.com/1592/



Slides online: [www.cs.unh.edu/~dietz](http://www.cs.unh.edu/~dietz)

KG4IR Workshop at SIGIR (+mailinglist)  
TREC Complex Answer Retrieval track  
Tutorial Utilizing KGs for Text-centric IR

<http://kg4ir.github.io>

<http://trec-car.cs.unh.edu>

[github.com/laura-dietz/tutorial-utilizing-kg](https://github.com/laura-dietz/tutorial-utilizing-kg)

Looking for students & postdocs!

[dietz@cs.unh.edu](mailto:dietz@cs.unh.edu)