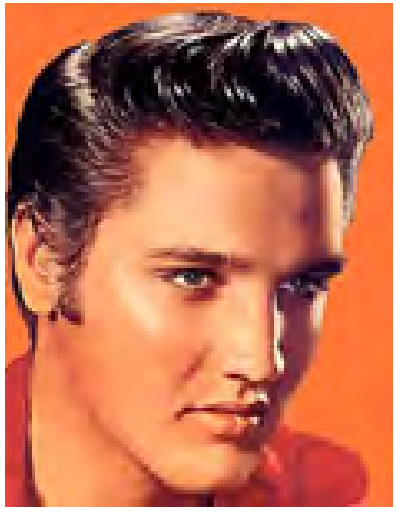
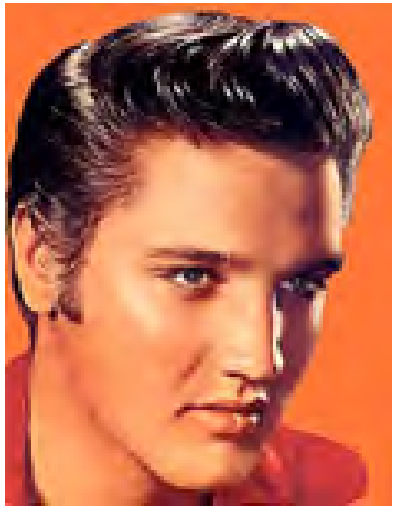
Intelligence artificielle et les Bases de Connaissances

Fabian M. Suchanek

I am an Elvis Fan!



We visited Elvis in New Zealand



Elvis' lodge in Murchison

New Technology: Knowledge Bases

Apple Siri

New Technology: Knowledge Bases

Apple Siri

What is the capital of New Zealand?
“Wellington”

Amazon Echo

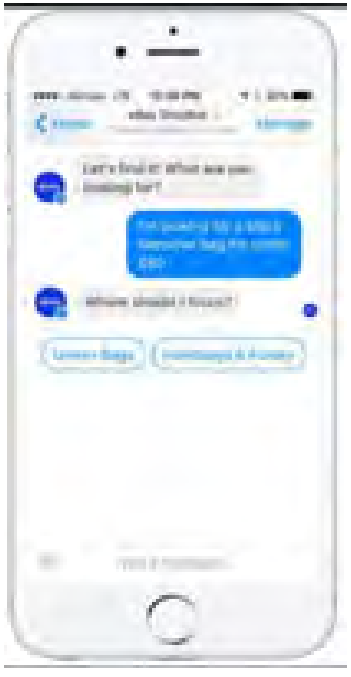
Discovered 6 kineasis proteins that relate to cancer

IBM Watson

New Technology: Knowledge Bases

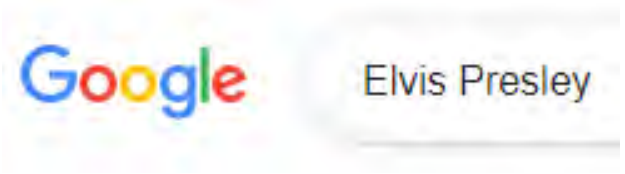
Microsoft Academic

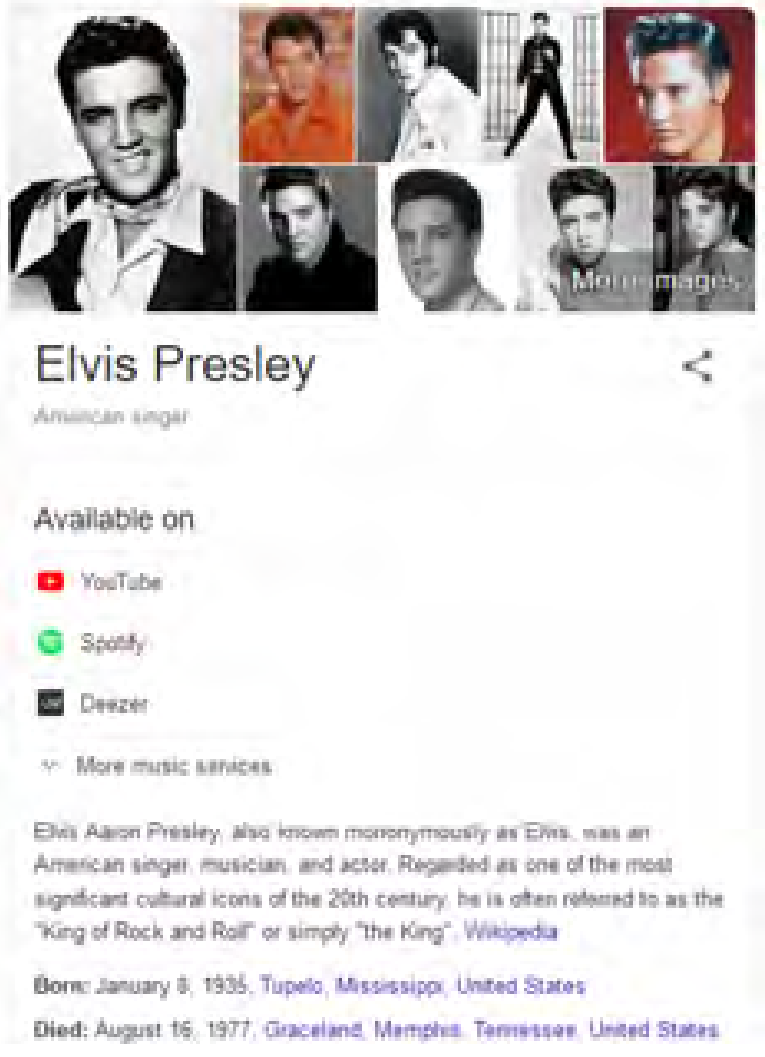
Ebay Knowledge Graph



Facebook

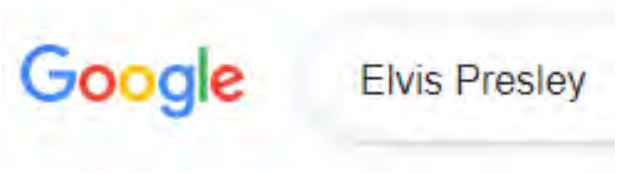
New Technology: Knowledge Bases

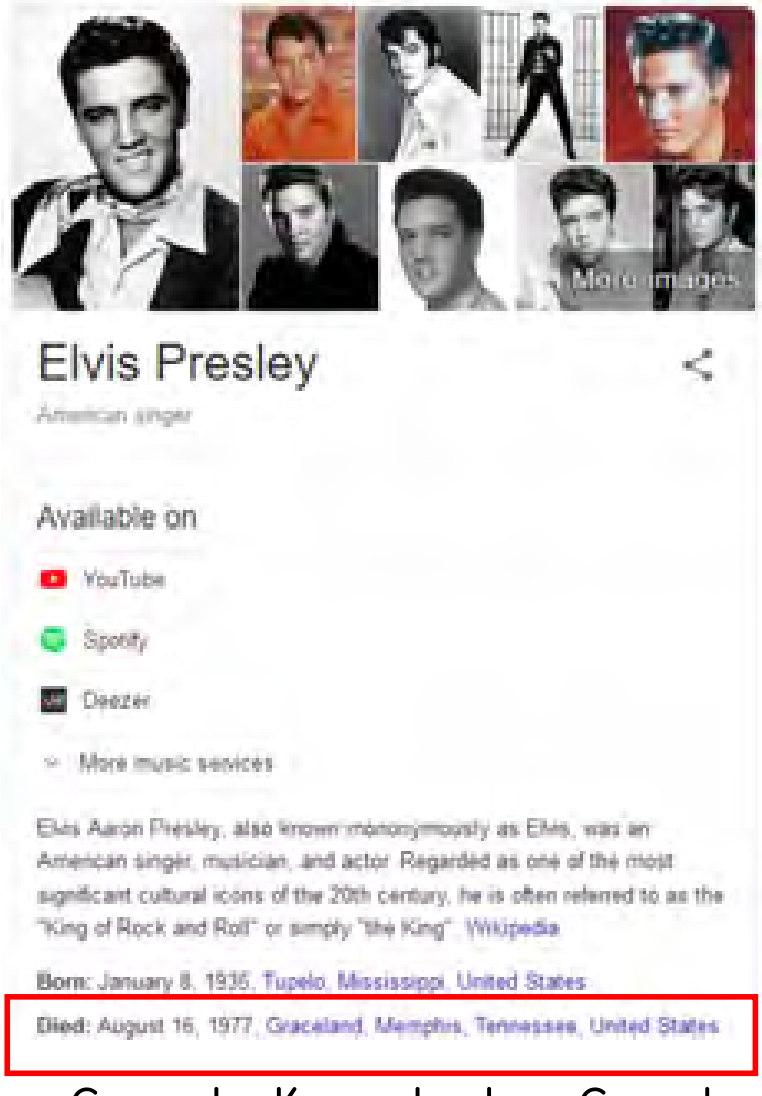




Google Knowledge Graph

New Technology: Knowledge Bases

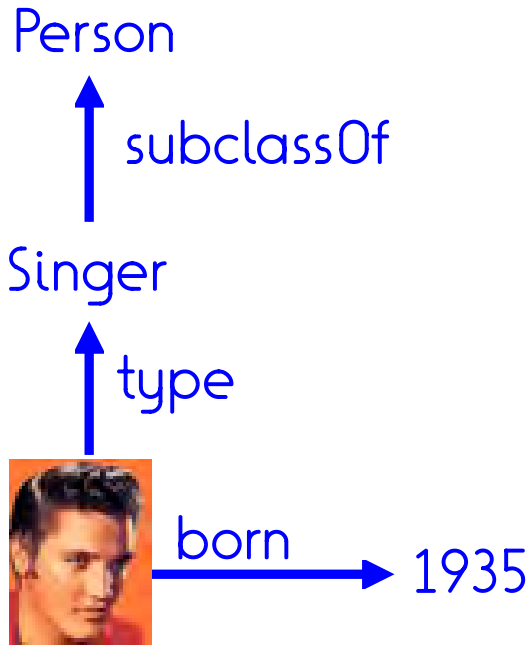




???

Google Knowledge Graph

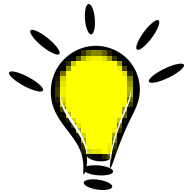
Knowledge Bases



For us, a knowledge base (KB) is a graph, where the nodes are entities and the edges are relations.

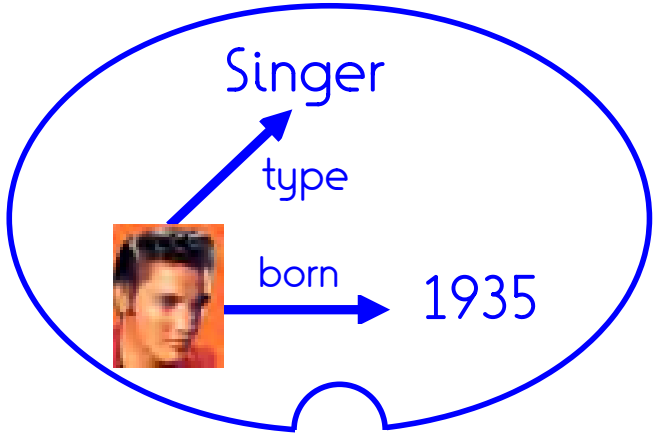
(We do not distinguish T-Box and A-Box.)

Knowledge Base Life Cycle



Using
knowledge bases ->

Constructing
knowledge bases ->



?

Querying
knowledge bases
->

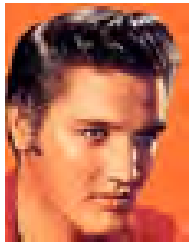
Mining
incompleteness
->

Extracting from Wikipedia

Elvis Presley

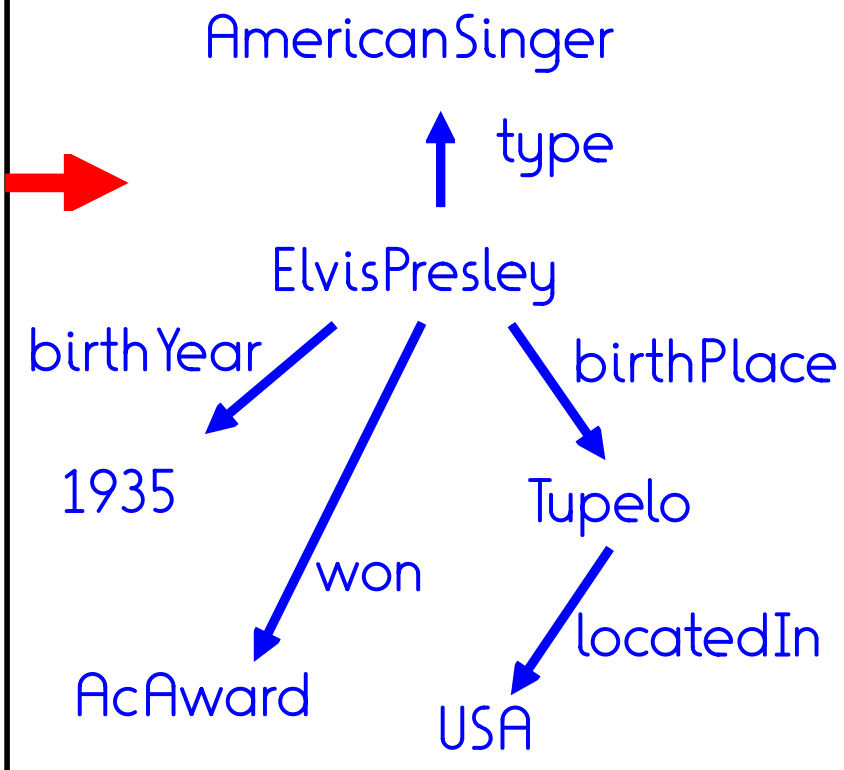
WIKIPEDIA
The Free Encyclopedia

Elvis Presley was one of the best blah blah blub blah don't read this, listen to the speaker! blah blah blah blubl blah you are still reading this! blah blah blah blah blabbel blah

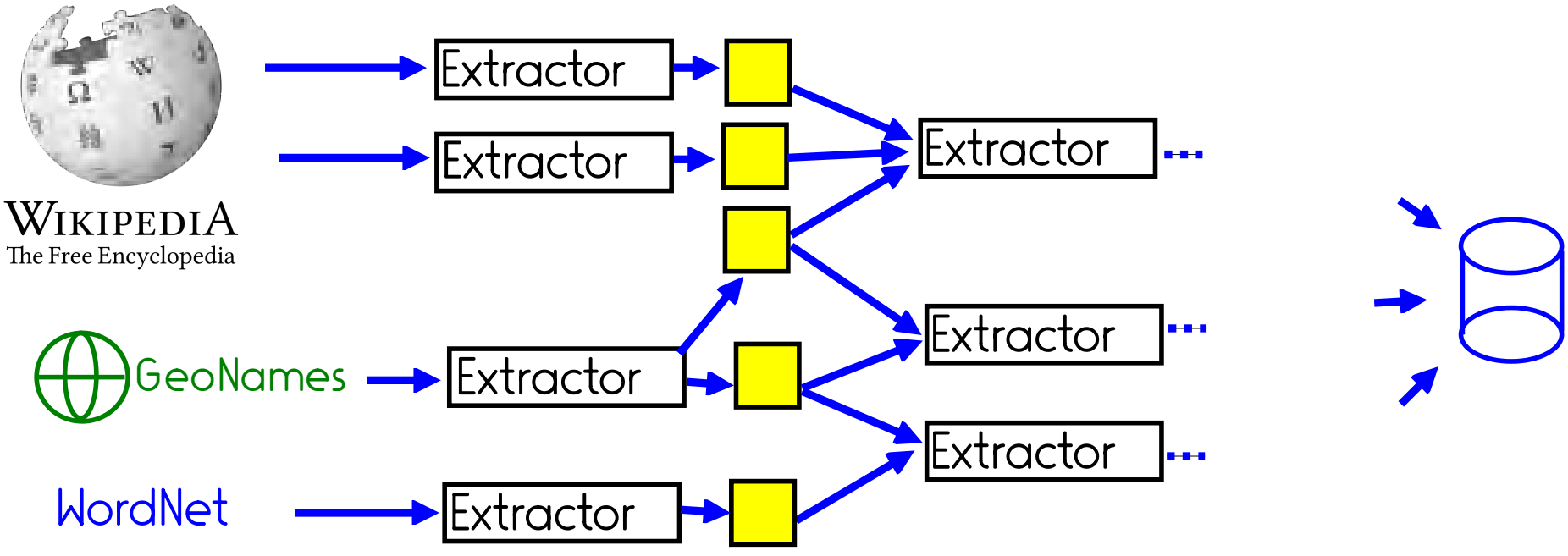


Born: 1935
In: Tupelo
...

Categories:
Rock&Roll, American Singers, Academy Award winners...



Creating a large knowledge base



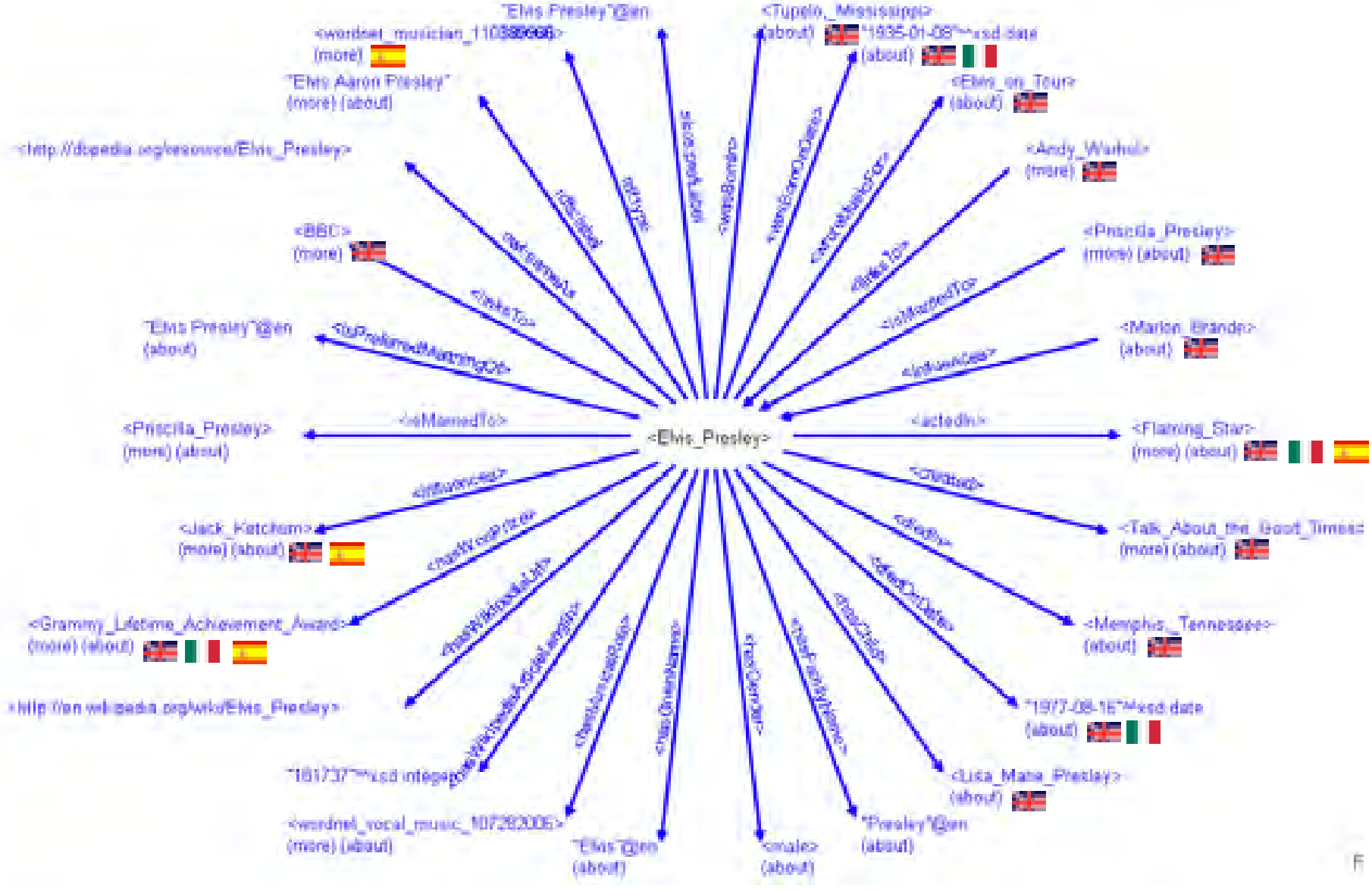
Intermediate extractors

- clean facts
- deduplicate facts and entities
- check consistency

> ensuring high quality (95%)

See it online See it locally

Example: YAGO about Elvis



YAGO: a large knowledge base

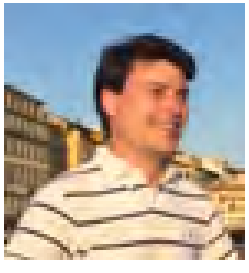
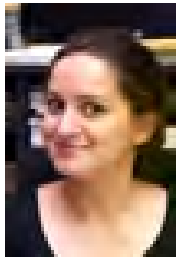
http://yago-knowledge.org
open code and open data

Wikipedia + WordNet
time and space
10 languages
100 relations
100m facts
10m entities
95% accuracy
used by DBpedia
and IBM Watson

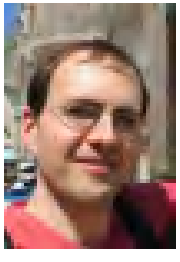


[WWW'07,JWS'08,WWW'11 demo,AIJ'13,WWW'13 demo,CIDR'15,ISWC'16]

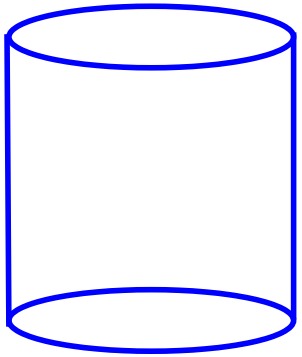
>IBEX



Goal: Harvest entities from the Web

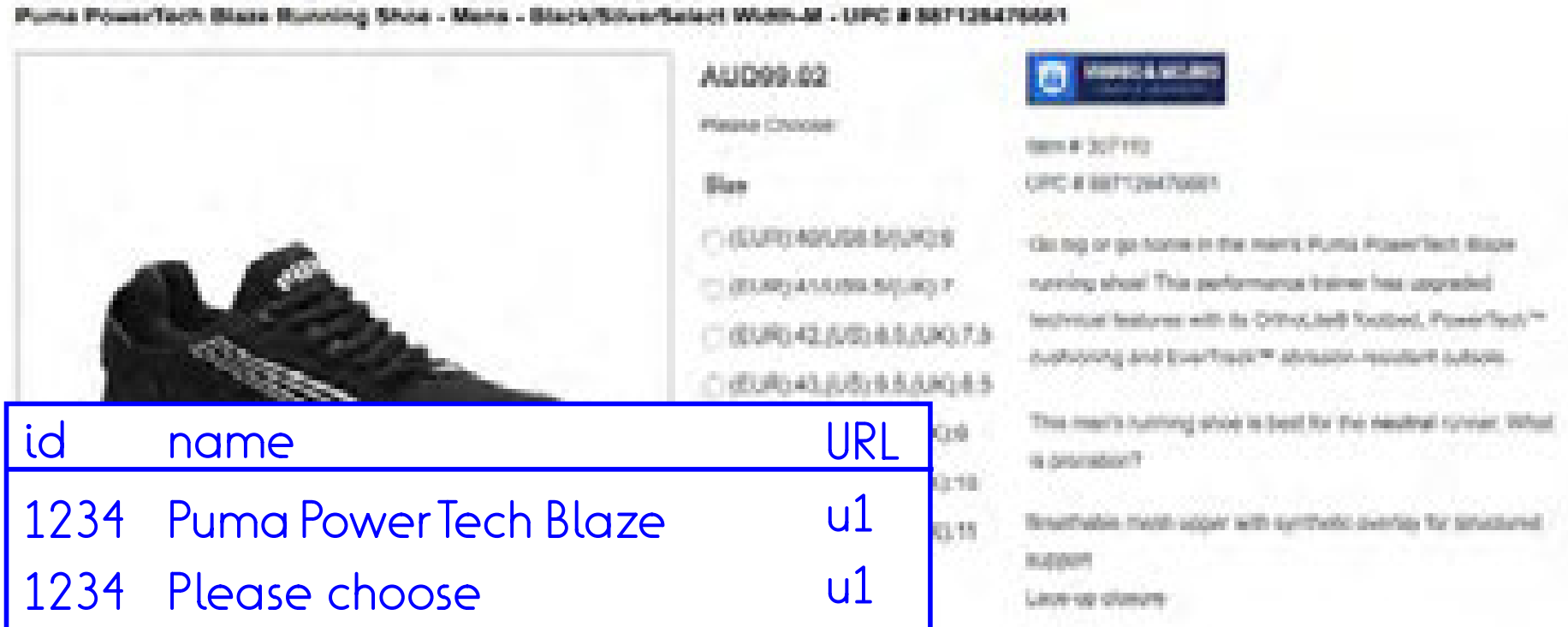


IBEX: Collect unique ids

© Whittleseacitylac

Unique identifiers can be found by a regular expression, and verified by a checksum.

IBEX: Collect unique ids



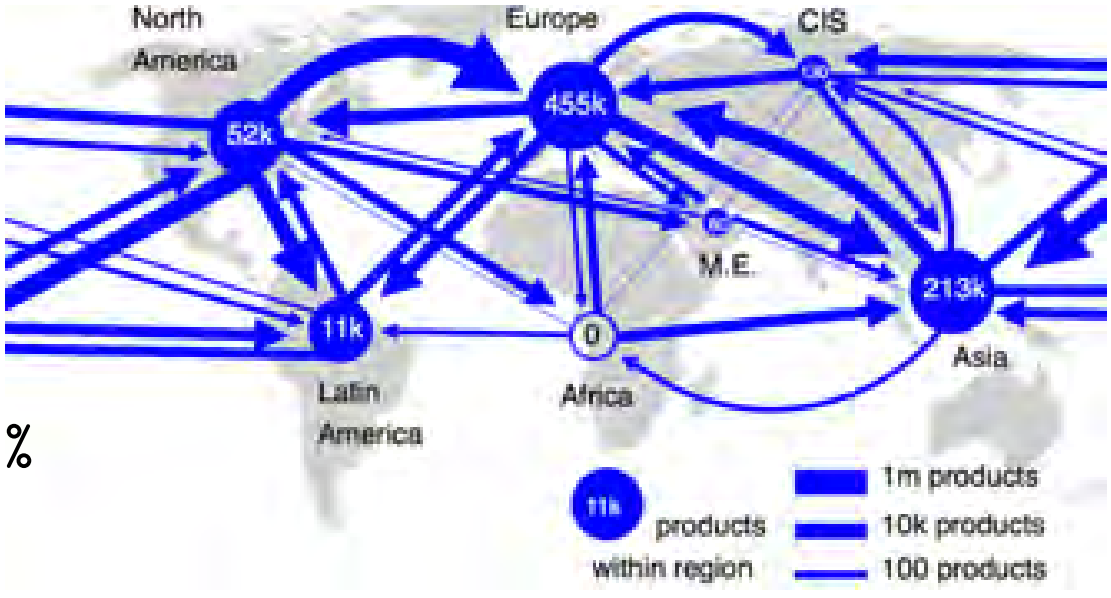
id	name	URL
1234	Puma PowerTech Blaze	u1
1234	Please choose	u1
1234	Puma PowerTech Blaze	u2
1234	Puma Shoe	u2
5678	Please choose	u3
5678	Sony Cybershot TS100	u3
...	...	...

IBEX: analyses

Found

- 13m email addresses with their name
- 235K chemical products
- 1.4m books
- 1.1m products

... with an accuracy of 73%-96%



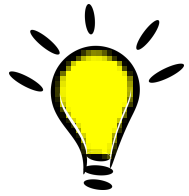
Analyzed

- Global trade flow
- frequent email providers
- frequent people names and more

[WebDB 2015]

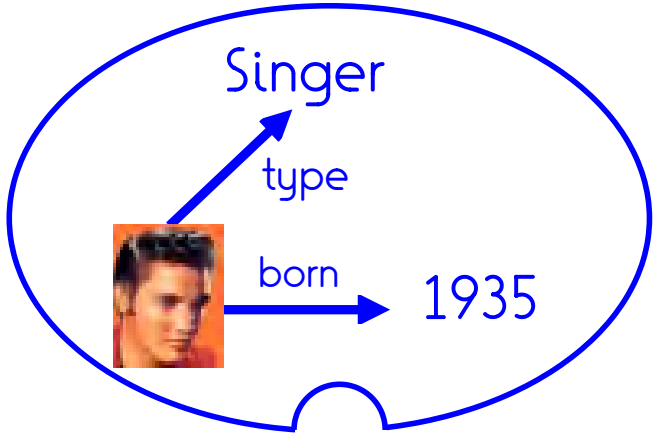
All data available online at
http://resources.mpi-inf.mpg.de/d5/ibex/

Knowledge Base Life Cycle



Using
knowledge bases ->

Constructing ✓
knowledge bases ->



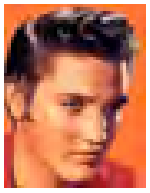
?

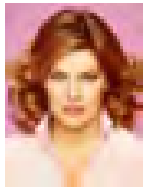
Querying
knowledge bases
->

Mining
incompleteness
->

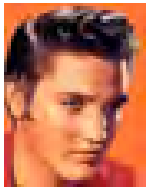
->end

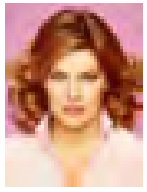
Are all people married?

 √

 √

Are all people married?

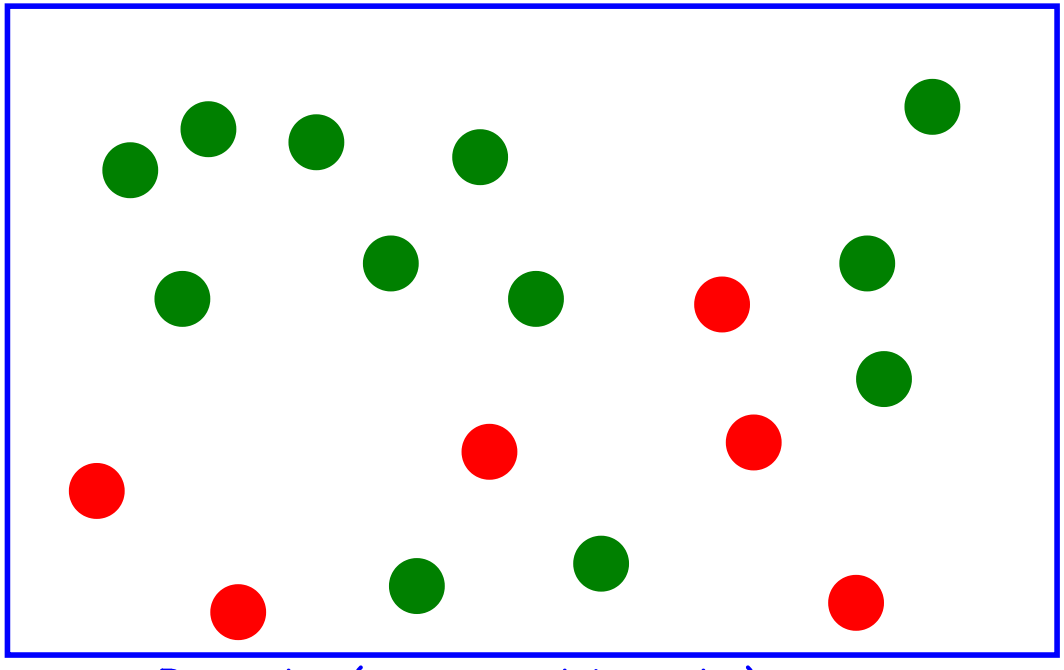
 √

 √

 ?

>details

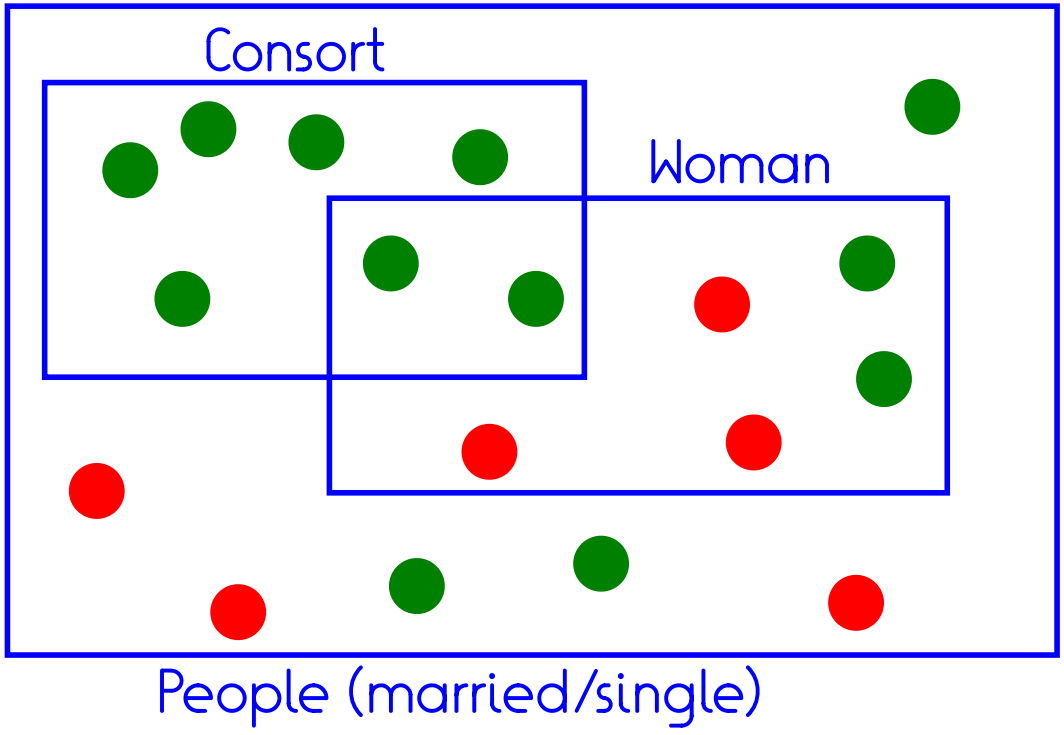
Are all people married?



People (married/single)

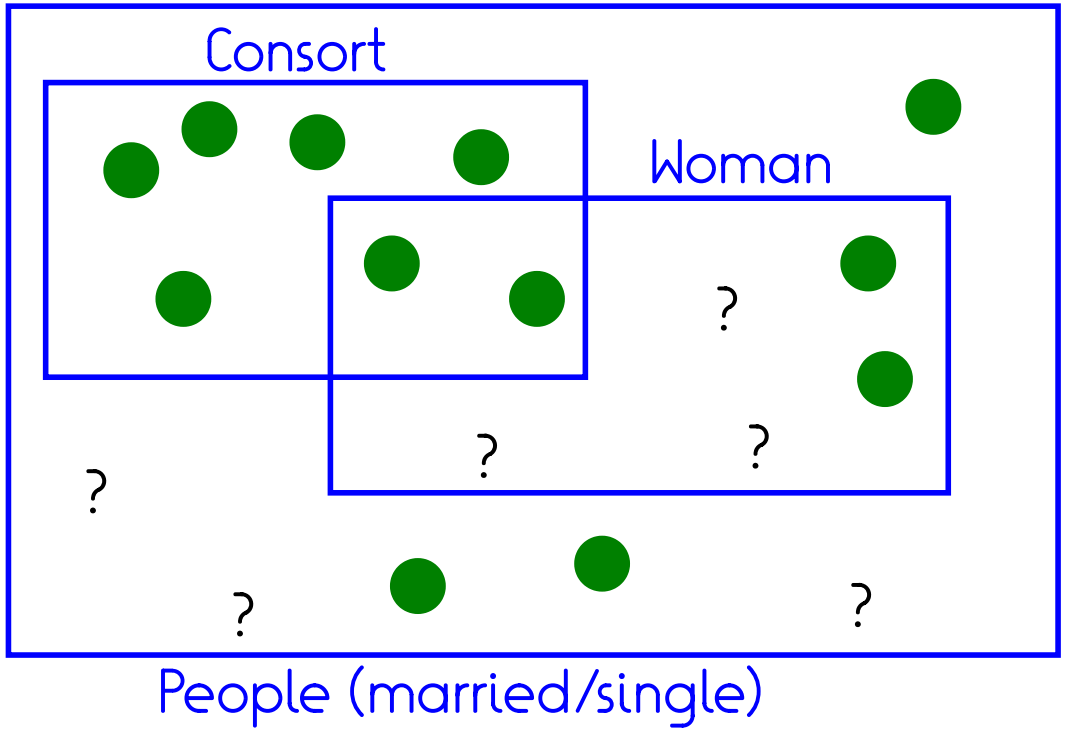
Real World

Are all people married?



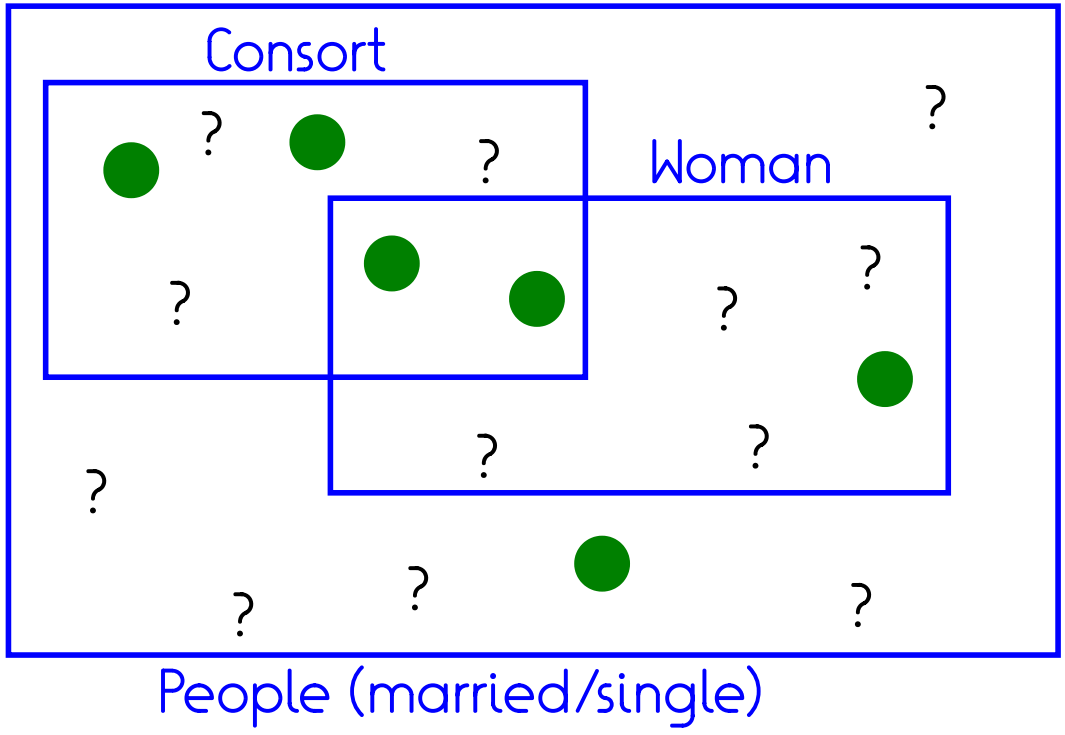
Real World

Are all people married?



Knowledge base without negative information

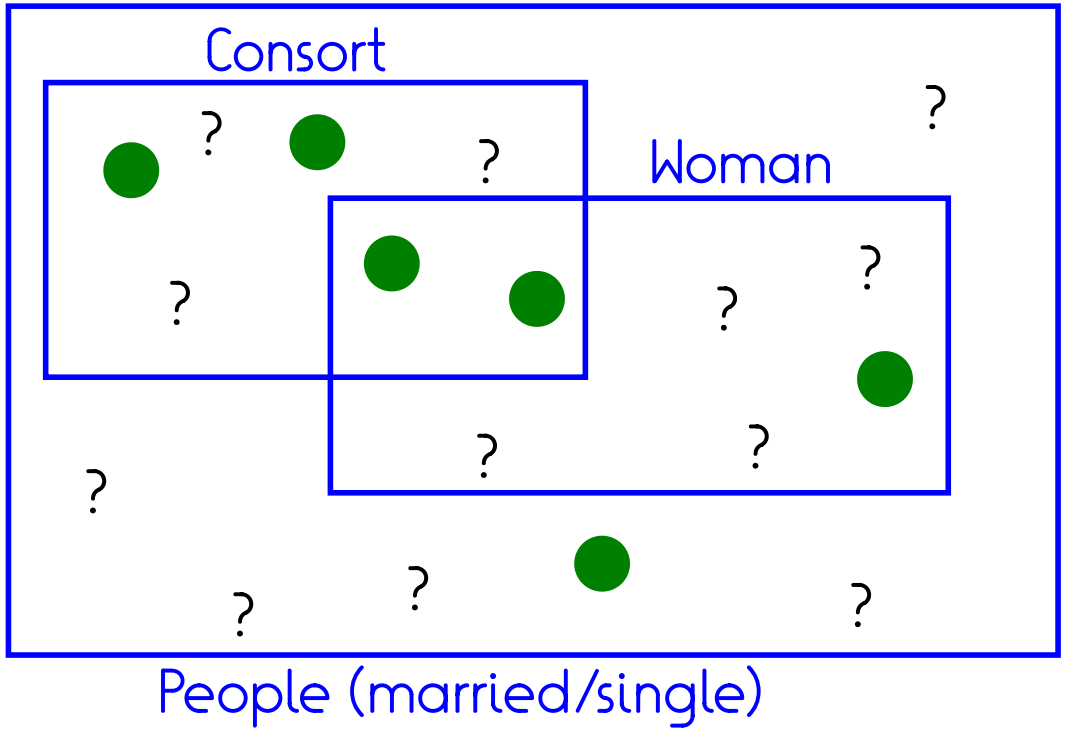
Are all people married?



Knowledge base without negative information and with incompleteness

>details
>baselines

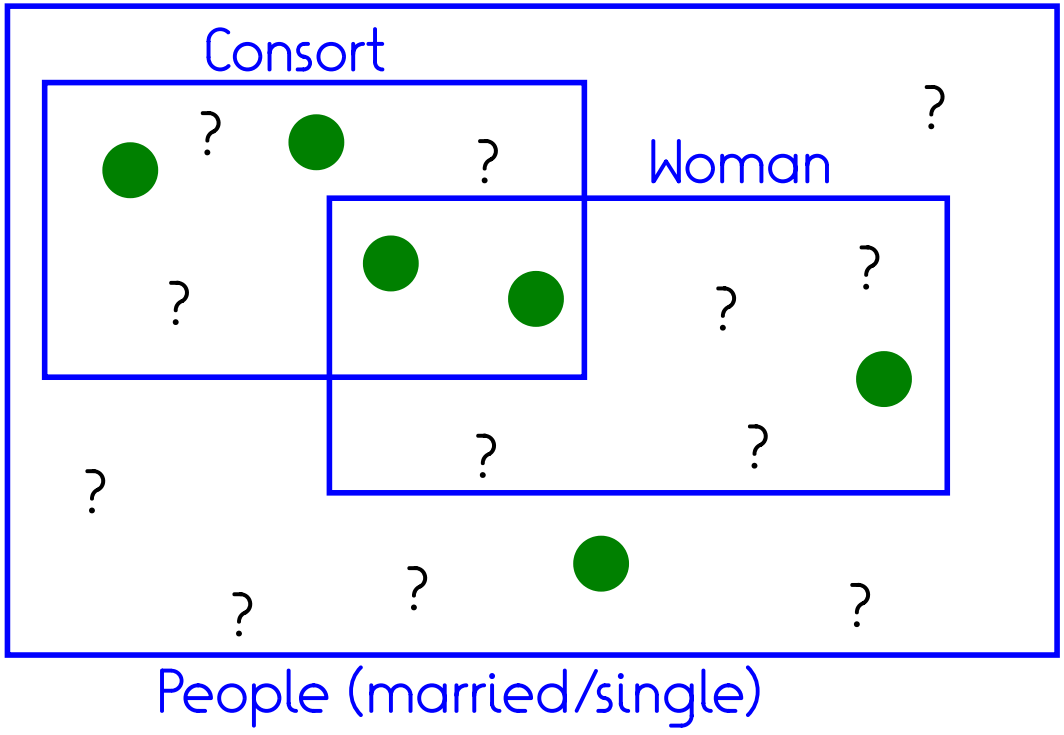
Are all people married?



Knowledge base without negative information and with incompleteness

Baseline 1: Obligatory if all instances have it

Are all people married?

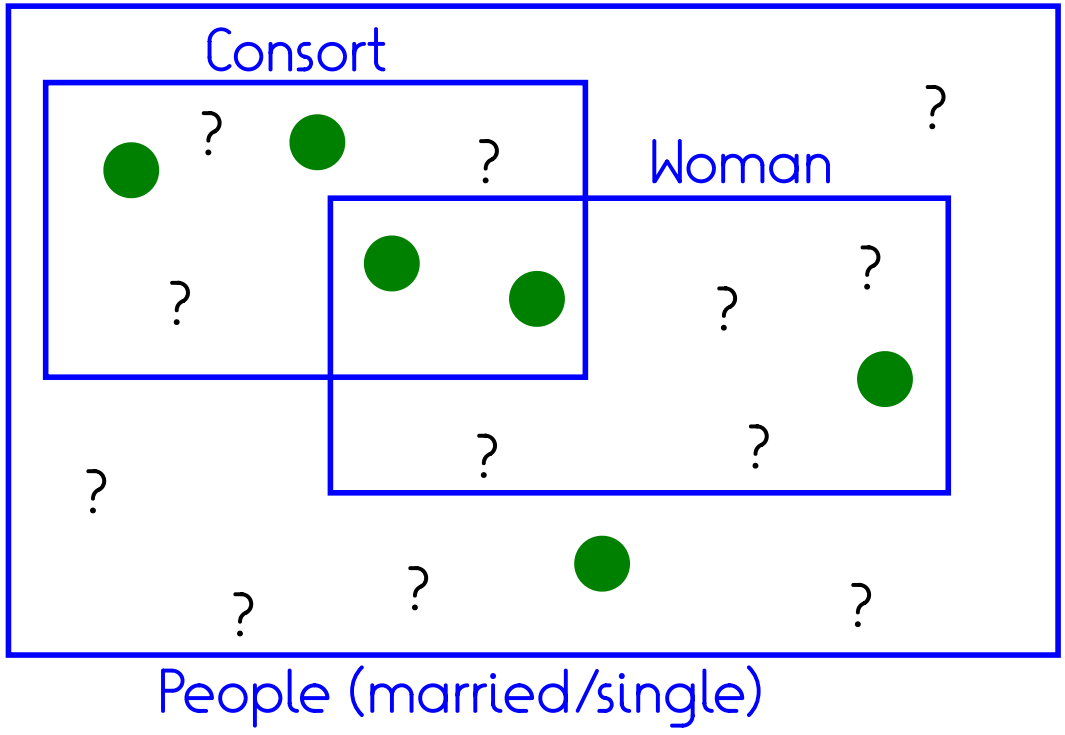


Knowledge base without negative information and with incompleteness

Baseline 1: Obligatory if all instances have it ✗

Baseline 2: Obligatory if at least n% of instances have it

Are all people married?



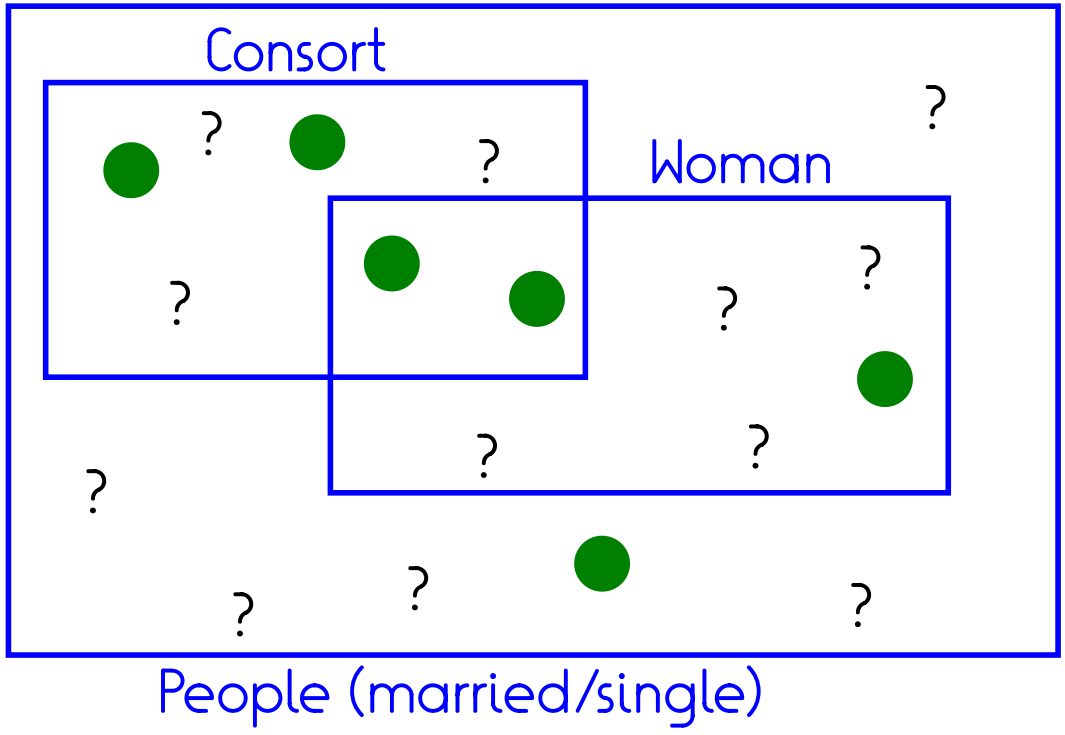
Knowledge base without negative information and with incompleteness

Baseline 1: Obligatory if all instances have it ✗

Baseline 2: Obligatory if at least n% of instances have it =>Woman ✗

Baseline 3: Obligatory if all instances that have it fall in the class

Are all people married?



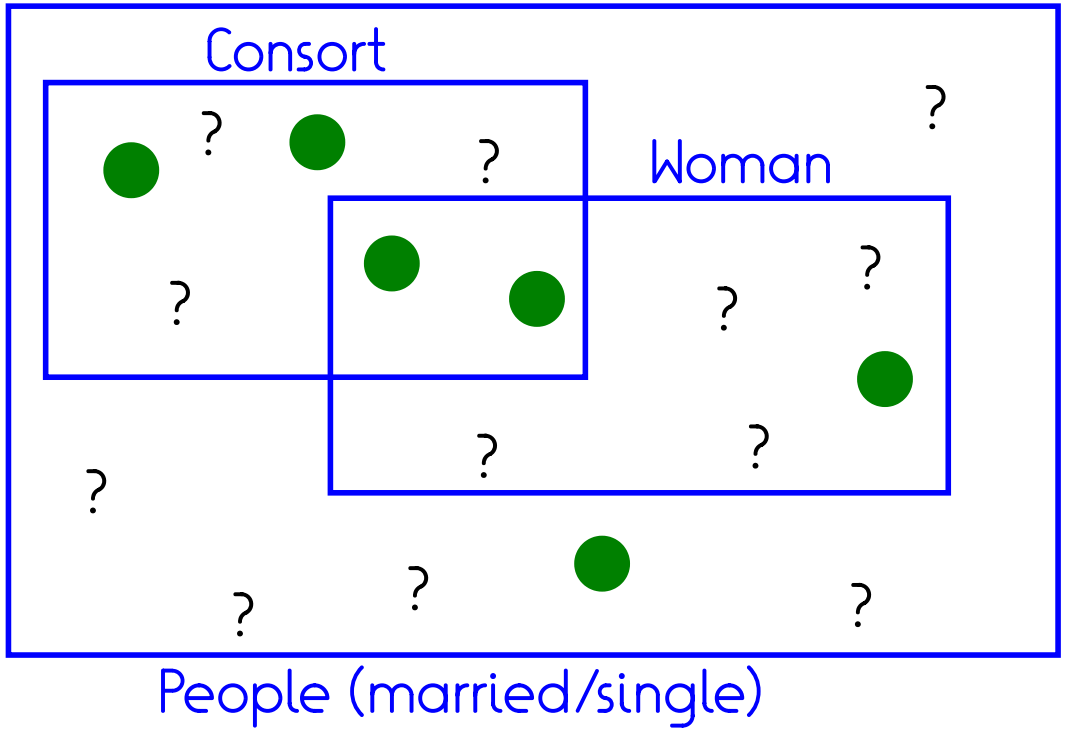
Knowledge base without negative information and with incompleteness

Baseline 1: Obligatory if all instances have it ✗

Baseline 2: Obligatory if at least n% of instances have it =>Woman ✗

Baseline 3: Obligatory if all instances that have it fall in the class =>Person ✗

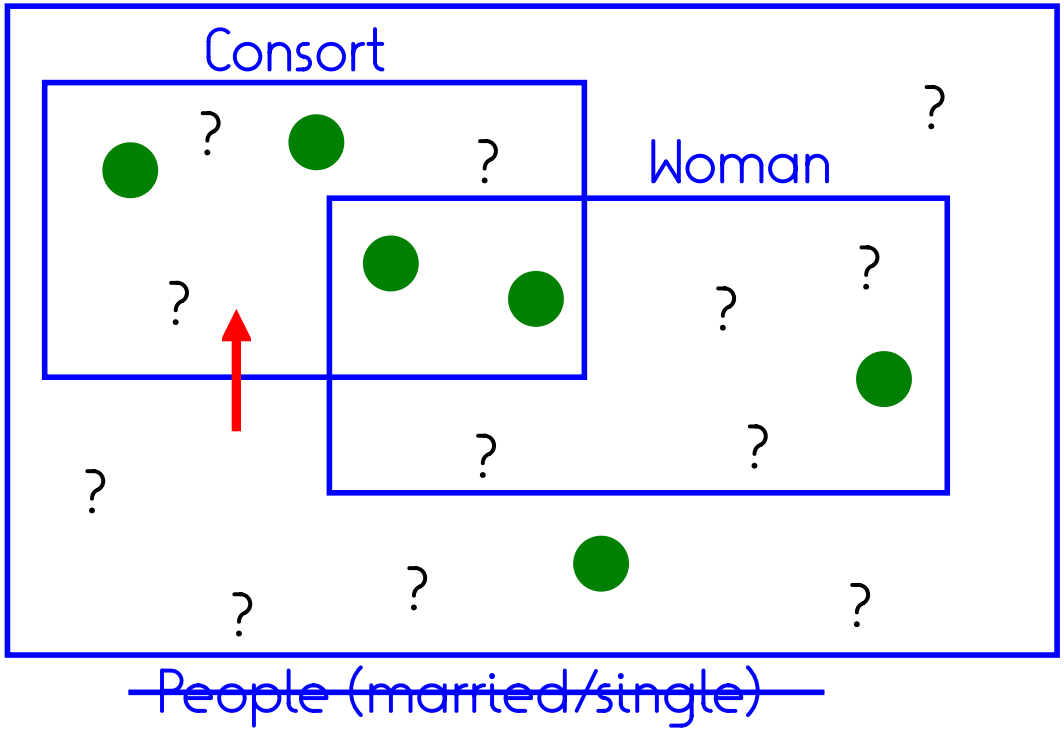
Are all people married?



Knowledge base without negative information and with incompleteness

Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

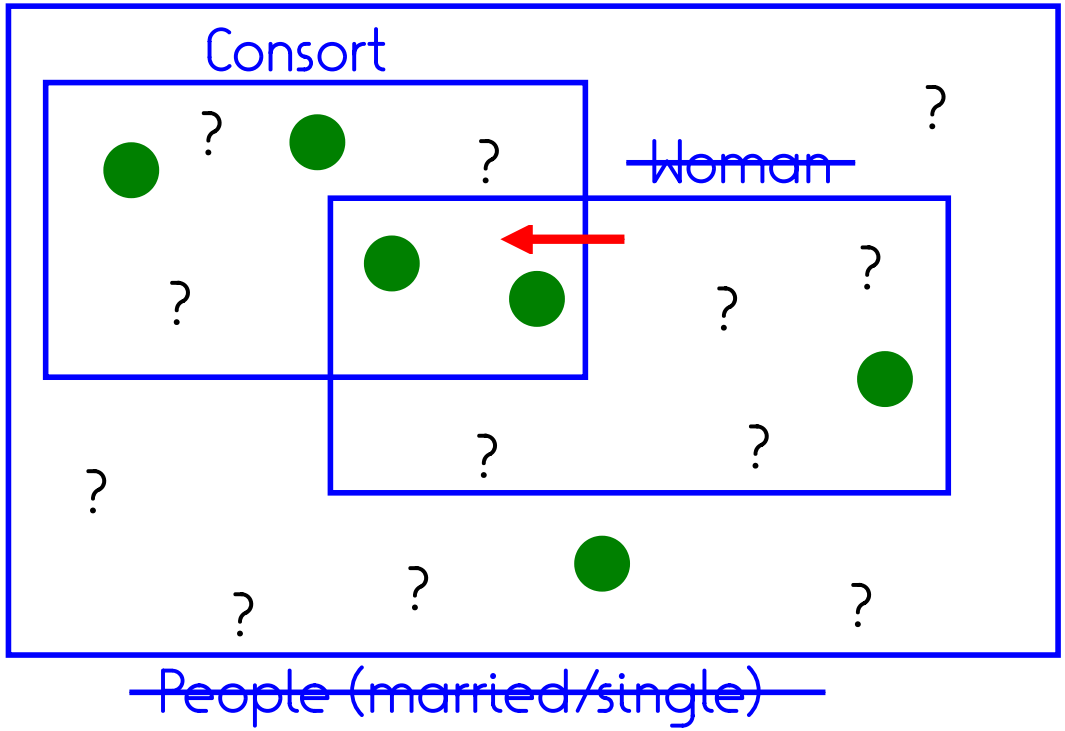
Are all people married?



Knowledge base without negative information and with incompleteness

Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

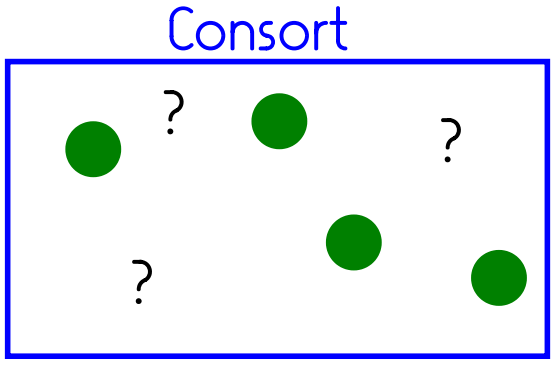
Are all people married?



Knowledge base without negative information and with incompleteness

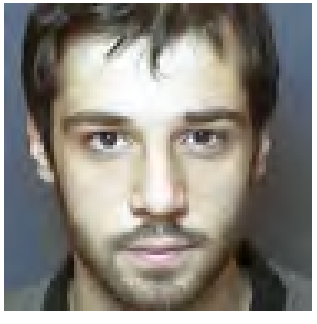
Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

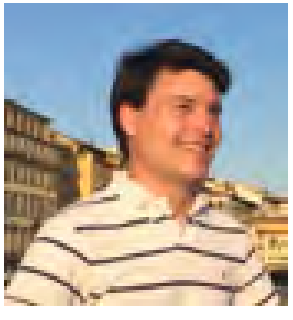
Determining obligatory attributes



We can predict obligatory attributes of classes with up to 80% precision (at 40% recall).

Caveat:
We do not actually predict, but exclude.

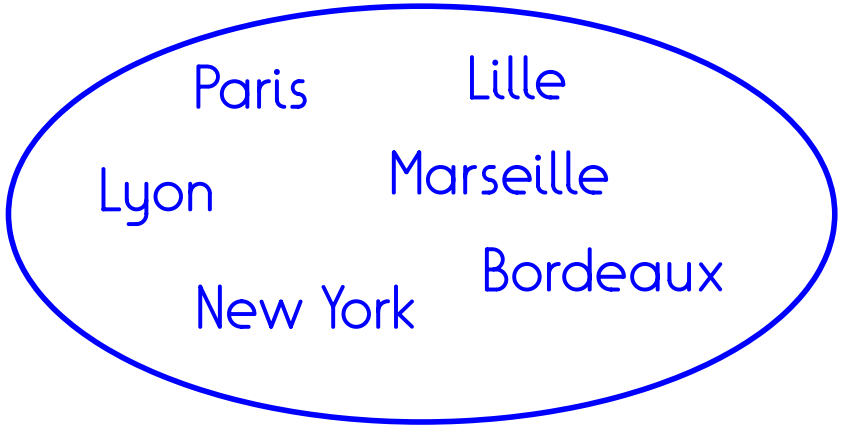




[WWW 2018]

Incompleteness: Missing entities

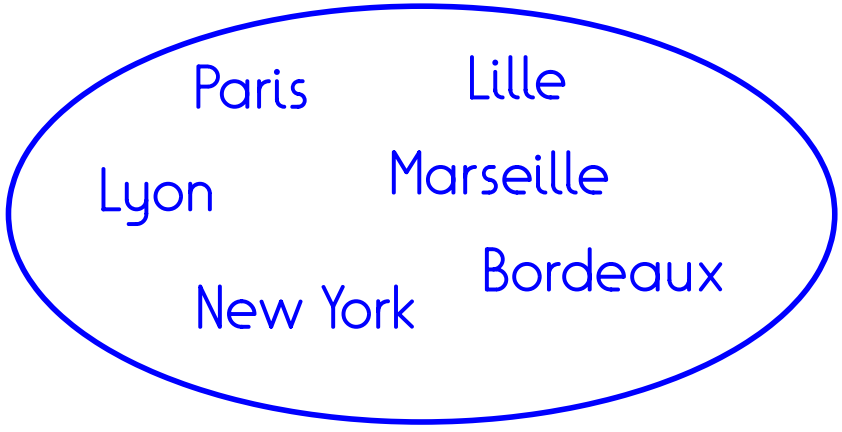
We have the following cities in our knowledge base:



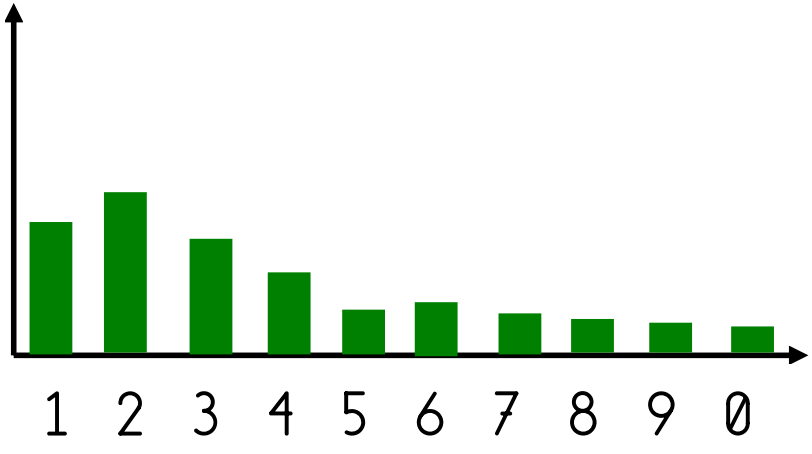
Are there any cities missing?

Incompleteness: Missing entities

We have the following cities in our knowledge base:



Are there any cities missing?



1) Take the number of inhabitants of each city
2) Take the first digit
3) Plot the number of cities per first digit

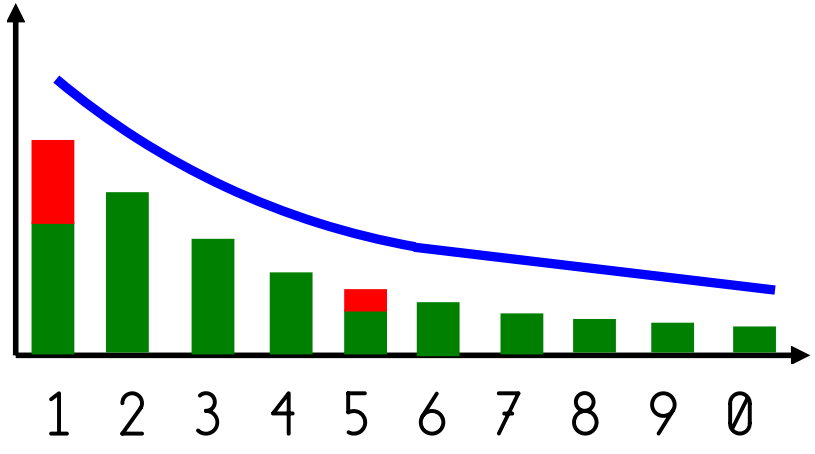
Incompleteness: Missing entities

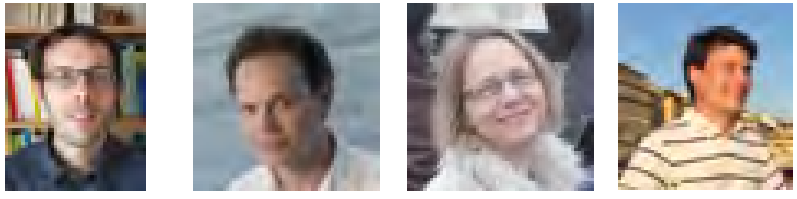
Benford's law says that the first digit d appears with probability

$$log_{10}(1 + \frac{1}{d})$$

=> We can give a minimum numbers of cities that are missing to make the distribution representative of the real world.

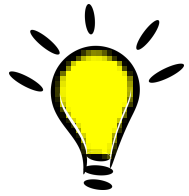
(For other classes, we can learn a parameter for a variant of the law.)





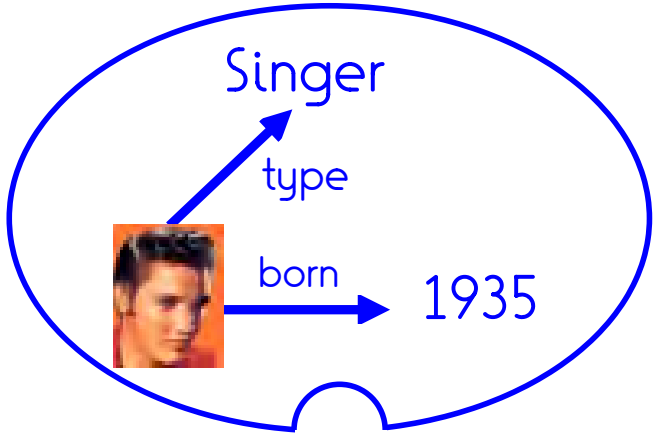
[ISWC 2018]

Knowledge Base Life Cycle



Using
knowledge bases ->

Constructing ✓
knowledge bases ->



?

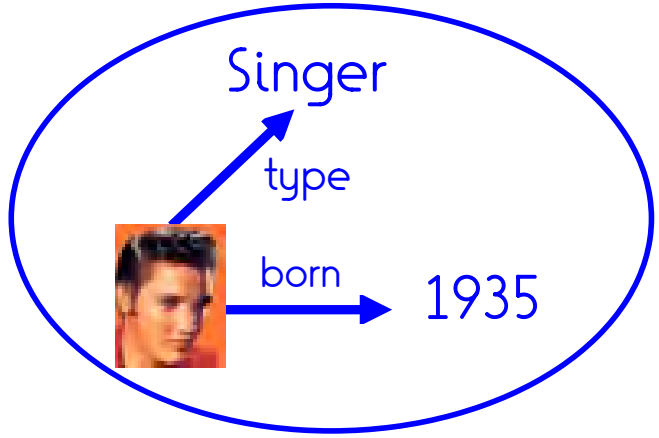
Querying
knowledge bases
->

Mining ✓
incompleteness
->

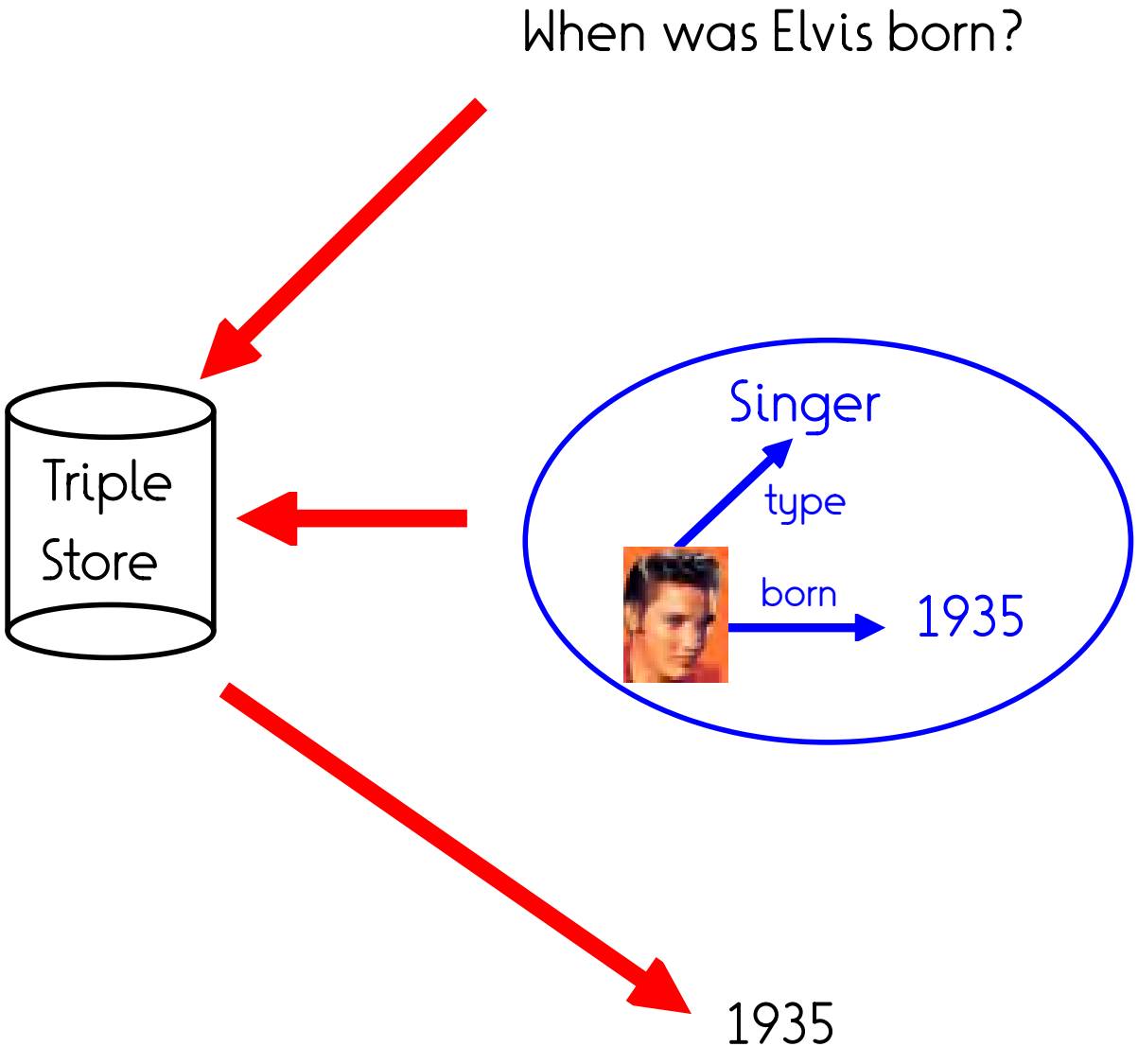
->end

Querying Knowledge Bases

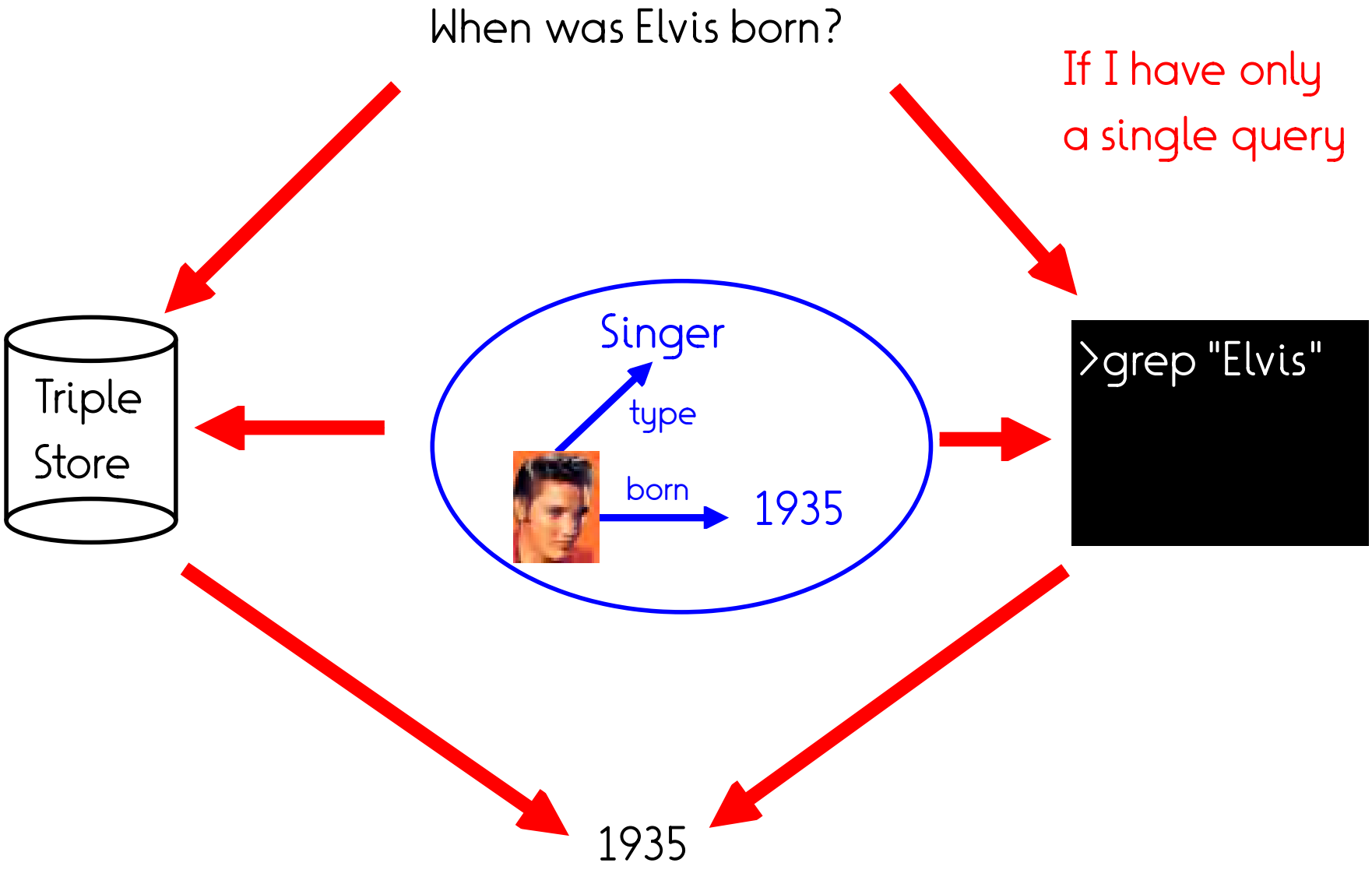
When was Elvis born?



Querying Knowledge Bases



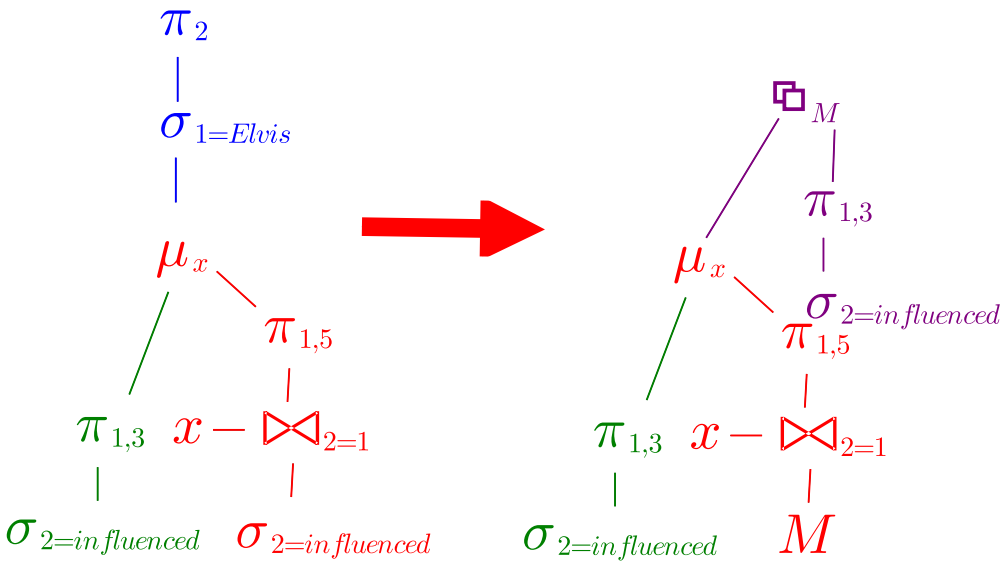
Querying Knowledge Bases



Querying Knowledge Bases

Who were the people that Elvis influenced (transitively)?

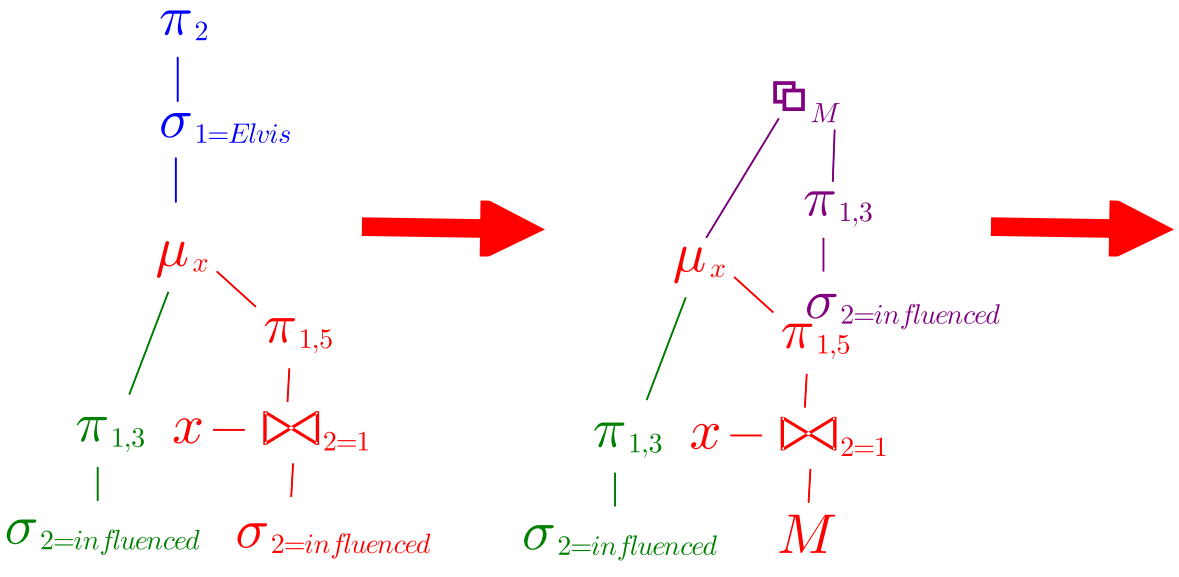
```
SELECT ?a {
 <Elvis> <influenced>+ ?a
}
```



Querying Knowledge Bases

Who were the people that Elvis influenced (transitively)?

```
SELECT ?a {
 <Elvis> <influenced>+ ?a
}
```




```
while sort -t $'\t' -k 1 -k
 <(cat tmp/lock_mat1 1
 join -t $'\t' -1 2 -2 1 -o 1
 <(sort -t $'\t' -k 1 tmp/d
| comm -23 - tmp/full2 >
 mv tmp/new2 tmp/delt
 sort -u —merge -o tmp/
 [ -s tmp/delta2 ];
 do continue; done
```

Querying Knowledge Bases

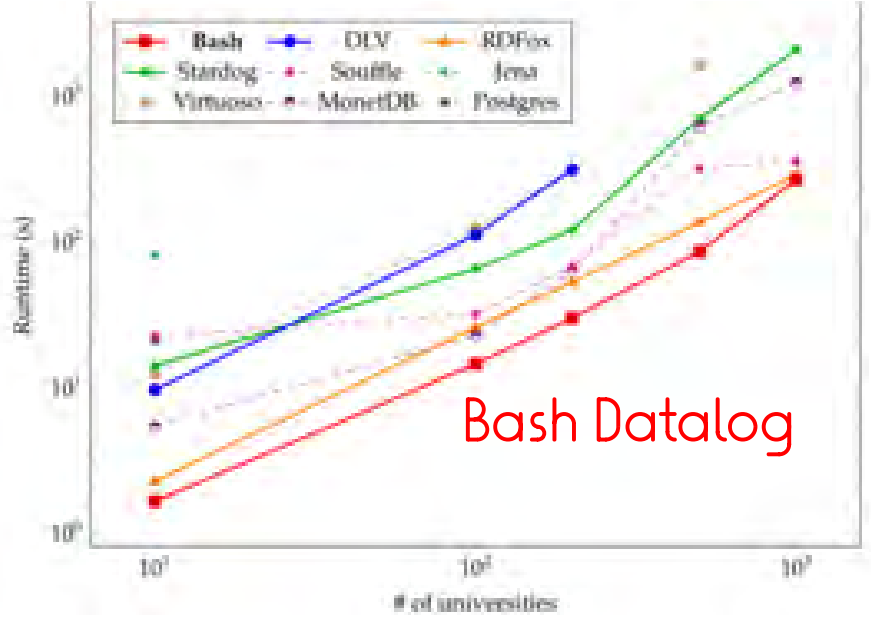
VLDB reviewer: Unix Shell commands to run queries is not very innovative.

The DB community exists to overcome the shortcomings of this approach.

Querying Knowledge Bases

VLDB reviewer: Unix Shell commands to run queries is not very innovative.

The DB community exists to overcome the shortcomings of this approach.



including
the loading time

	Bash	RDFox	BigDatalog	Stardog
LiveJournal (69M edges)	2min	1min	9min	15min
Orkut (117M edges)	2.5min	2min	30min	18min
Friendster (2586M edges)	4:30h	OOM	OOS	>10h
WikiData (2100M edges)	1h	OOM		

Convert SPARQL to Bash online

Bash Datalog

Answering Datalog Queries with Unix Shell Commands

About | Datalog mode | SPARQL/OWL mode | API

SPARQL query

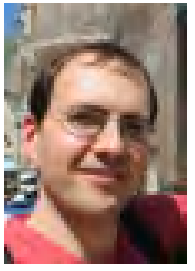
```
BASE <http://yago-knowledge.org/resource/>
SELECT ?a WHERE {
  <Elvis_Presley> <influenced>+ ?a.
}
```

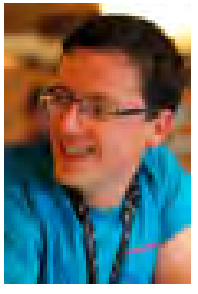
Convert to bash script | Download script

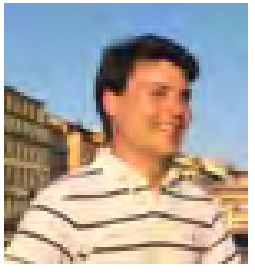
Bash script

```
read_ntriples() { $awk -F" " '{ sub(" ", "\t"); sub(" ",
"\t"); sub(/ \.$/, ""); print $0 }' "$@"; }
conv_ntriples() { $awk -F$'\t' '{ print $1 " " $2 " " $3 "
." }'; }


touch tmp/mat0 tmp/mat1
$awk -v FS=$'\t' ' ($1 == "<http://yago-knowledge.org
/resource/Elvis_Presley>" && $2 == "<http://yago-
knowledge.org/resource/influenced>") { print $3 >>
"tmp/mat0" }
```



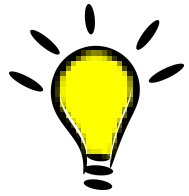




[ISWC 2018]

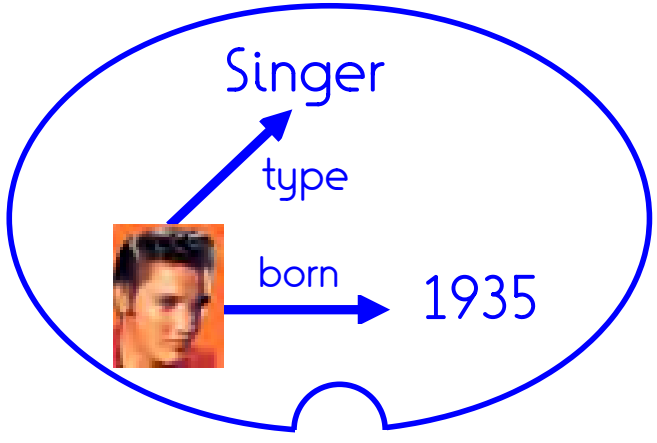
https://thomasrebele.org/projects/bashlog/

Knowledge Base Life Cycle



Using
knowledge bases ->

Constructing ✓
knowledge bases ->



?

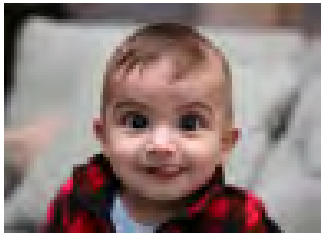
Querying ✓
knowledge bases
->

Mining ✓
incompleteness
->

->end

Combinatorial Creativity

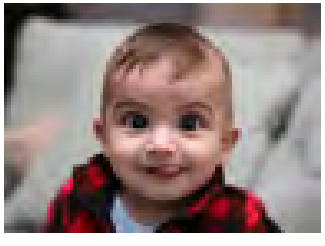
+ ( − ) +



Description Logics do not work

$$Mop \equiv Tool \sqcap \exists has.Stick \sqcap \exists has.Strings$$

$$BabyMop \equiv$$
$$Romper \sqcap \exists has.(Mop \sqcap \neg\exists has.Stick) \sqcap \exists has.Baby$$
$$\equiv \ldots \sqcap \exists has.\bot \sqcap \ldots$$



Language for Combinatorial Creativity

$$Mop \equiv Tool \sqcap \exists\, has.Stick \sqcap \exists\, has.Strings$$

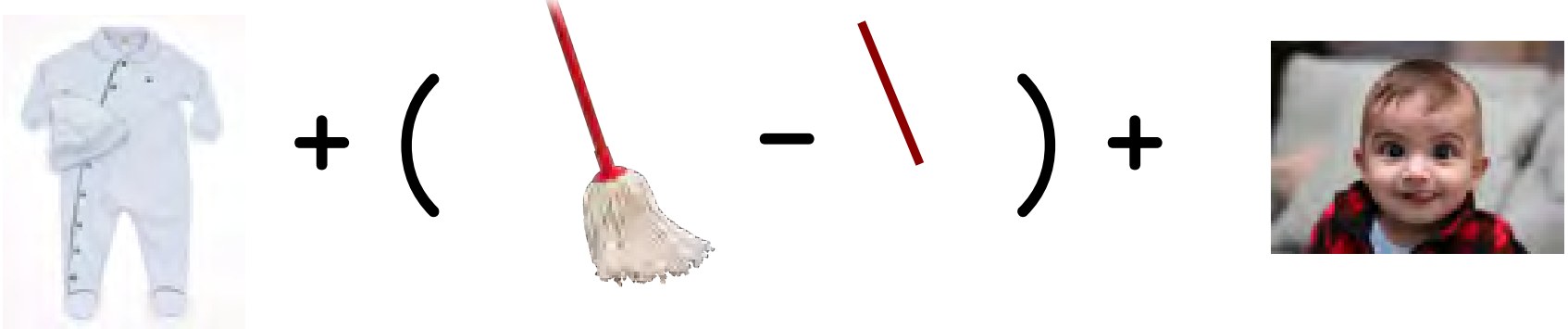
Subtraction: $Mop - \exists\, has.\top \equiv Tool \sqcap \exists\, has.Strings$

Addition: $Mop + \exists\, has.\top \equiv Mop$

Succession: $Mop \rightarrow \exists\, u.\top \equiv Stick$

Selection*: $Mop \uparrow \exists\, has.\top \equiv \exists\, has.Stick$

$$Romper + \exists\, has.(Mop - \exists\, has.Stick) + \exists\, has.Baby \equiv BabyMop$$



Language for Combinatorial Creativity

$$Mop \equiv Tool \sqcap \exists\, has.Stick \sqcap \exists\, has.Strings$$

Subtraction: $Mop \;-\; \exists\, has.\top \equiv Tool \sqcap \exists\, Strings$

Addition: $Mop \;+\; \exists\, has.\top \equiv Mop$

Succession: $Mop \;\rightarrow\; \exists\, r.\top \equiv Stick$

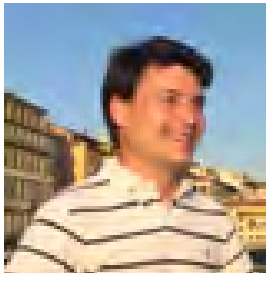
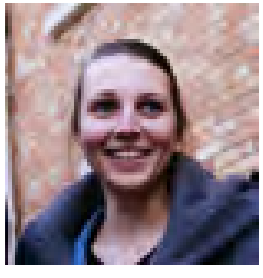
Selection*: $Mop \uparrow \exists\, has.\top \equiv \exists\, has.Stick$

1) Descriptive experiments

2) Generative experiments

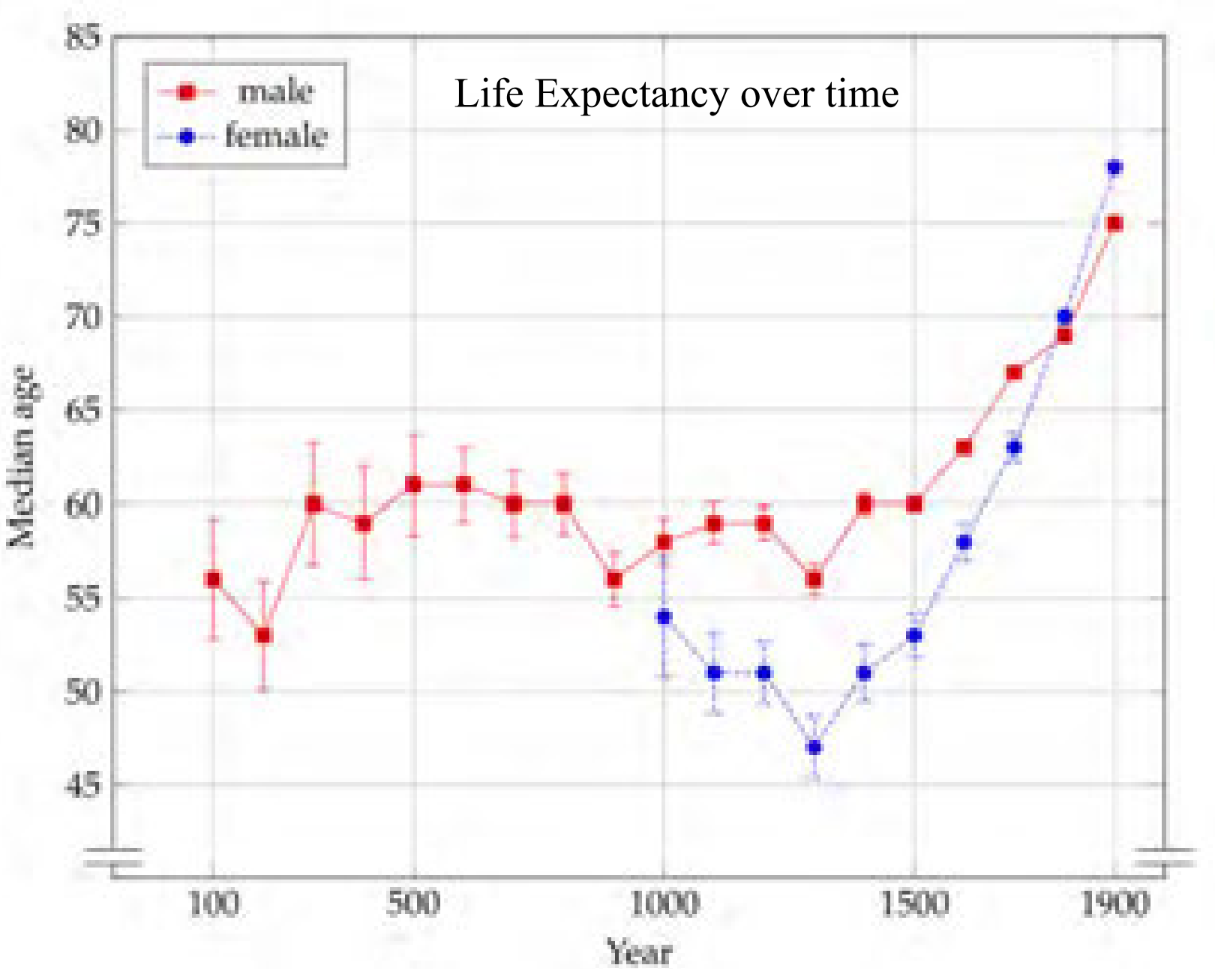
1/3 nonsense, 1/3 exists,

1/3 "imaginable"

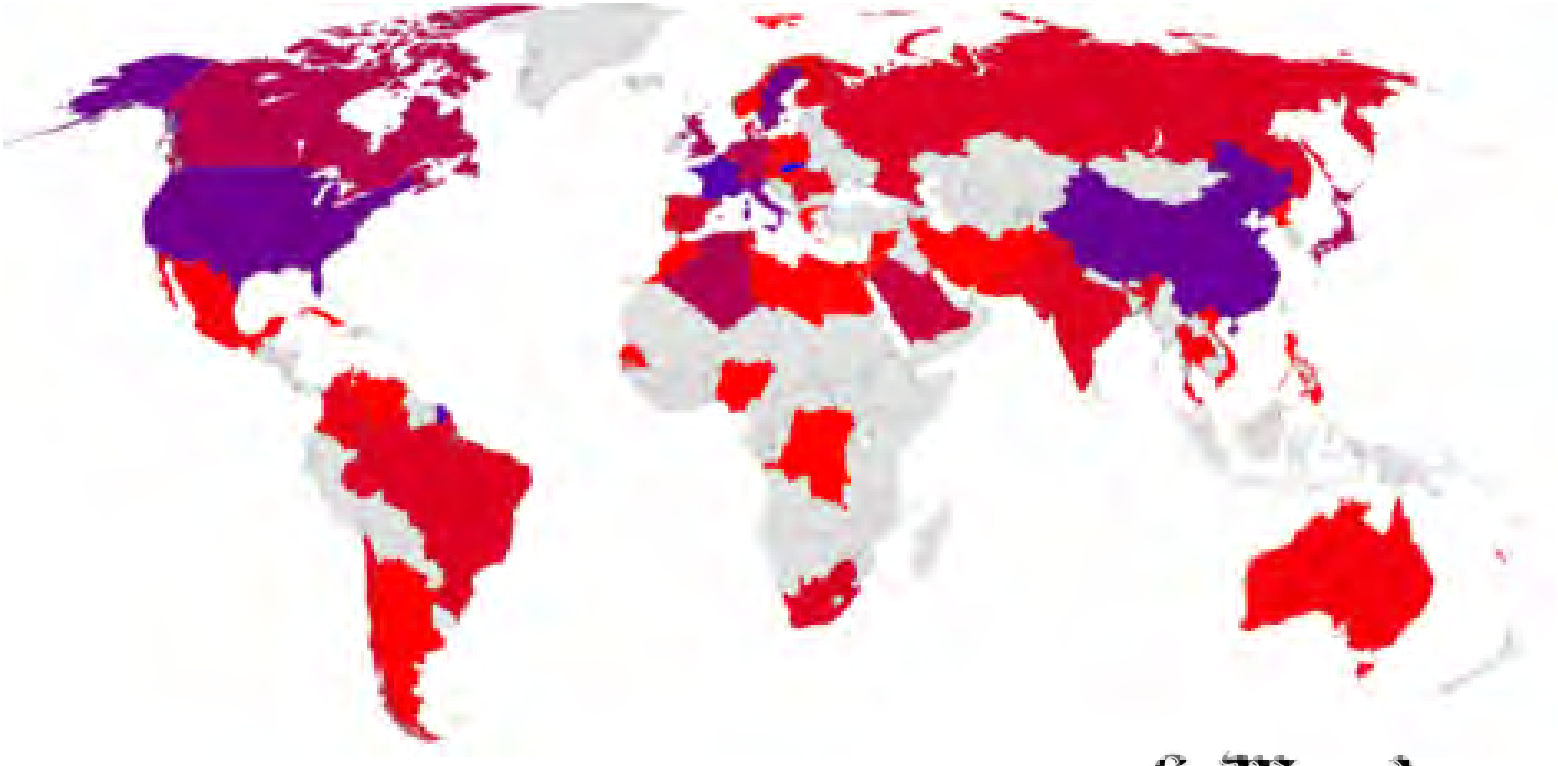



[ISWC 2016
paper & demo]

YAGO for the Digital Humanities



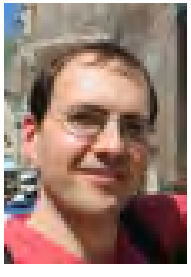
Presence of foreign companies

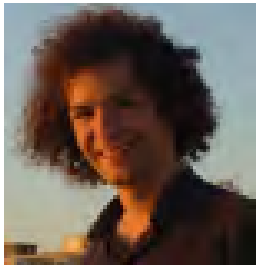
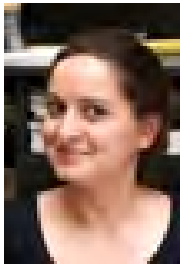
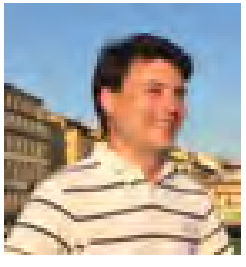


red: many foreign companies mentioned

blue: few foreign companies mentioned

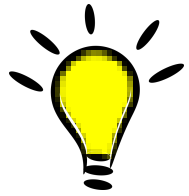
Le Monde + YAGO



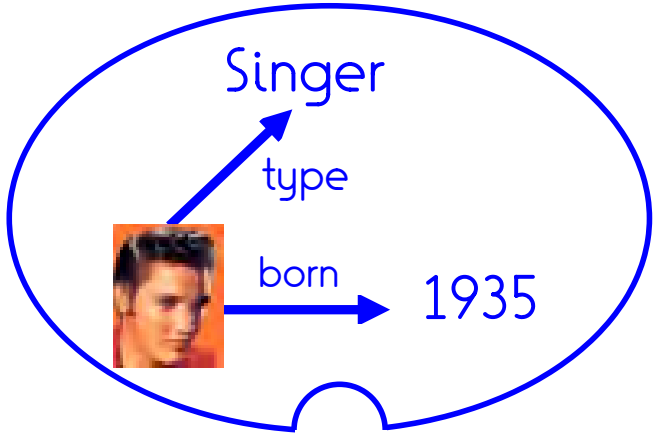
[AKBC 2013, VLDB 2014 vision, LDOW 2018]

Knowledge Base Life Cycle



Using knowledge bases -> ✓

Constructing ✓ knowledge bases ->

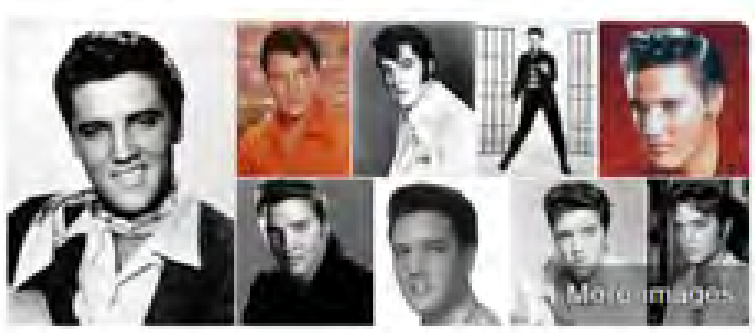


?

Querying ✓ knowledge bases ->

Mining ✓ incompleteness ->

Is Elvis dead?



Elvis Presley

American singer

Available on

YouTube

Spotify

Deezer

More music services

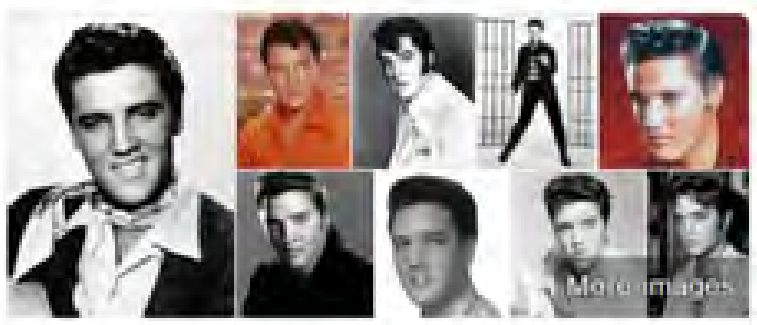
Elvis Aaron Presley, also known mononymously as Elvis, was an American singer, musician, and actor. Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as the "King of Rock and Roll" or simply "the King". Wikipedia

Born: January 8, 1935, Tupelo, Mississippi, United States

Died: August 16, 1977, Graceland, Memphis, Tennessee, United States

???

Is Elvis dead?



Elvis Presley

American singer

Available on

YouTube

Spotify

Deezer

More music services

Elvis Aaron Presley, also known mononymously as Elvis, was an American singer, musician, and actor. Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as the "King of Rock and Roll" or simply "the King". Wikipedia

Born: January 8, 1935, Tupelo, Mississippi, United States

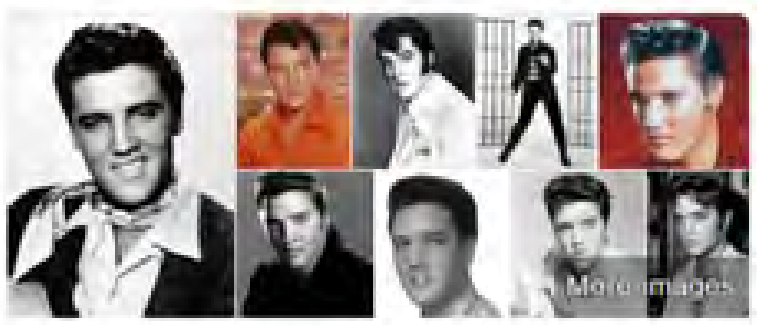
Died: August 16, 1977, Graceland, Memphis, Tennessee, United States

???

yago
select knowledge

Elvis —diedIn→ 1977

Is Elvis dead?



Elvis Presley

American singer

Available on

YouTube

Spotify

Deezer

More music services

Elvis Aaron Presley, also known mononymously as Elvis, was an American singer, musician, and actor. Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as the "King of Rock and Roll" or simply "the King". Wikipedia

Born: January 8, 1935, Tupelo, Mississippi, United States

Died: August 16, 1977, Graceland, Memphis, Tennessee, United States

???

yago
select knowledge

Elvis —diedIn→ 1977

100m statements

95% accuracy

-> 5m wrong statements

Knowledge Base Life Cycle