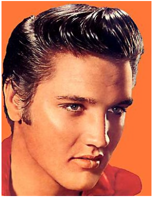
Making knowledge bases more complete

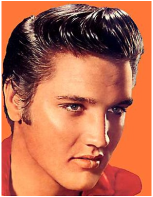
Wikimedia Showcase 2019-12-18

Fabian M. Suchanek
Télécom Paris University, France

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

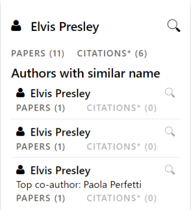
What is the capital of New Zealand? "Wellington"

Amazon Echo

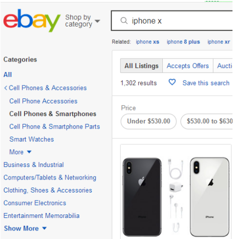
Discovered 6 kineasis proteins that relate to cancer

IBM Watson

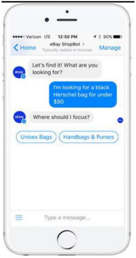
New Technology: Knowledge Bases



Microsoft Academic

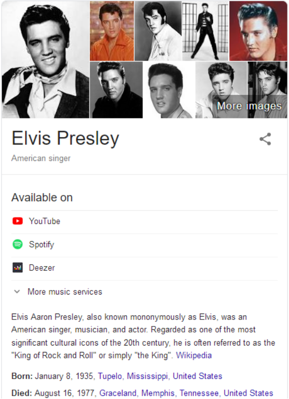


Ebay Knowledge Graph



Ebay Shopbot on Facebook Messenger

New Technology: Knowledge Bases



Google Knowledge Graph

New Technology: Knowledge Bases

Elvis Presley

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

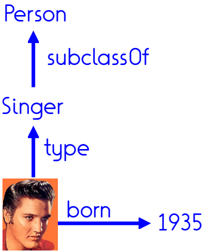
Born: January 8, 1935, Tupelo, Mississippi, United States

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

???

Google Knowledge Graph

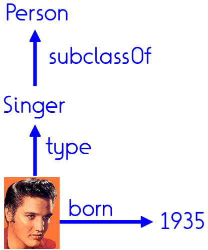
Knowledge Bases



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.)

This talk



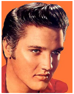
1. Constructing Knowledge Bases
2. Completing Knowledge Bases

Extracting from Wikipedia

Elvis Presley

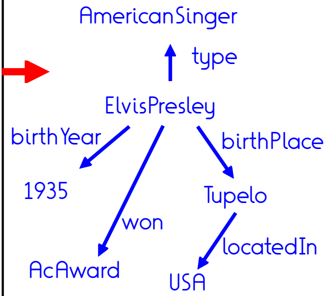
WIKIPEDIA
The Free Encyclopedia

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

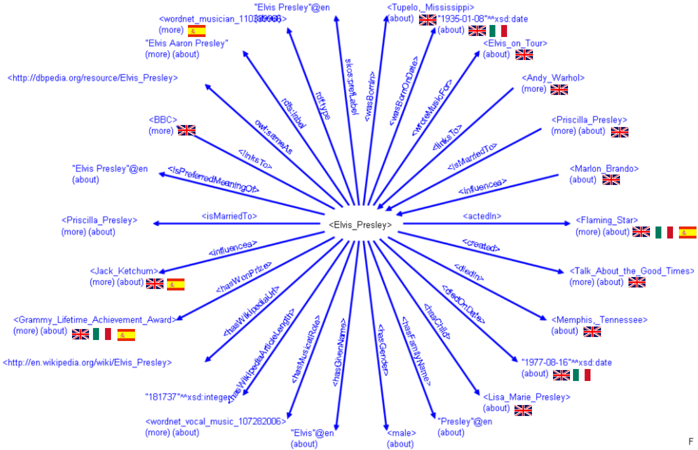


Born: 1935
In: Tupelo
...

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



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

New version in preparation!

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

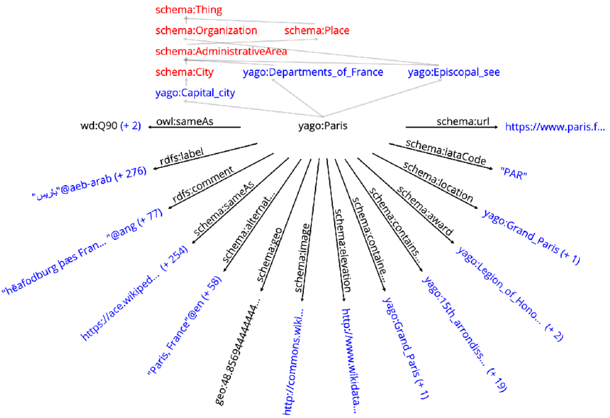




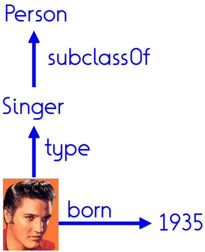


YAGO 4: A "reason-able" KB

YAGO 4 combines schema.org + Wikidata + constraints

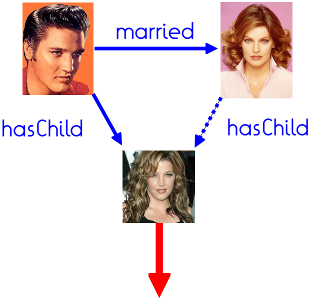


This talk



1. Constructing Knowledge Bases ✓
2. Completing Knowledge Bases

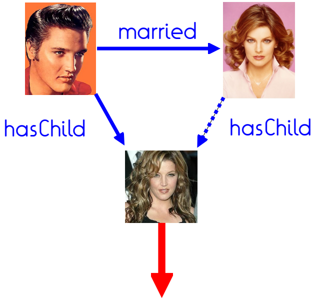
Incompleteness: Concrete facts



$$married(x,y) \wedge hasChild(x,z) \Rightarrow hasChild(y,z)$$

>PCA

Incompleteness: Concrete facts



$$married(x,y) \wedge hasChild(x,z) \Rightarrow hasChild(y,z)$$

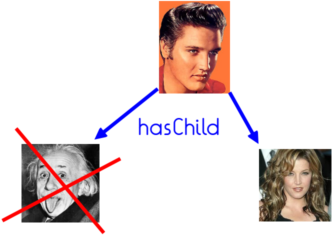
But: Rule mining needs counter examples
and RDF ontologies are positive only

>PCA

Partial Completeness Assumption

Assumption:
If we know $r(x,y_1),\ldots,\ r(x,y_n)$,
then all other $r(x,z)$ are false.

Partial Completeness Assumption

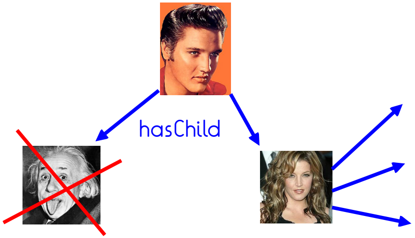


Assumption:

If we know $r(x,y_1),\ldots,\ r(x,y_n)$,

then all other $r(x,z)$ are false.

Partial Completeness Assumption



Assumption:
If we know $r(x,y_1),\ldots,\ r(x,y_n)$,
then all other $r(x,z)$ are false.

AMIE finds rules in knowledge bases

yago
select knowledge

AMIE

(2min)

$$r(x, y) \wedge r'(z, y) \Rightarrow r''(x, z)$$

AMIE is based on an efficient in-memory database implementation.

Caveat: rules cannot predict the unknown with high precision

AMIE finds rules in knowledge bases

AMIE

$$type(x,pope) \Rightarrow diedIn(x,Rome)$$

[WWW 2013, VLDB journal 2015]

New version
in preparation!

Incompleteness: Existence of facts

 marriedTo →  √

the quality of YAGO w
ls a precision of 95%, as
ıks to our brilliant algori

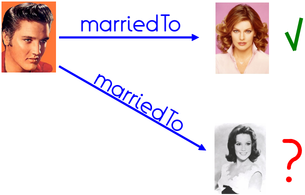
Incompleteness: Existence of facts

marriedTo

√

the quality of YAGO w
ls a precision of 95%, as
ıks to our brilliant algori

Incompleteness: Existence of facts

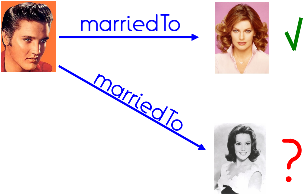


the quality of YAGO w
ls a precision of 95%, as
ıks to our brilliant algori

Given a subject s and a relation r, do we know all o with $r(s, o)$?

© Girl Happy

Signals for Incompleteness



Closed World Assumption
Partial Completeness Assumption
Popularity oracle

No-change oracle
Star-pattern oracle
Class-oracle

AMIE oracle: Learn rules such as

$$moreThan_1(x, hasParent) \Rightarrow complete(x, hasParent)$$

CrowdFlower

>details

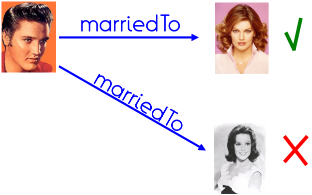
Signals for Incompleteness (F1)

Relation	CWA	PCA	$card_2$	Popularity	No change	Star	Class	AMIE
diedIn	60%	22%	—	4%	15%	50%	**99%**	96%
directed	40%	96%	19%	7%	71%	0%	0%	**100%**
graduatedFrom	89%	4%	2%	2%	10%	89%	**92%**	87%
hasChild	71%	1%	1%	2%	13%	40%	**78%**	**78%**
hasGender	78%	**100%**	—	2%	—	86%	95%	**100%**
hasParent*	1%	54%	**100%**	—	—	0%	0%	**100%**
isCitizenOf*	4%	98%	11%	1%	4%	10%	5%	**100%**
isConnectedTo	87%	34%	19%	—	—	68%	88%	**89%**
isMarriedTo*	55%	7%	0%	3%	12%	37%	**57%**	46%
wasBornIn	28%	**100%**	—	5%	8%	0%	0%	**100%**

Relation	CWA	PCA	$card_2$	Popularity	Star	Class	AMIE
alma_mater	**90%**	14%	5%	1%	87%	87%	87%
brother	93%	1%	—	1%	94%	**96%**	**96%**
child	70%	1%	—	1%	**79%**	72%	73%
country_of_citizenship*	42%	97%	10%	3%	0%	0%	**98%**
director	81%	**100%**	—	3%	94%	89%	**100%**
father*	5%	**100%**	6%	9%	89%	8%	**100%**
mother*	3%	**100%**	3%	10%	67%*	5%	**100%**
place_of_birth	53%	**100%**	7%	5%	55%	0%	**100%**
place_of_death	89%	35%	1%	2%	81%	81%	**96%**
sex_or_gender	81%	**100%**	6%	3%	92%	91%	**100%**
spouse*	**57%**	7%	—	1%	54%	54%	55%

• = biased training sample

Incompleteness: Existence of facts



AMIE can predict incompleteness

- bornIn: 100% F1-measure
- diedIn: 96%
- directed: 100%
- graduatedFrom: 87%
- hasChild: 78%
- isMarriedTo: 46%

... and more.




[WSDM 2017]

>rep&married

>married

Are all people married?

 √

 √

Are all people married?

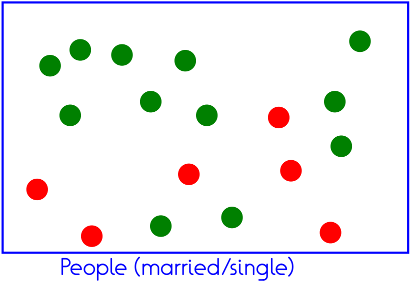
 √

 √

 ?

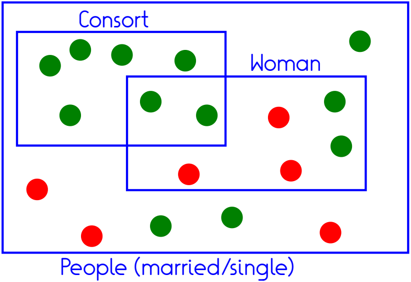
>details

Are all people married?



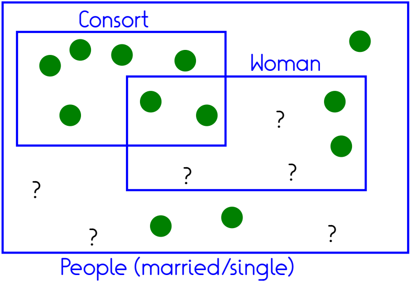
Real World

Are all people married?



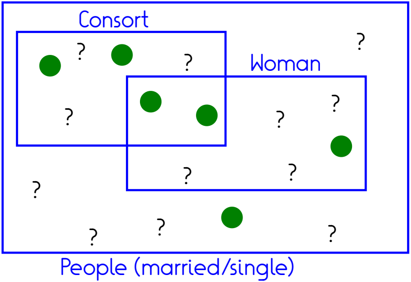
Real World

Are all people married?



Knowledge base
under the
Open World Assumption

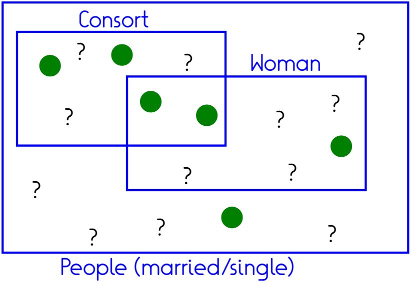
Are all people married?



Knowledge base
under the
Open World Assumption
and incompleteness

>details
>baselines

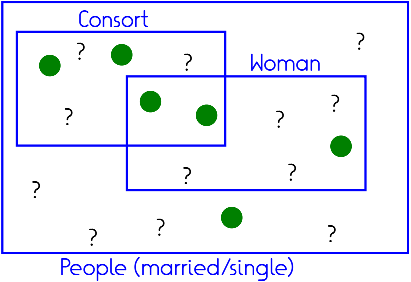
Are all people married?



Knowledge base
under the
Open World Assumption
and incompleteness

Baseline 1: Obligatory if all instances have it

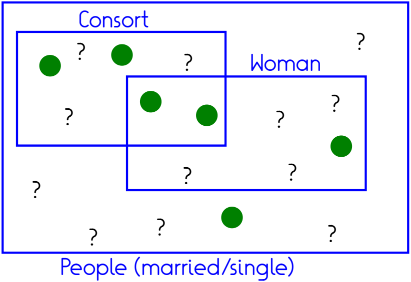
Are all people married?



Knowledge base
under the
Open World Assumption
and incompleteness

Baseline 1: Obligatory if all instances have it ✗
Baseline 2: Obligatory if at least n% of instances have it

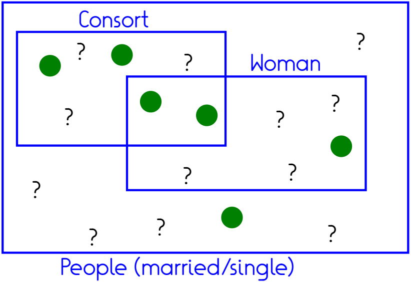
Are all people married?



Knowledge base
under the
Open World Assumption
and 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
under the
Open World Assumption
and incompleteness

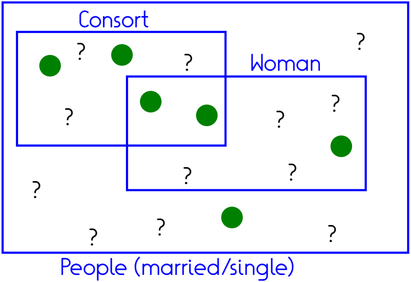
Baseline 1: Obligatory if all instances have it

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

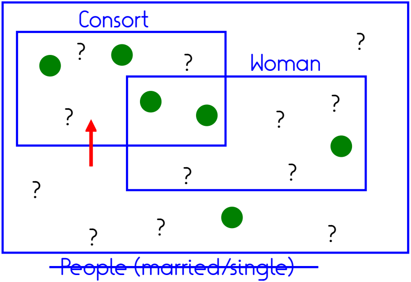
Are all people married?



Knowledge base under the Open World Assumption and 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.

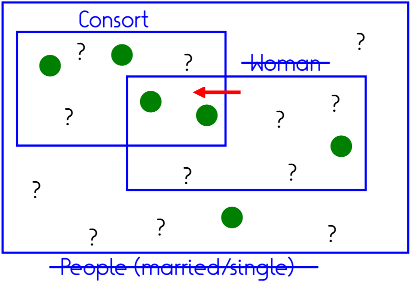
Are all people married?



Knowledge base
under the
Open World Assumption
and 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.

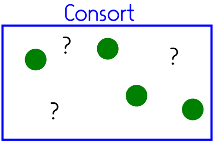
Are all people married?



Knowledge base
under the
Open World Assumption
and incompleteness

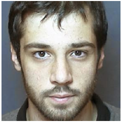
Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

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.

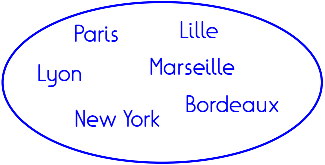


[WWW 2018]

>rep

Incompleteness: Missing entities

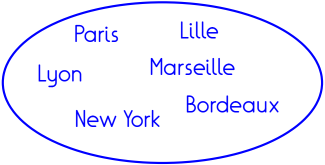
We have the following cities in our knowledge base:



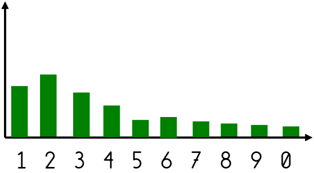
Are there any cities missing?

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

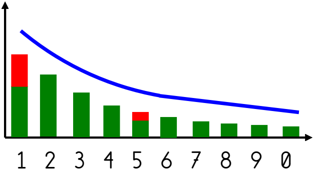
Incompleteness: Missing entities

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

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

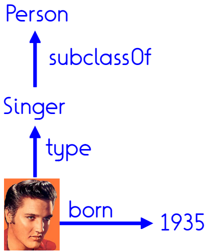
=> We can give a minimum numbers of cities that are missing
to make the distribution representative of the real world.

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



[ISWC 2018]

This talk



1. Constructing Knowledge Bases
2. Completing Knowledge Bases

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

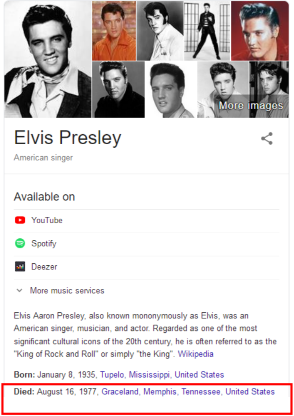
Died: August 16, 1977, Graceland, Memphis, Tennessee, United States

???

yago select knowledge

Elvis $\xrightarrow{\text{diedIn}}$ 1977

Is Elvis dead?



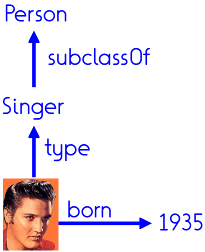
???

yago select knowledge

Elvis $\xrightarrow{\text{diedIn}}$ 1977

100m statements
95% accuracy
-> 5m wrong statements

Knowledge Bases



1. Constructing Knowledge Bases ✓
2. Completing Knowledge Bases ✓