A Hitchhiker's guide to Ontology

Fabian M. Suchanek
Télécom ParisTech University
Paris, France

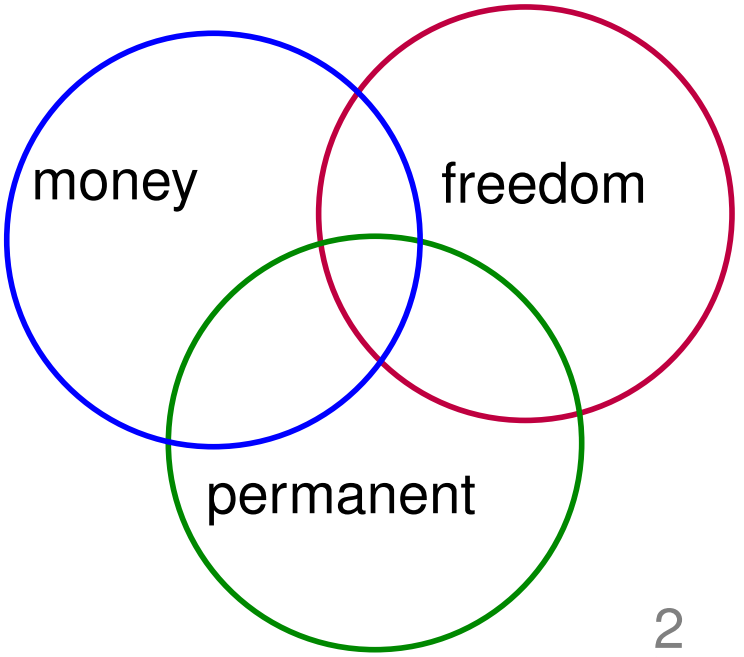
me>13
intro>19

Fabian M. Suchanek

2003:

2005:

2008:



Fabian M. Suchanek

2003: BSc in Cognitive Science

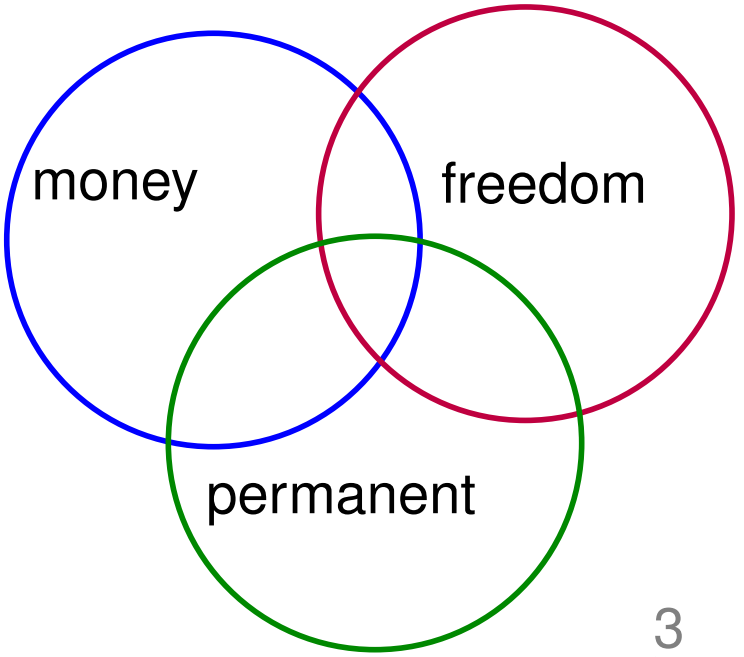
Osnabrück University/DE

2005: MSc in Computer Science

Saarland University/DE

2008: PhD in Computer Science

Max Planck Institute/DE



Fabian M. Suchanek

2003: BSc in Cognitive Science

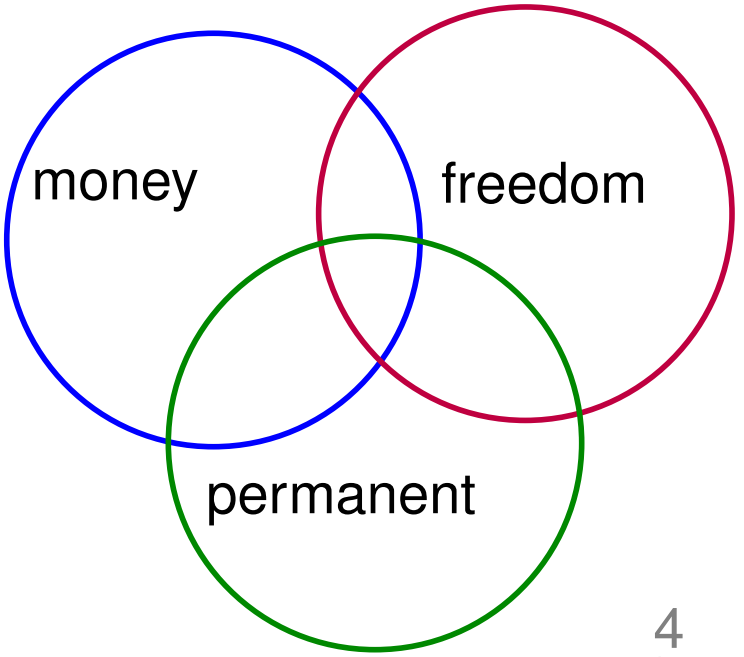
Osnabrück University/DE

2005: MSc in Computer Science

Saarland University/DE

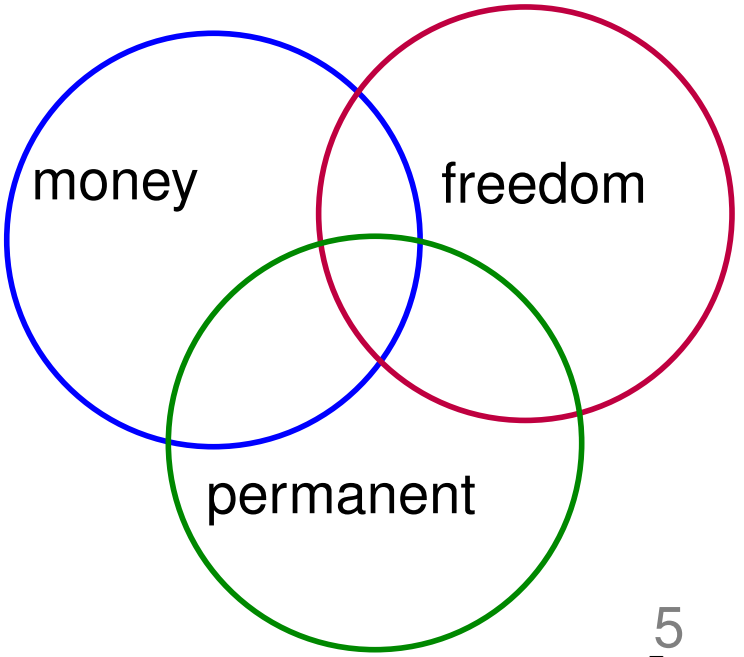
2008: PhD in Computer Science

Max Planck Institute/DE



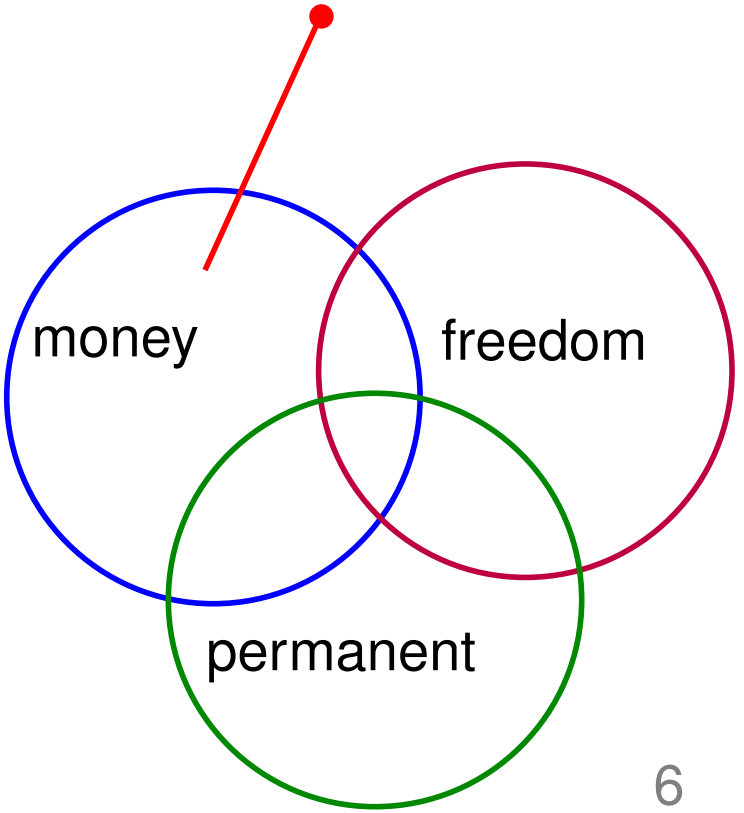
Fabian M. Suchanek

2009: PostDoc at Microsoft Research
Silicon Valley/US



Fabian M. Suchanek

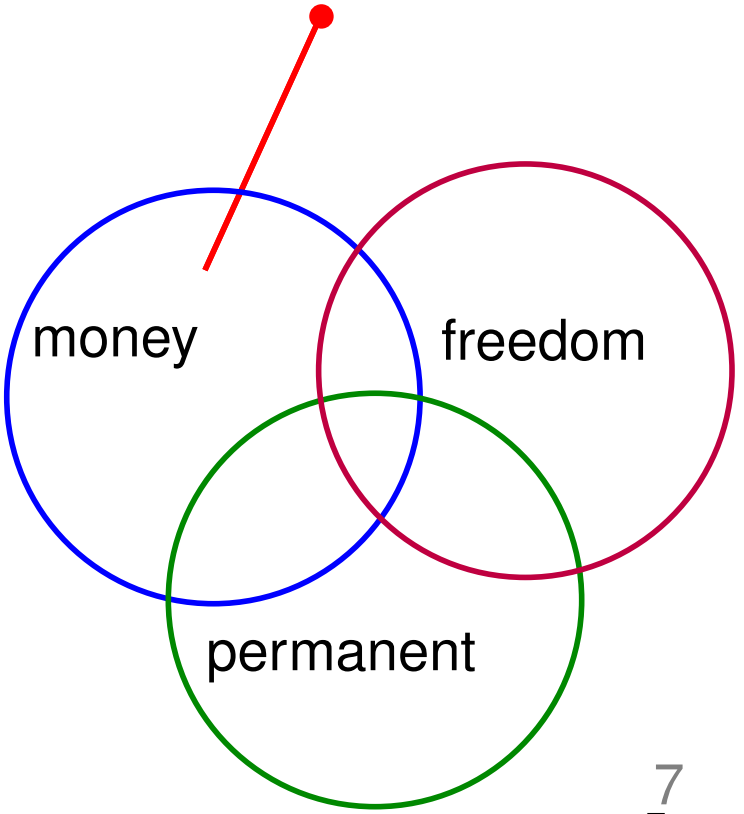
2009: PostDoc at Microsoft Research
Silicon Valley/US



Fabian M. Suchanek

2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR



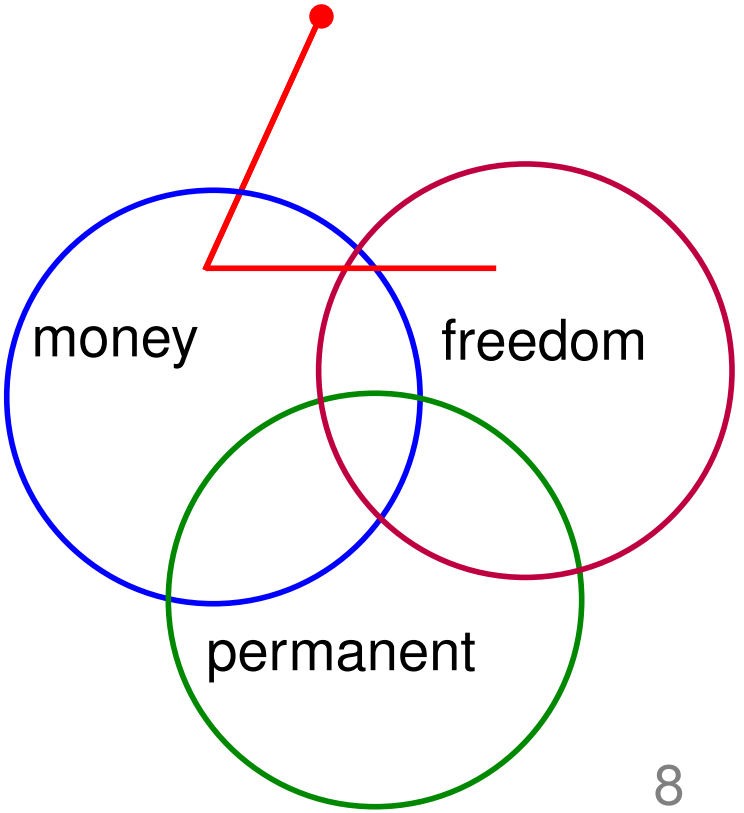
Fabian M. Suchanek

2009: PostDoc at Microsoft Research

Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR



Fabian M. Suchanek

2009: PostDoc at Microsoft Research

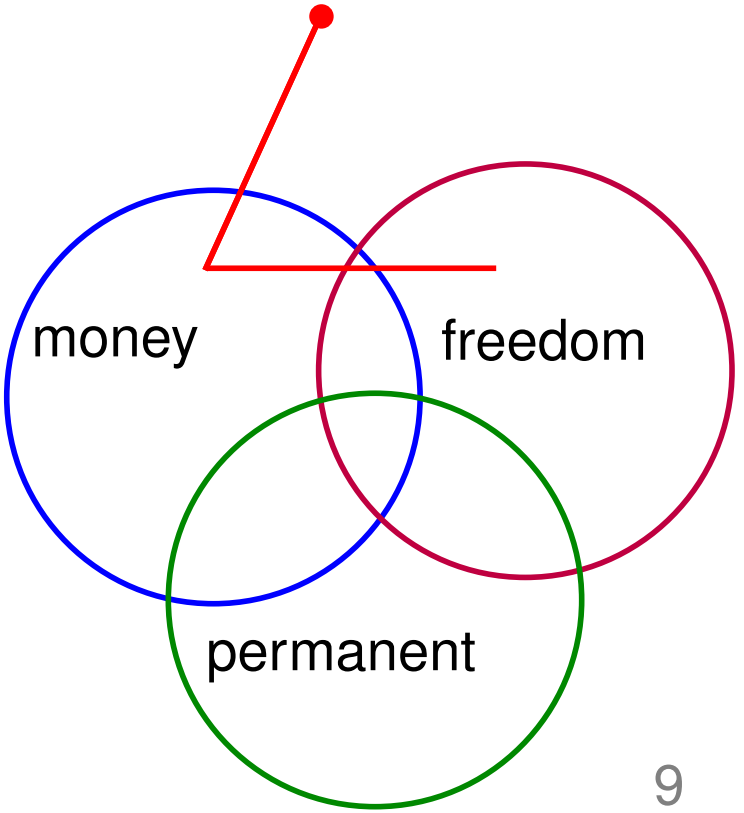
Silicon Valley/US

2010: PostDoc

INRIA Saclay/FR

2012: Research group leader

Max Planck Institute/DE

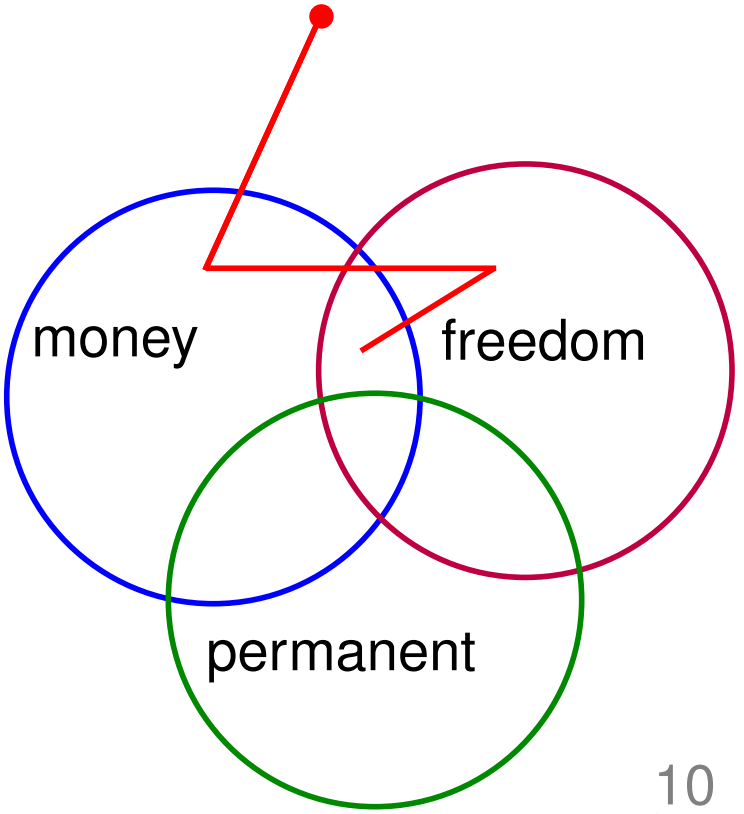


Fabian M. Suchanek

2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DE



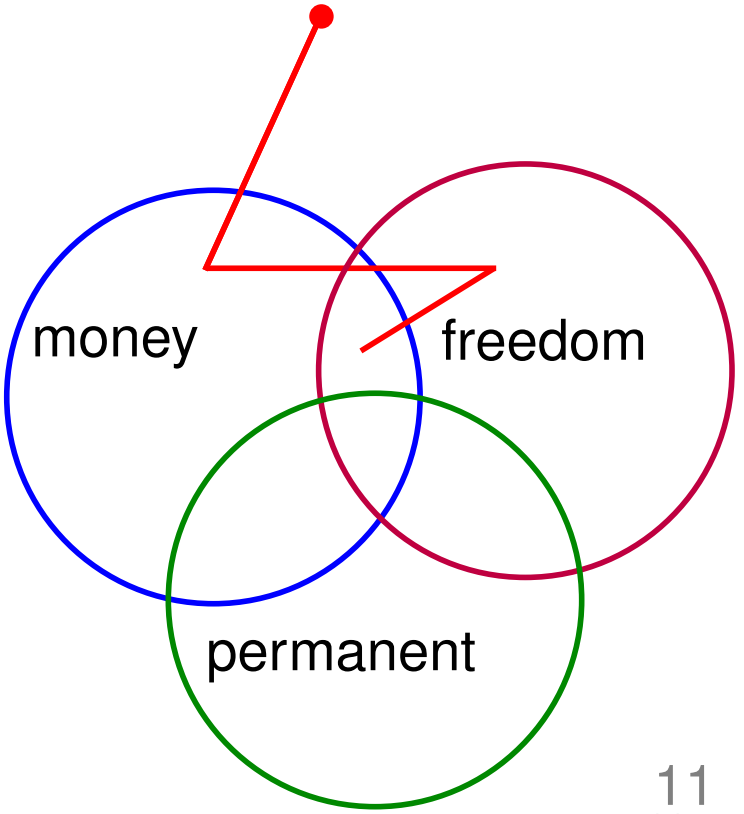
Fabian M. Suchanek

2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DE

2013: Associate Professor
Télécom ParisTech/FR



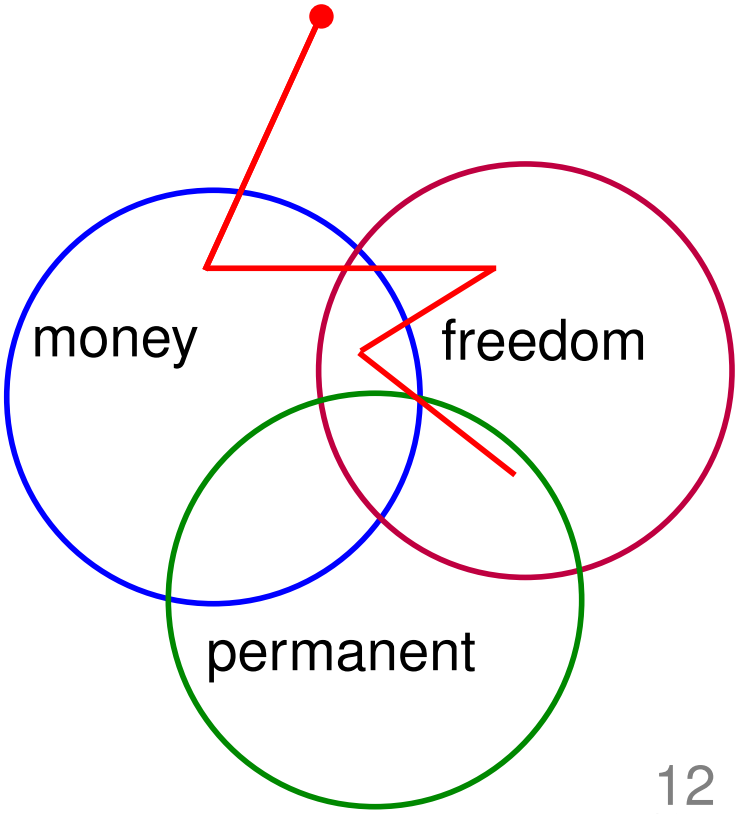
Fabian M. Suchanek

2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DE

2013: Associate Professor
Télécom ParisTech/FR



Fabian M. Suchanek

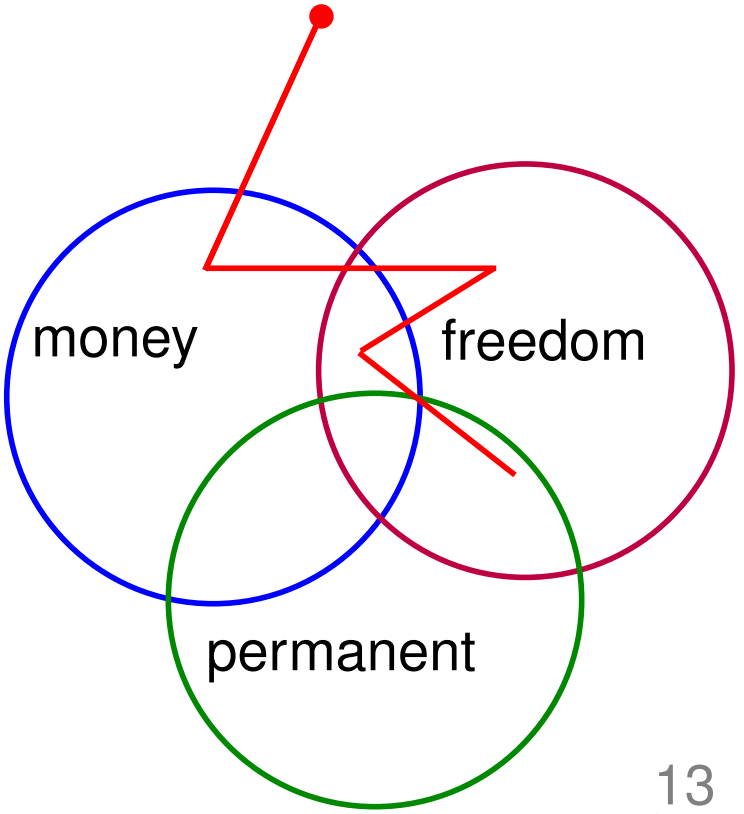
2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DE

2013: Associate Professor
Télécom ParisTech/FR

2016: Full Professor
Télécom ParisTech/FR



Fabian M. Suchanek

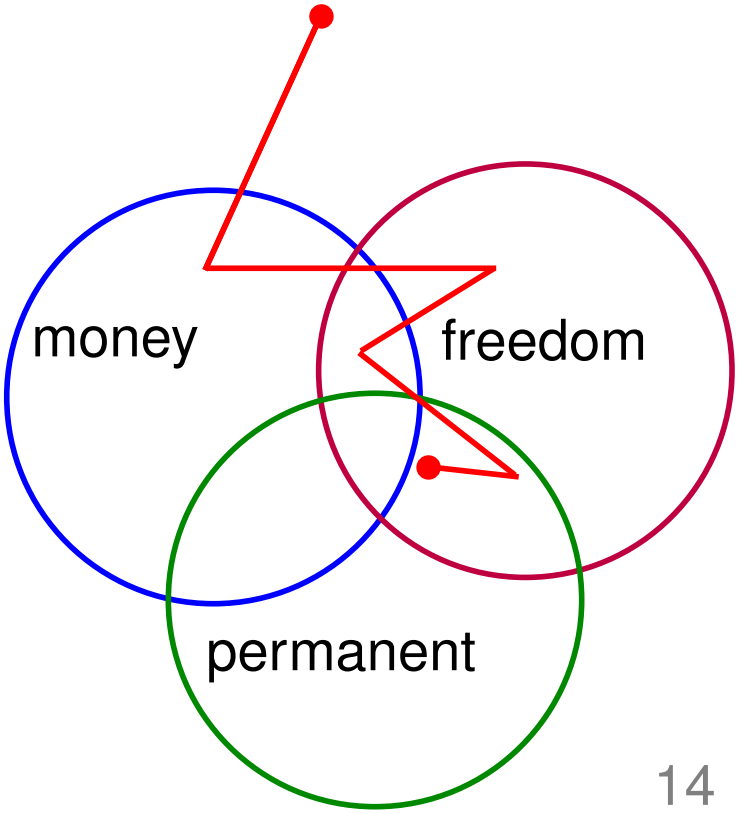
2009: PostDoc at Microsoft Research
Silicon Valley/US

2010: PostDoc
INRIA Saclay/FR

2012: Research group leader
Max Planck Institute/DE

2013: Associate Professor
Télécom ParisTech/FR

2016: Full Professor
Télécom ParisTech/FR



I am an Elvis Fan!

1961

Meeting Elvis Presley

1961

2015

Old Elvis from Sachsmedia

Meeting Elvis Presley

1961

2015

At Elvis' ranch
in New Zealand

Old Elvis from Sachsmedia

Ranch is Murchison Lodge in New Zealand

Knowledge Cards

Google

Elvis Presley

Elvis Presley - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Elvis_Presley** ▾ Wikipedia ▾
Elvis Aaron **Presley** (January 8, 1935 – August 16, 1977) was an American singer and actor. Regarded as one of the most significant cultural icons of the 20th ...
Priscilla Presley - Albums discography - Singles discography - Michael Lockwood

Elvis Presley Biography - Elvis Life Story - **Elvis Presley**
www.**elvis**.com/about-the-king/biography_.aspx ▾ Elvis Presley ▾
Elvis Presley The incredible Elvis life story began when Elvis Aaron Presley was born to Vernon and Gladys Presley in a two-room house in Tupelo, Mississippi, ...

Elvis Presley Official Web Site
www.**elvis**.com/ ▾ Elvis Presley ▾
Play the **Elvis** Slots on Slotomania. Play slots like a king with the brand new "**ELVIS**" Slotomania video slots game now available through Facebook! Prev Next.

In the news

Listening to **Elvis Presley's** Bizarre Album of Stage Banter
Mental Floss - 20 hours ago
In 1974, **Elvis Presley** released Having Fun With Elvis on Stage, an album consisting of

Elvis Presley

Singer

Elvis Aaron Presley was an American singer 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

Died: August 16, 1977, Memphis, Tennessee, United States
Spouse: Priscilla Presley (m. 1967–1973)
Children: Lisa Marie Presley

Songs

My Way	1973	Aloha From Hawaii via ...
Jailhouse Rock		
Love Me Tender	1972	As Recorded at Madiso...
Love Me	1972	As Recorded at Madiso...

Knowledge Cards

Google

Elvis Presley

Elvis Presley - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Elvis_Presley** ▾ Wikipedia ▾
Elvis Aaron **Presley** (January 8, 1935 – August 16, 1977) was an American singer and actor. Regarded as one of the most significant cultural icons of the 20th ...
Priscilla Presley - Albums discography - Singles discography - Michael Lockwood

Elvis Presley Biography - Elvis Life Story - **Elvis Presley**
www.**elvis**.com/about-the-king/biography_.aspx ▾ Elvis Presley ▾
Elvis Presley The incredible Elvis life story began when Elvis Aaron Presley was born to Vernon and Gladys Presley in a two-room house in Tupelo, Mississippi, ...

Elvis Presley Official Web Site
www.**elvis**.com/ ▾ Elvis Presley ▾
Play the **Elvis** Slots on Slotomania. Play slots like a king with the brand new "**ELVIS**" Slotomania video slots game now available through Facebook! Prev Next.

In the news

Listening to **Elvis Presley's** Bizarre Album of Stage Banter
Mental Floss - 20 hours ago
In 1974, **Elvis Presley** released Having Fun With Elvis on Stage, an album consisting of

Elvis Presley

Singer

Elvis Aaron Presley was an American singer 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

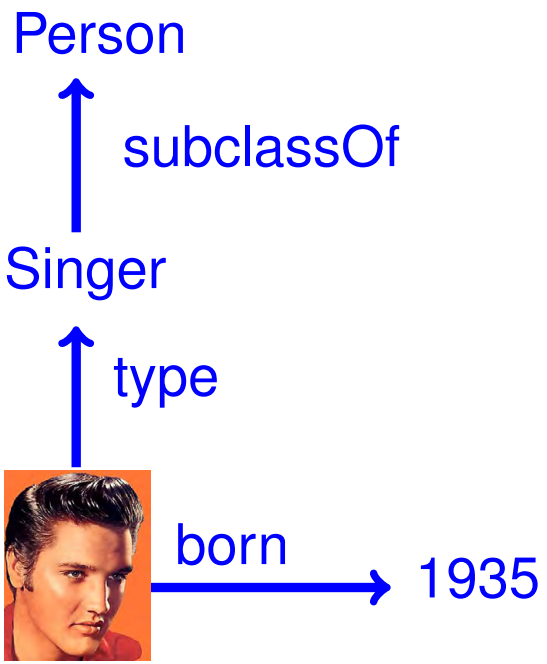
Died: August 16, 1977, Memphis, Tennessee, United States
Spouse: Priscilla Presley (m. 1967–1973)
Children: Lisa Marie Presley

???

Songs

My Way	1973	Aloha From Hawaii via ...
Jailhouse Rock		
Love Me Tender	1972	As Recorded at Madiso...
Love Me	1972	As Recorded at Madiso...

Knowledge Bases



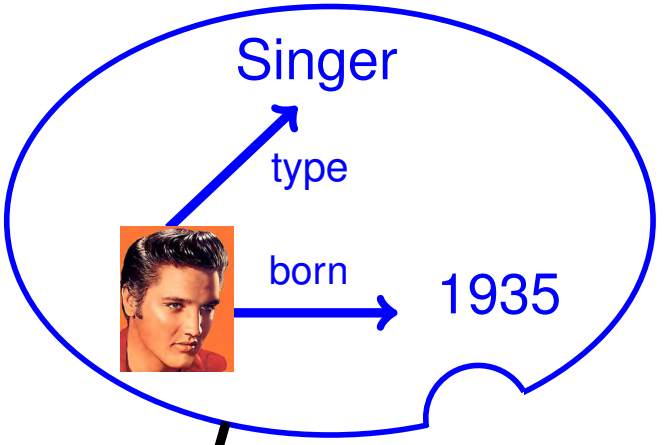
For us, a knowledge base (“KB”, “ontology”) is a graph, where the nodes are entities and the edges are relations.

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

Knowledge Base Life Cycle

Creativly extending knowledge bases ->

Constructing knowledge bases



Protecting knowledge bases ->

Mining knowledge bases ->

$A \wedge B \Rightarrow C$

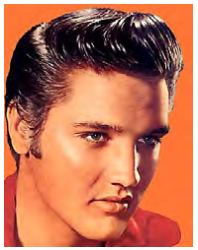
Determining Incompleteness ->

Extracting from Wikipedia

Elvis Presley

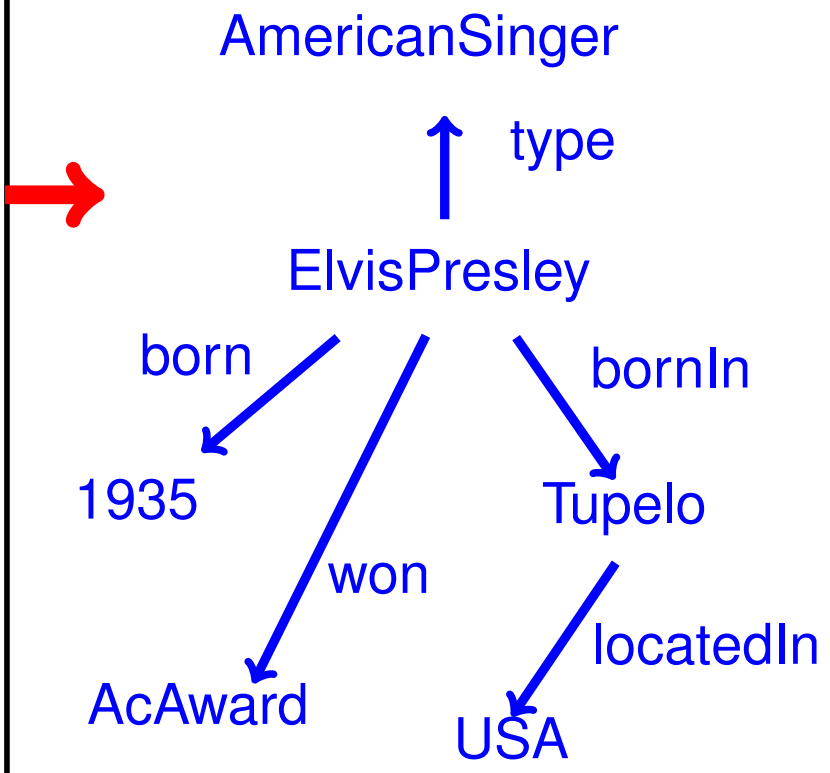
WIKIPEDIA
The Free Encyclopedia

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

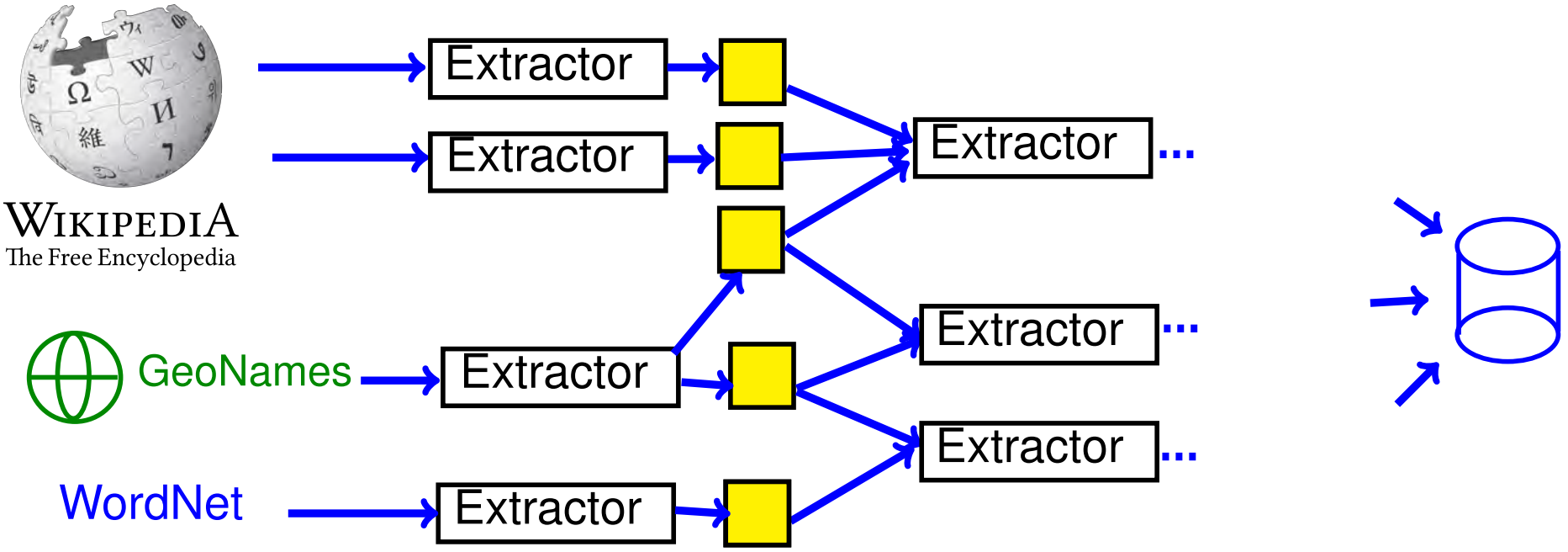


Born: 1935
In: Tupelo
...

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



Creating a large knowledge base

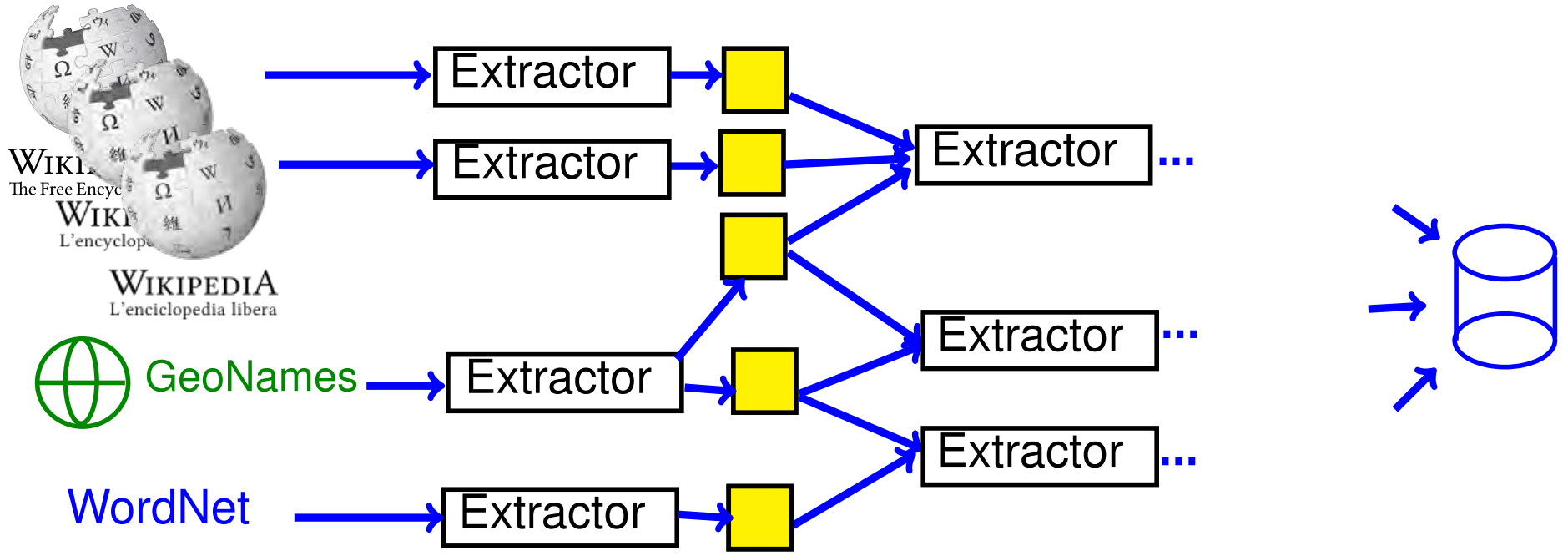


Intermediate extractors

- clean facts
- deduplicate facts and entities
- check consistency

ensuring high quality (95%)

Creating a large knowledge base



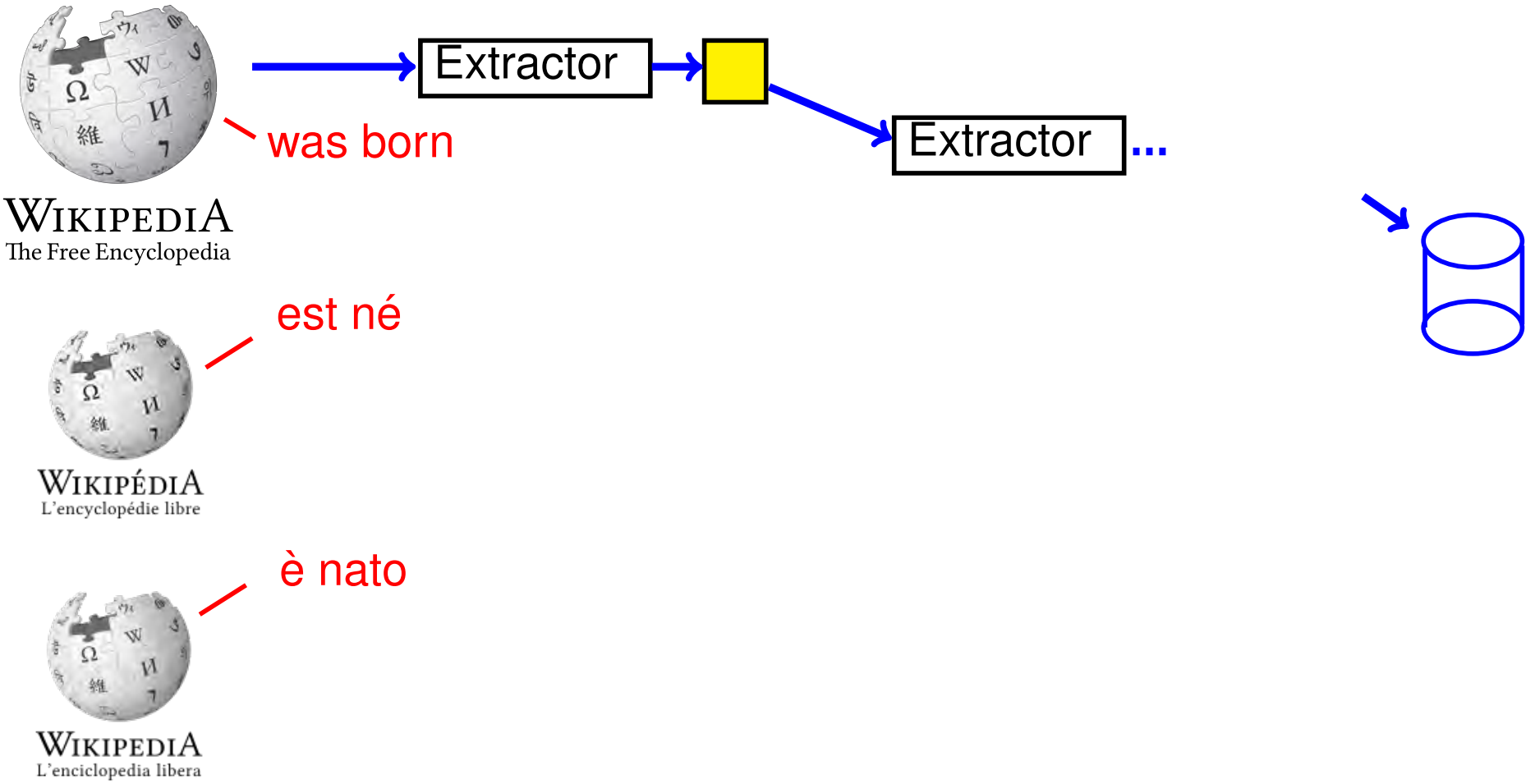
Intermediate extractors

- clean facts
- deduplicate facts and entities
- check consistency

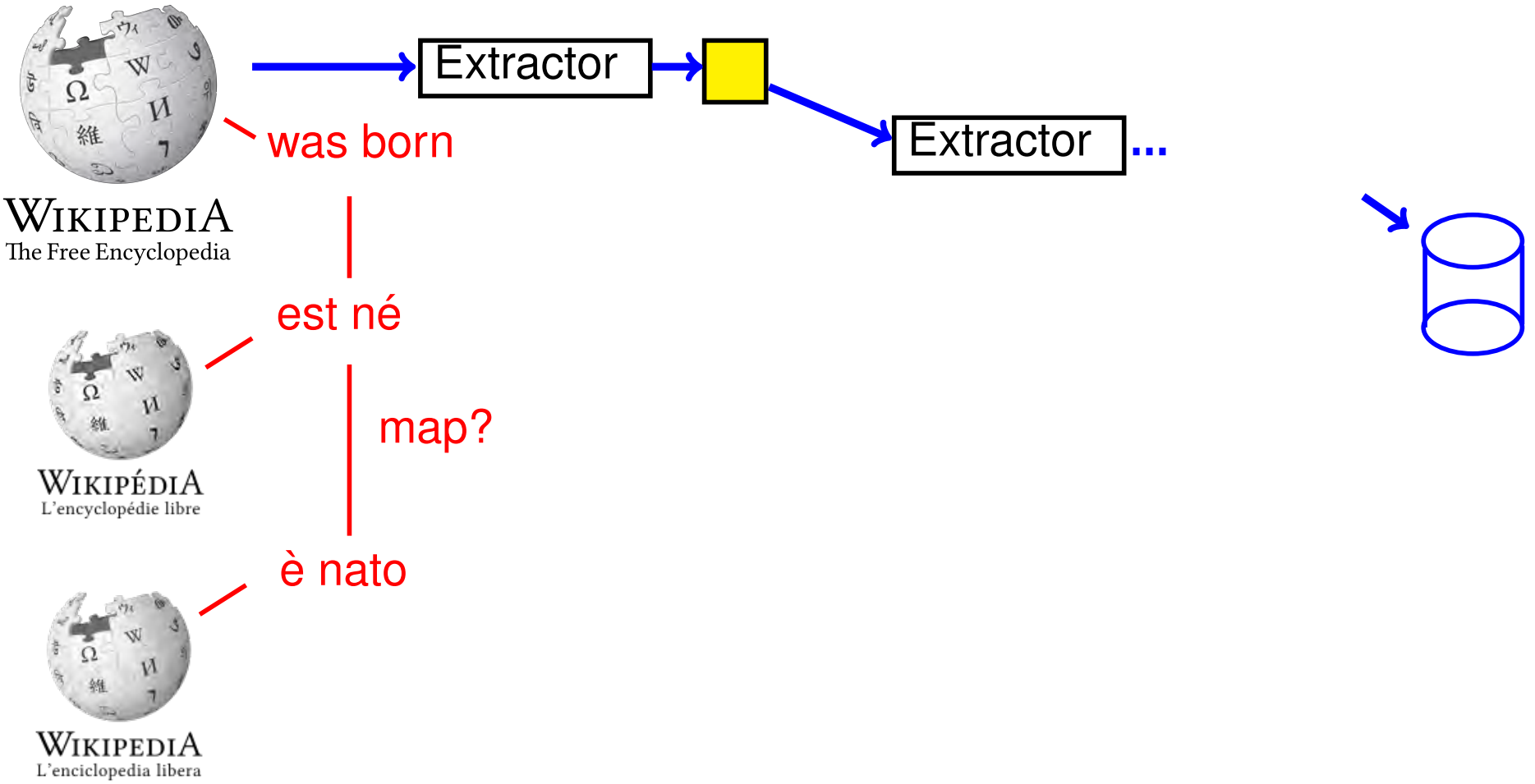
ensuring high quality (95%)

multilingual>29

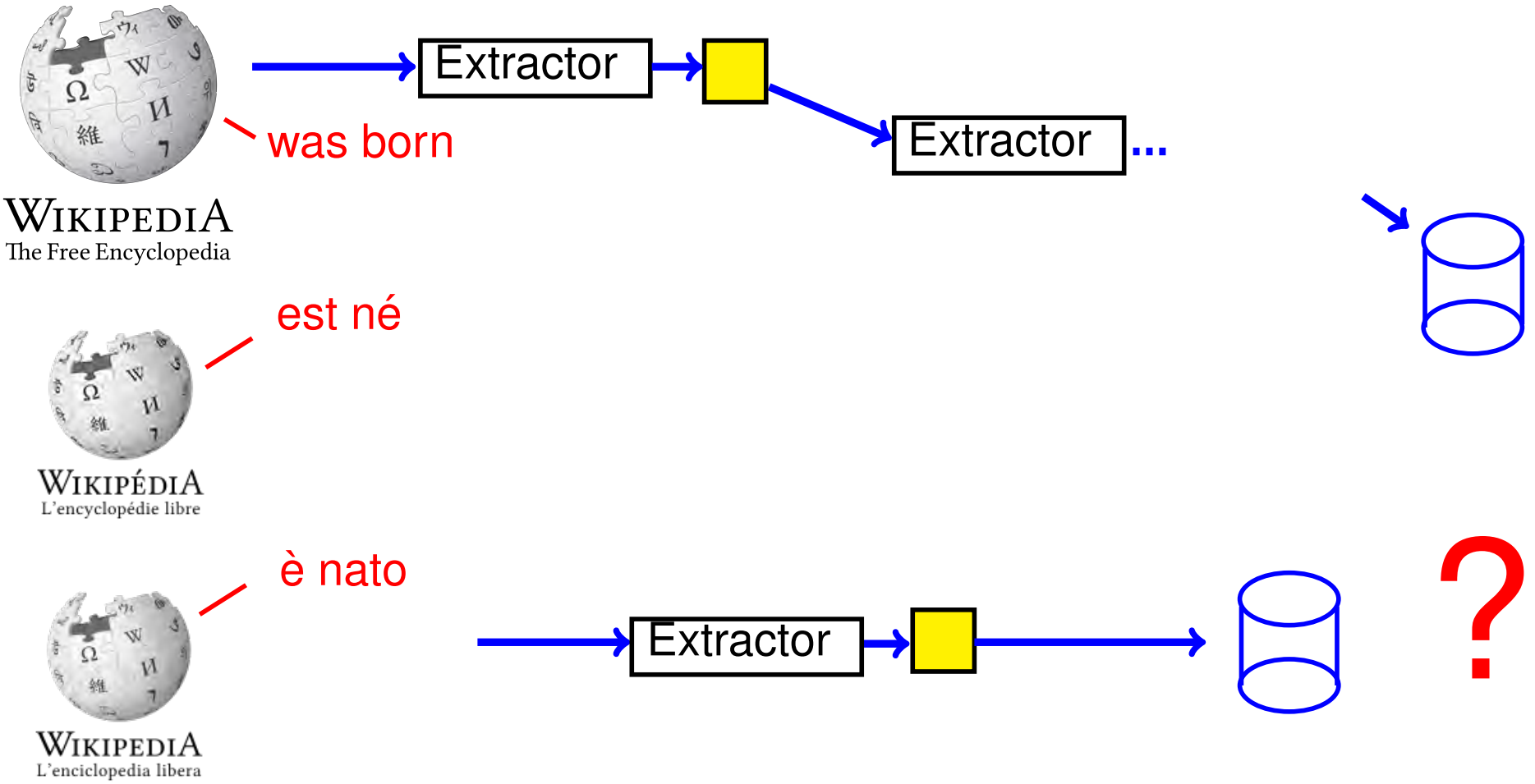
Integrating multilingual Wikis



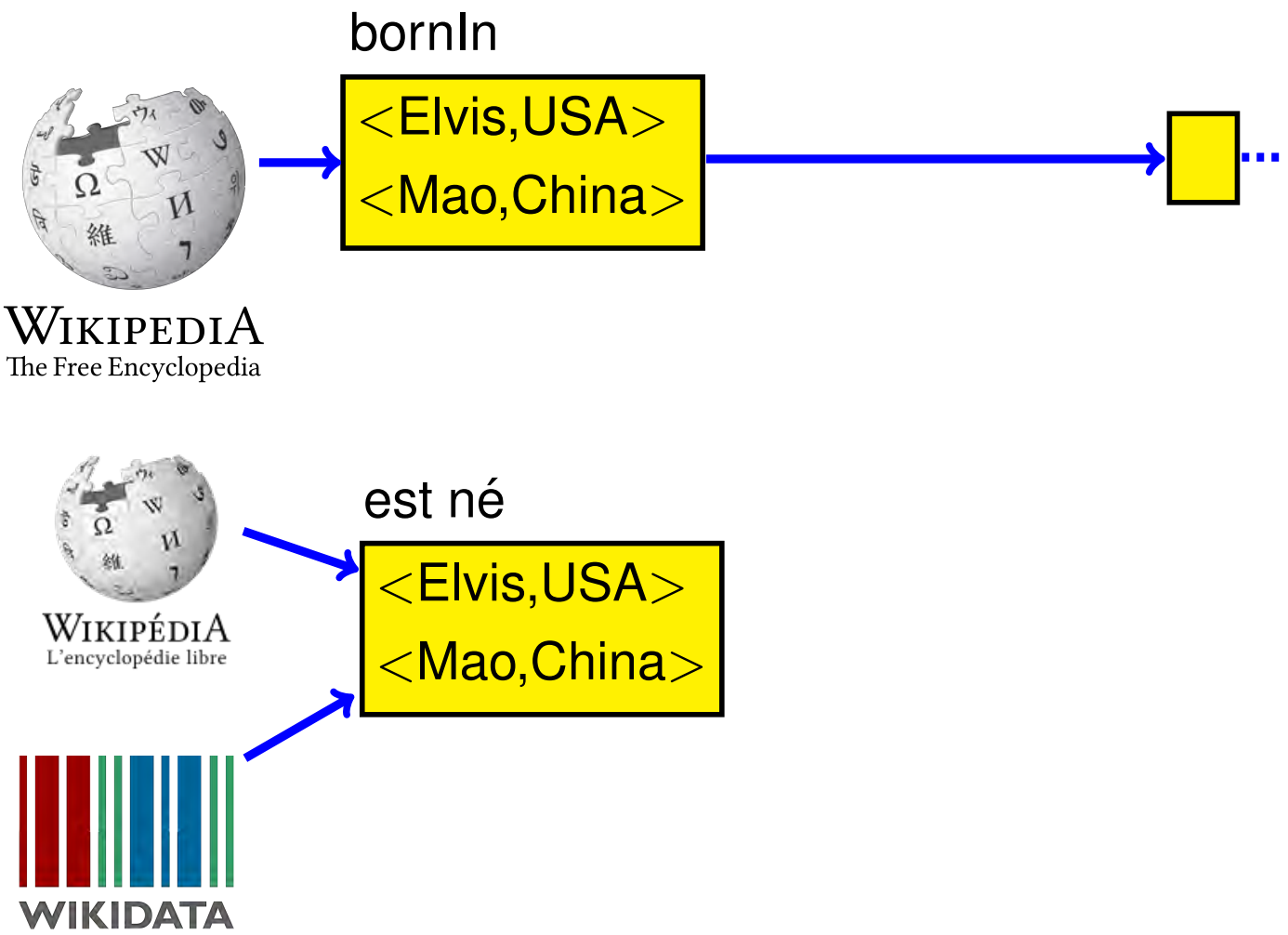
Integrating multilingual Wikis



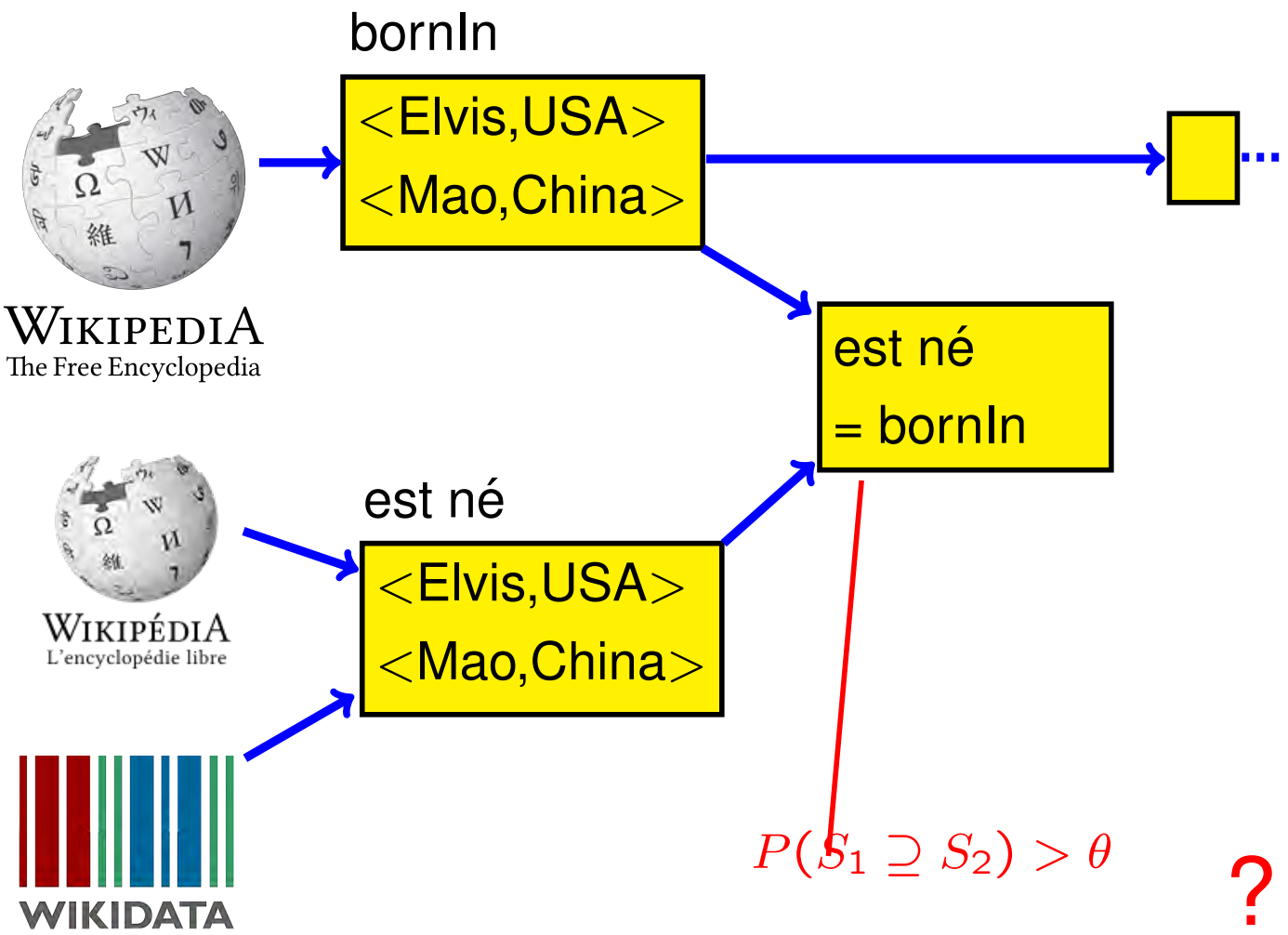
Integrating multilingual Wikis



Integrating multilingual Wikis

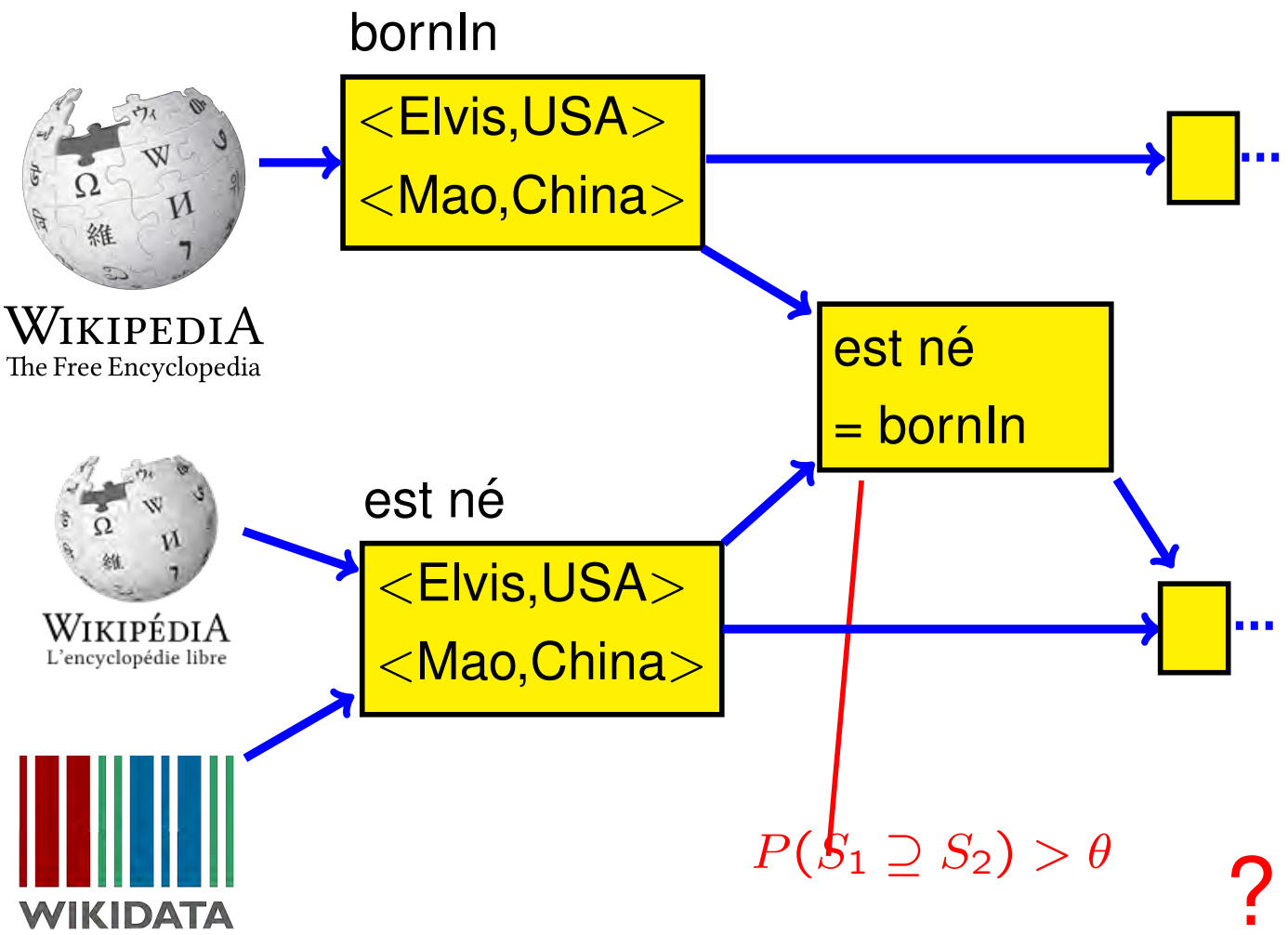


Integrating multilingual Wikis



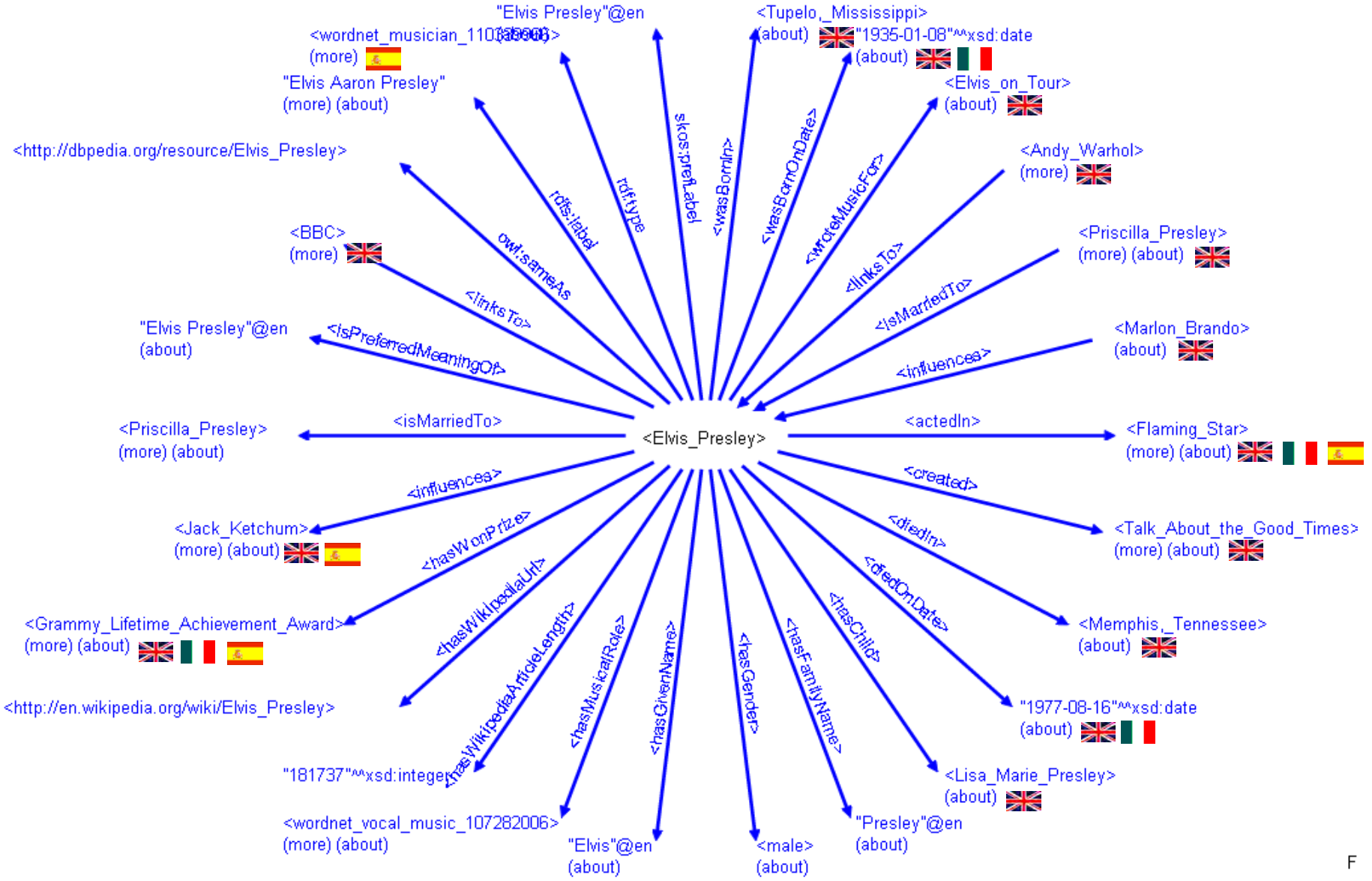
[CIDR 2015]

Integrating multilingual Wikis



[CIDR 2015]

Example: YAGO about Elvis



YAGO: a large knowledge base

http://yago-knowledge.org

soon open-source!

Wikipedia + WordNet

time and space

10 languages

100 relations

100m facts

10m entities

95% accuracy

used by DBpedia

and IBM Watson

Caveat:

focus on

precision!

[WWW2007,JWS2008,WWW2011 demo,AIJ2013,WWW2013 demo,CIDR20

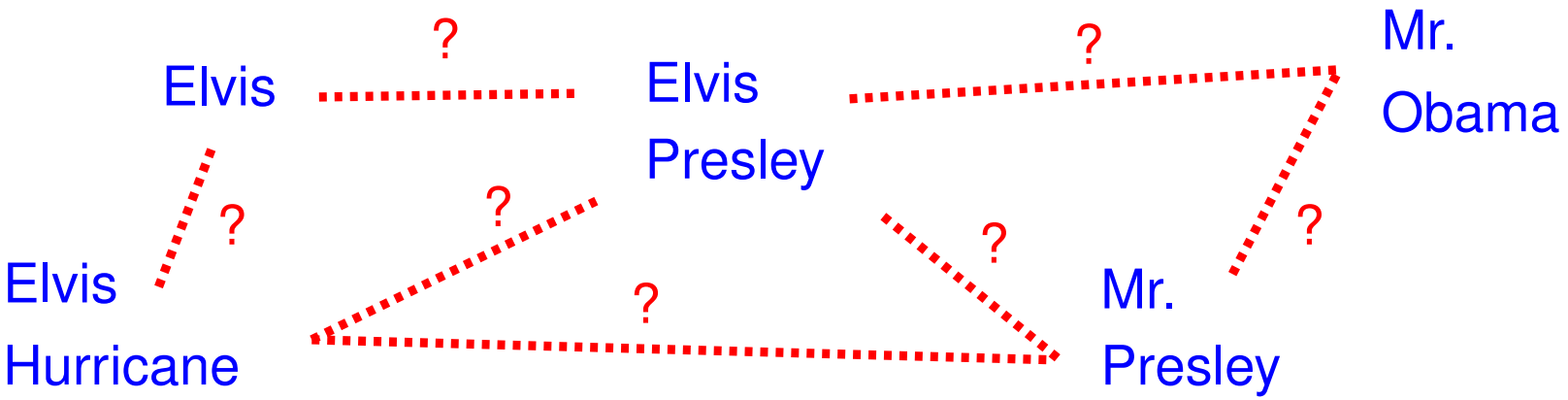




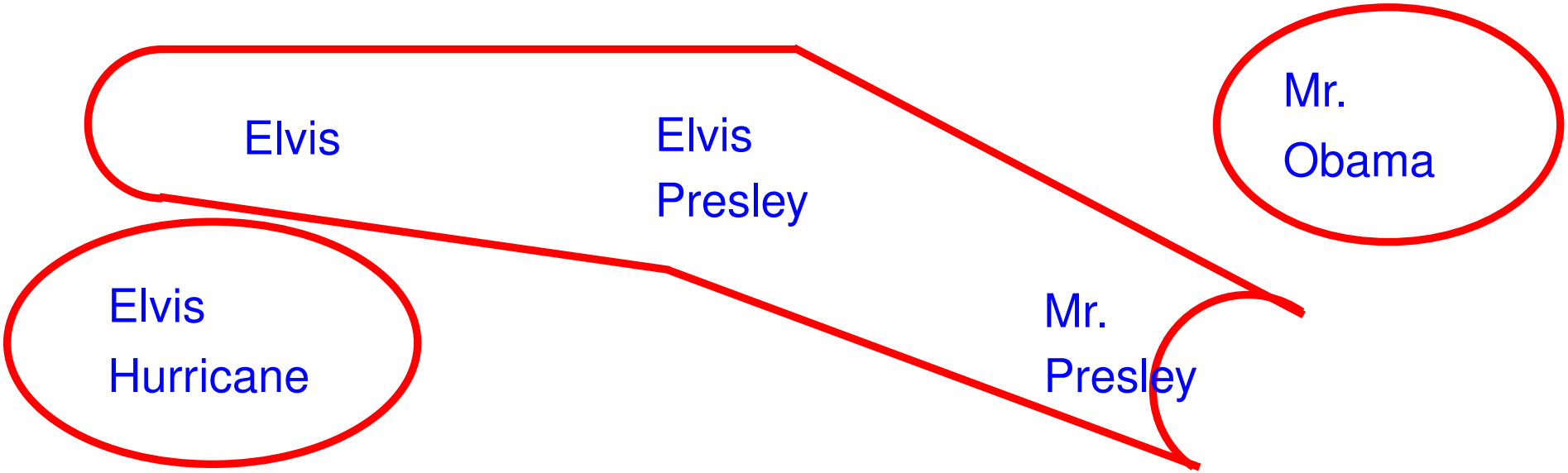


canonic&IBEX>42 canonic>37

Canonicalizing New Entities



Canonicalizing New Entities

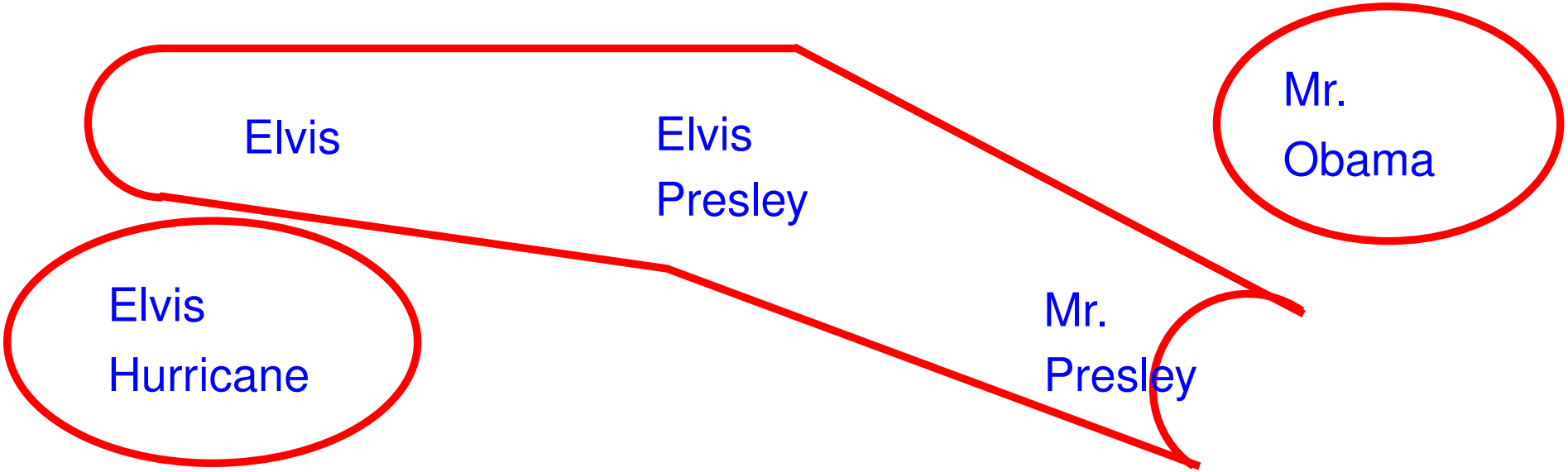


Use hierarchical agglomerative clustering

- TF-IDF token overlap
- Triple overlap
- String Similarity
- Word overlap in source docs
- Type overlap
- Overlap of co-mentions

details>36

Canonicalizing New Entities

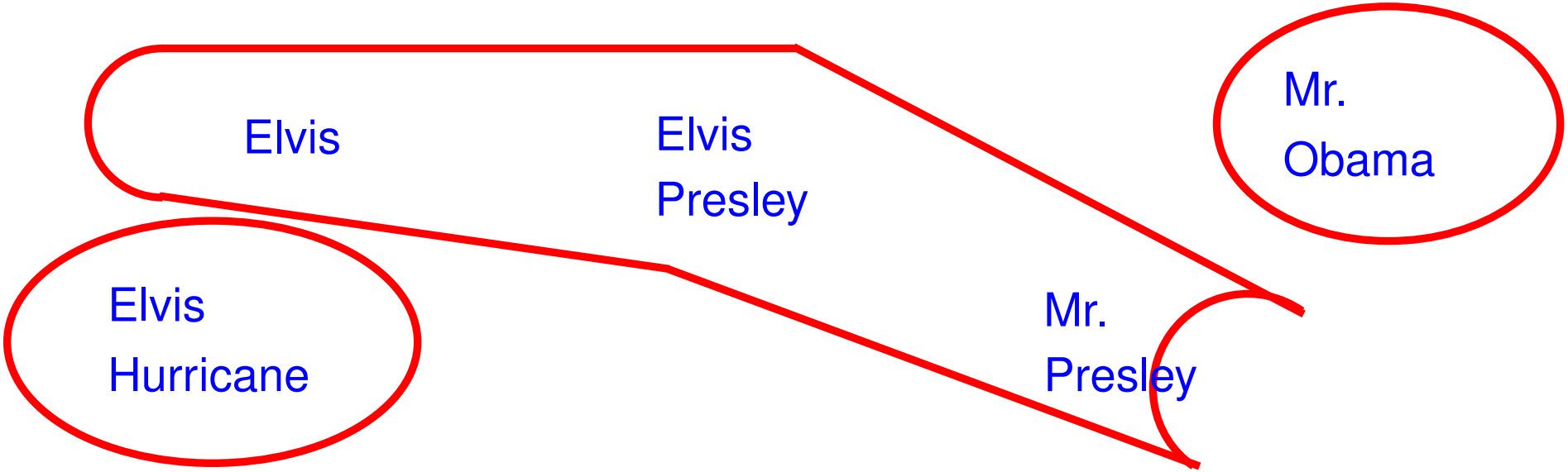


Use hierarchical agglomerative clustering

- TF-IDF token overlap
- Triple overlap
- String Similarity
- Word overlap in source docs
- Type overlap
- Overlap of co-mentions

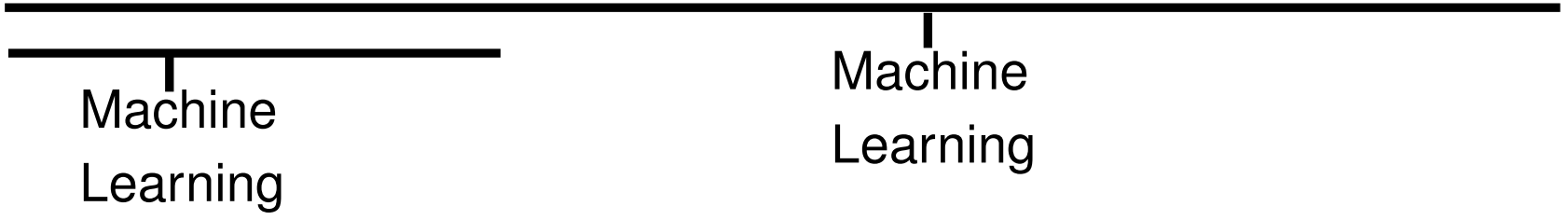
Machine Learning

Canonicalizing New Entities

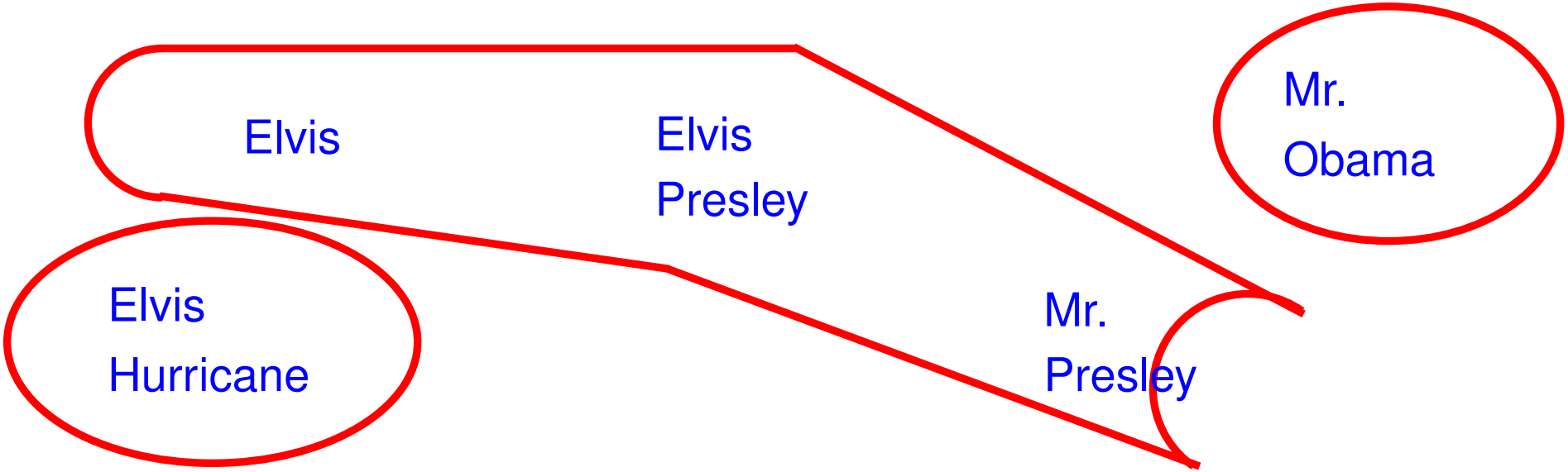


Use hierarchical agglomerative clustering

- TF-IDF token overlap
- Triple overlap
- String Similarity
- Word overlap in source docs
- Type overlap
- Overlap of co-mentions



Canonicalizing New Entities



Use hierarchical agglomerative clustering

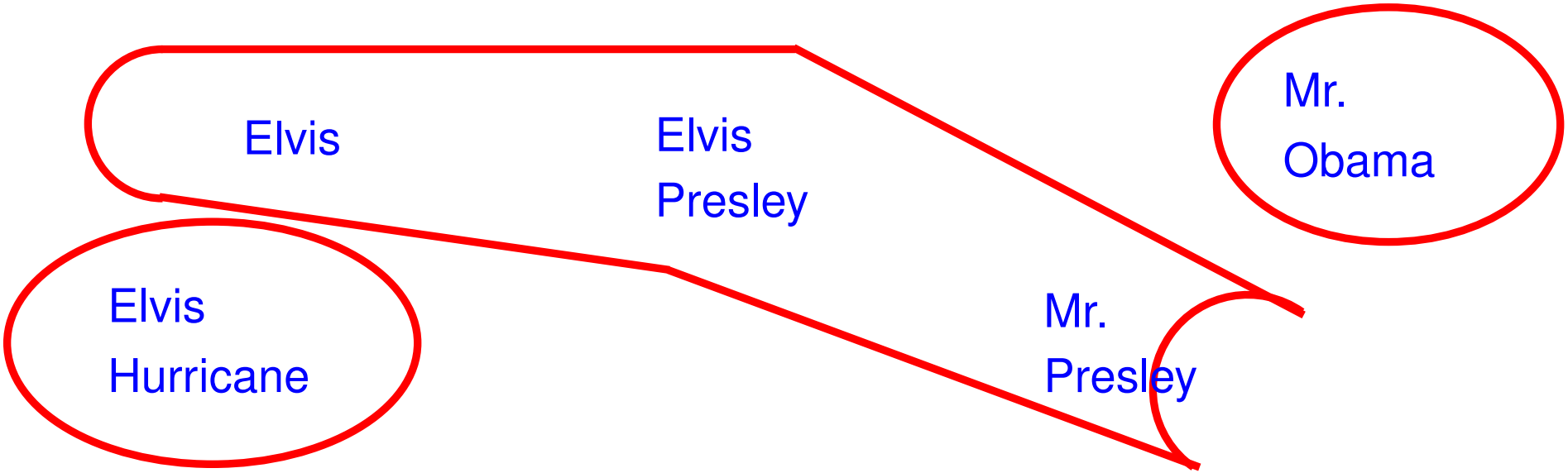
- TF-IDF token overlap
- Triple overlap
- String Similarity
- Word overlap in source docs
- Type overlap
- Overlap of co-mentions

Machine Learning

Machine Learning

[CIKM 2014]

Canonicalizing New Entities

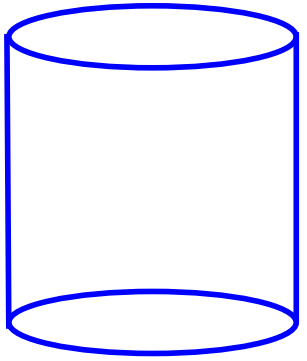





Google

[CIKM 2014] IBEX>4

Goal: Harvest entities from the Web



IBEX: Collect unique ids

Puma PowerTech Blaze Running Shoe - Mens - Black/SilverSelect Width-M - UPC # 887128476661

© Whittleseacitylac larger image

AUD99.02

Please Choose:

Size

- (EUR):40/US8.5/(UK):6
- (EUR):41/US9.5/(UK):7
- (EUR):42,(US):8.5,(UK):7.5
- (EUR):43,(US):9.5,(UK):8.5
- (EUR):44,(US):10,(UK):9
- (EUR):45,(US):11,(UK):10
- (EUR):46,(US):12,(UK):11

Add to Cart: 1

VERIFIED & SECURED

Item # 307110

UPC # 887128476661

Go big or go home in the men's Puma PowerTech Blaze running shoe! This performance trainer has upgraded technical features with its OrthoLite® footbed, PowerTech™ cushioning and EverTrack™ abrasion-resistant outsole.

This men's running shoe is best for the **neutral** runner. What is pronation?

Breathable mesh upper with synthetic overlay for structured support

Lace-up closure

887128476661

Unique identifiers can be verified by a checksum.

They exist for products, books, documents, chemicals,...

IBEX: Collect unique ids

Puma PowerTech Blaze Running Shoe - Mens - Black/SilverSelect Width-M - UPC # 887128476661

AUD99.02

Please Choose:

Size

- (EUR):40/US8.5/(UK):6
- (EUR):41/US9.5/(UK):7
- (EUR):42,(US):8.5,(UK):7.5
- (EUR):43,(US):9.5,(UK):8.5
- K):9
- K):10
- K):11

VERIFIED & SECURED
VERIFY SECURITY

Item # 307110

UPC # 887128476661

Go big or go home in the men's Puma PowerTech Blaze running shoe! This performance trainer has upgraded technical features with its OrthoLite® footbed, PowerTech™ cushioning and EverTrack™ abrasion-resistant outsole.

This men's running shoe is best for the **neutral** runner. What is pronation?

Breathable mesh upper with synthetic overlay for structured support

Lace-up closure

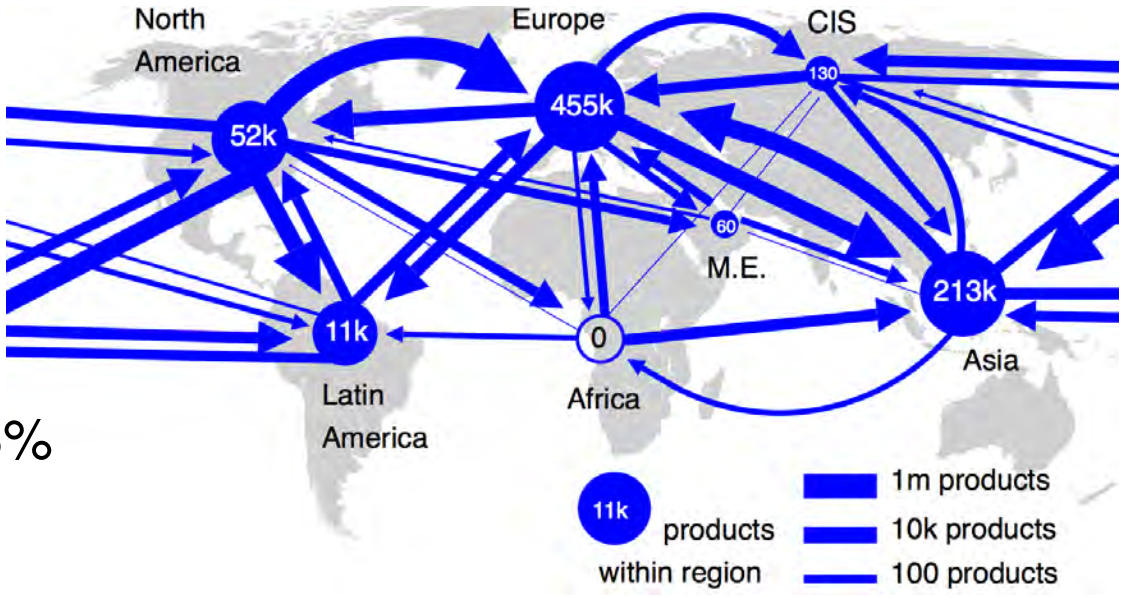
id	name	URL
1234	Puma PowerTech Blaze	u1
1234	Please choose	u1
1234	Puma PowerTech Blaze	u2
1234	Puma Shoe	u2
5678	Please choose	u3
5678	Sony Cybershot TS100	u3
...	...	...

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

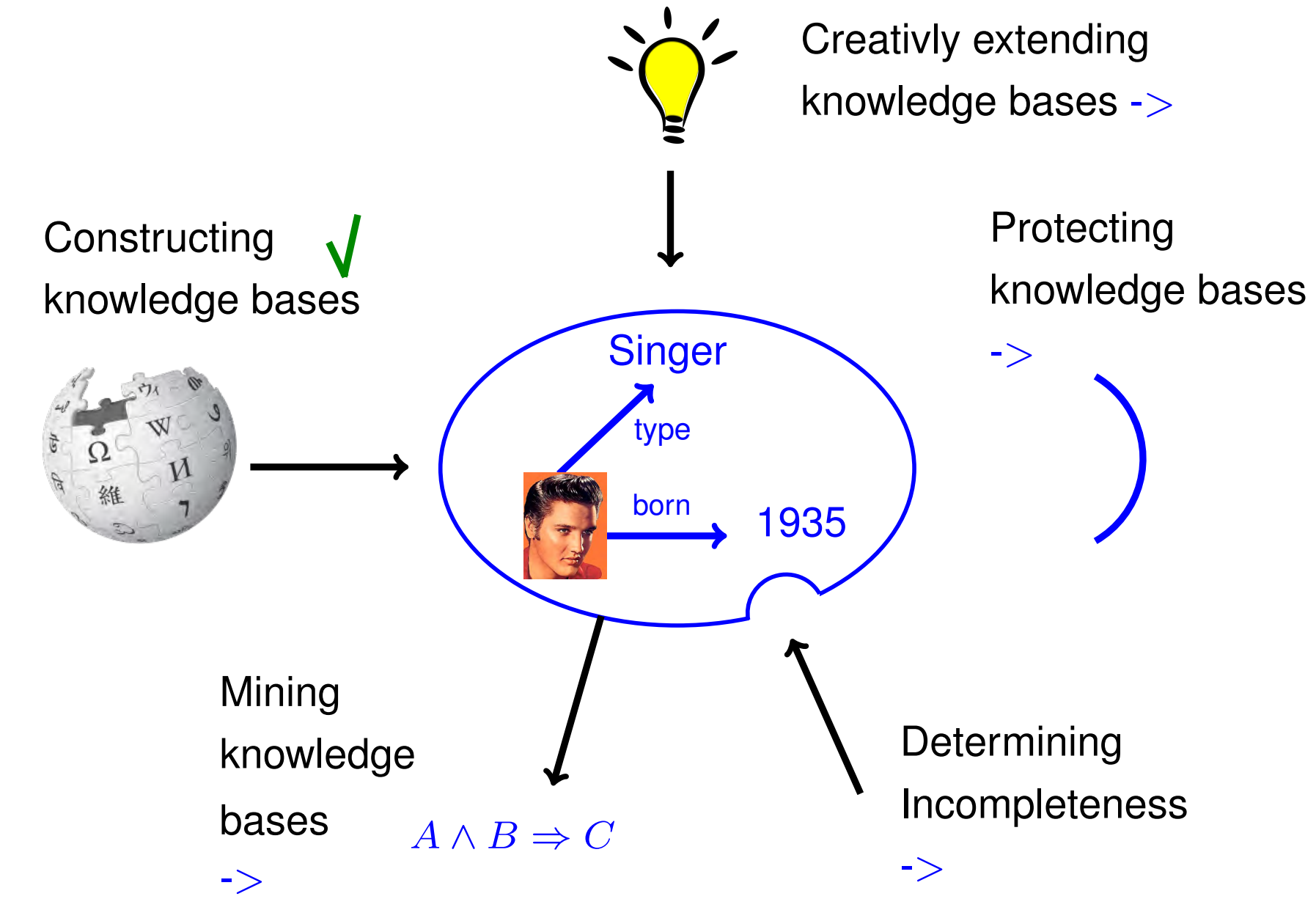
- Global trade flow
- frequent email providers
- frequent people names and more



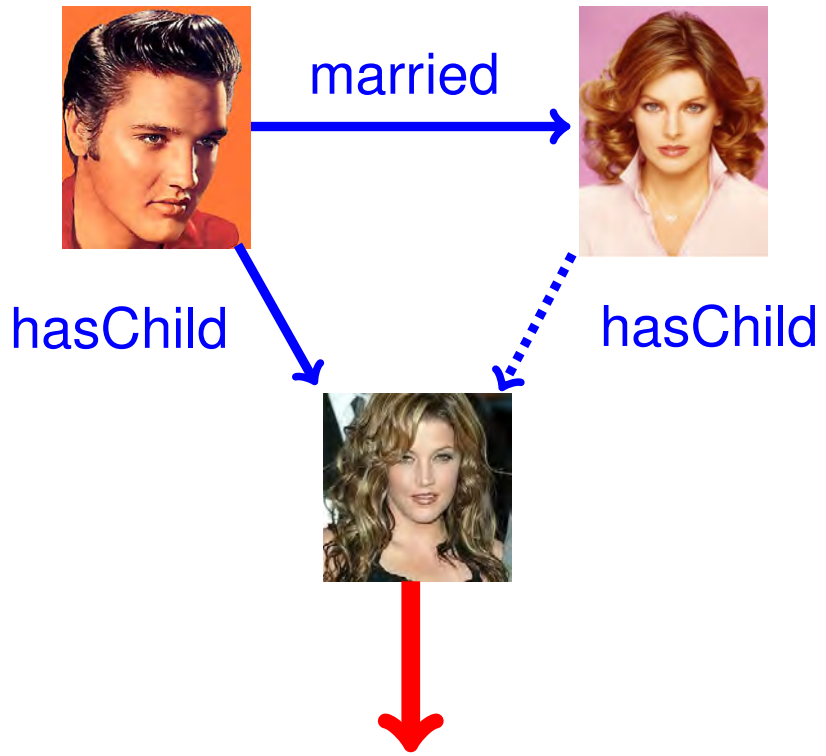

[WebDB 2015]

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

Knowledge Base Life Cycle

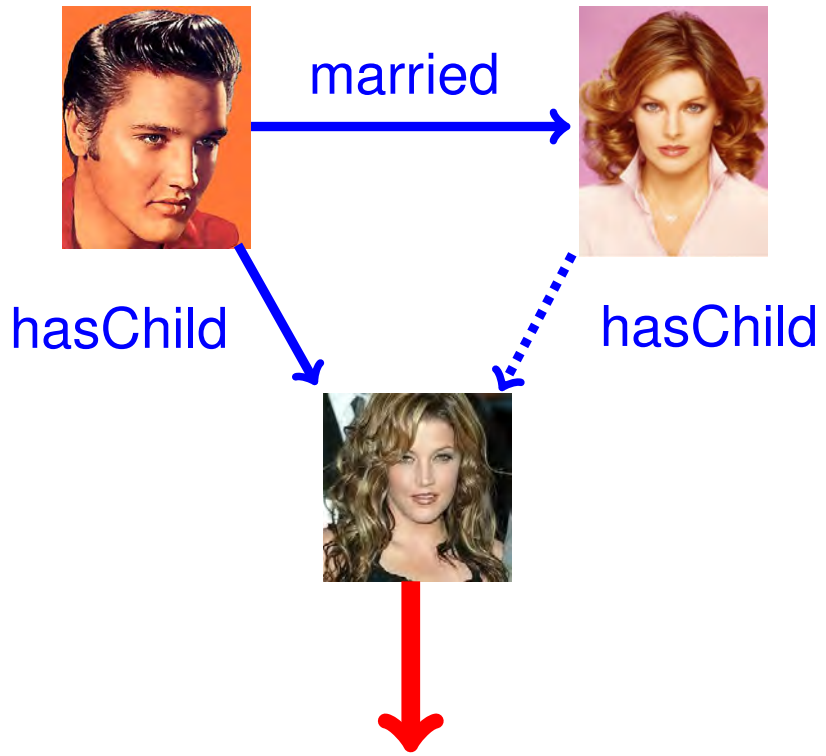


Rule Mining finds patterns



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

Rule Mining finds patterns



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

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

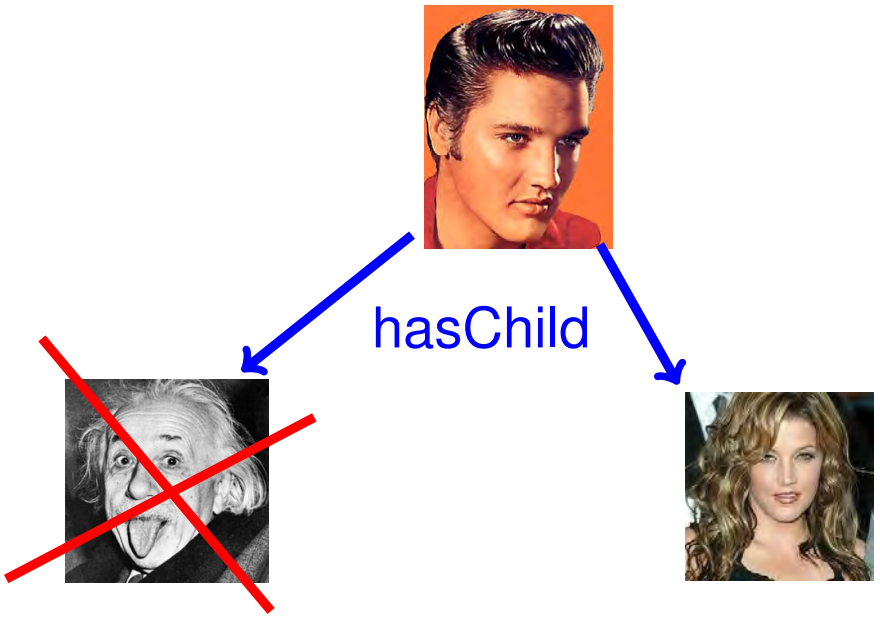
Partial Completeness Assumption

Assumption:

If we know r(x,y1),..., r(x,yn),

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

Partial Completeness Assumption

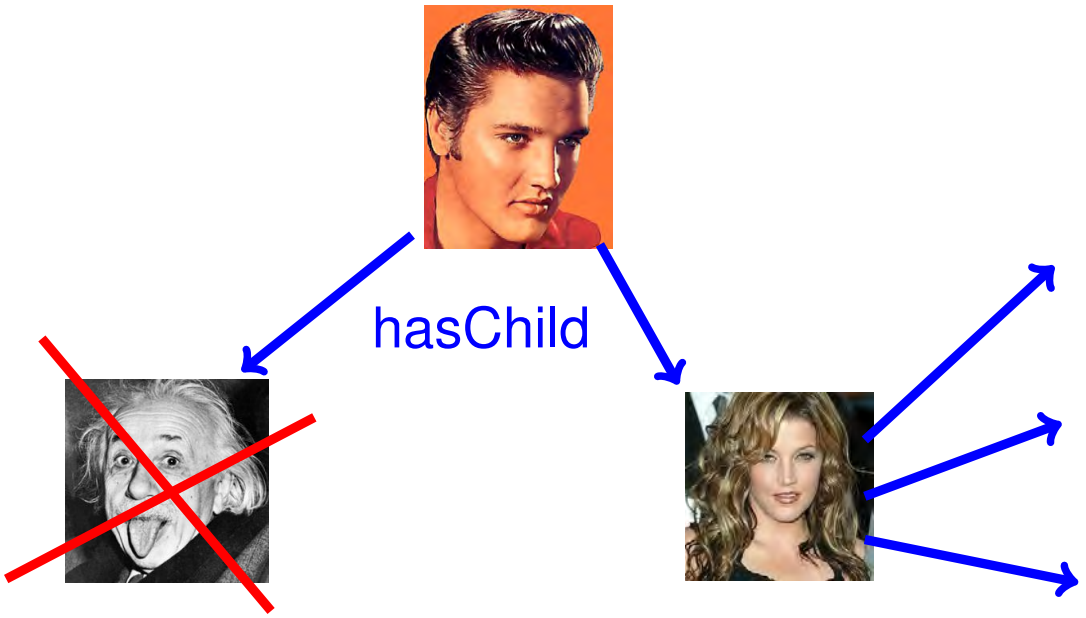


Assumption:

If we know r(x,y1),..., r(x,yn),

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

Partial Completeness Assumption



Assumption:

If we know r(x,y1),..., r(x,yn),

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

Partial Completeness Assumption

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

$$\vec{B} \Rightarrow r(x, y)$$

$$conf(\vec{B} \Rightarrow r(x, y)) = \frac{\#x,y{:}\vec{B} \wedge r(x,y)}{\#x,y{:}\exists y'{:}\vec{B} \wedge r(x,y')}$$

# of parent-child relations that we predict correctly in the KB

# of parent-child relations that we predict and where a child is known

AMIE finds rules in ontologies

AMIE

(5min)

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

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

Caveat: rules cannot predict the unknown with high precision

AMIE finds rules in ontologies

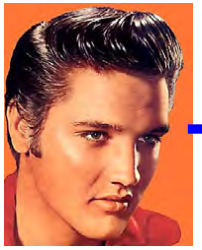
AMIE

(5min)

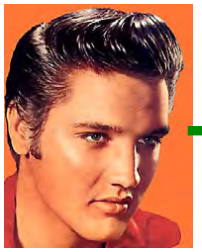
$$type(x, pope) \Rightarrow$$
$$diedIn(x, Rome)$$

[WWW 2013, VLDB journal 2015]

Matching heterogeneous KBs

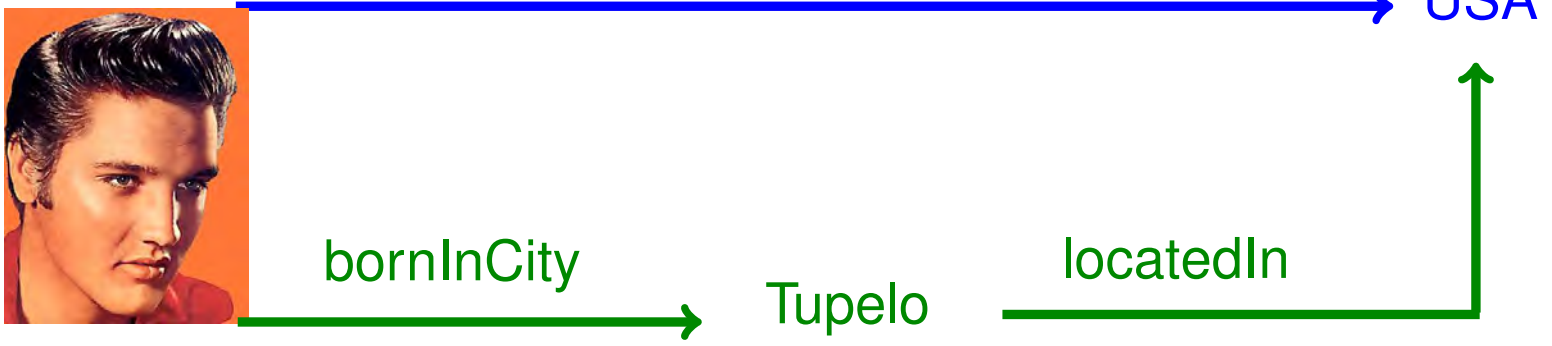


bornInCountry → USA



bornInCity → Tupelo → locatedIn → USA

1. Coalesce the KBs



2. Mine rules

bornInCountry

USA

bornInCity

Tupelo

locatedIn

AMIE

$$bornInCity(x, y) \wedge locatedIn(y, z) \not\Rightarrow bornInCountry(x, z)$$

"ROSA rule"

Caveat:
Spurious
correlations

ROSA rules match ontologies

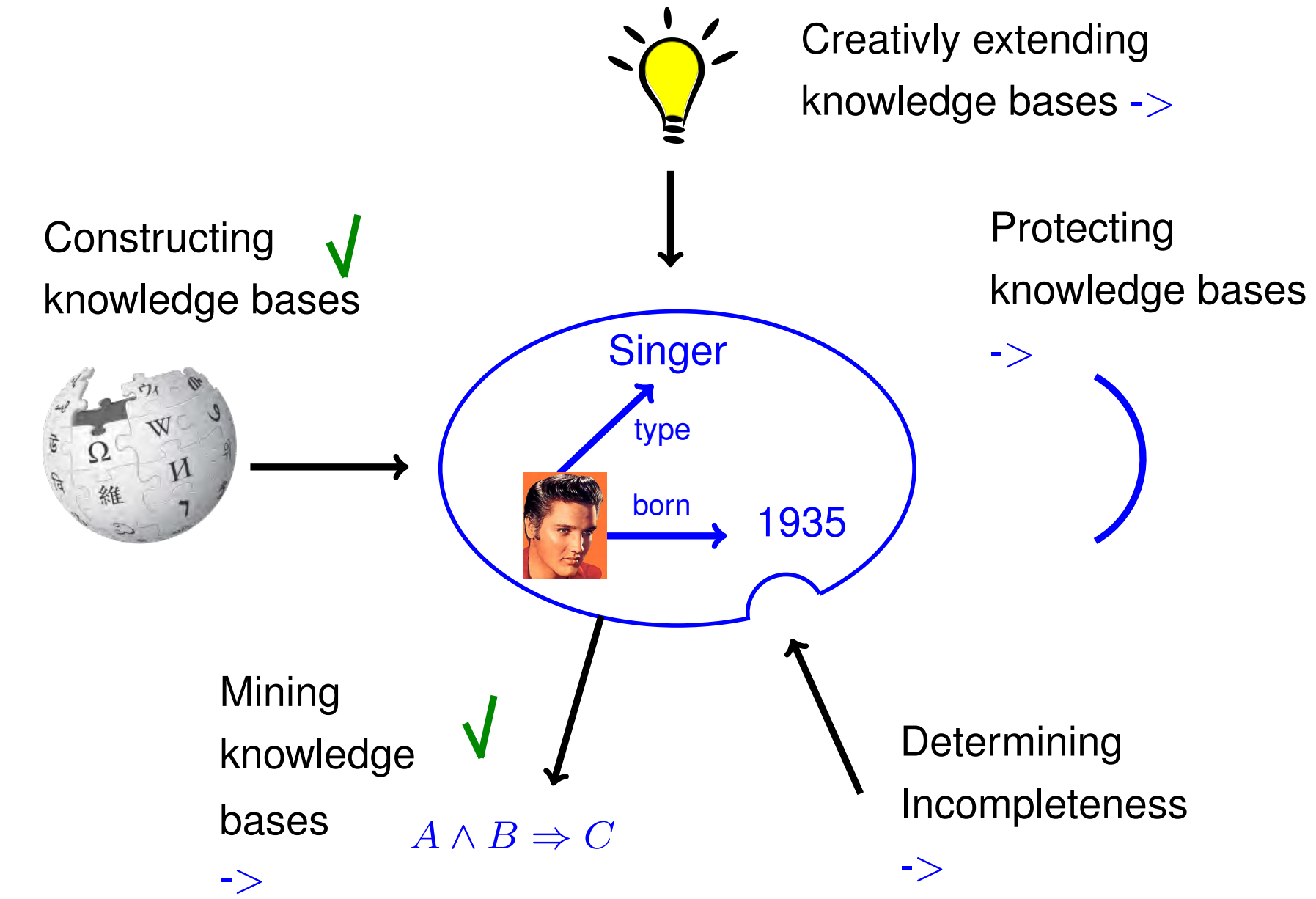


[AKBC 2013]

$$bornInCity(x, y) \wedge locatedIn(y, z) \nRightarrow bornInCountry(x, z)$$

"ROSA rule"

Knowledge Base Life Cycle



Incompleteness

marriedTo

√

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

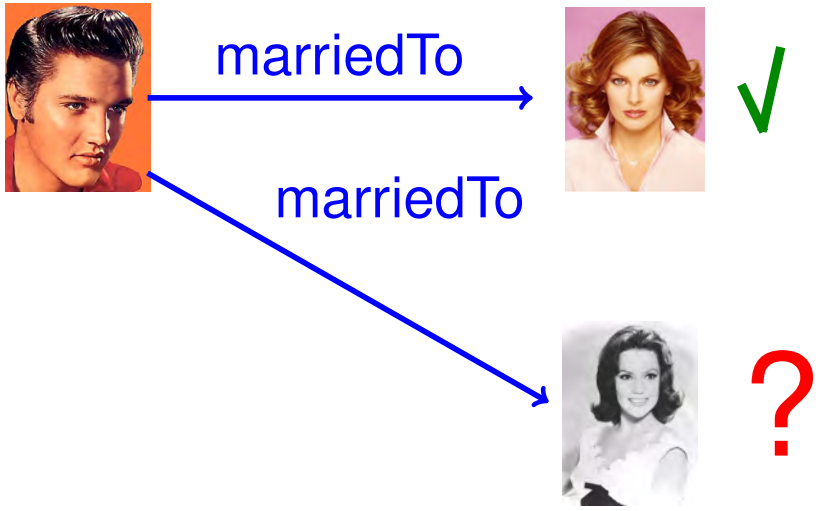
Incompleteness

marriedTo

√

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

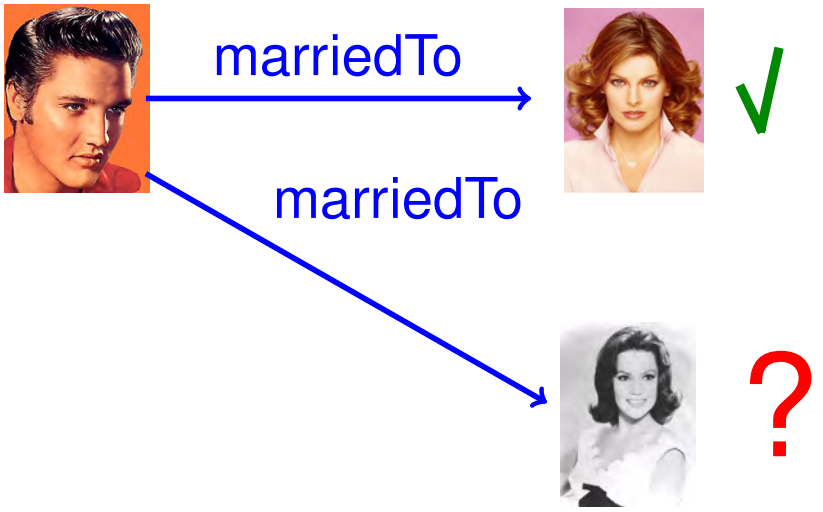
Incompleteness



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 $r(s, o)$?

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

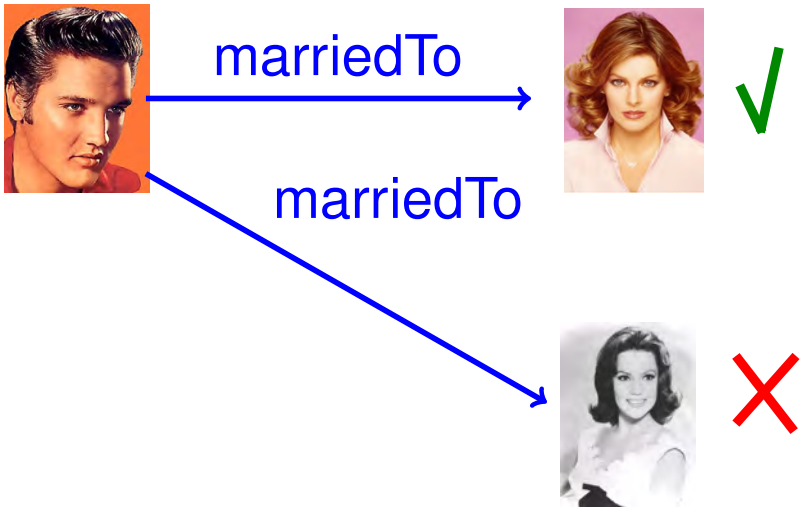
Signals for Incompleteness (F1)

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

Signals for Incompleteness



AMIE can predict incompleteness

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

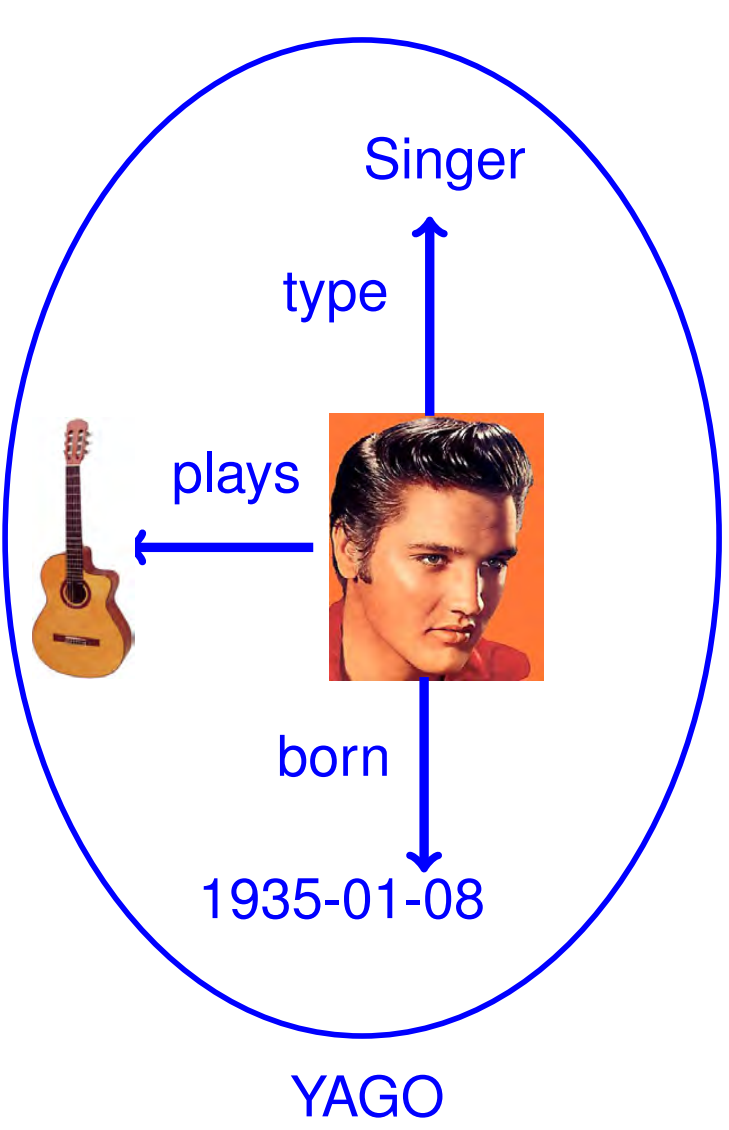
... and more.

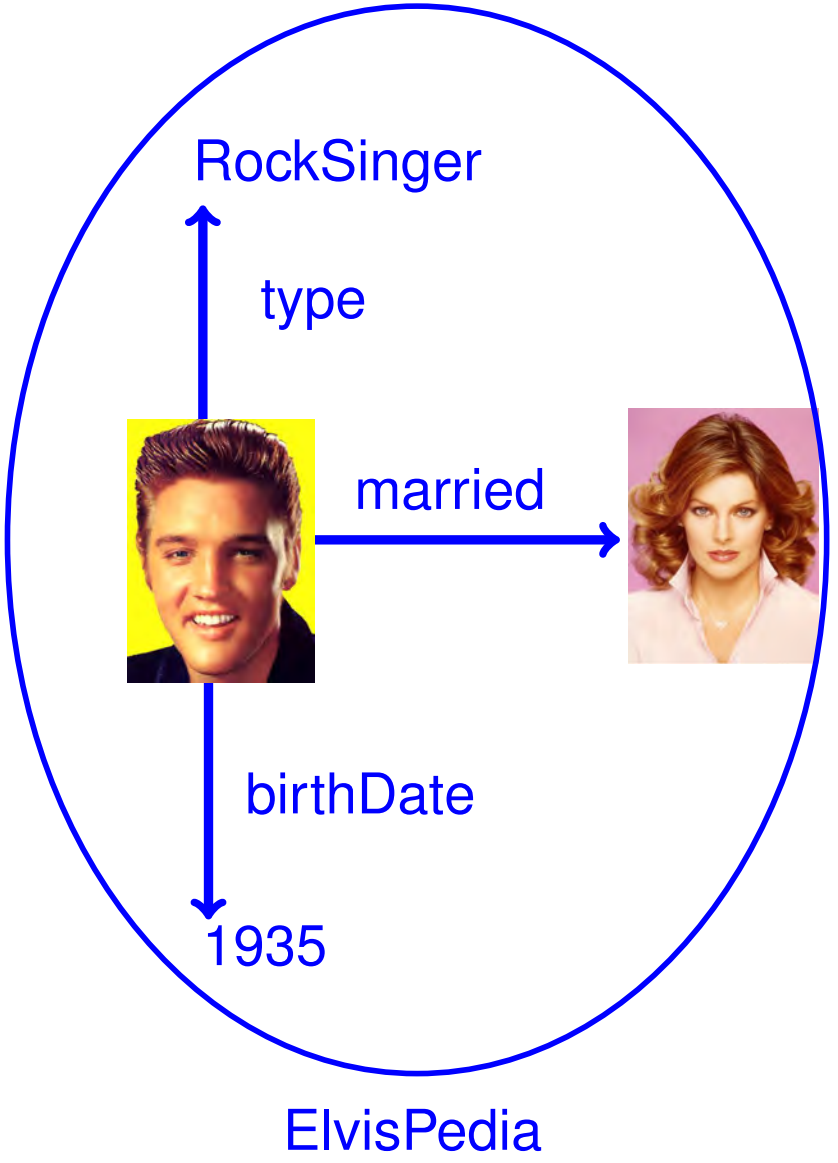



[WSDM 2017]

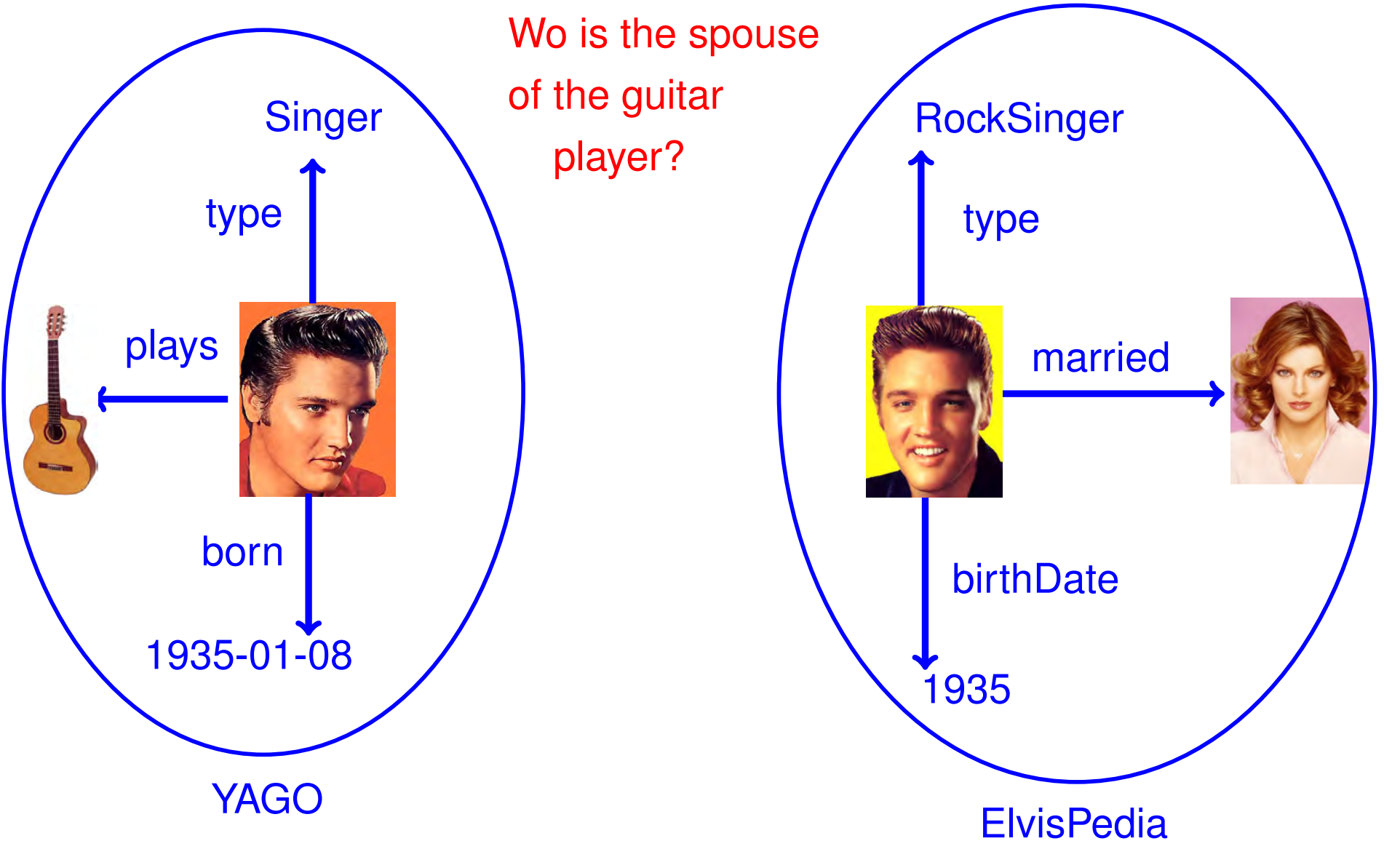
>paris

If incomplete, use another KB

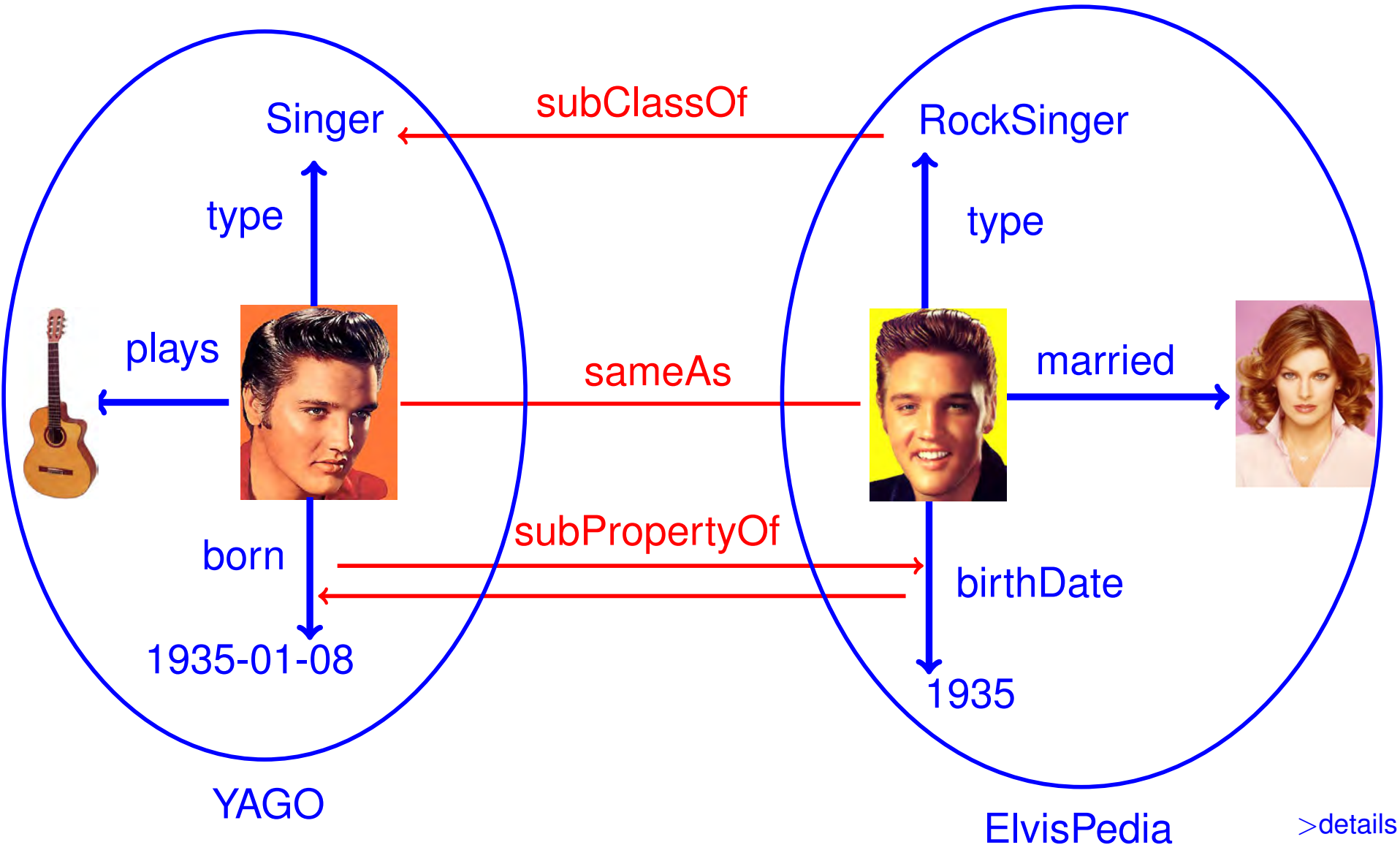




No Links => No Use

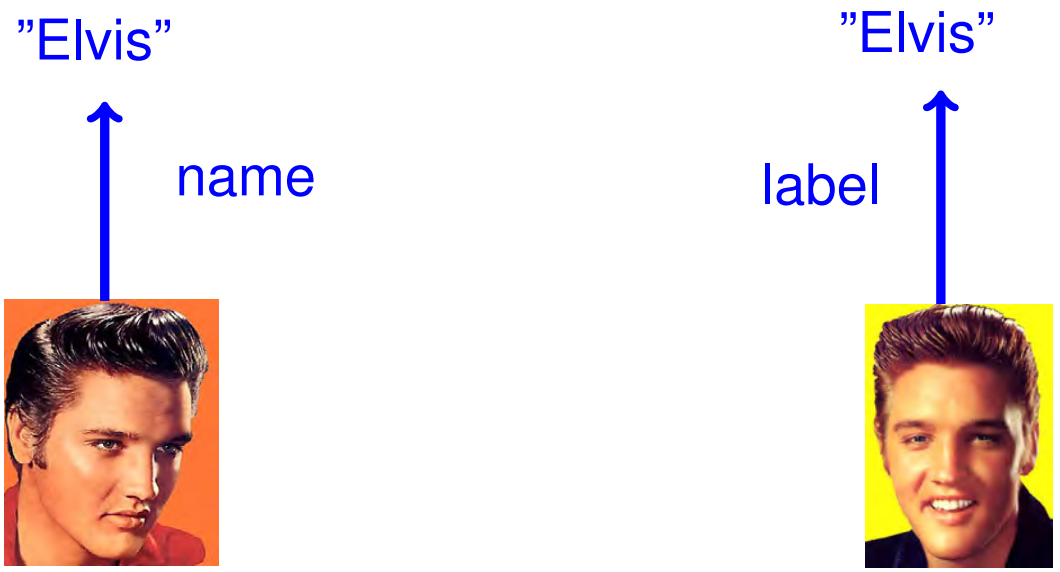


Match classes, entities, & relations

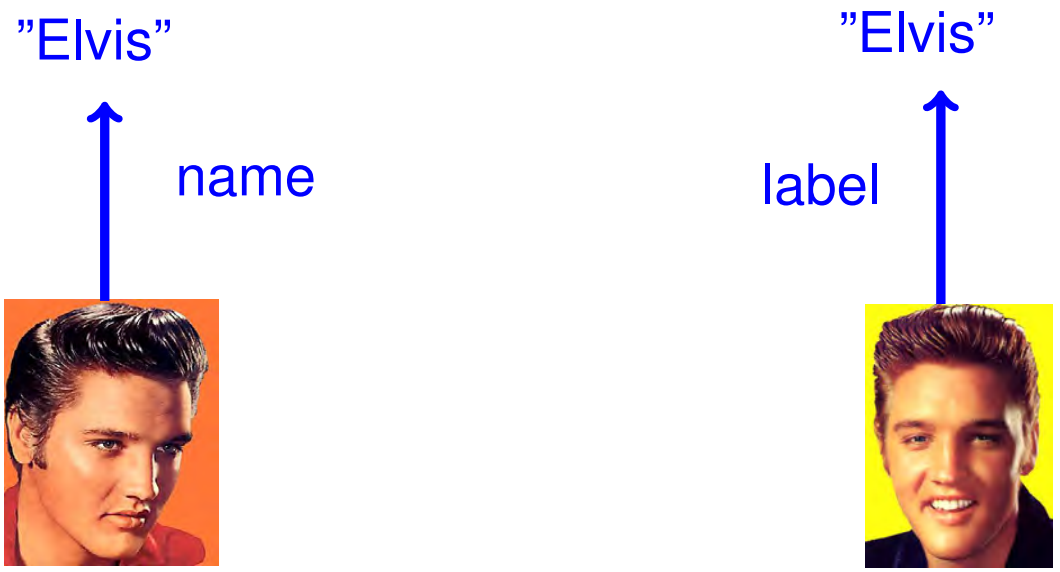


>details

Match classes, entities, & relations

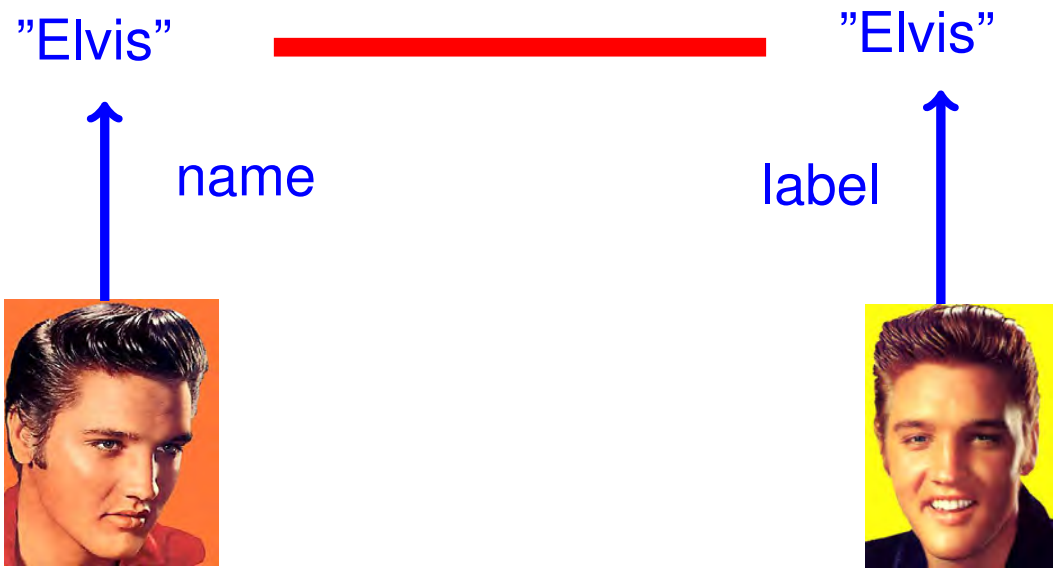


Match classes, entities, & relations



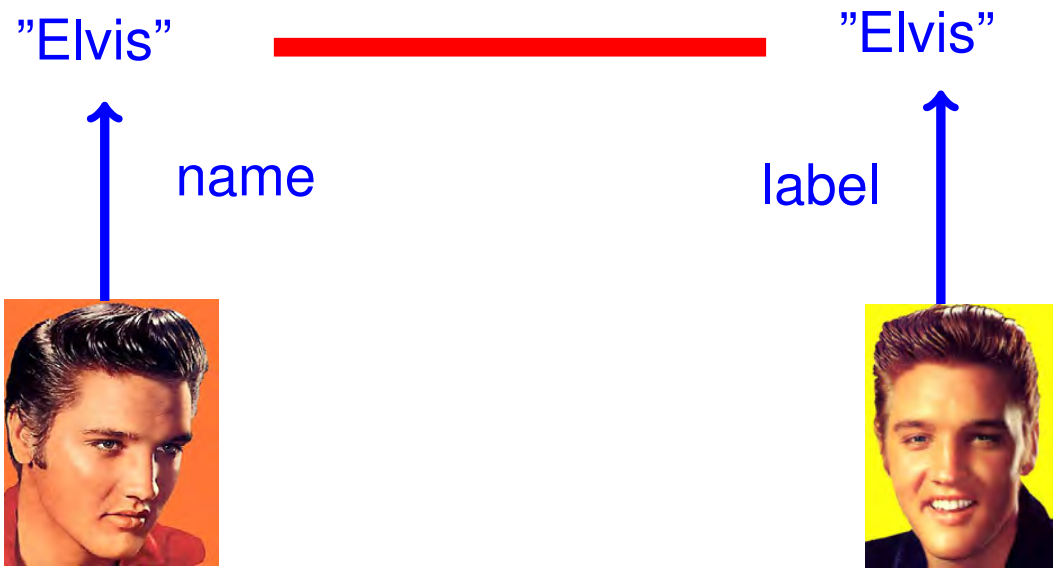
1. Match literals

Match classes, entities, & relations



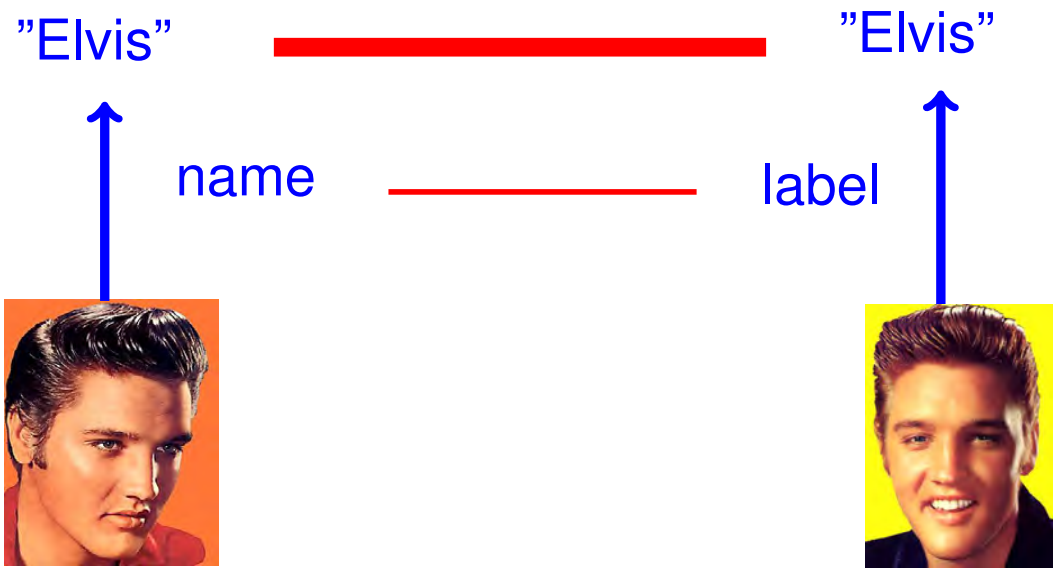
1. Match literals

Match classes, entities, & relations



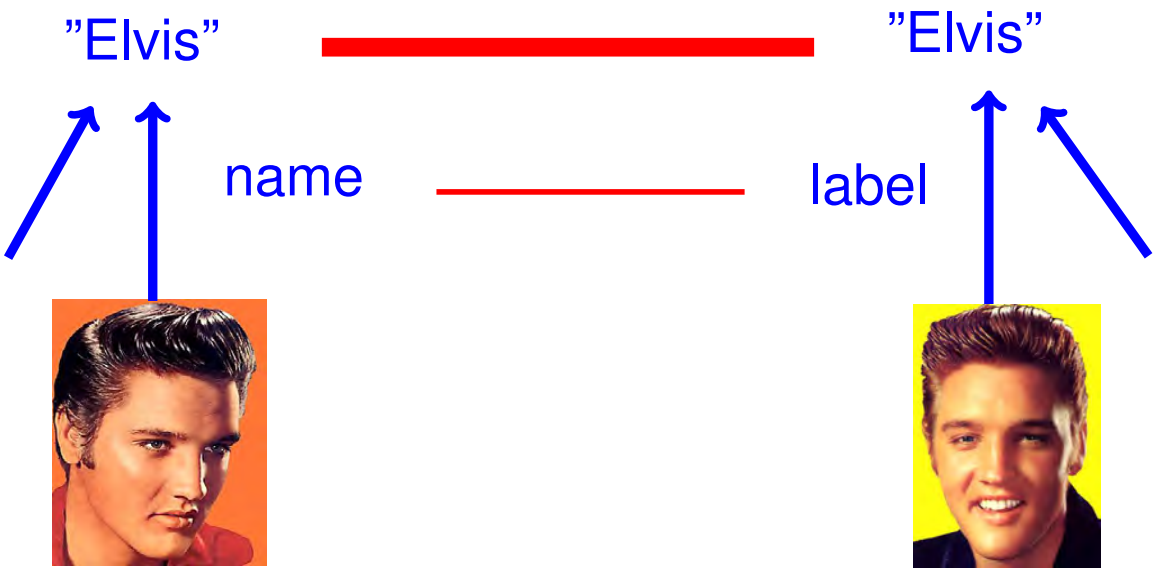
2. Assume small equivalence of all relations

Match classes, entities, & relations



2. Assume small equivalence of all relations

Match classes, entities, & relations

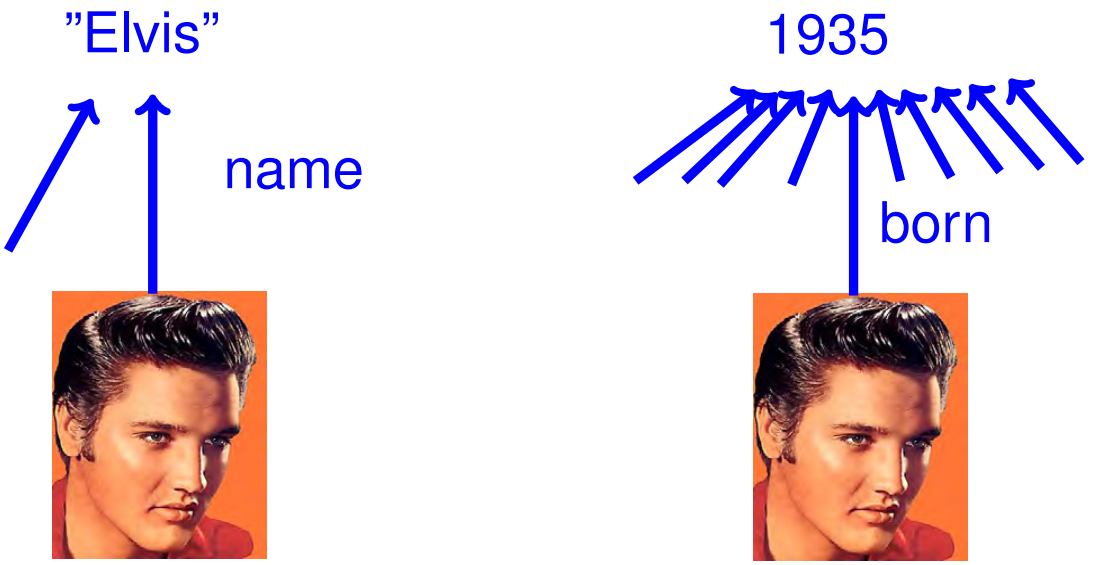


What about matching the entities?

What does it mean that both Elvises share the same name?

>fun

Inverse Functionality



$$ifun(r, y) = \frac{1}{\#x{:}r(x,y)}$$

ifun(name, Elvis)= 1 / 2

ifun(born, 1935)= 1/10

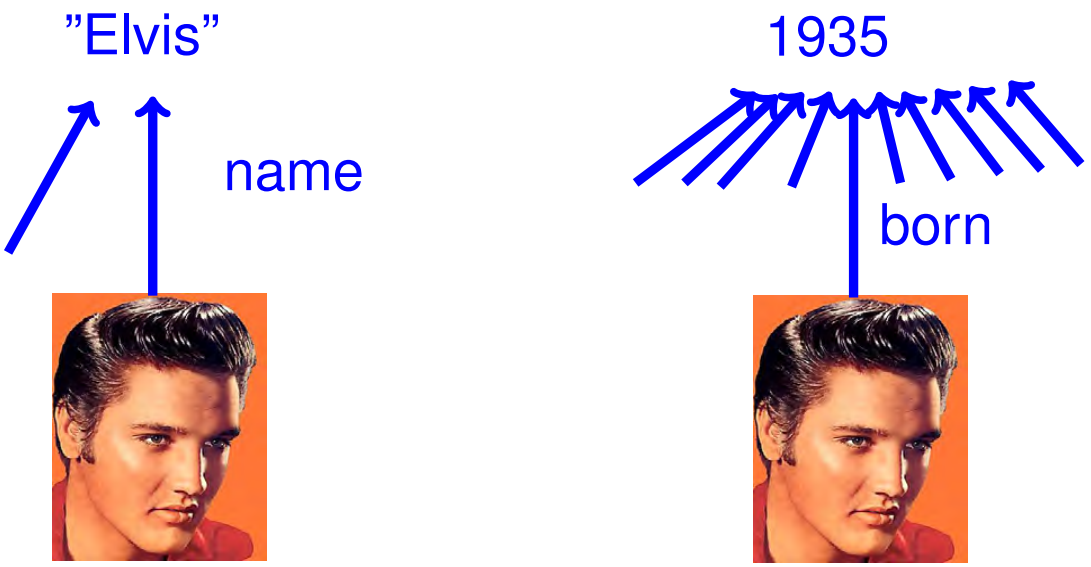
$$ifun(r) = HM_y ifun(r, y)$$

ifun(name)= 0.9

ifun(born)= 0.1

>fun

Inverse Functionality

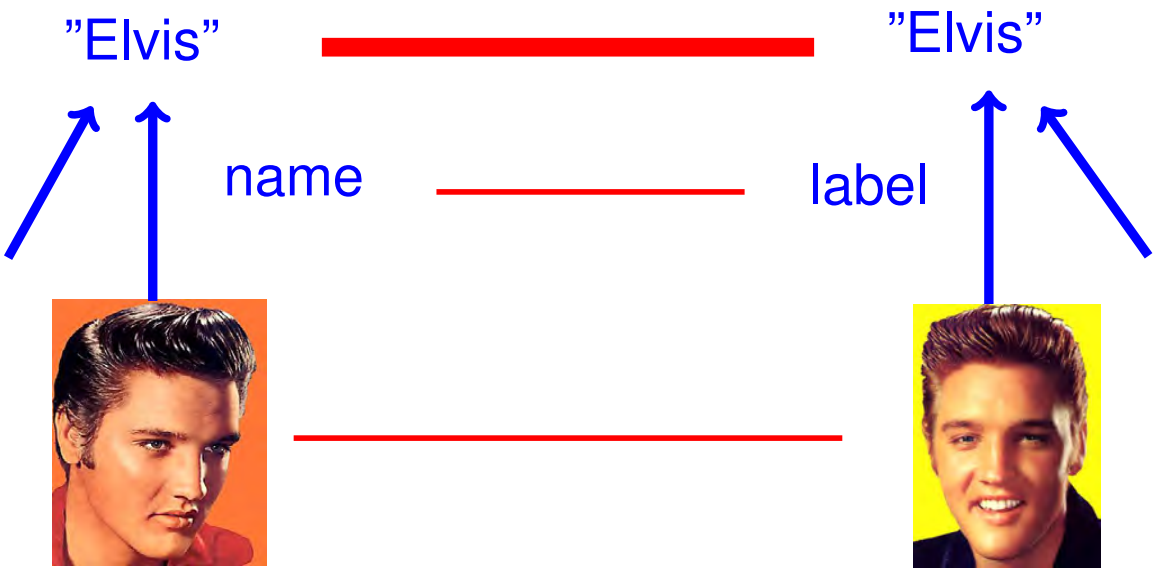


$$ifun(r) = \frac{\#x{:}\exists y{:}r(x,y)}{\#x,y{:}r(x,y)}$$

# of subjects
divided by
# of facts

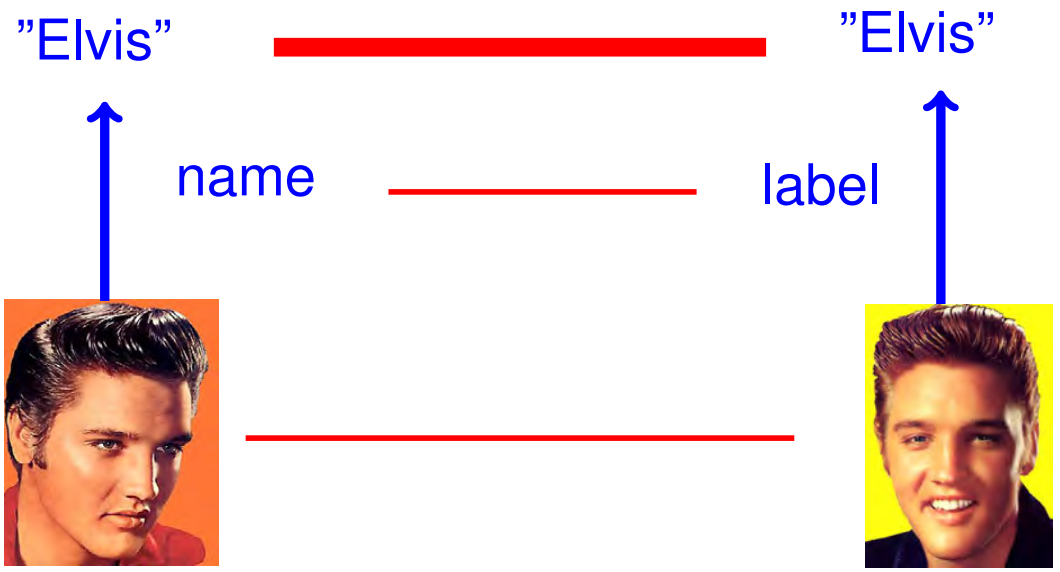
$$ifun(r) = HM_y ifun(r,y)$$

Match classes, entities, & relations



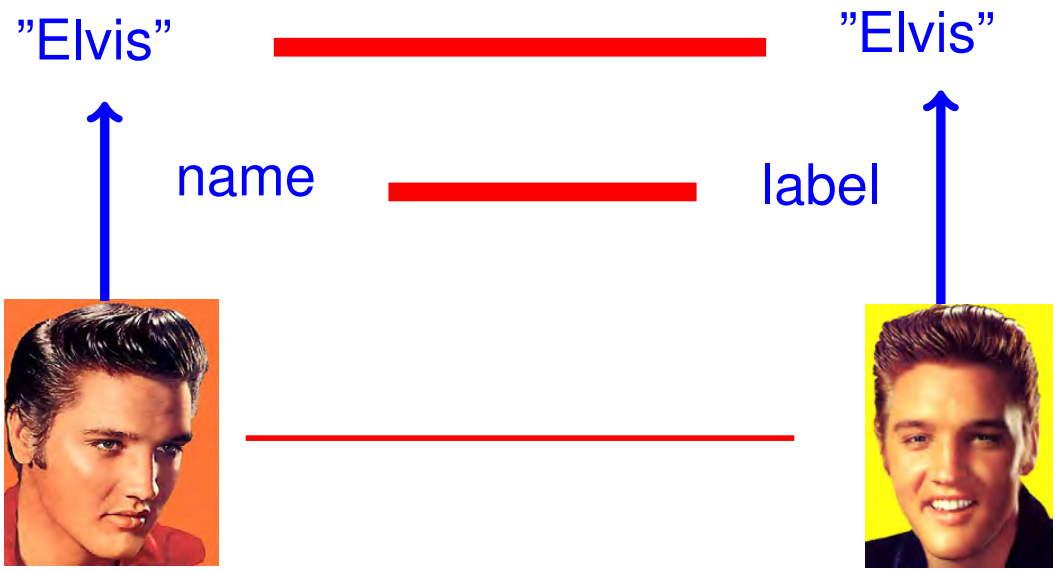
3. If subjects share a relation that is highly inverse functional, and the object is matched, then match the subjects.

Match classes, entities, & relations



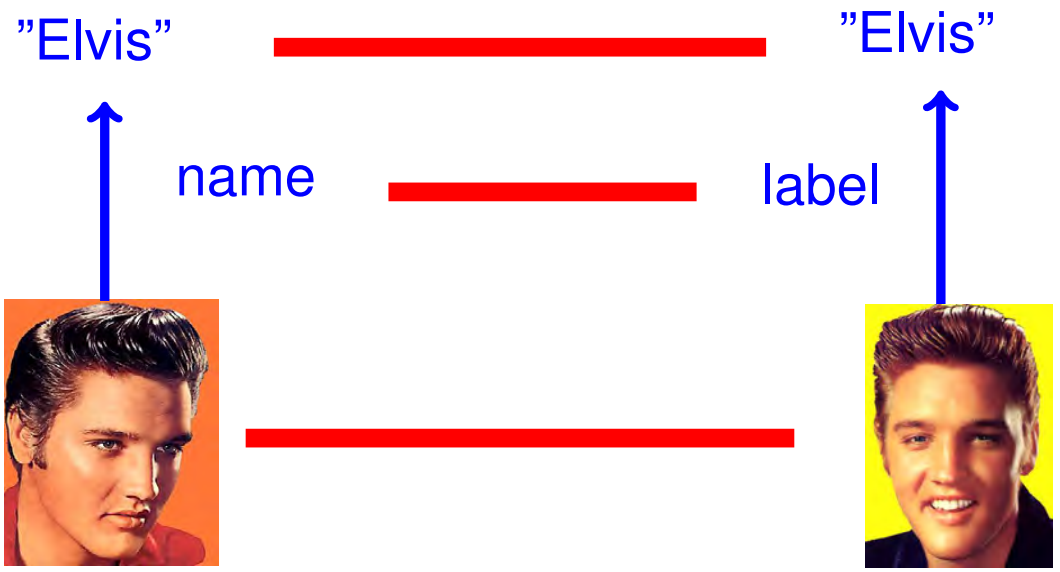
4. If relations share many pairs,
   increase their match

Match classes, entities, & relations



4. If relations share many pairs,
   increase their match

Match classes, entities, & relations



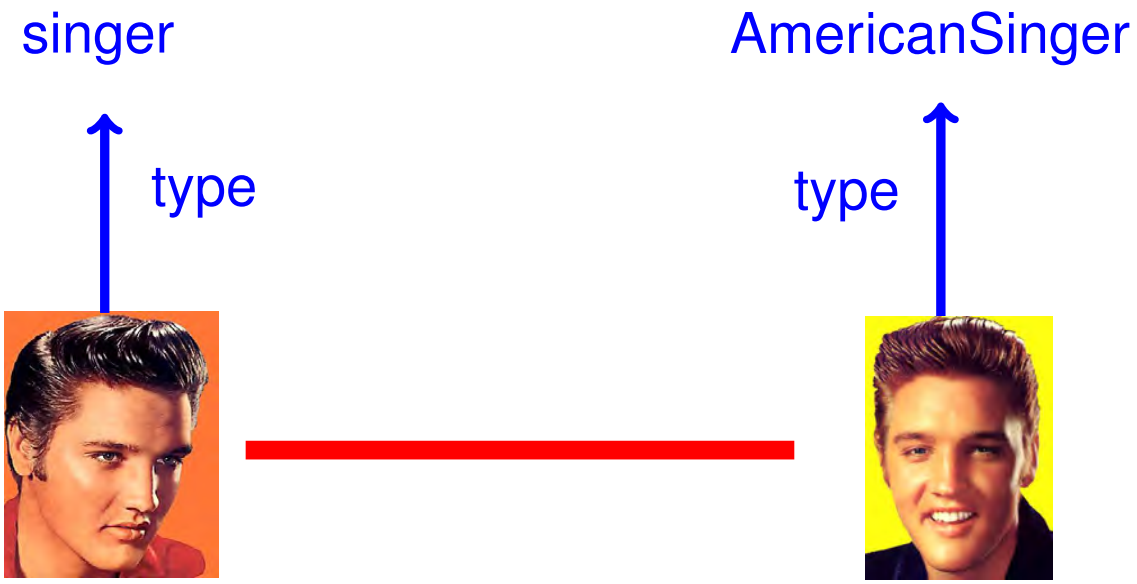
5. Iterate

$$P(e_1 \equiv e_2) = \prod_{42}^{1} \alpha^{\beta} ... P(r_1 \subseteq r_2) ...$$

$$P(r_1 \subseteq r_2) = 42\phi^{\gamma} \sum ... P(e_1 = e_2) ...$$

Caveat: Convergence proof only in subcase

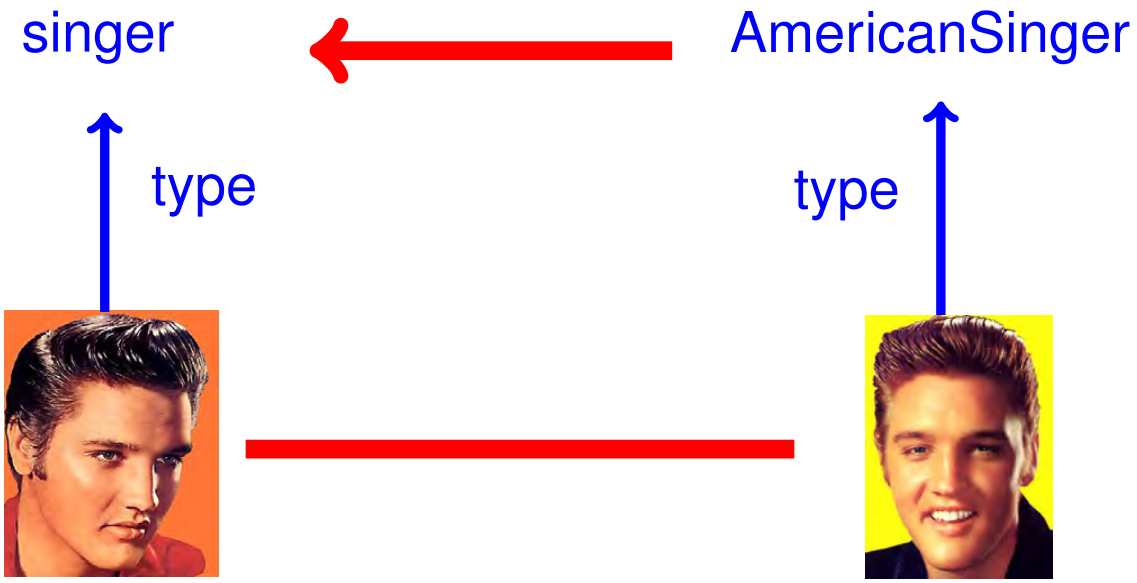
Match classes, entities, & relations



6. Compute class subsumption

$$P(c_1 \subseteq c_2) = arcsin(4.1125) \times P(e_1 \equiv e_2) \times ...$$

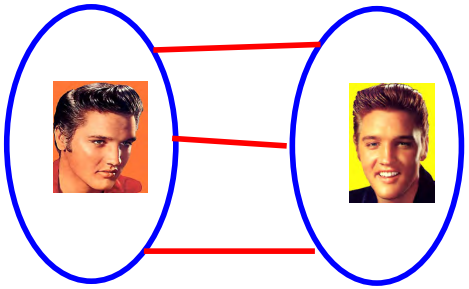
Match classes, entities, & relations



6. Compute class subsumption

Caveat: Class alignment works with “only” 85% precision due to spurious correlation.

PARIS:match entities,classes,relations



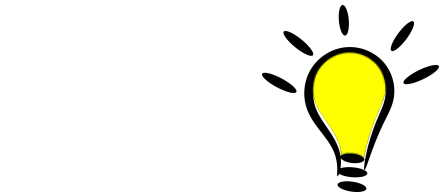
PARIS matches DBpedia & YAGO

- in 2 hours
- with 90% accuracy

[VLDB 2012, APWeb 2014 invited paper]



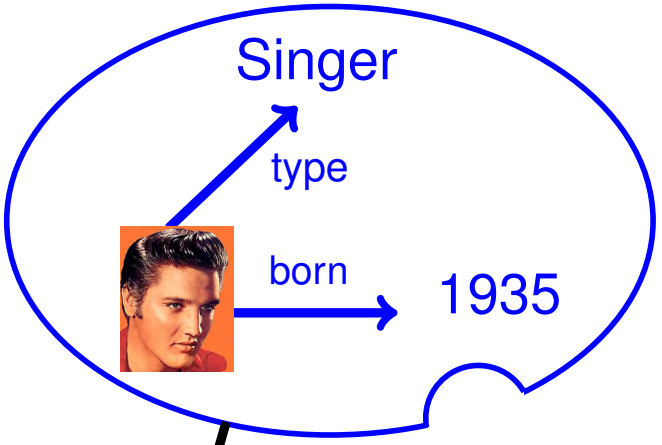

Knowledge Base Life Cycle



Creativly extending knowledge bases ->

Constructing knowledge bases ✓

Protecting knowledge bases ->



Mining knowledge bases ✓ ->

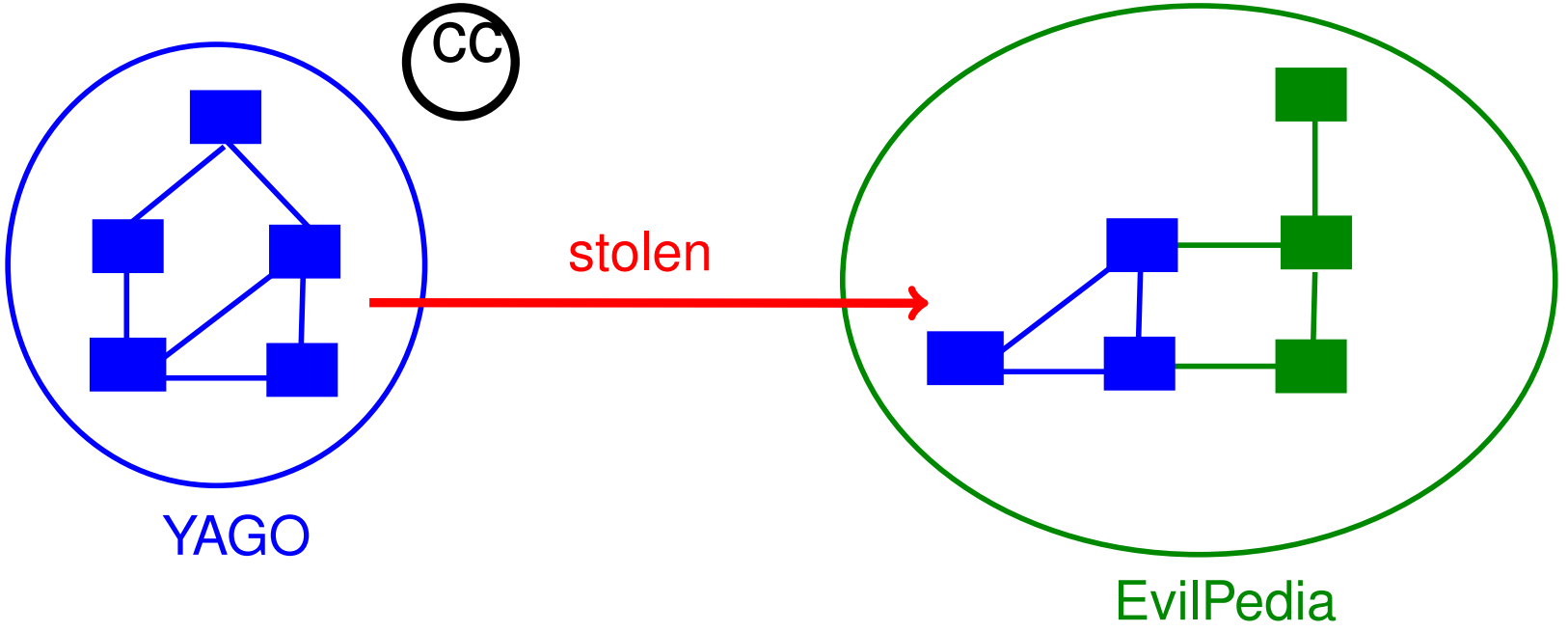
$A \wedge B \Rightarrow C$

Determining Incompleteness ✓ ->

Plagiarism

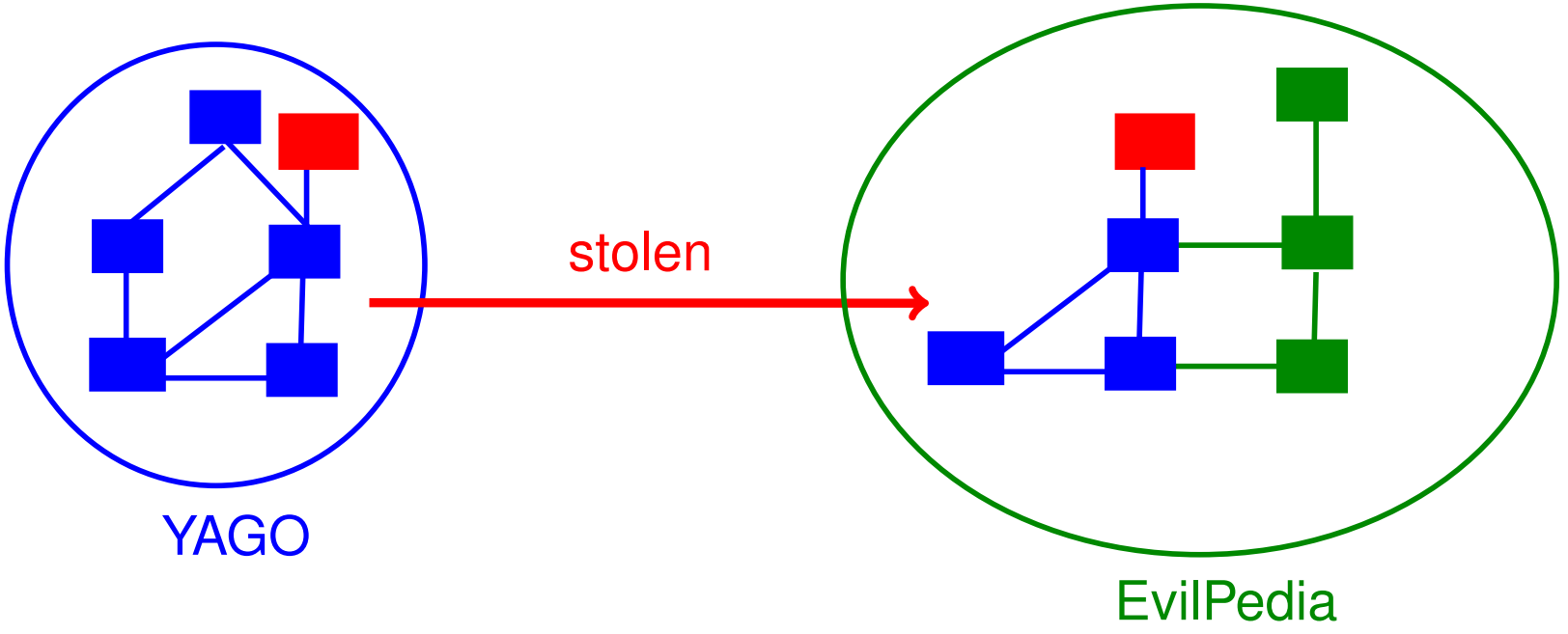
People may “steal” from other ontologies without giving due credit.

Most ontologies have licenses that require attribution.



Additive Watermarking

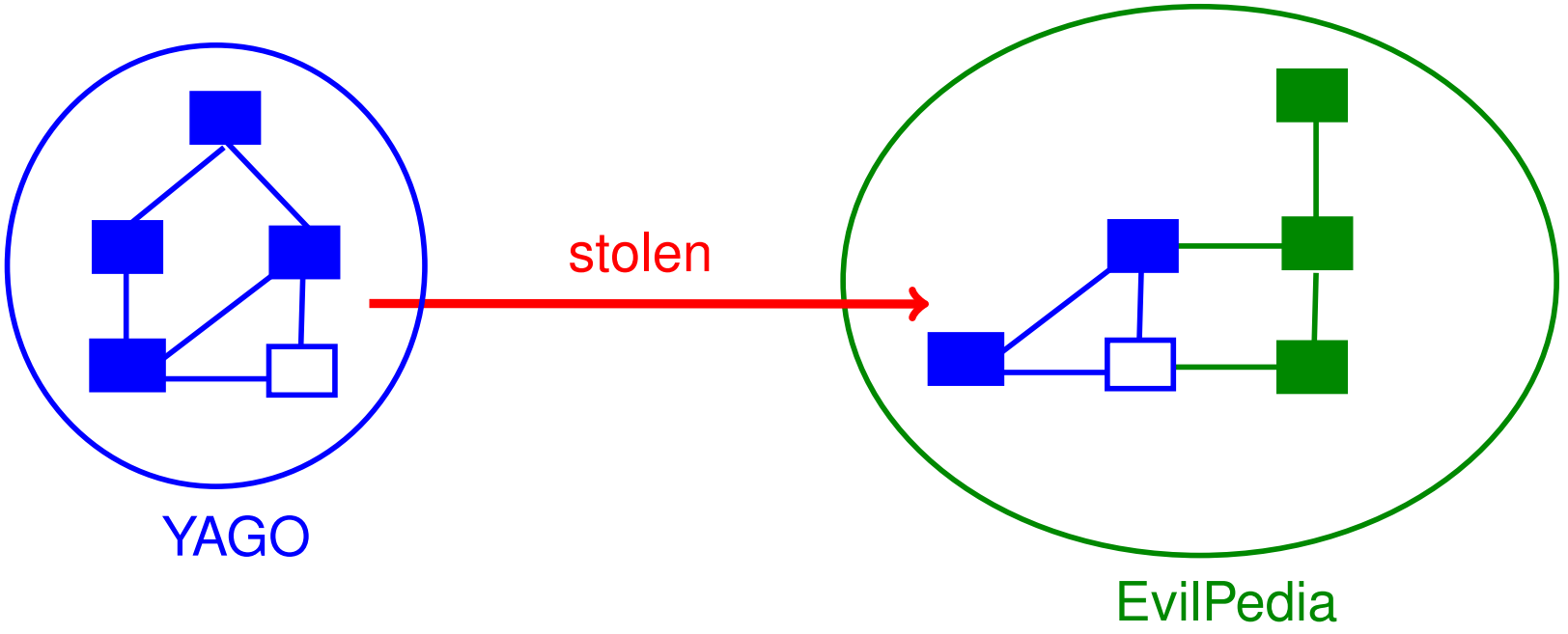
By adding a few fake facts to the source ontology,
one can prove theft in the target ontology.



[WWW 2012 demo]

Subtractive Watermarking

One can also prove theft by selectively removing facts from the source ontology.

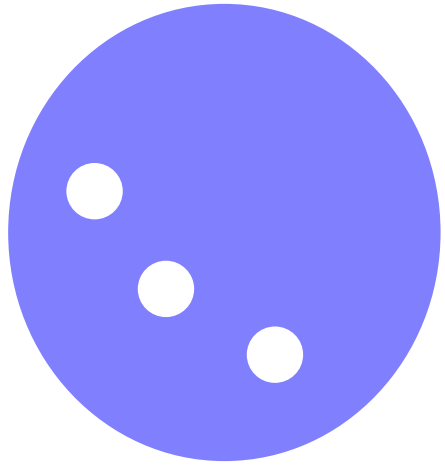


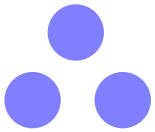


[ISWC 2011]

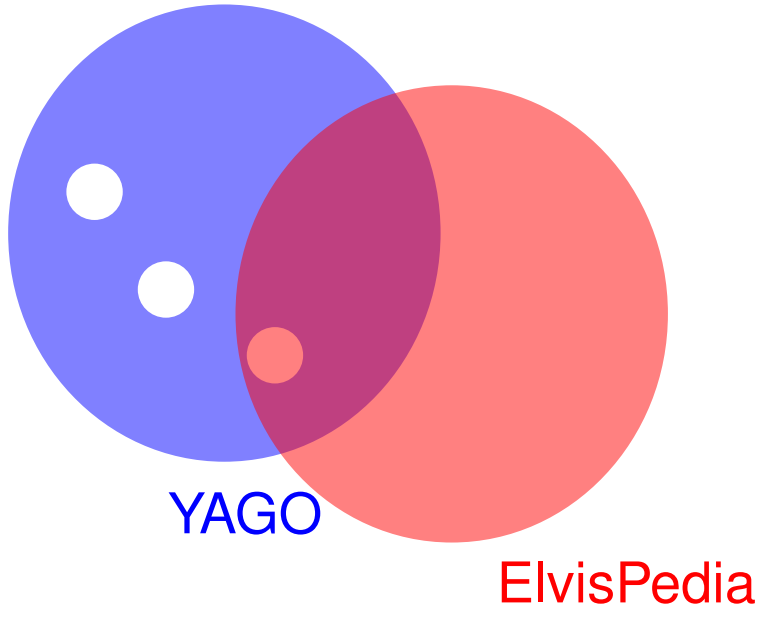
details>76
details&DIVINA>99

Subtractive Watermarking





An Innocent KB

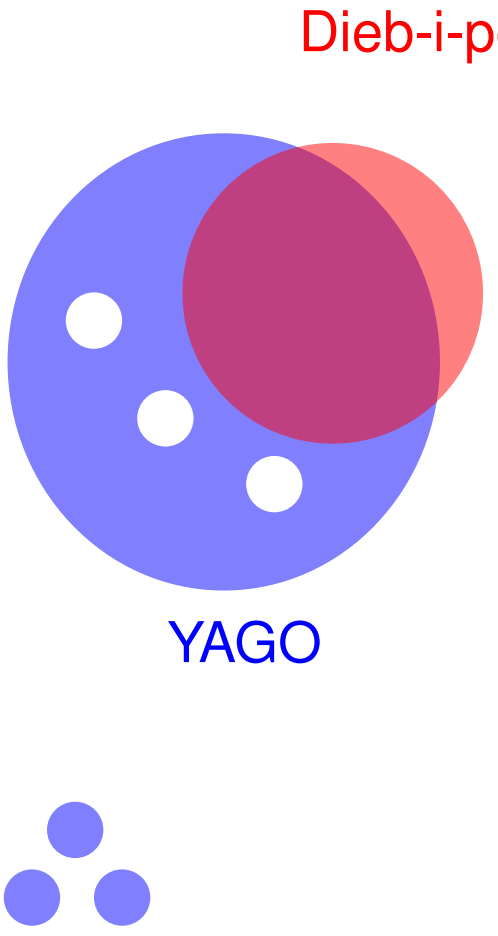


ratio of overlap

≈

ratio of covered holes

A Stolen KB



high overlap

but

no holes covered

Probability that this happens by chance

$$\sim (1-\alpha)^n$$

α : overlap ratio

n : number of holes

(Argument by Mauro Sozio; original paper uses chi square test)

Protecting the user

Mat Honan

amazon

Protecting the user

Mat Honan

amazon

Protecting the user

Mat Honan

Protecting the user

Mat Honan

reset
password
by credit card

reset
password
by backup email

Protecting the user

Mat Honan

reset
password
by credit card

reset
password
by backup email

reset
password
by backup email

Protecting the user

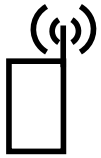
Mat Honan

reset
password
by credit card

reset
password
by backup email

reset
password
by backup email

Two-factor
authentication



Protecting the user

Mat Honan

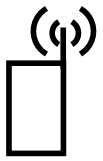
amazon

reset password by credit card

reset password by backup email

reset password by backup email

Two-factor authentication

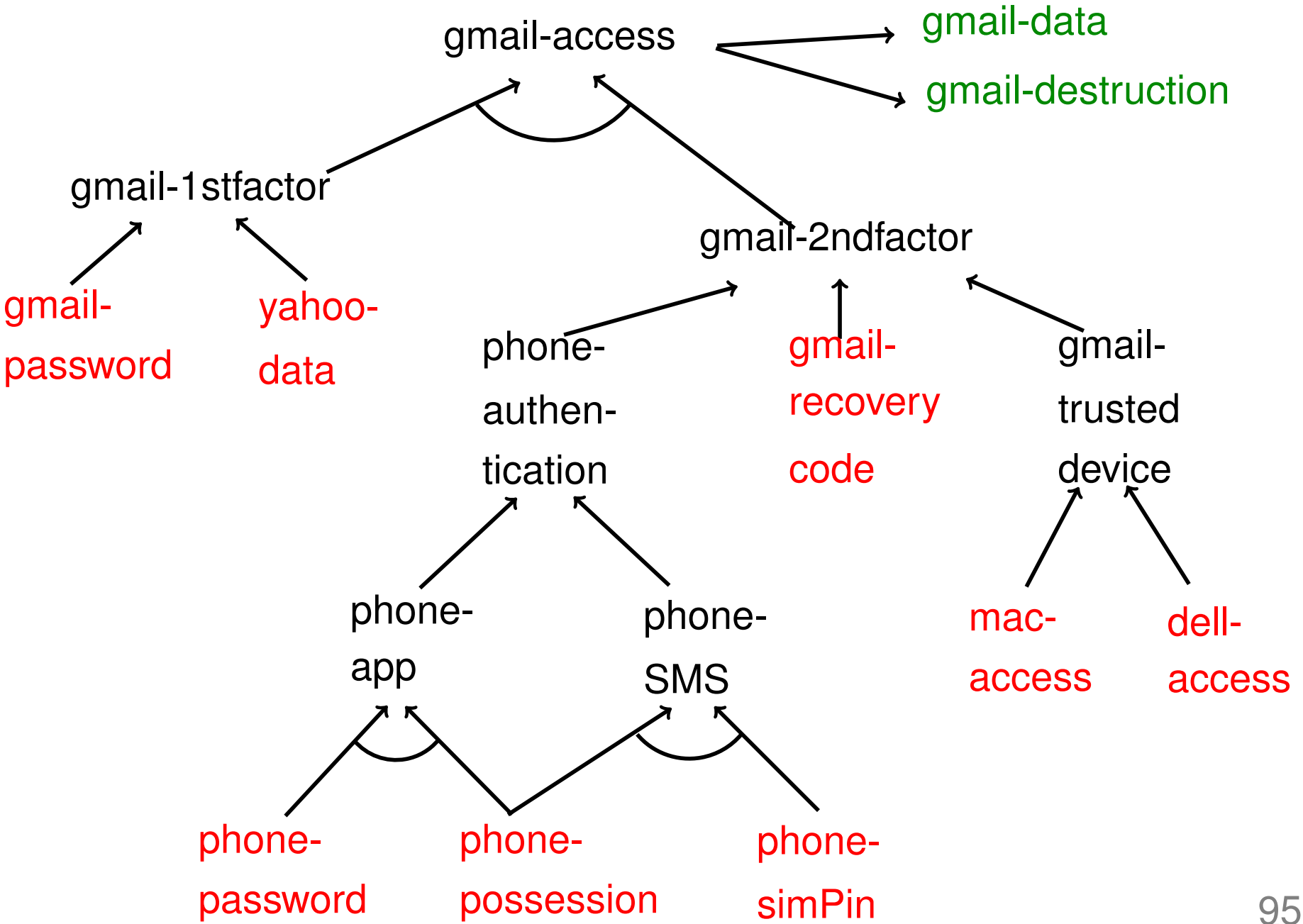


Security: How many keys does a hacker need?

Safety: How many keys can I afford to lose?

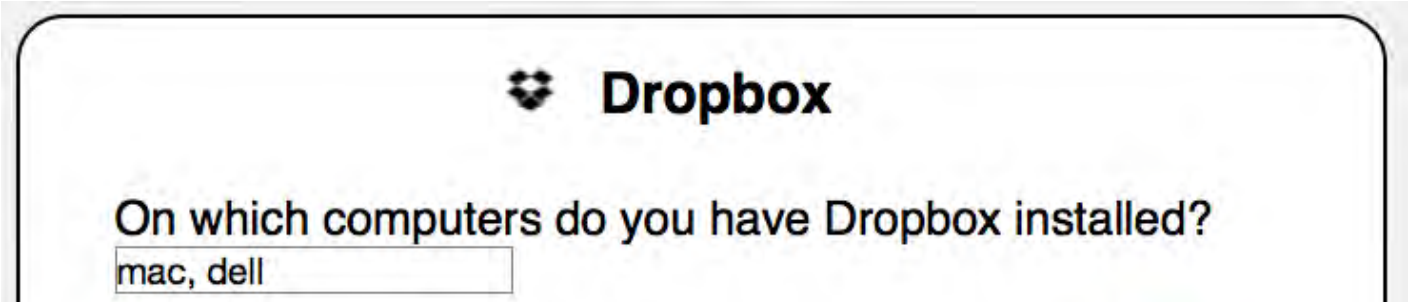
Protection: How many keys to destroy my data?

Account Dependencies



DIVINA

Discovering Vulnerabilities of Internet Accounts



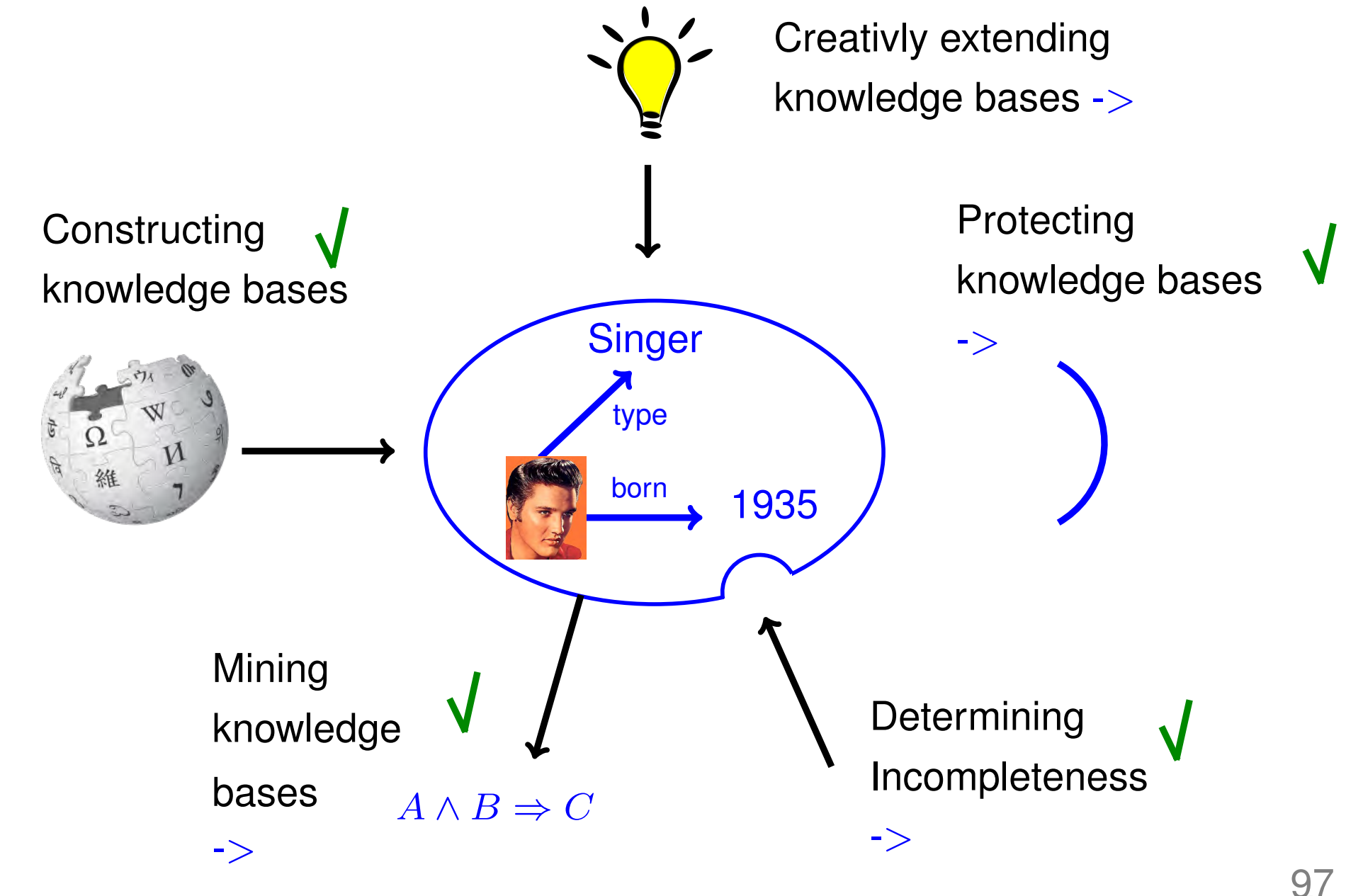
	Safety	Security	Protection
	✓	✓	✗

For online accounts: Consider two-factor

[WWW 2015 demo]

https://suchanek.name/programs/divina

Knowledge Base Life Cycle



Combinatorial Creativity

+ ( - ) +

© Stone Mens Wear © BetterThanPants.com © Vileda © Aysar Aras

Description Logics do not work

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

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

+ (

-

) +

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

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

+ (

-

) +

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


TECH

The 25 Best Inventions

TIME Staff · Nov. 19, 2015

2) Generative experiments

1/3 nonsense, 1/3 exists,

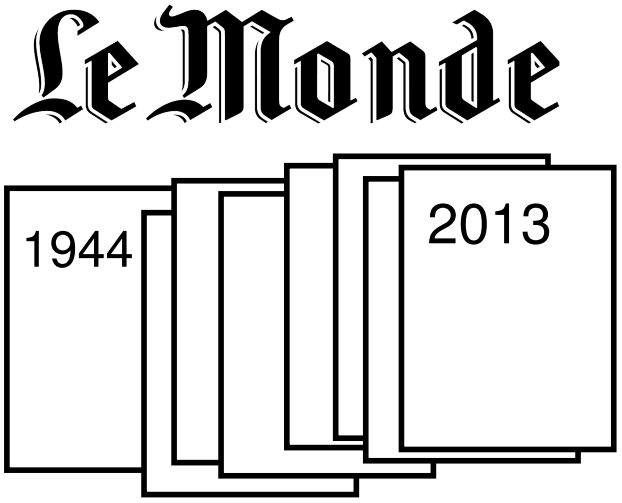
1/3 “imaginable”




[ISWC 2016
paper & demo]

>LeMonde

Another creative idea...



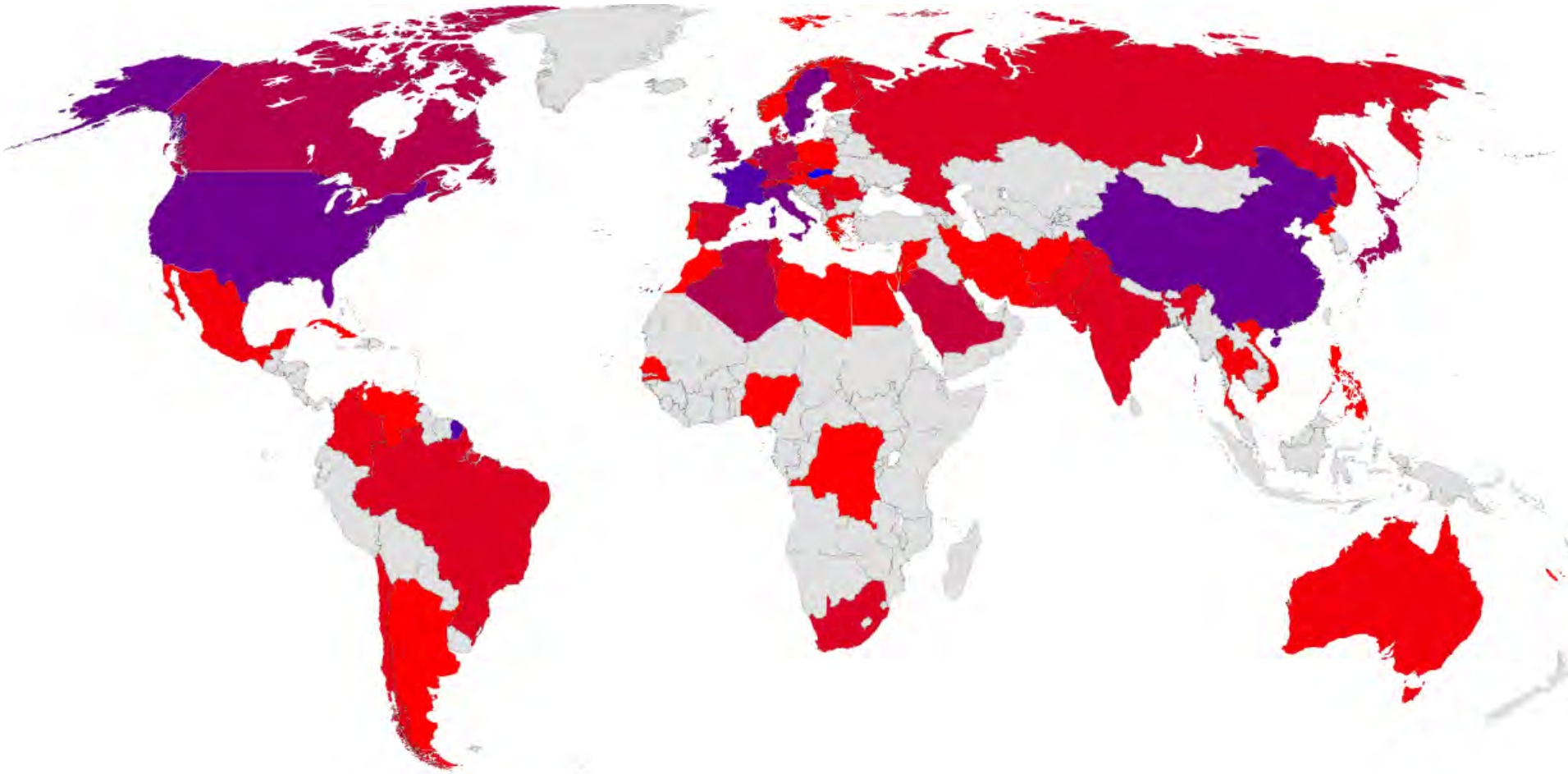
Mining Le Monde

Le Monde

time	place	entity
1967	USA	

yago
select knowledge

Presence of foreign companies



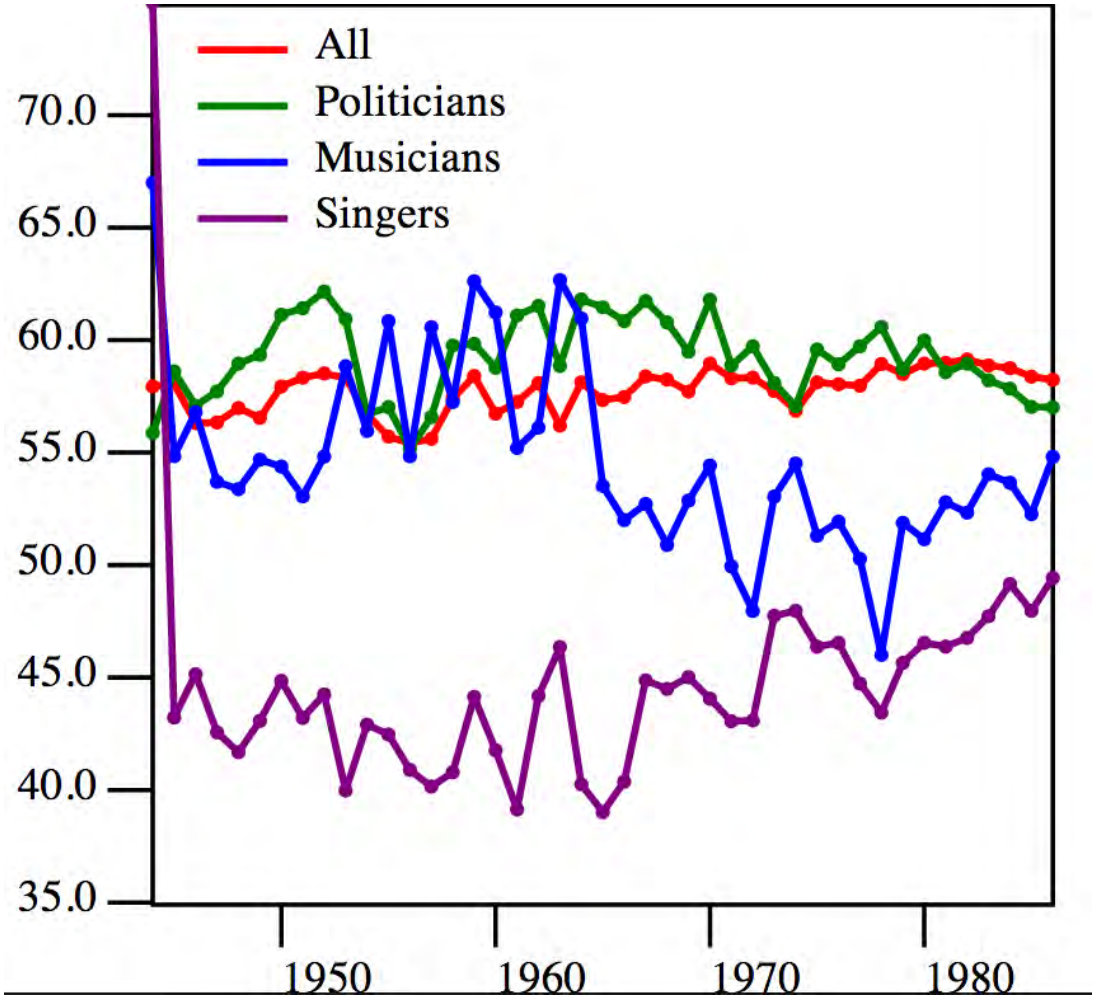
red: many foreign companies mentioned

blue: few foreign companies mentioned

Average age of people mentioned

Le Monde

yago
select knowledge



[AKBC 2013]

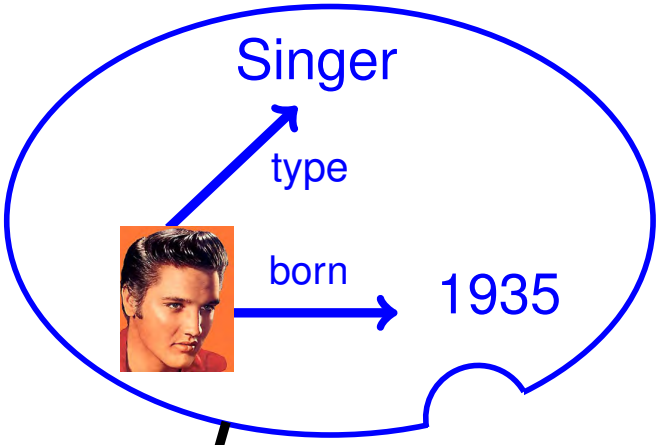
[VLDB 2014 vision]

Knowledge Base Life Cycle

Creativly extending knowledge bases ->

√

Constructing knowledge bases √



Protecting knowledge bases ->

√

Mining knowledge bases ->

√

$A \wedge B \Rightarrow C$

Determining Incompleteness ->

√

Is Elvis dead?

Elvis Presley

Singer

Elvis Aaron Presley was an American singer 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

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

Spouse: Priscilla Presley (m. 1967–1973)

Children: Lisa Marie Presley

???

Elvis —died→ 1977

yago
select knowledge

Is Elvis dead?

Elvis Presley

Singer

Elvis Aaron Presley was an American singer 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

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

Spouse: Priscilla Presley (m. 1967–1973)

Children: Lisa Marie Presley

???

Elvis —died→ 1977

yago
select knowledge

100m statements

95% precision

-> 5m wrong statements

Knowledge Base Life Cycle