Atelier VERITA : VERacity In daTA
22 Janvier 2019, Metz

DATA VERACITY ASSESSMENT: HOW A-PRIORI KNOWLEDGE CAN IMPROVE TRUTH DISCOVERY MODELS

Valentina Beretta

IMT Mines Alès
École Mines-Télécom

Espace Dev

Data veracity assessment: Enhancing Truth Discovery using *a priori* knowledge

Outline:

1. Motivations behind data veracity assessment
2. Truth Discovery: problem and positioning
3. Enhancing Truth Discovery models using *a priori* knowledge
4. Conclusion

Data veracity assessment: Enhancing Truth Discovery using *a priori* knowledge

Outline:

1. Motivations behind data veracity assessment

1.1 Why studying data veracity?

1.2 What are the tasks that can benefit from data veracity?

2. Truth Discovery: problem and positioning

3. Enhancing Truth Discovery models using *a priori* knowledge

4. Conclusion

Why studying data veracity?

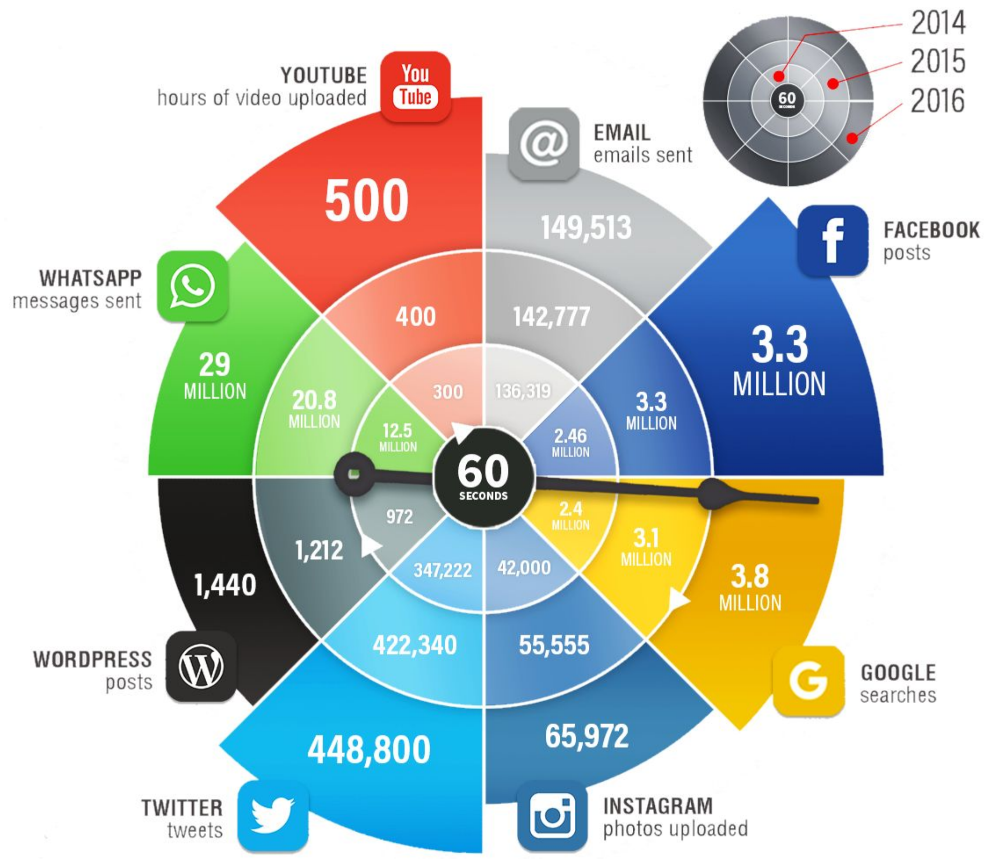


Figure 1: This infographic shows what happens online each minute (source https://www.smartinsights.com/wp-content/uploads/2016/08/What-happens-online-in-60-seconds.png)

The amount of information published on the Web increases over the years

Why studying data veracity?

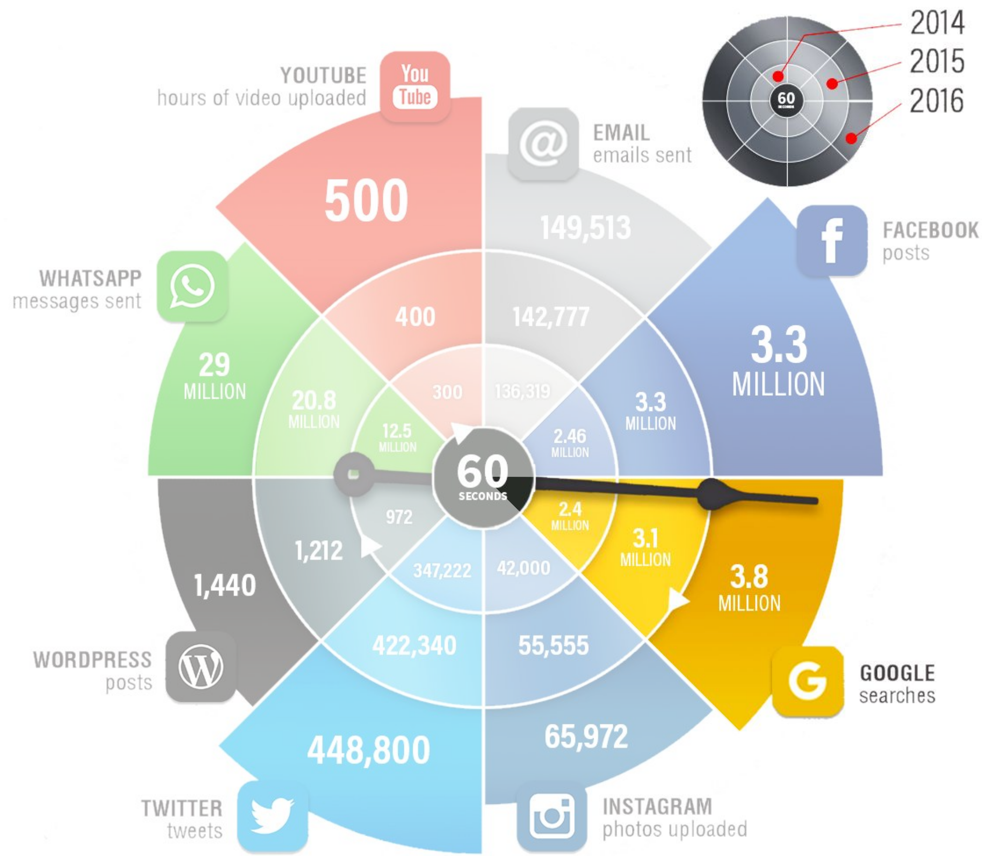


Figure 1: This infographic shows what happens online each minute (source https://www.smartinsights.com/wp-content/uploads/2016/08/What-happens-online-in-60-seconds.png)

The amount of information published on the Web increases over the years

At the same time, the number of searches to consume this information grows

Why studying data veracity?

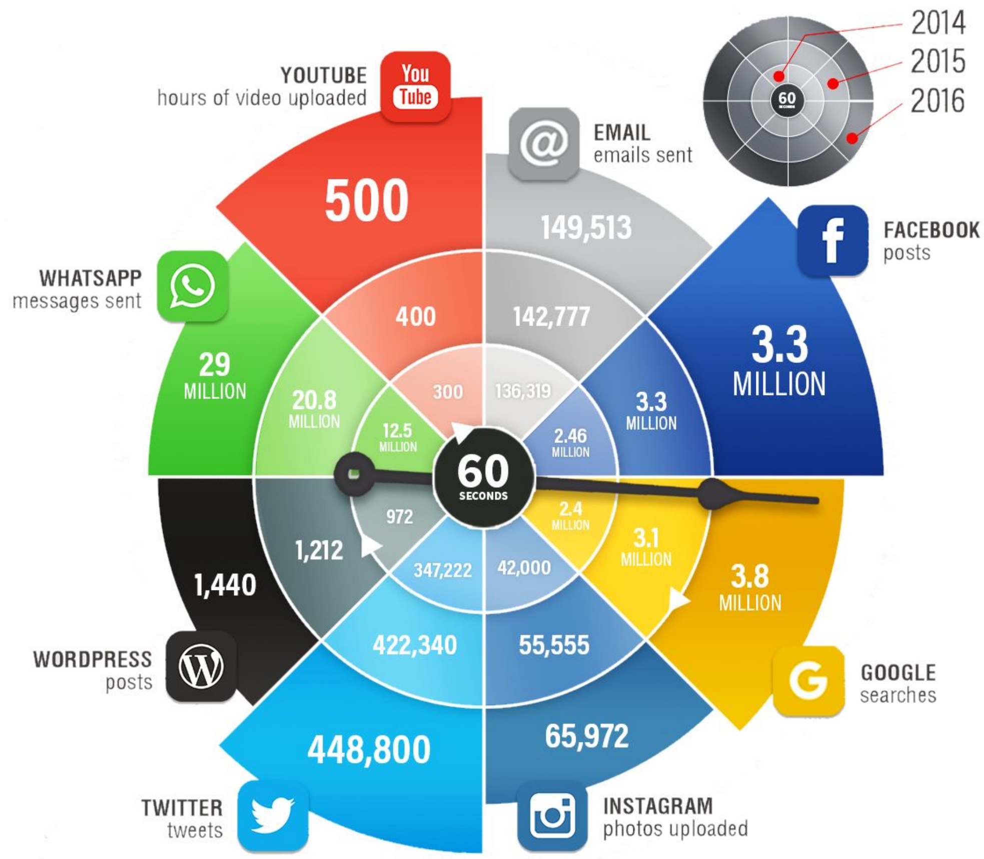


Figure 1: This infographic shows what happens online each minute (source https://www.smartinsights.com/wp-content/uploads/2016/08/What-happens-online-in-60-seconds.png)

The amount of information published on the Web increases over the years

Time consuming task

At the same time, the number of searches to consume this information grows

Confirmation bias

Why studying data veracity?

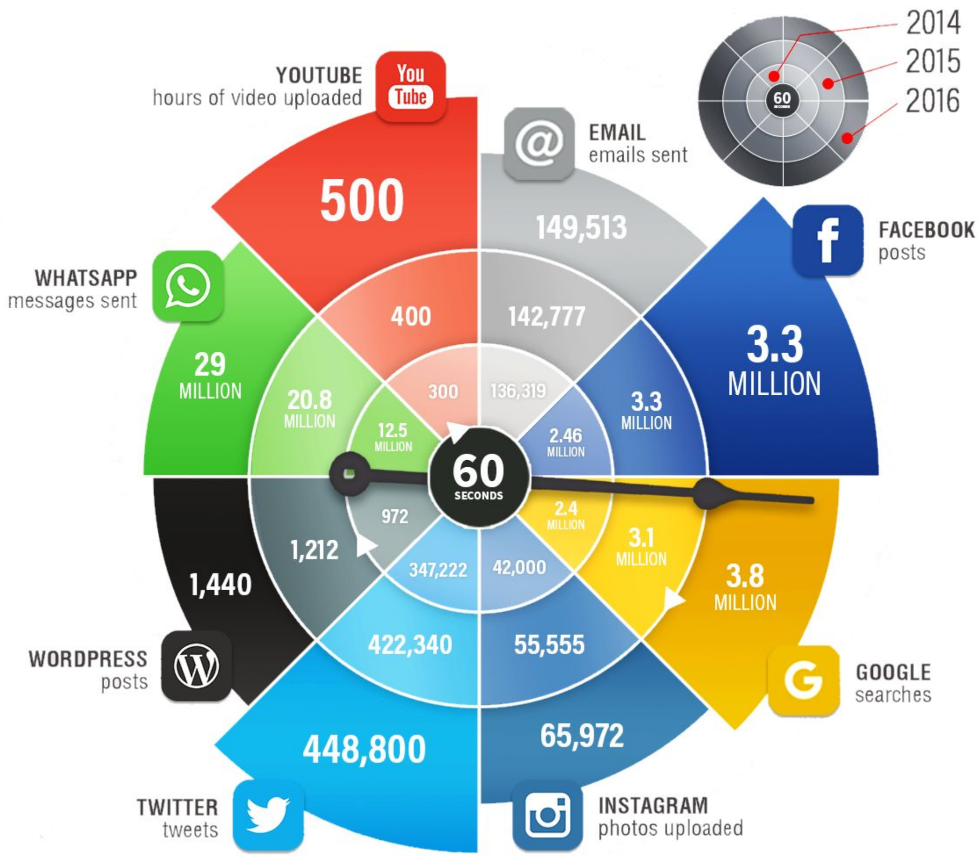


Figure 1: This infographic shows what happens online each minute (source https://www.smartinsights.com/wp-content/uploads/2016/08/What-happens-online-in-60-seconds.png)

The amount of information published on the Web increases over the years

Time consuming task

At the same time, the number of searches to consume this information grows

Confirmation bias

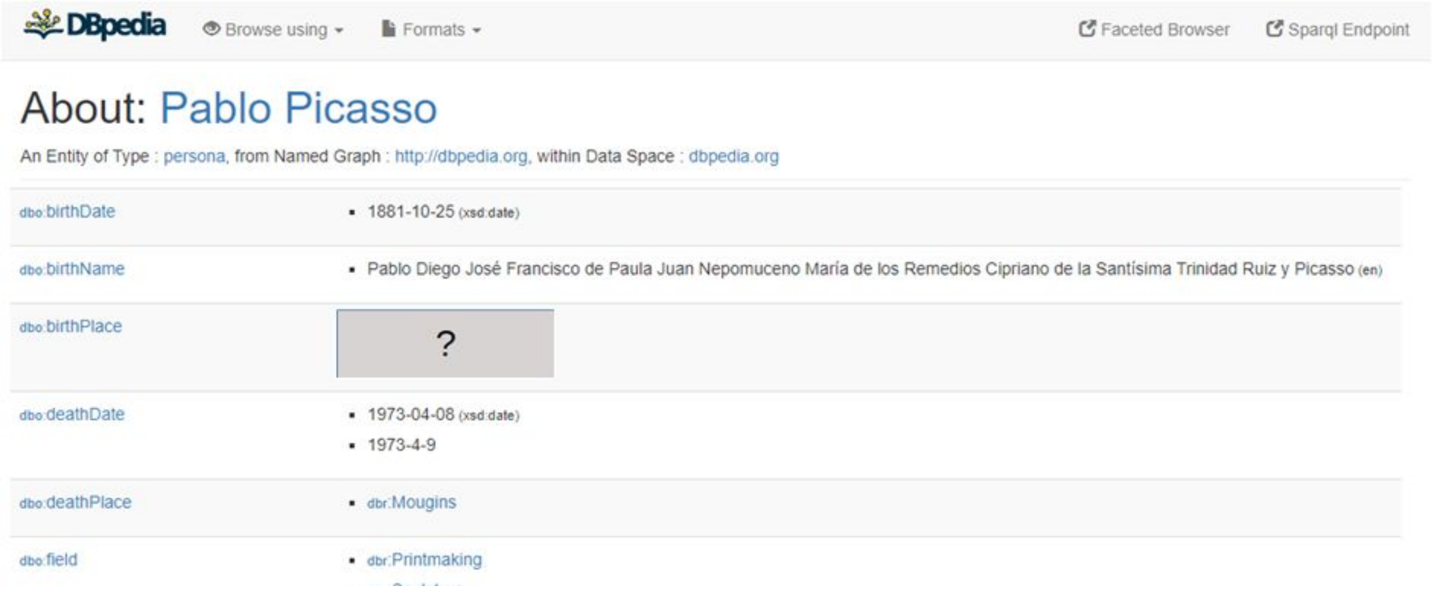
Automatically assessing data veracity is a critical task

What are the tasks that can benefit from data veracity?

- Information retrieval
- Business intelligence applications
- Decision-making processes
- Knowledge base population

What are the tasks that can benefit from data veracity?

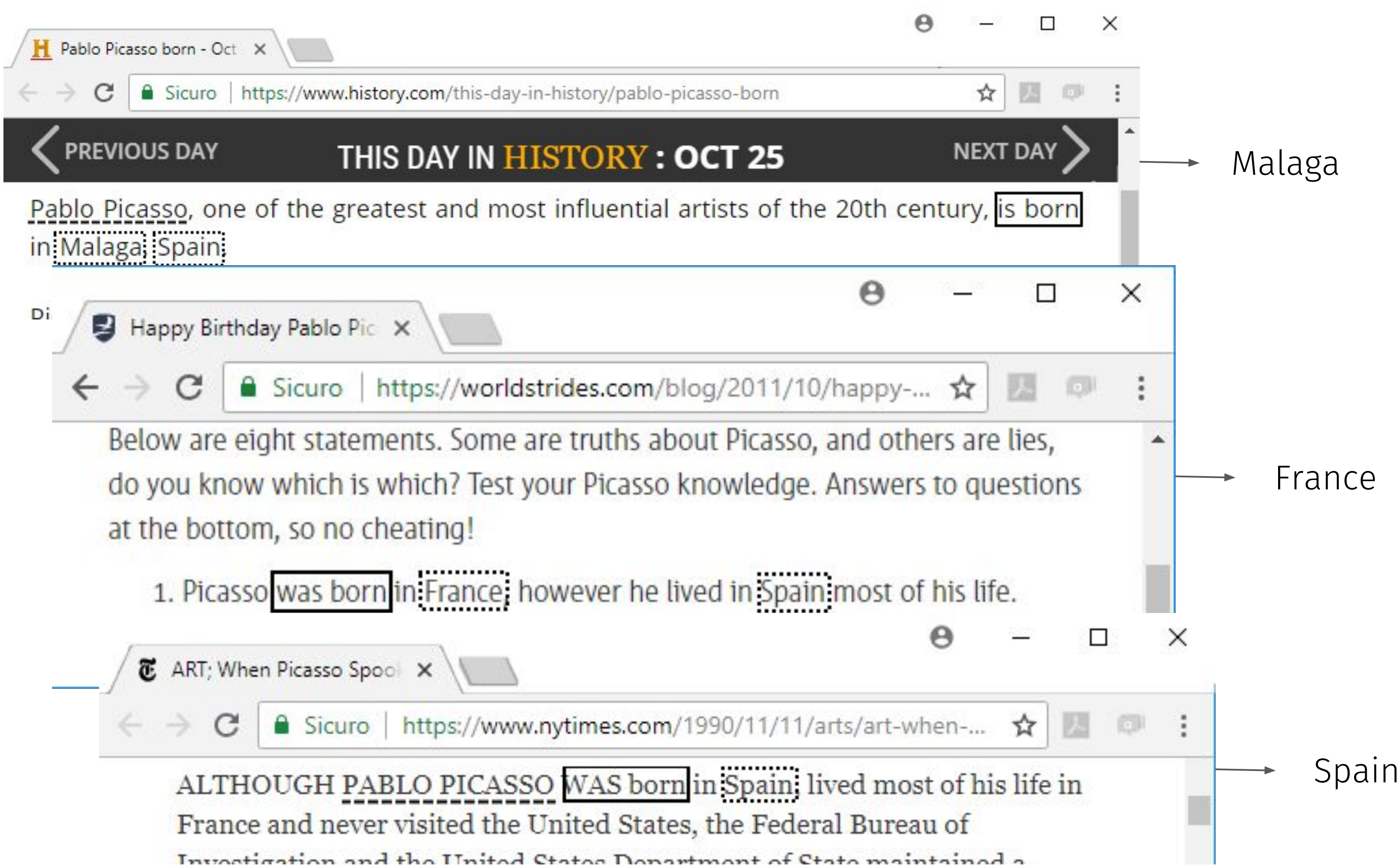
- Information retrieval
- Business intelligence applications
- Decision-making processes
- **Knowledge base population**



What are the tasks that can benefit from data veracity?

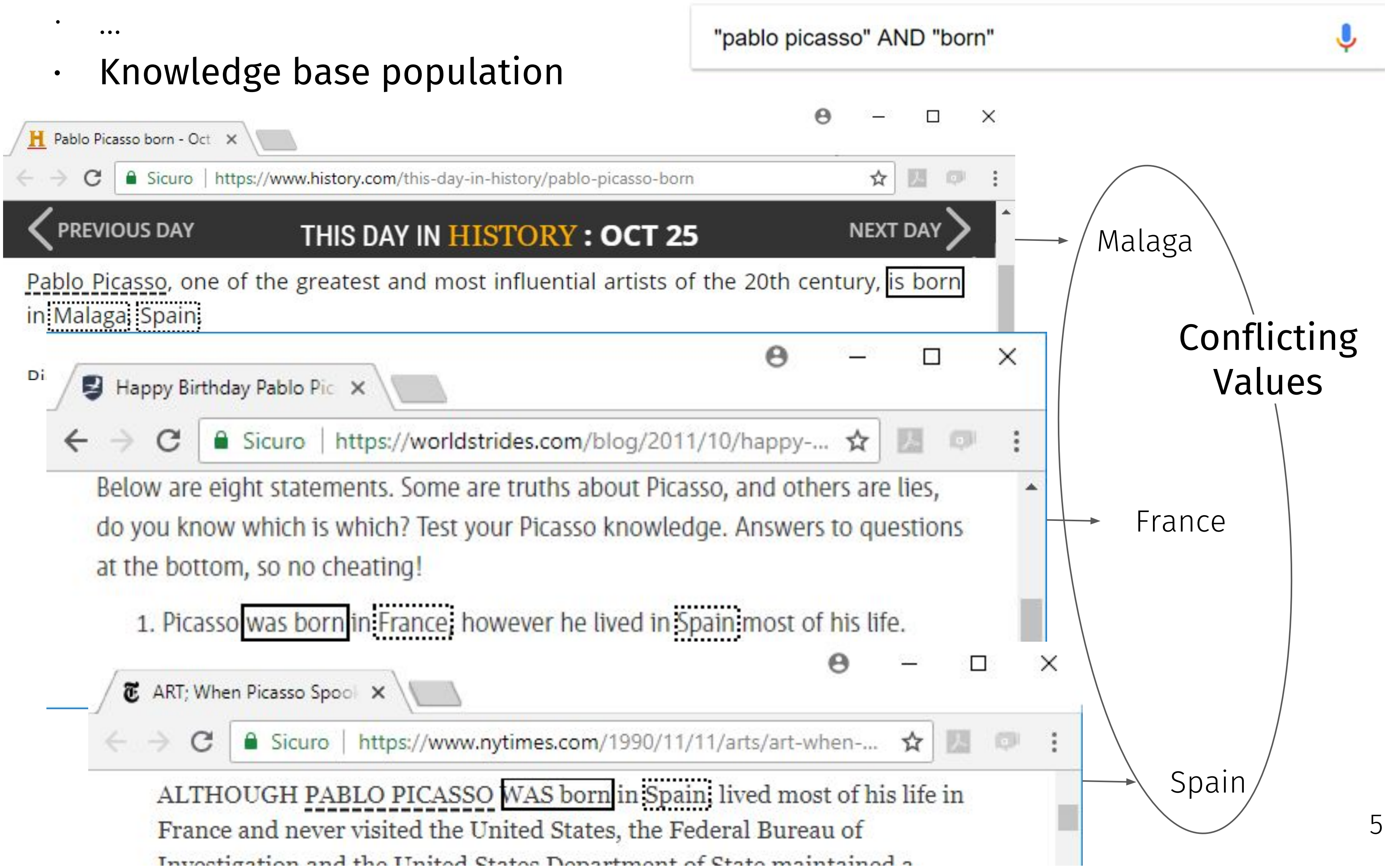
- ...
- Knowledge base population

"pablo picasso" AND "born"



What are the tasks that can benefit from data veracity?

- ...
- Knowledge base population



Data veracity assessment: Enhancing Truth Discovery using *a priori* knowledge

Outline:

1. Motivations behind data veracity assessment

2. Truth Discovery: problem and positioning

2.1 Truth Discovery: state-of-the-art

2.2 Ontologies: definition and enforcement

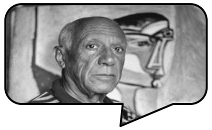
3. Enhancing Truth Discovery models using *a priori* knowledge

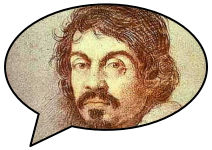
4. Conclusion

Truth Discovery

Truth Discovery

Where were these painters born?





	Data item ($d \in D$)		
	Subject	Predicate	

Truth Discovery

Where were these painters born?

	Data item ($d \in D$)		
	Subject	Predicate	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	...	...	

Truth Discovery

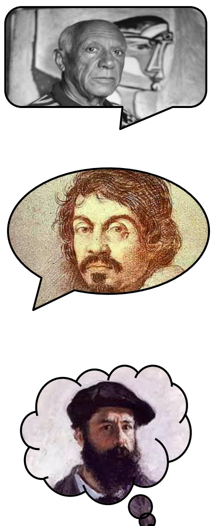
Where were these painters born?

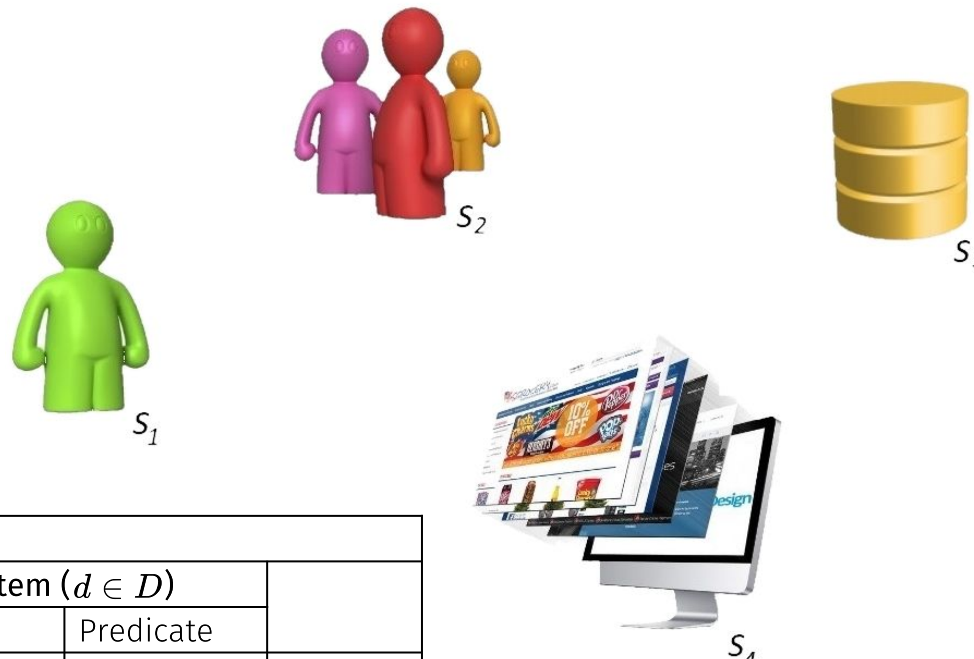
In this study we focus on functional properties

	Data item ($d \in D$)		
	Subject	Predicate	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	...	...	

Truth Discovery

Where were these painters born?





Source ($s \in S$)	Data item ($d \in D$)		
	Subject	Predicate	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	Pablo Picasso	bornIn	
	...	...	

Truth Discovery

Where were these painters born?



Source ($s \in S$)	Data item ($d \in D$)		
	Subject	Predicate	
A	Pablo Picasso	bornIn	
B	Pablo Picasso	bornIn	
C	Pablo Picasso	bornIn	
D	Pablo Picasso	bornIn	
E	Pablo Picasso	bornIn	
...	...	...	

Truth Discovery

Where were these painters born?

S_1

S_2

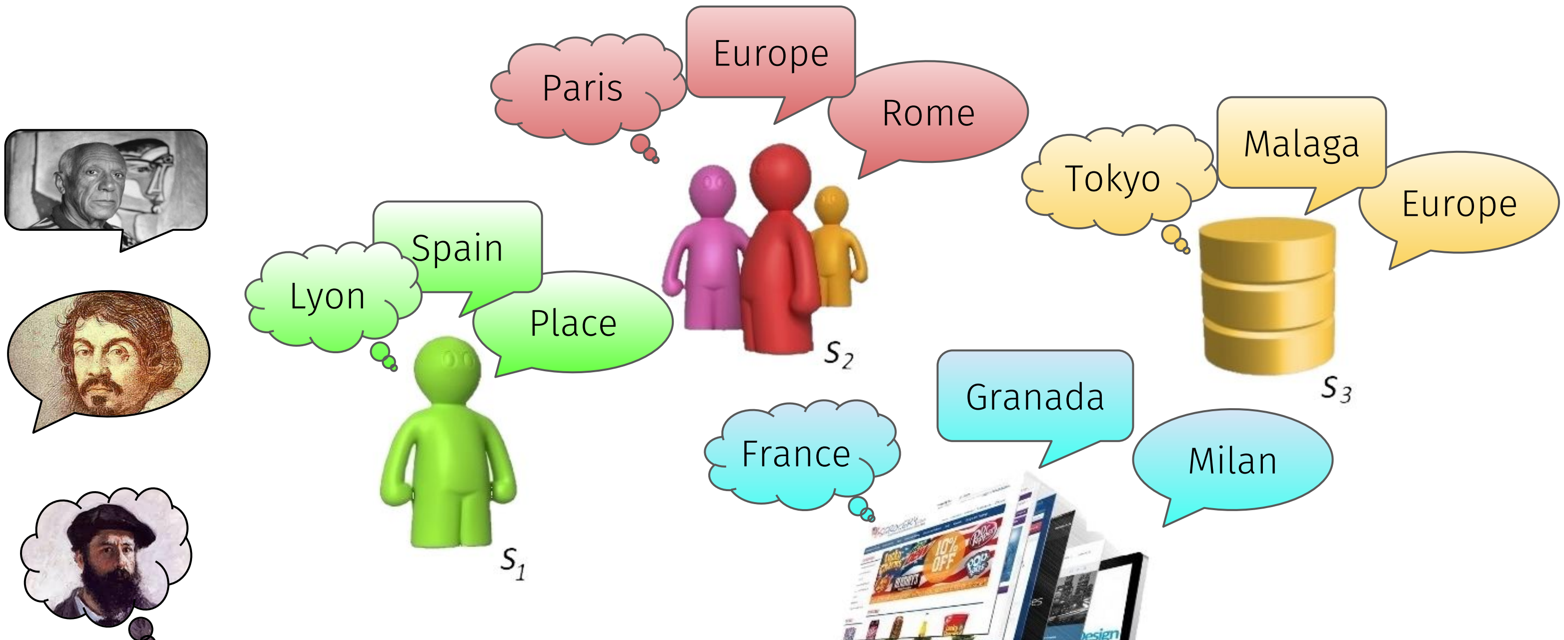
S_3

S_4

Source ($s \in S$)	Claim (abbreviated as v_d)		
	Data item ($d \in D$)		
	Subject	Predicate	
A	Pablo Picasso	bornIn	
B	Pablo Picasso	bornIn	
C	Pablo Picasso	bornIn	
D	Pablo Picasso	bornIn	
E	Pablo Picasso	bornIn	
...	...	...	

Truth Discovery

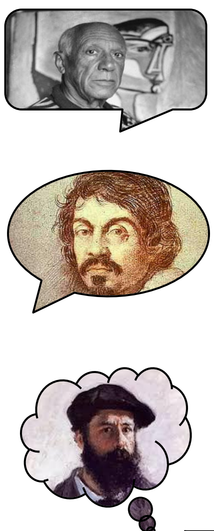
Where were these painters born?

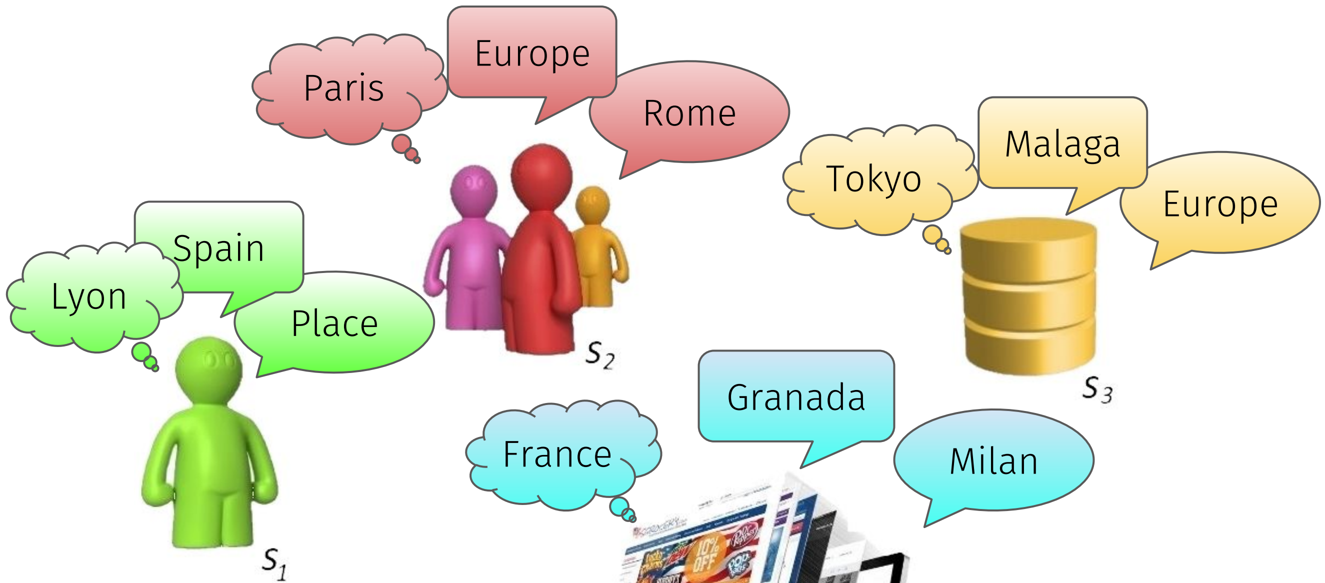


Source ($s \in S$)	Claim (abbreviated as v_d)		
	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	
B	Pablo Picasso	bornIn	
C	Pablo Picasso	bornIn	
D	Pablo Picasso	bornIn	
E	Pablo Picasso	bornIn	
...	...	...	

Truth Discovery

Where were these painters born?

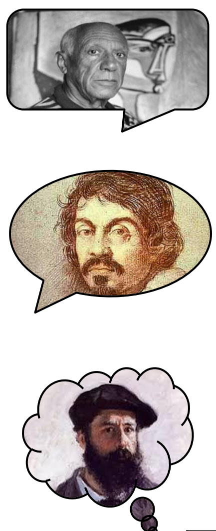


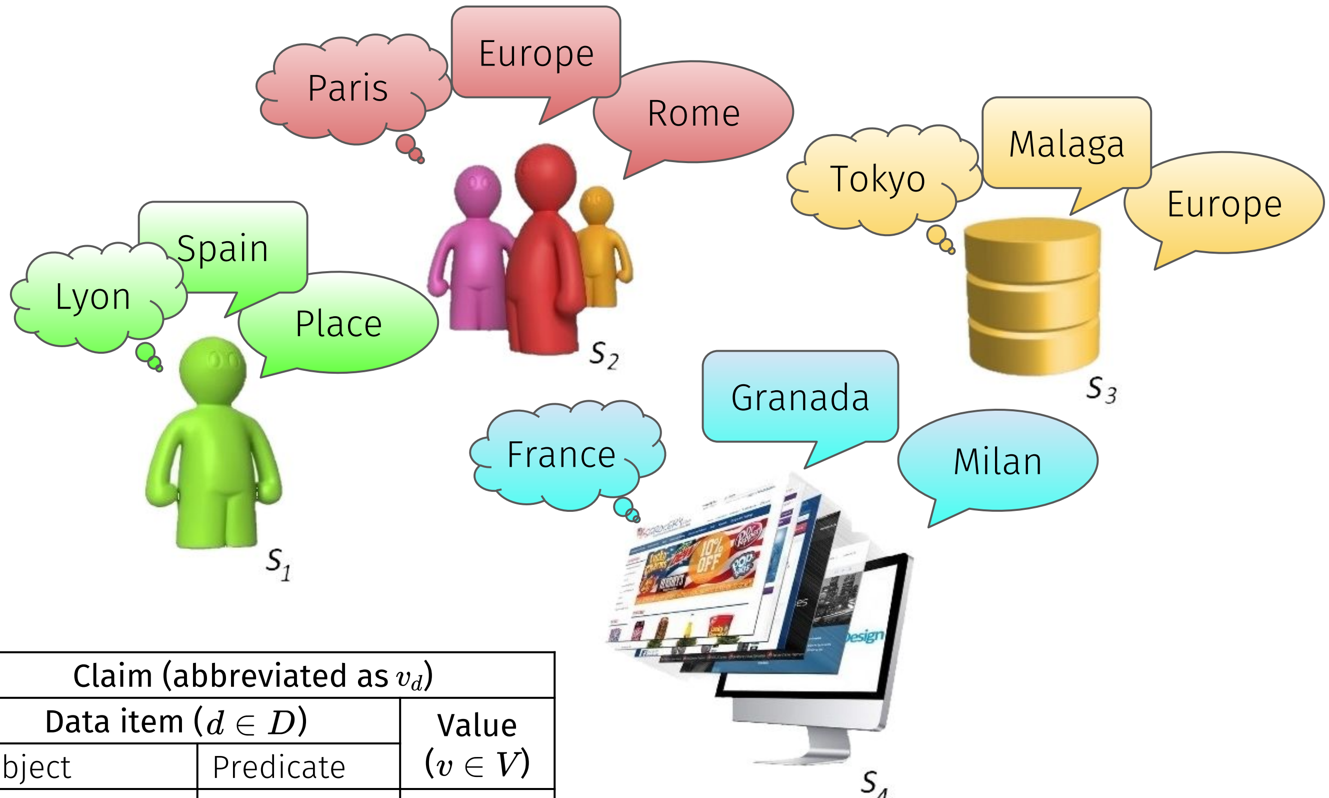


Source ($s \in S$)	Claim (abbreviated as v_d)		
	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Truth Discovery

Where were these painters born?





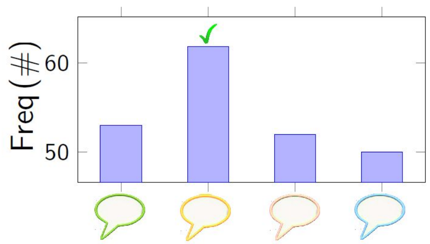
Source ($s \in S$)	Claim (abbreviated as v_d): Data item ($d \in D$): Subject	Predicate	Value ($v \in V$)
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Facts
(Pablo Picasso, bornIn, Málaga)
(Caravaggio, bornIn, Milan)
(Claude Monet, bornIn, Paris)

How to identify facts among conflicting claims?

Conflicts can be solved using as discriminating factors:

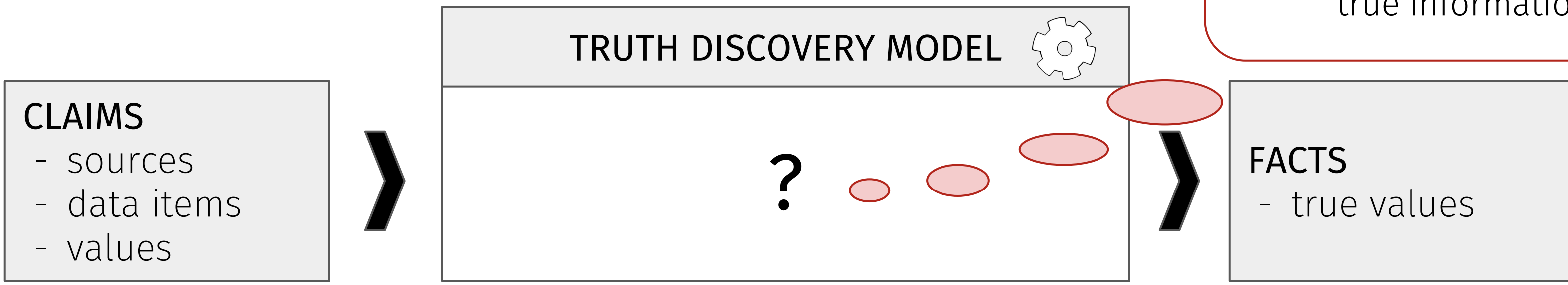
- frequency of answers
    - voting



- source reliability (or *source trustworthiness*) that can be evaluated based on source content

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information



How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

❯

TRUTH DISCOVERY MODEL

?

❯

FACTS
- true values

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

❯

TRUTH DISCOVERY MODEL

?

❯

FACTS
- true values

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

❯

TRUTH DISCOVERY MODEL

?

❯

FACTS
- true values

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Value confidence
c(Spain)
c(Madrid)
c(Europe)
c(Málaga)
c(Arles)
...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

TRUTH DISCOVERY MODEL

?

FACTS
- true values

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Value confidence
c(Spain)
c(Madrid)
c(Europe)
c(Málaga)
c(Arles)
...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

TRUTH DISCOVERY MODEL

?

FACTS
- true values

Source trustworthiness
t(A)
t(B)
t(C)
t(D)
t(E)
...

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Value confidence
c(Spain)
c(Madrid)
c(Europe)
c(Málaga)
c(Arles)
...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

TRUTH DISCOVERY MODEL

?

FACTS
- true values

Source trustworthiness
$t(A)$
$t(B)$
$t(C)$
$t(D)$
$t(E)$
...

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Spain
B	Pablo Picasso	bornIn	Madrid
C	Pablo Picasso	bornIn	Europe
D	Pablo Picasso	bornIn	Málaga
E	Pablo Picasso	bornIn	Arles
...	...	...	...

Value confidence
$c(Spain)$
$c(Madrid)$
$c(Europe)$
$c(Málaga)$
$c(Arles)$
...

How does Truth Discovery work?

True information is provided by reliable sources and, in turn, reliable sources claim true information

CLAIMS
- sources
- data items
- values

TRUTH DISCOVERY MODEL

1. Estimations
Iterative computations of value confidence and source trustworthiness

?

FACTS
- true values

Source trustworthines
$t(A)$
$t(B)$
$t(C)$
$t(D)$
$t(E)$
...

Source trustworthiness
$t(A)$
$t(B)$
$t(C)$
$t(D)$
$t(E)$
...

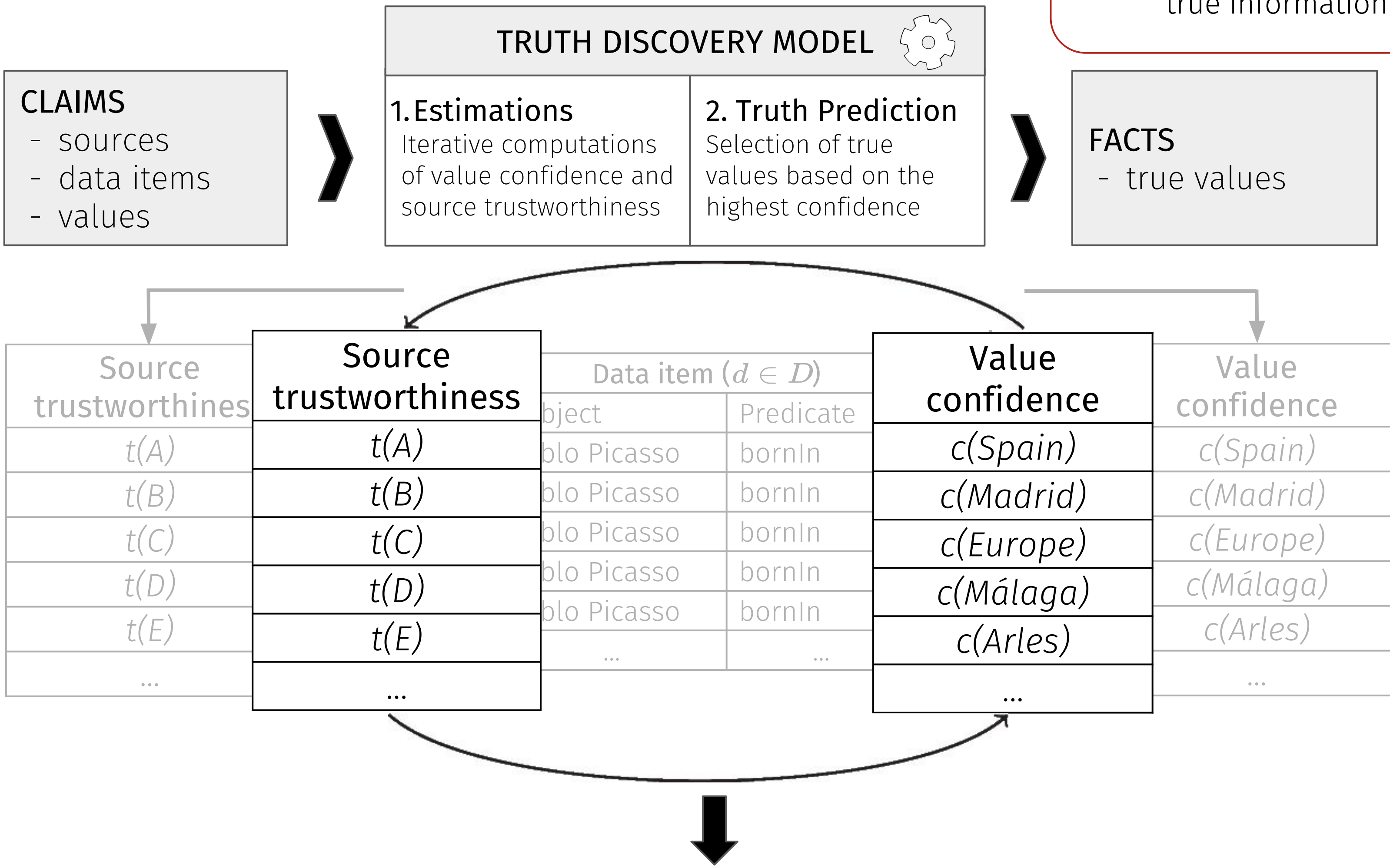
Data item ($d \in D$)	
bject	Predicate
blo Picasso	bornIn
blo Picasso	bornIn
blo Picasso	bornIn
blo Picasso	bornIn
blo Picasso	bornIn
...	...

Value confidence
$c(Spain)$
$c(Madrid)$
$c(Europe)$
$c(Málaga)$
$c(Arles)$
...

Value confidence
$c(Spain)$
$c(Madrid)$
$c(Europe)$
$c(Málaga)$
$c(Arles)$
...

How does Truth Discovery work?

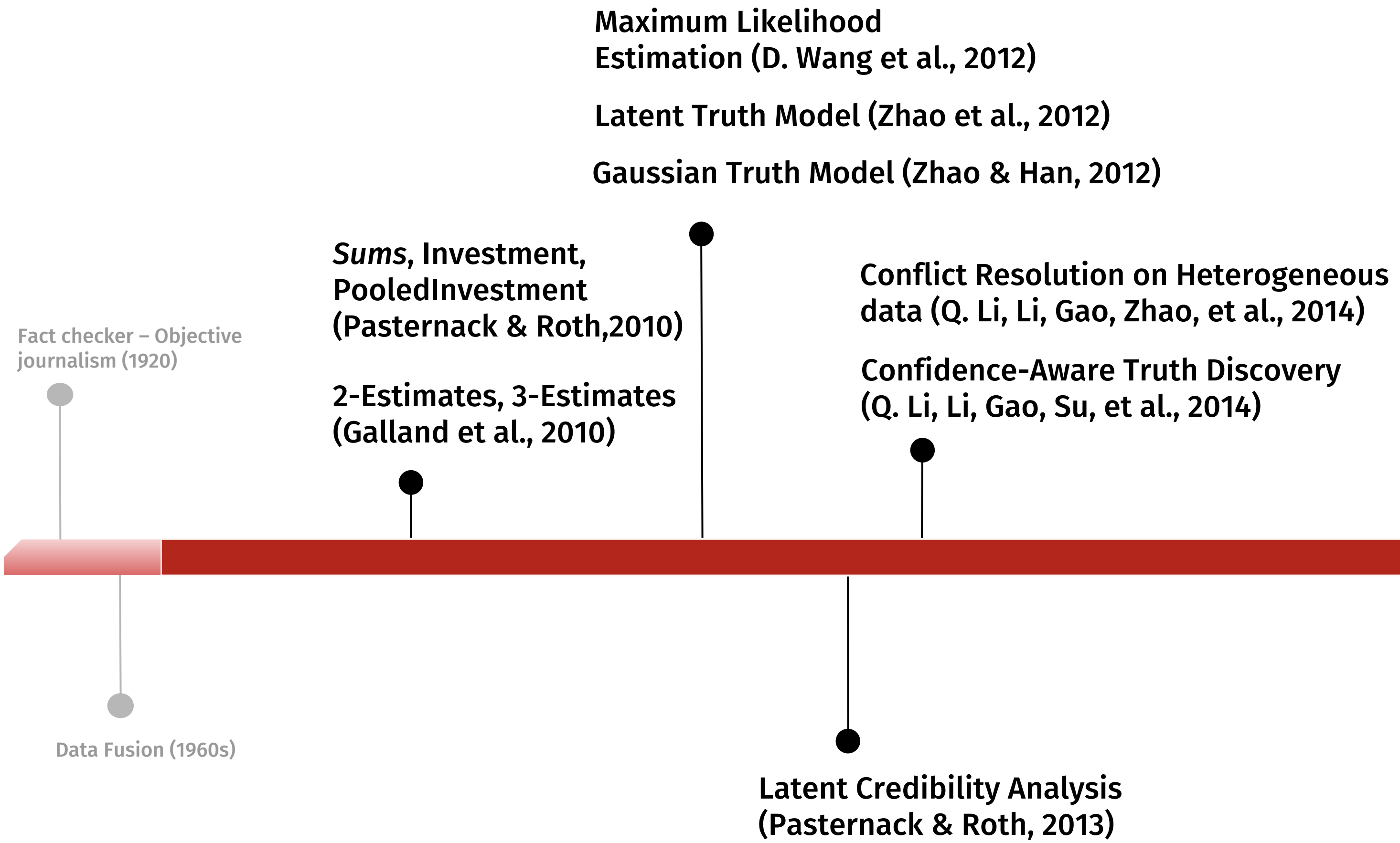
True information is provided by reliable sources and, in turn, reliable sources claim true information



True claims are the ones having the highest confidence

How are trustworthiness and confidence computed?

How are trustworthiness and confidence computed?



How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

Example

How to compute ***t(A)***?

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Málaga
B	Pablo Picasso	bornIn	Málaga
C	Pablo Picasso	bornIn	Madrid
A	Claude Monet	bornIn	Paris
B	Claude Monet	bornIn	France
...	...	...	...

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in \boxed{V_s}} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

Example

How to compute ***t(A)***?

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Málaga
B	Pablo Picasso	bornIn	Málaga
C	Pablo Picasso	bornIn	Madrid
A	Claude Monet	bornIn	Paris
B	Claude Monet	bornIn	France
...	...	...	...

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

Example

How to compute

c(bornIn(Pablo Picasso, Málaga))?

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Málaga
B	Pablo Picasso	bornIn	Málaga
C	Pablo Picasso	bornIn	Madrid
A	Claude Monet	bornIn	Paris
B	Claude Monet	bornIn	France
...	...	...	...

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

Example

How to compute *c(bornIn(Pablo Picasso, Málaga))*?

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Málaga
B	Pablo Picasso	bornIn	Málaga
C	Pablo Picasso	bornIn	Madrid
A	Claude Monet	bornIn	Paris
B	Claude Monet	bornIn	France
...	...	...	...

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

How are trustworthiness and confidence computed? An example

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in \boxed{S_{v_d}}} t^i(s)$$

with $d \in D$

Example

How to compute

c(bornIn(Pablo Picasso, Málaga))?

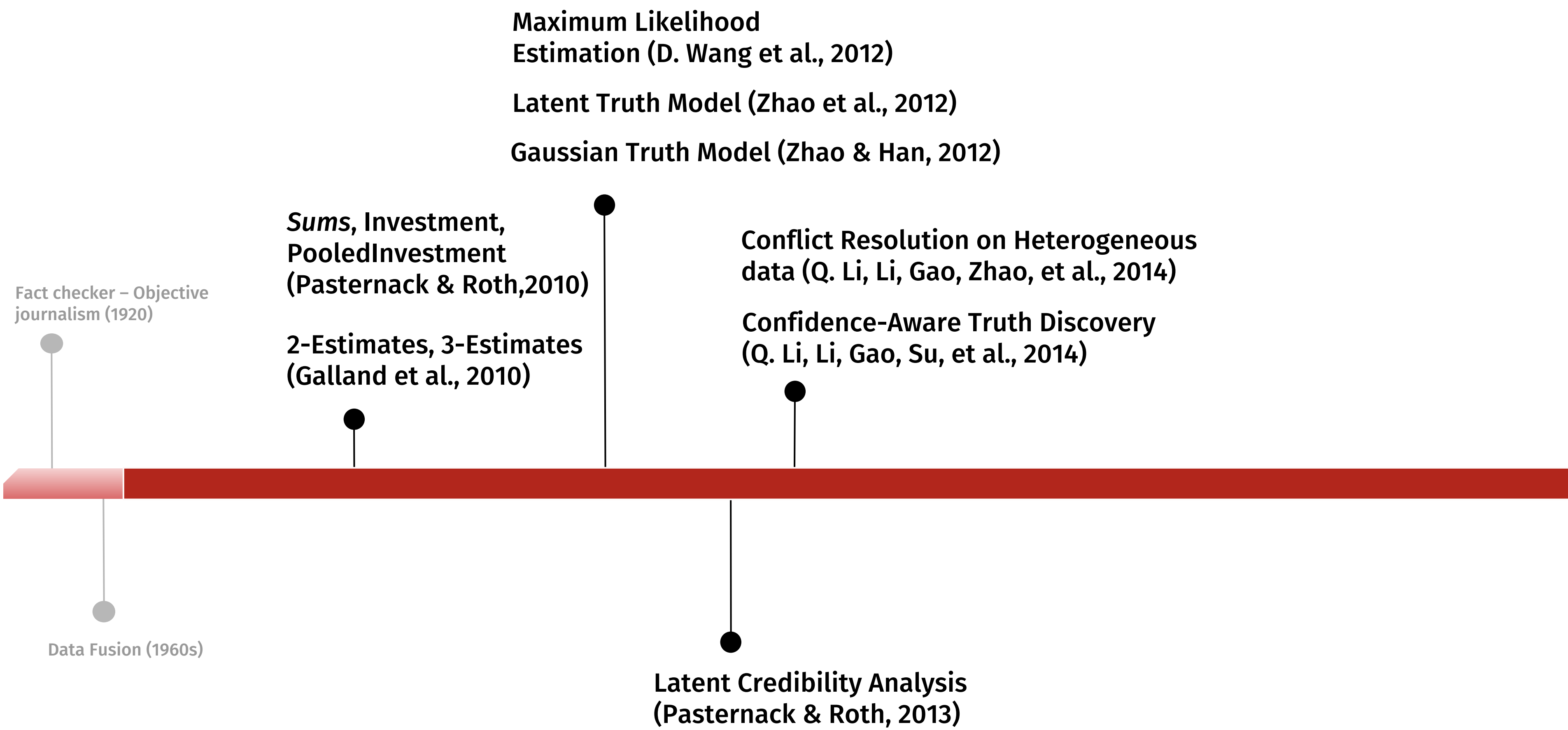
Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso	bornIn	Málaga
B	Pablo Picasso	bornIn	Málaga
C	Pablo Picasso	bornIn	Madrid
A	Claude Monet	bornIn	Paris
B	Claude Monet	bornIn	France
...	...	...	...

D = set of data items

V_s = set of claims provided by source s

S_{v_d} = set of sources that claim v_d

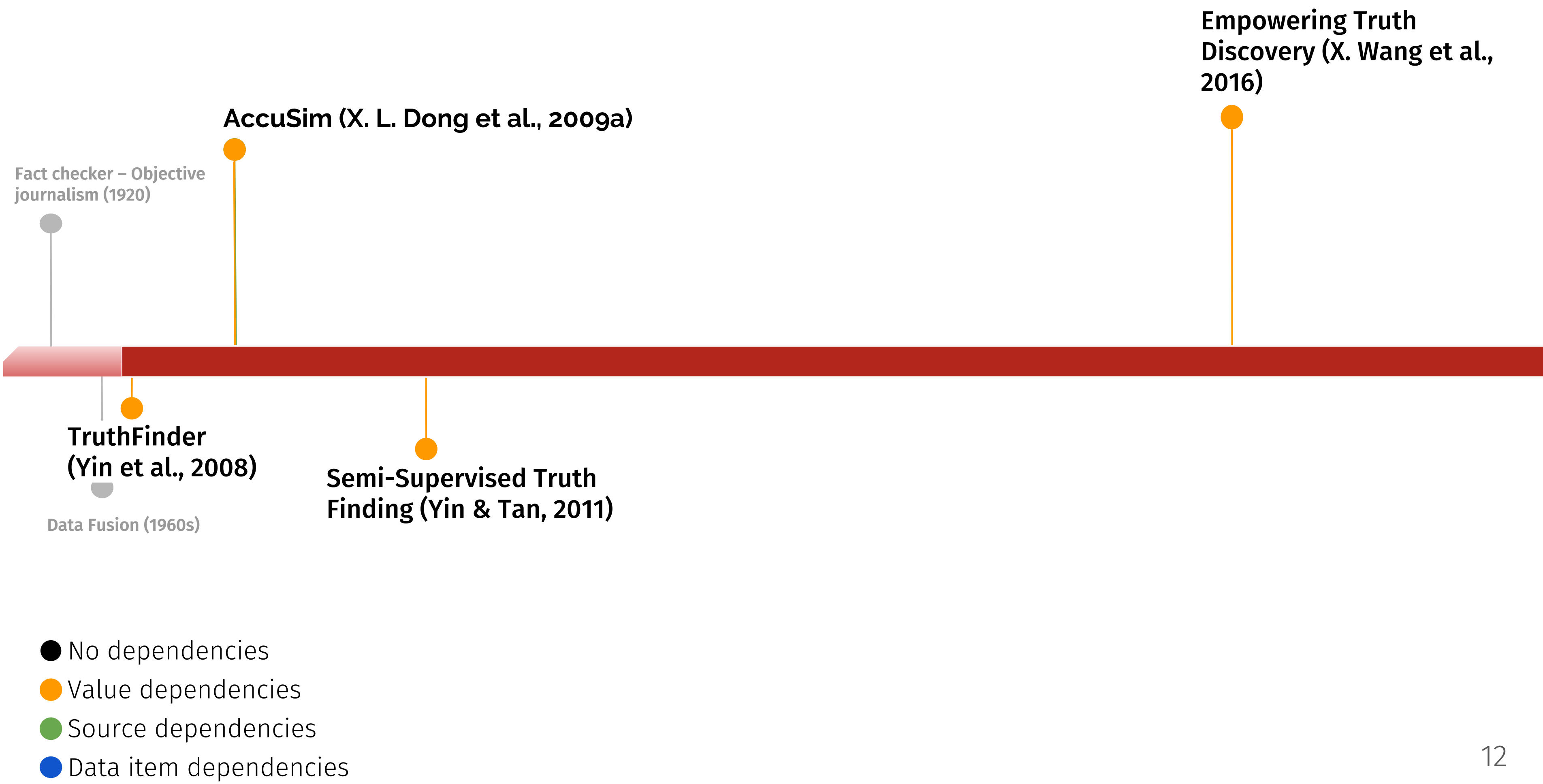
How are trustworthiness and confidence computed?



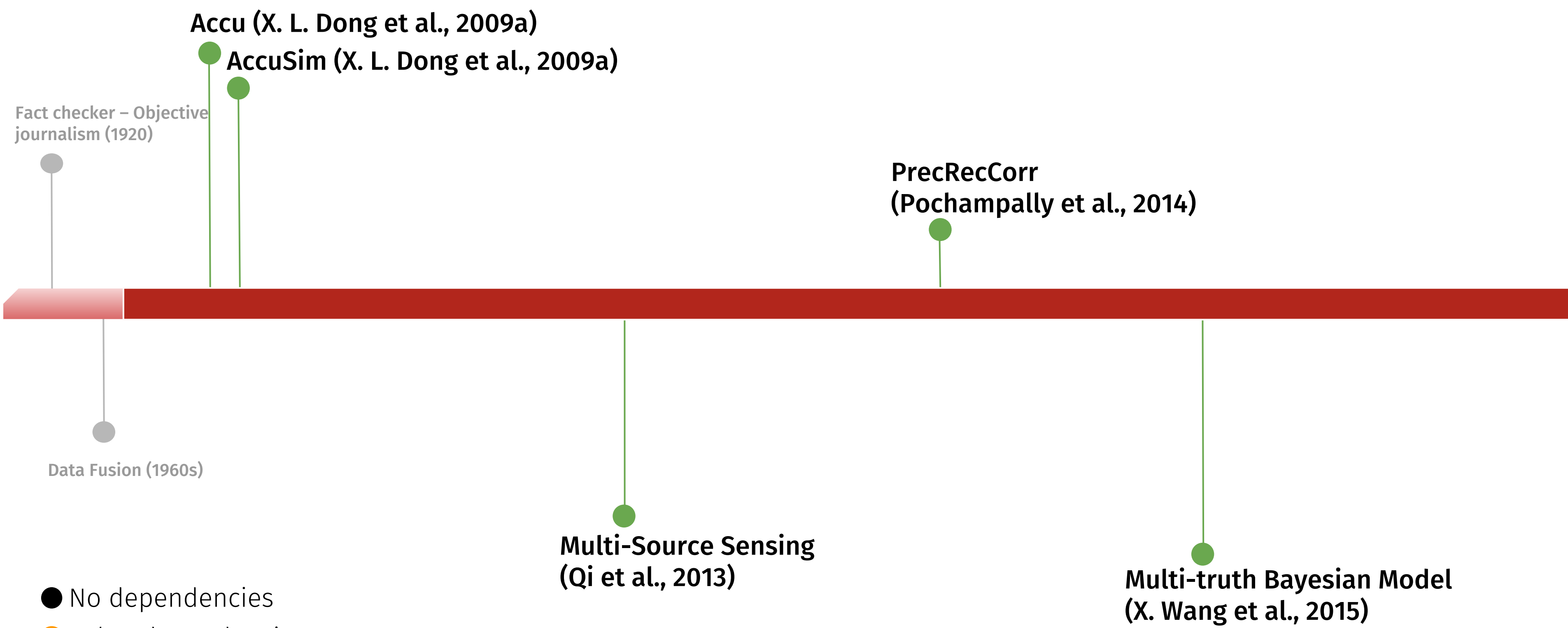
- No dependencies
- Value dependencies
- Source dependencies
- Data item dependencies

A **dependence** is a relationship that exists between two entities when an entity influences the other (whether they are sources, values or data items).

How are trustworthiness and confidence computed?

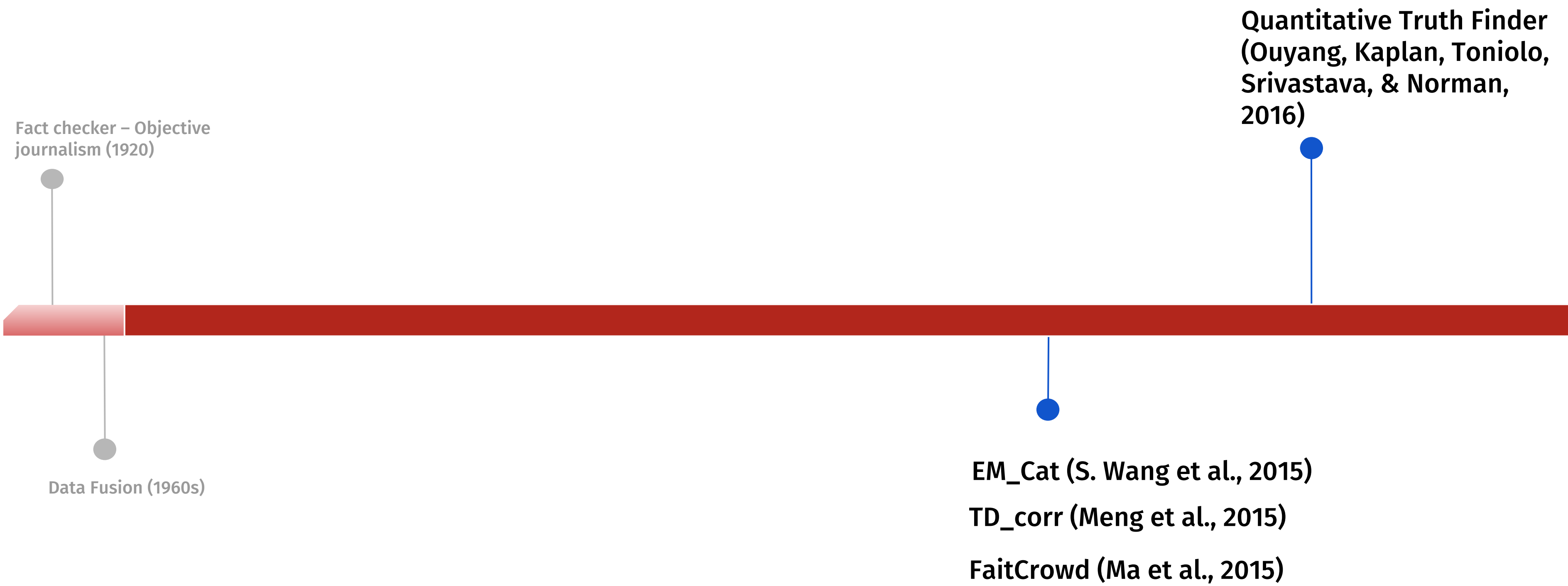


How are trustworthiness and confidence computed?



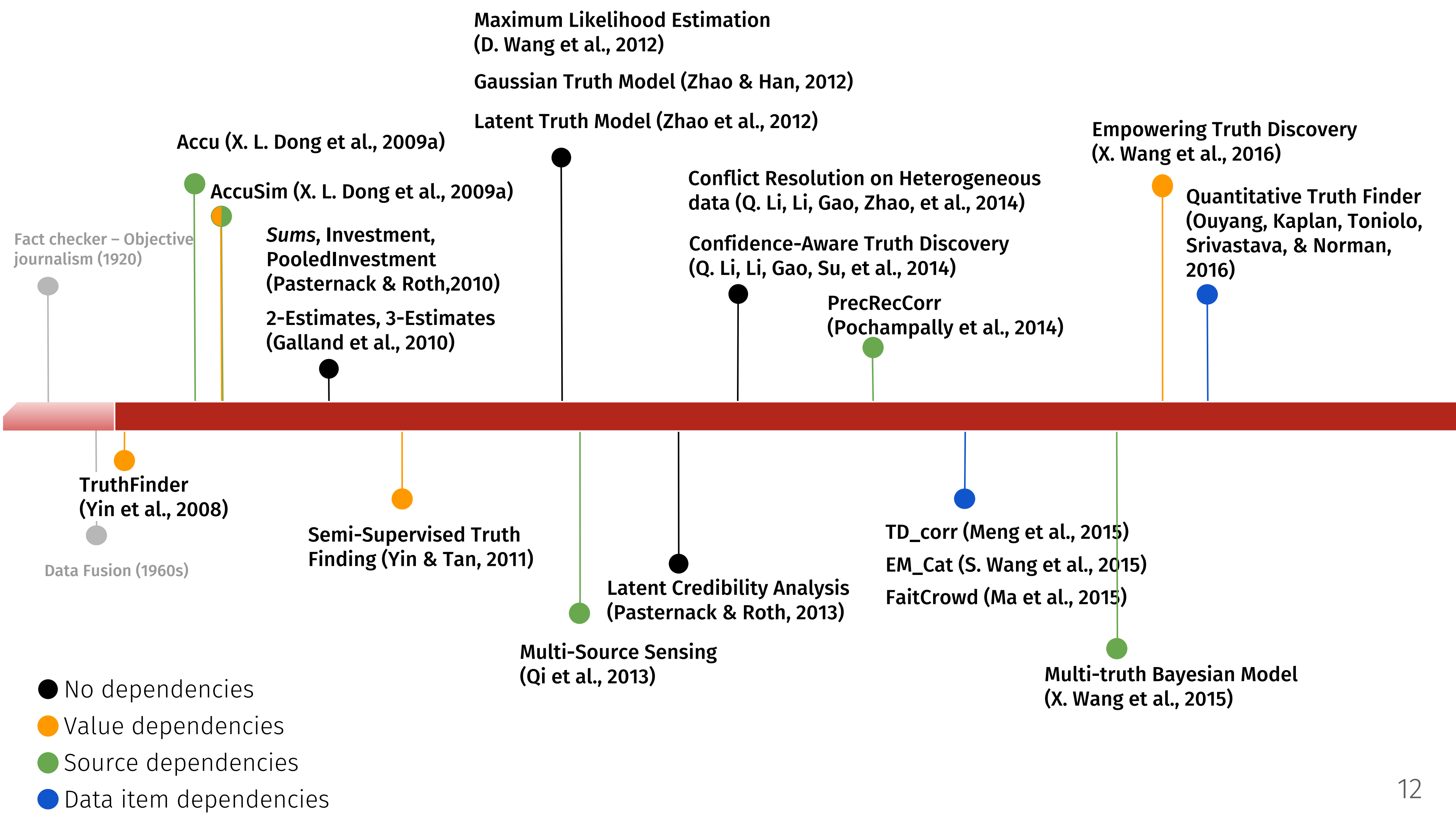
- No dependencies
- Value dependencies
- Source dependencies
- Data item dependencies

How are trustworthiness and confidence computed?



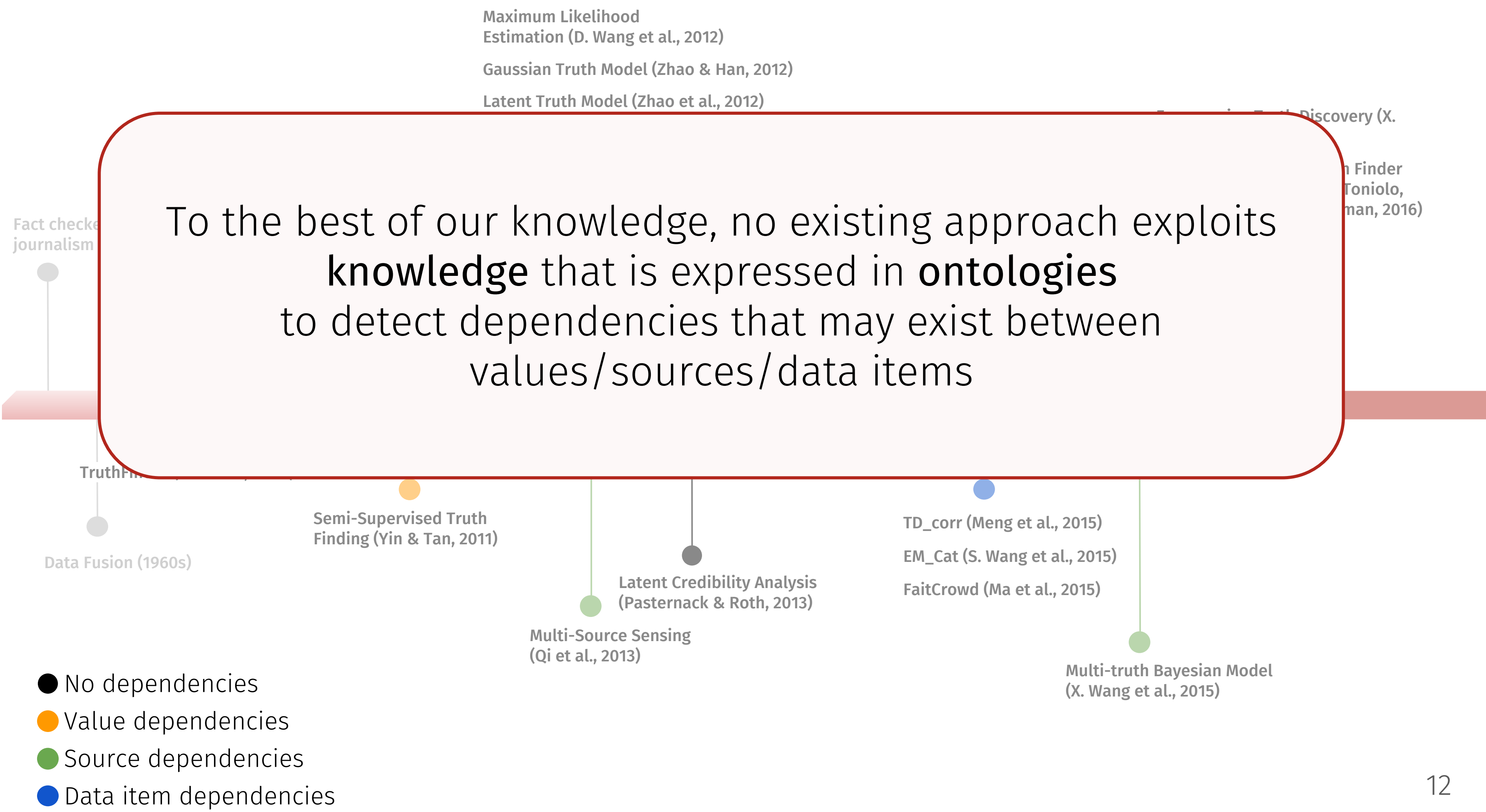
- No dependencies
- Value dependencies
- Source dependencies
- Data item dependencies

How are trustworthiness and confidence computed?



How are trustworthiness and confidence computed?

To the best of our knowledge, no existing approach exploits **knowledge** that is expressed in **ontologies** to detect dependencies that may exist between values/sources/data items



What is an ontology?

It is a well-defined knowledge representation that intends to represent knowledge in the most formal and reusable possible way.

"An ontology is an explicit specification of a conceptualization" (Gruber, 1993).

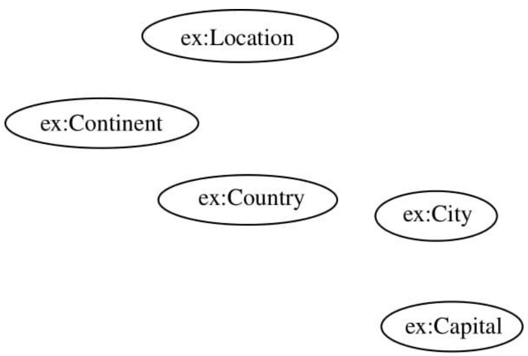
What is an ontology?

It is a well-defined knowledge representation that intends to represent knowledge in the most formal and reusable way possible.

"An ontology is an explicit specification of a conceptualization" (Gruber, 1993).

The basic elements of an ontology are:

- **Concepts** that represent classes of individuals sharing some properties;



ex: "http://www.example.com/"

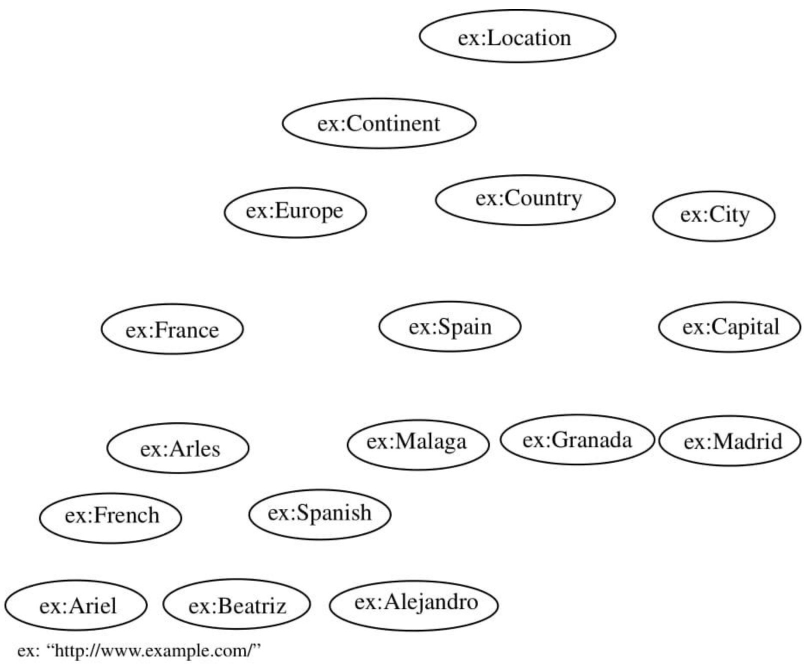
What is an ontology?

It is a well-defined knowledge representation that intends to represent knowledge in the most formal and reusable way possible.

"An ontology is an explicit specification of a conceptualization" (Gruber, 1993).

The basic elements of an ontology are:

- **Concepts** that represent classes of individuals sharing some properties;
- **Instances** that are actual occurrences of concepts;



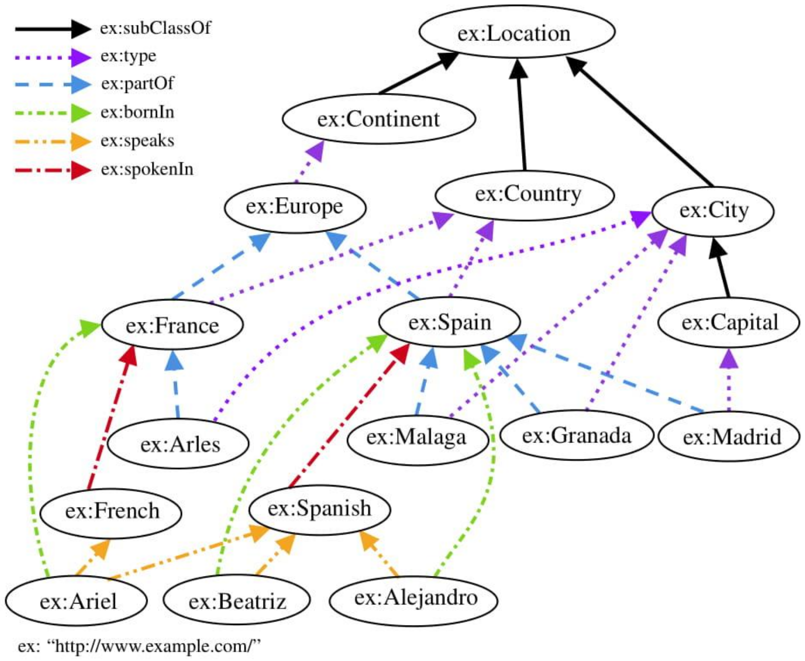
What is an ontology?

It is a well-defined knowledge representation that intends to represent knowledge in the most formal and reusable way possible.

"An ontology is an explicit specification of a conceptualization" (Gruber, 1993).

The basic elements of an ontology are:

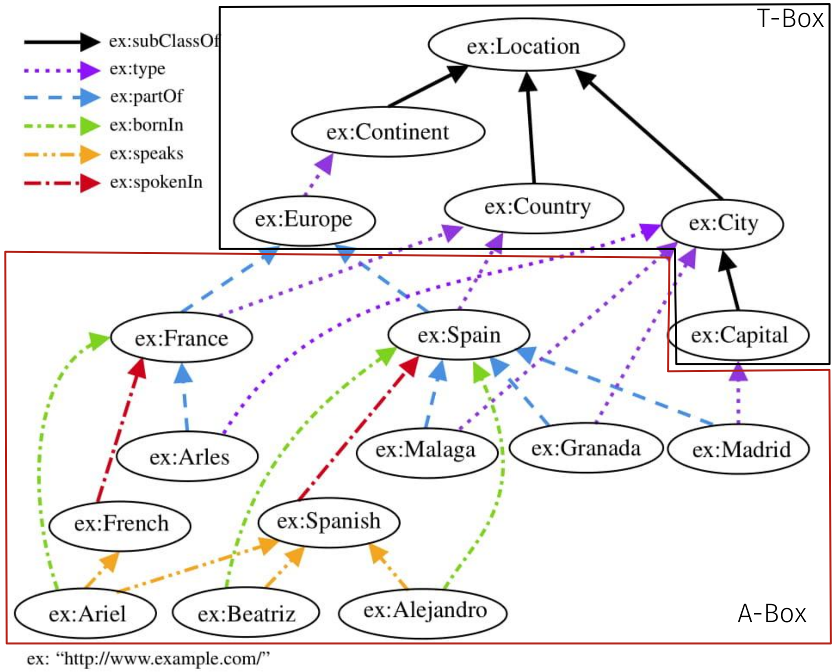
- **Concepts** that represent classes of individuals sharing some properties;
- **Instances** that are actual occurrences of concepts;
- **Relations** that are links or connections between instances, and between instances and concepts.



Which knowledge is contained in ontologies?

Ontologies based on Description Logics consist of:

- **T-Box** or terminological knowledge, knowledge that includes assertions about concepts and relations;
- **A-Box** or assertional knowledge, knowledge that includes assertions related to instances of the concepts and relations; the assertions should be expressed conforming to the T-Box.



Data veracity assessment: Enhancing Truth Discovery using *a priori* knowledge

Outline:

1. Motivations behind data veracity assessment

2. Truth Discovery: problem and positioning

3. Enhancing Truth Discovery models using *a priori* knowledge

3.1 Enhancing confidence estimations using value dependencies

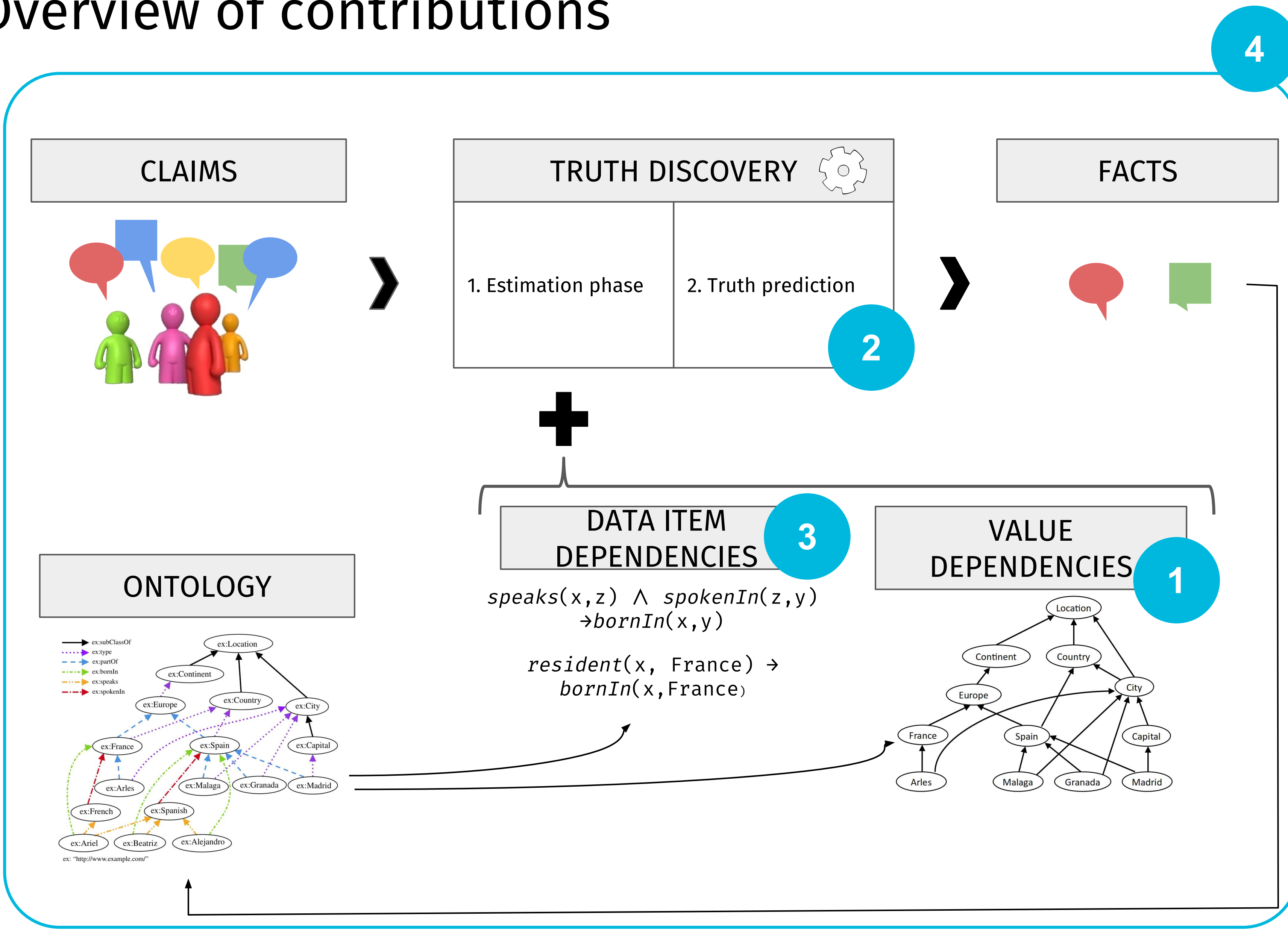
3.2 Truth prediction when considering value dependencies

3.3 Enhancing confidence estimations using data item dependencies

3.4 A case study on real-world data

4. Conclusion

Overview of contributions



Enhancing confidence estimations using value dependencies

CLAIMS

TRUTH DISCOVERY

1. Estimation phase	2. Truth prediction

FACTS

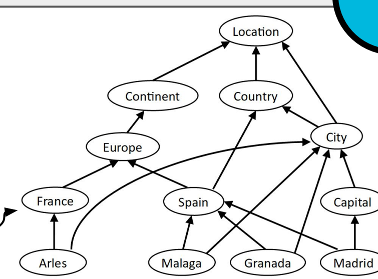
DATA ITEM DEPENDENCIES

speaks(x,z) ∧ *spokenIn*(z,y) →*bornIn*(x,y)

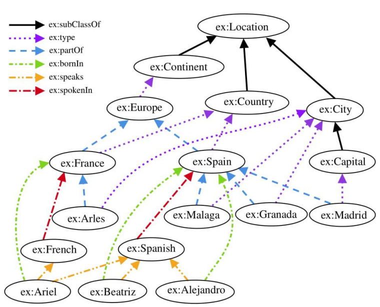
resident(x, France) → *bornIn*(x,France)

VALUE DEPENDENCIES

1

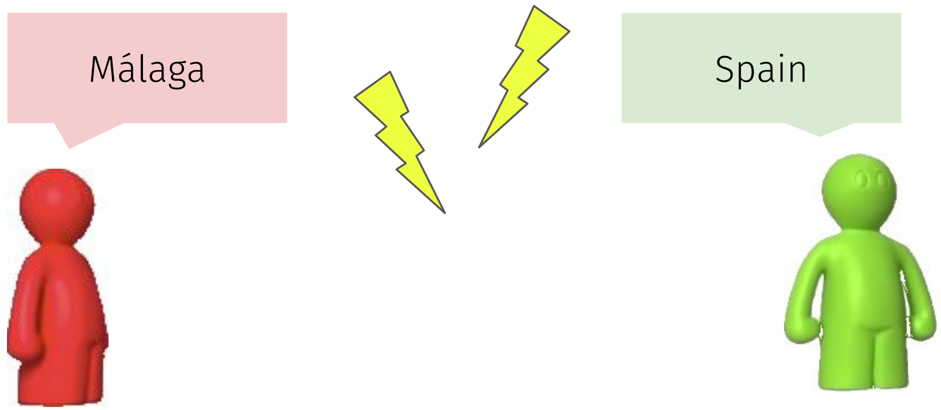


ONTOLOGY



How to exploit value dependencies?

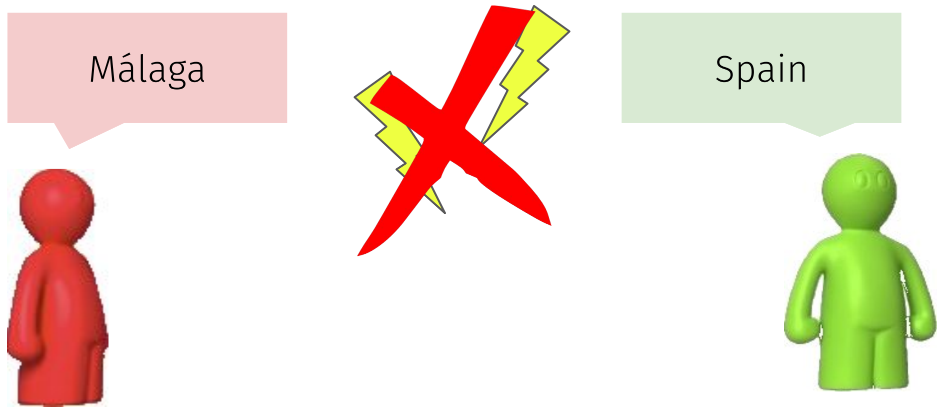
Intuition: *Where was Pablo Picasso born?*



The two sources answer with different values,
BUT ...

How to exploit value dependencies?

Intuition: *Where was Pablo Picasso born?*



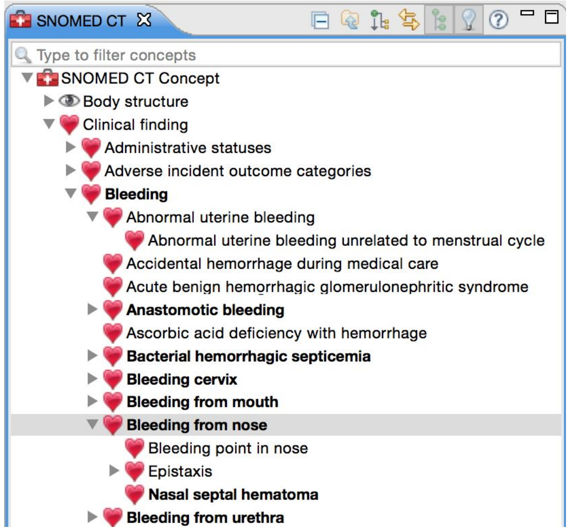
BUT ...
knowing that Málaga is in *Spain*
we can say that the source **explicitly** stating that "Pablo Picasso was born in *Málaga*" **implicitly** supports that "Pablo Picasso was born in *Spain*/*Europe*/...".

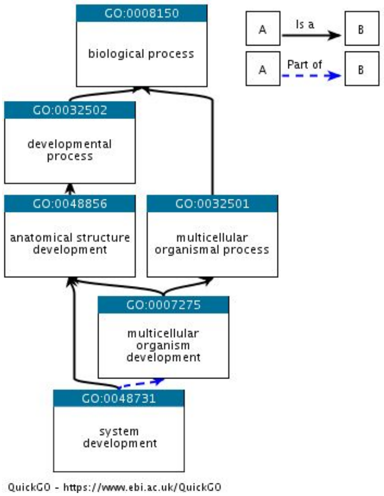
This interpretation is in accordance with the mathematical framework of belief function introduced by Dempster-Shafer theory.

How can value dependencies be modeled?

Several standards organizations attempt to create resources that contain **dedicated terms** (values) of specific domains; for instance:

- SNOMED for medical terms
- Gene Ontology for biological terms



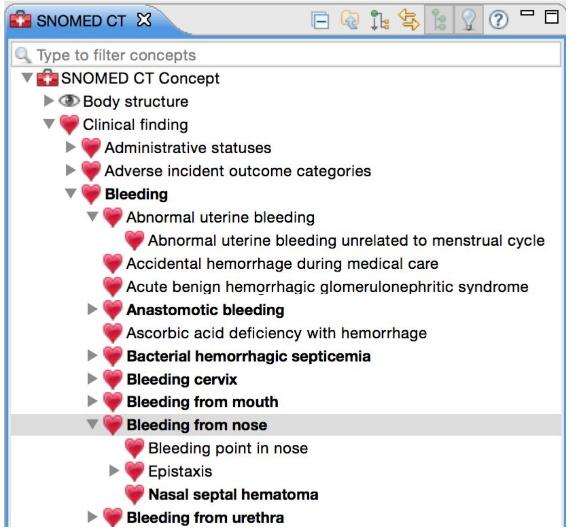


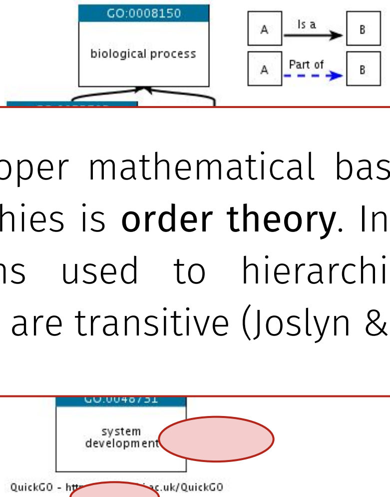
Usually these terms are organized in **hierarchies** based on taxonomic, meronomic and implication relations.

How can value dependencies be modeled?

Several standards organizations attempt to create resources that contain **dedicated terms** (values) of specific domains; for instance:

- SNOMED for medical terms
- Gene Ontology for biological terms





The proper mathematical basis to deal with hierarchies is **order theory**. Indeed, the main relations used to hierarchically structure objects are transitive (Joslyn & Hogan, 2010).

Usually these terms are organized in **hierarchies** based on taxonomic, meronomic and implication relations.

TD-*poset* approach

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

with $d \in D$

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

TD-*poset* approach

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$\boxed{c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)}$$

with $d \in D$

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

TD-*poset* approach

Sums (Pasternack & Roth, 2010)

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in \boxed{S_{v_d}}} t^i(s)$$

with $d \in D$

Sums$_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$

D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

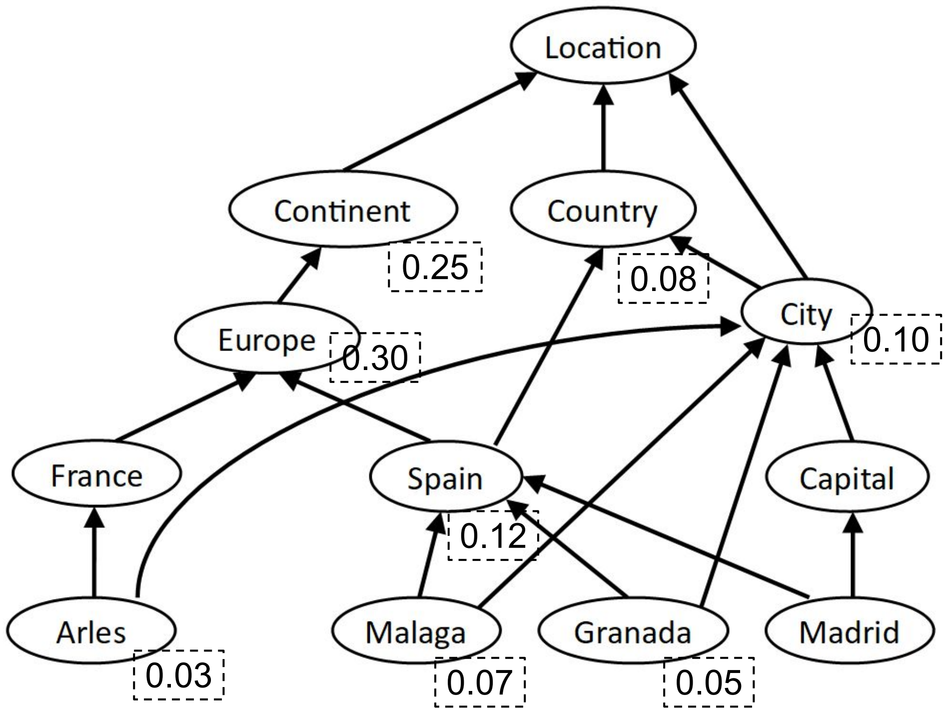
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

 trustworthiness of source that provides a value

 value confidence

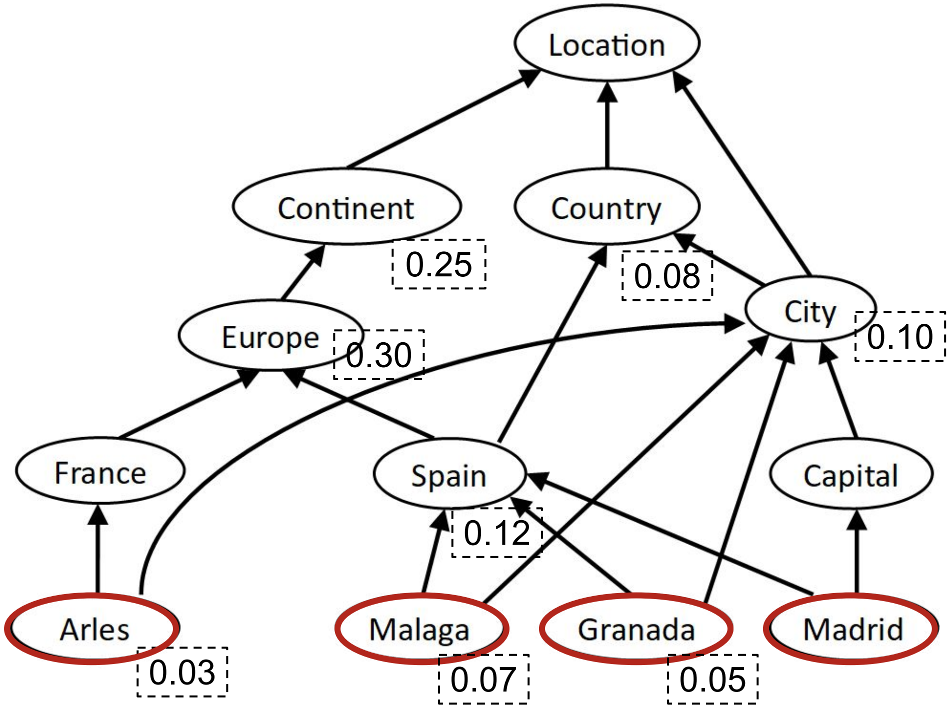
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

trustworthiness of source that provides a value

value confidence

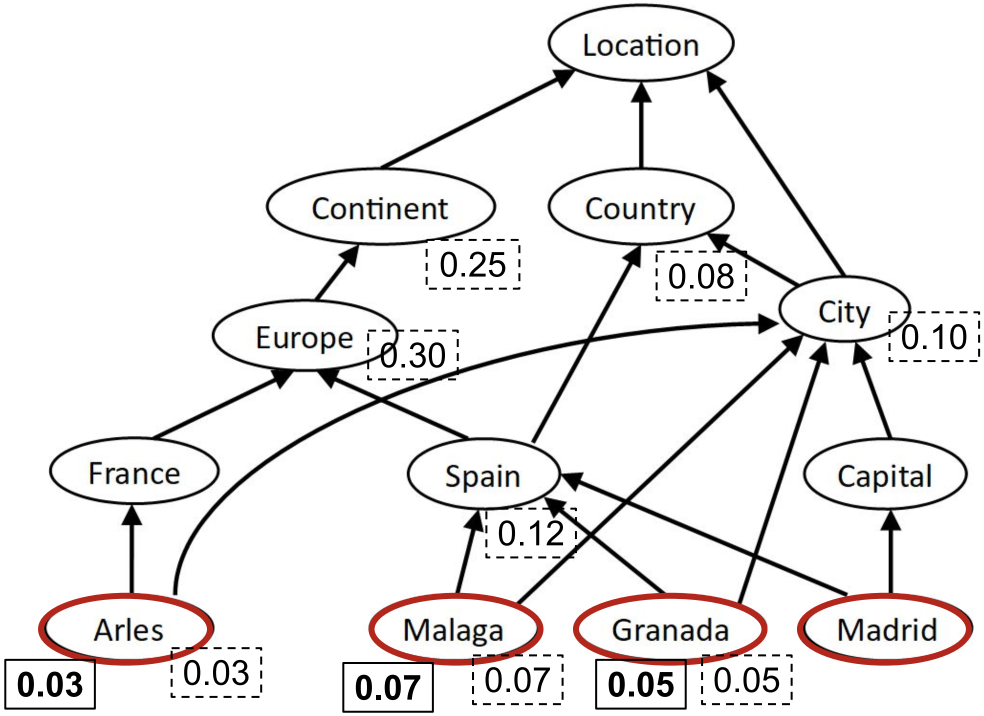
TD-*poset* approach: a practical example

Sums$_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

$$\text{with}\, S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

trustworthiness of source that provides a value

value confidence

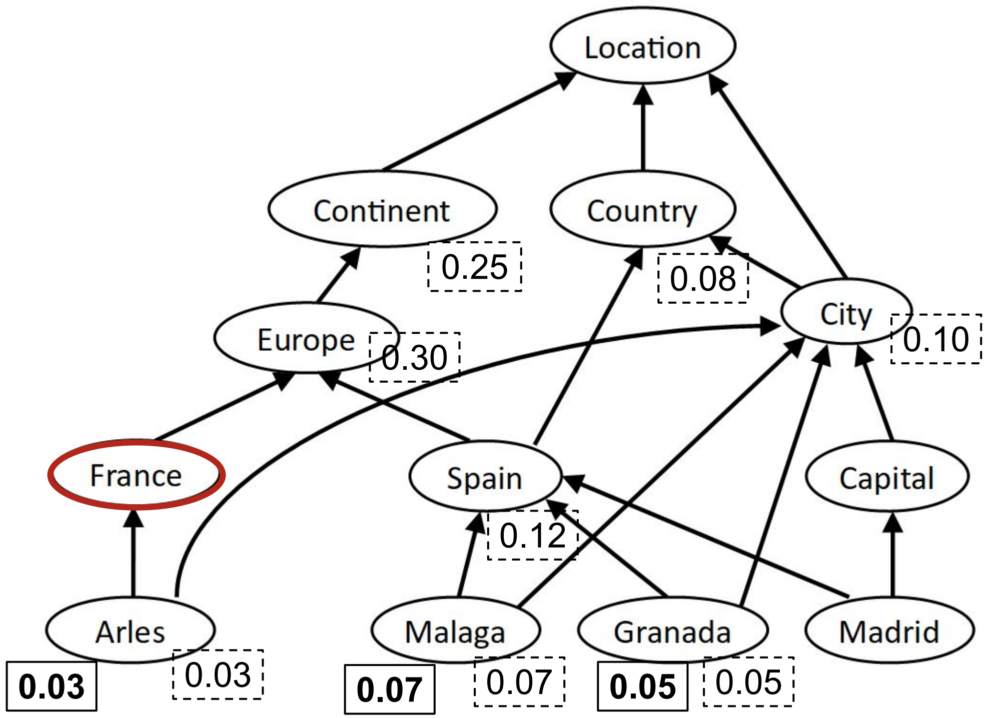
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

trustworthiness of source that provides a value (dashed box)
value confidence (solid box)

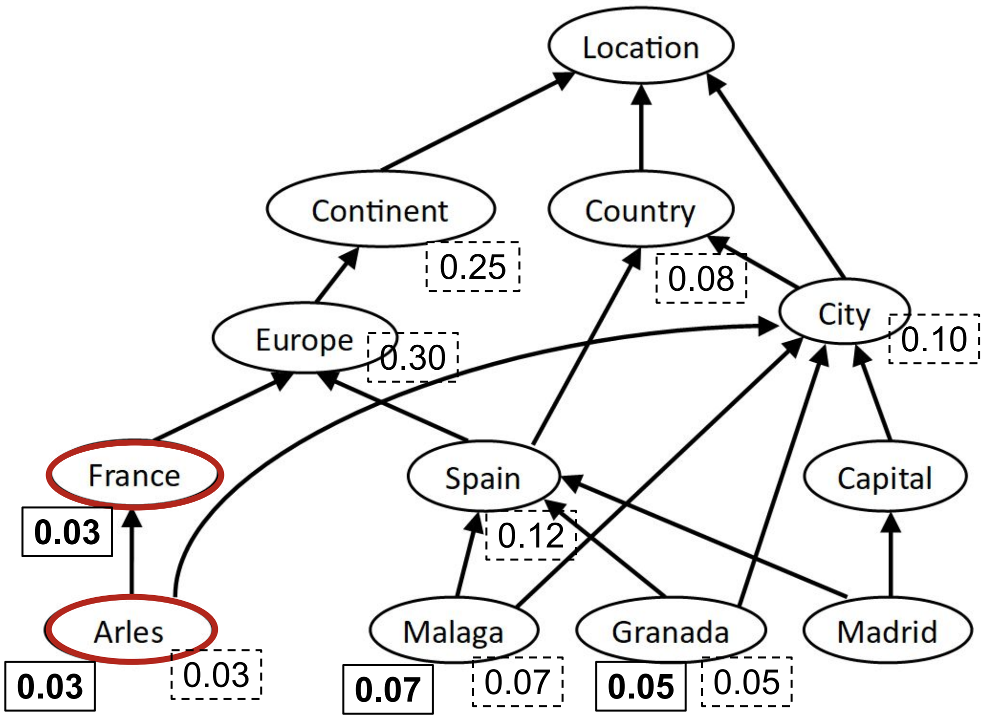
TD-*poset* approach: a practical example

Sums$_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

(dashed box) trustworthiness of source that provides a value

(solid box) value confidence

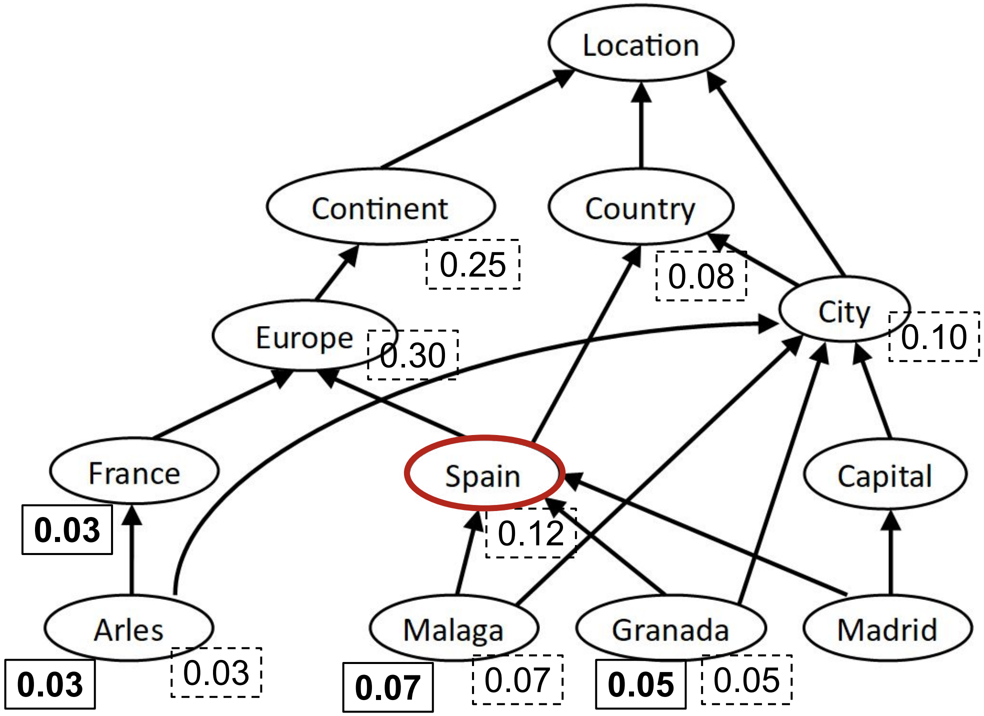
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

[dashed box] trustworthiness of source that provides a value

[solid box] value confidence

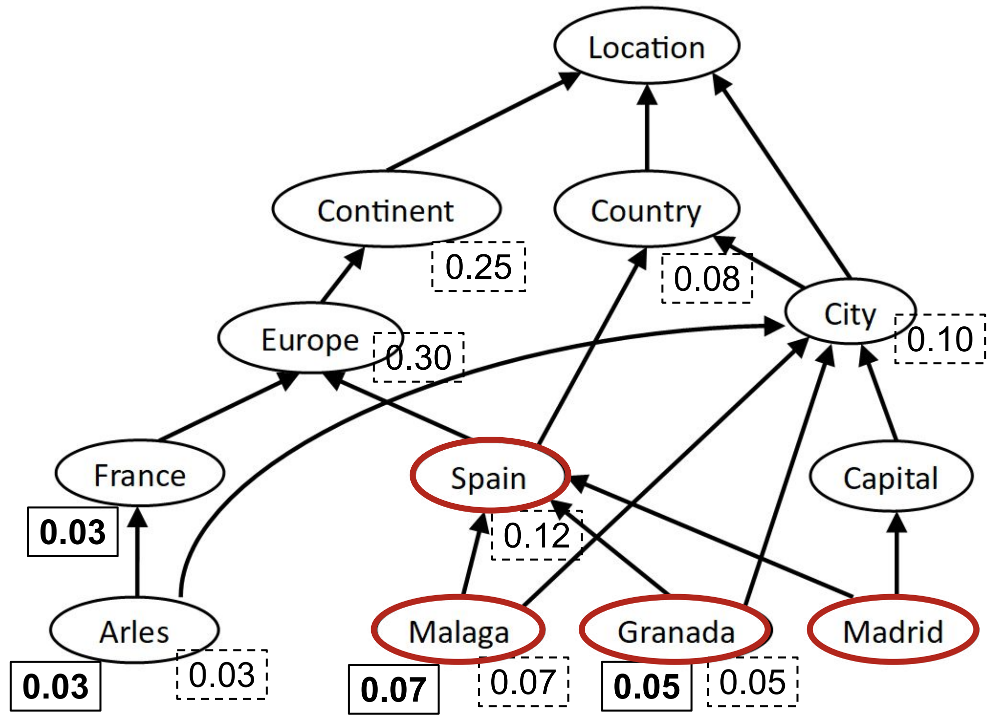
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

Dashed box: trustworthiness of source that provides a value

Solid box: value confidence

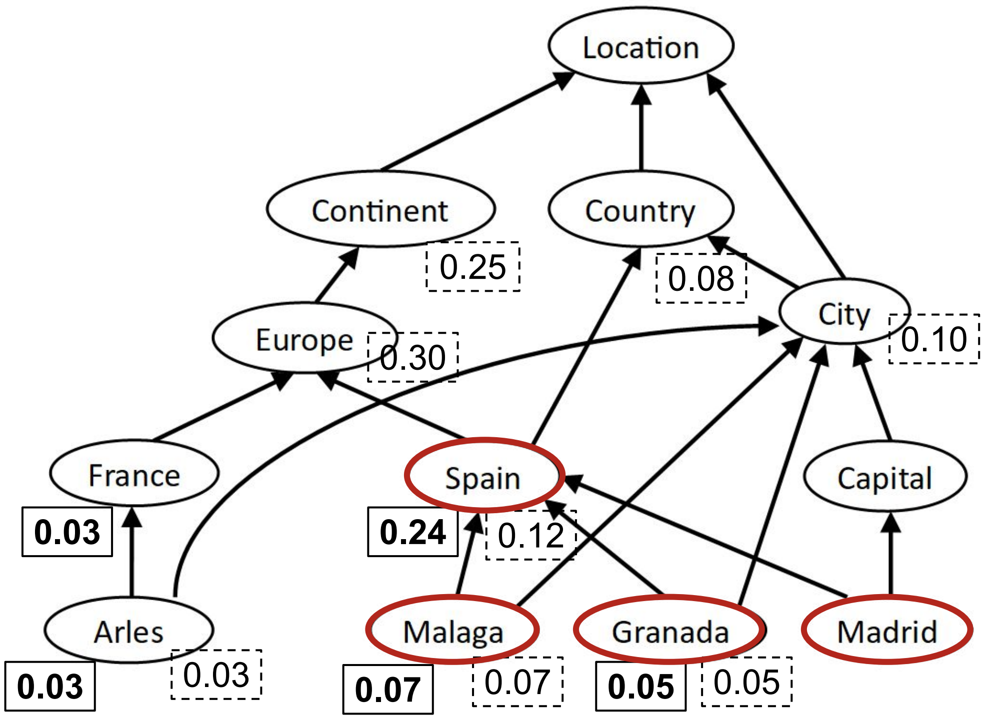
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

trustworthiness of source that provides a value

value confidence

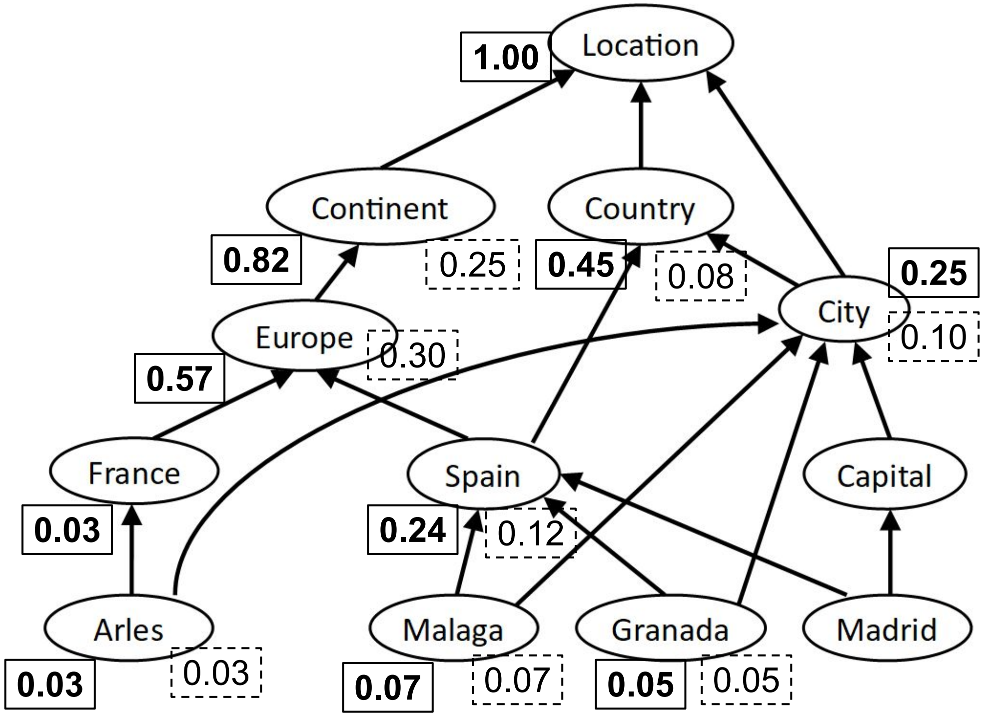
TD-*poset* approach: a practical example

Sums$_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

with $S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

[dashed box] trustworthiness of source that provides a value
[solid box] value confidence

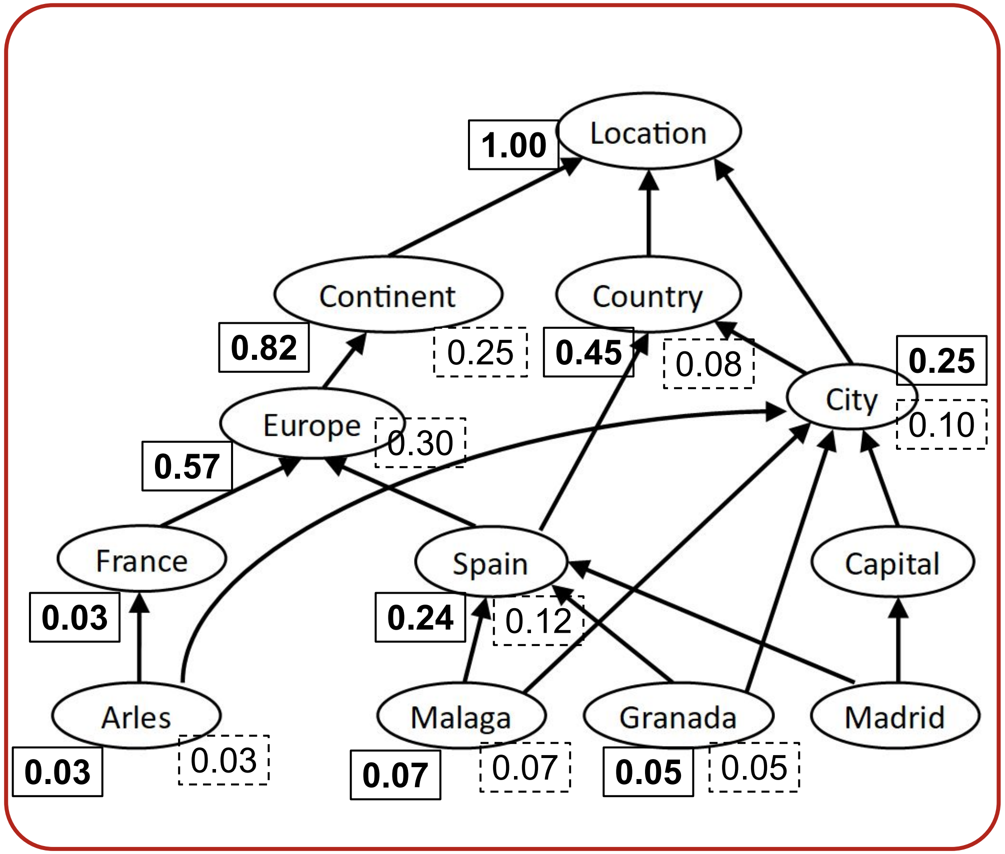
TD-*poset* approach: a practical example

$Sums_{PO}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d+}} t^i(s)$$

$$\text{with } S_{v_d+} = \{s \in S_{v'_d} | v'_d \in V_d \wedge v'_d \preceq v_d\}$$



D = set of data items
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d
S_{v_d+} = set of sources that claim v_d or more specific values (that implicitly support v_d)

trustworthiness of source that provides a value (dashed box)

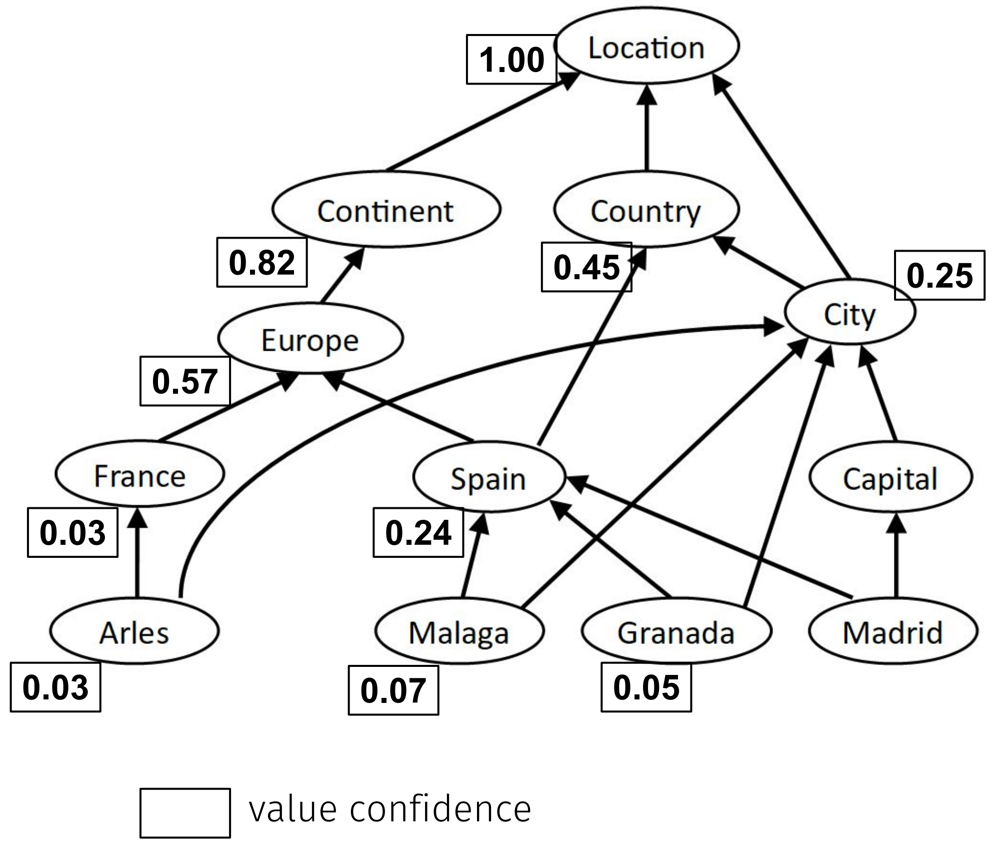
value confidence (solid box)

TD-*poset* approach: which are the consequences of considering partial order of values within TD?

Confidence **estimations monotonically increase** with respect to the partial order of values

The **highest confidence** is always assigned to the **most general value**; it is always supported by all the others

The truth consists in a **true value set** and not in a single value anymore, i.e. all generalizations of a true value are true as well



Truth prediction when considering value dependencies

CLAIMS

TRUTH DISCOVERY

1. Estimation phase

2. Truth prediction

2

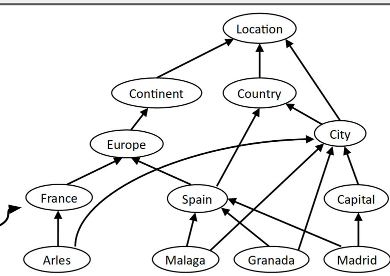
FACTS

DATA ITEM DEPENDENCIES

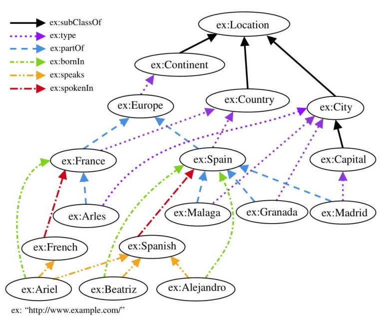
speaks(x,z) ∧ *spokenIn*(z,y) →*bornIn*(x,y)

resident(x, France) → *bornIn*(x,France)

VALUE DEPENDENCIES



ONTOLOGY



Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

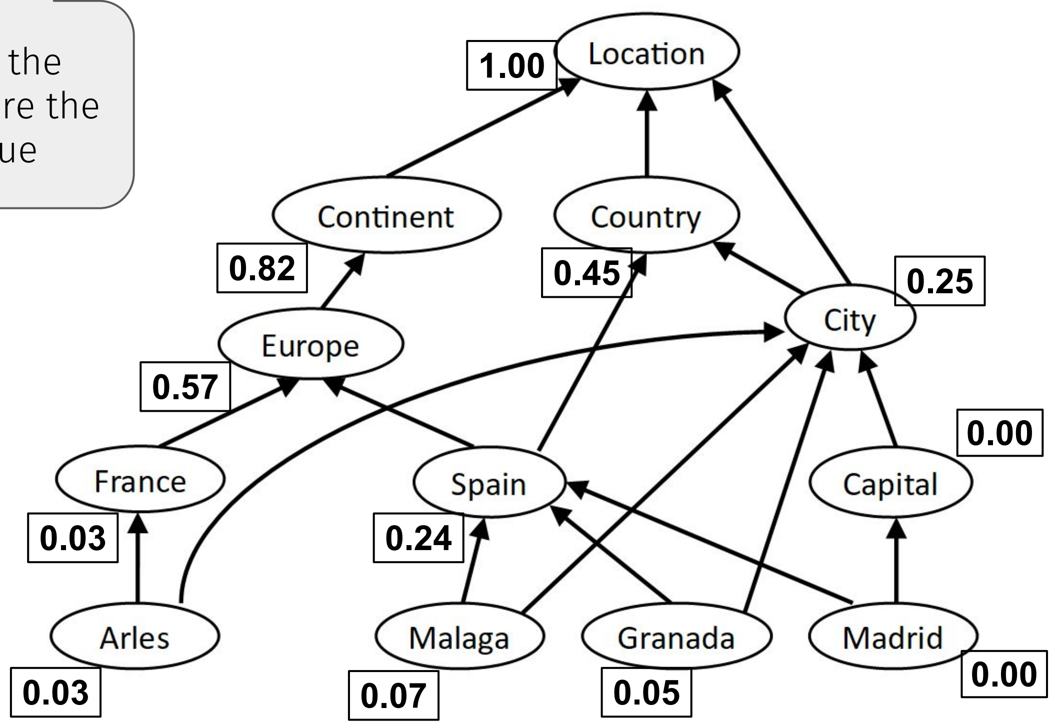
Selection phase ❯ Ranking phase ❯ Filtering phase

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

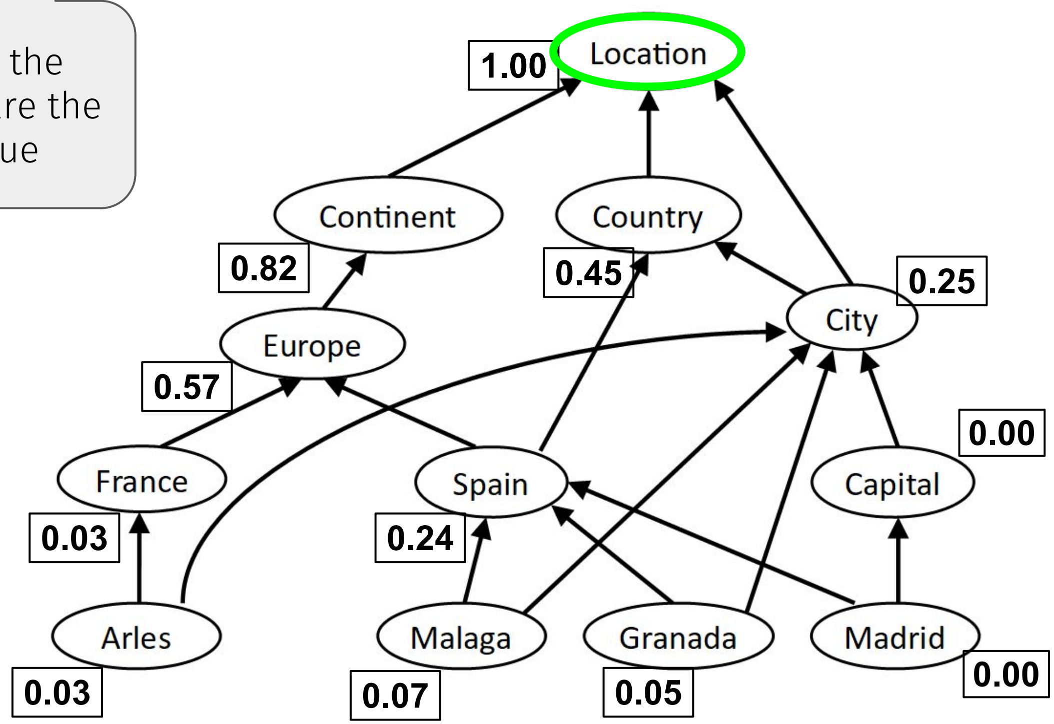


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

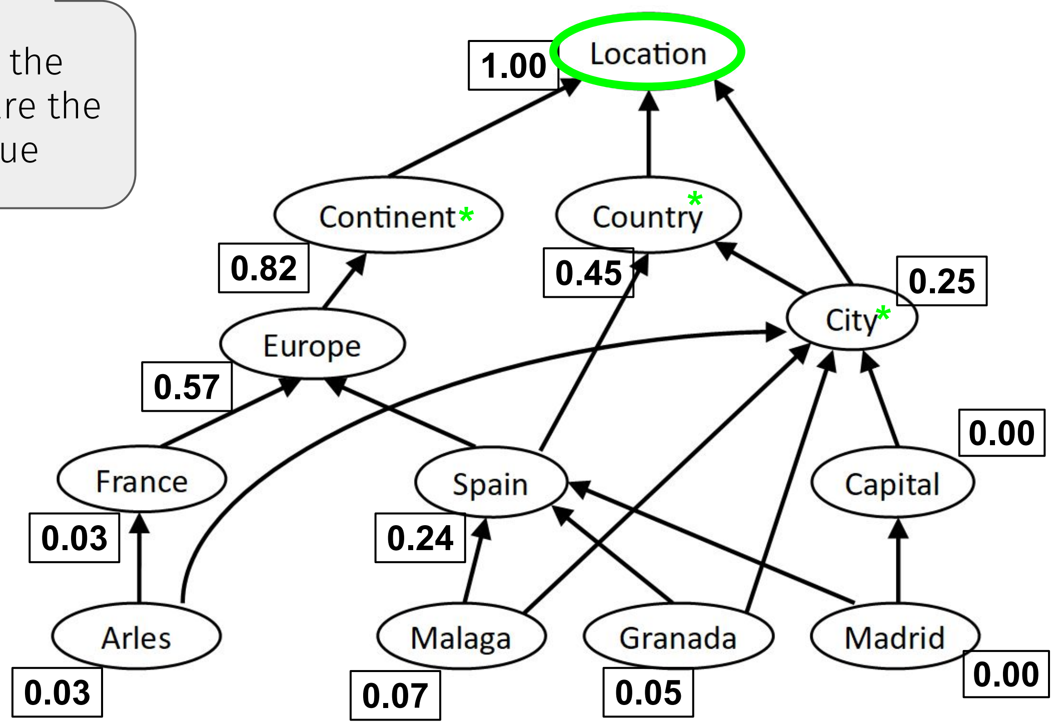


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

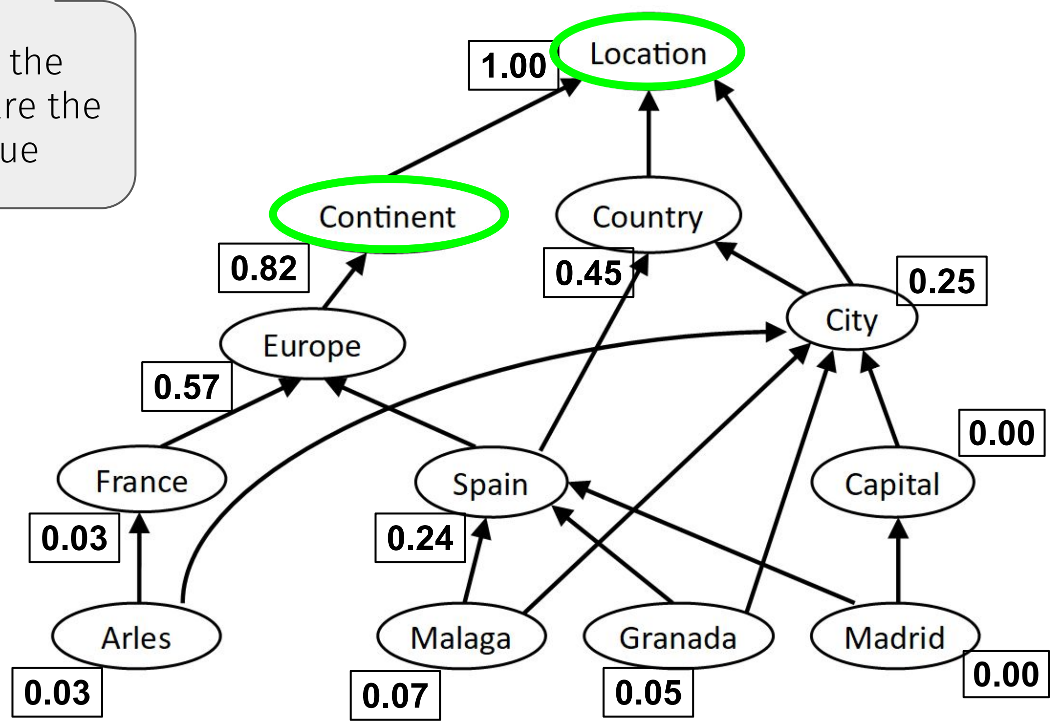


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

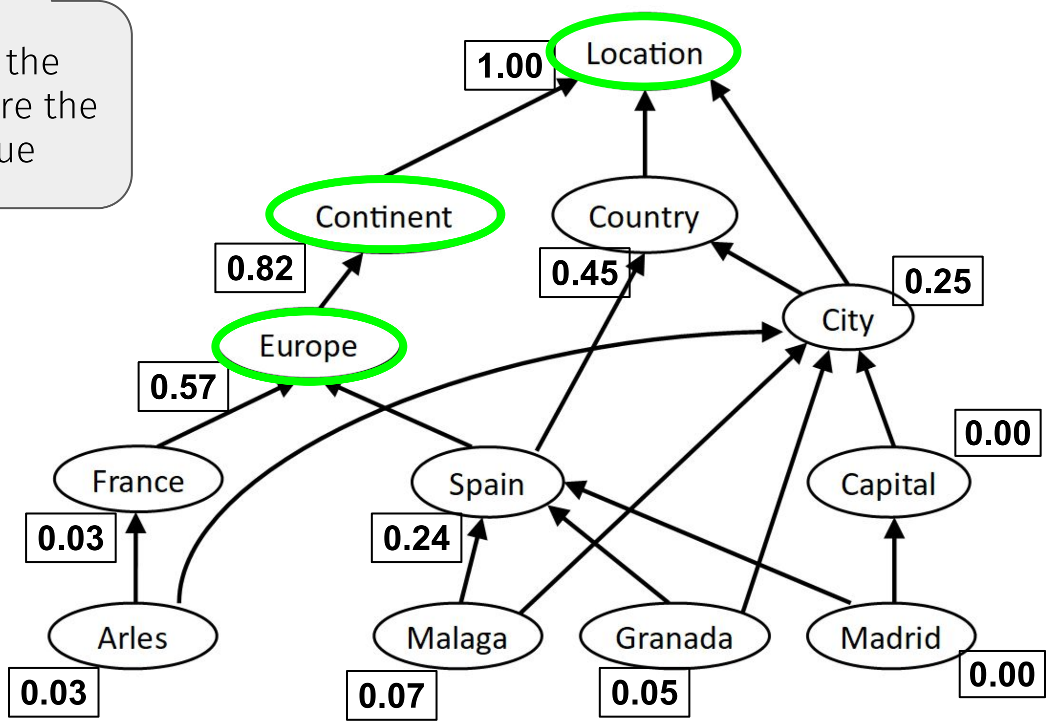


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

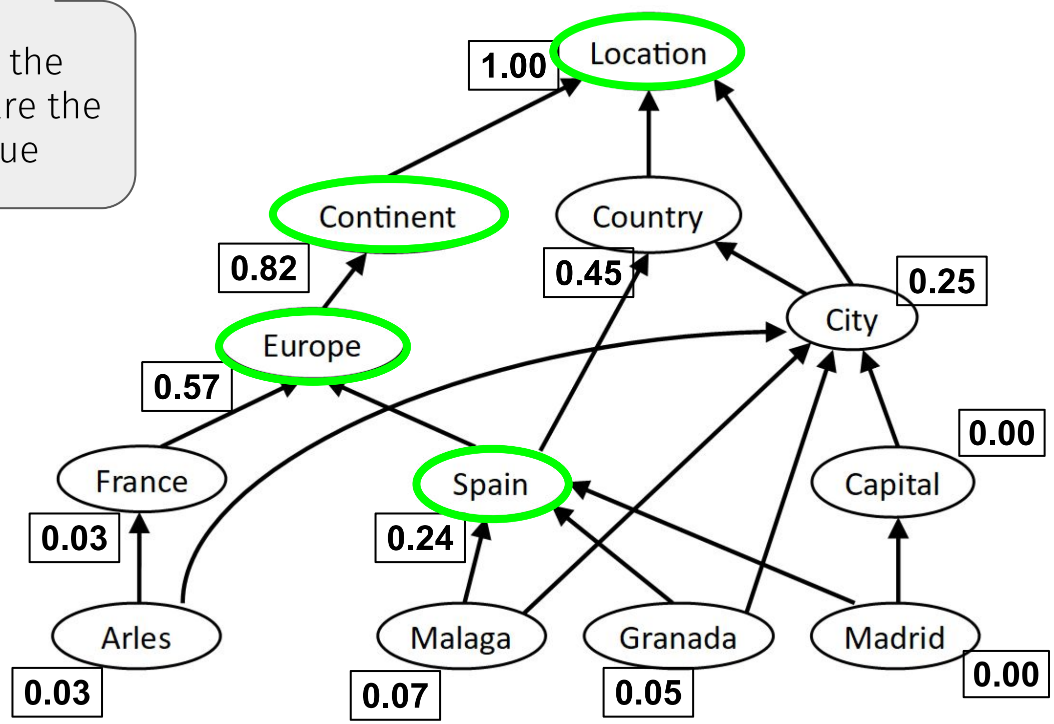


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

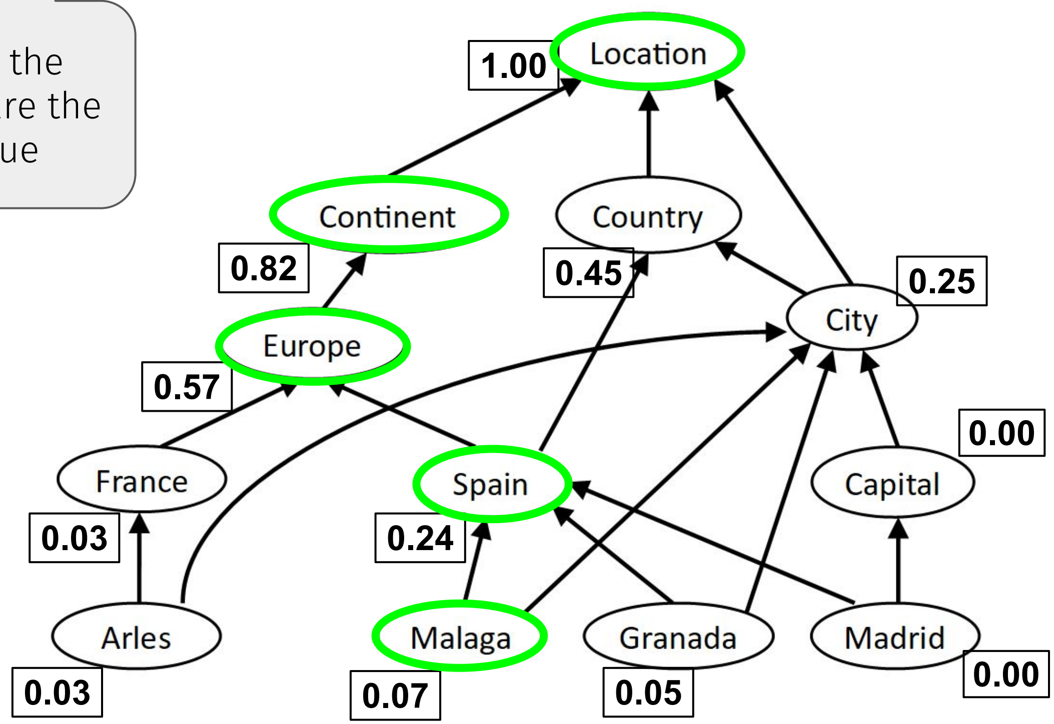


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true

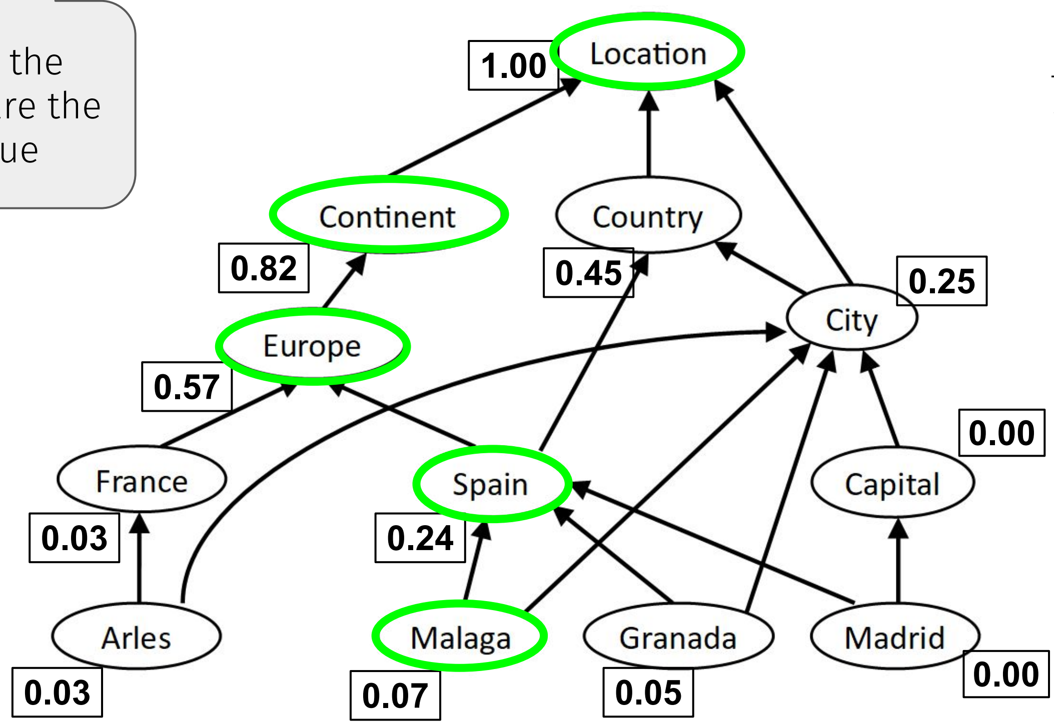


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true



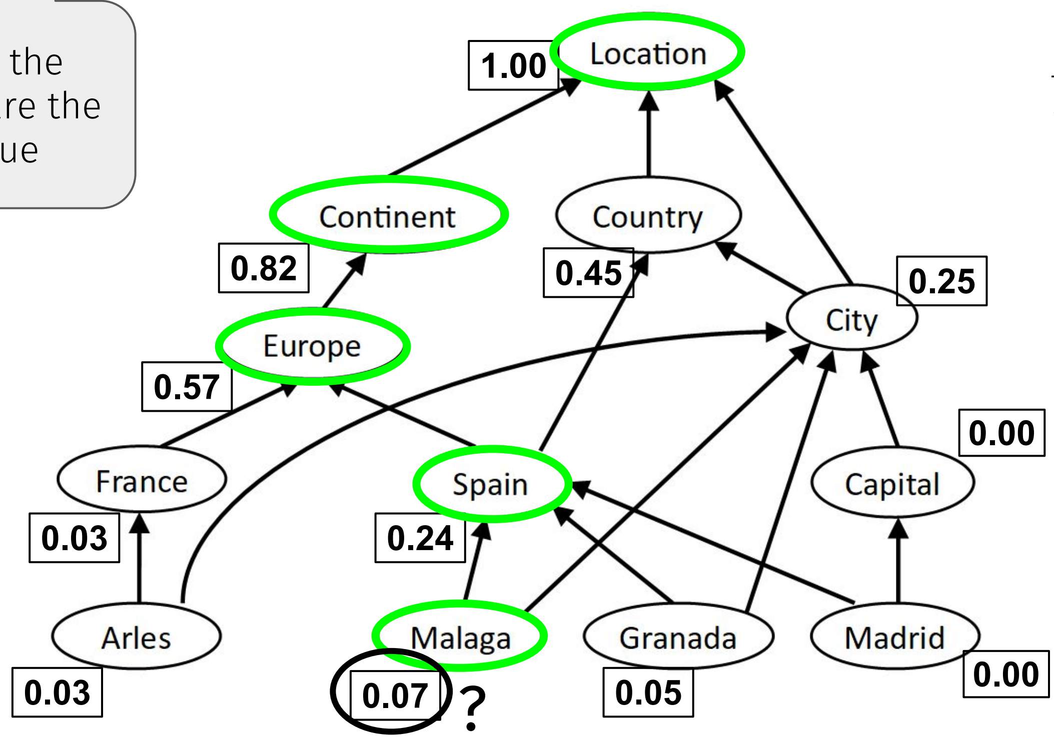
Two thresholds were introduced:

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true



Two thresholds were introduced:

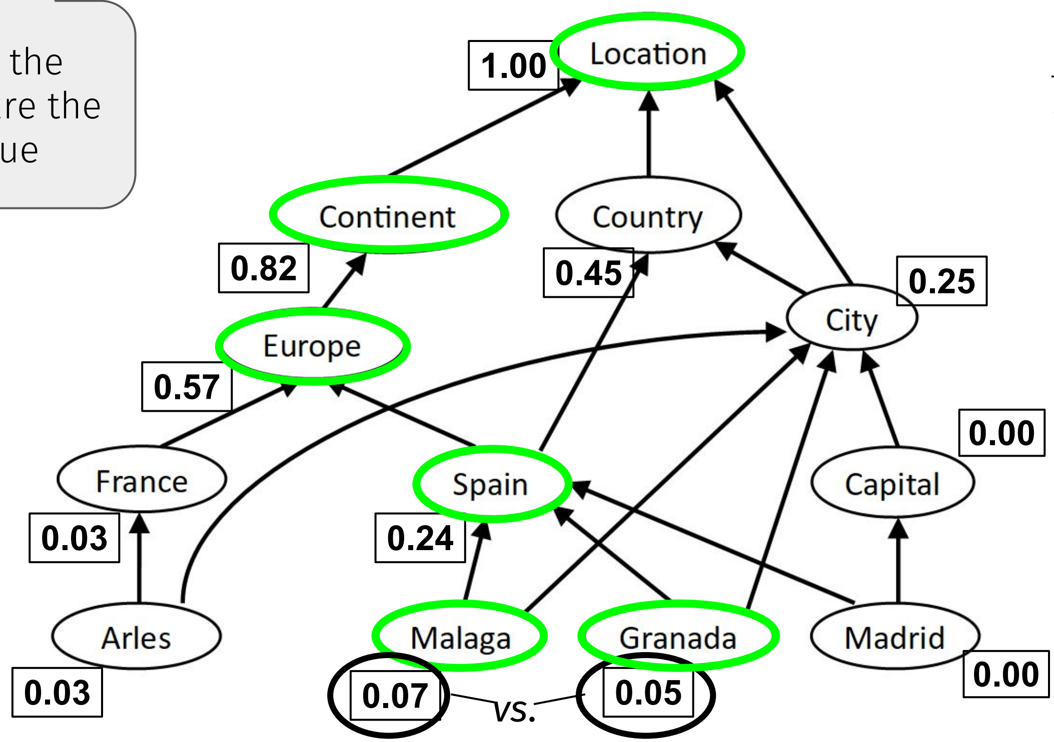
- θ specifies the minimum confidence score that is required for a value to be part of the set of true values

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the set of values that are the most likely to be true



Two thresholds were introduced:

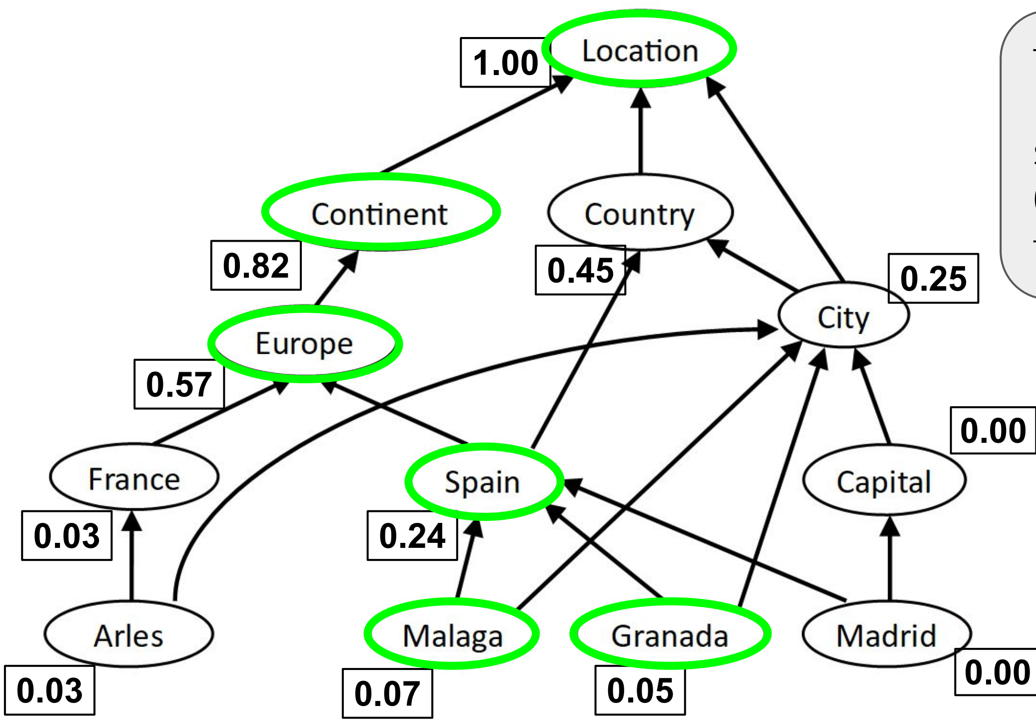
- θ specifies the minimum confidence score that is required for a value to be part of the set of true values
- δ represents the maximum admitted difference between the highest confidence and the confidence of any other selected values.

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

The selected values are ordered based on pre-defined criteria such as Information Content (IC) and source average trustworthiness (WA_{trust}).

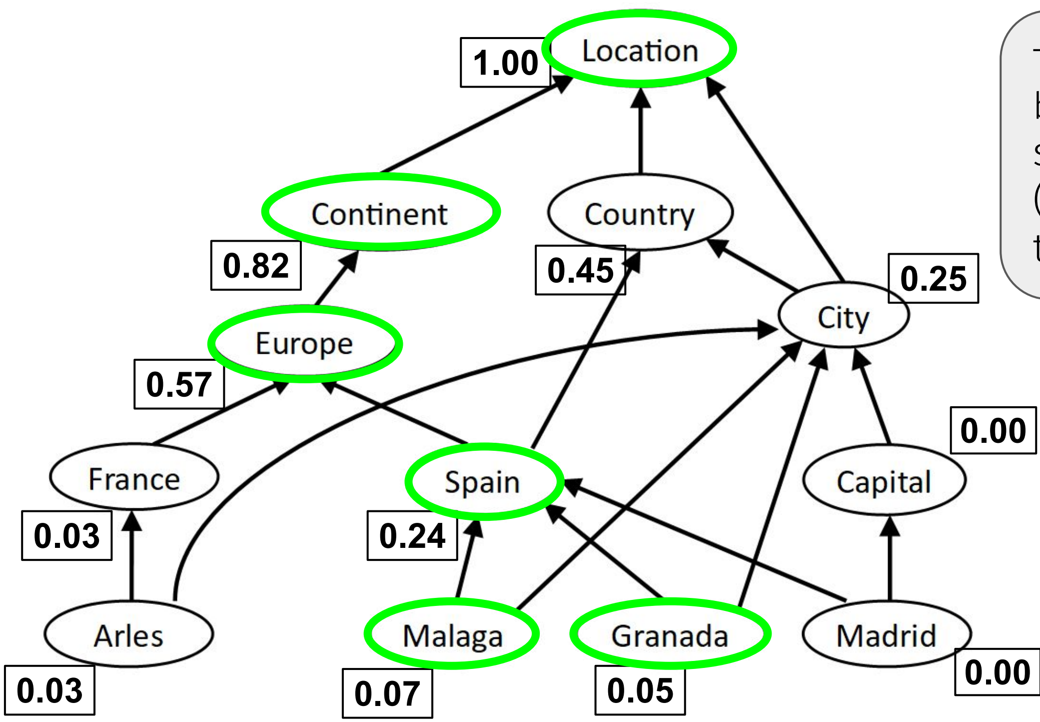


Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

The selected values are ordered based on pre-defined criteria such as Information Content (IC) and source average trustworthiness (WA_{trust}).



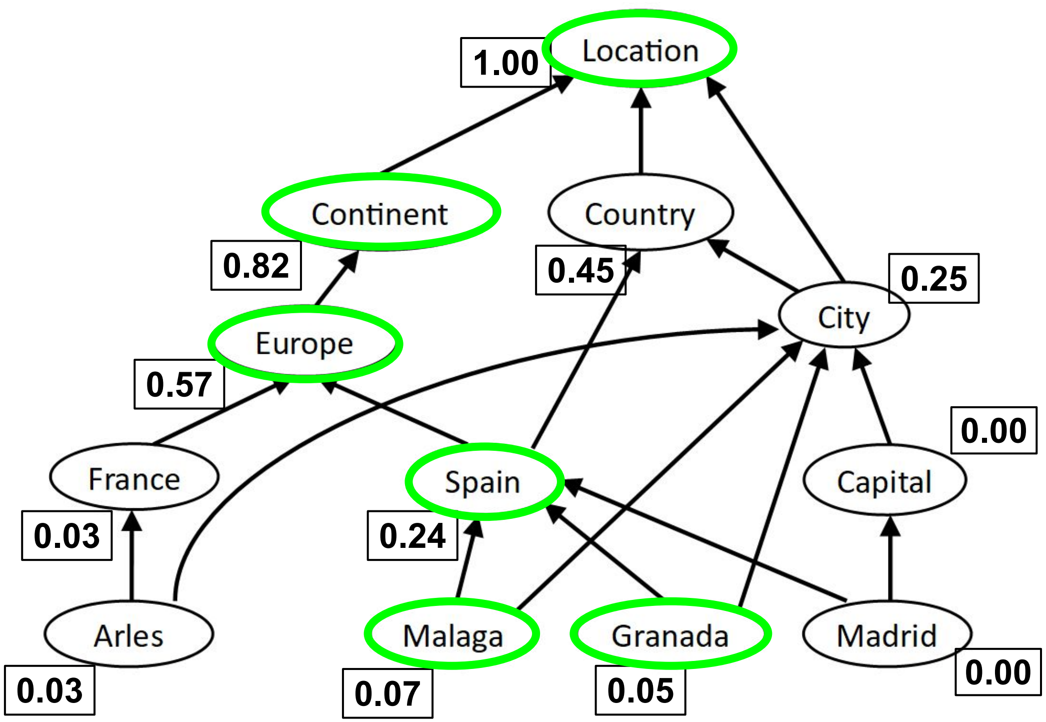
Rank based on IC:

1. Malaga
2. Granada (same IC than Malaga, but lower value confidence)
3. Spain
4. Europe
5. Continent
6. Location

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase



It aims at selecting the top-*k* true values that respect predefined properties such as selection of ordered values only or unordered values only.

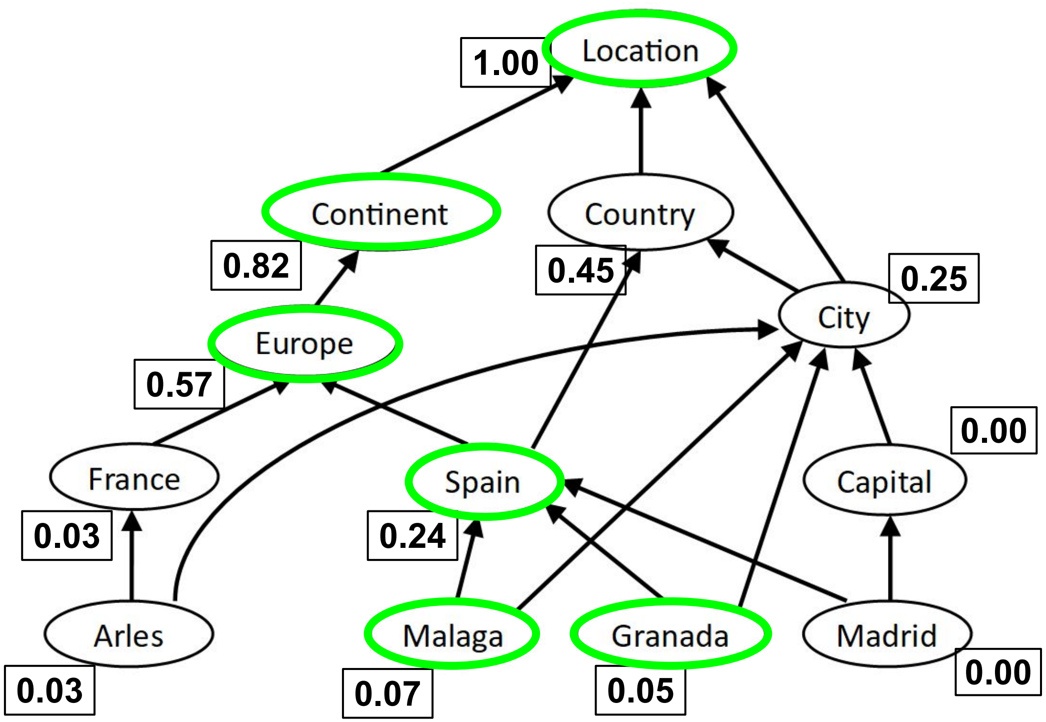
Rank based on IC:

1. Malaga
2. Granada
3. Spain
4. Europe
5. Continent
6. Location

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase



It aims at selecting the top-*k* true values that respect predefined properties such as selection of <u>ordered values</u> only or unordered values only.

Rank based on IC:

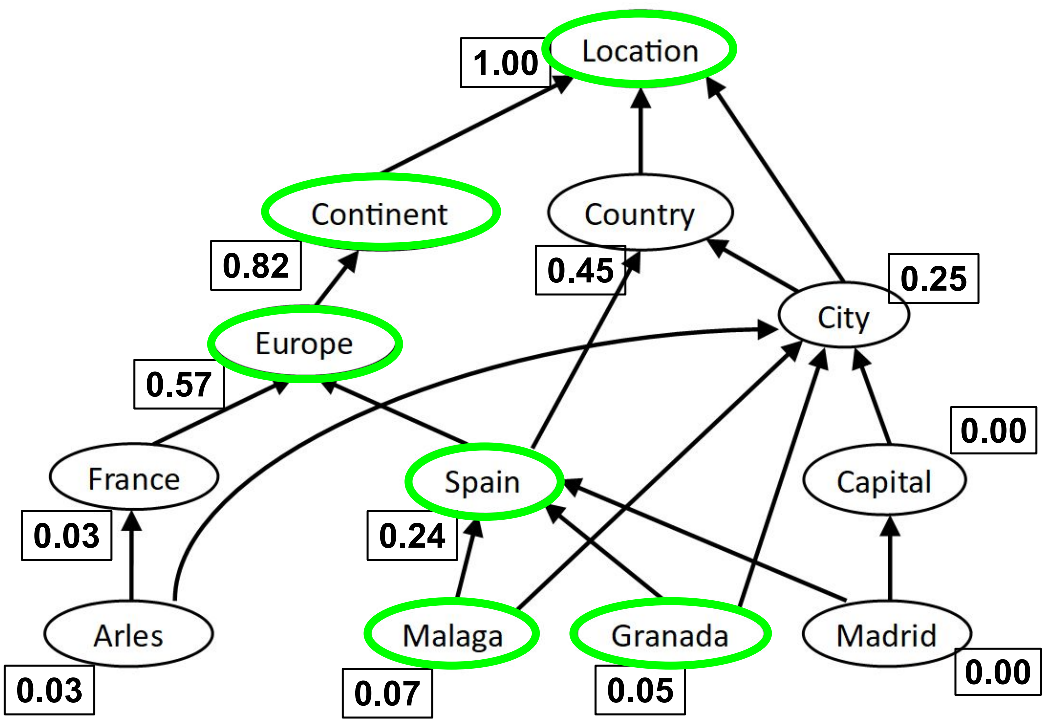
1. Malaga
2. ~~Granada~~
3. Spain
4. Europe
5. Continent
6. Location

Truth Prediction using partial order of values

2. Truth Prediction
Selection of expected true values

Selection phase ❯ Ranking phase ❯ Filtering phase

It aims at selecting the top-*k* true values that respect predefined properties such as selection of ordered values only or unordered values only.



Rank based on IC:

1. Malaga
2. ~~Granada~~
3. Spain
4. Europe
5. Continent
6. Location

Rank based on IC:

1. Malaga
2. Granada
3. ~~Spain~~
4. ~~Europe~~
5. ~~Continent~~
6. ~~Location~~

Experiments: generation of synthetic datasets

The generation of synthetic datasets is controlled by the several parameters, e.g. source coverage, source trustworthiness, ...).

The parameter that permits to generate different kinds of datasets is:

- Granularity of the provided true value

Coarse-grained	Location
↑	Europe
	Spain
fine-grained	Malaga

Experiments: generation of synthetic datasets

The generation of synthetic datasets is controlled by the several parameters, e.g. source coverage, source trustworthiness, ...).

The parameter that permits to generate different kinds of datasets is:

- <ins>Granularity of the provided true value</ins>

Coarse-grained	Location
	Europe
	Spain
fine-grained	Malaga

Considering a given predicate, sources can be:

- **EXPERTS** on a topic
  - they often provide specific (fine-grained) true values since they well known the topic, e.g. an art expert should know the city of birth of Picasso
- **NON-EXPERTS** on a topic
  - they provide fine-grained true values when they known the information, but they provide general true values rather than false values when they do not know the topic, e.g. I known the exact birth location of Picasso even if I am not an art expert

Experiments: generation of synthetic datasets

The generation of synthetic datasets is controlled by the several parameters, e.g. source coverage, source trustworthiness, ...).

The parameter that permits to generate different kinds of datasets is:

- Granularity of the provided true value

Coarse-grained	Location
	Europe
	Spain
fine-grained	Malaga

Considering a given predicate, sources can be:

- **EXPERTS** on a topic
  - they often provide specific (fine-grained) true values since they well known the topic, e.g. an art expert should know the city of birth of Picasso
- **NON-EXPERTS** on a topic
  - they provide fine-grained true values when they known the information, but they provide general true values rather than false values when they do not know the topic, e.g. I known the exact birth location of Picasso even if I am not an art expert

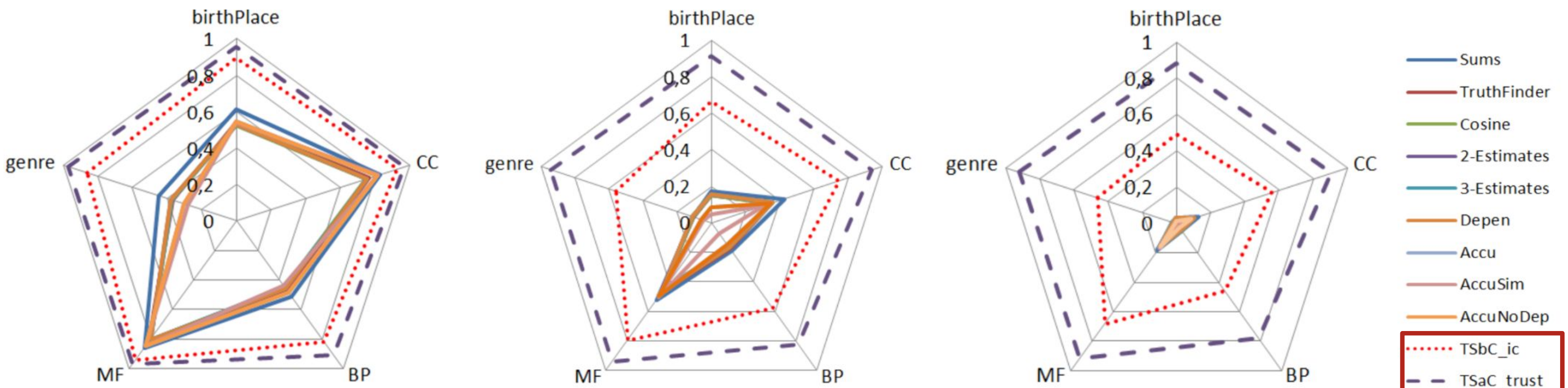
EXPERTS | EXPERTS and NON-EXPERTS | NON-EXPERTS

Experiments: Recall of $Sums_{PO}$ on synthetic datasets

EXPERTS

EXPERTS and NON-EXPERTS

NON-EXPERTS



	1. Estimations	2. Truth Prediction		
		Selection phase	Ranking phase	Filtering phase
$TSbC_{IC}$	$Sums_{PO}$	θ = 0.0 et δ= 0.0	IC used as ranking criterion	returning ordered values
$TSaC_{TRUST}$		θ = 0.0 et δ= 1.0	WA_{trust} used as ranking criterion	returning not ordered values

$Sums_{PO}$: what have we learned?

Considering *a priori* knowledge in the form of a **partial order of values,** we exploit the **deductive reasoning** capabilities offered by ontologies:

- different values are not always independent
- a partial order of values enables to distinguish when two different values have similar semantics
- a partial order of values is useful and permits to make performance more robust independently from the type of sources that is considered

Enhancing confidence estimations using value dependencies

CLAIMS

TRUTH DISCOVERY

1. Estimation phase

2. Truth prediction

FACTS

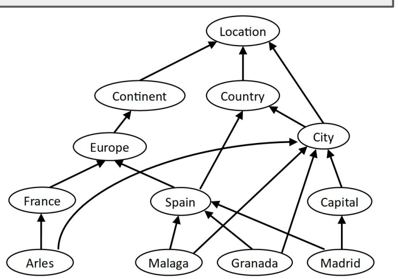
DATA ITEM DEPENDENCIES

3

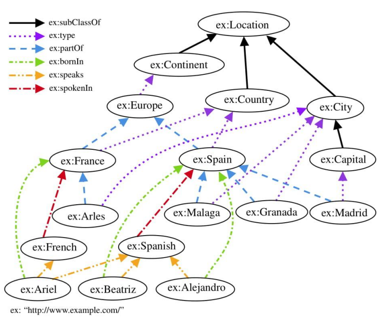
$speaks(x,z) \wedge spokenIn(z,y) \rightarrow bornIn(x,y)$

$resident(x, France) \rightarrow bornIn(x, France)$

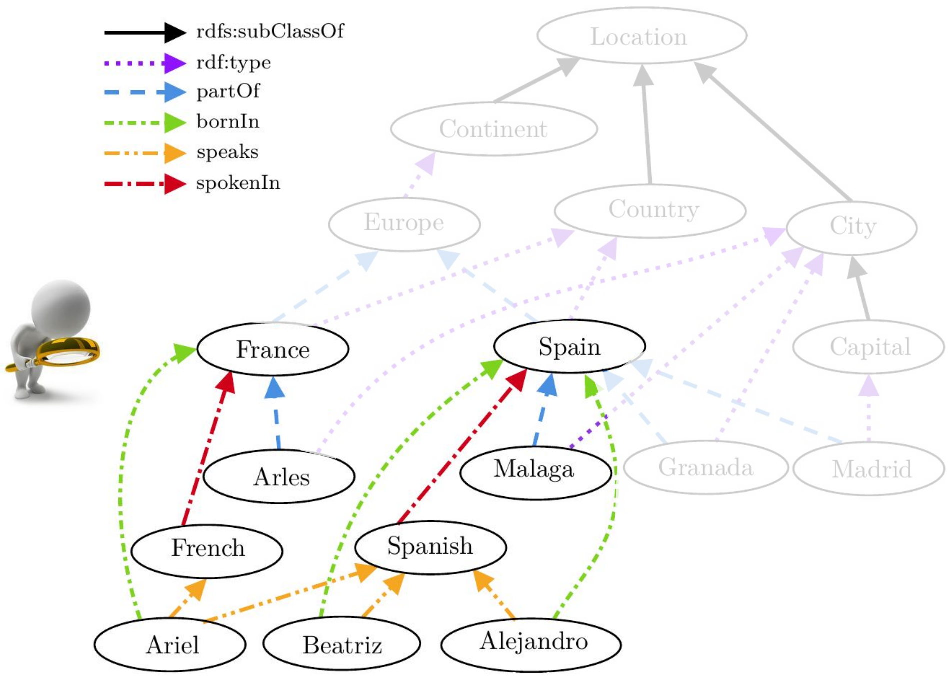
VALUE DEPENDENCIES



ONTOLOGY



How to exploit data item dependencies?



People speaking the official language of a country were usually born in that country.

This knowledge can be used to increase the confidence about some information.

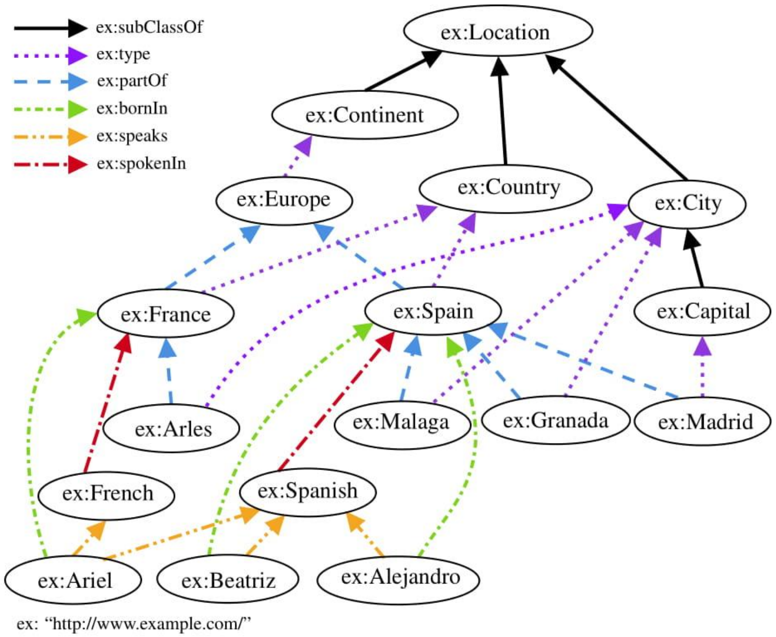
How data item dependencies can be modeled?

Recurrent patterns are modeled by **rules** that are inferred from an ontology.

A **rule** *r* is an implication from a set of atoms called body to a set of atoms called head:

$$r : B_1 \wedge B_2 \wedge \ldots \wedge B_{|B|} \rightarrow H_1 \wedge H_2 \wedge \ldots \wedge H_{|H|}$$

It indicates that when observing some conditions (reported in the body), the occurrence of other conditions (reported in the head) is expected.



speaks(x,y)∧ spokenIn(y,z)→ bornIn(x,z)

How to exploit rules to improve TD?

Similar entity subjects should have similar property values.

ONTOLOGY → recurrent patterns expressed by rules → increasing confidence of values supported by rules

AMIE+ (Galárraga et al, 2015)

`speaks(x,y)` ∧ `spokenIn(y,z)` → `bornIn(x,z)`

The rule body indicates on which basis (which properties and values) entities are considered to be similar.

How to exploit rules to improve TD?

Similar entity subjects should have similar property values.

ONTOLOGY → recurrent patterns expressed by rules → increasing confidence of values supported by rules

AMIE+ (Galárraga et al, 2015)

$$\texttt{speaks(x,y)} \land \texttt{spokenIn(y,z)} \rightarrow \texttt{bornIn(x,z)}$$

The rule body indicates on which basis (which properties and values) entities are considered to be similar.

Source ($s \in S$)	Data item ($d \in D$)		Value ($v \in V$)
	Subject	Predicate	
A	Pablo Picasso,	bornIn	Spain
B	Pablo Picasso,	bornIn	Madrid
C	Pablo Picasso,	bornIn	Europe
A	Claude Monet,	bornIn	Málaga
B	Claude Monet,	bornIn	Arles
...	...	...	...

Example: If we known that Picasso speaks Spanish, then the confidence of value Spain increases

TD using Rules: TDR approach

Sums

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

Sums$_{RULES}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}_{RULES}(v_d)$$

$$c^i_{RULES}(v_d) = \frac{1}{norm_{v_d}} \left[(1-\gamma) c^i(v_d) + \gamma \, boost(d, v_d) \right]$$

ontology point of view

source point of view

D = set of claims provided by source
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

TD using Rules: TDR approach

Sums

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}(v_d)$$

$$c^i(v_d) = \beta \sum_{s \in S_{v_d}} t^i(s)$$

Sums$_{RULES}$

$$t^i(s) = \alpha \sum_{v_d \in V_s} c^{i-1}_{RULES}(v_d)$$

$$c^i_{RULES}(v_d) = \frac{1}{norm_{v_d}} \left[(1-\gamma)c^i(v_d) + \gamma\, \underline{boost(d, v_d)}\right]$$

$$boost(d, v_d) \approx \frac{\sum_{r \in R^v_d} score(r)}{\sum_{r \in R_d} score(r)}$$

D = set of claims provided by source
V_s = set of claims provided by source s
S_{v_d} = set of sources that claim v_d

TD using Rules: which are the rules to consider?

The increase of confidence must be proportional to the percentage of rules that support the provided value (called **approving rules**) among the considered rules (called **eligible rules**).

Rules that support the considered claim

Rules that can potentially support the considered claim since they concern similar data item subject

TD using Rules: which are the rules to consider?

The increase of confidence must be proportional to the percentage of rules that support the provided value (called **approving rules**) among the considered rules (called **eligible rules**).

Rules that support the considered claim

Rules that can potentially support the considered claim since they concern similar data item subject

Example: evaluating the confidence of *bornIn(Pablo Picasso, Spain)*

K

...
spokenIn(Spanish, Spain)
speaks(Picasso,Spanish)
resident(Picasso,France)
...

r_1: *speaks*(x,y) ∧ *spokenIn(y,z)* → *bornIn*(x,z)

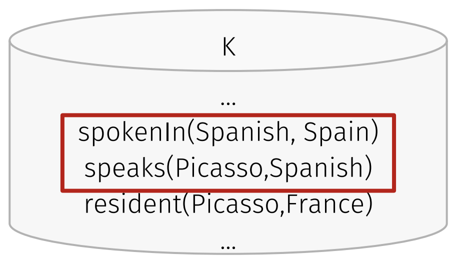
TD using Rules: which are the rules to consider?

The increase of confidence must be proportional to the percentage of rules that support the provided value (called **approving rules**) among the considered rules (called **eligible rules**).

Rules that support the considered claim

Rules that can potentially support the considered claim since they concern similar data item subject

Example: evaluating the confidence of *bornIn(Pablo Picasso, Spain)*



r_1: *speaks*(x,y) $\wedge$ *spokenIn(y,z)* $\rightarrow$ *bornIn*(x,z)

ELIGIBLE RULE

TD using Rules: which are the rules to consider?

The increase of confidence must be proportional to the percentage of rules that support the provided value (called **approving rules**) among the considered rules (called **eligible rules**).

Rules that support the considered claim

Rules that can potentially support the considered claim since they concern similar data item subject

Example: evaluating the confidence of *bornIn(Pablo Picasso, Spain)*

K

...

~~spokenIn(Spanish, Spain)~~

speaks(Picasso,Spanish)

resident(Picasso,France)

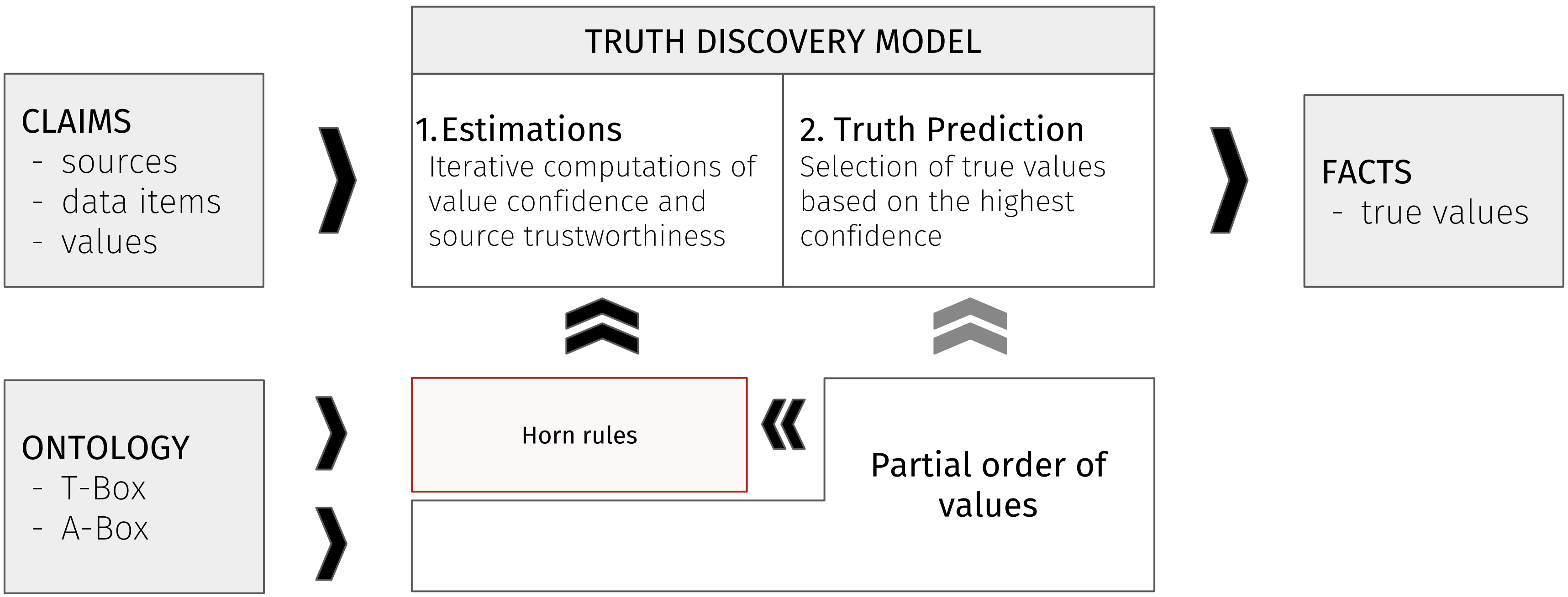
...

r_1: *speaks*(x,y) ∧ *spokenIn(y,z)* → *bornIn*(x,z)

NOT ELIGIBLE RULE

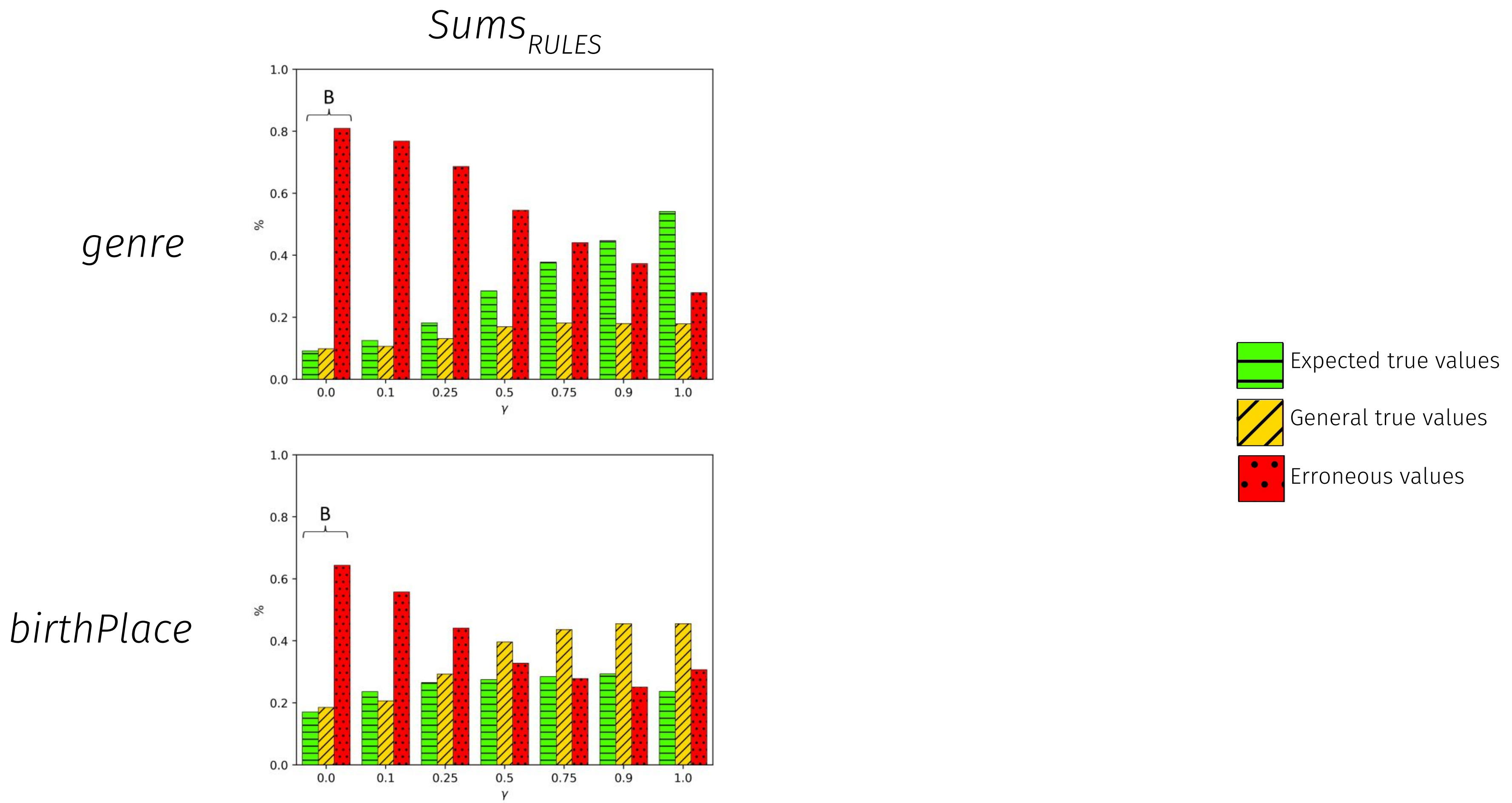
the body is not verified when instantiating the variables with respect to the subject *Pablo Picasso*

How to combining value partial order and rules to improve TD?



The rule ***resident***(x, France) → ***bornIn***(x,France) explicitly support the value France.
It implicitly supports the value Europe and all the other generalizations.

Experiments: Results of $Sums_{RULES}$ and $Sums_{RULES\&PO}$ on synthetic datasets

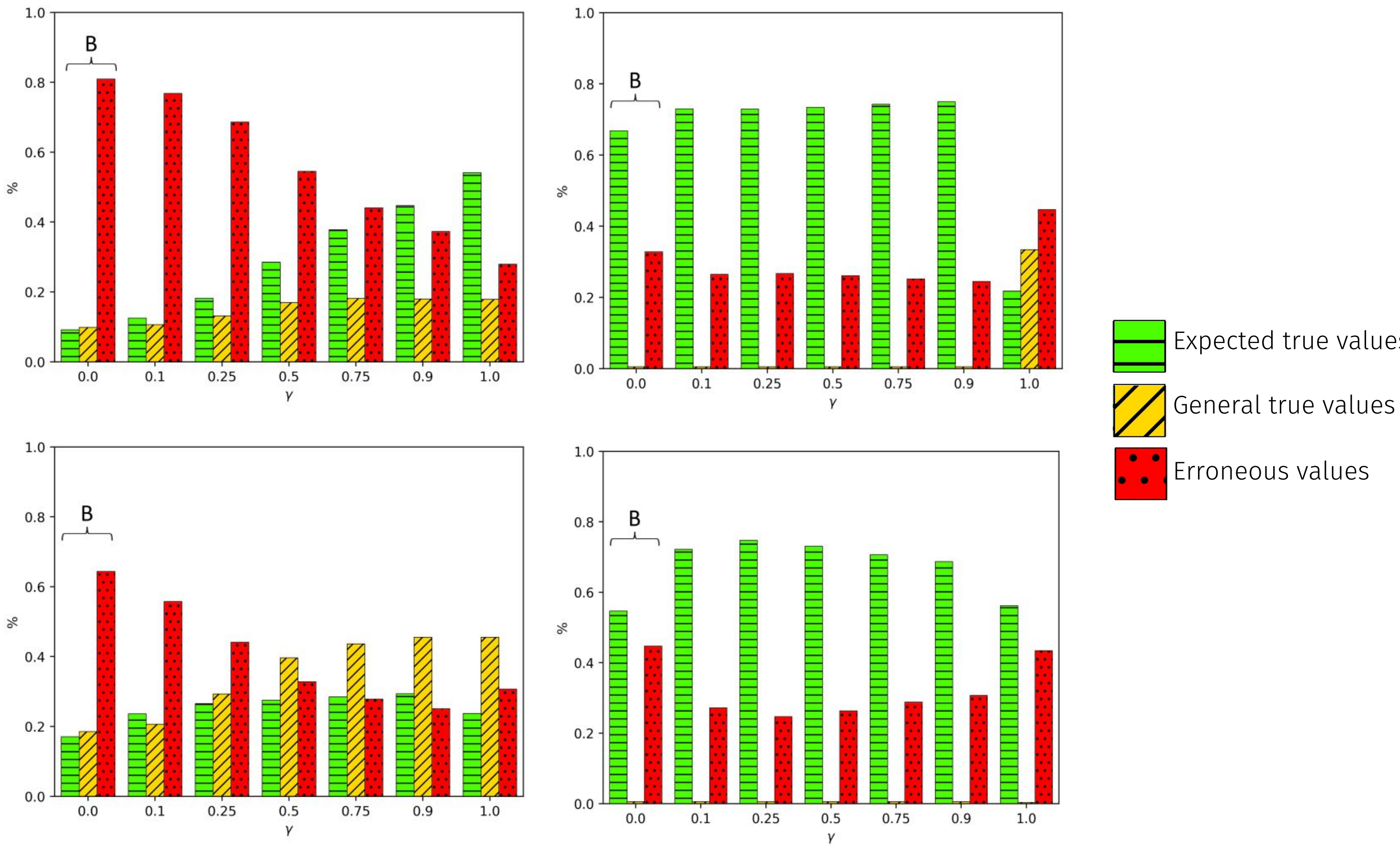


Experiments: Results of $Sums_{RULES}$ and $Sums_{RULES\&PO}$ on synthetic datasets

$Sums_{RULES}$ $Sums_{RULES\&PO}$

genre

birthPlace



$Sums_{RULES}$ and $Sums_{RULES\&PO}$: what have we learned?

Considering *a priori* knowledge in the form of **rules** exploits **inductive reasoning** capabilities offered by ontologies:

- Rules help to identify similar data items and identify the most probable true value for them
- Rules are more effective when the IC of the values they infer is high

A case study on real-world data

4

CLAIMS

TRUTH DISCOVERY

1. Estimation phase

2. Truth prediction

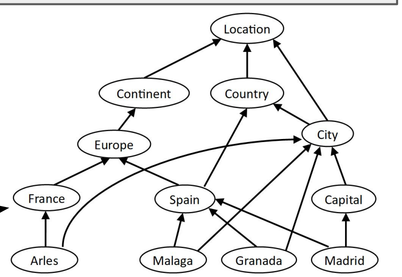
FACTS

DATA ITEM DEPENDENCIES

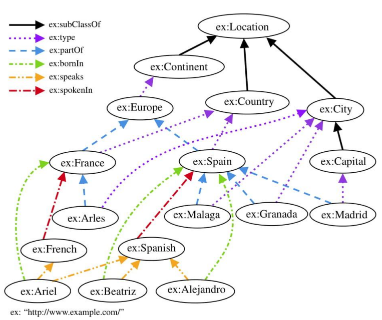
speaks(x,z) ∧ *spokenIn*(z,y) → *bornIn*(x,y)

resident(x, France) → *bornIn*(x,France)

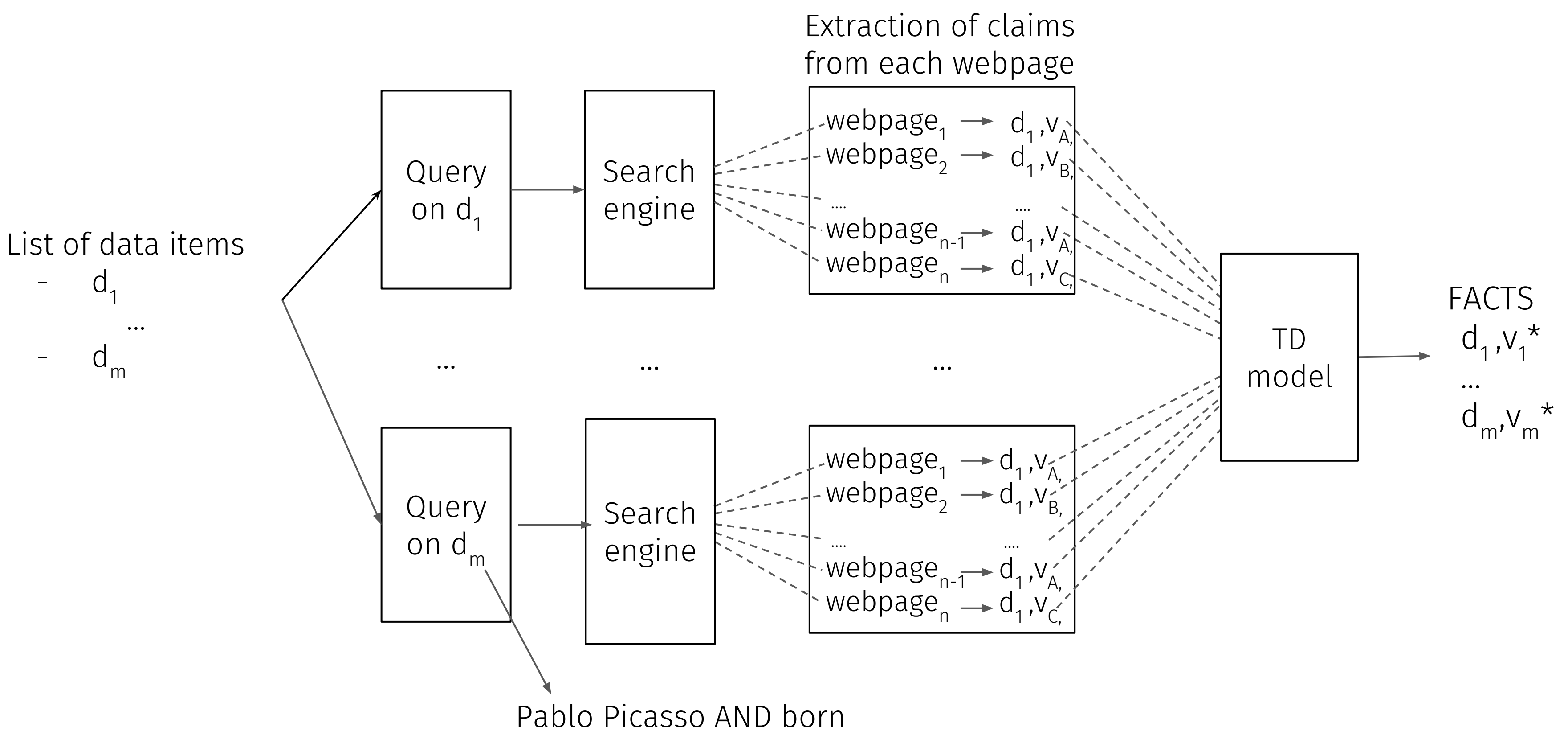
VALUE DEPENDENCIES



ONTOLOGY



Real-world datasets: application context



Recall obtained on real-world data

Sums...	Data A	Data B
Original model	*0.448*	*0.473*
with partial ordering	0.517	0.566
with partial ordering and rules	0.565	0.590
with partial ordering and rules + post-processing	**0.631**	**0.614**

$Sums_{RULES\&PO}$ on a real world scenario: what have we learned?

- *Sums* rewards sources having high coverage and penalizes sources having low coverage;
- It makes a distinction between **reliable sources** (that always provide true values) of **different coverage levels**
  - wikipedia.org (high coverage) is correctly considered a highly reliable source
  - a fan club website (low coverage), specialized on its favorite actor, is incorrectly considered unreliable
- In real-world datasets there are very few sources having high coverage, while the majority of them have a low coverage, i.e power law phenomenon is over expressed.

Recall obtained on real-world data

Sums...	Data A	Data B
Original model	*0.448*	*0.473*
with partial ordering	0.517	0.566
with partial ordering and rules	0.565	0.590
with partial ordering and rules + post-processing	**0.631**	**0.614**

Existing Model	Data A	Data B
Truth Finder	**0.646**	0.622
2-Estimates	0.631	0.635
3-Estimates	0.008	0.612
Cosine	0.640	0.635
AccuCopy	0.638	0.640
Accu	0.638	**0.660**
Depen	0.431	0.494
AccuSim	0.413	0.448
SimpleLCA	0.631	**0.660**
GuessLCA	0.644	0.646

Data veracity assessment: Enhancing Truth Discovery using *a priori* knowledge

Outline:

1. Motivations behind data veracity assessment

2. Truth Discovery: problem and positioning

3. Enhancing Truth Discovery models using *a priori* knowledge

4. Conclusion

4.1 Summary

4.2 Limitations and future studies

Summarising

Enhancing truth discovery models using **value dependencies**

- Value dependencies are modeled using a **partial order of values** extracted from an ontology
- Confidence estimations formulas are modified taking this additional information into account
  - Adaptation of *Sums* approach
- Definition of a new **truth prediction phase** that is able to select the expected true value

Summarising

Enhancing truth discovery models using **value dependencies**

- Value dependencies are modeled using a **partial order of values** extracted from an ontology
- Confidence estimations formulas are modified taking this additional information into account
  - Adaptation of *Sums* approach
- Definition of a new **truth prediction phase** that is able to select the expected true value

Enhancing truth discovery models using **data item dependencies**

- Data item dependencies are modeled using **rules** extracted from an ontology
- Confidence estimations formulas are modified taking this additional information into account, as well as partial order of values
  - Adaptation of *Sums* approach

Summarising

Enhancing truth discovery models using **value dependencies**

- Value dependencies are modeled using a **partial order of values** extracted from an ontology
- Confidence estimations formulas are modified taking this additional information into account
  - Adaptation of *Sums* approach
- Definition of a new **truth prediction phase** that is able to select the expected true value

Enhancing truth discovery models using **data item dependencies**

- Data item dependencies are modeled using **rules** extracted from an ontology
- Confidence estimations formulas are modified taking this additional information into account, as well as partial order of values
  - Adaptation of *Sums* approach

Evaluation of the proposed approaches on **synthetic and real-world dataset**

- Definition of protocols to generate synthetic datasets and collect real-world data
- Experiments conducted using the several proposed models

Limitations and future studies

Some interesting further studies are:

- incorporating the proposed rationale in other existing models
- adapting the proposed models in order to deal with non-functional and dynamic predicates
- assessing the reliability of the considered *a priori* knowledge
- proposing a solution to automatically estimate the best configuration of model parameters, such as the weight that controls the importance given to rules during the confidence estimation phase
- proposing models that deal with the over expression of power-law phenomena in real-world datasets

Data veracity assessment: Enhancing Truth Discovery using a priori knowledge

Take-away messages:

- All tasks/applications that consume Web data require to check data veracity automatically
- We propose to incorporate into the existing models prior knowledge in the form of partial order among values and rules
- The proposed approaches open up new research perspectives that should be explored in order to make data increasingly reliable