Semantic Data Enrichment meets Neural-Symbolic Integration

Vincenzo Cutrona

University of Milano - Bicocca
vincenzo.cutrona@unimib.it

City, University of London
Vincenzo.Cutrona@city.ac.uk

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

DISCo
DIPARTIMENTO
DI INFORMATICA
SISTEMISTICA
E COMUNICAZIONE

CITY
UNIVERSITY OF LONDON
EST 1894

Problem Statement: Semantic Data Enrichment

Inputs:

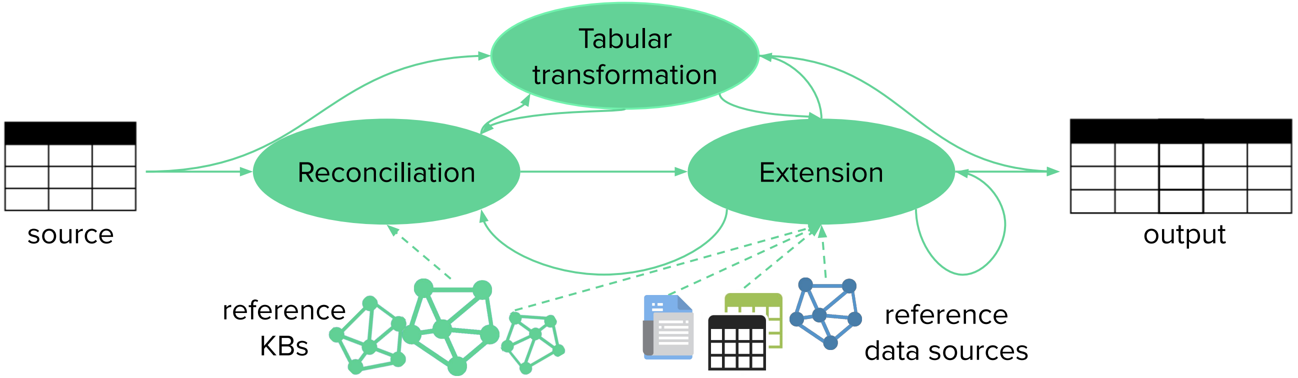
- a **source** dataset
- a pool of **reference data sources**

Output:

- source dataset extended with more columns

Semantic data enrichment: a walk on the **data transformations** graph G^T where at least one node is a **reconciliation**

$$t_1(s_0), ...t_n(s_{n-1}) \text{ where } s_0 = \text{source dataset and } t_n(s_{n-1}) = \text{target dataset}$$
$$t_i = \text{tabular transformation} \mid \text{reconciliation} \mid \text{extension}$$



Problem Statement: Semantic Data Enrichment

- **large-scale** tabular data
  ➔ big data processing

Inputs:

- a **source** dataset
- a pool of **reference data sources**

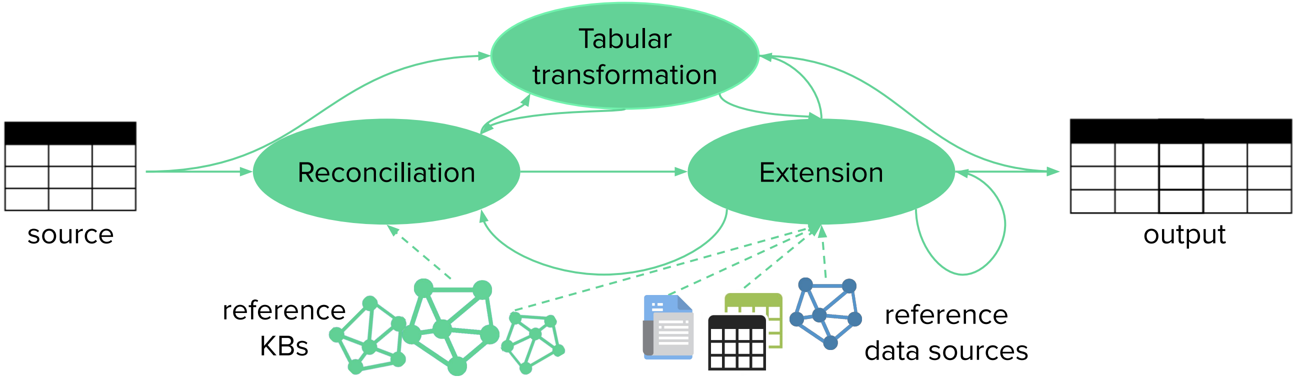
Output:

- source dataset extended with more columns

- against **large-scale** KBs
  ➔ fast execution

Semantic data enrichment: a walk **ns** graph G^T where at least one node is a **reconciliation**

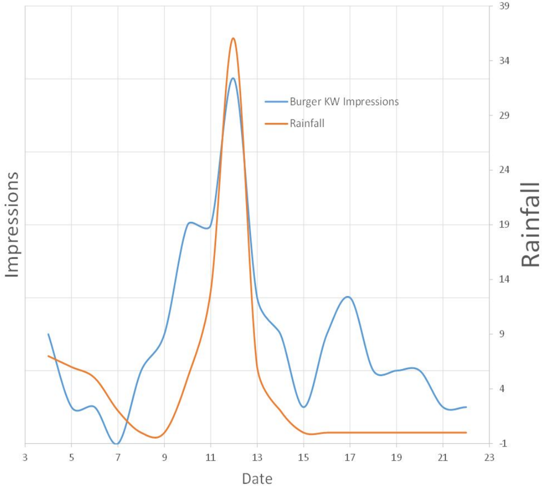
$$t_1(s_0), \ldots t_n(s_{n-1}) \text{ where } s_0 = \text{source dataset and } t_n(s_{n-1}) = \text{target dataset}$$
$$t_i = \text{tabular transformation} \mid \text{reconciliation} \mid \text{extension}$$

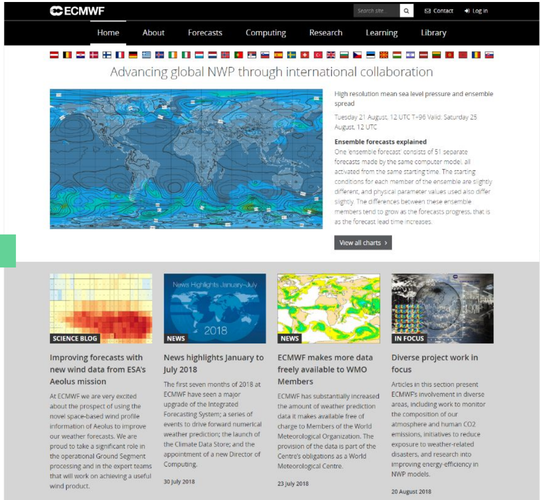


Relevancy: Enrichment for Data Analytics

- Useful in a variety of data science applications based on the analysis

  1) **Weather-aware scheduler for digital marketing campaigns**





Advancing global NWP through international collaboration

High resolution mean sea level pressure and ensemble spread

Tuesday 21 August, 12 UTC T+96 Valid: Saturday 25 August, 12 UTC

Ensemble forecasts explained
One 'ensemble forecast' consists of 51 separate forecasts made by the same computer model, all activated from the same starting time. The starting conditions for each member of the ensemble are slightly different, and physical parameter values used also differ slightly. The differences between these ensemble members tend to grow as the forecasts progress, that is as the forecast lead time increases.

SCIENCE BLOG

Improving forecasts with new wind data from ESA's Aeolus mission

At ECMWF we are very excited about the prospect of using the novel space-based wind profile information of Aeolus to improve our weather forecasts. We are proud to take a significant role in the operational Ground Segment processing and in the expert teams that will work on achieving a useful wind product.

NEWS

News highlights January to July 2018

The first seven months of 2018 at ECMWF have seen a major upgrade of the Integrated Forecasting System; a series of events to drive forward numerical weather prediction; the launch of the Climate Data Store; and the appointment of a new Director of Computing.

30 July 2018

NEWS

ECMWF makes more data freely available to WMO Members

ECMWF has substantially increased the amount of weather prediction data it makes available free of charge to Members of the World Meteorological Organization. The provision of the data is part of the Centre's obligations as a World Meteorological Centre.

23 July 2018

IN FOCUS

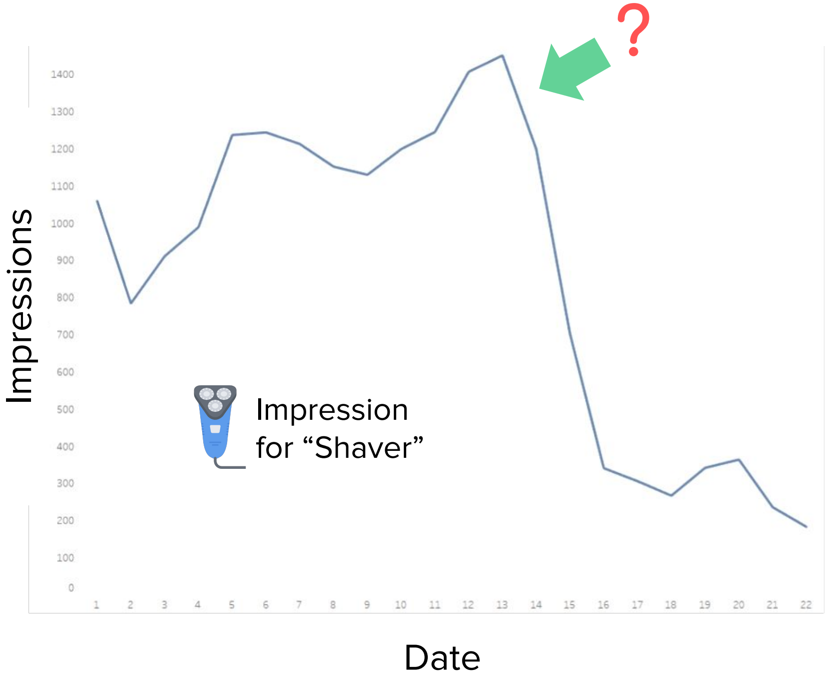
Diverse project work in focus

Articles in this section present ECMWF's involvement in diverse areas, including work to monitor the composition of our atmosphere and human CO2 emissions, initiatives to reduce exposure to weather-related disasters, and research into improving energy-efficiency in NWP models.

20 August 2018

Relevancy: Enrichment for Data Analytics

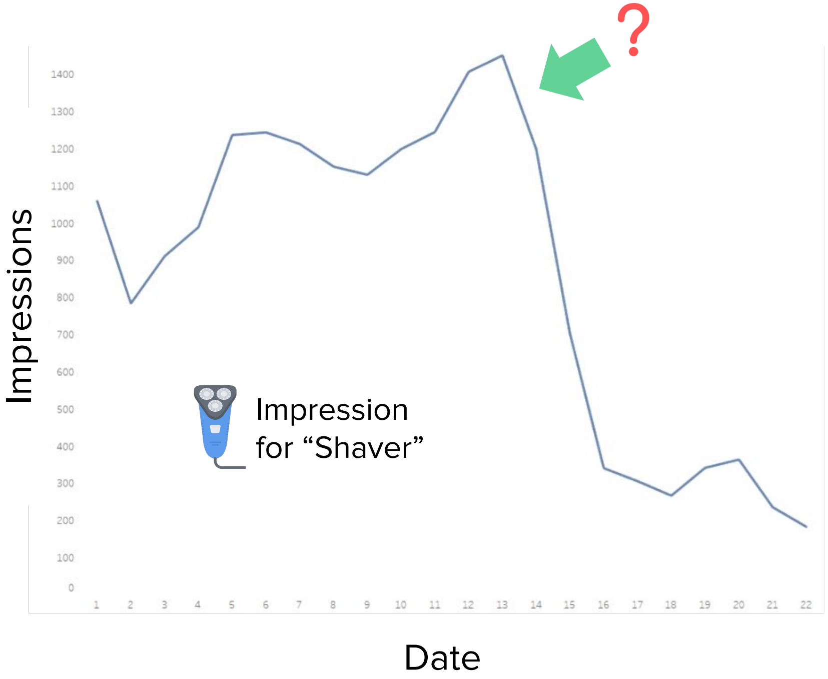
- Useful in a variety of data science applications based on the analysis
  2) **Event-aware scheduler for digital marketing campaigns**

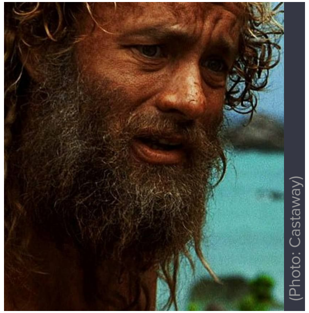


Relevancy: Enrichment for Data Analytics

- Useful in a variety of data science applications based on the analysis

2) Event-aware scheduler for digital marketing campaigns





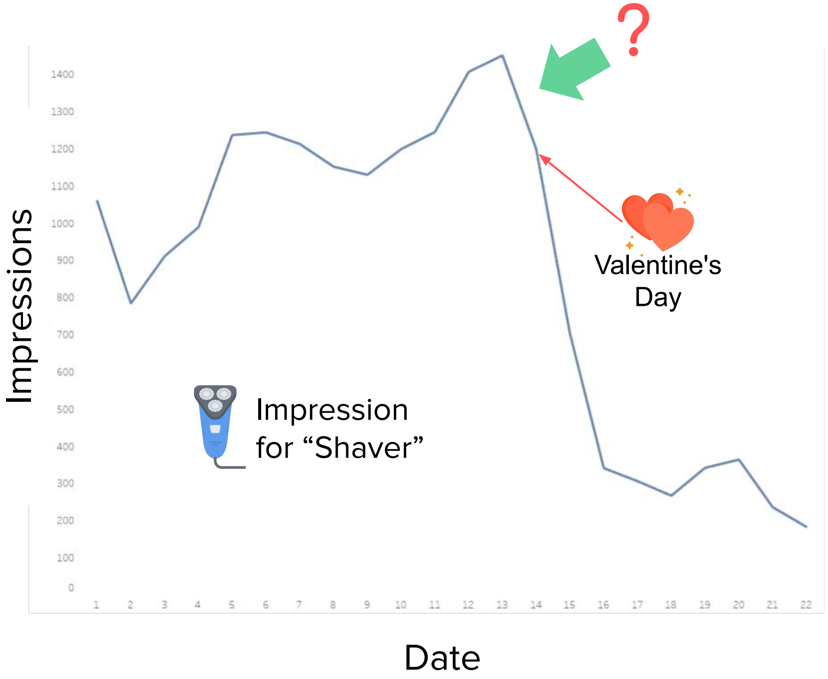
(Photo: Castaway)

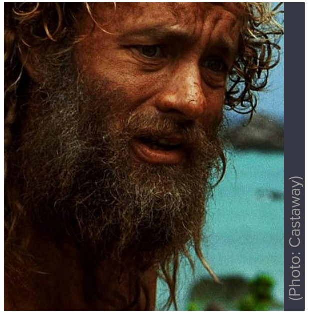
Explanation 1
Men stop shaving

Relevancy: Enrichment for Data Analytics

- Useful in a variety of data science applications based on the analysis

2) Event-aware scheduler for digital marketing campaigns



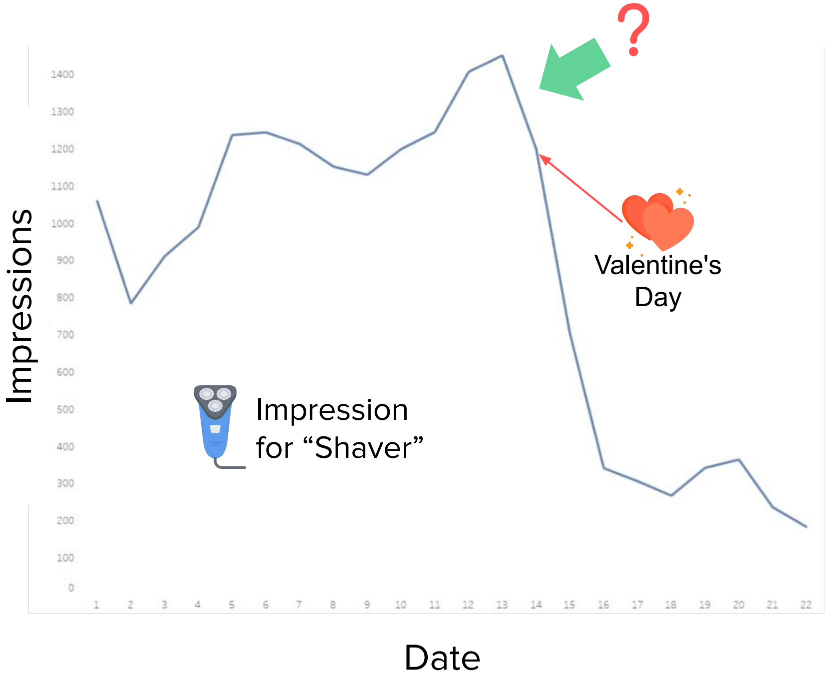


Explanation 1
Men stop shaving

Relevancy: Enrichment for Data Analytics

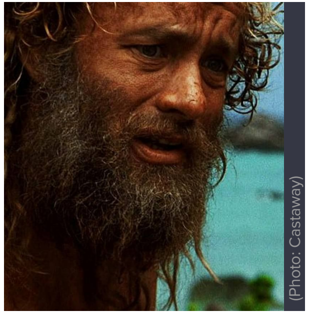
- Useful in a variety of data science applications based on the analysis

2) Event-aware scheduler for digital marketing campaigns



Valentine's Day

Impression for “Shaver”



(Photo: Castaway)

Explanation 1
Men stop shaving

www.spreadshirt.com

Explanation 2
Women interest for shavers starts decreasing the day before Valentine's Day

Relevancy: Enrichment for Data Analytics

- Useful in a variety of data science applications based on the analysis

2) Event-aware scheduler for digital marketing campaigns

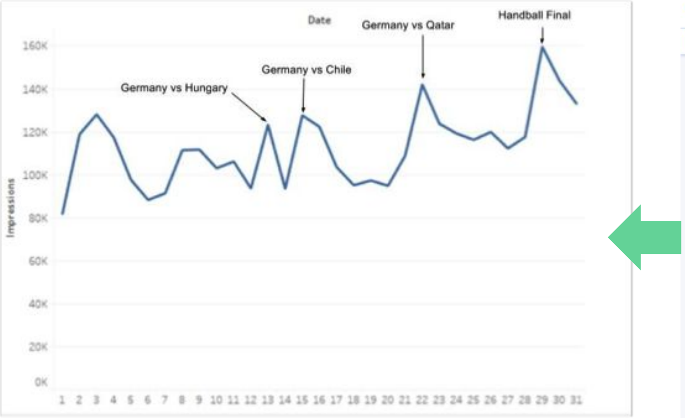
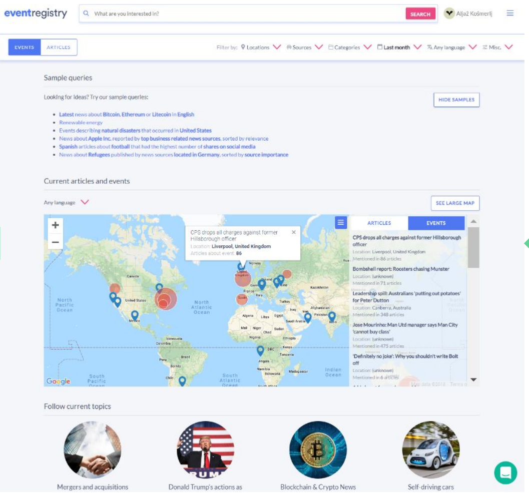


Figure 12: Correlation between German national handball team matches and online digital indicators in the "SportFitness" category.



Sample queries

Looking for ideas? Try our sample queries:

- Latest news about Bitcoin, Ethereum or Litecoin in English
- Renewable energy
- Events describing natural disasters that occurred in United States
- News about Apple Inc. reported by top business related news sources, sorted by relevance
- Spanish articles about football that had the highest number of shares on social media
- News about Refugees published by news sources located in Germany, sorted by source importance

Current articles and events

CPS drops all charges against former Hillsborough officer
Location: Liverpool, United Kingdom
Articles about event: 86

CPS drops all charges against former Hillsborough officer
Location: Liverpool, United Kingdom
Mentioned in 86 articles

Bombshell report: Roosters chasing Munster
Location: (unknown)
Mentioned in 71 articles

Leadership spill: Australians 'putting out potatoes' for Peter Dutton
Location: Canberra, Australia
Mentioned in 348 articles

Jose Mourinho: Man Utd manager says Man City 'cannot buy class'
Location: (unknown)
Mentioned in 475 articles

'Definitely no joke': Why you shouldn't write Bolt off
Location: (unknown)
Mentioned in 6 articles

Follow current topics

Mergers and acquisitions
Donald Trump's actions as
Blockchain & Crypto News
Self-driving cars

WORLD MEN'S HANDBALL CHAMPIONSHIP MUST SEE MATCHES

By BeIN SPORTS January 11, 2017 11:38 AM

As the 2017 World Handball Championship in France draws near, Sports fans in general and handball fans in particular are looking forward to intense first round showdowns that rival knockout stages.

An Example of Data Enrichment

JOT

Google AdWords

KEYWORD	CITY	REGION	Clicks	Category	Date
517827	Ingolstadt	Bavaria	50	NewsMediaPublications	12/03/2018
459143	Berlin	Berlin	42	TravelTourism	12/03/2018
891139	Munich	Bavaria	36	HomeGarden	11/03/2018

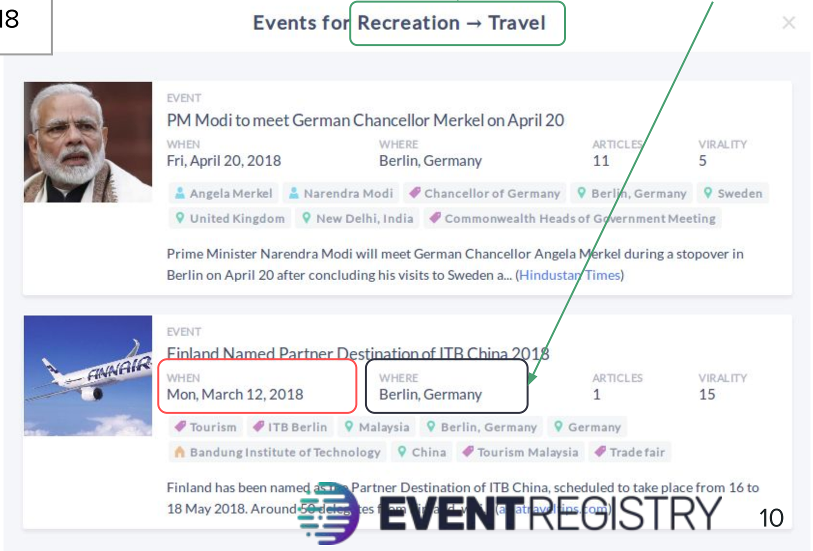
Heterogeneous data (different systems of identifiers)

dmoz

GeoNames

Intuitively: more data attached to input data

Basically, a **LEFT OUTER JOIN** between datasets...



Events for Recreation → Travel

EVENT

PM Modi to meet German Chancellor Merkel on April 20

WHEN: Fri, April 20, 2018 | WHERE: Berlin, Germany | ARTICLES: 11 | VIRALITY: 5

Angela Merkel · Narendra Modi · Chancellor of Germany · Berlin, Germany · Sweden · United Kingdom · New Delhi, India · Commonwealth Heads of Government Meeting

Prime Minister Narendra Modi will meet German Chancellor Angela Merkel during a stopover in Berlin on April 20 after concluding his visits to Sweden a... (Hindustan Times)

EVENT

Finland Named Partner Destination of ITB China 2018

WHEN: Mon, March 12, 2018 | WHERE: Berlin, Germany | ARTICLES: 1 | VIRALITY: 15

Tourism · ITB Berlin · Malaysia · Berlin, Germany · Germany · Bandung Institute of Technology · China · Tourism Malaysia · Trade fair

Finland has been named as the Partner Destination of ITB China, scheduled to take place from 16 to 18 May 2018. Around 50 delegates from Finland w... (asiatraveltips.com)

EVENTREGISTRY

An Example of Data Enrichment

JOT

Google AdWords

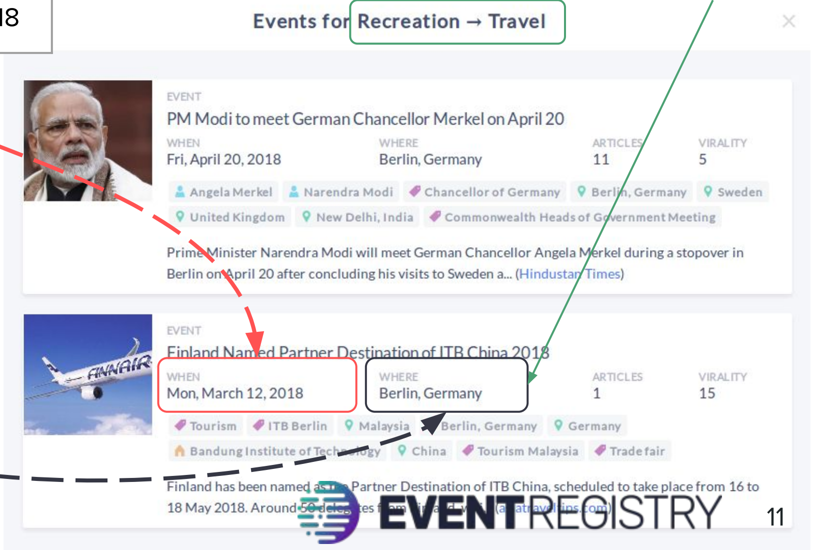
KEYWORD	CITY	REGION	Clicks	Category	Date
517827	Ingolstadt	Bavaria	50	NewsMediaPublications	12/03/2018
459143	Berlin	Berlin	42	TravelTourism	12/03/2018
891139	Munich	Bavaria	36	HomeGarden	11/03/2018

dmoz

GeoNames

PROBLEM

No common identifiers → no JOIN



Events for Recreation → Travel

EVENT

PM Modi to meet German Chancellor Merkel on April 20

WHEN Fri, April 20, 2018

WHERE Berlin, Germany

ARTICLES 11

VIRALITY 5

Angela Merkel · Narendra Modi · Chancellor of Germany · Berlin, Germany · Sweden · United Kingdom · New Delhi, India · Commonwealth Heads of Government Meeting

Prime Minister Narendra Modi will meet German Chancellor Angela Merkel during a stopover in Berlin on April 20 after concluding his visits to Sweden a... (Hindustan Times)

EVENT

Finland Named Partner Destination of ITB China 2018

WHEN Mon, March 12, 2018

WHERE Berlin, Germany

ARTICLES 1

VIRALITY 15

Tourism · ITB Berlin · Malaysia · Berlin, Germany · Germany · Bandung Institute of Technology · China · Tourism Malaysia · Trade fair

Finland has been named as the Partner Destination of ITB China, scheduled to take place from 16 to 18 May 2018. Around 50 delegates from Finland will... (asiatraveltips.com)

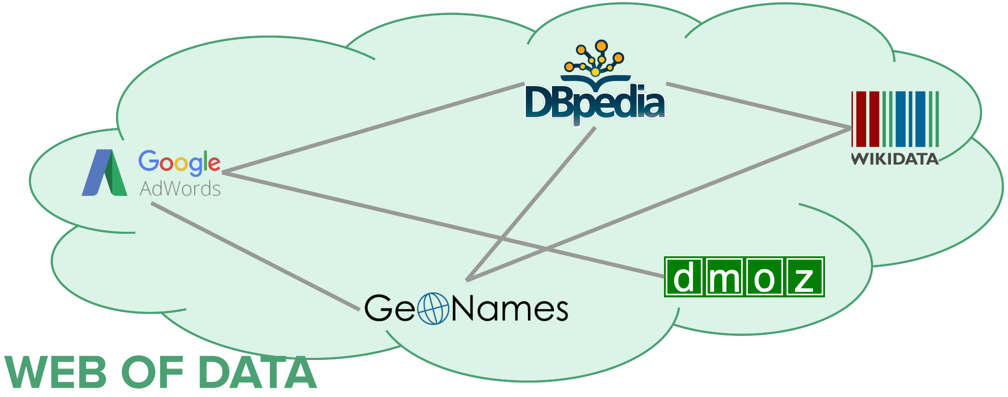
EVENTREGISTRY

An Example of **Semantic** Data Enrichment

JOT

KEYWORD	CITY	REGION	Clicks	Category	Date
517827	Ingolstadt	Bavaria	50	NewsMediaPublications	12/03/2018
459143	Berlin	Berlin	42	TravelTourism	12/03/2018
891139	Munich	Bavaria	36	HomeGarden	11/03/2018

Semantic Web of Data to support the enrichment task



An Example of **Semantic** Data Enrichment

JOT

KEYWORD	CITY	REGION	Clicks	Category	Date
517827	Ingolstadt	Bavaria	50	NewsMediaPublications	12/03/2018
459143	Berlin	Berlin	42	TravelTourism	12/03/2018
891139	Munich	Bavaria	36	HomeGarden	11/03/2018

Semantic Web of Data to support the enrichment task

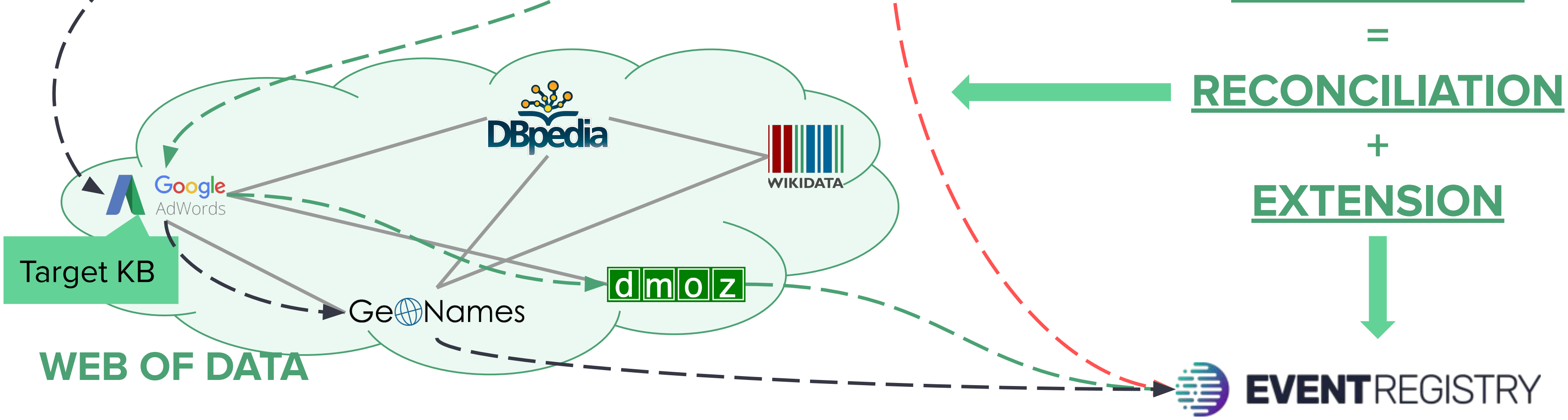
SEMANTIC ENRICHMENT

=

RECONCILIATION

+

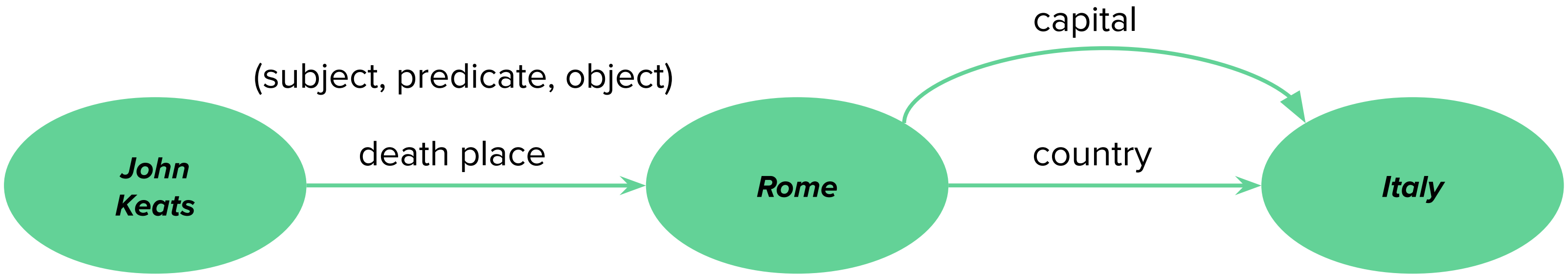
EXTENSION



Fundamentals

Knowledge Graphs

Knowledge Graph: Instances



(subject, predicate, object)

John Keats

death place

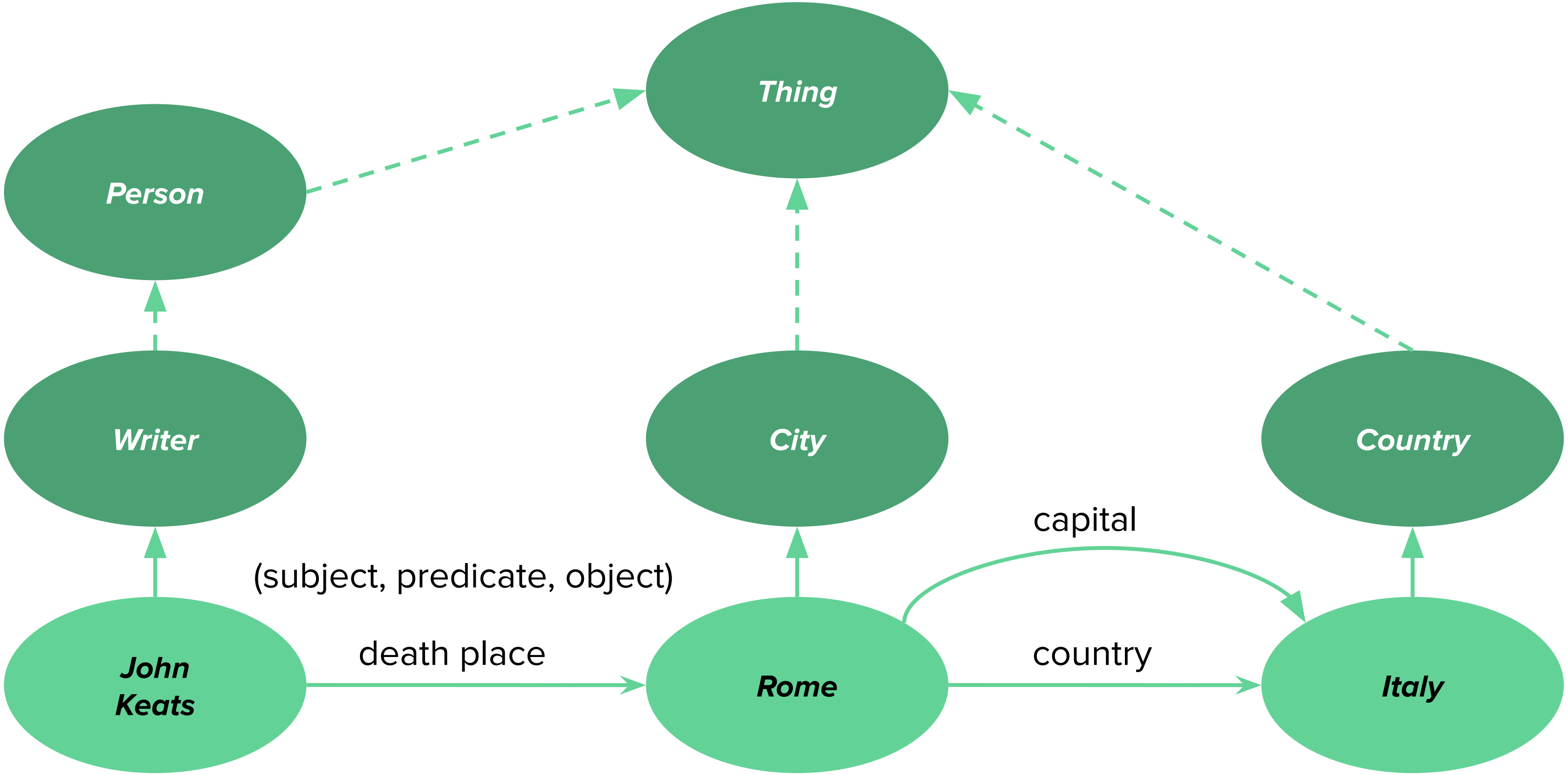
Rome

capital

country

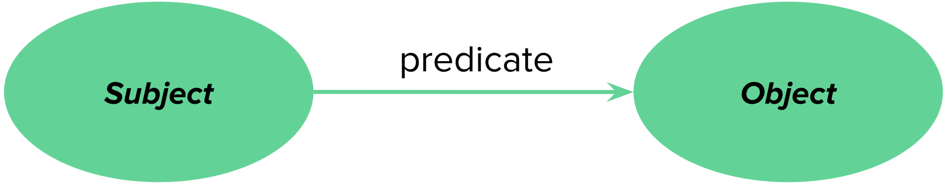
Italy

Knowledge Graph: Types



Knowledge Graph: FOL View vs. Graph View

Predicate(Subject,Object)

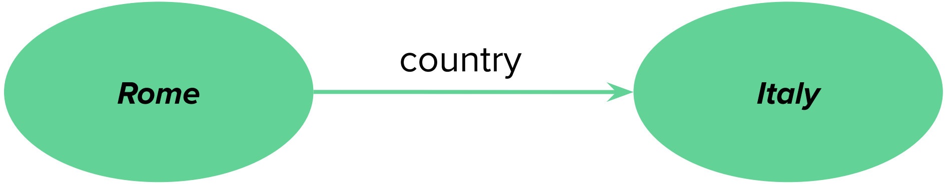


<Subject,predicate,Object>

FOL VIEW

GRAPH VIEW

Country(Rome,Italy)



<Rome,country,Italy>

FOL VIEW

GRAPH VIEW

Semantics in KGs

Schema (≍ ontology) defines the meaning of general terms in the KG

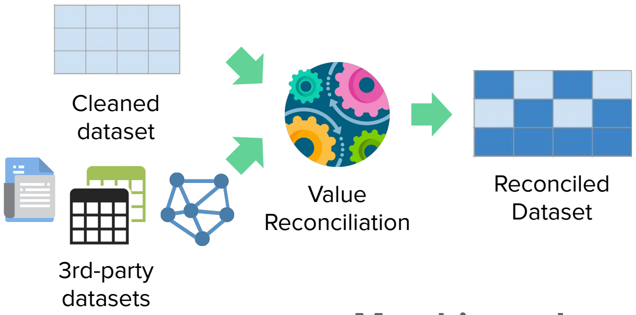
- Types, e.g., City(x)
- Relations, e.g., country(x,y)

Schema definition supports inference (by deduction, or by induction, etc.)

- E.g., $\forall x,y$ country(x,y) $\Rightarrow$ (City(x) $\wedge$ Country(y))
- E.g., $\forall x,y$ capital(x,y) $\Rightarrow$ country(x,y)

Reconciliation

Reconciliation



Cleaned dataset

3rd-party datasets

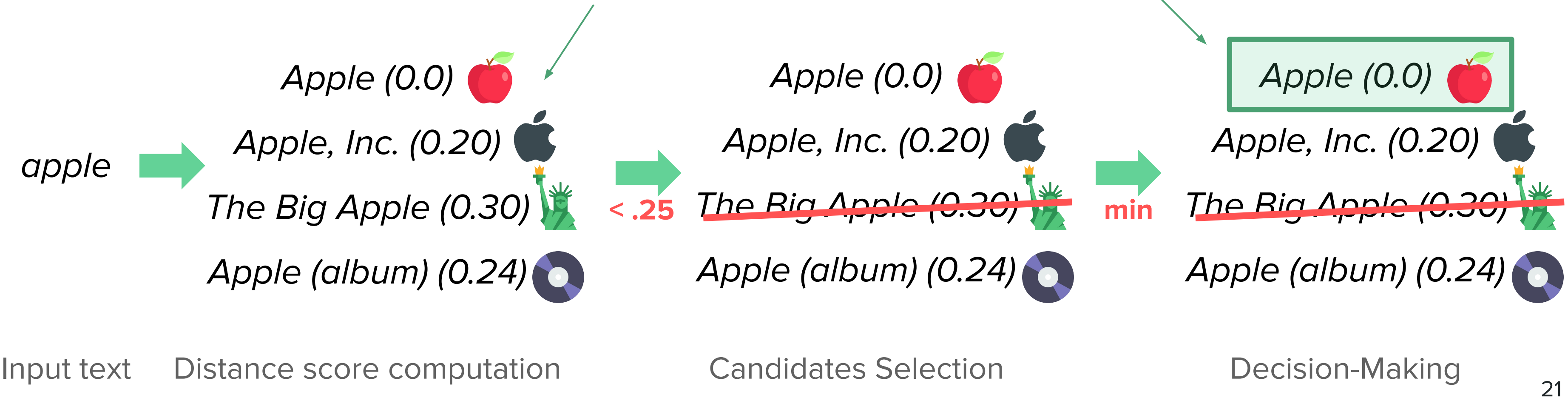
Value Reconciliation

Reconciled Dataset

Given a text that is a name or label for something, returns a ranked list of **potential entities** (based on some matching rules)

Matching rules assign a score to the entities (e.g., distance)

Fully-automated approaches must also **select the best candidate** to return



apple

Apple (0.0) 🍎

Apple, Inc. (0.20)

The Big Apple (0.30)

Apple (album) (0.24)

< .25

Apple (0.0) 🍎

Apple, Inc. (0.20)

~~*The Big Apple (0.30)*~~

Apple (album) (0.24)

min

Apple (0.0) 🍎

Apple, Inc. (0.20)

~~*The Big Apple (0.30)*~~

Apple (album) (0.24)

Input text

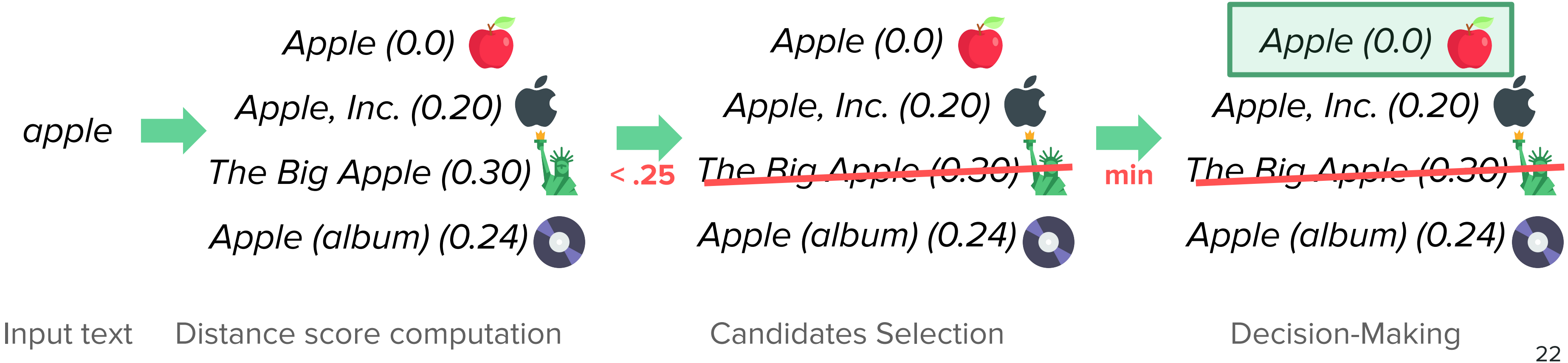
Distance score computation

Candidates Selection

Decision-Making

Reconciliation: Main issues

- Precision depends on the **ambiguities**
- Impossible to explore the whole **candidate entities space**
- Human experts can not **check all results**



apple

Input text

Apple (0.0)

Apple, Inc. (0.20)

The Big Apple (0.30)

Apple (album) (0.24)

Distance score computation

< .25

Apple (0.0)

Apple, Inc. (0.20)

~~*The Big Apple (0.30)*~~

Apple (album) (0.24)

Candidates Selection

min

Apple (0.0)

Apple, Inc. (0.20)

~~*The Big Apple (0.30)*~~

Apple (album) (0.24)

Decision-Making

Reconciliation in Tables

Google AdWords

KEYWORD	REGION	Clicks	Category	Date
194906	Thuringia	64	BusinessManagement	11/03/2018
517827	Bavaria	50	NewsMediaPublications	12/03/2018
459143	Berlin	42	TravelTourism	12/03/2018
891139	Bavaria	36	Vehicles	11/03/2018
459143	Bavaria	30	HomeGarden	10/03/2018

Reconcile region names versus Geonames identifiers (11.7M entities)

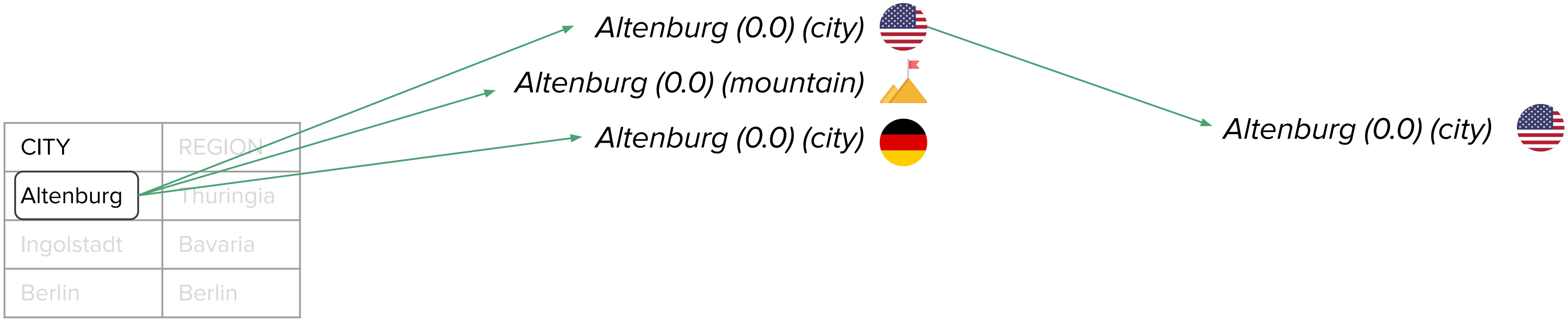
Reconcile category names versus DMOZ taxonomy (1M entities)

GeoNames

dmoz

KEYWORD	REGION	Geonames ID	Clicks	Category	DMOZ ID	Date
194906	Thuringia	2822542	64	BusinessManagement	dmoz/Business/Management	11/03/2018
517827	Bavaria	2951839	50	NewsMediaPublications	dmoz/News	12/03/2018
459143	Berlin	2950157	42	TravelTourism	dmoz/Recreation/Travel	12/03/2018
891139	Bavaria	2951839	36	Vehicles	dmoz/Shopping/Vehicles	11/03/2018
459143	Bavaria	2951839	30	HomeGarden	dmoz/Home/Gardening	10/03/2018

Reconciliation in Tables - Cell-by-Cell



CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	Berlin

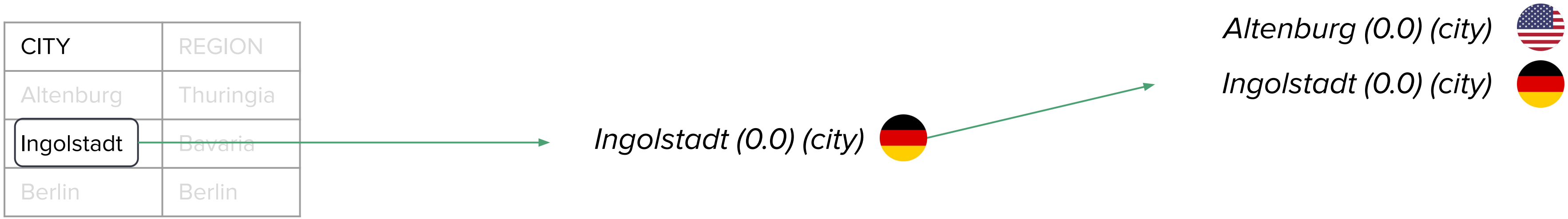
Altenburg (0.0) (city)

Altenburg (0.0) (mountain)

Altenburg (0.0) (city)

Altenburg (0.0) (city)

Reconciliation in Tables - Cell-by-Cell



CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	Berlin

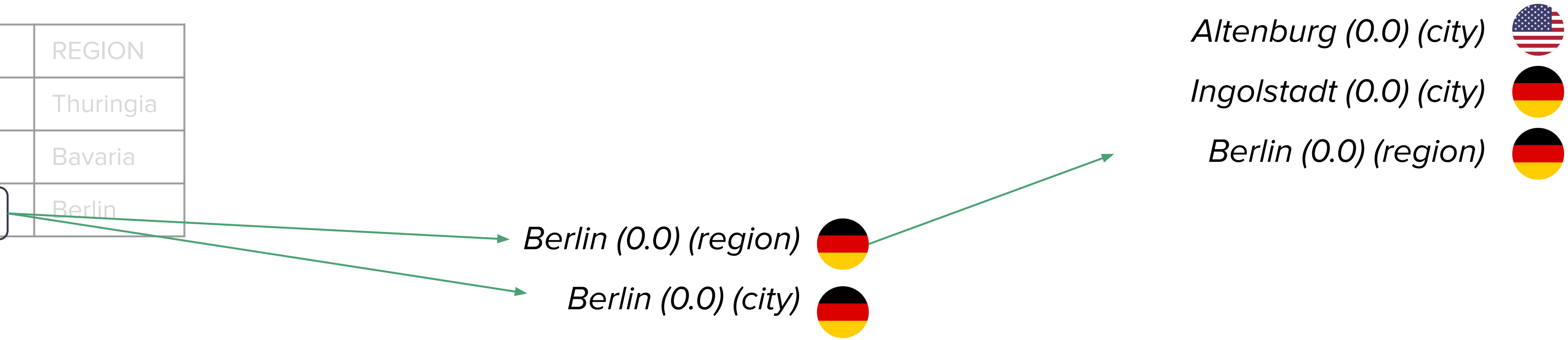
Ingolstadt (0.0) (city) 🇩🇪

Altenburg (0.0) (city) 🇺🇸

Ingolstadt (0.0) (city) 🇩🇪

Reconciliation in Tables - Cell-by-Cell

CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	Berlin



Berlin (0.0) (region) 🇩🇪

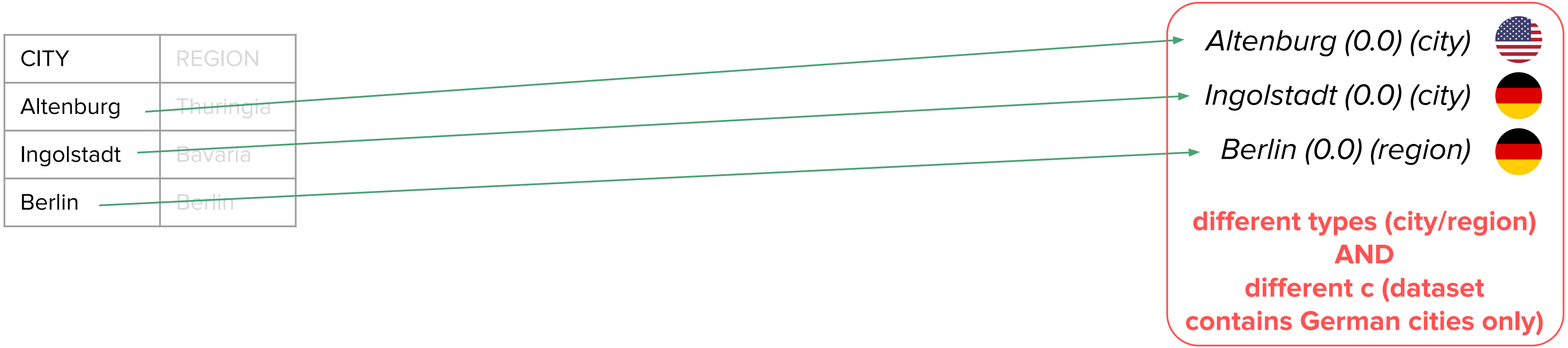
Berlin (0.0) (city) 🇩🇪

Altenburg (0.0) (city) 🇺🇸

Ingolstadt (0.0) (city) 🇩🇪

Berlin (0.0) (region) 🇩🇪

Reconciliation in Tables - Cell-by-Cell



CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	Berlin

Altenburg (0.0) (city) 🇺🇸

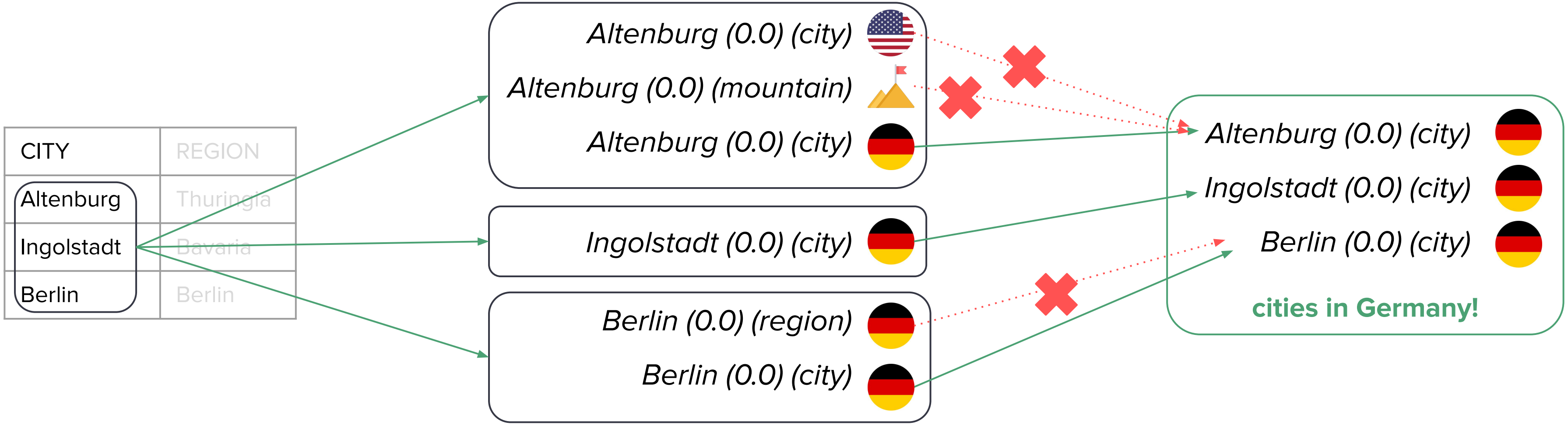
Ingolstadt (0.0) (city) 🇩🇪

Berlin (0.0) (region) 🇩🇪

different types (city/region)
AND
different c (dataset contains German cities only)

- We did not exploit the tabular structure!
- Cells in the same column talk about the same things
    - Not always true! Sometimes data are very noisy...

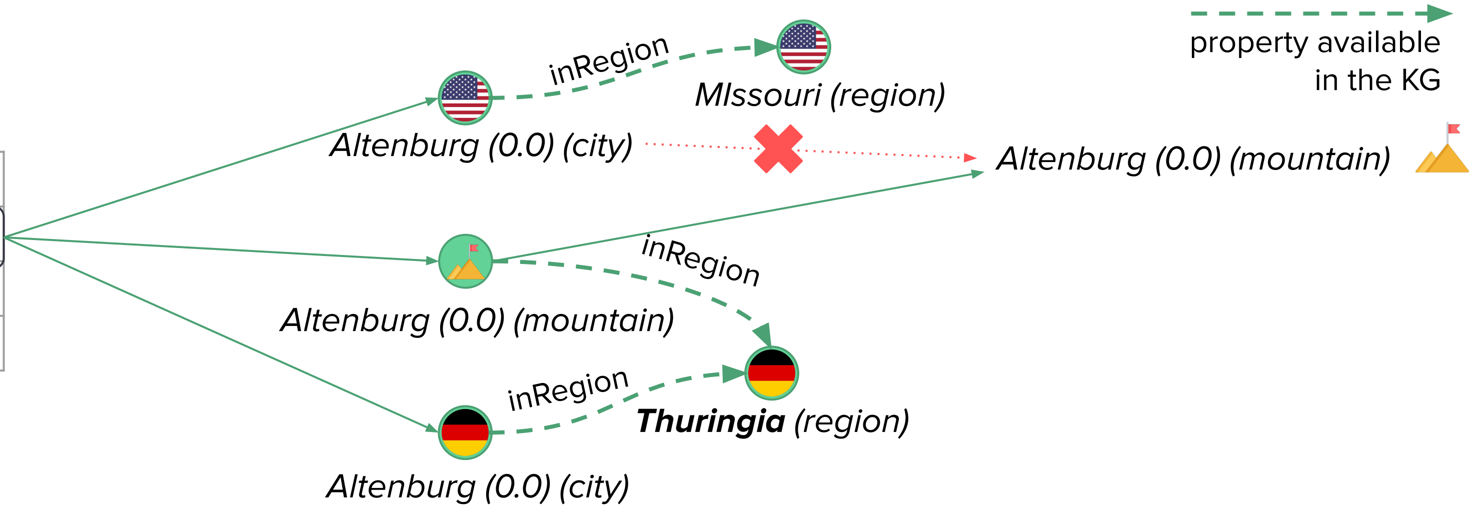
Reconciliation in Tables - Column-by-Column



- By looking at the columns, we are focusing on **CATEGORIES**
- We have to identify which is the **category that has at least one candidate in each subgroup**
- How many categories exist? cities, cities in Europe, cities in Italy …
  - ~2^(m · n), where m = #attributes and n = #possible values for each attribute

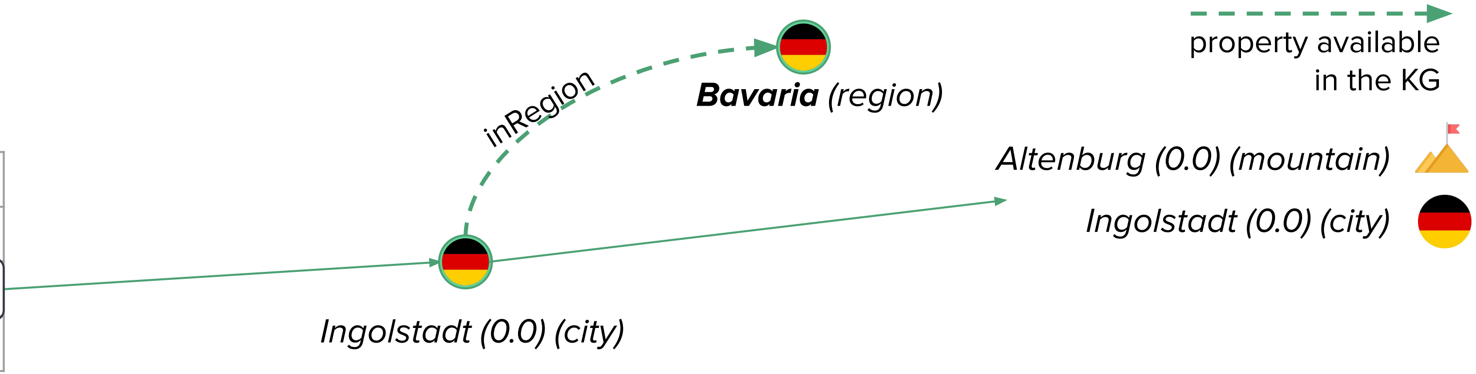
Reconciliation in Tables - Row-by-Row

CITY	REGION
Altenburg	**Thuringia**
Ingolstadt	Bavaria
Berlin	Berlin



Reconciliation in Tables - Row-by-Row

CITY	REGION
Altenburg	Thuringia
Ingolstadt	**Bavaria**
Berlin	Berlin



property available in the KG

inRegion

Bavaria *(region)*

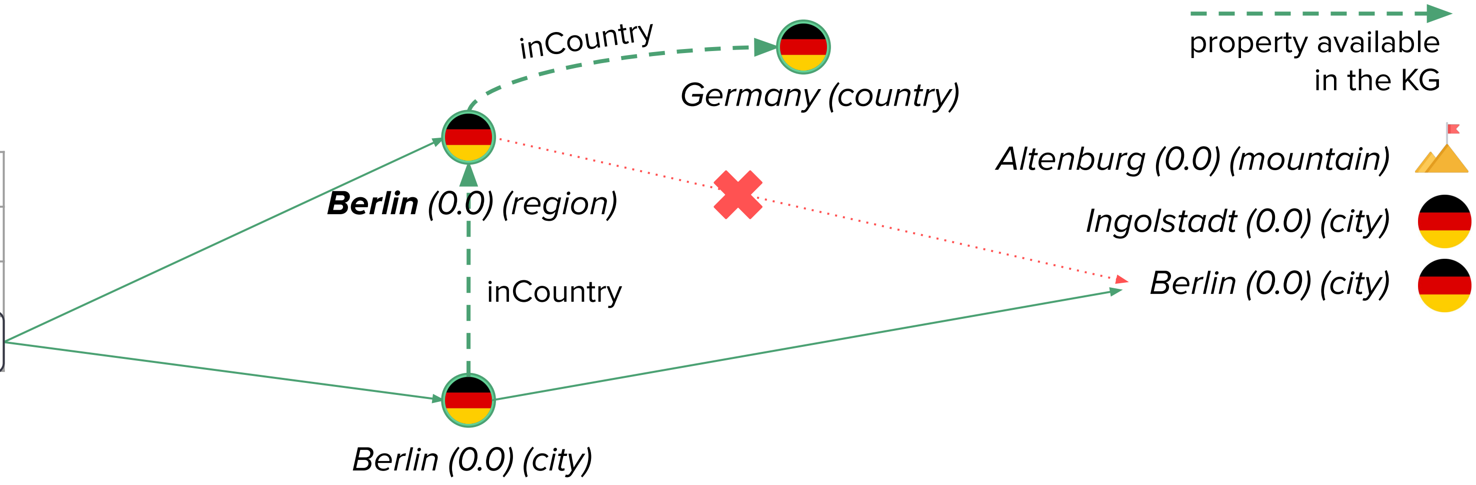
Altenburg (0.0) (mountain)

Ingolstadt (0.0) (city)

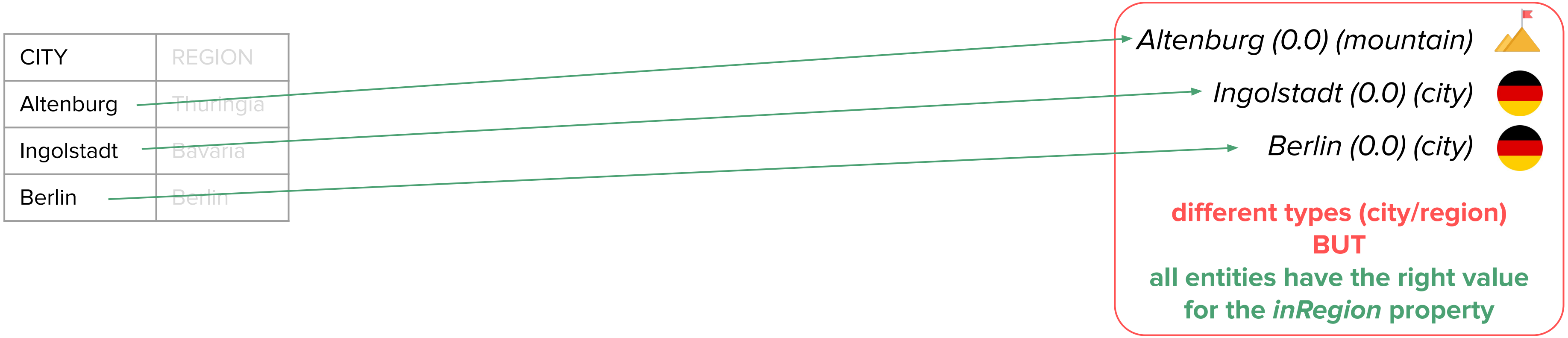
Ingolstadt (0.0) (city)

Reconciliation in Tables - Row-by-Row

CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	**Berlin**



Reconciliation in Tables - Row-by-Row



CITY	REGION
Altenburg	Thuringia
Ingolstadt	Bavaria
Berlin	Berlin

Altenburg (0.0) (mountain)

Ingolstadt (0.0) (city)

Berlin (0.0) (city)

different types (city/region)
BUT
all entities have the right value for the *inRegion* property

- By looking at the rows, we are focusing on **PROPERTIES**
- We have to identify which are the **most discriminative properties to consider**
- How many properties to compare for each row?
  - ~(m · n), where m = #attributes and n = #candidates

Logic Tensor Networks

Next slides adapted from a seminar given by Federico Bianchi at UNIMIB (f.bianchi@unibocconi.it)

Terminological Recap

A **constant** is an element of a domain (set) taken in consideration

S : {Rome, Paris, …}
T : {Italy, France, …}

A **function** is a relation f: S → T between sets that associates to every element of a first set exactly one element of the second set.

Capital: T → S
Capital(Italy) = Rome

A **predicate** is a Boolean-valued function P: S → {1 (= True), 0 (=False)}.

city: S → {1, 0}
city(Rome) = 1

country: S x T → {1, 0}
country(Rome, Italy) = 1

I ❤

Terminological Recap (cont)

An **axiom:** a statement in a logical language:

R(a, b)

A **grounded axiom** contains grounded constants:

country(Rome, Italy)

A **quantified axiom** is an axiom that contains quantified variables:

∀ x,y *capital(x, y)*

A **formula** is a combination of grounded and quantified axioms:

∀ x,y *country(Rome, Italy) & country(Paris, France) & capital(x, y)*

Logic Tensor Networks

Logic Tensor Networks [Serafini+,2016] (LTNs) => neuro-symbolic [Garcez+,2008;Garcez+,2012] combines neural network and symbolic AI.

LTNs = **Neural Networks** + **First Order Fuzzy Logic**

Key Aspects:

- LTNs **ground fuzzy logic in a vector space: continuous values in [0,1]**
- LTNs assign truth values to formulas using neural networks
- LTNs can learn from both data and rules
- LTNs can be used to do inferences over rules after training

Key Idea: LTNs provide a method to learn reasoning over vector spaces

Logic Tensor Networks

Logic Tensor Networks [Serafini+,2016] (LTNs) => neuro-symbolic [Garcez+,2008;Garcez+,2012] combines neural network and symbolic AI.

LTNs = **Neural Networks** + **First Order Fuzzy Logic**

Key Aspects:

- LTNs ground fuzzy logic in a vector space: continuous values in [0,1]
- LTNs **assign truth values to formulas using neural networks**
- LTNs can learn from both data and rules
- LTNs can be used to do inferences over rules after training

Key Idea: LTNs provide a method to learn reasoning over vector spaces

Logic Tensor Networks

Logic Tensor Networks [Serafini+,2016] (LTNs) => neuro-symbolic [Garcez+,2008;Garcez+,2012] combines neural network and symbolic AI.

LTNs = Neural Networks + First Order Fuzzy Logic

Key Aspects:

- LTNs ground fuzzy logic in a vector space: continuous values in [0,1]
- LTNs assign truth values to formulas using neural networks
- LTNs **can learn from both data and rules**
- LTNs can be used to do inferences over rules after training

Key Idea: LTNs provide a method to learn reasoning over vector spaces

Logic Tensor Networks

Logic Tensor Networks [Serafini+,2016] (LTNs) => neuro-symbolic [Garcez+,2008;Garcez+,2012] combines neural network and symbolic AI.

LTNs = **Neural Networks** + **First Order Fuzzy Logic**

Key Aspects:

- LTNs ground fuzzy logic in a vector space: continuous values in [0,1]
- LTNs assign truth values to formulas using neural networks
- LTNs can learn from both data and rules
- LTNs **can be used to do inferences over rules after training**

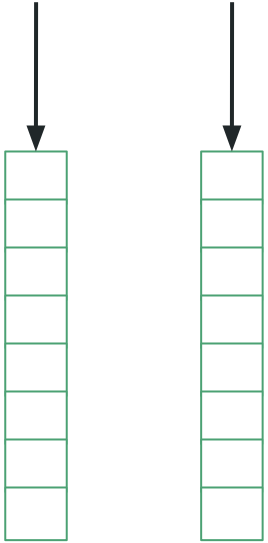
Key Idea: LTNs provide a method to learn reasoning over vector spaces

Logic Tensor Networks: General Idea

parent(Susan, Ann)

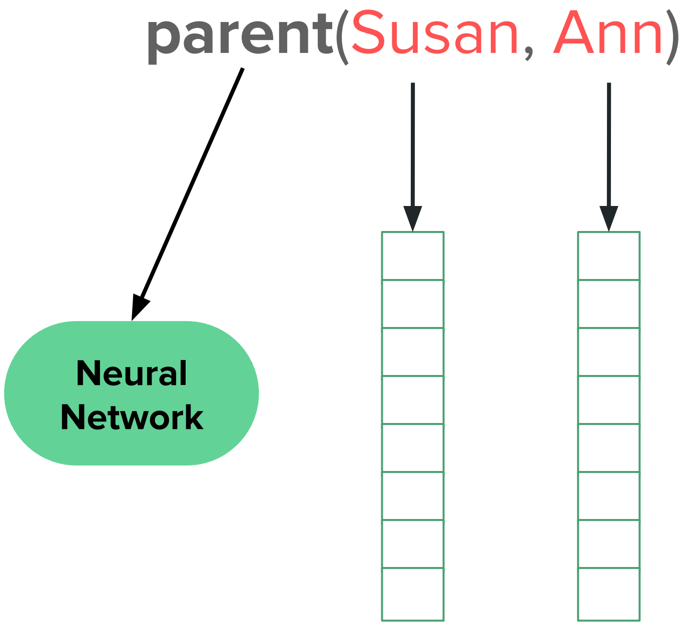
Logic Tensor Networks: General Idea

parent(Susan, Ann)



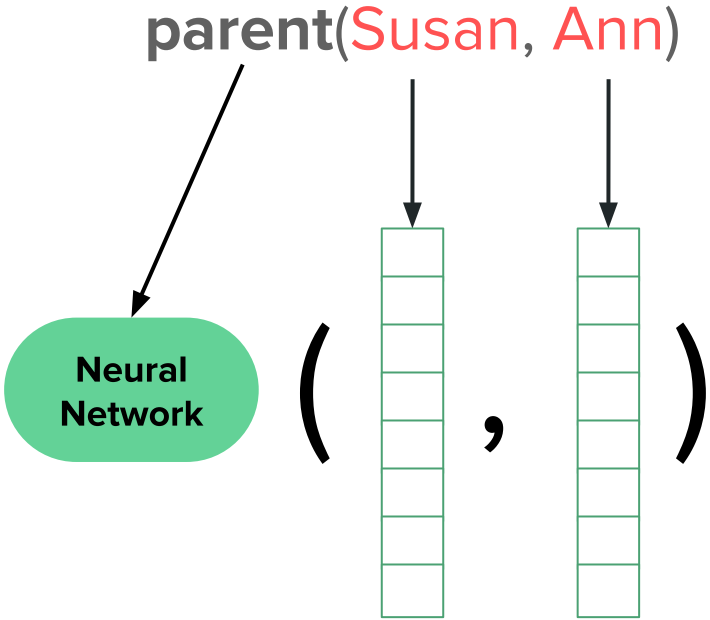
Constants are points in $\mathbf{R}^k$

Logic Tensor Networks: General Idea

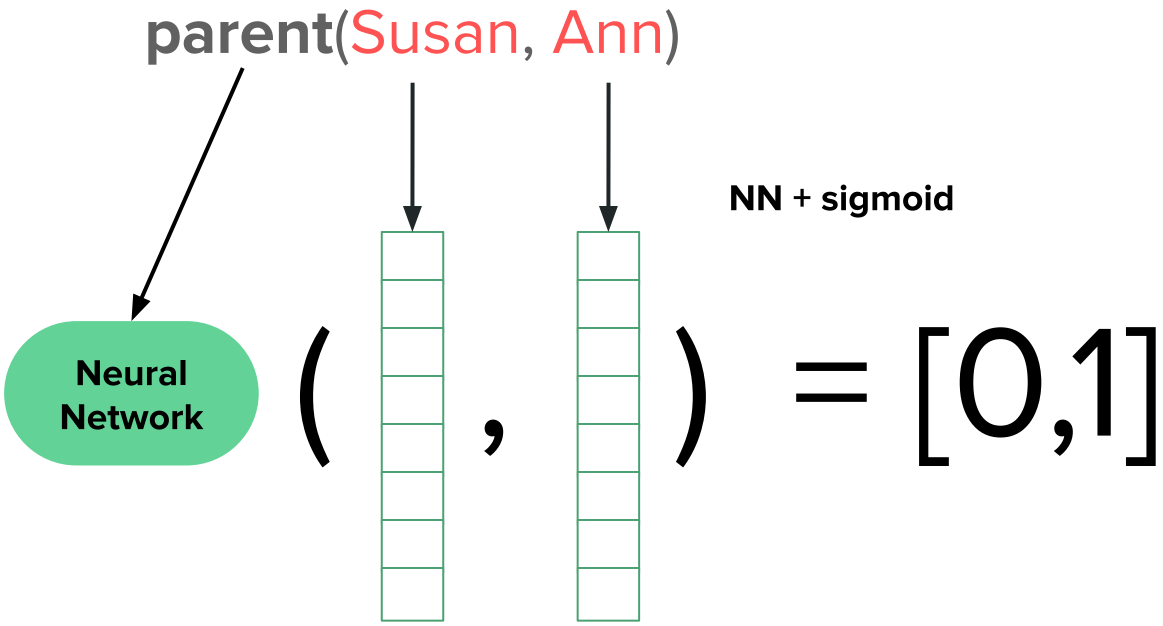


Each predicate in LTN is a NN (the training phase is on many networks as predicates)

Logic Tensor Networks: General Idea



Logic Tensor Networks: General Idea

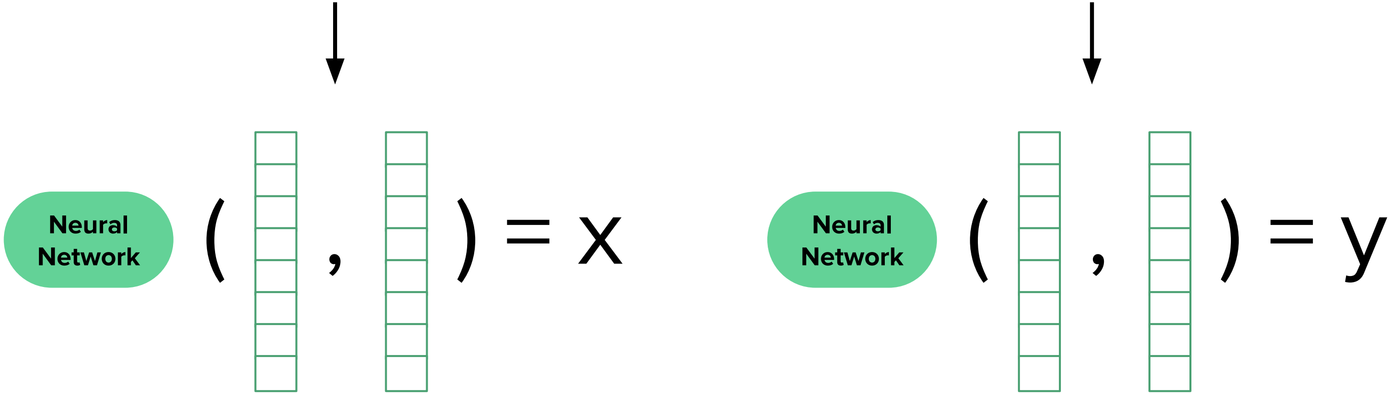


Logic Tensor Networks: General Idea

parent(Susan, Ann) & **parent(**Mike, Robert)

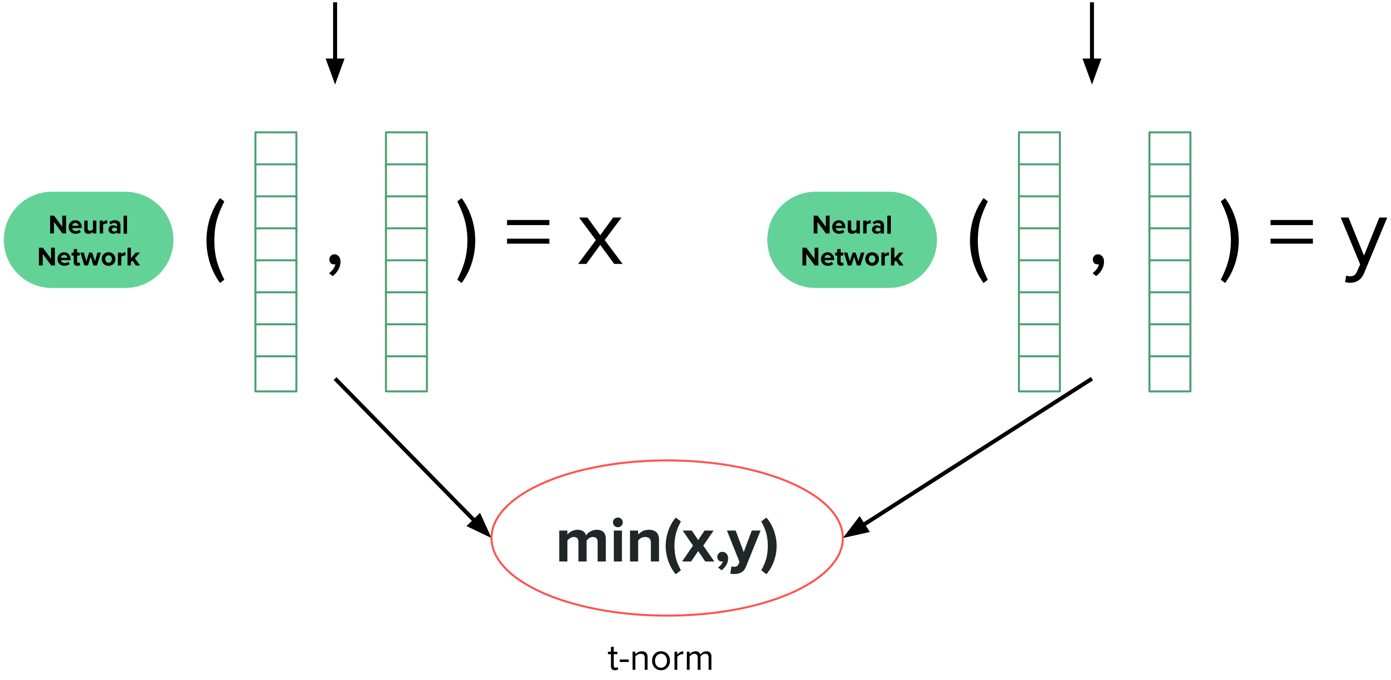
Logic Tensor Networks: General Idea

parent(Susan, Ann) & **parent(**Mike, Robert)



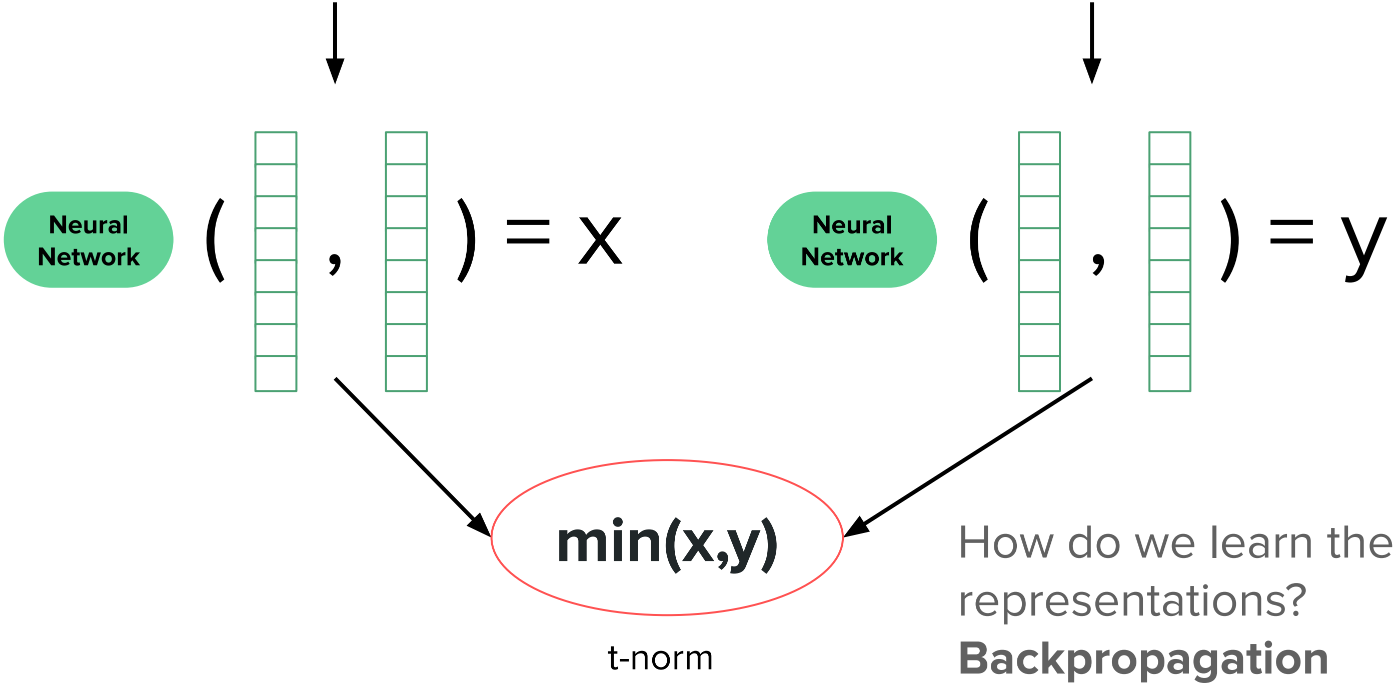
Logic Tensor Networks: General Idea

parent(Susan, Ann) & **parent(**Mike, Robert)



Logic Tensor Networks: General Idea

parent(Susan, Ann) & **parent(**Mike, Robert)

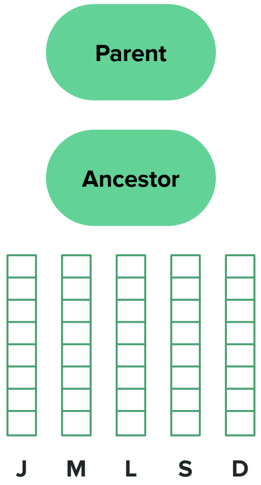


How do we learn these representations?
Backpropagation

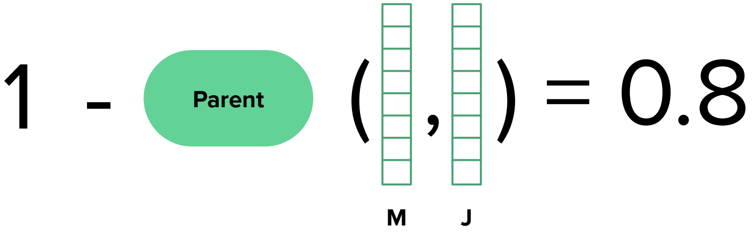
Example

KB:

- ¬parent(mark, john)
- parent(john,mark)
- ancestor(mark, lucas)
- parent(john, susan) | parent(john, dania)



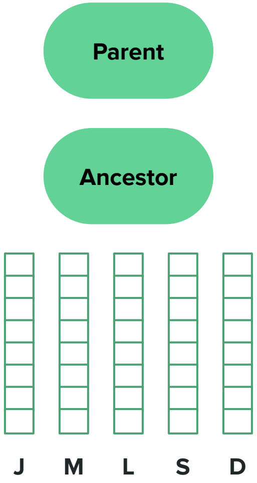
Example: Forward Pass



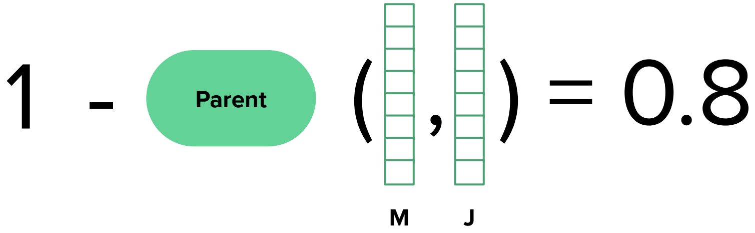
We want to maximize this

KB:

- **¬parent(mark, john)**
- parent(john,mark)
- ancestor(mark, lucas)
- parent(john, susan) | parent(john, dania)



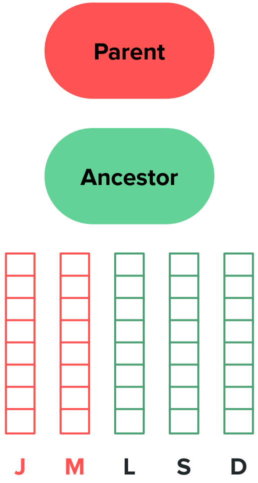
Example: Back Pass



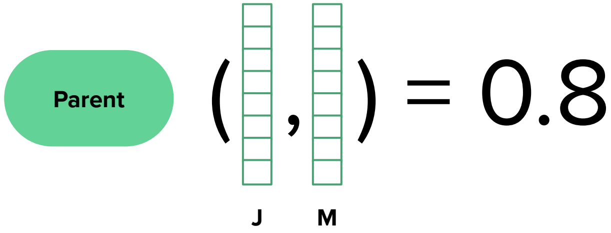
Update using backpropagation

KB:

- **¬parent(mark, john)**
- parent(john,mark)
- ancestor(mark, lucas)
- parent(john, susan) | parent(john, dania)



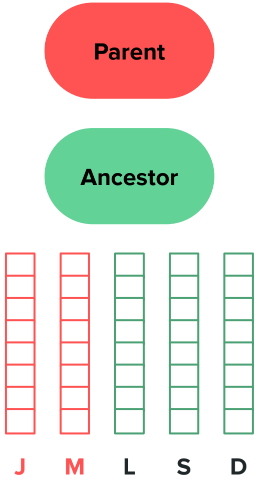
Example: Forward and Back Pass



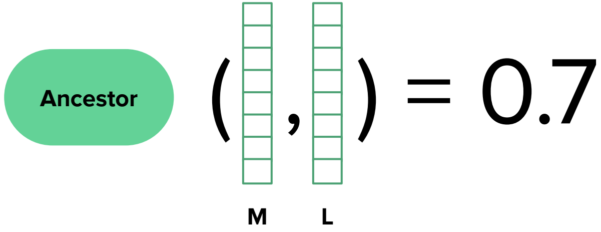
We want to maximize this and thus we update the respective values

KB:

- ¬parent(mark, john)
- **parent(john,mark)**
- ancestor(mark, lucas)
- parent(john, susan) | parent(john, dania)



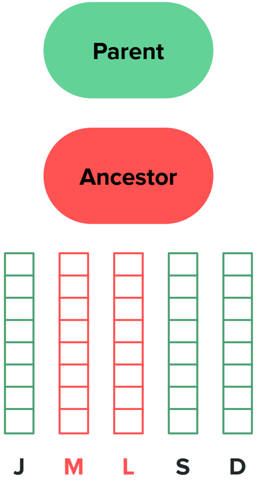
Example: Forward and Back Pass



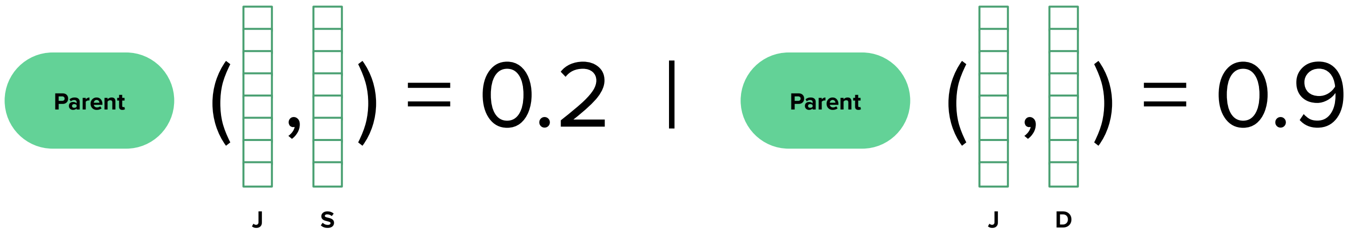
We want to maximize this and thus we update the respective values

KB:

- ¬parent(mark, john)
- parent(john,mark)
- **ancestor(mark, lucas)**
- parent(john, susan) | parent(john, dania)



Example: Forward and Back Pass

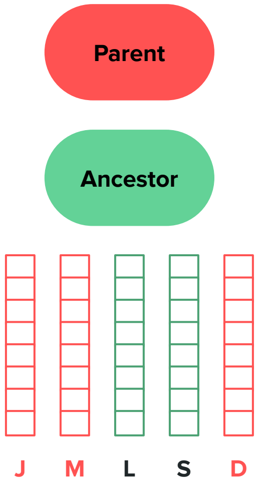


$$\max(0.2, 0.9) = 0.9$$

We want to maximize this and thus we update the respective values

KB:

- ¬parent(mark, john)
- parent(john,mark)
- ancestor(mark, lucas)
- **parent(john, susan) | parent(john, dania)**

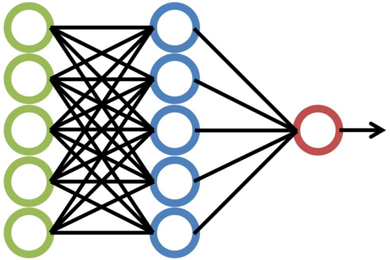


Logic Tensor Networks: Learning

The network is trained on a **best satisfiability task**:

- Learn the representations
    - **vectors** for the constants
    - **parameters** for the predicates

in such a way that **the axioms are satisfied in the best possible way**.

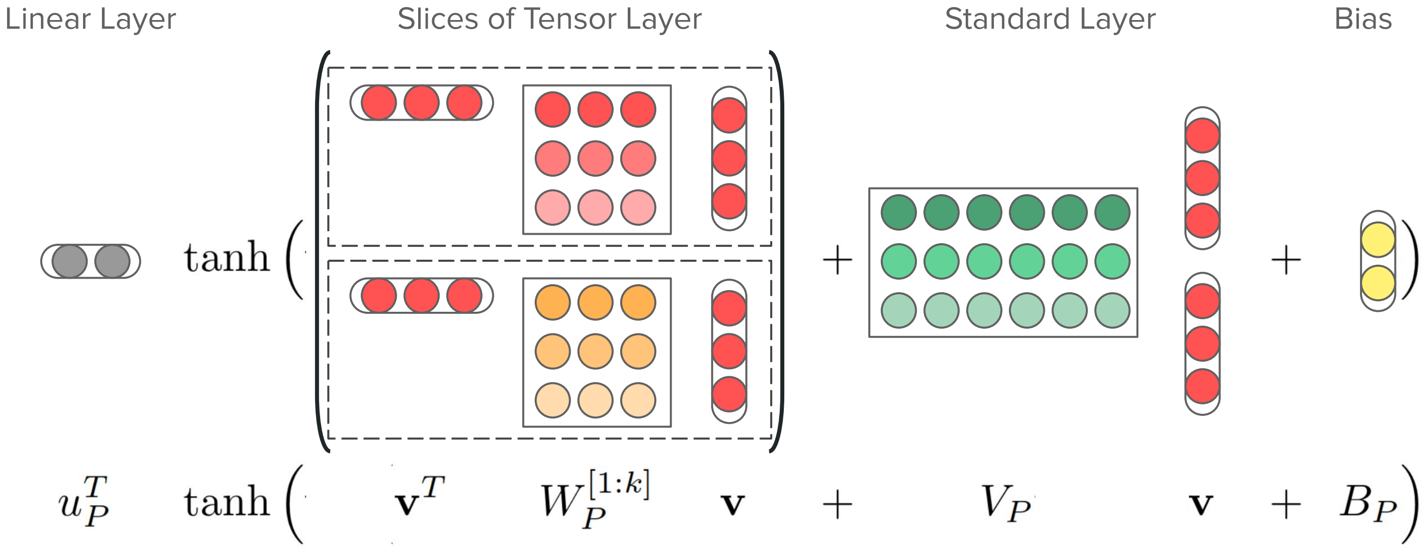


Given **parent(Ann, Susan)** we expect the network to **learn representations** for **Ann**, **Susan** and **parent** in such a way that the predicted value is close to 1

Implementing Logic in Tensor Networks

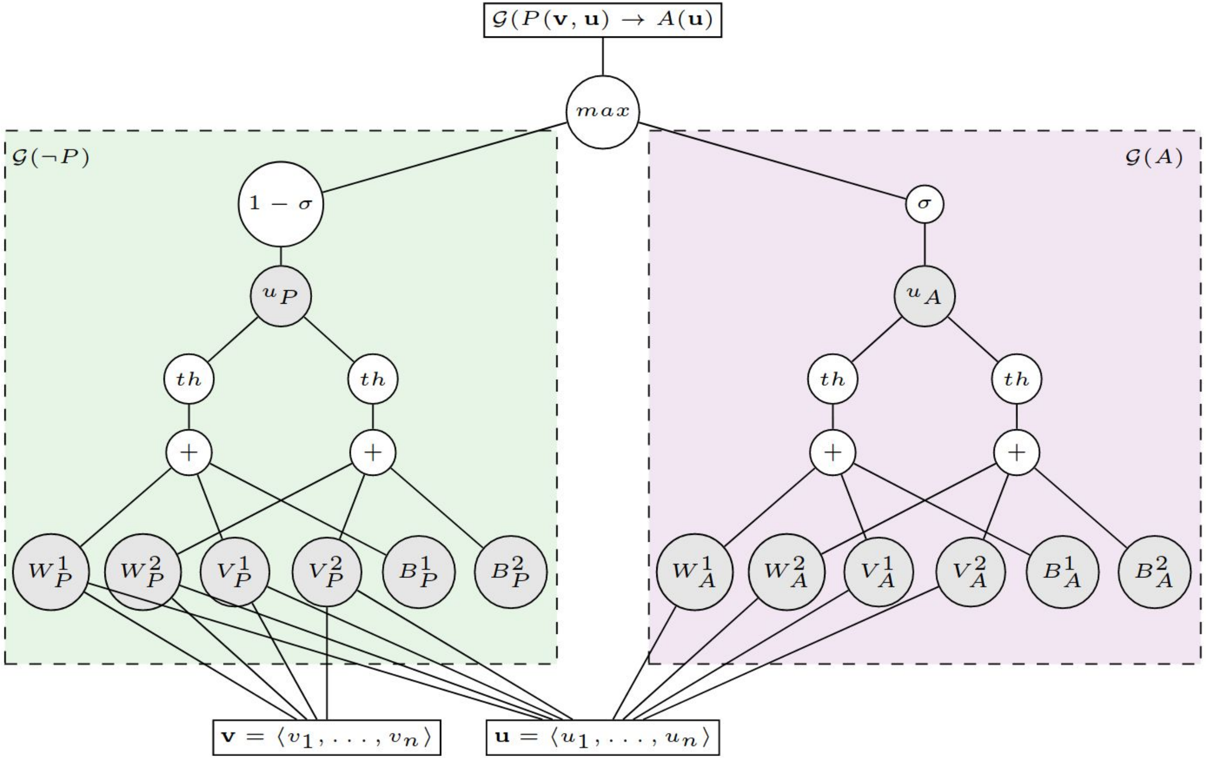
The grounding of *m*-ary predicate P, **G(P)**, is defined as a function from R^{mn} to [0,1]

$$\mathcal{G}(P) = \sigma\left(u_P^T \tanh\left(\mathbf{v}^T W_P^{[1:k]}\mathbf{v} + V_P\mathbf{v} + B_P\right)\right)$$ [Serafini+,2016]



(Image adapted from [Socher+,2013])

Implementing Logic in Tensor Networks: an example



Tensor net for P(x, y) → A(y), with G(x) = v and G(y) =u and k= 2 (from [Serafini+,2016])

Logic Tensor Networks: Data and Rules

LTNs can learn from both data and rules.

Quantifiers are defined **over a domain sample.**

parent(Mark,Susan)

parent(Ron,Susan)

∀ x,y parent(x,y) → ancestor(x,y)

Optimize the **representations** of the parameters to support the axioms

Quantifiers interpreted using an **aggregation function** (e.g., average):

∀ x P(x) = average value of P(x) in LTNs.

Logic Tensor Networks: After Training Inference

The trained network defines a new **compositional language** built on constants, functions and predicates, which can be combined arbitrary.

The trained network can be used for discovering novel inferences.

Suppose we train using a dataset of ***parents*** and ***ancestors*** relationships.

Logic Tensor Networks: After Training Inference

The trained network defines a new **compositional language** built on constants, functions and predicates, which can be combined arbitrary.

The trained network can be used for discovering novel inferences.

Suppose we train using a dataset of ***parents*** and ***ancestors*** relationships.

After training we can query LTNs on:

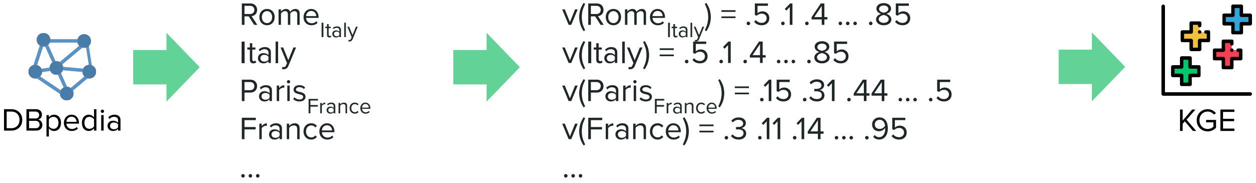
∀ x,y ancestor(x,y) → parent(x,y) has truth value close to 0

Reconcile tables with LTNs

LTN-based Reconciliation

1 Embed the KB in a vector space KGE

- Each entity in the graph is mapped to a *n*-dimensional point in R^n
  - e.g., by Graph Embedding [Wang+,2017]



DBpedia

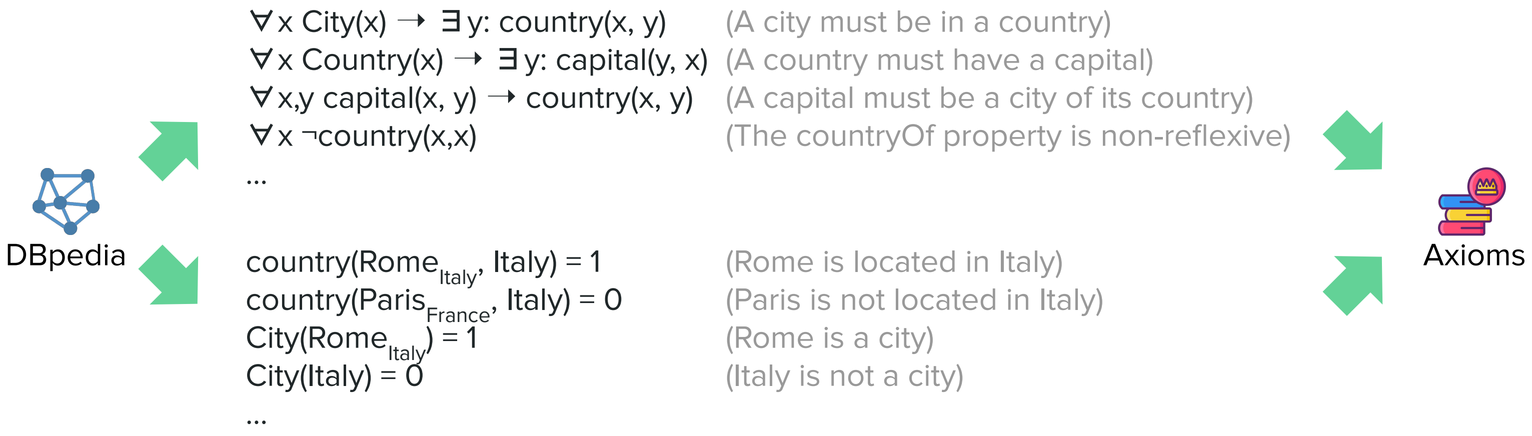
$\text{Rome}_{\text{Italy}}$
Italy
$\text{Paris}_{\text{France}}$
France
...

$\text{v(Rome}_{\text{Italy}}\text{)}$ = .5 .1 .4 ... .85
v(Italy) = .5 .1 .4 ... .85
$\text{v(Paris}_{\text{France}}\text{)}$ = .15 .31 .44 5
v(France) = .3 .11 .14 ... .95
...

KGE

LTN-based Reconciliation

2 Get axioms from the KB ontology

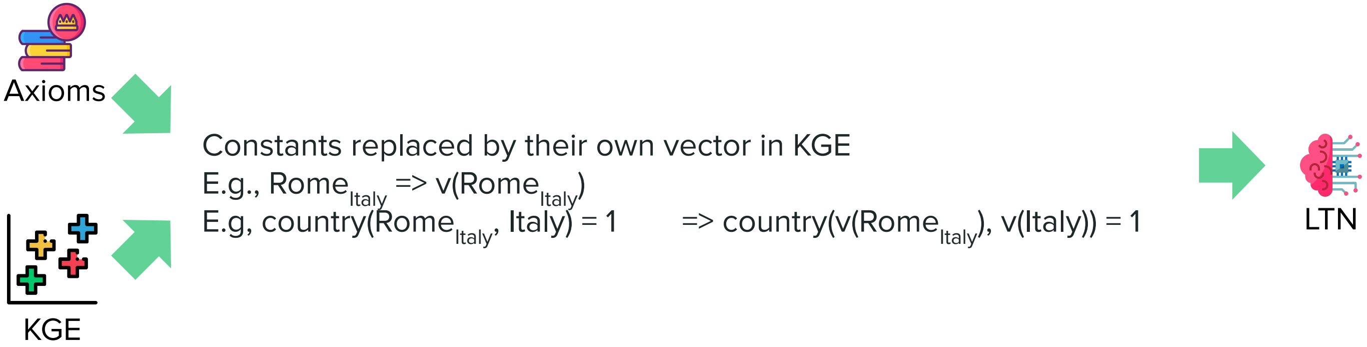


$\forall x\ City(x) \rightarrow \exists y: country(x, y)$	(A city must be in a country)
$\forall x\ Country(x) \rightarrow \exists y: capital(y, x)$	(A country must have a capital)
$\forall x,y\ capital(x, y) \rightarrow country(x, y)$	(A capital must be a city of its country)
$\forall x\ \neg country(x,x)$	(The countryOf property is non-reflexive)
…	

$country(Rome_{Italy}, Italy) = 1$	(Rome is located in Italy)
$country(Paris_{France}, Italy) = 0$	(Paris is not located in Italy)
$City(Rome_{Italy}) = 1$	(Rome is a city)
$City(Italy) = 0$	(Italy is not a city)
…	

LTN-based Reconciliation

3 Train the LTN with axioms and KGE, and obtained the trained model (which represents a new language!)



Axioms

KGE

Constants replaced by their own vector in KGE

E.g., $\text{Rome}_{\text{Italy}}$ => $\text{v}(\text{Rome}_{\text{Italy}})$

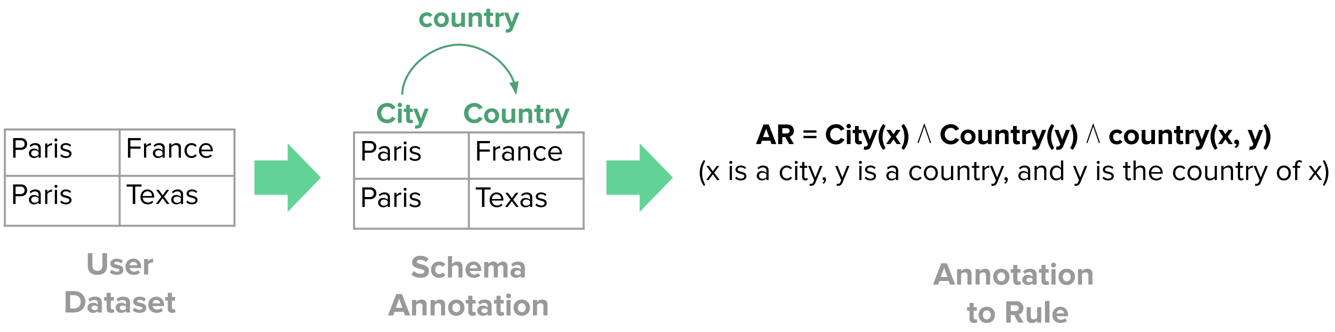
E.g, $\text{country}(\text{Rome}_{\text{Italy}}, \text{Italy}) = 1$ => $\text{country}(\text{v}(\text{Rome}_{\text{Italy}}), \text{v}(\text{Italy})) = 1$

LTN

LTN-based Reconciliation

4 Schema-level table annotation

- With the language defined by the LTN we can made infinite annotations by combining symbols



Paris	France
Paris	Texas

User Dataset

City	Country
Paris	France
Paris	Texas

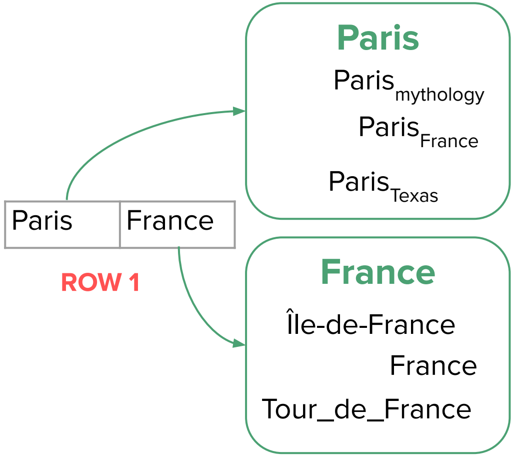
Schema Annotation

AR = City(x) ∧ Country(y) ∧ country(x, y)
(x is a city, y is a country, and y is the country of x)

Annotation to Rule

LTN-based Reconciliation

5 Iterate over table rows and test AR for all candidates (pairwise)



Paris	France

ROW 1

Paris

$\text{Paris}_{\text{mythology}}$

$\text{Paris}_{\text{France}}$

$\text{Paris}_{\text{Texas}}$

France

Île-de-France

France

Tour_de_France

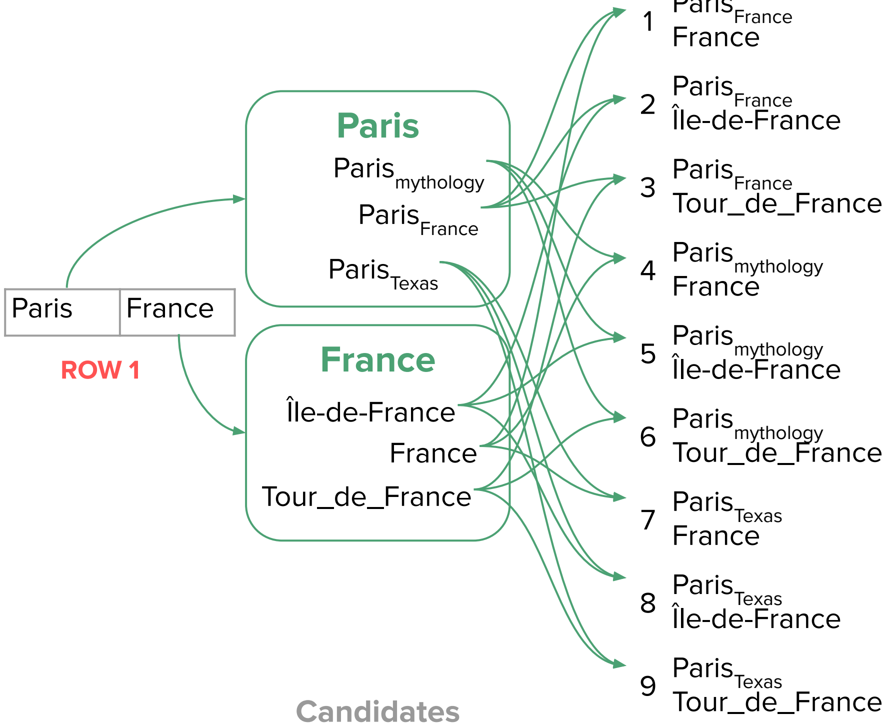
Candidates Generation

Pairwise Test

Scoring

LTN-based Reconciliation

5 Iterate over table rows and test AR for all candidates (pairwise)



Paris	France

ROW 1

Paris

Paris$_{\text{mythology}}$

Paris$_{\text{France}}$

Paris$_{\text{Texas}}$

France

Île-de-France

France

Tour_de_France

1 Paris$_{\text{France}}$ France

2 Paris$_{\text{France}}$ Île-de-France

3 Paris$_{\text{France}}$ Tour_de_France

4 Paris$_{\text{mythology}}$ France

5 Paris$_{\text{mythology}}$ Île-de-France

6 Paris$_{\text{mythology}}$ Tour_de_France

7 Paris$_{\text{Texas}}$ France

8 Paris$_{\text{Texas}}$ Île-de-France

9 Paris$_{\text{Texas}}$ Tour_de_France

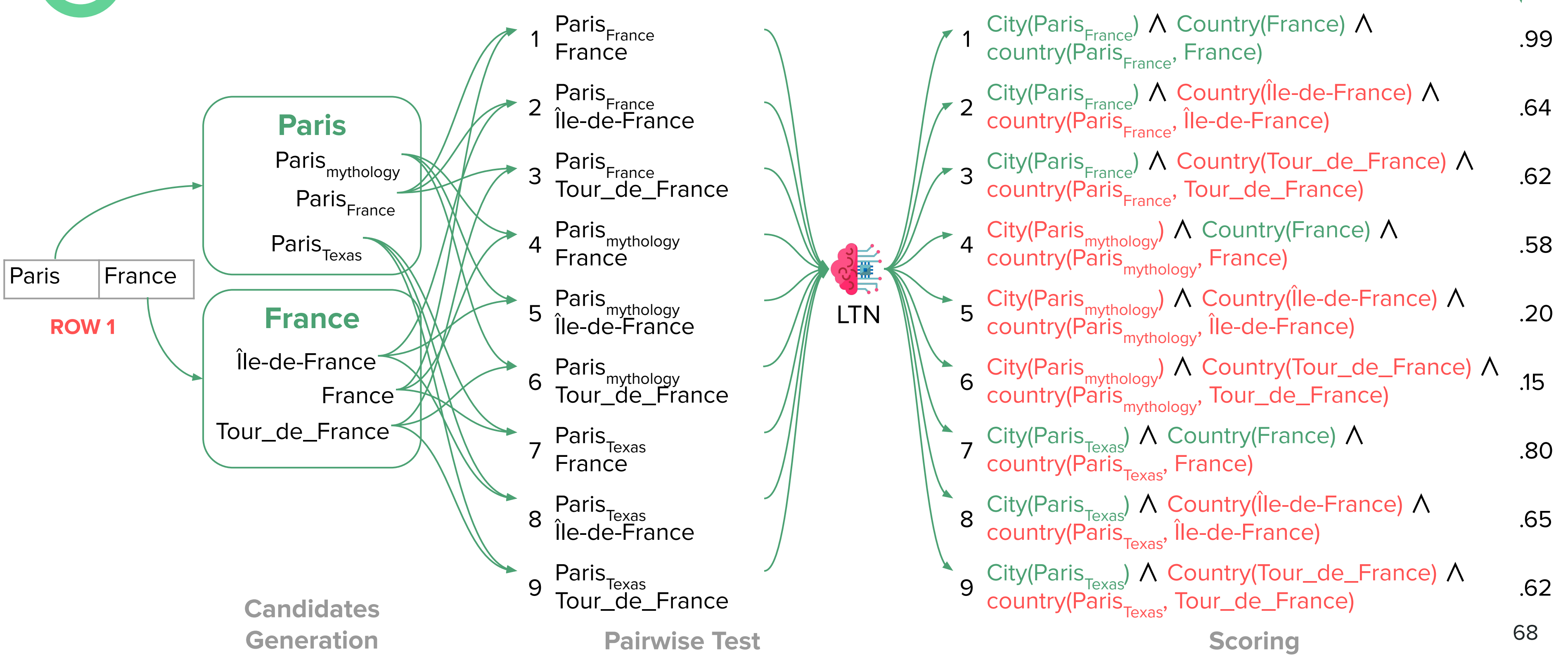
Candidates Generation

Pairwise Test

Scoring

LTN-based Reconciliation

5 Iterate over table rows and test AR for all candidates (pairwise)



the higher the better

Paris	France

ROW 1

Paris
- $\text{Paris}_{\text{mythology}}$
- $\text{Paris}_{\text{France}}$
- $\text{Paris}_{\text{Texas}}$

France
- Île-de-France
- France
- Tour_de_France

Candidates Generation

1. $\text{Paris}_{\text{France}}$ France
2. $\text{Paris}_{\text{France}}$ Île-de-France
3. $\text{Paris}_{\text{France}}$ Tour_de_France
4. $\text{Paris}_{\text{mythology}}$ France
5. $\text{Paris}_{\text{mythology}}$ Île-de-France
6. $\text{Paris}_{\text{mythology}}$ Tour_de_France
7. $\text{Paris}_{\text{Texas}}$ France
8. $\text{Paris}_{\text{Texas}}$ Île-de-France
9. $\text{Paris}_{\text{Texas}}$ Tour_de_France

Pairwise Test

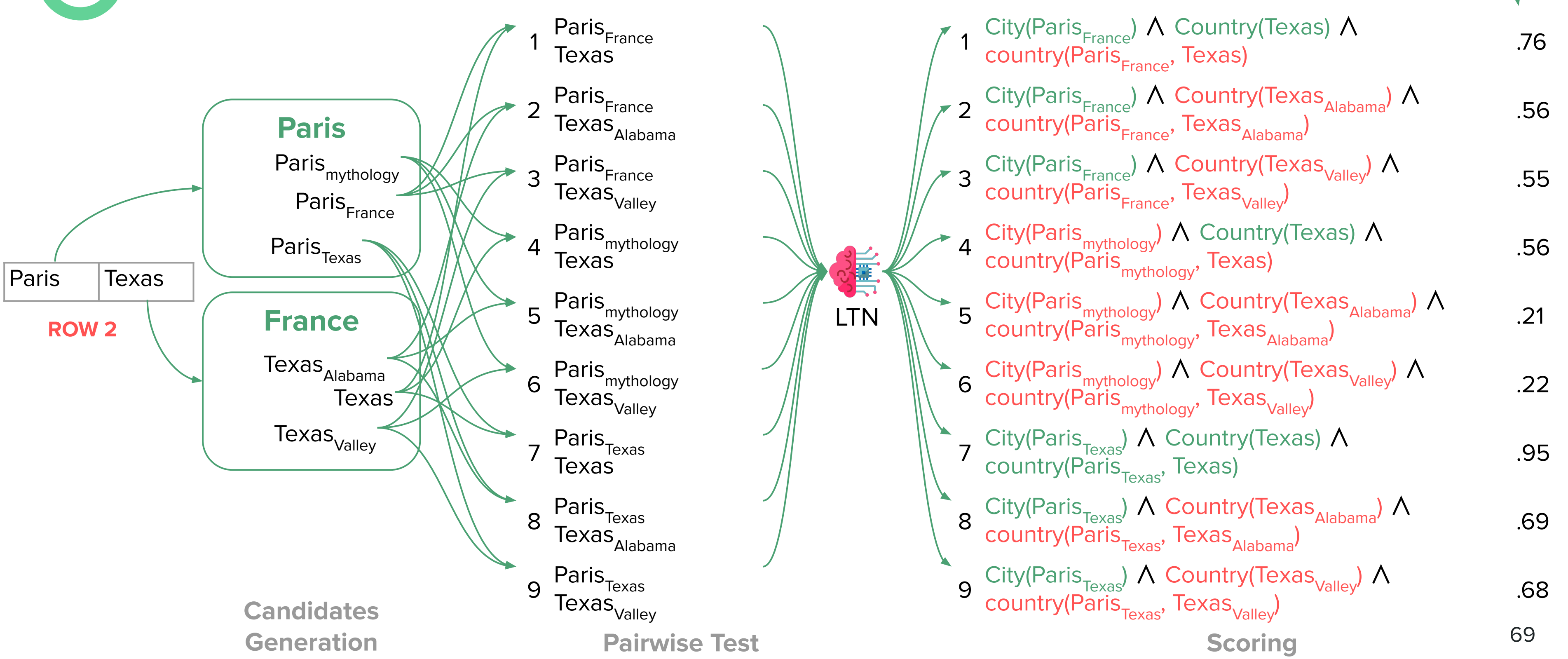
LTN

1	City($\text{Paris}_{\text{France}}$) ∧ Country(France) ∧ country($\text{Paris}_{\text{France}}$, France)	.99
2	City($\text{Paris}_{\text{France}}$) ∧ Country(Île-de-France) ∧ country($\text{Paris}_{\text{France}}$, Île-de-France)	.64
3	City($\text{Paris}_{\text{France}}$) ∧ Country(Tour_de_France) ∧ country($\text{Paris}_{\text{France}}$, Tour_de_France)	.62
4	City($\text{Paris}_{\text{mythology}}$) ∧ Country(France) ∧ country($\text{Paris}_{\text{mythology}}$, France)	.58
5	City($\text{Paris}_{\text{mythology}}$) ∧ Country(Île-de-France) ∧ country($\text{Paris}_{\text{mythology}}$, Île-de-France)	.20
6	City($\text{Paris}_{\text{mythology}}$) ∧ Country(Tour_de_France) ∧ country($\text{Paris}_{\text{mythology}}$, Tour_de_France)	.15
7	City($\text{Paris}_{\text{Texas}}$) ∧ Country(France) ∧ country($\text{Paris}_{\text{Texas}}$, France)	.80
8	City($\text{Paris}_{\text{Texas}}$) ∧ Country(Île-de-France) ∧ country($\text{Paris}_{\text{Texas}}$, Île-de-France)	.65
9	City($\text{Paris}_{\text{Texas}}$) ∧ Country(Tour_de_France) ∧ country($\text{Paris}_{\text{Texas}}$, Tour_de_France)	.62

Scoring

LTN-based Reconciliation

5 Iterate over table rows and test AR for all candidates (pairwise)



1 Paris$_{France}$ Texas

2 Paris$_{France}$ Texas$_{Alabama}$

3 Paris$_{France}$ Texas$_{Valley}$

4 Paris$_{mythology}$ Texas

5 Paris$_{mythology}$ Texas$_{Alabama}$

6 Paris$_{mythology}$ Texas$_{Valley}$

7 Paris$_{Texas}$ Texas

8 Paris$_{Texas}$ Texas$_{Alabama}$

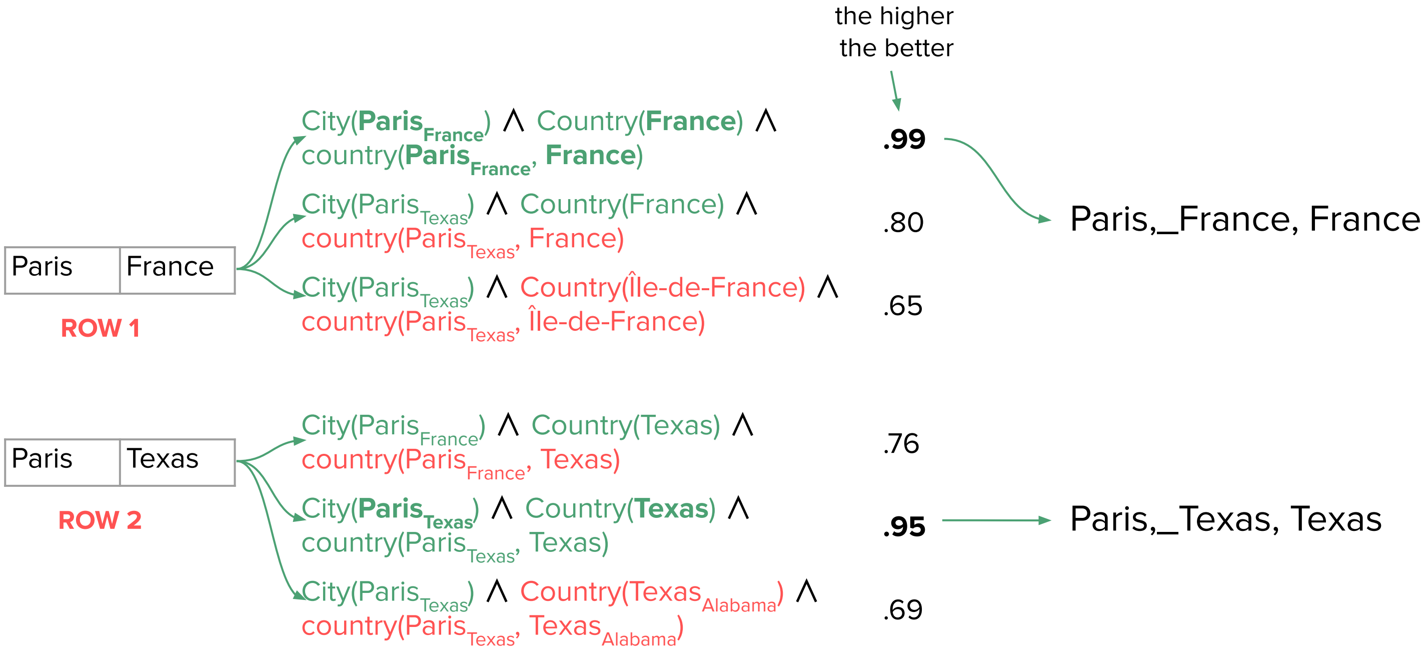
9 Paris$_{Texas}$ Texas$_{Valley}$

the higher the better

1	$\text{City}(\text{Paris}_{\text{France}}) \land \text{Country}(\text{Texas}) \land \text{country}(\text{Paris}_{\text{France}}, \text{Texas})$	.76
2	$\text{City}(\text{Paris}_{\text{France}}) \land \text{Country}(\text{Texas}_{\text{Alabama}}) \land \text{country}(\text{Paris}_{\text{France}}, \text{Texas}_{\text{Alabama}})$	.56
3	$\text{City}(\text{Paris}_{\text{France}}) \land \text{Country}(\text{Texas}_{\text{Valley}}) \land \text{country}(\text{Paris}_{\text{France}}, \text{Texas}_{\text{Valley}})$	.55
4	$\text{City}(\text{Paris}_{\text{mythology}}) \land \text{Country}(\text{Texas}) \land \text{country}(\text{Paris}_{\text{mythology}}, \text{Texas})$	.56
5	$\text{City}(\text{Paris}_{\text{mythology}}) \land \text{Country}(\text{Texas}_{\text{Alabama}}) \land \text{country}(\text{Paris}_{\text{mythology}}, \text{Texas}_{\text{Alabama}})$	.21
6	$\text{City}(\text{Paris}_{\text{mythology}}) \land \text{Country}(\text{Texas}_{\text{Valley}}) \land \text{country}(\text{Paris}_{\text{mythology}}, \text{Texas}_{\text{Valley}})$	.22
7	$\text{City}(\text{Paris}_{\text{Texas}}) \land \text{Country}(\text{Texas}) \land \text{country}(\text{Paris}_{\text{Texas}}, \text{Texas})$	.95
8	$\text{City}(\text{Paris}_{\text{Texas}}) \land \text{Country}(\text{Texas}_{\text{Alabama}}) \land \text{country}(\text{Paris}_{\text{Texas}}, \text{Texas}_{\text{Alabama}})$	.69
9	$\text{City}(\text{Paris}_{\text{Texas}}) \land \text{Country}(\text{Texas}_{\text{Valley}}) \land \text{country}(\text{Paris}_{\text{Texas}}, \text{Texas}_{\text{Valley}})$	.68

LTN-based Reconciliation

6 Select the best candidates for each row



the higher the better

Paris	France

ROW 1

$\text{City}(\mathbf{Paris}_{\mathbf{France}}) \wedge \text{Country}(\mathbf{France}) \wedge \text{country}(\mathbf{Paris}_{\mathbf{France}}, \mathbf{France})$ **.99** → Paris,_France, France

$\text{City}(\text{Paris}_{\text{Texas}}) \wedge \text{Country}(\text{France}) \wedge \text{country}(\text{Paris}_{\text{Texas}}, \text{France})$ .80

$\text{City}(\text{Paris}_{\text{Texas}}) \wedge \text{Country}(\text{Île-de-France}) \wedge \text{country}(\text{Paris}_{\text{Texas}}, \text{Île-de-France})$ .65

Paris	Texas

ROW 2

$\text{City}(\text{Paris}_{\text{France}}) \wedge \text{Country}(\text{Texas}) \wedge \text{country}(\text{Paris}_{\text{France}}, \text{Texas})$ .76

$\text{City}(\mathbf{Paris}_{\mathbf{Texas}}) \wedge \text{Country}(\mathbf{Texas}) \wedge \text{country}(\text{Paris}_{\text{Texas}}, \text{Texas})$ **.95** → Paris,_Texas, Texas

$\text{City}(\text{Paris}_{\text{Texas}}) \wedge \text{Country}(\text{Texas}_{\text{Alabama}}) \wedge \text{country}(\text{Paris}_{\text{Texas}}, \text{Texas}_{\text{Alabama}})$ .69

Experimental results: Datasets

Dataset:

- 8 african countries
- No more than 50 cities for each country

Embedding:

- Pretrained DBpedia embeddings from RDF2VEC (200 dimensions)
    - Only cities and countries
- Embeddings downsized to 40 dimensions (using PCA)
    - Cosine similarity between vectors is preserved (similar vectors are still similar in the new space)

Experimental results: Training capital() and locatedIn()

Universally quantified axiom:

- ∀ ?a,?c,?d: locatedIn(?a,?c) -> (¬ equals(?c,?d) & ¬ locatedIn(?a,?d))
- ∀ ?a,?b,?c: capital(?a,?c) -> (¬ equals(?a,?b) & ¬ capital(?b,?c))
- ∀ ?a,?c: capital(?a,?c) -> locatedIn(?a,?c)
- ∀ ?a,?c: ¬ locatedIn(?a,?c) -> ¬ capital(?a,?c)

Legend:
a-b: all cities
c-d: all countries

TRAINING

locatedIn():	**Precision: 0.95**
threshold = 0.80	**Recall: 0.97**
capital():	**Precision: 0.62**
threshold = 0.95	**Recall: 1.00**

(In almost all cases the right pair is the one with the highest score)

TEST

16 cities never trained (about 2 cities for each country)

- **16 / 16** cities properly assessed (**locatedIn()**)
- **16 / 16** cities properly assessed (**capital()**)

Experimental results: Training unary predicates

Universally quantified axiom:

- ∀ ?a,?c,?d: locatedIn(?a,?c) -> (¬ equals(?c,?d) & ¬ locatedIn(?a,?d))
- ∀ ?a,?c: locatedIn(?a, ?c) -> City(?a) & Country(?c)
- ∀ ?a: ¬ Country(?a)
- ∀ ?c: ¬ City(?c)

Legend:
a-b: all cities
c-d: all countries

TRAINING

locatedIn(): **Precision: 1.00**
threshold = 0.80 **Recall: 0.99**

377/378 cities satisfy City()
8/8 countries satisfy Country()

TEST

16 cities never trained (about 2 cities for each country)

- **16 / 16** cities properly assessed (**locatedIn()**)
- **all cities** have **City()** value > **0.5** and **Country()** value < **0.5**
- **all countries** have **Country()** value > **0.5** and **City()** value < **0.5**

Experimental results: Training all predicates

Universally quantified axiom:

- ∀ ?a,?c,?d: locatedIn(?a,?c) -> (¬ equals(?c,?d) & ¬ locatedIn(?a,?d))
- ∀ ?a,?b,?c: capital(?a,?c) -> (¬ equals(?a,?b) & ¬ capital(?b,?c))
- ∀ ?a,?c: capital(?a,?c) -> locatedIn(?a,?c)
- ∀ ?a,?c: ¬ locatedIn(?a,?c) -> ¬ capital(?a,?c)
- ∀ ?y: Capital(?y)
- ∀ ?x: ¬ Capital(?x)
- ∀ ?a: City(?a)
- ∀ ?a: ¬ Country(?a)
- ∀ ?c: Country(?c)
- ∀ ?c: ¬ City(?c)
- ∀ ?c: ¬ Capital(?c)

Legend:
a-b: all cities
c-d: all countries
y: all capitals
x: all non capitals

Experimental results: Training all predicates (cont)

	TRAINING
locatedIn():	**Precision: 0.94**
threshold = 0.70	**Recall: 0.95**
capital():	**Precision: 0.60**
threshold = 0.90	**Recall: 1.00**

377/378 cities that satisfy City()
8/8 countries that satisfy Country()
8/8 cities that satisfy Capital()

TEST

16 cities never trained (about 2 cities for each country)

- **16 / 16** cities properly assessed (**locatedIn()**)
- **16 / 16** cities properly assessed (**capital()**)
- **all cities** have:
  - **City()** value > **0.5**
  - **Country()** value < **0.5**
  - **Capital()** value < **0.5**
- **all countries** have:
  - **Country()** value > **0.5**
  - **City()** value < **0.5**
  - **Capital()** value < **0.5**

Thanks!

Thanks to Freepik, Nikita Golubev, itim2101, Pixel perfect, Smashicons, and mynamepong for their awesome icons available for free at Flati

References

d’Avila Garcez, A., Lamb, L. C., & Gabbay, D. M. (2008). Neural-symbolic cognitive reasoning. Springer Science & Business Media.

d’Avila Garcez, A., Broda, K. B., & Gabbay, D. M. (2012). Neural-symbolic learning systems: foundations and applications. Springer Science & Business Media.

Serafini, L., & d'Avila Garcez A. (2016). Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge. 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy16) arXiv:1606.04422.

Socher, R., Chen, D., Manning, C. D., & Ng, A. Y. (2013). Reasoning with neural tensor networks for knowledge base completion. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS’13). pp. 926–934.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724-2743.