# Explainable, Data-Efficient, Verifiable Representation Learning

Pasquale Minervini, UCL
p.minervini@ucl.ac.uk

ICL, Nov 27th

# Symbolic and Sub-Symbolic AI

# Symbolic and Sub-Symbolic AI

$$\forall X, Y, Z : \textbf{grandFather}(X, Y) \Leftarrow \textbf{father}(X, Z), \textbf{parent}(Z, Y)$$

**Symbolic**

*Ontologies, First-Order Logic, Logic Programming, Knowledge Bases, Theorem Proving..*

# Symbolic and Sub-Symbolic AI

$$\forall X, Y, Z : \textbf{grandFather}(X, Y) \Leftarrow \textbf{father}(X, Z), \textbf{parent}(Z, Y)$$

✓ Data-Efficient

✓ Interpretable and Explainable

✓ Easy to Incorporate Knowledge

✓ Verifiable

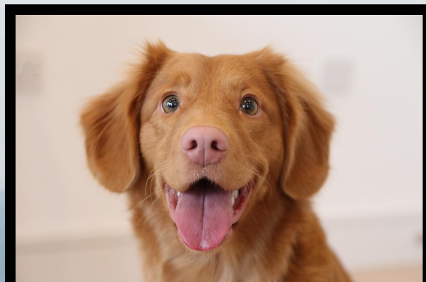
**Symbolic**

*Ontologies, First-Order Logic, Logic Programming, Knowledge Bases, Theorem Proving..*

# Symbolic and Sub-Symbolic AI

**Sub-Symbolic/Connectionist**

*Neural Representation Learning, (Deep) Latent Variable Models..*

$f_1(f_2(\ldots f_n($  $)\ldots)) = \begin{cases} 0.15 \equiv \textbf{cat} \\ 0.85 \equiv \textbf{dog} \end{cases}$

$$\forall X, Y, Z : \textbf{grandFather}(X, Y) \Leftarrow \textbf{father}(X, Z), \textbf{parent}(Z, Y)$$

✓ Data-Efficient

✓ Interpretable and Explainable

✓ Easy to Incorporate Knowledge

✓ Verifiable

**Symbolic**

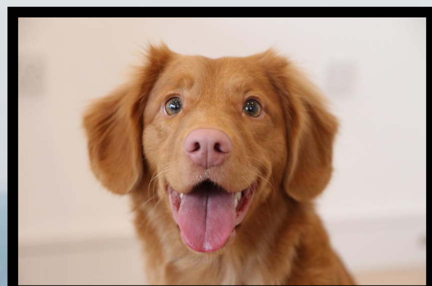
*Ontologies, First-Order Logic, Logic Programming, Knowledge Bases, Theorem Proving..*

# Symbolic and Sub-Symbolic AI

**Sub-Symbolic/Connectionist**

*Neural Representation Learning, (Deep) Latent Variable Models..*

- ✓ Noisy, Ambiguous, Sensory Data
- ✓ Highly Parallel
- ✓ High Predictive Accuracy

$f_1(f_2(\ldots f_n($  $)\ldots)) = \begin{cases} 0.15 \equiv \textbf{cat} \\ 0.85 \equiv \textbf{dog} \end{cases}$

$$\forall X, Y, Z : \textbf{grandFather}(X, Y) \Leftarrow \textbf{father}(X, Z), \textbf{parent}(Z, Y)$$

- ✓ Data-Efficient
- ✓ Interpretable and Explainable
- ✓ Easy to Incorporate Knowledge
- ✓ Verifiable

**Symbolic**

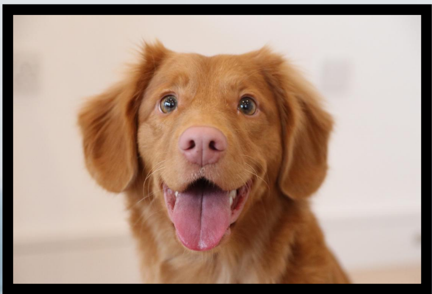
*Ontologies, First-Order Logic, Logic Programming, Knowledge Bases, Theorem Proving..*

# Symbolic and Sub-Symbolic AI

**Sub-Symbolic/Connectionist**

*Neural Representation Learning, (Deep) Latent Variable Models..*

- ✓ Noisy, Ambiguous, Sensory Data
- ✓ Highly Parallel
- ✓ High Predictive Accuracy

$f_1(f_2(\ldots f_n($  $)\ldots)) = \begin{cases} 0.15 \equiv \textbf{cat} \\ 0.85 \equiv \textbf{dog} \end{cases}$

$$\forall X, Y, Z : \textbf{grandFather}(X, Y) \Leftarrow \textbf{father}(X, Z), \textbf{parent}(Z, Y)$$

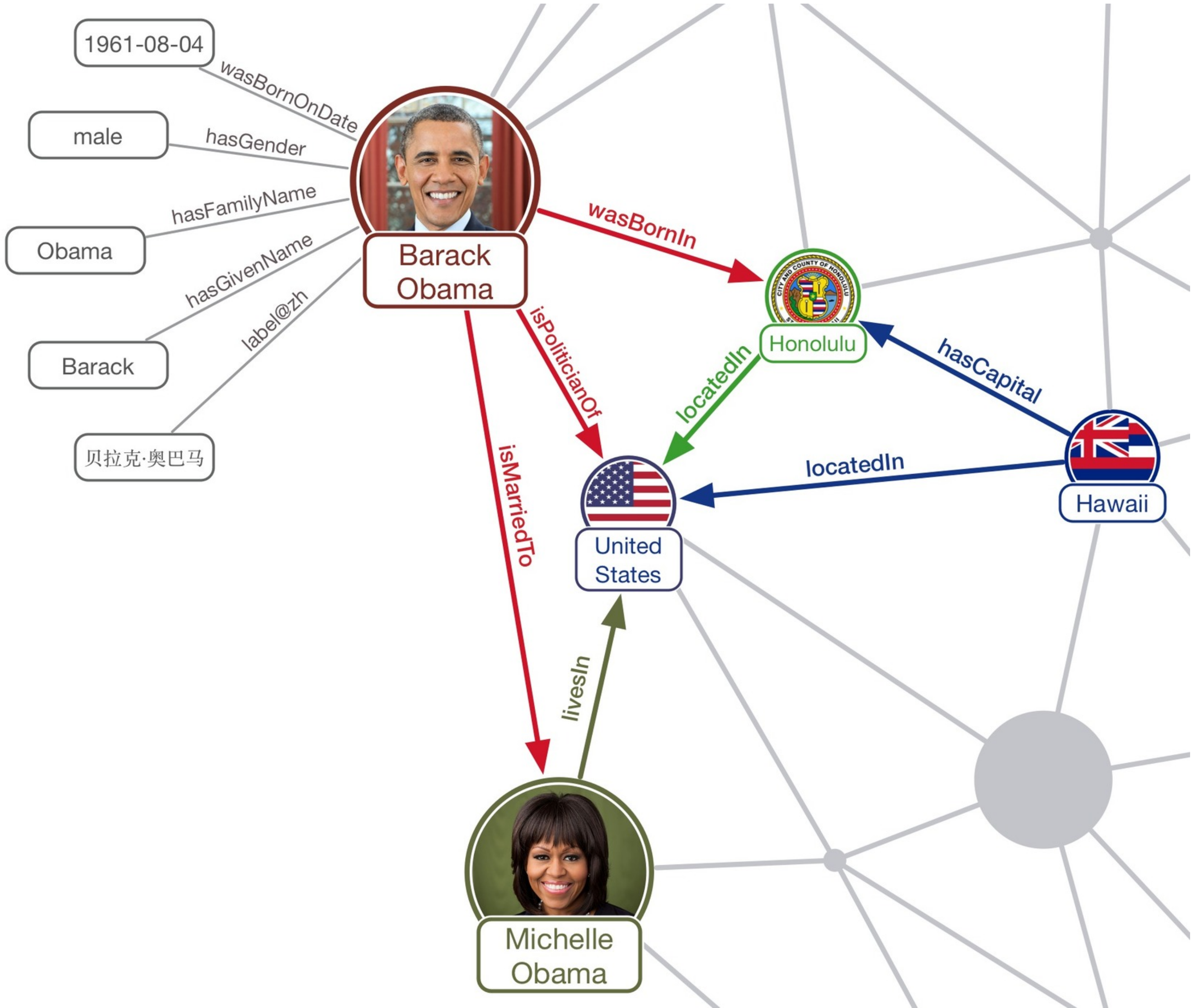
- ✓ Data-Efficient
- ✓ Interpretable and Explainable
- ✓ Easy to Incorporate Knowledge
- ✓ Verifiable

**Symbolic**

*Ontologies, First-Order Logic, Logic Programming, Knowledge Bases, Theorem Proving..*

# Knowledge Graphs

**Knowledge Graph** — graph structured Knowledge Base, where knowledge is encoded by relationships between entities.

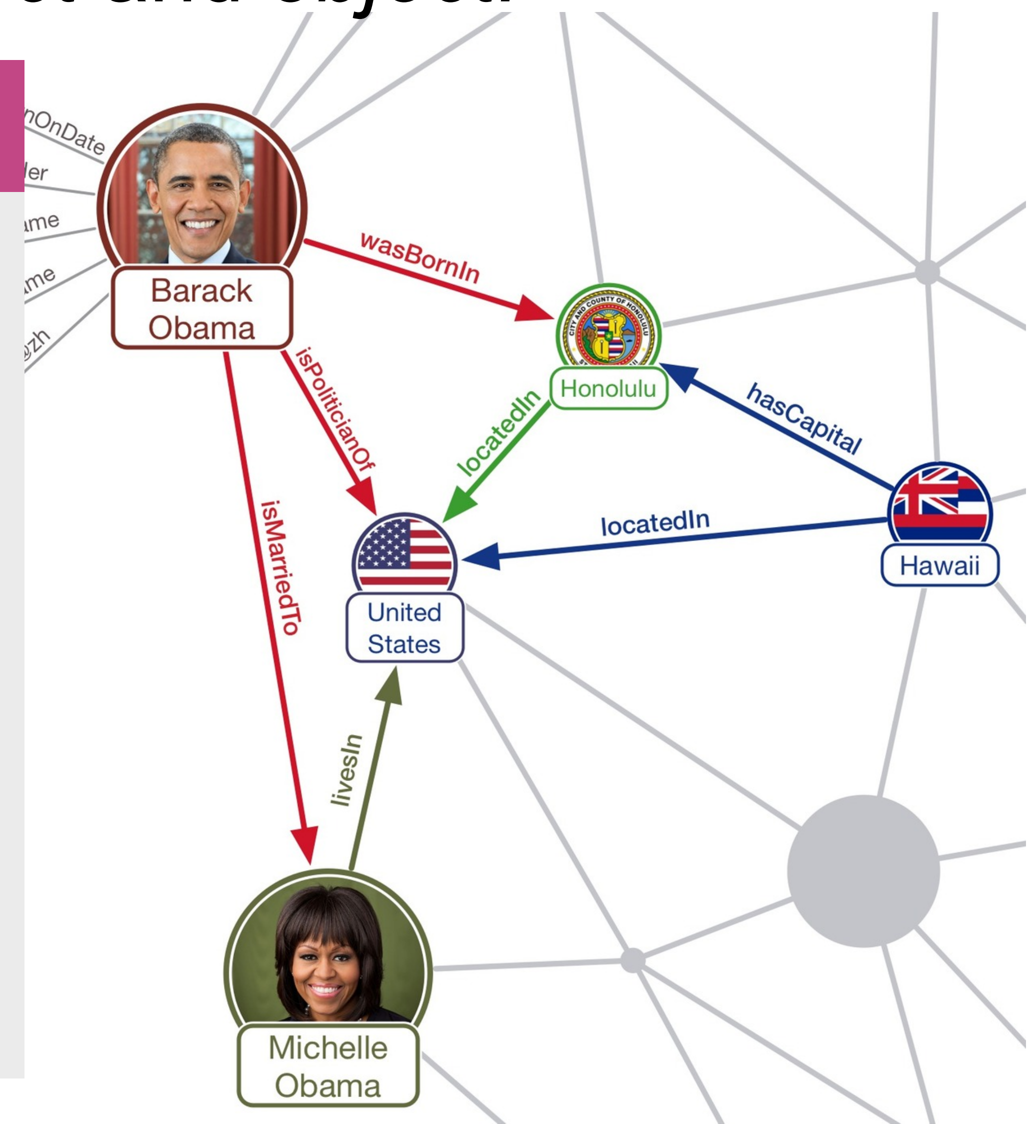


# Knowledge Graphs

**Knowledge Graph** — graph structured Knowledge Base, where knowledge is encoded by relationships between entities.

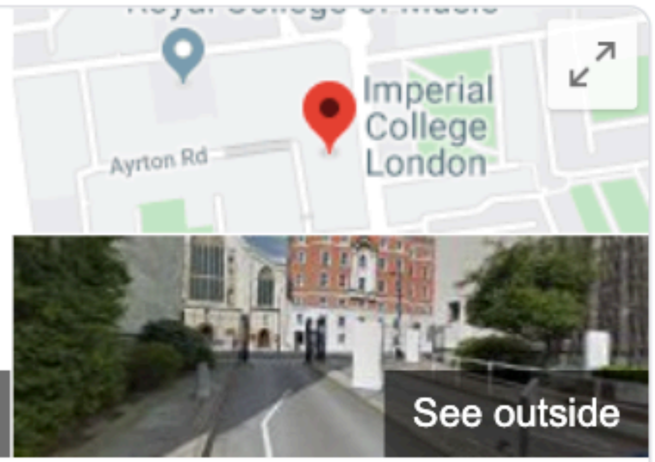
In practice — set of **subject-predicate-object** triples, denoting a relationship of type *predicate* between *subject* and *object*.

| subject | predicate | object |
|---|---|---|
| Barack Obama | was born in | Honolulu |
| Hawaii | has capital | Honolulu |
| Barack Obama | is politician of | United States |
| Hawaii | is located in | United States |
| Barack Obama | is married to | Michelle Obama |
| Michelle Obama | is a | Lawyer |
| Michelle Obama | lives in | United States |



# Knowledge Graphs

Imperial College London




See photos

## Imperial College London

Website | Directions | Save

Public university in London, England · 2.8 mi

Imperial College London is a public research university located in London. In 1851, Prince Albert built his vision of an area for culture, including the Victoria and Albert Museum, Natural History Museum, Royal Colleges, Royal Albert Hall, and the Imperial Institute.
Wikipedia

**Address:** South Kensington, London SW7 2AZ

**Acceptance rate:** 14.3% (2015)

**Subsidiaries:** Imperial College Business School, MORE

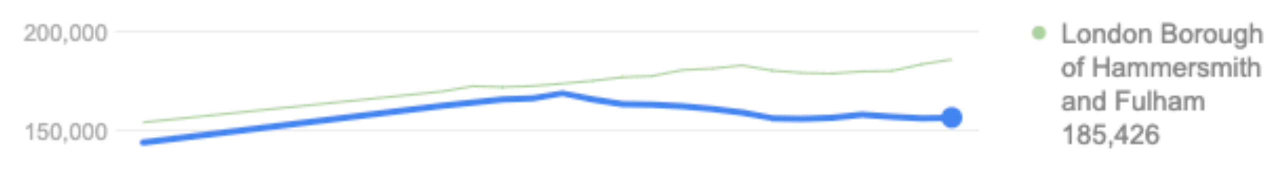
population kensington and chelsea

All | Images | News | Maps | Shopping | More | Settings | Tools

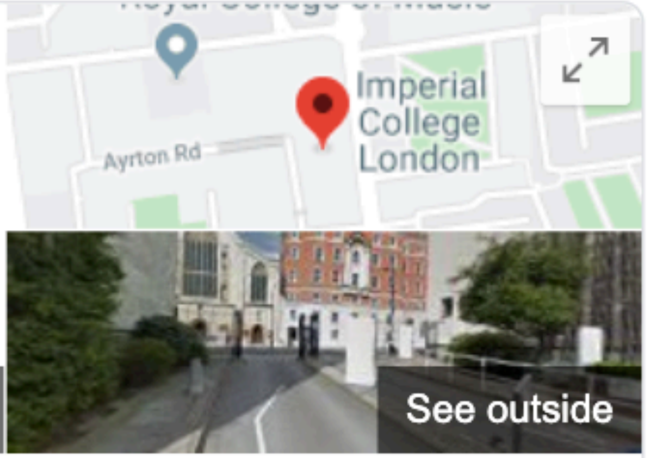
About 4,410,000 results (0.57 seconds)

Royal Borough of Kensington and Chelsea / **Population**

156,197 (2018)



# Knowledge Graphs




## Imperial College London

Website | Directions | Save

Public university in London, England · 2.8 mi

Imperial College London is a public research university located in London. In 1851, Prince Albert built his vision of an area for culture, including the Victoria and Albert Museum, Natural History Museum, Royal Colleges, Royal Albert Hall, and the Imperial Institute. Wikipedia

**Address:** South Kensington, London SW7 2AZ

**Acceptance rate:** 14.3% (2015)

**Subsidiaries:** Imperial College Business School, MORE

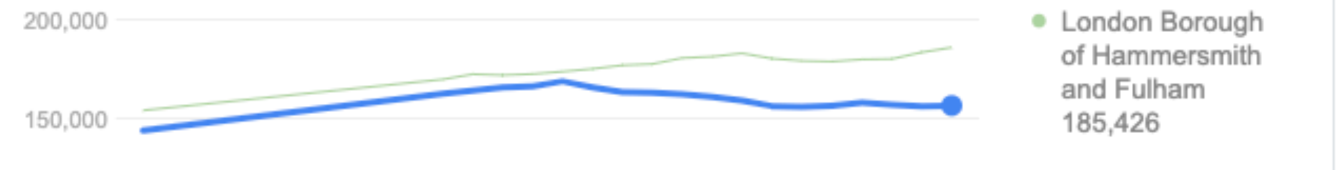
population kensington and chelsea

All · Images · News · Maps · Shopping · More · Settings · Tools

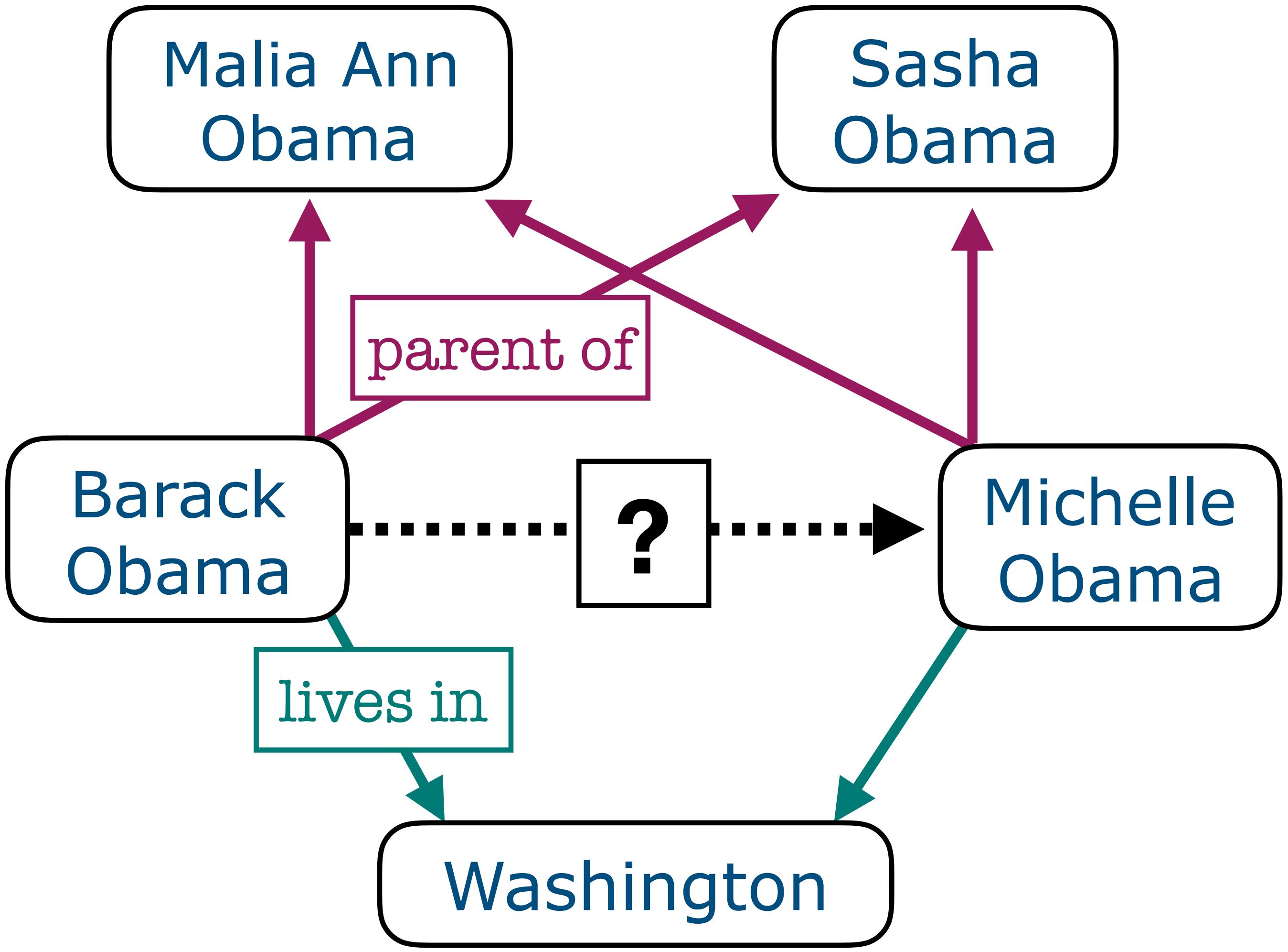
About 4,410,000 results (0.57 seconds)

Royal Borough of Kensington and Chelsea / **Population**

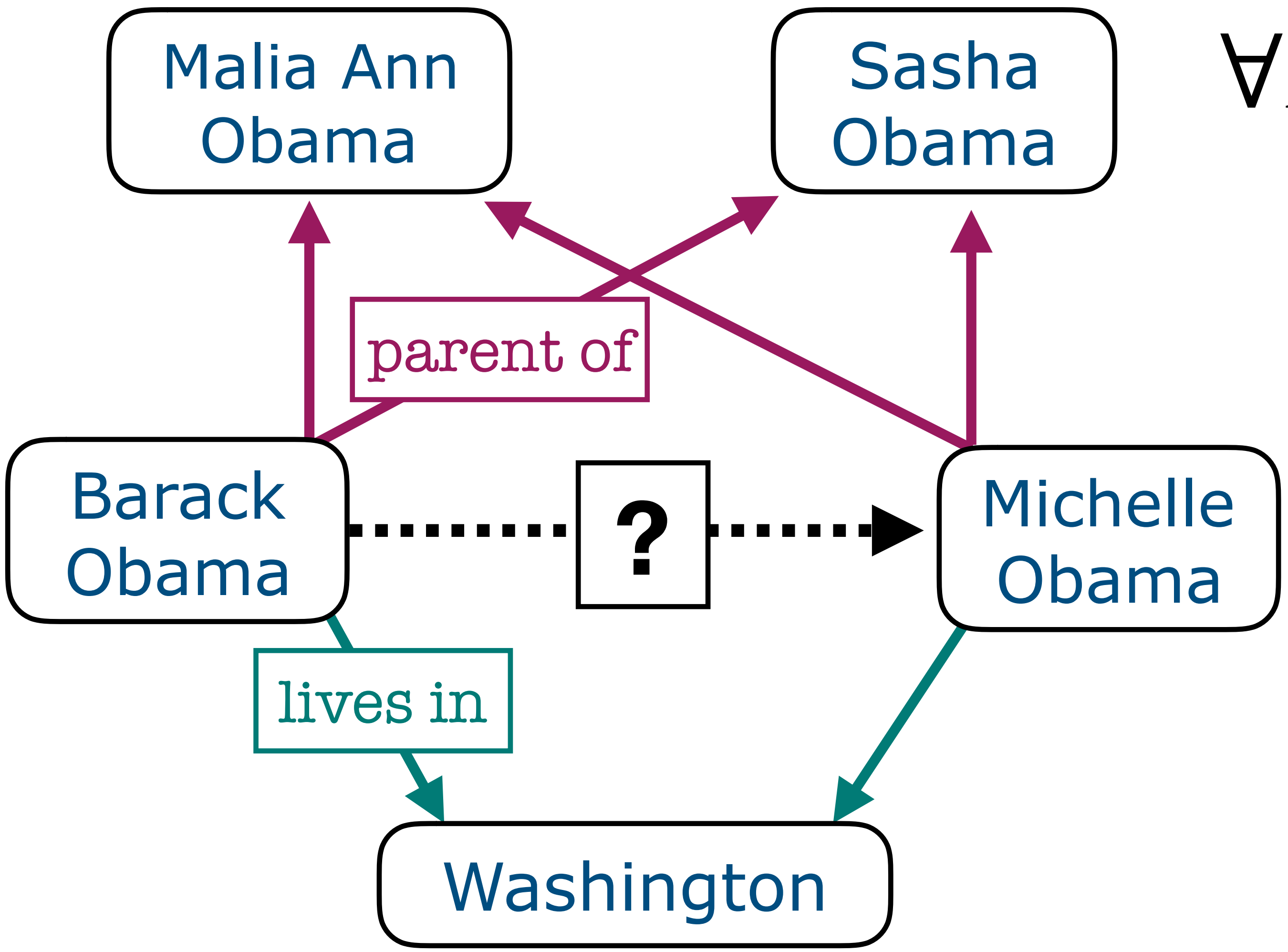
156,197 (2018)



# Link Prediction in Knowledge Graphs

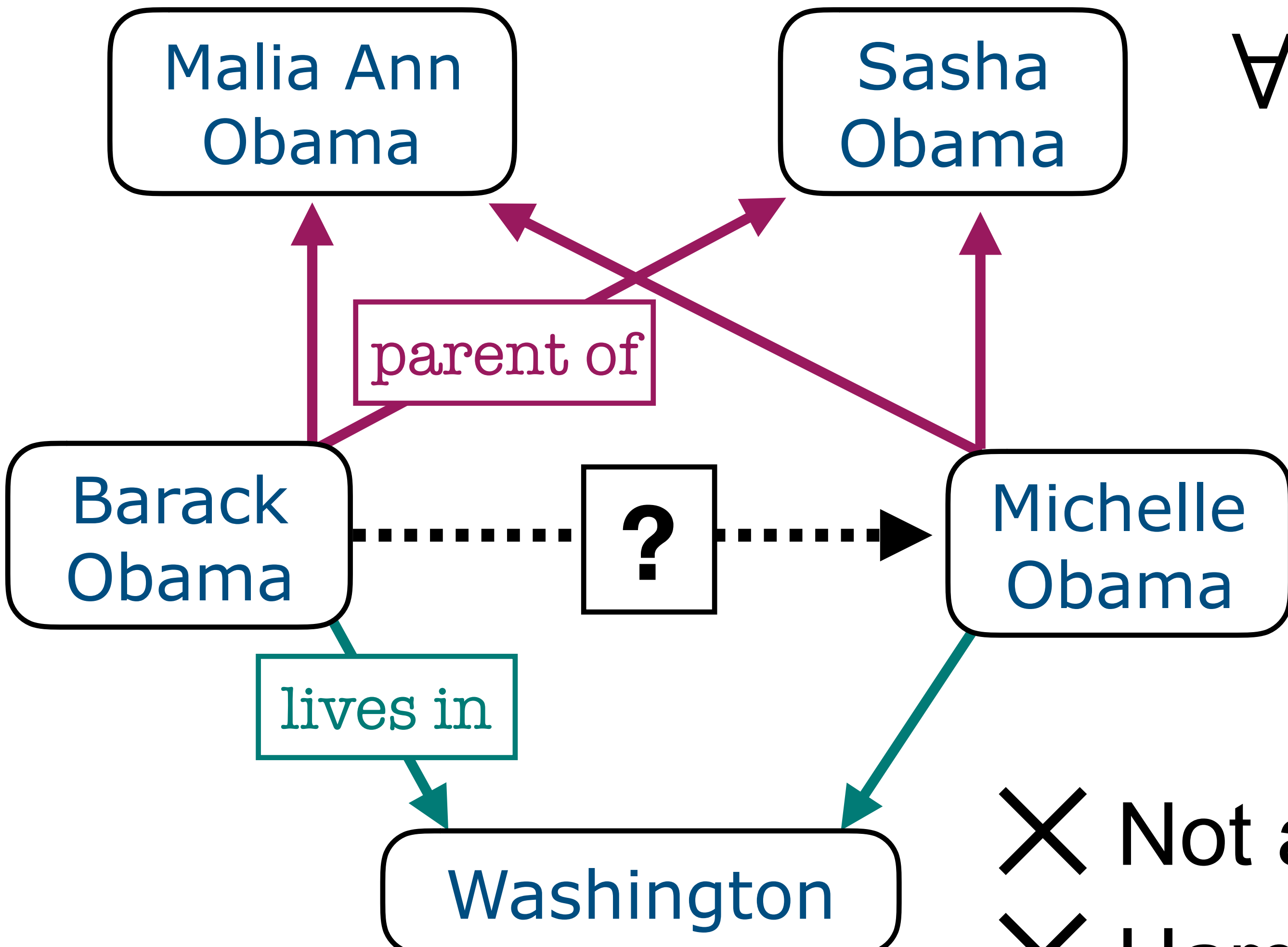


# Rule-Based Link Prediction




$$\forall X, Y, Z : \text{married with}(X, Y) \Leftarrow \text{parent of}(X, Z), \text{parent of}(Y, Z)$$

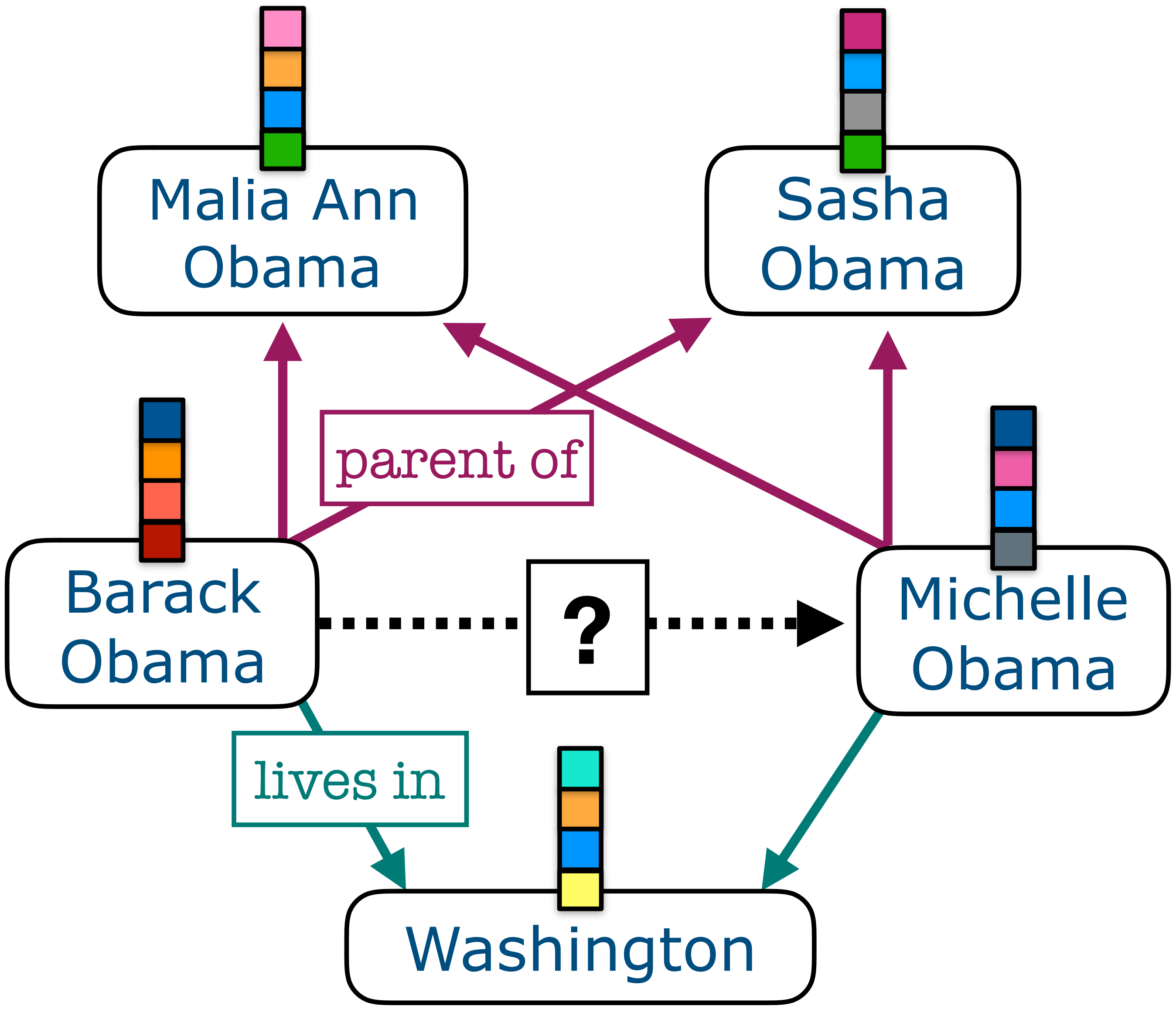
# Rule-Based Link Prediction



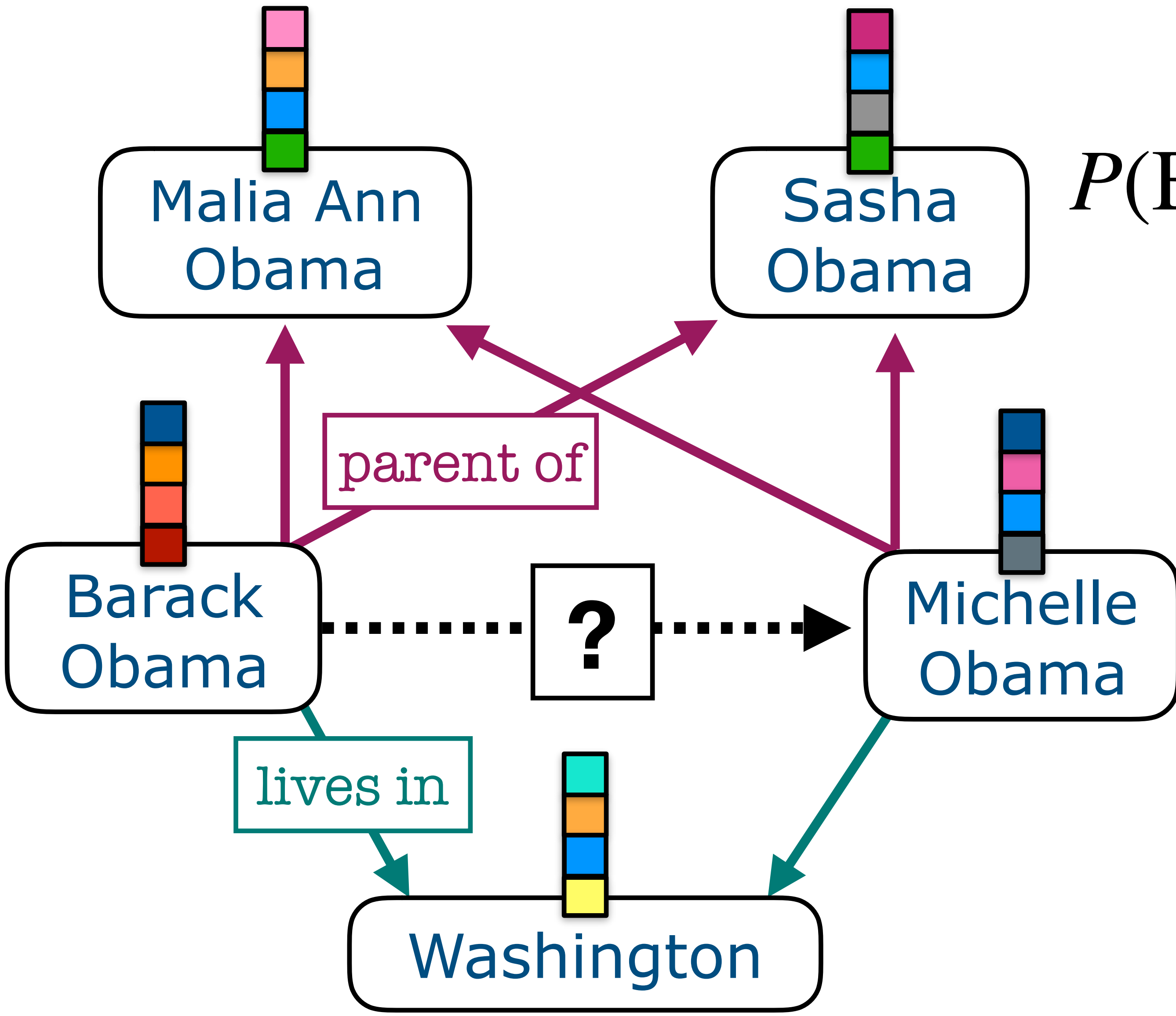

$$\forall X, Y, Z : \text{married with}(X, Y) \Leftarrow \text{parent of}(X, Z), \text{parent of}(Y, Z)$$

- ✗ Not always true
- ✗ Hard to learn from data
- ✗ Hard to formalise for other modalities

# Neural Link Prediction

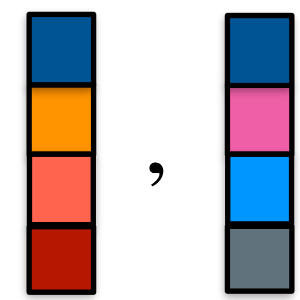


# Neural Link Prediction

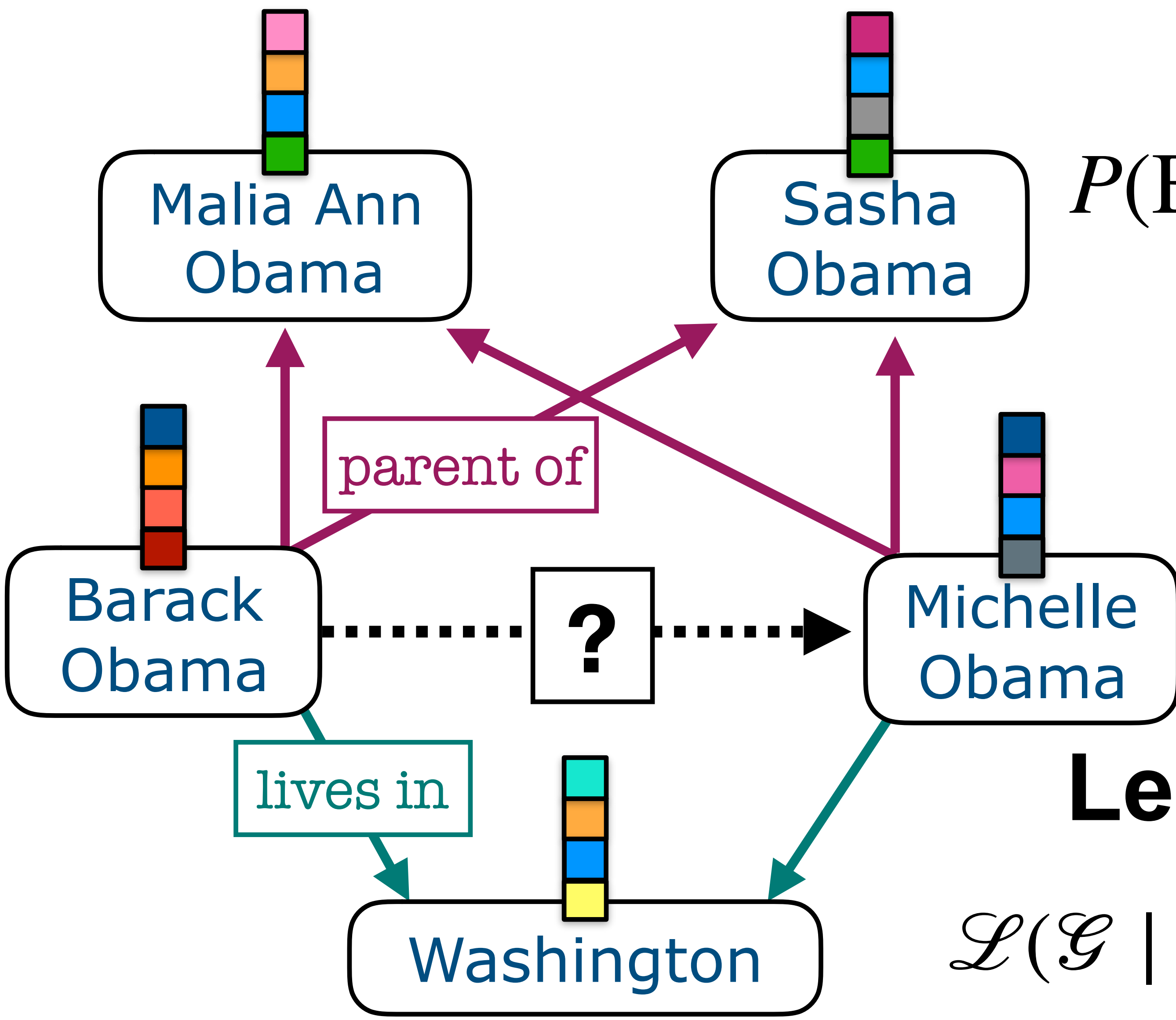



$$P(\mathrm{BO} \xrightarrow{\text{married}} \mathrm{MO}) \propto$$

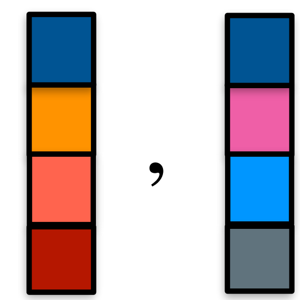
$$f_{\text{married}}\Big(\quad , \quad\Big)$$



# Neural Link Prediction




$$P(\text{BO} \xrightarrow{\text{married}} \text{MO}) \propto$$

$f_{\text{married}}$ (  )

## Learning Representations

$$\mathscr{L}(\mathscr{G} \mid \Theta) = \sum_{(s,p,o) \in \mathscr{G}} \log \sigma\Big(f_p(\mathbf{e}_s, \mathbf{e}_o)\Big)$$

$$+ \sum_{(s,p,o) \notin \mathscr{G}} \log \Big[1 - \sigma\Big(f_p(\mathbf{e}_s, \mathbf{e}_o)\Big)\Big]$$

# Neural Link Prediction — Scoring Functions

The interaction between the *latent features* is defined by the scoring function $f(\cdot)$ — several variants in the literature:

| Models | Scoring Functions | Parameters |
|---|---|---|
| RESCAL [Nickel et al. 2011] | $\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o$ | $\mathbf{W}_p \in \mathbb{R}^{k \times k}$ |
| TransE [Bordes et al. 2013] | $- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_p^2$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| DistMult [Yang et al. 2015] | $\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| HolE [Nickel et al. 2016] | $\mathbf{r}_p^\top \left( \mathscr{F}^{-1} \left[ \overline{\mathscr{F}[\mathbf{e}_s]} \odot \mathscr{F}[\mathbf{e}_o] \right] \right)$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| **ComplEx [Trouillon et al. 2016]** | $\text{Re}\left( \langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle \right)$ | $\mathbf{r}_p \in \mathbb{C}^k$ |
| **ConvE [Dettmers et al. 2017]** | $f\left( \text{vec}\left( f\left( [\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega \right) \right) \mathbf{W} \right) \mathbf{e}_o$ | $\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$ |

# Neural Link Prediction — Scoring Functions

The interaction between the *latent features* is defined by the scoring function $f(\cdot)$ — several variants in the literature:

| Models | Scoring Functions | Parameters |
|---|---|---|
| RESCAL [Nickel et al. 2011] | $\boxed{\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o}$ | $\mathbf{W}_p \in \mathbb{R}^{k \times k}$ |
| TransE [Bordes et al. 2013] | $- \left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_p^2$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| DistMult [Yang et al. 2015] | $\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| HolE [Nickel et al. 2016] | $\mathbf{r}_p^\top \left( \mathcal{F}^{-1} \left[ \overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o] \right] \right)$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| **ComplEx [Trouillon et al. 2016]** | $\mathrm{Re} \left( \langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle \right)$ | $\mathbf{r}_p \in \mathbb{C}^k$ |
| **ConvE [Dettmers et al. 2017]** | $f \left( \mathrm{vec} \left( f \left( [\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega \right) \right) \mathbf{W} \right) \mathbf{e}_o$ | $\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$ |

# Neural Link Prediction — Scoring Functions

The interaction between the *latent features* is defined by the scoring function $f(\cdot)$ — several variants in the literature:

| Models | Scoring Functions | Parameters |
|---|---|---|
| RESCAL [Nickel et al. 2011] | $\boxed{\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o}$ | $\mathbf{W}_p \in \mathbb{R}^{k \times k}$ |
| TransE [Bordes et al. 2013] | $-\left\| \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\|_p^2$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| DistMult [Yang et al. 2015] | $\boxed{\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle}$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| HolE [Nickel et al. 2016] | $\mathbf{r}_p^\top \left( \mathcal{F}^{-1} \left[ \overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o] \right] \right)$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| **ComplEx [Trouillon et al. 2016]** | $\text{Re}\left( \langle \mathbf{e}_s, \mathbf{r}_p, \overline{\mathbf{e}}_o \rangle \right)$ | $\mathbf{r}_p \in \mathbb{C}^k$ |
| **ConvE [Dettmers et al. 2017]** | $f\left( \text{vec}\left( f\left( [\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega \right) \right) \mathbf{W} \right) \mathbf{e}_o$ | $\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$ |

# Neural Link Prediction — Scoring Functions

The interaction between the *latent features* is defined by the scoring function $f(\cdot)$ — several variants in the literature:

| Models | Scoring Functions | Parameters |
|---|---|---|
| RESCAL [Nickel et al. 2011] | $\boxed{\mathbf{e}_s^\top \mathbf{W}_p \mathbf{e}_o}$ | $\mathbf{W}_p \in \mathbb{R}^{k \times k}$ |
| TransE [Bordes et al. 2013] | $-\left\Vert \mathbf{e}_s + \mathbf{r}_p - \mathbf{e}_o \right\Vert_p^2$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| DistMult [Yang et al. 2015] | $\boxed{\langle \mathbf{e}_s, \mathbf{r}_p, \mathbf{e}_o \rangle}$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| HolE [Nickel et al. 2016] | $\mathbf{r}_p^\top \left( \mathcal{F}^{-1} \left[ \overline{\mathcal{F}[\mathbf{e}_s]} \odot \mathcal{F}[\mathbf{e}_o] \right] \right)$ | $\mathbf{r}_p \in \mathbb{R}^k$ |
| **ComplEx [Trouillon et al. 2016]** | $\boxed{\mathrm{Re}\left( \langle \mathbf{e}_s, \mathbf{r}_p, \bar{\mathbf{e}}_o \rangle \right)}$ | $\mathbf{r}_p \in \mathbb{C}^k$ |
| **ConvE [Dettmers et al. 2017]** | $f\left( \mathrm{vec}\left( f\left( [\overline{\mathbf{e}_s}; \overline{\mathbf{r}_p}] * \omega \right) \right) \mathbf{W} \right) \mathbf{e}_o$ | $\mathbf{r}_p \in \mathbb{R}^k, \mathbf{W} \in \mathbb{R}^{c \times k}$ |

# Neural Link Prediction — Accuracy

**Evaluation Metrics** — Area Under the Precision-Recall Curve (AUC-PR), Mean Reciprocal Rank (MRR), Hits@k. In MRR and Hits@k, for each test triple:

- Modify its subject with all the entities in the Knowledge Graph,
- Score all the triple variants, and *compute the rank* of the original test triple,
- Repeat for the object.

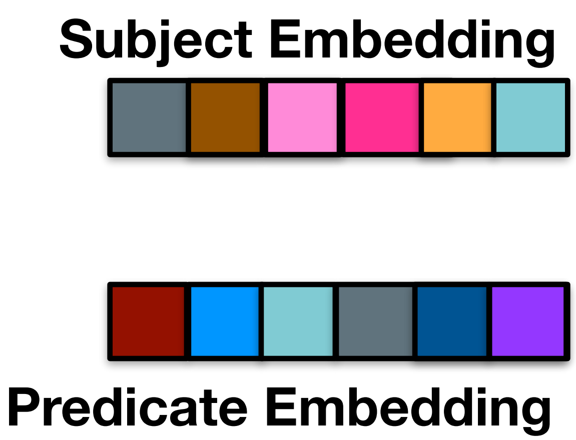
$$\text{MRR} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} \frac{1}{\text{rank}_i}, \quad \text{HITS}@k = \frac{|\{\text{rank}_i \leq 10\}|}{|\mathcal{T}|}$$

From [Lacroix et al. ICML 2018]

| | Model | WN18 | | WN18RR | | FB15K | | FB15K-237 | | YAGO3-10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR | H@10 | MRR | H@10 | MRR | H@10 | MRR | H@10 | MRR | H@10 |
| Reciprocal | CP-FRO | **0.95** | 0.95 | 0.46 | 0.48 | **0.86** | 0.91 | 0.34 | 0.51 | 0.54 | 0.68 |
| | CP-N3 | **0.95** | 0.96 | 0.47 | 0.54 | **0.86** | 0.91 | 0.36 | 0.54 | 0.57 | **0.71** |
| | ComplEx-FRO | **0.95** | 0.96 | 0.47 | 0.54 | **0.86** | 0.91 | 0.35 | 0.53 | 0.57 | **0.71** |
| | ComplEx-N3 | **0.95** | 0.96 | **0.48** | **0.57** | **0.86** | 0.91 | **0.37** | **0.56** | **0.58** | **0.71** |

# Convolutional 2D Knowledge Graph Embeddings

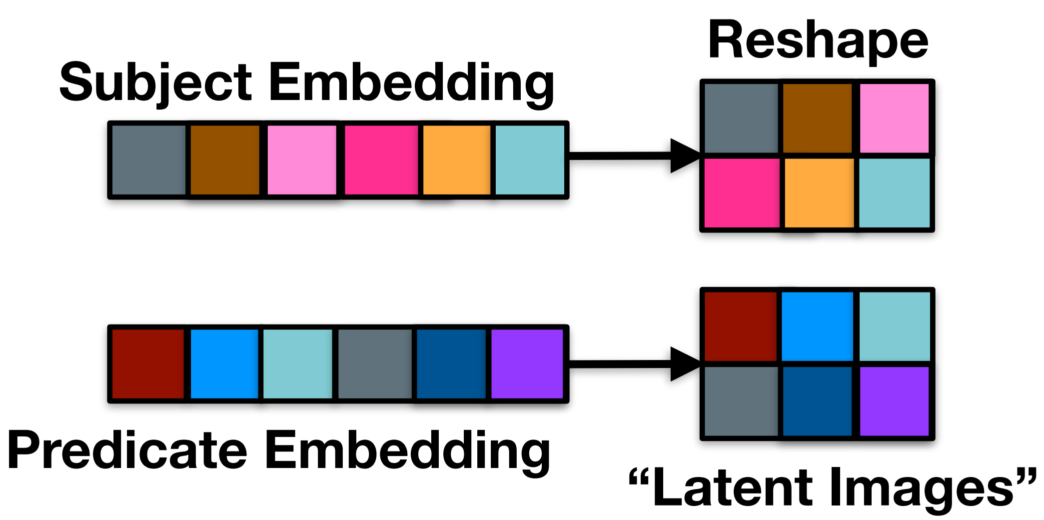
**Idea** — use ideas from *computer vision* for modeling the interactions between latent features.




[AAAI 2018]

# Convolutional 2D Knowledge Graph Embeddings

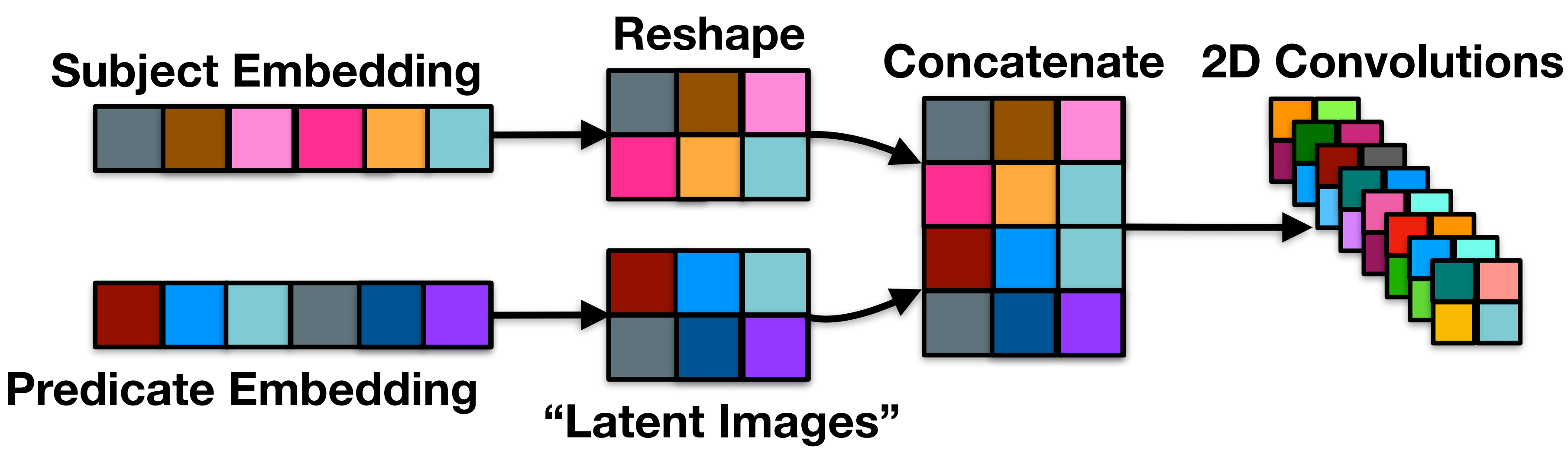
**Idea** — use ideas from *computer vision* for modeling the interactions between latent features.




[AAAI 2018]

# Convolutional 2D Knowledge Graph Embeddings

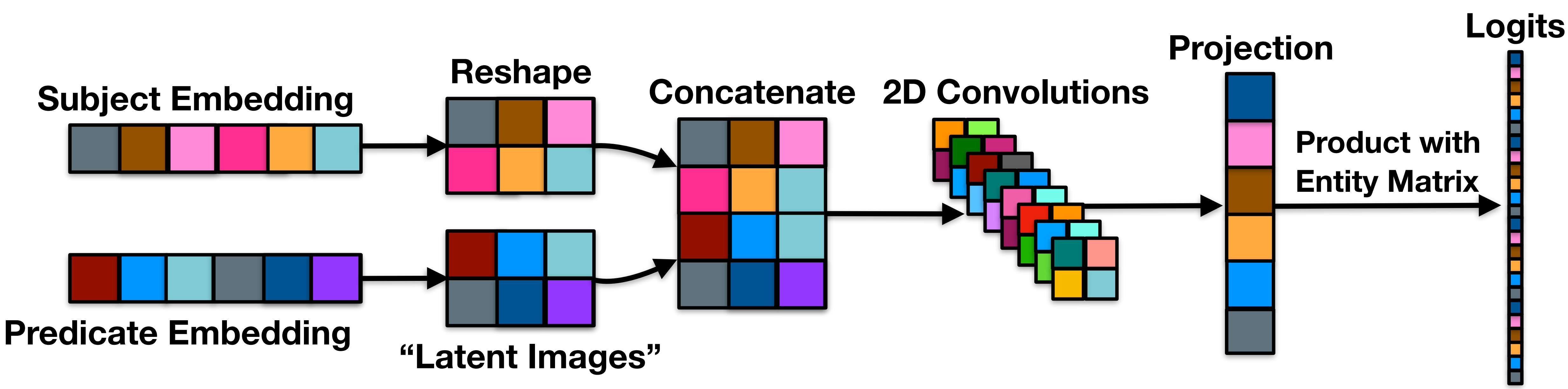
**Idea** — use ideas from *computer vision* for modeling the interactions between latent features.




[AAAI 2018]

# Convolutional 2D Knowledge Graph Embeddings

**Idea** — use ideas from *computer vision* for modeling the interactions between latent features.



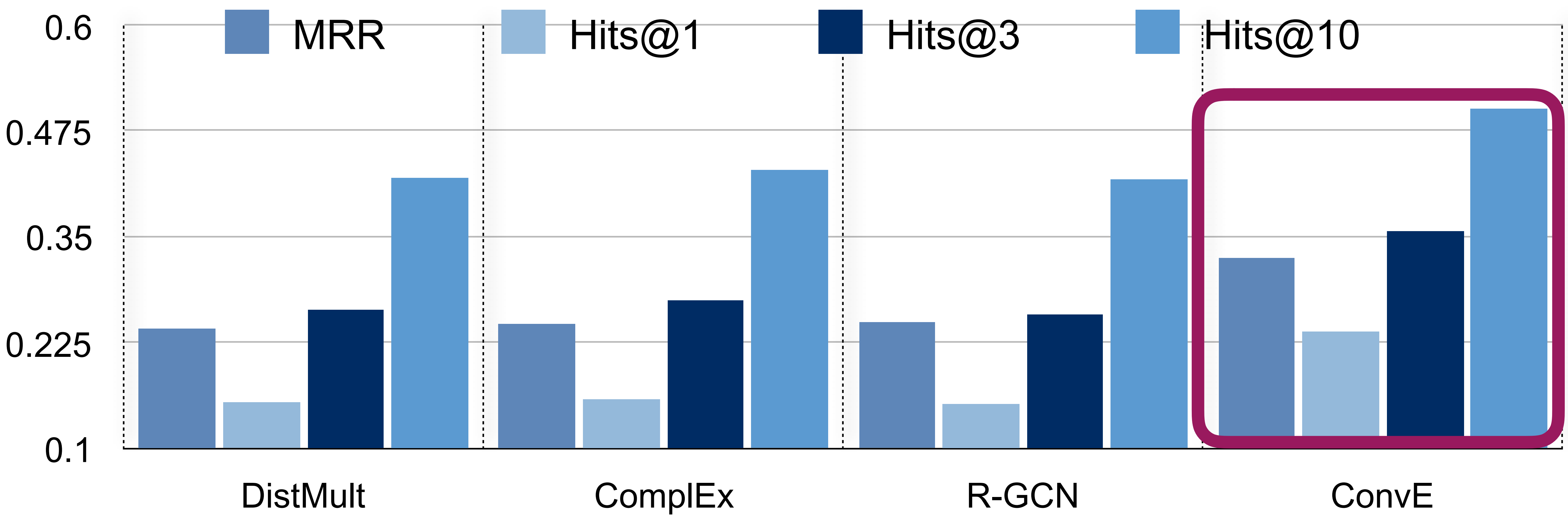

✓ Scalable

✓ State-of-the-art Results

[AAAI 2018]

# Convolutional 2D Knowledge Graph Embeddings

**Idea** — use ideas from *computer vision* for modeling the interactions between latent features.




✓ Efficiency via parameter sharing

✓ State-of-the-art Results

[AAAI 2018]

# Interpreting Knowledge Graph Embeddings

Quite hard to understand the *semantics* of the learned representations..

| | Real Part | | | | | Imaginary Part | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| hypernym | 1.0 | 3.0 | -3.1 | 2.5 | -2.7 | 3.2 | 2.9 | 1.7 | -3.0 | -3.0 |

[Minervini et al. ECML 2017]

# Interpreting Knowledge Graph Embeddings

Quite hard to understand the *semantics* of the learned representations..

| | Real Part | | | | | Imaginary Part | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| hypernym | 1.0 | 3.0 | -3.1 | 2.5 | -2.7 | 3.2 | 2.9 | 1.7 | -3.0 | -3.0 |
| hyponym | 1.0 | 3.1 | -3.1 | 2.6 | -2.7 | -3.4 | -2.8 | -1.7 | 2.9 | 3.0 |

.. but we can use their geometric relationships for identifying — and *incorporating* — semantic relationships between them.

[Minervini et al. ECML 2017]

# Regularising Knowledge Graph Embeddings

Quite hard to understand the *semantics* of the learned representations..

| | Real Part | | | | | Imaginary Part | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| hypernym | 1.0 | 3.0 | -3.1 | 2.5 | -2.7 | 3.2 | 2.9 | 1.7 | -3.0 | -3.0 |
| hyponym | 1.0 | 3.1 | -3.1 | 2.6 | -2.7 | -3.4 | -2.8 | -1.7 | 2.9 | 3.0 |
| instance hypernym | -1.1 | -2.8 | 1.6 | 2.7 | -2.5 | 3.0 | -2.6 | 2.6 | -1.1 | -2.8 |
| instance hyponym | -1.0 | -2.9 | 1.5 | 2.9 | -2.4 | -2.9 | 2.8 | -2.6 | 1.1 | 2.8 |
| part of | -2.4 | 3.2 | 2.7 | -1.5 | 3.0 | -2.4 | -0.6 | -2.6 | 2.9 | -1.9 |
| has part | -2.5 | 3.2 | 2.9 | -1.5 | 3.0 | 2.4 | 0.7 | 2.8 | -3.0 | 1.9 |

.. but we can use their geometric relationships for identifying — and *incorporating* — semantic relationships between them.

[Minervini et al. ECML 2017]

# Regularising Knowledge Graph Embeddings

Quite hard to understand the *semantics* of the learned representations..

| | Real Part | | | | | Imaginary Part | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| hypernym | 1.0 | 3.0 | -3.1 | 2.5 | -2.7 | 3.2 | 2.9 | 1.7 | -3.0 | -3.0 |
| hyponym | 1.0 | 3.1 | -3.1 | 2.6 | -2.7 | -3.4 | -2.8 | -1.7 | 2.9 | 3.0 |
| instance hypernym | -1.1 | -2.8 | 1.6 | 2.7 | -2.5 | 3.0 | -2.6 | 2.6 | -1.1 | -2.8 |
| instance hyponym | -1.0 | -2.9 | 1.5 | 2.9 | -2.4 | -2.9 | 2.8 | -2.6 | 1.1 | 2.8 |
| part of | -2.4 | 3.2 | 2.7 | -1.5 | 3.0 | -2.4 | -0.6 | -2.6 | 2.9 | -1.9 |
| has part | -2.5 | 3.2 | 2.9 | -1.5 | 3.0 | 2.4 | 0.7 | 2.8 | -3.0 | 1.9 |

.. but we can use their geometric relationships for identifying — and *incorporating* — semantic relationships between them.

✗ $\text{is\_a}(x, y) \wedge \text{is\_a}(y, z) \Rightarrow \text{is\_a}(x, z)$

**[Minervini et al. ECML 2017]**

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

- An *adversary* searches for inputs where the model violates constraints

**e.g.** $x, y, z$ **such that**

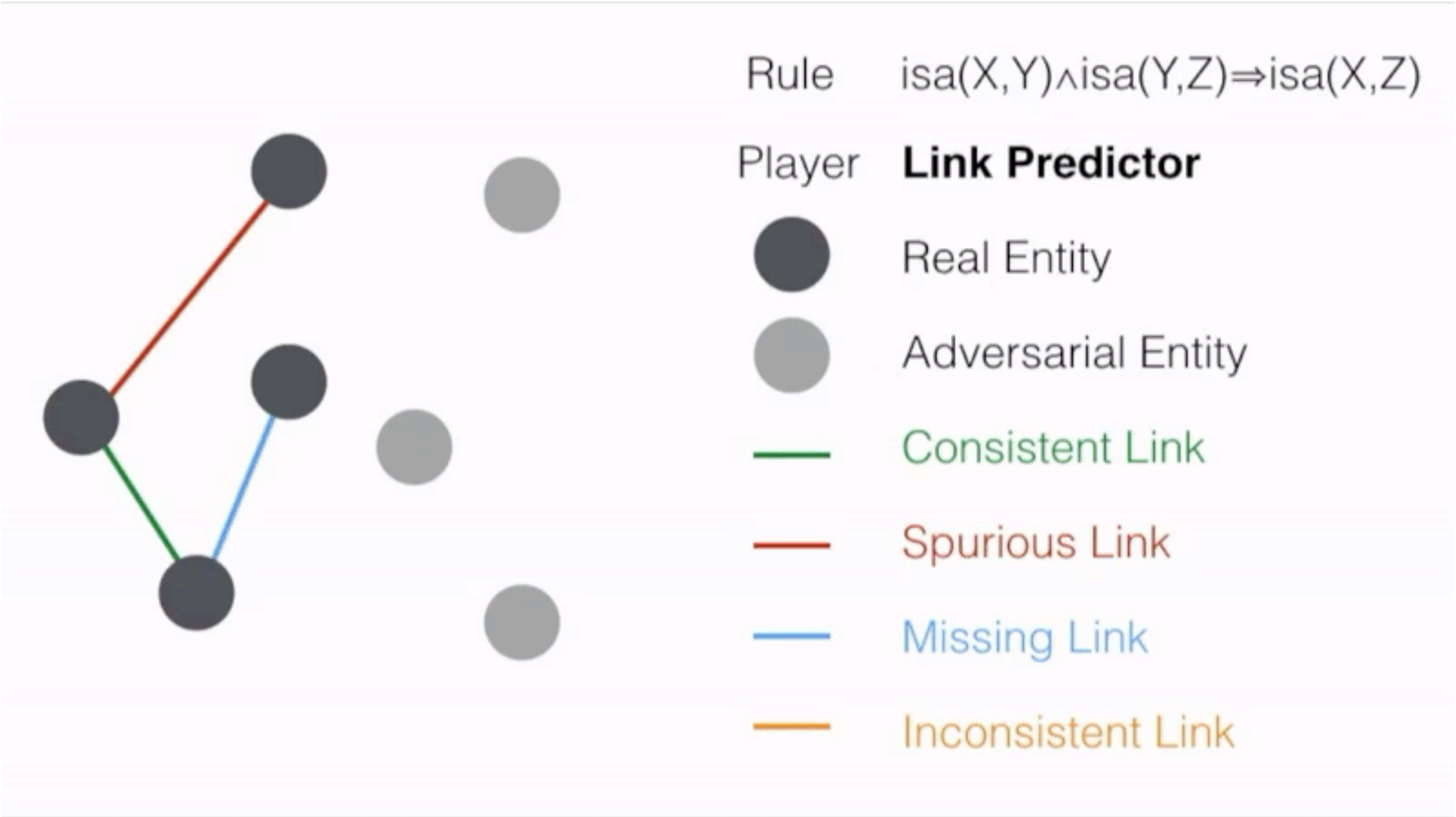
$$\text{is\_a}(x, y) \wedge \text{is\_a}(y, z) \wedge \neg\text{is\_a}(x, z)$$

**[Minervini et al. UAI 2017]**

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

- An *adversary* searches for inputs where the model violates constraints
- The model is *regularised* to correct such violations.




[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

- An *adversary* searches for inputs where the model violates constraints,
- The model is *regularised* to correct such violations.

**e.g.** $S = \{x, y, z\}$ **such that**

$\text{is\_a}(x, y) \wedge \text{is\_a}(y, z) \wedge \neg\text{is\_a}(x, z)$

Formally:

$$\min_{\Theta} \mathscr{L}_{\text{data}}(D \mid \Theta) + \lambda \max_{S} \mathscr{L}_{\text{violation}}(S, D \mid \Theta)$$

[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

- An *adversary* searches for *inputs* where the model violates constraints,
- The model is *regularised* to correct such violations.

**e.g.** $S = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ **such that**

$$\text{is a}(\mathbf{e}_x, \mathbf{e}_y) \wedge \text{is a}(\mathbf{e}_y, \mathbf{e}_z) \wedge \neg\text{is a}(\mathbf{e}_x, \mathbf{e}_z)$$

Formally:

$$\min_{\Theta} \mathscr{L}_{\text{data}}(D \mid \Theta) + \boxed{\lambda \max_{S} \mathscr{L}_{\text{violation}}(S, D \mid \Theta)}$$

- Inputs $S$ can be either *input space* or *embedding space*
- In most interesting cases, *max* has <u>*closed form solutions*</u>
- <u>Constraints are *guaranteed to hold everywhere*</u> in embedding space.

[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training

**Idea** — *adversarial training* process where, iteratively:

- An *adversary* searches for *inputs* where the model violates constraints,
- The model is *regularised* to correct such violations.

**e.g.** $S = \{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ **such that**

$\text{is a}(\mathbf{e}_x, \mathbf{e}_y) \wedge \text{is a}(\mathbf{e}_y, \mathbf{e}_z) \wedge \neg\text{is a}(\mathbf{e}_x, \mathbf{e}_z)$

Formally:

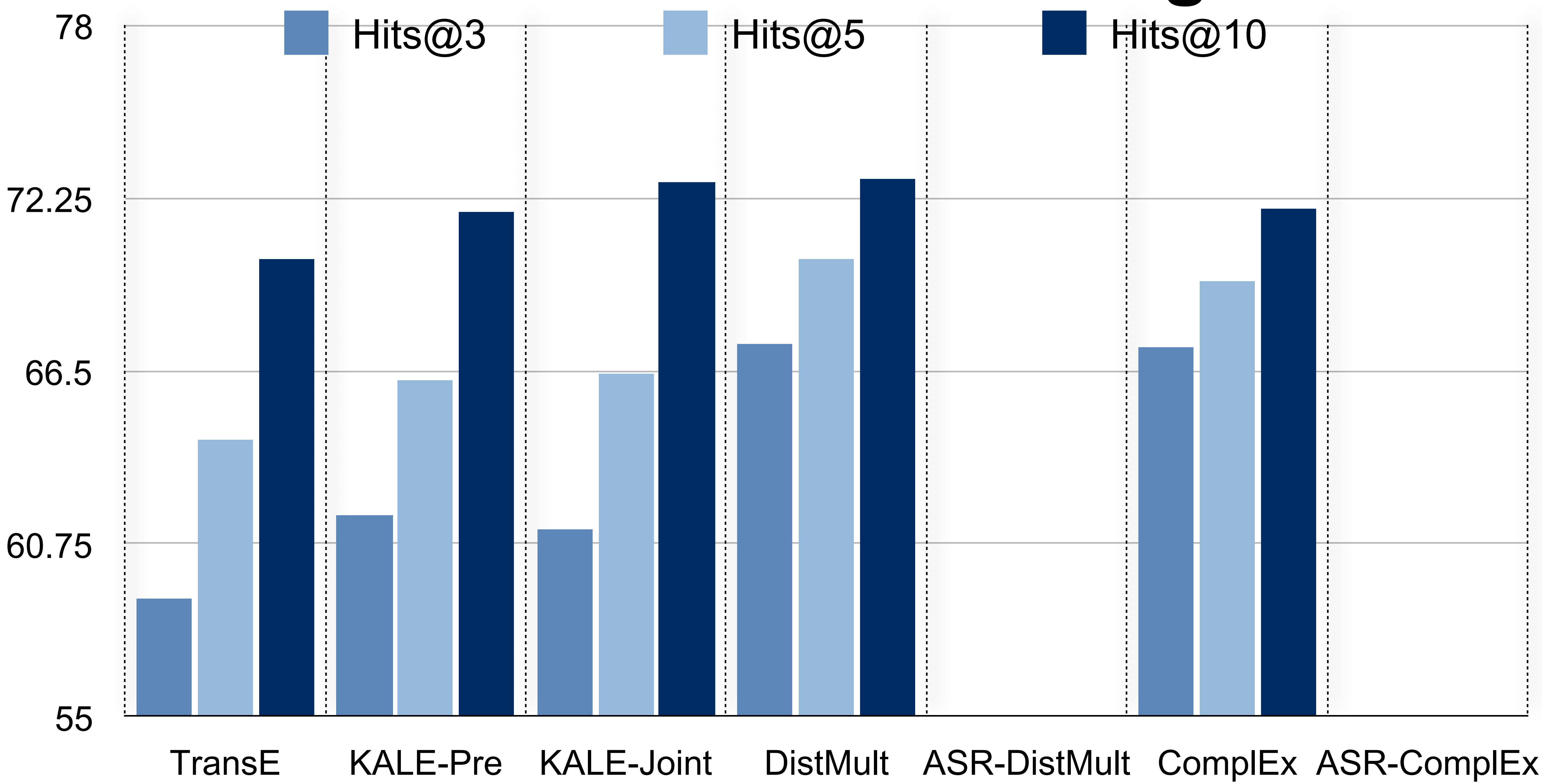
$$\min_{\Theta} \mathscr{L}_{\text{data}}(D \mid \Theta) + \boxed{\lambda \max_{S} \mathscr{L}_{\text{violation}}(S, D \mid \Theta)}$$

✓ Incorporates Background Knowledge

✓ Verifiable

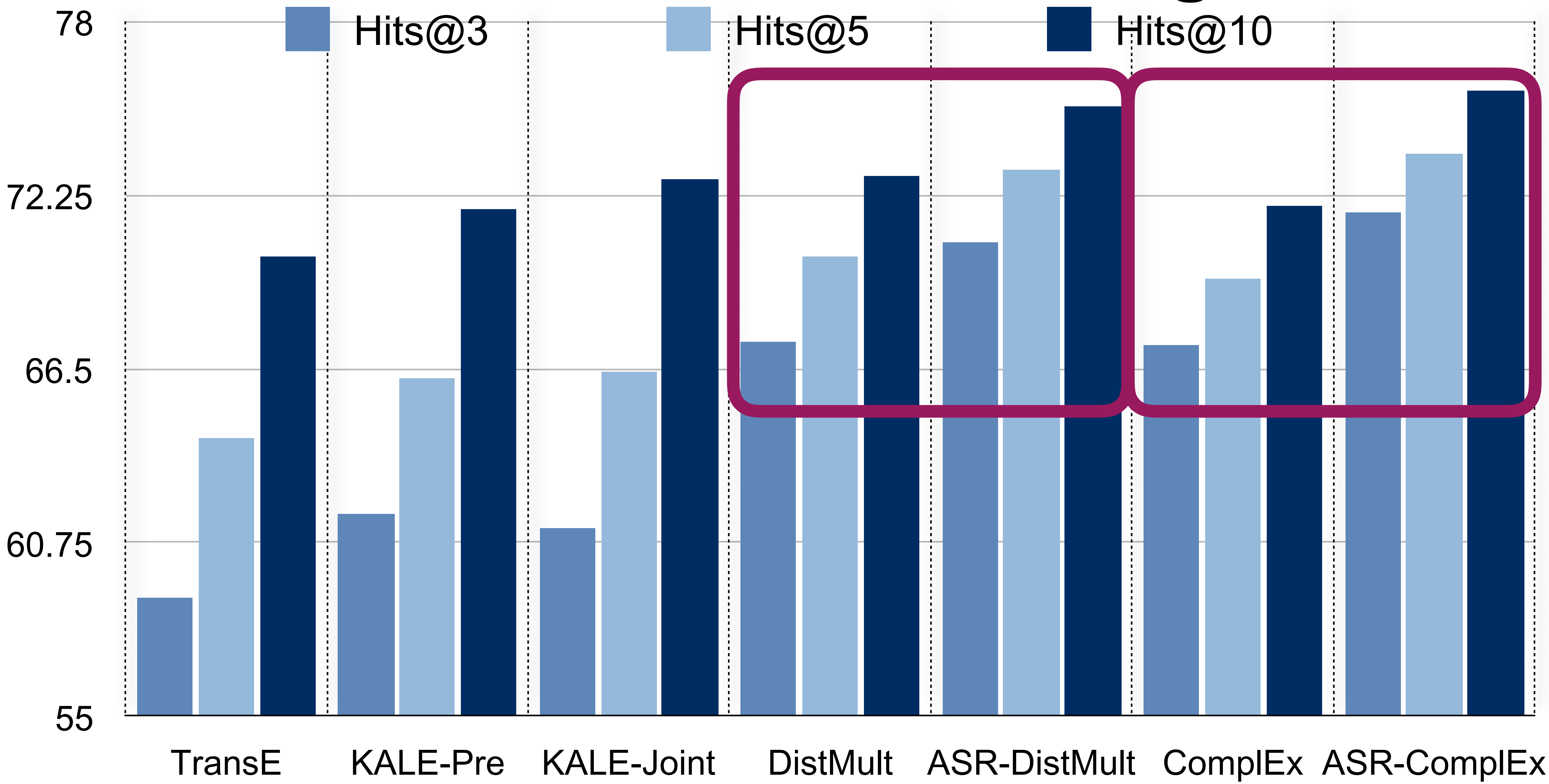
[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training




[Minervini et al. UAI 2017]

# Incorporating Background Knowledge via Adversarial Training




[Minervini et al. UAI 2017]

# Incorporating Background Knowledge in Natural Language Inference Models

**Natural Language Inference** — detect the type of relationship, *i.e. entailment*, *contradiction, neutral,* between two sentences.

[Minervini et al. CoNLL 2018]

# Incorporating Background Knowledge in Natural Language Inference Models

**Natural Language Inference** — detect the type of relationship, *i.e. entailment*, *contradiction, neutral,* between two sentences.

If a stentence **x** contradicts **y**, then also **y** contradicts **x**.
If **x** entails **y**, and **y** entails **z**, then **x** also entails **z**.

[Minervini et al. CoNLL 2018]

# Incorporating Background Knowledge in Natural Language Inference Models

**Natural Language Inference** — detect the type of relationship, *i.e. entailment*, *contradiction, neutral,* between two sentences.

If a stentence **x** contradicts **y**, then also **y** contradicts **x**.
If **x** entails **y**, and **y** entails **z**, then **x** also entails **z**.

**x)** A man in uniform is pushing a medical bed.
**y)** A man is ~~pushing~~ **carrying** something.

[Minervini et al. CoNLL 2018]

# Incorporating Background Knowledge in Natural Language Inference Models

**Natural Language Inference** — detect the type of relationship, *i.e. entailment*, *contradiction, neutral,* between two sentences.

If a stentence $\mathbf{x}$ contradicts $\mathbf{y}$, then also $\mathbf{y}$ contradicts $\mathbf{x}$.
If $\mathbf{x}$ entails $\mathbf{y}$, and $\mathbf{y}$ entails $\mathbf{z}$, then $\mathbf{x}$ also entails $\mathbf{z}$.

**x)** A man in uniform is pushing a medical bed.
**y)** A man is ~~**pushing**~~ **carrying** something.

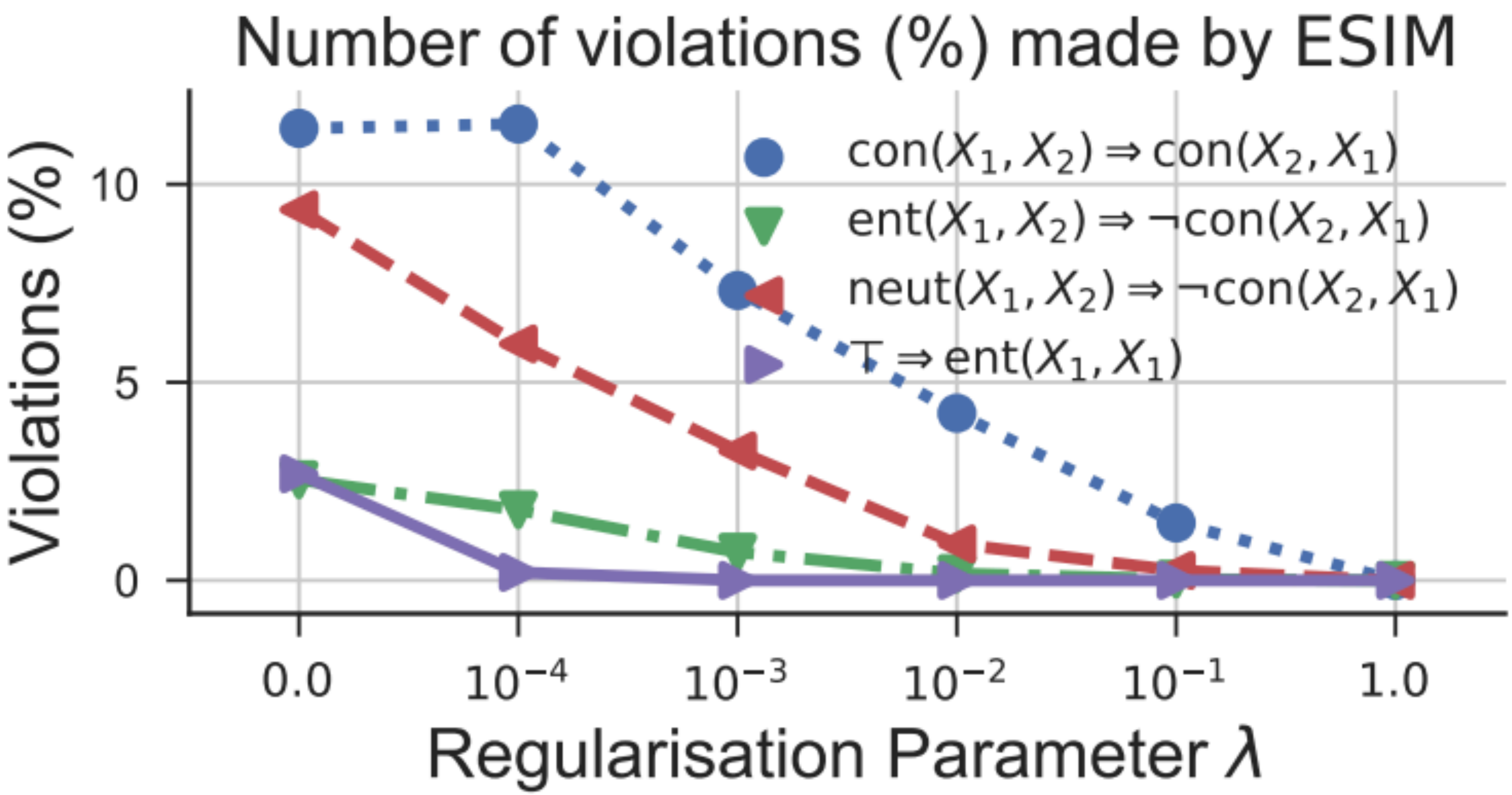
$$P(\mathbf{x} \xrightarrow{\text{entails}} \mathbf{y}) = 0.72$$

$$P(\mathbf{y} \xrightarrow{\text{contradicts}} \mathbf{x}) = 0.93$$

$$\mathcal{L}_{\text{violation}}(\{\mathbf{x}, \mathbf{y}\}) : 0.01 \rightsquigarrow 0.92$$

[Minervini et al. CoNLL 2018]

# Incorporating Background Knowledge in Natural Language Inference Models




[Minervini et al. CoNLL 2018]

# End-to-End Differentiable Reasoning

# End-to-End Differentiable Reasoning

Core idea — we can combine *neural networks* and *symbolic models* by re-implementing classic reasoning algorithms using end-to-end differentiable (neural) architectures.

**(Black-Box) Neural Models** ⟷ **Symbolic Reasoning Models**

- Can generalise from noisy and ambiguois modalities
- Can learn representations from data
- SOTA on a number of tasks

- Data efficient
- Interpretable
- Explainable
- Verifiable
- Can incorporate background knowledge and constraints

# Reasoning via Backward Chaining

**Backward Chaining —** start with a list of *goals*, and work backwards from the *consequent* Q to the *antecedent* P to see if any data supports any of the consequents.

$$q(X) \leftarrow p(X)$$

$$p(a) \qquad q(a)?$$

$$p(b)$$

$$p(c)$$

$$\cdots$$

You can see backward chaining as a *query reformulation strategy.*

# Reasoning via Backward Chaining

**Backward Chaining —** start with a list of *goals*, and work backwards from the *consequent Q* to the *antecedent P* to see if any data supports any of the consequents.

$$q(X) \leftarrow p(X)$$

$$p(a) \qquad\qquad q(a)?$$

$$p(b)$$

$$p(c) \qquad\qquad p(a)$$

$$\ldots$$

You can see backward chaining as a *query reformulation strategy.*

# Reasoning via Backward Chaining

**Backward Chaining —** start with a list of *goals*, and work backwards from the *consequent Q* to the *antecedent P* to see if any data supports any of the consequents.

$$q(X) \leftarrow p(X)$$

$$p(a) \qquad\qquad q(a)?$$

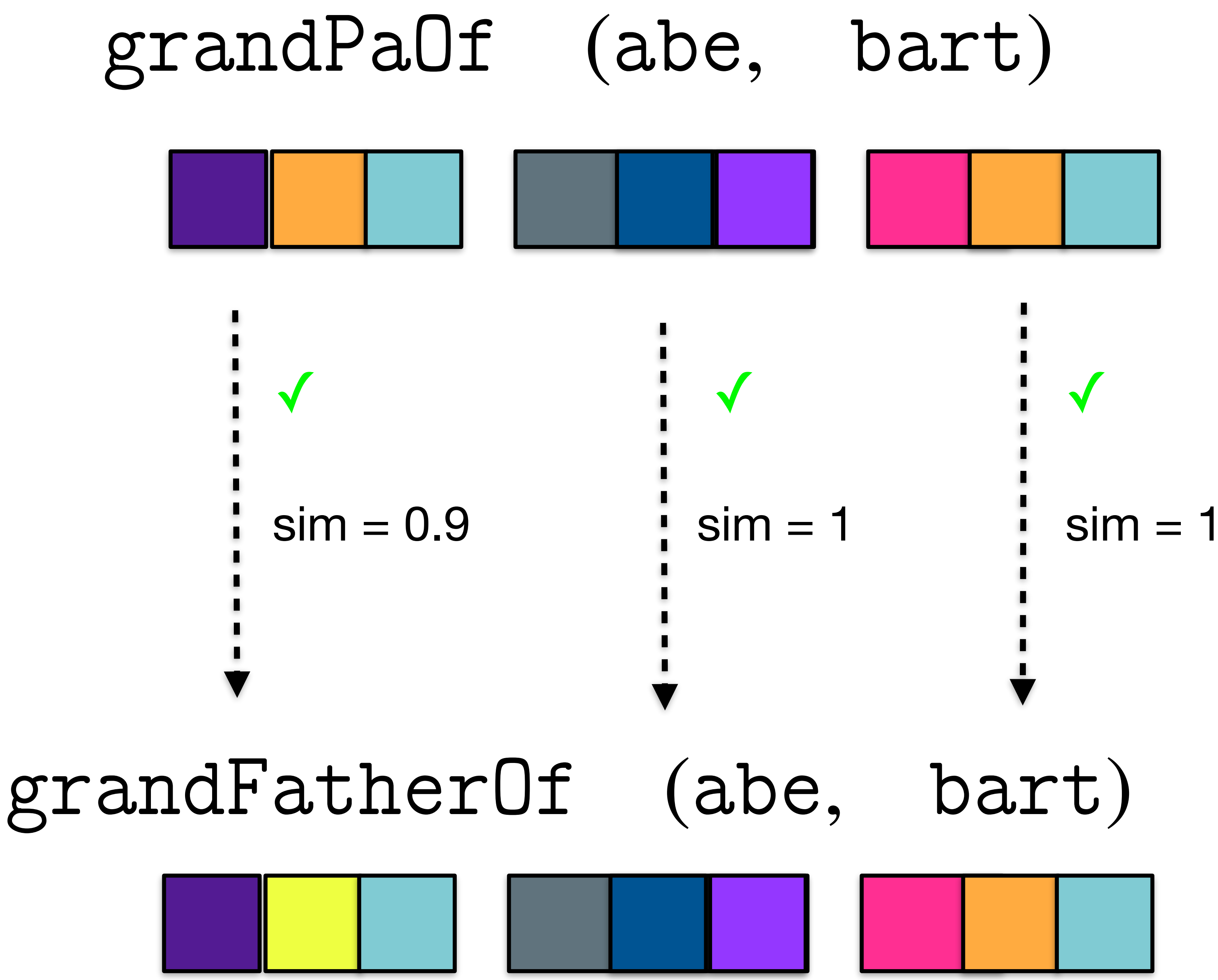
$$p(b)$$

$$p(c) \qquad\qquad p(a)$$

$$\ldots$$

You can see backward chaining as a *query reformulation strategy.*

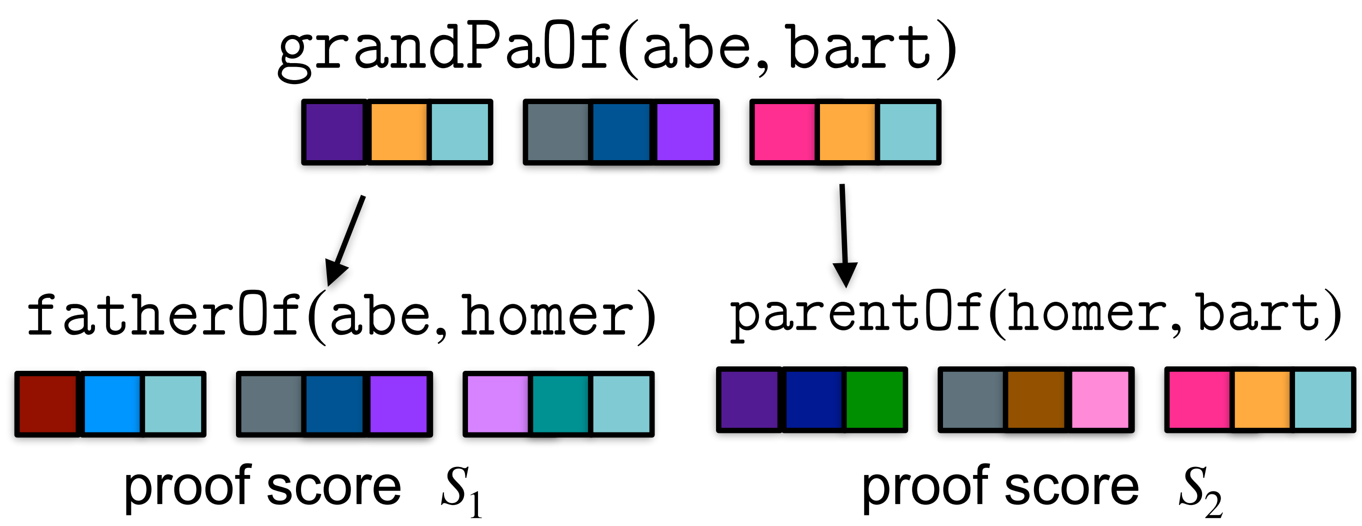
# End-to-End Differentiable Reasoning



# End-to-End Differentiable Reasoning

**Knowledge Base:**

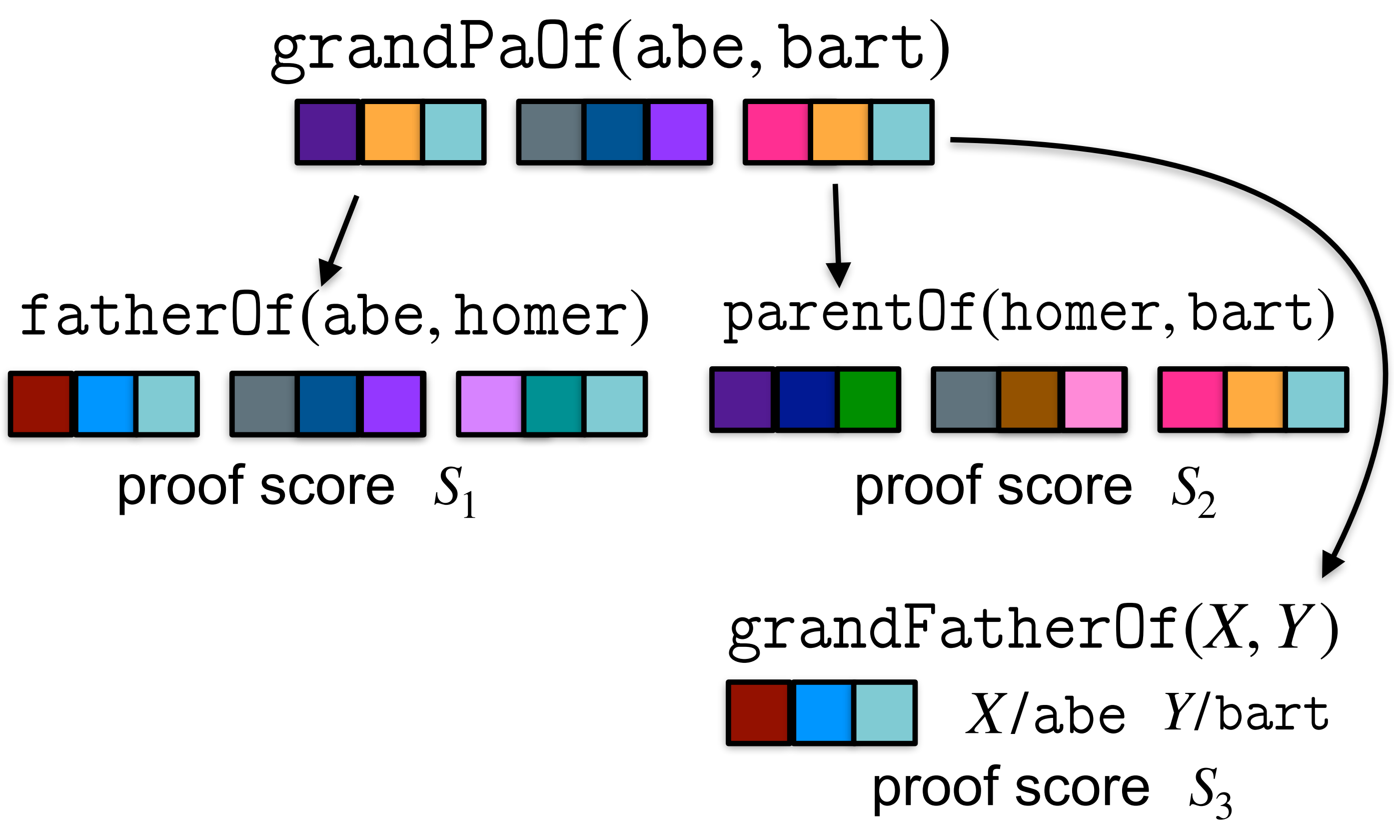
$$\mathtt{fatherOf}(\mathtt{abe}, \mathtt{homer})$$
$$\mathtt{parentOf}(\mathtt{homer}, \mathtt{bart})$$
$$\mathtt{grandFatherOf}(X, Y) \Leftarrow \mathtt{fatherOf}(X, Z), \mathtt{parentOf}(Z, Y) .$$



# End-to-End Differentiable Reasoning

**Knowledge Base:**

$$
\begin{aligned}
&\mathtt{fatherOf}(\mathtt{abe}, \mathtt{homer})\\
&\mathtt{parentOf}(\mathtt{homer}, \mathtt{bart})\\
&\mathtt{grandFatherOf}(X, Y) \Leftarrow\\
&\quad \mathtt{fatherOf}(X, Z),\\
&\quad \mathtt{parentOf}(Z, Y).
\end{aligned}
$$




**Subgoals:**

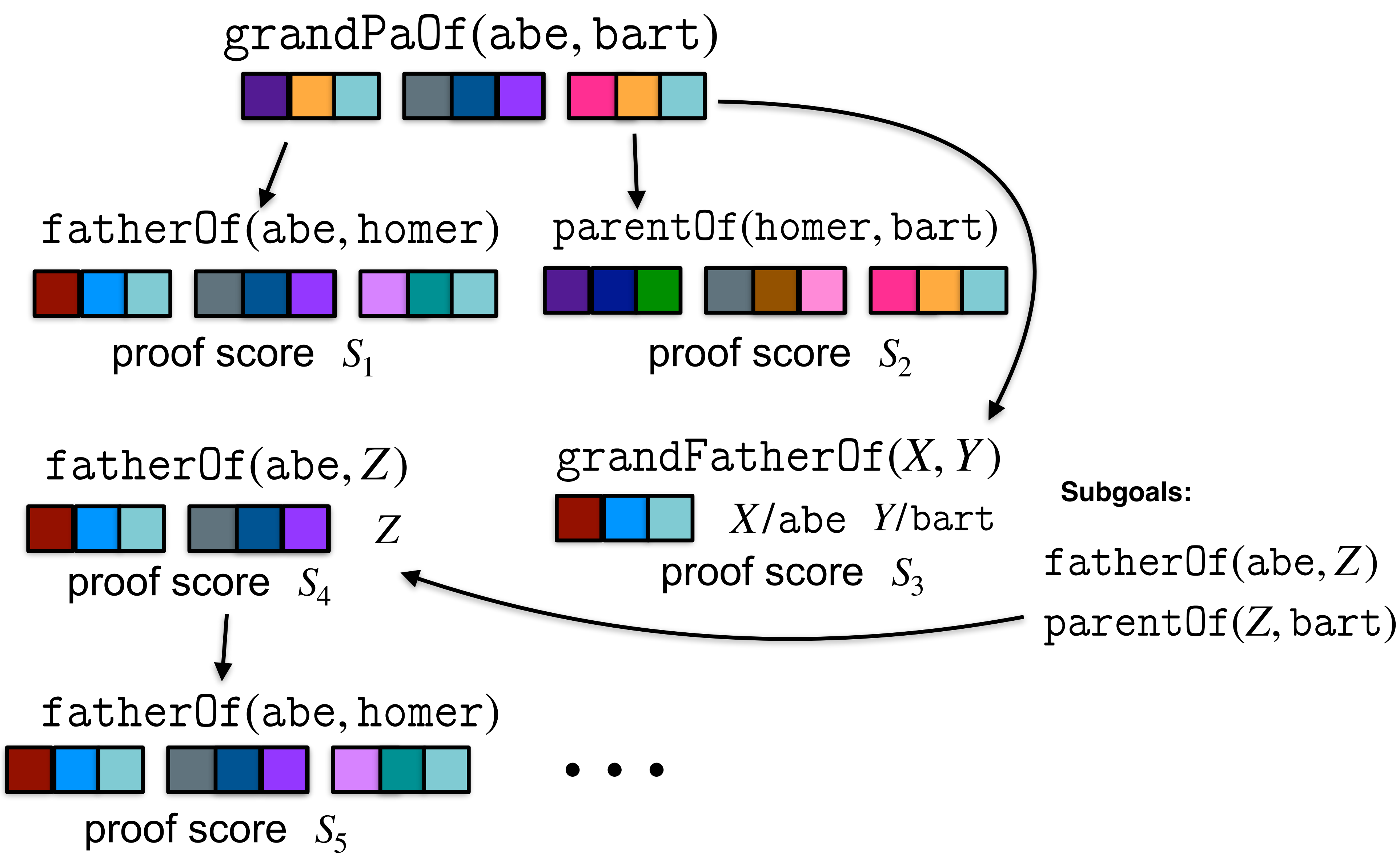
$\mathtt{fatherOf}(\mathtt{abe}, Z)$

$\mathtt{parentOf}(Z, \mathtt{bart})$

# End-to-End Differentiable Reasoning

**Knowledge Base:**

$$\texttt{fatherOf}(\texttt{abe}, \texttt{homer})$$
$$\texttt{parentOf}(\texttt{homer}, \texttt{bart})$$
$$\texttt{grandFatherOf}(X, Y) \Leftarrow \texttt{fatherOf}(X, Z), \texttt{parentOf}(Z, Y).$$



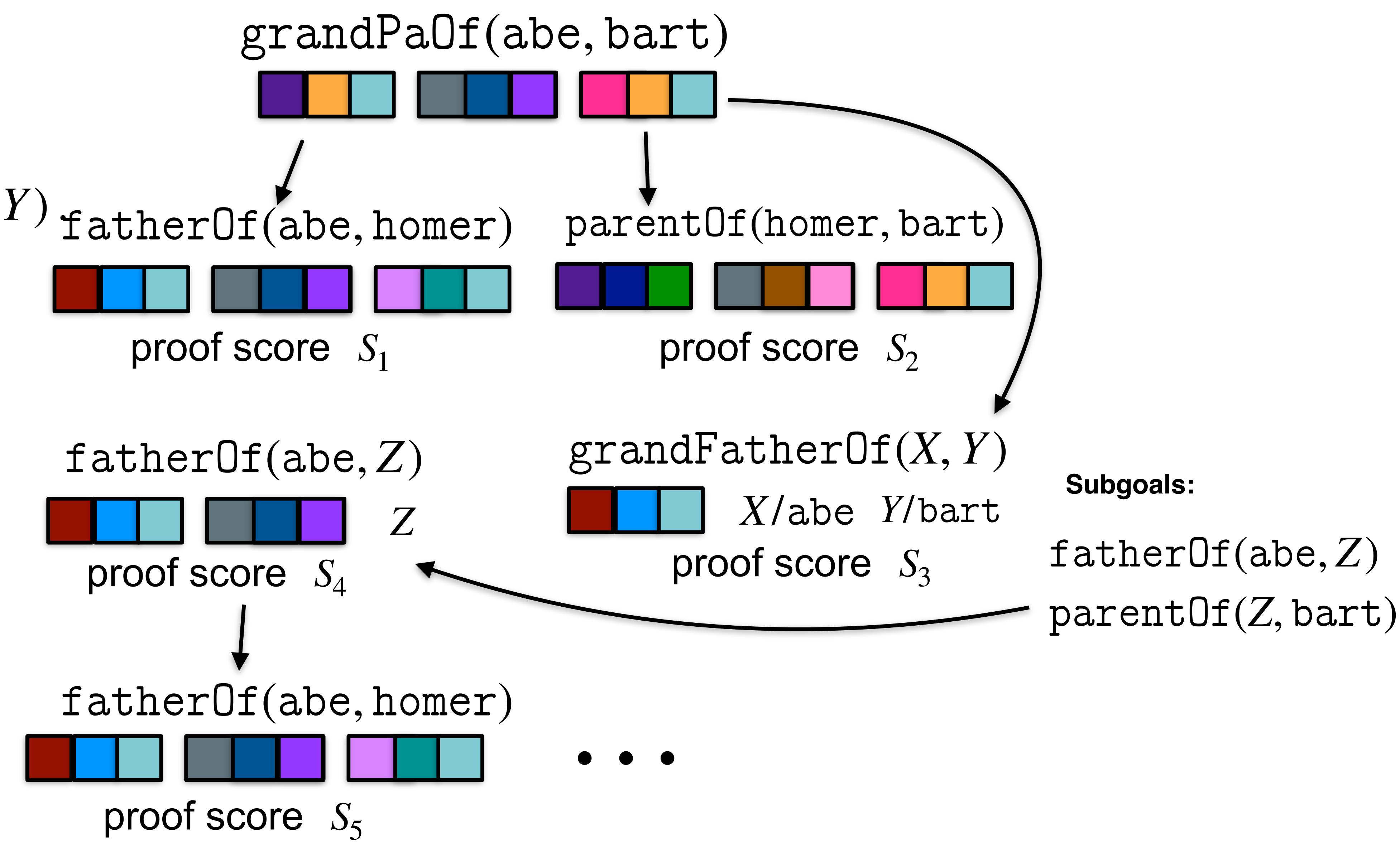
# End-to-End Differentiable Reasoning

**Knowledge Base:**

$\mathtt{fatherOf}(\mathtt{abe}, \mathtt{homer})$

$\mathtt{parentOf}(\mathtt{homer}, \mathtt{bart})$

$\theta_1(X, Y) \Leftarrow \theta_2(X, Z), \theta_3(Z, Y)$



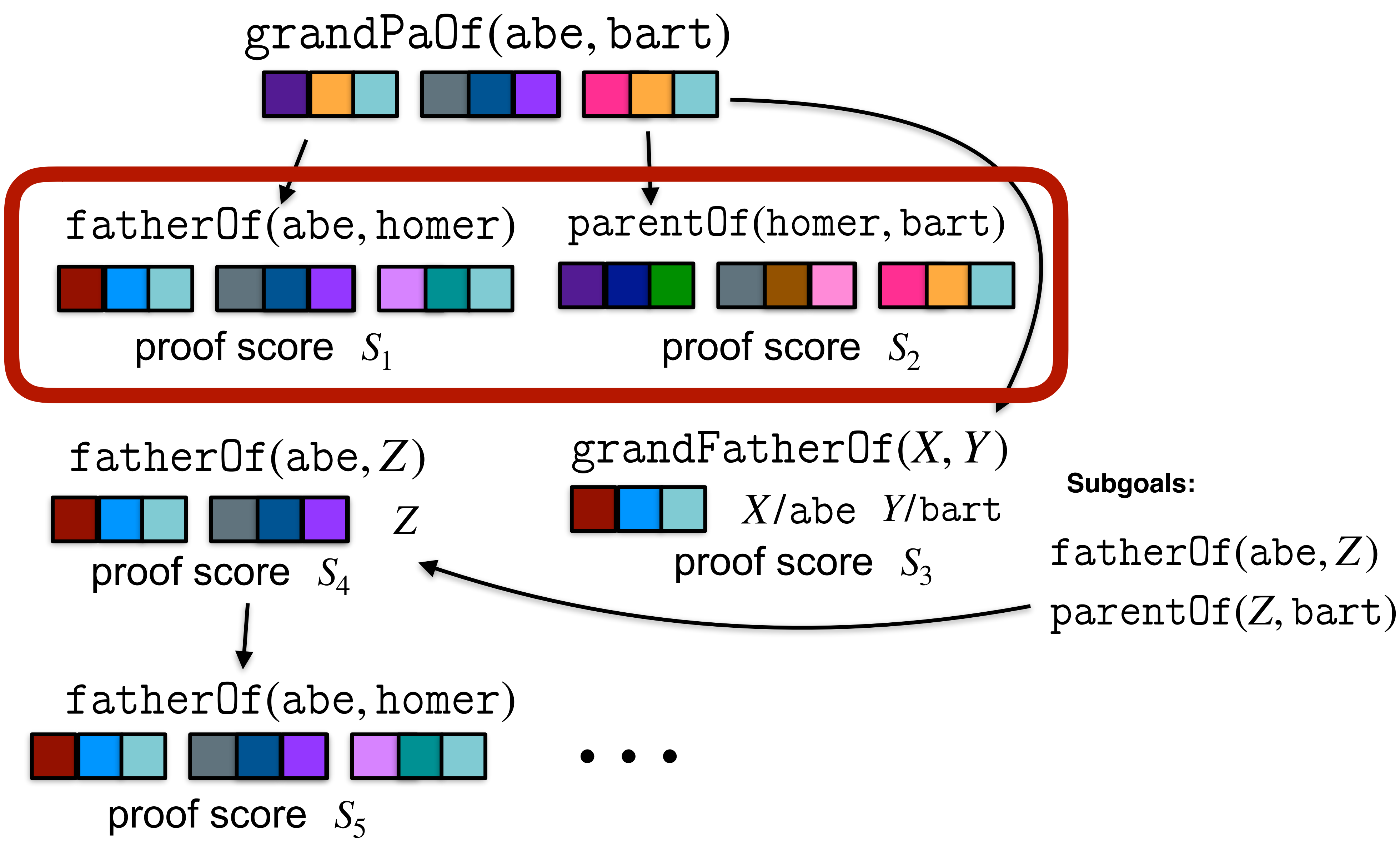

**Train via Self-Supervision:**

$$\sum_{F \in K} \log p^{KB \setminus F}(F) - \sum_{\tilde{F} \sim corr(F)} \log p^{KB}(\tilde{F})$$

# End-to-End Differentiable Reasoning

**Knowledge Base:**

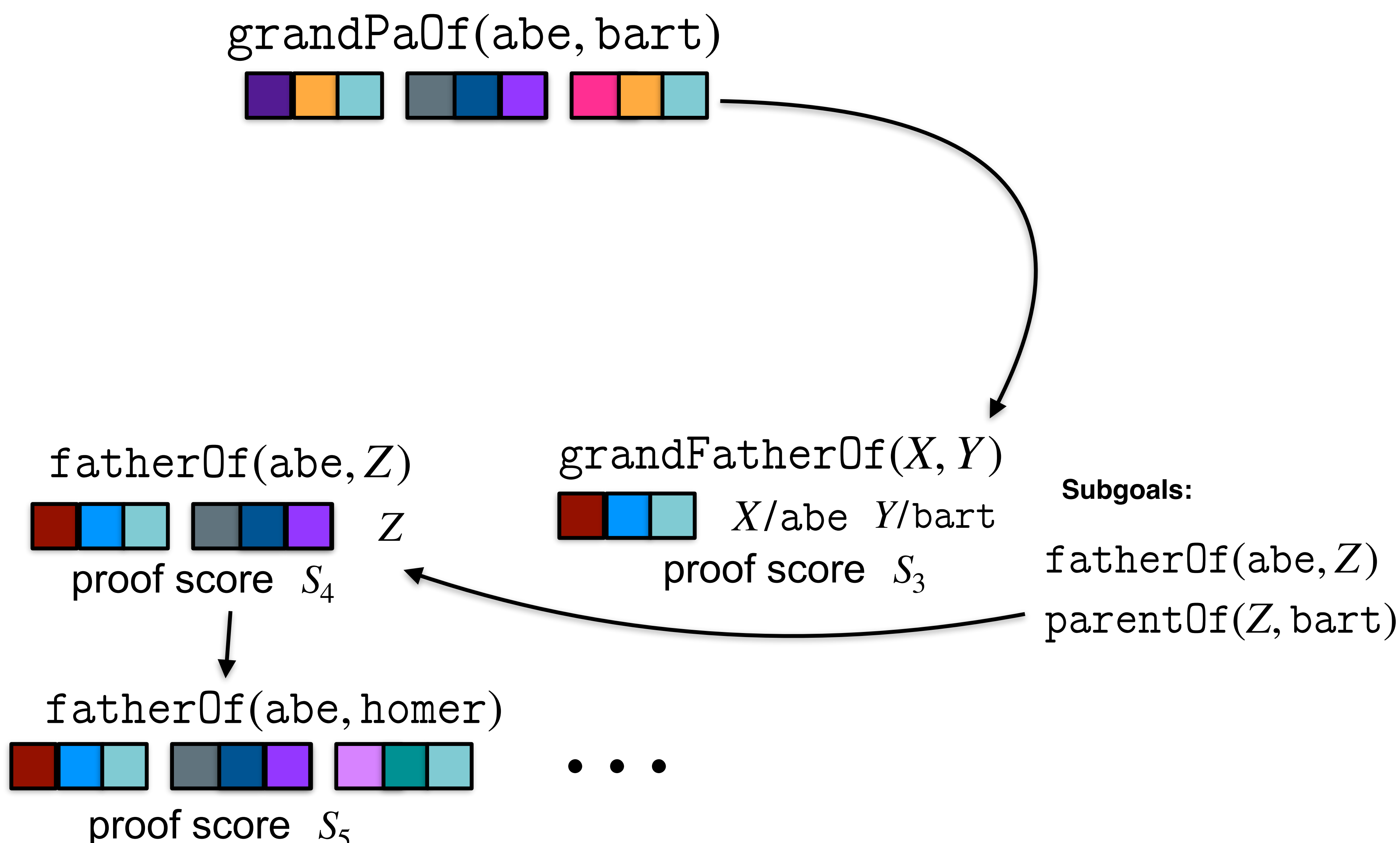
$$\texttt{fatherOf}(\texttt{abe}, \texttt{homer})$$
$$\texttt{parentOf}(\texttt{homer}, \texttt{bart})$$
$$\texttt{grandFatherOf}(X, Y) \Leftarrow \texttt{fatherOf}(X, Z), \texttt{parentOf}(Z, Y).$$



# End-to-End Differentiable Reasoning

**Knowledge Base:**

$$\texttt{fatherOf}(\texttt{abe}, \texttt{homer})$$
$$\texttt{parentOf}(\texttt{homer}, \texttt{bart})$$
$$\texttt{grandFatherOf}(X, Y) \Leftarrow \texttt{fatherOf}(X, Z), \texttt{parentOf}(Z, Y).$$




[Minervini et al. 2018, AAAI 2020,
Welbl et al. ACL 2019]

# End-to-End Differentiable Reasoning

| | Query | Score $S_\rho$ | Proofs / Explanations |
|---|---|---|---|
| WN18 | part_of(CONGO.N.03, AFRICA.N.01) | 0.995 | part_of(X, Y) :– has_part(Y, X)<br>has_part(AFRICA.N.01, CONGO.N.03) |
| | | 0.787 | part_of(X, Y) :– instance_hyponym(Y, X)<br>instance_hyponym(AFRICAN_COUNTRY.N.01, CONGO.N.03) |
| | hyponym(EXTINGUISH.V.04, DECOUPLE.V.03) | 0.987 | hyponym(X, Y) :– hypernym(Y, X)<br>hypernym(DECOUPLE.V.03, EXTINGUISH.V.04) |
| | | 0.920 | hypernym(SNUFF_OUT.V.01, EXTINGUISH.V.04) |
| | part_of(PITUITARY.N.01, DIENCEPHALON.N.01) | 0.995 | has_part(DIENCEPHALON.N.01, PITUITARY.N.01) |
| | has_part(TEXAS.N.01, ODESSA.N.02) | 0.961 | has_part(X, Y) :– part_of(Y, X)<br>part_of(ODESSA.N.02, TEXAS.N.01) |
| | hyponym(SKELETAL_MUSCLE, ARTICULAR_MUSCLE) | 0.987 | hypernym(ARTICULAR_MUSCLE, SKELETAL_MUSCLE) |
| | deriv_related_form(REWRITE, REWRITING) | 0.809 | deriv_related_form(X, Y) :– hypernym(Y, X)<br>hypernym(REVISE, REWRITE) |
| WN18RR | also_see(TRUE.A.01, FAITHFUL.A.01) | 0.962 | also_see(X, Y) :– also_see(Y, X)<br>also_see(FAITHFUL.A.01, TRUE.A.01) |
| | | 0.590 | also_see(CONSTANT.A.02, FAITHFUL.A.01) |
| | also_see(GOOD.A.03, VIRTUOUS.A.01) | 0.962 | also_see(VIRTUOUS.A.01, GOOD.A.03) |
| | | 0.702 | also_see(RIGHTEOUS.A.01, VIRTUOUS.A.01) |
| | instance_hypernym(CHAPLIN, FILM_MAKER) | 0.812 | instance_hypernym(CHAPLIN, COMEDIAN) |

[Minervini et al. AAAI 2020]

# End-to-End Differentiable Reasoning

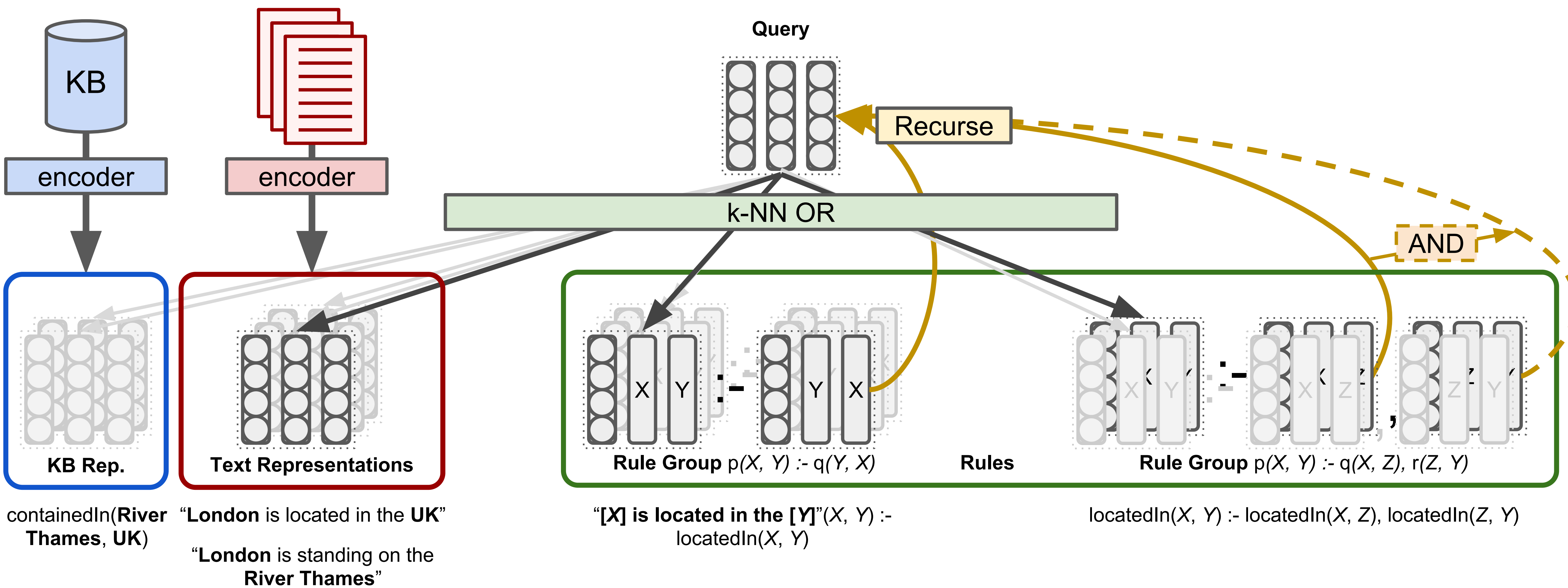
| | | Test-I | | | | Test-II | | | | Test-ALL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Hits@N (%) | | | MRR | Hits@N (%) | | | MRR | Hits@N (%) | | | MRR |
| | | 3 | 5 | 10 | | 3 | 5 | 10 | | 3 | 5 | 10 | |
| With Rules | KALE-Pre (Guo et al. 2016) | 35.8 | 41.9 | 49.8 | 0.291 | 82.9 | 86.1 | 89.9 | 0.713 | 61.7 | 66.2 | 71.8 | 0.523 |
| | KALE-Joint (Guo et al. 2016) | **38.4** | **44.7** | **52.2** | 0.325 | 79.7 | 84.1 | 89.6 | 0.684 | 61.2 | 66.4 | 72.8 | 0.523 |
| | *ASR*-DistMult (Minervini et al. 2017) | 36.3 | 40.3 | 44.9 | 0.330 | 98.0 | 99.0 | 99.2 | 0.948 | 70.7 | 73.1 | 75.2 | 0.675 |
| | *ASR*-ComplEx (Minervini et al. 2017) | 37.3 | 41.0 | 45.9 | **0.338** | **99.2** | **99.3** | **99.4** | **0.984** | 71.7 | 73.6 | 75.7 | 0.698 |
| | KBLR (García-Durán and Niepert 2018) | – | – | – | – | – | – | – | – | **74.0** | **77.0** | **79.7** | **0.702** |
| Without Rules | TransE (Bordes et al. 2013) | 36.0 | 41.5 | 48.1 | 0.296 | 77.5 | 82.8 | 88.4 | 0.630 | 58.9 | 64.2 | 70.2 | 0.480 |
| | DistMult (Yang et al. 2015) | 36.0 | 40.3 | 45.3 | 0.313 | **92.3** | **93.8** | **94.7** | 0.874 | **67.4** | **70.2** | **72.9** | 0.628 |
| | ComplEx (Trouillon et al. 2016) | **37.0** | **41.3** | **46.2** | **0.329** | 91.4 | 91.9 | 92.4 | **0.887** | 67.3 | 69.5 | 71.9 | **0.641** |
| | GNTPs | 33.7 | 36.9 | 41.2 | 0.313 | **98.2** | **99.0** | **99.3** | **0.977** | **69.2** | **71.1** | **73.2** | **0.678** |

| Datasets | | Metrics | Models | | | | | Rules Learned by GNTP |
|---|---|---|---|---|---|---|---|---|
| | | | NTP [3] | GNTP | | NeuralLP | MINERVA | |
| | | | | Standard | Attention | | | |
| Countries | S1 | AUC-PR | 90.83 ± 15.4 | 99.98 ± 0.05 | **100.0 ± 0.0** | **100.0 ± 0.0** | **100.0 ± 0.0** | `locatedIn`(X,Y) :– `locatedIn`(X,Z), `locatedIn`(Z,Y) |
| | S2 | | 87.40 ± 11.7 | 90.82 ± 0.88 | **93.48 ± 3.29** | 75.1 ± 0.3 | 92.36 ± 2.41 | `neighborOf`(X,Y) :– `neighborOf`(X,Z), `locatedIn`(Z,Y) |
| | S3 | | 56.68 ± 17.6 | 87.70 ± 4.79 | 91.27 ± 4.02 | 92.20 ± 0.2 | **95.10 ± 1.20** | `neighborOf`(X,Y) :– `neighborOf`(Y,X) |
| Kinship | | MRR | 0.35 | 0.719 | **0.759** | 0.619 | 0.720 | `term0`(X, Y) :– `term0`(Y, X) |
| | | HITS@1 | 0.24 | 0.586 | **0.642** | 0.475 | 0.605 | `term4`(X, Y) :– `term4`(Y, X) |
| | | HITS@3 | 0.37 | 0.815 | **0.850** | 0.707 | 0.812 | `term13`(X,Y) :– `term13`(X, Z), `term10`(Z, Y) |
| | | HITS@10 | 0.57 | 0.958 | **0.959** | 0.912 | 0.924 | `term2`(X,Y) :– `term4`(X, Z), `term7`(Z, Y) |

[Minervini et al. AAAI 2020]

# End-to-End Differentiable Reasoning with Natural Language

We can embed facts from the KG and facts from text in a *shared embedding space*, and learn to reason over them *jointly:*




[Minervini et al. AAAI 2020]

# End-to-End Differentiable Reasoning with Natural Language

We can embed facts from the KG and facts from text in a *shared embedding space*, and learn to reason over them *jointly:*

Control Myself record_label Jam Recordings

$\text{record\_label}(X, Z) \leftarrow p_1(X, Y)$

$p_1(X, Z) \leftarrow p_2(X, Y) \wedge p_3(Y, Z)$

Control Myself [...] is a song by american rapper [...] Ell

Ell cools 1989 album [...] was released by [...] Jam Recordings

[Welbl et al. ACL 2019, Minervini et al. AAAI 2020]

# End-to-End Differentiable Reasoning with Natural Language

We can embed facts from the KG and facts from text in a *shared embedding space*, and learn to reason over them *jointly:*

Thrasyvoulos F.C. country Greece

country$(X, Z) \leftarrow p_1(X, Y)$

$p_1(X, Z) \leftarrow p_2(X, Y) \wedge p_3(Y, Z)$

Thrasyvoulos Fylis is a football club based in Fyli, Attica [...]

Fyli is a town and a municipality in the northwestern part of Attica, Greece

**[Welbl et al. ACL 2019, Minervini et al. AAAI 2020]**

# Thank you!