Embedding-Based Techniques

MATRICES, TENSORS, AND NEURAL NETWORKS

Probabilistic Models: Downsides

Embeddings

Limitation to Logical Relations

- Representation restricted by manual design
    - Clustering? Assymetric implications?
    - Information flows through these relations
- Difficult to generalize to unseen entities/relations

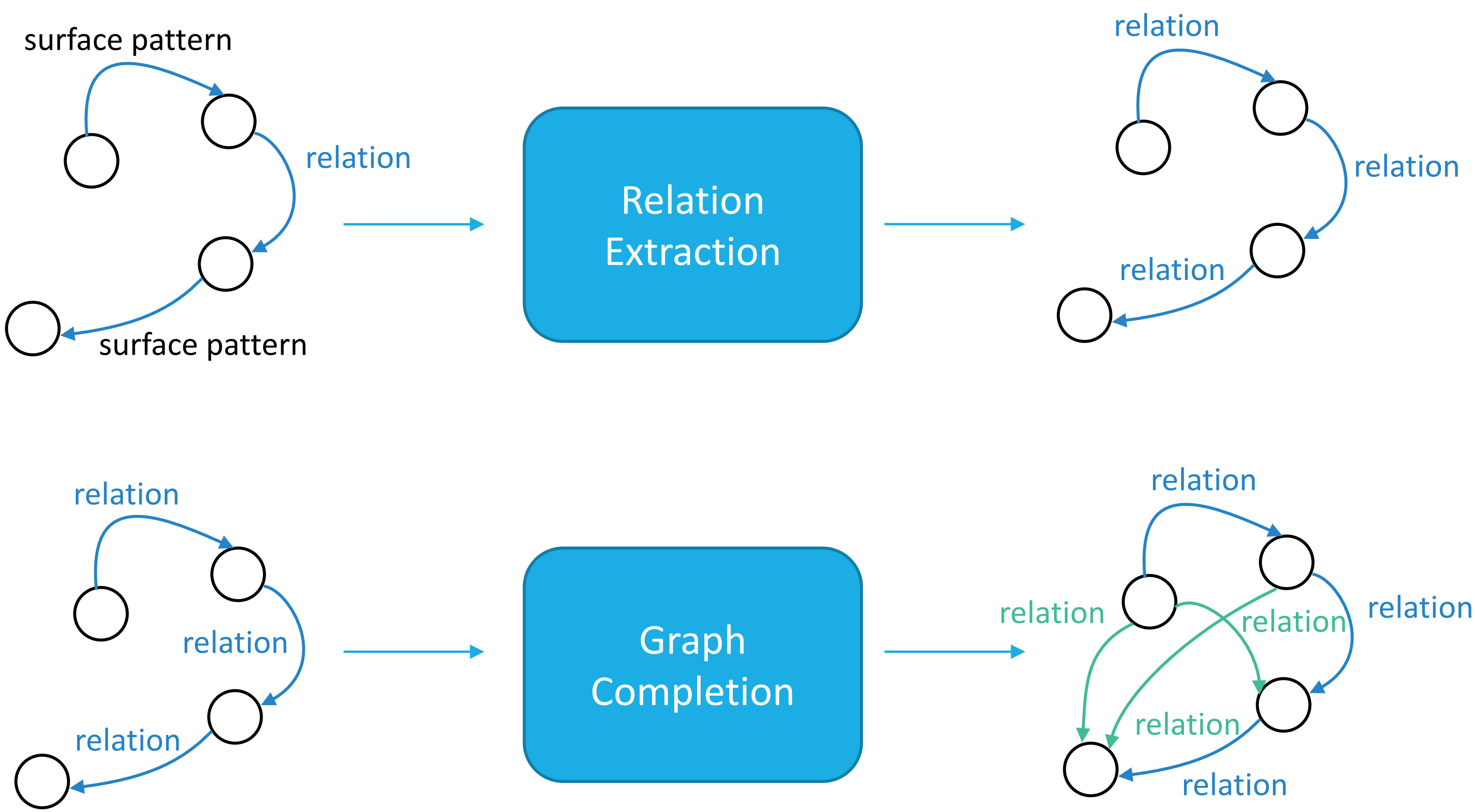
- Everything as dense vectors
- Can capture many relations
- Learned from data

Computational Complexity of Algorithms

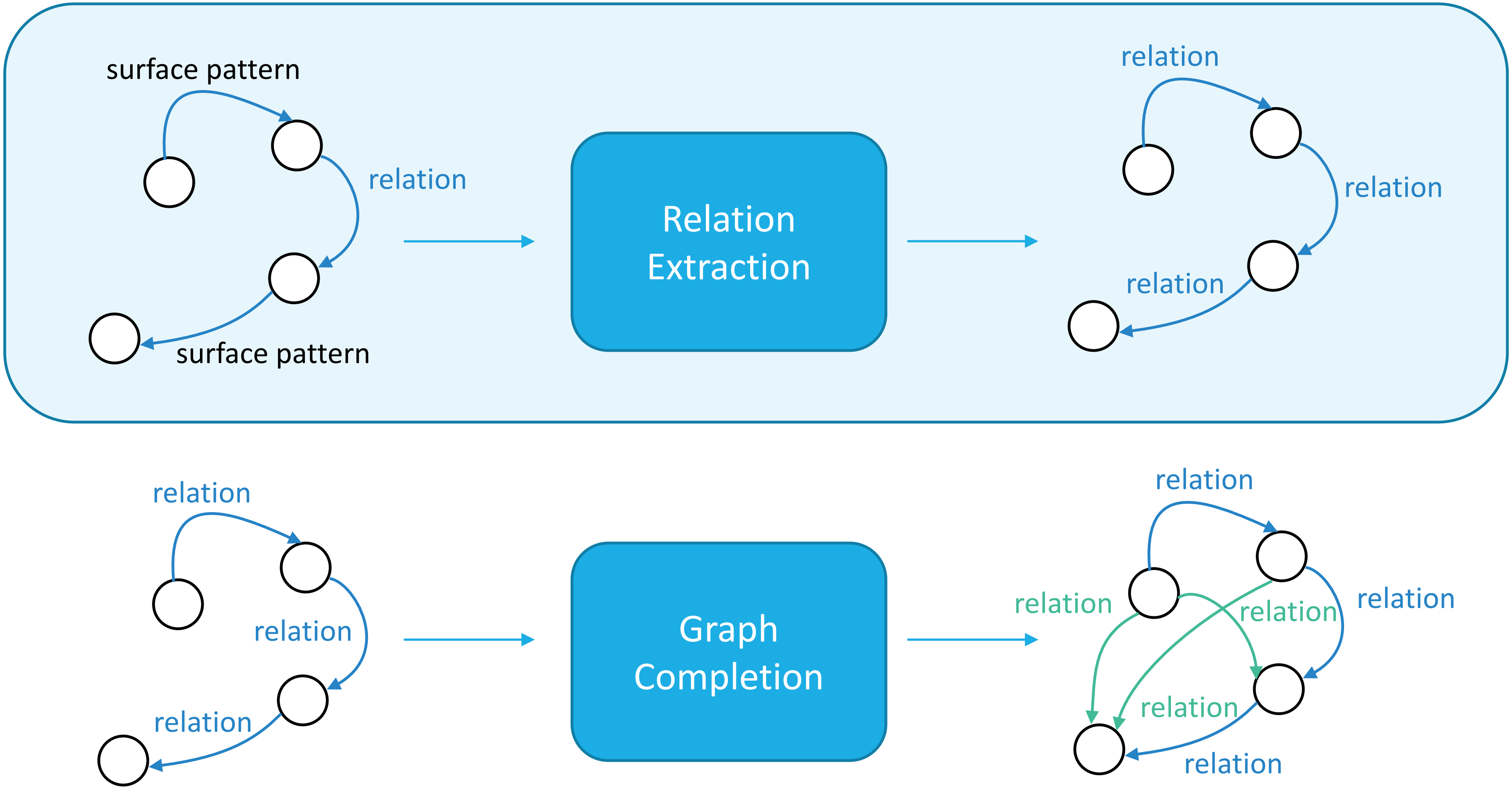
- Complexity depends on explicit dimensionality
    - Often NP-Hard, in size of data
    - More rules, more expensive inference
- Query-time inference is sometimes NP-Hard
- Not trivial to parallelize, or use GPUs

- Complexity depends on latent dimensions
- Learning using stochastic gradient, back-propagation
- Querying is often cheap
- GPU-parallelism friendly

Two Related Tasks

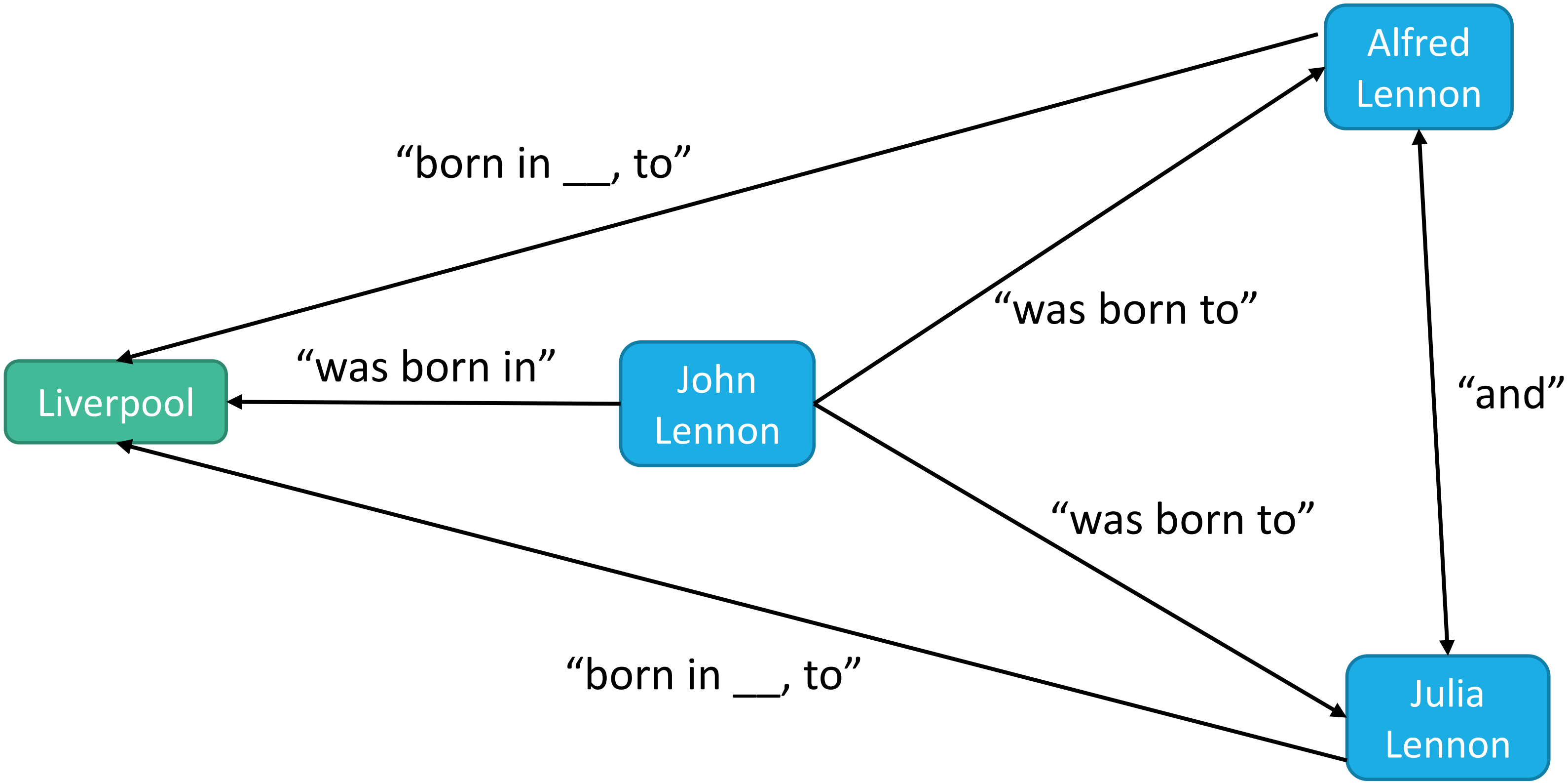


Two Related Tasks



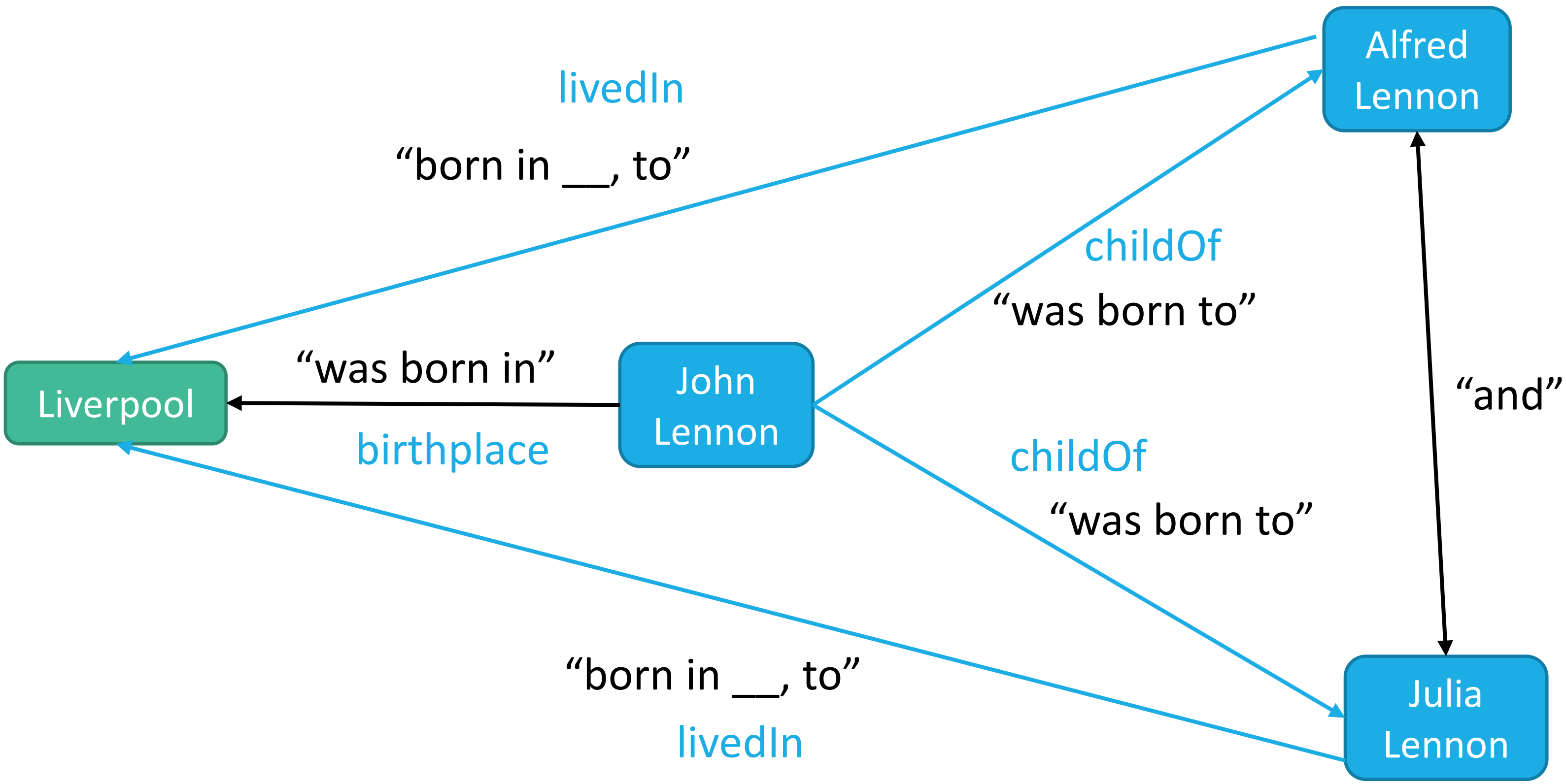
Relation Extraction From Text

John was born in Liverpool, to Julia and Alfred Lennon.



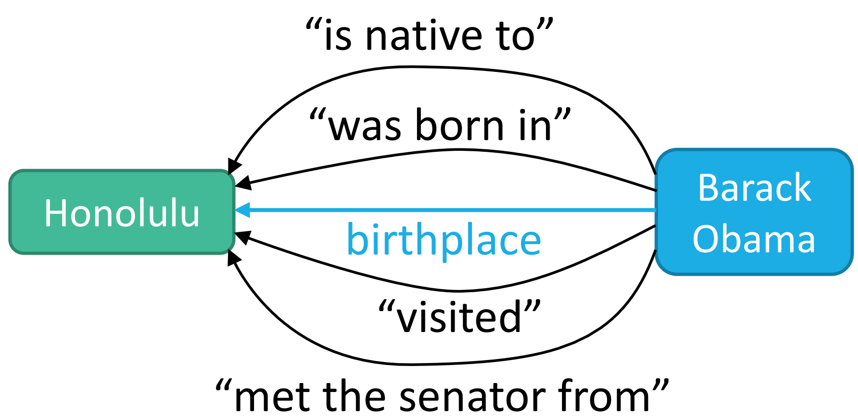
Relation Extraction From Text

John was born in Liverpool, to Julia and Alfred Lennon.



“Distant” Supervision

No direct supervision gives us this information.
Supervised: Too expensive to label sentences
Rule-based: Too much variety in language
Both only work for a small set of relations, i.e. 10s, not 100s

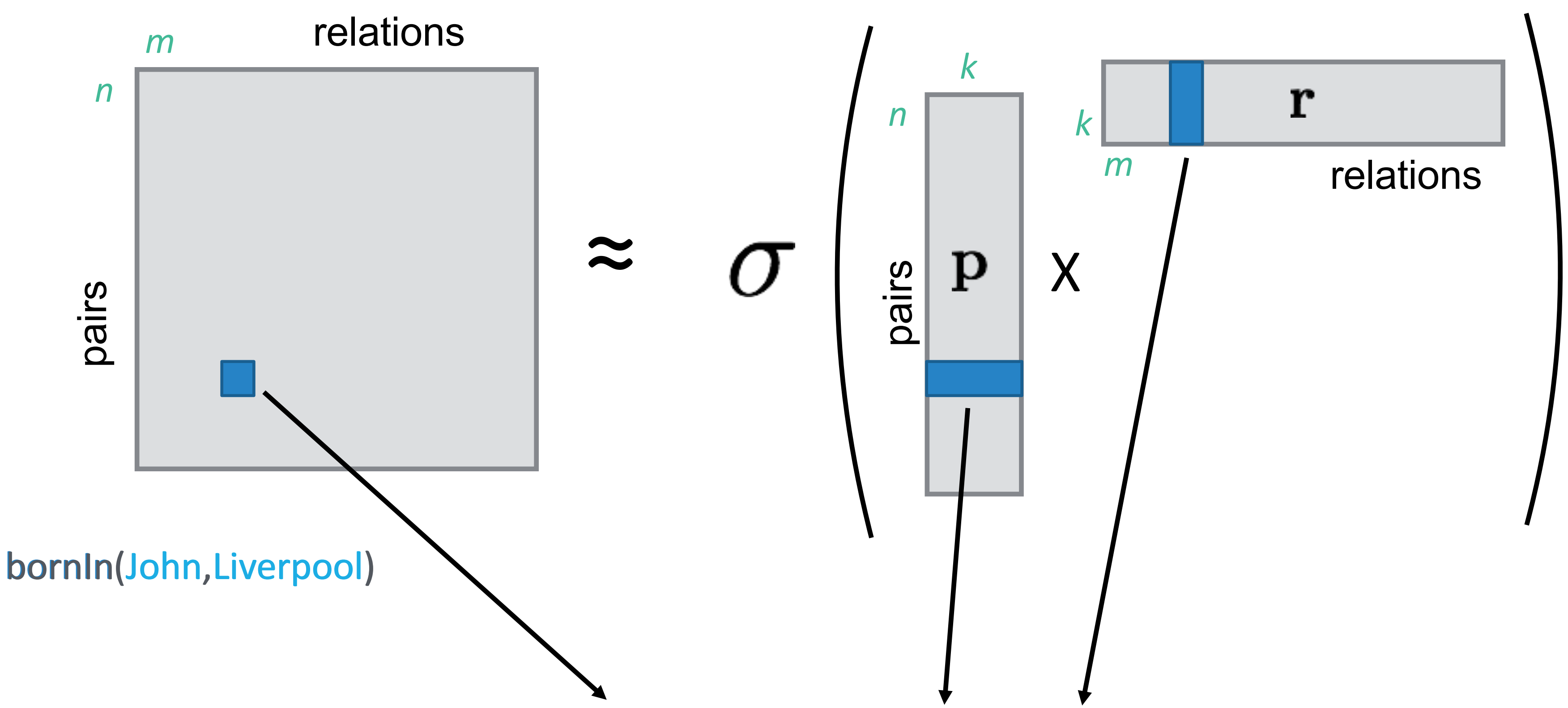


Relation Extraction as a Matrix

John was born in Liverpool, to Julia and Alfred Lennon.

Entity Pairs	was born in <-nsubjpass-born<-nmod:in-	was born to	and	birthplace(X,Y)	spouse(X,Y)
John Lennon, Liverpool	1			?	
John Lennon, Julia Lennon		1			
John Lennon, Alfred Lennon		1			
Julia Lennon, Alfred Lennon			1		?
Barack Obama, Hawaii	1			1	
Barack Obama, Michelle Obama			1		1

Matrix Factorization



$$P(R(i,j)) = \sigma(\mathbf{p}_{i,j} \cdot \mathbf{r}_R)$$

Training

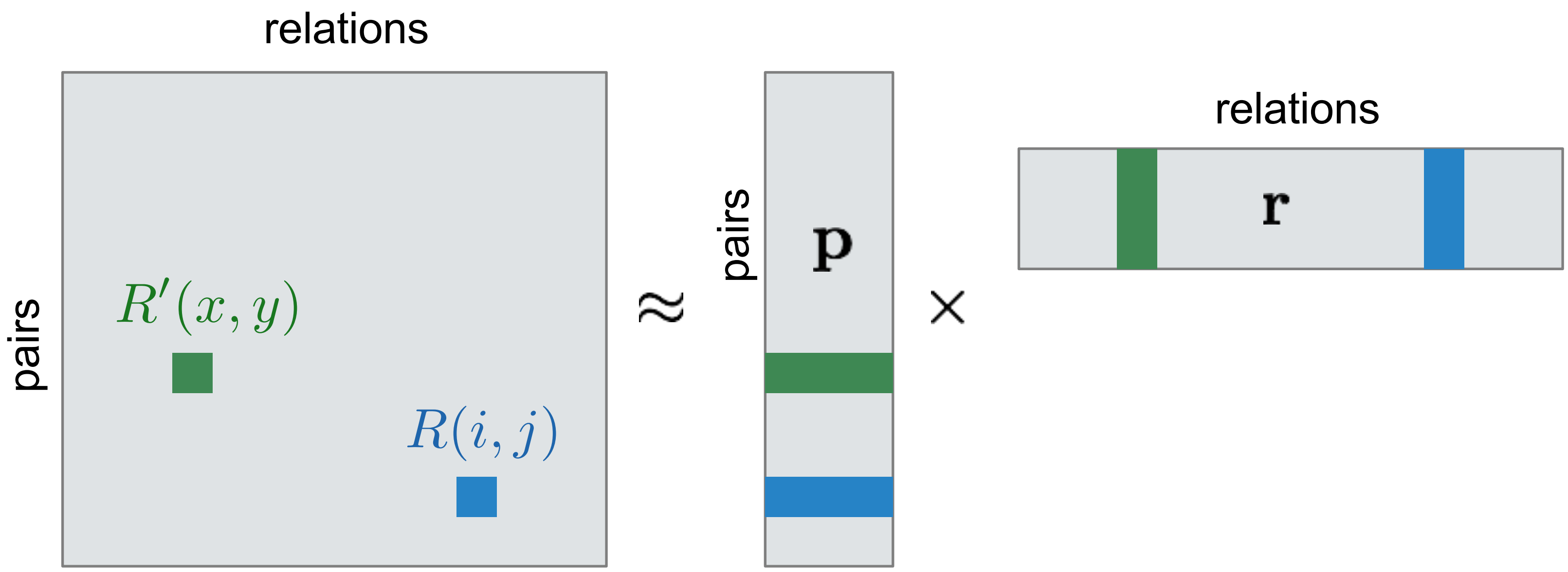
Loss Function: $$\max_{\mathbf{v},\mathbf{w}} \log \prod_{x,y,r} \exp\langle \mathbf{v}^{x,y}, \mathbf{w}_r \rangle - \lambda(||\mathbf{v}||_2^2 + ||\mathbf{w}||_2^2)$$

Desiderata from the training algorithm:

- Do not instantiate the whole matrix!
- Do not hold all the observed cells in memory
- Each iteration linear in the no. of observations

Solution: Stochastic Gradient Descent!

Training: Stochastic Updates



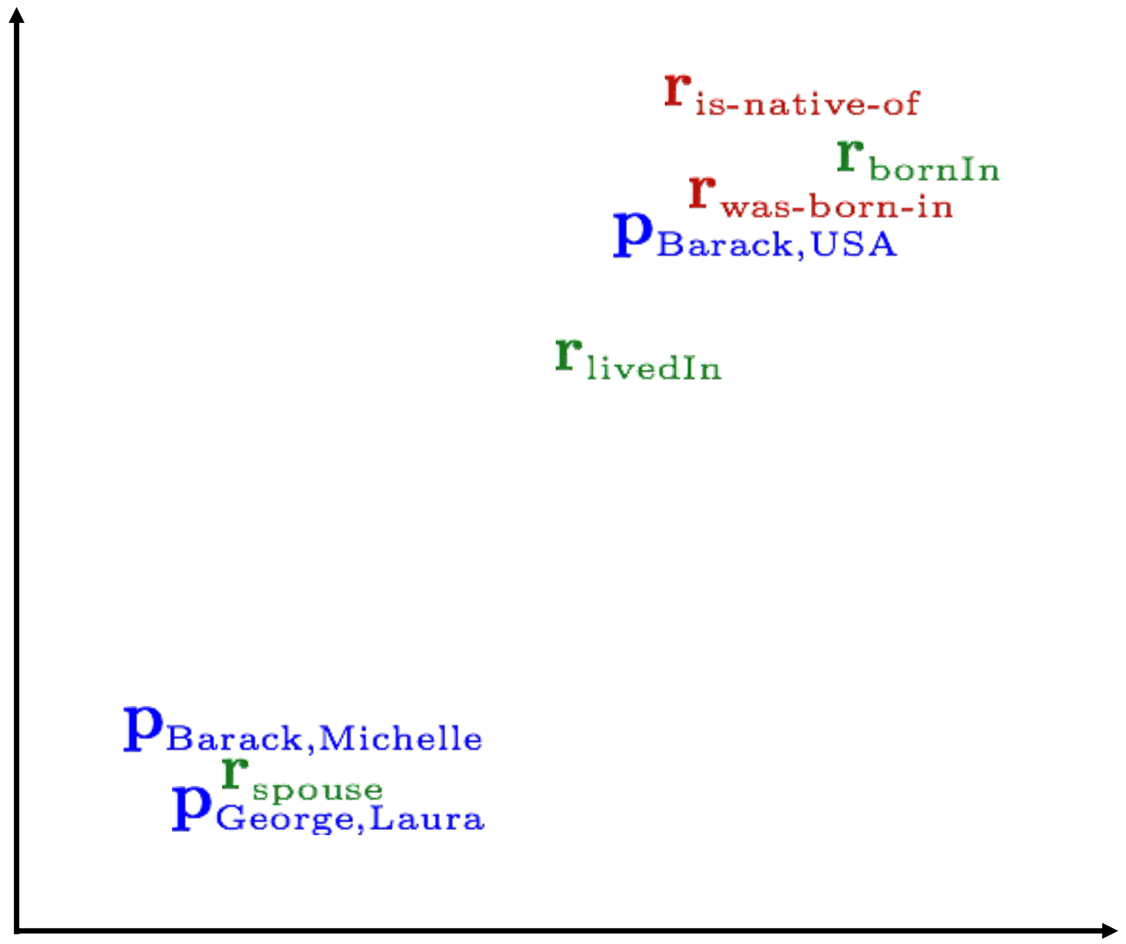
Pick an observed cell, $R(i,j)$:

- Update $\mathbf{p}_{ij}$ & $\mathbf{r}_R$ such that $R(i,j)$ is higher

Pick any random cell, assume it is negative:

- Update $\mathbf{p}_{xy}$ & $\mathbf{r}_{R'}$ such that $R'(x,y)$ is lower

Relation Embeddings



Embeddings ~ Logical Relations

Relation Embeddings, w

- Similar embedding for 2 relations denote they are paraphrases
  - is married to, spouseOf(X,Y), /person/spouse
- One embedding can be contained by another
  - w(topEmployeeOf) ⊂ w(employeeOf)
  - topEmployeeOf(X,Y) → employeeOf(X,Y)
- Can capture logical patterns, without needing to specify them!

Entity Pair Embeddings, v

Similar entity pairs denote similar relations between them

Entity pairs may describe multiple "relations"

independent foundedBy and employeeOf relations

Similar Embeddings

similar underlying embedding

similar embedding

	X **own percentage of** Y	X **buy stake in** Y
Time, Inc Amer. Tel. and Comm.	1	1
Volvo Scania A.B.		1
Campeau Federated Dept Stores		
Apple HP		

Successfully predicts “Volvo owns percentage of Scania A.B.” from “Volvo bought a stake in Scania A.B.”

Implications

X historian at Y → X professor at Y

(Freeman,Harvard)
→ (Boyle,OhioState)

	X **professor at** Y	X **historian at** Y
Kevin Boyle Ohio State		1
R. Freeman Harvard	1	

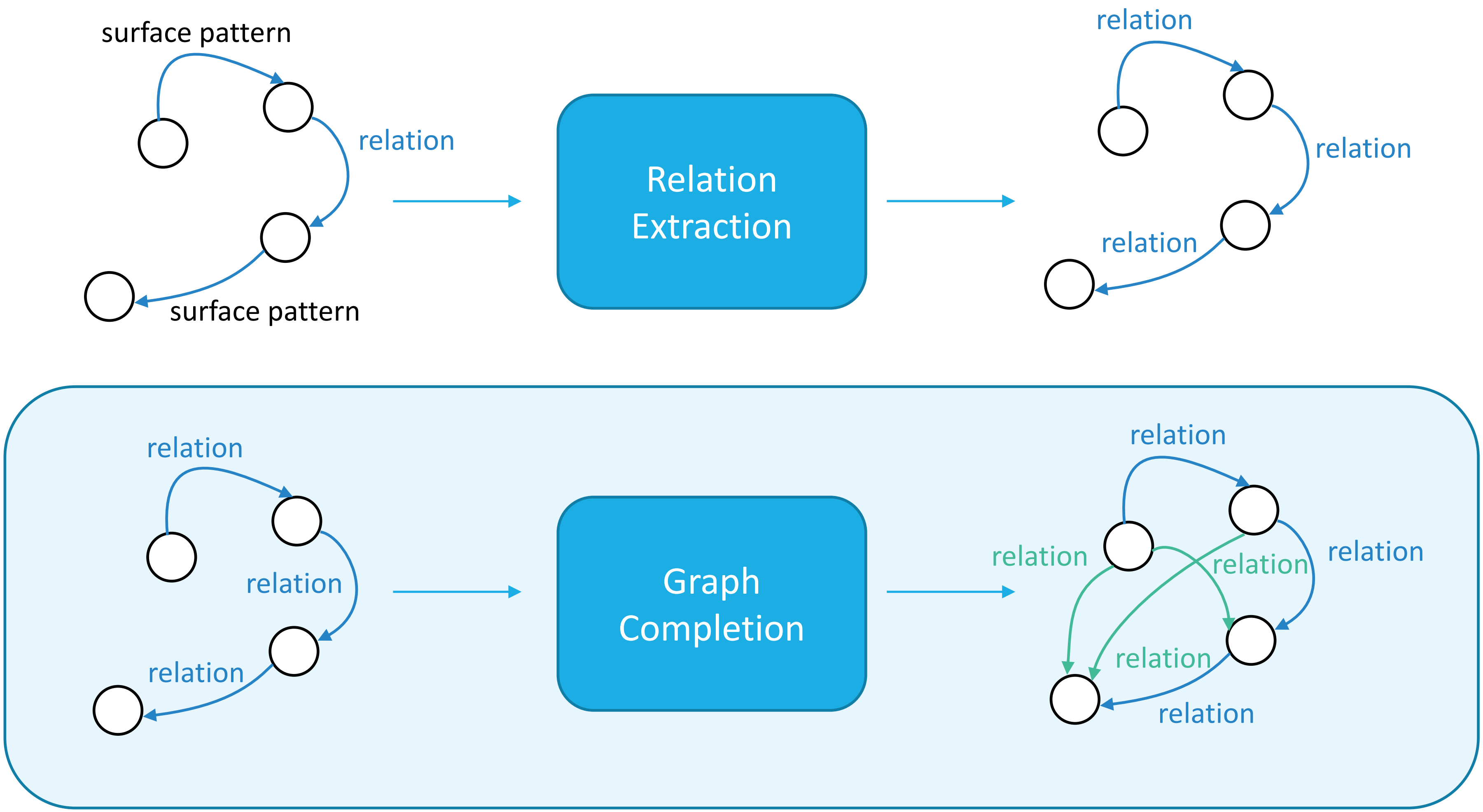
Learns asymmetric entailment:

PER historian at UNIV → PER professor at UNIV

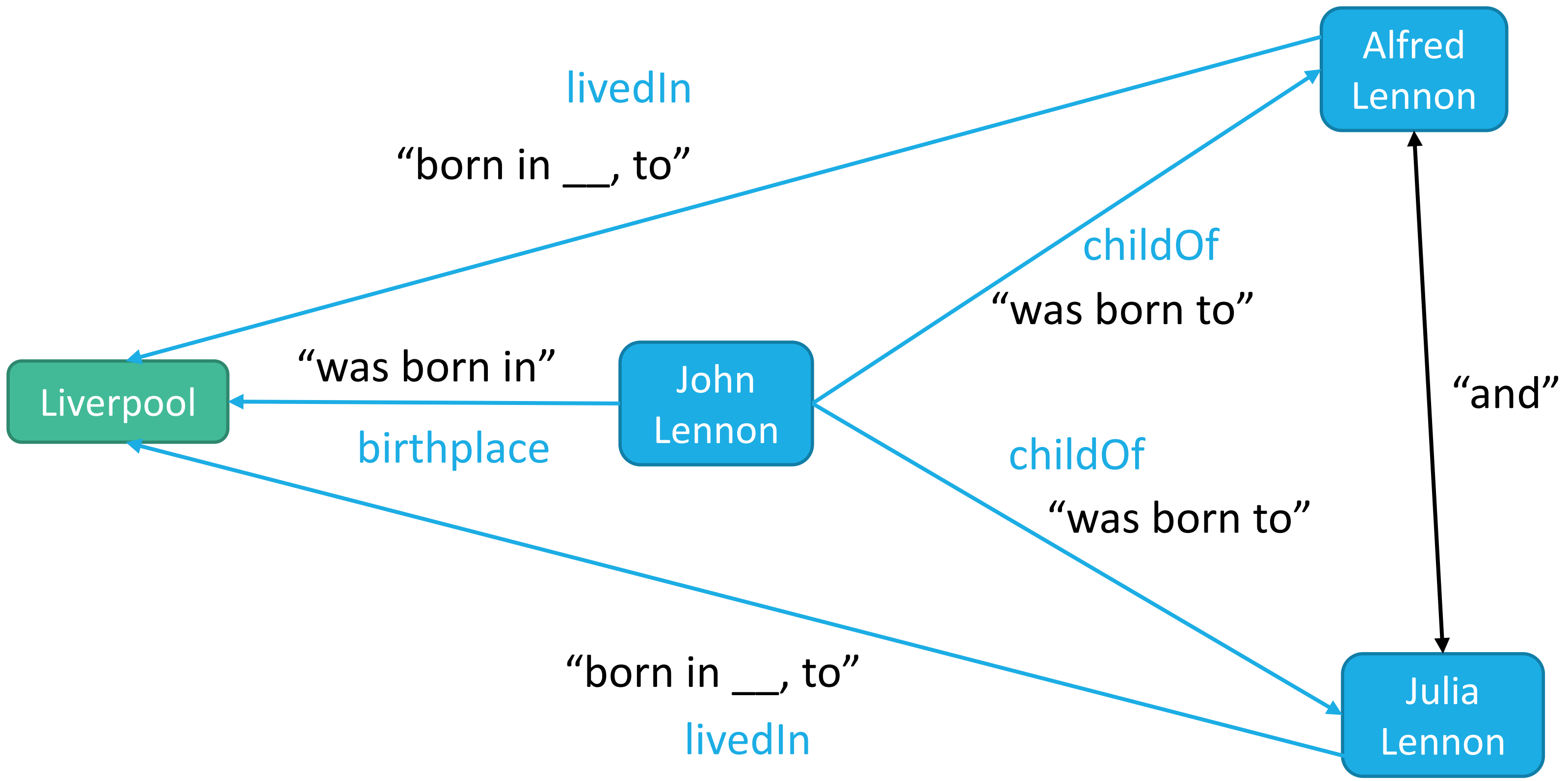
But,

PER professor at UNIV $\nrightarrow$ PER historian at UNIV

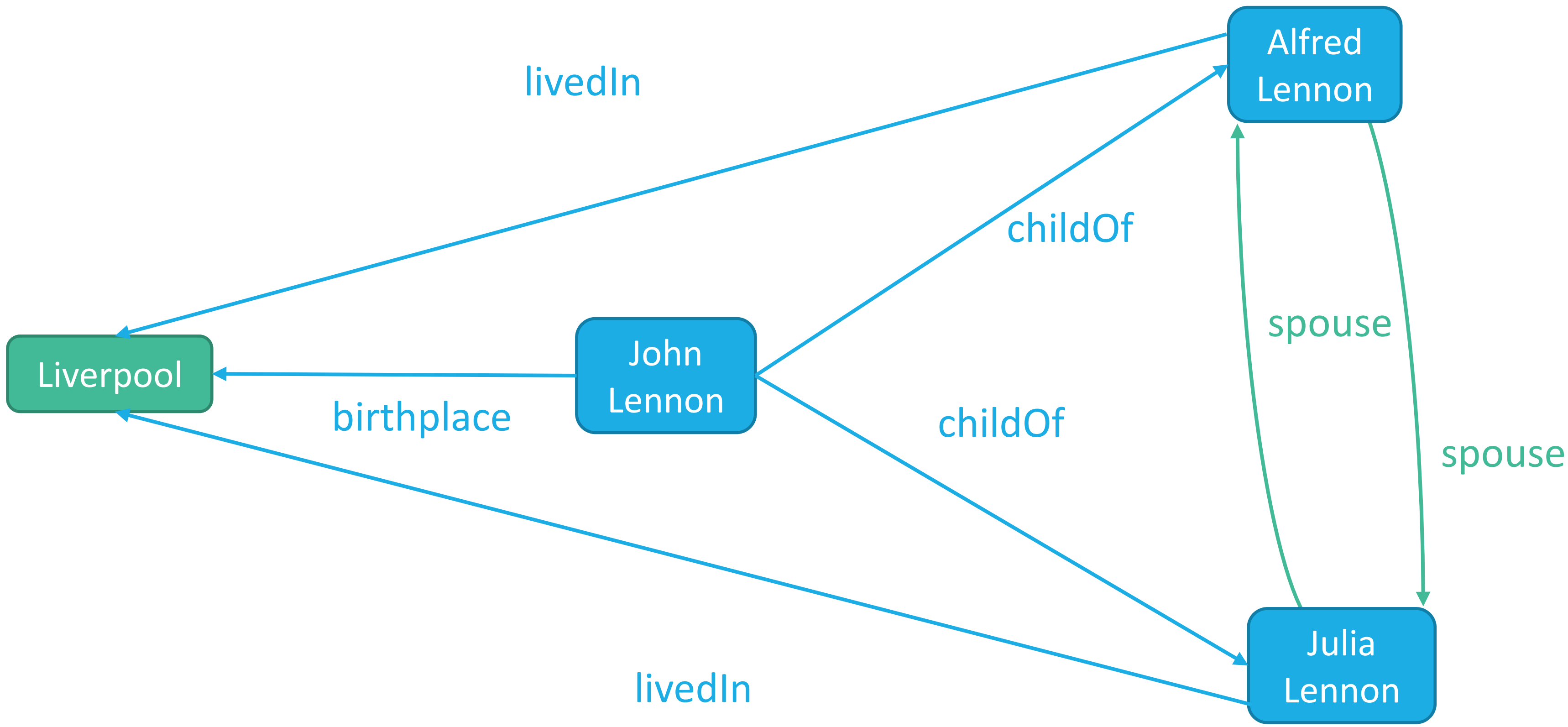
Two Related Tasks



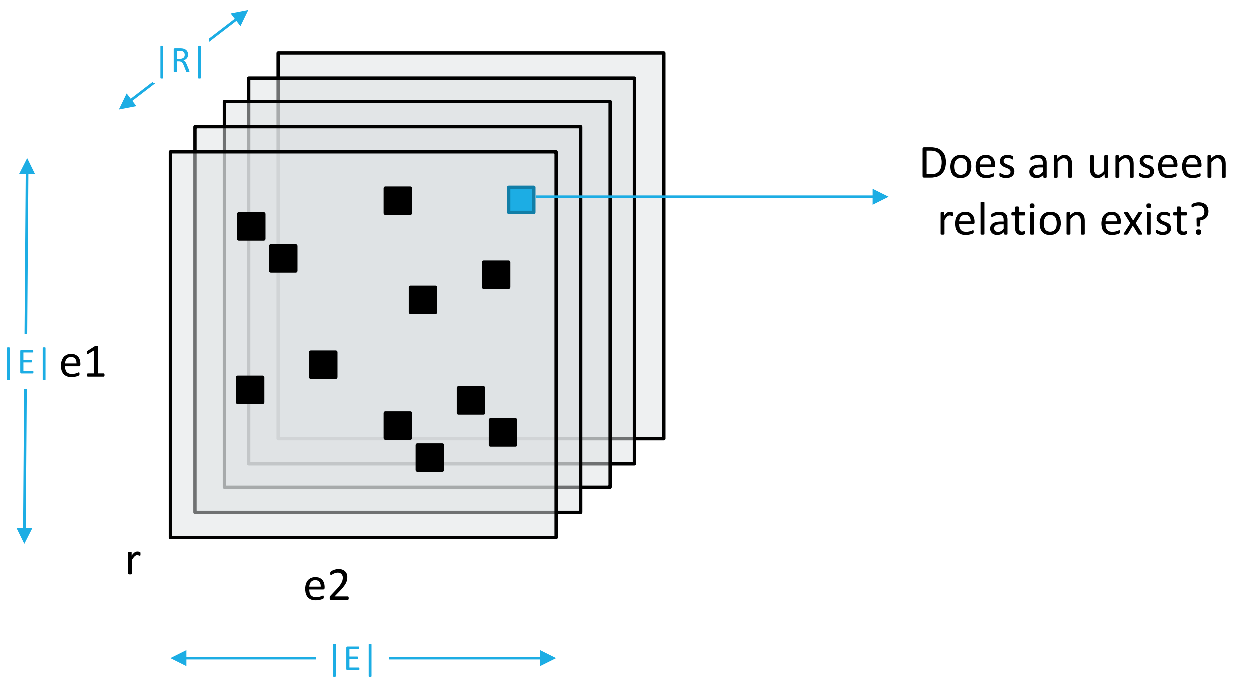
Graph Completion



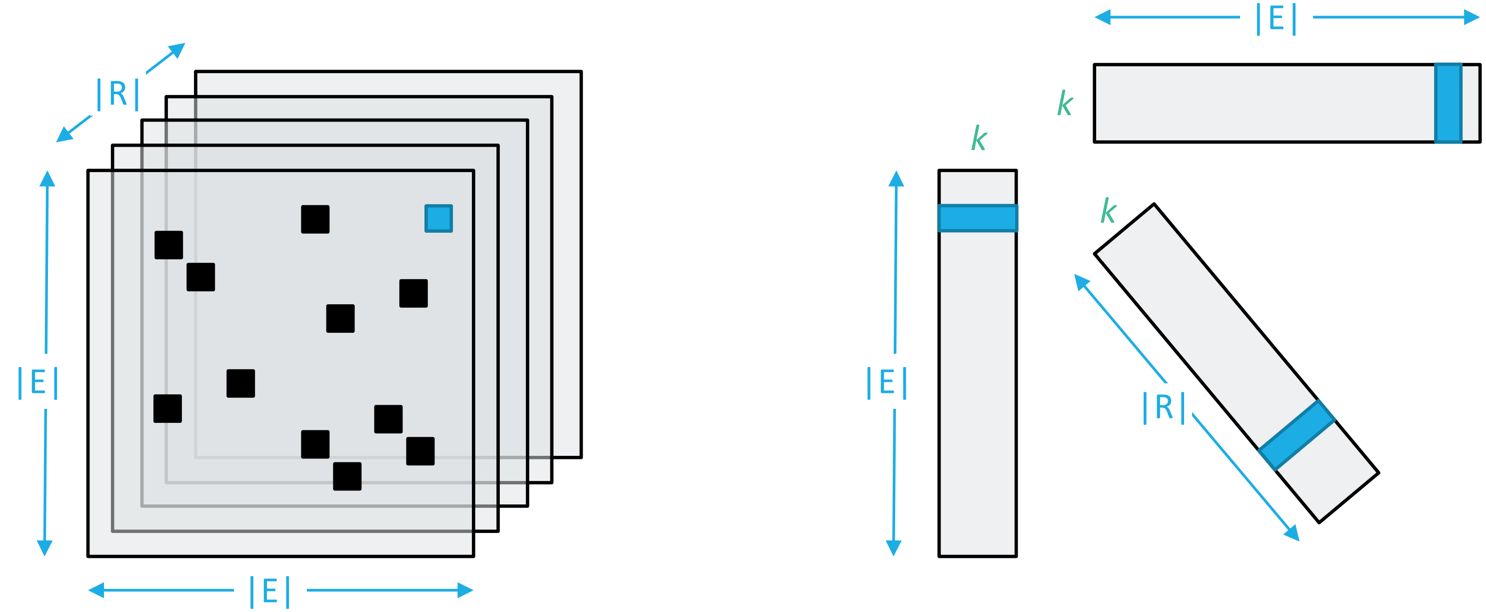
Graph Completion



Tensor Formulation of KG



Factorize that Tensor



$$S(r(a,b)) = f(\mathbf{v}_r, \mathbf{v}_a, \mathbf{v}_b)$$

Many Different Factorizations

CANDECOMP/PARAFAC-Decomposition

$$S\left(r(a,b)\right) = \sum_k R_{r,k} \cdot e_{a,k} \cdot e_{b,k}$$

Tucker2 and RESCAL Decompositions

$$S\left(r(a,b)\right) = \left(\mathbf{R}_r \times \mathbf{e}_a\right) \times \mathbf{e}_b$$

Model E

$$S\left(r(a,b)\right) = \mathbf{R}_{r,1} \cdot \mathbf{e}_a + \mathbf{R}_{r,2} \cdot \mathbf{e}_b$$

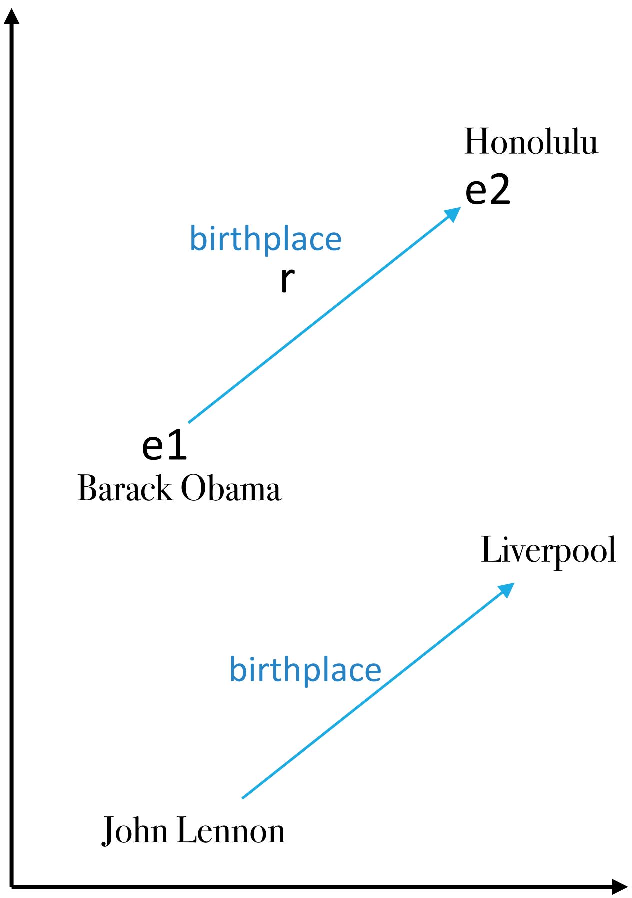
Holographic Embeddings

$$S(r(a,b)) = \mathbf{R}_r \times (\mathbf{e}_a \star \mathbf{e}_b)$$

Not tensor factorization (per se)

HOLE: Nickel et al, AAAI (2016), Model E: Riedel et al, NAACL (2013), RESCAL: Nickel et al, WWW (2012), CP: Harshman (1970), Tucker2: Tucker (1966)

Translation Embeddings



TransE

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a + \mathbf{R}_r - \mathbf{e}_b\|_2^2$$

TransH

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a^{\perp} + \mathbf{R}_r - \mathbf{e}_b^{\perp}\|_2^2$$

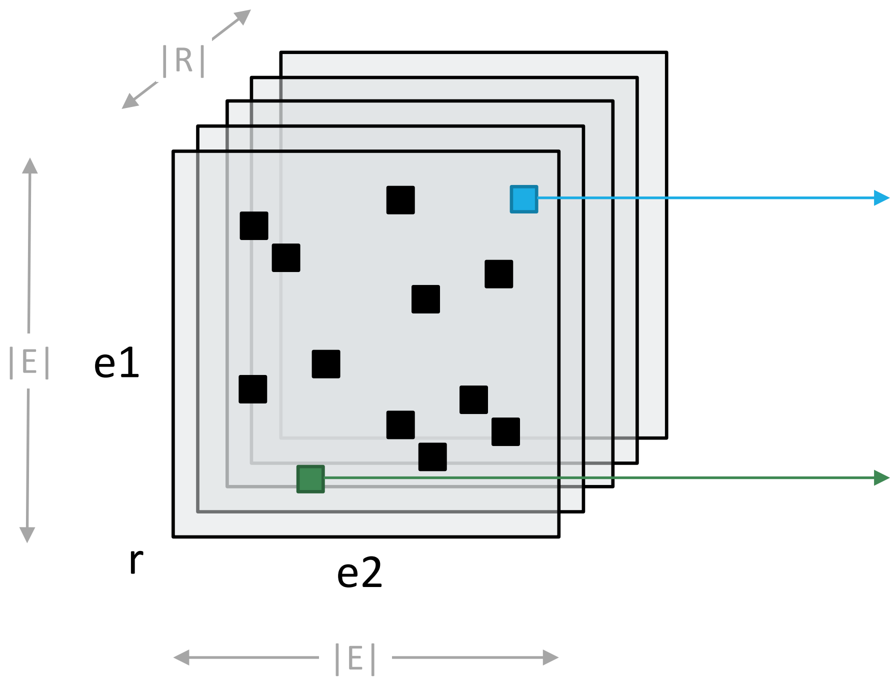
$$\mathbf{e}_a^{\perp} = \mathbf{e}_a - \mathbf{w}_r^T \mathbf{e}_a \mathbf{w}_r$$

TransR

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a \mathbf{M}_r + \mathbf{R}_r - \mathbf{e}_b \mathbf{M}_r\|_2^2$$

TransE: Bordes et al. XXX (2011), TransH: Bordes et al. XXX (2011), TransR: Bordes et al. XXX (2011)

Parameter Estimation



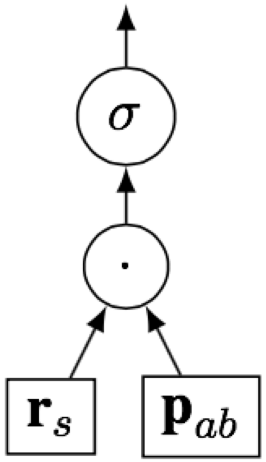
Observed cell: increase score

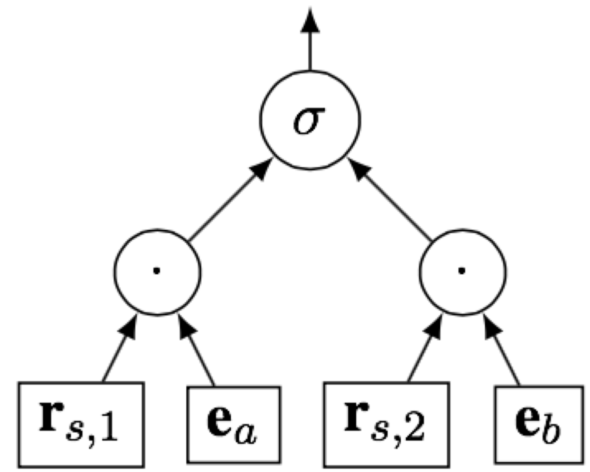
$$S\left(r(a,b)\right)$$

Unobserved cell: decrease score

$$S\left(r'(x,y)\right)$$

Matrix vs Tensor Factorization

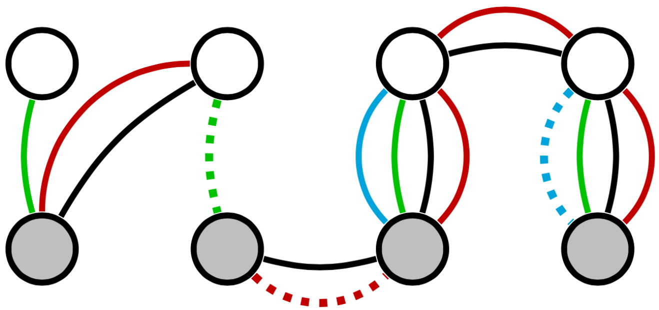




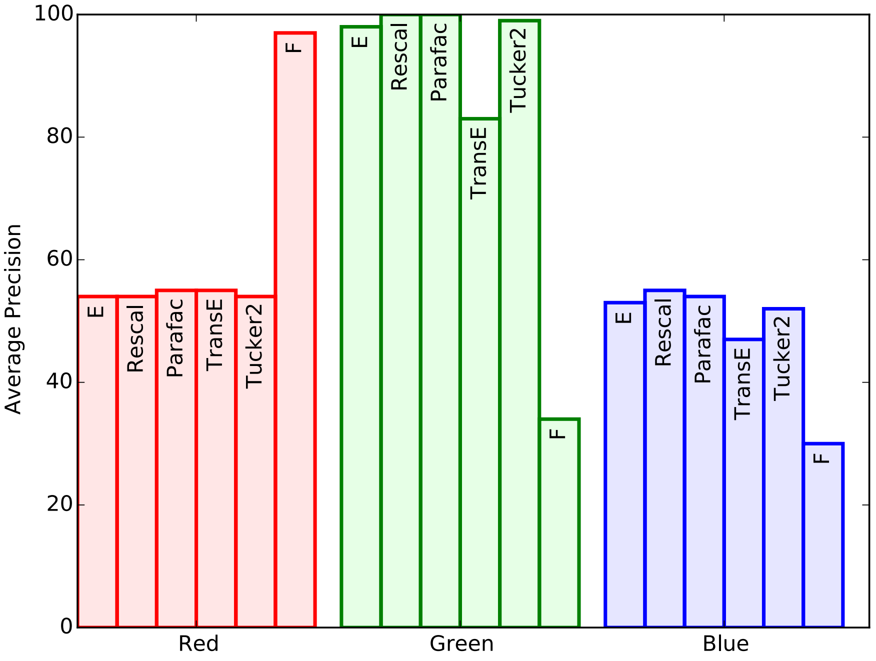
- No encoding of type information
- Can only predict for entity pairs that appear in text together
- Sufficient evidence has to be seen for each entity pair

- Assume low-rank for pairs
- But many relations are not!
- Spouse: you can have only ~1
- Cannot learn pair specific information

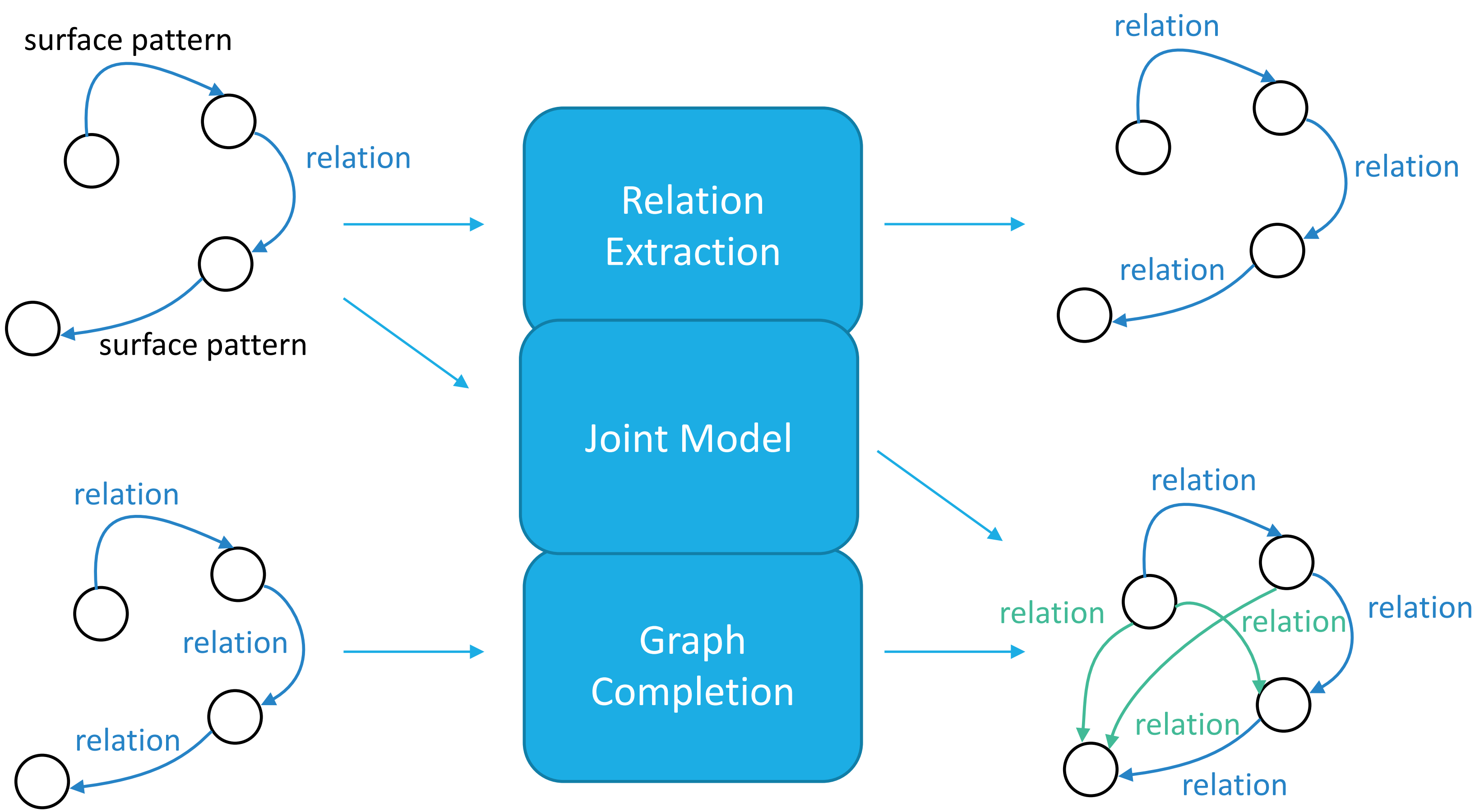
What they can, and can't, do..



- **Red**: deterministically implied by **Black**
  - needs *pair-specific* embedding
  - Only **F** is able to generalize
- **Green**: needs to estimate entity types
  - needs *entity-specific* embedding
  - Tensor factorization generalizes, **F** doesn't
- **Blue**: implied by **Red** and **Green**
  - Nothing works much better than random



Joint Extraction+Completion



Compositional Neural Models

So far, we're learning vectors for each entity/surface pattern/relation..

But learning vectors independently ignores "composition"

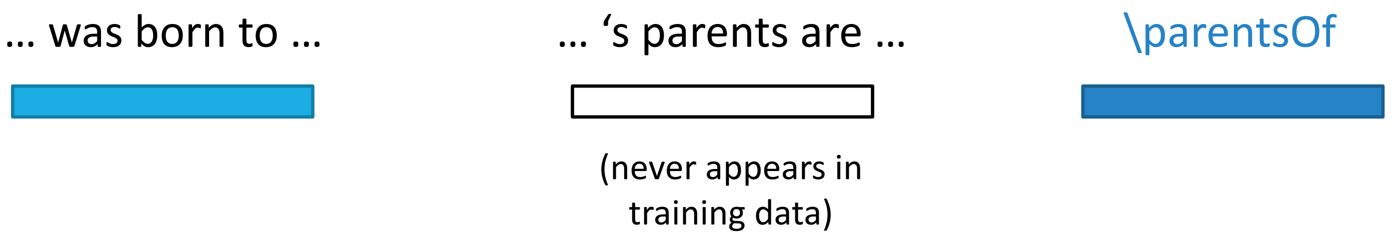
Composition in Surface Patterns

- Every surface pattern is not unique
- Synonymy: A is B's spouse.
  A is married to B.
- Inverse: X is Y's parent.
  Y is one of X's children.
- Can the representation learn this?

Composition in Relation Paths

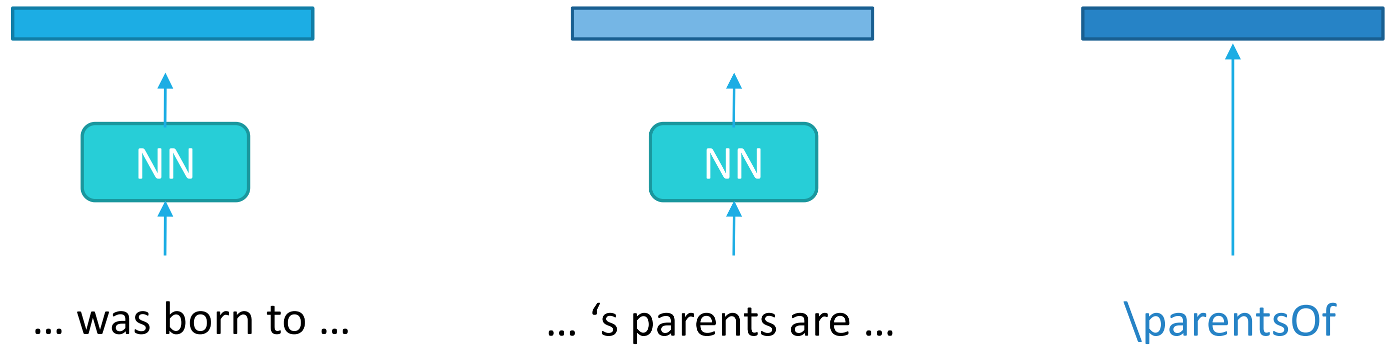
- Every relation path is not unique
- Explicit: A parent B, B parent C
  A grandparent C
- Implicit: X bornInCity Y, Y cityInState Z
  X "bornInState" Z
- Can the representation capture this?

Composing Dependency Paths

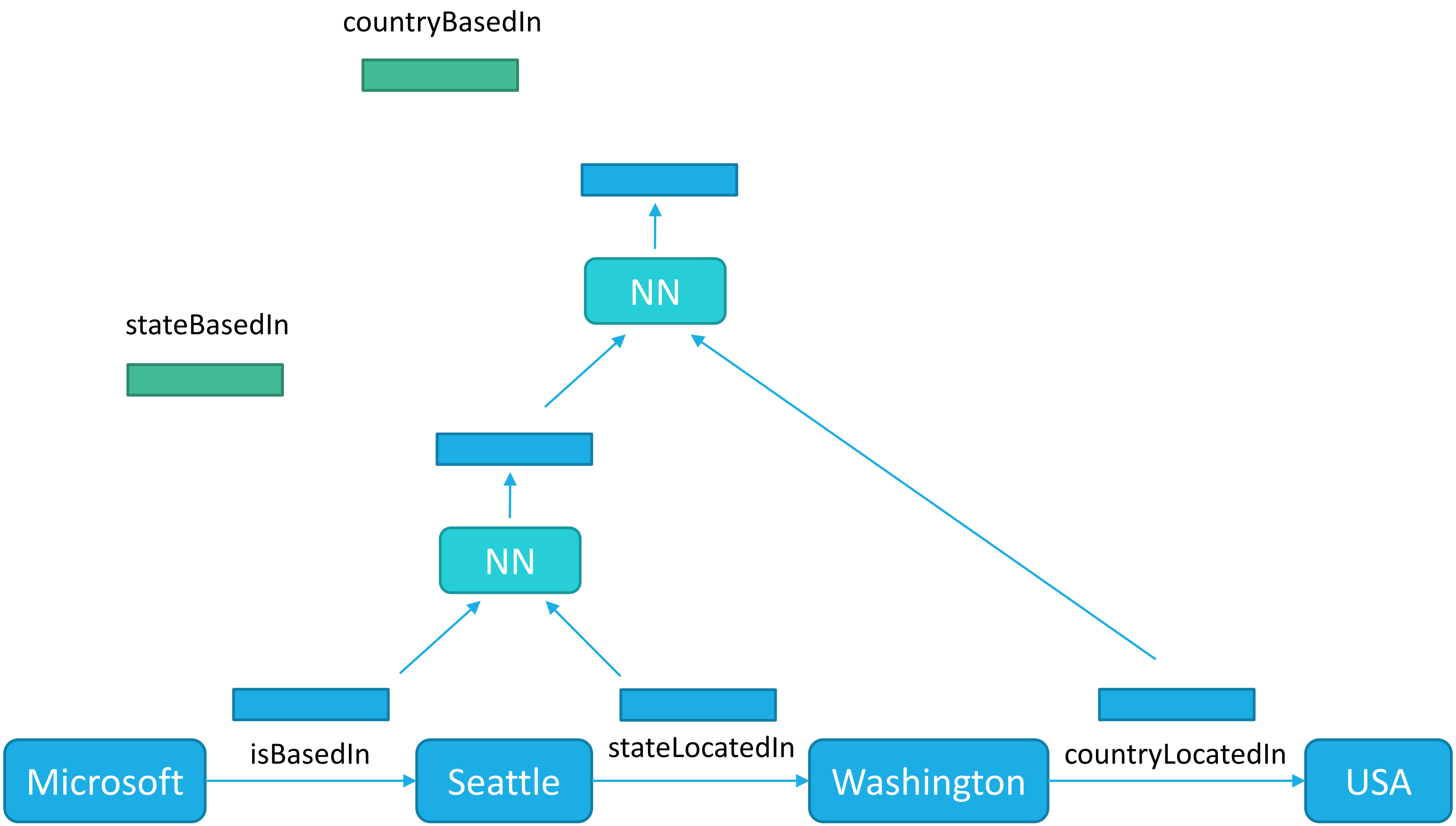


But we don’t need linked data to know they mean similar things…

Use neural networks to produce the embeddings from text!



Composing Relational Paths



Neelakantan et al (2015), http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/viewFile/10254/10032

Lin et al, EMNLP (2015), https://arxiv.org/pdf/1506.00379.pdf

Review: Embedding Techniques

Two Related Tasks:

- Relation Extraction from Text
- Graph (or Link) Completion

Relation Extraction:

- Matrix Factorization Approaches

Graph Completion:

- Tensor Factorization Approaches

Compositional Neural Models

- Compose over dependency paths
- Compose over relation paths

Tutorial Overview