

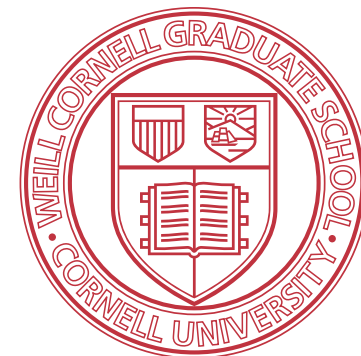
(adapted from a talk I gave in Alicante at the “International Workshop on Embeddings and Semantics”, Sept 15th)


A Generative Model of Words and Relationships from Multiple Sources

Stephanie L. Hyland^{1,2}, Theofanis Karaletsos¹, Gunnar Rätsch¹

¹Computational Biology Program, Memorial Sloan Kettering Cancer Centre, New York

²Weill Cornell Graduate School of Medical Sciences, New York



 @__hylandSL

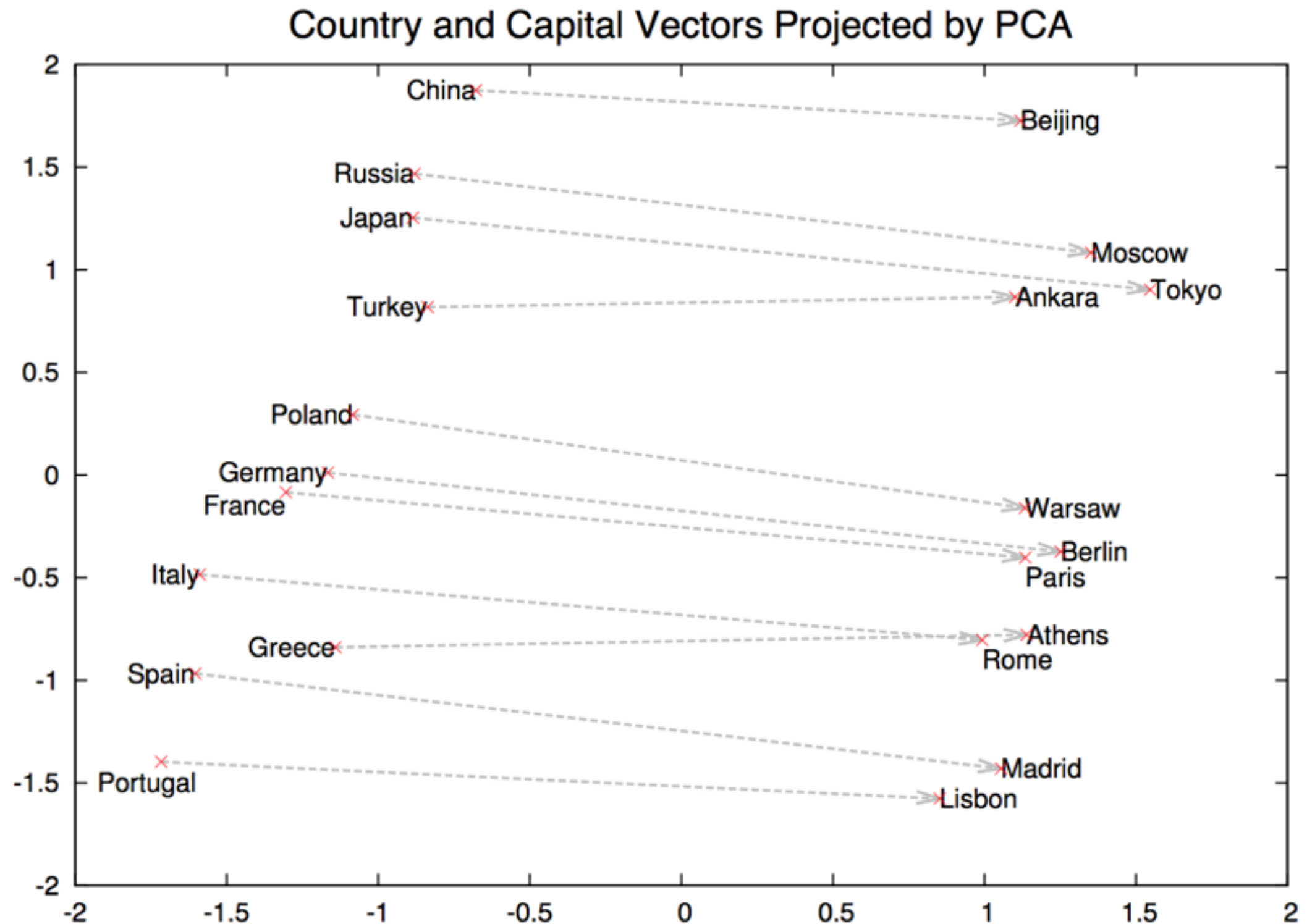
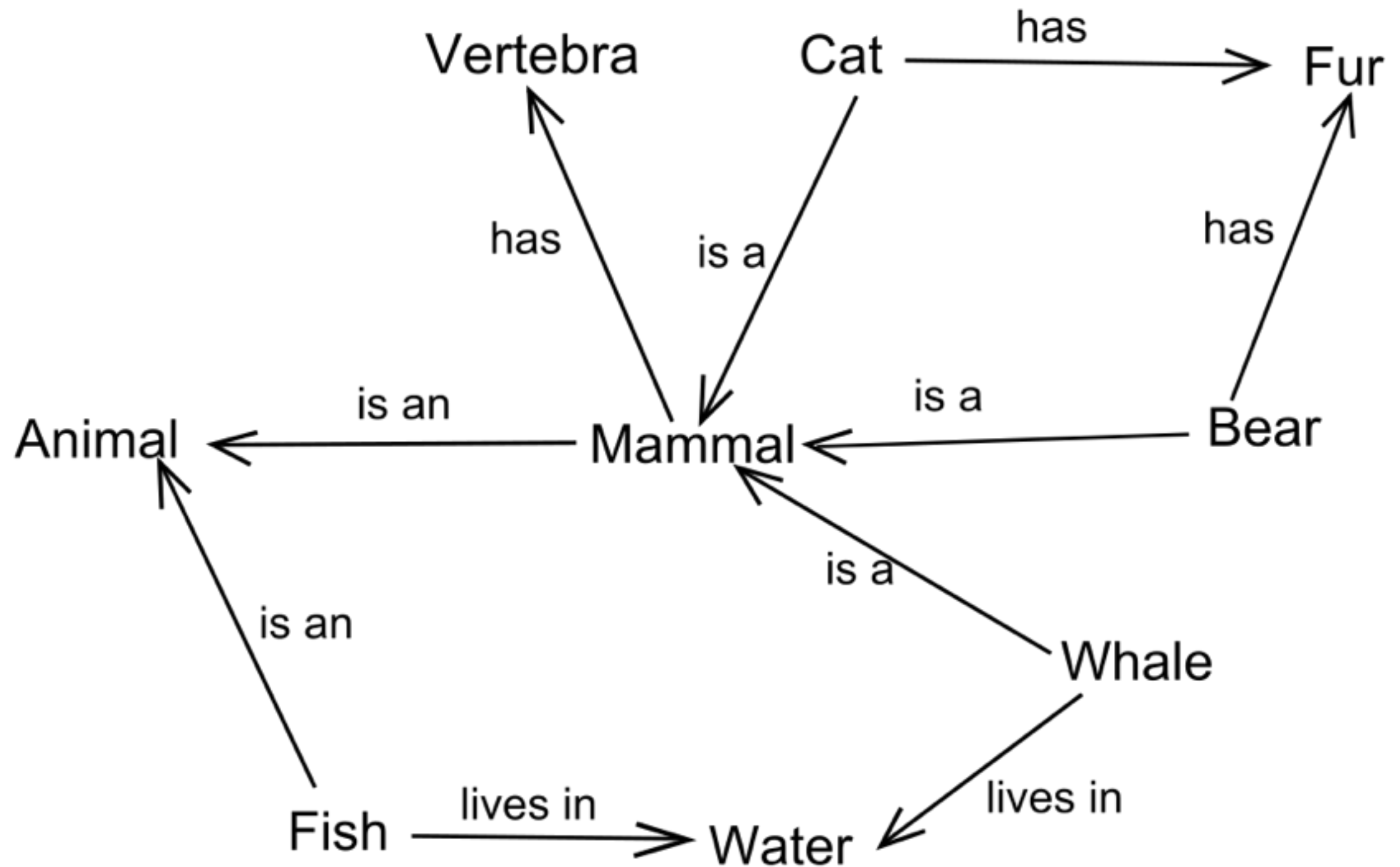


figure 2 from Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality", NIPS 2013

er was allegedly first published in 1730 by philip j. berg, a swedish officer who traveled in the eastern russian empire of war after the great northern war. so this attempt also fails the axiom of dependent choice (dc). results requiring ac (or v) but weaker than it. the subject of much debate, a short description, but gives all the credit of the successful negotiation to le coastline dates back to about 4000 bc. however, each of these ins after the release of stanley kubrick's popular 1971 film adaptation "k orange", itself the subject of much controversy. unhappy v to abandon that score, he heavily criticised the band's experiment ip hop, liturgical and gothic music. shortly before the first world gan expanding, and new suburbs were built. there are three financial districts in amsterdam. hotels with 4 or 5 stars contribute total beds available and 41% of the overnight stays in amsterdam team amstel tigers play in the jaap eden ice rink. august horch position at the ministry of transport, bu

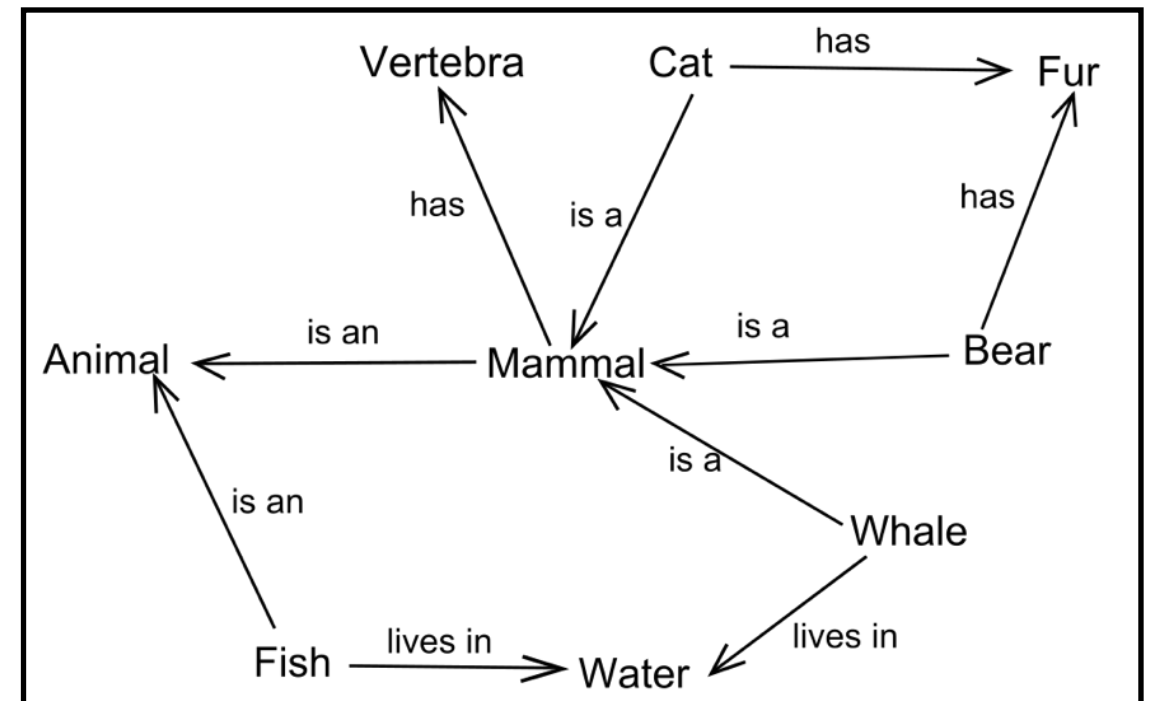
er was allegedly first published in 1730 by philip j...
erg, a swedish officer who traveled in the eastern russian empire
r of war after the great northern war. so this attempt also fail
is the axiom of dependent choice (dc). results requiring ac (or v)
ut weaker than it. **the subject of much debate**, a short descrip
ting, but gives all the credit of the successful negotiation to le
oastline dates back to about 4000 bc. however, each of these ins
ter the release of stanley kubrick's popular 1971 film adaptatio
rk orange", itself **the subject of much controversy**. unhappy w
to abandon that score, he heavily criticised the band's exper
ip hop, liturgical and gothic music. shortly before the first world
gan expanding, and new suburbs were built. there are three
financial districts in amsterdam. hotels with 4 or 5 stars contrib
otal beds available and 41% of the overnight stays in amsterdam
team amstel tigers play in the jaap eden ice rink. august horch
position at the ministry of transport, bu



example semantic network from Wikipedia:
https://en.wikipedia.org/wiki/Semantic_network

combine **structured** and **unstructured** data

On Stramonio...
thern war. so this attempt also fails
orms) but weaker than it. the sub
e successful negotiation to leo. t
came after the release of stanl
ct of much controversy. unhappy
mix of hip hop, liturgical and go
burbs were built. there are three o
total beds available and 41%



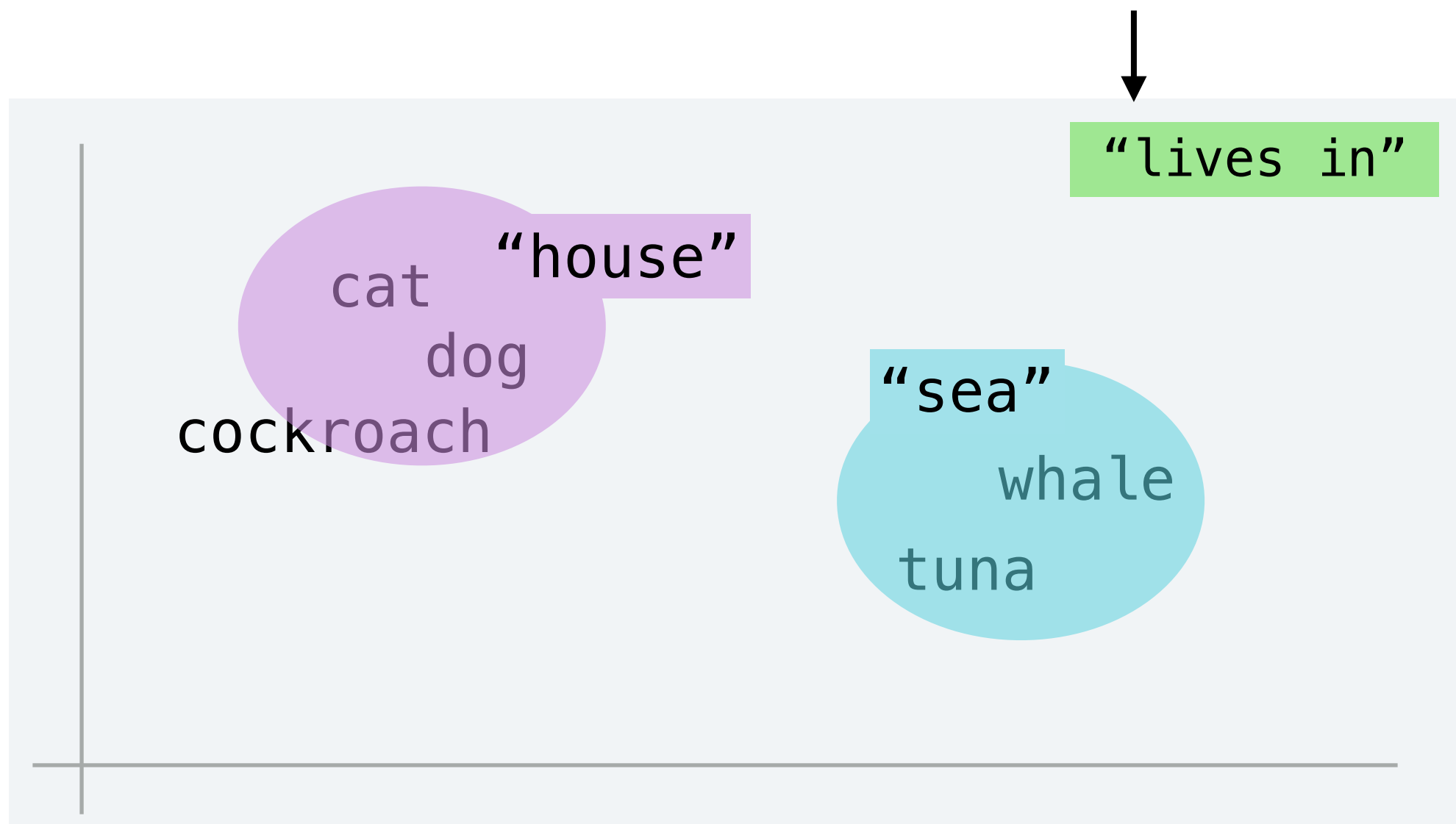
learn representations of **entities** (words) as well as
relationships between them



transformations of vectors

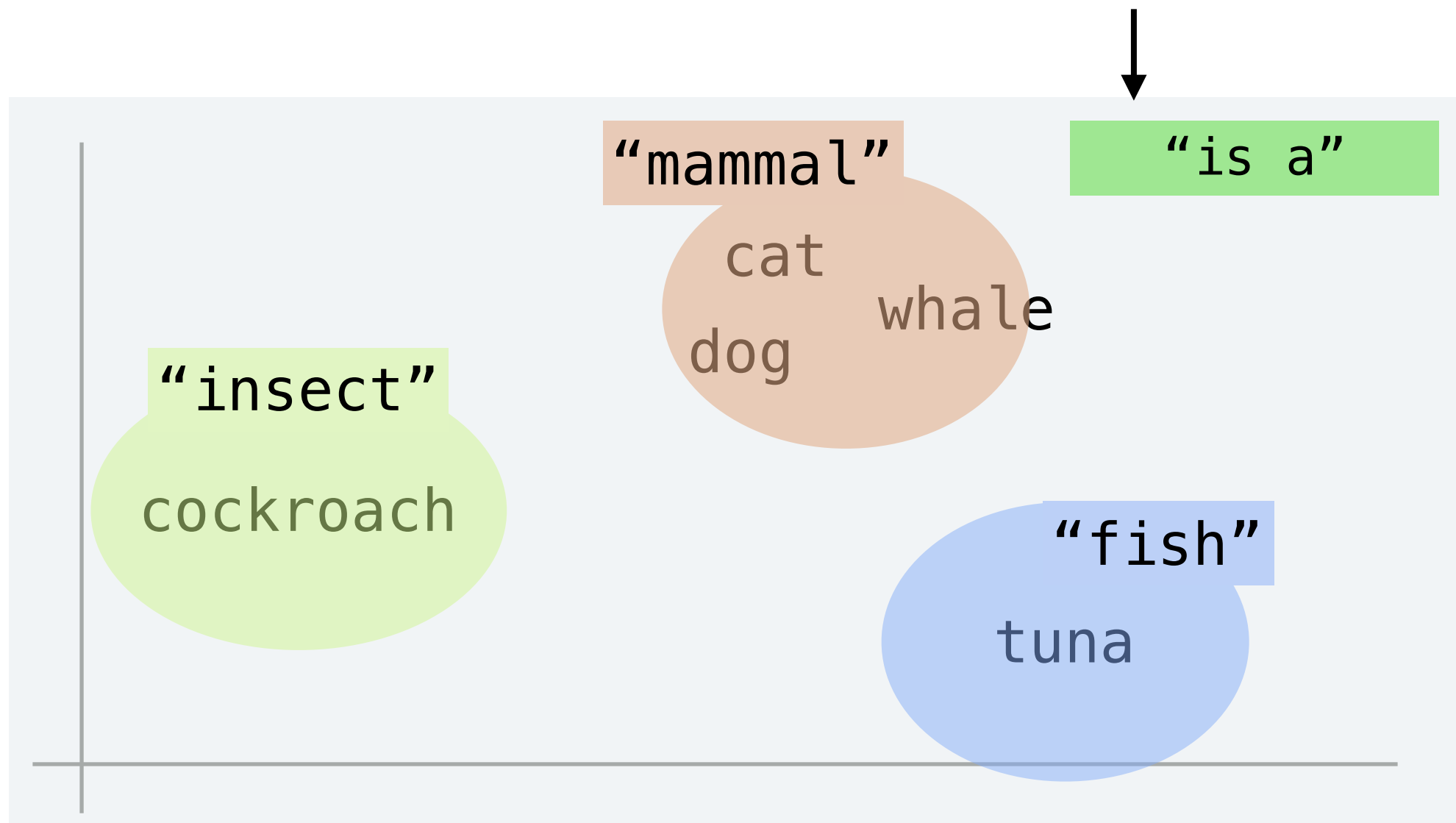
intuition: closeness in the vector space encodes **semantic similarity**, broadly interpreted

obtain different **types** of similarity when considering different **contexts**



intuition: closeness in the vector space encodes **semantic similarity**, broadly interpreted

obtain different **types** of similarity when considering different **contexts**



relationships as transformations

For generality, we use **affine transformations**,
allowing for:

- rotations
- reflections
- shear
- squeeze
- translations

linear transformation

$$G_R = \begin{bmatrix} A_R & \mathbf{b}_R \\ 0 & 1 \end{bmatrix}$$

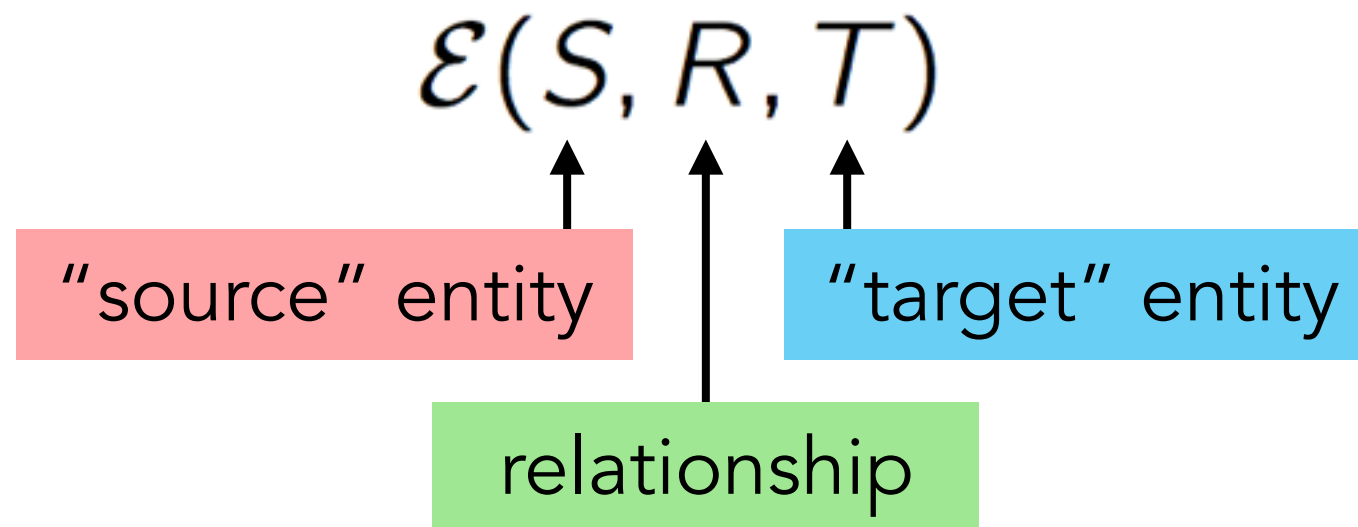
representation of relationship R

representation of entity S

$$G_R \mathbf{c}_S = \begin{bmatrix} A_R & \mathbf{b}_R \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{c}}_S \\ 1 \end{bmatrix} = \begin{bmatrix} A_R \tilde{\mathbf{c}}_S + \mathbf{b}_R \\ 1 \end{bmatrix}$$

enforcing similarity

Define an **energy function**



Energy is **low** if 'S is related to T through R' is **true**

note: R is often not symmetric, e.g. 'is a'

enforcing similarity

Energy in terms of entity and relationship representations:

$$\mathcal{E}(S, R, T|\Theta) = -\mathbf{v}_T \cdot G_R \mathbf{c}_S$$

... susceptible to maximising norms to minimise energy, so...

$$\mathcal{E}(S, R, T|\Theta) = -\frac{\mathbf{v}_T \cdot G_R \mathbf{c}_S}{\|\mathbf{v}_T\| \|G_R \mathbf{c}_S\|}$$

probabilistic model

Given energy, obtain **probability distribution**

$$P(S, R, T|\Theta) = \frac{e^{-\mathcal{E}(S,R,T|\Theta)}}{\sum_{s,r,t} e^{-\mathcal{E}(s,r,t|\Theta)}}$$

\Leftrightarrow low energy triples have high probability

Also get **conditional distributions**, e.g.:

$$P(S|R, T; \Theta) = \frac{e^{-\mathcal{E}(S,R,T|\Theta)}}{\sum_s e^{-\mathcal{E}(s,R,T|\Theta)}}$$


training

Parameters learned using stochastic maximum likelihood

$$\Theta^* = \operatorname{argmax} \sum_n^N \log P((S, R, T)_n | \Theta)$$

Exact gradients are computationally intractable due to partition function term, but reduce to expression of the form:

$$\frac{\partial \log P(\mathcal{D} | \Theta)}{\partial \Theta_i} = \mathbb{E}_{P_D(S, R, T)} \left[\frac{-\partial \mathcal{E}(S, R, T | \Theta)}{\partial \Theta_i} \right] - \mathbb{E}_{P_M(S, R, T)} \left[\frac{-\partial \mathcal{E}(S, R, T | \Theta)}{\partial \Theta_i} \right]$$



data distribution model distribution

persistent contrastive divergence

$$\frac{\partial \log P(\mathcal{D}|\Theta)}{\partial \Theta_i} = \mathbb{E}_{P_D(S,R,T)} \left[\frac{-\partial \mathcal{E}(S,R,T|\Theta)}{\partial \Theta_i} \right] - \mathbb{E}_{P_M(S,R,T)} \left[\frac{-\partial \mathcal{E}(S,R,T|\Theta)}{\partial \Theta_i} \right]$$

↑
data distribution ↑
model distribution

Obtain samples from model distribution with **Gibbs sampling** on conditional distributions

$$P(S|R,T;\Theta) \quad P(R|S,T;\Theta) \quad P(T|S,R;\Theta)$$

Run independent Markov chains, **retain** between batches

experiments

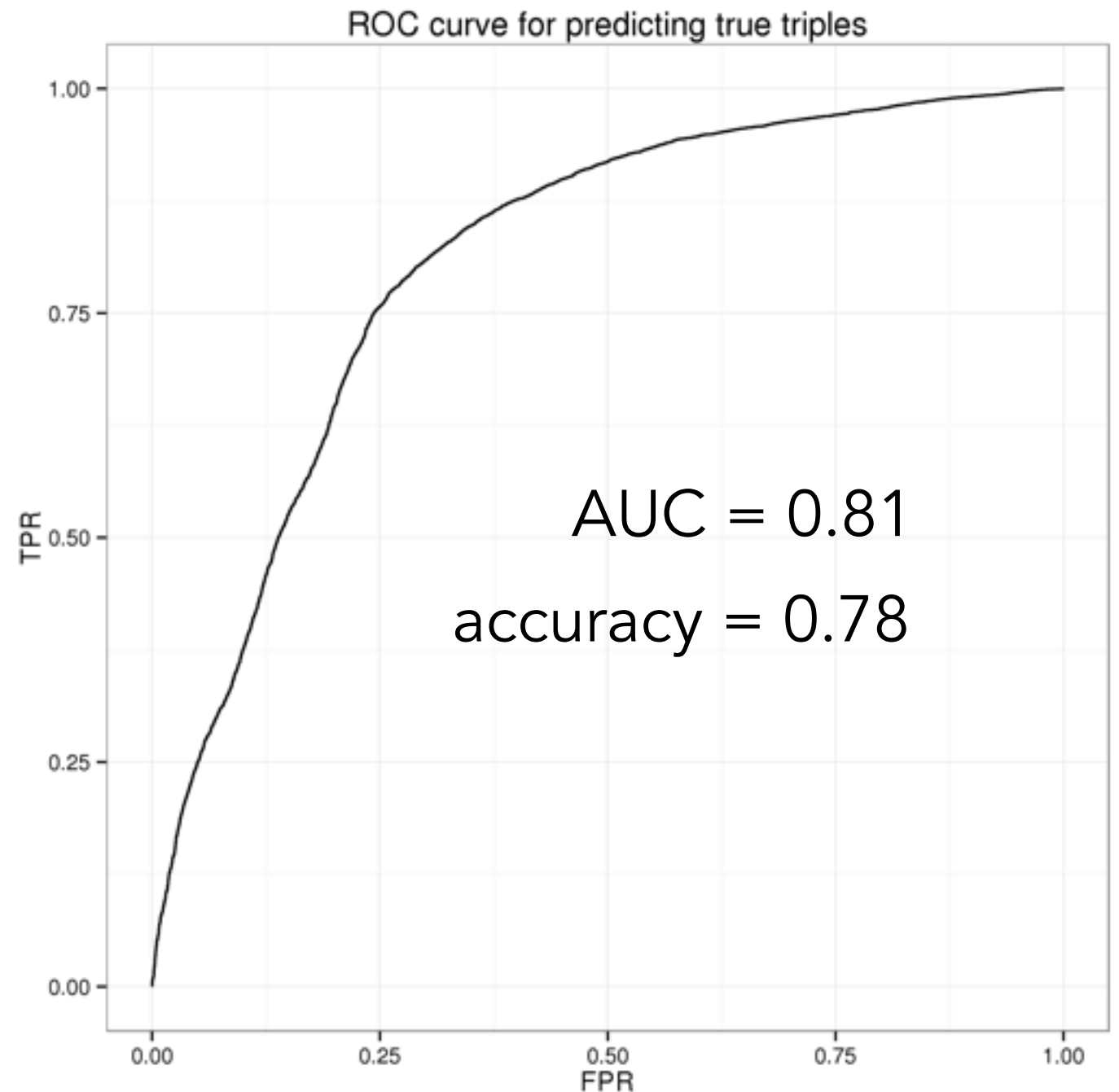
WordNet task: predict whether a triple (S, R, T) is 'true' by looking at $P(R \mid S, T)$

example:

is **rye** a **type of** **whiskey?**



is **liebfraumilch** an **instance of** **dry land?**



Socher et al.(2013) report accuracy of 0.74

Socher et al., "Reasoning with Neural Tensor Networks for Knowledge Base Completion", NIPS 2013

experiments - semi-supervised

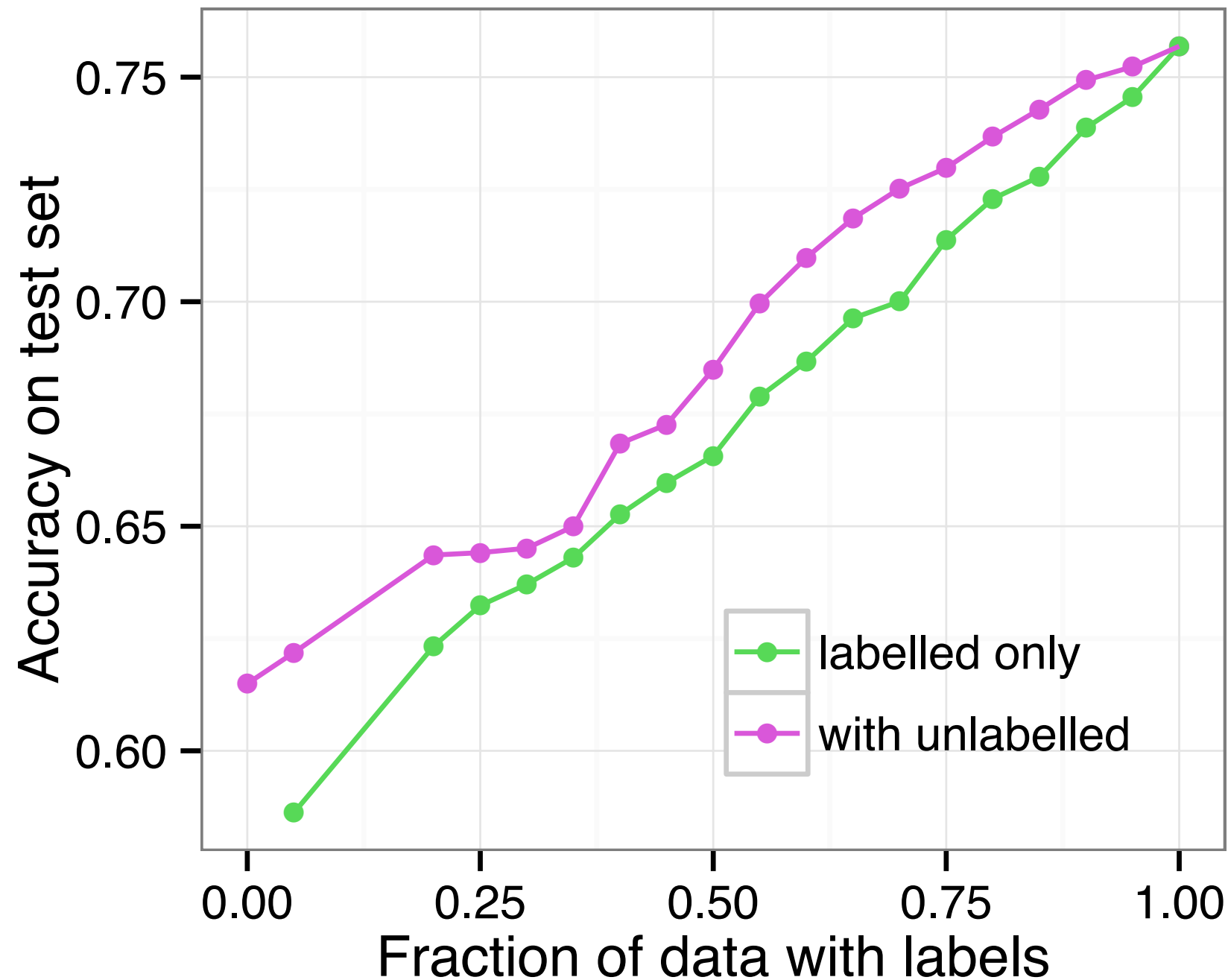
Since we have a joint model, we can integrate over unseen labels

$$P(S, T) = \sum_r P(S, r, T)$$

We can do semi-supervised training, forming gradient updates as weighted averages of the unseen label

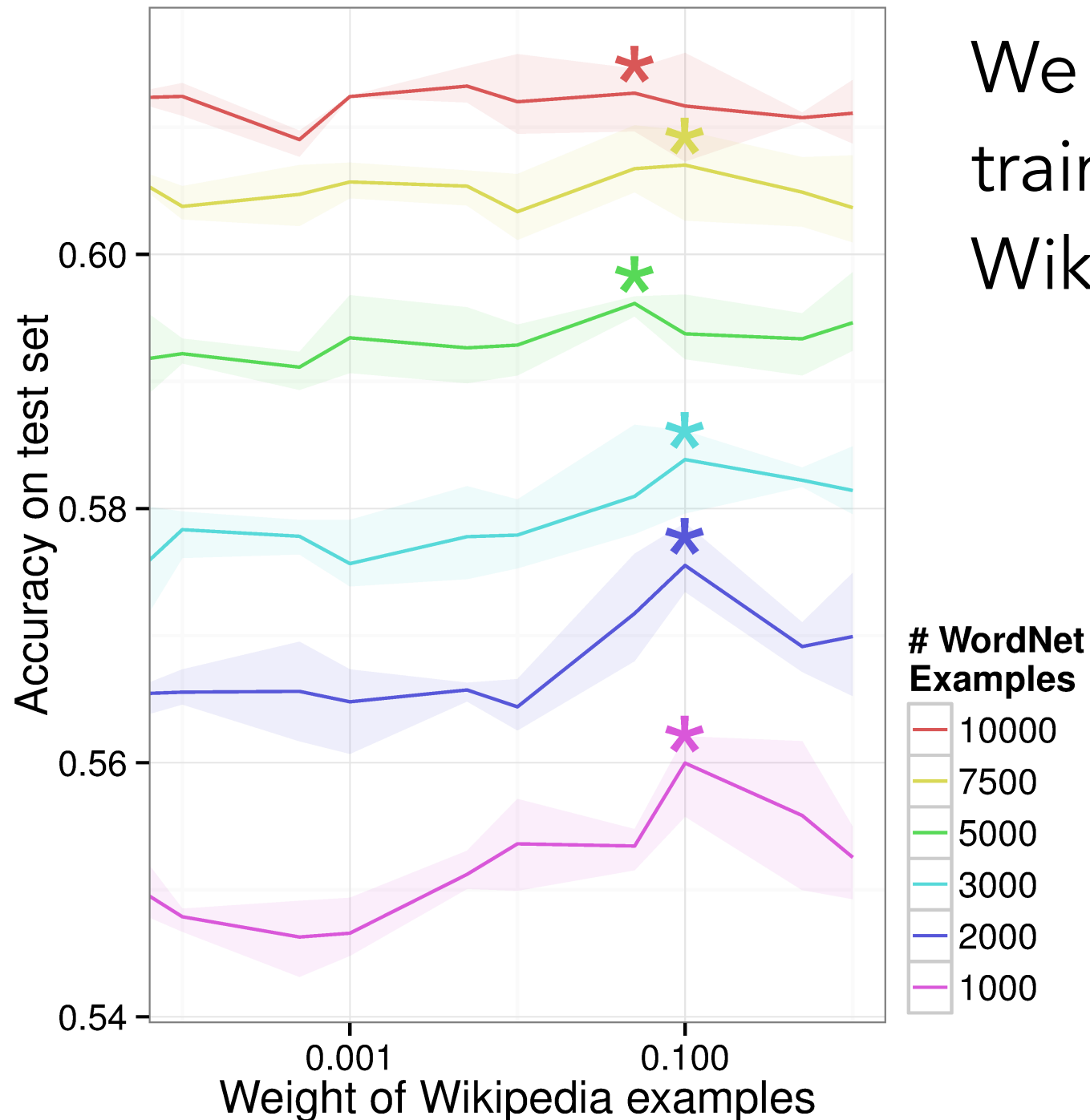
Given a fixed amount of labelled data, we see if adding **unlabelled** data helps the prediction task

experiments - semi-supervised

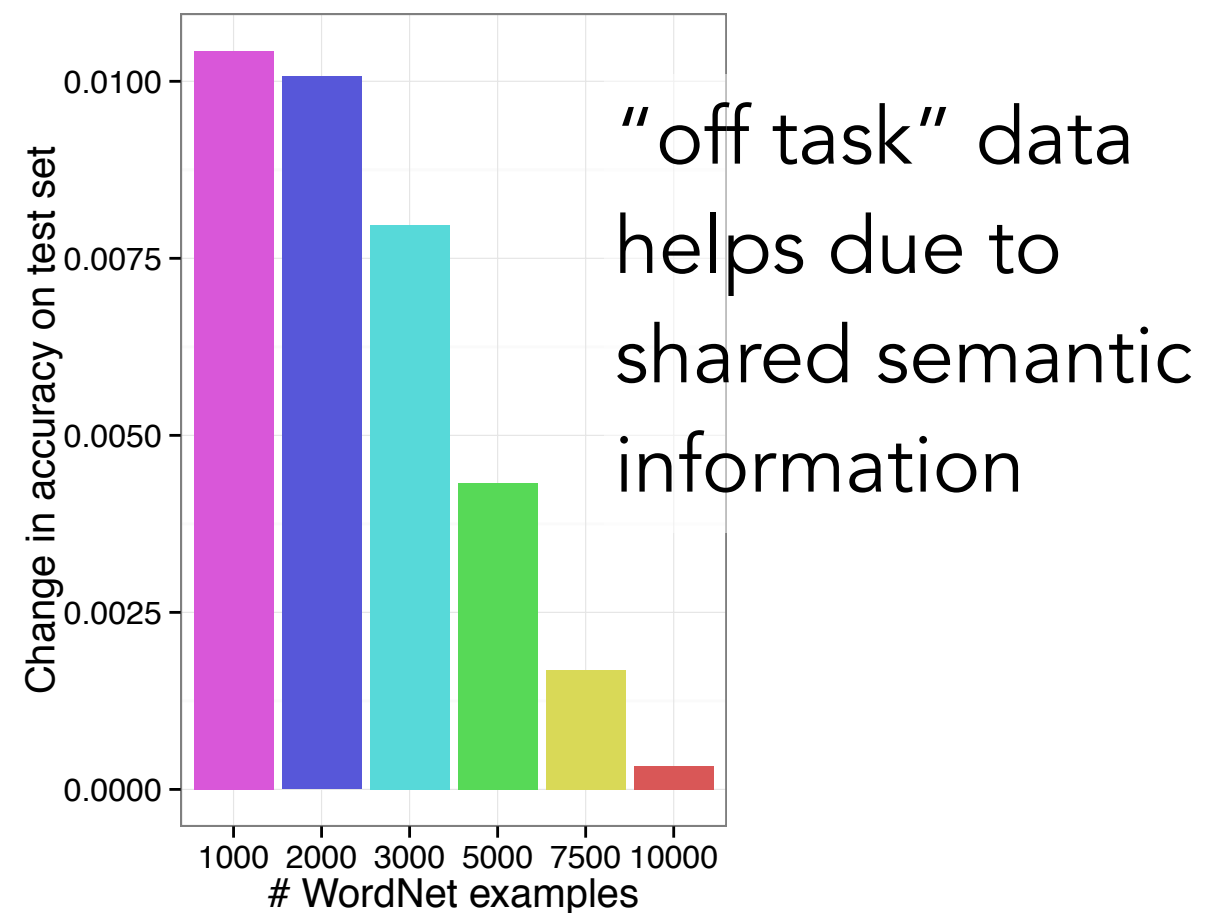


Adding examples with **no** labelled relationship can improve predictive performance

experiments



We can also add **unstructured** training examples (here from Wikipedia)



experiments

We can ask the model to **find** relationships in **free** data by giving it 'unlabelled' examples from Wikipedia and telling it how many relationships we think there are

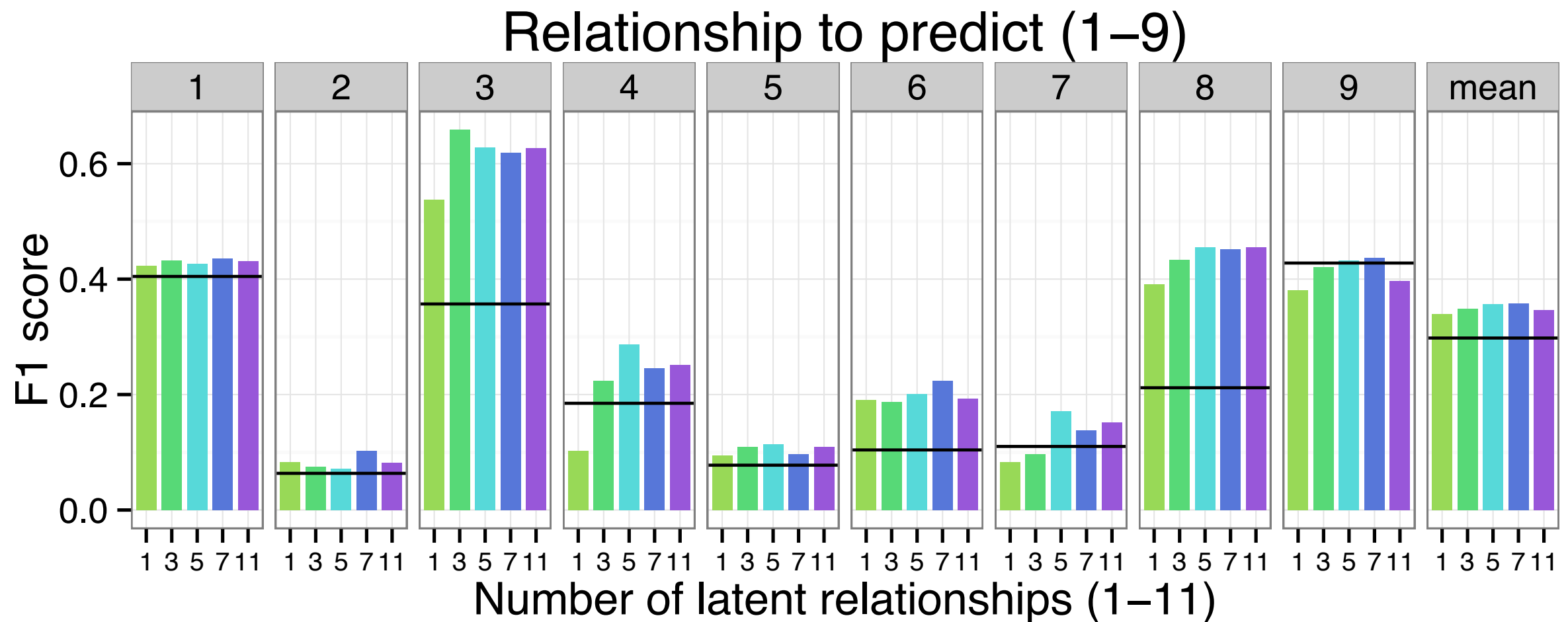
Then evaluate the raw **word embeddings** to see if the presence of latent relationships allows them to capture more semantic information

...we do this by using them as inputs to multi-class classifier (random forest) predicting the **WordNet** relationships

experiments

Reporting F1 score for varying number of latent relationship...

This is preliminary, but more latent relationships **possibly** help!



future work

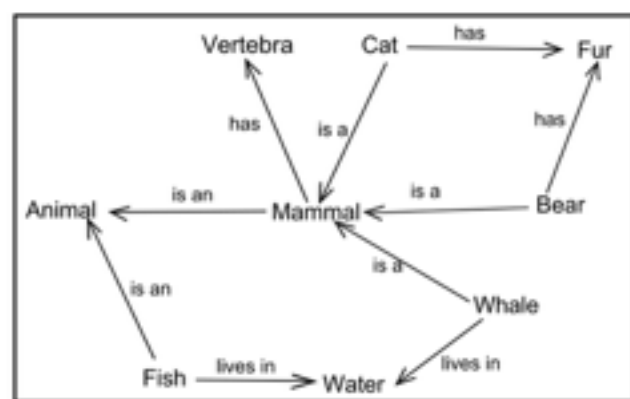
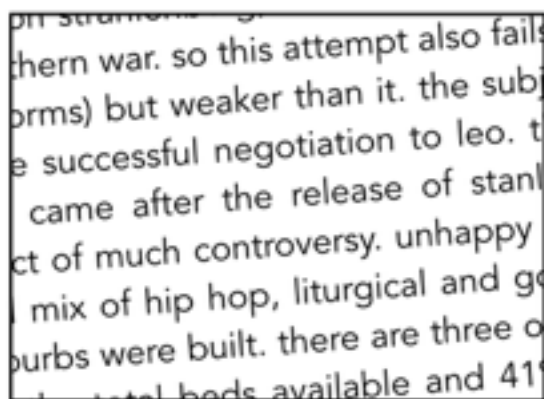
- Analysis of learned latent relationships
- Relationship representation: are affine transformations enough?
- Energy function: cosine distance is not sensitive to small deviations between vectors
- GPU implementation using Theano

summary

I have presented a
**probabilistic
generative model**

$$P(S, R, T|\Theta) = \frac{e^{-\mathcal{E}(S,R,T|\Theta)}}{\sum_{s,r,t} e^{-\mathcal{E}(s,r,t|\Theta)}}$$

for combining **structured** and **unstructured** data,
potentially only partially labelled



... to learn representations
of word entities as **vectors**
and relationships as **affine
transformations**



code and data here:

<https://github.com/corcra/bf2>

Collaborators:

Gunnar Rätsch and Theofanis Karaletsos

Acknowledgements:

The rest of the Rätschlab at MSKCC

Tri-Institutional Training Program in Computational Biology and Medicine

thanks for listening!

paper: <http://arxiv.org/abs/1510.00259>



code and data here:

<https://github.com/corcra/bf2>