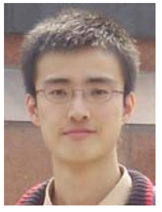
Automated Feature Generation from Structured Knowledge



Weiwei Cheng

Gjergji Kasneci

Thore Graepel

David Stern

Ralf Herbrich

Microsoft Research Cambridge

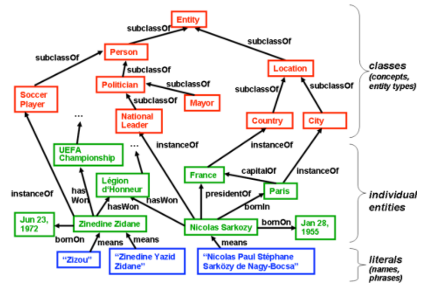
Most machine learning researches focus on the modeling, while the construction of features is as crucial.

Can we design a mechanism that compactly describes and extracts relevant features?

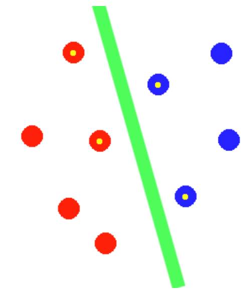
learning task

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY
sunny	85	85	FALSE
sunny	80	90	TRUE
overcast	83	78	FALSE
rain	70	96	FALSE
rain	68	80	FALSE
rain	65	70	TRUE
overcast	64	65	TRUE
sunny	72	95	FALSE
sunny	69	70	FALSE
rain	75	80	FALSE
sunny	75	70	TRUE
overcast	72	90	TRUE
overcast	81	75	FALSE
rain	71	80	TRUE

knowledge base



learner



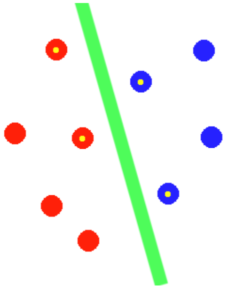
In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.

learning task

knowledge base

learner



In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.

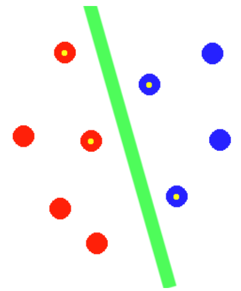
learning task

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY
sunny	85	85	FALSE
sunny	80	90	TRUE
overcast	83	78	FALSE
rain	70	96	FALSE
rain	68	80	FALSE
rain	65	70	TRUE
overcast	64	65	TRUE
sunny	72	95	FALSE
sunny	69	70	FALSE
rain	75	80	FALSE
sunny	75	70	TRUE
overcast	72	90	TRUE
overcast	81	75	FALSE
rain	71	80	TRUE

knowledge base

ESPN

learner



In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.

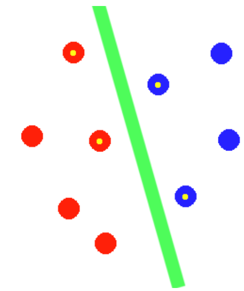
learning task

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY
sunny	85	85	FALSE
sunny	80	90	TRUE
overcast	83	78	FALSE
rain	70	96	FALSE
rain	68	80	FALSE
rain	65	70	TRUE
overcast	64	65	TRUE
sunny	72	95	FALSE
sunny	69	70	FALSE
rain	75	80	FALSE
sunny	75	70	TRUE
overcast	72	90	TRUE
overcast	81	75	FALSE
rain	71	80	TRUE

knowledge base

msdn
Microsoft® Developer Network

learner



In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.

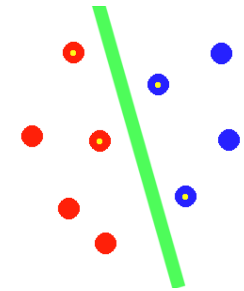
learning task

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY
sunny	85	85	FALSE
sunny	80	90	TRUE
overcast	83	78	FALSE
rain	70	96	FALSE
rain	68	80	FALSE
rain	65	70	TRUE
overcast	64	65	TRUE
sunny	72	95	FALSE
sunny	69	70	FALSE
rain	75	80	FALSE
sunny	75	70	TRUE
overcast	72	90	TRUE
overcast	81	75	FALSE
rain	71	80	TRUE

knowledge base

YAGO

learner



In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.

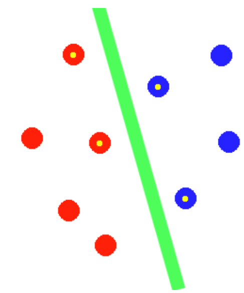
learning task

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY
sunny	85	85	FALSE
sunny	80	90	TRUE
overcast	83	78	FALSE
rain	70	96	FALSE
rain	68	80	FALSE
rain	65	70	TRUE
overcast	64	65	TRUE
sunny	72	95	FALSE
sunny	69	70	FALSE
rain	75	80	FALSE
sunny	75	70	TRUE
overcast	72	90	TRUE
overcast	81	75	FALSE
rain	71	80	TRUE

knowledge base

YAGO

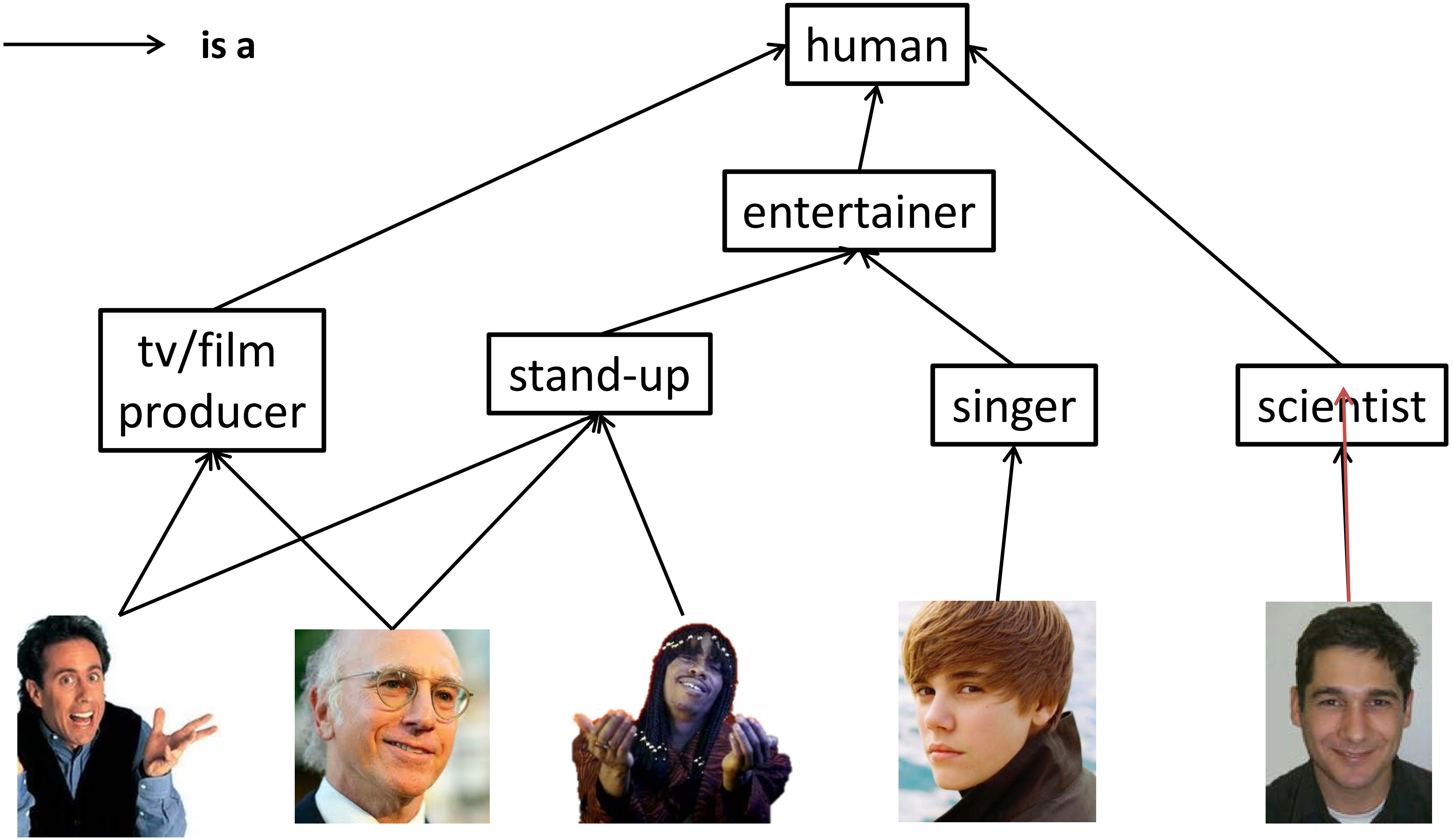
learner



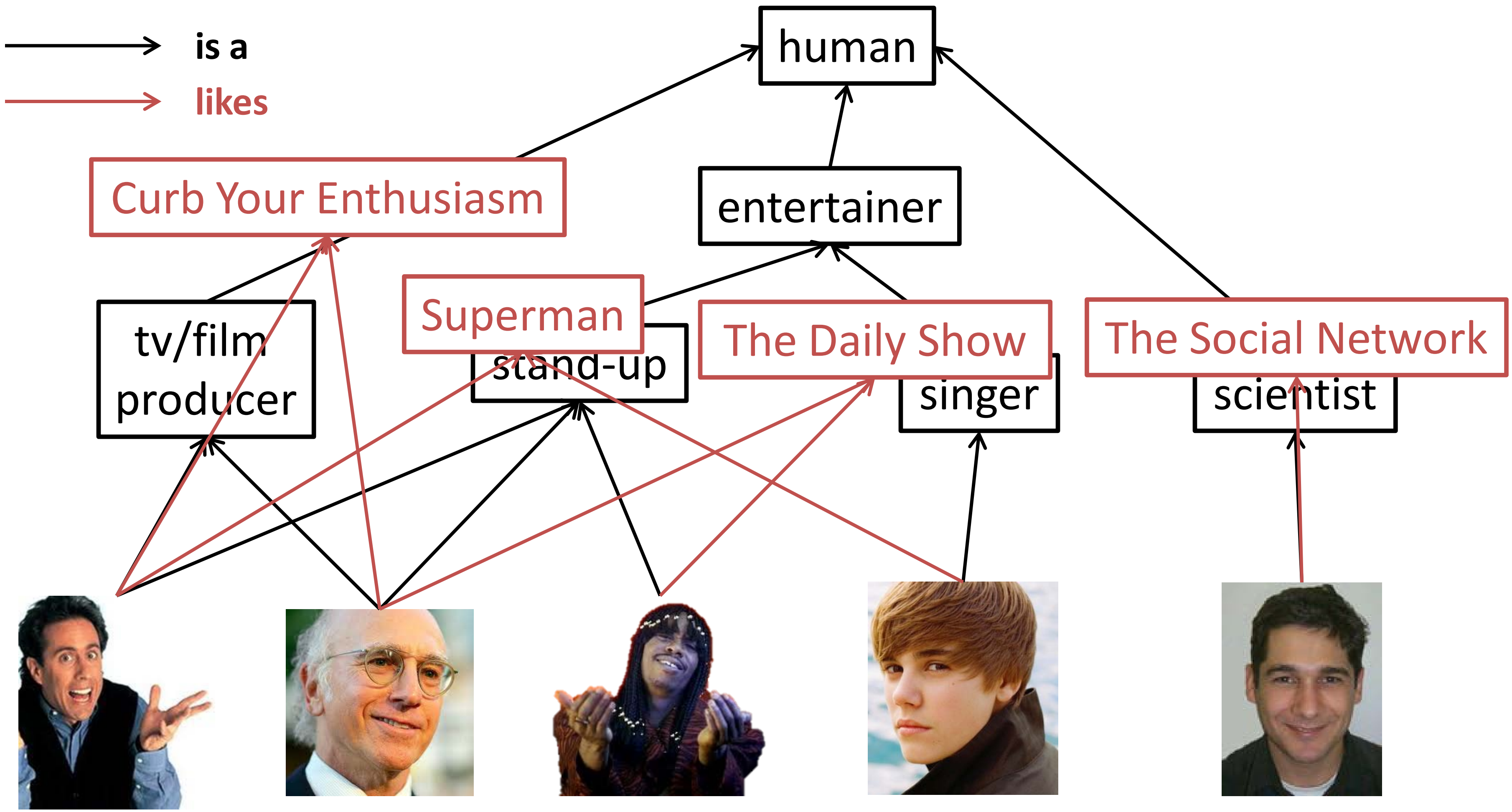
Modularity!

In our work, we introduce

1. a theoretical framework for constructing semantic features from a given knowledge base;
2. various strategies for incorporating these features into a prediction model.



isa	producer	stand-up	singer	entertainer	scientist	human
	1	1	0	1	0	1
	0	0	1	1	0	1



isa	proc	*likes*	curb	superman	network	daily		human
			1	1	0	1		1
	(		0	1	0	0		1

To query semantic features, we propose

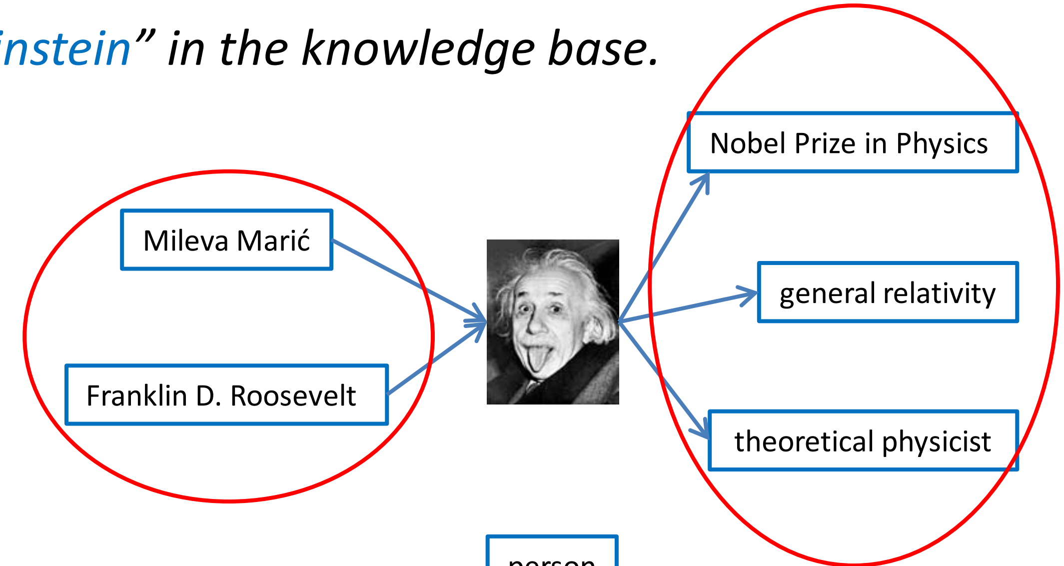
ESPARQL

(Extended SPARQL Query Language), an extension of SPARQL with NAGA features.

Exemplary Queries

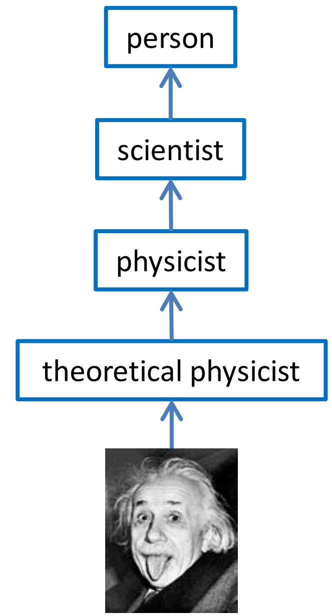
Retrieve all neighbors of "Albert Einstein" in the knowledge base.

```
select ?x
where {
    (albert_einstein  ?r  ?x)  union
    (?x  ?r  albert_einstein)
}
```



Retrieve all classes "Albert Einstein" belongs to.

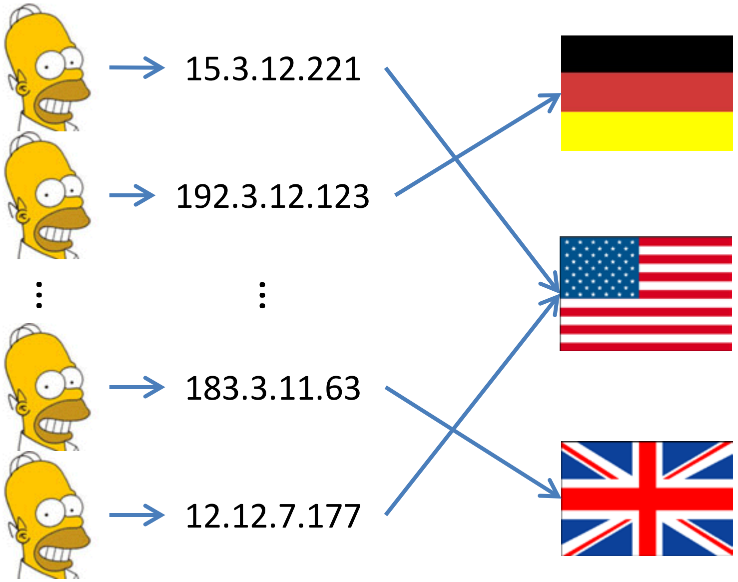
```
select ?x
where {
    albert_einstein  type subClass*  ?x
}
```



Exemplary Queries

Find country information based on the user's IP address.

```
select ?z
where {
        ?y  hasIPAddress  userIP .
        userIP  belongsToLocation  ?z .
        ?z  type  subClass*  country
}
```



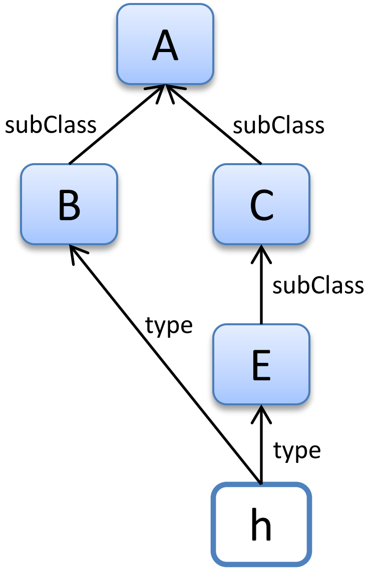
Count the followers of a particular user in Twitter.

```
select count(?x)
where {
    ?x  follows  user
}
```

Constructing Feature Vectors

select ?c

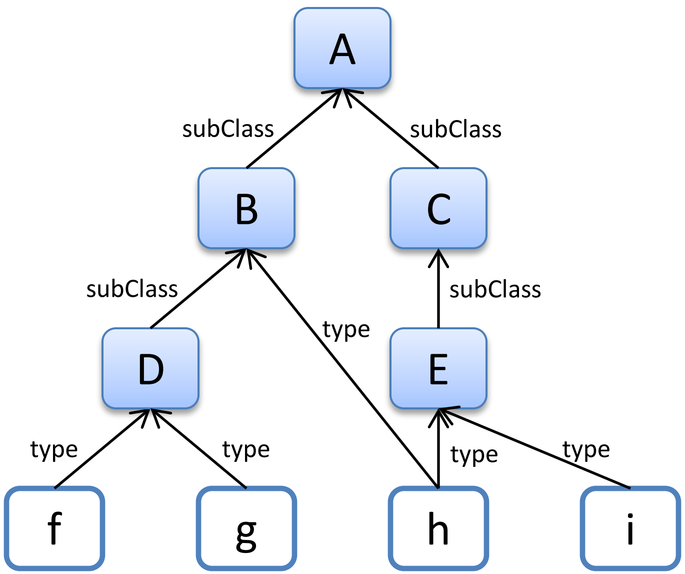
where { h type subClass* ?c }



1. Based on the training corpus and the knowledge base, choose appropriate ESPARQL queries to extract semantic information about the entities in the training corpus.

Constructing Feature Vectors

```
select ?c
where {  h  type subClass*  ?c  }
```

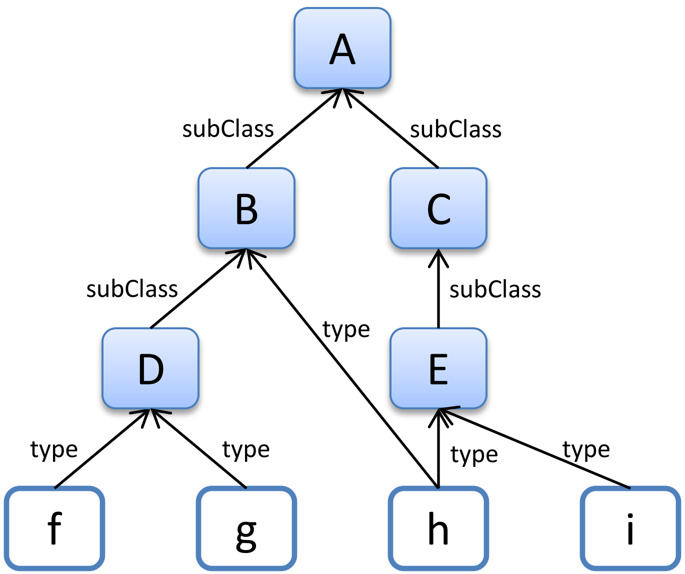


A B C D E

2. Unify the answer sets of the queries to construct the set of semantic features that indexes the dimensions of the effective feature space.

Constructing Feature Vectors

```
select ?c
where {  h  type subClass*  ?c  }
```

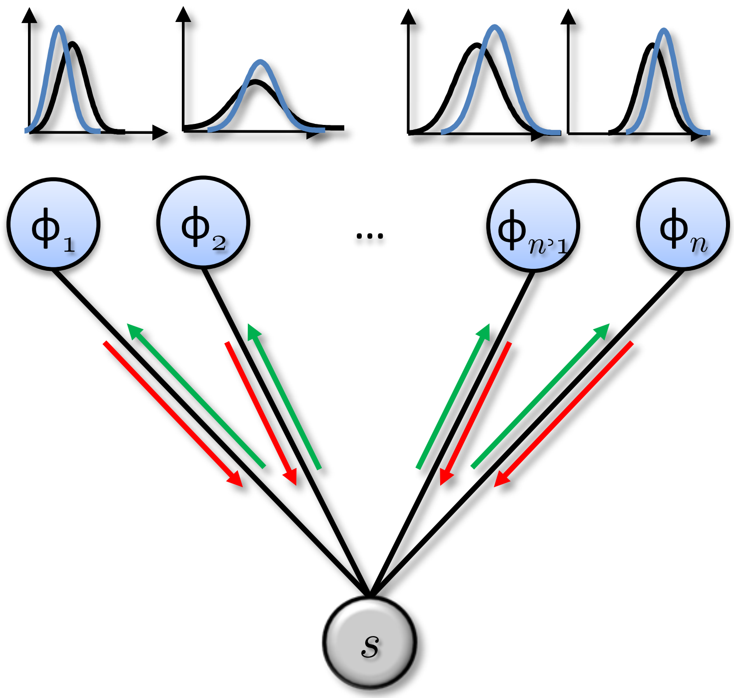


	A	B	C	D	E
$\phi(h) =$	1	1	1	0	1
$\phi(g) =$	1	1	0	1	0

3. The feature vector of an entity, which has the dimensionality of the feature space, can be built by setting those dimensions which correspond to query answers for the entitiy to 1.

Generalized Linear Bayesian Probit Model

- Given the binary feature vector $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_n(x))$, the importance of features is represented by a weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)$, where the belief on w_i is given by a Gaussian distribution with mean μ_i and σ_i^2. We have
$$p(\boldsymbol{w}) = \prod_i \mathcal{N}(w_i;\, \mu_i, \sigma_i^2).$$
- The score of instance (i.e., entity) x is $s(x) = \sum_{i=1}^n w_i \phi_i(x)$, and we have
$$p(s \,|\, \boldsymbol{w}, x) = \mathcal{N}\left(s;\, \sum_{i=1}^n \phi_i(x)\mu_i, \sum_{i=1}^n \phi_i(x)\sigma_i^2\right).$$
- The output y is modeled as a probit function
$$P(y \,|\, s) = \Phi\left(\frac{ys}{\beta}\right)$$
with noise variance β^2, where $\Phi(t) = \int_{-\infty}^t \mathcal{N}(v;\, 0, 1)\mathrm{d}v$.


ϕ_1
ϕ_2
...
ϕ_{n-1}
ϕ_n
s

Classification of Tweets

Dataset:

Twitter messages were categorized by Amazon Mechanical Turk and then manually verified by researchers in our lab.

Business / Finance	2,182	9.56%
Entertainment	4,235	18.56%
Lifestyle	4,085	17.90%
Politics	1,199	5.25%
Science / Environment	789	3.45%
Sport	1,145	5.01%
Technology	1,880	8.23%
World Events	2,122	9.30%
Other / Miscellaneous	12,838	56.26%
22,816 tweets in total		

Classification of Tweets

Ballmer on iPad: “they've sold certainly more than I'd like them to have sold”

Technology

Obama blames Bush for all of his misdeeds and then takes credit for the successful war in Iraq http://is.gd/e0iVM (via @PennyStarrDC)

Politics   *World Events*

Classific

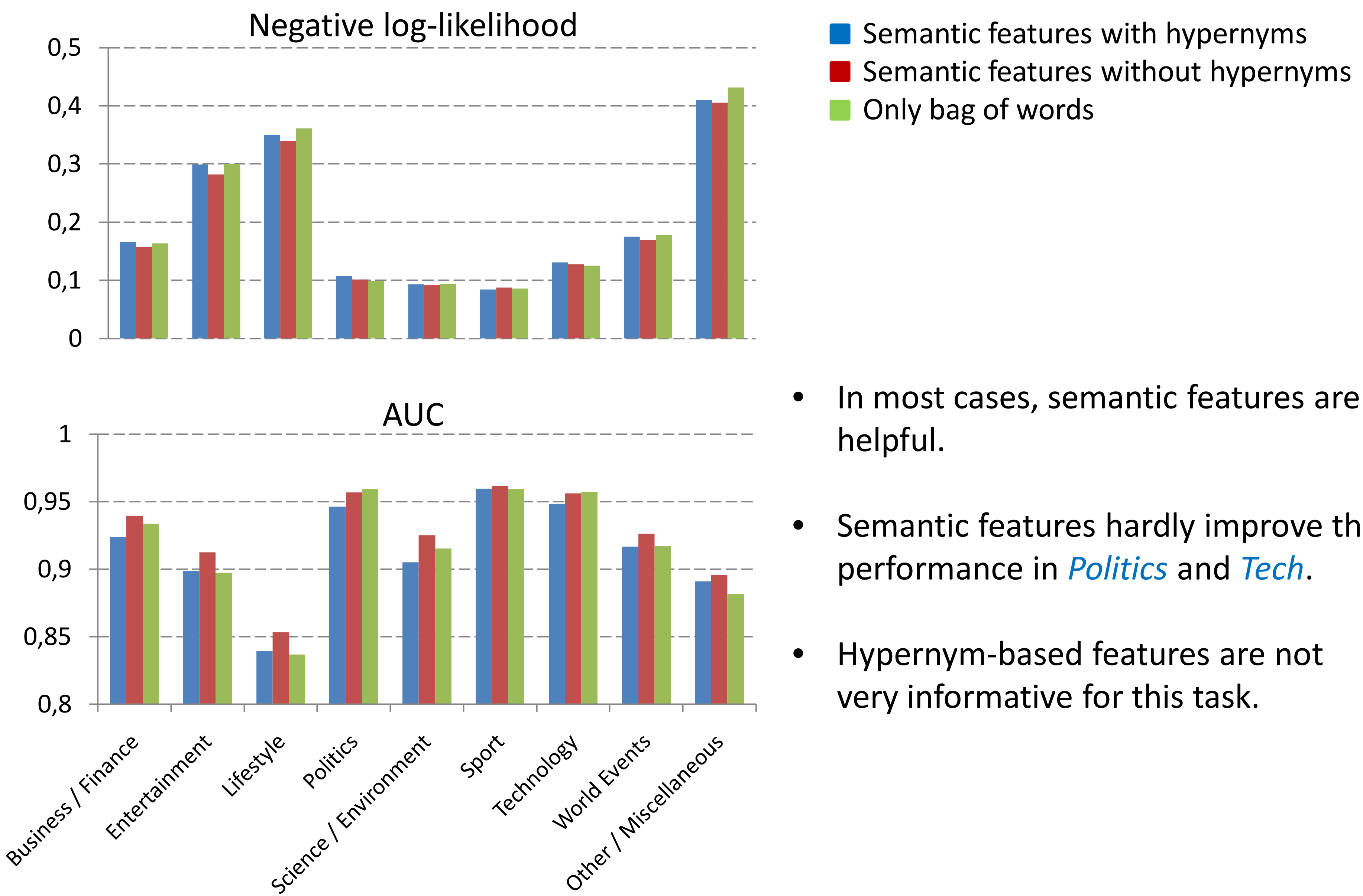
Ballmer on iPad: “they'
them to have sold”

Obama blames Bush fo
credit for the successfu
@PennyStarrDC)

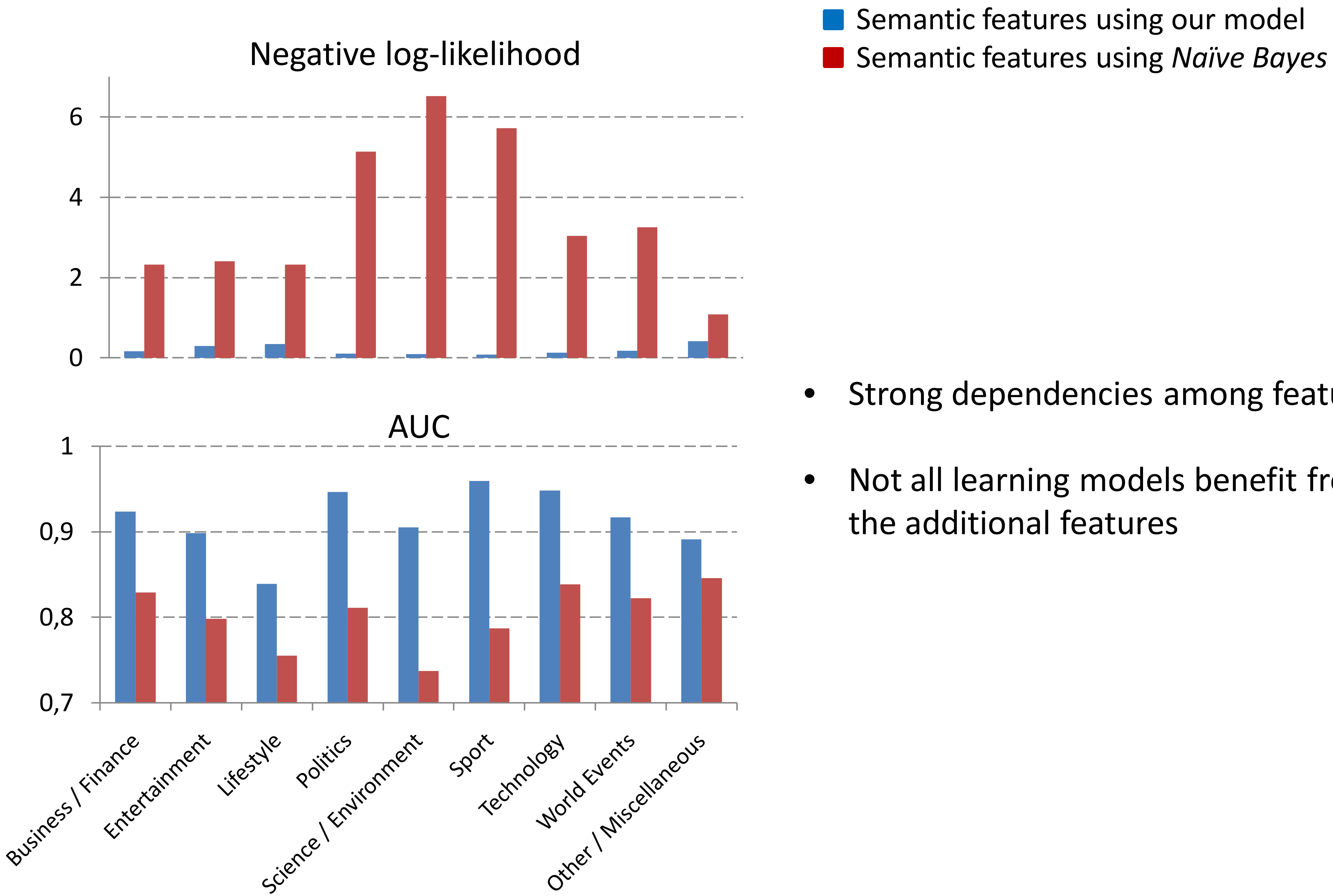
...
wikicategory_Governors_of_Texas
wikicategory_Harvard_Business_School_alumni
wikicategory_Harvard_University_alumni
wikicategory_Phillips_Academy_alumni
wikicategory_Presidents_of_the_United_States
wikicategory_Texas_Republicans
wikicategory_Time_magazine_Persons_of_the_Year
wikicategory_United_States_Air_Force_officers
wikicategory_Yale_University_alumni
wordnet_businessperson_109882716
wordnet_president_110468559
wordnet_person_100007846
wordnet_physical_entity_100001930
wordnet_politician_110451263
wordnet_republican_110522495
wordnet_scholar_110557854
wordnet_serviceman_110582746
wordnet_skilled_worker_110605985
wordnet_worker_109632518
wordnet_yagoActor_0
wordnet_yagoActorGeo_1
wordnet_capitalist_109609232
wordnet_causal_agent_100007347
wordnet_convert_109962414
wordnet_corporate_executive_109966255
wordnet_executive_110069645
wordnet_governor_110140314
wordnet_head_110162991
wordnet_intellectual_109621545
wordnet_leader_109623038
wordnet_military_officer_110317007
wordnet_administrator_109770949
wordnet_alumnus_109786338
...

first order
semantic features

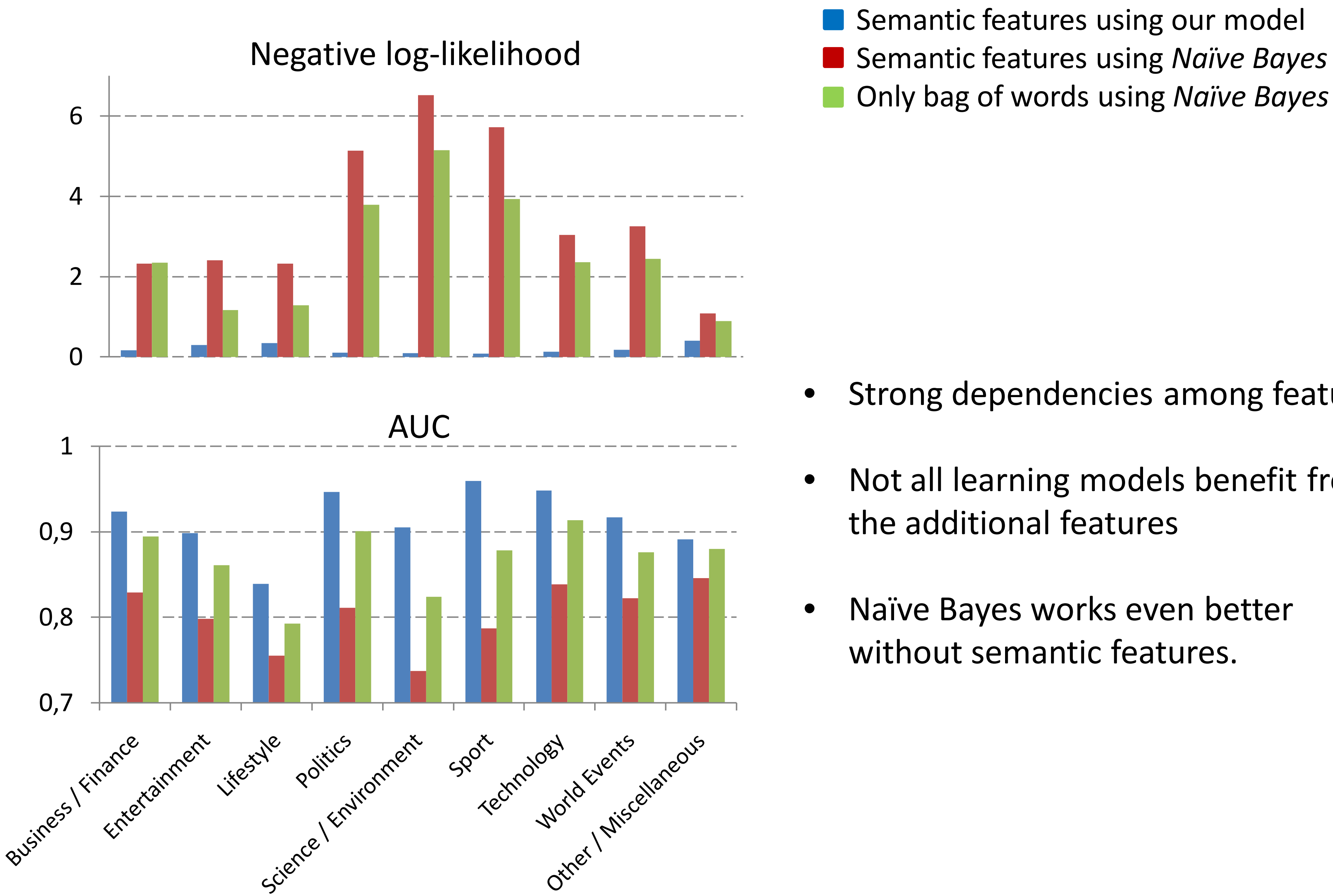
hypernyms / high order
semantic features



- In most cases, semantic features are helpful.
- Semantic features hardly improve the performance in *Politics* and *Tech*.
- Hypernym-based features are not very informative for this task.

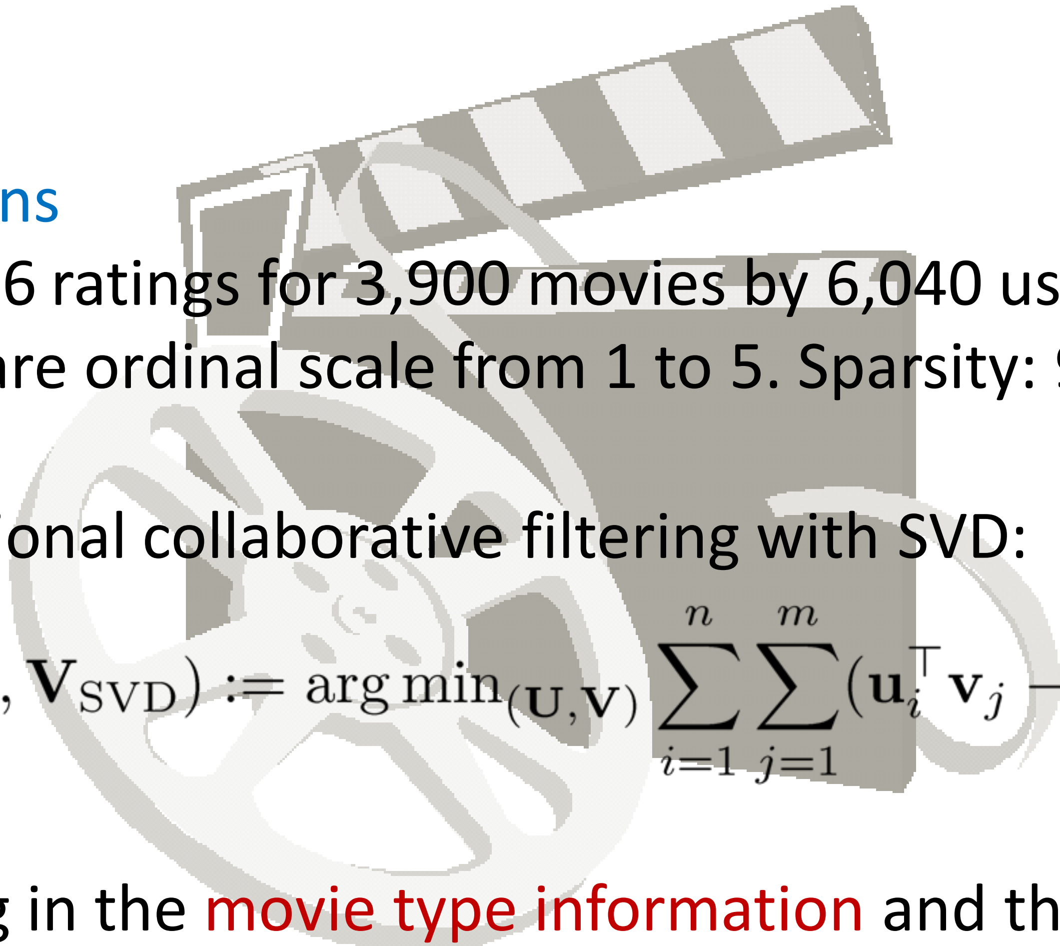


- Strong dependencies among features
- Not all learning models benefit from the additional features



- Strong dependencies among features
- Not all learning models benefit from the additional features
- Naïve Bayes works even better without semantic features.

Movie Recommendation



Dataset:

MovieLens

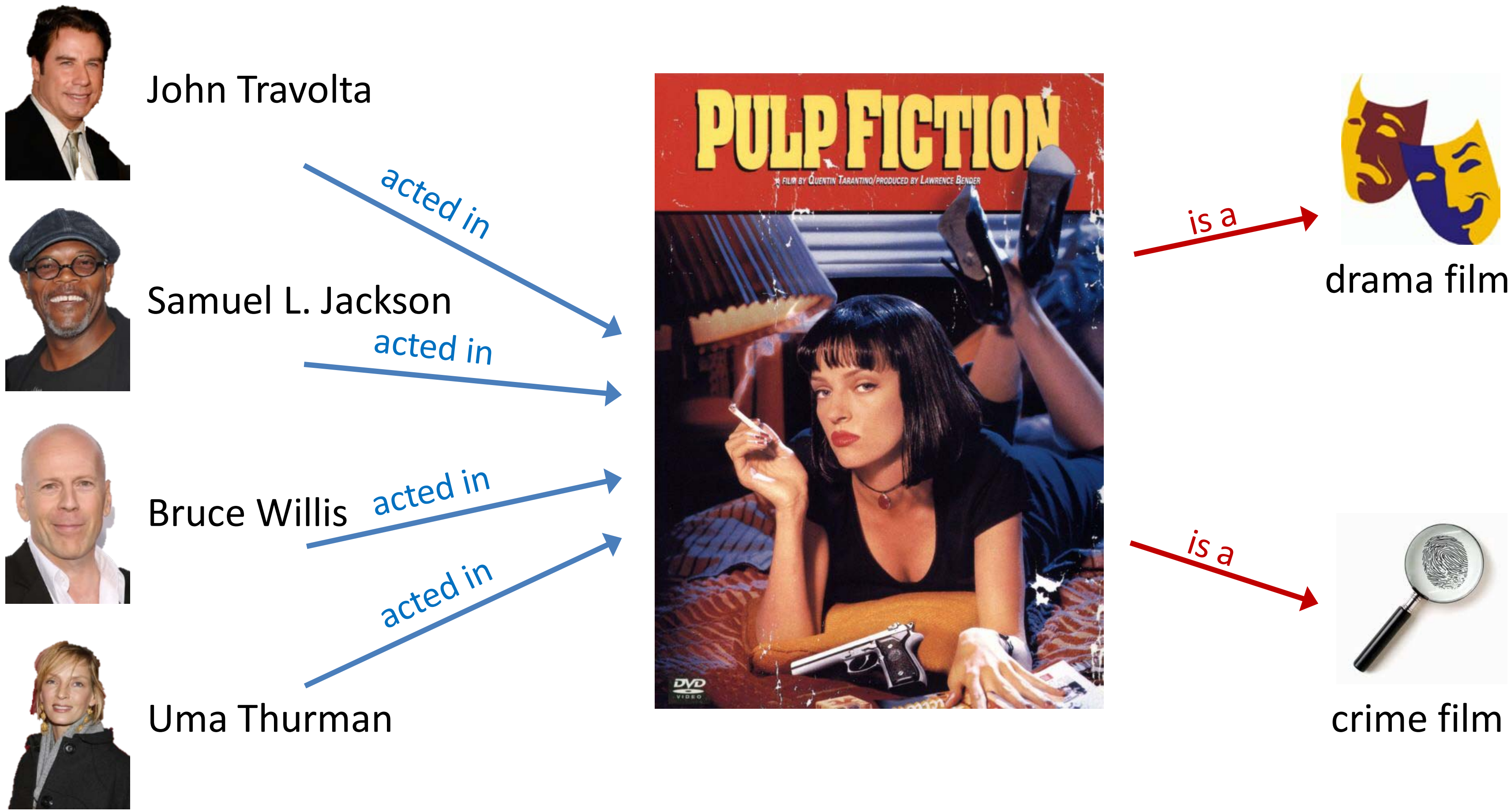
1,000,206 ratings for 3,900 movies by 6,040 users

Ratings are ordinal scale from 1 to 5. Sparsity: 95.7%.

Conventional collaborative filtering with SVD:

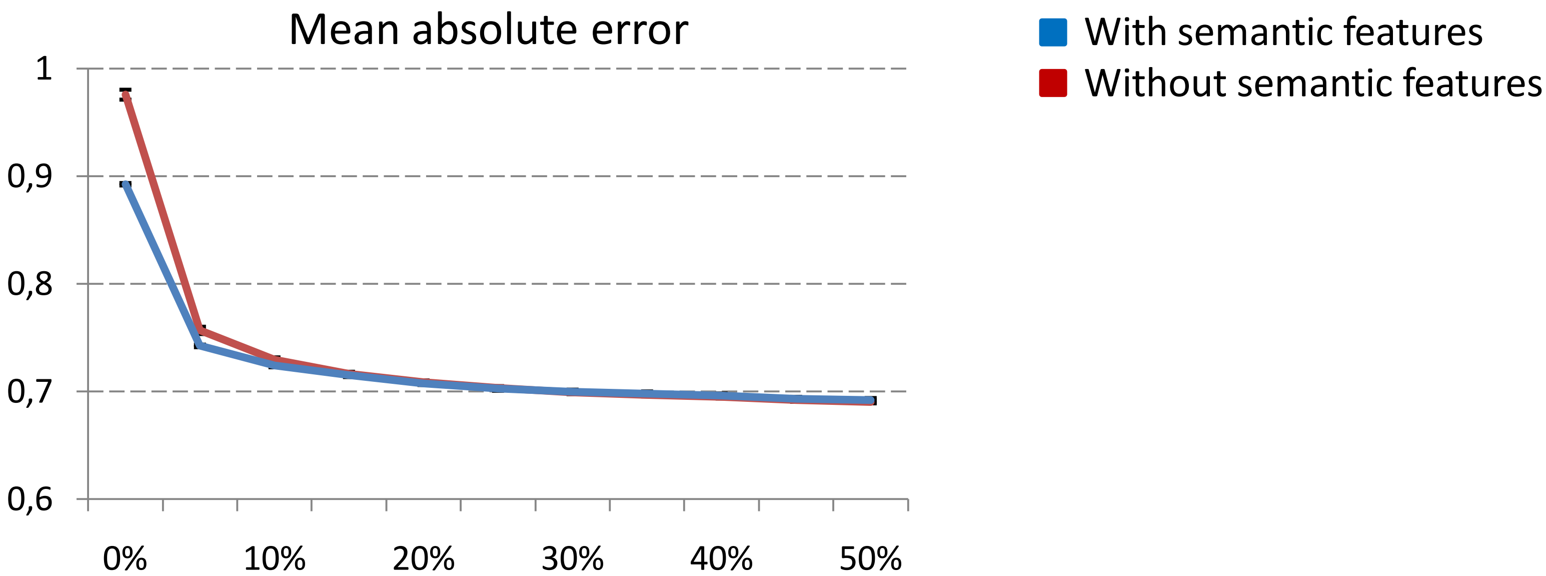
$$(\mathbf{U}_{\mathrm{SVD}}, \mathbf{V}_{\mathrm{SVD}}) := \arg\min_{(\mathbf{U},\mathbf{V})} \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathbf{u}_i^\top \mathbf{v}_j - r_{ij})^2$$

We bring in the movie type information and the actor information with YAGO.



SELECT ?x
WHERE { ?x actedIn pulp_fiction }

SELECT ?x
WHERE { pulp_fiction hasGenre ?x }



- Semantic features are very helpful during cold start, i.e., for movies that have not been rated frequently.

- The effect becomes insignificant when more and more ratings are unveiled.

Summary

- Build the link between semantic technology and machine learning
- Propose framework that compactly describes and extracts relevant features
- Test semantic features in learning (tweet classification, movie recommendation)
- Modularity is a key feature. Choose the knowledge base wisely.