Machine Learning with Knowledge Graphs

Volker Tresp

Siemens Corporate Technology
Ludwig Maximilian University of Munich

Joint work with Maximilian Nickel
With contributions from Xueyan Jiang and Denis Krompass

Prelude

- My background is in Machine Learning and I got involved in Semantic Web projects maybe 6 years ago
- Learning about the Semantic Web clarified my thinking about many things dramatically
- Immediate love affaire with RDF
  - Nothing is ever wrong
  - No contradictions

Prelude

- My background is in Machine Learning and I got involved in Semantic Web projects maybe 6 years ago
- Learning about the Semantic Web clarified my thinking about many things dramatically
- Immediate love affaire with RDF
  - Nothing is ever wrong
  - No contradictions

Overview

- **Why Machine Learning needs Knowledge Graphs**
- Statistical Relational Learning
- Learning with the YAGO Knowledge Graph
- Towards Relevant Use Cases

What is Machine Learning?

Machine Learning versus Statistics versus Data Mining

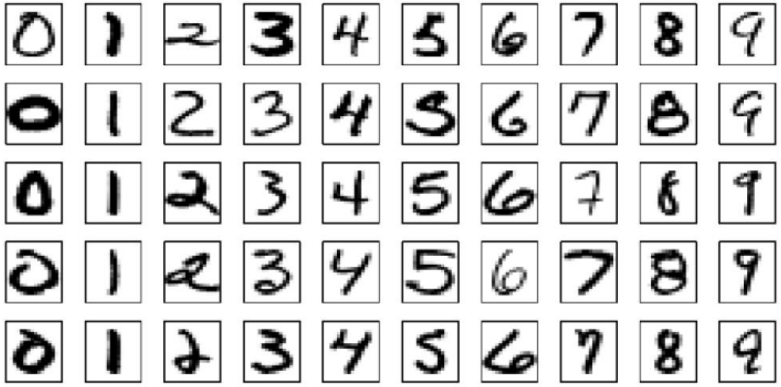
- Statistics focuses on interpretable parameters
- Data mining focuses on the discovery of meaningful patterns
- Machine Learning focuses on prediction accuracy

Classification

Classification is the work horse of machine learning

- Predict class memberships for many objects

- Very powerful
- Surprisingly general



Typical Classifiers

Predicting class *k* for input z_l $\quad P(x^k(z_l)=1) \leftarrow f^k(z_l)$

Fixed basis functions $\quad f^k(z_l) = \sum_{m=1}^{M} w_m^k b_m(z_l)$

Kernels $\quad f^k(z_l) = \sum_{n=1}^{N} v_m^k k(z_l, z_n)$

Really the same things; deep learners would call the shallow

Neural Networks $\quad f^k(z_l) = NN_{deep}(z_l)$

- 10 layers with 1000 neurons per layer
- Currently the hottest thing!

Deep Learning Neural Networks

Scientists See Promise in Deep-Learning Programs

A voice recognition program translated a speech given by Richard F. Rashid, Mi
Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theori
how the brain recognizes patterns, technology companies
reporting startling gains in fields as diverse as computer vi
speech recognition and the identification of promising new
for designing drugs.

点击查看本文中文版。

Connect With Us on Social Media
@nytimesscience on Twitter.

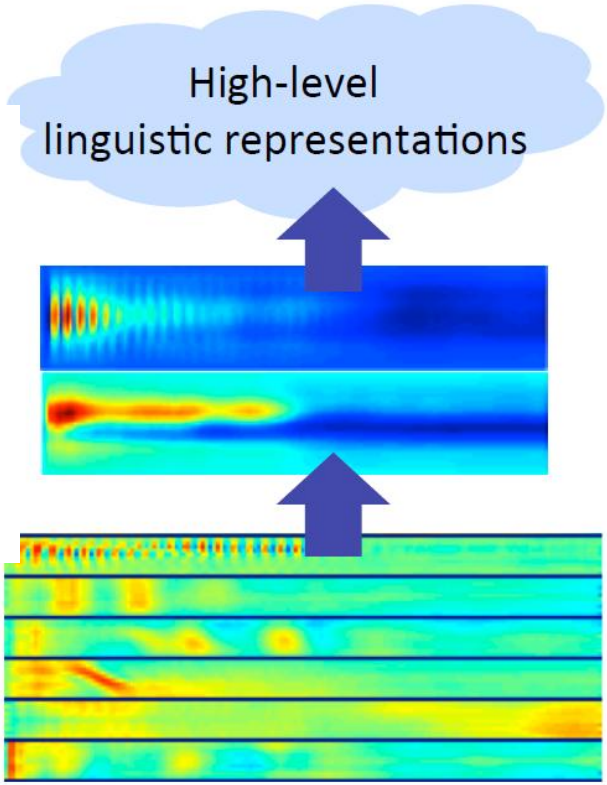
· Science Reporters and Editors on Twitter

Like the science desk on Facebook.

The advances have led to w
enthusiasm among researc
design software to perform human activities like seeing, listening and thinking. They offer the promise of machines that converse with humans and perform tasks like driving cars and working in factories, raising the specter of automated robots that could replace human workers.

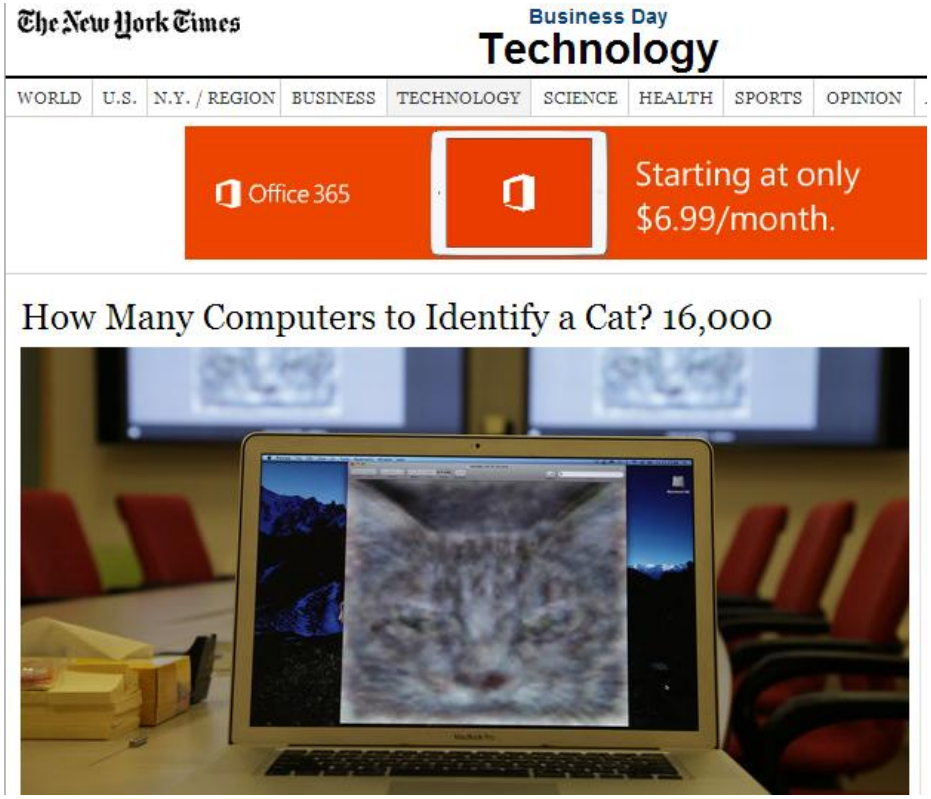
PRINT
SINGLE PAGE
REPRINTS

- Google, Microsoft, Facebook, Baidu are all investing heaviliy in deep learning



Detecting Cats in Images

- Best performing in detecting cats in images and videos (Andrew Ng)



Where from here?

- A deep learning network sees more cats than any child but is not as good at this task
- Deep Learning community: we need better unsupervised learning to pre-structure the network

<Image of cats>

- Maybe we would say: we need background knowledge
- Also: we do not just want to *detect* cats!

Challenges

Predict all classes: „This is a cat!“ „This is a dog!“
„This is a house!“ …

<Image of cats>

Recognize specific entities: „This my cat Max!“

[In our experiments 10^7]

<Image of cats>

Predict all attributes: „Max is evil!“

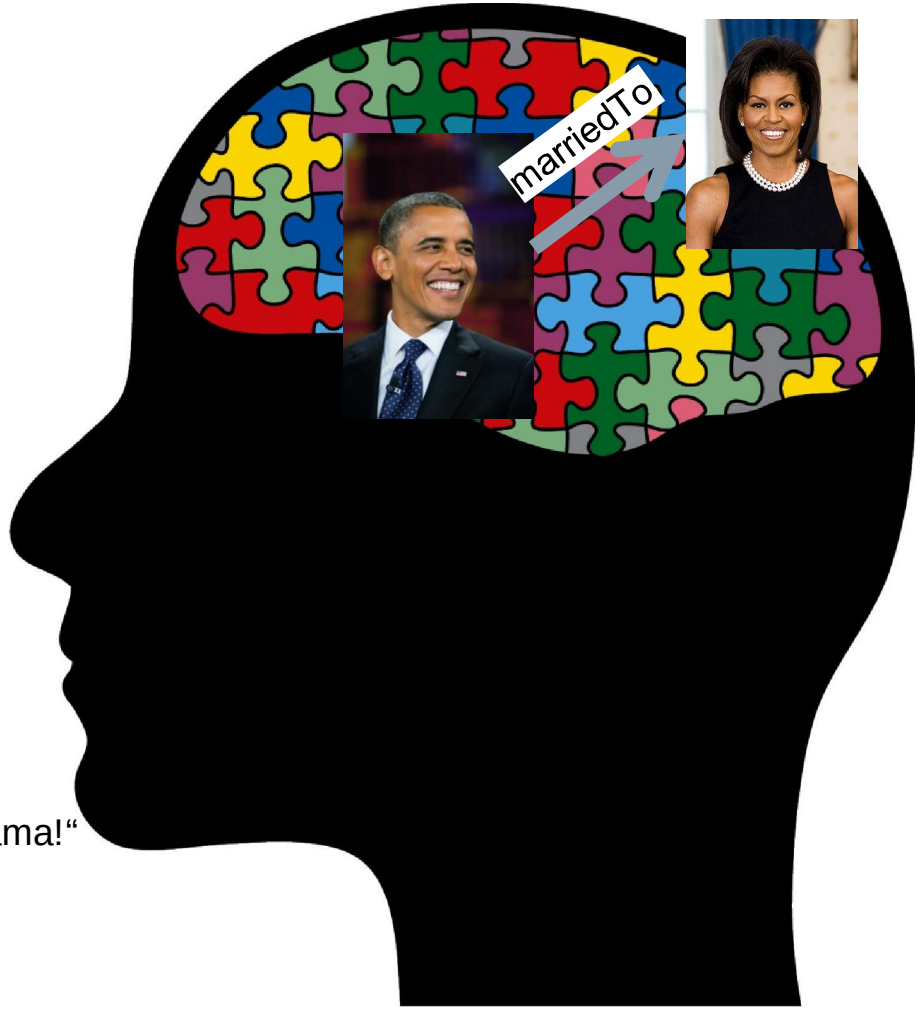
<Image of cats>

Predict all relationships: „Max likes Mary!“

[In our experiments 10^{14}] [#of synapses]

<Image of cats>

Vision



„You must be president Obama!“

„How is your wife Michelle?“

γλαῦκας εἰς Ἀθήνας κομίζειν

Requirement: Understanding of the World

- We need to know about the entities, attributes and classes in the world, and the various relationships that do or might exist between those
- We need ontologies!

Biomedical Ontologies

International Statistical Classification of Diseases and Related Health Problems (ICD)

- Used extensively in billing

SNOMED Clinical Terms (SNOMED CT)

- A systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting.
- Application: *EHR*

RadLex

- Unified language of radiology terms for standardized indexing and retrieval of radiology information resources

Open Biomedical Ontologies (OBO)

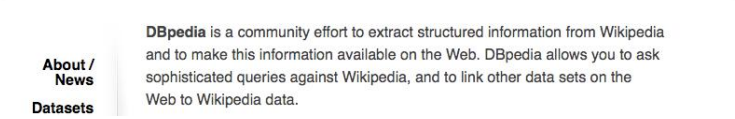
- Controlled vocabularies for shared use across different biological and medical domains
- Gene Ontology (GO) is a part (genes and gene products)

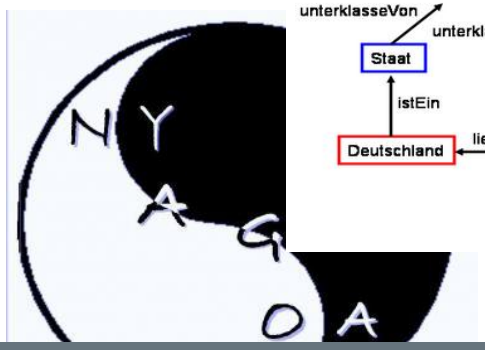
Example GO term [edit]

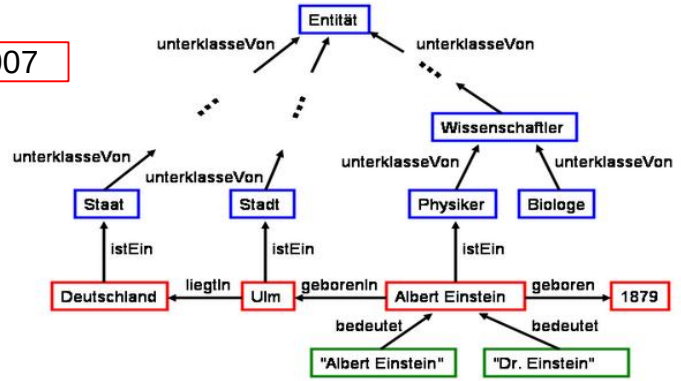
```
id:         GO:0000016
name:       lactase activity
namespace:  molecular_function
def:        "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
synonym:    "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym:    "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref:       EC:3.2.1.108
xref:       MetaCyc:LACTASE-RXN
xref:       Reactome:20536
is_a:       GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

For the First Time there Exist Sizable General Ontologies: DBpedia, YAGO, Freebase, Knowledge Graph

Auer, Bizer, Kobilarov, Lehmann, Cyganiak, Ives: 2007



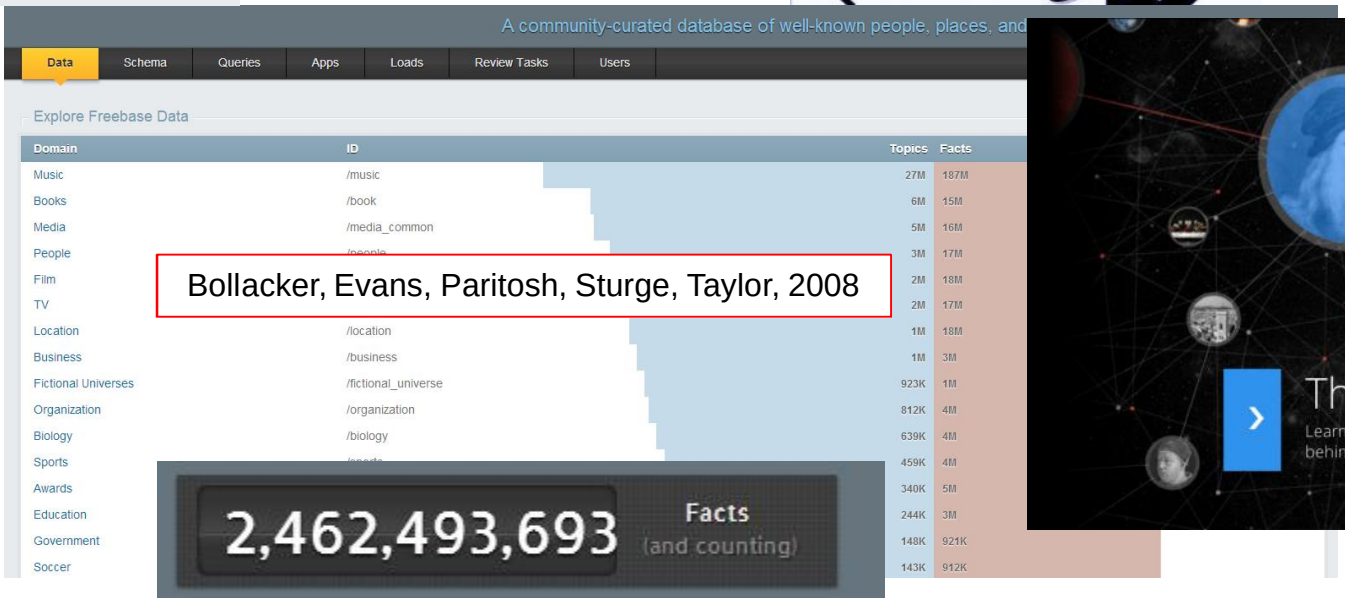




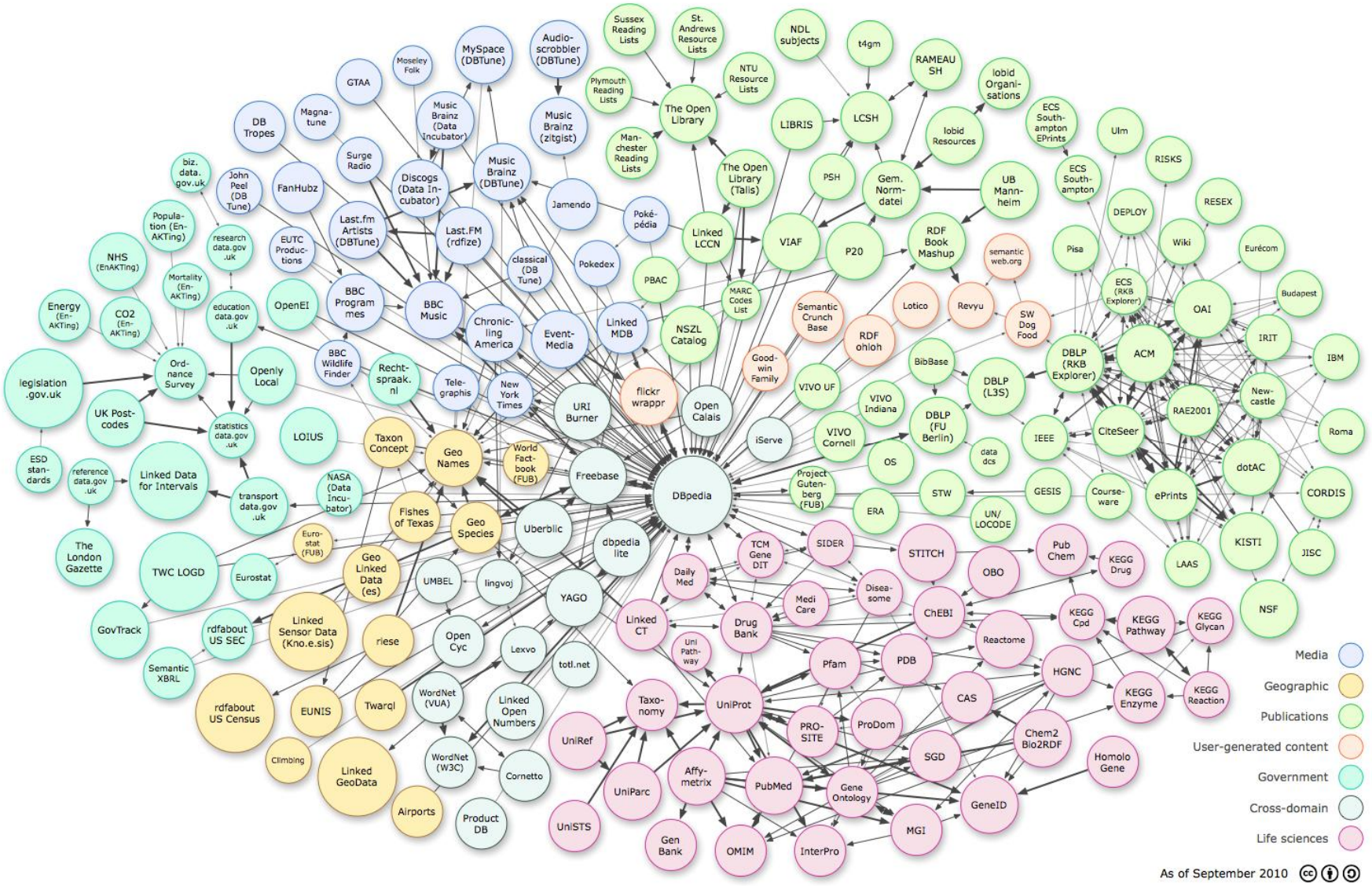
Suchanek, Kasneci, Weikum: 2007

Freebase

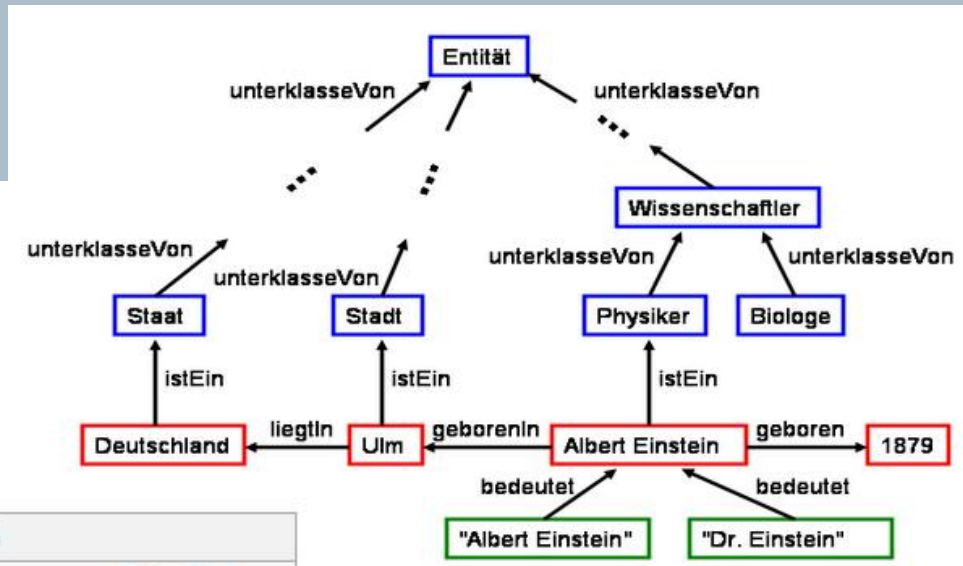
Bollacker, Evans, Paritosh, Sturge, Taylor, 2008

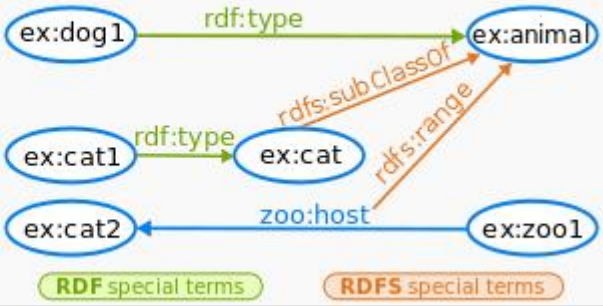


Linked Open Data (Semantic Web)



Triple Graphs



In english	The graph
• Dog1 is an animal • Cat1 is a cat • Cats are animals • Zoos host animals • Zoo1 hosts the Cat2	

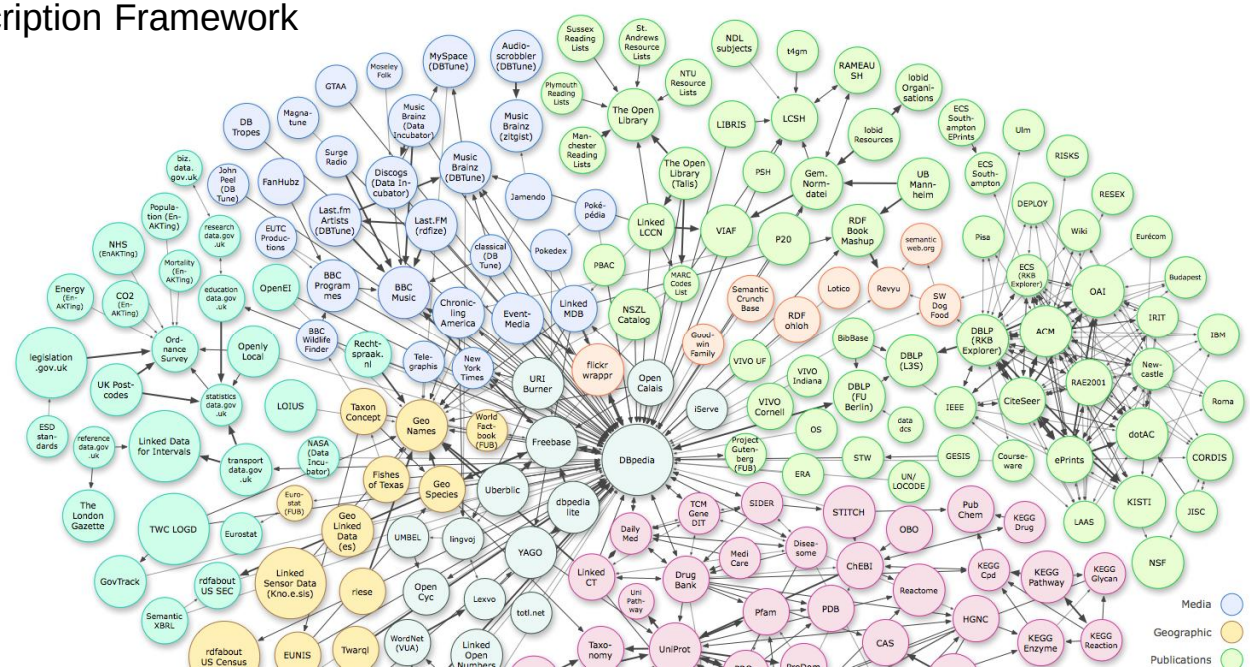
RDF/turtle

```
@PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@PREFIX rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@PREFIX ex:   <http://example.org/> .
@PREFIX zoo:   <http://example.org/zoo/> .
ex:dog1    rdf:type         ex:animal .
ex:cat1    rdf:type         ex:cat .
ex:cat     rdfs:subClassOf  ex:animal .
zoo:host   rdfs:range       ex:animal .
ex:zoo1    zoo:host         ex:cat2 .
```

Knowledge Bases are Triple Graphs

- Linked Open Data (LOD) and large ontologies like DBpedia, Yago, Knowledge Graph are graph-based knowledge representations using light-weight ontologies, and are ***accessible to machine learners***

- They are all triple oriented and more or less follow the RDF standard
  - RDF: Resource Description Framework



Overview

- Why Machine Learning needs Knowledge Graphs
- **Statistical Relational Learning**
- Learning with the YAGO Knowledge Graph
- Towards Relevant Use Cases

Canonical Relational Machine Learning Task

$\langle e_i, r^k, e_j \rangle$ true or false?

$$P\left(\langle e_i, r^k, e_j \rangle = 1\right) \leftarrow f^k(z_l)$$

- So, very simple, we build one classifier for each relation type *k* and we are done
- But what is the input z_l ?

I. Relational Learning with *Known* Features

features (age, sex, *features derived from a neighborhood of the entity in the environment of the RDF-graph*)

$$(a_{i,1}, a_{j,2}, \ldots, a_{i,r})^T$$

$$\langle e_i, r^k, e_j \rangle \rightarrow (a_{j,1}, a_{j,2}, \ldots, a_{j,r})^T$$

$$x^k(z_{l(i,j)}) \qquad z_{l(i,j)} = (a_{i,1}, a_{j,2}, \ldots, a_{i,r}, a_{j,1}, a_{j,2}, \ldots, a_{j,r})^T$$

$$f^k(z_l) = \sum_{m=1}^{M} w_m^k b_m(z_l)$$

$$f^k(z_l) = \sum_{n=1}^{N} v_n^k k(z_l, z_n)$$

Popular in learning from the Semantic Web

$$f^k(z_l) = NN_{deep}(z_l)$$

II. Relational Learning with *Latent* Features

Same, but features are treated as *latent (unknown) variables*

$$\langle e_i, r^k, e_j \rangle \rightarrow (a_{i,1}, a_{j,2}, \ldots, a_{i,r})^T$$

$$\langle e_i, r^k, e_j \rangle \rightarrow (a_{j,1}, a_{j,2}, \ldots, a_{j,r})^T$$

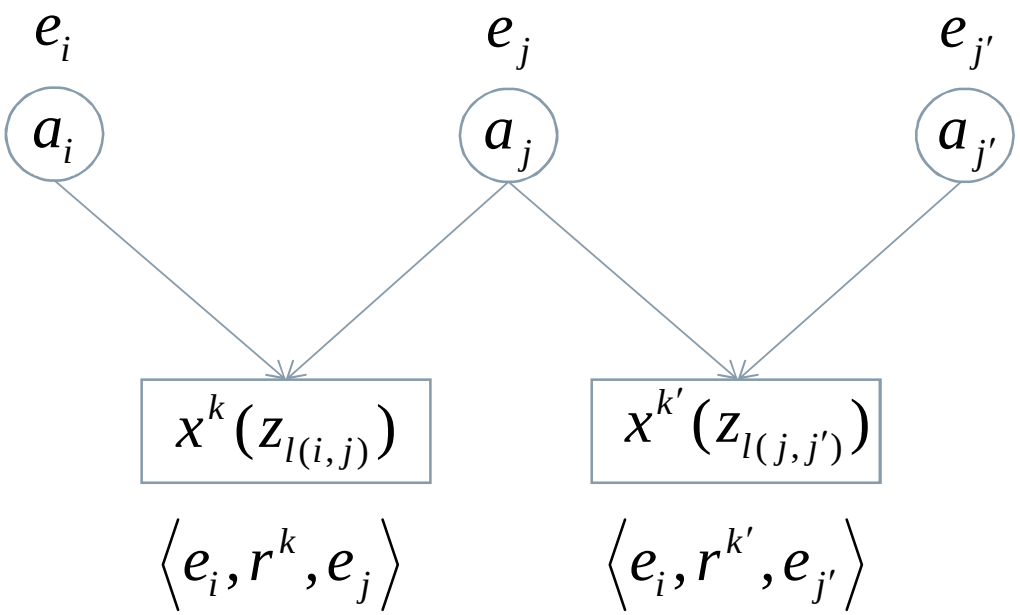
$$x^k(z_{l(i,j)})$$

$$z_{l(i,j)} = (a_{i,1}, a_{j,2}, \ldots, a_{i,r}, a_{j,1}, a_{j,2}, \ldots, a_{j,r})^T$$

unknowns!

$$f^k(z) = \sum_{m=1}^{M} w_m^k b_m(z_l)$$

With Latent Features We Get *Collective Learning*



Information flow open since *unknown* parents states a_j

Information flow open since *known* children states $x^{k'}(z_{l(j,j')})$

- Information can globally propagate in the network of random variables
- Thus one can learn that: *Jack is rich since the father of his father is rich*

Model with Polynomial Basis Functions

- But what are good basis functions?
- We need to represent the interactions between all feature components
- Binary interactions

$$f^k(z_l) = \sum_{s=1}^{r} \sum_{t=1}^{r} w_{s,t}^k b_{s,t}(z_l)$$

$$b_{s,t}(z_l) = a_{i,s} a_{j,t}$$

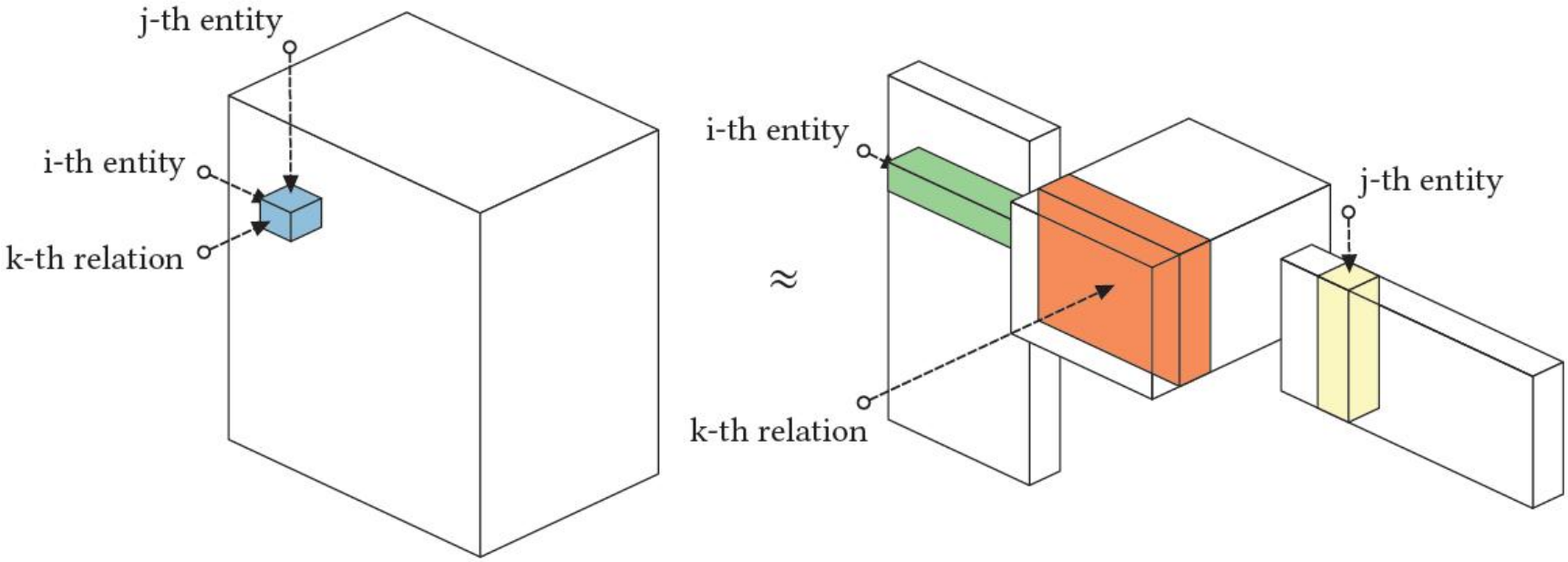
Mapping to a Tensor Factorization Problem

$$f^k(z_l) = \sum_{s=1}^{r}\sum_{t=1}^{r} w^k_{s,t} a_{i,s} a_{j,t} = a_i^T R_k a_j \qquad \left(R_k\right)_{s,t} = w^k_{s,t}$$

- Here, R_k is a *r x r* matrix
- We can take the matrices for the different relations $R_1, R_2, R_3, \ldots$ on to of each other and obtain the core tensor *R*
- In tensor notation: We factorize the tensor *X*

$$X \leftarrow R \times_1 A \times_2 A \qquad \left(X\right)_{i,j,k} = x^k(z_{l(i,j)})$$

RESCAL Factorization



$\mathbf{X}$ $\quad\quad$ $\mathbf{R} \times_1 A \times_2 A$

$$\mathcal{X}_{ijk} = \begin{cases} 1, & \text{if triple (i-th entity, k-th relation, j-th entity) exists} \\ 0, & \text{otherwise} \end{cases}$$

Cost Functions

Frobenius norm

$$\arg\min_{A,\mathbf{R}} \|\mathbf{X} - \mathbf{R} \times_1 A \times_2 A\|^2 + \lambda_A \|A\|^2 + \lambda_{\mathbf{R}} \|\mathbf{R}\|^2$$

Probabilistic View

$$\mathrm{P}(\mathbf{X} \mid A, \mathbf{R}) = \prod_{i=1}^{n} \prod_{j=1}^{n} \prod_{k=1}^{m} \mathrm{P}\left(x_{ijk} \mid \boldsymbol{a}_i^T R_k \boldsymbol{a}_j\right)$$

$$\boldsymbol{a}_i \sim \mathcal{N}(0, \sigma_A^2 I)$$

$$R_k \sim \mathcal{N}(0, \sigma_R^2 I)$$

Gaussian $x_{ijk} \sim \mathcal{N}(\boldsymbol{a}_i^T R_k \boldsymbol{a}_j, \sigma^2)$

Bernoulli $x_{ijk} \sim Bernoulli(\boldsymbol{a}_i^T R_k \boldsymbol{a}_j)$

Iterative Update

- Most efficient: Alternating Least Squares (ALS)
  - Can exploit data sparsity
- (stochastic gradient descent, …)

$$A \leftarrow \left(\sum_{k=1}^{m} X_k A R_k^T + X_k^T A R_k\right)\left(\sum_{k=1}^{m} B_k + C_k + \lambda_A I\right)^{-1}$$

$$B_k = R_k A^T A R_k^T, \quad C_k = R_k^T A^T A R_k$$

$$\mathrm{vec}\,(R_k) \leftarrow \left(Z^T Z + \lambda_{\mathbf{R}} I\right)^{-1} Z^T \mathrm{vec}\,(X_k)$$

$$Z = A \otimes A$$

RESCAL for Different -arities

Unary Relations

$$P(r_k(e_i)) \leftarrow r_k^T a_i = \sum_{n=1}^{r} r_{k,n} a_{i,n}$$

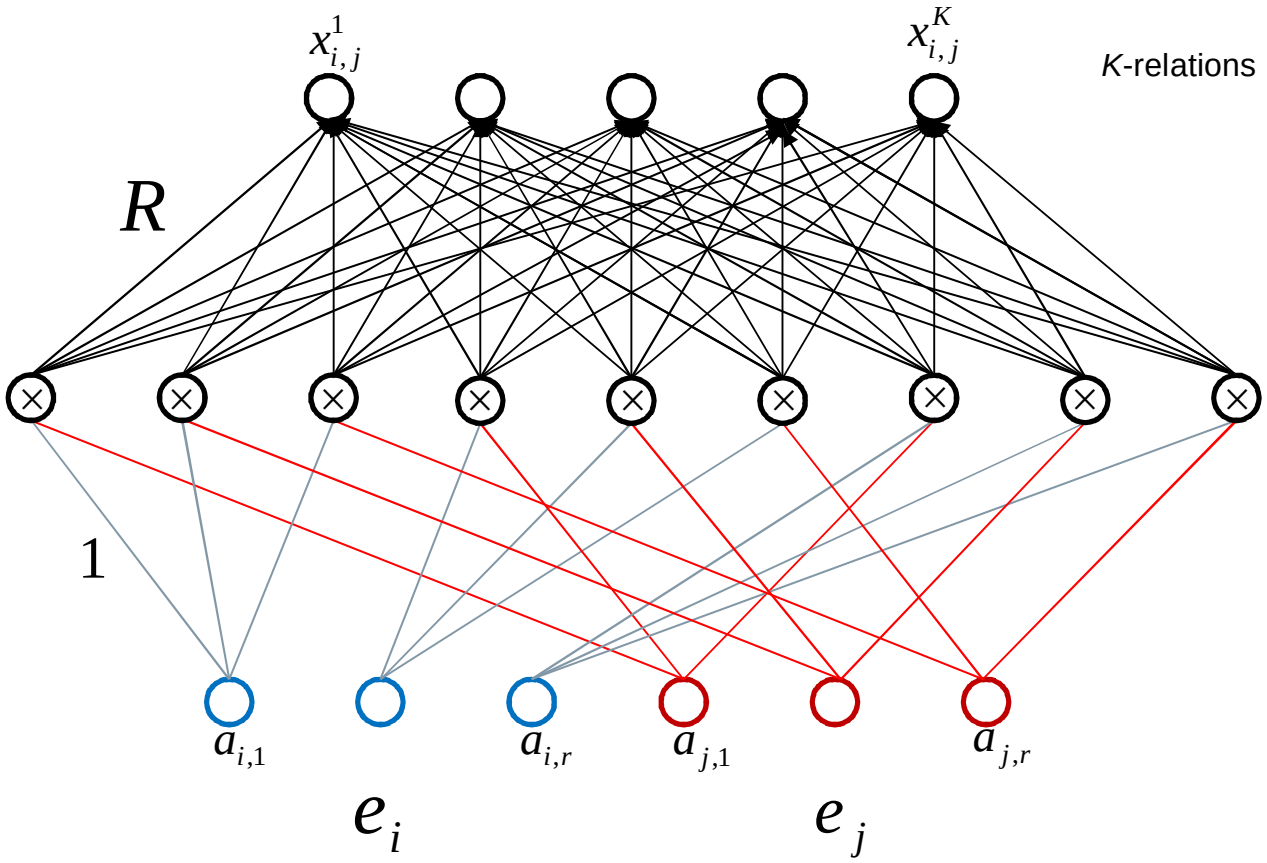
Binary Relations

$$P(r_k(e_i, e_j)) \leftarrow a_i^T R_k a_j = \sum_{n_1=1}^{r} \sum_{n_2=1}^{r} R_{k,n_1,n_2} a_{i,n_1} a_{j,n_2}$$

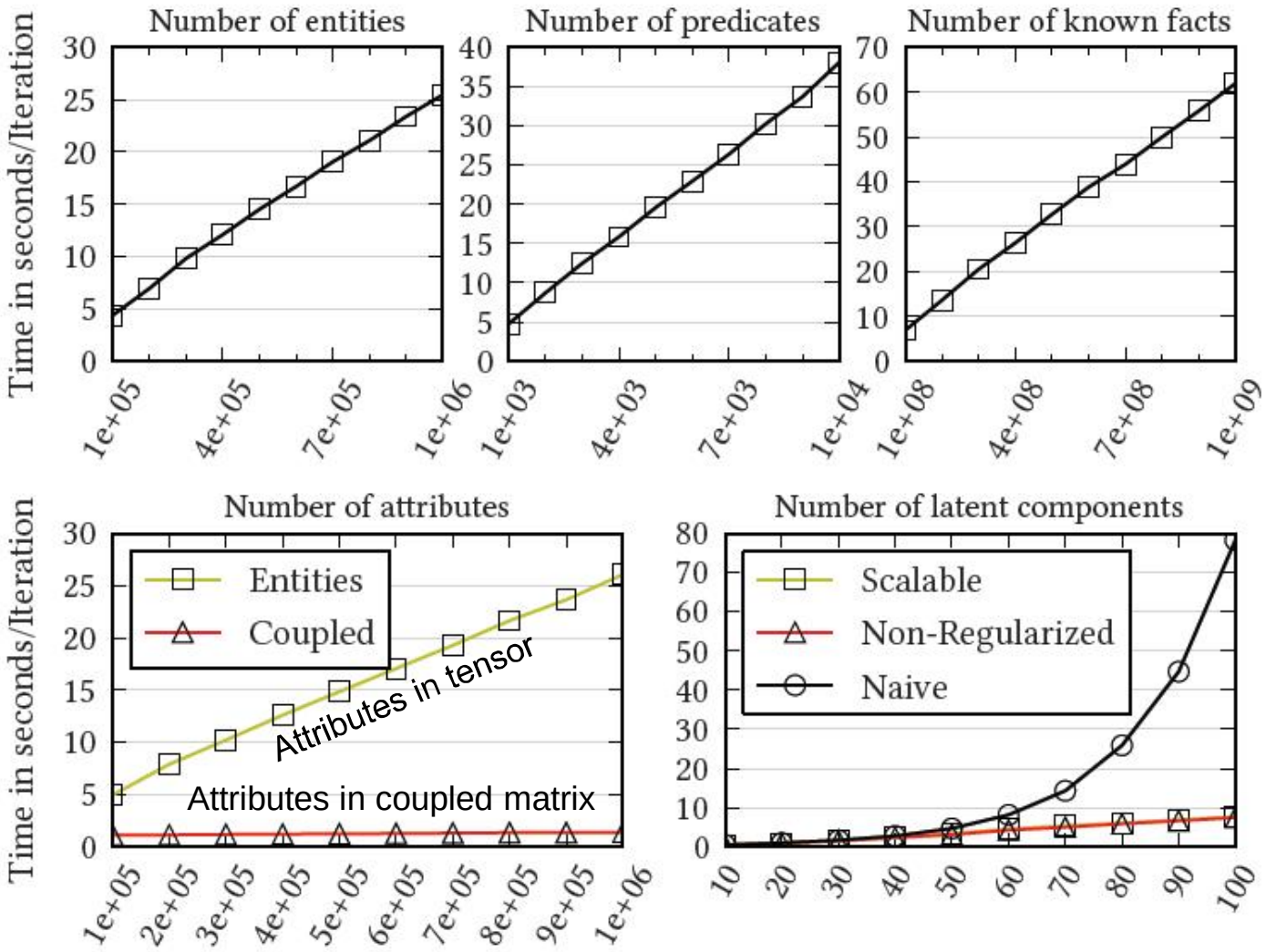
Ternary Relations

$$P(r_k(e_i, e_j, e_l)) \leftarrow \sum_{n_1=1}^{r} \sum_{n_2=1}^{r} \sum_{n_3=1}^{r} R_{k,n_1,n_2,n_3} a_{i,n_1} a_{j,n_2} a_{l,n_3}$$

RESCAL for Binary Relations



Scalabilty


Number of entities
Number of predicates
Number of known facts
Time in seconds/Iteration
Number of attributes
Number of latent components
Entities
Coupled
Attributes in tensor
Attributes in coupled matrix
Scalable
Non-Regularized
Naive

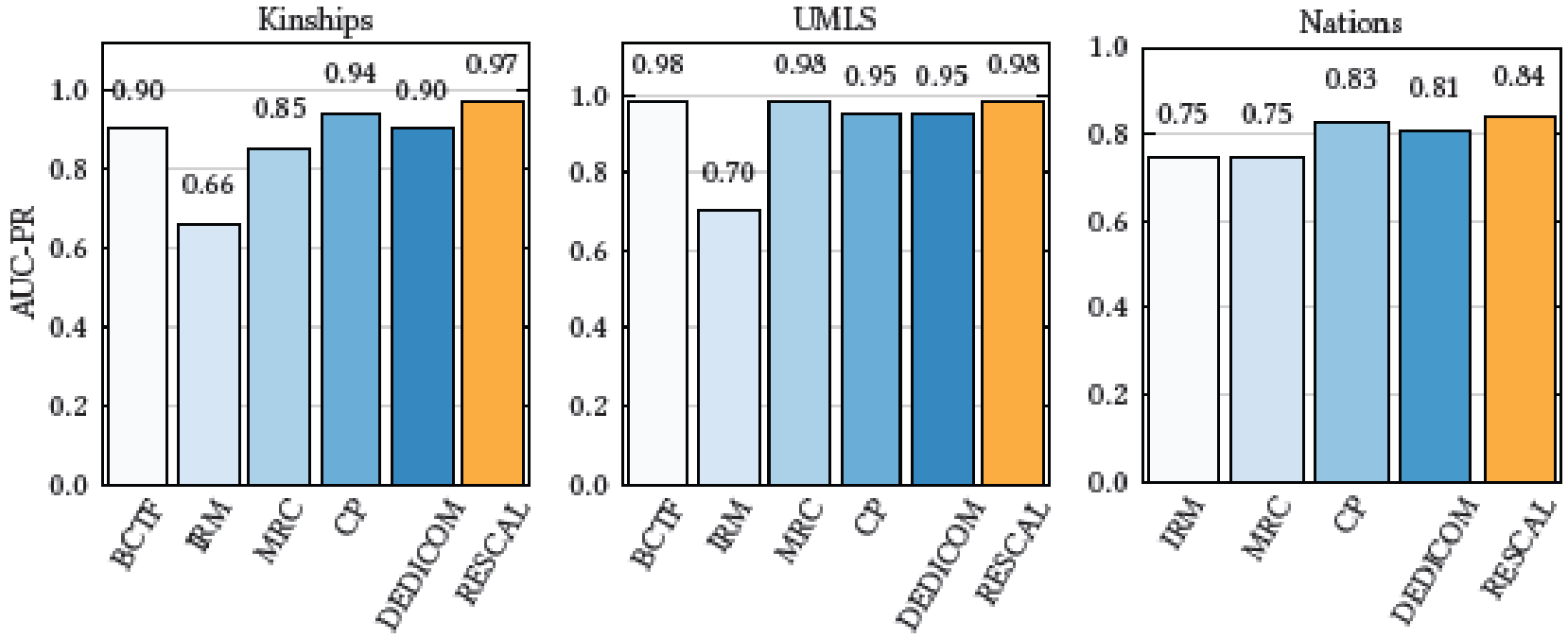
Leading Performance in Link prediction on benchmark data sets

Predicting relationships:
„Max likes Mary“

Kinship: multiple kinship relations between members of the Alyawarra tribe in central Australia (10,790 kinship relationships (facts) between 104 persons over 26 relations)

UMLS: The UMLS data set consists of a small semantic network which is part of the Unified Medical Language System (UMLS) ontology. 6,752 relationships (facts) between 135 concepts over 49 relations

Nations: The Nations data set describes political interactions of countries between 1950 and 1965 . It contains information such as military alliances, trade relationships or whether a country maintains an embassy in a particular country. 2,024 relationships between 14 countries over 56 dyadic relations

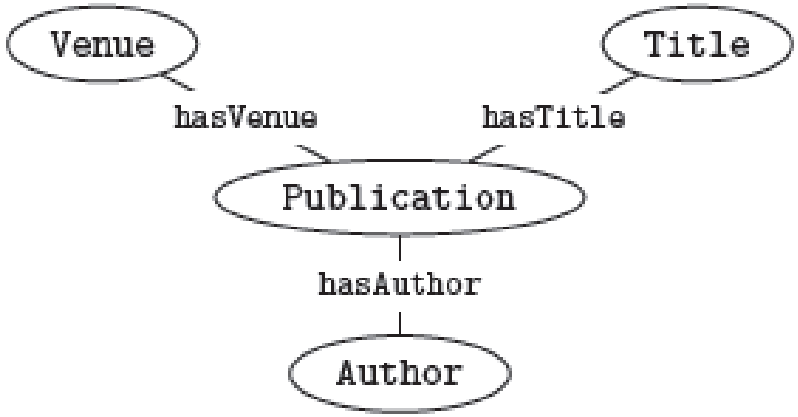


BCTF: Bayesian clustered tensor factorization; MRC: Multi-View Relational Classification

Cora Data: Entity Resolution

Recognizing specific entities:
„This my cat Max!“

- 1295 publication records, where each publication is the subject of a relationship to its first author, a relationship to its title, and a relationship to its publication venue
- Task: identify which authors, entities and venues refer to identical entities

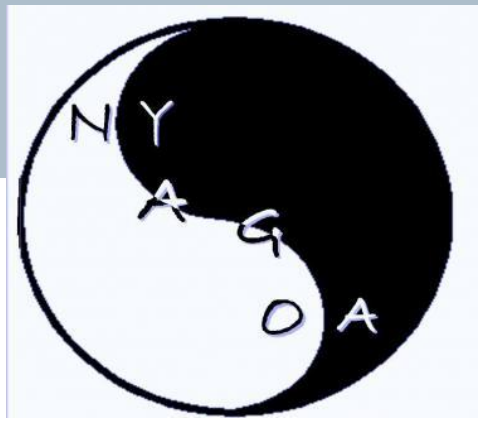


	AUC-PR				
Entity Type	Naive Bayes	MLN (B) (basic rules)	MLN (BCTS) (complex rules)	CP	RESCAL
Publications	0.913	0.915	0.988	0.991	0.991
Authors	0.986	0.987	0.992	0.984	0.997
Venue	0.738	0.736	0.807	0.746	0.810

Overview

- Why Machine Learning needs Knowledge Graphs
- Statistical Relational Learning
- **Learning with the YAGO Knowledge Graph**
- Towards Relevant Use Cases

Yago2 Core Ontology



YAGO2 core ontology	
Number of Resources	2.6 million
Number of Classes	340,000
Number of Predicates	87
Number of Known Facts	33 million

The tensor has 10^{14} entries!

Siemens – MPII cooperation

Classification: Type Prediction

Type	Number of entities
wordnet:person	884,261
wordnet:location	429,828
wordnet:movie	62,296

Predicting concepts:
„This is a cat“

Table 3.9.: Link-prediction experiments on YAGO2.

	AUC-PR		
	wordnet:person	wordnet:location	wordnet:movie
Random	0.32	0.18	0.06
Setting a)	0.99	1.0	0.75
Setting b)	0.96	0.98	0.51
With attributes	-	-	0.85

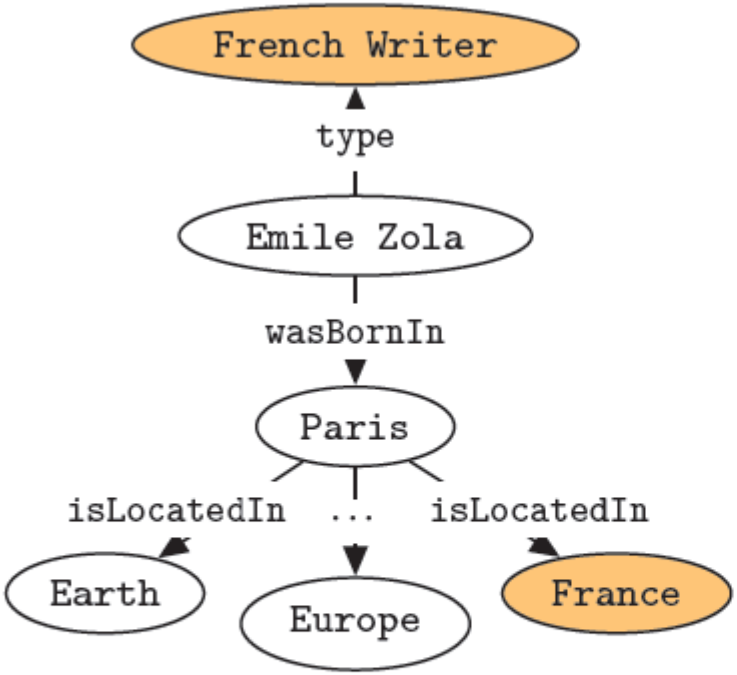
(text attributes)

a) Only those `rdf:type` triples that include the class C that should be predicted were removed from the test fold. All other type triples, including subclasses of C, are still present in the data.

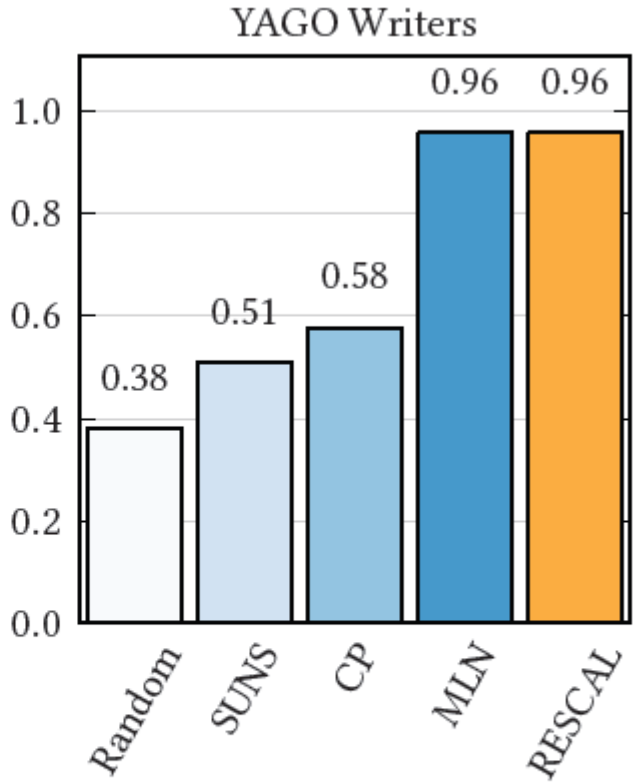
b) *All* `rdf:type` triples were deleted in the test fold.

Writer‘s Nationality: Demonstrating Collective Learning

Predicting concepts/attributes:
„Max is evil“



(a) Collective learning example on YAGO. The objective is to learn the correlation between `France` and `French Writer` from examples like `Emile Zola`.



(b) Results for link prediction on YAGO2 writers data set over ten-fold cross-validation.

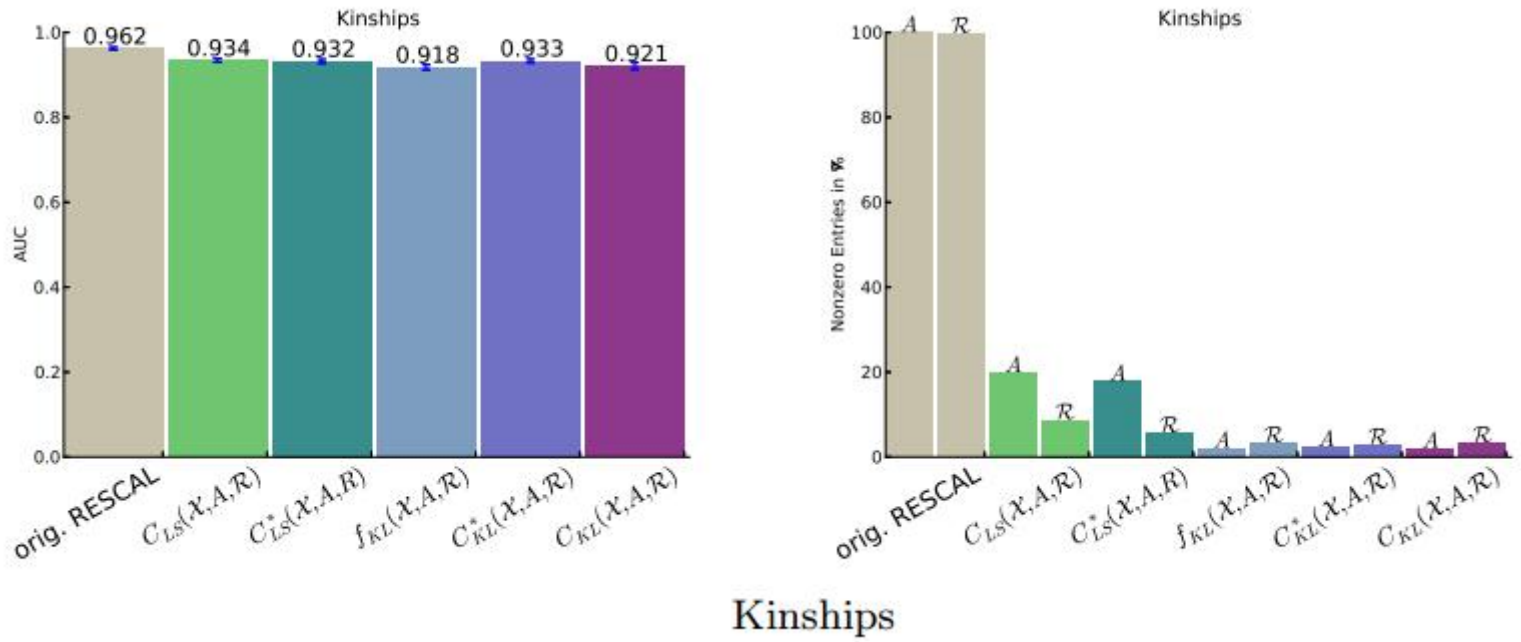
Learning a Taxonomy (-> Ontology)

- IIMB 2010 benchmark provided by the Ontology Alignment Evaluation
- Around 1400 entities of a movie domain
- 5 distinct top-level concepts
- On the top level: every concept is represented by a sufficient number of entities, while e.g. some level 2 movie concepts only include two or three entities and therefore are hard to recognize.

Table 3.10.: F-measure for selected concepts and weighted F-measure for all concepts per subclass-level

Level 1		**Level 2**		**Level 3**	
Locations	0.95	City	0.99	Capital	0.99
Films	1.0	Anime	0.67	Director	0.78
Creature	1.0	Character	0.73	Character Creator	0.53
Budget	1.0	Person	1.0	Actor	0.98
Language	1.0	Country	0.80		
All	0.982	**All**	0.852	**All**	0.947

Extensions: Nonnegative RESCAL



Kinships

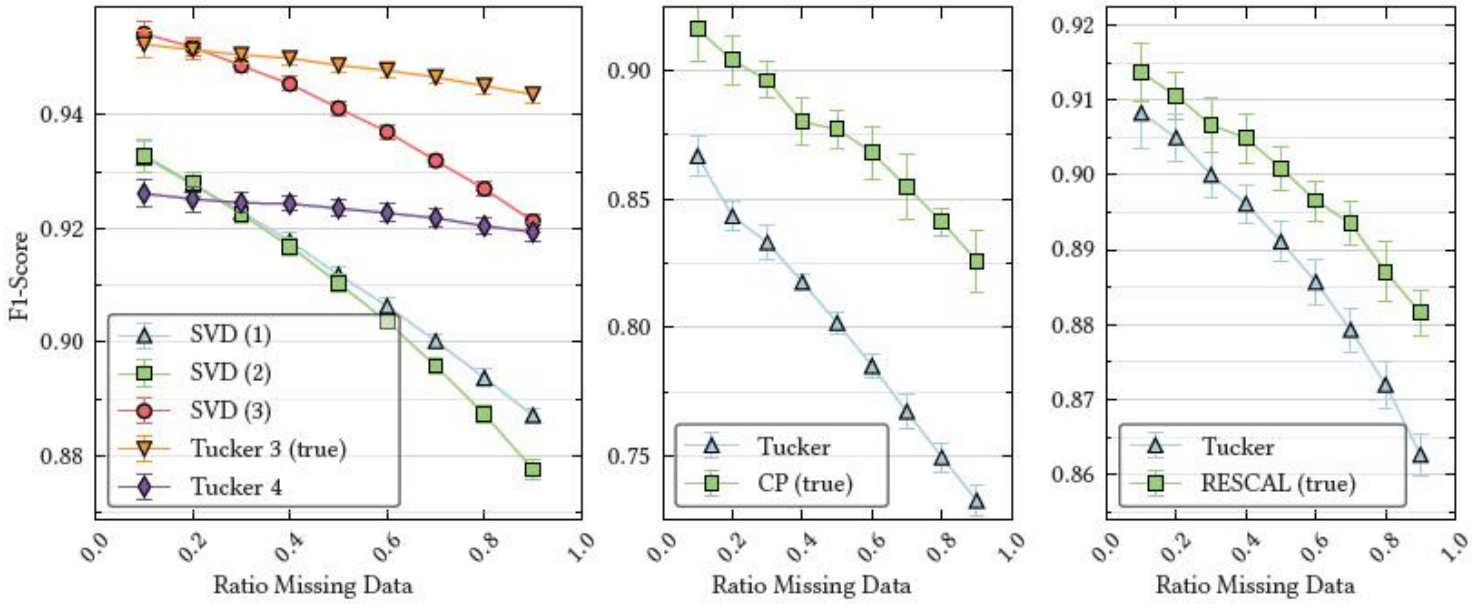
Nonnegatve RESCAL (Krompass, Nickel, Tresp)

- sparse solutions with clustering properties

Extensions: Proofs and Bounds

- Analysis of generalization bounds when order of the tensor match or do not match
- Matricization results in a loss of generalization performance

Maximilian Nickel and Volker Tresp. An Analysis of Tensor Models for Learning on Structured Data. Proceedings of the ECML/PKDD, 2013



Overview

- Why Machine Learning needs Knowledge Graphs
- Statistical Relational Learning
- Learning with the YAGO Knowledge Graph
- **Towards Relevant Use Cases**

Machine Learning with Structured Data and Ontologies

Within the domain:

- Prediction of triples
- Classification (defining type)
- Clustering
- Taxonomy Learning
- Entity Resolution
- Visualization
- Querying
  - Who wants to be Trelenas friends
  - Can be generalized towards more complex probabilistic queries *(Krompass, Nickel, Tresp, ISWC 2014)*

Outside of the domain (new entities):

- Calculate the latent factors for the new entity
  - Can do all of the tasks above
  - Object recognition becomes entity resolution
- Formulate the new object as a query
  - Object recognition as a query
  - Queries can become complex

Clinical Data Intelligence

Bundesministerium für Wirtschaft und Technologie

Goals

- Personalized medicine: modeling the patient in her/his full complexity -> patient specific recommendations
- Global modeling of the clinical data / clinical decision processes: clinical ontology (concepts and instances)

Use Cases

- All data from all patients
- Breast cancer
- Nephrology
- Data from clinical studies

Challenges

- Ontologies
- Complex relational data (patient in a clinic)
- Representing time; sequential data
- Decision modeling: decision optimization (confounders, causality)
- Including unstructured data (reports, images)
- Including OMICS data

SIEMENS

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

averbis medical language technology

Institut für FrauenGesundheit (IFG®)

Predicting Diagnoses and Procedures

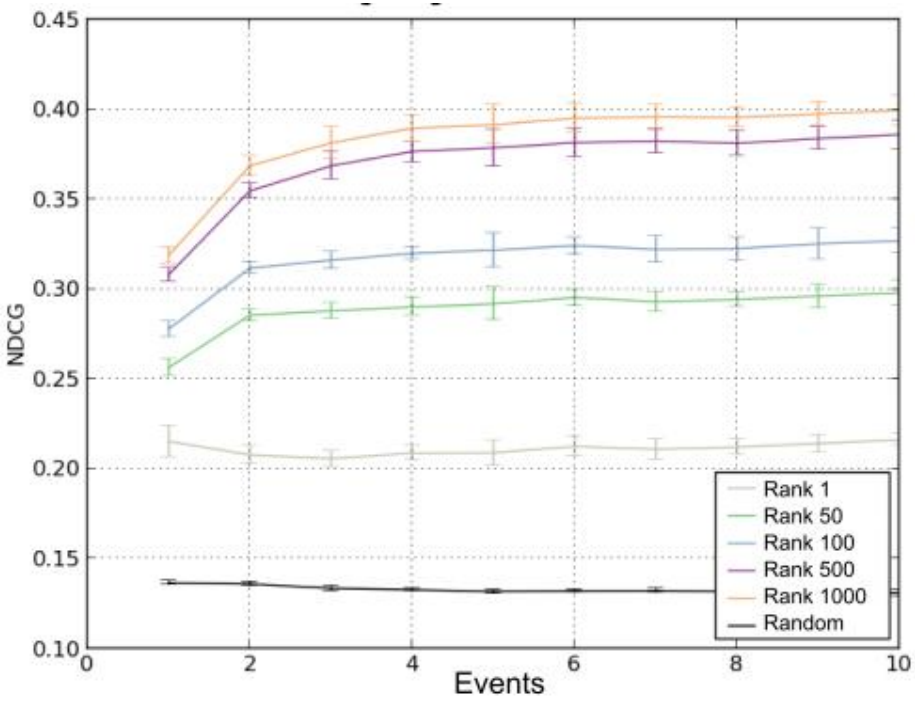


Figure 1: Data from 10000 patients were used. We considered 2331 possible diagnoses, 1634 possible procedures, 2721 possible lab results, 209 possible therapies and 281 general patient data. In total the data contained 5.9 million facts. We predicted the next decision (diagnosis, procedure) as a function of the information available for each patient. Plotted is the NDCG score (a popular score for evaluating ranking results [11]) as a function of the information available for each patient (a large number is desirable). An event corresponds to an instance in time where patient data is recorded. With increasing information, the prediction improves. We see plots for different approximation ranks: the highest rank gives best scores which reflects the high degree of data complexity.

Machine Learning with Images and Ontologies

Linking textual descriptions in radiology reports to medical images

References and Related Work

RESCAL

- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proceedings of the 28th International Conference on Machine Learning*, 2011
- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing YAGO: Scalable Machine Learning for Linked Data. In *Proceedings of the 21st International World Wide Web Conference (WWW1012)*, 2012
- Maximilian Nickel and Volker Tresp. An Analysis of Tensor Models for Learning on Structured Data. *Proceedings of the ECML/PKDD*, 2013
- Denis Krompaß, Maximilian Nickel, Xueyan Jiang, and Volker Tresp. Non-Negative Tensor Factorization with RESCAL. *ECML/PKDD 2013 Workshop on Tensor Methods for Machine Learning*, 2013

Extensions from other groups

- R. Jenatton, N. Le Roux, A. Bordes, G. Obozinski (contributed equally). A latent factor model for highly multi-relational data. Advances in Neural Information Processing Systems, *NIPS*, 2012
- Richard Socher, Danqi Chen, Christopher D. Manning, Andrew Y. Ng. Reasoning With Neural Tensor Networks for Knowledge Base Completion, NIPS, 2013

SUNS (First application of factorization approaches to relational Semantic Web domains)

- Volker Tresp, Yi Huang, Markus Bundschus, and Achim Rettinger. Materializing and querying learned knowledge. *IRMLeS, 2009*

Triplerank (Application of PARAFAC for ranking; no collective learning)

- T. Franz, A. Schultz, S. Sizov, and S. Staab. “Triplerank: Ranking semantic web data by tensor decomposition”. *ISWC,* 2009

Factorization Machines

- S. Rendle et al.: Different factorization approaches for preference prediction and relational learning (2009 and later)

Knowledge Vault (Google Team)

- X. Dong , E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, ND W. Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion KDD 2014.

Conclusions

- **Knowledge Graphs**
  - First time: large general ontologies available
  - Useful for solving machine learning tasks
- **Relational Machine Learning with RESCAL**
  - Scalable relational learning with very competitive performance
  - Collective Learning
  - We are working on many improvements/extensions
- **RESCAL Learning with the YAGO Knowledge Graph**
  - Experimental results in a number of relational learning tasks
- **Towards Relevant Use Cases**
  - Text understanding
  - Image understanding
  - Clinical data