Intelligent Agents

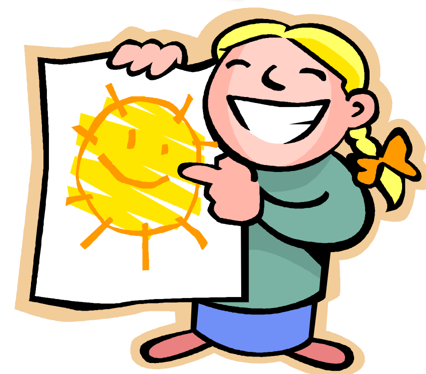
Multi-Relational Latent Semantic Analysis

Ralf Möller
Universität zu Lübeck
Institut für Informationssysteme

Semantics Needs More Than Similarity

Tomorrow will be rainy.

Tomorrow will be sunny.



$similar$(r[illegible], sunny)?

$antonym$(rainy, sunny)?

Leverage Linguistic Knowledge

- Can’t we just use the existing thesauri for information about synonyms and antonyms?
  - Knowledge in these resources is never complete
  - Often lack of “membership degree” for relations
    - Various ways to measure “membership degree”

- Goal: Create a representation that
  - leverages existing rich linguistic resources,
  - discovers new relations, and
  - enables us to measure the “degree” of multiple relations (not just similarity)

Roadmap

- Two opposite relations:
  - Polarity Inducing Latent Semantic Analysis
- Multiple relations:
  - Multi-Relational Latent Semantic Analysis
- Relational domain knowledge

- Yih, Zweig & Platt. *Polarity Inducing Latent Semantic Analysis*. In EMNLP-CoNLL-12.
- Chang, Yih & Meek. *Multi-Relational Latent Semantic Analysis*. In EMNLP-13.
- Chang, Yih, Yang & Meek. *Typed Tensor Decomposition of Knowledge Bases for Relation Extraction*. In EMNLP-14.

EMNLP: Empirical Methods in Natural Language Processing
CoNLL: Computational Natural Language Learning
ACL; Annual Meeting of the Association for Computational Linguistics

Problem: Handling Two Opposite Relations

- Can cope to some extent with homonyms and synonyms due to word context
- Embedding techniques cannot clearly distinguish antonyms
  - “Distinguishing synonyms and antonyms is still perceived as a difficult open problem.” [Poon & Domingos 09]

- Idea #1: Change the data representation

Synonyms: Different words, same meaning
Homonyms: Same words, different meanings
Antonyms: Opposite words

Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In Proceedings EMNLP ‘09. **2009**.

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Target word: row-vector

	joy	gladden	sorrow	sadden	goodwill
Group 1: “joyfulness”	1	1	1	1	0
Group 2: “sad”	1	1	1	1	0
Group 3: “affection”	0	0	0	0	1

Wen-tau Yih, Geoffrey Zweig, John Platt. Polarity Inducing Latent Semantic Analysis. In Proceedings EMNLP ‘12. **2012**.

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Target word: row-vector

Inducing polarity

	joy	gladden	sorrow	sadden	goodwill
Group 1: "joyfulness"	1	1	-1	-1	0
Group 2: "sad"	-1	-1	1	1	0
Group 3: "affection"	0	0	0	0	1

Cosine Score: $+ \mathit{Synonyms}$

Encode Synonyms & Antonyms in Matrix

- Joyfulness: joy, gladden; sorrow, sadden
- Sad: sorrow, sadden; joy, gladden

Target word: row-vector

Inducing polarity

	joy	gladden	sorrow	sadden	goodwill
Group 1: “joyfulness”	1	1	-1	-1	0
Group 2: “sad”	-1	-1	1	1	0
Group 3: “affection”	0	0	0	0	1

Cosine Score: $-\ Antonyms$

Problem: How to Handle More Relations?

- Limitation of the matrix representation
  - Each entry captures a particular type of relation between two entities, or
  - Two opposite relations with the polarity trick
- Encoding other binary relations
  - Is-A (hyponym) – ostrich *is a* bird
  - Part-whole – engine is a *part of* car
- Idea #2
  - Encode multiple relations in a 3-way tensor (3-dim array)!

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Encode Multiple Relations in Tensor

- Represent word relations using a tensor
  - Each slice encodes a relation between terms and target words.

	joy	gladden	sadden	feeling
joyfulness	1	1	0	0
gladden	1	1	0	0
sad	0	0	1	0
anger	0	0	0	0

Synonym layer

	joy	gladden	sadden	feeling
joyfulness	0	0	0	0
gladden	0	0	1	0
sad	1	0	0	0
anger	0	0	0	0

Antonym layer

Construct a tensor with two slices

Encode Multiple Relations in Tensor

- Can encode multiple relations in the tensor

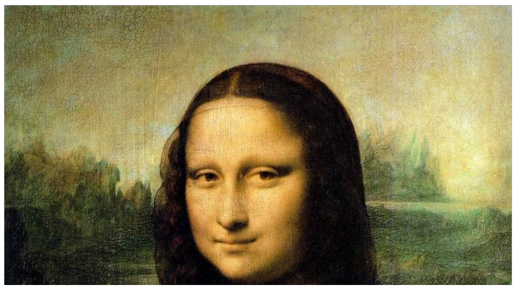
1	1	0	0
1	1	0	0
0	0	1	0
0	0	0	0

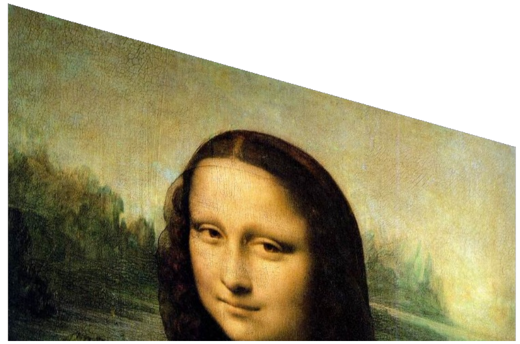
	joy	gladden	sadden	feeling
joyfulness	0	0	0	1
gladden	0	0	0	0
sad	0	0	0	1
anger	0	0	0	1

Hyponym layer

Hyponym IS-A/TYPE-OF hypernym
Metonym: Substitute for another term
(substitute usually used for sth else)

Wiederholung: Abbildung von Daten





- Beispiel: Scherung
- Der rote Pfeil ändert sich nicht

Matrixdarstellung [Bearbeiten | Quelltext bearbeiten]

Wählt man in der Ebene ein kartesisches Koordinatensystem, bei dem die x-Achse mit der Achse der Scherung zusammenfällt, dann wird diese Scherung durch die lineare Abbildung

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x + my \\ y \end{pmatrix} = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

mit der Abbildungsmatrix

$$\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$$

dargestellt. Ist die Achse der Scherung hingegen die y-Achse, tauschen 0 und m in der Abbildungsmatrix ihre Plätze. Beide Abbildungen verändern den Winkel zwischen den Koordinatenachsen jeweils um $\arctan m$.

[Wikipedia]

Eigenwerte und Eigenvektoren

- Eigenvektoren (für eine quadratische m×m Matrix **S**)

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v}$$

(rechter) Eigenvektor $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

Eigenwert $\lambda \in \mathbb{R}$

Beispiel

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- Wie viele Eigenwerte gibt es maximal?

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v} \iff (\mathbf{S} - \lambda \mathbf{I})\, \mathbf{v} = \mathbf{0}$$

Determinante

Hat eine von 0 verschiedene Lösung falls $|\mathbf{S} - \lambda \mathbf{I}| = 0$

Gleichung m-ter Ordnung in λ mit maximal m verschiedenen Lösungen (Nullstellen des charakteristischen Polynoms)
– möglicherweise komplex, obwohl **S** real ist.

Singulärwertzerlegung

Für eine $m \times n$ Matrix **A** vom Rang r gibt es eine Faktorisierung (Singulärwertzerlegung, engl. Singular Value Decomposition = **SVD**) wie folgt:

$$A = U \Sigma V^T$$

U: m×m, Σ: m×n, V^T: n×n

Spalten von **U**: links-singuläre Eigenvektoren von $\mathbf{AA^T}$

Spalten von **V**: rechts-singuläre Eigenvektoren von $\mathbf{A^TA}$

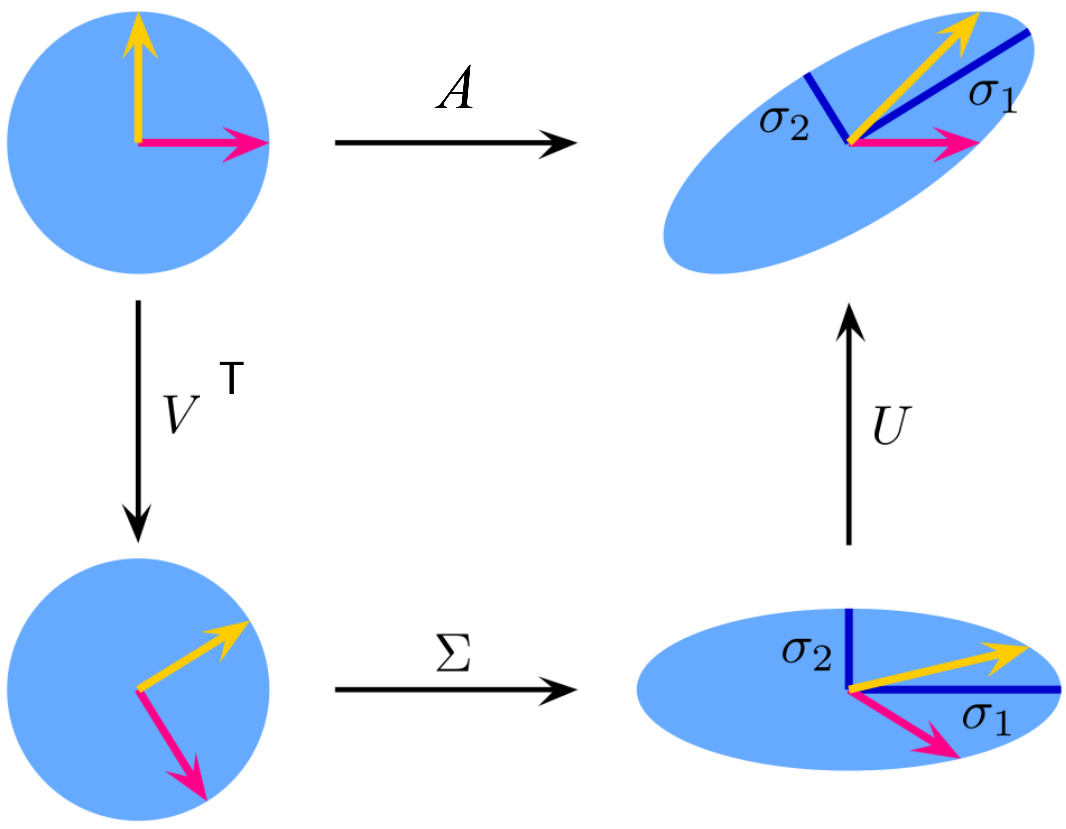
Eigenwerte $\lambda_1 \ldots \lambda_r$ von $\mathbf{AA^T}$ sind Eigenwerte von $\mathbf{A^TA}$

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1, \cdots, \sigma_r)$$

← Singulärwerte

Scherung mit Einheitsvektoren



$$A = U \cdot \Sigma \cdot V^T$$

[Wikipedia]

Approximation durch Matrix mit kleinem Rang

- SVD kann zur Berechnung einer optimalen Approximation einer Matrix A vom Rang r durch eine Matrix A_k mit kleinerem Rang k verstanden werden

$$A_k = \underset{X:rank(X)=k}{\arg\min} \|A - X\|_F$$

← Frobenius-Norm

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}.$$

- A_k und X sind beides m×n Matrizen
- Typischerweise $k << r$

Approximation durch Matrix mit kleinem Rang

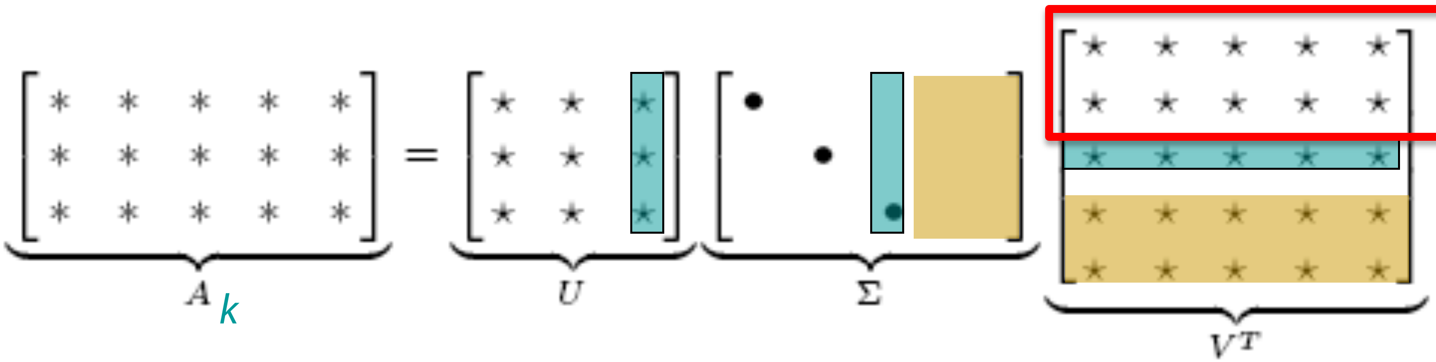
Optimierungsproblem $A_k = \arg\min_{X:rank(X)=k} \|A - X\|_F$ k fix

Lösung mittels SVD

$$A_k = U \cdot diag(\sigma_1, \cdots, \sigma_k, \underbrace{0, \cdots, 0}) \cdot V^T$$

Setze kleinste r-k
Eigenwerte auf 0

Neue Dokumente

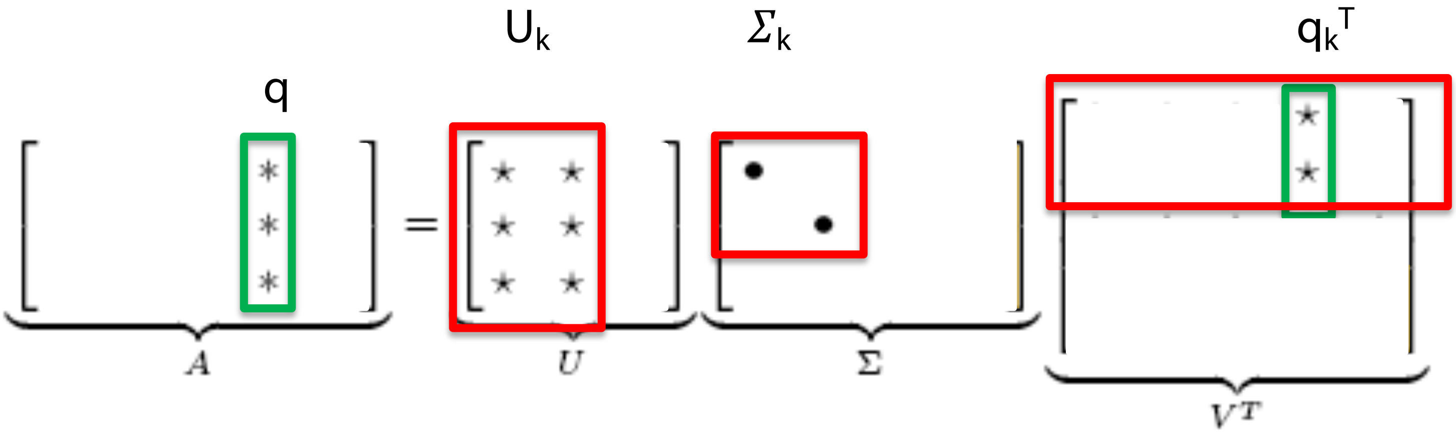


C. Eckart, G. Young, The approximation of a matrix by another of lower rank. Psychometrika, 1, 211-218, **1936**

Anwendung zur Informationsrecherche

- Eine Term-Dokument-Matrix kann m=50000, n=10 Millionen Einträge haben (Rang nah bei 50000)
- Wir können eine Approximation A_{100} konstruieren mit Rang 100 und kleinstem Frobenius-Fehler
  - Auch Hauptkomponentenanalyse genannt (engl. Principle Component Analysis, PCA)
- Die neue Matrix (siehe vorigen Präsentation) definiert latente Merkmale (keine verstehbaren Terme mehr) für die Informationsrecherche (Latent Semantic Indexing, LSI)

Wie behandeln wir Anfragen?



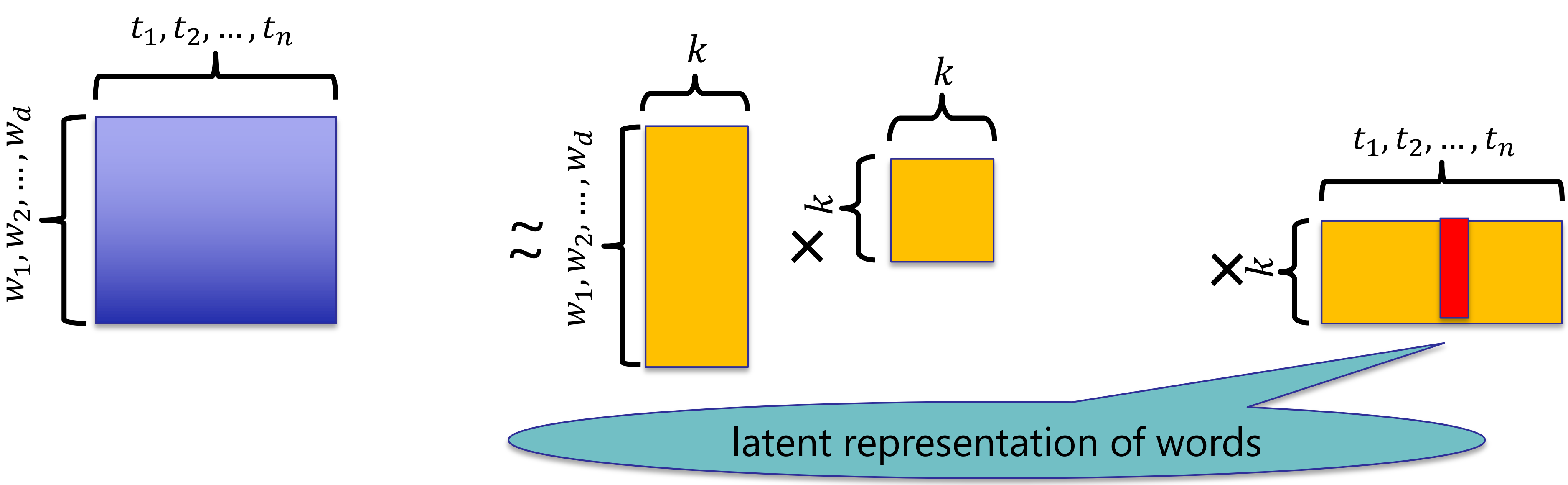
- Anfrage q (dünn besetzt)
- Eine Anfrage q wird wie folgt in den LSI-Raum abgebildet

$$q_k = q^T U_k \Sigma_k^{-1\,T}$$

- Anfrage q_k ist nicht dünn besetzt
- Anfragebeantortung über k nächste Nachbarn (Cosinusabstand)

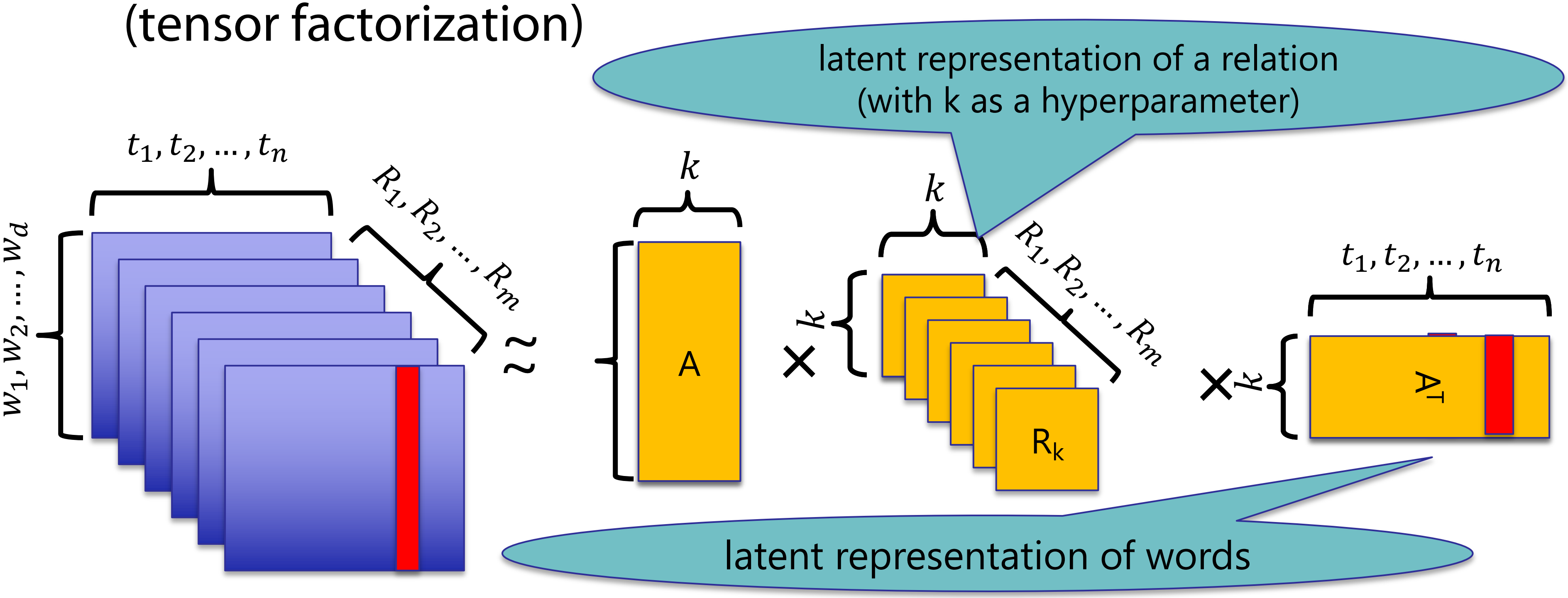
Tensor Decomposition – Analogy to SVD

- Derive a low-rank approximation to generalize the data and to discover unseen relations
- SVD



Tensor Decomposition – Analogy to SVD

- Derive a low-rank approximation to generalize the data and to discover unseen relations
- Apply Tucker decomposition and reformulate the results (tensor factorization)



Ledyard R. Tucker. "Some mathematical notes on three-mode factor analysis". Psychometrika. 31 (3): 279–311, **1966**.

Measure Degree of Relation: Raw Representation

- $ant(\text{joy}, \text{sadden}) = \cos\left(\boldsymbol{w}_{:,\text{joy},syn}, \boldsymbol{w}_{:,\text{sadden},ant}\right)$

	joy	gladden	sadden	felling
joyfulness	1	1	0	0
gladden	1	1	0	0
sad	0	0	1	0
anger	0	0	0	0

Synonym layer

	joy	gladden	sadden	felling
joyfulness	0	0	0	0
gladden	0	0	1	0
sad	1	0	0	0
anger	0	0	0	0

Antonym layer

Measure Degree of Relation: Raw Representation

- $ant(\text{joy}, \text{sadden}) = \cos(\boldsymbol{w}_{:,\text{joy},syn}, \boldsymbol{w}_{:,\text{sadden},ant})$

	joy	gladden	sadden	felling
joyfulness	1	1	0	0
gladden	1	1	0	0
sad	0	0	1	0
anger	0	0	0	0

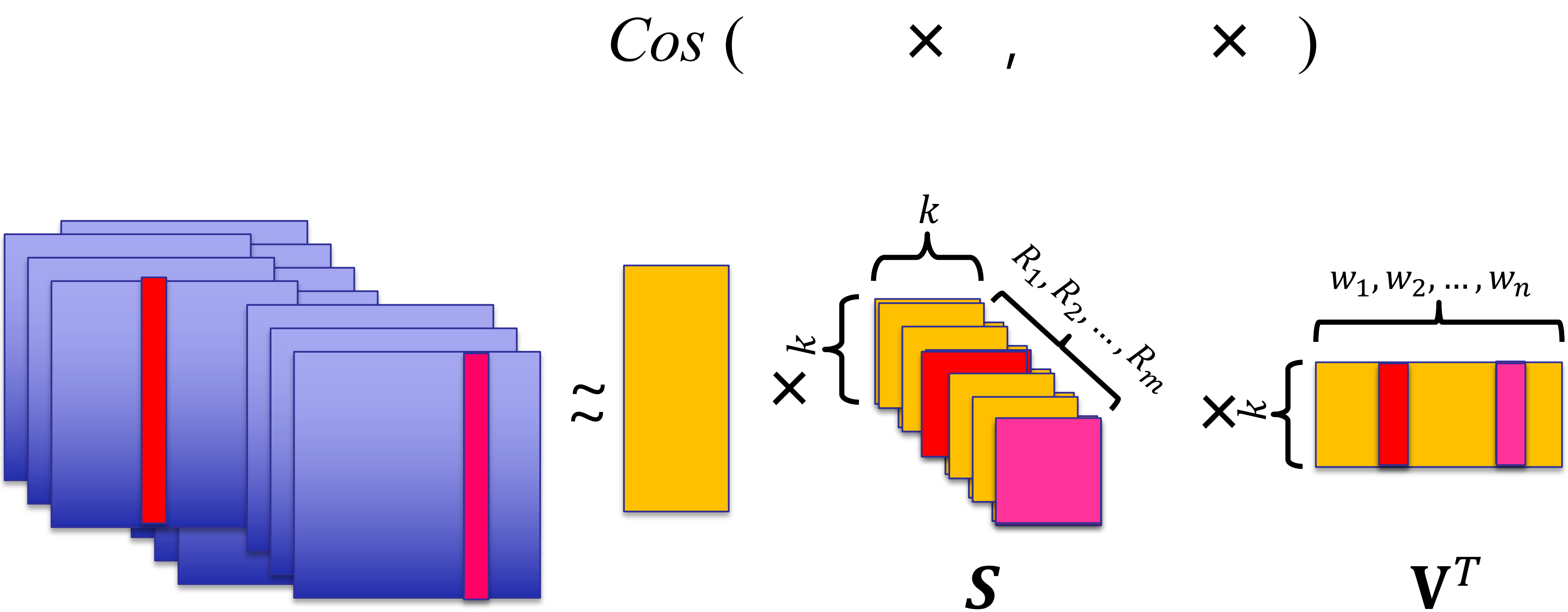
Synonym layer

	joy	gladden	sadden	felling
joyfulness	0	0	0	0
gladden	0	0	1	0
sad	1	0	0	0
anger	0	0	0	0

Antonym layer

Measure Degree of Relation: Latent Representation

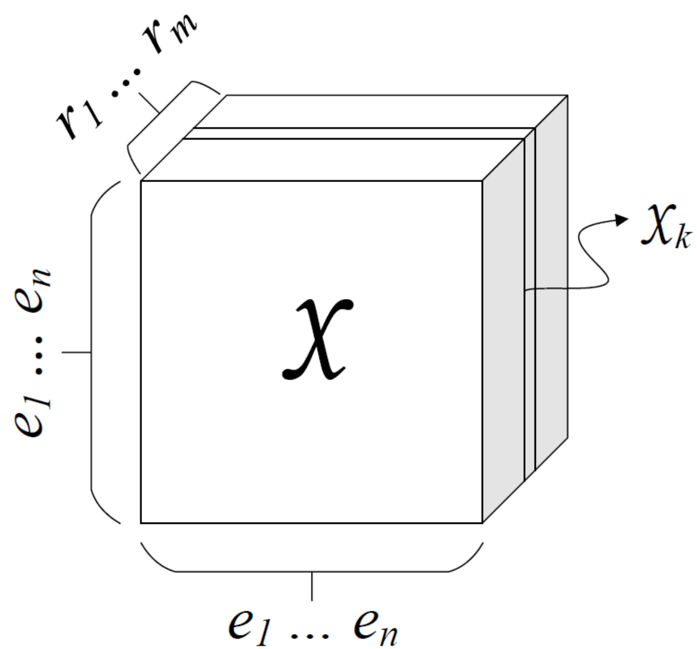
- $rel(\mathrm{w}_i, \mathrm{w}_j) = \cos(\boldsymbol{S}_{:,:,syn}\mathbf{V}_{i,:}^T, \boldsymbol{S}_{:,:,rel}\mathbf{V}_{j,:}^T)$



Knowledge Graphs (1/2)

- Collection of subj-pred-obj triples – (e_1, r, e_2)

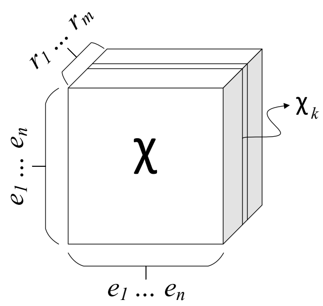
Subject	Predicate	Object
Obama	Born-in	Hawaii
Bill Gates	Nationality	USA
Bill Clinton	Spouse-of	Hillary Clinton
Satya Nadella	Work-at	Microsoft
…	…	…



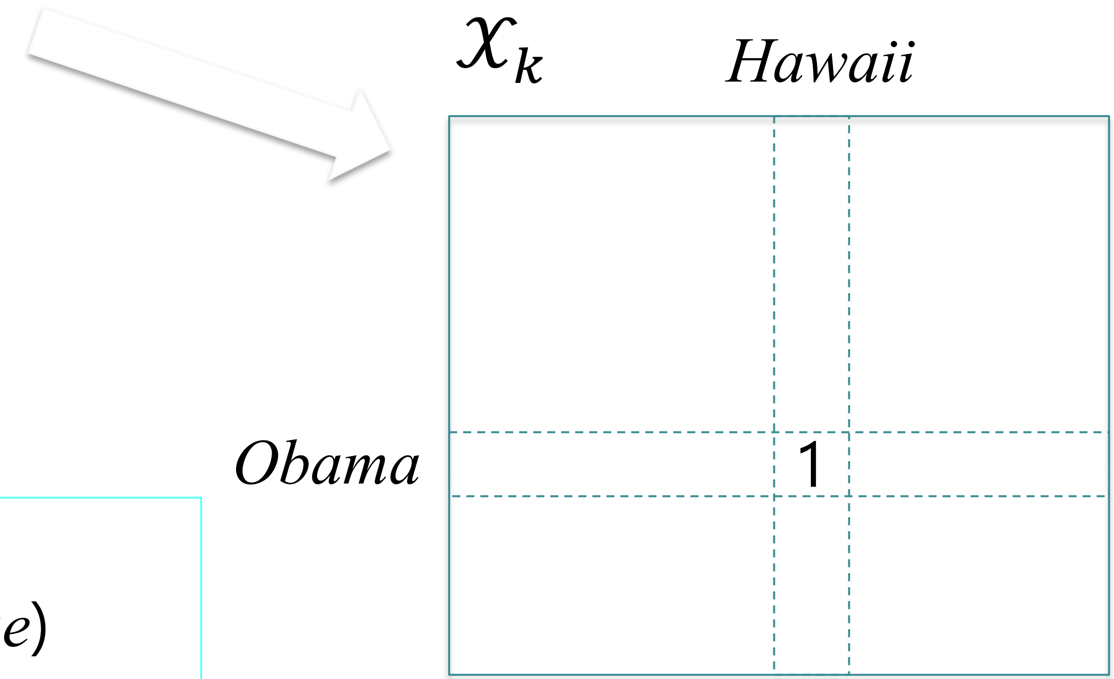
n: # entities, m: # relations

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Knowledge Graphs (2/2)



k-th slice



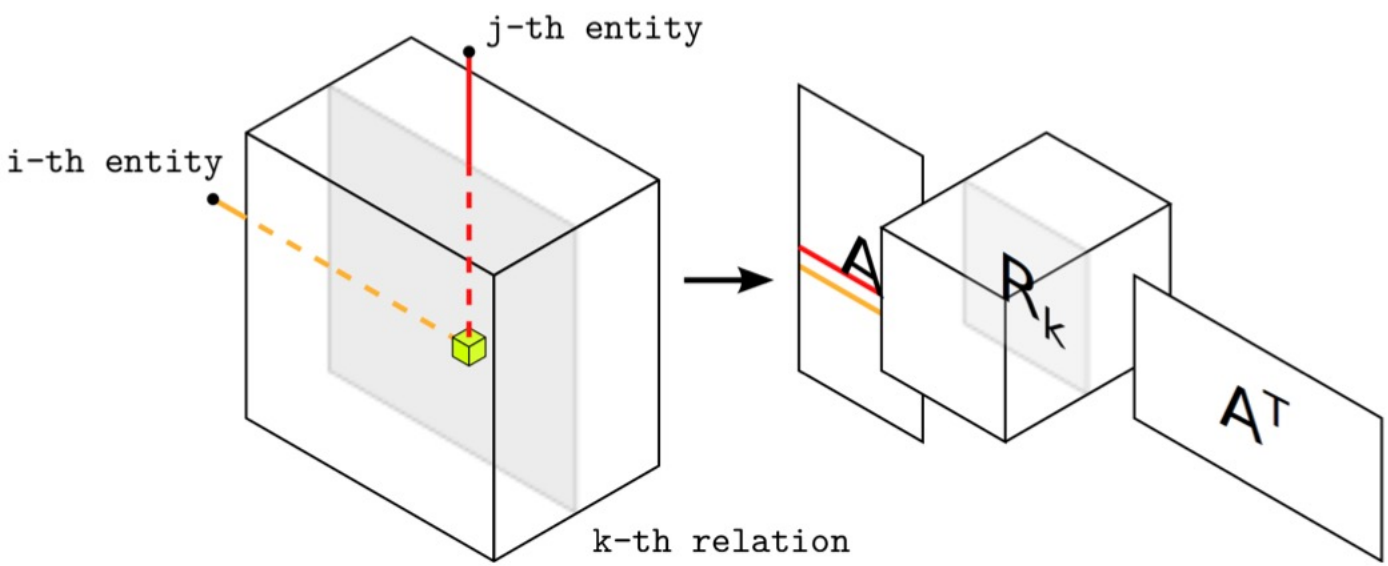
R_k : *born-in*

A 0 entry means:

- Incorrect (*false*)
- Unknown

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Factorization



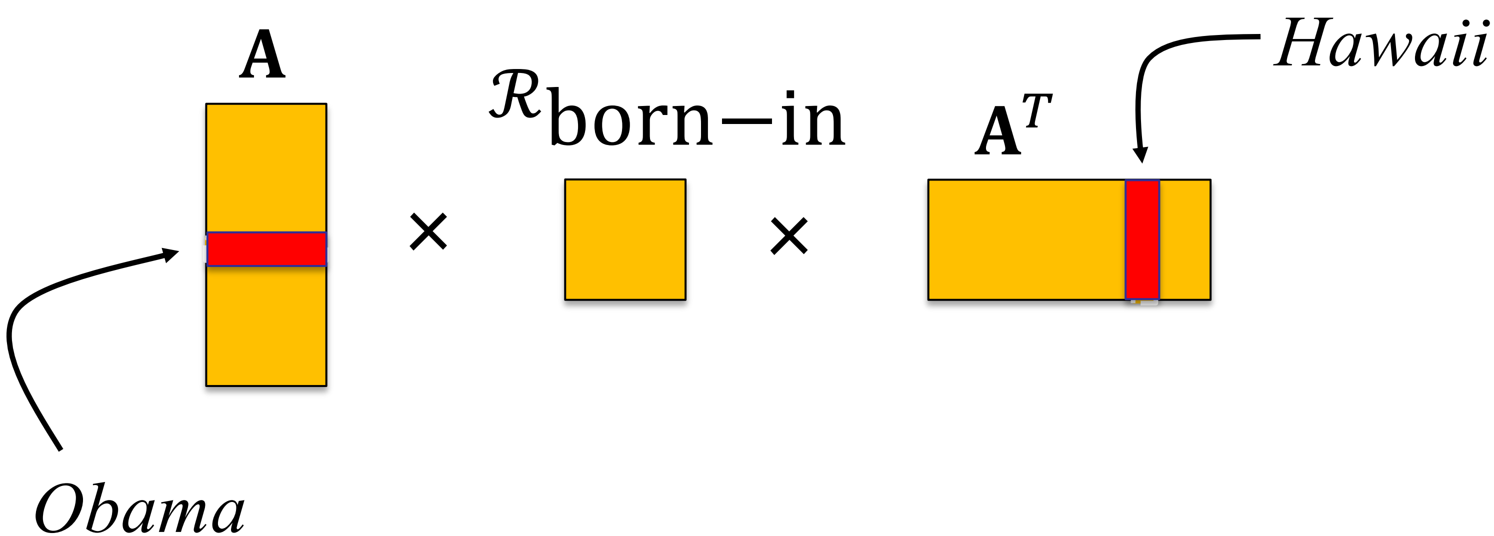
M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 809–816, **2011**.

Measure the Degree of a Relationship

$$f_{\text{born-in}}(\text{Obama}, \text{Hawaii})$$

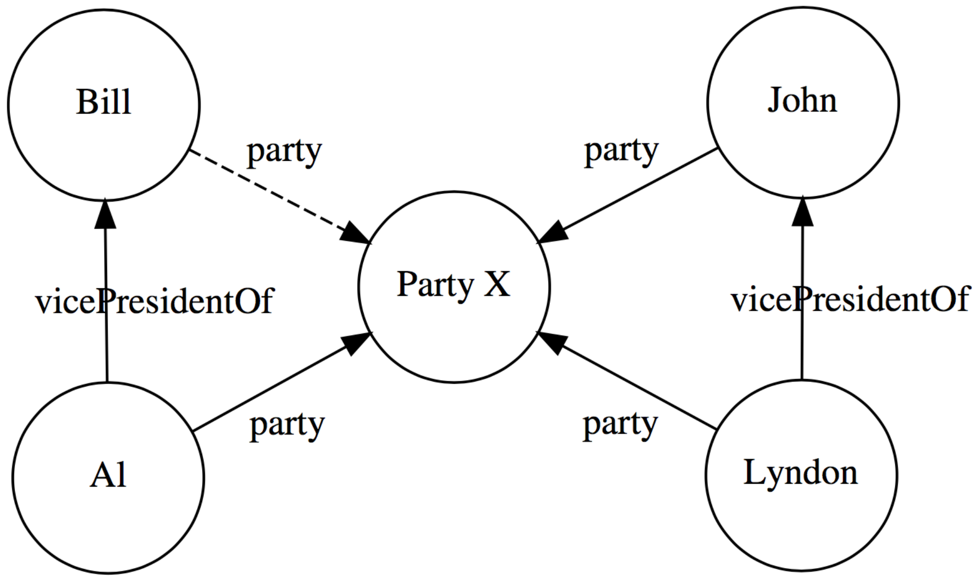
$$=$$

$$\mathbf{A}_{\text{Obama},:} \mathcal{R}_{\text{born-in}} \mathbf{A}^{\mathrm{T}}_{\text{Hawaii},:}$$



Prediction of Unknown Facts

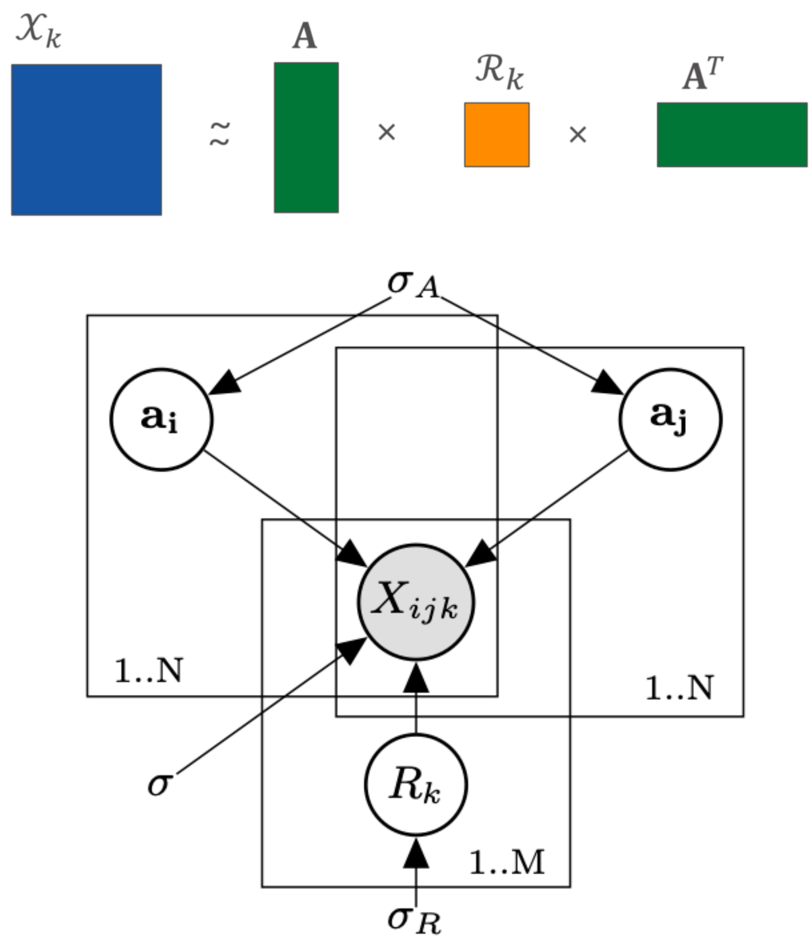
- Predict party membership of US (vice) presidents



Prediction of unknown fact party(Bill, Party X)

RESCAL: Graphical Model in Plate Notation

- Tensor factorization can be seen as a probabilistic model
  - Specified here in plate notation
- With appropriate CPTs, queries for the distribution $P(R(e_i, e_j))$ can be answered
- Can be used for prediction of unknown facts

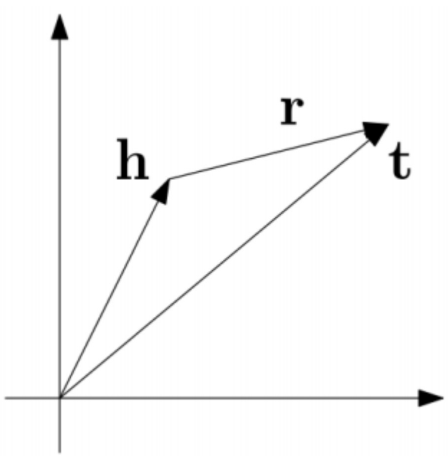


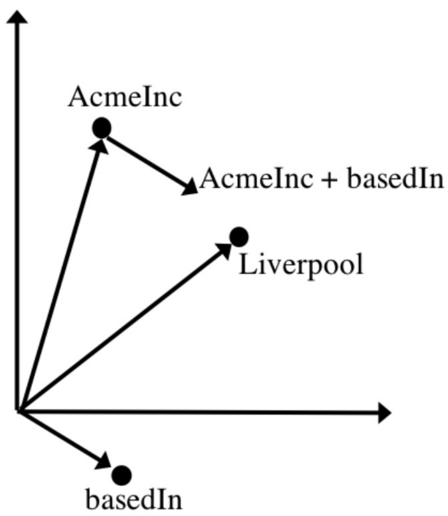
Nickel, M, Tresp, V, Kriegel, HP: Factorizing YAGO. Scalable Machine Learning for Linked Data. In Proceedings of the 21st International World Wide Web Conference, **2012**.

TransE: KG-Completion

- Inspired by word2vec

$$\mathrm{score}(\mathcal{R}_p(\mathrm{e}_s, \mathrm{e}_o)) = -\|\boldsymbol{e}_s + \boldsymbol{r}_p - \boldsymbol{e}_o\|_1$$

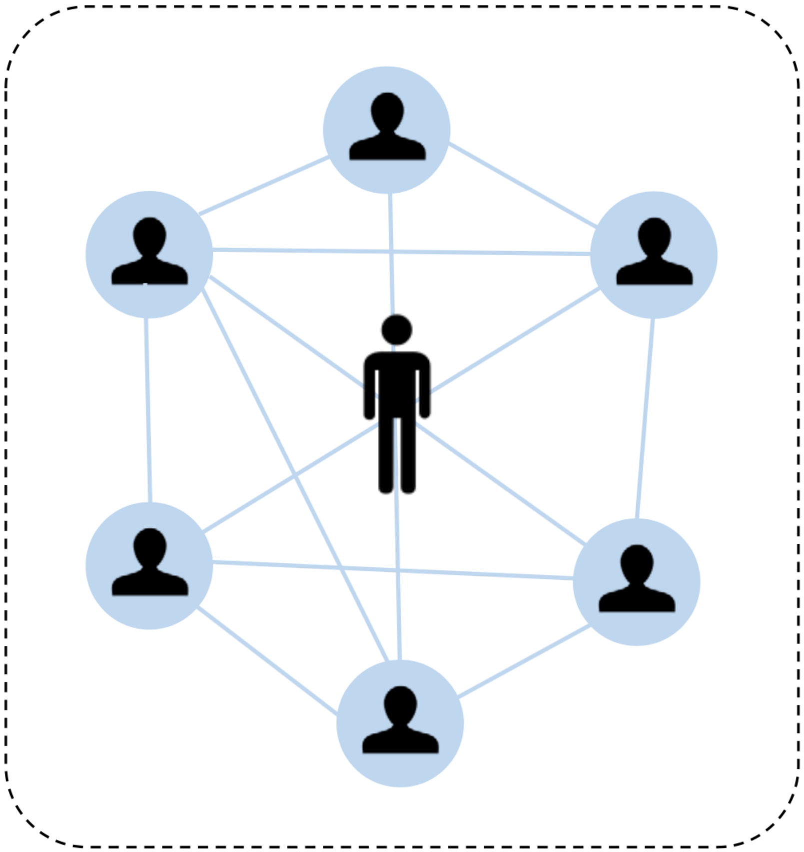


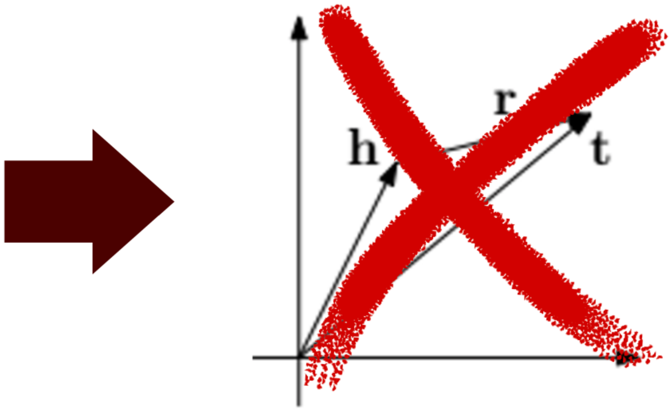


Learning objective: **h + r = t**

TransE: KG-Completion

However...





- In real world, we construct many relationships with many subjects.
- TransE can't represent more than one relationship between entities.

Summary

- Very many RESCAL- and TransE-like approaches for handcrafted embeddings of relational data
- None of the many approaches covers what's in a text
- Forget about handcrafted approaches