# PROBABILISTIC AND DIFFERENTIABLE PROGRAMMING

## V5: Embeddings

Özgür L. Özçep
**Universität zu Lübeck**
**Institut für Informationssysteme**

# Today‘s Agenda

From word embeddings (word2Vec)
to ontology embeddings

Or: Advertising Own Work

# WORD EMBEDDINGS WITH WORD2VEC

# Going deep in two other ways

- Text-related applications (such as document retrieval) require a deep, here: semantical, processing of text
- Word Embeddings achieve this by representing words as vectors in a low-dimensional continuous space
- Semantical similarity of words is reflected by nearness (e.g. cosine) of the words' vectors
- Prominent example is word2Vec (Mikolov et al. 13)
  - Learning in word2Vec based on a shallow feed forward network
  - But, the idea is deep: reading a window of words, relating each word with surrounding words (= context)

=> The idea of self-supervised learning

# Let LeCun speak again

*„I now call it „self-supervised learning" because unsupervised is both a loaded and confusing term. In self-supervised learning, the system learns to predict part of its input from other parts of its input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input."*

(somewhere at Facebook, cited according to Hare )

# Approaches for Representing Word Semantics

Beyond bags of words

**Distributional Semantics (*Count*)**

- Used since the 90's
- Sparse word-context PMI/PPMI matrix
- Decomposed with SVD

**Word Embeddings (*Predict*)**

- Inspired by deep learning
- `word2vec` *(Mikolov et al., 2013)*
- GloVe *(Pennington et al., 2014)*

Underlying Theory: **The Distributional Hypothesis** *(Harris, '54; Firth, '57)*

"Similar words occur in similar contexts"

https://www.tensorflow.org/tutorials/word2vec

https://nlp.stanford.edu/projects/glove/

# The Contributions of Word Embeddings

## Novel Algorithms

*(objective + training method)*

- Skip Grams + Negative Sampling
- CBOW + Hierarchical Softmax
- Noise Contrastive Estimation
- GloVe
- …

## New Hyperparameters

*(preprocessing, smoothing, etc.)*

- Subsampling
- Dynamic Context Windows
- Context Distribution Smoothing
- Adding Context Vectors
- …

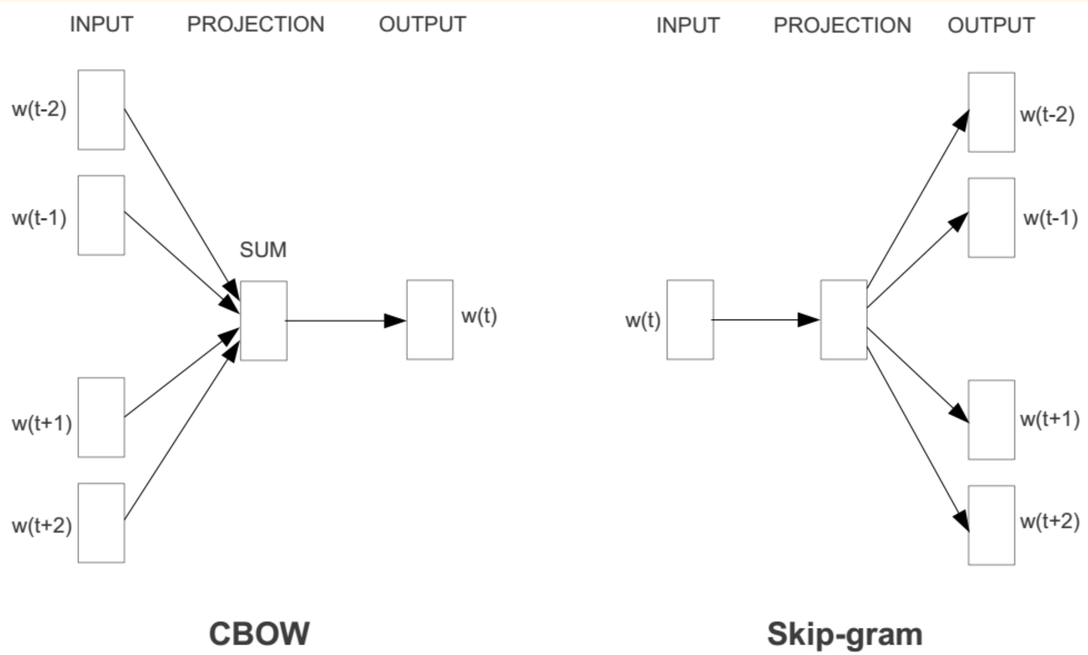
What's really improving performance?



Improving Distributional Similarity with Lessons Learned from Word Embeddings, Omer Levy, Yoav Goldberg, Ido Dagan

# **Represent** the meaning of **word** – word2vec

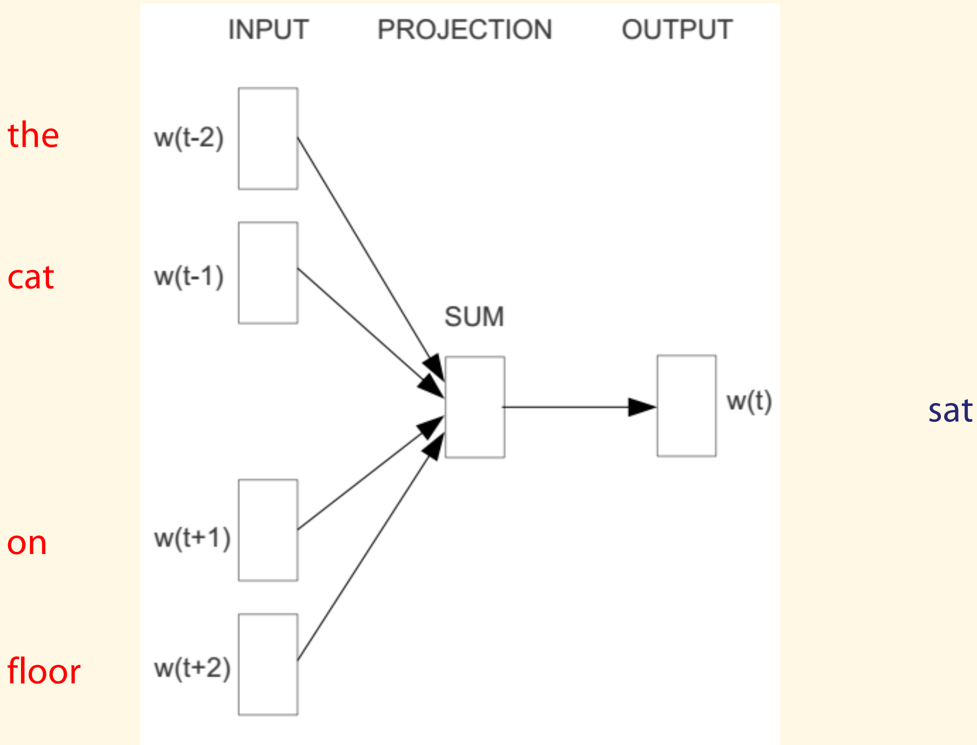
- 2 basic network models:
  - Continuous Bag of Word (CBOW): use a window of words to predict the middle word
  - Skip-gram (SG): use a word to predict the surrounding ones in window.

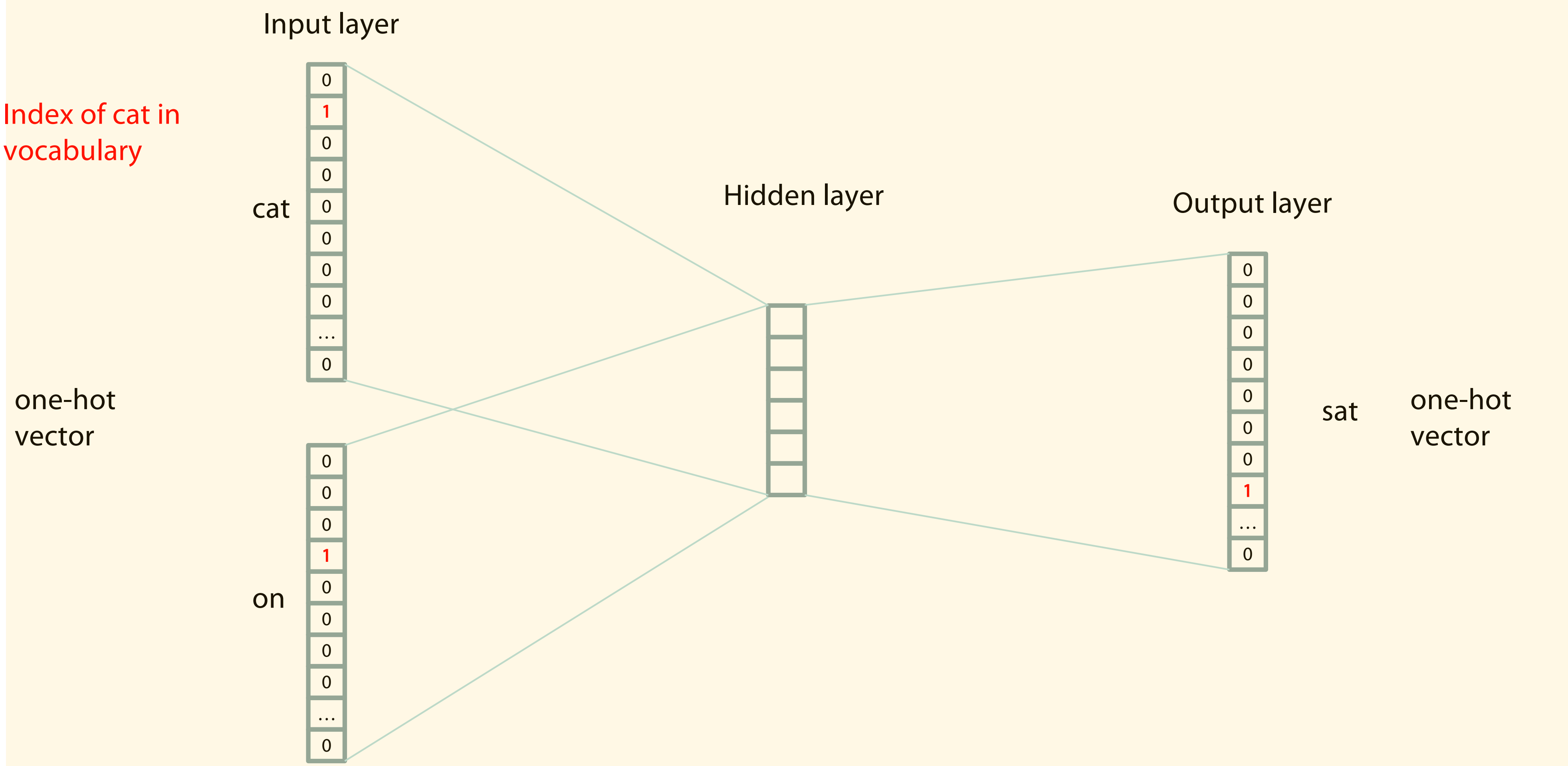


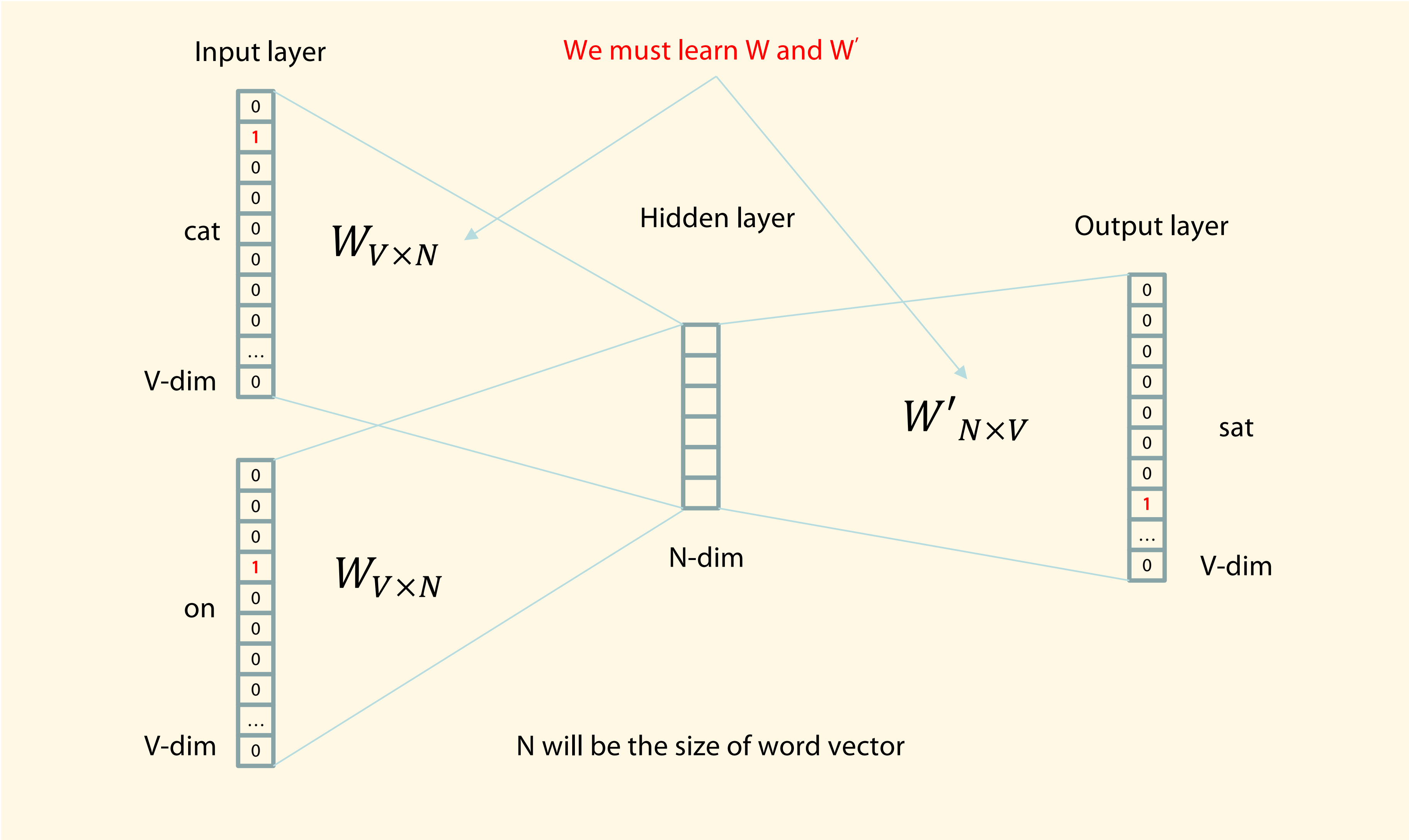

We shortly recap the CBOW mode. For SG see e.g. Goldberg/Levy 14

# Word2vec – Continuous Bag of Word (CBOW)

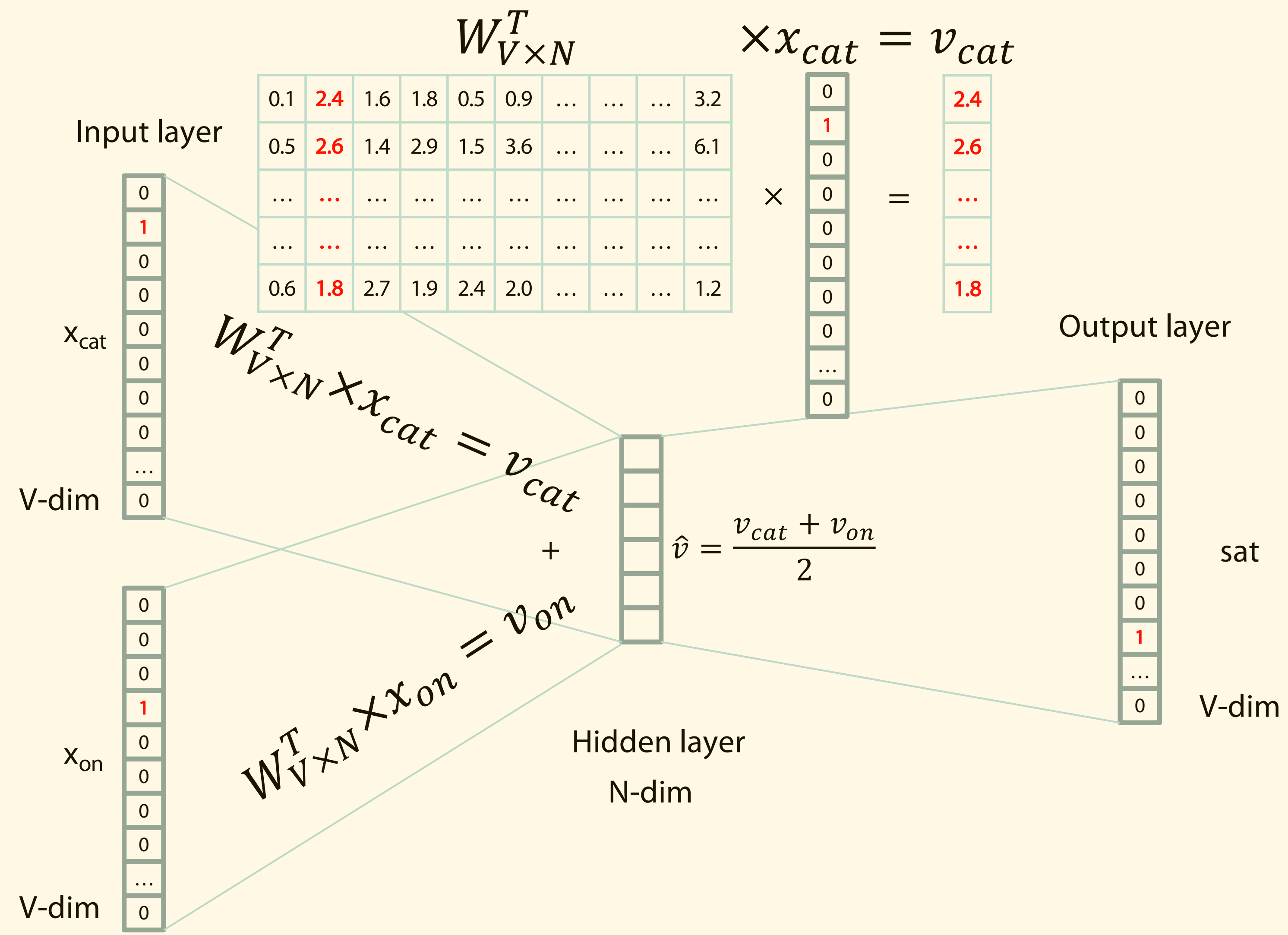
- E.g. "The cat sat on floor"
  - Window size = 2




Input layer
Index of cat in vocabulary
cat
0
1
0
0
0
0
0
0
…
0
Hidden layer
Output layer
0
0
0
0
0
0
0
1
…
0
sat
one-hot vector
one-hot vector
on
0
0
0
1
0
0
0
0
…
0

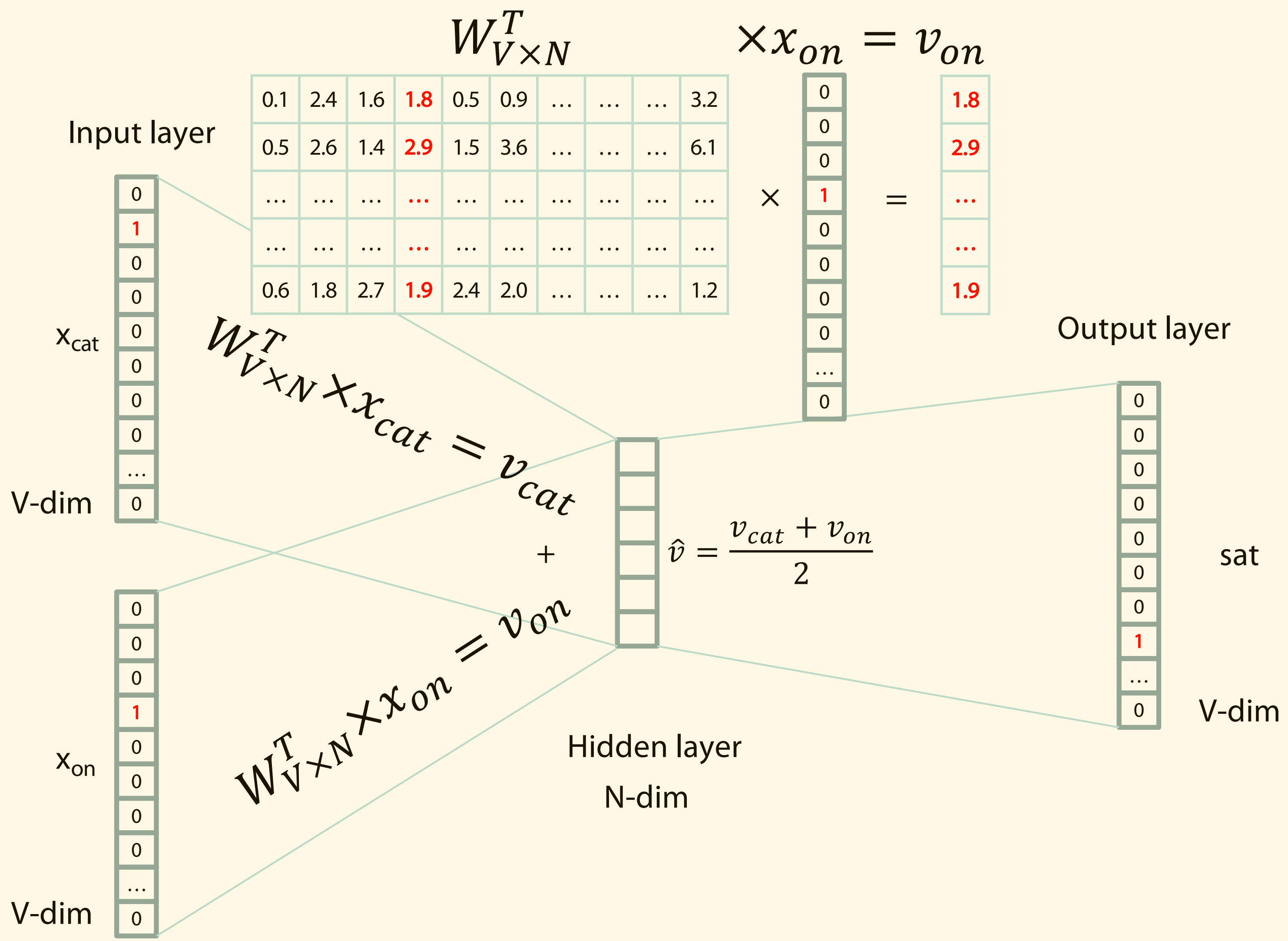


Input layer
We must learn W and W′
0
1
0
0
cat
0
0
0
0
…
V-dim
0
$W_{V\times N}$
Hidden layer
Output layer
0
0
0
0
0
0
0
1
…
0
$W'_{N\times V}$
sat
V-dim
0
0
0
1
on
0
0
0
0
…
V-dim
0
$W_{V\times N}$
N-dim
N will be the size of word vector

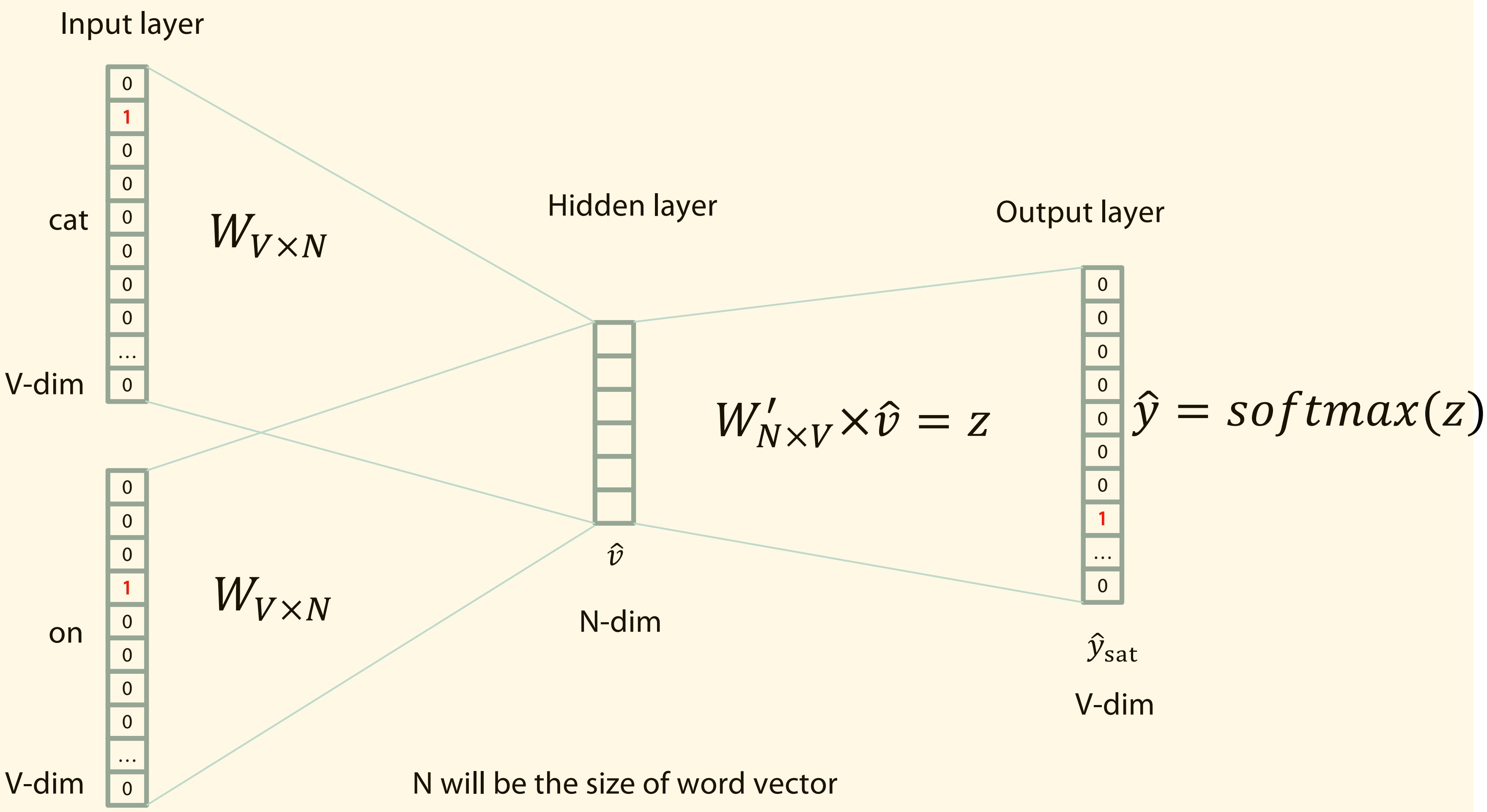



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 2.4 | 1.6 | 1.8 | 0.5 | 0.9 | … | … | … | 3.2 |
| 0.5 | 2.6 | 1.4 | 2.9 | 1.5 | 3.6 | … | … | … | 6.1 |
| … | … | … | … | … | … | … | … | … | … |
| … | … | … | … | … | … | … | … | … | … |
| 0.6 | 1.8 | 2.7 | 1.9 | 2.4 | 2.0 | … | … | … | 1.2 |

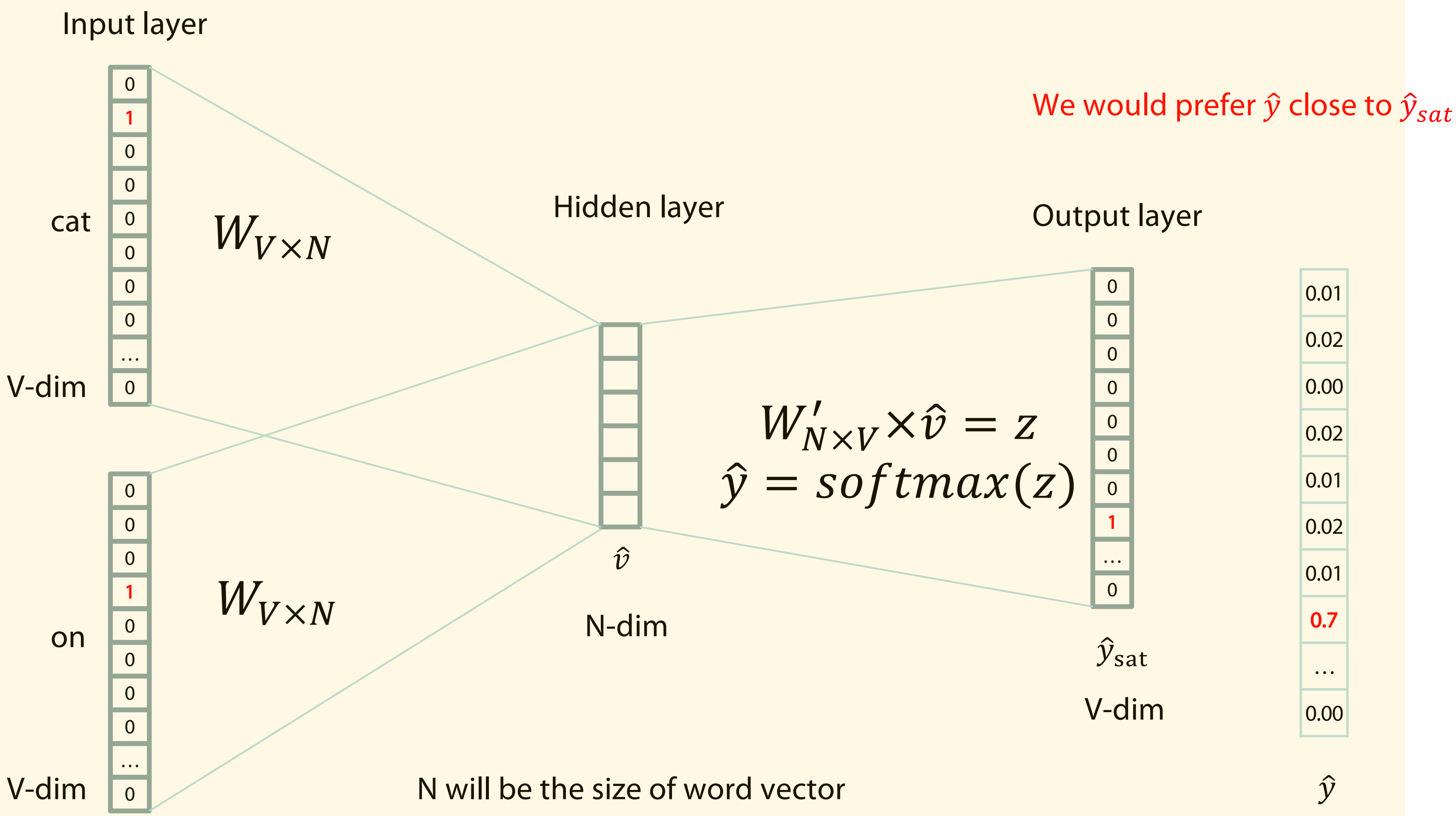
$\times$ $(0, 1, 0, 0, 0, 0, 0, 0, \ldots, 0)^T$ $=$ $(2.4, 2.6, \ldots, \ldots, 1.8)^T$

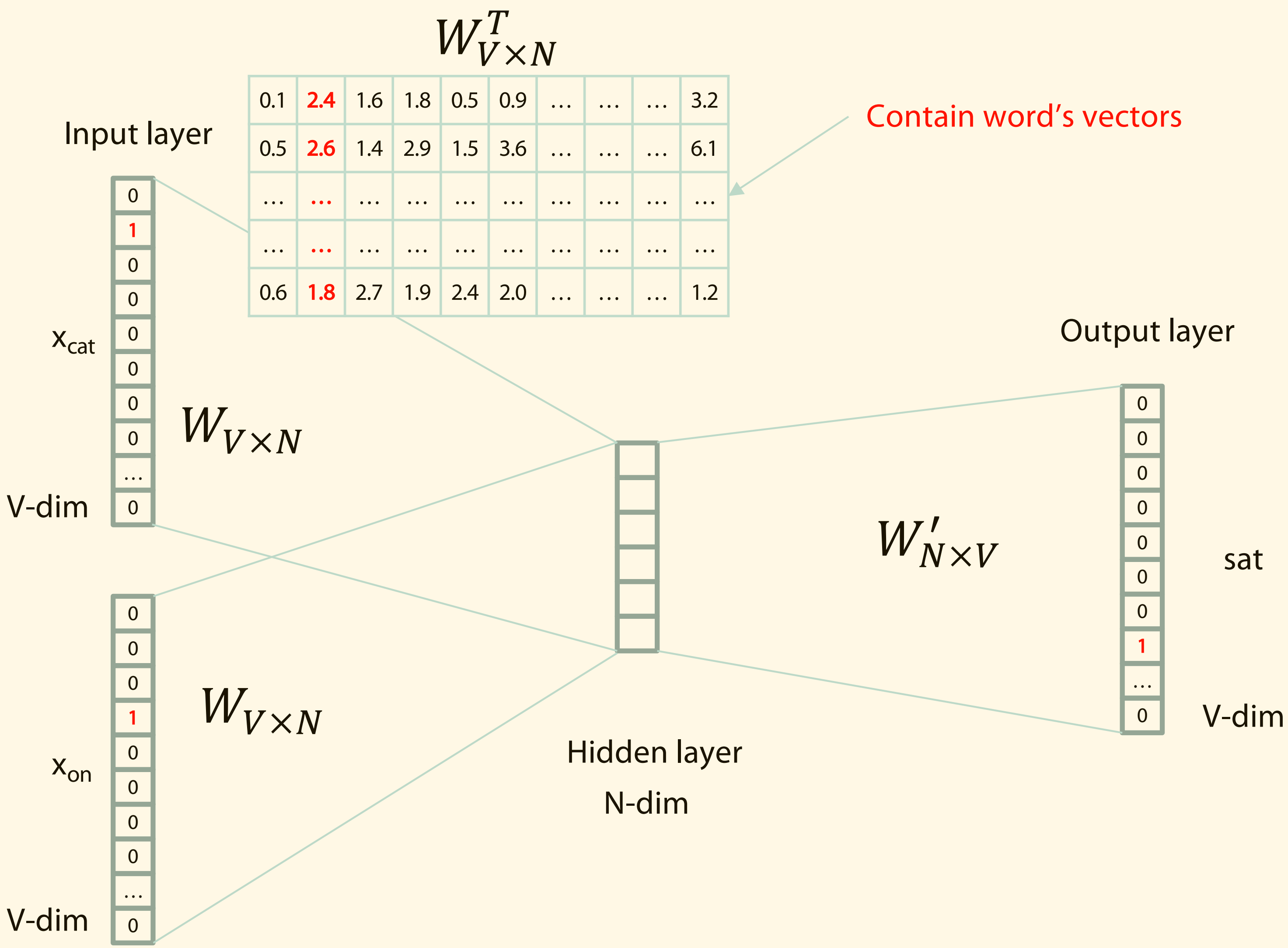


$W^T_{V\times N}$
$\times x_{on} = v_{on}$
Input layer
0.1 2.4 1.6 1.8 0.5 0.9 … … … 3.2
0.5 2.6 1.4 2.9 1.5 3.6 … … … 6.1
0.6 1.8 2.7 1.9 2.4 2.0 … … … 1.2
1.8
2.9
1.9
$x_{cat}$
V-dim
$W^T_{V\times N} \times x_{cat} = v_{cat}$
+
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
Output layer
sat
V-dim
$x_{on}$
V-dim
$W^T_{V\times N} \times x_{on} = v_{on}$
Hidden layer
N-dim




N will be the size of word vector


We would prefer $\hat{y}$ close to $\hat{y}_{sat}$
Input layer
cat
V-dim
on
V-dim
$W_{V\times N}$
$W_{V\times N}$
Hidden layer
$\hat{v}$
N-dim
$W'_{N\times V} \times \hat{v} = z$
$\hat{y} = softmax(z)$
Output layer
$\hat{y}_{sat}$
V-dim
$\hat{y}$
0.01
0.02
0.00
0.02
0.01
0.02
0.01
0.7
…
0.00
N will be the size of word vector




We can consider either W ("context-word role") or W′ ("middle word role") as the word's representation.

# Word Analogies

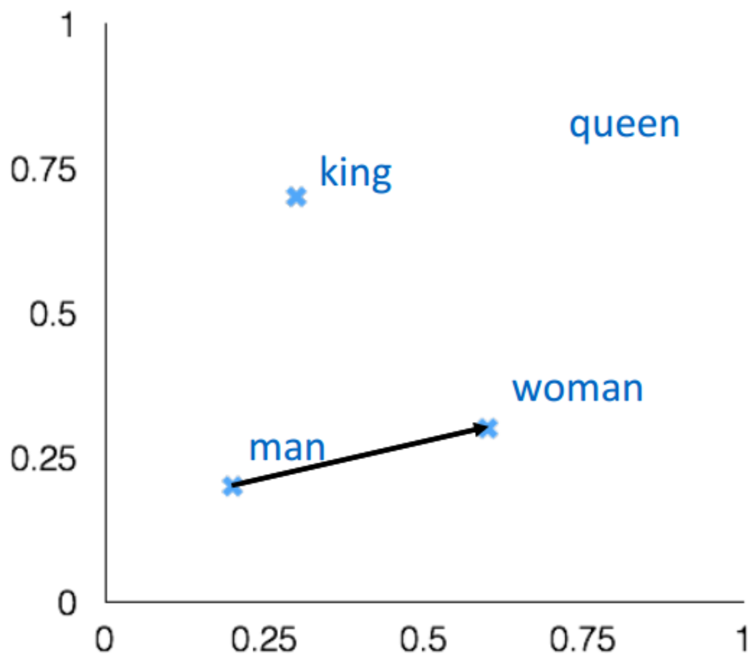
Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:?

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

| | | |
|---|---|---|
| + | king | [ 0.30 0.70 ] |
| - | man | [ 0.20 0.20 ] |
| + | woman | [ 0.60 0.30 ] |
| | queen | [ 0.70 0.80 ] |



# What is word2vec?

- Word2vec **is not** a single algorithm
- It is a software package for representing words as vectors, containing:
  - Two distinct models
    - CBoW
    - Skip-Gram (SG)
  - Various training methods
    - Negative Sampling (NS)
    - Hierarchical Softmax
  - A rich preprocessing pipeline
    - Dynamic Context Windows
    - Subsampling
    - Deleting Rare Words

# KNOWLEDGE GRAPH EMBEDDING

# Knowledge Graph (KG)

- Set of assertions in triple form $KG = \{ \ (subj\ rel\ obj\ ) \ \}$
- Convenient logical notation $KG = \{ \ rel(subj, obj) \ \}$

**Example**

$KG = \{ \text{worksIn(alice,AI), subfield(AI,CS),manyPubls(alice,bob)} \}$
$\cup \{ \text{worksIn(bob,AI)} \}$

- Usually KGs highly incomplete
- ⇒ “Learn” new triples assuming regularities in data

# Knowledge Graph Embedding

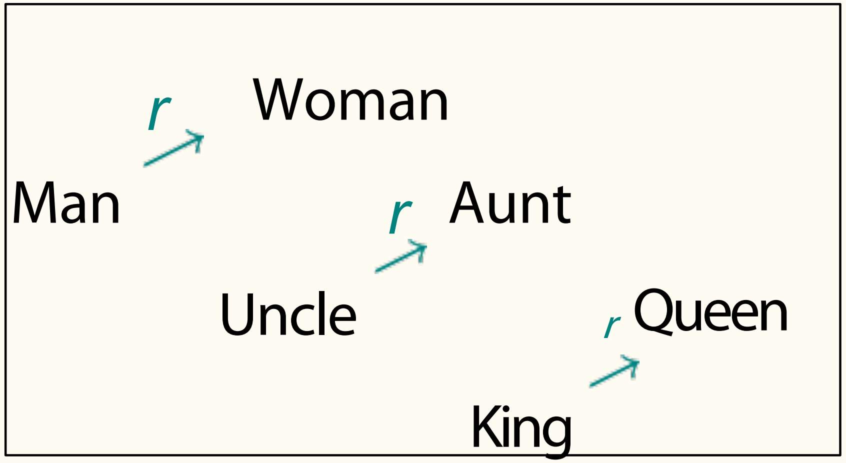
**Main Idea (for embedding/interpretation $(\ \ )^I$)**

- Learn
  1. vector representation $o^I$ of objects $o$ in (low-dimensional) continuous space $E = \mathbb{R}^n$
  2. scoring function $\mathrm{r}^I = s_r : E \times E \rightarrow \mathbb{R}$ for relations $r$
- Completion: If $s_r(o_1^I, o_2^I)$ small, add $r(o_1, o_2)$ to KG.

- Various approaches differing in type of $s_r$
  - TRansE (Bordes et al. 13), TransR (Lin et al. 15), STransE (Nguyen et al. 16)
  - DistMult (Yang et al. 15)
  - ComplEx (Trouillon et al. 16)
  - SimplE (Kazemi/Poole 18)
  - RESCAL (Nickel et al. 11)

# Problems of Classical Embeddings

## Example (TransE (Bordes et al. 13))




- $s_r(u, v) = \|u + r - v\|$ ($\|\cdot\|$: Euclidean Norm)
- Limitation: Relations $r$ = vector translations, hence functional

- Similar problems for other embedding approaches: not fully expressive. (Kazemi/Poole 18)

# Expressivity Criterion

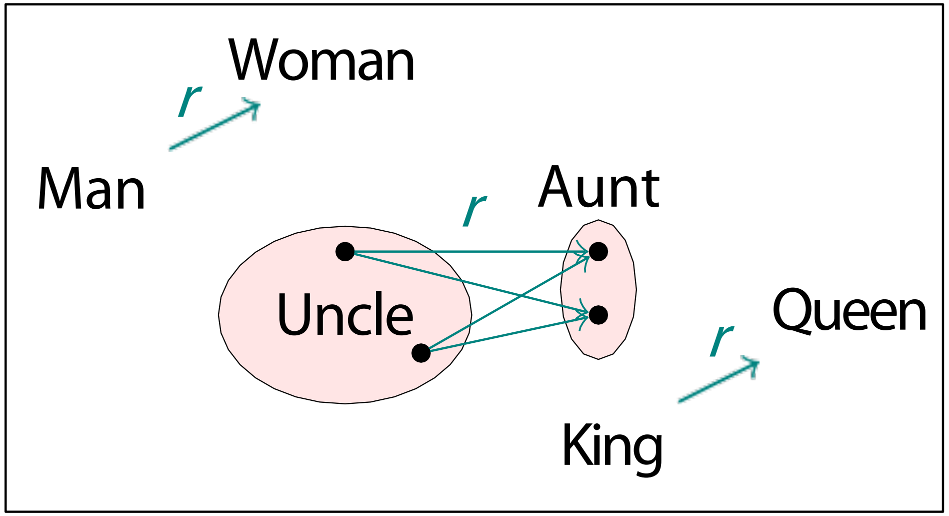
- Usually one considers low-dimensional spaces
- But nonetheless must be sufficiently high to embed knowledge expressed in KGs
  - $P$: set of valid triples in KG
  - $N$: set of non-valid triples in KG $(N \cap P = \emptyset)$

**Definition (Kazemi/Poole 18)**

An embedding model is fully expressive iff there are a dimension $n$, an embedding $e$, and a threshold $\lambda_R$ such that:

- for all $R(u,v) \in P$: $s_R(e(u), e(v)) \leq \lambda_R$
- for all $R(u,v) \in N$: $s_R(e(u), e(v)) > \lambda_R$

# Solution: Logico-geometrical Semantics




- Represent concepts as geometrically shaped sets (sets of vectors, not single vector)

- Represent binary relations as geometrically shaped sets of pairs of objects

- Benefit: Can add background kowledge

# Adding Background Knowledge

## Example

$$\{\ \forall X, Y, Z. subfield(X, Y) \land worksIn(Z, X) \rightarrow worksIn(Z, Y)\ \}$$

KG = {worksIn(alice,AI), subfield(AI,CS), manyPubls(alice,bob),

worksIn(bob,AI), (by induction)

worksIn(alice,CS) (by deduction)}

Adding background knowledge really means a benefit...

# In Favour of Controlled Explainable AI

**nature machine intelligence** View all Nature I

**Explore our content ⌄** **Journal information ⌄**

nature > nature machine intelligence > perspectives > article

Perspective | Published: 13 May 2019

## Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

*Nature Machine Intelligence* **1**, 206–215(2019) | Cite this article

**29k** Accesses | **206** Citations | **239** Altmetric | Metrics

NOT: Explain parameter tweaking

BUT: More control via background knowledge/ontology

## Research program

Find appropriate pairs

background knowledge language / geometrical structure for

knowledge graph embedding

- Quasi-chained Datalog / convex (Gutierrez-Basulto, Schockaert 18)
- $\mathcal{EL}^{++}$ / spheres (Kulmanov et al. 19)
- $\mathcal{ALC}$ / convex al-cones (this lecture)
- $L_{cone}$ / convex cones (this lecture)

# Agenda for the Following

- Description Logics
- Convex Cones
- Faithful Embedding for Al-cones
- Towards a Logic of Convex Cones

# DESCRIPTION LOGICS

## Definition (Description logics (DLs))

Logics for use in knowledge representation with special attention on a good balance of expressibility and feasibility of reasoning services

- Can be mapped to fragments of FOL
- Usage
  - Ontology representation language
  - Foundation for standard web ontology language (OWL)
- Have been investigated for ca. 30 years now
  - Many theoretical insights on various different purpose DLs
  - Various reasoners

# Description Logics (DLs)

## Example (ALC Concepts)

- $Students$ "Students"
- $Students \sqcap Male$ "Male students"
- $Female \sqcap Male$ "Female or Male"
- $\exists attends.LogicCourse$ "Those attending a logic course"
- $\neg\exists attends.LogicCourse$ "Those not attending a logic course"

Full concept negation allowed in DL $\mathcal{ALC}$ .

# Tbox and Abox

- Terminological box (tbox): { concept inclusions }

**Example**

$\{\ MasterStudent \sqsubseteq\ Student, Student\ \sqcap Professor \sqsubseteq \bot \}$

- Assertional box (abox): { assertions }

**Example**

$\{\ MasterStudent(peter), attends(peter, CS4711)\ \}$

- Ontology: tbox $\cup$ abox

# Example (Certain Answers for Conjunctive Queries)

- $T \ = \{ \ \top \sqsubseteq Male \sqcap Female, \ \ Male \sqcap Female \sqsubseteq \bot \}$
- $A \ \ = \{ \ friend(john, susan), \ friend(john, andrea), \ female(susan),$
  $likes(susan, andrea), likes(andrea, bill), Male(bill) \ \}$
- $Q(x) = \ \exists y, z(friend(x, y) \ \wedge Female(y) \ \wedge likes(y, z) \ \wedge Male(z))$

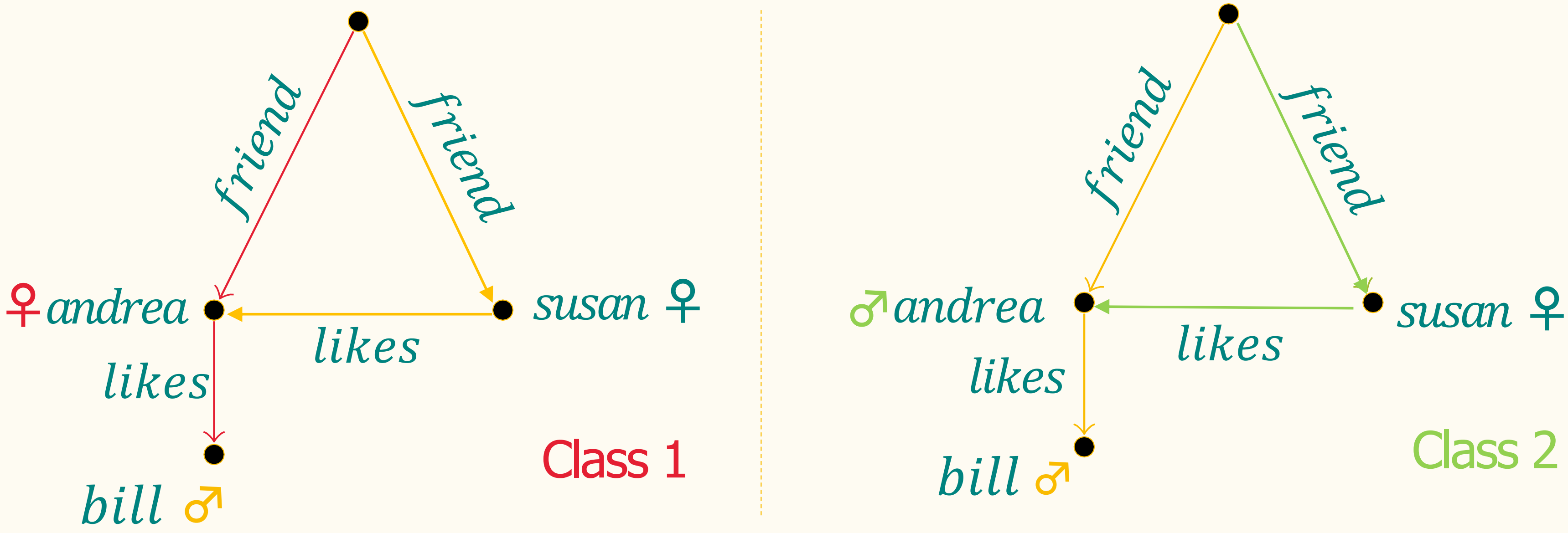
---

- $cert(Q(x), O) \ \ =?$

- We have to consider all possible models of the ontology
- But here there are actually two classes: Andrea is male vs. Andrea is not male.

## Example (Certain Answers for Conjunctive Queries)

- $T = \{ \top \sqsubseteq Male \sqcap Female, \ Male \sqcap Female \sqsubseteq \bot \}$
- $A = \{ friend(john, susan), \ friend(john, andrea), \ female(susan), likes(susan, andrea), likes(andrea, bill), Male(bill) \}$
- $Q(x) = \exists y,z(friend(x,y) \wedge Female(y) \wedge likes(y,z) \wedge Male(z))$




$cert(Q(x), O) = \{john\}$

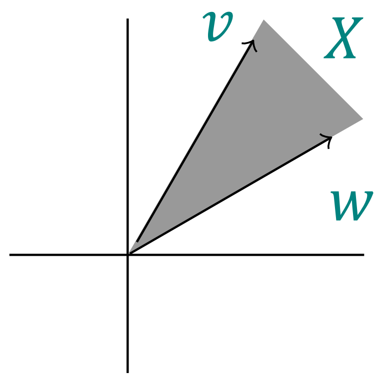
# CONVEX CONES

# Idea: Interpret Concepts by Convex Cones

- $X \subseteq \mathbb{R}^n$ is a convex cone iff for all

  $v, w \in X, \lambda, \mu \in R_{\geq 0}: \lambda v + \mu w \in X$

- $hull(X)$ = smallest cone containing $X$

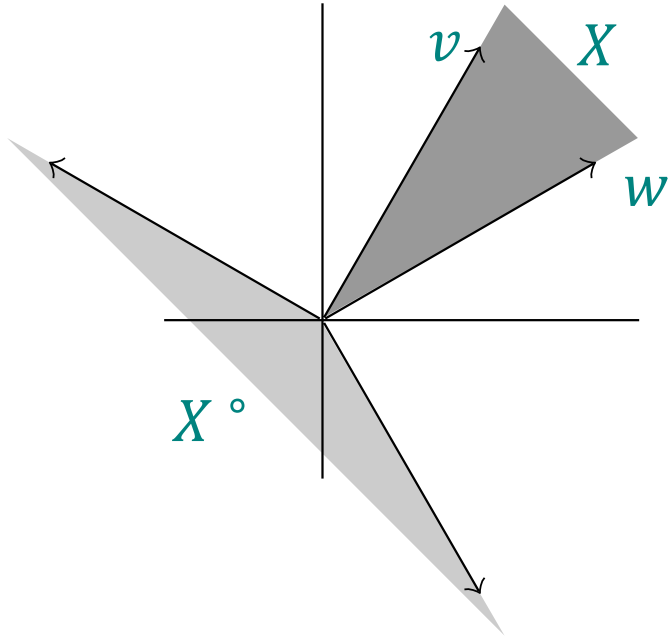


# Idea: Interpret Concepts by Convex Cones

- $X \subseteq \mathbb{R}^n$ is a convex cone iff for all

  $v, w \in X, \lambda, \mu \in \mathbb{R}_{\geq 0}: \lambda v + \mu w \in X$
- $hull(X)$ = smallest cone containing $X$
- Polar cone

  $X^\circ = \{v \in \mathbb{R}^n | \forall w \in X, v \cdot w \leq 0\}$




**Proposition**

*For closed convex cones $X$, $Y$:*

- $X^\circ$ *is a closed convex cone*
- $(X^\circ)^\circ = X$
- $hull(X \cup Y) = (X^\circ \cap Y^\circ)^\circ$

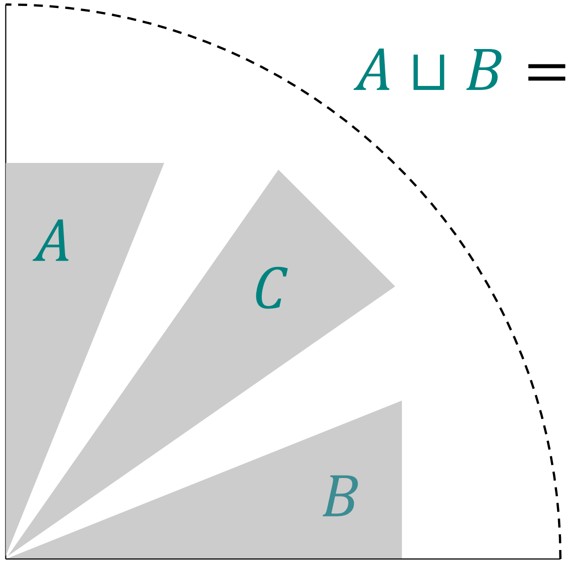
Why convex cones ?

- $\neg X = X^\circ$

  $X \sqcap Y = X \cap Y$

  $X \sqcup Y = hull(X \cup Y)$
- Convex/conic optimization

# Not all Cones Appropriate for $\mathcal{ALC}$




$A \sqcup B = hull(A \cup B)$

$$C \sqcap (A \sqcup B) = C$$
$$\neq (C \sqcap A) \sqcup (C \sqcap B) = \bot$$

- Distributivity law not fulfilled
- What should we do?
  1. Restrict family of cones (in the following)
  2. Search for a (the) logic of cones (thereafter)

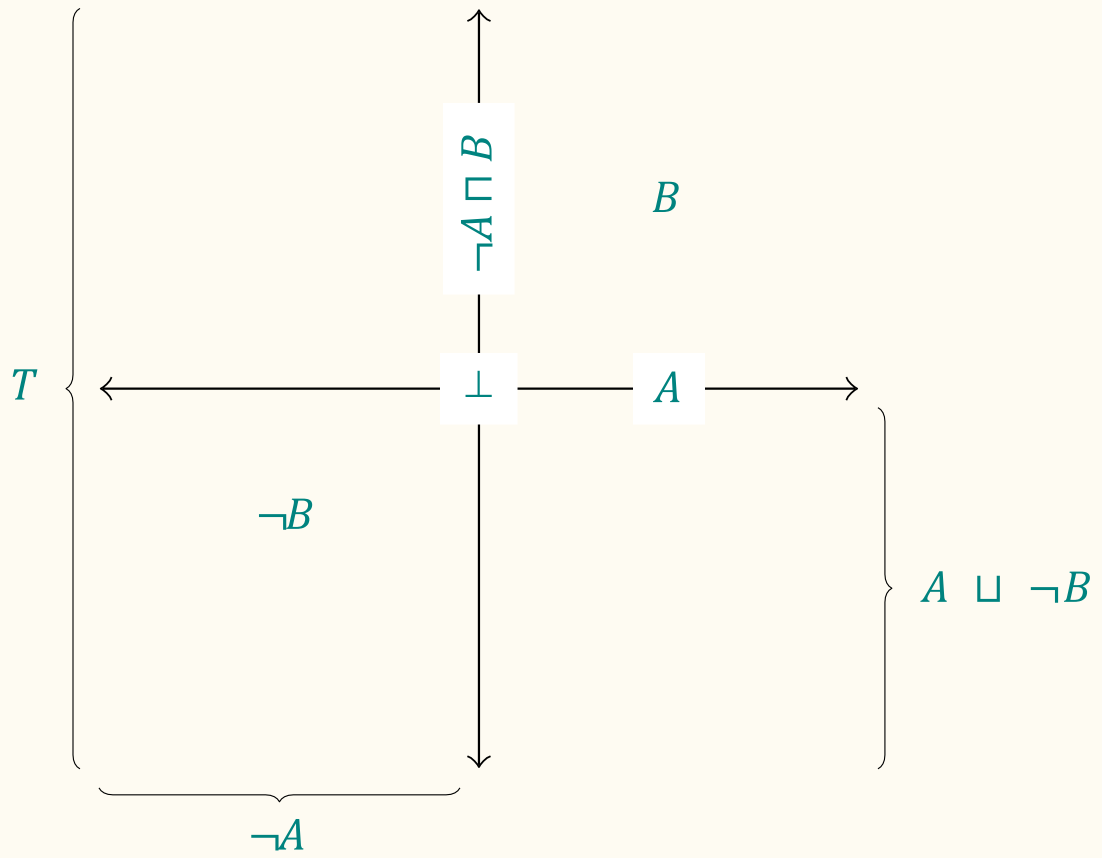
# FAITHFUL EMBEDDING FOR AL-CONES

# Searching for ALC-cones (Nomen est Omen)

**Definition (Axis-aligned Cone (al-cone))**

$X$ is an al-cone in $\mathbb{R}^n$ iff $X = X_1 \times \cdots \times X_n$
where each $X_i \in \{\mathbb{R}, \mathbb{R}_+, \mathbb{R}_-, \{0\}\}$

## Example (Embedding for tbox $\{A \sqsubseteq B\}$)



# Al-Cones are Appropriate for Boolean $\mathcal{ALC}$

**Proposition (Ö., Leemhuis, Wolter 2020)**

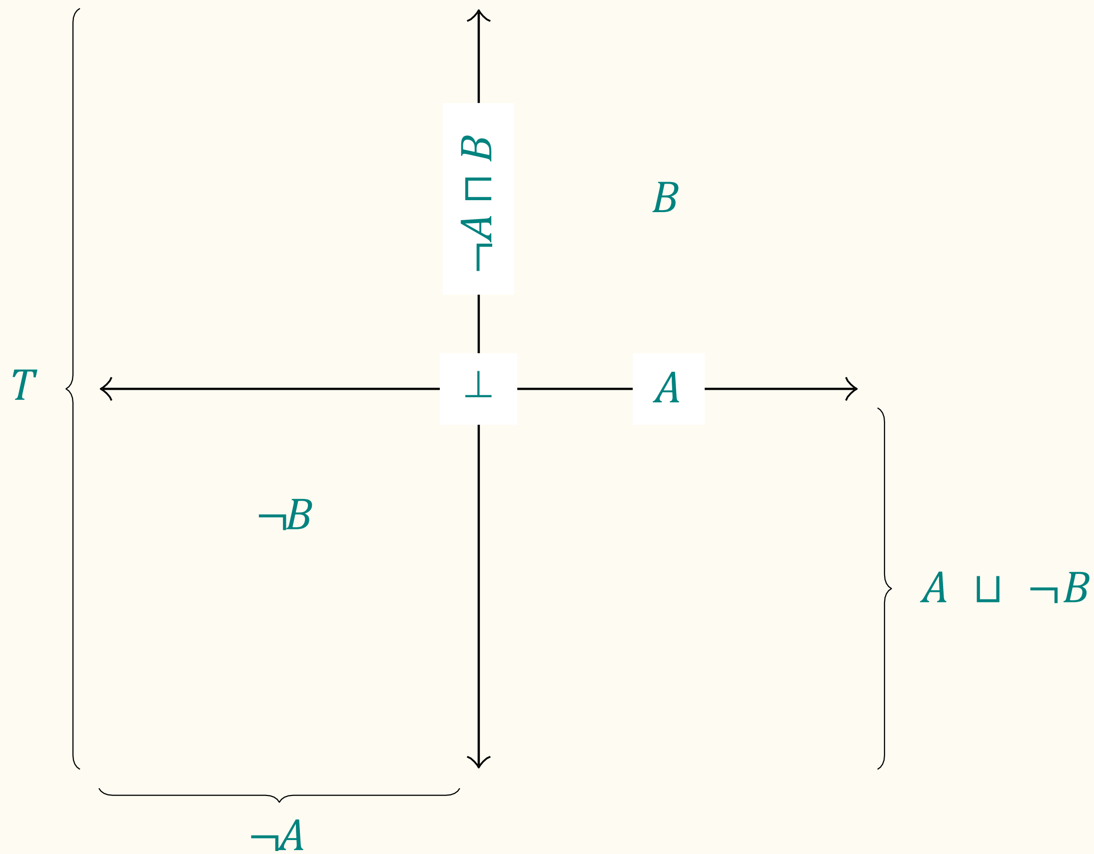
*For Boolean*[1] $\mathcal{ALC}$ *-ontologies:*

*classically satisfiable = satisfiable with faithful al-cone model*

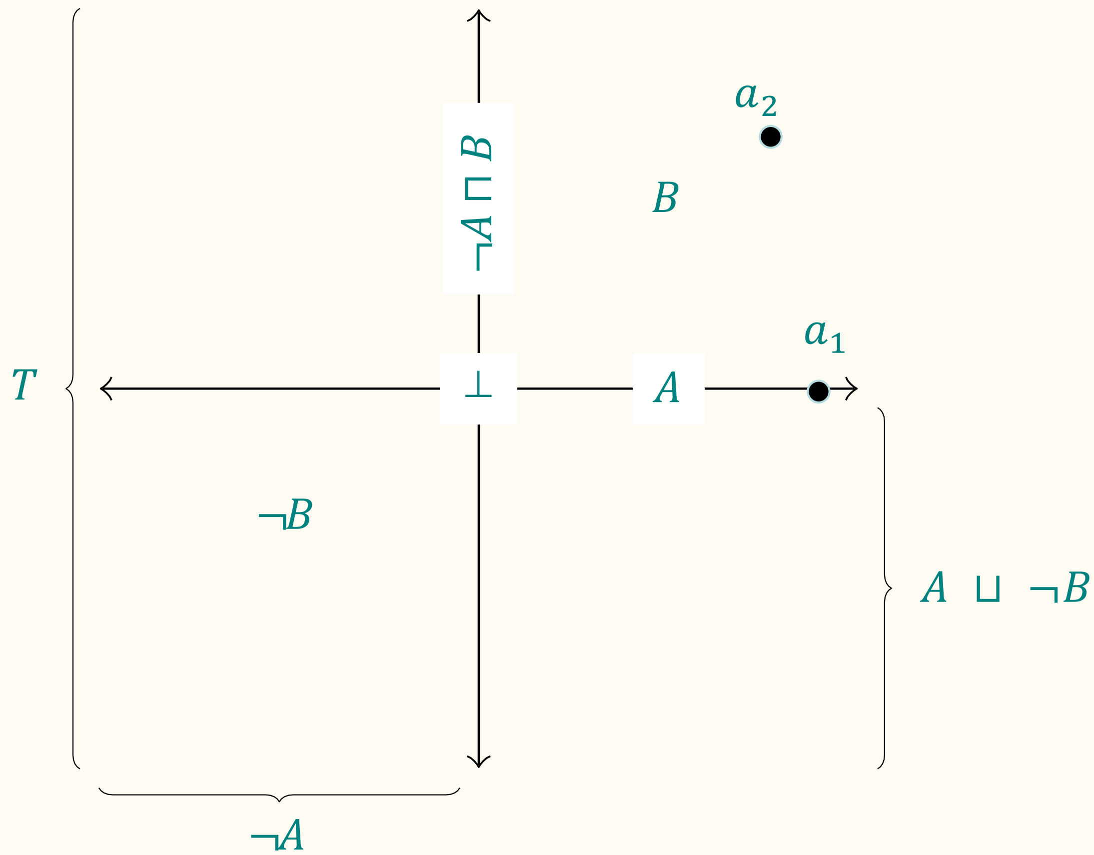
What does faithfulness mean here?

[1] No roles (binary relations) allowed.

# Example (Embedding for tbox $\{A \sqsubseteq B\}$)

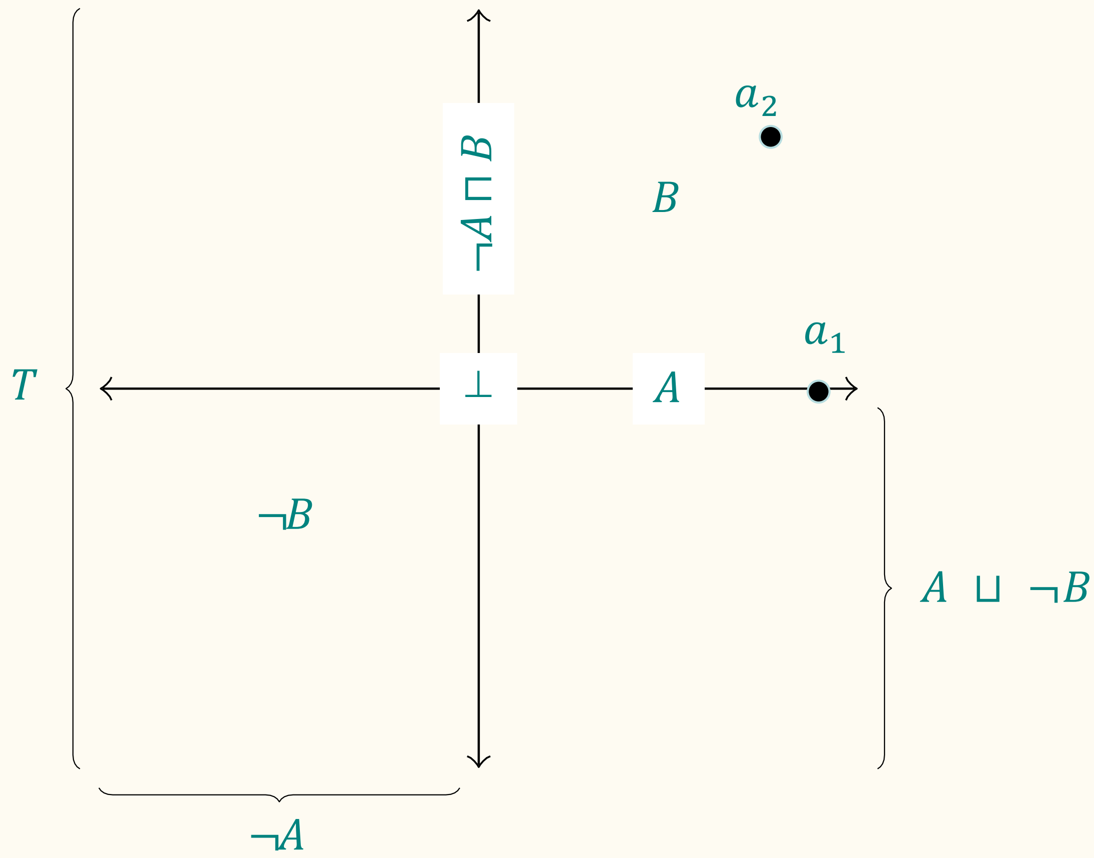


# Example (Embedding for tbox $\{A \sqsubseteq B\}$)




Our geometric models are partial: can model uncertainty in ontology ($a_2$ not known to be $A$ or $\neg A$; in contrast: $a_1$ completely determined )

# Example (Embedding for tbox $\{A \sqsubseteq B\}$)




**Faithfulness:** $o^I \in C^I$ iff ontology entails $C(o)$.

# Al-Cone models for full $\mathcal{ALC}$

**Proposition (Ö., Leemhuis, Wolter 2020)**

*For possibly non-Boolean $\mathcal{ALC}$ -ontologies up to some rank[1]:*

*classically satisfiable = satisfiable with faithful al-cones model*

- This is an approximative solution
- Relation not interpreted geometrically (not conic)

[1] Rank = nesting depth of quantifiers in a formula

# TOWARDS A LOGIC OF CONES

# Searching for an Appropriate Logic

**Definition (Minimal orthologic (Goldblatt 74))**

Minimal orthologic is characterised by $\sqcap, \sqcup$ fulfilling rules of a lattice and the existence of an orthonegation $\neg$

- If $A \sqsubseteq B$ then $\neg B \sqsubseteq \neg A$
- $\neg\neg A \sqsubseteq A$
- $A \sqcap \neg A \sqsubseteq \bot$

**Proposition (Leemhuis, Ö., Wolter 2020)**

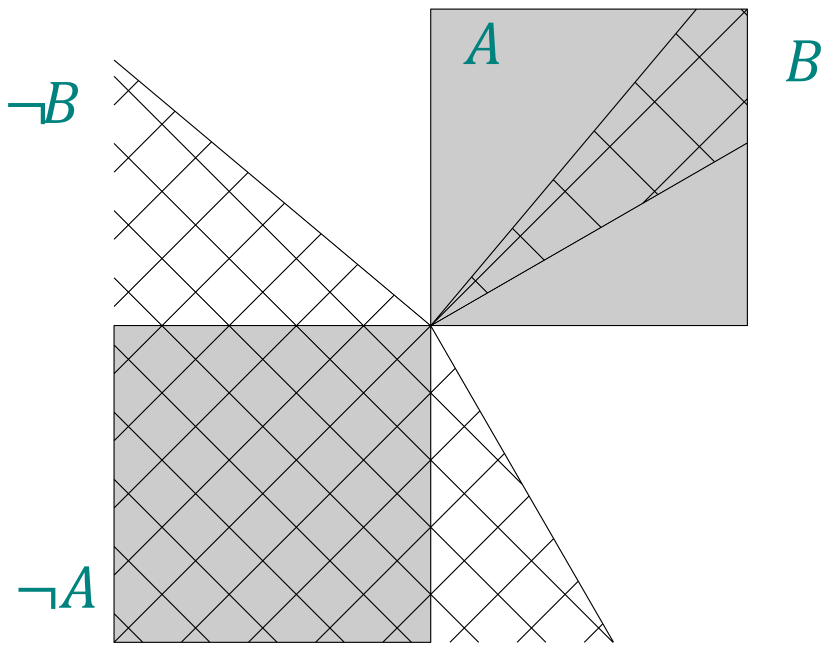
*Convex cones fulfil the rules of minimal orthologic.*

- Hence: $L_{cone}$ must extend minimal orthologic
- Which additional rules hold?

# Known Weakenings of Distributivity are Falsified

**Definition (Orthomodularity)**

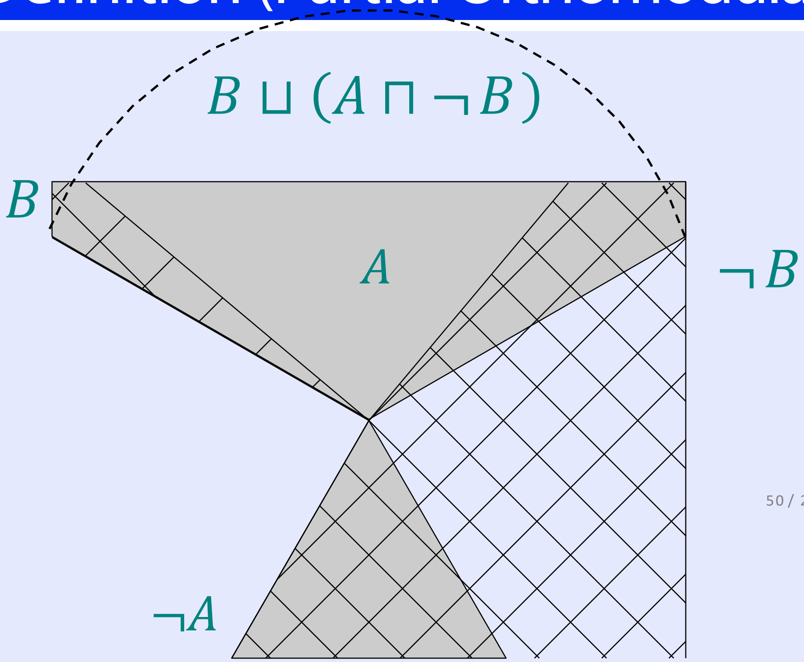
If $B \sqsubseteq A$ then $A \sqcap (\neg A \sqcup B) \sqsubseteq B$




$B \sqsubseteq A$, but

$A \sqcap (\neg A \sqcup B) = A \neq B$

## Definition (Partial Orthomodularity (pOM))




If $B \sqsubseteq A$

and $A \sqsubseteq B \sqcup (A \sqcap \neg B)$

then $A \sqcap (\neg A \sqcup B) \sqsubseteq B$



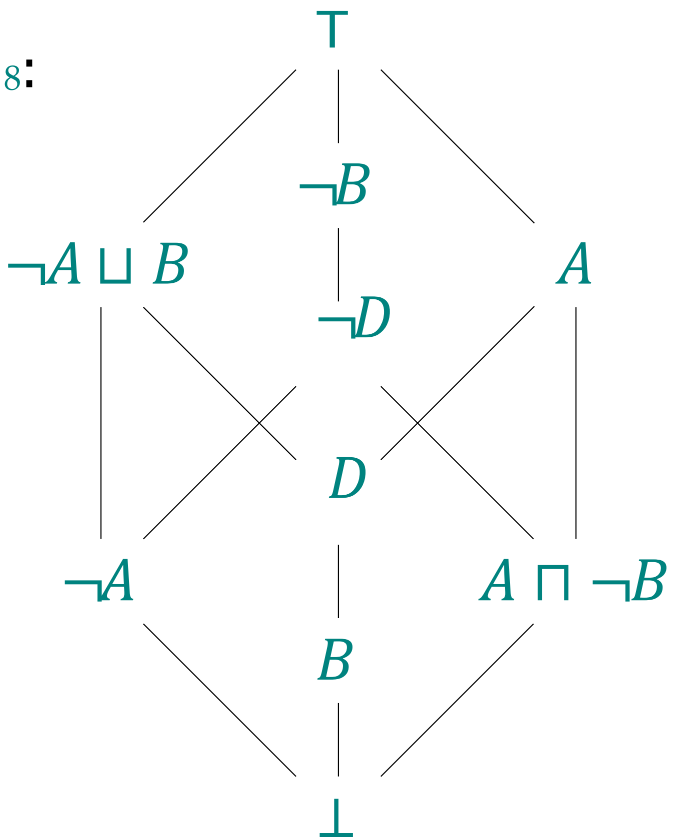
## Theorem (Leemhuis, Ö., Wolter 2020)

*Convex cones in 2D fulfill (pOM).*

Already in 3D have counterexamples

# Subalgebra Theorem

$MC_8$:




If $B \sqsubseteq A$

and $A \sqsubseteq B \sqcup (A \sqcap \neg B)$

But $A \sqcap (\neg A \sqcup B) = D \neq B$

**Theorem (Leemhuis, Ö., Wolter 2020)**

*A logic fulfils (pOM) iff it does not contain* $MC_8$.

# Open questions

- $L_{cone} = L_{min}+?$ / convex cones
- ML implementation with (al)-cones[1)]

1) (Leemhuis, Ö., Wolter ICCS 20)

Uhhh, a lecture with a hopefully useful

# APPENDIX

# Color Convention in this course

- Formulae, when occurring inline
- Newly introduced terminology and definitions
- Important results (observations, theorems) as well as emphasizing some aspects
- Examples are given with standard orange with possibly light orange frame
- Comments and notes
- Algorithms

# Today‘s lecture is based on the following

- Jonathon Hare: Lecture 14 of course „COMP6248 Differentiable Programming (and some Deep Learning)“ http://comp6248.ecs.soton.ac.uk/
- Möller/Özcep: Lecture „Word semantics and Latent Relational Structures“ of Course „Web-Mining Agents“
- Özcep: Knowledge Graph Embeddings, Talk at KI-Kolloqium https://www.ifis.uni-luebeck.de/~moeller/KI-Kolloquium/2019-12-02-Oezcep.pdf

# References


- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. ArXiv e-prints, Oct. 2013.
- J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543, 2014.
- Y. Goldberg and O. Levy. word2vec Explained: deriving Mikolov et al.'s negative-sampling word- embedding method, Arxiv eprints,Feb. 2014.
- Nickel, M.; Tresp, V.; and Kriegel, H.-P. A three-way model for collective learning on multi-relational data. In Proc. ICML, 809–816, 2011.
- A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In Proc. of NIPS-13, pp. 2787–2795, 2013.
- Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of AAAI-15, pages 2181–2187, 2015.
- B. Yang, W. tau Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. CoRR, abs/1412.6575, 2014
- Nguyen, D. Q.; Sirts, K.; Qu, L.; and Johnson, M. STransE: a novel embedding model of entities and relationships in knowledge bases. In Proc. of NAACL-HLT, 460–466, 2016.
- Trouillon, T., Welbl, J., Riedel, S., Gaussier, E. and Bouchard, G. Complex embeddings for simple link prediction. In Proc. of ICML, 2071–2080, 2016
- S. Mehran Kazemi and D. Poole. SimplE Embedding for Link Prediction in Knowledge Graphs. arXiv e-prints, page arXiv:1802.04868, Feb 2018.
- V. Gutierrez-Basulto and S. Schockaert. From knowledge graph embedding to ontology embedding? An analysis of the compatibility between vector space representations and rules. In Proceedings of KR 2018, 379–388, 2018.
- M. Kulmanov, W. Liu-Wei, Y. Yan, and R. Hoehndorf. $EL$ embeddings: Geometric construction of models for the description logic $EL^{++}$. In Proceedings of IJCAI-19, 2019.

# References


- Ö. L. Özçep, M. Leemhuis, and D. Wolter. Cone semantics for logics with negation. In C. Bessiere, editor, Proc. of IJCAI-20, pages 1820–1826. ijcai.org, 2020.
- M. Leemhuis, Ö. L. Özçep, and D. Wolter. Multi-label learning with a cone-based geometric model. In Proc. of ICCS-20, 2020.
- R. I. Goldblatt. Semantic analysis of orthologic. Journal of Philosophical Logic, 3(1):19–35, 1974.
- M. Leemhuis, Ö. L. Özçep, and D. Wolter. Orthologics for Cones. arXive-prints, arXiv:2008.03172, Aug. 2020.