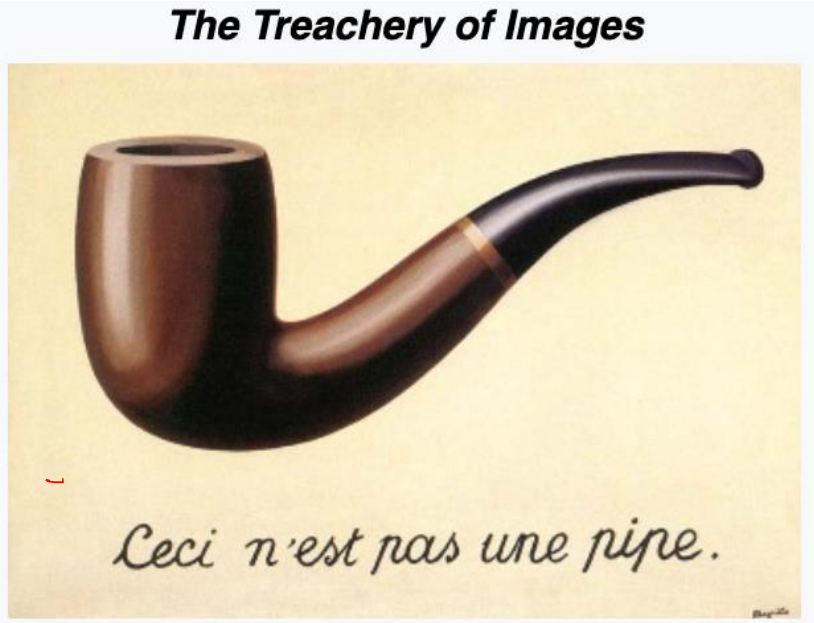

Ceci n'est pas une pipe.
Magritte

WORD VECTOR SPACE MODELS



- Phonetics and Phonology
  knowledge about linguistic sounds
- Morphology
  knowledge of the formation and internal structure of words
- Syntax
  knowledge of the structural relationships between words
- Semantics
  knowledge of meaning
- Pragmatics
  knowledge of the relationship of meaning to the goals & intentions of the speaker
- Discourse
  knowledge about linguistic units larger than a single utterance

Task: Engaging in Natural Language Communication

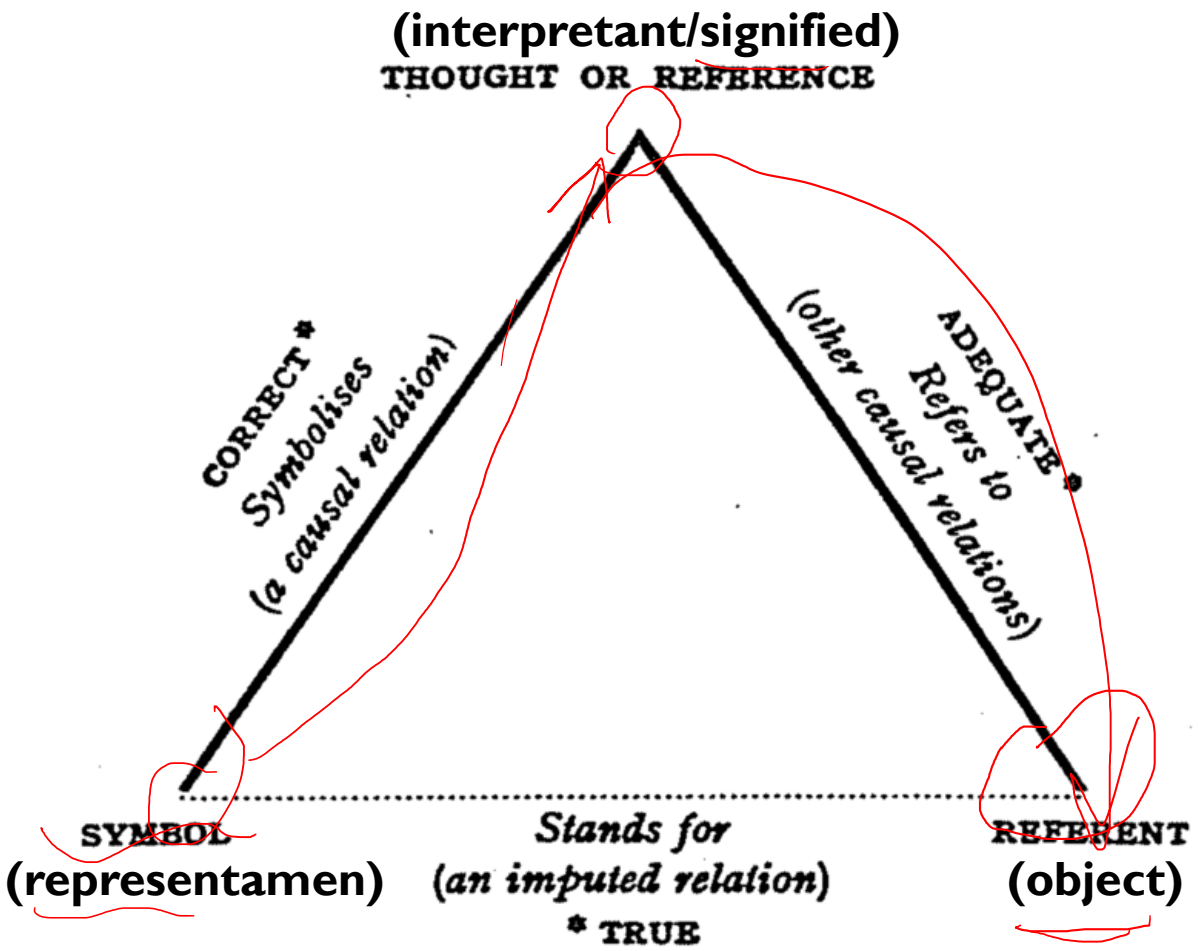
Semiotics: The Science of Symbols

Semantics: Relation between signs and things to which they refer: meaning; sense

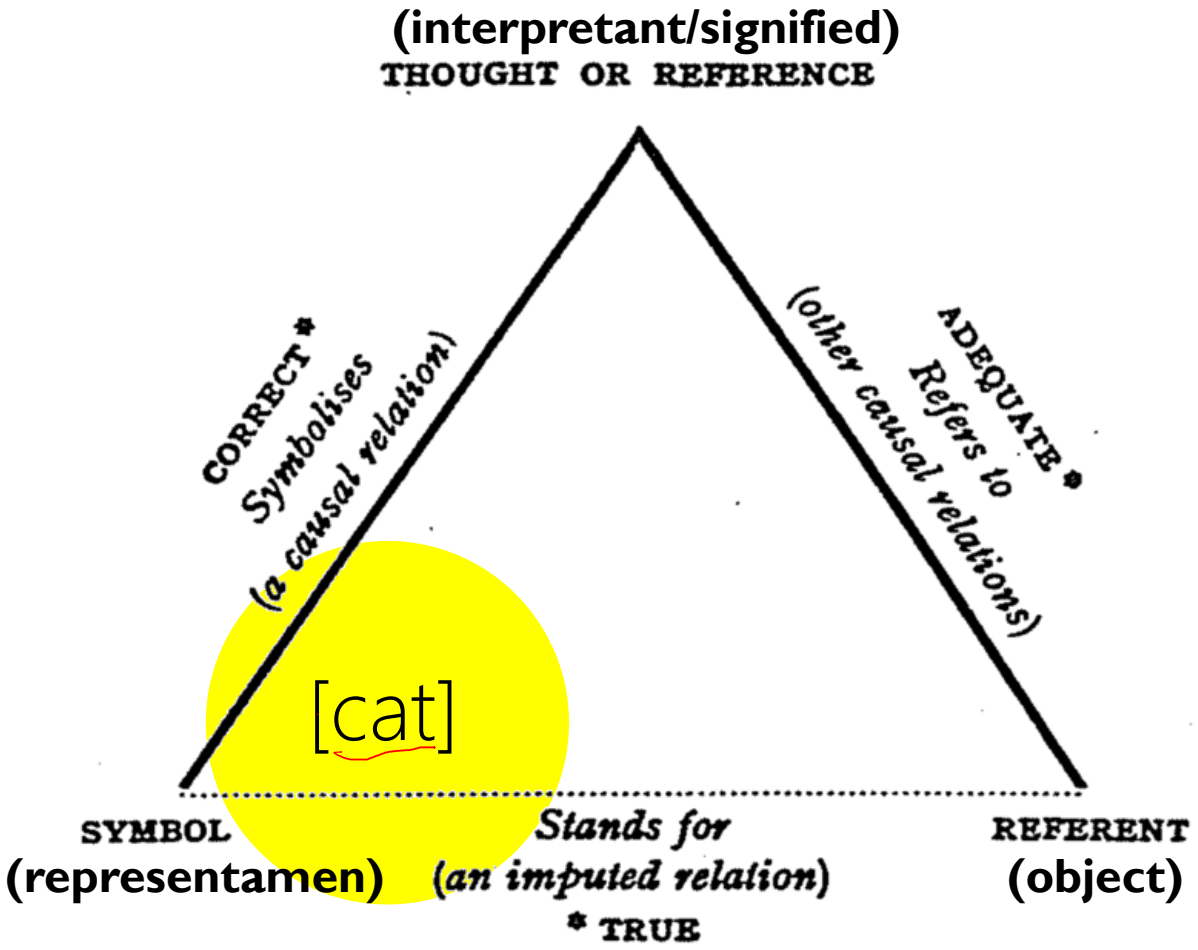
Syntactics: Relations among signs in formal structures

Pragmatics: Relation between signs and sign-using agents

Triangle of Semiotics

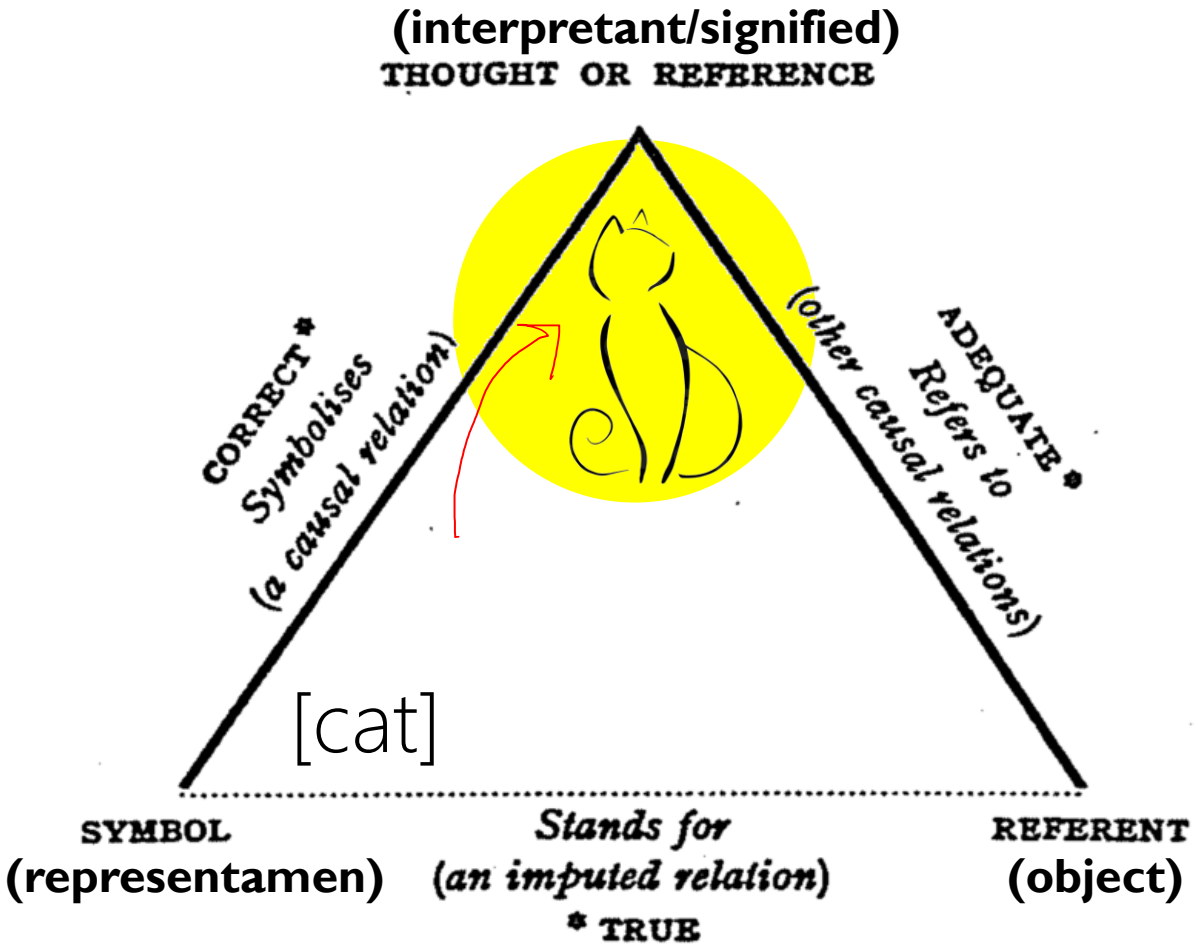


Triangle of Semiotics



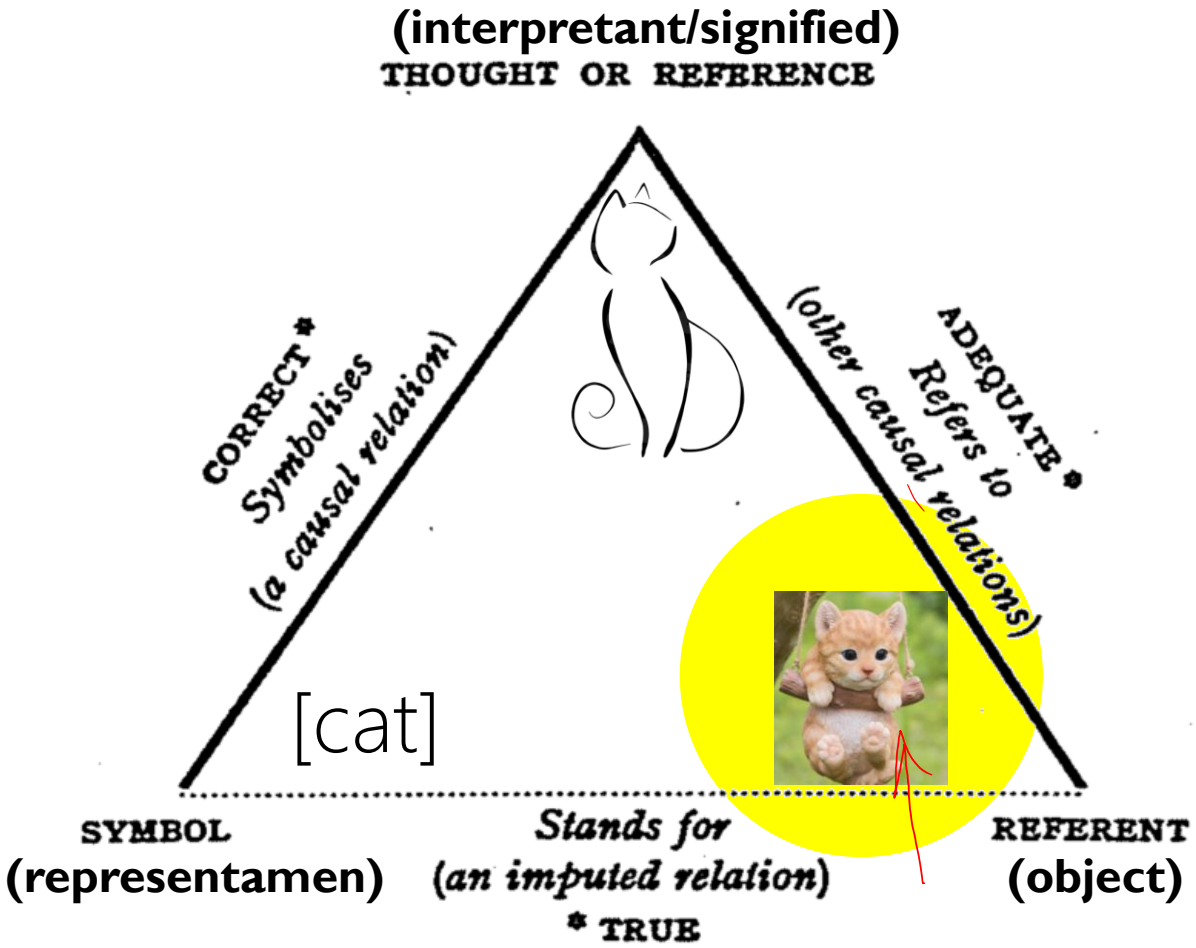
The representamen (sign vehicle, signifier, symbol) represents the object

Triangle of Semiotics



The interpretant (signified) is the sense/meaning made of the representamen and the object

Triangle of Semiotics

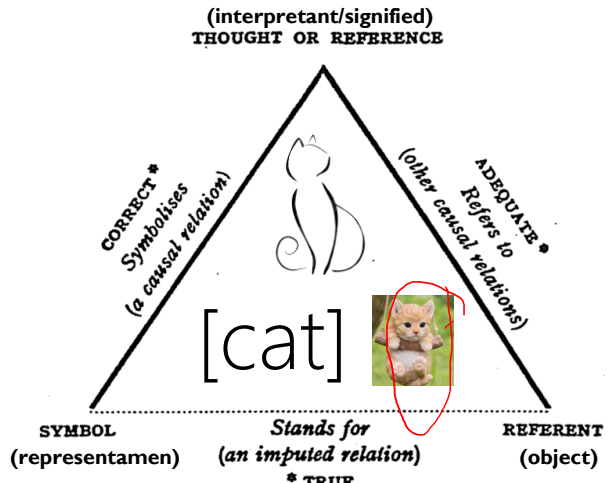


The object of a sign is always hidden!

Sign

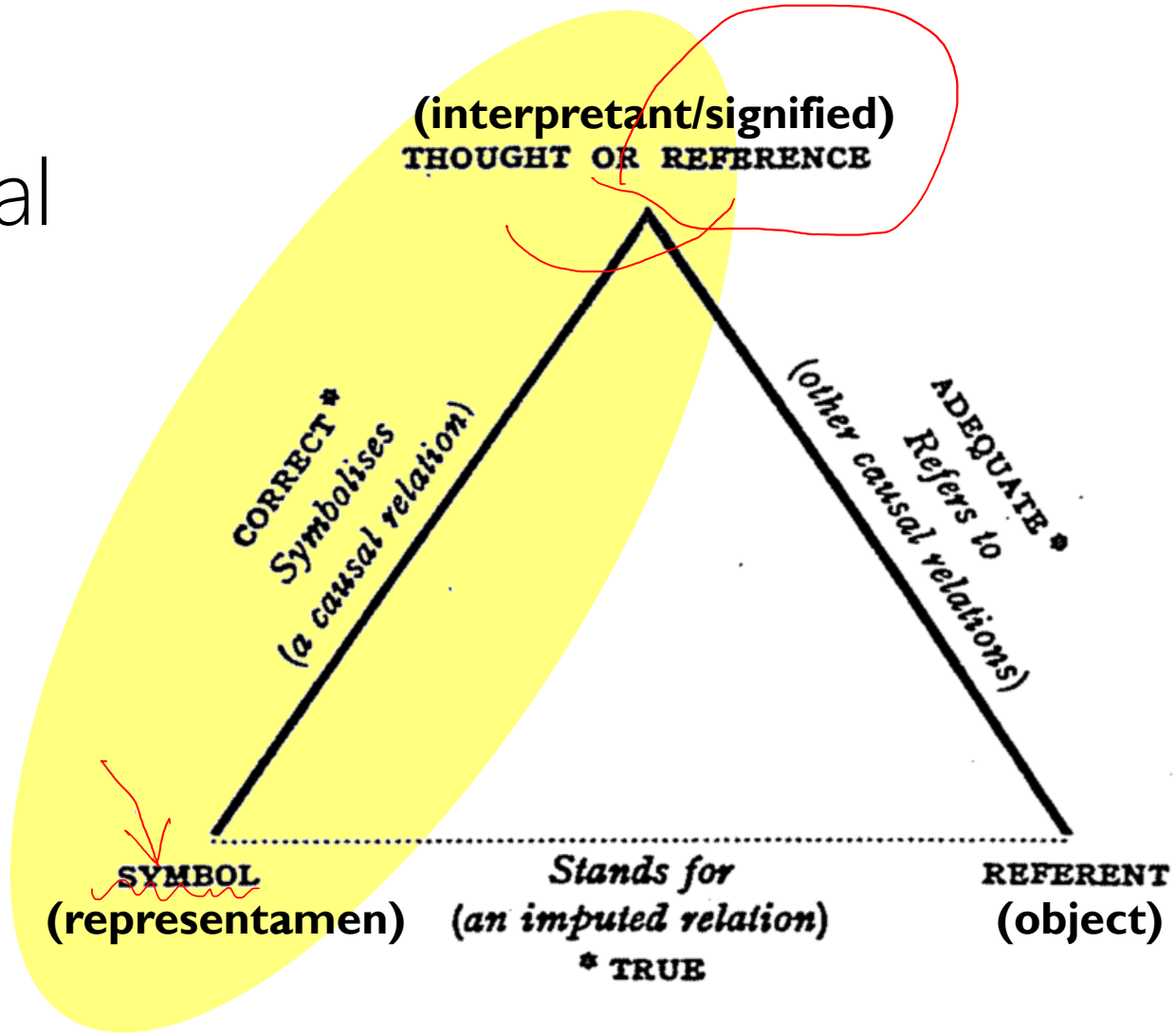
A sign is a triadic unity of:

- Object
- Representamen (signifier/symbol)
- Interpretant (signified/mean/sense)



Computational Semantics

The objective of computational semantics research is to *automatically* find the relation between a signifier and the signified/sense/meaning.

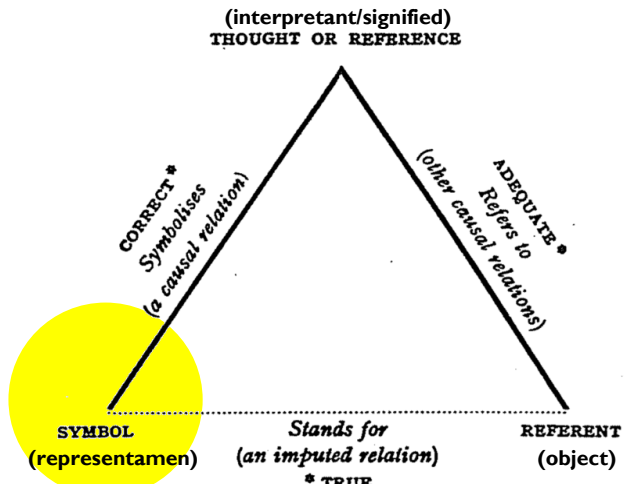


Representament Learning

Representation Learning

From school:

- ['c', 'a', 't']

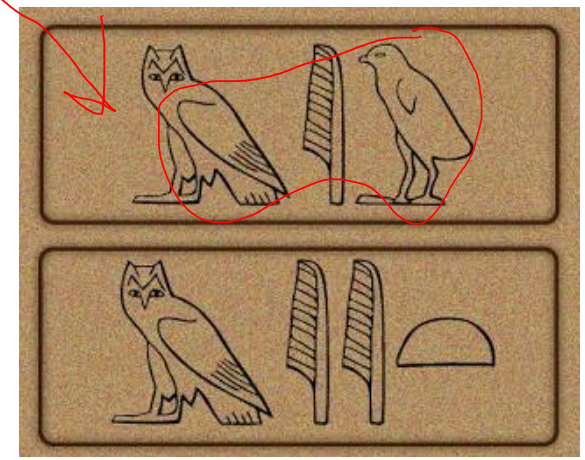


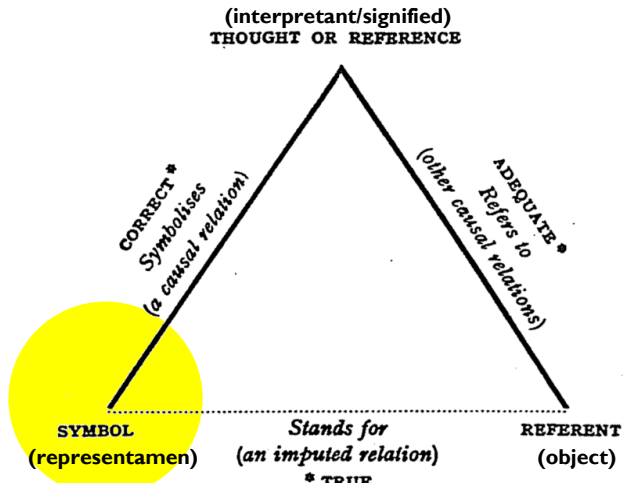
Representament Learning

Representation Learning

In hieroglyphs:

- miu/mii (male)
- Miit (female)



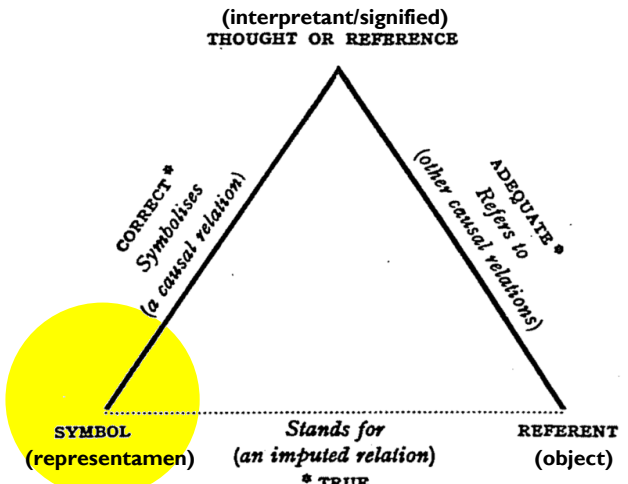


Representament Learning

Representation Learning

In computer ASCII code:

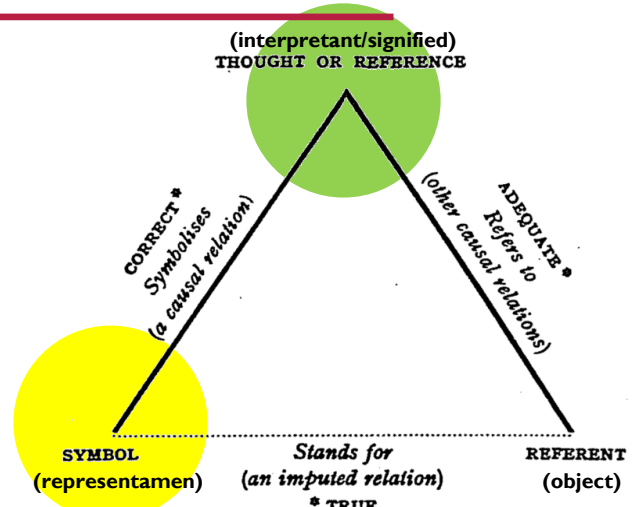
- [99, 97, 116]



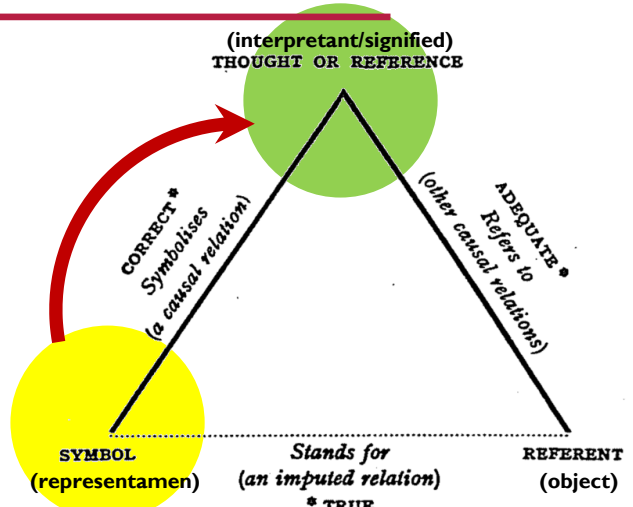
Representament Learning

Representation Learning

Help with signified/meaning/sense!

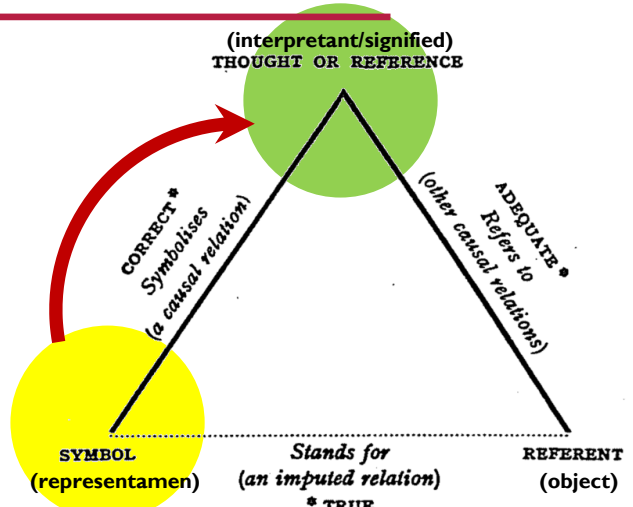


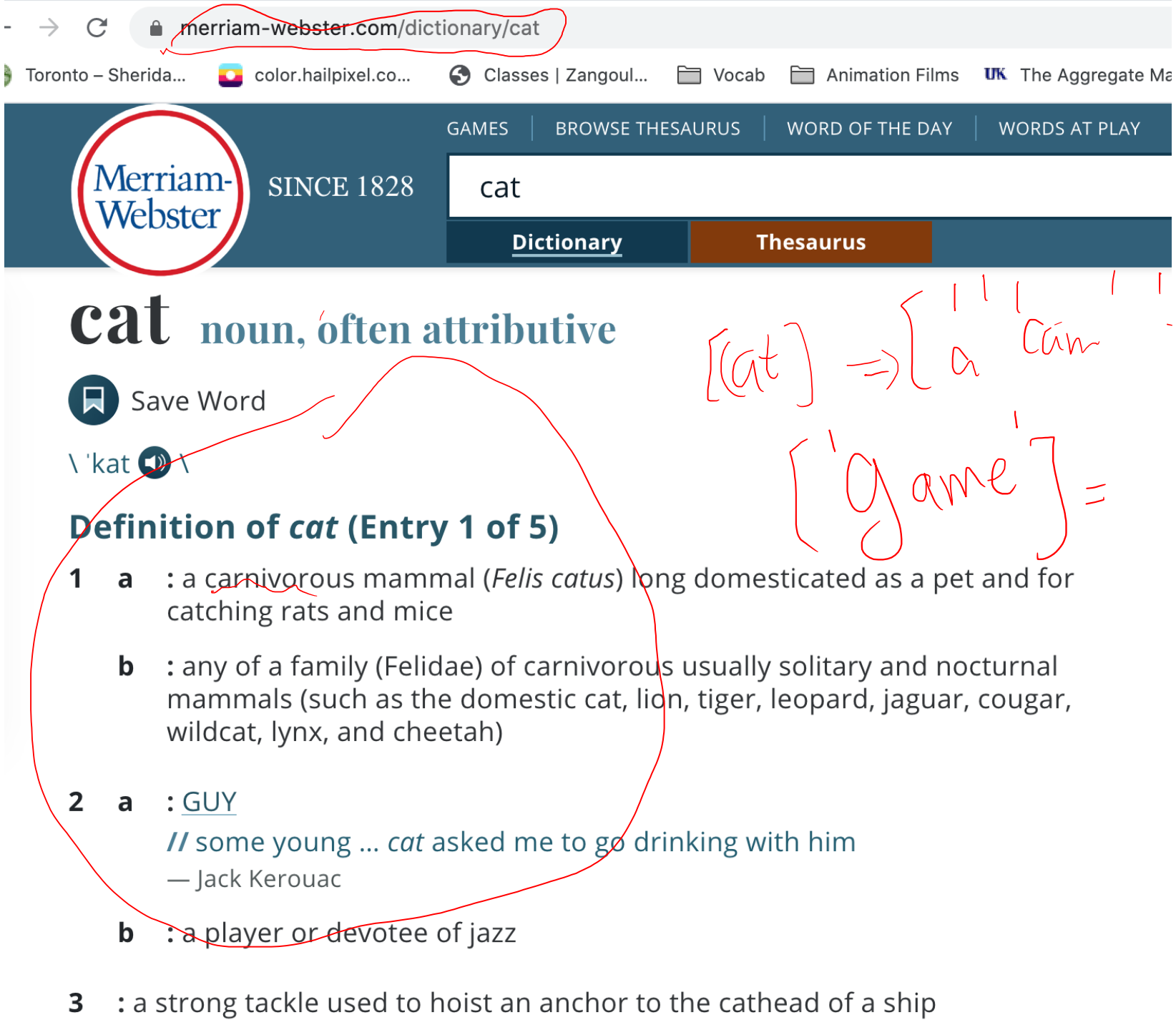
Representament → Interpretant



Representament → ?

We/machine see a token/symbol/signifier
→ What does it mean?
→ Comprehension (analyze the context)




merriam-webster.com/dictionary/cat
Toronto – Sherida...
color.hailpixel.co...
Classes | Zangoul...
Vocab
Animation Films
The Aggregate Ma
GAMES
BROWSE THESAURUS
WORD OF THE DAY
WORDS AT PLAY
Merriam-Webster
SINCE 1828
cat
Dictionary
Thesaurus
cat noun, often attributive
Save Word
\ 'kat \
Definition of cat (Entry 1 of 5)
1 a : a carnivorous mammal (Felis catus) long domesticated as a pet and for catching rats and mice
b : any of a family (Felidae) of carnivorous usually solitary and nocturnal mammals (such as the domestic cat, lion, tiger, leopard, jaguar, cougar, wildcat, lynx, and cheetah)
2 a : GUY
// some young ... cat asked me to go drinking with him
— Jack Kerouac
b : a player or devotee of jazz
3 : a strong tackle used to hoist an anchor to the cathead of a ship

Ludwig Josef Johann Wittgenstein

/ˈvɪtgənʃtaɪn, -staɪn/

1889 –1951

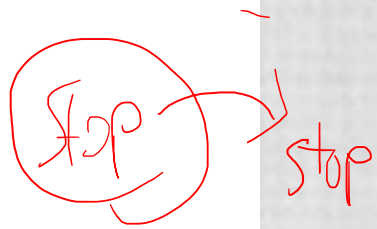
Austrian-British Philosopher

Worked primarily in:

- Logic
- The philosophy of mathematics
- The philosophy of mind
- The philosophy of language

Ludwig Josef Johann Wittgenstein

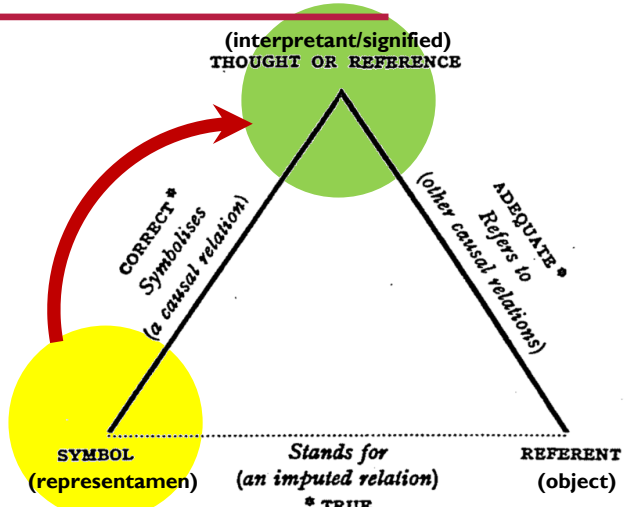
/ˈvɪtgənʃtaɪn, -staɪn/
1889 –1951
Austrian-British Philosopher



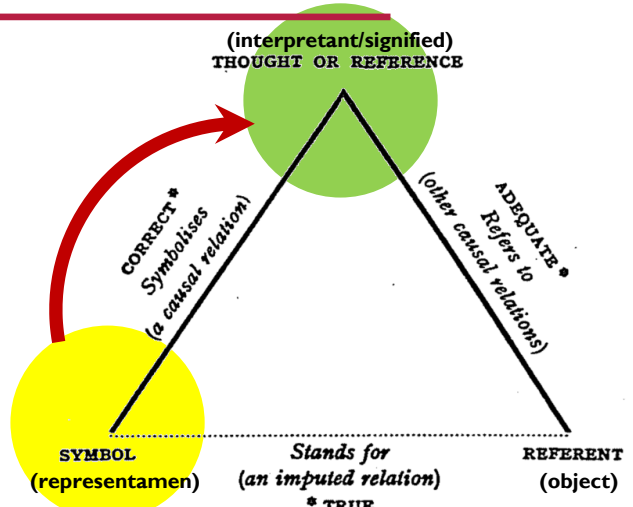
Skeptical of a completely formal theory of meaning definitions for each word

→ "the meaning of a word is its use in the language" - Philosophical Investigations.

Token → ?

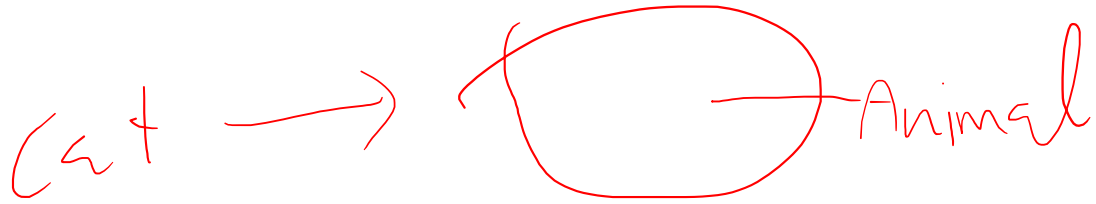

(interpretant/signified)
THOUGHT OR REFERENCE
CORRECT *
Symbolises
(a causal relation)
ADEQUATE *
Refers to
(other causal relations)
SYMBOL
(representamen)
Stands for
(an imputed relation)
REFERENT
(object)

Token → Relations with other Tokens → Meaning



Lexical vs. Vector Semantics

- Lemma
- Wordforms
- Synonyms
- Antonyms
- Connotations
- Similar Tokens (Word Similarity)
- Related Tokens (Word Relatedness)
- Distribution (co-occurrences)

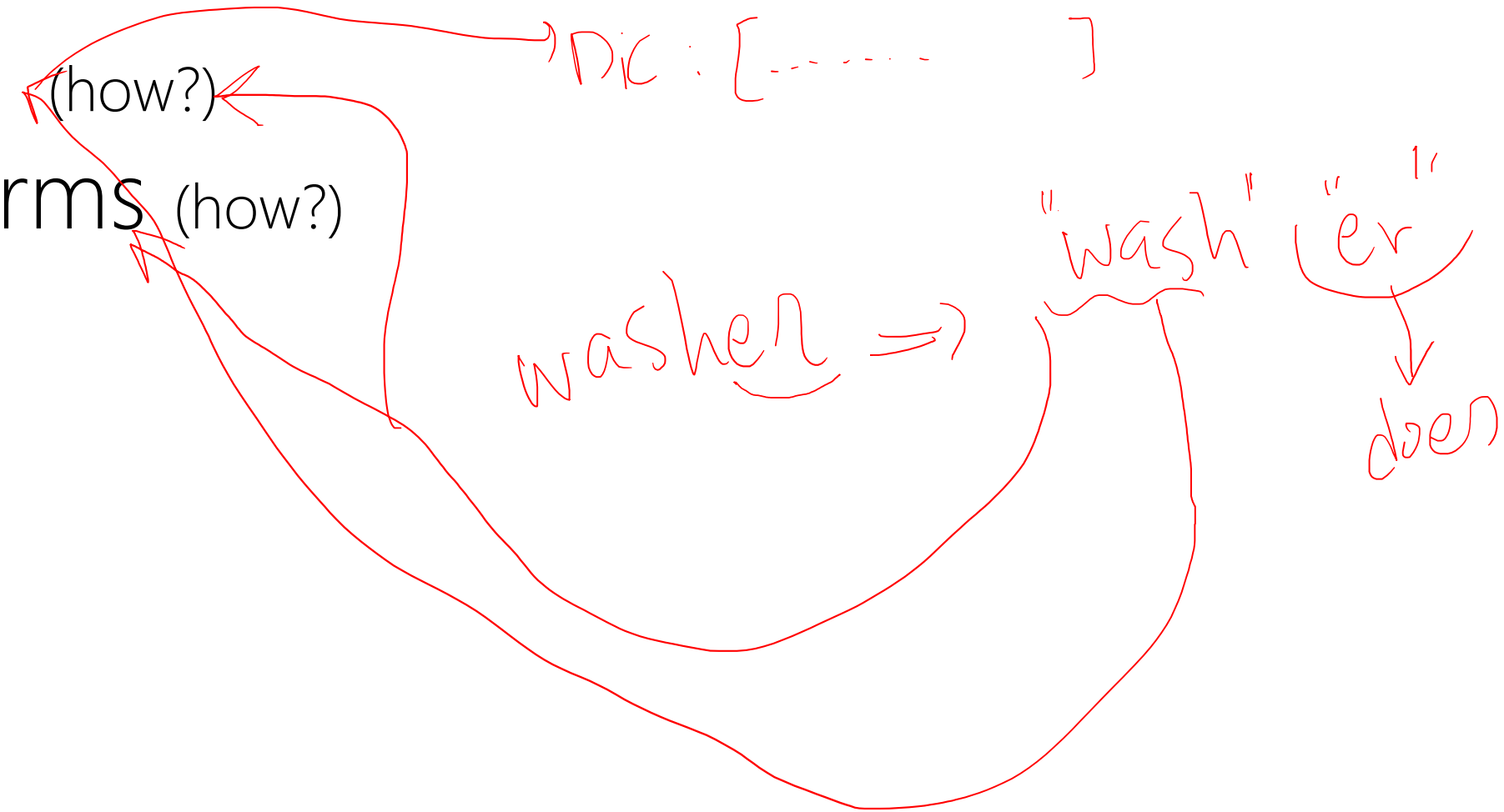


Lexical Semantics

- Lemma
- Wordforms
- Synonyms
- Antonyms
- Connotations
- Similar Tokens (Word Similarity)
- Related Tokens (Word Relatedness)
- Distribution (co-occurrences)

Lexical Semantics

- Lemma (how?)
- Wordforms (how?)



Lexical Semantics

- Synonyms: different signifiers near same signified, e.g., water:H_2O

The principle of contrast: difference in linguistic form [signifier] is always associated with at least some difference in meaning.
(Girard 1718, Breal 1897, Clark 1987)

[H2O] → scientific contexts
[water] → hiking guide

Difference in genre is part of the meaning of the word.

Lexical Semantics

- Similar Tokens (Word Similarity)

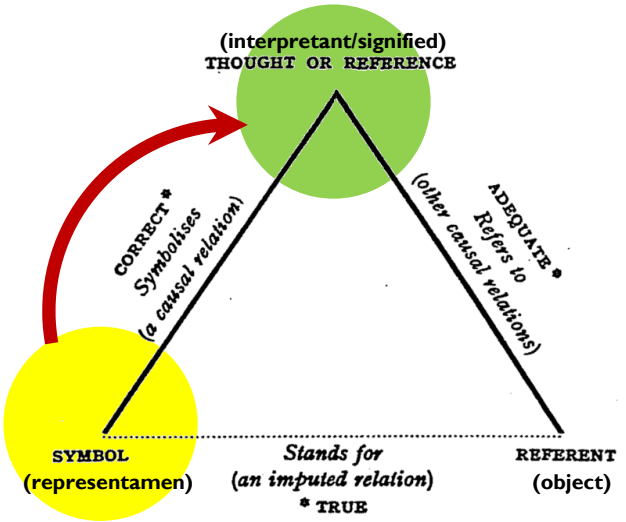
Similarity in the signified and object

SimLex-999

SimLex-999 is a gold standard resource for the evaluation of models that learn the meaning of words and concepts.

SimLex-999 provides a way of measuring how well models capture *similarity*, rather than *relatedness* or *association*. The scores in SimLex-999 therefore differ from other well-known evaluation datasets such as *WordSim-353* (Finkelstein et al. 2002). The following two example pairs illustrate the difference - note that *clothes* are not similar to *closets* (different materials, function etc.), even though they are very much related:

Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00



Lexical Semantics

- Related Tokens (Word Relatedness aka. Word Associations)

Semantic Field: School, Student, Book, Teacher

- Semantic Frame: events or transactions in field

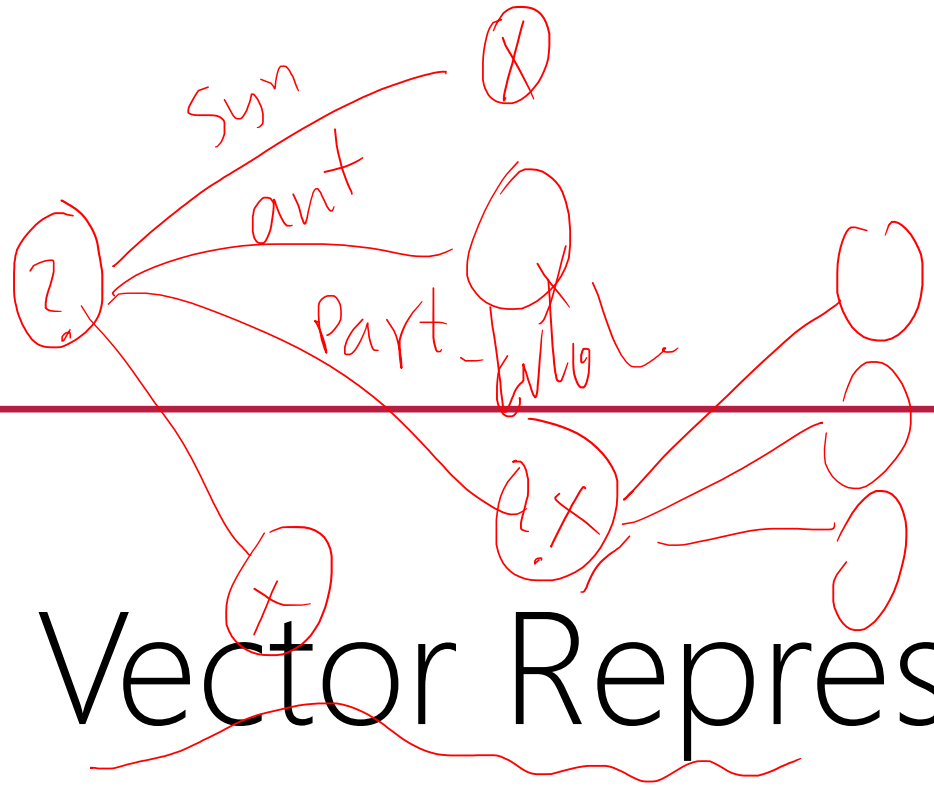
  registration, registrar, student, course

- Semantic Role: Who does What

  Student register a course. A course is taken by students.

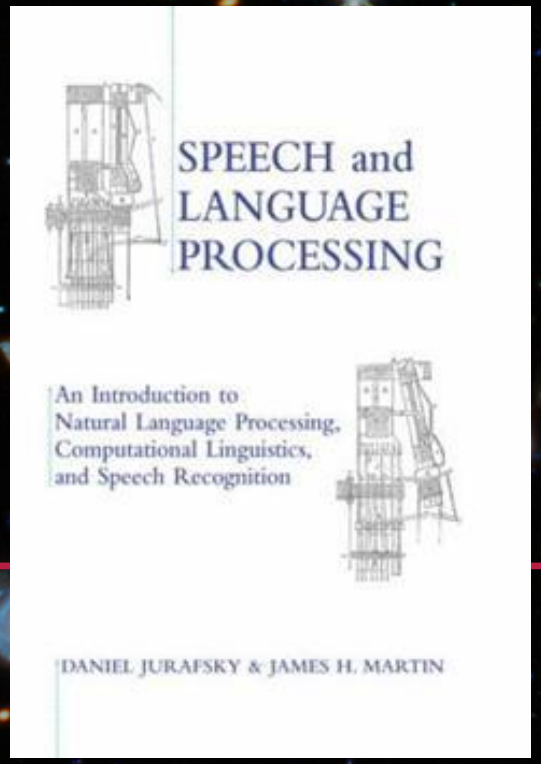
Lexical Semantics

- Hypernymy: Hierarchical Relation,
  color → red/blue
- Meronymy: Part-Whole Relation
  engine → Car



No Vector Representation

Not very helpful to machines

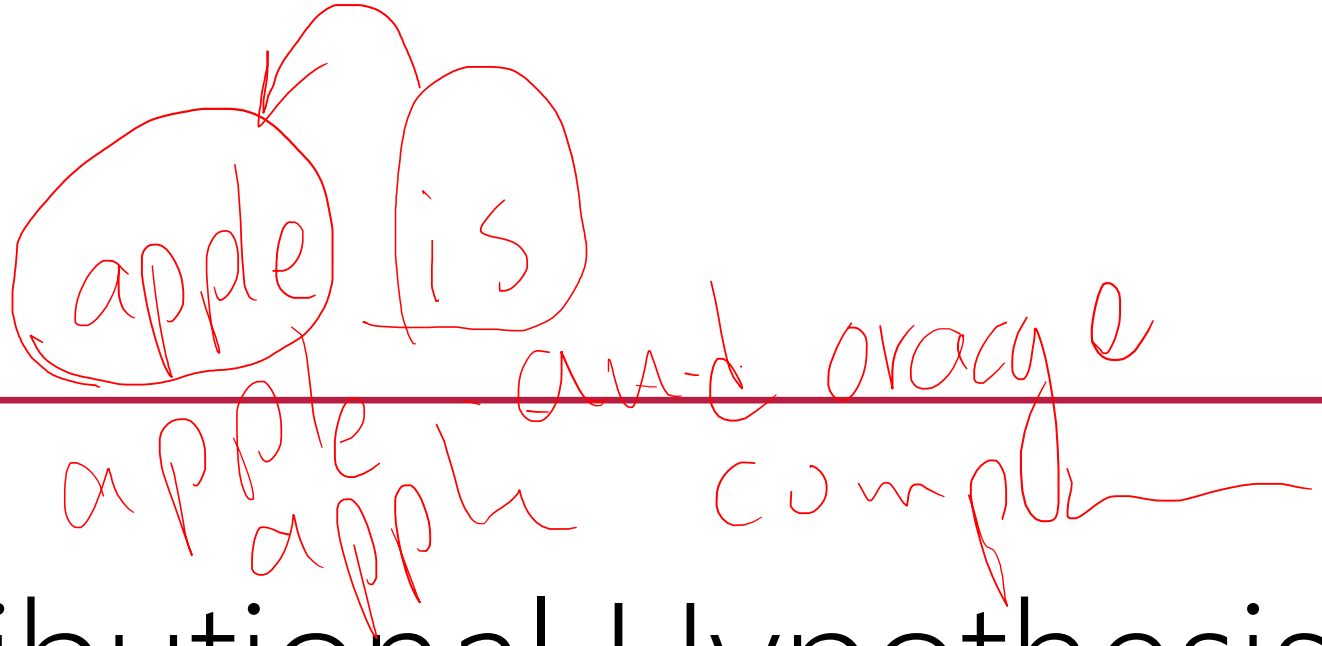


Vector Semantics & Embeddings

CH06

Vector Semantics

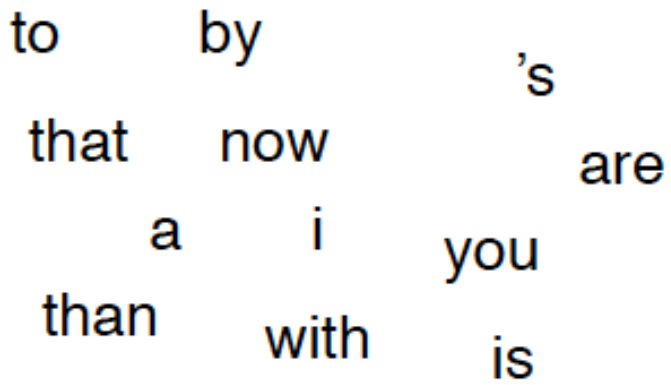
- Distribution (co-occurrences)
  - No supervised lexical connection for tokens
  - Statistical connections for tokens

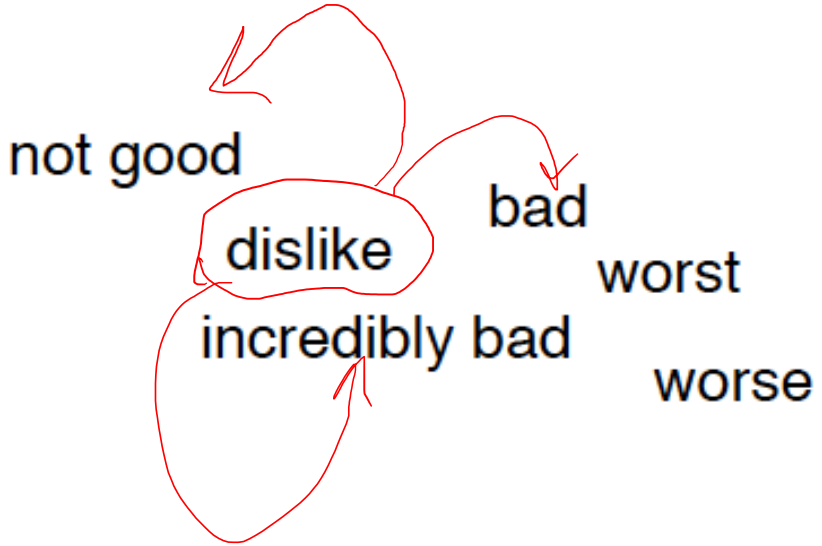


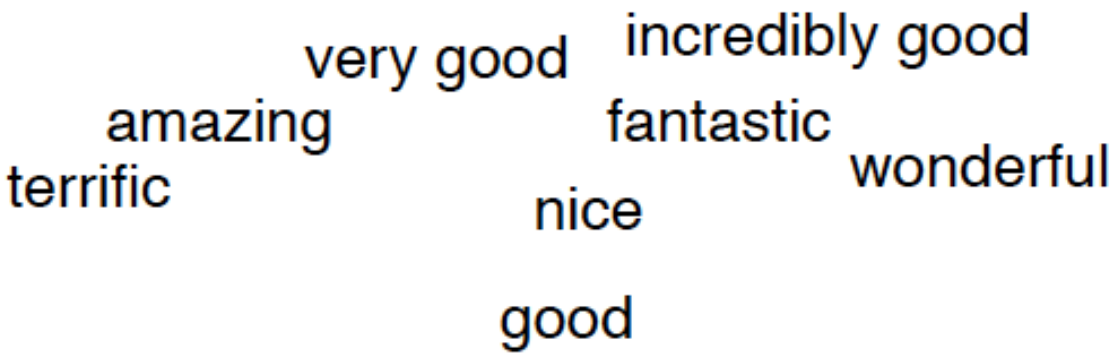
Distributional Hypothesis

Words that occur in similar contexts tend to have similar meanings.
Meaning difference corresponds to difference in environments.

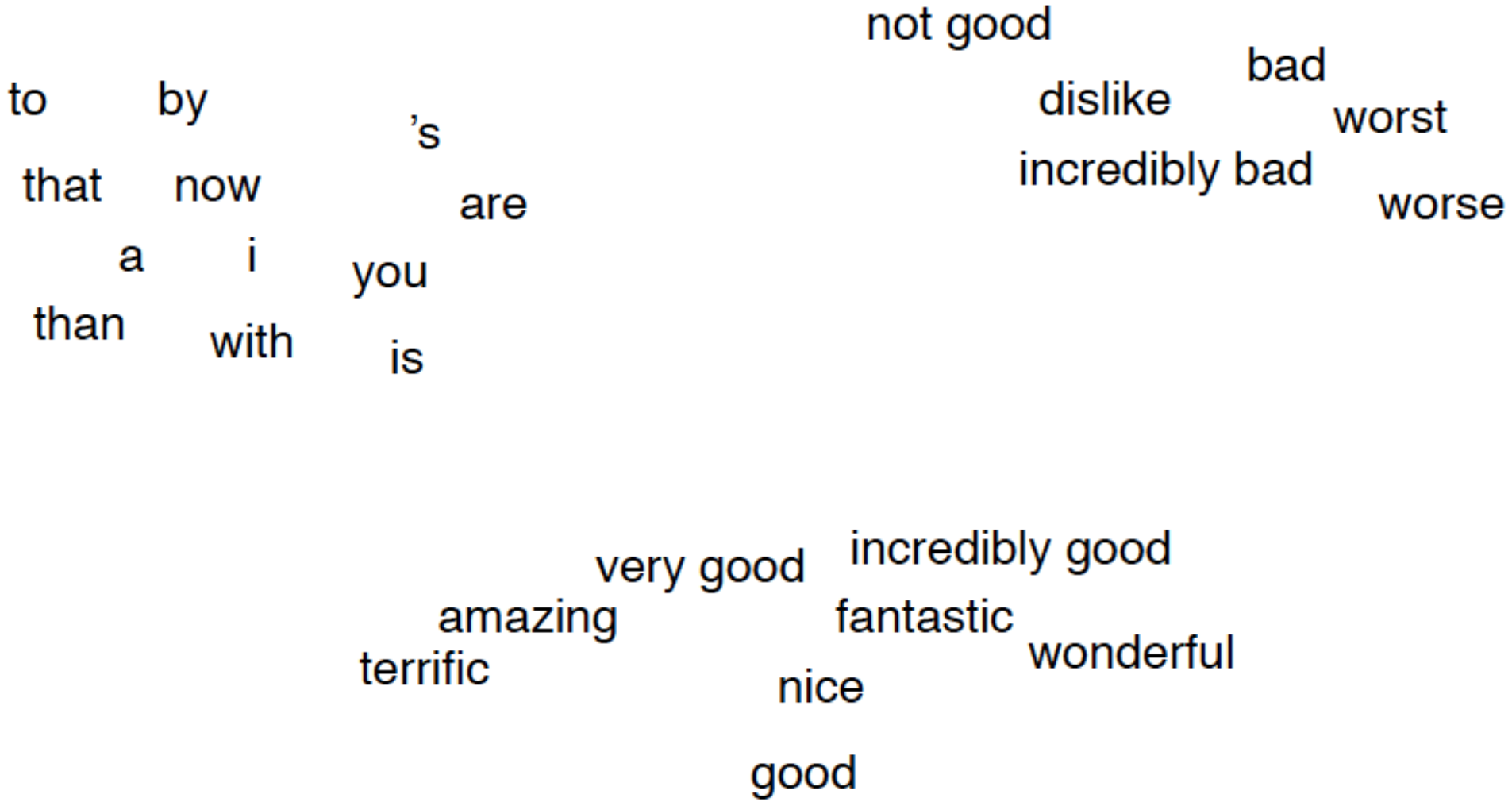
Joos, M. (1950). Description of language design. JASA, 22, 701–708.
Harris, Z. S. (1954). Distributional structure. Word, 10, 146–162.
Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955.



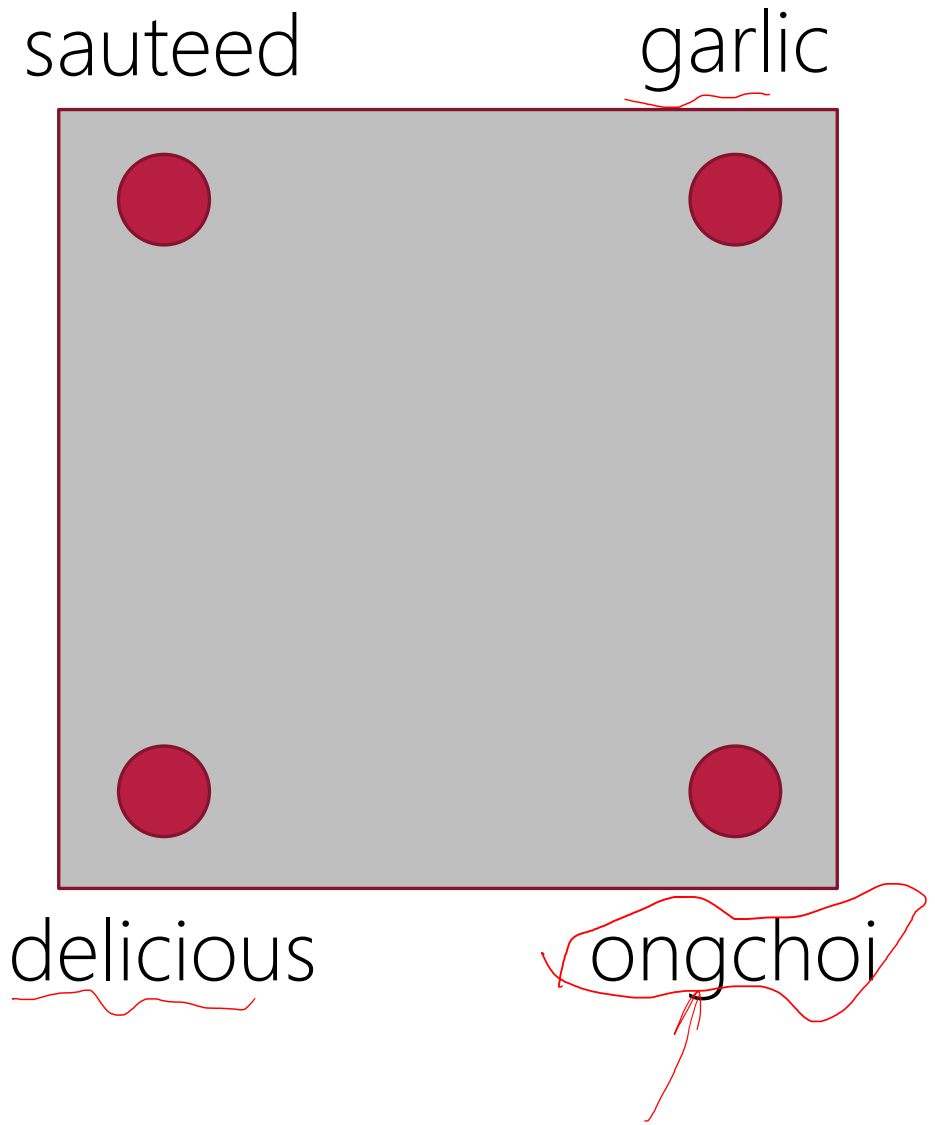




Our Mind Space!

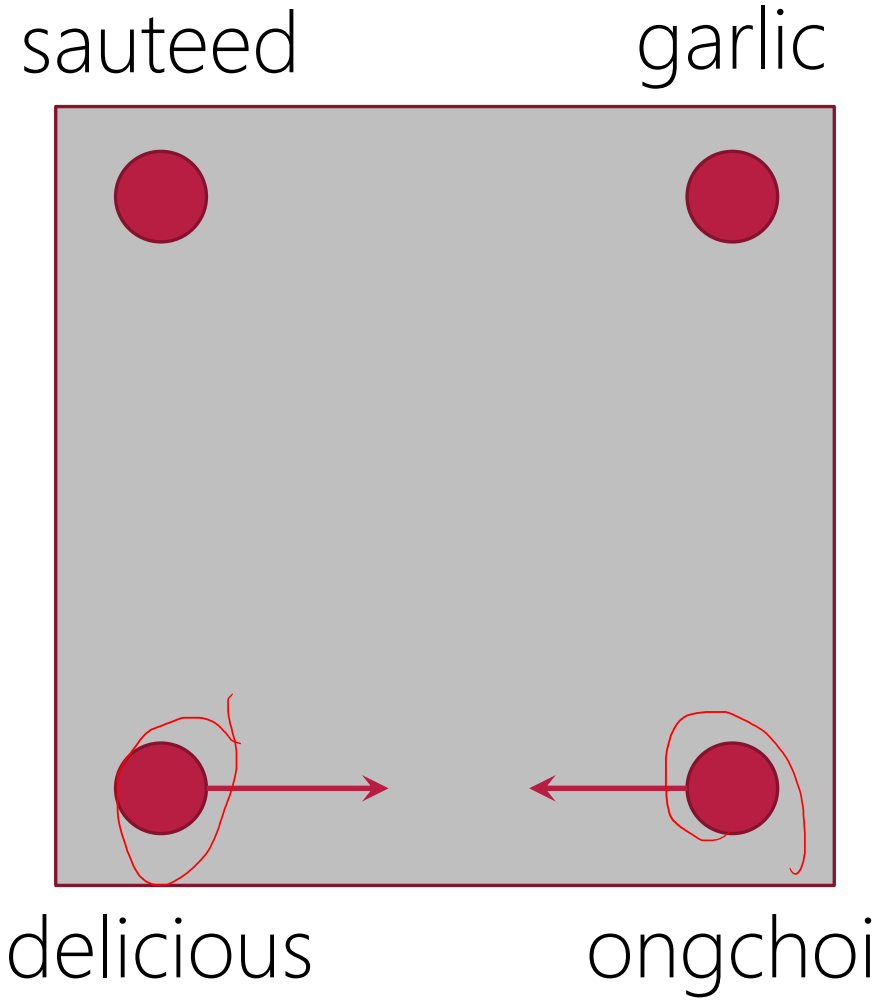

not good
bad
dislike
worst
incredibly bad
worse
to
by
's
that
now
are
a
i
you
than
with
is
very good
incredibly good
amazing
fantastic
wonderful
terrific
nice
good

2-D Space!



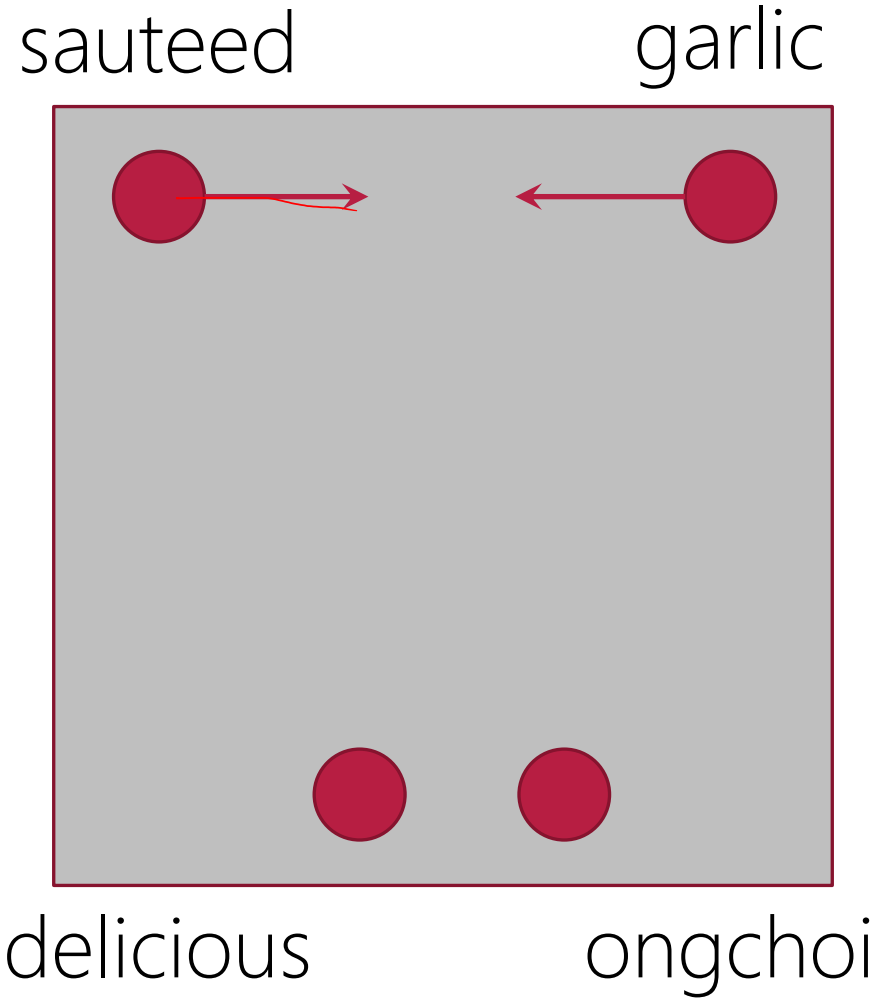
2-D Space!


ongchoi is delicious
context window of size 3 tokens


sauteed
garlic
delicious
ongchoi

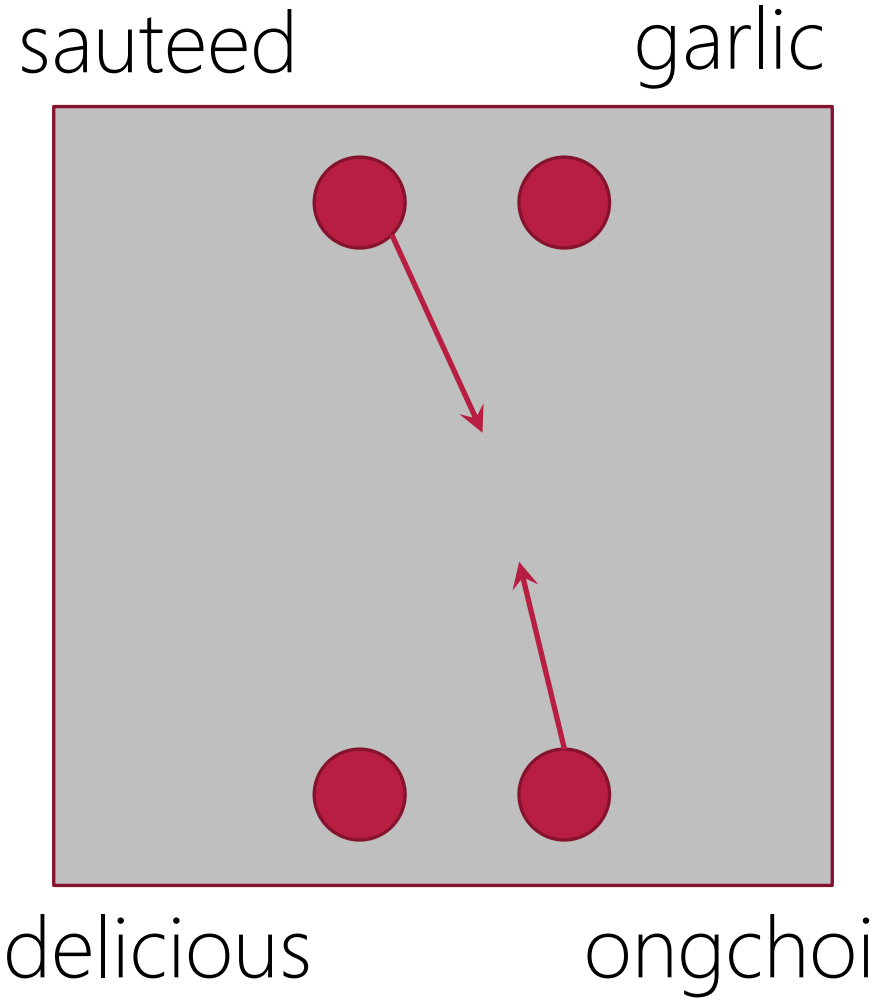
2-D Space!

ongchoi is delicious
ongchoi can be sauteed with garlic



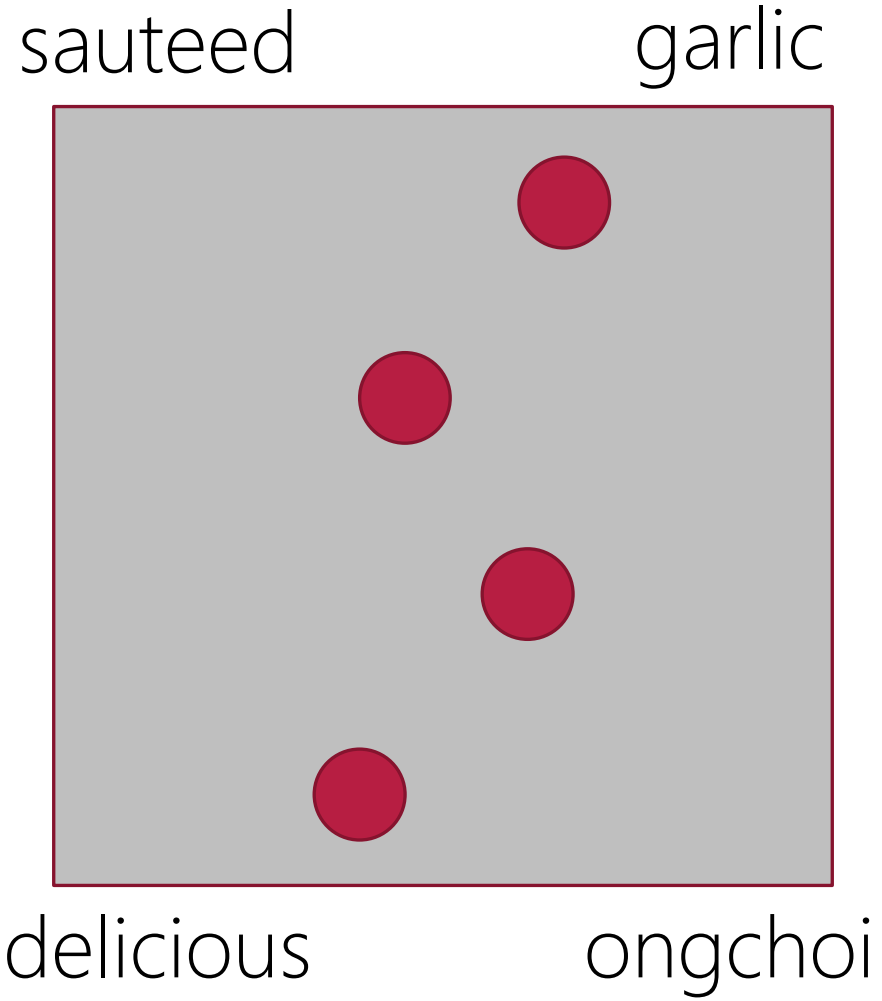
2-D Space!

ongchoi is delicious
ongchoi can be sauteed with garlic
sauteed ongchoi on rice



2-D Space!

ongchoi is delicious
ongchoi can be sauteed with garlic
sauteed ongchoi on rice



2-D Space!

Word Vector
Word Embedding
Word Point

Vector semantics is to represent/embed a word as a point in some multidimensional vector space as the semantic space.

Word Vector
Word Embedding
Word Point

Meaning Similarity → Geometric Proximity

Context Window

Term-Document Matrix

Context Window = Document

Term-Document Matrix

Context Window = Document

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

$$|\text{Vocabs}| \times |\text{Documents}|$$

Term-Document Matrix

Context Window = Document

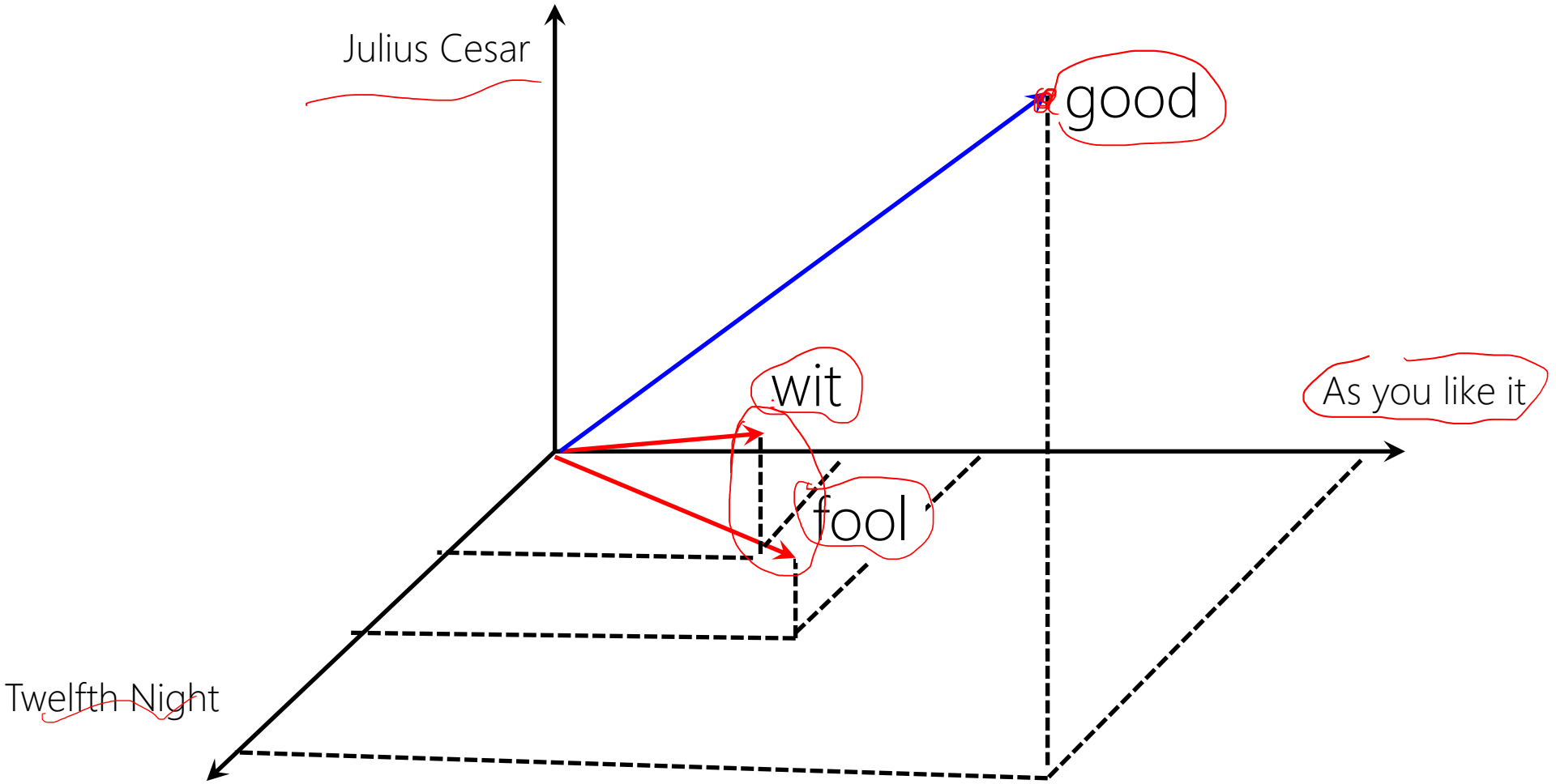
	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

$$|\text{Vocabs}| \times |\text{Documents}|$$

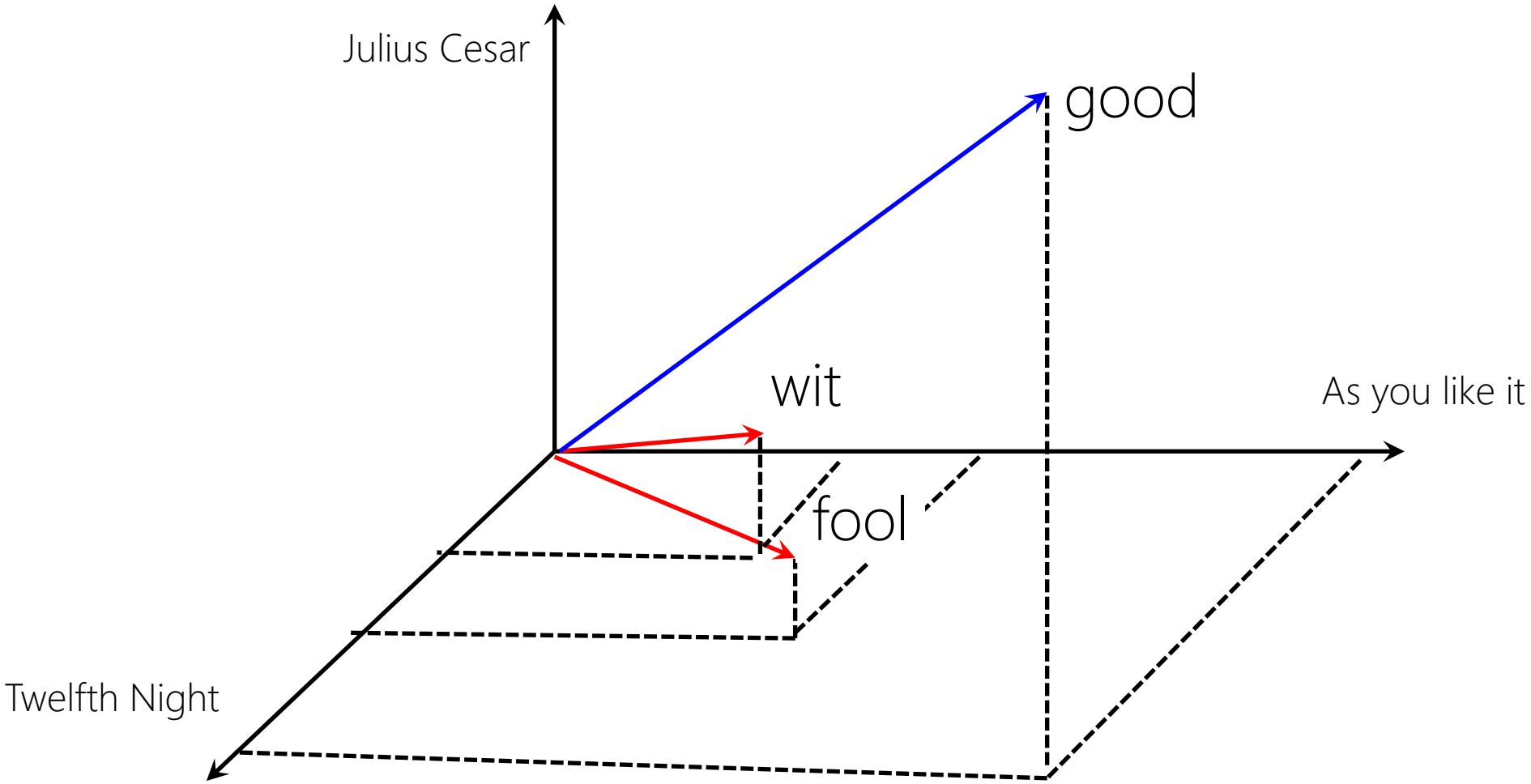
Term-Document Matrix

Words are points (vectors) in document space!



Term-Document Matrix

Distribution of words in documents!



Term-Document Matrix

Distribution of documents in words.
Documents in word space!

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

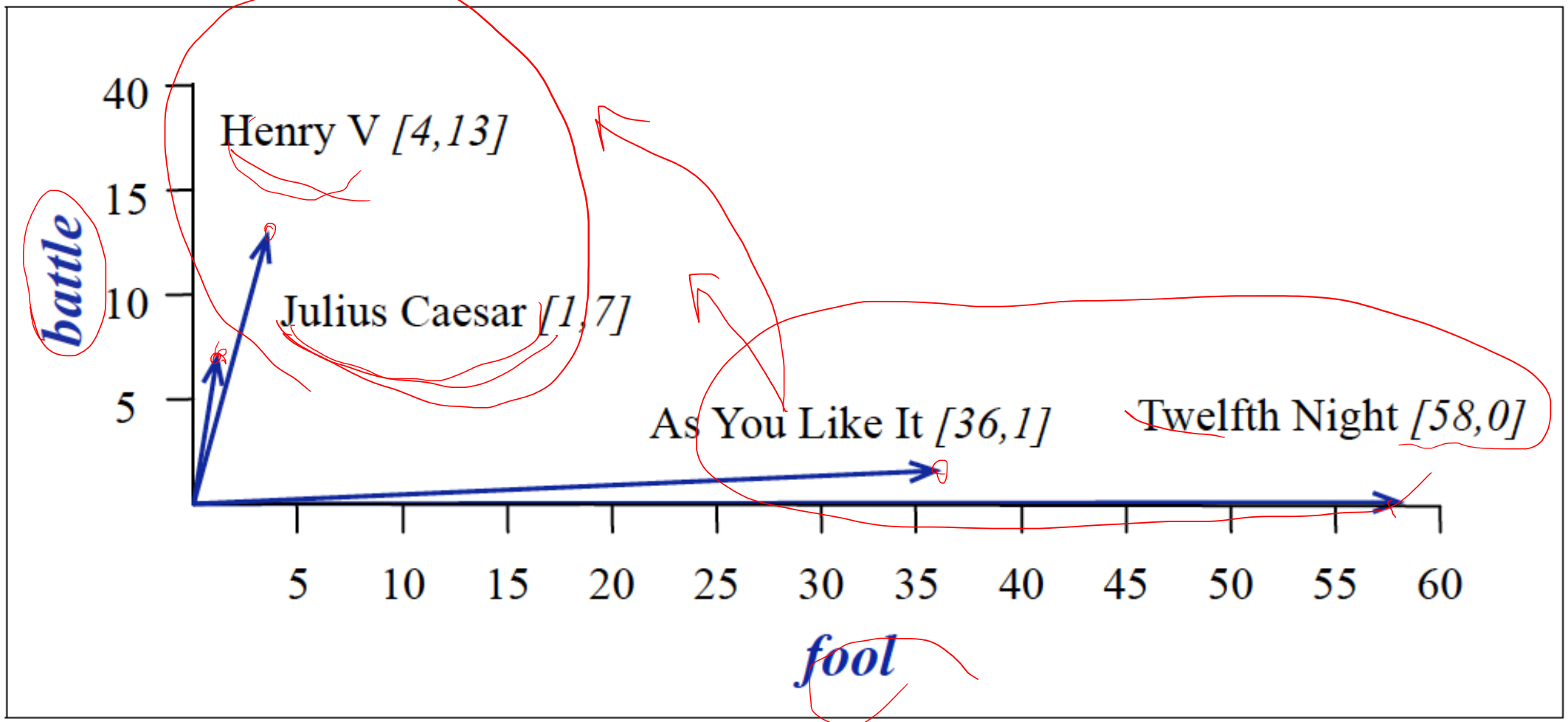


Figure 6.4 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Term-Document Matrix

Context Window = Document

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

Is it sparse or dense matrix? Why?

Term-Term Matrix

aka word-word matrix
aka word-context matrix

Context Window =

- Fixed (Window): ±n tokens = unordered n-gram
- Dynamic: Document, Paragraph, Sentence, Tweet

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

Figure 6.5 Co-occurrence vectors for four words in the Wikipedia corpus, showing six of the dimensions (hand-picked for pedagogical purposes). The vector for *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

Context Window: ± 4 words around the word.

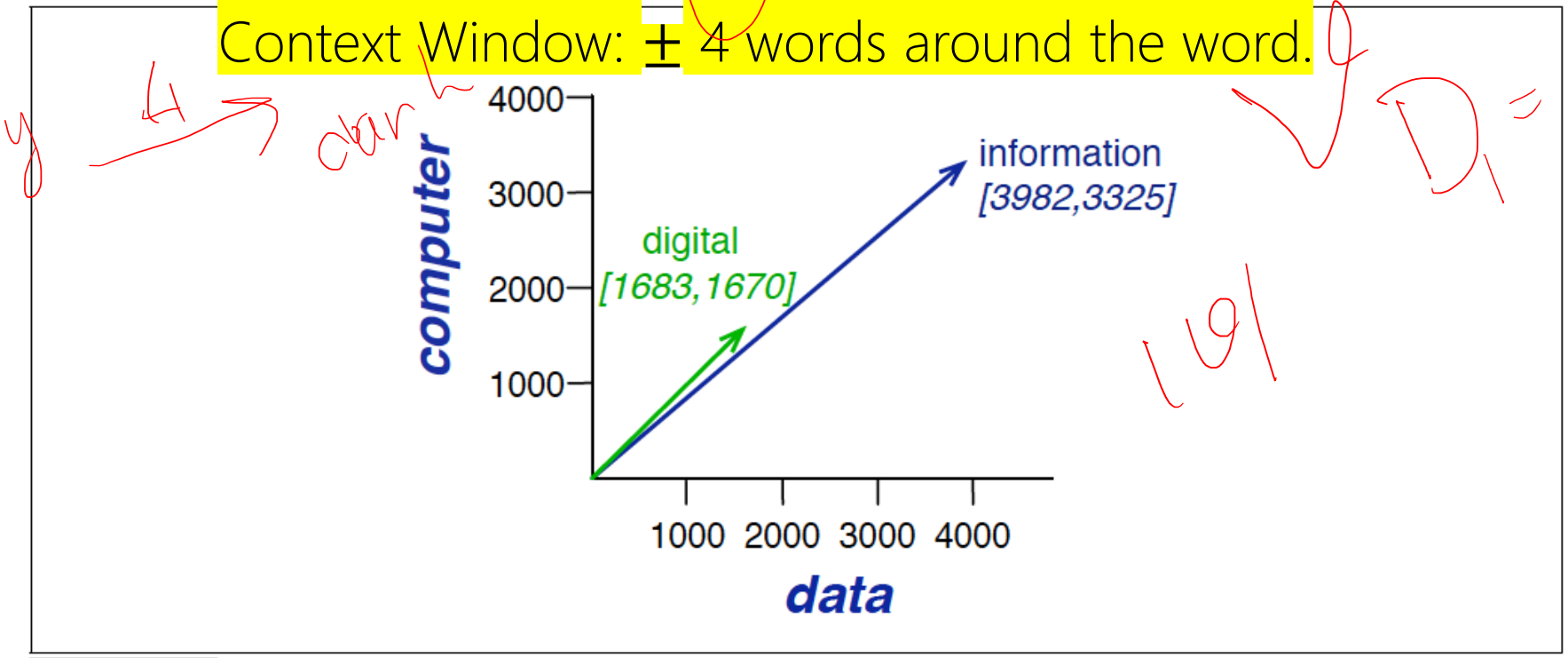
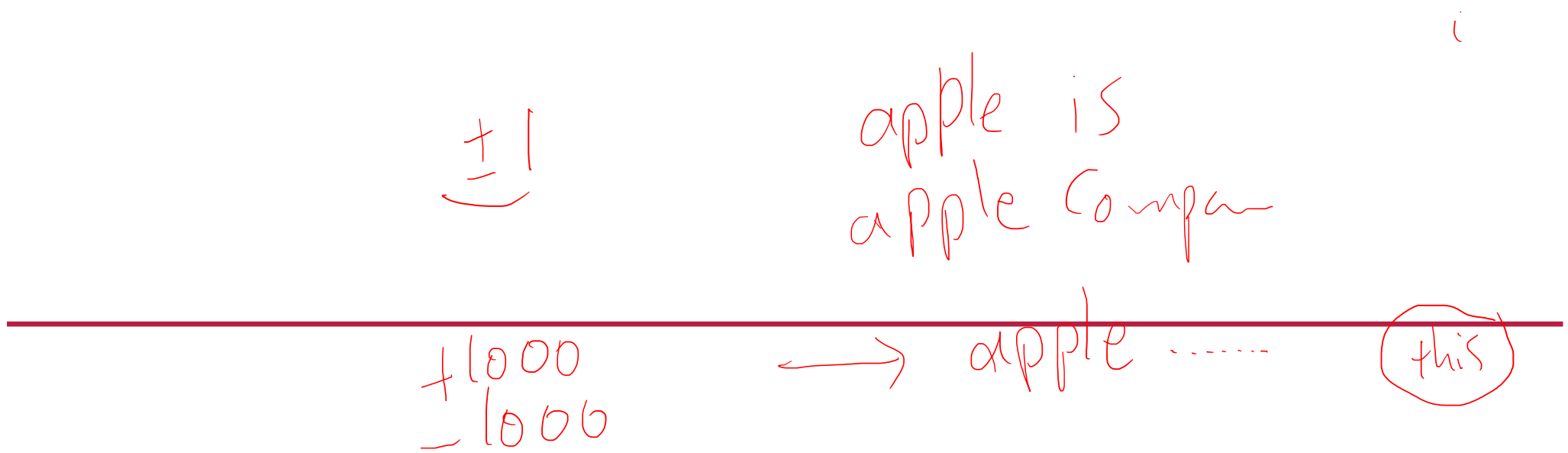


Figure 6.6 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *computer*.



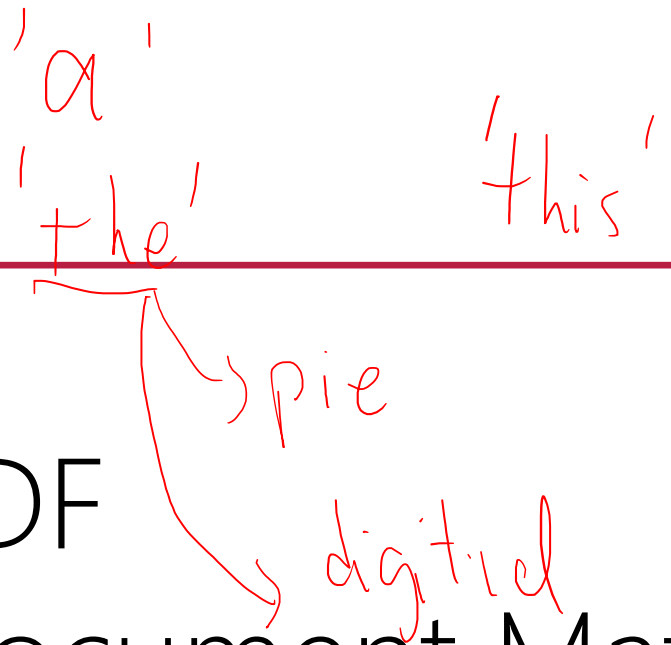
Term-Term Matrix

Matrix Size = $|Vocabs|^2$

Is it sparse or dense? Why? Does context size matter?

Term-Term Matrix

Term-Document Matrix = Term-Term Matrix
if Context Window = ± 1



TF-IDF
Weighted Term-Document Matrix

Context Window = Document

|Vocabs| × |Documents|

Term Frequency = TF = Term-Document Matrix

Term-Document Matrix

Context Window = Document

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

Term-Document Matrix

Non-Discriminative Words

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Figure 6.2 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

iDF = inverse Document Frequency

$$\text{idf}_t = \left(\frac{N}{\text{df}_t}\right) \rightarrow \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

Word	df	idf
Romeo	1	1.57
fool	36	0.012
good	37	0

where N is the total number of documents in the collection

TF-iDF

$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

$$\text{tf}_{t.d} = \log_{10}(\text{count}(t,d) + 1)$$

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

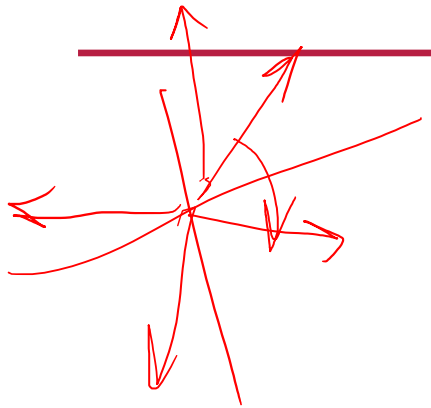
TF-iDF

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Figure 6.8 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.049 value for *wit* in *As You Like It* is the product of $\text{tf} = \log_{10}(20+1) = 1.322$ and $\text{idf} = .037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.

Cosine Similarity

the angle ∈ [0,360], Cosine Similarity ∈ [-1, 1]



$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

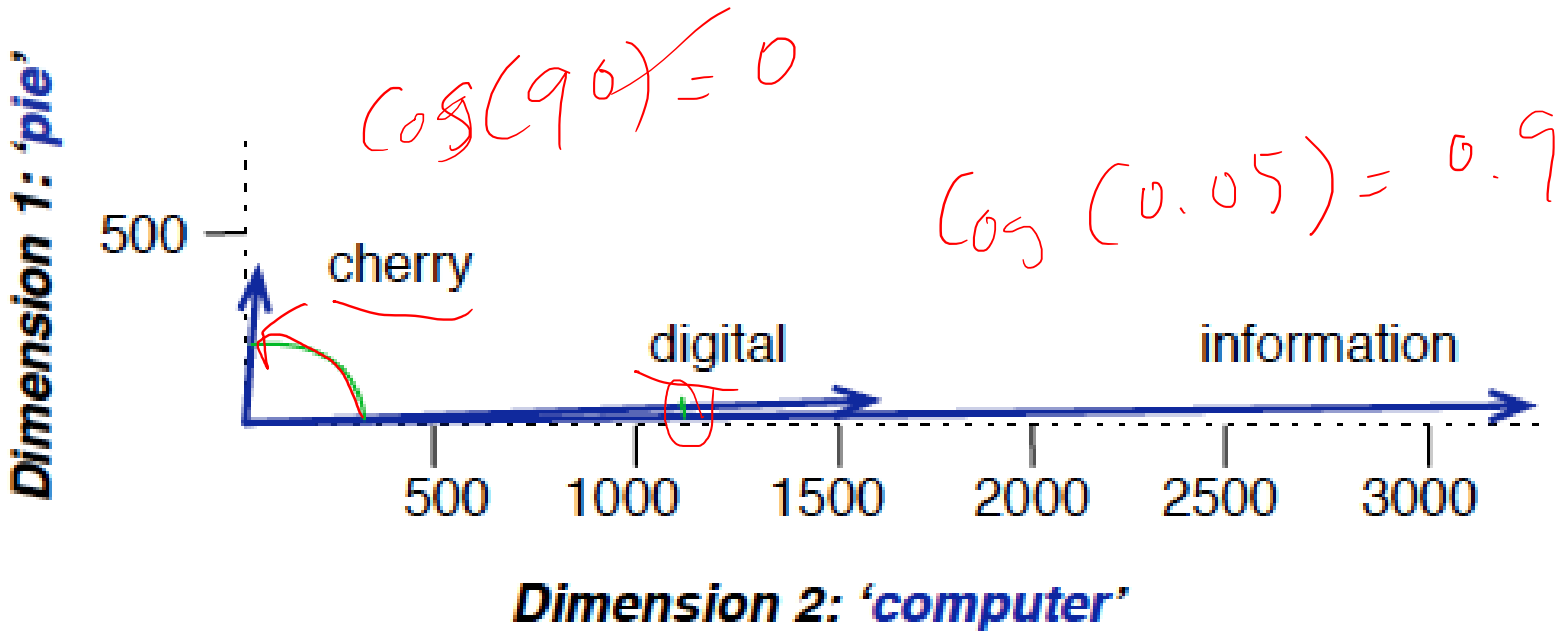
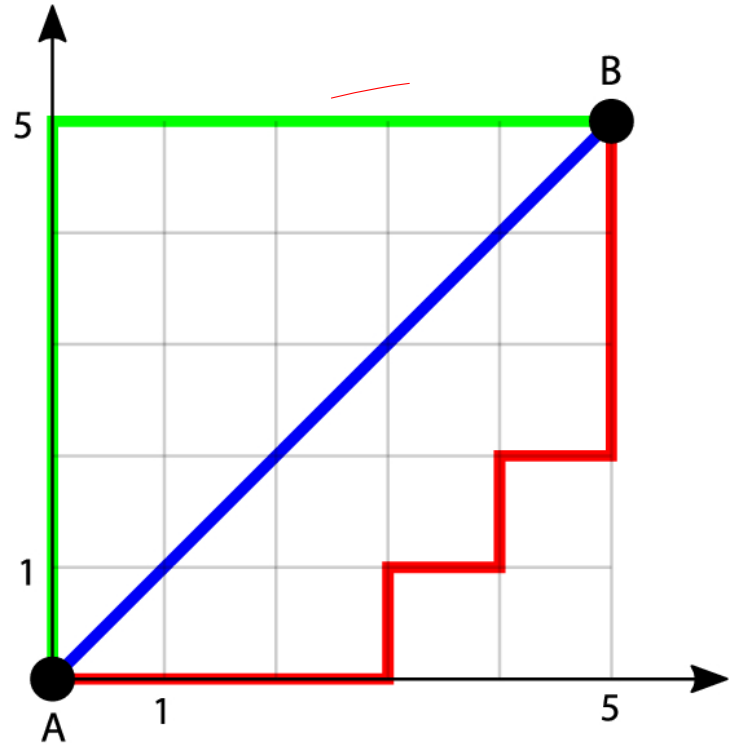


Figure 6.7 A (rough) graphical demonstration of cosine similarity, showing vectors for three words (*cherry*, *digital*, and *information*) in the two dimensional space defined by counts of the words *computer* and *pie* nearby. Note that the angle between *digital* and *information* is smaller than the angle between *cherry* and *information*. When two vectors are more similar, the cosine is larger but the angle is smaller; the cosine has its maximum (1) when the angle between two vectors is smallest (0°); the cosine of all other angles is less than 1.

Minkowski Distance

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

p = 1, Manhattan Distance
p = 2, Euclidean Distance
p = ∞, Chebychev Distance



Euclidean distance

Manhattan distance

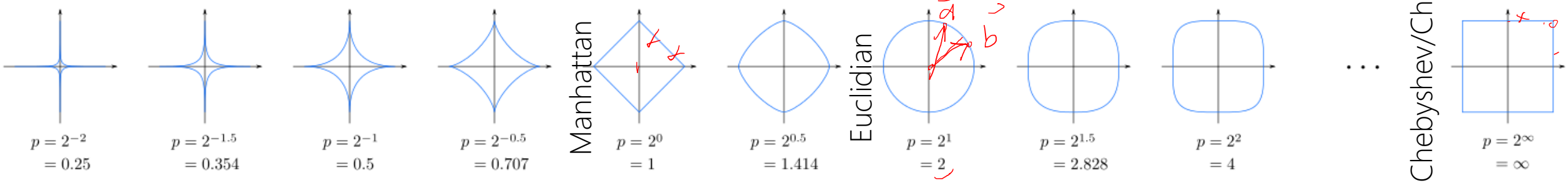
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

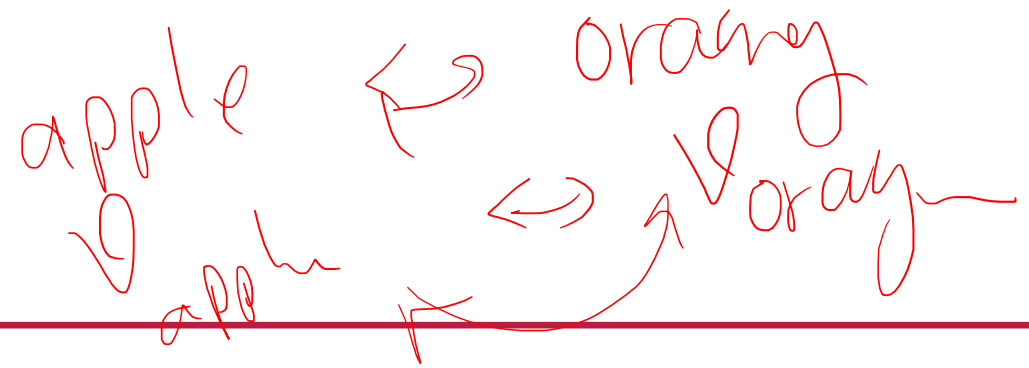
p = 1, Manhattan Distance

p = 2, Euclidean Distance

Minkowski Distance

Blue lines show all points (x,y) with same distance to the center (0,0)



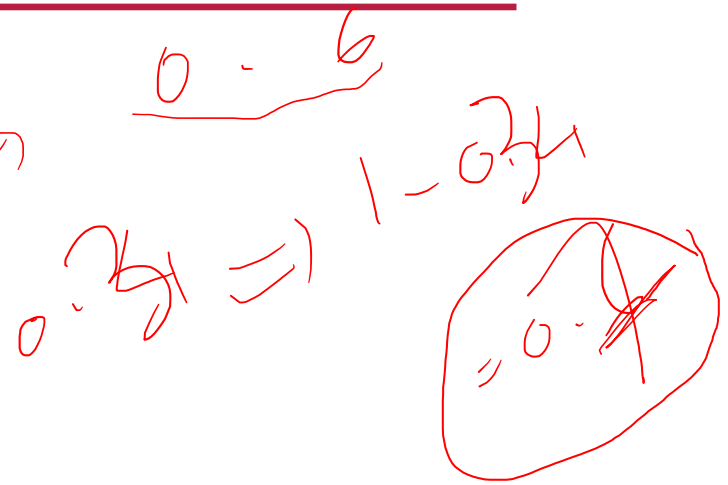


Does it matter? Why? How if yes?

Research Question (RQ)

Cosine Similarity

Minkowski Distance



Vector Semantics

Sparse

Method	Size of word/token/term vector	Sparse/Dense
Word-Documents (TF)	|Documents|	Sparse (Integer)
Term-Term	|Vocabs|	Sparse (Integer)
TF-iDF	|Vocabs|	Sparse (Real)

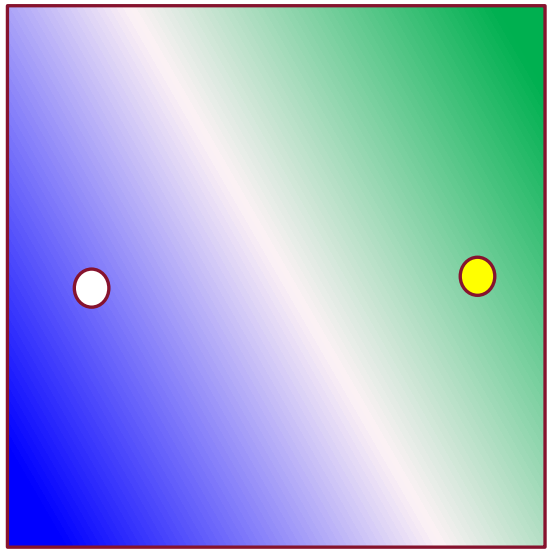
Vector Semantics Plug into LR

Sparse

Logistic Regression

Optimization: Min $-\sum_{(x,y)\in D} \text{Log } P(y|x_y)] \rightarrow \begin{cases} y = 1: P(+|x_+) = Sigmoid(f(x_+)) \\ y = 0: P(-|x_-) = 1 - Sigmoid(f(x_-)) \end{cases}$

- Function *f* is linear function of weights



$$\text{Sigmoid}([x_0\ x_1\ \ldots x_d\ 1] \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_{d+1} \end{bmatrix})$$

Vector Semantics

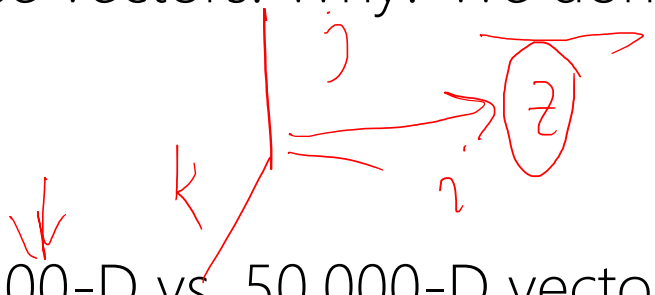
Sparse vs. Dense

Method	Size of word/token/term vector	Sparse/Dense
Word-Documents (TF)	|Documents|	Sparse (Integer)
Term-Term	|Vocabs|	Sparse (Integer)
TF-iDF	|Vocabs|	Sparse (Real)
?	10, 100, …	Dense (Real)

Vector Semantics

Sparse vs. Dense

Dense vectors work better in every NLP task than sparse vectors. Why? We don't completely understand!



Some guesses:

- Dense vectors lead to a model with less parameter: 100-D vs. 50,000-D vectors for a simple binary classifier
  - Generalize better
  - Avoid overfitting
- Captures word semantic dependencies
  - Do a better job of capturing synonymy than sparse vectors.
  - In word space, each dimension is a word. However, these dimensions may not be independent!

Dimensionality Reduction

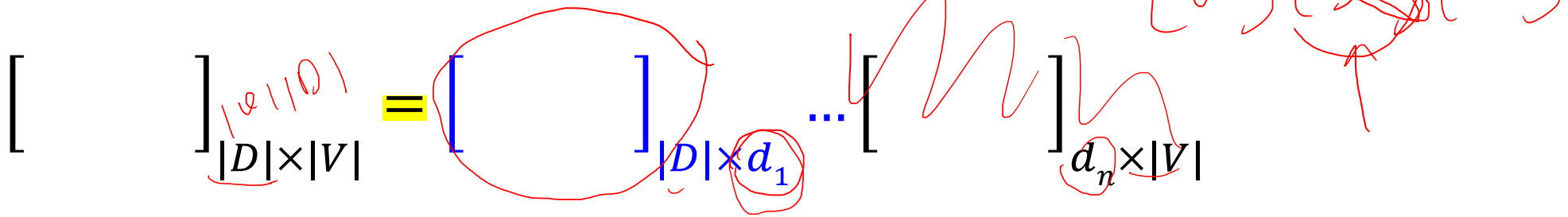
Drop less informative dimensions (columns)

Stop-words

Dimensionality Reduction

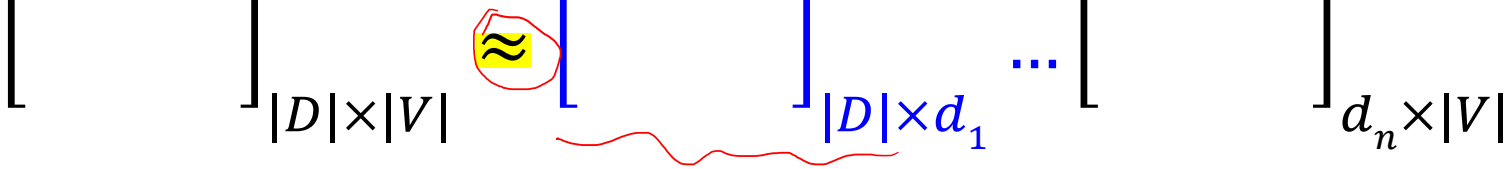
Matrix Factorization (Decomposition): LU, QR, SVD, …

https://en.wikipedia.org/wiki/Matrix_decomposition



Dimensionality Reduction

Matrix Factorization (Decomposition)



Word2Vec

Learning Word Representations

Word2Vec as LR for w_i

Extreme Distributional Semantics: Bigrams

$w_i w_{i+1}$ are semantically similar → Same Class (+)

$w_i w_{i+2}$ are not semantically similar → Different Class (-)

$$P(+|w_i w_{i+1}) = \text{sigmoid}[Vw_i \cdot Vw_{i+1}][\text{Weights}] = 1.0$$

$$P(-|w_i w_{i+2}) = 1 - P(+|w_i w_{i+2}) = 1 - \text{sigmoid}[Vw_i \cdot Vw_{i+2}][\text{Weights}] = 1.0$$

if we fix Vw_i → [Weights] of Vw_i

Word2Vec as LR for w_j

Extreme Distributional Semantics: Bigrams

w_j w_{j+1} are semantically similar → Same Class (+)

w_j w_{j+2} are not semantically similar → Different Class (-)

$P(+|\ w_j w_{j+1}) = \text{sigmoid}[Vw_j : Vw_{j+1}][\text{Weights}] = 1.0$

$P(-|w_j w_{j+2}) = 1 - P(+|w_j w_{j+2}) = 1 - \text{sigmoid}[Vw_j : Vw_{j+2}][\text{Weights}] = 1.0$

if we fix Vw_j → [Weights] of Vw_j

Word2Vec as LR

$Vw_j \propto Vw_j \Rightarrow$ close

[Weights] of $Vw_i \propto$ [Weights] of Vw_j

What can we tell about the [Weights]?

Word2Vec as LR

$$V w_j \propto V w_j$$

$$[\text{Weights}] \text{ of } V w_i \cdot [\text{Weights}] \text{ of } V w_j$$

What can we tell about the cosine of [Weights]?

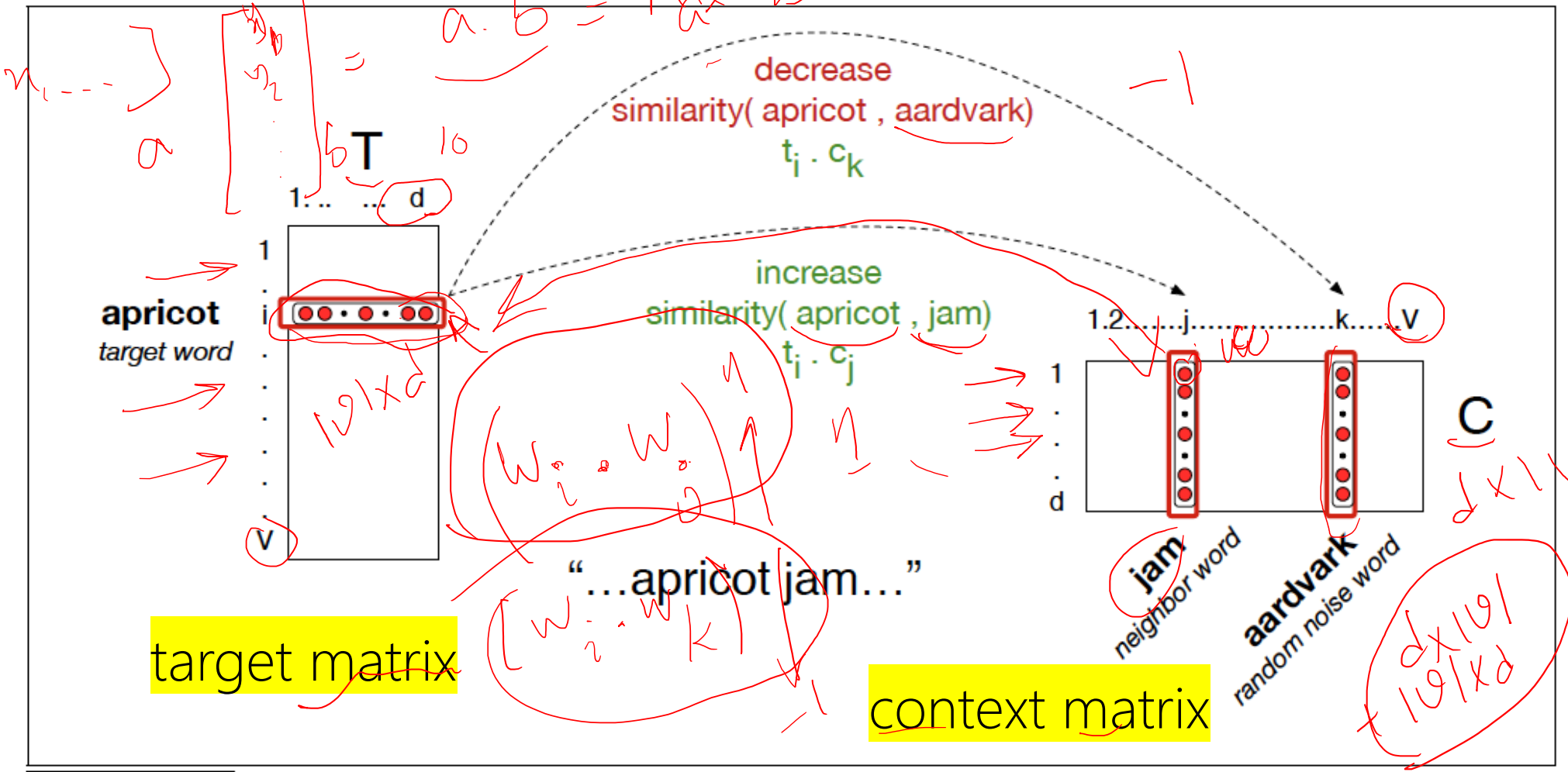


Figure 6.12 The skip-gram model tries to shift embeddings so the target embeddings (here for *apricot*) are closer to (have a higher dot product with) context embeddings for nearby words (here *jam*) and further from (have a lower dot product with) context embeddings for words that don't occur nearby (here *aardvark*).

Is it possible to use only one matrix?

Word2Vec

Given a context: … [tablespoon of apricot jam, a] …

Choose a word as target word *t*: apricot

Choose others as context word c_i: jam, tablespoon

Estimate d-dimensional vectors for *t* and all c_i

Such that they are close to each other in d-dimensional space

d ≪ |Vocabs| or |Documents|

Close V_t *and* V_{c_i} → $V_t \cdot V_{c_i} \approx 1$ → $\sigma(V_t \cdot V_{c_i}) = \frac{1}{1+e^{-(V_w \cdot V_{c_i})}}$

Word2Vec

Given a context: … [tablespoon of apricot jam, a] …

Choose a word as target word *t*: apricot

Choose *random* word n_i from out of context: car, phone, …

Estimate d-dimensional vectors for *t* and all n_i

Such that they are far from each other in d-dimensional space

d ≪ |Vocabs| or |Documents|

distant V_t *and* $V_{n_i} \rightarrow V_t \cdot V_{n_i} \approx -1$

$$V_t \cdot -V_{n_i} \approx +1 \rightarrow \sigma(V_t \cdot -V_{n_i}) = \frac{1}{1+e^{+(V_w \cdot V_{n_i})}}$$

Word2Vec

$$P(+|t,c) = \frac{1}{1+e^{-t\cdot c}}$$

Independent Assumption: P(x,y) = p(x)p(y)

$$P(+|t,c_{1:k}) = \prod_{i=1}^{k} \frac{1}{1+e^{-t\cdot c_i}}$$

$$\log P(+|t,c_{1:k}) = \sum_{i=1}^{k} \log \frac{1}{1+e^{-t\cdot c_i}}$$

$$L(\theta) = \sum_{(t,c)\in+} \log P(+|t,c) + \sum_{(t,c)\in-} \log P(-|t,c)$$

$$L(\theta) = \log P(+|t,c) + \sum_{i=1}^{k} \log P(-|t,n_i)$$

$$= \log \sigma(c\cdot t) + \sum_{i=1}^{k} \log \sigma(-n_i\cdot t)$$

Google Scholar

Tomas Mikolov

FOLLOW

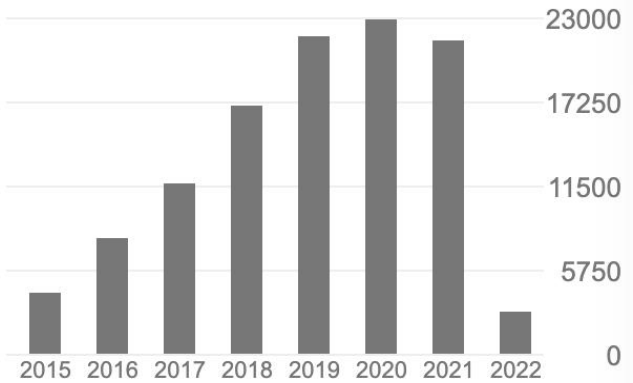
Senior Researcher, CIIRC CTU
Verified email at fb.com

Artificial Intelligence Machine Learning Language Modeling Natural Language Processing

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	32433	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	27631	2013
Distributed representations of sentences and documents Q Le, T Mikolov International conference on machine learning, 1188-1196	9054	2014

Cited by VIEW ALL

	All	Since 2017
Citations	113645	98036
h-index	46	44
i10-index	82	72



Word2Vec as MF

$$\mathrm{V}w_j \propto \mathrm{V}w_j$$

$$[\text{Weights}] \text{ of } \mathrm{V}w_i \cdot [\text{Weights}] \text{ of } \mathrm{V}w_j$$

$$|\mathrm{V}w_j \cdot \mathrm{V}w_j - ([\text{Weights}]_{\mathrm{V}wi} \cdot [\text{Weights}]_{\mathrm{V}wj})| \approx 0$$

Word2Vec

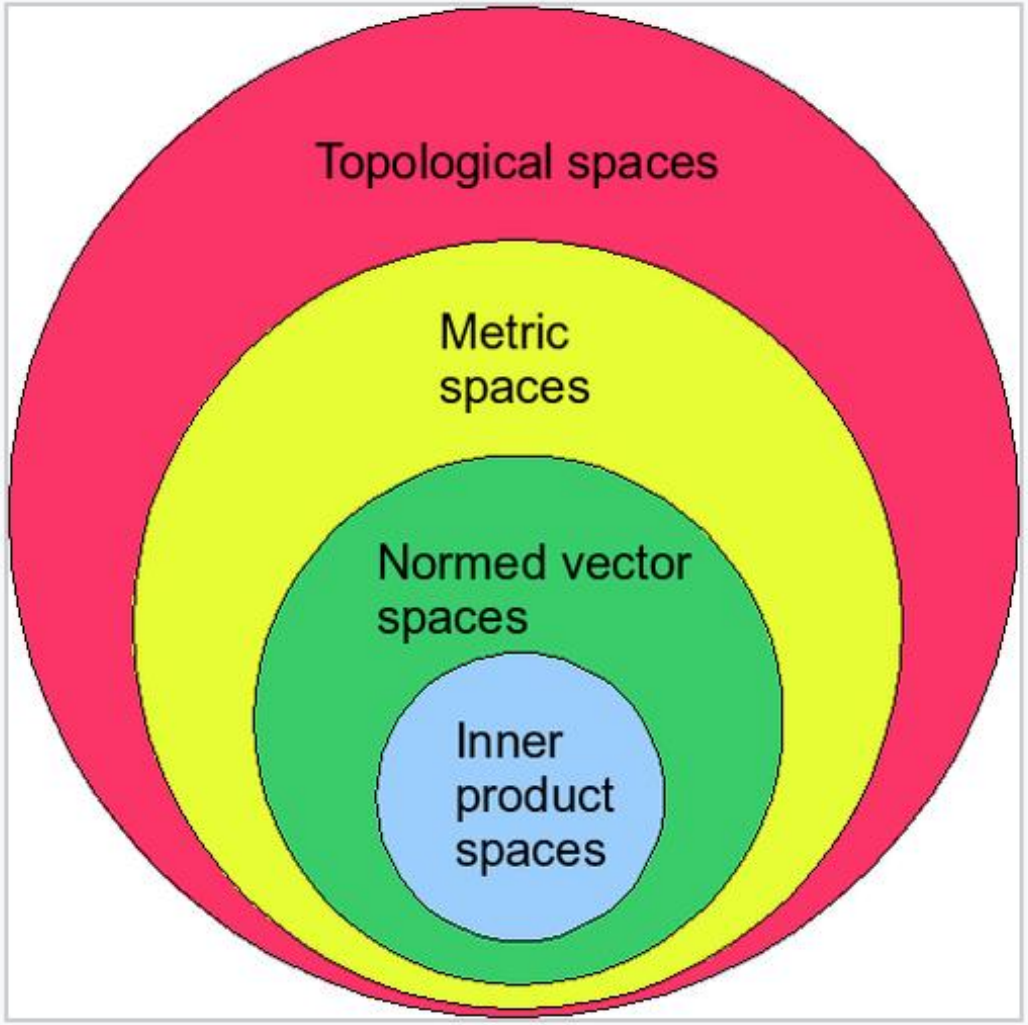
- **Context Window?** Longer vs. Shorter?
- **Deterministic?** Any runs of training ended with same set of vectors?
- **Transformation?** rotation, flips, shear (skew), …
- **Which signifier:**
  1. [cat], [miu], [*image_of_cat*], [ascii_cat],
  2. Count-based: [*tf*], [*tf-idf*], …
  3. Learning methods: [word2vec]

Pre-trained Word Vectors

Available in genism python libarary:

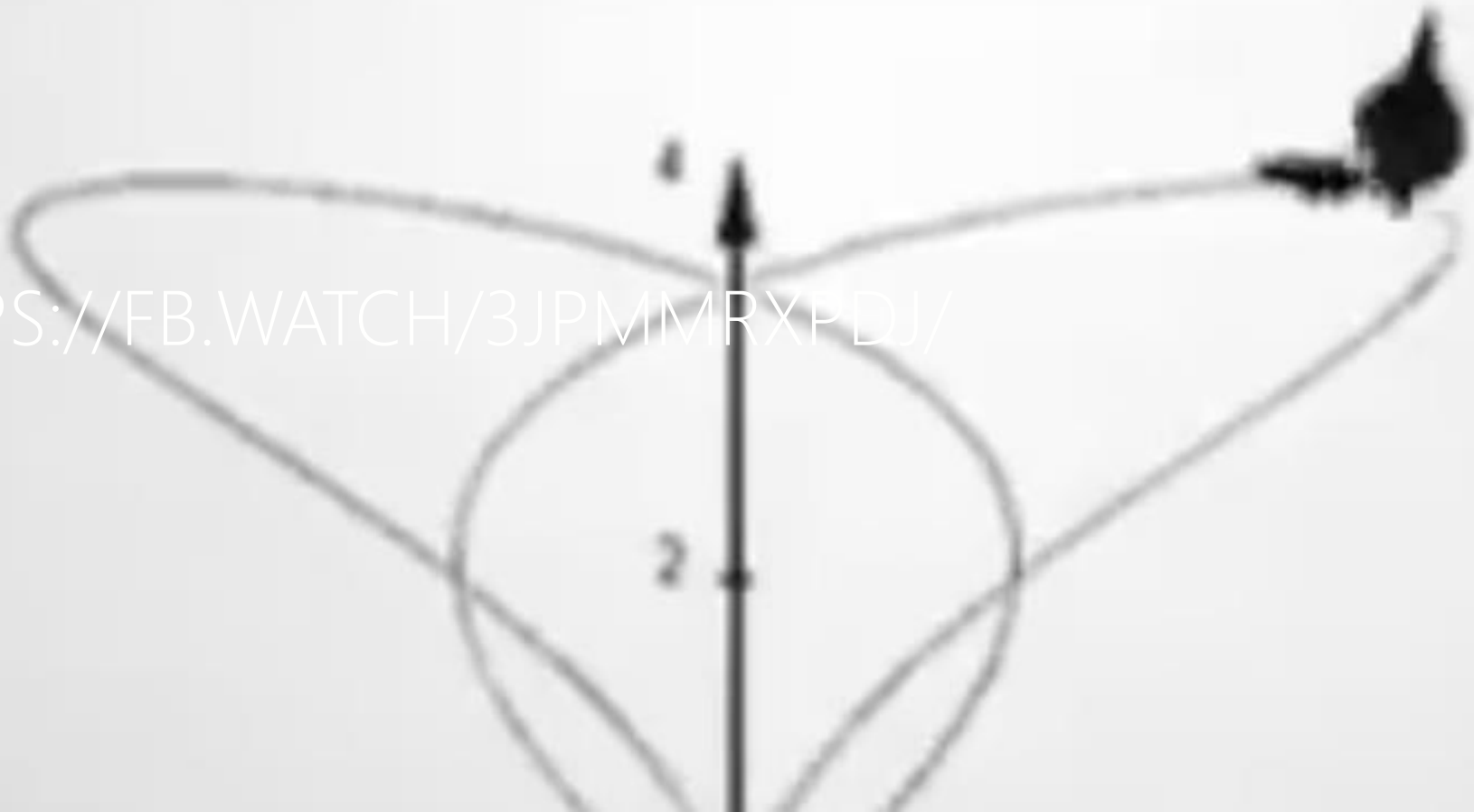
- conceptnet-numberbatch-17-06-300 (1917247 records): ConceptNet Numberbatch consists of state…
- fasttext-wiki-news-subwords-300 (999999 records): 1 million word vectors trained on Wikipe…
- glove-twitter-100 (1193514 records): Pre-trained vectors based on 2B tweets,…
- glove-twitter-200 (1193514 records): Pre-trained vectors based on 2B tweets, …
- glove-twitter-25 (1193514 records): Pre-trained vectors based on 2B tweets, …
- glove-twitter-50 (1193514 records): Pre-trained vectors based on 2B tweets, …
- glove-wiki-gigaword-100 (400000 records): Pre-trained vectors based on Wikipedia 2…
- glove-wiki-gigaword-200 (400000 records): Pre-trained vectors based on Wikipedia 2…
- glove-wiki-gigaword-300 (400000 records): Pre-trained vectors based on Wikipedia 2…
- glove-wiki-gigaword-50 (400000 records): Pre-trained vectors based on Wikipedia 2…
- word2vec-google-news-300 (3000000 records): Pre-trained vectors trained on a part of…
- word2vec-ruscorpora-300 (184973 records): Word2vec Continuous Skipgram vectors tra…

Vector Semantics
Vector Space
Transformation
Linear Algebra



Hierarchy of mathematical spaces. Normed vector spaces are a superset of inner product spaces and a subset of metric spaces, which in turn is a subset of topological vector space.

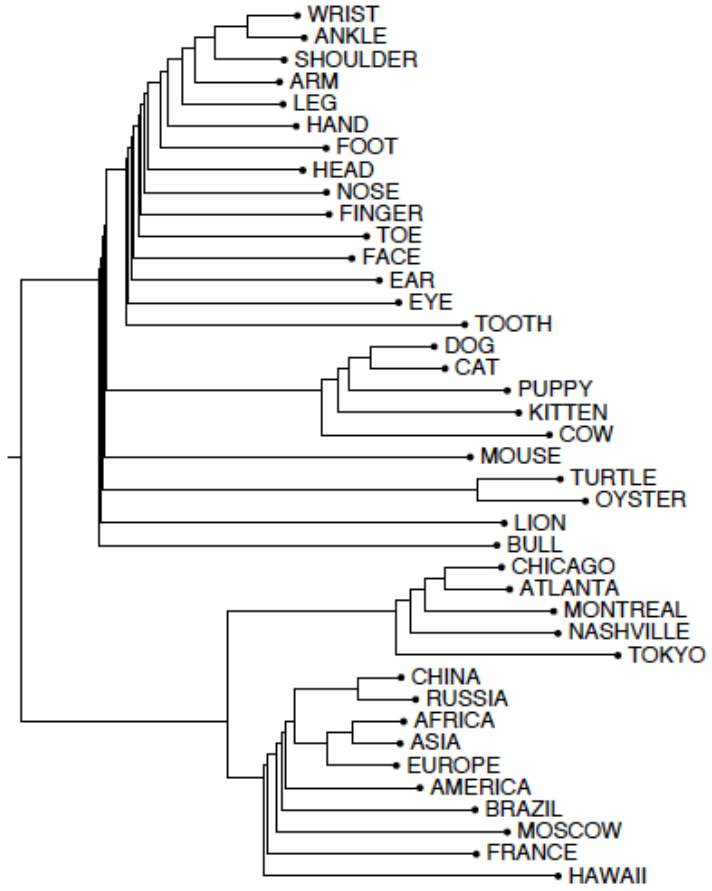
"ONE POINT OF VIEW DOES NOT SHOW THE WHOLE PICTURE"



HTTPS://FB.WATCH/3JPMMRXPDJ/

Visualization

Intuition, Geometry



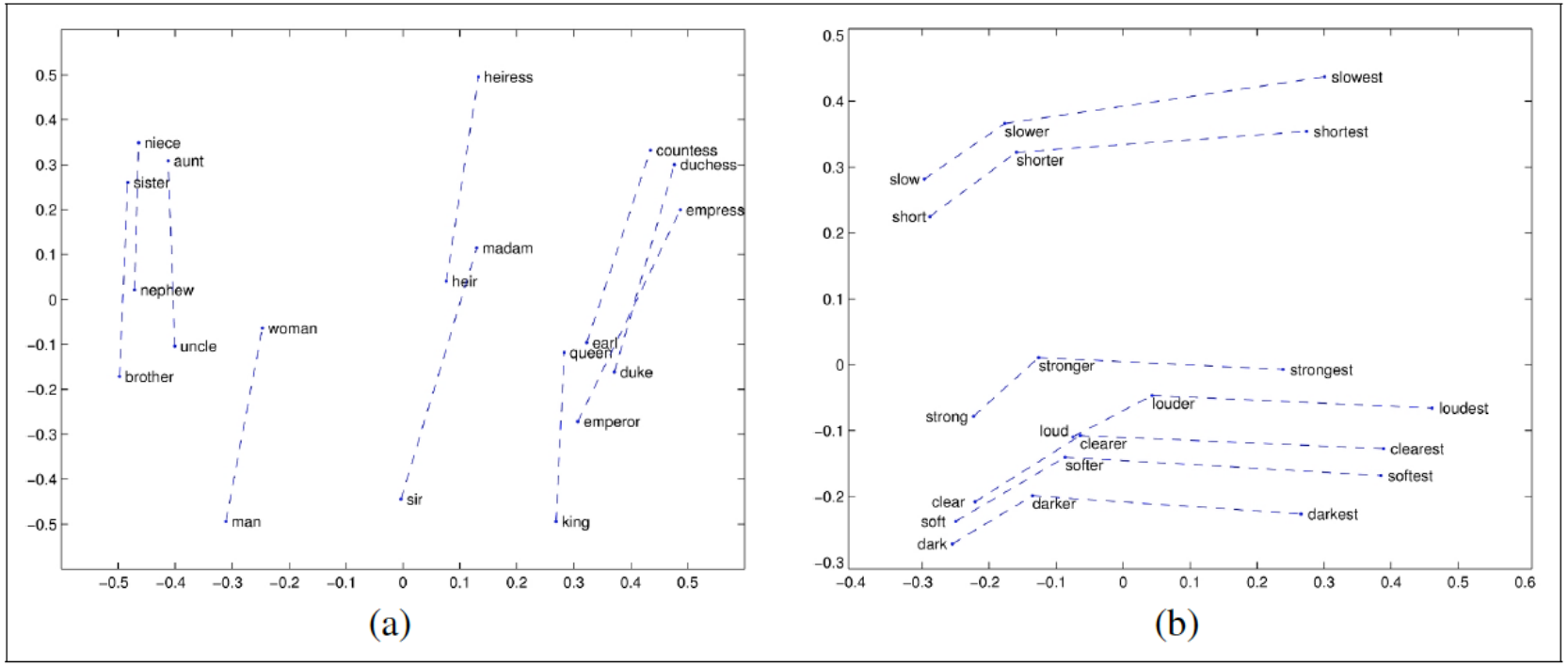


Figure 6.13 Relational properties of the vector space, shown by projecting vectors onto two dimensions. (a) 'king' - 'man' + 'woman' is close to 'queen' (b) offsets seem to capture comparative and superlative morphology (Pennington et al., 2014).

Movement

Temporality (How?)

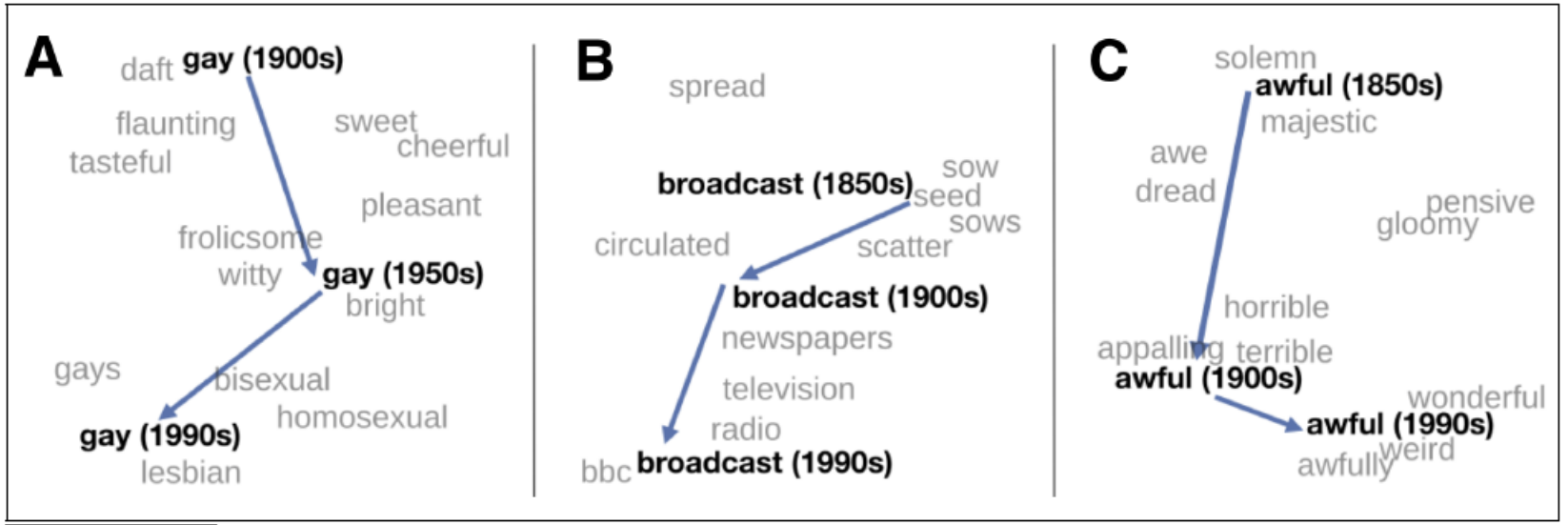


Figure 6.14 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016b).

Biases

Inherent/Latent/Hidden Distribution

- (sare, mom, nurse), (mr., ahmed, doctor, president)
- (drug, mexican), (education, usa, canada)
- (flowers, pleasant, {European-American}), (insects, ugly, {African-American})

Debiasing

- Gender-base: [he] remains masculine, [she] remains feminine, but [nurse],[doctor],[president] becomes neutral

Study of Bias in History

Evaluation

Intrinsic

- Golden Standards for Semantic Similarity/Distance
  - No Context: just pair of words
    - WordSim-353
    - SimLex-999
  - With Context:
    - Stanford Contextual Word Similarity (SCWS) (Huang et al., 2012) and the
    - Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019)

Extrinsic:

- Improve the performance of underlying task
  - Information Retrieval (IR), Document Classification, Sentiment Analysis, …

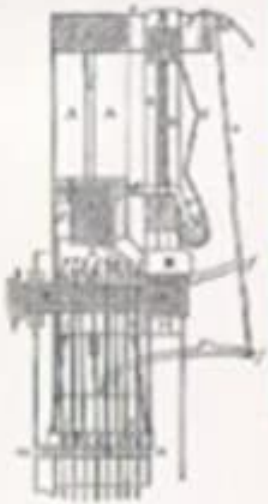
CROSS-LINGUAL WORD EMBEDDINGS

Words from two or more languages are represented in the same shared low-dimensional vector space.
Level of supervision:

- sentence-level: Machine Translation (MT) Corpora
- document-level: Wikipedia
- word-level: Bilingual Dictionaries
- unsupervised: Distribution of words in monolingual corpora in a bilingual dictionary

How to learn representation for:

- Character (autocorrection)
- Sentence/Paragraph/Documents (Doc2Vec)



SPEECH and LANGUAGE PROCESSING

An Introduction to
Natural Language Processing,
Computational Linguistics,
and Speech Recognition

DANIEL JURAFSKY & JAMES H. MARTIN