
Ceci n'est pas une pipe.
Magritte

WORD
VECTOR SPACE MODELS

The Treachery of Images

Artist	René Magritte
Year	1929
Medium	Oil on canvas
Movement	Surrealism
Dimensions	60.33 cm × 81.12 cm (23.75 in × 31.94 in)
Location	Los Angeles County Museum of Art[1]

Language Modeling

Building a model that can generate an accurate stream of tokens (words + inflection rules, punctuations, fillers, …)

Language Modeling

Building a model that can generate an accurate stream of tokens (words + inflection rules, punctuations, fillers, ...)

No semantics were involved!

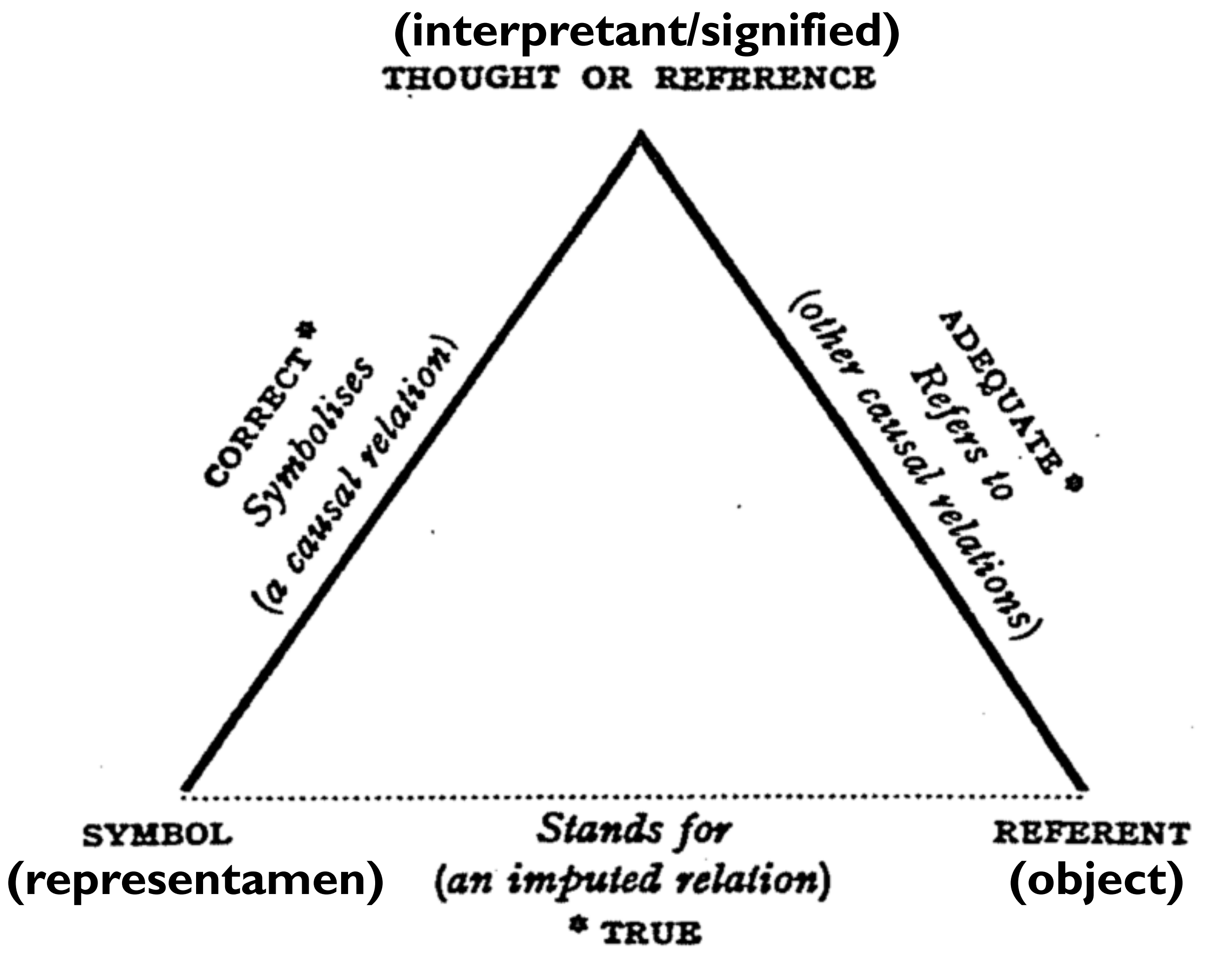
Semiotics

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

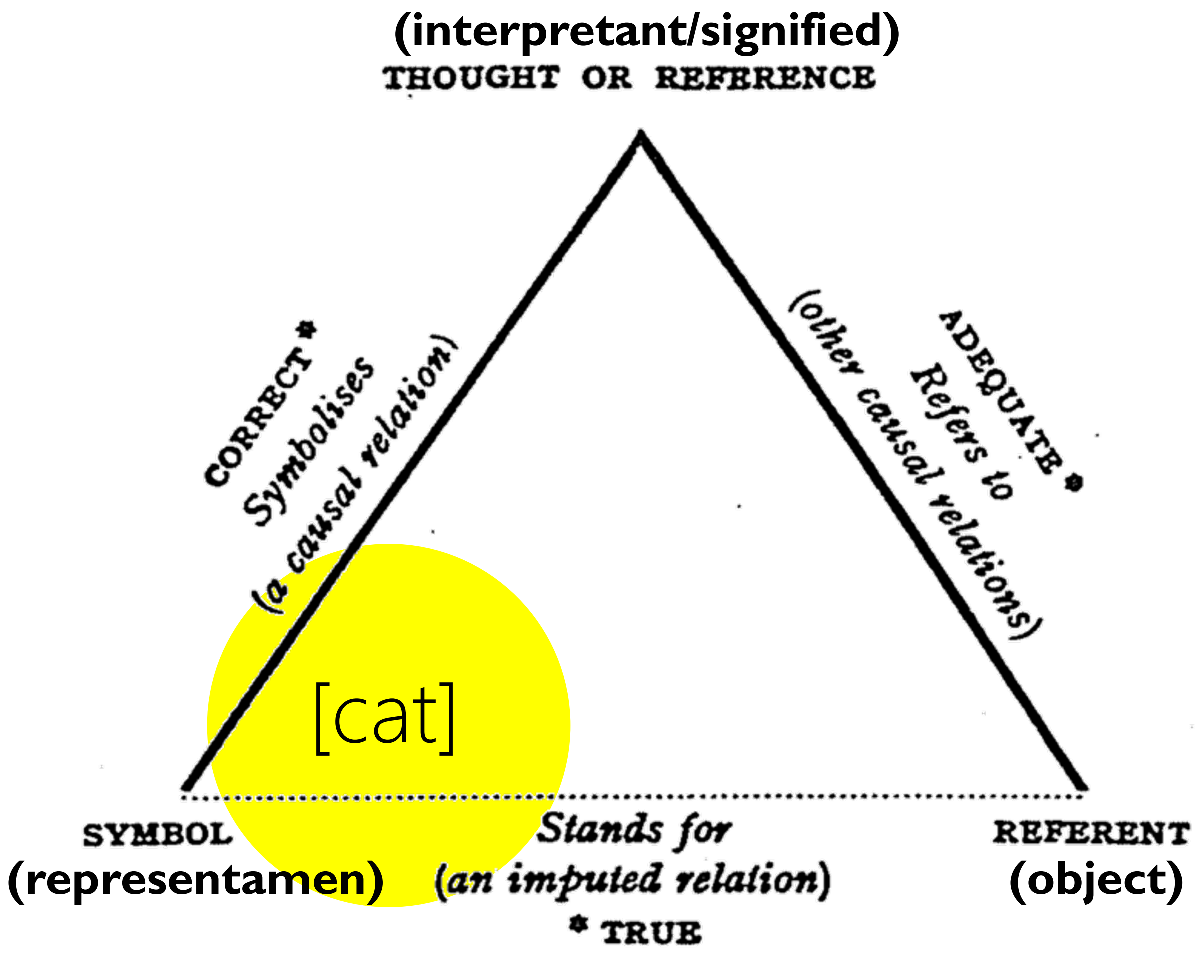
Syntactics: Relations among signs in formal structures

Pragmatics: Relation between signs and sign-using agents

Triangle of Semiotics

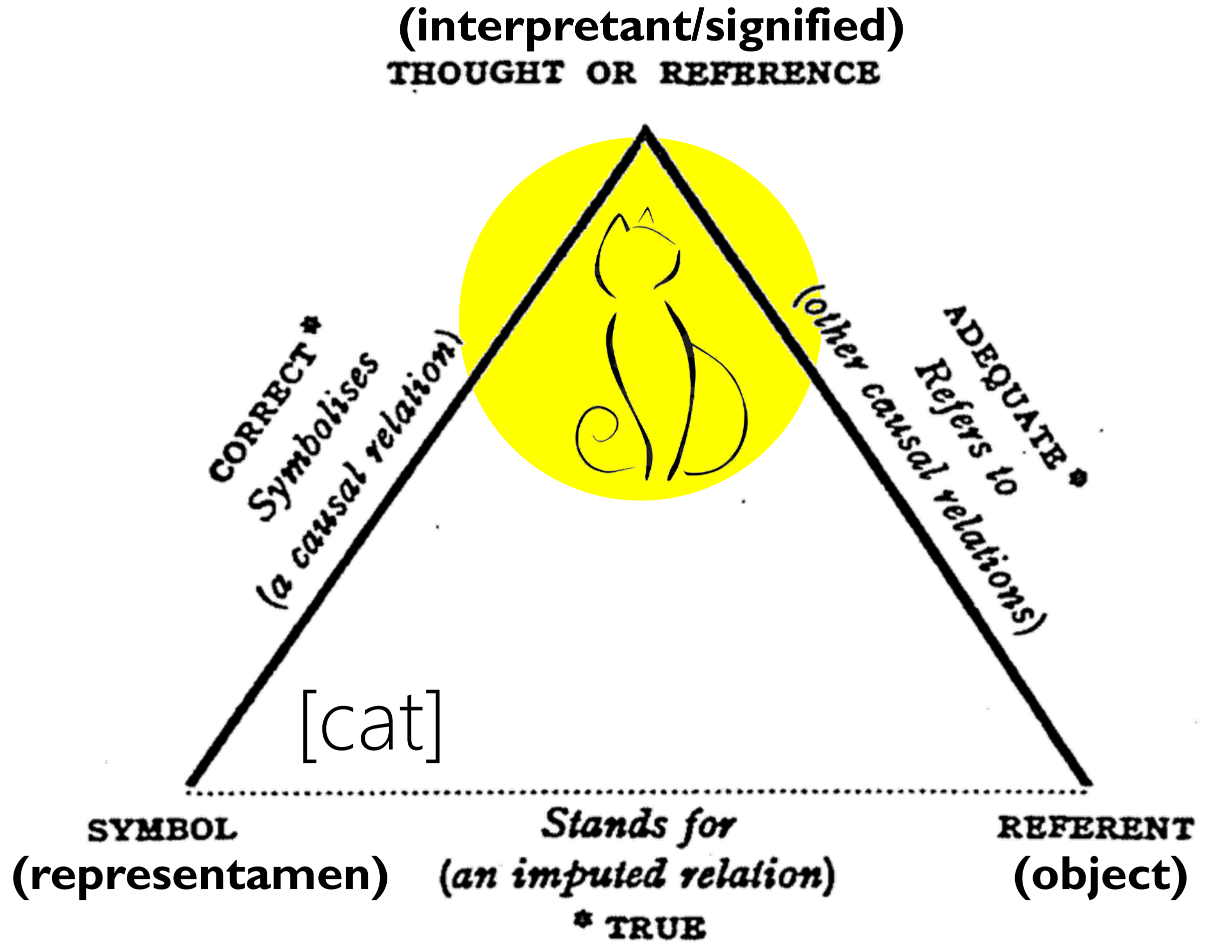


Triangle of Semiotics



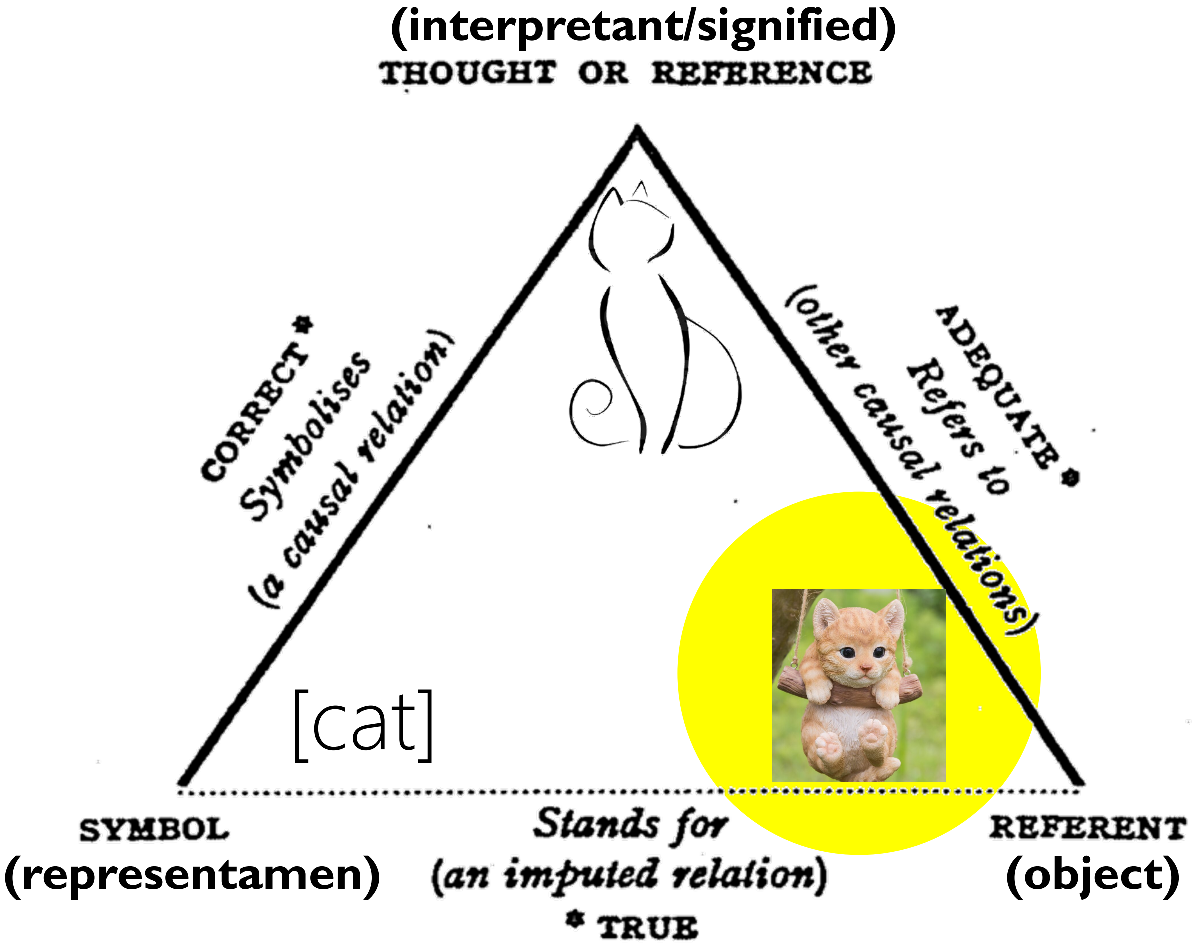
The representamen (sign vehicle, signifier, symbol) represents the object

Triangle of Semiotics



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

Triangle of Semiotics

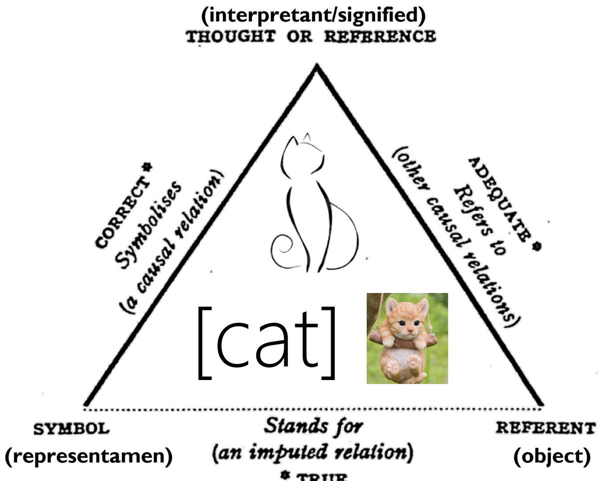


The object of a sign is always hidden!

Sign

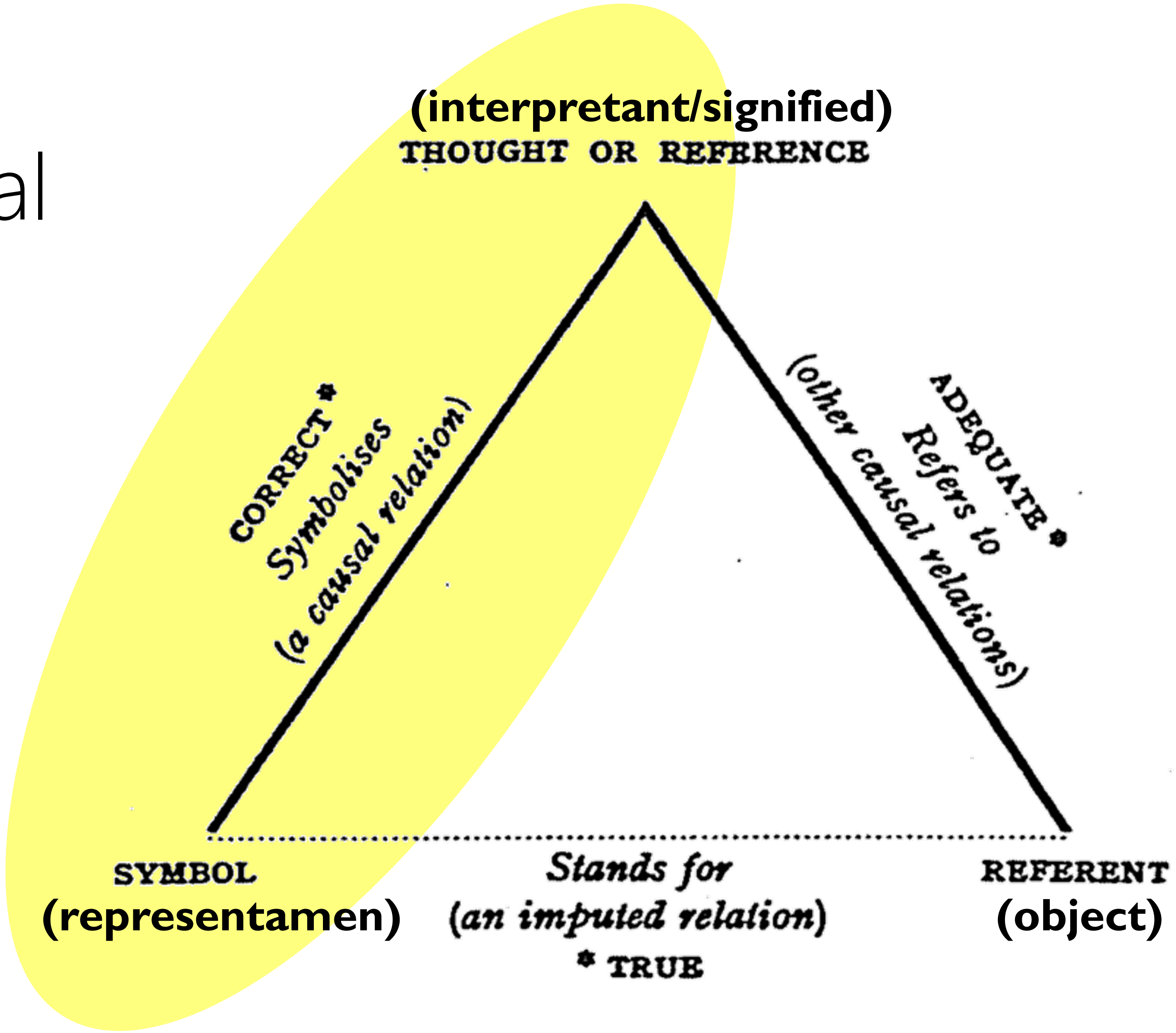
A sign is a triadic unity of:

- 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.

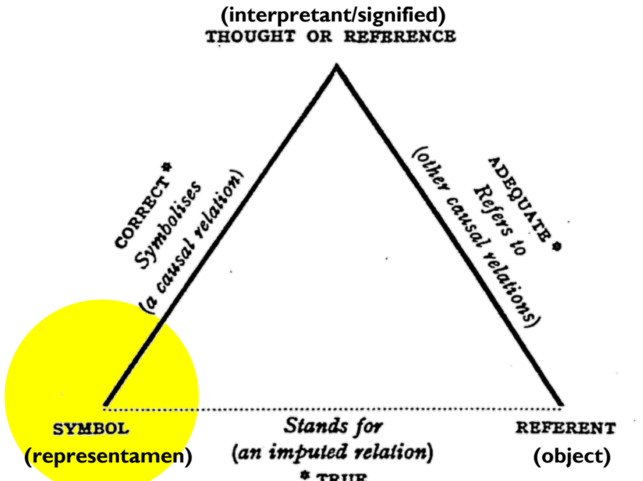


Representament Learning

Representation Learning

From school:

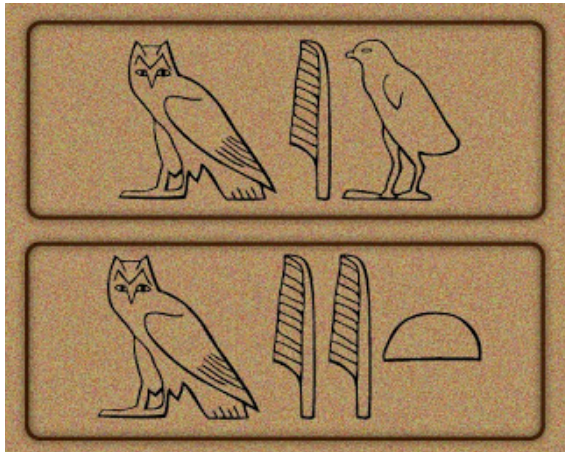
- ['c', 'a', 't']

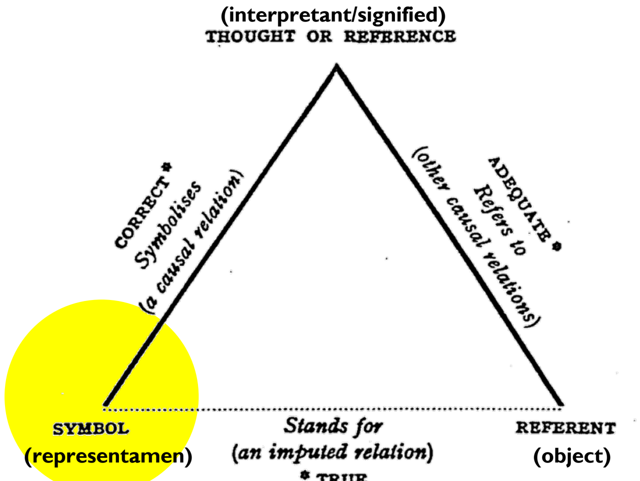


Representament Learning

Representation Learning

In hieroglyphs:
- miu/mii (male)
- Miit (female)



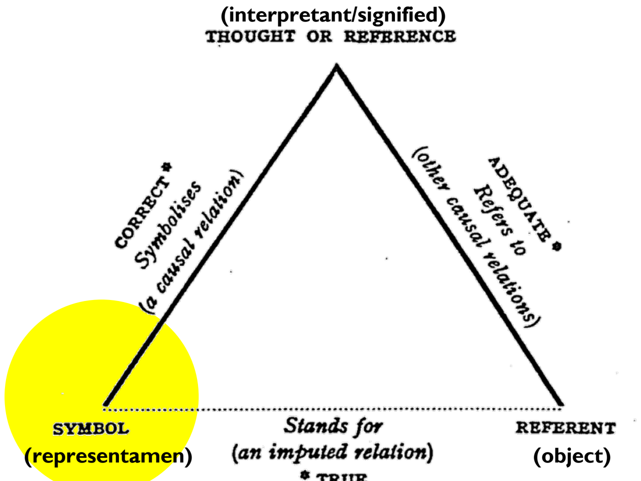


Representament Learning

Representation Learning

In computer ASCII code:

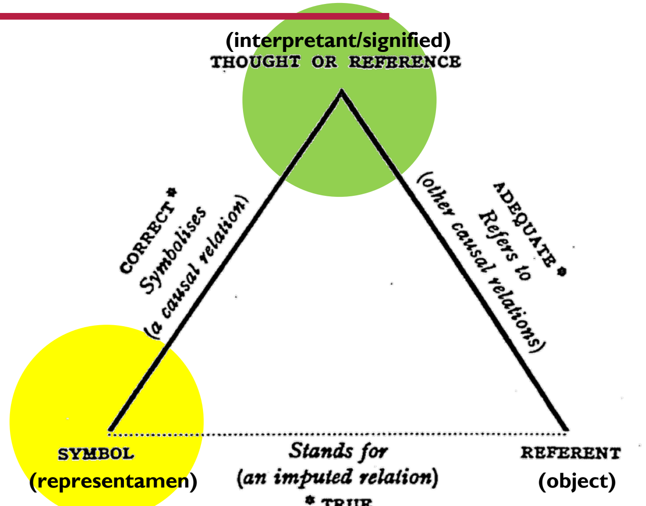
- [99, 97, 116]



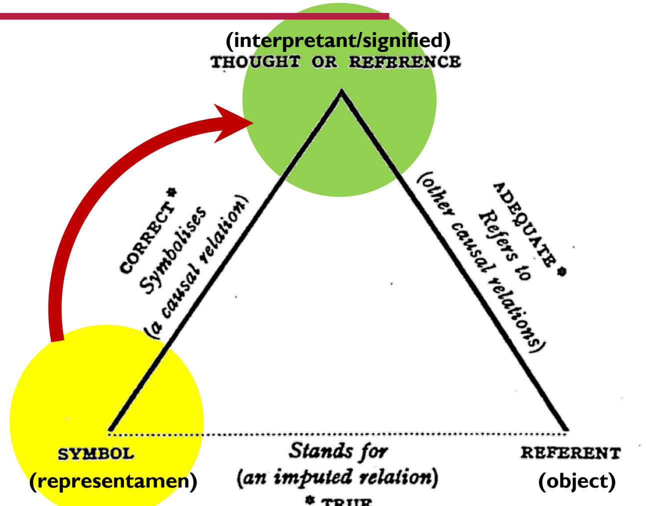
Representament Learning

Representation Learning

Help with signified/meaning/sense!

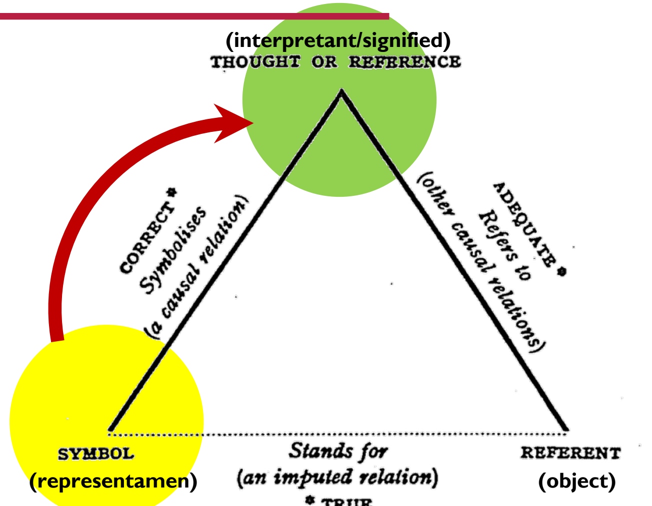


Representament → Interpretant



Representament → ?

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



cat noun, often attributive

Save Word

\ ˈkat \

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

/ˈvɪtɡənʃtaɪn, -staɪn/

1889 –1951

Austrian-British Philosopher

Worked primarily in:

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

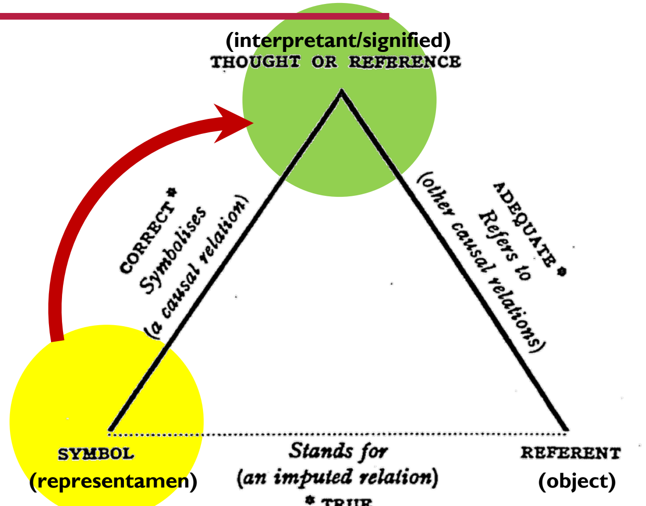
Ludwig Josef Johann Wittgenstein

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

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

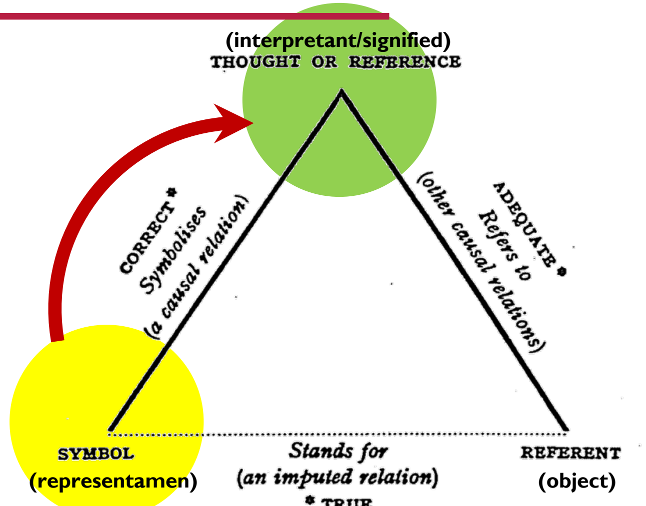
→ "the meaning of a word is its use in the language" - Philosophical Investigations.

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

- Connotations

The aspects of a word's meaning that are related to a writer or reader's emotions, sentiment, opinions, or evaluations.

E.g., happy vs. sad

Evaluation (sentiment)

E.g., great/love vs. terrible/hate

Lexical Semantics

- Similar Tokens (Word Similarity)

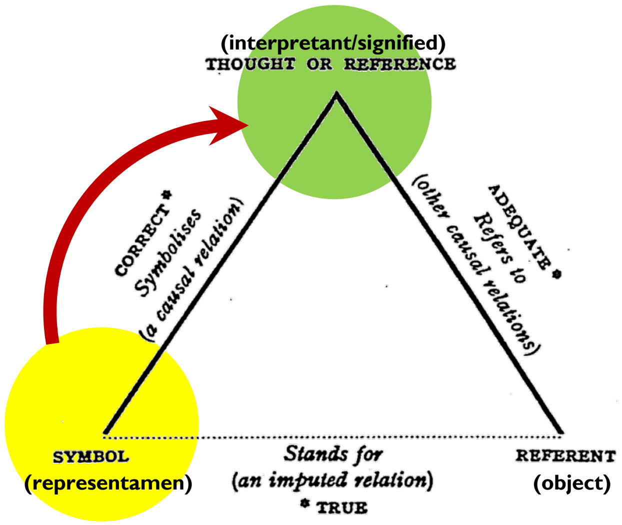
Similarity in the signified and object

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

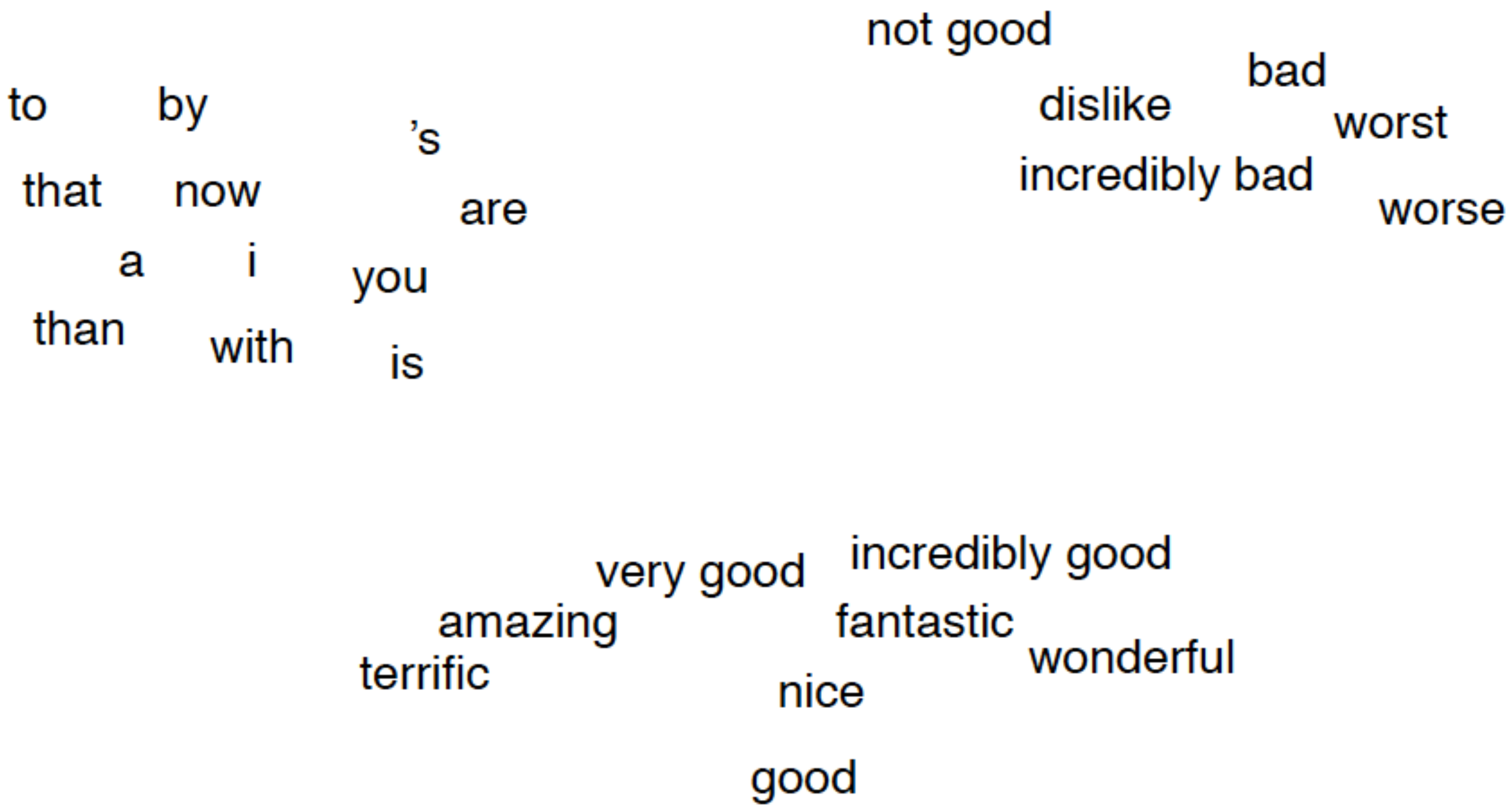
Vector Semantics

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

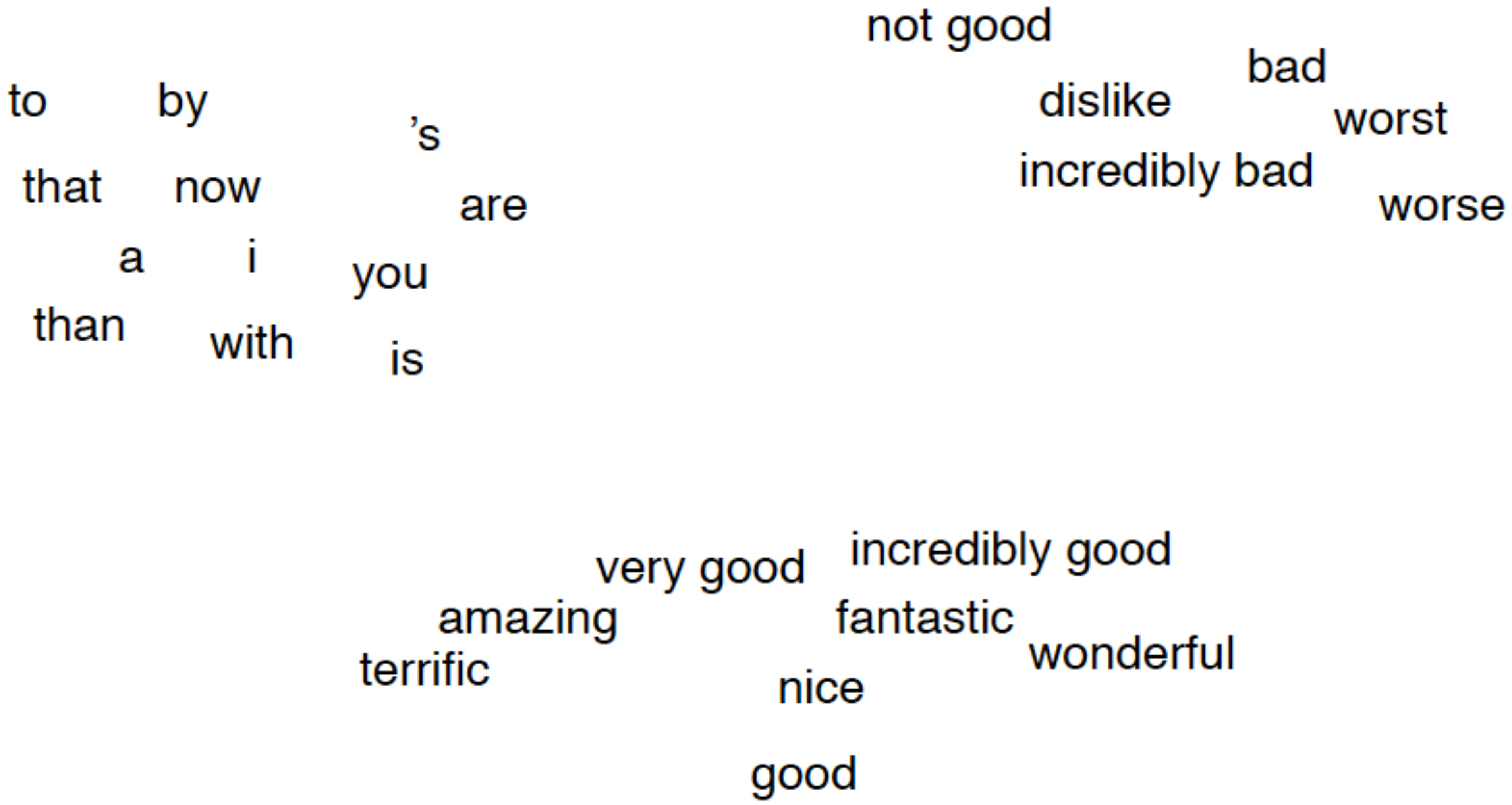
Distributional Hypothesis

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

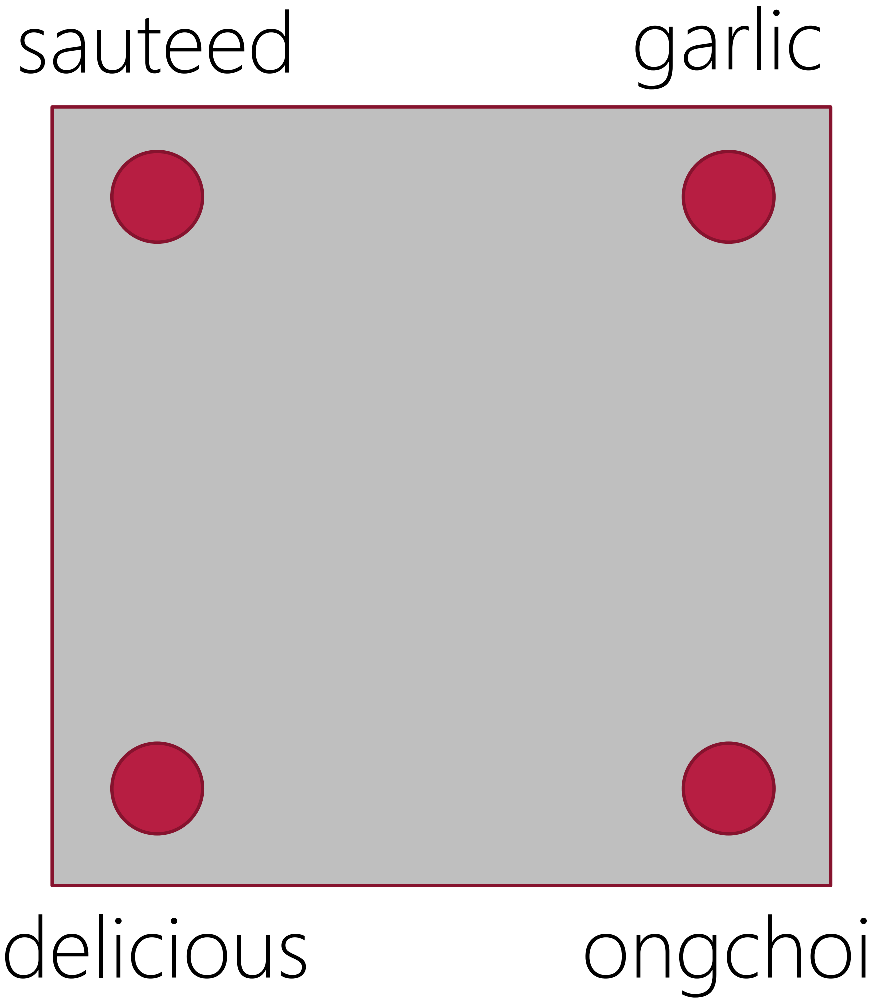
Joos, M. (1950). Description of language design. JASA, 22, 701–708.
Harris, Z. S. (1954). Distributional structure. Word, 10, 146–162.
Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955.



Our Mind Space!

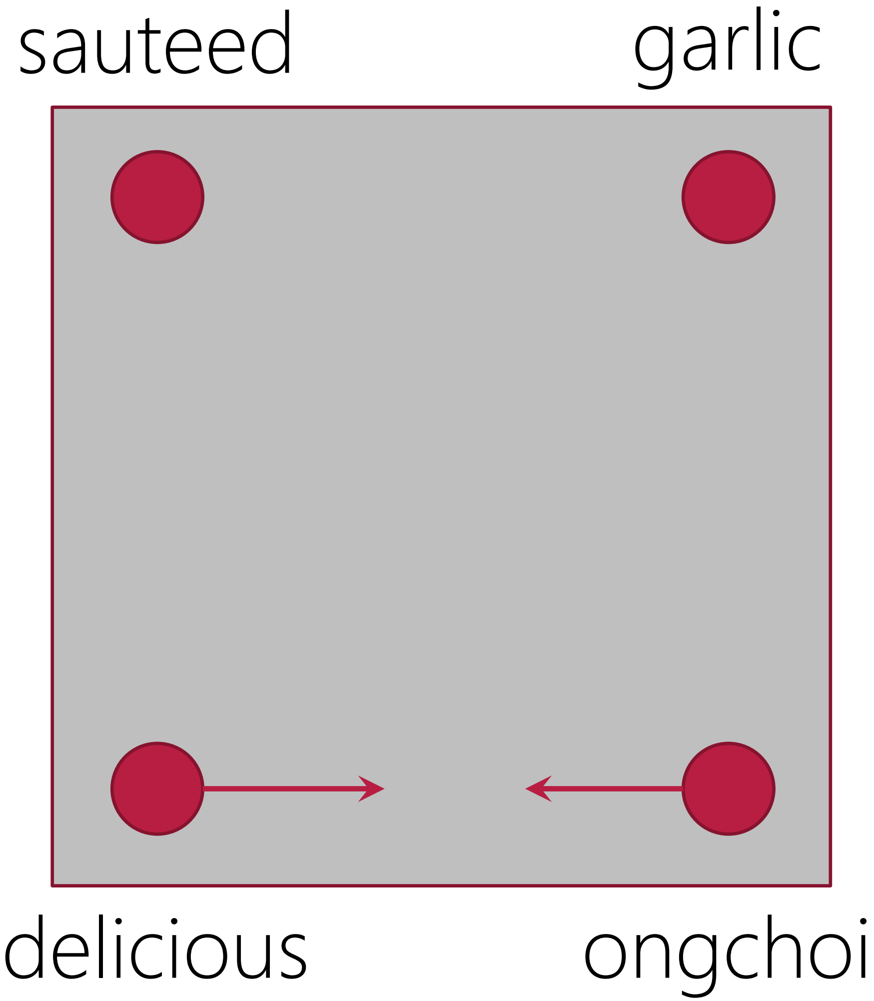

not good
to
by
bad
dislike
’s
worst
that
now
incredibly bad
are
worse
a
i
you
than
with
is
very good
incredibly good
amazing
fantastic
wonderful
terrific
nice
good

2-D Space!



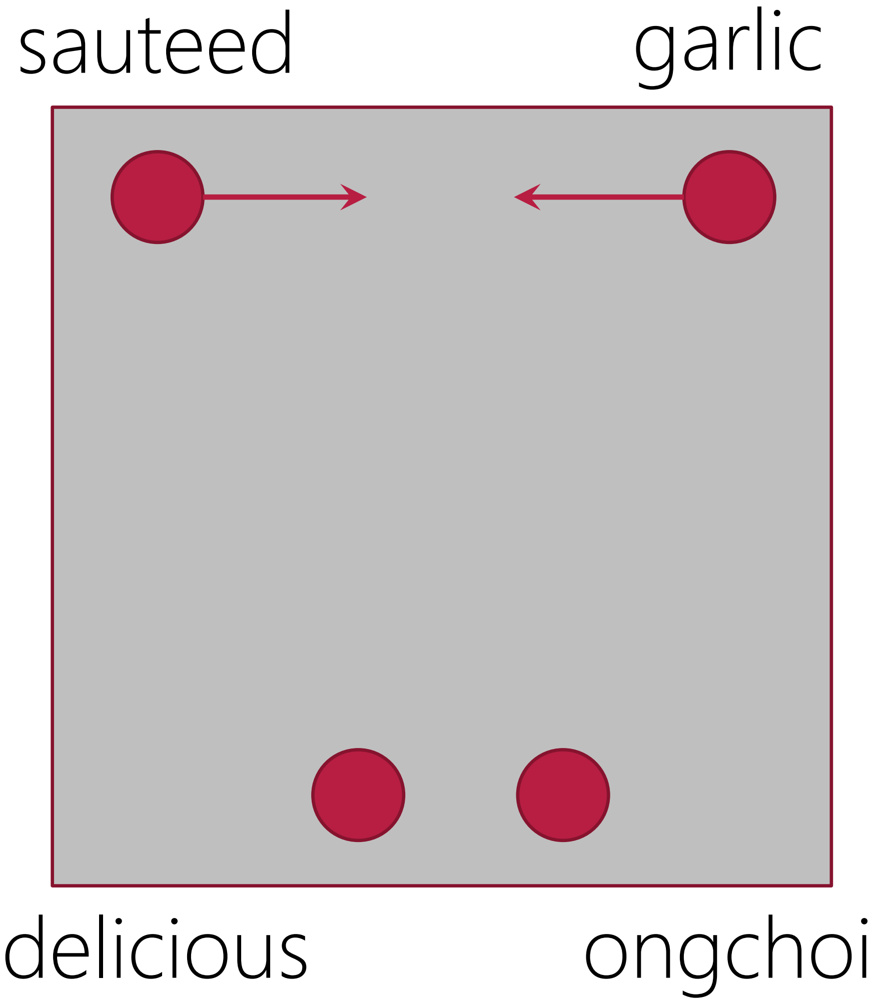
2-D Space!

$$\underbrace{\text{ongchoi is delicious}}_{\text{context window of size 3 tokens}}$$



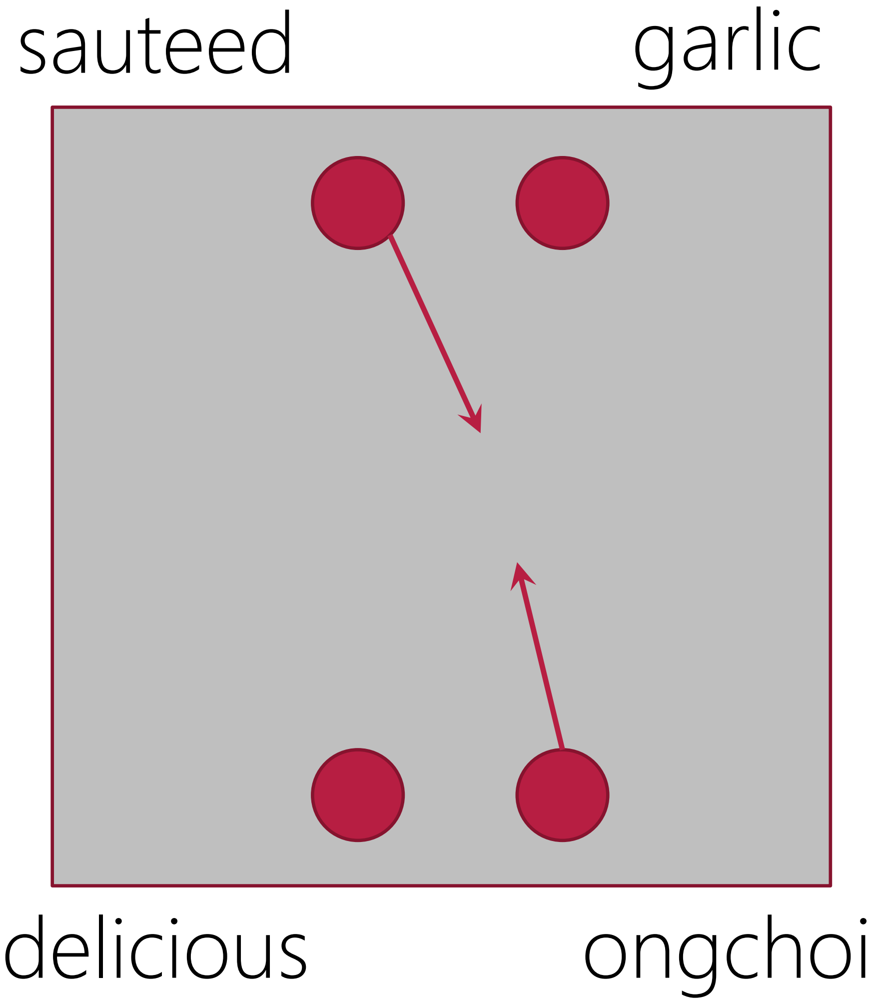
2-D Space!

ongchoi is delicious
ongchoi can be sauteed with garlic



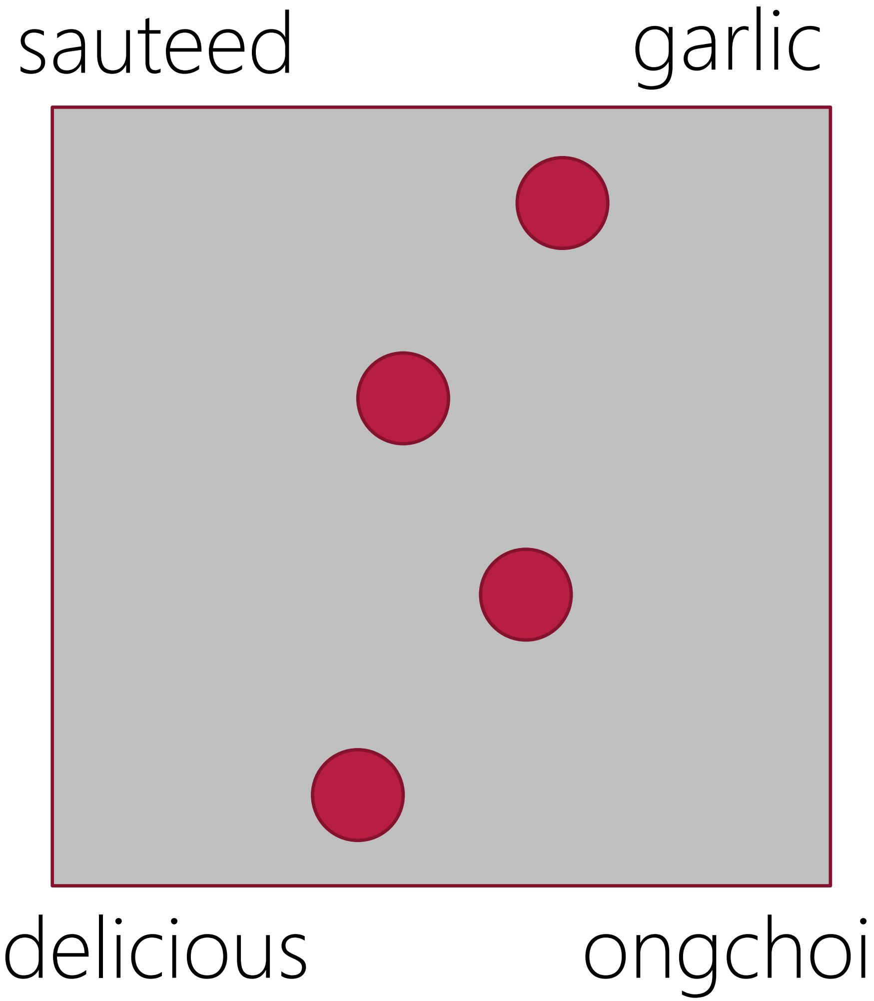
2-D Space!

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

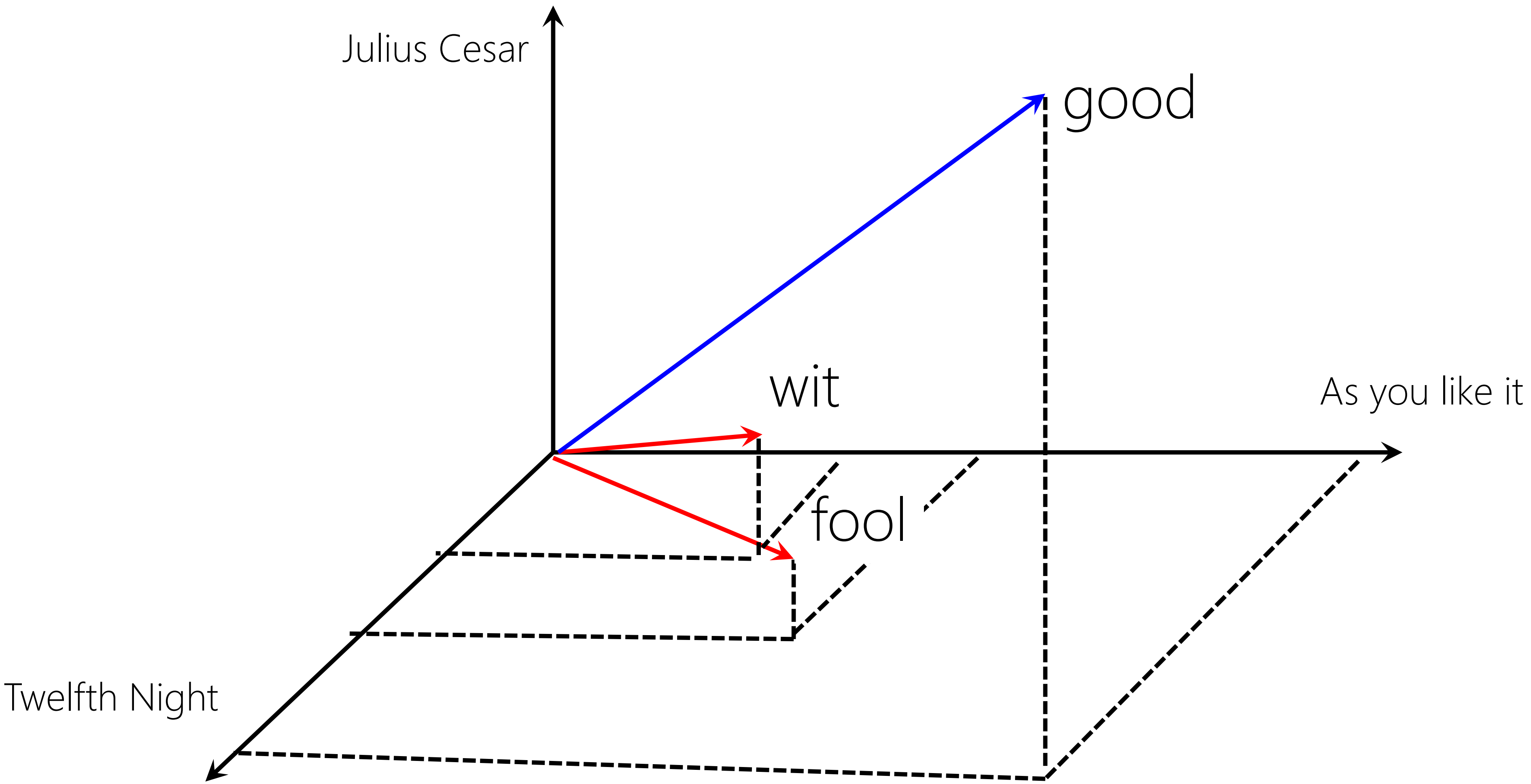
	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

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.

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

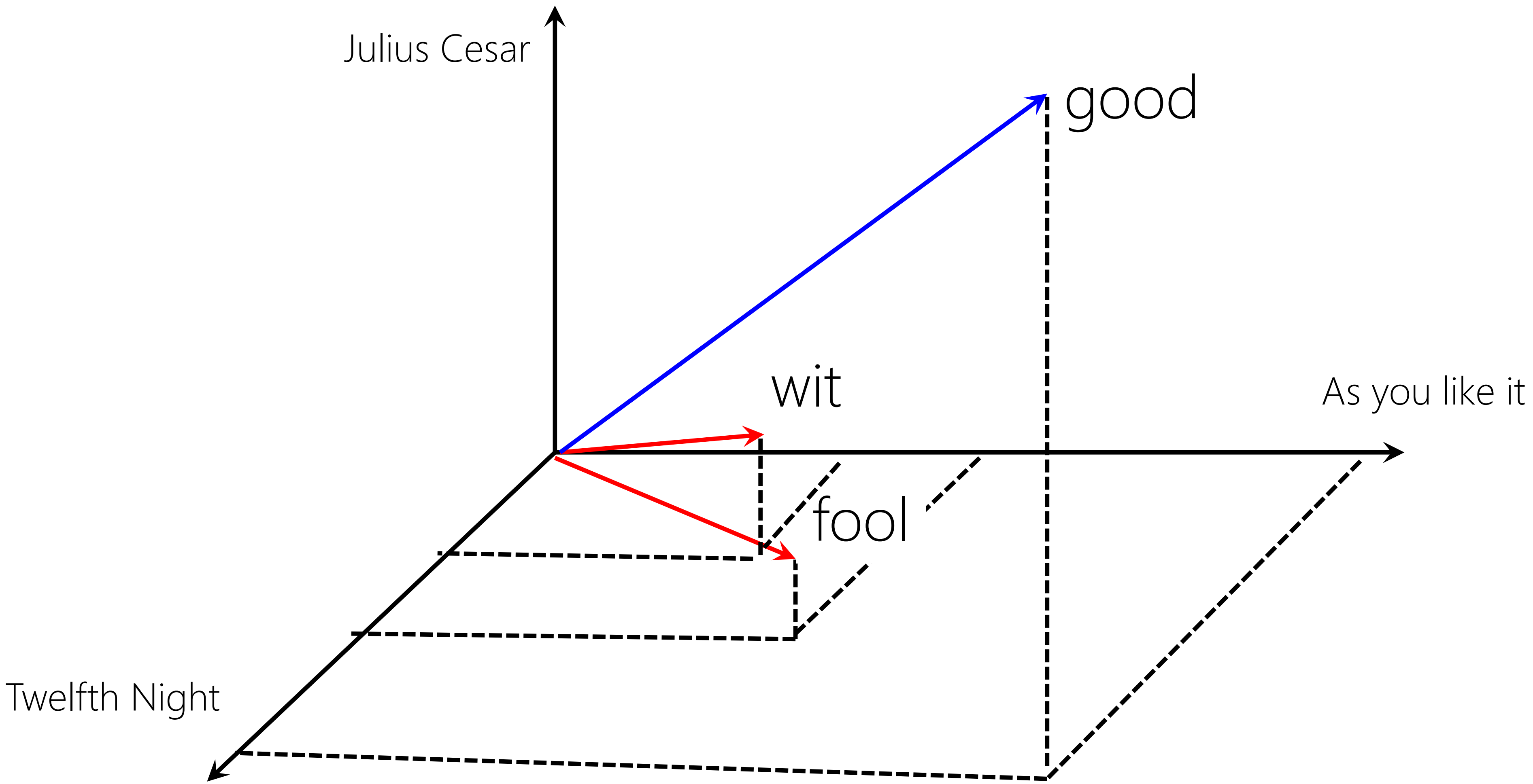
Term-Document Matrix

Words are points (vectors) in document space!



Term-Document Matrix

Distribution of words in documents!



Term-Document Matrix

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-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.

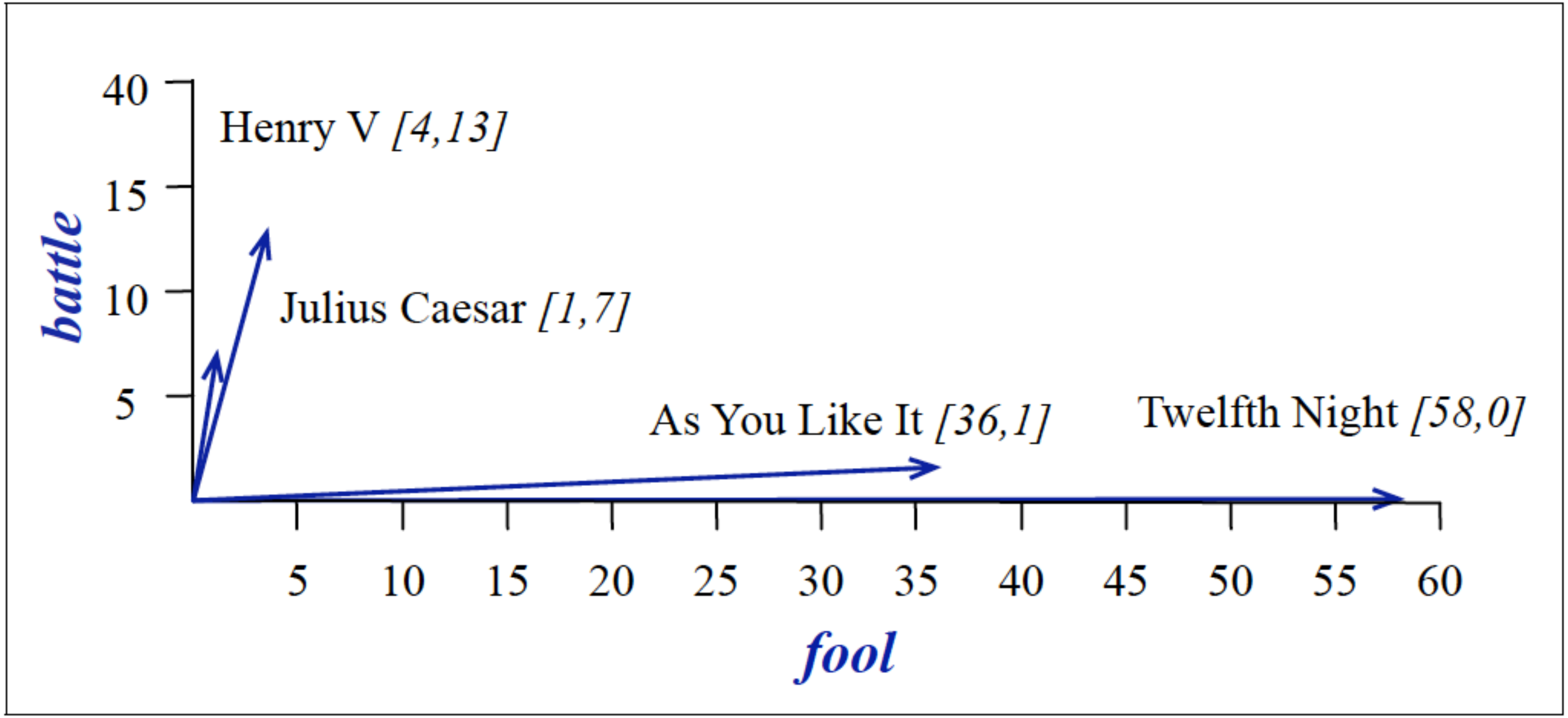


Figure 6.4 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Term-Term Matrix

aka word-word matrix
aka word-context matrix

Context Window =

- Fixed (Window): $\pm$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.

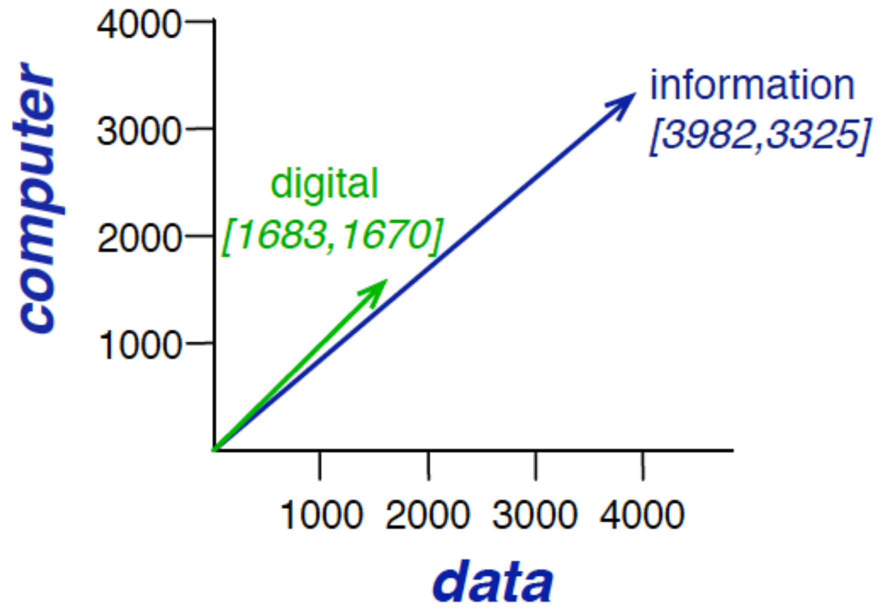


Figure 6.6 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *computer*.

Term-Term Matrix

Matrix Size = $|\text{Vocabs}|^2$

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

Term-Term Matrix

Matrix Size = $|\text{Vocabs}|^2$

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

TF-IDF
Weighted Term-Document Matrix

Context Window = Document

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

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.

Term-Term Matrix

As such, non-discriminative words co-occur with all terms: high term-term values with all other tokens!

iDF = inverse Document Frequency

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

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

TF-iDF

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

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

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

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 tf $= \log_{10}(20+1) = 1.322$ and 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*.

Semantic Similarity = Vector Similarity
Semantic Dissimilarity = Vector Distance

Cosine Similarity
Minkowski Distance

Cosine Similarity

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

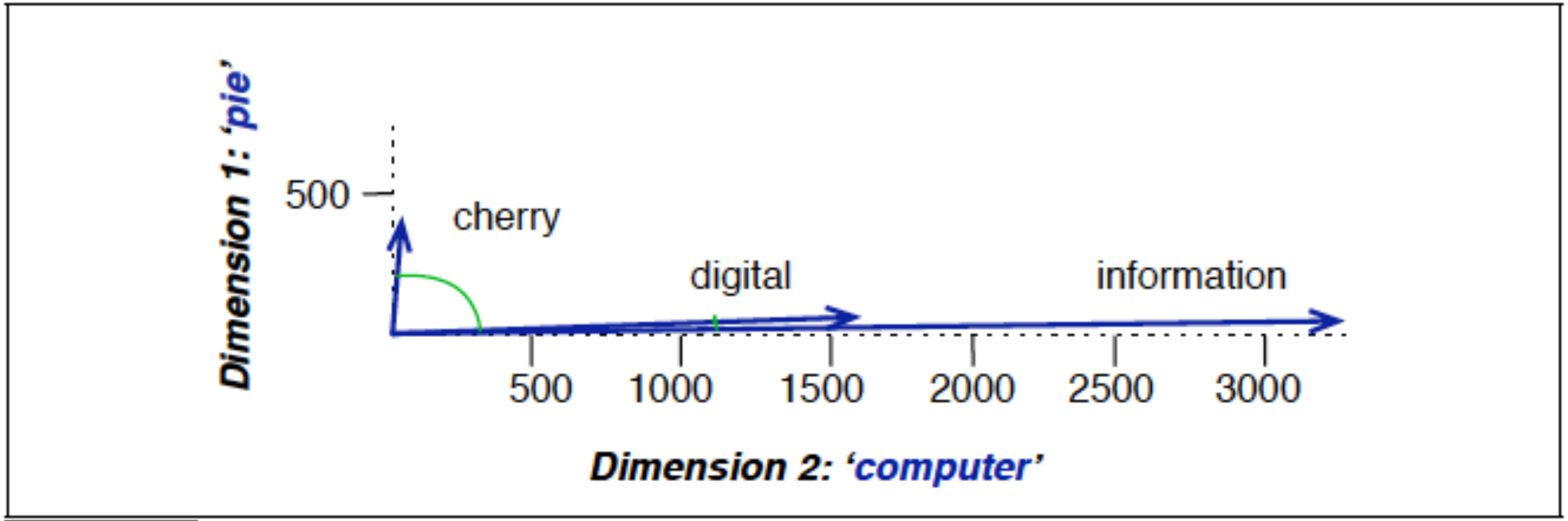
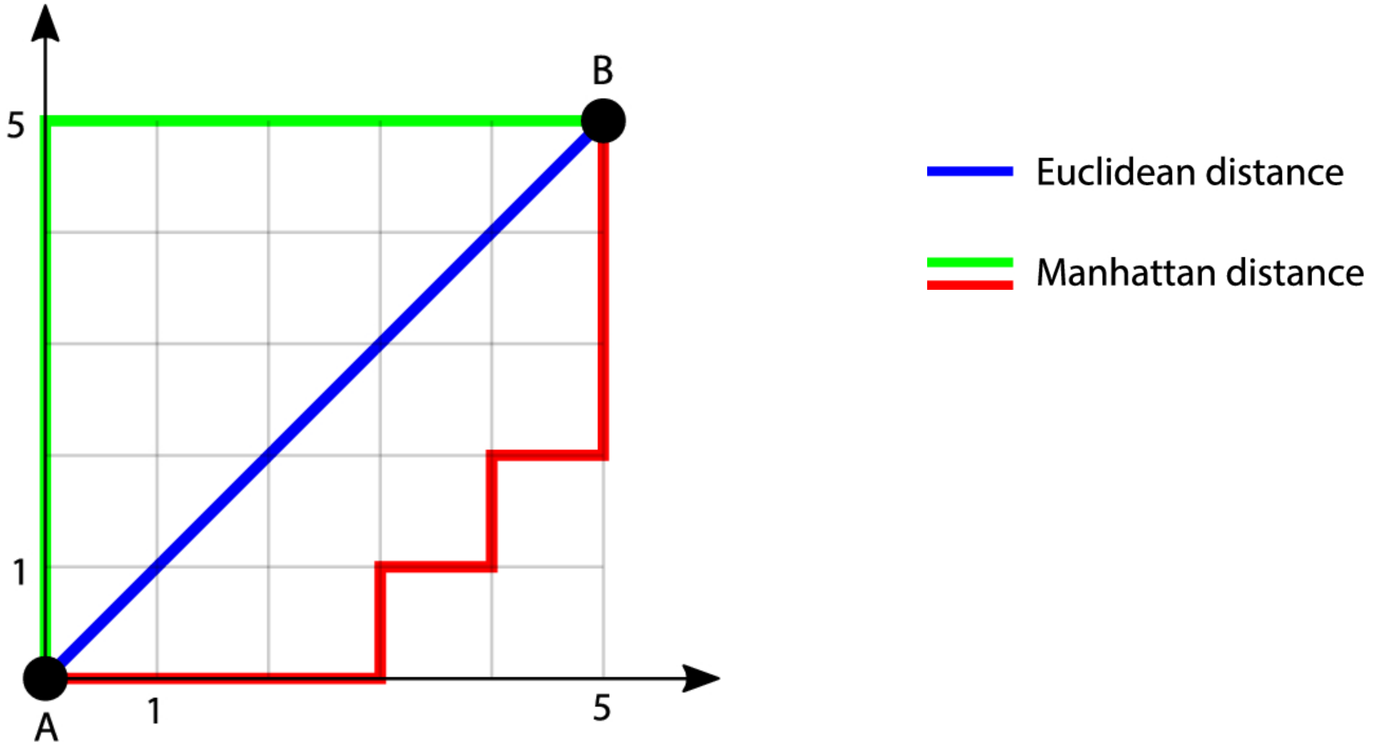


Figure 6.7 A (rough) graphical demonstration of cosine similarity, showing vectors for three words (*cherry*, *digital*, and *information*) in the two dimensional space defined by counts of the words *computer* and *pie* nearby. Note that the angle between *digital* and *information* is smaller than 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



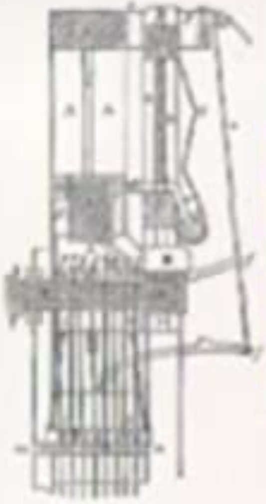
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

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

HTTPS://FB.WATCH/3GAMOB06A8/

Linear Algebra
Transformation between Spaces

How to learn representation for sentence/paragraph/documents?



SPEECH and LANGUAGE PROCESSING

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

DANIEL JURAFSKY & JAMES H. MARTIN