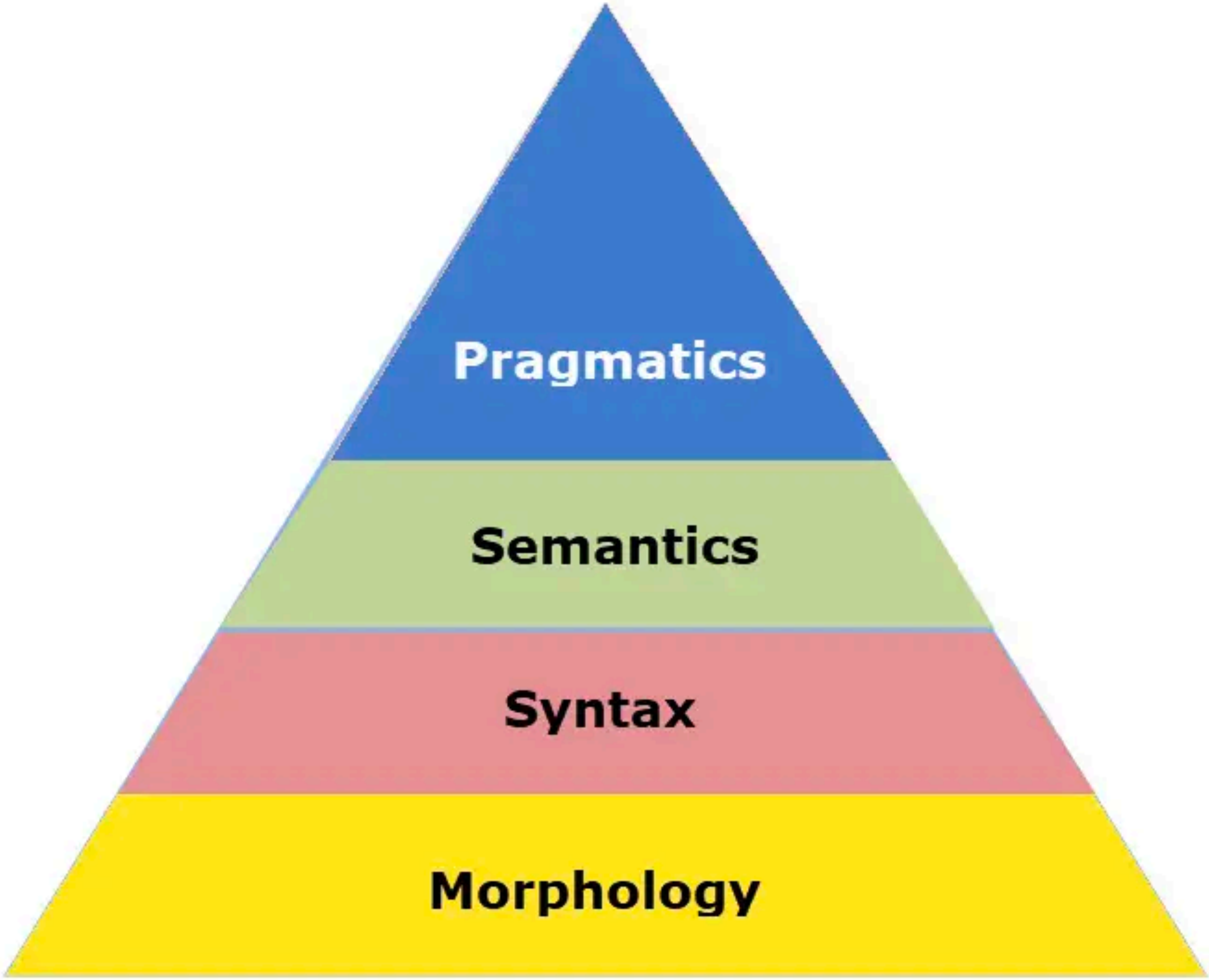
Lecture 11
Embedding: Representations of the meaning of words

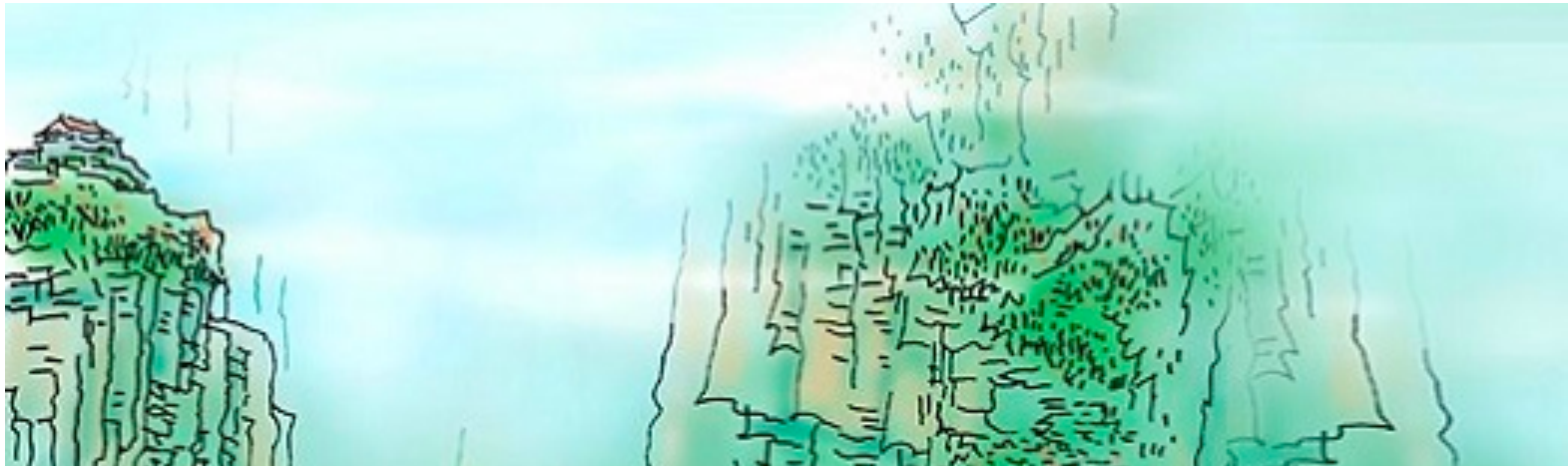
Zhizheng Wu

Agenda

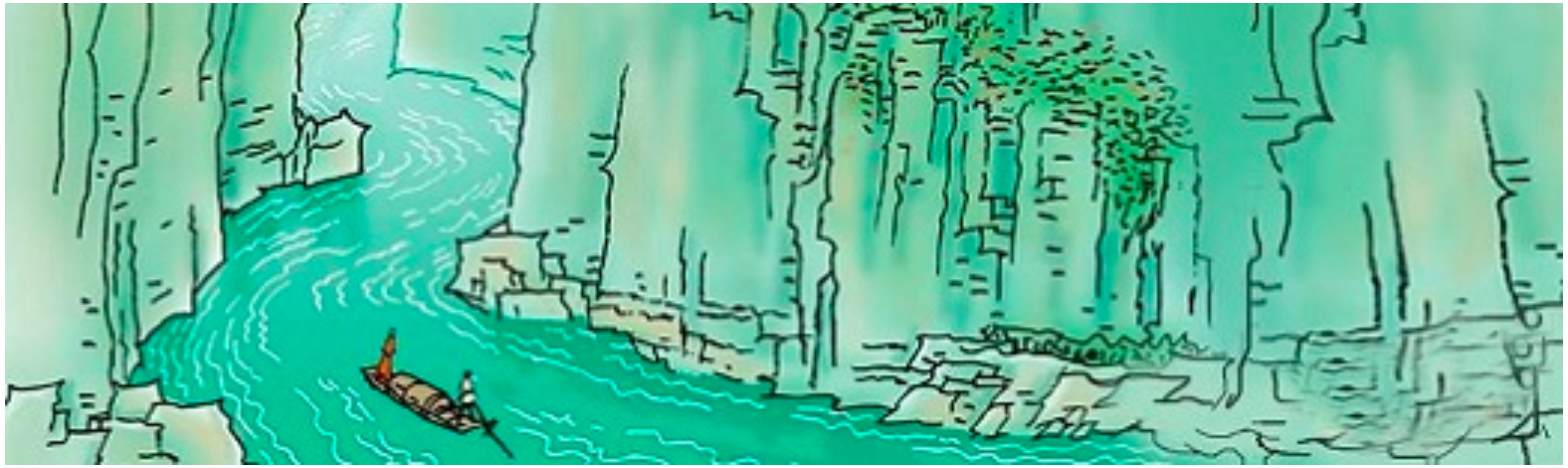
- Recap
- Word sense and their relations
- Word representation and embedding
- Measuring semantic similarity



Natural Language Processing Pyramid



朝辞白帝彩云间，千里江陵一日还。
两岸猿声啼不住，轻舟已过万重山。



言者所以在意，得意而忘言

Words are for meaning; Once you get the meaning, you can forget the words

Bank

Bank

Bank

Word sense (词义)

- Word sense vs Lemma

Lemma

Play (N)

Sense

- a theatrical performance of a drama, "the play lasted two hours"
- a preset plan of action in team sports, "the coach drew up the plays for her team"
- a state in which action is feasible, "the ball was still in play"; "insiders said the company's stock was in play"
- utilization or exercise, "the play of the imagination"

Word sense (concept)

- He wrote several **plays** but only one was produced on Broadway
- Insiders said the company's stock was in **play**
- The runner was out on a **play** by the shortstop

Recommended podcast on play (玩儿）: https://etw.fm/2036

Relations between senses: Synonymy(同义词)

- Synonyms have the same meaning in some or all contexts
  - couch/sofa
  - large/big
  - water/H_2O

Relations between senses: Similarity

- Words with similar meanings
- Not synonyms, but sharing some element of meaning
  - Car, bicycle
  - Cow, horse

Relations between senses: Relatedness

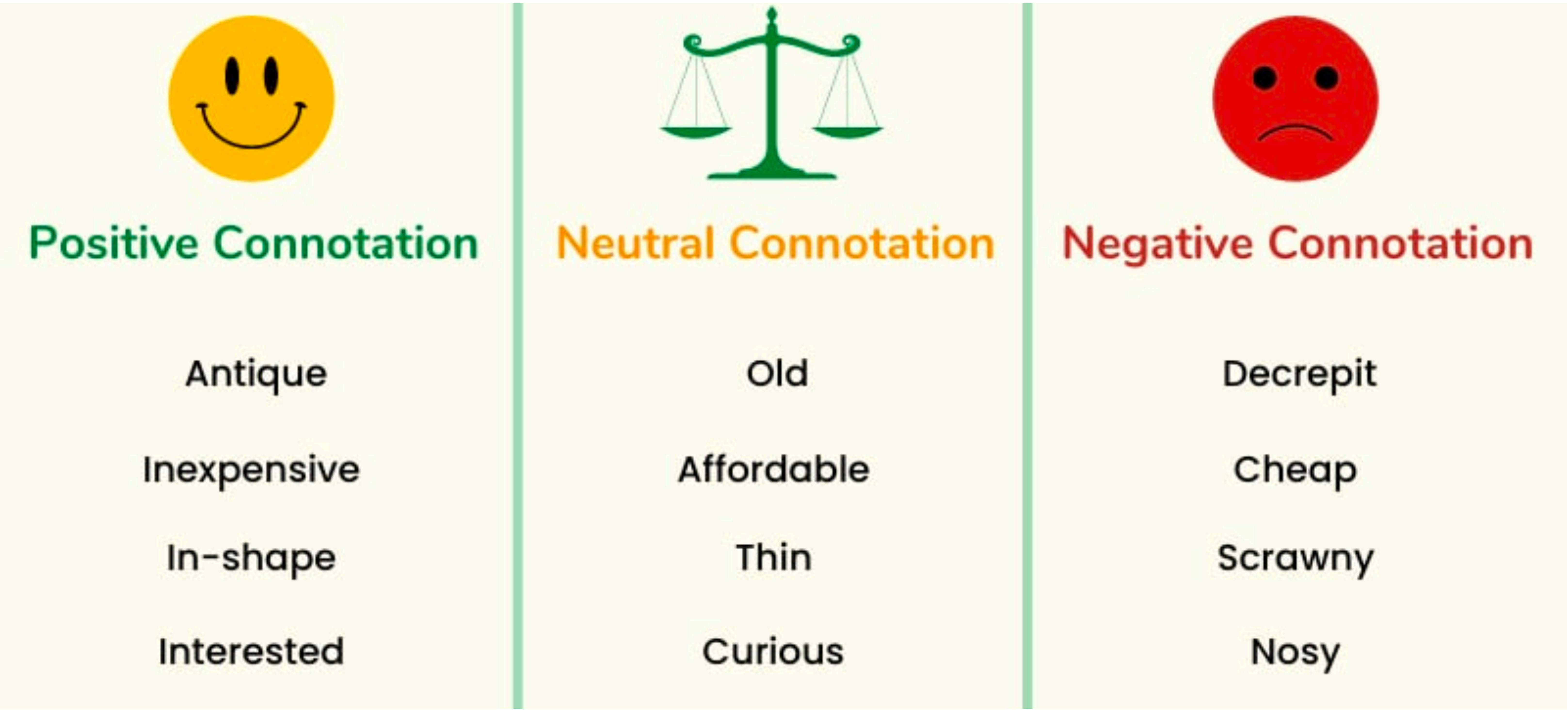
- Also named as *word association*
- Words can be related in any way, perhaps via a semantic frame or field
    - Similar: coffee, tea
    - Related (but not similar)
        - coffee, cup

Relations between senses: Antonymy（反义词）

- Senses that are opposites with respect to only one feature of meaning
  - Examples
    - Short/long
    - Hot/cold
    - In/out

Relations between senses: Connotation（含义）

- Affective meaning of words
  - fake, knockoff, forgery
  - copy, replica, reproduction



Evolution of word sense

汤：Soup

汤（湯），热水也

《说文解字》

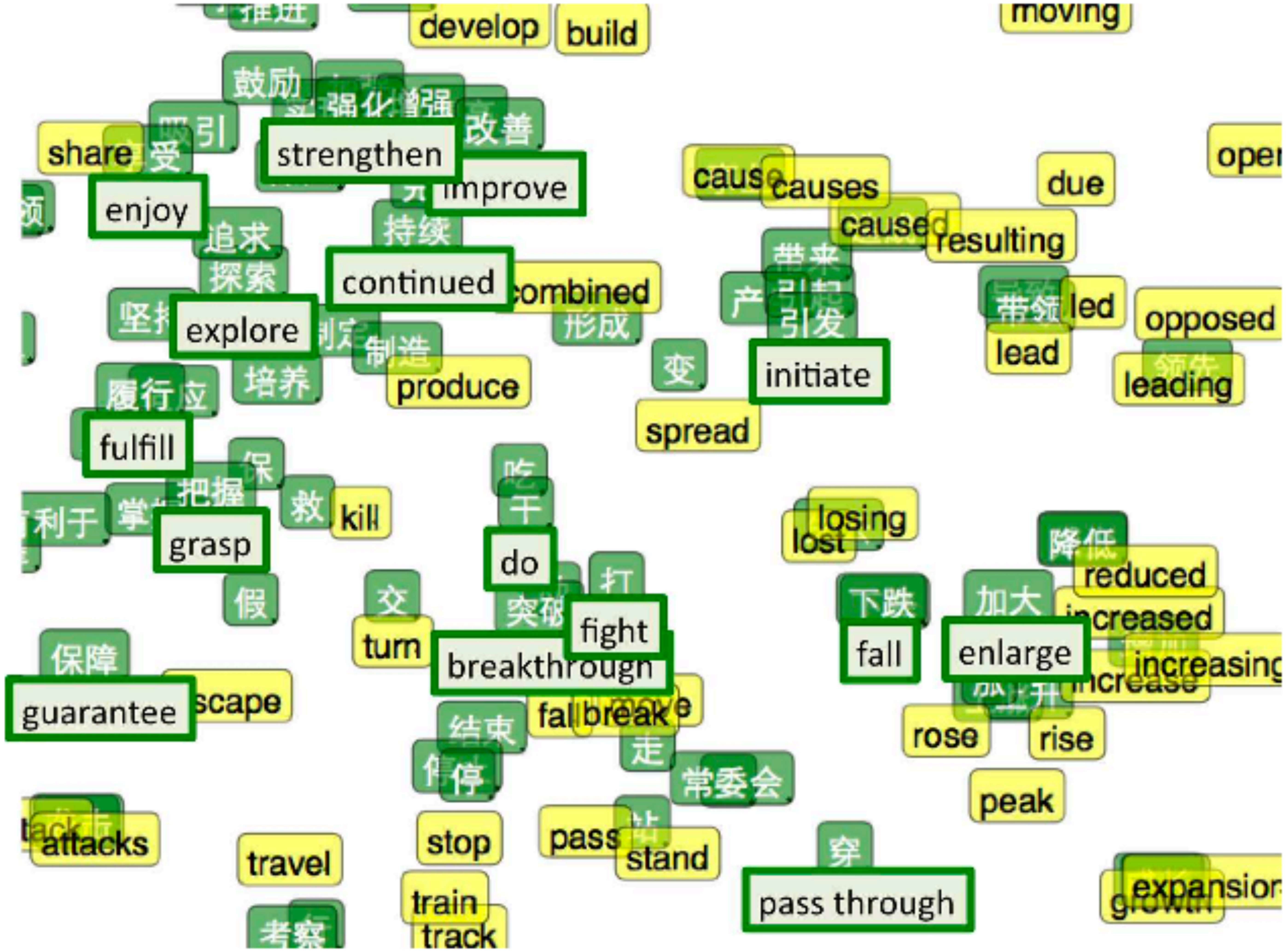
Word representation

- Five words vocabulary: man, walk, wowan, swim, ask
  - 1-of-N encoding/one-hot encoding
    - [1, 0, 0, 0, 0]: man
    - [0, 1, 0, 0, 0]: walk
    - [0, 0, 1, 0, 0]: woman
    - [0, 0, 0, 1, 0]: swim
    - [0, 0, 0, 0, 1]: ask

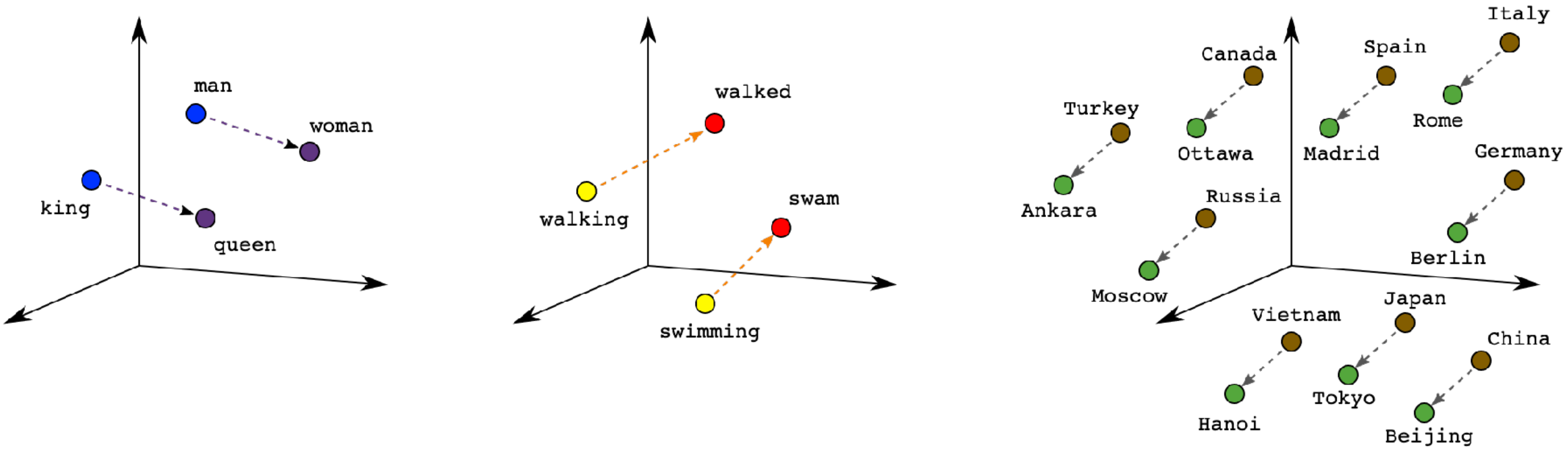
Cross lingual

- Banana
- 香蕉
- バナナ
- 바나나
- plátano
- quả chuối

Cross-lingual word embedding



Semantic similarity



Male-Female   Verb Tense   Country-Capital

Embedding representations

Dense Matrix

1	2	31	2	9	7	34	22	11	5
11	92	4	3	2	2	3	3	2	1
3	9	13	8	21	17	4	2	1	4
8	32	1	2	34	18	7	78	10	7
9	22	3	9	8	71	12	22	17	3
13	21	21	9	2	47	1	81	21	9
21	12	53	12	91	24	81	8	91	2
61	8	33	82	19	87	16	3	1	55
54	4	78	24	18	11	4	2	99	5
13	22	32	42	9	15	9	22	1	21

Sparse Matrix

1	.	3	.	9	.	3	.	.	.
11	.	4	.	.	.	.	.	2	1
.	.	1	.	.	.	4	.	1	.
8	.	.	.	3	1	.	.	.	.
.	.	.	9	.	.	1	.	17	.
13	21	.	9	2	47	1	81	21	9
.	.	.	.	.	.	.	.	.	.
.	.	.	.	19	8	16	.	.	55
54	4	.	.	.	11	.	.	.	.
.	.	2	.	.	.	.	22	.	21

Co-occurrence matrix

- term-document matrix
  - each row represents a word in the vocabulary
  - each column represents a document from some collection of documents

- Term-term matrix
  - the columns are labeled by words rather than documents

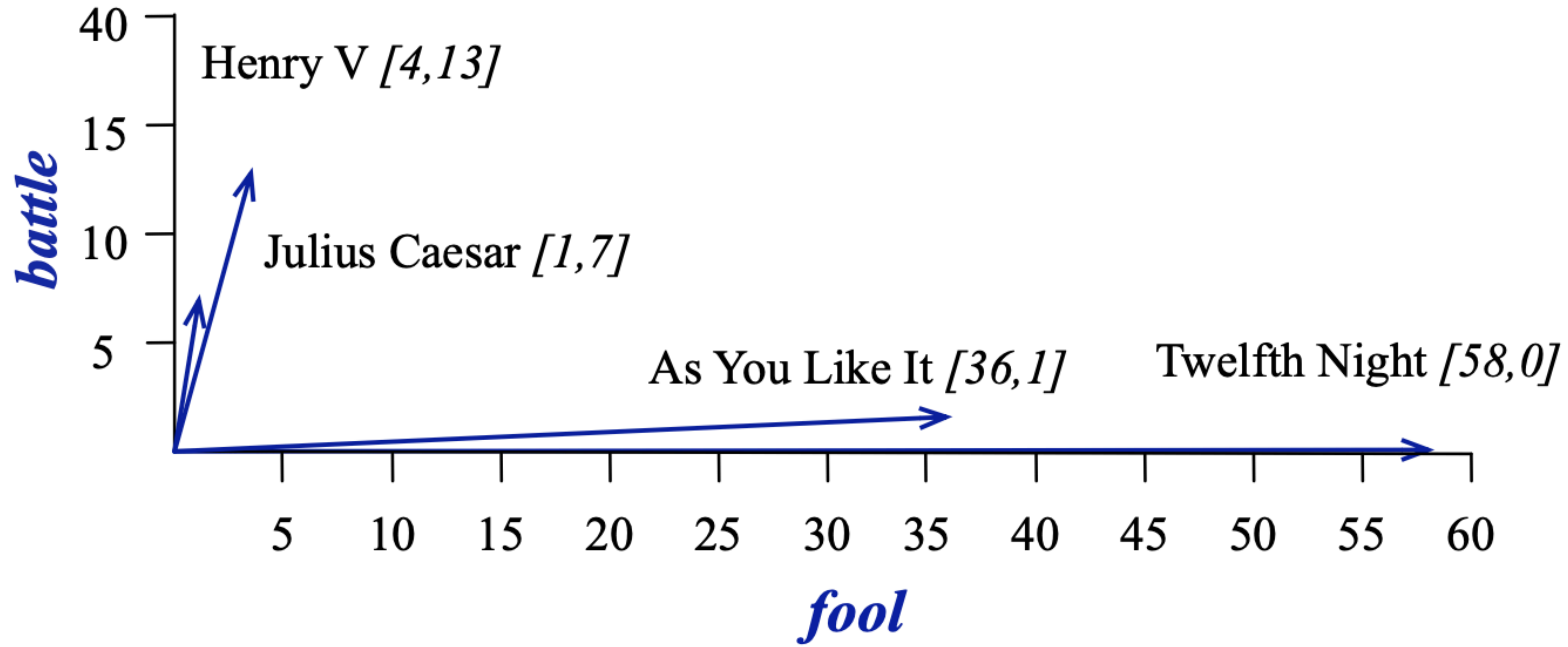
Term-document matrix

- originally defined as a means of finding similar documents

	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

similar documents had similar vectors

Spatial visualization



Words as vectors: Document dimensions

	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

similar words have similar vectors
because they tend to occur in similar documents

Term-term matrix

- the columns are labeled by words rather than documents

- Two words are similar in meaning if their context vectors are similar

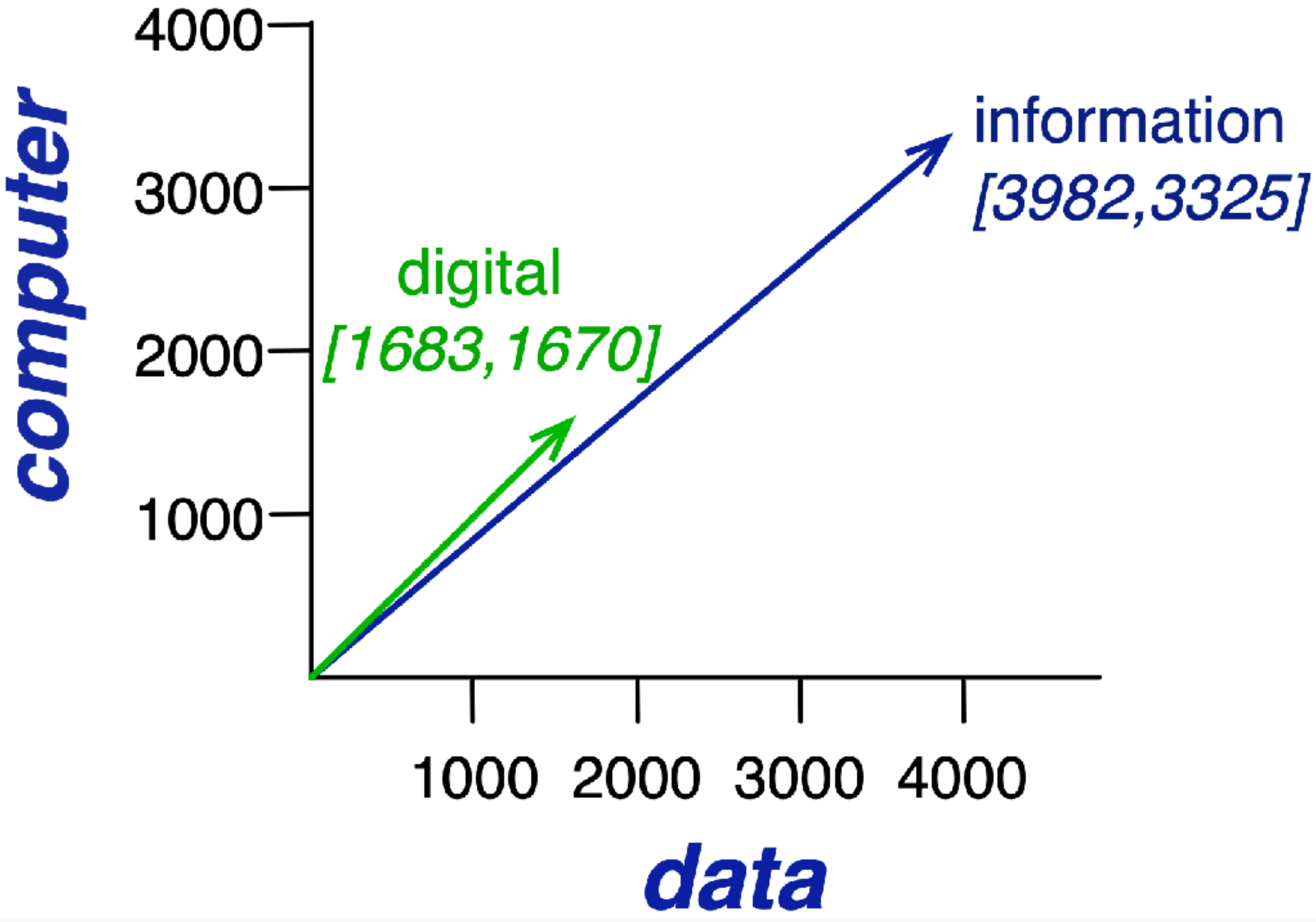
is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

Words as vectors: Word dimensions

- word-word co-occurrence matrix

	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	...

Spatial visualization



Is the raw frequency a good representation?

- Motivation
  - Frequency is clearly useful
  - However, overly frequent words like

    the and *it*

    are not very informative about the context

We need to balance

TF-IDF

- Term frequency

$$\mathrm{tf}_{t,d} = \mathrm{count}(t,d)$$

Instead of using raw count, we squash a bit:

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

TF-IDF

- Document frequency
  - df is a term t is the number of documents it occurs in

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

TF-IDF

- Inverse document frequency

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

N is the total number of documents

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

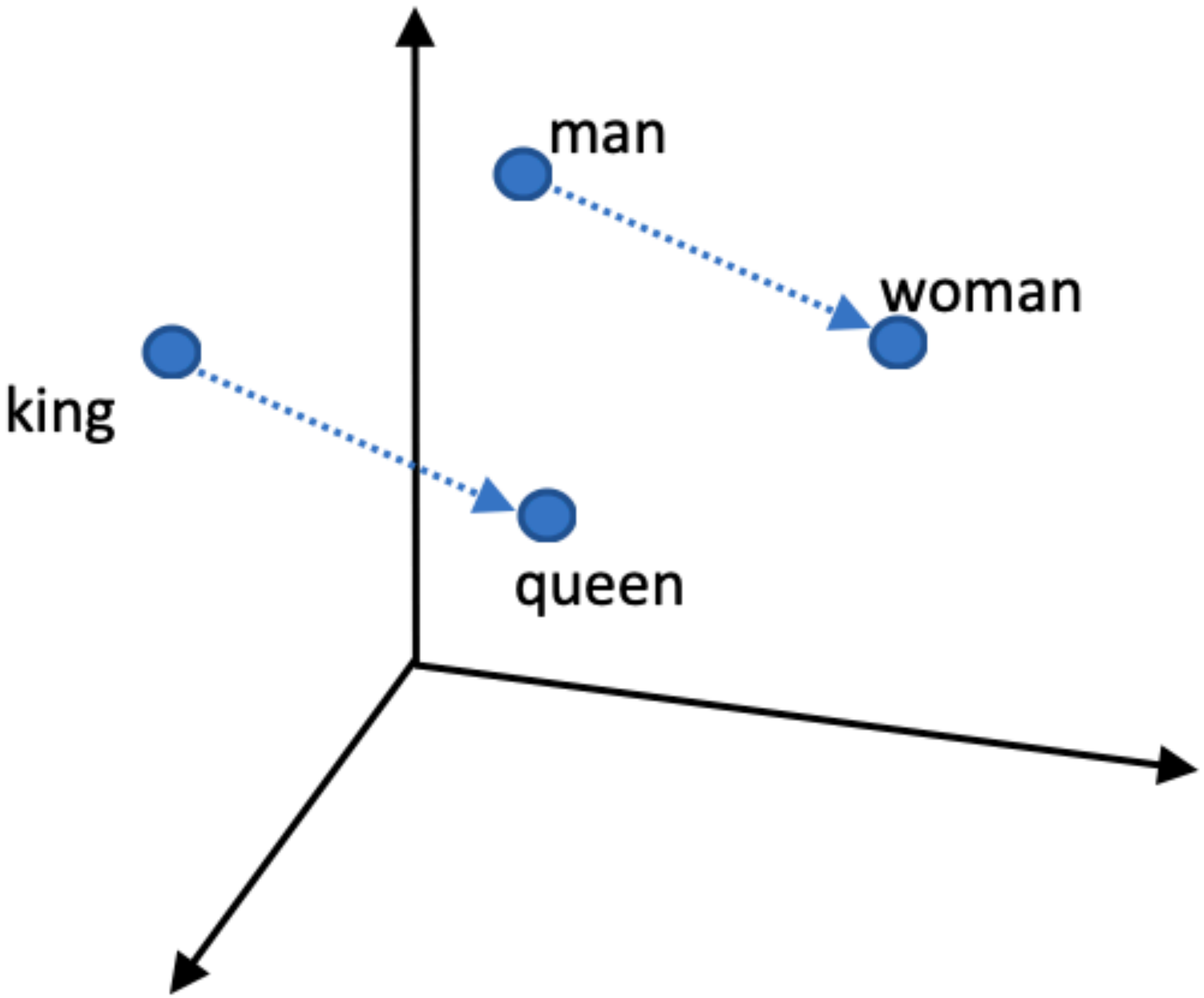
TF-IDF

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

	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

	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

Semantic similarity measurement



Inner/dot product

- The dot product between two vectors is a scalar

$$\text{dot product}(\mathbf{v}, \mathbf{w}) = \mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

The dot product tends to be **high**
when the two vectors have **large** values in the **same** dimensions

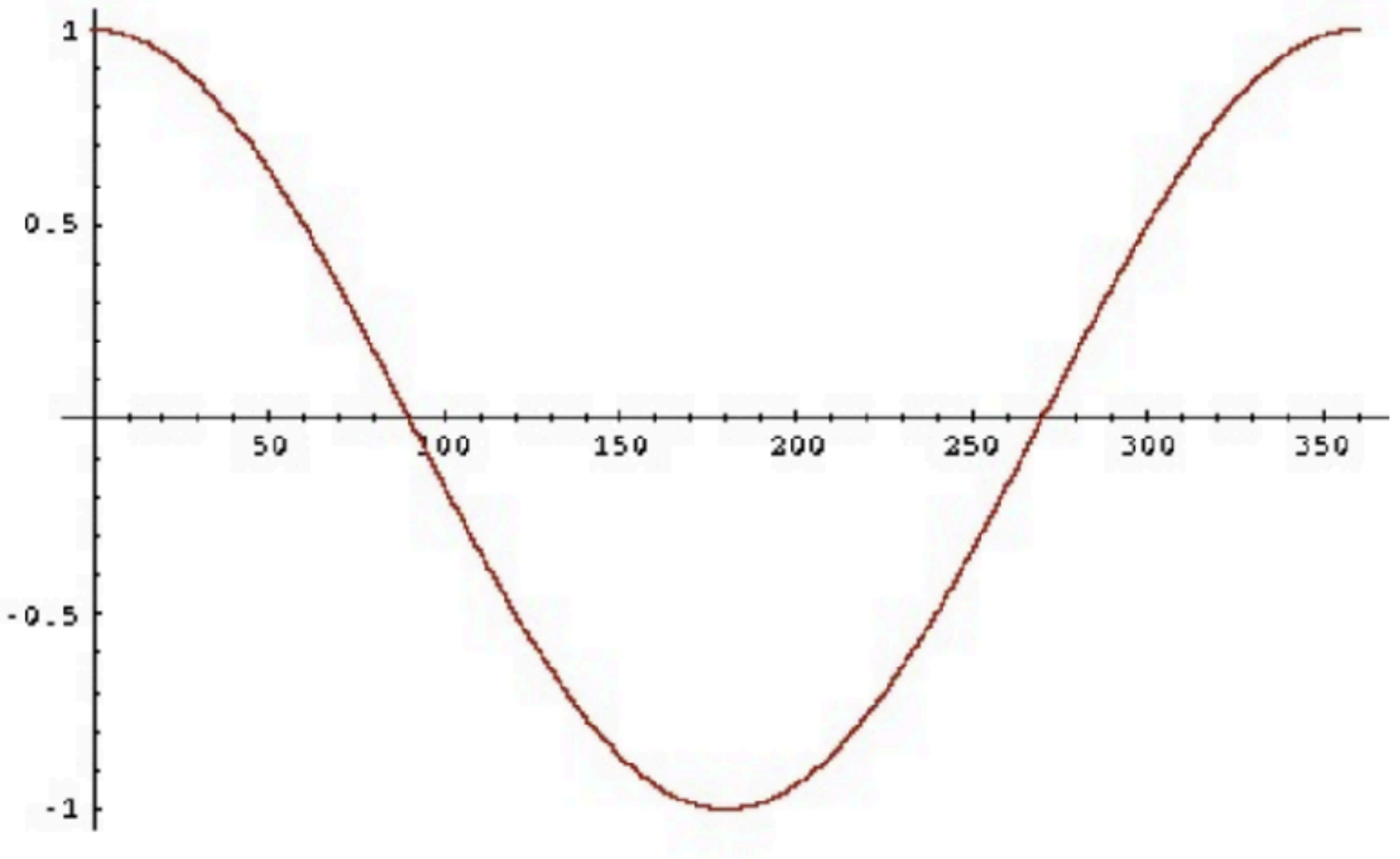
Dot-product: problem

- Dot-product favors long vectors (i.e. vectors with larger norm)

$$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

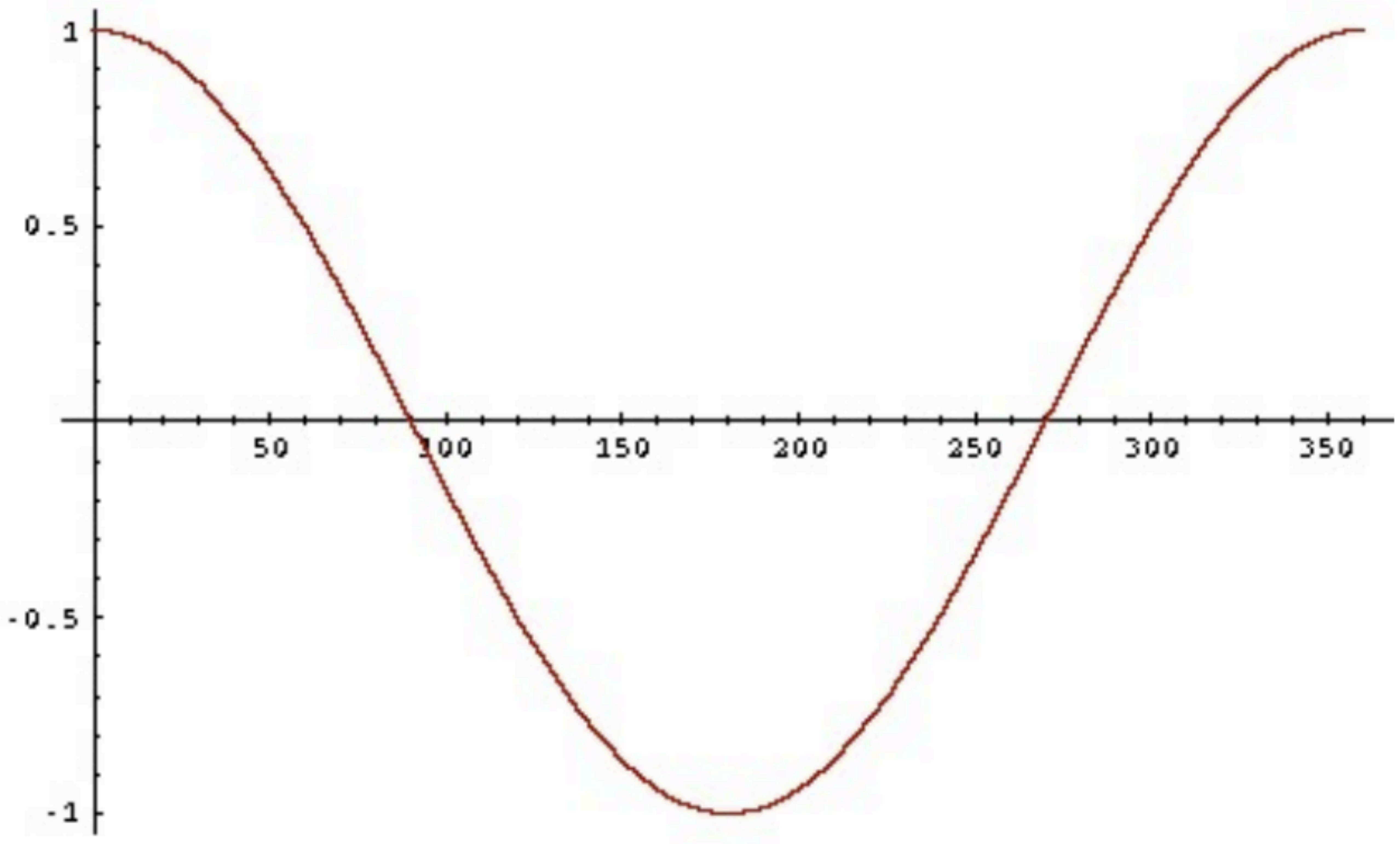
Cosine similarity

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



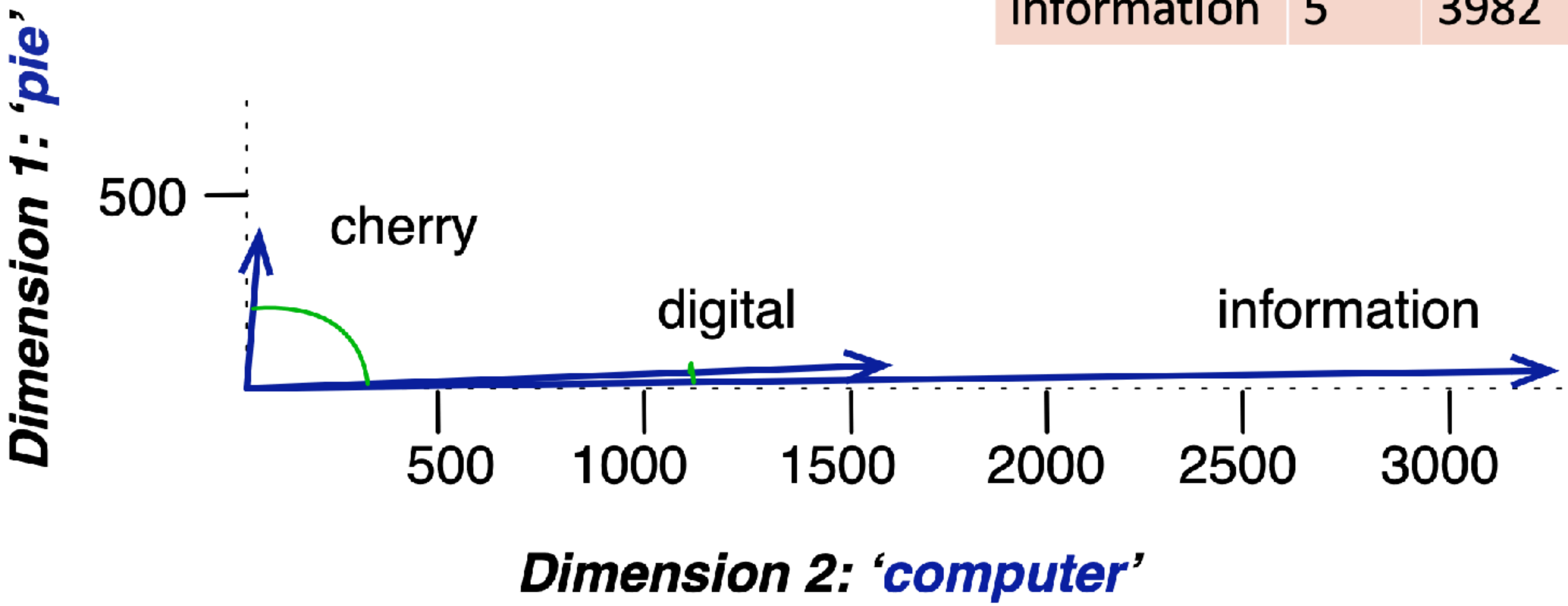
Cosine similarity: Interpretation

- -1: opposite directions
- +1: same direction
- 0: orthogonal



Cosine similarity

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325



Summary

- Word sense and their relations
- Word representation
  - We focus on sparse representation in today's lecture
  - Term-document matrix
  - Term-term matrix
  - TF-IDF
- Measure semantic similarity

Reading and tools

- Word embedding colab
    - https://colab.research.google.com/github/pytorch/tutorials/blob/gh-pages/_downloads/363afc3b7c522e4e56981679c22f1ad6/word_embeddings_tutorial.ipynb
    - https://colab.research.google.com/github/tensorflow/text/blob/master/docs/guide/word_embeddings.ipynb

- Chapter 6: Vector Semantics and Embeddings
    - https://web.stanford.edu/~jurafsky/slp3/6.pdf