COMP90042
Web search and text analysis

Workshop Week 6

xudong.han@unimelb.edu.au
https://github.com/HanXudong/COMP90042_Workshops

Review

1. N-gram model
2. Backoff and interpolation

1. <s> <s> how much wood would a wood chuck chuck if a wood chuck would chuck wood </s>

2. <s> <s> a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood </s>

<s> <s> a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood </s></s>

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \,|\, w_{i-2}\ w_{i-1})$$

$$P_{add1}(w_i \,|\, w_{i-2}\ w_{i-1}) = \frac{C(w_{i-2}\ w_{i-1}\ w_i) + 1}{C(w_{i-2}\ w_{i-1}) + V}$$

Q3: What does back–off mean, in the context of smoothing a language model? What does interpolation refer to?

- The idea in a Backoff model is to build an Ngram model based on an (N-1) model

- https://en.wikipedia.org/wiki/Katz%27s_back-off_model

- Interpolation: instead of just backing off to the non-zero Ngram, it is possible to take into account all Ngrams.

- Estimate lambdas from held-out dataset.

This workshop

- Words and senses
- Wordnet and lexical semantics
- Distributional semantics and word embedding

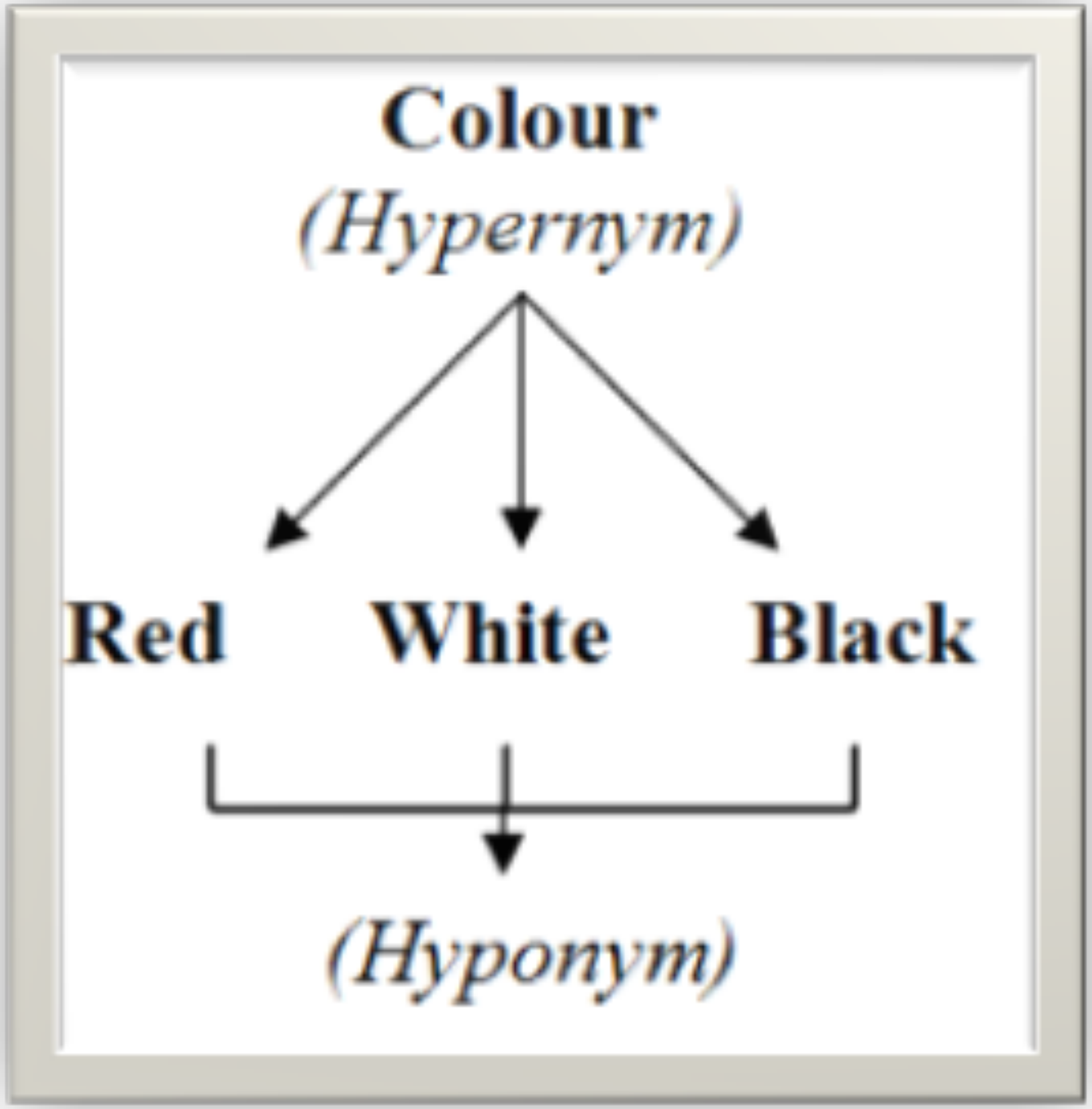
Senses

Synonyms

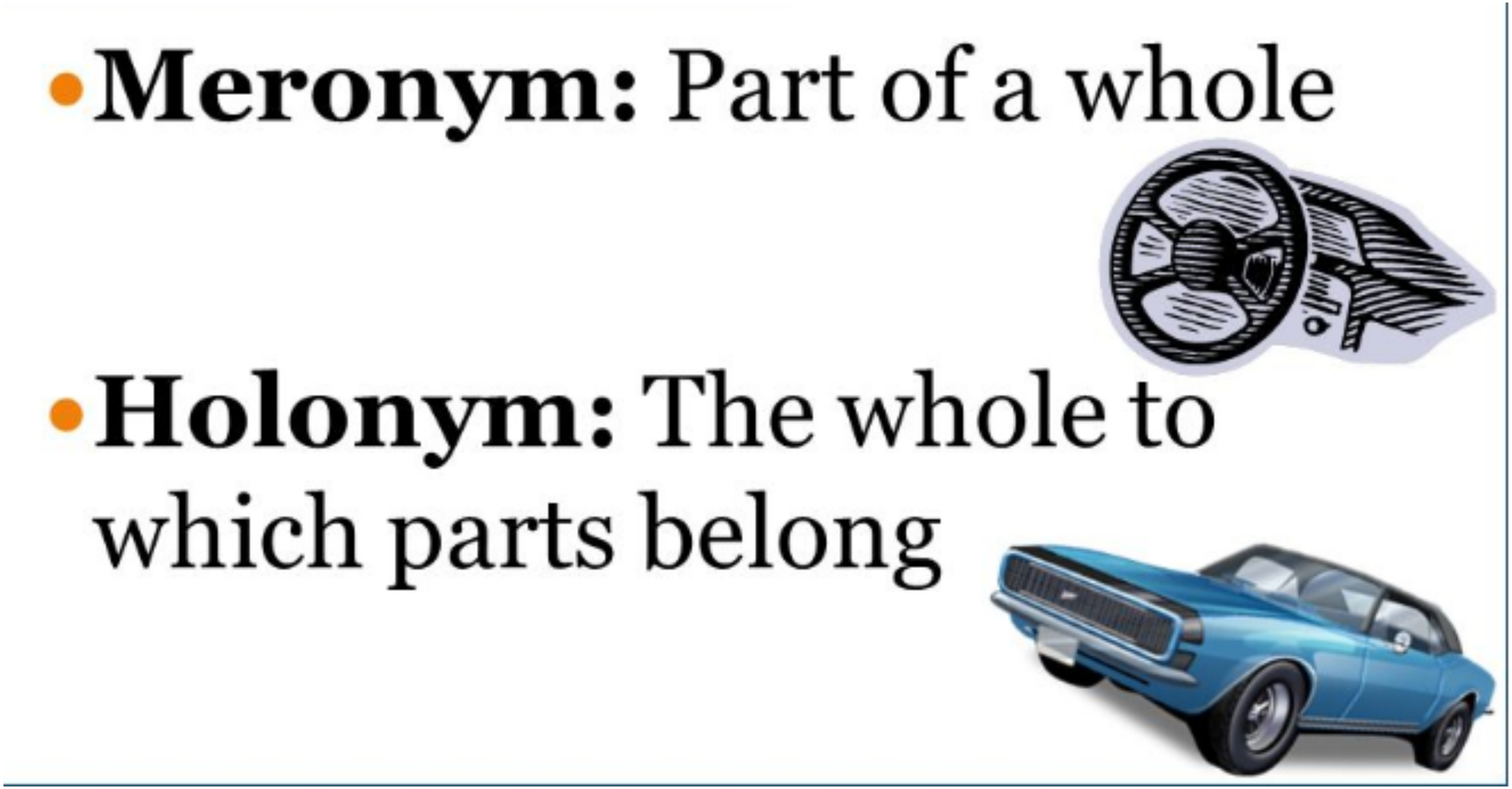
words that have the same meaning

gift
kids
present
children

Senses



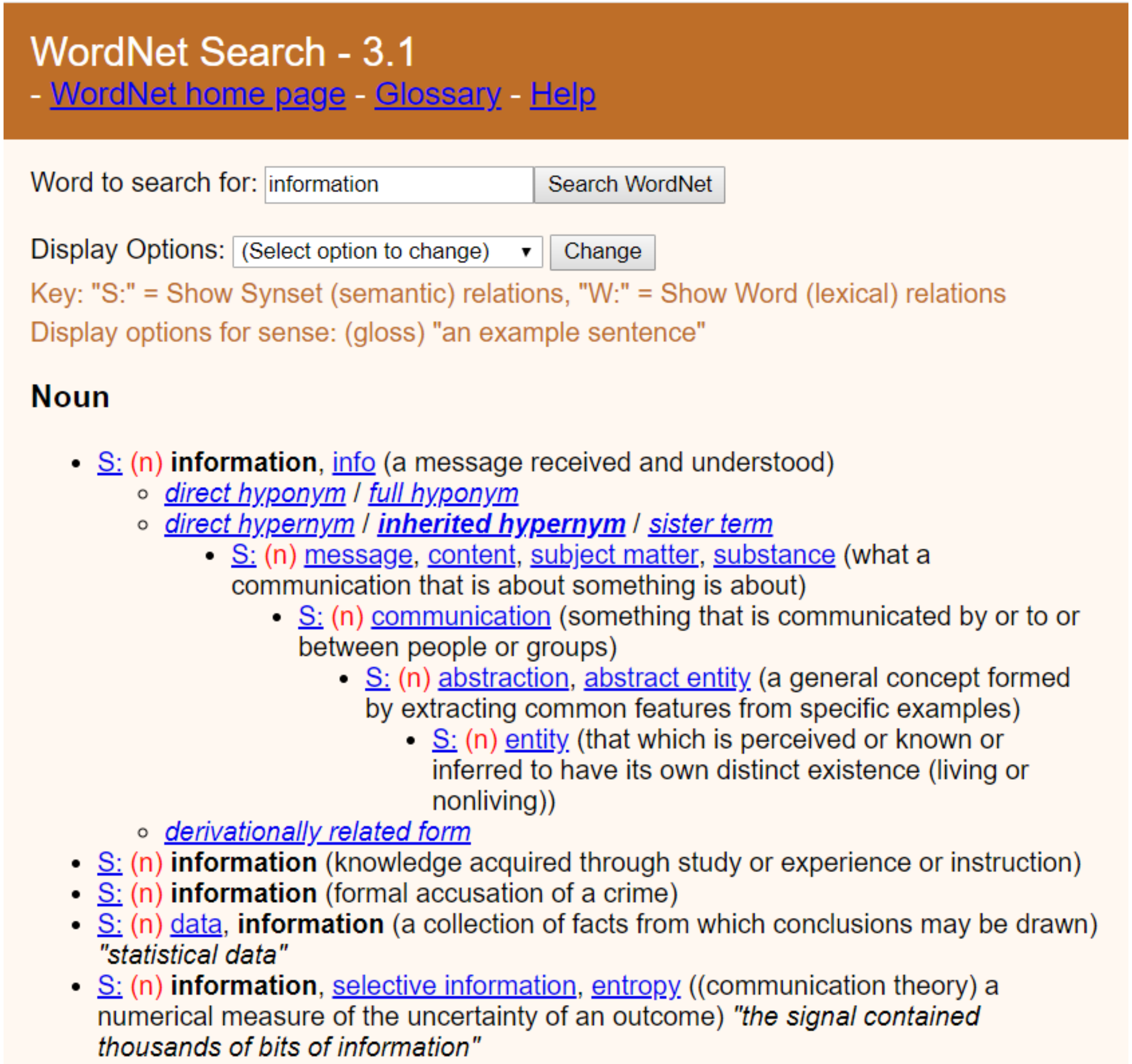
Senses



- **Meronym:** Part of a whole
- **Holonym:** The whole to which parts belong

Q2

WordNet



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: information

Display Options:

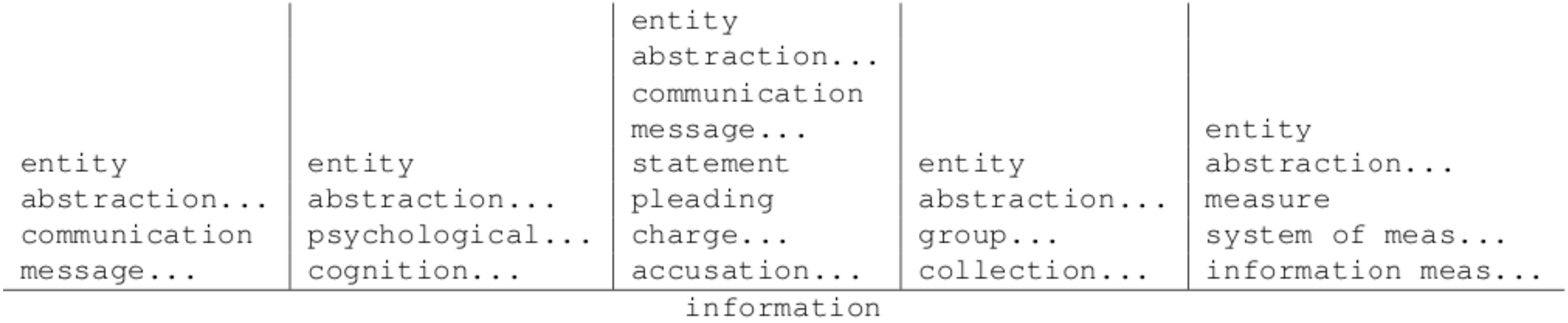
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

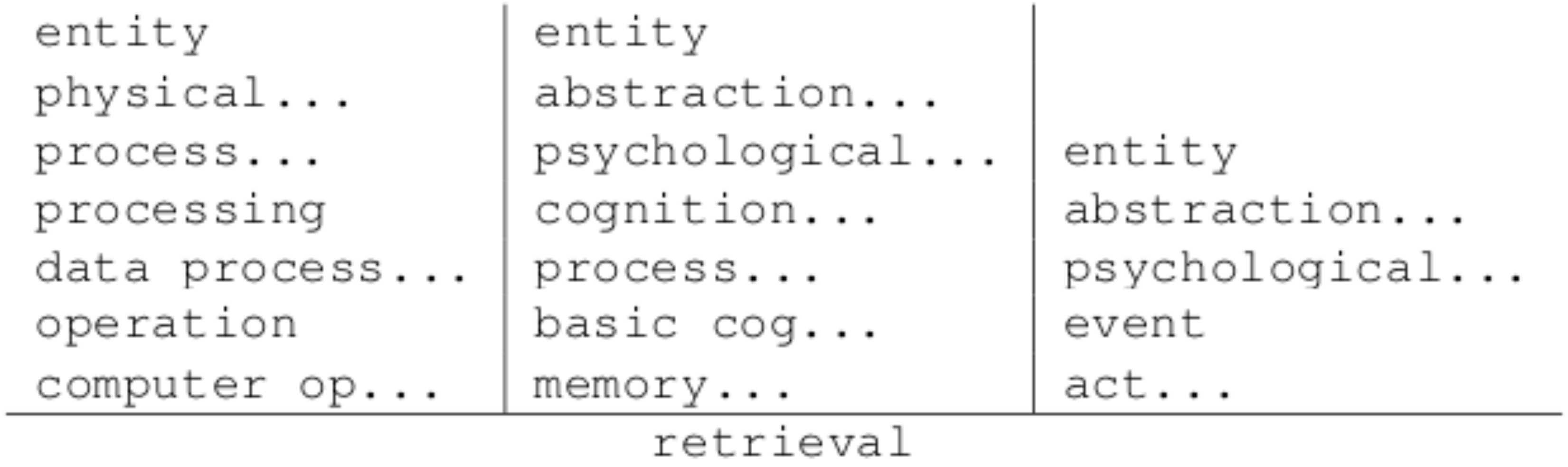
Noun

- S: (n) **information**, info (a message received and understood)
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) message, content, subject matter, substance (what a communication that is about something is about)
      - S: (n) communication (something that is communicated by or to or between people or groups)
        - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
          - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
  - *derivationally related form*
- S: (n) **information** (knowledge acquired through study or experience or instruction)
- S: (n) **information** (formal accusation of a crime)
- S: (n) data, **information** (a collection of facts from which conclusions may be drawn) *"statistical data"*
- S: (n) **information**, selective information, entropy ((communication theory) a numerical measure of the uncertainty of an outcome) *"the signal contained thousands of bits of information"*

https://github.com/evanmiltenburg/images/tree/master/wordnet%20graphs



```
                     |                     | entity              |                     |
                     |                     | abstraction...      |                     |
                     |                     | communication       |                     |
                     |                     | message...          |                     | entity
entity               | entity              | statement           | entity              | abstraction...
abstraction...       | abstraction...      | pleading            | abstraction...      | measure
communication        | psychological...    | charge...           | group...            | system of meas...
message...           | cognition...        | accusation...       | collection...       | information meas...
-----------------------------------------------------------------------------------------------------------
                                              information
```



```
entity                | entity                |
physical...           | abstraction...        |
process...            | psychological...      | entity
processing            | cognition...          | abstraction...
data process...       | process...            | psychological...
operation             | basic cog...          | event
computer op...        | memory...             | act...
-----------------------------------------------------------------
                          retrieval
```

information is more similar to the word *retrieval* or the word *science*

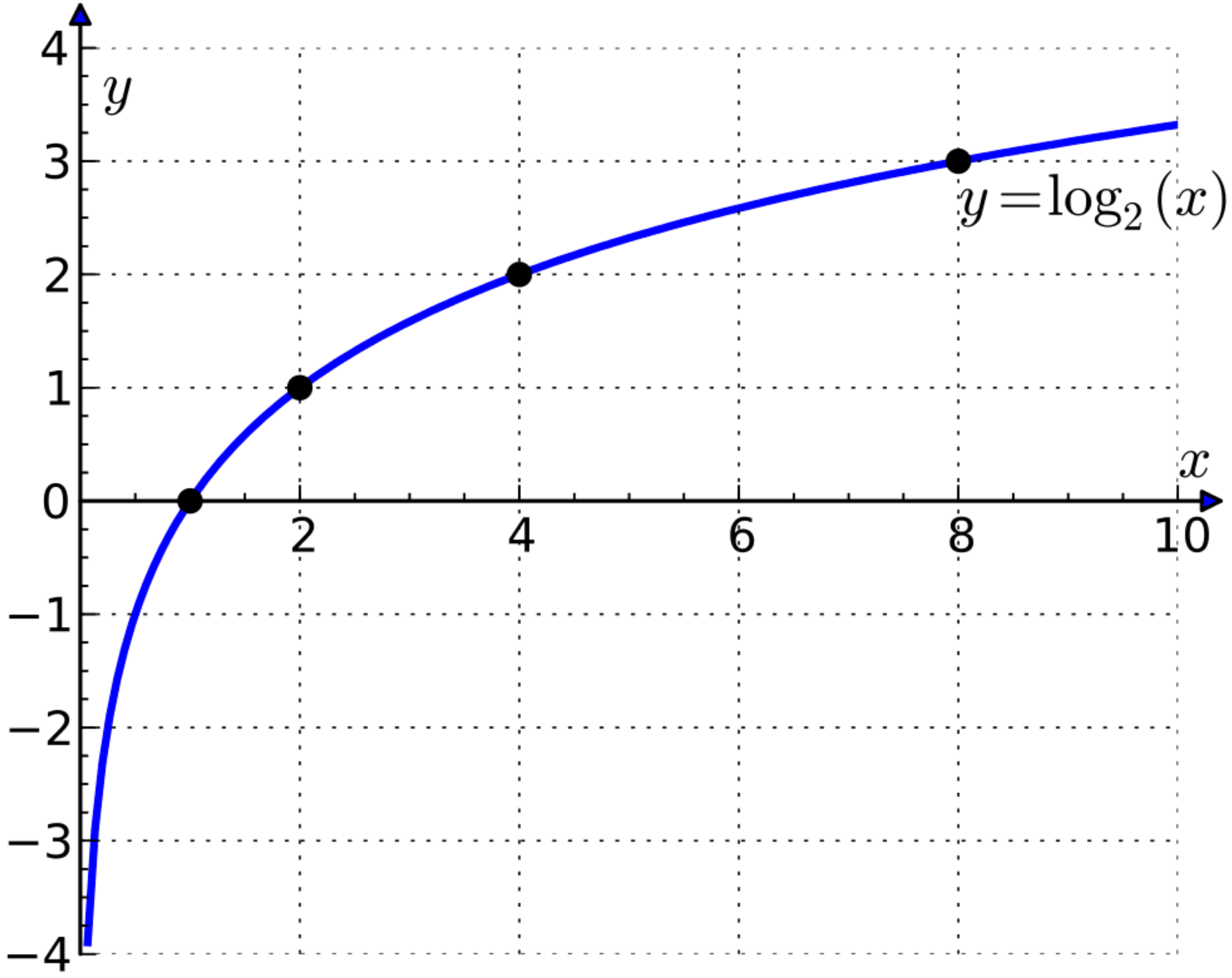
$$WuP\_sim(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

		information				
		1	2	3	4	5
retrieval	1	0.154	0.154	0.118	0.154	0.143
	2	0.308	0.615	0.235	0.308	0.286
	3	0.364	0.545	0.267	0.364	0.333

Q4a PMI

	cup	not (cup)	Total
world	55	225	**280**
not (world)	315	1405	**1720**
Total	**370**	**1630**	**2000**

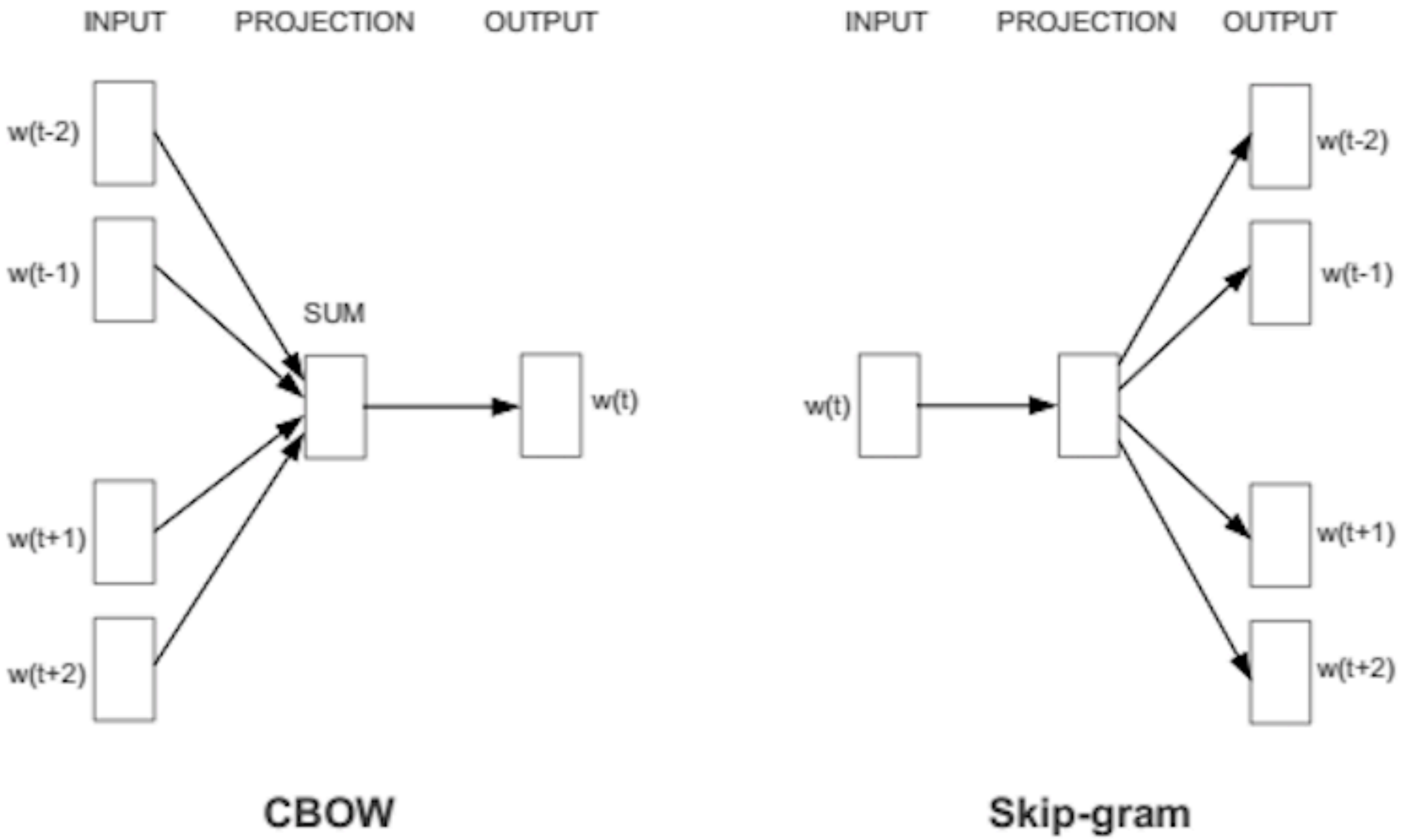
$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)} = log_2 P(x, y) - log_2 p(x) - log_2 P(y)$$

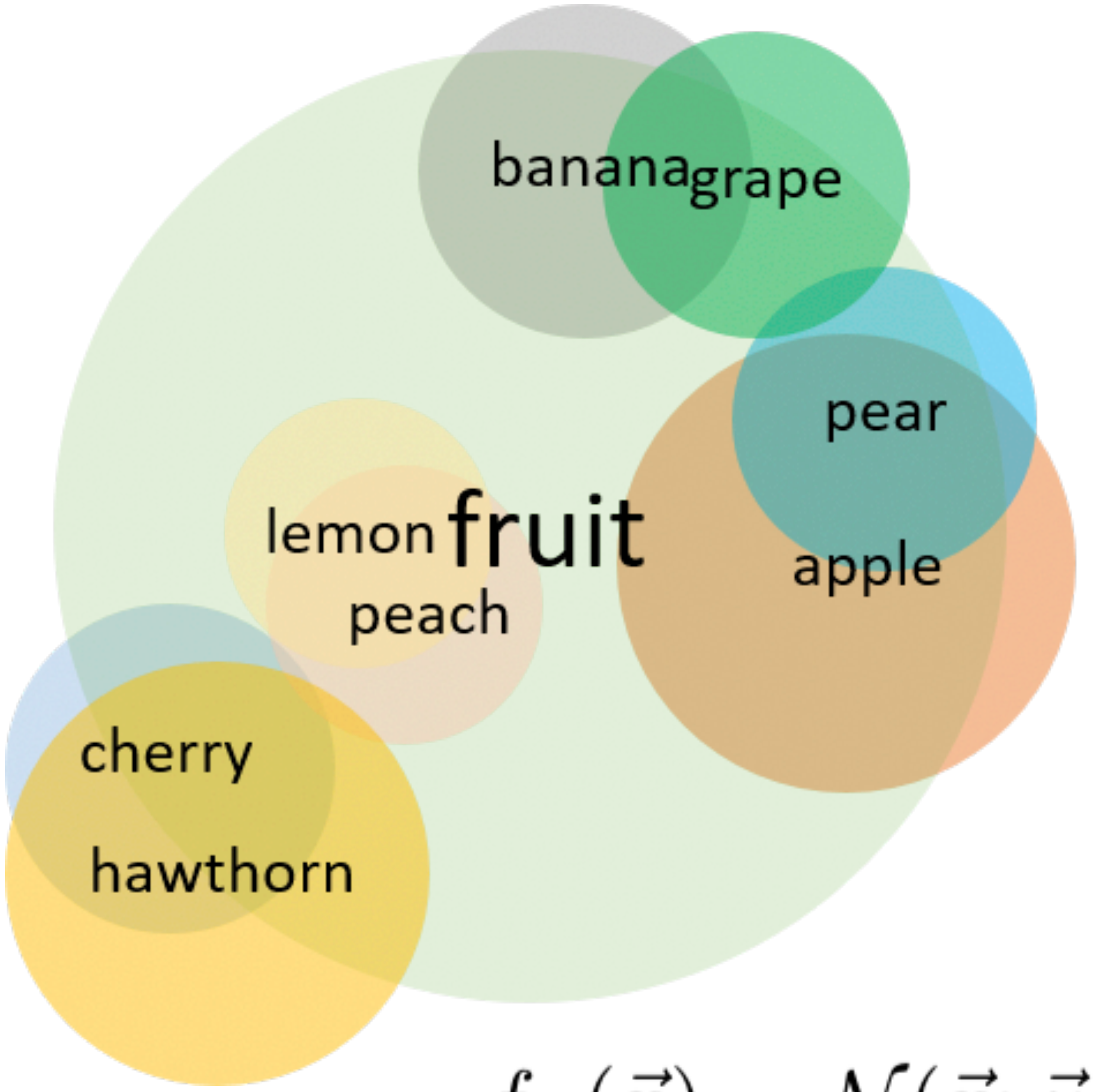


$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)} = log_2 \frac{p(y|x)}{p(y)} = log_2 \frac{p(x|y)}{x}$$

Q6 word to vector

https://skymind.ai/wiki/word2vec





Gaussian Word Embedding

$$f_w(\vec{x}) = \mathcal{N}(\vec{x}; \vec{\mu}_w, \Sigma_w)$$

$$= \frac{1}{\sqrt{(2\pi)^D |\Sigma_w|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu_w})^\top \Sigma_w^{-1} (\vec{x}-\vec{\mu_w})}$$