COMP90042
Web search and text analysis

Workshop Week 6

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

Review

<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>

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

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

This workshop

- Words and senses(Q1)

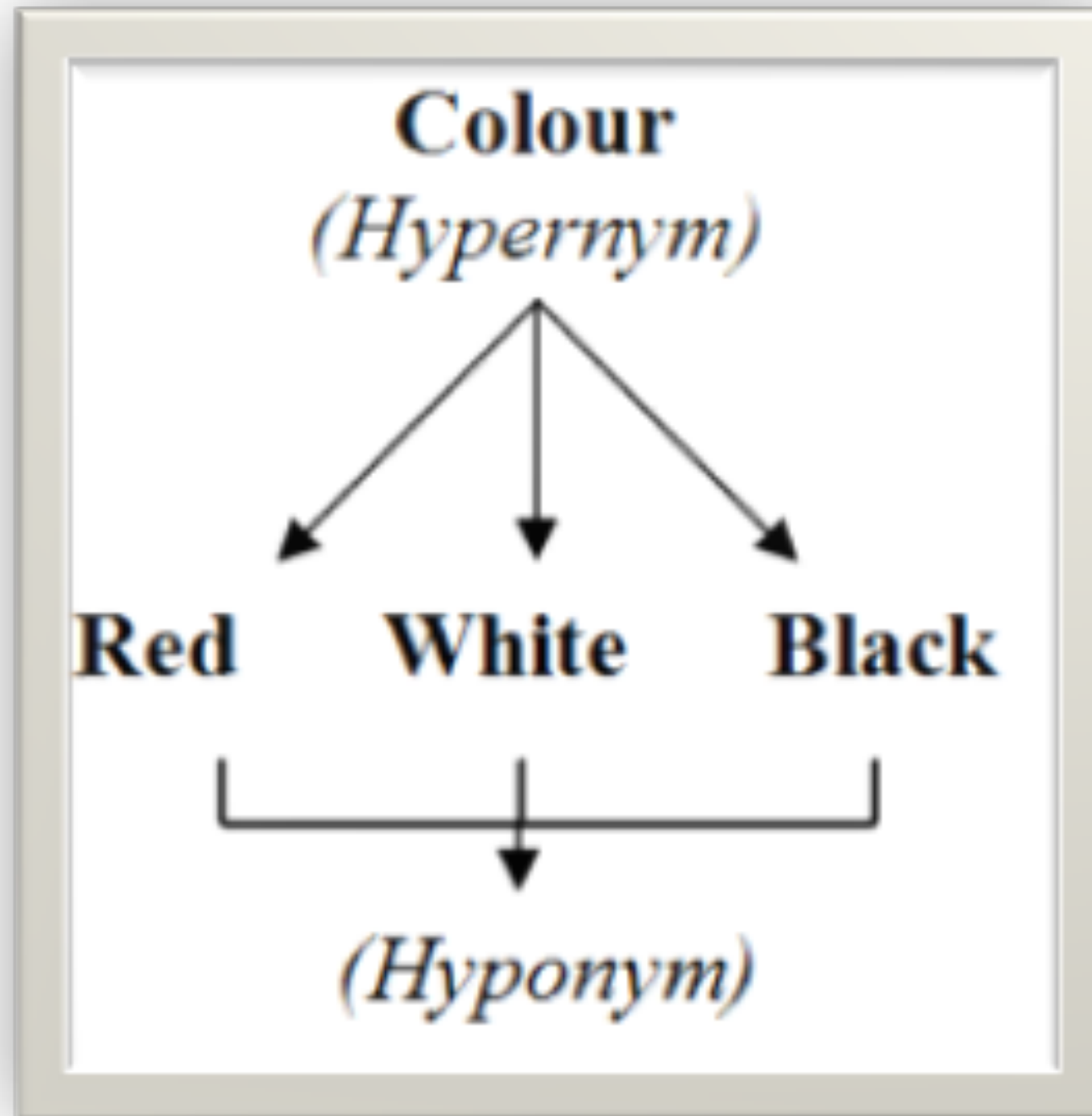- Wordnet and lexical semantics(Q2, Q4)

- Word embedding(Q6)

At the end of this tutorial you will be able to…

1. explain 6 different basic lexical relations

2. explain what it WordNet

3. compute word similarity

4. tell the difference between skip-gram model and CBOW model

Senses—
Synonyms and antonyms

Senses—
Hypernym and Hyponym

Senses—
Holonym and Meronym

- **Meronym:** Part of a whole

- **Holonym:** The whole to which parts belong

Q2

WordNet

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

Q4a PMI

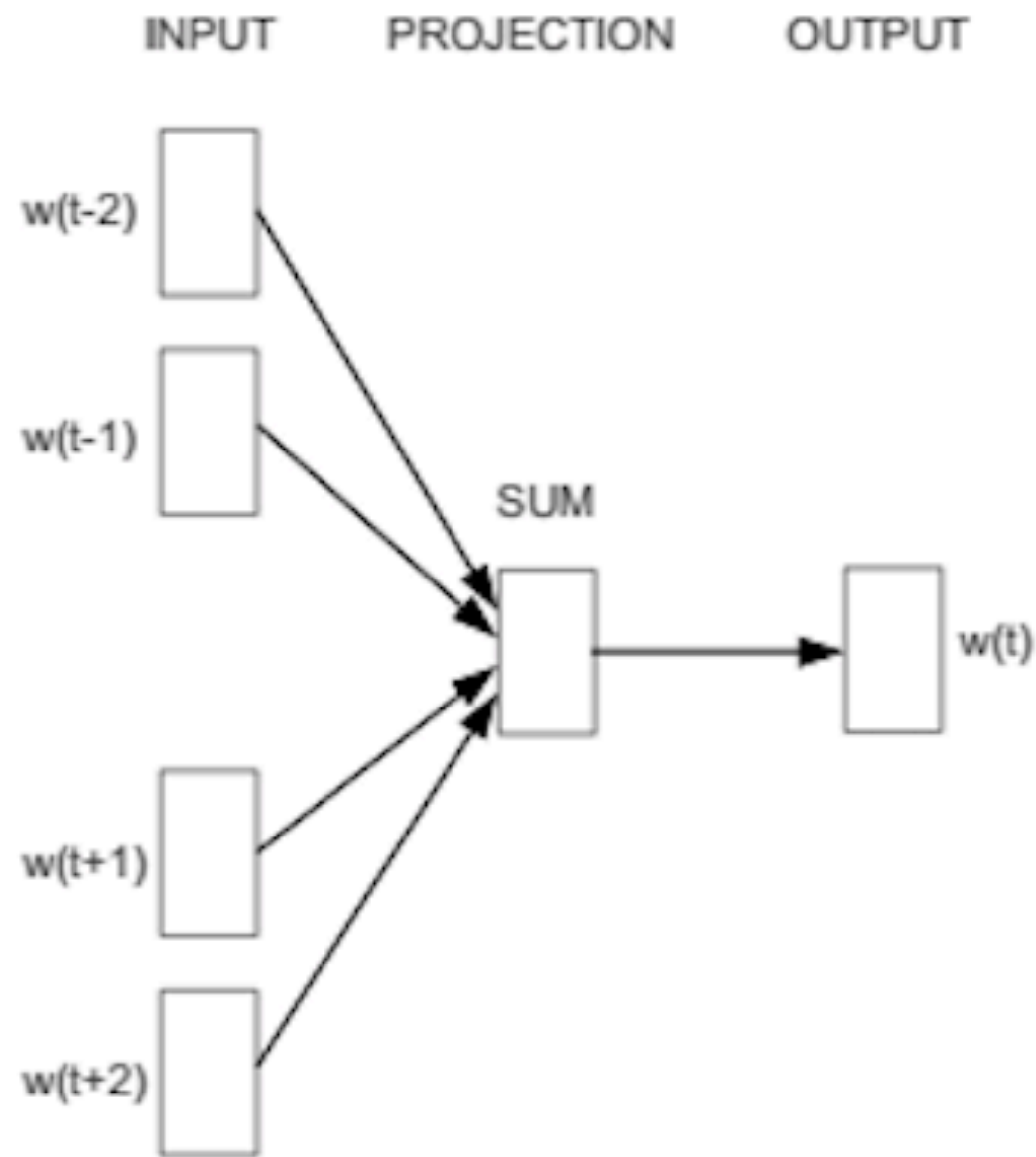|	cup	not (cup)	Total
world	55	225	**280**
not (world)	315	1405	**1720**
Total	**370**	**1630**	**2000**

$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)} = log_2 P(x, y) - log_2 p(x) - log_2 P(y)$$
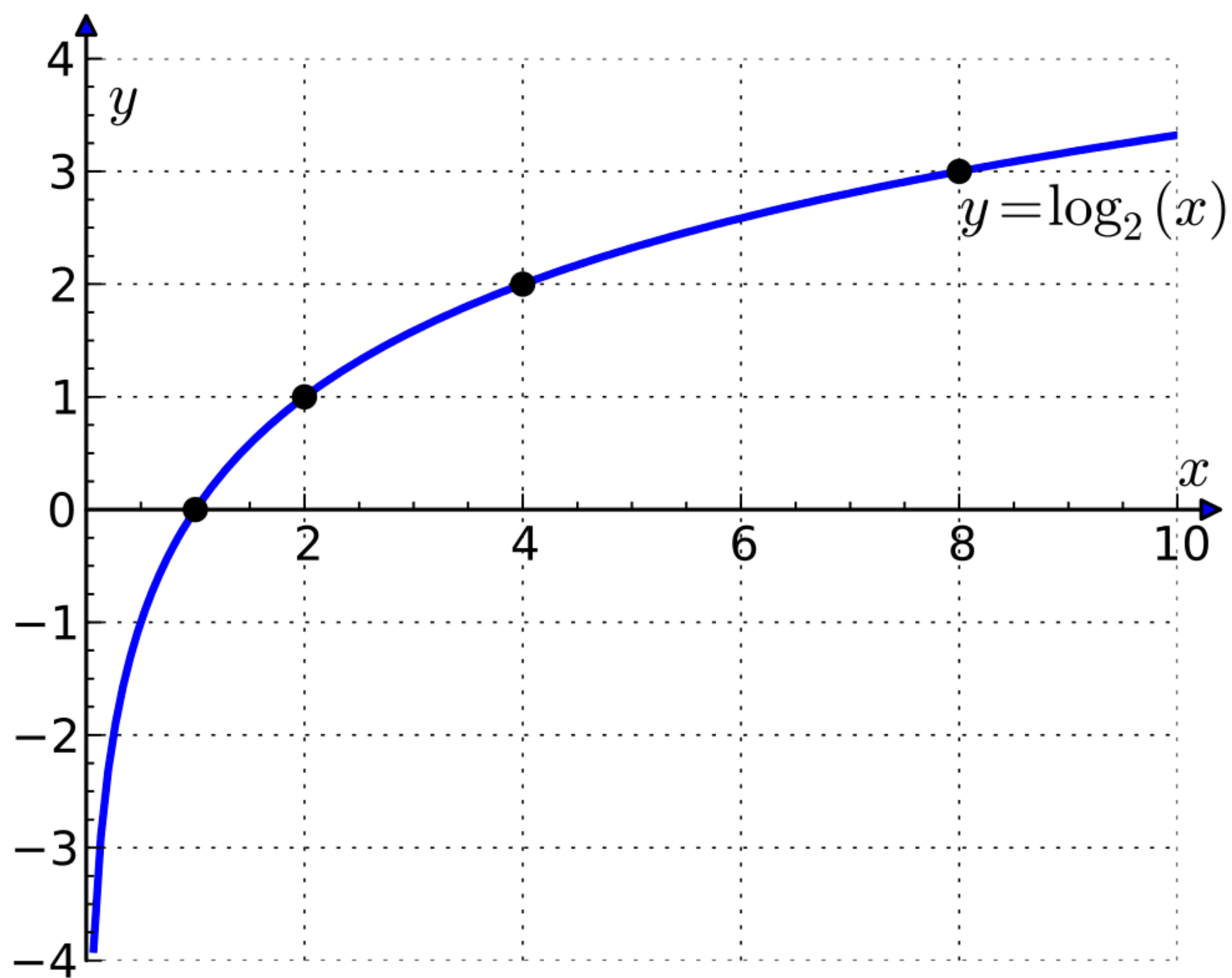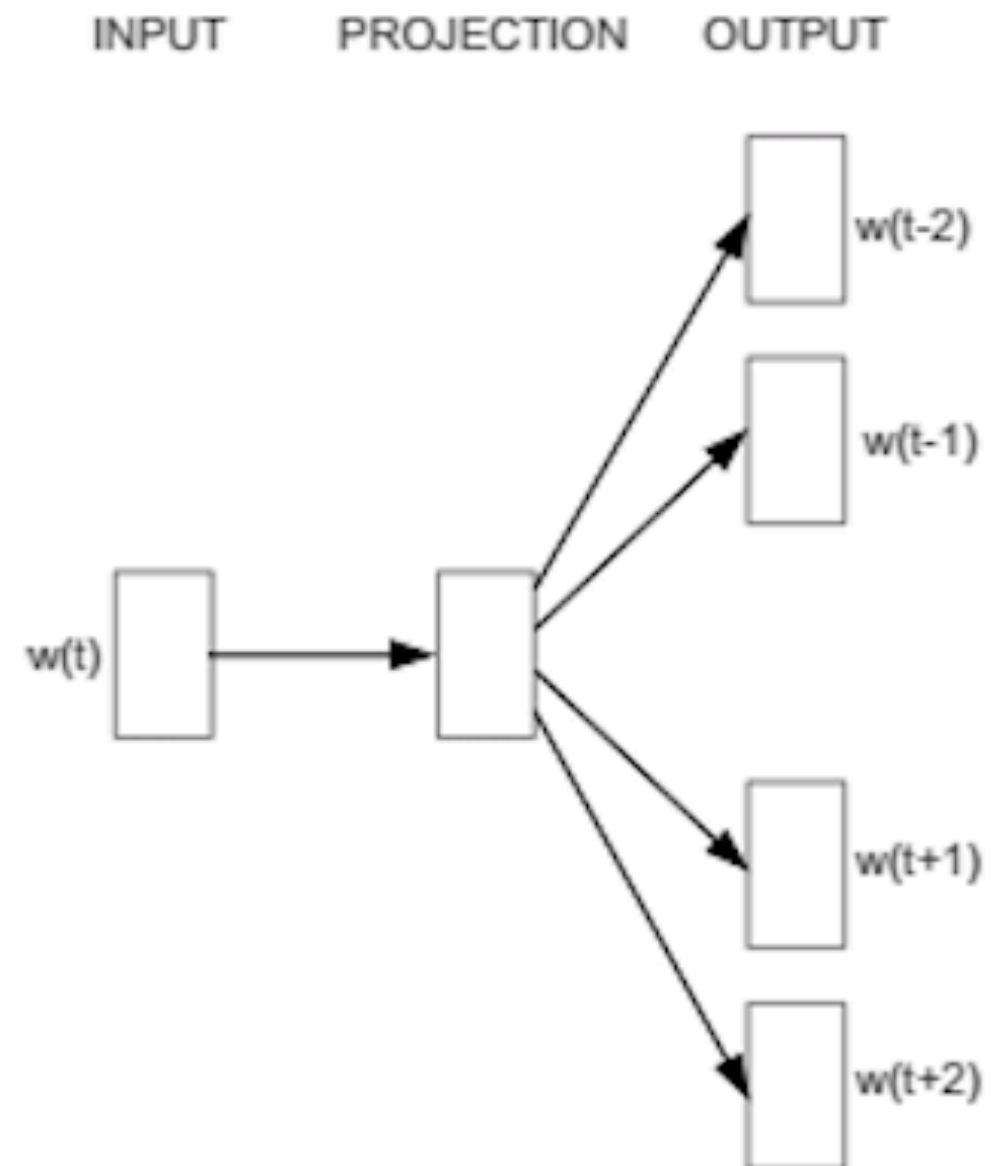
$$PMI(x, y) = log_2\frac{p(x, y)}{p(x)p(y)} = log_2\frac{p(y\,|\,x)}{p(y)} = log_2\frac{p(x\,|\,y)}{x}$$
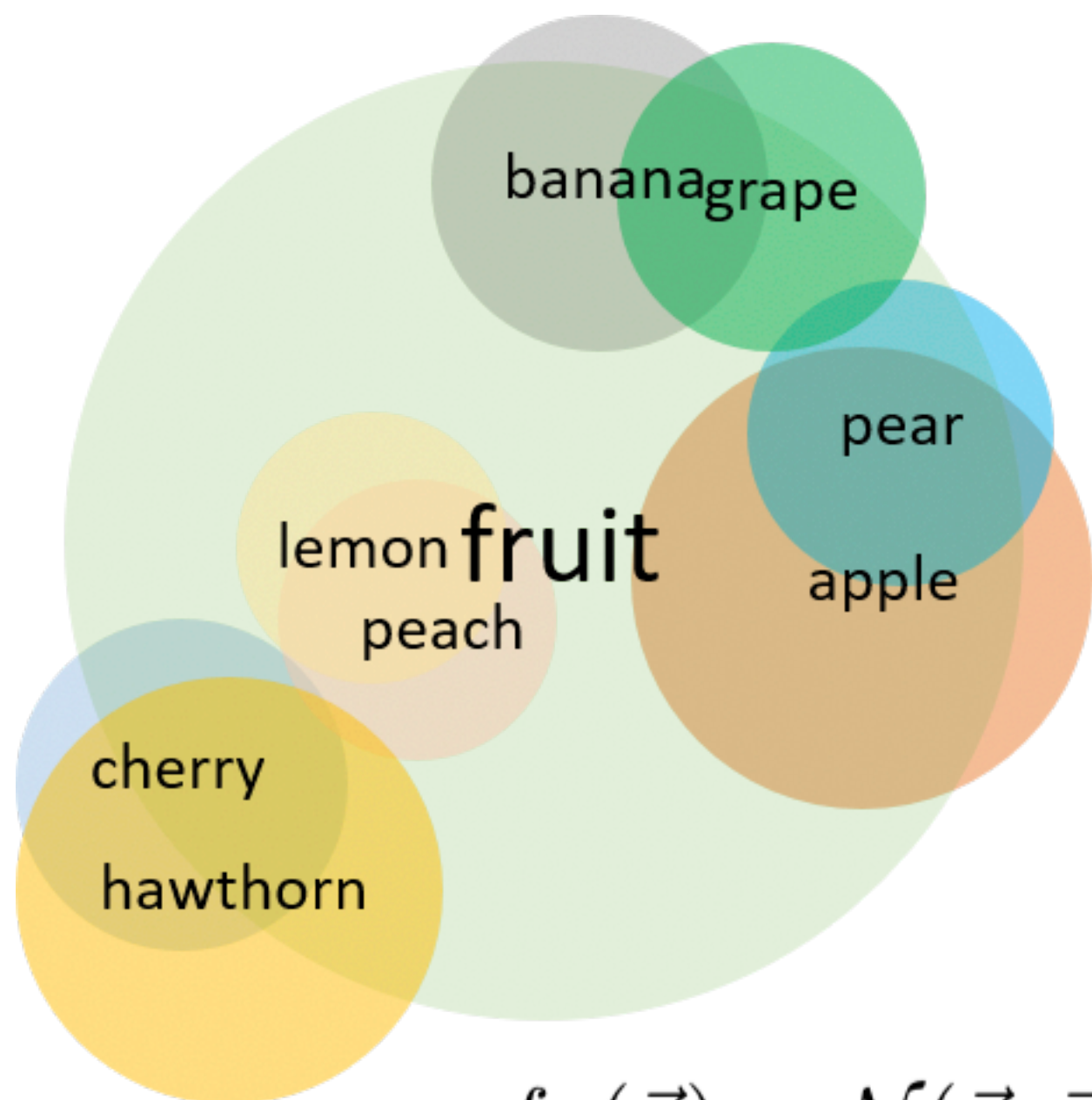
Q6 word to vector

CBOW   Skip-gram

Gaussian Word Embedding

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

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