

# Text Mining and Data Viz

---

2018-05-12

leoluyi@iii

---

Slides <http://pcse.pw/6WHWJ>

# 關於我

- 呂奕 Leo Lu
- 台大工管
- 目前於金融業服務
- Build data products
  - ETL
  - Models
  - Text mining
  - Viz
  - ...



# Text Mining



## 流程 與 工具們

舊時代的工具  
vs.  
新世代的工具

以前我們都用外國人寫的東西

---

tm + tmcn  
Rwordseg



但是這些套件往往在中文

---

會有未知的雷



今天我們要用一些新的工具

# 流程

Get data → Tokenize → Embedding → Viz → Model

PTT 是宅宅的好朋友



每天都有很多很多的廢文語料

# 自己的爬蟲自己寫

---

```
devtools::install_packages(  
  "leoluyi/PTTr")
```



# Get data

Get data → Tokenize → Embedding → Viz → Model

12

# Tokenize

---

Transform whole text  
into parts (of  
distributions)

Get data → Tokenize → Embedding → Viz → Model



# For English

- normalization
- stemming (詞幹提取)
- lemmatization (詞型還原)
- POS tagging
- ...



# 中文似乎比較簡單

- 斷詞
- 不斷詞
- POS tagging
- ...



# R tools

- **stringr**
- **jiebaR**



# Embedding



(Encode, Feature Extraction)

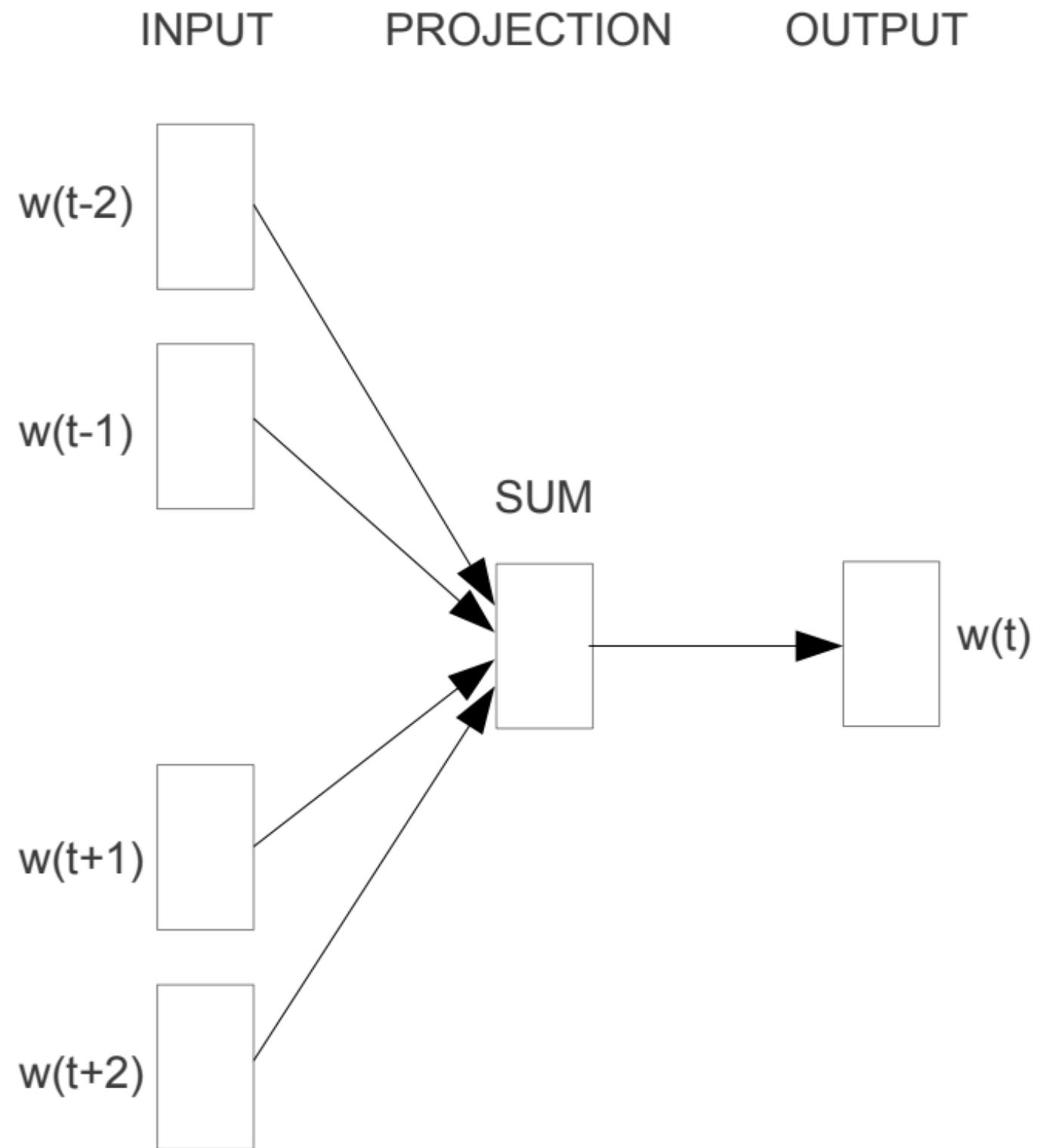
# Embedding

In a nutshell, Word Embedding turns text into numbers.

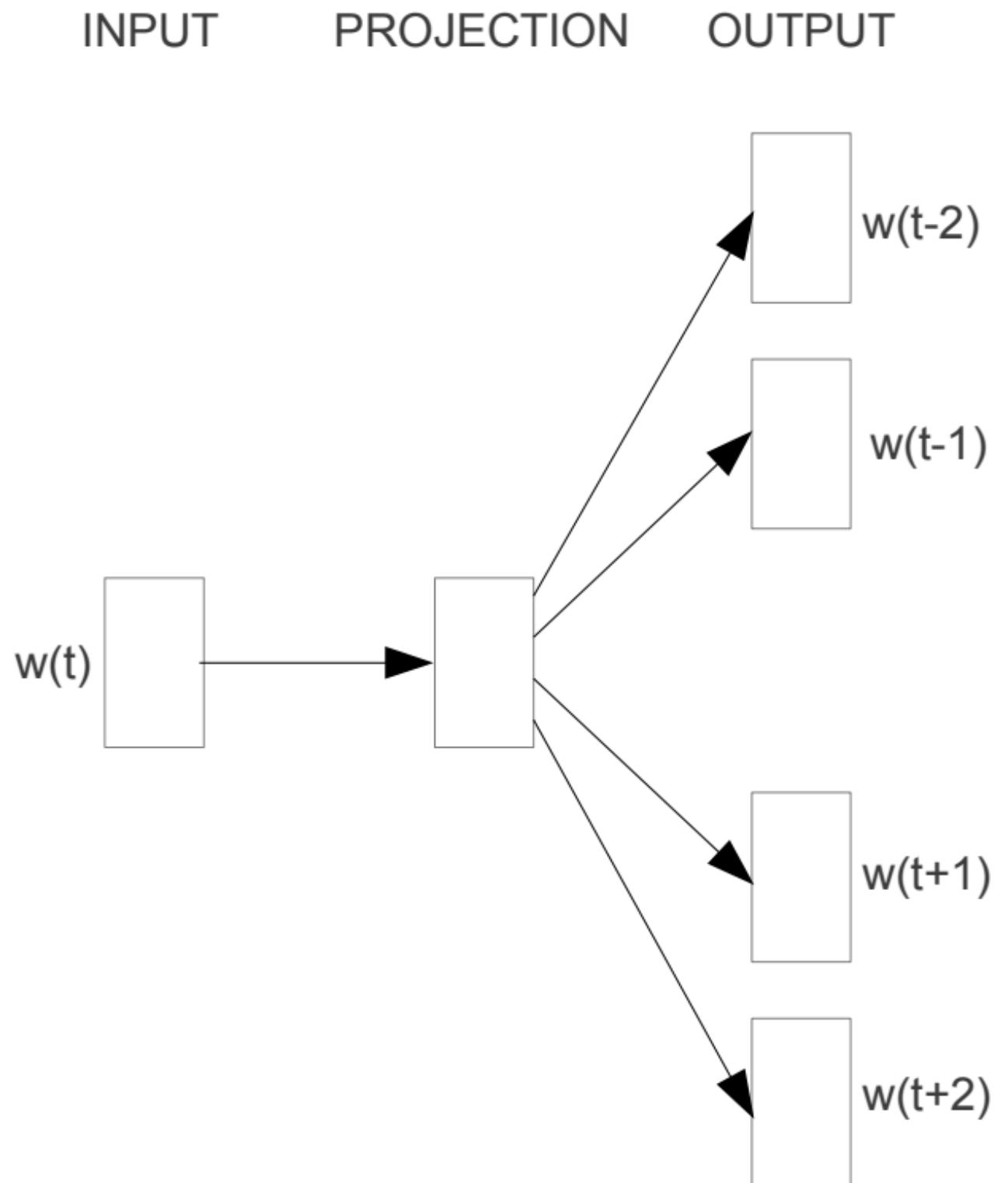
- Embedding Layer<sup>1</sup>
- Word2Vec
- GloVe
- doc2vec
- sense2vec

---

<sup>1</sup><https://machinelearningmastery.com/what-are-word-embeddings/>



**CBOW**



**Skip-gram**

# Demo

...

## Information Retrieval



# Visualize

- Dimension Reduction
  - t-sne
  - PCA
- Clustering
- Interactive or static plots



# Visualize

- `tsne::tsne()`
- `prcomp()`



# Model

# Tasks

- Classification
  - 文本分類
- Clustering
  - 找尋相似文本
- Generative models
  - 文本自動生成



# 用到最後都會想要寫自己的 toolkit

- Sparse Matrix manipulation
- Information retrieval tools
- ...



# 數據科學

呂奕 [leoluyi@github](mailto:leoluyi@github.com)

<https://leoluyi.github.io>