Bukalapak

Natural Language Processing

Intermediate Class

Afif A. Iskandar

About Me

Name : Afif Akbar Iskandar

Role : AI Scientist

Company : Bukalapak

Specialization :

- Computer Vision
- Machine Learning
- Deep Learning
- Natural Language Processing

About Me

Educational Background :

- Bachelor of Mathematics at Universitas Indonesia (2011)
- Master of Computer Science at Universitas Indonesia (2015)

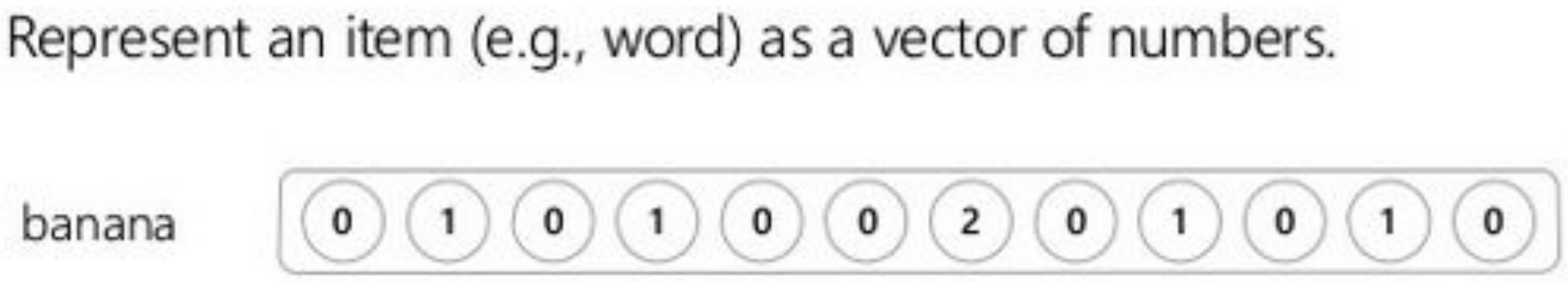
Working Experience :

- Data Scientist (2015-Now)

OUTLINE

- Word Embedding
- Word2Vec
- Recurrent Neural Network

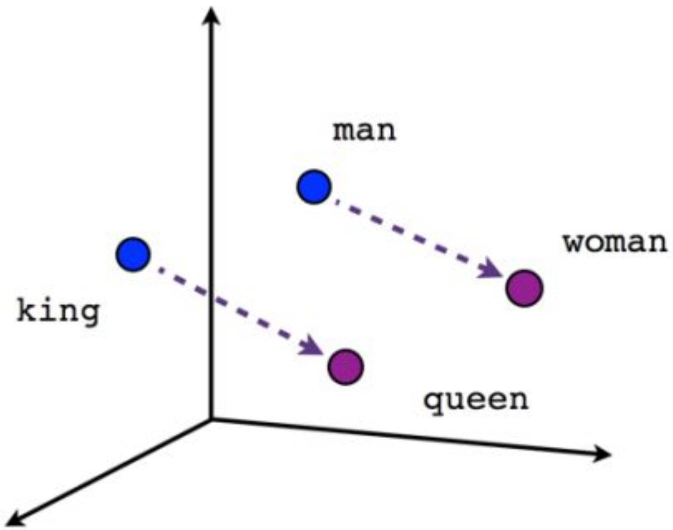
Vector Space Model



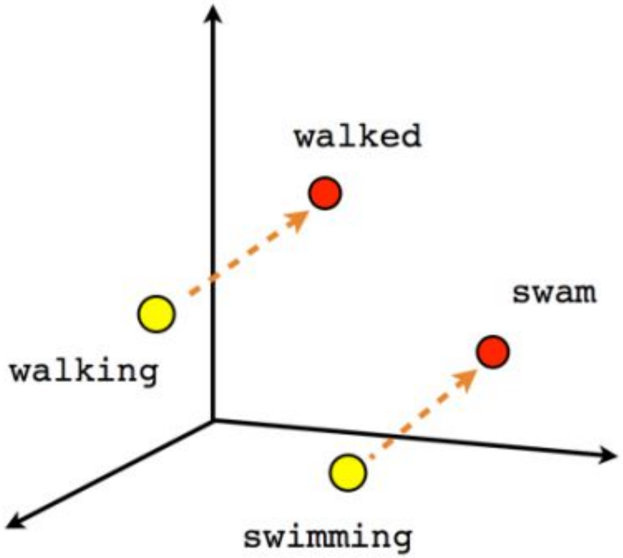
Represent an item (e.g., word) as a vector of numbers.

banana	0	1	0	1	0	0	2	0	1	0	1	0

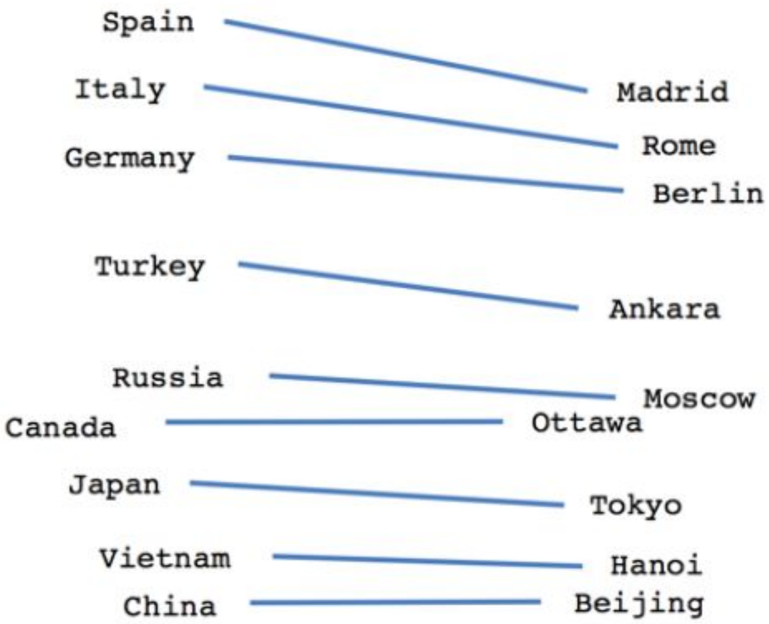
Word Embedding : word -> real vector



Male-Female



Verb tense

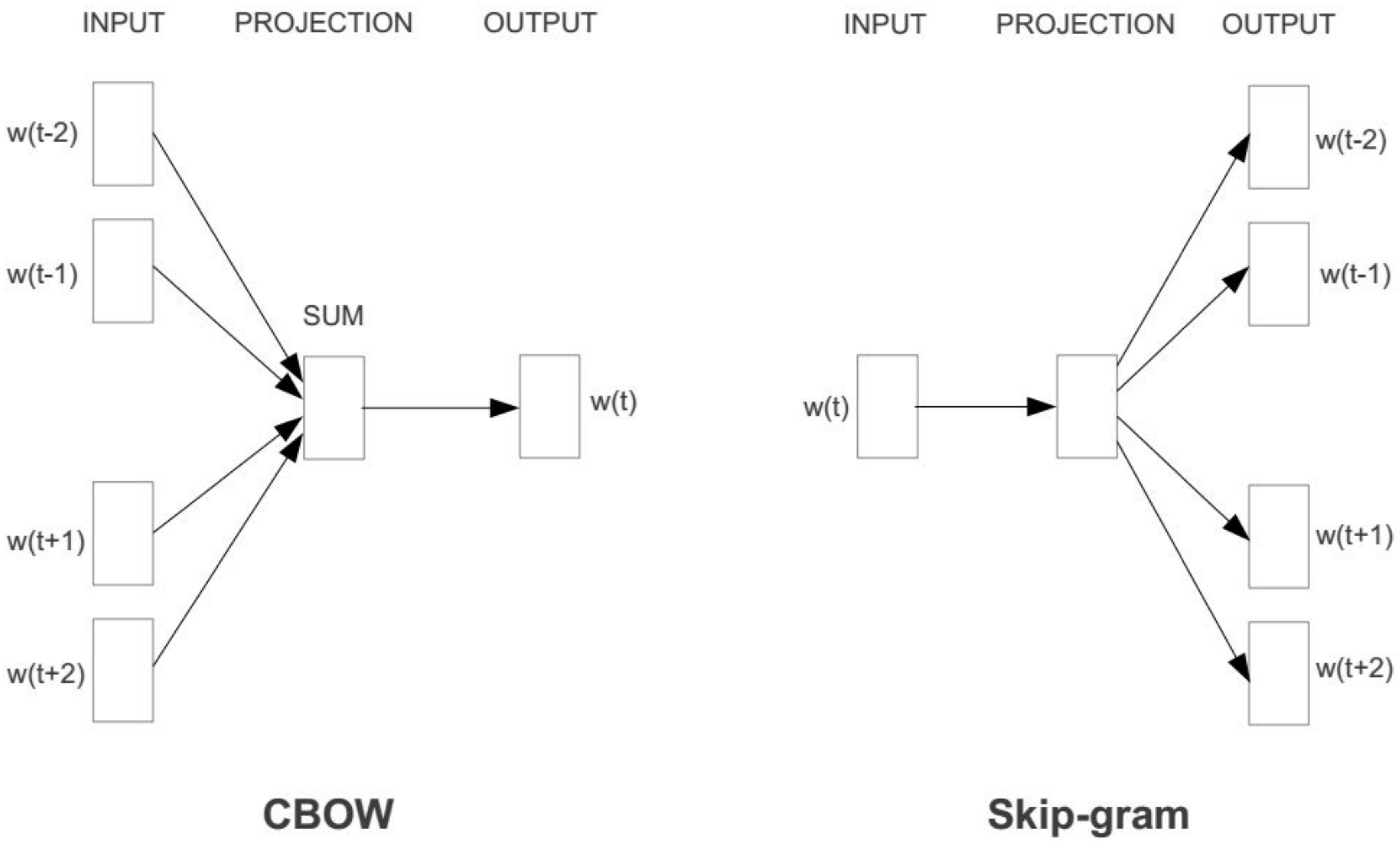


Country-Capital

Word2vec

- Represent each word with a low-dimensional vector
- Word similarity = vector similarity
- Key idea: Predict surrounding words of every word
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary

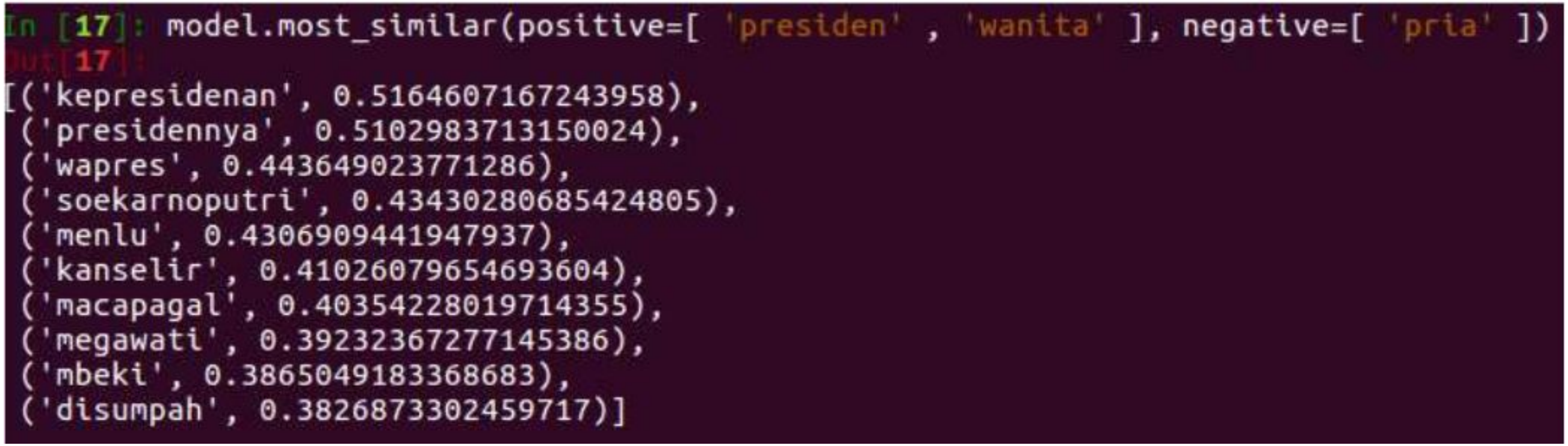
Word2Vec Architecture



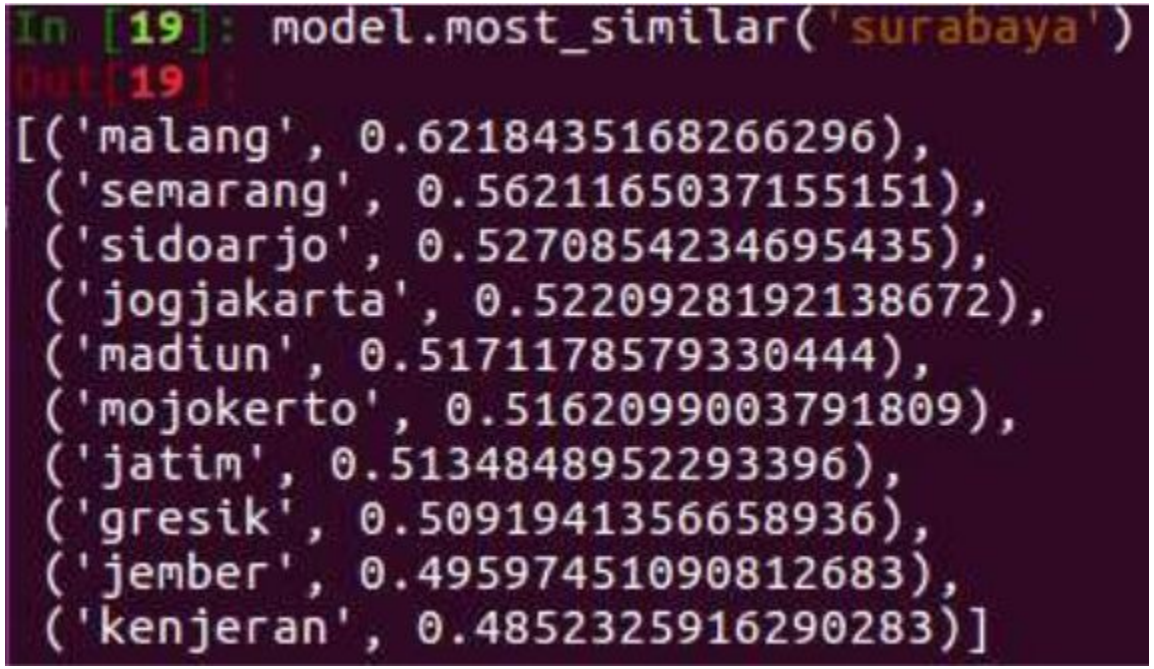
Word2vec Application(s)

- Search, e.g., query expansion
- Sentiment analysis
- Classification
- Clustering

Most Similar Words

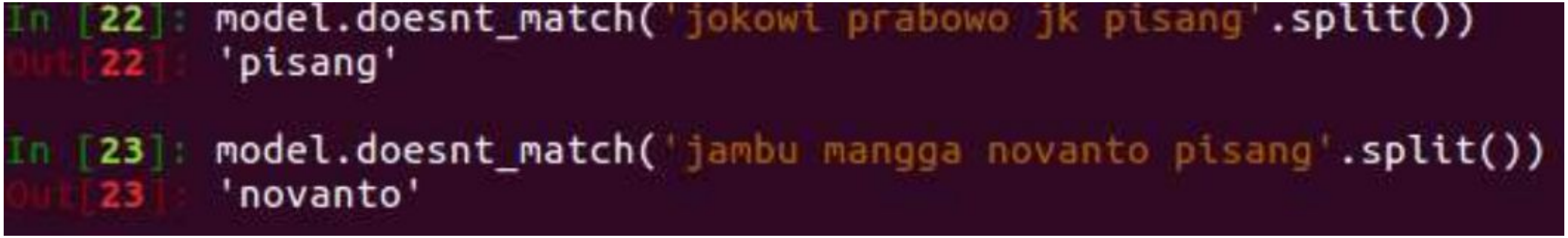


```
In [17]: model.most_similar(positive=[ 'presiden' , 'wanita' ], negative=[ 'pria' ])
Out[17]:
[('kepresidenan', 0.5164607167243958),
 ('presidennya', 0.5102983713150024),
 ('wapres', 0.443649023771286),
 ('soekarnoputri', 0.43430280685424805),
 ('menlu', 0.4306909441947937),
 ('kanselir', 0.41026079654693604),
 ('macapagal', 0.40354228019714355),
 ('megawati', 0.39232367277145386),
 ('mbeki', 0.3865049183368683),
 ('disumpah', 0.3826873302459717)]
```



```
In [19]: model.most_similar('surabaya')
Out[19]:
[('malang', 0.6218435168266296),
 ('semarang', 0.5621165037155151),
 ('sidoarjo', 0.5270854234695435),
 ('jogjakarta', 0.5220928192138672),
 ('madiun', 0.5171178579330444),
 ('mojokerto', 0.5162099003791809),
 ('jatim', 0.5134848952293396),
 ('gresik', 0.5091941356658936),
 ('jember', 0.4959745109081268 3),
 ('kenjeran', 0.4852325916290283)]
```

Word Matching

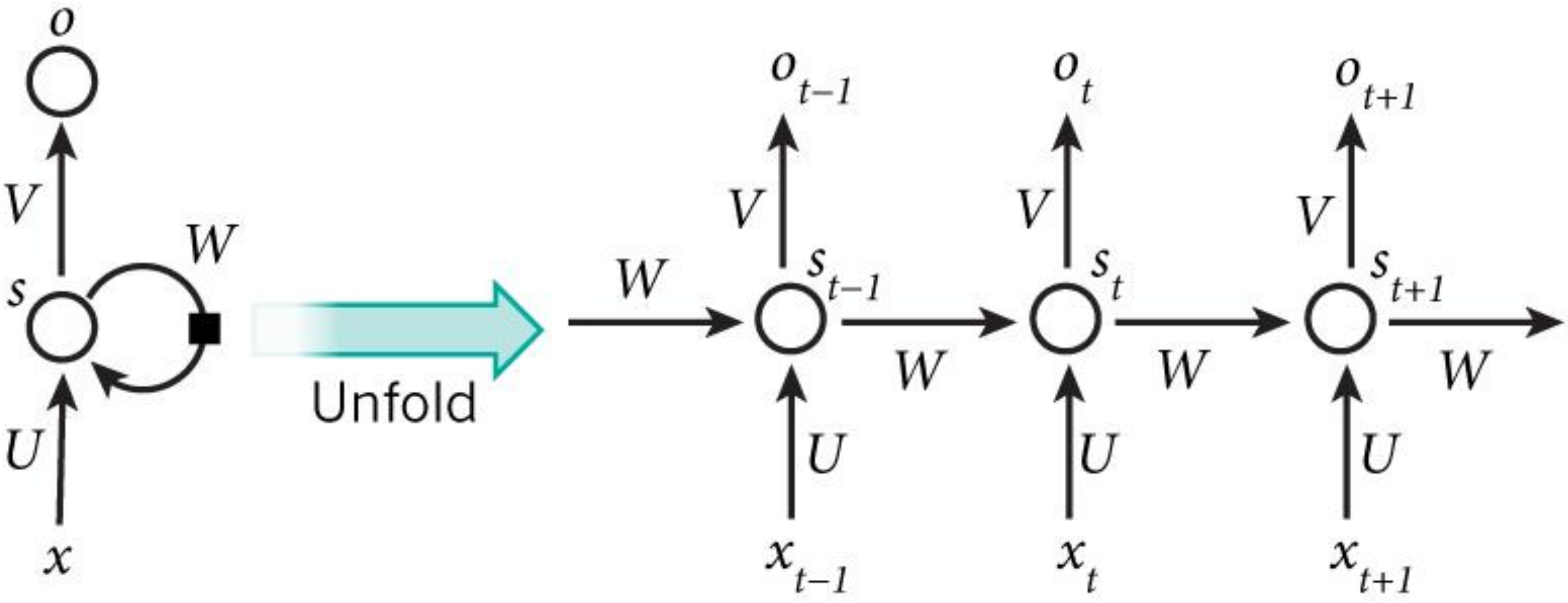


```
In [22]: model.doesnt_match('jokowi prabowo jk pisang'.split())
Out[22]: 'pisang'

In [23]: model.doesnt_match('jambu mangga novanto pisang'.split())
Out[23]: 'novanto'
```

Recurrent Neural Network

What is RNN ?



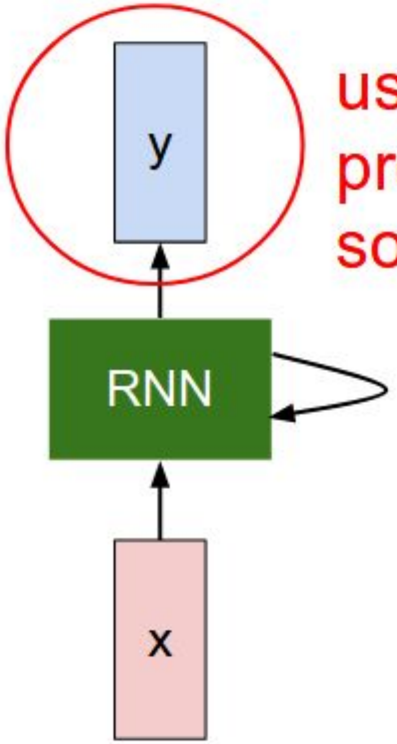
$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state

some function with parameters W

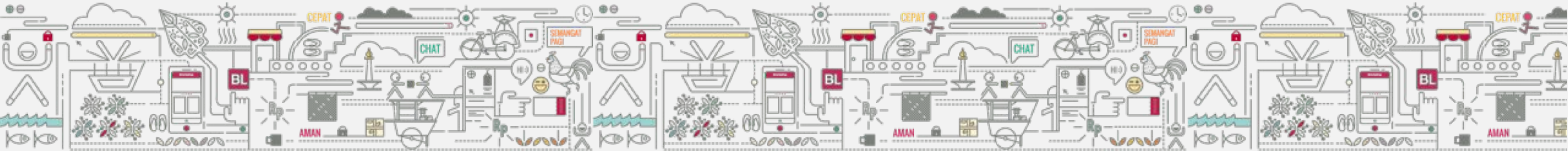
old state

input vector at some time step



usually want to predict a vector at some time steps

Let’s Get Our Hands Dirty



Thank You

Bukalapak