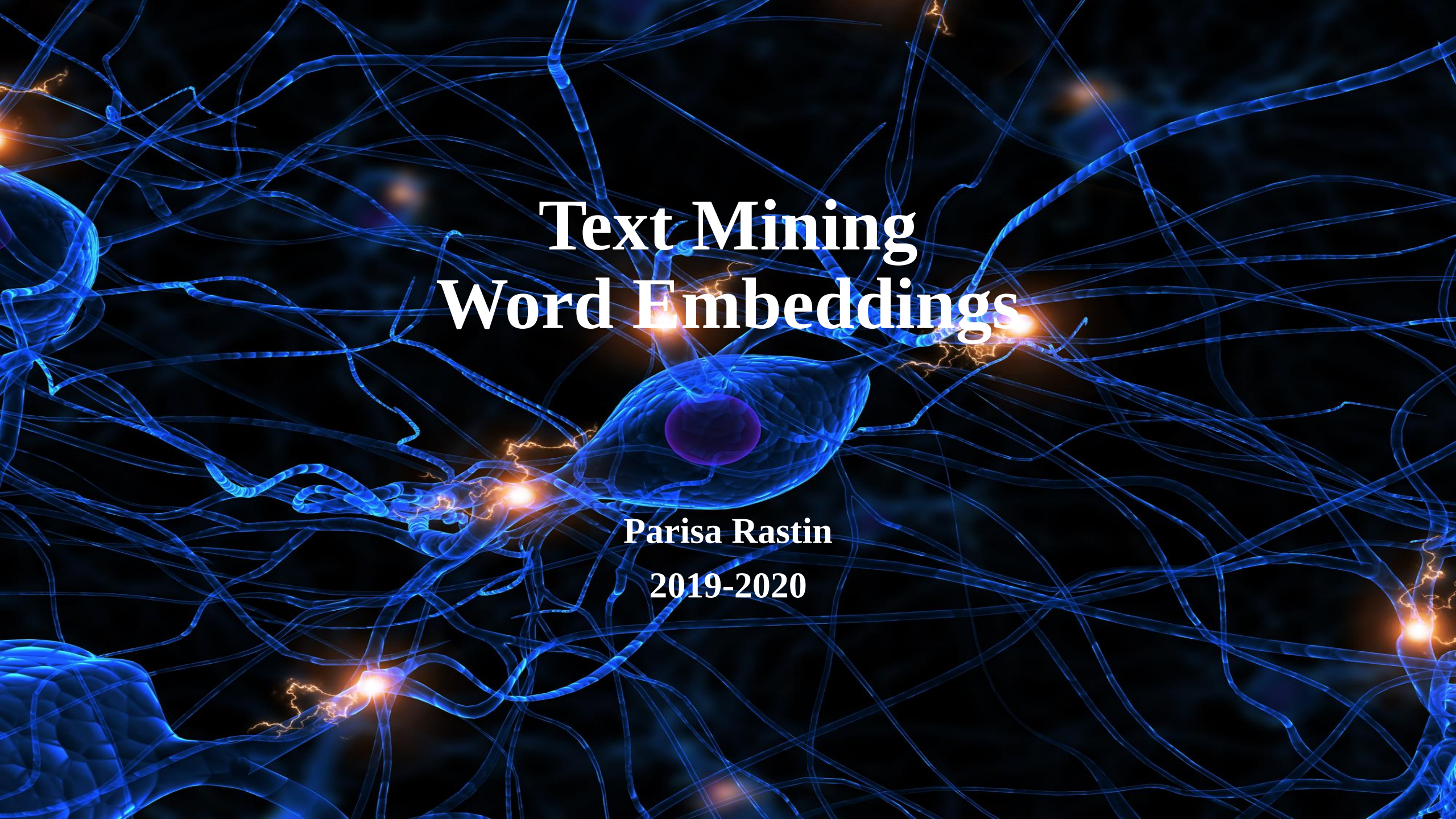
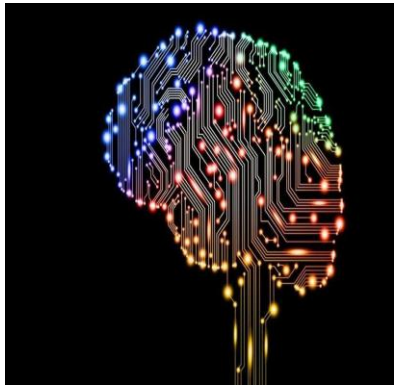


Text Mining
Word Embeddings

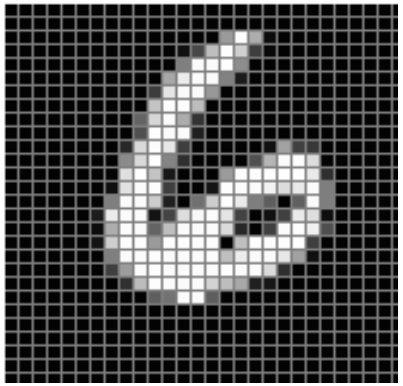
Parisa Rastin

2019-2020

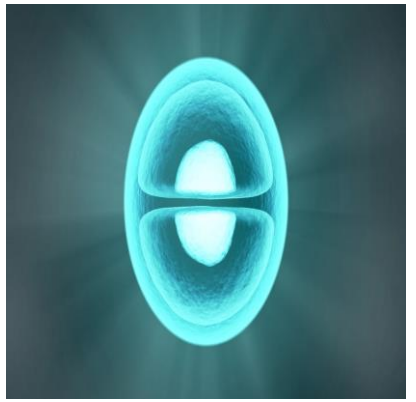
Lecture Schedule



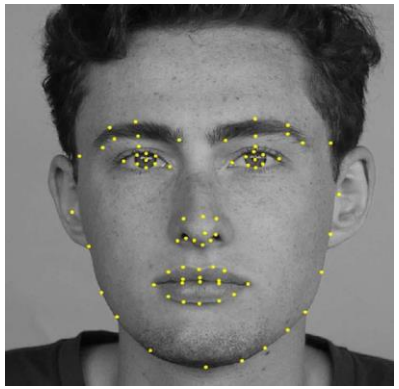
Introduction to
Deep **L**earning



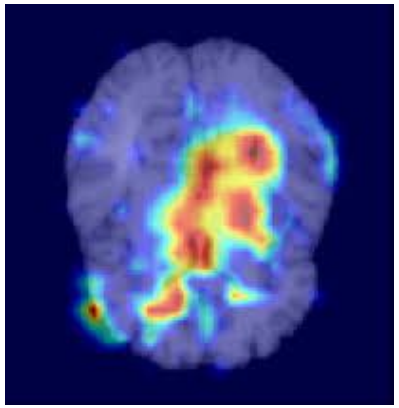
Feed Forward Networks
Convolutional Neural Networks



RNN
Seq2Seq, LM



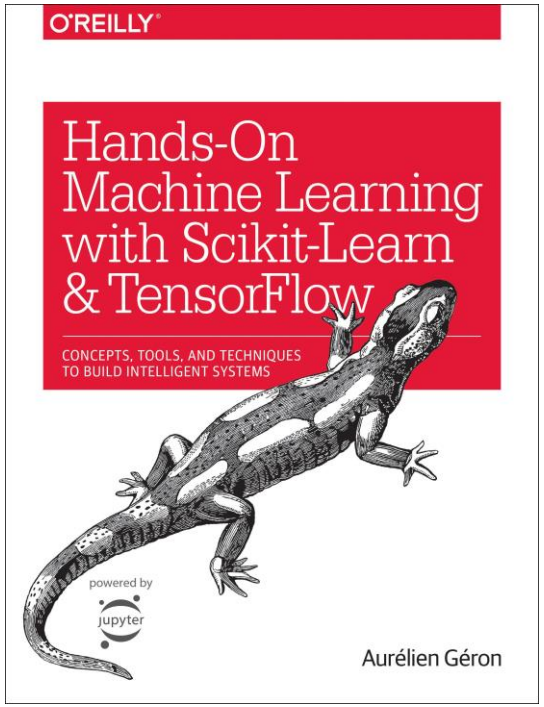
Transformers
DAN

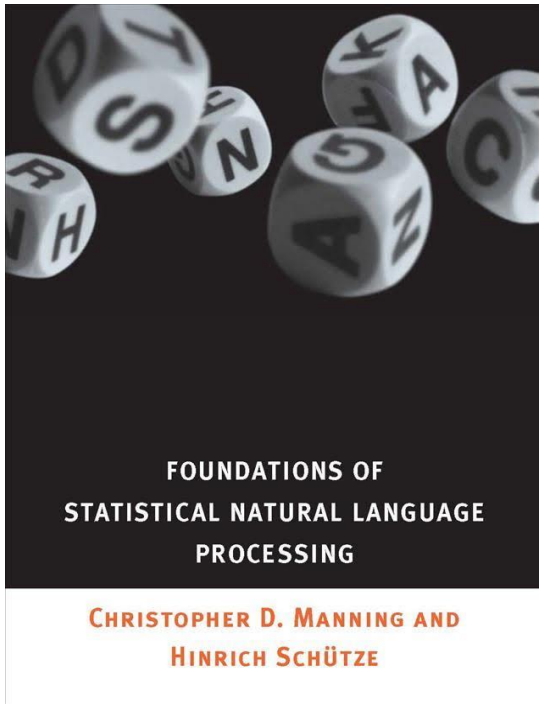


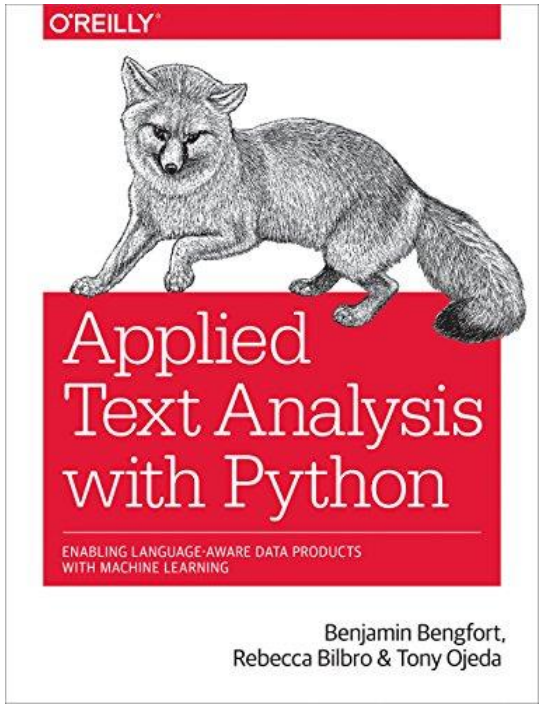
Text Mining
Word Embeddings

Autoencoders
Generative
Adversarial
Networks

Some sources:







Working with Text is… important, under-discussed, and HARD

- Text Mining or Text data Mining or approximately **Text Analytics,** is the process of deriving high-quality information from text.

- First is (not sure!) Thomas C.Mendenhall in 1887, Stylometry, the quantitative analysis of writing style.

SCIENCE.-

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

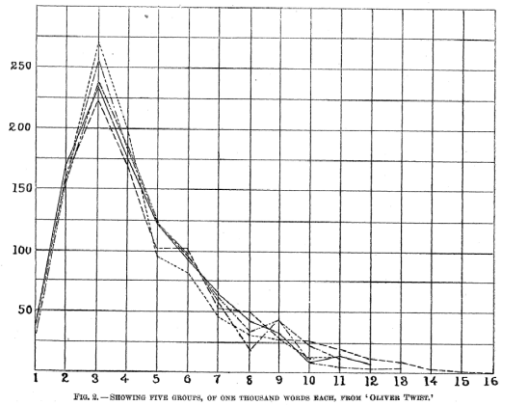
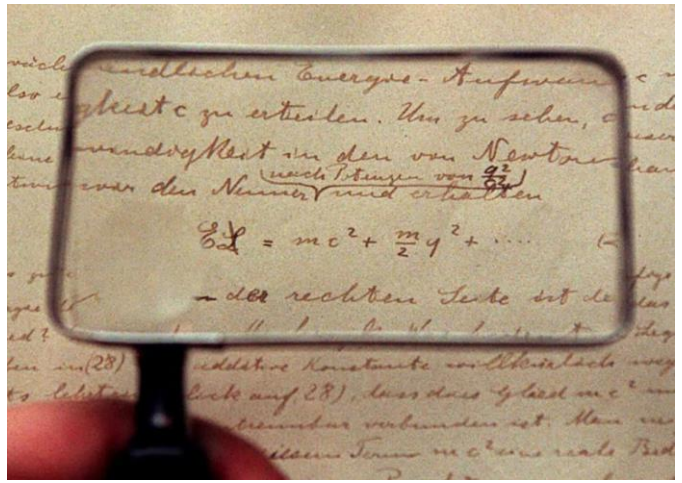


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'



Age of too much information!

We can now easily retrieve far more relevant information than is humanly possible to read.

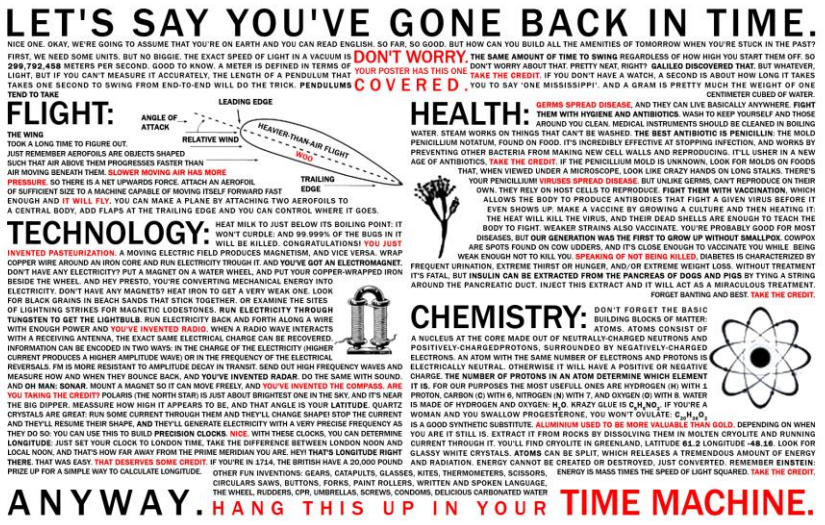
Machine translation, Part-of-speech tagging, Information extraction, Question-answering, Text categorization, Disputed authorship (stylometry), etc.

Text Mining and Language Processing

Text Mining deals with the text itself	NLP deals with the underlying/latent metadata.
Frequency counts of words	Linguistics
length of the sentence	Stylistics
presence/absence of certain words	Content analysis
etc	etc

Mining the texts? What we trying to understand?

- The texts themselves?
- The writer of the texts?
  - The writer as a writer?
  - The writer as an entity in the world?
- Things in the world?
  - Directly linked to texts?
  - Described by texts?



Which interpretation?

- Correct interpretation of a text is not evident!
  - Lexical Ambiguity (Reilly 1991; Walton 1996)
    - She bagged two **silver** medals!
    - She made a **silver** speech!
  - Syntactical Ambiguity (Kooij 1971)
    - How do you stop a fish from **smelling**?
      - Cut off its nose!!!
  - Inflective Ambiguity (Walton 1996; Fowler and Aaron 1998)
    - Bob has devised a scheme to save costs by recycling paper. Therefore, Bob is a schemer, and should not be trusted.

words have more than one meaning

Which meaning? Odor or Smell?

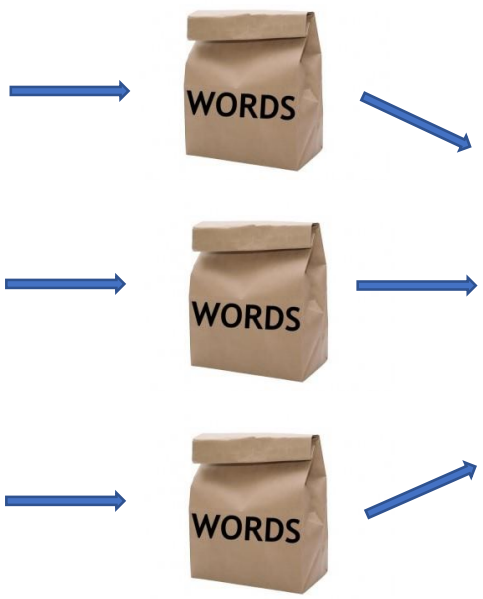
A word is used more than once in a sentence or paragraph, but with different meanings each time

Text data

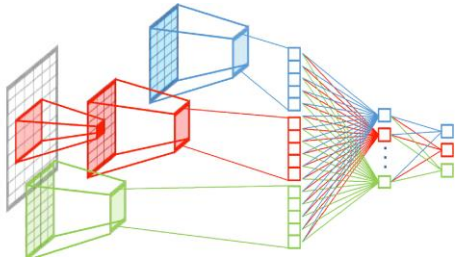
- Text documents in a natural language
- Unstructured
- Documents in plain text, Word or PDF format
- Emails, online chat log and phone transcripts
- Online news and forums, blogs, micro-blogs and social media
- …

Typical Process of Text Mining

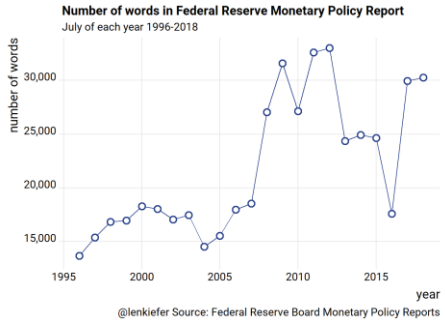
Documents (unstructured)



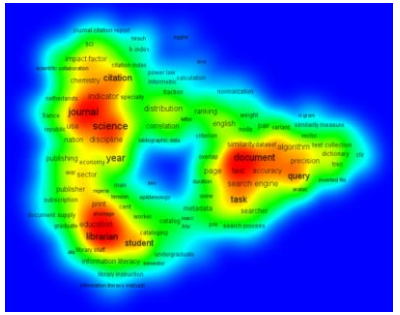
Bag of words



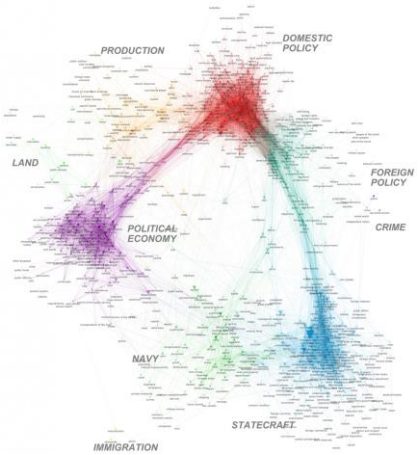
Modelling



Statistical analysis



Visualisation



Clustering

Typical Process of Text Mining

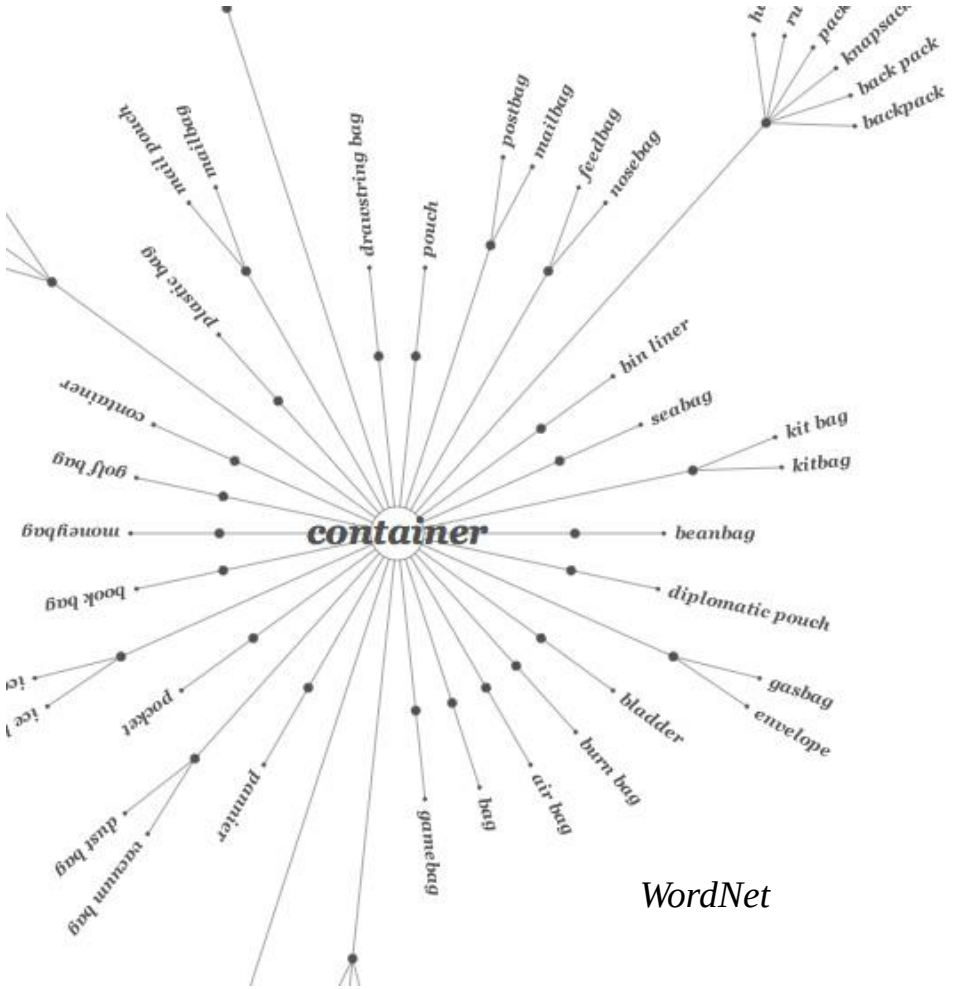
- Transform text into structured data
  - Term-Document Matrix (TDM)
  - Entities and relations
  - …
- Apply traditional data mining techniques to the above structures data
  - Clustering
  - Classification
  - Social Network Analysis
  - …

document-term matrix

	I	Like	Hate	Databases
Doc 1	1	1	0	1
Doc 2	1	0	1	1

Text Mining Tasks

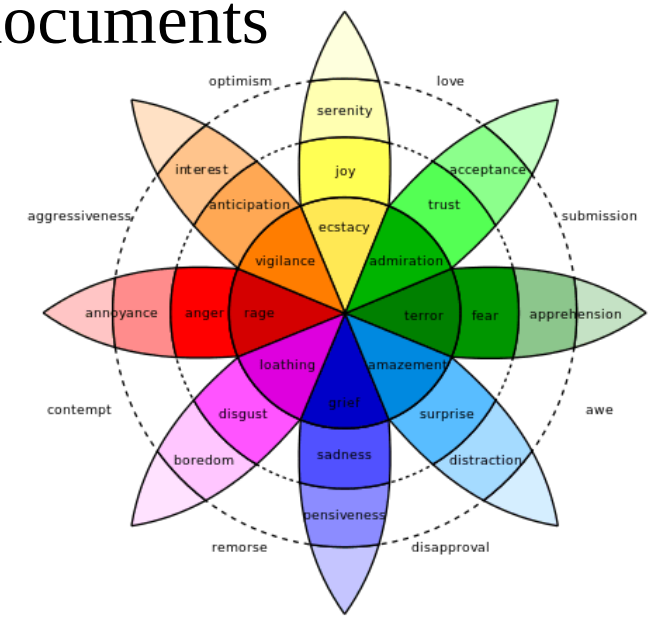
- Text classification
- Text clustering and categorization
- Topic modelling
- Sentiment analysis
- Document summarization
- Entity and relation extraction
- ….



WordNet

Sentiment Analysis

- Also known as *opinion mining*
- To determine attitude, polarity or emotions from documents
- Polarity: positive, negative, netural



- Emotions: angry, sad, happy, bored, afraid, etc.
- Method:
  1. Identify individual words and phrases and map them to different emotional scales
  2. Adjust the sentiment value of a concept based on modifications surrounding it

Document Summarization

- Approaches:
  - Extraction:
    - Select a subset of existing words, phrases or sentences to build a summary
  - Abstraction:
    - Use natural language generation techniques to build a summary that is similar to natural language

Entity and Relationship Extraction

- Named Entity Recognition (NER): identify named entities intext into pre-defined categories, such as person names, organizations, locations, date and time, etc.

- Relationship Extraction: identify associations among entities.

- Example:
  - Tom lives at 7 Champs Elysées, Paris.

Entity and Relationship Extraction

- Named Entity Recognition (NER): identify named entities intext into pre-defined categories, such as person names, organizations, locations, date and time, etc.

- Relationship Extraction: identify associations among entities.

- Example:
  - Tom lives at 7 Champs Elysées, Paris.

Entity and Relationship Extraction

- Named Entity Recognition (NER): identify named entities intext into pre-defined categories, such as person names, organizations, locations, date and time, etc.

- Relationship Extraction: identify associations among entities.

- Example:
  - Tom lives at 7 Champs Elysées, Paris.

Tom 7 Champs Elysées, Paris.

Bag of Words

"Tokenization" process: Splitting a stream of characters into sentences, symbols, words. We are particularly interested in words.

1. Identify the words (tokens) in the documents

The word delimiter identification is important. It will often be space, punctuations, … Some characters are less obvious (e.g. “-” in bio-informatic)

2. Who will constitute the dictionary?

Potentially, the number of words is very important. There may be redundancies in the dictionary, it will be necessary to treat them (sometimes obvious: car vs. cars, sometimes less: car vs. car ...).

3. The absence or presence of words is then associated with each document.

It can also be the number of appearance of the words, or other type of values. We are talking about weighting

Using a lexical database

- List, classify and relate the semantic and lexical content of a language (eg ***WORDNET***).
- Synset
  - The synset (synonym set) correspond to groups of interchangeable words. Ex. In English: car, automobile, machine, motorcar, etc. In the bag of word representation, they can significantly reduce dimensionality.
- Ontologies (Concept trees)

car, auto, automobile, machine, motorcar motor vehicle, automotive vehicle

vehicle

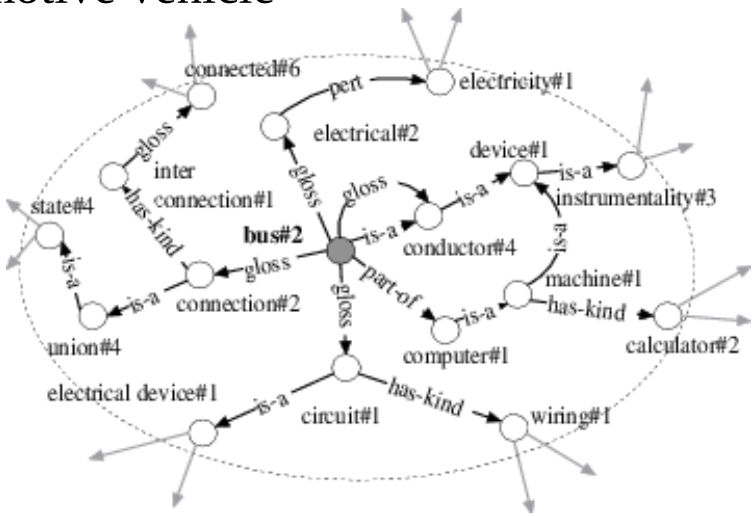
conveyance, transport

instrumentality, instrumentation

artifact, artefact

object, physical object

entity, something



Part of speech (POS)

- Part of speech (POS) proposes to distinguish the words of a sentence according to their lexical categories (e.g. noun, verb, adjective, etc).
  - They refuse to permit us to obtain the refuse permit.

>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")

>>> nltk.pos_tag(text)[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]

Lemmatization

- Lemmatization consists in analyzing the terms in order to identify its canonical form (lemma). The idea is to reduce the different forms (plural, feminine, conjugation, etc.) into one.

Ex. am, are, is → be

car, cars, car's, cars' → car

Thus, the phrase :

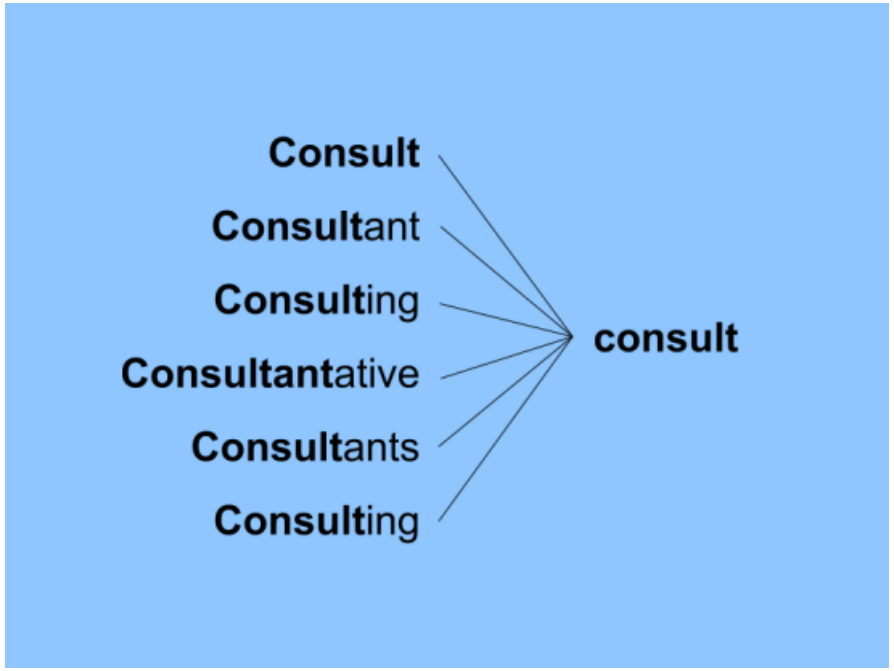
"*the boy's cars are different colors*"

Becomes:

"*the boy car be different color*".

Stemming

- The stemming is to reduce a word to its root (stem).
- The stemming is a final treatment, which no longer allows post treatments on the words.



TF-IDF

- Term Frequency (TF) $tf_{i,j}$: the number of occurrences of term t_i in document d_j
- Inverse Document Frequency (IDF) for term t_i is:

$$\mathrm{idf}_i = \log_2 \frac{|D|}{|\{d \mid t_i \in d\}|}$$

|D|: the total number of documents

$|\{d \mid t_i \in d\}|$: the number of documents where term t_i appears

- Term Frequency - Inverse Document Frequency (TF-IDF)

$$\mathrm{tfidf} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i$$

- IDF reduces the weight of terms that occur frequently in documents and increases the weight of terms that occur rarely.

Example

- Doc 1: I love Paris
- Doc 2: I love Rome

Term Frequency

	Doc 1	Doc 2
I	1	1
Love	1	1
Paris	1	0
Rome	0	1

IDF

	IDF
I	0
Love	0
Paris	1
Rome	1

TFIDF

	Doc 1	Doc 2
I	0	0
Love	0	0
Paris	1	0
Rome	0	1

Terms that can distinguish different documents are given greater weights.

Cleaning the text

1. Load Data

```
# load text
filename = 'metamorphosis_clean.txt'
file = open(filename, 'rt')
text = file.read()
file.close()
```

2. Split by Whitespace

```
# load text
filename = 'metamorphosis_clean.txt'
file = open(filename, 'rt')
text = file.read()
file.close()
# split into words by white space
words = text.split()
print(words[:100])
```

3. Select Words

```
# load text
filename = 'metamorphosis_clean.txt'
file = open(filename, 'rt')
text = file.read()
file.close()
# split based on words only
import re
words = re.split(r'\W+', text)
print(words[:100])
```

['One', 'morning,', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams,', 'he', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back,', 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly,', 'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs,', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him,', 'waved', 'about', 'helplessly', 'as', 'he', 'looked.', '"What\'s', 'happened', 'to', 'me?"', 'he', 'thought.', 'It', "wasn't", 'a', 'dream.', 'His', 'room,', 'a', 'proper', 'human']

Cleaning the text

4. Split by Whitespace and Remove Punctuation

```
print(string.punctuation)
```

```
!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~
```

```
# load text
filename = 'metamorphosis_clean.txt'
file = open(filename, 'rt')
text = file.read()
file.close()
# split into words by white space
words = text.split()
# remove punctuation from each word
import string
table = str.maketrans('', '', string.punctuation)
stripped = [w.translate(table) for w in words]
print(stripped[:100])
```

5. Normalizing Case

```
filename = 'metamorphosis_clean.txt'
file = open(filename, 'rt')
text = file.read()
file.close()
# split into words by white space
words = text.split()
# convert to lower case
words = [word.lower() for word in words]
print(words[:100])
```

['One', 'morning,', 'when', 'Gregor', 'Samsa', 'woke', 'from', 'troubled', 'dreams,', 'he', 'found', 'himself', 'transformed', 'in', 'his', 'bed', 'into', 'a', 'horrible', 'vermin.', 'He', 'lay', 'on', 'his', 'armour-like', 'back,', 'and', 'if', 'he', 'lifted', 'his', 'head', 'a', 'little', 'he', 'could', 'see', 'his', 'brown', 'belly,', 'slightly', 'domed', 'and', 'divided', 'by', 'arches', 'into', 'stiff', 'sections.', 'The', 'bedding', 'was', 'hardly', 'able', 'to', 'cover', 'it', 'and', 'seemed', 'ready', 'to', 'slide', 'off', 'any', 'moment.', 'His', 'many', 'legs,', 'pitifully', 'thin', 'compared', 'with', 'the', 'size', 'of', 'the', 'rest', 'of', 'him,', 'waved', 'about', 'helplessly', 'as', 'he', 'looked.', '"What\'s', 'happened', 'to', 'me?"', 'he', 'thought.', 'It', "wasn't", 'a', 'dream.', 'His', 'room,', 'a', 'proper', 'human']

Additional Text Cleaning Considerations

- Handling large documents and large collections of text documents that do not fit into memory.
- Extracting text from markup like HTML, PDF, or other structured document formats.
- Transliteration of characters from other languages into English.
- Decoding Unicode characters into a normalized form, such as UTF8.
- Handling of domain specific words, phrases, and acronyms.
- Handling or removing numbers, such as dates and amounts.
- Locating and correcting common typos and misspellings.
- …

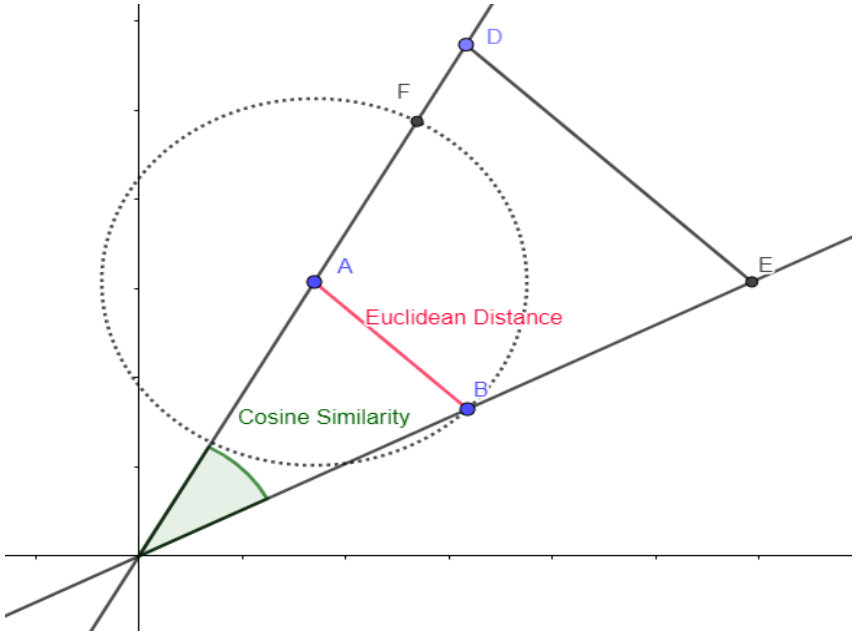
Similarity measure

Similarity measures underlie many data mining methods (visualization, supervised and unsupervised classification). They characterize the similarities between objects.

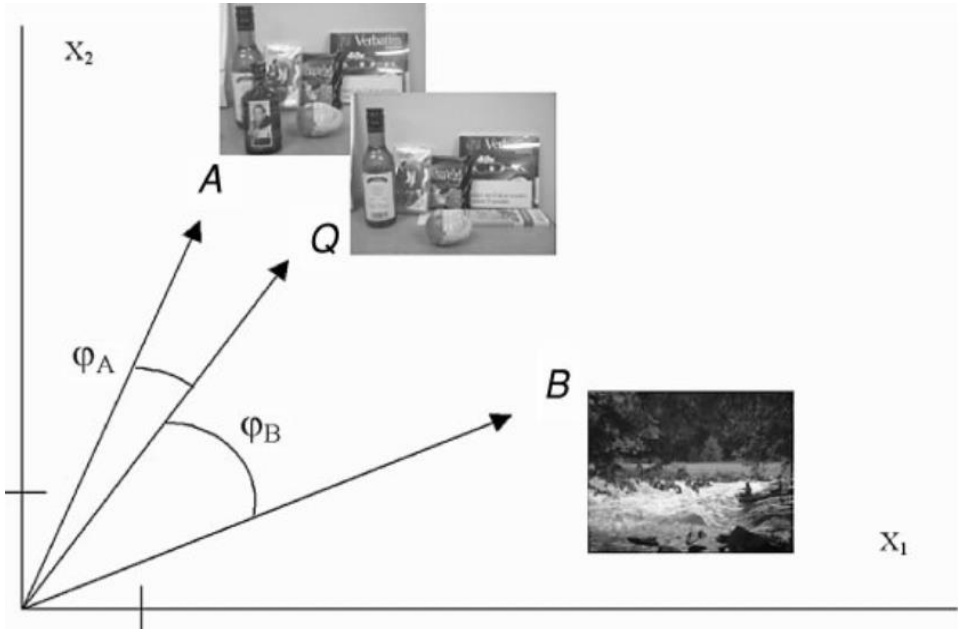
In the particular case of text mining, we need to measure similarities between documents.

Which similarity measure?

Similarity measure



Euclidean distance & Cosine distance



Cosine distance

Similarity measure

```
# Define the documents
doc_trump = "Mr. Trump became president after winning the political election. Though he lost t
he support of some republican friends, Trump is friends with President Putin"

doc_election = "President Trump says Putin had no political interference is the election outco
me. He says it was a witchhunt by political parties. He claimed President Putin is a friend wh
o had nothing to do with the election"

doc_putin = "Post elections, Vladimir Putin became President of Russia. President Putin had se
rved as the Prime Minister earlier in his political career"

documents = [doc_trump, doc_election, doc_putin]
```

```
# Scikit Learn
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

# Create the Document Term Matrix
count_vectorizer = CountVectorizer(stop_words='english')
count_vectorizer = CountVectorizer()
sparse_matrix = count_vectorizer.fit_transform(documents)

# OPTIONAL: Convert Sparse Matrix to Pandas Dataframe if you want to see the word frequencies.
doc_term_matrix = sparse_matrix.todense()
df = pd.DataFrame(doc_term_matrix,
                  columns=count_vectorizer.get_feature_names(),
                  index=['doc_trump', 'doc_election', 'doc_putin'])
df
```

	after	as	became	by	career	claimed	do	earlier	election	elections	...	the	though	to	trump	vladimir
doc_trump	1	0	1	0	0	0	0	0	1	0	...	1	1	0	2	0
doc_election	0	0	0	1	0	1	1	0	2	0	...	2	0	1	1	0
doc_putin	0	1	1	0	1	0	0	1	0	1	...	1	0	0	0	1

3 rows × 48 columns

```
# Compute Cosine Similarity
from sklearn.metrics.pairwise import cosine_similarity
print(cosine_similarity(df, df))
#> [[ 1.          0.48927489  0.37139068]
#>  [ 0.48927489  1.          0.38829014]
#>  [ 0.37139068  0.38829014  1.        ]]
```

Word Embedding

- **Word embedding** is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

Word2Vec is one of the most popular technique to learn word embeddings using deep neural network. It was developed by *Tomas Mikolov in 2013 at Google*.

Can a text be input in deep learning?

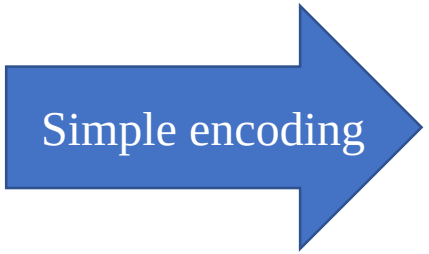
Word Embedding

Can text be input in deep learning? **No!!!**

Encoding

Can number be input in deep learning? **Yes!!!**

Good night!
Good morning!



Unique word	encoding
Good	0
night	1
morning	2

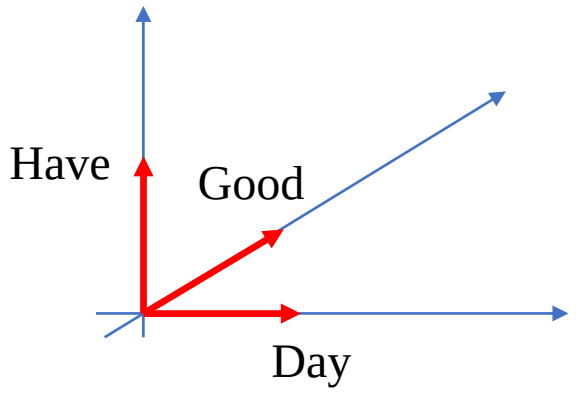
How does Word2Vec work?

Have a good day and Have a great day. ⟶ V = {Have, a, good, great, day}

One-hot encoded vector

Have = [1,0,0,0,0]`; a=[0,1,0,0,0]` ; good=[0,0,1,0,0]` ; great=[0,0,0,1,0]` ; day=[0,0,0,0,1]` (` represents transpose)

If we try to visualize these encodings, we can think of a 5-dimensional space, where each word occupies one of the dimensions and has nothing to do with the rest (no projection along the other dimensions). This means 'good' and 'great' are as different as 'day' and 'have', which is not true.



Have = [1,0,0]

Good = [0,1,0]

Day = [0,0,1]

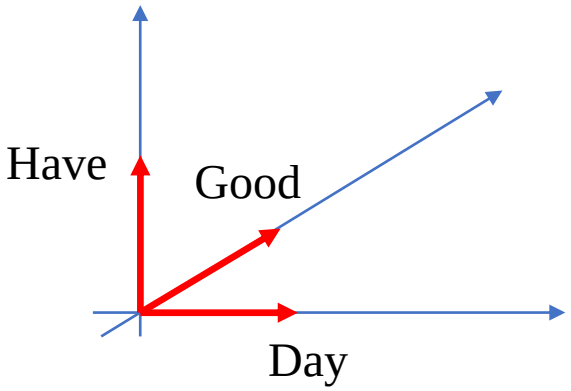
How does Word2Vec work?

Have a good day and Have a great day. ⟶ V = {Have, a, good, great, day}

One-hot encoded vector

Have = [1,0,0,0,0]`; a=[0,1,0,0,0]` ; good=[0,0,1,0,0]` ; great=[0,0,0,1,0]` ; day=[0,0,0,0,1]` (` represents transpose)

If we try to visualize these encodings, we can think of a 5-dimensional space, where each word occupies one of the dimensions and has nothing to do with the rest (no projection along the other dimensions). This means 'good' and 'great' are as different as 'day' and 'have', which is not true.



Have = [1,0,0]

Good = [0,1,0]

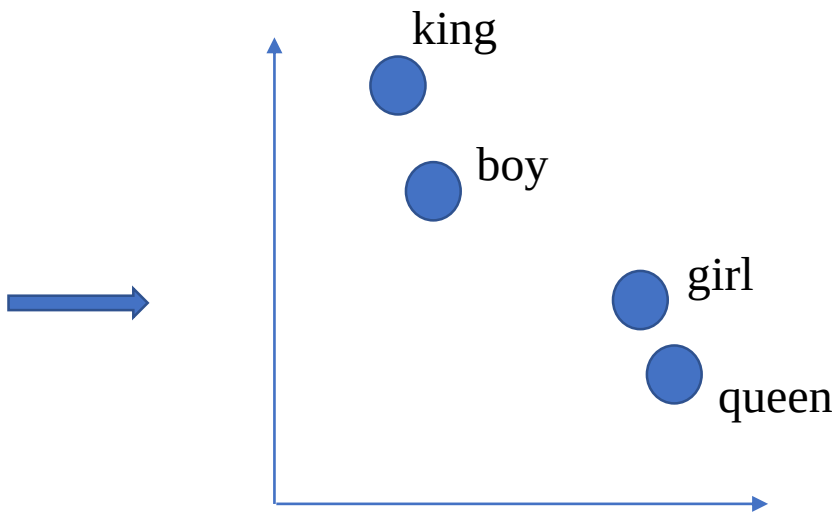
Day = [0,0,1]

Does not show the similarity!

Embedding

Is a dense vector with similarity.

Unique word	encoding	Embedding
king	[1,0,0,0]	[1,2]
queen	[0,1,0,0]	[2,3]
girl	[0,0,1,0]	[1,5]
boy	[0,0,0,1]	[3,4]



Word2vec is word embedding which consider **similarity** comes from neighbor words.

How does Word2Vec work? (Skip-Gram model, window size = 1)

“King brave man”
“Queen intelligent woman”

Word	neighbor
King	brave
brave	king
brave	man
man	brave
Queen	intelligent
intelligent	Queen
intelligent	woman
Woman	intelligent

Train data

How does Word2Vec work? (Skip-Gram model, window size = 2)

“King brave man”
“Queen intelligent woman”

Word	neighbor
King	brave
King	man
brave	king
brave	man
man	brave
man	King
Queen	intelligent
Queen	woman
intelligent	Queen
intelligent	woman
Woman	intelligent
Woman	Queen

Train data

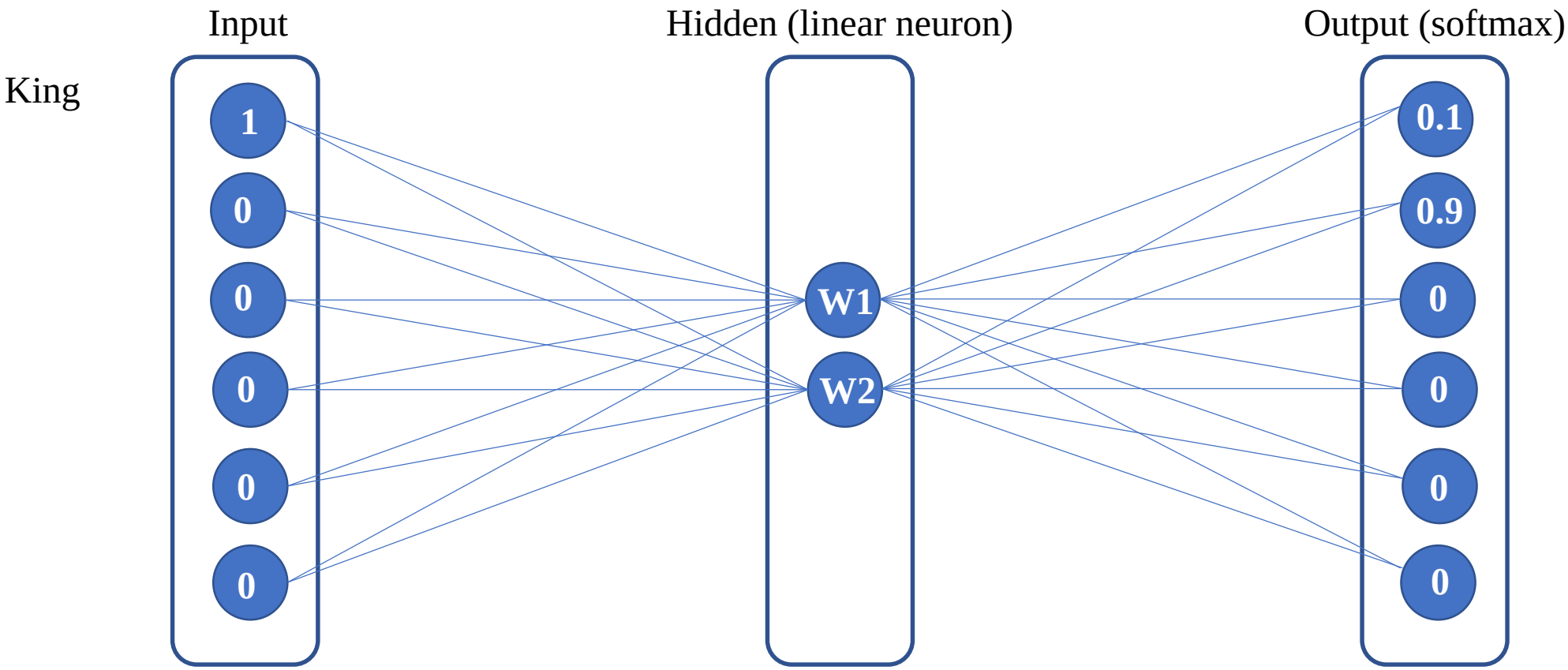
How does Word2Vec work? (Skip-Gram model, window size = 2)

Word	Word one hot encoding	neighbor	Neighbor one hot encoding
King	[1,0,0,0,0,0]	brave	[0,1,0,0,0,0]
King	[1,0,0,0,0,0]	man	[0,0,1,0,0,0]
brave	[0,1,0,0,0,0]	king	[1,0,0,0,0,0]
brave	[0,1,0,0,0,0]	man	[0,0,1,0,0,0]
man	[0,0,1,0,0,0]	brave	[0,1,0,0,0,0]
man	[0,0,1,0,0,0]	King	[1,0,0,0,0,0]
Queen	[0,0,0,1,0,0]	intelligent	[0,0,0,0,1,0]
Queen	[0,0,0,1,0,0]	woman	[0,0,0,0,0,1]
intelligent	[0,0,0,0,1,0]	Queen	[0,0,0,1,0,0]
intelligent	[0,0,0,0,1,0]	woman	[0,0,0,0,0,1]
Woman	[0,0,0,0,0,1]	intelligent	[0,0,0,0,1,0]
Woman	[0,0,0,0,0,1]	Queen	[0,0,0,1,0,0]

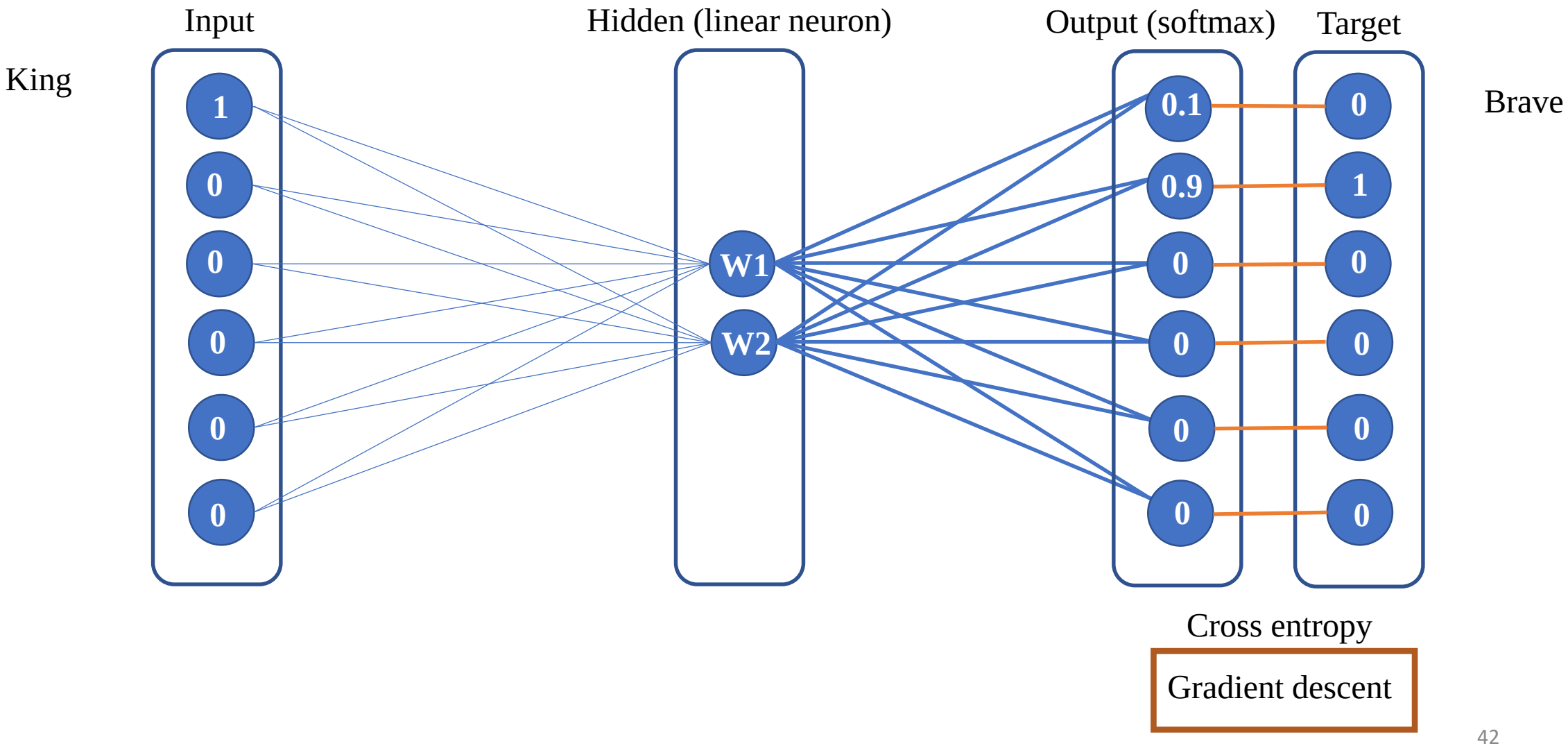
How does Word2Vec work? (Skip-Gram model, window size = 2)

Input (Word one hot encoding)	Target (Neighbor one hot encoding)
[1,0,0,0,0,0]	[0,1,0,0,0,0]
[1,0,0,0,0,0]	[0,0,1,0,0,0]
[0,1,0,0,0,0]	[1,0,0,0,0,0]
[0,1,0,0,0,0]	[0,0,1,0,0,0]
[0,0,1,0,0,0]	[0,1,0,0,0,0]
[0,0,1,0,0,0]	[1,0,0,0,0,0]
[0,0,0,1,0,0]	[0,0,0,0,1,0]
[0,0,0,1,0,0]	[0,0,0,0,0,1]
[0,0,0,0,1,0]	[0,0,0,1,0,0]
[0,0,0,0,1,0]	[0,0,0,0,0,1]
[0,0,0,0,0,1]	[0,0,0,0,1,0]
[0,0,0,0,0,1]	[0,0,0,1,0,0]

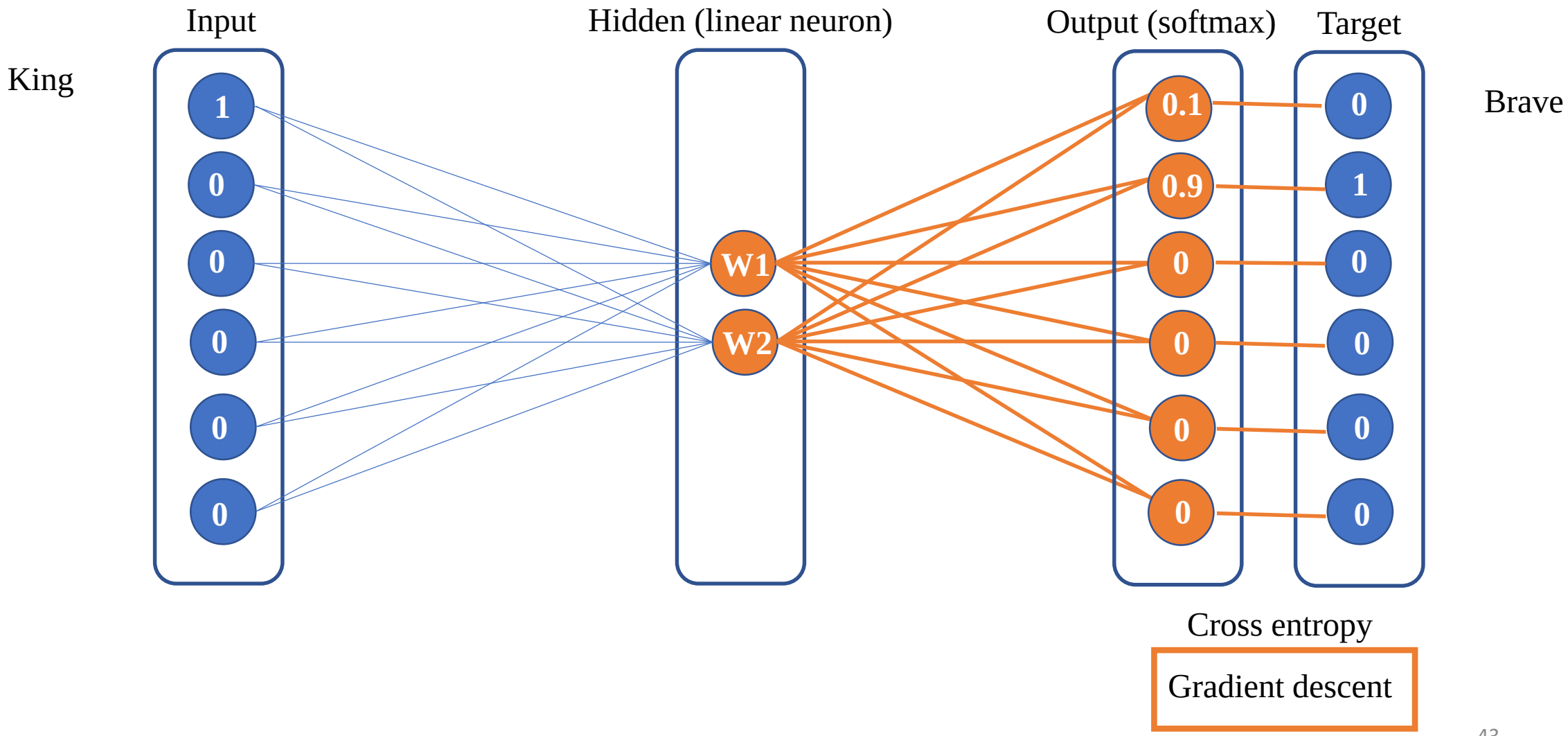
How does Word2Vec work?



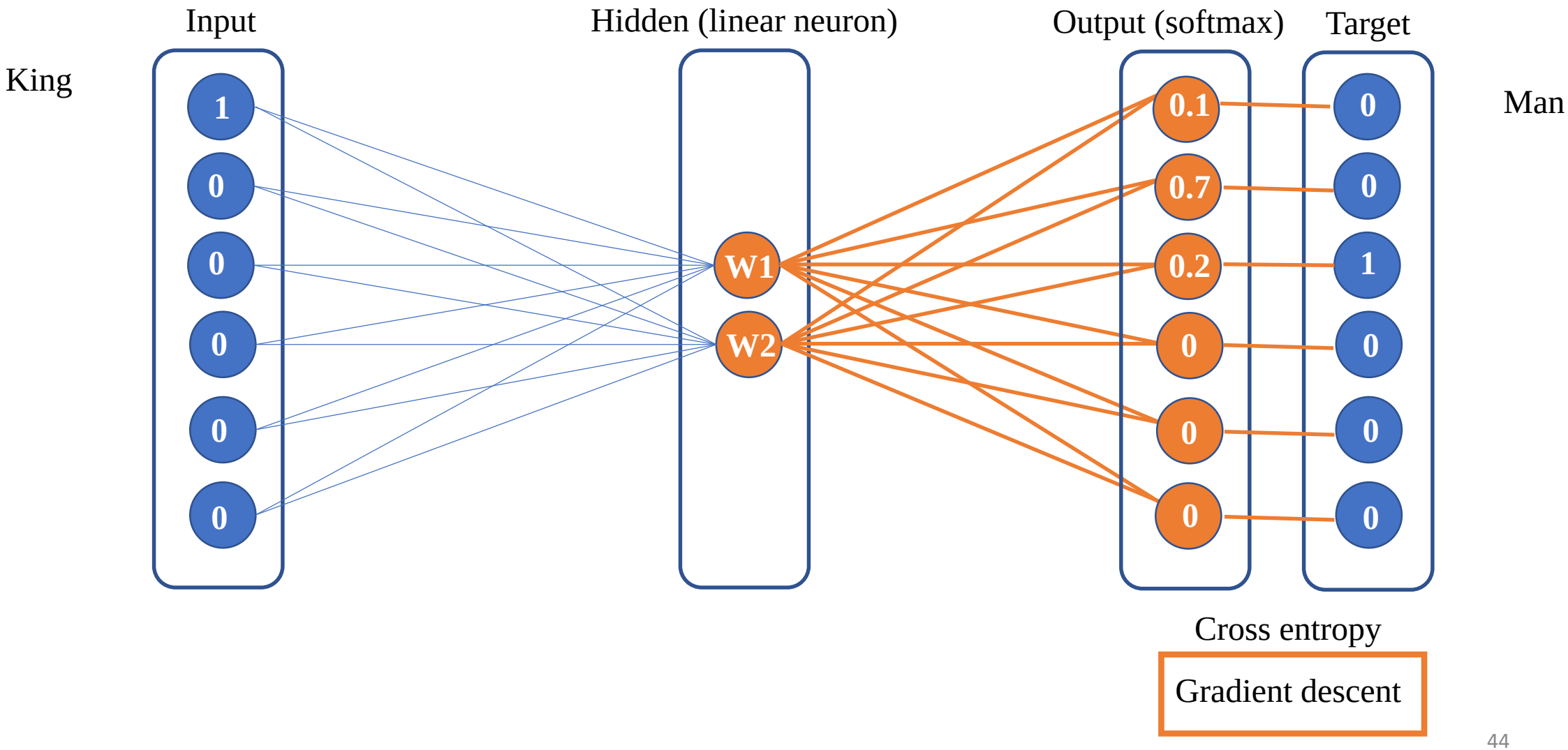
Word2vec training



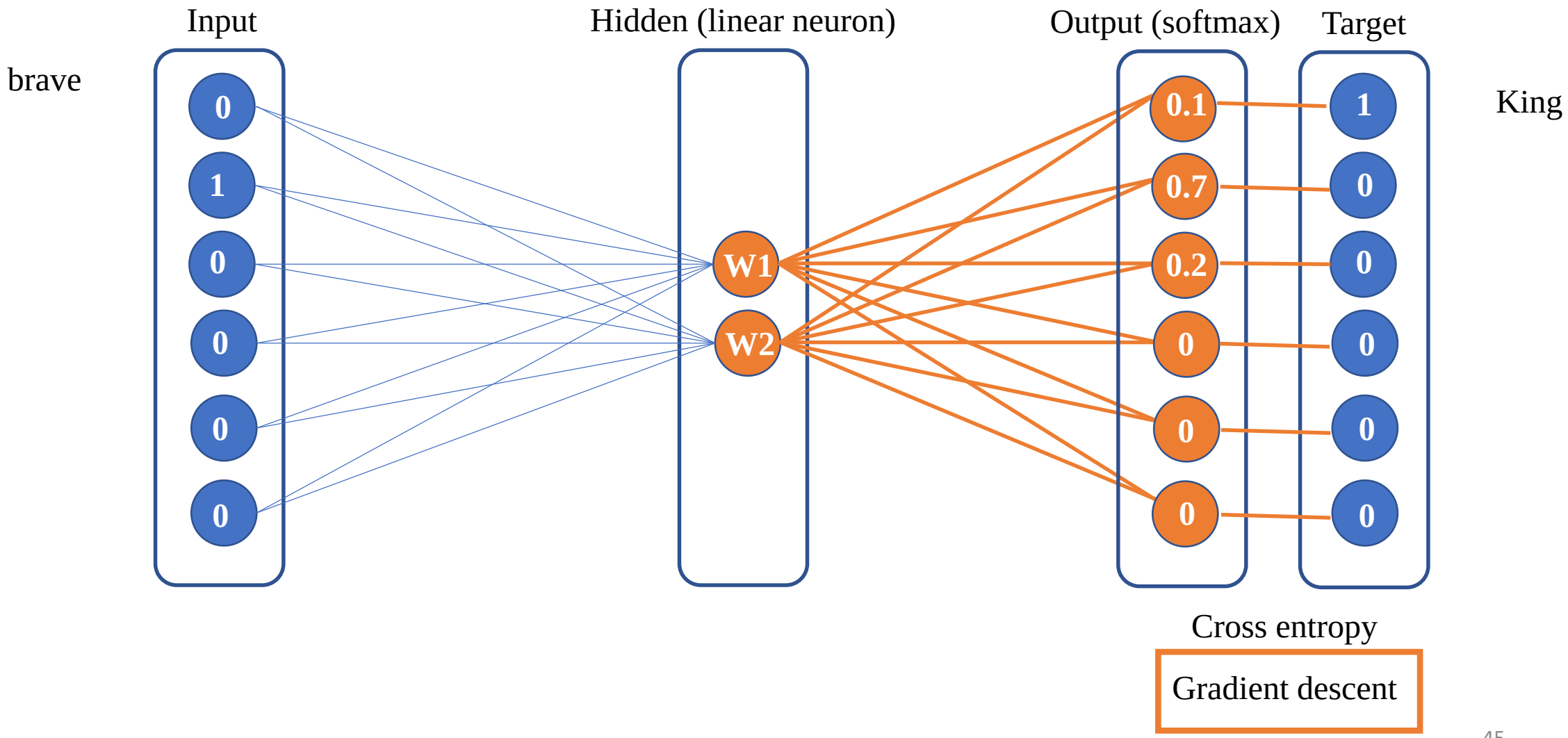
Word2vec training



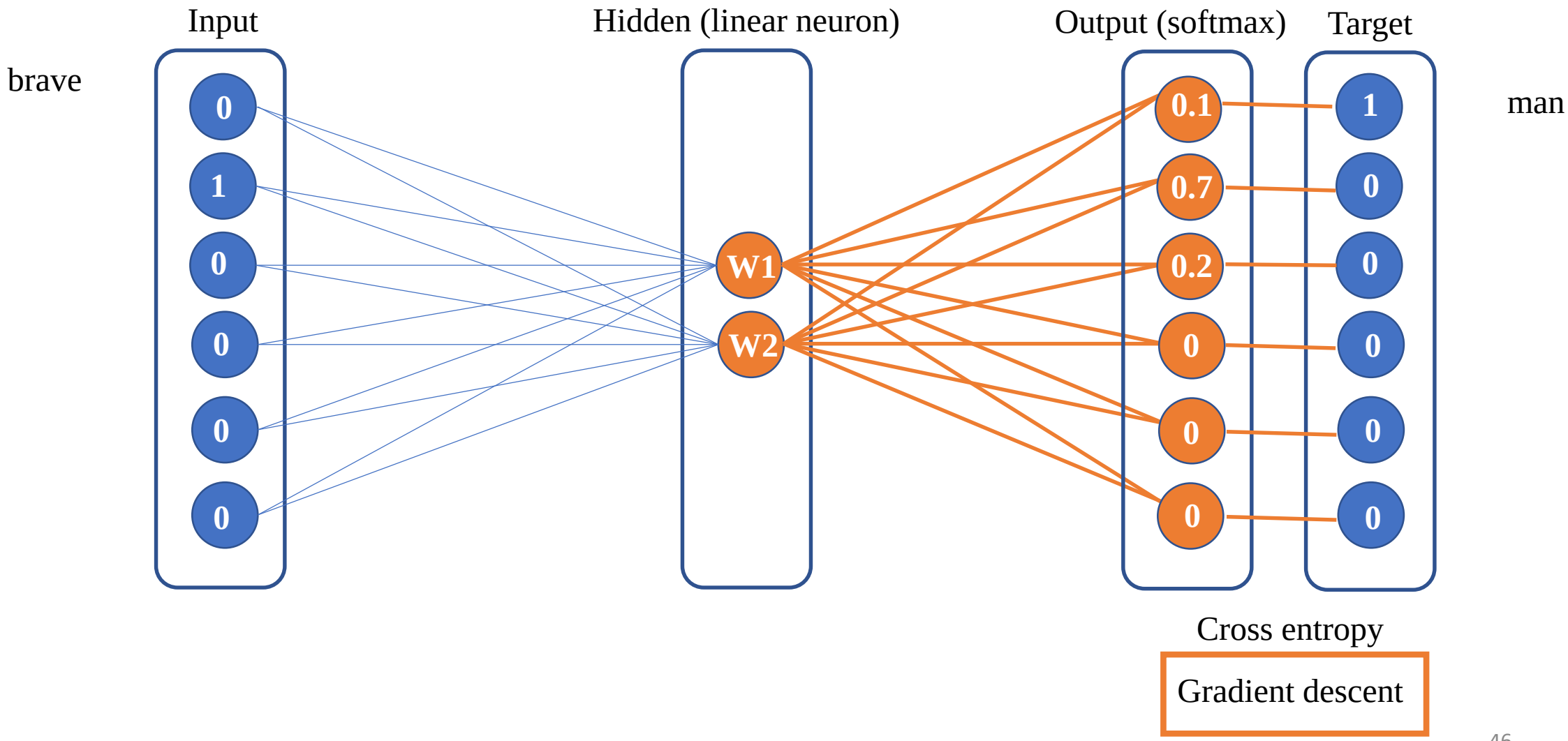
Word2vec training



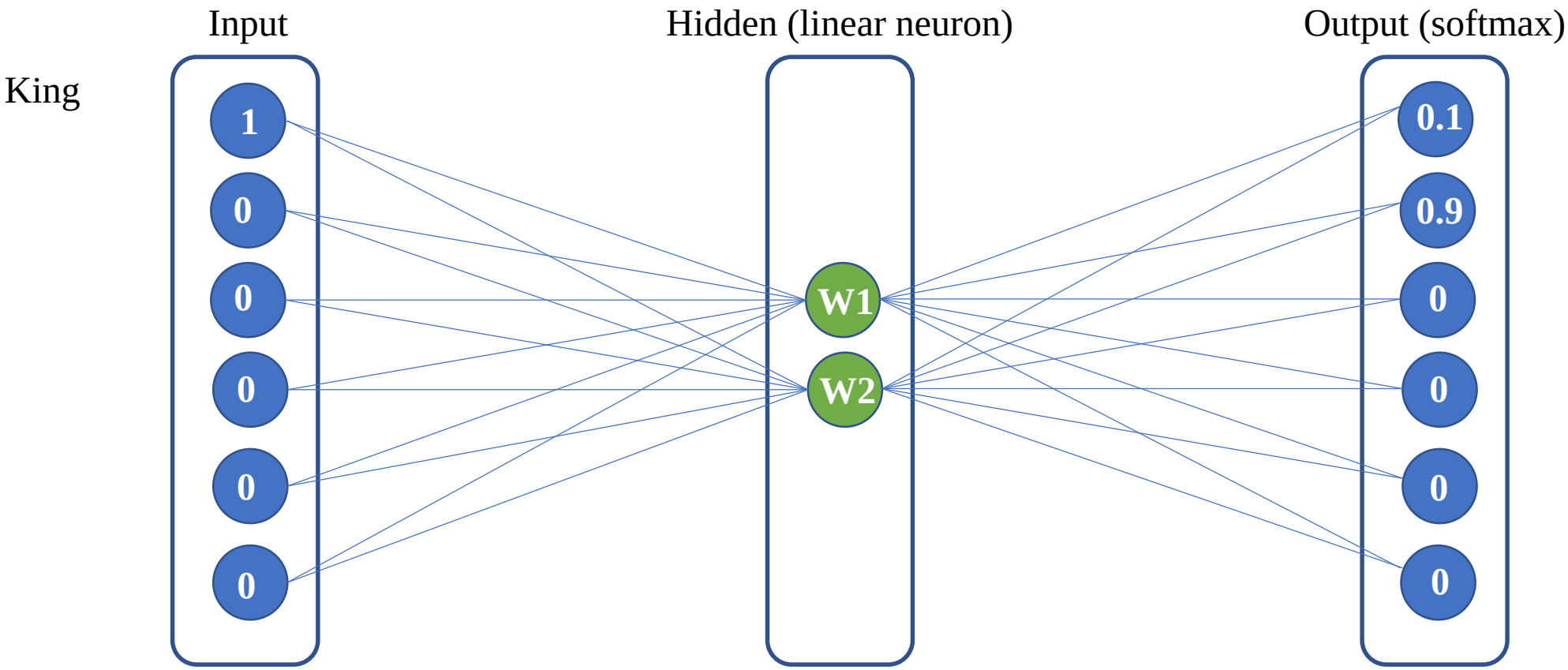
Word2vec training



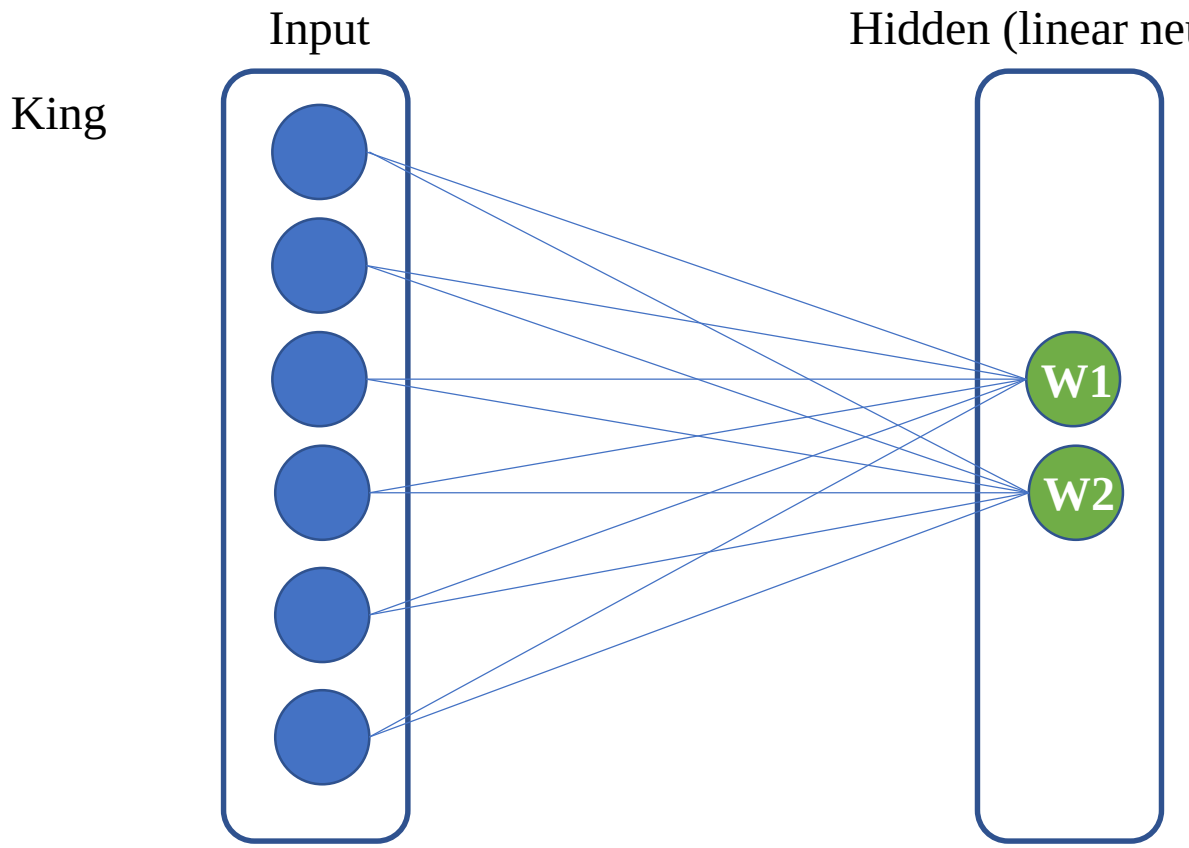
Word2vec training



Hidden layer is our Word2vec!!!



Hidden layer is our Word2vec!!!

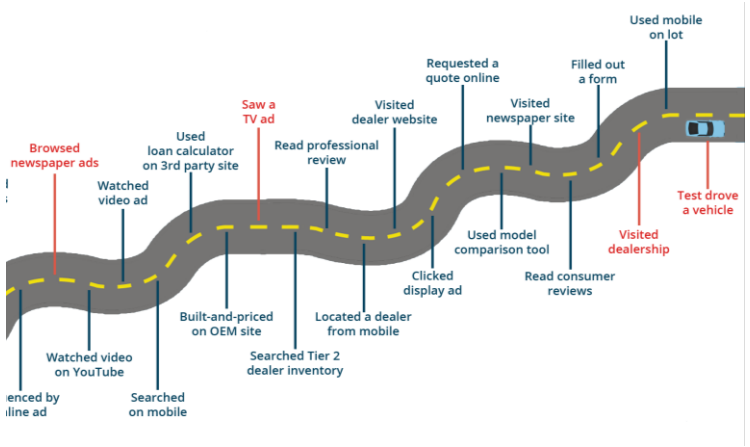


Unique Word	Embedding
King	[1,1]
brave	[1,2]
man	[1,3]
Queen	[5,5]
intelligent	[5,4]
Woman	[5,6]

6-dimensional vector to two-dimensional vector

Text mining and marketing

Purpose:



Touch Points → **Users Profiling** → **Mindset**

Data:

User-ID	URLs	Time Stamp	Postal Codes
2354	https://www.seloger-construire.com	12635	75009 (Paris)
8768	http://www.goal.com/fr/news	12655	75017 (Paris)
2562	https://www.lemonde.fr/culture	12675	54100 (Nancy)
5656	https://www.lemonde.fr/sciences	12676	33200 (Bordeaux)
5656	https://www.instagram.com	13677	35700 (Rennes)
8563	https://portail.lipn.univ-paris13.fr	13700	93430 (Villetaneuse)
3698	https://www.manototv.com/shows	13909	67200 (Strasbourg)
5650	https://www.leboncoin.fr/accessoires	13100	69009 (Lyon)
4447	https://boutique.orange.fr/mobile	13101	75016 (Paris)
4447	https://www.maisonsdumonde.com	13105	75016 (Paris)

Needs:

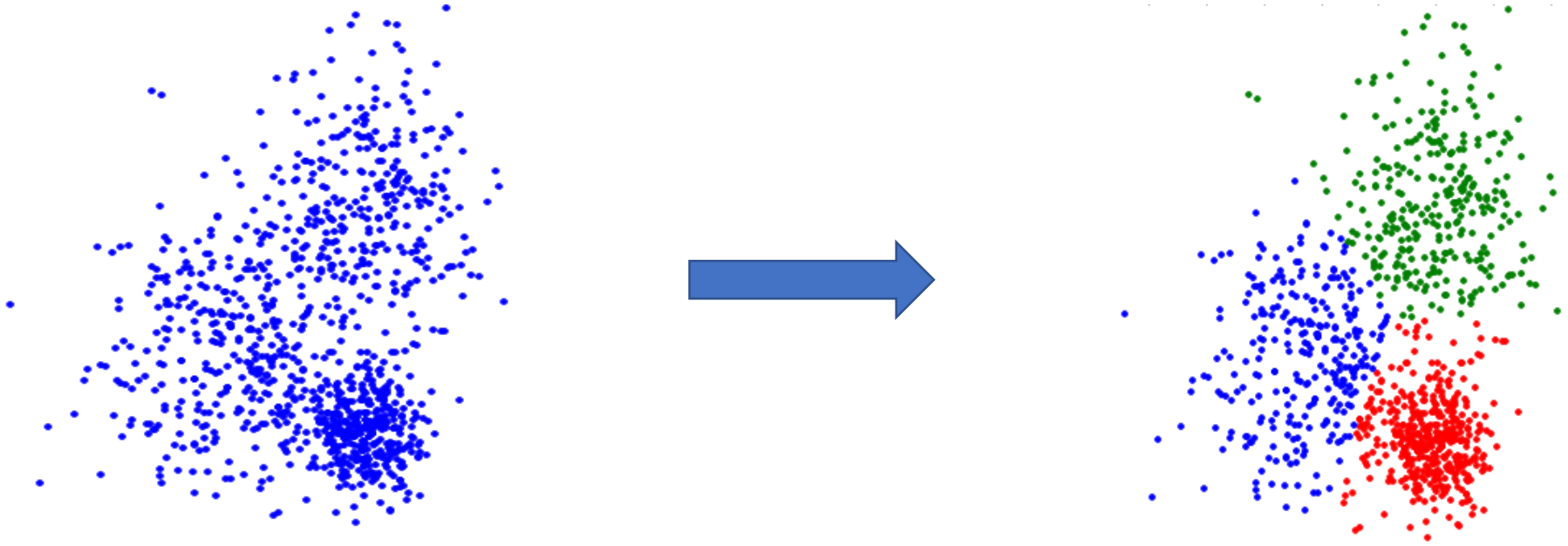
- To deal with very large, complex and dynamic databases.
- To follow the changes in behaviors of the connected users (Concept Drift).
- To detect "mindset" of a user (profiles).

Challenges:

- Must be very fast and have a low memory complexity.
- URLs information are Semantic and Contextual.
- The similarity measure is not necessarily Euclidean so the solution therefore requires an adapted representation space.

Clustering

A cluster is a group of relatively homogeneous data that share more common characteristics between each other than with the data belonging to other clusters.



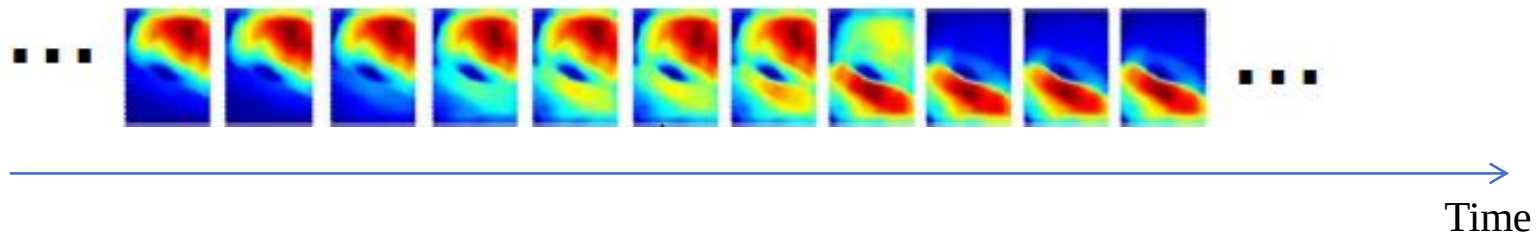
The notion of cluster therefore relies on the notion of similarity and dissimilarity between the data.

Data Stream

A data set with structure changing over time

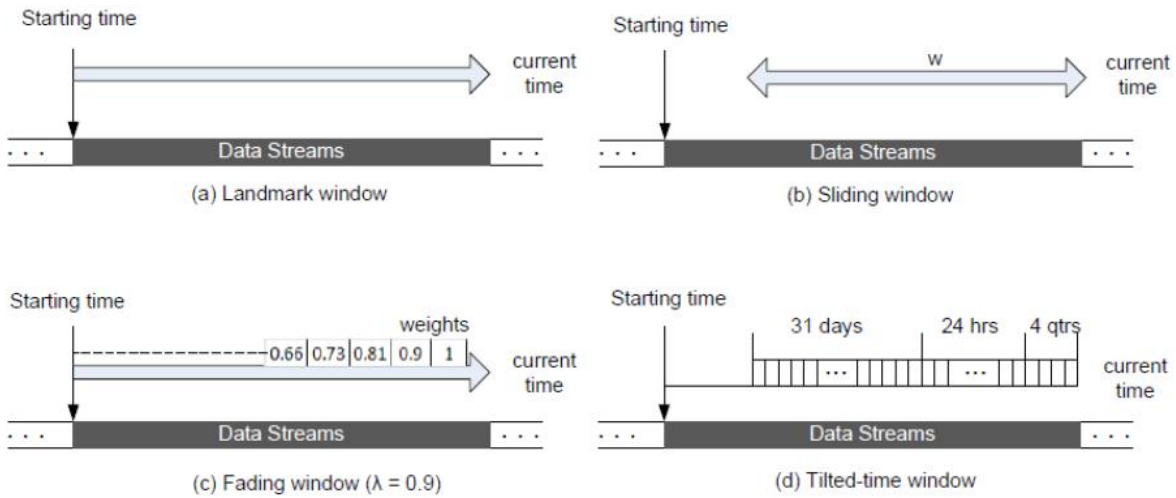
Problematics

- Need low complexity algorithms processing data on the fly
- The detection of "concept drift" is a difficult problem



Data Stream

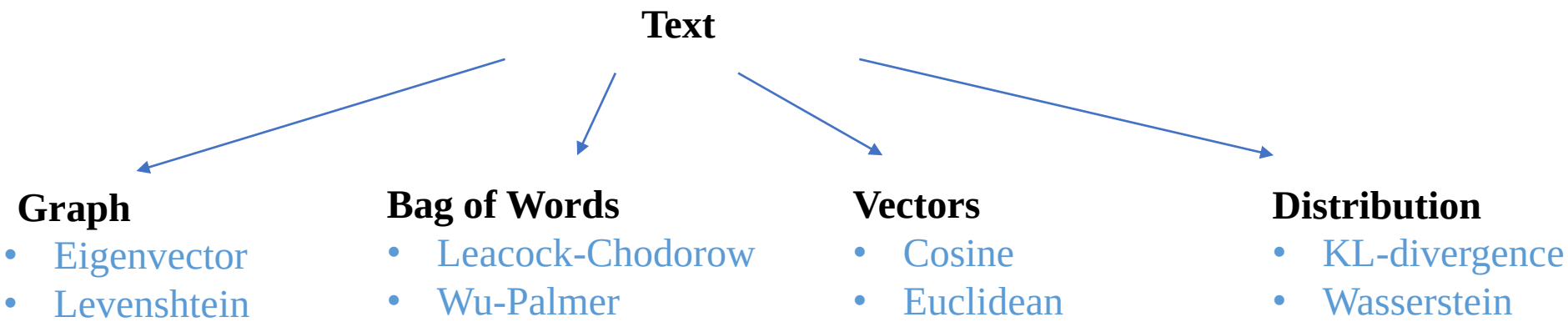
- **Computational Strategies:**
  - Online Learning
  - Two-phase Learning



- **Time Window Models**
  - Landmark window; the entire data stream from starting to the current time instant.
  - Sliding window; the most recent objects are important; the others are eliminated.
  - Fading window; the new objects receive higher weights than old ones.
  - Tilted-time window; is between the fading window and sliding window.

Relational Data

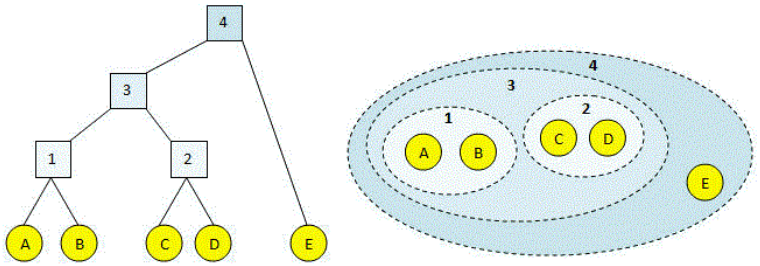
- Relational objects can represent virtually anything:

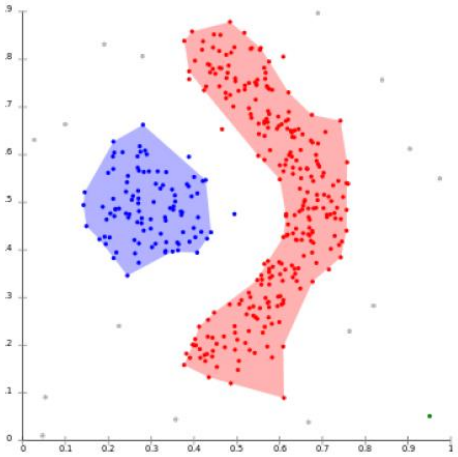


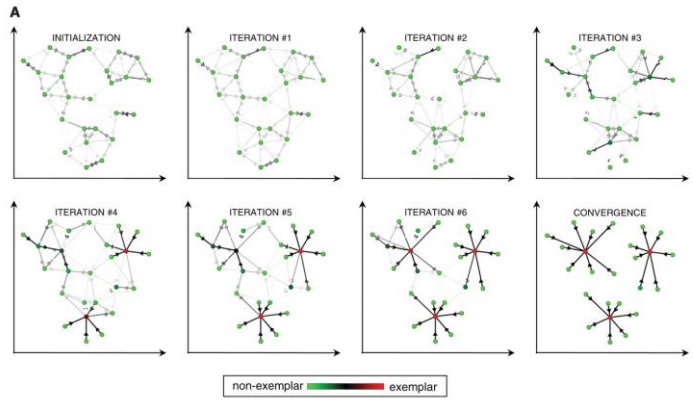
- Defined by their relations : R = [relation($o^{(i)}$, $o^{(j)}$)] with $1 \leq i, j \leq n$.
- The relational matrix often takes the form of a dissimilarity matrix **D**
- A dissimilarity matrix **D** is: square, symmetric, non-negative and hollow

Clustering of Relational Data

- A few are suited for relational data (from a distance matrix):
  - Single-Linkage, Affinity Propagation, Spectral Clustering, DBSCAN…







Single-Linkage DBSCAN Affinity Propagation

- Most prototype-based algorithms are only adapted to vector data.

Clustering of Relational Data

Main difficulties:

- ✓ High complexity (time and memory)

Proposed Solution:

- ✓ Prototype-Based Algorithm Using Barycentric Coordinates

Main difficulties:

- ✓ Not adapted to data streams

Proposed Solution:

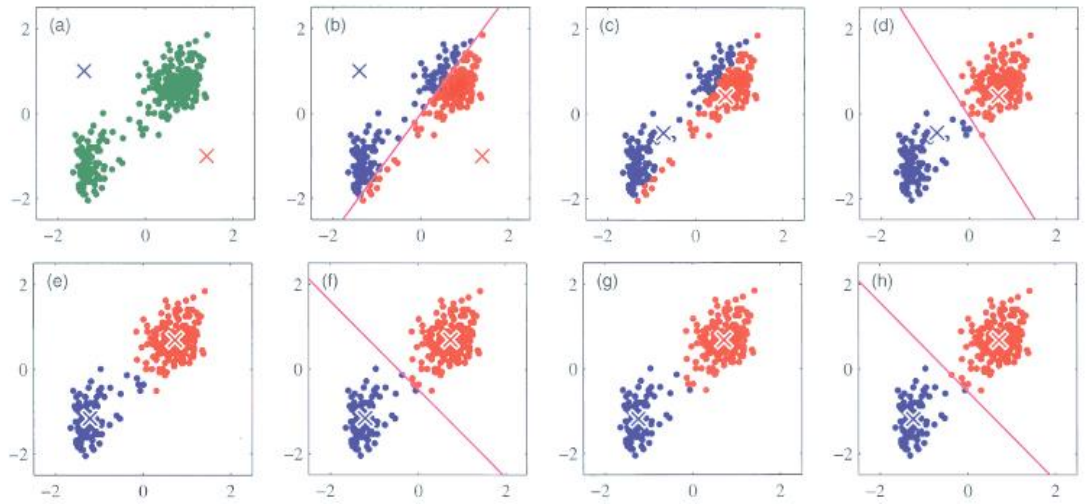
- ✓ New Clustering Algorithm for Data Stream and Relational Data-set

Clustering using Barycentric Coordinates

- Objectives:
  - Clustering of complex data-sets (URLs, Users)

- Solution:
  - Prototype-based clustering

K-Means



Affectation:

$$c^i = \arg\min_j \|x^i - \mu^j\|^2$$

Update:

$$\mu^k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

Complexity: O(NK)

- Number of Observation = N
- Number of cluster = K

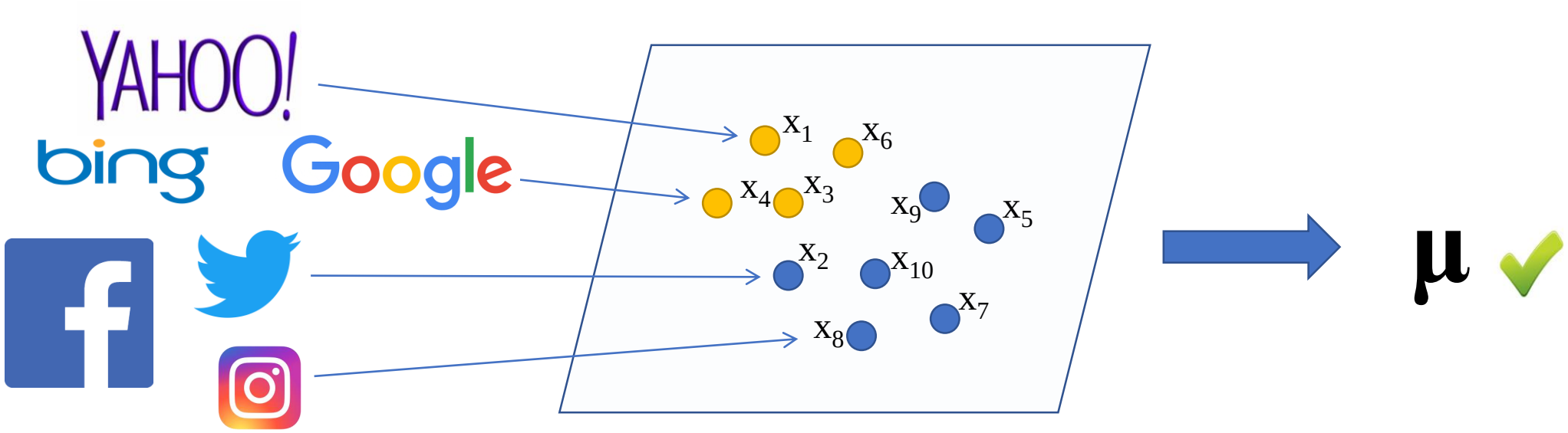
μ^k is the point which minimize the sum square distances

How to compute the prototypes from objects?

w ?

Objects + Distance function **Prototype**

How to compute the prototypes from objects?

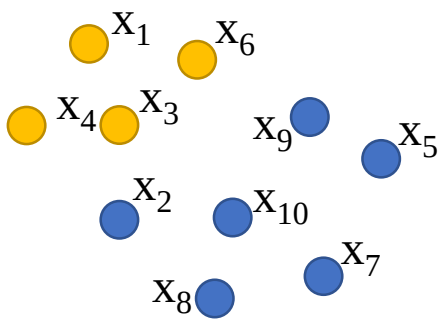


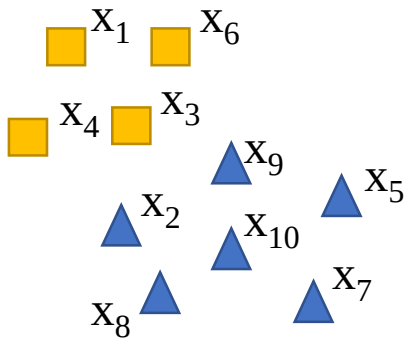
Objects
+
Distance function

Vectorial Representation of objects in unknown space

Prototype

Naive Relational K-Means (Hattaway 1989)





Representation of objects

Computing prototypes from all of objects

$$\text{Prototype } \mu^k = \alpha_1^k \,.\, x_1 + \alpha_2^k \,.\, x_2 + \alpha_3^k \,.\, x_3 + \ldots + \alpha_{N-1}^k \,.\, x_{N-1} + \alpha_N^k \,.\, x_N$$

Naive Relational K-Means (Hattaway 1989)

Affectation:

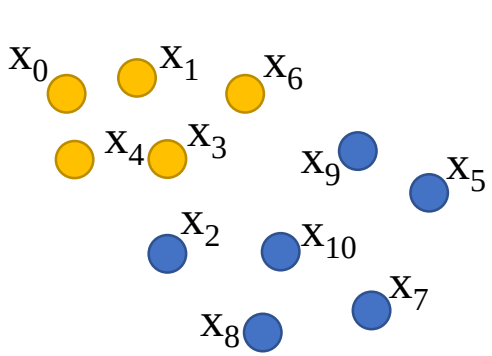
$$d(w^k, o^i) = (D\alpha^k)_i - \frac{1}{2}\alpha^{kT} D\alpha^k$$

Update:

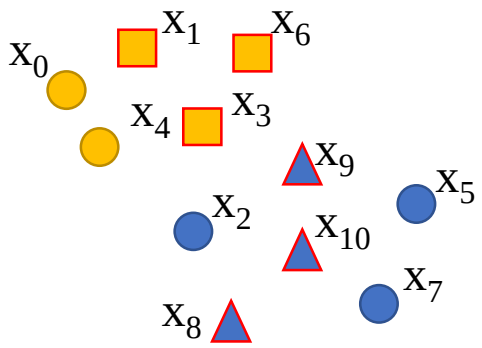
$$\alpha^k = \frac{1}{|C_k|}(\delta_{k,1}, ..., \delta_{k,N}) \qquad \sum_{i=1}^{N} \alpha_i^k = 1$$

Complexity (time) : $O(KN^2)$
Complexity (memory) : $O(N^2)$

Sparse K-Means Relational (Rossi 2007)



Representation of objects



Computing prototypes from support points

Prototype News: $\mu^{News} = \beta^{News}_1 . s^{News}_1 + \beta^{News}_2 . s^{News}_2 + \beta^{News}_3 . s^{News}_3$

Prototype Sport: $\mu^{Sport} = \beta^{Sport}_1 . s^{Sport}_1 + \beta^{Sport}_2 . s^{Sport}_2 + \beta^{Sport}_3 . s^{Sport}_3$

$$\sum_{p=1}^{P} \beta_p^k = 1$$

Complexity (time): O(KNP)
Complexity (memory): $O(N^2)$

Sparse K-Means Relational (Rossi 2007)

Affectation:

$$d(w^k, o^i) = (D\alpha^k)_i - \frac{1}{2}\alpha^{kT} D\alpha^k \qquad \sum_{i=1}^{N} \alpha_i^k = 1$$

Update:

Minimize $\longrightarrow \sum_{o^i \in C_k} d(w^k, o^i) \longrightarrow \sum_{o^i \in C_k} (D\alpha^k)_i - \frac{1}{2}\alpha^{kT} D\alpha^k$

$$\nabla L_k = s_{k,J} - |C_k| D_{J_k} \beta^k + \lambda \vec{1} = 0$$

$$\sum_{p=1}^{P} \beta_p^k = 1$$

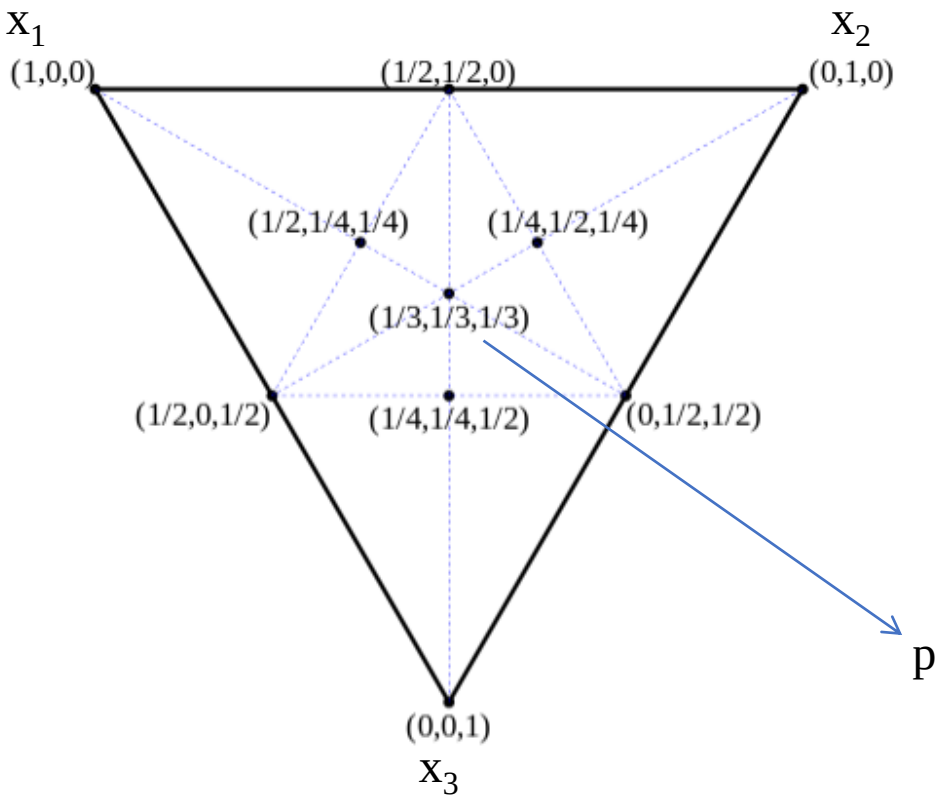
$$\longrightarrow \begin{bmatrix} |C_k| D_{J_k} & -\vec{1} \\ \vec{1}^T & 0 \end{bmatrix} \begin{bmatrix} \beta^k \\ \lambda \end{bmatrix} = \begin{bmatrix} s_{k,J} \\ 1 \end{bmatrix}$$

Relational Clustering Based on Barycentric Coordinate System

Objectives:

- Incremental clustering
- No need to store the whole data set in memory
- Use barycentric coordinate formalism

Barycentric Coordinate System



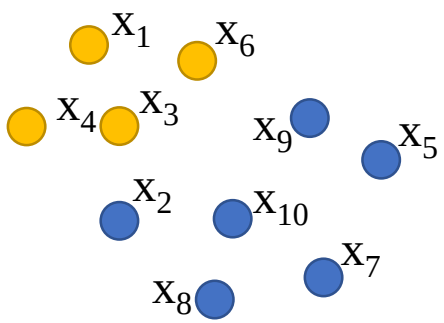
$$p = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

$$\sum a = 1$$

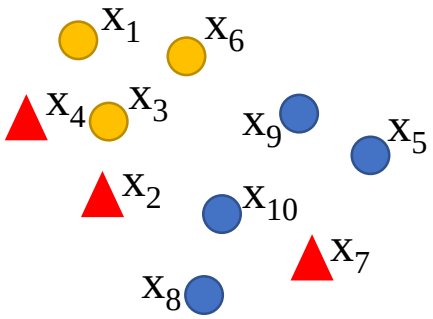
$$a_{x_1} = (1, 0, 0, \ldots, 0)$$
$$a_{x_2} = (0, 1, 0, \ldots, 0)$$
$$a_{x_3} = (0, 0, 1, \ldots, 0)$$
$$\vdots$$
$$a_{x_n} = (0, 0, 0, \ldots, 1)$$

- The location of a point of a simplex (ex. triangle) is specified as the center of mass, or barycenter, of usually unequal masses placed at its vertices.

Relational Clustering
Based on Barycentric Coordinate System



Representation of objects



Computing prototypes from fixed support points

- **Fixed and independent**

$$\text{Prototype News: } \mu^{\text{News}} = \beta^{\text{News}}{}_1 . \mathbf{s_1} + \beta^{\text{News}}{}_2 . \mathbf{s_2} + \beta^{\text{News}}{}_3 . \mathbf{s_3}$$

$$\text{Prototype Sport: } \mu^{\text{Sport}} = \beta^{\text{Sport}}{}_1 . \mathbf{s_1} + \beta^{\text{Sport}}{}_2 . \mathbf{s_2} + \beta^{\text{Sport}}{}_3 . \mathbf{s_3}$$

$$\text{Object } O^i : x^i = \beta_1^i . \mathbf{s_1} + \beta_2^i . \mathbf{s_2} + \beta_3^i . \mathbf{s_3}$$

Relational Clustering
Based on Barycentric Coordinate System

Barycentric coordinates of o^i:

$$A \cdot \beta^i = J^i \Rightarrow \beta^i = A^{-1} \cdot J^i$$

$$A = \begin{pmatrix} d(s^1,s^1)\text{-}d(s^2,s^1) & \dots & d(s^1,s^P)\text{-}d(s^2,s^P) \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ d(s^1,s^1)\text{-}d(s^P,s^1) & \dots & d(s^1,s^P)\text{-}d(s^P,s^P) \\ \\ 1 & \dots & 1 \end{pmatrix}, \quad J^i = \begin{pmatrix} d(o^i,s^1)\text{-}d(o^i,s^2) \\ . \\ . \\ . \\ d(o^i,s^1)\text{-}d(o^i,s^P) \\ 1 \end{pmatrix}$$

Relational Clustering
Based on Barycentric Coordinate System

Affectation:

$$d\ (o^i, w^k) = -\frac{1}{2}(\beta^i - \beta^k)^T \cdot D_S \cdot (\beta^i - \beta^k)$$

Update:

$$\frac{\delta \sum_{o^i \in C_k} d(o^i, w^k)}{\delta \beta^k} = -D_S \sum_{i|o^i \in C_k} (\beta^i - \beta^k) = 0 \qquad \sum_{p=1}^{P} \beta_p^k = 1$$

Batch

$$\beta^k = \frac{1}{|C_k|} \sum_{i|o^i \in C_k} \beta^i$$

Stochastic

$$\beta^k{}_{t+1} = \beta_t^k + \gamma(\beta^i - \beta_t^k)$$

Gradient descent

Data-sets Description

data set	# Objects	# Type	# Classes
Art1	10000	Vector (artificial)	4
Art2	1000	Vector (artificial)	4
Art3	800	Vector (artificial)	5
Iris	150	Vector (real)	3
Digits	1797	Vector (real)	10
Wine	178	Vector (real)	3
Prot1	115	Sequence	3
Prot4	129	Sequence	5
Prot5	98	Sequence	3
Hist-mean	1000	Distribution	5
Hist-shape	1000	Distribution	5
Hist-std	1000	Distribution	5
People	300	Concept	3

Data-set People

Quality

ARI	S=1	S=2	S=3	S=5	S=10	S=20	S=100
Art1	0.00	0.76	0.96	0.98	0.98	0.98	0.98
Art2	0.00	0.89	0.85	1.00	1.00	1.00	1.00
Art3	0.00	0.57	0.93	0.96	0.95	0.67	0.67
Iris	0.00	0.87	0.85	0.79	0.77	0.75	0.79
Digits	0.00	0.17	0.28	0.41	0.62	0.57	0.61
Wine	0.00	0.37	0.36	0.36	0.35	0.36	0.36
Prot1	0.00	0.98	0.98	0.98	0.98	0.98	0.98
Prot2	0.00	0.94	0.94	0.94	0.94	0.94	0.94
Prot3	0.00	0.94	0.94	0.94	0.94	0.94	0.94
Hist-mean	0.00	0.97	1.00	1.00	1.00	1.00	1.00
Hist-shape	0.00	0.31	0.52	0.72	0.72	0.75	0.76
Hist-std	0.00	0.60	1.00	1.00	1.00	1.00	1.00
People	0.00	0.56	0.76	0.93	0.94	0.95	0.95

Silhouette	S=1	S=2	S=3	S=5	S=10	S=20	S=100
Art1	-1.00	0.52	0.55	0.56	0.56	0.56	0.56
Art2	-1.00	0.66	0.69	0.79	0.79	0.79	0.79
Art3	-1.00	0.44	0.63	0.63	0.65	0.62	0.63
Iris	-1.00	0.55	0.55	0.55	0.56	0.56	0.55
Digits	-1.00	0.01	0.05	0.11	0.16	0.18	0.18
Wine	-1.00	0.52	0.52	0.52	0.52	0.52	0.52
Prot1	-1.00	0.96	0.96	0.96	0.96	0.96	0.96
Prot2	-1.00	0.95	0.95	0.95	0.83	0.95	0.95
Prot3	-1.00	0.99	0.99	0.99	0.99	0.99	0.99
Hist-mean	-1.00	0.62	0.64	0.64	0.64	0.64	0.64
Hist-shape	-1.00	0.12	0.18	0.24	0.25	0.25	0.26
Hist-std	-1.00	0.30	0.65	0.65	0.65	0.65	0.65
People	-1.00	0.27	0.33	0.34	0.27	0.25	0.34

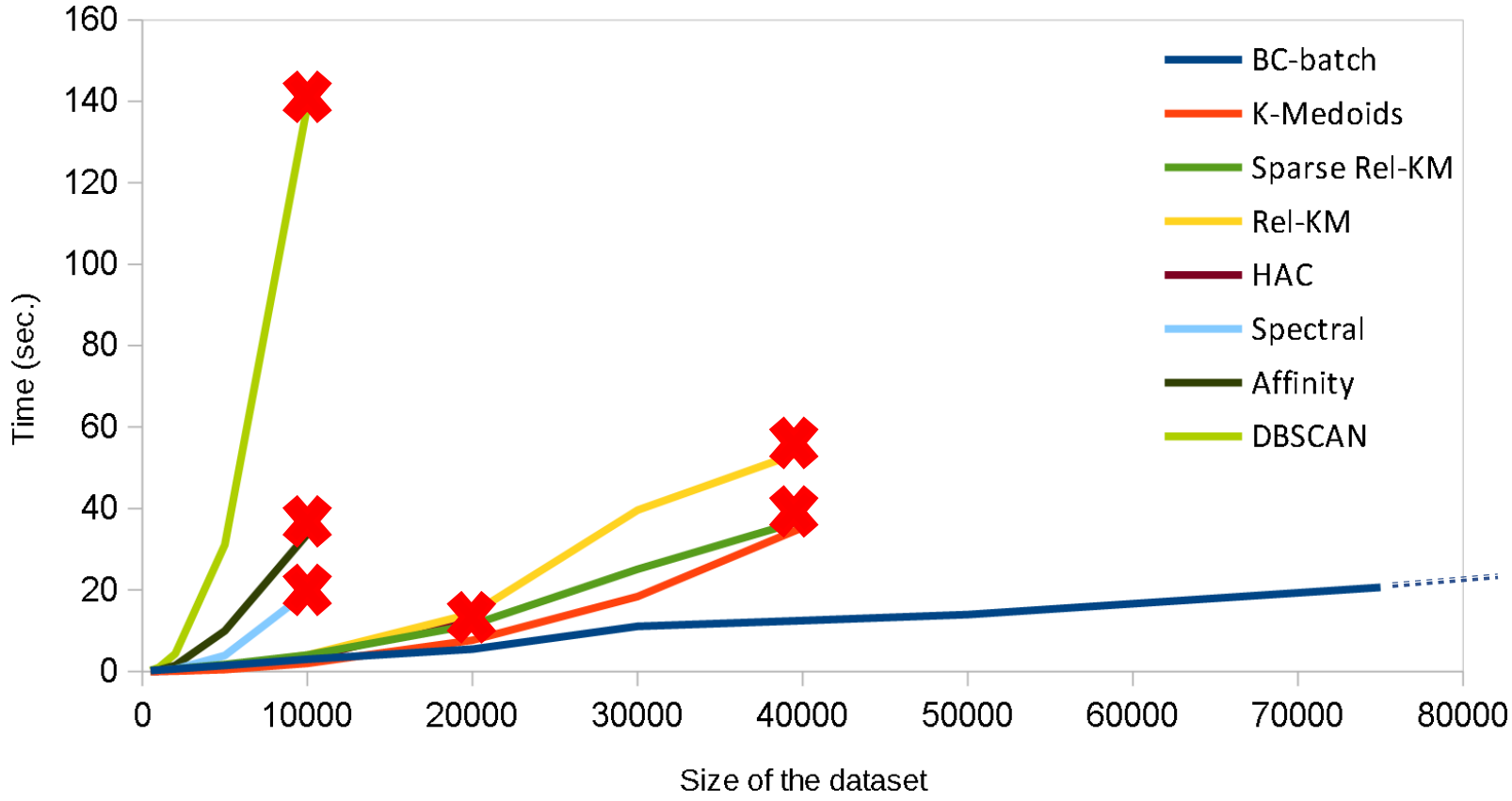
- Increasing the number of support points usually does not have a major effect on the results' internal and external quality.
- Consequently our algorithm does not need a high number of support points to work; this reduce furthermore the complexity of the approach.

Quality

	HAC	Affinity	HDBSCAN	Spectral	KMed	Rel-KM	S-Rel-KM	BC-batch	BC-stoch
Art1	0.86	0.56	0.84	0.97	0.97	0.97	0.97	0.97	0.97
Art2	1.00	0.74	1.00	0.89	1.00	1.00	1.00	1.00	1.00
Art3	0.94	0.63	0.98	0.80	0.87	0.83	0.86	0.92	0.92
Iris	0.74	0.47	0.76	0.75	0.79	0.76	0.80	0.80	0.80
Digits	0.04	0.48	0.70	0.69	0.64	0.73	0.70	0.69	0.65
Wine	0.11	0.16	0.09	0.41	0.42	0.39	0.32	0.40	0.41
Prot1	0.96	0.50	0.90	0.96	0.96	0.96	0.96	0.96	0.96
Prot2	0.86	0.77	0.86	0.86	0.86	0.86	0.86	0.81	0.86
Prot3	0.92	0.38	0.78	0.92	0.92	0.92	0.92	0.92	0.92
Hist-mean	1.00	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hist-shape	0.03	0.19	0.38	0.73	0.76	0.75	0.70	0.77	0.80
Hist-std	1.00	0.46	1.00	1.00	1.00	1.00	1.00	1.00	1.00
People	0.02	0.00	0.64	0.57	0.79	0.70	0.70	0.91	0.94

- The internal and external qualities of our algorithms (BC-batch & BC-stoch) are at least **as good as** the competitors on the experimental data sets for the three indexes tested (here only NMI is shown).

Computational Time



- The proposed algorithm can deal with massive data-sets
- The K-medoid algorithm is very fast when the number of observations is low. But with the increase of the number of observations the computational time increase much faster than in our approach.

Selection of the Support Points

- **How ?**
  - Randomly
  - K-Means++
  - In function of order

Cayley-Menger determinants: $CM_{D_S} = \begin{bmatrix} D_S & \vec{1} \\ \vec{1}^T & 0 \end{bmatrix}$

- **How many?**
  - Dimension +1 (usually unknown)
  - Random Projection

$$(1-\epsilon)\|i-j\|_2 \leq \|f(i)-f(j)\|_2 \leq (1+\epsilon)\|i-j\|_2 \qquad d' \geq \frac{4}{\mathbf{1/2\epsilon^2 - 1/3\epsilon^3}}.log(N)$$

  - In function of speed

Barycentric Coordinates Algorithm For Data Stream

Objectives:

- A fast online Algorithm with low complexity
- Deal with complex data and large volumes of data
- Deal with the evolution of the structure of data (Concept Drift)

Algorithm; Step 1

Projection of object into Barycentric Coordinate space

ex. www.Google.fr

New object o^i

$$A = \begin{pmatrix} d(\mathbf{s}^1,\mathbf{s}^1)\text{-}d(\mathbf{s}^2,\mathbf{s}^1) & \dots & d(\mathbf{s}^1,\mathbf{s}^P)\text{-}d(\mathbf{s}^2,\mathbf{s}^P) \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ d(\mathbf{s}^1,\mathbf{s}^1)\text{-}d(\mathbf{s}^P,\mathbf{s}^1) & \dots & d(\mathbf{s}^1,\mathbf{s}^P)\text{-}d(\mathbf{s}^P,\mathbf{s}^P) \\ 1 & \dots & 1 \end{pmatrix}, \quad J^i = \begin{pmatrix} d(o^i,\mathbf{s}^1)\text{-}d(o^i,\mathbf{s}^2) \\ . \\ . \\ . \\ d(o^i,\mathbf{s}^1)\text{-}d(o^i,\mathbf{s}^P) \\ 1 \end{pmatrix}$$

$$A \cdot \boldsymbol{\beta}^i = J^i \Rightarrow \boldsymbol{\beta}^i = A^{-1} \cdot J^i$$

Compute Matrix **A** once and $\mathbf{J^i}$ for each o^i

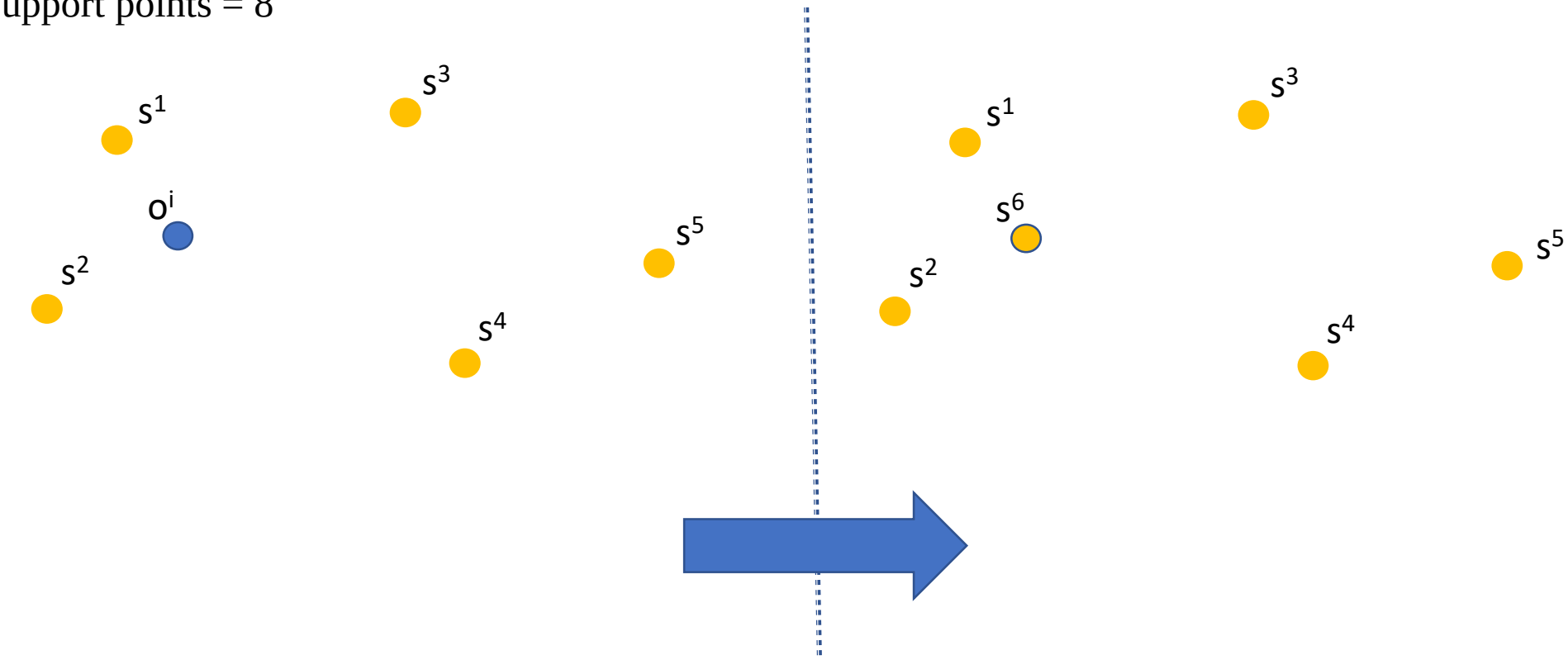
Compute the distance between object o^i and prototype μ^k :

$$d(o^i,\boldsymbol{\mu}^k) = -\frac{1}{2}(\boldsymbol{\beta}^i - \boldsymbol{\beta}^k)^T \cdot D_S \cdot (\boldsymbol{\beta}^i - \boldsymbol{\beta}^k)$$

Algorithm; Step 2

Support Point
New Object

ex. Max Support points = 8



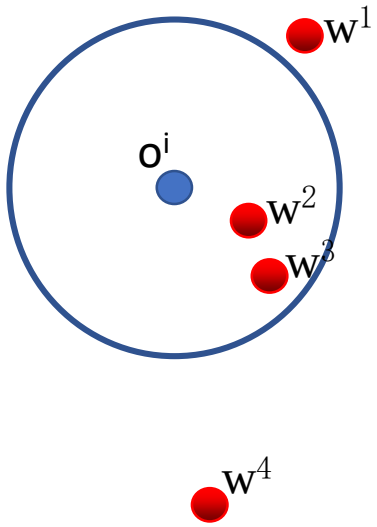
If Support point numbers < Max support points:
Add this new object o^i in support point list

Algorithm; Step 3

Prototype w^k

Object o^i

Case 1

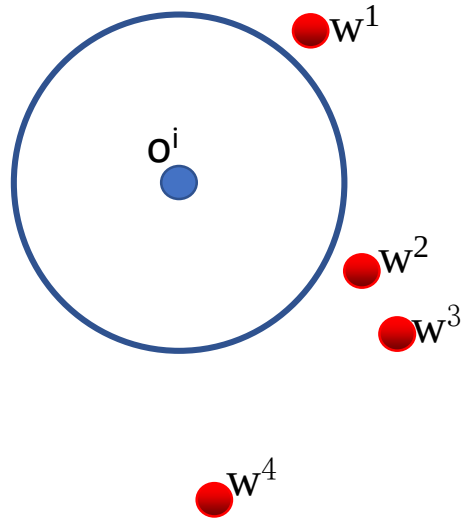


If Distance(o^i,w^k) < Max Radius:

Assign o^i to the closest prototype;

Update the prototype and set the age to zero

Case 2



If Distance(o^i,w^k) > Max Radius:

Create a new prototype with $\beta^k = \beta^i$

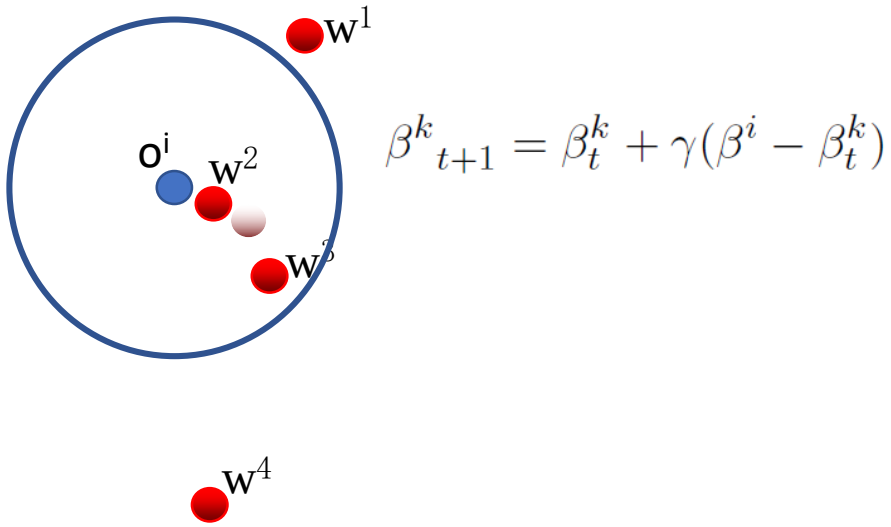
Set the age to zero

Algorithm; Step 3

Prototype w^k

Object o^i

Case 1



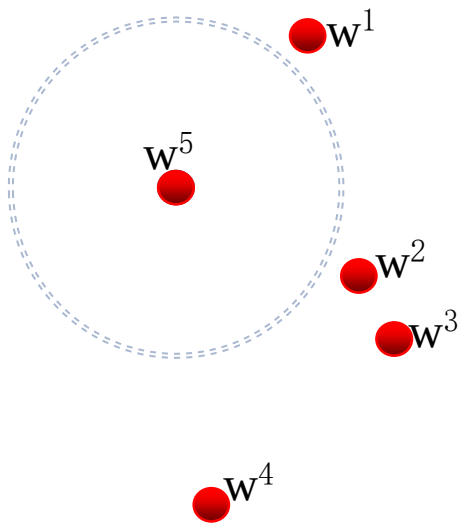
$$\beta^k{}_{t+1} = \beta^k_t + \gamma(\beta^i - \beta^k_t)$$

If Distance(o^i,w^k) < Max Radius:

Assign o^i to the closest prototype;

Update the prototype and set the age to zero

Case 2



If Distance(o^i,w^k) > Max Radius:

Create a new prototype with $\beta^k = \beta^i$

Set the age to zero

Algorithm; Step 4

ex. timeStamp of o_t - timeStamp of o_{t-1} = 3

Prototype w^k

$w^1 = 5 \rightarrow 8$

$w^5 = 0$

$w^2 = 3 \rightarrow 6$

$w^3 = 4 \rightarrow 7$

$w^4 = 9 \rightarrow 12$

$N^k = N^k + 1$ for cluster with data

$N^k = 1$ for cluster without data

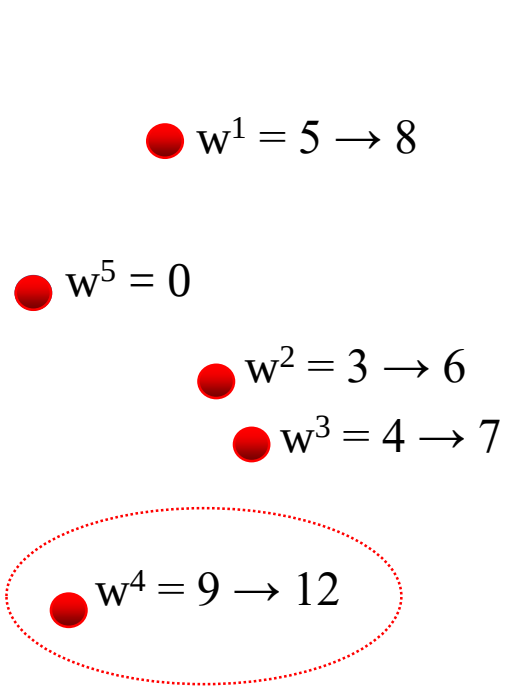
$N^k = N^k \times \epsilon^{(\text{timeStamp of } o_t - \text{timeStamp of } o_{t-1})}$

$\epsilon^{\text{window size}} = 0.01$

Age of all prototypes += (timeStamp of o_t - timeStamp of o_{t-1})

Algorithm; Step 5

ex. Max Age (window size) = 10



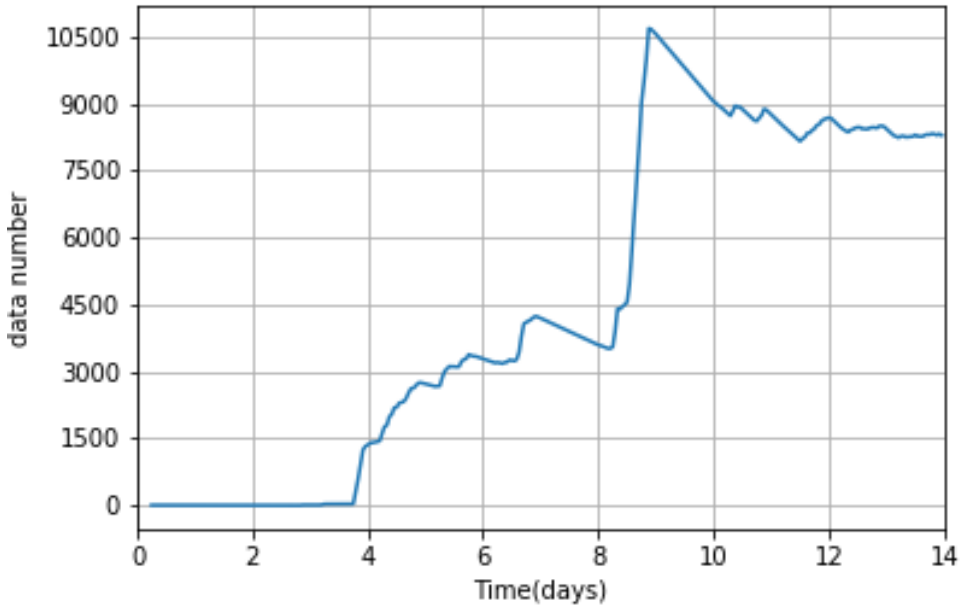
For each prototype w^k:

If Age of w^k > window size

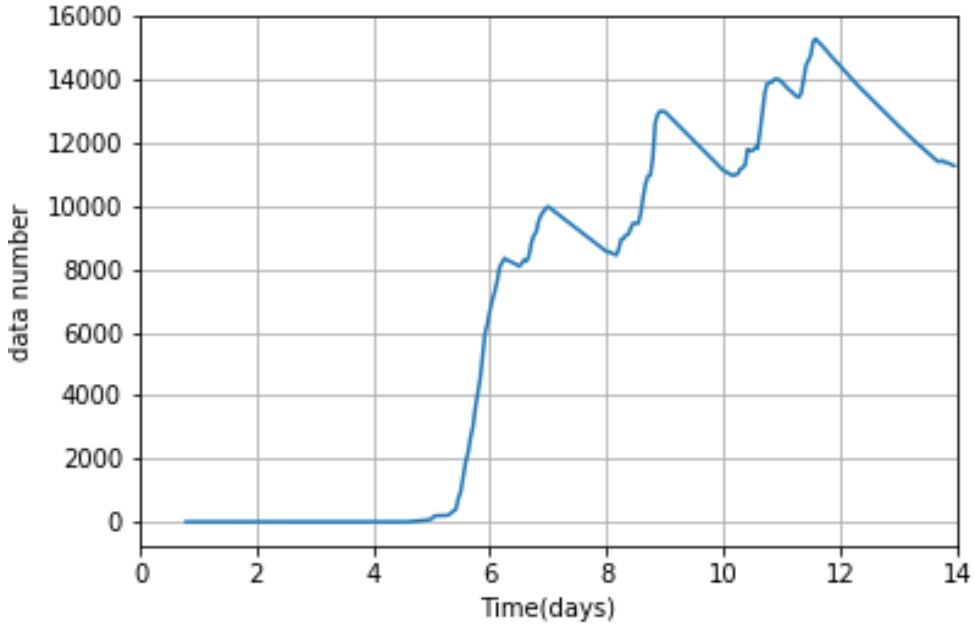
Remove the prototype w^k

Change Detection

First time Stamp : 22/8/2017 at 2:00:08 , Two last weeks of August



a. Prototype 113 : Maxifoot.fr
(world cup qualification for France)



b. Prototype 384 : opodo.fr
(Purchasing the tickets; begining of the scholar year)

Clustering of Urls

Cluster number	URLs	Label	Concept
1	lacentrale.fr/occasion-voiture-marque-mini.html lacentrale.fr/occasion-voiture-modele-porsche-911.html lacentrale.fr/occasion-voiture-marque-dacia.html lacentrale.fr/occasion-moto-marque-triumph-9.html lacentrale.fr/occasion-voiture-marque-bmw-4.html lacentrale.fr/occasion-moto-marque-bmw-21.html lacentrale.fr/occasion-voiture-modele-ford-mustang-6.html lacentrale.fr/occasion-voiture-marque-toyota.html lacentrale.fr/occasion-voiture-marque-lexus-41.html lacentrale.fr/occasion-voiture-marque-mini-19.html lacentrale.fr/occasion-voiture-modele-ford-mustang-10.html lacentrale.fr/occasion-voiture-marque-mercedes-5.html lacentrale.fr/voiture-occasion-porsche-911-911/type/997-21.html lacentrale.fr/occasion-voiture-marque-peugeot-4.html lacentrale.fr/occasion-voiture-marque-ford-5.html lacentrale.fr/occasion-voiture-modele-kia-sportage-2.html	Peugeot 5008 Peugeot308 Jumpy Renault Clio RS Cabriolet	Car

An example of clustering results of Semantic data set Labeled using Wikipedia Pages

Clustering of Urls

Cluster Number	Associated URLs for each cluster	Label	Concept
1	afrologize.blogspot.fr/2013/07/comment-prendre-soin-des-tresses-avec.html extensionstopchrono.over-blog.com/article-extensions-a-clip-de-cheveux-naturels-lisses-meches-53213780.html madmoizelle.com/youtubeuses-beaute-afro-francophones-436919 curlidole.fr/4-idees-coiffures-a-realiser-sur-les-cheveux-des-enfants-aux-cheveux-crepus-frises-et-boucles aufeminin.com/idees-maquiller.html,madmoizelle.com/maquiller-yeux-marrons-779447 coiffure-simple.com/2016/11/50-magnifiques-couleurs-cheveux-tendance-2017 beautiful-boucles.com/comment-embarquer-ses-cosmetiques-en-voyageavion-conseils-coiffure-produits	pretty hair for curly hair hairdressing tutor hairstyle for hair hair hairstyles	Hairdressing

a. An example of clustering results of Semantic data set labeled using French Words

Concept	Sport	Music	Travel	Education	News
Clusters	sport24.lefigaro.fr	skyrock.net	opodo.fr	linkedin.com	lepoint.fr
	sports.fr	skyrock.com	expedia.fr	linguee.fr	europe1.fr
	footmercato.net	music.skyrock.com	govoyages.com	researchgate.net	bfmtv.com
	rmcsport.bfmtv.com	skyrock.mobi	flights-results.liligo.fr	la-conjugaison.nouvelobs.com	lexpress.fr

b. Cluster Labelization using Navigation Data Set labeled Using Host names

Question?