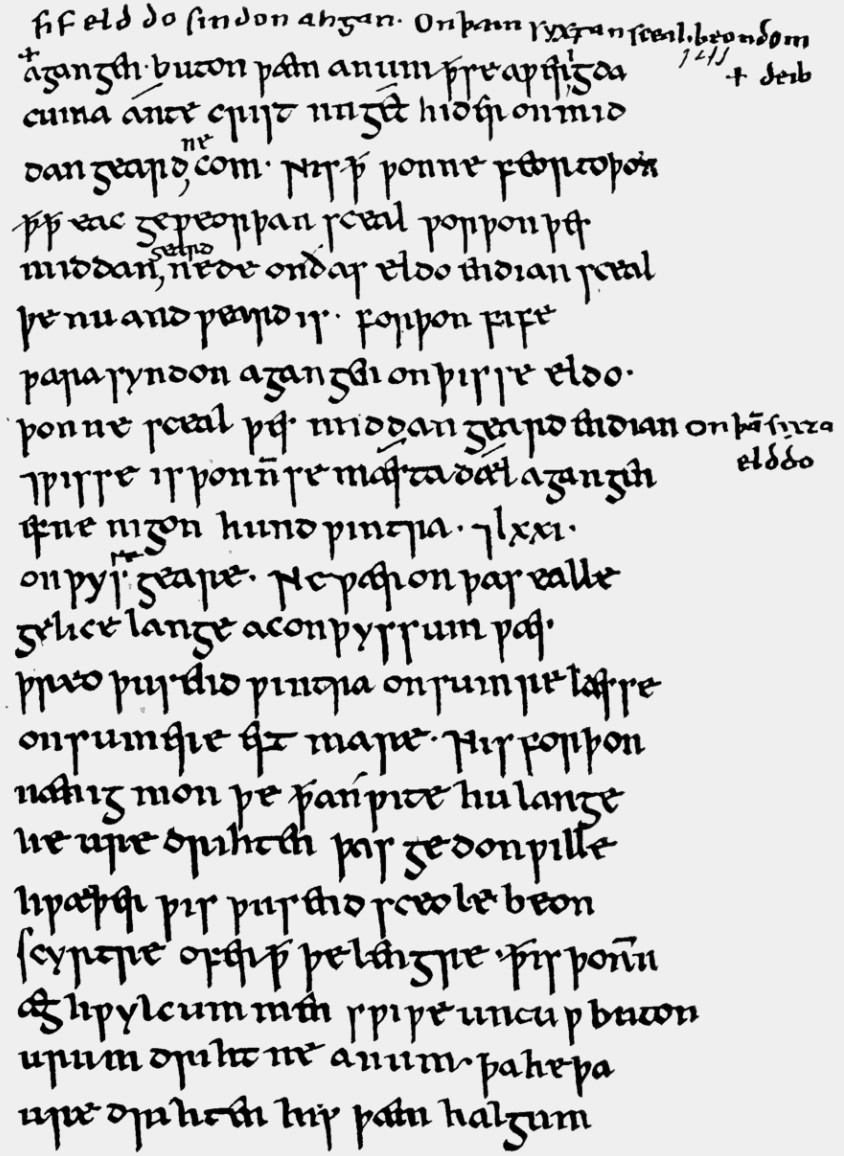
Natural Language Processing e Computational Linguistics

@Vitor Meriat



BLICKLING HOMILIES, MARQUIS OF LOTHIAN'S MS. (A.D. 971)
leaf 141.

About me

Vitor is a computer scientist who is passionate about creating software that will positively change the world we live in.

Currently, he works as **Data Scientist and Machine Learning Engineer** at **ESX**, where he is helping to shape new disruptive services based in Cloud Computing, Big Data and Artificial Intelligence. **Microsoft MVP Artificial Intelligence**.

vitormeriat.com.br

linkedin.com/in/vitormeriat

twitter.com/vitormeriat

github.com/vitormeriat

youtube.com/vitormeriat

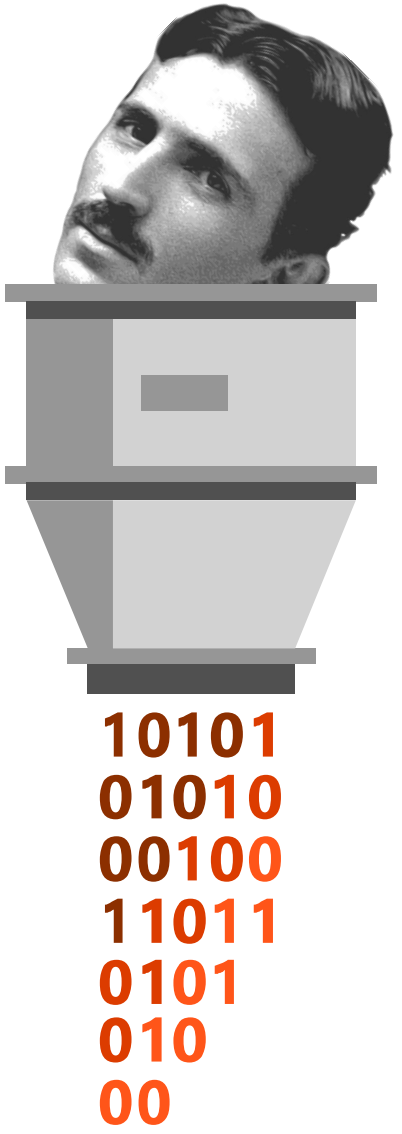
esx

Microsoft

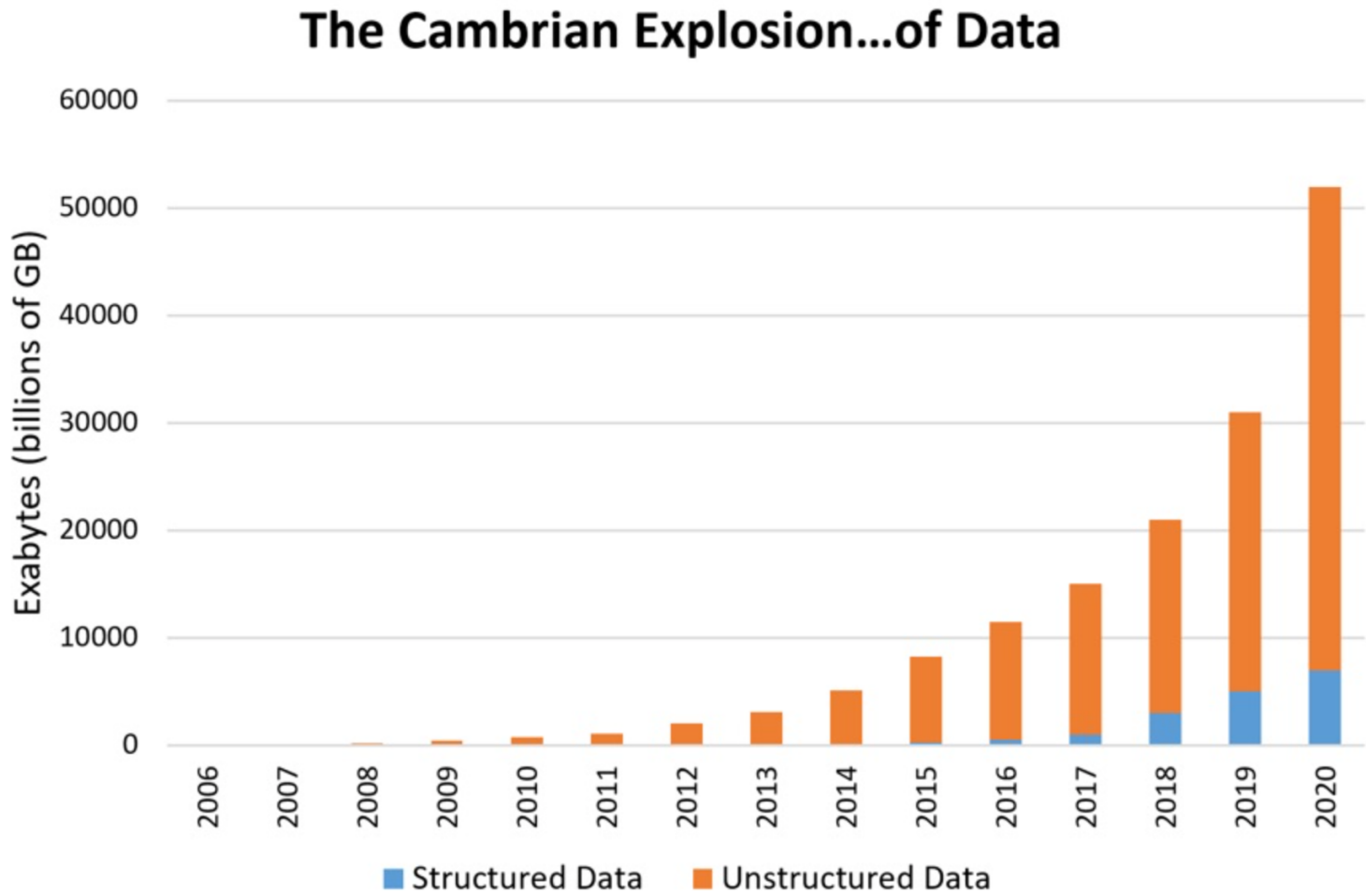
MVP Microsoft® Most Valuable Professional

Agenda

- O que é análise de texto;
- Natural Language Processing;
- Garbage in, garbage out;
- Computational Linguistics.



Big Data



Fonte: https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462

Text Analysis

Text Analysis

"...utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the 'action points' in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points."

October 1958 ***IBM Journal article*** *by H. P. Luhn,* ***A Business Intelligence System***

Fonte: https://dl.acm.org/citation.cfm?id=1662381

Garbage in, garbage out

Fonte: https://xkcd.com/1838

Natural Language Processing

Exposing impersonators of a Romanian writer using stopwords

Pastiche detection based on stopword rankings. Exposing impersonators of a Romanian writer

Liviu P. Dinu
Faculty of Mathematics and Computer Science
University of Bucharest
ldinu@fmi.unibuc.ro

Vlad Niculae
Faculty of Mathematics and Computer Science
University of Bucharest
vlad@vene.ro

Octavia-Maria Șulea
Faculty of Foreign Languages and Literatures
Faculty of Mathematics and Computer Science
University of Bucharest
mary.octavia@gmail.com

Abstract

We applied hierarchical clustering using Rank distance, previously used in computational stylometry, on literary texts written by Mateiu Caragiale and a number of different authors who attempted to impersonate Caragiale after his death, or simply to mimic his style. Their pastiches were consistently clustered opposite to the original work, thereby confirming the performance of the method and proposing an extension of the method from simple authorship attribution to the more complicated problem of pastiche detection.

The novelty of our work is the use of frequency rankings of stopwords as features, showing that this idea yields good results for pastiche detection.

1 Introduction

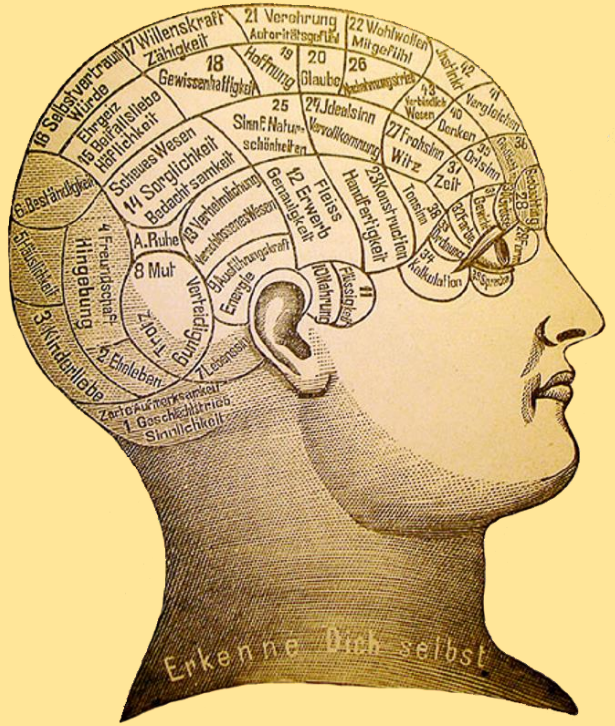
The postulated existence of the human stylome has been thoroughly studied with encouraging results. The term *stylome*, which is currently not in ercise in mimicking another's style. Even in this case, the best confirmation that the author of the pastiche can get is if he manages to fool an authorship attribution algorithm, even if the ground truth is known and there is no real question about it.

Marcus (1989) identifies the following four situation in which text authorship is disputed:

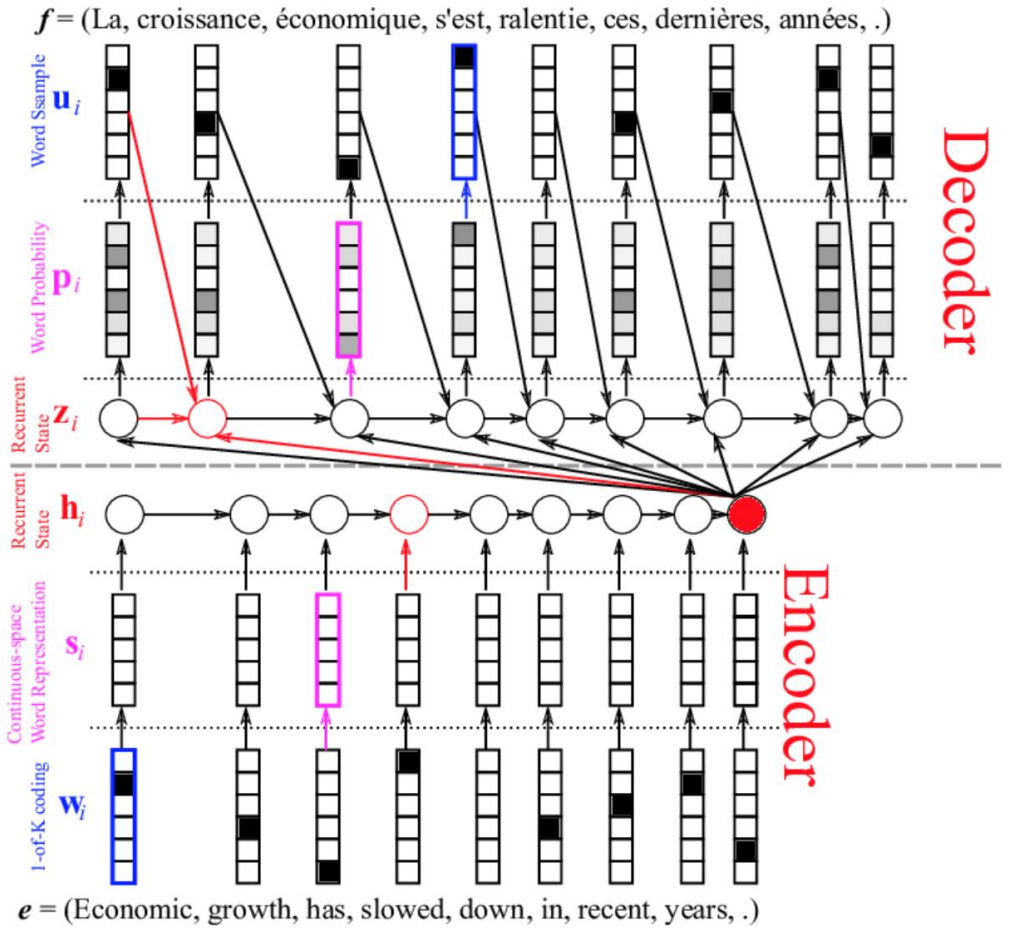
- A text attributed to one author seems non-homogeneous, lacking unity, which raises the suspicion that there may be more than one author. If the text was originally attributed to one author, one must establish which fragments, if any, do not belong to him, and who are their real authors.
- A text is anonymous. If the author of a text is unknown, then based on the location, time frame and cultural context, we can conjecture who the author may be and test this hypothesis.
- If based on certain circumstances, arising

Fonte: http://www.aclweb.org/anthology/W12-0411

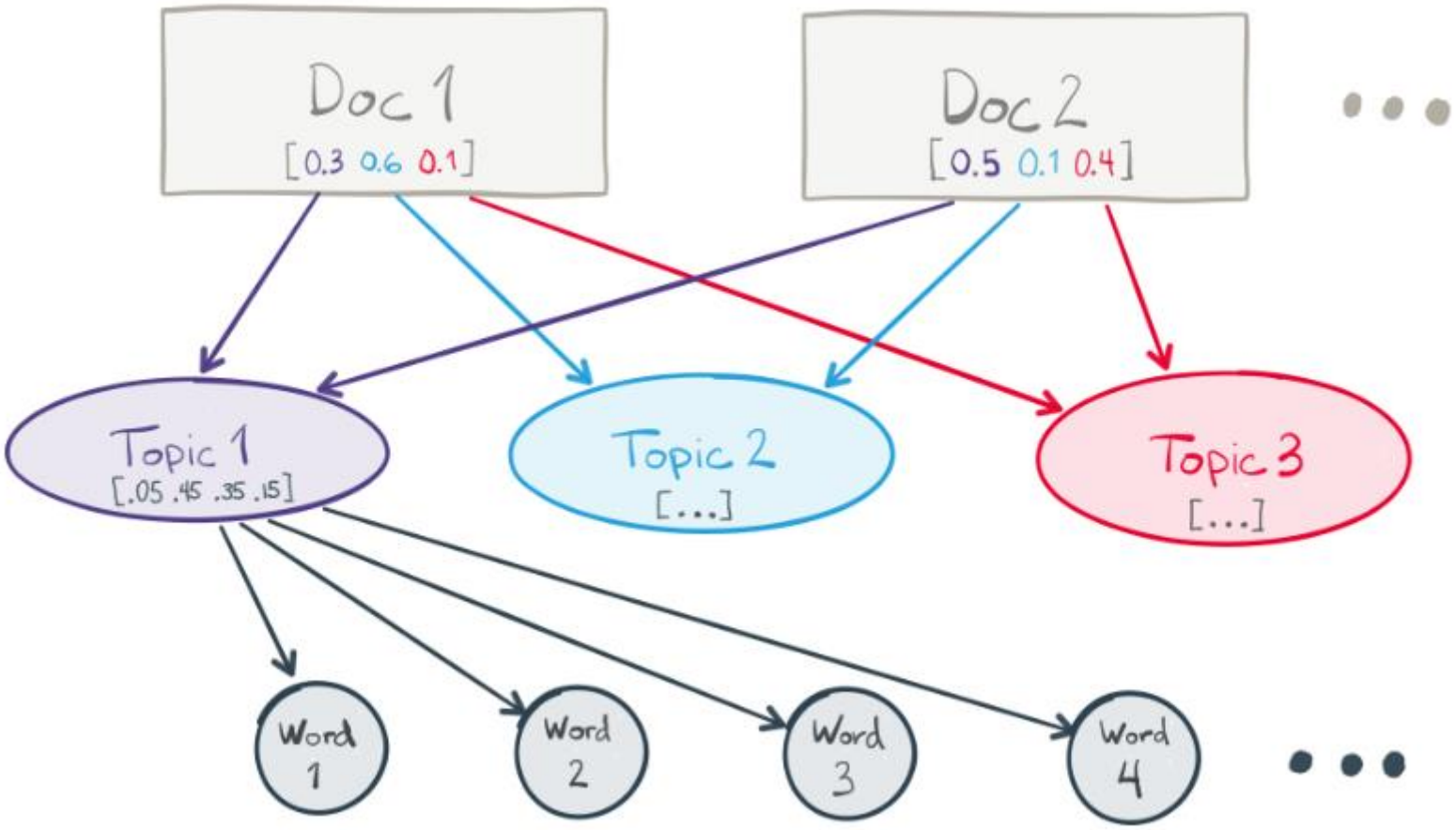
Computational Linguistics



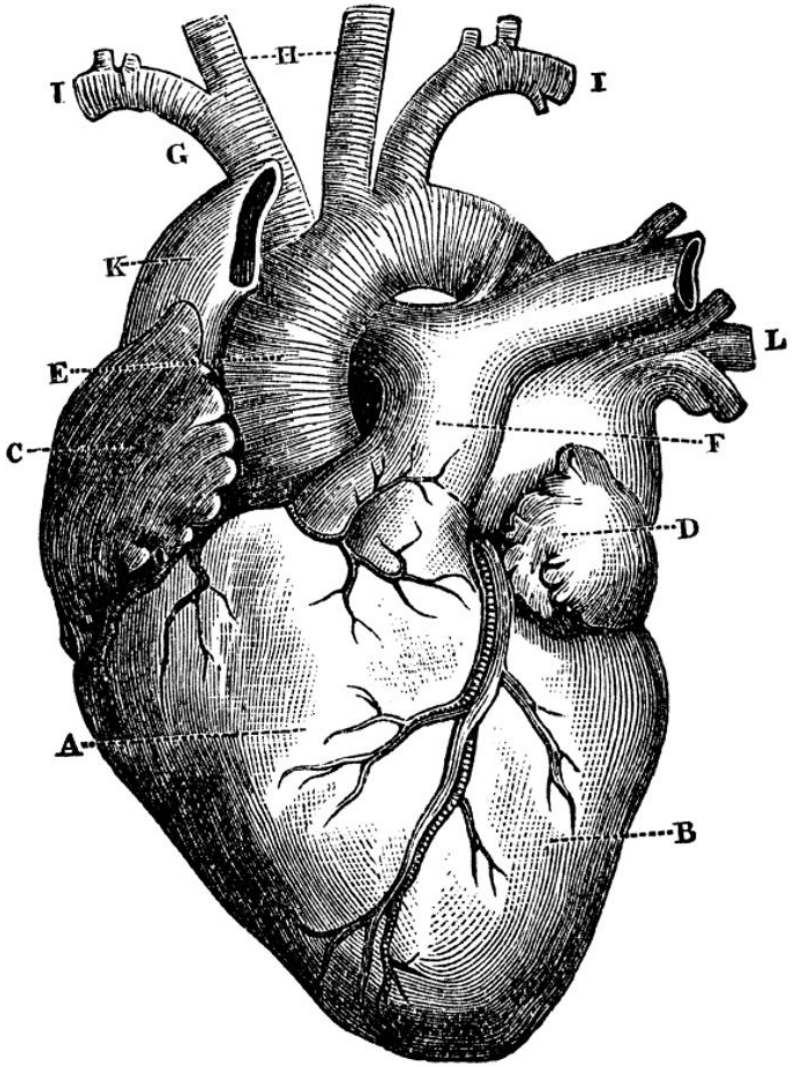
Neural Machine Translation



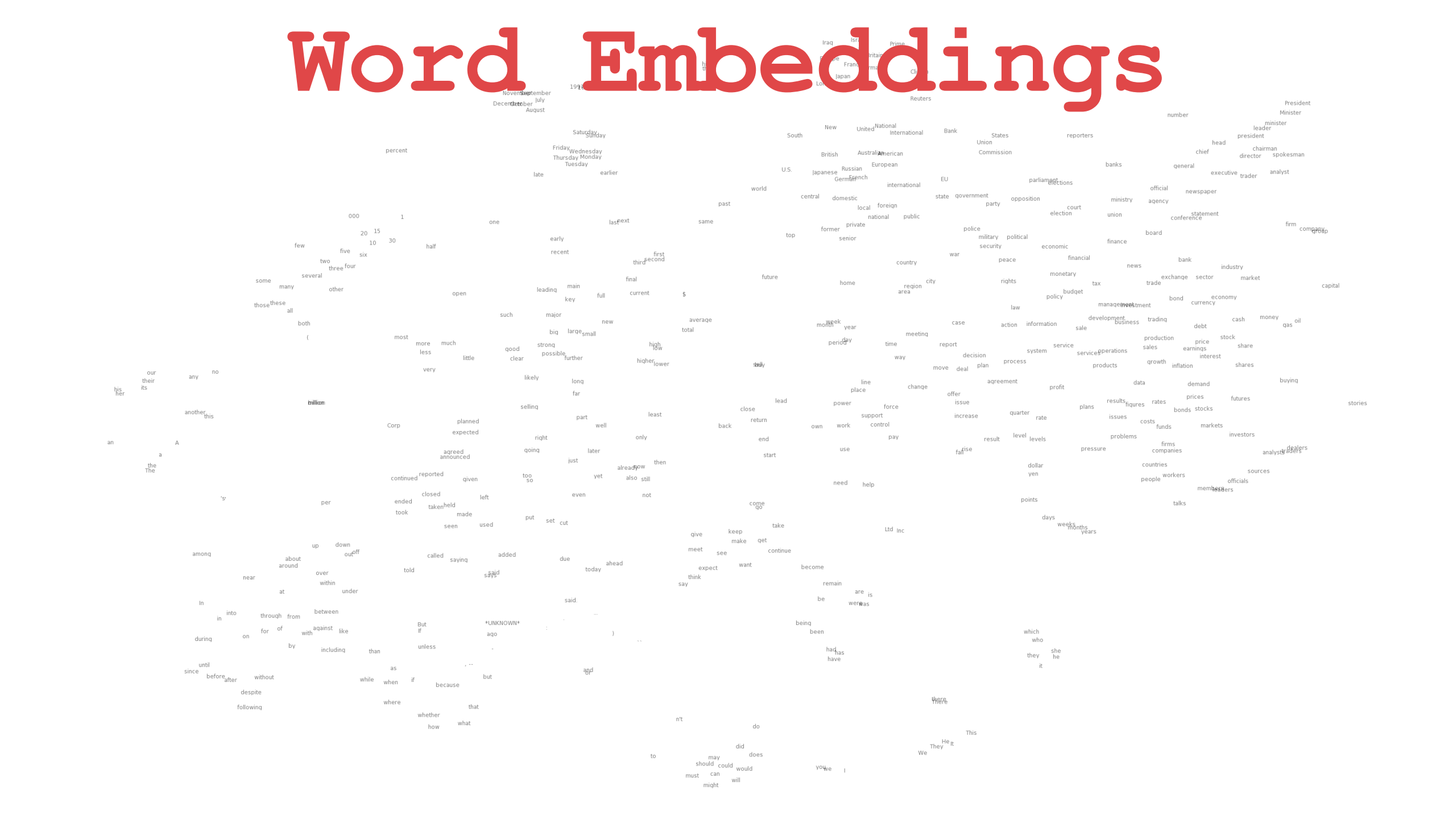
Topic Modeling



Emotions and needs



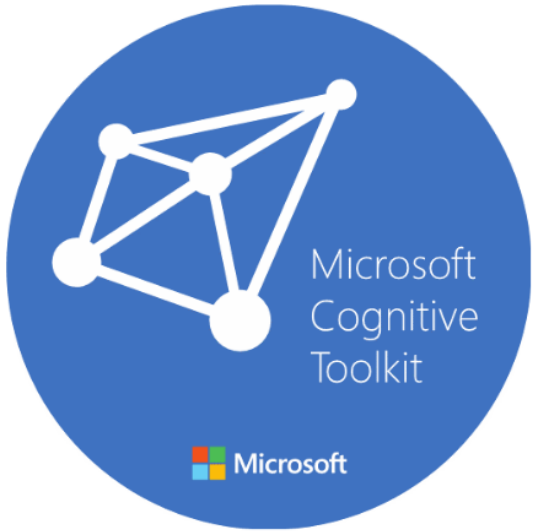
Word Embeddings

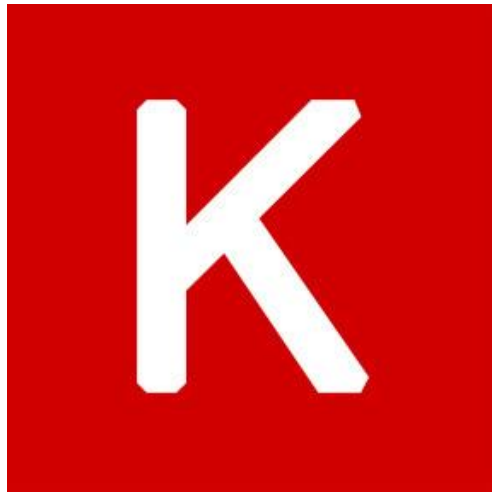


Frameworks

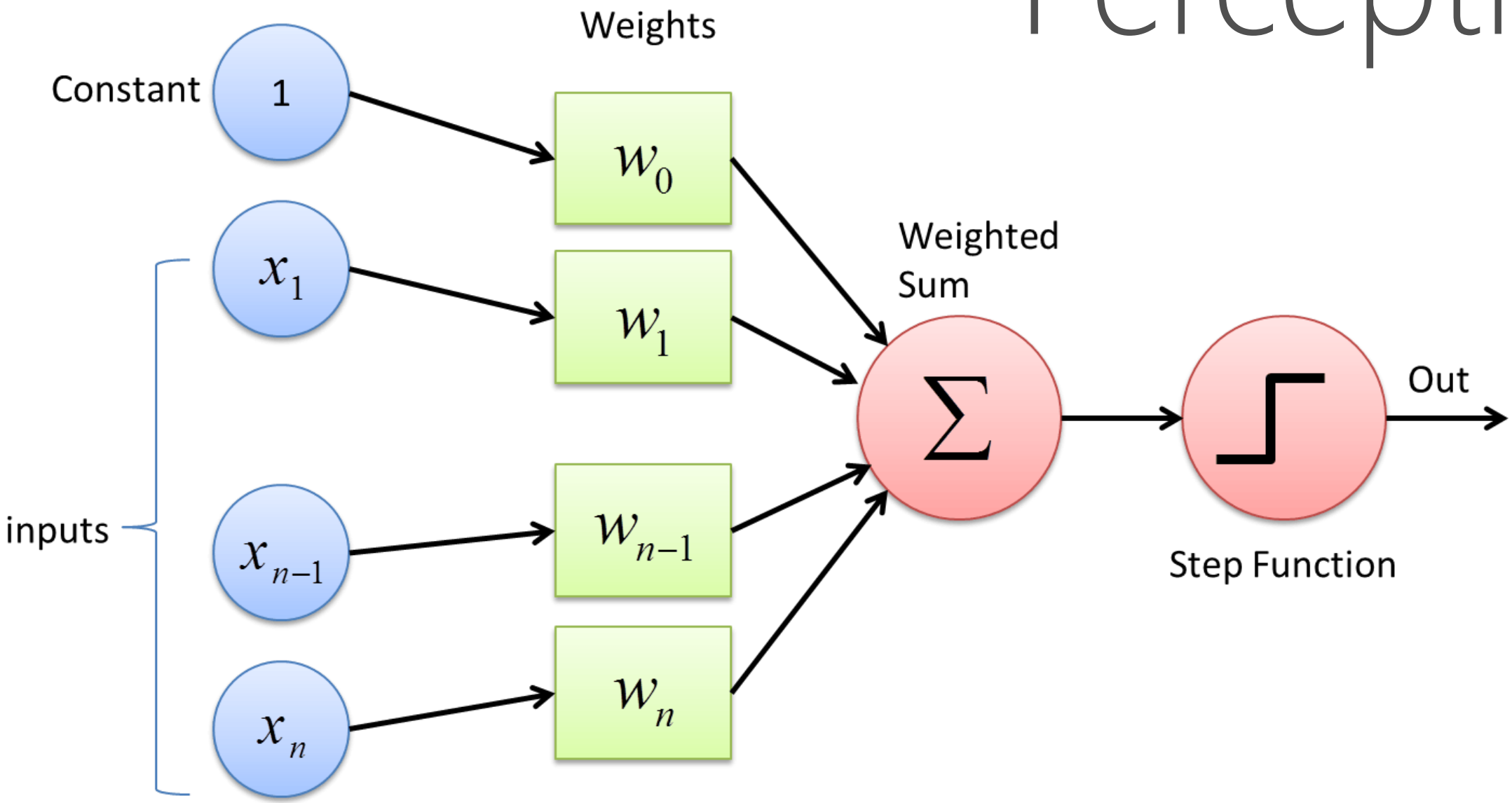
	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Deep Learning Frameworks

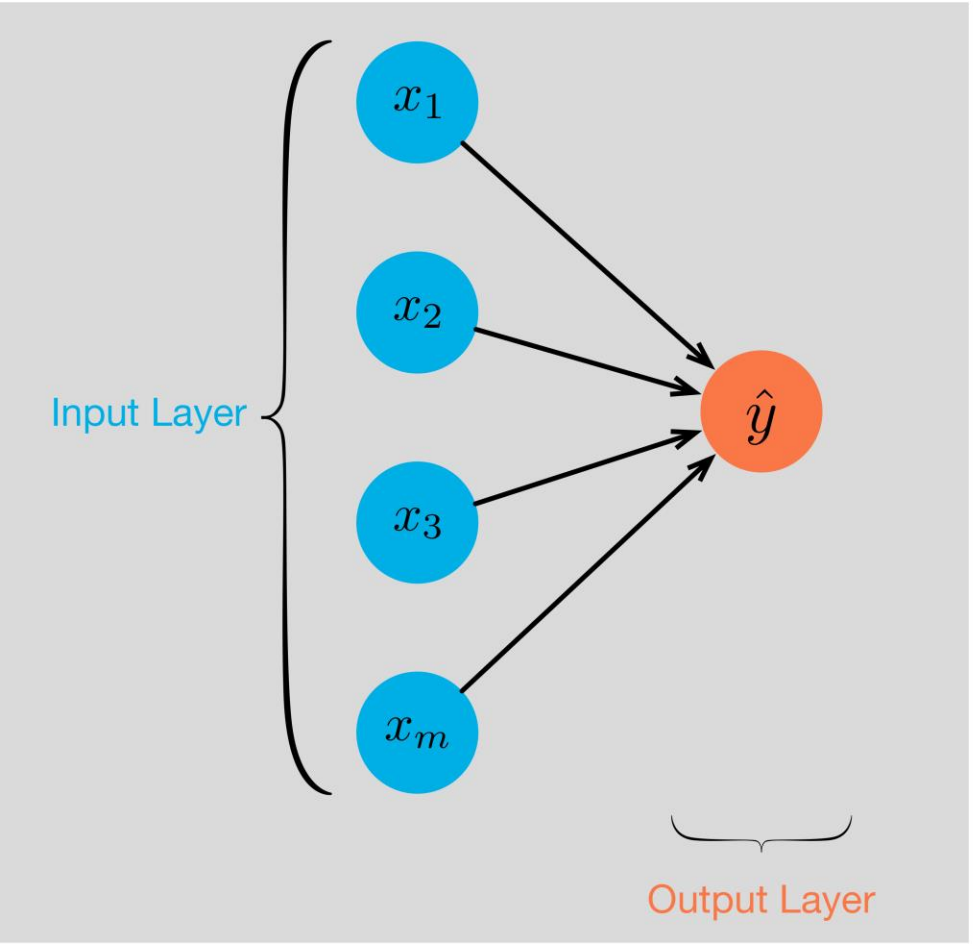


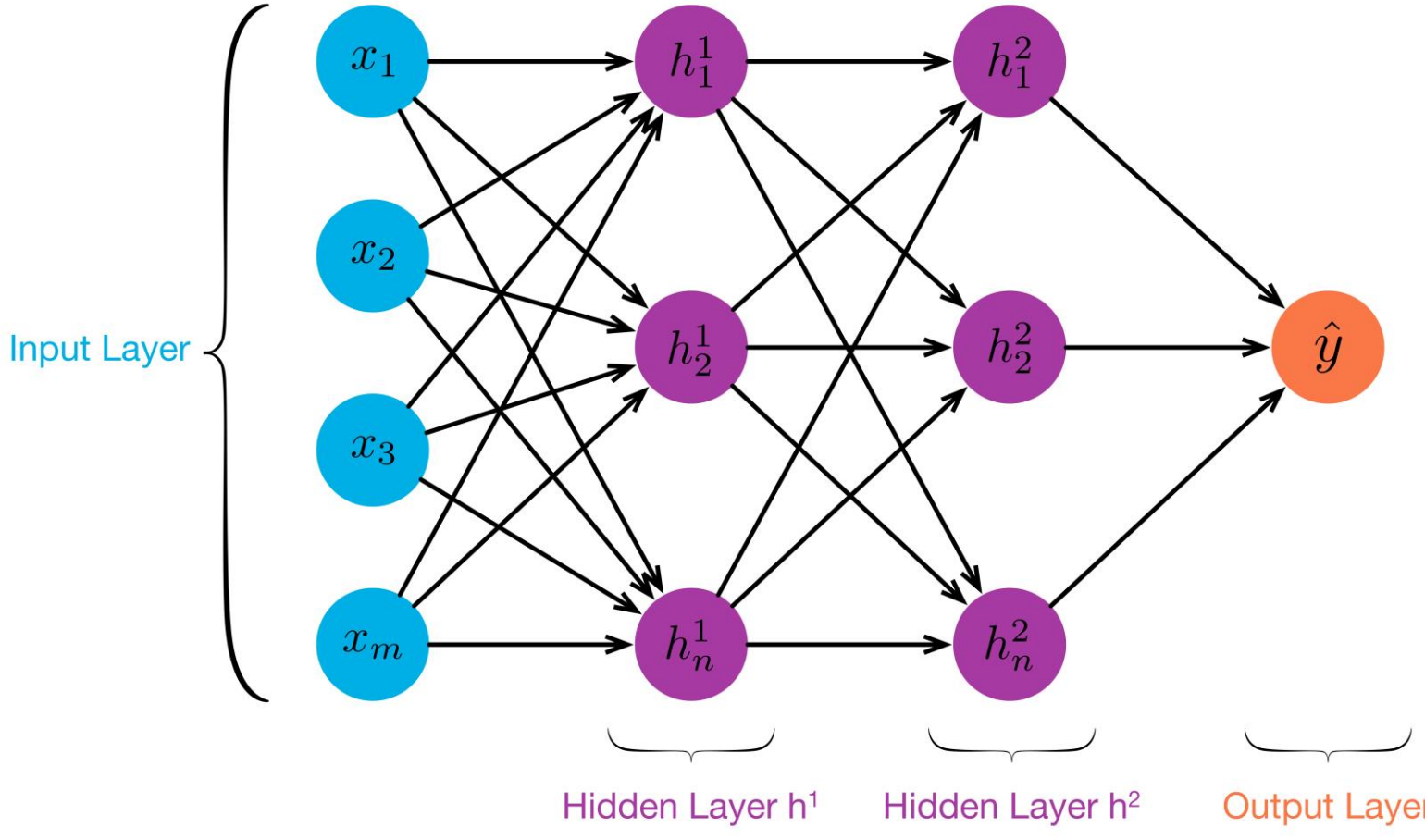


Perceptron


Weights
Constant
1
w_0
x_1
w_1
Weighted
Sum
Σ
Out
inputs
x_{n-1}
w_{n-1}
x_n
w_n
Step Function

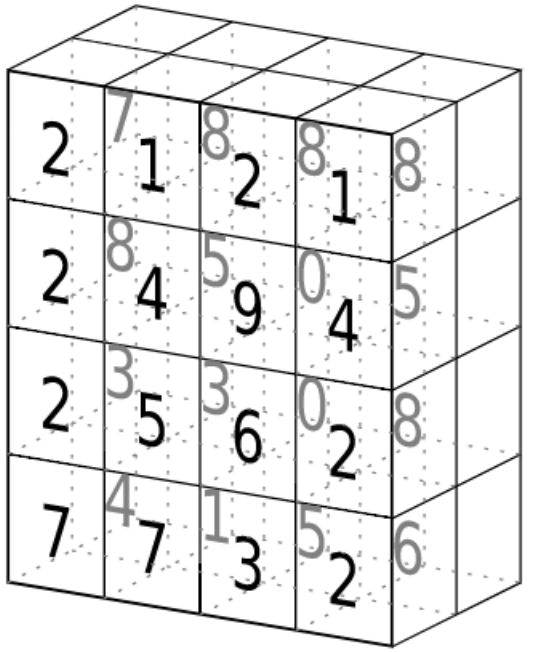
Single-Layer and Multi-Layer





't'
'e'
'n'
's'
'o'
'r'

3	1	4	1
5	9	2	6
5	3	5	8
9	7	9	3
2	3	8	4
6	2	6	4



Vetor

1

Matriz

2

Tensor

N

A faster, more efficient, more intelligent cloud

The need for **SCALE**

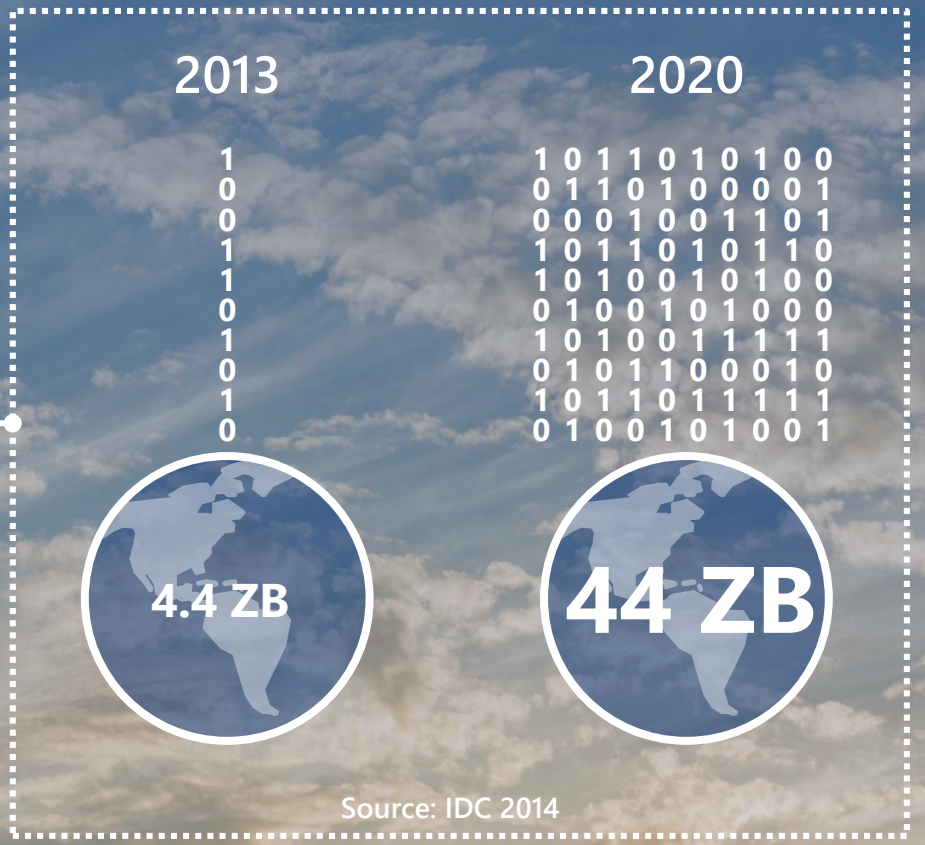
Data explosion: 2013 4.4 ZB - 2020 44 ZB
ML, DNN, AI are driving requirements up faster

The need for **LOW-LATENCY**

Autonomous decision making
Real-time insights into connected devices
Interactive user experiences

The need for **THROUGHPUT**

Cloud-scale services
Searches and recommendations (Indexing the Internet!)



1 SHOW
2 ME
3 THE
4 CODE

Thank you!

vitormeriat.com.br

linkedin.com/in/vitormeriat

twitter.com/vitormeriat

github.com/vitormeriat

youtube.com/vitormeriat