

# Natural Language Processing e Computational Linguistics

@Vitor Meriat

fiƿ eld do findon ahtan. On þam ƿætan ſceal beon ðom  
+ aƿanƿi bucon þam anum þre aƿanƿi ða 141 + ƿe  
cuna anre eƿiſe na ƿe hiðſi on mid  
dan ƿearð com. þiſ þe ƿonne ƿeoƿeoƿon  
þe eac ƿeoƿpan ſceal ƿoƿon þe  
midan, ne ðe on ða eldo ðiðian ſceal  
þe nu and ƿearð iſ. ƿoƿon þe  
ƿara ƿyndon aƿanƿi on þiſe eldo.  
þonne ſceal þe midan ƿearð ðiðian on þa ƿara  
þiſe iſ þon þe maſta ða aƿanƿi eldo  
þe ne nigon hund ƿinea. 7 lxxi.  
on ƿiſ ƿearð. þe ƿaſ on ƿaſealle  
ſelice lange acon ƿyſſum ƿaſ.  
þa ƿo ƿiſ ðið ƿinea on ſum þe laſſe  
on ſum þe þe maſe. þiſ ƿoƿon  
nahtiz mon þe ƿan ƿiſe hu lange  
ne ƿe oƿihte þaſ ƿe ðon ƿille  
hi ƿaſ þiſ ƿiſ ðið ſceole beon  
ſcƿeƿe oƿiſ þe lahtiz. þiſ þon  
aſ hi ƿileum mid ƿiſe uncu þ bucon  
uƿum oƿihte ne anum. þa he ƿa  
uƿe oƿihte hi þam halzum

# 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](http://vitormeriat.com.br)



[linkedin.com/in/vitormeriat](https://linkedin.com/in/vitormeriat)



[twitter.com/vitormeriat](https://twitter.com/vitormeriat)



[github.com/vitormeriat](https://github.com/vitormeriat)

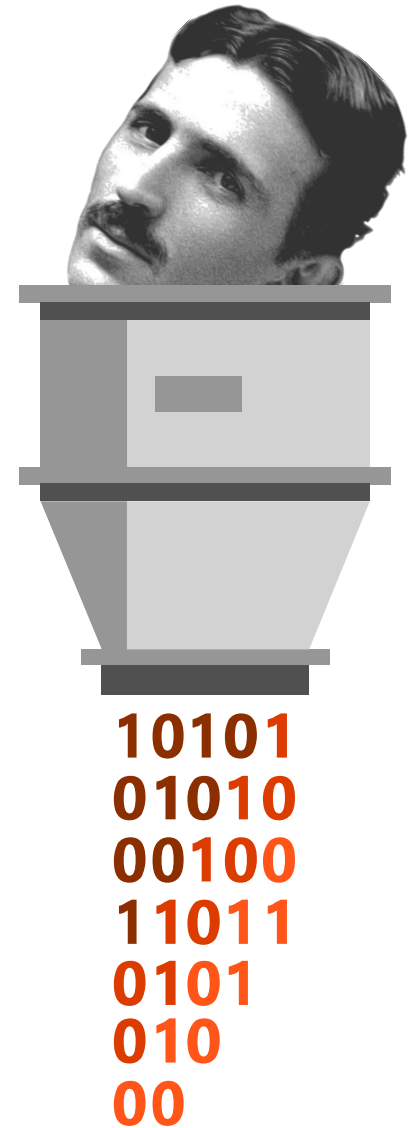


[youtube.com/vitormeriat](https://youtube.com/vitormeriat)



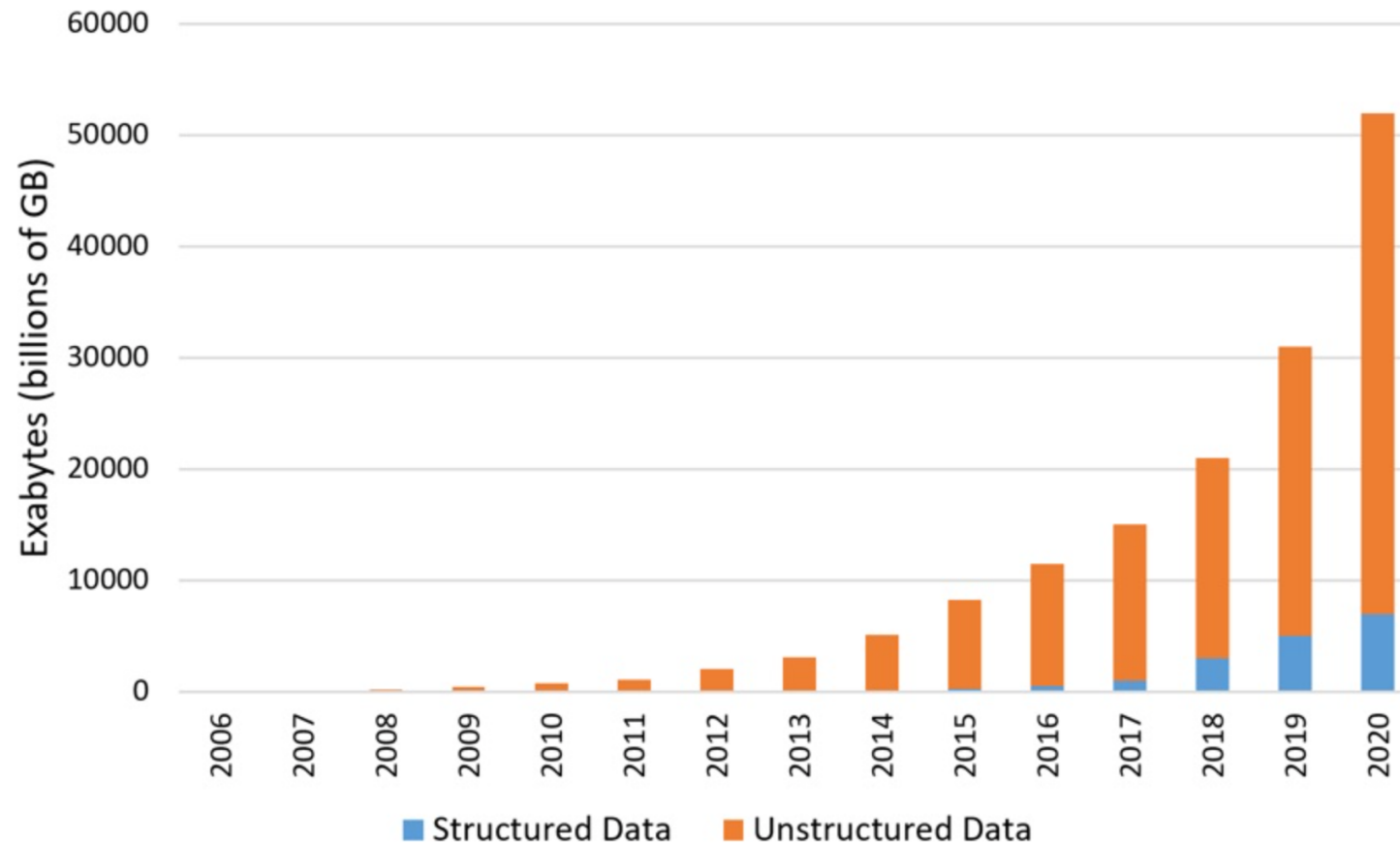
# Agenda

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



# Big Data

## The Cambrian Explosion...of Data



# 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**

# Garbage in, garbage out



# 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

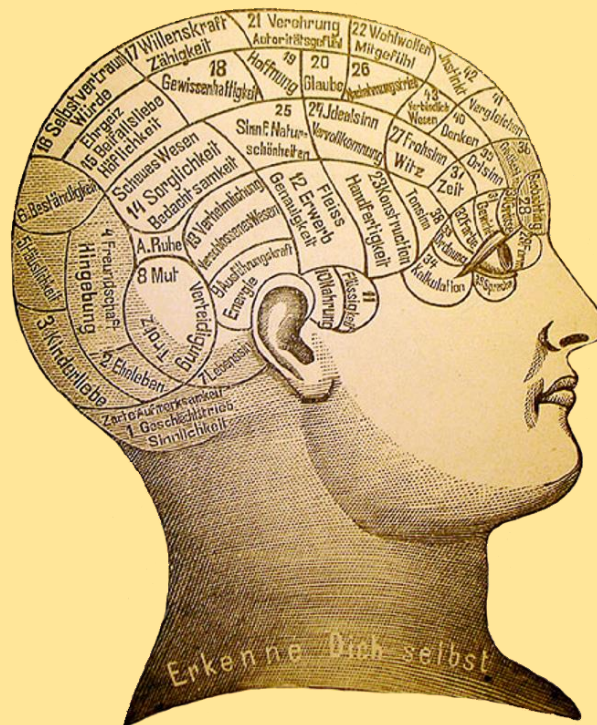
The postulated existence of the human stylome has been thoroughly studied with encouraging results. The term *stylome*, which is currently not in

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.

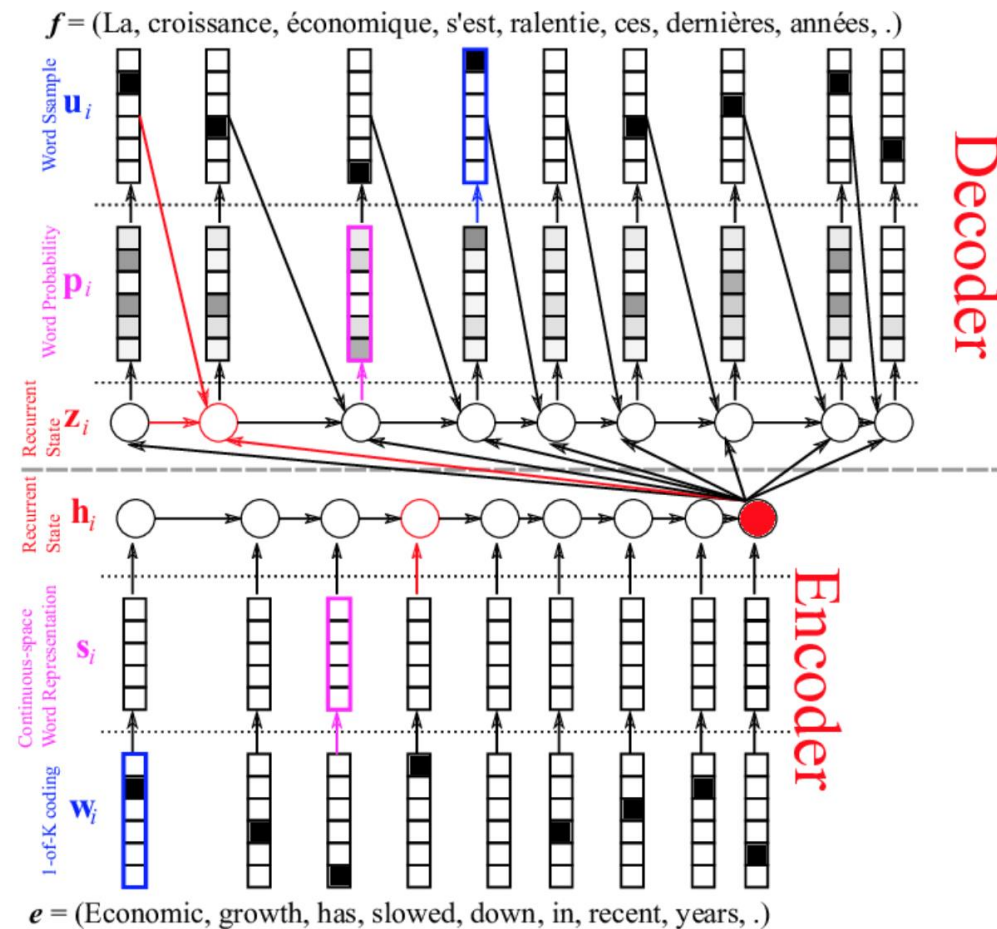
Marcus (1989) identifies the following four situation in which text authorship is disputed:

- 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

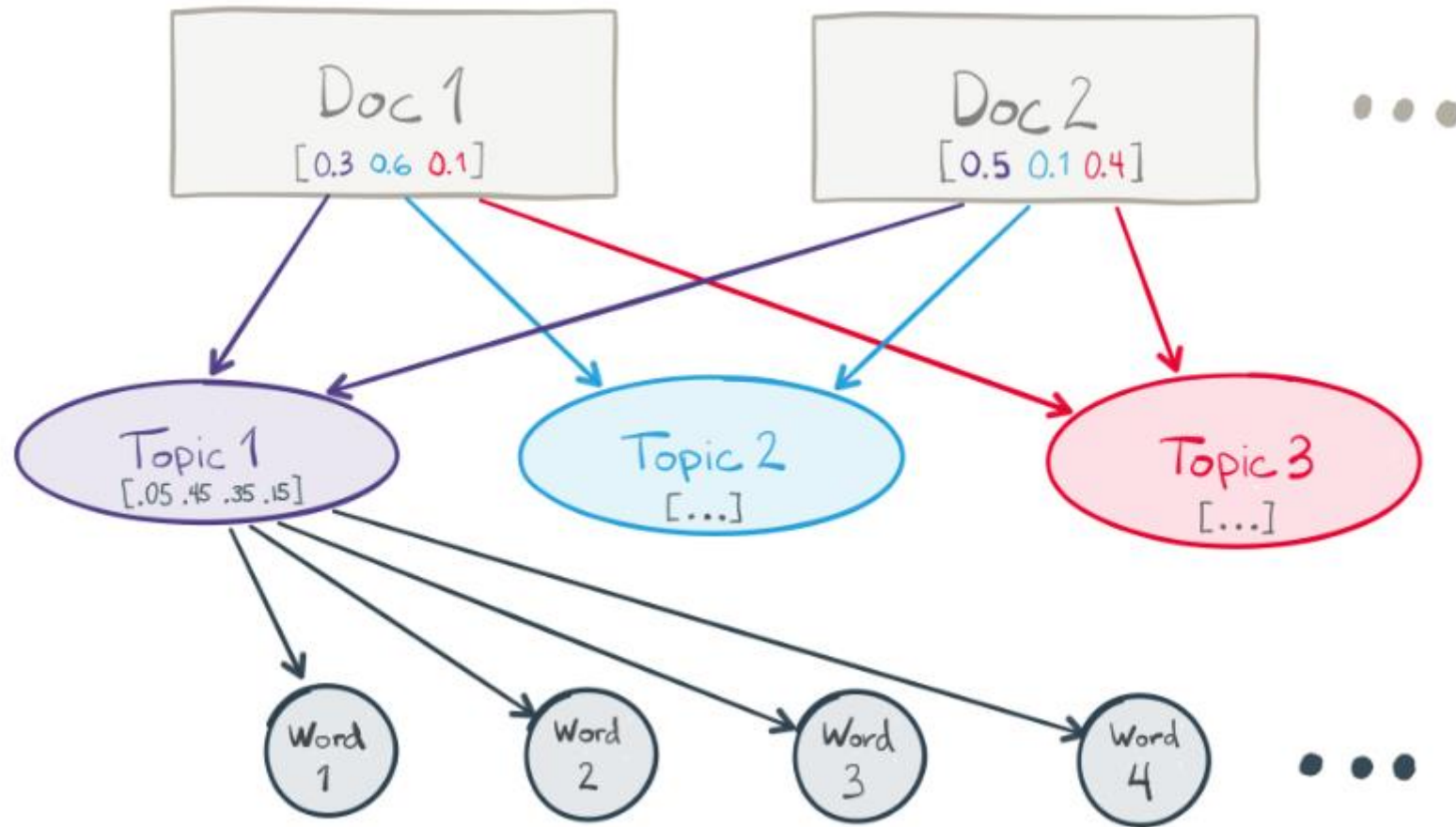
# Computational Linguistics



# Neural Machine Translation



# Topic Modeling



**T-REX IN:  
"COMPUTATIONAL  
LINGUISTICS"**



Computational linguistics is the study of computer-based language processing!



A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language - English, for example - can have more than one meaning!



Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?



Many have focused on statistical modelling of language, but this approach is approximate. I agree!



What do YOU know about computational linguistics?

Ever read a little paper called "Non-Statistical Models for Unsupervised Prepositional Phrase Attachment?"

That was me!

It was some of my earliest work on head word tuples!



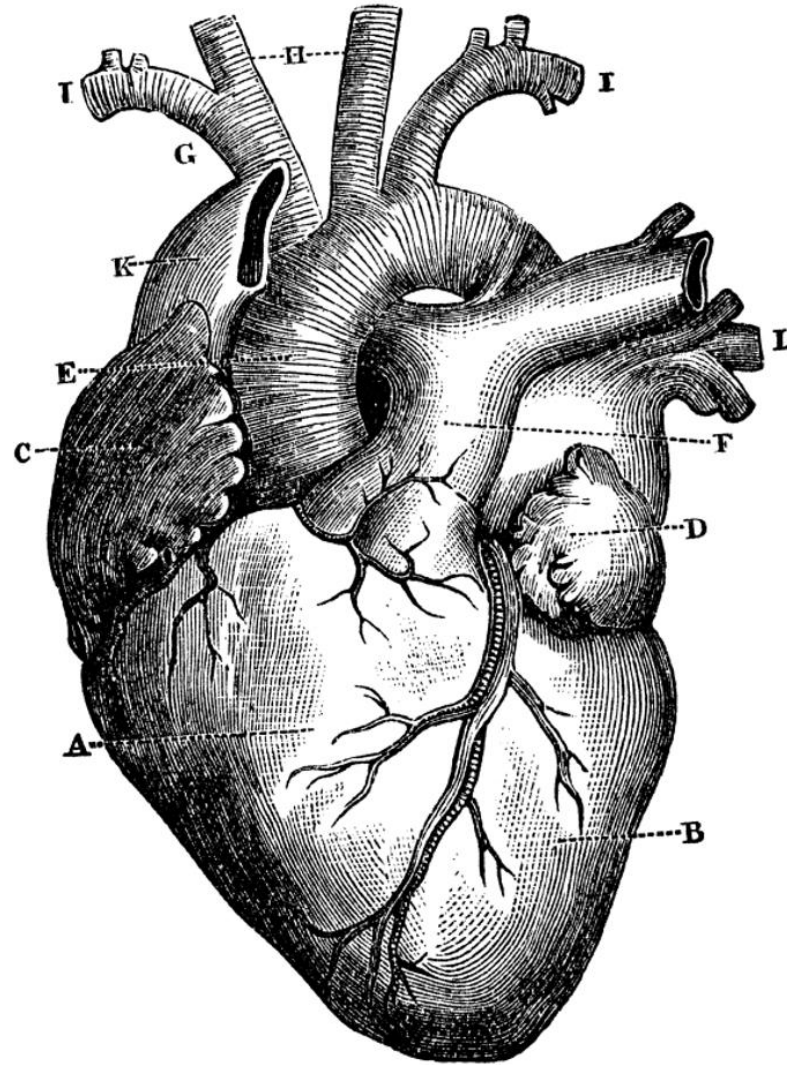
Shit man, you know more about this than I do!



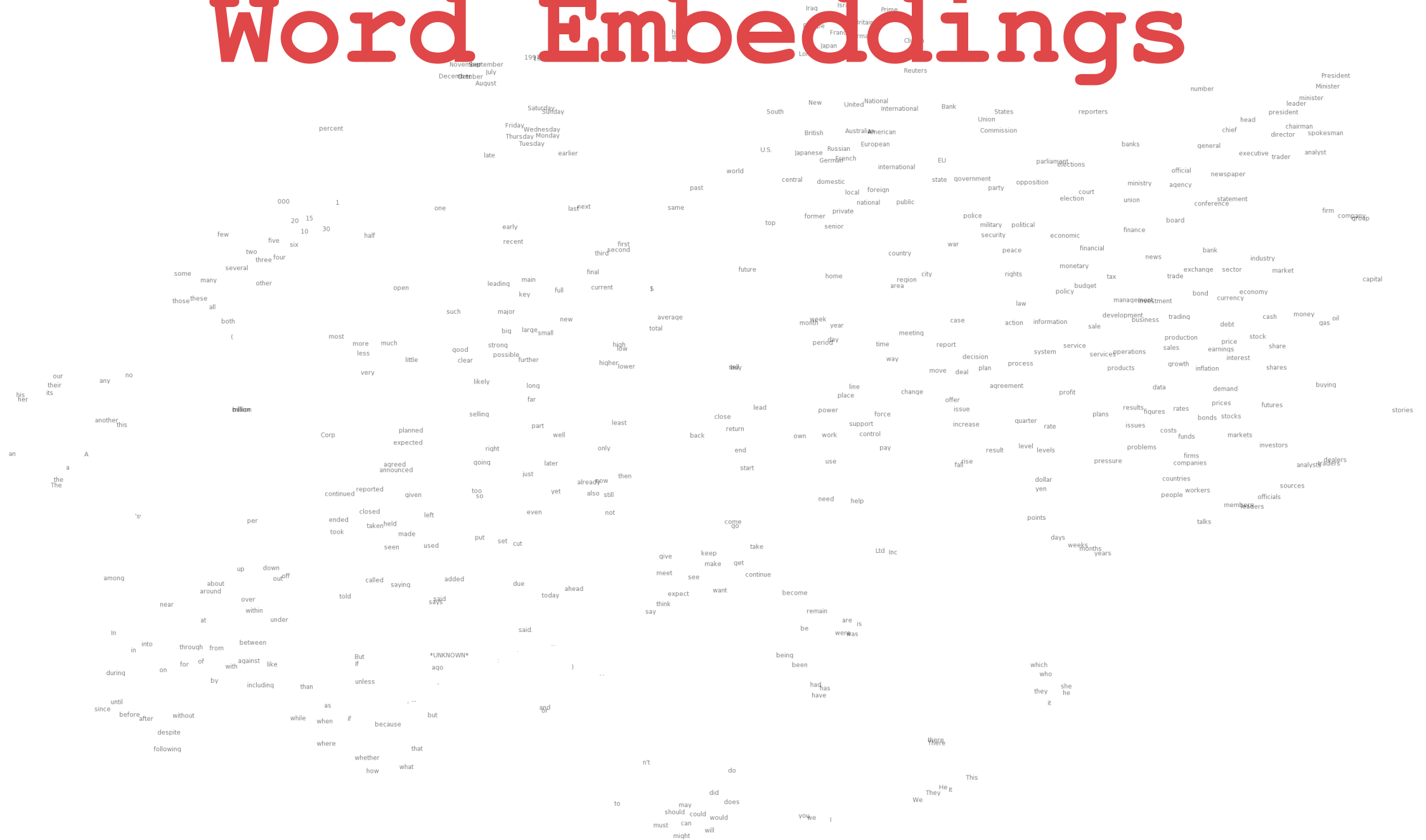
You know what? You should be the one doing the talking here!



# Emotions and needs



# Word Embeddings

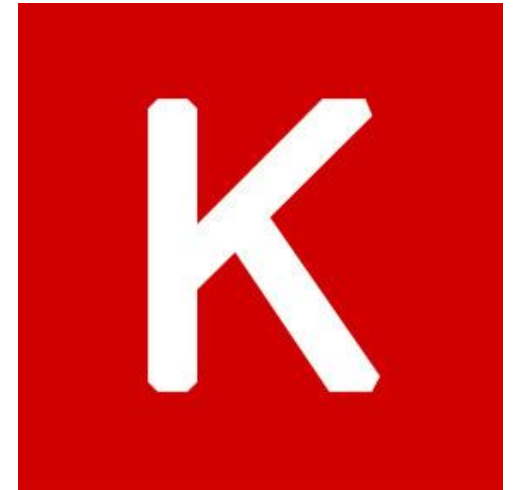


# Frameworks

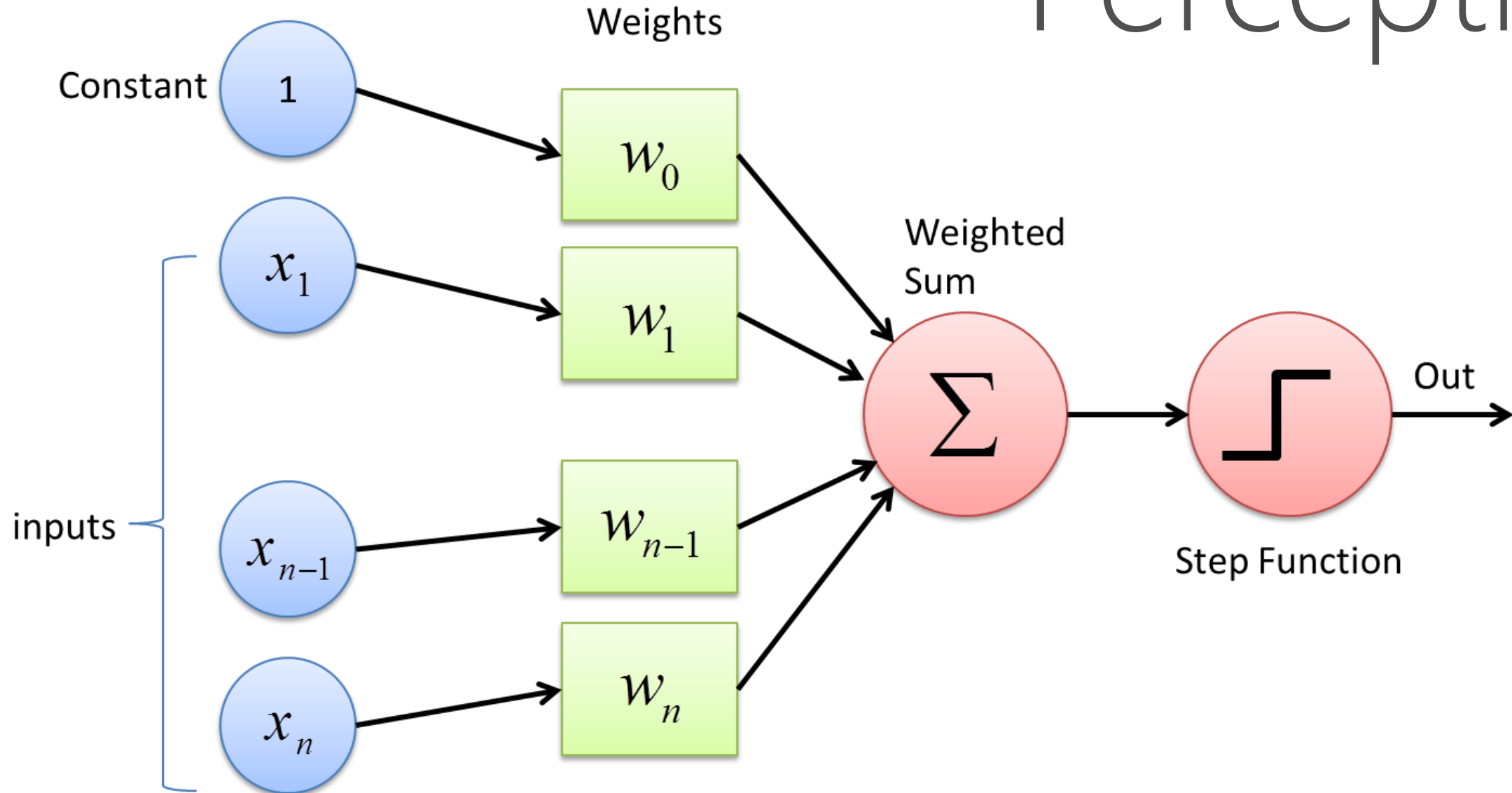
	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓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



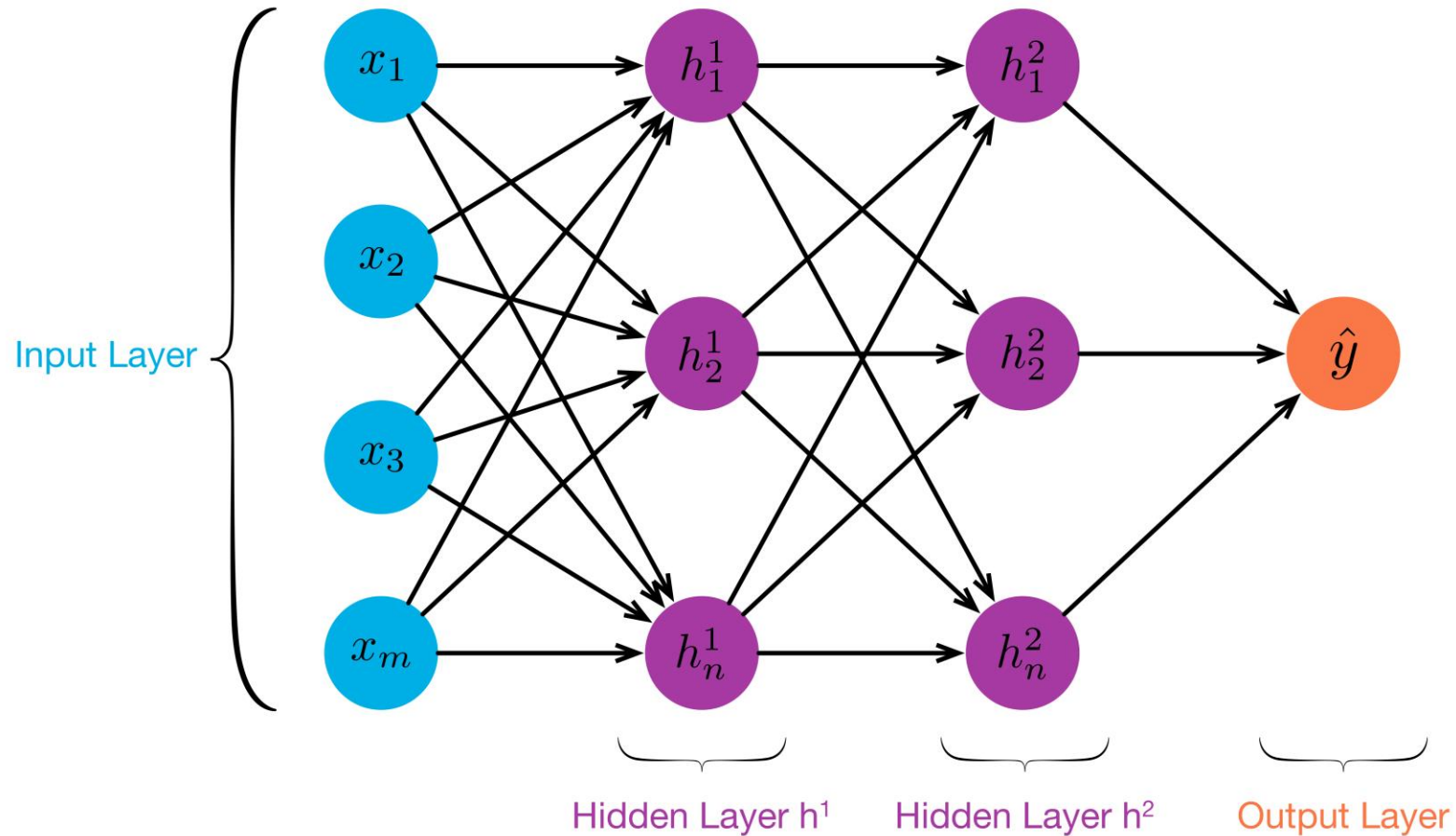
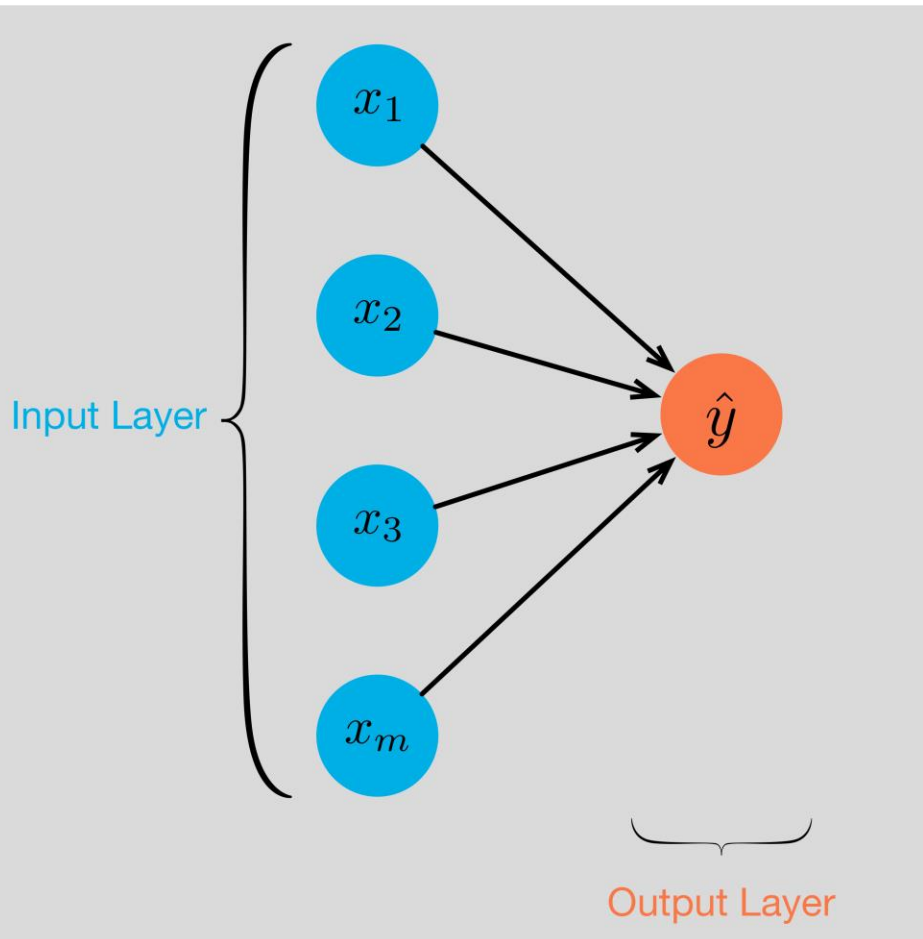
# Deep Learning Frameworks



# Perceptron



# Single-Layer and Multi-Layer



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

Vetor  
1

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

Matriz  
2

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

Tensor  
N

# A faster, more efficient, more intelligent cloud

## ➔ The need for **SCALE**

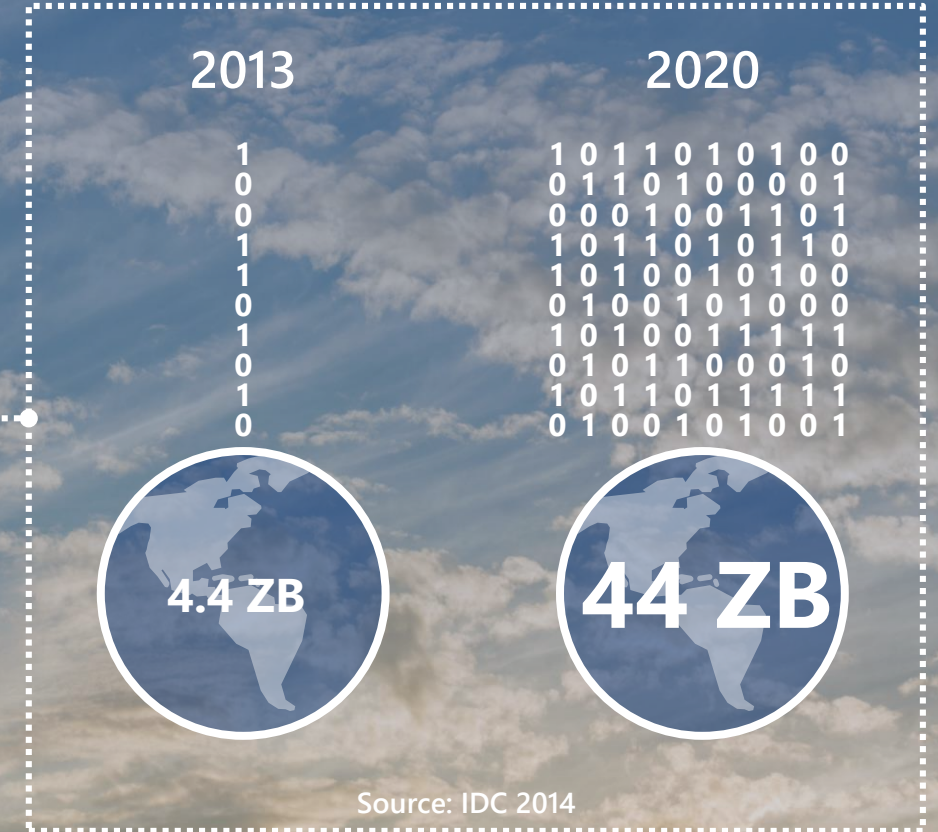
Data explosion: 2013 4.4 ZB - 2020 44 ZB  
ML, DNN, AI are driving requirements up faster

## ➔ 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](http://vitormeriat.com.br)



[linkedin.com/in/vitormeriat](https://linkedin.com/in/vitormeriat)



[twitter.com/vitormeriat](https://twitter.com/vitormeriat)



[github.com/vitormeriat](https://github.com/vitormeriat)



[youtube.com/vitormeriat](https://youtube.com/vitormeriat)

