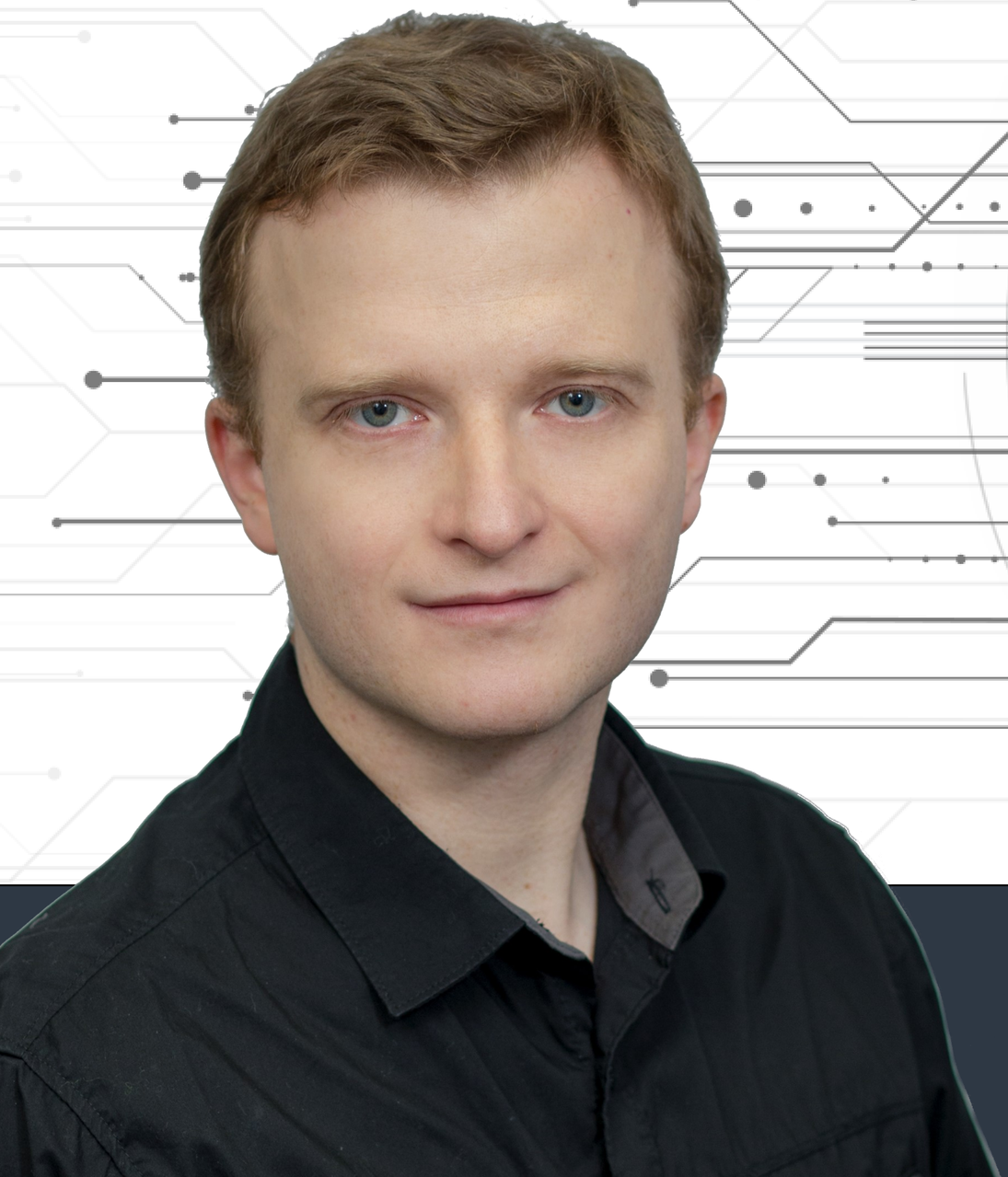


XTM AI

November 2020

Rafał Jaworski

Linguistic AI Expert

XTM International

XTM International

Assessing Cross-lingual Word Similarities Using Neural Networks

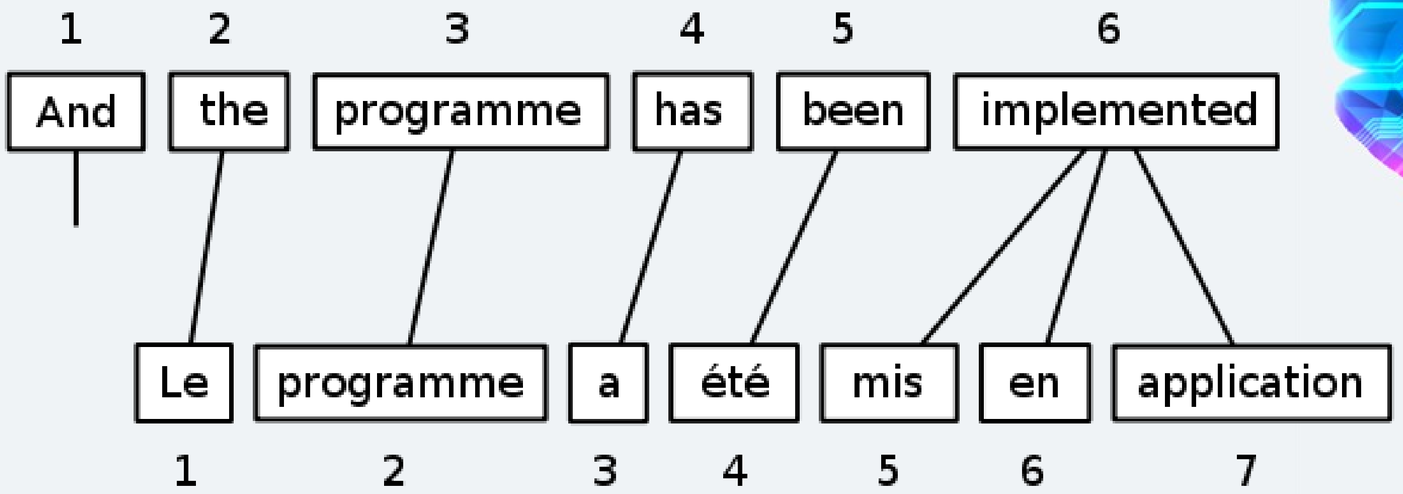
XTM International

Word Similarities

Word-level alignment

- Word-level alignment is a process of automatic matching of similar words.

- For a pair of sentences the algorithm has to decide which words from the source and target sentences are each other's counterparts:



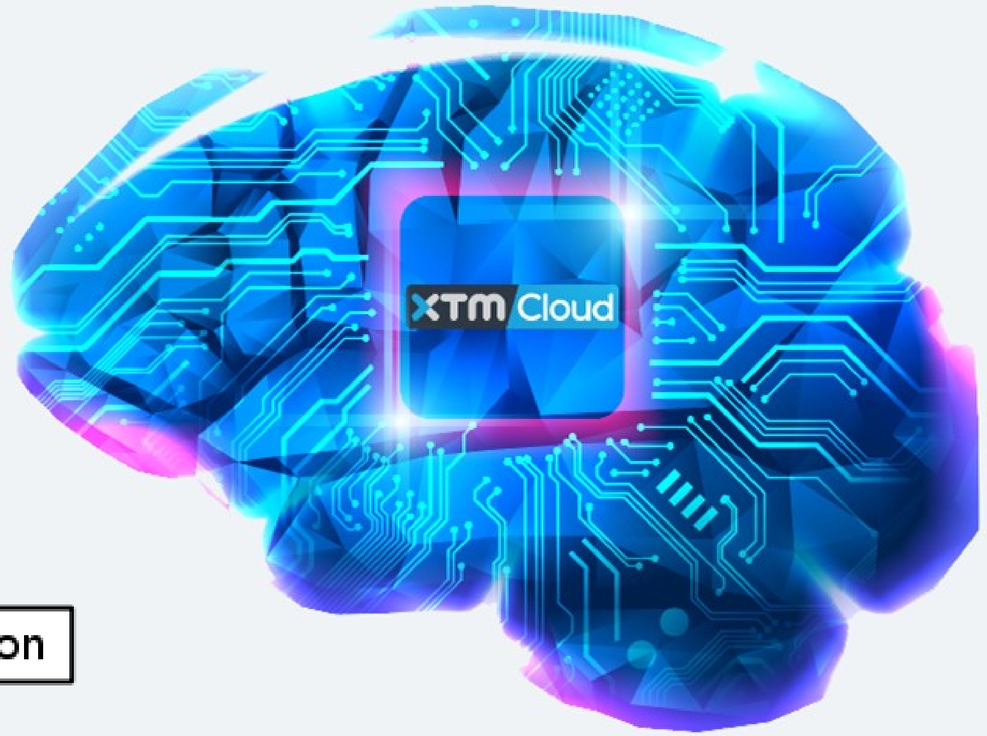


Image source: wikipedia.org

XTM International

Word Similarities

Word-level alignment - objectives

The information about word alignments between the source and target sentence allows for the implementation of the following intelligent features:

- Bilingual terminology extraction
- Automatic placement of HTML tags in translation
- Computer-assisted review (highlighting unmatched words)
- and many others…

Word Similarities

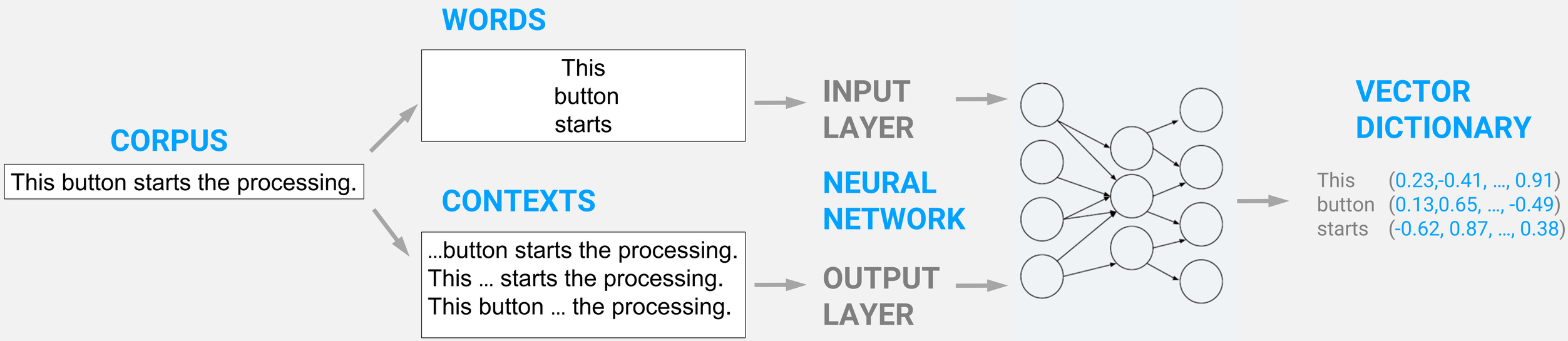
Vector Space

- Based on the concept of Vector Space, first introduced by Google Research Center in 2013
- Using a deep neural network analysis of a corpus: Google News Corpus to produce a language model for English
  - Neural network to predict the surrounding words based on the current word
  - By-product – representation of words with vectors of 300 numeric values that govern the relationship of the word with surrounding words

XTM International

Word Similarities

Vector Space



Word Similarities

Vector Space

- Vector Space is limited by the corpus that was used for training
- It can assess both *semantic* and *syntactic* similarity for a pair of words
- For *syntactic* information it can verify the following related words as similar:
  - Adjective to adverb: apparent -> apparently
  - Opposites: possible -> impossible
  - Comparative: great -> greater -> greatest
  - Superlative: easy -> easiest
  - Present participle: think -> thinking
  - Plural nouns: mouse -> mice

XTM International

Word Similarities

Vector Space

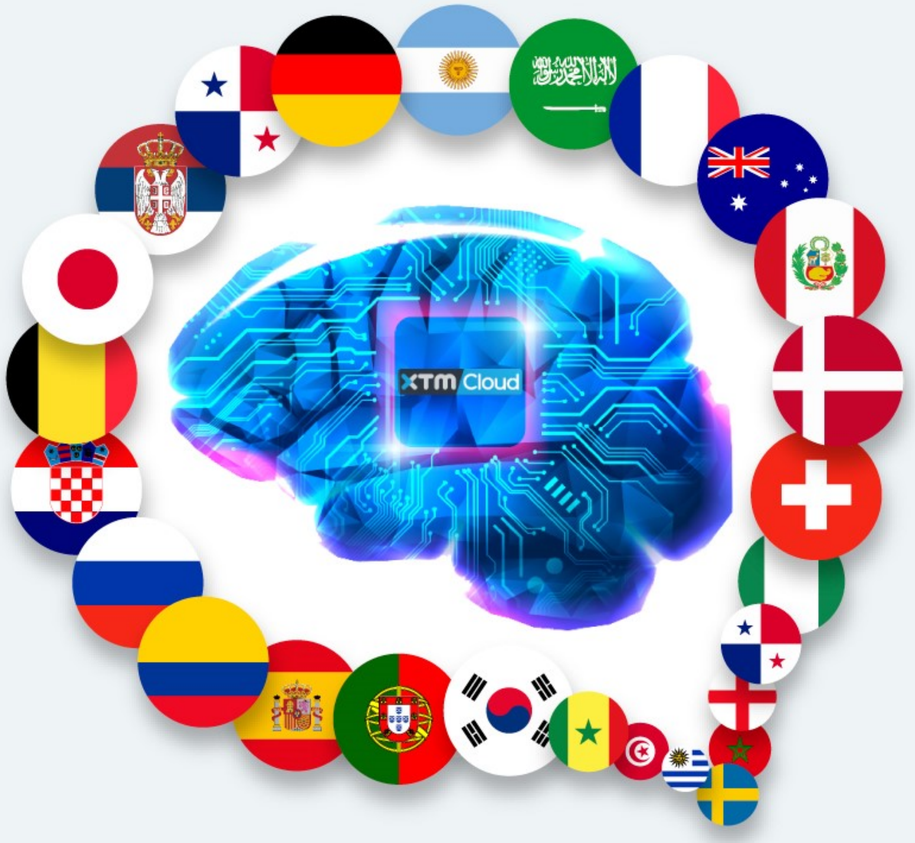
- For *semantic* information it can answer the following type of question:
  - If Einstein was a scientist, what was Mozart
  - Are Apples fruit or vegetables
  - If Athens is to Greece, then what is the equivalent for Norway
  - If king is to man, what is the equivalent for woman
  - Identify concept cluster having a close relationship:
    - potato, salad, radish, broccoli, tomato
    - apple, pear, orange, lemon, raspberry, blueberry, strawberry

XTM International

Word Similarities

Vector Space

- In 2016 Facebook Research published further work on Vector Space
- Based on a crawl of the entire Internet
- Vector Space for 157 languages
- The most *complete Vector Space language model* for each of the 157 languages
- Individual language Vector Spaces are *unique*: not directly comparable
- In 2017 Babylon Health produced a paper showing a possible way of '*normalizing*' the Vector Space between 2 languages



XTM International

Word Similarities

Inter-language Vector Space

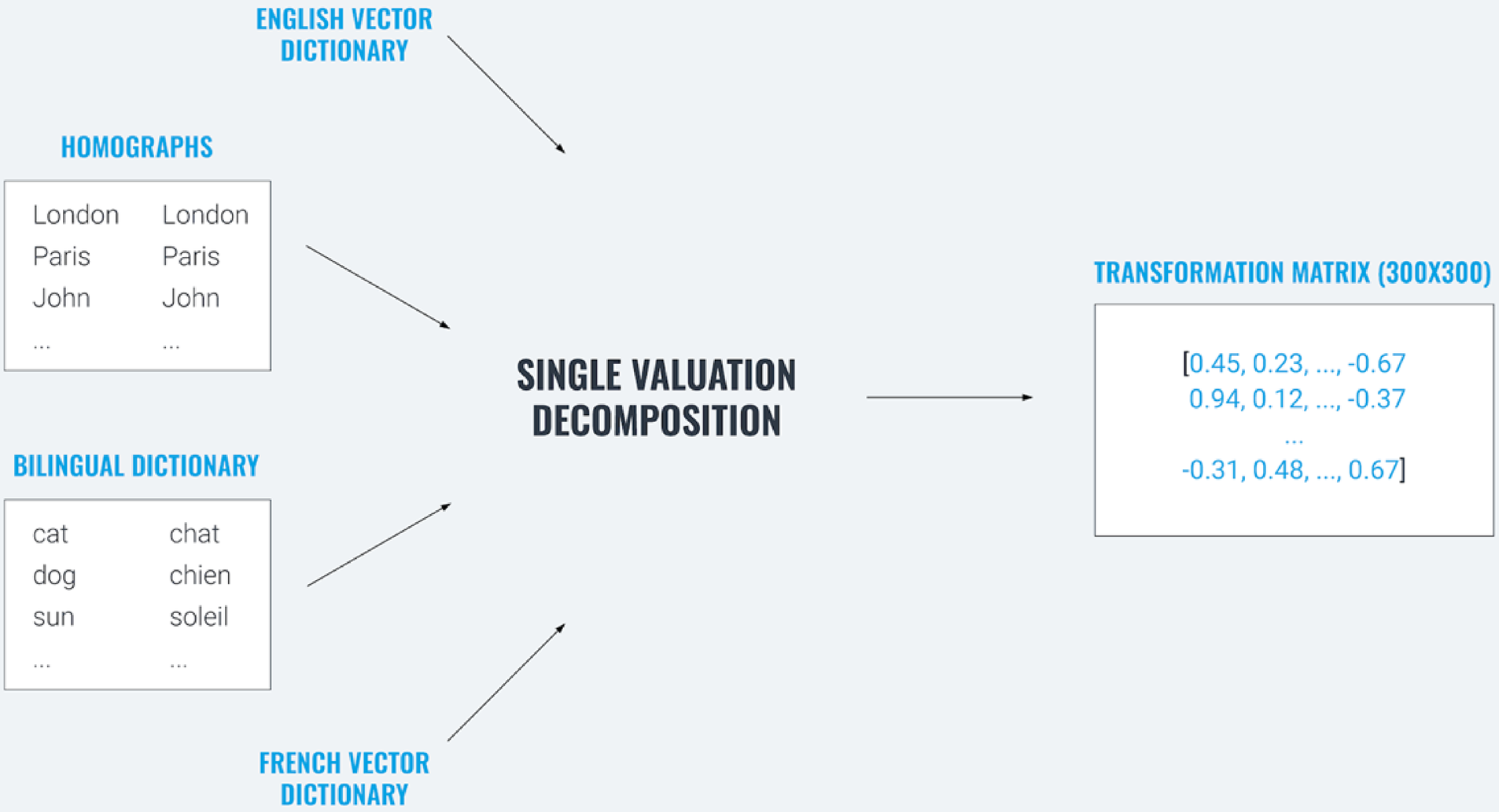
Powered by XTM AI

- Unique XTM development/contribution
- Normalized Vector Space for 50 languages onto the same plane
- Uses the XTM bilingual dictionaries generated from BabelNet and other online resources
- Patent applied for
- Allows for a probability score for each target word as a translation of a source word in a given segment

XTM International

Word Similarities

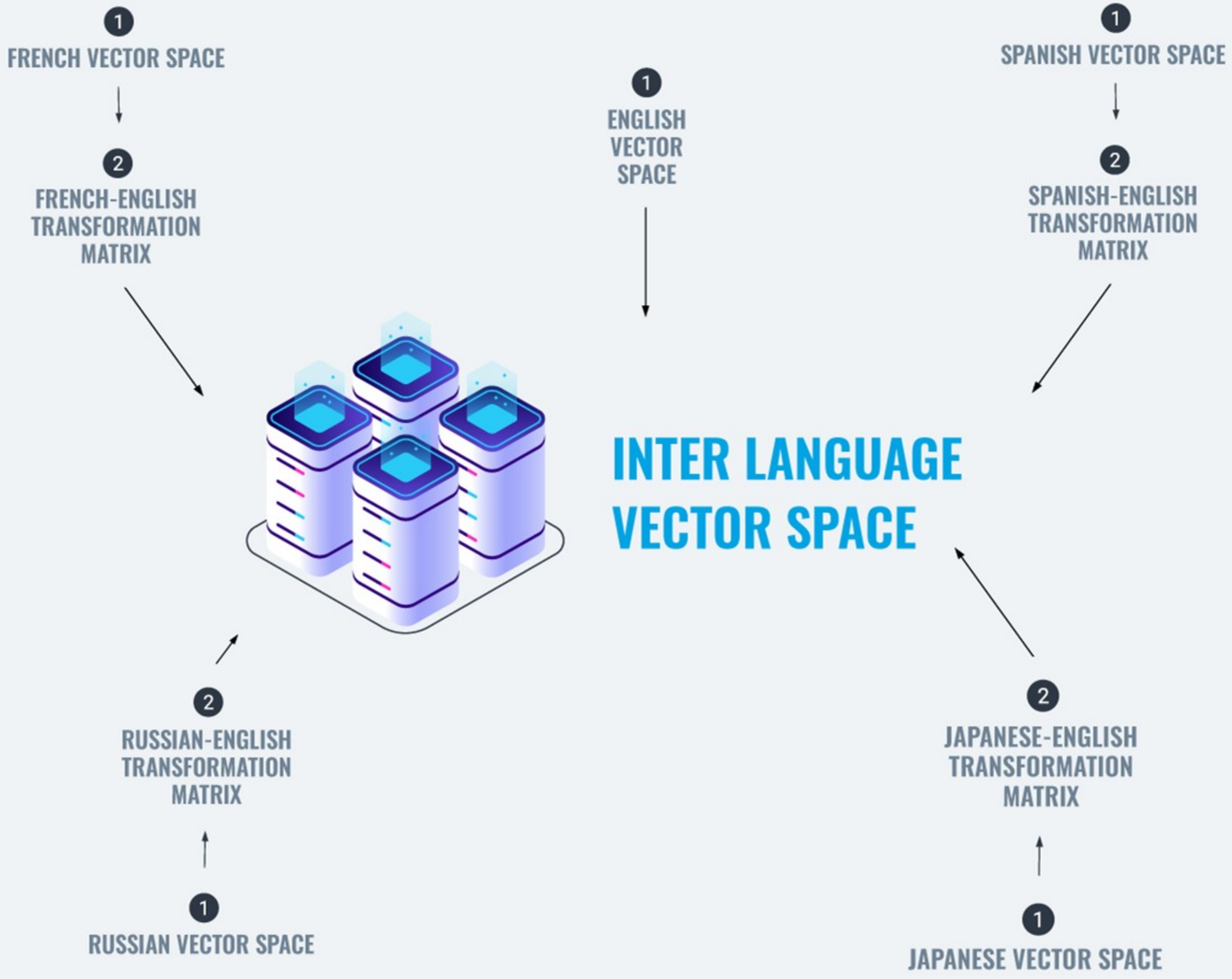
Inter-language Vector Space | Transformation Matrix



XTM International

Word Similarities

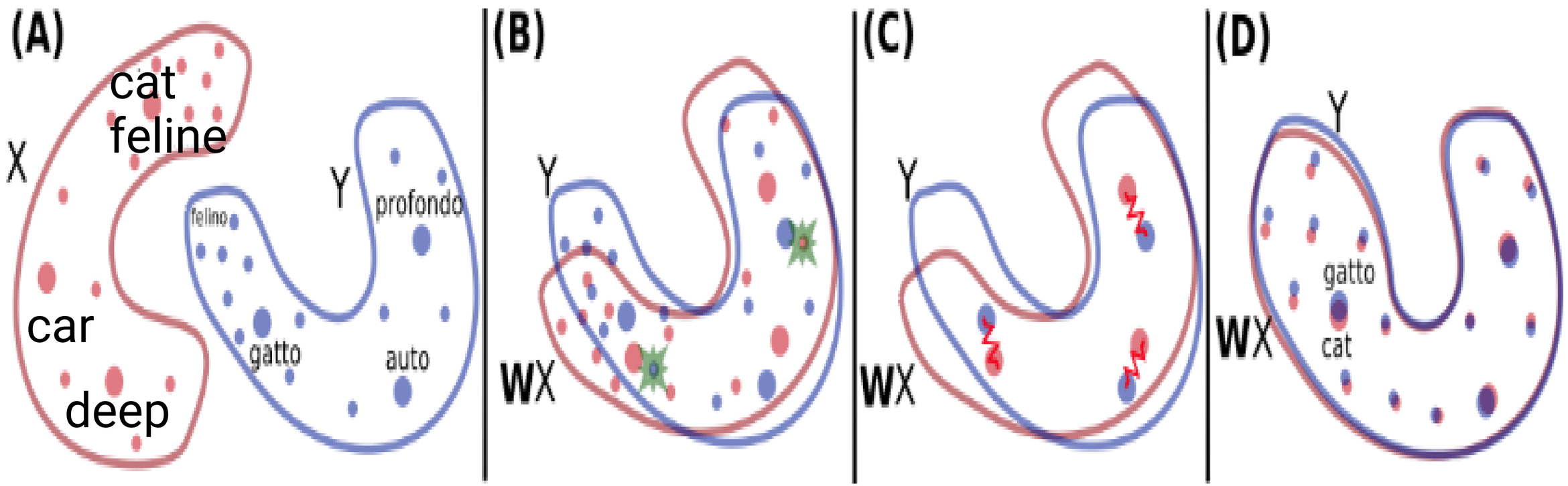
Inter-language Vector Space | Transformation Matrix



XTM International

Word Similarities

Inter-language Vector Space



Word Similarities

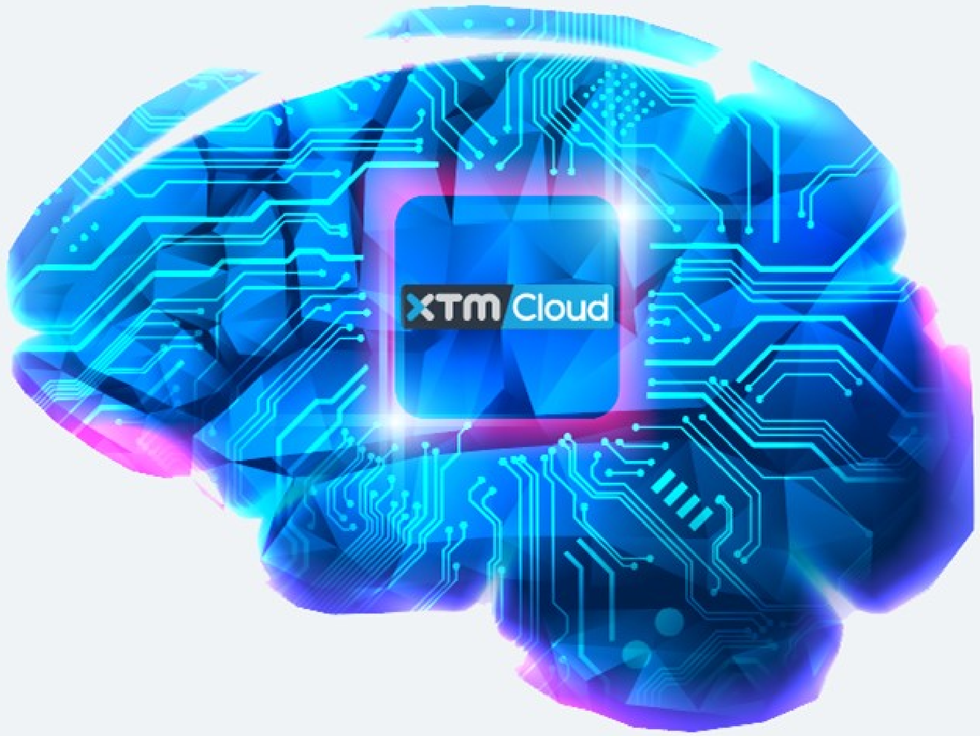
Inter-language Vector Space

English word	Italian word	Similarity
cat	gatto	0.696
cat	gatta (female cat)	0.552
cat	giorno (day)	0.164 (as expected, low similarity)
day	giorno	0.692
day	fuoco (fire)	0.193 (as expected, low similarity)
fire	fuoco	0.590

Word Similarities

Inter-language Vector Space | Recap

- Advanced new direction for linguistic AI
- Based on neural network analysis of vast amounts of textual data: *comprehensive language models*
- XTM uses a crawl of the whole Internet for each language – terabytes of data

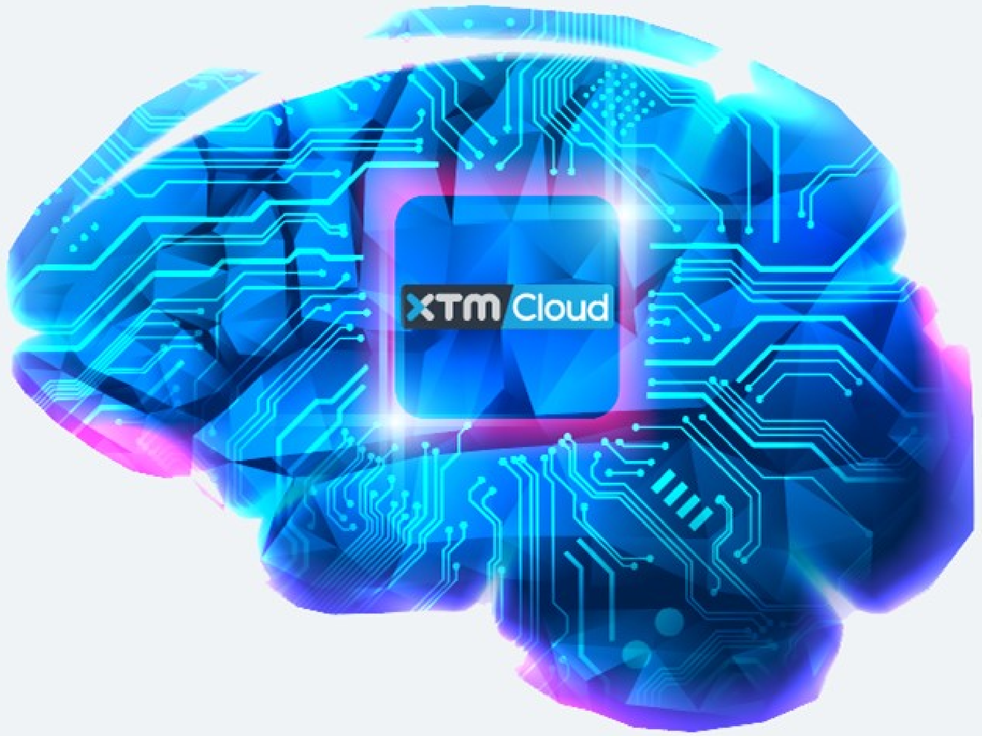


XTM International

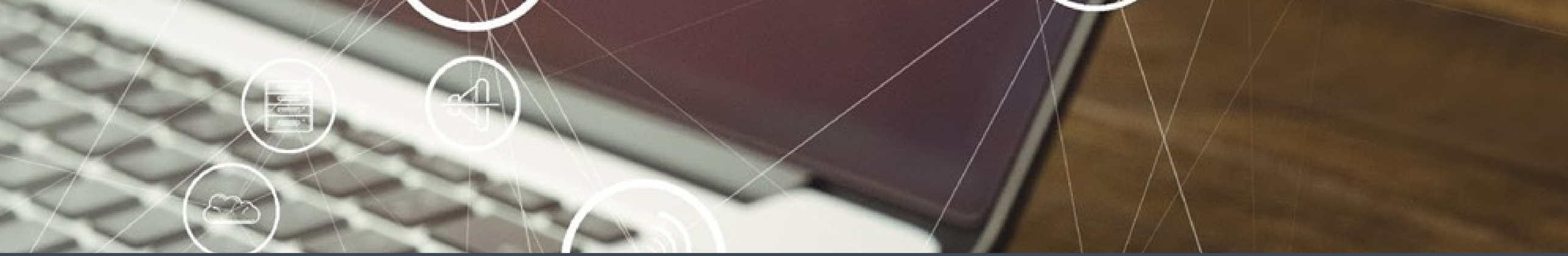
Word Similarities

Inter-language Vector Space | Recap

- Mono-lingual Vector Space calculates the relationships of words that can then be used to answer the following type of question:
  - If king is to man, then what is the equivalent for woman?
  - If Einstein was a scientist then what was Mozart?
- Each language produces a unique mono-lingual Vector Space
- XTM is able to ‘normalize’ the values so that they are comparable between two languages. This is unique:
  - This allows for a probabilistic assessment regarding the equivalence of words and phrases between a source and target language segment



XTM International



FREE 30 day trial account

www.xtm.cloud/trial

XTM International

More information

www.xtm.cloud
training@xtm.cloud