Word2Vec

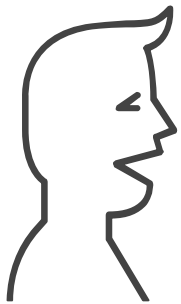
O caso mais famoso de word embeddings

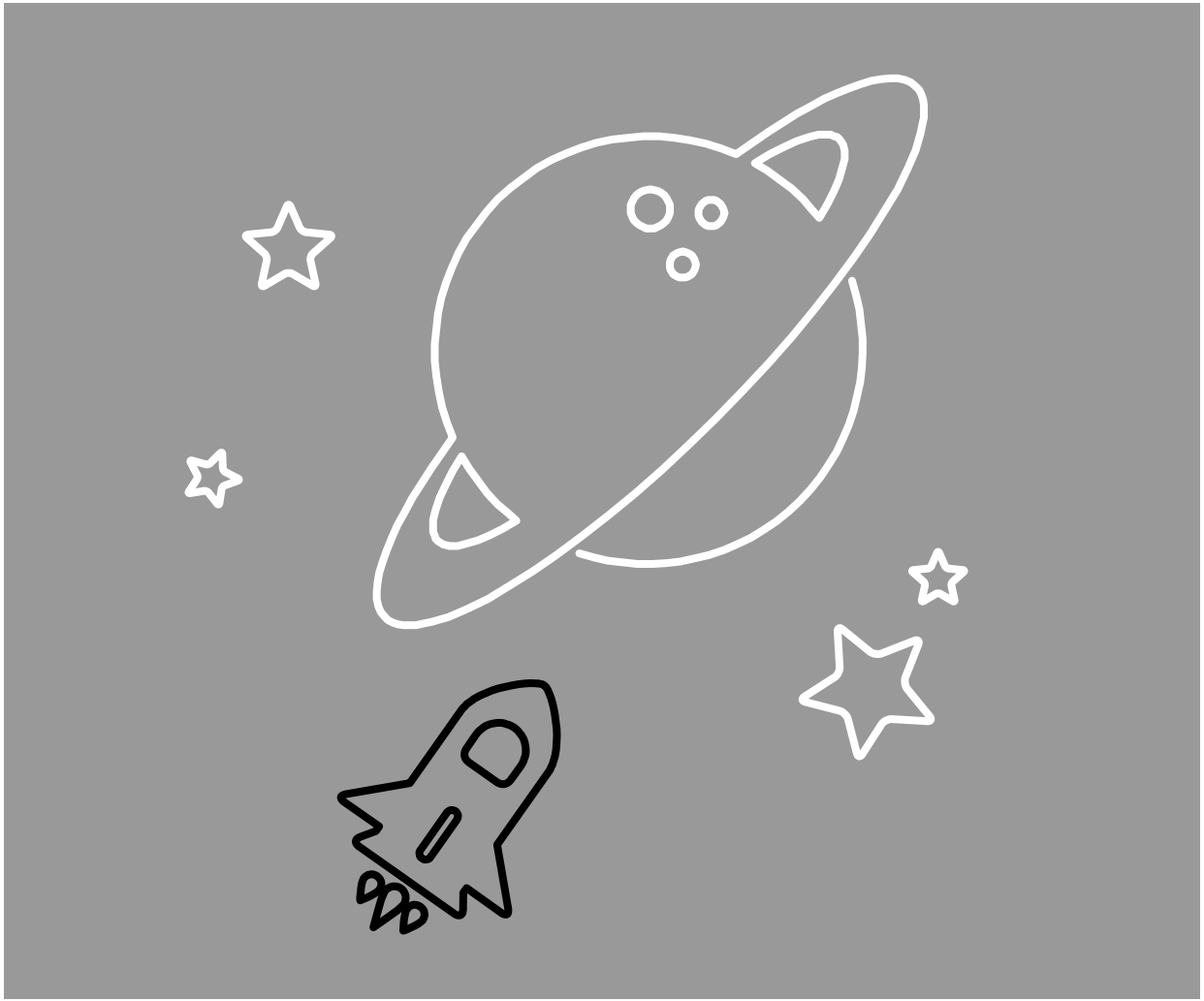
Larissa Sayuri Futino Castro dos Santos

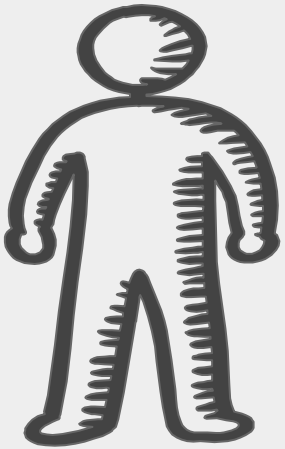
18 maio 2019

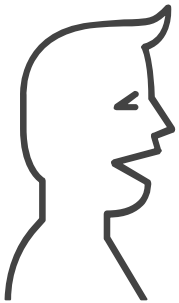
Parece mágica!

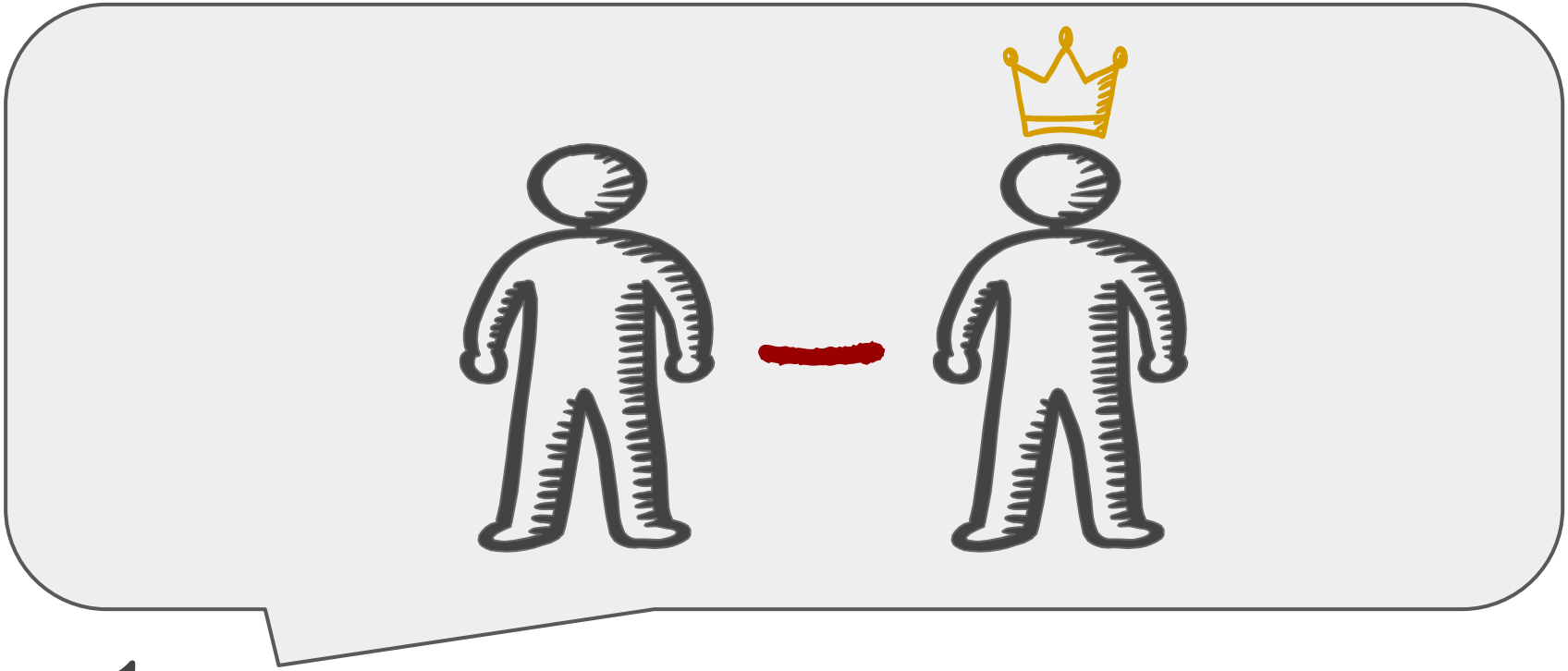
Usa matemática,
Jeitão de linguística

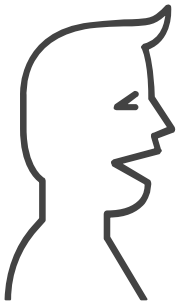


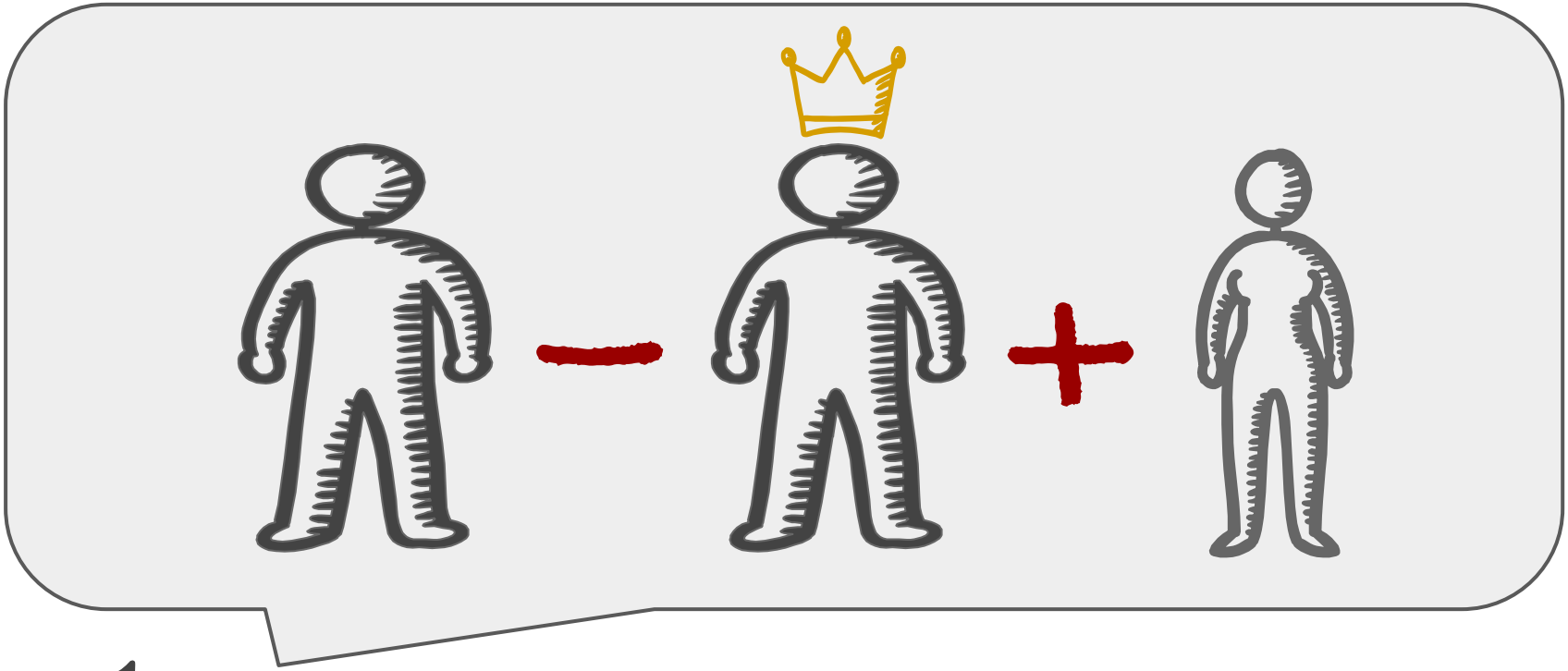


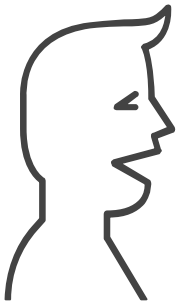


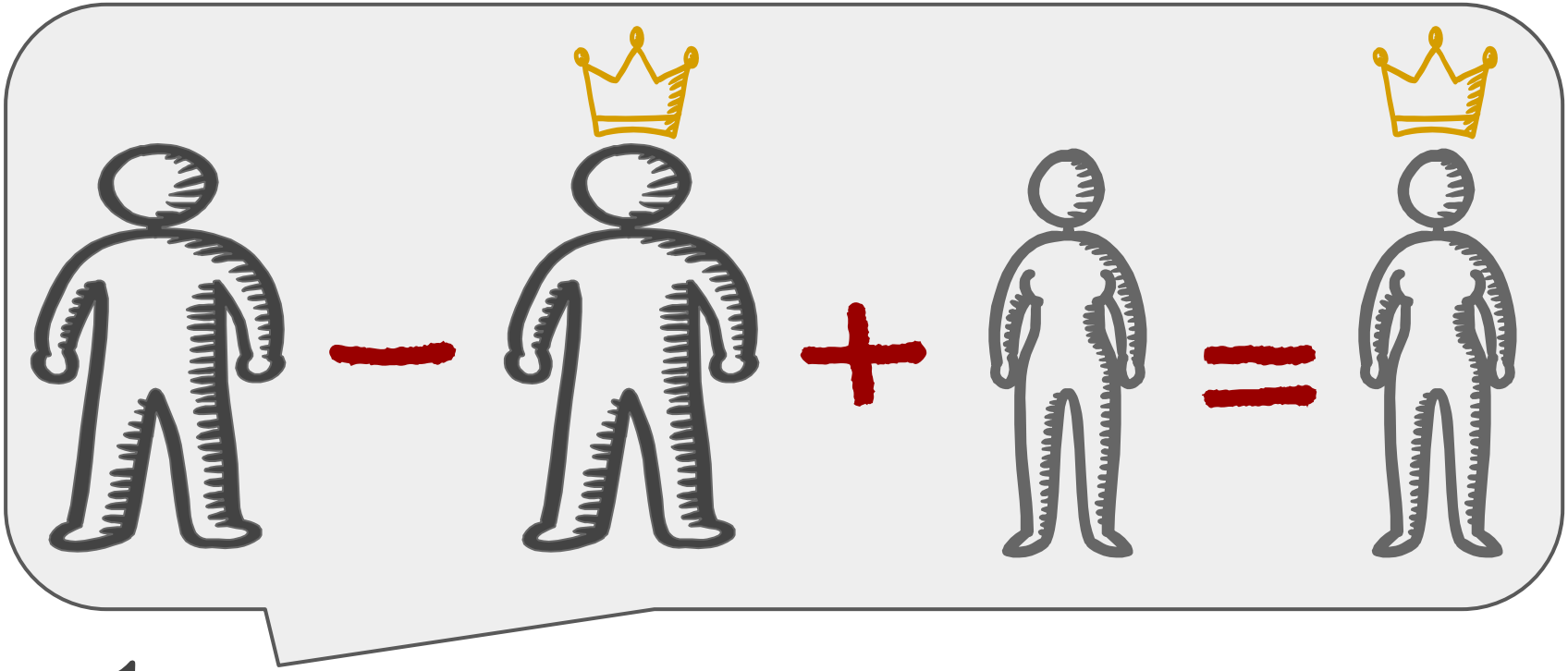


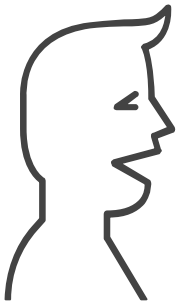




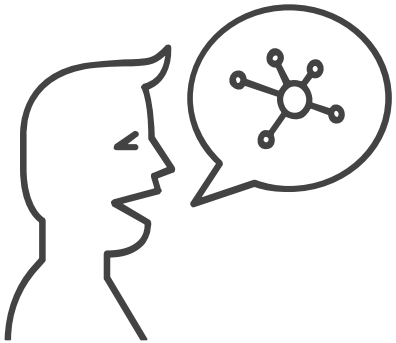




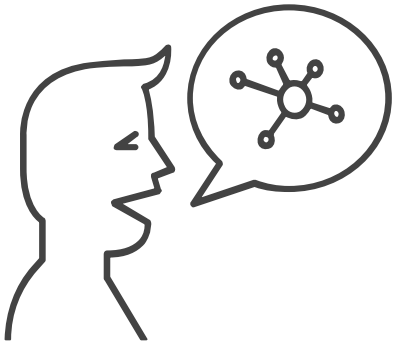


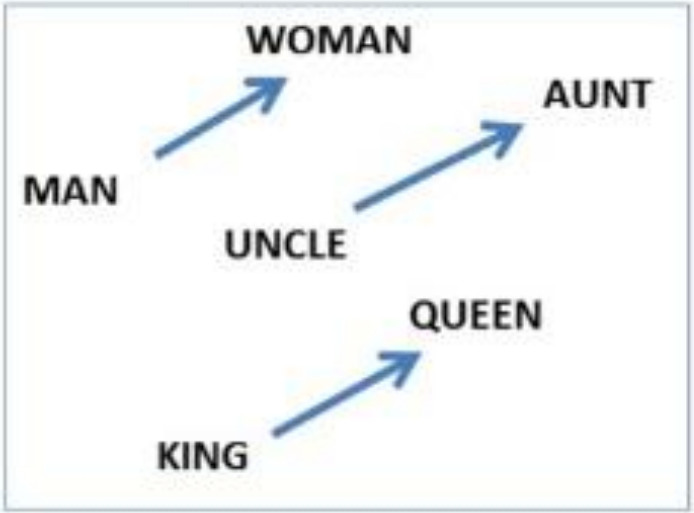


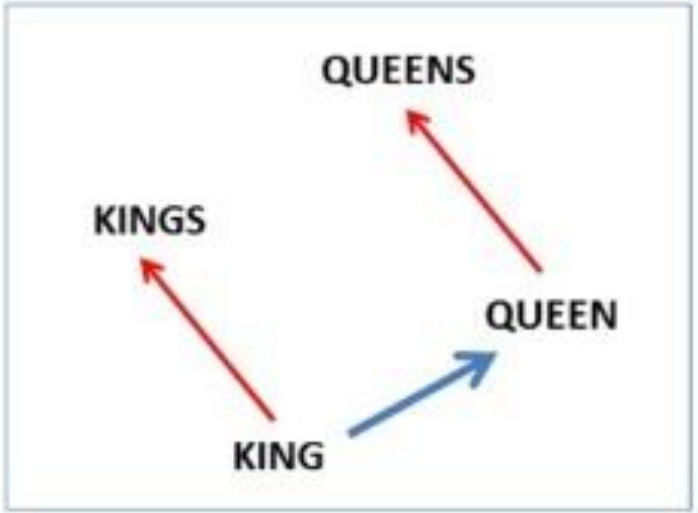
vec(“man”) – vec(“king”) + vec(“woman”) = vec(“queen”)

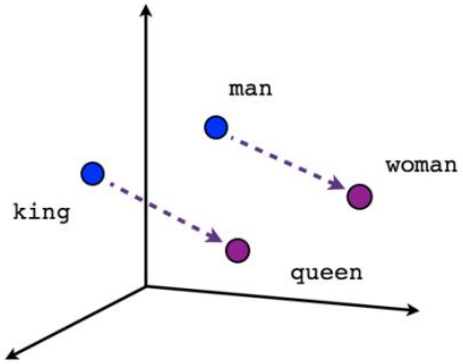


vec(“man”) – vec(“king”) + vec(“woman”) = vec(“queen”)

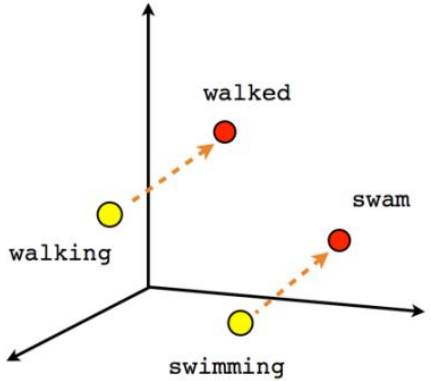




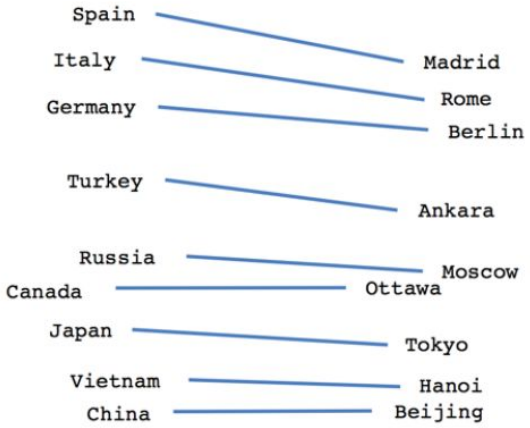




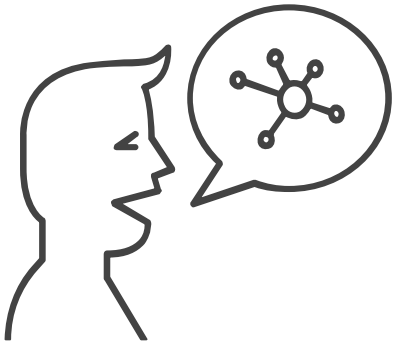
Male-Female

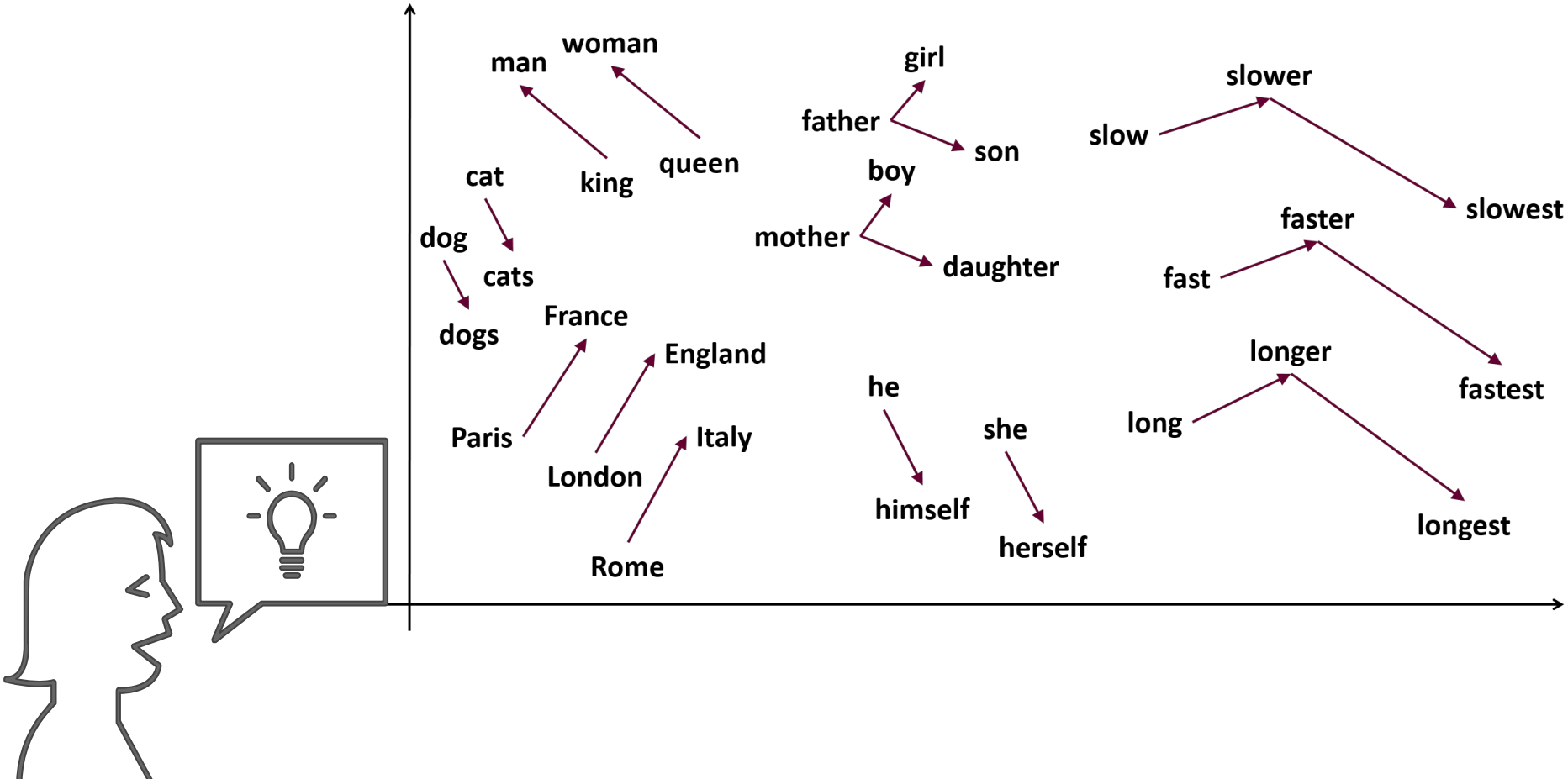


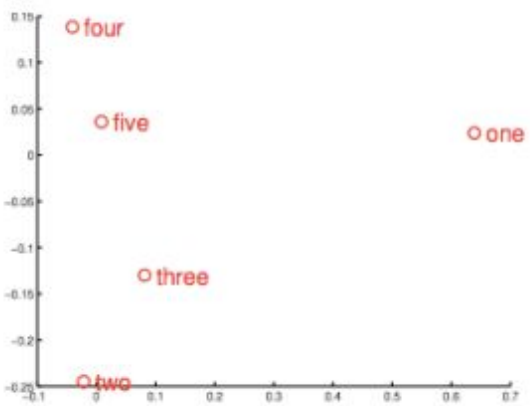
Verb tense

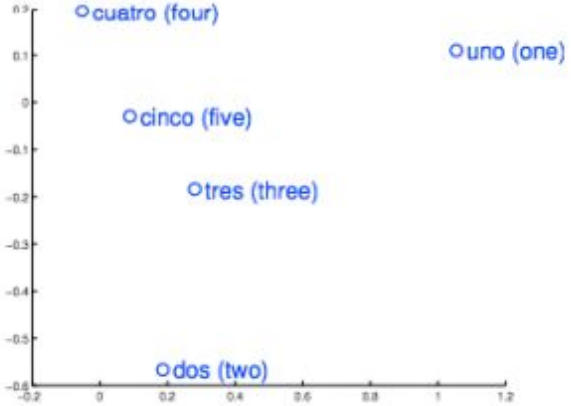


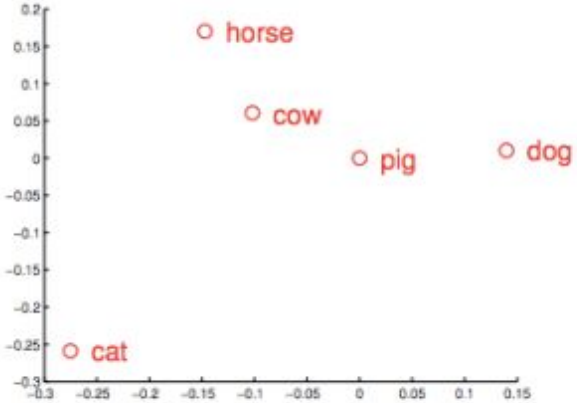
Country-Capital



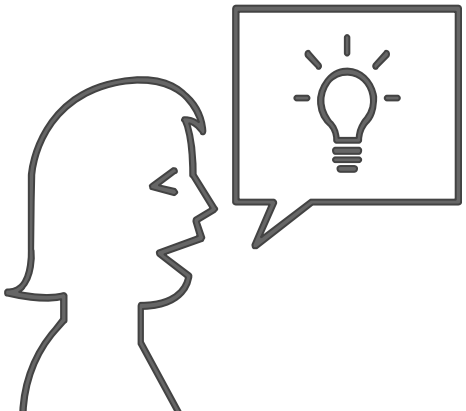

man
woman
king
queen
cat
cats
dog
dogs
France
Paris
England
London
Italy
Rome
girl
father
son
boy
mother
daughter
he
himself
she
herself
slow
slower
slowest
fast
faster
fastest
long
longer
longest


four
five
one
three
two


cuatro (four)
uno (one)
cinco (five)
tres (three)
dos (two)


horse
cow
pig
dog
cat


caballo (horse)
vaca (cow)
perro (dog)
cerdo (pig)
gato (cat)



Trabalhos

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MIKOLOV, Tomas; YIH, Wen-tau; ZWEIG, Geoffrey. Linguistic regularities in continuous space word representations. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. 2013. p. 746-751.

MIKOLOV, Tomáš et al. Recurrent neural network based language model. In: **Eleventh Annual Conference of the International Speech Communication Association**. 2010.

BENGIO, Yoshua et al. A neural probabilistic language model. **Journal of machine learning research**, v. 3, n. Feb, p. 1137-1155, 2003.


EU FAÇO PARTE
de uma das 100 empresas
mais inovadoras do país.
100
STARTUPS
TO WATCH
maio
MARKETING TECHNOLOGY AND SERVICE


MAIO
2019

Doutorado Sanduíche
UFMG/CAPES/
University of
Wolverhampton

2017 2018

Mestrado UFMG

CAPES

Doutorado UFMG

2013

2015

Bacharelado
UNB

2007

2010

About: MIKOLOV, Tomas

Tomas Mikolov

A grande referência

42.756 citações (17/05/19)

Doutorado Brno University of Technology (Czech Republic)

Bons exemplos

Exemplos com diferentes formas de similaridade (semântica e sintática)

Popular

Google desenvolveu uma ferramenta, tensorflow, e a disponibilizou

Grandes datasets

A técnica lida bem com bases grandes, típicas da realidade atual
(6B tokens, 1M termos mais freq)

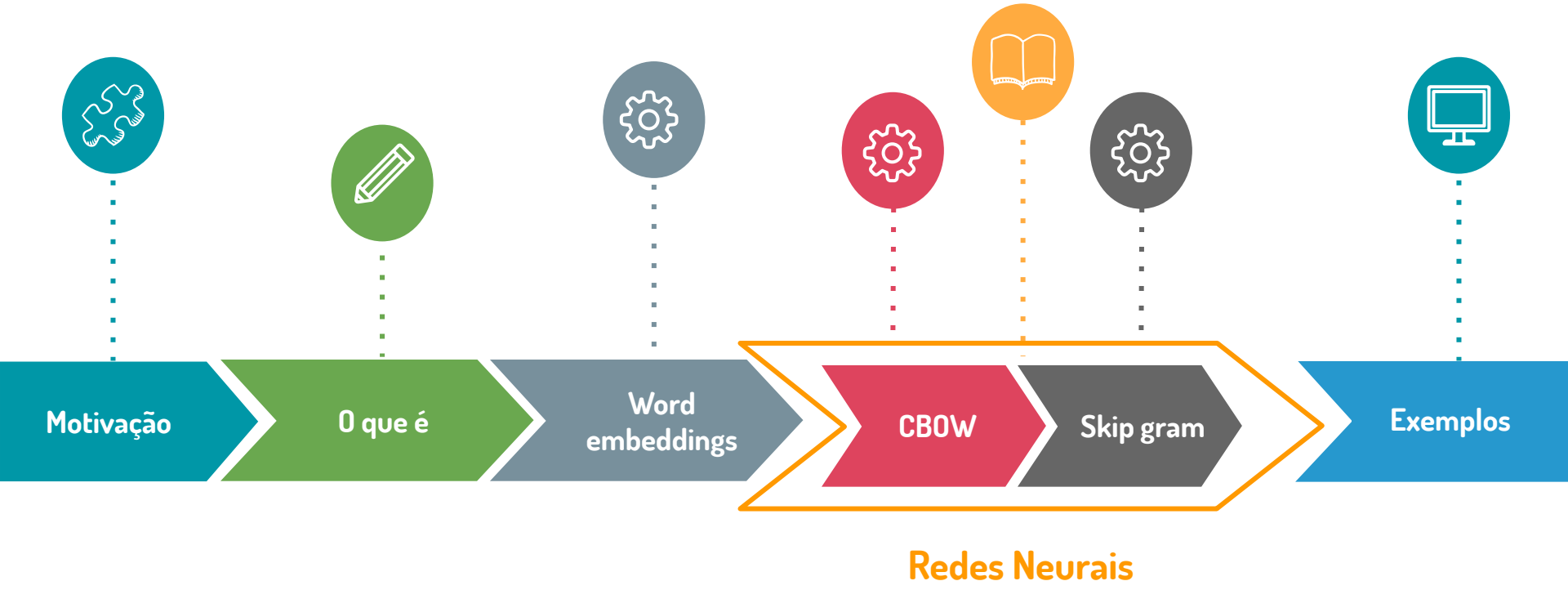
RÁPIDO

De: 08 semanas
Para: 01-03 dias

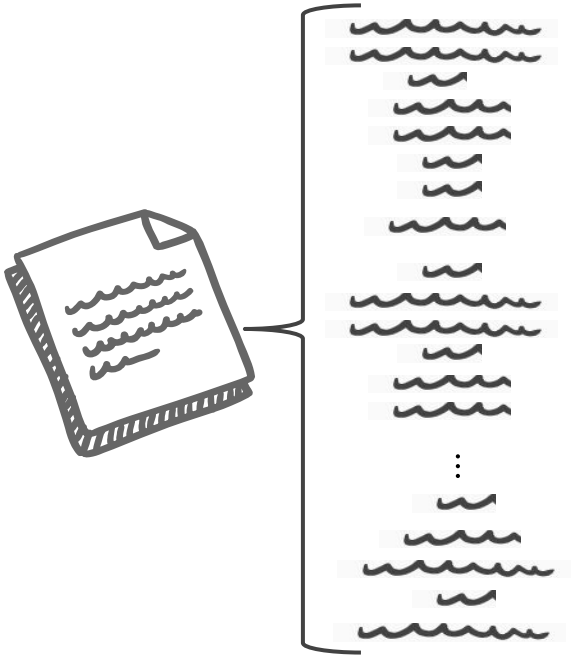
Relações preservadas entre línguas

Permite tradução termo-a-termo

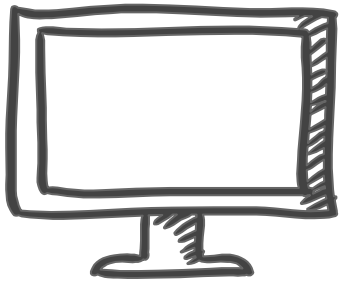
Agenda


Motivação
O que é
Word embeddings
CBOW
Skip gram
Exemplos
Redes Neurais

Linguagem de Máquina

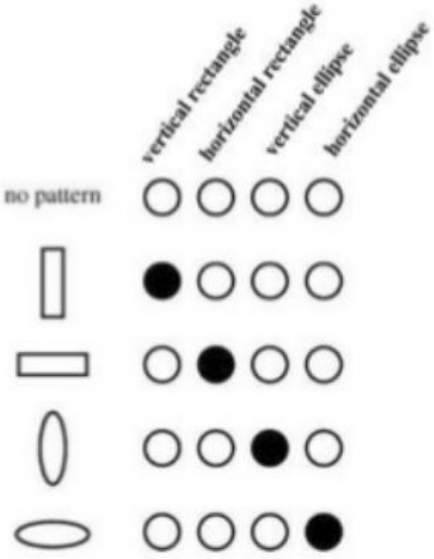


Representação do texto



Linguagem de Máquina

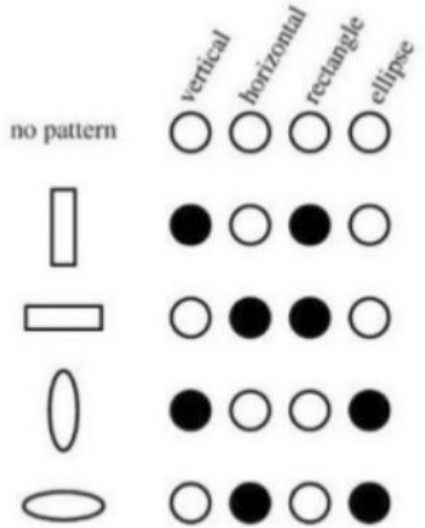
sparse representations



"one hot" vector

A distributed representation is dense

1.One concept is represented by more than one neuron firing
2.One neuron represents more than one concept



Fonte: slideshare_UnivGothenburg

O QUE FAZ?

Cria vetores que são representações numéricas dos atributos (termos/palavras) sem intervenção humana

Objetivo

Criar representação de termos (palavras) que capte seus significados, relações semânticas e os diferentes tipos de contexto em que são empregadas

Objetivo (geometria)

Posiciona termos (palavras) em um espaço dimensional tal que as localizações dos termos são determinadas pelos seus significados e os vazios e distâncias no espaço também tem significado

Como funciona?

Rede neural de duas camadas que processa texto

Para quê serve?

Cria representação de strings para algoritmos de Machine Learning ou Deep Learning

Resultado

Agrupar vetores de termos similares no espaço de atributos

Word Embeddings

Definição:

Representações numéricas para textos

Textos convertidos em números

D1: He is a lazy boy. She is also lazy.

D2: Neeraj is a lazy person.

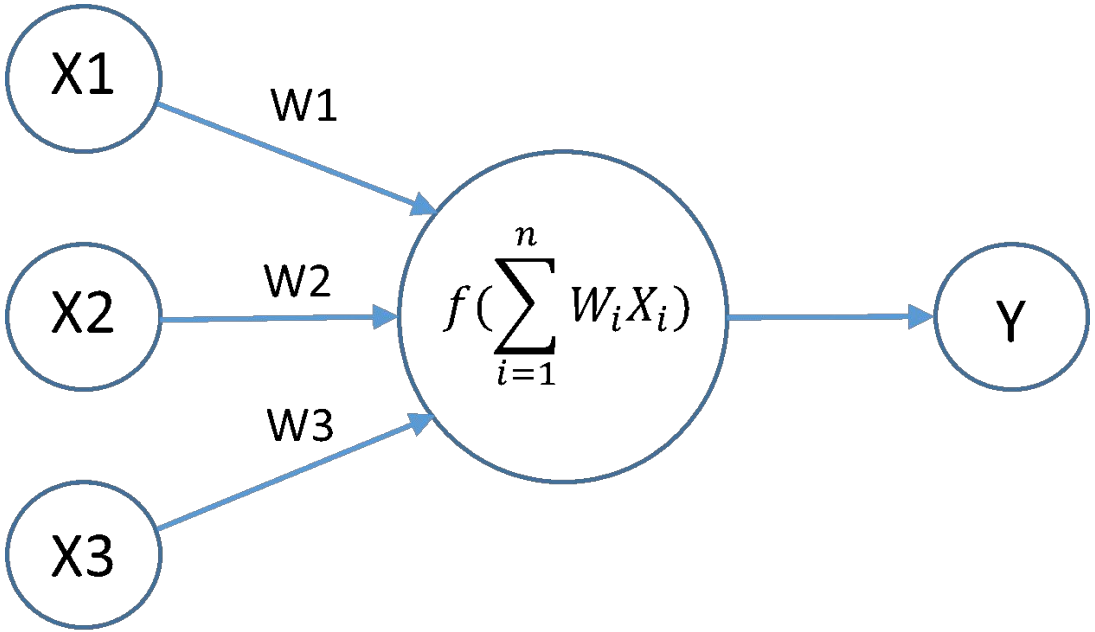
	He	She	lazy	boy	Neeraj	person
D1	1	1	2	1	0	0
D2	0	0	1	0	1	1

Fonte: analyticsvidhya

Redes Neurais

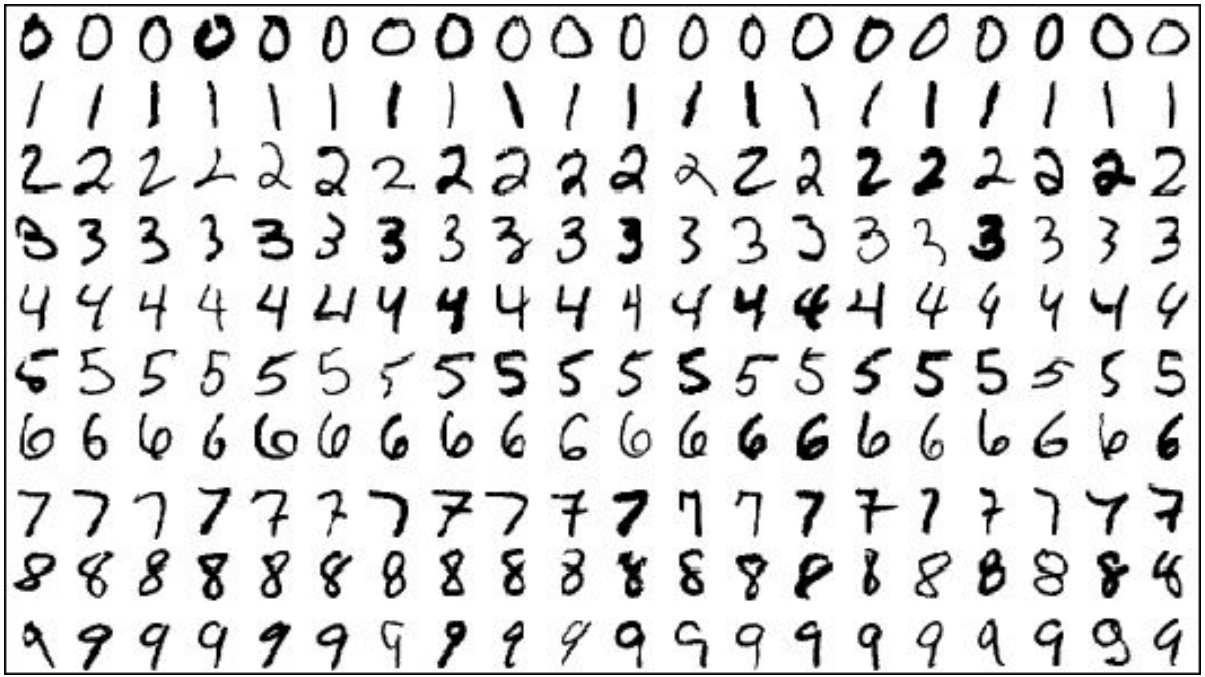
Definição:

Técnicas **computacionais** inspiradas na estrutura neural de organismos inteligentes e que **adquirem conhecimento através da experiência**

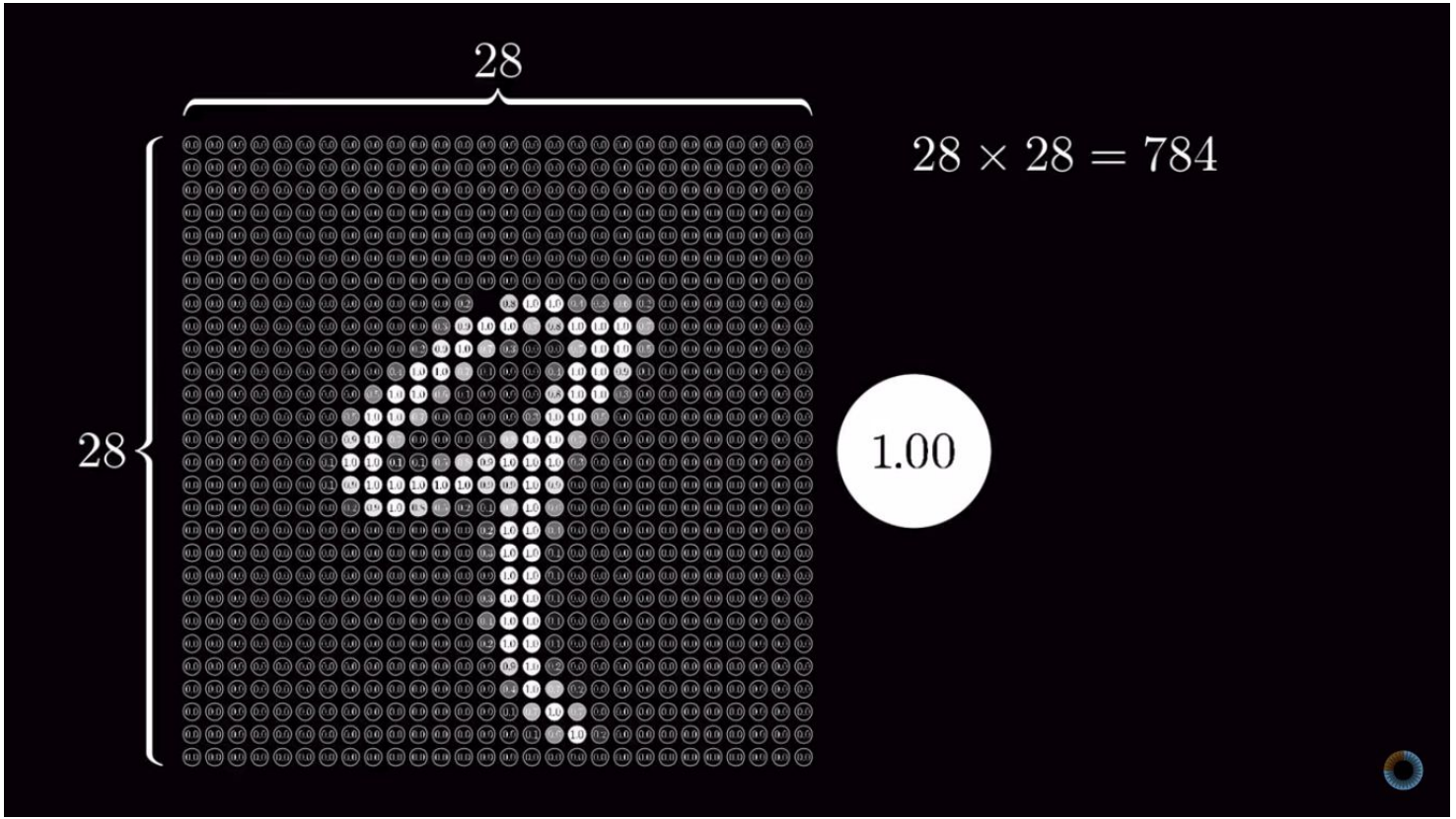


Redes Neurais

Motivação: Um modelo para reconhecer dígitos numéricos escritos à mão.

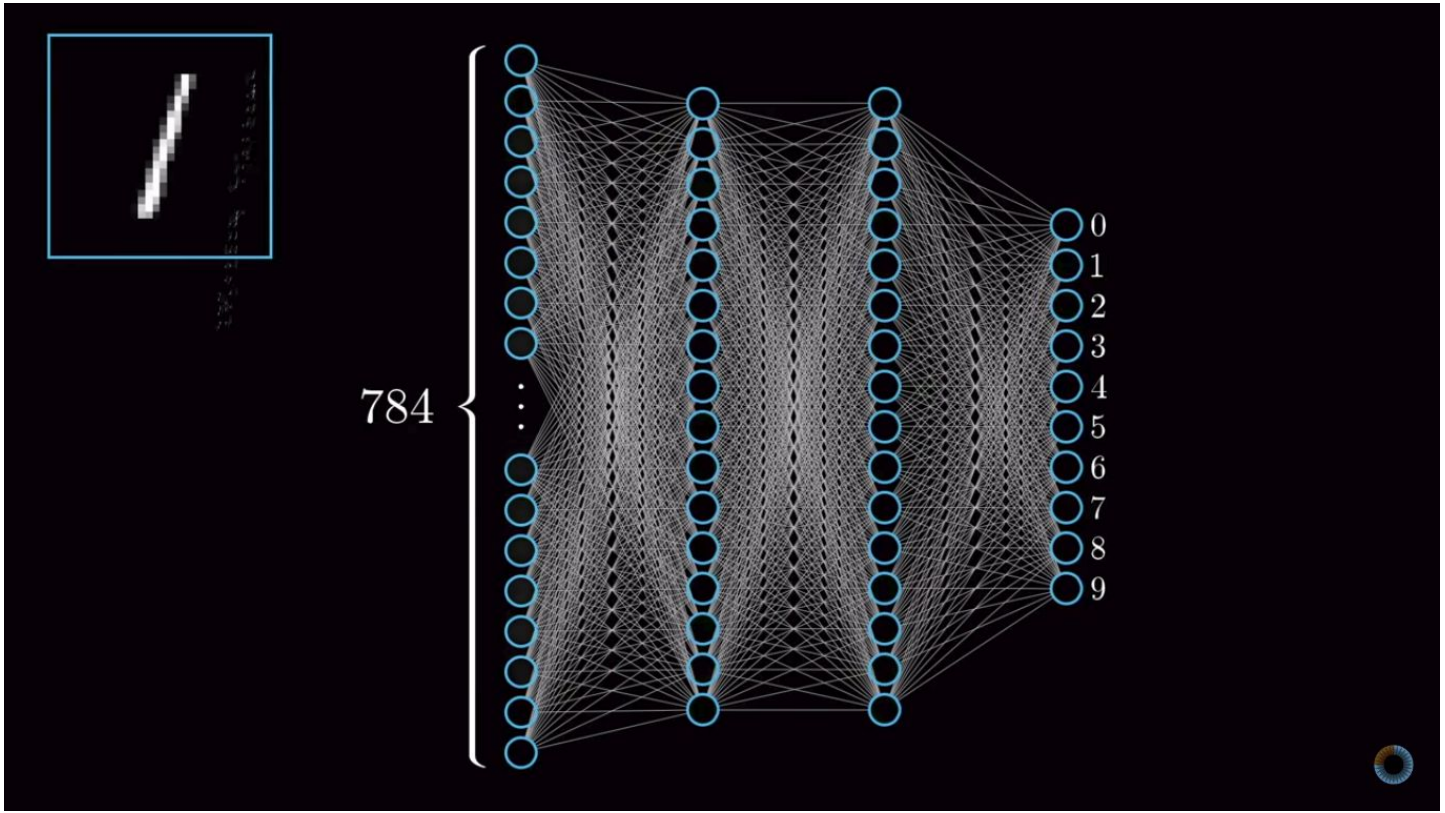


Redes Neurais – Dígitos escritos à mão



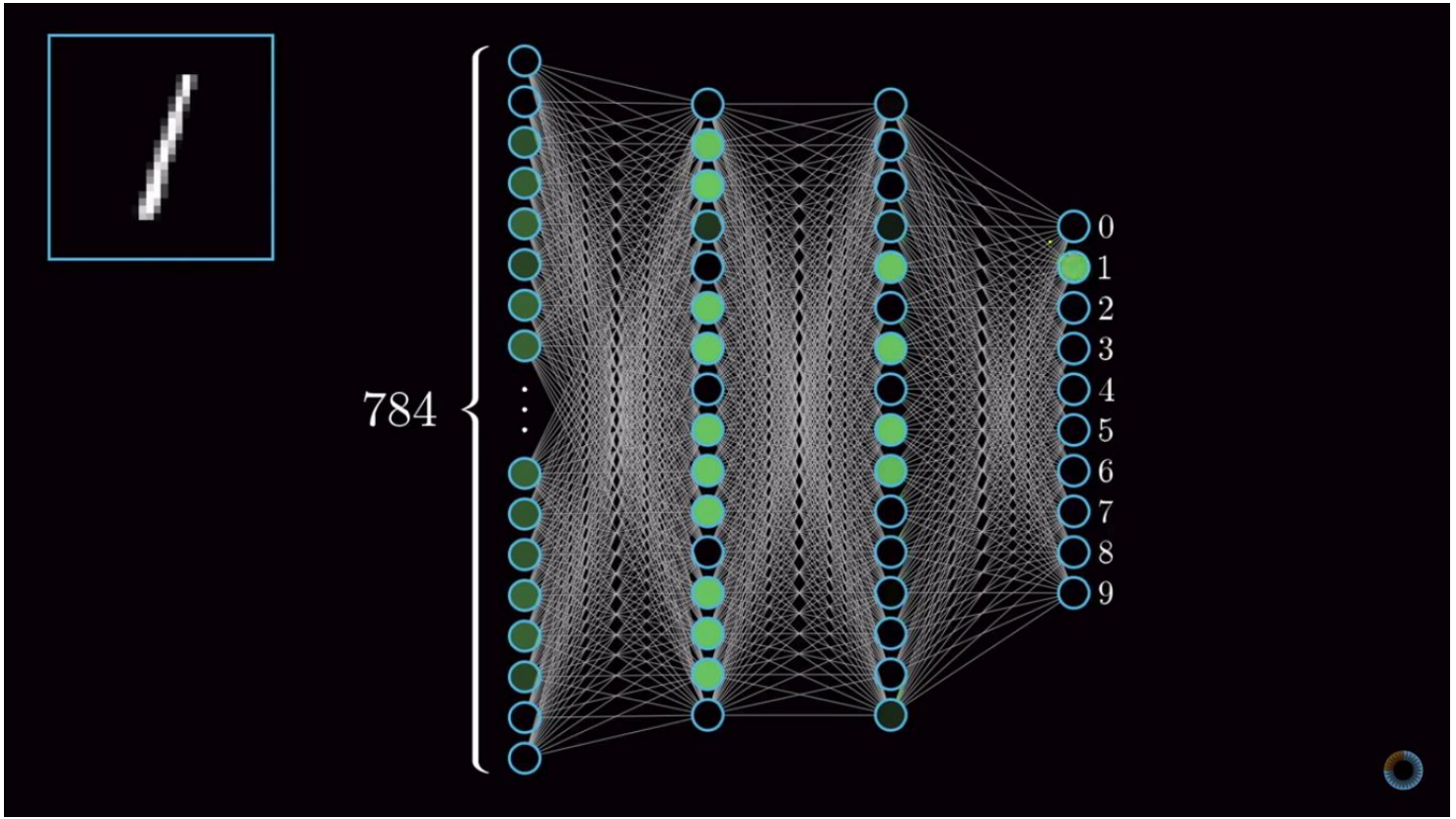
Fonte: 3blue1brown

Redes Neurais – Dígitos escritos à mão

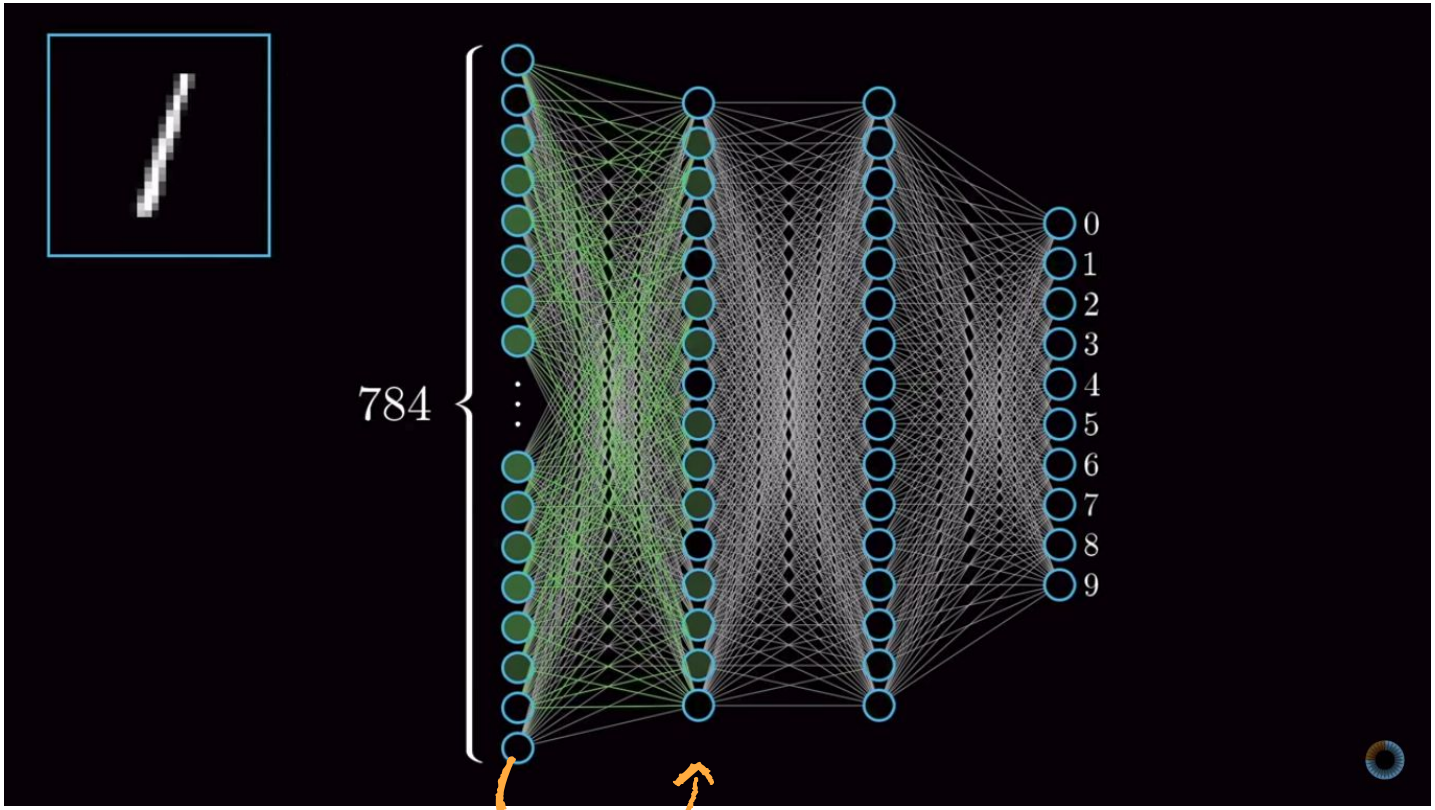


Fonte: 3blue1brown

Redes Neurais – Dígitos escritos à mão

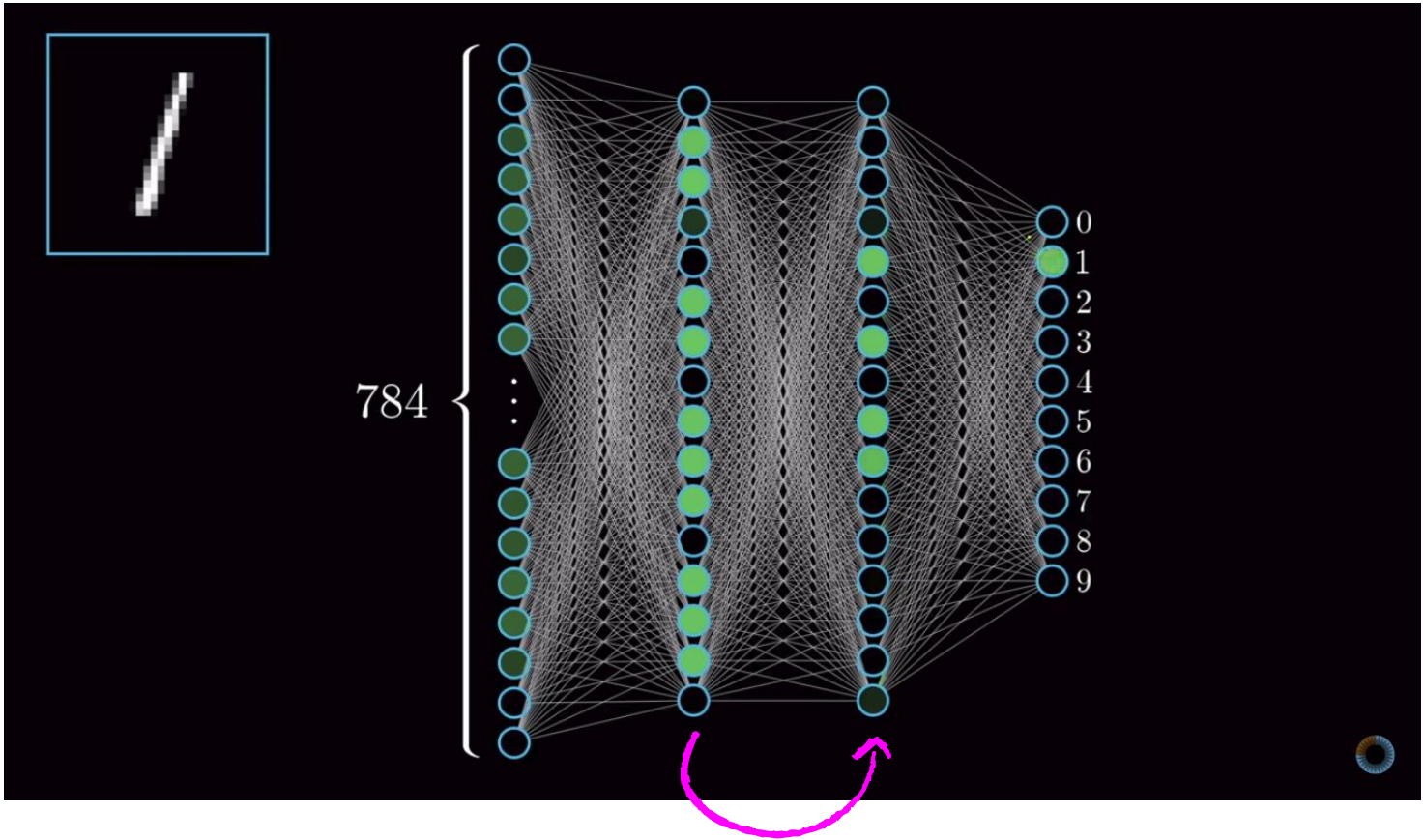


Fonte: 3blue1brown



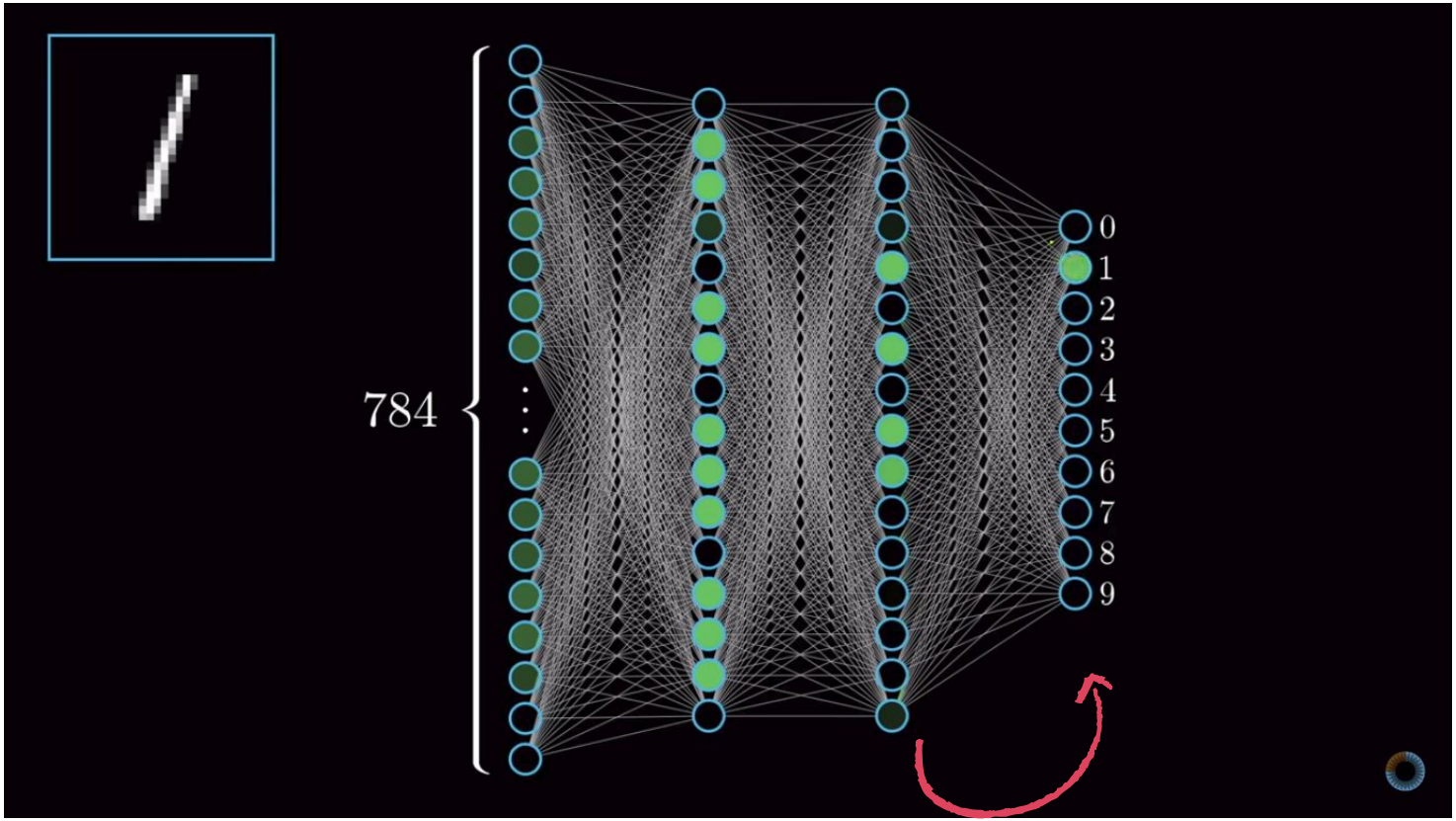
Um padrão de neurônios na camada inicial ativa um padrão de neurônios na 1ª camada oculta

Fonte: 3blue1brown



Um padrão de neurônios na 1ª camada oculta ativa um padrão de neurônios na 2ª camada oculta

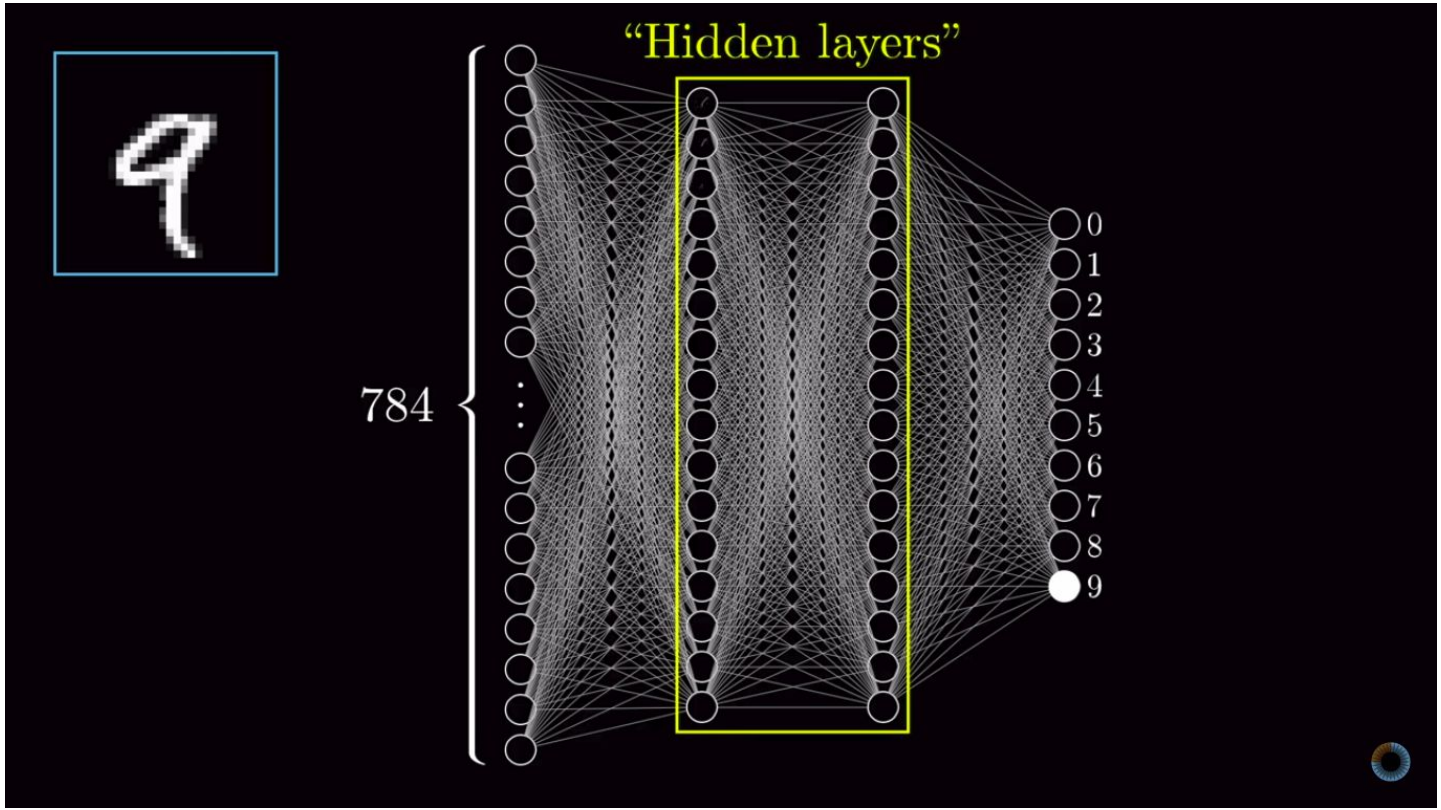
Fonte: 3blue1brown



Um padrão de neurônios na 2ª camada oculta ativa um label final

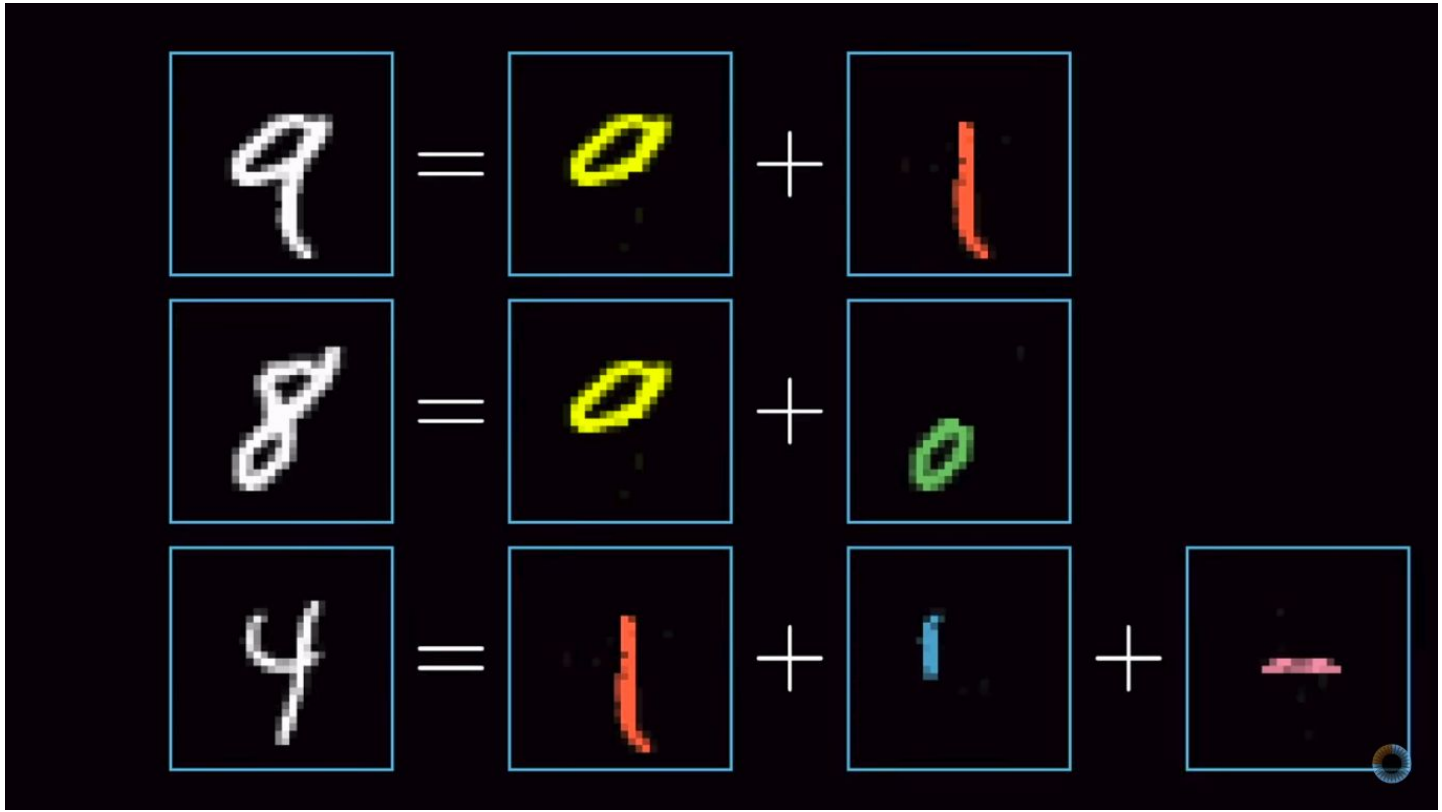
Fonte: 3blue1brown

Redes Neurais – Camadas ocultas



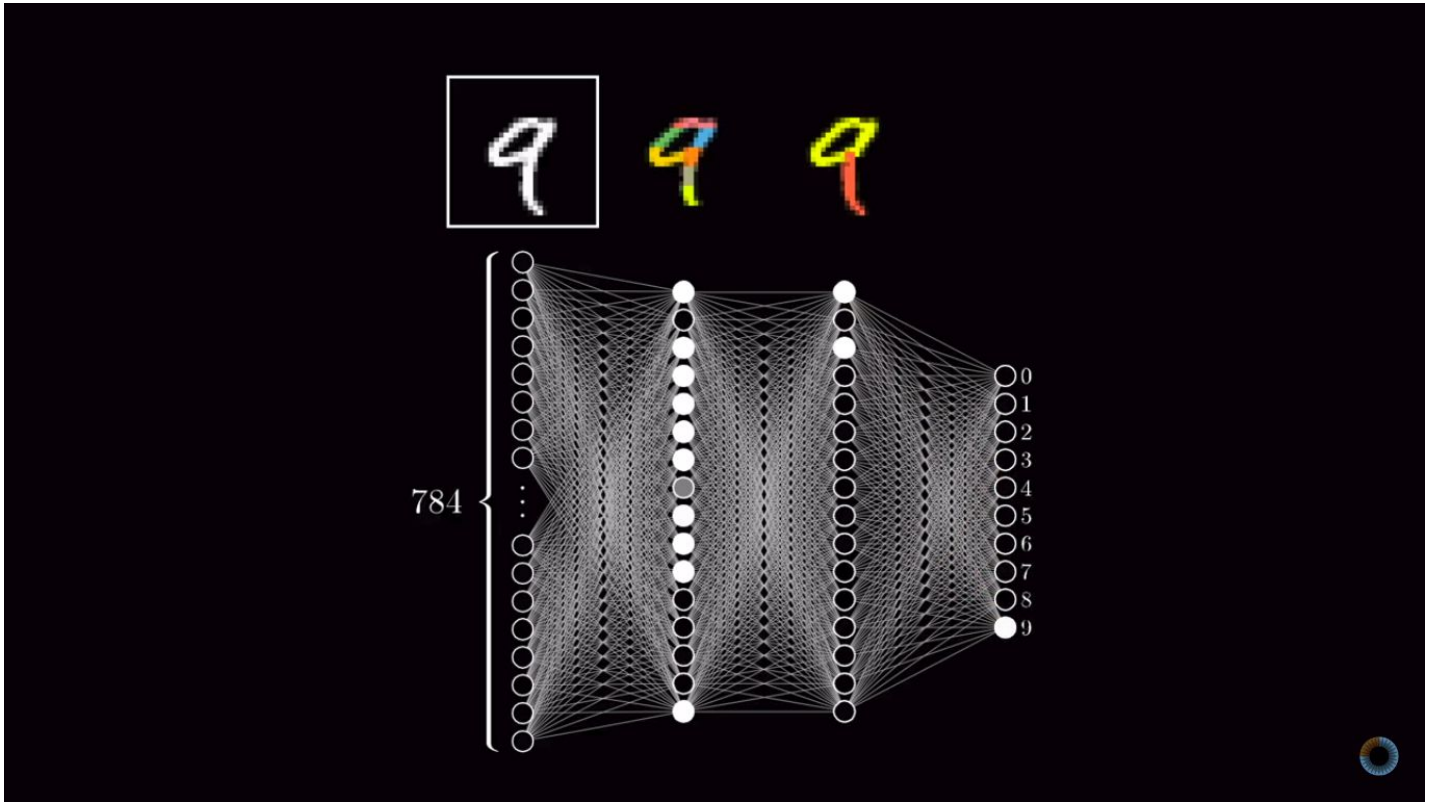
Fonte: 3blue1brown

Redes Neurais – Camadas ocultas



Fonte: 3blue1brown

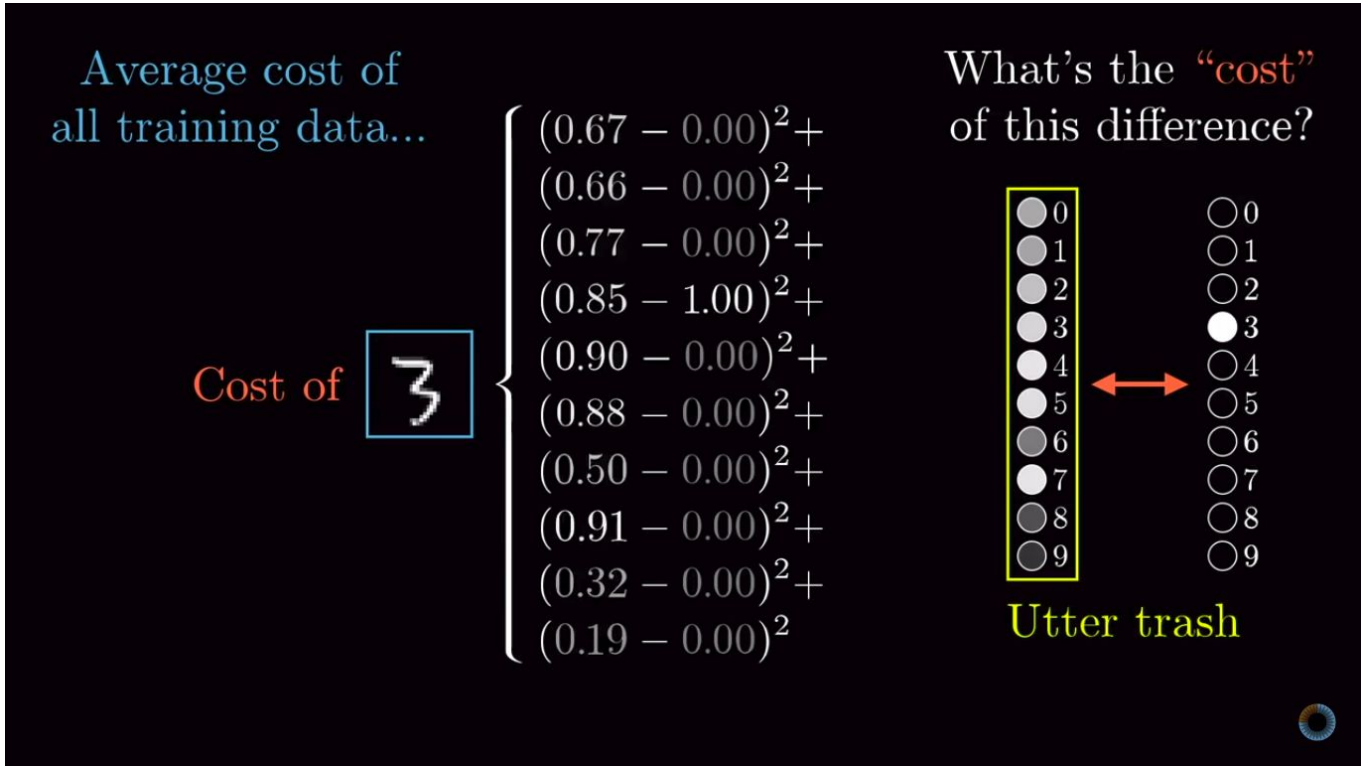
Redes Neurais – Camadas ocultas



(forte) Suposição: a 1ª camada reconhece vértices,
a 2ª camada reconhece padrões de formas

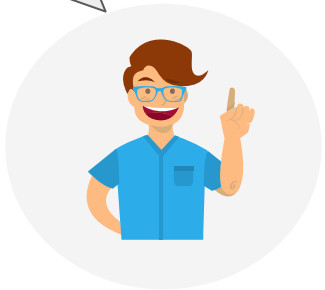
Fonte: 3blue1brown

Redes Neurais - Erro/Custo



Existe um rótulo para comparar.

Trata-se de uma abordagem supervisionada



Erro: diferença de ativação na camada final e como deveria ser a ativação para classificação correta

Fonte: 3blue1brown

Word2Vec

Estamos interessados na noção de **significado**.

Vamos aprender os significados das palavras com uma ideia já usual:

A partir do **contexto**

CBOW

Skip gram

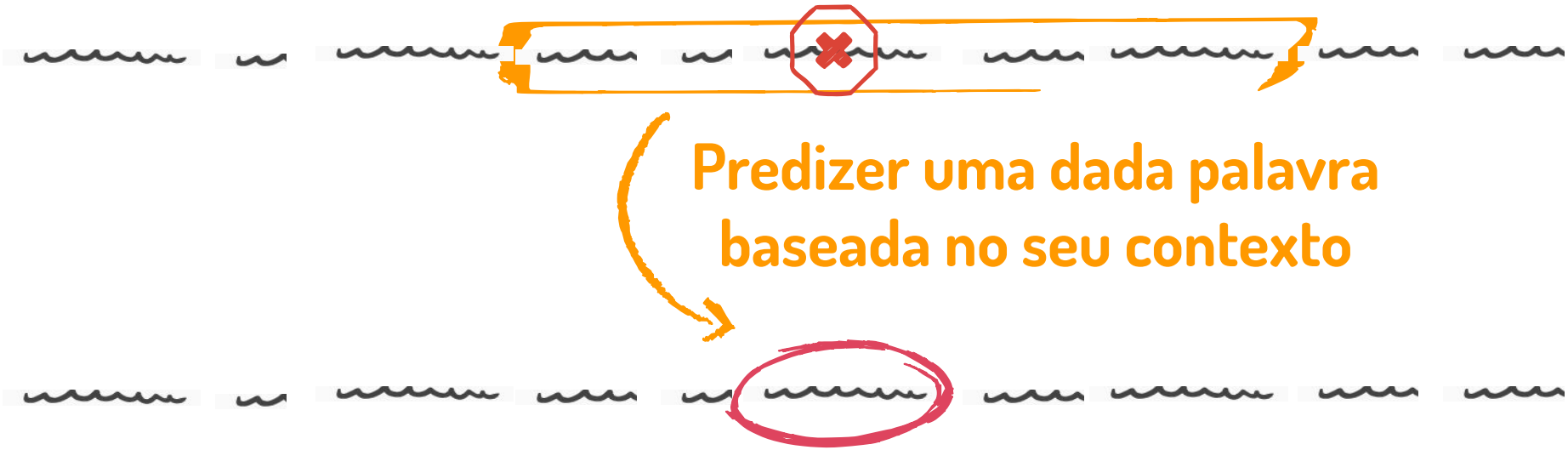
Word2Vec

CBOW

Skip gram

Fonte: Artigo 2013

Word2Vec

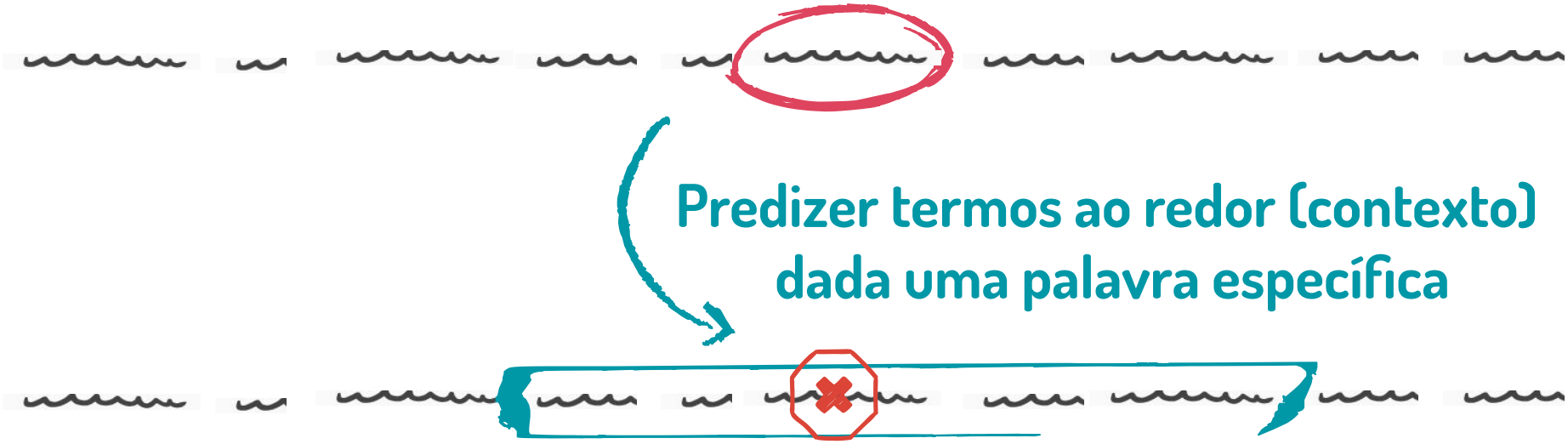


CBOW

Skip gram

Fonte: Artigo 2013

Word2Vec

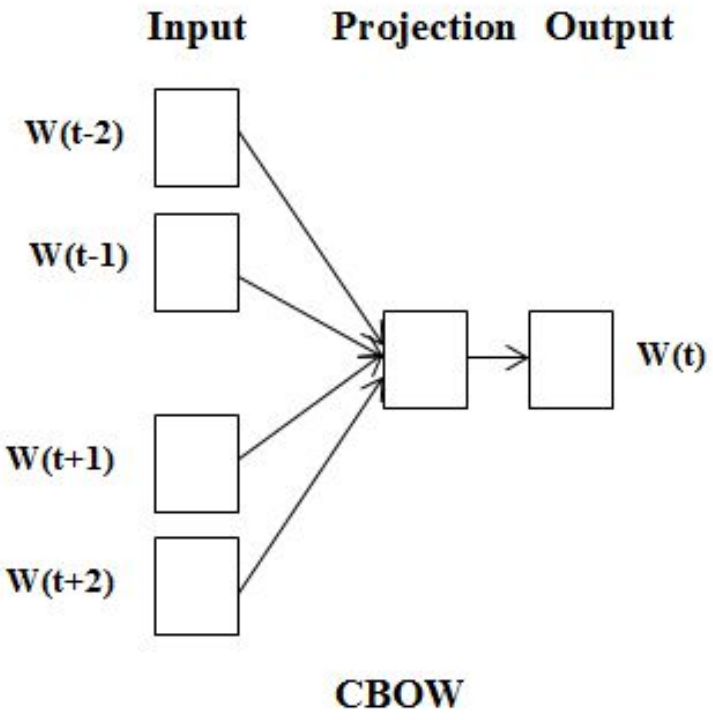


CBOW

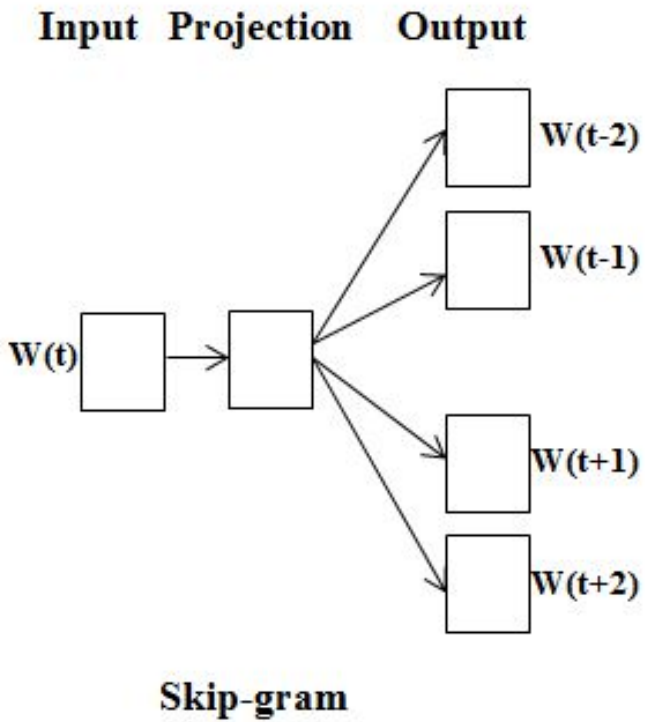
Skip gram

Fonte: Artigo 2013

Arquiteturas para modelar palavras



CBOW



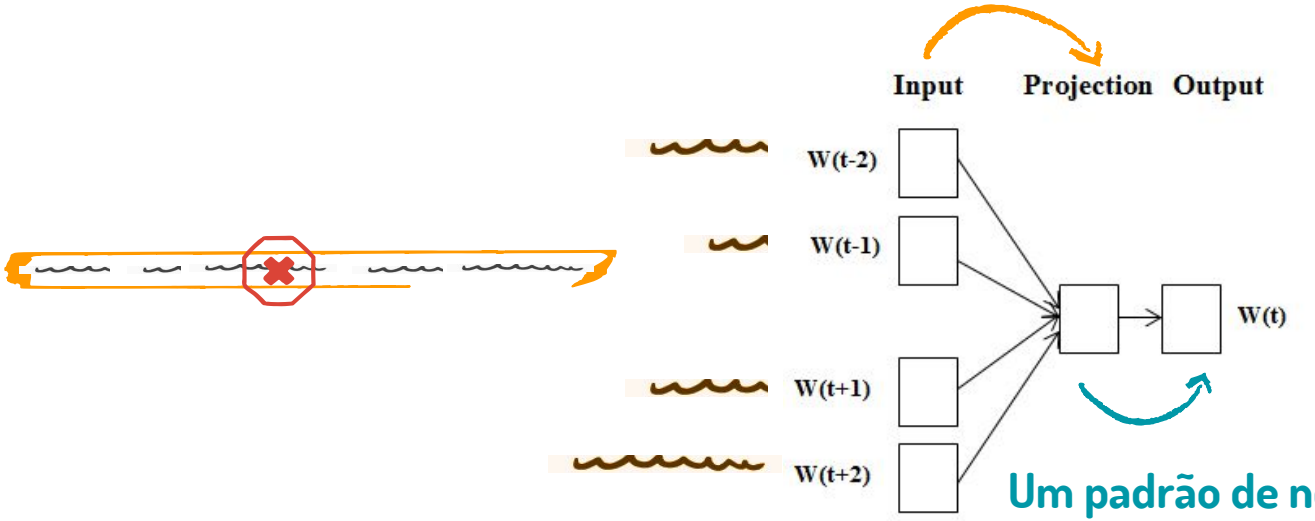
Skip-gram

CBOW Skip gram

Fonte: Artigo 2013

CBOW

Um padrão de neurônios na camada inicial ativo a partir dos termos vizinhos ativa
um padrão de neurônios na camada oculta

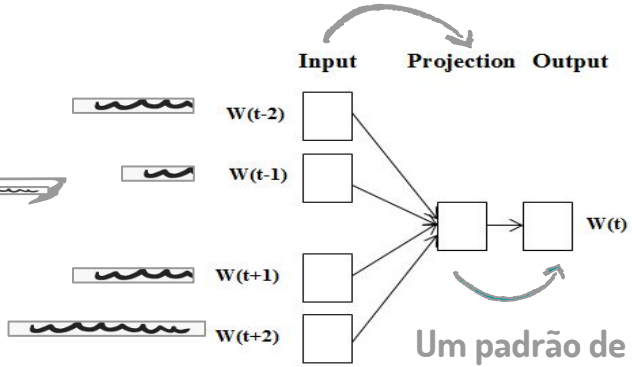


Erro/Custo do emprego dos parâmetros utilizados

Um padrão de neurônios na camada oculta ativa um label final

CBOW

Um padrão de neurônios na camada inicial ativo a partir dos termos vizinhos ativa
um padrão de neurônios na camada oculta



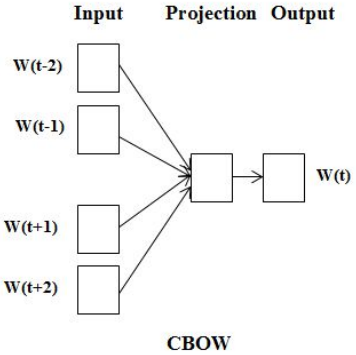
Um padrão de neurônios na camada oculta ativa um label final

Erro/Custo do emprego dos parâmetros utilizados

BACKPROPAGATION

CBOW | Skip gram

CBOW



Predizer uma dada palavra baseada no seu contexto

Melhor para associações

Sintáticas

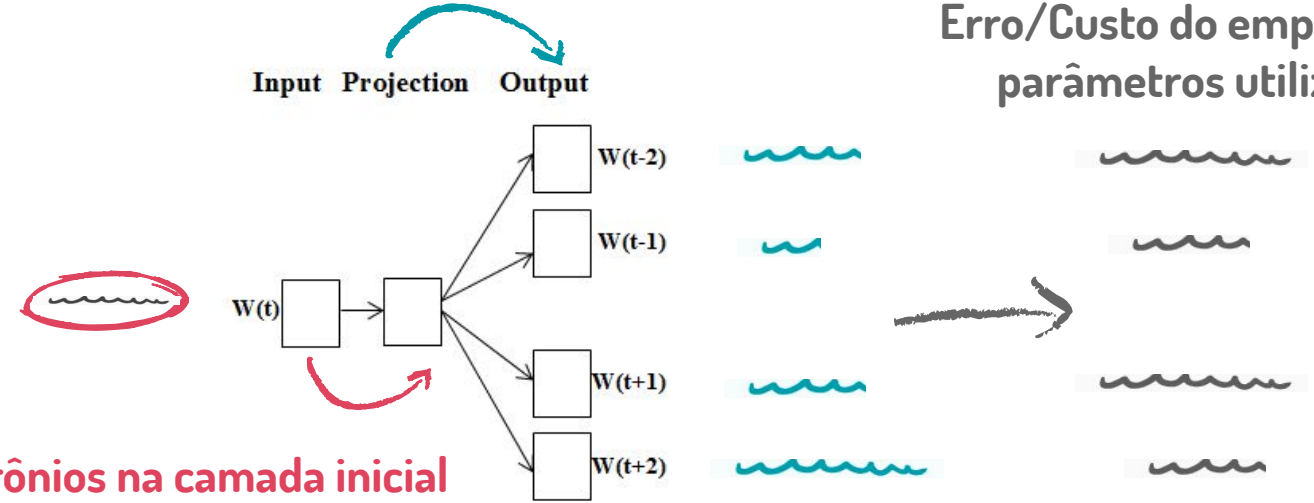
Type of relationship	Word Pair 1		Word Pair 2	
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Fonte: Artigo 2013

CBOW | Skip gram

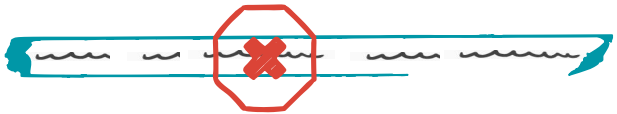
Skip-gram

Um padrão de neurônios na camada oculta ativa N vizinhos padrões de label final



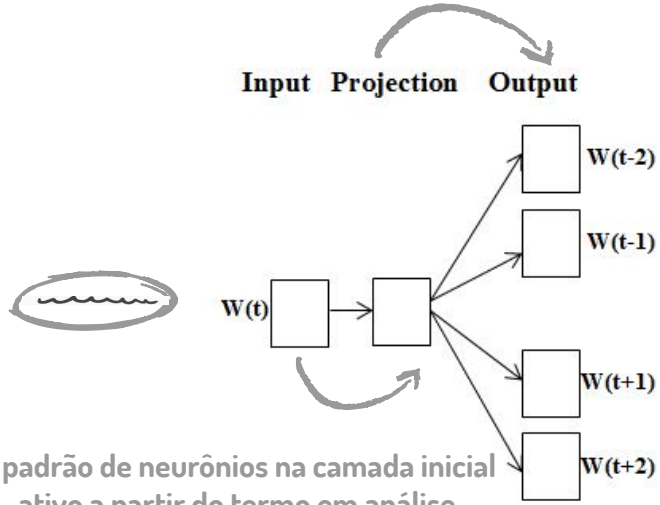
Erro/Custo do emprego dos parâmetros utilizados

O padrão de neurônios na camada inicial ativo a partir do termo em análise ativa um padrão de neurônios na camada oculta



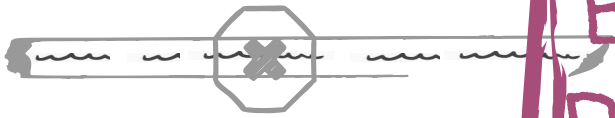
Skip-gram

Um padrão de neurônios na camada oculta ativa N vizinhos padrões de label final



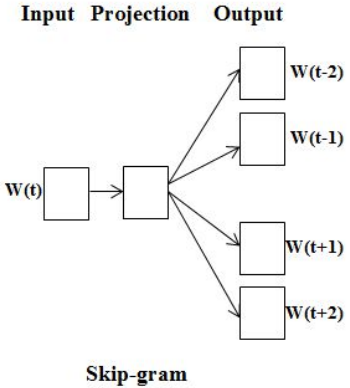
O padrão de neurônios na camada inicial ativo a partir do termo em análise ativa
um padrão de neurônios na camada oculta

Erro/Custo do emprego dos parâmetros utilizados

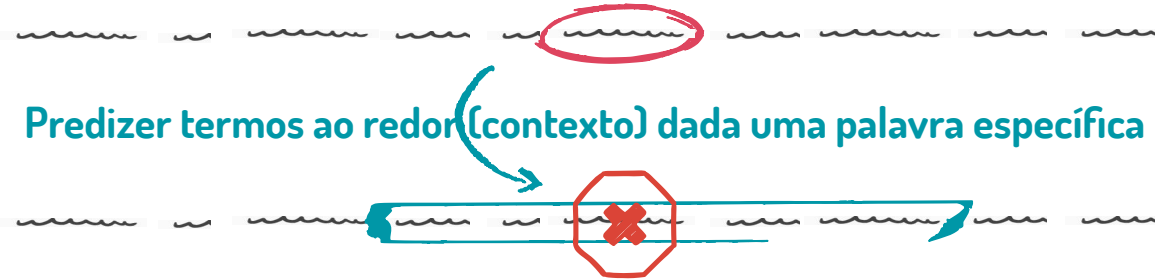


CBOW | Skip gram

Skip-gram



Skip-gram



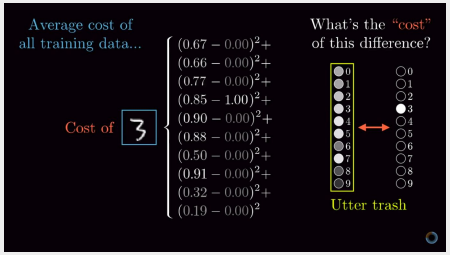
Melhor para associações
Semânticas

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter

CBOW | Skip gram

Fonte: Artigo 2013

Redes Neurais - Erro/Custo e Word2Vec



Existe um rótulo para comparar.

Trata-se de uma abordagem supervisionada

É impossível rotular contexto, ainda mais com relações sintáticas e semânticas

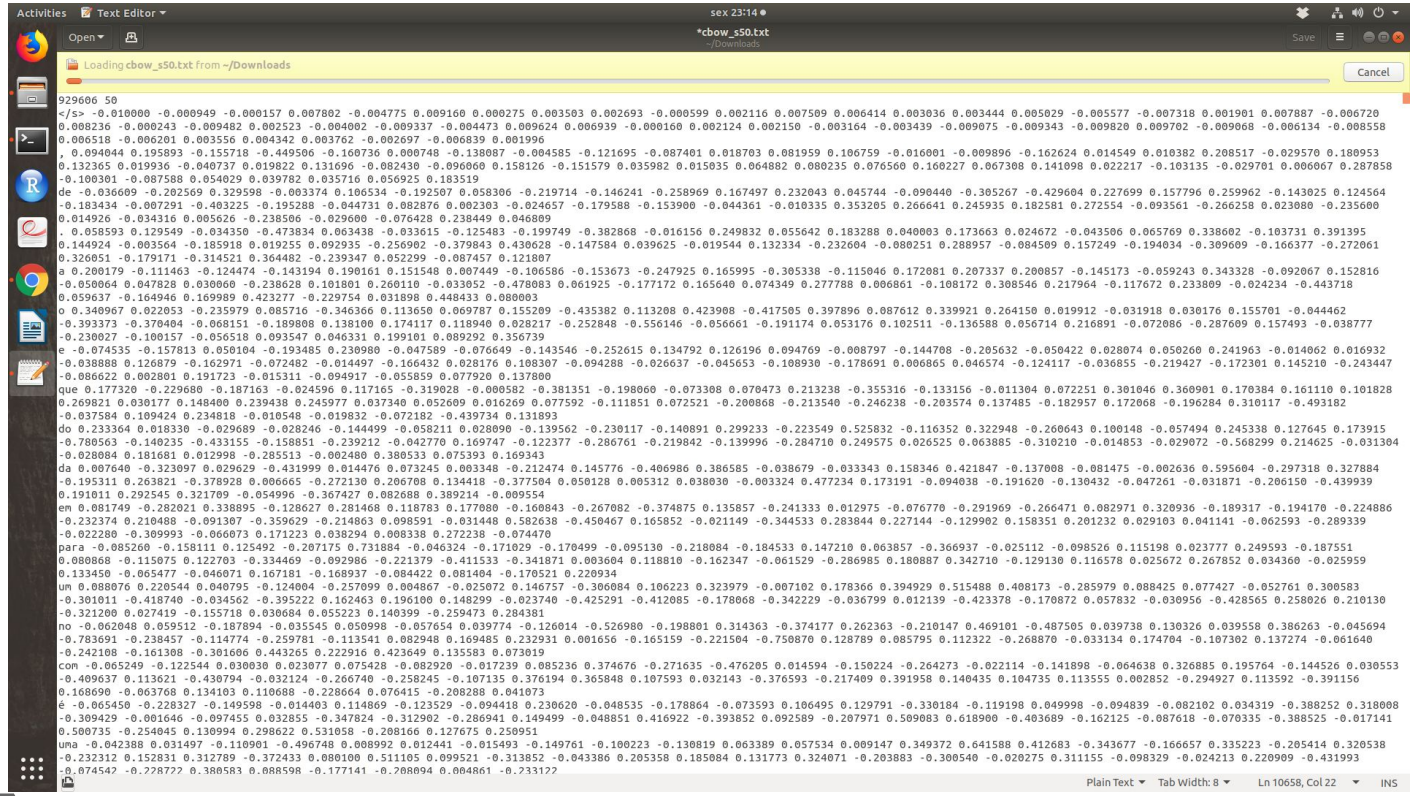
CBOW

Skip gram

Comentários

GloVe- Global Vectors	Referência teórica	Embeddings treinados em Português	Tensorflow	R
Representação de **Stanford** Muito utilizado Text2vec - Glove	IllustratedWord2Vec	linkUSP Parpinelli	Ferramenta da **Google** para Mac Learn detalhes	Rword2vec (Google) Train_word2vec wordVectors/ primaryobjects Tensorflow.rstudio

Word2Vec - Representação CBOW 50



CBOW

Skip gram

Recap - Definições

O que faz	Definição	Como funciona	Para quê serve	Resultado
Cria vetores que são representações numéricas dos atributos (termos/palavras) sem intervenção humana	Posiciona termos (palavras) em um espaço dimensional tal que as localizações dos termos são determinadas pelos seus significados e os vazios e distâncias no espaço também tem significado	Rede neural de duas camadas que processa texto	Criar representação de strings para algoritmos de Machine Learning ou Deep Learning	Agrupar vetores de termos similares no espaço de atributos

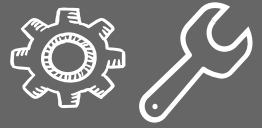
Recap – Vantagens

Bons exemplos

Exemplos com diferentes formas de similaridade (semântica e sintática)

Popular

Google desenvolveu uma ferramenta, tensorflow, e a disponibilizou



Grandes Datasets

A técnica lida bem com bases grandes, típicas da realidade atual (6B tokens, 1M termos mais freq)

Rápido

De: 08 semanas
Para: 01-03 dias

Relações preservadas entre línguas

Permite tradução termo-a-termo

Obrigada!
Dúvidas?

larissa.sayuri.fcs@gmail.com

larissa@maiomarketing.com.br