



Desvendando os bancos de dados vetoriais e a busca vetorial



@alexsalgadoprof

Alex Salgado
Developer Advocate @ Elastic



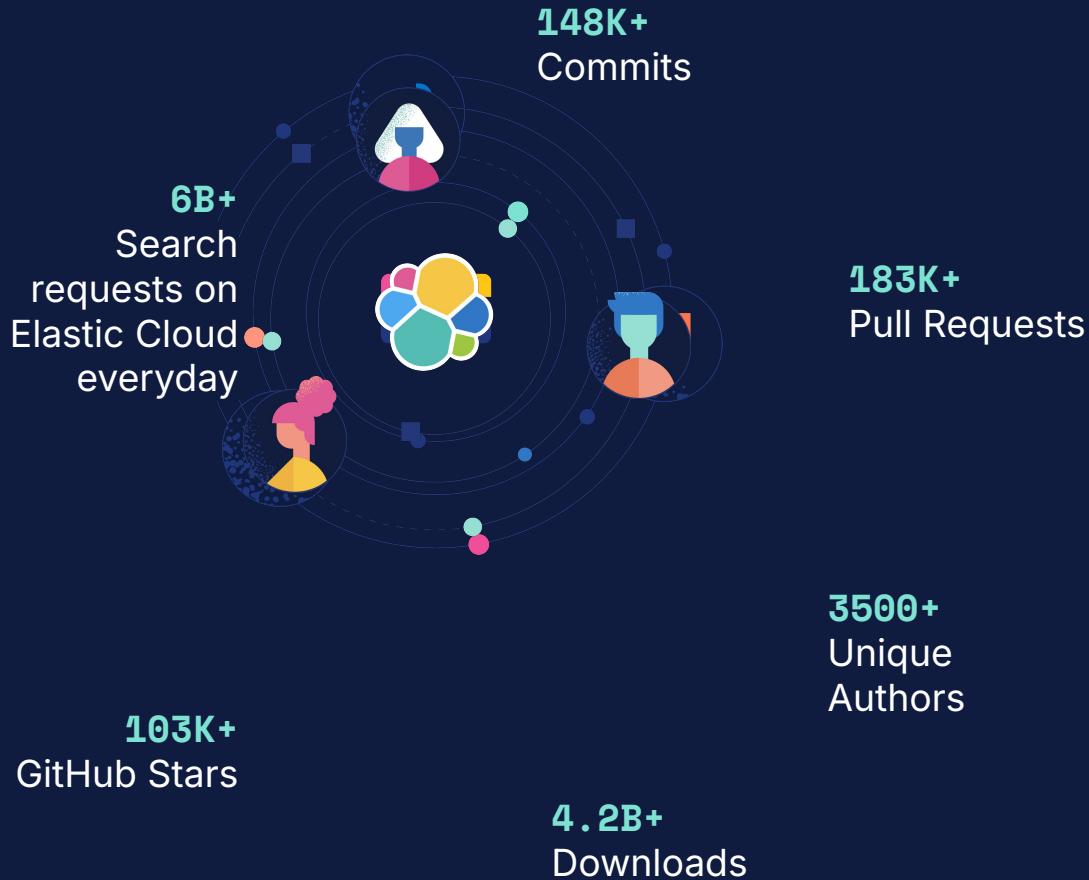
Alex Salgado
Senior Developer
Advocate LATAM

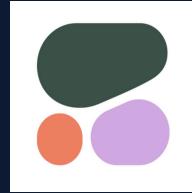
 @alexsalgadoprof
 salgado
 @alexsalgadoprof
 /in/alex-salgado/

- **Mestre** em Ciência da Computação pela UFF (Games)
 - **MBA** UFF
 - **PhD Candidate UFF: Robótica/Visão Computacional**
-
- **+ 25 anos** de experiência na área de desenvolvimento de software
 - Ocupei diversos cargos, trabalhando em **startups**, pequenas e grandes empresas como Oracle, CSN, BRQ/IBM, **Chemtech/Siemens (9 anos)**.
 - **8 anos** como professor universitário



A busca mudou a forma como o mundo utiliza os dados.





A Busca é **muito mais** que uma **caixa de pesquisa**



{ }

```
Select *\nFrom cadastro\nWhere nome =\n"fulano"
```

• • •

O Desafio das Buscas Tradicionais

Limitações das buscas baseadas em texto e palavras-chave

- Falta de compreensão semântica e contexto
- Dificuldade em integrar diferentes tipos de dados (texto, imagem, áudio)

O Desafio das Buscas Tradicionais

1. Correspondência Exata de Palavras-chave

- Dificuldade em encontrar resultados relevantes se as palavras exatas não forem usadas.

2. Falta de Compreensão do Contexto

- Não distingue entre diferentes significados de uma mesma palavra.
- Exemplo: "Maçã" como fruta vs. "Apple" como empresa.

3. Sinônimos e Variações Linguísticas

- Não lida bem com sinônimos e diferentes formas de uma palavra.
- Exemplo: "Corrida" vs. "Maratona".

4. Busca por Similaridade ou Conceito

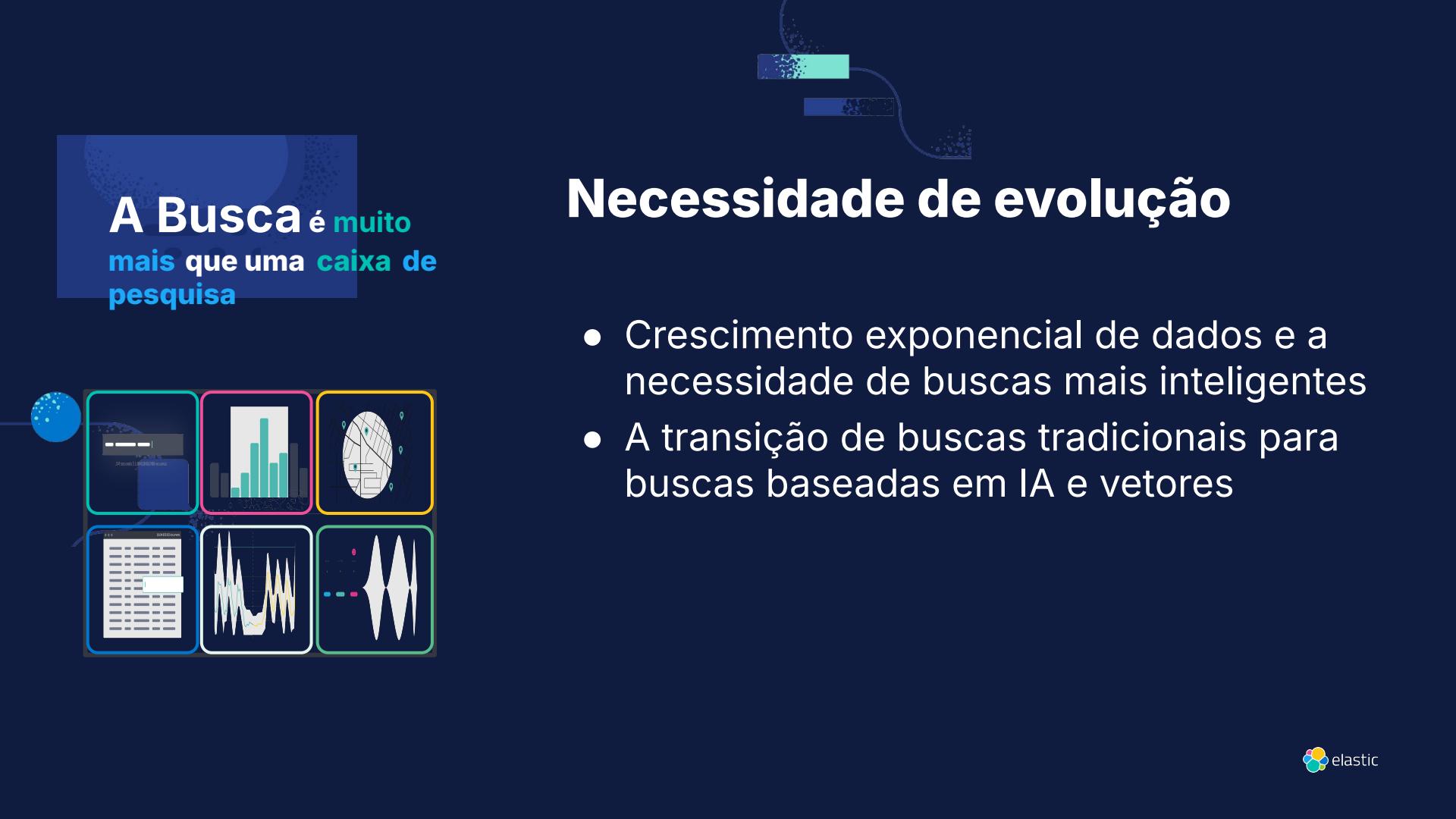
- Não identifica documentos conceitualmente similares sem termos em comum.
- Exemplo: "Roteiro de viagem" vs. "O que fazer em Paris".

5. Escalabilidade e Eficiência

- Diminuição da eficiência em grandes volumes de dados.
- Resultados menos relevantes e mais lentos.

6. Desempenho em Consultas Complexas

- Dificuldade em processar consultas que envolvem múltiplos conceitos.
- Exemplo: "Efeitos do aumento das taxas de juros na economia global".



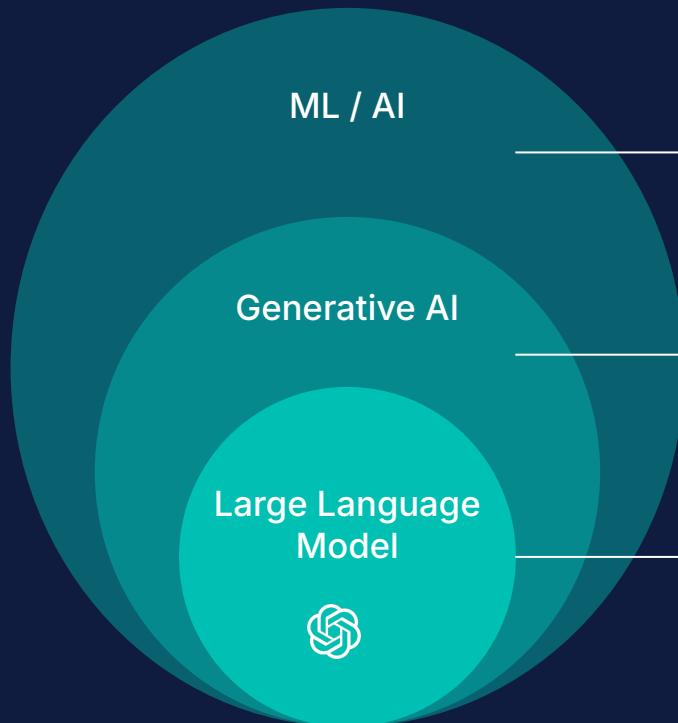
A Busca é muito
mais que uma caixa de
pesquisa



Necessidade de evolução

- Crescimento exponencial de dados e a necessidade de buscas mais inteligentes
- A transição de buscas tradicionais para buscas baseadas em IA e vetores

Aplicações interessantes de ML, IA generativa e LLMs



O que é ?

Algoritmos programados para fazer previsões com base em dados

Algoritmos de IA projetados para criar novos dados

Algoritmos de aprendizado profundo que podem gerar texto

Aplicações

Reconhecimento de imagem, processamento de linguagem natural, reconhecimento de fala

Chatbots, geradores de texto, geradores de imagem, geradores de música

Geradores de texto, tradução, escrita, resposta a perguntas

O poder da IA generativa

reside nos
seus dados
proprietários



Three solutions powered by one stack

3 solutions



Enterprise Search



Observability



Security

Powered by
the Elastic
Stack

Kibana

Elasticsearch

Agent

Beats

Logstash

Deployed
anywhere



Elastic Cloud



Elastic Cloud
Enterprise

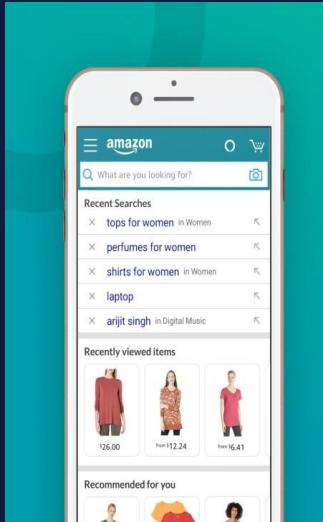


Elastic Cloud
on Kubernetes

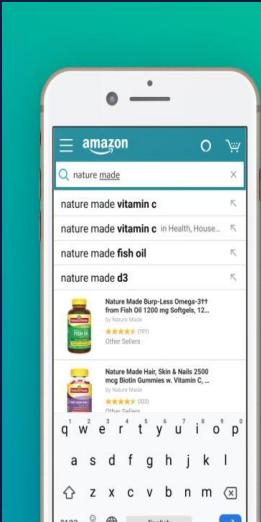
SaaS

Orchestration

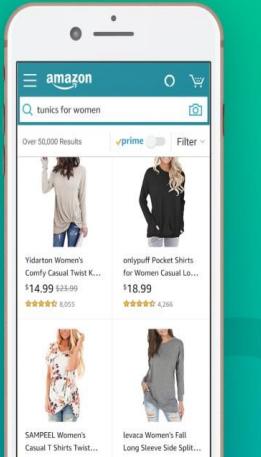
As pessoas estão mudando o jeito de pesquisar



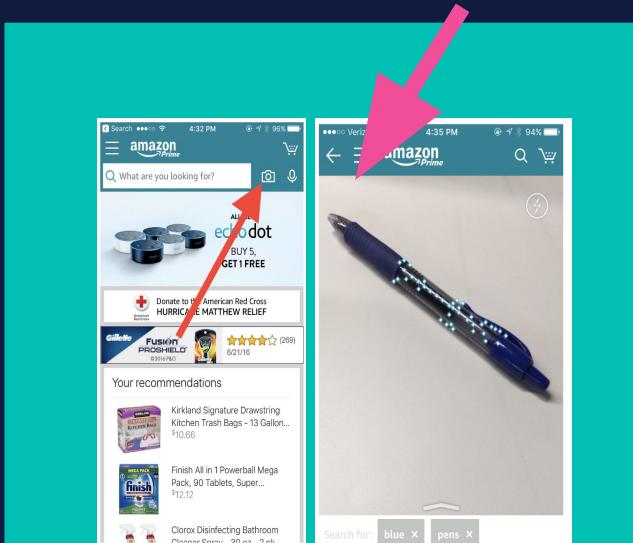
1-Busca textual



2-Busca textual + Semantica

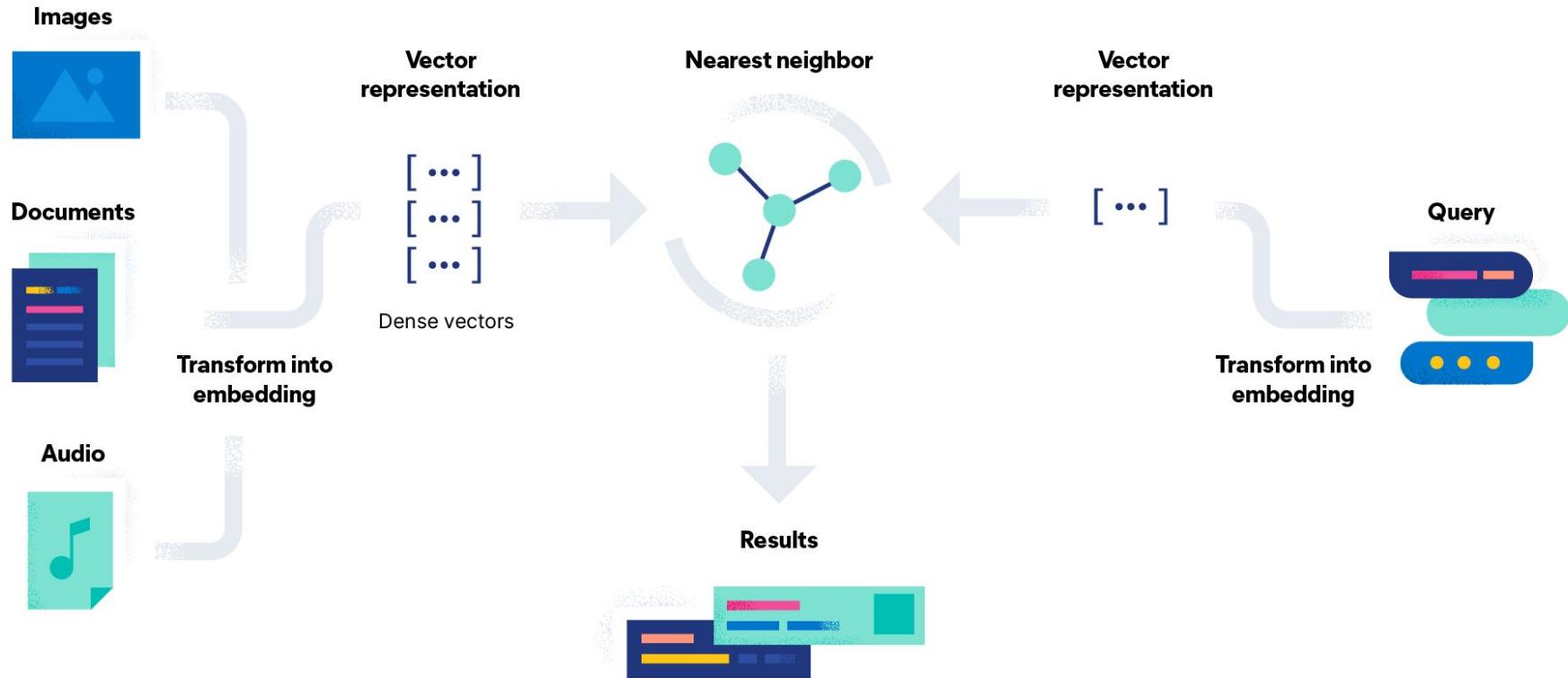


3-Busca por Imagem e Som



Elasticsearch: You Know, for **Vector** Search

Vector Search





O que é um Vetor?

Embeddings represent your data

Example: 1-dimensional vector



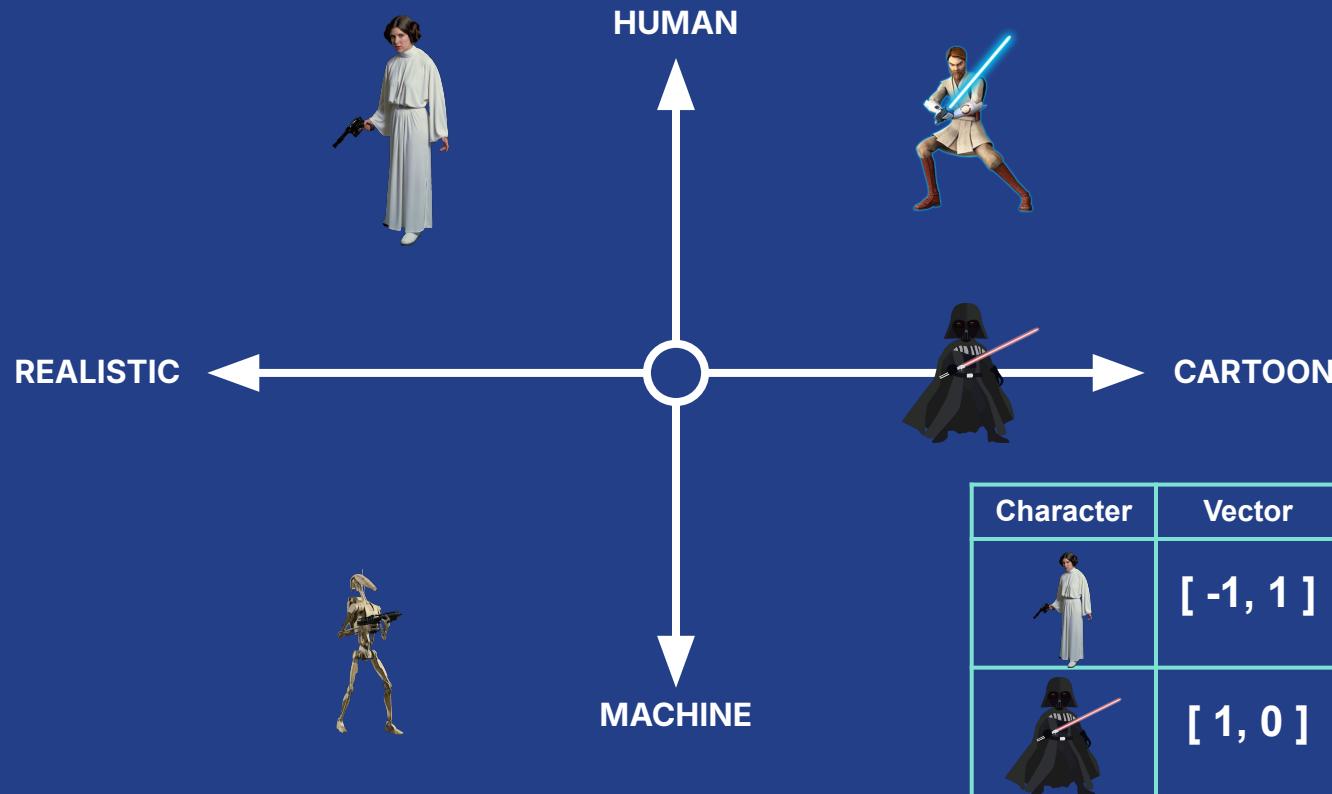
REALISTIC



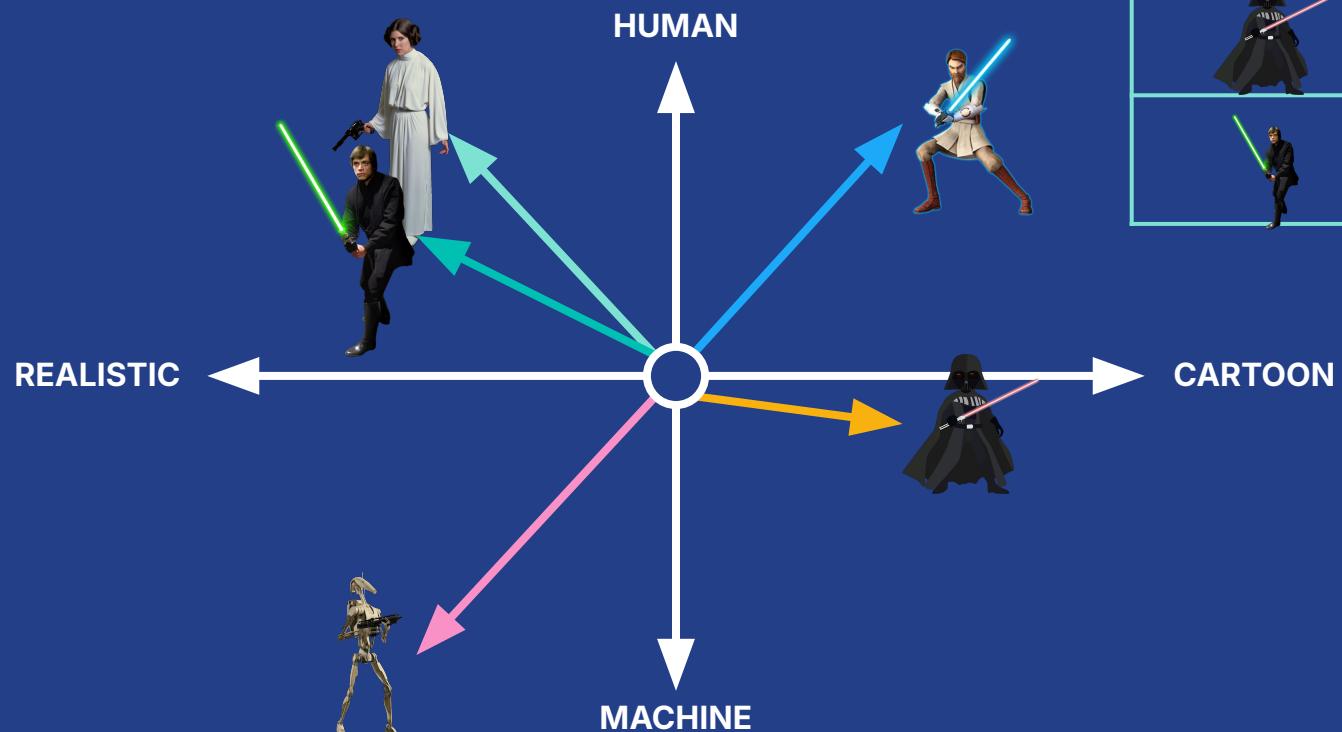
CARTOON

Character	Vector
	[-1]
	[1]

Multiple dimensions represent different data aspects

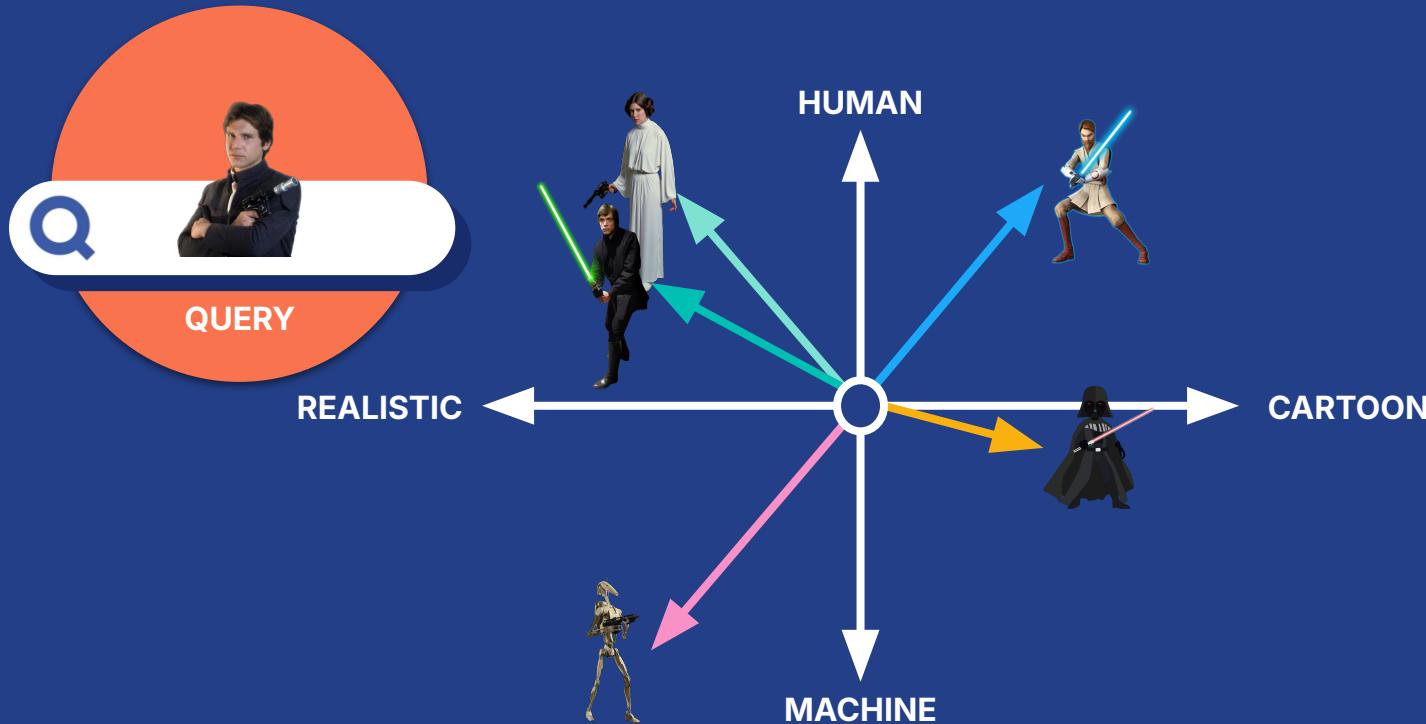


Similar data is grouped together



Character	Vector
	[-1.0, 1.0]
	[1.0, -0.1]
	[-1.0, 0.8]

Vector search ranks objects by similarity (relevance) to the query

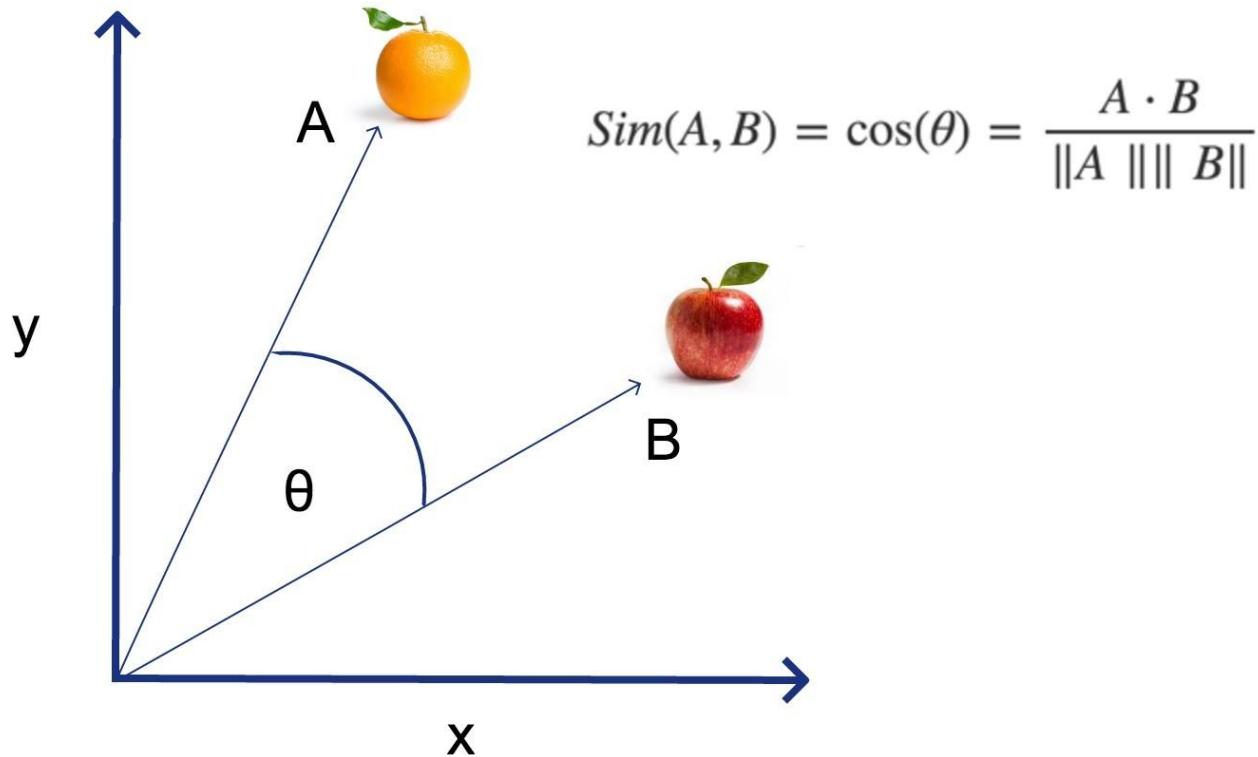


Relevance	Result
Query	
1	
2	
3	
4	
5	

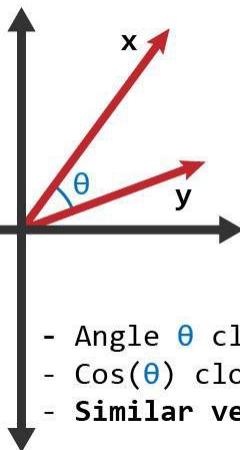


Como essa busca por similaridade realmente funciona?

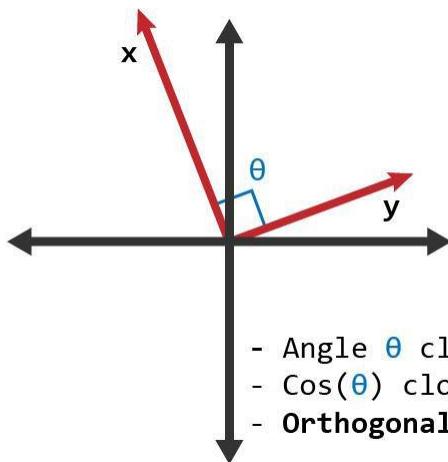
Similarity: Cosine



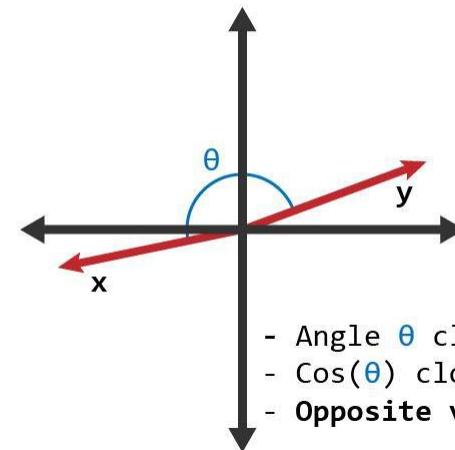
Similarity: Cosine



- Angle θ close to 0
- $\cos(\theta)$ close to 1
- Similar vectors

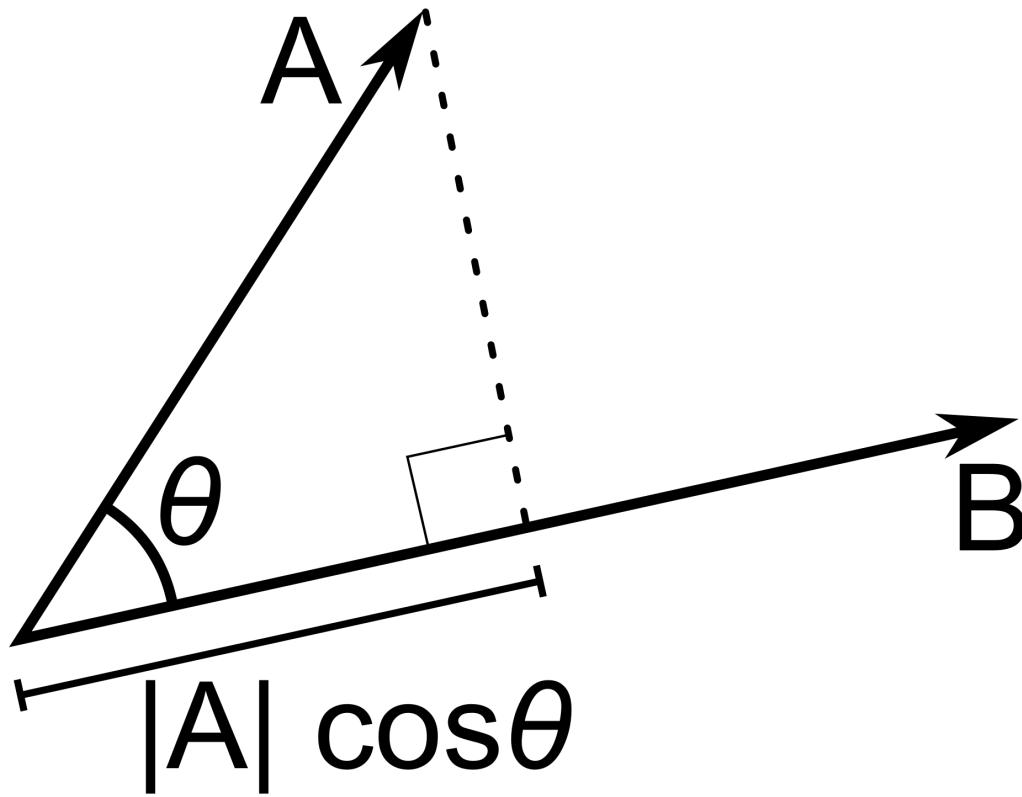


- Angle θ close to 90
- $\cos(\theta)$ close to 0
- Orthogonal vectors

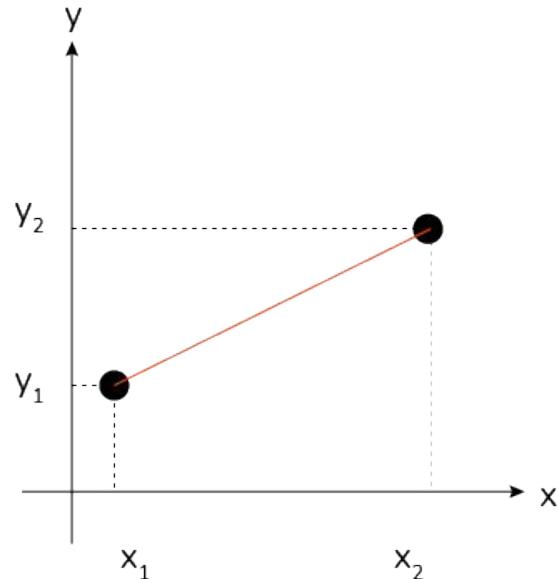


- Angle θ close to 180
- $\cos(\theta)$ close to -1
- Opposite vectors

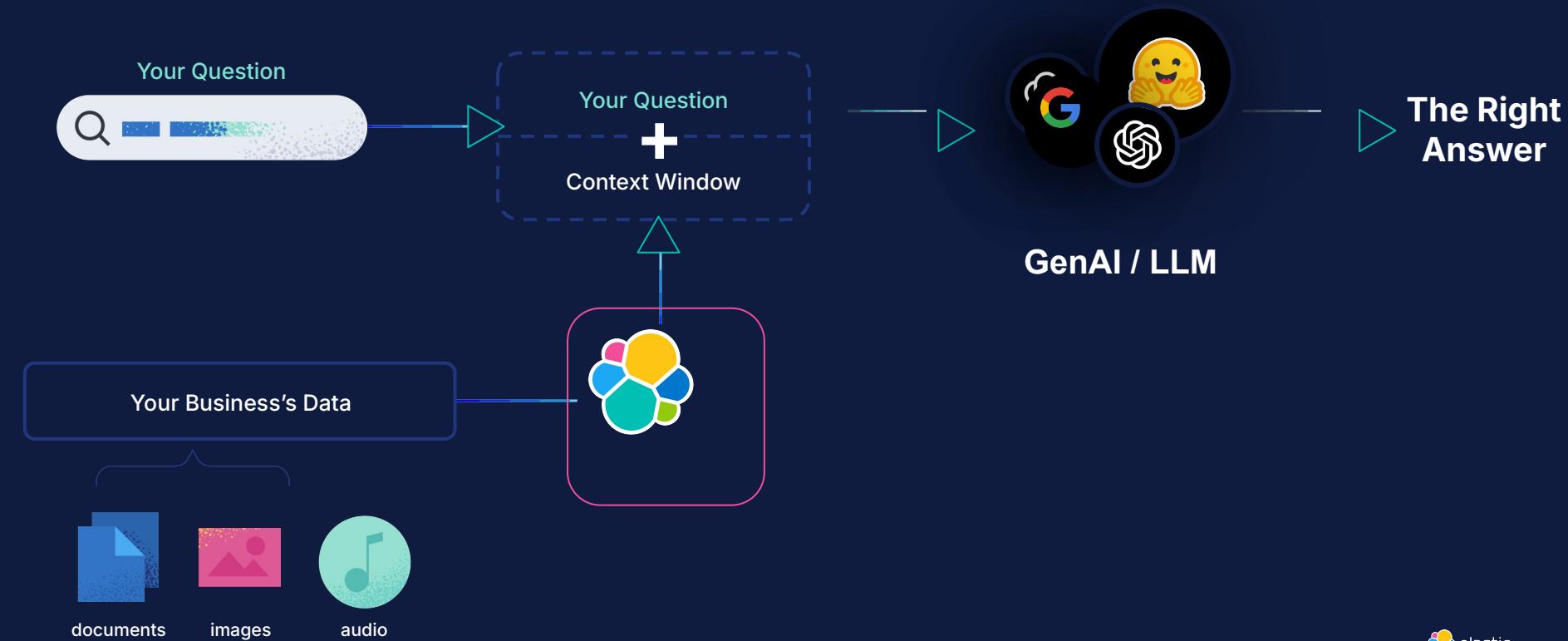
Similarity: Dot Product



Similarity: Euclidean / L2



A importância do banco vetorial para a Arquitetura RAG



Elasticsearch Relevance Engine™ (ESRE)

Oferece aos desenvolvedores um poderoso conjunto de ferramentas de **aprendizado de máquina** para construir aplicativos de busca **potencializados por IA** que se integram com **large language models**

- Pesquisa textual e banco de dados vetorial
- Capacidade de hospedar seu próprio modelo de transformer
- Capacidade de integração com LLMs de terceiros (OpenAI)
- RRF - modelo de pontuação híbrida (vetorial e pesquisa textual)
- Modelo de ML proprietário da Elastic
- Integração com ferramentas de terceiros como LangChain

Reflete anos de P&D

Os ingredientes da busca com IA generativa



+



+



+



Busca de texto,
vetorial e híbrida

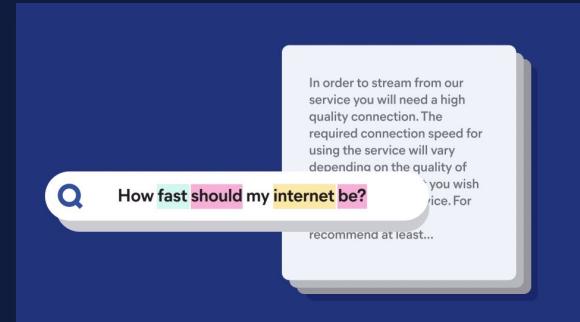
Capacidades de
banco de dados
vetorial

Escolha de
modelos de
embedding

Ferramentas e APIs
para construir
aplicativos de busca

Busca vetorial: A base para a relevância semântica

- Resultados de busca **mais rápidos e precisos**
- Consulta a dados **não estruturados, linguagem natural**
- Além das palavras-chave - vetores são termos de busca **armazenados como números**
- busca de similaridade vetorial ou distância numérica, classificação baseada no significado **semântico**
- Casos de uso para **texto, imagens, áudio, vídeo**



O que é um modelo multimodal?



How fast should my internet be?

In order to stream from our service you will need a high quality connection. The required connection speed for using the service will vary depending on the quality of

you wish to use the service. For

recommend at least...

Modelos Multimodais



(Audio)



(Video)

Laboris sunt autem ex exercitatione aliqua id commode sunt irure labore ut pariatur laborum sit. Aliquip tempor anim esse enim Lorem. Reprehenderit amet reprehenderit tempor deserunt cillum consectetur. Es tempor autem deserunt do veniam nostrud ea cillum aliquip minim eu veniam ex. Non in reprehenderit officia officia qui nisi et proident est cillum dolore labore non.
Lorem ullamco commodo consequat voluptate fugiat ea dolore sit magna sit reprehenderit id. Cupidatat magna occaecat irure esse esse dolore exercitatione duis elit eu. Velit consectetur ea culpa consequat elit sunt ex mollit. Nisi volutate aliqua sunt est proident dolor Lorem voluptate excepteur dolore ex. Magna fugiat velit elit exercitatione dolor nisi dolores ut proident qui qui et. Lorem magna labore eiusmod velit. Consequat adipisciing consequatur ex Lorem aliquip.
Consequat Lorem amet magna magna proident irure. Consequat deserunt ullamco mollit velit consequatur in laboris nisi occaecat cillum dolor irure adipisciing fugiat. Eiusmod duis id qui reprehenderit proident magna ut ex non sunt est ipsum. Do irure cupidatat amet non officia do aliqua sit et consectetur aute culpa incididunt. Culpa culpa ex culpa mollit. Eu minim consequat minim minim mollit velit. Officia sint fugiat ex deserunt consectetur voluptate esse sint velit aliquip exercitatione elit incididunt ut.
Est culpa exercitatione ullamco cillum nisi quis elit. Ea consequat ipsum occaecat labore quis veniam. Lorem ipsum labore eiusmod nostrud ex tempor aliqua labore culpa cillum deserunt Lorem.

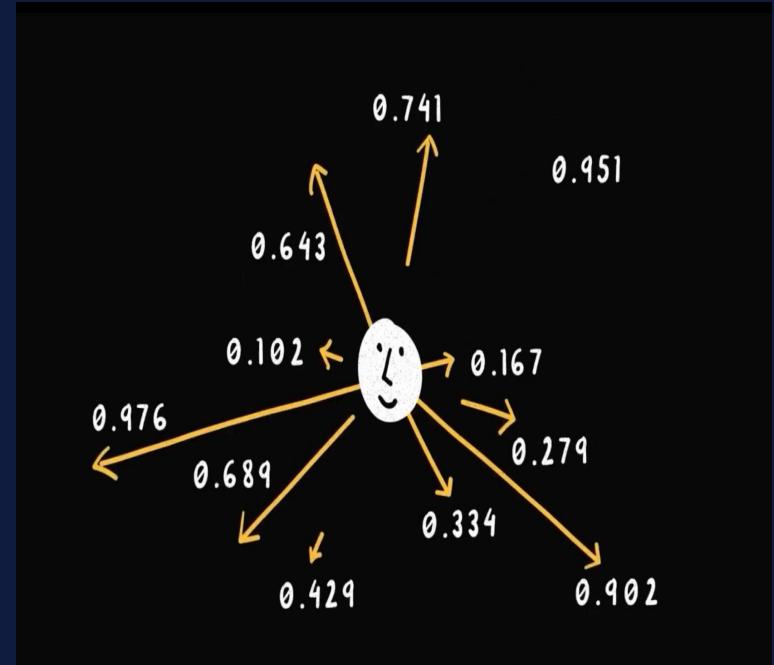
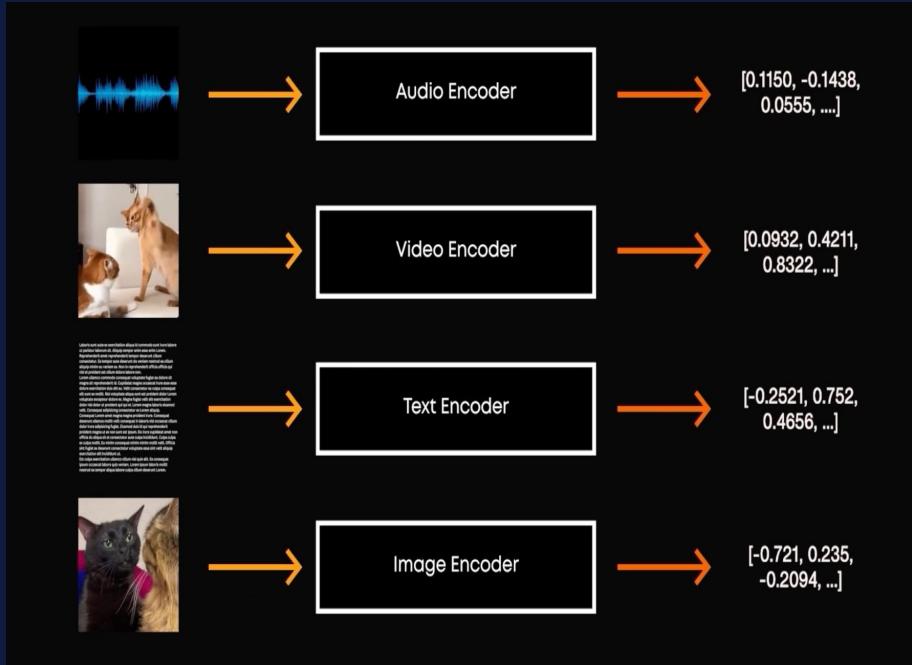
(Text)



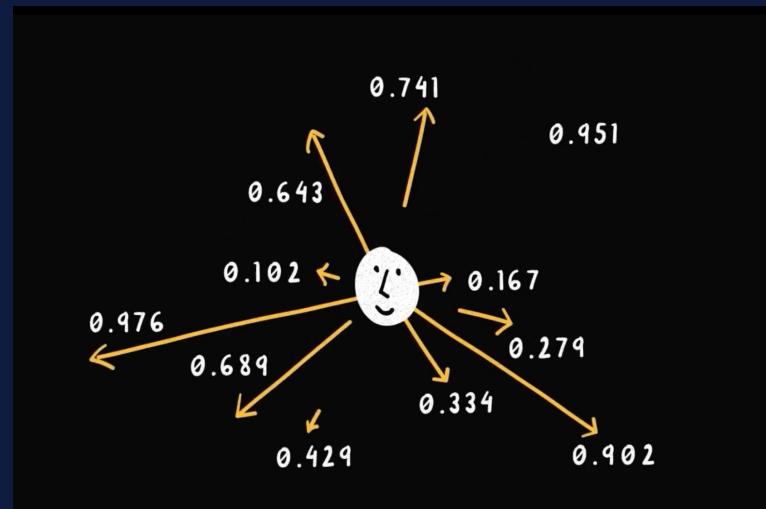
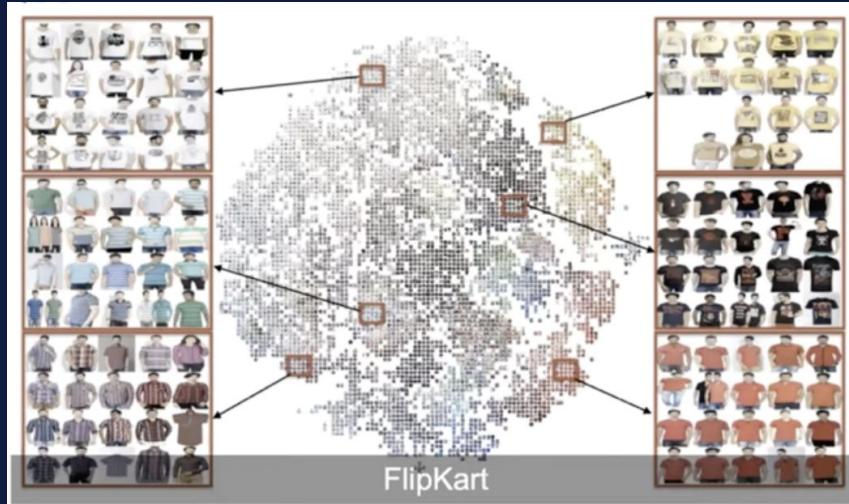
(Image)

Modelos Multimodais

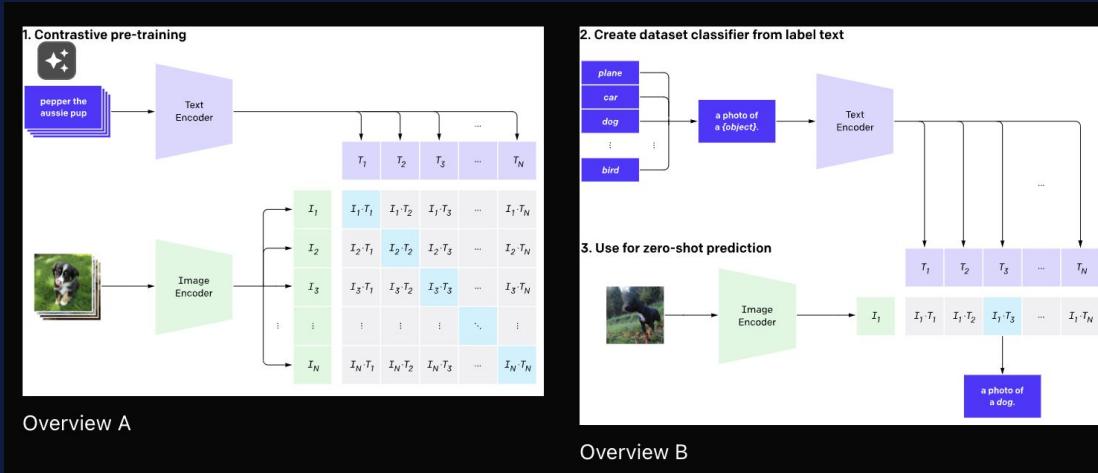
embeddings



Busca por Imagem



Anatomia de um modelo multimodal - CLIP da OpenAI



CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as “a photo of a dog” and predict the class of the caption CLIP estimates best pairs with a given image.

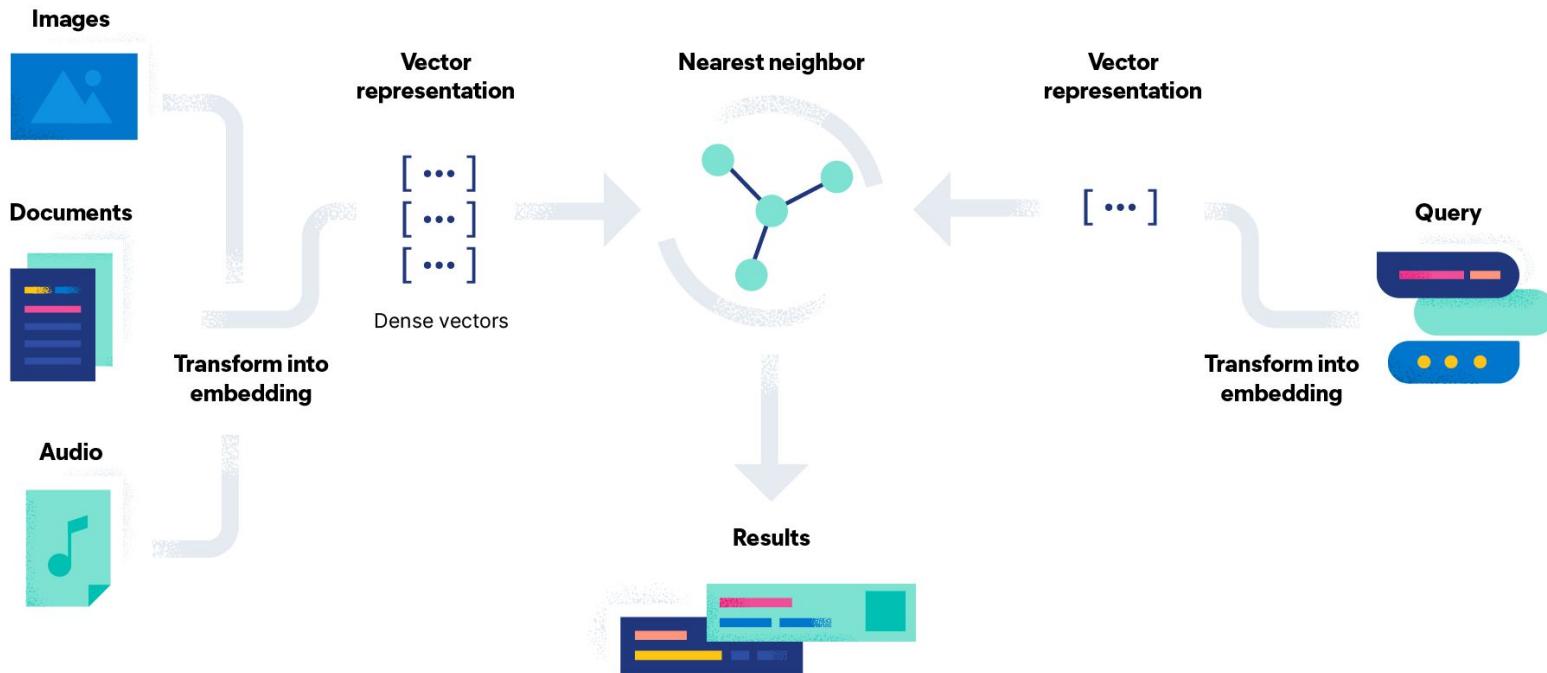
```
from sentence_transformers import SentenceTransformer  
  
model = SentenceTransformer('clip-ViT-B-32')
```

Pirâmide de Aprendizado de William Glasser

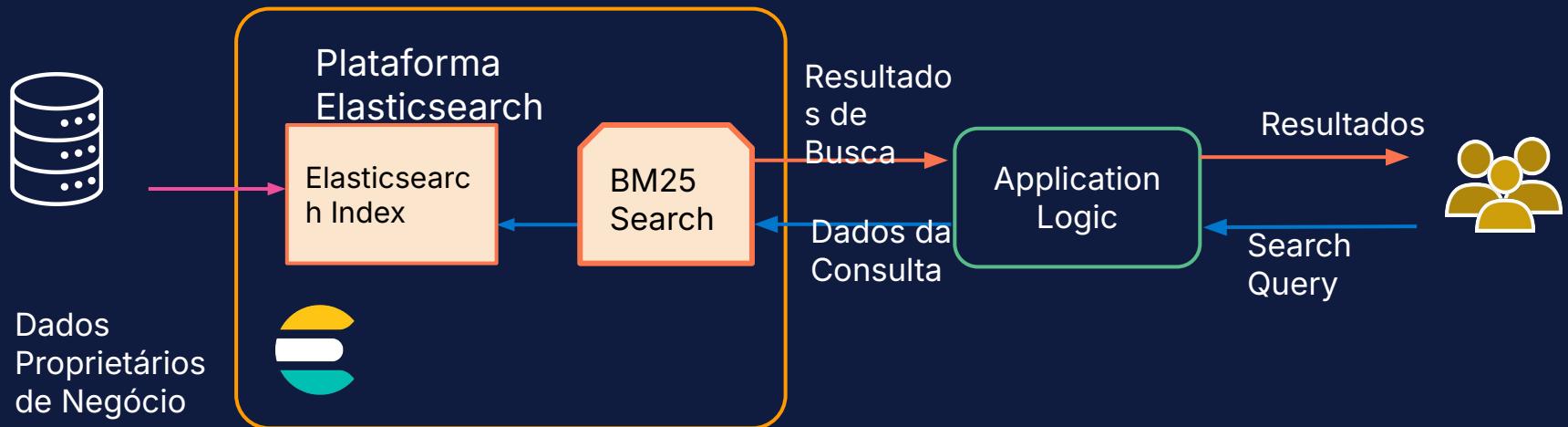


Evolução da arquitetura de busca

Consultas também são vetorizadas

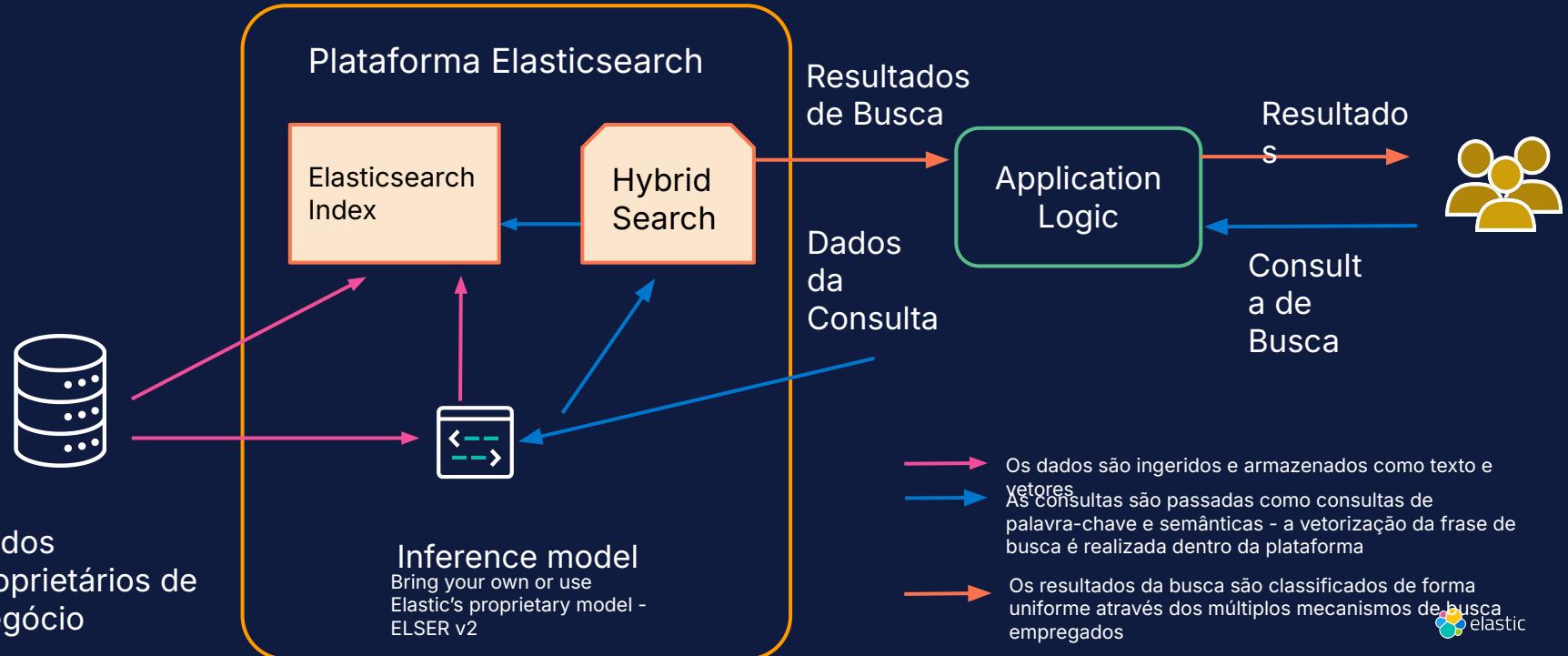


Arquitetura de Busca típica de busca textual (palavras-chave)



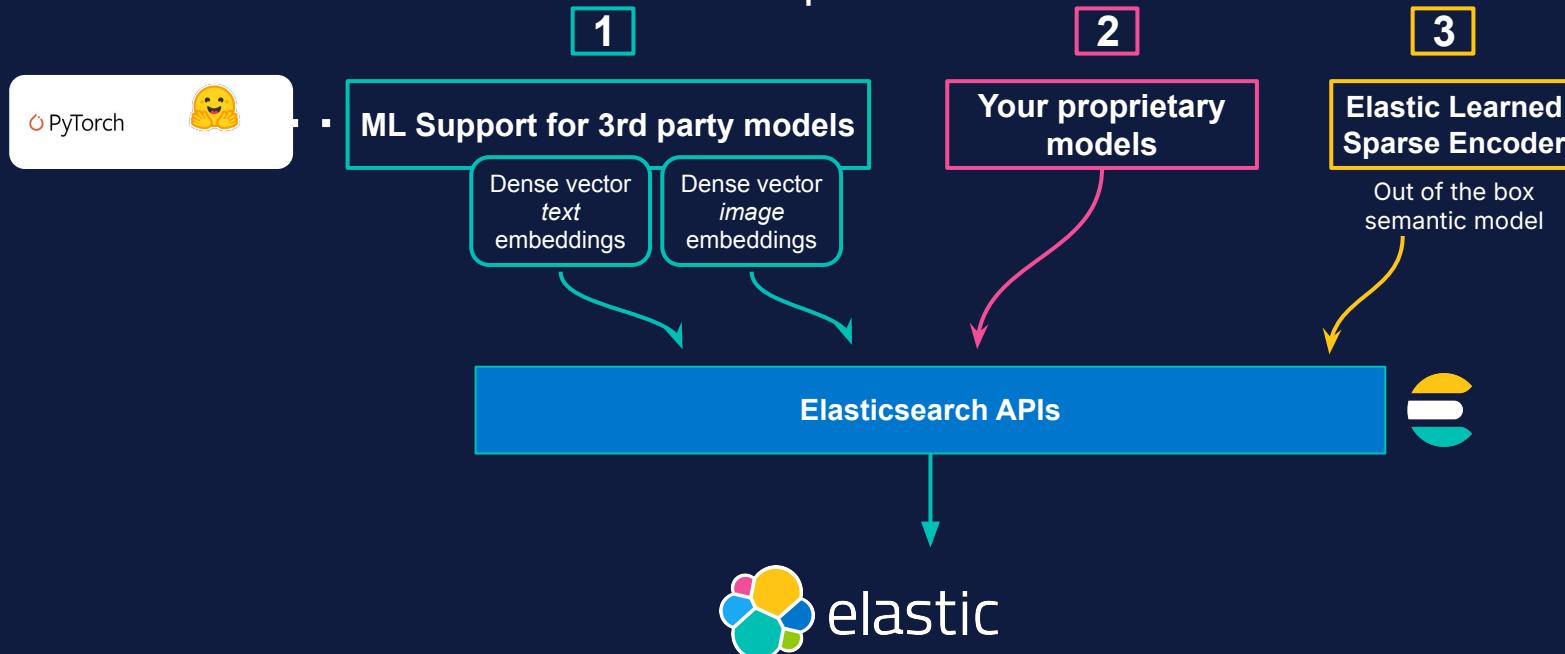
- A. Consulta de Busca
- B. Execução da Busca BM25

Arquitetura de busca evoluída para busca vetorial embeddings gerados dentro do Elasticsearch



Embeddings e Elasticsearch: três opções

modelos de terceiros, traga seus próprios modelos, nosso codificador esparso



Evolução dos banco de dados para suportar essas mudanças

Exemplos de aplicações

Busca em e-commerce

Gallivant WOMEN MEN SALE HOME

Q blue t-shirt with stripes

FILTER RESULTS Reset

DEPARTMENT

- Men 762
- Women 688

CATEGORY

- T-Shirt 489
- Shirt 280
- Jacket 234
- Dress 175
- Sweatsuit 83

+ SHOW 5 MORE

ON SALE

PRICE

0 \$999

COLORS

- Black 296
- White 102
- Dark Blue 93
- Grey 45
- Blue 42

+ SHOW 85 MORE

SIZE

- Large 1,450
- Medium 1,450
- Small 1,450

RATING

Showing 1 - 20 out of 1450 for: blue t-shirt with stripes

Sort by relevance

T-Shirt - blue \$11.99
Basic T-shirt - blue \$10.99
Print T-shirt - blue \$11.99
Basic T-shirt - blue \$11.99

Print T-shirt - blue \$11.99
Print T-shirt - blue \$10.99
Print T-shirt - blue \$11.99
Basic T-shirt - blue \$11.99

Print T-shirt - blue \$11.99
Print T-shirt - blue \$12.99
Print T-shirt - blue \$13.99
Basic T-shirt - blue \$7.99

Print T-shirt - blue \$11.99
Basic T-shirt - blue \$10.99
Basic T-shirt - light blue \$7.99
Basic T-shirt - blue mélange \$7.99

What are you looking for?

blue t-shirt with stripes

Search

P19220A0NKT1.png Print T-shirt - navy \$11.99 0.66015697

cotton; occasion:leisure; product model length:27.0 " (Size M); pattern: striped; mode: height:Our model is 74.0 " tall and is wearing size M; washing_instructions:do not tumble dry; washing_instructions: machine wash at 30°C; washing_instructions:Machine wash on gentle cycle; sleeve_length:option short correct; fit: regular; clothing_length: standard; neckline: round neck

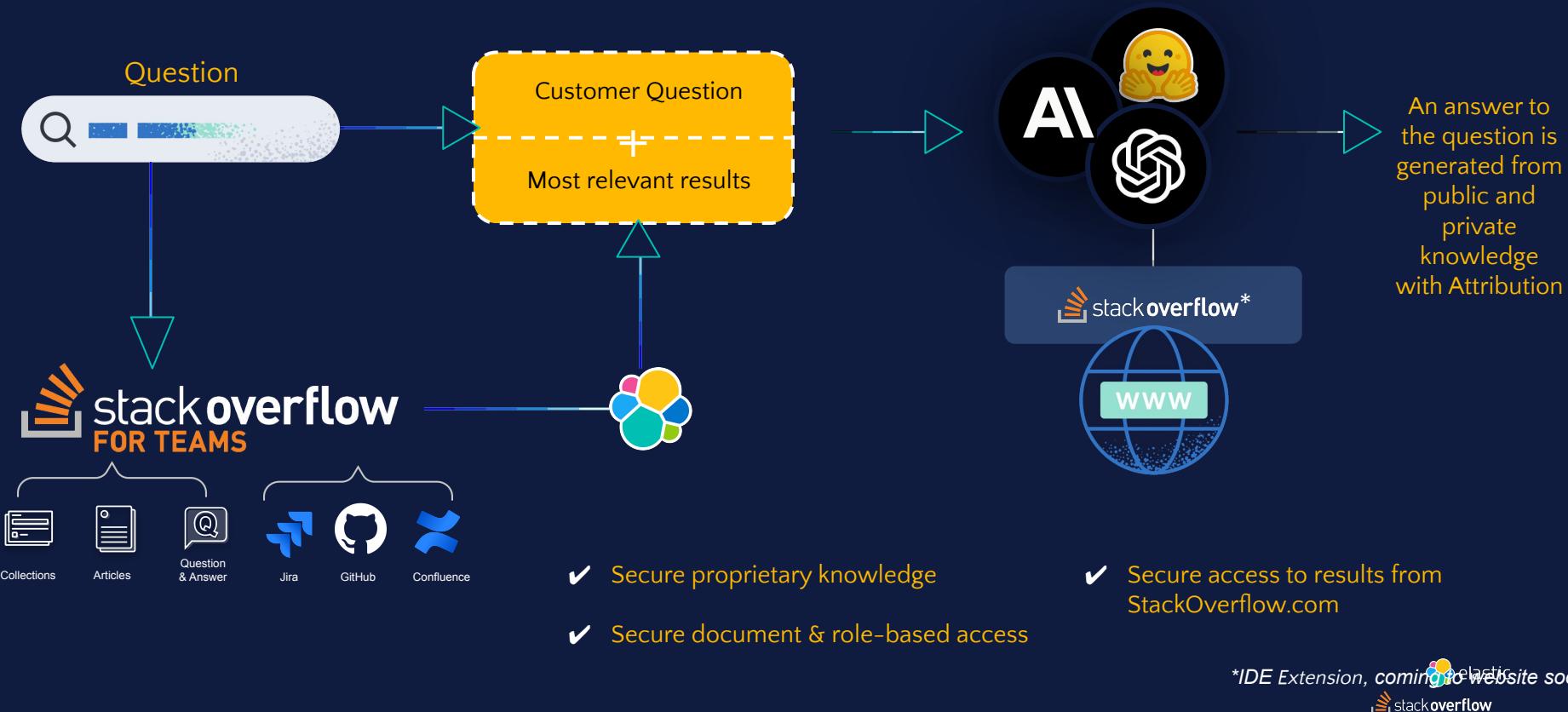
AN021C0288950.png Maxi dress \$24.99 0.66015697

upper material: clothing: 97% polyester, 3% spandex; product model length: 56.0 " (Size S); skirt_details: elasticated waist; washing_instructions: do not tumble black; white; dry; washing_instructions: machine wash at 30°C; clothing_length: long

Y01220A0CA11.png Print T-shirt - white/blue \$11.99

material: construction: Jersey; upper material: clothing: 100% cotton; occasion:leisure; product model length: 32.5 " (Size M); pattern: striped; mode: height:Our model is 74.5 " tall and is wearing size M; washing_instructions: do not tumble dry; washing_instructions: machine wash at 30°C; washing_instructions: Machine wash on gentle cycle; sleeve_length: option short correct; fit: large; clothing_length: long; neckline: round neck

Case : Stackoverflow Teams



Exemplo em Blogs



Finding your puppy with Image Search

Have you ever been in a situation where you found a lost puppy on the street and didn't know if it had an owner? Learn ho...

November 7, 2023 • Alex Salgado

Searching by music: Leveraging vector search for audio information retrieval

By Alex Salgado

16 August 2023



Dec 22nd, 2023: [PT] Papai Noel Encontra a IA Generativa: Decifrando Cartas de Natal Escritas à Mão com LLM, LangChain e Elasticsearch

Elastic Community and Ecosystem Advent Calendar



Alex Salgado-Elastic Alex Salgado Elastic Team Member

Dec 2023



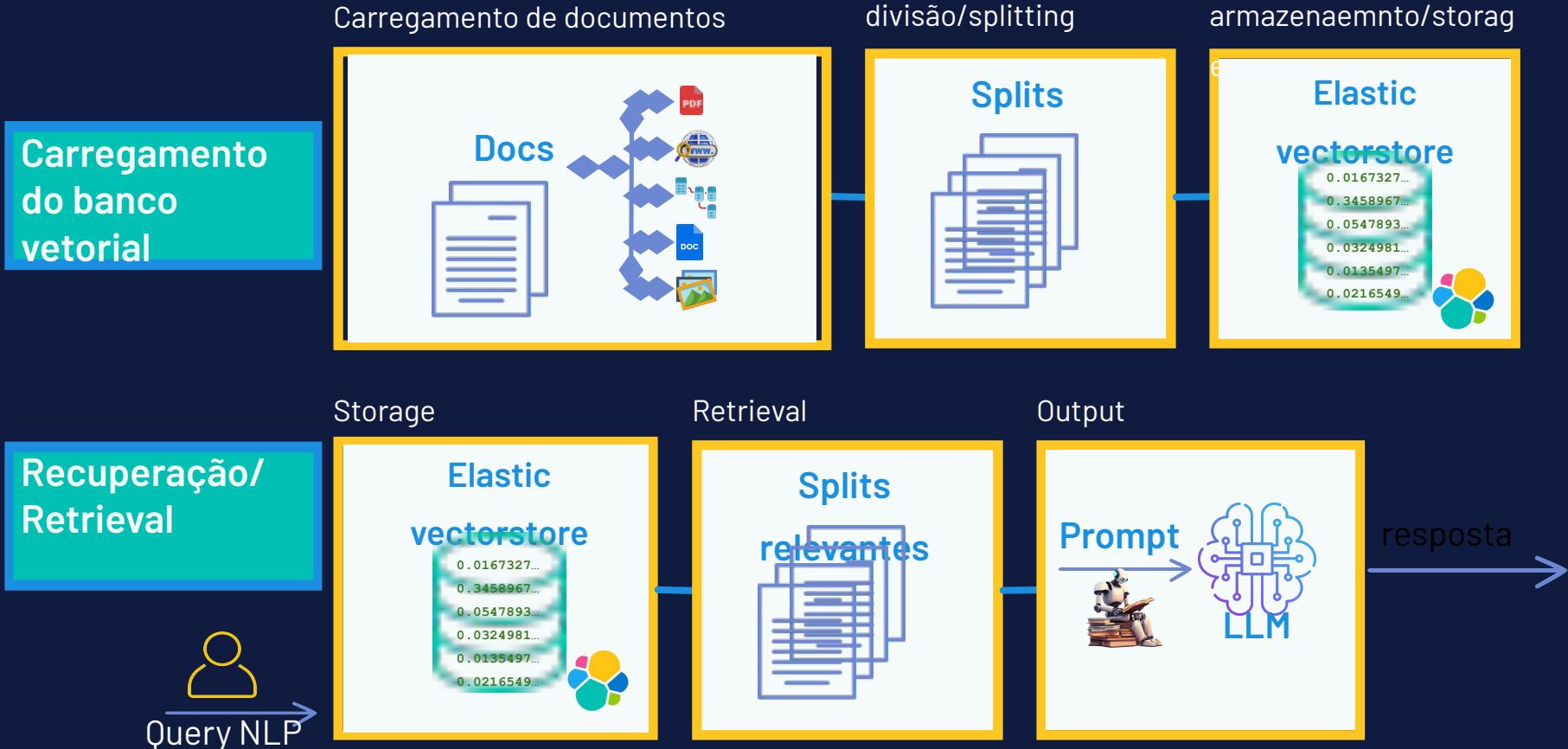
This post is also available in english.

No coração do Polo Norte, a equipe de duendes de Papai Noel enfrentava um desafio logístico formidável: como lidar com milhões de cartas de crianças de todo o mundo. Com um olhar determinado, Papai Noel decidiu que era hora de incorporar inteligência artificial na operação natalina.



1 / 2

Exemplo com chatbot



Novo campo *semantic_text*

- Modelos Internos: ELSER, E5.
- Modelos Externos: Cohere, Hugging Face, OpenAI, Mistral, etc.

```
# criar inference point
PUT _inference/sparse_embedding/my-elser-endpoint
{
  "service": "elser",
  "service_settings": {
    "num_allocations": 1,
    "num_threads": 1
  }
}
```

Modelo interno

```
PUT _inference/text_embedding/mistral_embeddings
{
  "service": "mistral",
  "service_settings": {
    "api_key": "<api_key>",
    "model": "mistral-embed"
  }
}
```

Modelo externo

Novo campo *semantic_text*

```
# 3. Indexação e Embeddings - Vamos adicionar documentos ao índice
PUT test-index/_doc/1
{
  "infer_field": """A pizza Margherita é feita com uma base crocante, molho de tomate
fresco, queijo mozzarella de alta qualidade e manjericão. É a escolha perfeita para os
amantes de sabores clássicos."""
}
```

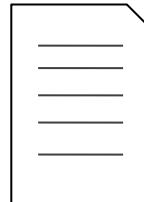
```
# Agora vamos realizar uma busca semantica
GET test-index/_search
{
  "query": {
    "semantic": {
      "field": "infer_field",
      "query": "pizza com muito queijo cremoso e sabores intensos"
    }
  }
}
```

Exemplo de busca híbrida

Traditional,
term-based score

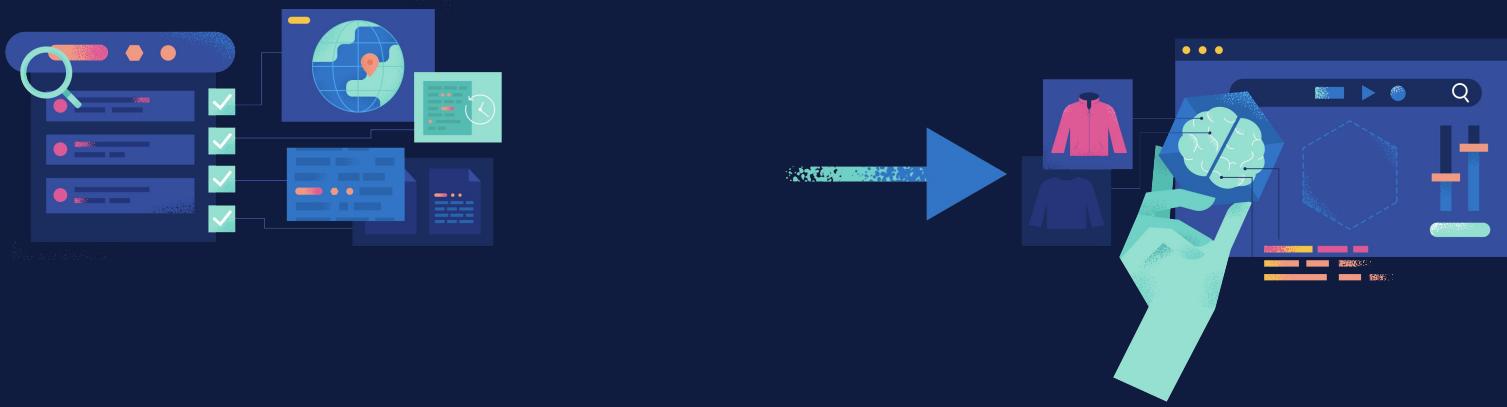
Vector similarity
score

Combine



```
GET product-catalog/_search
{
  "query": {
    "match": {
      "description": {
        "query": "summer clothes",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "desc_embedding",
    "query_vector": [0.123, 0.244, ...],
    "K": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  },
  "size": 10
}
```

Como estamos tornando o Elasticsearch ainda melhor?





Google Cloud

Microsoft Azure

NVIDIA



cohere



Hugging Face

MetaLLama

OpenAI

Tornar o **Lucene** o **MELHOR** banco de dados vetorial do mundo

Tornar o **Elasticsearch** a plataforma de busca **MAIS** abrangente e **simples** para aplicativos de IA Generativa

Ser o **membro MAIS aberto** do ecossistema de **IA Generativa**

Recursos para desenvolvedores: Elasticsearch Labs

elastic.co/search-labs

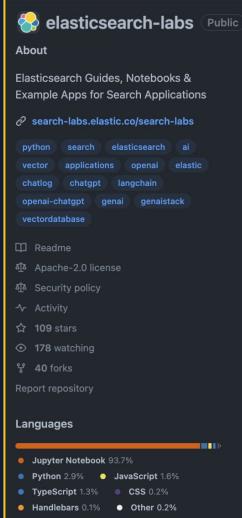
github.com/elastic/elasticsearch-labs

BLOG / ML RESEARCH

Evaluating RAG: A journey through metrics



In 2020, Meta published a paper titled “[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)”. This paper introduced a method for expanding the knowledge of Language



The screenshot shows the GitHub repository page for `search-labs.elastic.co/search-labs`. The repository is public and contains guides, notebooks, and example apps for search applications. It features a sidebar with tags like `python`, `search`, `elasticsearch`, `ai`, `vector`, `applications`, `openai`, `elastic`, `chatbot`, `chatgpt`, `langchain`, `openai-chatgpt`, `genai`, `genaistack`, and `vectordatabase`. The repository has 109 stars, 178 watchers, 40 forks, and a report repository. A languages chart shows Jupyter Notebook as the primary language at 93.7%, followed by Python (2.9%), JavaScript (1.6%), TypeScript (1.3%), CSS (0.2%), Handlebars (0.1%), and Other (0.2%).

Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

Recursos para desenvolvedores: Junte-se à Comunidade Elastic

Elastic User Groups

Estamos sempre em busca de organizadores, palestrantes e participantes.

Encontre mais eventos Elastic em todo o mundo em community.elastic.co



Meetup Elastic: Rio de Janeiro/RJ

[elastic.co/**community**](https://elastic.co/community)



Meetup Elastic: Blumenau/SC





Obrigado

Alex Salgado
Developer Advocate @ Elastic



@alexsalgadopro

f

salgado



@alexsalgadoprof

/in/alex-salgado/



@alexsalgadoprof