Desvendando os bancos de dados vetoriais e a busca vetorial

code<con> summit

@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

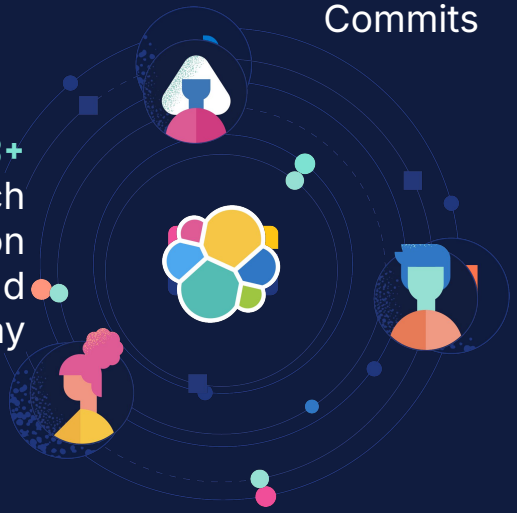
aws certified Solutions Architect Associate

aws certified Cloud Practitioner

LEVEL 2 CERTIFIED Educator Google for Education

Microsoft CERTIFIED

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



148K+
Commits

6B+
Search requests on Elastic Cloud everyday

183K+
Pull Requests

3500+
Unique Authors

103K+
GitHub Stars

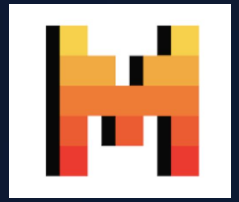
4.2B+
Downloads

elastic

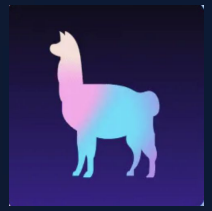




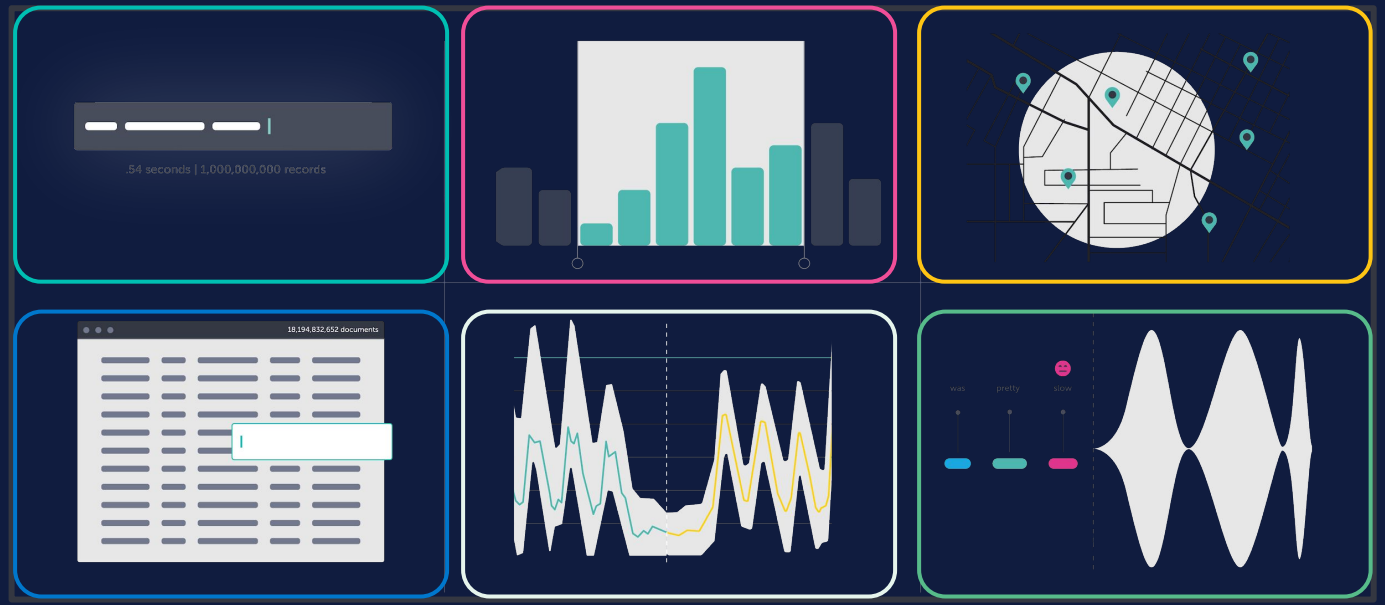









A Busca é muito mais que uma caixa de pesquisa



elastic

```
Select *
From cadastro
Where nome =
"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".

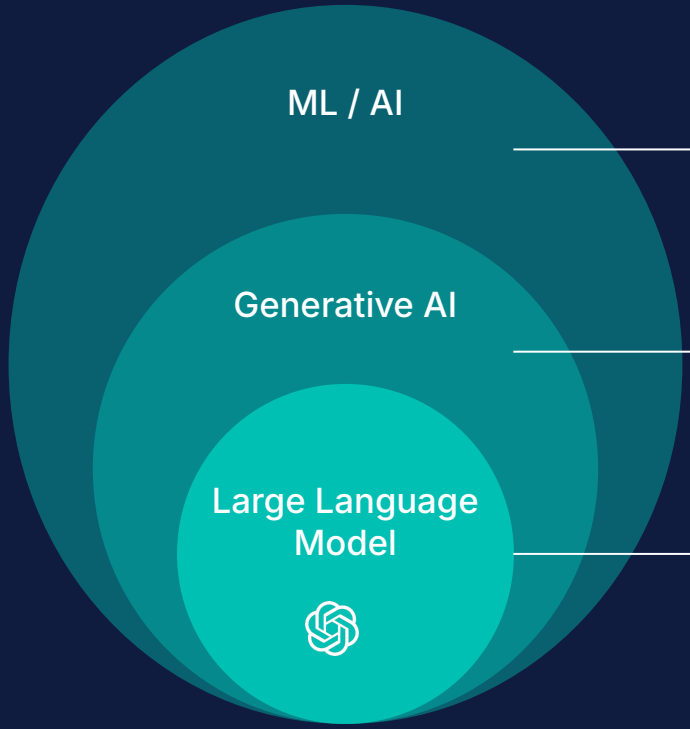
A Busca é muito mais que uma caixa de pesquisa

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

elastic

Aplicações interessantes de ML, IA generativa e LLMs



	O que é ?	Aplicações
ML / AI	Algoritmos programados para fazer previsões com base em dados	Reconhecimento de imagem, processamento de linguagem natural, reconhecimento de fala
Generative AI	Algoritmos de IA projetados para criar novos dados	Chatbots, geradores de texto, geradores de imagem, geradores de música
Large Language Model	Algoritmos de aprendizado profundo que podem gerar texto	Geradores de texto, tradução, escrita, resposta a perguntas

elastic

O poder da
IA generativa

reside nos
seus dados proprietários

Three solutions powered by one stack

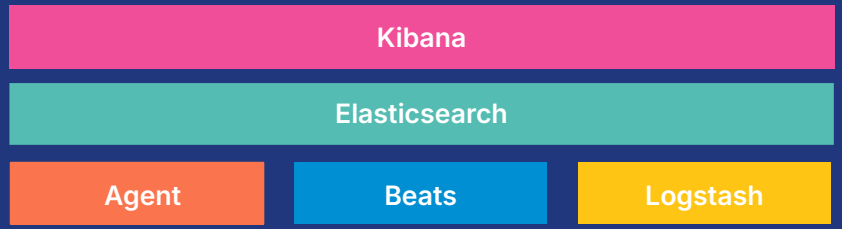
3 solutions

Enterprise Search

Observability

Security

Powered by the Elastic Stack



Deployed anywhere

Elastic Cloud

Elastic Cloud Enterprise

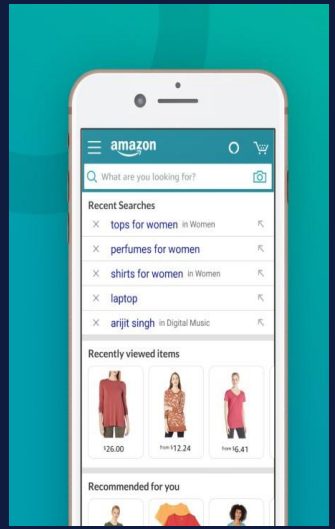
Elastic Cloud on Kubernetes

Saas

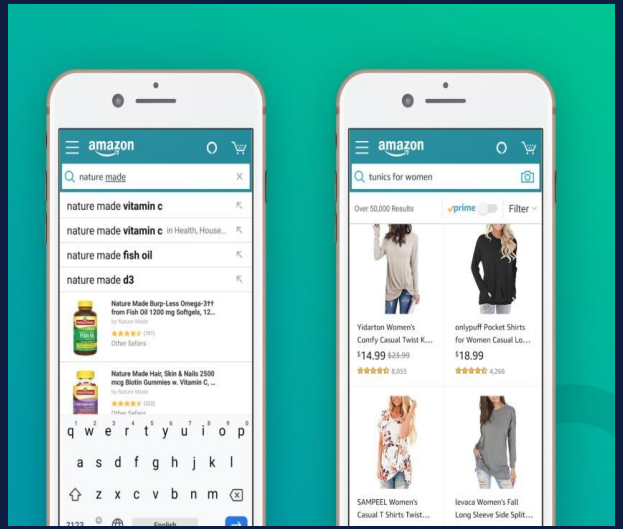
Orchestration

elastic

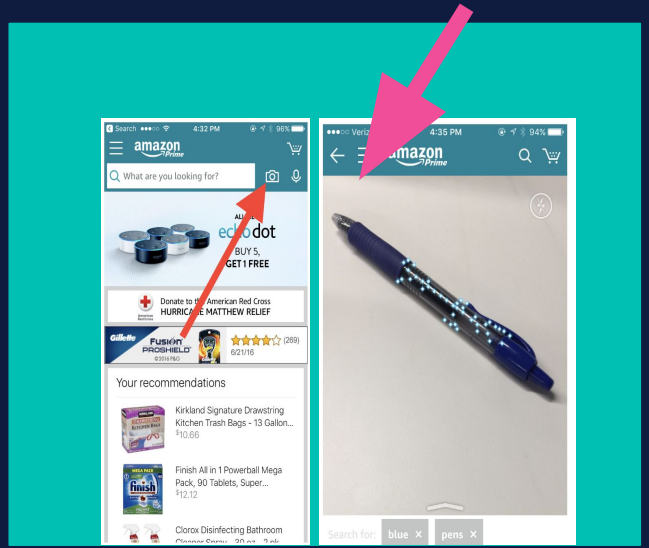
As pessoas estão mudando o jeito de pesquisar



1-Busca textual



2-Busca textual + Semantica

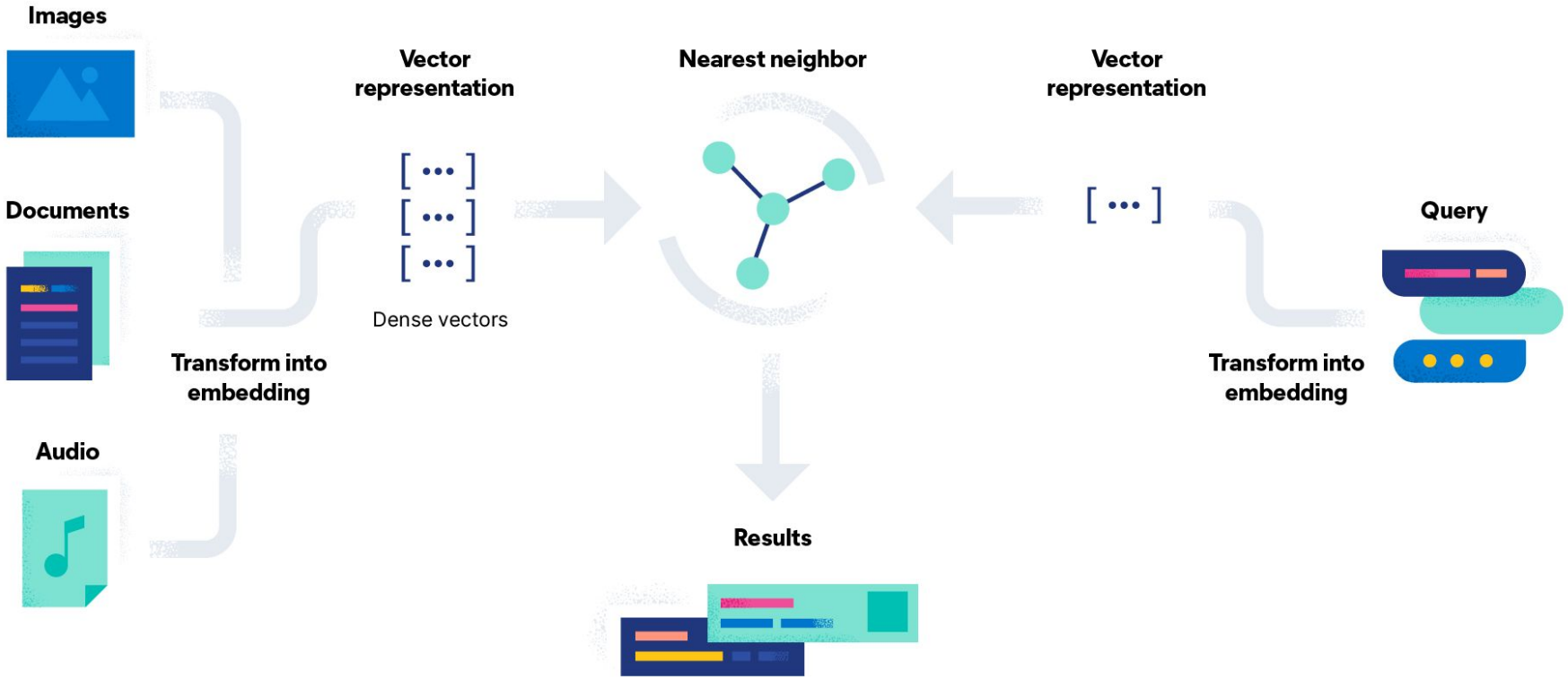


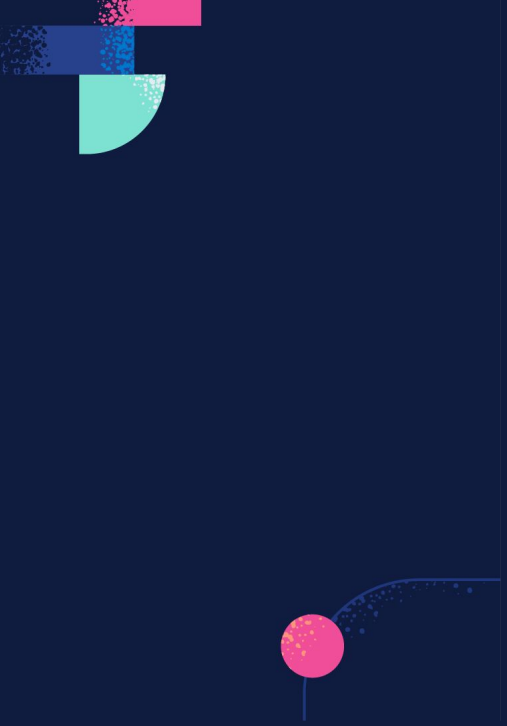
3-Busca por Imagem e Som

elastic

Elasticsearch: You Know, for Vector Search

Vector Search





O que é um Vetor?

elastic

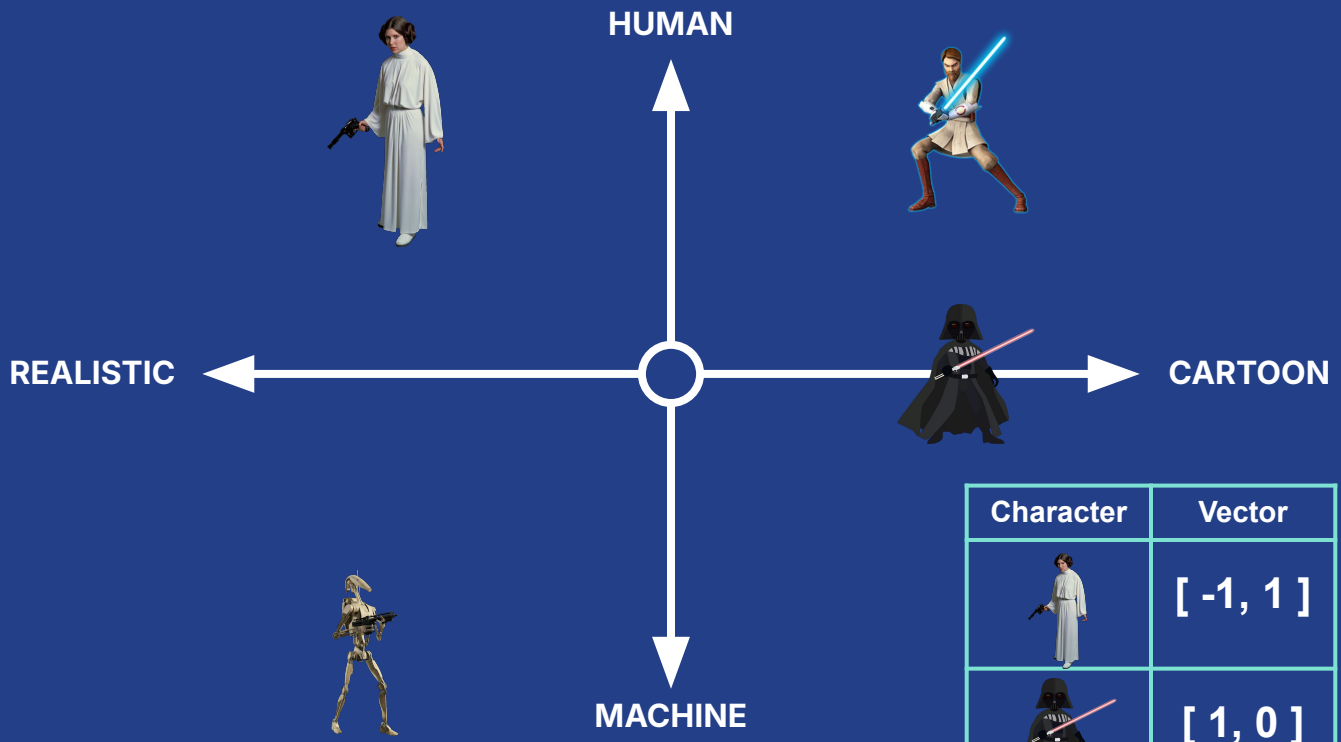
Embeddings represent your data

Example: 1-dimensional vector

Character	Vector
	[-1]
	[1]

elastic

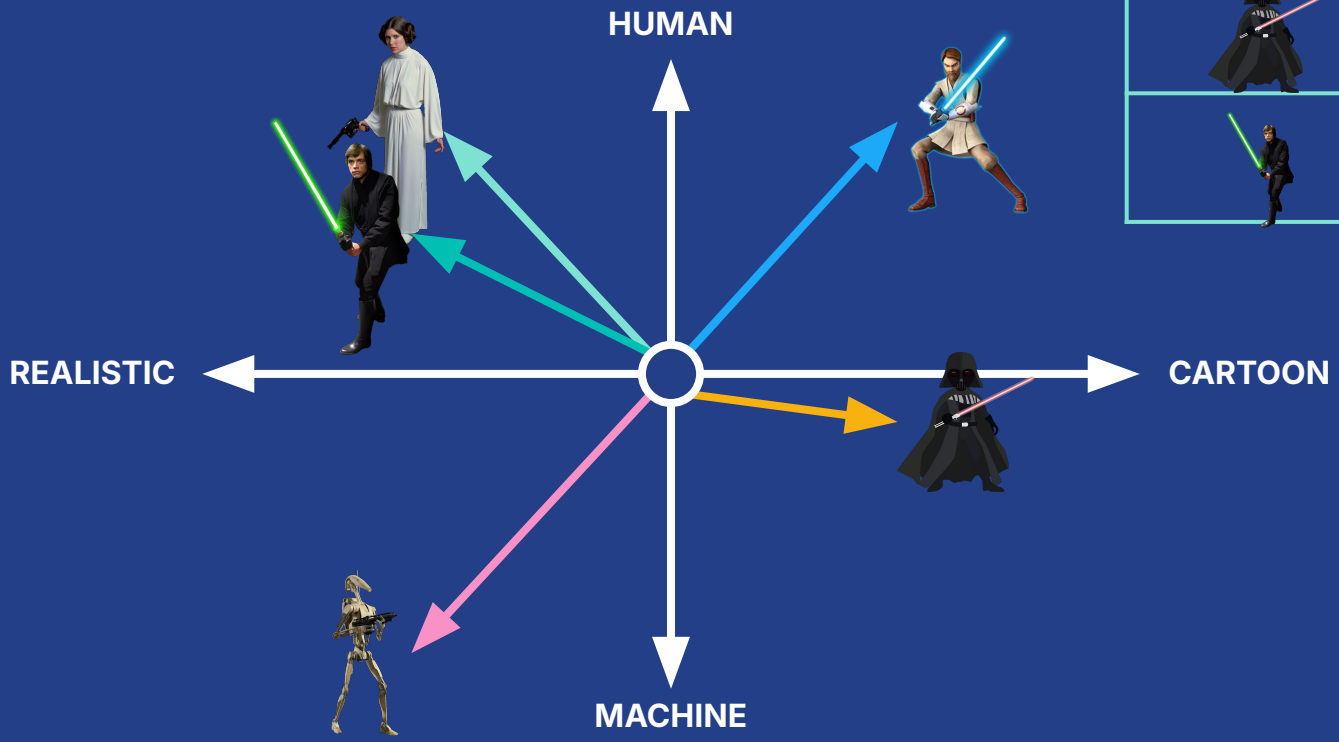
Multiple dimensions represent different data aspects



Character	Vector
	[-1, 1]
	[1, 0]

elastic

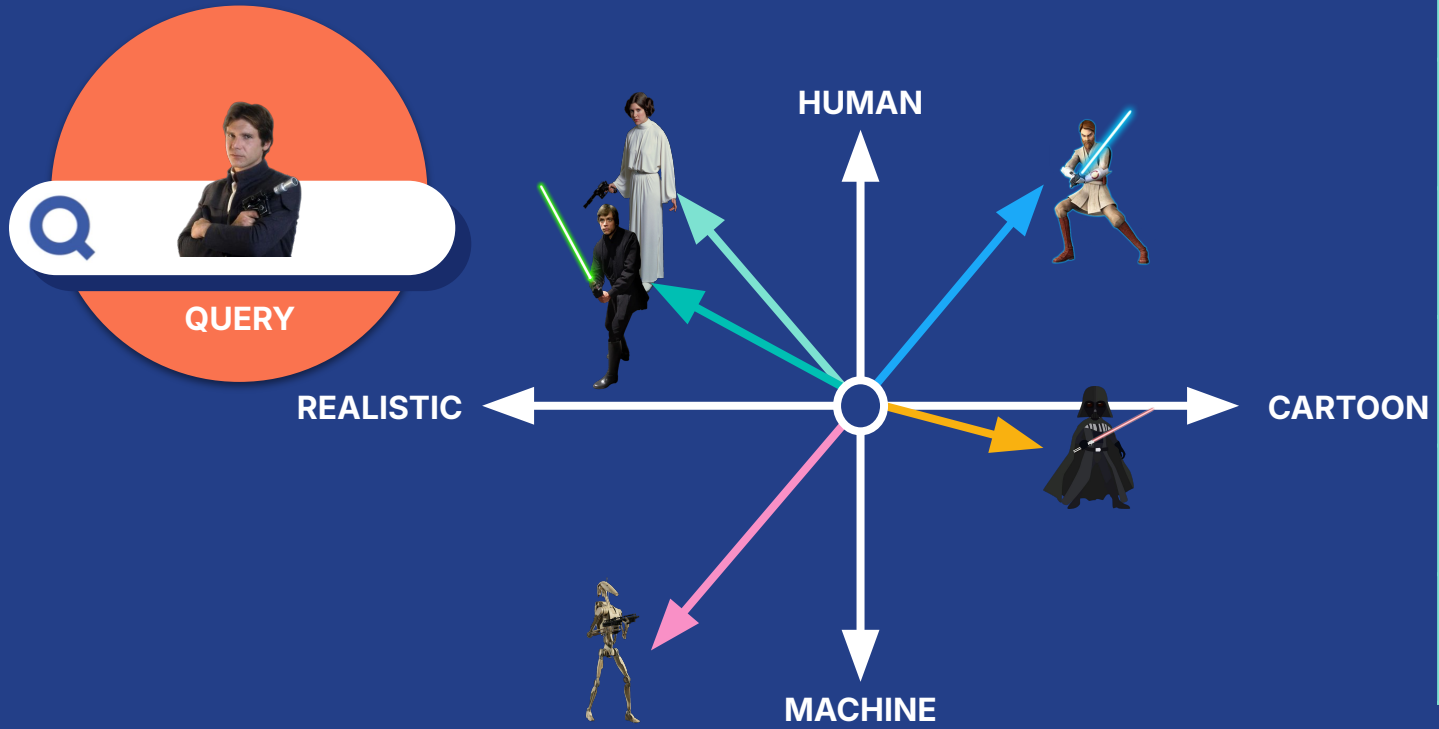
Similar data is grouped together



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

elastic

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



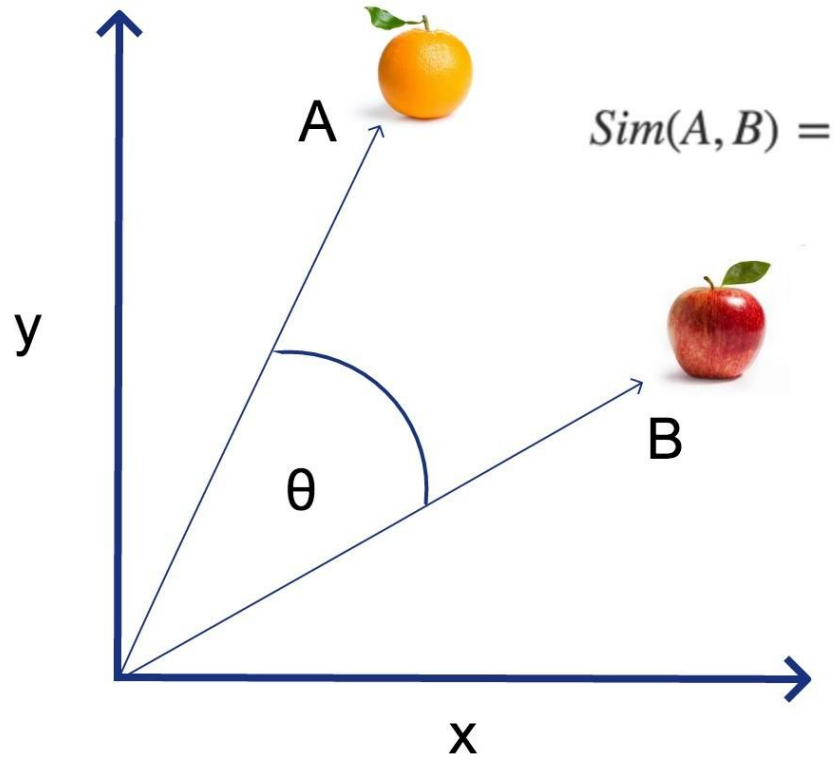
Relevance	Result
Query	
1	
2	
3	
4	
5	

elastic

Como essa busca por similaridade realmente funciona?

elastic

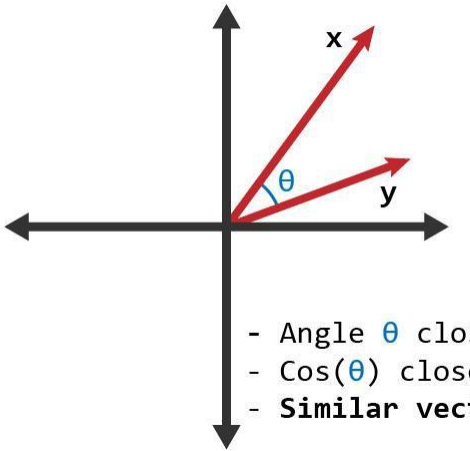
Similarity: Cosine



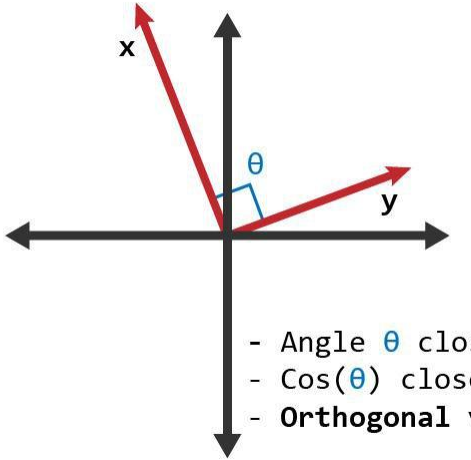
$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

elastic

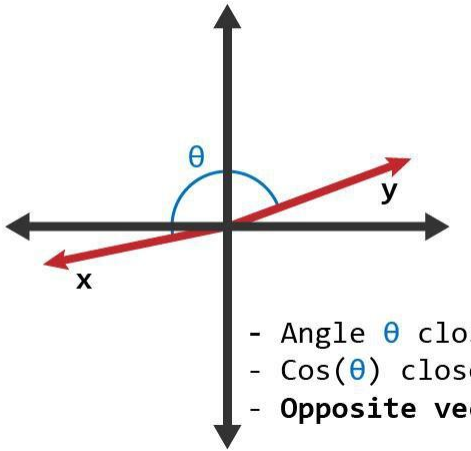
Similarity: Cosine



- Angle θ close to 0
- Cos(θ) close to 1
- **Similar vectors**



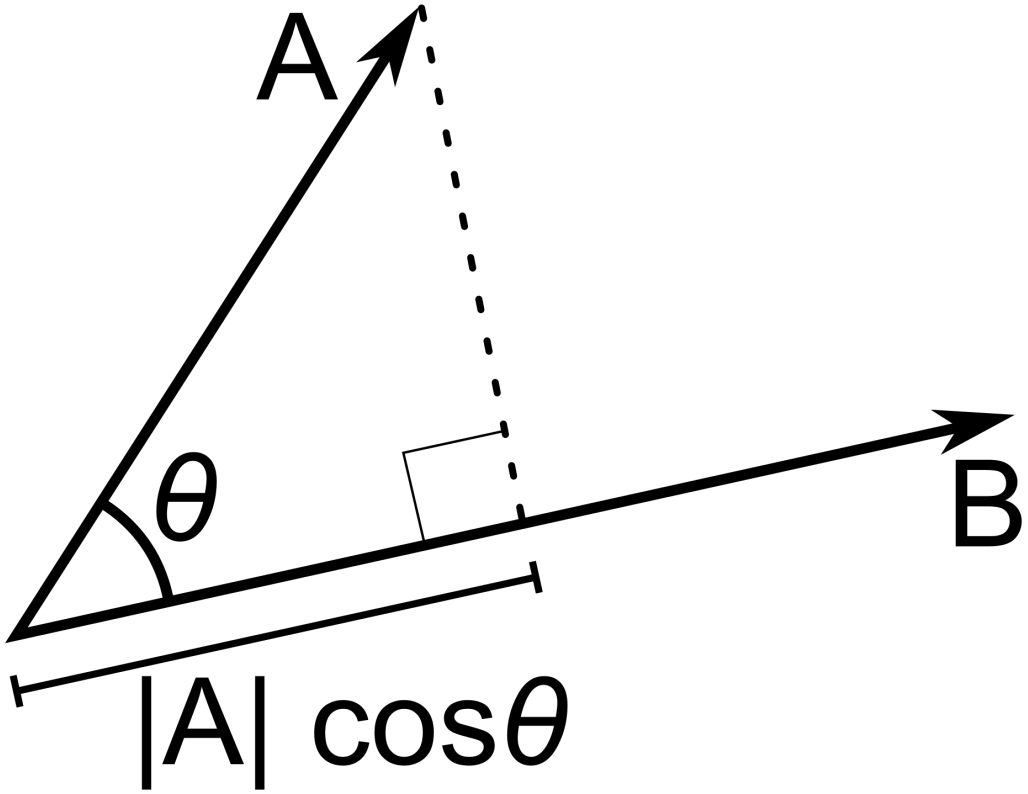
- Angle θ close to 90
- Cos(θ) close to 0
- **Orthogonal vectors**



- Angle θ close to 180
- Cos(θ) close to -1
- **Opposite vectors**

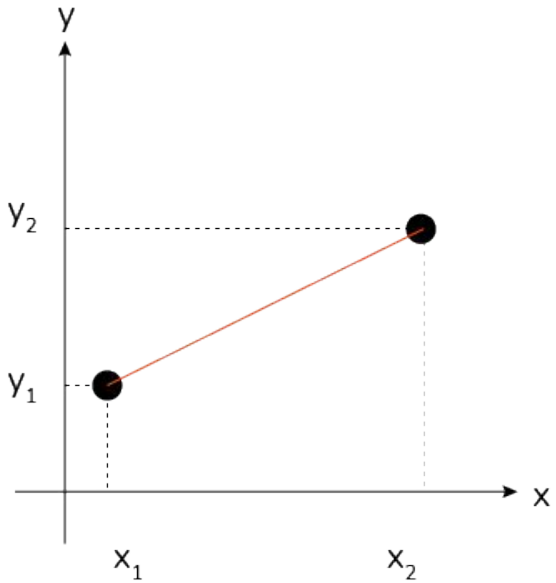
elastic

Similarity: Dot Product



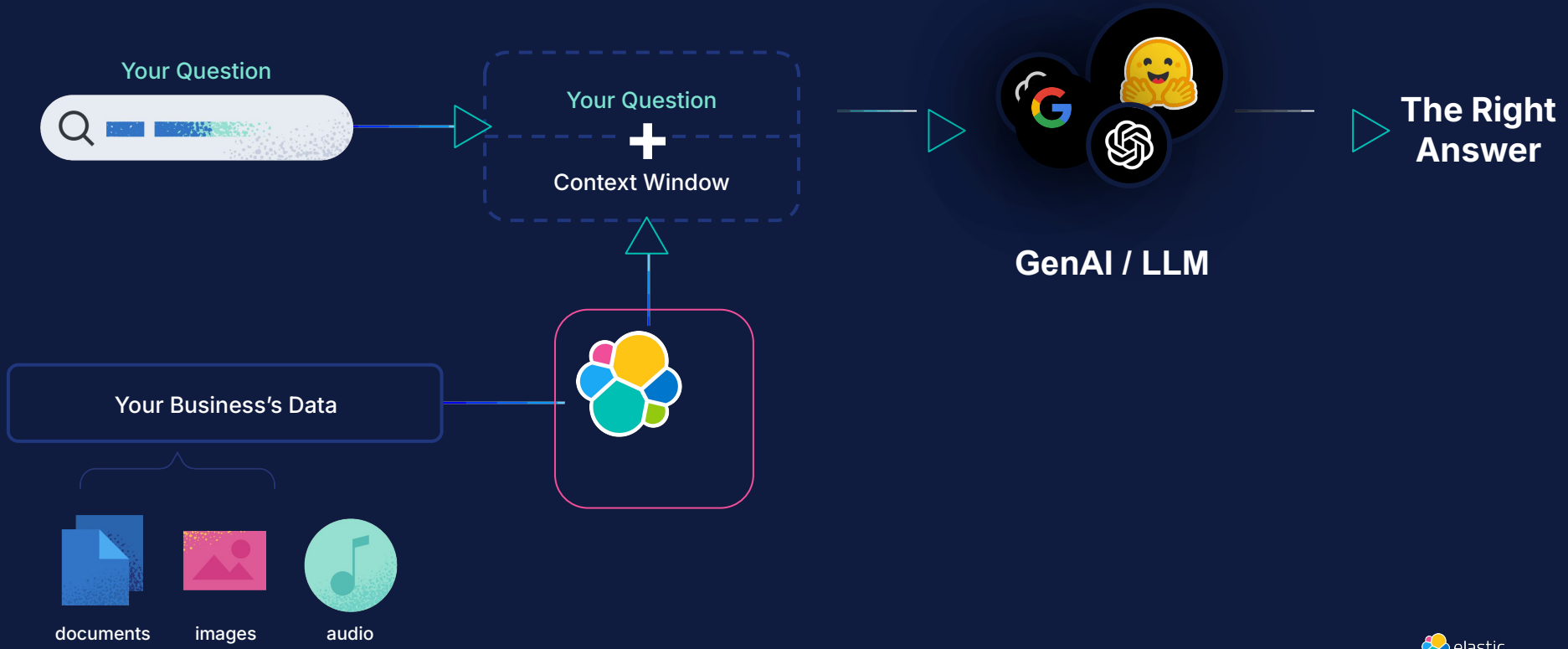
elastic

Similarity: Euclidean / L2



elastic

A importância do banco vetorial para a Arquitetura RAG



elastic

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

elastic

Os ingredientes da busca com IA generativa

Busca de texto, vetorial e híbrida

+

Capacidades de banco de dados vetorial

+

Escolha de modelos de embedding

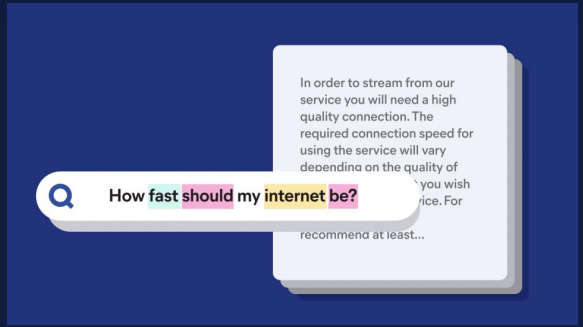
+

Ferramentas e APIs para construir aplicativos de busca

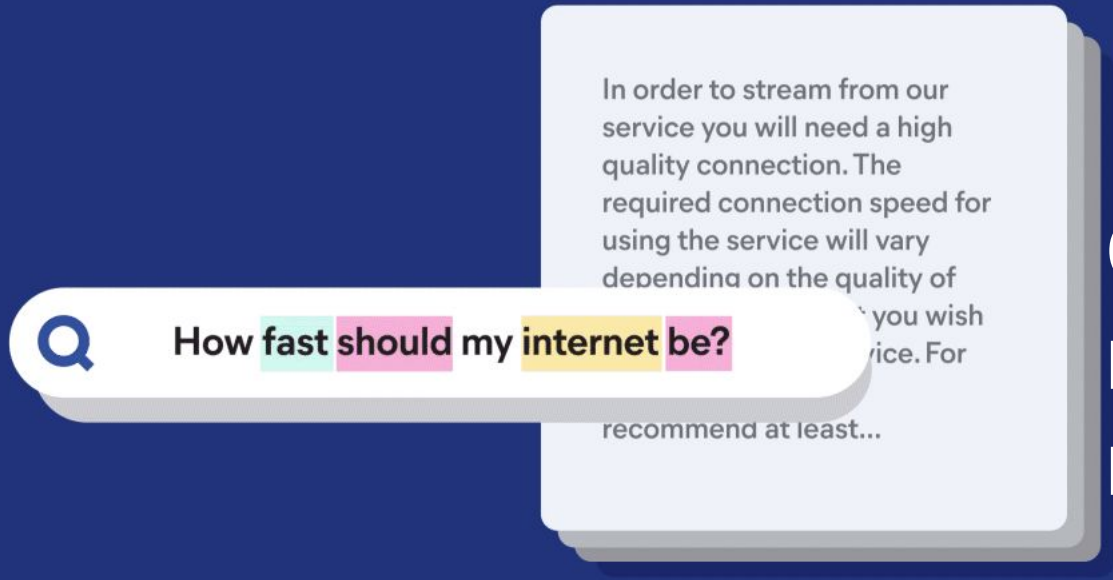
elastic

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**



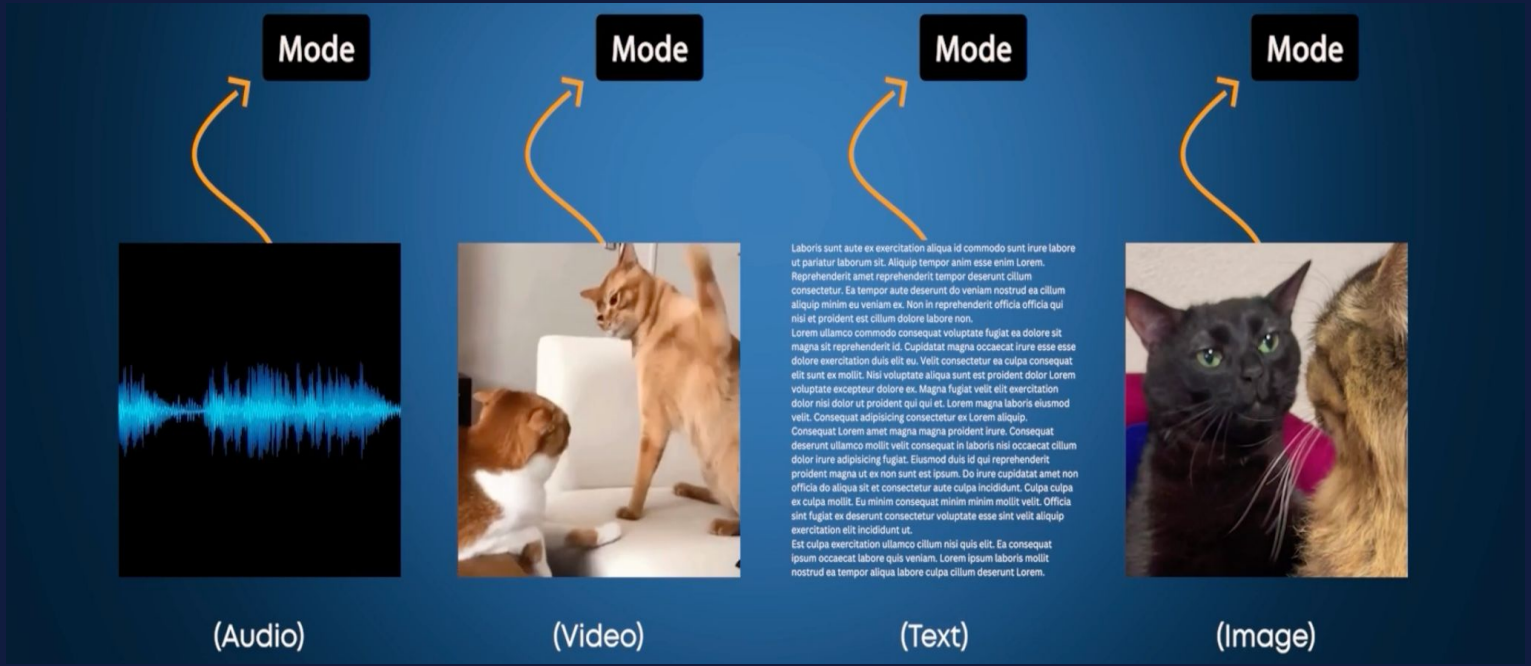
elastic



O que é um modelo multimodal?

elastic

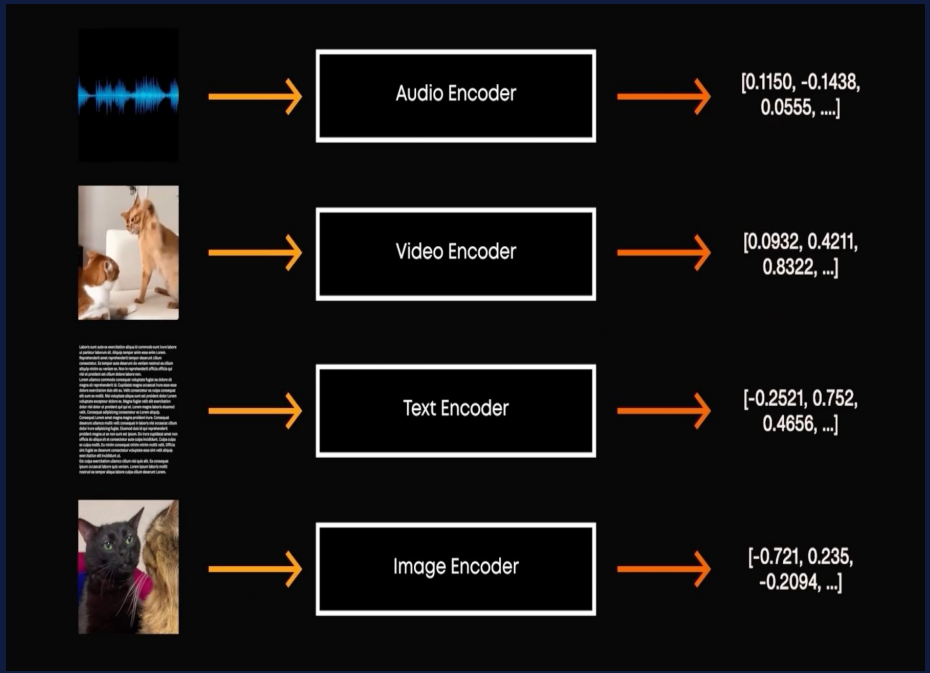
Modelos Multimodais

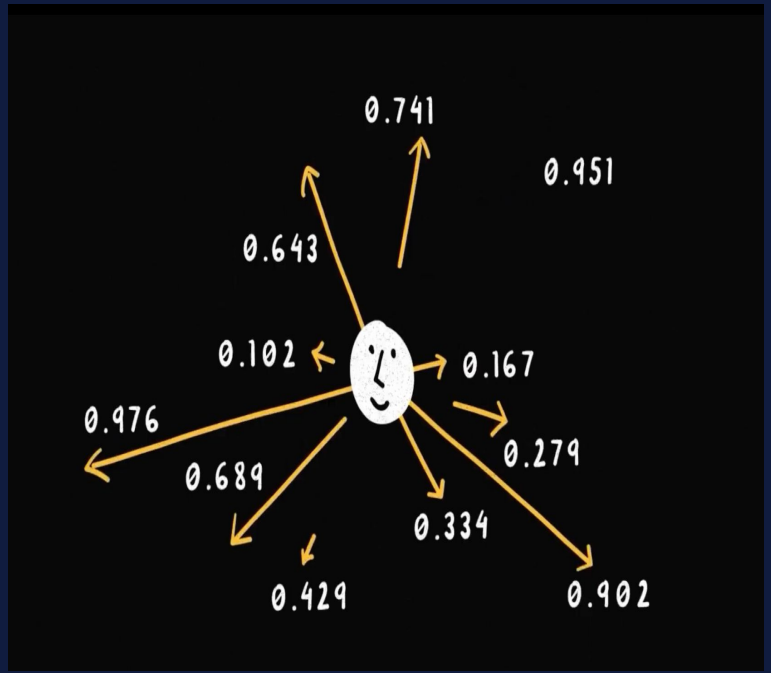


elastic

Modelos Multimodais

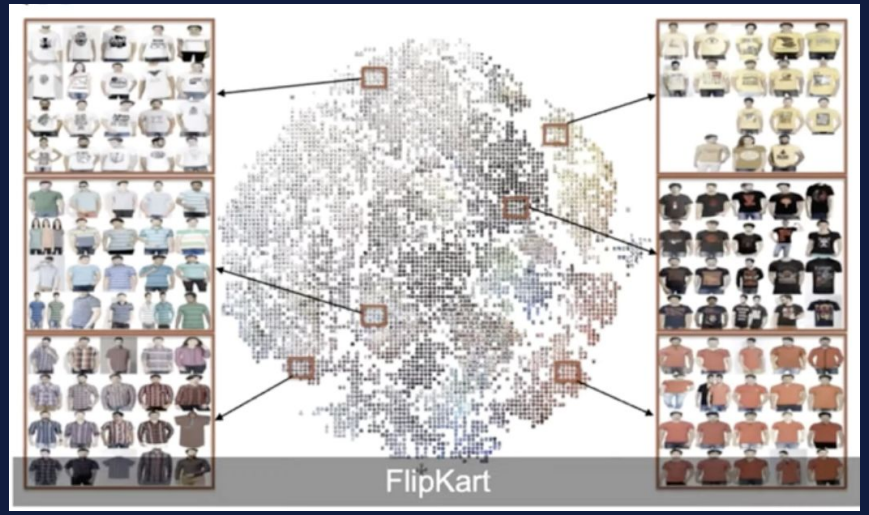
embeddings



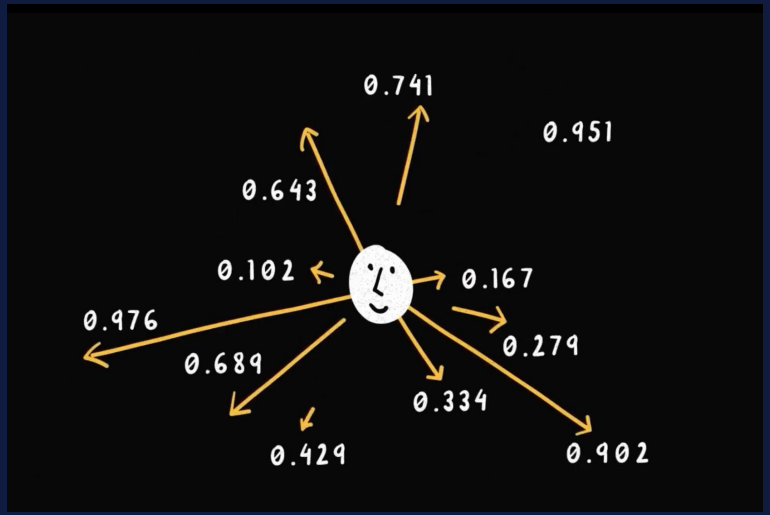


elastic

Busca por Imagem

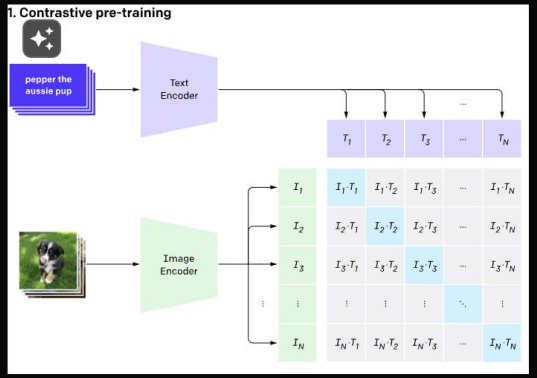


embeddings

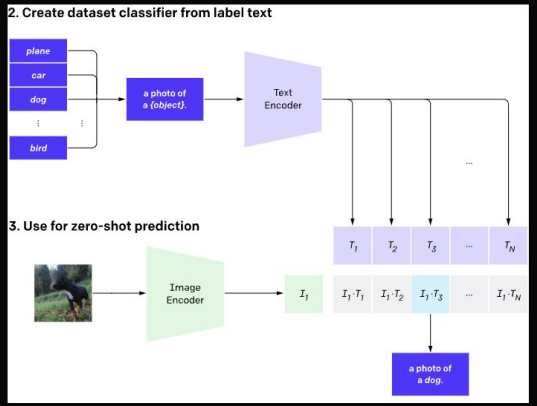


elastic

Anatomia de um modelo multimodal - CLIP da OpenAI



Overview A



Overview B

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 imag

```
from sentence_transformers import SentenceTransformer

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

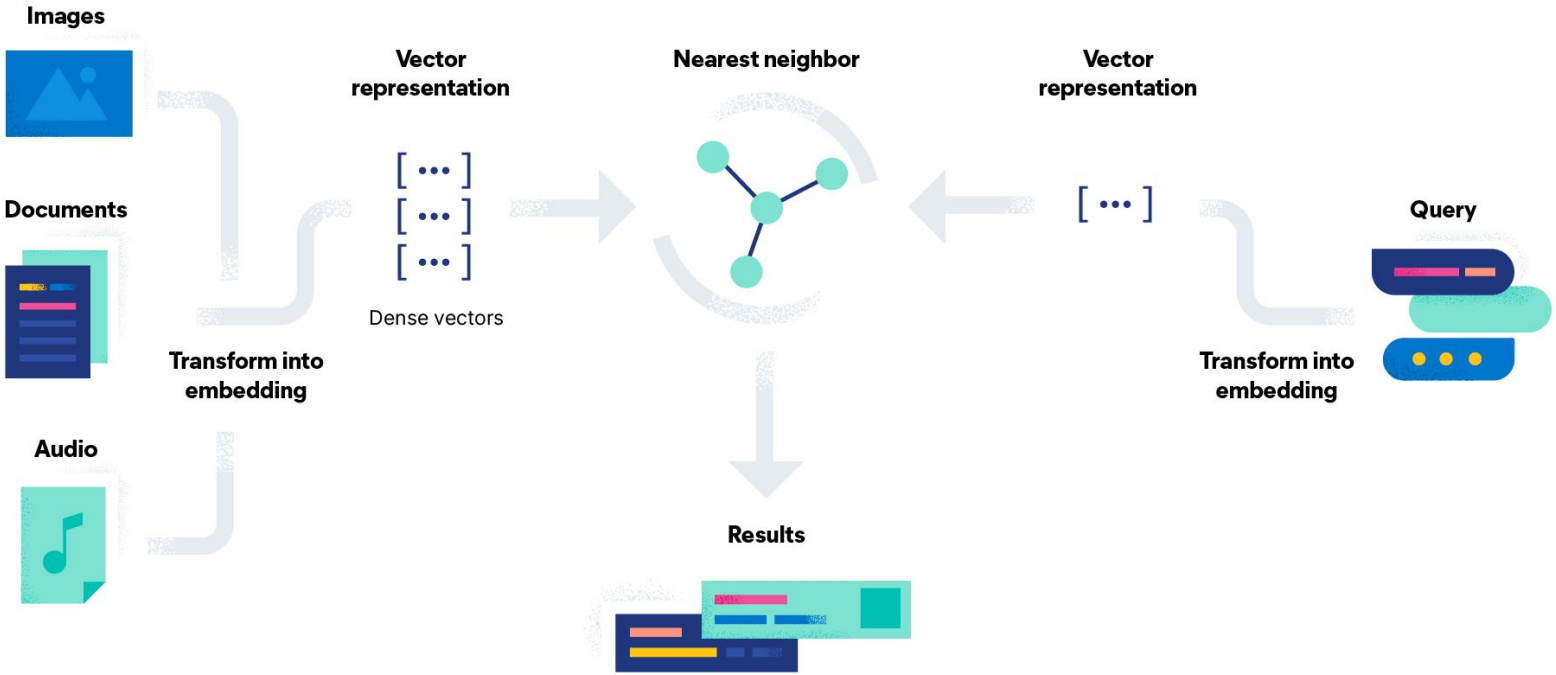
elastic

Pirâmide de Aprendizado de William Glasser

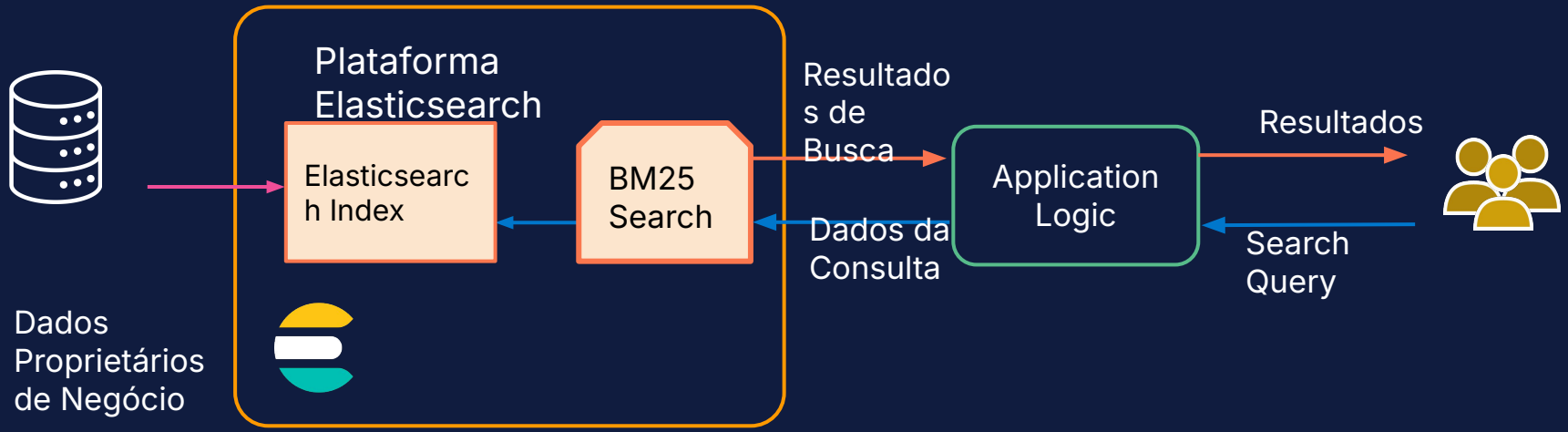
elastic

Evolução da arquitetura de busca

Consultas também são vetorizadas

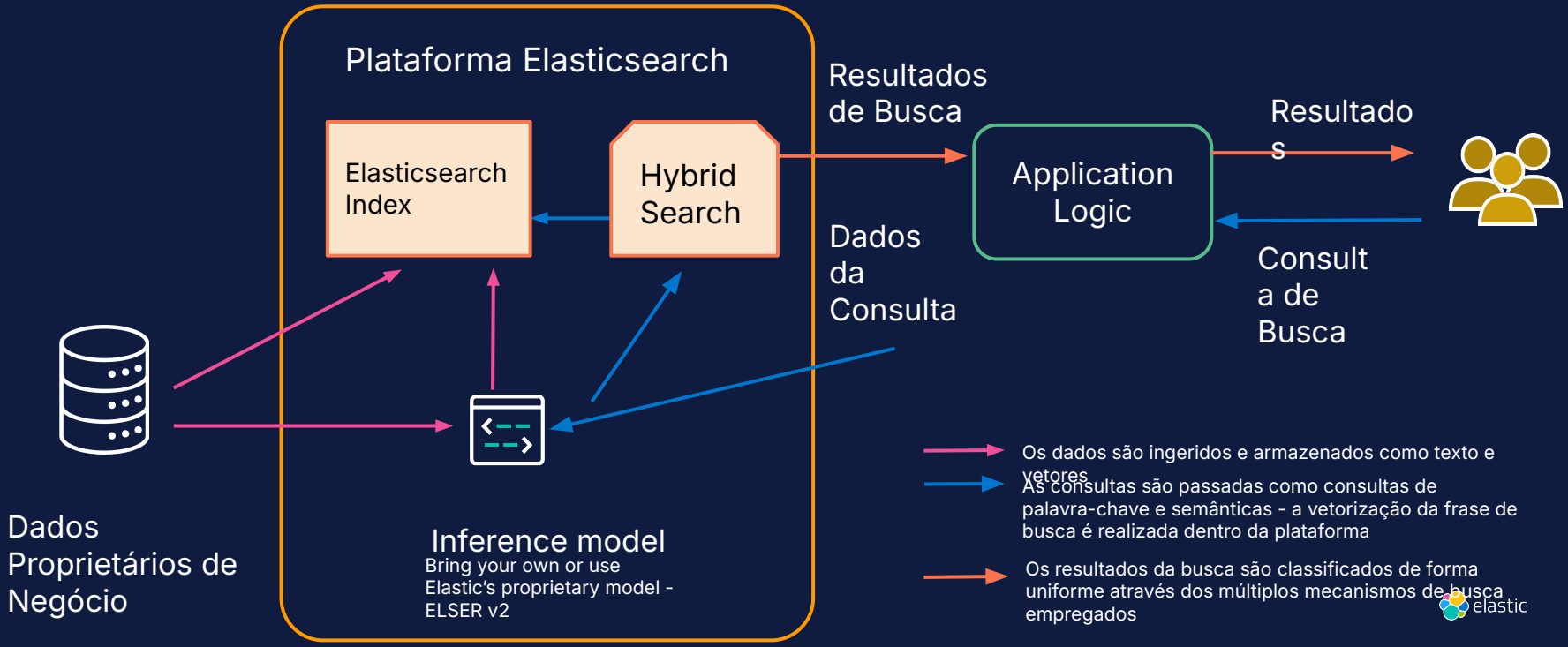


Arquitetura de Busca típica de busca textual (palavras-chave)



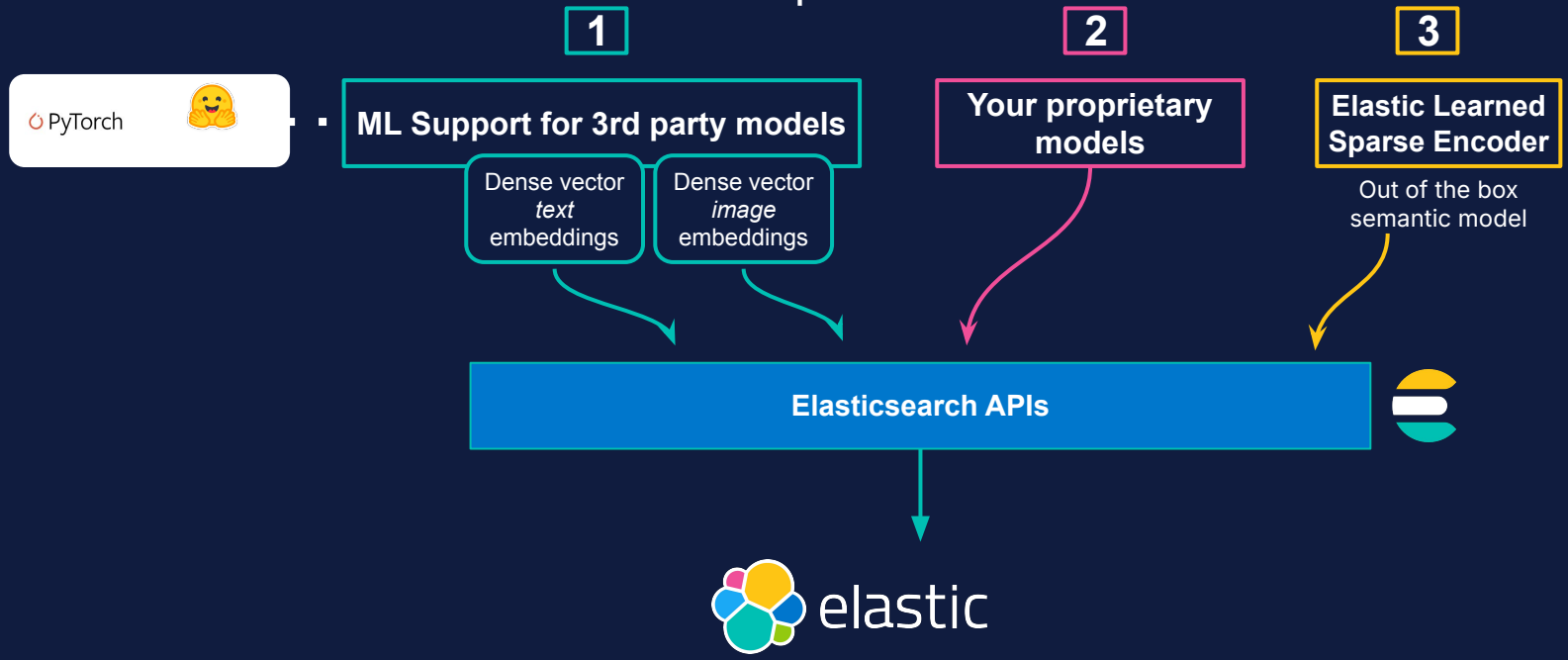
A. Consulta de Busca
B. Execução da Busca BM25

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



Embeddings e Elasticsearch: três opções

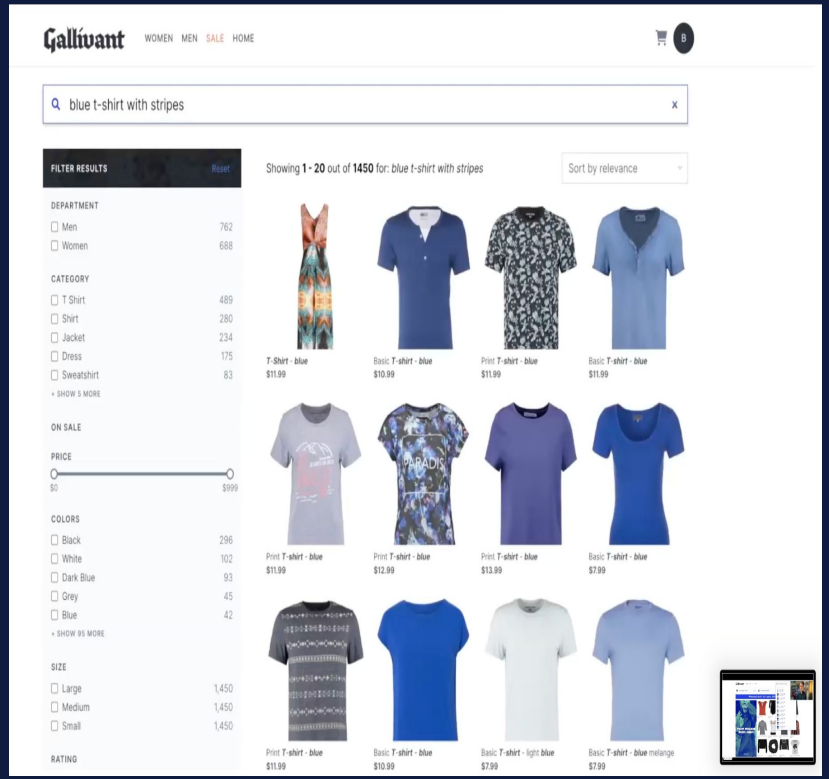
modelos de terceiros, traga seus próprios modelos, nosso codificador esparso



Evolução dos banco de dados para suportar essas mudanças

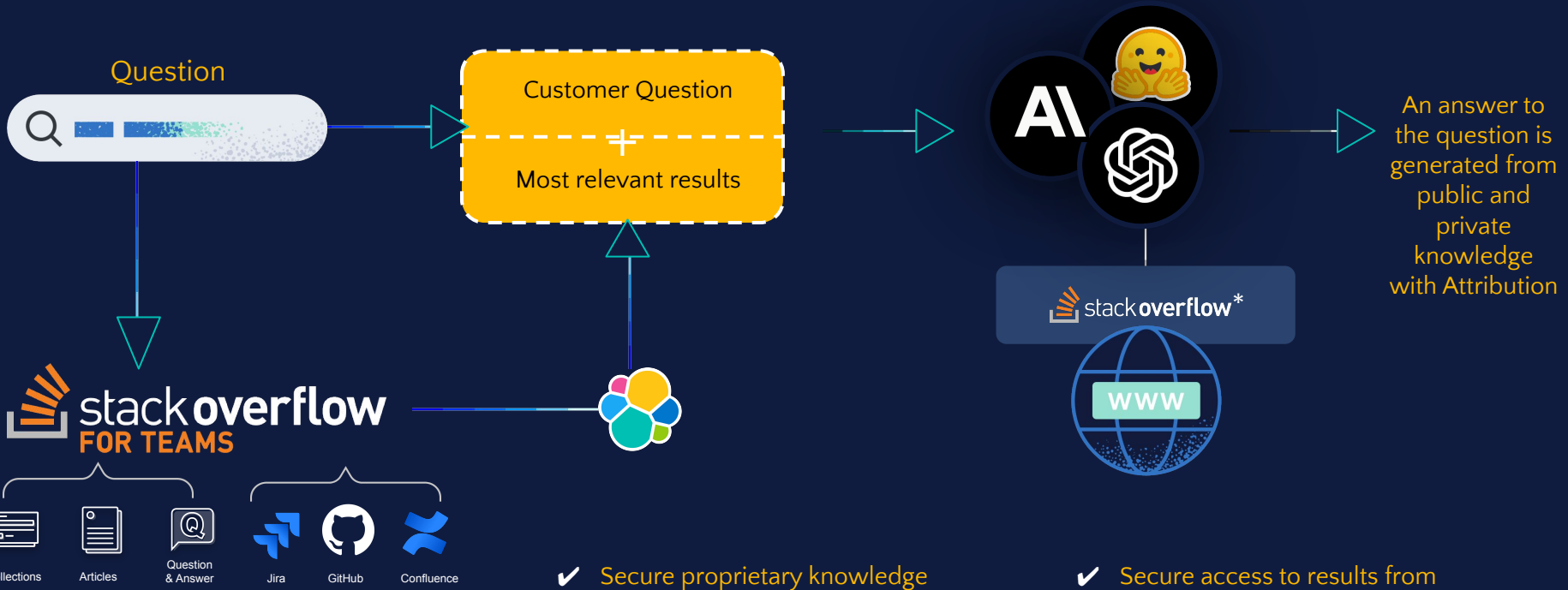
Exemplos de aplicações

Busca em e-commerce



elastic

Case : Stackoverflow Teams



- ✔ Secure proprietary knowledge
- ✔ Secure document & role-based access
- ✔ Secure access to results from StackOverflow.com

*IDE Extension, coming to website soon

elastic

stackoverflow

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.

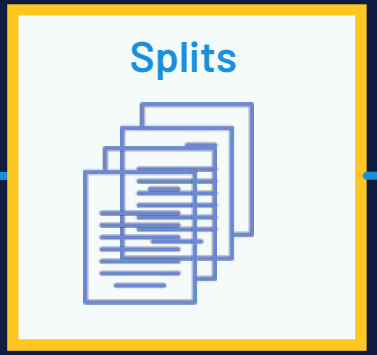
elastic

Exemplo com chatbot

Carregamento do banco vetorial

Carregamento de documentos

divisão/splitting



armazenaemnto/storag

Recuperação/ Retrieval

Storage

Retrieval

Output

resposta

Query NLP

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"
    }
}
```

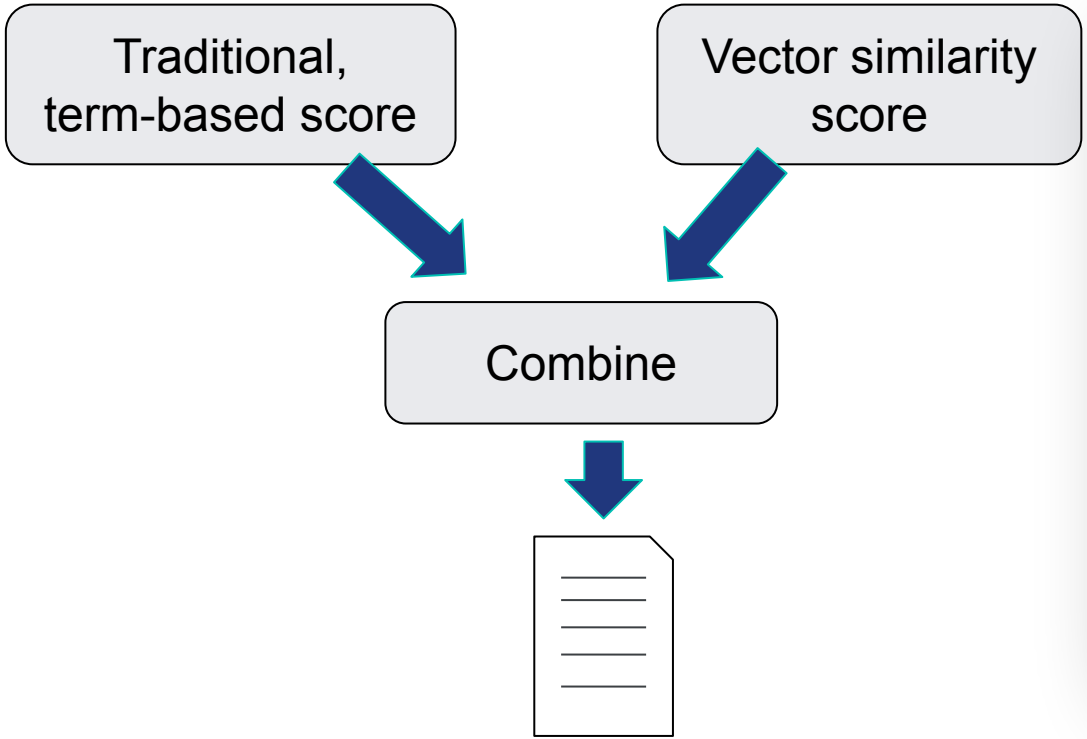
Modelo externo

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"
    }
  }
}
```

Exemblo de busca híbrida

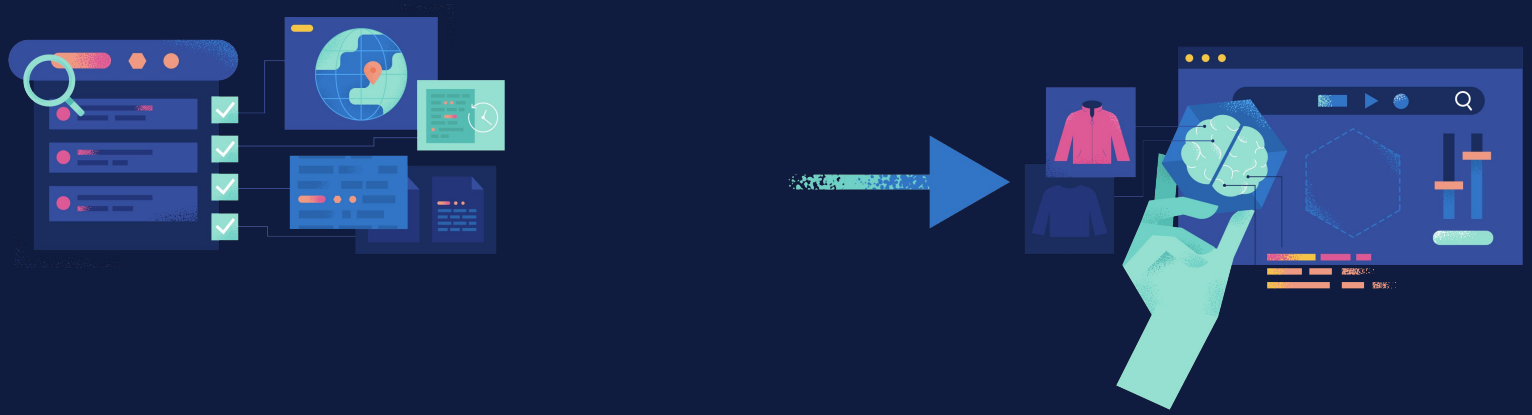



```
GET product-catalog/_search

{
  "query": {
    "match": {
      "description": {
        "query": "summer clothes",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "desc_embbeding",
    "query_vector": [0.123, 0.244, ...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
       "term": {
         "department": "women"
       }
     }
   },
  "size": 10
}
```

elastic

Como estamos tornando o Elasticsearch ainda melhor?



elastic

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

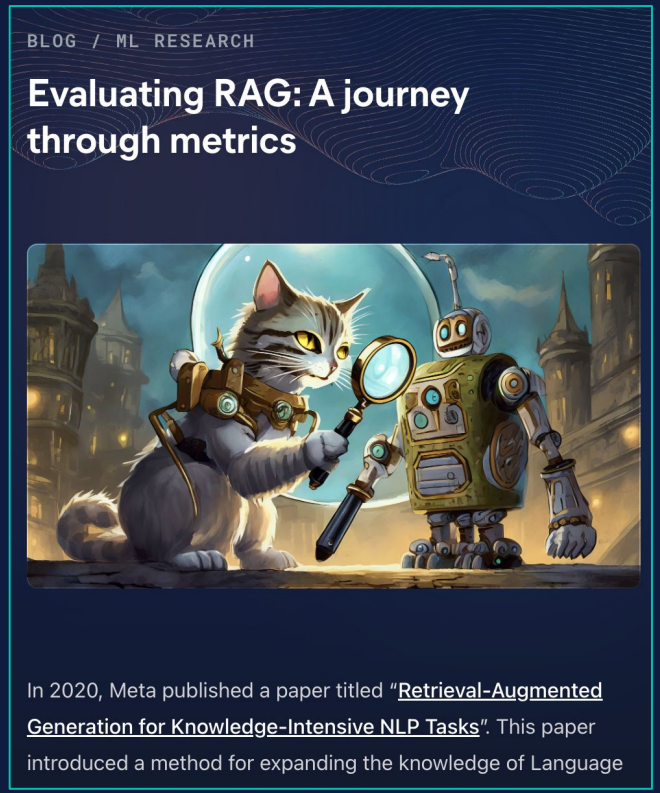
Tornar o **Elasticsearch** a plataforma de busca **MAIS** abrangente e **simples** para aplicativos de IA Generativa

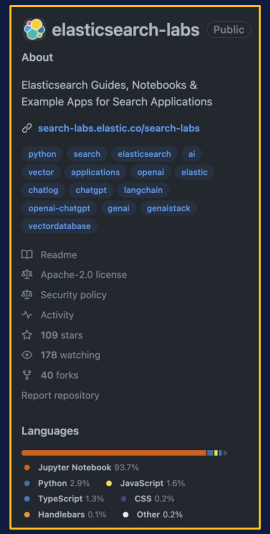
Ser o **membro MAIS aberto** do ecossistema de **IA Generativa**

Recursos para desenvolvedores: **Elasticsearch Labs**

elastic.co/**search-labs**

github.com/elastic/**elasticsearch-labs**





Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

elastic

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

Meetups

elastic.co/**community**

Meetup Elastic: Blumenau/SC

elastic

Obrigado

Alex Salgado
Developer Advocate @ Elastic

code<con> summit

@alexsalgadoprof

@alexsalgadopro

f

salgado

@alexsalgadoprof

/in/alex-salgado/