



# Image Search

Alex Salgado  
Developer Advocate @ Elastic

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





**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





**Alex Salgado**  
Senior Developer  
Advocate LATAM



@alexsalgadoprof



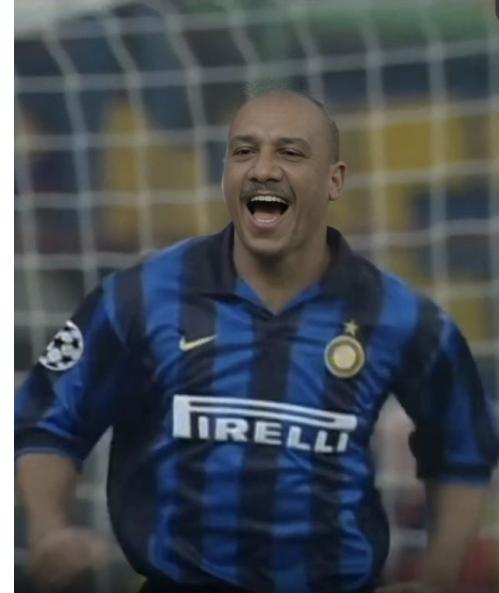
salgado



@alexsalgadoprof



/in/alex-salgado/



# Preocupações em torno da IA Generativa.

---

80%

Dados mundiais são não-estruturados

[KPMG Generative AI Survey](#)

[The Prompt: Generative AI survey | Google Cloud Blog](#)



# 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



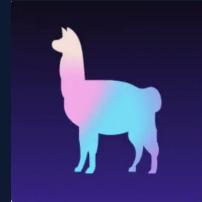
Elastic Cloud  
Enterprise



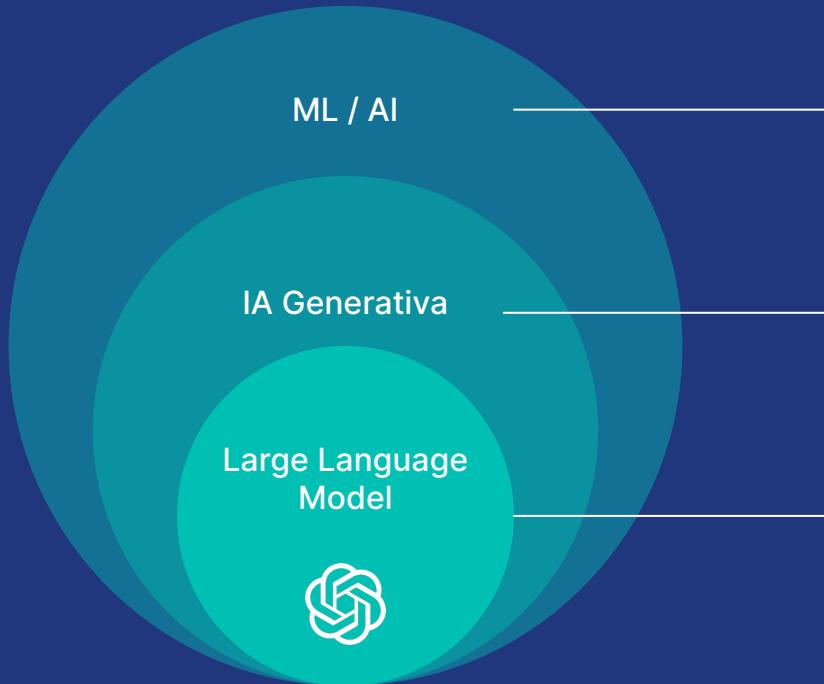
Elastic Cloud  
on Kubernetes

SaaS

Orchestration



# Conceitos básicos de ML, IA Generativa e LLMs



## O que?

Algoritmos programados para aprender o comportamento dos dados e fazer previsões

## Casos de uso

Detecção de anomalias, forecasting, reconhecimento de imagem, PLN

Algoritmos programados para criar novos dados

Chatbots, geradores de texto, imagem e música

Algoritmos (Deep Learning) treinados com grandes volumes de dados e programados para criar novos dados

Chatbots, geradores de texto, tradutores, geradores de código, aplicativos de pergunta e resposta

# Blog referência

<https://www.elastic.co/search-labs/finding-your-puppy-with-image-search>



## 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



# Elasticsearch: You Know, for Search



How fast should my internet be?

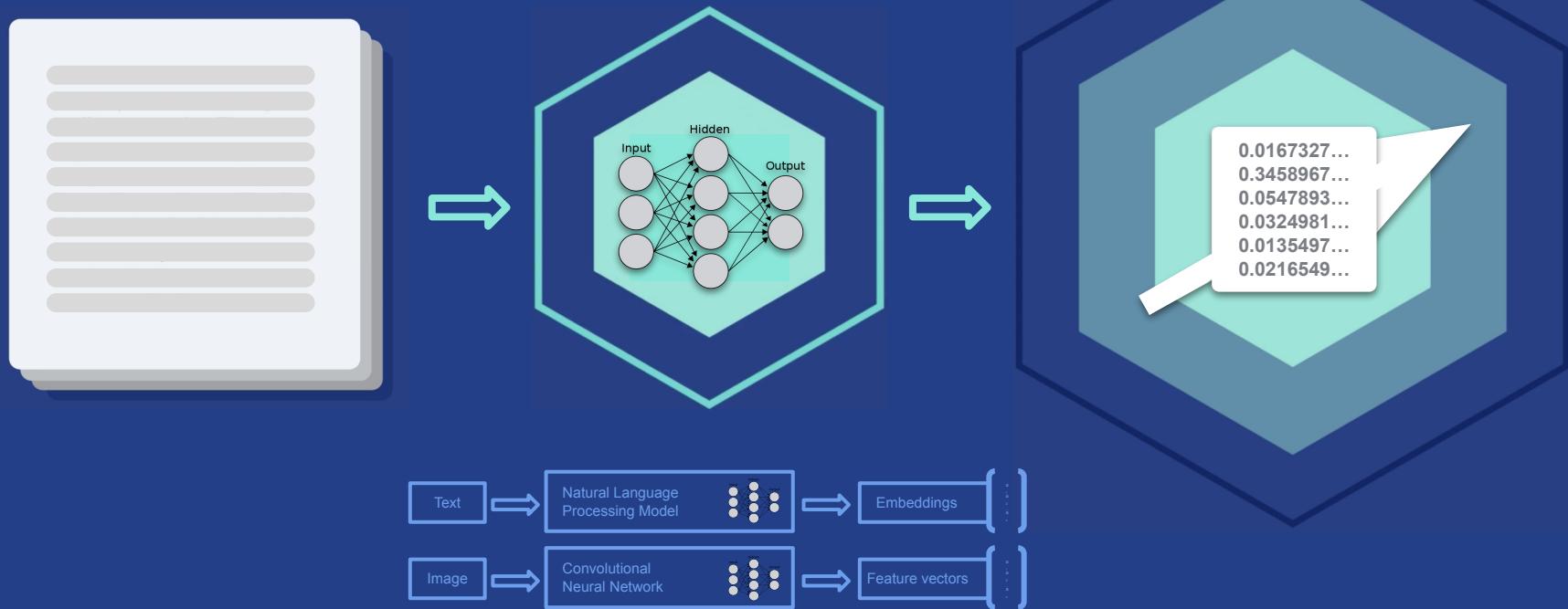
In order to stream from our service you will need a high quality connection. The required connection speed for using the service will vary depending on the quality of

you wish  
vice. For

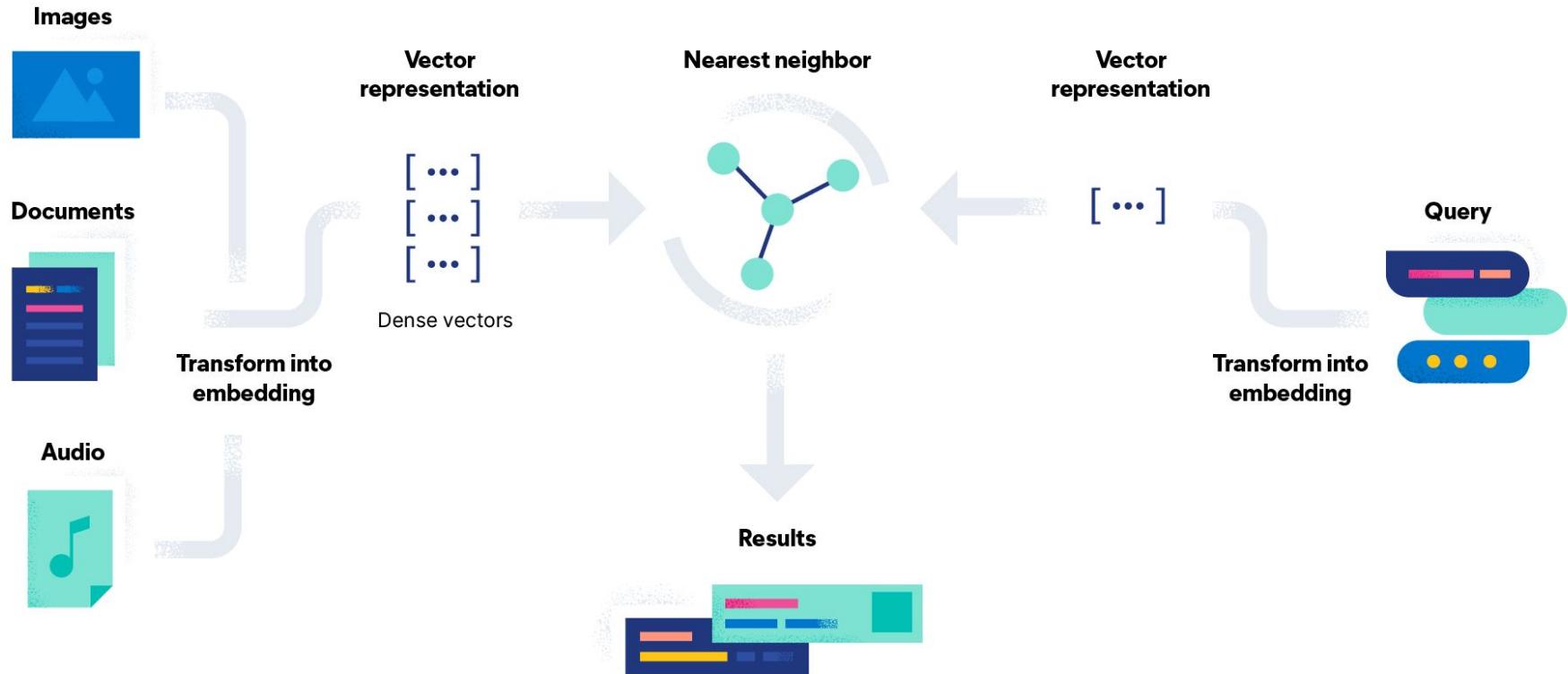
recommend at least...

# O que é similaridade de vetores?

Converta dados em representações vetoriais onde as distâncias representam similaridade.



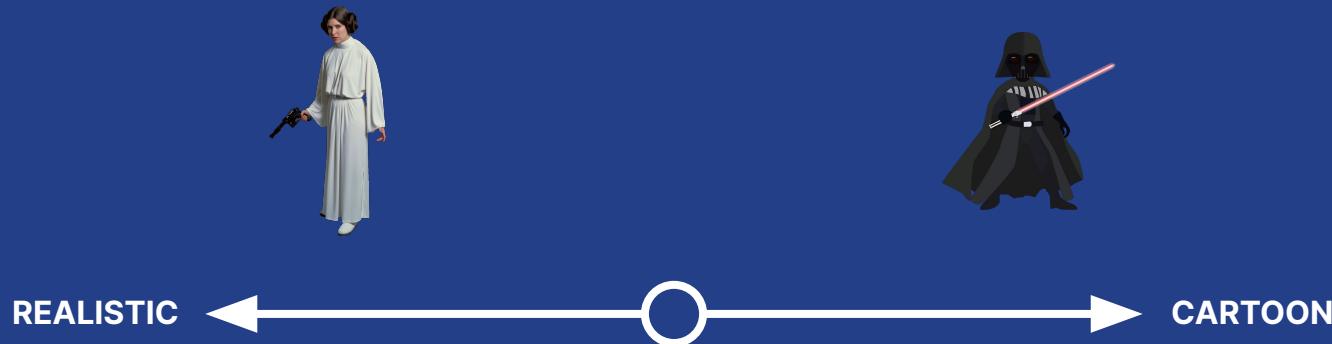
# Queries are also vectorized



# What is a Vector?

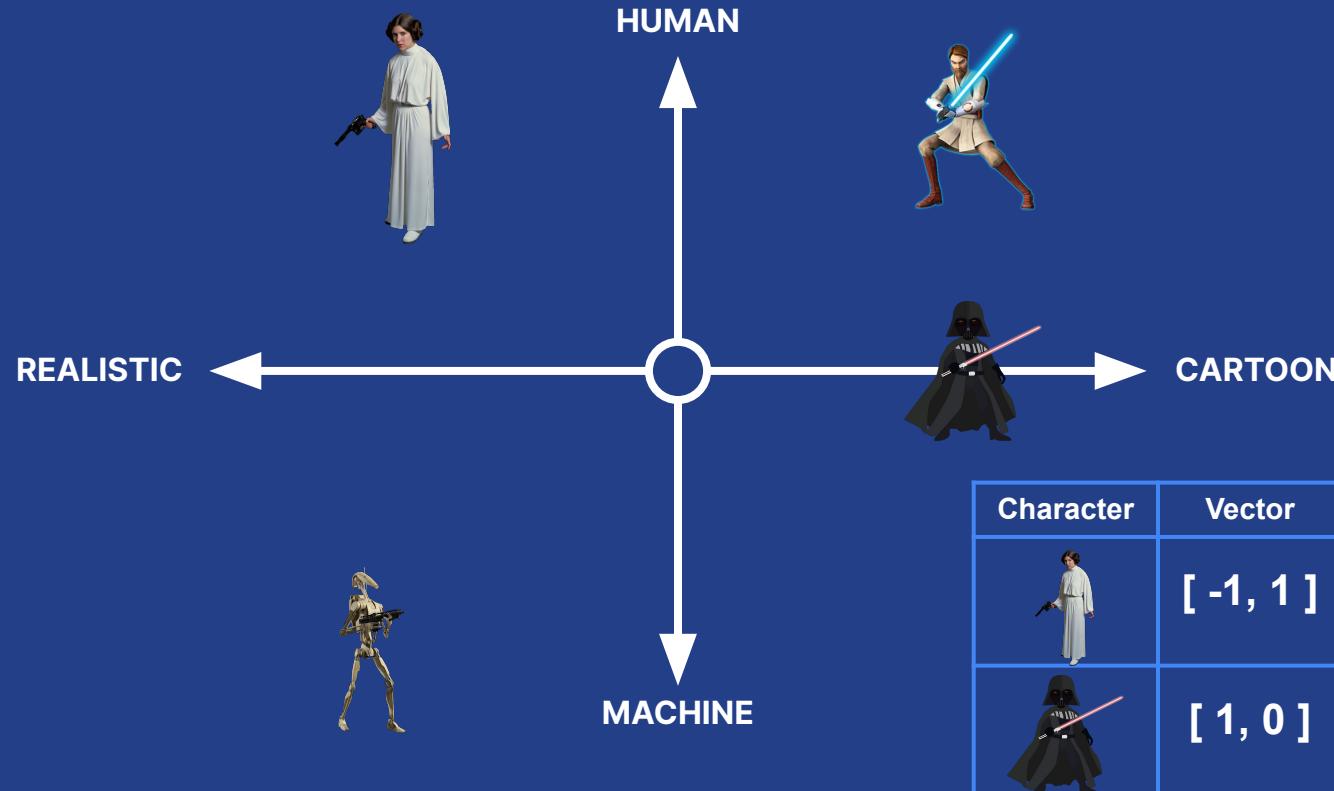
# Embeddings represent your data

Example: 1-dimensional vector

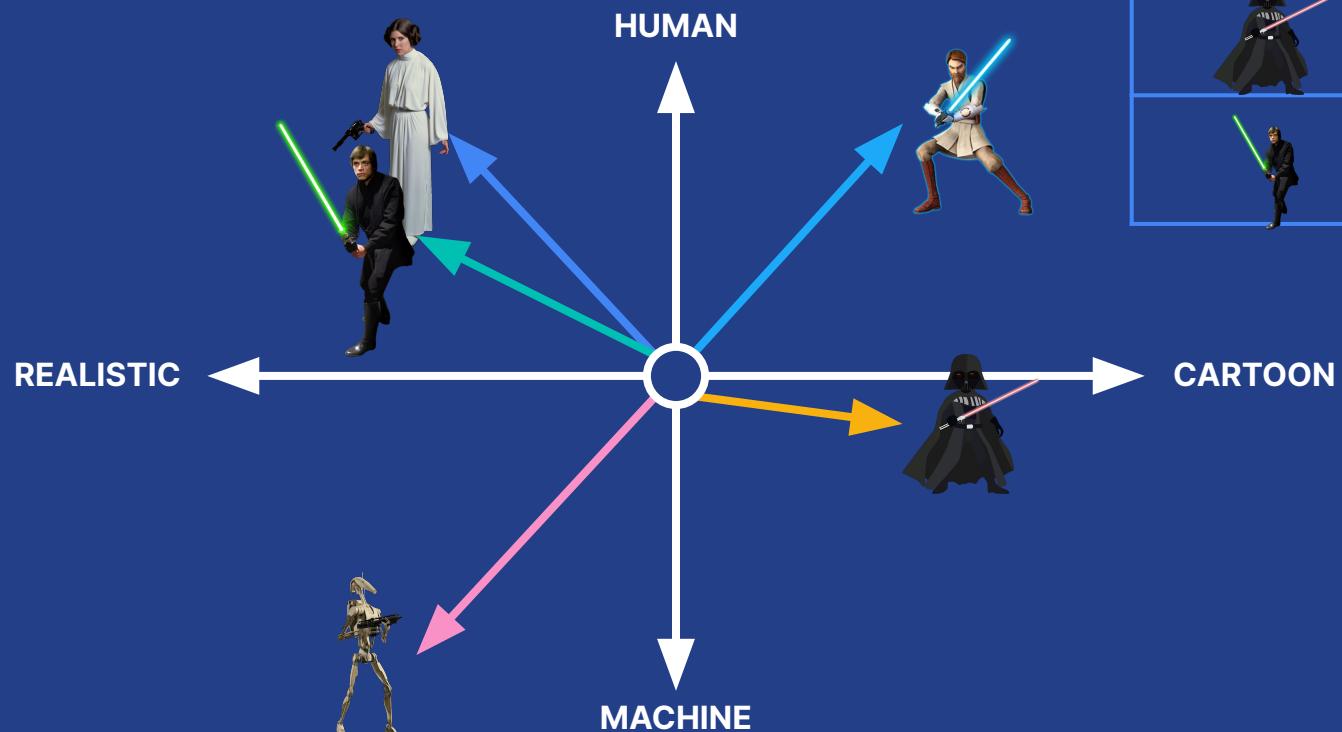


Character	Vector
	[ -1 ]
	[ 1 ]

# Multiple dimensions represent different data aspects

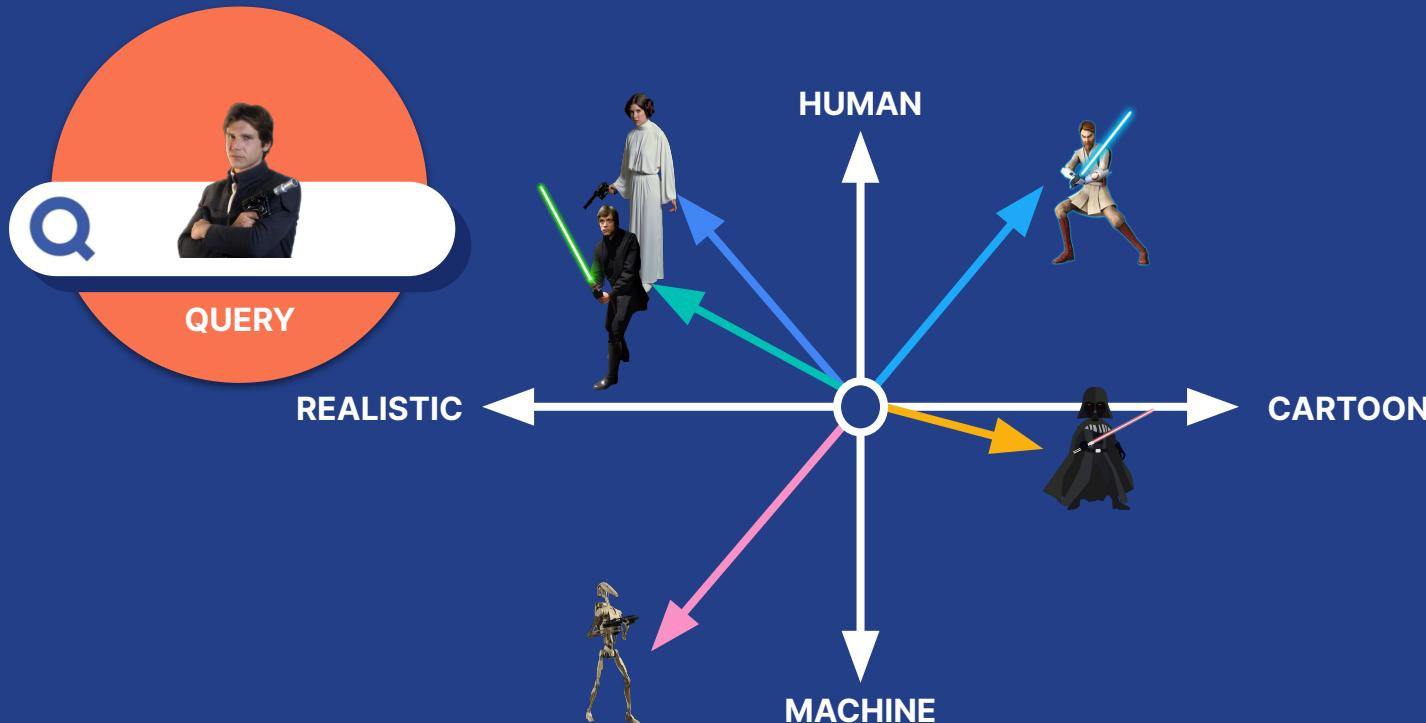


# Similar data is grouped together



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	

# Demo

<https://www.elastic.co/search-labs/finding-your-puppy-with-image-search>



**Finding your puppy with Image Search**

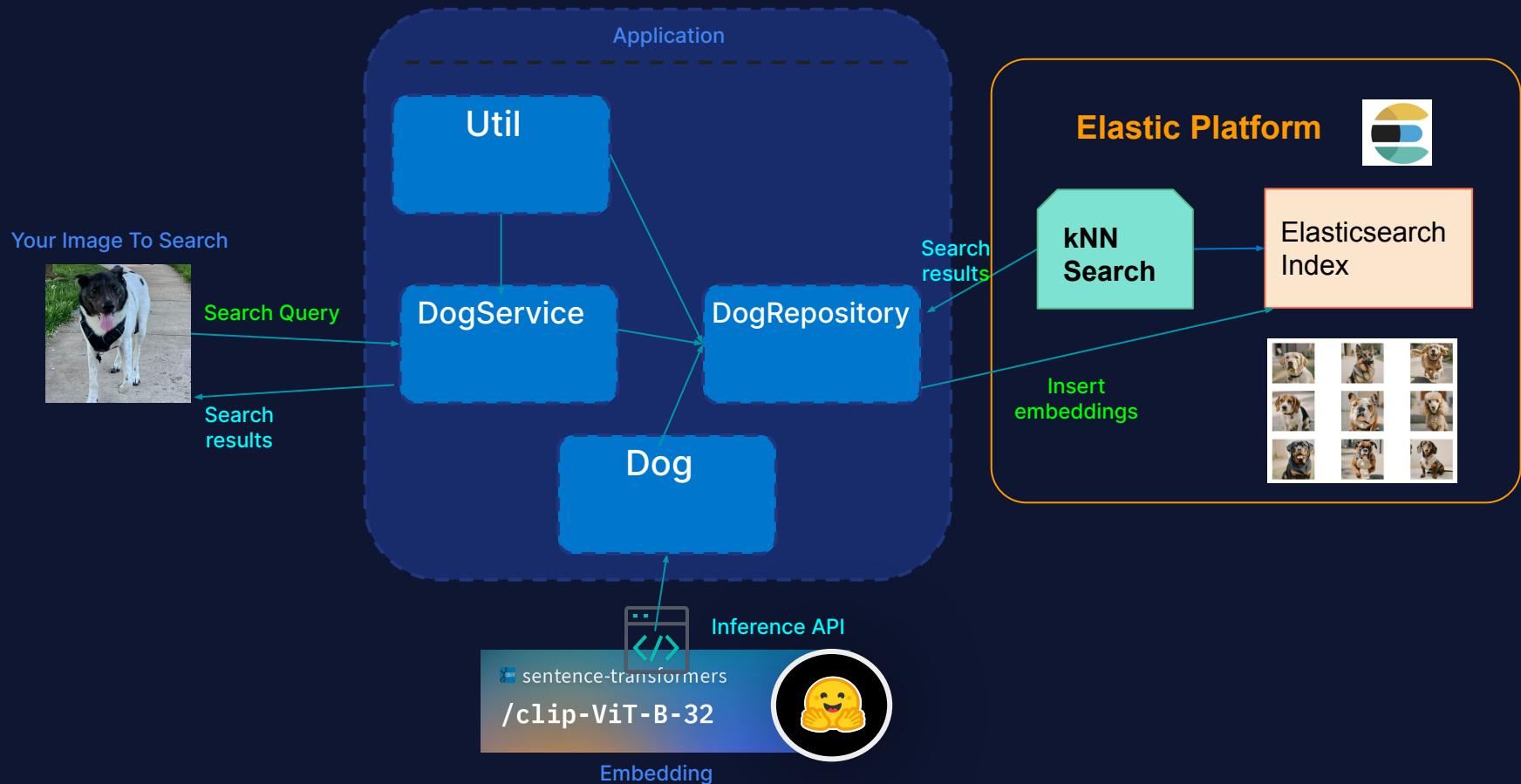
Have you ever been in a situation where you found a lost puppy on the street and didn't know if it had an owner? Learn ho...

November 7, 2023 • Alex Salgado

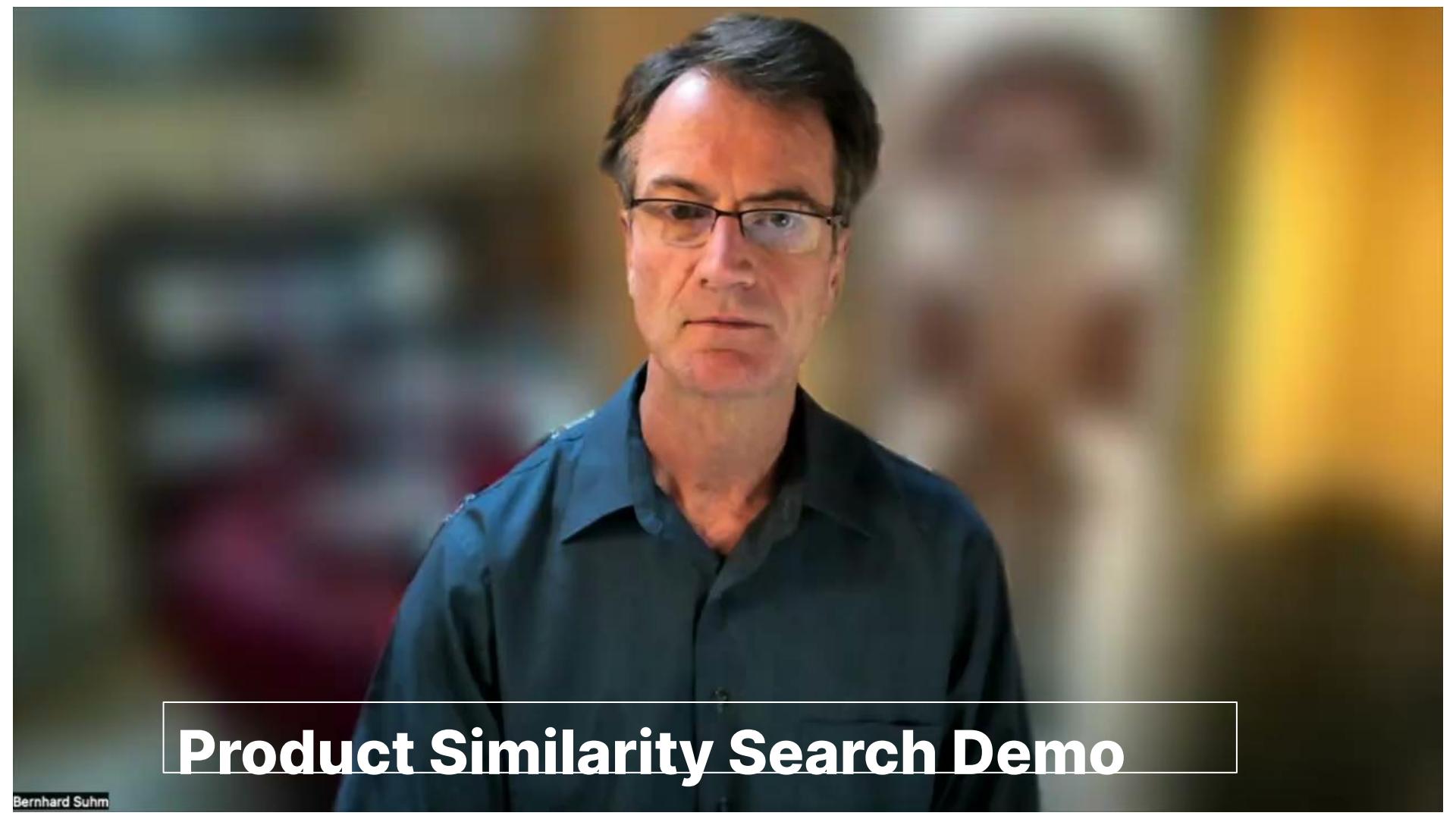


# Image Search Architecture

Generate embeddings outside Elasticsearch



# Case e-commerce



# Product Similarity Search Demo

You asked, we answered. Our best-selling classic wrap dress now comes in a cotton poplin that's wear-all-day perfect. Bonus: stripes (our favorite).

**FIT**  
• 39" from high point of shoulder

**DETAILS**  
• Cotton.  
• Linen.  
• Machine wash.  
• Import.



## Source data



Transformer  
model

POST /\_doc

2

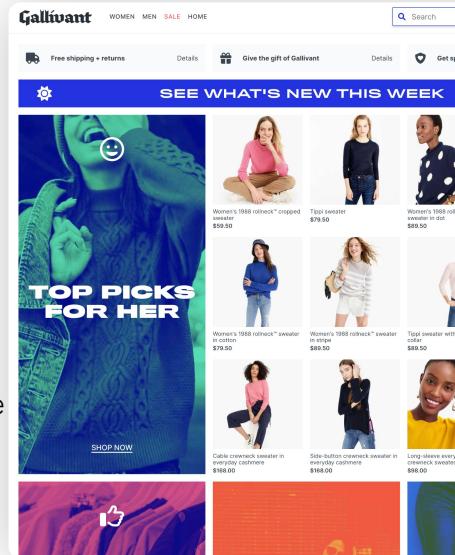
## Documents stored in Elasticsearch

```
{  
  "_id": "product-1234",  
  "product_name": "Summer Dress",  
  "description": "Our best-selling...",  
  "Price": 118,  
  "color": "blue",  
  "fabric": "cotton",  
  "desc_embedding": [0.452, 0.3242,...]
```

1

GET /\_search  
with kNN clause

3



 Search-powered  
application

# Step 1: Setting up the machine learning model

The screenshot shows the Hugging Face Model Hub interface. It features a search bar at the top and a sidebar with categories such as Tasks, Models, Datasets, Spaces, Docs, Solutions, Pricing, Login, and Sign Up. The main area displays a grid of cards for different models. One card for "BERT-MiniLM-L6" is highlighted in blue. Other visible models include "distilbert-base-uncased", "roberta-base", "T5-base", "distilbert-base-uncased-distilled-tiny", "bert-base-chinese", "distilbert-base-chinese", "sentence-transformers/all-mnli-MiniLM-L6-v2", and "apertium-MiniLM-L6-v2". Each card includes a brief description, a URL, and a "View Model" button.

```
$ eland_import_hub_model  
--url https://cluster_URL  
--hub-model-id BERT-MiniLM-L6  
--task-type text_embedding  
--start
```

The screenshot shows the eland UI's "Machine Learning" section under "Model Management". On the left, there's a sidebar with options like Overview, Notifications, Anomaly Detection, Job, Single Metric Viewer, Settings, Data Frame Analytics, Jobs, Results Explorer, Analytics Map, Model Management, Trainet Models, Nodes, Data Visualizer, File, Data View, AIOps Lab, Explainable AI, and Log Pattern Analysis. The main area is titled "Trained Models" and shows a list of trained models. One model, "sentence-transformers/all-minilm-l6-v2", is highlighted in blue. A "Test trained model" panel on the right contains a text input field with the sentence "I'm wearing a cotton polo that's wear-all-day perfect. Bonus: supercute stripes (our favorite!)" and a "Test" button.



PyTorch

Select the appropriate model



Load the model to the cluster



Manage models

# Step 2: Data ingestion and embedding generation



Source data



```
{  
  "_id": "product-1234",  
  "product_name": "Summer Dress",  
  "description": "Our best-selling...",  
  "Price": 118,  
  "color": "blue",  
  "fabric": "cotton",  
} "desc_embedding": [0.452, 0.3242, ...]  
}
```



ML Inference pipelines Add inference pipeline

Inference pipelines will be run as processors from the Enterprise Search Ingest Pipeline

ml-inference-embedding-generation	Actions
• Deployed pytorch text_embedding	⋮

ml-inference-emotional-analysis	Actions
• Deployed pytorch text_classification	⋮

Learn more about deploying ML models in Elastic [🔗](#)



# Step 3: Issuing a vector query

a

Query is submitted to the search-powered application

 summer clothes 



b

Query embedding is generated

```
POST /_ml/trained_models/my-model/_infer
{
  "docs": [
    {
      "description": "summer clothes"
    }
  ]
}
```



Transformer model

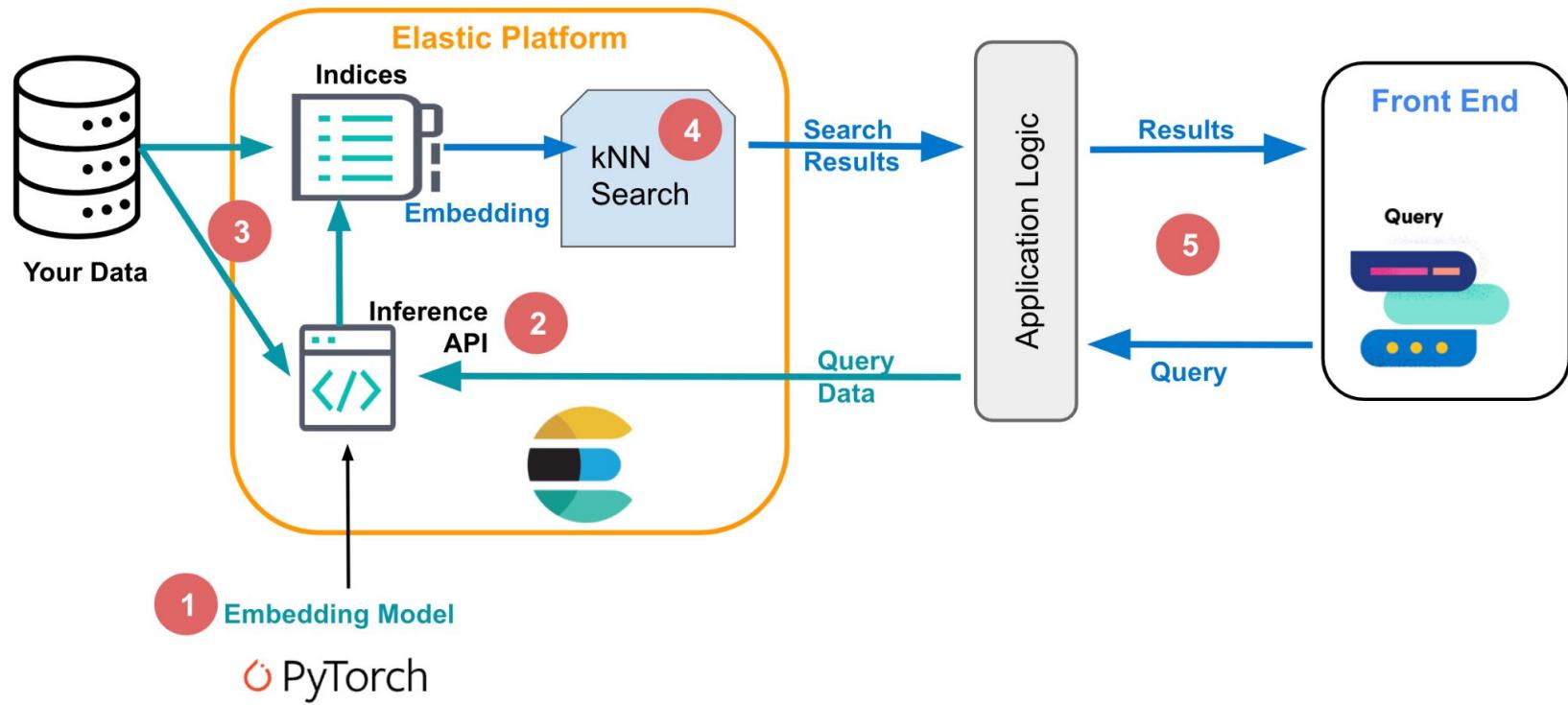
 PyTorch

c

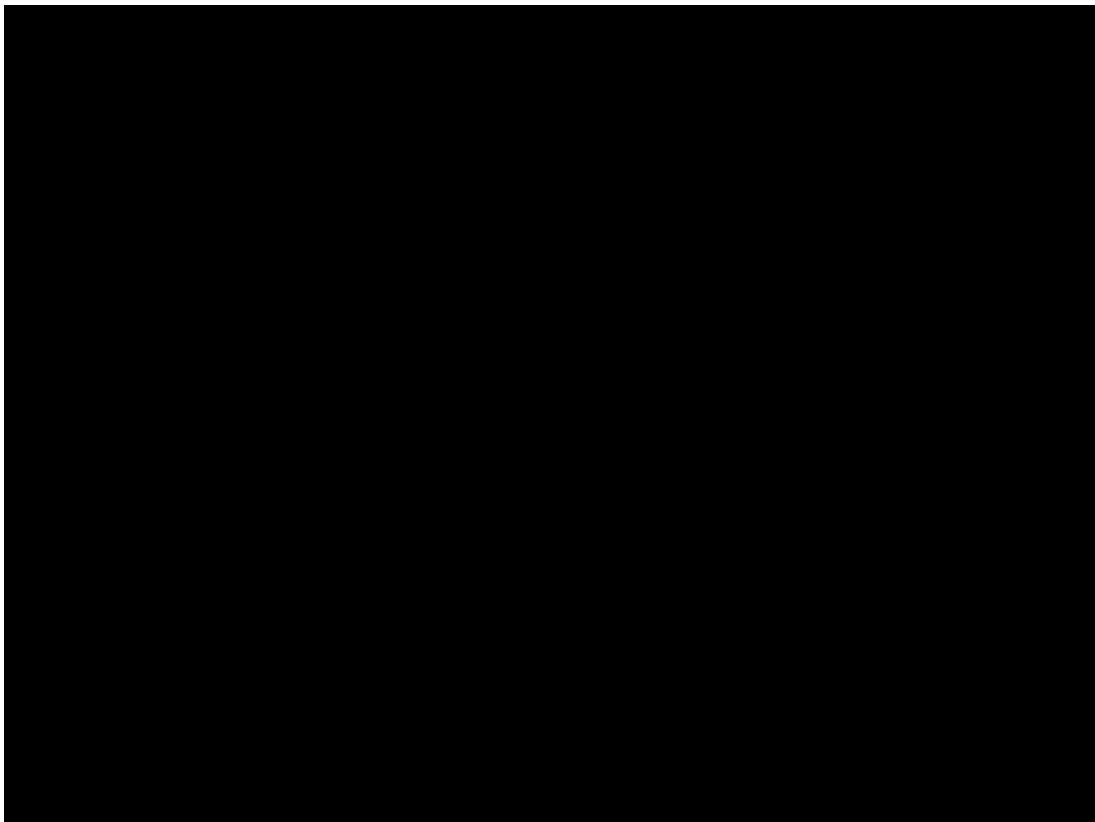
Issue query using the `_search` endpoint, with a kNN clause, using the previously generated embedding

```
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
  }
}
```

# Architecture of Vector Search



# How this works in our Ecommerce Demo



# Obrigado



@alexsalgadoprof



salgado



@alexsalgadoprof



/in/alex-salgado/

