elastic

Image Search

Alex Salgado
Developer Advocate @ Elastic

@alexsalgadoprof
salgado
@alexsalgadoprof
/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

aws certified Solutions Architect Associate

aws certified Cloud Practitioner

LEVEL 2 CERTIFIED Educator Google for Education

Microsoft CERTIFIED

Alex Salgado

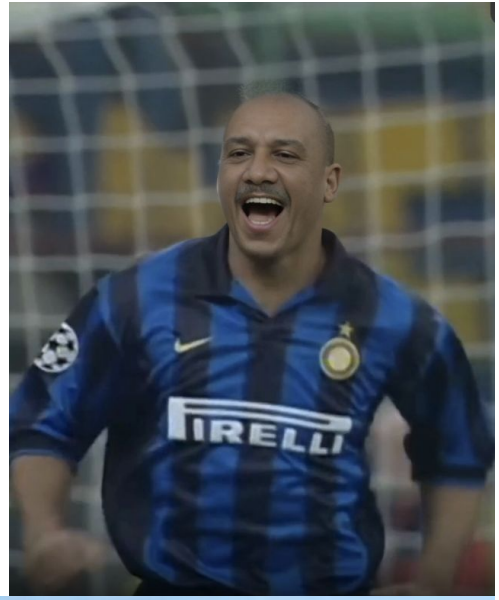
Senior Developer Advocate LATAM

@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

elastic

Three solutions powered by one stack

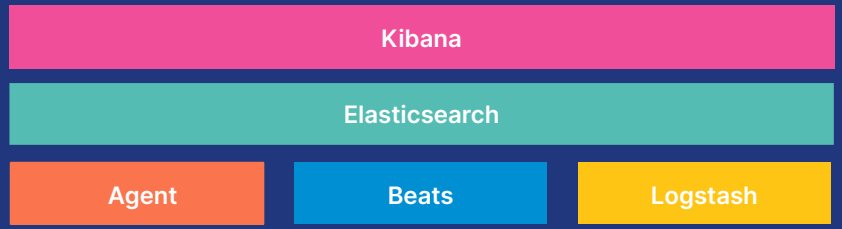
3 solutions

Enterprise Search

Observability

Security

Powered by the Elastic Stack



Deployed anywhere

Elastic Cloud

Elastic Cloud Enterprise

Elastic Cloud on Kubernetes

Saas

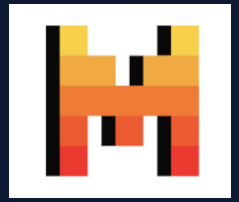
Orchestration

elastic

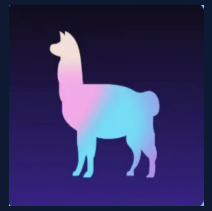




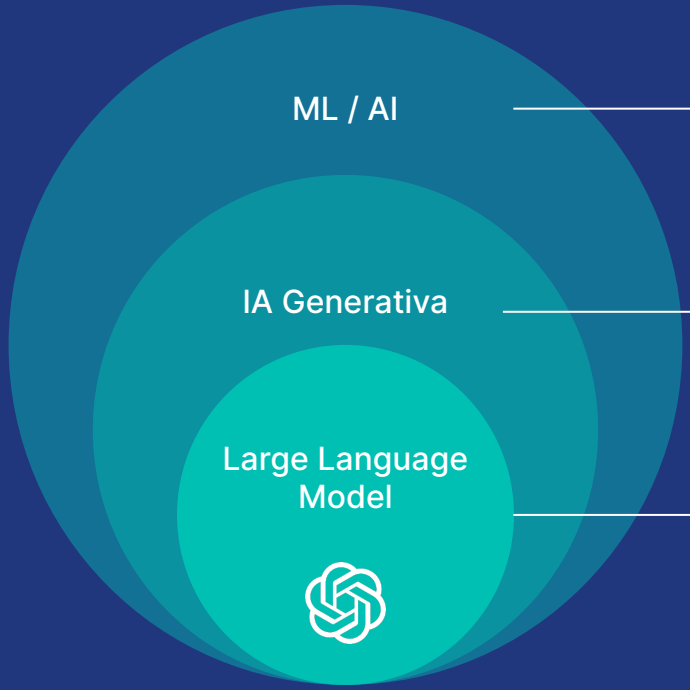









Conceitos básicos de ML, IA Generativa e LLMs

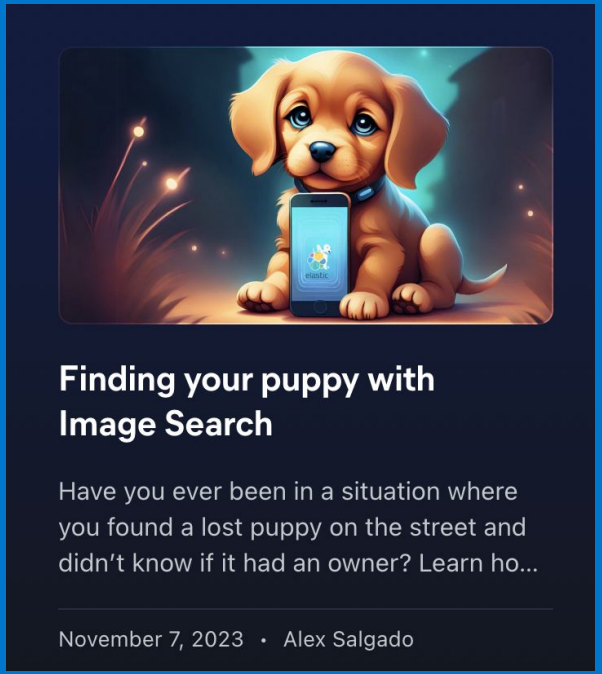


	O que?	Casos de uso
ML / AI	Algoritmos programados para aprender o comportamento dos dados e fazer previsões	Detecção de anomalias, forecasting, reconhecimento de imagem, PLN
IA Generativa	Algoritmos programados para criar novos dados	Chatbots, geradores de texto, imagem e música
Large Language Model	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

elastic

Blog referência

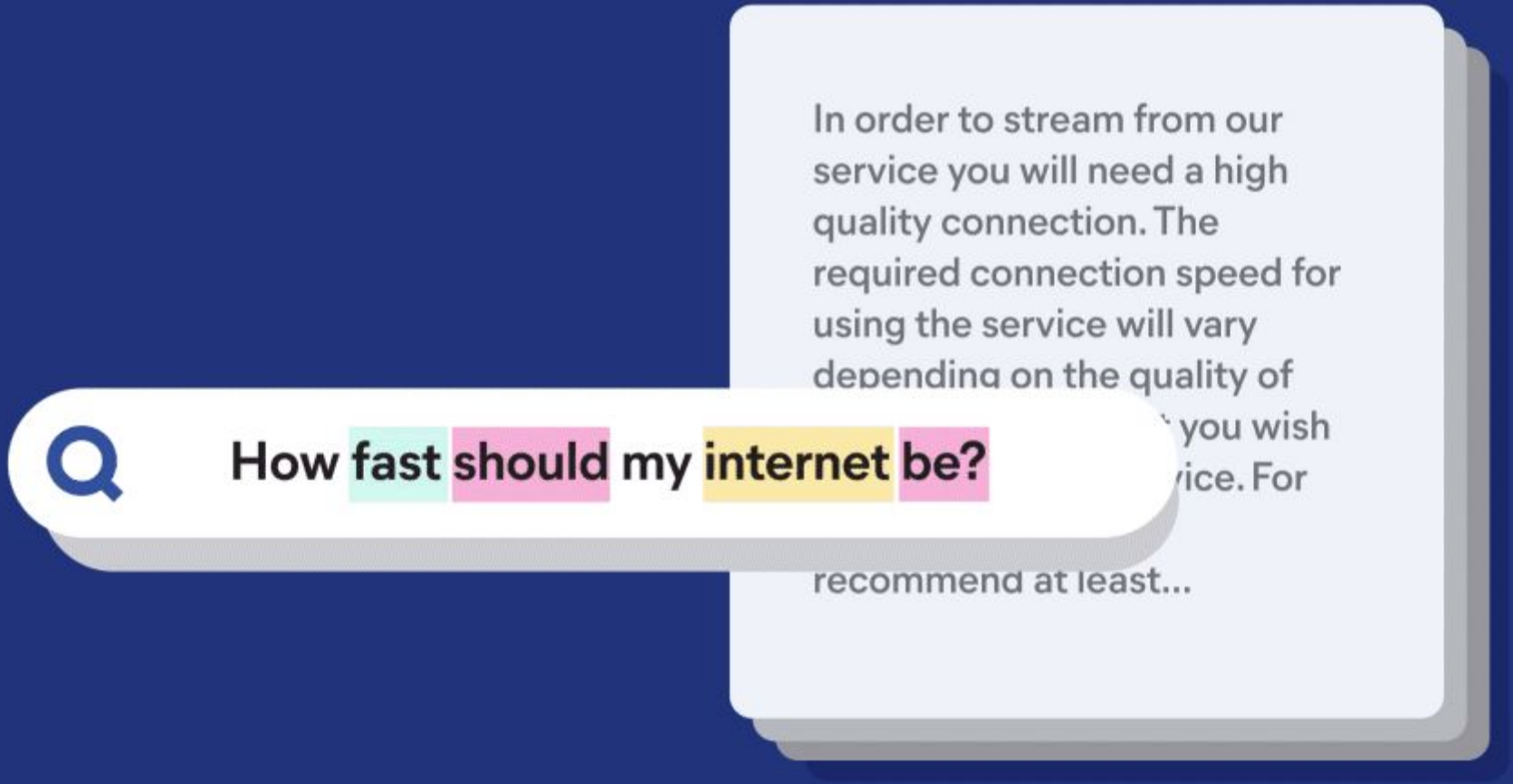
https://www.elastic.co/search-labs/finding-your-puppy-with-image-search



elastic

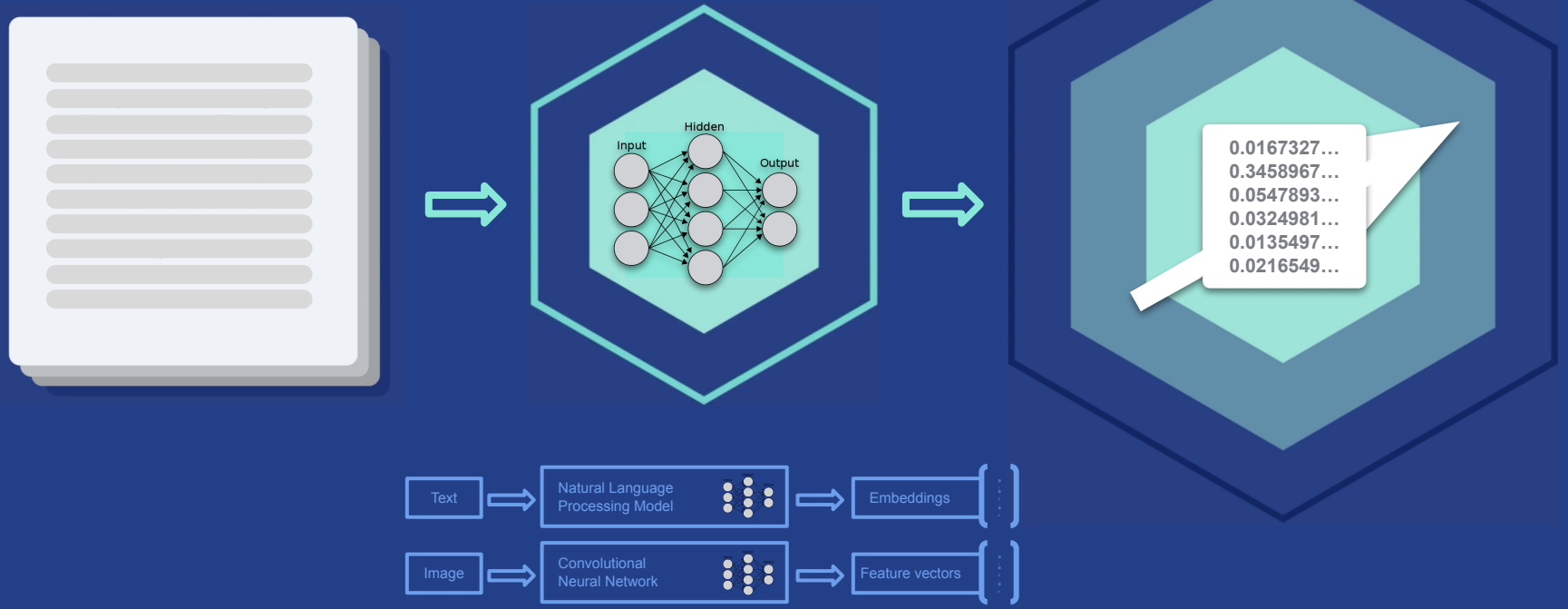
Elasticsearch:
You Know, for Search

elastic


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...

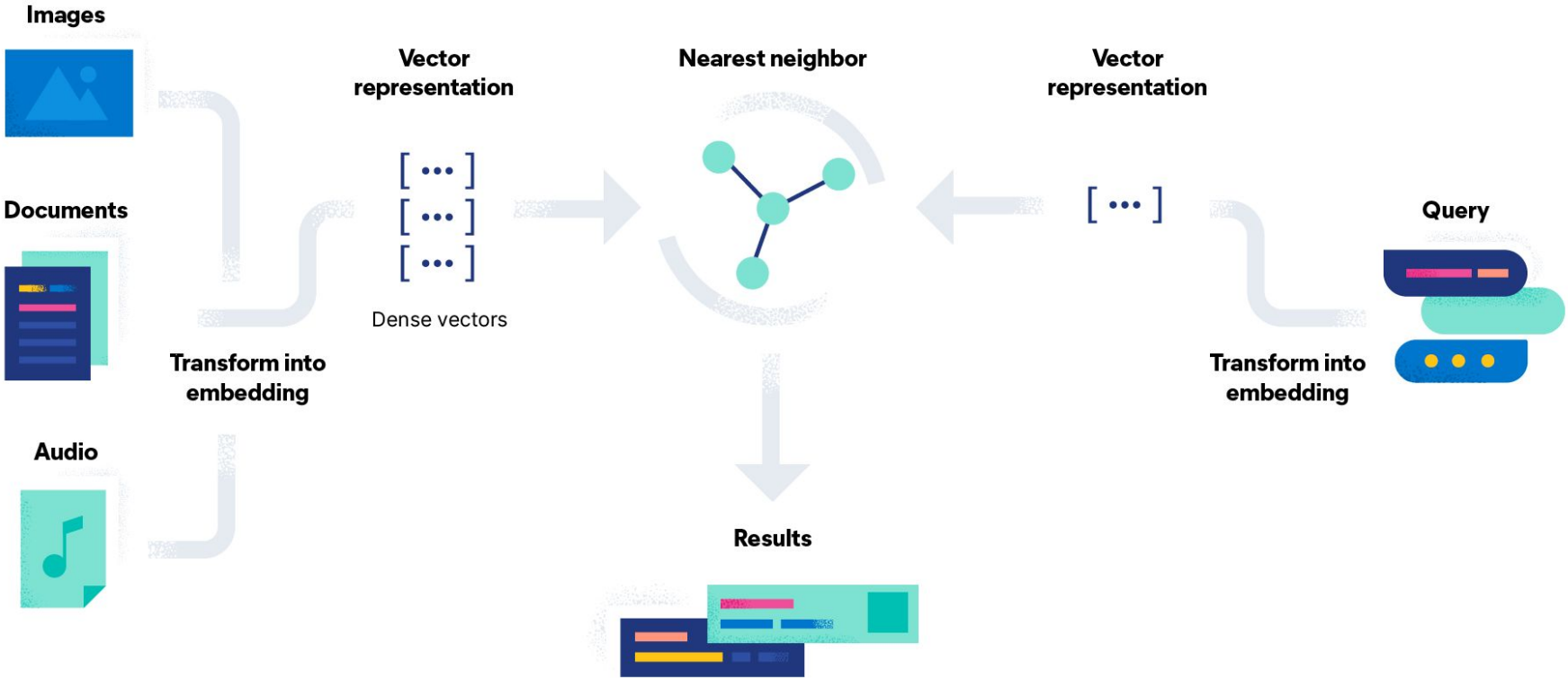
O que é similaridade de vetores?

Converta dados em representações vetoriais onde as distâncias representam similaridade.



elastic

Queries are also vectorized



What is a Vector?

elastic

Embeddings represent your data

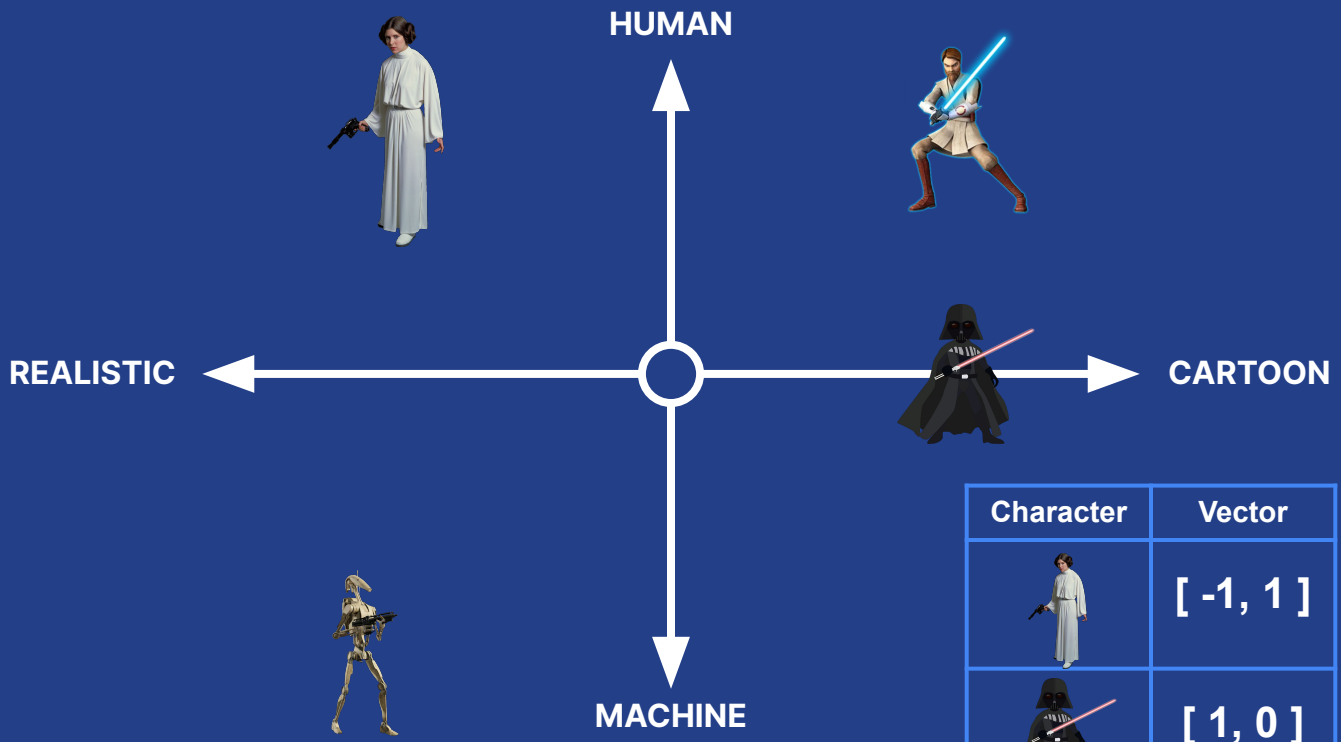
Example: 1-dimensional vector

REALISTIC ◀——●——▶ CARTOON

Character	Vector
	[-1]
	[1]

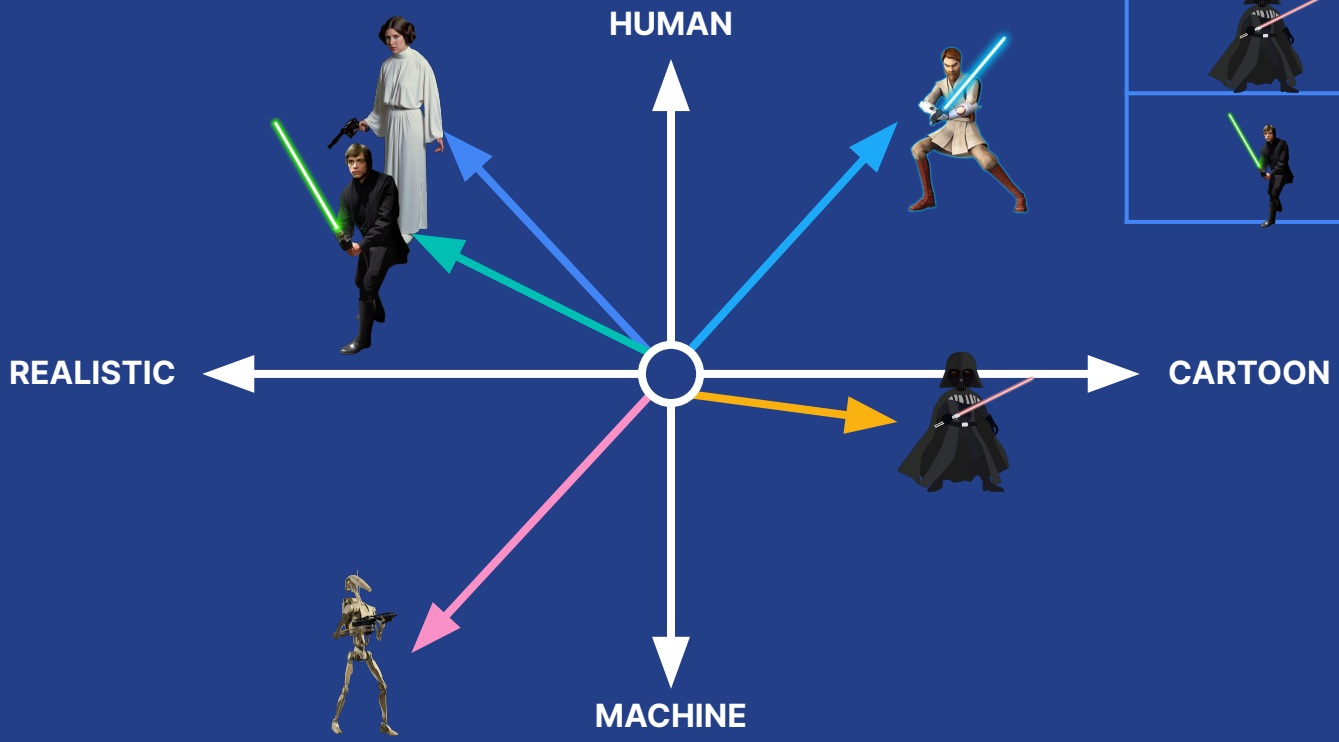
elastic

Multiple dimensions represent different data aspects



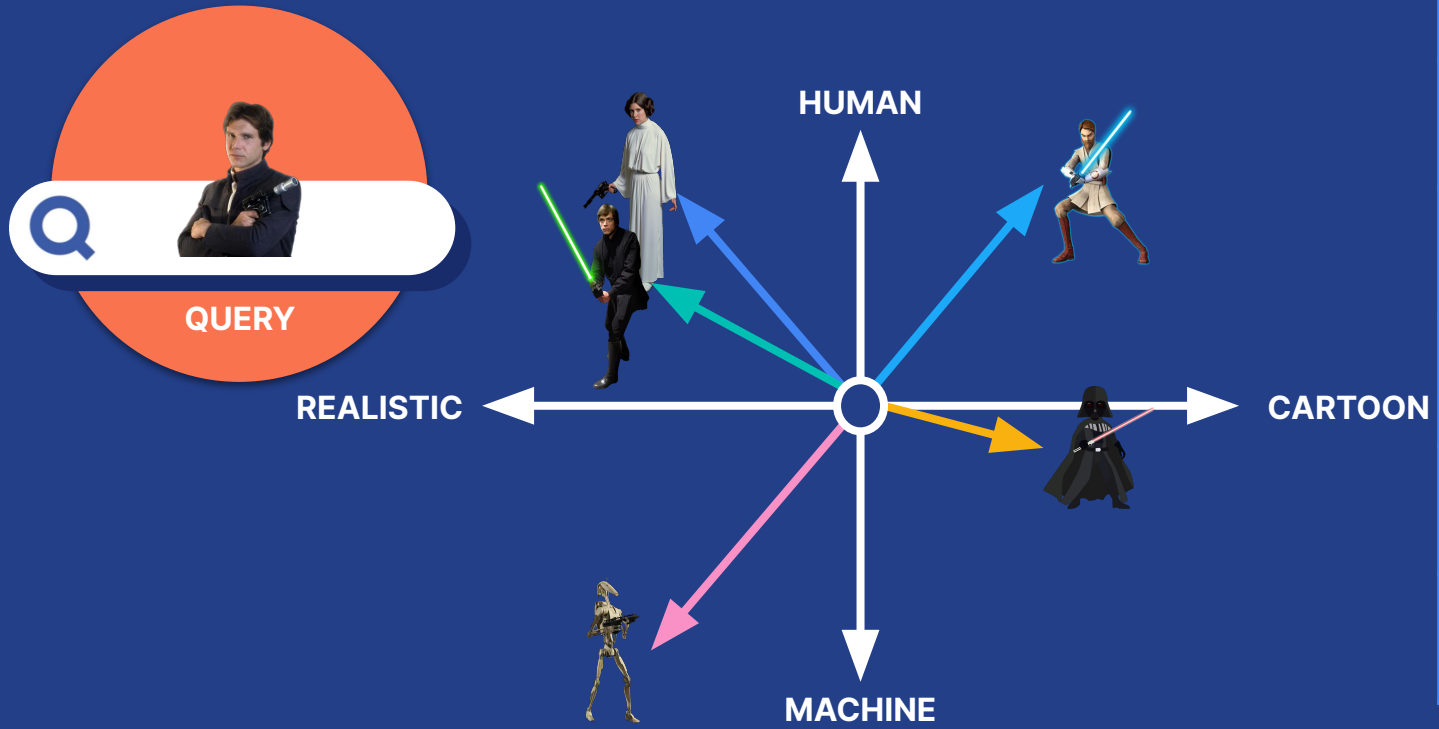
Character	Vector
	[-1, 1]
	[1, 0]

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	

elastic

Demo

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

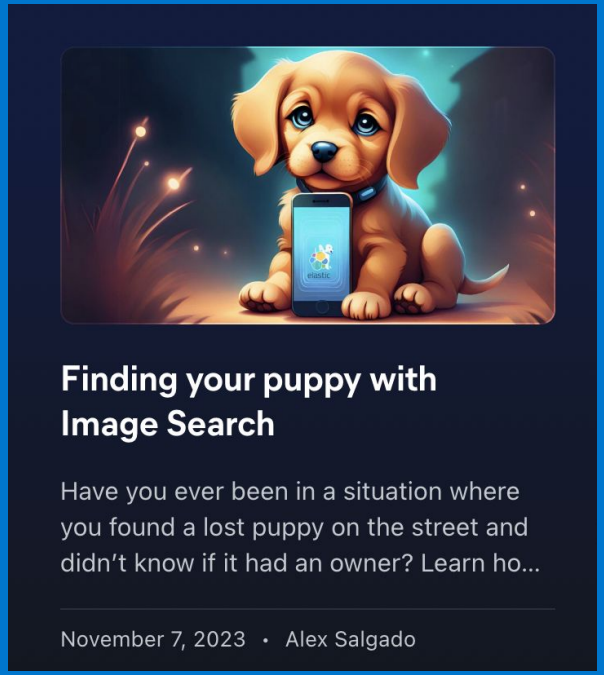
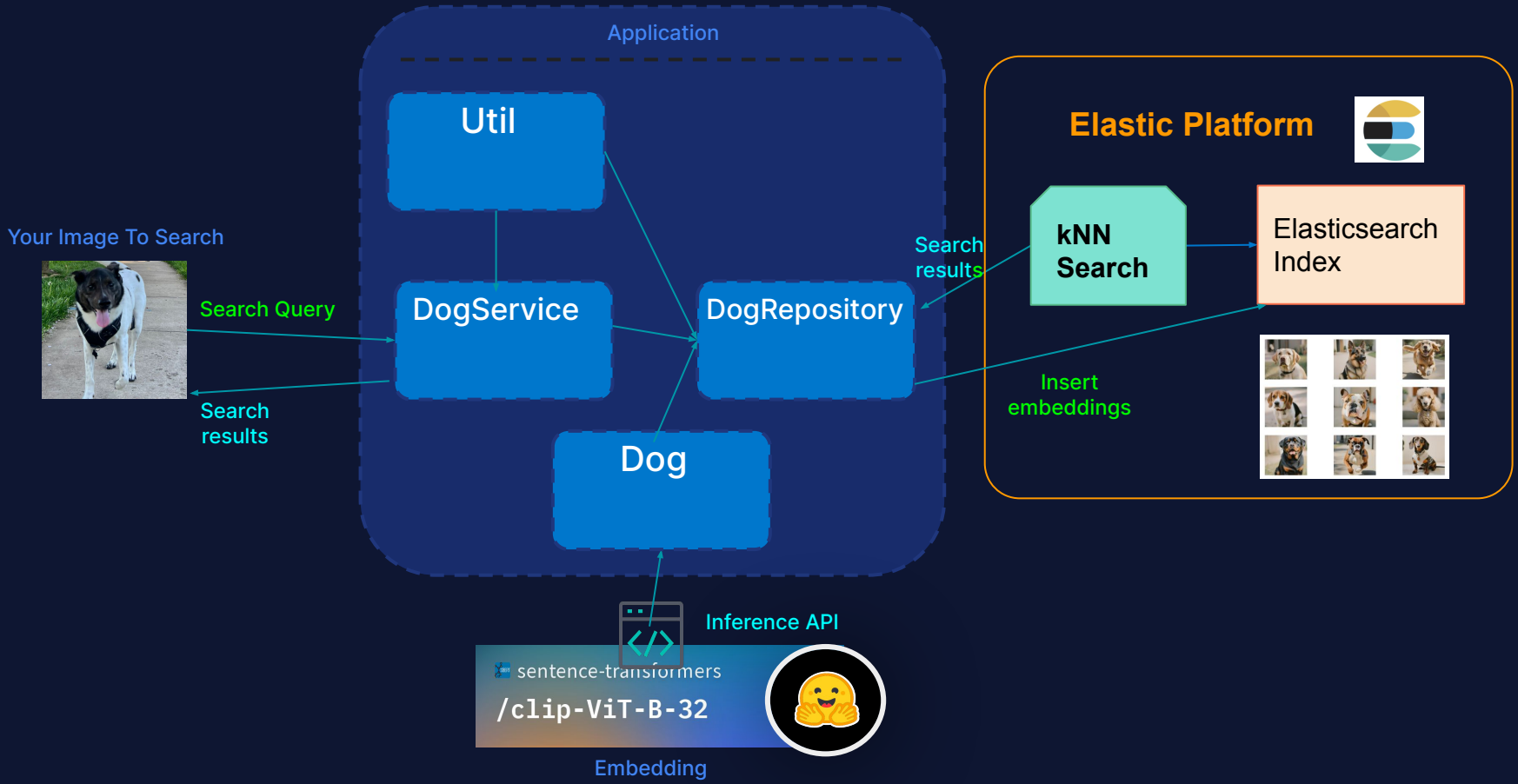
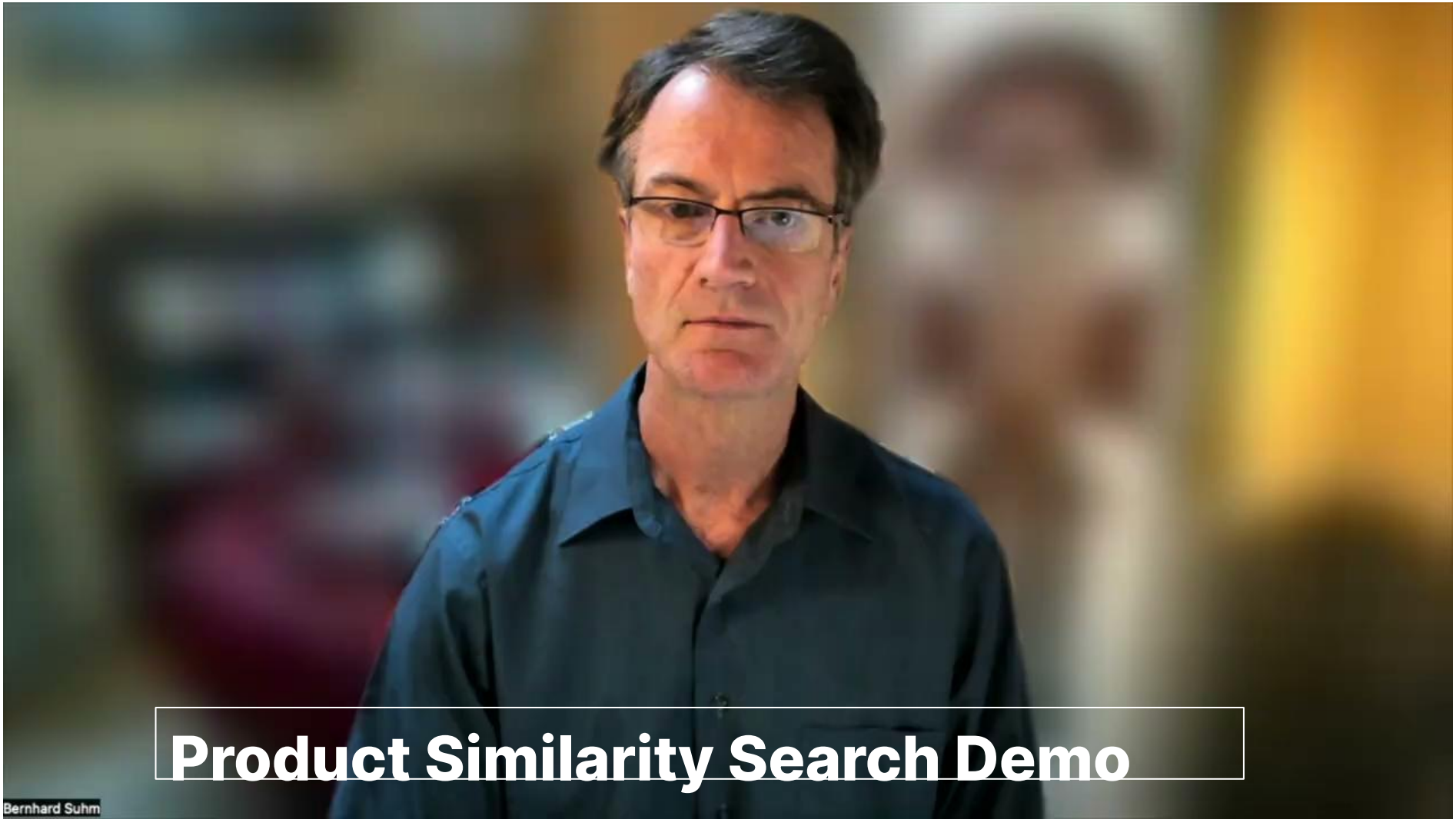


Image Search Architecture

Generate embeddings outside Elasticsearch



Case e-commerce


Product Similarity Search Demo
Bernhard Suhm

cloud

Source data

2

POST /_doc

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,…]
}
```

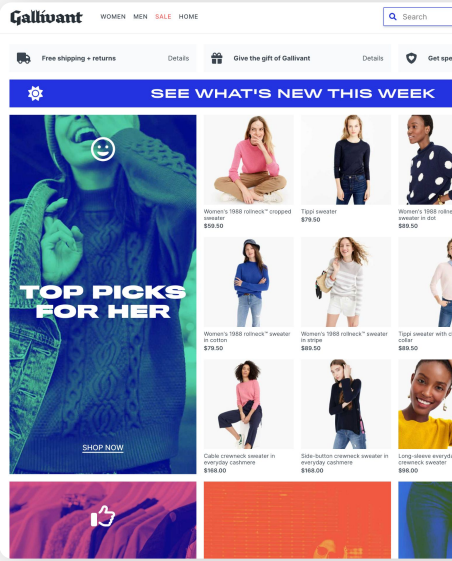
1

Transformer model

3

GET /_search

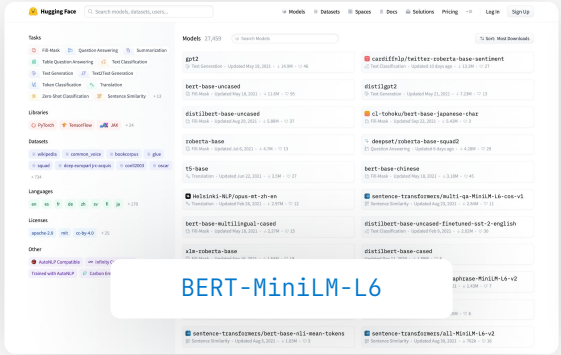
with kNN clause



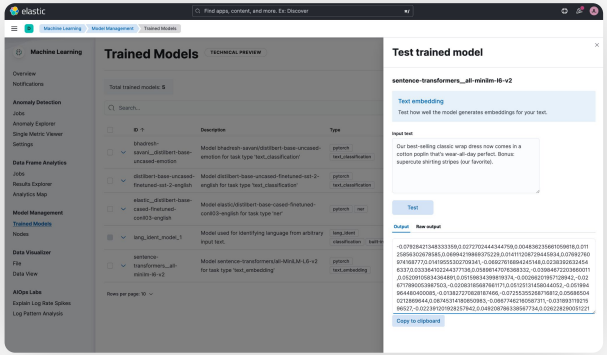
Search-powered application

elastic

Step 1: Setting up the machine learning model




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



PyTorch → eland →

Select the appropriate model

Load the model to the cluster

Manage models

elastic

Step 2: Data ingestion and embedding generation

Source data

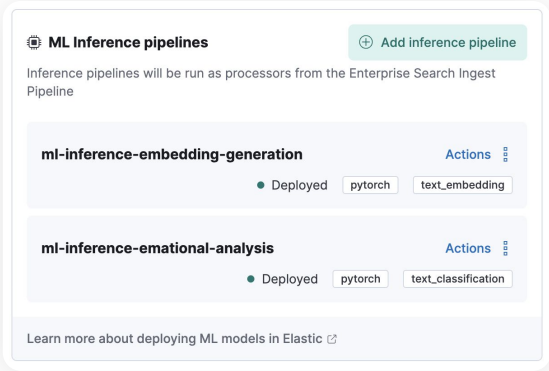
POST /_doc

Standard field indexing for non-vector types

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

POST /_doc

Encoding via Inference Processor



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

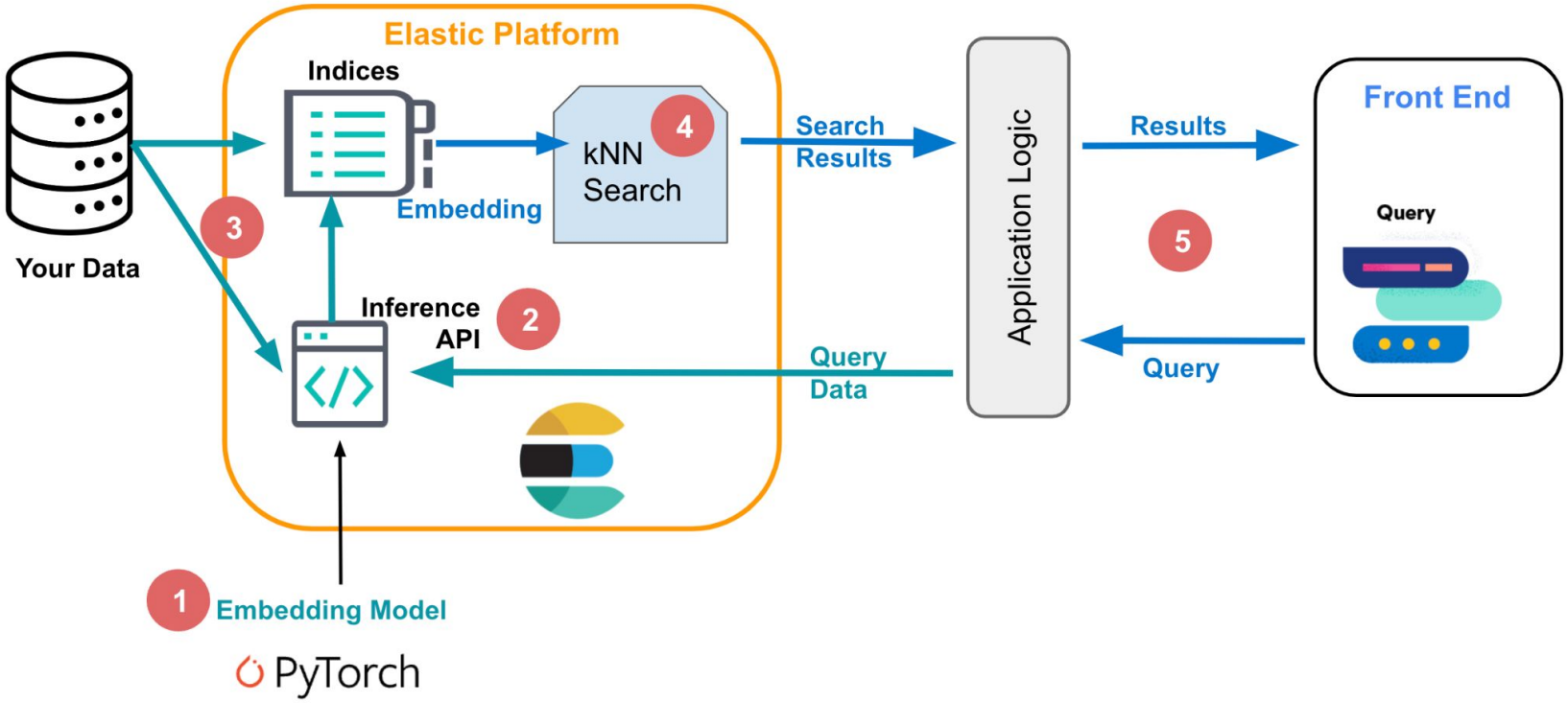
PyTorch

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

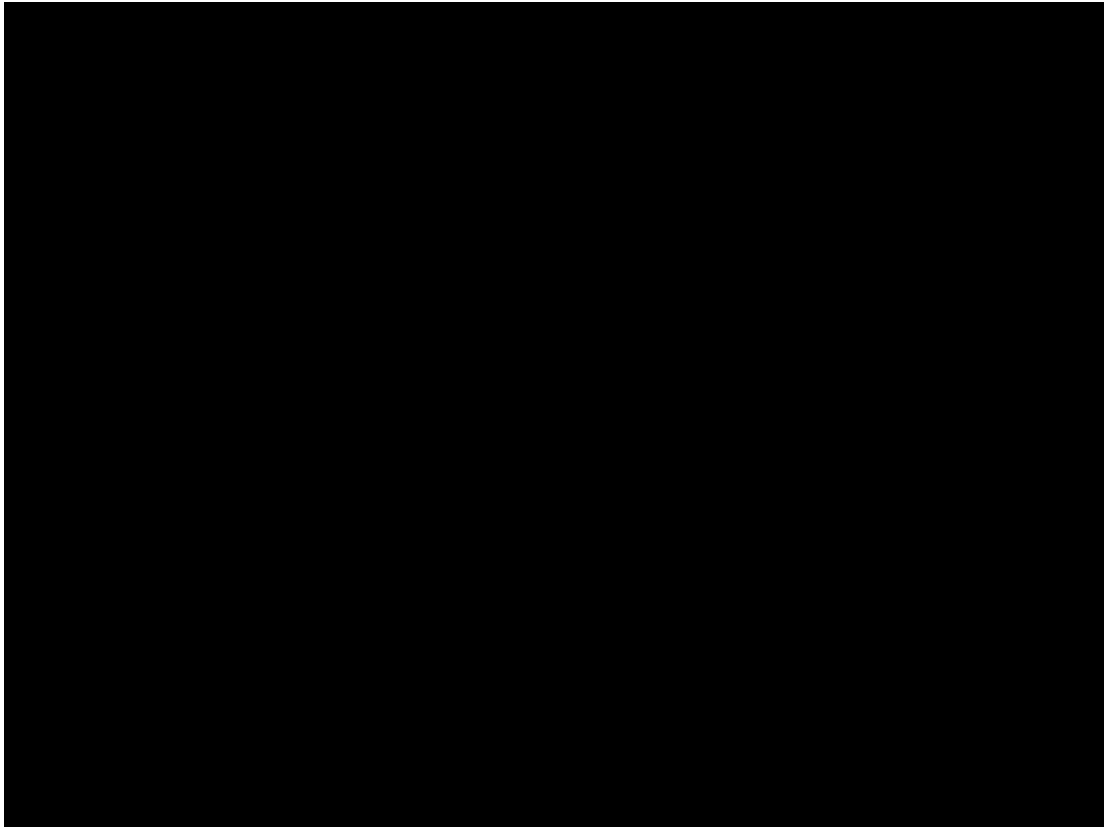
```
GET product-catalog/_search

{
  "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
}
```

elastic

Architecture of Vector Search

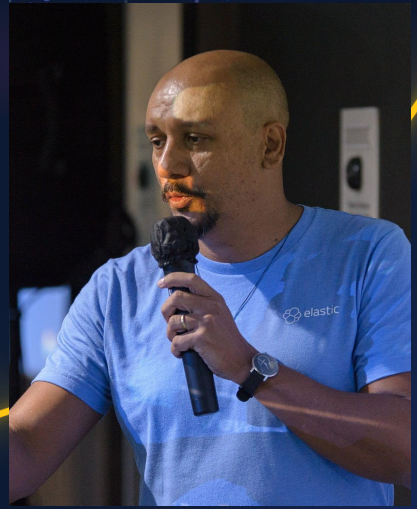


How this works in our Ecommerce Demo



elastic

Obrigado



@alexsalgadoprof

salgado

@alexsalgadoprof

/in/alex-salgado/