



IA e Busca: Desbloqueie o Poder do Gemini para Question Answering com Elasticsearch e Langchain

Alex Salgado

Senior Developer Advocate, Elastic

@alexsgadoprof - redes sociais



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

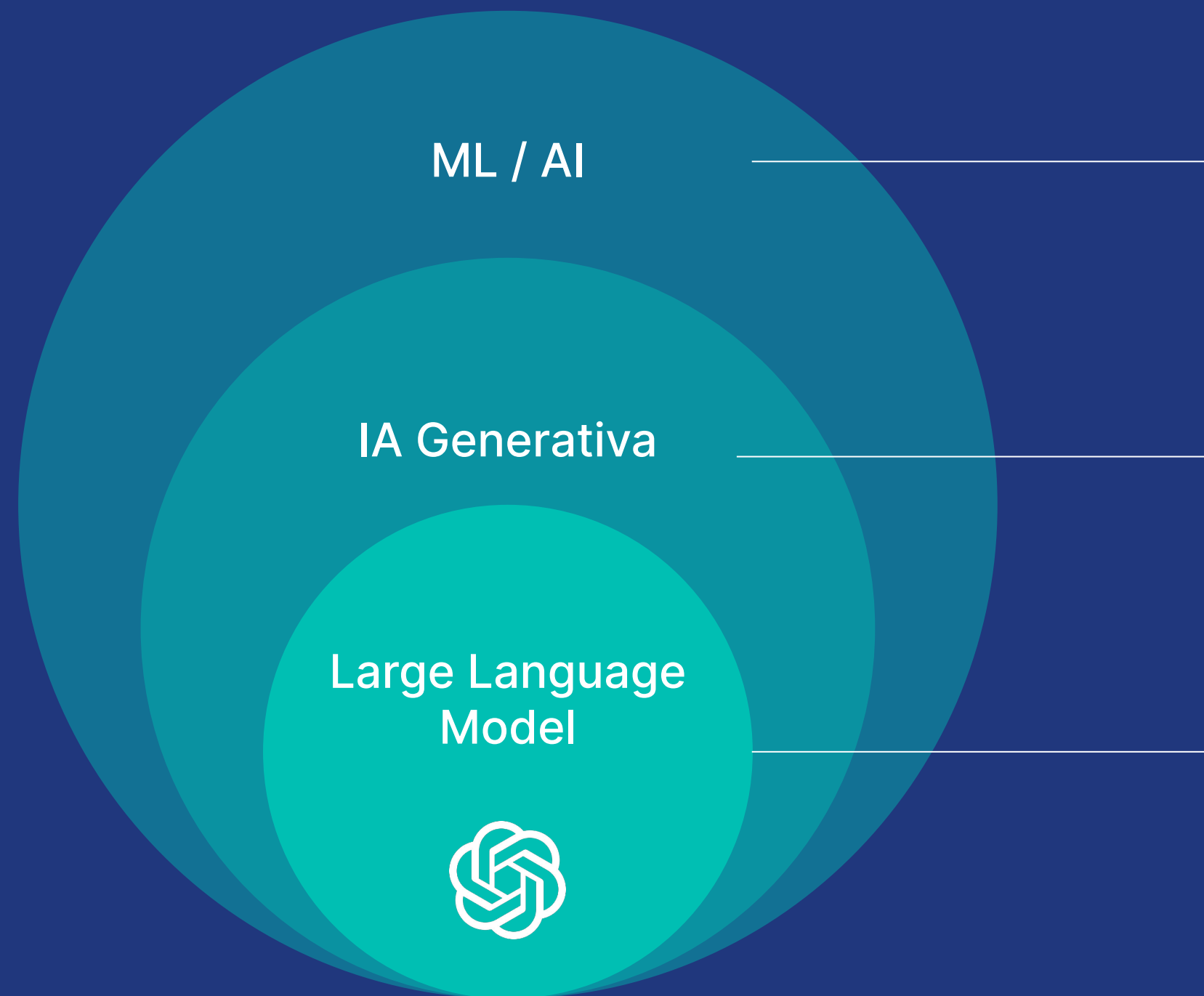
Orchestration

Preocupações em torno da IA Generativa.

80%

Dados mundiais são não-estruturados

Conceitos básicos de ML, IA Generativa e LLMs



O que?

Algoritmos programados para aprender o comportamento dos dados e fazer previsões

Algoritmos programados para criar novos dados

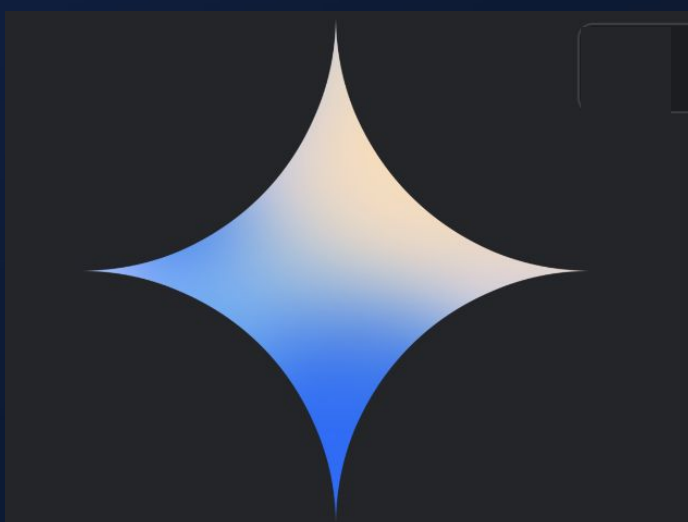
Algoritmos (Deep Learning) treinados com grandes volumes de dados e programados para criar novos dados

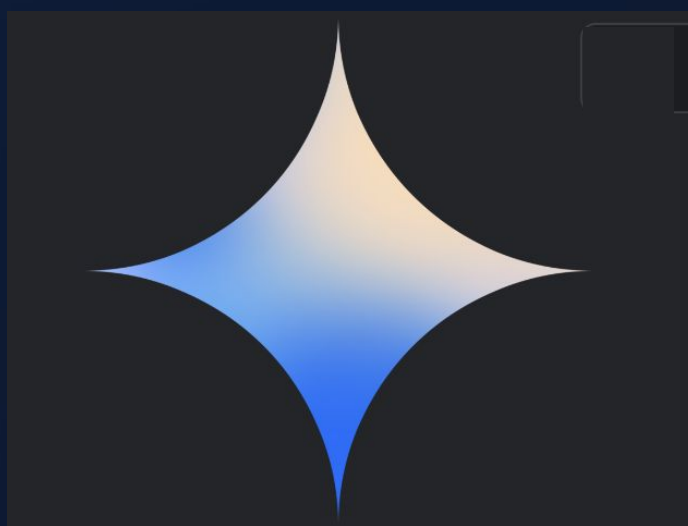
Casos de uso

Detecção de anomalias, forecasting, reconhecimento de imagem, PLN

Chatbots, geradores de texto, imagem e música

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





O que é Gemini?

Gemini é uma família de LLMs desenvolvida pelo Google DeepMind, sucessora do LaMDA e PaLM 2. É um modelo de IA poderoso, capaz de entender e processar informações em vários formatos, incluindo texto, código, imagens e áudio

Elasticsearch e Langchain

Langchain é um framework para desenvolvimento de aplicações com LLMs (modelos de linguagem grandes). Ele facilita a construção e o uso de LLMs para diversas tarefas, tornando o processo mais eficiente e acessível.

Com Langchain, você pode criar aplicações que classificam texto, extraem informações, geram conteúdo e muito mais.

What is a Vector?

Embeddings represent your data



Example: 1-dimensional vector



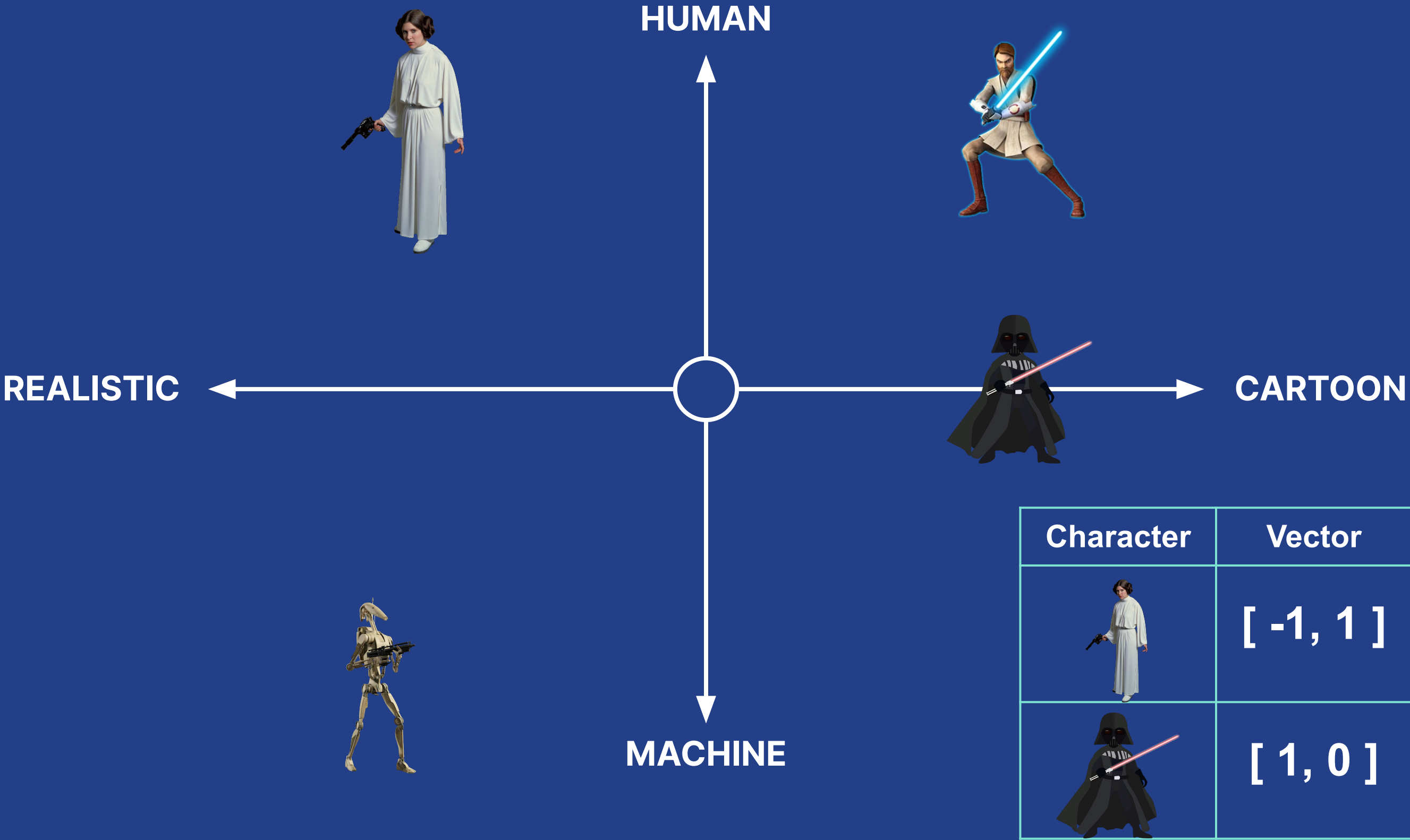
REALISTIC



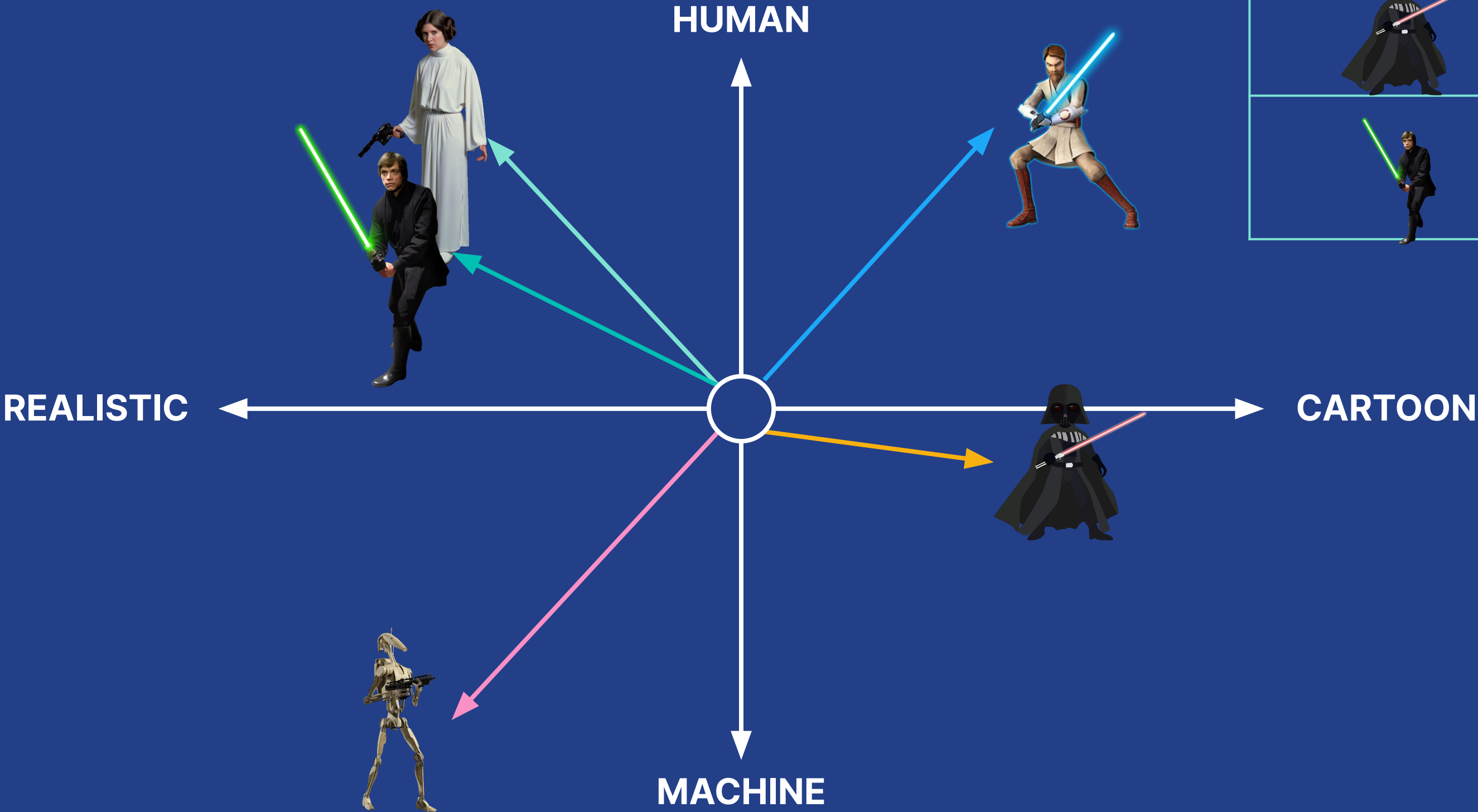
CARTOON




Character	Vector
	$[-1]$
	$[1]$

Multiple dimensions represent different data aspects

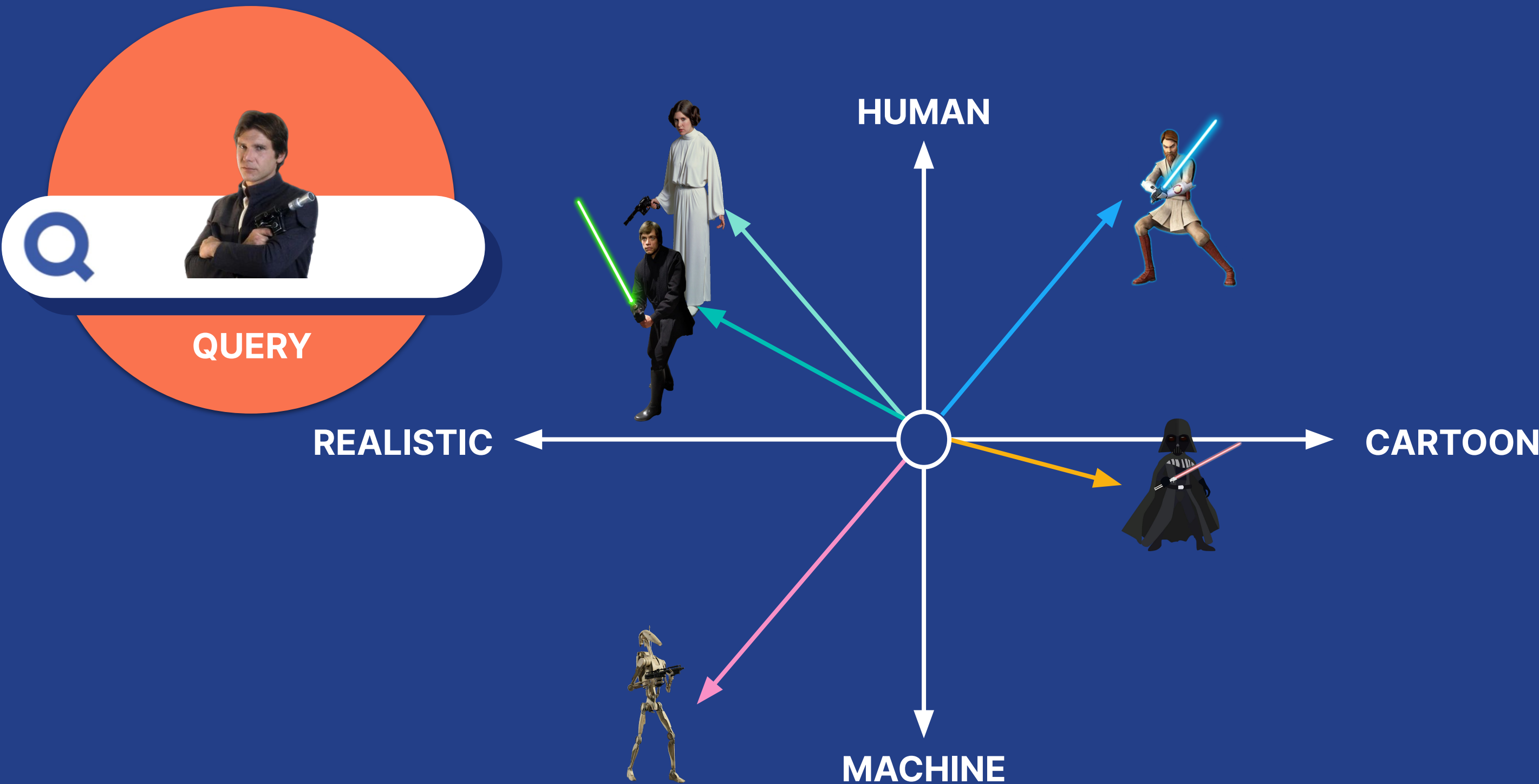







Similar data is grouped together



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

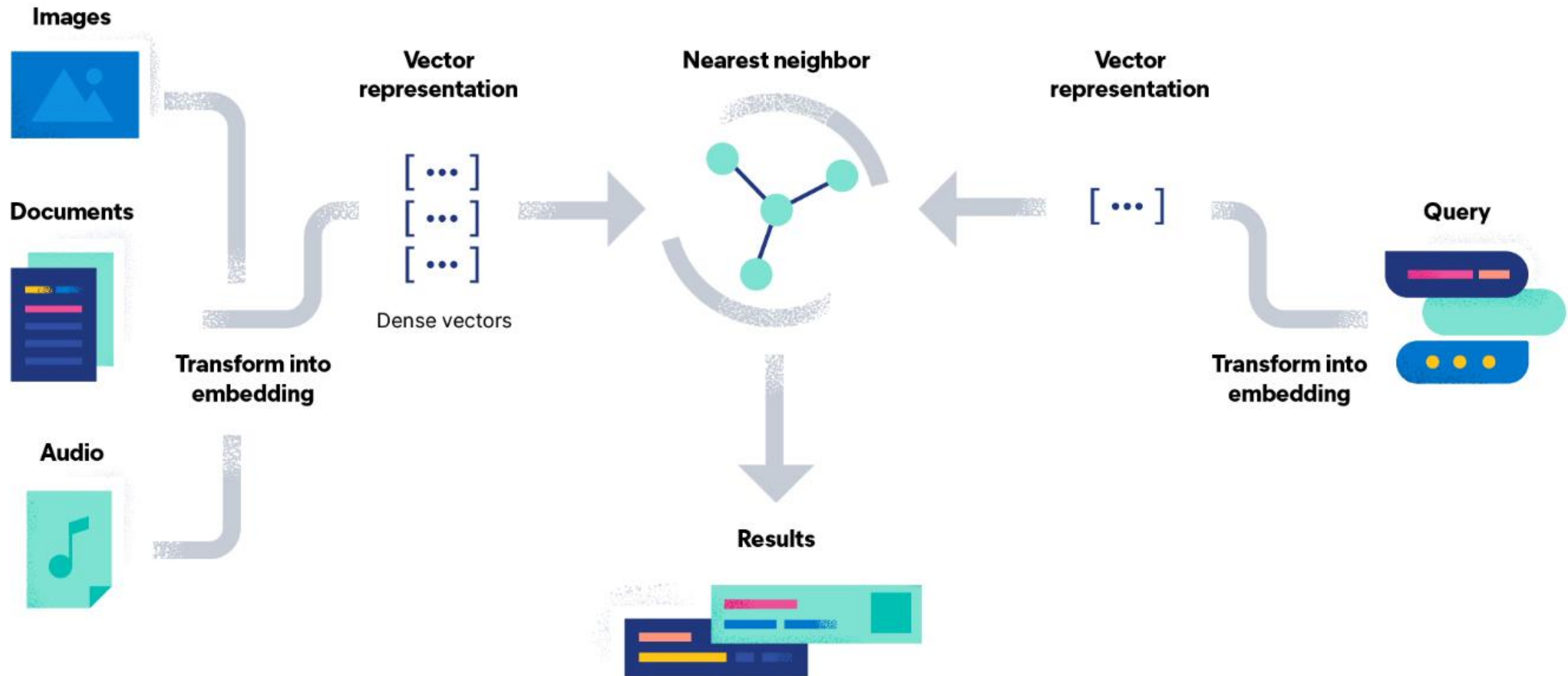
Vector search ranks objects by similarity (relevance) to the query



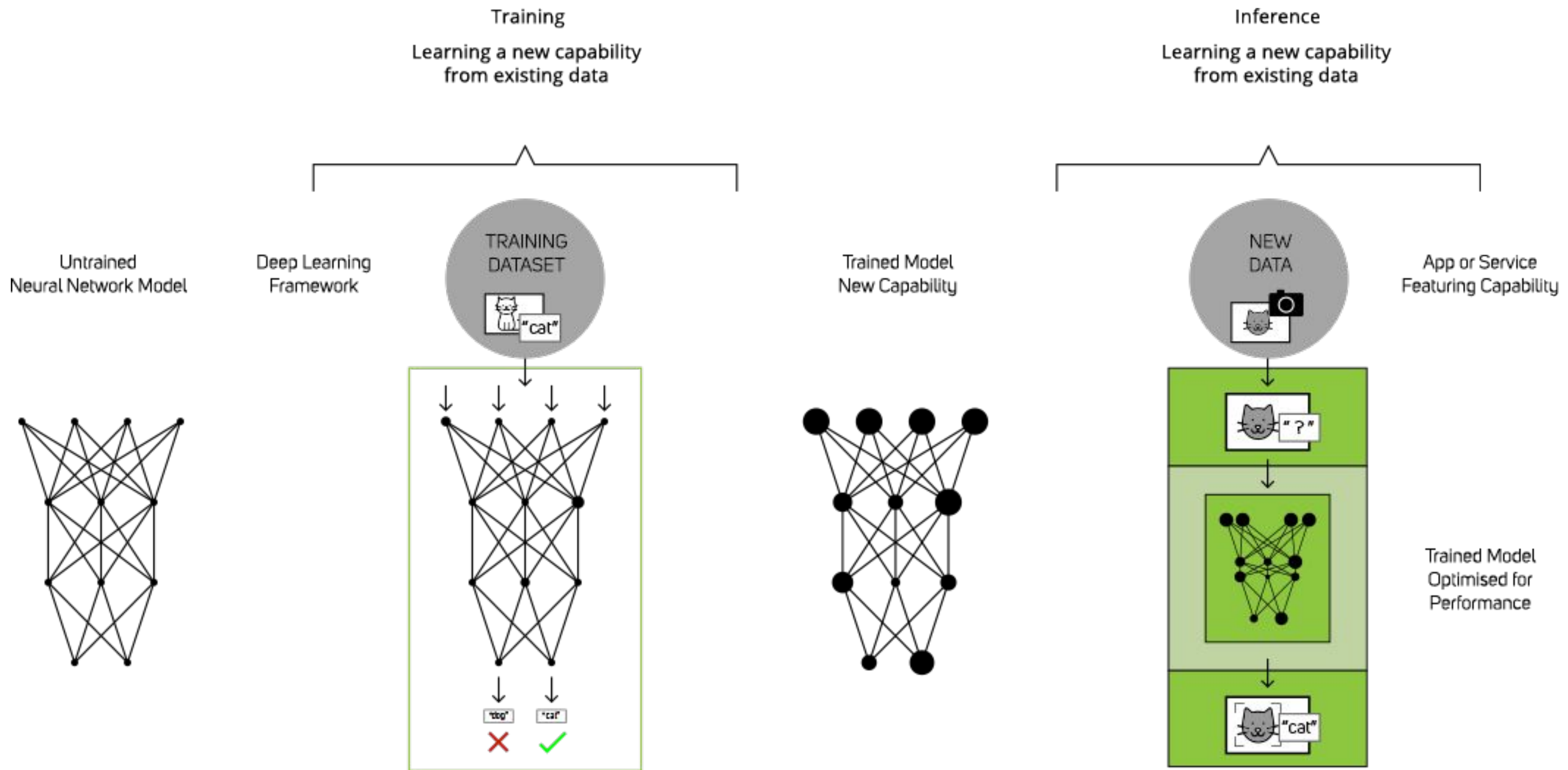
Relevance	Result
Query	
1	
2	
3	
4	
5	

Vector search conceptual architecture

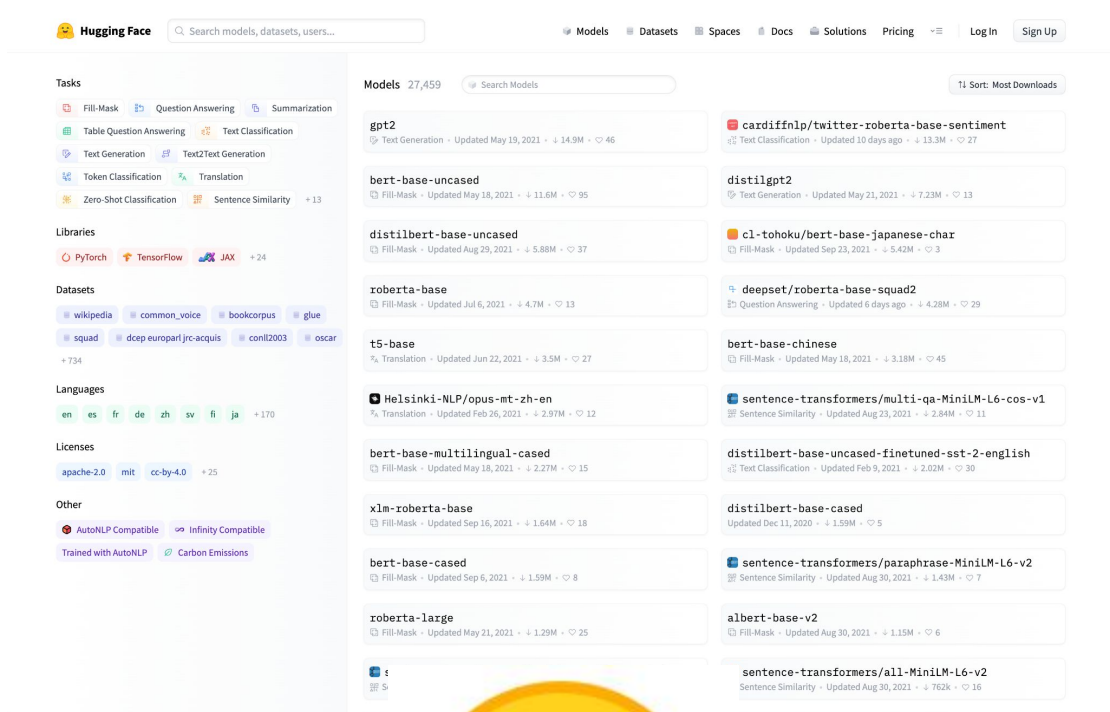
Use vector nearest neighbor to generate a search ranking



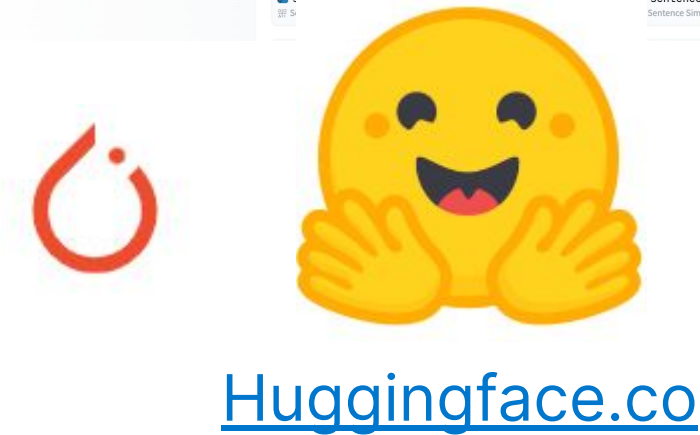
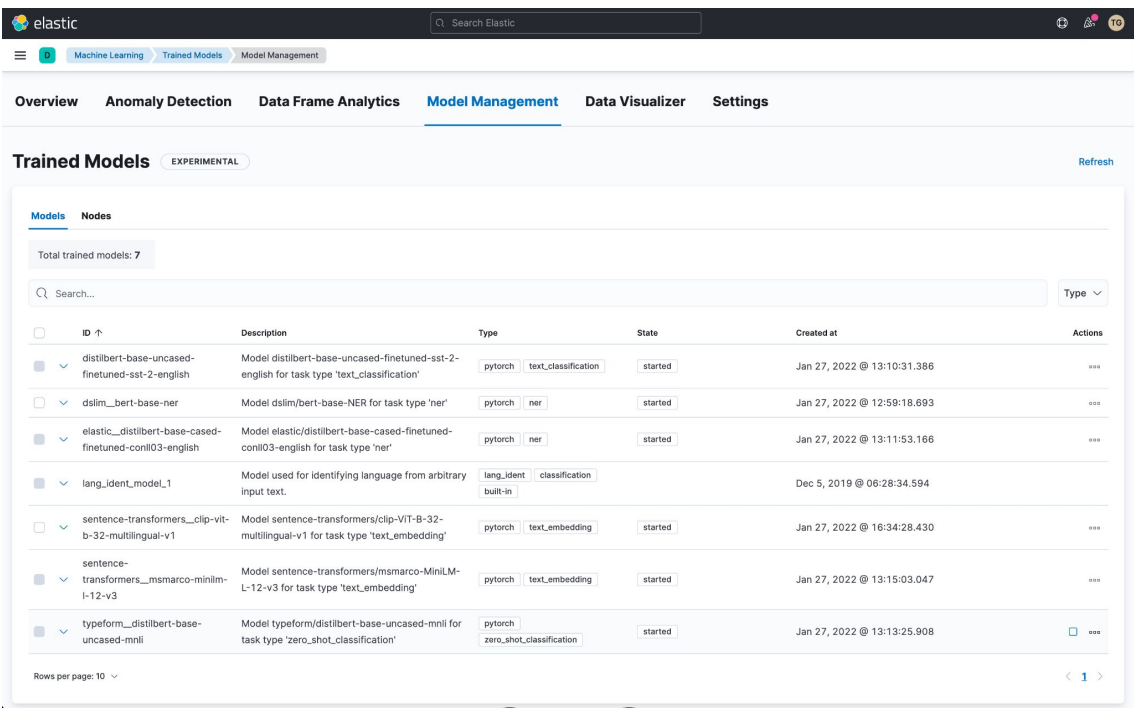
What is a Model?



Eland Imports PyTorch Models



```
$ eland_import_hub_model
--url https://Cluster_URL
--hub-model-id bert_model
--task-type text_embedding
--start
```



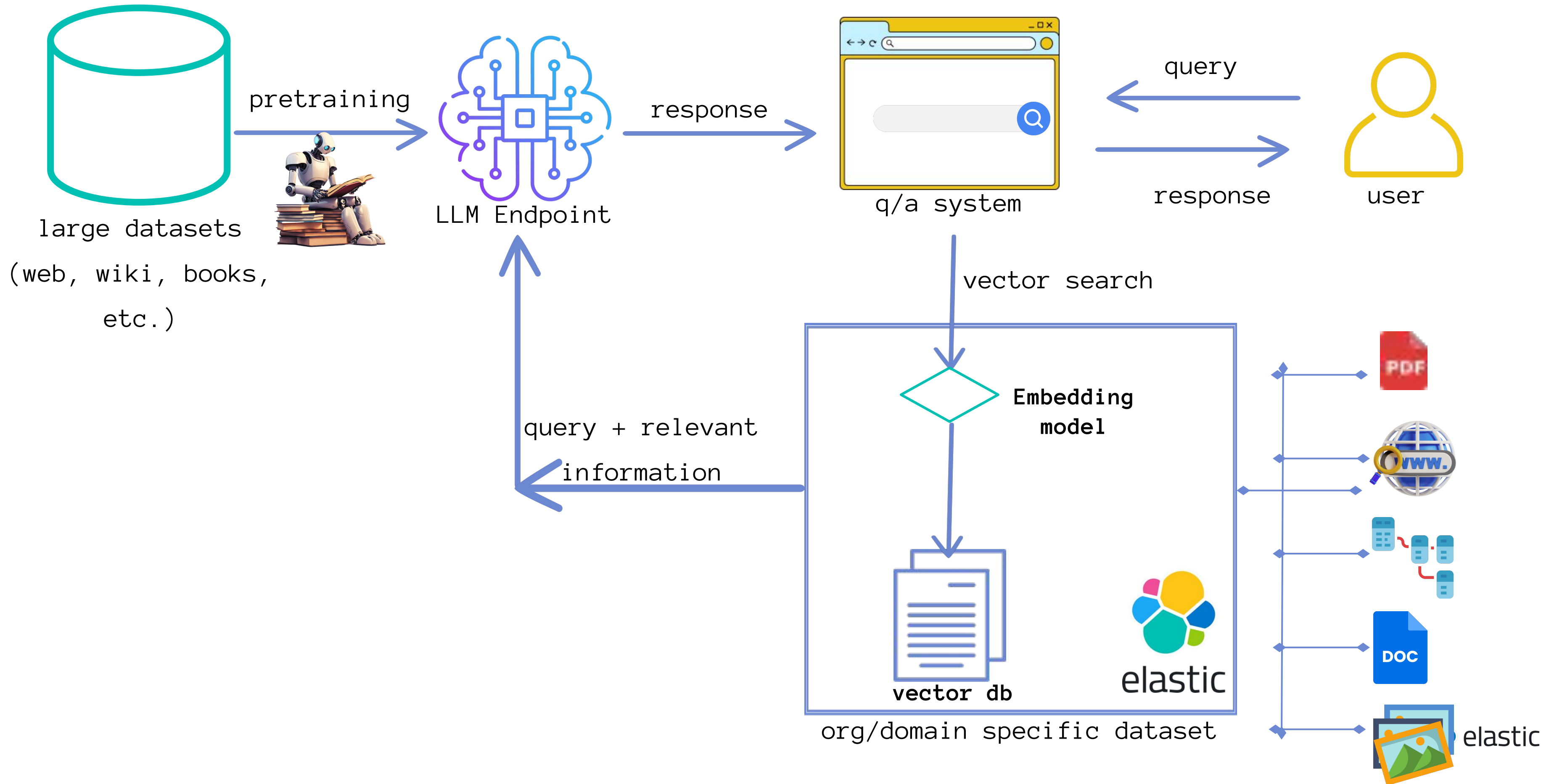
[Huggingface.co](https://huggingface.co)



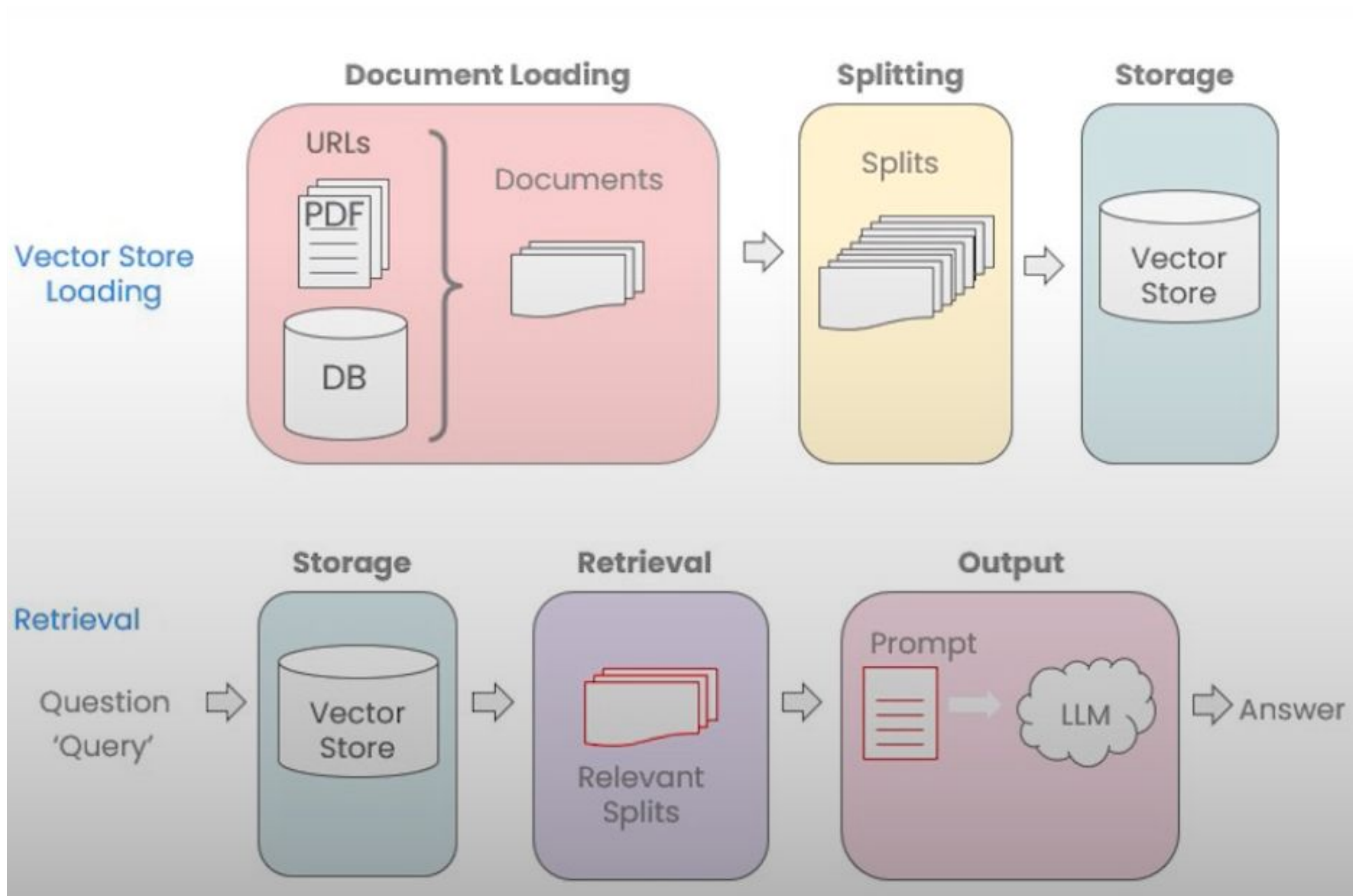
Inference, not training



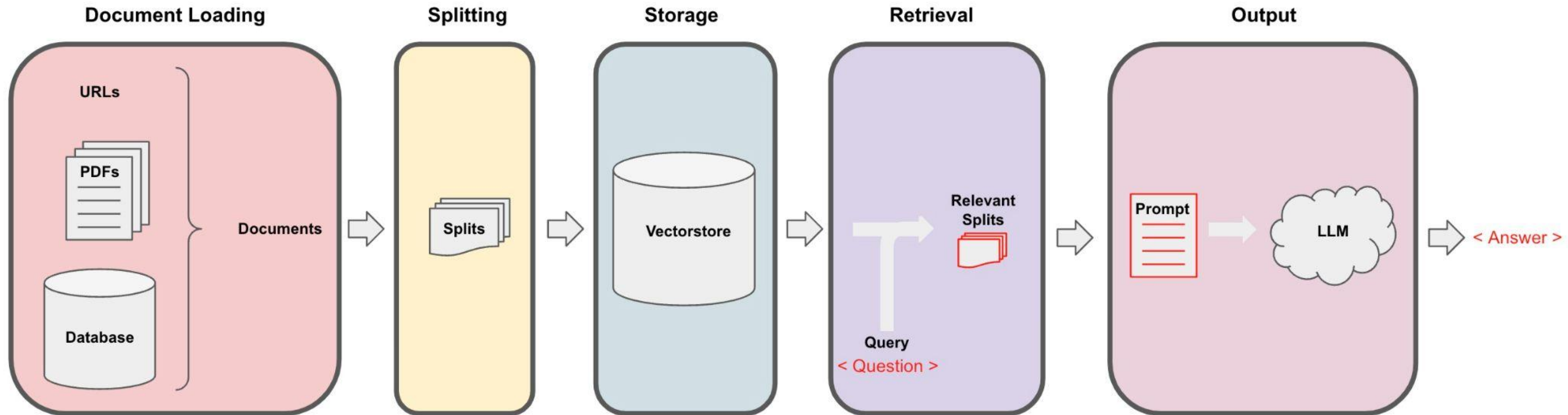
RAG Architecture



Chat with your data



Chat data pipeline





Demo



Connect with the Elastic Brazil Community

Find an Elastic group



<https://community.elastic.co/>

See local Elastic events



<https://ela.st/brvirtual>

Referências

<https://github.com/elastic/elasticsearch-labs/blob/main/notebooks/integrations/gemini/qa-langchain-gemini-elasticsearch.ipynb>

https://github.com/salgado/meetup_goiania/blob/main/Meetup_Goiânia_qa_langchain_gemini_elasticsearch.ipynb

Obrigado

Alex Salgado



@alexsgadoprof



salgado



@alexsgadoprof



/in/alex-salgado/