



Priscilla Parodi

Principal Developer Advocate at Elastic

Contact Info:

E-mail: priscilla.parodi@elastic.co

Linkedin: Priscilla Parodi



Improving Retrieval Quality for Search

Agenda

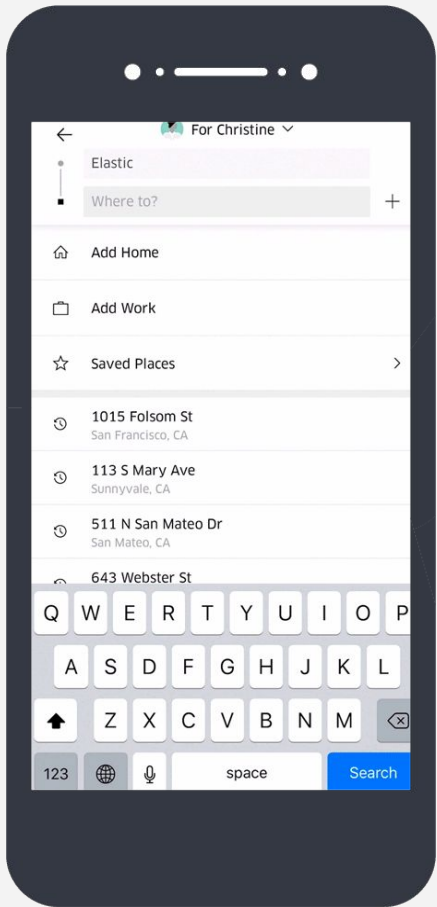
- **Revisit Text Search (Lexical Search)**
- **Vector Search**
- **Semantic Search with dense vectors (Dense Vector Retrieval)**
- **Semantic Search with sparse vectors (Learned Sparse Retrieval)**
- **Hybrid Retrieval**

First,

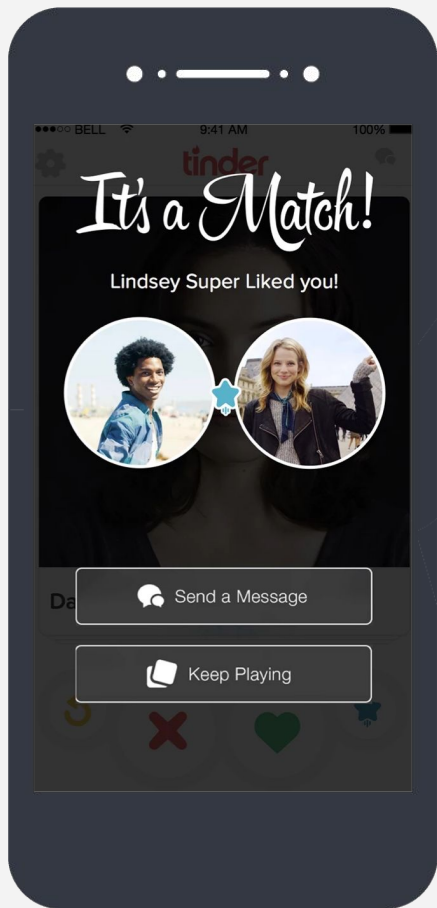
What is Elastic?

Elastic is a search company.





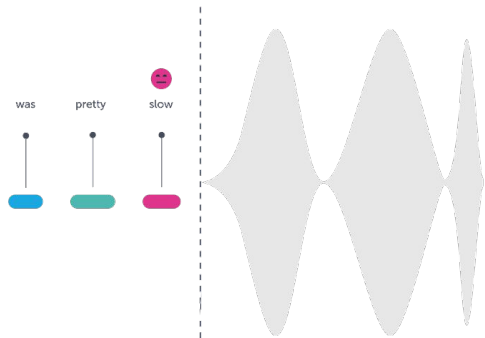
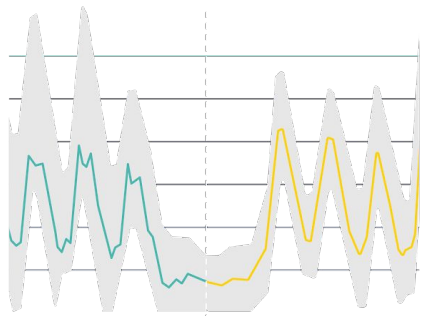
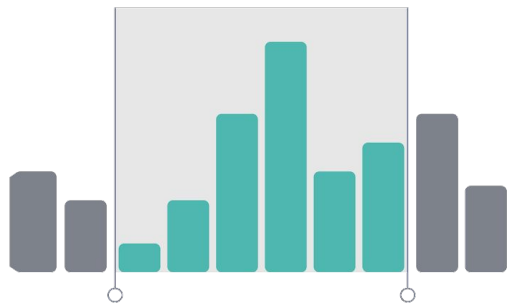
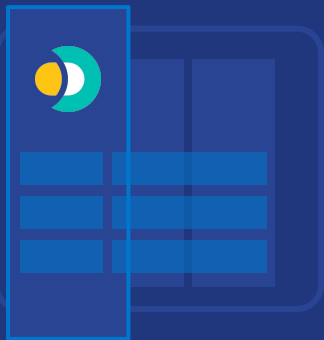
Uber



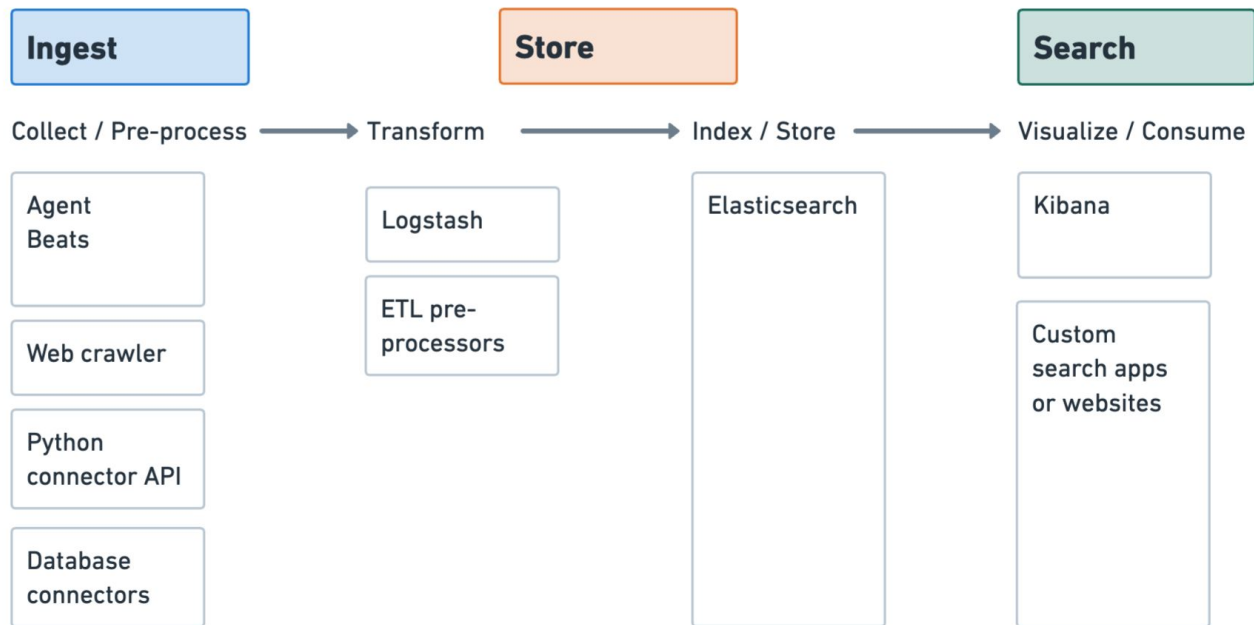
tinder™

Search

is constant



Typical **Search** Architecture



Text search revisited

"the best way to secure Elasticsearch"

tokenization

best way secure Elasticsearch

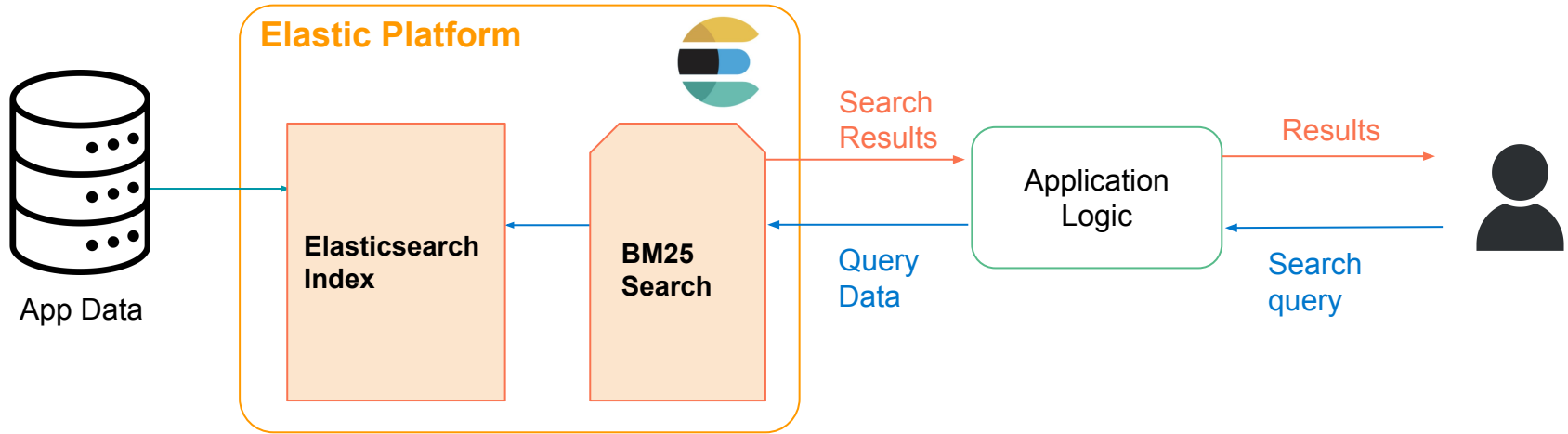
lexical matching

relevance is based on **frequency** and **rarity** of these terms

doc1: Tuning best practices
doc2: Elasticsearch: The best way to search
doc3: A guide to securing Elasticsearch

BM25 ranking function

Text search architecture revisited



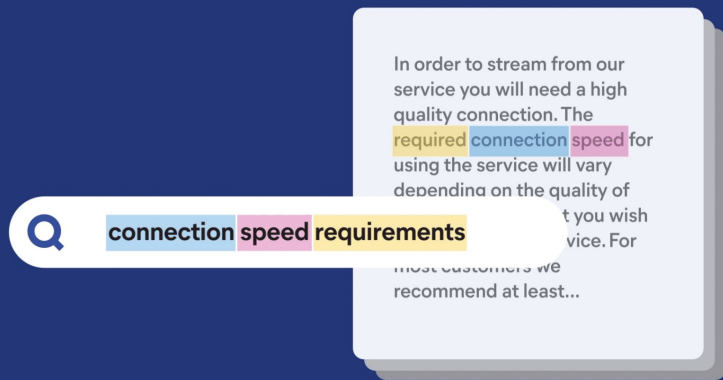
A. Search query

B. Execution of BM25 Search

Text Search is useful for many use cases

	Where it works...	Where it may fall short...
Text Search (BM25)	Well understood Interpretable	Vocabulary mismatch Context (semantic mismatch)

Users **expect more** from **Search**

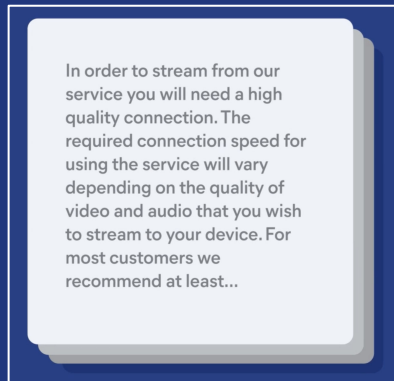


Retrieve results based on **intent** and **contextual meaning** of search queries, not just terms

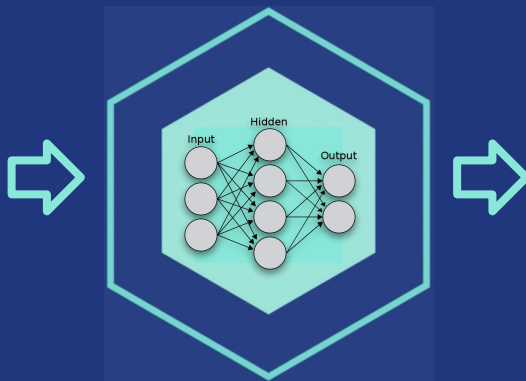
Vector Search enables Semantic Search

What is **Vector Search**?

Search based on Vector Representations (or “vector embeddings”)



Documents



Generate embeddings



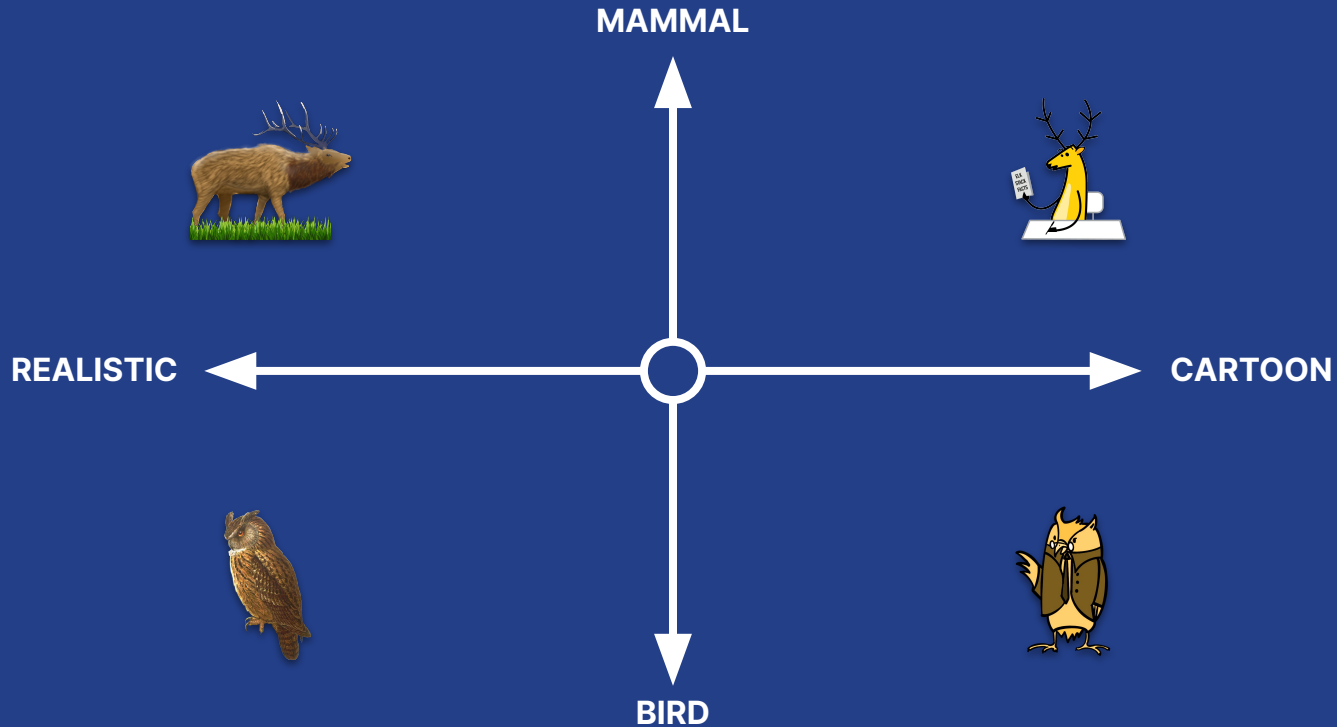
Vector Search

Embeddings represent your data

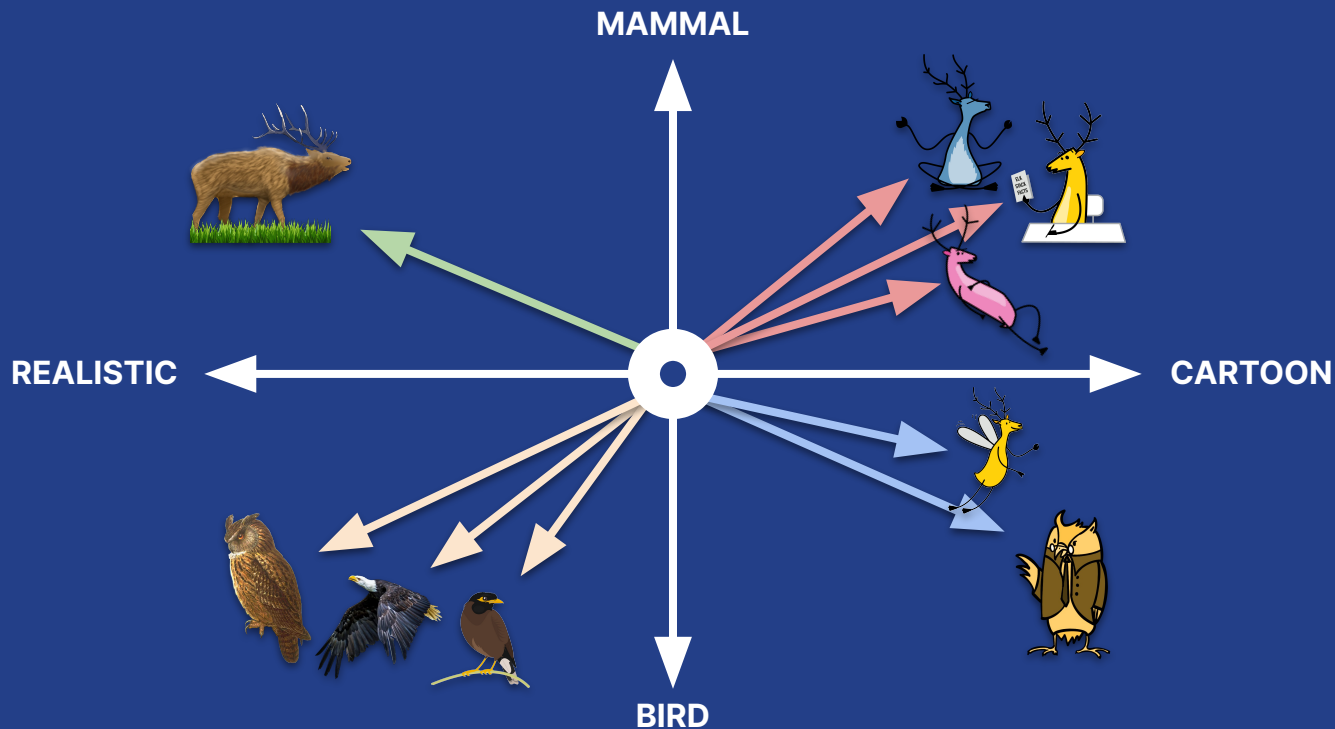
Example: 1-dimensional vector



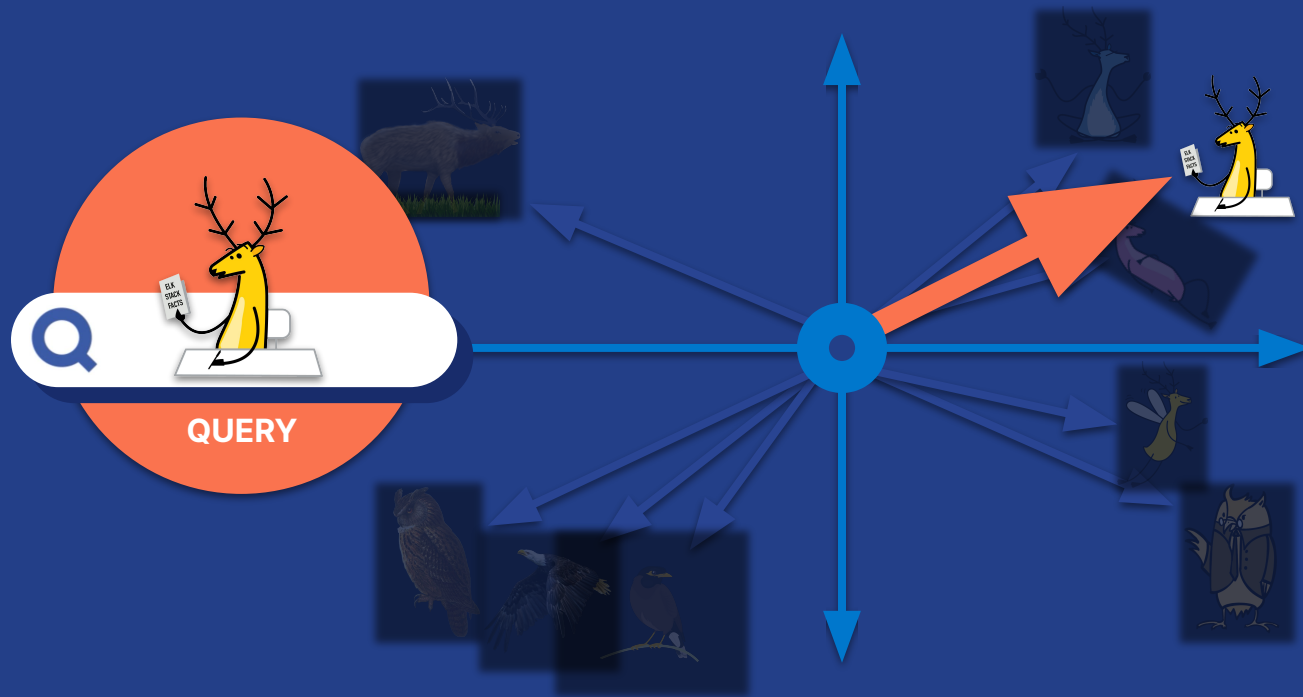
Multiple dimensions represent different aspects of data



In the “embedding space”, similar data are grouped together

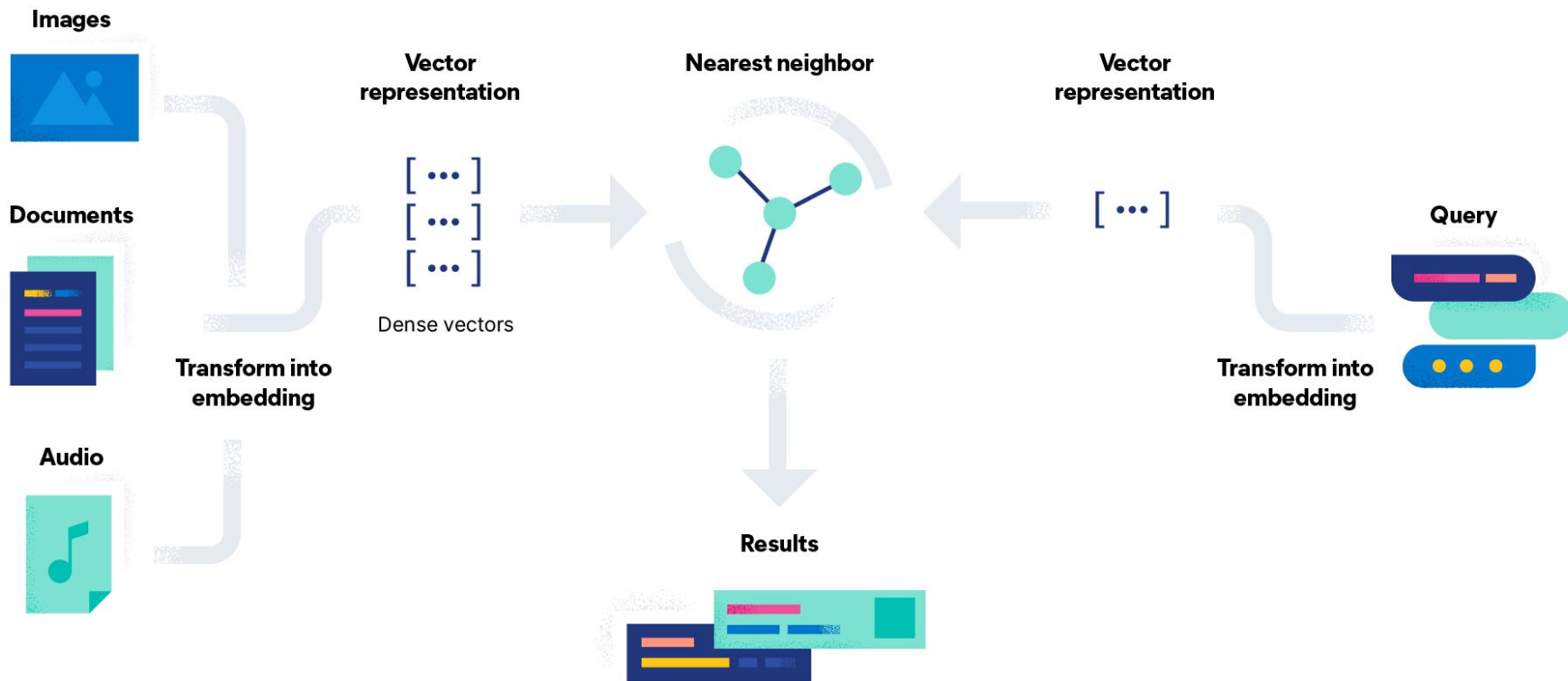


Vector search ranks objects by similarity (relevance) to the query



Relevance	Result
Query	
1	
2	
3	
4	
5	

Vector search: conceptual architecture

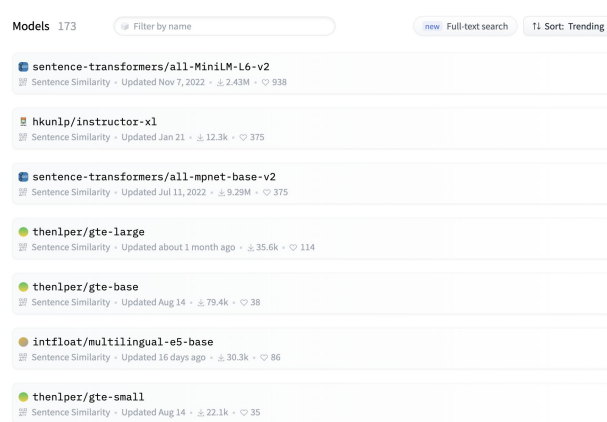


How to generate context aware text embeddings?

Apply a Natural Language Processing (NLP) model!

(+)With Elastic → Import and Deploy proprietary or third party NLP models.

Select the text embedding model



Models 173

Filter by name

Full-text search

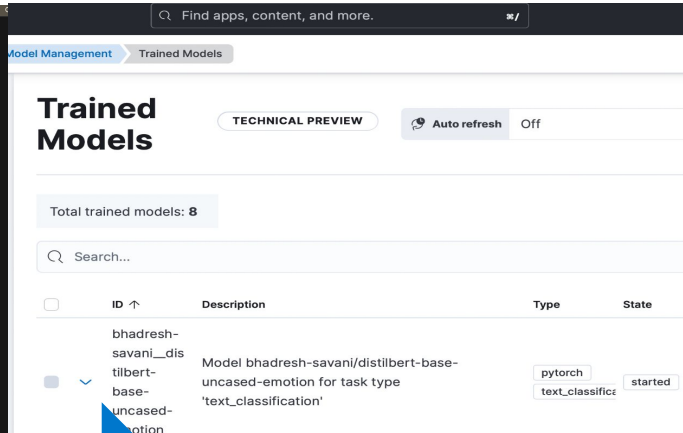
Sort: Trending

- sentence-transformers/all-MiniLM-L6-v2**
Sentence Similarity · Updated Nov 7, 2022 · 2.43M · 938
- hkunlp/instructor-x1**
Sentence Similarity · Updated Jan 21 · 12.3k · 375
- sentence-transformers/all-mpnet-base-v2**
Sentence Similarity · Updated Jul 11, 2022 · 9.29M · 375
- thenlper/gte-large**
Sentence Similarity · Updated about 1 month ago · 35.6k · 114
- thenlper/gte-base**
Sentence Similarity · Updated Aug 14 · 79.4k · 38
- intfloat/multilingual-e5-base**
Sentence Similarity · Updated 16 days ago · 30.3k · 86
- thenlper/gte-small**
Sentence Similarity · Updated Aug 14 · 22.1k · 35

Import into Elasticsearch

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

Deploy and use the model



Model Management

Trained Models

TECHNICAL PREVIEW

Auto refresh Off

Total trained models: 8

Search...

ID	Description	Type	State
bhadresh-savani_distilbert-base-uncased-emotion	Model bhadresh-savani/distilbert-base-uncased-emotion for task type 'text_classification'	pytorch	started



Hugging Face



eland
Python Elasticsearch
Client

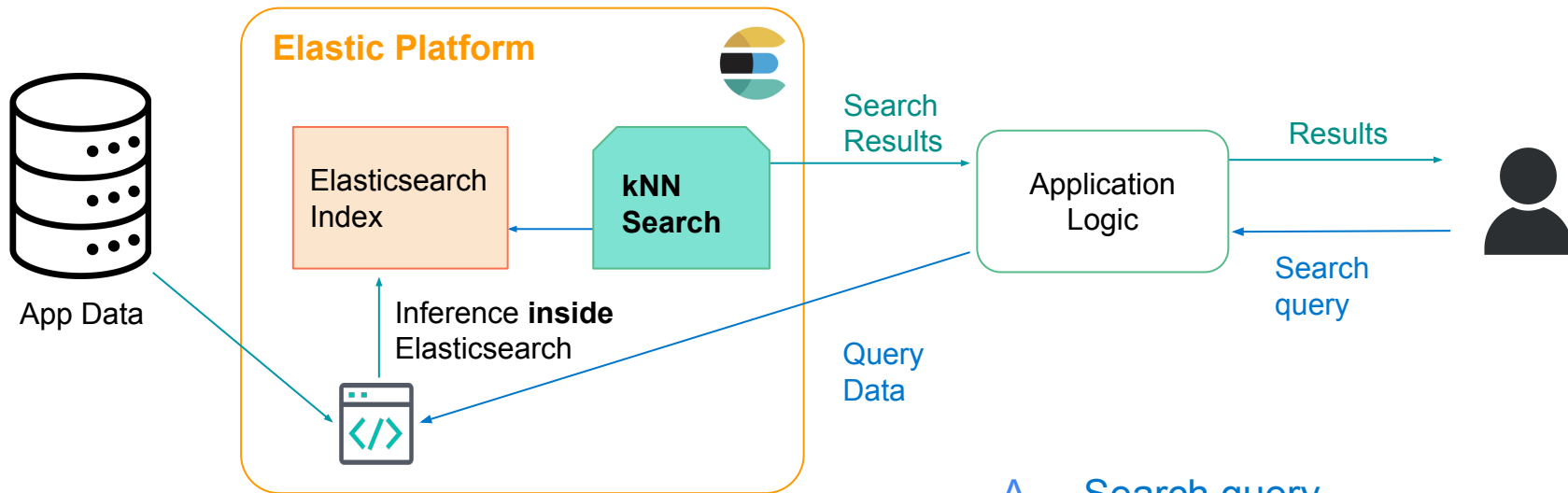


Elastic
Cluster

Open repository:
huggingface.co

Generating dense embeddings with Elastic: Two options

Generate embeddings **inside** Elasticsearch

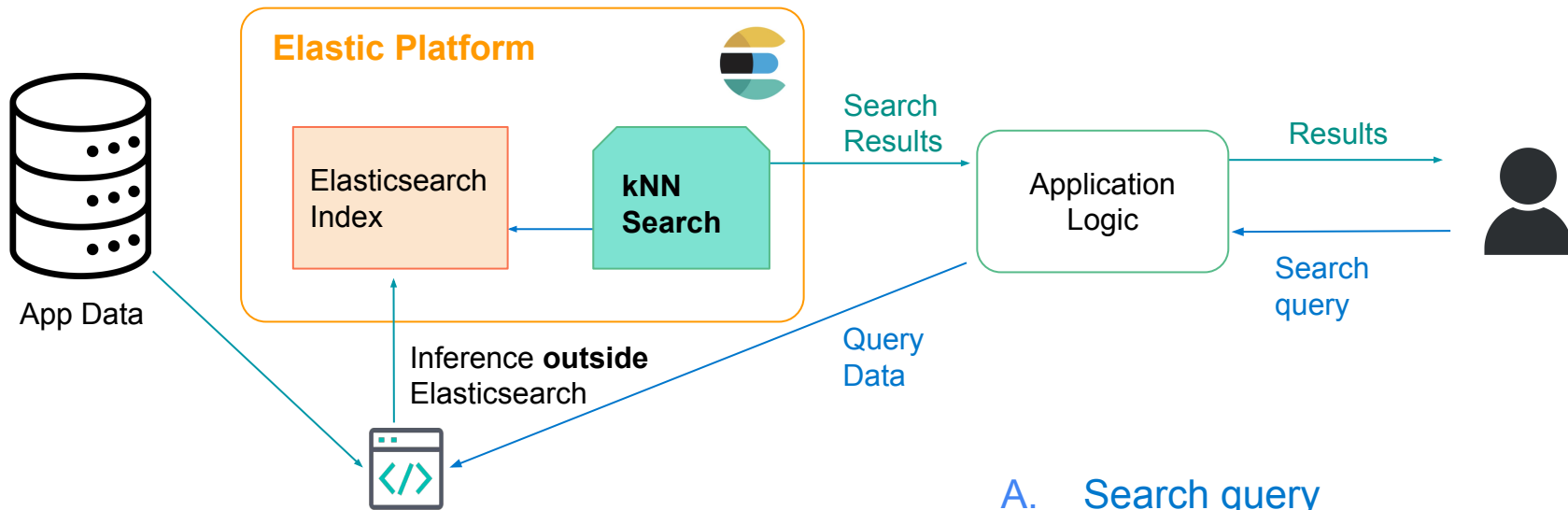


A. Search query

B. Execution of kNN Search

Generating dense embeddings with Elastic: Two options

Generate embeddings **outside** Elasticsearch



A. Search query

B. Execution of kNN Search

Dense Vector retrieval performs well (conditions apply*)

	Where it works...	Where it may fall short...
Dense Vector	Can beat other approaches for semantic search	Domain adaptation* Not easily interpretable (no exact match)

Learned sparse retrieval - an alternative approach for **Semantic Search**

Provides 'trade-offs' over dense retrieval and traditional sparse retrieval methods (BM25)

Term Expansion

By identifying contextual importance between terms, it utilizes that knowledge to improve sparse vector embeddings

Query: "Comfortable furniture for a large balcony"

Doc: "is a comfortable and stylish garden lounge set, including a sofa, chairs, and a side table for outdoor relaxation"

Term Expansion

By identifying contextual importance between terms, it utilizes that knowledge to improve sparse vector embeddings

Query: "Comfortable furniture for a large balcony"

without term expansion
(lexical search)

Doc: "is a comfortable and stylish garden lounge set, including a sofa, chairs, and a side table for outdoor relaxation"

Term Expansion

By identifying contextual importance between terms, it utilizes that knowledge to improve sparse vector embeddings

Query: "Comfortable furniture for a large balcony"

with term expansion

landscape

sofa

porch

relax

couch

garden

calm

chairs

outdoor

sleep

side table

leisure

Doc: "is a comfortable and stylish garden lounge set, including a sofa, chairs, and a side table for outdoor relaxation"

Elastic provides a 'built-in' option for this approach!

(1) Download the model



Trained Models

Auto refresh Off Refresh

Total trained models: 1

Search... Type ▾

<input type="checkbox"/>	ID ↑	Description	Type	State	Created at	Download model
<input type="checkbox"/>	.elser_model_1	Elastic Learned Sparse Encoder	curated	-	-	

(2) Start the model deployment



Deployment ID

Specify unique identifier for the model deployment.

Deployment ID:

Priority

Select low priority for demonstrations where each model will be very lightly used.

Priority: low normal

Number of allocations

Increase to improve throughput of all requests.

Number of allocations:

Threads per allocation

Increase to improve latency for each request.

Threads per allocation: 1 2 4 8 16

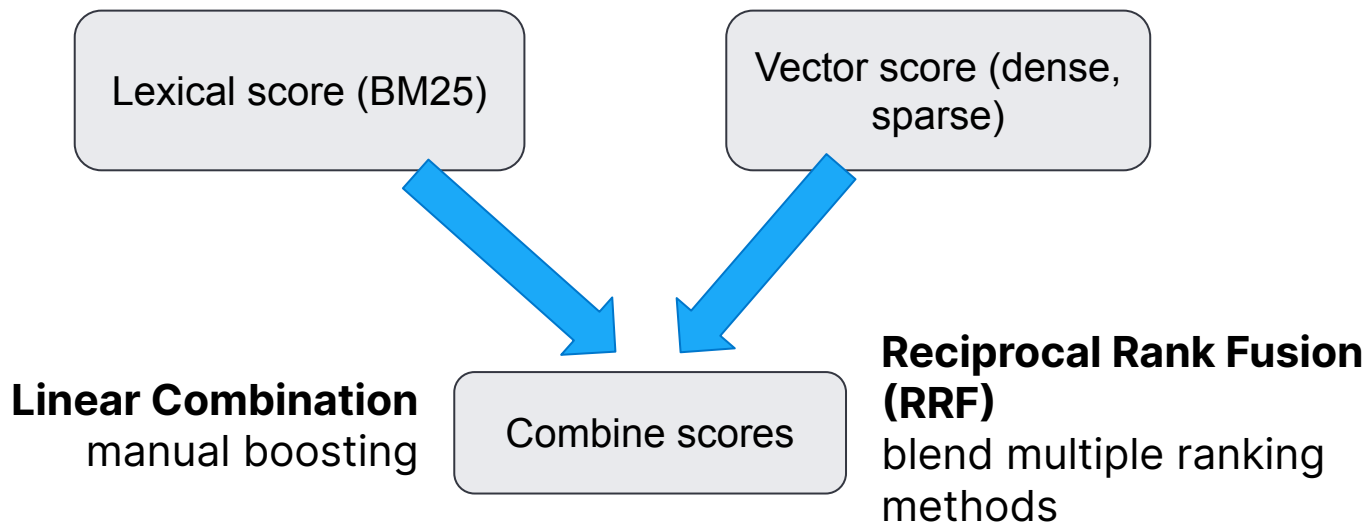
[Learn more](#) Cancel Start

Learned Sparse retrieval is an improvement over text search

	Where it works...	Where it may fall short...
Text Search (BM25)	Well understood Interpretable	Vocabulary mismatch Semantic mismatch
Learned Sparse Retrieval (ELSER)	+ Interpretable / Well understood (tokens) + Vocabulary/semantic matching - Perform Semantic Search	Larger index (terms/tokens) <u>Dense vector retrieval</u> can outperform learned sparse retrieval for <u>semantic search</u>

Is a combined approach a better idea?

Hybrid retrieval:



Powered by

Elasticsearch Relevance Engine™



Vector database



Ability to host your own transformer model



Ability to integrate with 3rd party transformer models (OpenAI)



RRF - hybrid scoring model (vector & textual search)



Elastic's proprietary ML model



Integration with 3rd party tooling like LangChain

Thank you!

Questions and demo at the Elastic booth

Contact Info:

E-mail: priscilla.parodi@elastic.co

Linkedin: Priscilla Parodi

