Priscilla Parodi

Principal Developer Advocate at Elastic

Contact Info:
E-mail: priscilla.parodi@elastic.co
Linkedin: Priscilla Parodi

elastic

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

elastic

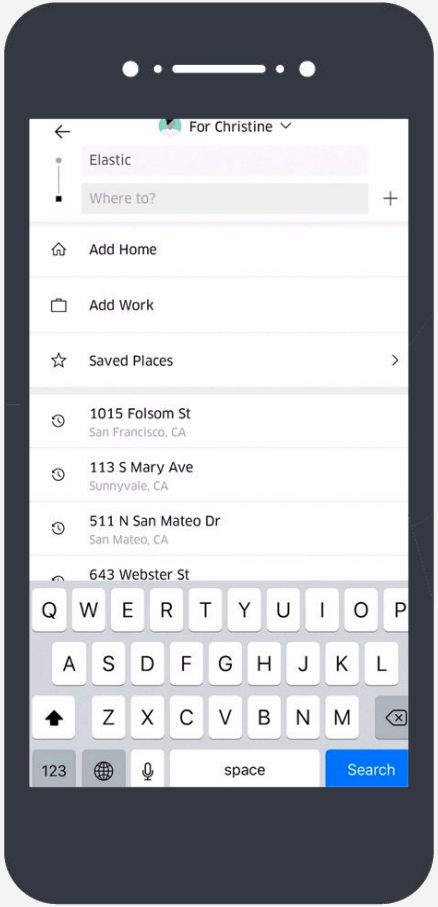
First,

What is Elastic?

elastic

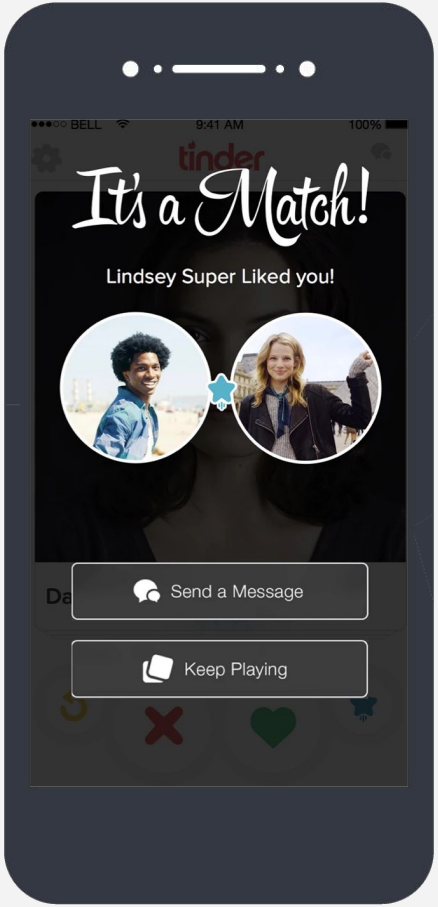
Elastic is a search company.

elastic



Uber

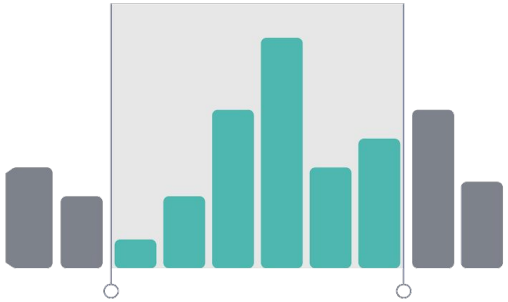
elastic

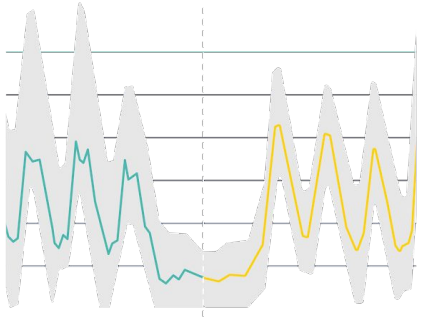


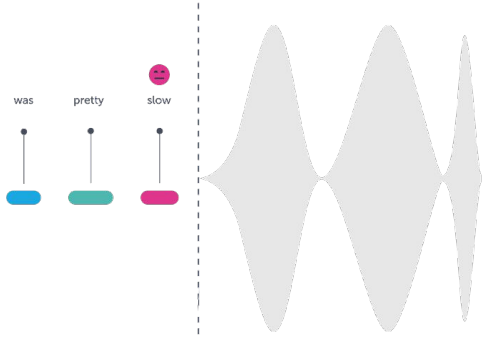
tinder™

elastic

Search
is constant

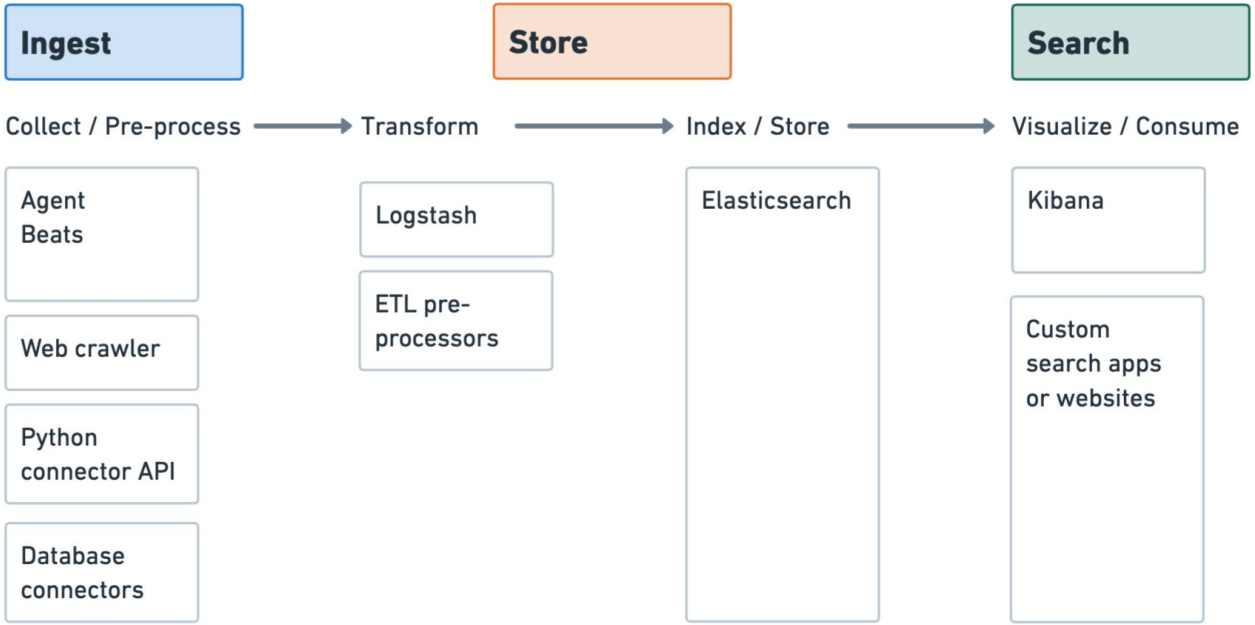






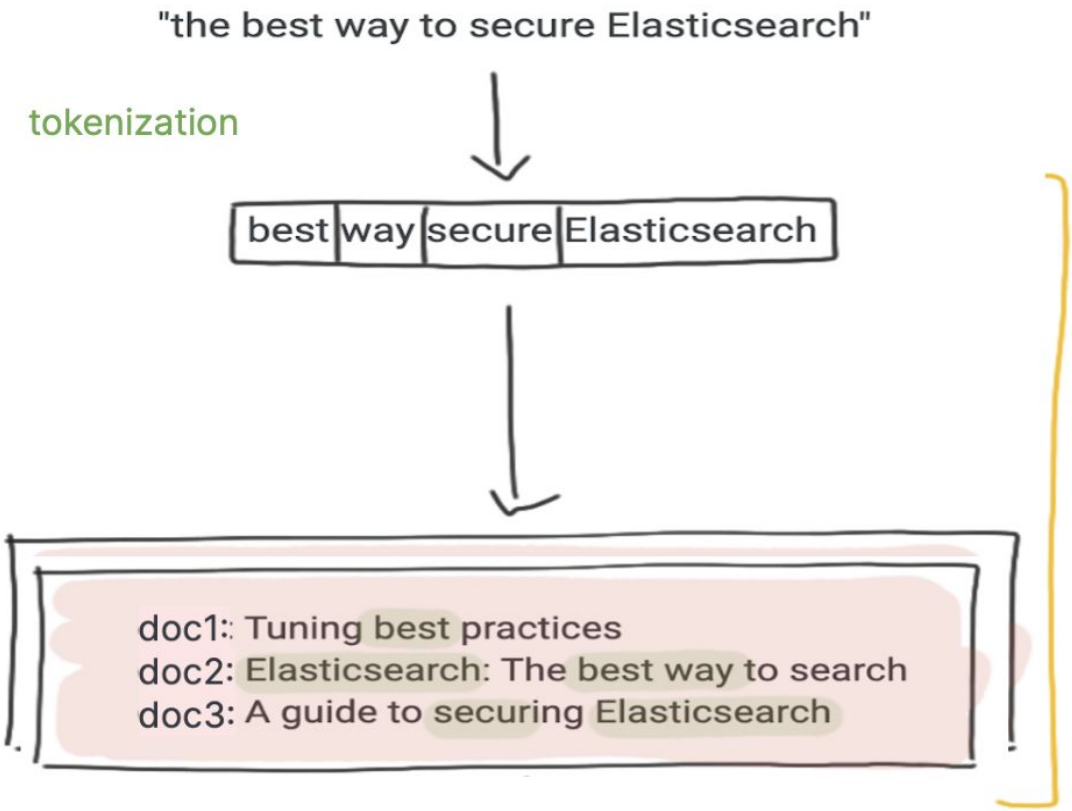
elastic

Typical **Search** Architecture



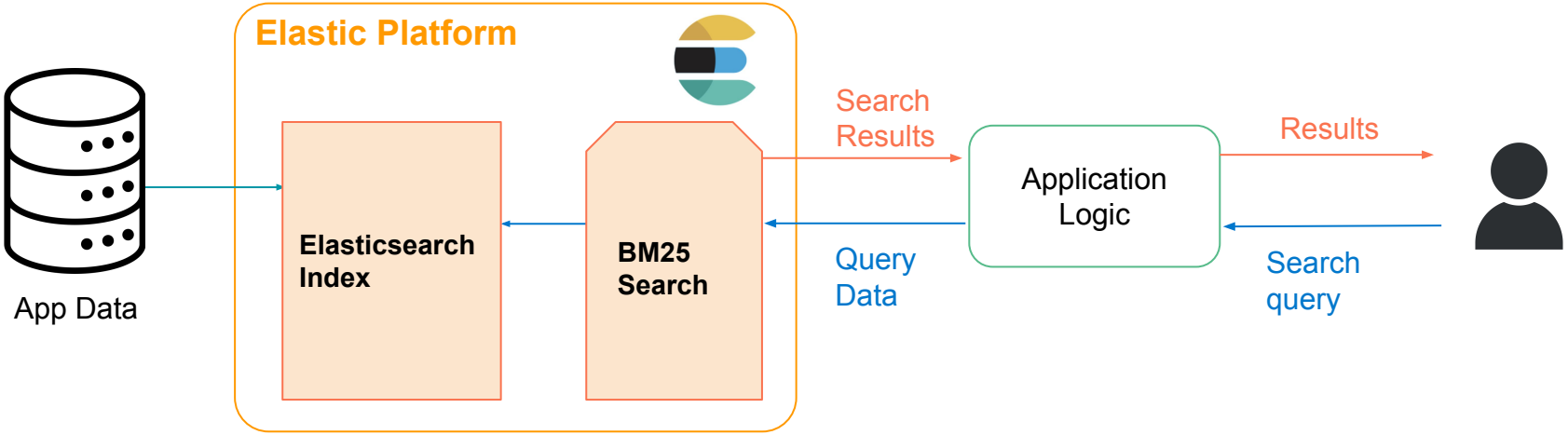
elastic

Text search revisited



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

Text search architecture revisited



A. Search query

B. Execution of BM25 Search

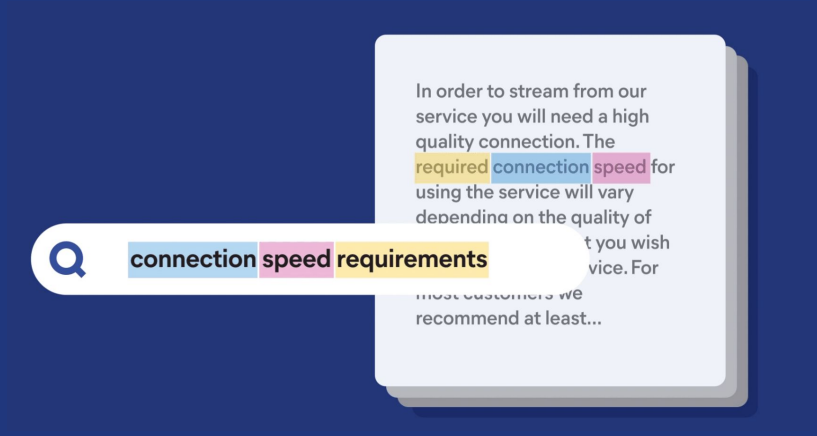
elastic

Text Search is useful for many use cases

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

elastic

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



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

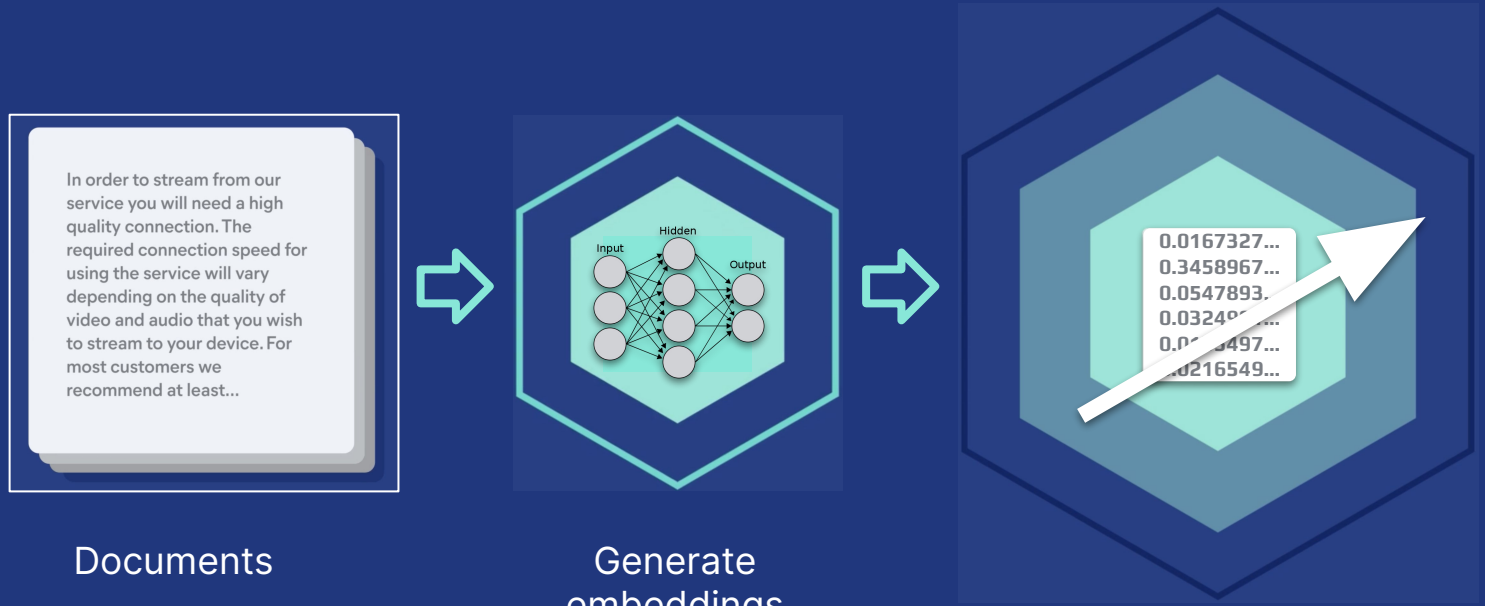
elastic

Vector Search enables Semantic Search

elastic

What is **Vector Search**?

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



Documents

Generate embeddings

Vector Search

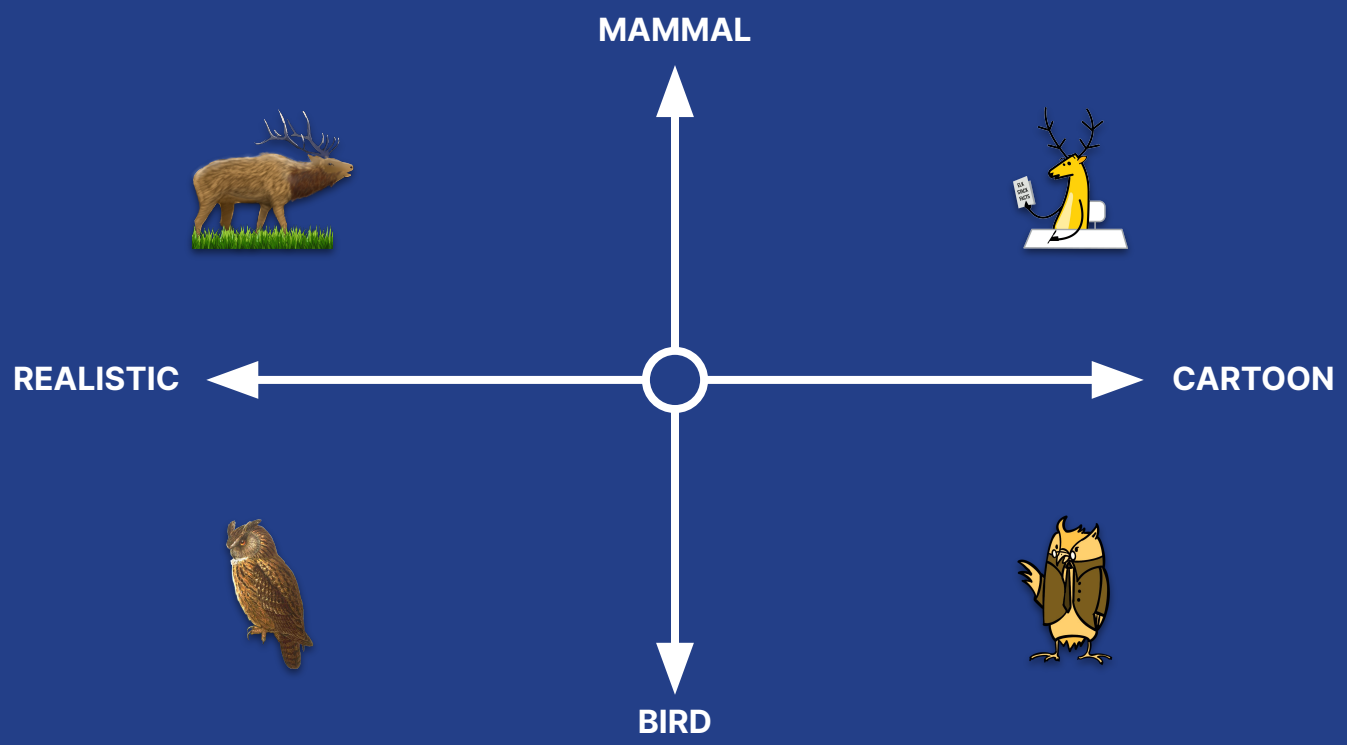
elastic

Embeddings represent your data

Example: 1-dimensional vector

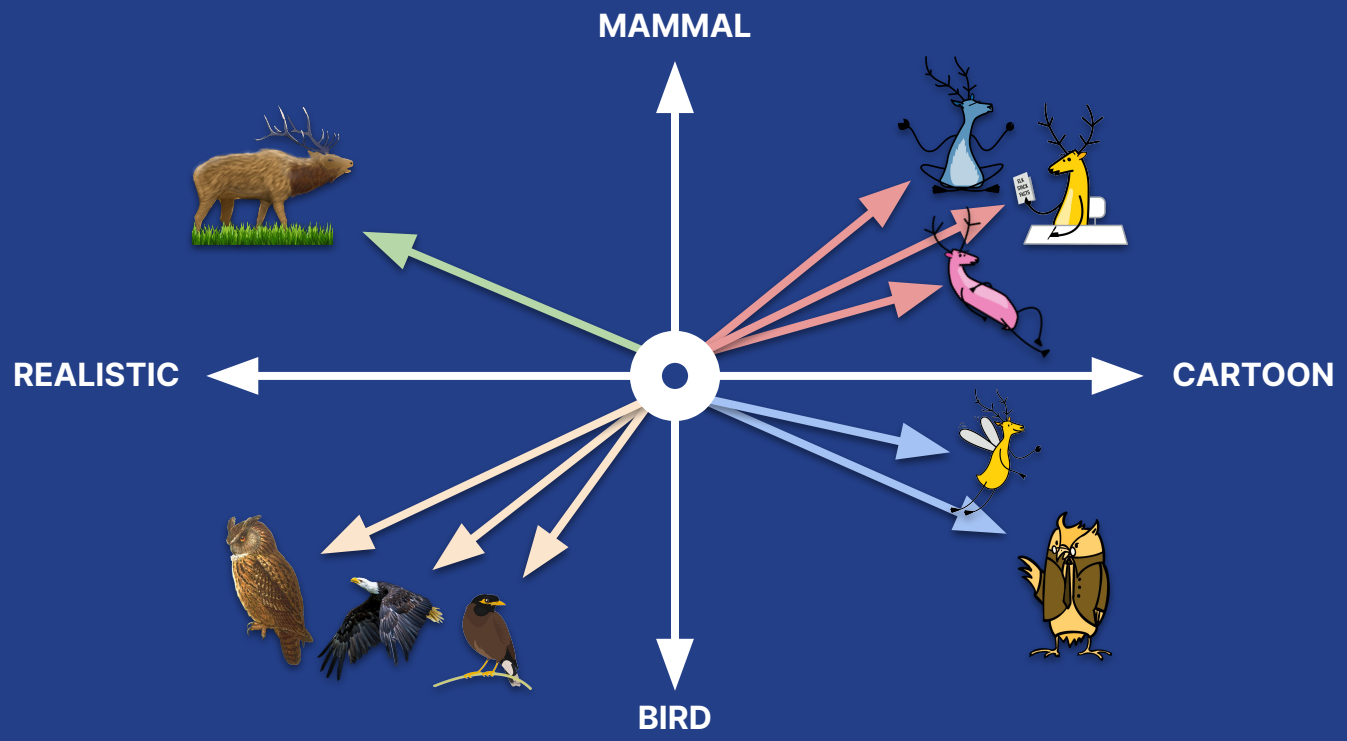
REALISTIC ← → CARTOON

Multiple dimensions represent different aspects of data



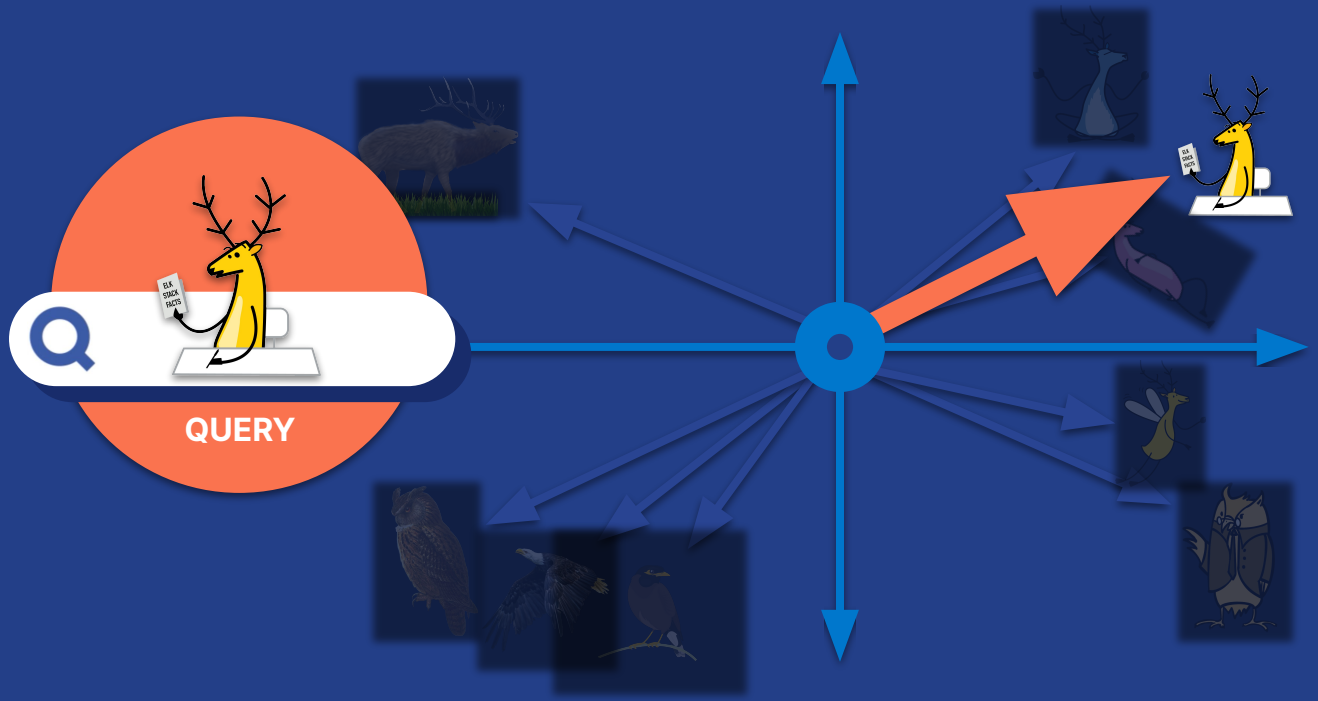
elastic

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



elastic

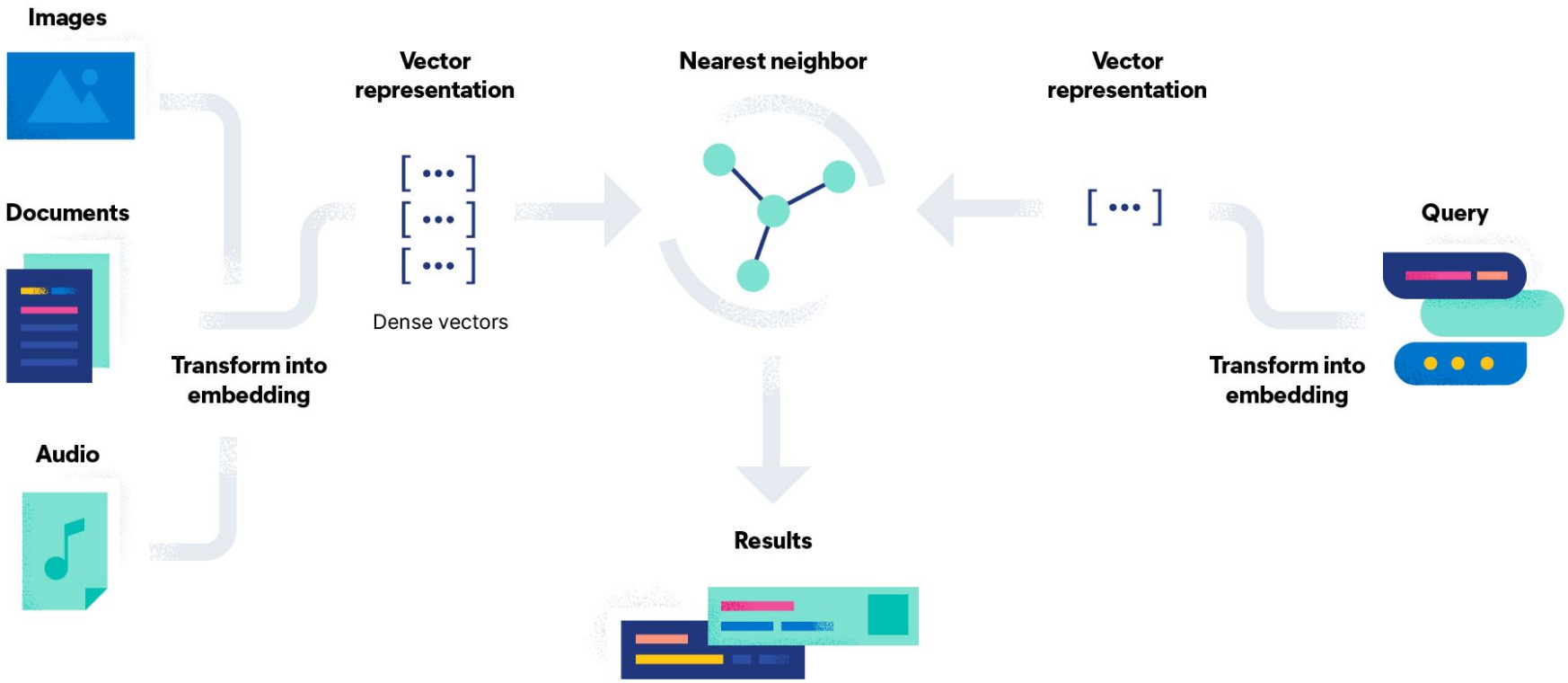
Vector search ranks objects by similarity (relevance) to the query



Relevance	*Result*
Query	
1	
2	
3	
4	
5	

elastic

Vector search: conceptual architecture

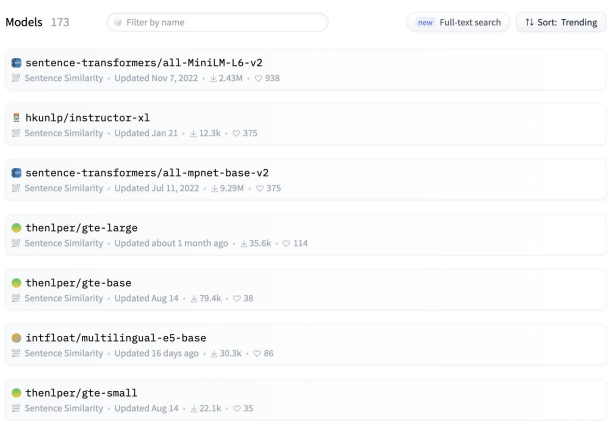


How to generate context aware text embeddings?

Apply a Natural Language Processing (NLP) model!

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

Select the text embedding model



Hugging Face

PyTorch

Open repository:
huggingface.co

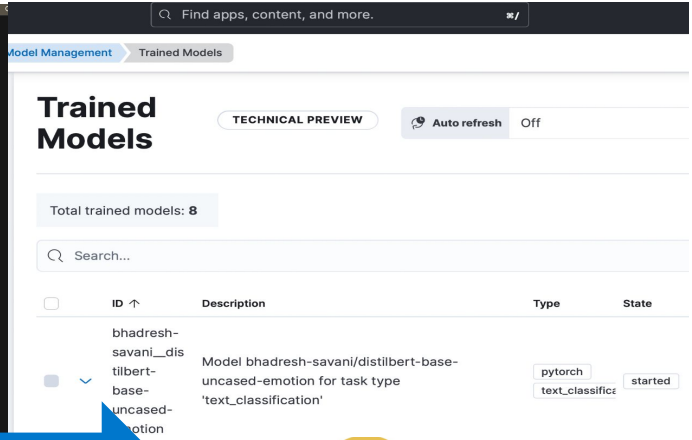
Import into Elasticsearch

```
$ eland_import_hub_model

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

eland
Python Elasticsearch Client

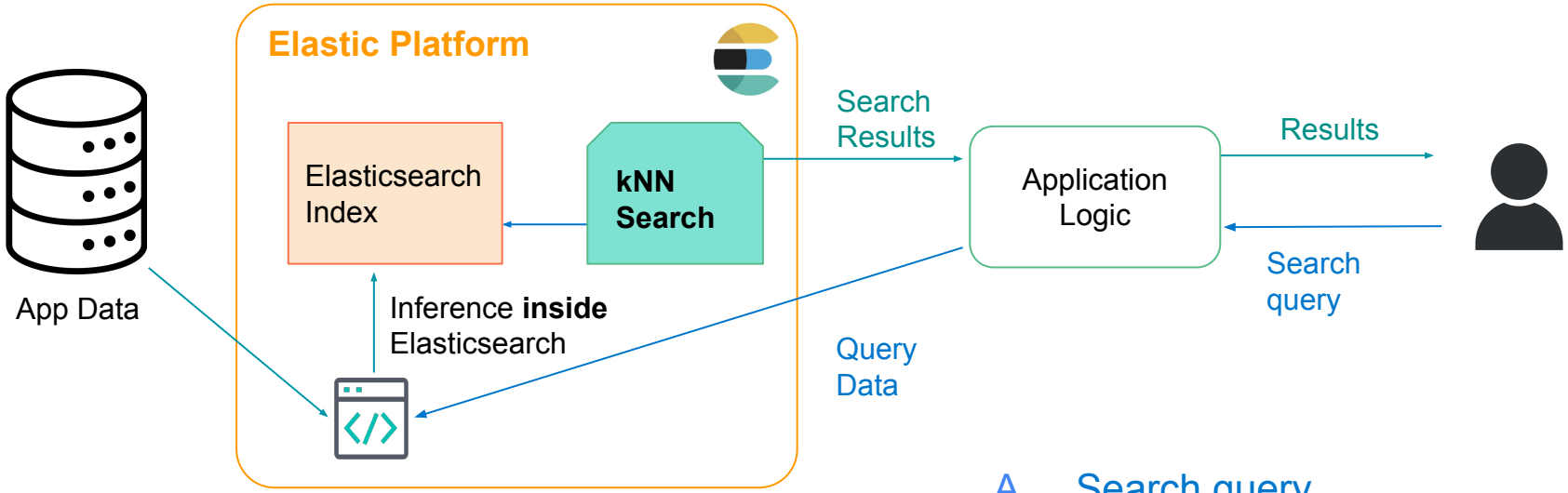
Deploy and use the model



Elastic Cluster

Generating dense embeddings with Elastic: Two options

Generate embeddings **inside** Elasticsearch

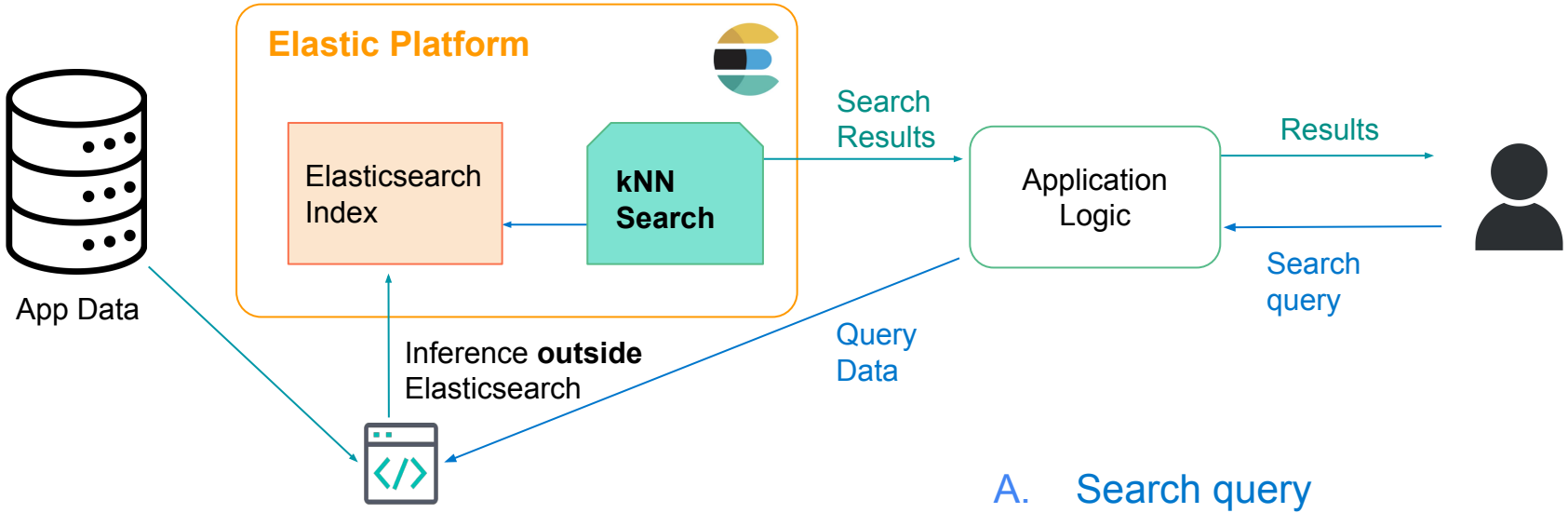


A. Search query

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)

elastic

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

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

elastic

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"
```

elastic

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"

elastic

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

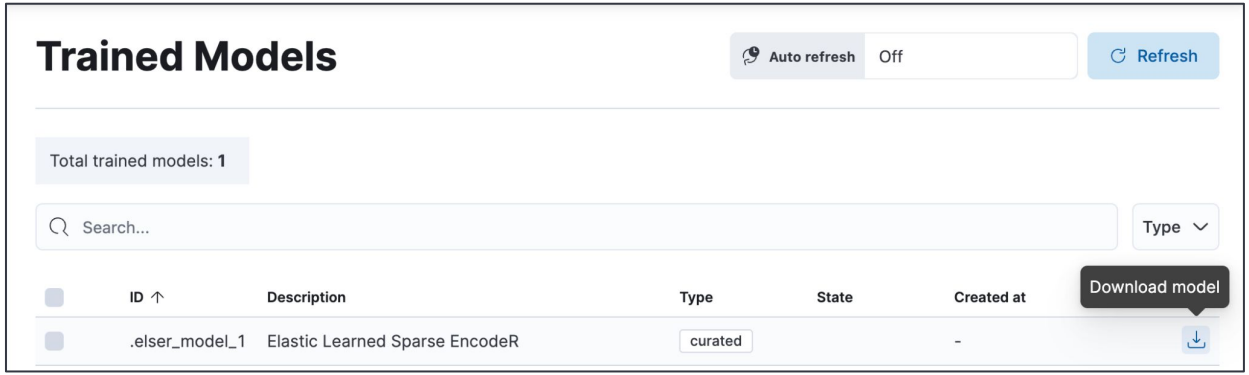
landscape	sofa	porch
relax	couch	garden
calm	chairs	outdoor
sleep	side table	leisure

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

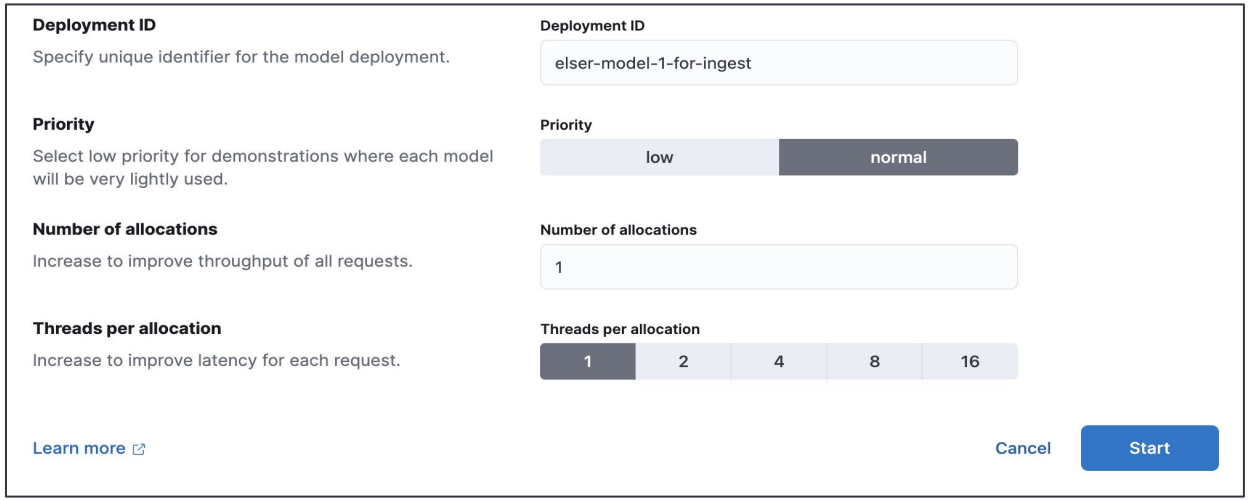
elastic

Elastic provides a ‘built-in’ option for this approach!

(1) Download the model



(2) Start the model deployment



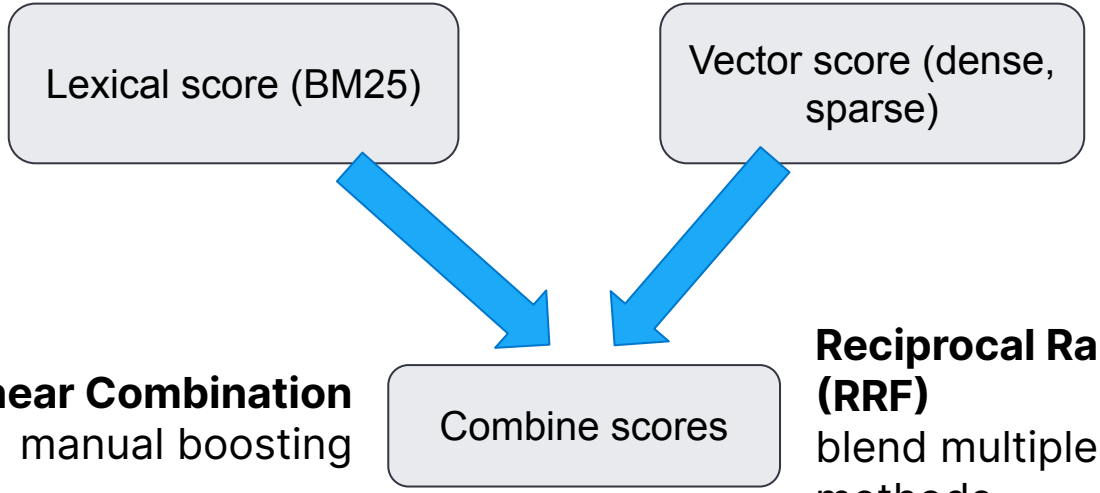
Learned Sparse retrieval is an improvement over text search

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

elastic

Is a combined approach a better idea?

Hybrid retrieval:



Linear Combination
manual boosting

Reciprocal Rank Fusion (RRF)
blend multiple ranking methods

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

elastic

Join the Elastic community!

1

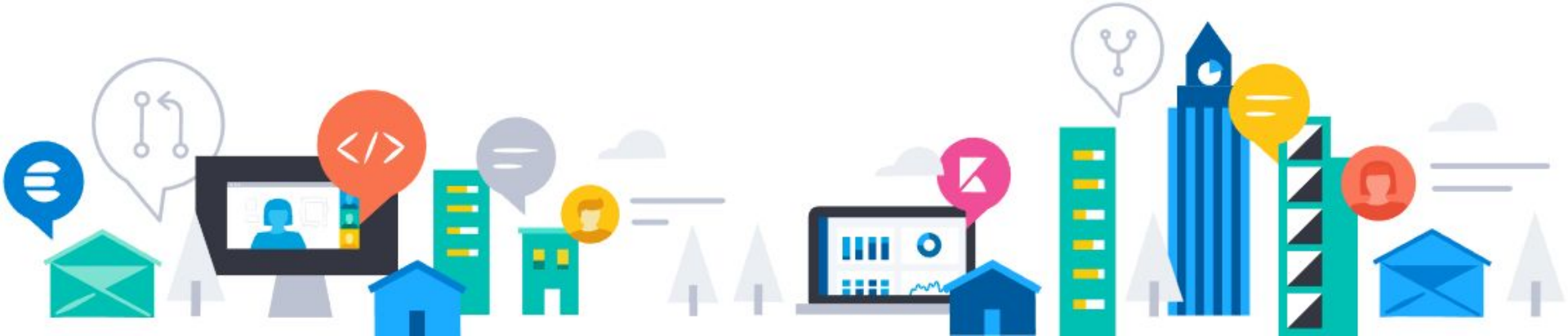
Learn More:
elastic.co/search-labs

2

Try free on Cloud:
elastic.co/cloud

3

Connect on Slack:
ela.st/slack



Thank you!

Questions and demo at the Elastic booth

Contact Info:
E-mail: priscilla.parodi@elastic.co
Linkedin: Priscilla Parodi

elastic