Understanding Vector Databases

PyLadies Meetup NYC

elastic

Jessica Garson

@JessicaGarson

@JessicaGarson@macaw.social

Senior Developer Advocate

elastic

This talk is aimed for developers who may have some experience with machine learning but looking to skill up.

elastic

This talk will explore the traditional "Classic" search, highlight its constraints, and provide an overview of vector databases.

Elasticsearch:
You Know, for Search

elastic

elasticsearch

What's the well-known sentence of this scene in Star Wars?

elastic

These are not the droids
you are looking for.

elastic

```
GET /_analyze
{
  "char_filter": [ "html_strip" ],
  "tokenizer": "standard",
  "filter": [ "lowercase", "stop", "snowball" ],
  "text": "These are <em>not</em> the droids
           you are looking for."
}
```

elastic

```
"char_filter": "html_strip"
```

```
These are <em>not</em> the droids you are
   looking for.
```

```
These are not the droids you are
   looking for.
```

elastic

```
"tokenizer": "standard"
```

These are not the droids you are looking for.

```
These
are
not
the
droids
you
are
looking
for
```

```
"filter": "lowercase"
```

These		these
are		are
not		not
the		the
droids		droids
you		you
are		are
looking		looking
for		for

elastic

```
"filter": "stop"
```

These		
are		
not		
the		
droids		droids
you		you
are		
looking		looking
for		

elastic

```
"filter": "snowball"
```

droid**s**	droid
you	you
look**ing**	look

elastic

```
These are <em>not</em> the droids you are looking for.


{ "tokens": [{
      "token": "droid",
      "start_offset": 27, "end_offset": 33,
      "type": "<ALPHANUM>", "position": 4
    },{
      "token": "you",
      "start_offset": 34, "end_offset": 37,
      "type": "<ALPHANUM>", "position": 5
    }, {
      "token": "look",
      "start_offset": 42, "end_offset": 49,
      "type": "<ALPHANUM>", "position": 7
    }]}
```

elastic

PROS

- Fast and cheap
- Scalable
- Easy to understand

CONS

- Synonyms and homonyms
- Images, videos, audio
- User intent

Semantic Search:
Meaning, not literal matches

elastic

X-wing starfighter squadron

What ships and crews do I need to destroy an almost finished death star?

Or is there a secret weakness?

Elasticsearch:
You Know, for Vector Search

elastic

What do you need to get started with vector search?

elastic

- Elasticsearch instance optimized for machine learning (8.13 or later)
- Python 3.8 (or later)
- Elasticsearch Python Client

elastic


What is a vector database?

Why use vector search?

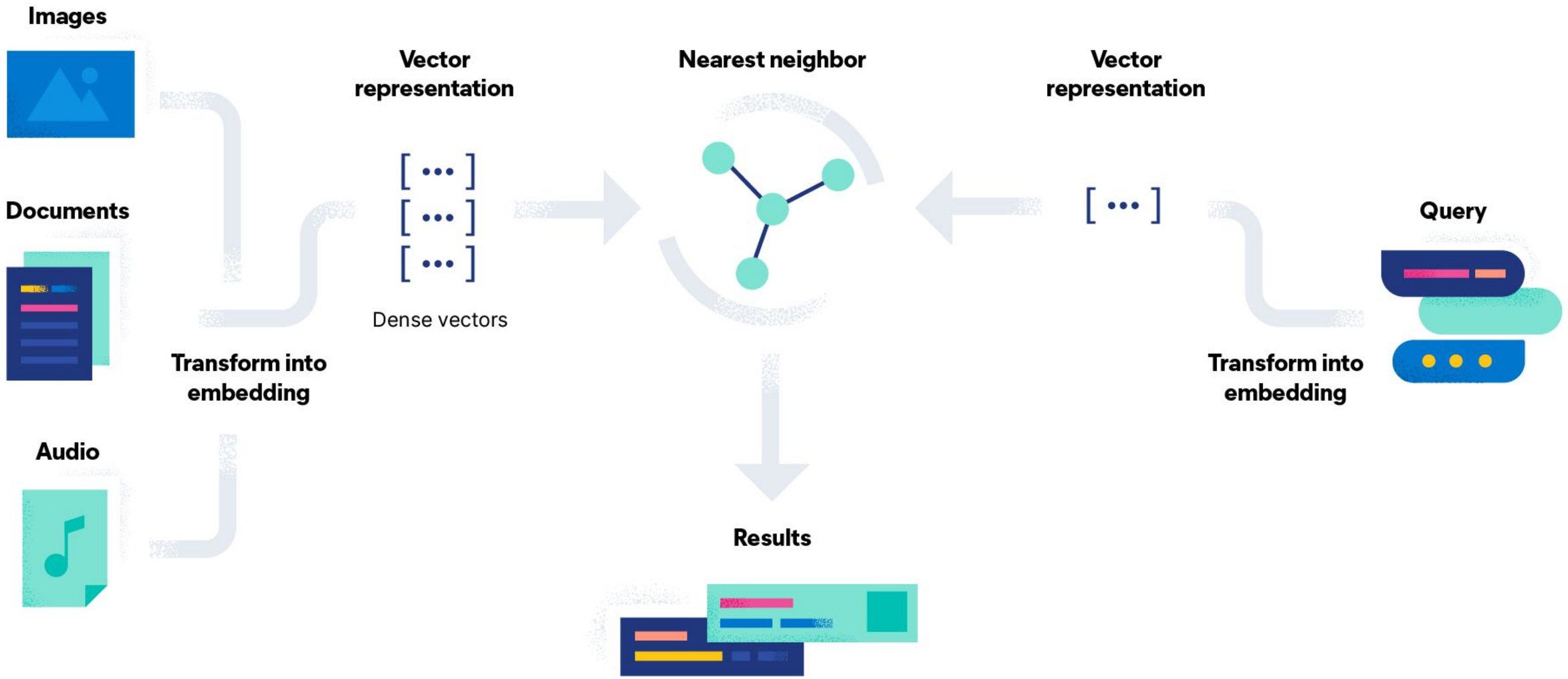
Vector search provides the foundation for implementing semantic search for text or similarity search for images, videos, or audio.

elastic

A vector database allows you to find similar data quickly.

elastic

Searching in a vector database



What’s a vector?

elastic

What are embeddings?

Embeddings leverage a machine learning model to translate text into numbers, allowing you to perform vector searches.

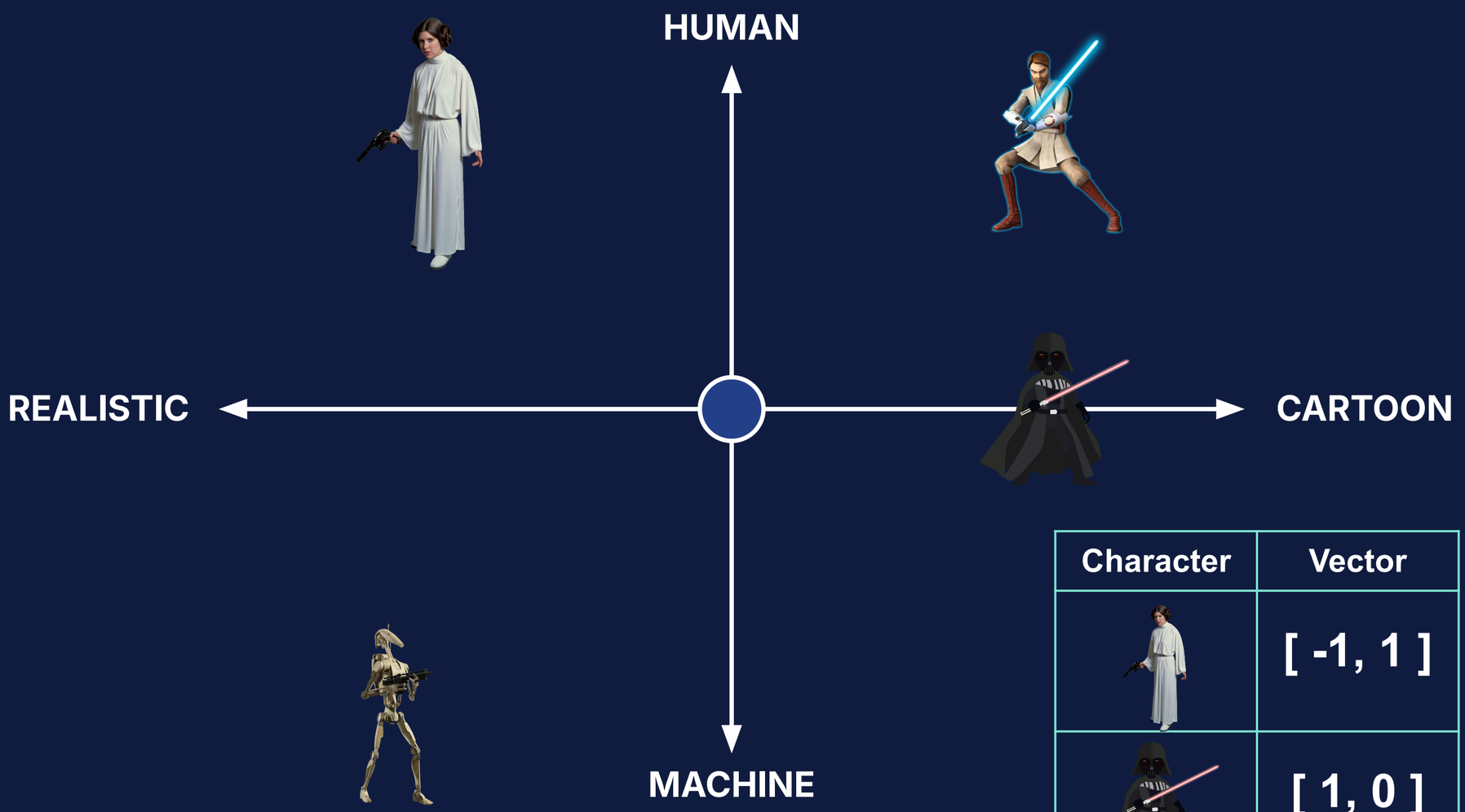
Embeddings represent your data

Example: 1-dimensional vector

Character	Vector
	[-1]
	[1]

elastic

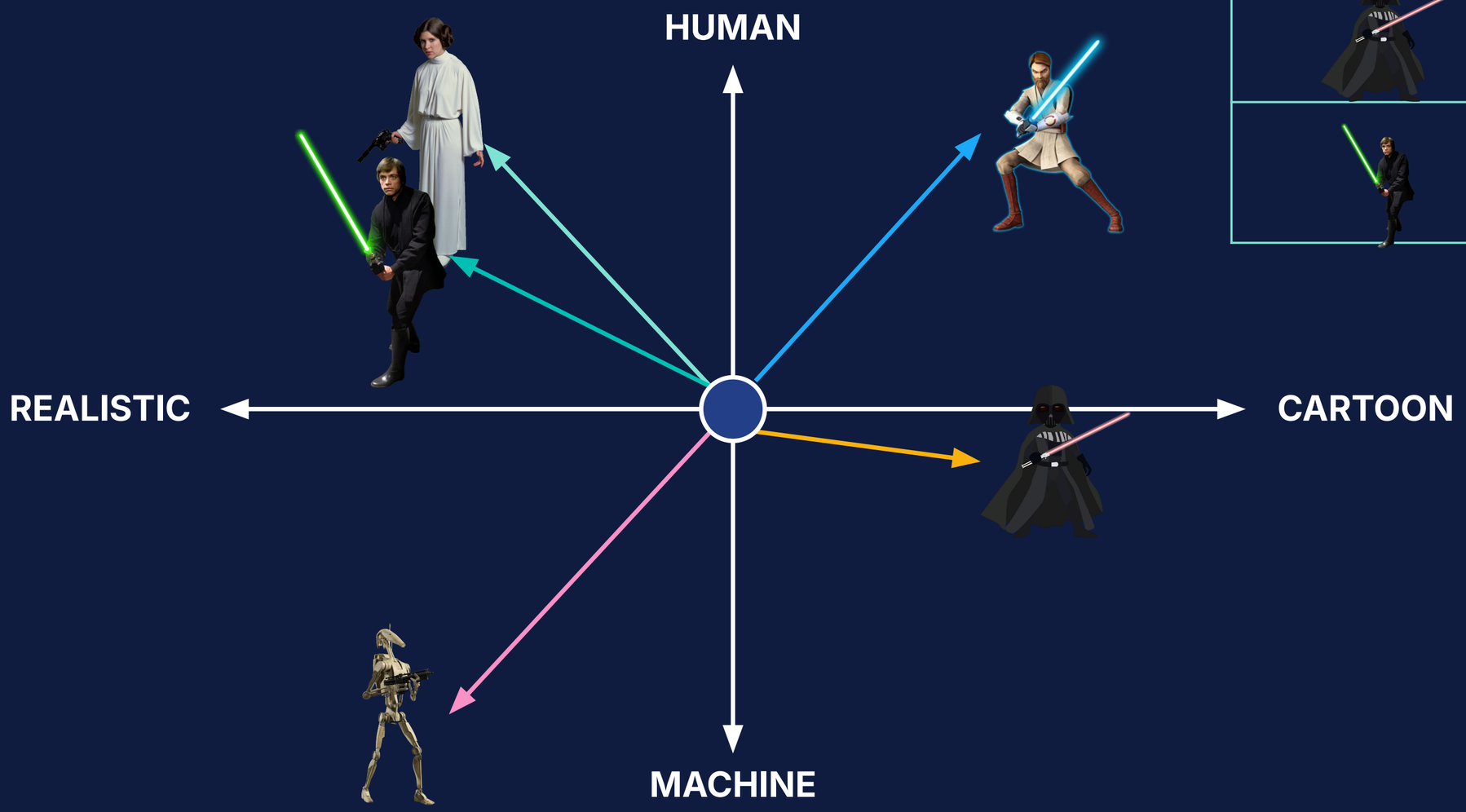
Multiple dimensions represent different data aspects



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

elastic

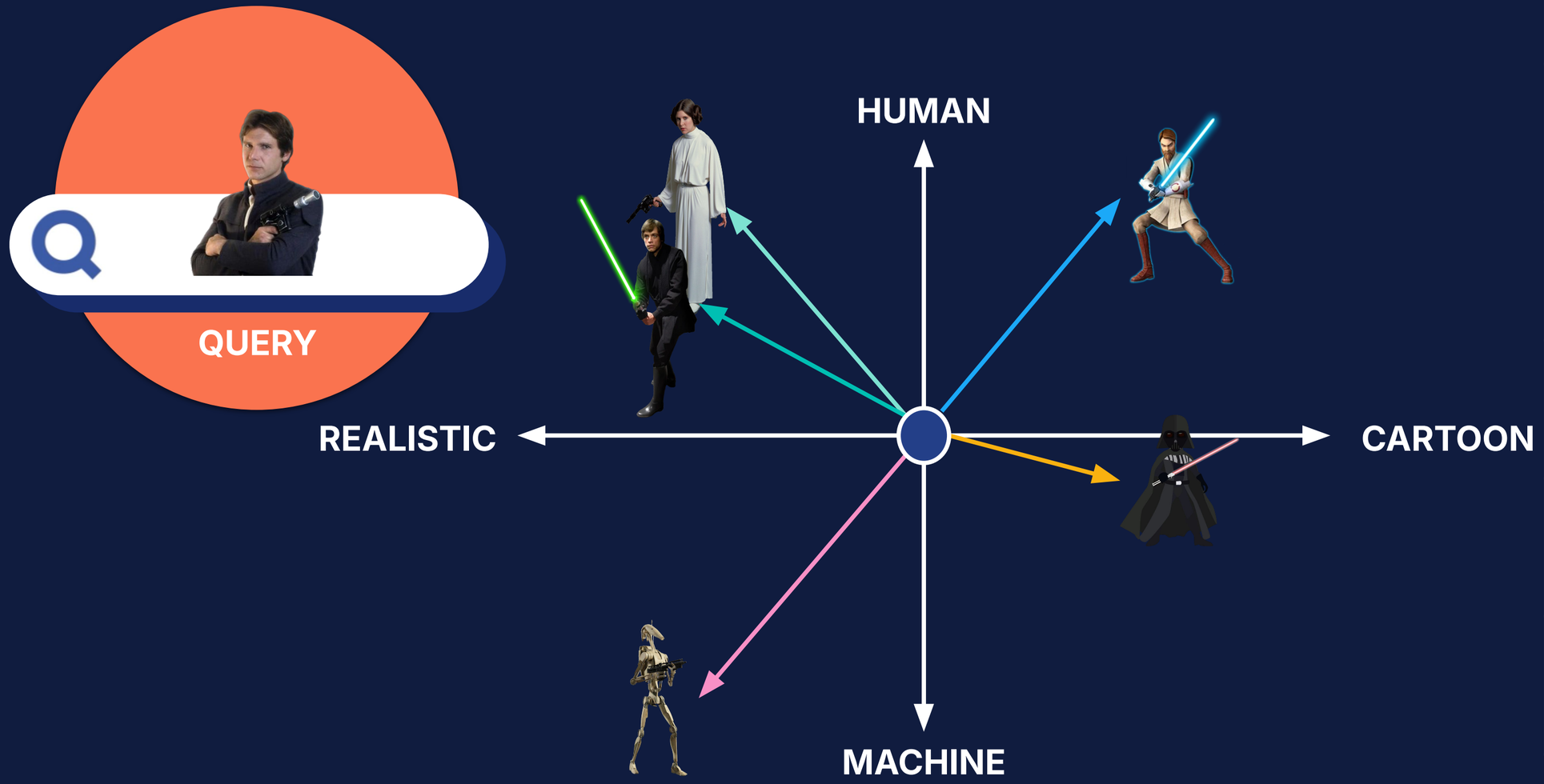
Similar data is grouped together



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

elastic

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



Relevance	Result
Query	
1	
2	
3	
4	
5	

elastic

How to select a model?

elastic

Choice of Embedding Model

Start with Off-the Shelf Models

- Text data: Hugging Face (like Microsoft’s E5)
- Images: OpenAI’s CLIP

Extend to Higher Relevance

- Apply hybrid scoring
- Bring Your Own Model: requires expertise + labeled data

What is chunking?

Chunking refers to the process of breaking content up.

elastic

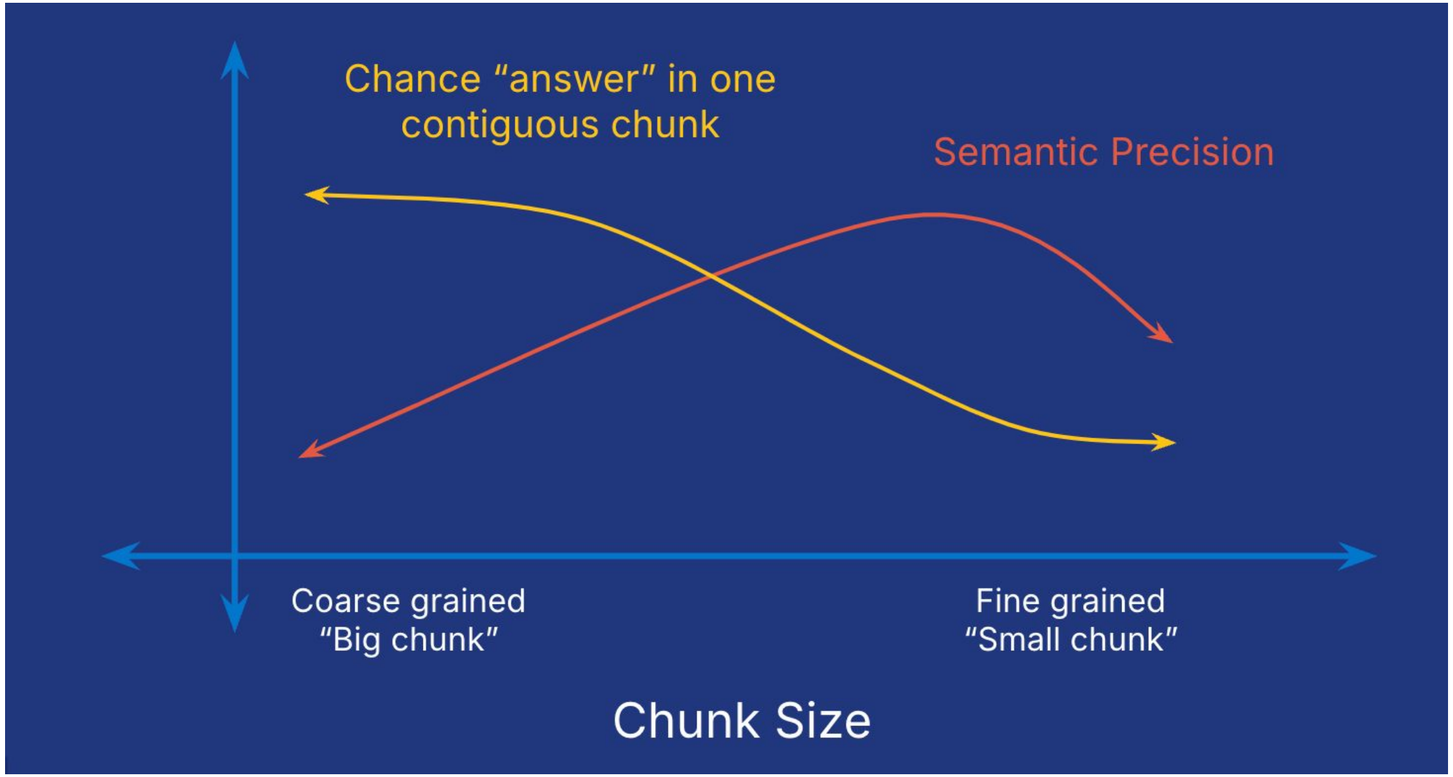
It is necessary for representing complex texts in vector embeddings.

elastic

```
helpers.bulk(es, actions, pipeline="text-embedding", chunk_size=1000)
```

elastic

Impact of chunk size



elastic

How and when to run a hybrid search?

elastic

Demo

elastic

Demo

https://www.elastic.co/search-labs/blog/elastic-vector-database-practical-example

elastic

Let’s take a look at the book data

elastic

Pipeline

elastic

```
import json
from getpass import getpass

from elasticsearch import Elasticsearch, helpers
from sentence_transformers import SentenceTransformer
```

elastic

```
cloud_endpoint = getpass("Elastic deployment Cloud Endpoint: ")
cloud_api_key = getpass("Elastic deployment API Key: ")
INDEX_NAME = "books-pipeline"

es = Elasticsearch(
    hosts=cloud_endpoint,
    api_key=cloud_api_key,
)
```

elastic

```
resp = es.ingest.put_pipeline(
    id="text-embedding",
    description="converts book description text to a vector",
    processors=[
        {
            "inference": {
                "model_id": "sentence-transformers__msmarco-minilm-l-12-v3",
                "input_output": [
                    {
                        "input_field": "book_description",
                        "output_field": "description_embedding",
                    }
                ],
            }
        }
    ],
    on_failure=[
        {
            "set": {
                "description": "Index document to 'failed-<index>'",
                "field": "_index",
                "value": "failed-{{{_index}}}",
            }
        },
        {
            "set": {
                "description": "Set error message",
                "field": "ingest.failure",
                "value": "{{_ingest.on_failure_message}}",
            }
        },
    ],
)

print(resp)
```

elastic

```
mappings = {
    "mappings": {
        "properties": {
            "book_title": {"type": "text"},
            "author_name": {"type": "text"},
            "rating_score": {"type": "float"},
            "rating_votes": {"type": "integer"},
            "review_number": {"type": "integer"},
            "book_description": {"type": "text"},
            "genres": {"type": "keyword"},
            "year_published": {"type": "integer"},
            "url": {"type": "text"},
        }
    }
}
# Delete any previous index
es.indices.delete(index=INDEX_NAME)
es.indices.create(index=INDEX_NAME, body=mappings)
print(f"Index '{INDEX_NAME}' created.")
```

elastic

```python
file_path = "../data/books.json"
with open(file_path, "r") as file:
    books = json.load(file)

# create an array of index actions, with each element holding one document
actions = [
    {"_index": INDEX_NAME, "_id": book.get("id", None), "_source": book}
    for book in books
]

try:
    helpers.bulk(es, actions, pipeline="text-embedding", chunk_size=1000)
    print(f"Successfully added {len(actions)} books into the '{INDEX_NAME}' index.")

except helpers.BulkIndexError as e:
    print(f"Error occurred while ingesting books: {e}")
    print(e.errors)
```

elastic

What is RAG?

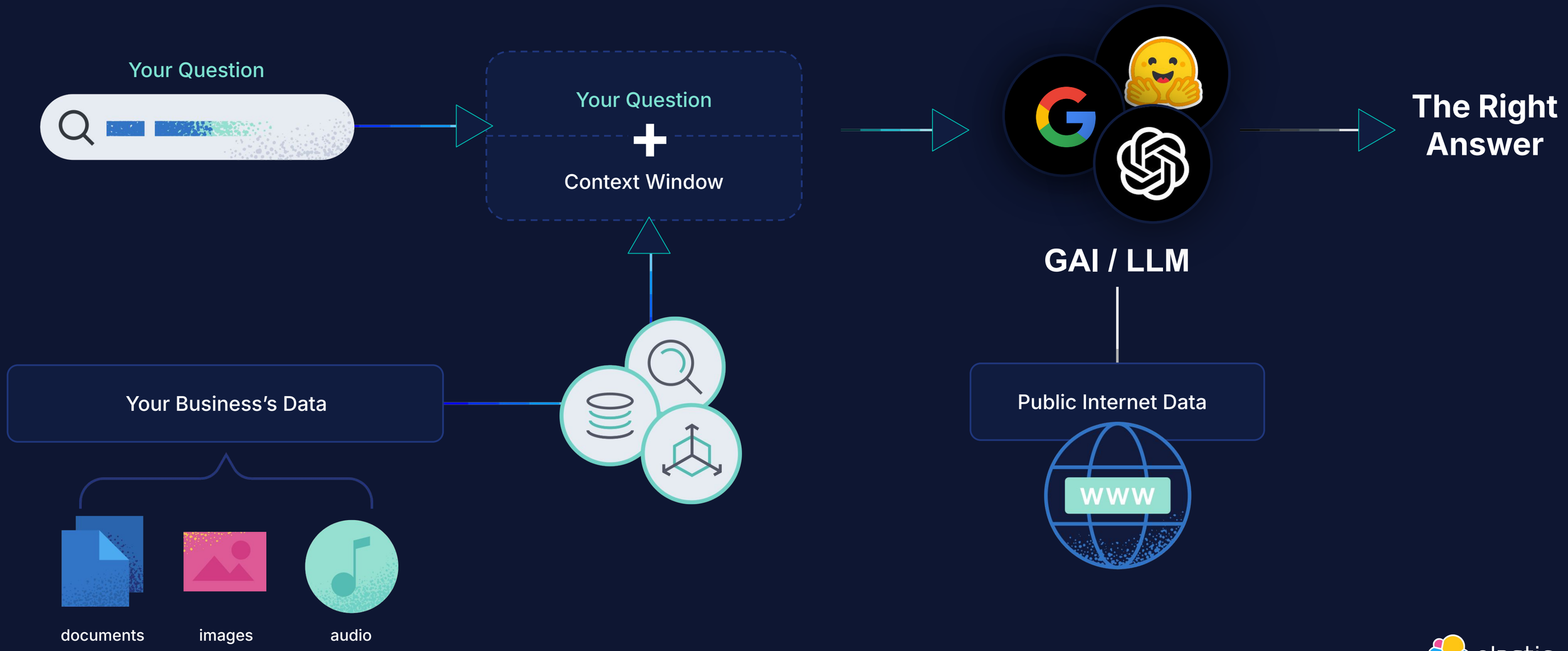
Retrieval-augmented generation (RAG) integrates external information retrieval into generating responses by Large Language Models (LLMs).

elastic

Vector Databases vs RAG

elastic

Retrieval Augmented Generation



elastic

Elastic's playground

A low-code interface for you to explore LLMs

elastic


es_client = elasticsearch.Elasticsearch(
getpass.getpass("Host: "),
api_key=getpass.getpass("API Key: "),
)

elastic

```
embedding = OpenAIEmbeddings(model="text-embedding-3-large")
elastic_vector_search = ElasticsearchStore(
    index_name="vr_tour_data",
    es_connection=es_client,
    embedding=embedding,
)
```

```
translated_texts = df['translated_text'].tolist()
combined_text = "\n".join(translated_texts)
text_splitter = CharacterTextSplitter(chunk_size=1100, chunk_overlap=0)
docs = text_splitter.split_documents([Document(page_content=combined_text)])
```

Next steps

elastic

RESOURCES FOR DEVELOPERS | BY DEVELOPERS LIKE YOU!

Search Labs

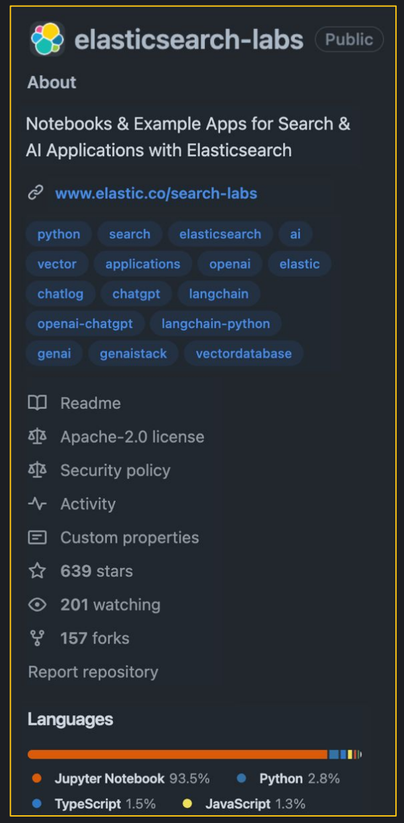
ML Research December 1, 2023

RAG evaluation metrics: A journey through metrics

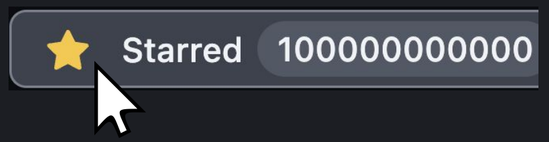
Explore RAG evaluation metrics like BLEU score, ROUGE score, PPL, BARTScore, and more. Discover...

By: **Quentin Herreros**, **Thomas Veasey** and **Thanos Papaoikonomou**

elastic.co/**search-labs**



github.com/elastic/**elasticsearch-labs**



Tutorials
GitHub Example Code
Integrations
Blogs

elastic

ElasticON is back!

ElasticON New York

Nov. 13, 2024

Manhattan Center

https://www.elastic.co/events/elasticon/new-york-city

elastic

Elastic Contributors

This community program is designed to recognize and reward the hard work of our awesome contributors. Join this friendly competition and earn points for:

Code contributions | Presentations | Video tutorials | Event organization
Translations | Technical Q&A | Written content | Content validation

The top contributors will win cool prizes such as **Elastic trainings**, **certificate exams, cloud credits** and more. Check out the rules on our website and start submitting today!

elastic.co/community/contributor

The current cycle runs from February 1st 2024 to January 31st 2025

elastic

Get involved

The community would love to hear from you!

If you have a cool use case to share at a meetup, please let us know! We would love to make that happen.

And if your company would like to host a meetup - we have user groups in other cities as well, please let us know.

Send an email to meetups@elastic.co or ully@elastic.co

Let me know if this talk inspires you to build anything. I'm @JessicaGarson on most platforms.

If you have any questions on our Discuss forums and the community Slack channel.

elastic


Thank you!
elastic