elastic

Search & AI:
a new era

David Pilato | @dadoonet

Agenda

- "Classic" search and its limitations
- ML model and usage
- Vector search or hybrid search in Elasticsearch
- OpenAI's ChatGPT or LLMs with Elasticsearch

elastic

Elasticsearch

You Know, for Search

elastic

Elasticsearch

Lucene

elastic

> These are not the droids you are looking for.

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

These are not 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
    }]}
```

Semantic
search
≠
Literal
matches

TODAY

X-wing starfighter squadron

TOMORROW

What ships and crews do I need to destroy an almost finished death star?
Or is there a secret weakness?

elastic

Elasticsearch

You Know, for Vector Search

elastic

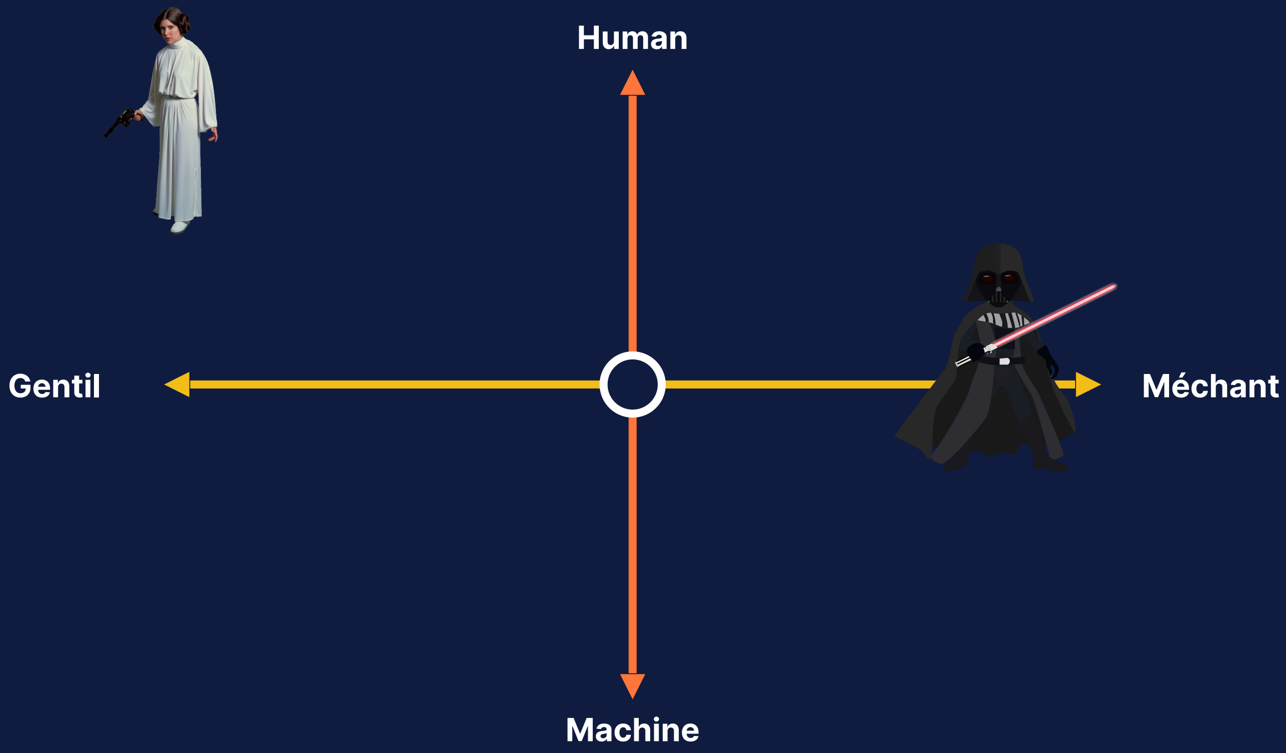
What is a **Vector**?

Embeddings represent your data

Example: 1-dimensional vector

Character	Vector
	[-1]
	[1]

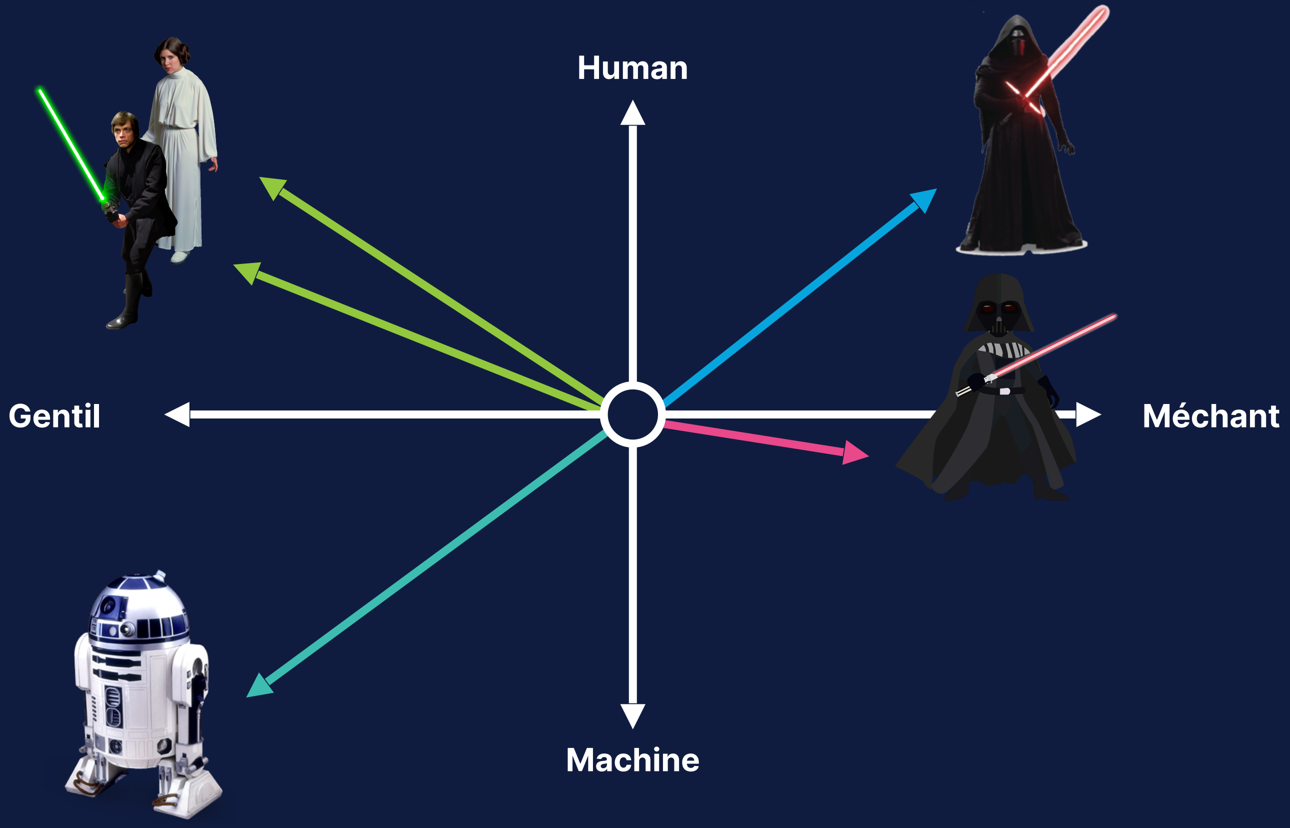
Multiple dimensions
represent different data aspects



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

elastic

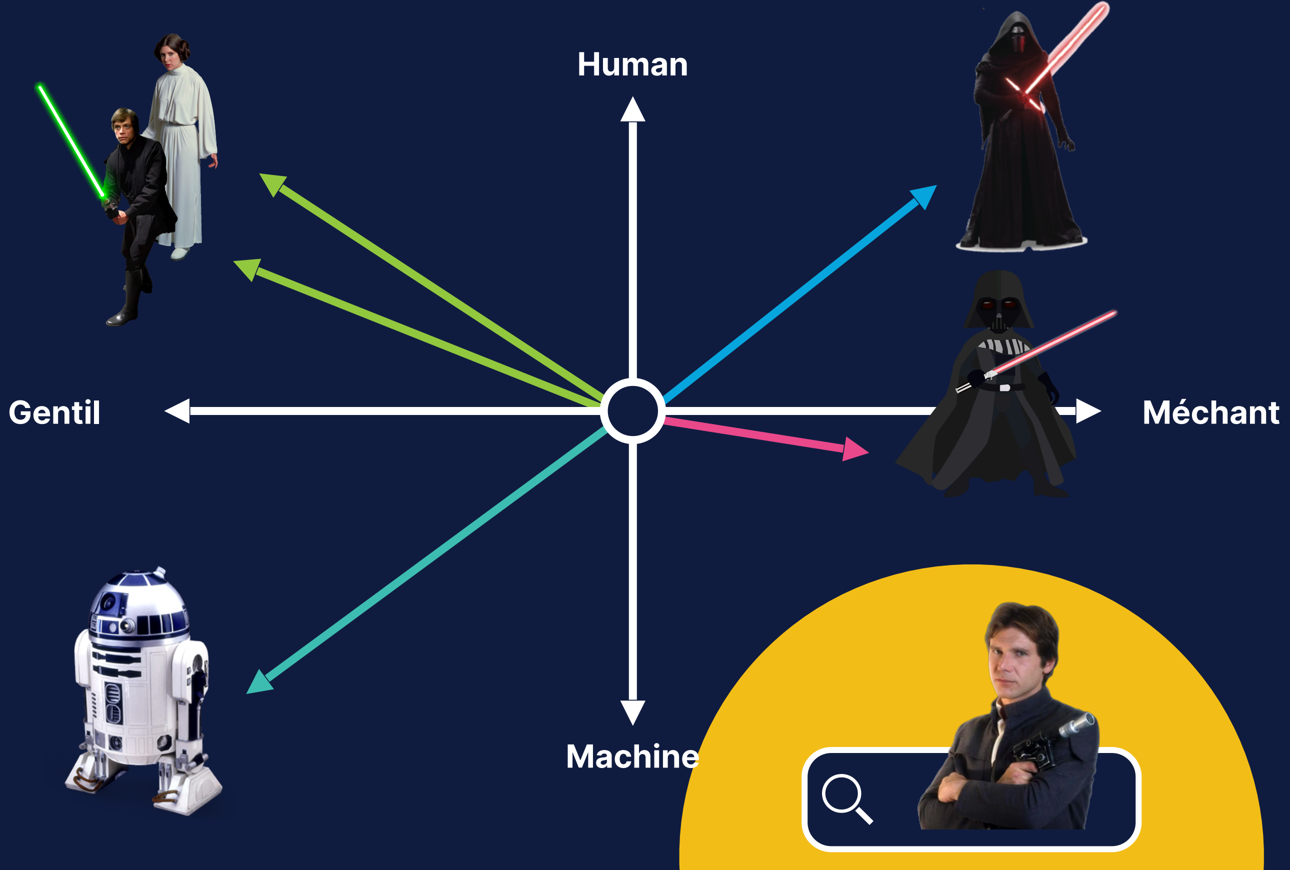
Similar data
is grouped together



Character	Vector
	[-1.0, 1.0]
	[1.0, 0.0]
	[-1.0, 0.8]
	[1.0, 1.0]
	[-1.0, -1.0]

elastic

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



Rank	Result
Query	
1	
2	
3	
4	
5	

Choice of Embedding Model

Start with Off-the Shelf Models

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

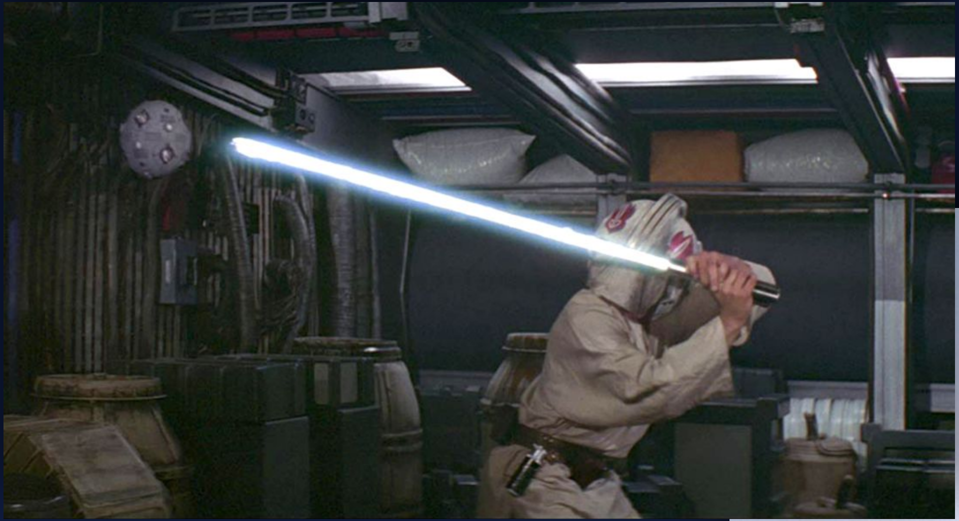
Extend to Higher Relevance

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

elastic

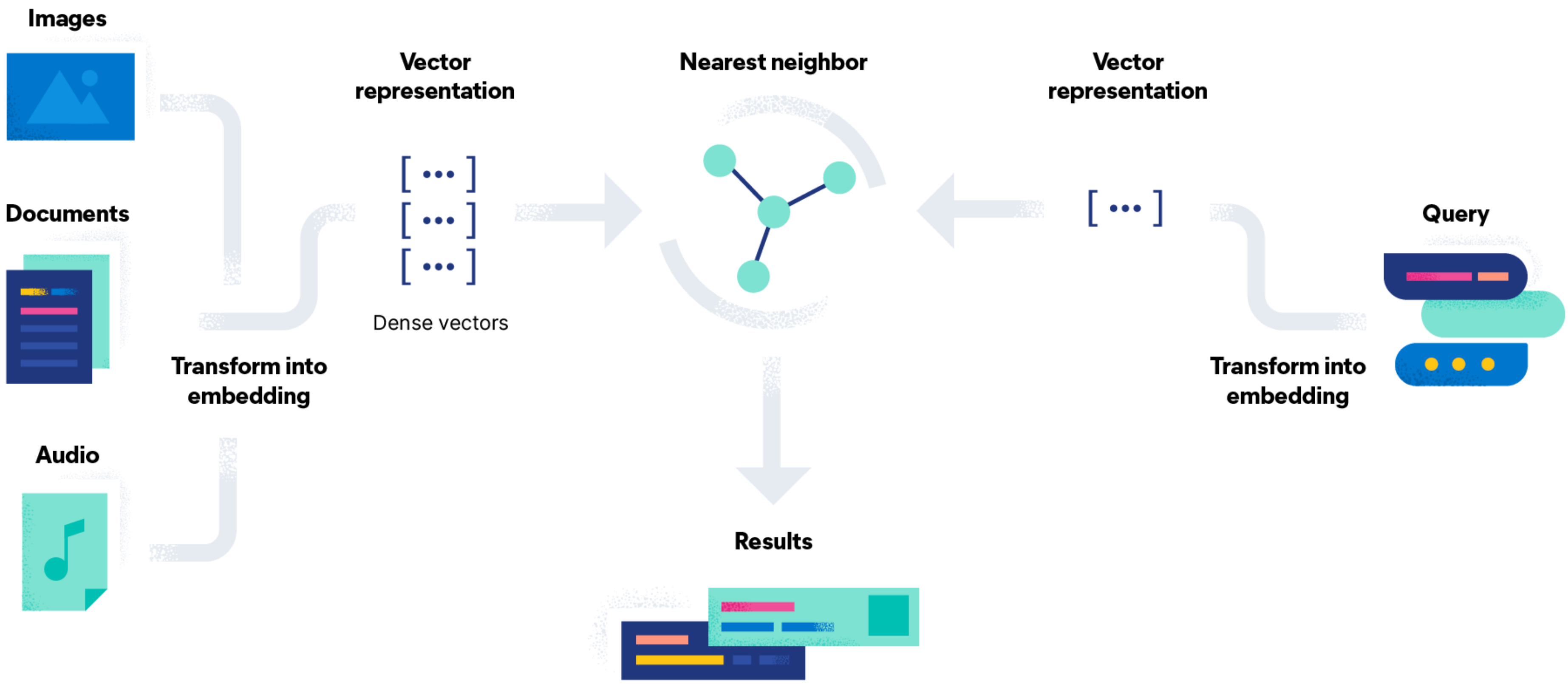
Problem

training vs actual use-case



elastic

Architecture of Vector Search

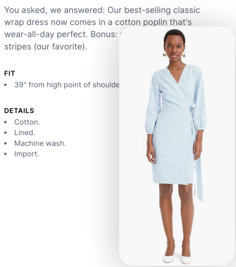


How do you

index vectors?

elastic

Data Ingestion and Embedding Generation



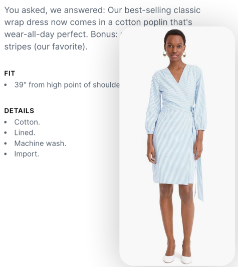
Source data

POST /_doc


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


elastic

Data Ingestion and Embedding Generation



Source data

PyTorch

python™


```
{
  "_id":"product-1234",
  "product_name":"Summer Dress",
  "description":"Our best-selling…",
  "Price": 118,
  "color":"blue",
  "fabric":"cotton",
  "desc_embedding":[0.452,0.3242,…],
  "img_embedding":[0.012,0.0,…]
}
```


POST /_doc

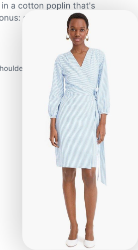
elastic

With Elastic ML

You asked, we answered: Our best-selling classic wrap dress now comes in a cotton poplin that's wear-all-day perfect. Bonus: stripes (our favorite).

FIT
- 39" from high point of shoulde

DETAILS
- Cotton.
- Lined.
- Machine wash.
- Import.



Source data

POST /_doc

ML Inference pipelines

Add inference pipeline

Inference pipelines will be run as processors from the Enterprise Search Ingest Pipeline

ml-inference-embedding-generation Actions

Deployed | pytorch | text_embedding

ml-inference-emational-analysis Actions

Deployed | pytorch | text_classification

Learn more about deploying ML models in Elastic

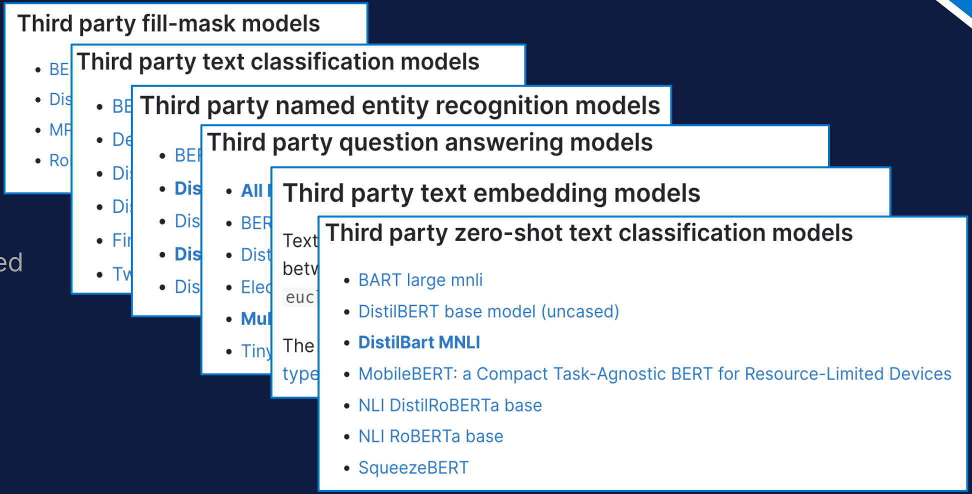

```
{
  "_id":"product-1234",
  "product_name":"Summer Dress",
  "description":"Our best-selling…",
  "Price": 118,
  "color":"blue",
  "fabric":"cotton",
  "desc_embedding":[0.452,0.3242,…]
}
```


elastic

Commercial

Elastic’s range of supported NLP models

- **Fill mask model**
  Mask some of the words in a sentence and predict words that replace masks
- **Named entity recognition model**
  NLP method that extracts information from text
- **Text embedding model**
  Represent individual words as numerical vectors in a predefined vector space
- **Text classification model**
  Assign a set of predefined categories to open-ended text
- **Question answering model**
  Model that can answer questions given some or no context
- **Zero-shot text classification model**
  Model trained on a set of labeled examples, that is able to classify previously unseen examples



Full list at: ela.st/nlp-supported-models

How do you

search vectors?

elastic

Vector Query

summer clothes

PyTorch

python™

```
GET product-catalog/_search
{
 "query" : {
    "bool": {
      "must": [{
         "knn": {
           "field": "desc_embbeding",
           "num_candidates": 50,
           "query_vector": [0.123, 0.244,...]



         }
      }],
      "filter": {
        "term": {
          "department": "women"
        }
      }
    }
  },
  "size": 10
}
```

elastic

Vector Query

Commercial

summer clothes

Transformer model

PyTorch

```
GET product-catalog/_search
{
 "query" : {
    "bool": {
      "must": [{
         "knn": {
           "field": "desc_embbeding",
           "num_candidates": 50,
           "query_vector_builder": {
             "text_embedding": {
               "model_text": "summer clothes",
               "model_id": <text-embedding-model>
             }
           }
         }
      }],
      "filter": {
        "term": {
          "department": "women"
        }
      }
    }
  },
  "size": 10
}
```

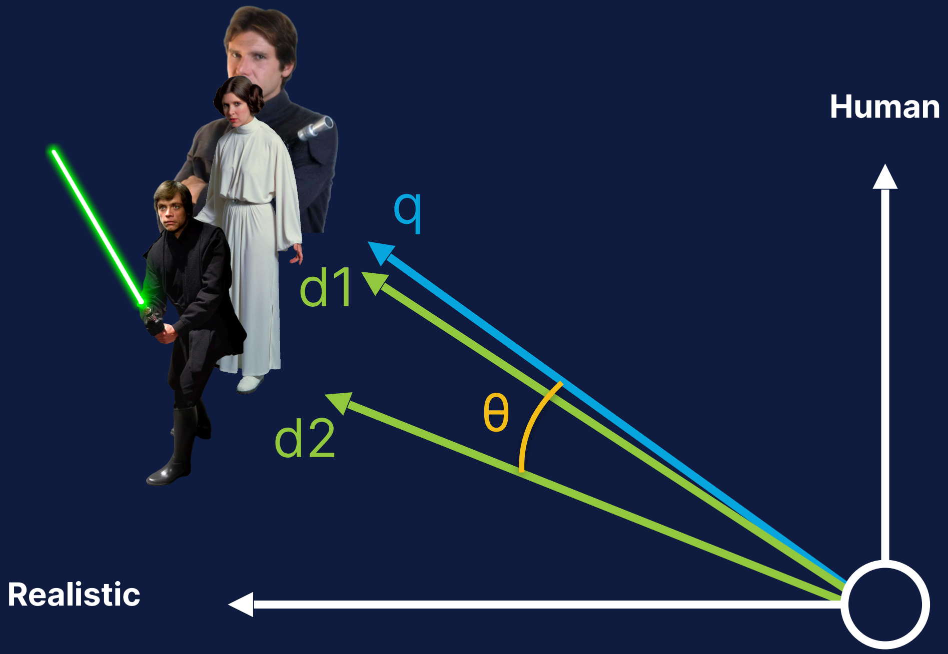
elastic

But how does it

really work?

elastic

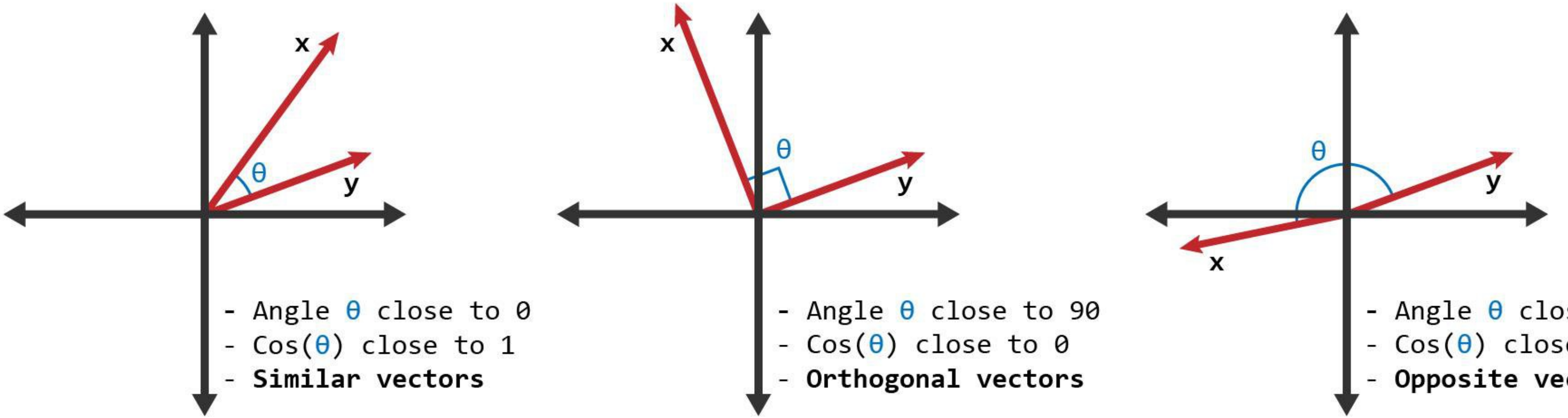
Similarity: cosine (cosine)



$$cos(\theta) = \frac{\vec{q} \times \vec{d}}{|\vec{q}| \times |\vec{d}|}$$

$$\_score = \frac{1 + cos(\theta)}{2}$$

Similarity: cosine (cosine)



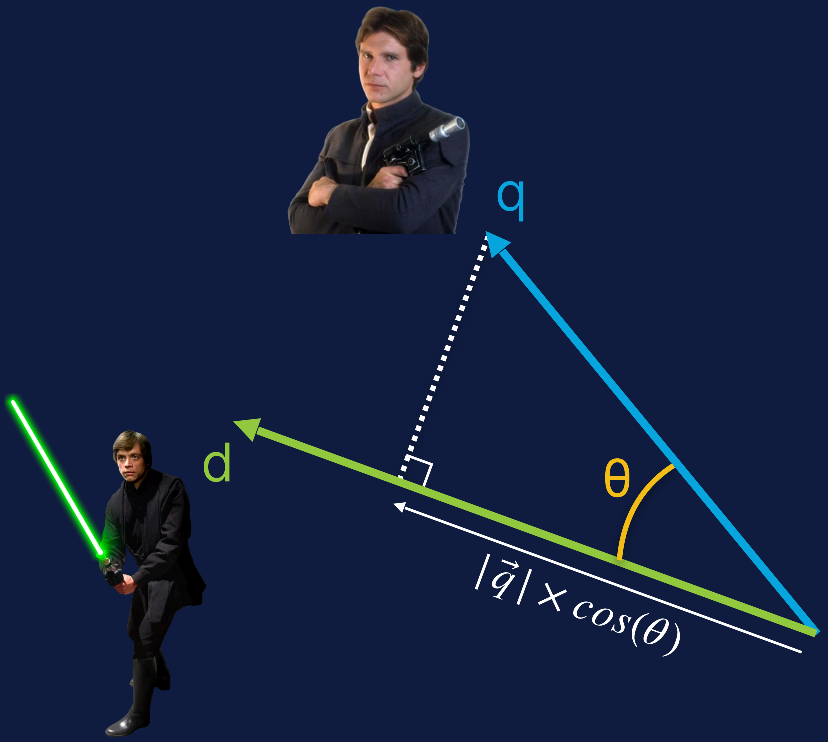
$$\_score = \frac{1+1}{2} = 1$$

$$\_score = \frac{1+0}{2} = 0.5$$

$$\_score = \frac{1-1}{2} = 0$$

elastic

Similarity: Dot Product (dot_product)



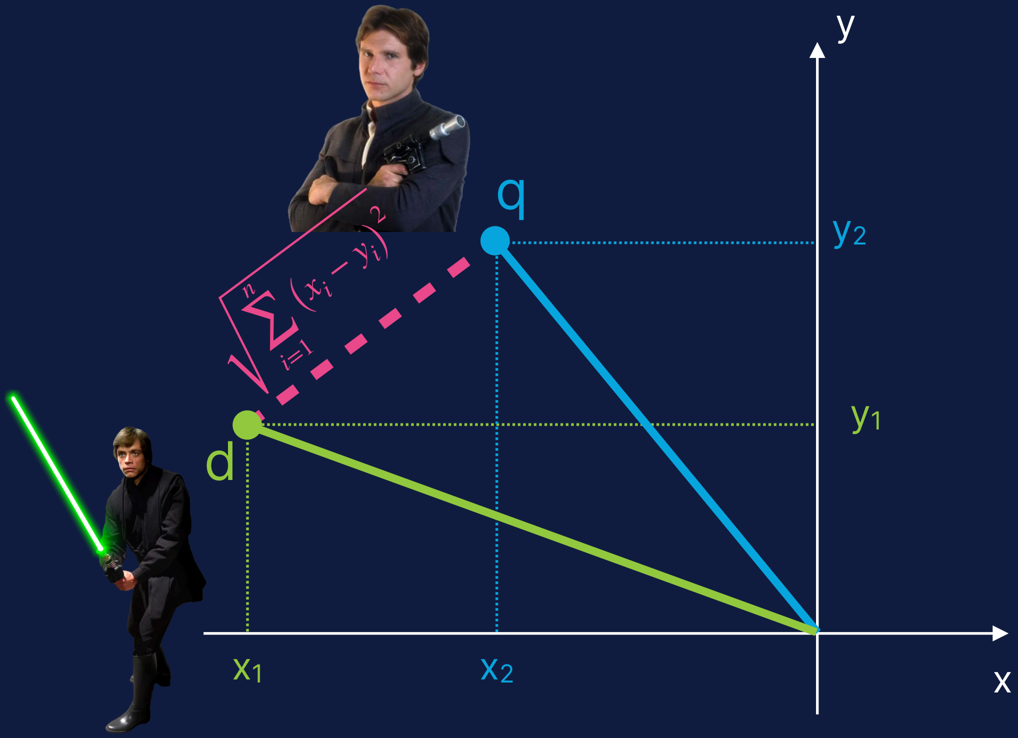
$$\vec{q} \times \vec{d} = |\vec{q}| \times cos(\theta) \times |\vec{d}|$$

$$\_score_{float} = \frac{1 + dot\_product(q, d)}{2}$$

$$\_score_{byte} = \frac{0.5 + dot\_product(q, d)}{32768 \times dims}$$

elastic

Similarity: Euclidean distance (l2_norm)



$$l2\_norm_{q,d} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

$$\_score = \frac{1}{1 + (l2\_norm_{q,d})^2}$$

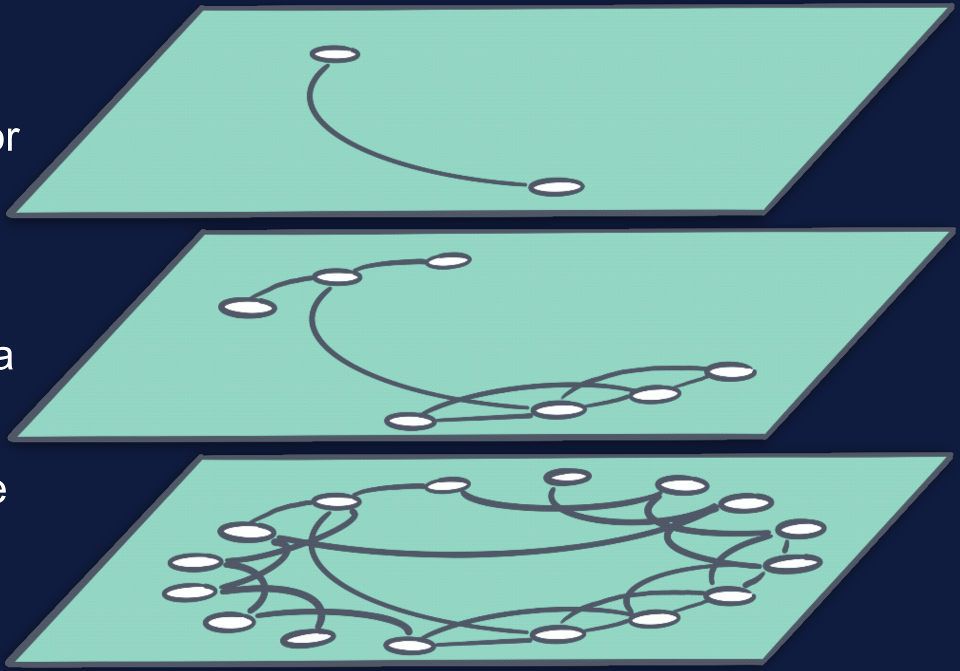
Brute Force

elastic

Hierarchical Navigable Small Worlds (HNSW)

One popular approach

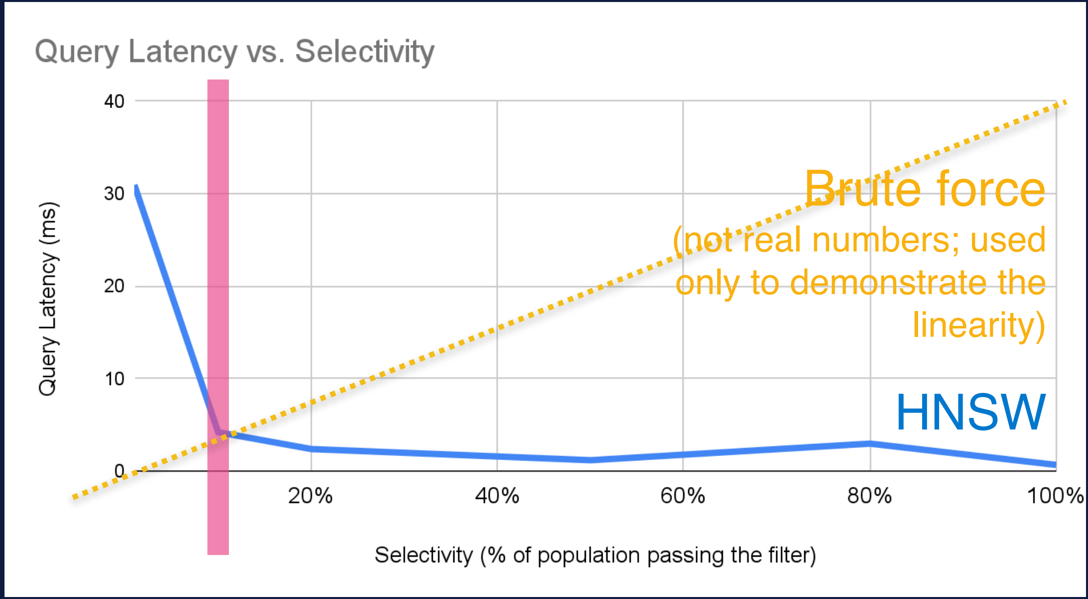
- **HNSW:** a layered approach that simplifies access to the nearest neighbor
- **Tiered:** from coarse to fine approximation over a few steps
- **Balance:** Bartering a little accuracy for a lot of scalability
- **Speed:** Excellent query latency on large scale indices



elastic

Filtering KNN Vector Similarity

Automatically choose between brute force and HNSW

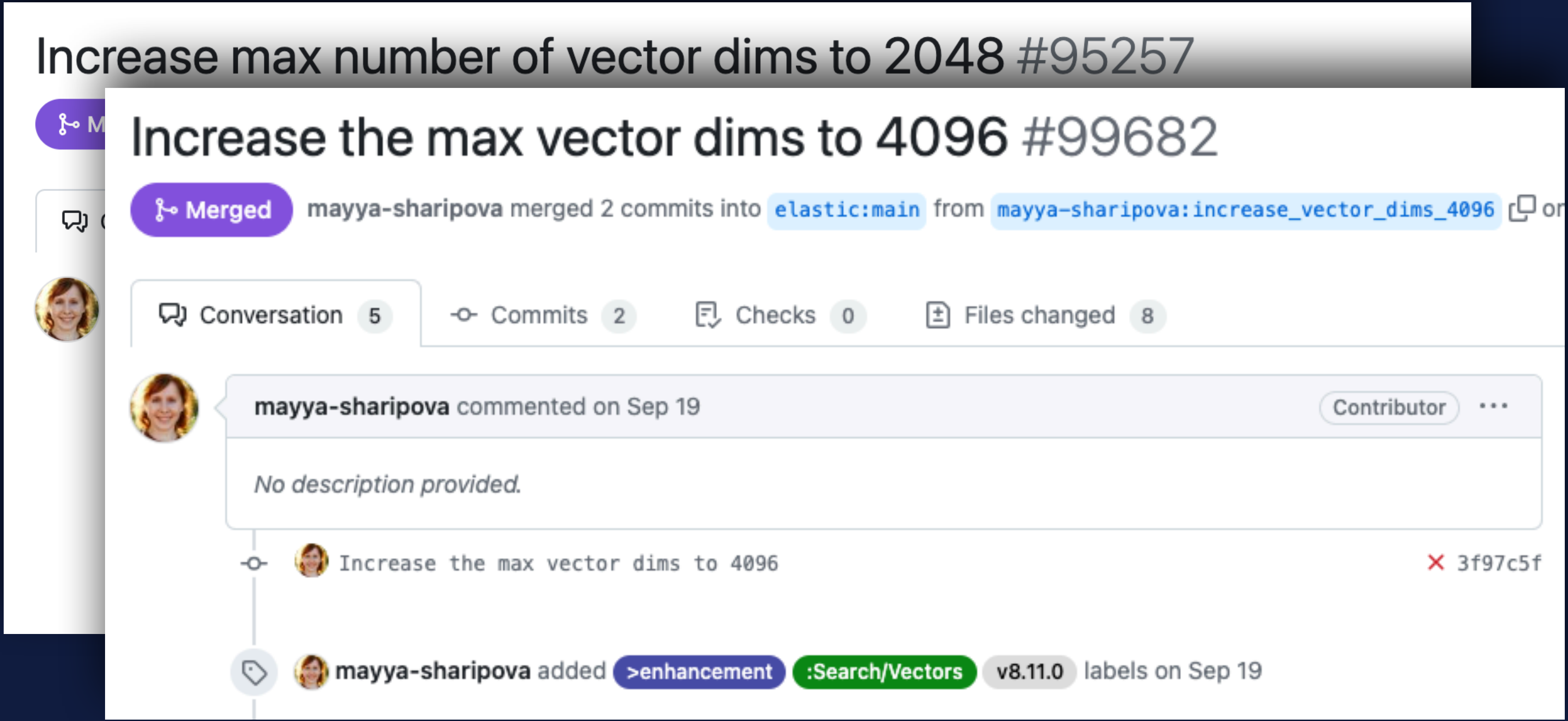


Bound worst case to 2*(brute force)

- Brute force scales O(n) of filtered
- HNSW scales ~O(log(n)) of all docs

elastic

Elasticsearch + Lucene = fast progress ❤️



Scaling Vector Search

Vector search

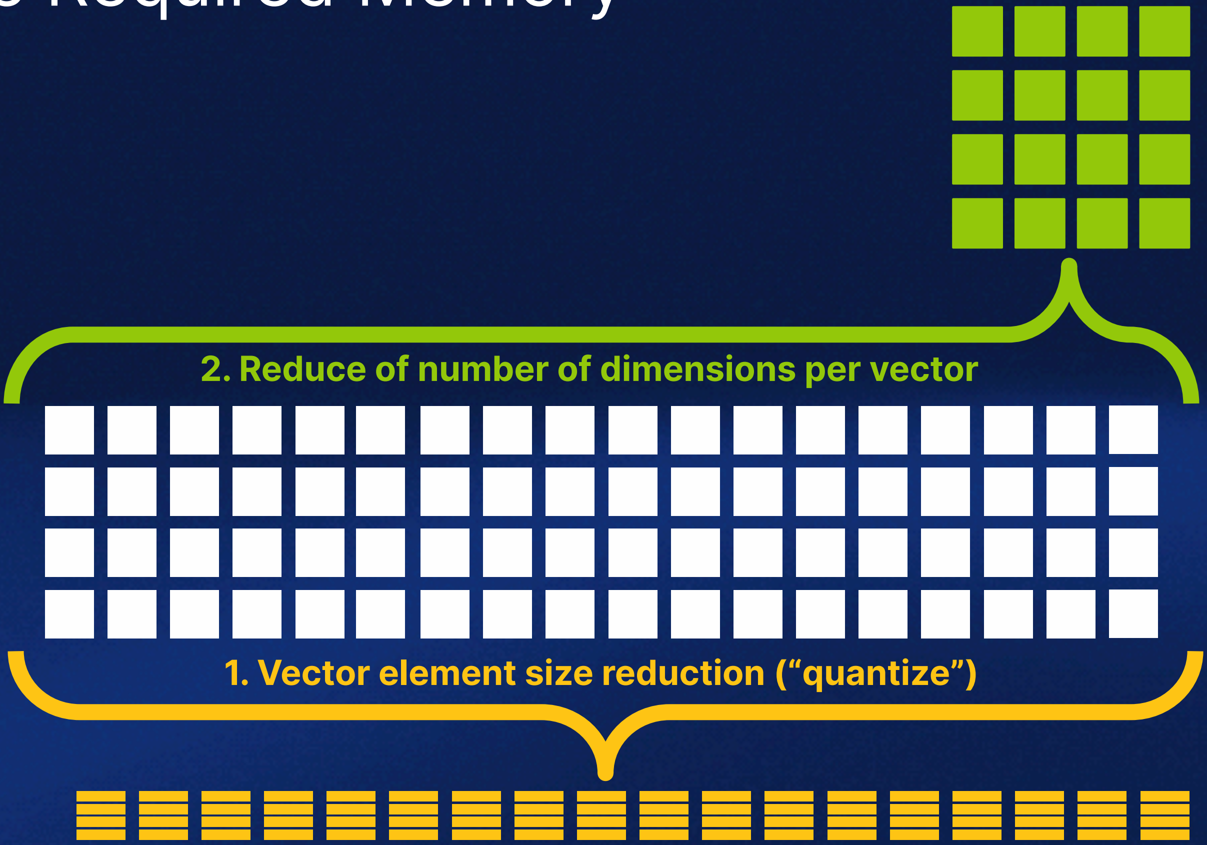
1. Needs lots of memory
2. Indexing is slower
3. Merging is slow

* Continuous improvements in Lucene + Elasticsearch

Best practices

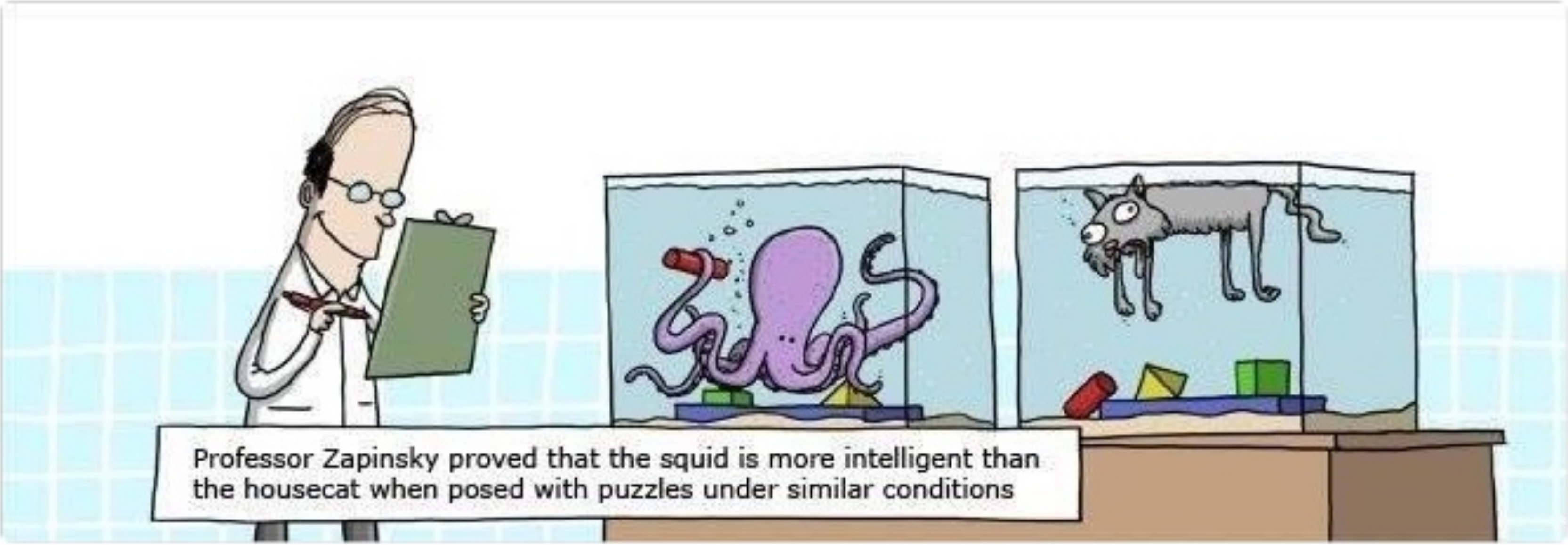
1. Avoid searches during indexing
2. Exclude vectors from _source
3. Reduce vector dimensionality
4. Use byte rather than float

Reduce Required Memory



elastic

Benchmarketing



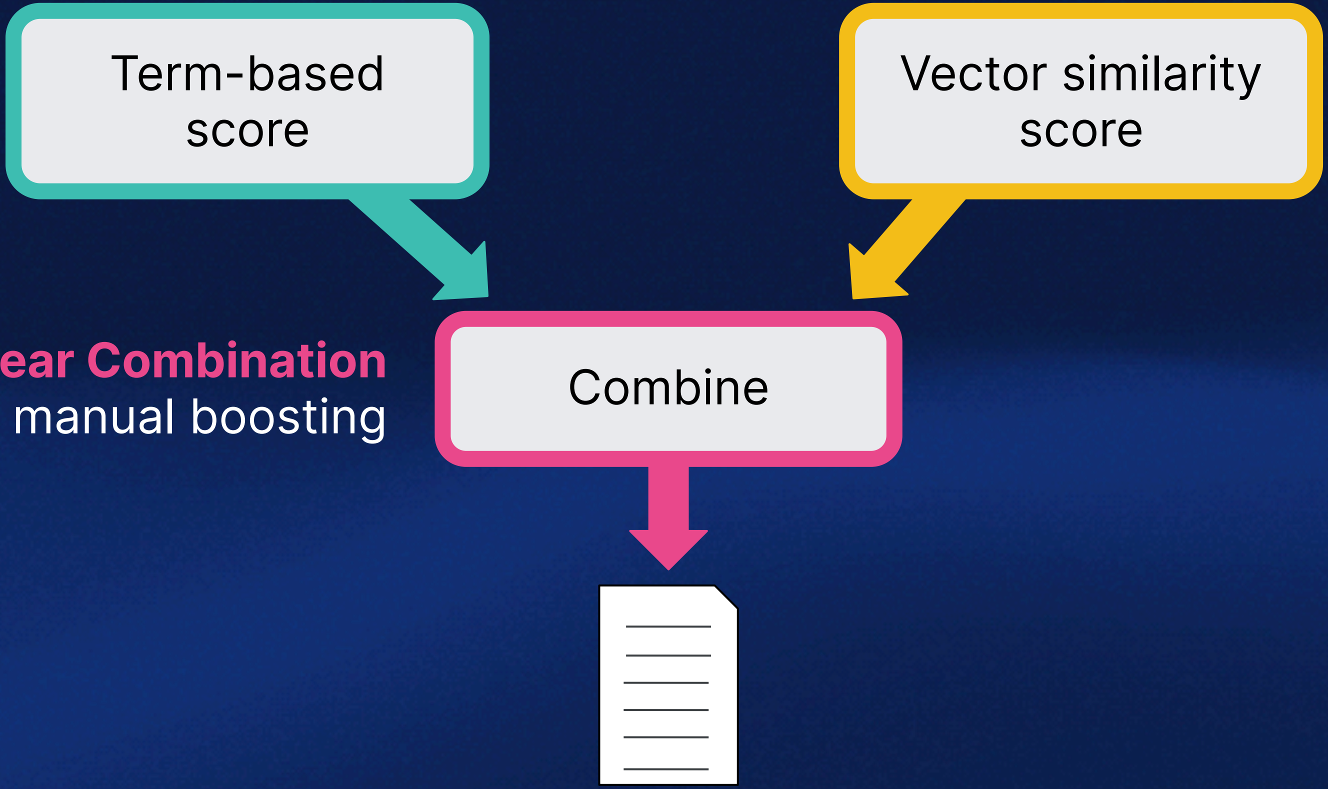
elastic

Elasticsearch

You Know, for **Hybrid** Search

elastic

Hybrid scoring



Linear Combination
manual boosting

```
GET product-catalog/_search
{
  "query" : {
    "bool" : {
      "must" : [{
        "match": {
          "description": {
            "query": "summer clothes",
            "boost": 0.9
          }
        }
      },{
        "knn": {
          "field": "desc_embbeding",
          "query_vector": [0.123, 0.244,...],
          "num_candidates": 50,
          "boost": 0.1,
          "filter": {
            "term": {
              "department": "women"
            }
          }
        }
      }],
      "filter" : {
        "range" : { "price": { "lte": 30 } }
      }
    }
  }
}
```

summer clothes

pre-filter

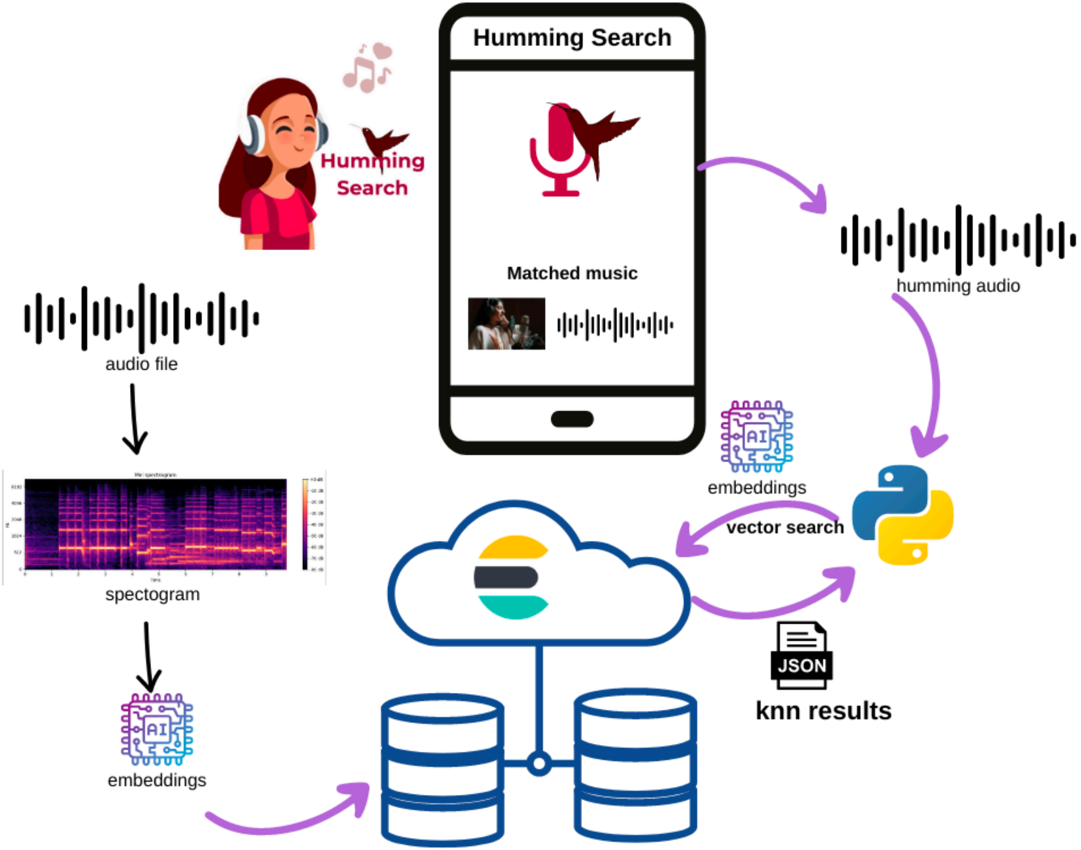
post-filter

elastic

```
GET product-catalog/_search
{
  "query" : {
    "bool" : {
      "must" : [{
        "match": {
          "description": {
            "query": "summer clothes",
            "boost": 0.9
          }
        }
      },{
        "knn": {
          "field": "image-vector",
          "query_vector": [54, 10, -2],
          "num_candidates": 50,
          "boost": 0.1
        }
      },{
        "knn": {
          "field": "title-vector",
          "query_vector": [1, 20, -52, 23, 10],
          "num_candidates": 10,
          "boost": 0.5
        }
      }]
    }
  }
}
```

https://djdadoo.pilato.fr/

elastic



https://github.com/dadoonet/music-search/

elastic

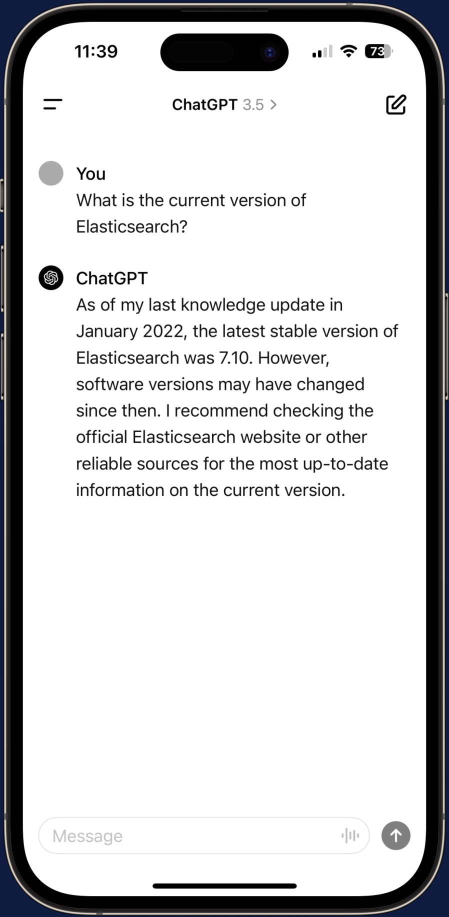
ChatGPT

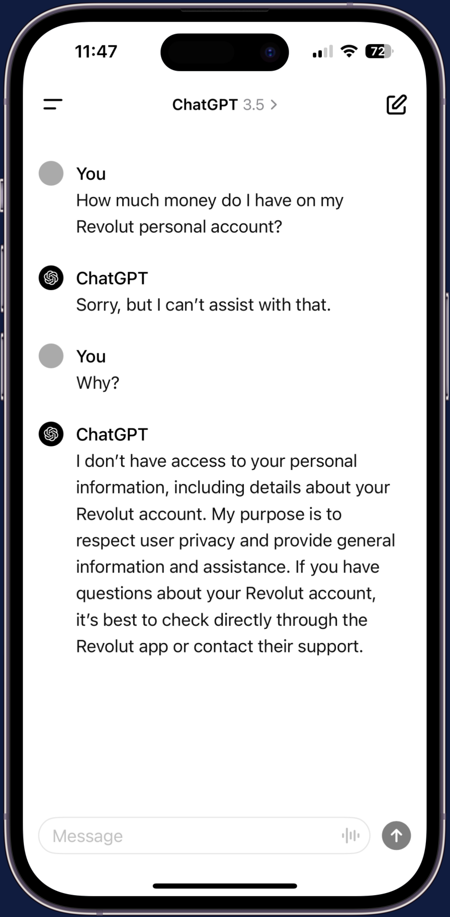
Elastic and LLM


Gen AI
Search engines

LLM: opportunities and limits

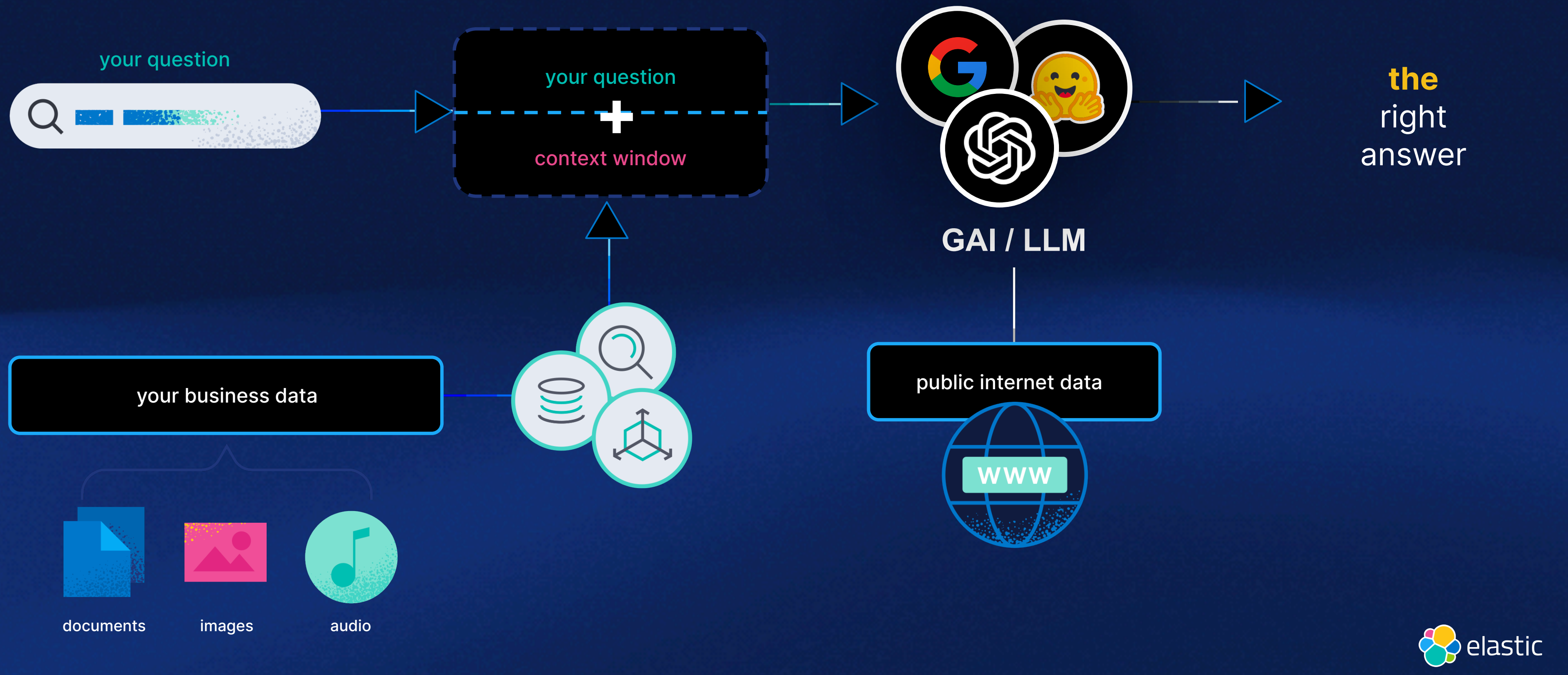
elastic


11:39
ChatGPT 3.5 >
You
What is the current version of Elasticsearch?
ChatGPT
As of my last knowledge update in January 2022, the latest stable version of Elasticsearch was 7.10. However, software versions may have changed since then. I recommend checking the official Elasticsearch website or other reliable sources for the most up-to-date information on the current version.
Message


11:47
ChatGPT 3.5 >
You
How much money do I have on my Revolut personal account?
ChatGPT
Sorry, but I can't assist with that.
You
Why?
ChatGPT
I don't have access to your personal information, including details about your Revolut account. My purpose is to respect user privacy and provide general information and assistance. If you have questions about your Revolut account, it's best to check directly through the Revolut app or contact their support.
Message

elastic

Retrieval Augmented Generation



11:47

ChatGPT 3.5 >

You
How much money do I have on my Revolut personal account?

ChatGPT
Sorry, but I can't assist with that.

You
Why?

ChatGPT
I don't have access to your personal information, including details about your Revolut account. My purpose is to respect user privacy and provide general information and assistance. If you have questions about your Revolut account, it's best to check directly through the Revolut app or contact their support.

Message

Home | Online banking | Enviroment setup

me | Transaction search | Financial summary | Customer support

Search your transactions:

This search is not enabled by Elastic and reflects the kind of functionality available to customers today.

Submit

Date	Account	Description	Value	Opening balance	Closing balance
18/06/24	EL03-130981-Transmission	Inbound payment made from EL03-130981-Transmission, St.james's Plac (STJ): 864dce1b-bb95-47d5-87dd-7d02f3b10c3f	7419.0	-825.0	6594.0
18/06/24	EL03-130981-Transmission	Purchase at merchant: Southeastern Grocers, LLC, location: Fayetteville,AR	82.0	6594.0	6512.0
18/06/24	EL03-130981-Transmission	Purchase at merchant: Müller Holding Ltd. & Co. KG, location: Glendale,AZ	188.0	6512.0	6324.0
17/06/24	EL03-130981-Transmission	Payment made from EL03-130981-Transmission to Elwood Erickson, Mitie Grp. (MTO): d37085fc-1382-4593-9cb8-26e5526bd9a0	533.0	20.0	-513.0
17/06/24	EL03-130981-Transmission	Payment made from EL03-130981-Transmission to Classie Johns, Barclays (BARC): 75b603a2-1c1b-45e9-a7ec-4a551bf98a8d	312.0	-513.0	-825.0
16/06/24	EL03-130981-Transmission	Purchase at merchant: E-MART Inc., location: Fayetteville,AR	31.0	51.0	20.0
14/06/24	EL03-130981-Transmission	Purchase at merchant: Dick's Sporting Goods, Inc., location: Montgomery,AL	182.0	329.0	147.0
14/06/24	EL03-130981-Transmission	Purchase at merchant: Valor Holdings Co., Ltd., location: Louisville,KY	96.0	147.0	51.0
13/06/24	EL03-130981-Transmission	Purchase at merchant: The Save Mart Companies, location:	34.0	363.0	329.0

elastic

Elasticsearch

You Know, for **Semantic** Search

elastic

Search & AI:
a new era

David Pilato | @dadoonet