

The background features a vibrant blue gradient with subtle, concentric wavy lines. In the bottom right corner, there are abstract, flowing shapes in shades of purple, pink, and orange, creating a modern and dynamic aesthetic.

# aws SUMMIT

INDIA | MAY 25, 2023

AIML003

# Innovations from Elasticsearch - Drive speed, scale & relevance

The latest innovations from Elasticsearch: How to implement recent developments in search to drive speed scale, and relevance

Ravindra Ramnani

Principal Solutions Architect  
Elastic



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.













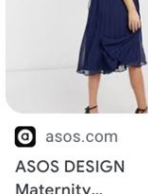
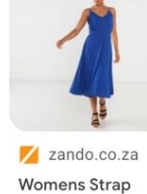


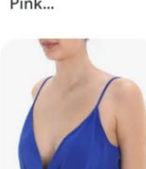
# Agenda

- Relevance innovations powered by vector search
  - The power of vector search
  - How to implement in Elastic
  - Best of both worlds with hybrid scoring
  - Improving search experiences with NLP and Personalization
  - Ingestion
- Performance improvements
- Elastic solutions & differentiators

# Want to copy your famous influencer?



Visual matches

 lyst.com Boohoo Strappy Pleated Midi... €35.00	 boohoo.com Blue Dresses   Royal, Light Blu...	 all4-gp.us Eve Mauve See through blue...	 lyst.co.uk Mango Polka-dot Pleated Dress...	 corporacion... düzenlemek Burger devirme...
 boohoo.com Strappy Pleated Midi Skater Dress ★★★★★ (1)	 torrid.com Plus Size - Midi Chiffon Pleated...	 shopstyle.co... ASOS Tall ASOS DESIGN Tall...	 rolypolyappa... V Neck Pleated Midi Sundress - ...	 zalando.ie Mango FORTUNY - Day dress - ...
 lulus.com Royal Blue Clothing Perfec...	 gethattend... Playful Promises Pink...	 asos.com ASOS DESIGN Maternity...	 zando.co.za Womens Strap Dress   Shop &...	 pinterest.com ASOS DESIGN pleated cami...
 freeshop.gr Wildwood	 shein.com SHEIN Solid			

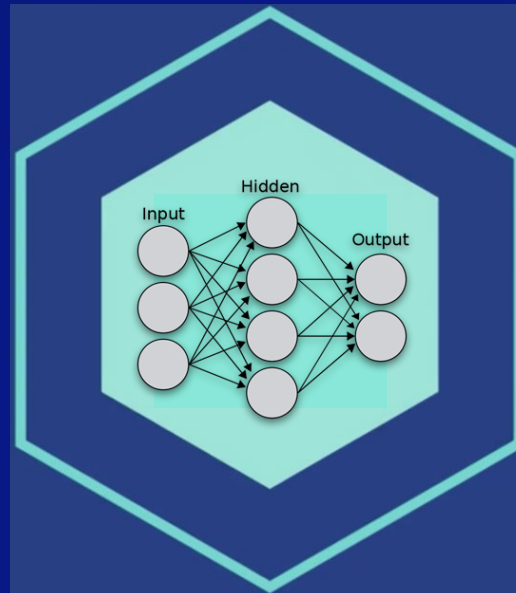
# Vector search lets you find what you mean, with higher relevance

- ✓ **Semantic search:** understand meaning
- ✓ **Multimedia:** not just text, also on pictures or sounds
- ✓ **Precision**

# Convert data into vector representation

## UNDERLYING APPROACH OF VECTOR SEARCH

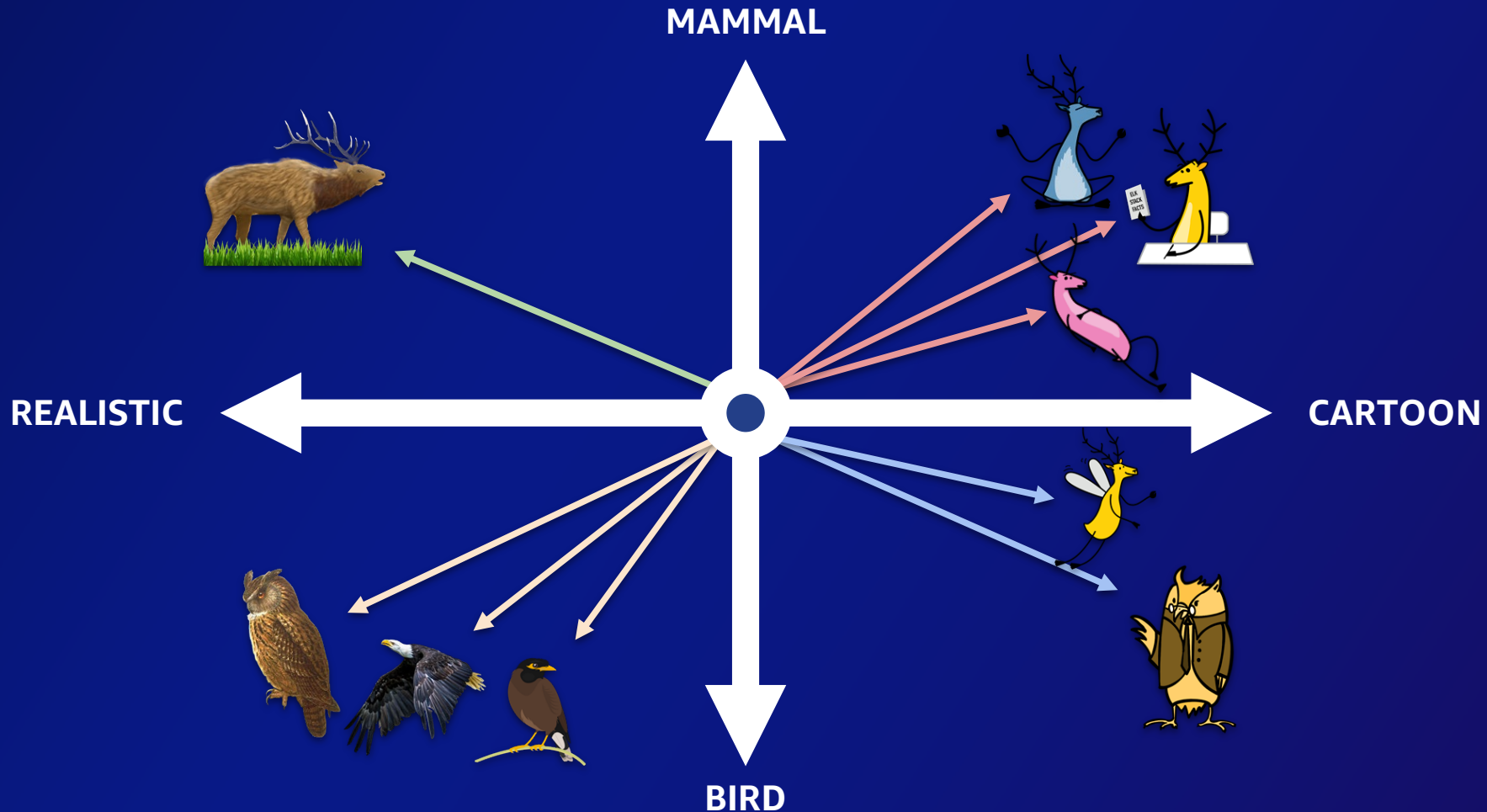
In order to stream from our service you will need a high quality connection. The required connection speed for using the service will vary depending on the quality of video and audio that you wish to stream to your device. For most customers we recommend at least...



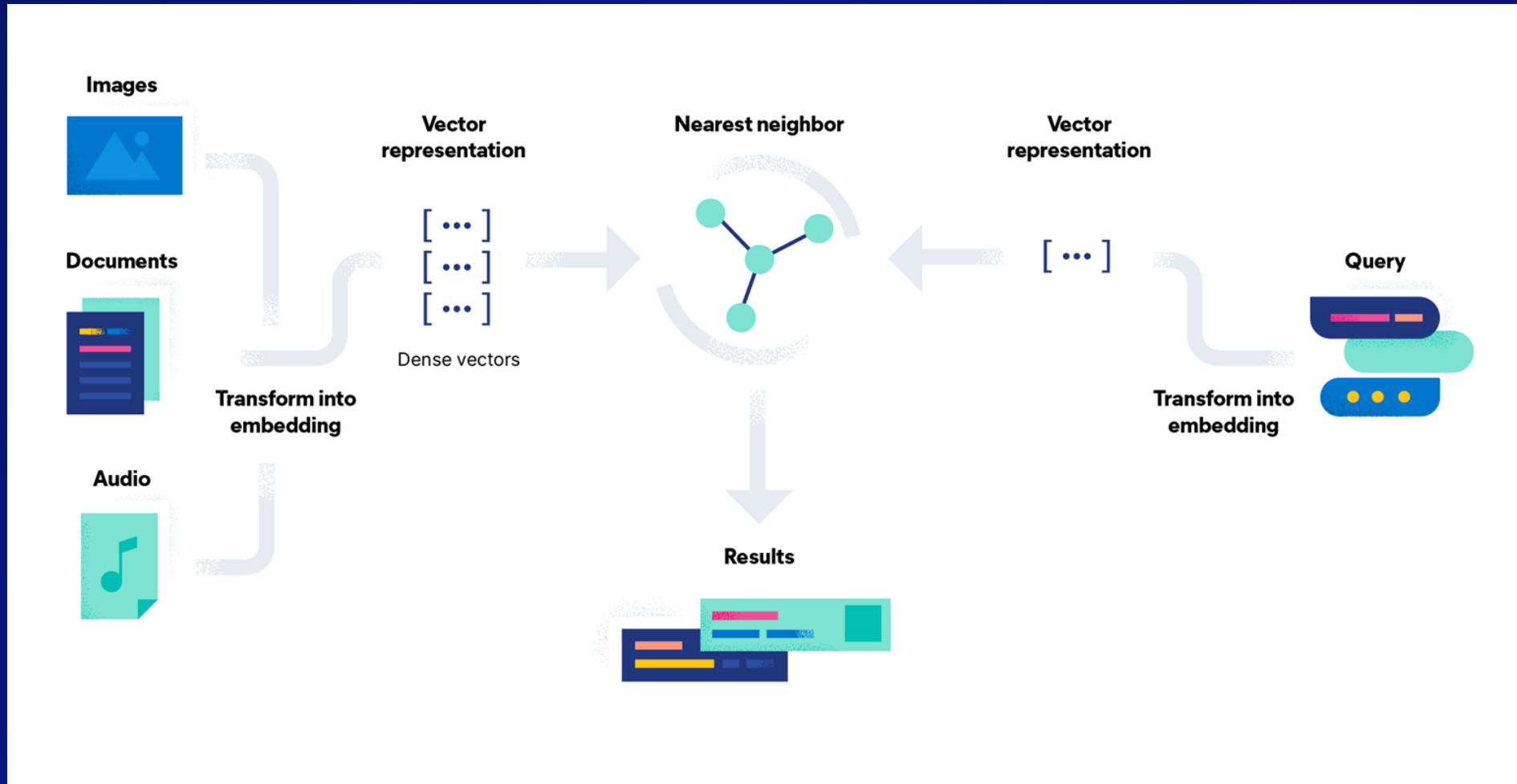
0.0167327...  
0.3458967...  
0.0547893...  
0.0324981...  
0.0135497...  
0.0216549...



# In vector space, similar data are nearby



# Vector search returns nearest neighbors of “vectorized” query





# Vector search and NLP may be intimidating:



**Lots of data, labelled**

**Expertise**

**Scale processing**

# Elastic makes vector search easy



Just bring the data

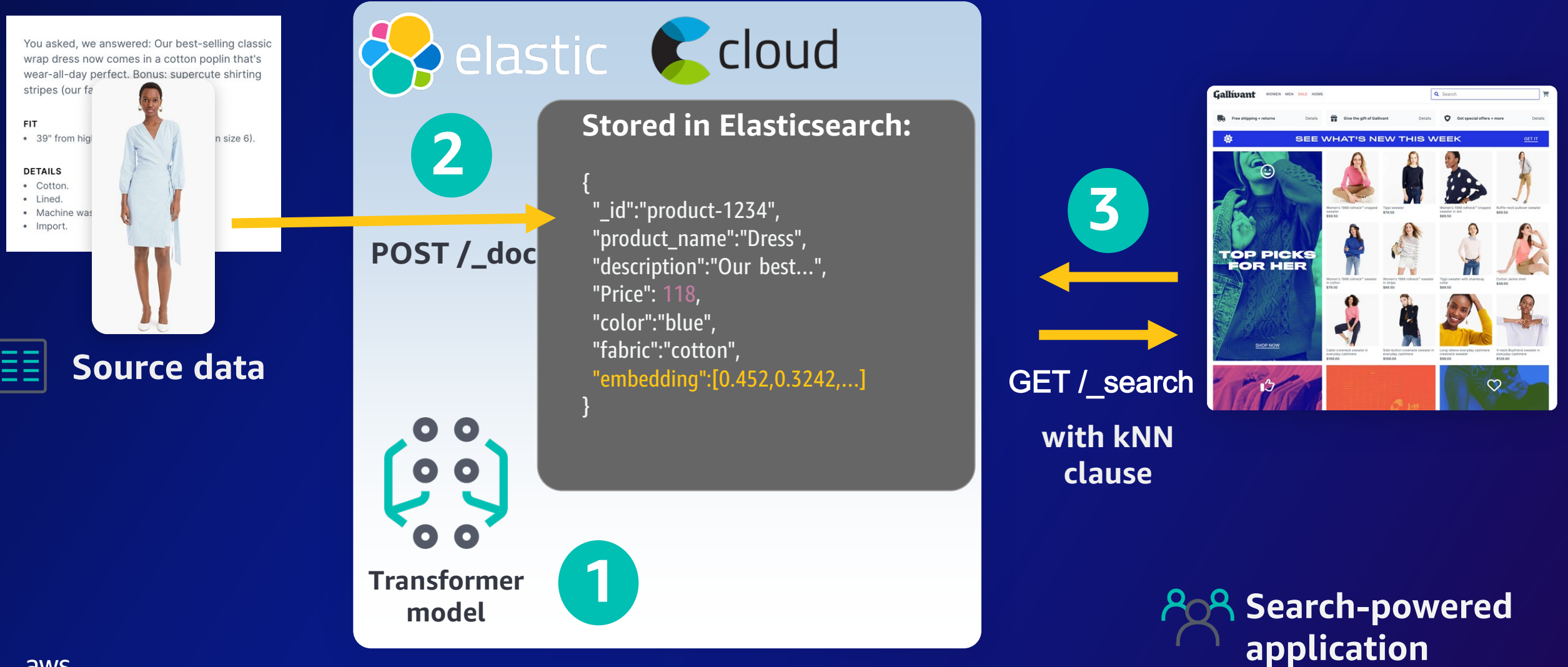


Inherit scalability

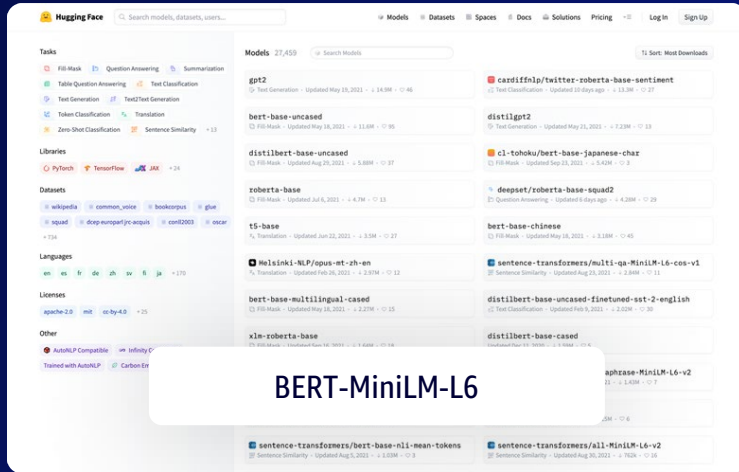


Flexibility

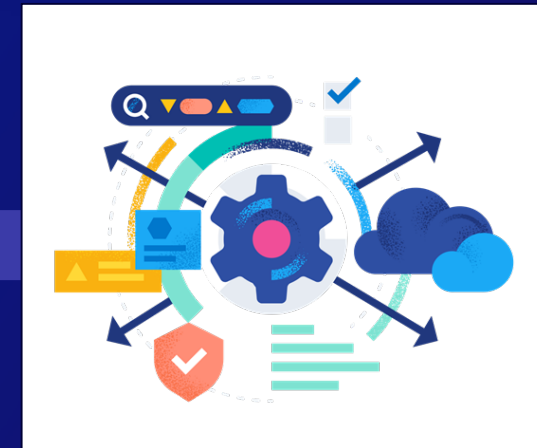
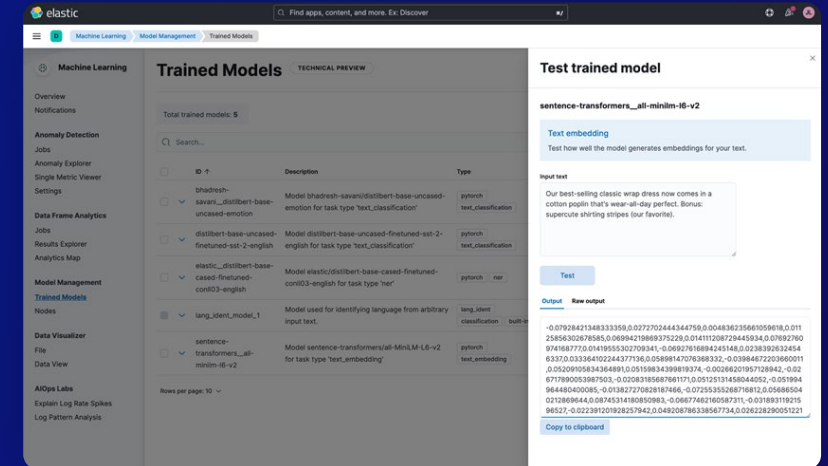
# Apply vector search in 3 steps



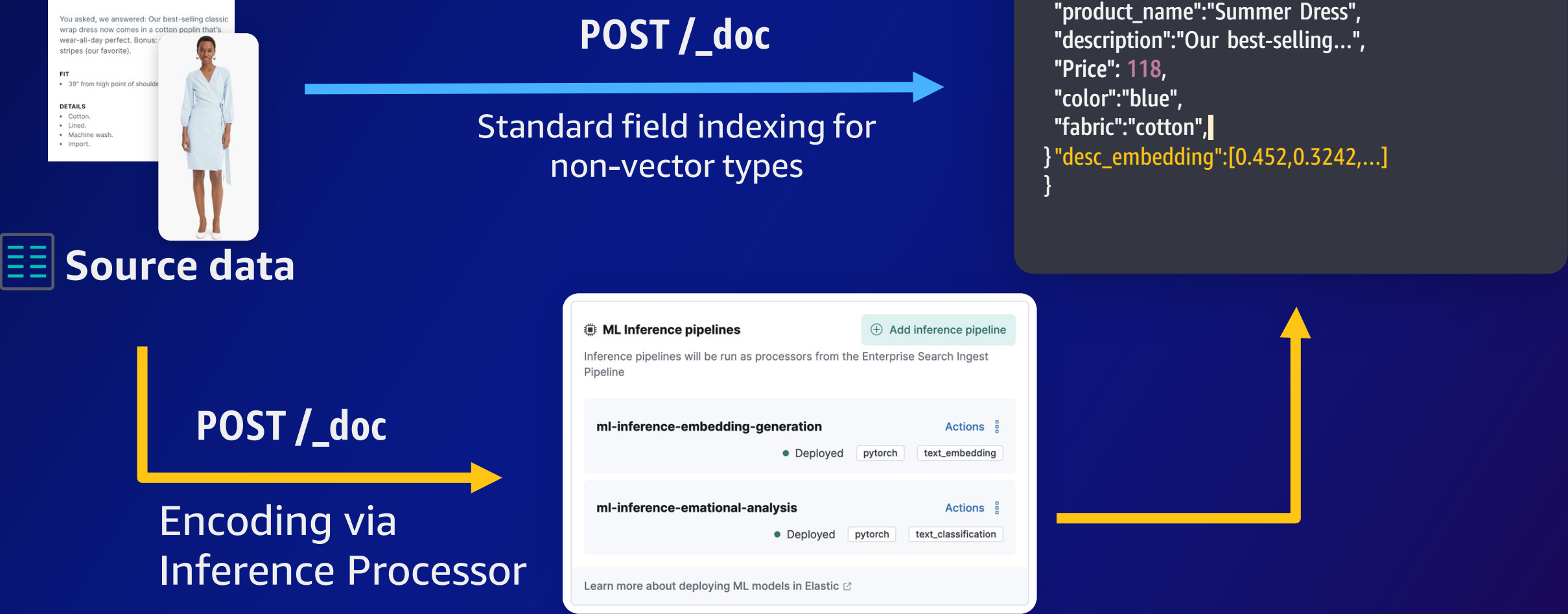
# Step 1: Import pre-trained model



```
$ eland_import_hub_model  
--url https://cluster_URL --hub-model-id  
BERT-MiniLM-L6  
--task-type text_embedding --start
```



# Step 2: Vectorize during data ingestion



# Step 3: Find nearest neighbor, approximately

Query is submitted to the search-powered application:

Generate embedding:

`POST /_ml/trained_models/my-model/_infer`

```
{
  "docs": {
    "description": "summer clothes"
  }
}
```



Transformer model

 PyTorch

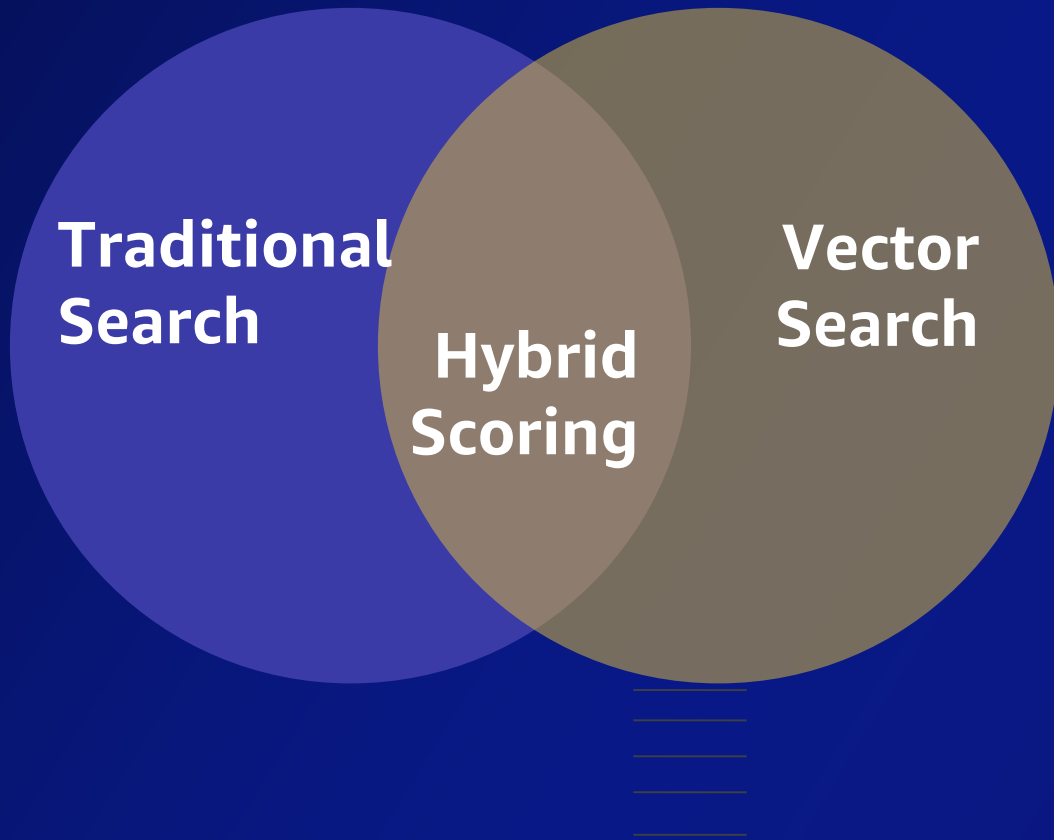


Issue knn query using the `_search` endpoint

`GET product-catalog/_search`

```
{
  "knn": {
    "field": "desc_embedding",
    "query_vector": [0.123, 0.244,...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  }
}
```

# Hybrid scoring gets you the best of both worlds



```
GET product-catalog/_search
{
  "query": {
    "match": {
      "description": {
        "query": "summer clothes",
        "boost": 0.9
      }
    }
  },
  "knn": {
    "field": "desc_embedding",
    "query_vector": [0.123, 0.244,...],
    "k": 5,
    "num_candidates": 50,
    "boost": 0.1,
    "filter": {
      "term": {
        "department": "women"
      }
    }
  }
}
```



# How will you apply vector search?



**Product similarity search**

“Do you sell black v-neck shirts that look like this?”



**Answer technical support**

“What are the troubleshooting steps for \_\_\_\_?”



**Query medical knowledge**

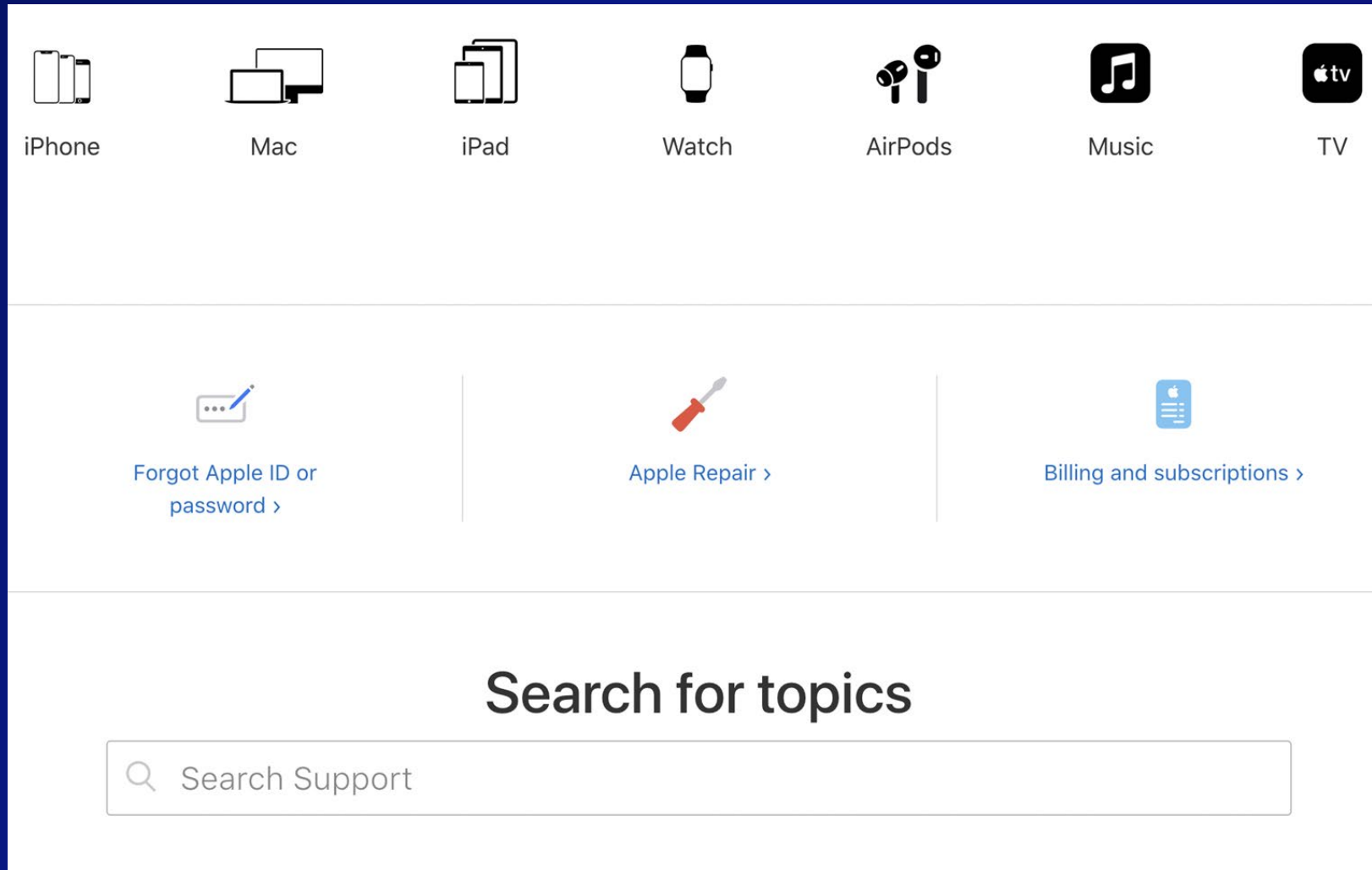
“Is lithium used to treat bipolar disorder?”



**React to user sentiments**

Identify poor customer interactions before they lead to escalations

# NLP #1: No more browsing manuals / FAQs



# Query with question-answering model

Google search results for "what does the engine light mean on my 2018 Mini cooper?". The search bar shows the query and the Google logo. Below the search bar, there are tabs for All, Images, Shopping, Videos, News, and More. The search results show "About 1,400,000 results (0.54 seconds)". The first result is from "https://www.miniofwarwick.com" and is titled "Discover the MINI Cooper Warning Lights Meanings". The snippet states: "Engine Light: The MINI Cooper Yellow engine warning light means **your emission system needs attention**. If there is engine misfiring, the light will flash instead. Tire Pressure Monitor: A yellow light will turn on when you have one or more tires with low pressure." Below the snippet, there is a link to "https://www.miniofwarwick.com" and a "Feedback" button. The "People also ask" section lists several related questions: "Why is my check engine light on in my Mini Cooper?", "Can I drive my Mini with the engine light on?", "What is the most common reason for check engine light?", and "What does a solid check engine light mean?". Below the "People also ask" section, there is a link to "https://www.miniofstevenscreek.com" and a "Feedback" button. The snippet for this link states: "MINI Check Engine Light On? | Common Symptoms & What to ... Common Causes For MINI Check Engine Light · Loose Gas Cap: **Your gas cap is loose, broken, or simply missing.** · Failing Catalytic Converter: **Your catalytic ...**"

```
POST _ml/trained_models/deepset_minilm-uncased-squad2/_infer
{
  "docs": [{ "text_field": "My name is Peter and I live in London" }],
  "inference_config": {
    "question_answering": {
      "question": "Where do I live?"
    }
  }
}
```

```
{
  "inference_results": [
    {
      "predicted_value": "London",
      "start_offset": 31,
      "end_offset": 37,
      "prediction_probability": 0.9948918266325819
    }
  ]
}
```

# NLP #2: Prevent customer service escalations

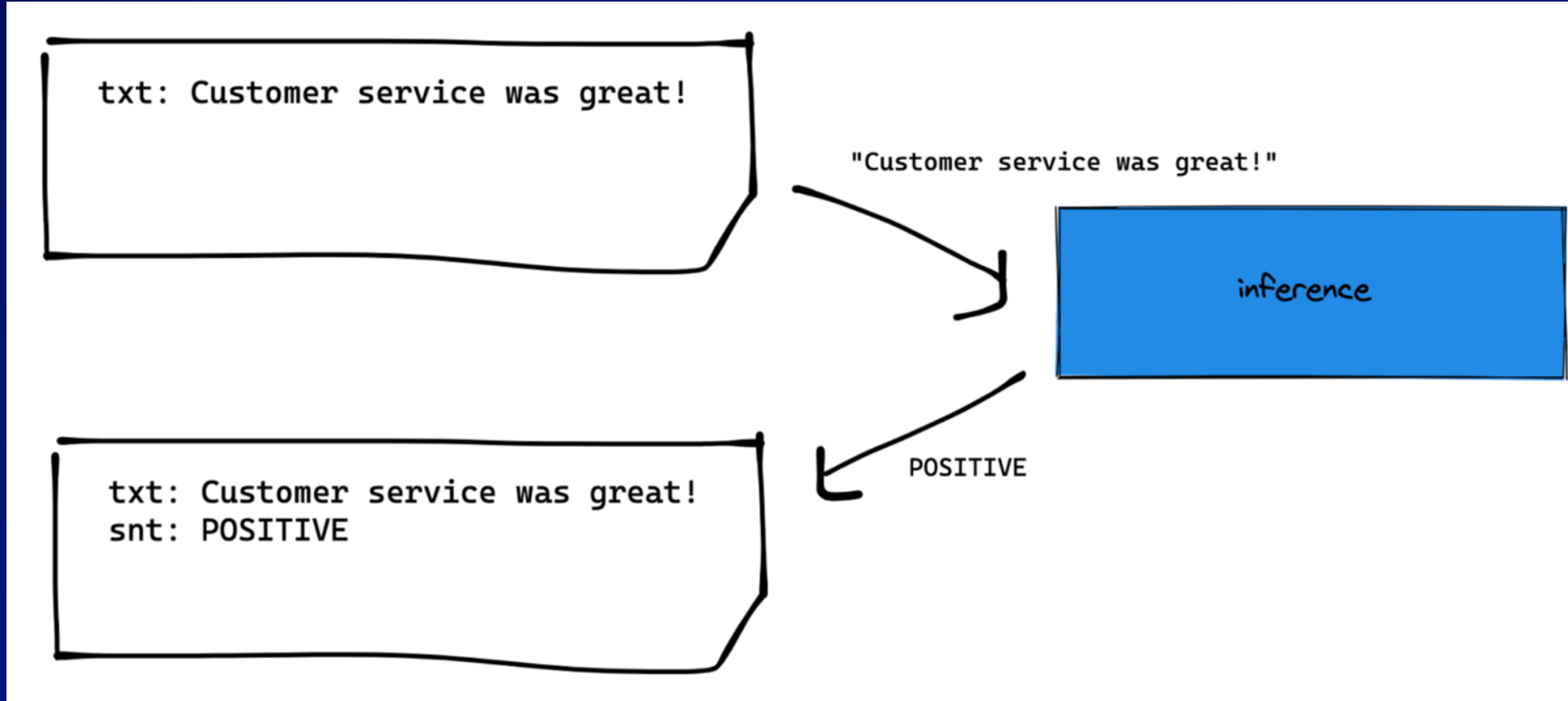
Possibly one of the worst service experiences at [REDACTED] We checked in at 5pm and were seated for about 20 minutes and no waiter. Not busy at all. We saw a table being seated 10 minutes after us that was served right away. We had to go back to host desk to ask for our waiters name and host was flustered at best.

I received a text from Open Table saying I missed my reservation almost an hour after we checked in. On my way out I asked them to check their system and note that they would have a table ready for me shortly. I said no just check me out. They showed them my receipt.



# Sentiment models classify feedback

MONITOR YOUR BRAND IN SOCIAL MEDIA, GET NOTIFIED AFTER FEEDBACK!



# Personalization drives digital commerce outcomes

Online shoppers today

88%

are more likely to continue shopping on websites that offer a personalized experience

84%

report personalization already influences their shopping decisions

68%

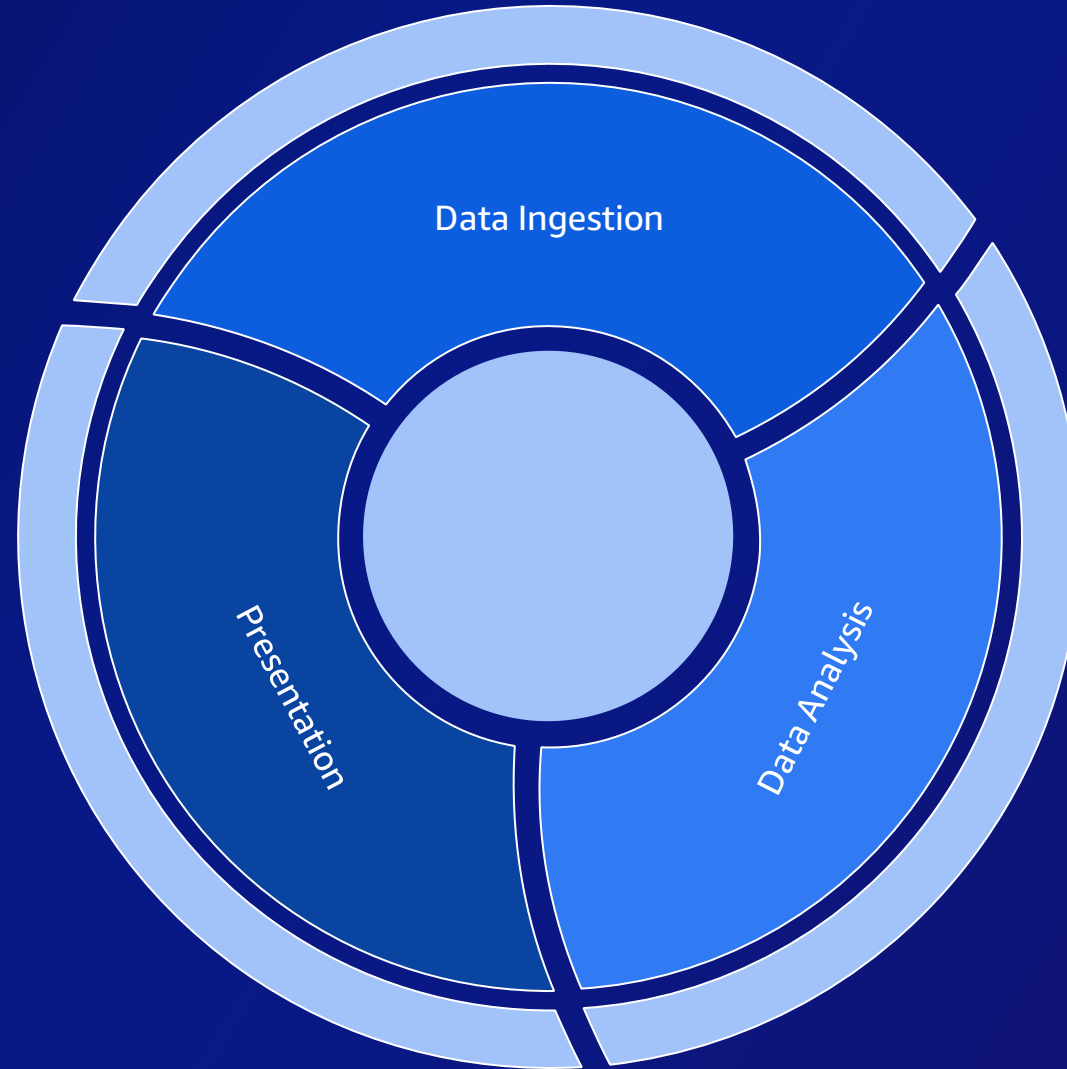
have purchased items they did not intend to initially, due to personalized recommendations

Source: *Product Over Price: The Critical Role Personalization Plays in Converting Online Searches into Sales*, commissioned by Elastic and conducted by Wakefield Research



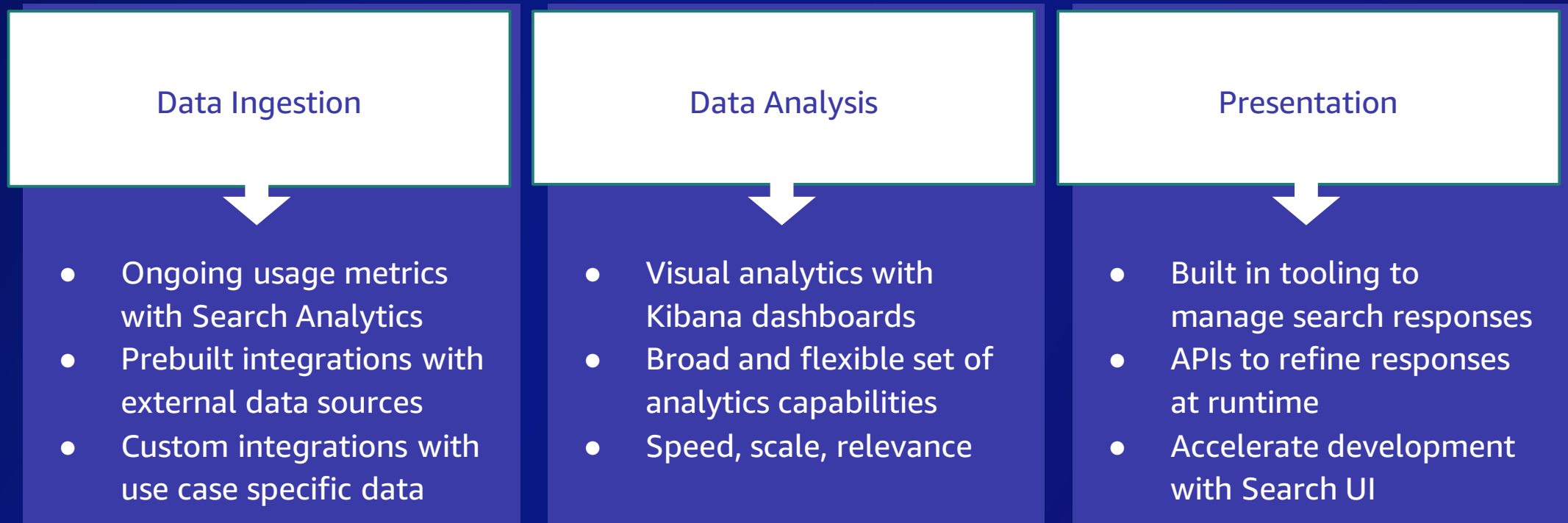
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Personalization is a **\*data\*** problem





# Elastic capabilities enable all stages of a personalization implementation



# Personalization journey with elastic

Add/test with the latest in search

Extend data & analysis over time

Get started with ease

Machine learning, vector search powered recommendations

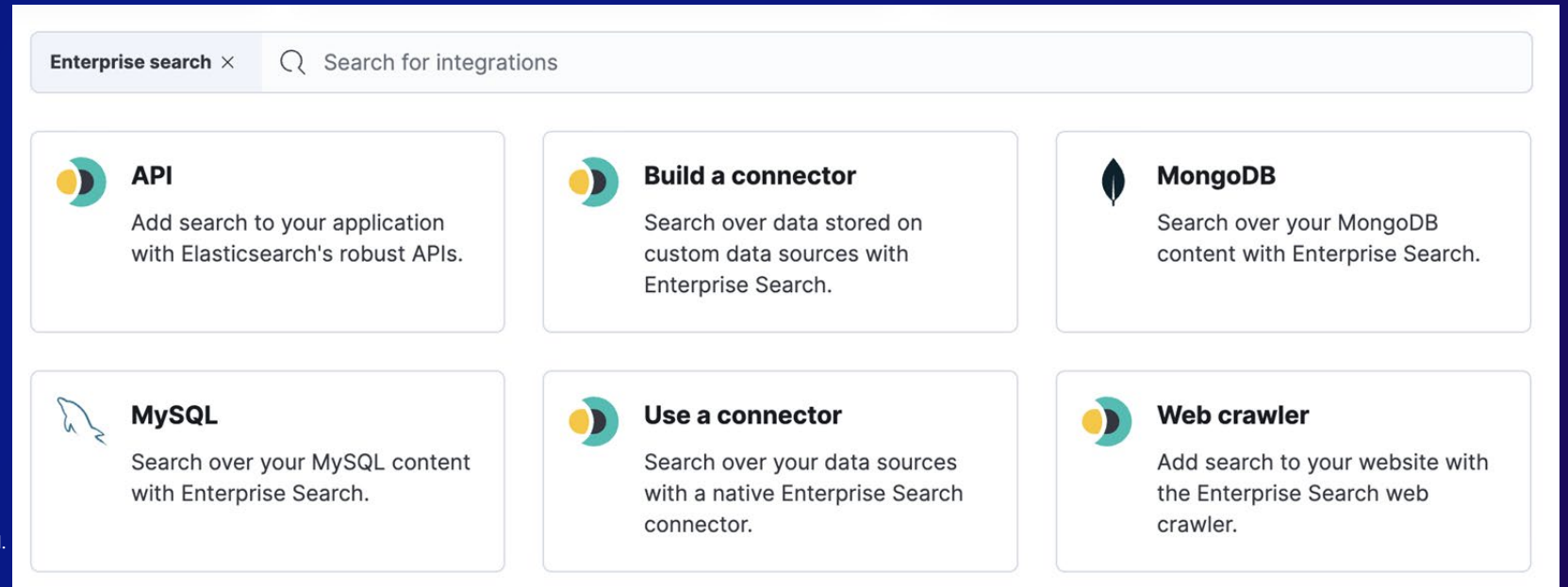
Additional data sources outside Elastic, and build holistic user insights

Built in usage metrics tracking and Analytics API, with a client experience built on Search UI

# Crawl or connect searchable data

*Accelerate building search experiences with ingestion flexibility and scalability*

- **Native web crawler** with PDF extraction and simple authenticated crawls
- Native connector clients for **MongoDB, Gitlab, and MySQL**
- Open code connector clients in **Ruby and Python**
- New **connectors-py framework**



# Pipeline management for ML models

- Choose what to ingest, how to transform ingested data, and ML models needed
- Managed, custom, and inference pipelines
- Tune relevance at ingest time

The screenshot displays the AWS OpenSearch console interface for managing pipelines. The top navigation bar includes tabs for Overview, Documents, Index Mappings, Manage Domains, Scheduling, and Pipelines. The main content area is divided into two sections: Ingest Pipelines and ML Inference pipelines. The Ingest Pipelines section shows a list of pipelines, including 'ent-search-generic-ingestion', with a 'Settings' button. The ML Inference pipelines section includes a '+ Add inference pipeline' button and a description of how inference pipelines are used as processors. Below these sections, two modal windows are shown. The 'Pipeline settings' modal for 'ent-search-generic-ingestion' allows users to optimize content for search by enabling 'Content extraction', 'Reduce whitespace', and 'ML Inference Pipelines'. The 'Add an inference pipeline' modal prompts users to enter a unique name, select a trained ML model, and specify source and destination fields for data processing.

**search-test** Search engines Crawl

Overview Documents Index Mappings Manage Domains Scheduling **Pipelines**

**Ingest Pipelines**  
Ingest pipelines optimize your index for search applications

ent-search-generic-ingestion Settings Managed

Learn more about using pipelines in Enterprise Search

**ML Inference pipelines** + Add inference pipeline  
Inference pipelines will be run as processors from the Enterprise Search Ingest Pipeline  
Learn more about deploying ML models in Elastic

**Pipeline settings**  
ent-search-generic-ingestion  
This pipeline runs automatically on all Crawler and Connector indices created through Enterprise Search.  
Learn more about Enterprise Search ingest pipelines

Optimize your content for search

- ☒ **Content extraction**  
Extract content from images and PDF files
- ☒ **Reduce whitespace**  
Trim extra whitespace from your documents automatically
- ☒ **ML Inference Pipelines**  
Enhance your data using compatible trained ML models

With a platinum license, you can create an index-specific version of this configuration and modify it for your use case.

Copy and customize Cancel Save

**Add an inference pipeline**  
Once created, this pipeline will be added as a processor on your Enterprise Search Ingestion Pipeline. You'll also be able to use this pipeline elsewhere in your Elastic deployment.  
Learn more about using ML models in Enterprise Search

Name  
Enter a unique name for this pipeline  
Pipeline names are unique within a deployment and can only contain letters, numbers, underscores, and hyphens. The pipeline name will be automatically prefixed with "ml-inference-".

Select a trained ML Model  
Select a model

Source field  
Select a schema field

Destination field (optional)  
custom\_field\_name  
Your field name will be prefixed with "ml.inference.", if not set it will be defaulted to "ml.inference."

Cancel Create

# Additional features to explore

- ✓ **Named entity recognition**
- ✓ **Zero-shot classification**
- ✓ **Applying NLP and vector search to Observability or Security**

# Performance enhancements



# Storage savings

7.14

match only text  
Save up to 10%

8.1

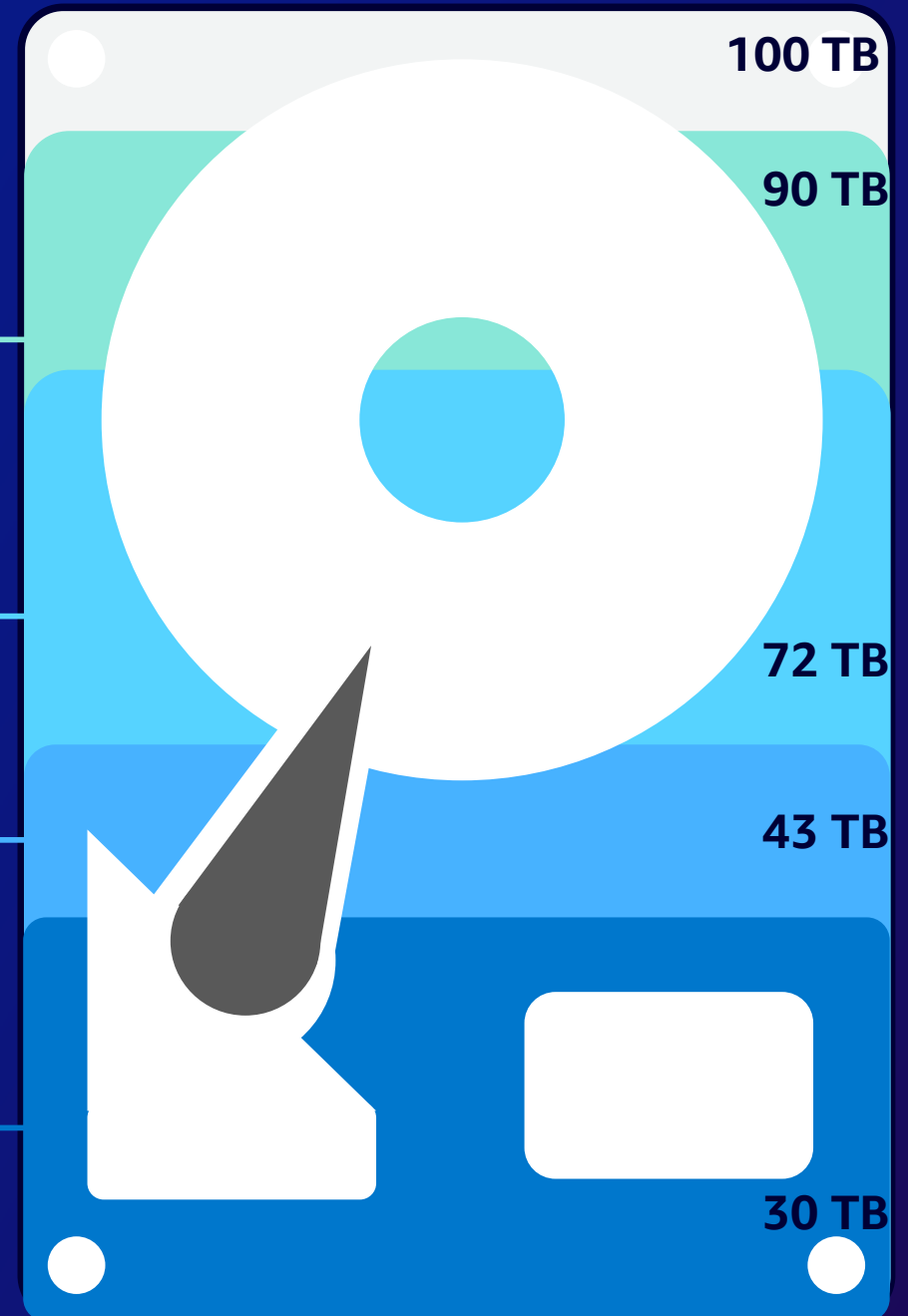
doc-value-only fields  
Save up to 20%

8.4

synthetic\_source  
Save up to 40%

8.5

routing and sorting by TSID  
Save up to 30%



Version wise detailed information on enhancements is available here - <https://www.elastic.co/guide/en/elasticsearch/reference/current/es-release-notes.html>



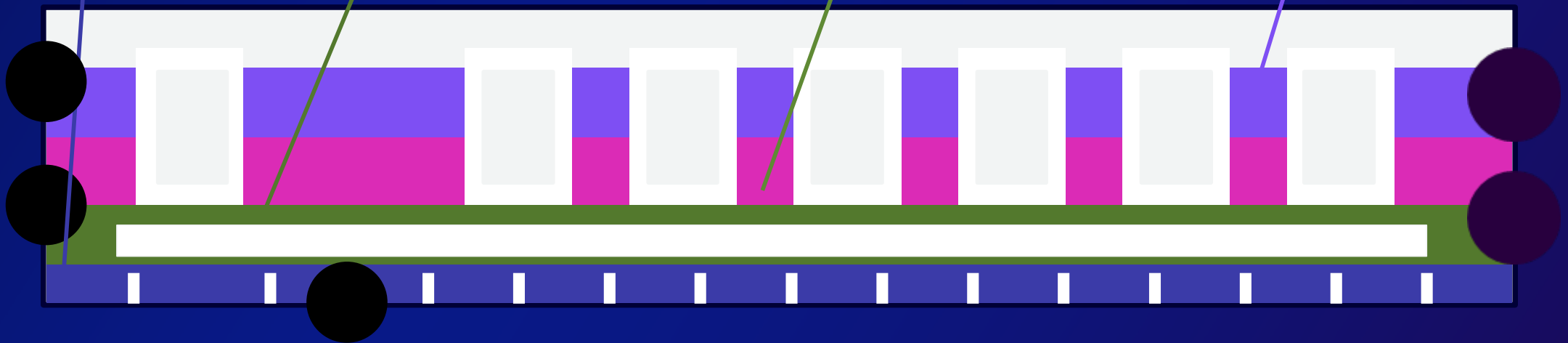
# Memory savings

**7.16** 92% Memory Savings  
on idle shards

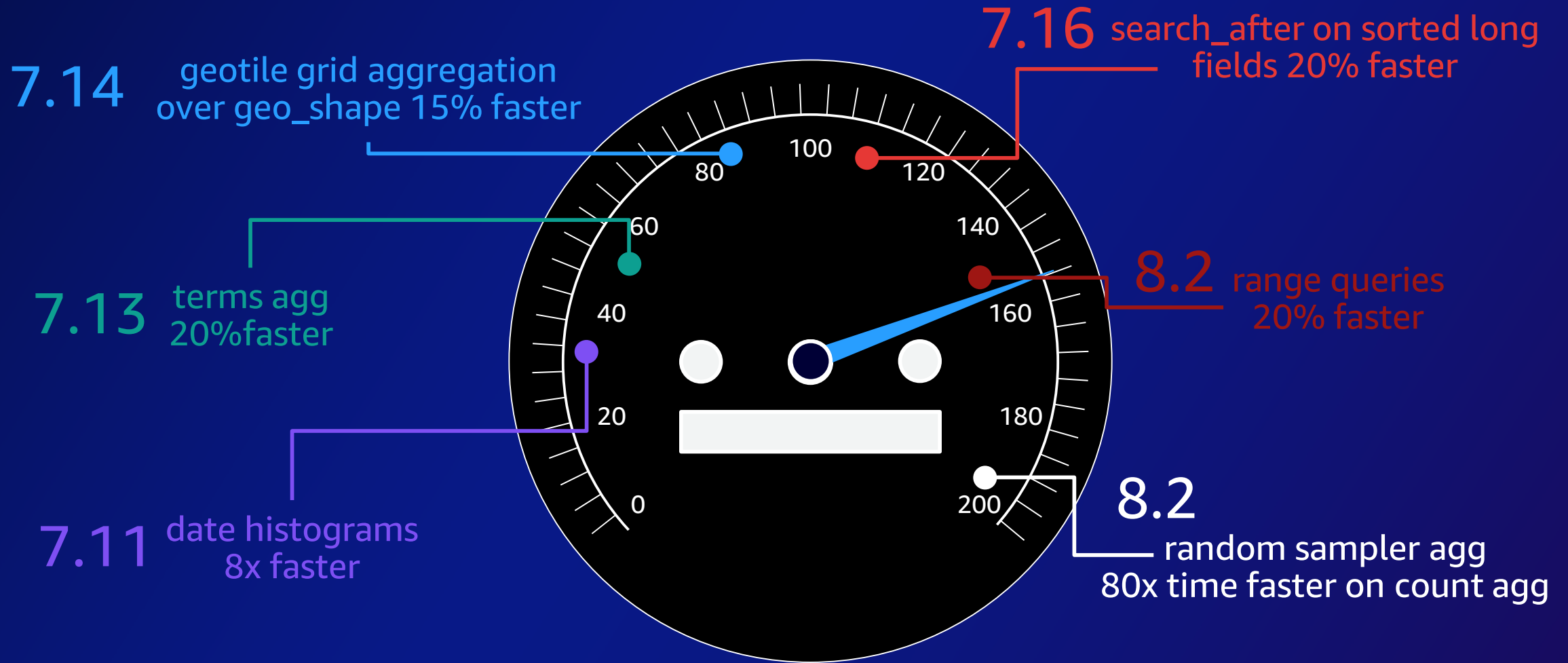
**7.14** Avoid global ordinals in  
composite aggregation

**8.3** New sharding  
3,000 indices per GB of heap

**8.3** New sharding guidance  
1kb per field



# Performance



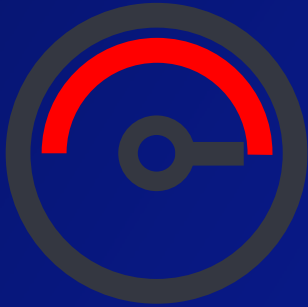
# Search 1PB in minutes stored on S3



# Intelligently store and search everything

## Data Usage

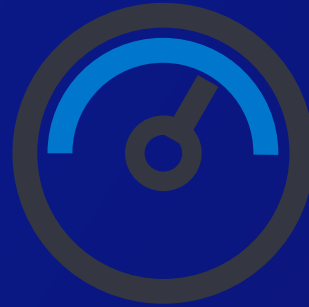
Accessed  
Frequently



Accessed  
Less Frequently



Accessed  
Less Frequently



Accessed  
Intermittently



Accessed  
Rarely



## Performance

milliseconds -  
seconds

seconds -  
10's seconds

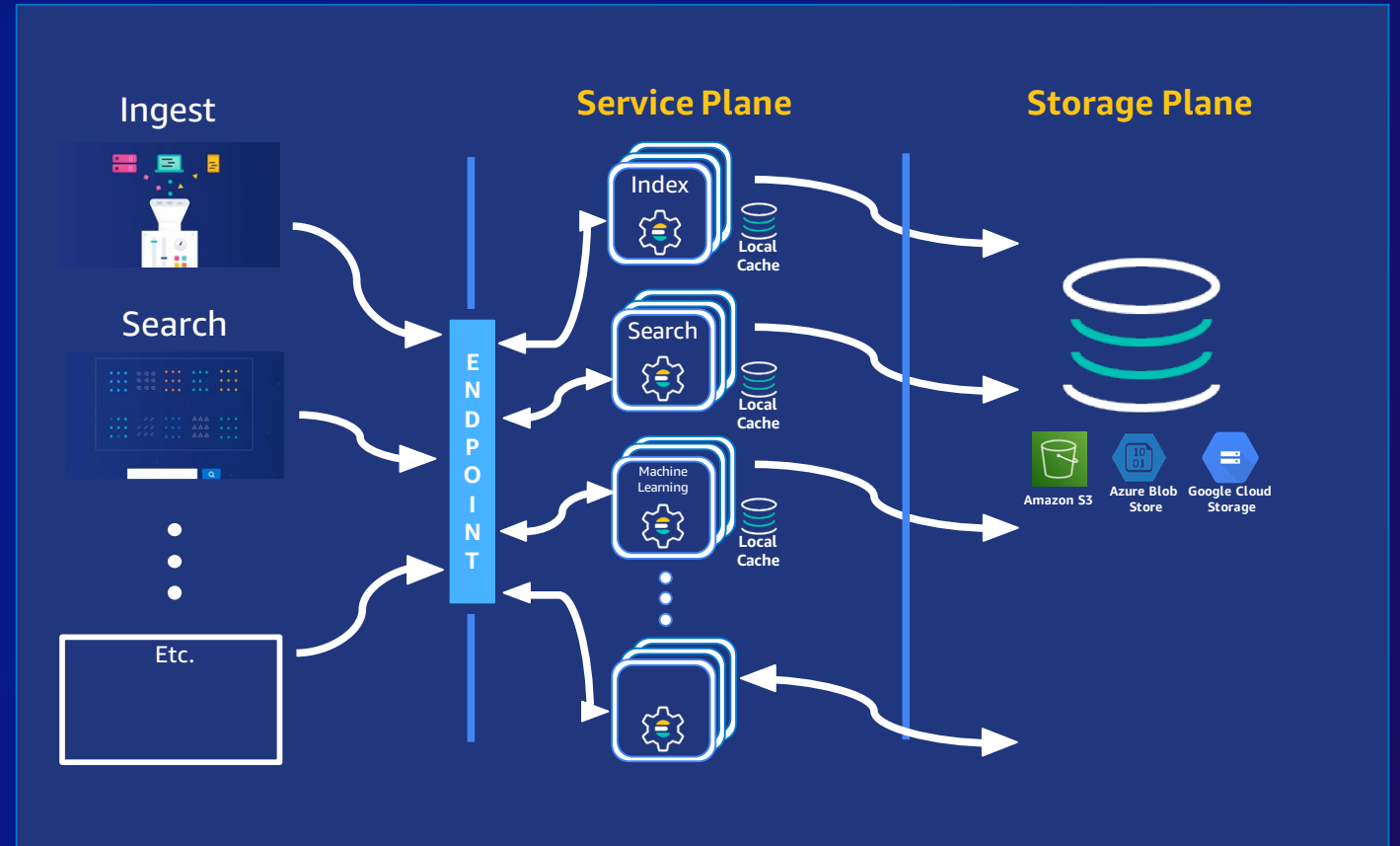
seconds -  
10's seconds

10's seconds -  
minutes

minutes -  
10's minutes

# Serverless Elasticsearch

- Next generation managed service
- Scale compute & storage independently
- Consume only what you need
- Versionless
- Autoscaling - index and search



# Elastic differentiators

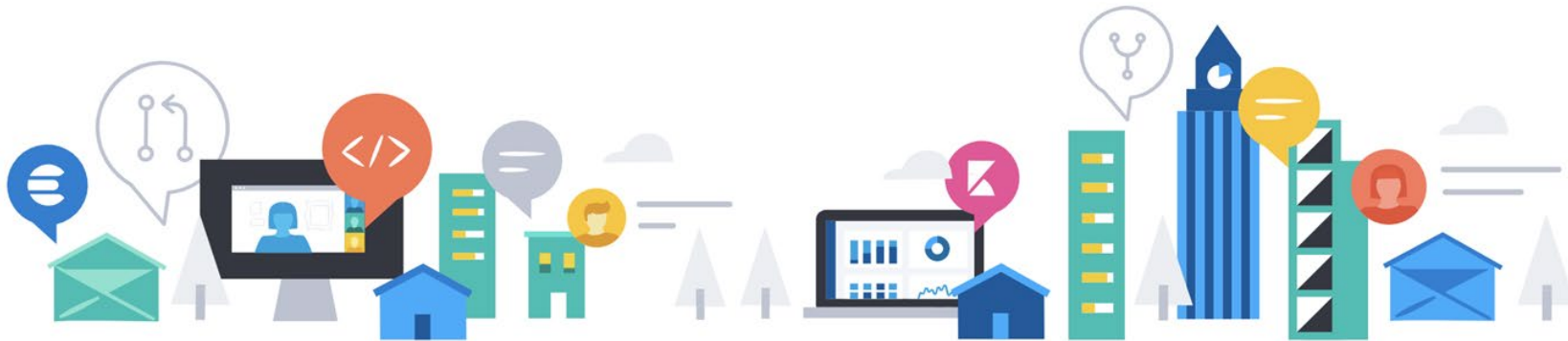
Elastic Cloud ES feature(s)	Customer benefit
Enterprise Search solution	Provides ready-to-use products and integrations for app search and workplace search
Observability solution	Provides ready-to-use log monitoring, infrastructure monitoring, APM, distributed tracing, all within the same UI; single pane of glass
Security solution	Provides ready-to-use SIEM with detection engine, rules, endpoint security, XDR, etc
Machine learning	Save time and tool bloat by using Elasticsearch to store, transform, build, test, and deploy machine learning models natively. Anomaly detection and supervised ML for sentiment analysis.
Elastic Agent and Fleet	Delivers a single unified agent with Fleet management capabilities, enabling automations for observability and security at scale from a single UI
Kibana Maps	Allows for easier analysis of geospatial data
Kibana actions	Integrates with several popular third-party systems
Breadth of integrations	Hundreds of Cloud and other native integrations

# Community resources

**Elastic Contributor  
Program**

**Elastic User  
Groups**

**Elastic for Students and  
Educators**



**[discuss.elastic.co](https://discuss.elastic.co)**

**Community YouTube  
Channel**

**Elastic Stack Community  
Slack Workspace**



# Thank you!

Ravindra Ramnani  
Principal Solutions Architect,  
Elastic



Please complete the  
session survey