aws SUMMIT

INDIA | MAY 25, 2023

AIML003

Innovations from Elasticsearch - Drive speed, scale & relevance

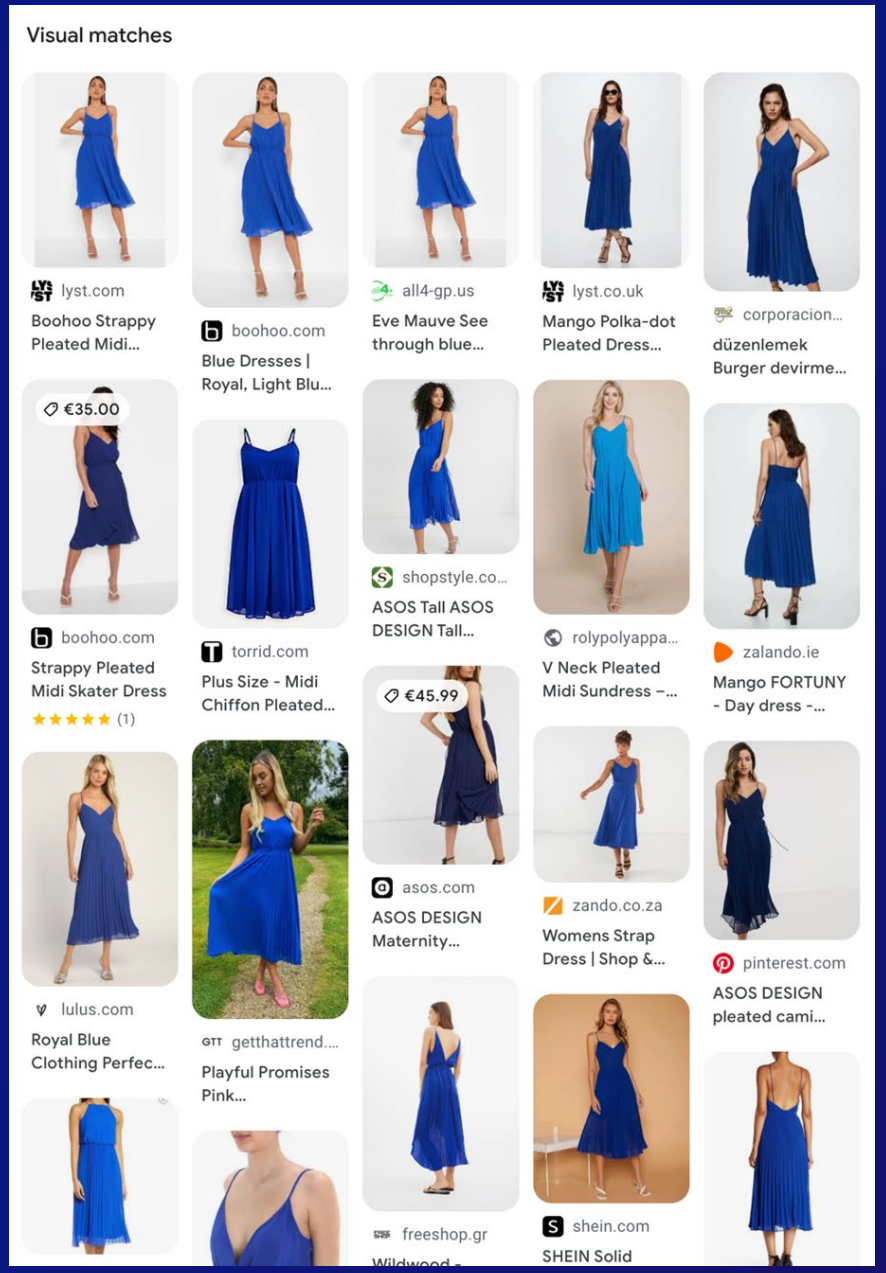
The latest innovations from Elasticsearch: How to implement recent developments in search to drive speed scale, and relevance

Ravindra Ramnani
Principal Solutions Architect
Elastic

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?

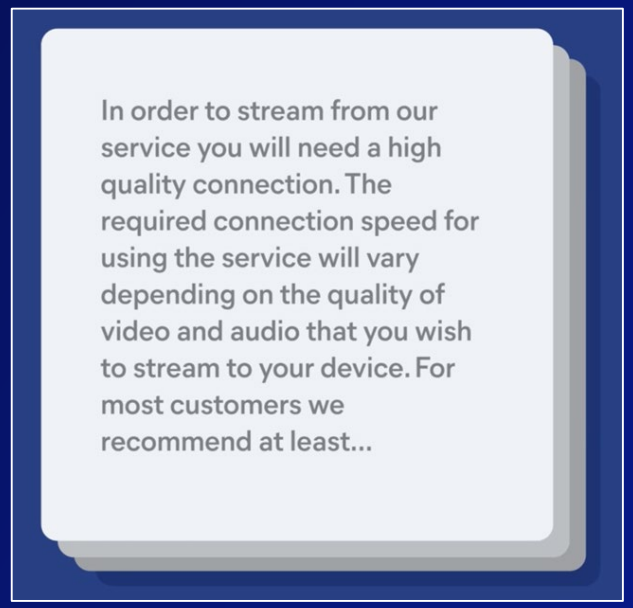


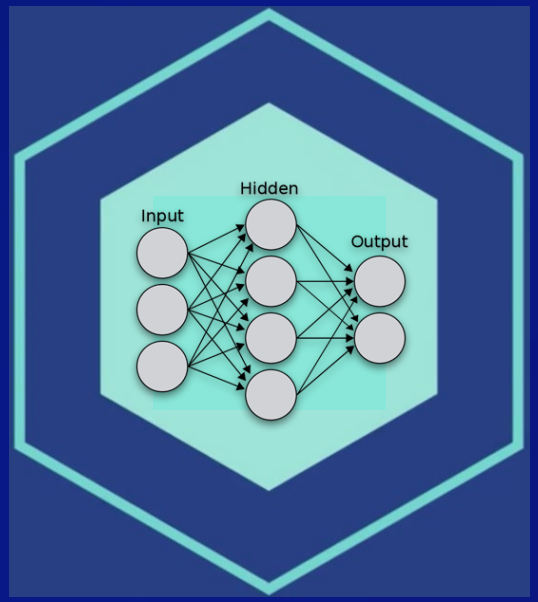
Vector search lets you find what you mean, with higher relevance

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

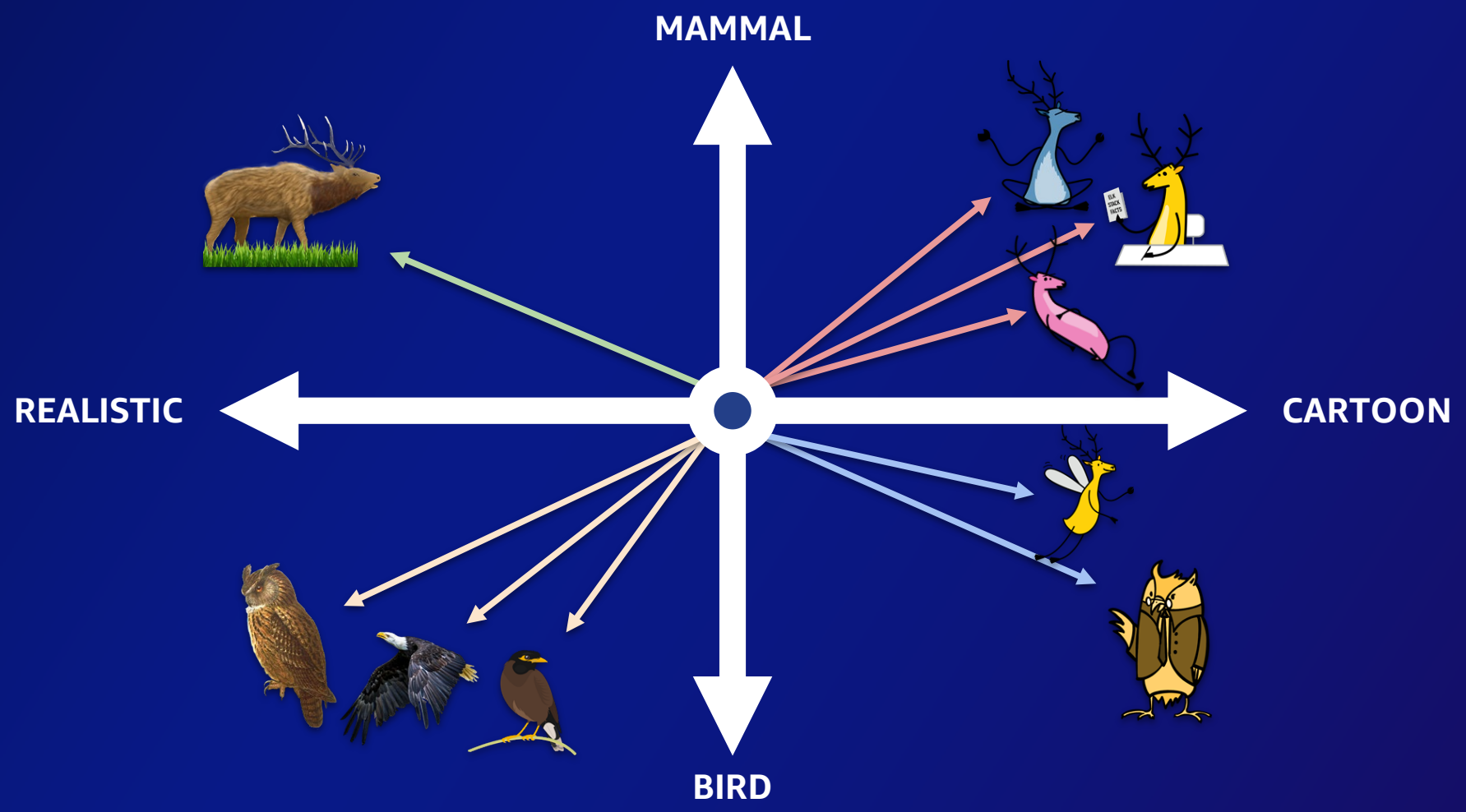
Convert data into vector representation

UNDERLYING APPROACH OF VECTOR SEARCH

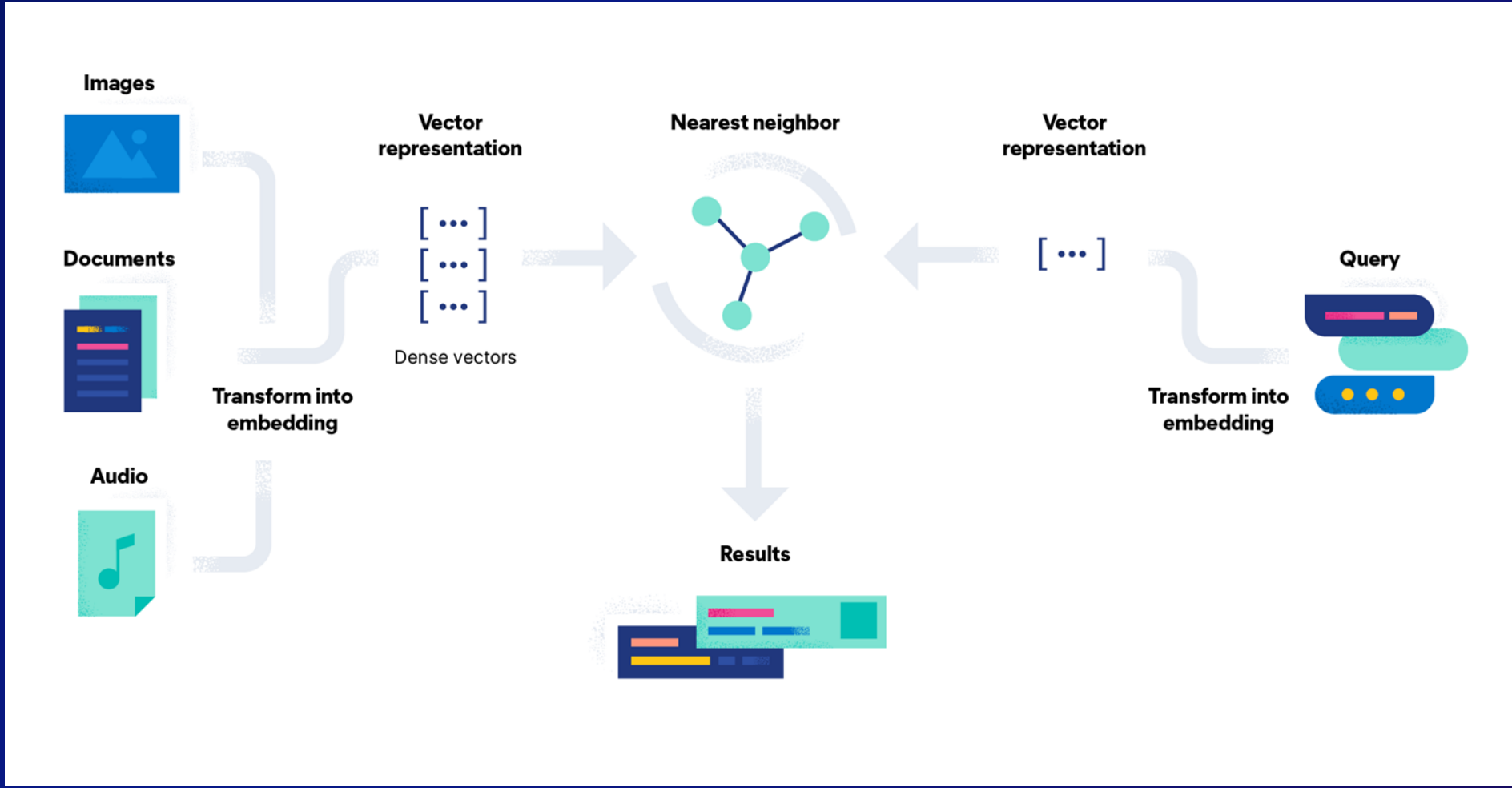




In vector space, similar data are nearby



Vector search returns nearest neighbors of “vectorized” query



Vector search and NLP may be intimidating:

- Lots of data, labelled
- Expertise
- Scale processing

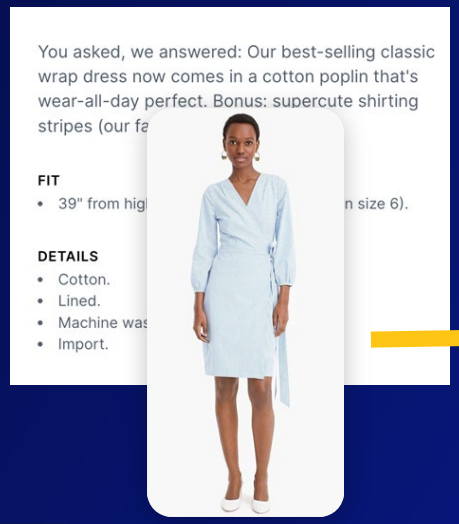
Elastic makes vector search easy

Just bring the data

Inherit scalability

Flexibility

Apply vector search in 3 steps



Source data

elastic cloud

2

POST /_doc

Stored in Elasticsearch:

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

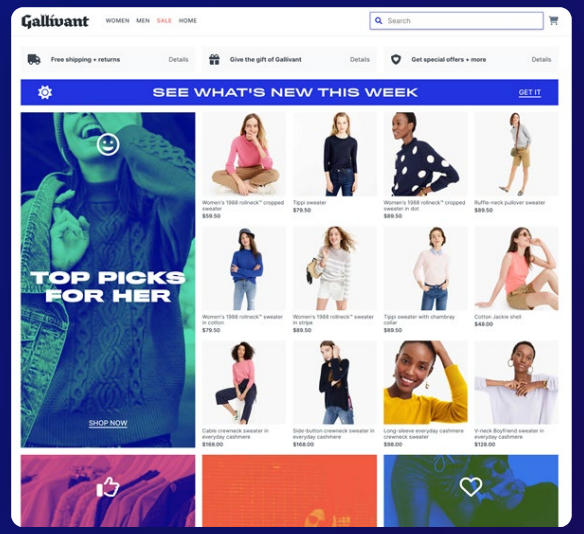
Transformer model

1

3

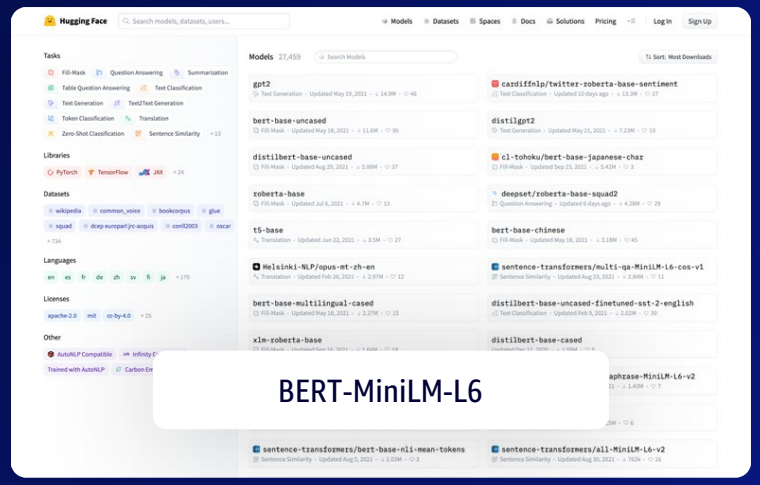
GET /_search

with kNN clause

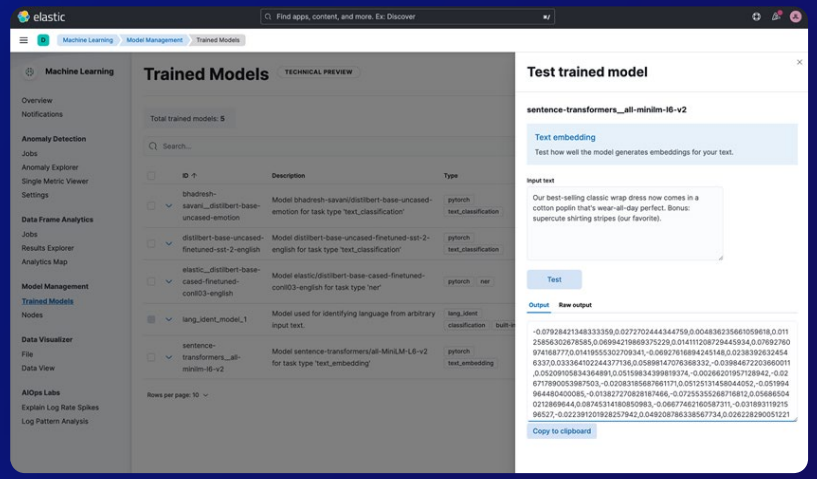


Search-powered application

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
```

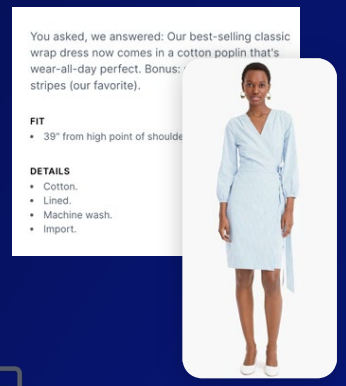


PyTorch

eland

Load the model to the cluster

Step 2: Vectorize during data ingestion



Source data

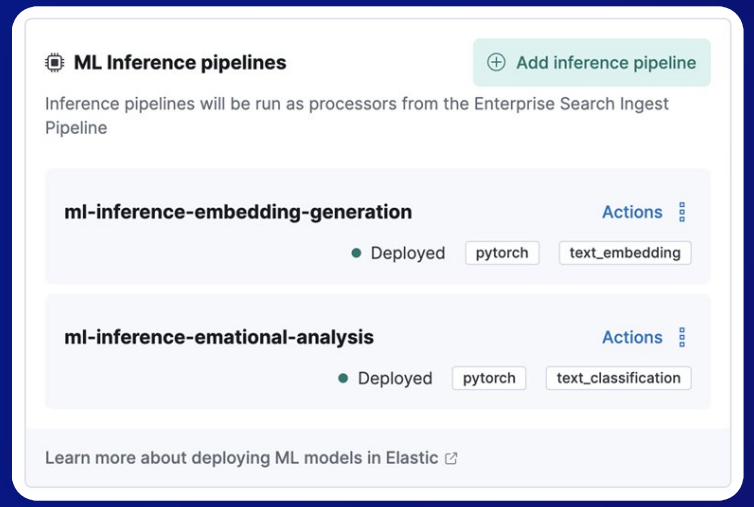
POST /_doc

Standard field indexing for non-vector types

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

POST /_doc

Encoding via Inference Processor



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

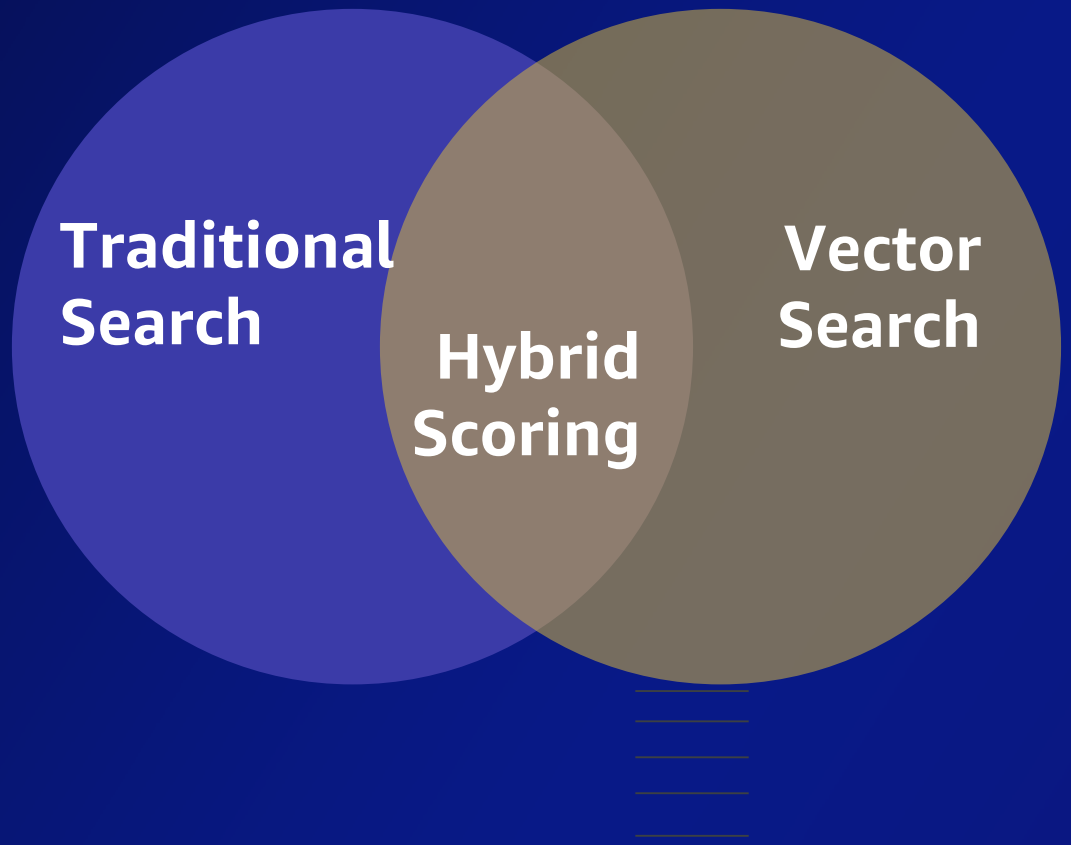
PyTorch

Issue knn query using the _search endpoint

```
GET product-catalog/_search

{
  "knn": {
    "field": "desc_embbeding",
    "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_embbeding",
    "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 ___?"

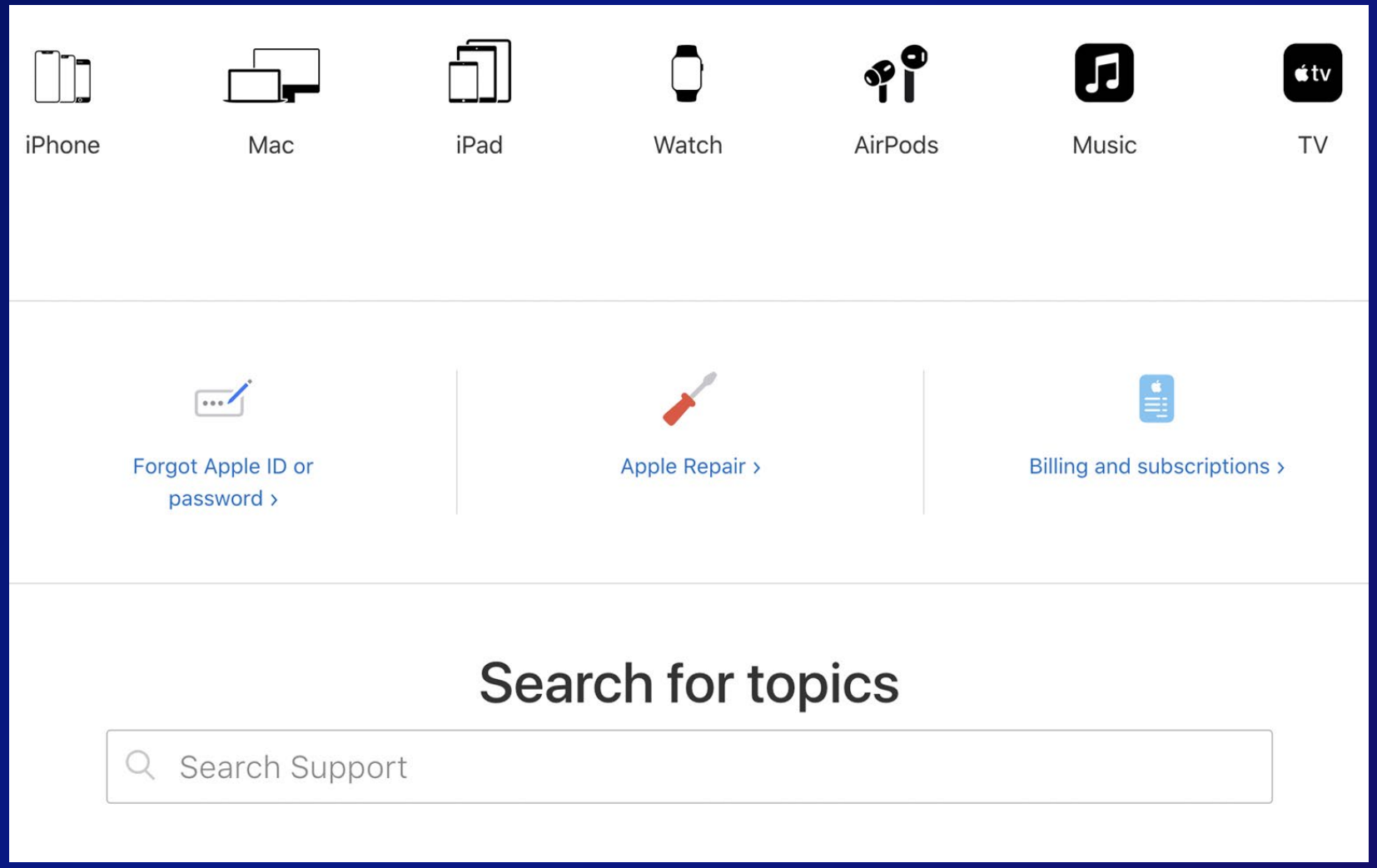
Query medical knowledge

"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

what does the engine light mean on my 2018 Mini cooper?

All | Images | Shopping | Videos | News | More | Tools

About 1,400,000 results (0.54 seconds)

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.

https://www.miniofwarwick.com › service-tips-and-tricks

Discover the MINI Cooper Warning Lights Meanings

About featured snippets • Feedback

People also ask

- 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?
- What does a solid check engine light mean?

Feedback

https://www.miniofstevenscreek.com › service › check-...

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
    }
  ]
}
```

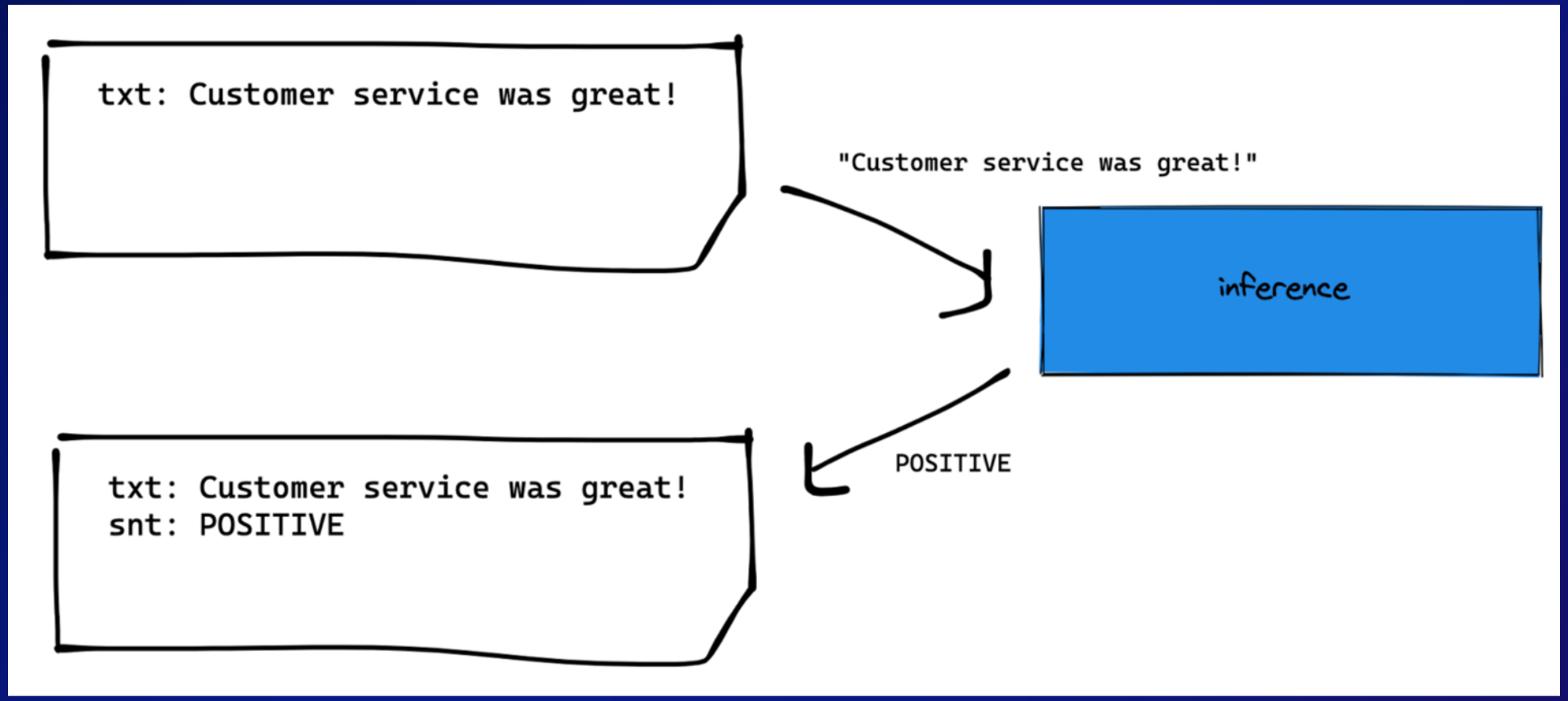
NLP #2: Prevent customer service escalations

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 m 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%

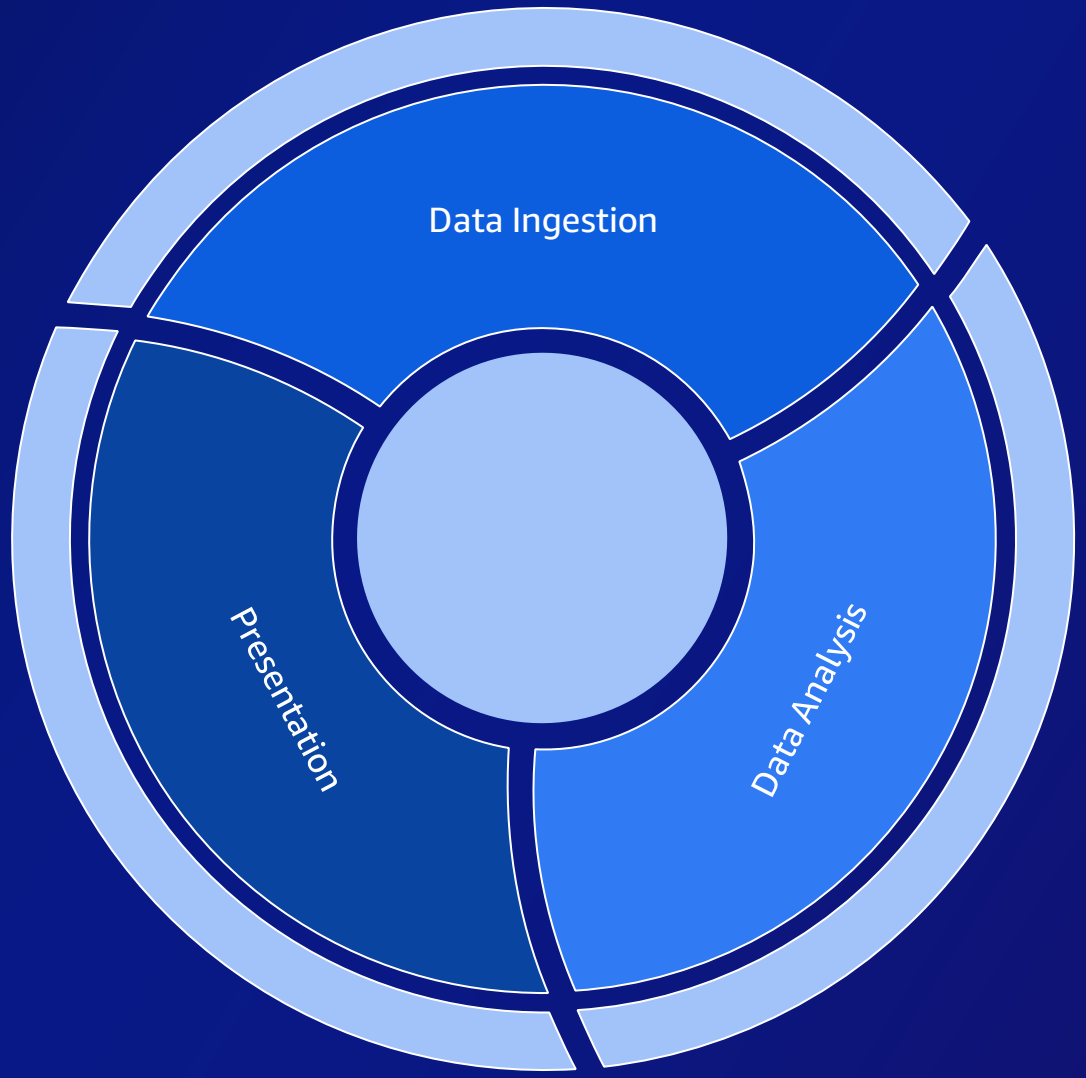
report personalization already influences their shopping decisions

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

Personalization is a *data* problem



Elastic capabilities enable all stages of a personalization implementation

Data Ingestion

- Ongoing usage metrics with Search Analytics
- Prebuilt integrations with external data sources
- Custom integrations with use case specific data

Data Analysis

- Visual analytics with Kibana dashboards
- Broad and flexible set of analytics capabilities
- Speed, scale, relevance

Presentation

- Built in tooling to manage search responses
- APIs to refine responses at runtime
- Accelerate development with Search UI

Personalization journey with elastic

Add/test with the latest in search

Machine learning, vector search powered recommendations

Extend data & analysis over time

Additional data sources outside Elastic, and build holistic user insights

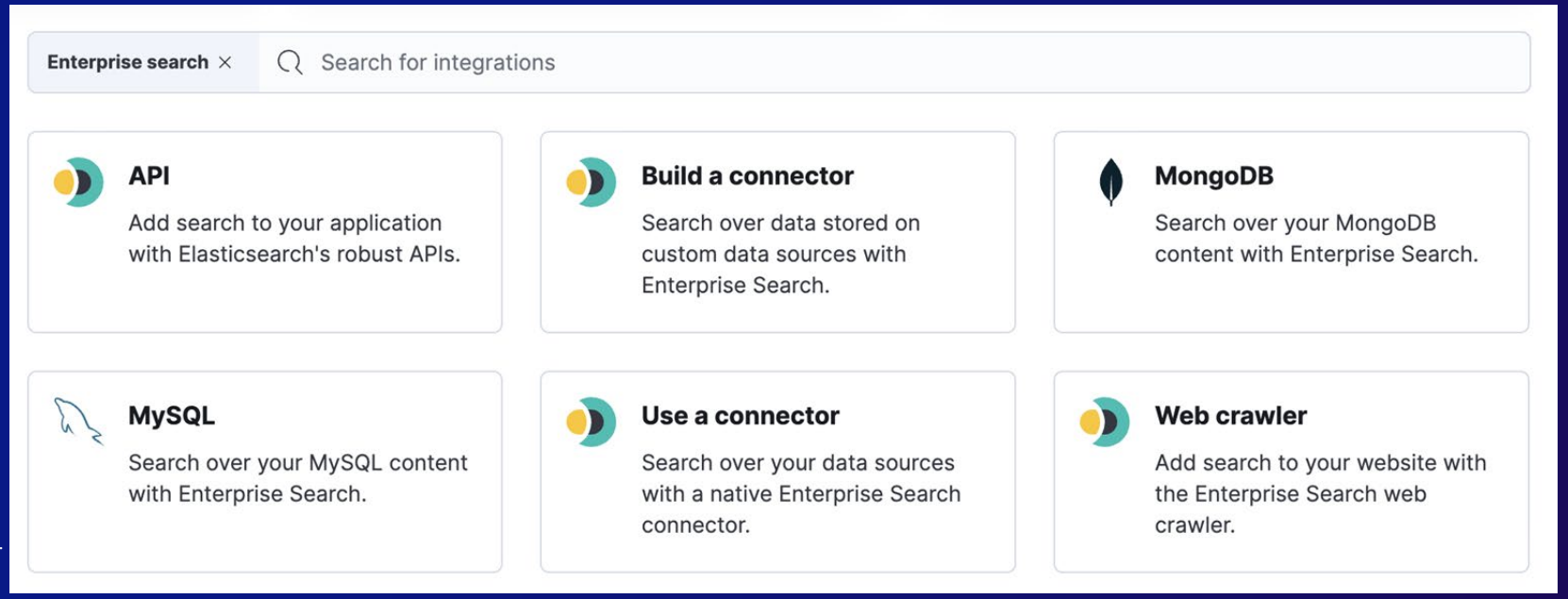
Get started with ease

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

Crawl or connect searchable data

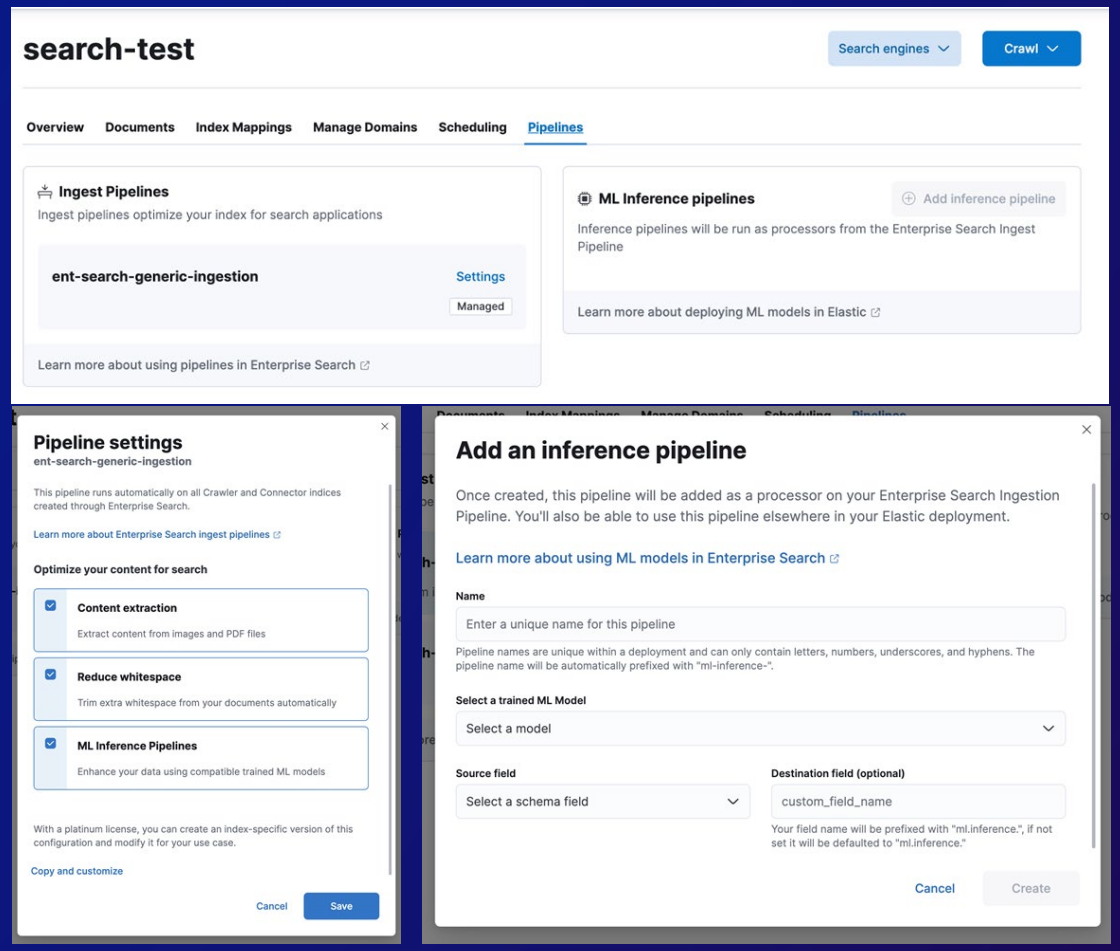
Accelerate building search experiences with ingestion flexibility and scalability

- **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



Additional features to explore

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

Performance enhancements

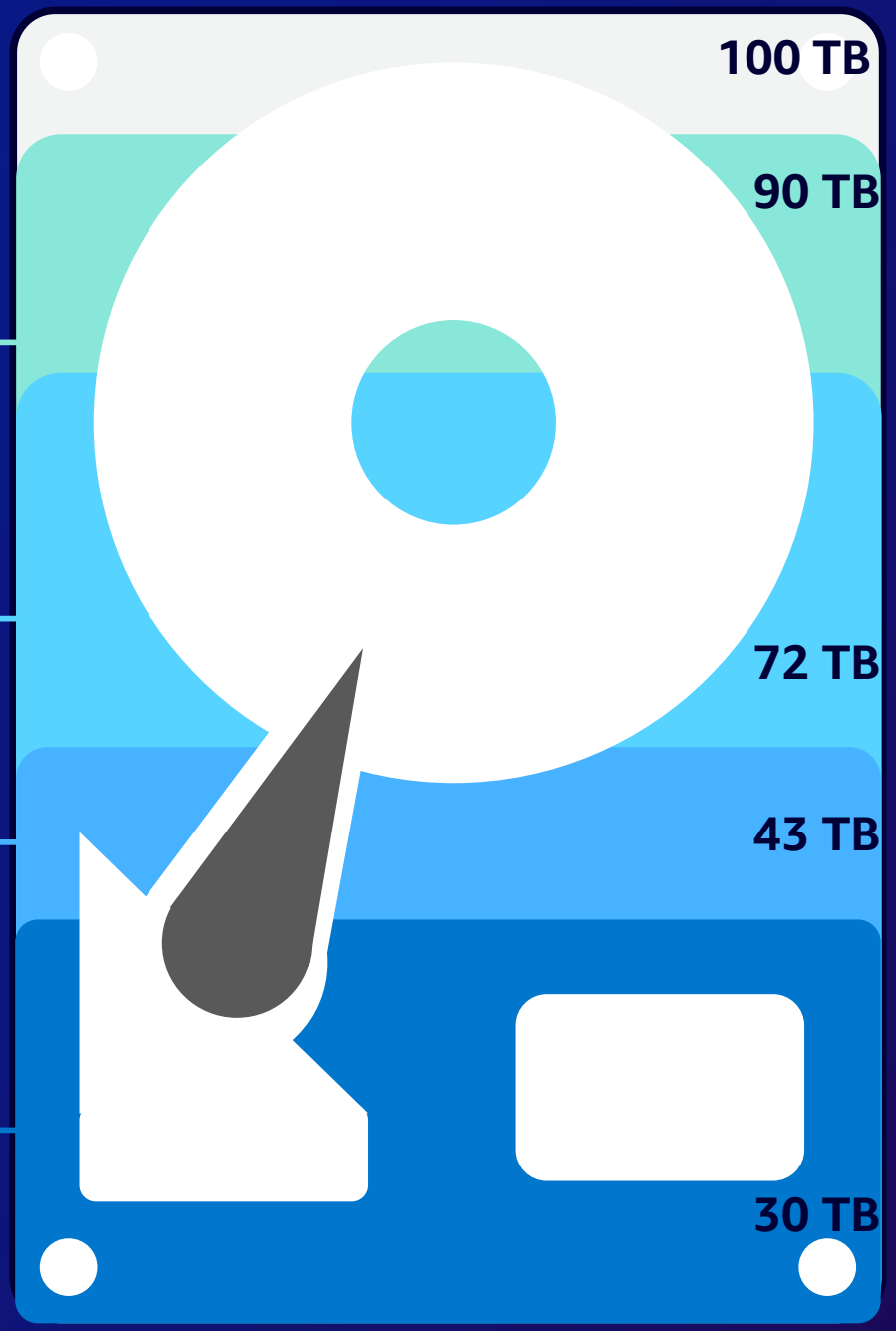
Storage savings

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

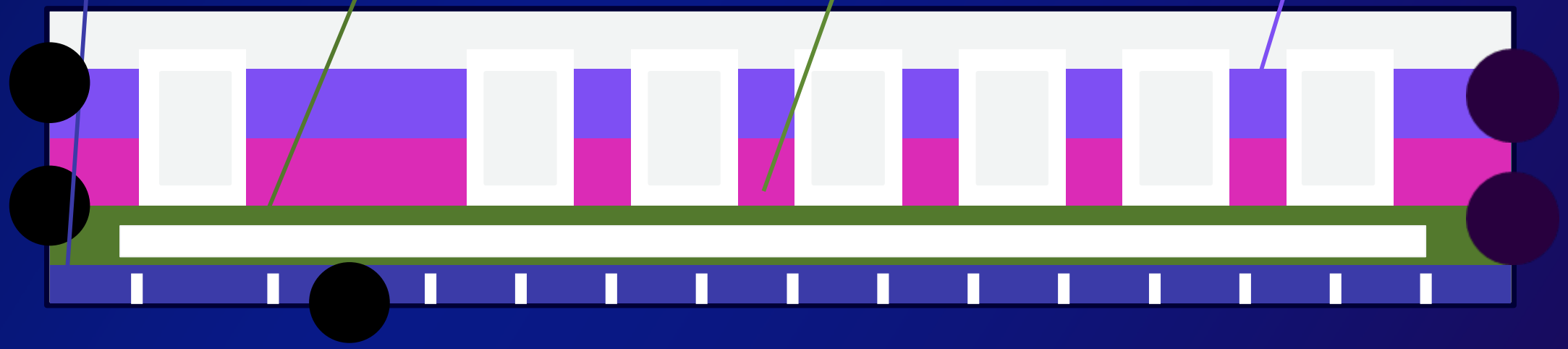
Memory savings

7.16 92% Memory Savings on idle shards

8.3 New sharding 3,000 indices per GB of heap

7.14 Avoid global ordinals in composite aggregation

8.3 New sharding guidance 1kb per field



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

Performance

7.14 geotile grid aggregation over geo_shape 15% faster

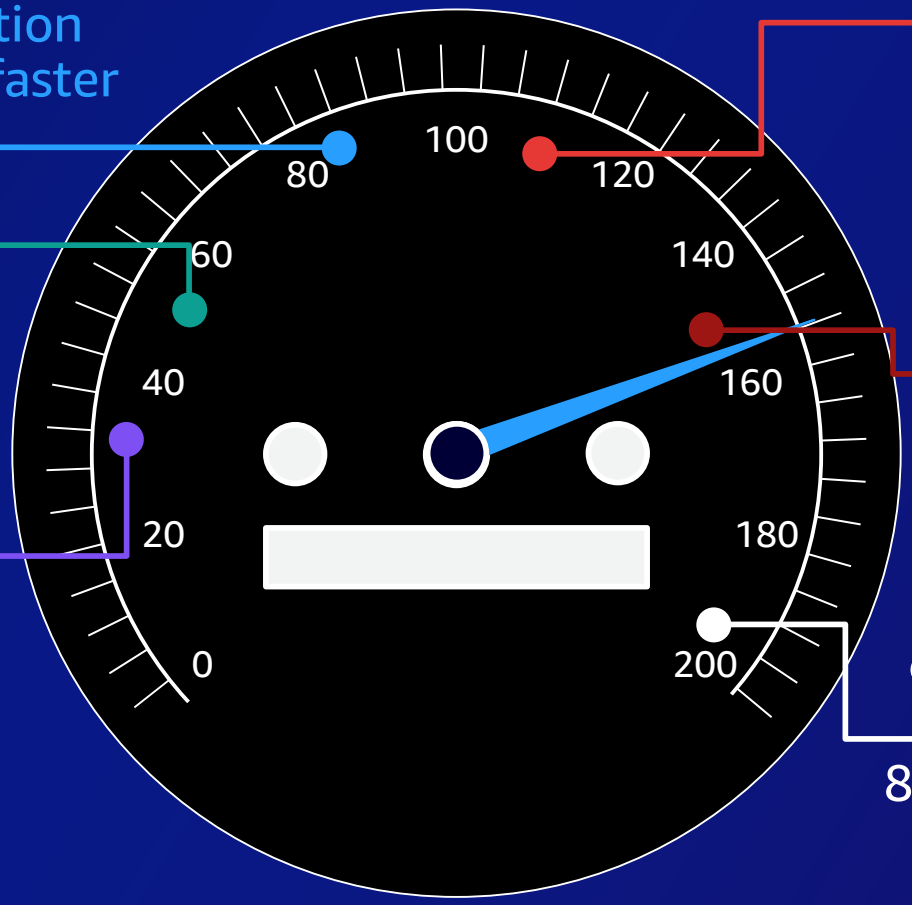
7.16 search_after on sorted long fields 20% faster

7.13 terms agg 20%faster

8.2 range queries 20% faster

7.11 date histograms 8x faster

8.2 random sampler agg 80x time faster on count agg



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

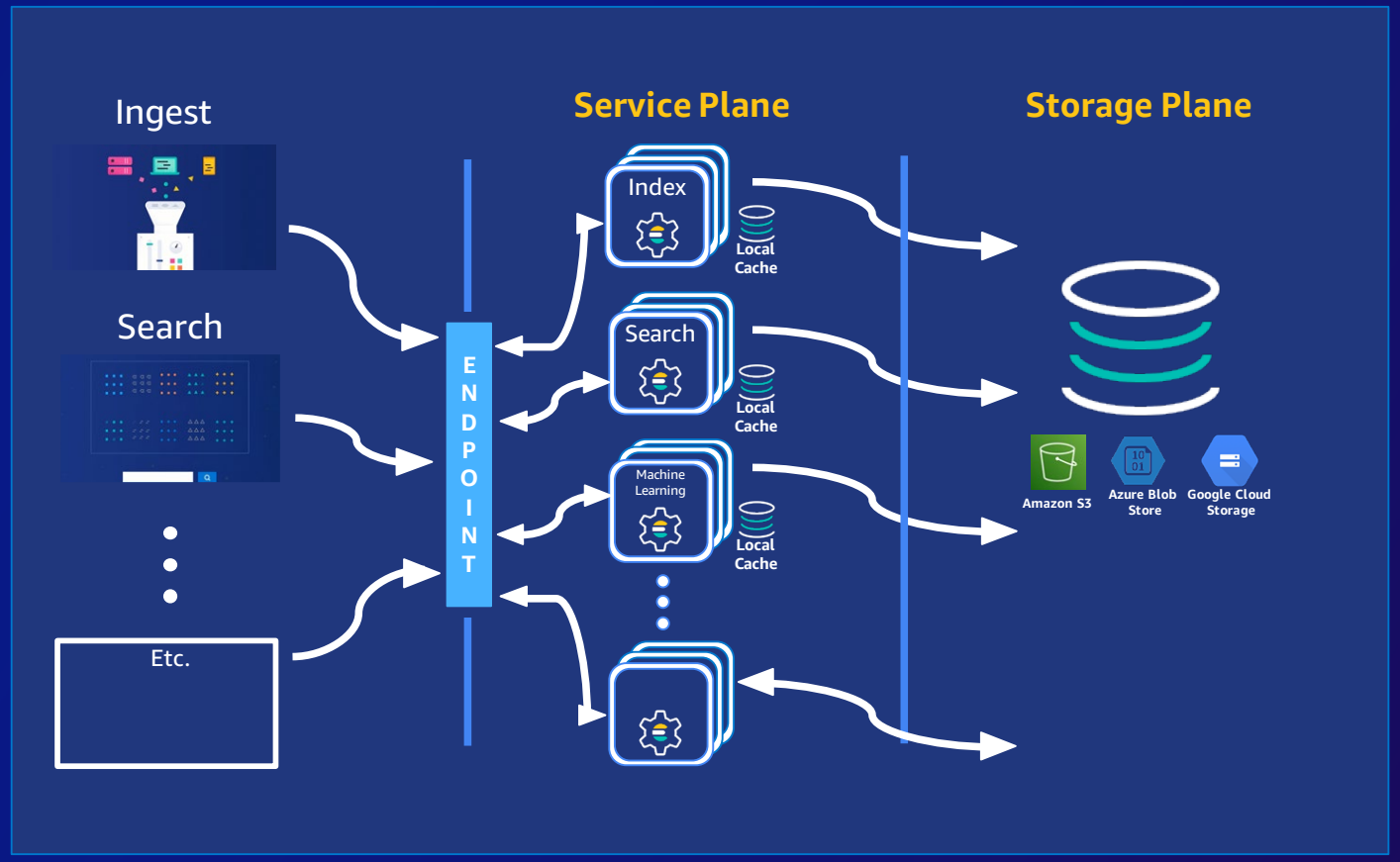
Search 1PB in minutes stored on S3

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

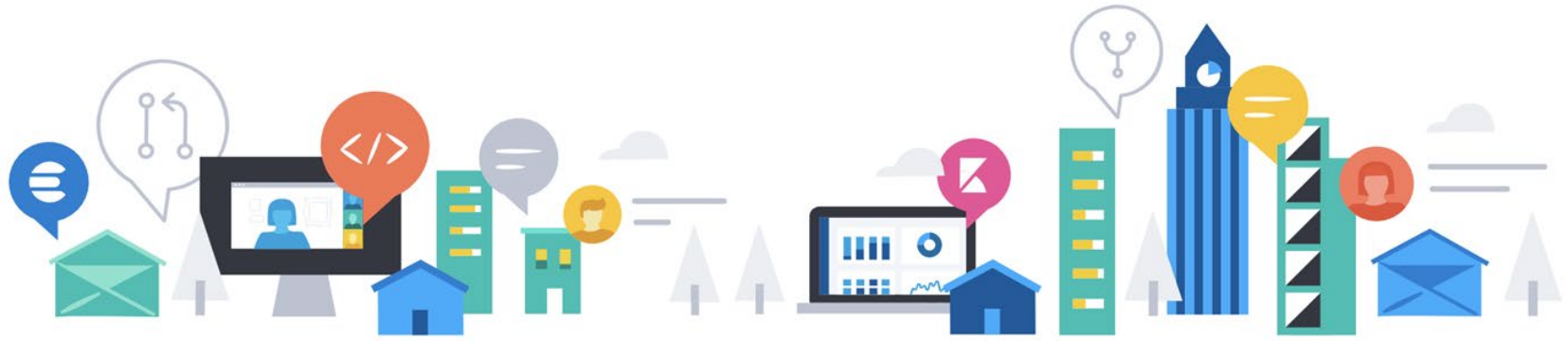
Elastic Cloud ES feature(s)	Customer benefit
Enterprise Search solution	Provides ready-to-use products and integrations for app search and workplace search
Observability solution	Provides ready-to-use log monitoring, infrastructure monitoring, APM, distributed tracing, all within the same UI; single 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

Community YouTube Channel

Elastic Stack Community Slack Workspace

Thank you!

Please complete the session survey

Ravindra Ramnani

Principal Solutions Architect, Elastic