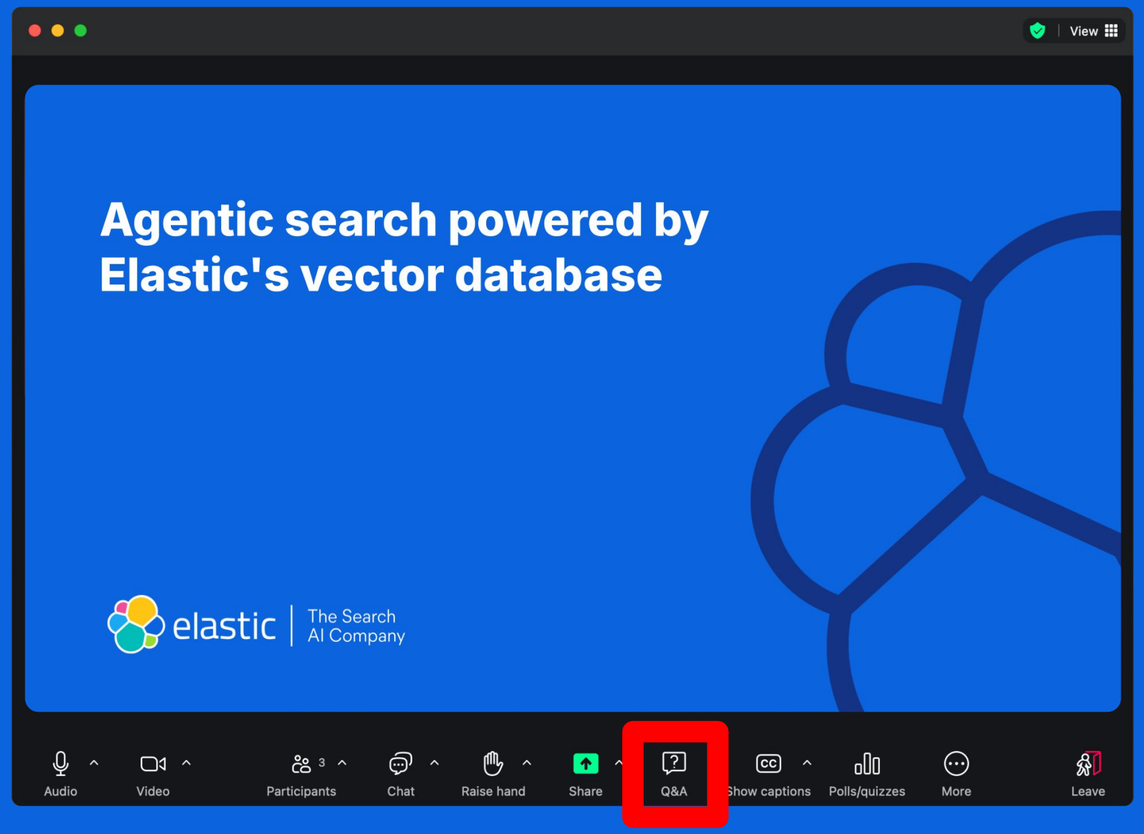
Agentic search powered by Elastic's vector database

elastic | The Search AI Company

Housekeeping & Logistics

- Please submit your questions via the ‘ask a question’ box or in the chat
- The event is recorded and the recording will be shared via email
- Ask more questions on our discuss forum: **discuss.elastic.co**



elastic

Lily Adler

Team Lead, GenAI
Search Specialist

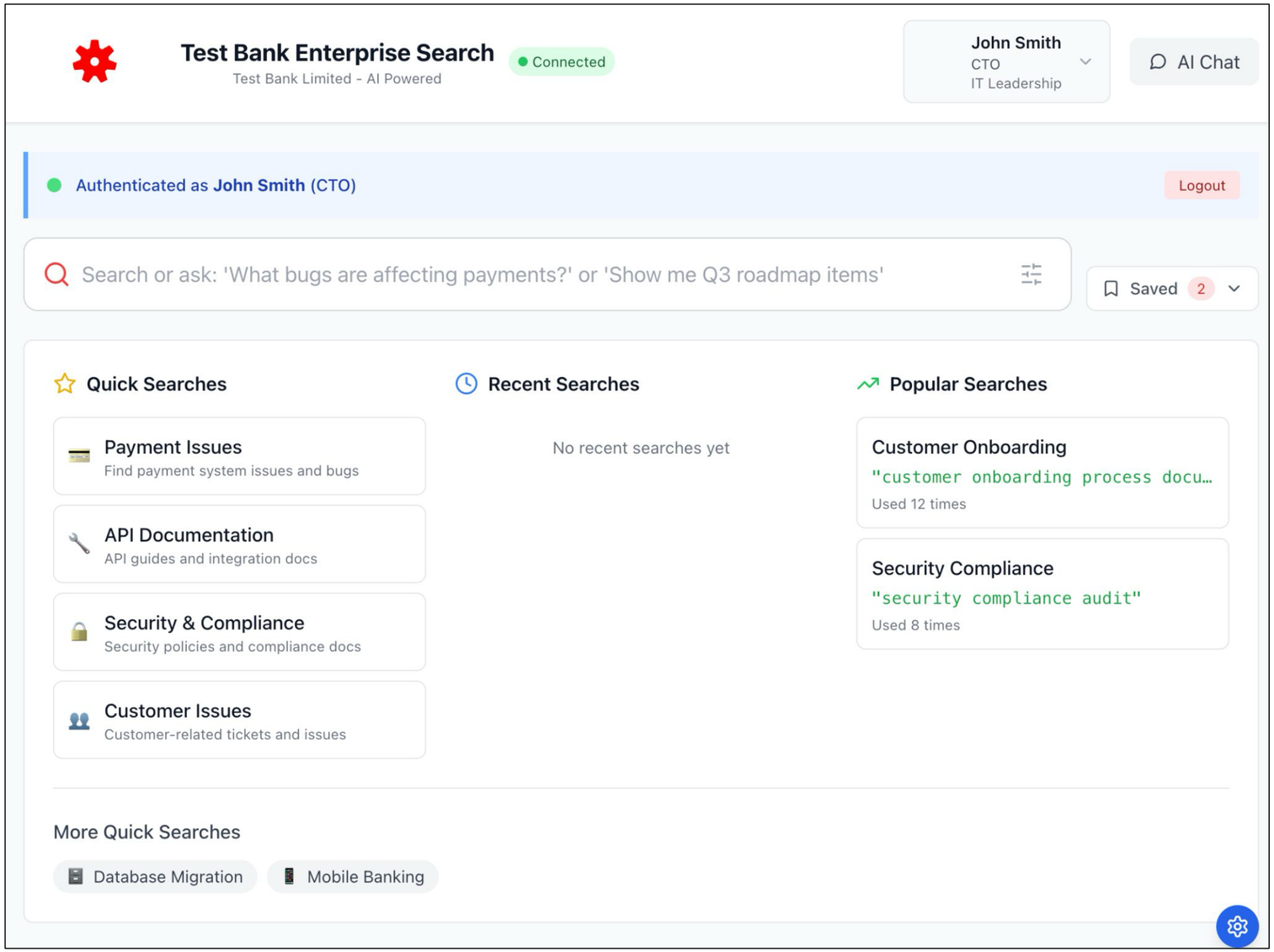
Peter Steenbergen

GenAI Search
Specialist

elastic

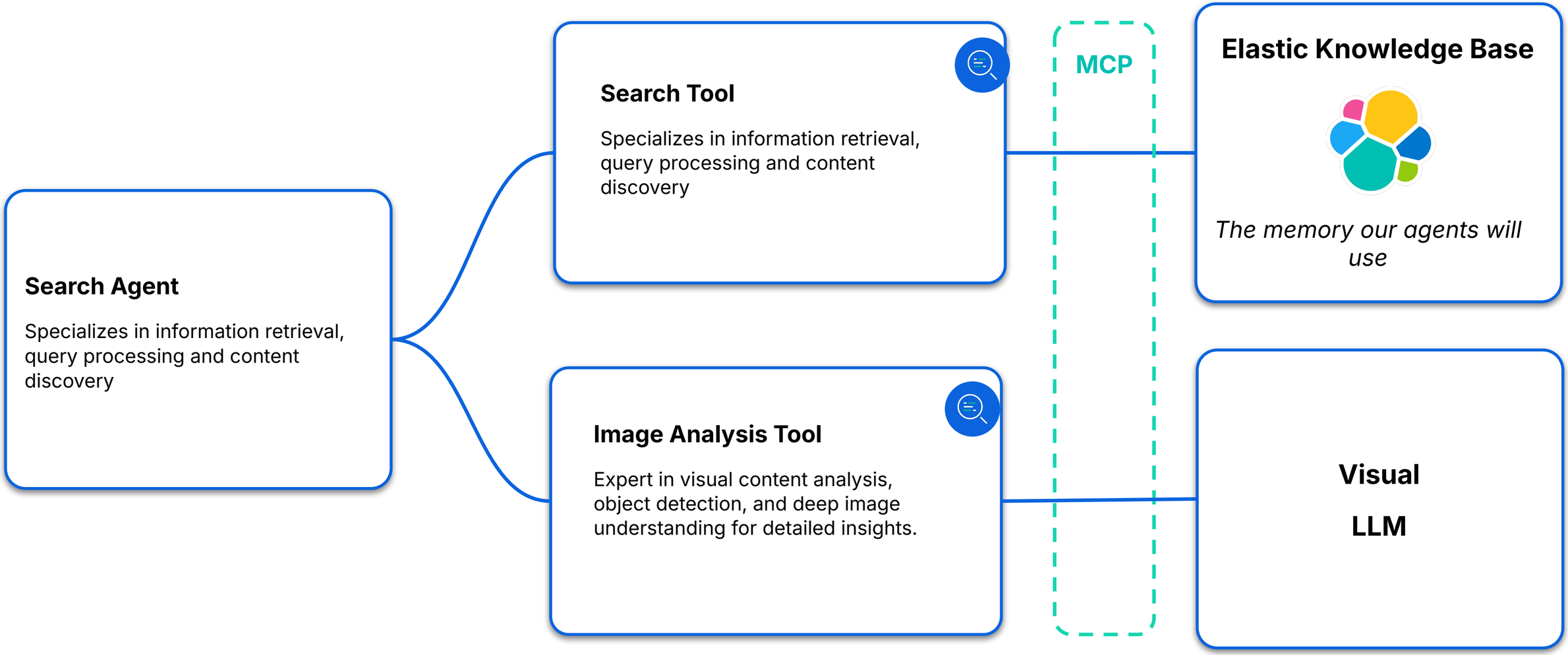
As a developer you are asked to create a new search for a repository of technical documents. How would you do this?

What does a typical Enterprise search look like

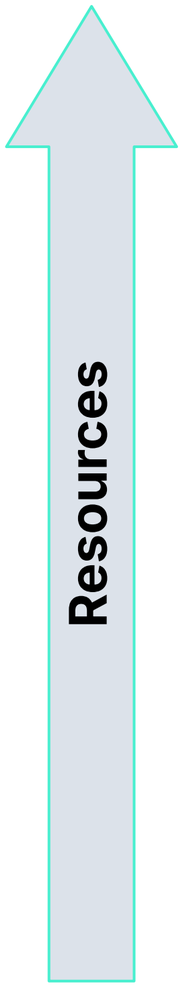


elastic

Elastic as Knowledge base used by agents

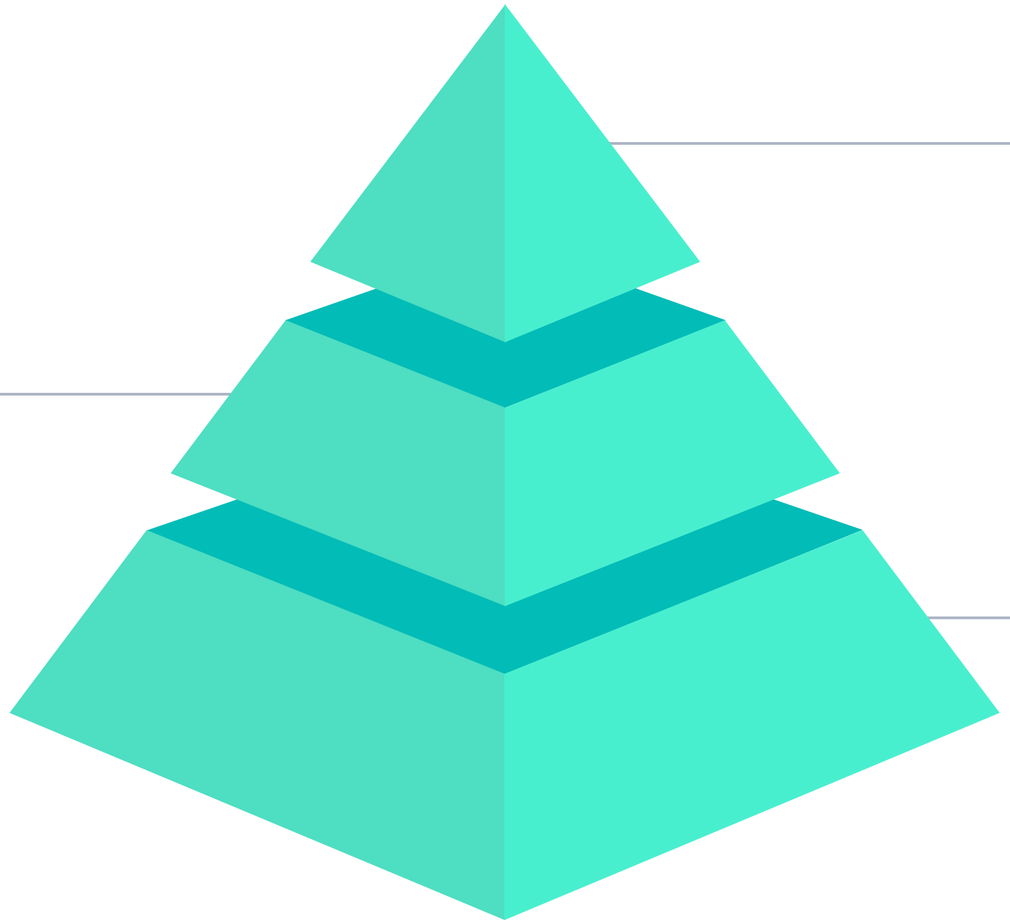


Let’s start with a Search Engine



Fine tuning
THOUSANDS

- Query Rules
- Query Rescorers
- Algorithm Rerankers
- Learning To Rank (LTR)



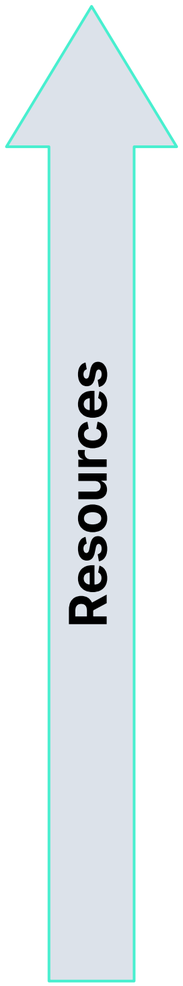
Late stage reranking
HUNDREDS

- Semantic ReRanking
- Multi-Vector Rescore

Retrieval
BILLIONS+

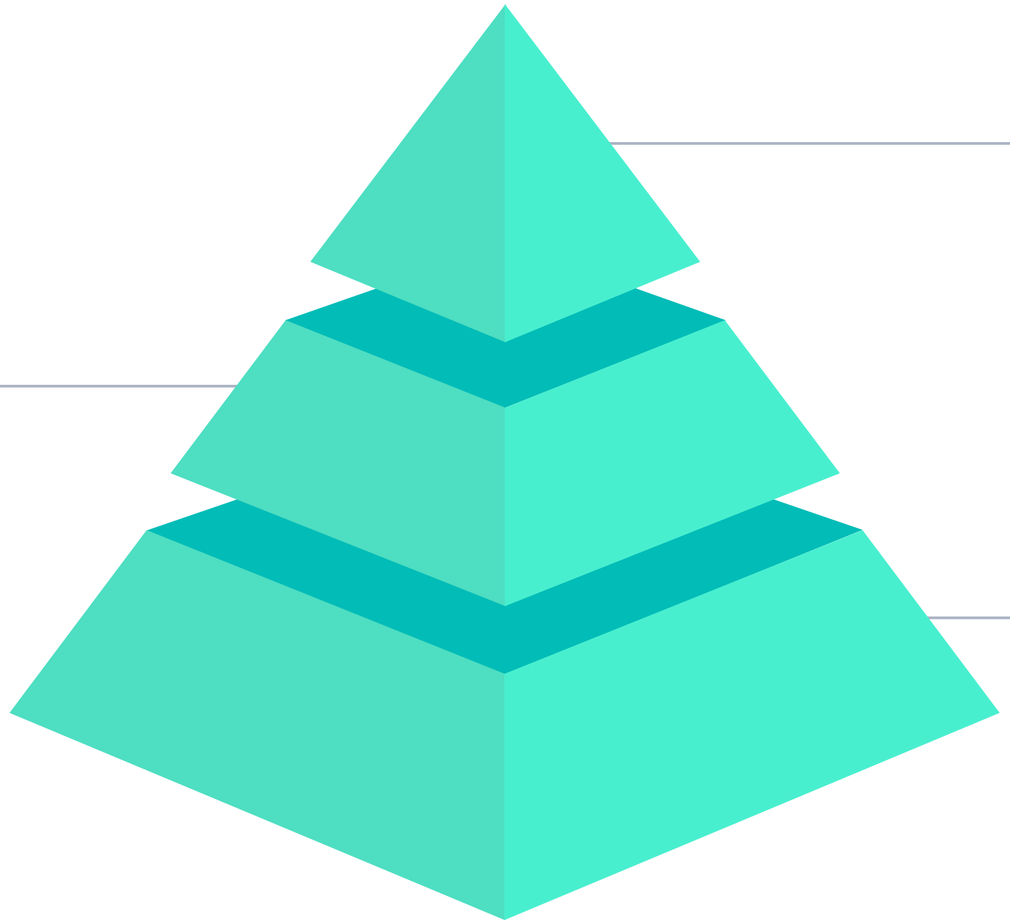
- BM25
- Vector
- Search Templates
- Ingest Pipelines
- _inference APIs
- Semantic_text
- Rank Features

Let’s start with a Search Engine



Fine tuning
THOUSANDS

- Personalise results
- Boost certain docs



Late stage reranking
HUNDREDS

- Recommend based on who I am!

Retrieval
BILLIONS+

- Overall document search based on what is typed in the search box

Basic requirements for this?

Keyword - Semantic

Be able to ask the question of search. Ask for an outcome and search is a byproduct.

Image Search

Users want to search for things that are maybe only in the diagrams.

Geolocation

Where is the user searching from?

Personalize

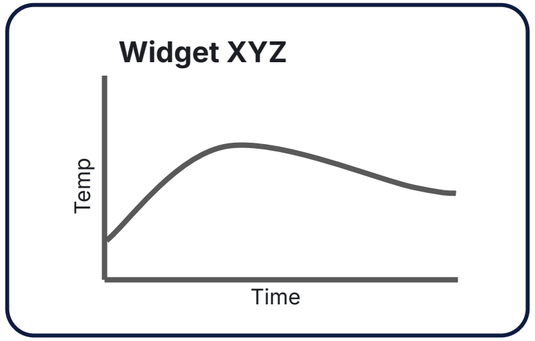
Make sure the results are relevant for me.

Let’s ask some questions

What is the process for releasing new releases into production near me?

The release process is...

What is the temperature of widget “XYZ” after two minutes?



23 Degrees

The Agentic Revolution

We're moving from reactive AI tools to proactive AI partners that understand context, maintain relationships, and drive outcomes independently.

What are AI Agents

- Systems to reason, plan and act
- Go beyond simple RAG and complete complex tasks
- Proactive decision-making partners
- Minimal human oversight

Capabilities

- Tool integration
- Memory
- Dynamic goals
- Context awareness -
- Reasoning & decision-making
- Autonomous execution

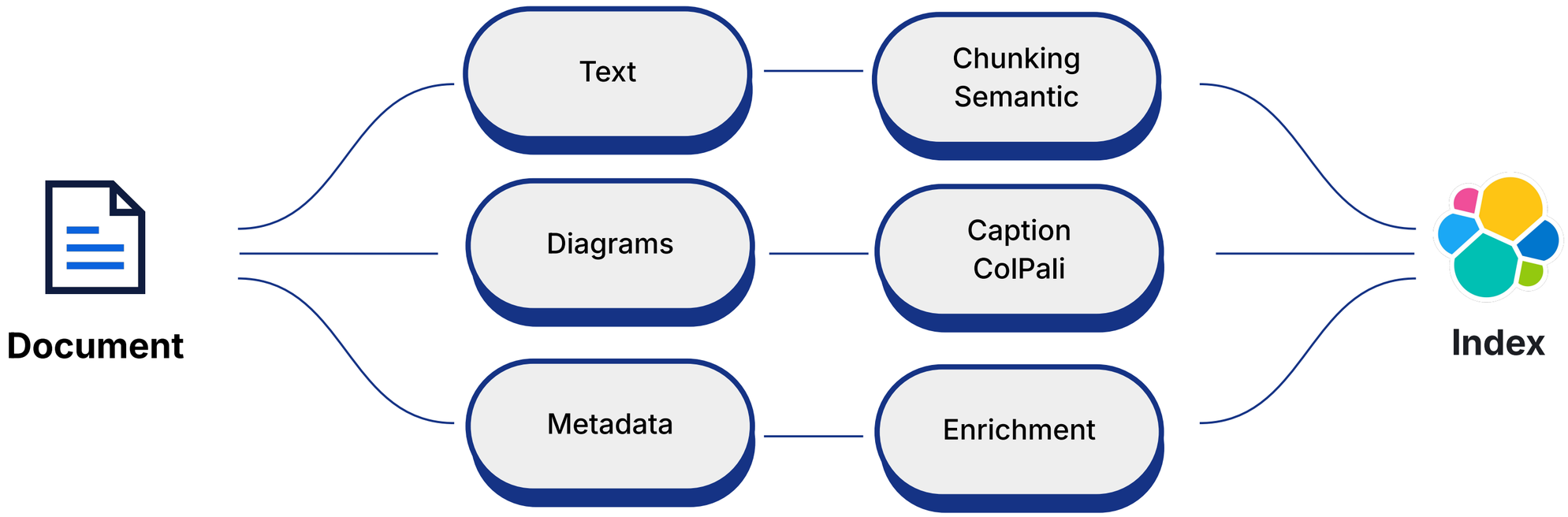
Impact

- 24/7 productivity
- **Faster decision-making** with data-driven insights and autonomous actions
- **Efficiency Gain** by automating complex workflows

Let's first look at how we ingest data

elastic

It all starts on ingestion of the document



Why vectors?

Semantic Understanding

Vectors capture meaning and context, not just exact keyword matches.

Similarity Search

Efficiently find similar items using distance

Multimodal Capabilities

Single vector representation works across text, images, audio, and video

AI-Powered Results:

- Modern embedding models create high-quality vectors that understand nuance, synonyms, and intent - powering recommendation systems and semantic search.
- Real-World Applications

E-commerce: "Find similar products"
Content discovery: "More like this"
RAG systems: Retrieve relevant context for LLMs
Image search: Find visually similar content

Vector search transforms "finding exact matches" into "finding what users actually want."

Why Elastic is a best vector DB choice?

Keeping your data close (Embeddings + Metadata)

Rescoring vectors is needed to improve recall → precision when quantization is aggressive (binary, 2/4 bits)

Advanced Quantizations

- Donated **BBQ** (Better binary quantizations) to Lucene
- 1-bit released, 2/4/8-bit coming
- Save memory by ~95%

Hybrid Search

- Avoid double storage (metadata and content)
- Retrieve the full documents once

A-CORN for Filtered Search(ANN)

- A-CORN adaptive pruning on HNSW graphs
- Reduces candidate explosion

Searching the text is easy!

Automatic embedding and chunking with transparent mapping

Model aware inference integration

Hybrid search ready

What type of **model** do I pick?

What is a **zero shot** model?

How do I turn stuff into **vectors**?

What is a **vector database**?

What is **chunking**?

semantic_text

How do I write a **query** with vectors?

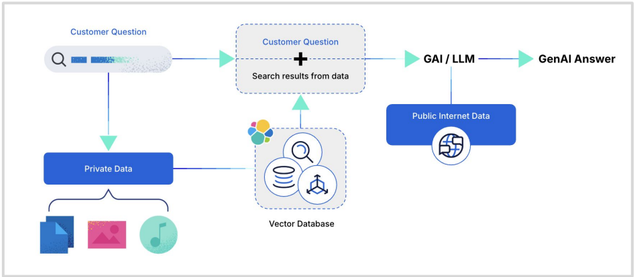
Dense vs sparse vectors?

How many **dimensions**?

What if my chunks are **too big**?

How do we ingest images?

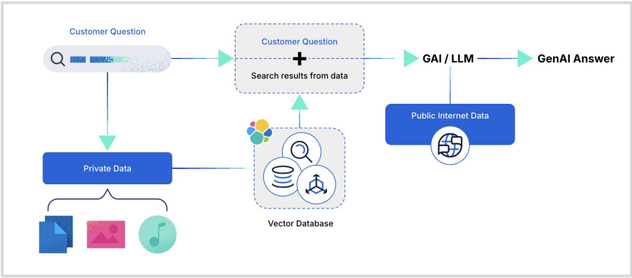
Captioning / Feature Extraction Agent



"A typical RAG model, using Elastic....."

Further information from image URL/Model call

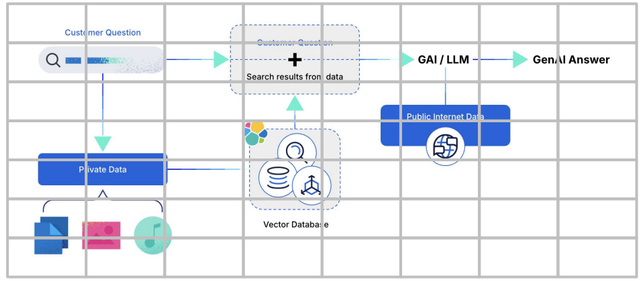
Single Vector + Agent



Vector - {0.373,...}

Further information from image URL/Model call

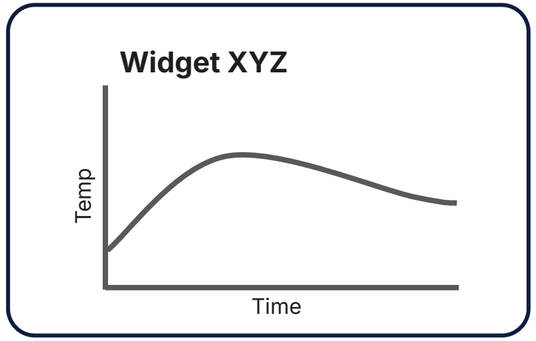
ColPali - Multi Vector



Multiple vectors - {{0.373,...},{0.45,...},...}

elastic

Let’s caption - why?



Gemma 3:
Describe this image

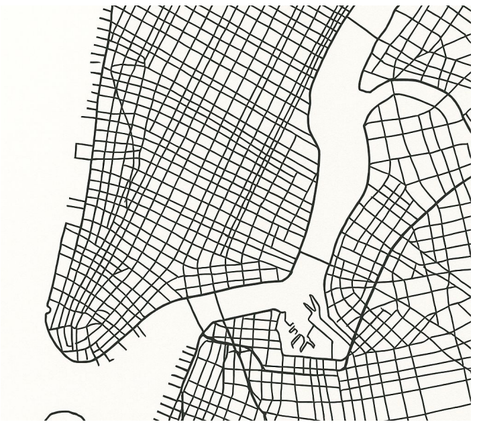
A graph of widget XYZ with temp. vs. time

Embed with model of choice

Index as chunk along with our text

elastic

Let’s talk about metadata



GEO

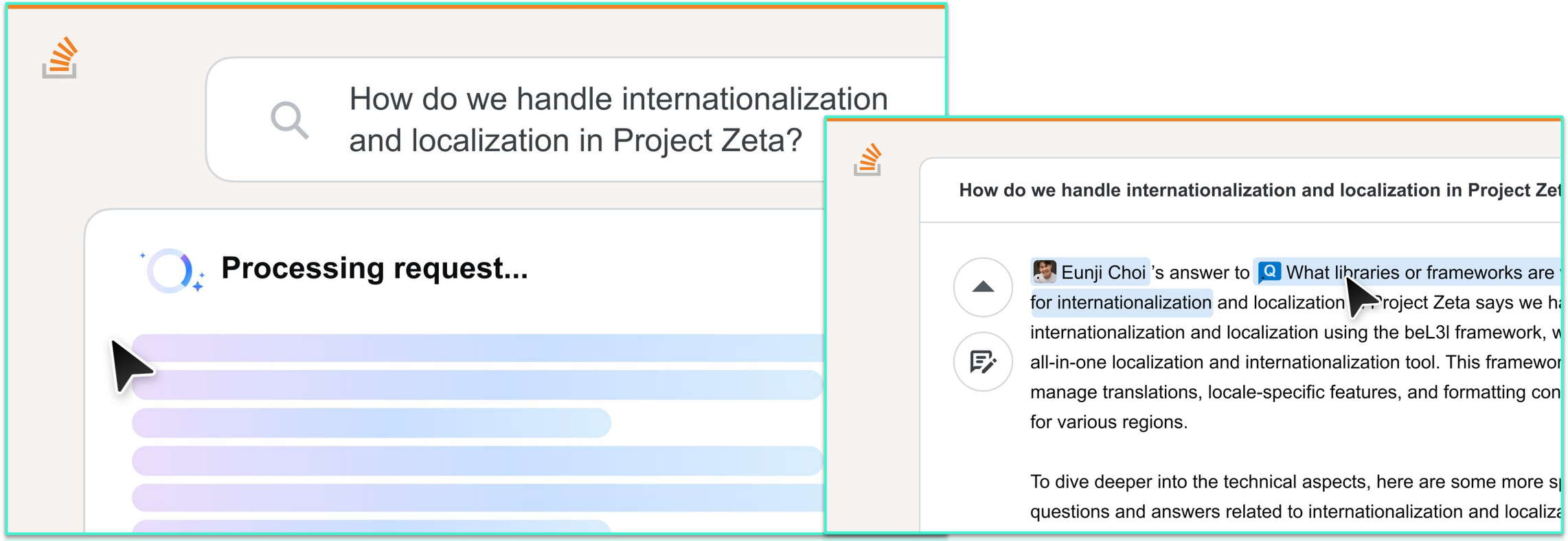
Tagging and Filtering

Document Level Security

Now let's question this data

elastic

We want to be able to ask the questions of what we've ingested



elastic

Using an agentic flow will allow us to ask more complex questions

Question

How many support tickets have been raised for project Alpha?

Prompt

Question
Use Tool

Tool

Aggregation query to Elastic

First we want to search our text

What type of **model** do I pick?

What is a **zero shot** model?

How do I turn stuff into **vectors**?

What is a **vector database**?

semantic_text

What is **chunking**?

How do I write a **query** with vectors?

Dense vs sparse vectors?

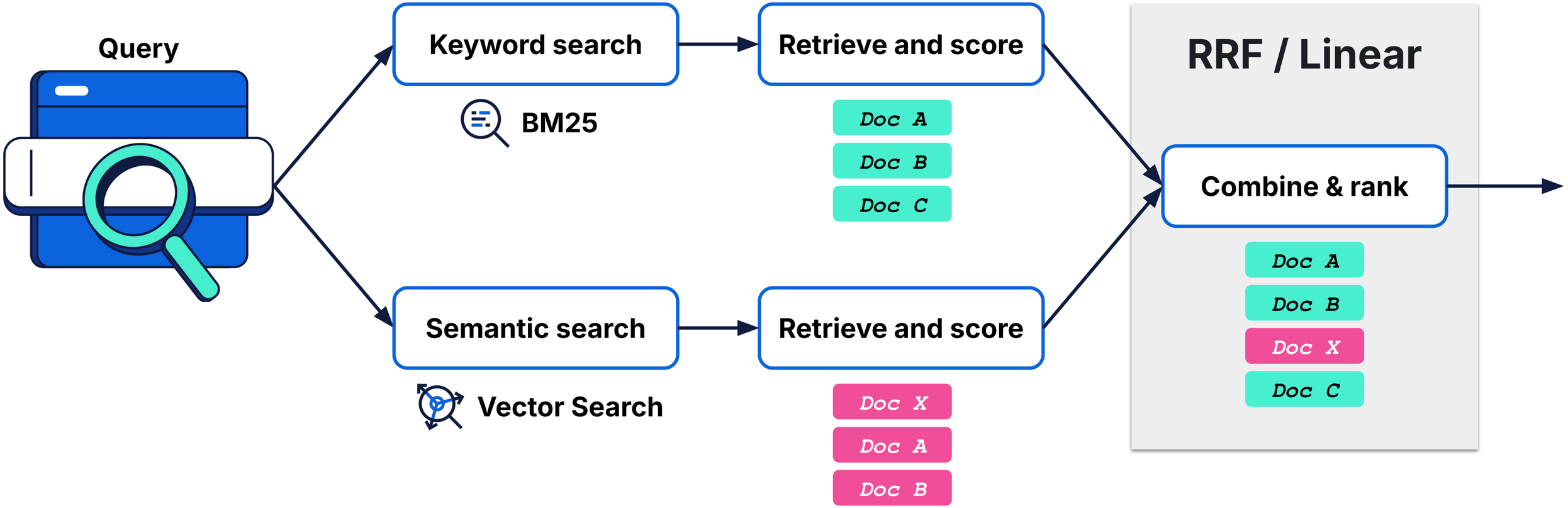
How many **dimensions**?

What if my chunks are **too big**?

But we can also use **Elastic's lexical search** to help us when people search for specific items, like part numbers

Hybrid Search: Combining keywords and meaning

The best of both worlds



Search Tool

Search Tool

Linear Retriever

- Semantic
- Lexical
-

Or RRF, now with weights!

Returns chunks and image urls

Elastic Knowledge Base

The memory our agents will use

How do we question images?

What is the temperature of widget xzy after 2 minutes?

Captioning/Feature Extraction Agent	Single Vector + Agent	ColPali - Multi Vector
Search elastic	Vector - {0.373,...}	Vector - {0.373,...}
Further information from image URL/Model call	Further information from image URL/Model call	Multiple vectors - {{0.373,...},{0.45,...},...}

elastic

How do we ask questions of the images?

Captioning - Fast search, may not even need an LLM

What is the temperature of widget “XYZ” after two minutes?

Hybrid query to find chunk with graph in it

Use VLLM to ask question based on chunk data

23 degrees Celsius

elastic

Image search tool

Image Analysis Tool

Takes a question and a URL

Returns an answer

Visual

LLM

How do we bring this together?

Prompt (for workflow)

```
You are an intelligent search agent that can process user
questions by searching knowledge bases and analyzing images
when relevant. Your goal is to provide comprehensive,
accurate responses by leveraging both textual and visual
information.

Core Workflow

When you receive a user question, follow this process:

1.  Initial Analysis: Understand the user's question and
    determine what type of information they're seeking
2.  Search Execution: Use the search tool to query the
    knowledge base
3.  Result Processing: Examine search results for both
    textual content and image URLs
4.  Image Analysis: If image URLs are found, analyze them
    using the image analysis tool with the original question
    context
5.  Response Synthesis: Combine all findings into a
    coherent, comprehensive response

Use search_tool

Use image_analysis_tool
```

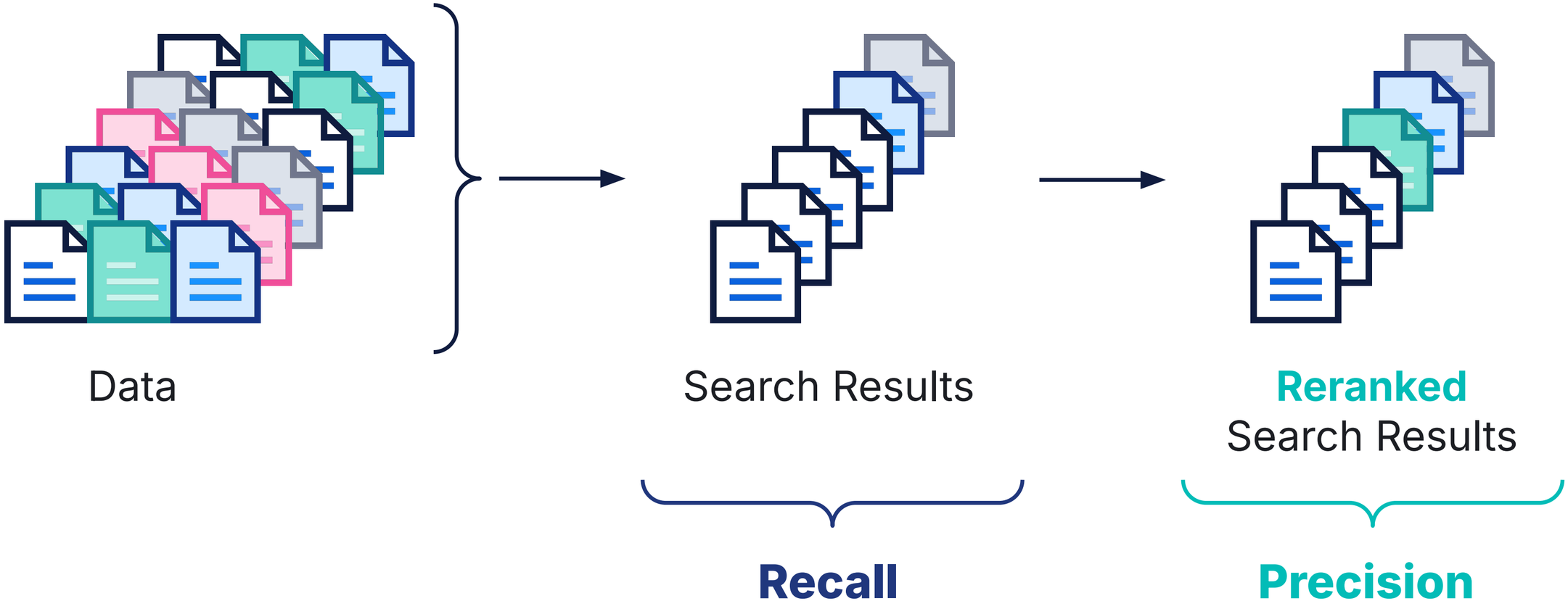
Search Tool

Image Analysis tool

Let enhance our search relevance

elastic

Why use a reranker to tune relevance?



Rank Features

Document-level ranking signals

Elasticsearch allows you to store relevance features per document and boost document scores at query time for enhanced search results.

Popularity & Engagement

Products with many views or purchases can automatically rank higher for relevant queries.

Personalized Scoring Control

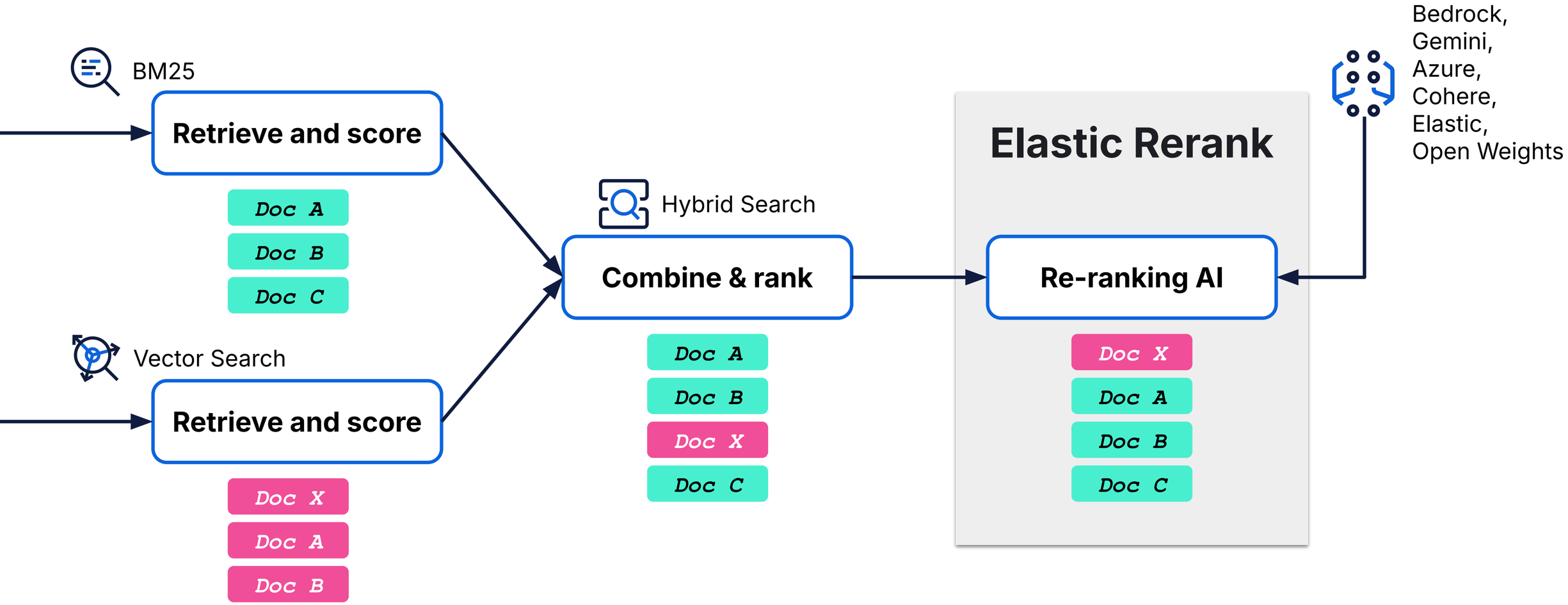
Gain granular control over how scores are calculated, allowing for custom ranking logic and personalized experiences.

Contextual Boosting

Fine-tune results based on specific query context, user intent, or current trends to hit relevance targets.

Semantic re-ranking

The final step for the best results



Open Q&A!

elastic | The Search AI Company

Thank you!

elastic | The Search AI Company