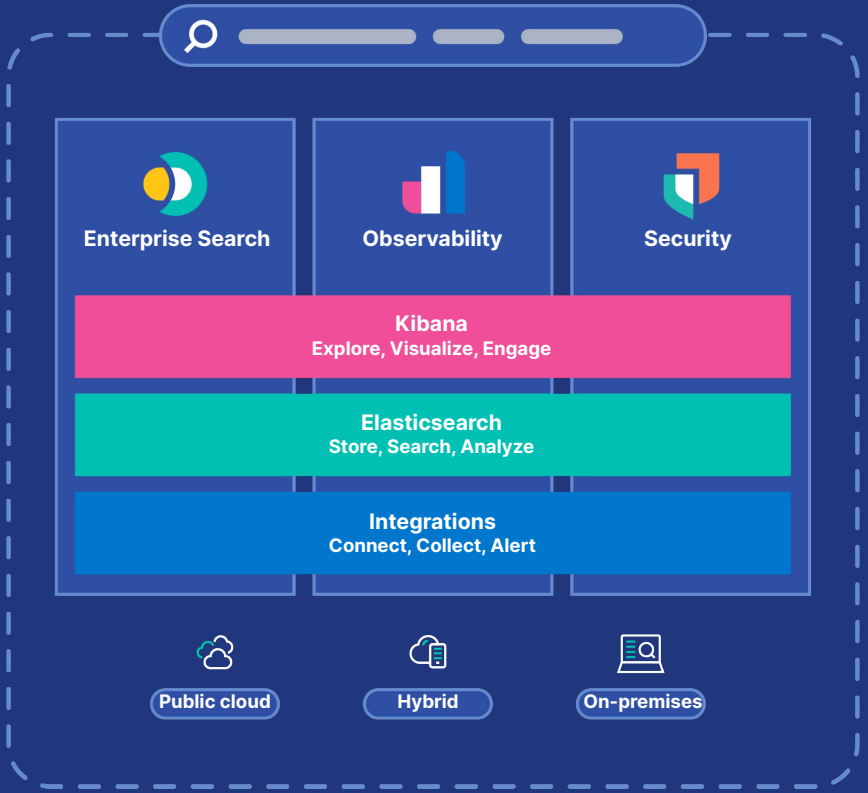
The Elastic Search Platform



elastic

Let's Create a Search Experience

from scratch!

elastic

Let's Create a Search Experience!

Let's Create a Search Experience

That's a lot of responsibility!

- 88% of online visitors won't return to a site after a bad experience
- 1.8 hours wasted by employees searching for information every day
- 76% online shoppers abandon a retail website after an unsuccessful search
- 43% of site visitors go immediately to search boxes
- Shoppers using search spend 2.6x more

Usability Geek, McKinsey, Harris Poll, Forrester research, Salesforce

Choosing Elasticsearch

Speed

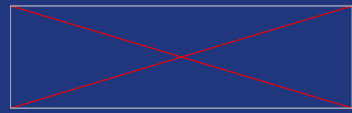
Find matches in milliseconds within structured and unstructured datasets

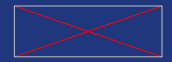
Scale

Scale massively and horizontally across hundreds of systems

Relevance

Generate highly relevant results and actionable insights from data





Let's Create a Search Experience

Enterprise Search helps you building and tuning your Search Experiences faster

Enterprise Search helps you with:

- Ingesting Searchable Data
- Search Relevance
- User Experience
- Feedback and Tuning

Elastic Enterprise Search

elastic

Ingesting Searchable Data

Ingesting Searchable Data

Get data into Elasticsearch that we'll be able to search

- Identify our Data Sources
- Extract Data
- Enrich Data
- Store Data for Search

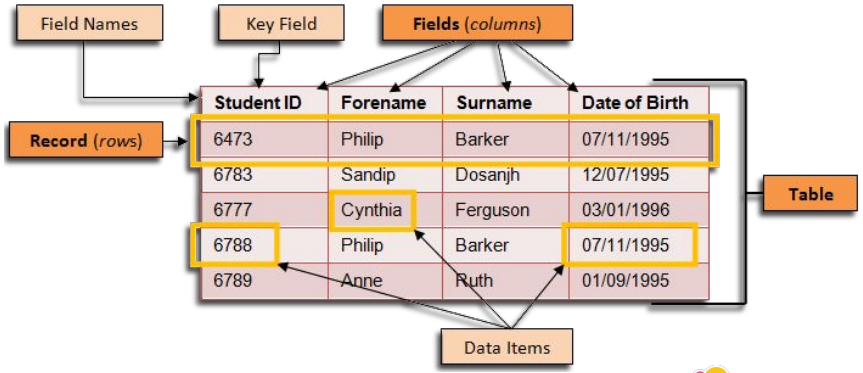
elastic

Ingesting Searchable Data

Identify our Data Sources

- RDBMS
- Files
- NoSQL
- Blob storage
- CMS
- Websites
- etc...

- What data to import?
  - All docs?
  - All fields?
- Resulting schema



elastic

Ingesting Searchable Data

Extract Data

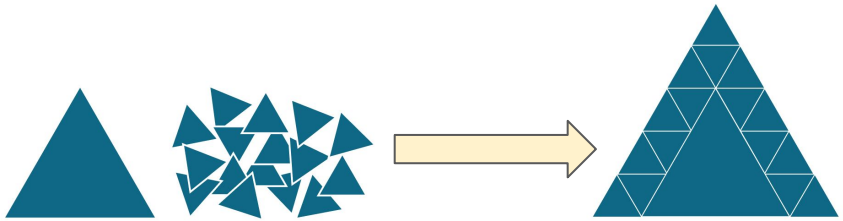
- Need specific methods for retrieving data
  - Client libraries
  - APIs
  - File traversal
- Authentication and Authorization
- Select data we want to import
- Schedule extraction
  - Full extraction
  - Refreshes

elastic

Ingesting Searchable Data

Enrich Data

- Format / Trim content
- Binary content extraction (PDF)
- Cross reference other sources
- Machine Learning
  - Sentiment Analysis
  - Named Entity Recognition
  - Text classification
  - Text embedding for Vector Search



elastic

Ingesting Searchable Data

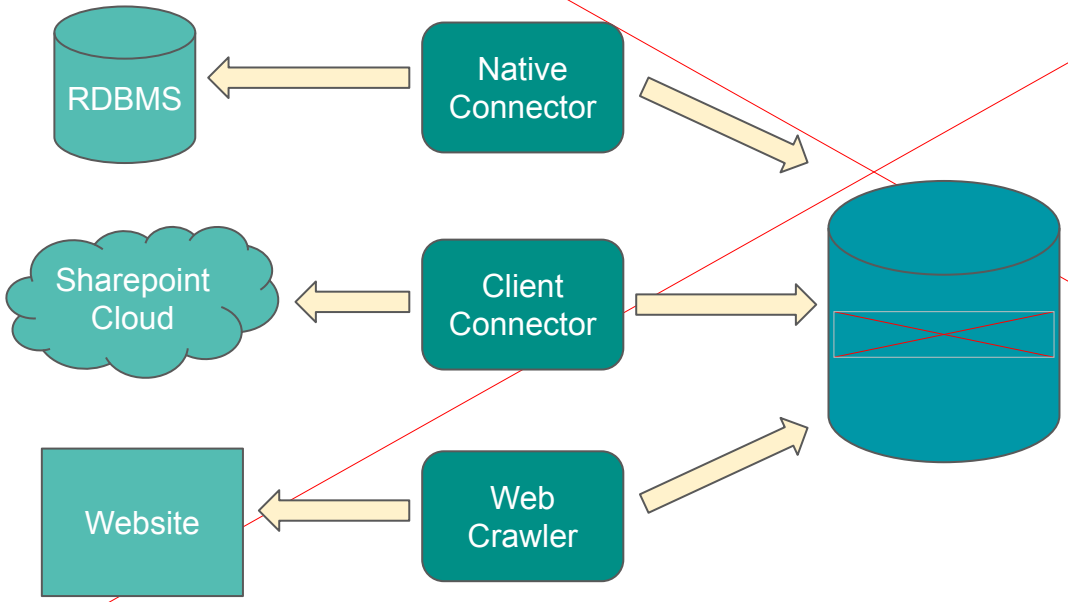
Store Data for Search

- Store in our searchable data store
- Make sure it is ready for searching
  - Stopwords removal
  - Stemming (language dependent!)
  - Tokenizing

elastic

Enter Connectors!

Connectors ingest data into Elasticsearch from data sources

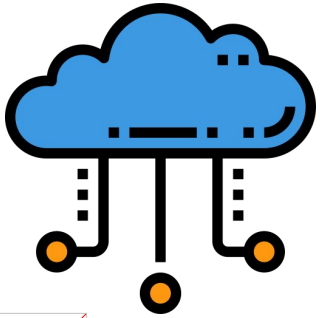


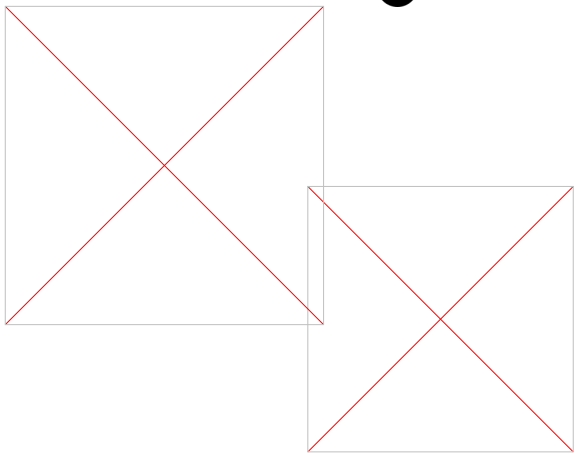
elastic

Connectors - Data Sources

Different types of connectors

- **Native**: Cloud ready
  - MongoDB
  - MySQL
- **Web Crawler**
- **Connector Clients**: Open Source
  - Deployed by you on your infrastructure
  - Connectors Framework: Python, Ruby
  - Object Storage (S3, Azure Blob, GCS)
  - Databases (PostgreSQL, Oracle, MySQL, MSSQL)
  - MongoDB



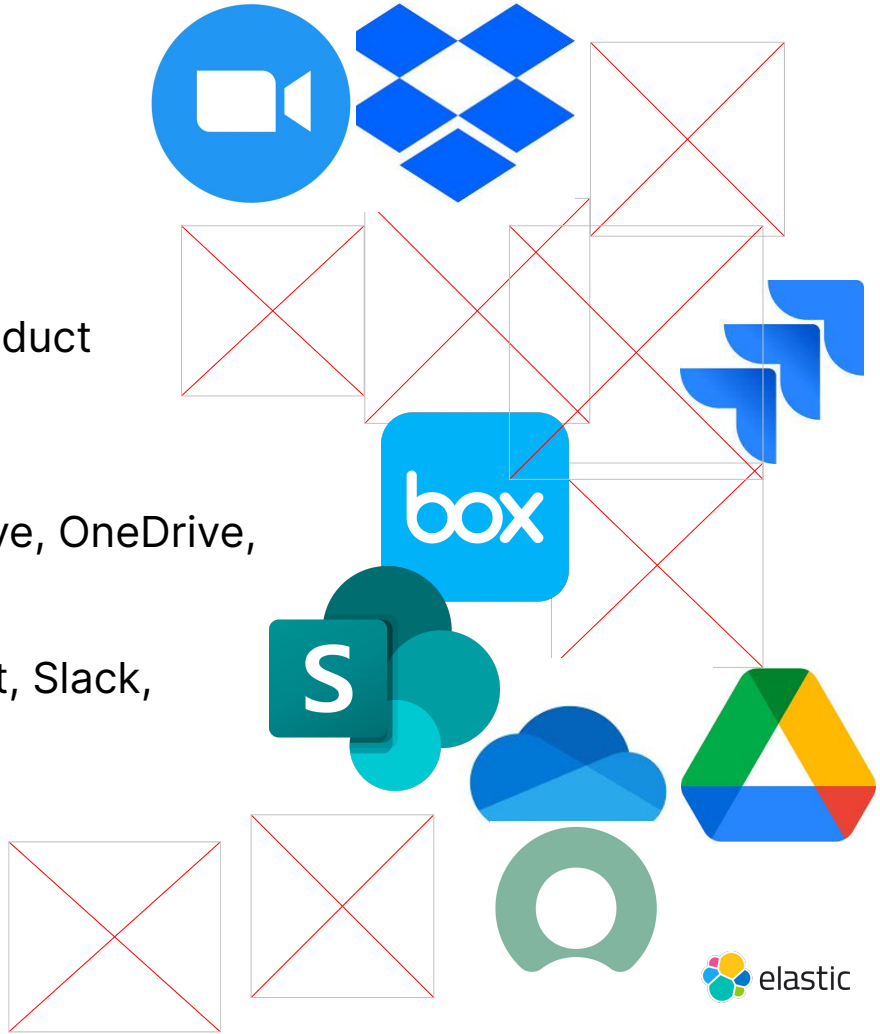


elastic

Connectors - Data Sources

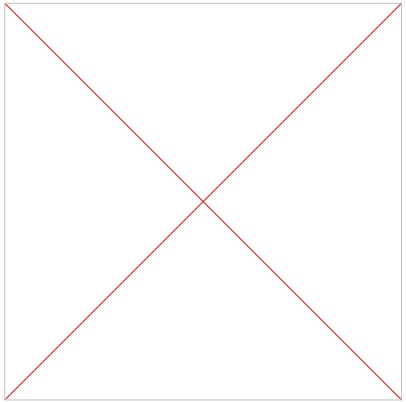
Different types of connectors

- **Workplace Search**
  - Available through Workplace Search product
  - Confluence (Cloud, Server)
  - Jira (Cloud, Server)
  - Github
  - Document Storage (Dropbox, Box, GDrive, OneDrive, Network drives)
  - Mail (GMail, Outlook)
  - Other Services (ServiceNow, SharePoint, Slack, Teams, Zendesk, Zoom)



Connectors - Extract Data

- Authentication and Authorization
- Select data we want to import
  - Sync rules
- Manage syncs from Kibana
  - Once
  - Recurring
  - Cancel
  - Check status

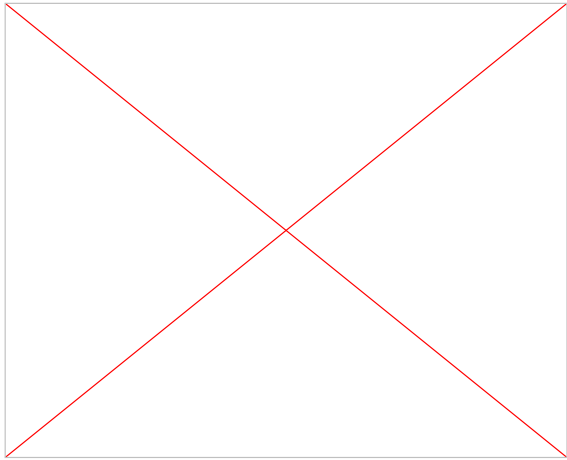


elastic

Connectors - Enrich Data

Connectors can create Ingest Pipelines to enrich data

- Binary content extraction (PDF)
- Machine Learning Inference
  - Transformer models that conform to the standard BERT model
  - Third Parties
  - Train your own
  - ELSER
- Custom Pipelines

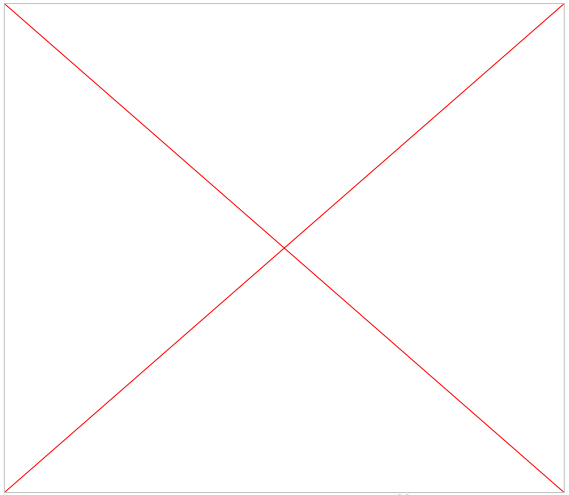


elastic

Connectors - Store Data

Use default text mappings to ensure compatibility with App Search Engines

- Language optimized
  - Stemming
  - Stop words
- Autocomplete ready
- Delimiters / Joined



elastic

Search Relevance

elastic

Search Relevance

Retrieve the best possible search results

- Typo tolerance
- Field weights
- Boosting
- Synonyms
- Promote and hide results

elastic

Search Relevance - Typo tolerance

Retrieve results for misspellings

- Added, changed or missing characters (ifone)
- Contractions (i phone)
- Delimiters (i-phone)

elastic

Search Relevance - Field weights and boosts

Ensure fields get proper importance

- Example: Title vs description

Boost specific values

- Exact value
- Distance to a point
- Function over numbers, dates or locations

Search Relevance - Synonyms

Synonyms are important for relevance

- E-commerce: PlayStation, PS
- Technical: TypeScript, TS

elastic

Search Relevance - Result promotion

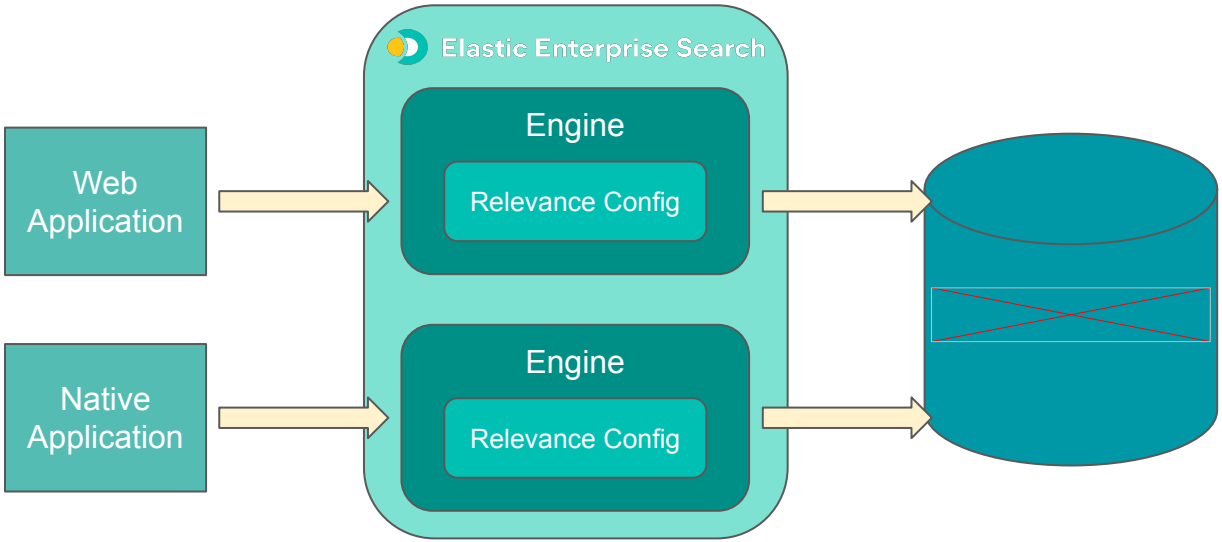
We may want to manually promote results, or hide them

- Campaigns
- Paid results
- Business knowledge

elastic

Enter App Search!

App Search simplifies searches and relevance management



App Search - Dynamic Settings

App Search allows non-technical users to manage relevance settings

- Removes developers from the relevance loop
- Check before going live
- Allow different permissions for engines

elastic

App Search - Typo tolerance

Precision Tuning manages:

- Fuzzy search for typo tolerance
- Hyphenations
- Contractions
- Prefix matching
- Minimum terms to match

elastic

App Search - Field weights and boosts

- Text fields can be assigned weight
- Field Boosts
  - Value
  - Proximity (geo, date fields)
  - Function (gaussian, exponential, linear)

elastic

App Search - Synonyms

Synonyms are dynamic

- No need to change analyzers
- No need to update synonym files in Elasticsearch

elastic

App Search - Result promotion

For a query:

- Set results appearing at the top
- Hide results
- Add the organic results to the mix

elastic

User Experience

User Experience

User expectations

- Result display
- Pagination
- Highlighting
- Autocomplete
- Sorting
- Filtering
- Faceting

elastic

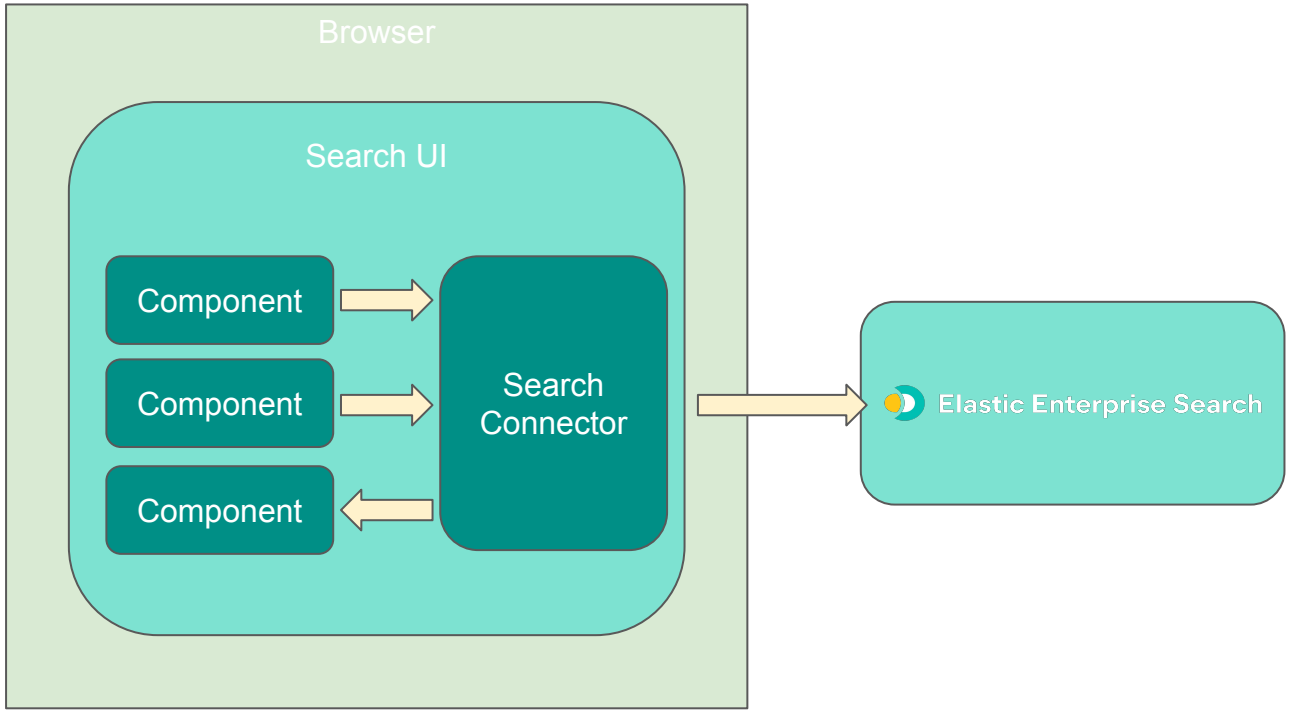
User Experience - Web Apps

Need to integrate our search experience in a web application

elastic

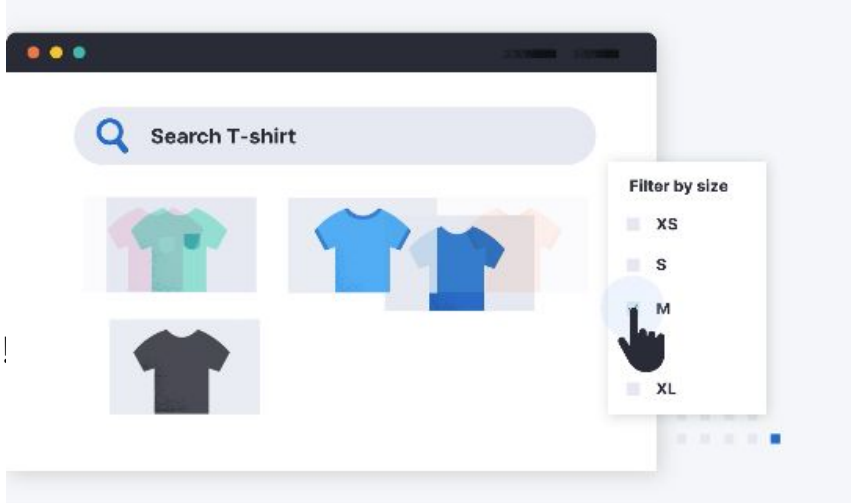
Enter Search UI!

Search UI is a search interface framework



Search UI

- Open Source
- Interoperability
  - App Search
  - Elasticsearch
- Compatibility
  - Vanilla JS
  - React
- Complete Search features
  - Filtering, faceting, sorting
  - Autocomplete, search-as-you-type
- Customizable
  - Download a pre-built experience from App Search!



elastic

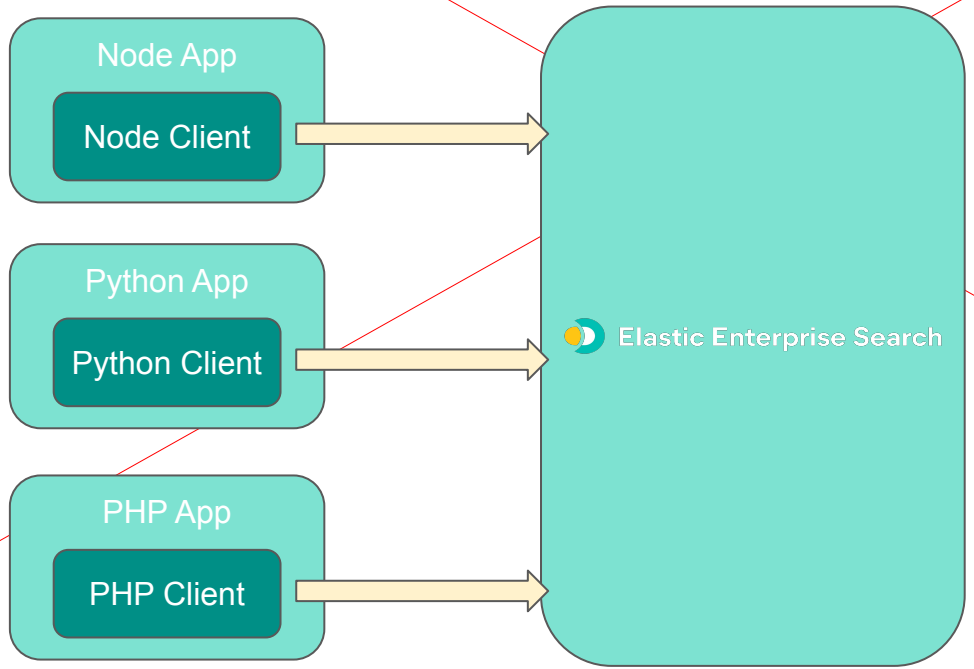
User Experience - Non-Web Apps

Need to integrate our search experience in a non-web application

- Native Applications (iOS, Android)
- Backend / CLI

Enter Language Clients!

Clients for accessing App Search in different programming languages



elastic

Language Clients

Allow interoperability with App Search...

- Node.js
- PHP
- Python
- Ruby

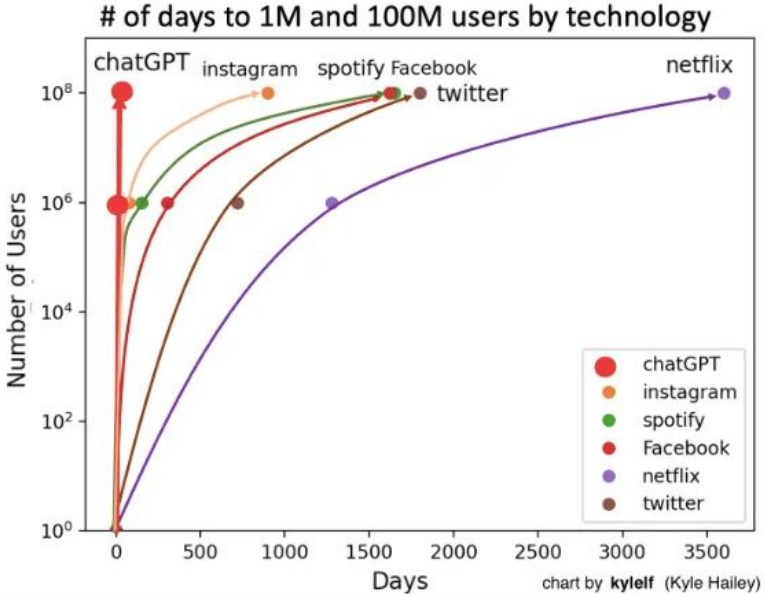
... and additional ones for Elasticsearch

- Java
- Go
- .NET
- Perl
- eland

Elasticsearch Relevance Engine (ESRE)

ChatGPT:
Elastic and LLM

The sudden rise of LLMs and generative AI



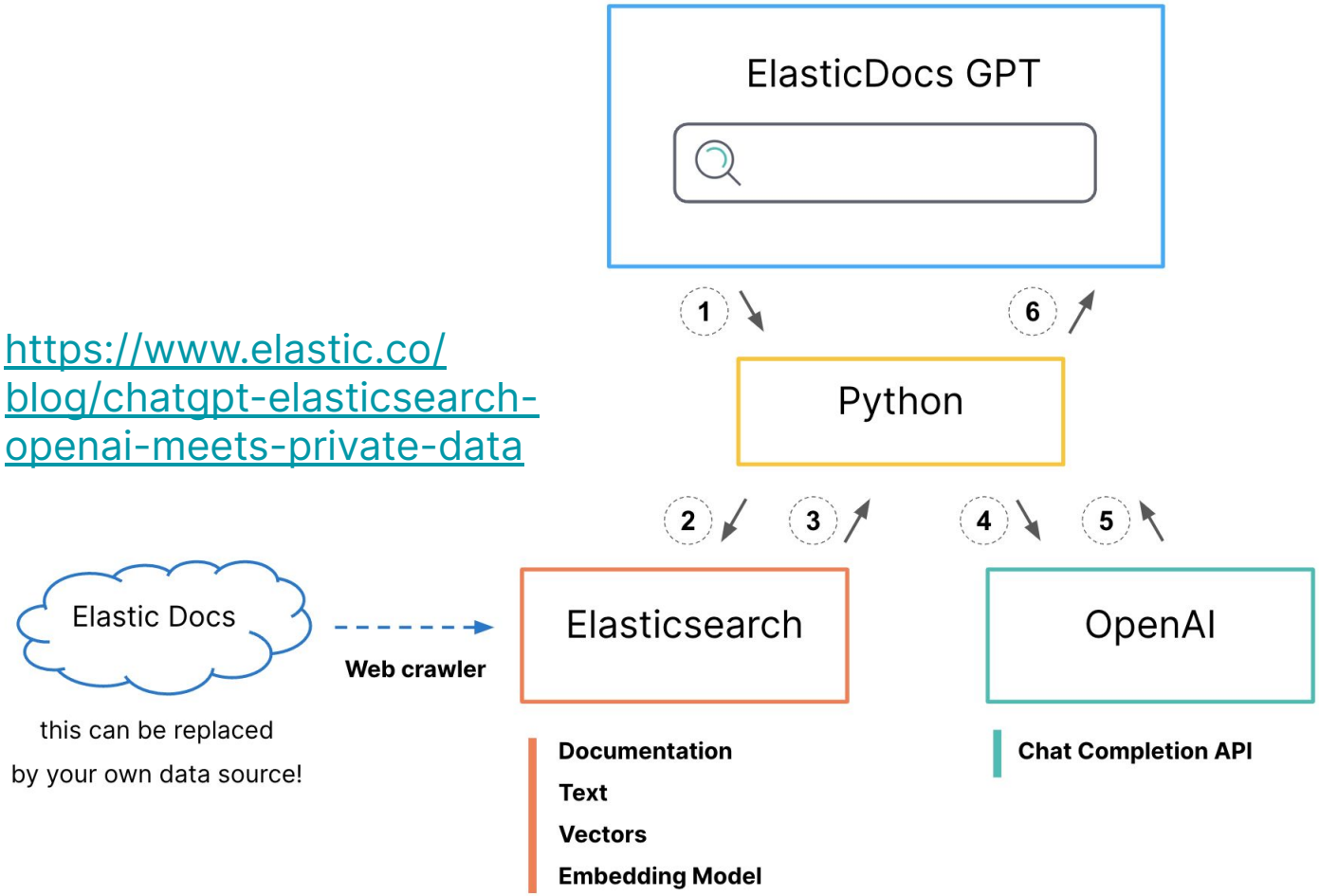
elastic

What’s (Not) Great about ChatGPT?

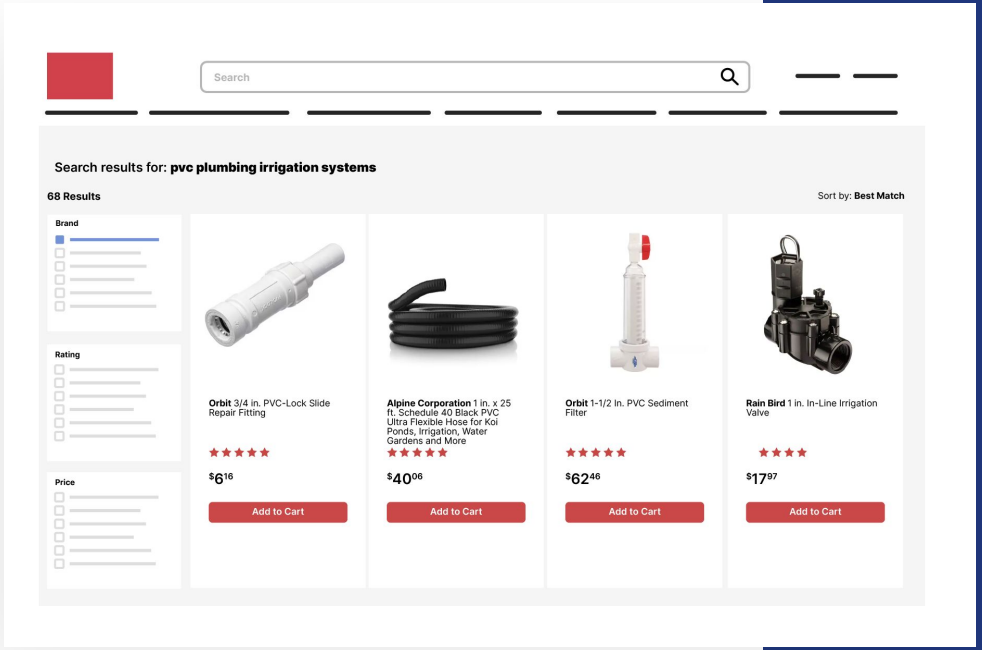
What’s (Not) Great about ChatGPT?

An error occurred. Either the engine you requested does not exist or there was another issue processing your request. If this issue persists please contact us through our help center at help.openai.com.

https://www.elastic.co/blog/chatgpt-elasticsearch-openai-meets-private-data



Search, very soon.



TODAY

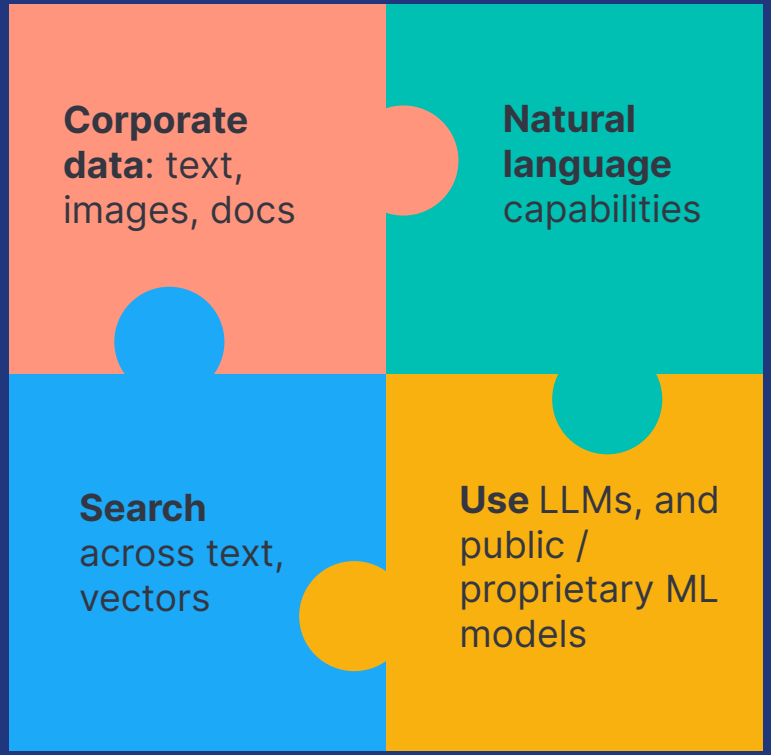
pvc plumbing irrigation systems

TOMORROW

What material list and tools do I need to build an irrigation system for my 1 acre back yard in Detroit, MI?

elastic

The ingredients enterprises need for **AI-search**



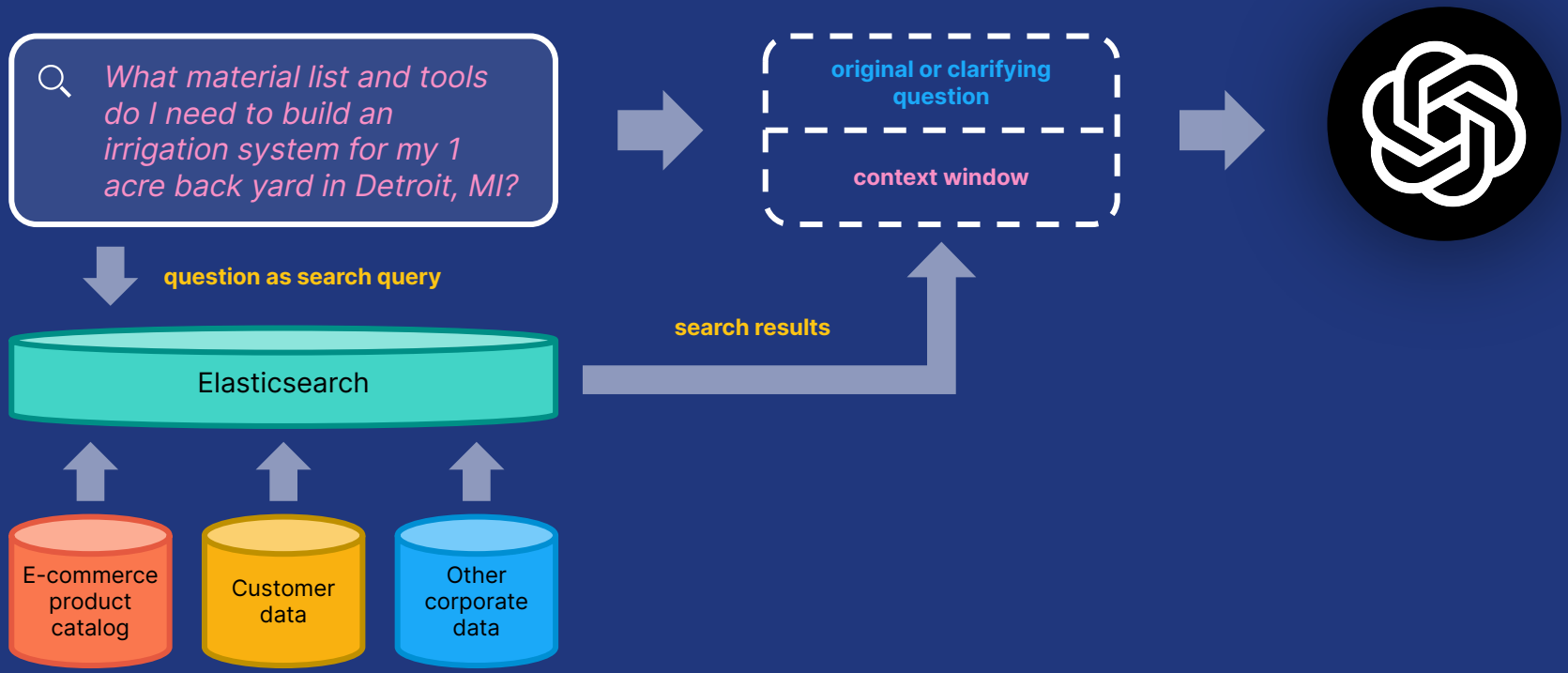
elastic

Enterprise adoption of LLMs **has some barriers**

Clear opportunity, but know before you go:

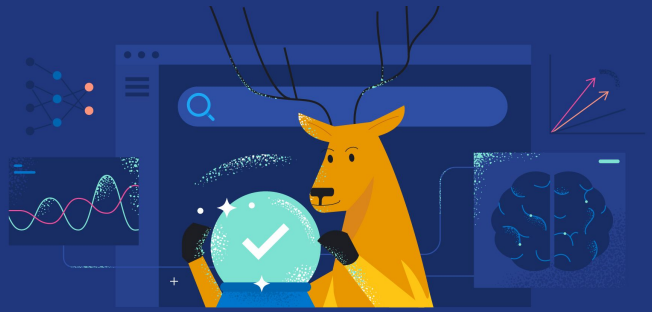
- **Enterprise data, context aware**
  May not have domain-specific knowledge, depends on data set the model is trained on
- **Timeliness & relevance**
  Models are only as fresh and relevant as the data they are trained on
- **Size and cost**
  LLMs can be prohibitive due to data volumes, required computing power and memory
- **Hallucinations**
  May invent facts that sound trustworthy but are projections

Context-window at work



elastic

ESRE



- **Elasticsearch Relevance Engine™ - ESRE (pronounced ez-ray):** Designed to power AI search applications
- **Highlights Elastic's Leadership:** Builds on two years of research and development
- **Developer tools for all:** Gives developers a range of tools for integrating with LLMs and creating highly relevant AI search applications [regardless of ML resources, skills]
- Available today for **Platinum** and **Enterprise** customers

Elasticsearch Relevance Engine™

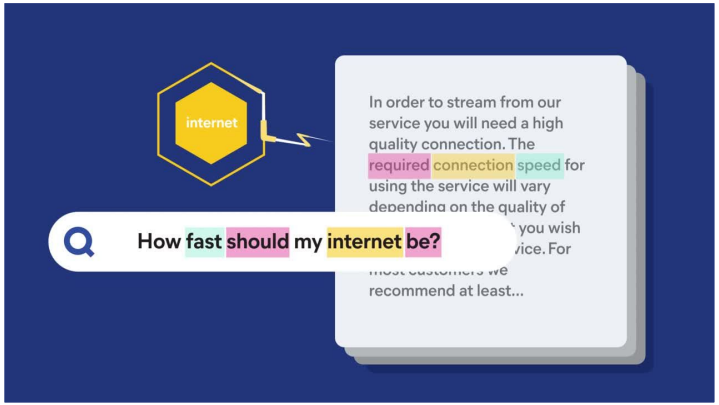
Gives developers a powerful set of **machine learning tools** to **build AI-powered search applications** that **integrate with large language models**

- Vector database
- Ability to host your own transformer model
- Ability to integrate with 3rd party transformer models (OpenAI)
- RRF - hybrid scoring model (vector & textual search)
- Elastic's proprietary ML model
- Integration with 3rd party tooling like LangChain

Reflects two years of R&D

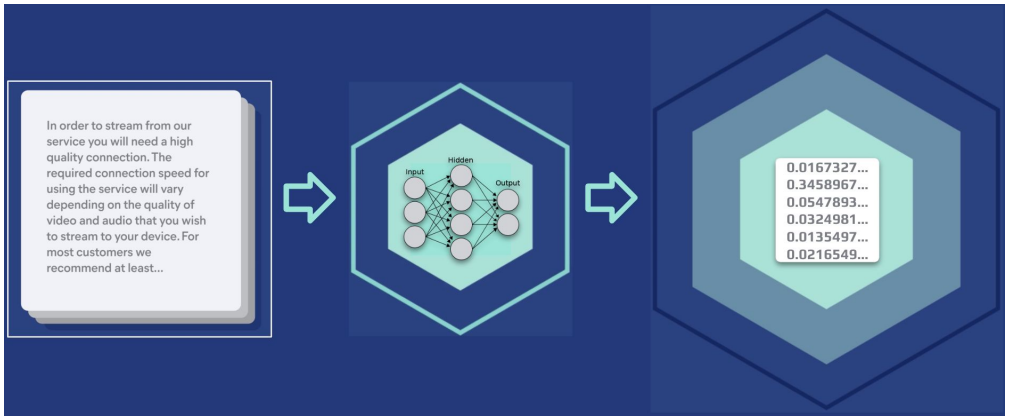
Vector search - for semantic relevance

- Traditional ‘keyword’ search: mentions, lexical similarity, frequency
- However, user behaviours are shifting:
  - ”How fast should my internet be?”
- Enter *vectors* - search terms stored as numbers
  - *vector similarity* or numerical distance search
- *Vectorizing* works on text, imagery, unstructured data



Vector database - storage optimised for vector operations

- Efficiently store numerical vectors, perform CRUD operations
- Fast search recall for vectors
- Purpose built for enabling vector search at scale
- **Elastic is a *superset* of vector database capabilities**



Elasticsearch is a **superset** of vector database capabilities

Elasticsearch is capable of:

- **Creating vector embeddings** (numeric representations of data)
- Is **optimised for storing** sparse and dense vectors, at scale

Elasticsearch also has:

- **Filters and aggregations:** over result set in vector database
- **Document level security built in:** used in production by enterprise customers
- **Data types:** Geo, full text, multiple languages
- **Ingestion tooling:** connectors, API, web-crawler and 3rd party integrations
- That's not all!.. Community, At scale enterprise adoption, Track record, Elastic stack..

Elastic Learned Sparse Encoder - out of the box relevance, across domains

- Allows devs to implement semantic search without the need to train their own model
- Out-performs on: question-answer pairs, weather records, medical text
- Generalises across domains without training
- Great relevance, out of the box

Search results ranking models

Data sets (BEIR benchmark)

	Average	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA	ArguAna	Touche-2020	DBPedia	SCIDOCS	FEVER	Climate-FEVER	SciFact
BM25	0.416	0.688	0.327	0.326	0.602	0.254	0.472	0.347	0.287	0.165	0.649	0.186	0.69
RRF (BM25/Dense)	0.449	0.697	0.317	0.445	0.611	0.318	0.474	**0.354**	0.353	0.159	0.746	0.238	0.671
Linear (BM25/Dense)	0.471	0.787	0.335	0.485	0.62	0.341	0.444	0.346	0.378	0.164	**0.778**	**0.272**	0.698
SPLADE	N/A	0.726	0.347	**0.537**	0.687	**0.347**	**0.526**	0.246	**0.436**	0.158			0.703
Elastic Learned Sparse Encoder	0.471	0.747	**0.351**	0.524	0.67	0.339	0.5	0.263	0.415	0.156	0.777	0.218	0.695
RRF (BM25 / Elastic Learned Sparse Encoder)	**0.478**	**0.797**	**0.352**	0.468	0.674	0.311	0.497	0.347	0.411	**0.166**	0.762	0.24	**0.712**

Search quality *worse* than benchmark (BM25)

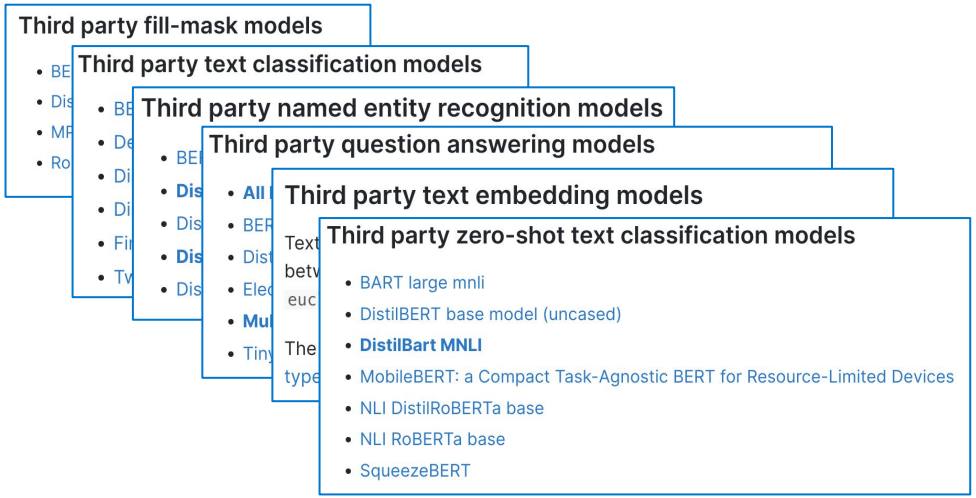
Search quality *better* than benchmark (BM25)

[Check out the public demo]

field enablement

Model management - integrate 3rd party or proprietary models

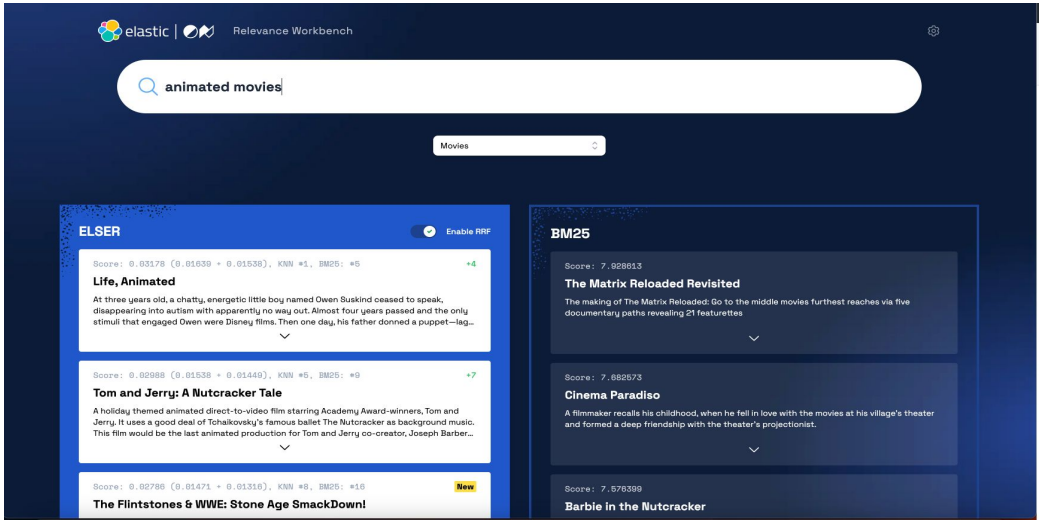
- This is a fast evolving space; flexibility allows you to keep up
- Use 3rd-party PyTorch models (e.g. from repo such as HuggingFace)
- Several types of NLP models supported



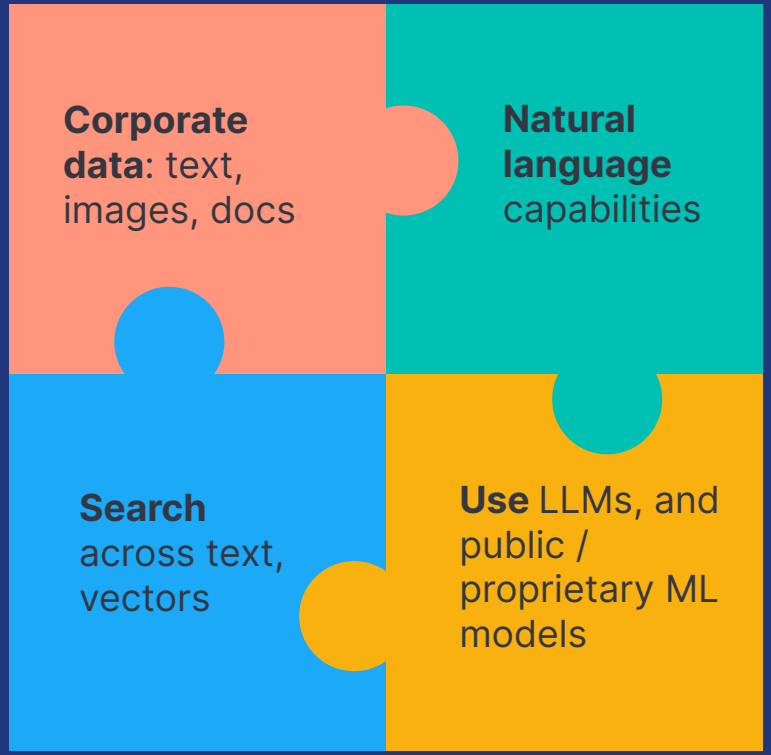
Full list of Elastic supported models:
ela.st/nlp-supported-models

Hybrid retrieval with RRF - reciprocal rank fusion to combine retrieval methodologies

- Hybrid ranking model for textual and vector search
- Combines multiple rankings without search tuning or score normalization
- Gives developers control to optimize search

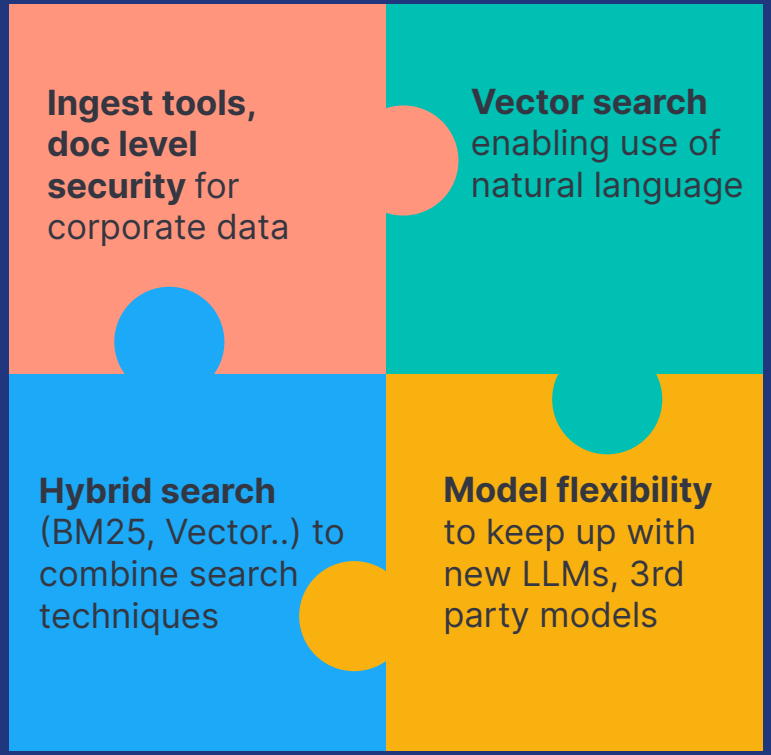


The ingredients enterprises need for **AI-search** ...



elastic

… are the capabilities we ship with **Elasticsearch Relevance Engine**



elastic

Our customers' search relevance journey

Getting started

Advanced

Industry-leading BM25 search provides relevant keyword search out of the box.

Out of the box, cutting edge semantic search that combines textual and vector search to identify query intent and meaning.

EN-only. Works well across domains.

Store numerical representations of data, across multiple dimensions. Run semantic search based on vector similarity.

Stores embeddings in Elastic. Vectors are numerical values assigned to keywords that use algorithms to identify contextual relationships.

May host your own machine learning model or select a third-party model including support for Pytorch and HuggingFace repos.

Retrieval ranking model for textual and vector search that gives developers control to optimize search.

How ESRE solves business challenges of using LLMs

Challenge	Why?	ESRE ™ Advantage
Using private data with LLMs	To provide LLMs with relevant, business-specific results, data must be represented as vectors.	**Built-in vector database** Can create embeddings directly in your Elasticsearch deployment at scale.
Access to LLMs and other models	Need to select a LLM best suited to business needs.	**Flexibility to select and host a LLM** Use your own model, supports range of models. Support for Pytorch and HugginFace repos.
High cost of LLMs	LLMs have high computing costs. Need to reduce costs at scale.	**Search results at Elastic scale** Deliver relevant results via context window to LLMs and keep compute and storage costs at a minimum.
Data privacy & security	Need to ensure proprietary training data is secure and private.	**Document-level security, integration with 3rd party tools like LangChain** Control data access with Elastic, integrate with LangChain for added security.
Providing search relevance with AI to LLMs	AI isn't suitable to all use cases. Need semantic / hybrid search powered by with ML.	**Best-in-class semantic search via BM25, RRF, and/or Learned Sparse Encoder** Hybrid search to optimize multiple ranking approaches without tuning. Cost effective solutions for best relevance.

Elastic Learned Sparse Encoder (ELSER)

AI-powered Search: ELSER [Tech Preview]

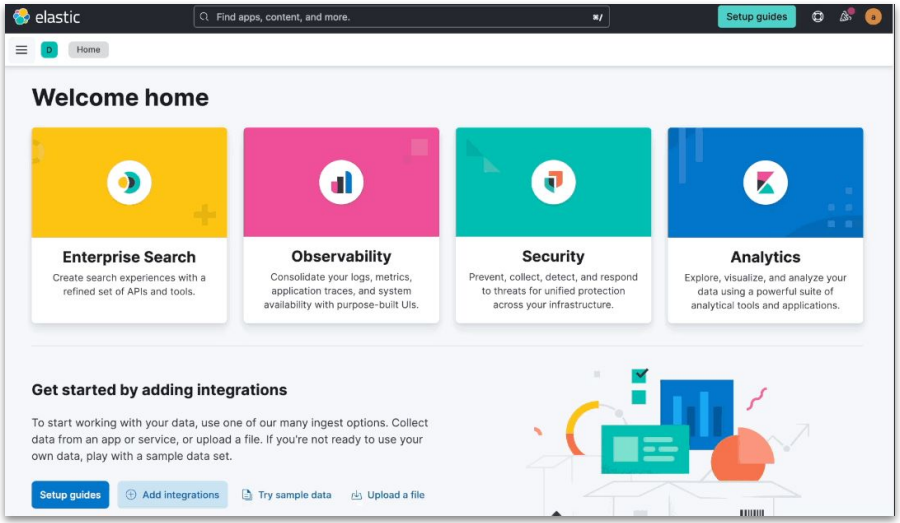
‘Elastic Learned Sparse EncodeR’ for Semantic Search

Drivers

- You know, for… Semantic Search!
- Elastic’s new relevance engine
- Democratise AI-powered search

Customer Value

- Superior Semantic relevance: Search based on *meanings*, instead of just terms
- Start using it OOTB with one click with our streamlined search APIs



Native Semantic Search OOTB

Platinum | Experimental

elastic

AI-powered Search: ELSER [Tech Preview]

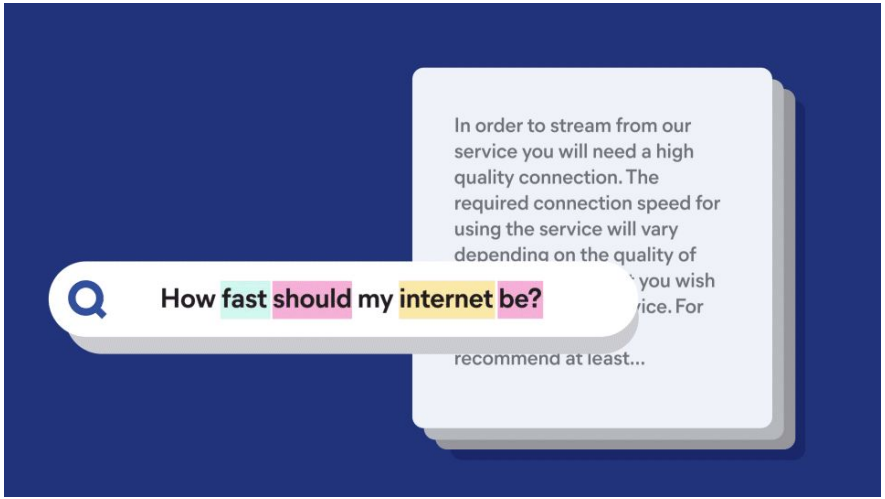
‘Elastic Learned Sparse EncodeR’ for Semantic Search

The insurmountable barrier to AI-search

- Significant expertise and resources
- Processes significantly different than conventional software productization

Enter ELSER

- Out-of-domain → Out-of the box
- Solves the vocabulary mismatch problem
- More interpretable than dense vector search results
- State of the art: Outperforms SPLADE, the previous champ in the category



Platinum | Experimental

elastic

ELSER Benchmarks: vs BM25 vs SPLADE

Overview of latest benchmarks (industry/research standard):

- **ELSER** outperforms BM25 in **11** out of the 12 benchmarks
- **ELSER** outperforms SPLADE in **8** out of the 12 benchmarks
- More granular statistics also on the positive side and expecting further improvements

Data sets (BEIR benchmark)

Search results ranking models	Average	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA	ArguAna	Touche-2020	DBPedia	SCIDOCS	FEVER	Climate-FEVER	SciFact
BM25	0.416	0.688	0.327	0.326	0.602	0.254	0.472	0.347	0.287	0.165	0.649	0.186	0.69
RRF (BM25/Dense)	0.449	0.697	0.317	0.445	0.611	0.318	0.474	**0.354**	0.353	0.159	0.746	0.238	0.671
Linear (BM25/Dense)	0.471	0.787	0.335	0.485	0.62	0.341	0.444	0.346	0.378	0.164	**0.778**	**0.272**	0.698
SPLADE	N/A	0.726	0.347	**0.537**	0.687	**0.347**	**0.526**	0.246	**0.436**	0.158			0.703
ELSER	0.471	0.747	**0.351**	0.524	0.67	0.339	0.5	0.263	0.415	0.156	0.777	0.218	0.695
RRF (BM25/ELSER)	**0.478**	**0.797**	**0.352**	0.468	0.674	0.311	0.497	0.347	0.411	**0.166**	0.762	0.24	**0.712**

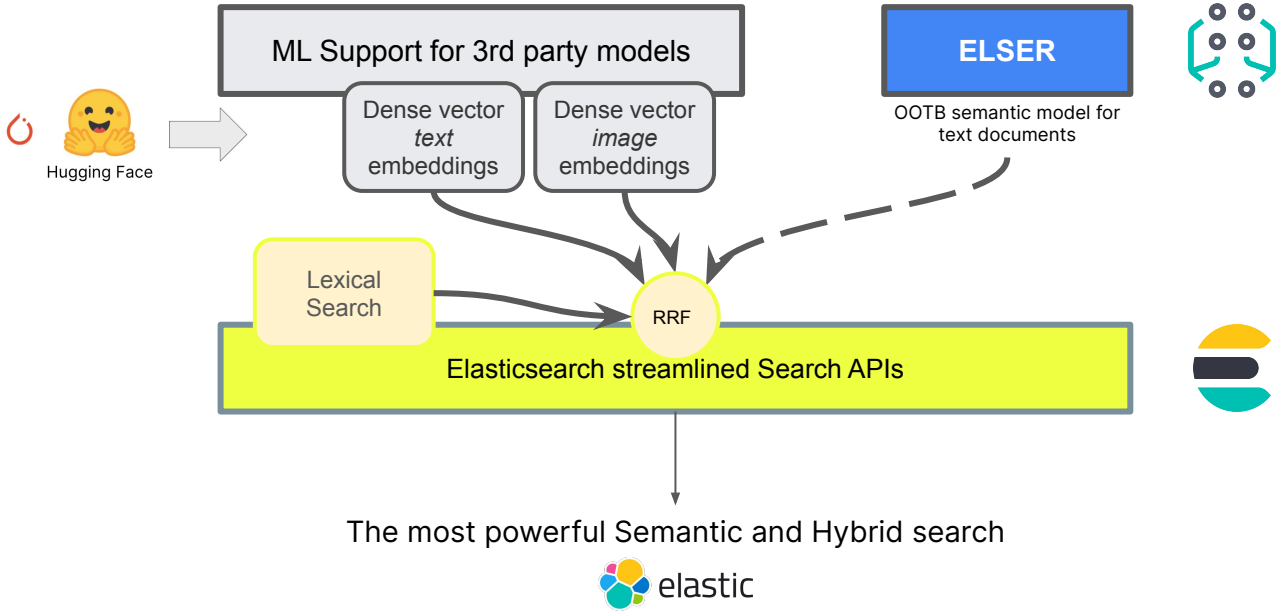
Search quality *worse* than benchmark (BM25)

Search quality *better* than benchmark (BM25)

elastic

AI-powered Search: ELSER [Tech Preview]

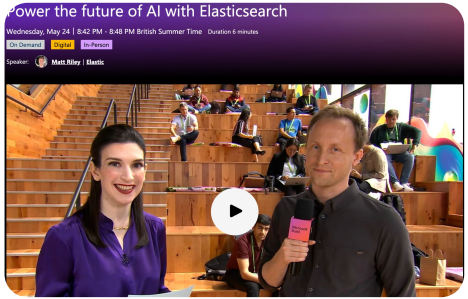
‘Elastic Learned Sparse EncodeR’ for Semantic Search



Platinum | Experimental

elastic

Videos



[5 min Elastic Gen AI pitch]
Great for: high level overview of ESRE, and Elastic's role in AI-search

[30 Min Demo at Microsoft Build]
Great for: *must-watch* magic, Elastic and GenAI working together

[45 min Breakout at Microsoft Build]
Great for: deeper dive take; why vector search is just the start, the depth and breadth of our investments

Featured blogs

- ChatGPT and Elasticsearch: OpenAI meets private data
- ChatGPT and Elasticsearch: A plugin to use ChatGPT with your Elastic data
- Privacy-first AI search using LangChain and Elasticsearch

Launch assets

- Press Release
- Elasticsearch Relevance Engine™ launch blog
- ESRE Landing Page
- ELSER Tech Blog: Elastic AI-powered Search
- ELSER Tech Blog #2: Introducing ELSER Elastic's new retrieval model
- Host Transformer Models/BYOModel Blog

Elasticsearch is a **superset** of vector database capabilities

Elasticsearch is capable of:

- **Creating vector embeddings** (numeric representations of data)
- Is **optimised for storing** sparse and dense vectors, at scale

Elasticsearch also has:

- **Filters and aggregations:** over result set in vector database
- **Document level security built in:** used in production by enterprise customers
- **Data types:** Geo, full text, multiple languages
- **Ingestion tooling:** connectors, API, web-crawler and 3rd party integrations
- That's not all!.. Community, At scale enterprise adoption, Track record, Elastic stack..

Our customers' search relevance journey

context window gen AI demo

Gen AI search

Gen AI search

Industry-leading BM25 search provides relevant keyword search out of the box.

Out of the box, cutting edge semantic search that combines textual and vector search to identify query intent and meaning.

EN-only. Works well across domains.

Store numerical representations of data, across multiple dimensions. Run semantic search based on vector similarity.

Stores embeddings in Elastic. Vectors are numerical values assigned to keywords that use algorithms to identify contextual relationships.

May host your own machine learning model or select a third-party model including support for Pytorch and HuggingFace repos.

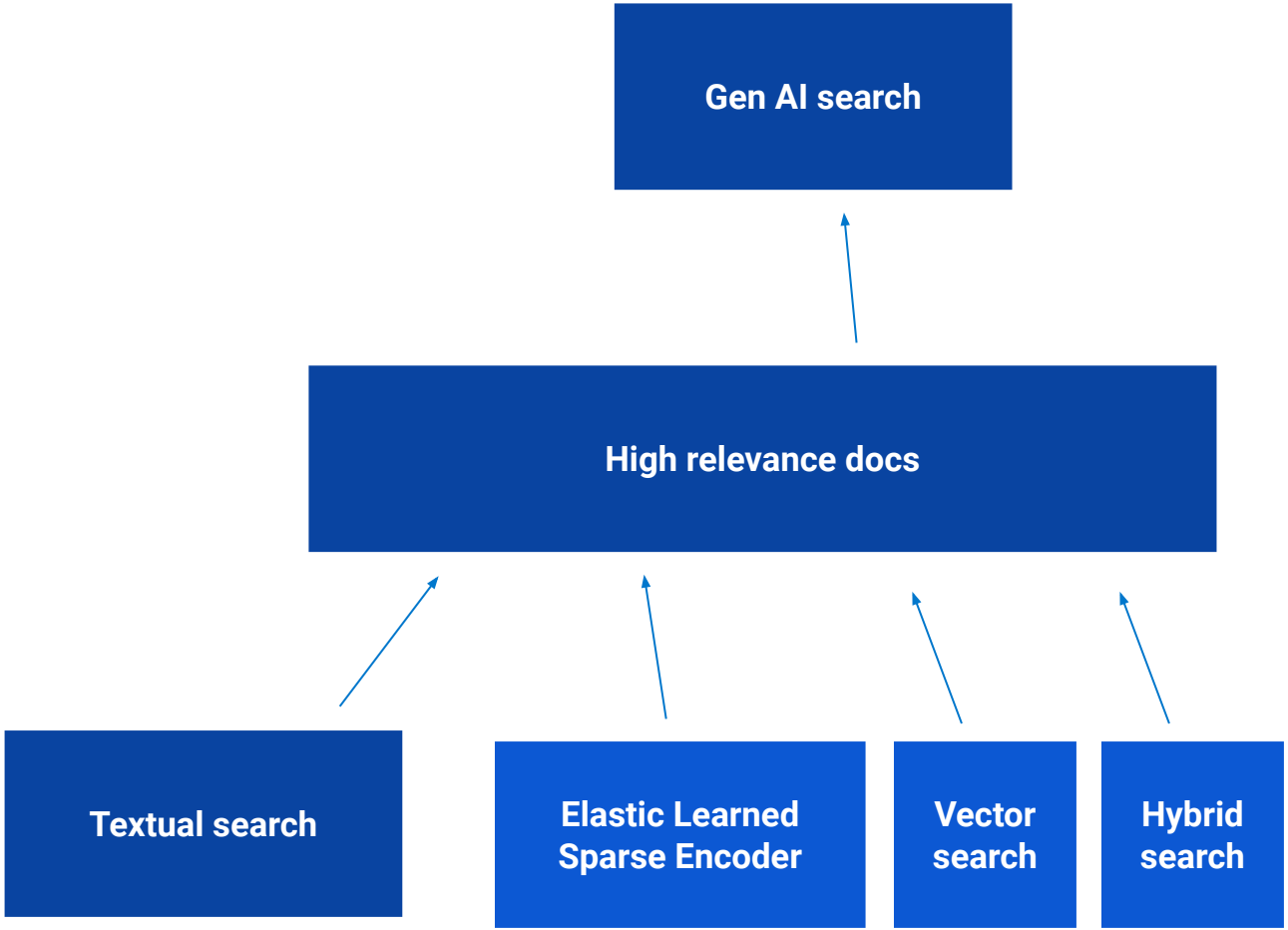
Retrieval ranking model for textual and vector search that gives developers control to optimize search.

Text search

ELSER demo

BYO-M, create embeddings, run Vector search

Hybrid search: full text + geo + semantic


Gen AI search
High relevance docs
Textual search
Elastic Learned Sparse Encoder
Vector search
Hybrid search

Any Data, Any Source

Business Outcomes for Everyone

Learn more:
ela.st/gen-ai

Read the blogs
ela.st/gen-ai-blog
ela.st/encoder-blog