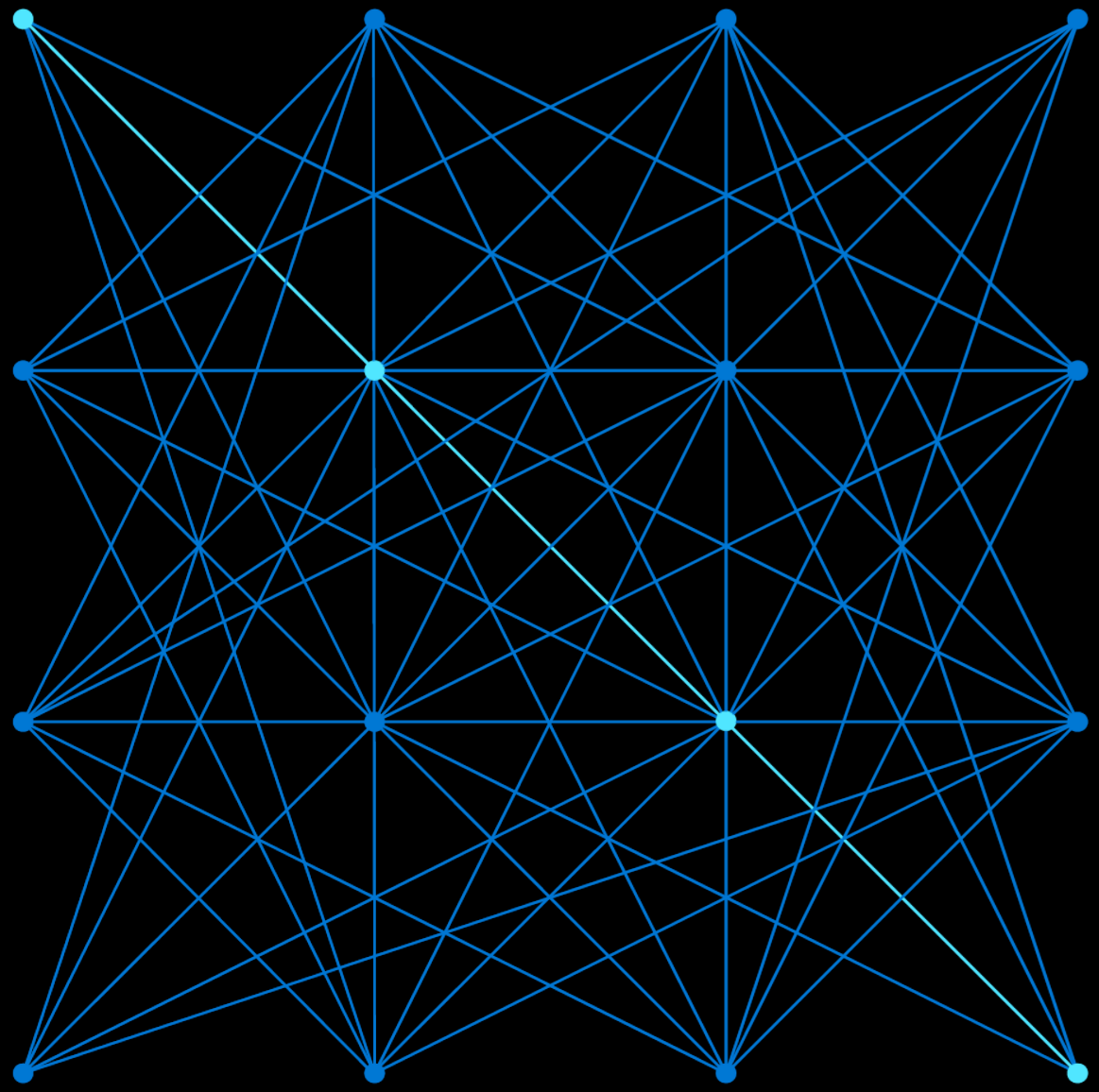
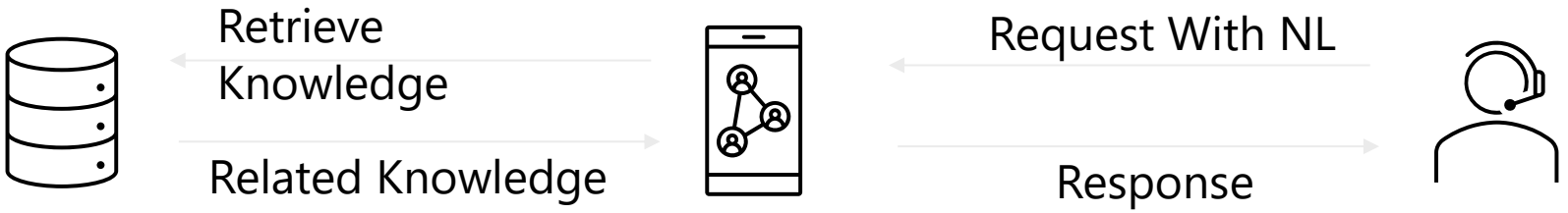
Microsoft Azure

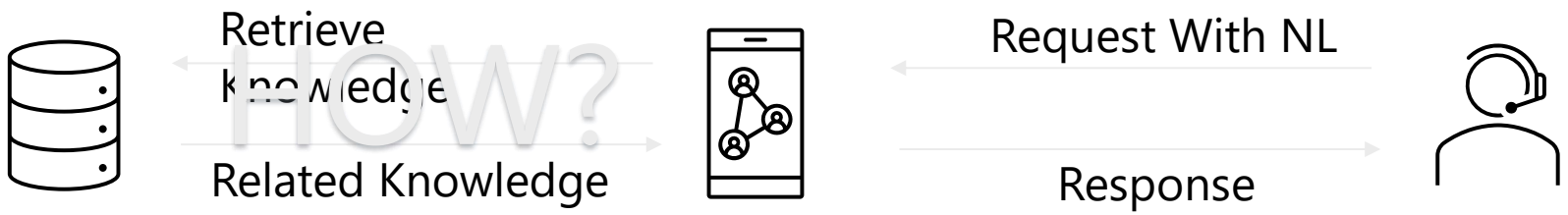
Azure Cognitive Search / Azure AI Search – Overview



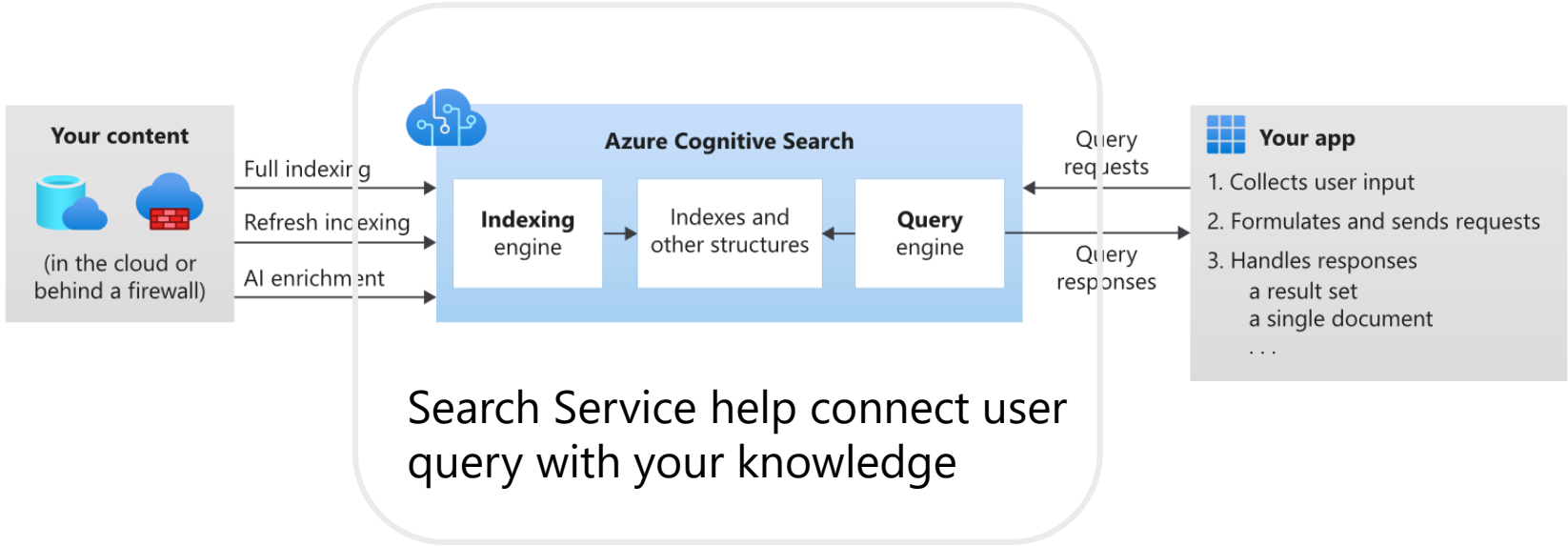
Why do we need search?



Why do we need search?

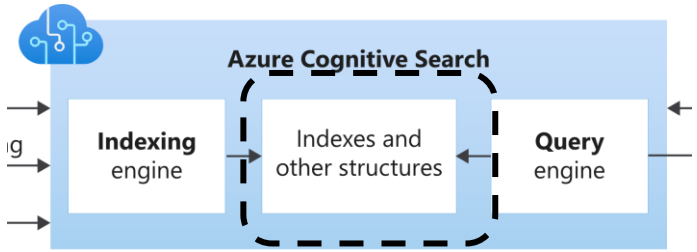


What is Azure Cognitive Search

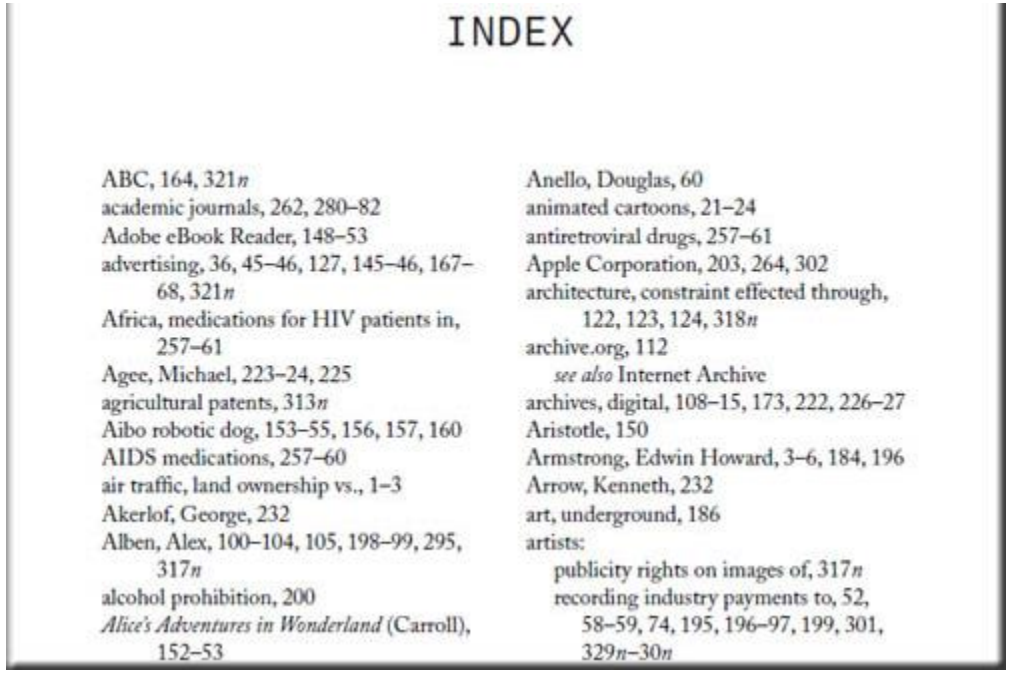


In your client app, the search experience is defined using APIs from Azure Cognitive Search, and can include **relevance tuning**, **semantic ranking**, **autocomplete**, **synonym matching**, **fuzzy matching**, **pattern matching**, **filter**, and **sort**.

What is Search Index



A set of structured data examined by a search engine looking for information **relevant** to a searcher's query.
Search indexing can transform all data and file types into searchable data.
The goal is to make searching as fast, accurate, and relevant as possible

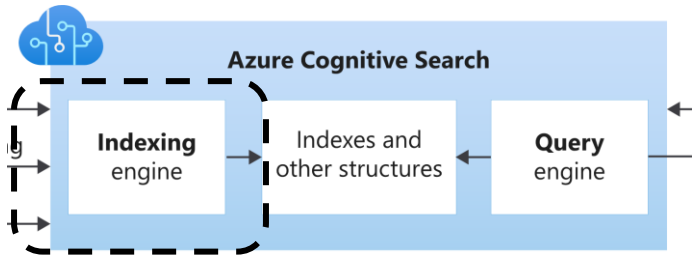


INDEX

ABC, 164, 321*n*
academic journals, 262, 280–82
Adobe eBook Reader, 148–53
advertising, 36, 45–46, 127, 145–46, 167–68, 321*n*
Africa, medications for HIV patients in, 257–61
Agee, Michael, 223–24, 225
agricultural patents, 313*n*
Aibo robotic dog, 153–55, 156, 157, 160
AIDS medications, 257–60
air traffic, land ownership vs., 1–3
Akerlof, George, 232
Alben, Alex, 100–104, 105, 198–99, 295, 317*n*
alcohol prohibition, 200
Alice's Adventures in Wonderland (Carroll), 152–53
Anello, Douglas, 60
animated cartoons, 21–24
antiretroviral drugs, 257–61
Apple Corporation, 203, 264, 302
architecture, constraint effected through, 122, 123, 124, 318*n*
archive.org, 112
see also Internet Archive
archives, digital, 108–15, 173, 222, 226–27
Aristotle, 150
Armstrong, Edwin Howard, 3–6, 184, 196
Arrow, Kenneth, 232
art, underground, 186
artists:
publicity rights on images of, 317*n*
recording industry payments to, 52, 58–59, 74, 195, 196–97, 199, 301, 329*n*–30*n*

Index of a book

What is Indexing



Index in ASC

- Inverted indexes
- vector indexes
- Ai enrichment indexes:
  1. attach image and language
  2. extract text embedded
  3. structure from non-text files

Parsing

- **Tokenization:** break keywords, phrases, symbols into *tokens*
- **Stemming:** expose stem of the word
- **Lemmatization:** Link similar as one word
- **...**

Indexing

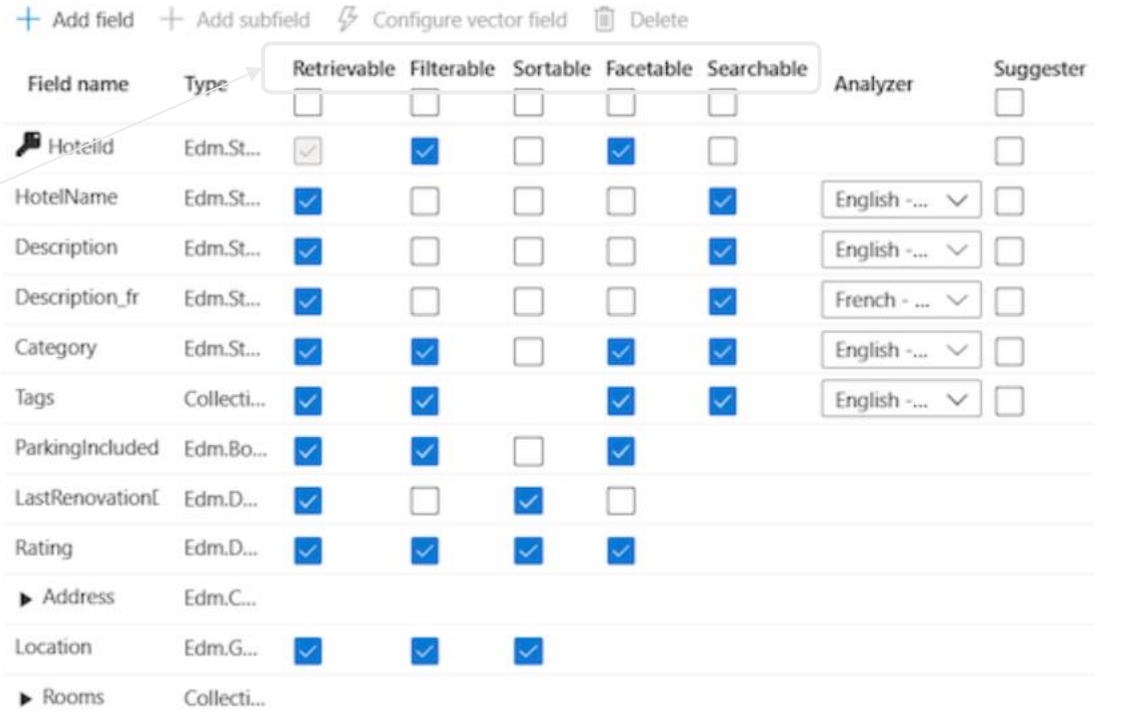
- An intake process that loads content into your search service and makes it searchable.
  Index is created, mapping included words, keywords, phrases, and terms to their source. Usually, such additional information as metadata, location within the source, and frequency of use is also included.

How to Index in ACS

1. **Define an Index Schema and create index**
   1. Define **Index name**
   2. Define a collection of **Fields**:
      - Unique Id *document key* for each doc must be defined.
      - Each field has a name, data type, and *attributes* that control how to use the field in the search index.
2. **Onboard data in the index**
   - **Pull Mode**
     Built in pipeline to update index automatically
   - **Push Mode**
     Manually ingest/update data through SDK / Rest API

How to Index data in ACS

1. **Define an Index Schema and create index**
   1. Define **Index name**
   2. Define a collection of **Fields**:
      - Unique Id *document key* for each doc must be defined.
      - Each field has a name, data type, and *attributes* that control how to use the field in the search index.
2. **Onboard data in the index**
   - **Pull Mode**
     Built in pipeline to update index automatically
   - **Push Mode**
     Manually ingest/update data through SDK / Rest API



How to Index data in ACS

1. **Define an Index Schema and create index**
2. **Onboard data in the index**
   - **Pull Mode**

Document cracking is the process of opening files and extracting content. Text-based content can be extracted from files on a service, rows in a table, or items in container or collection. If you add a skillset and image skills, document cracking can also extract images and queue them for image processing.

- Azure Blob Storage
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database
- Azure Table Storage
- Azure SQL Managed Instance
- SQL Server on Azure Virtual Machines
- Azure Files (in preview)
- Azure MySQL (in preview)
- SharePoint in Microsoft 365 (in preview)
- Azure Cosmos DB for MongoDB (in preview)
- Azure Cosmos DB for Apache Gremlin (in preview)

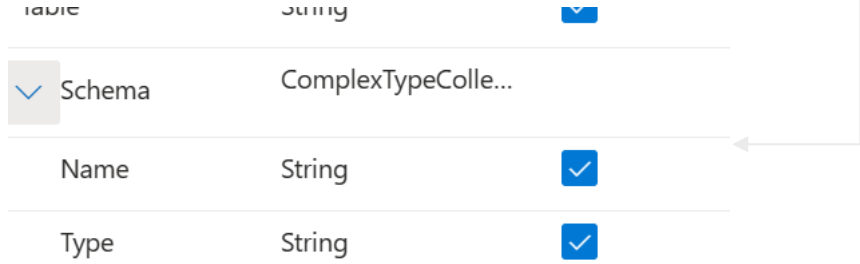
How to Index data in ACS

1. **Define an Index Schema and create index**
2. **Onboard data in the index**
   - **Pull Mode**

An indexer extracts text from a source field and sends it to a destination field in an index or knowledge store. When field names and data types coincide, the path is clear. However, you might want different names or types in the output, in which case you need to tell the indexer how to map the field.

```
[{"Name":"EnrollmentNumber","Type":"String"},
{"Name":"RatingAssetId","Type":"String"}]
```

Parse and extract field schema



Example of extracted text with complex data schema

Supported data type could be found in [Data type map for indexers - Azure Cognitive Search | Microsoft Learn]

How to Index data in ACS

1. **Define an Index Schema and create index**
2. **Onboard data in the index**
   - **Pull Mode**

Skillset execution is an optional step that invokes built-in or custom AI processing. Skillsets can add optical character recognition (OCR) or other forms of image analysis if the content is binary. Skillsets can also add natural language processing. For example, you can add text translation or key phrase extraction.


```
            "metadata_storage_name": "cGFnZS03LnBkZg2",
            "content": "\nF\no\n\nR\ne\n\nw\no\n\nR\nd\n\nE\nA\n\nR\nT\n\nH\n\nvi\n\nForeword\n\nof all celestial bodies within reach or view, as far as we can \n\nsee, out to the edge, the most wonderful and marvelous and \n\nmysterious is turning out to be our own planet earth. There is \n\nnothing to match it anywhere, not yet anyway. \n\n—Lewis Thomas \n\nSixty years ago, with the launch of Explorer 1, NASA made \n\nits first observations of Earth from space. Fifty years ago, \n\nastronauts left Earth orbit for the first time and looked back \n\nat our “blue marble.” All of these years later, as we send \n\nspacecraft and point our telescopes past the outer edges of  \n\nthe solar system, as we study our planetary neighbors and  \n\nour Sun in exquisite detail, there remains much to see and \n\nexplore at home. \n\nWe are still just getting to know Earth through the tools of \n\nscience. For centuries, painters, poets, philosophers, and \n\nphotographers have sought to teach us something about our \n\nhome through their art.\n\n\n",
            "persons": [
                "Lewis Thomas"
            ],
            "locations": [
                "earth",
                "Earth",
                "solar system",
                "Sun",
                "home"
            ],
            "organizations": [
                "NASA"
            ],
            "quantities": [
                "1",
                "first"
            ],
            "dateTimes": [
                "Sixty years ago",
                "Fifty years ago"
            ],
            "urls": [],
            "emails": [],
            "personTypes": [
                "astronauts",
                "painters",
                "poets",
                "philosophers",
                "photographers"
            ],
```

How to Index data in ACS

1. **Define an Index Schema and create index**
2. **Onboard data in the index**
   - **Pull Mode**

Flatten complex structures into a string collection.

Note: Updating progress depending on the data source, can honor the native change detection functionality of the underlying data source so that data refresh picks up just the changed data.

Basic Workflow for Indexer

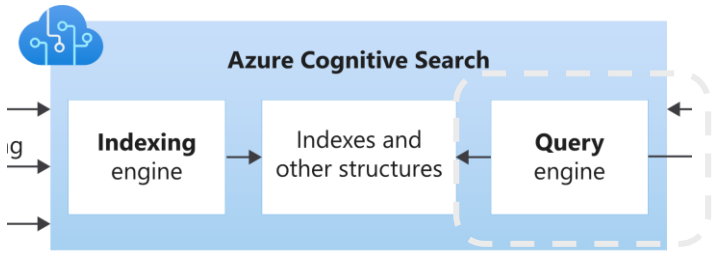
1. Create a data source
2. Create an index
3. Create and run (or schedule) the indexer

How to Index in ACS

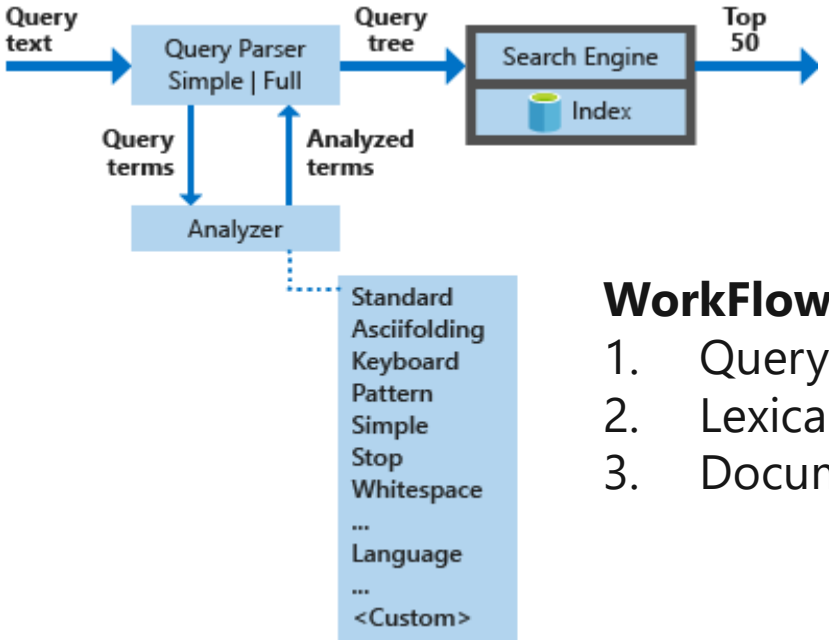
1. **Define an Index Schema and create index**
2. **Onboard data in the index**
   - **Pull Mode**
     - Update data source
     - Update indexer
   - **Push Mode**
     User can use Rest API / SDK to create index, upload data. Specially for .NET, you could find code sample for create index in Line 249 - 306, and upload data in Line 55

Manage your index
- Create index alias
- ….

What is Query in ACS



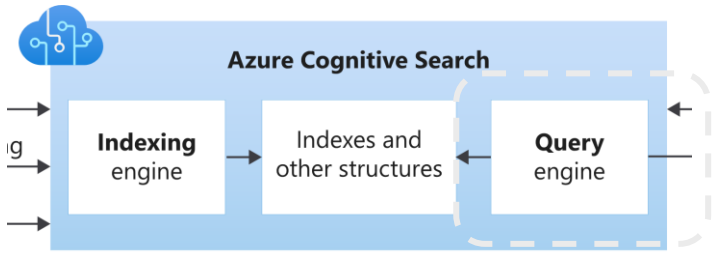
- **Full Text Search**
  BM25-ranked
- Semantic Search
- Vector search
- Hybrid Search



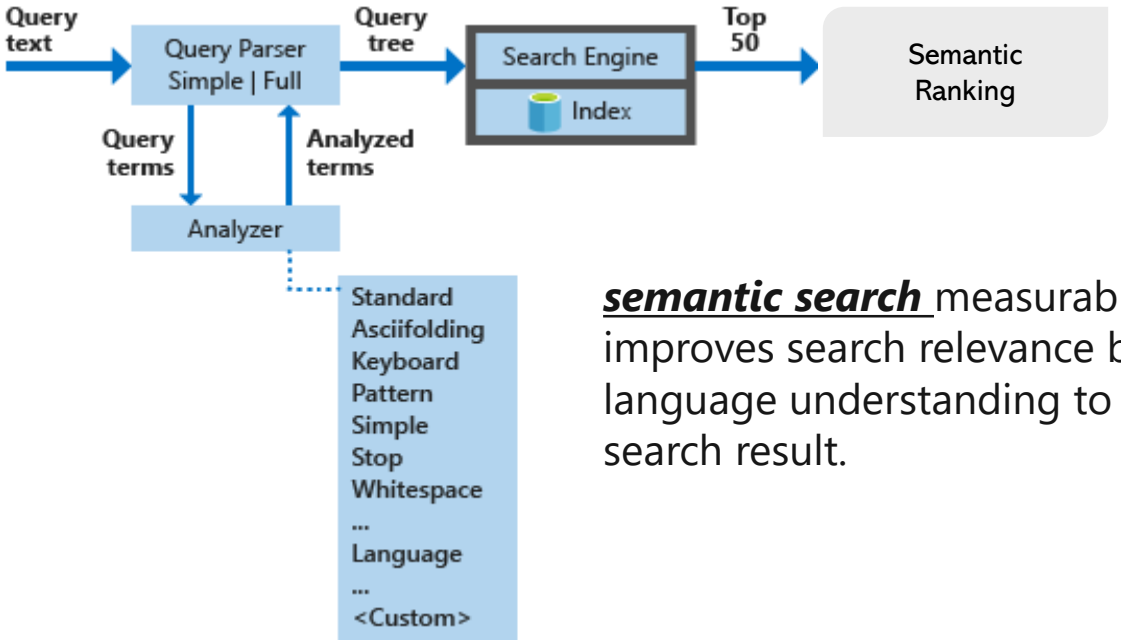
WorkFlow

1. Query Parsing
2. Lexical analysis
3. Document retrieval

What is Query in ACS

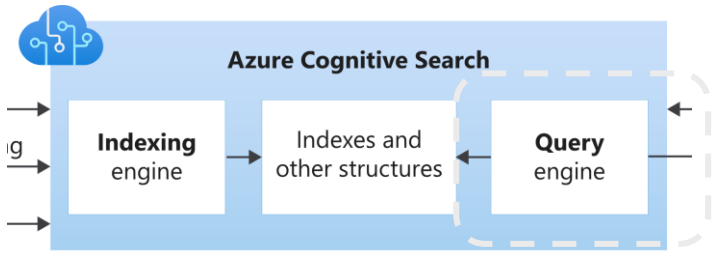


- Full Text Search
- **Semantic Search**
- Vector search
- Hybrid Search

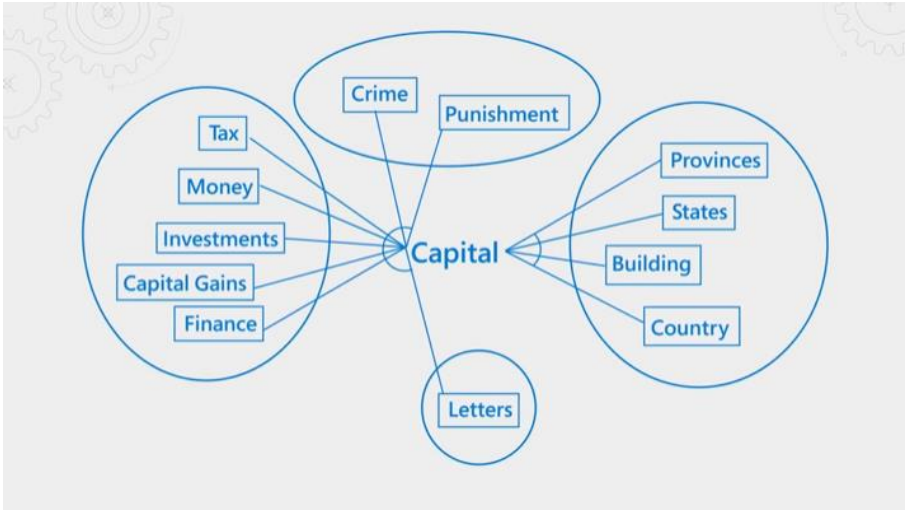


semantic search measurably improves search relevance by using language understanding to rerank search result.

What is Query in ACS

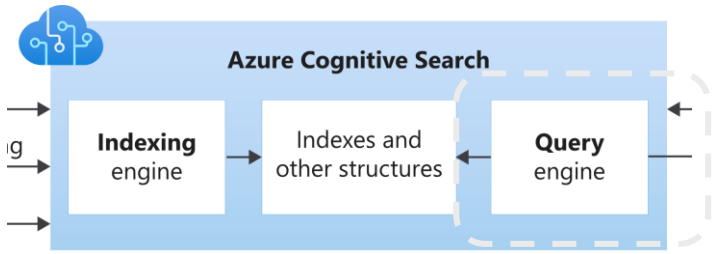


- Full Text Search
- **Semantic Search**
- Vector search
- Hybrid Search



Semantic ranking looks for context and relatedness among terms, elevating matches that make more sense for the query.

What is Query in ACS



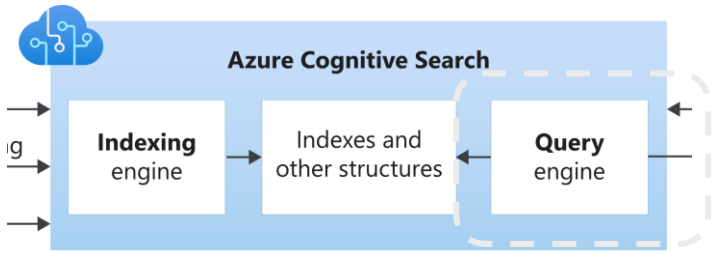
- Full Text Search
- **Semantic Search**
- Vector search
- Hybrid Search

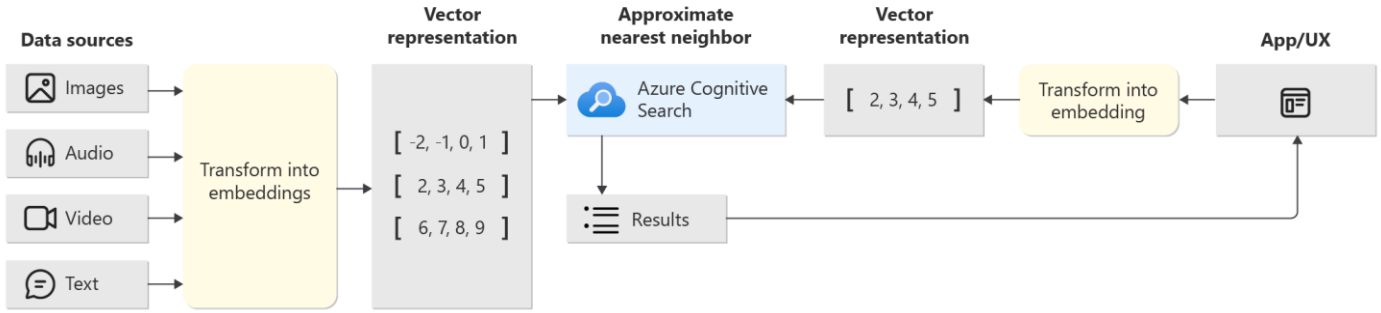
esults

```
1
2   data.context": "https://sotels-cognitive-search.search.windows.net/indexes('sotel-test-index')/$metadata#docs(*)
3   earch.answers": [],
4   lue": [
5
6    "@search.score": 8.279007,
7    "@search.rerankerScore": 1.423553466796875,
8    "@search.captions": [
9      {
10       "text": "# Blueshift Storage Stream Inventory\r \r `@@LastModified`\r \r ## Overview\r Blueshift storage me
11       "highlights": "#<em> Blueshift Storage Stream</em> Inventory\r \r `@@LastModified`\r \r ## <em>Overview\r B
12     }
13   ],
14   "content": "# Blueshift Storage Stream Inventory\r\n\r\n`@@LastModified`\r\n\r\n## Overview\r\nBlueshift storag
15   "metadata_storage_path": "aHR0cHM6Ly9jb21wdXRpbmd0ZWxlbWV0cnl0ZXN0LmJsb2IuY29yZS53aW5kb3dzLm5ldC9haS13b3Jrc3BhY
16 ,
17
18   "@search.score": 7.0862007,
19   "@search.rerankerScore": 1.393768310546875,
```

Query text:
semanticConfiguration=sotels-test-semantic-config&answers=extractive|count-3&queryLanguage=en-us&queryType=semantic&search=where I can find blueshift stream properites

What is Query in ACS

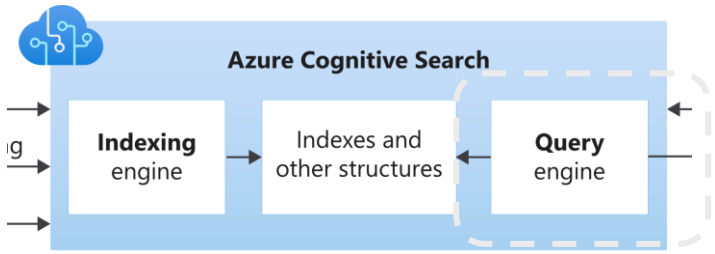




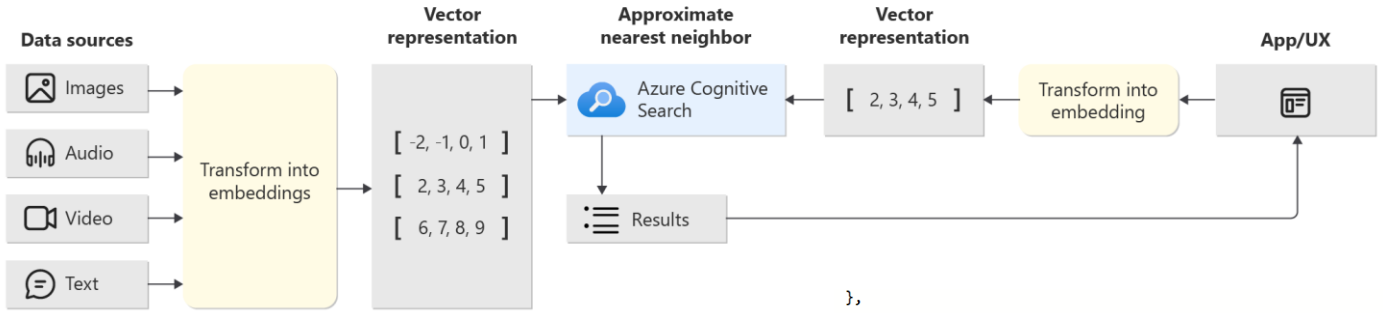
- Full Text Search
- Semantic Search
- **Vector search**
- Hybrid Search

On the indexing side,
Prepare source documents that contain embeddings. Cognitive Search doesn't generate embeddings.
On the query side,
Add a step that converts the query into a vector, and then send the vector query to your index on Cognitive Search for a similarity search. Cognitive Search returns documents with the requested k nearest neighbors (kNN) in the results.

What is Query in ACS



- Full Text Search
- Semantic Search
- **Vector search**
- Hybrid Search



Supported Measure:

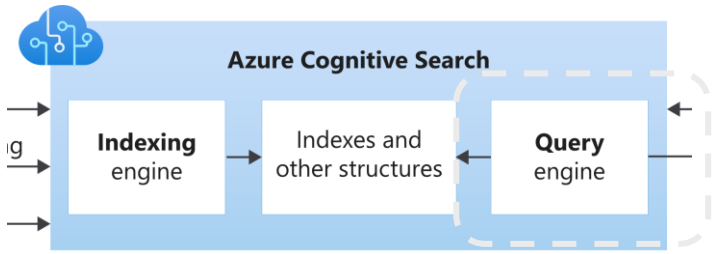
- **euclidean** (also known as L2 norm)
- **cosine**
- **dotProduct**

Supported Searching Method:

- **HNSW** (Hierarchical Navigable Small World)


```
},
"value": [
  {
    "@search.score": 1,
    "id": "0",
    "title": "ADLSG1-Merge Small DataChunk.md",
    "content": "---\r\ntitle: ADLSG1-Merge Small DataChunk Merge\r\
    "contentVector": [
      0.02842521,
      0.017407803,
      0.0018849034,
      -0.00850797,
      -0.022426676,
      0.036473494,
      -0.008907369,
      -0.023225477,
      -0.01871904,
      -0.04455192,
      0.015245016,
```

What is Query in ACS



- Full Text Search
- Semantic Search
- Vector search
- **Hybrid Search**
  - Full Text Search & Vector Search
  - Semantic Search & Vector Search

Common applications of Azure Cognitive Search

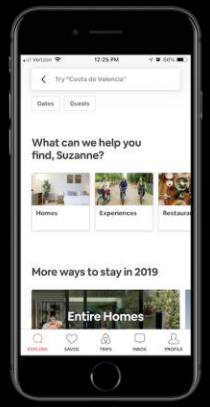
Workplace Search

Help internal teams explore databases and files

- Improve efficiency and productivity
- Enhance data accessibility
- Improve decision-making

SaaS Search

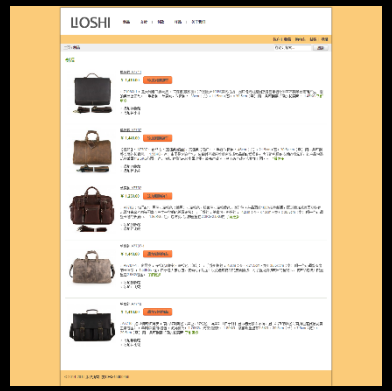
Build market-ready applications for customers



- Improve user experience
- Reduce development time

eCommerce

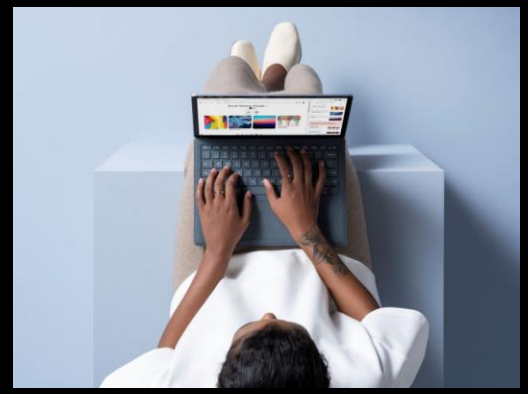
Help customers find and purchase products and services



- Provide personalized recommendations
- Improve user experience
- Enhance product discovery
- Increase conversion rates

Website Search

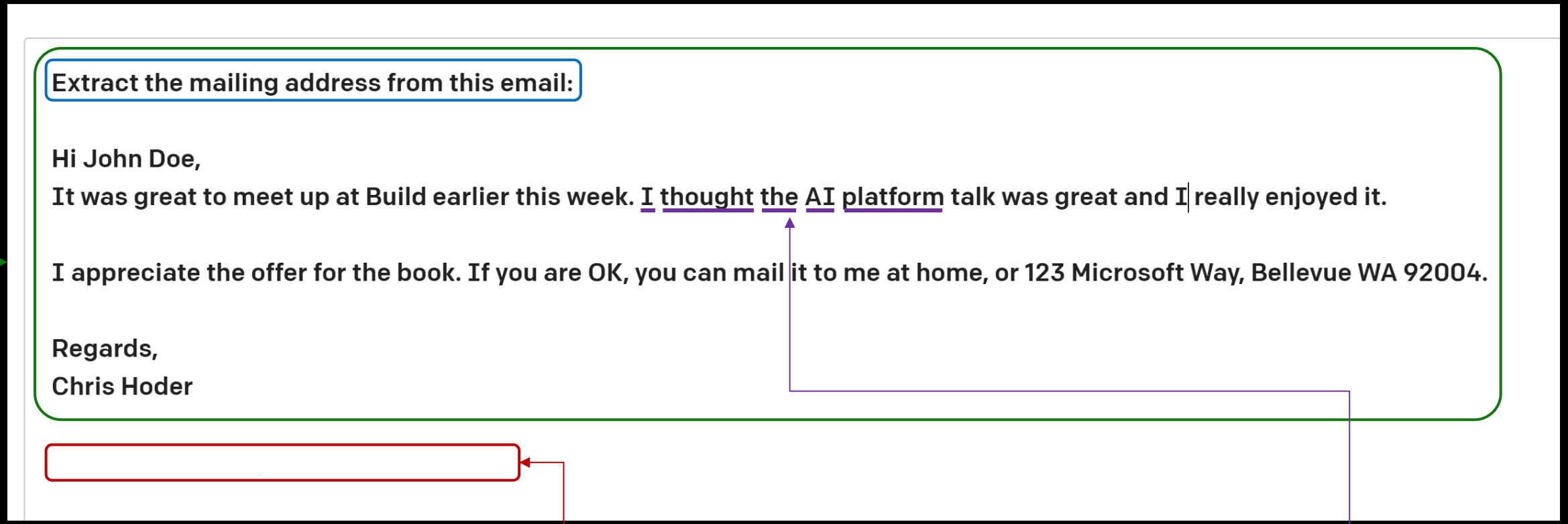
Help visitors find information quickly and easily



- Increase findability
- Better understand user behavior and needs

What about Generative AI scenarios?

Key Terms



Prompt—Text input that provides some context to the engine on what is expecting.

Completion—Output that GPT-3 generates based on the prompt.

Token — partial or full words processed and produced by the GPT models

Problem: Generative AI doesn’t have context for your data - and it doesn't know your data

Prompt

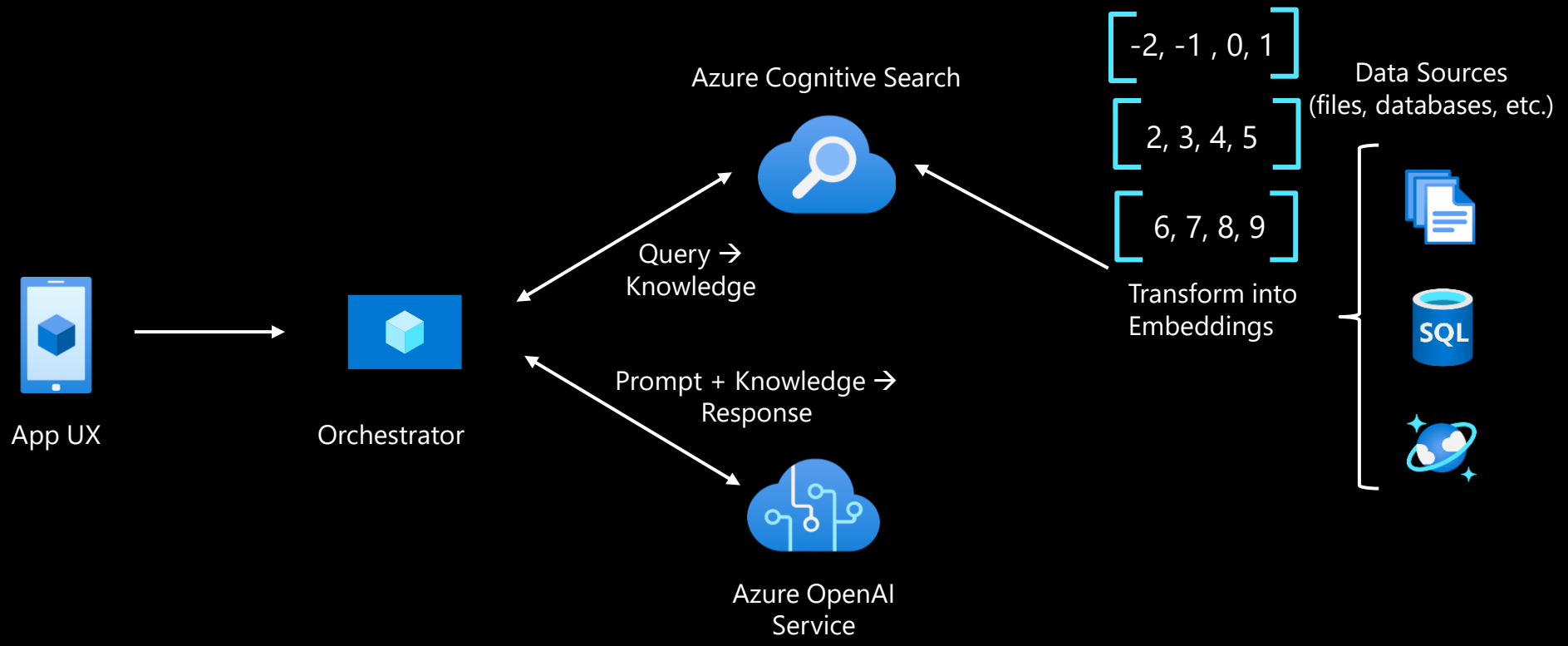
Does my health plan cover annual eye exams?

Response

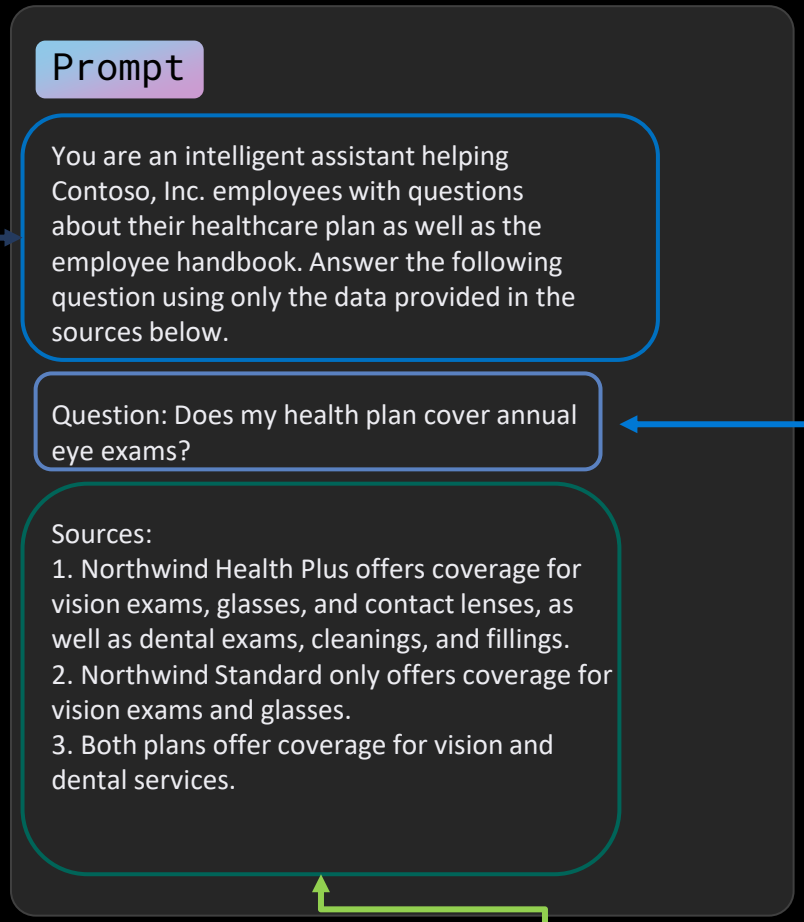
I'm an AI language model and don't have access to specific information about your health plan

Retrieval Augmented Generation (“RAG”)

Grounding for intelligent applications



Retrieval Augmented Generation Example: Bring your data to the prompt



Response

Based on the provided information, it can be determined that both health plans offered by Northwind Health Plus and Northwind Standard provide coverage for vision exams. Therefore, your health plan should cover annual eye exams.

Text input that provides some framing as to how the engine should behave

Sources used to answer the question

User provided question that needs to be answered

Enhancing RAG with Advanced Retrieval Features

Investing in cutting-edge retrieval technology for improved results

R

A

G

← The quality of the **retriever** is critical!

Azure Cognitive Search is committed to providing the BEST retrieval solution through:

- Vector Search capabilities
- Hybrid Search
- Advanced filtering
- Document security
- L2 reranking/optimization
- Built-in chunking
- Auto-Vectorization
- And much more!

How might we take our enterprise search or RAG scenarios to the next level?

Introducing Vector search

Revolutionize Indexing and Power Retrieval Augmented Generation for LLM Apps


Images


Audio


Video


Graphs

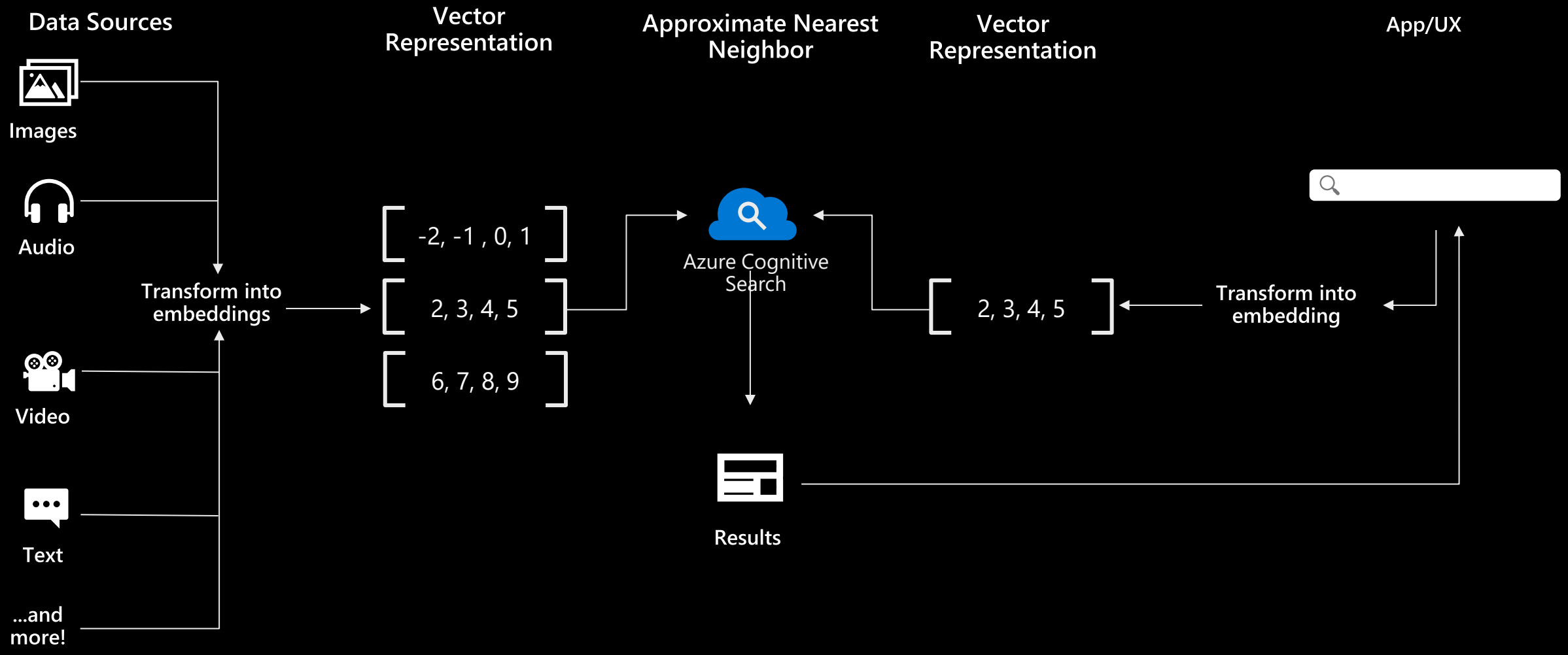

Text

- Leverage data from any data store
- Improve relevancy
- Query across multiple types of data
- Quickly search through large data sets

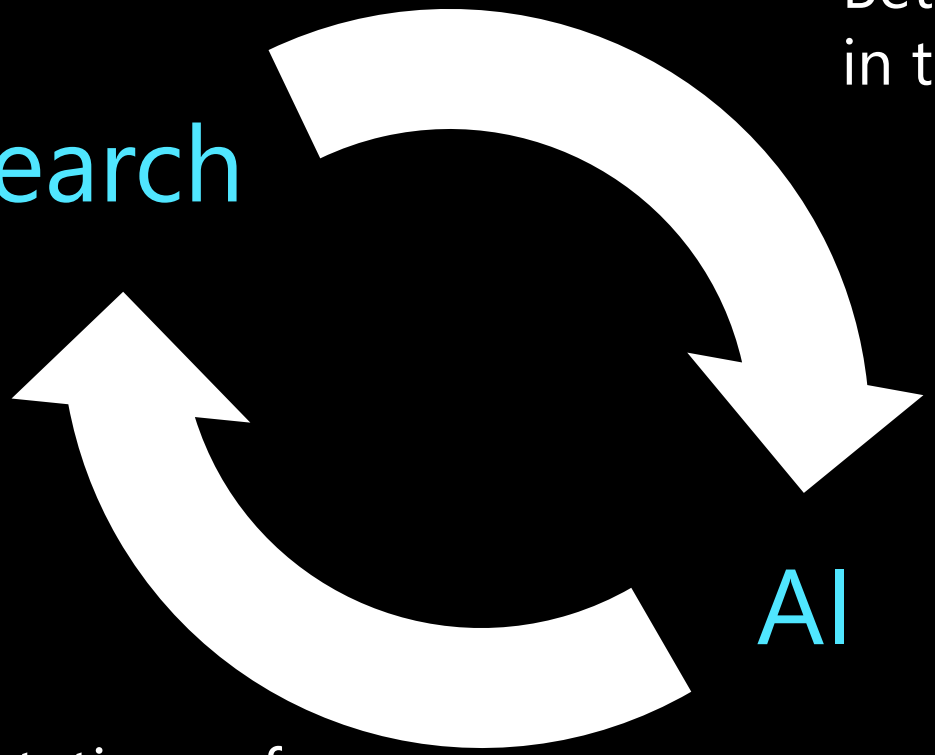
- Deploy with enterprise-grade security
- Easily scale with changing workloads
- Build retrieval plugins for OpenAI's ChatGPT using Azure OpenAI service

What is Vector search?

Convert data into vector representations where distances represent similarity



Search + AI
Better Together



Better information in the context

Better representations of data (embeddings)

Vector Search at a High Level

Scenario

A diverse collection of books, each containing unique insights and knowledge

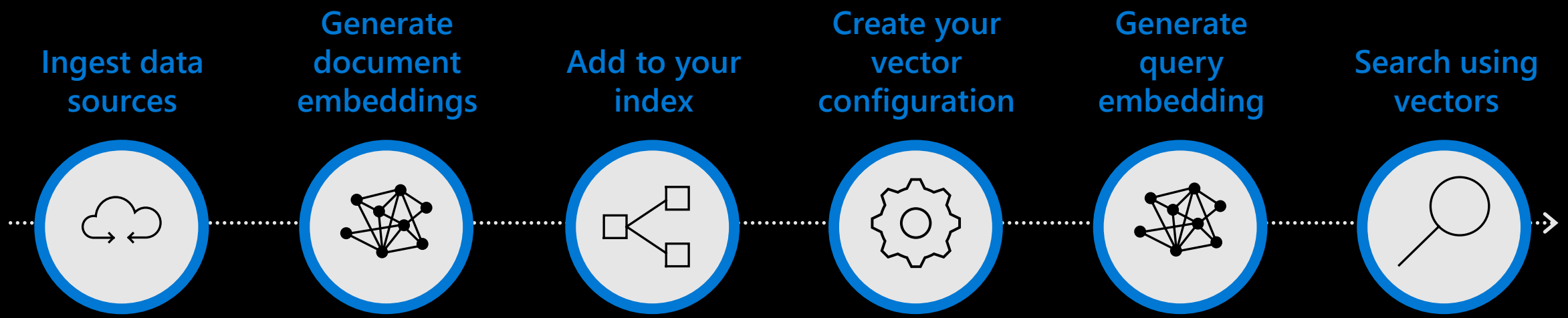
The Challenge

Finding a book on a specific topic or theme can be time-consuming and overwhelming, especially when the content is scattered.

The Solution

A skilled librarian can quickly connect you to books with similar topics or themes

How do I get started with Vector search?



Retrieval Modes

Vector search is good, but Hybrid search is even better!

	Full-text search (BM25)	Pure Vector search (ANN)	Hybrid search (BM25 + ANN)
Exact keyword match	✓	✗	✓
Proximity search	✓	✗	✓
Term weighting	✓	✗	✓
Semantic similarity search	✗	✓	✓
Multi-modal search	✗	✓	✓
Multi-lingual search	✓	✓	✓

Why is Hybrid Search important?

Hybrid Queries with BM25 and ANN Search Integration

- Hybrid search allows you to take advantage of multiple scoring algorithms such as BM25 and ANN vector similarity so you can get the benefits of **both keyword search and semantic search**

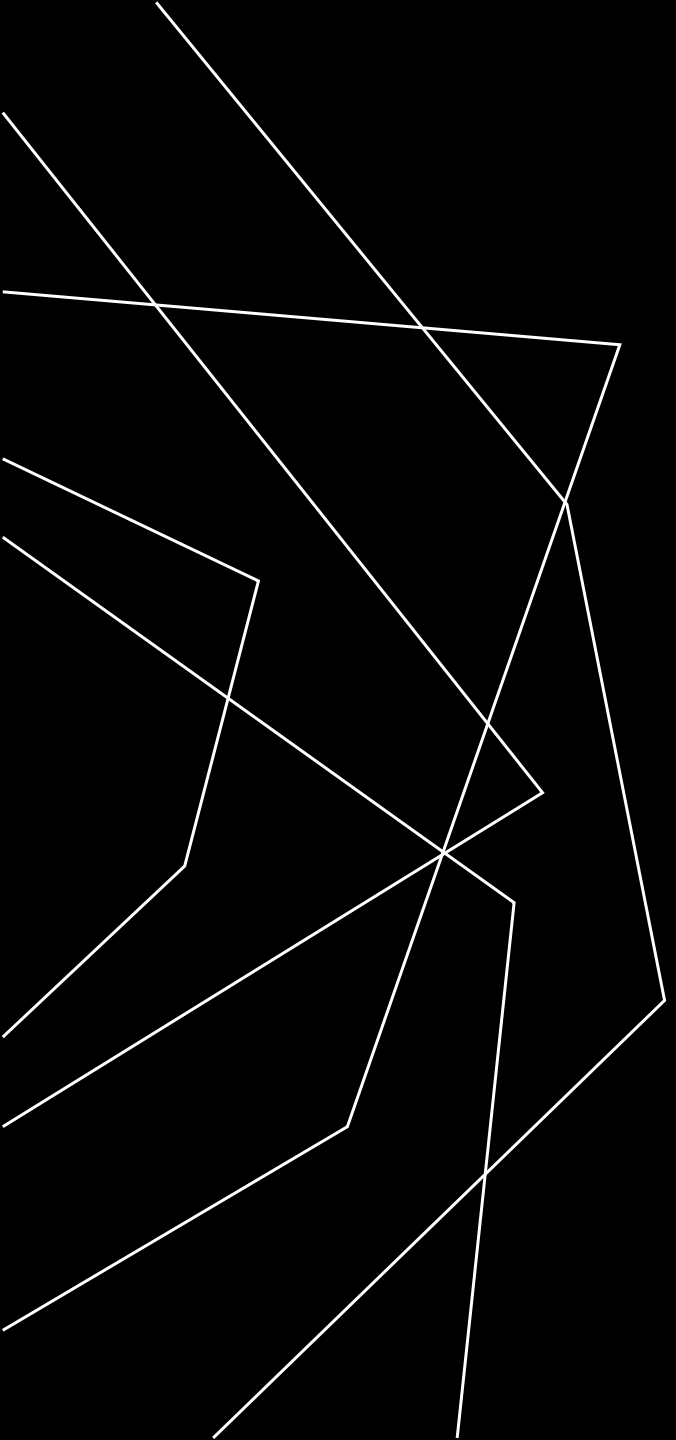
Achieve Better Relevance with Hybrid Search + Reranking

Optimal Search Relevance

Traditional “Full-text” search +  + 

Resources

https://aka.ms/IntroducingVectorSearch
https://learn.microsoft.com/en-us/azure/search/search-what-is-azure-search



THANK YOU