

# Wektory w akcji

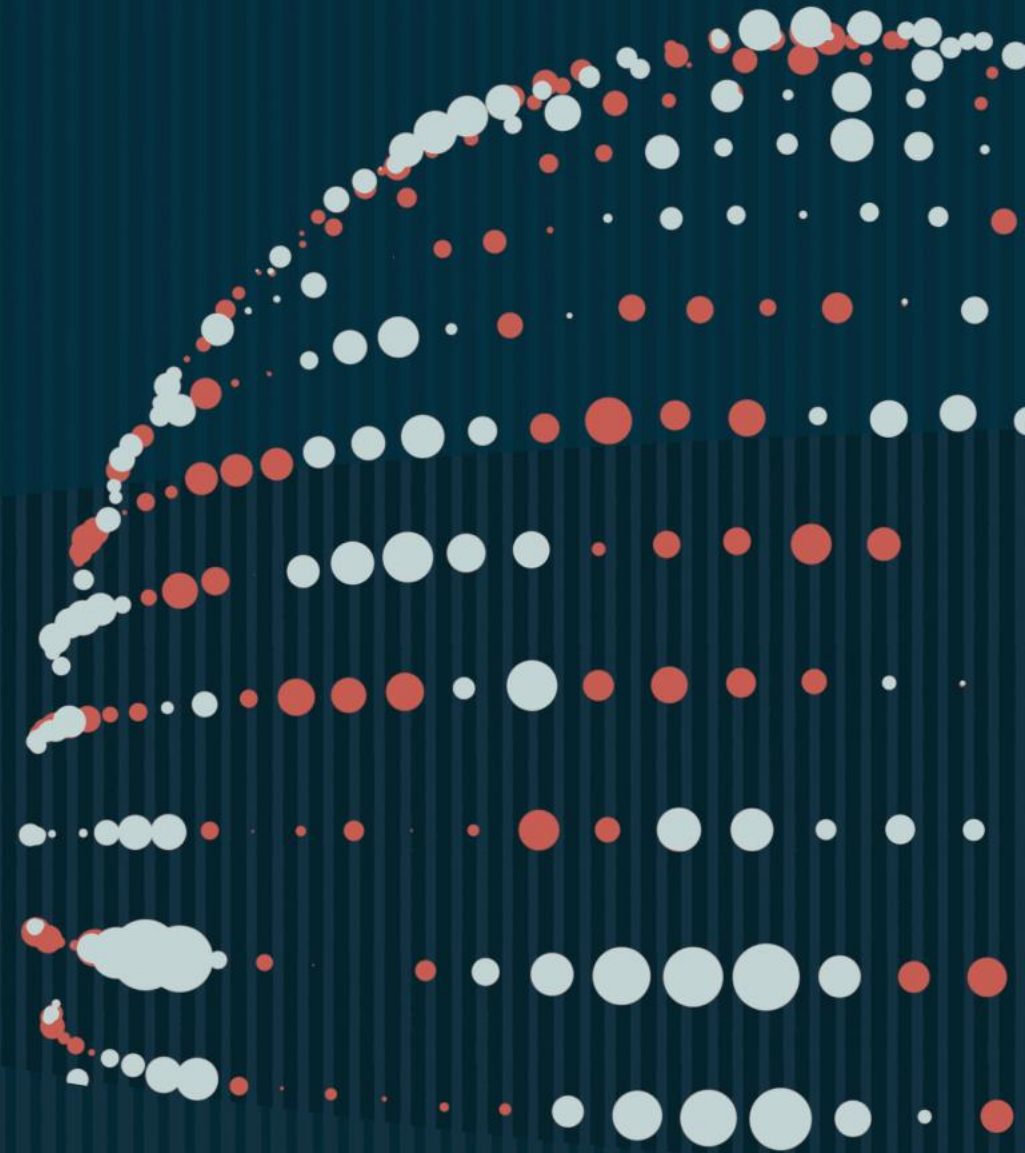
## Oracle 23ai, MySQL Vector i RAG w praktyce

Tomasz Żołnierzak

Consulting Technical Director, Oracle ECCC

Piotr Matejewski

Senior Principal Consultant, Oracle ECCC

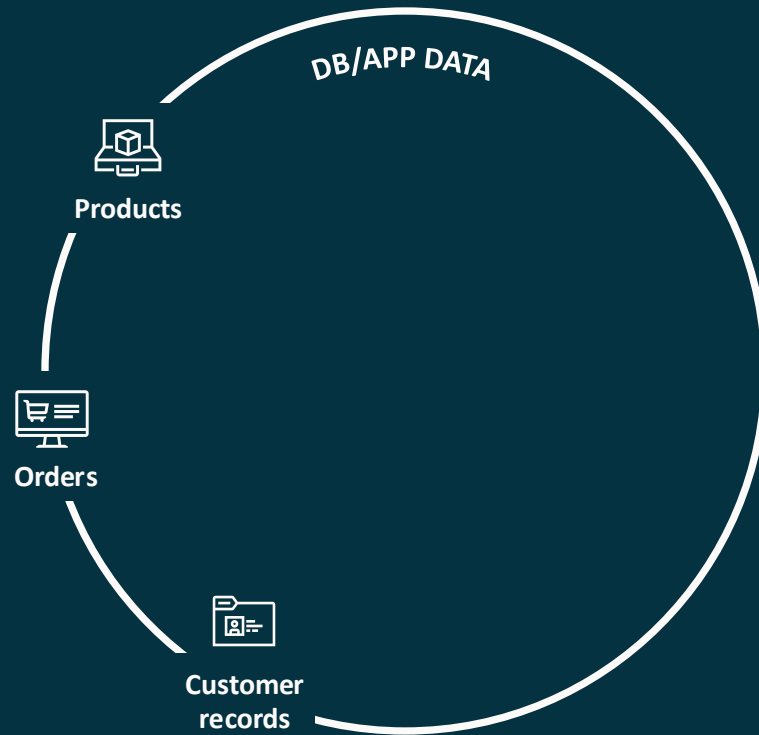


# Safe harbor statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

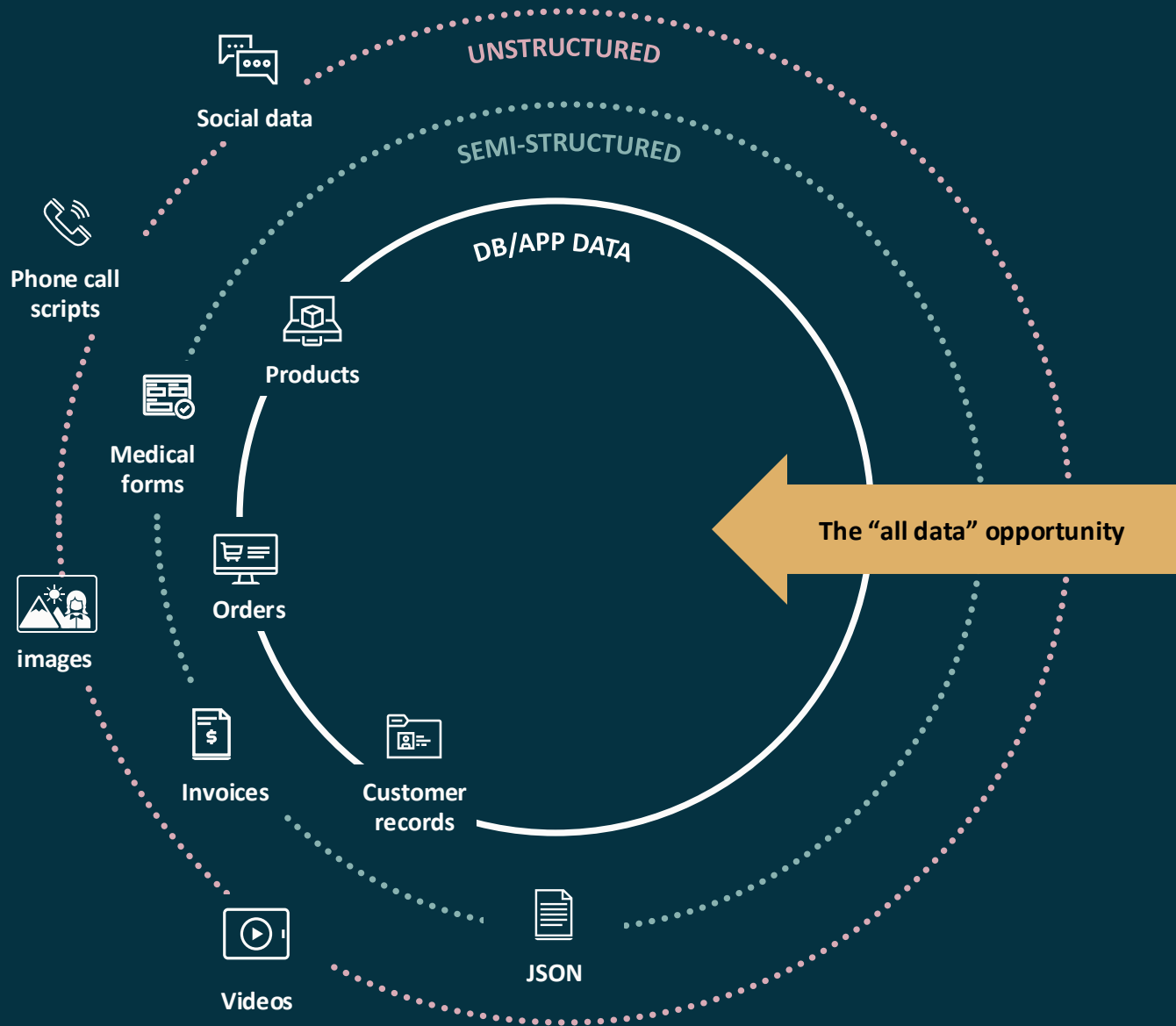
# Agenda

- Wektory w teorii
- Oracle AI Vector Search
- RAG & Oracle Select AI
- MySQL HeatWave
- (\*) Demo Select AI



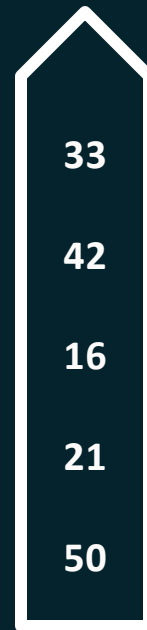
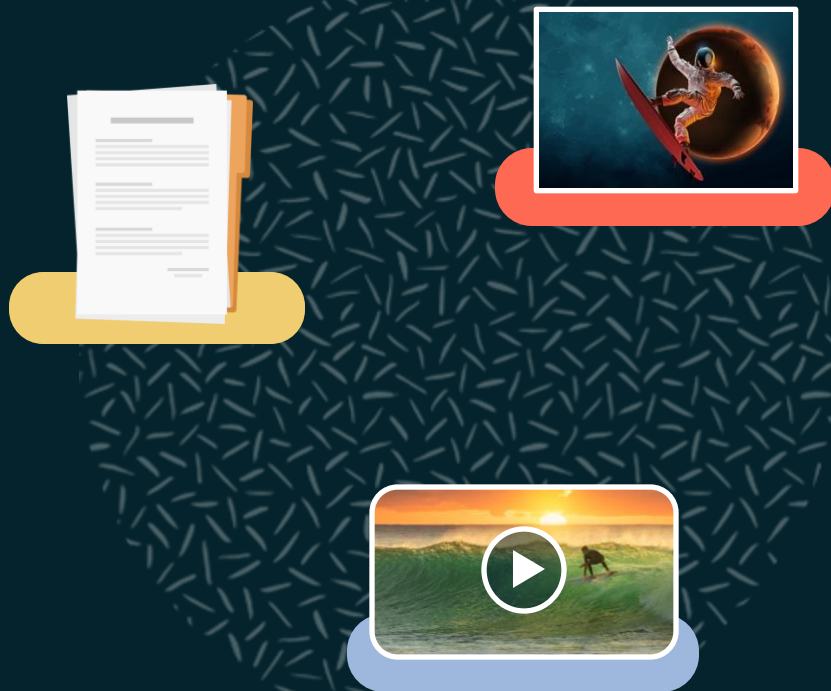
Databases are great at querying business data that is stored as strings, numbers, and dates

Find revenue by products



Growing volume of **unstructured** business data, which databases haven't been good at querying as it must be searched by **semantics**

Find products that match a photo or description



Vector

A **vector** is a sequence of numbers, called dimensions, that represent the **semantic content** of a document, image, audio, or video

Vectors represent the **semantic content** of data, not the underlying words or pixels

Deep learning **transformers** (or **embedding models**) generate vectors

The terms **vector** and **embedding** are often used interchangeably

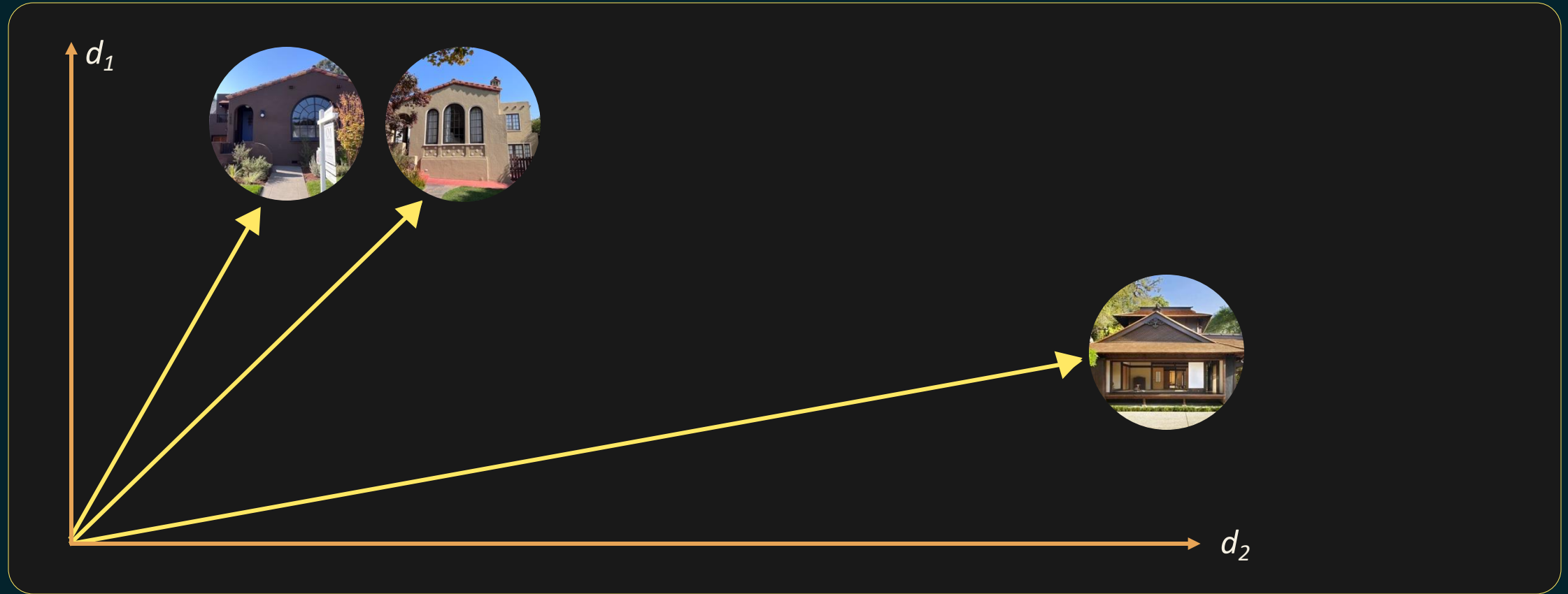


For example, the features of a house image could be:



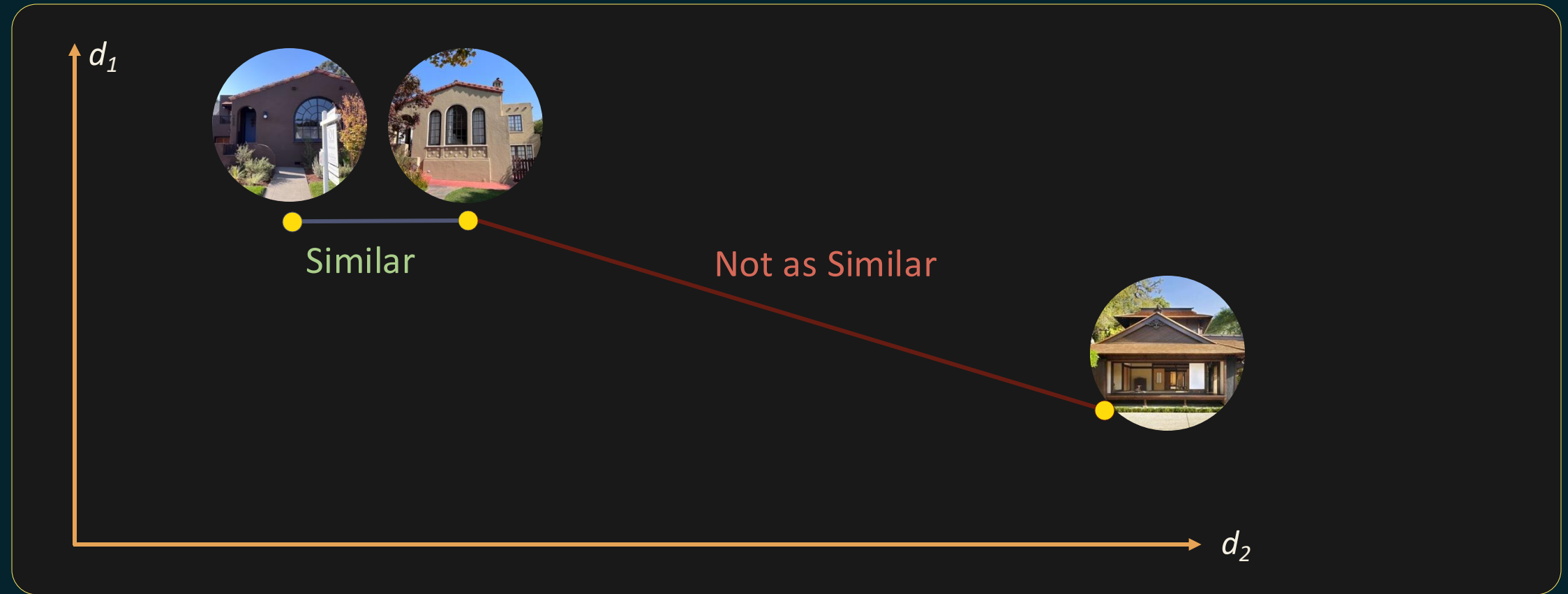
Note: Features are often chosen by ML algorithms and are not as simple as shown here

House vectors when collapsed into 2 dimensions instead of hundreds could look like this

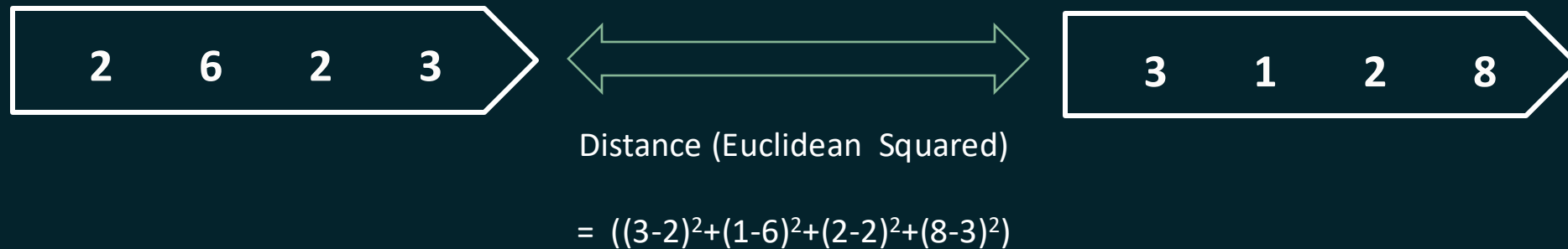




The **distance** between the vectors reflects their **semantic similarity**



The main operation on vectors is the **mathematical distance** between them



*\* multiple mathematical distance functions*

# Generating Vector Embeddings

Convert unstructured data (text, images, audio & video) into vectors for semantic similarity search

Inputs (Your Unstructured Data)

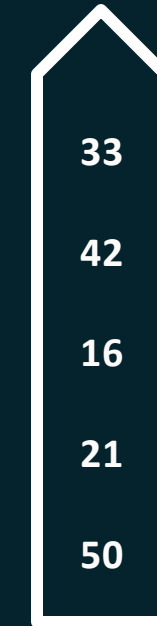
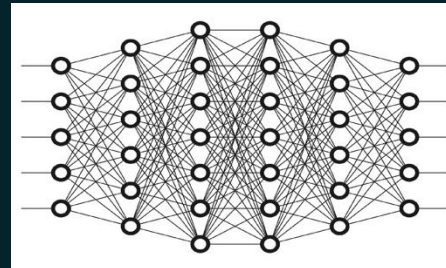
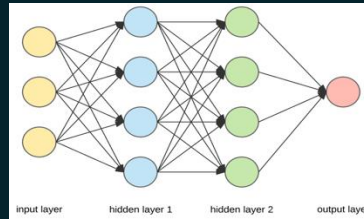


Text Data



Image Data

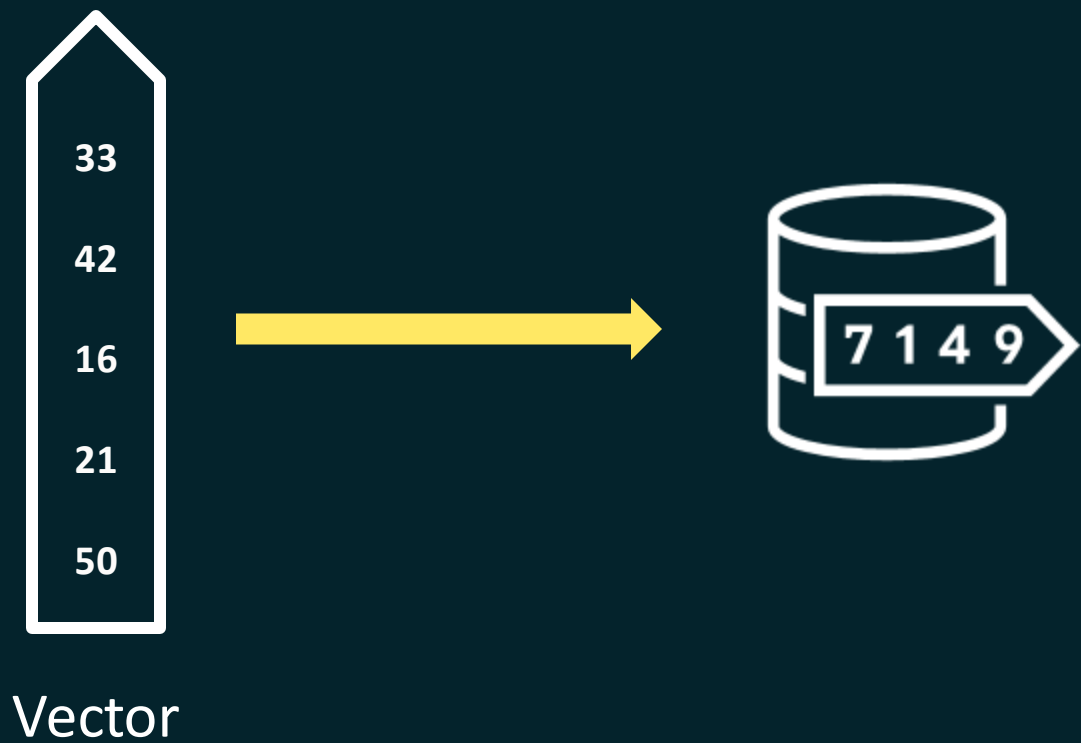
Embedding Models



Vector

Embedding vectors  
that represent similar  
content are closer in  
distance


# Vectors are stored in Vector Databases



Text Vector Table

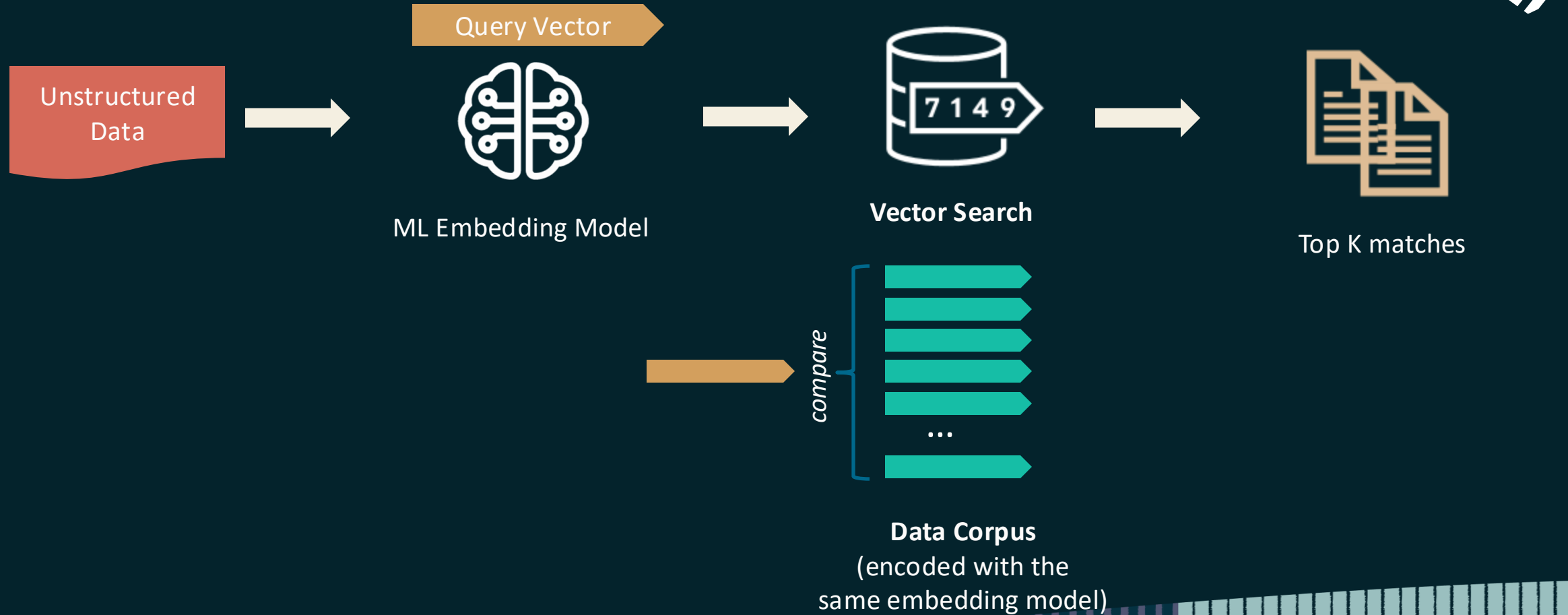
id	vector	text
1	[0.8, 0.5, 1.6, -2.5, ...]	"It was the best of times, it was the worst of times, it was.."
2	[1.1, 0.3, 0.6, -1.3, ...]	"It is a truth universally acknowledged, that a single man.."
3	[1.3, 0.1, 0.2, -1.1, ...]	"It was a bright cold day in April, and the clocks were striking.."
...	...	...

Image Vector Table

id	vector	Image
1	[0.5, 1.5, 2.6, -1.1, ...]	
2	[1.0, 0.9, 1.6, -1.3, ...]	
3	[0.6, 1.1, 1.3, -0.9, ...]	
...	...	...



# The Similarity Property powers **Vector Search**

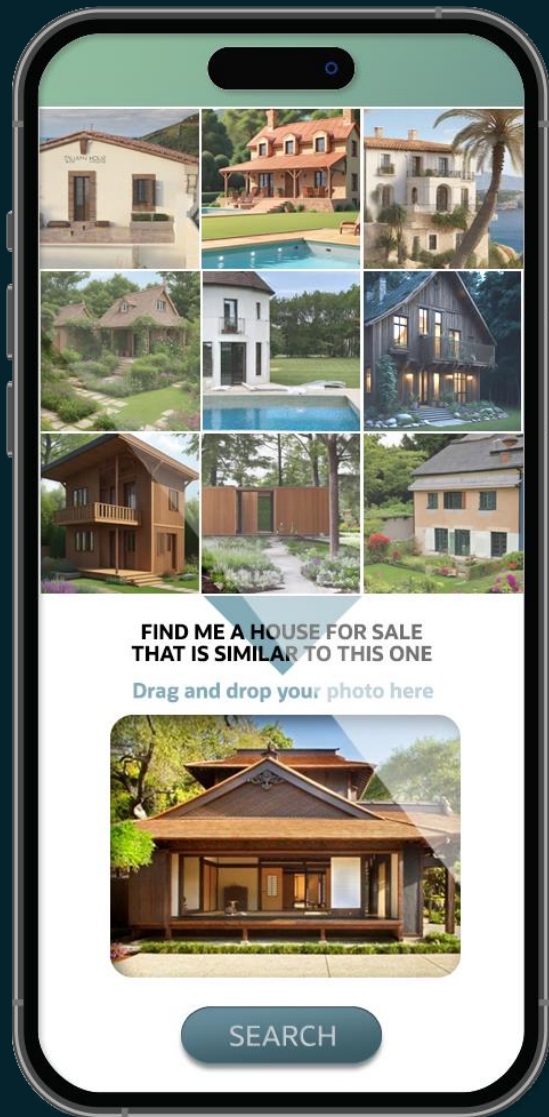


Now that we know what vectors are, let's talk about how they are used



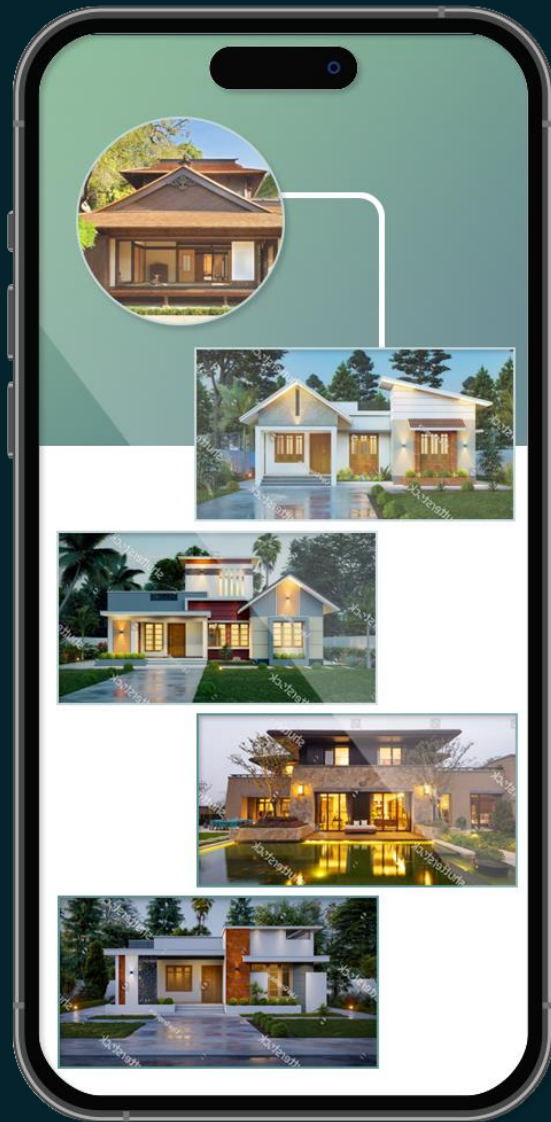
50 21 16 42 33

AI Vector Search on **images, documents, etc.** works best when combined with relational search on **business data** to solve business problems



Let's look at an example

Imagine an app that helps customers find houses for sale that are similar to a picture the customer uploads



Finding a good match requires combining semantic picture search with searches on business data including:

- **Customer data** such as the customer's preferred city and budget
- **Product data** such as houses available for sale in each city and their price

This is easy with Oracle Database 23ai



# Better Together: Business Data and Business Vectors



Converged AI Database

Oracle integrates **AI Vector Search** into your **business database**

Having AI Vector Search in the same database as your **customer** and **product data** facilitates sophisticated information retrieval

No need to move and synchronize data to a niche Vector Database

- Avoid challenges around data staleness, added complexity, comprised security etc.
- Benefit from Oracle Database's **mission-critical** capabilities

# AI Vector Search Highlights



# Oracle Database 23ai can store vectors using a new vector data type



```
CREATE TABLE house_for_sale (house_id      number,  
                              price         number,  
                              city          varchar2(400),  
                              house_photo   blob,  
                              house_vector  vector  
                              );
```

# Vector Embedding Generation | Your Way

AI vector search offers 3 alternatives for vector embedding generation

## 1 Use Pre-Created Embeddings

Load vectors directly from external files into database into VECTOR columns or map the data as external tables

## 2 Use an external embedding service

Generate embeddings using external callouts via **UTL\_TO\_EMBEDDING()** PLSQL function in the **DBMS\_VECTOR** package

## 3 Use a database resident embedding model

Generate embeddings using the **VECTOR\_EMBEDDING()** SQL function using an imported **ONNX** embedding model so that no data leaves the database

# New `VECTOR_EMBEDDING()` function to generate vectors

**Completeness:** Many customers want to be able to generate vectors **within** the database

Oracle Database supports the **Open Neural Net Exchange (ONNX)** framework to import models

The `VECTOR_EMBEDDING()` function can then generate vectors for unstructured data using the imported model

```

// import text model for documents
DBMS_VECTOR.load_onnx_model(
    model_name => "embedding-model",
    model_data => "embedding-model.onnx"
    ...
);
```

```

// generate vectors
SELECT VECTOR_EMBEDDING(CLIP_IMG USING :PHOTO_BLOB
as DATA) AS embedding;
```

# You can now find data that is semantically similar to an input

Find the top 10 houses  
that are similar to this picture



```
SELECT    ...  
FROM      house_for_sale  
ORDER BY  vector_distance(house_vector, :input_vector)  
FETCH FIRST 10 ROWS ONLY;
```

# You can now find data that is semantically similar to an input

Ultra simple

Data Professionals and  
Developers can learn to  
use AI Vector Search in  
minutes

No AI expertise required

Find the top 10 houses  
that are similar to this picture



```
SELECT    ...  
FROM      house_for_sale  
ORDER BY  vector_distance(house_vector, :input_vector)  
FETCH FIRST 10 ROWS ONLY;
```

# You can now run queries that combine AI Vector Search with business data about customers and products

Find houses that are similar to this picture **and** match the customer's preferred city and budget



```
SELECT ...  
FROM   house_for_sale  
WHERE  price <= (SELECT budget      FROM customer ...)  
AND    city  in (SELECT search_city FROM customer ...)  
ORDER BY vector_distance(house_vector, :input_vector)  
FETCH FIRST 10 ROWS ONLY;
```



# You can now run queries that combine AI Vector Search with business data about customers and products

Ultra simple and powerful

Combines customer data, product data, and AI search in 6 lines of SQL!

All data is fully consistent

Single integrated solution

Find houses that are similar to this picture **and** match the customer's preferred city and budget



```
SELECT ...  
FROM   house_for_sale  
WHERE  price <= (SELECT budget      FROM customer ...)  
AND    city  in (SELECT search_city FROM customer ...)  
ORDER BY vector_distance(house_vector, :input_vector)  
FETCH FIRST 10 ROWS ONLY;
```

# AI Vector Search | Ultra-Sophisticated SQL

Oracle is a converged database that supports all types of workloads and data models:

- Graph, Text, JSON, Spatial, Relational, etc.

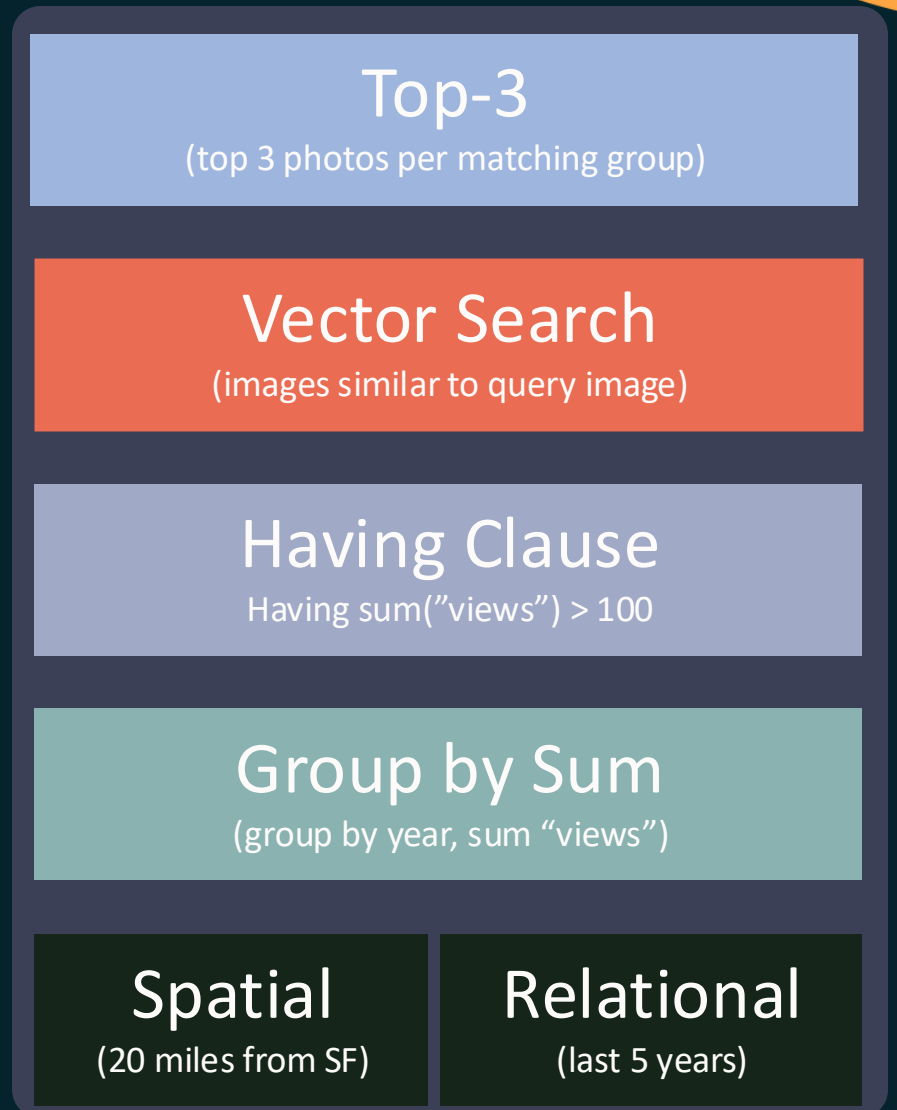
Oracle also has industry-leading SQL functionality

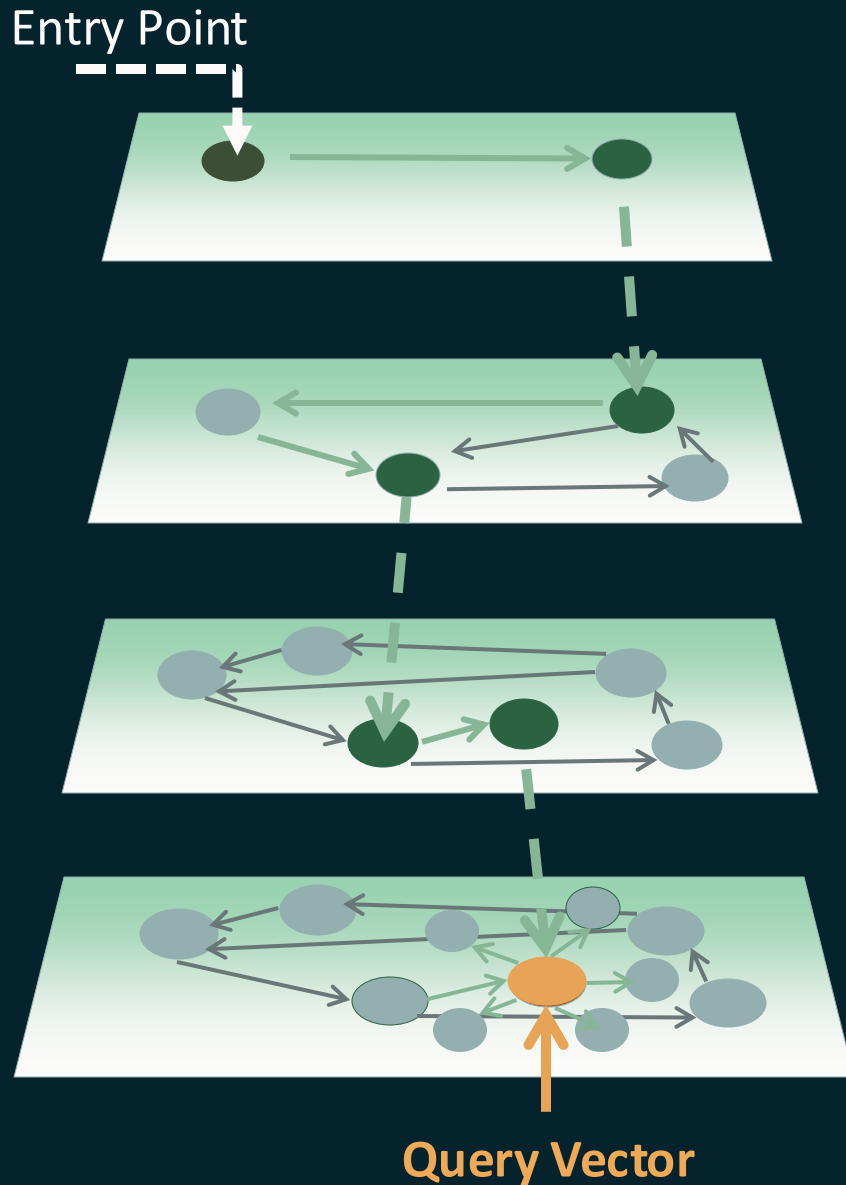
- Complex operators, group-by, aggregation ...
- Analytic functions, stored procedures, pattern matching ...

This allows vector search using **Ultra-Sophisticated SQL**:

*Show me the top 3 photos, grouped by year, over the past 5 years, based on similarity to a provided query image.*

*The photos should have been taken within 20 miles of San Francisco, and have been viewed by at least 100 different people*





Oracle database accelerates  
AI Vector Search using  
sophisticated new  
**vector indexes**

# Approximate Vector Indexes

Distance computation between every vector in a table and the query vector to find the Top-K matches will be 100% accurate but very slow

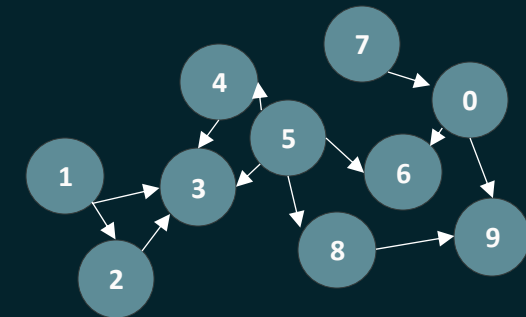
New vector indexes trade-off search accuracy for speed

- Vectors are clustered/connected based on similarity for accuracy
- Greedy search techniques limit accuracy for speed

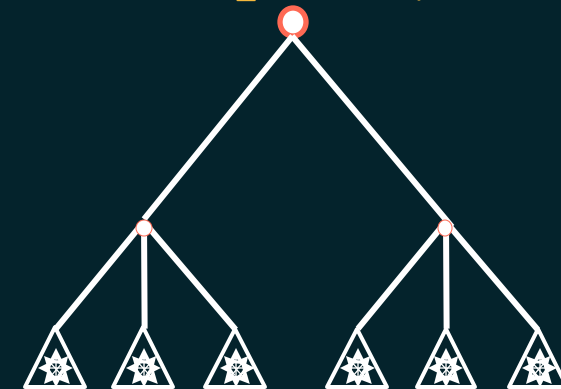
## Vector indexes

- **Neighbor Graph Vector Index** – Graph-based index where vertices represent vectors and edges between vertices represent *similarity*  
In-Memory only index - highly efficient for both accuracy and speed
- **Neighbor Partition Vector Index** – Partition-based index with vectors clustered into table partitions based on *similarity*  
Efficient scale-out index, with fast and seamless transactional support

Graph Vector Index (e.g.  
HNSW Index)



Partition Vector Index (e.g.  
IVF\_FLAT index)



# Vector Index Creation SQL Syntax

Basic index creation syntax:



```
CREATE VECTOR INDEX photo_idx ON house_for_sale(house_vector)
ORGANIZATION [INMEMORY NEIGHBOR GRAPH | NEIGHBOR PARTITIONS]
DISTANCE EUCLIDEAN | COSINE_SIMILARITY | HAMMING ...
```

Choosing the **ORGANIZATION** for an index is simple:

- If the index data will fit in-memory, it is best to use **INMEMORY NEIGHBOR GRAPH**
- Else use **NEIGHBOR PARTITIONS**

The **DISTANCE** function clause is optional (the default is Euclidean)

The distance function should be chosen based on the embedding model used to generate the vectors

# Enterprises are already using Oracle AI Vector Search



## Visual Search for Products

Find products that are similar to a user provided image



## Real-time offer management

Enable merchants to present the right offers to consumers at checkout



## Identify infection-causing bacteria

Compare bacteria genomes to determine cause of an infection



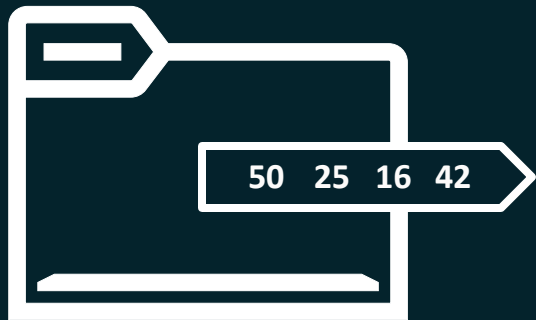
## Real-time intelligent assistant

Use RAG to answers customer questions about products

# Demo 1



# Vector Search enables Retrieval Augmented Generation



Vector Search improves Generative AI response by **augmenting LLM prompts** with **private database content**

This helps produce better answers to user questions

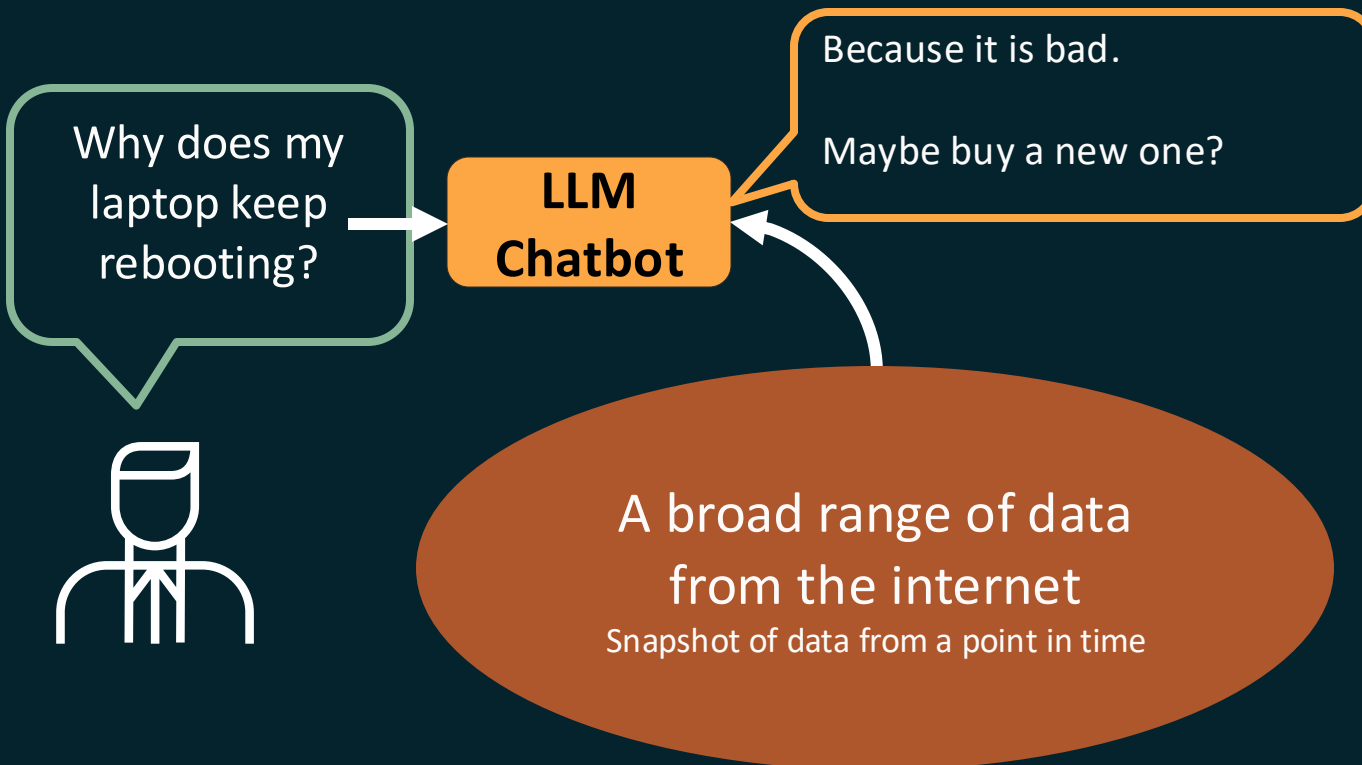
Why is that needed? ...



# Role of Vector Databases in Generative AI

LLMs are frozen on a past snapshot of the internet with no access to private enterprise data

LLMs by themselves therefore often provide **poor-quality** responses to support questions



LLMs need **relevant private enterprise data** in  
addition to the user question  
in the context to **ground** their responses

# Retrieval Augmented Generation (RAG)

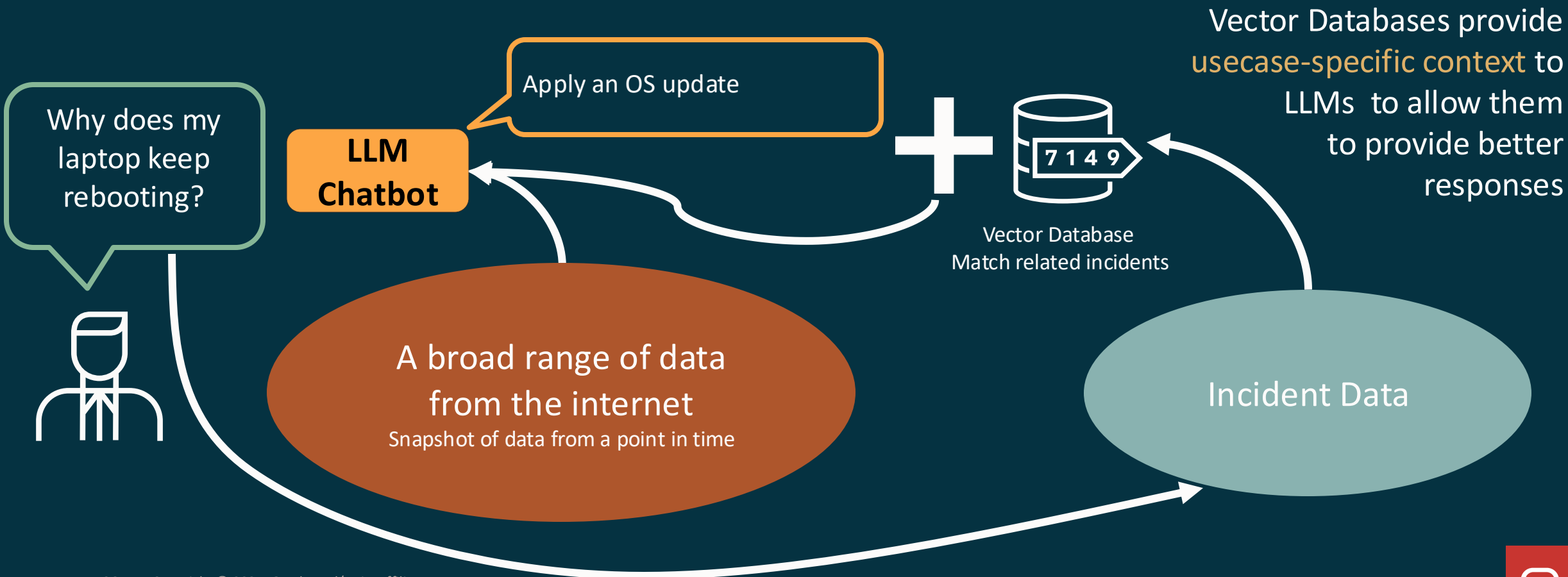


A technique that uses **vector database** content to **augment user-provided prompts** using semantic similarity search with LLMs

RAG enables LLMs to use business data to produce better and more accurate responses and not fine-tune LLMs using that data, which may introduce security concerns

# Role of Vector Databases in Generative AI

When augmented with enterprise information they provide better answers  
Known as Retrieval Augmented Generation (RAG)



Vector Databases provide **usecase-specific context** to LLMs to allow them to provide better responses



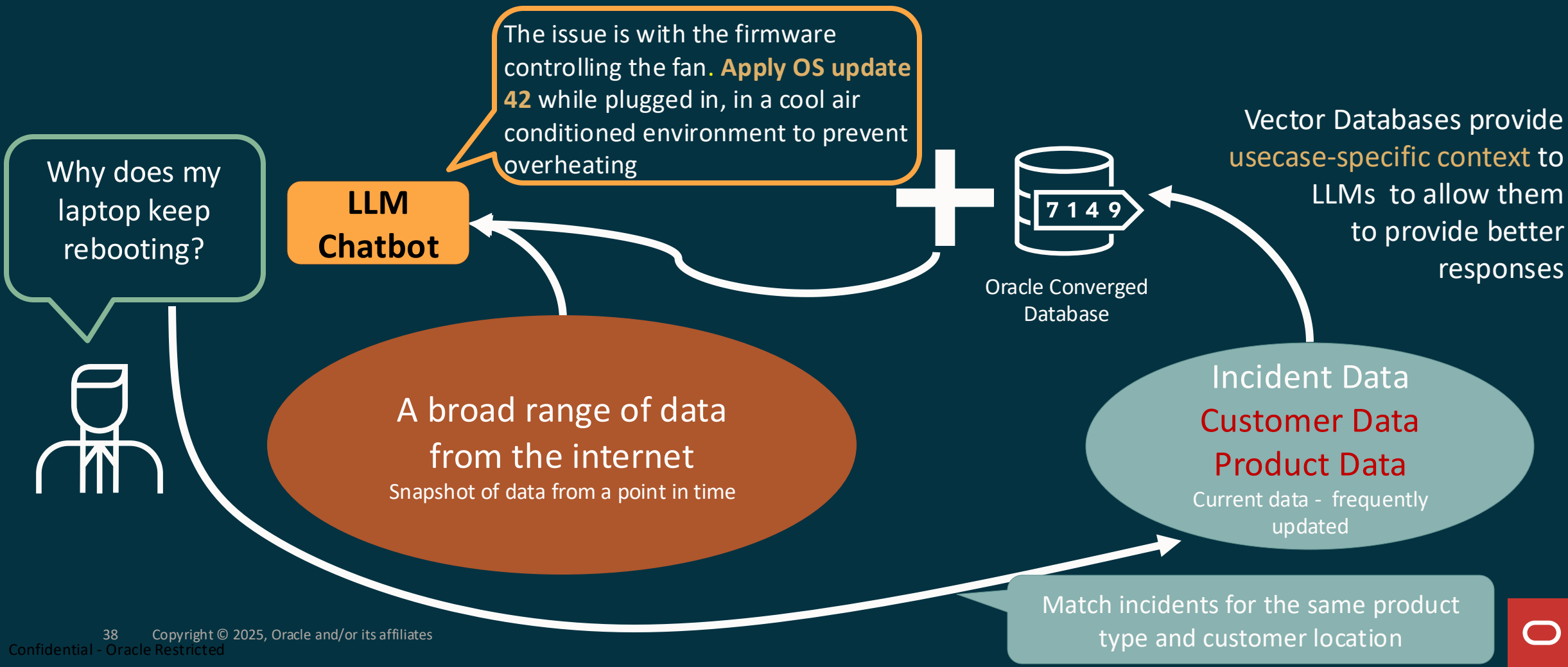
# Role of **Converged Oracle Database** in Generative AI

Oracle **Converged Database** has support for vectors in addition to Relational, JSON, Text, etc.  
No need for data movement, avoids the cost, complexity, and security risk of multiple systems  
Easily combine business data and vector data for ultra-sophisticated interactions with LLMs



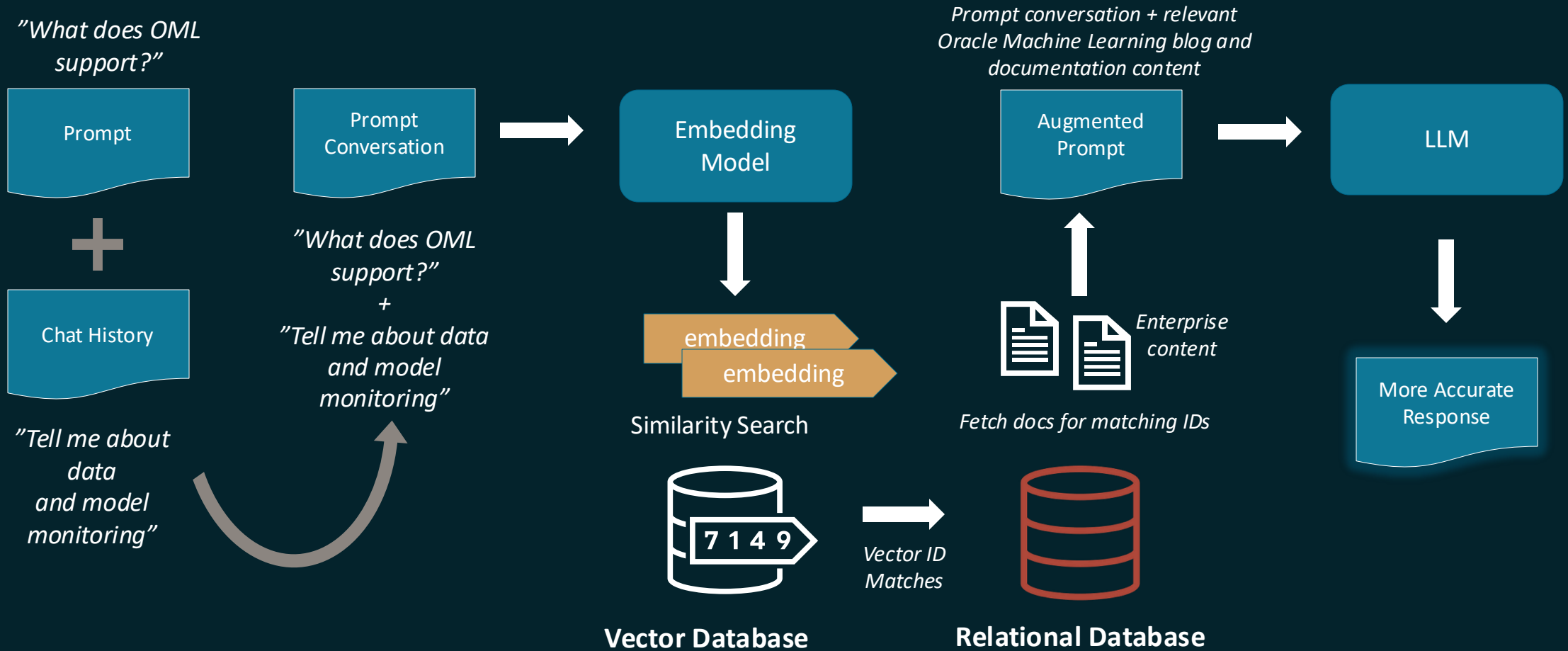
# Role of Converged Oracle Database in Generative AI

Converged Business Databases allow business rules, filters, security policies to be applied to RAG



# RAG pipeline example

Search Oracle Machine Learning blogs to answer question

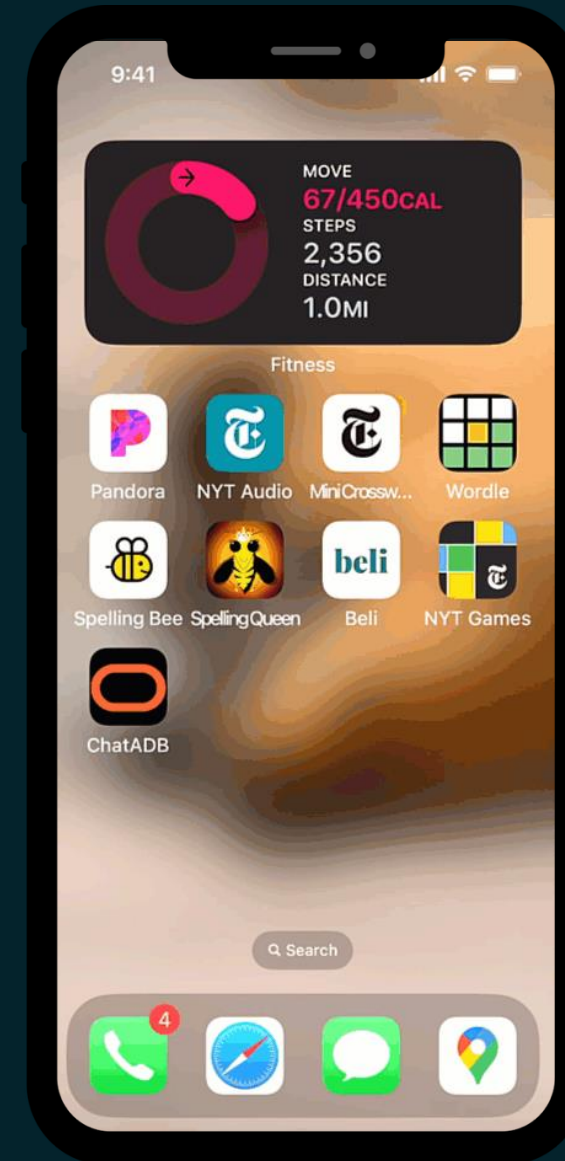


# Select AI

Supporting LLM-enhanced application development and productivity through natural language

NL2SQL

RAG





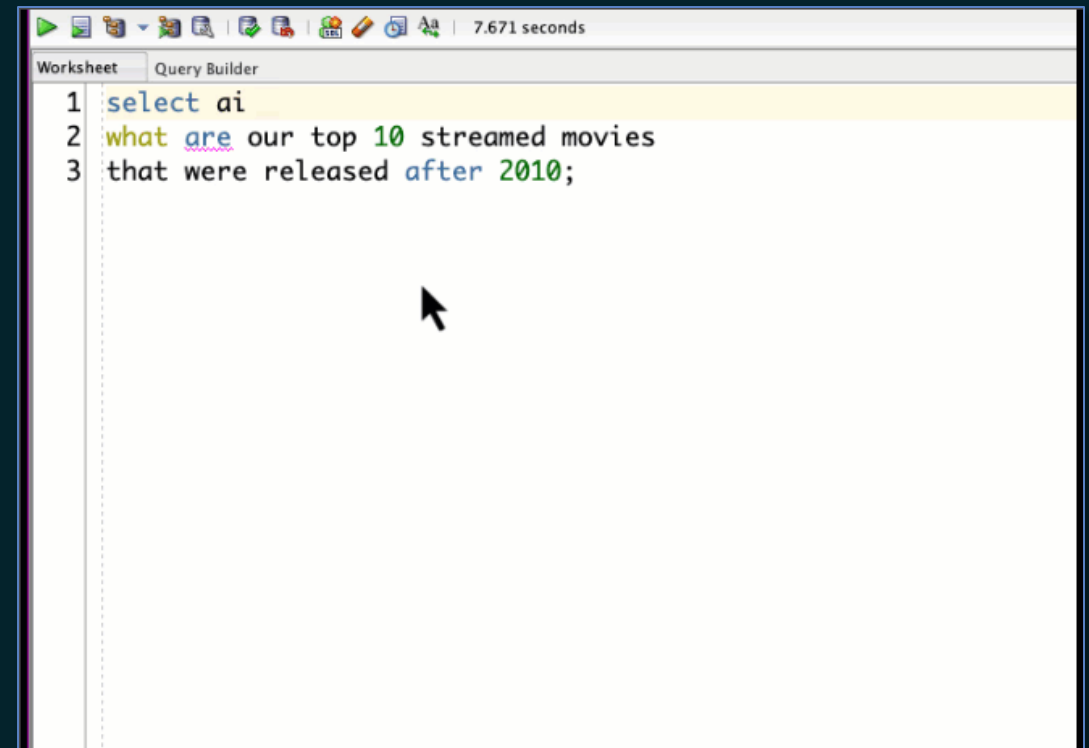
# Use natural language to query data and get responses using generative AI

## Easily access LLMs from multiple AI providers

- OCI GenAI Service
- OpenAI
- Azure OpenAI Service
- Cohere
- Google
- Anthropic
- Hugging Face
- ...more coming

## Actions

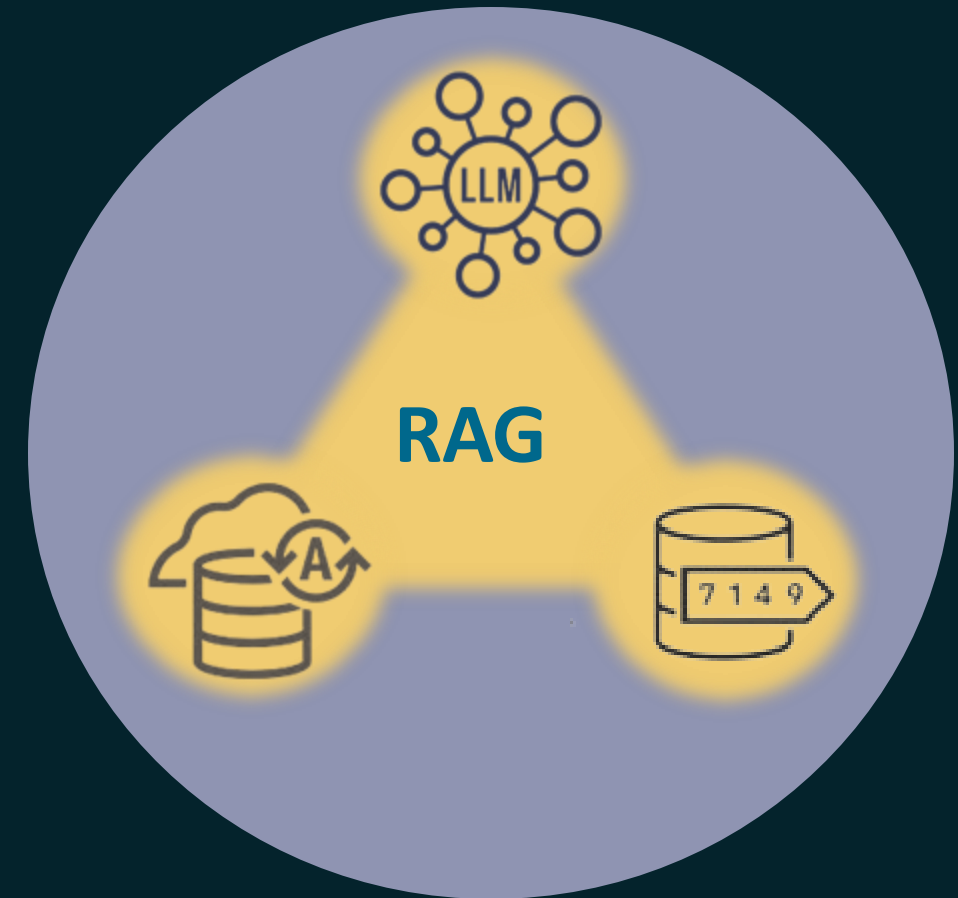
- **showsql** – return NL2SQL generated query
- **runsql** – (default) return SQL result set for NL2SQL  
return vector result set for RAG
- **explainsql** – explain NL2SQL generated query
- **narrate** – return a conversational result for NL2SQL/RAG
- **chat** – return LLM response to prompt – general AI chat



# Use natural language to query your documents

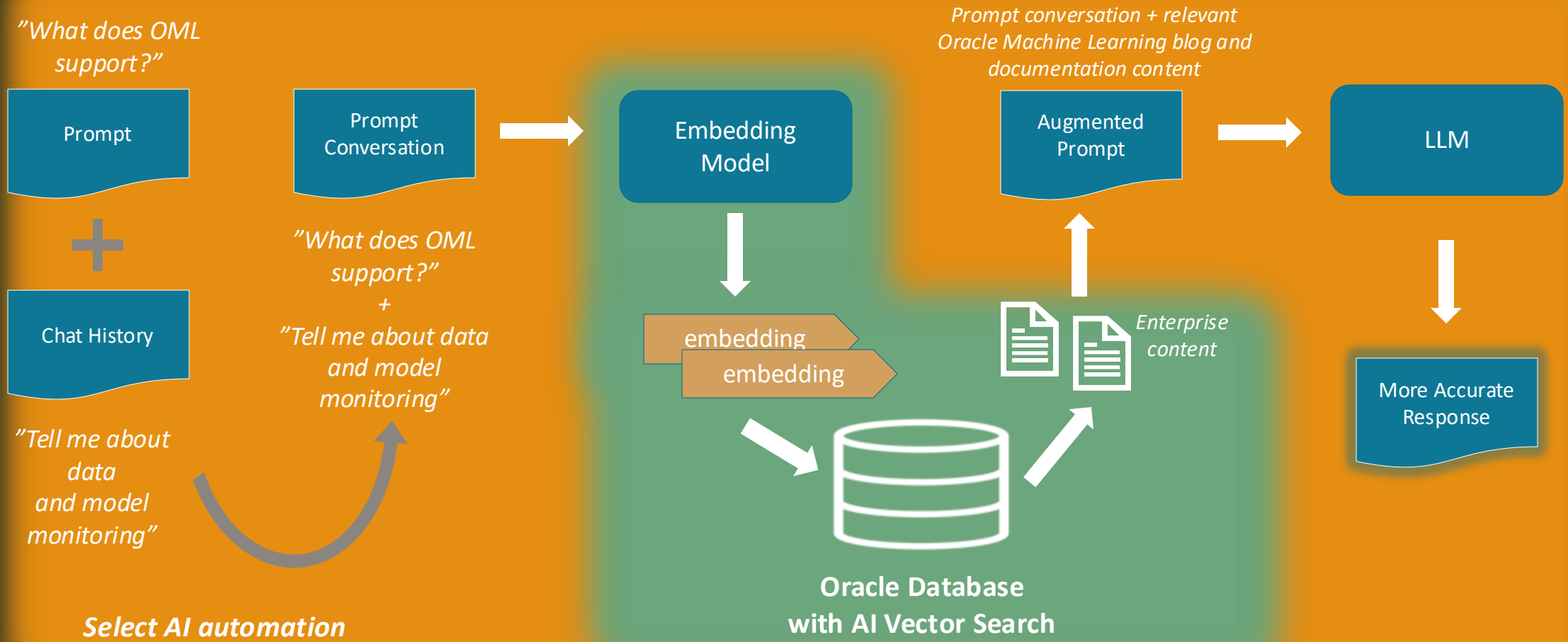
*Select AI Simplifies RAG workflow so any database user can use it!*

- ✓ Give the LLM new knowledge without fine-tuning
- ✓ Use natural language for semantic similarity search and LLM response generation
- ✓ Seamless integration with Oracle AI Vector Search
- ✓ Automate orchestration steps with fully managed Vector Index pipeline for new data



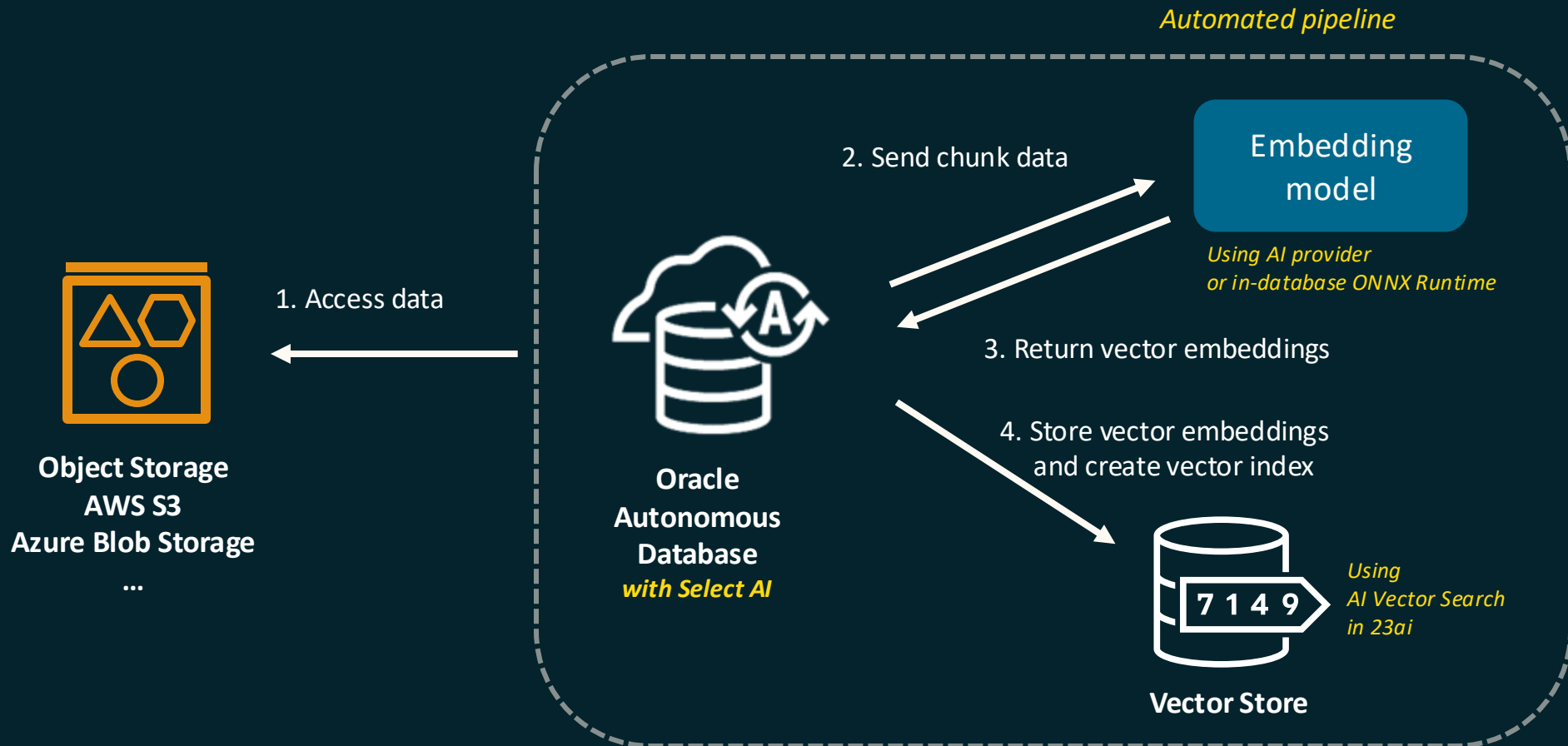
# RAG pipeline example

Search Oracle Machine Learning blogs to answer question



# Select AI Retrieval Augmented Generation (RAG)

Step 1: Create your vector content...automated by Select AI



# Create a RAG-enabled AI Profile for use with 'narrate' and 'runsql'

## Example

```
BEGIN
  DBMS_CLOUD_AI.CREATE_PROFILE(
    profile_name => 'OPENAI_GPT',
    attributes   => '{"provider": "openai",
                    "credential_name": "OPENAI_CRED",
                    "vector_index_name": "MY_VECTOR_INDEX",
                    "temperature": 0.2,
                    "max_tokens": 4096,
                    "model": "gpt-4o",
                    "embedding_model": "text-embedding-ada-002",
                    "enable_sources": true }');
END;
```

*Specify vector index*

*Specify LLM*

*Specify the transformer*

# Create a Vector Index

## Example

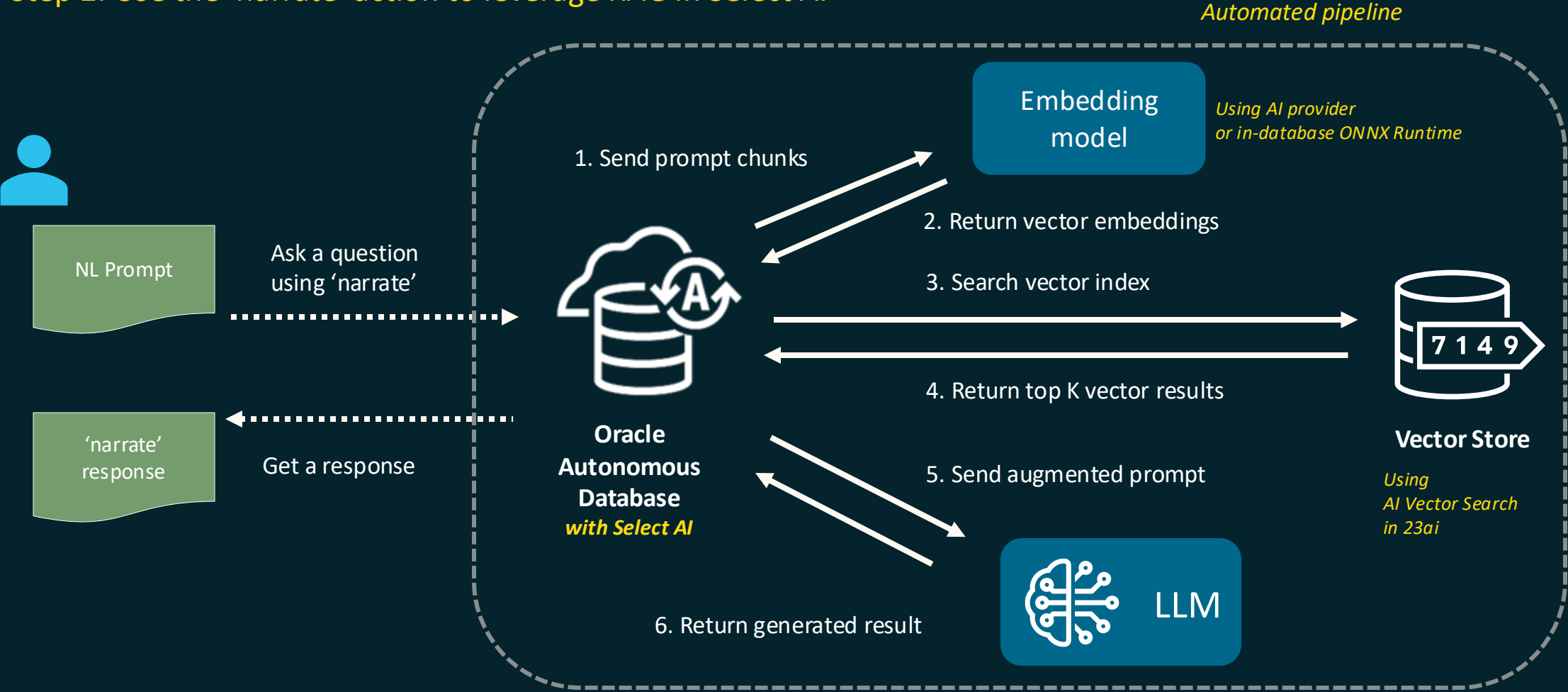
```
BEGIN
  DBMS_CLOUD_AI.CREATE_VECTOR_INDEX(
    index_name => 'MY_VECTOR_INDEX',
    attributes => '{"vector_db_provider": "oracle",
                  "location": "https://my_namespace/my_bucket/my_data_folder",
                  "object_storage_credential_name": "OCI_CRED",
                  "profile_name": "OPENAI_TEXT_TRANSFORMER",
                  "vector_dimension": 1536,
                  "vector_distance_metric": "cosine",
                  "chunk_overlap": 128,
                  "chunk_size": 1024}');
END;
```

*Specify object store location*

*Specify AI Profile*

# Select AI Retrieval Augmented Generation (RAG)

## Step 2: Use the 'narrate' action to leverage RAG in Select AI



# Example using 'narrate' on Oracle Machine Learning blogs

```
EXEC DBMS_CLOUD_AI.SET_PROFILE('OPENAI_GPT');
```

```
SELECT AI NARRATE What type of monitoring is enabled in Oracle Machine Learning
```

## RESPONSE

Oracle Machine Learning enables two types of monitoring: Data Monitoring and Model Monitoring.

Data Monitoring is a no-code user interface that provides insight into how enterprise data evolves over time. It helps to identify data drift and gain insight into individual data features and their interactions. This tool is essential for maintaining data integrity for enterprise applications and dashboards.

Model Monitoring, on the other hand, helps maintain the accuracy of machine learning models and the effectiveness of applications. It detects concept drift and quality drift, tracking and reporting changes in model prediction patterns and accuracy. This helps to know when models need to be rebuilt or when other causes such as data drift need to be investigated.

## Sources:

- [Announcing-OML-Data-Monitoring-User-Interface-on-Autonomous-Database.txt](https://objectstorage.../select_ai_rag_demo_1/o/Announcing-OML-Data-Monitoring-User-Interface-on-Autonomous-Database.txt)  
([https://objectstorage.../select\\_ai\\_rag\\_demo\\_1/o/Announcing-OML-Data-Monitoring-User-Interface-on-Autonomous-Database.txt](https://objectstorage.../select_ai_rag_demo_1/o/Announcing-OML-Data-Monitoring-User-Interface-on-Autonomous-Database.txt))
- [Announcing-OML-Monitoring-on-Autonomous-Database.txt](https://objectstorage.../select_ai_rag_demo_1/o/Announcing-OML-Monitoring-on-Autonomous-Database.txt) ([https://objectstorage.../select\\_ai\\_rag\\_demo\\_1/o/Announcing-OML-Monitoring-on-Autonomous-Database.txt](https://objectstorage.../select_ai_rag_demo_1/o/Announcing-OML-Monitoring-on-Autonomous-Database.txt))
- [Simplify-your-model-monitoring-and-MLOps-with-OML-Model-Monitoring-UI.txt](https://objectstorage.../select_ai_rag_demo_1/o/Simplify-your-model-monitoring-and-MLOps-with-OML-Model-Monitoring-UI.txt)  
([https://objectstorage.../select\\_ai\\_rag\\_demo\\_1/o/Simplify-your-model-monitoring-and-MLOps-with-OML-Model-Monitoring-UI.txt](https://objectstorage.../select_ai_rag_demo_1/o/Simplify-your-model-monitoring-and-MLOps-with-OML-Model-Monitoring-UI.txt))





# MySQL HeatWave

## Generative AI and Vector processing

HeatWave

Single platform for OLTP, OLAP, machine learning and GenAI

## HeatWave<sup>®</sup> Analytics



- Real-time analytics with no ETL
- Accelerate MySQL query by orders of magnitude
- Improve security

## HeatWave<sup>®</sup> Lakehouse



- Run queries across files in object store and database
- Data in object store remains in object store
- Unmatched performance & price-performance

## HeatWave<sup>®</sup> GenAI



- Automated, integrated, in-database GenAI and ML
- Build new applications with no AI/ML expertise required
- No additional cost for in-database LLMs and ML

## HeatWave<sup>®</sup> MySQL



- Managed MySQL EE
- Cloud only new innovations
- MySQL Enterprise support

# Implementation challenges

## Complexity

- External LLM integration
- Separate vector database
- Vector embedding generation
- Difficult to implement natural language capability

## AI expertise

- Embedding model selection
- LLM selection
- Meaningfully apply LLMs, embeddings to domain problems
- Performance optimization

## High costs

- Hiring AI experts
- Provisioning GPUs
- Storing vector embeddings
- Optimizing system resources

# HeatWave GenAI

## In-database, automated vector store

- Automate embedding generation
- AI expertise not required
- No need for a separate vector database

## In-database LLMs

- Invoke LLMs without complex integration
- Secure your data inside the database
- No additional cost for LLM invocation

## Scale-out vector processing

- Perform fast semantic searches on your organization's data
- Deliver rapid and accurate answers to questions

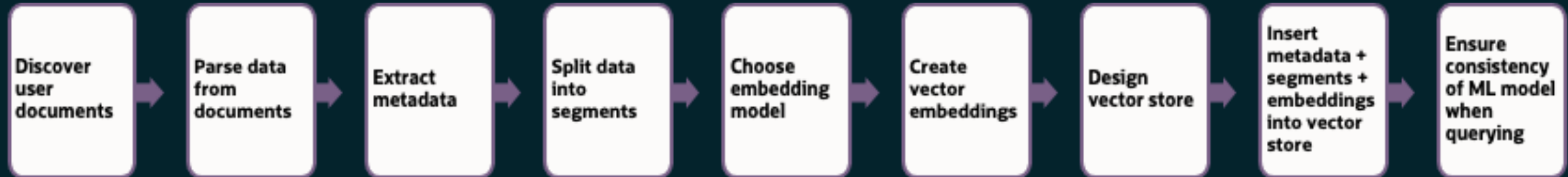
## HeatWave Chat

- Natural language conversations without manual operations
- Easily refine your searches

# In-database, automated vector store

# Building GenAI applications with non-Oracle databases is complex

## Create a vector store



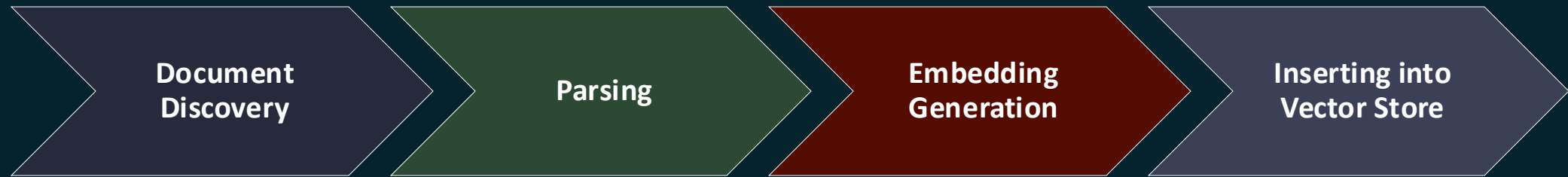
```
sys.heatwave_load(schema_name, @source_location)
```

## Use vector store with LLMs



```
sys.ML_RAG("What is HeatWave?", @NL_response, @optional_search_params)
```

# Automated, In-database Vector Store



# Benefits of HeatWave Vector Store

## Simplicity

- Reduced application complexity: in-database and single-step process
- No AI expertise required
- Data changes are incrementally updated to the vector store

## Lower cost

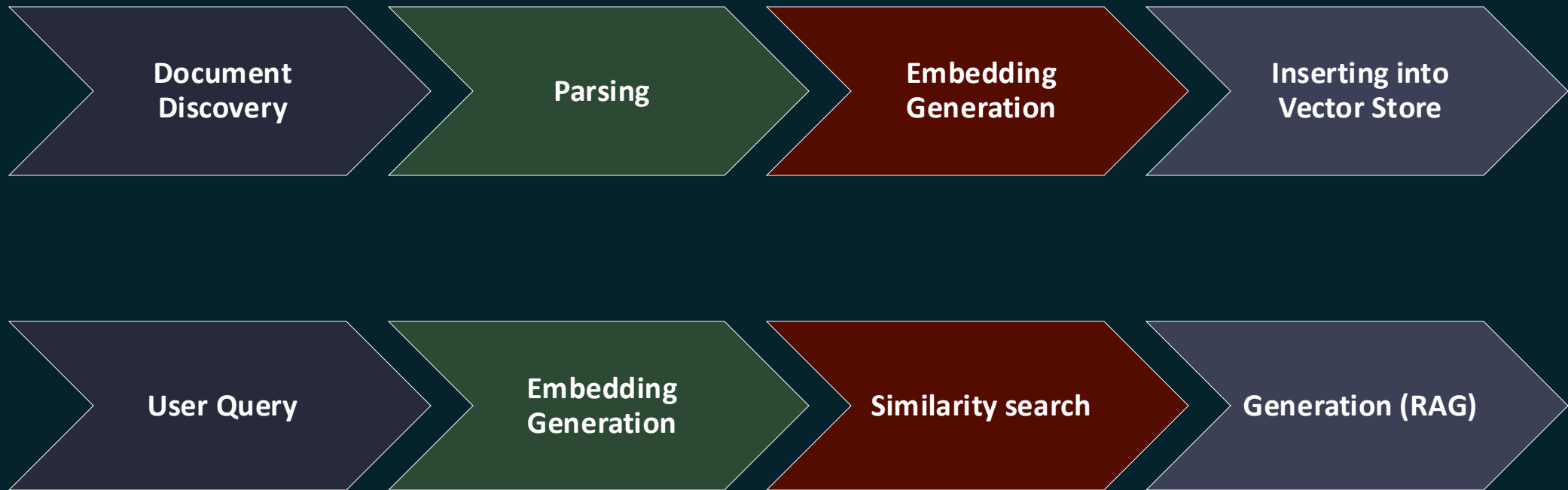
- No additional resources needed
- Vector embeddings are persisted in object storage
- No charge for using the embedding function

## Security and performance

- Data transformation completed inside the database
- Processing parallelized and tuned inside the database

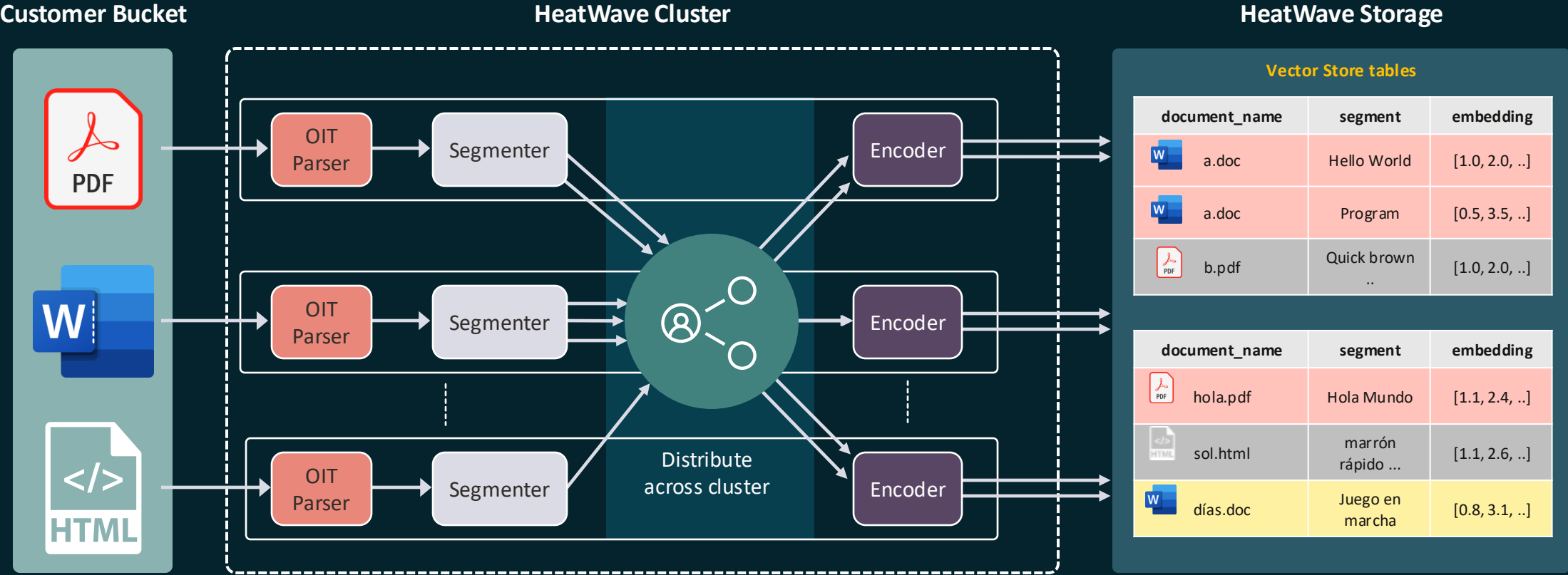


# Automated, In-database Vector Store



# All the steps for vector store creation are completed inside HeatWave

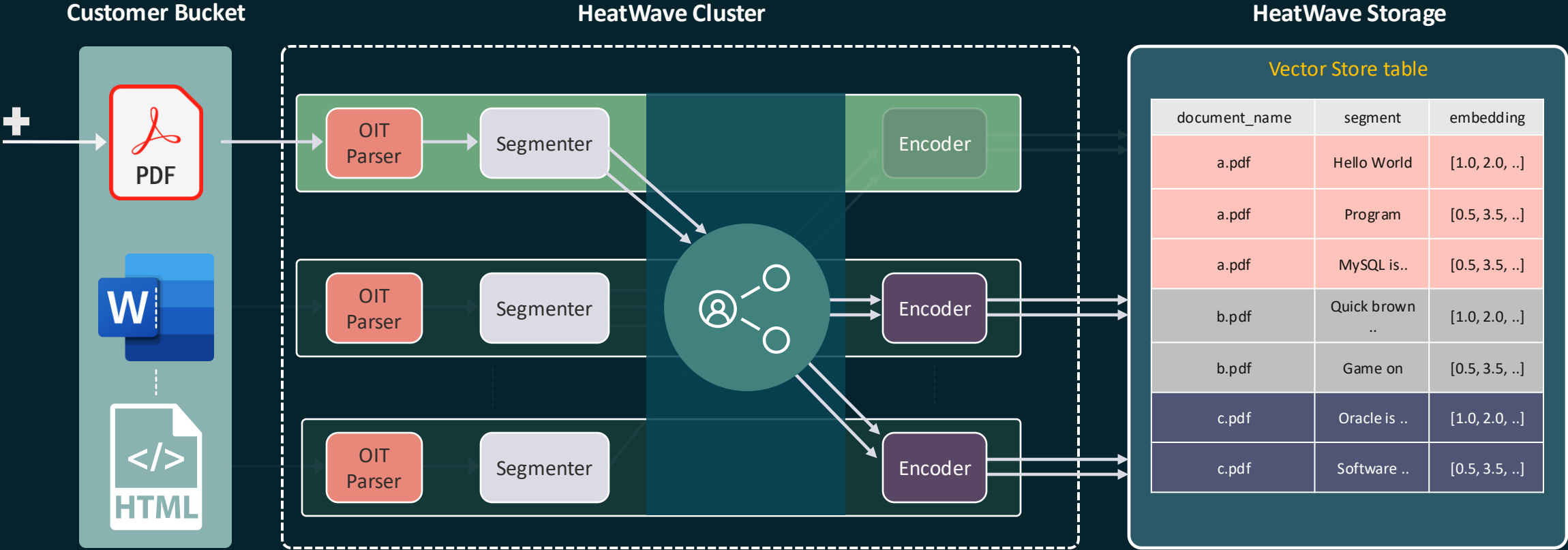
All system resources are optimized by HeatWave



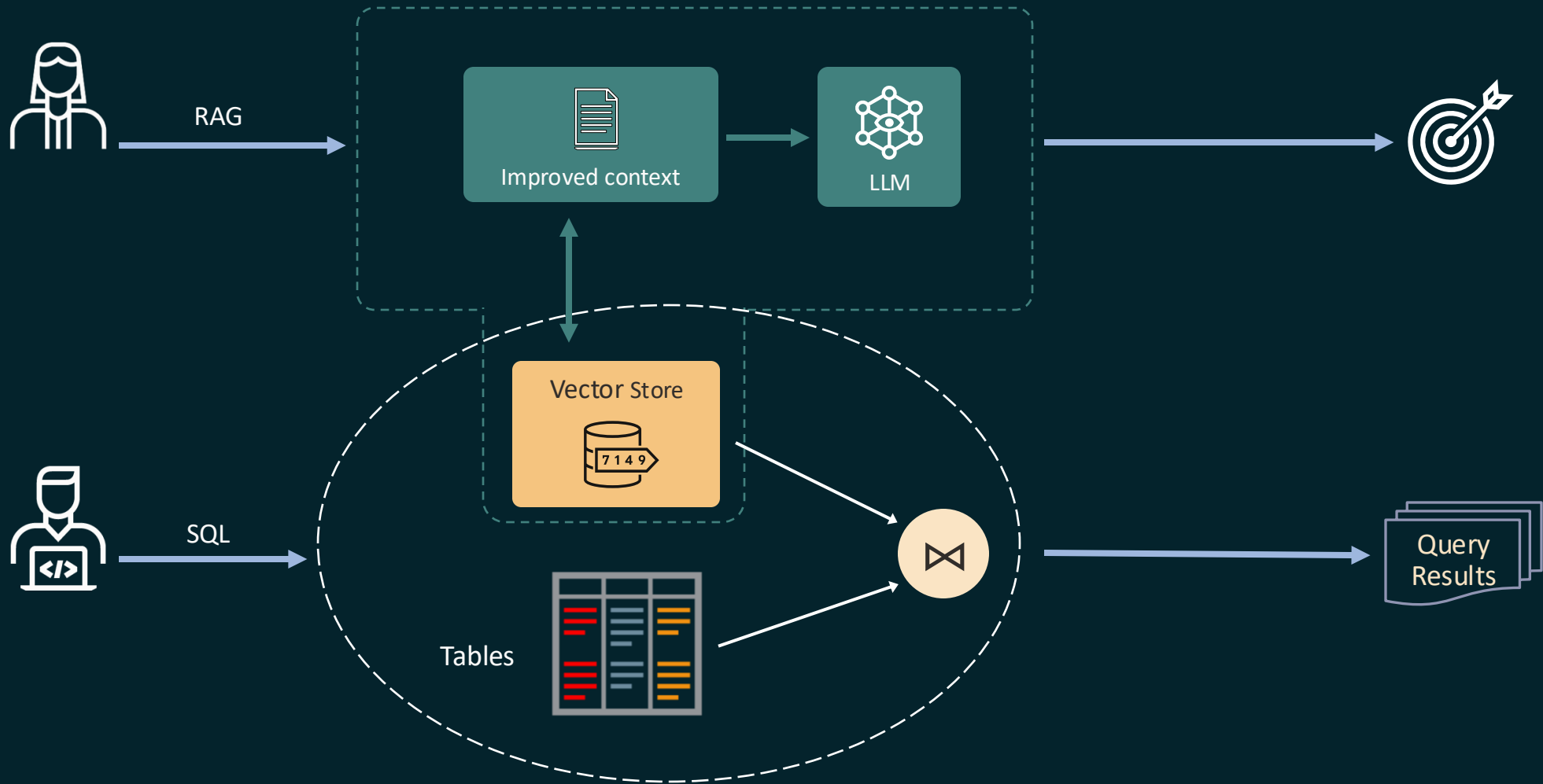
Documents can be in different languages



# Changes to data are automatically updated in the vector store



# Using HeatWave Vector Store: RAG and SQL queries



## New vector datatype in HeatWave MySQL

Vector as  
first-class  
data type

```
mysql> CREATE TABLE wikipedia (  
        title VARCHAR(1024),  
        page_data TEXT,  
        page_url TEXT,  
        page_embedding VECTOR(1024));
```

MySQL  
query  
syntax

```
mysql> SELECT page_url,  
        DISTANCE(page_embedding,  
                  @query_embedding, "COSINE")  
        as distance  
FROM wikipedia  
ORDER by distance DESC LIMIT 10;
```

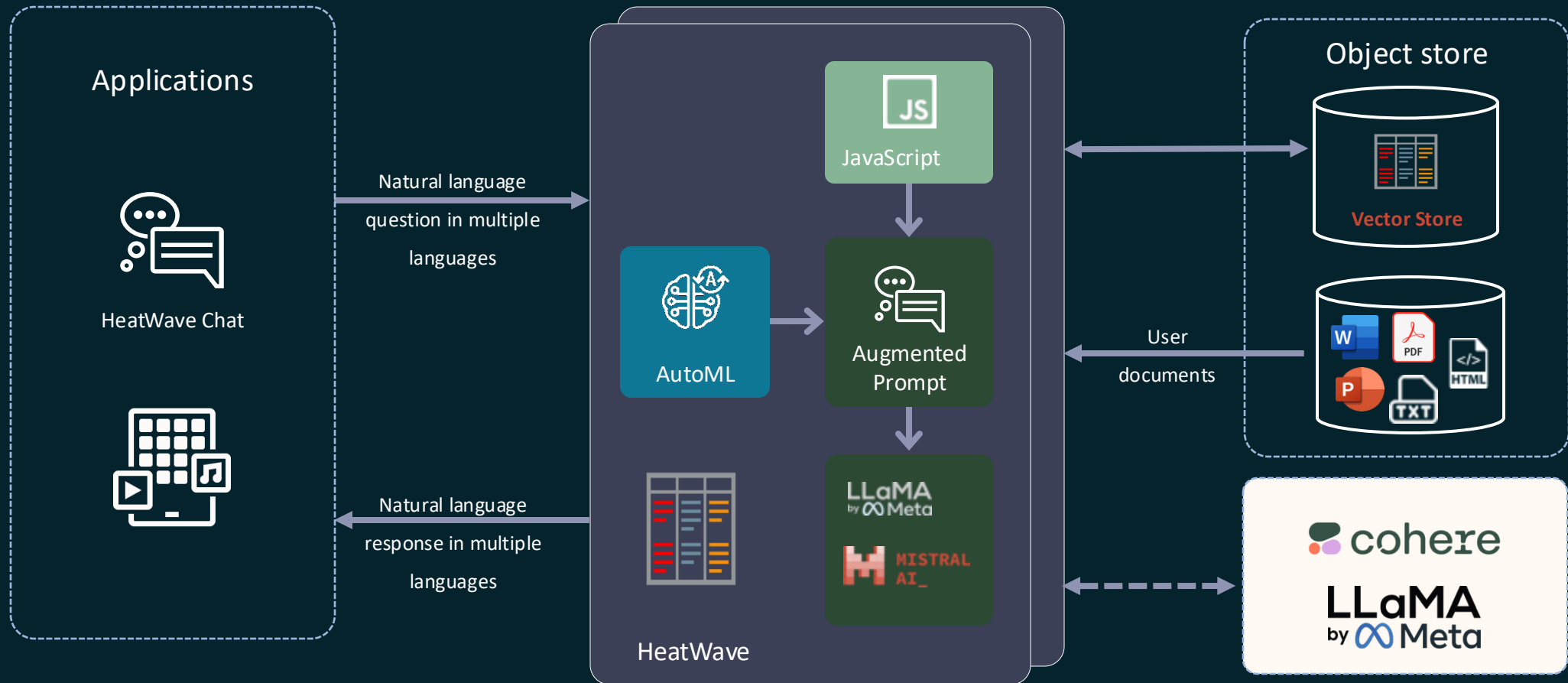
New distance function for similarity search

- L1/MANHATAN
- L2/EUCLIDIAN
- L1^2/MANHATAN\_SQUARED
- L2^2/EUCLIDIAN\_SQUARED
- COSINE
- DOT
- HAMMING

# In-database LLMs

# In-database LLMs and in-database embedding generation

Also integrated with OCI Generative AI service



```
> CALL sys.ML_RAG("¿Qué es HeatWave?", @NL_response, @optional_search_parameters);
```

# Benefits of in-database LLMs

## Simplicity

- No need to select and integrate external LLMs
- Develop turnkey GenAI apps, ready out-of-the-box
- Choose external LLMs if needed for your use case

## Lower cost

- No additional cost to use LLMs
- System resources are optimized

## Flexibility

- Use HeatWave GenAI across regions and clouds, with consistent results across deployments
- Integration with HeatWave AutoML enables new applications and higher quality results

## Security and performance

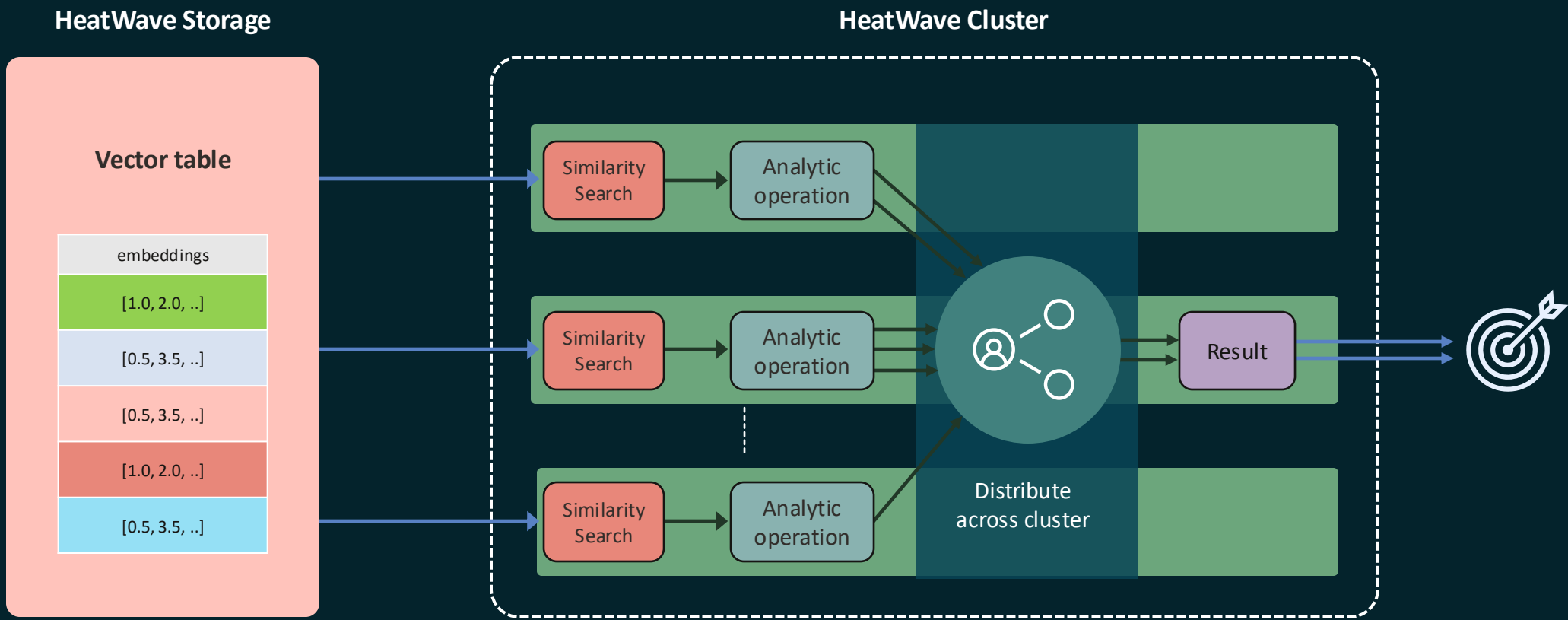
- Data doesn't leave the database - data isolation
- Not a shared service - performance isolation



# Scale-out vector processing

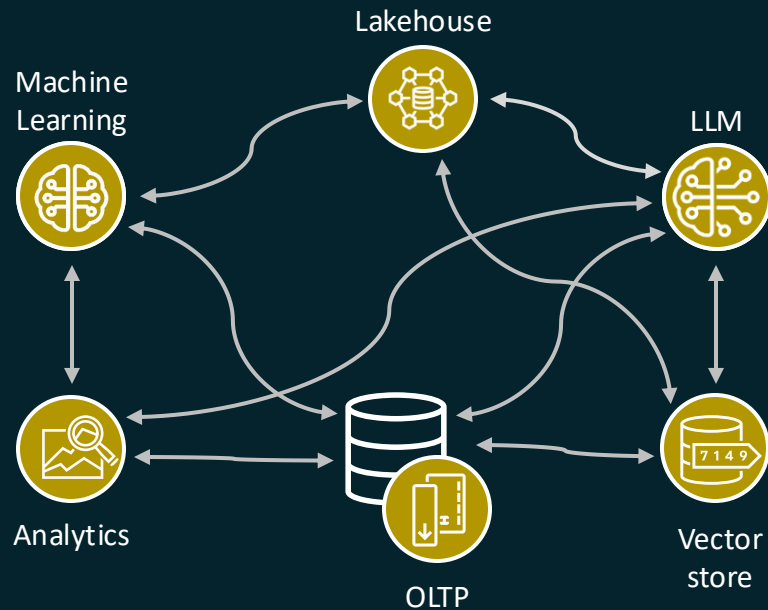
# Similarity search in HeatWave is exact and very efficient

Scales to 512 nodes and can be combined with other predicates



# HeatWave makes security easy

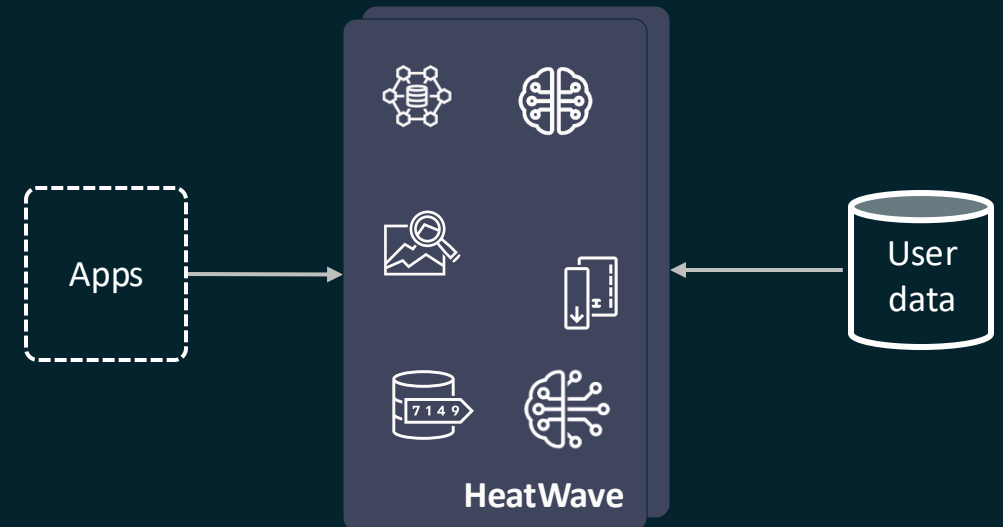
## Other Services



- Large surface area of data movement and exposure
- Different services with varying security postures: encryption keys, user access, authentication schemes
- User needs to configure, connect varied services



## HeatWave



- Data remains in one database system
- Uniform access controls and single configuration
- All communication is authenticated and encrypted

# Get started with HeatWave



## Continue the discussion with us or a partner

Let's further discuss your requirements and determine how we can help.



## Request a free workshop

To help you evaluate or get started with HeatWave.



## Try HeatWave

Build and run small-scale apps using free HeatWave resources for an unlimited time.

Learn more: [oracle.com/heatwave](https://oracle.com/heatwave)

ORACLE