Agenda

- Introduction to Oracle AI Vector Search
- Traditional Document Search
- Demo AI Vector Search with Generative AI APEX Application
- Code Innovate Program Overview

Copyright © 2024, Oracle and/or its affiliates

AI Vector Search

Breakthrough technology that **simplifies** and **optimizes** searching for documents, images, patterns, and data that have similar semantics



Copyright © 2024, Oracle and/or its affiliates

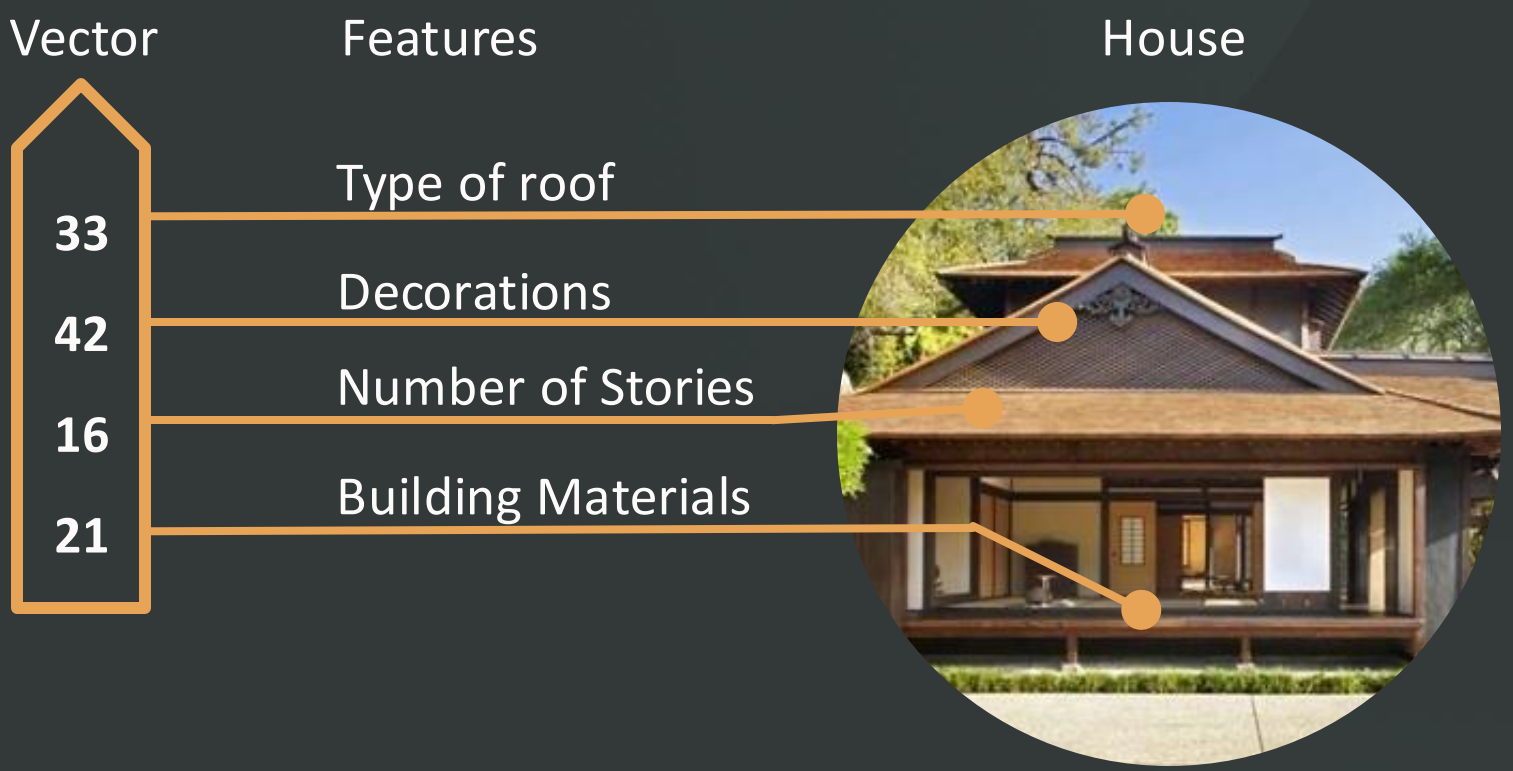
Vectors represent the content (semantics) inside images, documents, data, etc.



A vector is a sequence of numbers, called dimensions, used to capture the important “features” of the data

Copyright © 2024, Oracle and/or its affiliates

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

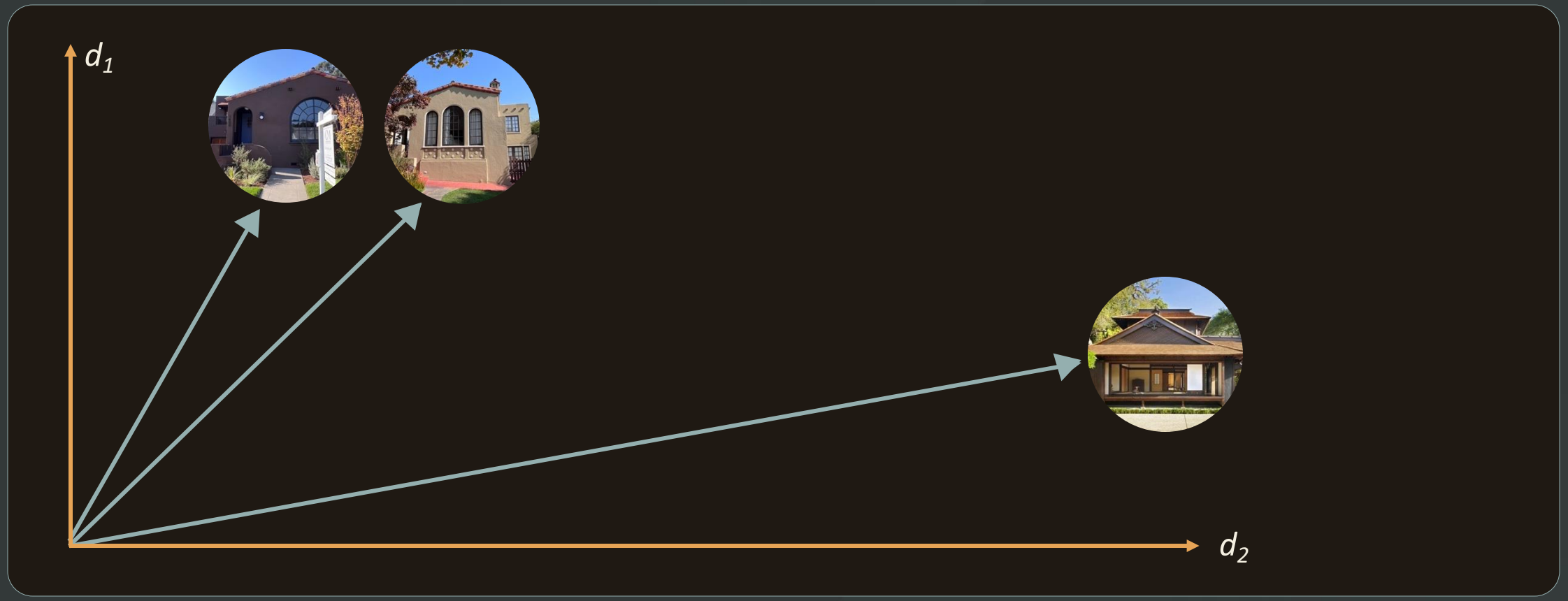


Each dimension (number), represents a feature of the house

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

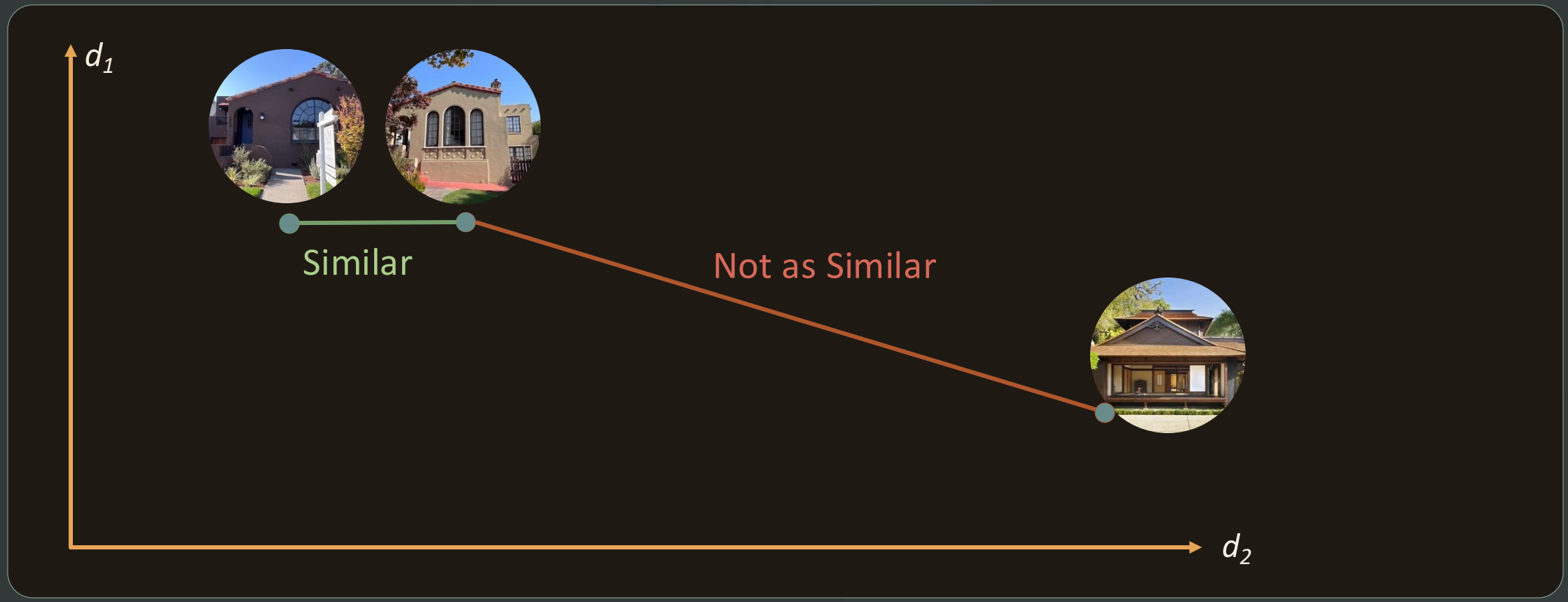
Copyright © 2024, Oracle and/or its affiliates

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



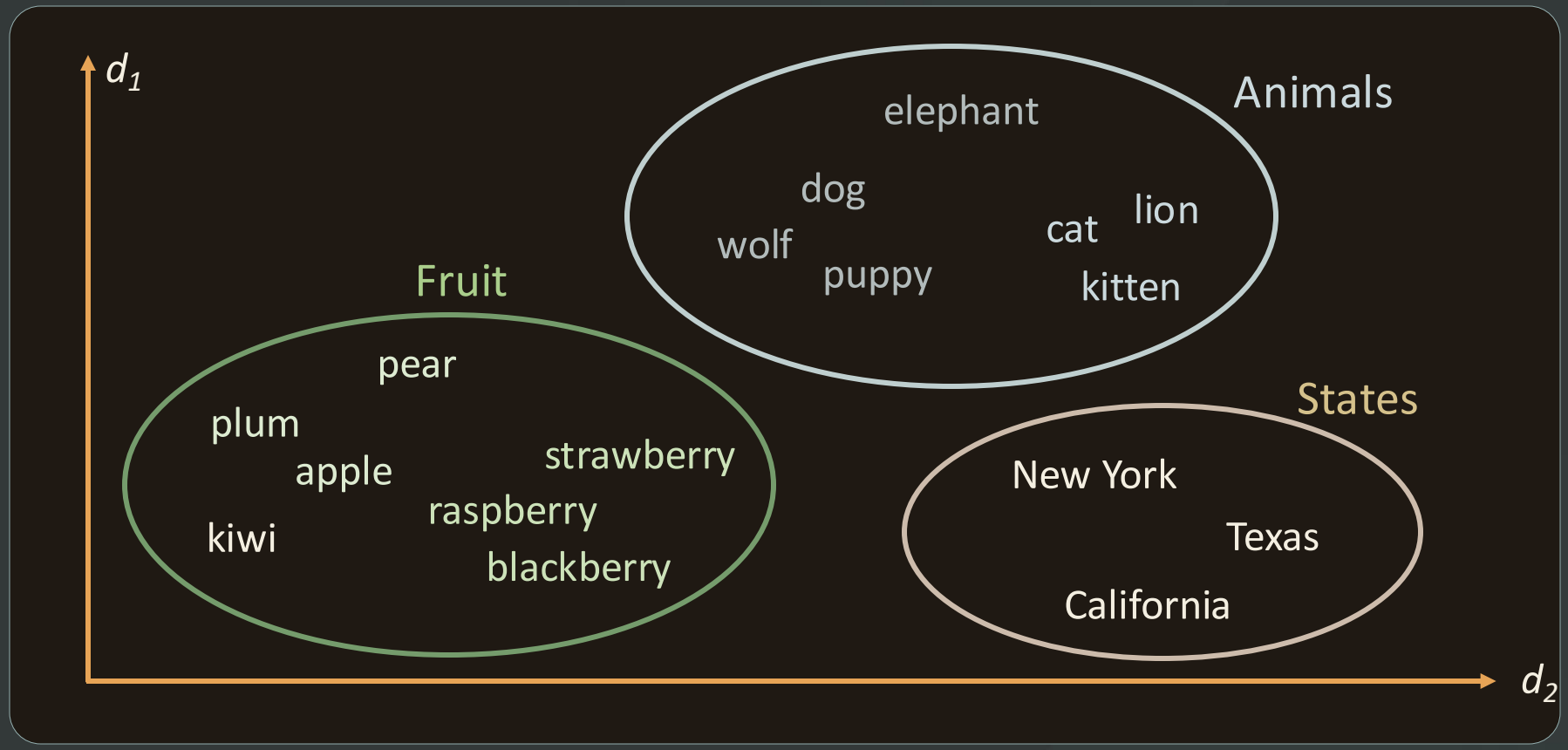
Copyright © 2024, Oracle and/or its affiliates

The distance between the vectors is proportional to their semantic similarity



Copyright © 2024, Oracle and/or its affiliates

Word similarity works the same way
Word vectors that are near each other are semantically similar



Documents also work the same way

Document vectors that represent similar content are closer in distance

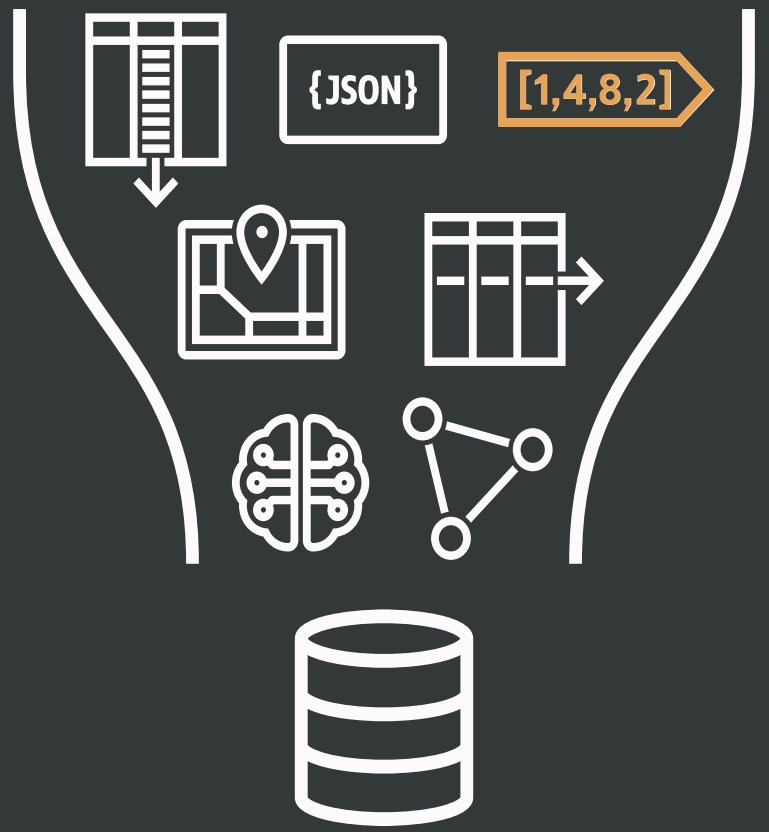
Copyright © 2024, Oracle and/or its affiliates

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



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

Copyright © 2024, Oracle and/or its affiliates



Converged Database

That’s why we added AI Vector Search to the leading repository of business data – Oracle

- Users can now combine business data and vectors to answering a question
- Eliminates the need to move and synchronize data, manage multiple products, etc.

Copyright © 2024, Oracle and/or its affiliates

Oracle Vector Search – Integrated Database 23ai

- AI Vector Search includes
  - Vector data type
  - Vector indexes
  - Vector SQL Functions and Operators
    - Chunking and Embedding
    - Calculate Distance
    - Constructors, Converters, and Descriptors
    - JSON Compatibility with the VECTOR Data Type
  - Retrieval Augmented Generation (RAG) support
- Ability to enable LLM-based natural language interface for applications built on Oracle Database



Copyright © 2024, Oracle and/or its affiliates



Adding semantic search to relational search is great, but we can do even better by adding Generative AI (LLMs)

Copyright © 2024, Oracle and/or its affiliates



Oracle AI Vector Search improves Generative AI
by augmenting prompts with private database content

This helps produce better and more accurate answers to user questions

Called: Retrieval Augmented Generation (RAG)

Copyright © 2024, Oracle and/or its affiliates

Retrieval Augmented Generation works like this

1 A user’s natural language question is encoded as a vector and sent to AI Vector Search



2 AI Vector Search finds private content such as documents stored in the database that match the user's question



User



3 The user's question is augmented with the content



GenAI

4 GenAI uses the question plus the content and general knowledge to provide an informed answer

Copyright © 2024, Oracle and/or its affiliates

Summary

Oracle holds most of the world's operational and enterprise data

Businesses want to query on both their business data and on vectors to answer questions meaningfully and accurately

With AI Vector Search, you can add vector search to your business data to accurately answer questions and augment Generative AI

- Don't have to move your business data out to a separate vector database
- Avoid data synchronizations issues
- Avoid complexity of managing multiple products and avoid added costs



Oracle leads AI Vector Search with its differentiators

- **Search Sophistication** – easily combine relational and vector search
- **Scale-Out** – scale your workload over RAC, Sharding, Partitioning
- **Indexing** – fast & accurate indexes w/ filtering & transaction support
- **Simplicity** – ground-breaking technology packaged in simple SQL that a DBA can understand and use in minutes!

Copyright © 2024, Oracle and/or its affiliates

Traditional Document Search

- Document/text search historically relied on words or phrases that either matched or sounded like the words and phrases you were interested in.
- Query functions use contains, stop words, matching, and document section queries, relying on document fragments, with no ability to consider the meaning of the entire document.
- Keyword search has many limitations, including lack of context and multiple meanings for the same word.
- Workarounds involved building classification rules and indexing.
- NONE of the text functions provide any interpretation of the query results.

Copyright © 2024, Oracle and/or its affiliates

Other Supported AI Vector Search Solutions

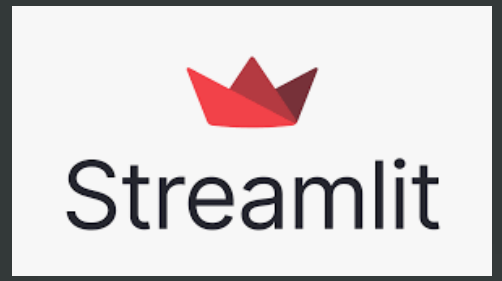


















Copyright © 2024, Oracle and/or its affiliates

Today’s Webinar Solution

A pure OCI solution leveraging SQL and PL/SQL to load, vectorize, search, and interpret questions about document content, and optionally combine that with other enterprise data. We'll use the popular APEX to show an application that allows end users to upload documents and search across those documents to understand and interpret their content using Large Language Models.

- Completely contained within the Oracle Cloud
- Leverages LLM models hosted by OCI Gen AI Service
- Targets database developers and DBAs that leverage SQL and PL/SQL
- Does not require any knowledge of Python or any use of open-source libraries such as LangChain
- Solution can be easily integrated into existing and new enterprise applications
- Demonstrates the entire process from beginning to end
- Provides code for download and local experimentation

Copyright © 2024, Oracle and/or its affiliates

Credentials Overview

dbms_cloud.create_credential

```
begin
  dbms_cloud.create_credential (
    credential_name => {enter_credential_name}’,
    user_ocid       => 'ocid1.user.oc1……’,
    tenancy_ocid    => 'ocid1.tenancy.oc1……’,
    private_key     => '-----BEGIN PRIVATE KEY-----
                        M……
                        -----END PRIVATE KEY-----’,
    fingerprint     => '3e:……:ee’);
end;
/
```

dbms_vector.create_credential

```
declare
  jo json_object_t;
begin
    jo := json_object_t();
    jo.put('user_ocid','ocid1.user.oc1……’);
    jo.put('tenancy_ocid','ocid1.tenancy.oc1……’);
    jo.put('compartment_ocid','ocid1.compartment.oc1…….’);
    jo.put('private_key','M----------------’);
    jo.put('fingerprint','3e:……:ee’);
    dbms_output.put_line(jo.to_string);
    dbms_vector.create_credential(
    credential_name   => '{enter_credential_name}’,
    params            => json(jo.to_string));
end;
/
```

Copyright © 2024, Oracle and/or its affiliates

Code Innovate Program



Copyright © 2024, Oracle and/or its affiliates


Legoland




Master Builder Class


Results

What is a **Virtual Code Innovate?**

5 Half Day Online Event to Jumpstart Cloud Development

- Hands on Introduction to Cloud Services
- Our Engineers Work Online with Your Team
- Facilitates Rapid Knowledge Sharing
- Leverages Your Use Cases
- No cost - if minimum requirements are met



Copyright © 2024, Oracle and/or its affiliates

Code Innovate Experience



Bring your use cases

We work on your use cases instead of showing canned demos of our offerings.



Rapid Knowledge Share

Oracle Engineers work with your team 4 hours each day.



Build a First Pass Prototype

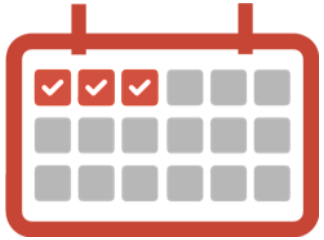
Our Engineers help build a first pass working prototype of your use case solution.

Copyright © 2024 Oracle and/or its affiliates

Requirements



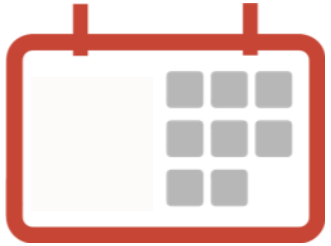
4 or more
Developers



20 hr
Time Commitment
½ Day Event,
5 Days



Bring your Use
Case’s



~3 Weeks
Lead Time



Contact Us to Schedule
Your Event
codeinnovate_us_grp@oracle.com

Copyright © 2024, Oracle and/or its affiliates

Contact us to schedule your event!



OR



Email Us at:

codeinnovate_us_grp@oracle.com

Enter your emails in the **zoom** chat and we will contact you

Copyright © 2024, Oracle and/or its affiliates