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.

Announcing Oracle Database 23c – The next Long Term Support Release

New In 23c

SQL Domains

Schema Privileges

Oracle Database 23c App Simple

Read-Only Per-PDB Standby

Property Graphs

Real-time SQL Plan Management

JSON Schema

Microservice Support

JSON / Relational Duality

AI Vector Search

True Cache

SQL Firewall

Priority Transactions

Globally Distributed Database

Rolling Patching

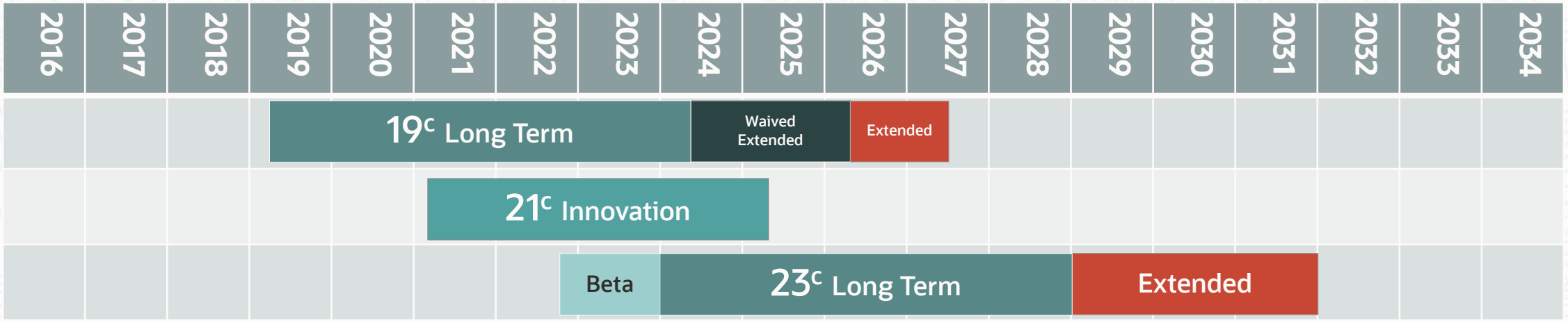
JS Stored Procedures

Developer Role

Shrink Tablespace

Boolean Datatype

Projected Database Release and Support Timeline



- Innovation Release - 2 years of Premier Support, and no Extended Support
- Long Term Release - 5 years of Premier Support, and 3 years of Extended Support
- Always refer to MOS Note: Release Schedule of Current Database Releases (Doc ID 742060.1)

ORACLE

Enabling Generative AI with Oracle AI Vector Search

Dominic Giles
Distinguished Product Manager
Oracle Database Development
@dominic_giles

What is similarity?

We are all very comfortable as human beings judging similarity

Are these two fruits similar?

Orange

Tangerine

Much more challenging to **teach** a computer what similarity is

Are these two fruits similar?

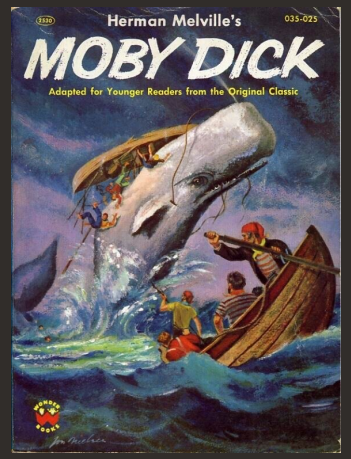
Orange

Tangerine

Much more challenging to **teach** a computer what similarity is

Are these books similar?

Moby Dick Finding Nemo



And yet we seemingly have....

Let’s clear up some terminology….

Straight to a Demo….

Experiments with LLM/Transformers

Let's start with a few terms

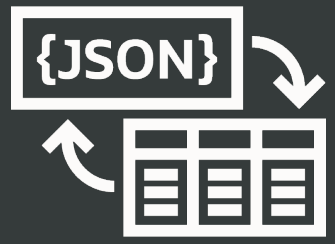
- **Transformers** : A transformer is a type of neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al. in 2017. Transformers are designed to process sequences of data, such as sentences in natural language, and have since become a fundamental building block in various deep learning applications.
- **LLMs** : a large language model, like GPT-3 (Generative Pre-trained Transformer 3), is a specific implementation of a transformer architecture that has been pre-trained on a massive amount of text data to perform language-related tasks. GPT-3, for instance, is a powerful language model developed by OpenAI that can generate coherent and contextually relevant text given a prompt. It can be fine-tuned for specific tasks, such as translation, summarization, question answering, and more.
- **Embeddings** : An embedding is a way to represent and encode objects, such as words or items, as vectors in a continuous vector space. Embeddings are commonly used in machine learning and natural language processing (NLP) to capture semantic relationships and similarities between objects.
- **Vector** : an embedding is essentially a vector, but the term "embedding" is often used in the context of machine learning to emphasize the idea that objects are being placed into a continuous vector space for the purpose of capturing meaningful relationships and patterns.

Source : Responses from questions to OpenAI ChatGPT

We'll be using Python (3.11), Pytorch (a Machine Learning Library), SentenceTransformer (A tool to generate text and image embeddings based

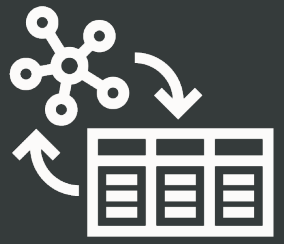
Game-Changing Technologies Coming Soon

Best Document Database



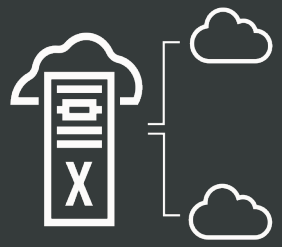
Oracle Database with JSON Duality

Best Graph Database



Oracle Database with Native Graph

Best Multicloud Database



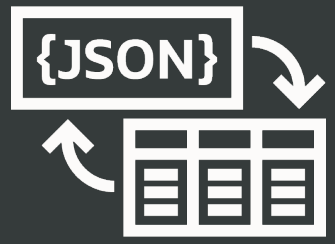
Oracle Database@Azure and others…

Best AI Database

Oracle Database with AI Vector Search

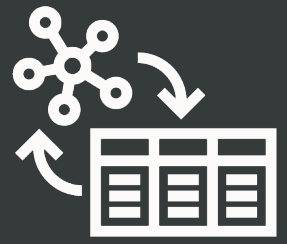
Game-Changing Technologies Coming Soon

Best Document Database



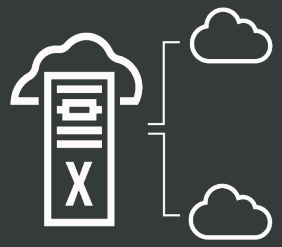
Oracle Database with JSON Duality

Best Graph Database



Oracle Database with Native Graph

Best Multicloud Database



Oracle Database@Azure and others…

Best AI Database

Oracle Database with AI Vector Search

Oracle Database is a Machine Learning Engine

Continuously Enhanced for over 20 years

Oracle makes it simple to make data-driven predictions

- Use declarative SQL to build models and run Machine Learning directly on business data

Over 30 in-database parallel and scalable ML algorithms

- Eliminates costly, risky, and slow data movement to separate ML engine

Business Use Case	ML Technique	Oracle ML Algorithms
Customer Loyalty and Retention	Classification	SVM, GLM, Random Forest, XGBoost, et al.
Customer Segmentation	Clustering Classification	K-Means, Expectation Maximization Decision Tree
Demand Forecasting	Time Series Regression	Exponential Smooth SVM, GLM, Neural Networks, XGBoost
Cross-sell / Up-sell	Association Rules Classification	A priori SVM, GLM, Random Forest, XGBoost, et al.
Credit Risk	Regression Classification	SVM, GLM, Neural Networks, XGBoost SVM, GLM, Random Forest, XGBoost, et al.

Oracle AI Vector Search

Example of importing ONNX (Open Neural Network eXchange) embedding model from Object Storage

```
DECLARE
  model_source BLOB := NULL;
BEGIN
   model_source := DBMS_CLOUD.get_object(
      credential_name => 'OBJ_STORE_CRED',
      object_uri      => 'https://objectstorage…bucketname/o/resnet50bundle.onnx');

  DBMS_DATA_MINING.import_onnx_model(
    model_name => "resnet50",
    model_data => model_source,
    metadata   => JSON('{ function : "embedding" }')
);
END;
```

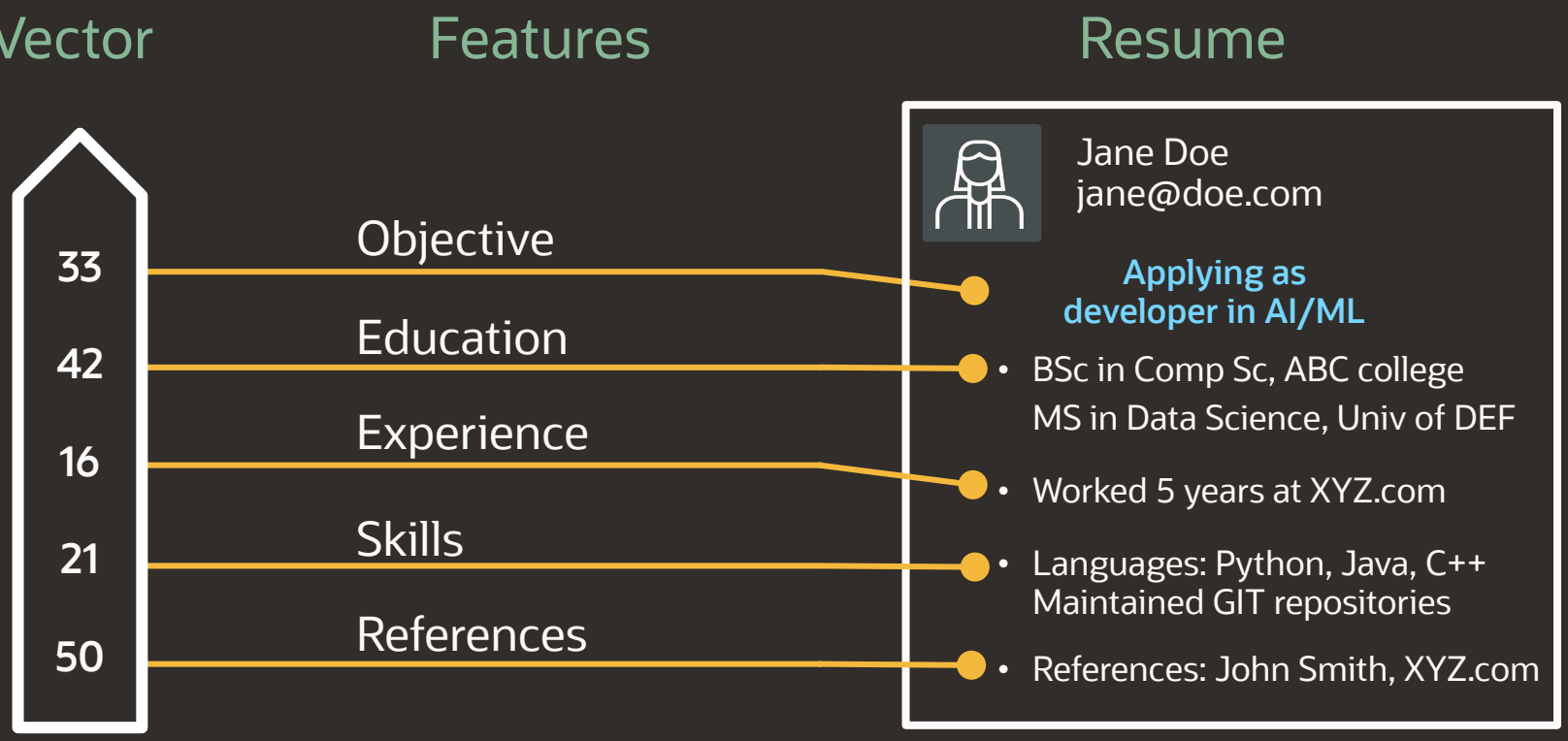
A new technology called AI Vector Search enables semantic searches on unstructured data

Vectors are used in AI to encode unstructured data such as images, documents, videos, etc.

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

Vectors represent the semantic content of data, not the actual words in a document or pixels in an image

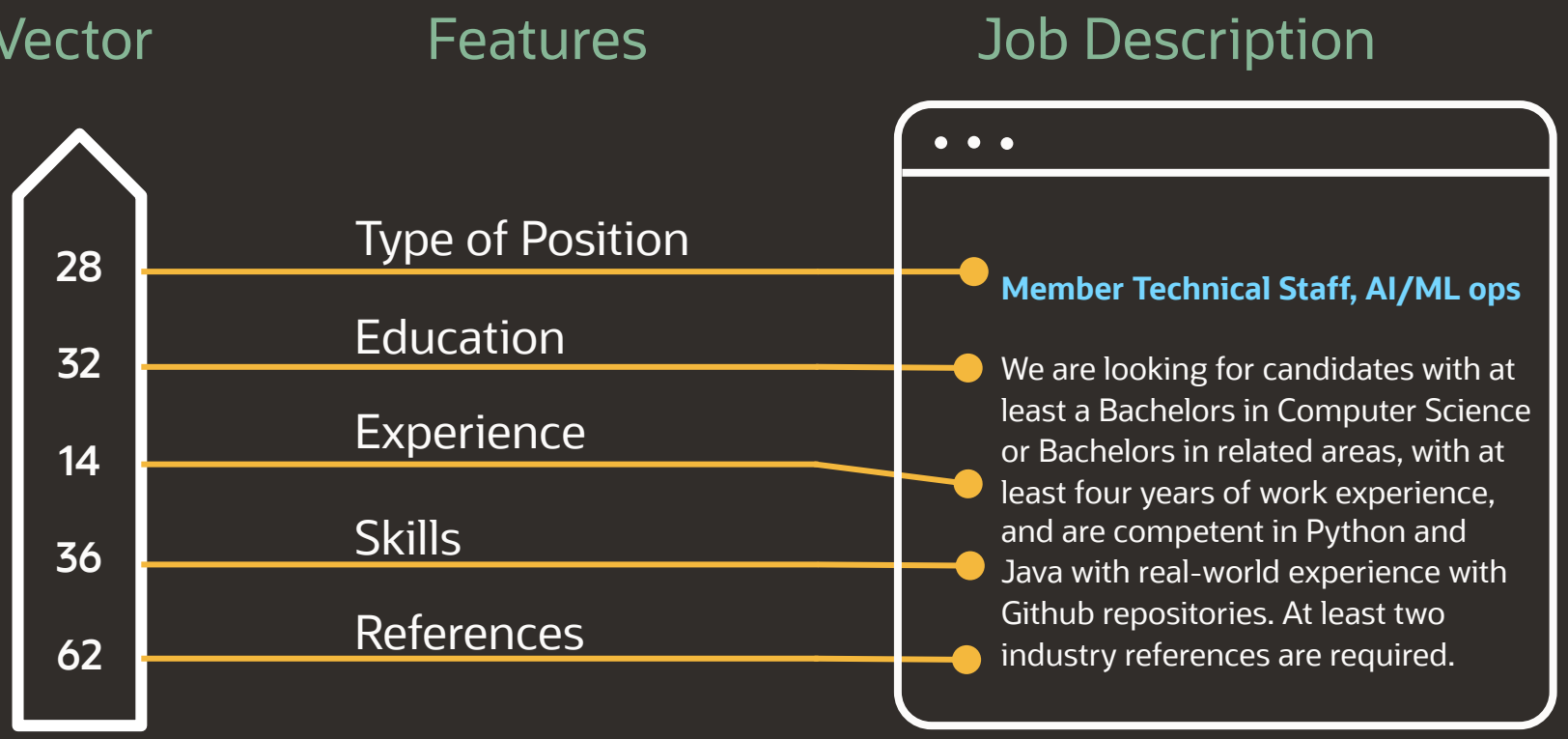
Example: The features for a resume could be



Each dimension (number), represents a different feature of the resume

Note: Features determined by actual AI models are not as simple as shown here

Example: … features for a job posting could also be



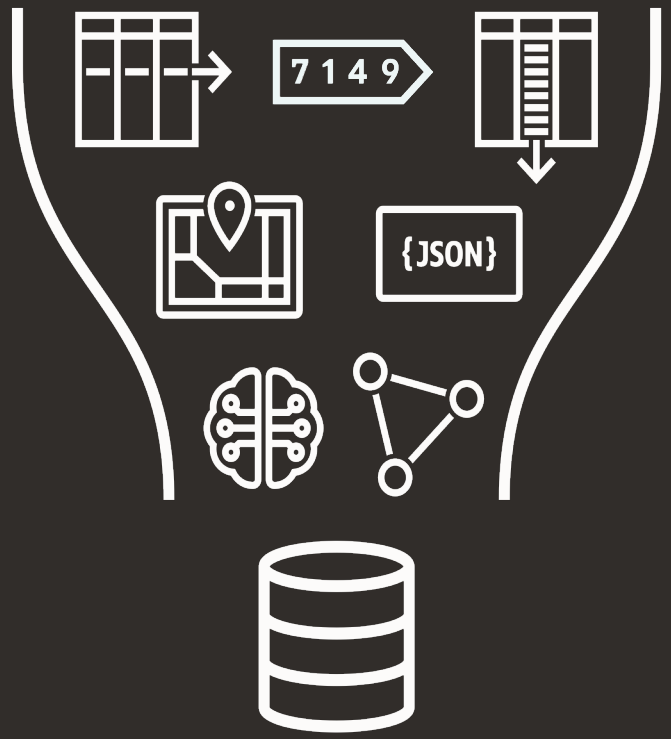
Each dimension (number), represents a different feature of the Job Posting

Note: Features determined by actual AI models are not as simple as shown here

The main operation on vectors is the mathematical distance between them

Distance (Euclidean Norm)
= SQRT((3-2)^2+(1-6)^2+(2-2)^2+(8-3)^2)

There are many mathematical distance formulas



Converged Database

Announcing:

AI Vector Search in Oracle Database

- Vector Data Type and Vector Operations Natively implemented in Oracle

Sign-up for preview

Another Demo...

Oracle AI Vector Search

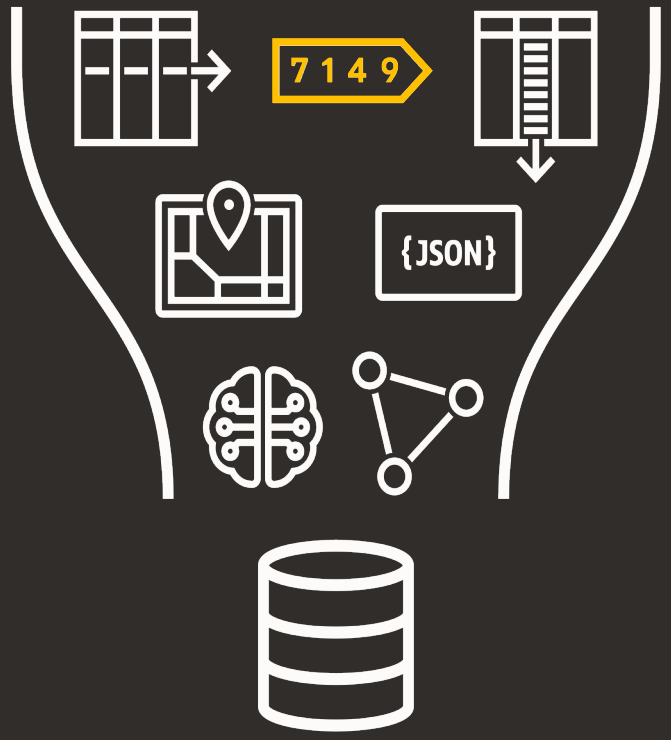
In this notebook we'll be using Oracle Database 23c's new AI Vector Search functionality to search for similar images using natural language. We do this by using the clip-ViT-B-32 model. This is an Image and Text model, which maps text and images to a shared vector space. This allows us to ask questions in natural language and return images in response. We'll store the images in the Oracle Database in the new `Vector` datatype and retrieve them with SQL ordering them based on a new function that determines their similarity based on their distance from one another.

We'll be using Python (3.11), Pytorch (a Machine Learning Library), SentenceTransformer (A tool to generate text and image embeddings based on Pytorch) and a few other utilities (PIL, Matplotlib, Numpy) to help display some of the objects we'll be using.

We'll first need to load some libraries and helper code we use later in the notebook

[ ]:
```
from oracledb import Connection
from IPython.display import HTML, display
from PIL import Image, ImageDraw
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats
from sentence_transformers import SentenceTransformer, util
```

Oracle AI Vector Search enables searches on business data to be combined with semantic searches



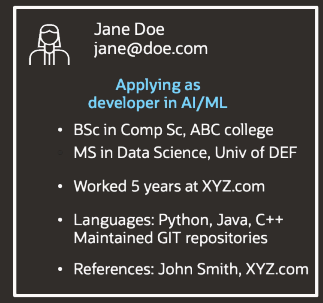
Converged Database

Best solution is to add vector search to your business database

- No need to move and synchronize data, manage multiple products, etc.

Enables combining AI vector search with search on business data about job positions and applicants

Find the top 10 job posts for a Software Engineer in New York ordered by relevance of the resume to the job description



Job Search Example

```
SELECT …
FROM   Job_Postings
WHERE  title = 'Software Engineer' AND location = 'New York'
ORDER BY VECTOR_DISTANCE(job_description_vector, :resume_vector)
FETCH FIRST 10 ROWS ONLY;
```

Enables combining AI vector search with search on business data about job positions and applicants

Combines candidate data, job data, and AI search in 5 lines of SQL!

Single integrated solution, all data fully consistent

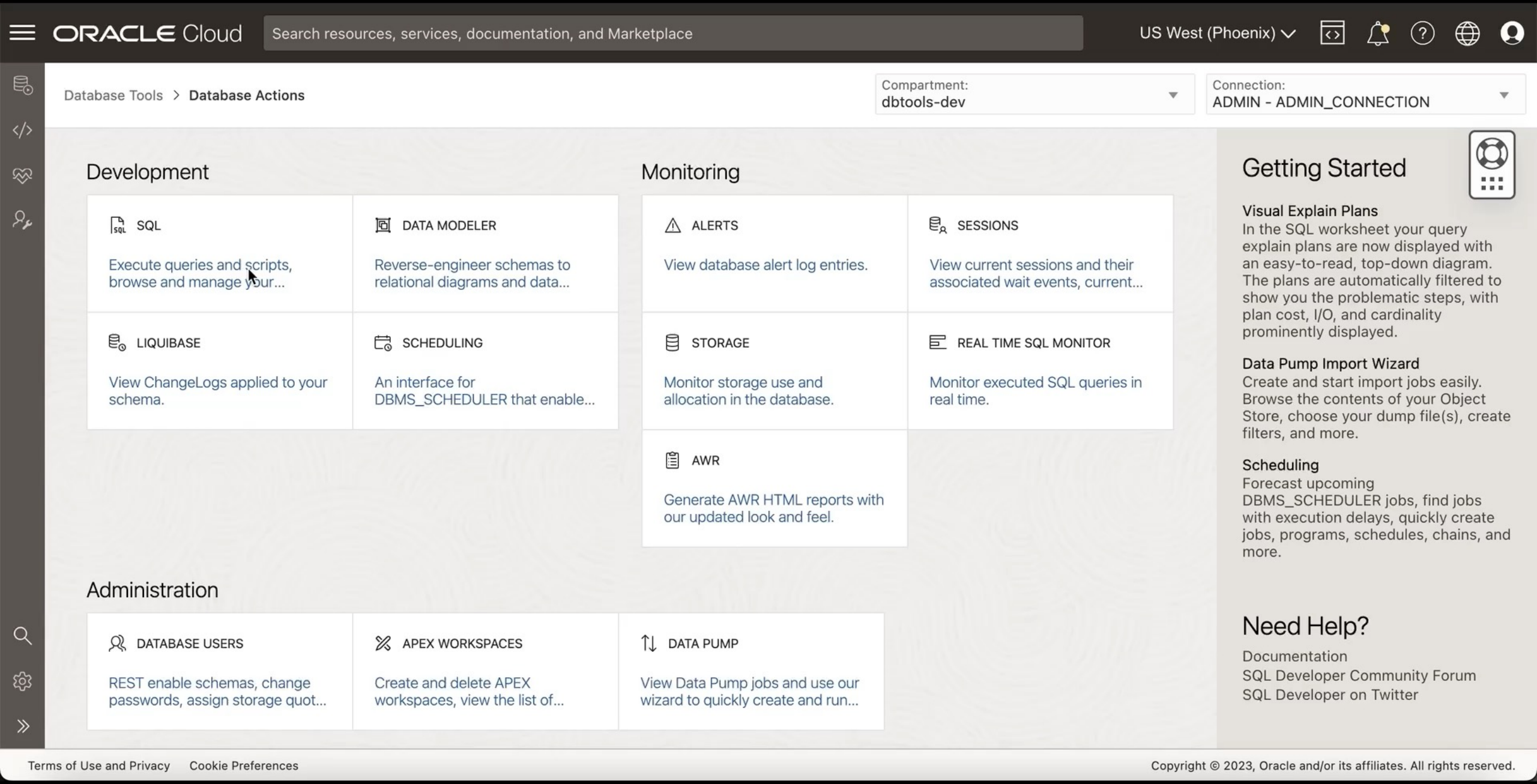
Any developer or DBA can learn to use it in 15 minutes

Find the top 10 job posts for a Software Engineer in New York ordered by relevance of the resume to the job description


```
SELECT …
FROM   Job_Postings
WHERE  title = 'Software Engineer' AND location = 'New York'
ORDER BY VECTOR_DISTANCE(job_description_vector, :resume_vector)
FETCH FIRST 10 ROWS ONLY;
```

But SQL is hard

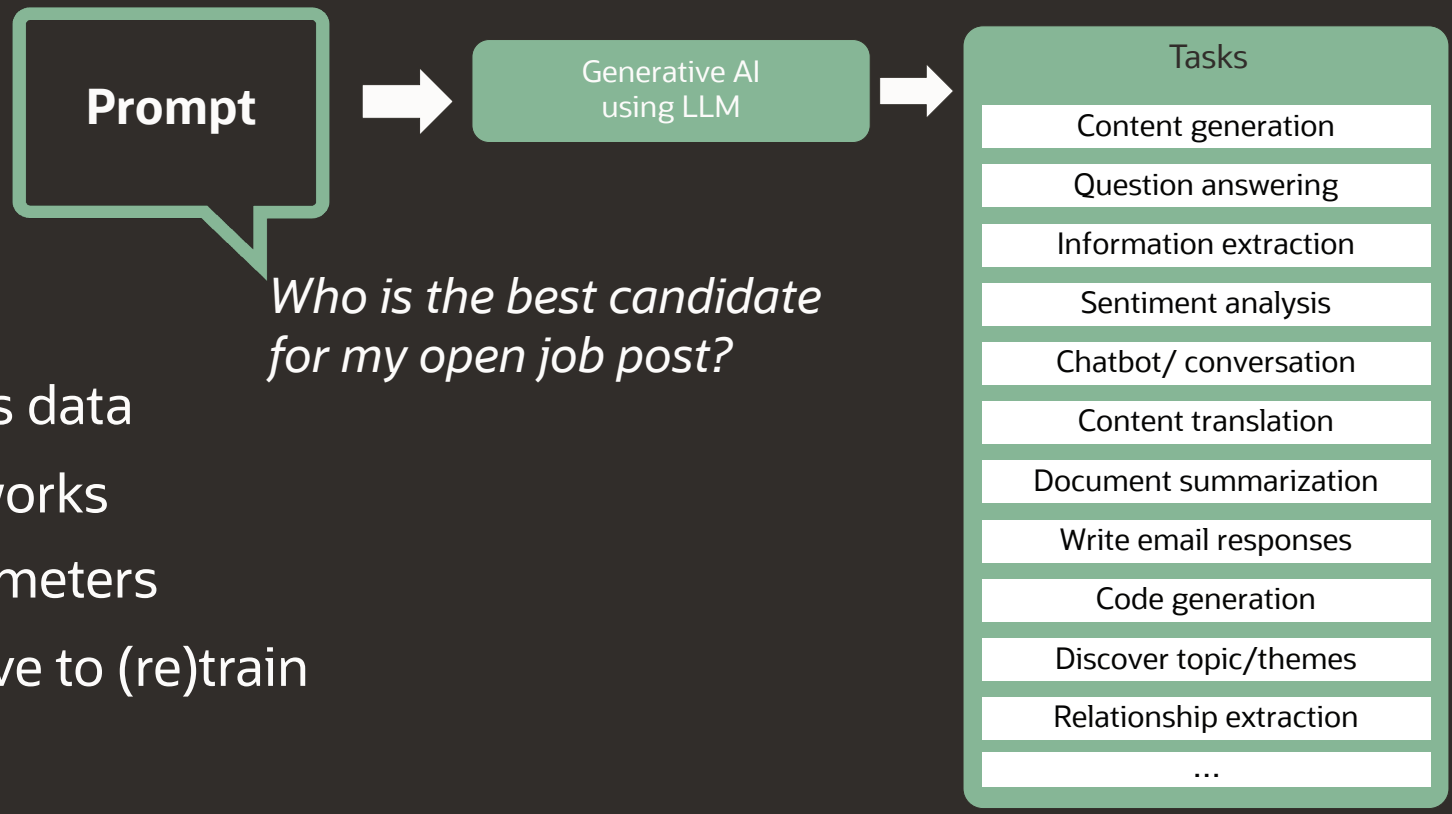
Time for another Demo…


ORACLE Cloud
Search resources, services, documentation, and Marketplace
US West (Phoenix)
Database Tools > Database Actions
Compartment:
dbtools-dev
Connection:
ADMIN - ADMIN_CONNECTION
Development
SQL
Execute queries and scripts, browse and manage your...
DATA MODELER
Reverse-engineer schemas to relational diagrams and data...
LIQUIBASE
View ChangeLogs applied to your schema.
SCHEDULING
An interface for DBMS_SCHEDULER that enable...
Monitoring
ALERTS
View database alert log entries.
SESSIONS
View current sessions and their associated wait events, current...
STORAGE
Monitor storage use and allocation in the database.
REAL TIME SQL MONITOR
Monitor executed SQL queries in real time.
AWR
Generate AWR HTML reports with our updated look and feel.
Administration
DATABASE USERS
REST enable schemas, change passwords, assign storage quot...
APEX WORKSPACES
Create and delete APEX workspaces, view the list of...
DATA PUMP
View Data Pump jobs and use our wizard to quickly create and run...
Getting Started
Visual Explain Plans
In the SQL worksheet your query explain plans are now displayed with an easy-to-read, top-down diagram. The plans are automatically filtered to show you the problematic steps, with plan cost, I/O, and cardinality prominently displayed.
Data Pump Import Wizard
Create and start import jobs easily. Browse the contents of your Object Store, choose your dump file(s), create filters, and more.
Scheduling
Forecast upcoming DBMS_SCHEDULER jobs, find jobs with execution delays, quickly create jobs, programs, schedules, chains, and more.
Need Help?
Documentation
SQL Developer Community Forum
SQL Developer on Twitter
Terms of Use and Privacy
Cookie Preferences
Copyright © 2023, Oracle and/or its affiliates. All rights reserved.

Oracle AI Vector Search also allows interactions with Large Language Models (LLMs) to be augmented with business data

Large Language Model (LLM)

Single general-purpose model capable of interactions on wide range of topics



Who is the best candidate for my open job post?

LLMs are trained on enormous data

Use deep learning neural networks

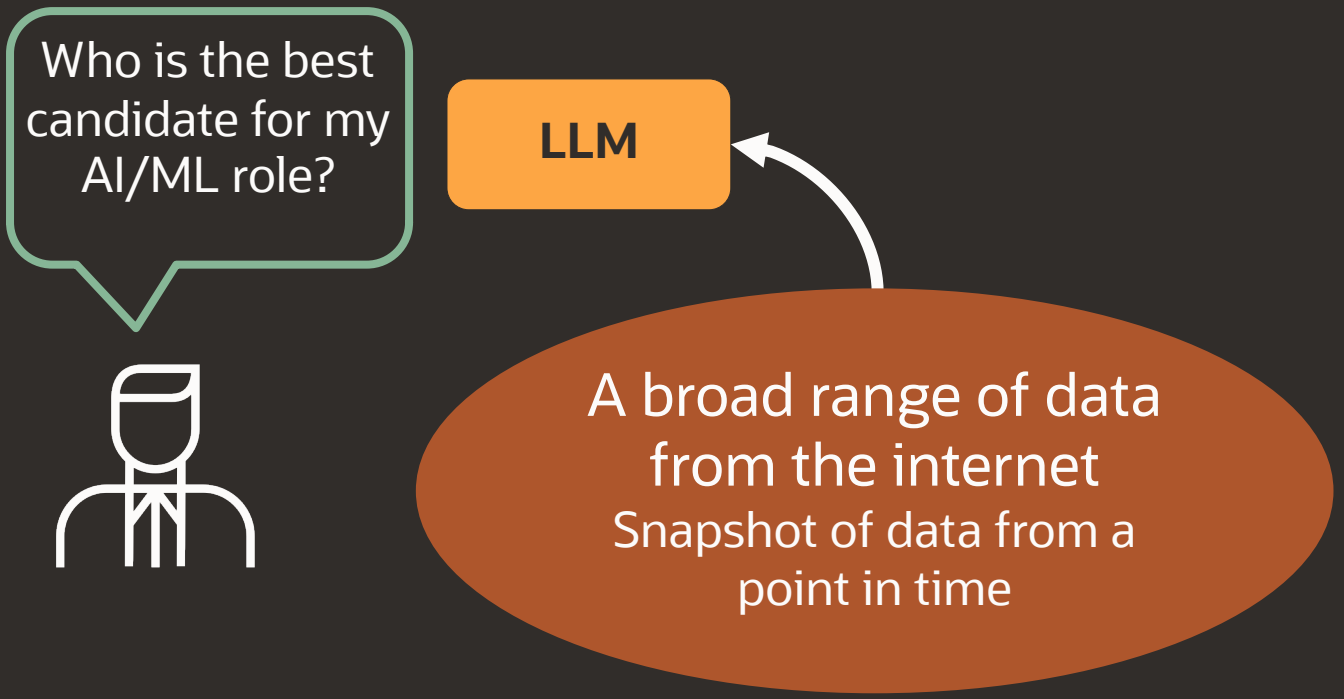
10s to 100s of Billions of Parameters

Computationally very expensive to (re)train

A *Generative Pre-trained Transformer* (GPT) is a specific type of LLM

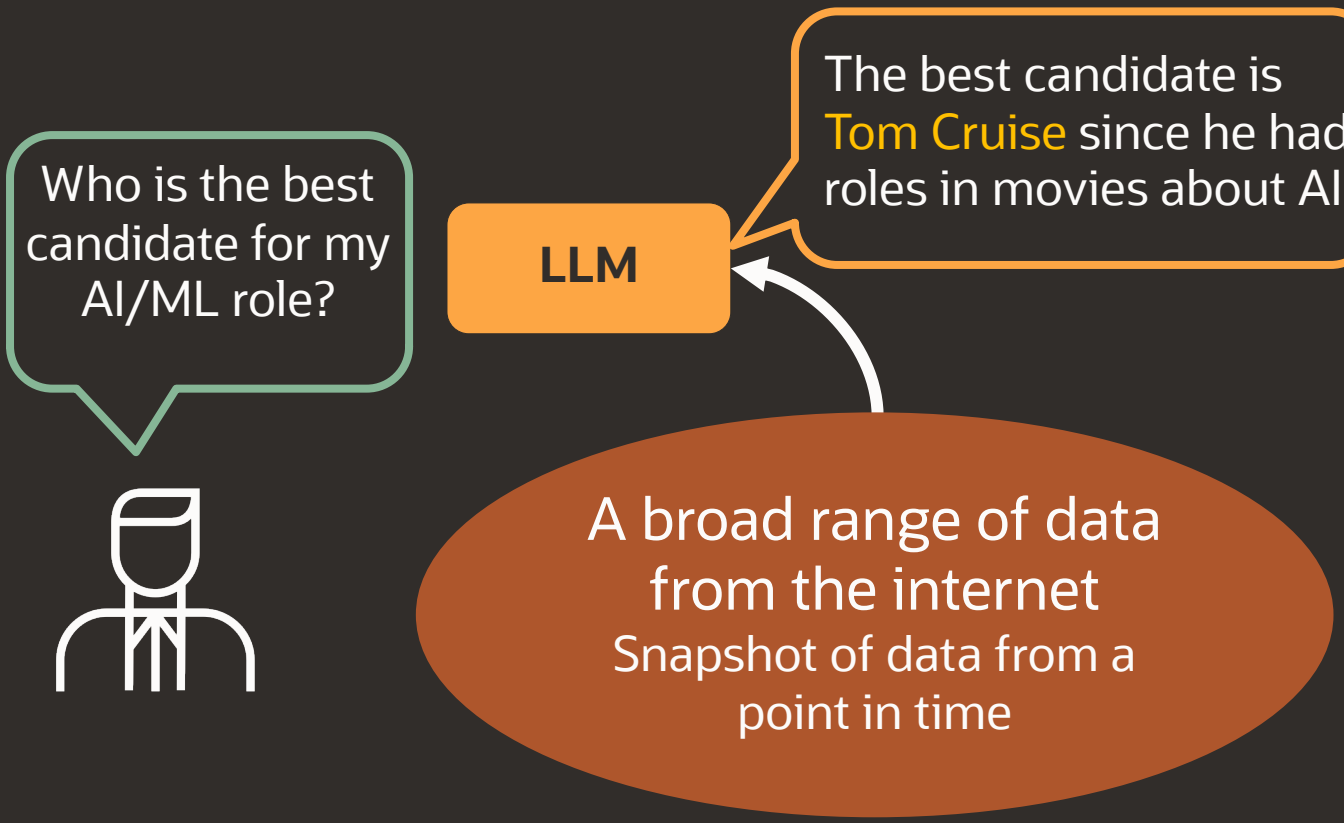
Role of Vector Databases with LLMs

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



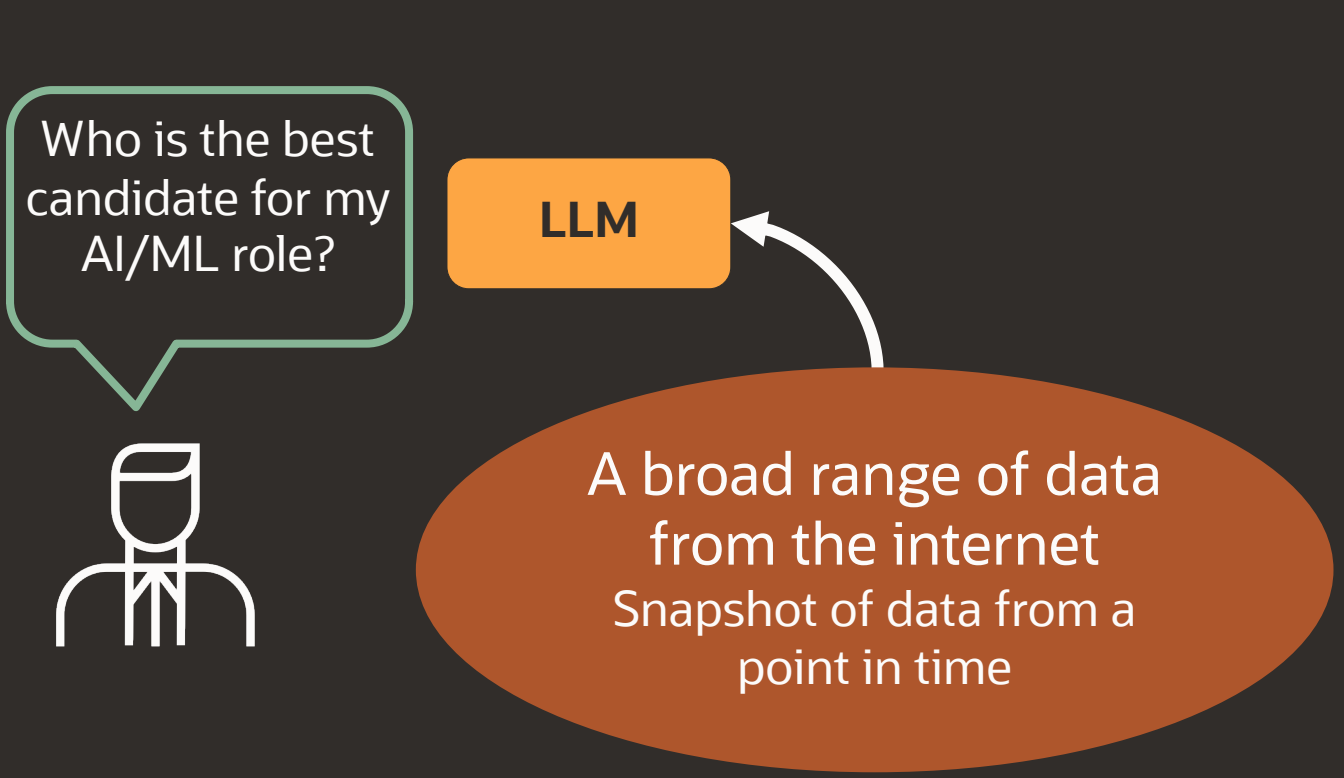
Role of Vector Databases with LLMs

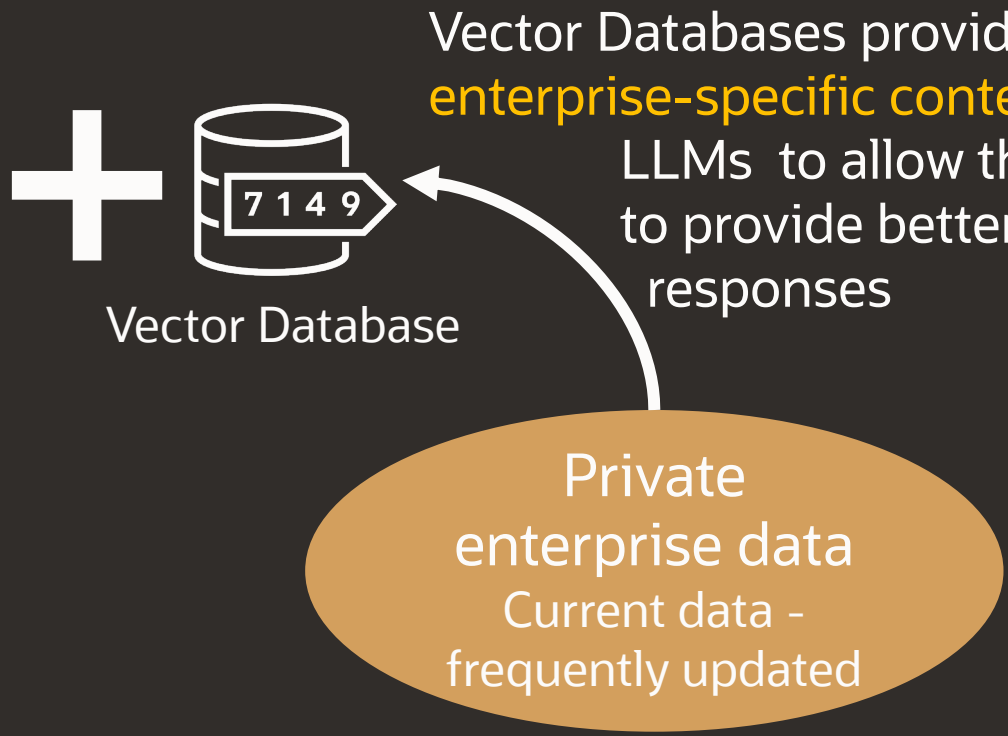
LLMs are frozen on a past snapshot of the internet with no access to private enterprise data
LLMs by themselves therefore often provide poor-quality responses to business questions



Role of Vector Databases with LLMs

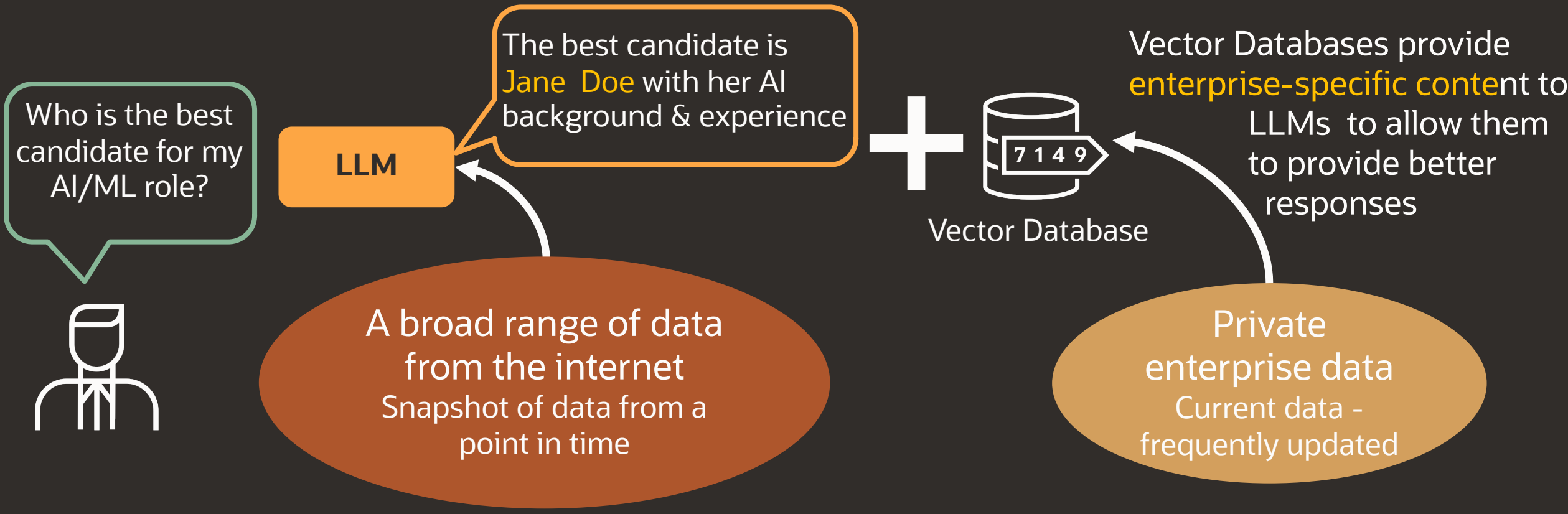
LLMs are frozen on a past snapshot of the internet with no access to private enterprise data
LLMs by themselves therefore often provide poor-quality responses to business questions





Role of Vector Databases with LLMs

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 business questions



Role of Vector Databases with LLMs

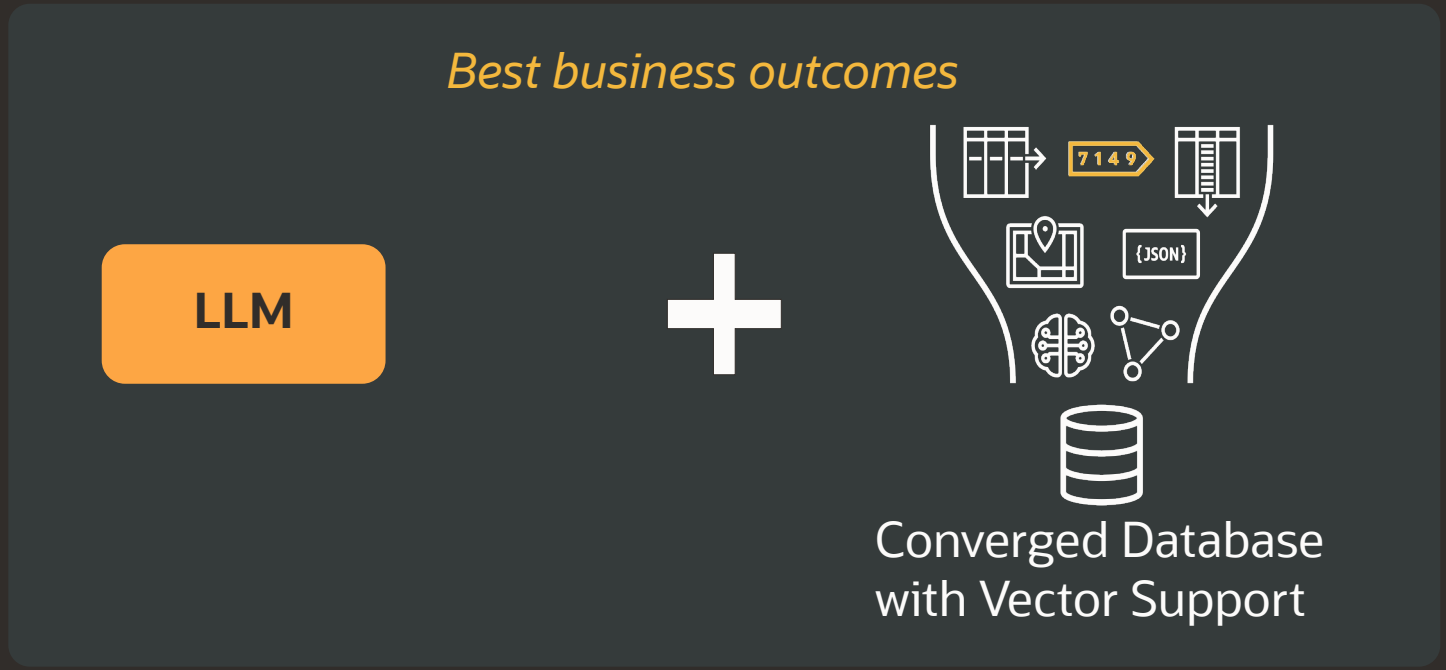
Provide enterprise content to enhance LLM interactions (retrieval augmentation)
Avoid having to train LLMs on sensitive enterprise data (not secure, expensive)
Cache previous LLM prompts/responses to improve performance and reduce costs

Role of Converged Vector Databases with LLMs

Real-time updated knowledge base for consistent responses

No need for data movement, avoids cost, complexity, and security risk of multiple systems

Easily combine business data and vector data for sophisticated interactions with LLMs



Vector Index

Fast, Approximate Searches

Vector Indexing

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

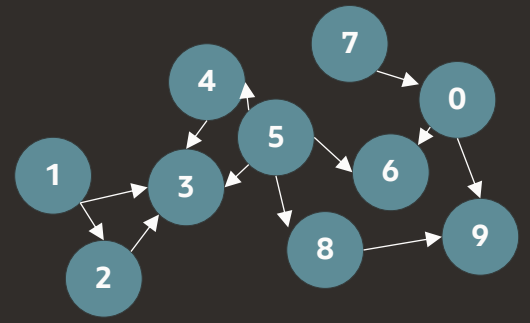
New vector indexes trade-off search accuracy for speed

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

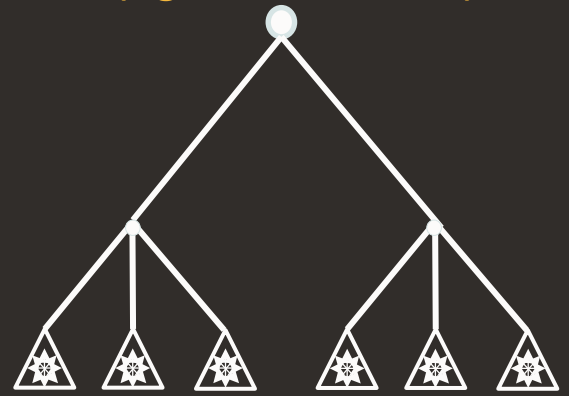
Vector indexes

- 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)

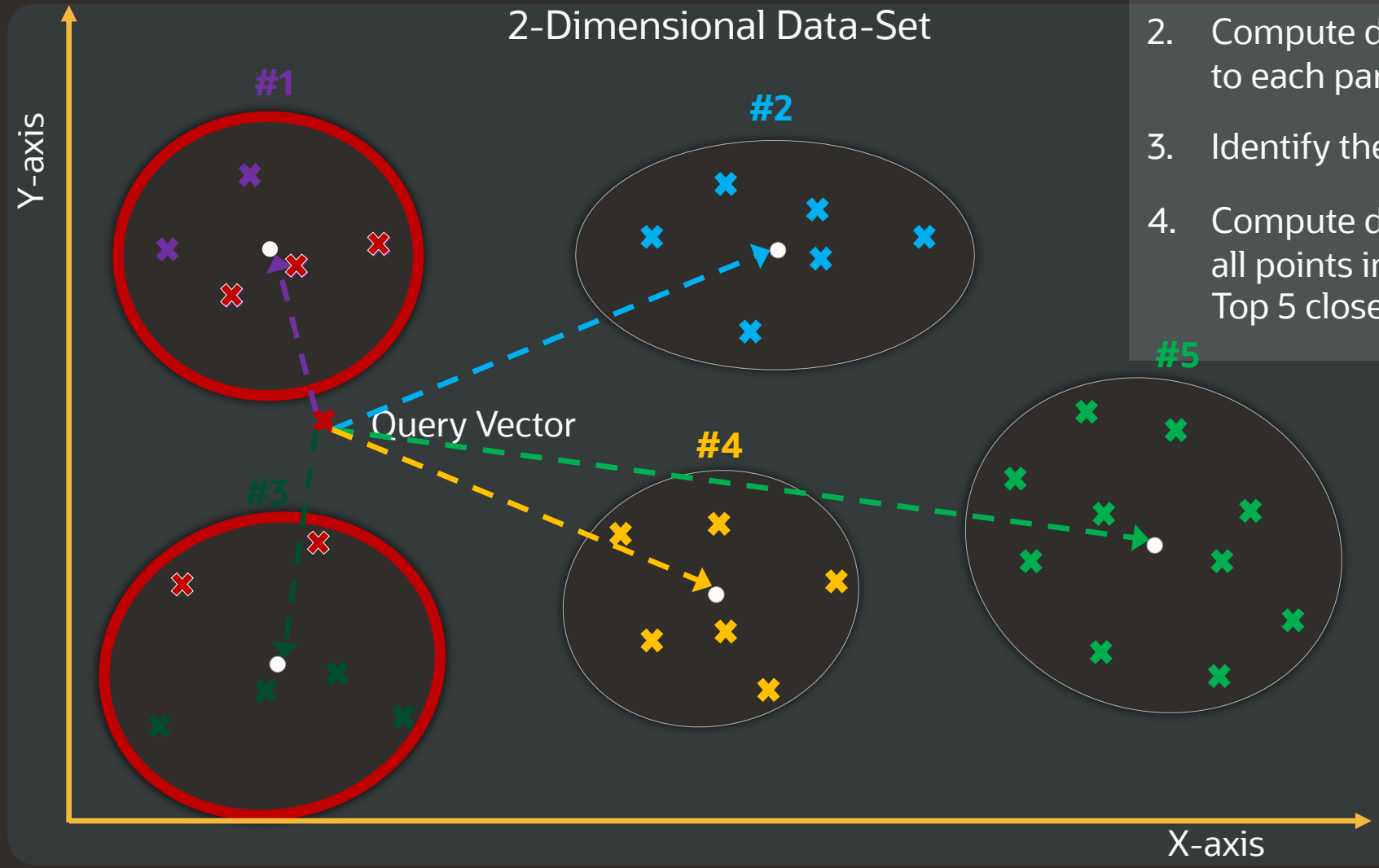


Partition Vector Index
(e.g. **IVF_FLAT** index)



Neighbor Partition Vector Index – Search

1. Group vectors into partitions using OML's K-means clustering algo (K = 5)
2. Compute distance from query vector to each partition's centroids
3. Identify the 2 nearest partitions
4. Compute distance from query vector to all points in Cluster #1 and #3 to find Top 5 closest matches (shown in red)



Graph Vector Index

Multi-layer in-memory graph index

In-memory index designed for speed and accuracy

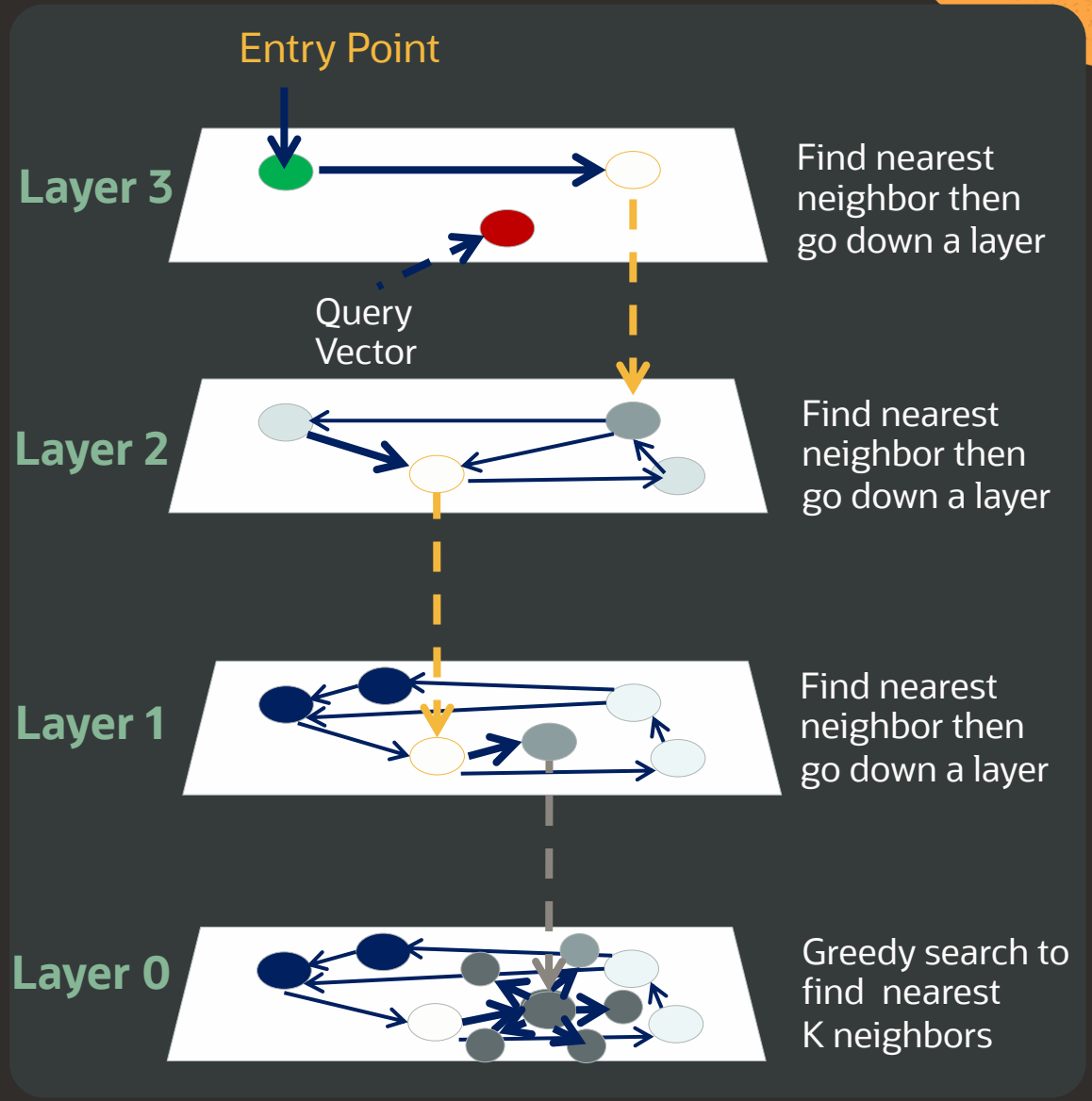
Considered the “B+ tree index for Vectors”

Construction

- Lowest layer of the graph has all the vectors
- Higher layers have a decaying fraction of vectors
- Vectors are connected based on similarity

Search

- Search begins from the top layer. When the nearest vector is found, the search continues in the layer below
- The search completes in the lowest layer when the Top K nearest vectors to the query vector are found



Summary

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

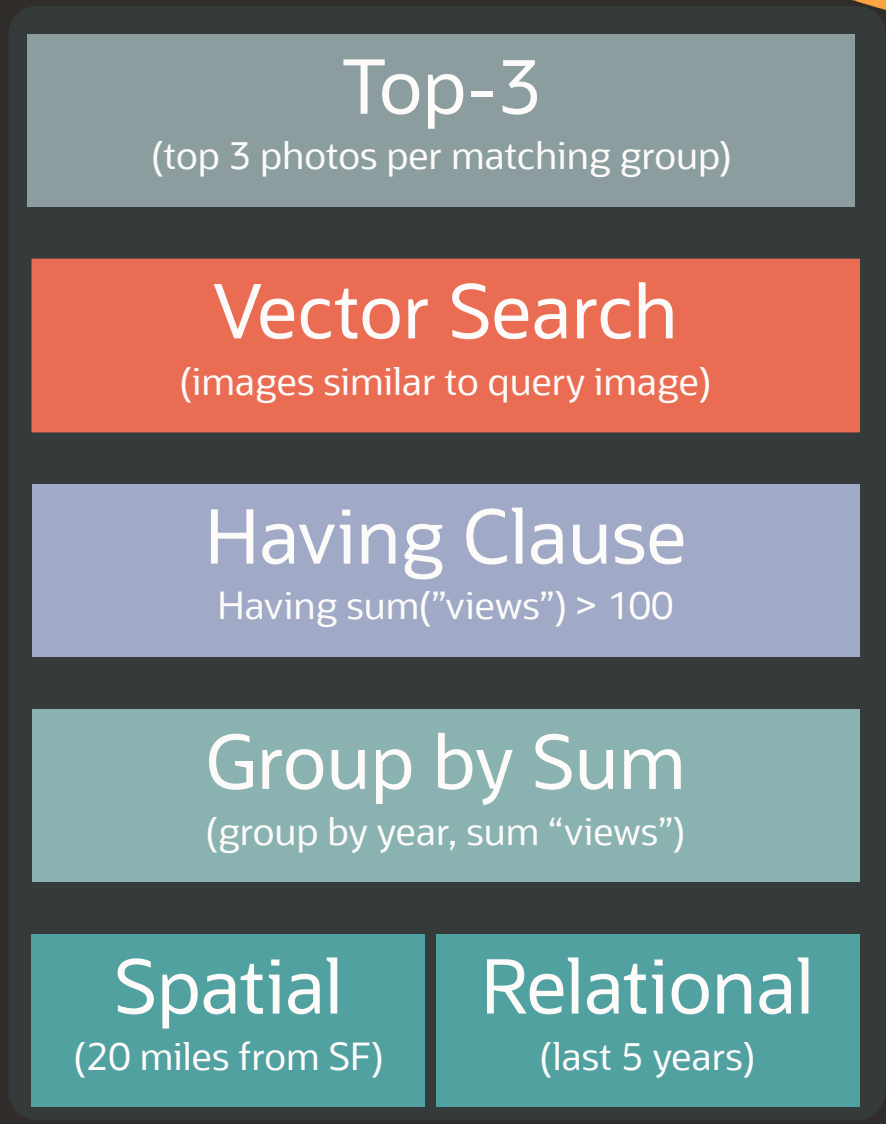
- Complex operators, group-by, aggregation …
- Analytic functions, stored procedures, pattern matching …

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

No purpose-built Vector Database Vector can do this



ORACLE

Thank you