ODSC EAST 2025

NetApp

Adaptive RAG Agents with Knowledge Graphs

Building Reinforcement-Learning-Driven AI Applications

David vonThenen
Senior AI/ML Engineer
@davidvonthenen

David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy...
- There is storage for that!

@davidvonthenen

Agenda

- RAG Agents: Vector DB vs Graph DB
  - Token Prediction vs Data Relationships
  - LLM-Generated vs Fixed Cypher Path
  - Reinforcement Learning
- Hands-On Workshop
  - Options: Laptop or Google Colab
- Q & A

Vector DB vs Graph DB

What to Use When? Weigh the Pros and Cons

Vector, Graph, NoSQL... Oh My!

Vector


Pinecone


Chroma


OpenSearch


Milvus

Graph


NebulaGraph

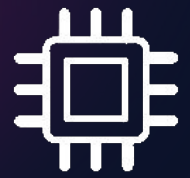

neo4j


JanusGraph

(No)SQL And Friends


instaclustr

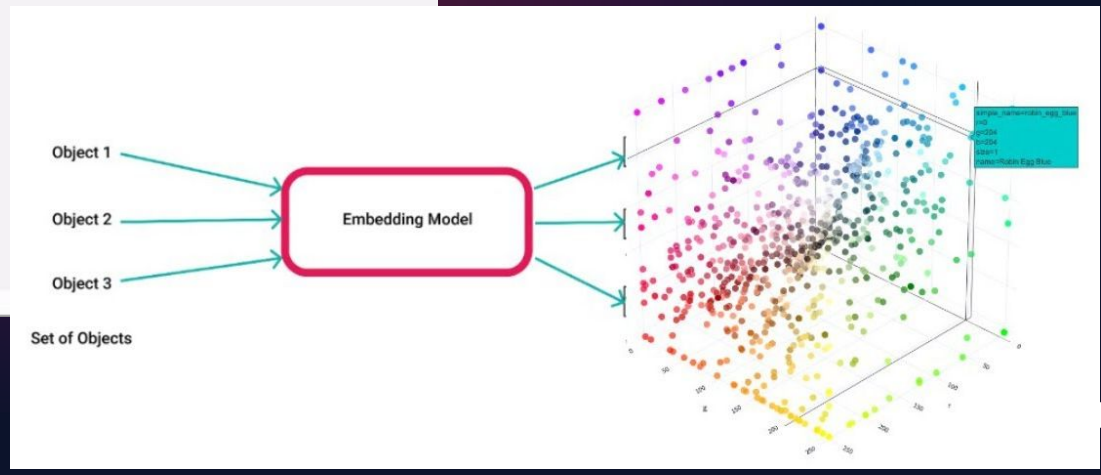

Redis


pgvector

Vector-based RAG: Pros

- Semantic Search Over Unstructured Text
- Associating Conceptually Relevant Info
  - Semantic Similarity (via Abstract Concepts)
- Semantic Search: Highly Scalable, Low Latency
- Diverse Data Types (Img, Audio)

Image Credit:
1. Colin Talks Tech - Introduction to Vector Embeddings



Vector-based RAG: Cons

- No Data Relationships, Exist as Isolated Vectors
  - All Knowledge is Flat
- Difficult to Reason Over Multiple Hops
  - No Holistic View
  - No Chain of Thought
- Miss Complex Entity Connections
  - Top K Limits
  - Top P Limits

Image Credit:
Darek Kleczek - A Gentle Introduction to LLM APIs

Graph-based RAG: Pros

- Excellent Presenting Relationships
  - Great for Structured Knowledge
  - Associations Between Data
- Retrieve Network of Facts vs Snippets
  - Gather Connected Info (All Hops!)
- Reduced Hallucinations!!
- Higher Retrieval Accuracy for RAG
  - Better Response/Answer!

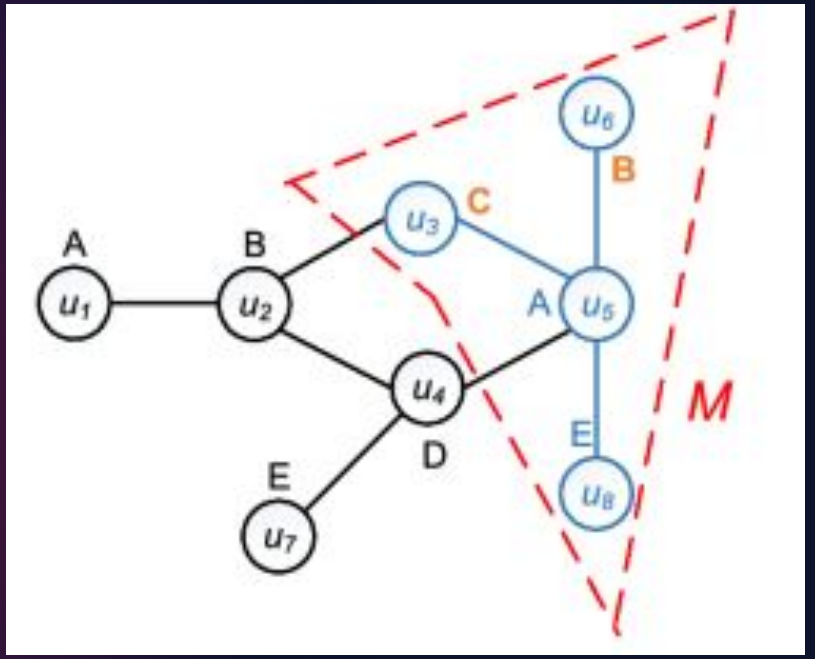


Image Credit:
Xi Wang, Qianzhen Zhang, Deke Guo & Xiang Zhao
A survey of continuous subgraph matching for dynamic graphs

Graph-based RAG: Cons

- Data Modeling & Structure
  - Manage Ontologies/Relationships
- Complexities of Maintenance
- Frequent Data Changes = Challenging
  - Data Consistency with Updates
- Performance Impacts vs Embeddings
  - More Relevant = More Time
  - In-Memory Cache/Optimization

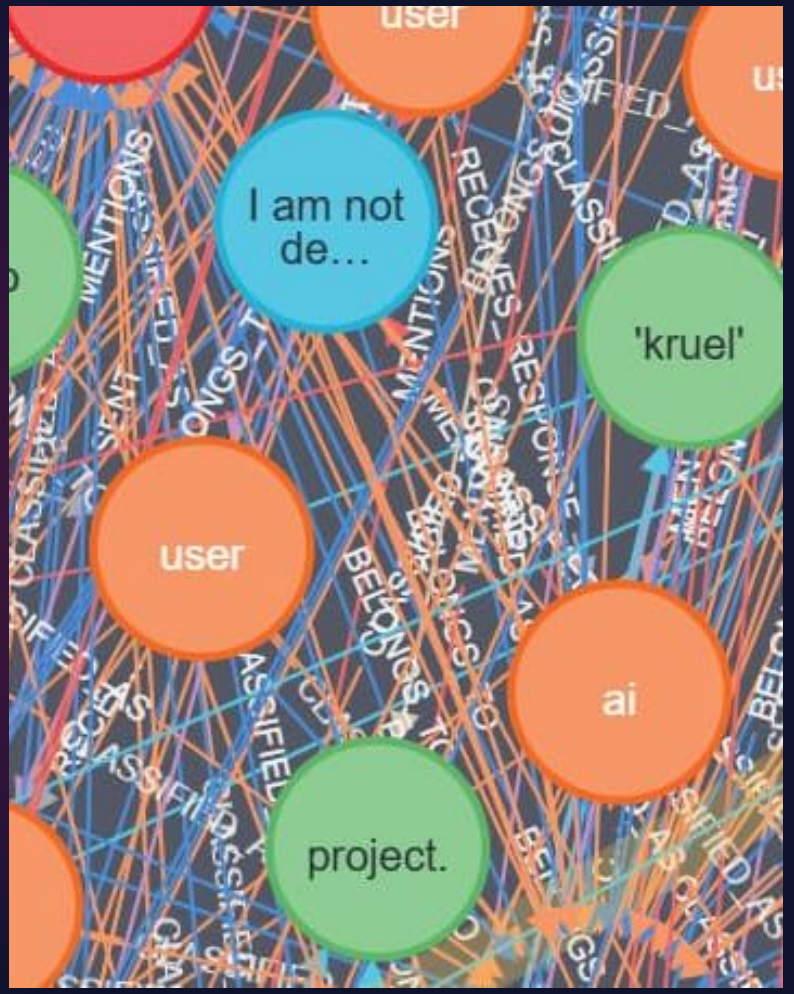
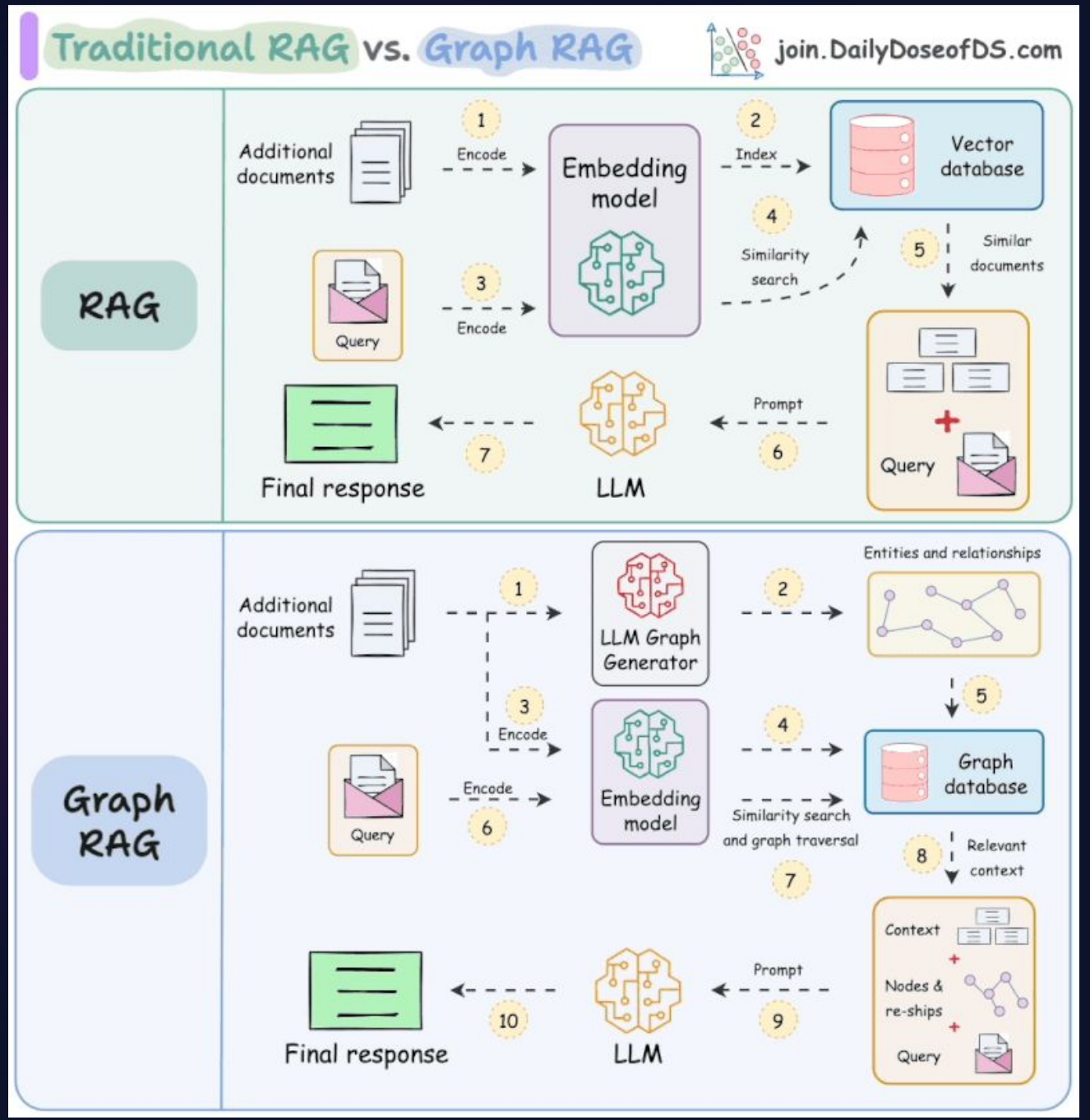


Image Credit:
What Ontology, RAG and Graph data do you use to develop Intelligent Assistants?

Vector vs Graph

Image Credit:
[Avi Chawla]
[LinkedIn Post - Vector vs Graph]



Tokens vs Relationships

Token Prediction vs Data Relationships

Prediction and Top-K

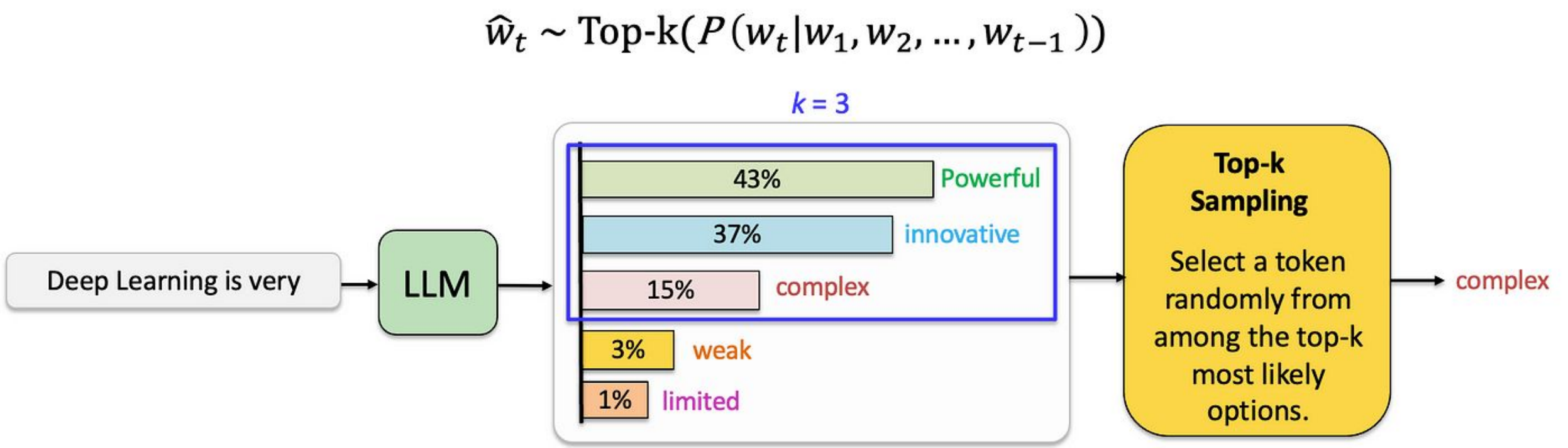


Image Credit:
LM Po - Understanding LLM Decoding Strategies

The Problem Is...

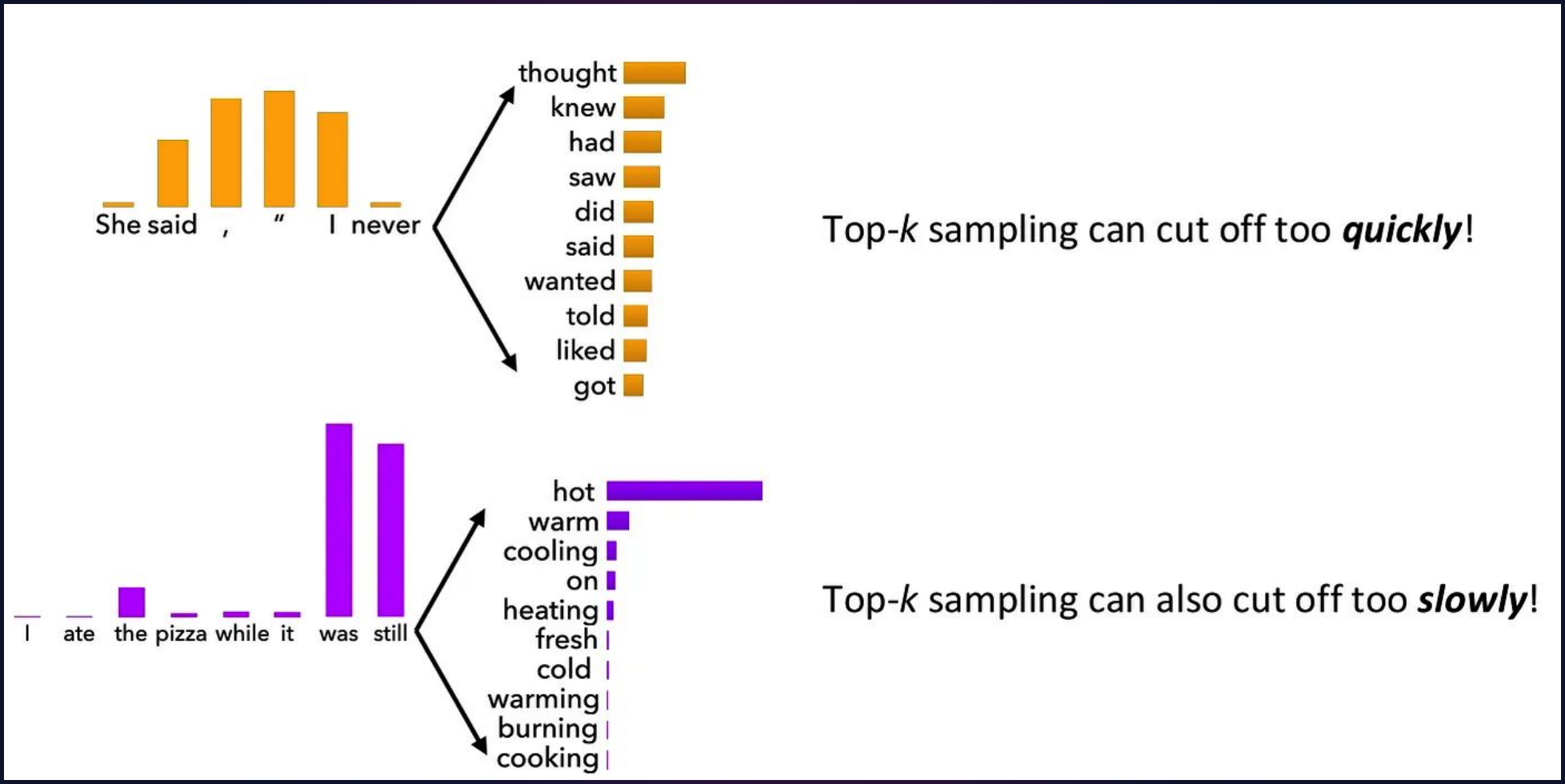


Image Credit:
Pasha Shaik - Natural Language Generation from Scratch in Large Language Models with PyTorch

Better: Graph Retrieval

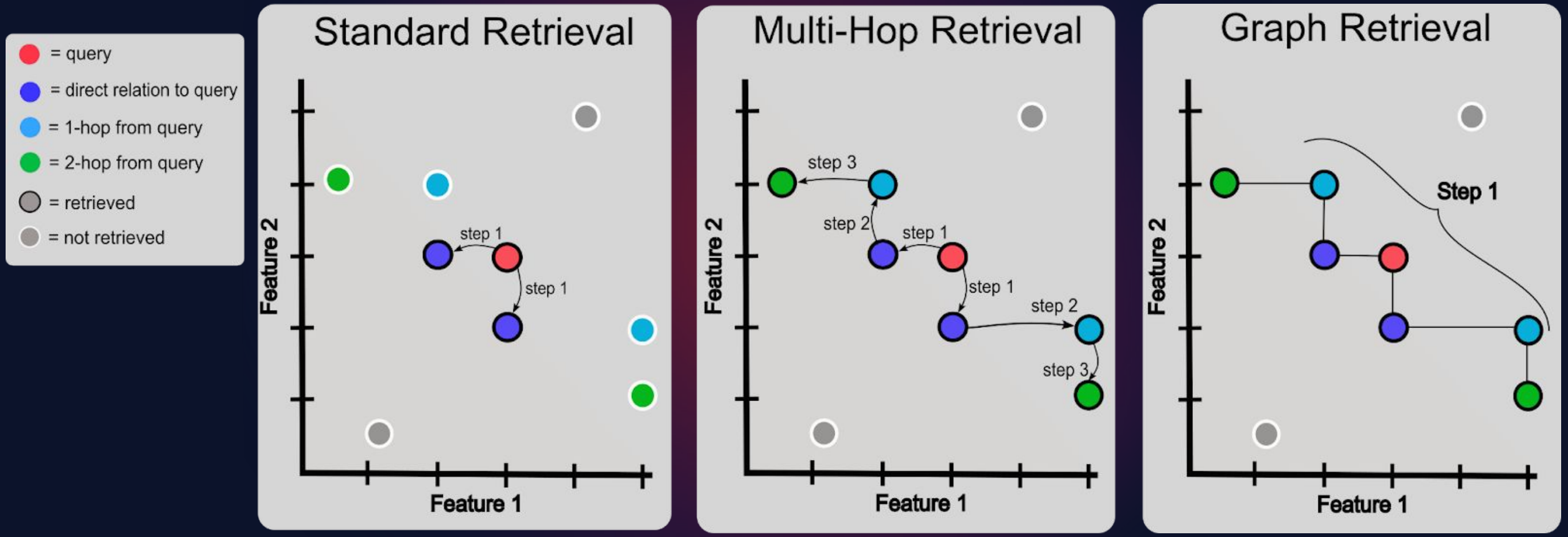


Image Credit:
Nick Alonso, Beren Millidge
Understanding Graph-based RAG and Multi-Hop Question Answering

LLM vs Fixed Cypher Path

How to Handle Information Lookup

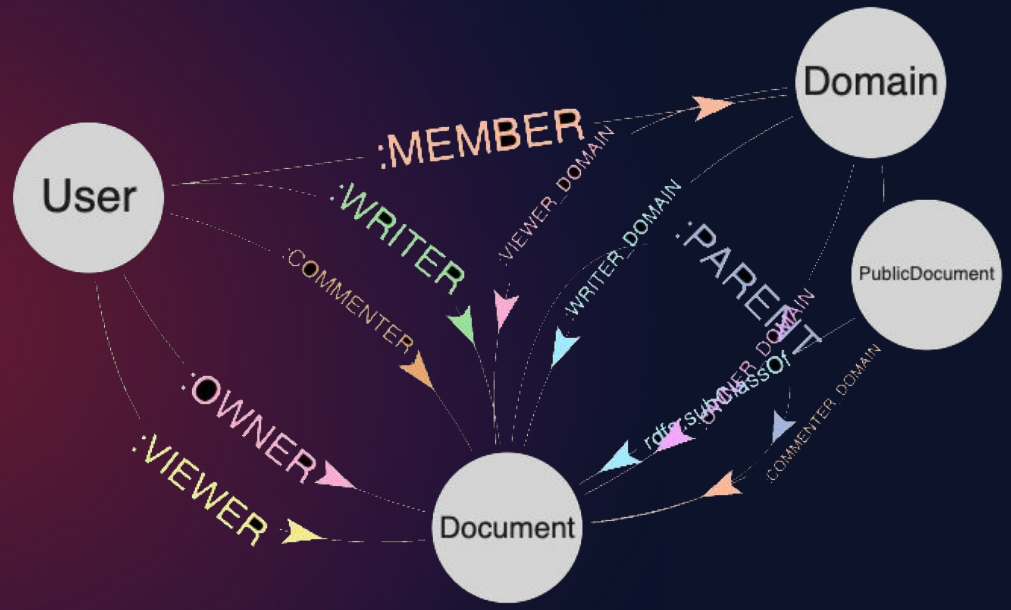
LLMs Are All The Rage, Man!

- LLM Generates the Cypher Paths (aka "Lookup")
  - User Question –> LLM digest –> Cypher Path
  - LangChain – GraphCypherQAChain
- Pros:
  - VERY Easy to Get Started
  - Demo Ready, Great for POC
- Cons:
  - Hides Cypher Gen. Details
  - LLM = Decent-ish GPU?
  - Brittle AF

LangChain

Crafting Your Own Cyphers

- "Manual" Process Based on Graph Schema
  - Don't Necessarily Need an LLM
    - Don't Need a GPU
- Pros:
  - Predictable Data Lookup
  - Predictable Performance
  - Easier Maintenance
- Cons:
  - Human Builds Data Associations
  - Must Be Conscious of Graph Schema



Reinforcement Learning

Short-Term / Long-Term Memory

Reinforcement Learning Differences

Mechanism

- Vector: Facts are Fixed-length Embeddings
- Graph: Facts/Entities Labeled Nodes/Edges

Expiration & Remembering

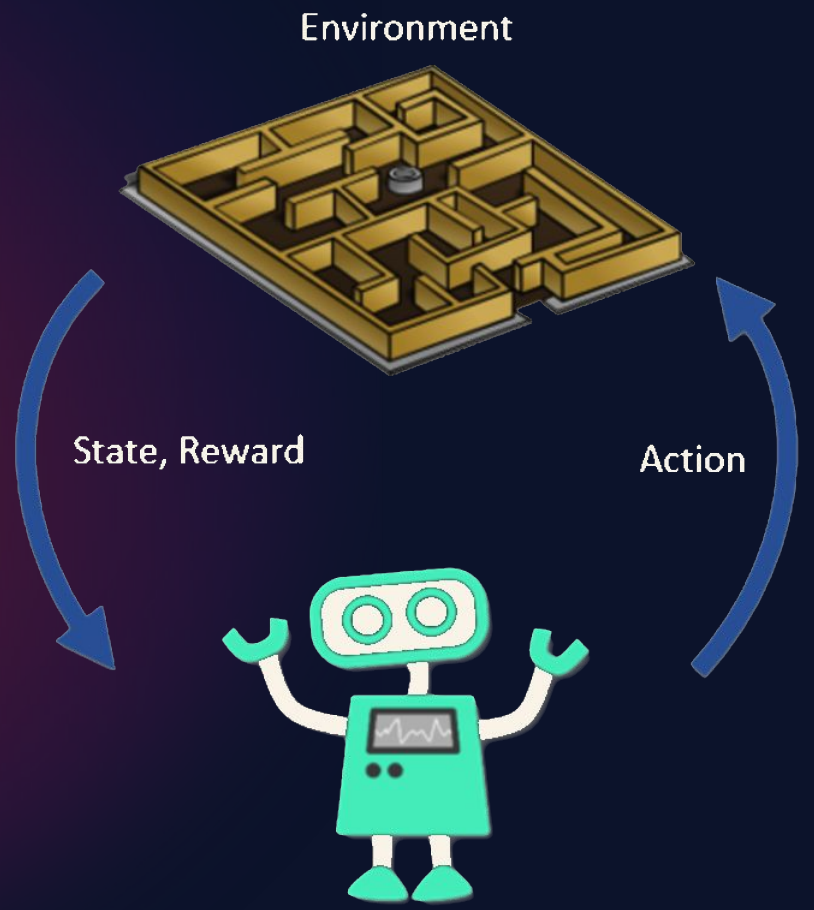
- Vector: Track IDs MUST Rebuild Index to "Forget"
- Graph: Expiration as Simple as Removing a Timestamp

Control & Granularity

- Vector: Blunt operations (delete entire vectors).
- Graph: Fine-grained TTL per Relationship; Promote or Expire Individual Facts on Demand.

Ability to Audit

- Vector: No Native History (Risk of Silent Data Loss)
- Graph: Full Provenance (Only Node/Edge Metadata Changes)



Graph Example: Short vs Long Term

Data Model

- :Document nodes for each fact
- :Entity nodes via spaCy NER
- (:Document)-[m:MENTIONS]->(:Entity) Relationships Has An Expiration Property

Short-Term Memory

- On 👍 : add data, then set m.expiration = now + 24h
- Query Edges expiration > now ⇒ default ephemerality

Promoting to Long-Term

- Remove expiration Field on Edge ⇒ Long-Term
- Document Stays Forever

Force-Expire

- Backdate m.expiration to past (e.g. now - 2 days) ⇒ Hide From Future Queries

Resources

Resources

All Materials/Code: github.com/davidvonthenen/2025-odsc-east-workshop

AIPod Mini (Cost Effective AI System): https://ntap.com/3F91LbO
Let's Chat on Discord: discord.gg/NetApp

Graph Database Options:

- Neo4j - github.com/neo4j/neo4j
- NebulaGraph - github.com/vesoft-inc/nebula

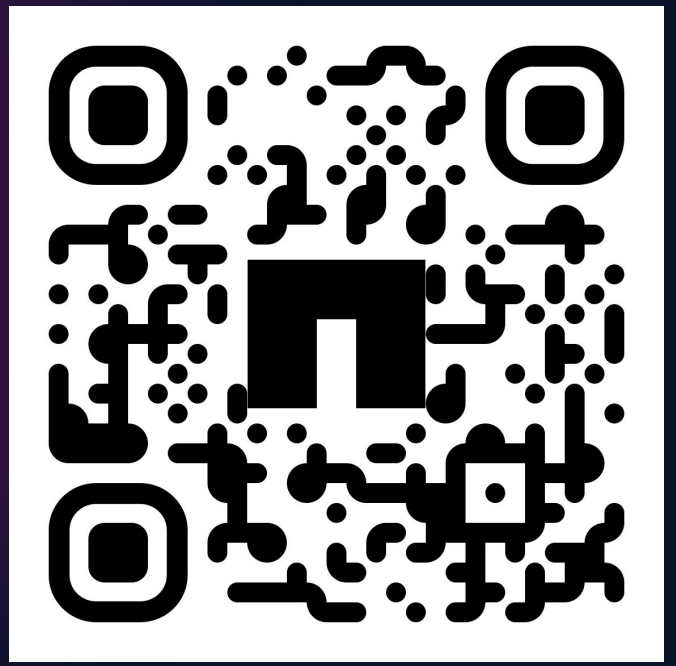
Document Store Options:

- OpenSearch - github.com/opensearch-project/OpenSearch

Vector Database:

- Milvus - github.com/milvus-io/milvus

Dataset (British Broadcasting Corporation) in Demo: bit.ly/4hBKNjp



Hands-On Workshop

Demo: Building a Graph RAG

https://youtu.be/WLEGg5zVwCQ

Demo: Explainable AI – Visualization

https://youtu.be/DDajZ5nS7aU

Demo: AI + Traditional Apps

https://youtu.be/FcEDJl1hDk4

Demo: Reinforcement Learning

https://youtu.be/B-M0swBtmgk

Demo: Agent2Agent Protocol

https://youtu.be/_XB2gmy9Gq0

To The GitHub Repo...

David vonThenen
Senior AI/ML Engineer
@davidvonthenen

Hands-On Instructions:
github.com/davidvonthenen/2025-odsc-east-workshop

Two Options:

- Easy But SUPER Slow = Google Colab
- Your Configuration Skills = Your Laptop

Worst Case Scenario:

- I AM HERE TO HELP!
- Bad Internet -> 20 USB Drives With Most Software

ODSC EAST 2025

NetApp

Thank You!

David vonThenen
Senior AI/ML Engineer
@davidvonthenen