Leveraging Knowledge Graphs for RAG

A Smarter Approach to Contextual AI Applications

David vonThenen

@davidvonthenen

Agenda

- RAG Agents: Vector DB vs Graph DB
- Deep Dive by Example
  - Token Prediction vs Data Relationships
  - Explainable AI
  - Leverage In Non-AI Apps
- Q&A

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

ALL THINGS OPEN.AI

DigitalOcean

Vector DB vs Graph DB

What to Use When? Weigh the Pros and Cons

Vector, Graph, NoSQL… Oh My!

Vector

Pinecone

Chroma

qdrant

Milvus

Graph

NebulaGraph

neo4j

JanusGraph

(No)SQL And Friends

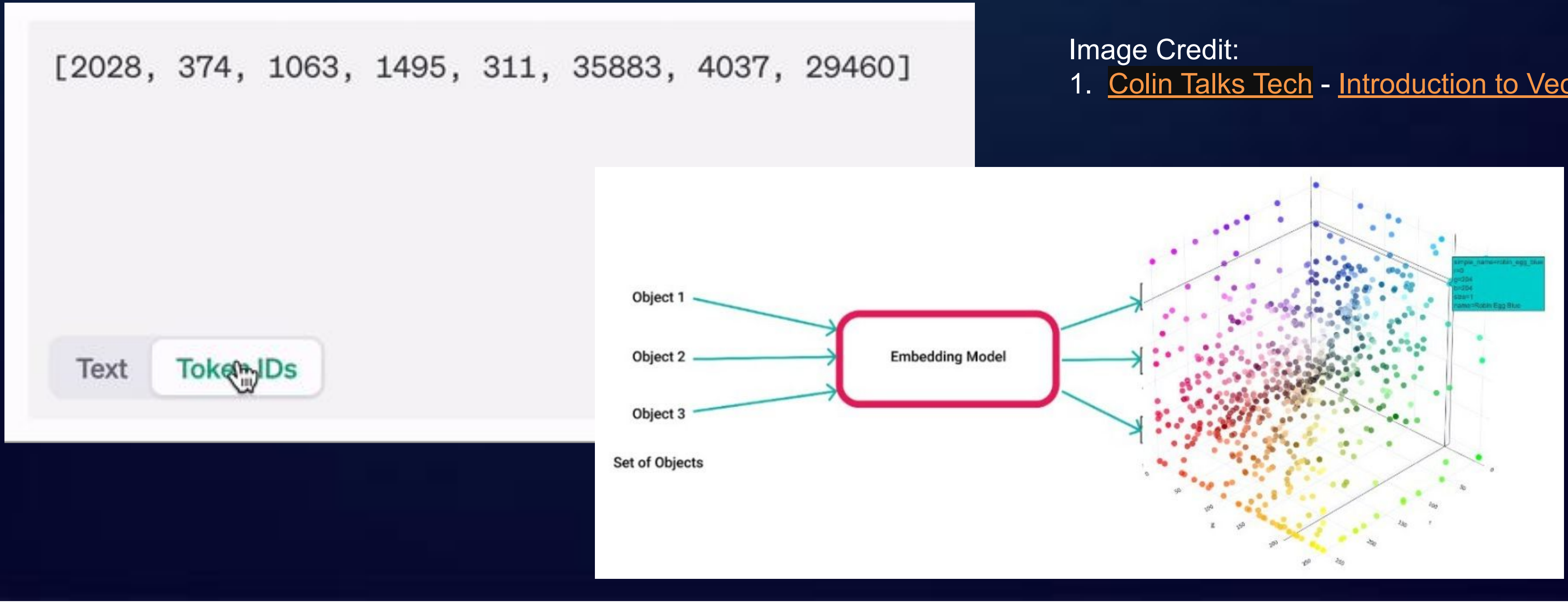
instaclustr

pgvector

Redis

Vector-based RAG: Pros

- Semantic Search Over Unstructured Text
  - Embedding/Semantic Similarity
- Finding Conceptually Relevant Info
- Highly Scalable, Low Latency
- Diverse Data Types (Img, Audio)



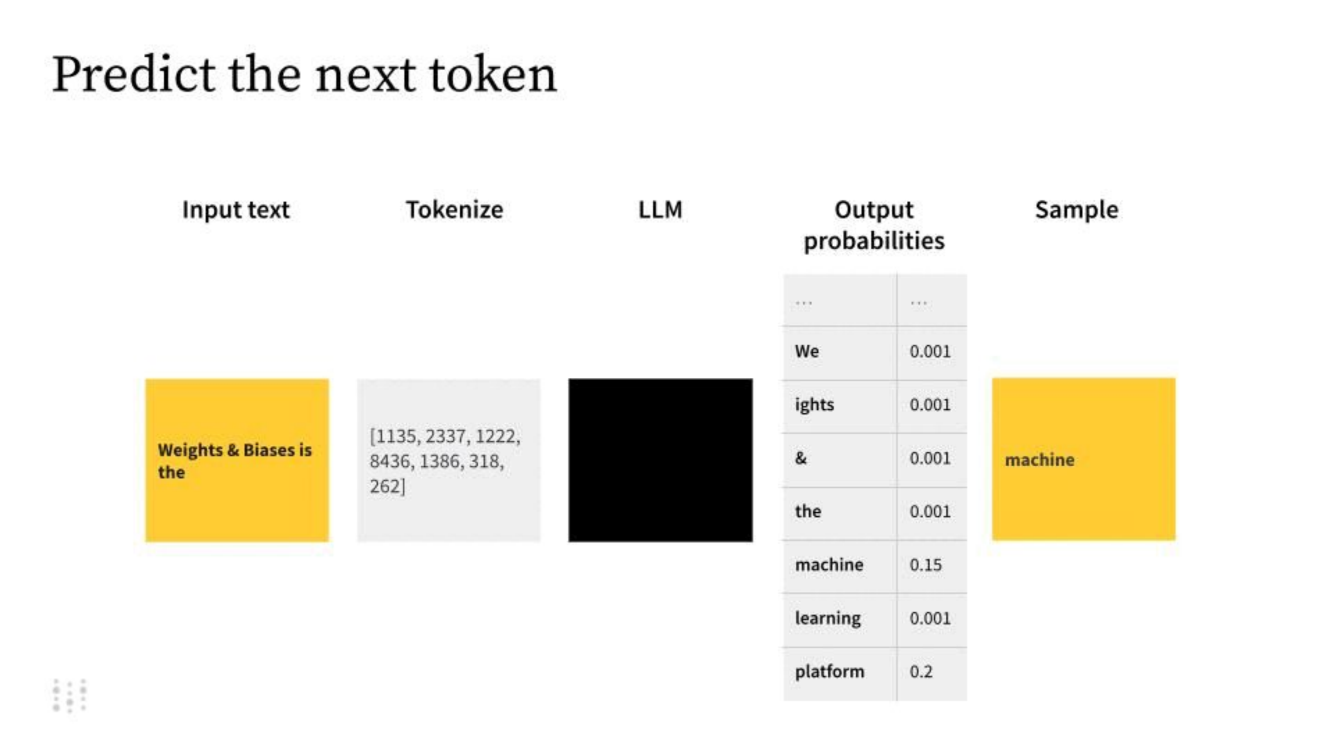
```
[2028, 374, 1063, 1495, 311, 35883, 4037, 29460]
```

Image Credit:

1. Colin Talks Tech - Introduction to Vector Embeddings

Vector-based RAG: Cons

- No Data Relationships
  - All Knowledge is Flat
- Difficult to Reason Over Multiple Hops
  - How Many Rs in Strawberry?
- Miss Complex Entity Connections
  - Top K Limits
  - Top P Limits



Predict the next token

Input text

Weights & Biases is the

Tokenize

[1135, 2337, 1222, 8436, 1386, 318, 262]

LLM

Output probabilities

...	...
We	0.001
ights	0.001
&	0.001
the	0.001
machine	0.15
learning	0.001
platform	0.2

Sample

machine

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!)
- Reduce Hallucinations
- Higher Retrieval Accuracy for RAG
    - Better Response/Answer!

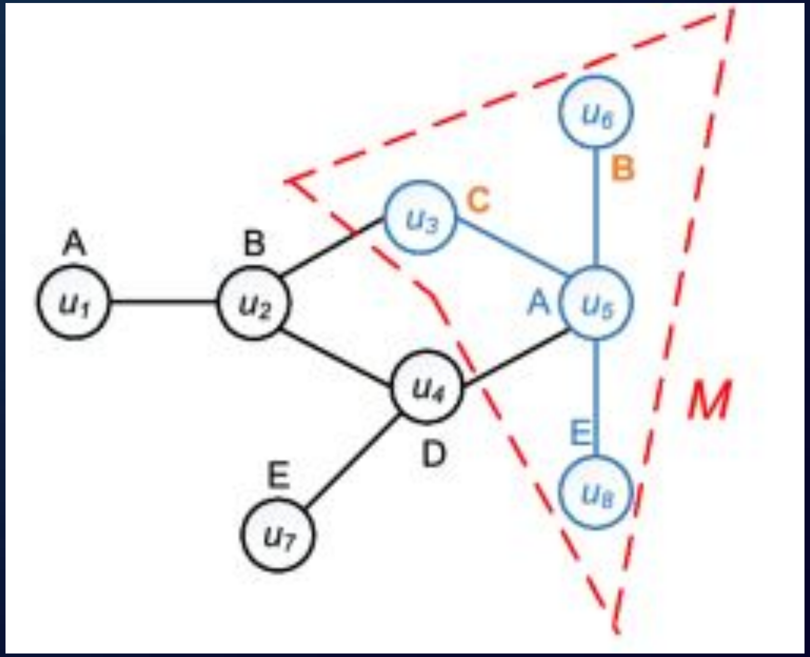


Image Credit:
Xi Wang, Qianzhen Zhang, Deke Guo & Xiang Zhao
A survey of continuous subgraph matching for dynamic graphs

Graph-based RAG: Cons

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

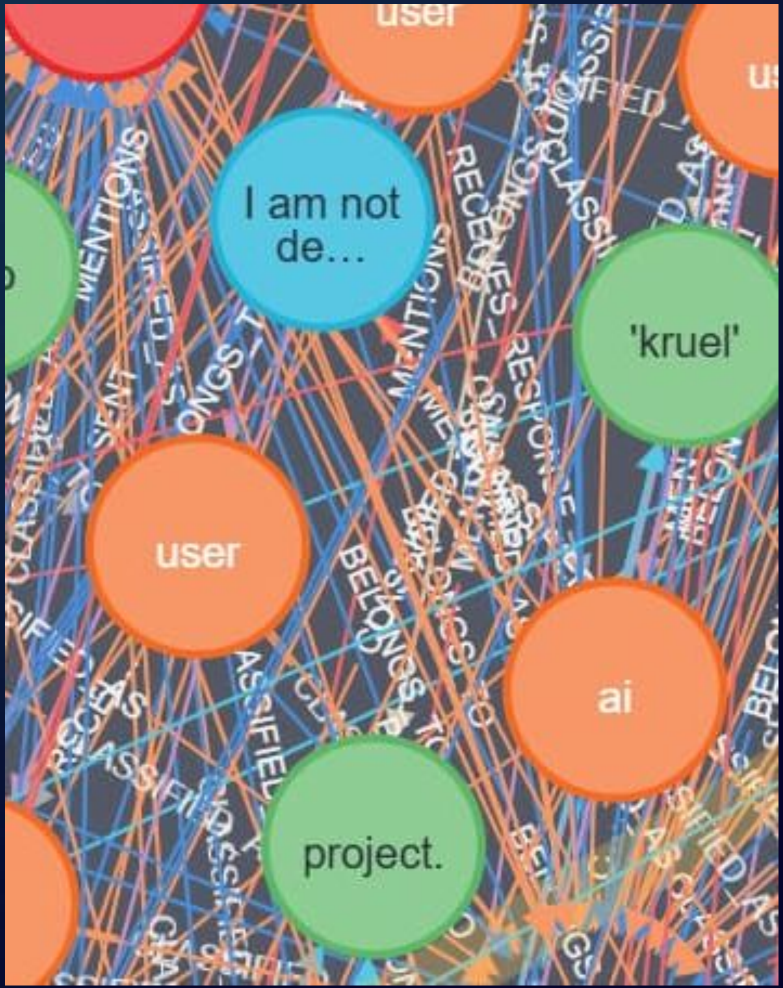
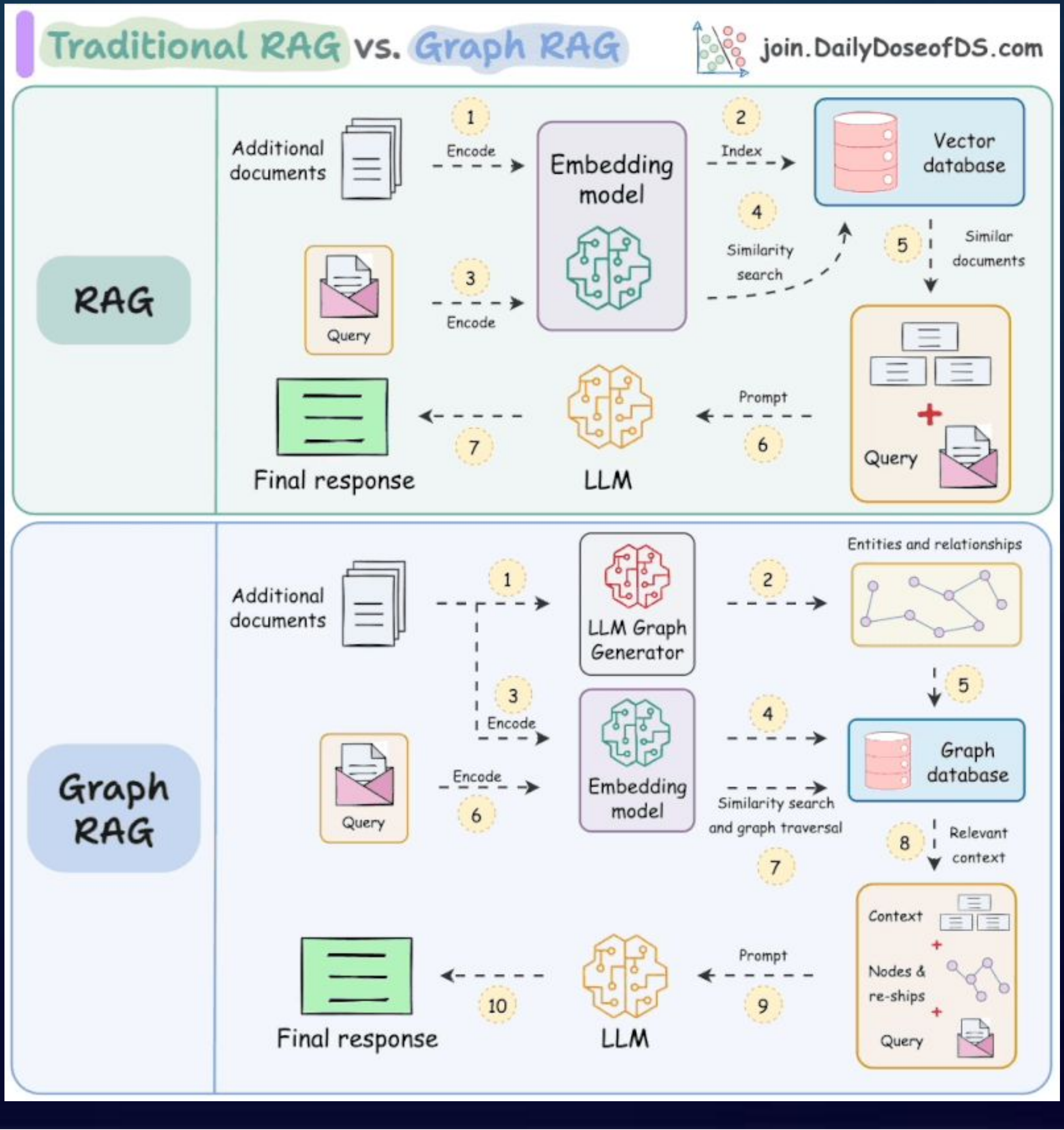


Image Credit:
What Ontology, RAG and Graph data do you use to develop Intelligent Assistants?

Vector vs Graph

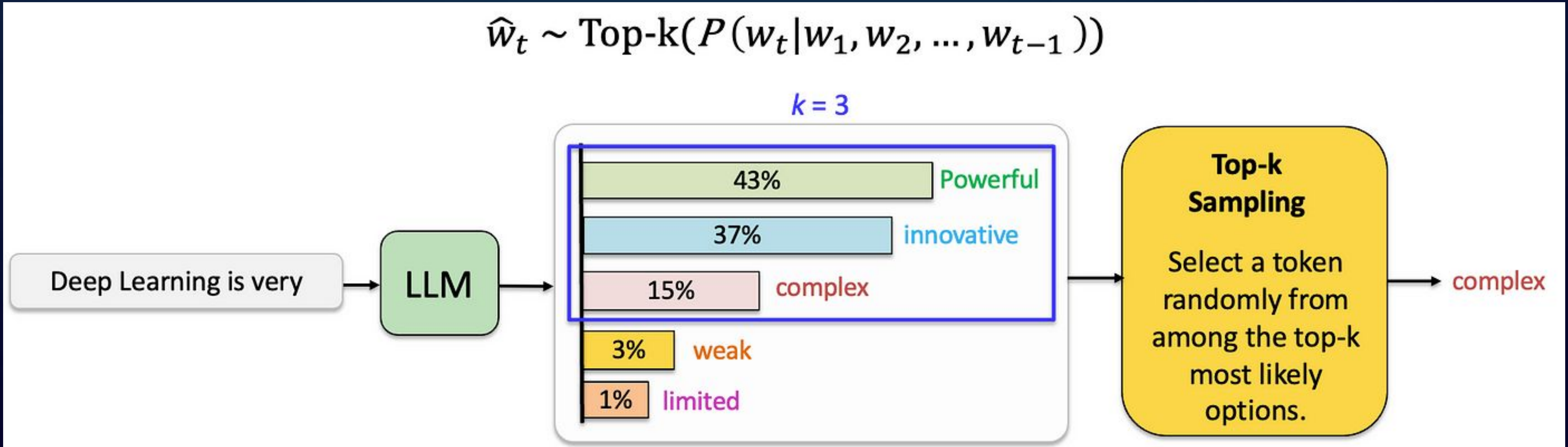
Image Credit:
Avi Chawla
LinkedIn Post - Vector vs Graph



Tokens vs Relationships

Token Prediction vs Data Relationships

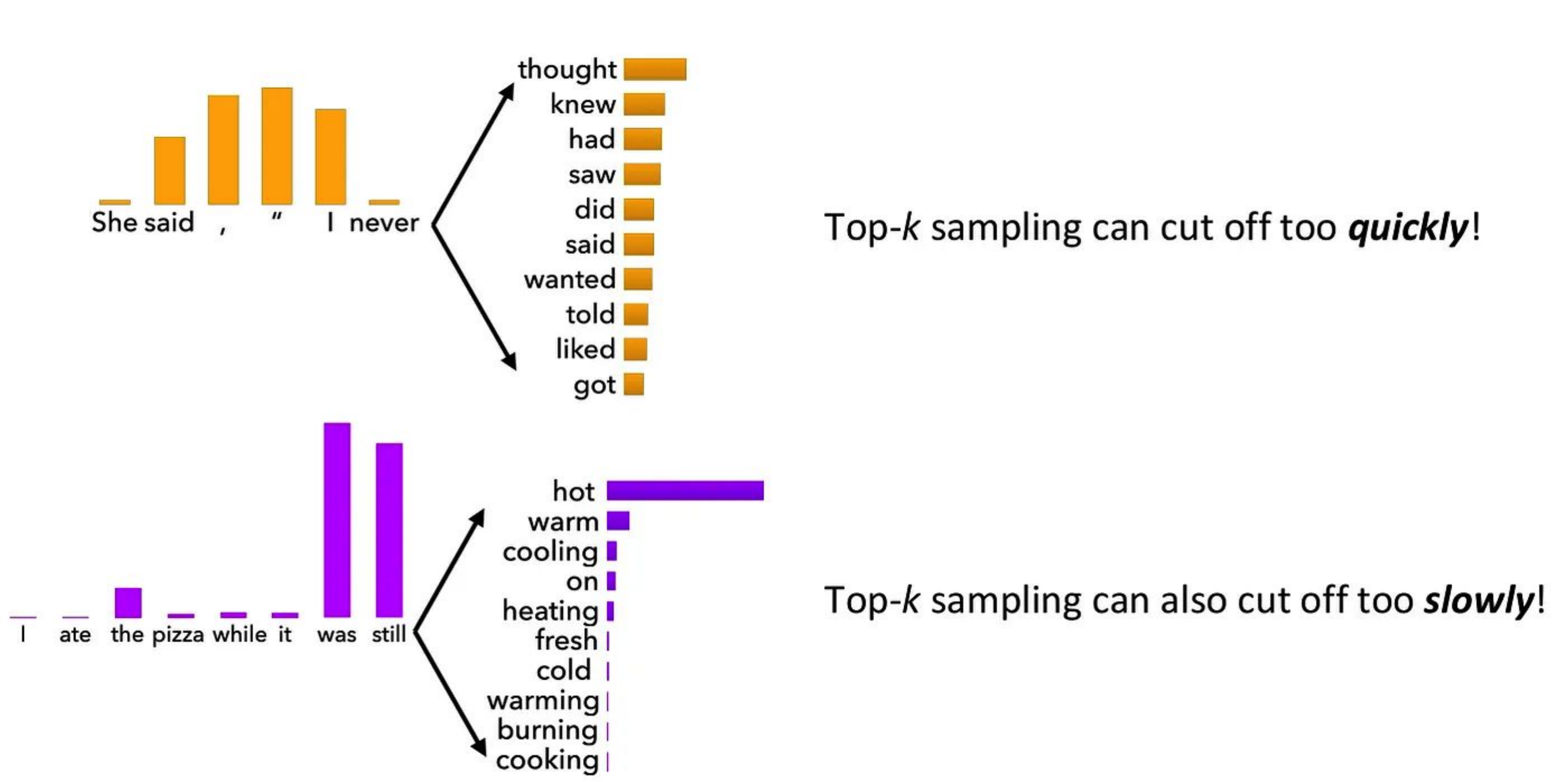
Prediction and Top-K



$$\hat{w}_t \sim \text{Top-k}(P(w_t|w_1, w_2, \ldots, w_{t-1}))$$

Image Credit:
LM Po - Understanding LLM Decoding Strategies

The Problem Is...



Top-*k* sampling can cut off too ***quickly***!

Top-*k* sampling can also cut off too ***slowly***!

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

Better: Graph Retrieval

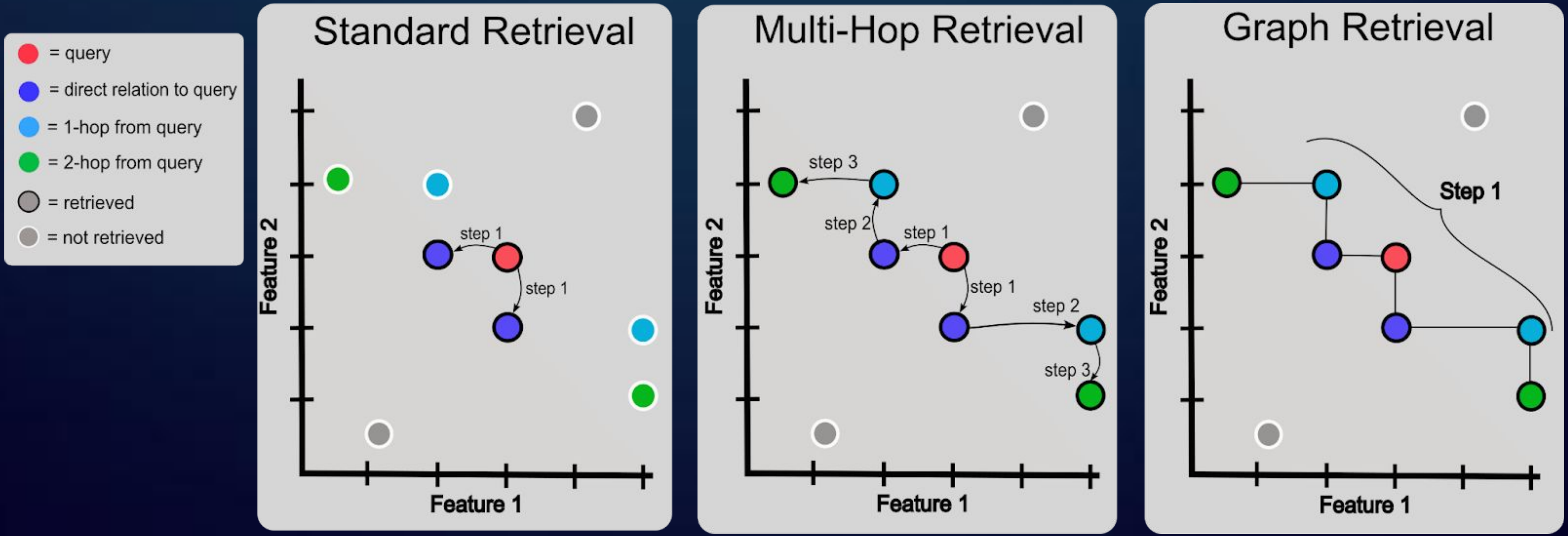


Image Credit:
Nick Alonso, Beren Millidge
Understanding Graph-based RAG and Multi-Hop Question Answering

Demo

https://youtu.be/WLEGg5zVwCQ

Explainable AI

Gain More Insights Into Your RAG Agents

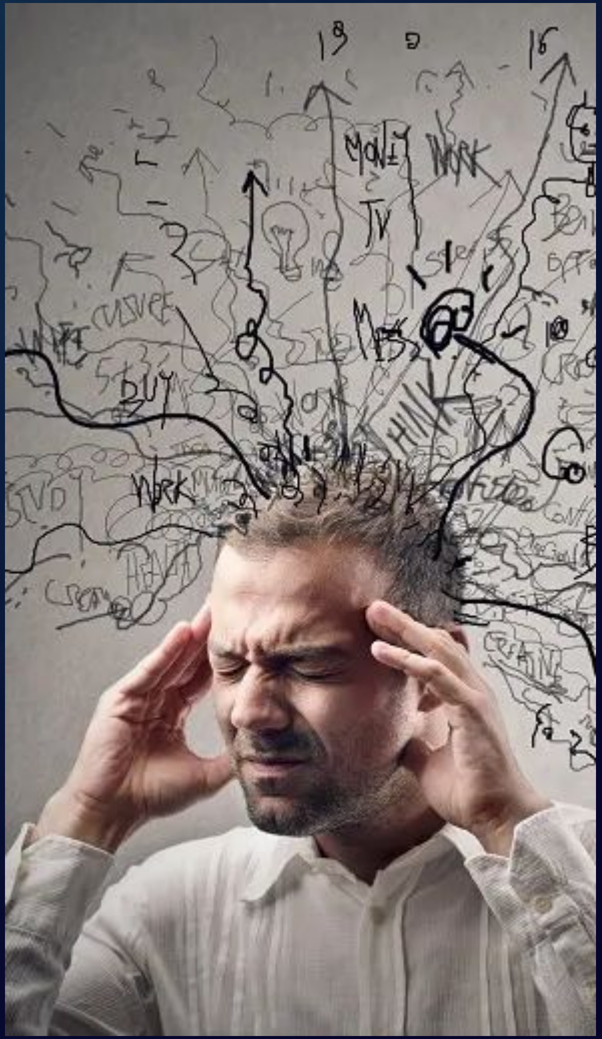
What is Explainable AI?

- Makes AI Decision Making Transparent & Understandable
  - LLMs/Embeddings = Black Box
  - Uncover the How and Why
- Goal is to Provide:
  - Trust and Validation
  - Bias & Error Detection
  - Collaboration (with Humans)
- My NVIDIA GTC Talk Video

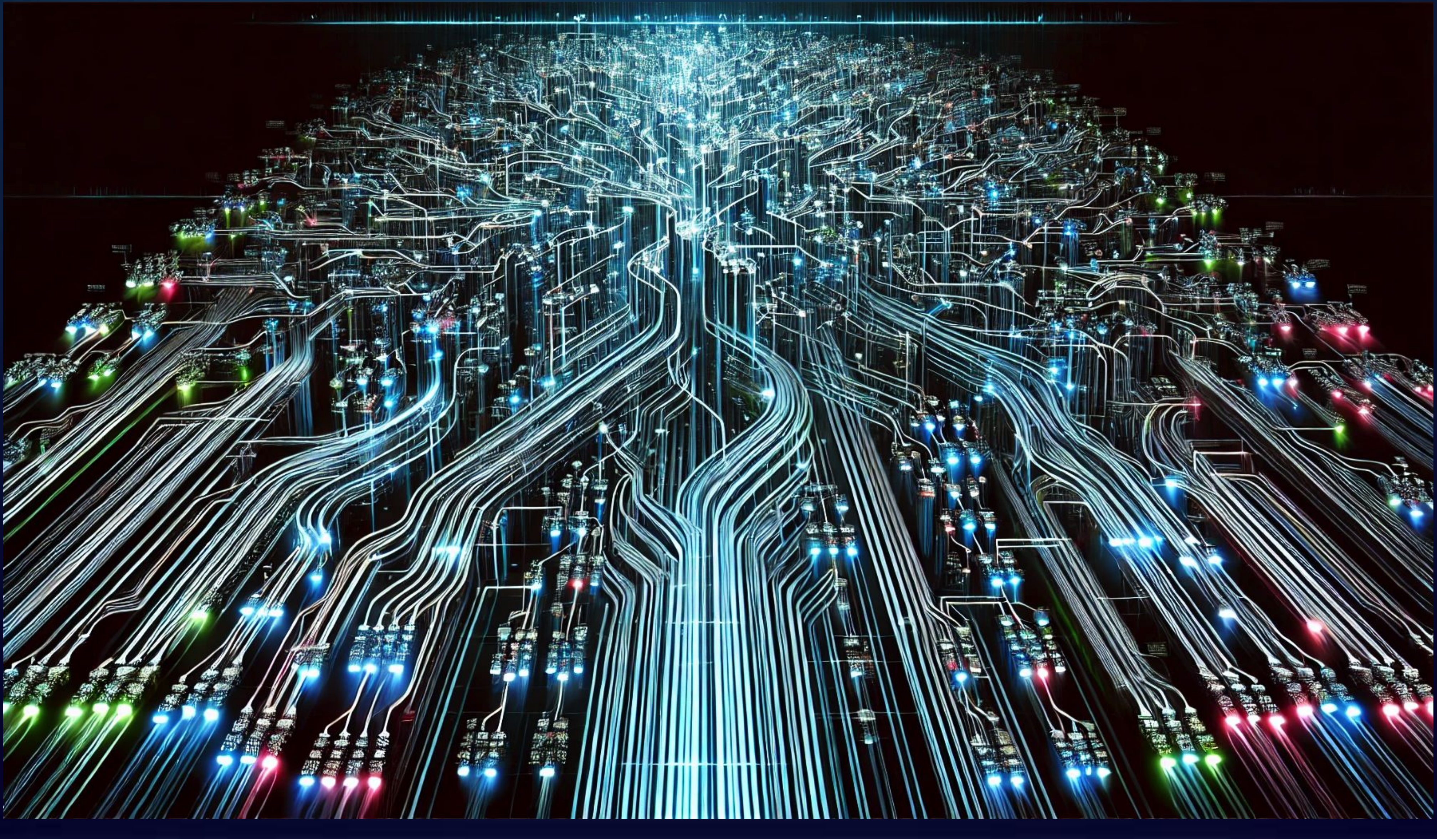
  Crack the AI Black Box: Practical Techniques for Explainable AI

Visualizing Data

- Graph DBs Offer:
  - Contextual Rep. -> Nodes, Edges, etc
  - Intuitive Visualization For Humans
  - Quick Glance Over Hops
  - How Everything Is Connected!
- Vector DBs:
  - Opaque: High-dimensional Embeddings
  - Flattened Connections: Related But How?
  - Limited Visibility: Difficult to Browse
  - "No Chain Of Nodes" Due To Weights

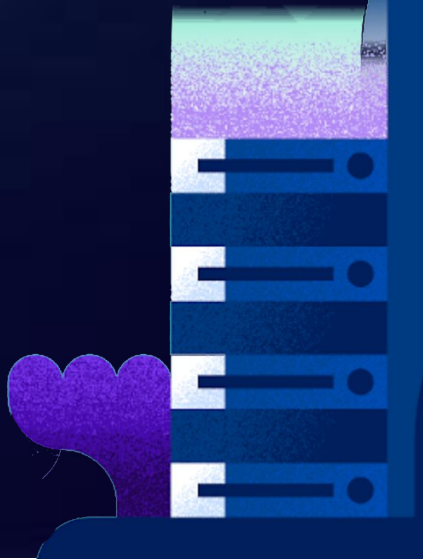


Visualizing Vector Data



Demo

https://youtu.be/DDajZ5nS7aU

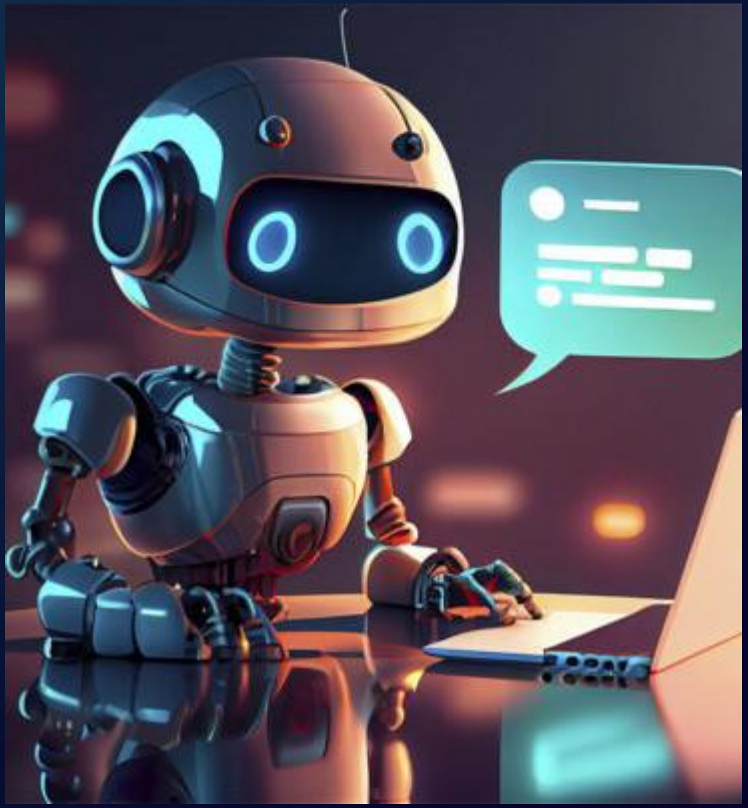


Leverage in Non-AI Apps

Consume (and Modify) Content in External Apps

Part 1: Chatbot Meets Graph Data

- Customer-Facing Chatbot for Retail Company
- Data is Stored in Graph Database
  - Product Info
  - User Purchase History
  - Supplier Inventory
- Chatbot Can Provide:
  - Contextually Relevant Info
  - Is the Item In Stock?
- Significantly Reduce Hallucinations



Part 2: Reports/Inventory Same Data

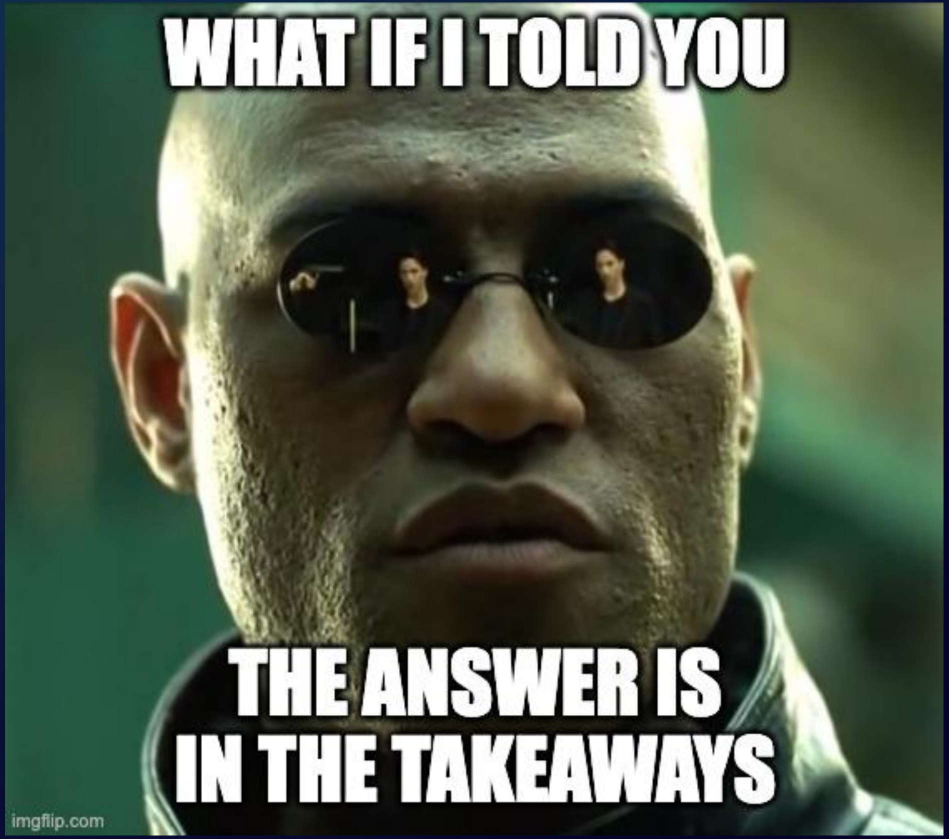
- Same Retail Company Can Use Directly Use Database
  - Inventory Management
  - Sales Reporting Tool
- Ex: Warehouse Dashboard of Orders
- Other Benefits:
  - Real-time Inventory Changes
  - User Data Instantly Updated
- Single Source of Truth!
- Conventional Business Aspects

Demo

https://youtu.be/FcEDJl1hDk4

Vector vs Graph: Takeaways

- Find the Right Tool For Your Use Case
- Graph Databases…
  - Can Use GPUs, CPU Optimized
  - Work Front Load = Modeling
  - Distribute the Ingest
- Vector Databases…
  - Need GPUs!!!!
  - Ingest/Embeddings Takes Time
  - Quick/Scalable



Resources

AI/ML Resources

[CLICK HERE] for All Material Contained in this Session [CLICK HERE]

DigitalOcean Bare Metal H200 Availability
https://www.digitalocean.com/blog/now-available-bare-metal-nvidia-hgx-h200-gpus

Continue the Conversation – DigitalOcean Discord
https://discord.com/invite/digitalocean

Graph Database Options:

- NebulaGraph - https://github.com/vesoft-inc/nebula
- Neo4j - https://github.com/neo4j/neo4j
- JanusGraph - https://github.com/JanusGraph/janusgraph

Dataset (BBC News) in Demo: https://bit.ly/4hBKNjp

Thank You!

David vonThenen

@davidvonthenen