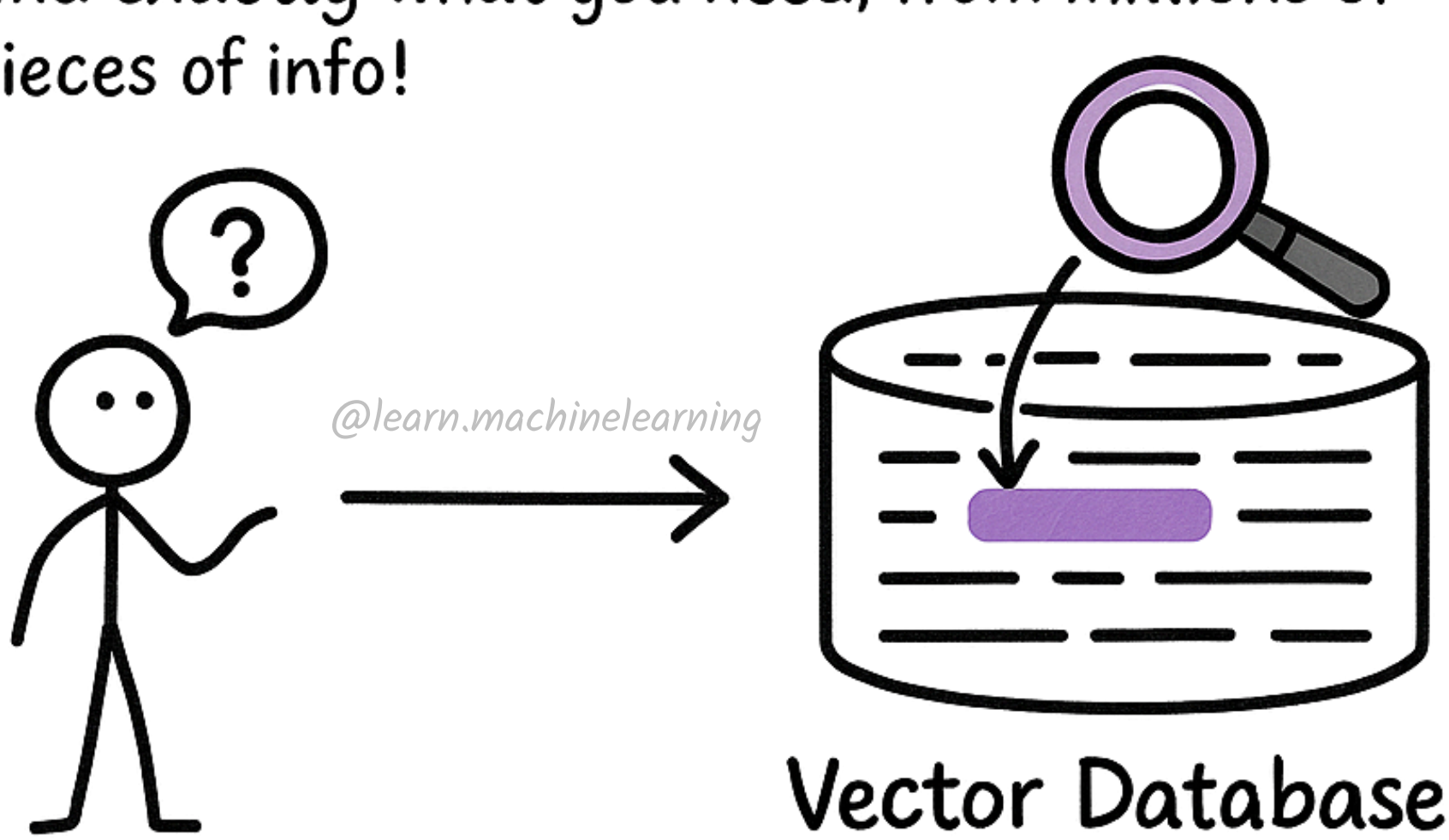
GENERATIVE AI SERIES: DAY 20

The Art of Retrieval: Finding the Right Information

Yesterday, we built the AI's knowledge library. Today, we learn how it searches that library to find *exactly* what you need, from millions of pieces of info!



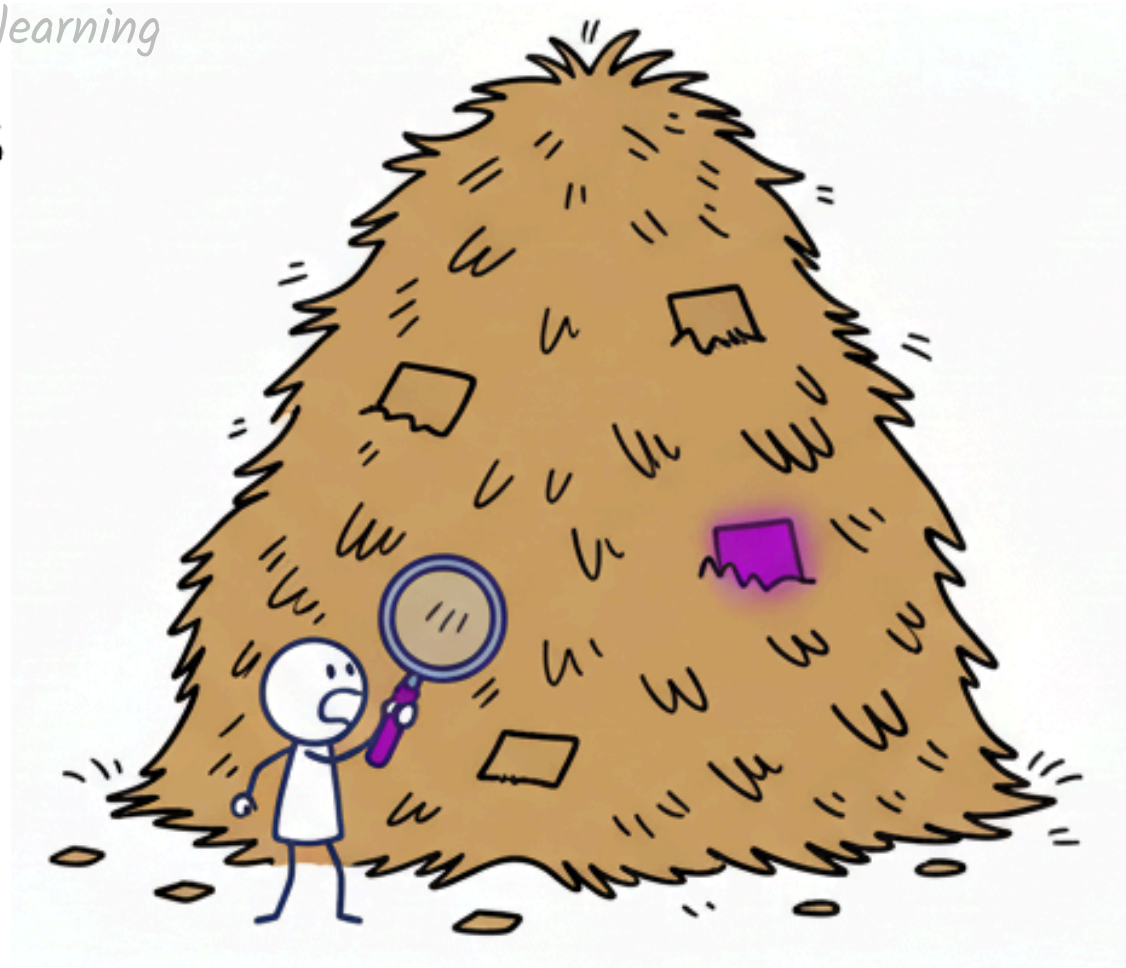
Your Question vs. & Millions of Chunks

Imagine your RAG system has processed thousands of documents, resulting in ***millions*** of small, embedded text chunks.

User Question: "What are the Q3 sales projections?"

The Problem: How does the system instantly find the few relevant chunks from its massive library?

The Goal: Fast, accurate, and contextually rich

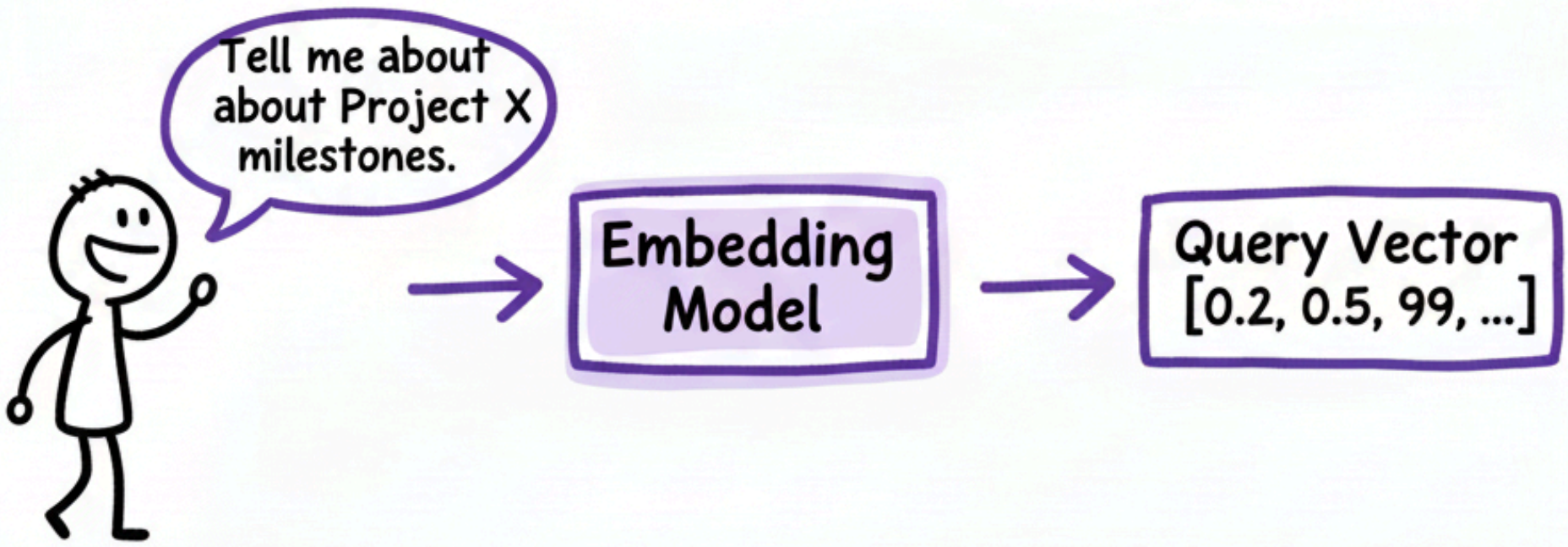


Day 20: The Art of Retreval

Step 3: The First Step: Turning Your Question into Numbers

- Just like we embedded our document chunks into vectors, we do the same for your question.
- **User Asks:** Tell me about Project X milestones.

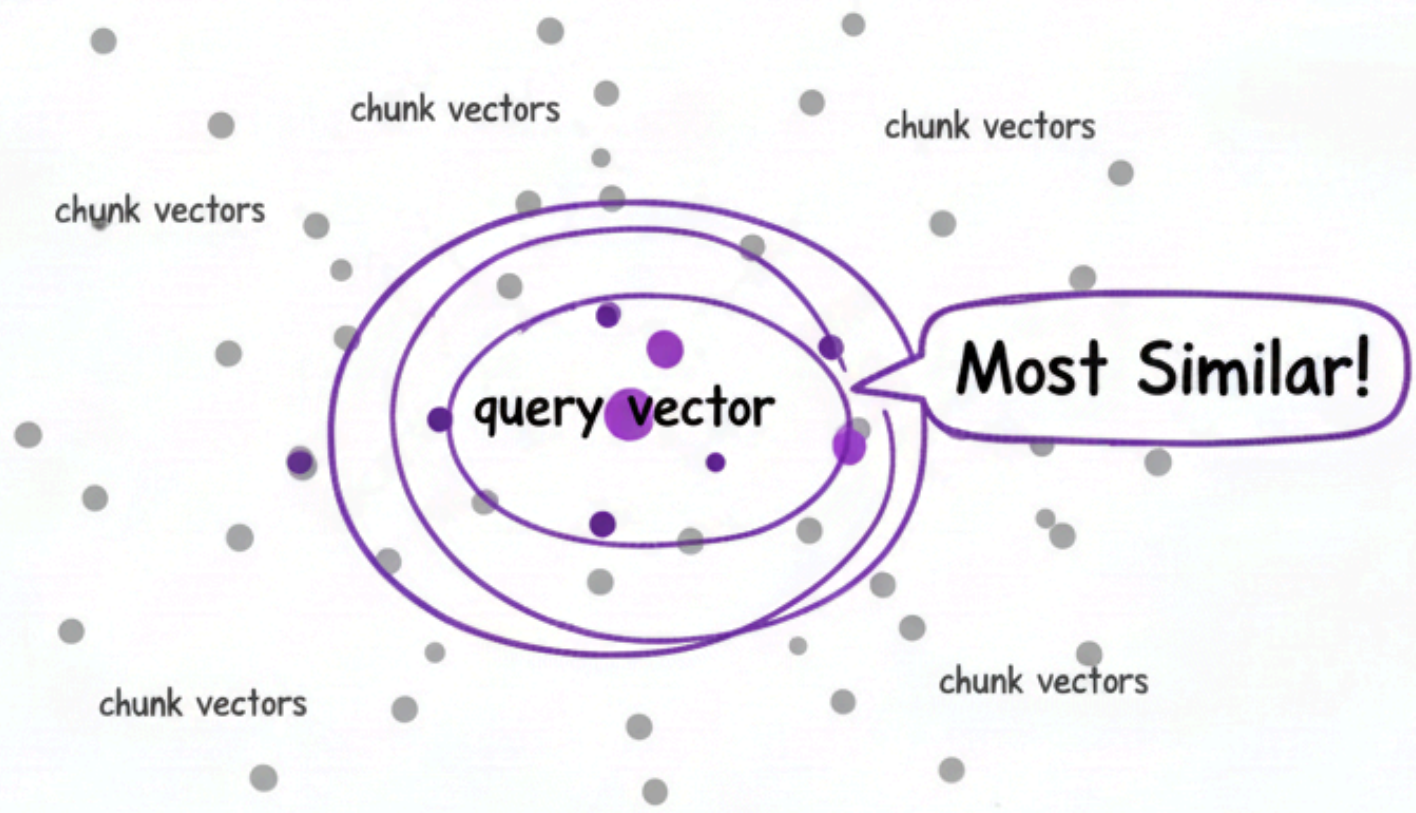
- **Embedding Model: Converts this question into query vector (a list of numbers).**
- **Why? Because computers great at comparing numbers to find find similarities!**



Vector Search: Finding Semantic Neighbors

- Now we compare the query vector to all the document chunk vectors storned in our database.
- Similarity Metrics: We use math (like Cosine Silliarity) to meaure how "close" two vectors are in multi-dimensional space.

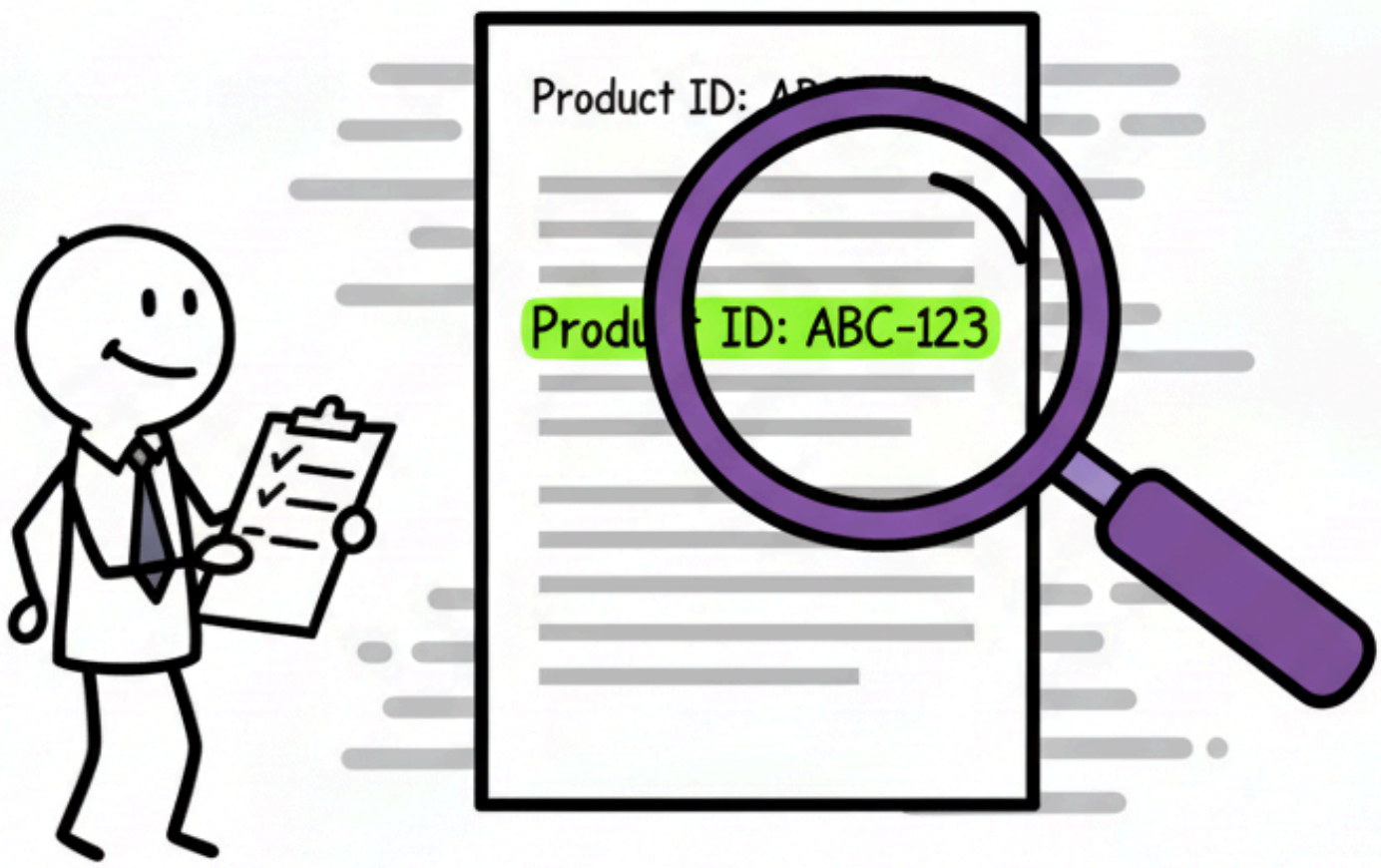
- Closer Vectors = More Relevant Meaning!
- Result: The system retrieves the top K (e.g, 5-10) most similar chunks. These are our potential answers!



Beyond Vectors: Keyword Search (BM25)

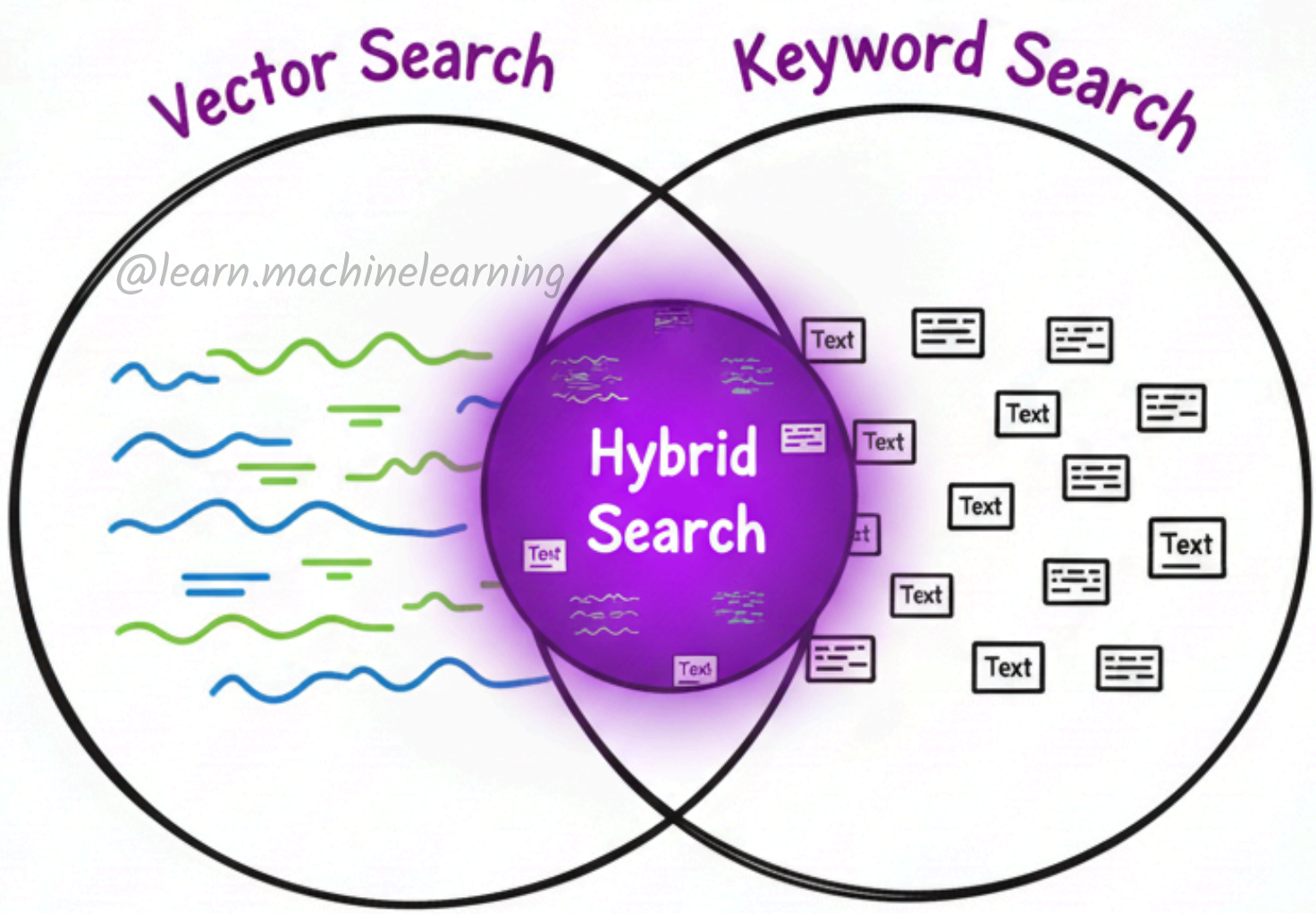
- Sometimes, pure semantic simiarity isn't enough. What if youre looking exact names, product codes, or very specific terms?
- Keyword Search (e.g.., BM25): This method focuss on finding exact matchs and their frequency within documents.
- Pros: Great for precise term matching,less prone "halluctating" if specific keywords it crucial.

- Cons: Can miss synnoms or context if the exact words arn't present.



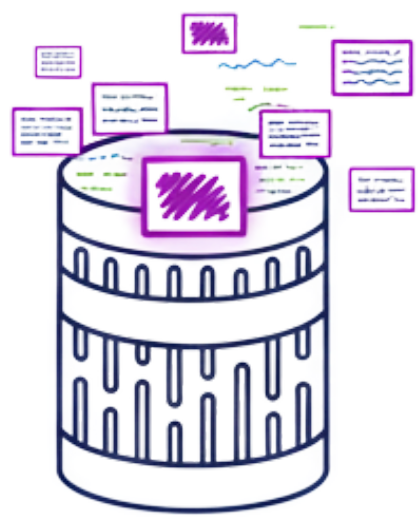
Hybrid Search: The Best of Both Worlds

- Why choose when you can have both?
- Hybrid Search: Combines the strengths of Vector Search (semantic undirstanding) and Keyword Search (exact matching).
- Benefit: Catches both conceptualy similar information and exact term matches, leading to higher quality answers.



Day 20 Takeaway:
Retrieval is a Smart Search Party!

- Retrieval is the critical "live" act of finding relevant information. By transforming questions into vectors, using
- Vector Search for semantic understanding, and sometimes combining it with *Keyword Search* in a Hybrid approach, your RAG system becomes an expert librarian, always finding the right book.



Vector Database

Next Up (Day 21):

Day 21: The Generation Step: Synthesizing Answers with Context

@learn.machinelearning

What kind to documents would you want your own AI to learn from? ⇩

Follow @learn.machinelearning for the full RAG Deep Dive!

SAVE this post if you want your yont your AI to find the right answers evers every time!