**Why use RAG:**

* Limited knowledge access
* Struggle to provide relevant information
* Hallucinations

**RAG Architecture:**

* Ingestion:

Document 🡪 Chunks 🡪 Embeddings 🡪 Index

* Retrieval:

Query 🡪 Index 🡪 Top K (result)

* Generation:

LLM 🡪 Response

**Retrieval Process:**  
 🡸 Query 🡸

Data 🡺 Chunks 🡺 Database 🡺 Top-K 🡺 LLM

The user is querying and based on similarity search we going to find out top-k responses.

We take out documents and embed them into this in 3D space we take our question and do the same we can then do a search like a local neighborhood search you can think about in this 3D space around our question to say what documents are nearby and then neighbor are retrieved because they have similar semantics similar to our question.

**Sentence-Window Retreival / Small to large Chunking:**