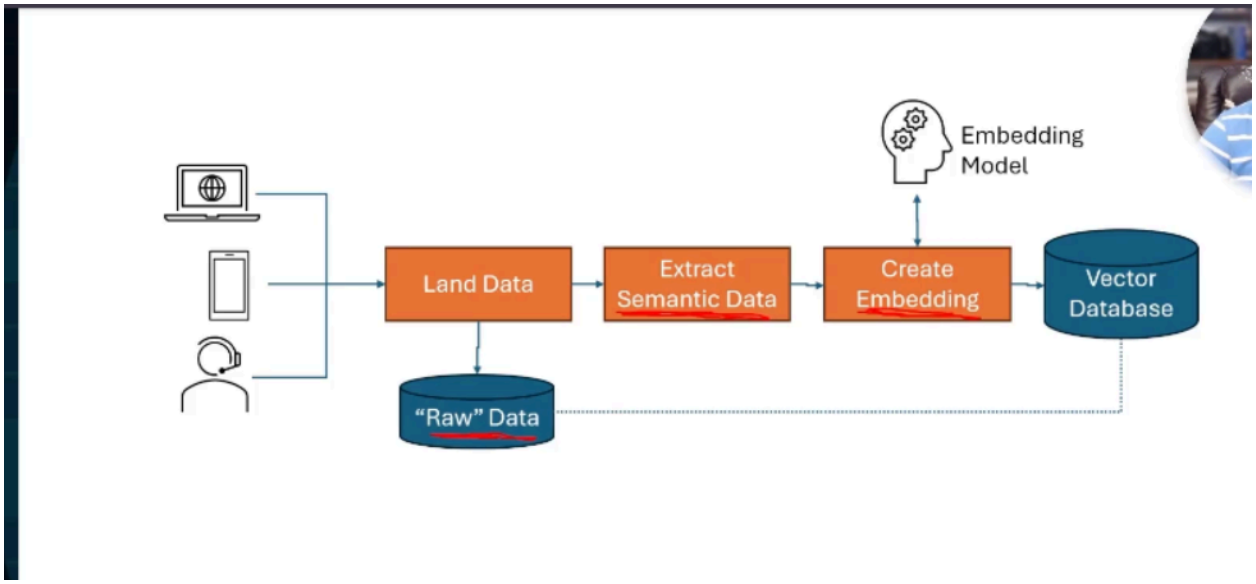


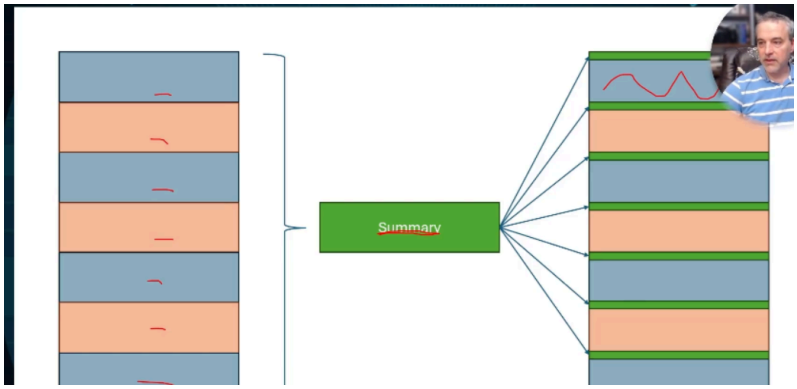
The process of converting raw data into a vector database ^
We are trying to capture the meaning of the document by looking at the words,phrases and sentences of the document and trying to map them in more than 3 dimensions, probably 100s or 1000s of dimensions, to capture the essence of what the document is trying to say.

Chunking:

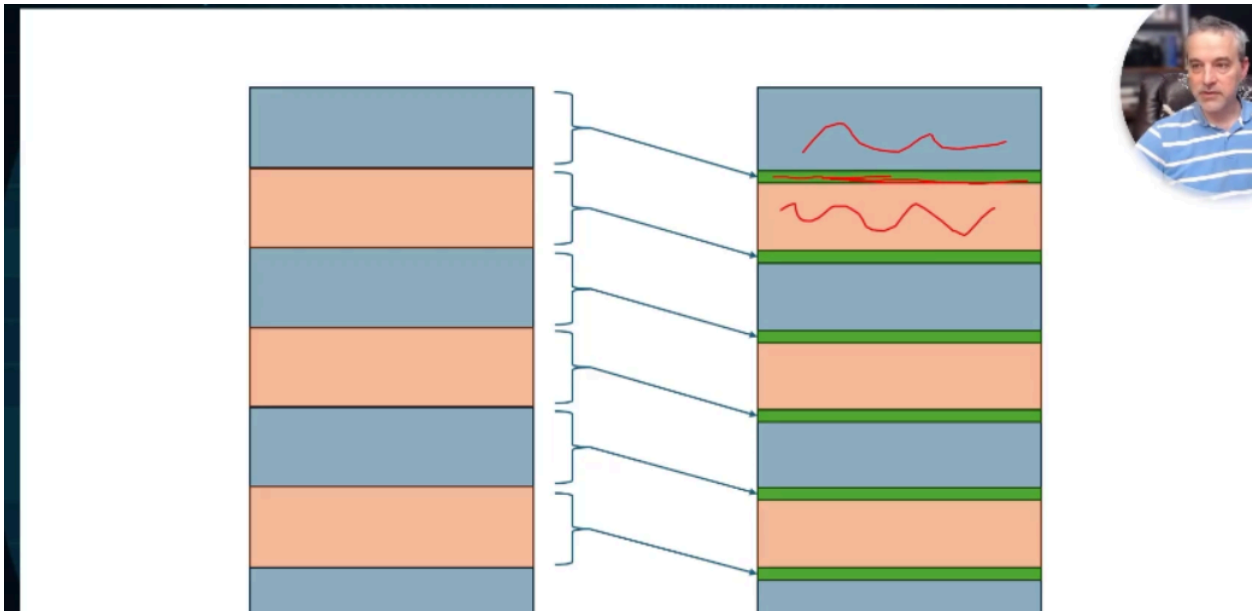
- Naive approach: split the text based on number of words or characters or lines, the drawback is that we will lose a lot of context and have a lack of continuity between the chunks, eg. one chunk might contain some info that continues into the next chunk, that is now not as useful as it is divided into parts

- Overlap: the text is split while maintaining an overlap between splits, eg. first chunk: "this is a chunk 1 abcdefghijk…."
  Second chunk: "a chunk1 abcdefghjk…."
  While this helps to maintain some level of continuity and context between chunks, there is still some loss of context. Eg. the last chunk does not have any link between first chunk which might contain some info about the last chunk

- Summary:



We summarize the whole paragraph and use that summary in portions along with each chunk.
Another method :



In this method, the each chunk is summarized and passed to the next chunk as an addition.

Popular Tools and VectorDBs:

Tool	Description
Elasticsearch	While primarily a search engine, Elasticsearch has incorporated vector search capabilities, allowing users to perform similarity searches using vectors with its dense vector data type and cosine similarity functions.
Faiss (Facebook AI Similarity Search)	Developed by Facebook AI Research, Faiss is a library for efficient similarity search and clustering of dense vectors. It is particularly well-known for its speed and efficiency in handling large vector datasets.
Milvus	An open-source vector database designed to handle large-scale vector similarity searches. Milvus supports multiple distance metrics and can be integrated with popular machine learning frameworks.
Weaviate	An open-source vector database that supports GraphQL and RESTful APIs. Weaviate uses machine learning models to index data and supports various vectorization techniques and similarity metrics.
Pinecone	A managed vector database service that is designed to scale and handle large volumes of vector similarity searches efficiently. Pinecone supports various metrics for similarity search and offers easy integration with existing data pipelines.
Annoy (Approximate Nearest Neighbors Oh Yeah)	Developed by Spotify, Annoy is a C++ library with Python bindings for approximate nearest neighbor searches. It allows users to create data structures that can be used for querying similar items in large datasets.
PostgreSQL with PGVector	This is a widely used open-source relational database. The PGVector extension enables vector searches that support distance metrics.
SQL Server 2022	SQL Server incorporates vector search capabilities through Vector Indexes. This feature allows efficient storage and retrieval of high-dimensional vectors. It enables SQL Server to perform similarity searches in real-time.