Vector Databases: Bridging Semantic Understanding in Data Storage

What is a Traditional Database?

- Traditional databases store data in structured formats
- Typically use tables with predefined columns
- Limited in handling unstructured data like images, text, and audio
- Metadata and tags provide basic information
- Struggles with semantic understanding of complex data

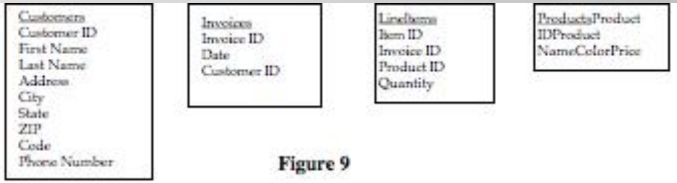


Figure 9

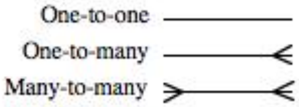


Figure 10

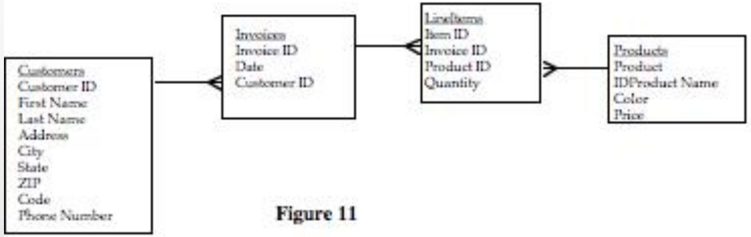
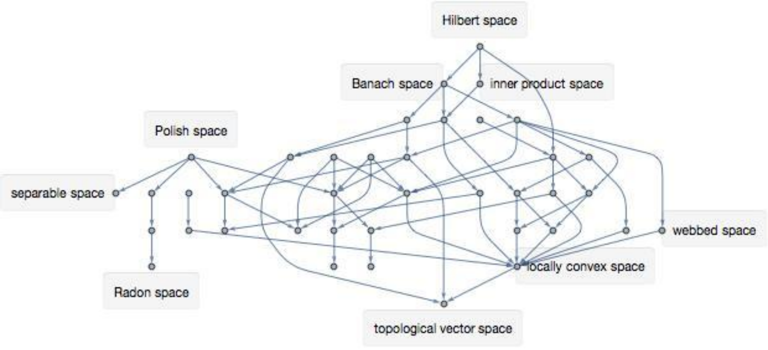


Figure 11

The Semantic Gap Challenge

- Traditional databases cannot capture nuanced data characteristics
- Difficulty in searching for conceptual similarities
- Limited by rigid, structured query methods
- Cannot easily find semantically related content
- Challenge: How do we represent complex data meaningfully?

Enter Vector Databases



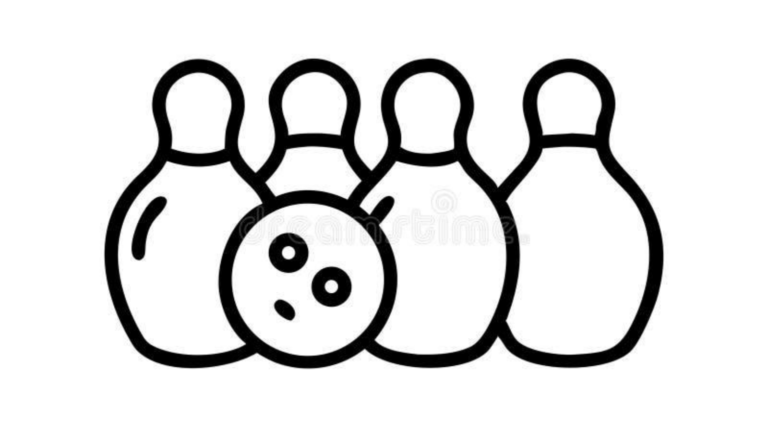
- Represent data as mathematical vector embeddings
- Convert unstructured data into arrays of numbers
- Each vector captures semantic essence of the data
- Similar items are positioned close together in vector space
- Dissimilar items are positioned far apart

What Can Be Stored in Vector Databases?

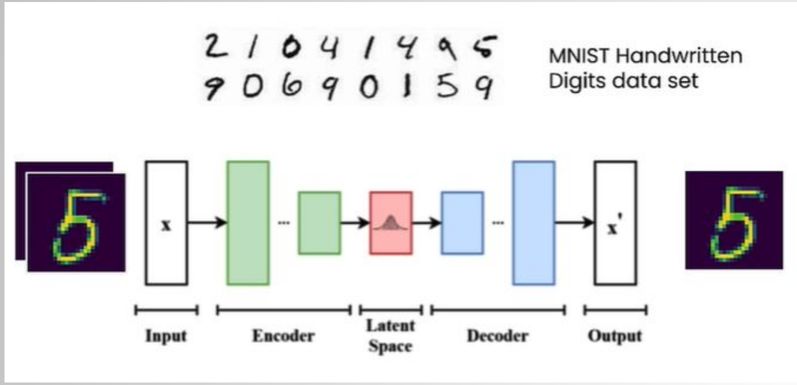
- Images
- Text files
- Audio recordings
- Video content
- Complex, unstructured digital objects

Understanding Vector Embeddings

- Arrays of numbers representing data features
- Each dimension captures a specific characteristic
- Can have hundreds or thousands of dimensions
- Learned through specialized embedding models
- Provides a multi-dimensional representation of data



Embedding Models



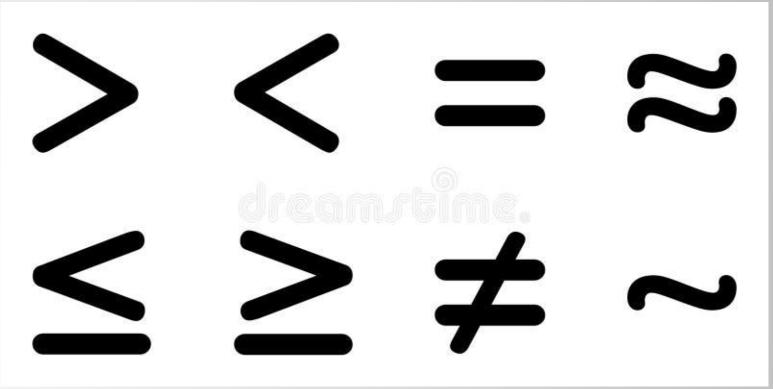
- Specialized models for different data types
- Image embeddings: CLIP
- Text embeddings: GloVe
- Audio embeddings: Wav2vec
- Process data through multiple layers
- Extract progressively abstract features

How Embedding Models Work

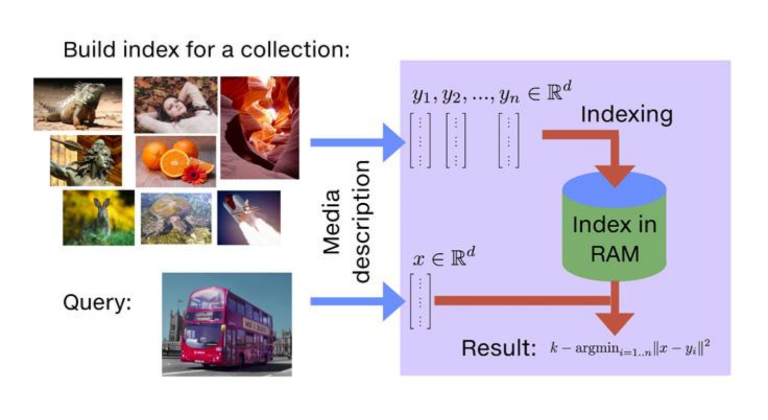
- Data passes through multiple neural network layers
- Early layers detect basic features
- Deeper layers recognize complex characteristics
- Produce high-dimensional vectors
- Capture essential input characteristics

Similarity Search

- Find items similar to a query item
- Compare vector positions in vector space
- Mathematically calculate proximity
- Enables semantic-based retrieval
- More nuanced than traditional keyword search



Vector Indexing Challenge


Build index for a collection:
$y_1, y_2, \ldots, y_n \in \mathbb{R}^d$
Indexing
Media description
Index in RAM
Query:
$x \in \mathbb{R}^d$
Result: $k - \mathrm{argmin}_{i=1..n} \|x - y_i\|^2$

- Millions of vectors with hundreds of dimensions
- Direct comparison is computationally expensive
- Solution: Approximate Nearest Neighbor (ANN) algorithms
- Trading slight accuracy for significant speed
- Examples: HNSW and IVF indexing methods

Practical Application: RAG

- Retrieval Augmented Generation (RAG)
- Store document chunks as vector embeddings
- When a question is asked, find relevant text chunks
- Use vector similarity to retrieve information
- Feed retrieved data to large language models

Key Advantages

- Handle unstructured data effectively
- Semantic understanding beyond keywords
- Fast and efficient similarity searches
- Support complex AI and machine learning applications
- Bridge gap between human and computer data interpretation

DATA
INNOVATION
SUMMIT
2019
asweden
expe
Technology in itself is neutral,
it, or not.
In the data driven world, both benefits
exponential - all organizations should
their use of data may impact society

Real-World Use Cases

- Recommendation systems
- Content search engines
- Medical image analysis
- Natural language processing
- Multimedia content management

Embedding Model Training

- Requires massive datasets
- Supervised and unsupervised learning techniques
- Continuous model refinement
- Domain-specific training
- Improves vector representation accuracy

Future of Vector Databases

- Growing importance in AI and machine learning
- Increasing computational power
- More sophisticated embedding models
- Broader application across industries
- Potential to revolutionize data understanding

Reflection and Discussion

- How might vector databases transform your field of study?
- What challenges do you see in implementing this technology?
- Can you imagine innovative applications for semantic data retrieval?
- Discuss the potential ethical implications of advanced data representation