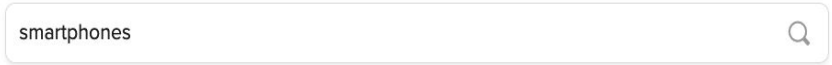
Vector Databases and ANN algorithm

Keyword search vs Semantic search



exact term matching

“Top smartphones on the market…”

“Smartphone industry struggles…”

synonyms and other similar/related terms

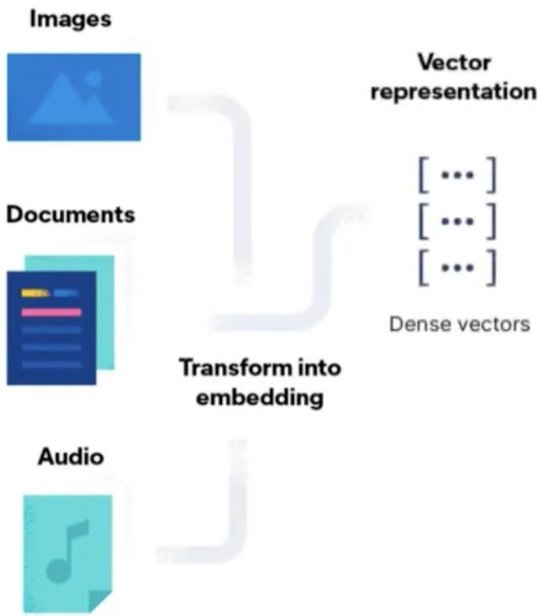
“Using a cell phone in Spain…”

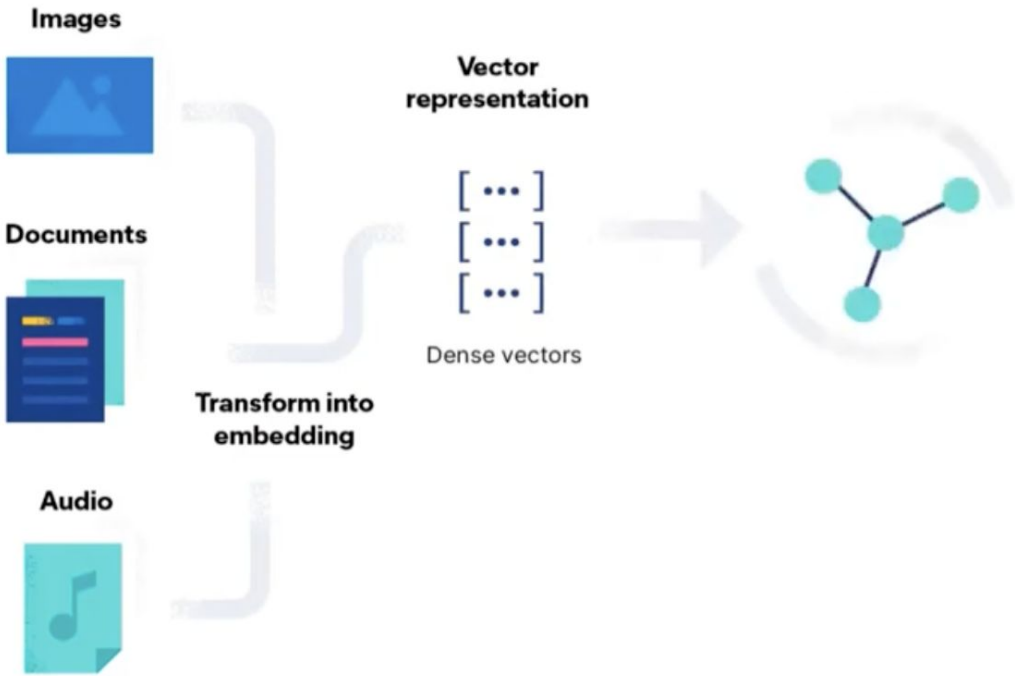
“Mobile device usage on the rise…”

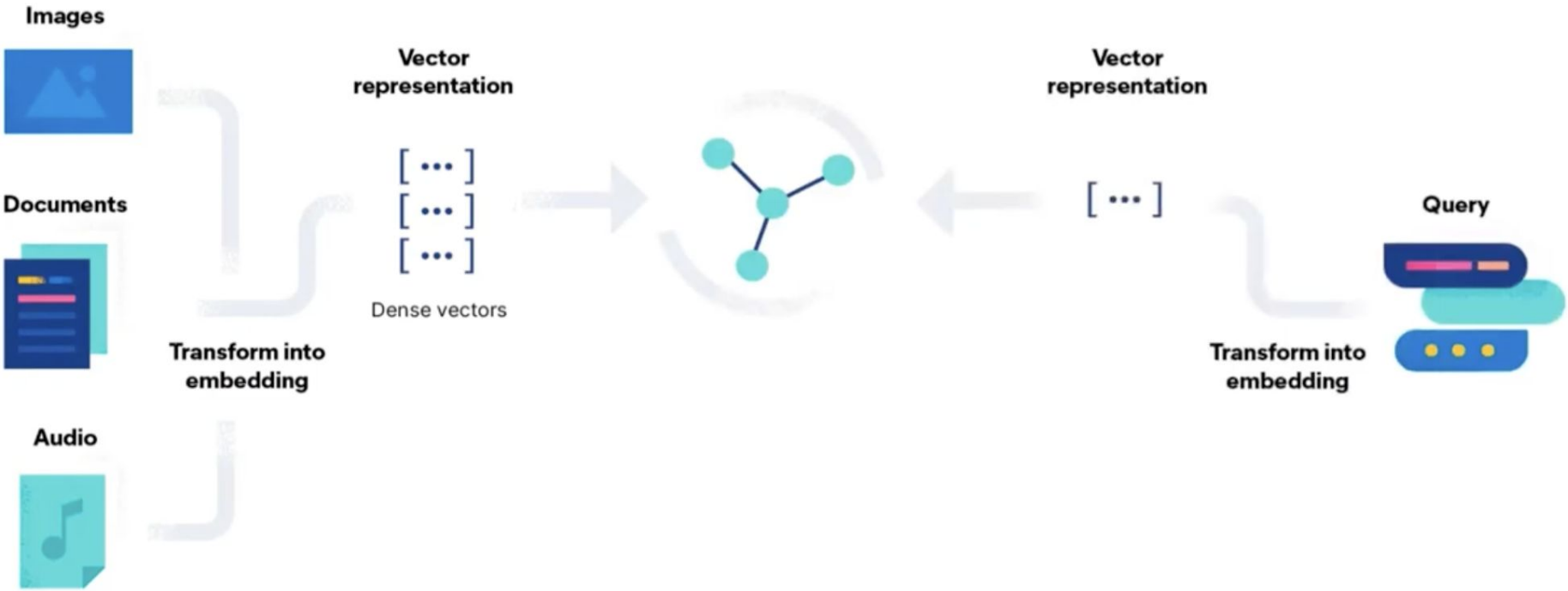
Images

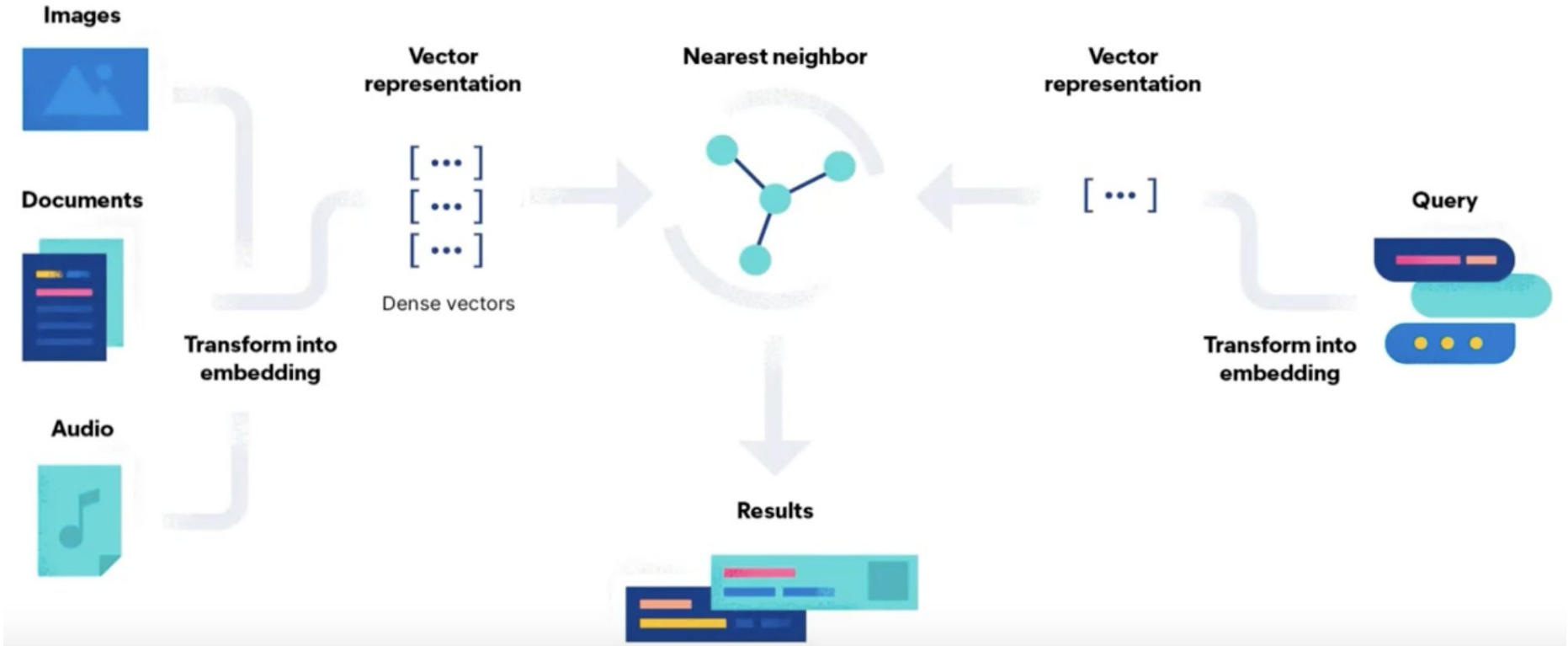
Documents

Audio


Images
Documents
Audio
Transform into embedding
Vector representation
[...]
[...]
[...]
Dense vectors


Images
Documents
Audio
Transform into embedding
Vector representation
Dense vectors


Images
Documents
Audio
Transform into embedding
Vector representation
Dense vectors
Vector representation
Query
Transform into embedding

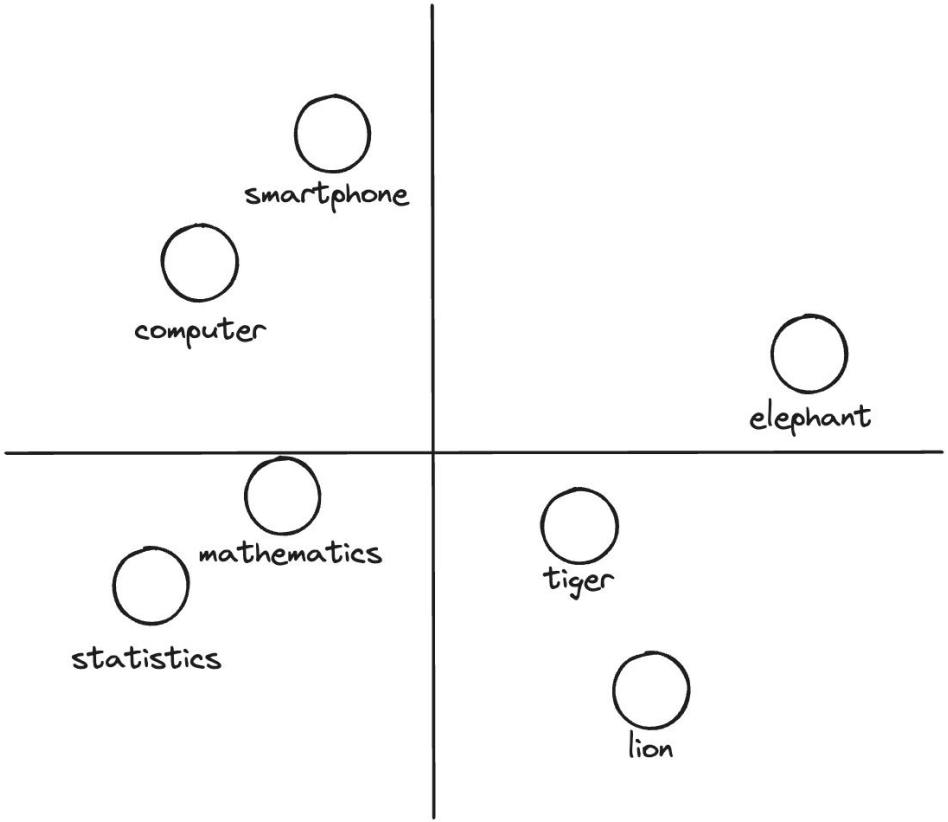

Images
Documents
Audio
Transform into embedding
Vector representation
Dense vectors
Nearest neighbor
Results
Vector representation
Transform into embedding
Query

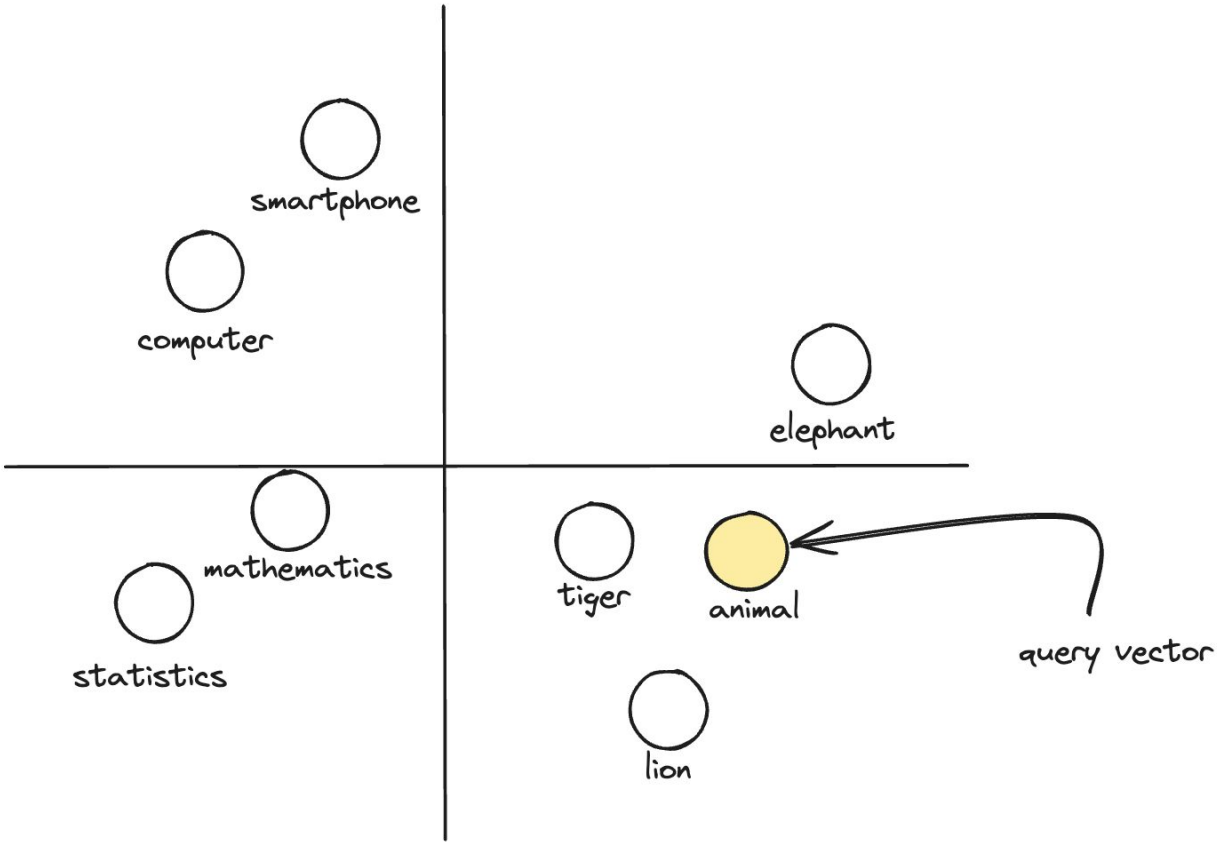
Semantic search is Vector search

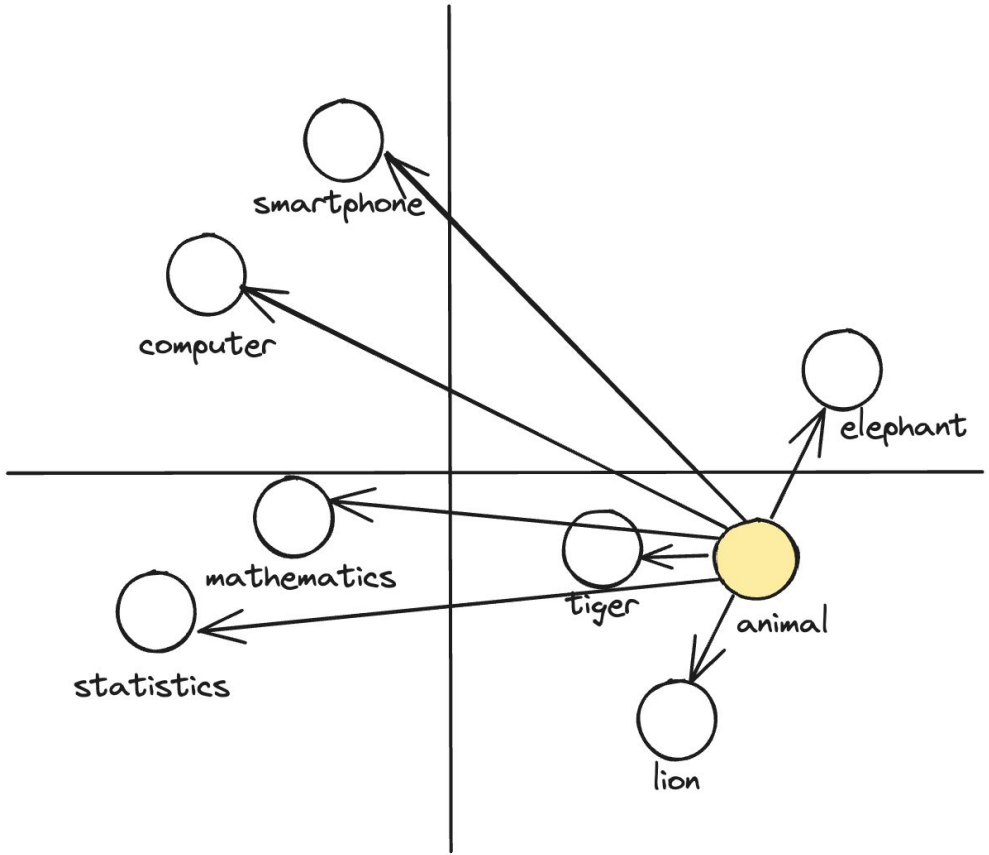
K Nearest Neighbour

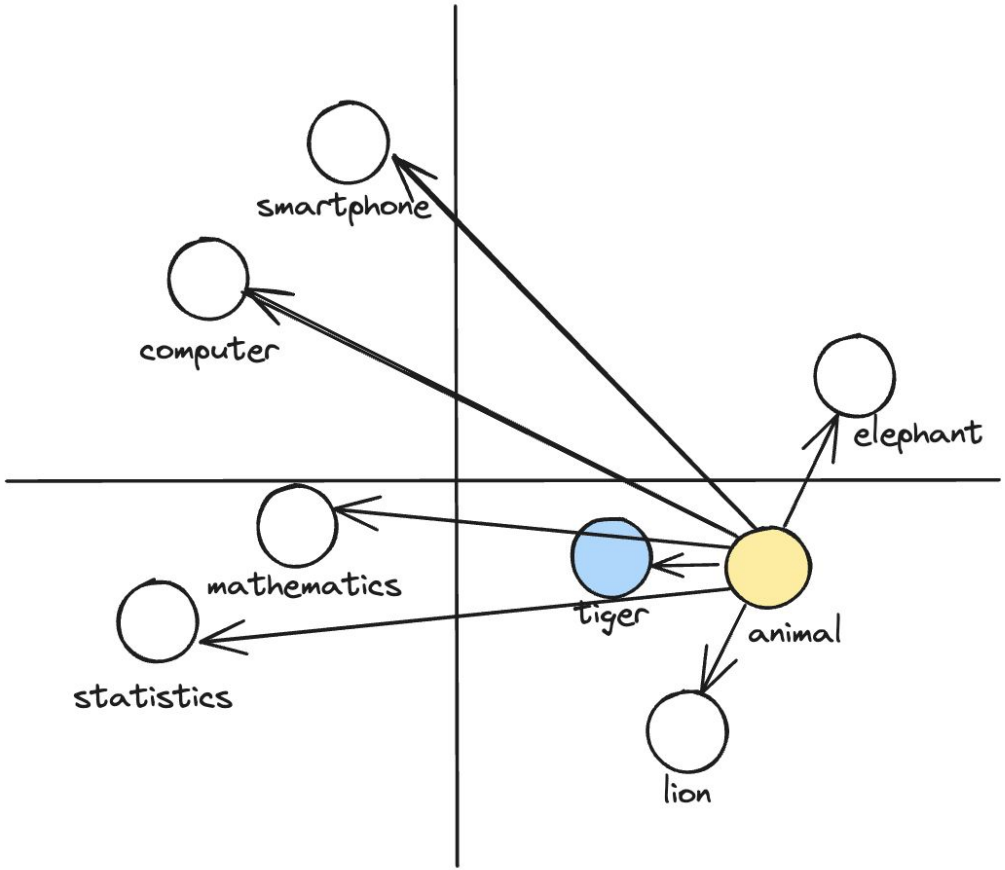
Exhaustive, brute-force approach

- Measure distance from query and all other vectors
- Sort distances
- Return top K results

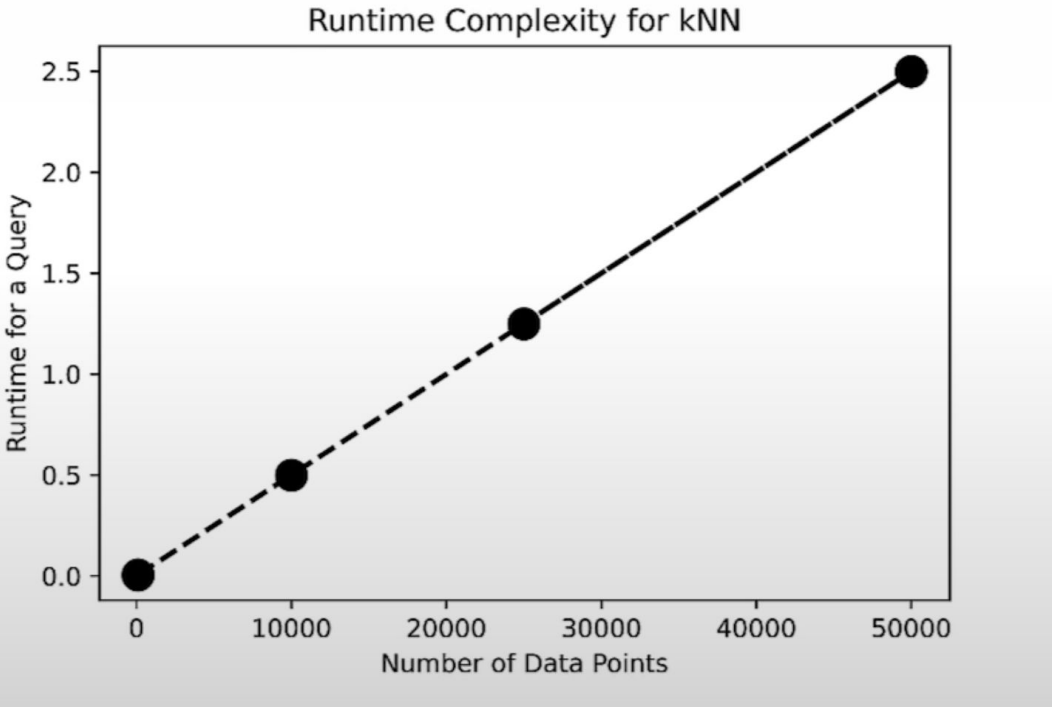

smartphone
computer
elephant
mathematics
statistics
tiger
lion


smartphone
computer
elephant
mathematics
statistics
tiger
animal
query vector
lion


smartphone
computer
elephant
mathematics
tiger
animal
statistics
lion


smartphone
computer
elephant
mathematics
tiger
animal
statistics
lion

Performance



- Number of dimension ~1000
- Number of vectors ~millions
- Typical query latency ~seconds

Trade off accuracy for performance

Approximate Nearest Neighbour

- Preprocess data into efficient index
- Speed up search using index

Approximate Nearest Neighbour

- Tree-based algorithm (ANNOY)
- Locality sensitive hashing
- Product Quantization
- Inverted File Index
- **Proximity Graph (Hierarchical Navigable Small World)**

Navigable Small World

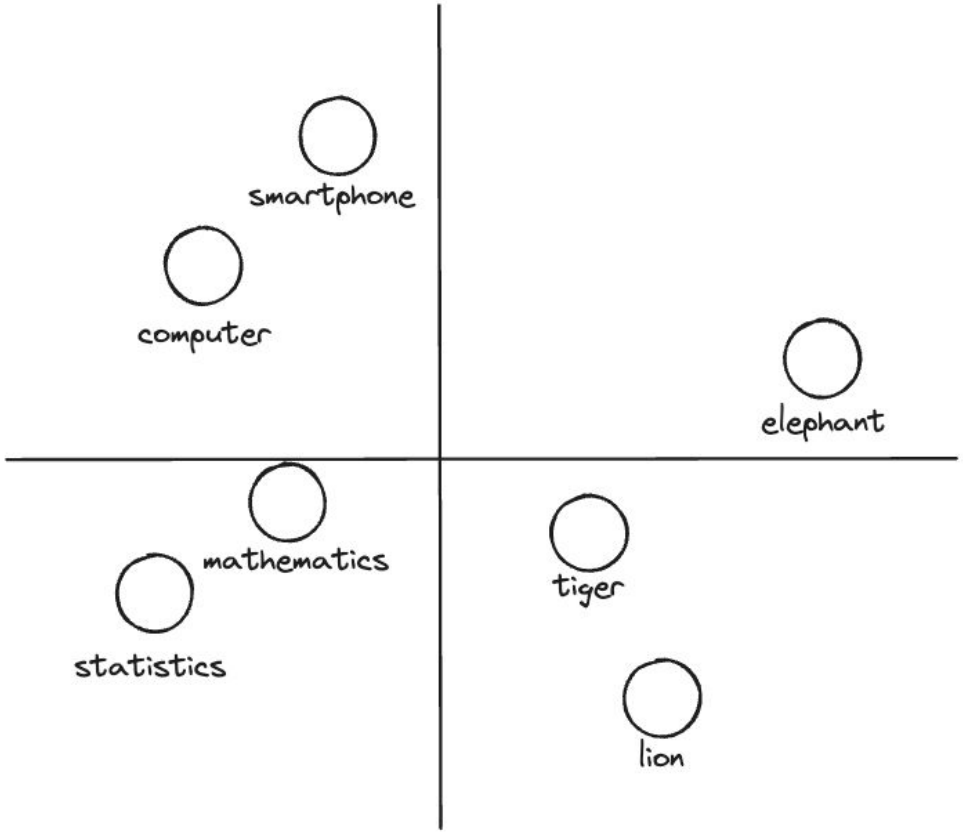
- Six degrees of separation
- Hungarian novelist – Karinthy Frigyes – 1929

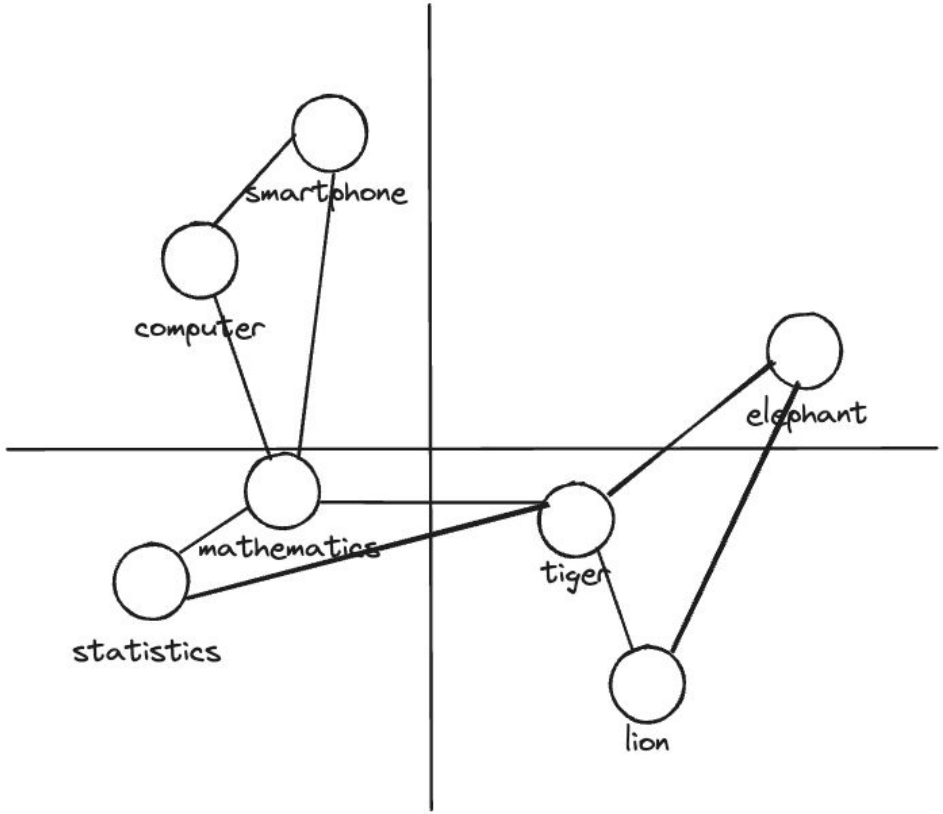
- Better communication tech and travel
- Migrations and increasing connectivity
- "Shrinking" modern world

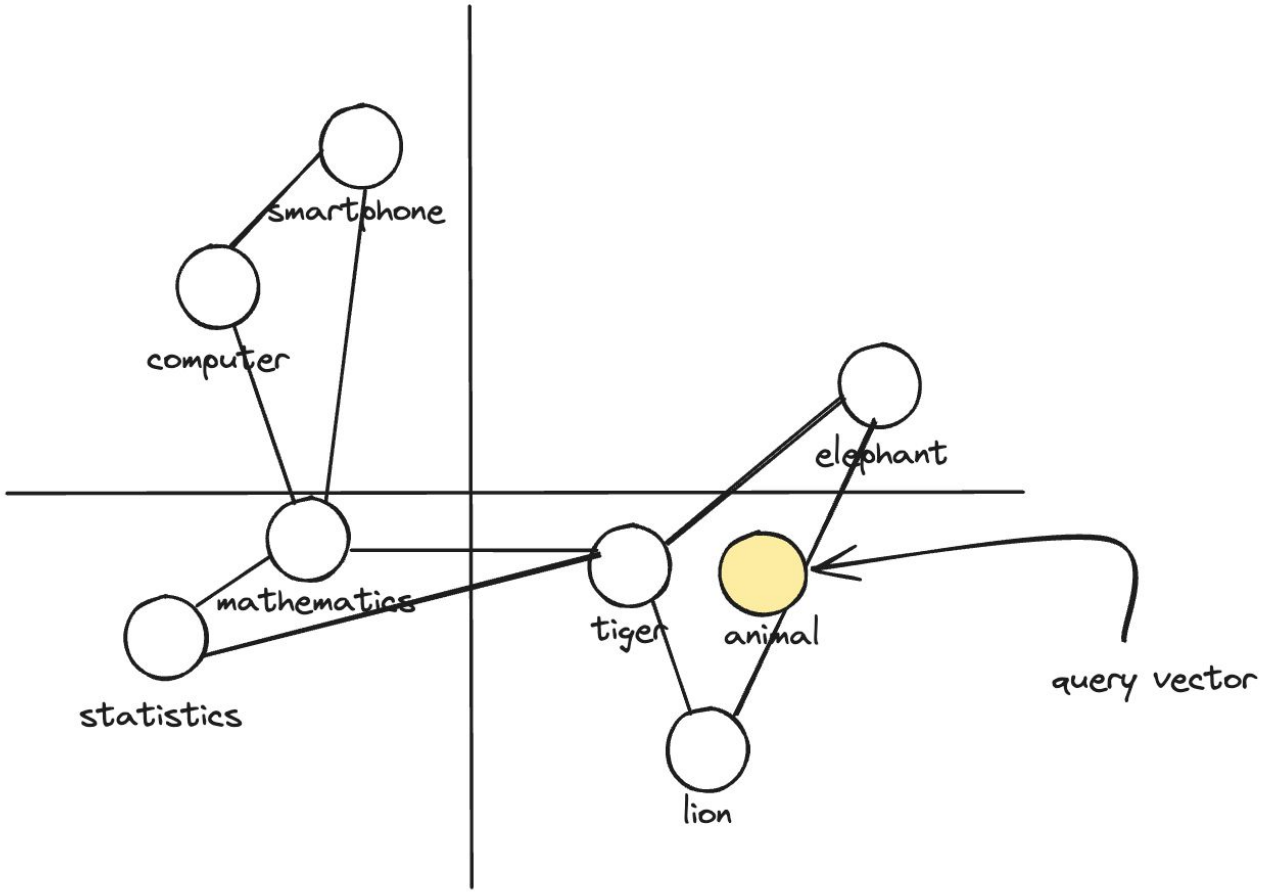
Navigable Small World

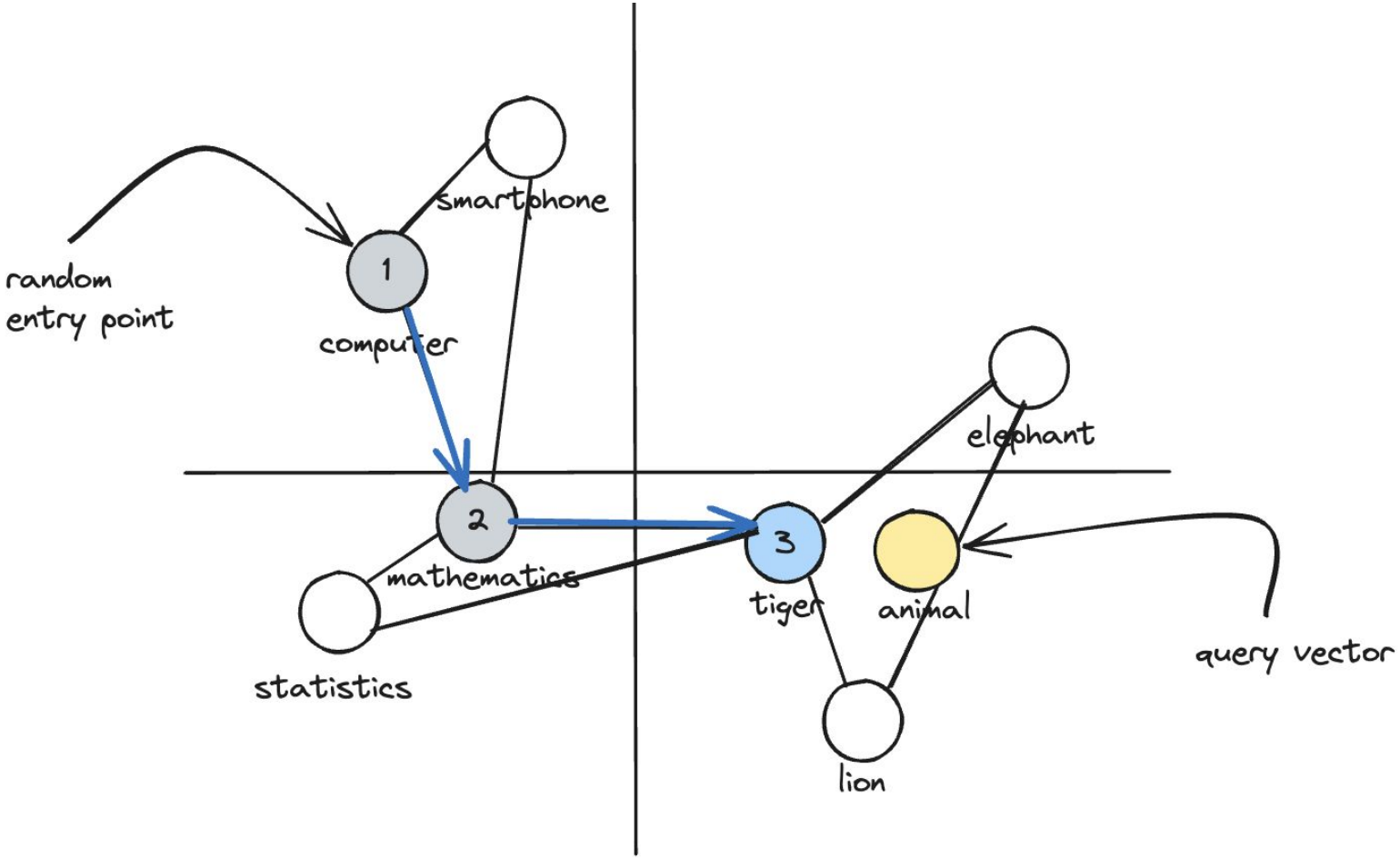
- every node can be reached with small number of hops from any other node

- Build a graph
- Every node has M connection to closest nodes (proximity list, “friend list”)
- Random entry node
- Greedy search, move to closest node
- Stop condition, no closer node in friend list


smartphone
computer
elephant
mathematics
statistics
tiger
lion

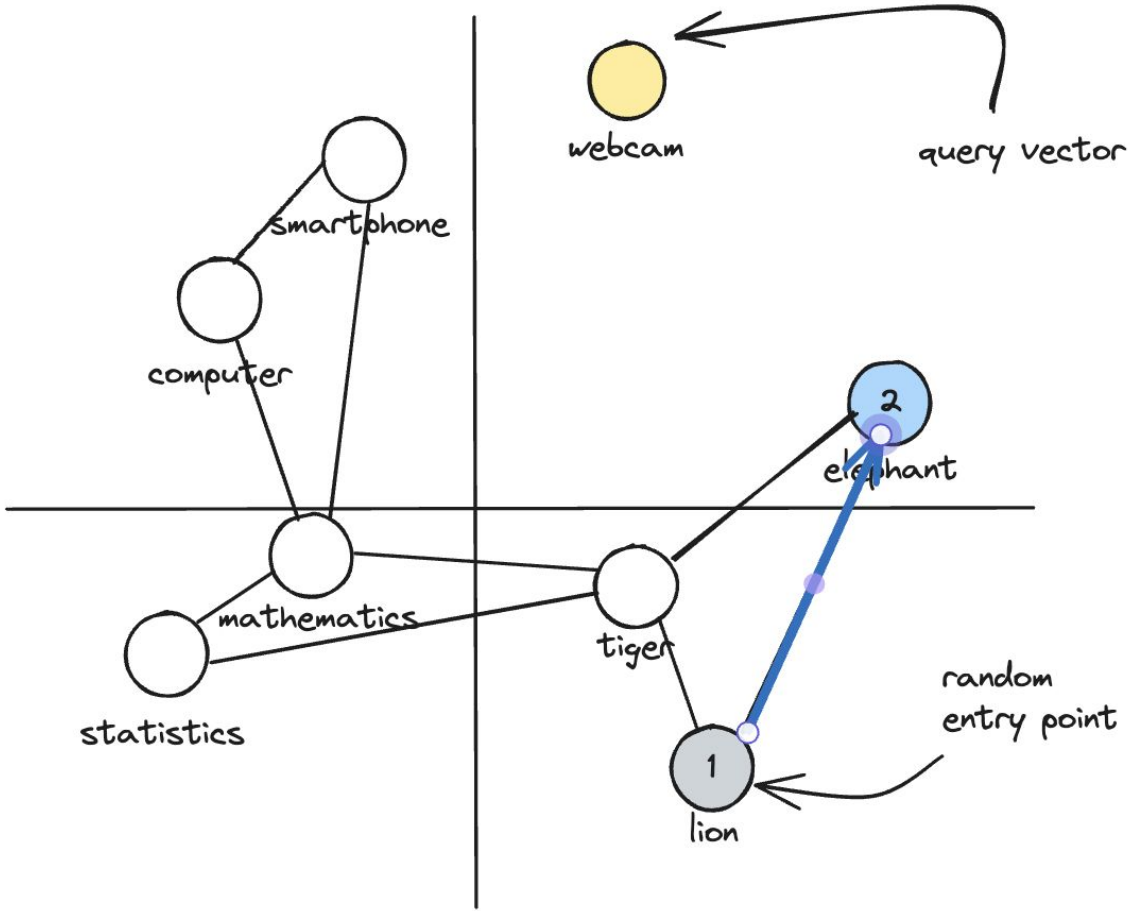

smartphone
computer
elephant
mathematics
tiger
statistics
lion


smartphone
computer
elephant
mathematics
tiger
animal
query vector
statistics
lion


random
entry point
1
smartphone
computer
2
mathematics
statistics
3
tiger
elephant
animal
lion
query vector

Drawbacks

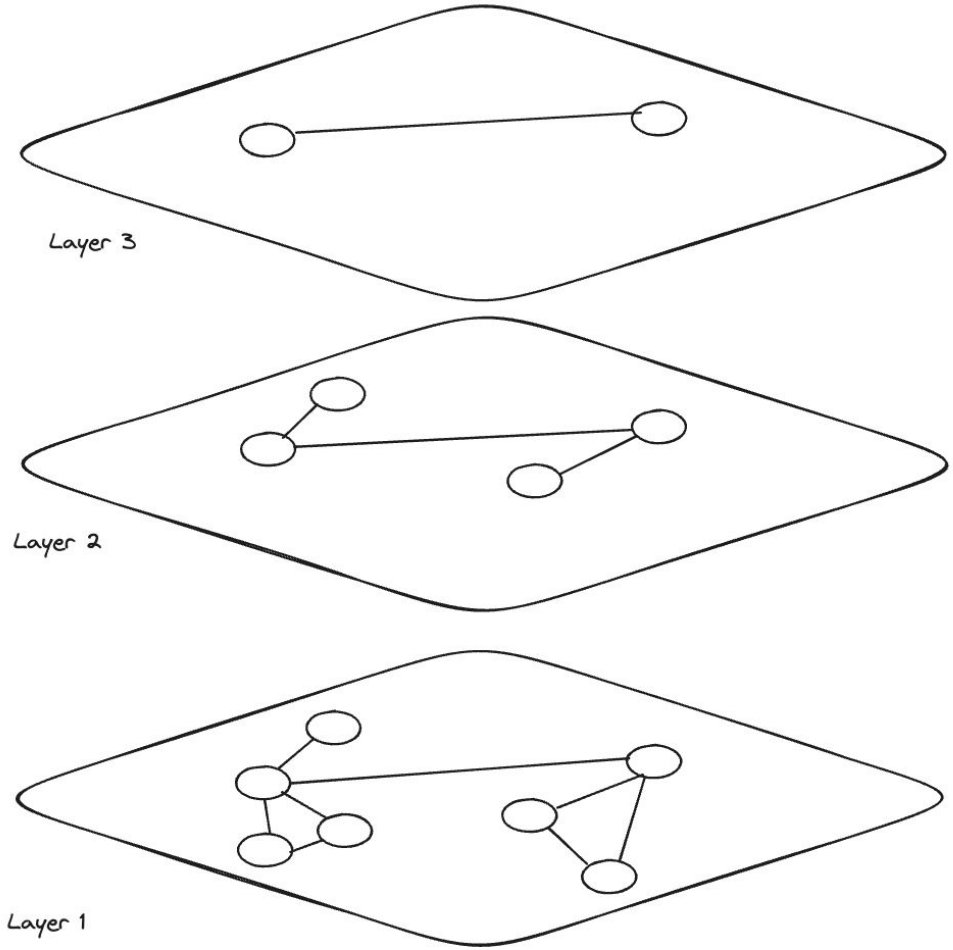
- Can get stuck in local minimum
- Polylogarithmic complexity, not performant at scale
  - Searches for multiple nodes, multiple times

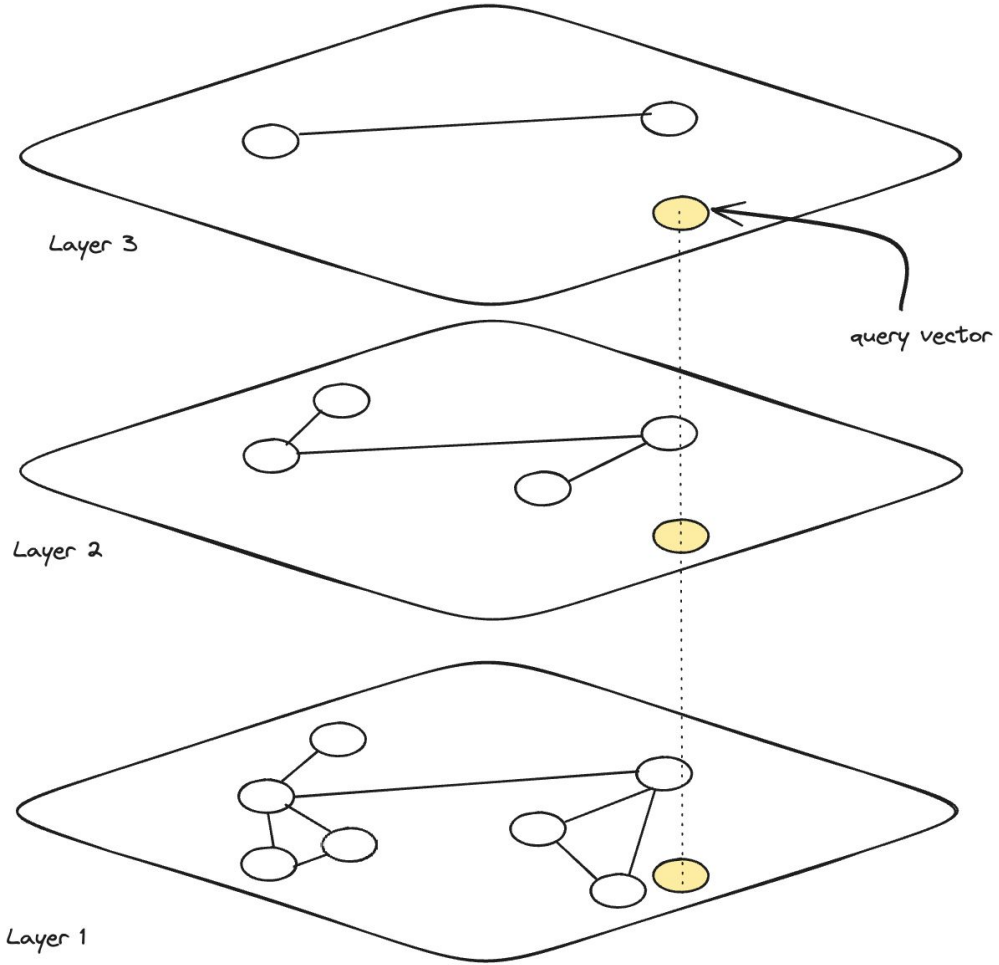

webcam
query vector
smartphone
computer
2
elephant
mathematics
tiger
statistics
random
entry point
1
lion

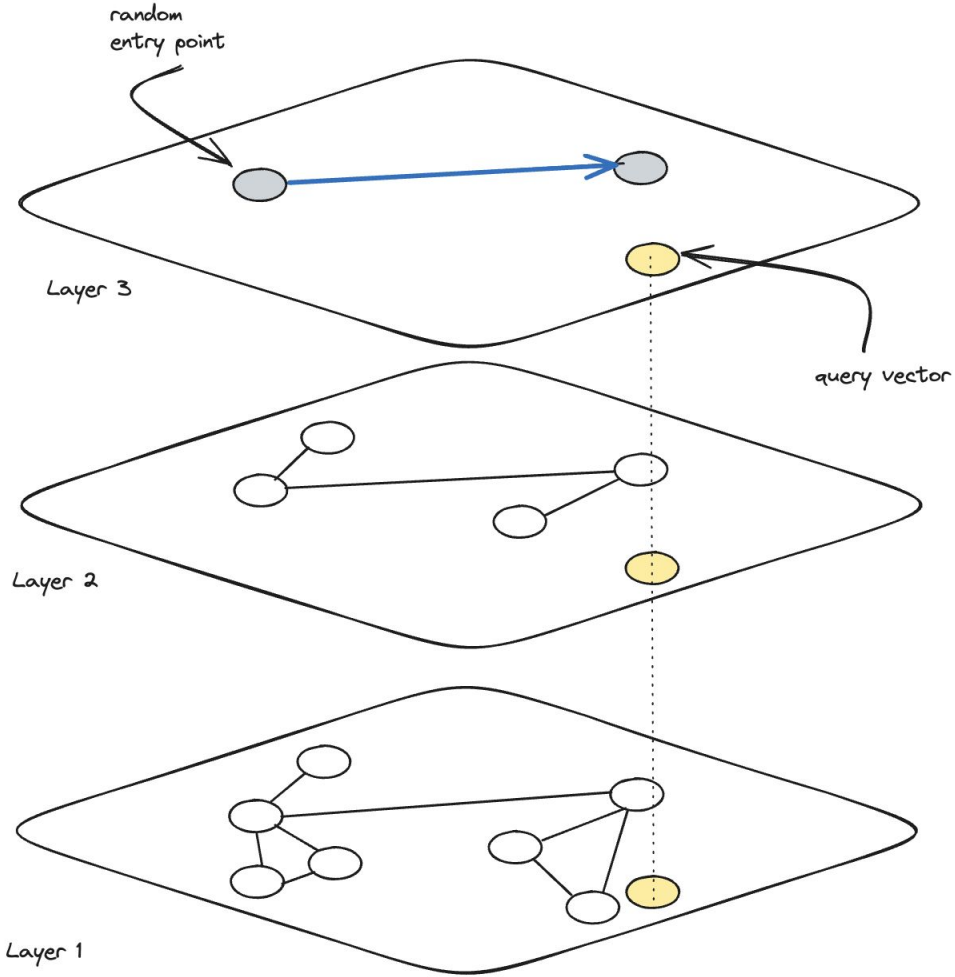
Hierarchical NSW

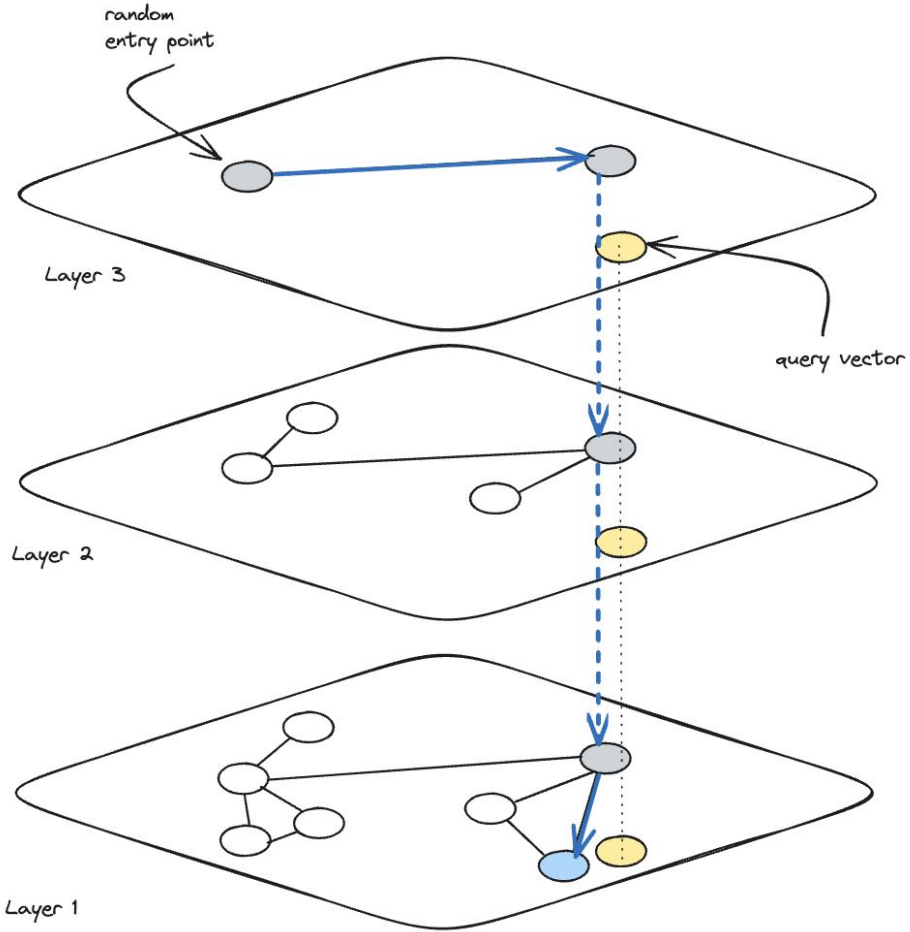
- Multi-layered graph
  - Base layer has all nodes (dense), more connections
  - Higher layer have fewer nodes (sparse) and fewer connections

- Search starts in top layer descending towards bottom layers

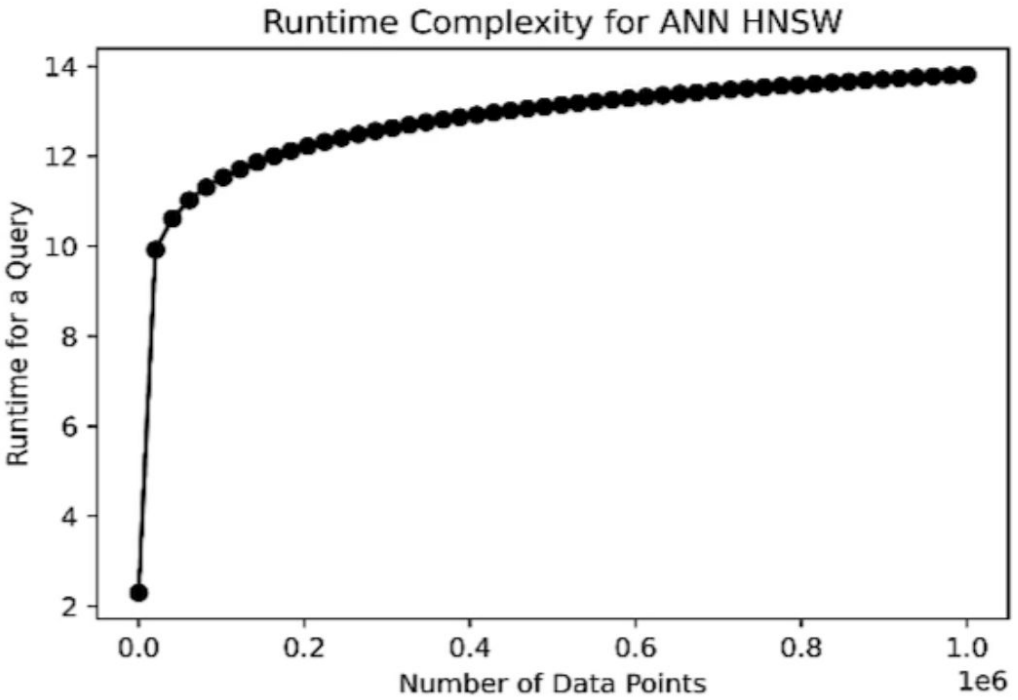

Layer 3
Layer 2
Layer 1


Layer 3
query vector
Layer 2
Layer 1


random
entry point
Layer 3
query vector
Layer 2
Layer 1


random
entry point
Layer 3
query vector
Layer 2
Layer 1

Performance



- Ensures consistent, reliable performance
- Query time increases logarithmically
- Typical query latency ~milliseconds