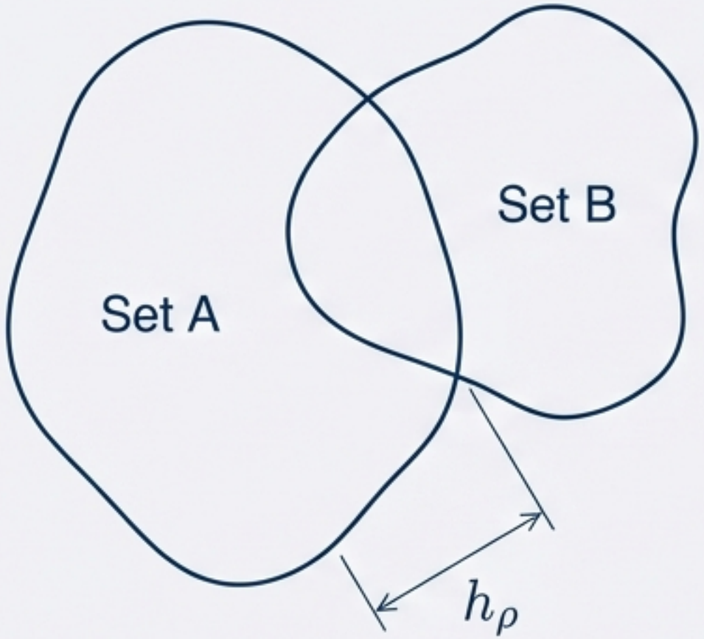
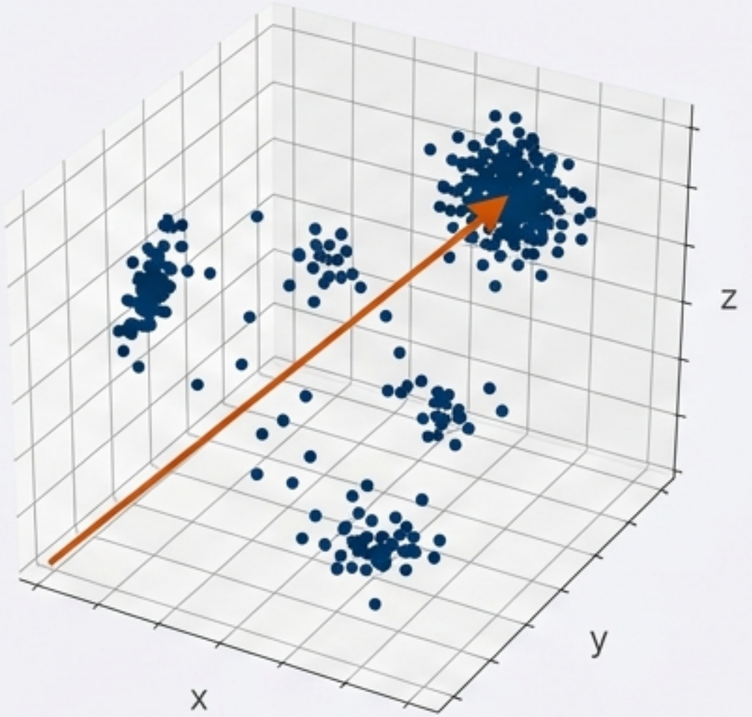


Geometric/Topological Foundation:
Hausdorff distance and set intersection.

The Mathematics of Similarity

From Geometric Axioms to High-Dimensional Vector Search

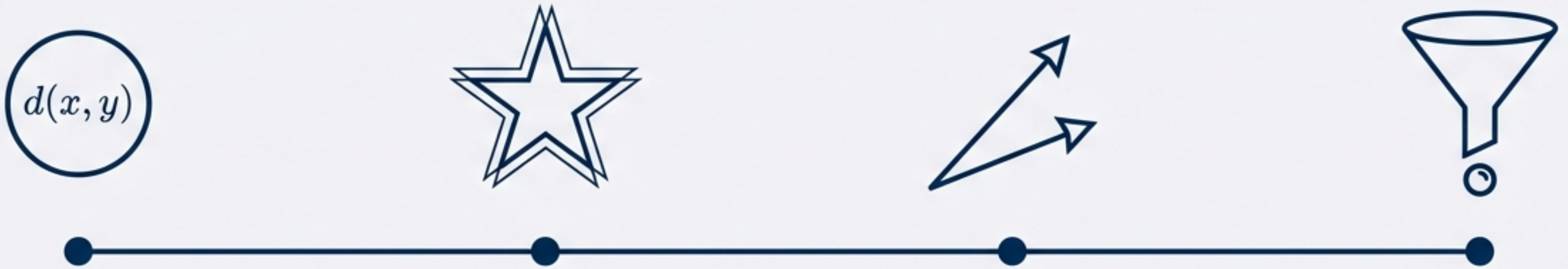


High-Dimensional Semantic Space:
Vector embeddings and cosine similarity.

RESEMBLANCE IS A FUNCTION, NOT A FEELING.

The Spectrum of Resemblance

A structural overview bridging topology, geometry, and engineering.



Foundations

- Metric vs. Pseudometric
- The 4 Axioms

Geometric Pattern

- Hausdorff & Visibility
- Robustness to Noise

Semantic Vectors

- Cosine vs. L2
- Direction > Magnitude

Optimization

- Convexity
- Hull-based Traversal

Narrative Arc: We begin with the mathematical rules, test them against physical shapes, translate them to semantic vectors, and solve for scale.

Defining the Distance Function

The Axioms of Engagement

$$d : S \times S \to \mathbb{R}$$

1. Self-Identity

$$d(x, x) = 0$$

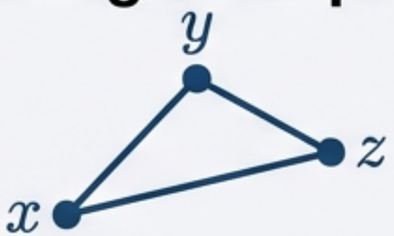
2. Positivity

$$\text{If } x \neq y, \text{ then } d(x, y) > 0$$

3. Symmetry

$$d(x, y) = d(y, x)$$

4. Triangle Inequality



$$d(y, z) \leq d(y, x) + d(x, z)$$

Metric vs. Pseudometric

A Metric satisfies all 4 axioms.

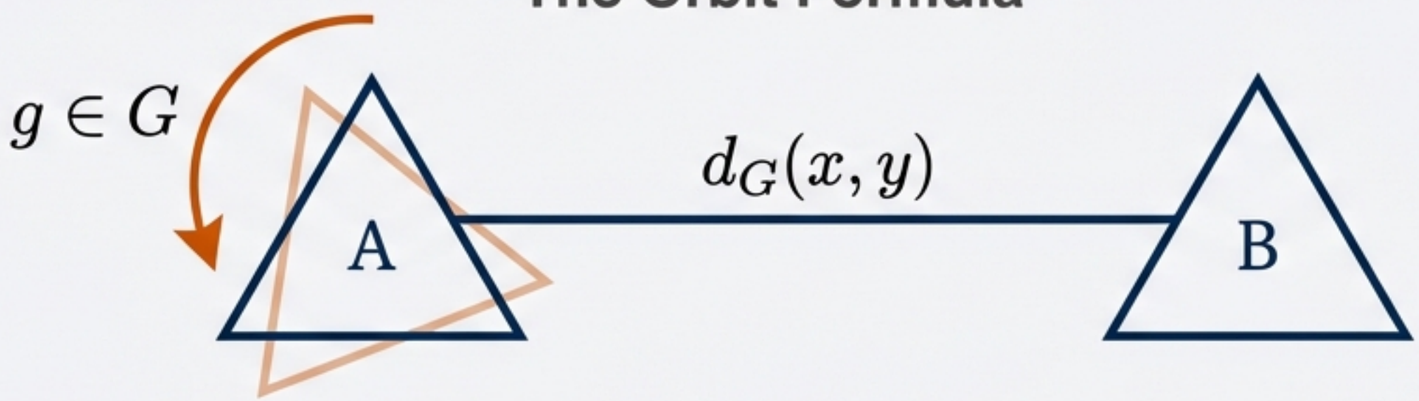
A Pseudometric allows distinct objects to have zero distance distance (d(x,y)=0 even if x≠y).

In semantic search, distinct sentences can be semantically identical, making embedding spaces often behave like pseudometric spaces.

Topology and Invariance

A robust similarity measure must recognize an object regardless of transformation (rotation, shift, or scale). This is the Invariance Principle.

The Orbit Formula

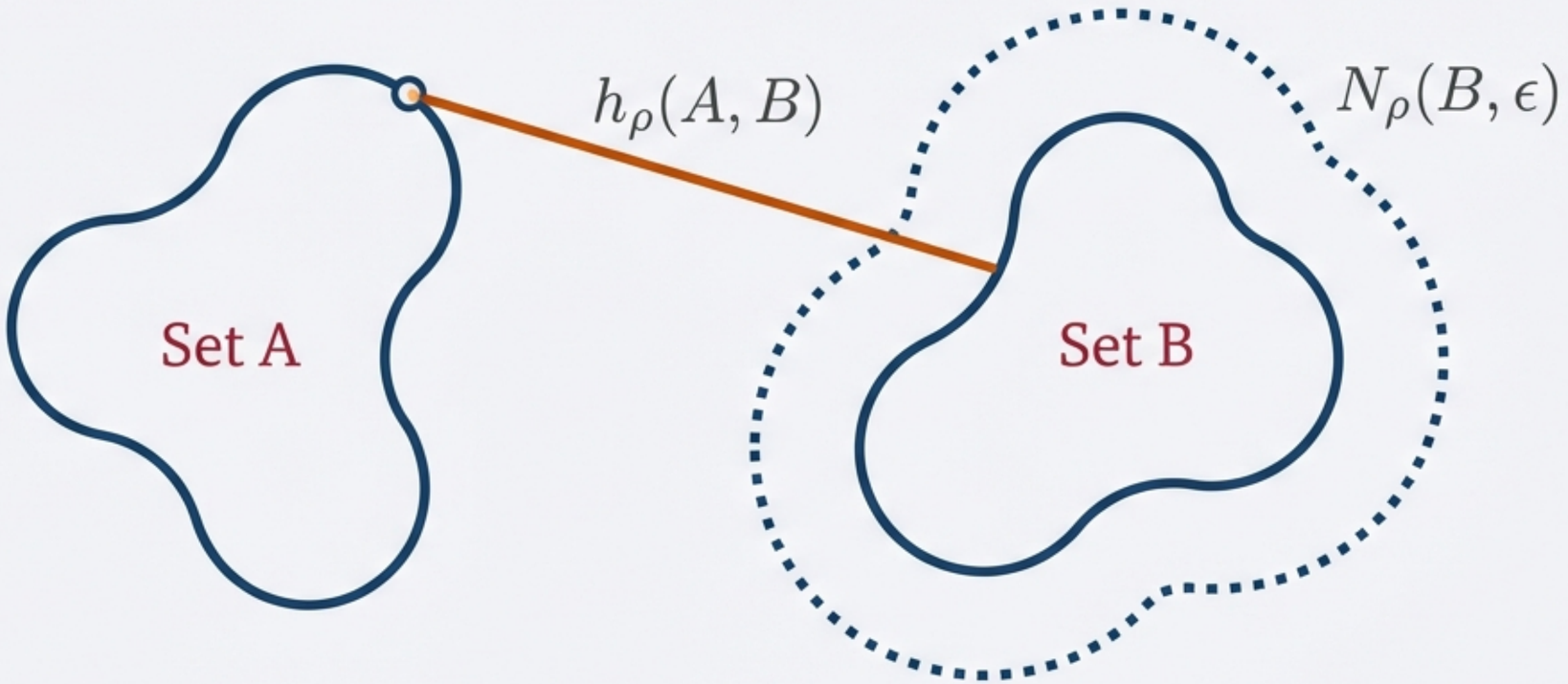


$$d_G(x, y) = \inf \{d(g(x), y) \mid g \in G\}$$

The distance is the minimum (infimum) achieved by rotating the shape to find the best alignment.

Measuring Shape: The Hausdorff Metric

The standard for comparing closed, bounded subsets. It measures the "worst-case" distance between two sets.



$$h_\rho(A, B) = \inf\{\epsilon > 0 \mid A \subseteq N_\rho(B, \epsilon) \text{ and } B \subseteq N_\rho(A, \epsilon)\}$$

The Four Axioms of Robustness

A metric must remain stable under topological distortions.

1. Deformation

2. Blur

3. Crack

4. Noise

Hausdorff is NOT Noise Robust. A single outlier can arbitrarily maximize the distance score.

Beyond Hausdorff: Reflection Visibility

To solve robustness issues, we measure the ‘view’ from inside the shape rather than just the edges. This is robust to noise and cracks.

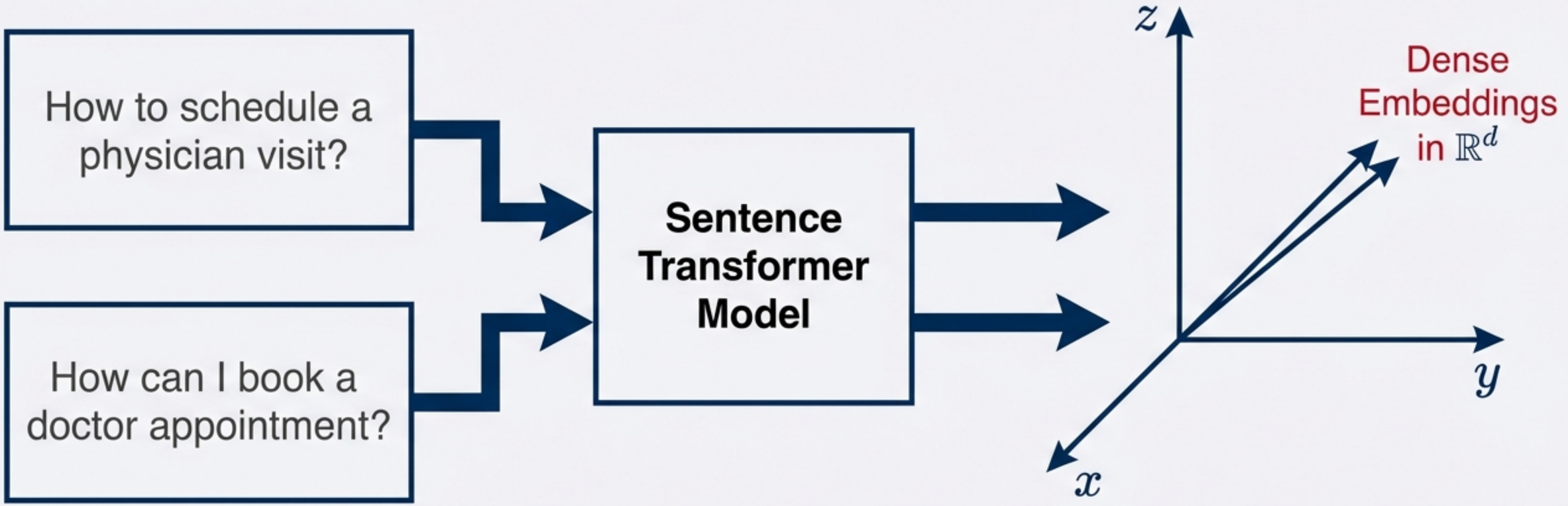
Pattern P

Visibility Star

Reflection Visibility Map

From Geometry to Meaning

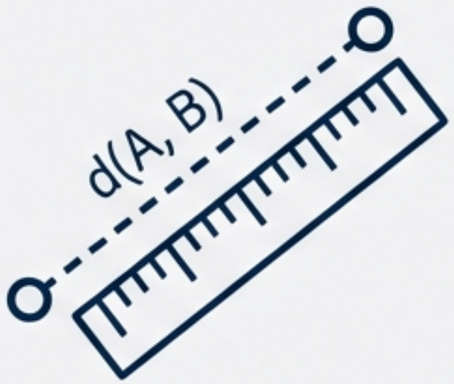
The Semantic Gap



Keywords fail because ‘Physician’ and ‘Doctor’ have zero character overlap. Semantic vectors capture the intent, placing them close in high-dimensional space.

The Metrics of Meaning

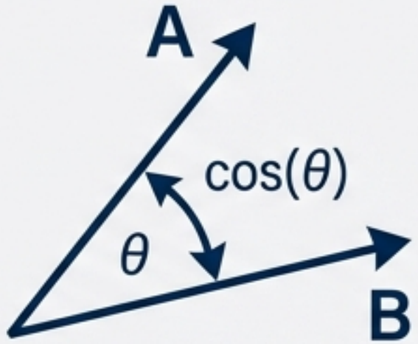
L2 (Euclidean)



Measures absolute distance. Magnitude (length) affects the score.

Use Case:
Drug Discovery (Potency).

Cosine Similarity



Measures orientation/angle. Magnitude is ignored.

Use Case:
Semantic Search (Intent).

Dot Product

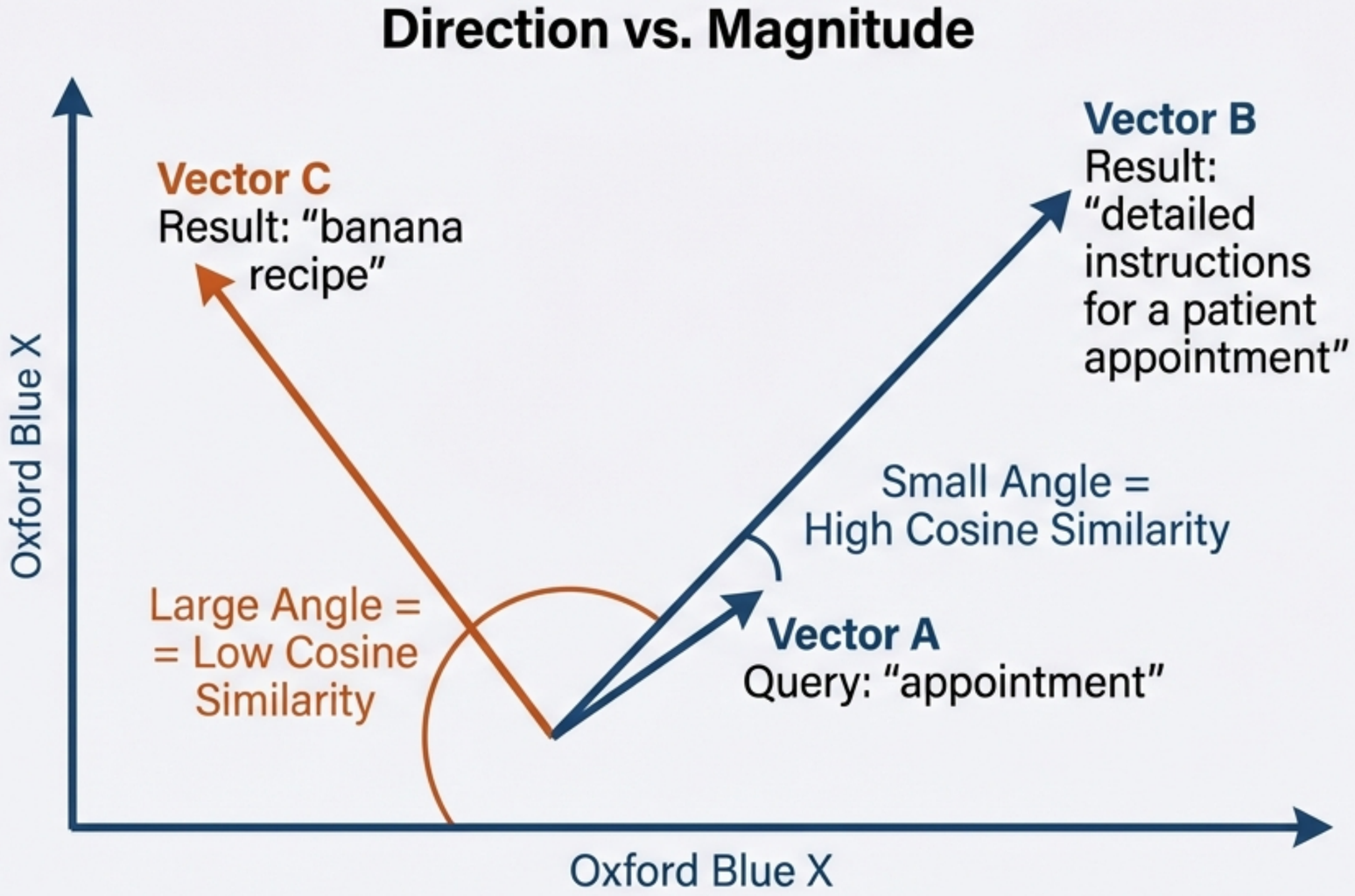
$$\vec{A} \cdot \vec{B}$$

Unnormalized Cosine. Faster to compute.

Use Case:
High-speed Ranking.

For normalized embeddings (unit vectors), Cosine Similarity ranking is identical to Dot Product ranking.

Why Cosine Wins for RAG



In text, "Length" often just means word count. "Direction" is meaning.

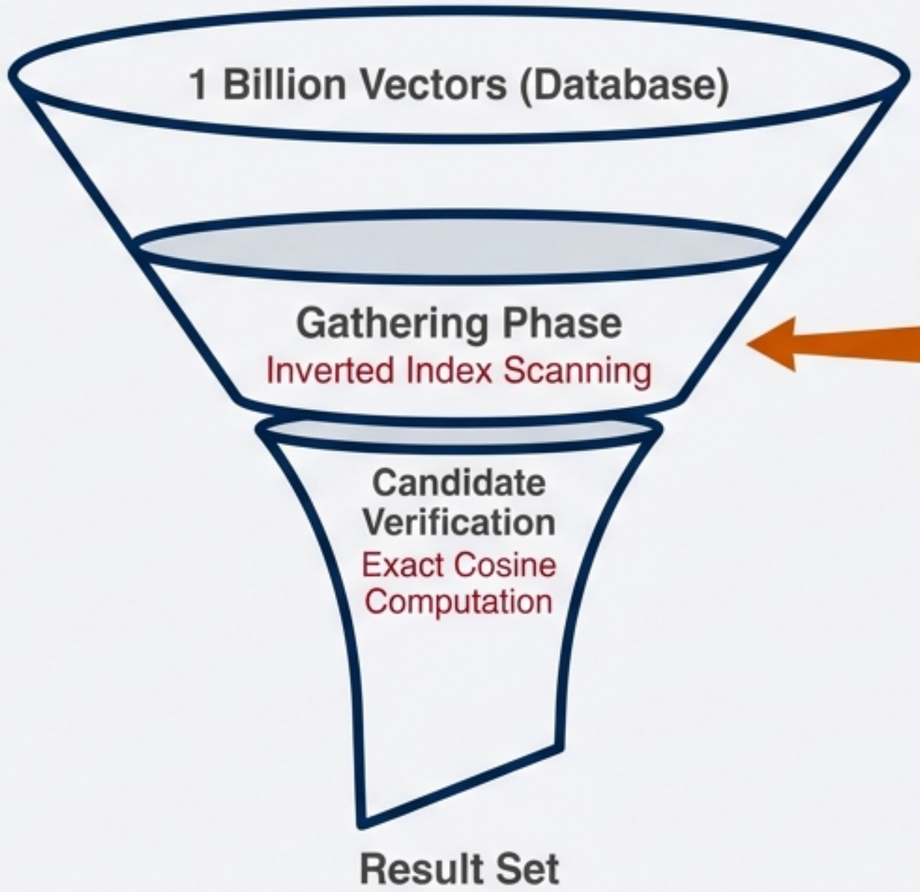
Therefore, we ignore magnitude.

Counter-Example: In Chemistry, two molecules with the same structure (direction) but different binding potency (magnitude) require L2 distance.

The Engineering Challenge: Scale

Cosine Threshold Querying

$$\text{Find all } s \in D \text{ such that } \cos(\boldsymbol{q}, \boldsymbol{s}) \geq \theta$$

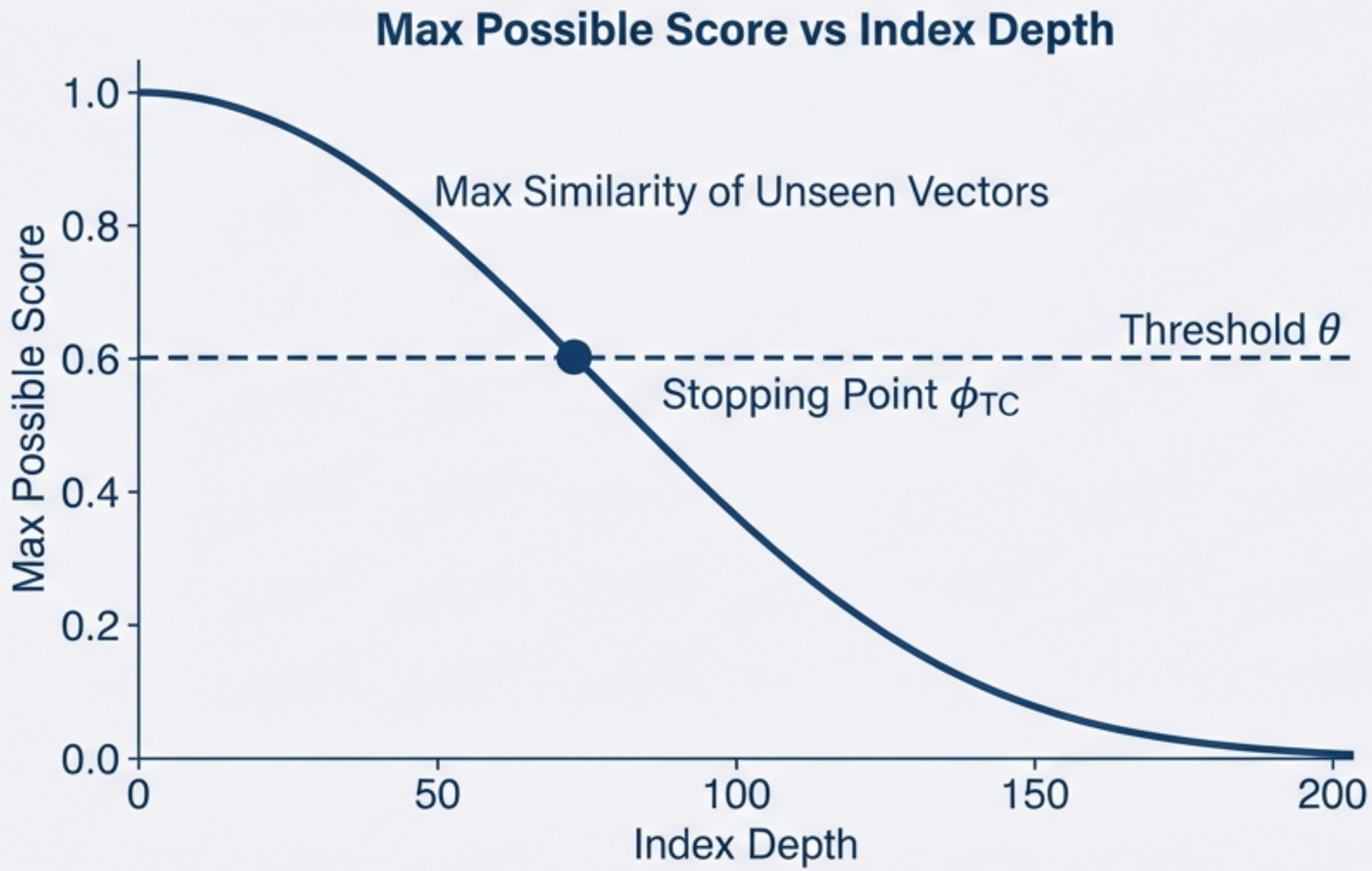


The Bottleneck

We must stop scanning the index as early as possible.

Optimizing the Stopping Condition

Standard algorithms miss a key constraint: all our vectors are normalized (Unit Vectors). By accounting for this, we can stop searching earlier.



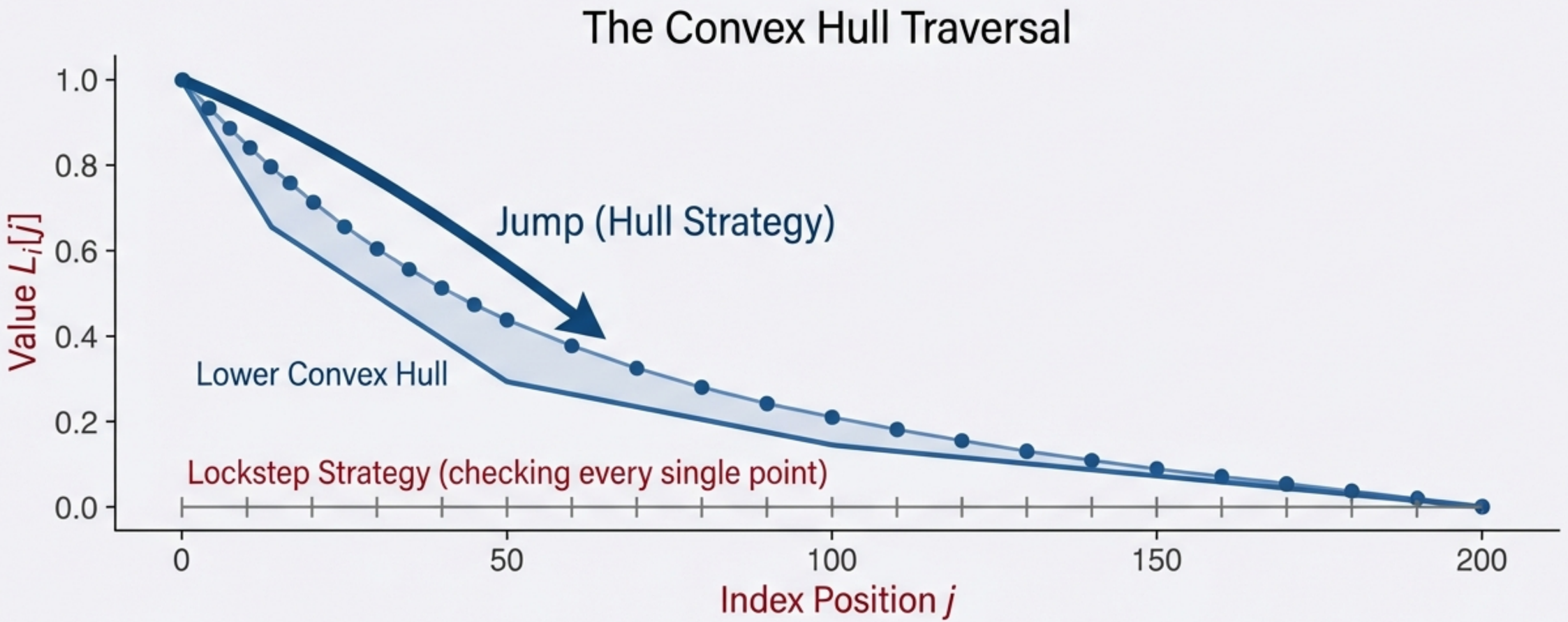
Solve for τ:

$$\sum \min(q_i\tau, L_i[b_i])^2 = 1$$

This condition is Tight (stops immediately when safe) and Complete (misses no results).

Exploiting Data Skew: The Hull Strategy

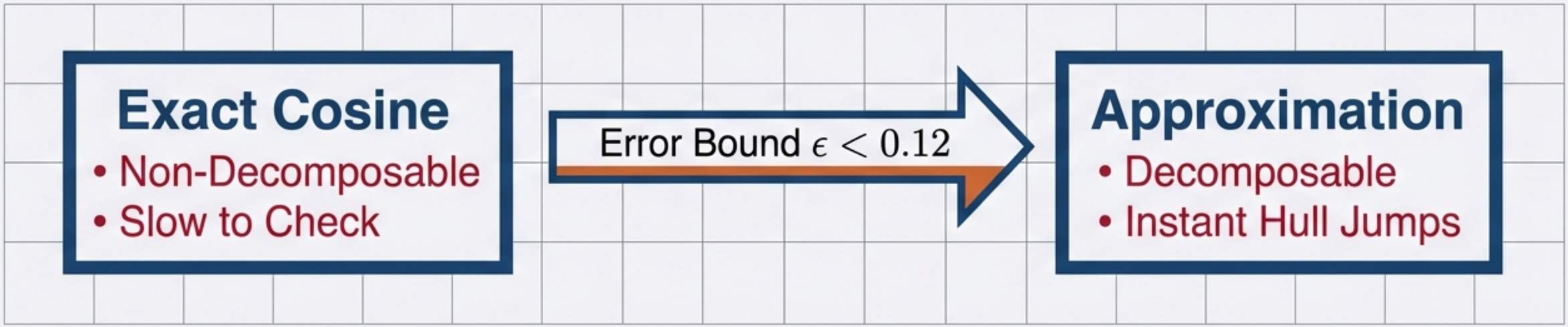
Real-world data is skewed (convex). We can pre-compute this shape to skip redundant checks.



In Mass Spectrometry data, this Hull Strategy reduces overhead to just **~1.3%** of access cost.

Approximating for Speed

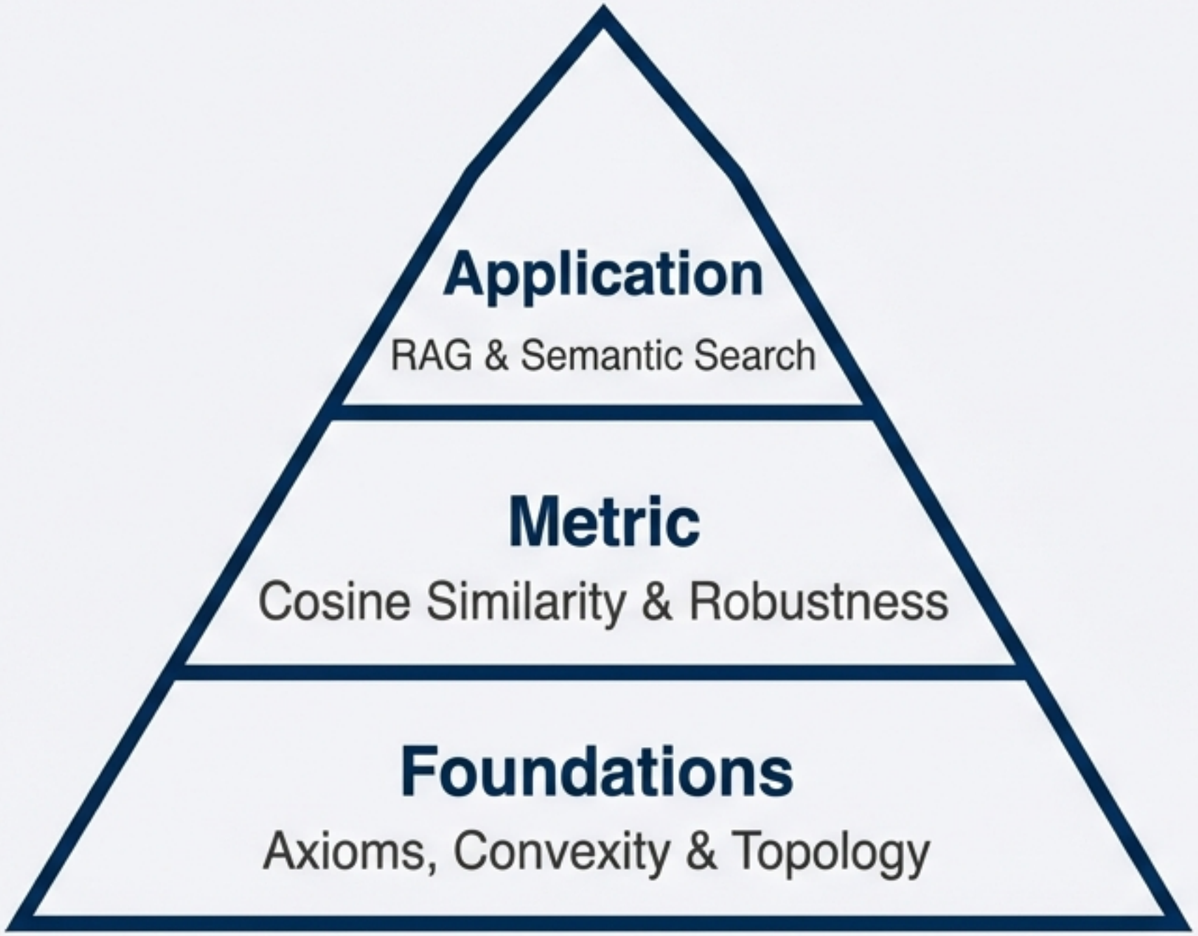
Exact Cosine stopping is computationally expensive. We use a decomposable approximation to maintain speed with negligible error.



$$\text{Cost} \leq \text{OPT}(\theta - \epsilon) + c$$

We achieve near-optimal speed by traversing for a slightly lower threshold $(\theta - \epsilon)$.

The Resemblance Framework



A great search engine is not just about indexing data; it is about choosing the mathematical geometry that best models the intent of the user.

Sources: DSpace (Theory of Similarity), DZone (Sentence Transformers), SAP Community (Cosine vs L2), CS@Purdue (Threshold Querying).