CSE2525 · DATA MINING · 2026–27

Distances & Similarity

Choosing a distance is choosing which patterns an algorithm can see — worked through two case studies.

A. Anand · study deck

IN THIS LECTURE

Distances & similarity — the roadmap

Almost every mining task — classify, cluster, flag outliers, search — reduces to one question: what is close to what? This lecture is about the decision hiding behind it: the distance.

— Design decisions: task → representation → distance → decision rule — and why choosing them *is* choosing the model.

— Metrics: the four axioms, why metric structure helps (pruning), and the pitfalls — cosine and squared-Euclidean are *not* true metrics.

— Representations: vector, set, sequence, graph — each with its own distance (even text can be a dense vector).

— Two real-world case studies: product recommendation (which differences *matter*) and near-duplicate detection (which to *ignore*).

PART I

Why distance underlies data mining

Before any formula: what does a distance actually let us do?

THE ONE IDEA TO HOLD ONTO

Choosing a distance is choosing what's similar

Every task today — classify, cluster, flag outliers, search — reduces to one question: what is close to what? Change the distance and the answer changes, though *no data point moved*.





The query ★ is a **running shoe**; the candidates are a **dress shoe** and **another running shoe**. Its nearest neighbour flips: dress by **price**, running by **purpose**.

"Similar" has **no default meaning** — it depends what you measure:

- Similar in price? — a running shoe lands next to a dress shoe.
- Similar in purpose? — it lands next to other running shoes.
- Similar in description text? — grouped by words, not by use.
- Similar in who buys it? — you get *complements*, not substitutes.

Same two candidates, same query — the nearest neighbour flips with the distance. The task decides which sense you want; the distance is how you encode it.

DISTANCE IS A SHARED BUILDING BLOCK

One dataset — four questions



Mining task	The question, expressed through distance
Classification	Which labelled examples are closest to this new object?
Clustering	Which objects are close enough to form groups?
Anomaly detection	Which objects are far from the usual observations?
Similarity search	Which existing objects are closest to this query?

YOUR TURN · 2 MIN WITH YOUR NEIGHBOUR

Design decision: the “more like this” button

SCENARIO

A music app wants a “more like this” button on every song. You decide what powers it.

— Which of today’s four questions is this — classification, clustering, anomaly, or similarity search?

— What should “similar” mean here — same genre, same tempo & mood, or the same listeners?

— Each answer is a different representation. Which one can you actually compute from the data you have?

There’s no correct answer — only a design decision you can defend. We’ll compare a few.

PART II

Task → representation → distance

The choice of distance is the model — so make the choice deliberately.

FROM INTUITION TO VOCABULARY

Features, and what we mean by a “model”

Feature

- A **measurable property** of an object — its price, its category, the words in its description, who buys it.

Representation

- The features you **choose** to describe an object — its encoding. We formalise this as a **representation space** next.

Model (here)

- **Representation + distance + decision rule.** Choosing the features and the distance *is* choosing the model.

Different features give different neighbours — so the modelling decision begins by choosing features, which we assemble into a representation space next.

BEFORE THE DISTANCE COMES THE ENCODING

What is a *representation space*?

DEFINITION

A **representation** encodes each object as an element of a chosen space $\mathcal{R}$ — a vector, a set, a sequence, a node in a graph. Every later step sees **only** what $\mathcal{R}$ records, so two objects identical in $\mathcal{R}$ are **indistinguishable** to the distance. Choosing $\mathcal{R}$ is the **first** modelling decision.



Vector · $\mathbb{R}^n$

object → a point of numeric features

`Euclidean · cosine`



Set

object → a set of elements (words, shingles)

`Jaccard`

`1 0 1 1 0`

Sequence

object → an ordered string of symbols

`edit · Hamming`



Graph / tree

object → a node in a graph

`shortest path`

YOUR TURN · MATCH THE DATA TO THE SPACE

What is a *representation space*?

QUESTION

You've met four spaces — **vector, set, sequence, graph**. For each, name a kind of data that fits it best. Then the twist: could the *same* object live in more than one?



Vector · $\mathbb{R}^n$

numeric data — sensor readings, pixels



Set

the words in a document



Sequence

a time series · text characters



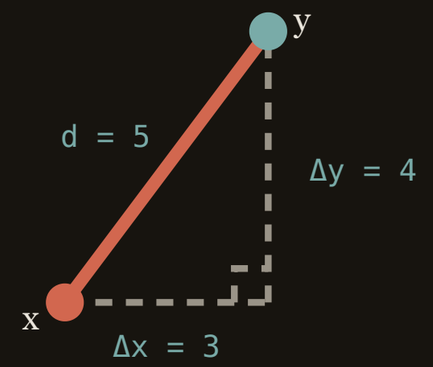
Graph / tree

a social network · the web

The catch: the *same* text is a set of words, a sequence of characters — or a dense vector (embeddings, next lecture). The representation is a choice, not a given.

VECTOR SPACE · THE STRAIGHT-LINE DISTANCE

Euclidean distance



The distance is the **hypotenuse**: $\sqrt{(3^2 + 4^2)} = 5$.

$$d(x, y) = \sqrt{\left(\Sigma_i \, (x_i - y_i)^2 \right)}$$

WORKED EXAMPLE

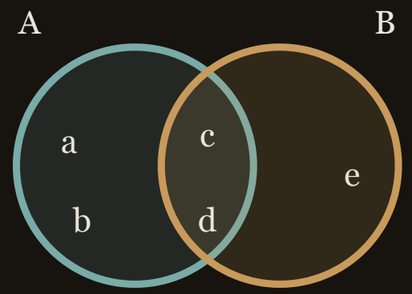
x = (1, 2), y = (4, 6).

d = $\sqrt{((4 - 1)^2 + (6 - 2)^2)} = \sqrt{(9 + 16)} = \sqrt{25} =$ **5**.

Cosine is its sibling for text — it compares *direction*, not magnitude. We'll meet both again: where they're used, and whether they're metrics.

SET · THE OVERLAP OF TWO SETS

Jaccard distance



∩ = {c, d} = 2 · ∪ = 5 → J = 0.4, d = 0.6.

$$J(A, B) = |A \cap B| \,/\, |A \cup B| \qquad d_J = 1 - J(A, B)$$

WORKED EXAMPLE

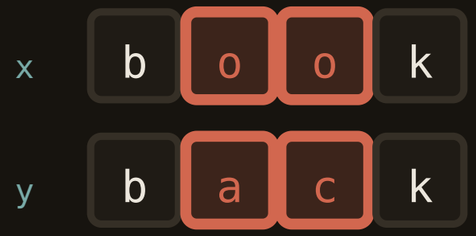
A = {a, b, c, d}, B = {c, d, e}.

A ∩ B = {c, d} → 2, A ∪ B = {a, b, c, d, e} → 5. J = 2/5 = 0.4, d_J = **0.6**.

Natural for documents seen as sets of words or shingles. We'll meet it again — in near-duplicate detection — and ask whether it's a metric.

SEQUENCE · HOW MANY EDITS APART

Edit (Levenshtein) distance



2 letters change

`book` → `back`: two substitutions → **2 edits**.

d(x, y) = fewest insertions, deletions & substitutions turning x into y

WORKED EXAMPLE

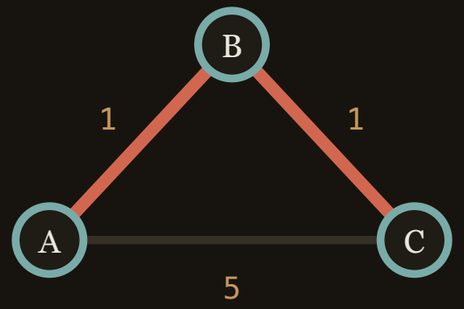
`book` → `back`: b·b ✓, o → a, o → c, k·k ✓ → **2** substitutions.

Equal length, substitutions only → this coincides with the **Hamming** distance. Insertions & deletions let it compare *different* lengths: `book` → `books` = 1.

We'll meet Hamming again in Part III — where we prove it *is* a metric — and in the video case study.

GRAPH · THE FEWEST STEPS BETWEEN NODES

Shortest-path (geodesic) distance



Direct A–C = 5, but A–B–C = 1 + 1 = **2** wins.

$$d(u, v) = \min \text{ over paths } u \to v \text{ of } \Sigma \text{ weight(edges)}$$

WORKED EXAMPLE

Edges: A–B = 1, B–C = 1, and a direct A–C = 5.

Paths A → C: direct = 5, via B = 1 + 1 = 2. So d(A, C) = **2** — the shortest path wins.

Unweighted graphs → just count edges. We'll work this out in full in Part III, and again in graph mining.

FOUR DECISIONS, IN ORDER

Task → representation → distance → decision

Choice	What it determines
Task	What “similar” should mean
Representation	What information is available for comparison
Distance	Which differences matter, and how much
Decision rule	What we do with the resulting distances (e.g. top-k)

And you don’t always need a vector first — trees, strings and sets have distances directly.

YOUR TURN · WALK THE SEQUENCE

Design decision: spot duplicate news articles

SCENARIO

A news aggregator scrapes the same story from dozens of sites, each with small edits. You must group the duplicates — fill in the four decisions.

Decision	Your call for this task
Task	What should “similar” mean — near-identical wording?
Representation	A set of words/*k*-grams? a vector of the text? or the whole string (a sequence)?
Distance	Each representation brings its own: Jaccard (set) · cosine (vector) · edit distance (string).
Decision rule	Group two articles when the distance is below which threshold?

Notice the dependency: the representation you pick decides which distances are even available to you.

PART III

Metrics as useful guarantees

What properties make a distance predictable — and fast?

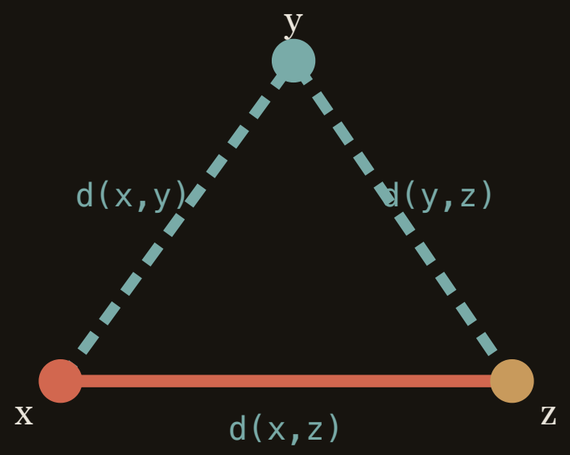
DEFINITION

When is a distance a *metric*?

METRIC SPACE

A function $d : X \times X \to \mathbb{R}$ is a **metric** on a set X if, for every $x, y, z \in X$, it satisfies the four axioms below. The pair **(X, d)** is then called a **metric space**.

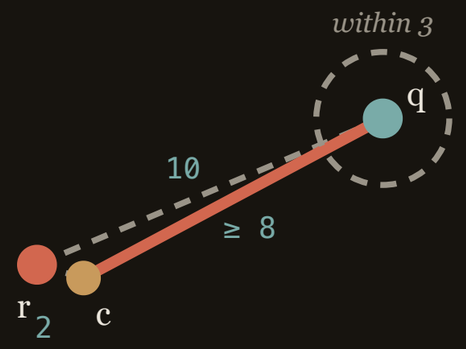
Axiom	Formal statement
Non-negativity	$d(x, y) \geq 0$
Identity of indiscernibles	$d(x, y) = 0 \iff x = y$
Symmetry	$d(x, y) = d(y, x)$
Triangle inequality	$d(x, z) \leq d(x, y) + d(y, z)$



Triangle inequality: the detour **x→y→z** is never shorter than going **direct**.

WHY THESE AXIOMS ARE WORTH DEMANDING

The axioms are guarantees an algorithm can exploit



c is ≥ 8 from **q**, but we only want neighbours within 3 — so **c** is skipped, unmeasured.

— Identity and symmetry make “the nearest object” well-defined and independent of comparison order.

— The triangle inequality lets known distances bound unknown ones, so candidates can be ruled out *without* measuring them.

PRUNING WITH THE TRIANGLE INEQUALITY

q is at distance 10 from a reference *r*; a candidate *c* is at distance 2 from *r*. The triangle inequality forces a **lower bound**: `d(q, c) ≥ |10 − 2| = 8`. Since **8 > 3**, *c* cannot be within 3 of *q* — discard it **without ever computing d(q, c)**.

This powers metric indexes — ball trees, VP-trees — that avoid the full $O(n^2)$ comparisons.

EXERCISE · CHECK THE AXIOMS

Is the Hamming distance a metric?



2 positions differ, so the distance = 2

CANDIDATE DISTANCE

For strings x, y of equal length n, let $d_H(x, y)$ be the number of positions at which they differ.

— Work through each of the four axioms for d_H — does it hold?

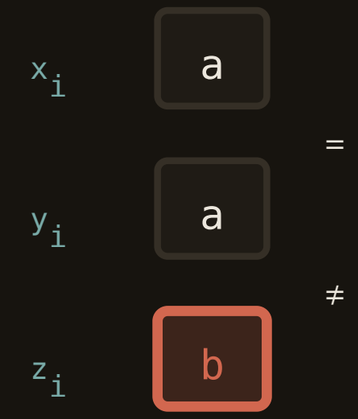
— For the ones that hold, give the reason; for any that fails, a counterexample.

Decide for yourself: is d_H a metric? We'll work it through together on the next slide.

Metric axioms to check: non-negativity $d(x,y) \geq 0$ · identity $d(x,y)=0 \Leftrightarrow x=y$ · symmetry $d(x,y)=d(y,x)$ · triangle $d(x,z) \leq d(x,y)+d(y,z)$

SOLUTION

Yes — the Hamming distance is a metric



At one position: $x_i = y_i$ but $y_i \neq z_i$, so $x_i \neq z_i$ — at least one pair always differs.

PROOF — VERIFY EACH AXIOM

Non-negativity. d_H is a count of positions, hence $d_H(x, y) \geq 0$.

Identity. $d_H(x, y) = 0 \iff$ no position differs $\iff x = y$.

Symmetry. "position i differs" is the same statement for (x, y) and (y, x), so $d_H(x, y) = d_H(y, x)$.

Triangle inequality. Fix a position i. If $x_i \neq z_i$, then x_i and z_i cannot both equal y_i, so $x_i \neq y_i$ *or* $y_i \neq z_i$. Hence at each position $[x_i \neq z_i] \leq [x_i \neq y_i] + [y_i \neq z_i]$. Summing over all n positions gives $d_H(x, z) \leq d_H(x, y) + d_H(y, z)$.

∎

At each index the distance is at most 1 (0 if equal, 1 if different), and $[x_i \neq z_i] \leq [x_i \neq y_i] + [y_i \neq z_i]$ holds index by index — so summing over all indices, the total $d(x,z) \leq d(x,y) + d(y,z)$.

EXERCISE · FIND THE BROKEN AXIOM

Is squared Euclidean distance a metric?

CANDIDATE DISTANCE

On $\mathbb{R}^n$, define $d_{sq}(x, y) = \|x - y\|^2 = \Sigma_i (x_i - y_i)^2$ — the **squared** Euclidean distance (the quantity k-means minimises).

— Three of the four axioms clearly hold. Which one can fail?

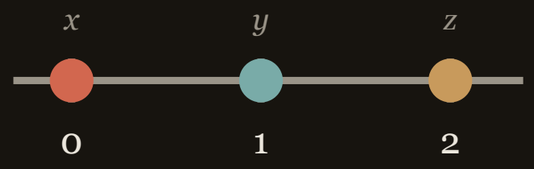
— Try to build a counterexample that breaks it.

Work on it first — the next slide gives a hint, and the one after that the full solution.

Metric axioms to check: non-negativity $d(x,y) \geq 0$ · identity $d(x,y)=0 \Leftrightarrow x=y$ · symmetry $d(x,y)=d(y,x)$ · triangle $d(x,z) \leq d(x,y)+d(y,z)$

HINT

Try three points on a line



Compare the **direct** jump x→z against the **two-step** path x→y→z.

HINT

Put three points $x = 0,\ y = 1,\ z = 2$ on the real line.

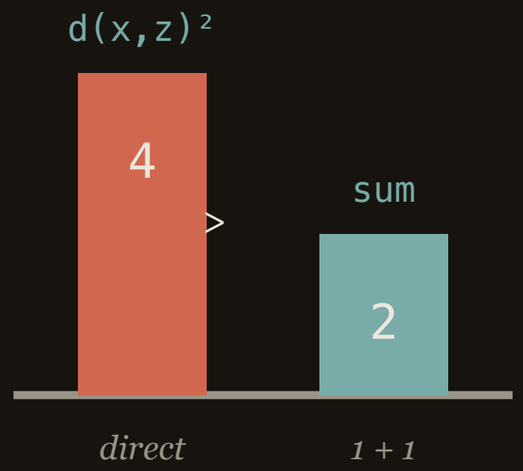
Work out the direct squared distance $d_{sq}(x,\ z)$ and the two-step sum $d_{sq}(x,\ y) + d_{sq}(y,\ z)$ — then compare them.

Which of the four axioms does that comparison test?

Metric axioms to check: non-negativity $d(x,y) \geq 0$ · identity $d(x,y)=0 \Leftrightarrow x=y$ · symmetry $d(x,y)=d(y,x)$ · triangle $d(x,z) \leq d(x,y)+d(y,z)$

SOLUTION

No — the triangle inequality fails



Direct (**4**) > detour sum (**2**) — so **d(x,z) ≤ d(x,y)+d(y,z)** is broken.

COUNTEREXAMPLE ON THE REAL LINE

Take $x = 0$, $y = 1$, $z = 2$. Then

$d_{sq}(x, z) = (2 - 0)^2 = \mathbf{4}$, while $d_{sq}(x, y) + d_{sq}(y, z) = 1^2 + 1^2 = \mathbf{2}$.

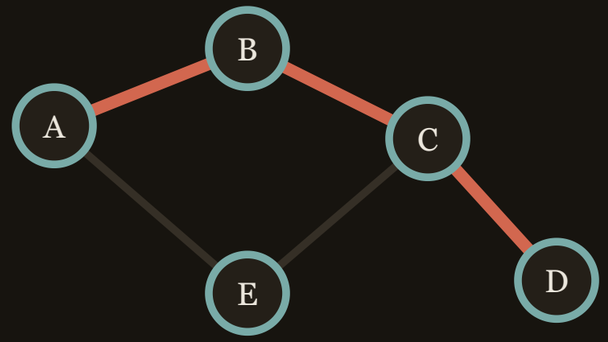
So $4 > 2$ — the triangle inequality is violated.

TAKEAWAY

Non-negativity, identity and symmetry all hold, yet d_{sq} is not a metric. Squaring breaks the triangle inequality — which is why the Euclidean distance (its square root) is used wherever metric structure is required.

A METRIC ON OBJECTS THAT ARE NOT VECTORS

Worked example: shortest-path distance on a graph



DEFINITION

In a connected graph, `d(u, v)` is the number of edges on a **shortest path** between u and v (the **geodesic** distance; for weighted graphs, the smallest total edge weight).

— BFS from A reads the distance straight off the layer number:

— `layer 0 = {A},  1 = {B, E},  2 = {C},  3 = {D}`.

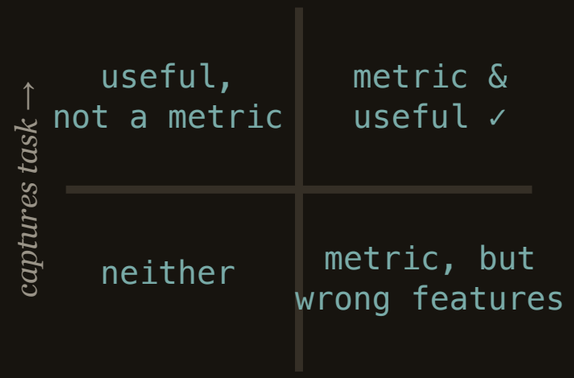
— So `d(A,D) = 3`, along A–B–C–D *or* A–E–C–D.

— Triangle check: `d(A,D) = 3 ≤ d(A,C) + d(C,D) = 2 + 1`. ✓

It is a metric: any path through a third node w is itself a valid $u \to v$ path, so the shortest cannot be longer — the triangle inequality holds by construction.

KEEP THESE SEPARATE

Valid metric ≠ useful distance



The two questions are **independent** — every box can be occupied.

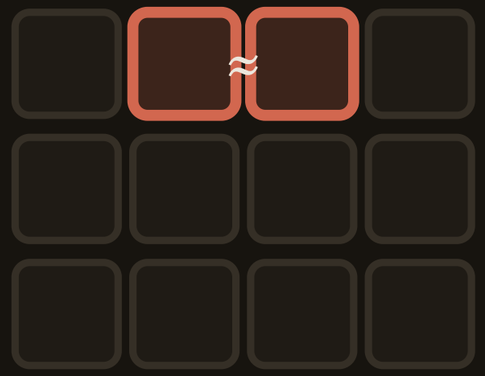
ASK BOTH, INDEPENDENTLY

1 · Is it a mathematically valid metric? 2 · Does it capture the similarity our task needs?

Neither answer settles the other. A valid metric can be useless for a task; a genuinely useful dissimilarity need not satisfy every axiom. We'll see both in the case studies.

YOUR TURN · DO YOU NEED A METRIC?

Design decision: 10 million face photos



Among millions of photos, find the **near-duplicate** pairs.

SCENARIO

You must find near-duplicate faces in a 10-million-photo archive — far too many to compare every pair.

— To prune candidates without computing every distance, which axiom must your distance obey — and why?

— Your score gives d(A,B) ≠ d(B,A). Which axiom breaks, and does it actually matter here?

— Keep two questions apart: is your score a valid metric, and does it capture "same person"?

Metric axioms to check: non-negativity `d(x,y) ≥ 0` · identity `d(x,y)=0 ⇔ x=y` · symmetry `d(x,y)=d(y,x)` · triangle `d(x,z) ≤ d(x,y)+d(y,z)`

ANSWERS · DO YOU NEED A METRIC?

10 million face photos — the calls

Question	The call
Prune without measuring every pair?	Needs the triangle inequality — known distances bound unknown ones, so far-away candidates are skipped unmeasured.
`d(A,B) ≠ d(B,A)`?	Symmetry breaks. “A is a duplicate of B” should equal the reverse — so symmetrise the score (e.g. average the two).
Valid metric vs. captures “same person”?	Two separate questions. A metric buys speed (indexing); only a task-fitted score buys correctness. You want both.

The point of this part: metric properties aren’t pedantry — the triangle inequality is what makes the search feasible at all.

PART IV · CASE STUDY 1

Products: which differences matter

Designing a similarity, as a sequence of decisions.

"I WANT TO SUGGEST SIMILAR ITEMS"

First decide what relevance means



What makes a product a useful recommendation for another — shared category? co-purchase? "related items" / clicks vs. purchases define the target before any distance is chosen.

WHAT INFORMATION IS AVAILABLE

Four representations of a product

Category

- Where it sits in the catalogue tree

Text

- Title & description terms

Price

- A single numeric feature

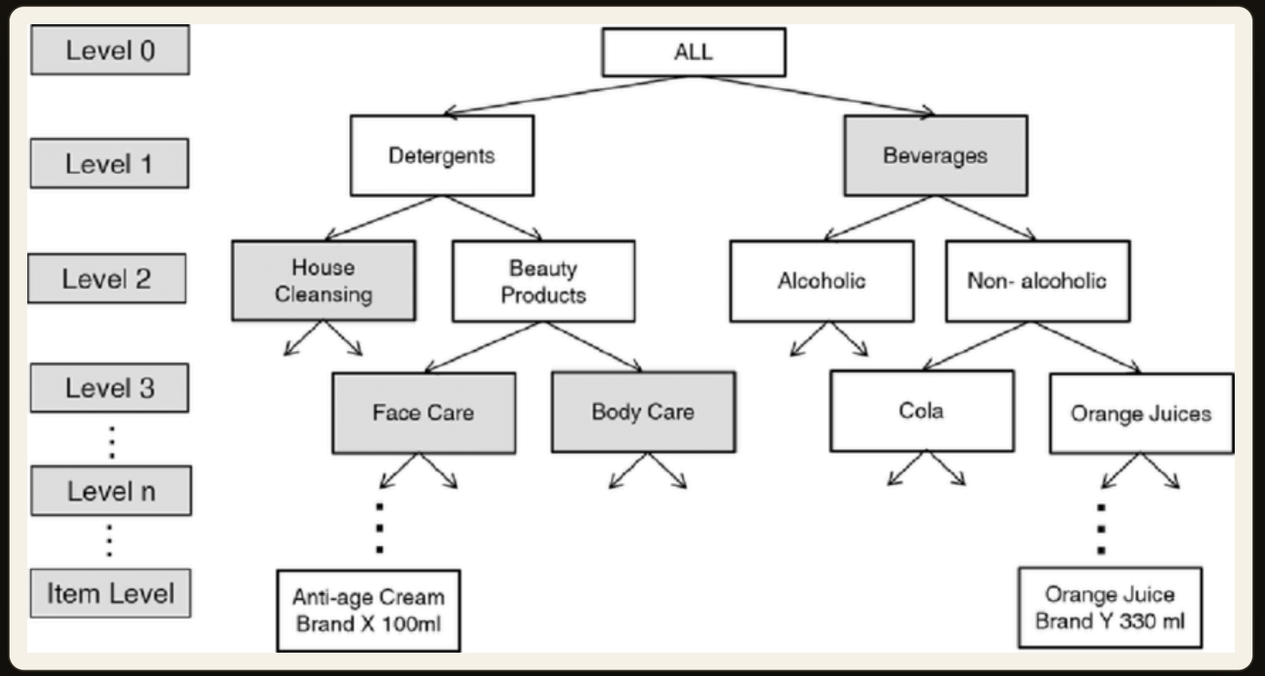
Reviews

- Ratings & their distribution

Each representation needs its own distance — and each will disagree about who the neighbours are.

CATEGORY → A TREE

Taxonomy distance = shortest path

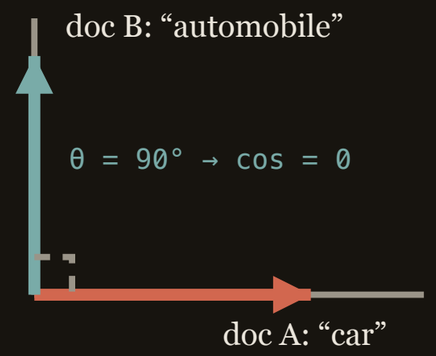


Distance = path length up to a common ancestor and back down. A tree is a graph, so this is exactly the shortest-path distance from Part III — and it is a metric on category nodes. But distinct products in the same category have distance 0, so the representation, not the metric, is the limit.

TEXT → DIRECTION OF A TERM VECTOR

Cosine: similar wording, any length

$$\cos(\theta) = (a \cdot b) / (\|a\| \, \|b\|) \quad d_{cos} = 1 - \cos(\theta)$$



Same meaning, **different words** → no shared terms → the vectors are **orthogonal**, cosine similarity **0**.

READ IT CAREFULLY

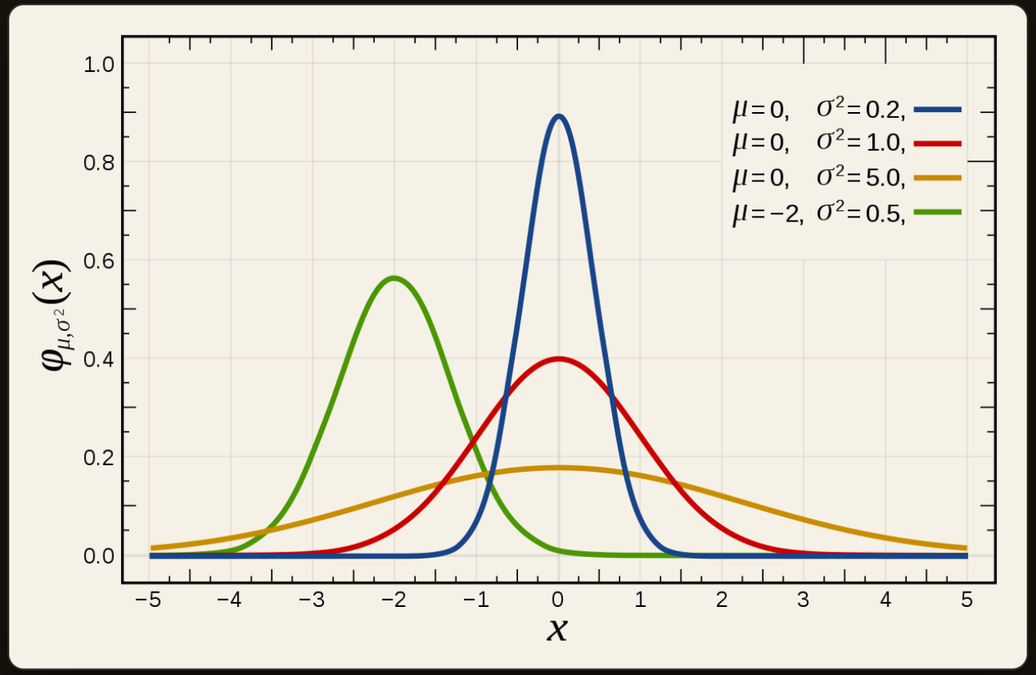
Only the direction of the term vectors matters, never their length — good for descriptions of different sizes.

Blind to synonyms: “car” and “automobile” share no terms, so cosine sees them as unrelated even though the docs mean the same thing. Dense embeddings (next lecture) fix this.

Also: distance 0 means same direction, not identical; and $1 - \cos\theta$ is not a true metric.

PRICE → EUCLIDEAN, BUT BEWARE UNITS

Standardize — then don’t over-trust it



Raw price dominates small-scale features, so standardize: $z = (x - \mu)/\sigma$. But equal scale ≠ equal relevance, and it does not remove outlier sensitivity — scaling is not a priority decision.

REVIEWS → SPARSE & COLD-START

What about products with no history?



Most products have few reviews; brand-new ones have none (cold start). The distance must degrade gracefully when a representation is missing or thin.

REVIEWS → A DISTRIBUTION OVER RATINGS

A distance over customer reviews



Each item → its **rating distribution**; the distance compares the two histograms.

REPRESENTATION + DISTANCE

Encode each item by its **rating distribution** $p = (p_1, \ldots, p_5)$ — the fraction of reviews at each star level. Then **d(A, B)** is a distance between the two histograms.

— Average rating alone is too lossy — very different distributions can share a mean.

— L1 / Euclidean on the histograms is simple, but treats every star as equally far apart.

— Earth Mover's (Wasserstein) respects order — 4★ vs 5★ is closer than 1★ vs 5★.

Two items have "similar reviews" when their distributions match — and few-review (cold-start) items need smoothing toward a prior.

QUESTION

Your query is a **running shoe**. Ranked by **price**, the single closest product is a formal **dress shoe**. Is that a good recommendation — and what decides the answer?

Same three candidates, three different distances.

Think — then → for the answer.

SEE THE CHOICE CHANGE THE ANSWER

One query, three candidates, three rankings

QUERY

a **running shoe** · €120 · *Footwear › Running* · “lightweight breathable running shoe”.

Candidate	By price	By category	By text
A · trail running shoe (€130)	near (Δ10)	nearest — same “Running”	near — “running shoe”
B · running socks (€15)	far (Δ105)	far — Apparel › Socks	near — “breathable lightweight running”
C · formal dress shoe (€125)	nearest (Δ5)	mid — shares “Footwear”	far — only “shoe”

Price crowns a dress shoe; text pulls in socks; only the right representation ranks the other running shoe first. The distance is the model.

FROM OBJECTIVE TO ALGORITHM

Combine distances — weights are priorities

$$d(x,y) = w_c \cdot d_{category} + w_t \cdot d_{text} + w_p \cdot d_{price}$$

TWO CAUTIONS, LINKED

The terms must be on comparable scales before you add them — exactly the standardization point again.

The weights w encode priorities; they are the model's hyperparameters. Standardization does not choose them for you — you decide what should matter. Then rank the top-k neighbours and evaluate against real clicks/purchases.

PART V · CASE STUDY 2

Videos: which differences to ignore

Invariance, thresholds, and a system that scales.

YOUTUBE: NEAR-DUPLICATES OF AN UPLOAD

Equivalent for our task, different in the bytes



Figure 1. Examples of queries and near-duplicate videos from CC_WEB_VIDEO dataset.

What should be ignored: re-encoding, modest colour changes, logos/overlays, small crops. The job is the opposite of case 1 — choosing differences to discard.

BYTE EQUALITY IS THE WRONG TOOL

Exact ≠ near

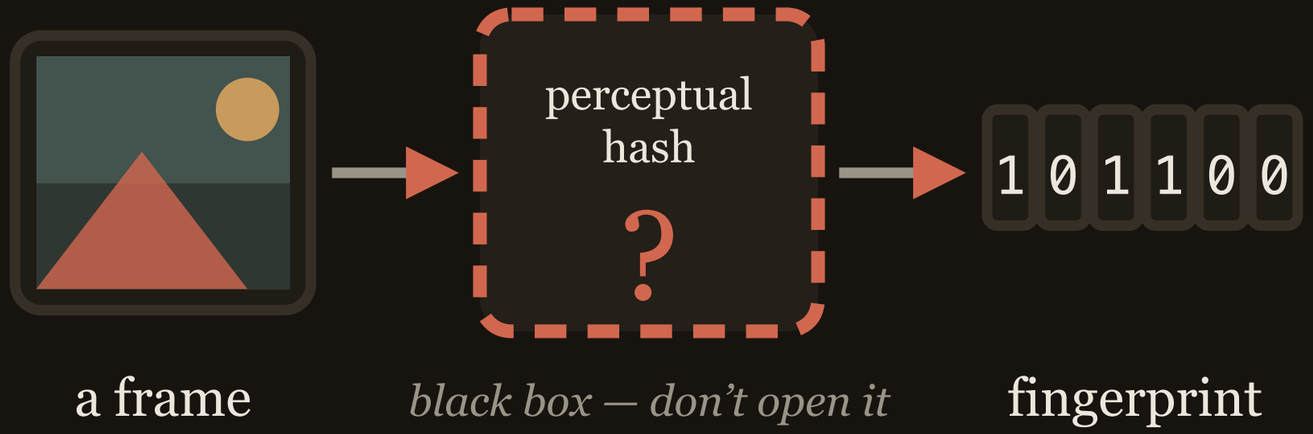
THE FAILURE

Byte equality and **cryptographic hashes** (MD5, SHA) are designed so one changed pixel flips the whole digest. They detect **identical files**, never **perceptually similar** ones.

We need a representation where small perceptual changes → small distance.

CONSTRUCT THE RIGHT REPRESENTATION

Perceptual hashing (pHash)



Each frame becomes a short **fingerprint**. We take the hash on faith — *how* it is computed is a **black box** here; all we need to **trust** is that **similar frames give similar fingerprints**.

The fingerprint survives re-encoding, small crops and overlays — so near-duplicate frames land on near-identical codes. We compare those codes next.

COMPARE → HAMMING DISTANCE

Count the differing bits

$a = 10110010 \quad b = 10111010 \rightarrow d_{Hamming} = 1$

BACK TO PART III

Hamming distance is a proper metric on fingerprints — we verified all four axioms in Part III.

Yet distinct videos can share a fingerprint. A metric *in the representation space* does not guarantee the originals are distinguishable — the same lesson as taxonomy-distance on products.

MAKE A DECISION, ACCEPT ERRORS

Pick a threshold — trade off the two mistakes

DECISION RULE

call it a near-duplicate when Hamming distance ≤ **τ**.

- Threshold too tight → miss real duplicates (false negatives).
- Threshold too loose → flag unrelated videos (false positives).
- τ is chosen from data — exactly the kind of knob the task, not the metric, decides.

MAKE IT SCALE

Fast filter → slow check

cheap candidate generation → expensive exact verification

A RECURRING PATTERN — FAST FILTER, SLOW CHECK

Generate a small candidate set cheaply (fingerprint buckets / LSH), then verify the few survivors with the expensive comparison. Never materialise all n^2 distances.

The same shape powers web search, retrieval and question answering — a recurring pattern this year.

SYNTHESIS

Change the distance, change the patterns

Why this one choice ripples through every task.

BACK TO THE FOUR TASKS

Changing the distance changes the answer

- Nearest neighbours change → classifications and recommendations change.
- Neighbourhoods change → clusters and density estimates change.
- What counts as “far” changes → detected anomalies change.

Case	Main lesson
Products	Choosing which differences matter
Videos	Choosing which differences to ignore

TAKE AWAY · AND A TRANSFER EXERCISE

Choosing a distance is choosing what you can see

THE MESSAGE

Task → representation → distance → decision. Metrics buy predictability and speed; but a valid metric and a task-useful distance are two separate questions.

Choosing a distance is choosing which patterns an algorithm can see.

Transfer: for detecting duplicate news articles — what representation would you choose, which changes should be ignored, and how would you evaluate the result?