Beyond Euclidean: A Conceptual Review of Distance Metrics and Dissimilarity Measures in Contemporary Machine Learning

This paper presents a broad conceptual review of distance metrics and dissimilarity measures fundamental to the field of Machine Learning (ML). While ubiquitous algorithms like k-nearest neighbours, clustering methods, and support vector machines heavily rely on measuring similarity, the default choice of the Euclidean distance often obscures a deeper, application-specific consideration of data geometry. We aim to provide a broad perspective across diverse data types and domains, moving from classical vector norms to covariance-aware, angular, temporal, topological, probabilistic, image-based, and manifold-aware measures. Beyond a simple catalogue, we address several theoretical implications in the ML pipeline. These include the relationship between distance metrics and positive definite kernels, distance metric learning as a subfield dedicated to learning task-specific metrics, latent space fidelity as a way to analyse what distances in low-dimensional latent spaces reveal about semantic separation in the observed space, induced metrics arising from feature maps and learned representations, identifiability of distances as an advanced perspective on latent geometry, and computational complexity aspects. Ultimately, this paper argues that researchers and practitioners should critically evaluate the underlying geometric assumptions in their models, and that the selection of the appropriate distance metric is, in fact, an explicit modelling decision.

Keywords: *Distance Metrics, Dissimilarity Measures, Similarity Learning, Machine Learning Geometry, Metric Learning*

Introduction

Distance measures are central to the conceptual and operational foundations of Machine Learning. Many widely used algorithms, such as k-nearest neighbours, clustering methods (e.g., k-means), and support vector machines, rely explicitly or implicitly on the notion of similarity between data points. Applied clustering studies, including fuzzy cluster analysis of educational data, illustrate how the chosen similarity structure shapes the resulting grouping (Bedalli & Ninka, 2015). In practice, however, the Euclidean distance is frequently used as a default choice, often without critical examination of its suitability for the data at hand.

Formally, a *distance metric* $d: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies four properties for all $x, y, z \in \mathcal{X}$:

$$d(x, y) \geq 0 \tag{1}$$

$$d(x, y) = 0 \Leftrightarrow x = y \tag{2}$$

$$d(x, y) = d(y, x) \tag{3}$$

$$d(x, z) \leq d(x, y) + d(y, z) \tag{4}$$

These correspond to non-negativity, identity of indiscernibles, symmetry, and the triangle inequality. A *dissimilarity measure*, also called *generalised metric*, may

relax one or more of these conditions. A *divergence* is often understood as a non-negative dissimilarity that vanishes exactly when the two inputs are identical, although the term is used somewhat differently across fields. A distance that relaxes (2) is a *pseudometric*. One that relaxes (3) is a *quasimetric*. One that relaxes (4) is a *semimetric*. Dissimilarity measures allow for greater flexibility in modelling complex data relationships. This distinction is particularly relevant in modern ML applications, where data often reside in structured, high-dimensional, or non-Euclidean spaces.

A growing body of research highlights that the choice of distance metric fundamentally shapes the geometry of the data space and, consequently, the behaviour of learning algorithms (Kaya & Bilge, 2019). This has led to the development of Distance Metric Learning (DML) (Suárez et al., 2018), which seeks to learn task-specific distance functions directly from data.

Comprehensive mathematical catalogues of distances, such as the work of Deza and Deza (Deza & Deza, 2016), provide a broad reference for definitions, properties, and examples across many areas. Recent guides also provide broad overviews of similarity measures and their data-science applications (Levy et al., 2025). The present review has a different focus: it does not aim to catalogue distances exhaustively, but to explain how metric choice operates as a modelling decision in contemporary ML pipelines, including kernel methods, learned embeddings, latent spaces, induced geometries, and practical evaluation settings.

A well-known phenomenon in high-dimensional spaces is the *distance concentration effect*, where the relative contrast between nearest and farthest neighbours diminishes as dimensionality grows (Aggarwal et al., 2001). This undermines the intuitive meaning of proximity and helps explain why Euclidean distance can become unreliable for nearest-neighbour search, clustering, and related distance-based methods. Section 2 revisits this effect when discussing classical vector norms.

Several theoretical considerations arise when selecting or designing a distance metric:

- **Kernel Compatibility:** The formal relationship between distance measures and positive definite kernels, which determines their compatibility with kernel methods.
- **Metric Learning:** The mechanism for learning optimised distance metrics directly from data distributions.
- **Latent Space Fidelity:** The fidelity of distances within learned, low-dimensional latent representations.
- **Induced Geometries:** The role of explicit feature transformations in inducing entirely new geometric structures.
- **Scale-Invariance:** The mathematical handling of distances within projective spaces where vectors exhibit inherent scale-invariance.
- **Identifiability:** The identifiability of learned representations, ensuring that an inferred geometry is uniquely determined rather than an artefact of optimisation or parametrisation.

The remainder of this paper is structured as follows. Section 2 presents a taxonomy of distance metrics and dissimilarity measures across domains. Section 3 discusses theoretical implications, including kernel connections, metric learning, latent space representations, geometric transformations, and identifiability. Section 4 addresses practical computational constraints and limitations to have in mind regarding each distance. Section 5 translates these themes into practical selection criteria, and Section 6 concludes.

Classical and Specialised Distance Metrics

Classical Vector Norms

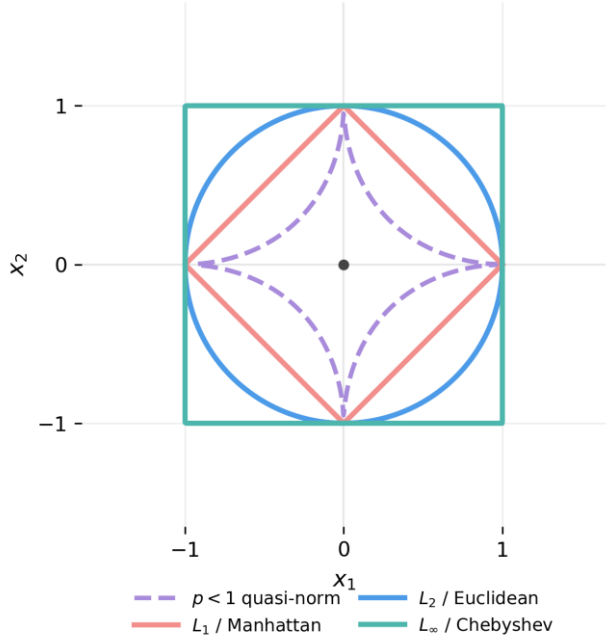
The most fundamental class of distance metrics arises from the Minkowski L_p norm:

$$\| \boldsymbol{x} - \boldsymbol{y} \|_p = \left(\sum_i | x_i - y_i |^p \right)^{\frac{1}{p}}. \quad (5)$$

Special cases of this formulation include the Euclidean (L_2), Manhattan (L_1), and Chebyshev (L_∞) distances. These metrics differ in how they aggregate coordinate-wise differences, leading to distinct geometric interpretations. For example, Euclidean distance assumes isotropy, while Manhattan distance is more robust for outlier detection or when working in high-dimensional sparse settings (Zhou et al., 2025). Figure 1 visualises these differences through the boundaries of the corresponding unit balls.

Despite their simplicity, these norms implicitly assume that features are independent and equally scaled. This assumption is rarely satisfied in real-world datasets, where correlations and heterogeneous feature scales are common.

Figure 1. *Boundaries of L_p unit balls in $\mathbb{R}^2$. The L_1 boundary is diamond-shaped, reflecting Manhattan distance. The L_2 boundary is circular and isotropic. The L_∞ boundary is square, reflecting dependence on the largest coordinate-wise difference. The dashed $p < 1$ curve illustrates a non-convex quasi-norm shape, for which the triangle inequality no longer holds*

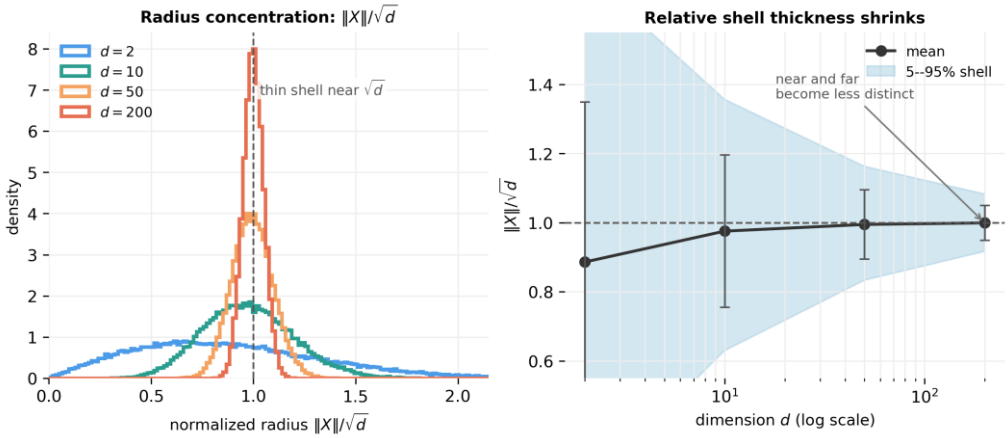


‘

The limitations of these classical norms become especially visible in high dimensions. As dimensionality increases, most pairwise Euclidean distances can fall within a narrow range, reducing the relative contrast between near and far points. This phenomenon can be understood through concentration of measure. For example, in a high-dimensional Gaussian distribution, probability mass concentrates in a thin shell at radius approximately $\sqrt{d}$ rather than near the origin. In such settings, almost all sampled points lie at nearly the same distance from the centre and, by extension, from each other. The result is that proximity becomes less informative for methods such as nearest-neighbour search and clustering.

Figure 2 illustrates this effect empirically. The plot was generated by drawing 45,000 samples from standard Gaussian distributions $X \sim \mathcal{N}(0, I_d)$ for dimensions $d \in \{2,10,50,200\}$, computing the Euclidean radius $\| X \|$, and normalising it by $\sqrt{d}$. As d increases, the distribution of $\| X \|/\sqrt{d}$ becomes increasingly concentrated around one, making the thin-shell geometry visible.

Figure 2. *Concentration of measure for high-dimensional Gaussian data. For* $X \sim \mathcal{N}(0, I_d)$*, the radius* $\| X \|$ *concentrates near* $\sqrt{d}$*. After normalisation by* $\sqrt{d}$*, the empirical distribution of radii collapses into an increasingly thin shell around one. This is the geometric intuition behind the "soap bubble" analogy: most mass lies near a shell rather than near the origin, making relative notions of near and far less informative in high dimensions*



Covariance-Aware Metrics

The Mahalanobis distance generalises the Euclidean distance by explicitly incorporating the covariance structure of the data:

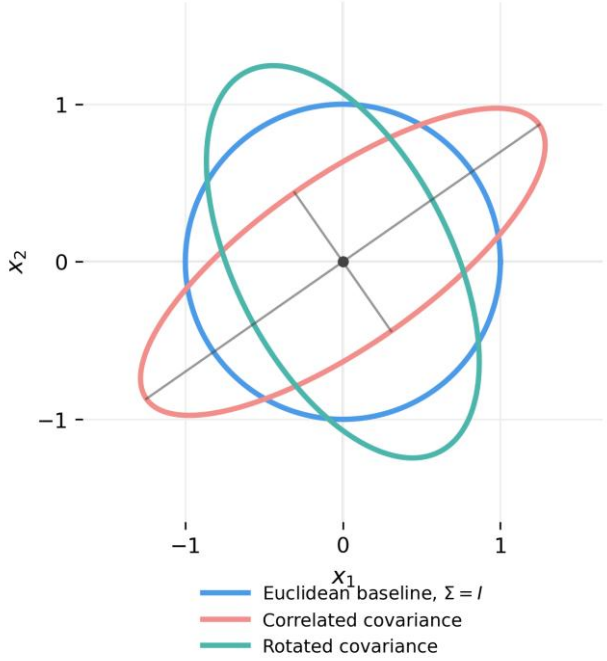
$$D_M(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{y})}. \tag{6}$$

Unlike the Euclidean distance, which assumes that all features are uncorrelated and equally scaled, the Mahalanobis distance accounts for both variance and correlation between features through the covariance matrix Σ. Given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with sample mean $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, this covariance is commonly estimated as

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T. \tag{7}$$

In practice, this means that differences along directions with high variance are considered less informative and are therefore down-weighted, while differences along low-variance directions are amplified. Intuitively, this reflects the idea that variation commonly observed in the data should contribute less to the notion of dissimilarity than rare or atypical variation.

Figure 3. *Mahalanobis distance contours for different covariance structures. While Euclidean distance treats all directions equally, Mahalanobis distance rescales and rotates the effective geometry according to the covariance matrix* Σ*, down-weighting high-variance directions and amplifying low-variance directions.*



Geometrically, this induces a transformation of the feature space in which the data is linearly transformed so that the covariance becomes the identity matrix. In this transformed space, the Mahalanobis distance reduces to the standard Euclidean distance. Consequently, distance contours that are spherical under Euclidean distance become ellipsoidal in the original space, aligning with the principal directions of the data distribution. This makes the Mahalanobis distance particularly suitable for anisotropic data, where the assumption of isotropy is violated. Figure 3 illustrates this covariance-aware deformation of distance contours.

The Mahalanobis distance plays a central role in statistical modelling, anomaly detection, and classification. For example, in multivariate Gaussian models, it naturally arises in the exponent of the likelihood function and is therefore directly linked to probabilistic interpretations of distance. In anomaly detection, points with large Mahalanobis distance from the mean are considered outliers relative to the underlying distribution. Furthermore, in Distance Metric Learning (DML), many approaches aim to learn a matrix $M = \Sigma^{-1}$ (or a related positive semi-definite matrix), thereby adapting the geometry of the space to optimise task-specific notions of similarity (Weinberger & Saul, 2009).

Several extensions have been proposed. The local Mahalanobis distance (Ghosh et al., 2025) adapts the matrix M depending on the region of the space, allowing for non-linear and heterogeneous data structures. Similarly, kernelised Mahalanobis distances implicitly define such transformations in high-dimensional feature spaces via kernel functions, bridging the gap between metric learning and kernel methods (Schölkopf et al., 2001). These extensions illustrate that the

Mahalanobis distance is not a single isolated metric but rather a representative of a broader class of geometry-aware distances that adapt to the structure of the data, either through statistical estimation or learning.

Cosine Similarity, Angular Distance and Pearson's Distance

Cosine similarity measures the directional alignment between two non-zero vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$S_{\cos}(\boldsymbol{x}, \boldsymbol{y}) = \cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\| \boldsymbol{x} \| \| \boldsymbol{y} \|}. \tag{8}$$

As a similarity score, $S_{\cos}$ reflects angle rather than magnitude. A common dissimilarity derived from it is the cosine dissimilarity

$$D_{\cos}(\boldsymbol{x}, \boldsymbol{y}) = 1 - S_{\cos}(\boldsymbol{x}, \boldsymbol{y}), \tag{9}$$

but $D_{\cos}$ is not a true metric because it can violate the triangle inequality. When vectors are normalised to unit length ($\| \boldsymbol{x} \| = \| \boldsymbol{y} \| = 1$), Euclidean distance is proportional to cosine similarity via

$$\| \boldsymbol{x} - \boldsymbol{y} \| = \sqrt{2(1 - S_{\cos}(\boldsymbol{x}, \boldsymbol{y}))}, \tag{10}$$

so cosine-based comparisons can be viewed as chordal or angular constraints on a spherical manifold.

The Pearson (correlation) distance is a centred analogue of cosine dissimilarity and accounts for mean offsets. Using the vector means $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$, the mean-adjusted vectors are $\boldsymbol{x} - \bar{\boldsymbol{x}}$ and $\boldsymbol{y} - \bar{\boldsymbol{y}}$. The Pearson (correlation) distance can be written as

$$D_{\text{corr}}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{(\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot (\boldsymbol{y} - \bar{\boldsymbol{y}})}{\| \boldsymbol{x} - \bar{\boldsymbol{x}} \| \| \boldsymbol{y} - \bar{\boldsymbol{y}} \|}. \tag{11}$$

Cosine- and correlation-based measures are particularly useful in high-dimensional embedding spaces (e.g., word and sentence embeddings (Mikolov et al., 2013; Pennington et al., 2014; Reimers & Gurevych, 2019)) because they emphasise angular relationships and are less sensitive to variations in vector norms. Directional high-dimensional models, such as discriminant analysis with the von Mises-Fisher distribution, provide a statistical example in which angular structure is central rather than incidental (Romanazzi, 2014). The choice among raw cosine similarity, angular distance, or Pearson distance should be guided by whether magnitude, absolute direction, or mean-centred pattern is most relevant to the task.

Point Cloud Dissimilarity

Before considering finite point cloud losses, it is useful to recall a classical distance between sets. Given a pointwise Minkowski norm $\|\cdot\|_p$, the distance from a point x to a non-empty set Y is $D_p(x,Y) = \inf_{y\in Y} \| x - y \|_p$. This induces the Hausdorff-type set distance

$$D_H^{(p)}(X,Y) = \max\left\{\sup_{x\in X} D_p(x,Y),\ \sup_{y\in Y} D_p(y,X)\right\}. \tag{12}$$

For non-empty compact sets this is a genuine metric, and it is not restricted to finite point clouds. For finite point sets, however, it is governed by the worst nearest-neighbour discrepancy, making it highly sensitive to outliers (Deza & Deza, 2016).

For comparing unordered 3D point sets, a common alternative is the Chamfer distance. Such point-set comparisons are central in 3D reconstruction settings, including object reconstruction pipelines and broader surveys of deep-learning-based architectural reconstruction (Enesi & Kuqi, 2023; Gorjian et al., 2025). Given two point sets X and Y, it measures bidirectional nearest-neighbour discrepancies:

$$D_{CD}(X,Y) = \sum_{x\in X} \min_{y\in Y} \| x - y \|^2 + \sum_{y\in Y} \min_{x\in X} \| x - y \|^2. \tag{13}$$

Intuitively, this measure penalises each point in one structure based on its closest counterpart in the other structure, capturing local geometric consistency without requiring explicit point correspondences. Naively, the nearest-neighbour searches require comparing all pairs of points, although spatial indexes or approximate search can reduce the practical cost. This makes it computationally efficient relative to full optimal transport and well-suited for large, unordered point clouds. Recent variants also attempt to learn or reweight Chamfer-style matching costs for point cloud reconstruction tasks (Huang et al., 2024). In the form given above, the Chamfer quantity is best understood as a dissimilarity measure rather than a metric in the strict axiomatic sense. In particular, the nearest-neighbour assignments are recomputed independently for each pair of point sets, so the resulting quantity does not in general satisfy the triangle inequality.

However, because the Chamfer distance relies purely on nearest-neighbour matching, it is primarily sensitive to local geometric deviations and does not explicitly encode global structural relationships. As a result, it may assign low distances to structures that share similar local features but differ significantly in their overall fold or global topology.

An alternative to the Chamfer distance is the Earth Mover's Distance (EMD), a point-set dissimilarity grounded in optimal transport theory (Peyré & Cuturi, 2019; Villani, 2009). To compare two point sets, EMD treats them as discrete distributions of mass rather than as collections of independent nearest-neighbour queries. Let $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$. Assign non-negative weights $a_i \geq 0$ to the points of X and $b_j \geq 0$ to the points of Y, with $\sum_i a_i = \sum_j b_j = 1$. These weights specify

how much mass is located at each point. Also let $c(x_i, y_j) \geq 0$ be the ground cost of moving one unit of mass from x_i to y_j.

The admissible transport plans are collected in

$$\Pi(a, b) = \left\{\pi \in \mathbb{R}_{\geq 0}^{m \times n} : \sum_{j=1}^{n} \pi_{ij} = a_i,\ \sum_{i=1}^{m} \pi_{ij} = b_j\right\}. \tag{14}$$

Thus each coefficient $\pi_{ij} \geq 0$ specifies how much mass is transported from x_i to y_j, while the row and column constraints ensure that all mass from X is sent and all mass required by Y is received. With these definitions, EMD is the minimum total transport cost

$$D_{\mathrm{EMD}}(X, Y) = \min_{\pi \in \Pi(a,b)} \sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij}\ c(x_i, y_j), \tag{15}$$

so the distance reflects a globally optimal redistribution rather than a set of local nearest-neighbour decisions. If we define the ground cost as $c(x, y) = \| x - y \|^p$, we obtain the common formulation of the discrete Wasserstein distance:

$$W_p(X, Y) = \left(\min_{\pi \in \Pi(a,b)} \sum_{i=1}^{m} \sum_{j=1}^{n} \pi_{ij}\ \| x_i - y_j \|^p \right)^{1/p}. \tag{16}$$

In this sense, EMD is the optimal transport formulation underlying Wasserstein-type distances. In common usage, EMD often refers specifically to the W_1 case with ground cost $\| x - y \|$. Optimal transport ideas have also been extended beyond Euclidean ground spaces, for example to tropical geometric settings where Wasserstein-type distances are defined on tropical projective spaces (Lee et al., 2022).

In protein backbone comparison, for example, each structure can be represented as a point set of 3D coordinates, typically corresponding to Cα atoms. Chamfer distance can provide a simple geometric baseline, but biologically meaningful similarity often depends on global fold, alignment, and length normalisation rather than local nearest-neighbour proximity (Zhang & Skolnick, 2004).

Information-Theoretic Measures for Probability Distributions

When the objects to be compared are probability distributions rather than individual data points, classical vector-space metrics are generally not applicable. Instead, a family of information-theoretic divergences and distances has been developed to quantify distributional discrepancy.

The Kullback–Leibler (KL) divergence (Cover & Thomas, 2006; Kullback & Leibler, 1951) measures the information lost when a distribution Q is used to approximate a reference distribution P:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathrm{supp}(P)} P(x) \log \frac{P(x)}{Q(x)}\, dx. \tag{17}$$

To ensure the integral is well-defined, it is assumed that the support of P is contained within the support of Q ($\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$). Otherwise, the divergence becomes infinite. It answers the question: *how much additional information is required to encode samples from P using a code optimised for Q?* Crucially, D_{KL} is not a metric: it is asymmetric ($D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$) and unbounded, which limits its use in settings that require a proper distance structure.

The Jensen–Shannon (JS) divergence (Lin, 1991) remedies the asymmetry of KL by measuring how far P and Q each deviate from their mixture $M = 1/2\ (P + Q)$:

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M). \tag{18}$$

The JS divergence is symmetric and bounded in $[0,1]$ (when using the base-2 logarithm), and its square root satisfies the triangle inequality, making it a true metric. It finds widespread application in generative modelling and natural language processing.

The Hellinger distance (Hellinger, 1909) is defined as

$$D_{\mathrm{H}}(P, Q) = \frac{1}{\sqrt{2}} \parallel \sqrt{P} - \sqrt{Q} \parallel_2. \tag{19}$$

Here $\sqrt{P}$ and $\sqrt{Q}$ denote the pointwise square-root representations of the two distributions. For discrete probability vectors $P = (p_k)$ and $Q = (q_k)$, this is

$$D_{\mathrm{H}}(P, Q) = \left(\frac{1}{2}\sum_k \left(\sqrt{p_k} - \sqrt{q_k}\right)^2\right)^{1/2}. \tag{20}$$

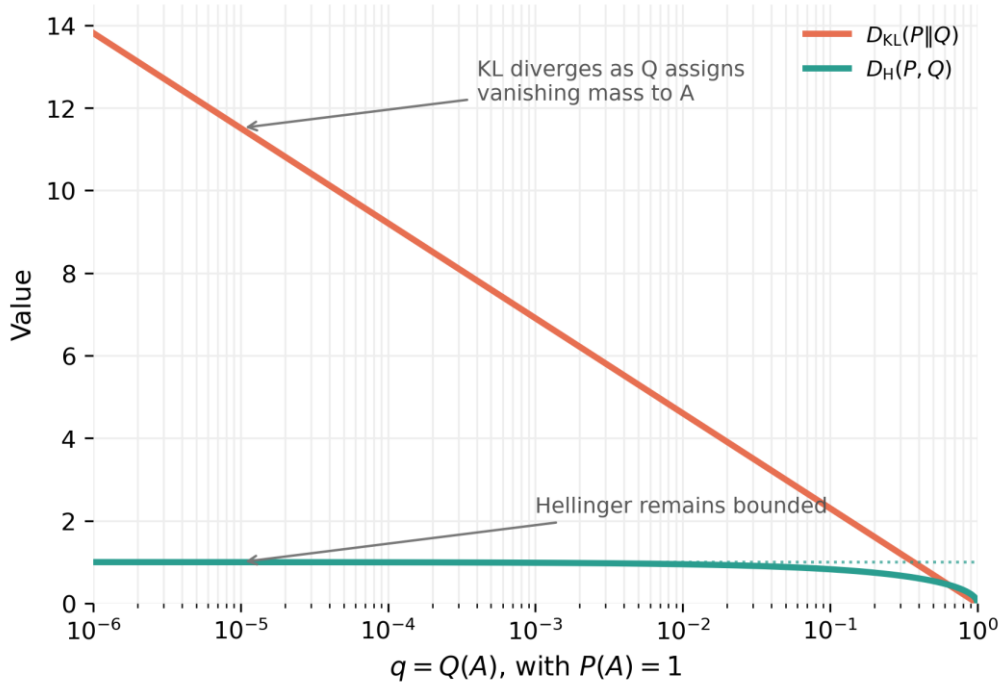
The square-root transformation embeds probability distributions into a Euclidean geometry in which ordinary ℓ_2 distance becomes a bounded statistical distance. The Hellinger distance is therefore a genuine metric that is symmetric, bounded in $[0,1]$, and particularly robust for comparing distributions with differing supports. It is often preferred over the KL divergence in practical data-science applications, for instance in the presence of class imbalance, because it does not diverge when one distribution assigns zero probability to events that the other does not. Figure 4 contrasts this bounded behaviour with the unbounded growth of KL divergence in a simple two-outcome case.

The Wasserstein distance extends the optimal-transport perspective introduced for point sets in Section 2.4 to probability distributions. Unlike KL, JS, and Hellinger distances, which compare probability mass through density values, Wasserstein distance also uses a ground metric on the sample space. It therefore measures how much work is required to move the mass of one distribution into the other. Formally, the p-Wasserstein distance is

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int \parallel x - y \parallel^p \ d\gamma(x, y)\right)^{1/p}, \tag{21}$$

where $\Gamma(P, Q)$ denotes the set of all couplings of P and Q (Peyré & Cuturi, 2019; Villani, 2009). The W_2 distance, often referred to as the *Fréchet distance* in the context of feature distributions, is of particular practical importance: it underlies the Fréchet Inception Distance (FID), a standard evaluation metric for generative image models (Heusel et al., 2017). A key advantage of the Wasserstein distance over KL and JS divergences is that it can remain meaningful even when two distributions have little or no overlap in support, a common situation in high-dimensional generative modelling. The trade-off is computational: exact optimal transport can be expensive for large empirical distributions.

Figure 4. *Comparison of KL divergence and Hellinger distance in a two-outcome example. Let* $P(A) = 1$ *and* $Q(A) = q$*. As* q *approaches zero,* $D_{\mathrm{KL}}(P \parallel Q) = \log(1/q)$ *diverges, while* $D_{\mathrm{H}}(P, Q) = \sqrt{1 - \sqrt{q}}$ *remains bounded by one. This illustrates why bounded distances can be more stable than KL divergence when distributions assign very different probability mass to the same event*



Together, these measures span a spectrum of trade-offs between theoretical rigour, computational tractability, and sensitivity to distributional properties, and the appropriate choice depends heavily on the application at hand.

Topological Data Analysis

Topological Data Analysis provides a framework for characterising the shape of data by tracking the birth and death of topological features (connected components, loops, and voids) across a range of scales (Chazal & Michel, 2021). This range is usually organised as a filtration, meaning a nested sequence of spaces

built by gradually increasing a scale parameter, for example by connecting points that lie within a growing distance threshold. These features are summarised in a persistence diagram $\mathcal{D}$, a multiset of points $(b, d) \in \mathbb{R}^2$ with $b < d$, where b and d denote the birth and death scales of a topological feature respectively. Comparing two such diagrams requires metrics that respect this combinatorial structure. The two most widely used are the bottleneck distance and the diagram Wasserstein distance. The latter uses the same optimal-matching intuition as the transport distances discussed in Section 2.4 and Section 2.5, but the objects being matched are topological features in persistence diagrams rather than mass elements in the original data space.

The bottleneck distance between two persistence diagrams $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as

$$D_\infty(\mathcal{D}_1, \mathcal{D}_2) = \inf_{\eta} \sup_{x \in \mathcal{D}_1} \| x - \eta(x) \|_\infty, \tag{22}$$

where the infimum is taken over all bijections $\eta: \mathcal{D}_1 \to \mathcal{D}_2$, with points on the diagonal $\{(a, a): a \in \mathbb{R}\}$ included in both diagrams to account for unmatched features. The bottleneck distance is thus governed by the *worst-case* correspondence: it records the largest displacement required to match any single feature across the two diagrams. This gives it strong stability guarantees under small perturbations of the filtration, but also means that a single large unmatched feature can dominate the distance.

The q-Wasserstein distance on persistence diagrams generalises this by aggregating the cost over *all* matched pairs:

$$D_{W,q}(\mathcal{D}_1, \mathcal{D}_2) = \left(\inf_{\eta} \sum_{x \in \mathcal{D}_1} \| x - \eta(x) \|_\infty^q \right)^{1/q}, \tag{23}$$

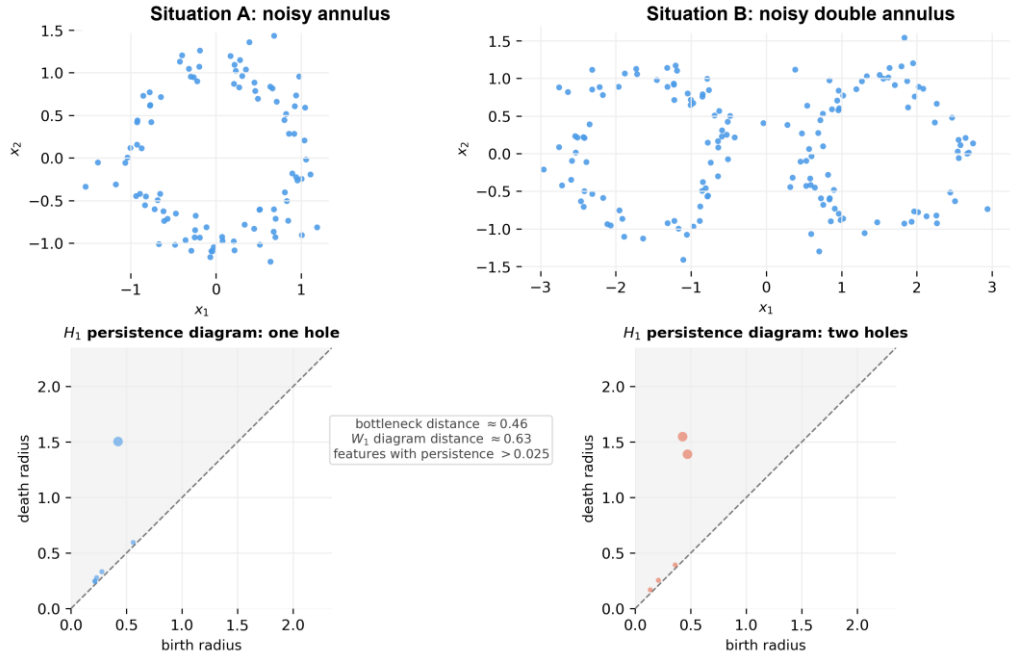
where again the infimum ranges over bijections that allow matching to the diagonal. In the limit $q \to \infty$ one recovers the bottleneck distance, while finite q penalises the total transport cost across the entire matching.

Crucially, the use of the term "Wasserstein" here refers to a discrete optimal assignment problem between point sets (extending the perspective of Section 2.4) rather than the comparison of probability distributions detailed in Section 2.5. While both formulations share the foundational concept of minimising a global transport or matching cost, the diagram Wasserstein distance operates directly on the combinatorial structure of persistence diagrams, preserving its status as a proper metric without requiring an information-theoretic or probabilistic framework.

Figure 5 illustrates this construction on a small synthetic experiment. We sampled two noisy point clouds in the plane: one annular cloud with a single central hole and a second cloud formed from two separated annular components. A Vietoris–Rips filtration was then built by increasing the distance threshold at which sampled points are connected, and the resulting one-dimensional persistence diagrams were computed from the birth and death of loops. The plotted points retain features with persistence greater than 0.025, so short-lived near-diagonal

fluctuations are suppressed while the dominant holes remain visible. The bottleneck and diagram Wasserstein distances are then obtained by optimally matching these diagram points, allowing unmatched features to be assigned to the diagonal.

Figure 5. *Persistence diagrams for two noisy annulus-like point cloud examples. The top row shows the sampled point clouds: one annular cloud with a single hole and a second cloud containing two separated annular components. The bottom row shows the corresponding H_1 persistence diagrams after filtering to features with persistence greater than 0.025. The single-annulus example has one dominant off-diagonal point, while the double-annulus example has two dominant off-diagonal points. The displayed bottleneck and W_1 diagram distances summarise the cost of optimally matching these topological features, including possible matches to the diagonal*



Both distances are stable with respect to small perturbations of the underlying data: a small change in the input filtration induces a correspondingly small change in the persistence diagram under either metric. This *stability property* is a cornerstone result of Topological Data Analysis and underpins the practical utility of these distances in noisy settings. Furthermore, both metrics are invariant to certain geometric transformations, such as rigid motions, provided the filtration is defined in a transformation-invariant manner. These properties make persistence-diagram distances well suited to applications in computational geometry, material science, and biological shape analysis, where data exhibit complex global structure that local, point-wise metrics fail to capture.

Temporal Alignment

Standard point-wise distances, such as the Euclidean distance, are ill-suited for comparing time series that exhibit temporal misalignment: two sequences that are

otherwise identical but shifted in phase, or scaled in speed, will incur a large Euclidean distance despite being semantically similar. Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978) addresses this limitation by seeking an optimal non-linear alignment between two sequences before measuring their discrepancy. Figure 6 shows how DTW avoids the misleading point-wise mismatch produced by a simple Euclidean comparison.

Formally, let $\boldsymbol{x} = (x_1, \dots, x_m)$ and $\boldsymbol{y} = (y_1, \dots, y_n)$ be two time series. A *warping path* $\mathcal{W} = (w_1, \dots, w_K)$ is a sequence of index pairs $w_k = (i_k, j_k) \in \{1, \dots, m\} \times \{1, \dots, n\}$ satisfying:

- *Boundary conditions:* $w_1 = (1,1)$ and $w_K = (m, n)$.
- *Monotonicity:* $i_k \leq i_{k+1}$ and $j_k \leq j_{k+1}$ for all k.
- *Step continuity:* $i_{k+1} - i_k \leq 1$ and $j_{k+1} - j_k \leq 1$ for all k.

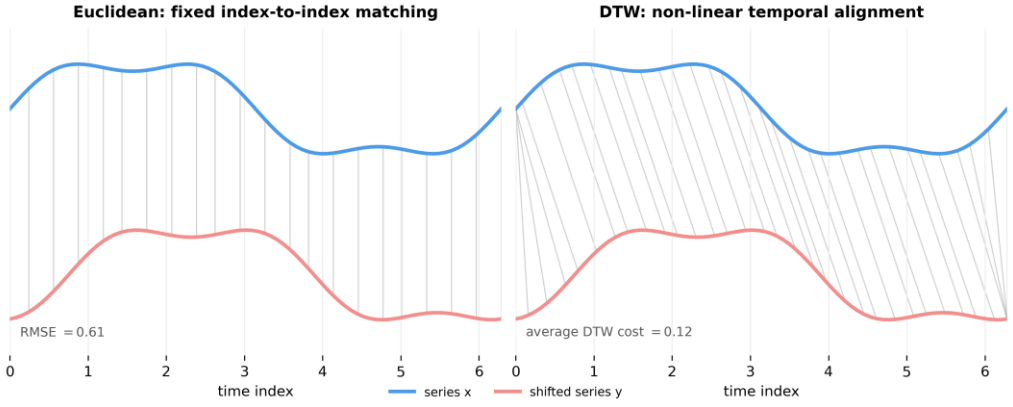
The DTW distance is then defined as the cost of the optimal warping path:

$$D_{\mathrm{DTW}}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\mathcal{W}} \sum_{k=1}^{K} D\left(x_{i_k}, y_{j_k}\right), \tag{24}$$

where $D(\cdot,\cdot)$ is a local distance measure, typically the squared or absolute difference. This optimisation is solved exactly in $\mathcal{O}(mn)$ time via dynamic programming, making DTW tractable for moderate sequence lengths.

The warping path encodes a many-to-one alignment: a single time step in one sequence may be matched to multiple consecutive steps in the other, allowing DTW to absorb differences in speed, rhythm, or phase that would otherwise inflate a point-wise distance. It should be noted, however, that DTW does not satisfy the triangle inequality in general and is therefore a dissimilarity measure rather than a true metric.

Figure 6. *Euclidean matching versus Dynamic Time Warping for two shifted time series. A point-wise Euclidean comparison aligns samples at the same time index, so phase shifts create large apparent discrepancies. DTW instead allows a non-linear warping path, aligning similar signal shapes even when they occur at different times*



DTW has found widespread application across domains where temporal structure is central: in speech recognition it enables robust comparison of utterances spoken at different rates. It has also been used for pattern discovery in time series databases, approximate matching in vocal-music retrieval and scoring, and analogy-based pandemic forecasting (Berndt & Clifford, 1994; Ye, 2026; Zhang & Ji, 2026). In bioinformatics it aligns gene-expression time courses or electrocardiographic signals. In finance it measures the similarity of price trajectories that may lead or lag one another. These diverse use cases illustrate that DTW is not merely a technical fix for phase misalignment, but a principled modelling choice reflecting the belief that *when* events occur matters less than *whether* they occur and in what order.

Discrete and Set-Based Metrics

Continuous vector-space distances are not meaningful when data are inherently discrete (binary strings, categorical feature vectors, sets of tokens, etc.) and dedicated metrics are required. Three measures are particularly prominent in this setting.

The Hamming distance between two sequences $x, y \in \Sigma^n$ over an alphabet Σ counts the number of positions at which the two sequences disagree:

$$D_{\mathrm{H}}(x, \boldsymbol{y}) = \sum_{i=1}^{n} \mathbf{1}\,[x_i \neq y_i]. \tag{25}$$

For binary strings this reduces to the number of bit-flips required to transform one string into the other, which coincides with the L_1 norm on $\{0,1\}^n$. The Hamming distance is a true metric and forms the foundation of classical error-correcting codes, where the minimum Hamming distance between codewords determines the error-detection and correction capacity of a code. In machine learning it is widely used for comparing categorical feature vectors and hashed representations. Generalised Hamming-type distances have also been proposed when near misses or structured symbol relabellings should receive partial credit rather than being counted as exact mismatches (Bookstein et al., 2002; Liu & Na, 2025). In perceptual hashing, recent work has similarly argued that plain Hamming distance can miss spatial structure in image hashes (McKeown, 2025).

The Hamming distance can be applied to text when strings are represented as equal-length sequences of discrete characters. More general string dissimilarities are provided by *edit distances* (Deza & Deza, 2016), which measure the minimum number or cost of operations required to transform one string into another. A widely used example is the Damerau–Levenshtein distance, defined recursively as:

$$D_{a,b}(i,j) = \min \begin{cases} 0 & \text{if } i = j = 0, \\ D_{a,b}(i-1,j) + 1 & \text{if } i > 0, \\ D_{a,b}(i,j-1) + 1 & \text{if } j > 0, \\ D_{a,b}(i-1,j-1) + \mathbf{1}_{(a_i \neq b_j)} & \text{if } i,j > 0, \\ D_{a,b}(i-2,j-2) + \mathbf{1}_{(a_i \neq b_j)} & \text{if } i,j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j, \end{cases} \tag{26}$$

with $\mathbf{1}_{(a_i \neq b_j)}$ being the indicator function equal to 0 when $(a_i = b_j)$ and equal to 1 otherwise. Each recursive call corresponds to some text alteration e.g., deletion, insertion or mismatch of character. Other examples of popular edit distances are the Longest Common Subsequence (LCS) and the Jaro distance. Although they lost relevancy regarding NLP with the development of text tokenizers, edit distances are still used in practical ML applications such as spelling correctors.

When the objects of interest are *sets* rather than ordered sequences, cardinality-based overlap measures are more natural. The Jaccard index quantifies the similarity between two finite sets A and B as the ratio of their intersection to their union:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \qquad J(A,B) \in [0,1]. \tag{27}$$

A value of 1 indicates identical sets and 0 indicates disjoint sets. The corresponding *Jaccard distance* $1 - J(A,B)$ is a true metric. The Jaccard index is particularly prevalent in information retrieval, document deduplication via MinHash, and ecological studies comparing species compositions. Related overlap measures have also been analysed in recommendation and diversity settings, where the Jaccard index appears alongside Sørensen, cosine, and entropy-derived measures (Jost, 2006; Verma & Aggarwal, 2020).

The Dice coefficient (also known as the Sørensen–Dice index) is a closely related overlap measure:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}, \qquad \mathrm{DSC}(A,B) \in [0,1]. \tag{28}$$

The factor of two in the numerator compensates for the fact that the shared elements $A \cap B$ are counted once in each of $|A|$ and $|B|$, rendering the denominator equivalent to $|A \cup B| + |A \cap B|$. Consequently the Dice coefficient up-weights the contribution of the overlap relative to the Jaccard index, and the two are related by

$$\mathrm{DSC}(A,B) = \frac{2\,J(A,B)}{1 + J(A,B)}. \tag{29}$$

Unlike the Jaccard distance, $1 - \mathrm{DSC}$ does not satisfy the triangle inequality and is therefore not a metric. Nevertheless, the Dice coefficient is the dominant evaluation criterion in medical image segmentation, where it directly measures the spatial overlap between a predicted and a ground-truth region, and it also serves as a differentiable training objective in segmentation networks when applied to soft, probabilistic predictions. Modified Dice coefficients continue to be proposed for medical-image settings where boundary location and spatial error matter in addition to raw overlap (Rainio & Klén, 2025).

Together, these three measures illustrate a recurring theme in metric design: the appropriate notion of dissimilarity is inseparable from the structure of the data. Imposing a continuous geometry on inherently discrete or combinatorial objects risks obscuring the very relationships one wishes to quantify.

Similarity Measures Between Images and Videos

Images and videos highlight the gap between signal-level and perceptual similarity: two images can be close under a pixel norm yet visually different, or visually similar despite small translations or intensity changes. This motivates a hierarchy of measures, from pixel-wise errors to structural, information-theoretic, and distribution-level criteria.

Pixel-wise measures such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) treat an image as a vector in $\mathbb{R}^{H\times W\times C}$. They are simple and useful in compression or restoration benchmarks, but they correlate poorly with human judgement when small spatial shifts or perceptually minor changes produce large pointwise errors.

Structural and statistical measures address part of this limitation. Edge-detection and facial-image-processing applications illustrate why the chosen image similarity measure or feature representation can strongly affect downstream recognition and matching pipelines (Arriagada et al., 2019; Rodrigues et al., 2016). The Structural Similarity Index (SSIM) compares local luminance, contrast, and structure, and was introduced precisely to better align image-quality assessment with human perception (Wang et al., 2004). Normalised cross-correlation (NCC) compares local intensity patterns after centring and scaling, making it useful for template matching and optical flow when brightness changes are approximately affine (Lewis, 1995). Mutual Information (MI) compares joint intensity statistics and is therefore common in multi-modal image registration (Maes et al., 1997; Viola & Wells, 1997).

For generative models, evaluation usually compares distributions of images rather than paired examples. The Fréchet Inception Distance (FID) fits Gaussians to real and generated Inception-v3 activations and computes the corresponding Wasserstein-2 distance (Heusel et al., 2017). For video, Fréchet Video Distance (FVD) applies the same principle to spatio-temporal features, making temporal coherence part of the comparison (Unterthiner et al., 2018). These distribution-level metrics are reference-free in the sense that they compare sets of generated and real samples rather than paired examples, but they depend strongly on the chosen feature encoder and its domain biases. Recent empirical studies of synthetic-image evaluation continue to examine how FID and Inception Score behave for generative models such as VAEs (Chan & Sithungu, 2025).

Geodesic Distances and Manifold-Aware Metrics

Many real-world datasets do not occupy a flat Euclidean space but instead lie on or near a low-dimensional *manifold* embedded in a high-dimensional ambient space. In such settings, the straight-line Euclidean distance between two points may be geometrically misleading: it cuts through empty regions of the ambient space that contain no data, whereas the true dissimilarity should follow the curvature of the manifold. The geodesic distance addresses this by measuring the length of the shortest path between two points *along the manifold*, rather than through the surrounding space.

Formally, given a Riemannian manifold $(\mathcal{M}, g)$ with metric tensor g, the geodesic distance between two points $p, q \in \mathcal{M}$ is:

$$D_{\text{geo}}(p, q) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\, dt, \tag{30}$$

where the infimum is taken over all smooth curves $\gamma: [0,1] \to \mathcal{M}$ with $\gamma(0) = p$ and $\gamma(1) = q$, and $\dot{\gamma}(t)$ denotes the tangent or velocity vector of the curve at time t. When $\mathcal{M}$ is a sphere S^{n-1}, this reduces to the great-circle distance. When $\mathcal{M} = \mathbb{R}^n$ with the standard metric, it coincides with the Euclidean distance.

Graph-based approximation and Isomap

In practice, the manifold $\mathcal{M}$ is unknown and must be inferred from a finite sample of data points. The standard approach is to construct a neighbourhood graph G in which each point x_i is connected to its k nearest Euclidean neighbours (or all neighbours within radius ε), with edge weights equal to the Euclidean distances between connected points. The geodesic distance between any two points is then approximated by the shortest path in G, computed efficiently via Dijkstra's or Floyd–Warshall's algorithm. Recent graph-algorithm work continues to refine all-pairs shortest-path computation, including Floyd–Warshall variants for sparse or disconnected graphs (Zugan et al., 2025). This approximation converges to the true geodesic distance as the sample density increases, under mild smoothness conditions on $\mathcal{M}$.

Isomap (Tenenbaum et al., 2000) builds directly on this idea: it replaces the Euclidean pairwise distance matrix used in classical Multidimensional Scaling (MDS) with the matrix of graph-based geodesic distances, then applies MDS to find a low-dimensional embedding that preserves these geodesic distances as faithfully as possible. This yields a nonlinear dimensionality reduction method that correctly "unfolds" curved manifolds where PCA and Euclidean MDS fail entirely. Figure 7 illustrates why the distinction between ambient and intrinsic distance matters in such settings on the famous Swiss roll example.

Heat kernels and diffusion distances

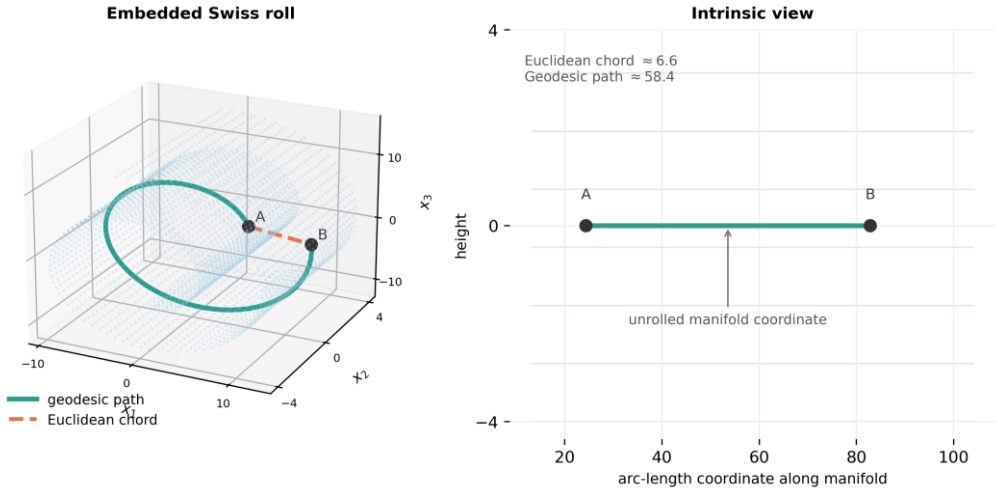
A complementary family of manifold-aware metrics is based on the heat equation on the data graph. The *diffusion distance* at time t between points x_i and x_j is defined as:

$$D_t(x_i, x_j)^2 = \sum_k \lambda_k^{2t} \left(\phi_k(x_i) - \phi_k(x_j)\right)^2, \tag{31}$$

where λ_k and ϕ_k are the eigenvalues and eigenvectors of the normalised graph Laplacian. To construct this operator, one first forms a weighted adjacency matrix W, where W_{ij} records the affinity between neighbouring data points, and a diagonal degree matrix D with $D_{ii} = \sum_j W_{ij}$. The unnormalised graph Laplacian is $L = D - W$, while common normalised variants include $L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}$ and $L_{\text{rw}} = I - D^{-1} W$. These matrices are discrete analogues of the Laplace–Beltrami

operator on a manifold: they compare the value of a function at a point with its values on neighbouring points, and therefore encode how signals, heat, or random walks propagate across the data graph.

Figure 7. *Ambient Euclidean distance versus intrinsic geodesic distance on a Swiss-roll manifold. In the embedded view, the Euclidean chord between two points cuts through the surrounding space and ignores the data manifold. The geodesic path instead follows the rolled surface. In the intrinsic unrolled coordinate system, this path corresponds to distance measured along the manifold itself*



Intuitively, the diffusion distance measures how differently heat (or a random walk) spreads from x_i versus x_j over t steps. This makes it robust to small perturbations and noise: two points connected by many short paths (i.e., well within the same region of the manifold) will have a small diffusion distance even if a single shortest path is corrupted. Diffusion Maps (Coifman & Lafon, 2006) use this distance to derive a low-dimensional embedding, analogously to Isomap but with stronger noise robustness and a probabilistic interpretation via random walks.

Geodesic distances on meshes and point clouds

When the manifold is explicitly represented as a triangulated surface mesh, exact geodesic distances can be computed via the Fast Marching Method (Sethian, 1996), which propagates a wavefront across triangle faces in $O(N\log N)$ time, or via the Heat Method (Crane et al., 2013), which solves two sparse linear systems on the mesh and is particularly efficient on GPU hardware. These methods are standard in 3D shape analysis and computational geometry, where geodesic distances underpin shape descriptors, surface parameterisation, and shape retrieval.

Limitations

Geodesic approximations via neighbourhood graphs are sensitive to the choice of k or ε. If it is too small, the graph becomes disconnected. If it is too large, short-circuit edges bridge distinct regions of the manifold, corrupting the distance. Furthermore, geodesic distances are expensive to compute for large datasets as all-pairs shortest paths scale as $O(N^2\log N)$ and are not straightforwardly

differentiable, complicating their use as training objectives in deep learning pipelines.

Theoretical Implications

Positive Definite Kernels and Their Relationship to Distance Metrics

Kernel methods occupy a central place in machine learning, underpinning support vector machines, Gaussian processes, kernel PCA, and a broad class of non-parametric estimators, with kernel-trick formulations also appearing in regression and classification schemes beyond standard SVMs (Huh, 2015). Their power derives from the *kernel trick*: by implicitly mapping data into a high-dimensional (possibly infinite-dimensional) feature space $\mathcal{H}$, linear operations in $\mathcal{H}$ capture nonlinear structure in the original input space without ever computing the map explicitly. The central object is therefore a positive definite kernel, and the key question for this review is when a distance can be used to construct one.

Positive definite kernels and RKHS embeddings

A symmetric function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if, for any finite collection of points $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathcal{X}$ and coefficients $c_1, \dots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0. \tag{32}$$

Equivalently, the Gram matrix G with entries $G_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi-definite for every finite sample. A canonical example is the Gaussian radial basis function (RBF) kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2\sigma^2}\right), \tag{33}$$

where nearby vectors receive large similarity values and distant vectors receive values close to zero. Under standard conditions, such kernels admit a feature map $\phi: \mathcal{X} \to \mathcal{H}$ into a reproducing kernel Hilbert space (RKHS) such that

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle_{\mathcal{H}}. \tag{34}$$

This factorisation is what makes the kernel trick possible: inner products in $\mathcal{H}$ can be evaluated directly through K without explicit construction of ϕ (Berg et al., 1984; Berlinet & Thomas-Agnan, 2004).

From distance metrics to kernels: conditionally negative definite functions

Not every distance metric induces a valid PD kernel, and understanding which distances do is essential for the correct application of kernel methods. The key concept is that of a *conditionally negative definite* (CND) distance: a symmetric function $d: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $d(\boldsymbol{x}, \boldsymbol{x}) = 0$ is CND if

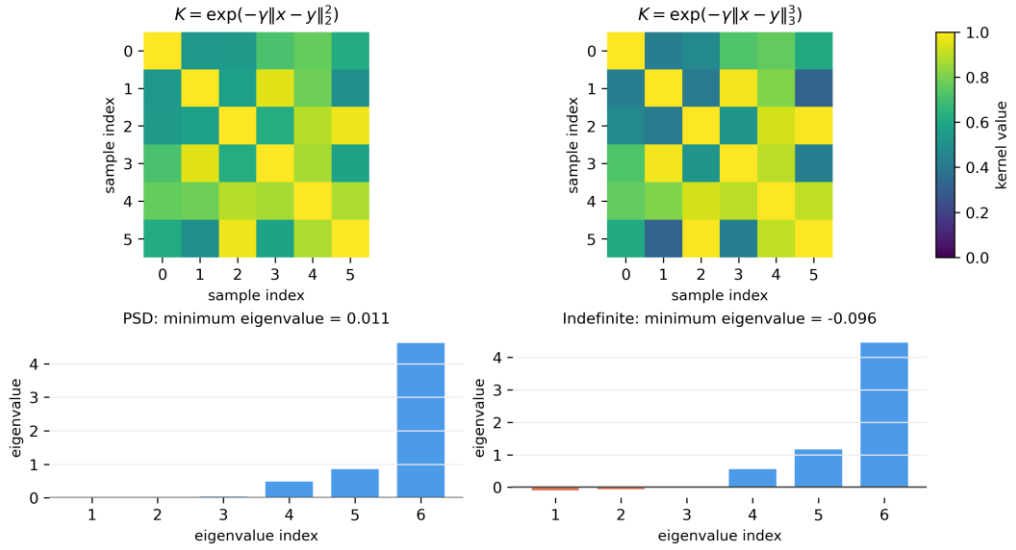
$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i\, c_j\, d(\boldsymbol{x}_i, \boldsymbol{x}_j) \le 0 \qquad \text{whenever} \qquad \sum_{i=1}^{n} c_i = 0. \tag{33}$$

Schoenberg's theorem states that d is CND exactly when

$$K_t(\boldsymbol{x}, \boldsymbol{y}) = \exp\big(-t\, d(\boldsymbol{x}, \boldsymbol{y})\big) \tag{34}$$

is PD for every $t > 0$ (Berg et al., 1984; Ressel, 1976; Schoenberg, 1938). This result has profound practical consequences: it precisely characterises which distance functions can be "kernelized" via the ubiquitous Gaussian (RBF) kernel, and explains why the RBF kernel constructed from a non-CND distance may yield an indefinite Gram matrix, rendering the kernel method ill-posed. Figure 8 illustrates this failure mode for an L_3-based radial kernel.

Figure 8. *Kernel validity depends on the distance used to construct the Gram matrix. For the same finite point set, a radial kernel built from squared Euclidean distance remains positive semi-definite, as expected from the CND property of that distance. Replacing it with an analogous construction based on an L_3 distance can produce a negative eigenvalue, showing that the resulting Gram matrix is indefinite and therefore not a valid positive definite kernel*



Which distances are conditionally negative definite?

Several canonical distances are known to be CND, and hence admit valid RBF kernelizations:

- The squared Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y}) = \| \boldsymbol{x} - \boldsymbol{y} \|_2^2$ is CND. Consequently, the standard Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\| \boldsymbol{x} - \boldsymbol{y} \|_2^2 / 2\sigma^2)$ is PD.

- For $0 < p \leq 2$, the L_p distance $\| \boldsymbol{x} - \boldsymbol{y} \|_p^p$ is CND. For $p > 2$ it is not, so the RBF kernel built on $\| \boldsymbol{x} - \boldsymbol{y} \|_p$ with $p > 2$ may not be PD.
- The geodesic distance on a sphere is CND, enabling valid kernel methods on spherical data. Related work extends Schoenberg-type characterisations of positive definite kernels on spherical and bundled domains (Kuryatnikova & Vera, 2019).
- The Hamming distance on binary strings is CND, and the corresponding RBF (diffusion) kernel is widely used in bioinformatics.

Conversely, distances that are *not* CND, such as the L_p norm for $p > 2$, or certain graph-theoretic distances, do not directly yield valid PD kernels via the RBF construction. In such cases, practitioners must either modify the distance (e.g., by taking a fractional power d^θ with $\theta \in (0,1)$, which can restore the CND property), use indefinite kernel machinery, or seek an alternative explicit feature map.

The practical implication for metric selection is clear. Choosing a distance metric is not merely a modelling decision about dissimilarity. It also determines whether the metric is compatible with the kernel machinery one intends to deploy downstream. A distance that faithfully captures domain geometry but fails the CND condition will corrupt the Gram matrix and invalidate theoretical guarantees for kernel SVMs, Gaussian processes, and related methods. Verifying or enforcing the CND property should therefore be considered an integral part of metric design in kernel-based pipelines (Berg et al., 1984; Schoenberg, 1938).

Distance Metric Learning

The distance metrics surveyed in Section 2 are largely hand-crafted, relying on domain knowledge or geometric intuition to define dissimilarity. Distance Metric Learning (DML) takes a data-driven perspective: rather than specifying a fixed metric, it *learns* a distance function from labelled or weakly labelled data such that the induced geometry reflects task-specific notions of similarity (Suárez et al., 2021). The core premise is that performance in downstream similarity-based tasks (nearest neighbour classification, clustering, etc.) is directly determined by the geometry of the distance, and that this geometry can and should be optimised jointly with or prior to the task itself.

Classical Mahalanobis metric learning

The most natural parametric family for DML is the class of Mahalanobis distances

$$D_M(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{y})}, \qquad \boldsymbol{M} \succcurlyeq 0, \tag{37}$$

where the positive semi-definite matrix $\boldsymbol{M}$ plays the role of Σ^{-1} in Equation 6. Learning $\boldsymbol{M}$ amounts to finding a linear transformation $\boldsymbol{L}$ (with $\boldsymbol{M} = \boldsymbol{L}^\top \boldsymbol{L}$) that maps the input space into a new Euclidean space in which the standard distance is task-appropriate (Goldberger et al., 2004; Weinberger & Saul, 2009).

Deep metric learning

Classical Mahalanobis methods are limited to linear transformations of the input space. Deep Metric Learning (DeepML) replaces the linear map $\boldsymbol{L}$ with a deep neural network $f_\theta: \mathcal{X} \to \mathbb{R}^d$, learning a non-linear embedding in which distances reflect semantic similarity (Hoffer & Ailon, 2015; Kaya & Bilge, 2019).

Angular margin losses

A further line of development replaces Euclidean distance in the embedding space with angular (cosine) distance and incorporates the margin directly into the classification objective. Methods such as ArcFace (Deng et al., 2022) and CosFace normalise both embeddings and weight vectors to lie on a hypersphere, then introduce an additive angular margin m into the softmax logit:

$$\mathcal{L}_{\text{Arc}} = -\log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)}+\sum_{j\neq y_i} e^{s\cos\theta_j}}, \tag{38}$$

where θ_j is the angle between the embedding and the j-th class centre, and s is a scale factor. Because the angular margin corresponds directly to a geodesic distance on the hypersphere, ArcFace enforces tighter intra-class compactness and larger inter-class separation than purely Euclidean objectives, and has become the de facto standard loss in large-scale face recognition (Deng et al., 2022).

Supervision regimes and sampling

DML methods differ not only in their loss function but also in the form of supervision they require. Pair- and triplet-based methods rely on relative constraints ("x_i is more similar to x_j than to x_k"), which are often cheaper to obtain than absolute class labels. This has motivated *self-supervised* and *contrastive representation learning* variants (such as SimCLR and MoCo) that generate positive pairs through data augmentation without any manual annotation, effectively learning a metric purely from the structure of the data (Chen et al., 2020; He et al., 2020).

Applications and broader impact

DML has achieved state-of-the-art results in face verification and recognition, person re-identification, image retrieval, few-shot learning, and recommendation systems (Kaya & Bilge, 2019). In contemporary ML, related metric choices also appear in transformer attention, where scaled dot products define token-level similarity, and in multimodal contrastive systems such as CLIP, where image and text embeddings are aligned by cosine-like objectives (Radford et al., 2021; Vaswani et al., 2017). Diffusion models likewise rely on feature-space and distributional distances for training diagnostics and evaluation, even when their generative objective is not itself a distance metric (Ho et al., 2020). Beyond performance, DML represents a conceptual shift in how distance is treated in the ML pipeline: rather than a fixed geometric assumption about the data, the metric becomes a *learned inductive bias*, shaped by the task, the data distribution, and the available supervision signal. This view reinforces the central argument of the

present paper: metric selection and design are modelling decisions with consequences for learning.

The same issue reappears in representation learning more broadly. DML explicitly trains an embedding so that distances serve a task objective. Latent-variable and generative models may also produce embeddings, but without a metric-learning objective there is no guarantee that Euclidean latent distance captures semantic similarity. This motivates the latent-space fidelity question considered next.

Latent Space Fidelity

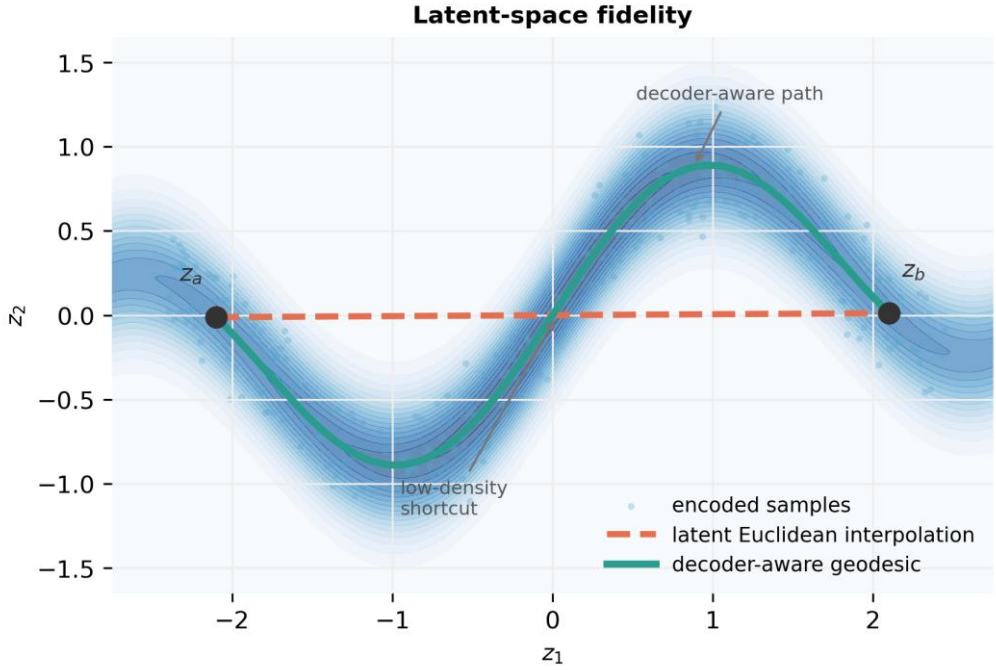
Representation learning methods such as autoencoders and variational autoencoders (VAE) compress high-dimensional observations $x \in \mathcal{X}$ into latent codes $z \in \mathcal{Z}$. A central question is whether distances in $\mathcal{Z}$ reflect meaningful semantic or geometric relationships in $\mathcal{X}$. This issue is directly relevant to downstream tasks such as interpolation, nearest-neighbour search, clustering, and anomaly detection.

The simplest assumption is that Euclidean distance in latent space is meaningful. In practice this often fails because the decoder is nonlinear, so straight lines in $\mathcal{Z}$ need not correspond to natural paths on the data manifold in observation space. As a result, latent Euclidean distance may reflect artefacts of the parametrisation rather than genuine similarity.

A principled alternative is to endow the latent space with the pullback Riemannian metric induced by the generator. Arvanitidis, Hansen, and Hauberg (Arvanitidis et al., 2018) show that this leads to geodesics that better respect the learned manifold geometry than naive Euclidean interpolation. Related work further argues that uncertainty is crucial for recovering meaningful latent geometry and that purely deterministic embeddings may distort or obscure the manifold structure (Arvanitidis et al., 2021; Hart et al., 2009). Figure 9 illustrates why a straight latent interpolation can differ from a decoder-aware geodesic path. The figure is synthetic: latent samples were drawn around a one-dimensional curved manifold $z_2 = 0.72\sin(1.75z_1) + 0.18z_1$ with additive Gaussian noise, and the background density was computed from distance to that same curve.

A separate failure mode is posterior collapse in VAEs, where the latent variables become uninformative because the decoder can model the data without using them. In that regime, latent distances lose semantic content altogether (Bowman et al., 2016; Ichikawa & Hukushima, 2024; Lucas et al., 2019). The broader lesson is that the geometry of a learned representation should not be assumed. It must be justified by the model and, when necessary, replaced by a geometry induced by the decoder or by explicit metric-preservation objectives.

Figure 9. *Latent-space fidelity in a learned representation. A straight Euclidean interpolation between latent codes z_a and z_b can pass through low-density or semantically invalid regions of the latent space. A decoder- or pullback-aware geodesic instead follows the learned high-density geometry, better preserving meaningful variation in the observation space*



Geometric Transformations and Induced Metrics

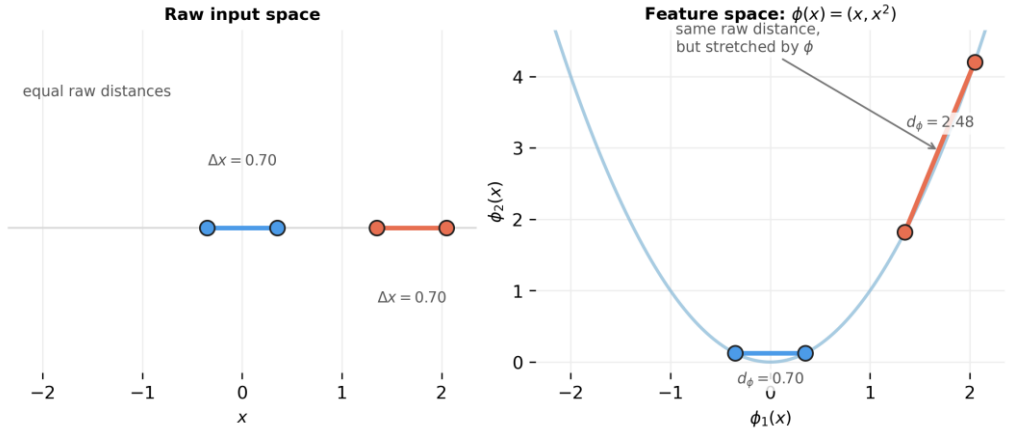
A central but often overlooked insight in machine learning is that every feature transformation implicitly defines a new distance. If data are mapped from an input space $\mathcal{X}$ into a feature space $\mathcal{F}$ by $\phi: \mathcal{X} \to \mathcal{F}$, then many algorithms compare $\phi(x)$ and $\phi(y)$ rather than x and y:

$$D_\phi(x, y) = \| \phi(x) - \phi(y) \|_2 \qquad (39).$$

Thus the modelling question is not only which distance is used, but in which representation it is used.

As a compact example, a polynomial feature map sends an input vector to monomials of its coordinates. Geometrically, this changes the notion of proximity by comparing points after nonlinear expansion rather than in the original coordinates. Two inputs that are moderately separated in the raw coordinates may become much farther apart after quadratic or higher-order terms are introduced, while other directions may be compressed. This is why a linear separator after a polynomial expansion corresponds to a nonlinear decision boundary in the original space: the algorithm is operating in the geometry of the transformed representation. Figure 10 illustrates this induced-distance effect for a simple quadratic feature map.

Figure 10. *A nonlinear feature map induces a new distance geometry. In the raw input space, two point pairs have the same Euclidean separation. After the quadratic feature map* $\phi(x) = (x, x^2)$*, Euclidean distance in feature space assigns different separations to the two pairs. The operative distance is therefore* $D_\phi(x, y) = \| \phi(x) - \phi(y) \|_2$*, not the raw input distance*



Kernel methods make the same idea implicit. A kernel $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ defines the RKHS distance

$$D_K(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}, \tag{40}$$

which may differ substantially from Euclidean distance in the input space (Schölkopf et al., 2001). For the Gaussian RBF kernel, for example, $K(x, y) = \exp(-\| x - y \|^2 / 2\sigma^2)$ maps Euclidean separation into a bounded similarity scale controlled by σ: small changes inside the bandwidth are treated as highly similar, while points beyond a few bandwidths become nearly indistinguishable from one another in terms of kernel similarity. The chosen bandwidth is therefore a geometric modelling choice, not merely a numerical hyperparameter.

Deep representations follow the same principle: contrastive, classification, and reconstruction losses can induce different notions of proximity. Even standard architectural operations such as batch normalisation and layer normalisation alter the scale and orientation of intermediate representations, and therefore change which directions in activation space are emphasised by subsequent dot products or Euclidean comparisons (Ba et al., 2016; Ioffe & Szegedy, 2015). The practical implication is straightforward: when a model transforms the data, the operative distance is the one induced by that transformation, not necessarily the Euclidean distance in the raw input coordinates.

Distances in Projective Geometry

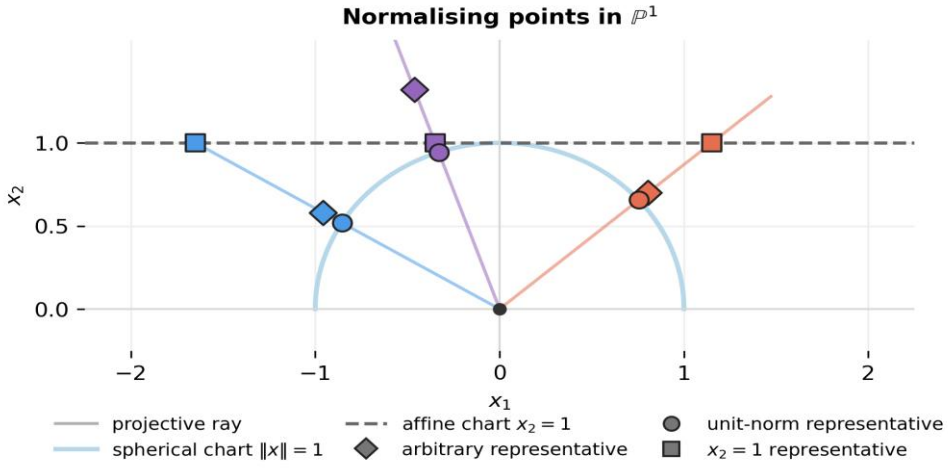
Projective representations illustrate a recurring theme of this review: the coordinates used by an algorithm are not always coordinates in which Euclidean distance is meaningful. Straight lines in 3-space, for example, can be represented by Plucker coordinates in $\mathbb{P}^5$ (De Boi et al., 2023; Pottmann et al., 1999), while planar

homographies are represented by 3×3 matrices defined only up to scale (Hartley & Zisserman, 2004). In both cases, multiplying the coordinate vector or matrix by a non-zero scalar leaves the represented geometric object unchanged.

This scale invariance makes naive Euclidean distances coordinate dependent. Normalising homogeneous coordinates by unit norm, by a chosen component such as $h_{33} = 1$, or by dehomogenisation leads to different numerical distances and may therefore change the outcome of matching, estimation, or clustering. Figure 11 illustrates this dependence on the chosen representative in the simple case of $\mathbb{P}^1$. Alternatives, such as angular distances (see also Section 2.3), geometry-aware line distances, cross-ratios (computed for 4 points rather than 2) or weighted projective geometry (in case the scale has a relevant interpretation) are often preferable when the task at hand is to compare the represented objects rather than their arbitrary coordinate representatives (Pottmann & Wallner, 2009).

In computer vision, homography estimation, line matching, and multi-view reconstruction compare objects that are naturally defined only up to projective scale. A distance on coordinate arrays can therefore disagree with the geometric error relevant to the task. Related line-geometric calibration and reconstruction settings illustrate the same point: the metric should compare the represented geometric object, not merely its coordinates (De Boi et al., 2021; De Boi et al., 2022). This section is included as a reminder that some data representations carry invariances before any learning algorithm is applied, and that those invariances should be reflected in the chosen dissimilarity. Similar representation-dependent issues arise in camera modelling and line-calculus-based optical calibration, where the operational distance is tied to the chosen geometric parametrisation rather than to raw image or actuator coordinates (De Boi et al., 2024; Penne et al., 2023).

Figure 11. *Different representatives of points in* $\mathbb{P}^1$. *Each coloured ray represents one projective point. Diamonds show arbitrary homogeneous representatives, circles spherical normalisation, and squares affine normalisation. Distances depend on the chosen representation: after affine normalisation, the distance between the blue and purple representatives is close to the distance between the purple and orange representatives, whereas after spherical normalisation these two distances become substantially different*

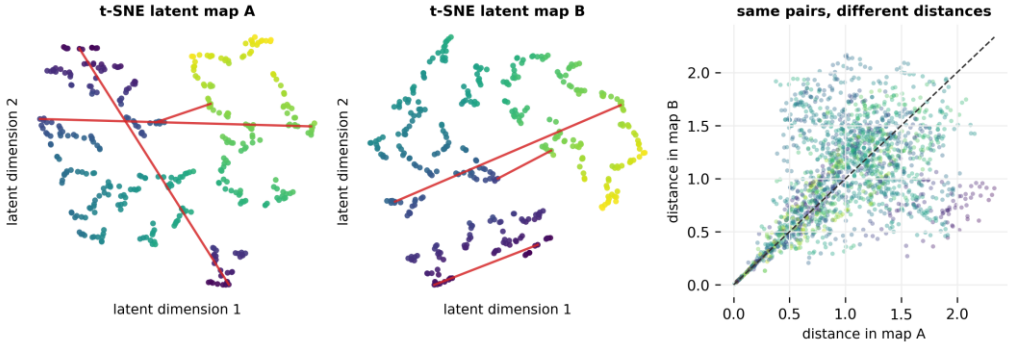


Identifiability of Distance Metrics

Identifiability asks whether geometric quantities inferred from data are uniquely determined, at least up to an accepted equivalence. In latent variable models, different latent coordinate systems may represent the same observed data distribution, so Euclidean distances between latent codes are not automatically interpretable. Related issues arise in nonlinear independent component analysis and identifiable variational autoencoders, where additional structure or auxiliary variables are needed to recover meaningful latent factors (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020).

A principled response is to focus on geometry induced by the model, such as pullback metrics and geodesic distances, rather than on raw coordinates. Recent work shows that, under suitable regularity conditions, such metric structure can be identifiable even when the latent parametrisation is not (Syrota et al., 2025). Figure 12 illustrates this issue with two t-SNE embeddings of the same high-dimensional Swiss-roll data. Although both embeddings are generated from the same observations, different random initialisations lead to different two-dimensional coordinates and different Euclidean distances between matched point pairs (Maaten & Hinton, 2008). Related uncertainty-aware extensions of multidimensional scaling reinforce the same caution for unstable or uncertain embeddings (Hagele et al., 2023).

Figure 12. *Two latent representations of the same observed high-dimensional points. The data are sampled from a Swiss-roll manifold embedded in 20 dimensions, then reduced to two dimensions by two t-SNE runs with different random initialisations. The colour encodes the same underlying manifold parameter in both maps. The right panel compares distances for the same pairs of observations in the two latent maps, showing that not every pairwise distance is preserved*



Computational Aspects of Distance Functions

Beyond their distinct use cases, the distances covered in this paper exhibit significant variation in computational cost. This is especially relevant for ML algorithms and models that are based on similarity measures, such as k-means or Gaussian processes, as they require computing distances not only between two specific points but between every possible pair of points in a potentially large

dataset. As a result, even inexpensive distances with pairwise complexity $\mathcal{O}(d)$ ultimately incur a total cost of $\mathcal{O}(N^2 d)$ when used in such models, where N is the size of the dataset or batch and d is the dimensionality of a sample. It is therefore important to understand both the cost of a single comparison and the number of times such comparisons are performed within a model. This section focuses on pairwise comparison complexities.

Providing a uniform frame for analysing the computational complexity of these distances is nearly impossible, as computational costs depend on multiple factors. The time required to compute a distance depends on the specific algorithm used, the structure and shape of the data on which it is computed, and the hardware used to perform the computations. For computationally expensive metrics specifically, the literature is constantly evolving, with new approximation methods, parallel algorithms optimised to run on GPUs, and memory optimisations. Therefore, this section only aims to provide a broad overview of computational considerations to keep in mind when choosing a metric for a specific application.

Minkowski L_p norms, cosine similarity, angular distances and Pearson distance are among the simplest distances in terms of computation. They require a single pass over the vector components, leading to an $\mathcal{O}(d)$ complexity, where d is the dimensionality of the vectors. GPU-optimised algorithms are available (Eslami et al., 2017).

Set-based metrics can be computed efficiently in $\mathcal{O}(|A| + |B|)$, where $|A|$ and $|B|$ are the cardinalities of the two sets A and B. For very large sets, this complexity can be drastically reduced using hashing algorithms such as MinHash (Broder, 1997).

Image distances that operate pixel-wise naturally incur an $\mathcal{O}(WHC)$ computational complexity, where W and H denote the width and height of the images, respectively, and C represents the number of colour channels. Because the Structural Similarity Index Measure (SSIM) relies on a sliding window, its complexity escalates to $\mathcal{O}(WHk^2)$, where k is the window's side length (Levy et al., 2025). To alleviate these computational overheads, domain-specific optimisations can be deployed, such as *Fast SSIM* (Chen & Bovik, 2011). Conversely, the Fréchet Inception Distance (FID) represents one of the most computationally demanding metrics, as its evaluation requires computing the matrix square root of a covariance product. This matrix operation scales cubically, resulting in an $\mathcal{O}(d^3)$ complexity, where d is the dimensionality of the underlying Inception-v3 latent embeddings.

Metrics that involve a minimisation problem often require more complex computations. Edit distances can be computed using a dynamic programming algorithm in $\mathcal{O}(l^2)$ time, where l is the length of the strings. DTW computation also has quadratic complexity, although constraints such as the Sakoe–Chiba band (Sakoe & Chiba, 1978) can provide cheaper lower-bound approximations in practice. Chamfer distance, as a sum of nearest-neighbour distances, can also be computed in $\mathcal{O}(N^2)$, where N is the number of samples.

For EMD, the optimisation problem involves a computation of complexity $\mathcal{O}(n^3 \log n)$, where n is the number of bins. This complexity can be reduced when working with one-dimensional data, or by using one of many approximation

methods (Shirdhonkar & Jacobs, 2008), such as the Sinkhorn algorithm (Cuturi, 2013).

As seen with the FID, metrics that involve covariance matrices often incur higher computational costs. The Mahalanobis distance requires a Cholesky decomposition of the covariance matrix, implying an $\mathcal{O}(d^3)$ preprocessing cost, where d is the dimensionality of the vectors. However, this decomposition only needs to be computed once and can then be reused for subsequent comparisons, reducing the complexity of each comparison to $\mathcal{O}(d^2)$ due to the matrix–vector multiplication.

Measures for probability distributions provide good examples of the difficulty of summarising the complexity of a metric with a single figure. In the case of discrete distributions, these measures can often be computed in linear time. Computing them on continuous distributions typically involves numerical integration, making them intractable in many cases (Pérez-Cruz, 2008). Recent literature has explored estimation methods based on k-nearest-neighbours to approximate the KL divergence (Zhao & Lai, 2020). On the other hand, restricting the problem to specific distributions (e.g., Gaussian distributions) often leads to analytical solutions that can be computed efficiently. For example, the KL divergence between two d-dimensional Gaussian distributions can be computed in $\mathcal{O}(d^3)$ time for full covariance matrices, but this reduces to $\mathcal{O}(d)$ for isotropic Gaussians.

Practical Synthesis: Choosing a Metric

The central practical question is not which distance is best in the abstract, but which assumptions about the data and task should be made explicit. Table 1 summarises choices that are mainly driven by the data representation, while Table 2 summarises choices that are mainly driven by modelling or theoretical constraints.

A hand-crafted metric is often appropriate when the data type already suggests a clear invariance: coordinate-wise comparison for vectors, covariance-aware comparison for correlated features, angular comparison for embeddings, projective normalisation for homogeneous coordinates, Chamfer or EMD for point clouds, divergence or transport distances for distributions, diagram distances for topological summaries, DTW for temporally misaligned sequences, Jaccard or Hamming distance for discrete objects, perceptual or feature-based criteria for images and videos, and geodesic or diffusion distances for manifold-like data. These choices are attractive because their assumptions are transparent, but they can fail when the task-specific notion of similarity differs from the generic geometry of the data type.

Metric learning becomes preferable when supervision or reliable weak supervision is available and the goal is predictive performance in a specific task. Kernel-compatible distances become important when the chosen dissimilarity will be used to form a positive definite kernel. Pullback and induced metrics become preferable when the central assumption is geometric rather than label-based: similarity should follow a feature map, a learned representation, or a decoder-induced surface. Common mistakes are to ignore feature scaling, apply Euclidean

distance after a nonlinear transformation without asking what geometry was induced, or use a computationally expensive metric without checking whether its additional structure changes the downstream decision.

Table 1. *Compact guide for selecting representation-driven metrics or dissimilarity measures*

Data or task	Useful choice	Metric status	Main caution
Scaled vectors	Euclidean, Manhattan, Chebyshev, Minkowski	Yes for $p \geq 1$, no for $p < 1$ because triangle inequality fails	Sensitive to scaling
Correlated features	Mahalanobis / learned PSD metric	Yes if matrix is positive definite, pseudo-metric if only semi-definite	Requires stable estimation
Sparse text or embeddings	Cosine similarity / angular distance	Cosine similarity is not a distance, angular distance is a metric	Scale information is removed
Projective data	Normalisation-aware projective distance	Conditional, coordinate distances depend on the chosen normalisation	Coordinates depend on gauge
Point clouds	Chamfer / EMD	Chamfer is not a metric because triangle inequality can fail, EMD is a metric under standard transport assumptions	Local and global costs differ
Distributions	KL, JS, Hellinger, Wasserstein	KL is not, JS is not unless square-rooted, Hellinger and Wasserstein are metrics	Support and symmetry differ
Persistence diagrams	Bottleneck / diagram Wasserstein	Yes on persistence diagrams with diagonal matching	Sensitive to filtration
Time series	Dynamic Time Warping	No, triangle inequality can fail	Can over-warp sequences
Sets or binary data	Jaccard / Hamming	Jaccard distance and Hamming distance are metrics	Ignores feature dependence
Strings or discrete ordered sequences	Damerau–Levenshtein / LCS / Jaro	Levenshtein is a metric. Damerau–Levenshtein and LCS-based distances depend on the precise variant, while Jaro relaxes the triangle inequality	Purely character-based, no grammar or semantic encoded
Images or videos	MSE, PSNR, NCC, SSIM, MI, FID, FVD	Root-MSE is a metric, most others are not because symmetry, identity, or triangle inequality fail	Encoder bias can dominate
Manifold-like data	Geodesic / diffusion distance	Usually yes when the graph or manifold construction is valid	Graph construction is fragile

Table 2. *Compact guide for model-dependent and theoretical metric choices.*

Data or task	Useful choice	Metric status	Main caution
Kernel methods	CND-compatible distances	Conditional, CND guarantees kernel validity but not all metric axioms	Indefinite Gram matrices
Supervised embeddings	Learned metric	Conditional, depends on whether the learned form enforces metric axioms	May overfit supervision
Latent generative models	Pullback / decoder metric	Conditional, valid when the induced geometry is regular	Euclidean latent distance can mislead
Feature-transformed data	Induced feature-space metric	Metric if the feature map is injective, otherwise pseudo-metric	Raw-space intuition may fail
Identifiability questions	Geometry-level comparison	Not a single metric, but a check on whether the distance is uniquely determined	Identifiability must be checked

The tables should be read as decision aids rather than rankings. The most reliable choice is usually the simplest metric whose invariances match the scientific or engineering question. More complex learned or geometric distances are justified when they encode information that the simpler metric cannot.

Conclusions

This review has highlighted the diversity and importance of distance metrics and dissimilarity measures in modern Machine Learning. Moving beyond Euclidean distance is not merely a matter of replacing one formula by another. It changes the geometry in which learning takes place, the invariances encoded by the model, and the meaning of similarity used by downstream algorithms.

Across the examples considered here, a common pattern emerges: each metric becomes appropriate only under particular assumptions about the data. When those assumptions are violated, the resulting notion of distance may be computationally convenient but semantically misleading.

The theoretical perspectives discussed in this paper reinforce the same point. In modern ML pipelines, the operative metric is often induced by a feature map, a learned embedding, a kernel, or a decoder-defined geometry rather than chosen directly in the raw input space. Open problems remain in making metric choice more auditable, especially for large foundation models, multimodal embeddings, and generative models whose learned spaces are difficult to interpret. For both researchers and practitioners, the main message is straightforward: distance is a modelling choice, and it should be justified with the same care as architecture, loss, or prior.

References

Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 1973

Arriagada RC, Apablaza RC, Pinninghoff MA (2019) Efficient Edge Detection by Adapting Artificial Bee Colony Algorithm. Athens Journal of Technology and Engineering 6(3) : 179–194

Arvanitidis G, Hansen LK, Hauberg S (2018) Latent Space Oddity: On the Curvature of Deep Generative Models. 6th international conference on learning representations

Arvanitidis G, Hauberg S, Schölkopf B (2021) Geometrically Enriched Latent Spaces. Proceedings of the 24th international conference on artificial intelligence and statistics 130 : 631–639

Ba JL, Kiros JR, Hinton GE (2016) Layer Normalization. arXiv preprint arXiv:1607.06450

Bedalli E, Ninka I (2015) Exploring an Educational System's Data through Fuzzy Cluster Analysis. Athens Journal of Sciences 2(1) : 33–44

Berg C, Christensen JPR, Ressel P (1984) Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions. Springer

Berlinet A, Thomas-Agnan C (2004) Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers

Berndt D, Clifford J (1994) Using Dynamic Time Warping to Find Patterns in Time Series. Workshop on knowledge discovery in databases 398

Bookstein A, Kulyukin VA, Raita T (2002) Generalized Hamming Distance. Information Retrieval 5(4)

Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S (2016) Generating Sentences from a Continuous Space. Proceedings of the 20th SIGNLL conference on computational natural language learning : 10–21

Broder AZ (1997) On the resemblance and containment of documents. Proceedings. Compression and complexity of SEQUENCES 1997 (cat. no.97TB100171) : 21–29

Chan DA, Sithungu SP (2025) Evaluating the Suitability of Inception Score and Frechet Inception Distance as Metrics for Quality and Diversity in Image Generation. Proceedings of the 7th international conference on computational intelligence and intelligent systems

Chazal F, Michel B (2021) An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. Frontiers in Artificial Intelligence 4

Chen M-J, Bovik AC (2011) Fast structural similarity index algorithm. Journal of Real-Time Image Processing 6(4) : 281–287

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th international conference on machine learning 119 : 1597–1607

Coifman RR, Lafon S (2006) Diffusion Maps. Applied and Computational Harmonic Analysis 21(1) : 5–30

Cover TM, Thomas JA (2006) Elements of Information Theory. Wiley-Interscience

Crane K, Weischedel C, Wardetzky M (2013) Geodesics in heat: A new approach to computing distance based on heat flow. ACM Transactions on Graphics 32(5)

Cuturi M (2013) Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems : 2292–2300

De Boi I, Ek CH, Penne R (2023) Surface Approximation by Means of Gaussian Process Latent Variable Models and Line Element Geometry. Mathematics 11(2)

De Boi I, Pathak S, Oliveira M, Penne R (2024) How to Turn Your Camera into a Perfect Pinhole Model. Progress in pattern recognition, image analysis, computer vision, and applications : 90–107

De Boi I, Sels S, De Moor O, Vanlanduit S, Penne R (2022) Input and Output Manifold Constrained Gaussian Process Regression for Galvanometric Setup Calibration. IEEE Transactions on Instrumentation and Measurement 71 : 1–8

De Boi I, Sels S, Penne R (2021) Semidata-Driven Calibration of Galvanometric Setups Using Gaussian Processes. IEEE Transactions on Instrumentation and Measurement 71 : 1–8

Deng J, Guo J, Yang J, Xue N, Kotsia I, Zafeiriou S (2022) ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(10)

Deza MM, Deza E (2016) Encyclopedia of Distances. Springer Berlin Heidelberg

Enesi I, Kuqi A (2023) Evaluation of the 3D Reconstruction Performance of Objects in Meshroom: A Case Study. Athens Journal of Technology and Engineering 10(1) : 49–70

Eslami T, Awan MG, Saeed F (2017) GPU-PCC: A GPU based technique to compute pairwise pearson's correlation coefficients for big fMRI data. Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics : 723–728

Ghosh A, Ghosh AK, SahaRay R, Sarkar S (2025) Classification using global and local Mahalanobis distances. Journal of Multivariate Analysis 207

Goldberger J, Hinton GE, Roweis S, Salakhutdinov RR (2004) Neighbourhood Components Analysis. Advances in neural information processing systems 17

Gorjian M, Caffey SM, Luhan GA (2025) Exploring Architectural Design 3D Reconstruction Approaches through Deep Learning Methods: A Comprehensive Survey. Athens Journal of Sciences 12(3) : 205–234

Hagele D, Krake T, Weiskopf D (2023) Uncertainty-Aware Multidimensional Scaling. IEEE Transactions on Visualization and Computer Graphics 29(1)

Hart GL, Zach C, Niethammer M (2009) Only Bayes Should Learn a Manifold. 2009 IEEE conference on computer vision and pattern recognition : 1206–1213

Hartley RI, Zisserman A (2004) Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518

He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum Contrast for Unsupervised Visual Representation Learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition : 9729–9738

Hellinger E (1909) Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik 136 : 210–271

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Advances in neural information processing systems 30

Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. Advances in neural information processing systems 33 : 6840–6851

Hoffer E, Ailon N (2015) Deep metric learning using triplet network. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 9370

Huang T, Liu Q, Zhao X, Chen J, Liu Y (2024) Learnable Chamfer Distance for Point Cloud Reconstruction. Pattern Recognition Letters 178

Huh M-H (2015) Kernel-Trick Regression and Classification. Communications for Statistical Applications and Methods 22(2)

Hyvärinen A, Morioka H (2016) Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. Advances in neural information processing systems 29

Ichikawa Y, Hukushima K (2024) Learning Dynamics in Linear VAE: Posterior Collapse Threshold, Superfluous Latent Space Pitfalls, and Speedup with KL Annealing. Proceedings of machine learning research 238

Ioffe S, Szegedy C (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd international conference on machine learning 37 : 448–456

Jost L (2006) Entropy and Diversity. Oikos 113(2)

Kaya M, Bilge HŞ (2019) Deep Metric Learning: A Survey. Symmetry 11(9)

Khemakhem I, Kingma DP, Monti RP, Hyvärinen A (2020) Variational Autoencoders and Nonlinear ICA: A Unifying Framework. Proceedings of the 23rd international conference on artificial intelligence and statistics 108 : 2207–2217

Kullback S, Leibler RA (1951) On Information and Sufficiency. The Annals of Mathematical Statistics 22(1) : 79–86

Kuryatnikova O, Vera JC (2019) Generalizations of Schoenberg's Theorem on Positive Definite Kernels. arXiv

Lee W, Li W, Lin B, Monod A (2022) Tropical Optimal Transport and Wasserstein Distances. Information Geometry 5(1)

Levy A, Shalom BR, Chalamish M (2025) A guide to similarity measures and their data science applications. Journal of Big Data 12(188)

Lewis JP (1995) Fast Normalized Cross-Correlation. Vision interface : 120–123

Lin J (1991) Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1) : 145–151

Liu P, Na J (2025) Word Motifs and a Generalized Hamming Distance. Contemporary Mathematics 6(1)

Lucas J, Tucker G, Grosse R, Norouzi M (2019) Understanding Posterior Collapse in Generative Latent Variable Models. Deep generative models for highly structured data, ICLR workshop

Maaten L van der, Hinton G (2008) Visualizing Data Using t-SNE. Journal of Machine Learning Research 9 : 2579–2605

Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P (1997) Multimodality Image Registration by Maximization of Mutual Information. IEEE Transactions on Medical Imaging 16(2) : 187–198

McKeown S (2025) Beyond Hamming Distance: Exploring Spatial Encoding in Perceptual Hashes. Forensic Science International: Digital Investigation 52

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781

Penne R, De Boi I, Vanlanduit S (2023) On the Angular Control of Rotating Lasers by Means of Line Calculus on Hyperboloids. Sensors 23(13)

Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. EMNLP 2014 - 2014 conference on empirical methods in natural language processing, proceedings of the conference

Pérez-Cruz F (2008) Kullback-leibler divergence estimation of continuous distributions. 2008 IEEE international symposium on information theory : 1666–1670

Peyré G, Cuturi M (2019) Computational Optimal Transport. Foundations and Trends in Machine Learning 11(5–6) : 355–607

Pottmann H, Peternell M, Ravani B (1999) An introduction to line geometry with applications. Comput. Aided Des. 31 : 3–16

Pottmann H, Wallner J (2009) Computational line geometry. Springer Science & Business Media

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th international conference on machine learning 139 : 8748–8763

Rainio O, Klén R (2025) Modified Dice Coefficients for Evaluation of Tumor Segmentation from PET Images: A Proof-of-Concept Study. Journal of Imaging Informatics in Medicine

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. Proceedings of the 2019 conference on empirical methods in natural language processing : 3982–3992

Ressel P (1976) A Short Proof of Schoenberg's Theorem. Proceedings of the American Mathematical Society 57(1)

Rodrigues M, Kormann M, Al Dulaimi M (2016) Data Protection and Privacy Issues Concerning Facial Image Processing in Public Spaces. Athens Journal of Technology and Engineering 3(1) : 39–52

Romanazzi M (2014) Discriminant Analysis with High Dimensional von Mises–Fisher Distribution. Athens Journal of Sciences 1(4) : 225–240

Sakoe H, Chiba S (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26(1) : 43–49

Schoenberg IJ (1938) Metric Spaces and Positive Definite Functions. Transactions of the American Mathematical Society 44(3) : 522–536

Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the Support of a High-Dimensional Distribution. Neural Computation 13(7) : 1443–1471

Sethian JA (1996) A fast marching level set method for monotonically advancing fronts. Proceedings of the National Academy of Sciences of the United States of America 93(4)

Shirdhonkar S, Jacobs DW (2008) Approximate earth mover's distance in linear time. 2008 IEEE conference on computer vision and pattern recognition : 1–8

Suárez JL, García S, Herrera F (2018) A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms and Software. arXiv:1812.05944 [cs.LG]

Suárez JL, García S, Herrera F (2021) A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. Neurocomputing 425

Syrota S, Zainchkovskyy Y, Xi J, Bloem-Reddy B, Hauberg S (2025) Identifying Metric Structures of Deep Latent Variable Models. Proceedings of machine learning research 267

Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500)

Unterthiner T, Steenkiste S van, Kurach K, Marinier R, Michalski M, Gelly S (2018) Towards Accurate Generative Models of Video: A New Metric and Challenges.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. Advances in neural information processing systems 30

Verma V, Aggarwal RK (2020) A Comparative Analysis of Similarity Measures Akin to the Jaccard Index in Collaborative Recommendations: Empirical and Theoretical Perspective. Social Network Analysis and Mining 10(1)

Villani C (2009) Optimal Transport: Old and New. Springer 338

Viola P, Wells WM (1997) Alignment by Maximization of Mutual Information. International Journal of Computer Vision 24(2) : 137–154

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing 13(4) : 600–612

Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10

Ye Q (2026) An Automatic Scoring Method for Vocal Singing Based on Approximate Matching Algorithm. International Journal of High Speed Electronics and Systems 35(3)

Zhang C, Ji D (2026) HAL-Net: A Historical Analogy Learning Network for Adaptive and Interpretable Pandemic Forecasting. Expert Systems with Applications 299

Zhang Y, Skolnick J (2004) Scoring Function for Automated Assessment of Protein Structure Template Quality. Proteins: Structure, Function, and Bioinformatics 57(4) : 702–710

Zhao P, Lai L (2020) Minimax optimal estimation of KL divergence for continuous distributions. IEEE Transactions on Information Theory 66(12) : 7787–7811

Zhou J, Hodge K, Dong W, Tamakloe E (2025) Distribution-Aware Outlier Detection in High Dimensions: A Scalable Parametric Approach. Mathematics 14(1)

Zugan D, Požar R, Brodnik A (2025) Floyd-Warshall Algorithm for Sparse Graphs. Algorithms 18(12)