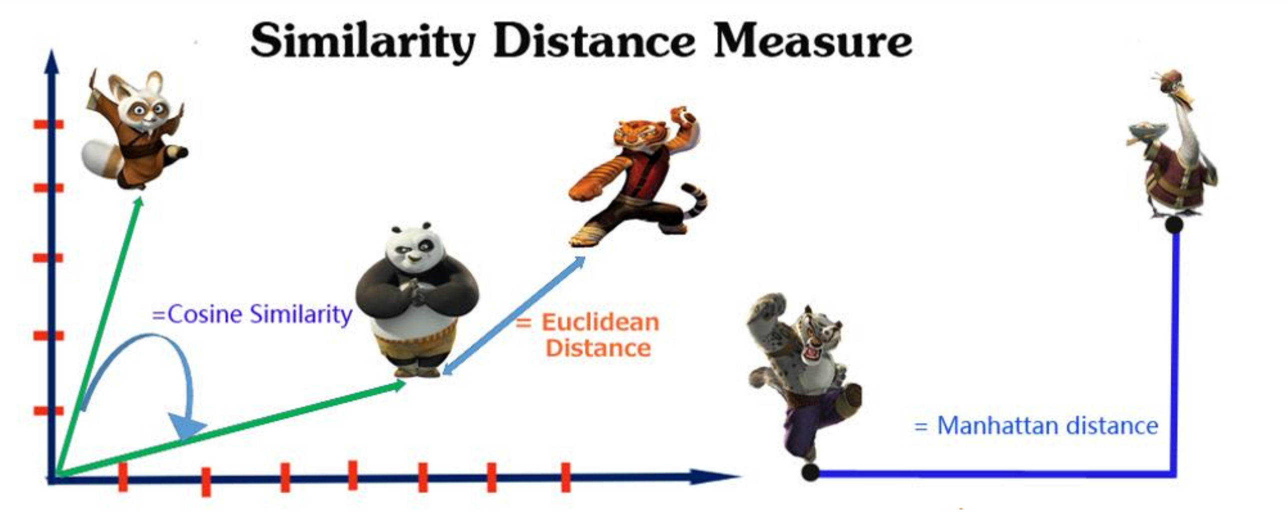
Euclidian Vs Manhattan Vs Cosine Distance

The choice between **Cosine**, **Euclidean**, and **Manhattan** distance depends on the nature of your data and the problem you're trying to solve. Each metric measures distance or similarity differently, making it suitable for specific applications.



When to Use Cosine Distance

Use **Cosine distance** when the **orientation** of the vectors is more important than their **magnitude**. This is especially true for high-dimensional data where the absolute values of the features are less meaningful than their relative proportions.

- **Preferred for:** Text analysis, document similarity, and recommender systems.
- **Why:** It measures the angle between two vectors, making it insensitive to the size or length of the data. For example, two documents on the same topic will have vectors pointing in a similar direction, even if one is much longer than the other. This prevents a longer document from being considered "further away" simply because it has a greater number of words.

When to Use Euclidean Distance

Use **Euclidean distance** when the **magnitude** of the vectors is a significant factor. It measures the straight-line distance between two points and is the most common and intuitive distance metric.

- **Preferred for:** Low-dimensional data, physical measurements, and algorithms where the "shortest path" is the most relevant concept.
- **Why:** It is best when the features are on a similar scale and are independent. It is a good default choice for many machine learning algorithms like **K-Means clustering** and **K-Nearest Neighbors (K-NN)**, especially when the data points are not too dense. It is also good for applications where the difference in each dimension is equally important.

When to Use Manhattan Distance

Use **Manhattan distance** when movement is restricted to a grid or when you need to emphasize differences in individual dimensions over their cumulative effect. It measures the distance by summing the absolute differences of the coordinates.

- **Preferred for:** Urban planning, grid-based pathfinding, and high-dimensional data with a large number of features.
- **Why:** It is more robust to outliers and is less sensitive to the **"curse of dimensionality"** than Euclidean distance. In high-dimensional spaces, Euclidean distance can become less meaningful as the distance between points tends to converge, a problem that Manhattan distance handles better. It is also useful when the cost of movement is linear rather than quadratic, such as in logistics and warehouse management where travel is restricted to a grid.