When dealing with plenty of data, we will more or less consider about the similarity for these data. Usually, such kind of similarity is based on distance.

There're 3 measures for similarity distance: Minkowski distance, Cosine Distance, Pearson Correlation

Minkowski Distance formula:

$$Minkowski\ distance = (\sum|x_i - y_i|^p)^{\frac{1}{p}}$$

[$p = 1, Manhattan\ Distance$

$p = 2, Euclidean\ Distance$

$p = \infty, Chebychev\ Distance$]

Cosine Distance:

$$cosine\ similarity\ = \cos(\theta) = \frac{AB}{|A||B|}$$

Pearson Correlation Coefficient:

$$Pearson\ Corr\ Coef\ =\ \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

When dealing with large dataset, the similarity problem will be transformed into similarity matrix.

Similarity matrices are broadly used in data analytics field, for example, Clustering (i.e. k-means), Classification (i.e. kNN), Dimensionality Reduction (i.e. PCA), Recommender System (i.e. Collaborative Filtering), NLP, etc.