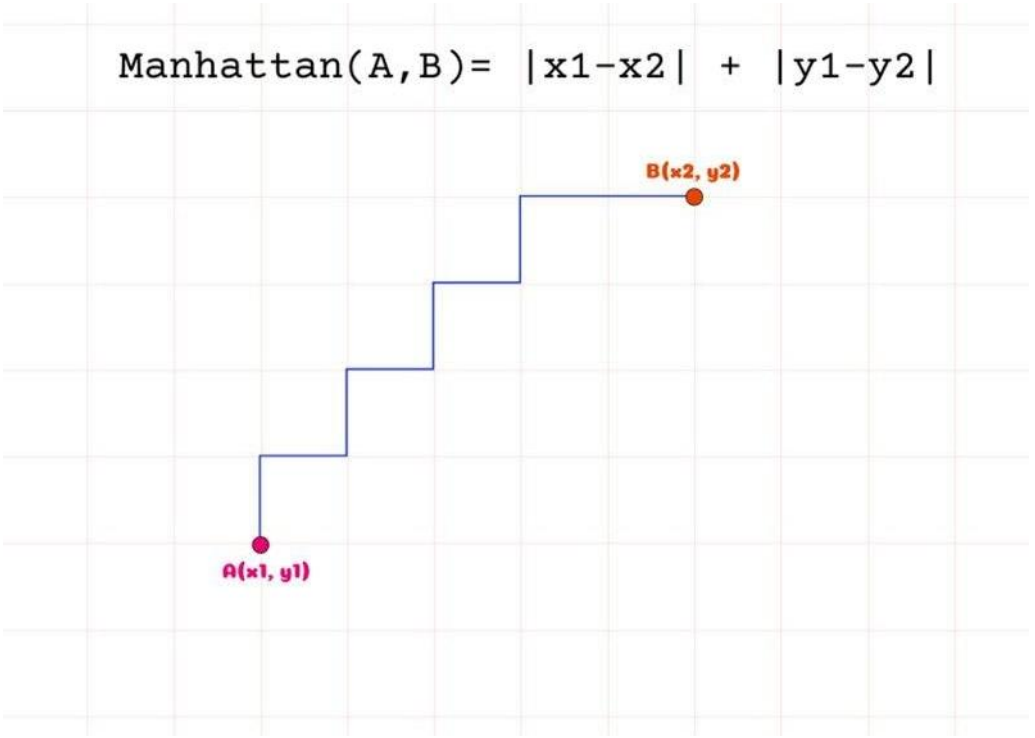
Measuring distance between datapoints

OLA 2

Christoffer Drejer Mikkelsen

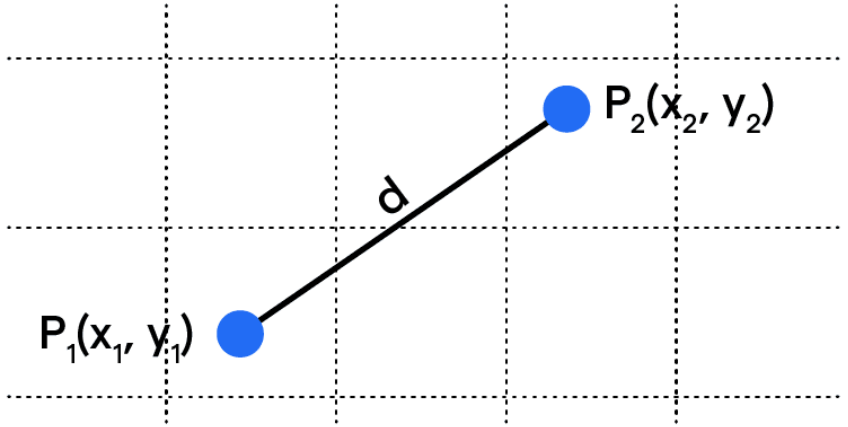
Manhattan Plot

- Also known as taxi cab distance
- Calculates distance in a grid similar as to how a taxi cab would be driving in a city
- Formula: $d_{\mathrm{T}}(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_{\mathrm{T}} = \sum_{i=1}^{n} |p_i - q_i|$
- Example: Point A: (1,6), Point B: (8,3)
- $|1 - 8| + |6 - 3| = 7 + 3 = 10$



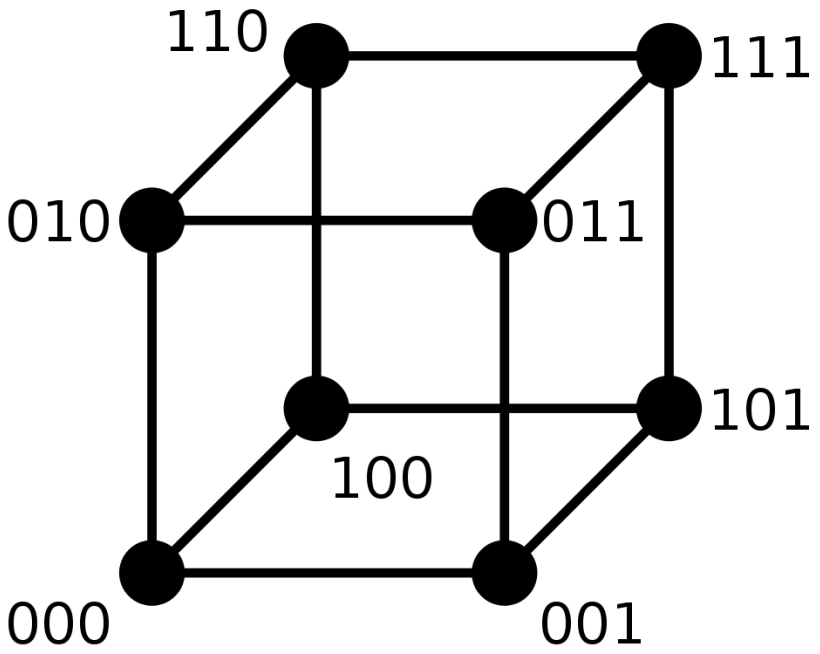
Euclidean Distance

- Also known as L2 norm
- Straight line distance between two points
- Formula: $d(p, q) = |p - q|$.
- Example: Point A: (1,6), Point B: (8,3)
- $\sqrt{(1-8)^2 + (6-3)^2} = \sqrt{49+9} = \sqrt{58}$



Hamming distance

- Calculates the minimum number of substitutions required to change one string into another of equal length
- Example: Consider two binary strings: "1101" and "1010“
- Hamming Distance = Number of positions with different symbols = 2 (the 2nd and 3rd positions)



Why are there different measures?

- Manhattan Distance is often used in circuit design, image processing, and robot navigation, where movements are restricted to certain directions or paths.
- Euclidean Distance is widely used in various fields such as machine learning (e.g., clustering algorithms like K-means), computer vision, and statistics.
- Hamming Distance is crucial in error detection and correction codes, DNA sequence analysis, and cryptography, where it helps measure the similarity or difference between two binary strings.