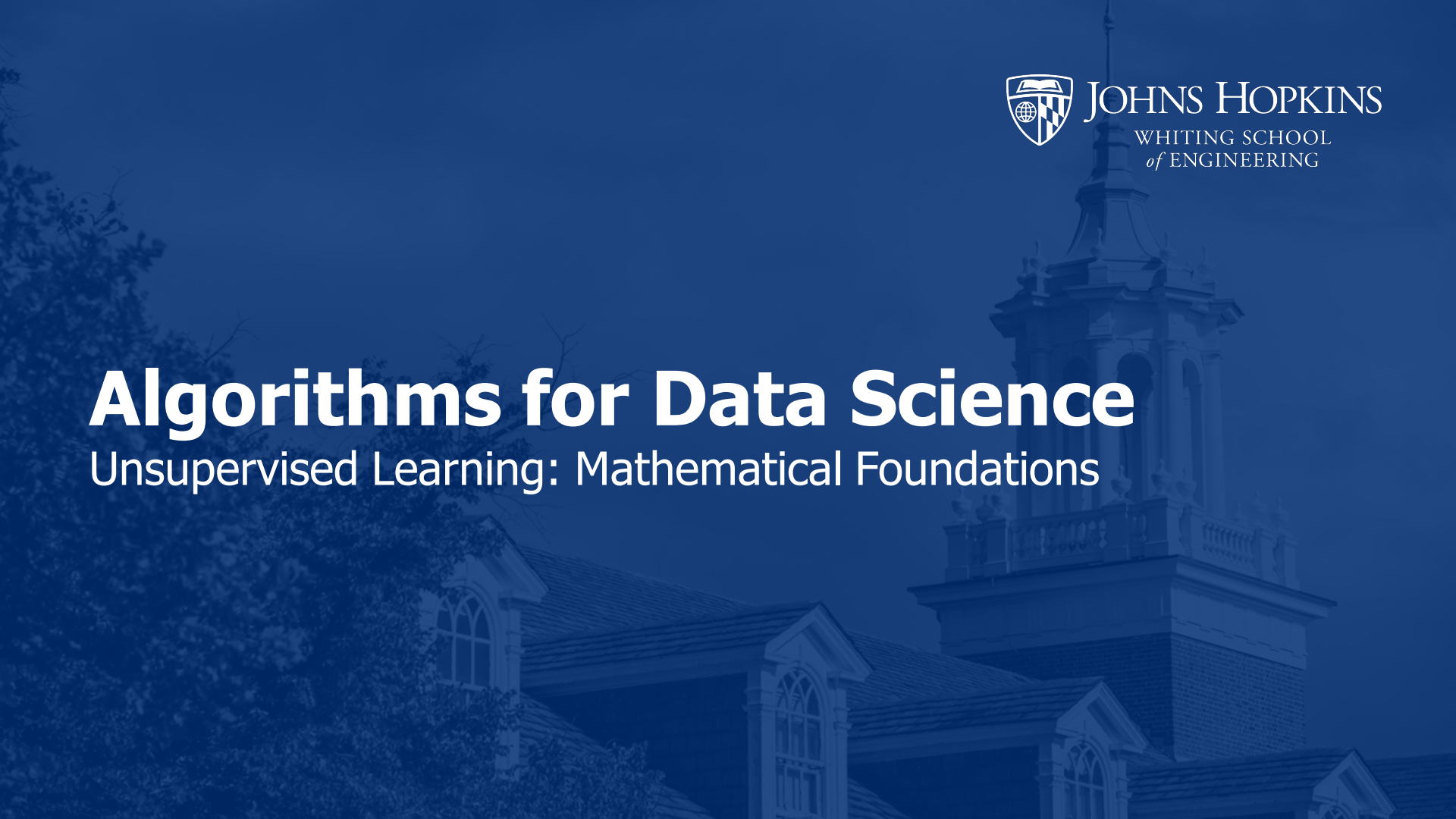


Algorithms for Data Science

Unsupervised Learning: Mathematical Foundations

Mathematics of Unsupervised Learning

Mathematics provides the foundation for algorithms to process and analyze data.

Linear Algebra	Probability & Statistics	Geometry
▪ Vectors	▪ Covariance	▪ Distance Metrics
▪ Matrices	▪ Variance	○ Euclidean Distance
▪ Eigenvalues & Eigenvectors	▪ Distributions	○ Manhattan Distance

Covariance and Variance

- Covariance
  - Measures how two variables vary together:

$$\mathrm{Cov}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

  - Variance
    - Special case of covariance for a single variable:

$$Var(X) = Cov(X,X)$$

Used in Principal Component Analysis (PCA) and Clustering

Distance Metrics

Distance metrics define the similarity between data points in clustering.

- Euclidean Distance
  - Straight-line distance between two points:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan Distance
  - Distance Measured along axes at rights angles:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Distance Metrics

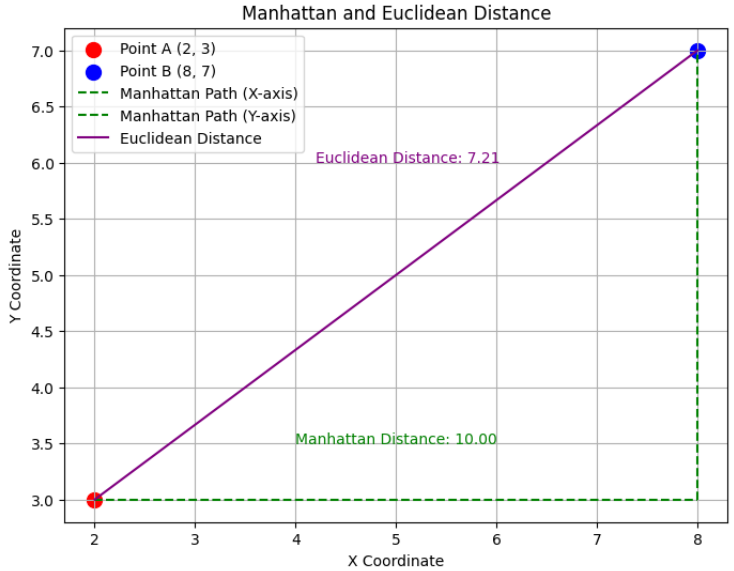
Distance metrics define the similarity between data points in clustering.

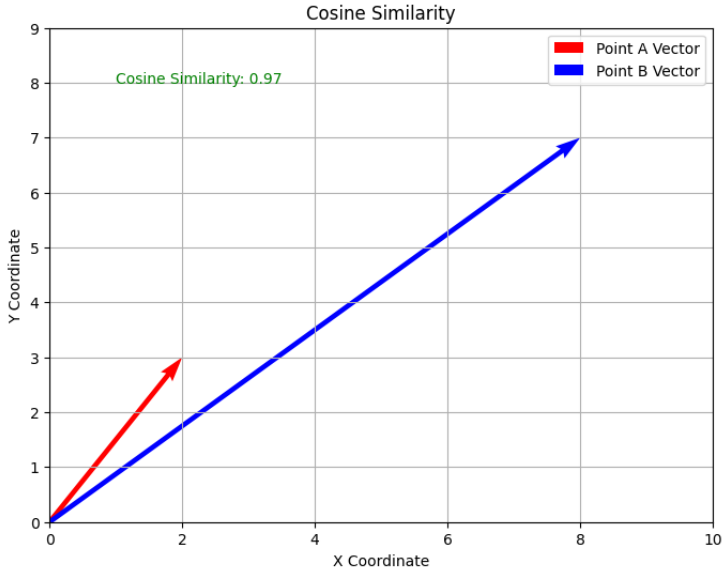
- Cosine Similarity
  - Measures angle between vectors, ignoring magnitude:

$$\text{CosSim} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

Distance Metrics

Distance metrics define the similarity between data points in clustering.





Eigenvalues and Eigenvectors

- Eigenvalues are scalars representing the variance captured by eigenvectors.
- Eigenvectors are the directions of maximum variance in the data.

$$Av = \lambda v$$

Where:

A is a matrix

v represents an eigenvector

λ represents an eigenvalue