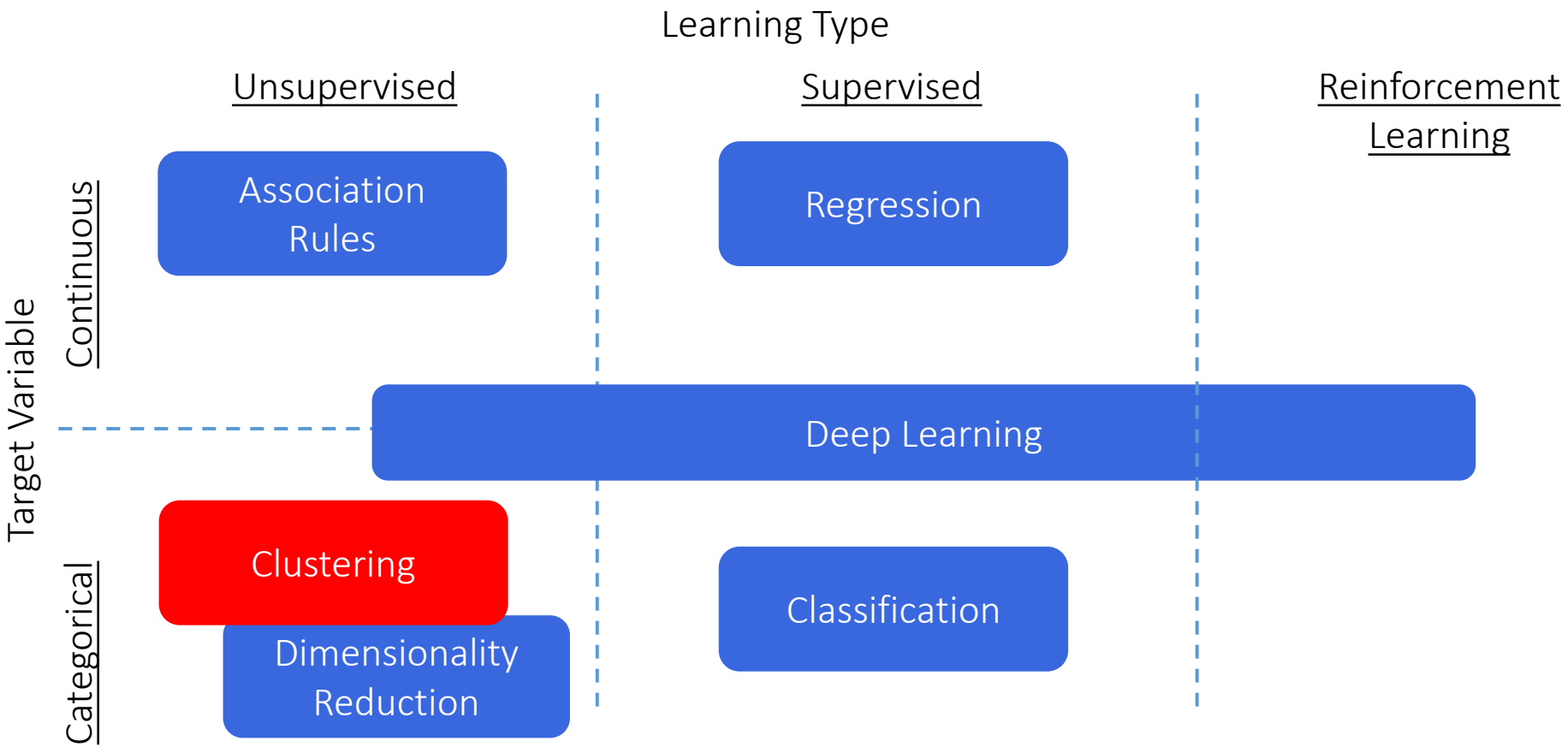
Clustering 101

Clustering

Machine Learning Types



Clustering

Introduction

- Unsupervised learning technique – no labels available
- Classify data points into groups
- Data within same group have similar properties
- Data within different groups have dissimilar properties
- Used for getting insights rather than prediction
- Provides classes / clusters without intrinsic value
  → you need to define a useful label

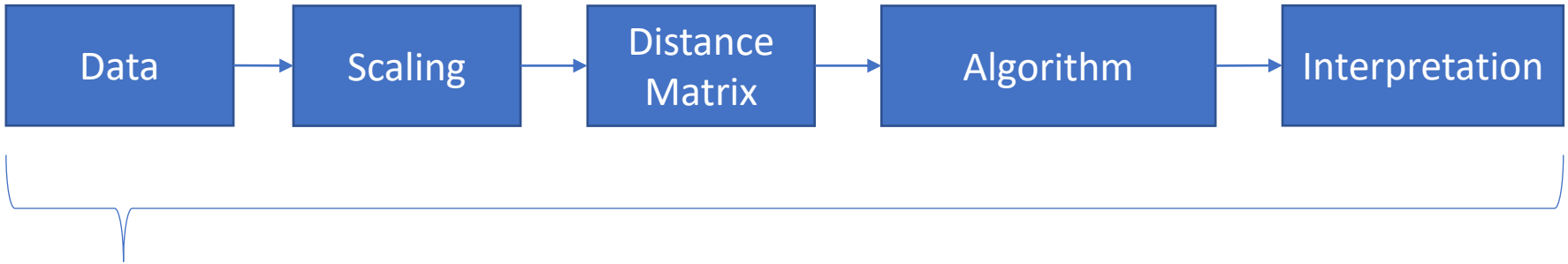
Clustering

Usecases

- Marketing
  - Focuses on data within clusters
  - customer and buying patterns
- Outlier detection
  - Focuses on data not within clusters
  - Cyber security (network intrusion)
- Dimensionality Reduction
  - Reduce number of features

Clustering

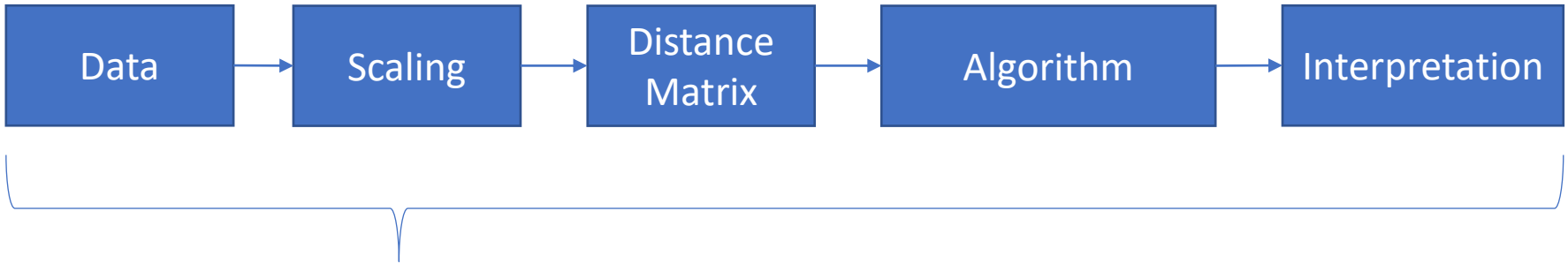
Workflow



- Variables can be all continuous or all categorical.
- Both types cannot be combined in one analysis.

Clustering

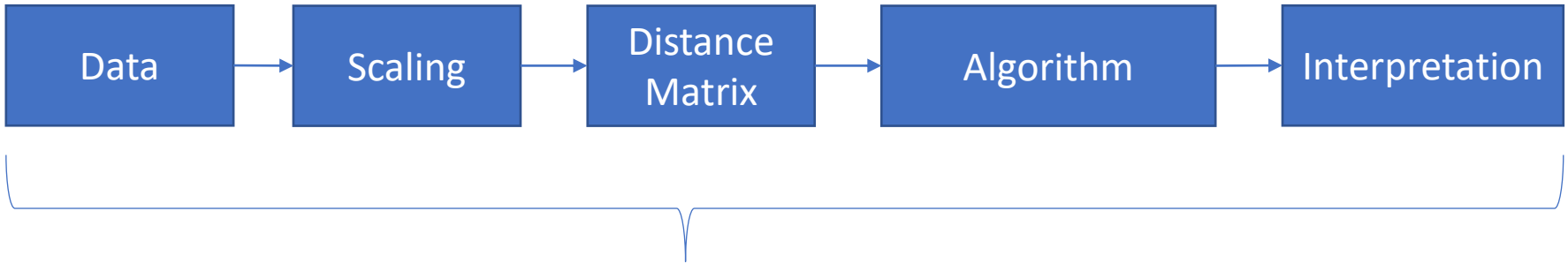
Workflow



- Scale data to avoid different scales

Clustering

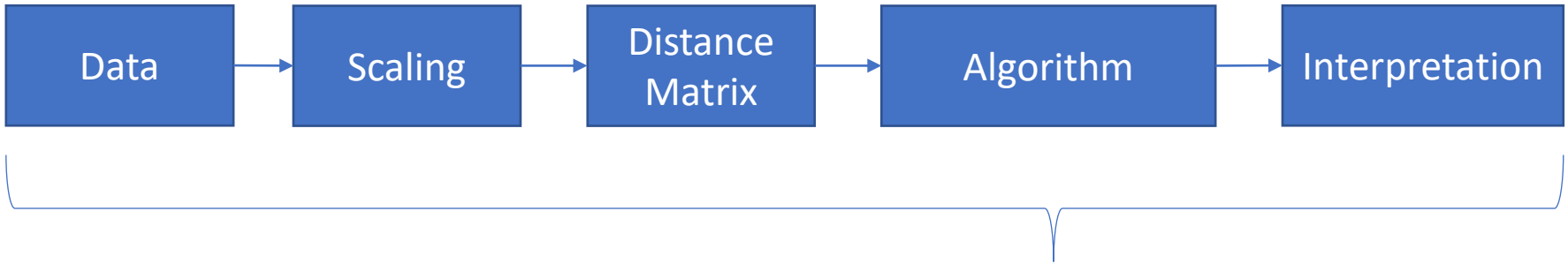
Workflow



- Distance matrix for continuous variables
- Similarity matrix for categorical variables

Clustering

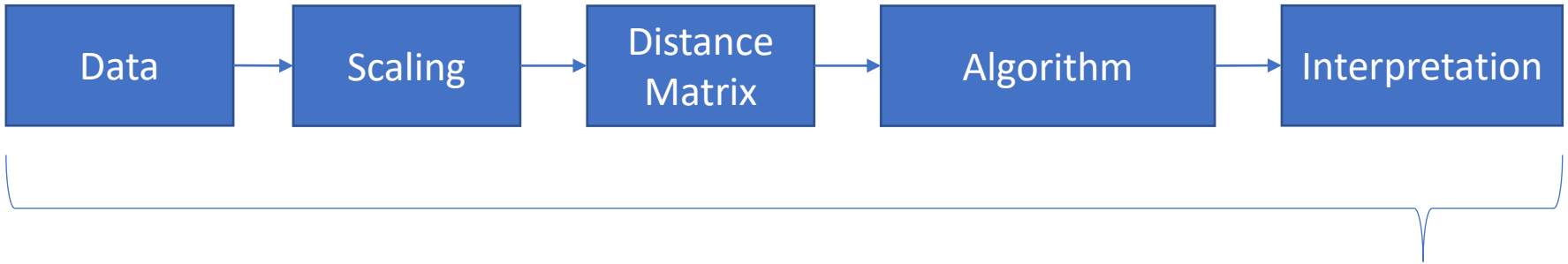
Workflow



- Algorithm selection

Clustering

Workflow



- Use the results to get deeper insights