Regression

PREDICTING A QUANTITY

START

Is sample size >50?

N

Gather more data

Y

Is sample size > 100K?

Arockia Liborious
Talk ML/DL with AL

Y

SGD Regressor

If not working

Dimensionality Reduction (PCA)

Y

Lasso ElasticNet

N

Fewer Features are important

N

Rigde Regression
SVR (kernel='linear')

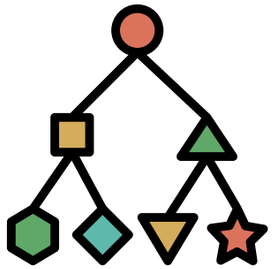
If not working

SVR (kernel='rbf')
Ensemble Regressors

Makes forecasts by estimating the relationship b/w variables
Answers questions like: How much or how many?

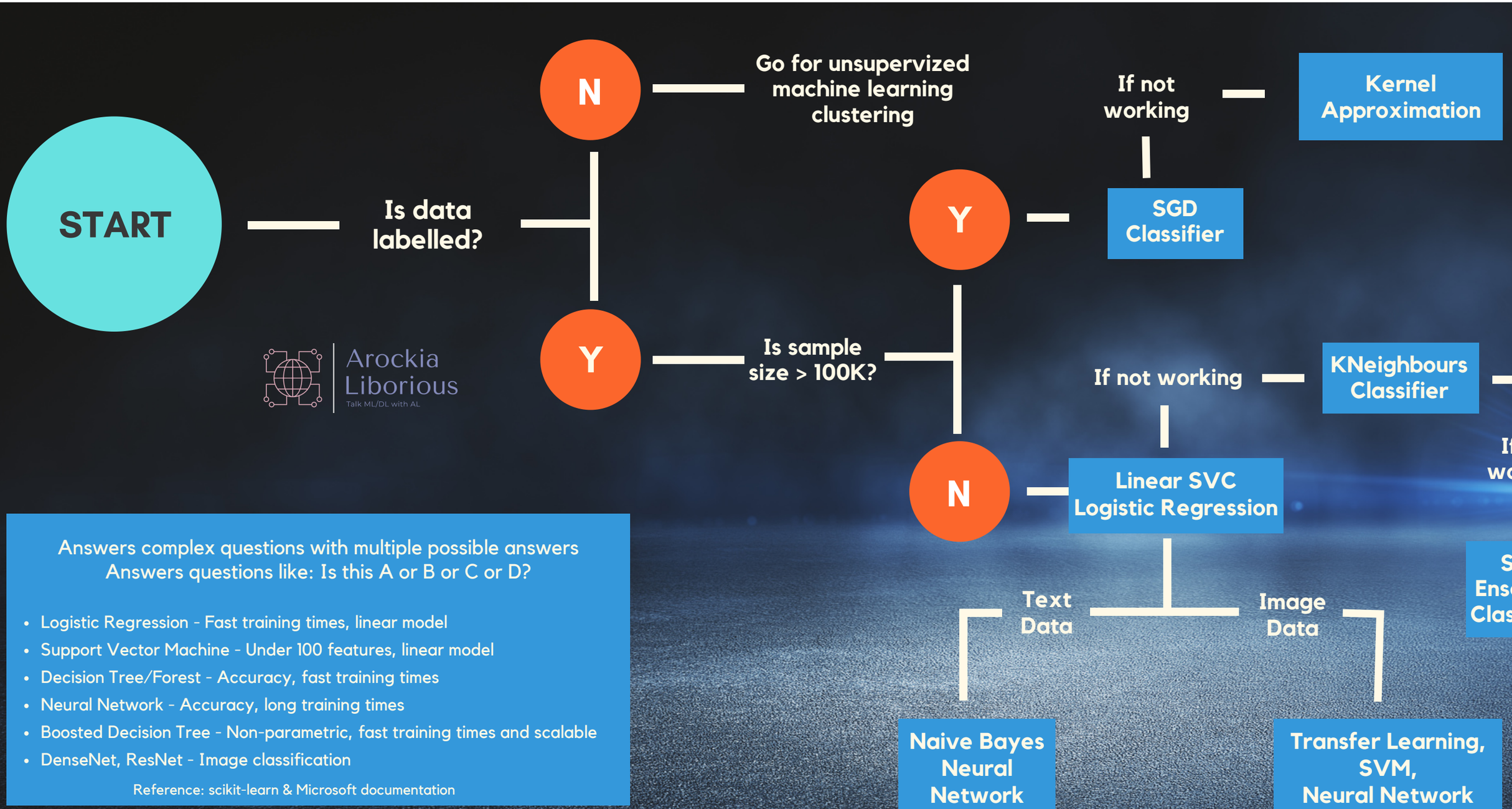
- Linear Regression - Fast training, linear model
- Poisson Regression - Predicts event counts
- Polynomial Regression - To model on-linear relationship
- Bayesian Linear Regression - Linear model, small data sets
- Decision Forest Regression - Accurate, fast training times
- Quantile Regression Forests - Predicts a distribution w/ outliers
- Neural Network Regression - Accurate, long training times
- Boosted Decision Tree Regression - Accurate, fast training times, large memory footprint

Reference: scikit-learn & Microsoft documentation



Classification

PREDICTING A CATEGORY



Arockia Liborious
Talk ML/DL with AL

Answers complex questions with multiple possible answers
Answers questions like: Is this A or B or C or D?

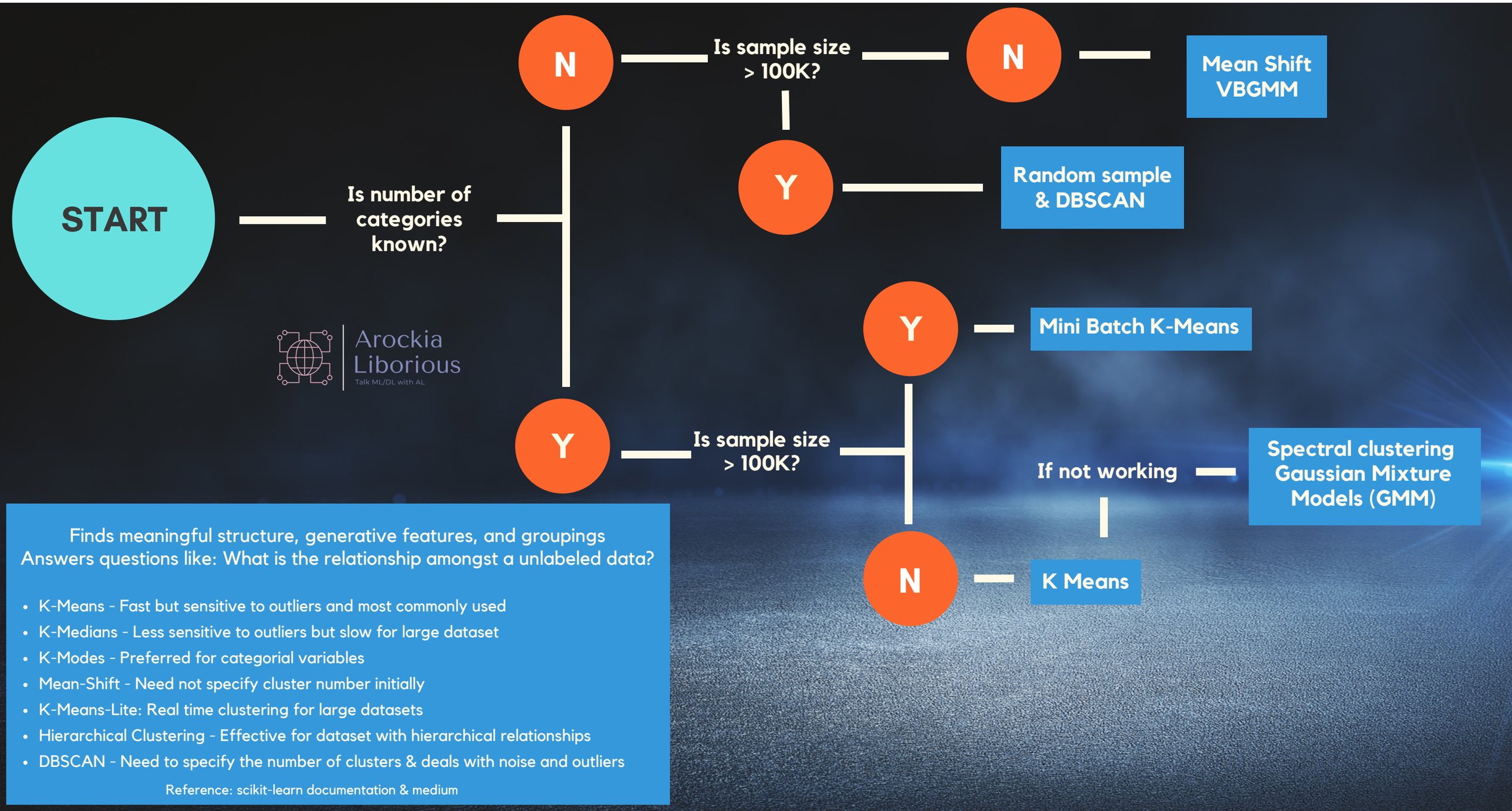
- Logistic Regression - Fast training times, linear model
- Support Vector Machine - Under 100 features, linear model
- Decision Tree/Forest - Accuracy, fast training times
- Neural Network - Accuracy, long training times
- Boosted Decision Tree - Non-parametric, fast training times and scalable
- DenseNet, ResNet - Image classification

Reference: scikit-learn & Microsoft documentation

Clustering

PREDICTING A (UNLABELED) CATEGORY / GROUP

Arockia Liborious
Talk ML with AL



Finds meaningful structure, generative features, and groupings
Answers questions like: What is the relationship amongst a unlabeled data?

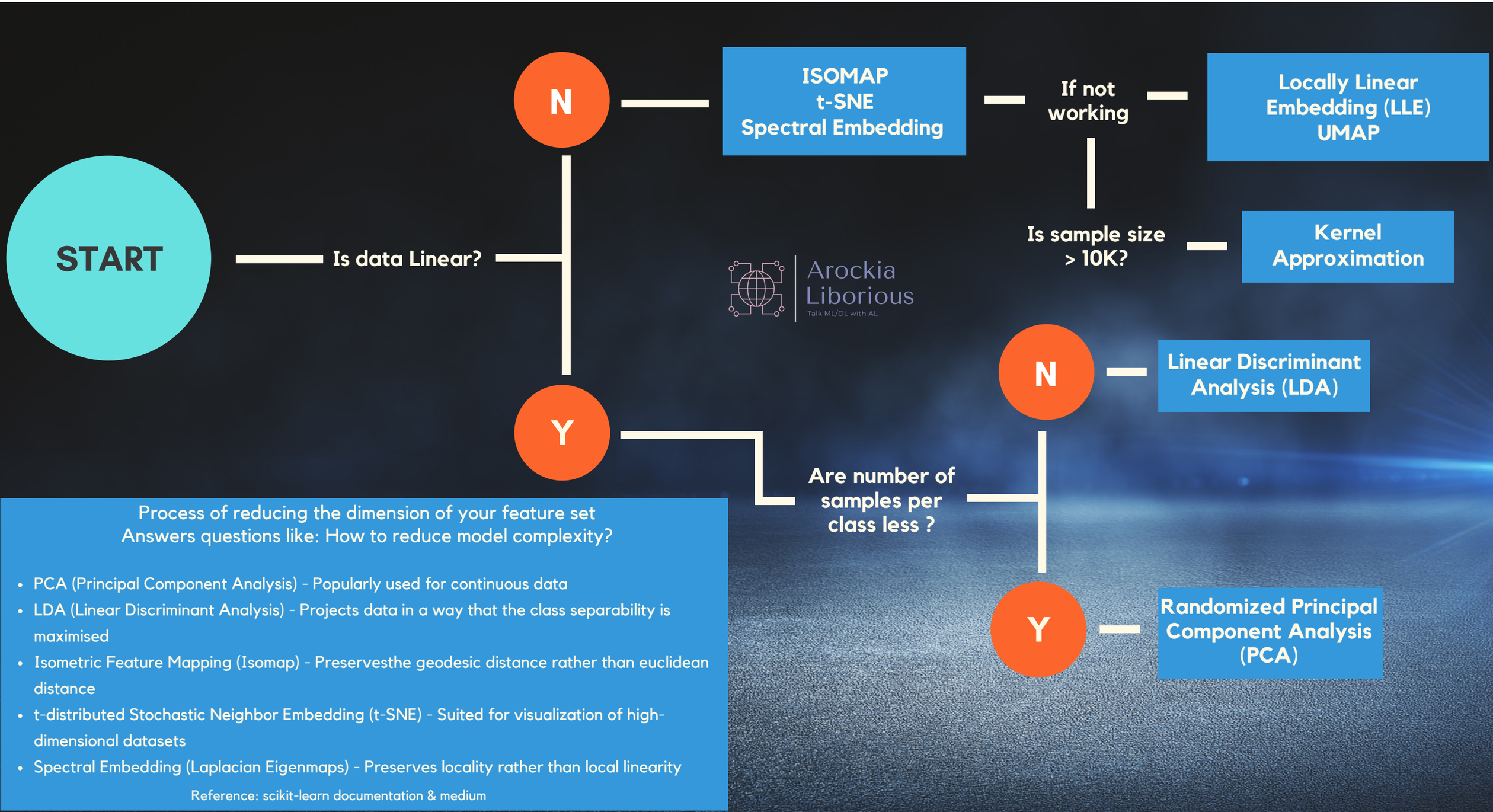
- K-Means - Fast but sensitive to outliers and most commonly used
- K-Medians - Less sensitive to outliers but slow for large dataset
- K-Modes - Preferred for categorial variables
- Mean-Shift - Need not specify cluster number initially
- K-Means-Lite: Real time clustering for large datasets
- Hierarchical Clustering - Effective for dataset with hierarchical relationships
- DBSCAN - Need to specify the number of clusters & deals with noise and outliers

Reference: scikit-learn documentation & medium

Dimensionality Reduction

IDENTIFY HIDDEN DEPENDENCIES & SIMPLIFY DATA

Arockia Liborious
Talk ML with AL



Process of reducing the dimension of your feature set
Answers questions like: How to reduce model complexity?

- PCA (Principal Component Analysis) - Popularly used for continuous data
- LDA (Linear Discriminant Analysis) - Projects data in a way that the class separability is maximised
- Isometric Feature Mapping (Isomap) - Preservesthe geodesic distance rather than euclidean distance
- t-distributed Stochastic Neighbor Embedding (t-SNE) - Suited for visualization of high-dimensional datasets
- Spectral Embedding (Laplacian Eigenmaps) - Preserves locality rather than local linearity

Reference: scikit-learn documentation & medium