Machine Learning in Astronomy

APS Hack Day
November 15, 2019

Protostellar classification using supervised machine learning algorithms

Galaxy detection and identification using deep learning and data augmentation.

Machine learning in APOGEE

Identification of stellar populations through chemical abundances★

A Machine-learning Data Set Prepared from the NASA *Solar Dynamics Observatory* Mission

Machine learning and Kolmogorov analysis to reveal gravitational lenses

Goals for this Presentation

Unfog some misconceptions about machine learning

Provide the framework for understanding machine learning

Outline how to implement machine learning

Goals for this Presentation

Unfog some misconceptions about machine learning

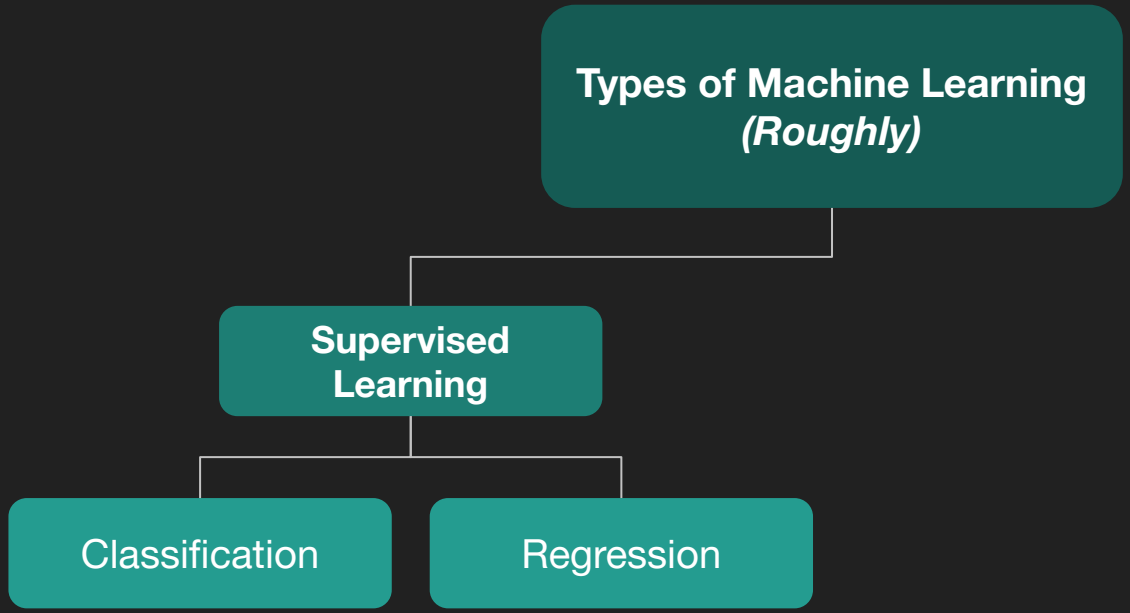
Provide the framework for understanding machine learning

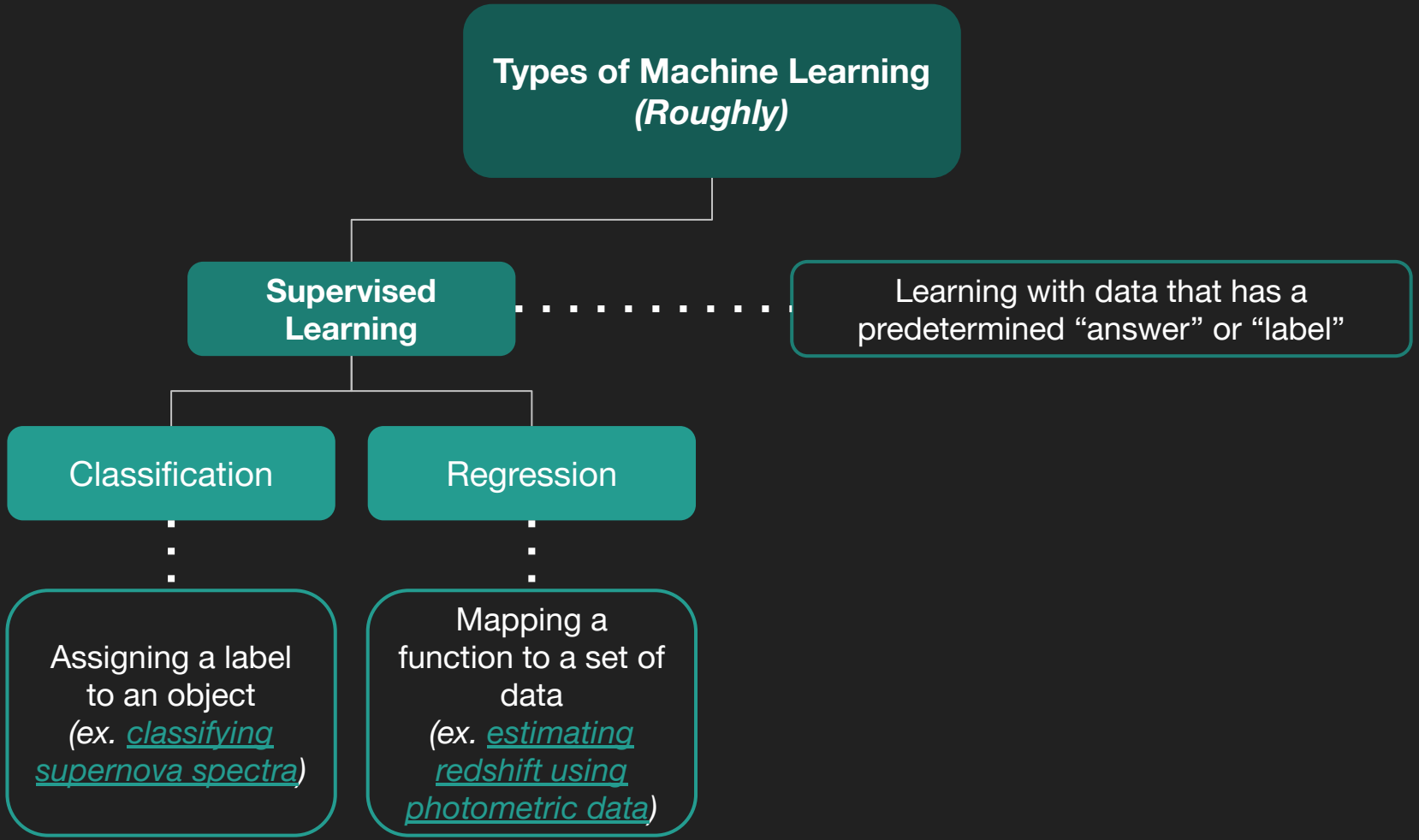
Outline how to implement machine learning

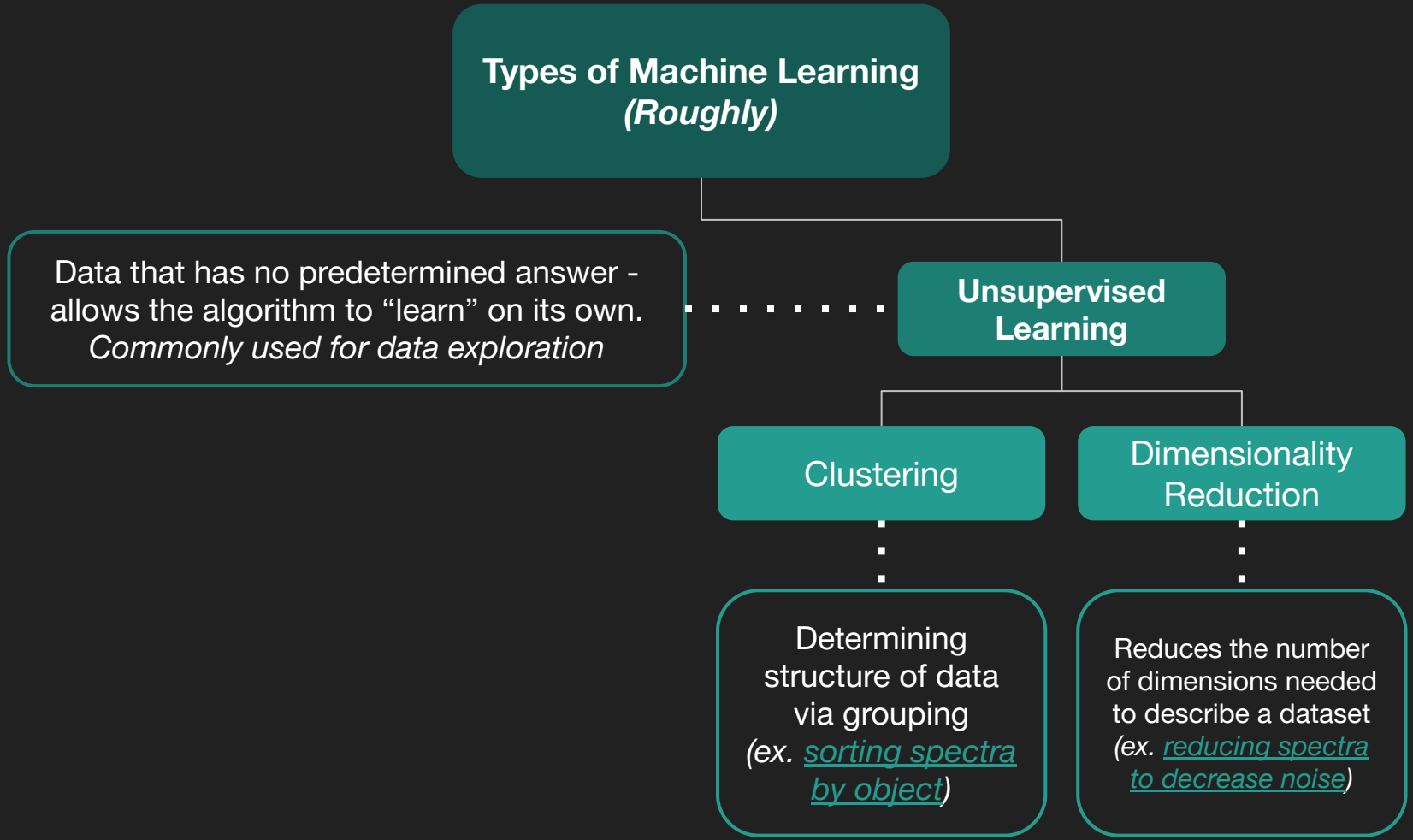
❗ Disclaimer:

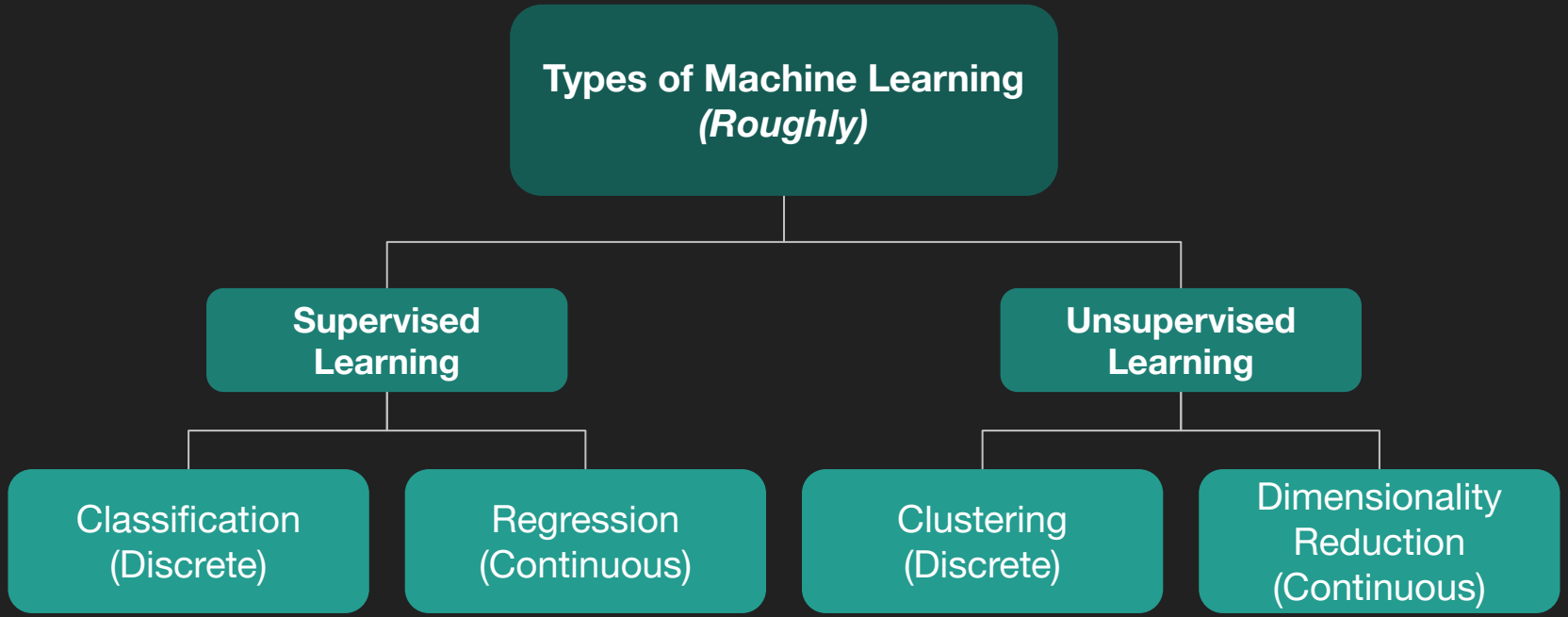
- I am not an expert.
- Everything I know is self-taught.
- Tons of resources are at the end of the presentation and I urge you to visit if you're wanting to learn more.

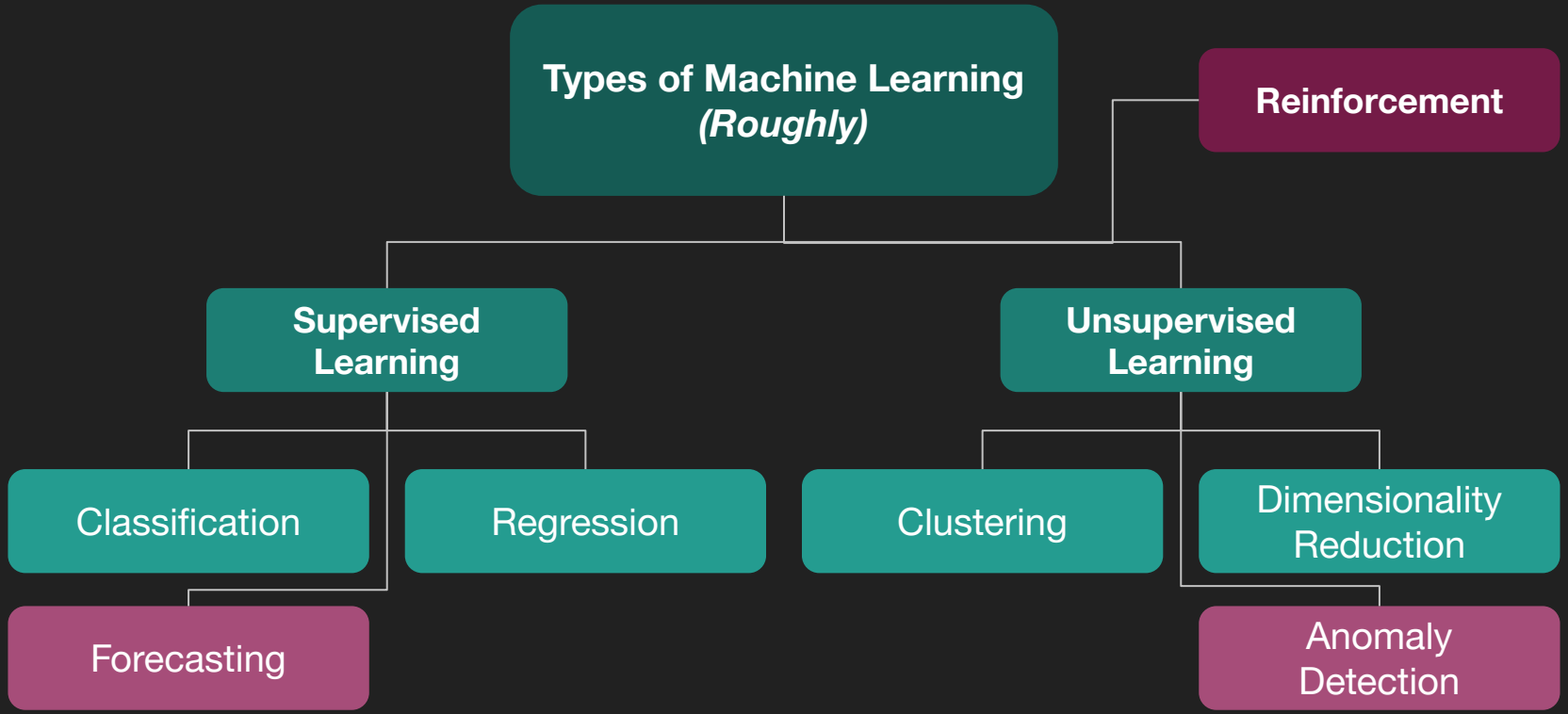
My goals are built around preparing you to be able to read about ML or use it in your own research.

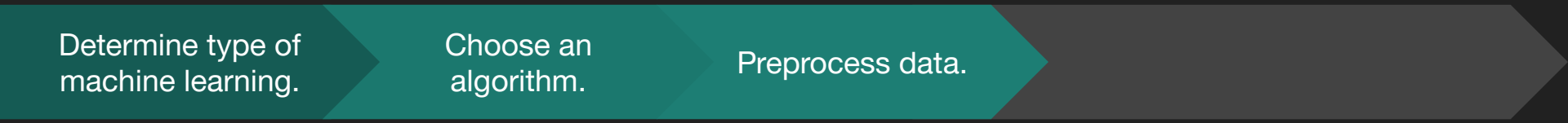

Types of Machine Learning
(Roughly)
Supervised Learning
Classification
Regression


Types of Machine Learning
(Roughly)
Supervised Learning
Learning with data that has a predetermined “answer” or “label”
Classification
Regression
Assigning a label to an object
(ex. classifying supernova spectra)
Mapping a function to a set of data
(ex. estimating redshift using photometric data)


Types of Machine Learning
(Roughly)
Data that has no predetermined answer - allows the algorithm to “learn” on its own.
Commonly used for data exploration
Unsupervised Learning
Clustering
Dimensionality Reduction
Determining structure of data via grouping
(ex. sorting spectra by object)
Reduces the number of dimensions needed to describe a dataset
(ex. reducing spectra to decrease noise)


Types of Machine Learning
(Roughly)
Supervised Learning
Unsupervised Learning
Classification (Discrete)
Regression (Continuous)
Clustering (Discrete)
Dimensionality Reduction (Continuous)


Types of Machine Learning
(Roughly)
Reinforcement
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering
Dimensionality Reduction
Forecasting
Anomaly Detection

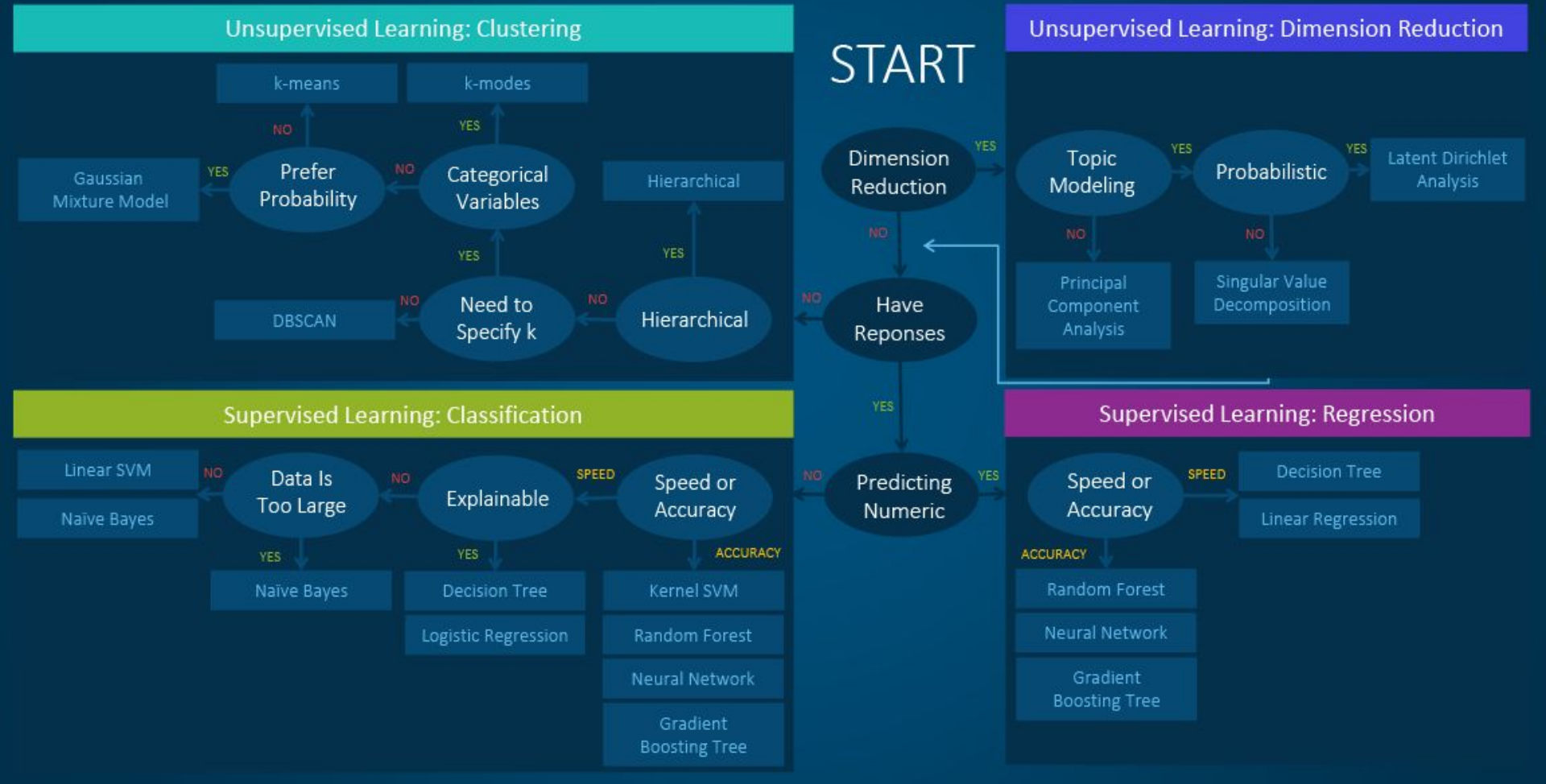


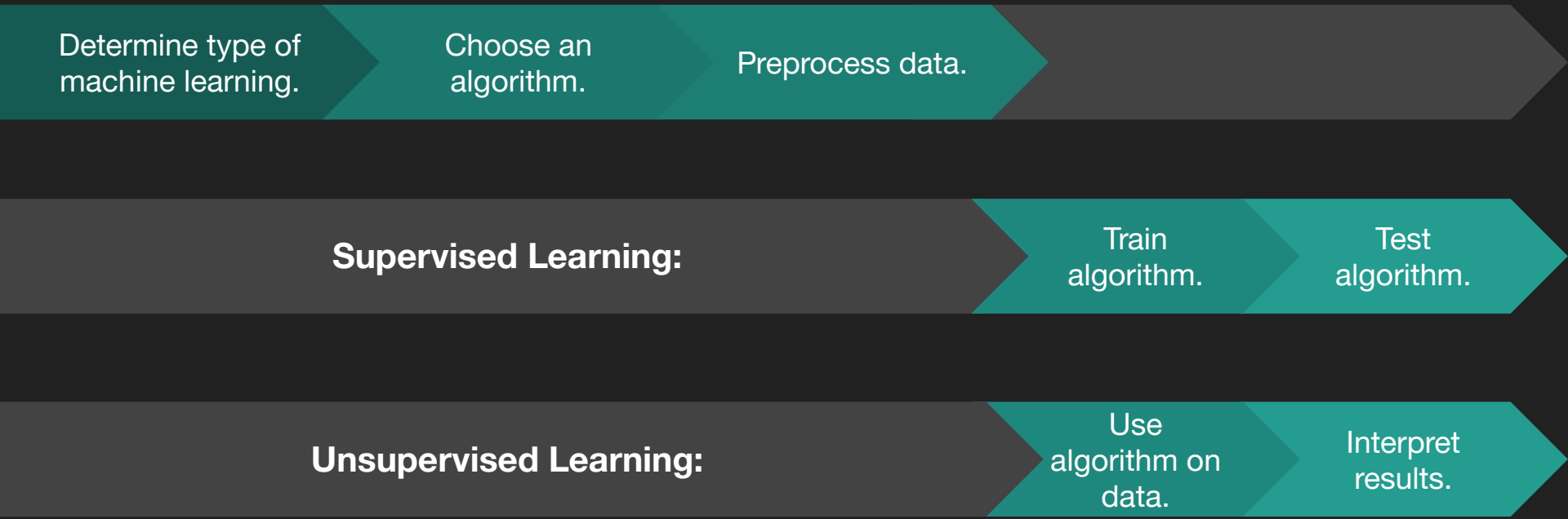
Based on your data and your goal, choose a type of machine learning to implement.

Each type of machine learning has many types of algorithms. Each has their strengths and weaknesses.

Machine learning algorithms do not play nicely with missing or extreme values. Many built in functions to help clean or normalize data.

Machine Learning Algorithms Cheat Sheet




Determine type of machine learning.
Choose an algorithm.
Preprocess data.
Supervised Learning:
Train algorithm.
Test algorithm.
Unsupervised Learning:
Use algorithm on data.
Interpret results.

Common Metrics for Machine Learning

Regression

MSPE
MSAE
R^2
Adjusted R^2

Classification

Precision-Recall
Accuracy
ROC-AUC
Log-Loss

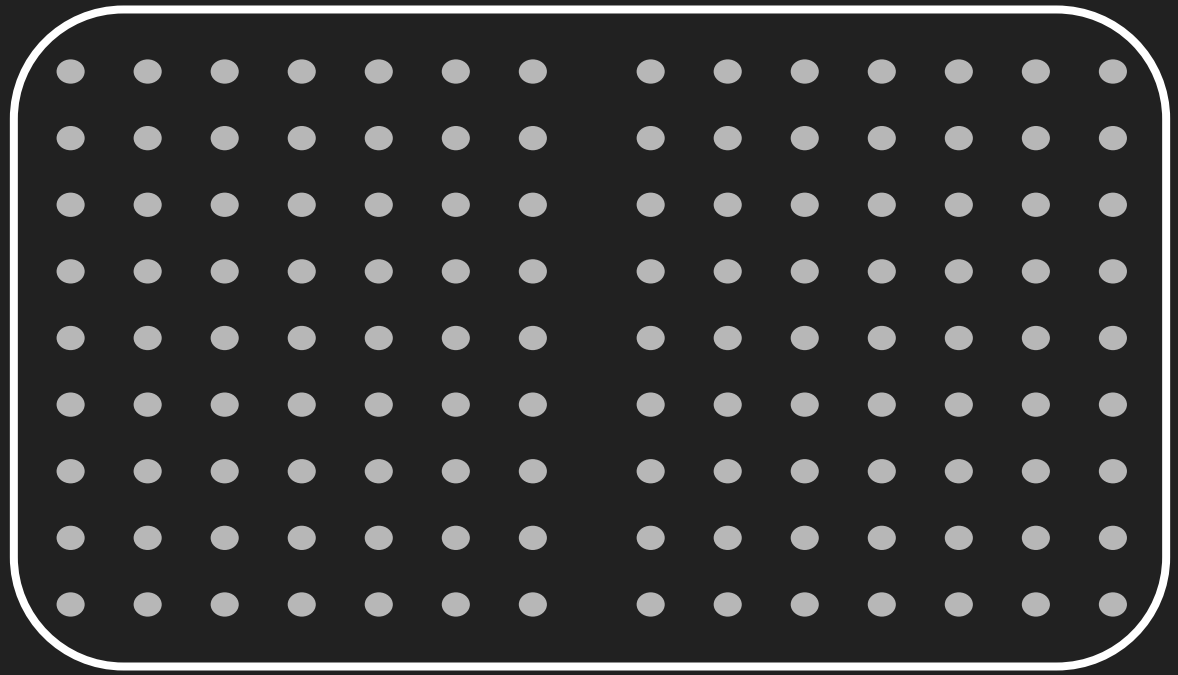
Unsupervised Models

Rand Index
Mutual Information

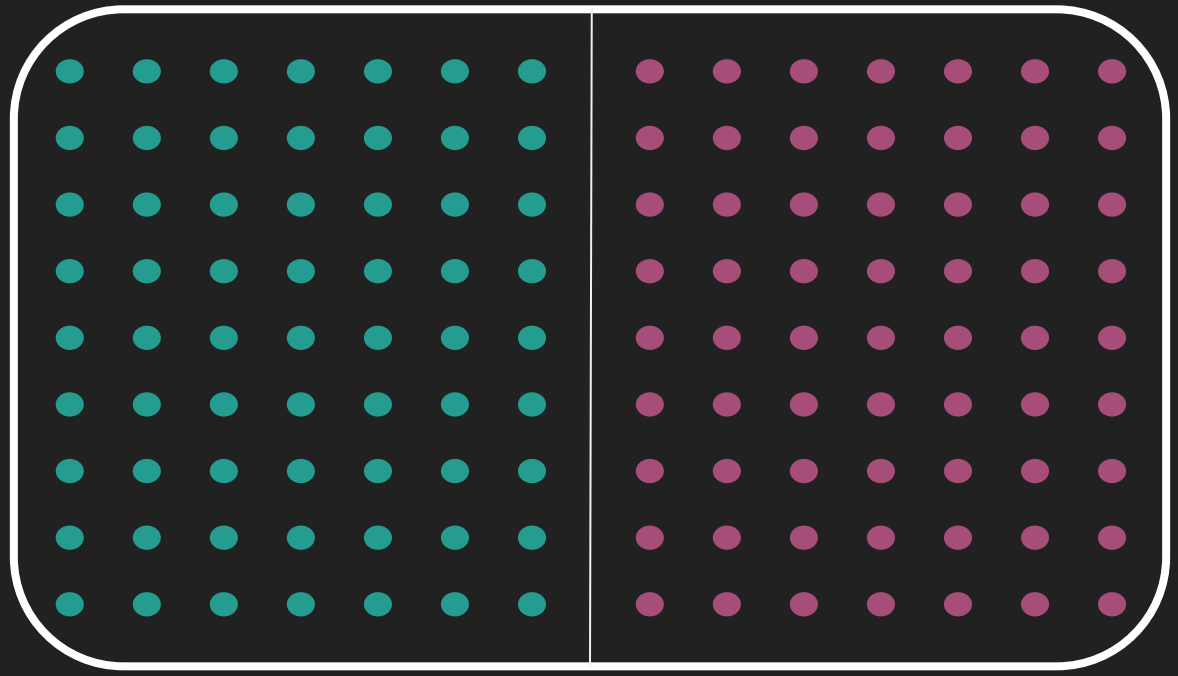
Other

CV Error
Heuristic Methods
BLEU Score

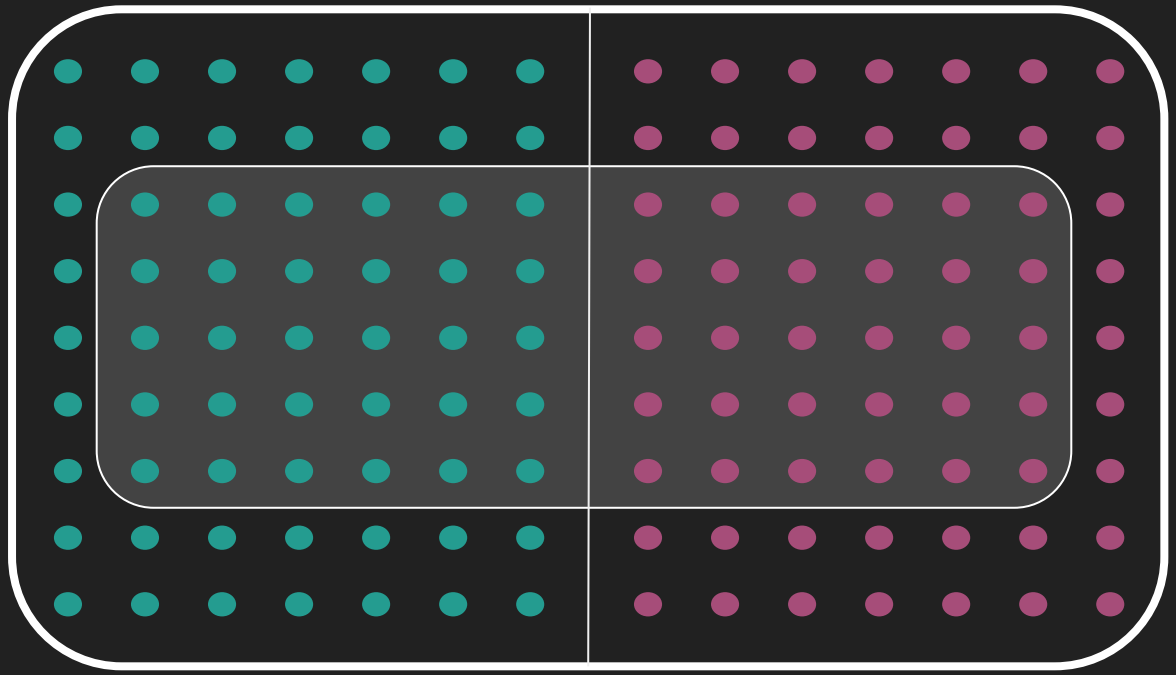
Classification Metric: Precision-Recall



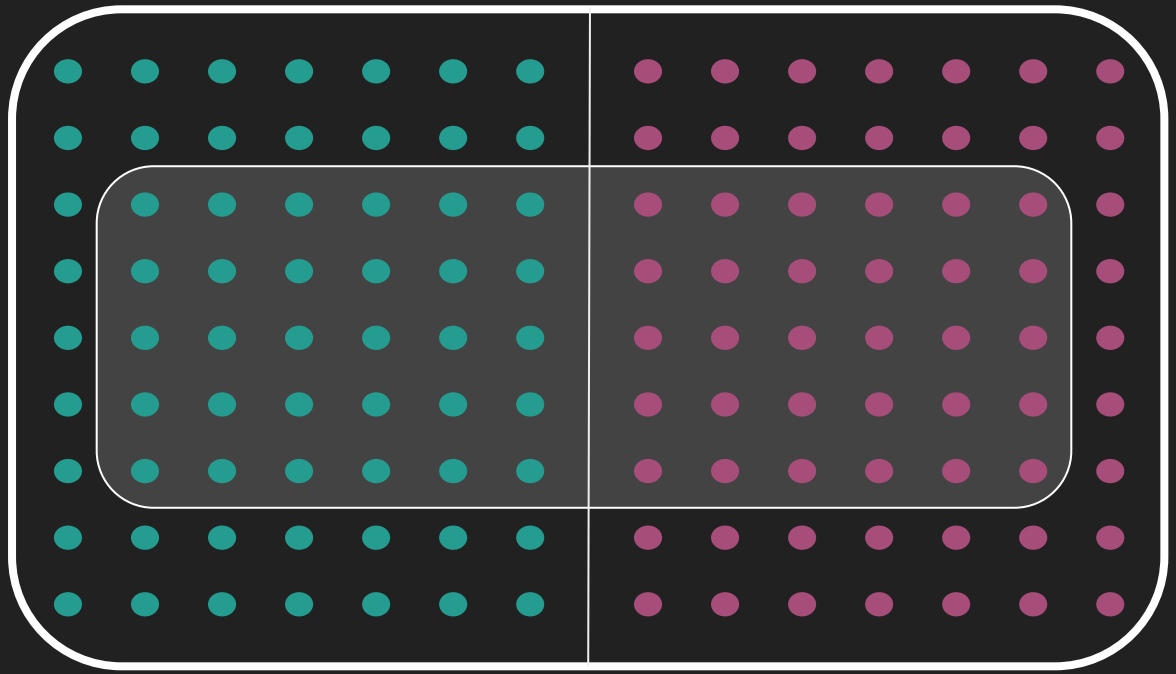
Classification Metric: Precision-Recall



Classification Metric: Precision-Recall



Classification Metric: Precision-Recall



Group 1

Group 2

Correctly Classified

Incorrectly Classified

True Positive

False Positive

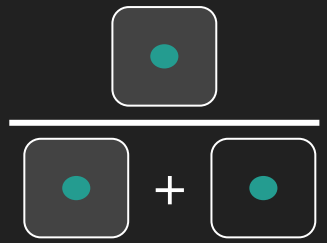
True Negative

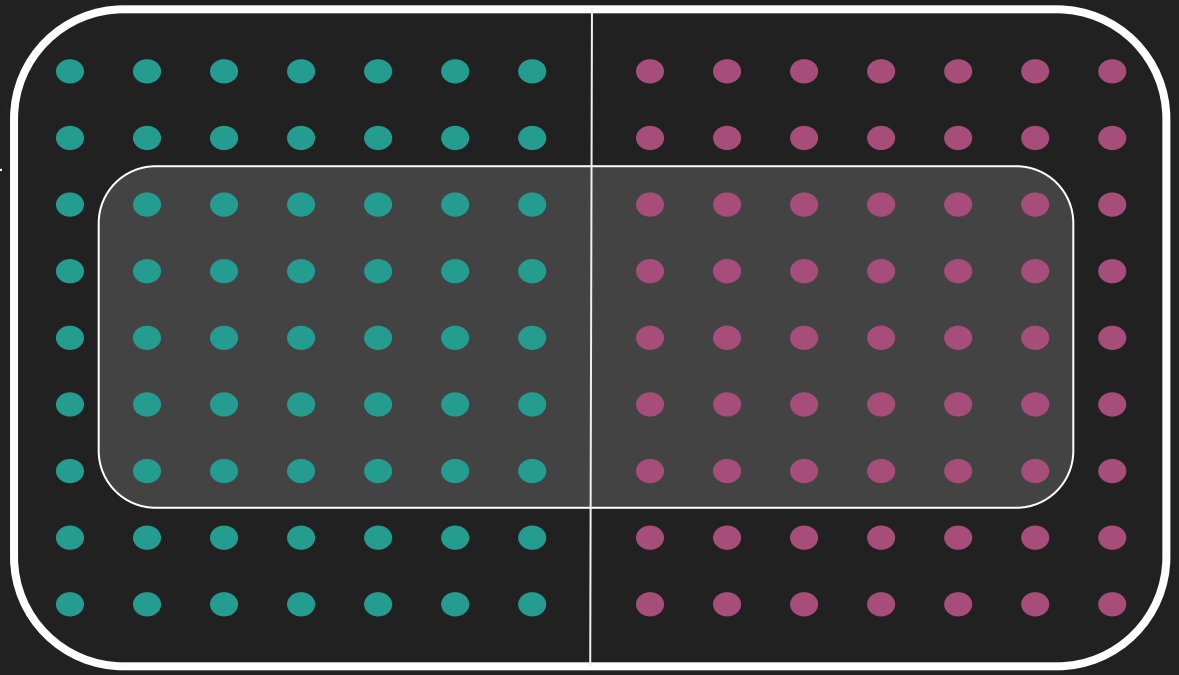
False Negative

Classification Metric: Precision-Recall

Precision

$$P = \frac{TP}{TP + FP}$$





Group 1

Group 2

Correctly Classified

Incorrectly Classified

True Positive

False Positive

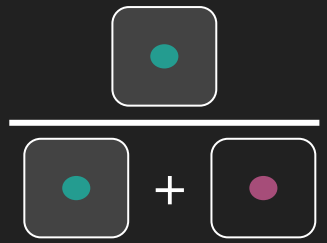
True Negative

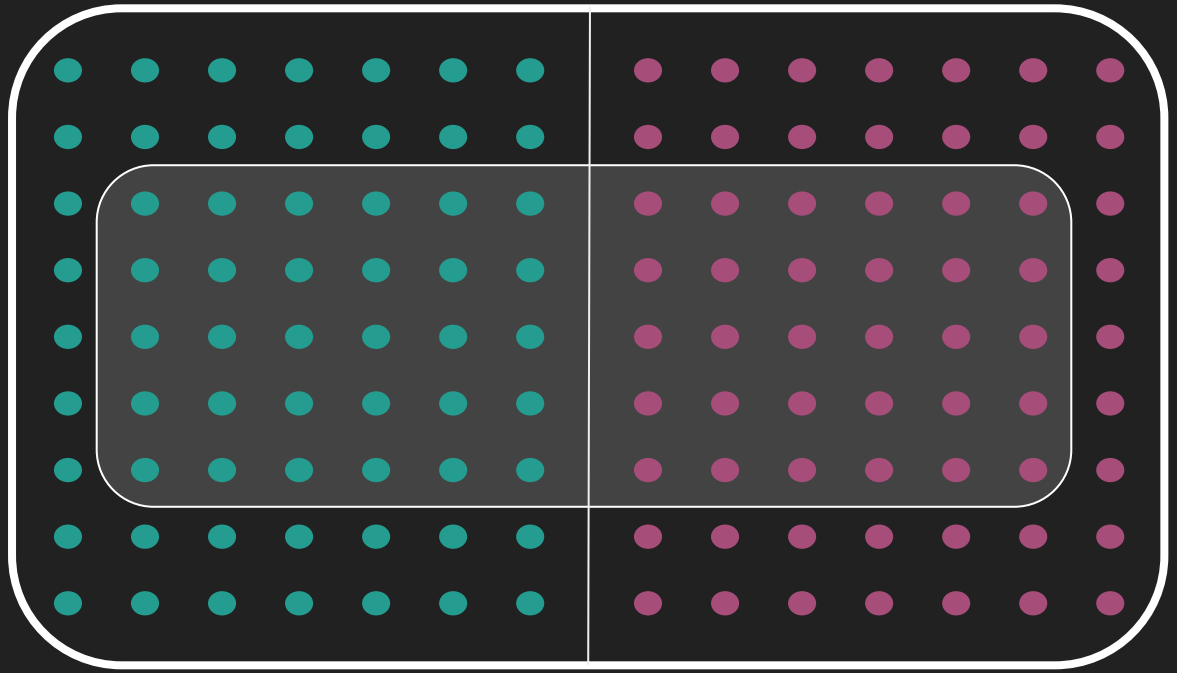
False Negative

Classification Metric: Precision-Recall

Recall

$$R = \frac{TP}{TP + FN}$$





Group 1

Group 2

Correctly Classified

Incorrectly Classified

True Positive

False Positive

True Negative

False Negative

Classification Metric: Precision-Recall

Precision

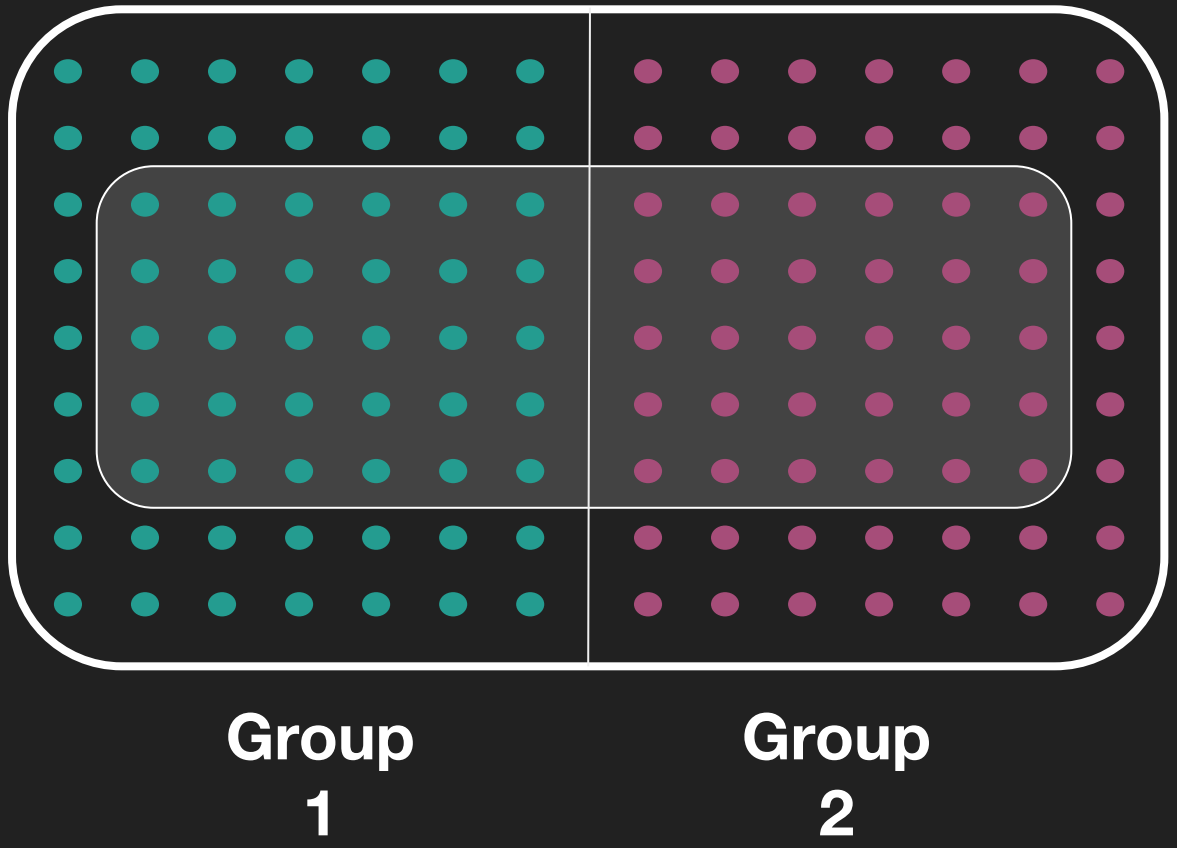
$$P = \frac{TP}{TP + FP}$$

Recall

$$R = \frac{TP}{TP + FN}$$

F_1-Score

$$F_1 = 2\frac{P \times R}{P + R}$$



Correctly Classified

Incorrectly Classified

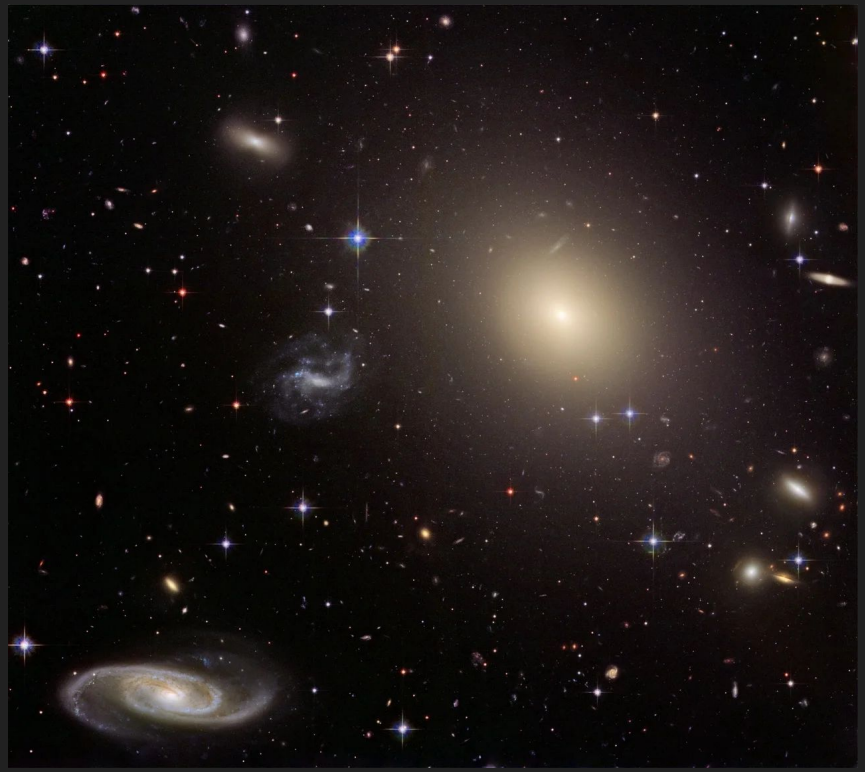
True Positive

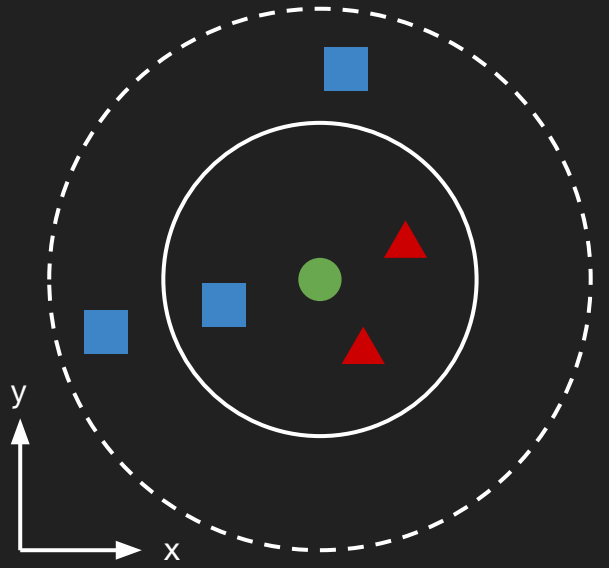
False Positive

True Negative

False Negative

Classification Tutorial:
Can we sort galaxies by morphology with SDSS?





K-Nearest Neighbors:
***k* = 3 (solid line):**
green dot = red triangle

***k* = 5 (dashed line):**
green dot = blue square

Resources

Python for Machine Learning Cheat Sheet

Machine Learning Algorithm Cheat Sheet

Sk-Learn Algorithm Cheat Sheet

Choosing the Right Metric for Machine Learning

Machine Learning in Astronomy Baron 2019
Particularly the sections on Unsupervised Learning

A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey - Catalog used for Tutorial