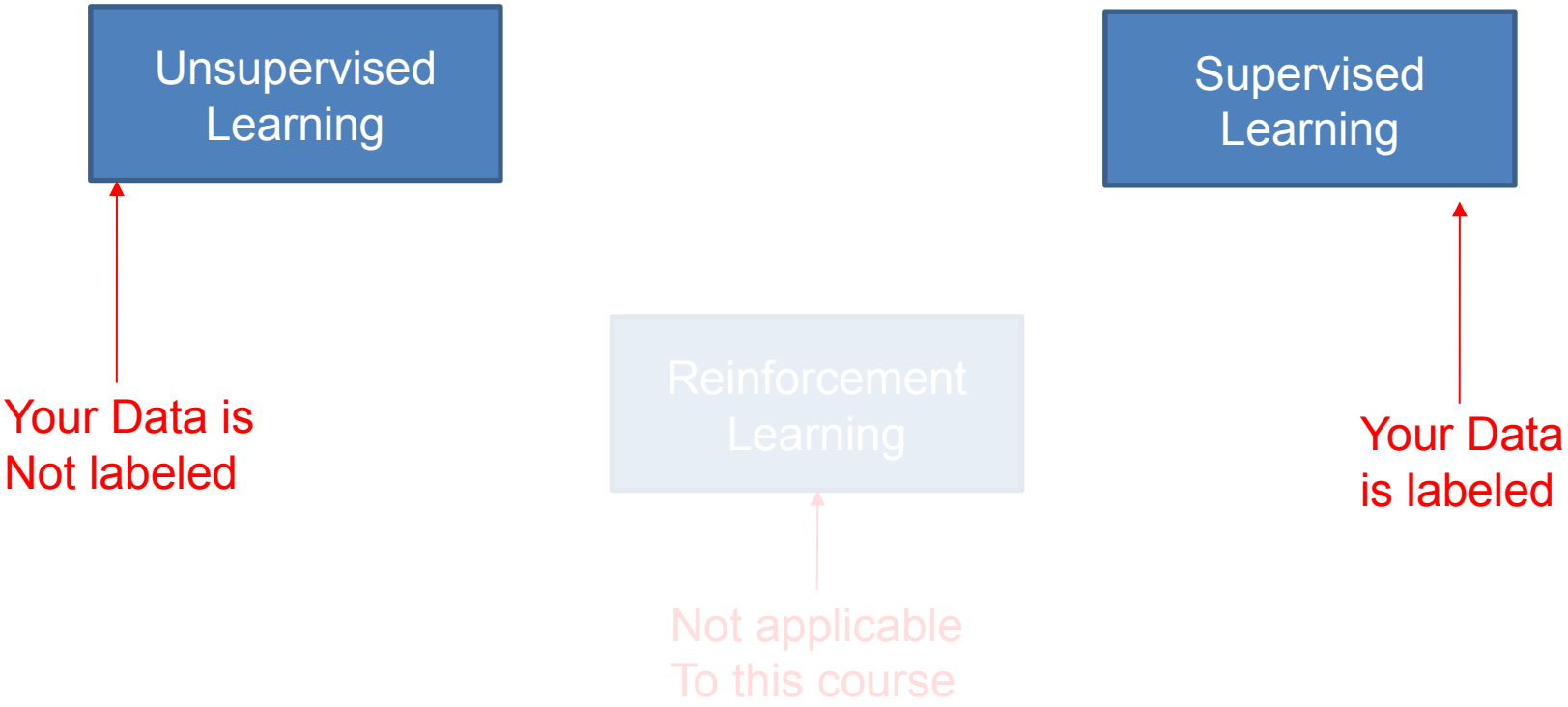
DATA 301: Machine Learning and Clustering Introduction

Three types of machine learning



Supervised and Unsupervised Learning

Unsupervised

- Unlabeled Data
- Model does not know correct result during training
- Model finds hidden patterns in data
- Goal is to train the model to detect these patterns in unseen data

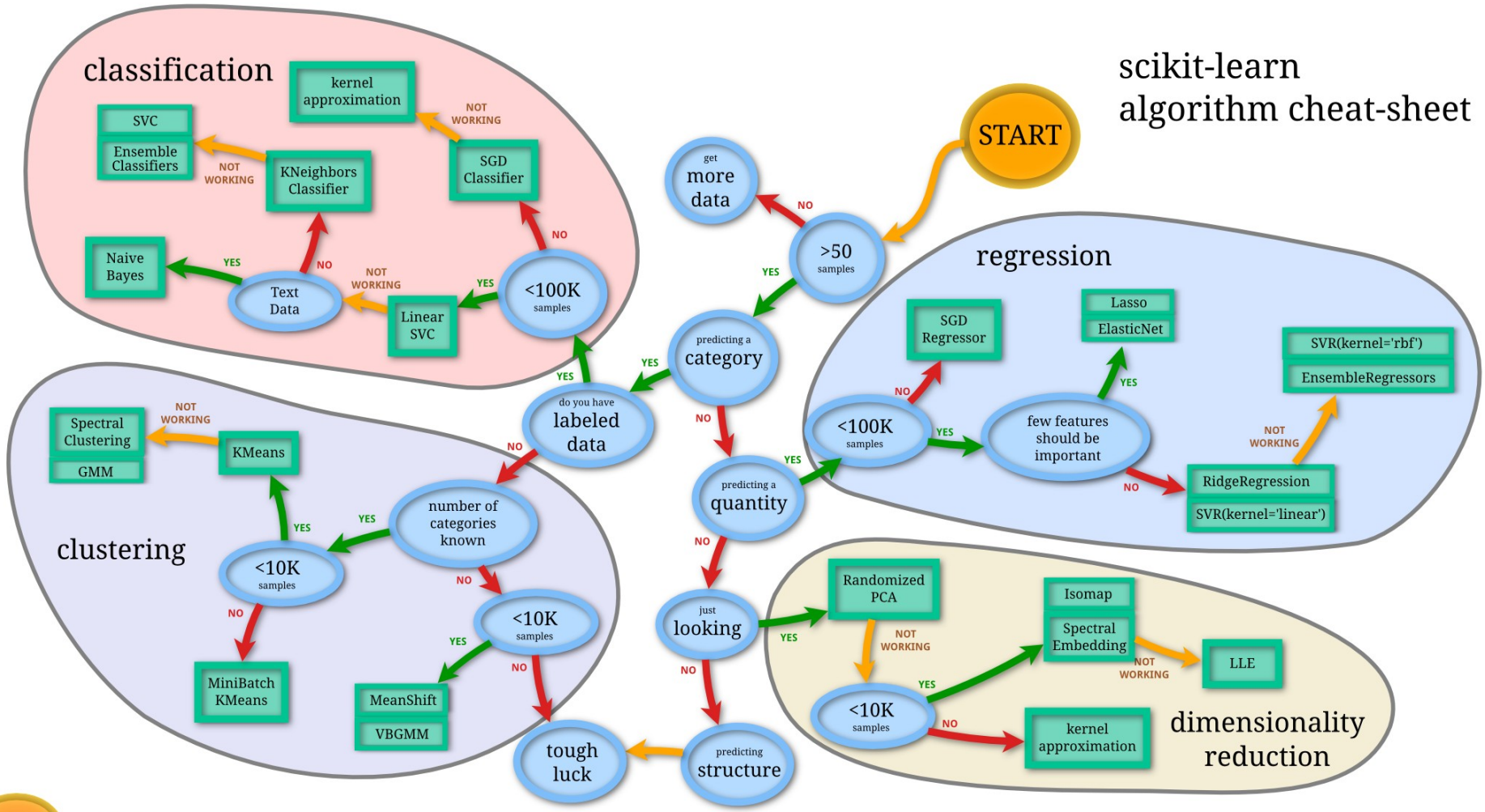
- Examples: Kmeans, DBSCAN, Mean Shift

Supervised

- Labeled Data
- Model knows correct result during training
- Uses correct result to adjust model params during training
- Goal is to train model to pick correct result on unseen data

- Examples: Regressions, Decision Trees, Neural Networks

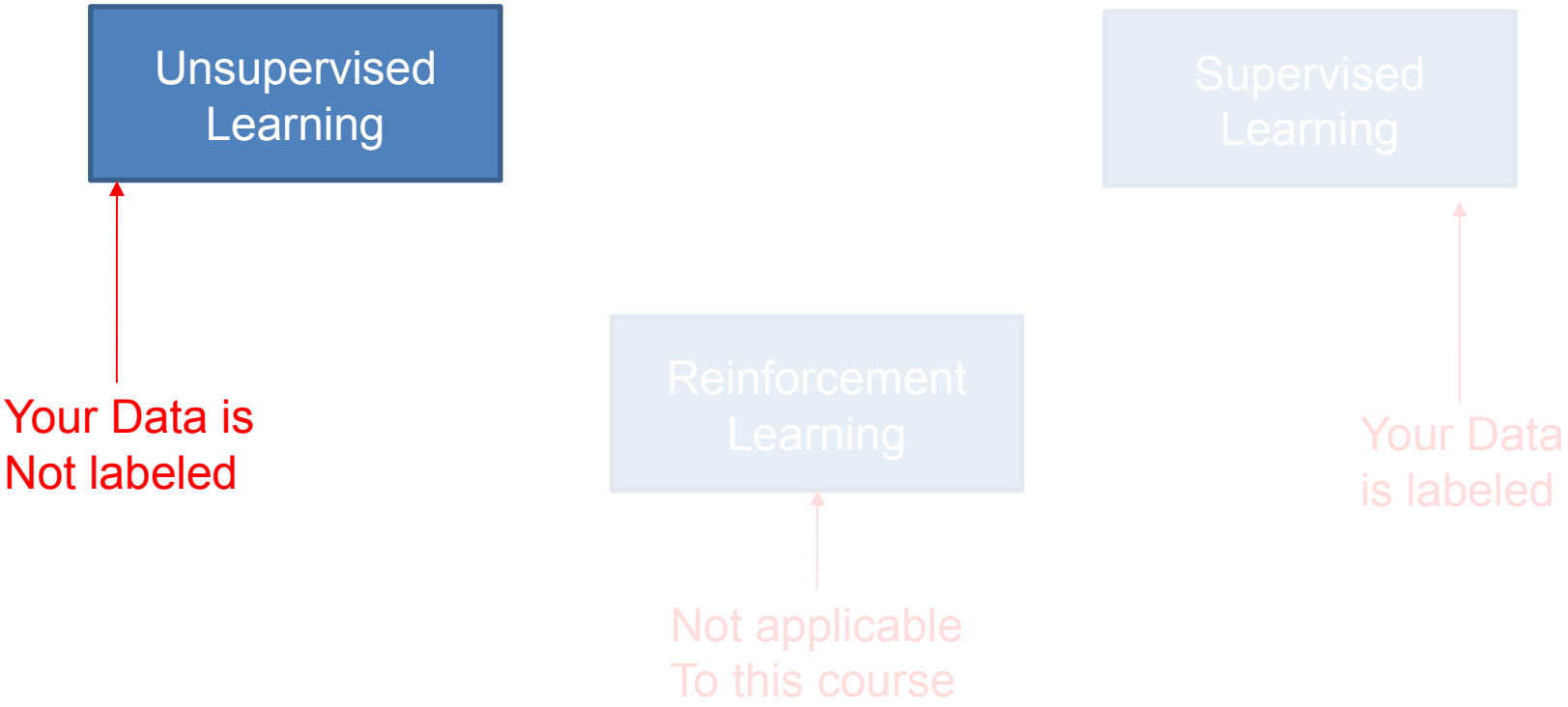
Supervised and Unsupervised Learning



Back

From https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Three types of machine learning



Unsupervised Learning – Clustering Introduction

For this course this means clustering.

Clustering is an exploratory analysis technique that organizes data into groups without any knowledge of what these clusters should be. The goal is that items in a cluster are more similar to each other than items in other clusters

But there is no free lunch.

- **You must tune clustering algorithms to the dataset they are applied to.**
- **It's difficult to tell when the algorithms are tuned properly when you have no knowledge of what the clusters should be.**
- **And finally, once clusters are selected, further analysis is required to determine what the groupings mean (if anything).**

Unsupervised Learning – Clustering, Some use cases

- **Market segmentation**
- **Recomender systems**
- **Email marketing**
- **Anomaly detection**
- **Producing new features for ML pipeline**

Unsupervised Learning – Clustering, What we will cover

- **K-means**
- **Hierarchical (maybe)**
- **DBSCAN and relatives**

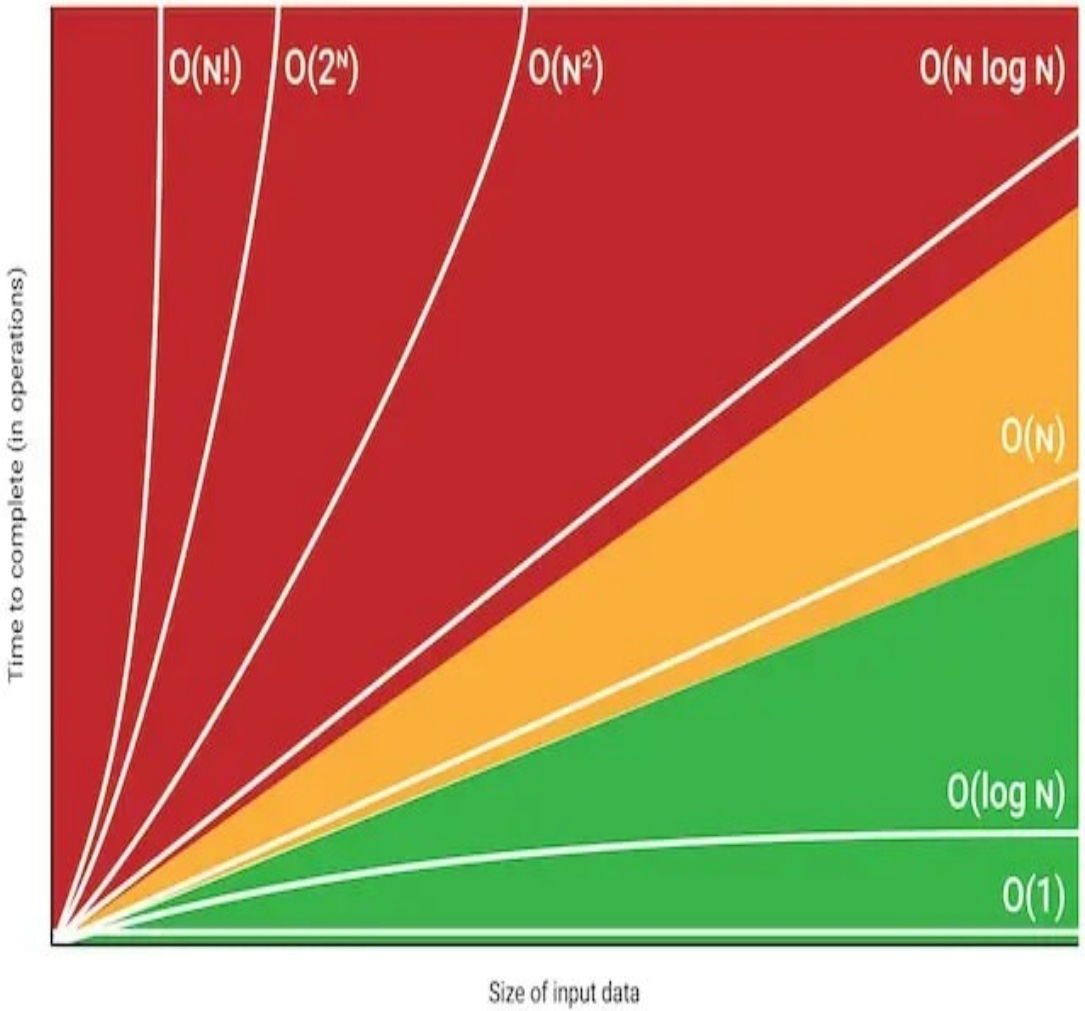
- **Go to https://scikit-learn.org/stable/modules/clustering.html**

- **For kmeans, DBSCAN, Agglomerative**
- **Note the ability of each algo to ID clusters**
- **Note the time it takes**
- **Note Outlier detection**

Unsupervised Learning – Clustering, Performance

- **Clustering algorithms are designed to do one thing; separate data into clusters.**
- **If 2 clustering algorithms generate roughly the same clusters, then which do you use?**
- **Likely the one that runs the fastest.**
- **In CS we measure this by Big-O notation.**
- **Big-O a proxy for runtime on an infinitely large dataset.**
- **Big-O can also be used to measure space complexity, the amount of memory used.**

Unsupervised Learning – Clustering, Performance - time



Fortunately, Big-O performance data is available for our algorithms of interest.

K-means	O(n)
Heiarchical	O(n**2)
DBSCAN	O(nlogn)

k-means looks mighty Attractive.

Until you consider how poorly it clustered non-linearly separable datasets*.

Chart from https://dev.to/karthikeyan676/brief-intro-to-big-o-5beg

*Scikitlearn performance info **https://scikit-learn.org/stable/modules/clustering.html**

Summary

Three types of machine learning
Clustering Introduction
Clustering Use cases
What we will cover
A bit about Big-O and time complexity