**Learning objective 1**

**Clustering**

refers to a set of specific algorithms

spark most commonly uses k-means

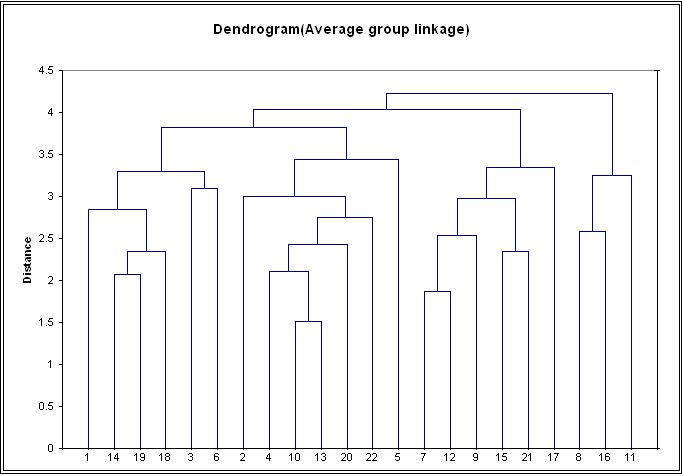
classes of clustering algorithms

eg hierarchical clustering

eg statistical clustering (k-means)

combination of the two

visualization tools



typical application

remove duplicates

https://spark.apache.org/docs/1.1.0/mllib-clustering.html

**Machine Learning**

refers to a set of algorithms and conceptual frameworks

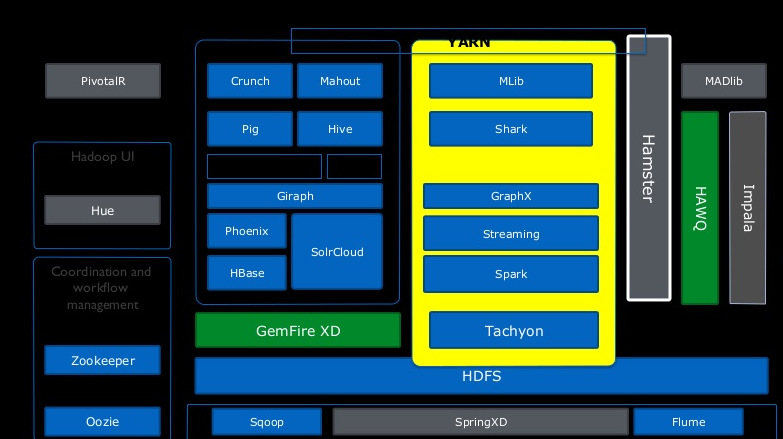
algorithms <https://spark.apache.org/docs/1.1.0/mllib-guide.html>

conceptual frameworks supervised. semi-supervised, …. in the end comes down to a mathematical domain for the conceptual basis of the algorithm.

**Data Science and Analysis**

refers to a set of algorithms and conceptual frameworks and tools

**Spark in context**



Summary Objective 1

**To**

**be able to describe Apache Spark as a solution in a given context. Clustering, Machine Learning, Data Science and Analysis, …..**

1. Clustering applications, eg remove duplicates
2. Machine Learning Mlib, training, recommendations
3. Spark components, Shark, GraphX
4. Runtime Modes, StandAlone, Yarn Cluster, Mesos Cluster