Guided Tour of Machine Learning in Finance

Week 3: Unsupervised Learning

Clustering algorithms

Igor Halperin

NYU Tandon School of Engineering, 2017

Why clustering?

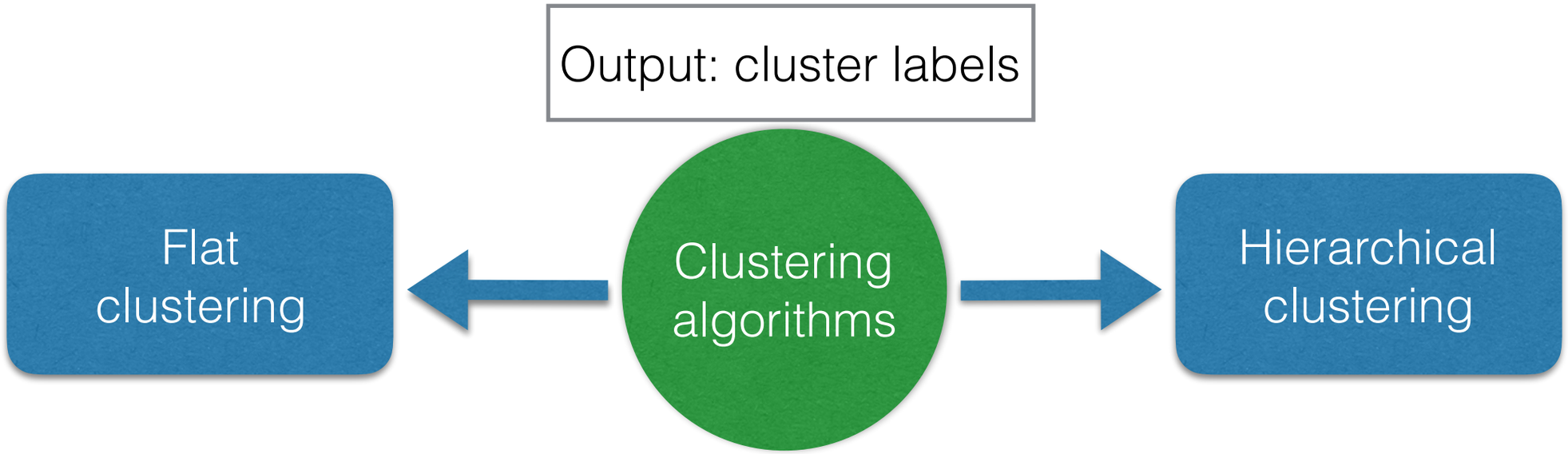
Aggregate a whole population of companies, stocks, credit card holders, mortgage holders etc. into a relatively homogeneous groups where intra-group variance is generally smaller than inter-group variance.

Purposes:

1. Visualization of data (when data is low-dimensional)
2. Conceptualization of clusters, model-building on clusters as homogeneous sets of data
3. Compact representation of data

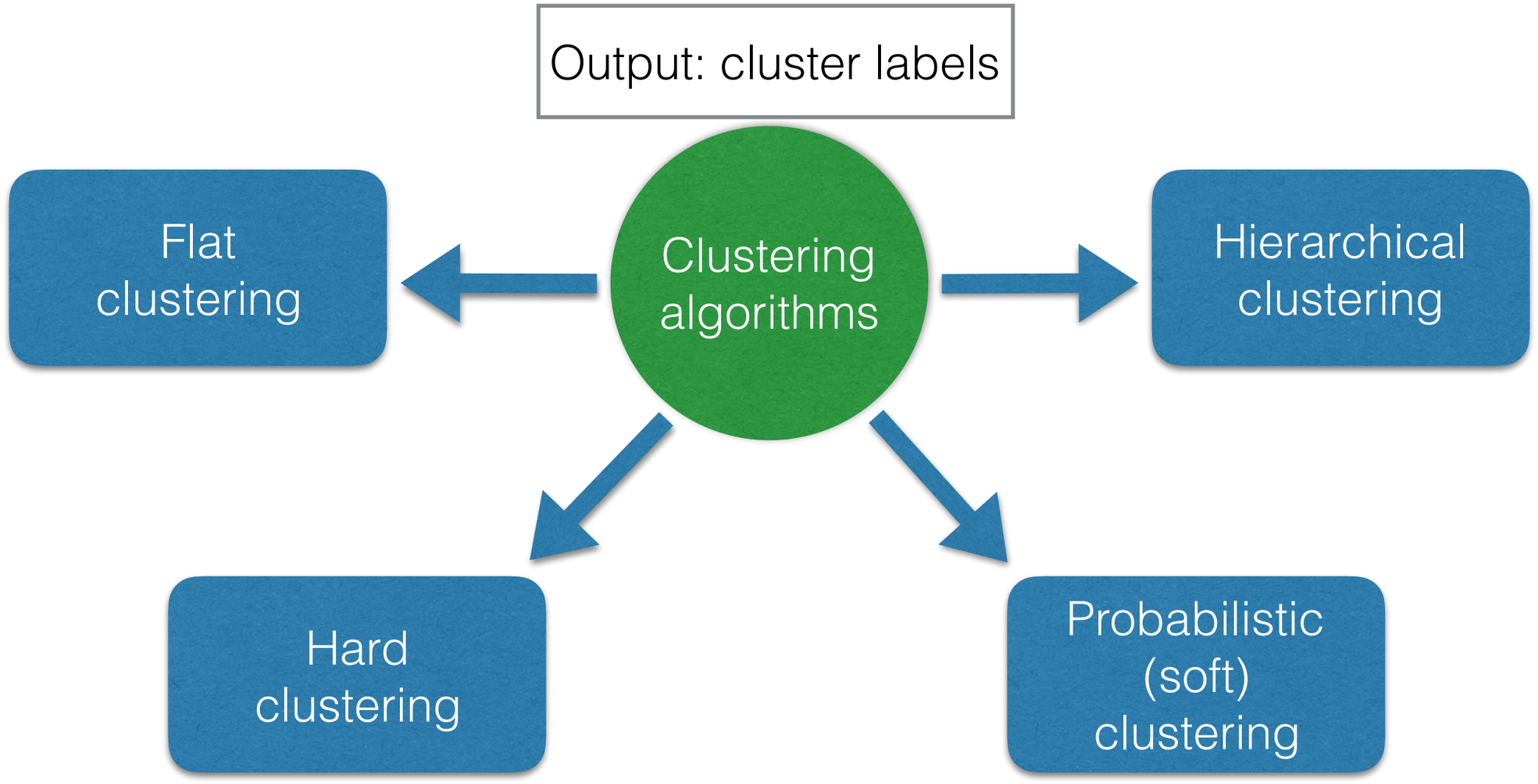
Types of clustering

Segment a set of companies, stocks, credit card holders, mortgage holders etc. into a relatively homogeneous groups where intra-group variance is generally smaller than inter-group variance.



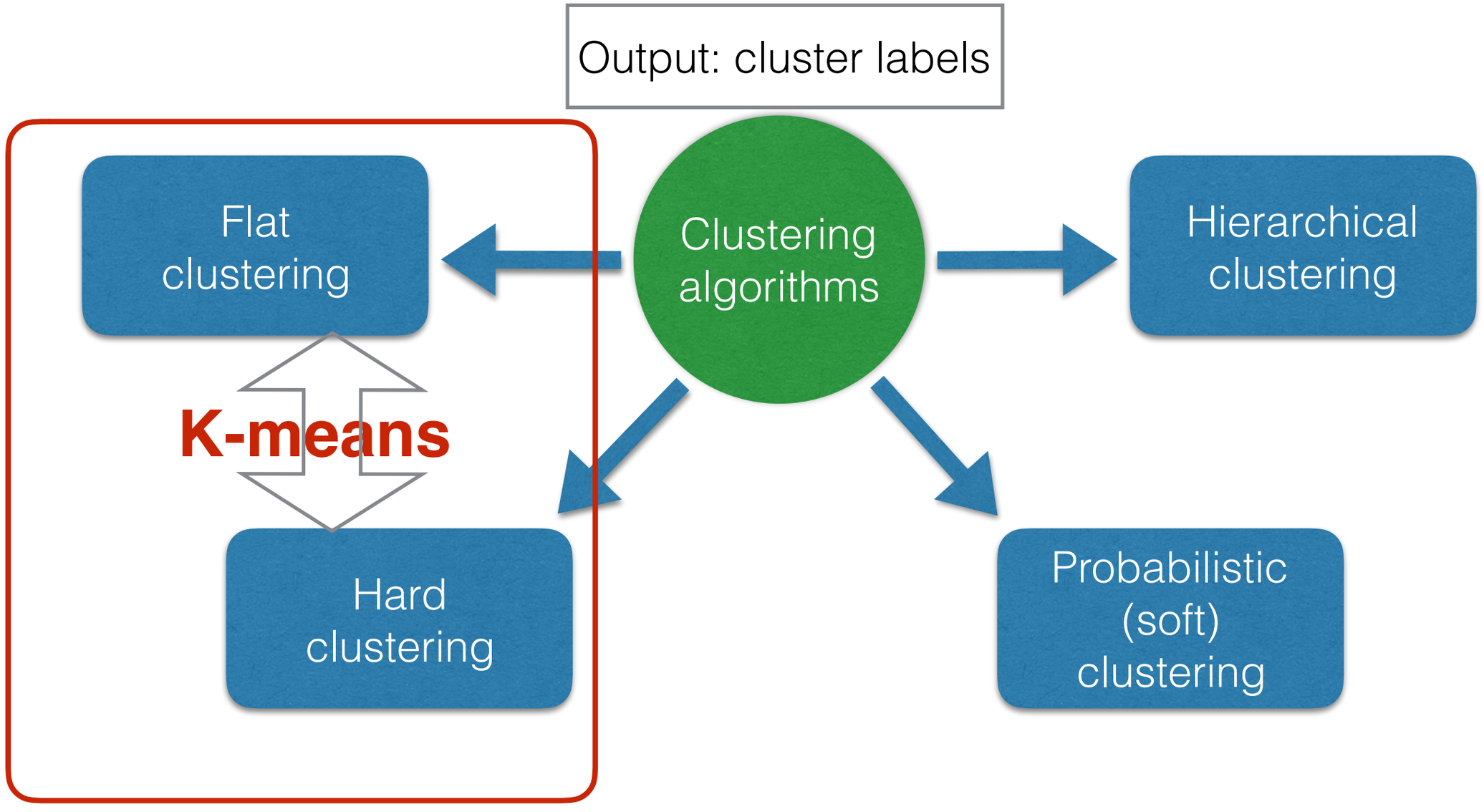
Types of clustering

Segment a set of companies, stocks, credit card holders, mortgage holders etc. into a relatively homogeneous groups where intra-group variance is generally smaller than inter-group variance.



K-means clustering

K-means is one of the most popular and scalable clustering algorithms



Control question

Select all correct answers

1. Clustering methods aggregate data points into relatively homogeneous groups of points that are more similar intra-cluster than inter-cluster.
2. With a Flat Clustering, all points within a given cluster are “equal”, but within a Hierarchical Clustering, some points within a cluster are “more equal” than the others, for example some points can form sub-clusters, etc.
3. A given clustering algorithm can be simultaneously Flat and Hard.
4. A given clustering algorithm can be simultaneously Hard and Soft.

Correct answers: 1, 2, 3