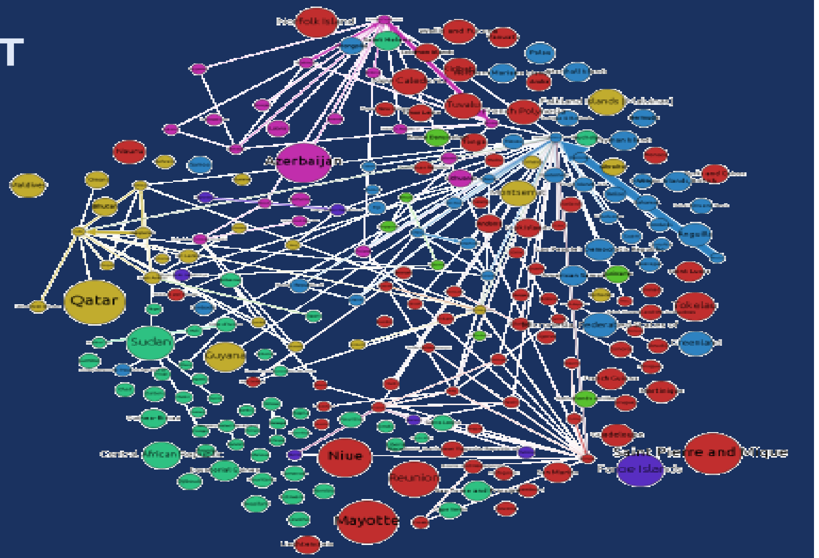
IS422P - DATA MINING
10 - CLUSTERING

AMIRA REZK
INFORMATION SYSTEM DEPARTMENT



AGENDA

The Basics
- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods
- K-Means

Hierarchical Methods
- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods
- DBSCAN

Grid-Based Methods

Evaluation of Clustering
- Assessing Clustering Tendency
- Measuring Clustering Quality

AGENDA

The Basics
- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods
- K-Means

Hierarchical Methods
- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods
- DBSCAN

Grid-Based Methods

Evaluation of Clustering
- Assessing Clustering Tendency
- Measuring Clustering Quality

WHAT IS CLUSTER ANALYSIS?

- Partitioning a set of data objects into subsets or clusters
  - objects in a cluster are similar, yet dissimilar to objects in other clusters
- **Goal**: discovery of previously unknown groups within the data
- Clusters are implicit classes
- **Applications** → business intelligence, image pattern recognition, web search, biology, security
- Clustering can be used for pre-processing and outlier detection

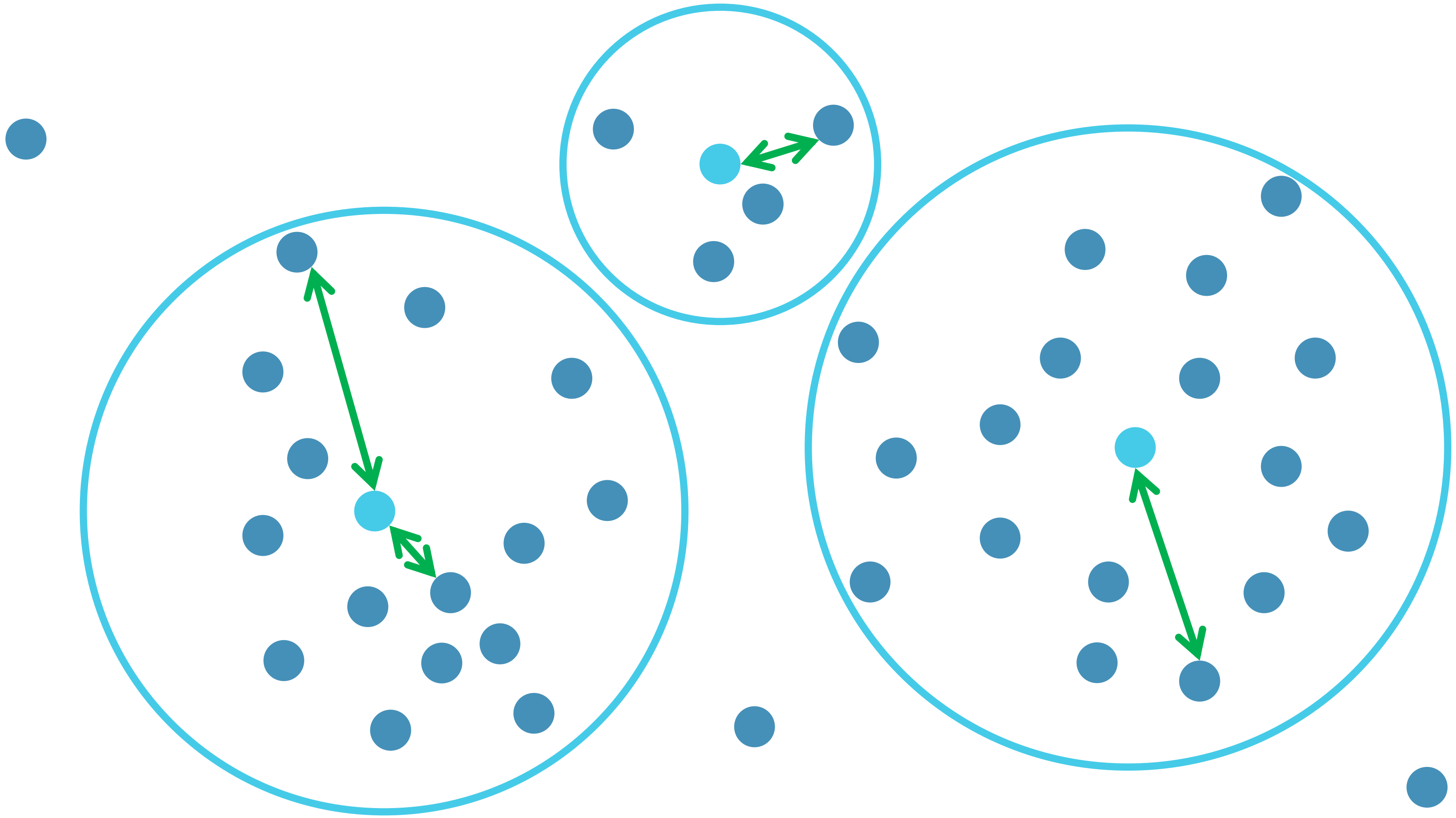
REQUIREMENTS FOR CLUSTER ANALYSIS

- Scalability → currently handles small datasets, uses sampling
- Handling different attribute types → mostly numerical
- Discovering clusters with arbitrary shape → currently mostly spherical
- Domain knowledge & input parameters → # clusters & clustering results
- Handling noisy data → currently sensitive to noise
- Incremental clustering & insensitivity to input order →new data requires re-computing clusters from scratch – sensitive to order
- Handling high-dimensionality data → mostly low Dimensionality
- Constraint-based clustering → little support for domain constraints
- Interpretability & usability → are results comprehensible & usable?

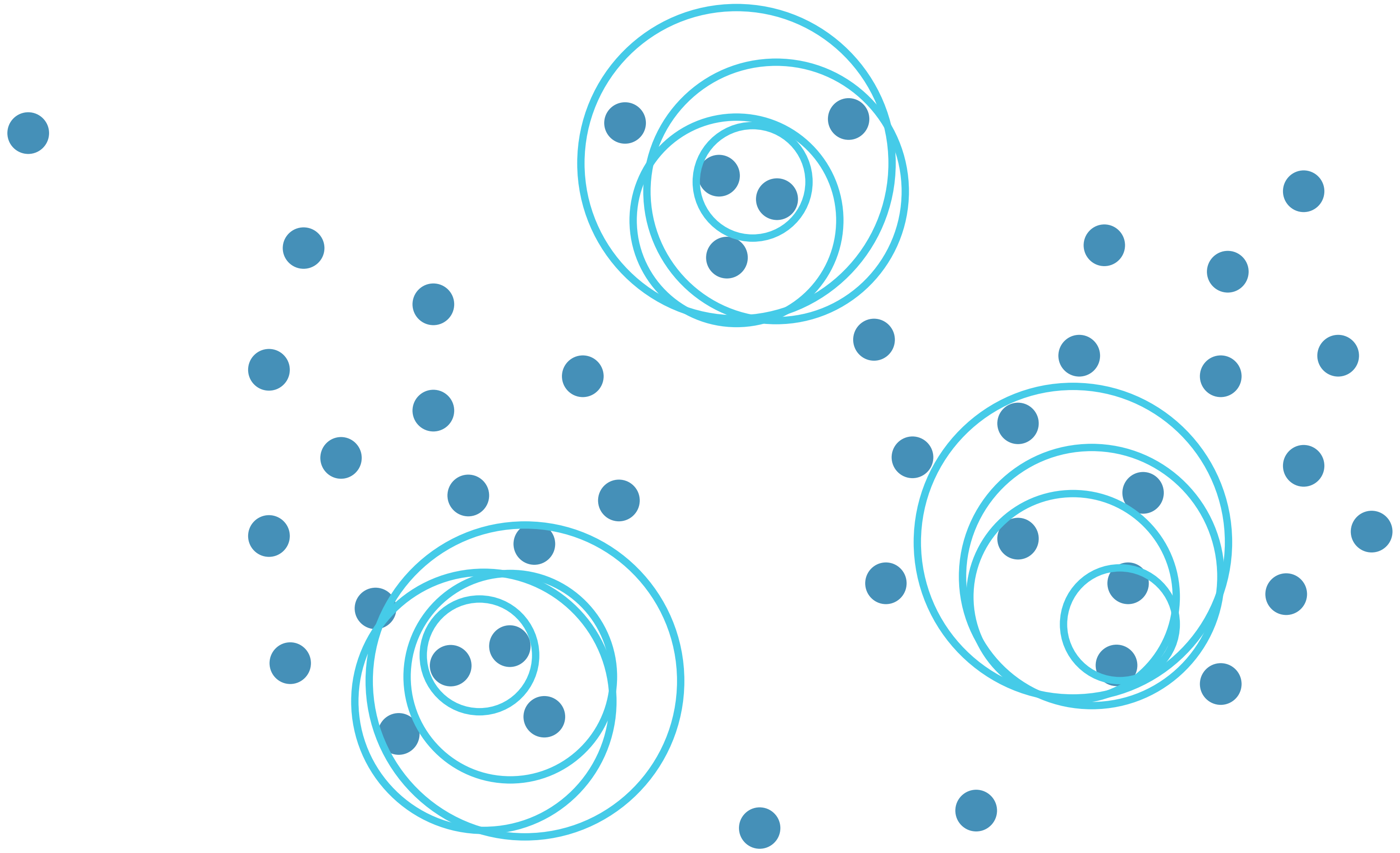
COMPARING CLUSTER ANALYSIS METHODS

- The partitioning criteria – **flat** or **hierarchical**?
- Separation of clusters – **mutually exclusive** or **overlapping**?
- Similarity measure – **distance** or **connectivity/density**?

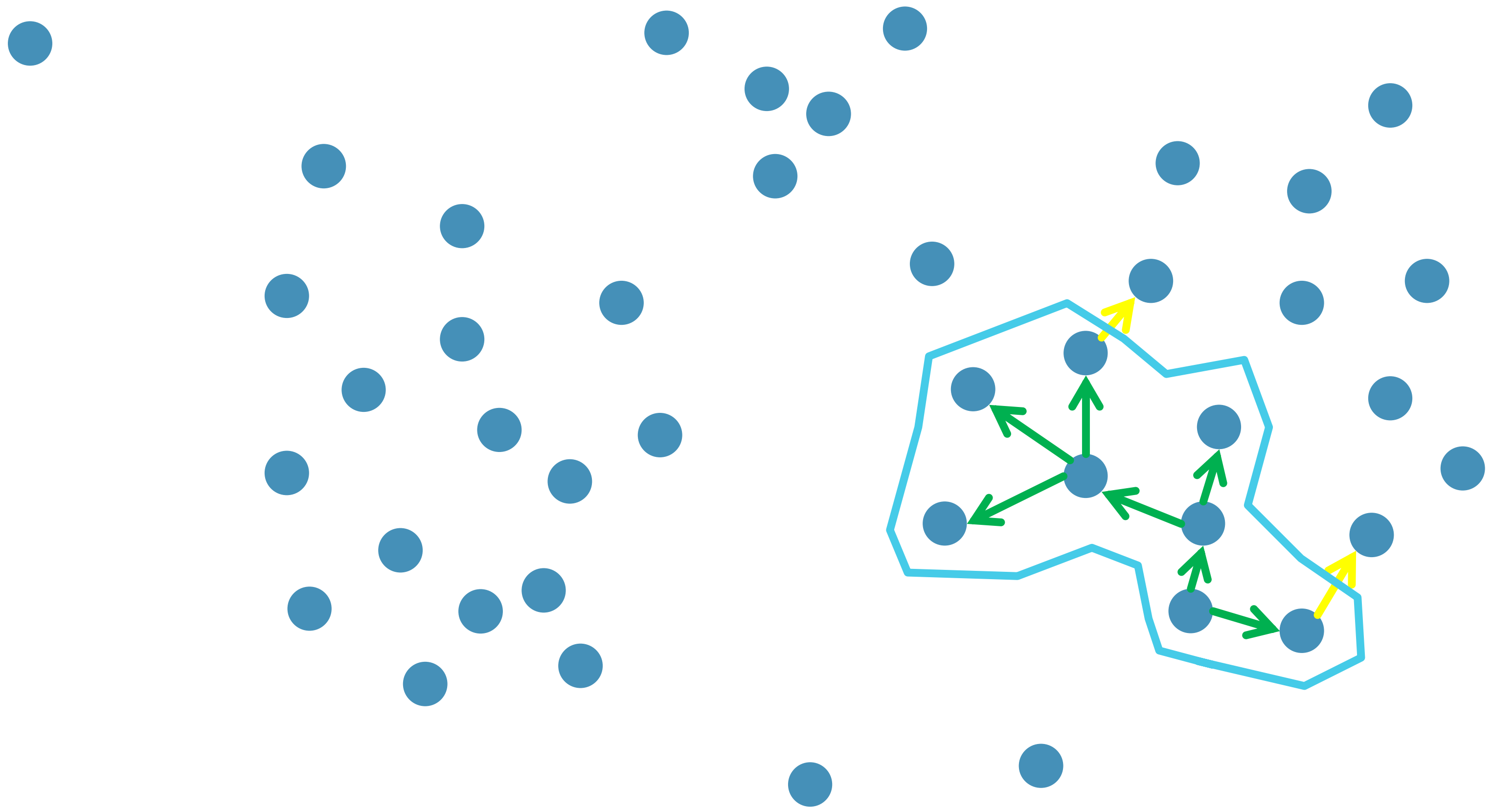
OVERVIEW OF CLUSTER ANALYSIS METHODS
PARTITIONING



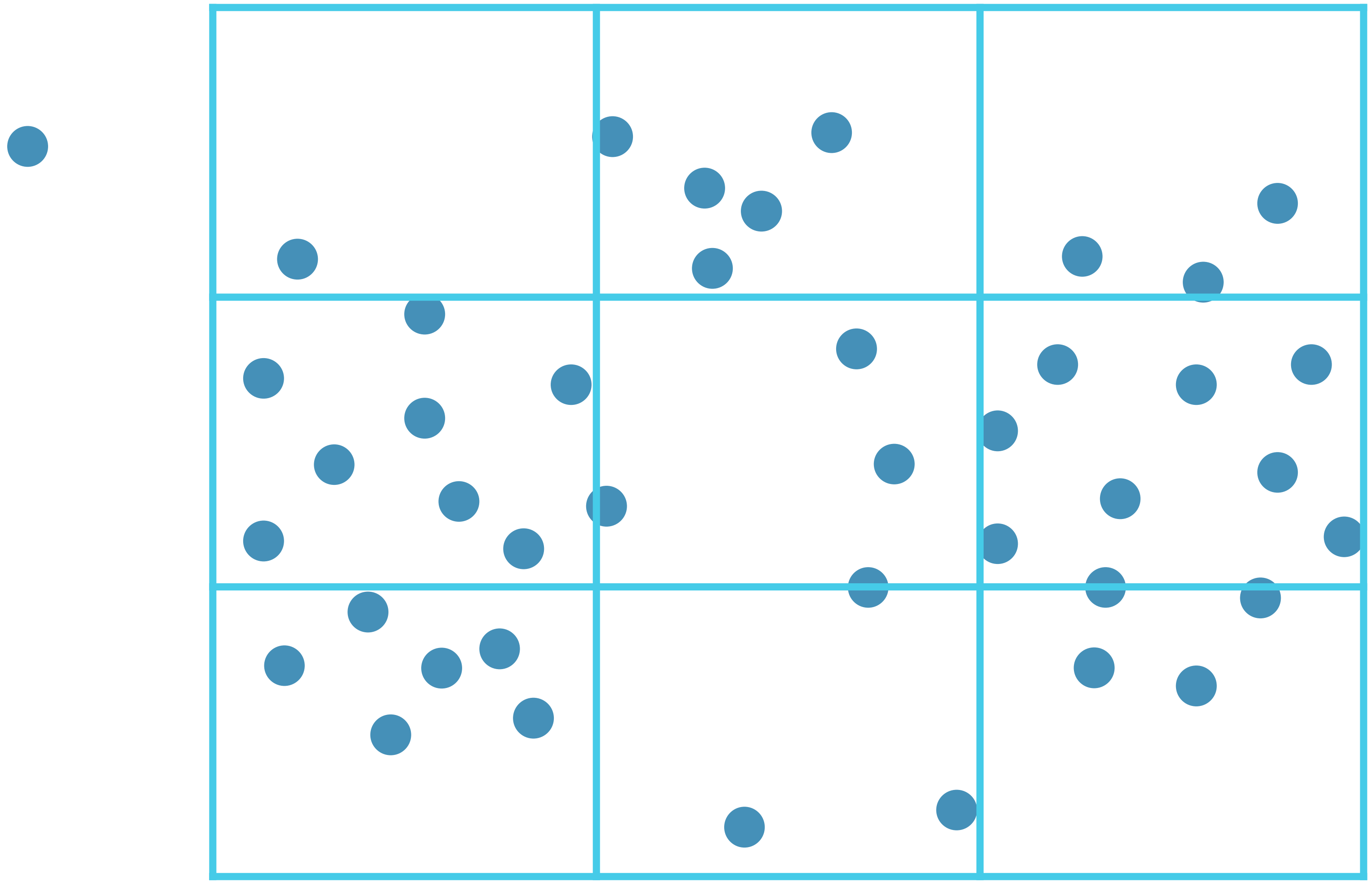
OVERVIEW OF CLUSTER ANALYSIS METHODS
HIERARCHICAL



OVERVIEW OF CLUSTER ANALYSIS METHODS DENSITY-BASED



OVERVIEW OF CLUSTER ANALYSIS METHODS
GRID-BASED



OVERVIEW OF CLUSTER ANALYSIS METHODS

Method	Characteristics
Partitioning methods	— Find *mutually exclusive* clusters of *spherical shape* — *Distance-based* — May *use mean or medoid* to represent cluster center — Effective for *small- to medium-size data sets*
Hierarchical methods	— Clustering is *hierarchy* involving multiple levels — Cannot correct *erroneous merges/splits* — May consider object "*linkages*"
Density-based methods	— Can find *arbitrarily shaped clusters* — Clusters are *dense regions* separated by *low-density regions* — Each point must have a *minimum number of points within its "neighborhood"* — May *filter out outliers*
Grid-based methods	— Use a multi-resolution *grid data structure* — *Fast processing time*

AGENDA

The Basics
- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods
- K-Means

Hierarchical Methods
- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods
- DBSCAN

Grid-Based Methods

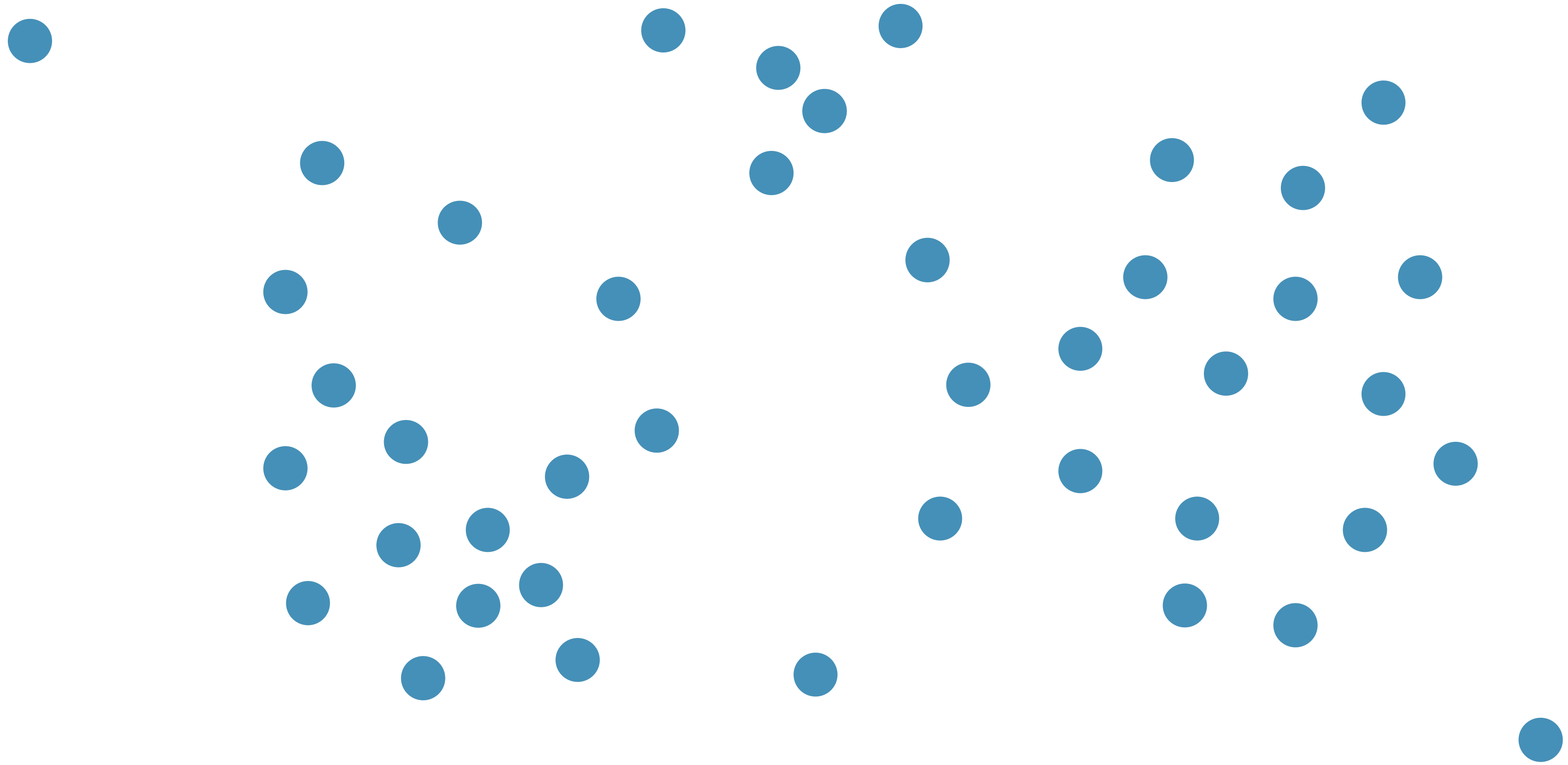
Evaluation of Clustering
- Assessing Clustering Tendency
- Measuring Clustering Quality

PARTITIONING METHODS
K-MEANS – A CENTROID-BASED TECHNIQUE

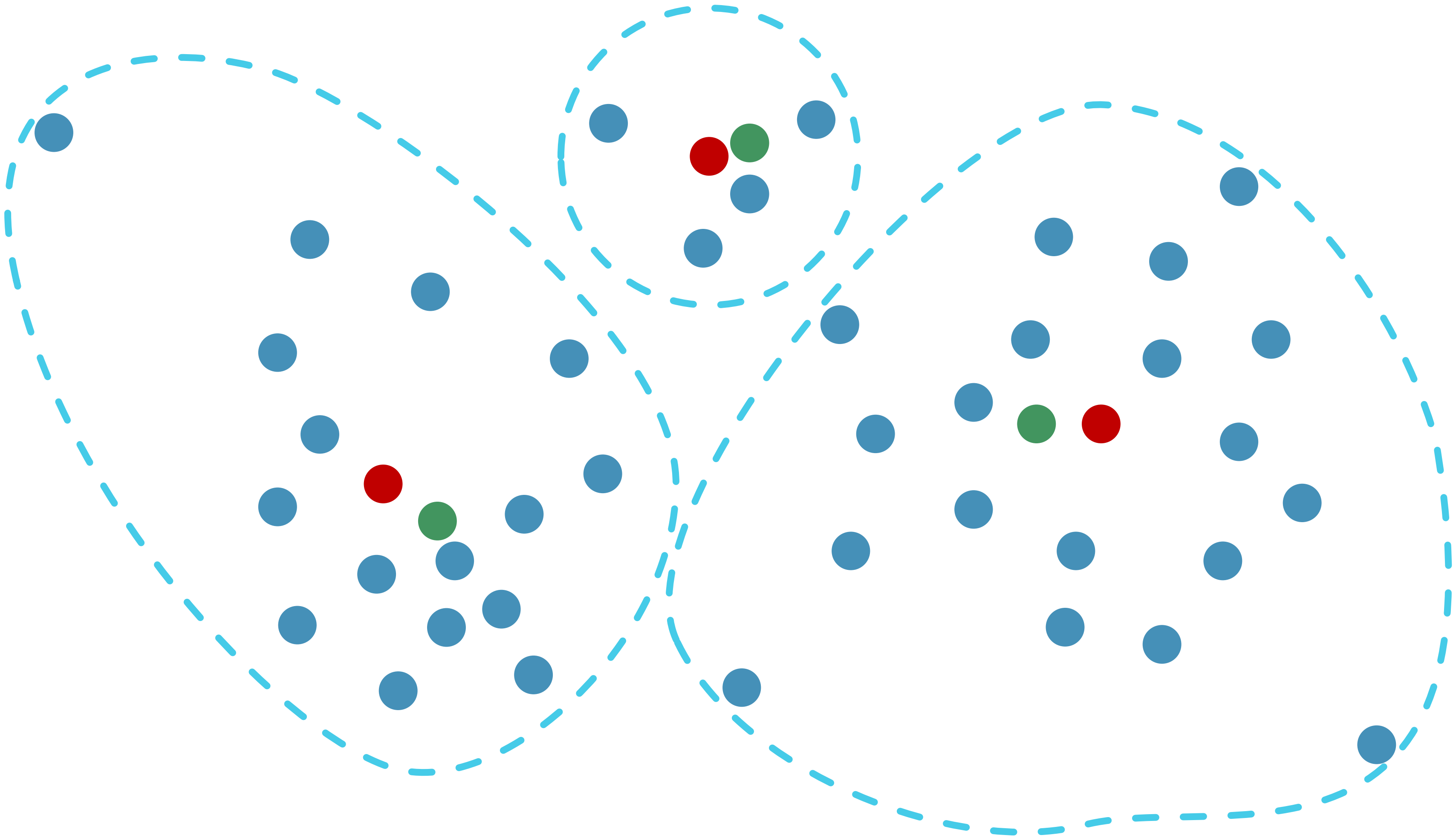
- Divide dataset into ***k*** **mutually exclusive** clusters
- Clusters are represented by their **centroids**
  - A centroid is a **cluster's center point**
- In *k*-means → centroid is **mean** of points within cluster
  - Each object ***x*** in cluster has a distance from centroid $\boldsymbol{c_i}$ → $dist(x, c_i)$
  - ***x*** is assigned to most similar cluster → $\boldsymbol{C_i}$ with ***min*** $dist(x, c_i)$
  - Cluster means are updated, then assignment is repeated
- To measure cluster quality → minimize sum of squared errors

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, c_i)^2$$

PARTITIONING METHODS
K-MEANS



PARTITIONING METHODS
K-MEANS



PARTITIONING METHODS
K-MEANS

Algorithm: ***k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

PARTITIONING METHODS
K-MEANS

Cluster the eight points in table using k-means. Assume that k = 3 and that initially the points are assigned to clusters as follows: C1 = {x1, x2, x3}, C2 = {x4, x5, x6}, C3 = {x7, x8}.

• Apply the k-means algorithm until convergence (i.e., until the clusters do not change), using the Manhattan distance.

(Hint: The Manhattan distance is: $d(i, j) = |x_{i1}-x_{j1}|+ |x_{i2}-x_{j2}|+ ....+ | x_{in}-x_{jn}|$.) Make sure you clearly identify the final clustering and show your steps.

	A1	A2
x1	2	10
x2	2	5
x3	8	4
x4	5	8
x5	7	5
x6	6	4
x7	1	2
x8	4	9

PARTITIONING METHODS
K-MEANS

- $C1 = \{x1, x2, x3\} = \{(2,10), (2,5), (8,4)\}$
  - Mean of $C1 = (\frac{2+2+8}{3}, \frac{10+5+4}{3}) = (4, 6\frac{1}{3})$
- $C2 = \{x4, x5, x6\} = \{(5,8), (7,5), (6,4)\}$
  - Mean of $C2 = (6, 5\frac{2}{3})$
- $C3 = \{x7, x8\} = \{(1,2), (4,9)\}$
  - Mean of $C3 = (2\frac{1}{2}, 5\frac{1}{2})$

	A1	A2
x1	2	10
x2	2	5
x3	8	4
x4	5	8
x5	7	5
x6	6	4
x7	1	2
x8	4	9

PARTITIONING METHODS
K-MEANS

	X1 (2,10)	X2 (2,5)	X3 (8,4)	X4 (5,8)	X5 (7,5)	X6 (6,4)	X7 (1,2)	X8 (4,9)	NEW MEAN
C1 $(4,6\frac{1}{3})$	$5\frac{2}{3}$	$3\frac{1}{3}$	$6\frac{1}{3}$	$2\frac{2}{3}$	$4\frac{1}{3}$	$4\frac{1}{3}$	$7\frac{1}{3}$	$2\frac{2}{3}$	$4\frac{1}{2}$, $8\frac{1}{2}$
C2 $(6,5\frac{2}{3})$	$8\frac{1}{3}$	$4\frac{2}{3}$	$3\frac{2}{3}$	$3\frac{1}{3}$	$1\frac{2}{3}$	$1\frac{2}{3}$	$8\frac{2}{3}$	$5\frac{1}{3}$	$7, 4\frac{1}{3}$
C3 $(2\frac{1}{2},5\frac{1}{2})$	5	1	7	5	5	5	5	5	$1\frac{2}{3}$, $5\frac{2}{3}$

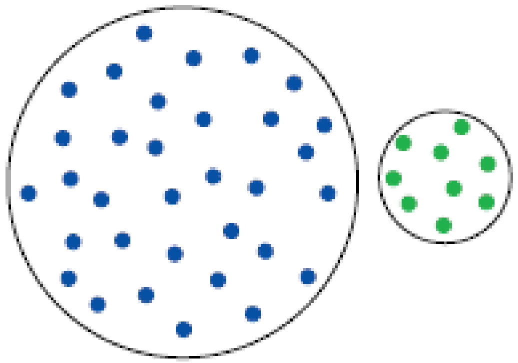
C1= {x1,x4,x8}={(2,10), (5,8), (4,9)} Mean of C1= $(2\frac{2}{3},9)$

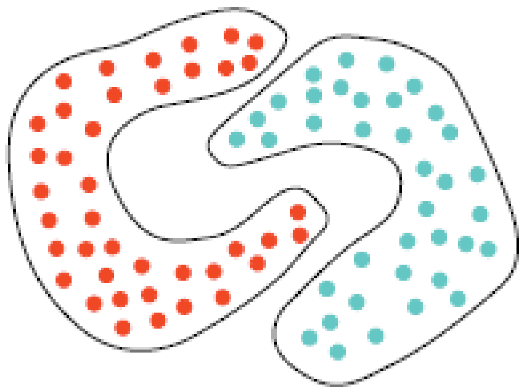
C2= {x3,x5,x6}={(8,4), (7,5), (6,4)} Mean of C2 = $(7, 4\frac{1}{3})$

C3= {x2,x7}={(2,5), (1,2)} Mean of C3 = $(1\frac{1}{2}, 3\frac{1}{2})$

PARTITIONING METHODS
K-MEANS

- Factors to consider:
- Selection of k
- Selection of initial centroids
- Calculation of dissimilarity
- Calculation of cluster means
- When it fails!
- Clusters with very different sizes & with concave shapes





AGENDA

The Basics
- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods
- K-Means

Hierarchical Methods
- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods
- DBSCAN

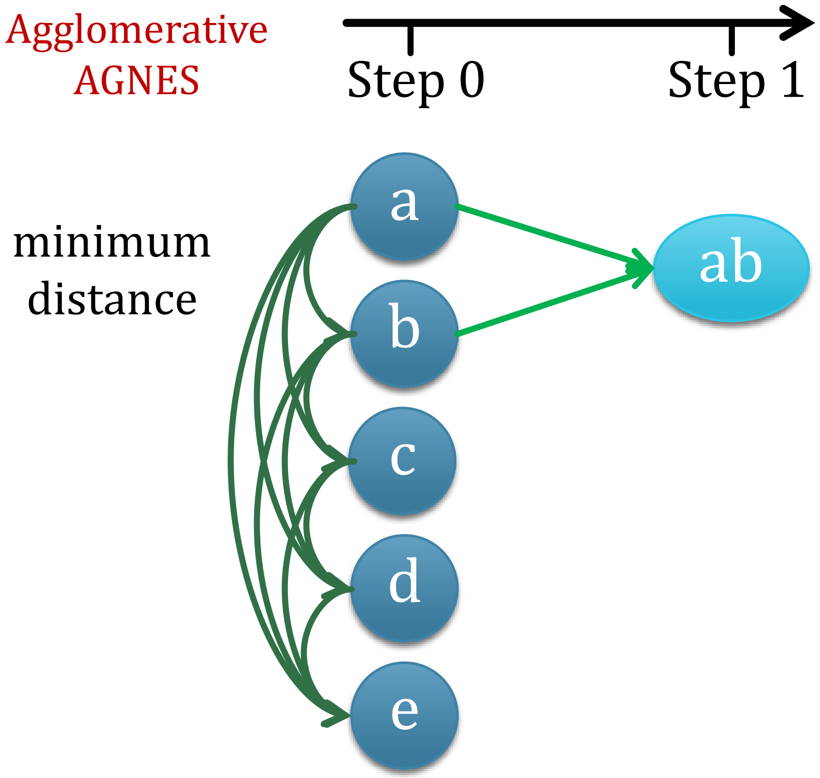
Grid-Based Methods

Evaluation of Clustering
- Assessing Clustering Tendency
- Measuring Clustering Quality

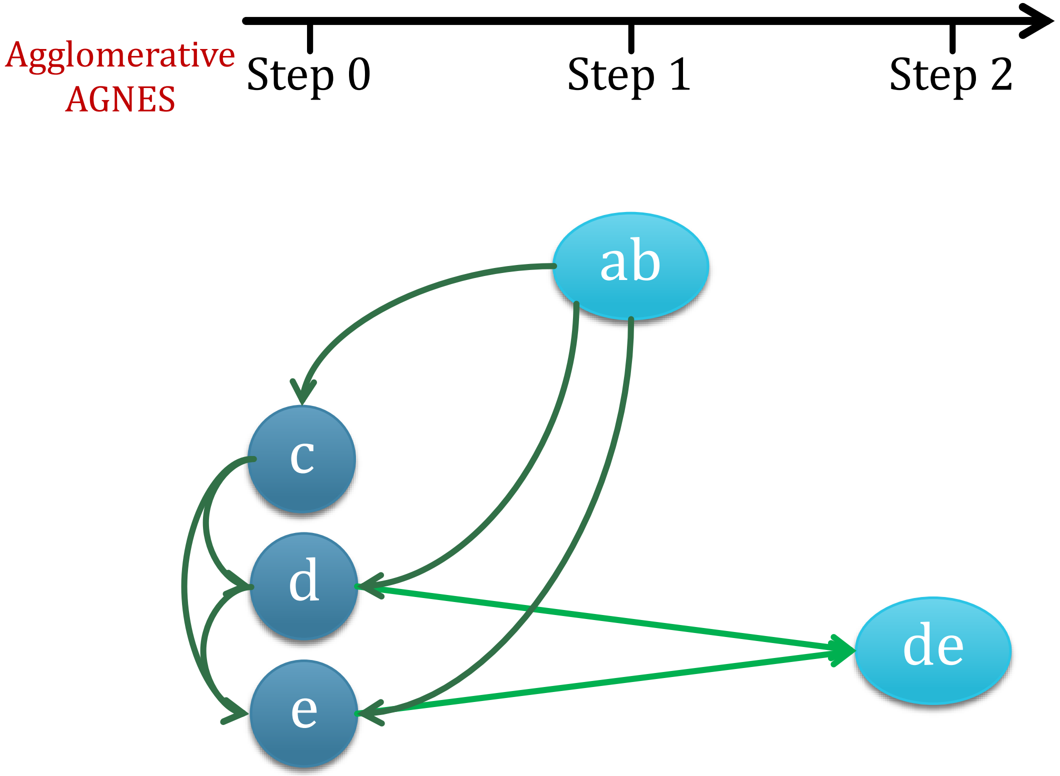
HIERARCHICAL METHODS
AGGLOMERATIVE VERSUS DIVISIVE CLUSTERING

- Hierarchical clustering → group data objects into a hierarchy or "tree" of clusters
- Agglomerative → bottom-up (merge) composition
  - Each object has its own cluster
  - Two clusters that are closest merged into a bigger cluster
  - Iteratively merge till termination condition or single cluster is formed
- Divisive → top-down (split) composition
  - All objects in one big cluster
  - Divide into subclusters
  - Recursively divide subclusters into even smaller subclusters
  - Terminate when each object has his own cluster or objects in clusters are similar "enough"

HIERARCHICAL METHODS
AGGLOMERATIVE CLUSTERING

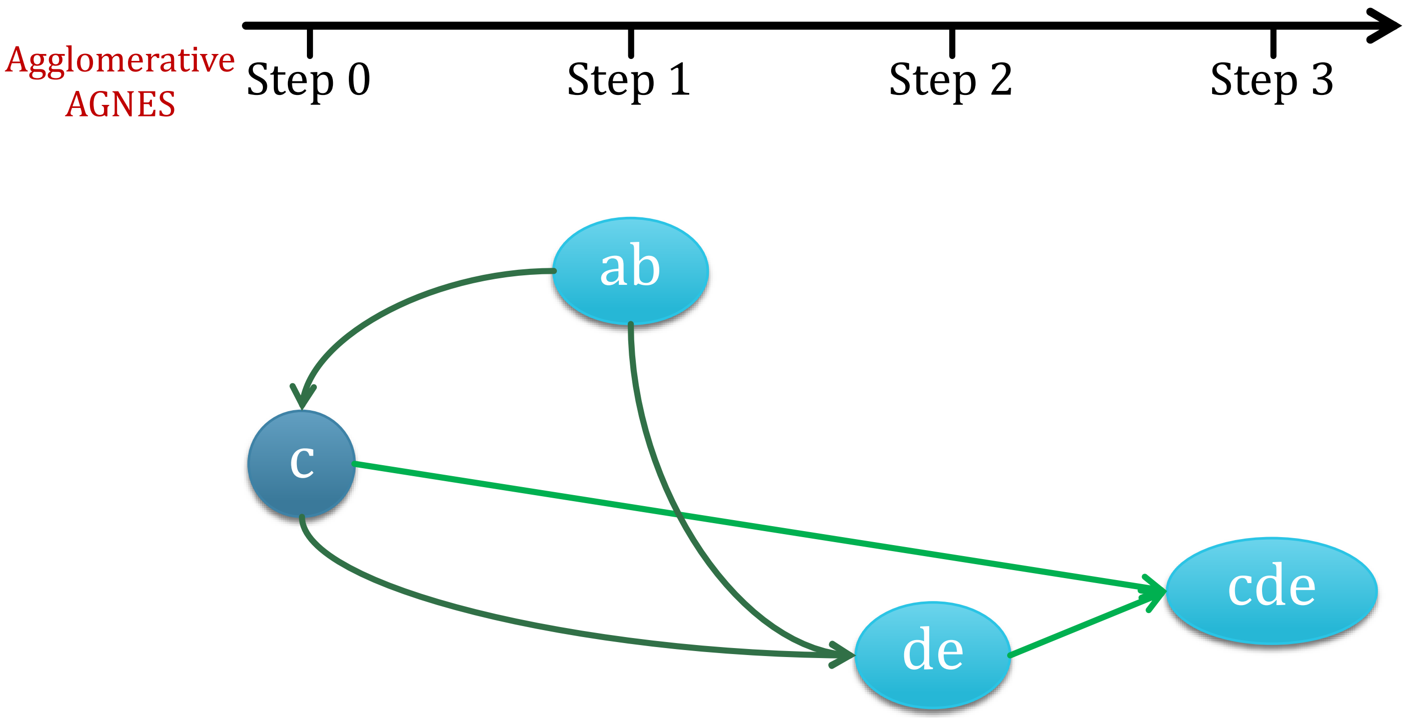


HIERARCHICAL METHODS
AGGLOMERATIVE CLUSTERING



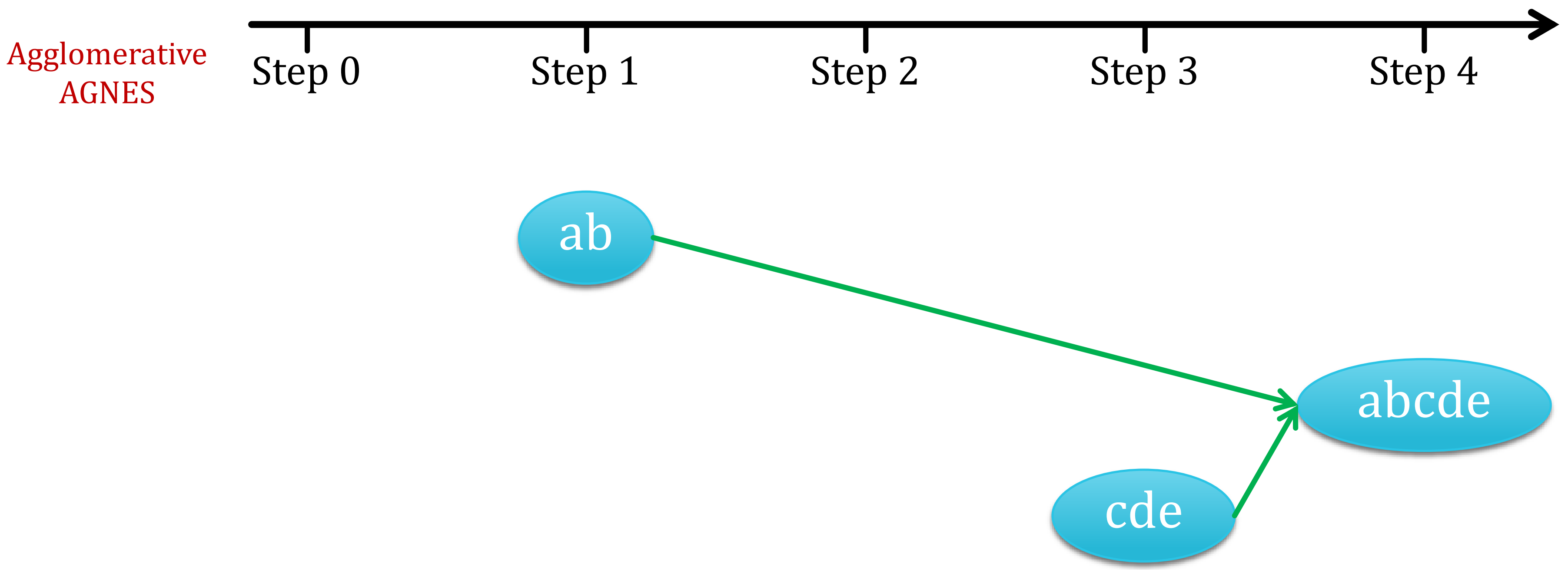
Measure distance between c, d, e and individual elements in cluster {a,b}, choose any with minimum distance (<u>single linkage</u>)

HIERARCHICAL METHODS
AGGLOMERATIVE CLUSTERING

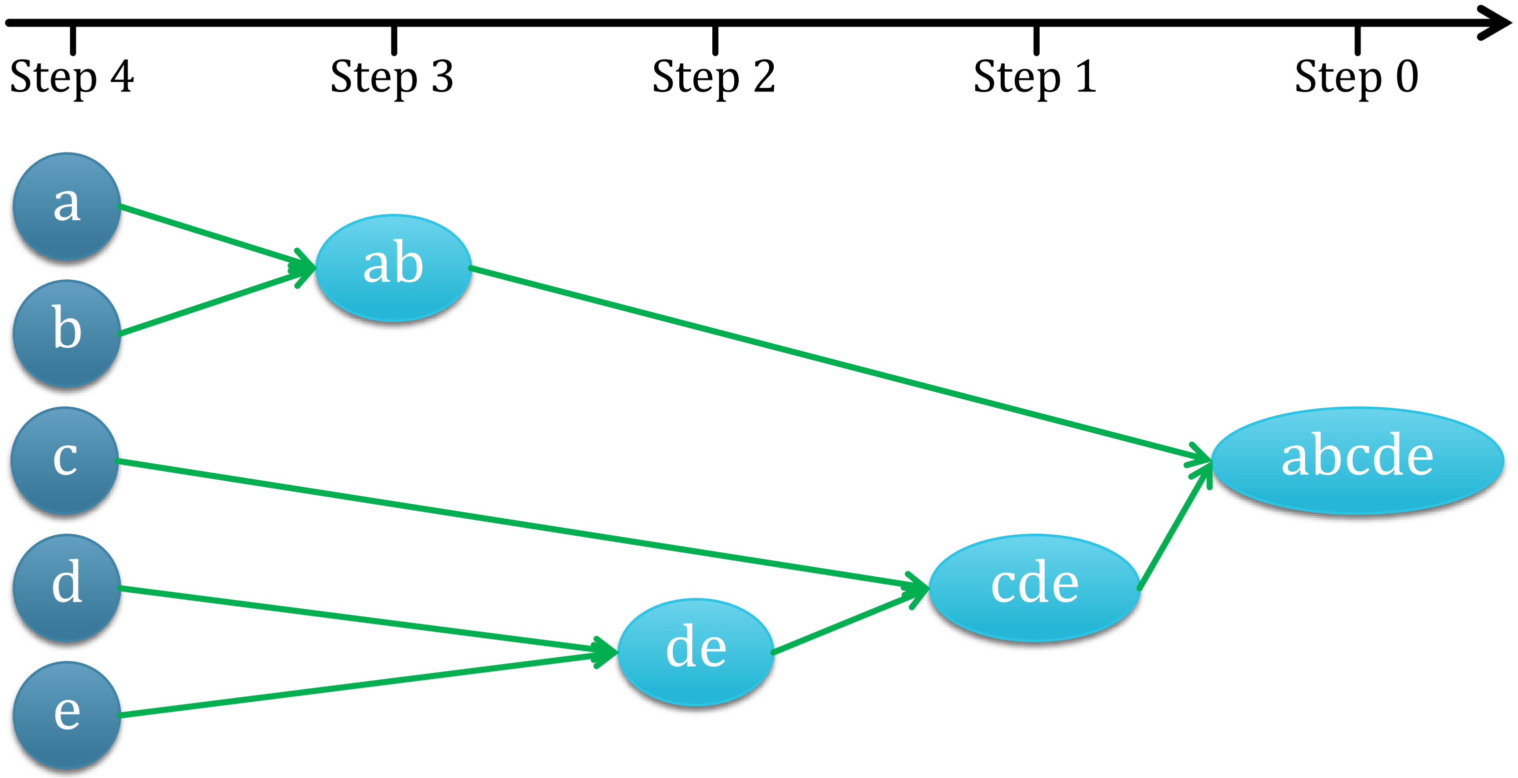


Measure distance between c and individual elements in cluster {a,b} and {d,e}, as well as distance between pairs in {a,b} and {d,e}, choose any with minimum distance (single linkage)

HIERARCHICAL METHODS
AGGLOMERATIVE CLUSTERING

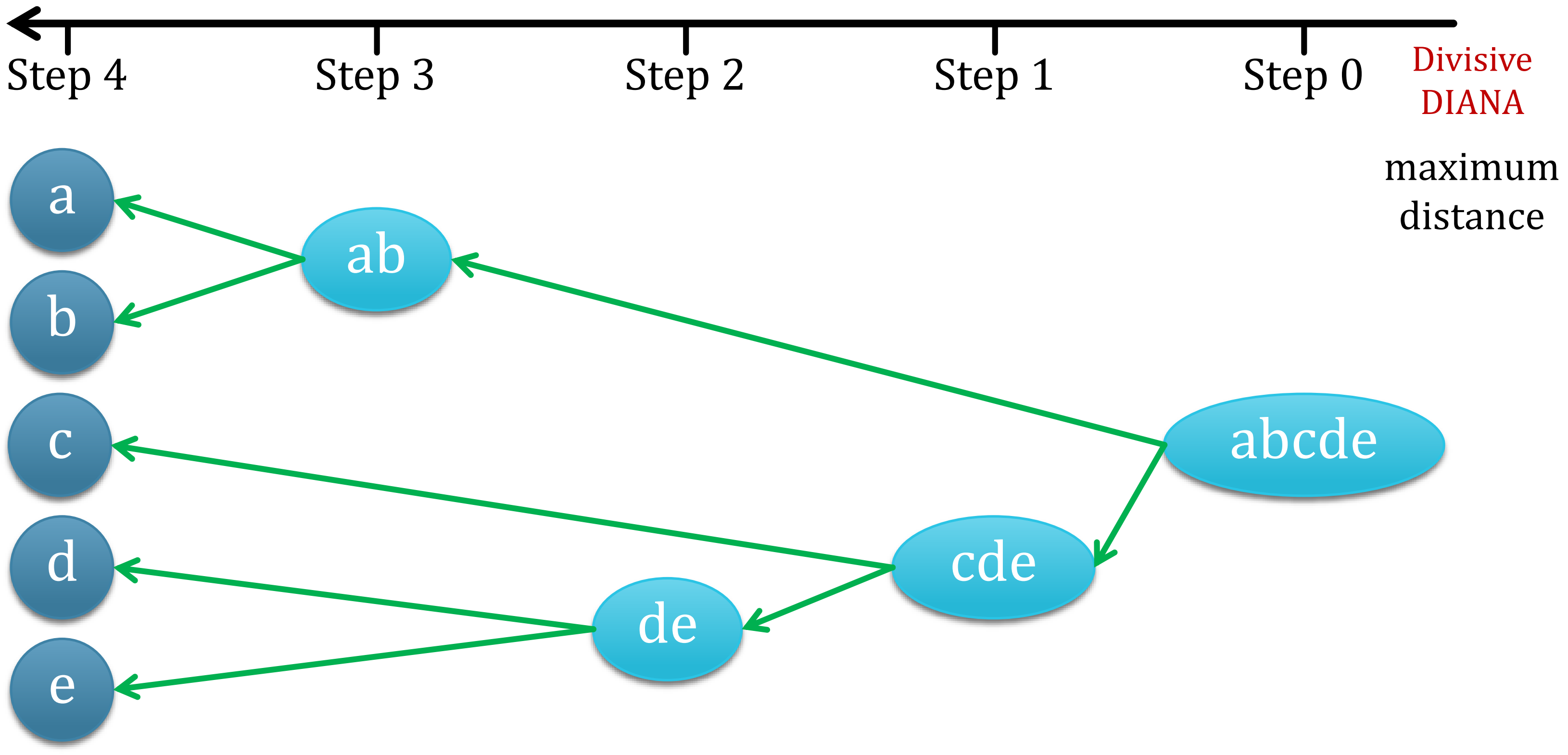


HIERARCHICAL METHODS
AGGLOMERATIVE CLUSTERING



Minimal spanning tree!

HIERARCHICAL METHODS
DIVISIVE CLUSTERING



How to divide a cluster is a challenge! Heuristic approaches may be used

QUESTION?

NEXT …….