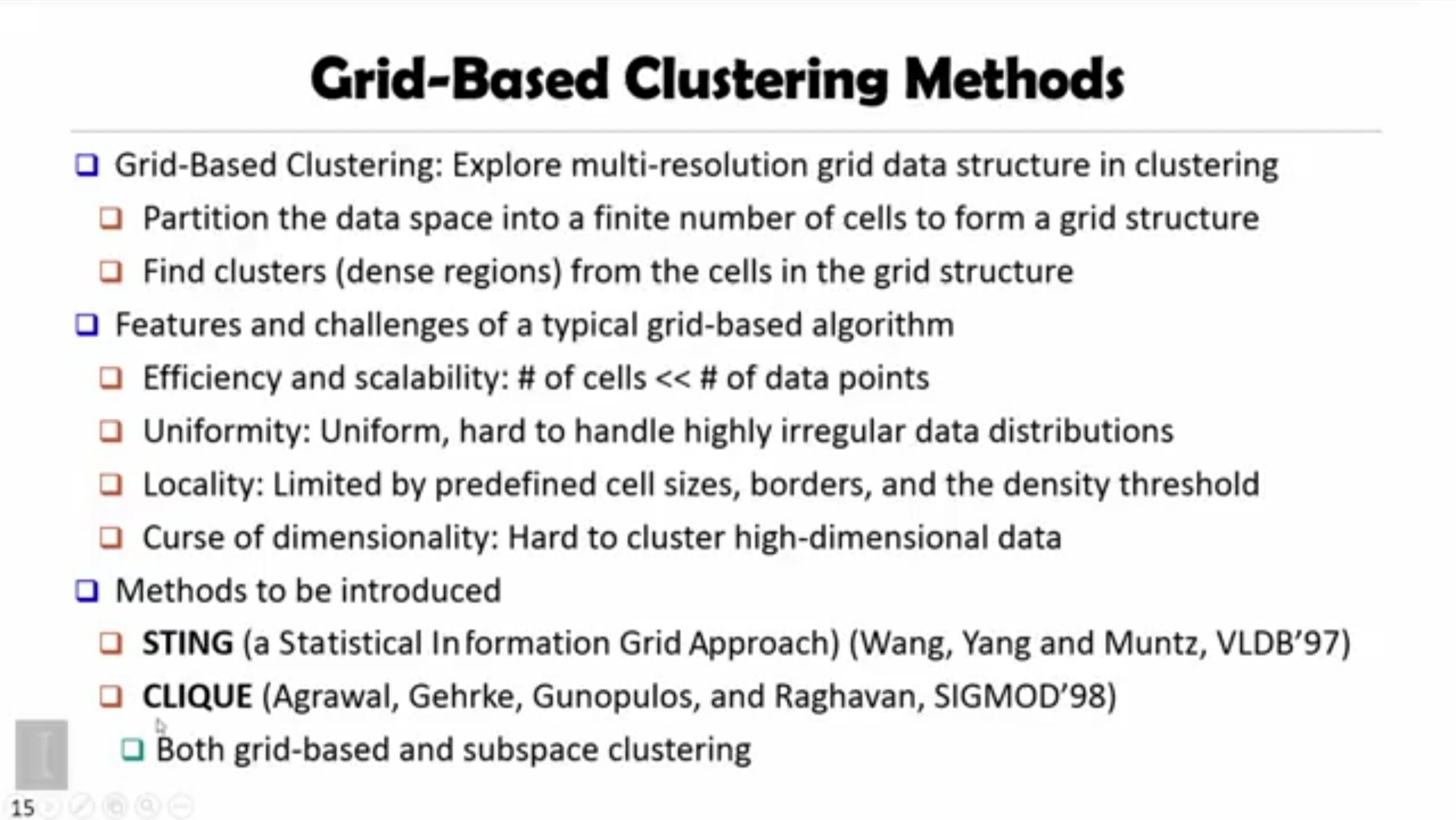
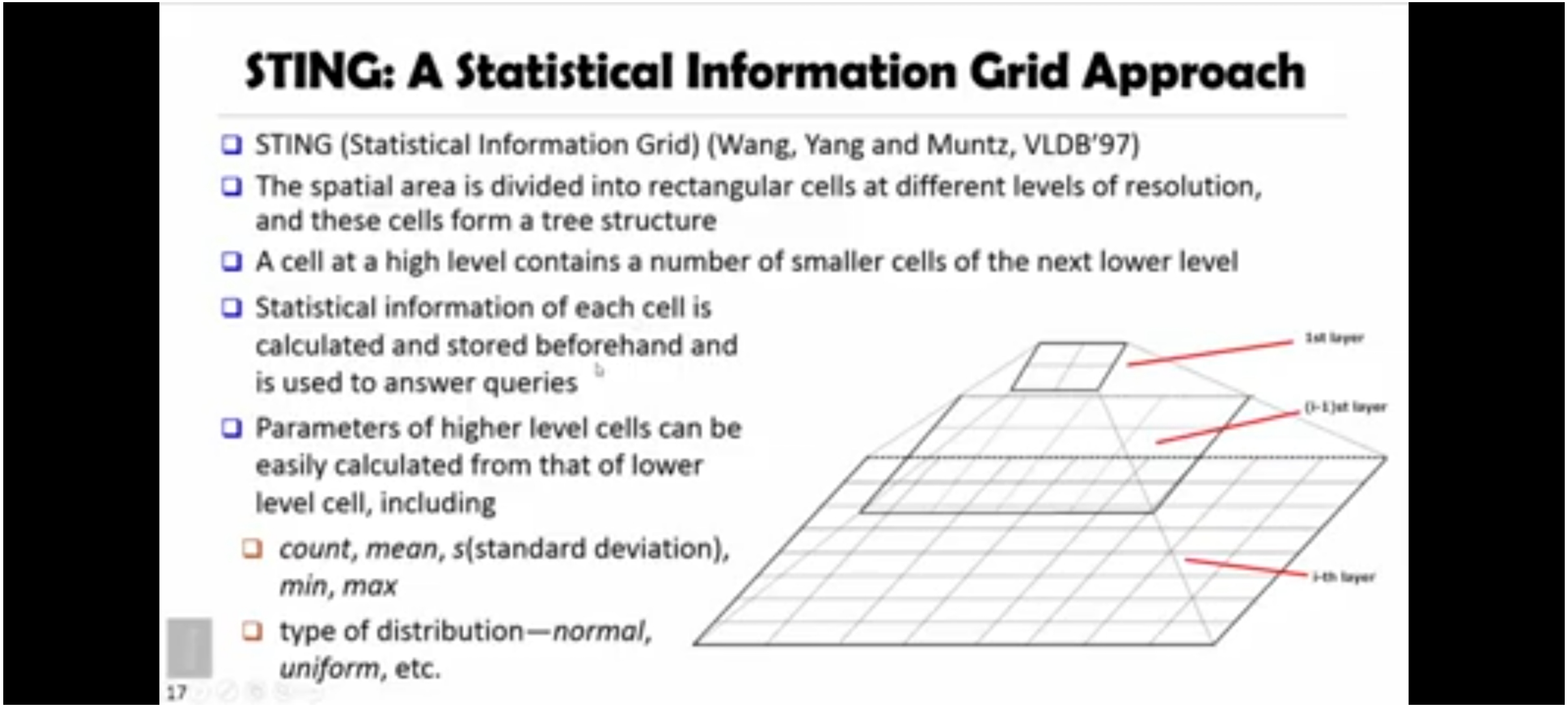


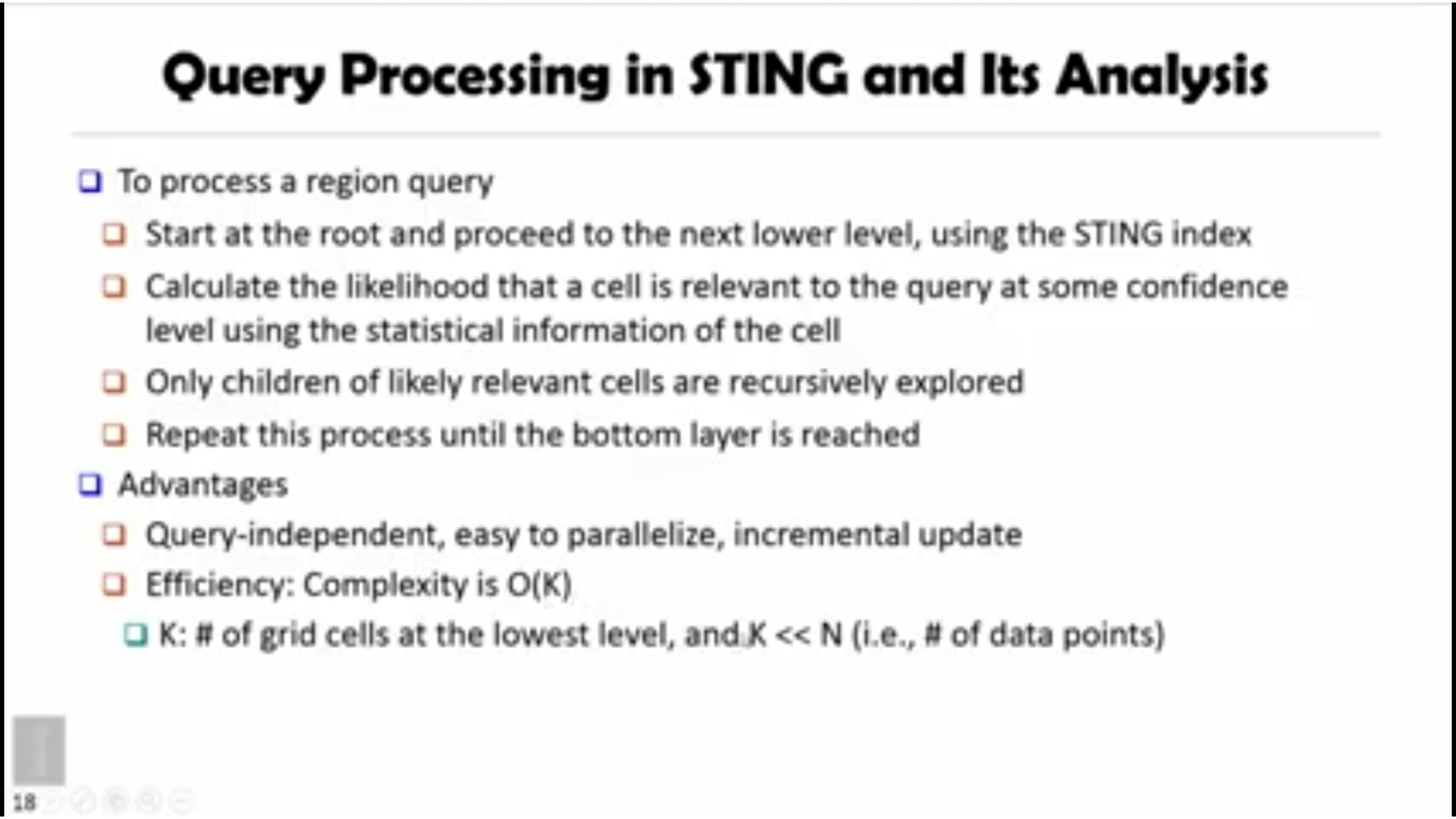
Grid-Based Clustering Methods

- ❑ Grid-Based Clustering: Explore multi-resolution grid data structure in clustering
  - ❑ Partition the data space into a finite number of cells to form a grid structure
  - ❑ Find clusters (dense regions) from the cells in the grid structure
- ❑ Features and challenges of a typical grid-based algorithm
  - ❑ Efficiency and scalability: # of cells << # of data points
  - ❑ Uniformity: Uniform, hard to handle highly irregular data distributions
  - ❑ Locality: Limited by predefined cell sizes, borders, and the density threshold
  - ❑ Curse of dimensionality: Hard to cluster high-dimensional data
- ❑ Methods to be introduced
  - ❑ **STING** (a Statistical Information Grid Approach) (Wang, Yang and Muntz, VLDB'97)
  - ❑ **CLIQUE** (Agrawal, Gehrke, Gunopulos, and Raghavan, SIGMOD'98)
    - ❑ Both grid-based and subspace clustering



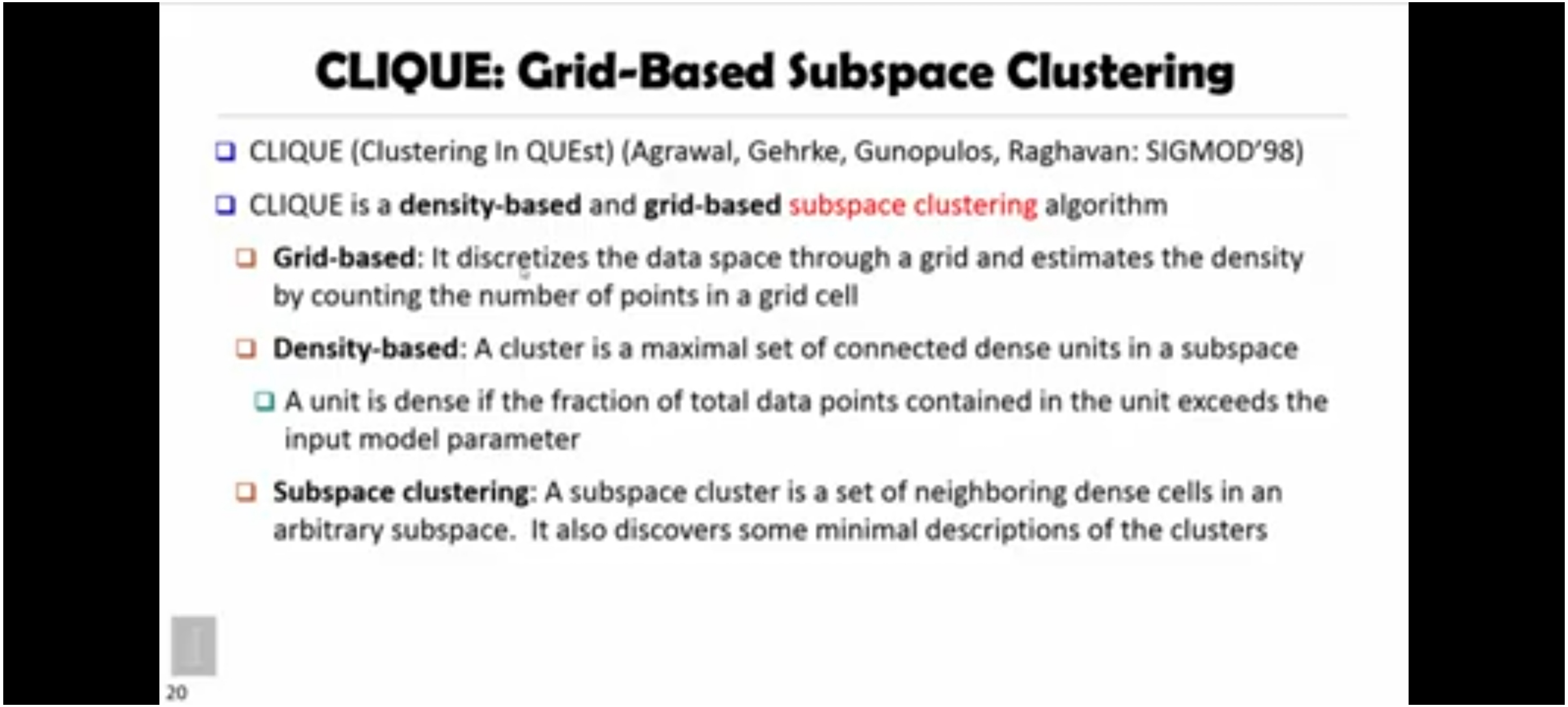
STING: A Statistical Information Grid Approach

- ❑ STING (Statistical Information Grid) (Wang, Yang and Muntz, VLDB'97)
- ❑ The spatial area is divided into rectangular cells at different levels of resolution, and these cells form a tree structure
- ❑ A cell at a high level contains a number of smaller cells of the next lower level
- ❑ Statistical information of each cell is calculated and stored beforehand and is used to answer queries
- ❑ Parameters of higher level cells can be easily calculated from that of lower level cell, including
  - ❑ *count*, *mean*, *s*(standard deviation), *min*, *max*
  - ❑ type of distribution—*normal*, *uniform*, etc.



Query Processing in STING and Its Analysis

- To process a region query
  - Start at the root and proceed to the next lower level, using the STING index
  - Calculate the likelihood that a cell is relevant to the query at some confidence level using the statistical information of the cell
  - Only children of likely relevant cells are recursively explored
  - Repeat this process until the bottom layer is reached
- Advantages
  - Query-independent, easy to parallelize, incremental update
  - Efficiency: Complexity is O(K)
    - K: # of grid cells at the lowest level, and K << N (i.e., # of data points)



CLIQUE: Grid-Based Subspace Clustering

- CLIQUE (Clustering In QUEst) (Agrawal, Gehrke, Gunopulos, Raghavan: SIGMOD'98)
- CLIQUE is a **density-based** and **grid-based** subspace clustering algorithm
  - **Grid-based**: It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell
  - **Density-based**: A cluster is a maximal set of connected dense units in a subspace
    - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
  - **Subspace clustering**: A subspace cluster is a set of neighboring dense cells in an arbitrary subspace. It also discovers some minimal descriptions of the clusters