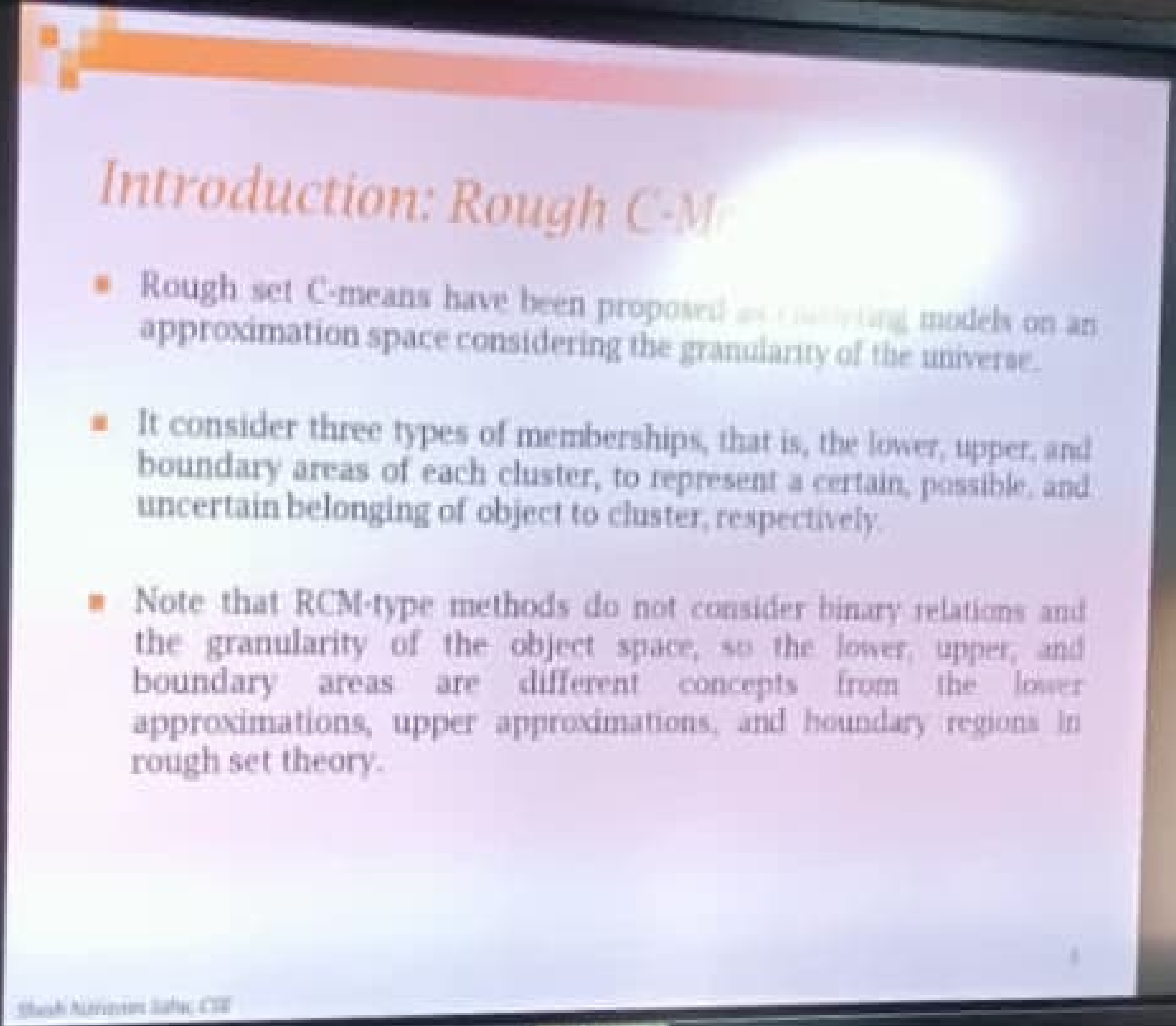


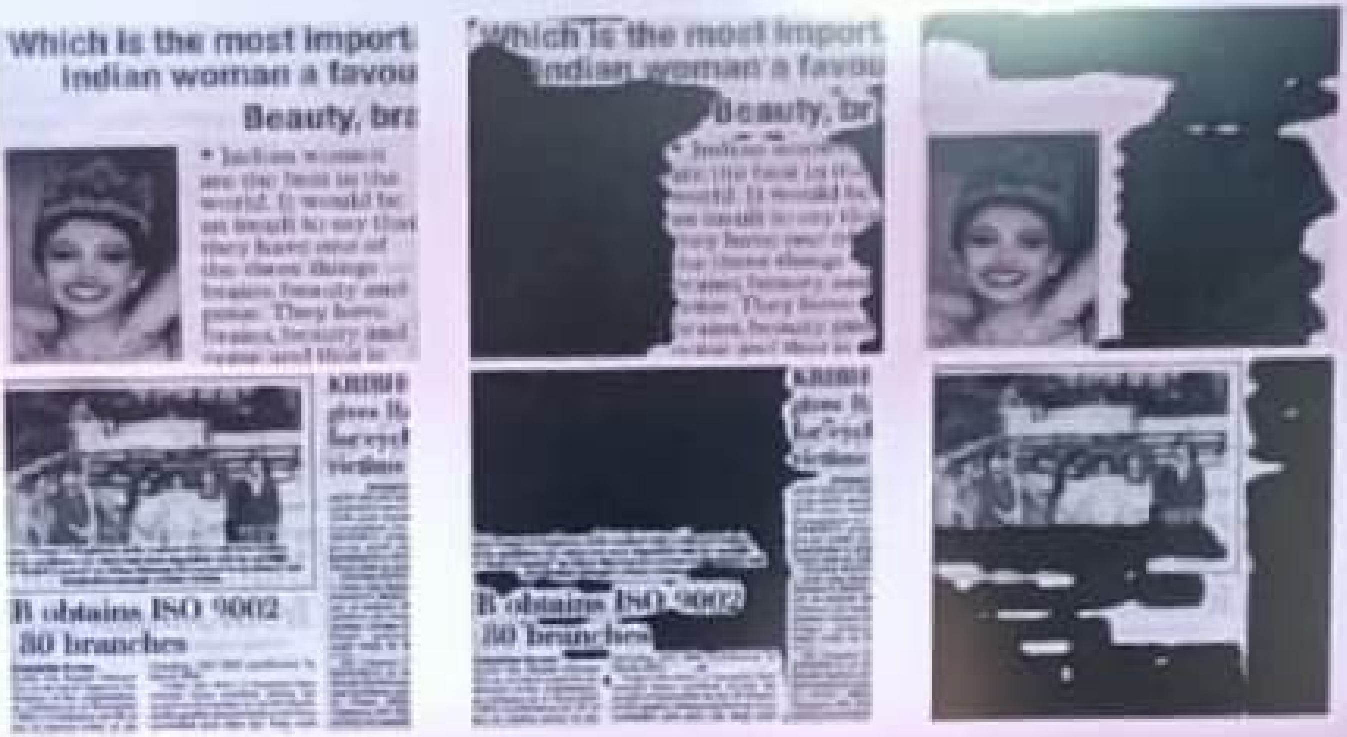
Introduction: Rough C-M

- Rough set C-means have been proposed as clustering models on an approximation space considering the granularity of the universe.

- It consider three types of memberships, that is, the lower, upper, and boundary areas of each cluster, to represent a certain, possible, and uncertain belonging of object to cluster, respectively.

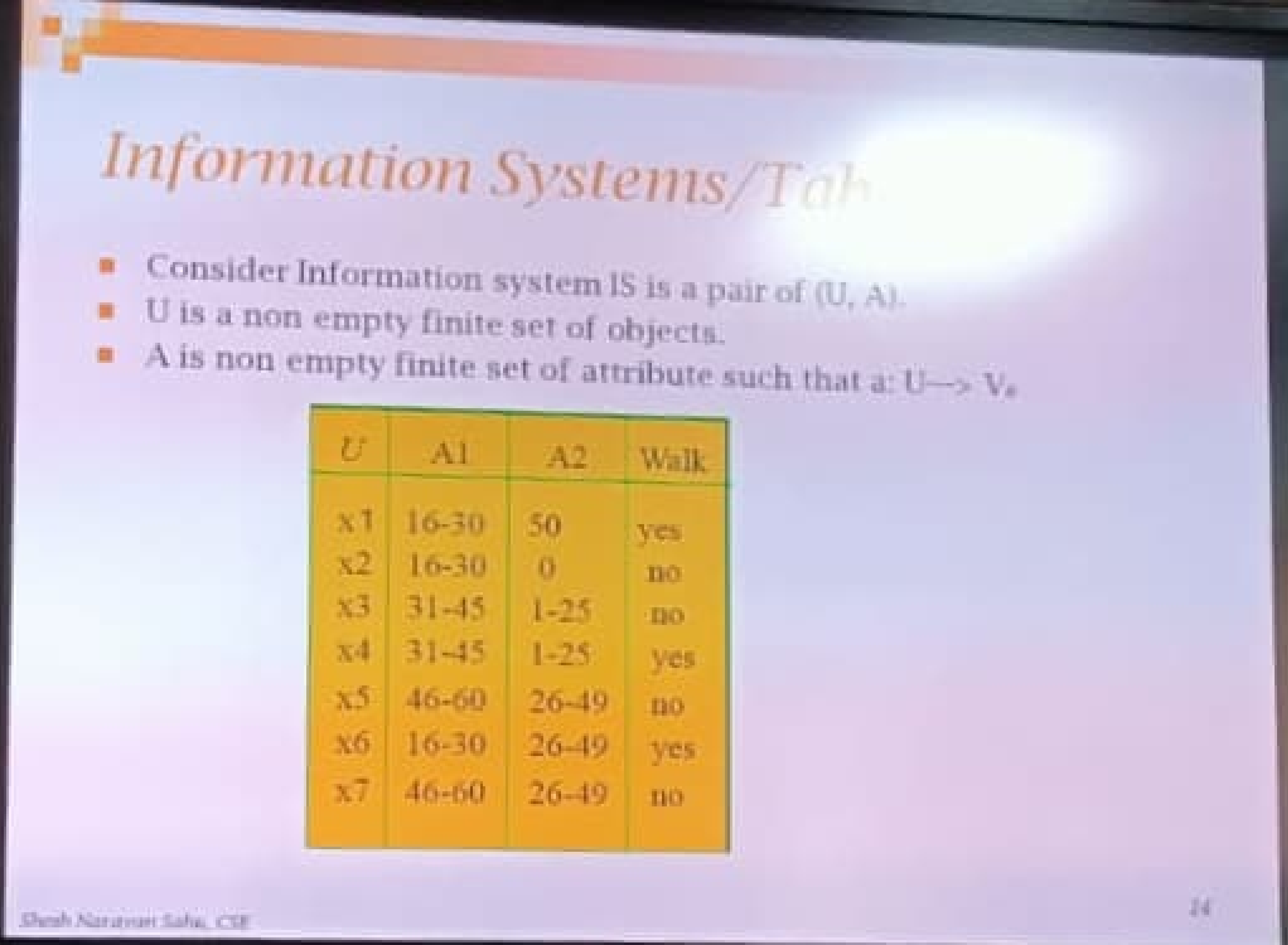
- Note that RCM-type methods do not consider binary relations and the granularity of the object space, so the lower, upper, and boundary areas are different concepts from the lower approximations, upper approximations, and boundary regions in rough set theory.

Example: Text-Graphics Segmentation Using FPCM



Concepts

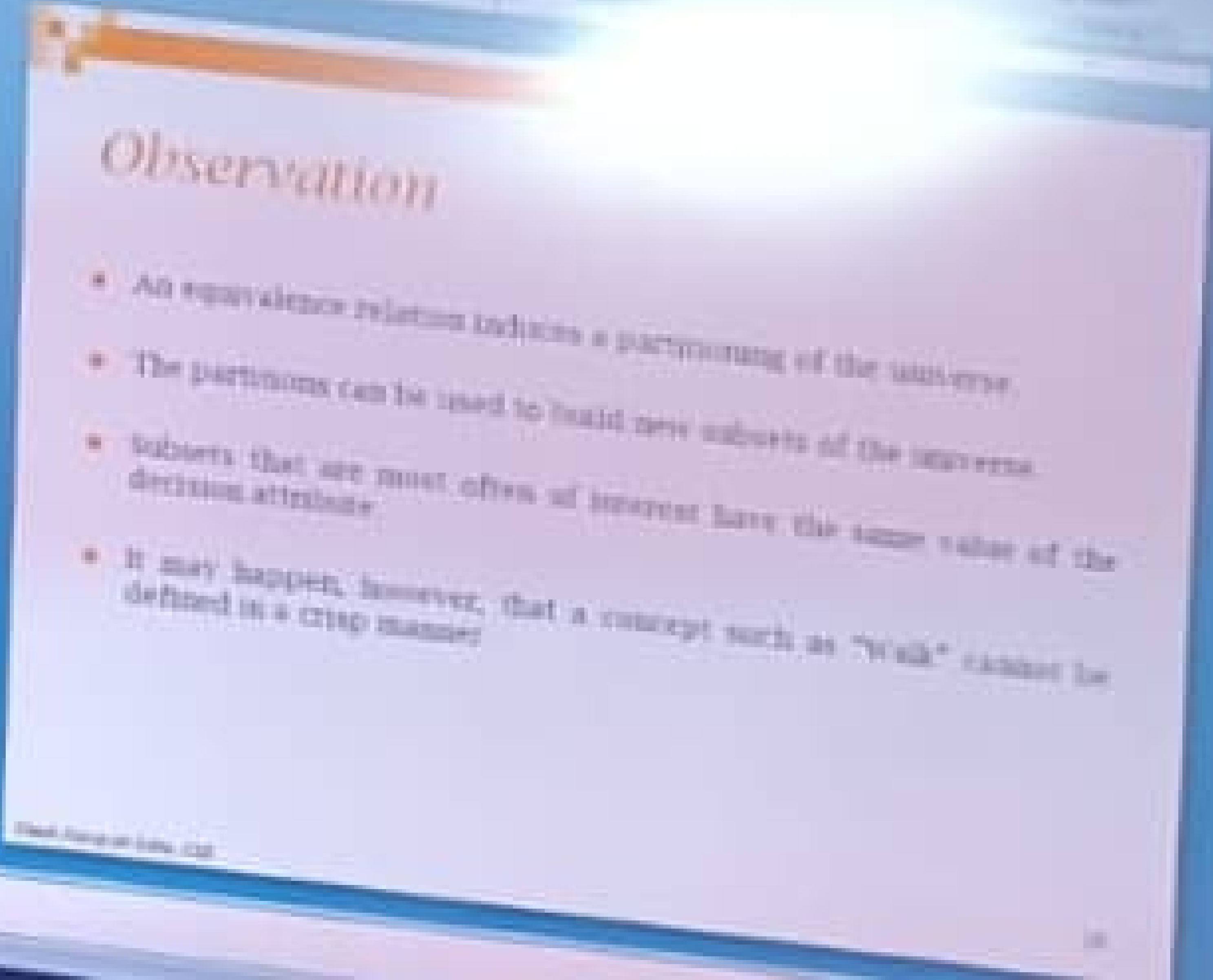
- Information/Decision system
- Indiscernibility
- Set Approximation
- Reduct and core
- Rough Membership



Information Systems/Tab

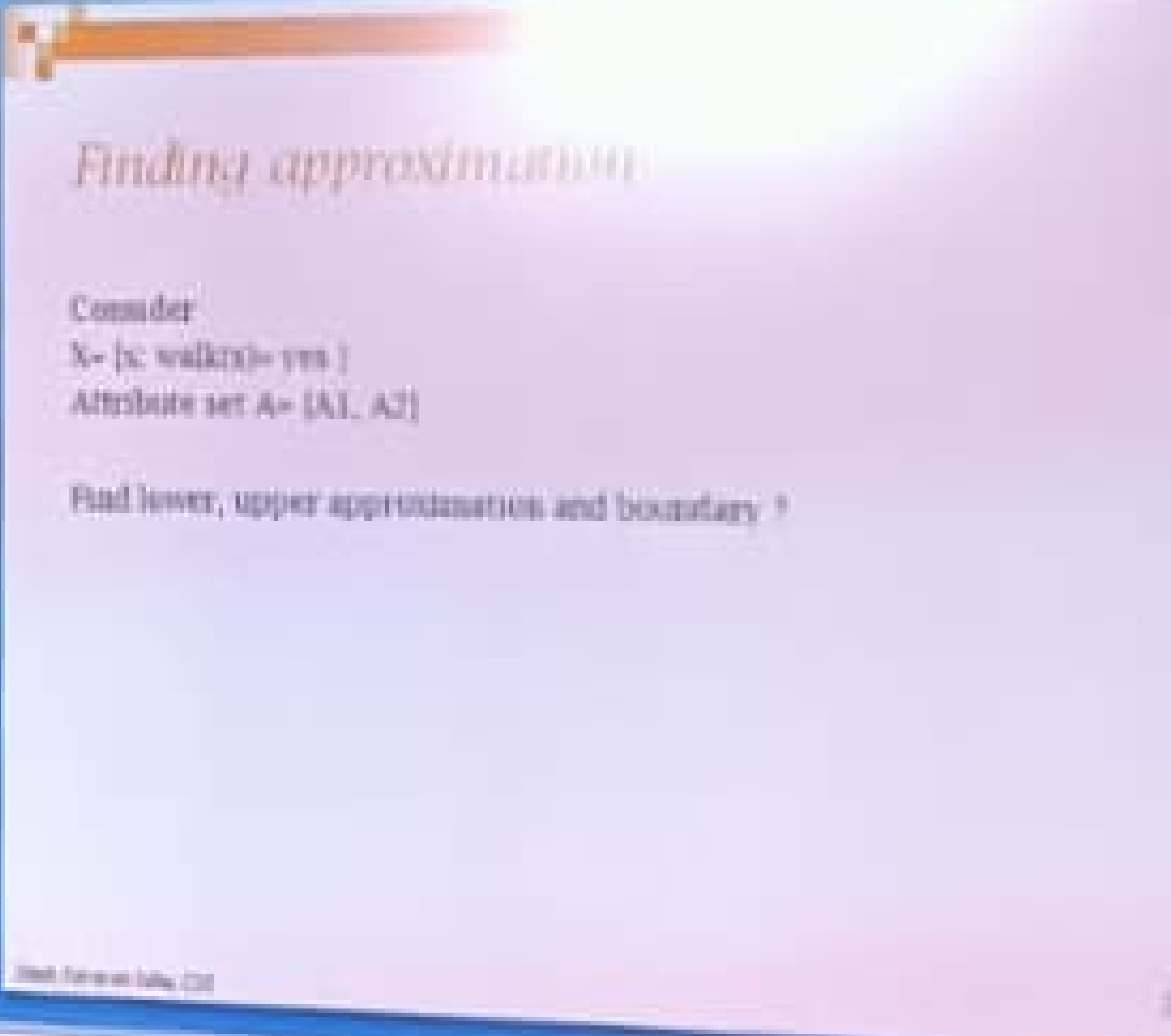
- Consider Information system IS is a pair of (U, A).
- U is a non empty finite set of objects.
- A is non empty finite set of attribute such that a: U—> V_a

U	A1	A2	Walk
x1	16-30	50	yes
x2	16-30	0	no
x3	31-45	1-25	no
x4	31-45	1-25	yes
x5	46-60	26-49	no
x6	16-30	26-49	yes
x7	46-60	26-49	no



Observation

- An equivalence relation induces a partitioning of the universe
- The partitions can be used to build new subsets of the universe
- Subsets that are most often of interest have the same value of the decision attribute
- It may happen, however, that a concept such as "Walk" cannot be defined in a crisp manner

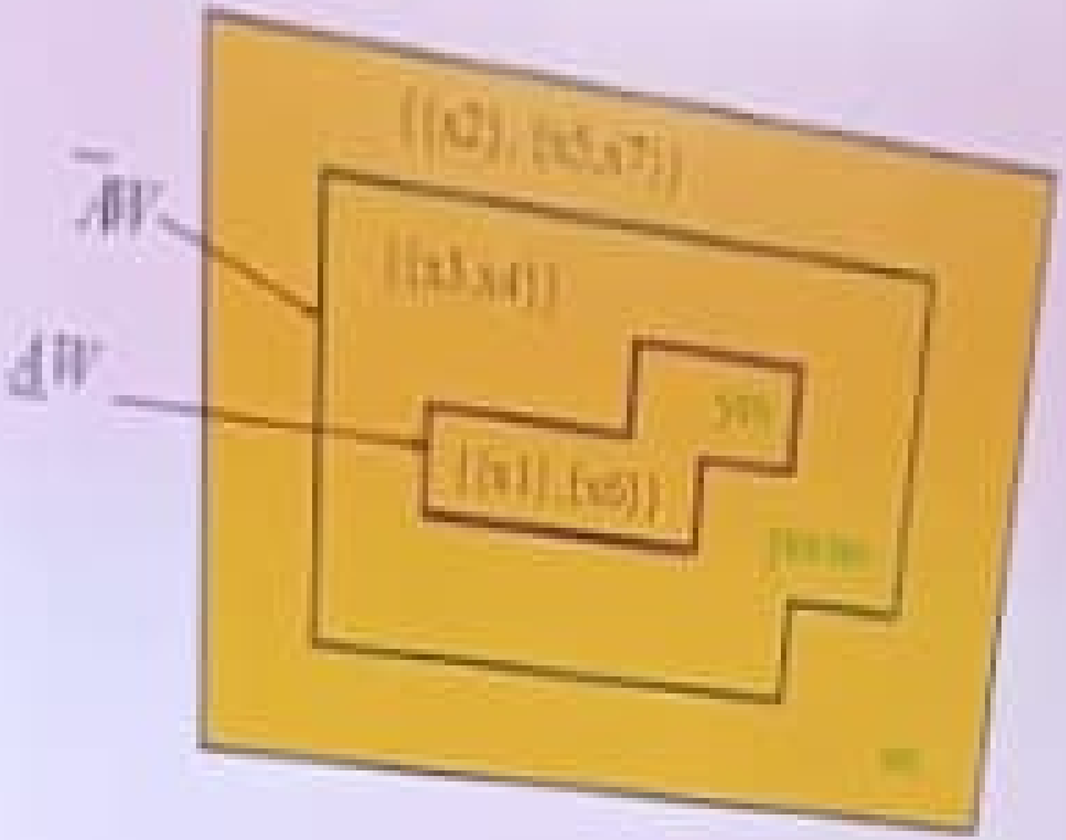


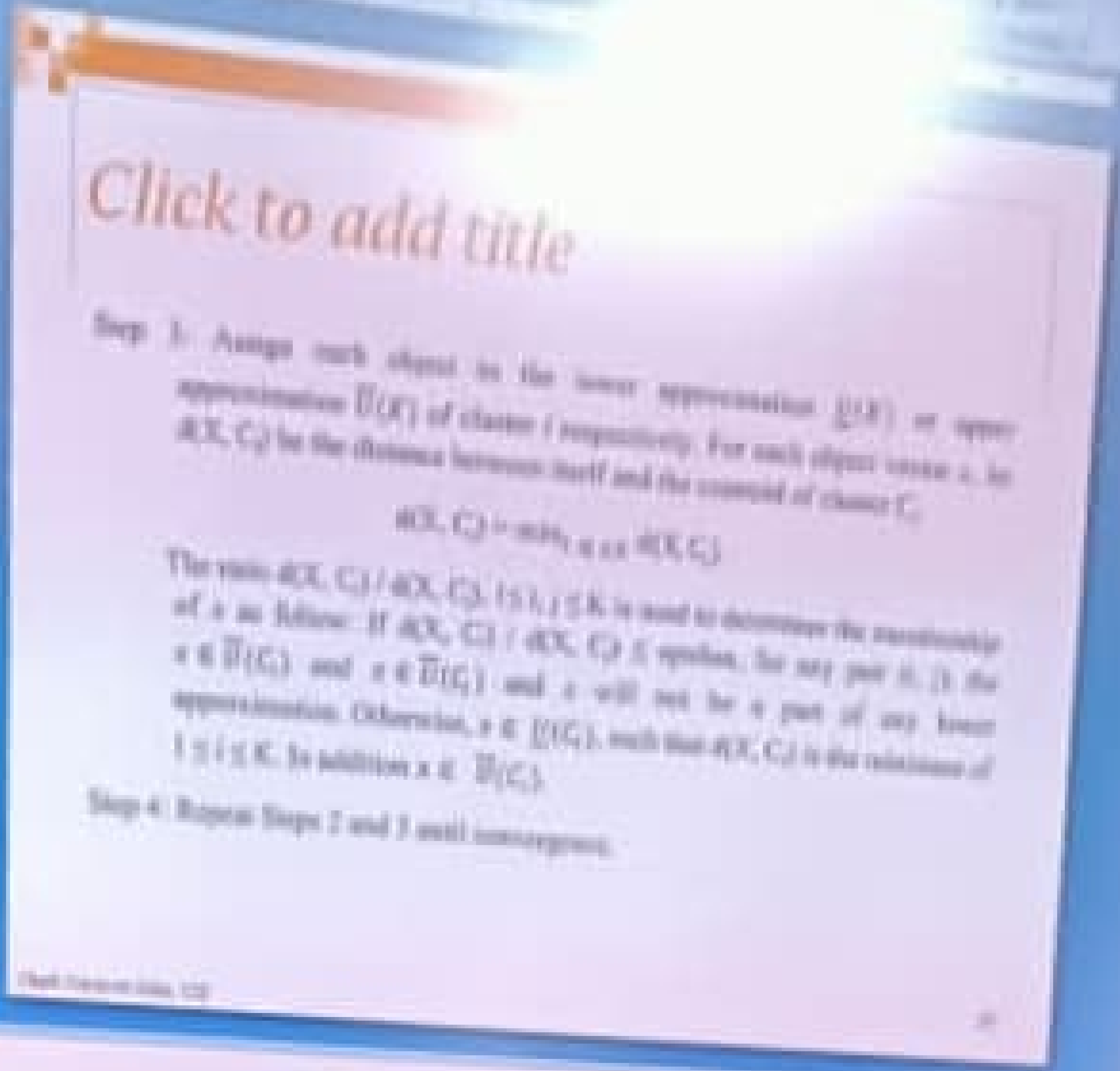
Finding approximation

Consider
X= {x: walk(x)= yes }
Attribute set A= {A1, A2}

Find lower, upper approximation and boundary ?

Contd.





Click to add title

Step 3: Assign each object in the lower approximation $\underline{A}(X)$ or upper approximation $\overline{A}(X)$ of cluster i respectively. For each object vector x, let $d(X, C_i)$ be the distance between itself and the centroid of cluster C_i

$$d(X, C_j) = \min_{1 \le i \le k} d(X, C_i)$$

The ratio $d(X, C_i) / d(X, C_j)$, $1 \le i, j \le K$ is used to determine the membership of x as follow: If $d(X, C_i) / d(X, C_j) \le$ epsilon, for any pair (i, j) the $x \in \overline{A}(C_i)$ and $x \in \overline{A}(C_j)$ and x will not be a part of any lower approximation. Otherwise, $x \in \underline{A}(C_i)$, such that $d(X, C_i)$ is the minimum of $1 \le i \le K$. In addition $x \in \overline{A}(C_i)$.

Step 4: Repeat Steps 2 and 3 until convergence.

Rough set: Indiscernibility

- Indiscernibility Relation is a central concept in Rough Set Theory, and is considered as a relation between two objects or more, where all the values are identical in relation to a subset of considered attributes.

- Indiscernibility relation is an equivalence relation, where all identical objects of set are considered as elementary (Pawlak, 1998).

- It can be expressed as-

$$\mathrm{IND}(P) = \left\{ (x, y) \in \mathbb{U}^2 \mid \forall a \in P, a(x) = a(y) \right\}$$

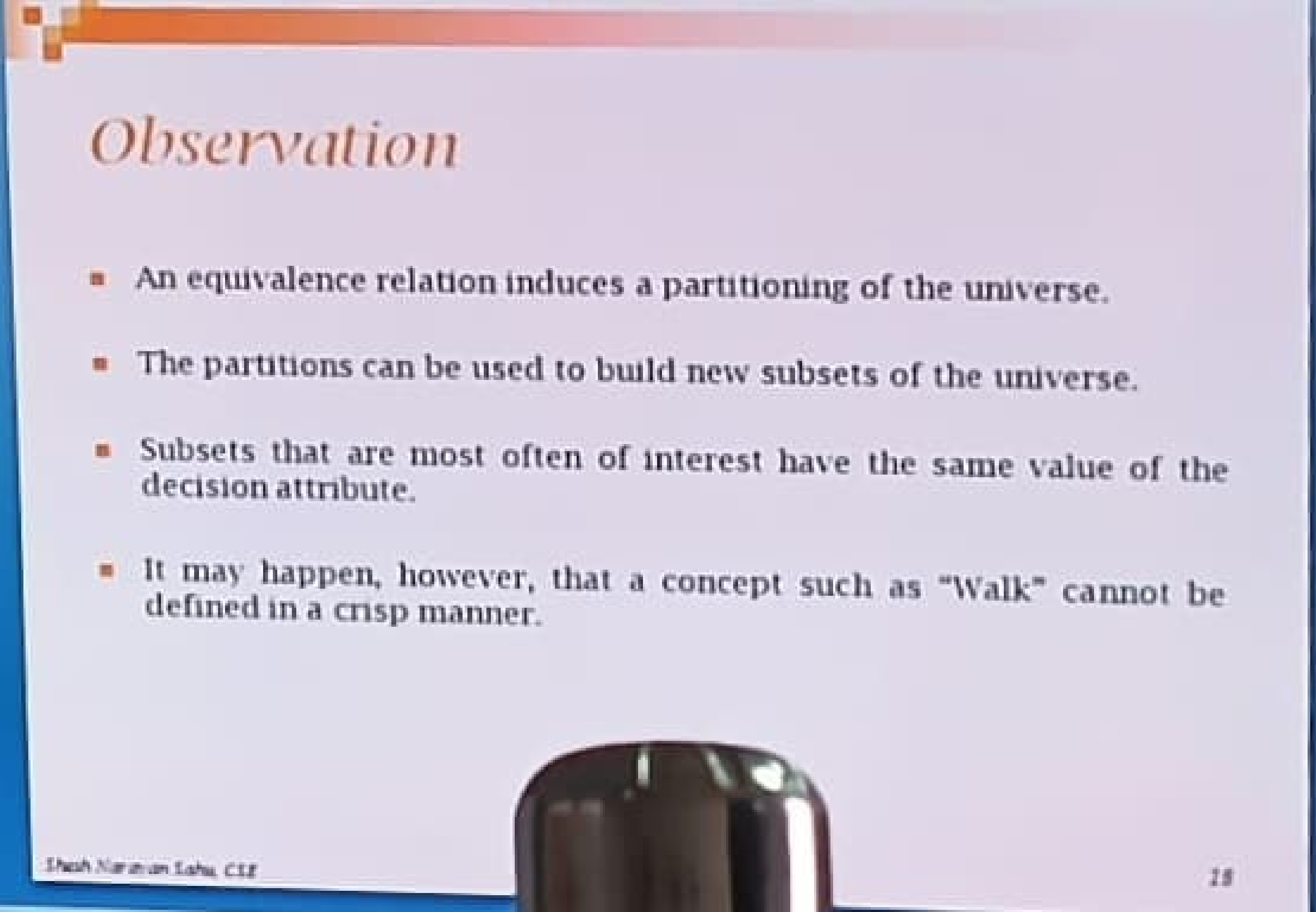
Rough Set

- Rough set theory is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It is proposed for indiscernibility in classification or clustering according to some similarity.

- Rough set theory is based on the establishment of equivalence classes within the given training data.

- All the data tuples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data.

- Given real-world data, it is common that some classes cannot be distinguished in terms of the available attributes. Rough sets can be used to approximately or "roughly" define such classes.



Observation

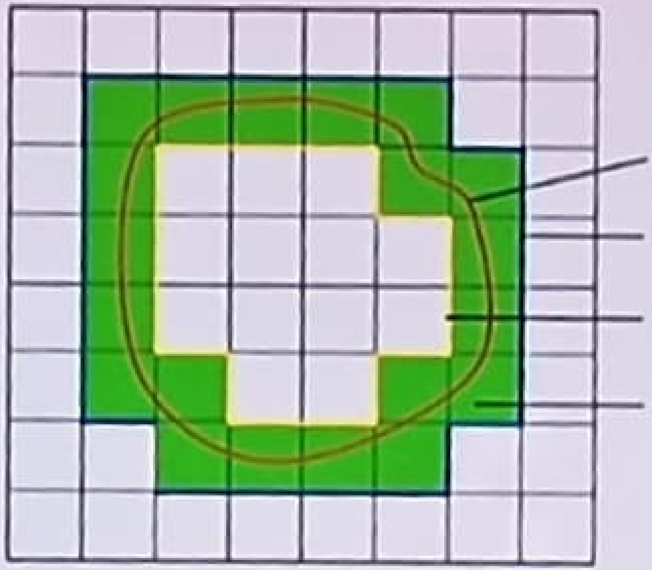
- An equivalence relation induces a partitioning of the universe.
- The partitions can be used to build new subsets of the universe.
- Subsets that are most often of interest have the same value of the decision attribute.
- It may happen, however, that a concept such as "Walk" cannot be defined in a crisp manner.

Set Approximation

- The indiscernibility relation is intended to express the fact that due to the lack of knowledge it is unable to discern some objects employing the available information Approximations is also an important concept in Rough Sets.

- The lower and the upper approximations of a set are interior and closure operations in a topology generated by the indiscernibility relation.

Contd..



Set X

Possible upper approximation $\overline{R}(X)$

Possible lower approximation $\underline{R}(X)$

Possible boundary region

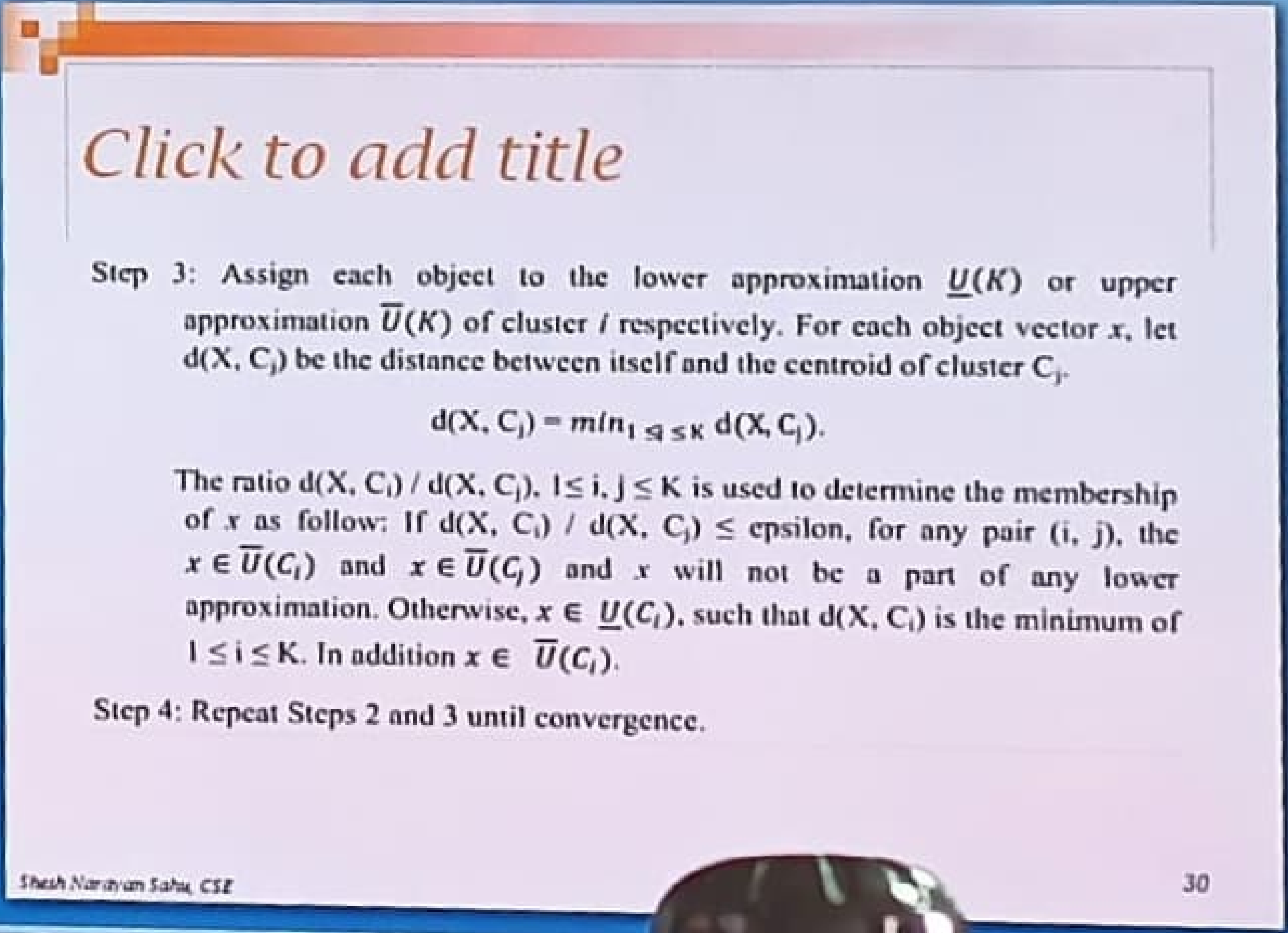
$BN_B(X) = \overline{R}(X) - \underline{R}(X)$

Example

Table 1 shows example information system with real-valued conditional attributes. It consists of six objects/genes, and two features/samples. k = 2, which is the number of clusters. Weight of the lower approximation W_{lower}=0.7, Weight of the upper approximation W_{upper} = 0.3 and Relative threshold = 2.

Table 1 Example dataset for Rough K-Means

U	X	Y
1	0	3
2	1	3
3	3	1
4	3	0.5
5	5	0
6	6	0



Click to add title

Step 3: Assign each object to the lower approximation $\underline{U}(K)$ or upper approximation $\overline{U}(K)$ of cluster I respectively. For each object vector x, let $d(X, C_j)$ be the distance between itself and the centroid of cluster C_j.

$$d(X, C_j) = min_{1 \leq i \leq K} d(X, C_j).$$

The ratio $d(X, C_i) / d(X, C_j)$, $1 \leq i, j \leq K$ is used to determine the membership of x as follow: If $d(X, C_i) / d(X, C_j) \leq$ epsilon, for any pair (i, j), the $x \in \overline{U}(C_i)$ and $x \in \overline{U}(C_j)$ and x will not be a part of any lower approximation. Otherwise, $x \in \underline{U}(C_i)$, such that $d(X, C_i)$ is the minimum of $1 \leq i \leq K$. In addition $x \in \overline{U}(C_i)$.

Step 4: Repeat Steps 2 and 3 until convergence.

Contd..

Step1: Randomly assign each data object one lower approximation$\underline{U}(K)$. By definition (property 2) the data object also belongs to upper approximation $\overline{U}K$ of the same Cluster.

Step 2: Compute Cluster Centroids C_j

If $\underline{U}(K) \neq \emptyset$ and $\overline{U}(K) - \underline{U}(K) = \emptyset$

$$C_j = \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|}$$

Else If $\underline{U}(K) = \emptyset$ and $\overline{U}(K) - \underline{U}(K) \neq \emptyset$

$$C_j = \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

Else

$$C_j = W_{lower} \times \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|} + W_{upper} \times \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$