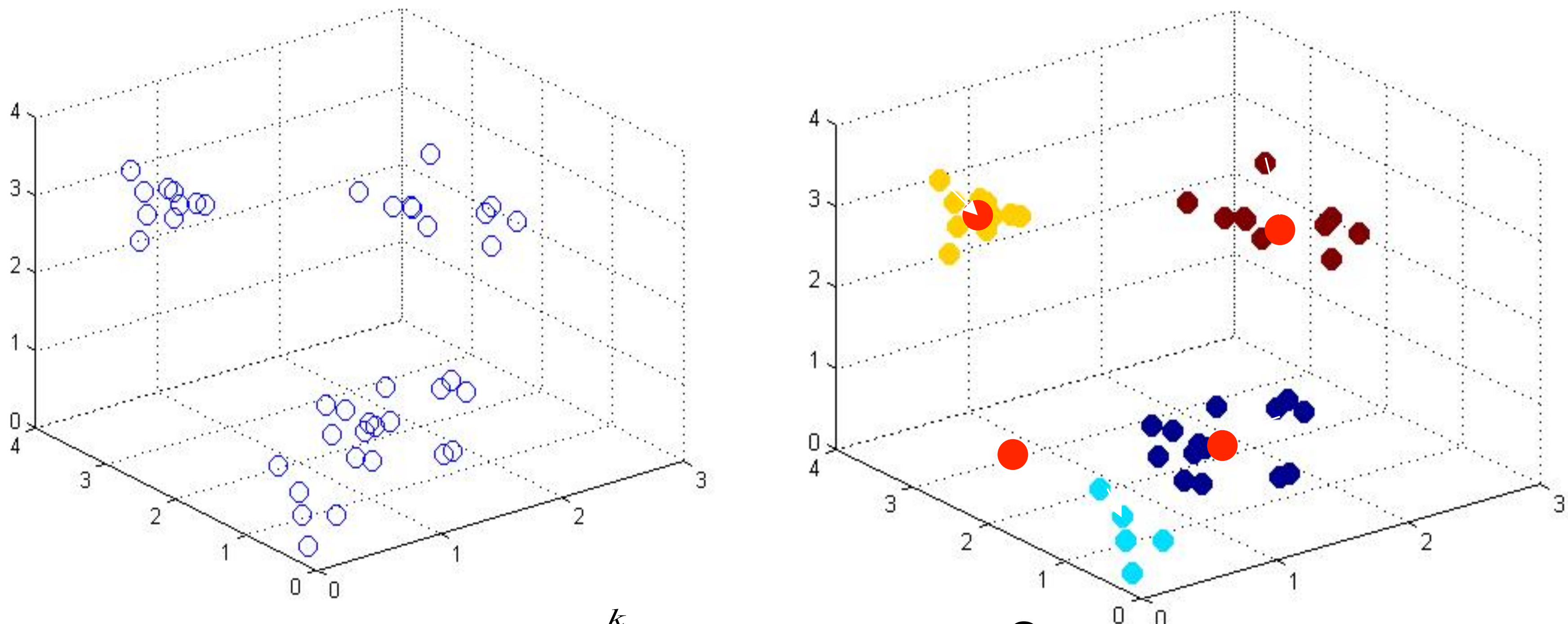
Bi-Clustering

CS 145
Fall 2015

The UNIVERSITY *of* CALIFORNIA *at* LOS ANGELES

Data Mining: Clustering



$$\sum_{t=1}^{k} \sum_{i \in c_t} dist(x_i, c_t)^2$$

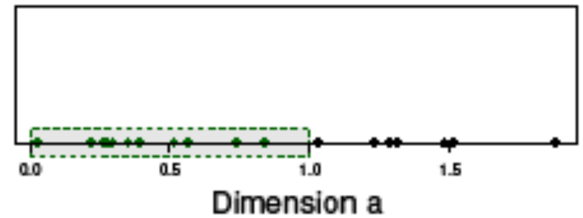
K-means clustering minimizes

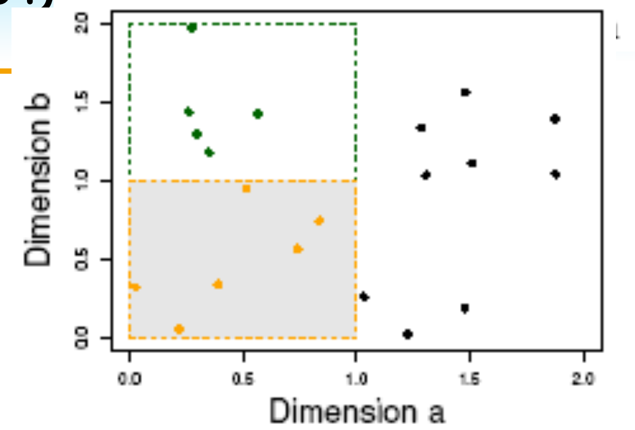
$$dist(x_i, c_t) = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{tj})^2}$$

The Curse of Dimensionality

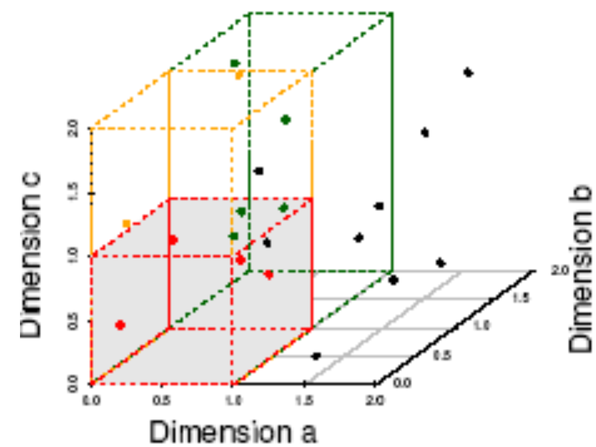
(graphs adapted from Parsons et al. KDD Explorations 2004)

- Data in only one dimension is relatively packed
- Adding a dimension "stretch" the points across that dimension, making them further apart
- Adding more dimensions will make the points further apart—high dimensional data is extremely sparse
- Distance measure becomes meaningless —due to equi-distance

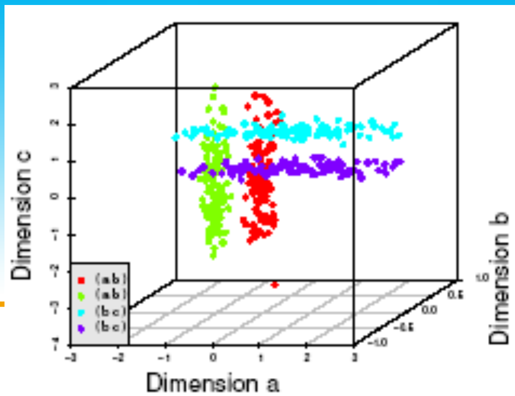




(b) 6 Objects in One Unit Bin



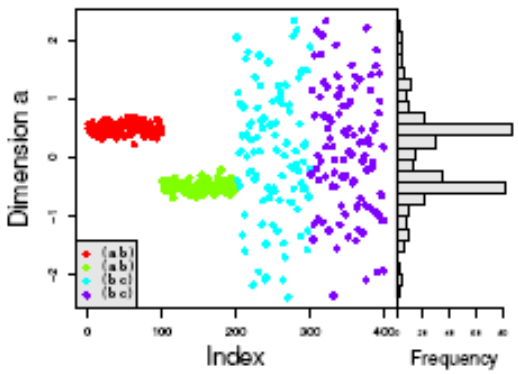
(c) 4 Objects in One Unit Bin



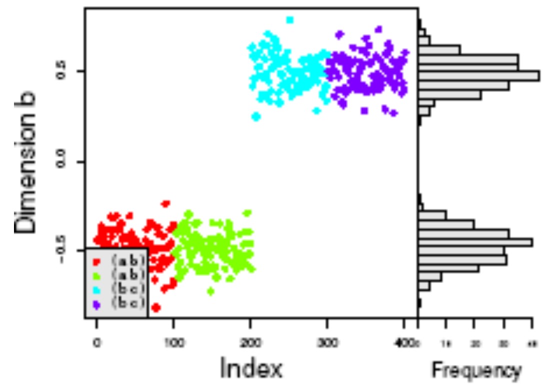
Why Subspace Clustering?

(adapted from Parsons et al. SIGKDD Explorations 2004)

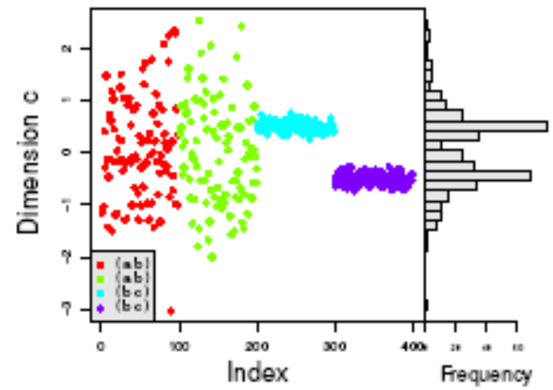
- Clusters may exist only in some subspaces
- Subspace-clustering: find clusters in all the subspaces



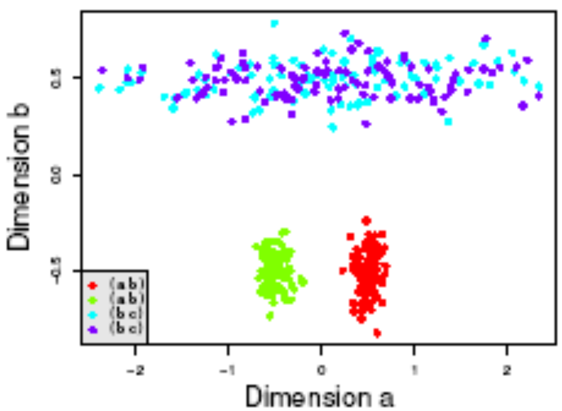
(a) Dimension a



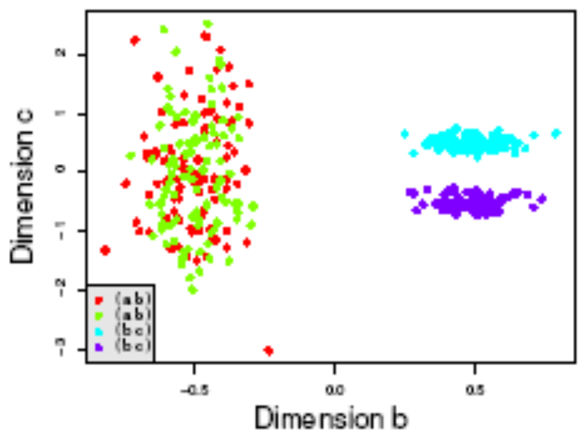
(b) Dimension b



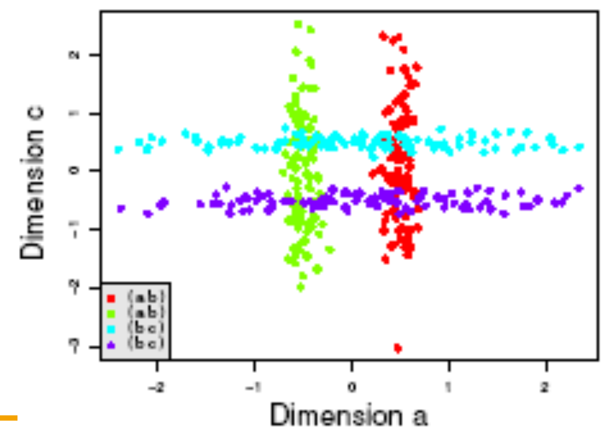
(c) Dimension c



(a) Dims a & b



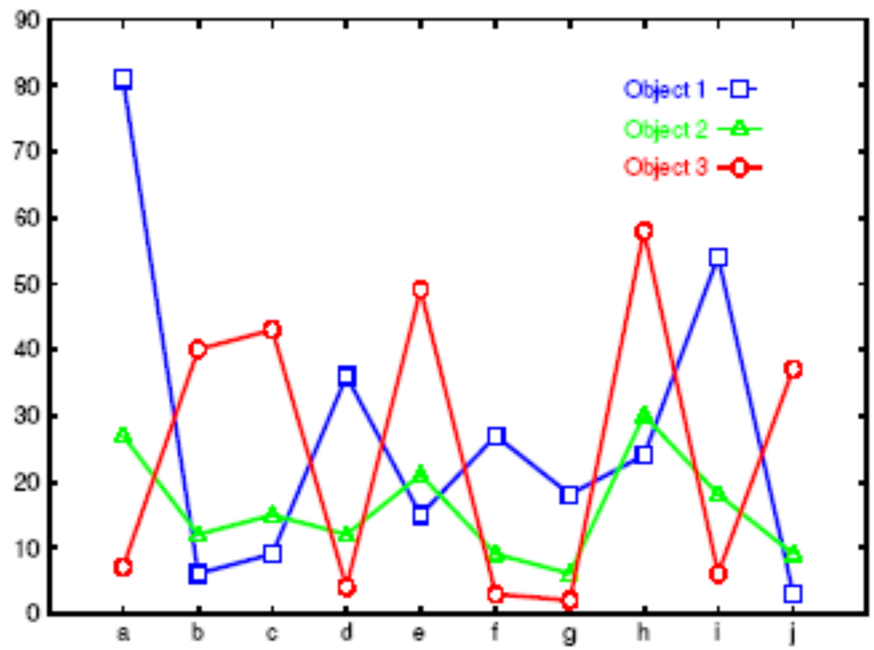
(b) Dims b & c



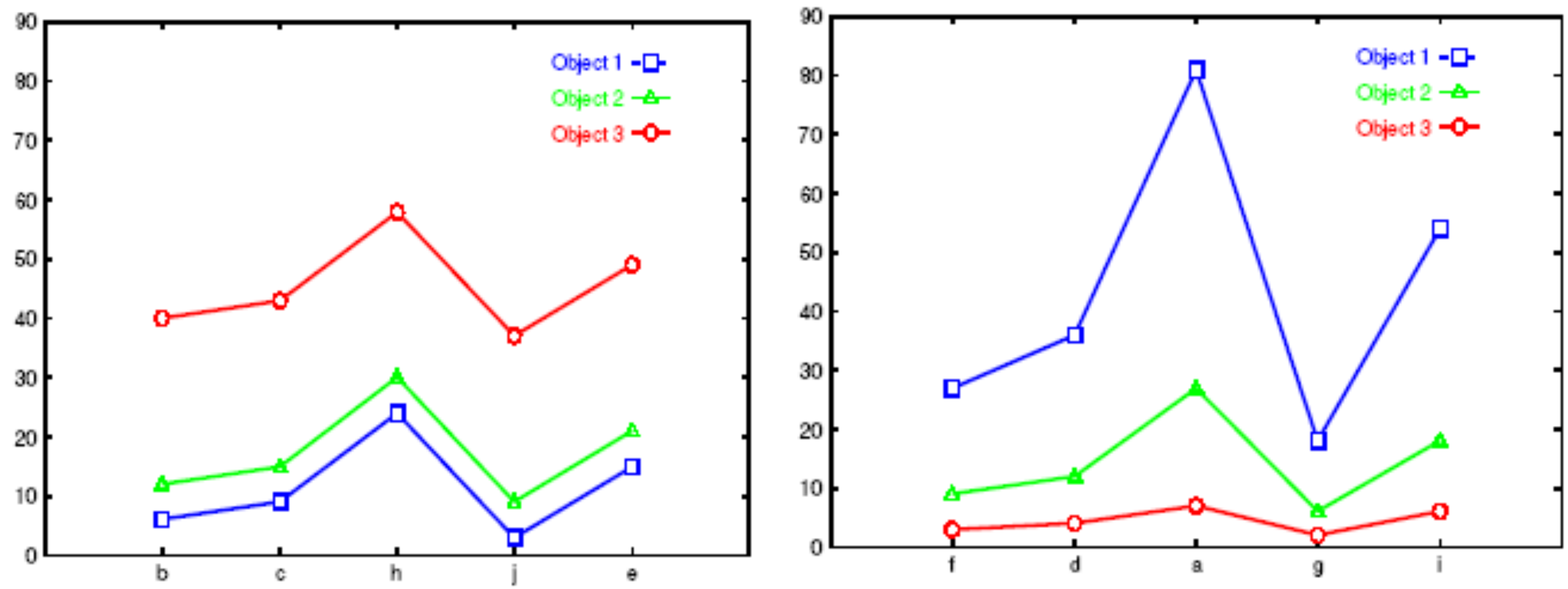
(c) Dims a & c

Clustering by Pattern Similarity (*p*-Clustering)

- **The micro-array "raw" data shows 3 genes and their values in a multi-dimensional space**
  - **Parallel Coordinates Plots**
  - **Difficult to find their patterns**
- **"non-traditional" clustering**



Clusters Are Clear After Projection

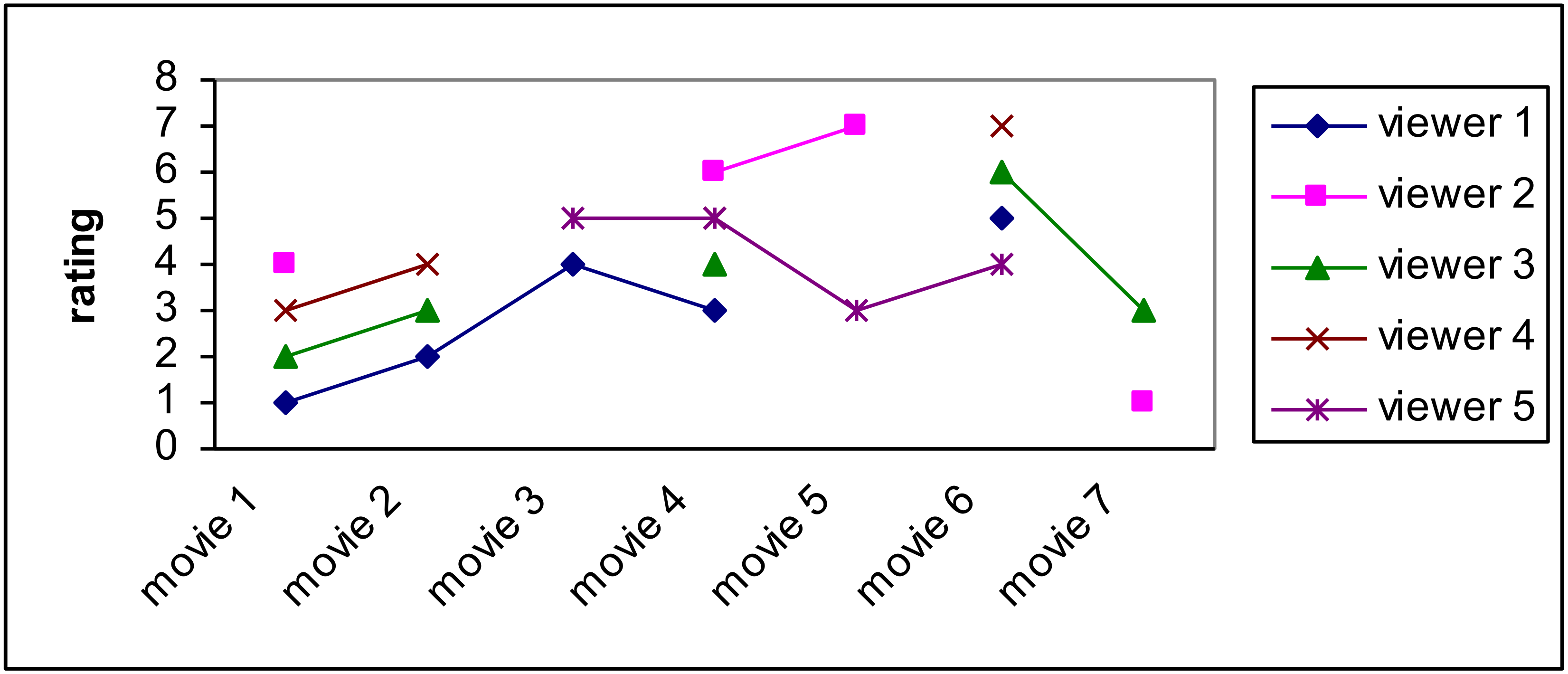


Motivation

- **E-Commerce: collaborative filtering**

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7
Viewer 1	1	2	4	3		5	
Viewer 2	4			6	7		1
Viewer 3	2	3		4		6	3
Viewer 4	3	4		5		7	
Viewer 5			5	5	3	4	

Motivation

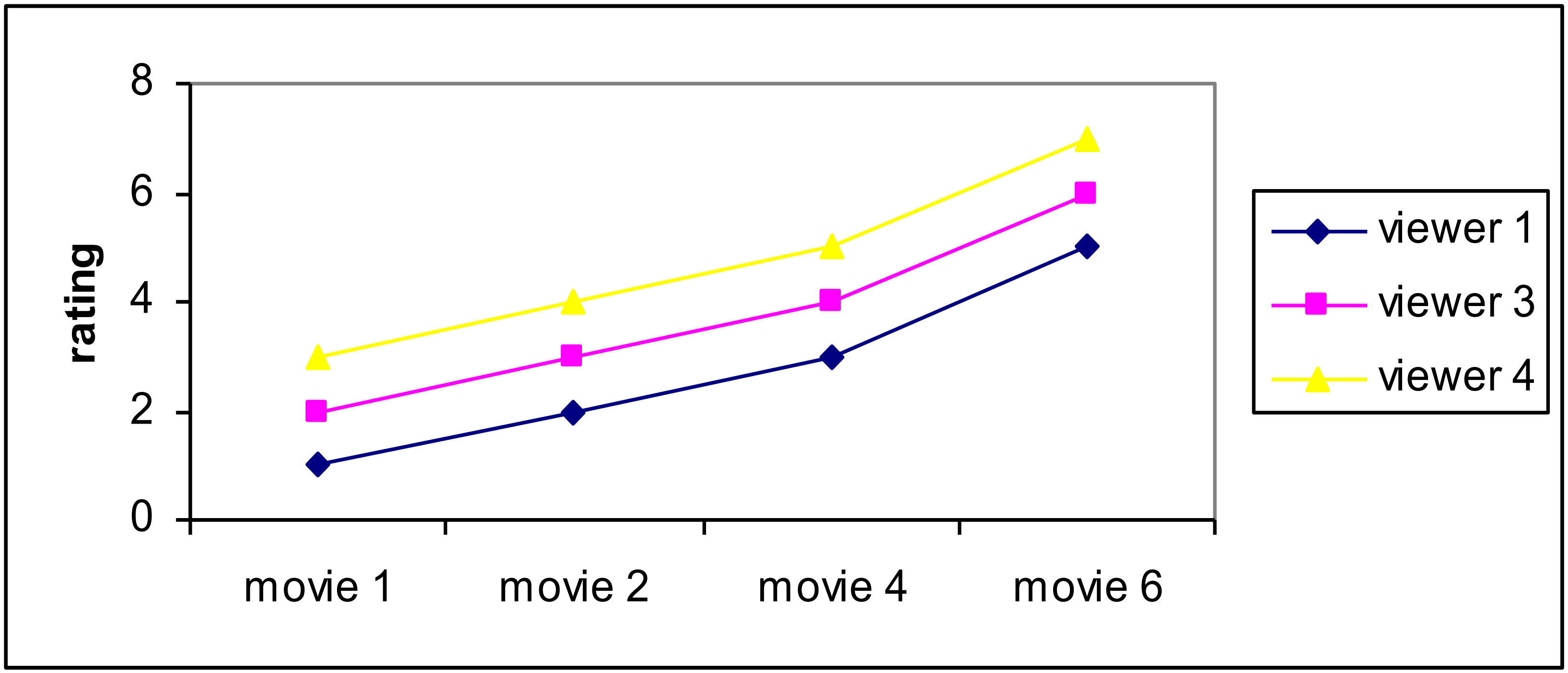


rating
0
1
2
3
4
5
6
7
8
movie 1
movie 2
movie 3
movie 4
movie 5
movie 6
movie 7
viewer 1
viewer 2
viewer 3
viewer 4
viewer 5

Motivation

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7
Viewer 1	1	2	4	3		5	
Viewer 2	4			6	7		1
Viewer 3	2	3		4		6	3
Viewer 4	3	4		5		7	
Viewer 5			5	5	3	4	

Motivation

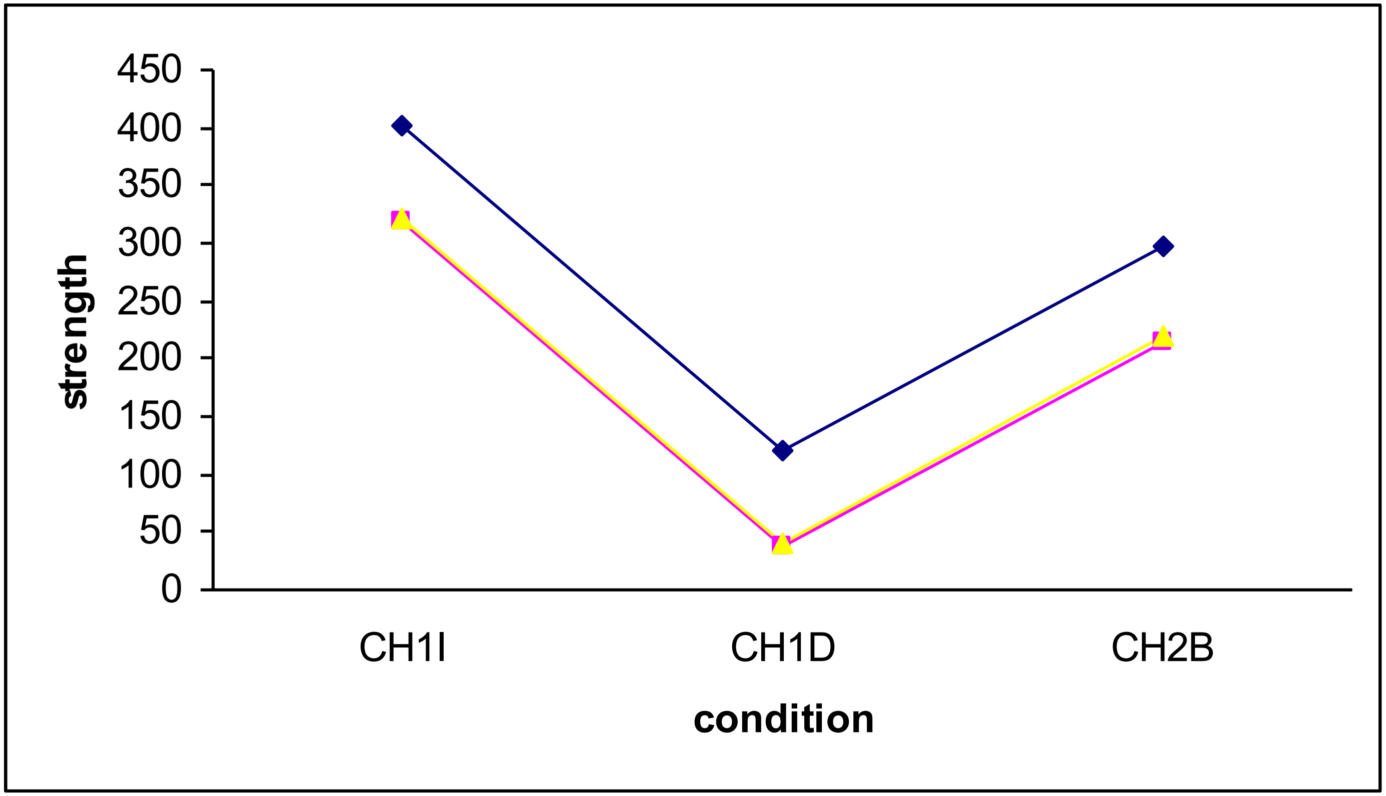

8
6
4
2
0
rating
movie 1
movie 2
movie 4
movie 6
viewer 1
viewer 3
viewer 4

Motivation

- **DNA microarray analysis**

	CH1I	CH1B	CH1D	CH2I	CH2B
CTFC3	4392	284	4108	280	228
VPS8	401	281	120	275	298
EFB1	318	280	37	277	215
SSA1	401	292	109	580	238
FUN14	2857	285	2576	271	226
SP07	228	290	48	285	224
MDM10	538	272	266	277	236
CYS3	322	288	41	278	219
DEP1	312	272	40	273	232
NTG1	329	296	33	274	228

Motivation


450
400
350
300
250
200
150
100
50
0
strength
CH1I
CH1D
CH2B
condition

Motivation

- Strong coherence exhibits by the selected objects on the selected attributes.
  - They are not necessarily close to each other but rather bear a constant shift.
  - Object/attribute bias
- bi-cluster

Challenges

- **The set of objects and the set of attributes are usually unknown.**
- **Different objects/attributes may possess different biases and such biases**
  - **may be local to the set of selected objects/ attributes**
  - **are usually unknown in advance**
- **May have many unspecified entries**

Previous Work

- **Subspace clustering**
  - **Identifying a set of objects and a set of attributes such that the set of objects are physically close to each other on the subspace formed by the set of attributes.**
- **Collaborative filtering: Pearson R**
  - **Only considers global offset of each object/attribute.**

$$\frac{\sum (o_1 - \bar{o}_1)(o_2 - \bar{o}_2)}{\sqrt{\sum (o_1 - \bar{o}_1)^2 \times \sum (o_2 - \bar{o}_2)^2}}$$

bi-cluster

- **Consists of a (sub)set of objects and a (sub)set of attributes**
  - **Corresponds to a** <u>submatrix</u>
  - **Occupancy threshold** α
    - **Each object/attribute has to be filled by a certain percentage.**
  - **Volume: number of specified entries in the submatrix**
  - **Base: average value of each object/attribute (in the bi-cluster)**

bi-cluster

	CH1I	CH1B	CH1D	CH2I	CH2B	Obj base
CTFC3						
VPS8	401		120		298	273
EFB1	318		37		215	190
SSA1						
FUN14						
SP07						
MDM10						
CYS3	322		41		219	194
DEP1						
NTG1						
Attr base	347		66		244	219

bi-cluster

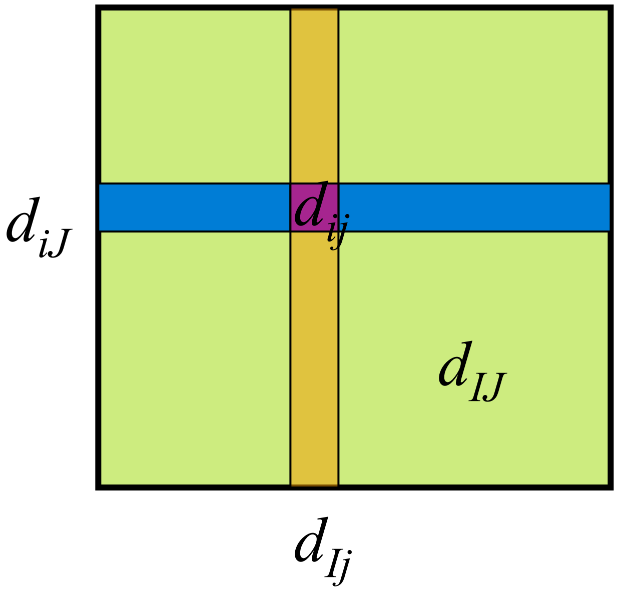
- **Perfect δ-cluster**

$$d_{ij} - d_{iJ} = d_{Ij} - d_{IJ}$$

$$d_{ij} - d_{Ij} = d_{iJ} - d_{IJ}$$

$$\boxed{d_{ij} = d_{iJ} + d_{Ij} - d_{IJ}}$$

- **Imperfect δ-cluster**
  - **Residue:**



$$r_{ij} = \begin{cases} d_{ij} - d_{iJ} - d_{Ij} + d_{IJ}, & d_{ij} \text{ is specified} \\ 0, & d_{ij} \text{ is unspecified} \end{cases}$$

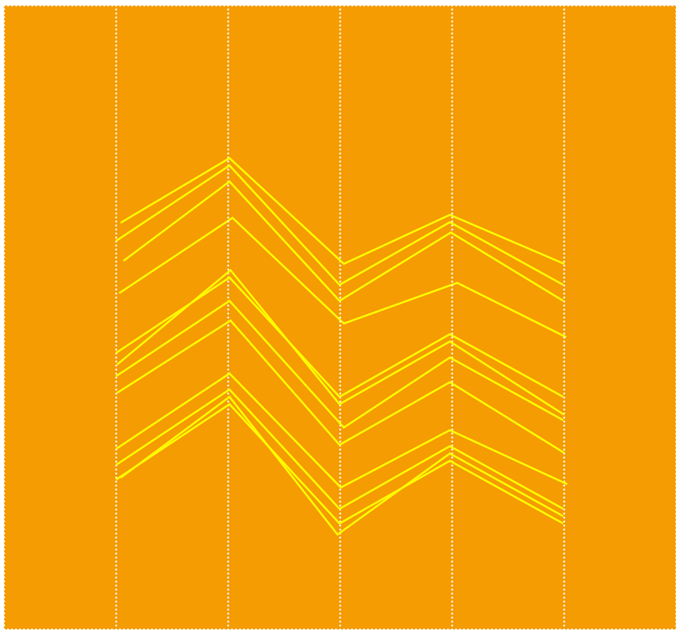
bi-cluster

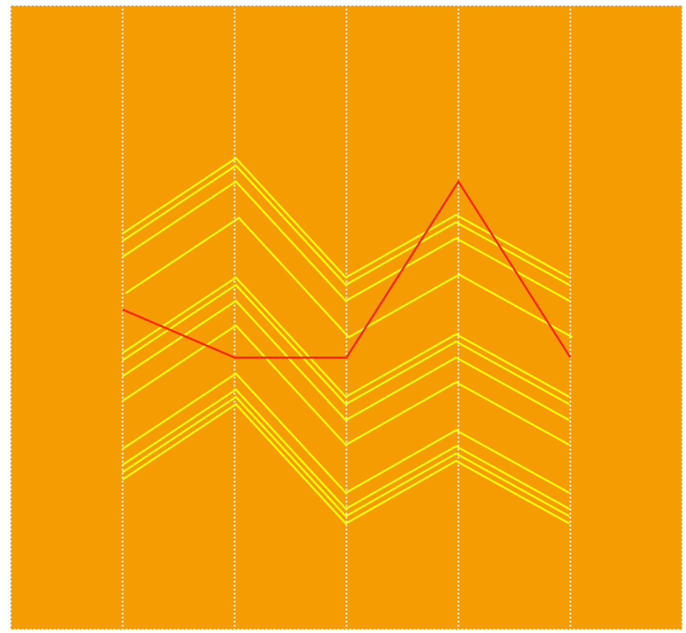
- The smaller the average residue, the stronger the coherence.
- **Objective**: identify δ-clusters with residue smaller than a given threshold

Cheng-Church Algorithm

- **Find one bi-cluster.**
- **Replace the data in the first bi-cluster with random data**
- **Find the second bi-cluster, and go on.**
- **The quality of the bi-cluster degrades (smaller volume, higher residue) due to the insertion of random data.**

Coherent Cluster



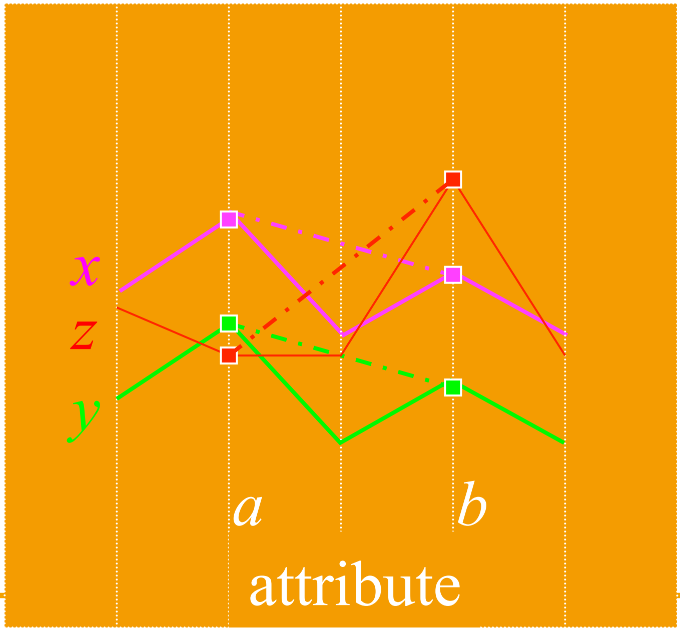


Want to accommodate noise but not outliers

Coherent Cluster

- **Coherent cluster**
  - **Subspace clustering**
- **pair-wise disparity**
  - **For a 2×2 (sub)matrix consisting of objects {*x*, *y*} and attributes {*a*, *b*}**

$$D\left(\begin{bmatrix} d_{xa} & d_{xb} \\ d_{ya} & d_{yb} \end{bmatrix}\right) = \left|\underbrace{(d_{xa} - d_{ya})}_{\text{mutual bias}} - \underbrace{(d_{xb} - d_{yb})}_{\text{mutual bias}}\right|$$



Coherent Cluster

- **A 2×2 (sub)matrix is a** δ-coherent cluster **if its D value is less than or equal to** δ**.**
- **An** $m \times n$ **matrix** X **is a** δ-coherent cluster **if every 2×2 submatrix of** X **is** δ**-coherent cluster.**
  - **A** δ**-coherent cluster is a** maximum δ**-coherent cluster if it is not a submatrix of any other** δ**-coherent cluster.**
- **Objective: given a data matrix and a threshold** δ**, find all maximum** δ**-coherent clusters.**

Coherent Cluster

- **Challenges:**
  - **Finding subspace clustering based on distance itself is already a difficult task due to the curse of dimensionality.**
    - **The (sub)set of objects and the (sub)set of attributes that form a cluster are unknown in advance and may not be adjacent to each other in the data matrix.**
  - **The actual values of the objects in a coherent cluster may be far apart from each other.**
    - **Each object or attribute in a coherent cluster may bear some relative bias (that are unknown in advance) and such bias may be local to the coherent cluster.**

Coherent Cluster

Compute the maximum coherent attribute sets for each pair of objects

↓

Two-way Pruning

↓

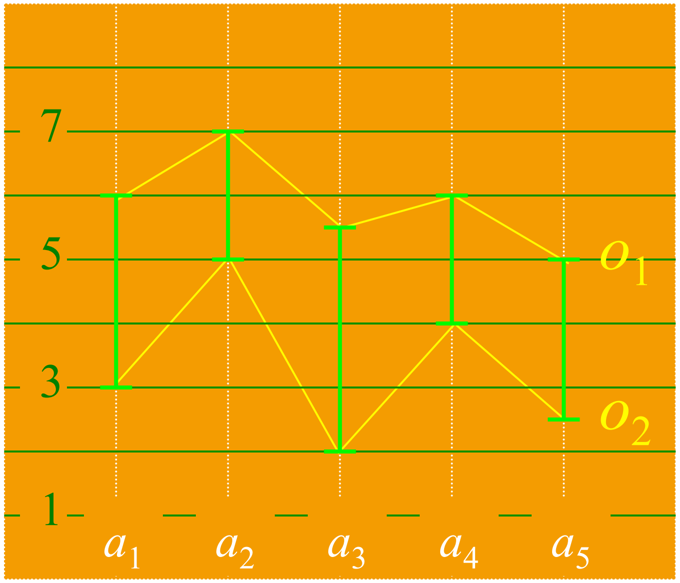
Construct the lexicographical tree

↓

Post-order traverse the tree to find maximum coherent clusters

Coherent Cluster

- **Observation**: **Given a pair of objects {o_1, o_2} and a (sub)set of attributes {a_1, a_2, …, a_k}, the 2×k submatrix is a δ-coherent cluster iff, for every attribute a_i, the mutual bias ($d_{o1ai} - d_{o2ai}$) does not differ from each other by more than δ.**

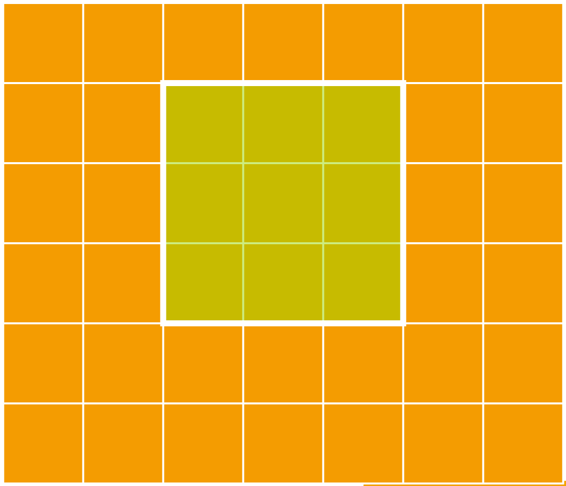


coherent attribute set (CAS) of (o_1,o_2).

3 2 3.5 2 2.5 $\in$ [2, 3.5]

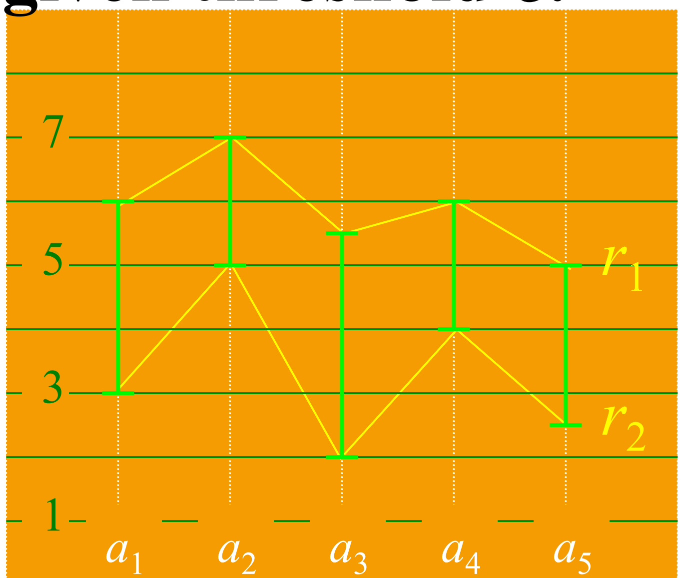
Coherent Cluster

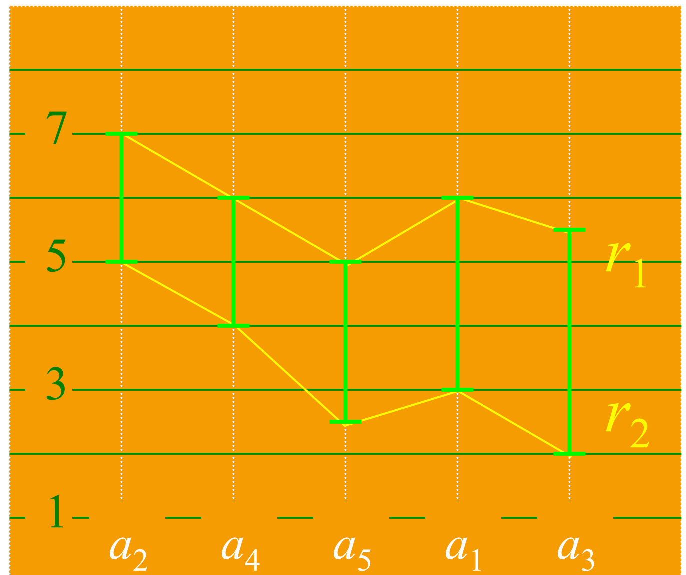
- **Observation**: given a subset of objects $\{o_1, o_2, \ldots, o_l\}$ and a subset of attributes $\{a_1, a_2, \ldots, a_k\}$, the $l \times k$ submatrix is a δ-coherent cluster iff $\{a_1, a_2, \ldots, a_k\}$ is a coherent attribute set for every pair of objects (o_i, o_j) where $1 \leq i, j \leq l$.



Coherent Cluster

- **Strategy**: **find the *maximum coherent attribute sets* for each pair of objects with respect to the given threshold δ.**





$\delta = 1$ 3 2 3.5 2 2.5 2 2 2.5 3 3.5

The maximum coherent attribute sets define the search space for maximum coherent clusters.

Two Way Pruning

	a0	a1	a2
o0	1	4	2
o1	2	5	5
o2	3	6	5
o3	4	200	7
o4	300	7	6

delta=1 nc =3 nr = 3

(o0,o2) →(a0,a1,a2)
(o1,o2) →(a0,a1,a2)

(a0,a1) →(o0,o1,o2)
(a0,a2) →(o1,o2,o3)
(a1,a2) →(o1,o2,o4)
(a1,a2) →(o0,o2,o4)

~~(o0,o2) →(a0,a1,a2)~~
~~(o1,o2) →(a0,a1,a2)~~

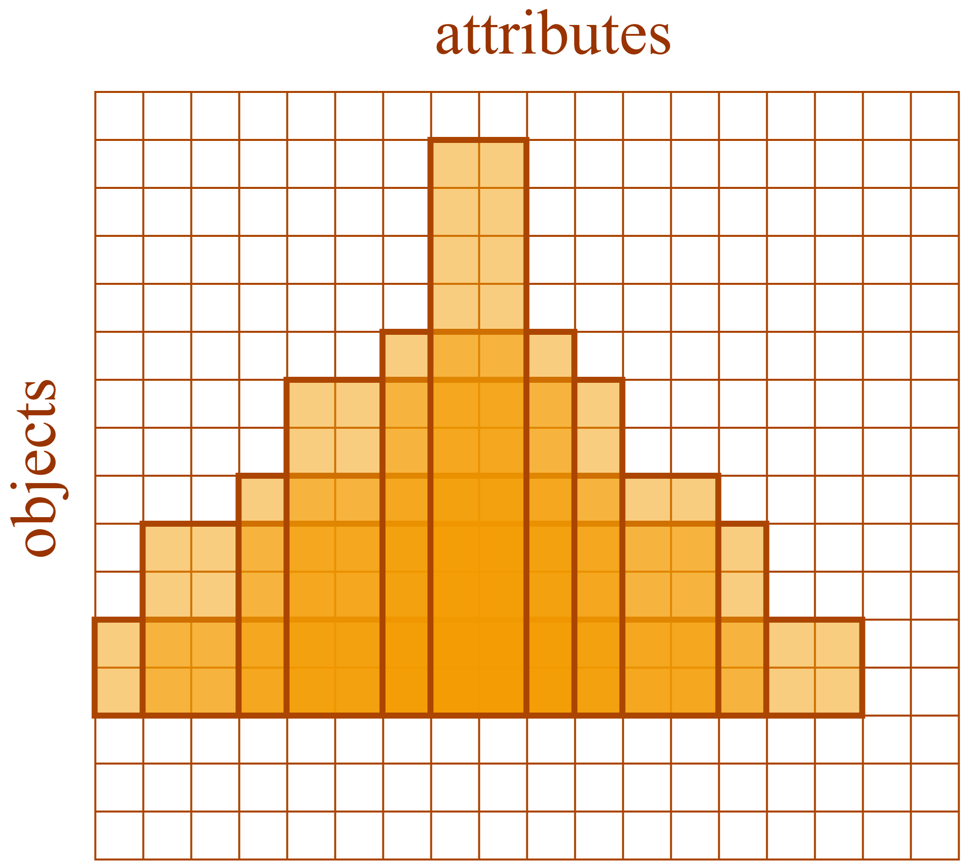
MCAS

~~(a0,a1) →(o0,o1,o2)~~
~~(a0,a2) →(o1,o2,o3)~~
~~(a1,a2) →(o1,o2,o4)~~
~~(a1,a2) →(o0,o2,o4)~~

MCOS

Coherent Cluster

- **Strategy**: grouping object pairs by their CAS and, for each group, find the maximum clique(s).
- **Implementation**: using a lexicographical tree to organize the object pairs and to generate all maximum coherent clusters with a single post-order traversal of the tree.



	a_0	a_1	a_2	a_3
o_0	1	4	2	5
o_1	2	5	5	8
o_2	3	6	5	7
o_3	4	20	7	2
o_4	30	7	6	6

$\{a_0,a_1\}$: (o_0,o_1) (o_1,o_2) (o_0,o_2)

$\{a_0,a_2\}$: $(o_1,o_3),(o_2,o_3)$ (o_1,o_2) (o_0,o_2)

$\{a_1,a_2\}$: $(o_0,o_4),(o_1,o_4),(o_2,o_4)$ (o_1,o_2) (o_0,o_2)

$\{a_2,a_3\}$: $(o_0,o_1),(o_1,o_2)$ (o_0,o_2)

$\{a_0,a_1,a_2\}$: (o_1,o_2) (o_0,o_2)

$\{a_0,a_1,a_2,a_3\}$: (o_0,o_2)

assume $\delta = 1$

(o_0,o_1) : $\{a_0,a_1\}$, $\{a_2,a_3\}$

(o_0,o_2) : $\{a_0,a_1,a_2,a_3\}$

(o_0,o_4) : $\{a_1,a_2\}$

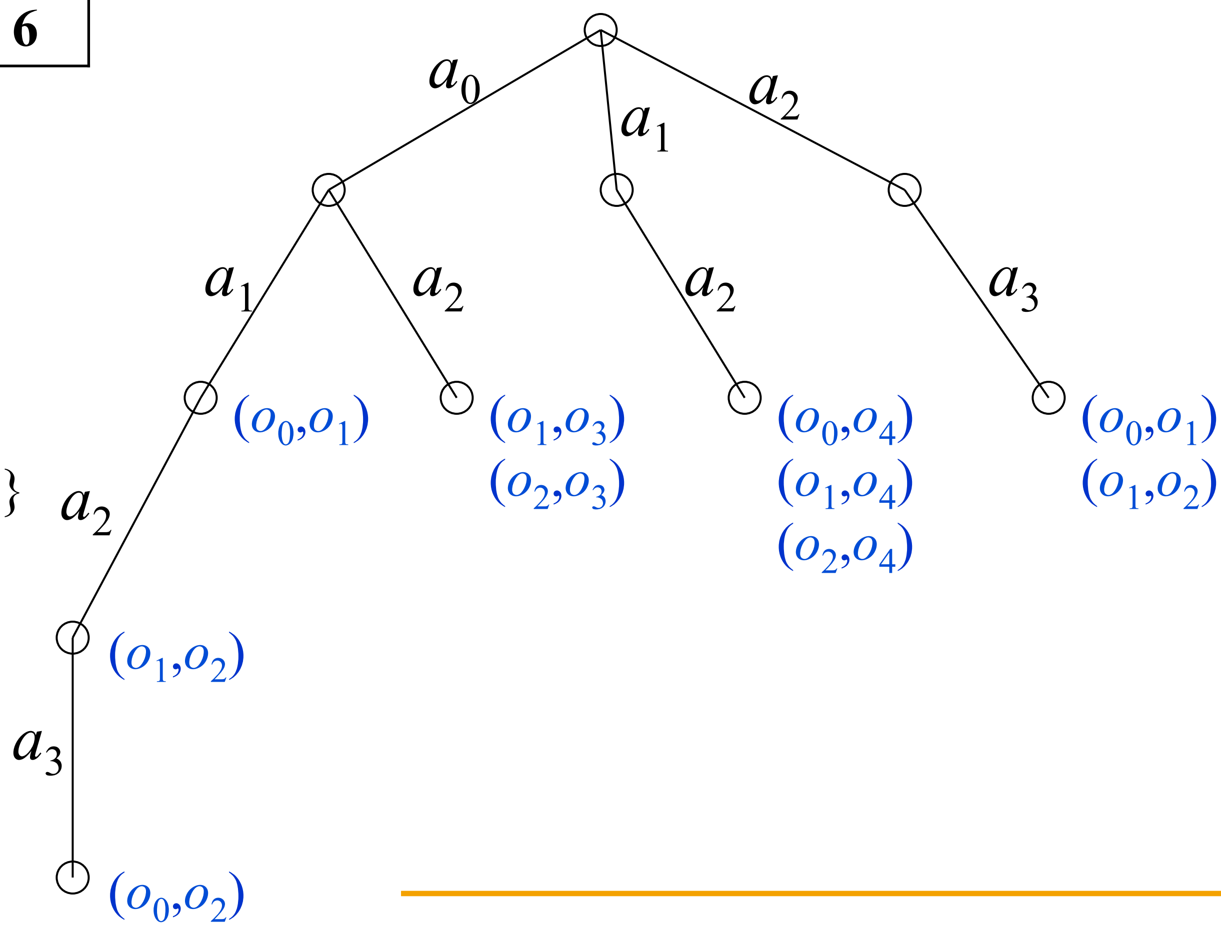
(o_1,o_2) : $\{a_0,a_1,a_2\}$, $\{a_2,a_3\}$

(o_1,o_3) : $\{a_0,a_2\}$

(o_1,o_4) : $\{a_1,a_2\}$

(o_2,o_3) : $\{a_0,a_2\}$

(o_2,o_4) : $\{a_1,a_2\}$



$\{o_0,o_2\} \times \{a_0,a_1,a_2,a_3\}$

$\{o_1,o_2\} \times \{a_0,a_1,a_2\}$

$\{o_0,o_1,o_2\} \times \{a_0,a_1\}$

$\{o_1,o_2,o_3\} \times \{a_0,a_2\}$

$\{o_0,o_2,o_4\} \times \{a_1,a_2\}$

$\{o_1,o_2,o_4\} \times \{a_1,a_2\}$

$\{o_0,o_1,o_2\} \times \{a_2,a_3\}$