Data Mining and Information Retrieval

Introduction to Data Mining

Why Data Mining?

- Thanks to the advances of data processing technologies, a lot of data can be collected and stored in databases efficiently
- New challenges: with a huge amount of data, how to analyze and understand the data?

Example: Data Collection

- Each customer has a club member card
- Transactions: customers’ purchases of commodities
  - {bread, milk, cheese} if they are bought together
  - Each transaction is associated with a customer, a store, time, total price
- Analytic question: what are product combinations that are frequently purchased together by customers?

Example: Data Pre-processing

- Data selection
  - Customer ids, stores are not interesting
  - Only the transactions are selected
- Data integration
  - Integrate data from all stores
  - Partition data by time, e.g., a data set per week
- Data cleaning
  - Normalize data sets for long weekends
  - Estimate effects of promotion campaigns

Example: Mining

- Identify proper data mining methods
  - Only the hot product combinations?
  - Changes of hot product combinations over time?
  - Only combinations having expensive items?
- Select appropriate algorithms
  - Centralized vs. distributed databases?
  - Mining in batch or online?
  - Incremental mining?

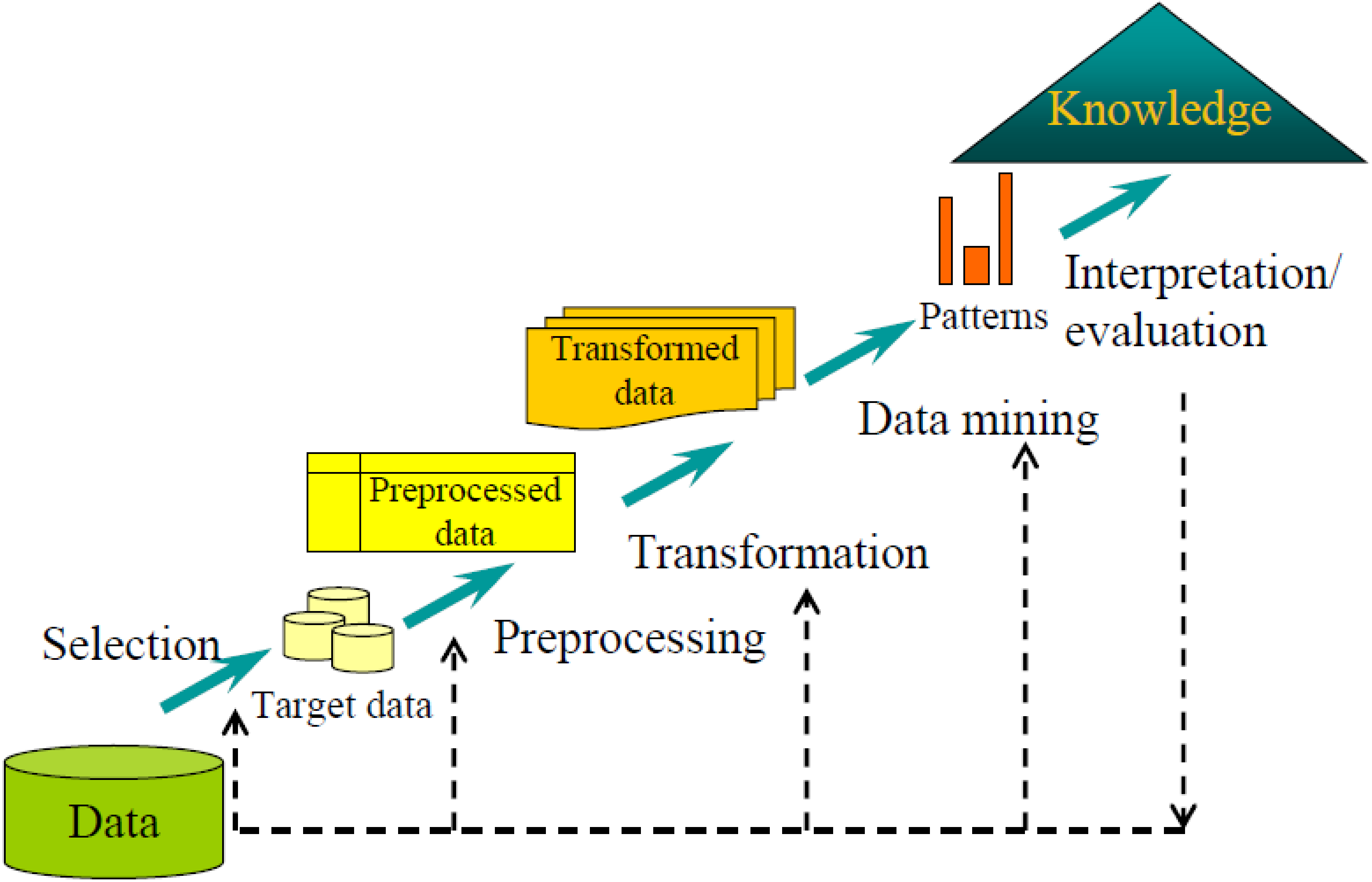
Example: Evaluation

- Find a pattern {digital camera, memory stick, image processing software}
- Explanation
  - Customers buying digital camera often purchase memory stick and image processing software at the same time
- Actions
  - Cluster digital cameras, memory sticks and image processing software together in stores
  - Promote memory stick to attract more digital camera purchases
- An interesting result:
  - “beer” and “baby diaper”

What is Data Mining?

- Mining data – mining knowledge
- Data mining is the *non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad, Piatetsky-Shapiro, Smyth, 96]*

The KDD Process



What Kinds of Data to Be Mined?

- Any data that are useful in practice
- Some examples
  - Relational databases
  - Transactional databases
  - Spatial data
  - Time-series
  - Semi-structured data & WWW
  - Streaming data
  - Bio-medical data
  - Network traffic data
  - Sensor network surveillance data

What Kinds of Patterns?

- Any meaningful complicated patterns that are not directly “query-able” from data
- Some examples
  - Association rules and sequential patterns
  - Classification
  - Clusters and outliers

Identify Interesting Patterns

- There can be a huge number of patterns
- Find all product combinations purchased by more than 1% of customers
  - {bread, milk} is trivial, common sense
  - {diamond, pearl necklace} is informative
- Various users may be interested in different patterns
- Find the patterns strongly wanted by users

Association Rules

- The Market-Basket Model
  - A large set of *items*, e.g., things sold in a supermarket.
  - A large set of *baskets*, each of which is a small set of the items, e.g., the things one customer buys on one day.

Support

- Simplest question: find sets of items that appear “frequently” in the baskets.
- *Support* for itemset *I* = the number of baskets containing all items in *I*.
- Given a support *threshold s*, sets of items that appear in $\geq$ *s* baskets are called *frequent itemsets*.

Example

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.
- B_1 = {m, c, b} B_2 = {m, p, j}
- B_3 = {m, b} B_4 = {c, j}
- B_5 = {m, p, b} B_6 = {m, c, b, j}
- B_7 = {c, b, j} B_8 = {b, c}
- Frequent itemsets: {m}, {c}, {b}, {j}, {m, b}, {c, b}, {j, c}.

Application (1)

- Real market baskets: chain stores keep terabytes of information about what customers buy together.
- Tells how typical customers navigate stores, lets them position tempting items.
- Suggests tie-in “tricks,” e.g., run sale on diapers and raise the price of beer.
- High support needed, or no $$’s .

Application (2)

- “Baskets” = documents; “items” = words in those documents.
- Lets us find words that appear together unusually frequently, i.e., linked concepts.
- “Baskets” = sentences, “items” = documents containing those sentences.
- Items that appear together too often could represent plagiarism.

Association Rules

- If-then rules about the contents of baskets.
- $\{i_1, i_2,\ldots,i_k\} \rightarrow j$ means: “if a basket contains all of $i_1,\ldots,i_k$ then it is *likely* to contain j.”
- *Confidence* of this association rule is the probability of j given $i_1,\ldots,i_k$.

Example

- $B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$
- $B_3 = \{m, b\}$ $B_4 = \{c, j\}$
- $B_5 = \{m, p, b\}$ $B_6 = \{m, c, b, j\}$
- $B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$
- An association rule: $\{m, b\} \rightarrow c$.
- Confidence = 2/4 = 50%.

Finding Association Rules

- A typical question: “find all association rules with support ≥ *s* and confidence ≥ *c*.”
- Note: “support” of an association rule is the support of the set of items it mentions.
- Hard part: finding the high-support (*frequent*) itemsets.
- Checking the confidence of association rules involving those sets is relatively easy.

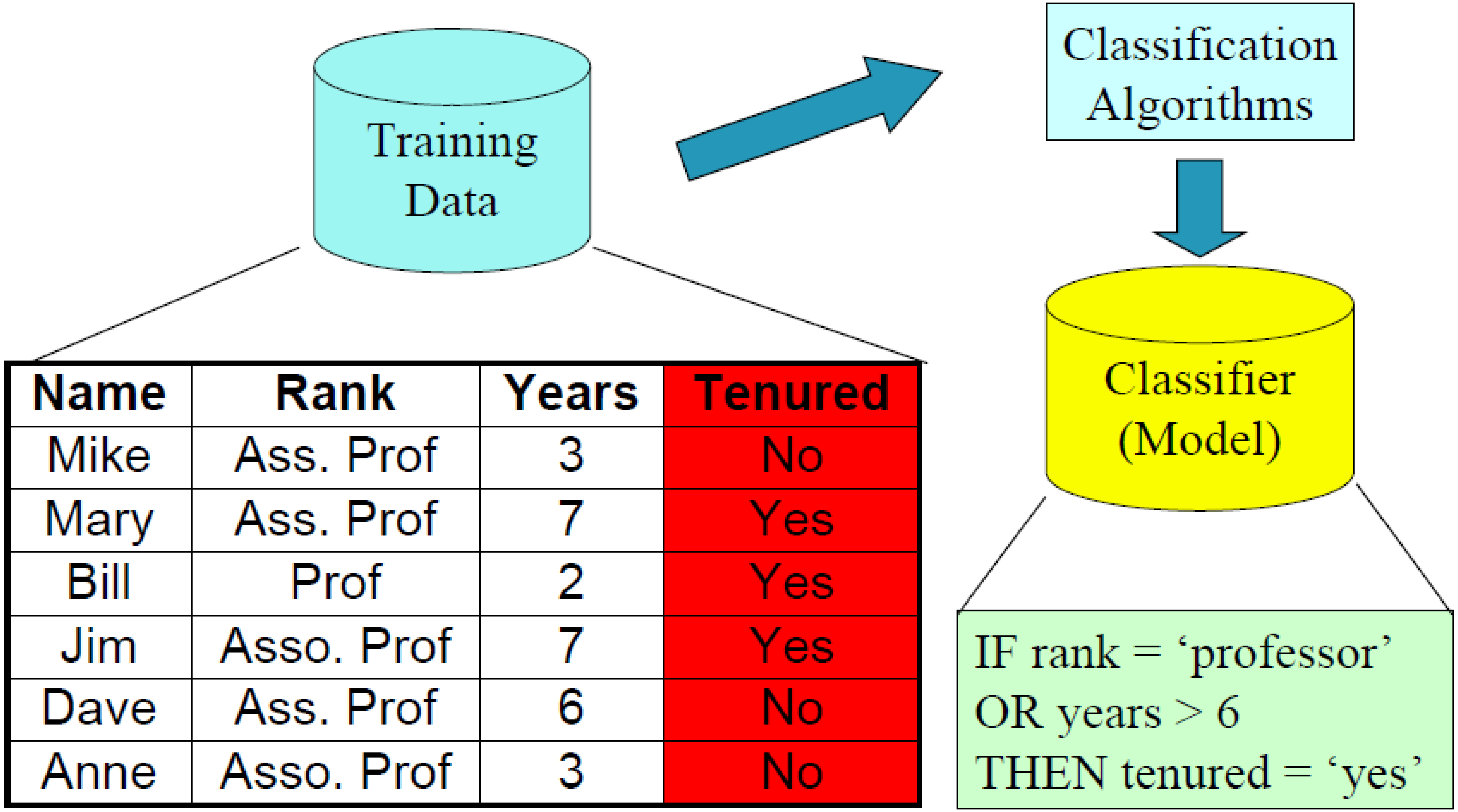
Classification

- Learning from Examples
  - Given a set of credit card frauds, can we detect frauds in the future?
    - Examples: credit card frauds
    - Goal: case predictions
  - Many applications
    - Fraud detection, intrusion detection, automatic credit approval, customer relationship management, spam detection, virus detection,

Classification: A 2-step Process

- Model construction: describe/summarize a set of predetermined classes
  - Training dataset: tuples for model construction
    - Each tuple/sample belongs to a predefined class
  - Model: classification rules, decision trees, or math formulae
- Model application: classify unseen objects
  - Estimate accuracy of the model using an independent test set
  - Acceptable accuracy -> apply the model to classify tuples with unknown class labels

Model Construction



Training Data

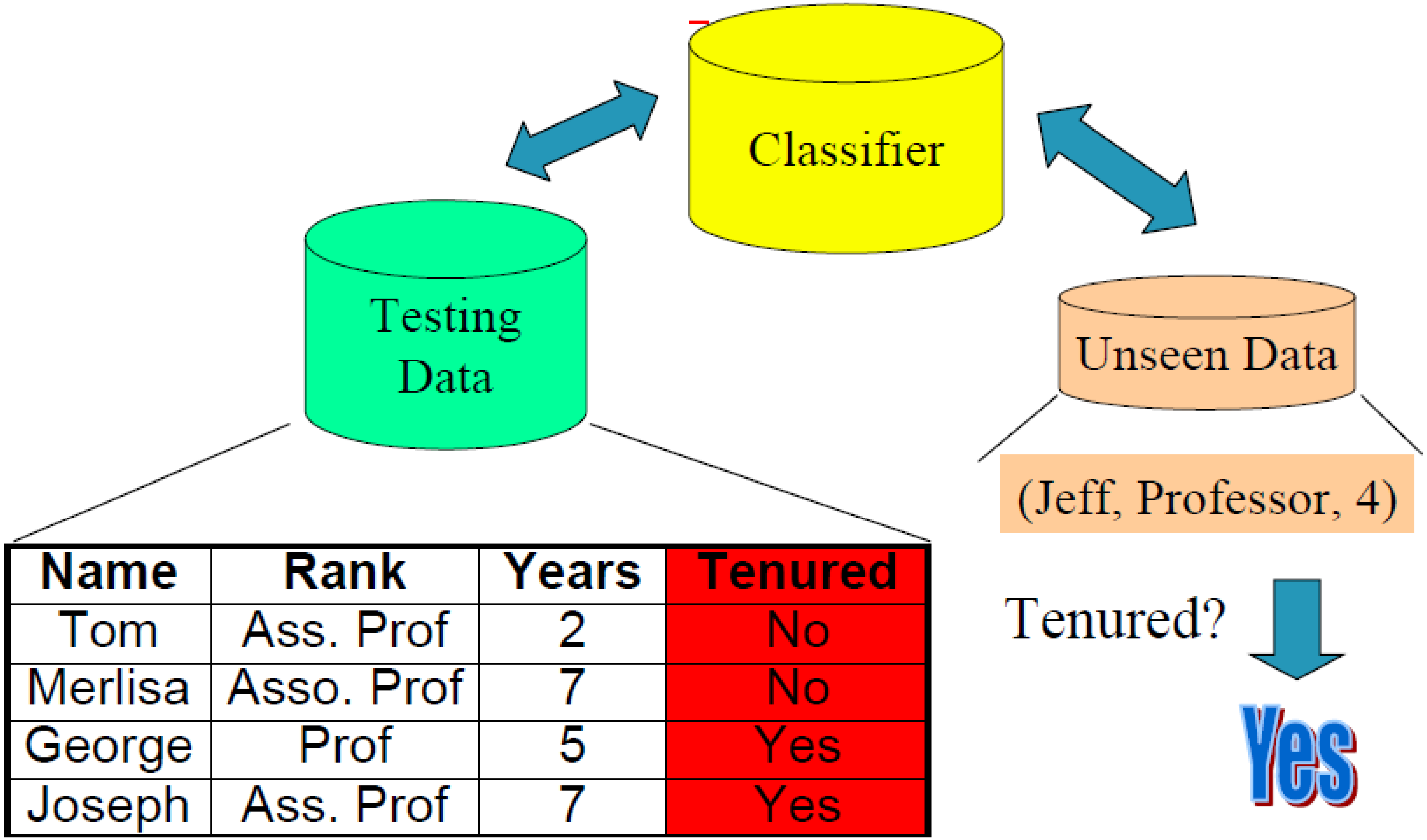
Classification Algorithms

Classifier (Model)

Name	Rank	Years	Tenured
Mike	Ass. Prof	3	No
Mary	Ass. Prof	7	Yes
Bill	Prof	2	Yes
Jim	Asso. Prof	7	Yes
Dave	Ass. Prof	6	No
Anne	Asso. Prof	3	No

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

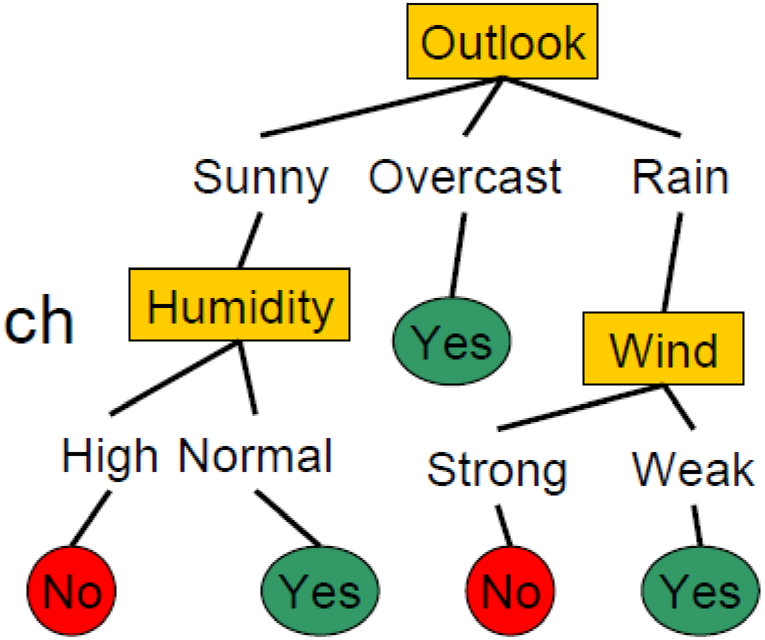
Model Application



Name	Rank	Years	Tenured
Tom	Ass. Prof	2	No
Merlisa	Asso. Prof	7	No
George	Prof	5	Yes
Joseph	Ass. Prof	7	Yes

An Example: Decision Tree

- A node in the tree – a test of some attribute
- A branch: a possible value of the attribute
- Classification
  - – Start at the root
  - – Test the attribute
  - – Move down to tree branch

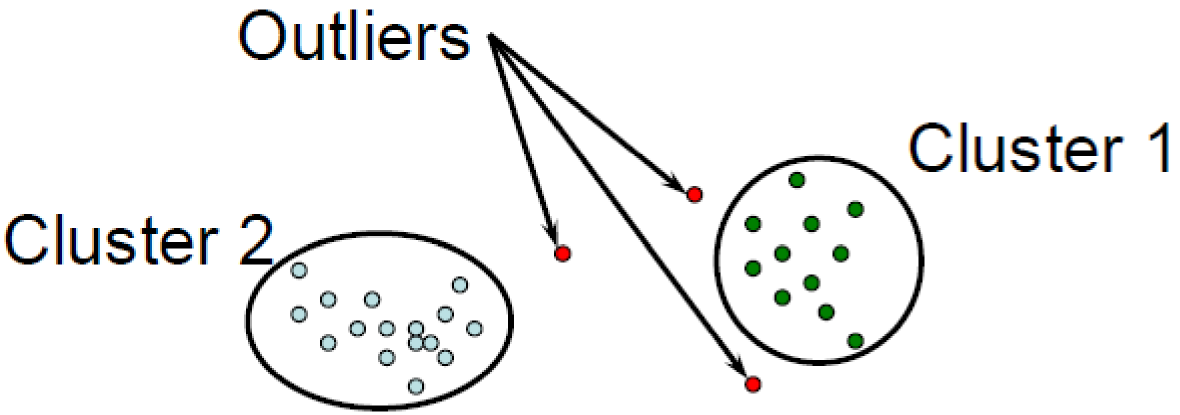


Clustering: Applications

- Customer segmentation
  - How to partition customers into groups so that customers in each group are similar, while customers in different groups are dissimilar?
- Pattern recognition in image
  - How to identify objects in a satellite image? The pixels of an object are similar to each other in some way

What is Clustering?

- Group data into clusters
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
  - Unsupervised learning: no predefined classes


Outliers
Cluster 1
Cluster 2