



Basic concepts of Data Mining, Clustering and Genetic Algorithms

Tsai-Yang Jea
Department of
Computer Science and Engineering
SUNY at Buffalo

Data Mining Motivation



- Mechanical production of data need for mechanical consumption of data
- Large databases = vast amounts of information
- Difficulty lies in accessing it

KDD and Data Mining



- **KDD:** Extraction of knowledge from data
 - “non-trivial extraction of implicit, previously unknown & potentially useful knowledge from data”
- **Data Mining:** Discovery stage of the KDD process

Data Mining Techniques



Any technique that helps to extract more out of data is useful

- Query tools
- Statistical techniques
- Visualization
- On-line analytical processing (OLAP)
- Clustering
- Classification
- Decision trees
- Association rules
- Neural networks
- Genetic algorithms

What's Clustering

- Clustering is a kind of unsupervised learning.
- Clustering is a method of grouping data that share similar trend and patterns.
- Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data.

– Example:



After clustering:



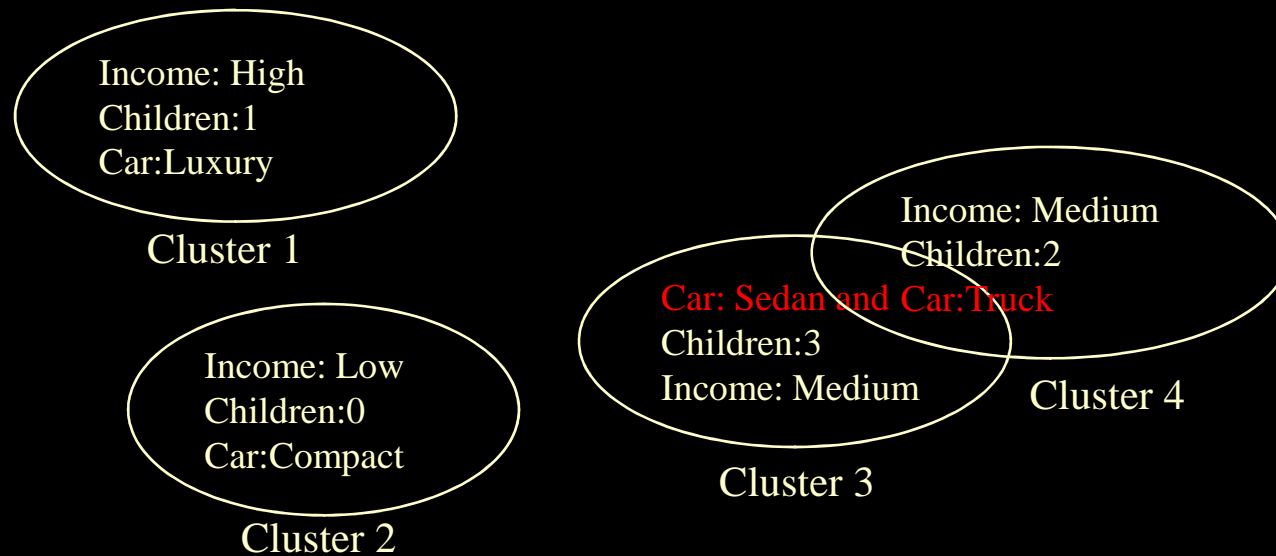
Thus, we see clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.

The usage of clustering

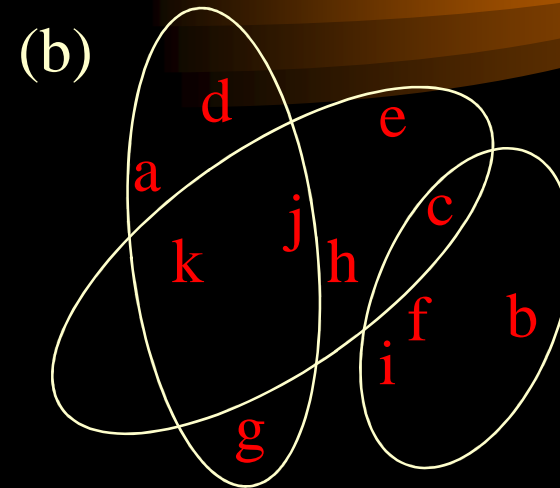
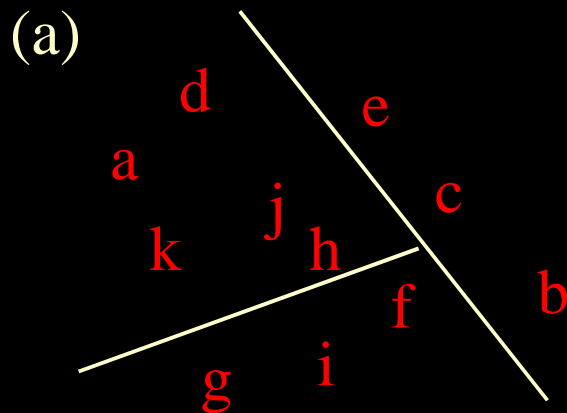


- Some engineering sciences such as pattern recognition, artificial intelligence have been using the concepts of cluster analysis. Typical examples to which clustering has been applied include handwritten characters, samples of speech, fingerprints, and pictures.
- In the life sciences (biology, botany, zoology, entomology, cytology, microbiology), the objects of analysis are life forms such as plants, animals, and insects. The clustering analysis may range from developing complete taxonomies to classification of the species into subspecies. The subspecies can be further classified into subspecies.
- Clustering analysis is also widely used in information, policy and decision sciences. The various applications of clustering analysis to documents include votes on political issues, survey of markets, survey of products, survey of sales programs, and R & D.

A Clustering Example



Different ways of representing clusters




(c)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
...			



K Means Clustering

(Iterative distance-based clustering)



- K means clustering is an effective algorithm to extract a **given number** of clusters of patterns from a training set. Once done, the cluster locations can be used to classify patterns into distinct classes.

K means clustering (Cont.)

Select the k cluster centers randomly.



Classify the entire training set. For each pattern X_i in the training set, find the nearest cluster center C^* and classify X_i as a member of C^* .



Loop until the change in cluster means is less the amount specified by the user.

For each cluster, recompute its center by finding the mean of the cluster :

$$M_k = \frac{1}{N_k} \cdot \sum_{j=1}^{N_k} X_{jk}$$

where M_k is the new mean, N_k is the number of training patterns in cluster k , and X_{jk} is the j -th pattern belonging to cluster k .



Store the k cluster centers.

The drawbacks of K-means clustering



- The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers. (fig. 1)
- We have to know how many clusters we will have at the first.

Drawback of K-means clustering (Cont.)

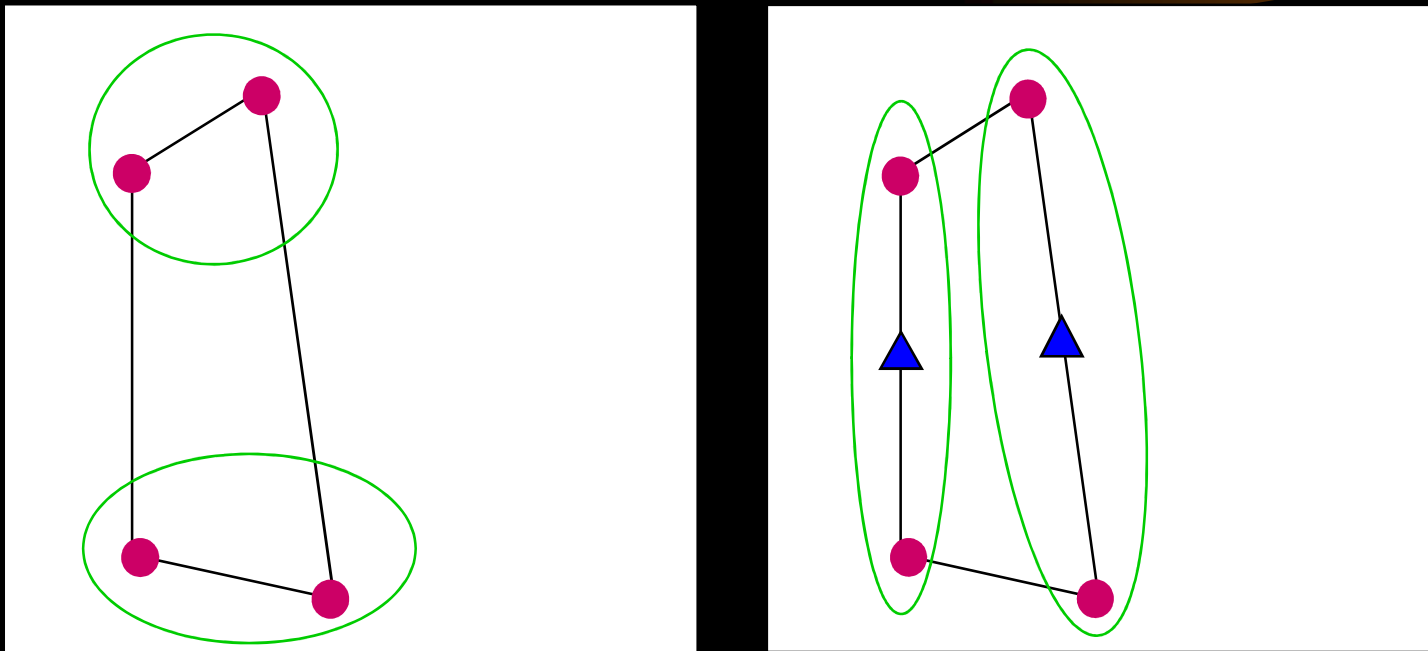


Figure 1

Clustering with Genetic Algorithm



- Introduction of Genetic Algorithm
- Elements consisting GAs
- Genetic Representation
- Genetic operators

Introduction of GAs

A decorative graphic consisting of a horizontal bar with a color gradient from dark blue on the left to bright yellow on the right. To the right of the bar is a large, stylized comet-like tail that tapers to a point, also following the same color gradient.

- Inspired by biological evolution.
- Many operators mimic the process of the biological evolution including
 - Natural selection
 - Crossover
 - Mutation

Elements consisting GAs



- Individual (chromosome):
 - feasible solution in an optimization problem
- Population
 - Set of individuals
 - Should be maintained in each generation

Elements consisting GAs



- Genetic operators. (crossover, mutation...)
- Define the fitness function.
 - The fitness function takes a single chromosome as input and returns a measure of the goodness of the solution represented by the chromosome.

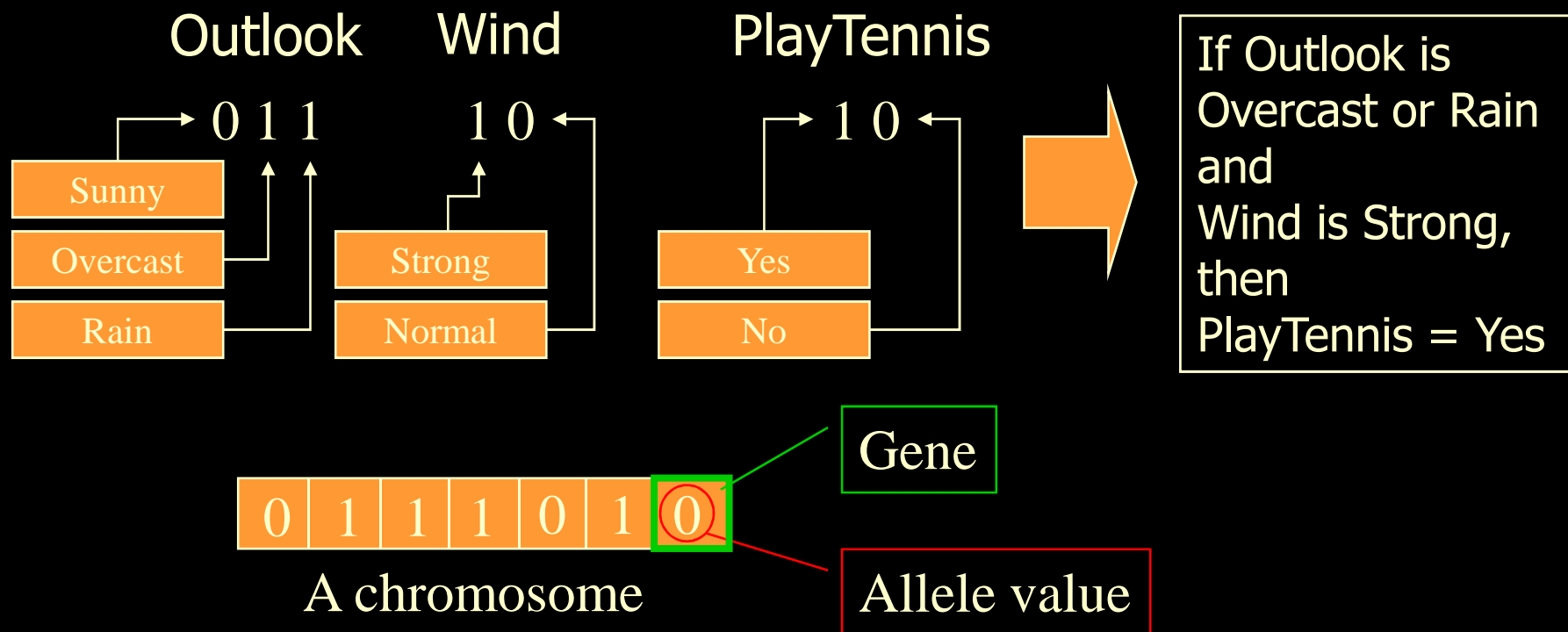
Genetic Representation



- The most important starting point to develop a genetic algorithm
- Each gene has its special meaning
- Based on this representation, we can define
 - fitness evaluation function,
 - crossover operator,
 - mutation operator.

Genetic Representation (Cont.)

- Examples 1



Genetic Representation (Cont.)

- **Examples 2** (In clustering problem)
 - Each chromosome represents a set of clusters; each gene represents an object; each allele value represents a cluster. Genes with the same allele value are in the same cluster.

1	2	1	4	3	5	5
A	B	C	D	E	F	G

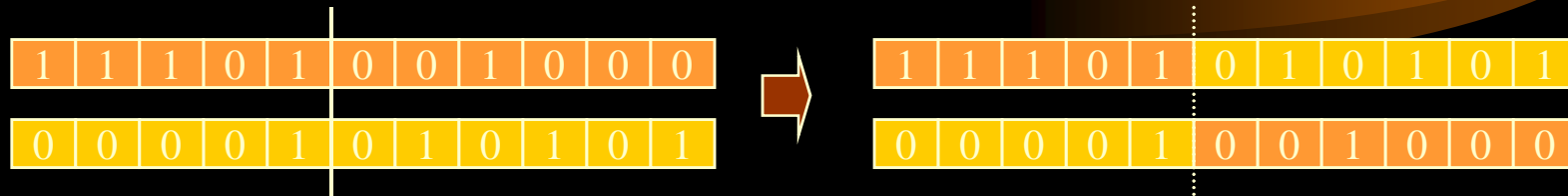
Crossover



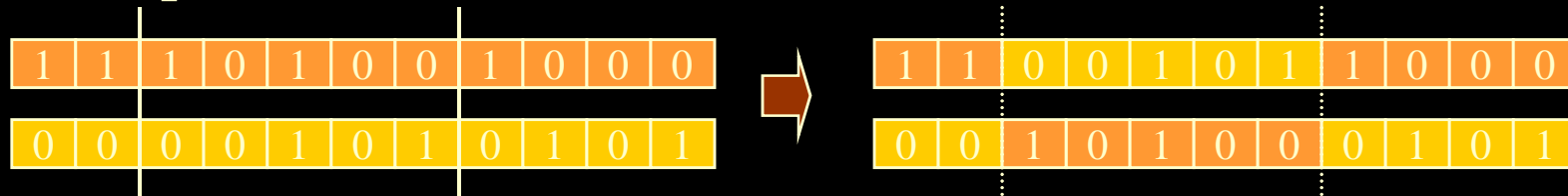
- *Exchange features* of two individuals to produce two offspring (children)
- Selected mates may have good properties to survive in next generations
- So, we can expect that exchanging features may produce other good individuals

Crossover (cont.)

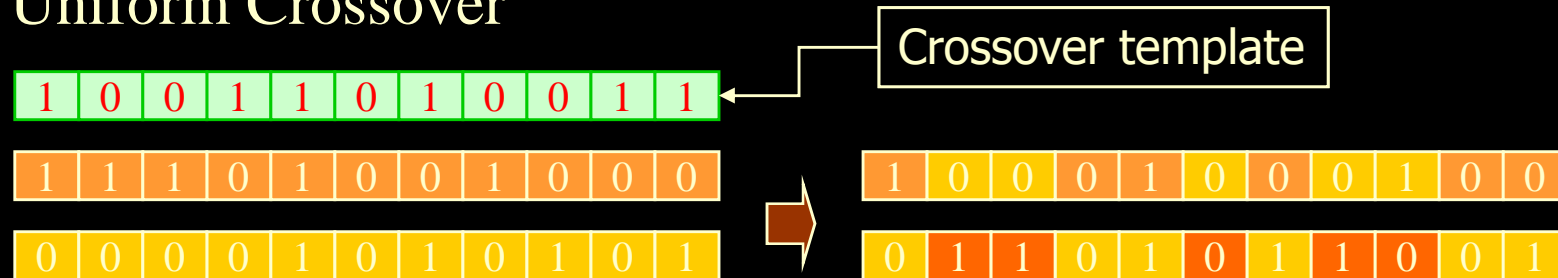
- Single-point Crossover



- Two-point Crossover

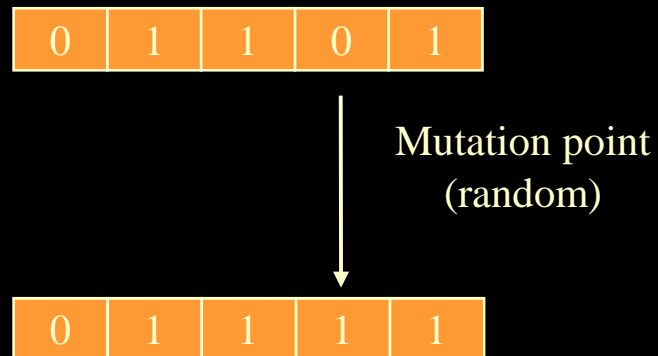


- Uniform Crossover

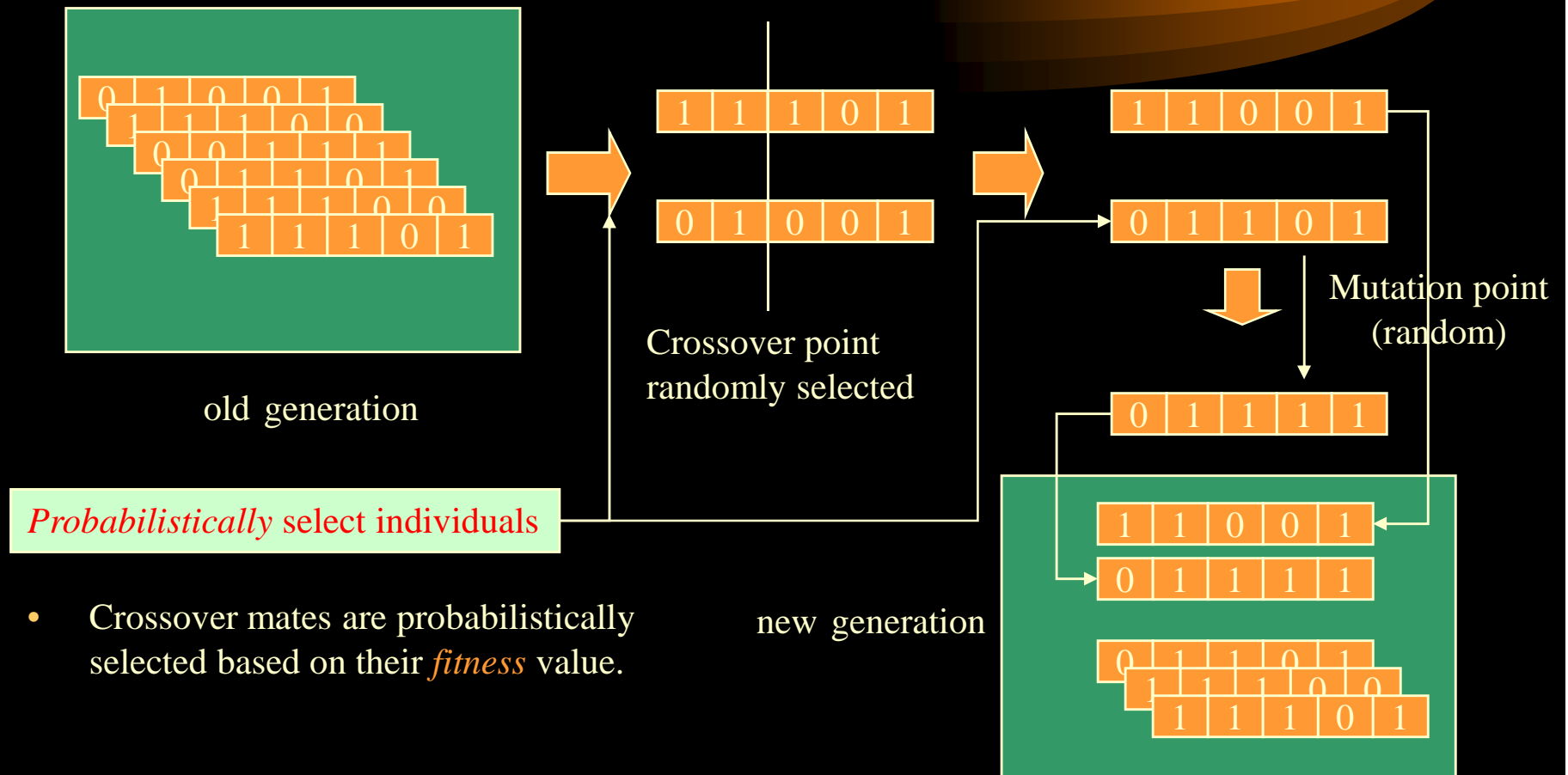


Mutation

- Usually change a single bit in a bit string
- This operator should happen with **very low** probability.



Typical Procedures



- Crossover mates are probabilistically selected based on their *fitness* value.

How to apply GA on a clustering problem

- Preparing the chromosomes



- Defining genetic operators

- Fusion: takes two unique allele values and combines them into a single allele value, combining two clusters into one.



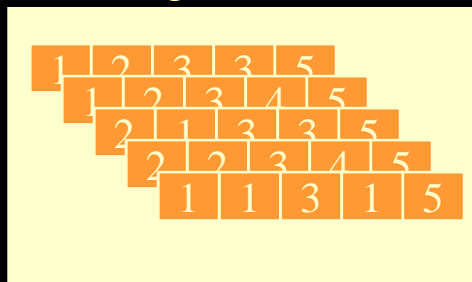
- Fission: takes a single allele value and gives it a different random allele value, breaking a cluster apart.



- Defining fitness functions

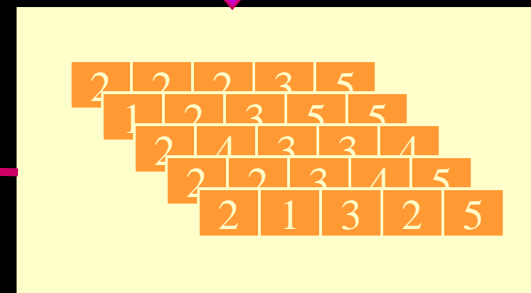
Example: (Cont.)

Old generation



Crossover
Mutation
Fusion
Fission

Select the chromosomes
according to the fitness
function.



New generation

Finally...

Thank You