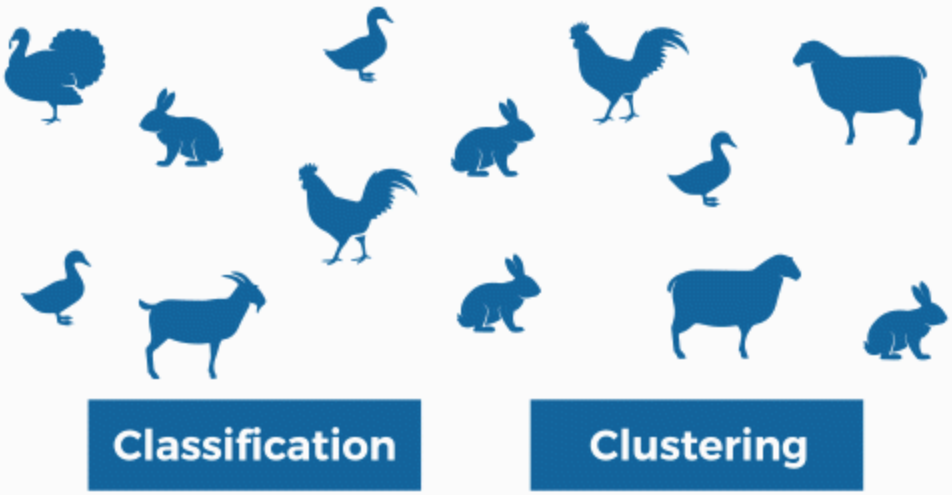
# Data Mining: Classification vs Clustering

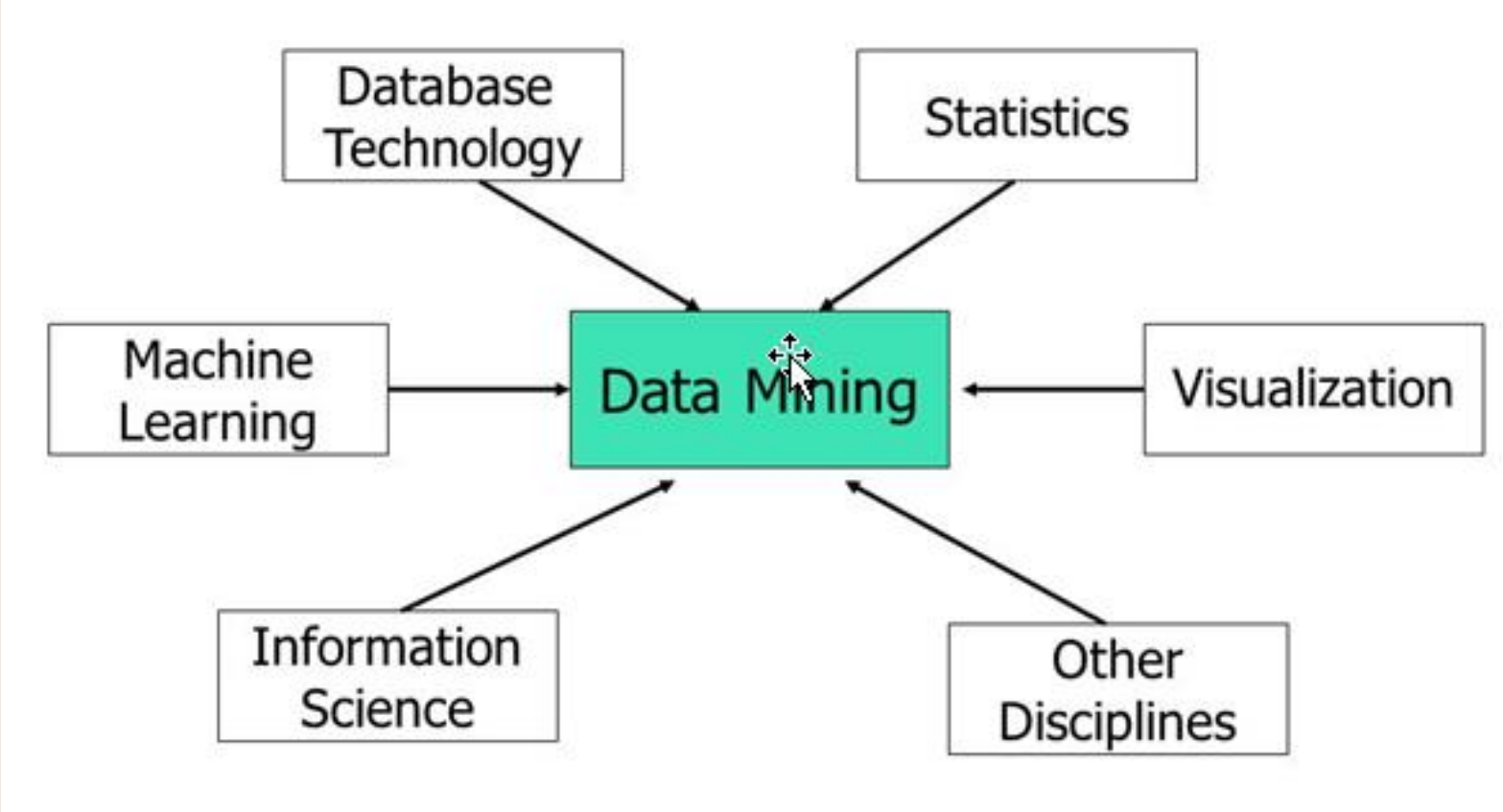
Yetli Oslan

Revisi @2020

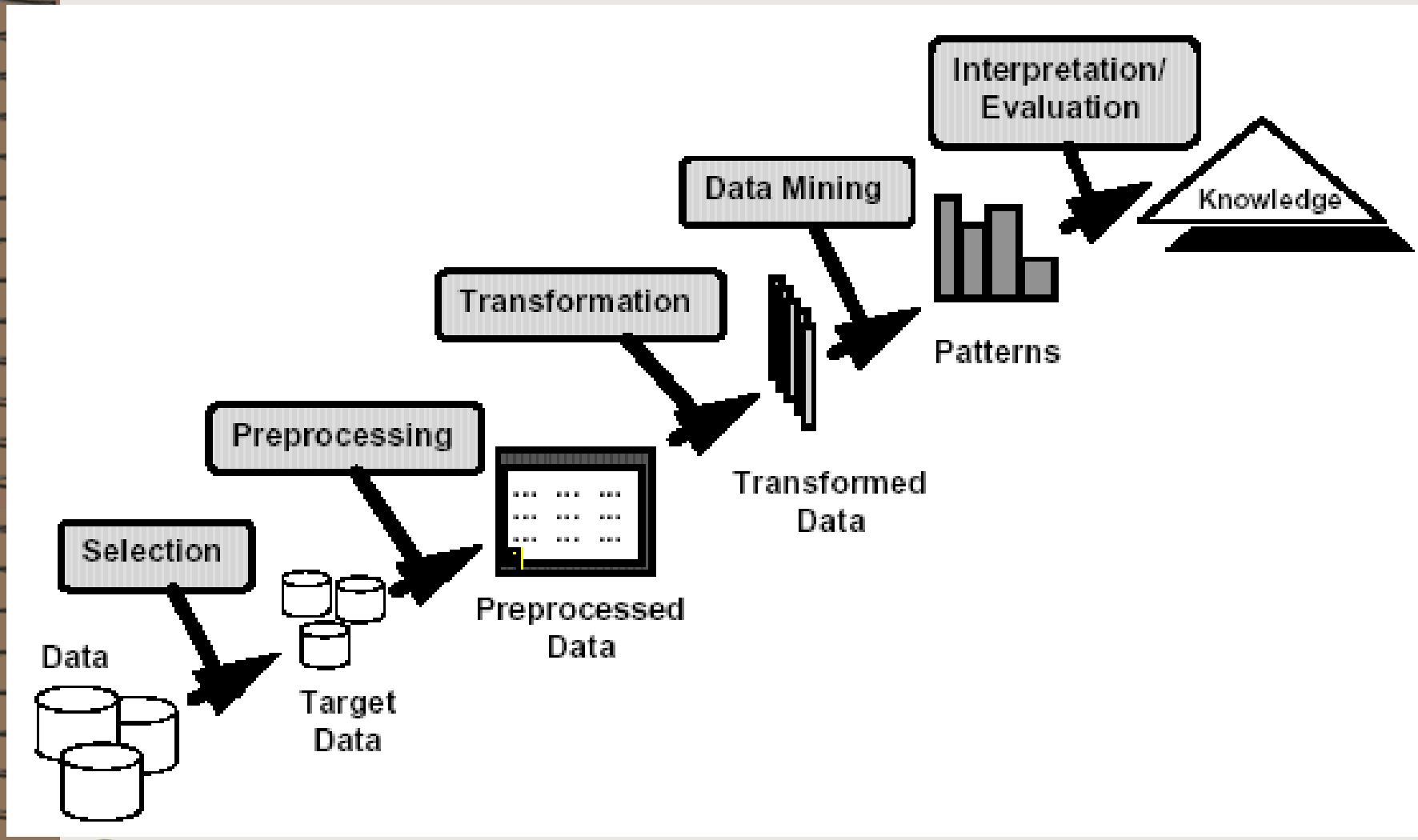
Sumber:
https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation#:~:text=Although%20both%20techniques%20have%20certain,which%20differentiate%20them%20from%20other



# Data Mining: confluence of Multiple Disciplines



# Knowledge Discovery



# Classification

Classification: memberikan label kelas atas satu set data yang belum diklasifikasikan

1. **Supervised Classification**

   Kumpulan kelas yang sudah dikenal sebelumnya.

2. **Unsupervised Classification**

   Kumpulan kelas yang mungkin tidak diketahui. Setelah klasifikasi, kita dapat mencoba memberikan nama ke kelas itu. **Unsupervised Classification** disebut **clustering**

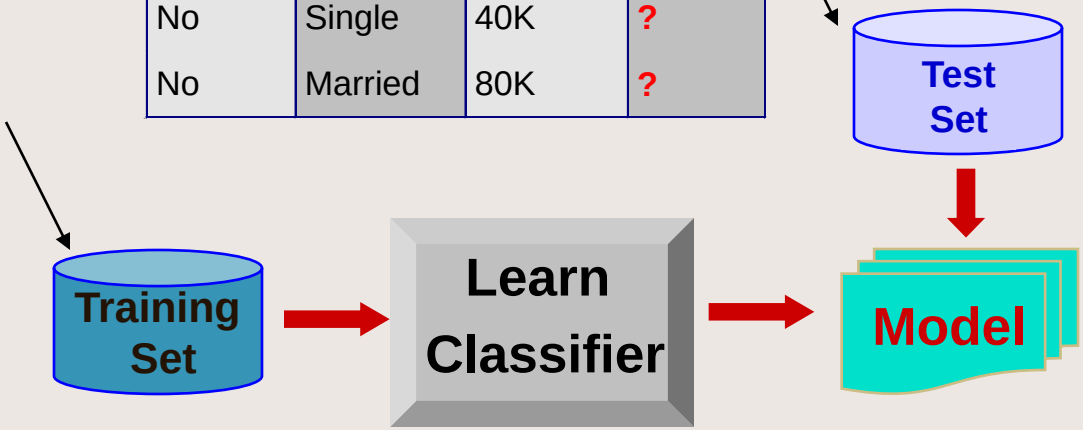
# Supervised Classification

- input data disebut ***training set***, merupakan kumpulan record yang masing-masing record memiliki beberapa atribut atau fitur
- Setiap record ditandi dengan sebuah label ***class***.
- Tujuan klasifikasi adalah untuk menganalisis data input dan mengembangkan model yang akurat untuk setiap kelas menggunakan fitur-fitur yang ada dalam data.
- Model ini digunakan untuk mengklasifikasikan ***test set*** yang deskripsi kelasnya tidak diketahui. (1)

# Classification Example

| Tid | Home Owner (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Default (class) |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

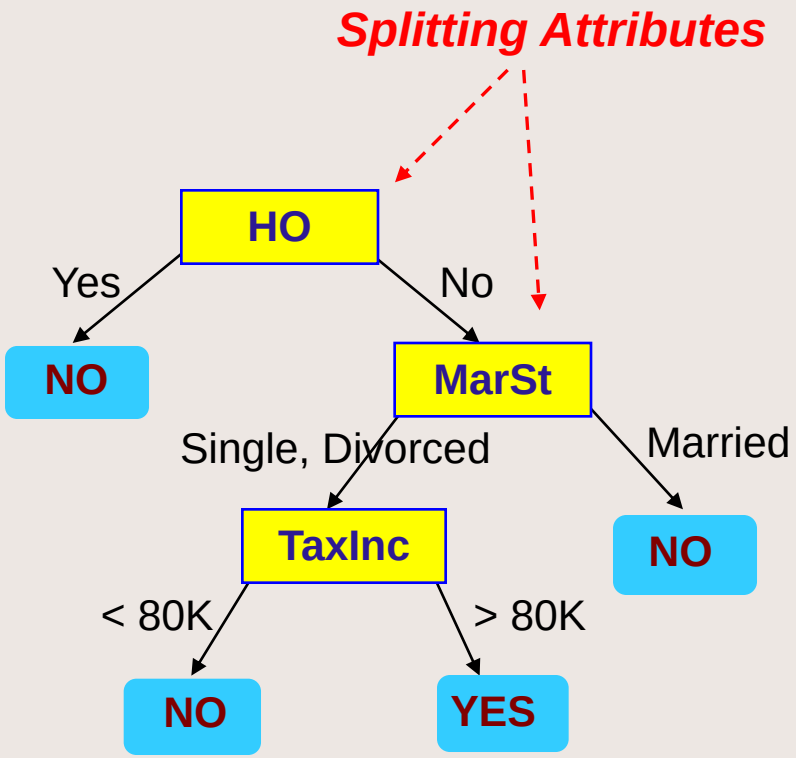
| Home Owner | Marital Status | Taxable Income | Default |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



# Example of a Decision Tree

| Tid | Home Owner | Marital Status | Taxable Income | Default |
|---|---|---|---|---|
| | categorical | categorical | continuous | class |
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

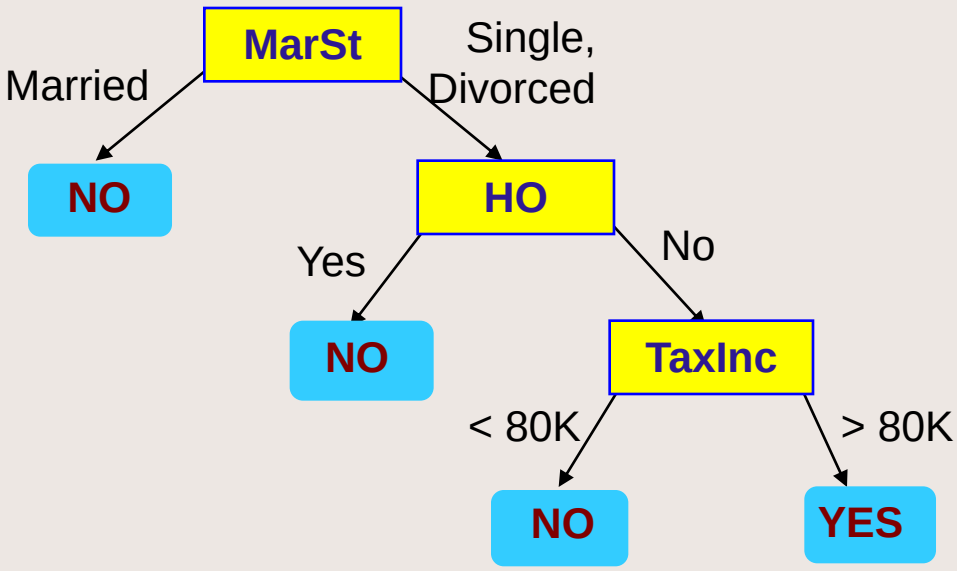
**Training Set**




**Model: Decision Tree**

# Another Example of Decision Tree

| Tid | Home Owner (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Default (class) |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

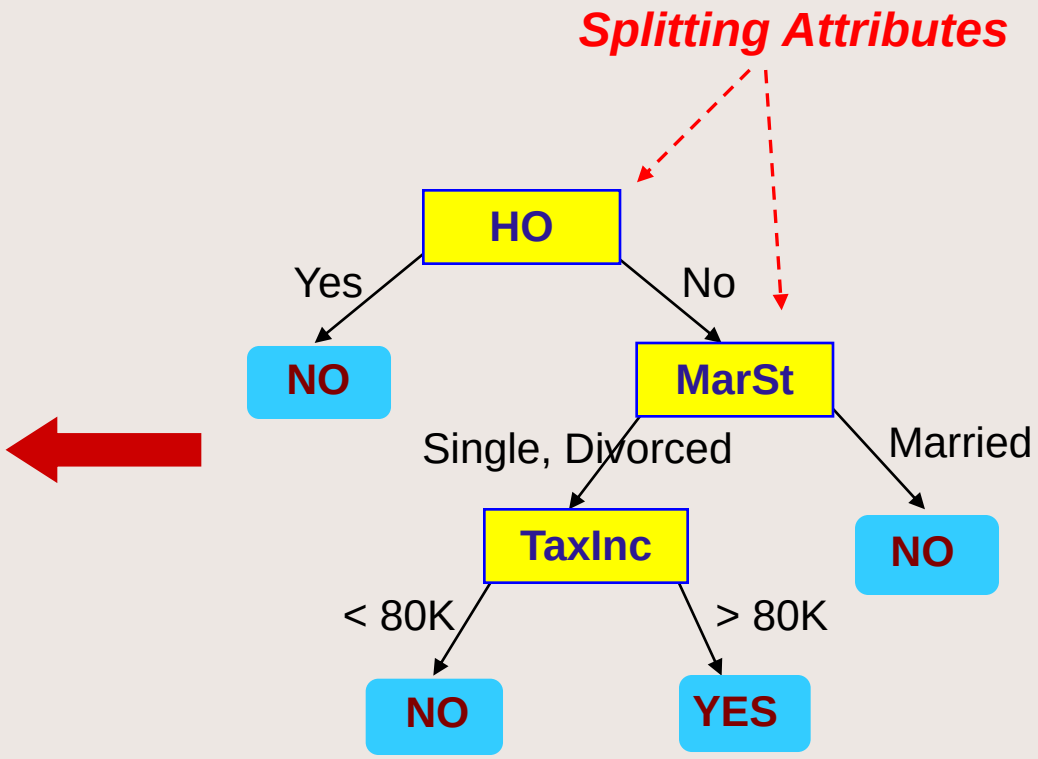



**There could be more than one tree that fits the same data!**

# Example of a Decision Tree

| Home Owner | Marital Status | Taxable Income | Default |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |

**Test Set**

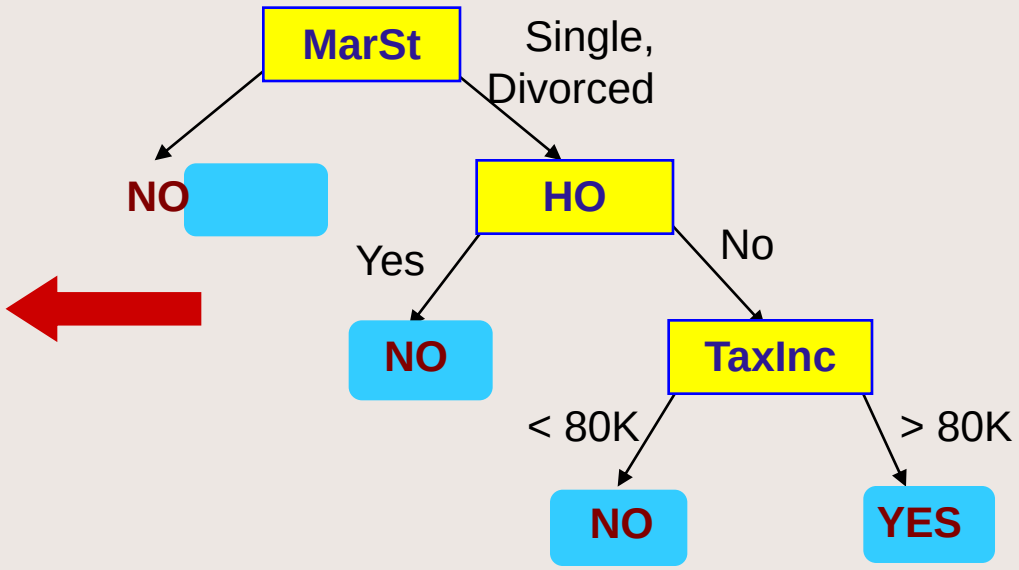



**Model: Decision Tree**

# Example of a Decision Tree

| Home Owner | Marital Status | Taxable Income | Default |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |

**Test Set**

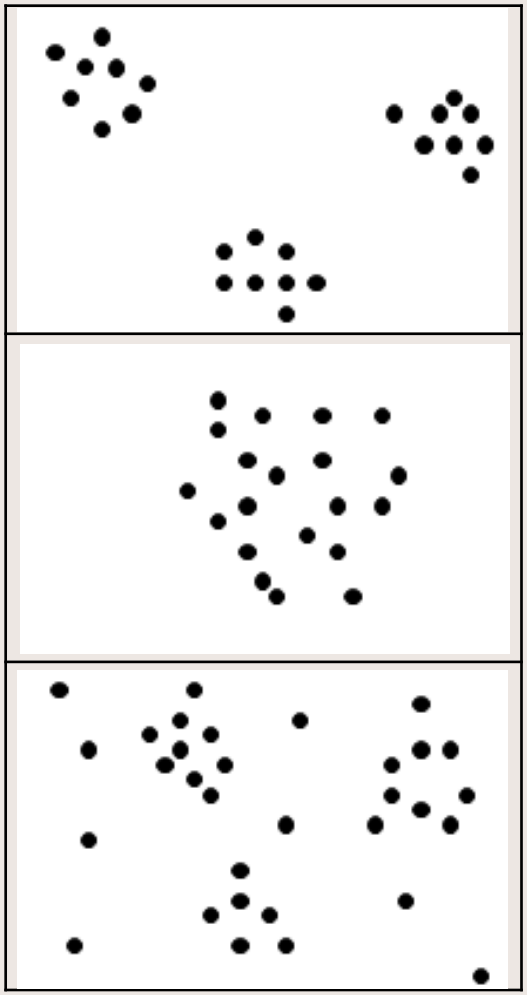



**Model: Decision Tree**

# Clustering

Clustering: partisi sekolompok data dalam satu cluster, lalu temukan model (kesamaan) dari sekelompok data tersebut yang berbeda dari kelompok data lainnya

# Clustering



Terkadang mudah
***(easy)***

Terkadang tidak mungkin
***(impossible)***

Terkadang membingungkan
***(in between)***

# Why is Clustering useful?

- "Discovery" pengetahuan baru dari sekumpulan data
  - Berbeda dengan *supervised classification* (dimana label Class diketahui)
  - Sejarah panjang dalam ilmu kategori, taksonomi, dll
  - Dapat sangat berguna untuk menyimpulkan (meringkas) kumpulan data yang besar
    - Untuk jumlah data yang banyak dan/ atau dimensi yang tinggi
- Penerapan teknik *clustering*
  - Clustering of documents produced by a search engine
  - Segmentation of customers for an e-commerce store
  - Discovery of new types of galaxies in astronomical data
  - Clustering of genes with similar expression profiles
  - Cluster pixels in an image into regions of similar intensity
  - dll

# Biomedical data mining and DNA analysis

- **DNA sequences consist of 4 basic building blocks (nucleotides):** adenine (A), cytosine (C), guanine (G), and thymine (T).
- **Gene**: a sequence of hundreds of individual nucleotides arranged in a particular order
- **Semantic integration of heterogeneous, distributed genome databases**
  - data cleaning and data integration methods developed in data mining will help

# Data mining systems (1)

- **IBM Intelligent Miner**
  - a wide range of data mining algorithms
  - scalable mining algorithms
  - **toolkits:** neural network algorithms, statistical methods, data preparation, and data visualization tools
  - tight integration with IBM's DB2 relational database system
- **SAS Enterprise Miner**
  - a variety of statistical analysis tools
  - data warehouse tools and multiple data mining algorithms

# Data mining systems (2)

- **SGI MineSet**
  - multiple data mining algorithms and advanced statistics
  - advanced visualization tools
- **Clementine (SPSS)**
  - an integrated data mining development environment for end-users and developers
  - multiple data mining algorithms and visualization tools

# Data mining systems (3)

- **DBMiner (DBMiner Technology Inc.)**
  - multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
  - efficient, association and sequential-pattern mining functions, and visual classification tool
  - mining both relational databases and data warehouses
- **Microsoft SQLServer 2000**
  - integrate DB and OLAP with mining
  - support OLEDB for DM standard