Konsep Data Mining

Konsep Data Mining

Outline

Pengertian Dasar

Arsitektur

Tugas Data Mining

Contoh Penggunaan Data Mining

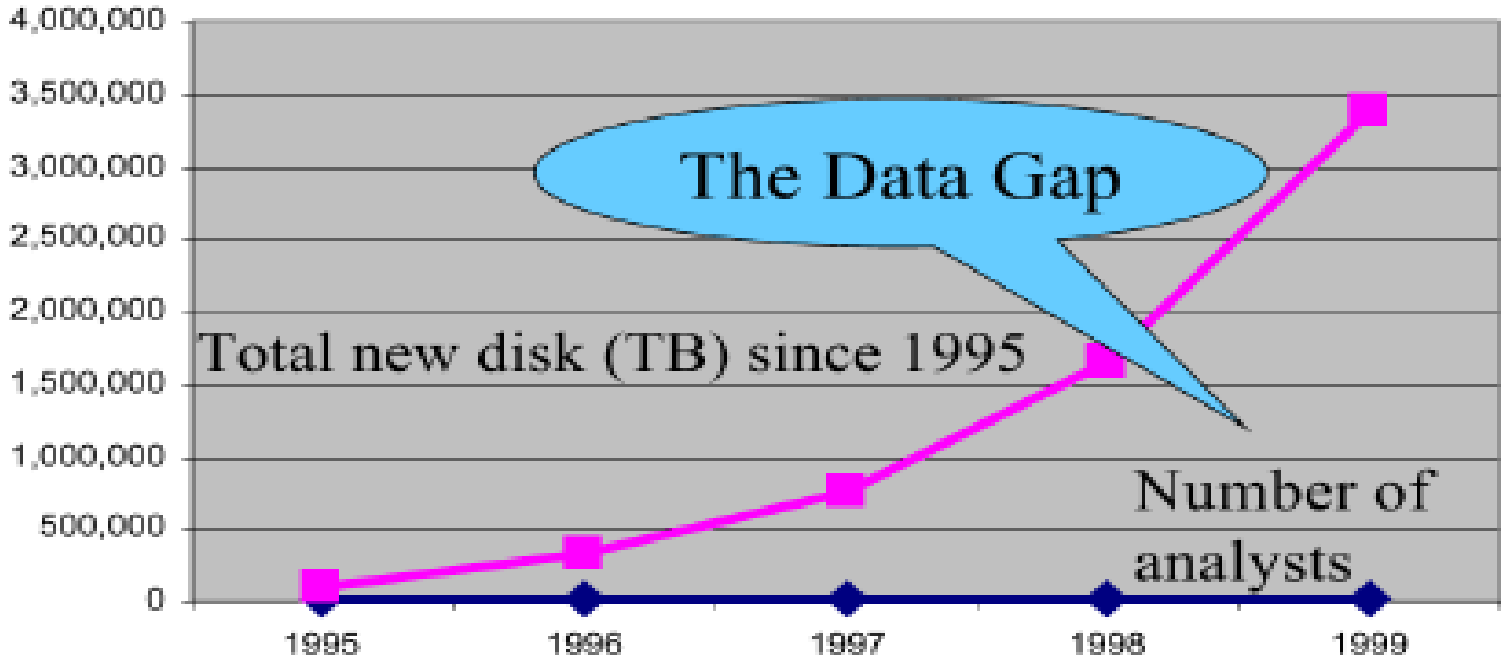
Latar Belakang

- Banyak data yang telah direkam dan disimpan:
  - Transaksi penjualan supermarket
  - Transaksi perbankan dan kartu kredit
  - Log kunjungan Web (access_log)
  - Akuisisi data dalam penelitian-penelitian seperti astronomi, kesehatan, dll
- Sistem komputer lebih murah dan cepat (Moore's Law)

- Kebutuhan untuk berkompetisi dengan strategi yang tepat menjadi lebih tinggi

Mengapa harus Data Mining?

- Data yang sedemikian besar kadang memiliki informasi yang tersembunyi
- Kemampuan manusia terbatas untuk "**mempelototi**" data-data tersebut dalam analisis



Definisi Data Mining

Data

Rekaman atau catatan terhadap fakta / transaksi / obyek

Definisi

- Ekstraksi informasi yang implisit, tidak diketahui sebelumnya, dan berpotensi berguna
- Eksplorasi dan analisis, secara otomatis atau tidak, data yang sangat besar untuk menemukan pola-pola yang berguna dan mempunyai arti

Pengertian Yang Salah

Bukan Data Mining

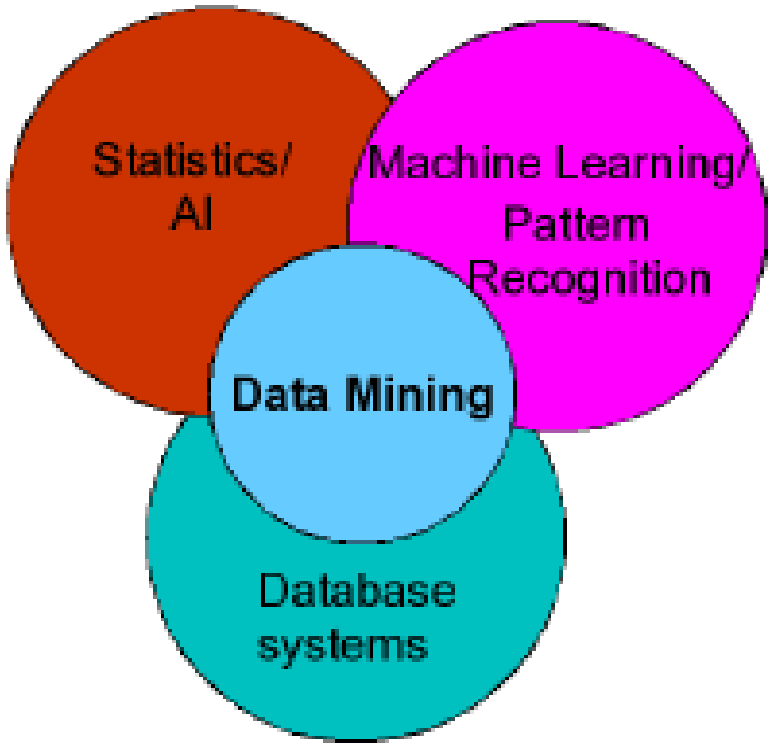
- Mencari nomor telepon "Bambang Gunawan" di buku telepon Indonesia
- Mencari informasi mengenai "Bunga" di google.com

Data Mining

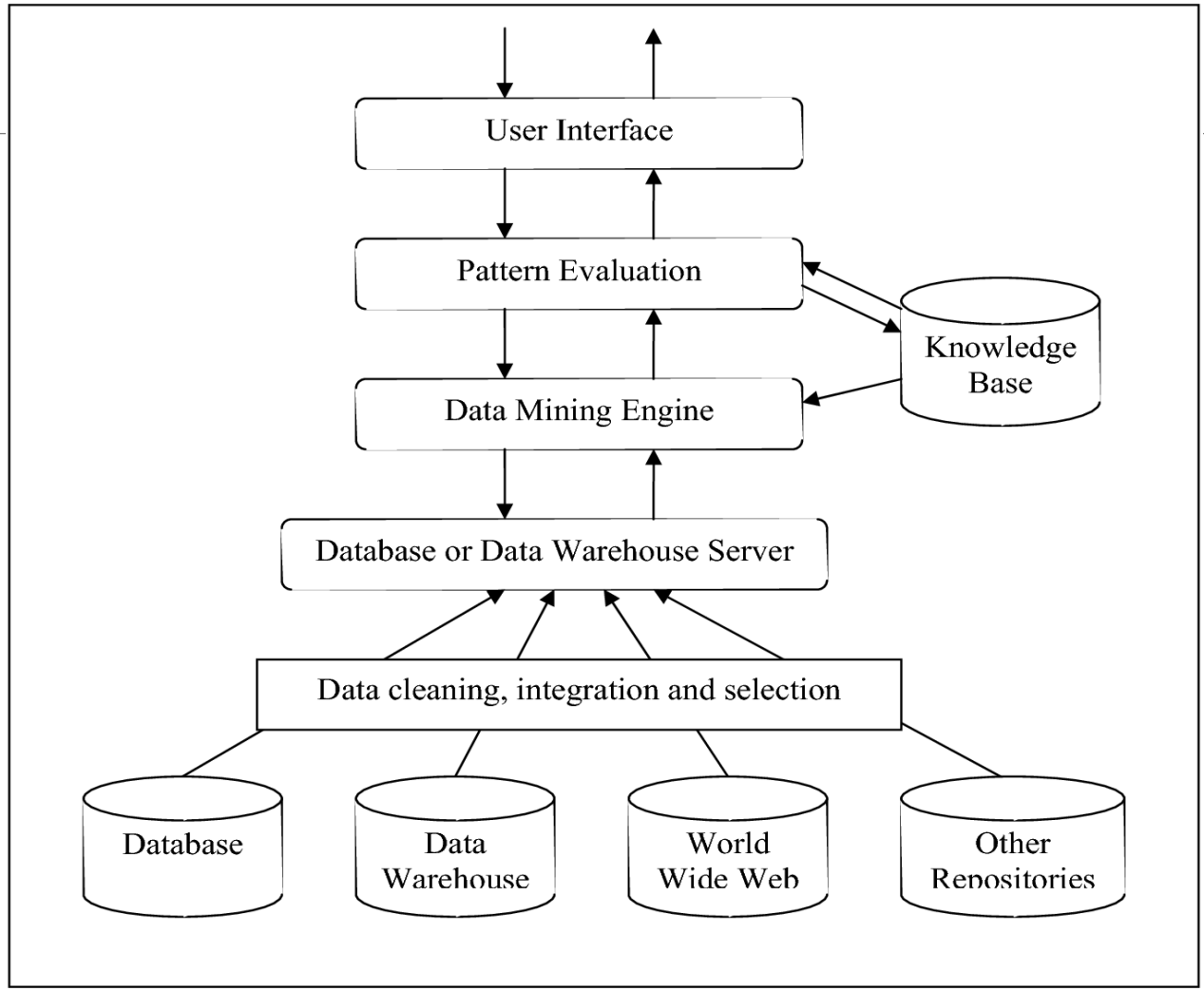
- Menemukan bahwa banyak orang bernama Bambang di daerah Jawa Timur
- Mengelompokkan dokumen web mengenai "Bunga" sesuai dengan konteks
  - Bunga Bank atau Kredit (Keuangan)
  - Bunga - Tanaman / Pertanian
  - BCL (Artis)

Ilmu Data Mining

- Gabungan dari beberapa bidang ilmu dalam Matematik dan Ilmu Komputer
- Diperlukan karena:
  - Data yang s(u)angat b(u)esar
  - Dimensi data yang besar
  - Data Heterogen



Arsitektur Data Mining -1-



Arsitektur Data Mining -2-

Knowledge Base

- Daerah knowledge yang digunakan untuk memberi petunjuk pencarian atau mengevaluasi hasil pola

Data Mining Engine

- Terdiri dari sekumpulan model fungsional seperti *characterization, association, classification, cluster analysis, evaluation and deviation analysis*

Arsitektur Data Mining -3-

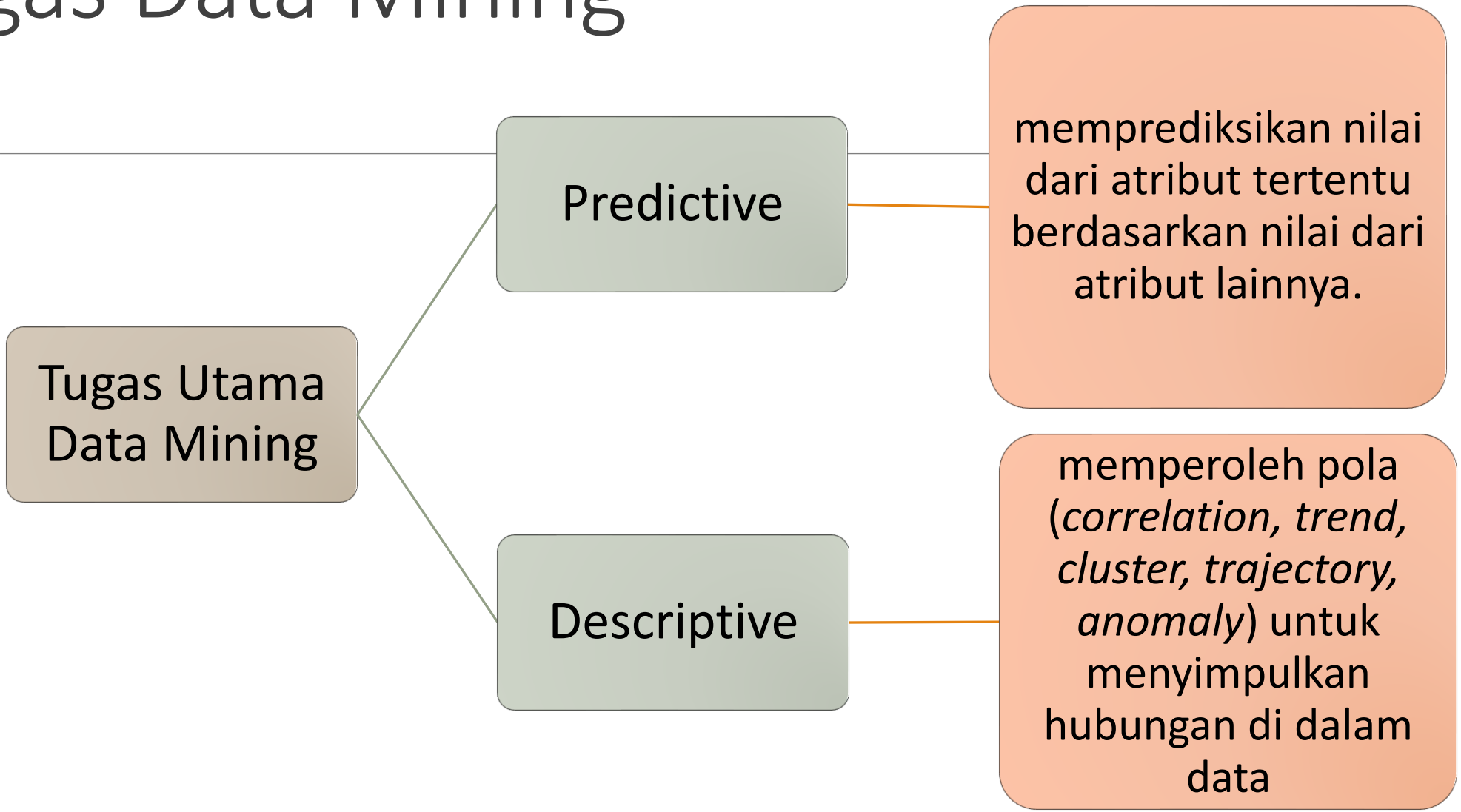
Pattern Evaluation Module

- Komponen yang berinteraksi dengan modul data mining untuk pencarian pola

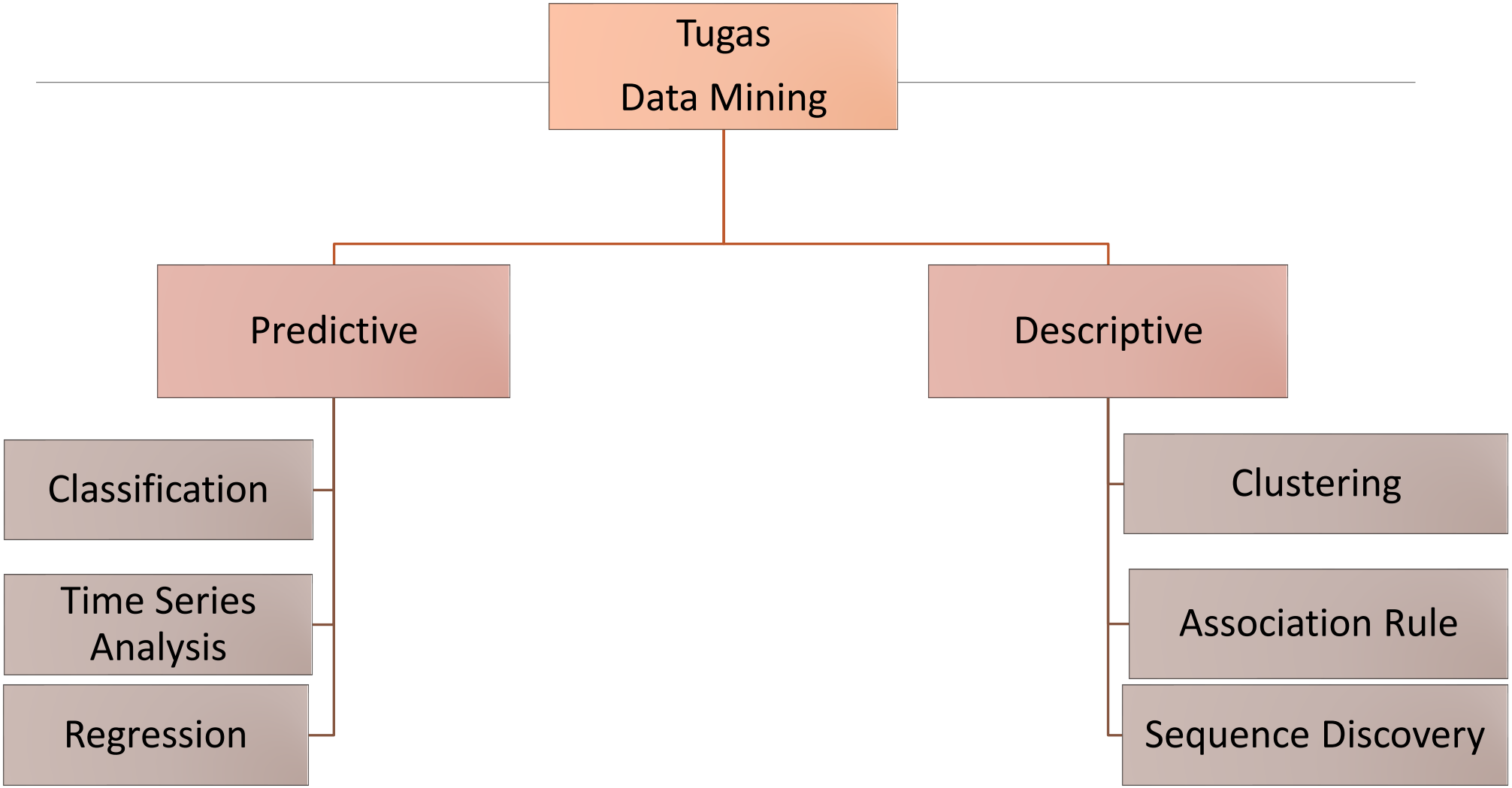
GUI *(Graphical User Interface)*

- modul yang mempermudah *user* berinteraksi dengan sistem data mining

Tugas Data Mining



Metode dalam Data Mining



Predictive - Classification

Menemukan fungsi atau model yang membedakan kelas data

Fungsi atau model tsb dapat berbentuk aturan if-else, descision tree, formula matematika, atau neural network

Tujuannya untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui

Algoritma : Decision Tree (C4.5), Artificial Neural Network, Naïve Bayes, Genetic Algorithm, Fuzzy, Case-Based Reasoning, k-Nearest Neighbor

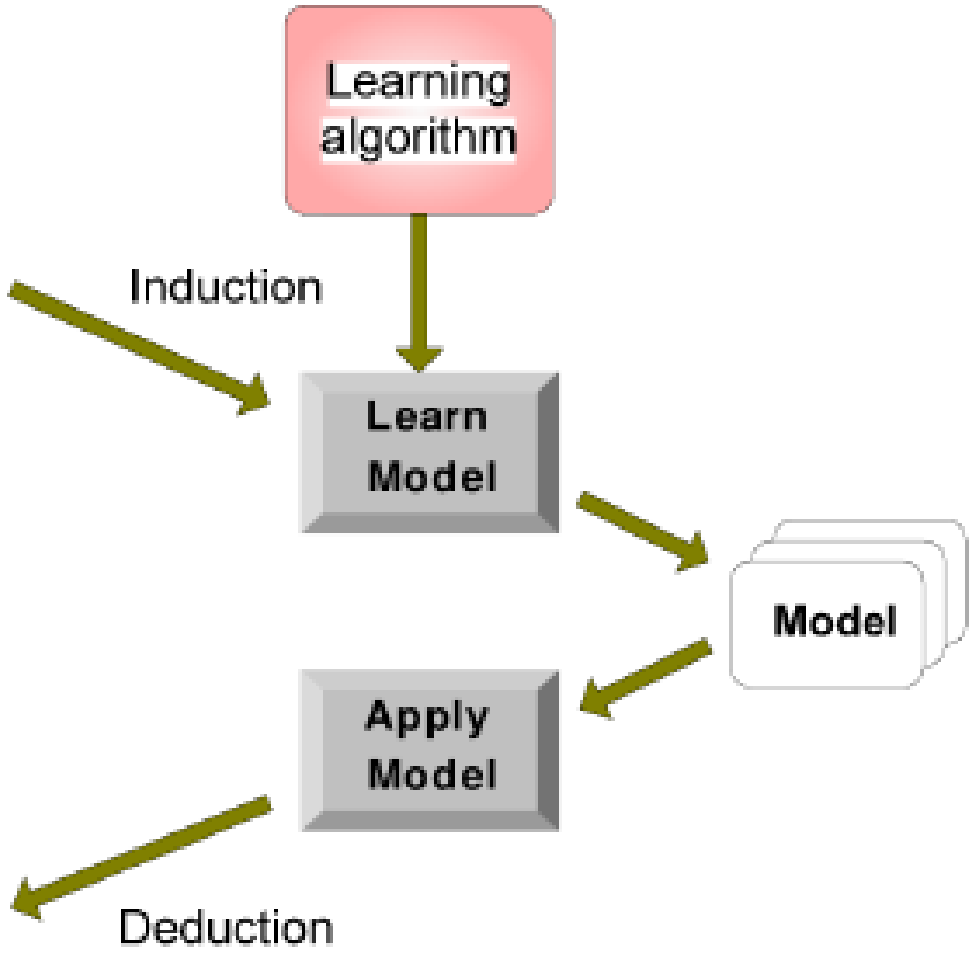
Supervised Method

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Contoh

- Pemakaian Kartu Kredit secara Ilegal
  - Tujuan : mendeteksi adanya penggunaan kartu kredit secara ilegal
  - Pendekatan :
    - Data transaksi sebelumnya (*lokasi & waktu transaksi, jenis barang yang dibeli, besar transaksi*)
    - Label data-data tersebut dengan Curang atau Aman
    - DM mencari model klasifikasi Curang atau Aman berdasarkan atribut transaksi
    - Menerapkan model tersebut jika ada transaksi baru untuk mempercepat / tepat tindakan preventif

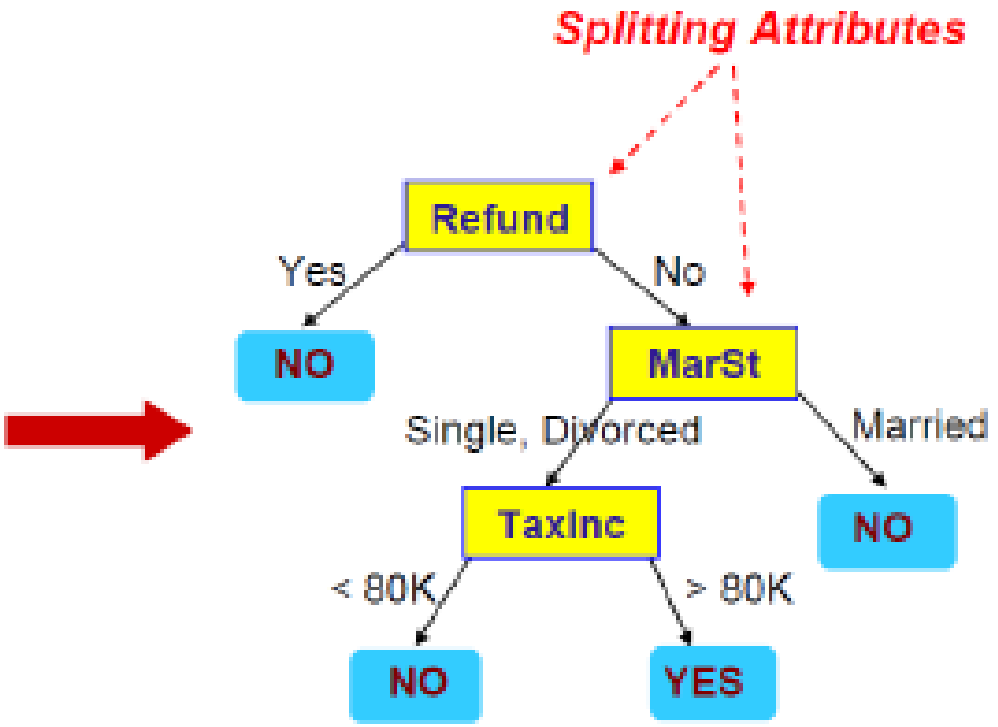
Contoh Lain..

- Deteksi SPAM
  - Tujuan : mendeteksi email yang tidak diharapkan secara dini
- Direct Marketing
  - Tujuan : mencari pengelompokan profil pelanggan agar target marketing sesuai
- Sky Survey Cataloging
  - Tujuan : mengelompokkan obyek langit hasil pemotretan teleskop ke dalam class-nya

Metode Pohon Keputusan

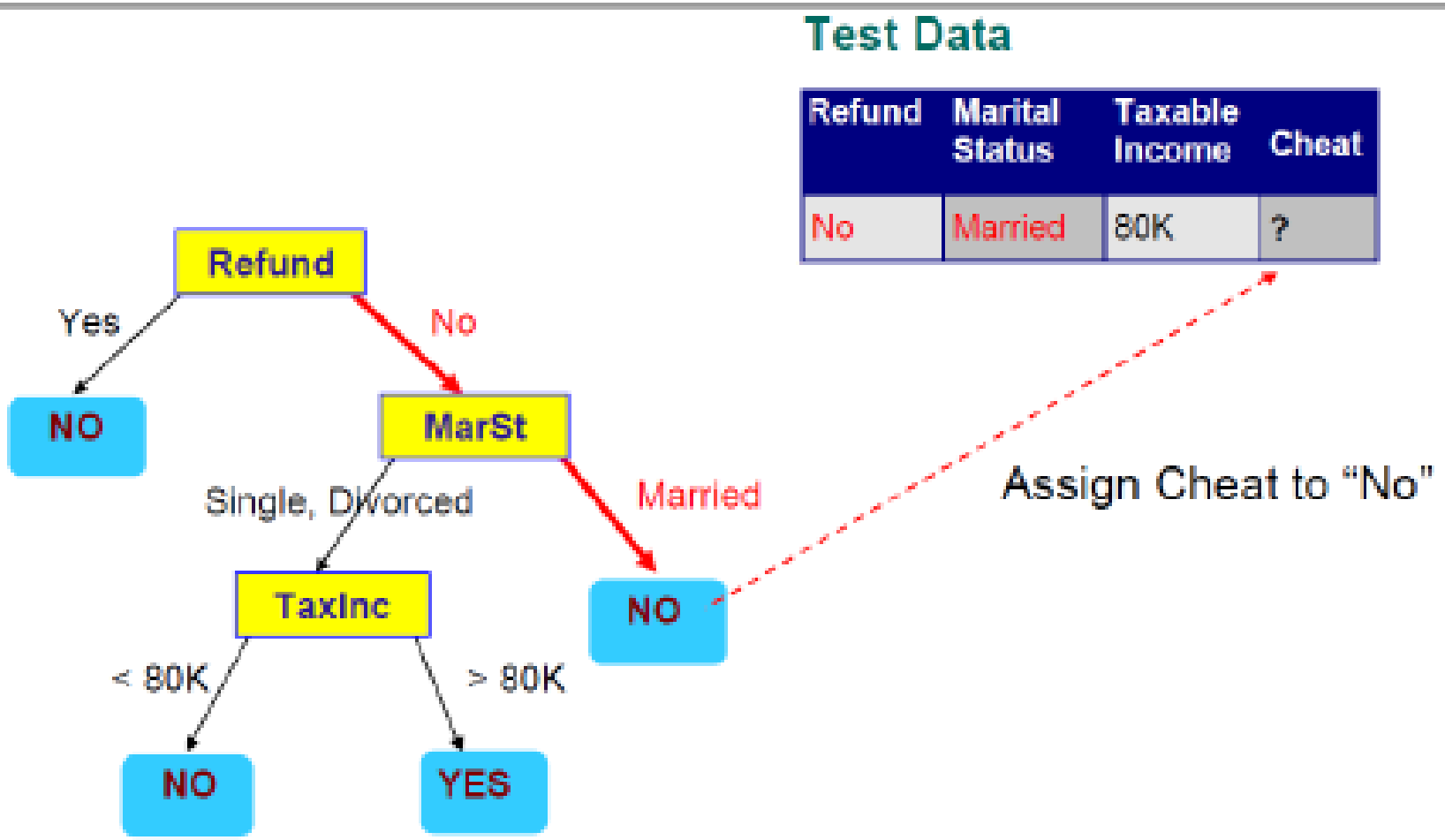
Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Lanj..


Test Data
Refund
Marital Status
Taxable Income
Cheat
No
Married
80K
?
Refund
Yes
No
NO
MarSt
Single, Divorced
Married
TaxInc
NO
< 80K
> 80K
NO
YES
Assign Cheat to "No"

Predictive – Time Series Analysis

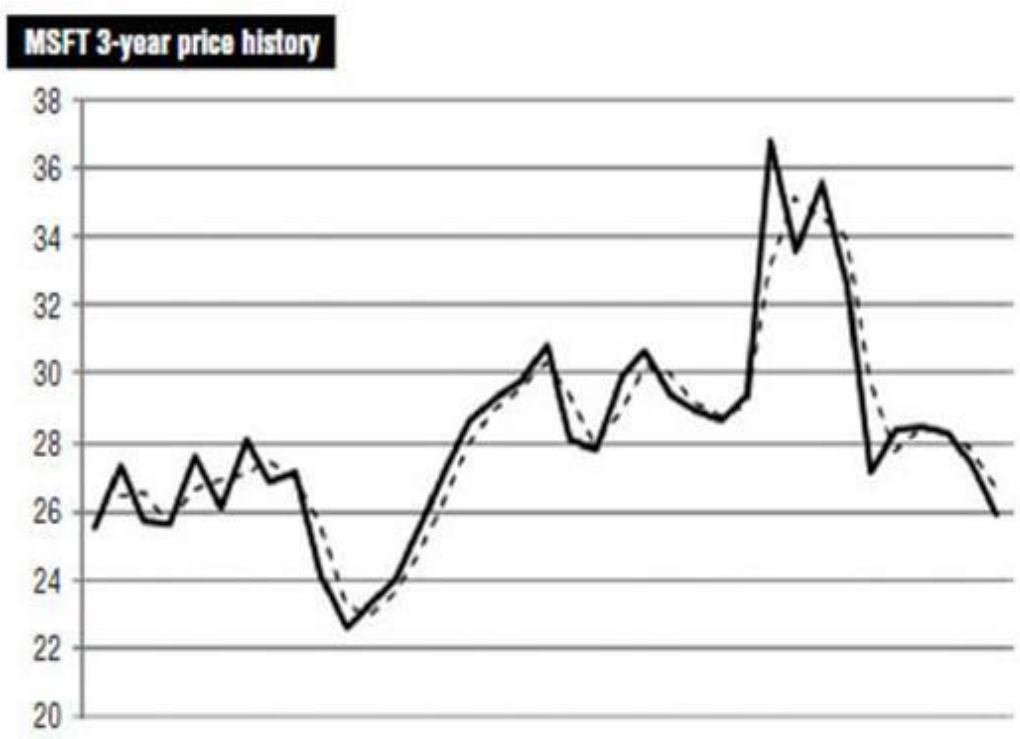
Time series data : sekuens data yang nilainya berubah setiap interval waktu tertentu.

Time series data dapat dipresentasikan dalam bentuk grafik atau kurva yang menunjukkan fungsi sebuah variabel data terhadap satuan waktu.

Metode : Neural Network (model Backpropagation, multi layer perceptron)

Aplikasi : memprediksikan indeks harga saham

Contoh : Prediksi dalam pasar saham



garis yang tegas adalah time-series data sebenarnya dari nilai saham Microsoft, dan garis putus-putus adalah time series model yang memprediksi nilai saham berdasarkan nilai saham pada masa lalu.

Predictive - Regression

Regression vs Classification :

- Regression diterapkan untuk mengklasifikasikan target data numerik
- Classification untuk mengklasifikasikan target data kategorial

Aplikasi : prediksi nilai penjualan yang akan datang berdasarkan trend data penjualan tahun sebelumnya.

Algoritma : Support Vector Machine (SVM), Generalized Linear Model (GLM)

Descriptive - Clustering

Mengidentifikasi kelompok alami dari data berdasarkan kemiripan atribut

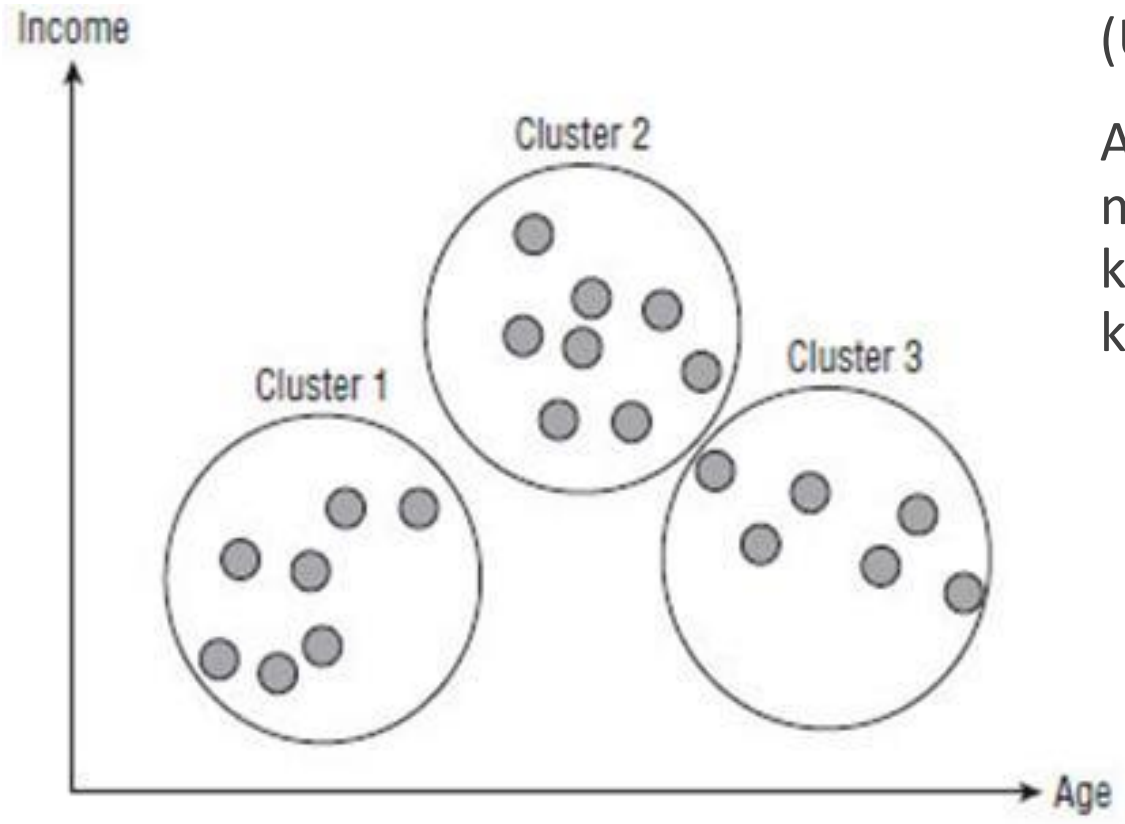
---

Disebut juga Segmentation

Unsupervised Method : tidak satupun atribut yang digunakan untuk memandu proses pembelajaran (tidak ada label)

Algoritma : k-Means, k-Medoid, Fuzzy C-Means, Subtractive, Mountain, Hierarki

Contoh : Data Pelanggan



Terdiri dari dua atribut, yaitu **Age** (Umur) dan **Income** (Pendapatan).

Algoritma Clustering mengelompokkan kelompok data kedalam tiga segment berdasarkan kedua atribut ini.

- Cluster 1 : populasi berusia muda dengan pendapatan rendah
- Cluster 2 : populasi berusia menengah dengan pendapatan yang lebih tinggi
- Cluster 3 : populasi berusia tua dengan pendapatan yang relatif rendah.

Contoh

- Web-Document Clustering:
  - Tujuan: mencari gugus dokumen-dokumen Web yang mirip berdasarkan kemunculan istilah penting
  - Pendekatan: mengidentifikasi istilah yang sering muncul pada setiap dokumen, mengukur kemiripan berdasarkan frekwensi kemunculan istilah pada dokumen lainnya
  - Hasil: Web search engine memunculkan dokumen-dokumen yang mirip (dalam 1 gugus) berdasarkan istilah yang dicari

Lanj..

- Segmentasi Pasar:
  - Tujuan: mencari gugus segmentasi pasar berdasarkan data transaksi untuk keperluan marketing
  - Pendekatan:
    - mempersiapkan data beserta atribut data pelanggan berdasarkan geografi dan data pribadi lainnya
    - mencari gugus pelanggan yang mirip berdasarkan atribut2 tsb
    - melakukan observasi perilaku pasar berdasarkan gugus-gugus pelanggan yang ditemukan
  - Hasil: strategi marketing yang tepat sasaran

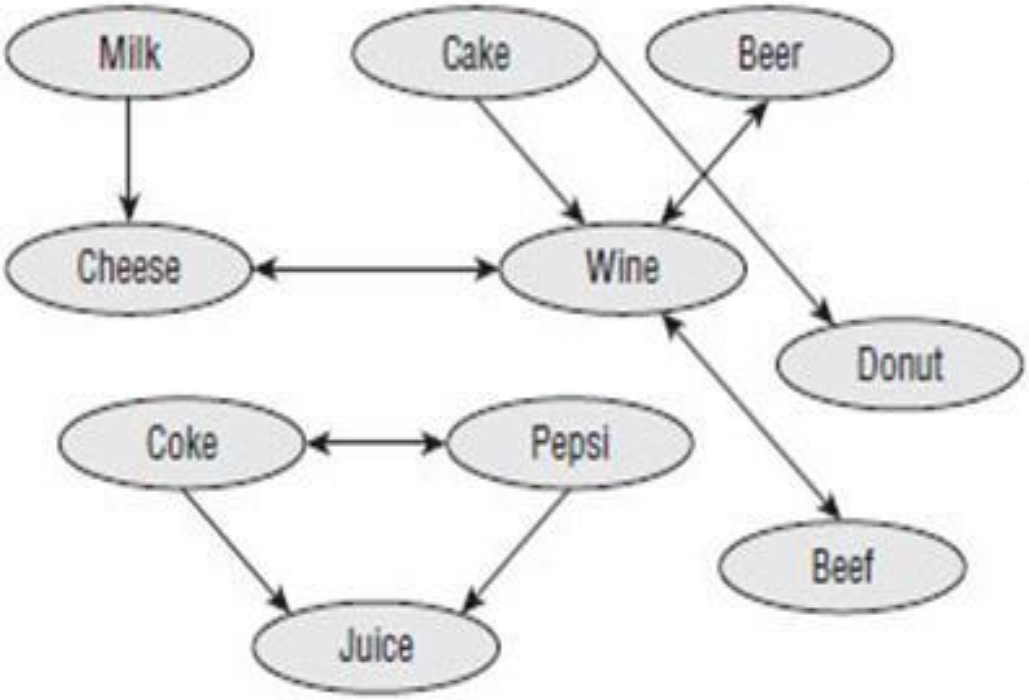
Descriptive – Association Rule

Disebut juga **Market Basket Analysis**.

Menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang seringkali dibeli bersamaan oleh customer

Untuk mengidentifikasi kelompok kesamaan dari produk dan kebiasaan apa yang terjadi guna kepentingan *cross-selling*

- Untuk mencari produk apa yang biasanya terjual bersamaan
- Untuk mencari tahu apa aturan yang menyebabkan kesamaan tersebut.



Ketika orang membeli susu, dia biasanya membeli keju

Ketika orang membeli pepsi atau coke, biasanya dia membeli juice

Contoh Lain

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

- Marketing & Sales Promotion
  - Misalnya pola yang ditemukan :
    $\{\text{Susu Anak}, \ldots\} \rightarrow \{\text{Kwaci}\}$
  - Kwaci sebagai konsekuen : bagaimana caranya menaikkan penjualan kwaci
  - Susu Anak sebagai anteseden : jika tidak lagi menjual susu anak, memprediksi produk lain yang ikut jatuh penjualannya
  - Dua-duanya : membuat paket promo Susu Anak, Kwaci, dll

- Pengelolaan Rak di Supermarket
  - Tujuan: memudahkan pelanggan berbelanja barang-barang yang sering dibeli bersama
  - Misalnya: ada rak kecil berisi kwaci diletakkan pada bagian susu anak
- Sistem Rekomendasi Pintar
  - Tujuan: memberikan rekomendasi kepada pelanggan toko buku on-line tentang buku-buku lain yang sering dibeli juga oleh pelanggan lainnya jika membeli buku tertentu

Descriptive – Sequence Analysis

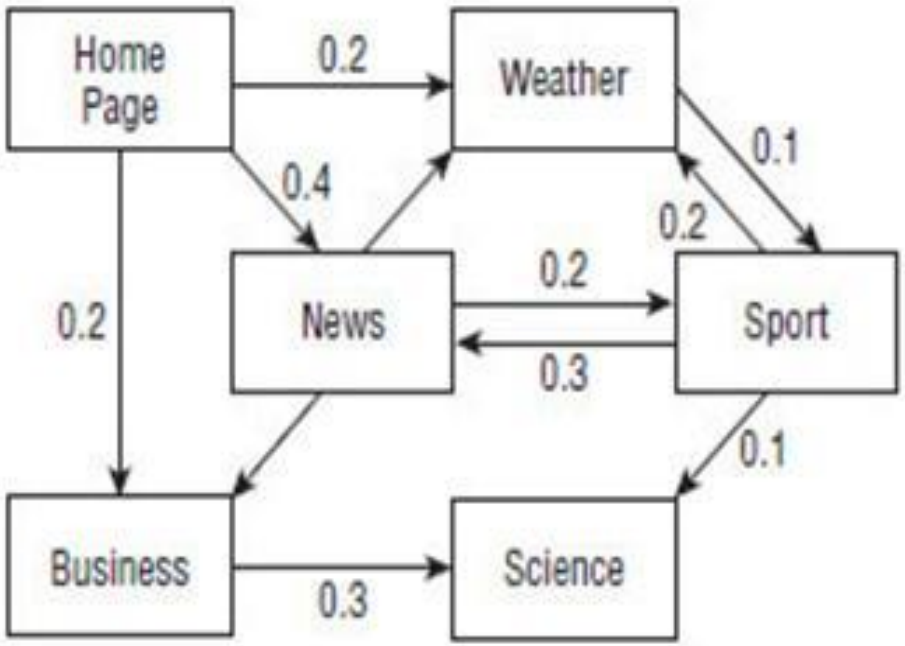
Digunakan untuk mencari pola pada serangkaian kejadian yang disebut dengan Sequence.

Contoh rangkaian klik pada sebuah website berisi rentetan URL.

Implementasi : memodelkan pembelian oleh pelanggan sebagai sebuah sequence (rangkaian) data :

- Pertama-tama seorang pelanggan membeli komputer
- kemudian membeli speaker
- dan akhirnya membeli sebuah webcam.

Contoh : Rangkaian Klik pada Sebuah Website Berita



Setiap node adalah sebuah kategori URL.

Garis melambangkan transisi antar kategori URL tersebut.

Setiap transisi dikelompokan dengan sebuah bobot yang menggambarkan kemungkinan transisi antara satu URL dan URL yang lain.

Penerapan Data Mining

- Clustering
- Tingkat kelulusan


thank
you!