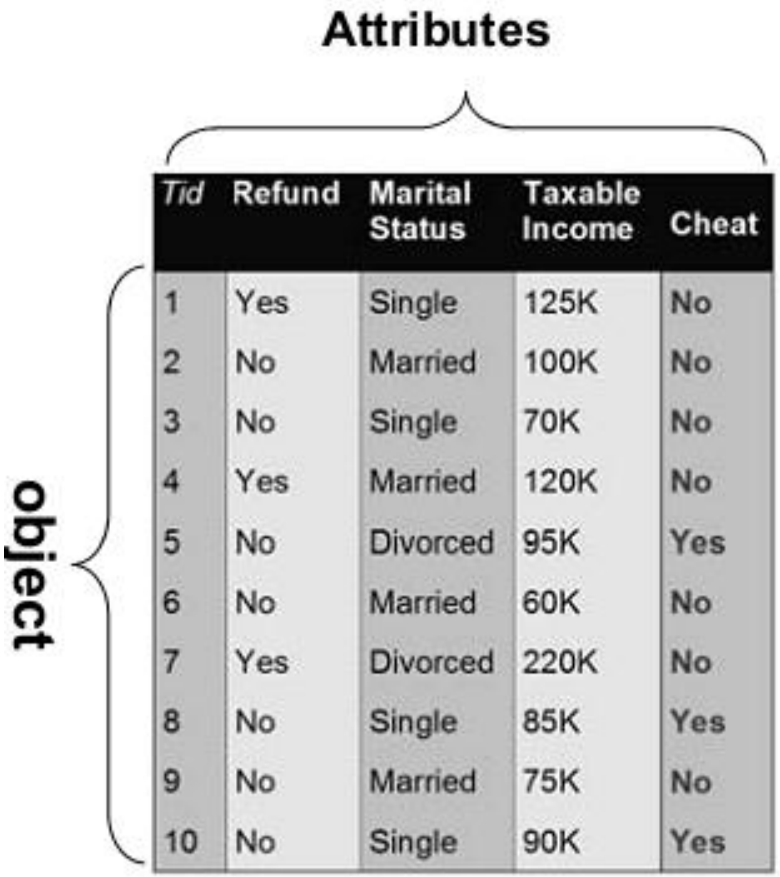
DATA

What is Data?

- Kumpulan data objek dan atributnya
- Karakter/properti dari sebuah objek
  - Juga dikenali sebagai variabel, karakteristik, feature, atau dimensi
- Sekumpulan atribut mendeskripsikan sebuah objek
  - Juga dikenal sebagai record, point, case, sampel, entitas, atau instance



Tid	Refund	Marital Status	Taxable Income	Cheat
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

NILAI ATRIBUT

- Atribut
  - Merupakan property atau karakteristik objek
  - Memiliki sifat berbeda antara satu objek dengan objek lain, atau satu waktu dengan waktu lain
- Nilai Atribut
  - Skala pengukuran yang berasosiasi dengan nilai numerik atau simbol dari atribut suatu objek
  - Merupakan bilangan atau simbol yang menggambarkan atribut

- Perbedaan antara atribut dan nilai atribut:
  - 1 atribut yang sama dapat di-map-kan ke dalam nilai atribut yang berbeda

Value	Measurement(s)	
Height	Meter(s)	Feet(s)

  - 2 atribut berbeda dapat memiliki nilai atribut yang sama

Value	Measurement(s)
ID	Integer
Age	Integer

TIPE ATRIBUT

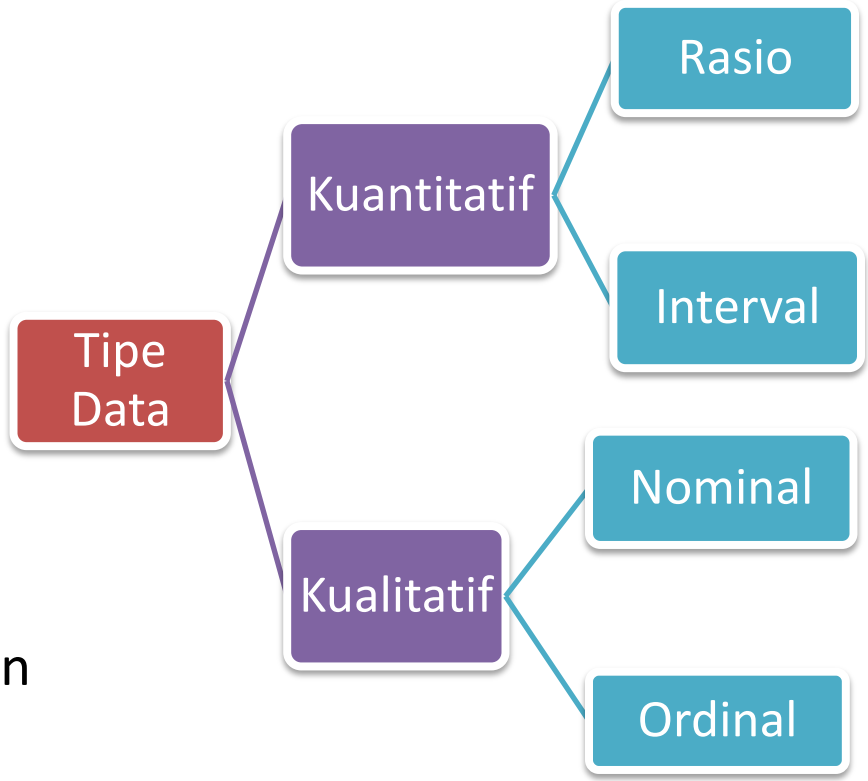
Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

PROPERTI DARI NILAI ATRIBUT

- Tipe atribut tergantung dari properti yang mengikuti:
  - Persamaan = ≠
  - Urutan < >
  - Penambahan + -
  - Perkalian * /
- Properti untuk masing2 atribut berbeda tergantung tipe atributnya:
  - Nominal: persamaan
  - Ordinal: persamaan & urutan
  - Interval: persamaan, urutan, & penambahan
  - Rasio: persamaan, urutan, penambahan, & perkalian

TIPE DATA

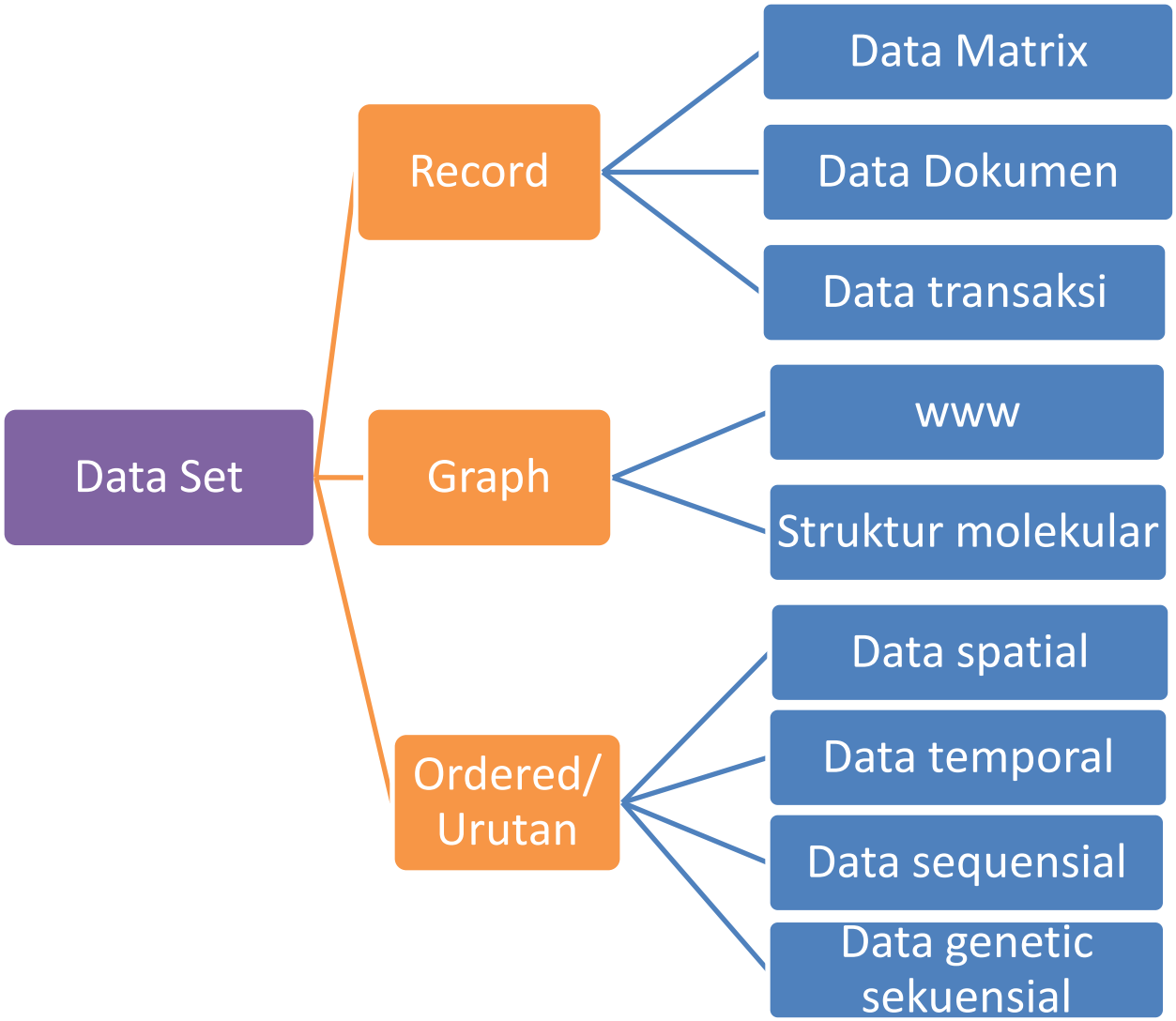
- Kualitas data:
  - Noise and outlier
  - Inconsistent
  - Duplicate
  - Biased
  - unrepresentative
- Preprocessing:
  - data harus diimprove atau dimodifikasi agar sesuai dengan teknik data mining atau tool tertentu
- Analisis relationship antara 2 object



ATRIBUT DISKRIT DAN CONTINUOUS

- Atribut diskrit
  - Memiliki nilai terbatas
  - Seringkali direpresentasikan sebagai bil. Integer
  - Special case: bil. Biner
  - Contoh: kode pos, jumlah kata di dalam dok., dll
- Atribut continuous
  - Memiliki nilai pasti yang dapat dihitung menggunakan jumlah digit yang terbatas
  - Biasanya direpresentasikan dalam bentuk var. float dengan nilai desimal
  - Contoh: temperatur, tinggi badan, BB

TIPE-TIPE DATA SET



KARAKTERISTIK DATA TERSTRUKTUR

- Dimensionality
  - Jumlah atribut yang dimiliki oleh object
  - Masalah: sulitnya menganalisis data berdimensi besar
- Sparsity
  - Mayoritas atribut memiliki nilai 0
- Resolusi
  - Pola tergantung dengan skala
  - Jika resolusi terlalu bagus, pola tidak dapat diamati atau mungkin dapat terkubur bersama noise
  - Jika resolusi terlalu kasar pola kemungkinan tidak terlihat

DATA RECORD

- Terdiri dari sekumpulan record yang terdiri dari sekumpulan atribut
- Tidak ada hubungan eksplisit antara record atau data field
- Perbedaan tipe data record
  - Data transaksi
  - Data matrix
  - Data matrix tersebar

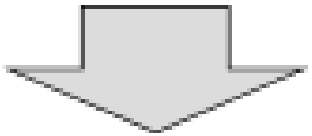
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

DATA TRANSAKSI

- Tipe khusus data record
- Masing-masing record transaksi melibatkan sekumpulan item
- Dapat dilihat sebagai sekumpulan record dengan atribut asymmetric → record data transaksi berhubungan dengan item penjualan
- Atribut dapat berupa data diskret atau continuous
- Contoh:

CONTOH DATA TRANSAKSI

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



TID	Bread	Coke	Milk	Beer	Diaper	…
1	1	1	1	0	0	
2	1	0	0	1	0	
3	0	1	1	1	1	
4	1	0	1	1	1	
5	0	1	1	0	1	

DATA MATRIX

- Jika data objek memiliki kesamaan atribut numerik maka data object dapat dianggap sebagai point dalam ruang multidimensi dimana masing-masing dimensi merepresentasikan atribut yang berbeda
- Beberapa data set direpresentasikan dalam bentuk matrix m x n, dimana m = baris dan n = kolom dari masing-masing atribut
- Operasi matrix standar dapat diaplikasikan untuk mentransformasikan dan memanipulasi data

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

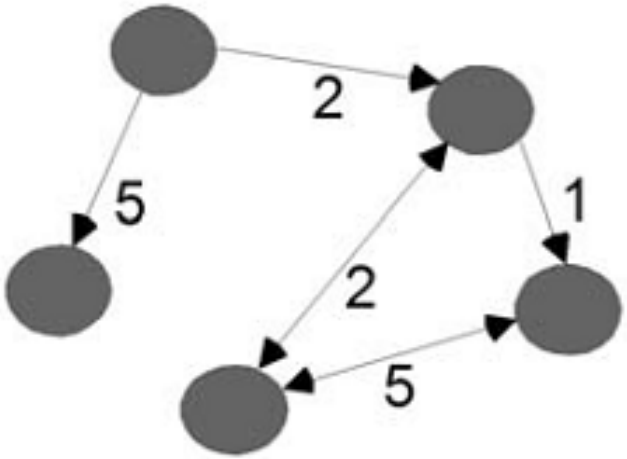
DATA MATRIX TERSEBAR

- Data matrix tersebar merupakan kasus khusus dari data matrix dimana hanya atribut yang bertipe sama dan bernilai≠0 yang penting
- Masing-masing dokumen merupakan vektor “term”
  - Masing-masing term merupakan komponen (atribut) vektor
  - Nilai dari masing-maisng komponen merupakan jumlah kemunculan di dokumen

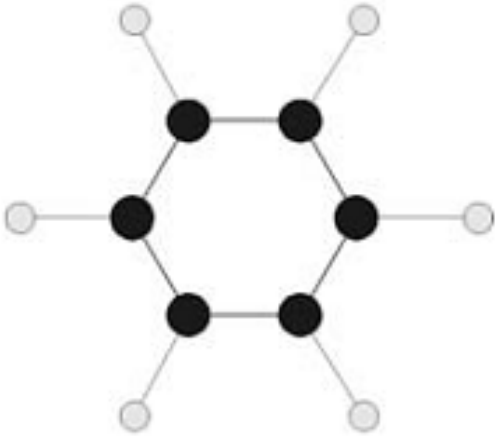
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

DATA BERBASIS GRAPH

- Data yang objeknya saling berhubungan
  - Contoh: graph generic
  - Link HTML

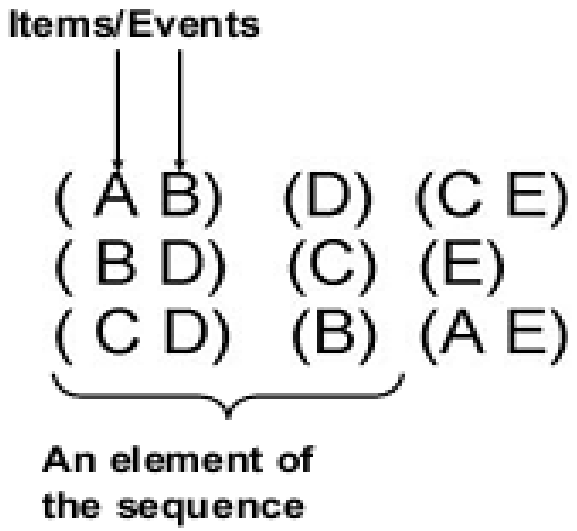


- Data yang objeknya berbentuk graph
  - Contoh: molekul benzena



DATA TERURUT

- Data sekuensial (temporal)
  - Masing-masing record memiliki asosiasi waktu
  - Contoh: transaksi sekuensial
- Data sekuensial
  - Mirip dengan sekuensial (temporal) hanya saja tidak ada batasan waktu
- Data time series
  - Tipe khusus dari data sekuensial dimana masing-masing record memiliki rangkaian waktu khusus
  - Contoh: data finansial, rata2 data temperatur bulanan
  - Metode: Moving Average
- Data spatial dengan atribut spasial
  - Contoh: posisi; area
- Spasial dengan autokorelasi
- Data spatio-temporal
  - Rata2 temperatur daratan dan lautan bulanan



KUALITAS DATA

- Masalah deteksi dan koreksi dari kualitas data (data cleaning)
- Penggunaan algoritma dapat mentoleransi kualitas data buruk
- Pengukuran kesalahan
  - Noise
  - Artifact
  - Bias
  - Presisi
  - Akurasi
- Masalah pengukuran dan pengumpulan data
  - Outlier
  - Missing dan inkonsisten
  - Duplikasi data

KESALAHAN PENGUKURAN DAN PENGUMPULAN DATA

- Error
  - Perbedaan numerik dari nilai yang terukur dan nilai secara nyata
- Kesalahan pengukuran
  - Masalah yang dihasilkan dari kesalahan proses pengukuran
- Kesalahan pengumpulan data
  - Mengacu pada kesalahan seperti penghilangan data object atau nilai atribut
- Kedua kesalahan bisa jadi merupakan kesalahan sistematik atau random

NOISE DAN ARTIFACTS

- Noise merupakan komponen acak dari pengukuran kesalahan
  - Noise mengacu pada modifikasi nilai asli
  - Contoh: distorsi suara user ketika berbicara menggunakan pesawat telepon yang buruk; bintik2 di pesawat televisi
  - Seringkali berhubungan dengan koneksi data yang memiliki komponen spatial atau temporal
  - Teknik dari sinyal atau image processing dapat digunakan untuk mengurangi noise
  - Mengingat sulitnya mengurangi noise, pengembangan algoritma yang robust dapat menghasilkan hasil yang dapat diterima dengan jumlah noise yang lebih feasible
- Kesalahan dapat terjadi sebagai akibat dari fenomena deterministik yang mengacu pada artifact

PRESISI, BIAS, DAN AKURASI

- Presisi
  - Kedekatan pengukuran yang berulang dari jumlah yang sama satu sama lain
  - Diukur menggunakan std. dev. dari sekumpulan nilai
- Bias
  - Variasi sistematik dari pengukuran terhadap kuantitas yang diukur.
  - Diukur dengan mencari perbedaan antara mean dan nilai quantity yang diukur
- Akurasi (vs. error)
  - Kedekatan pengukuran terhadap nilai kebenaran suatu kuantitas yang diukur
  - Akurasi tergantung pada presisi dan bias
  - Digit yang signifikan penting untuk akurasi

OUTLIER (PENCILAN)

- Merupakan data objek yang dianggap berbeda dari kebanyakan data objek di dalam sekumpulan data set
- Memiliki nilai atribut berbeda dari kebiasaan nilai atribut yang lain
- Anomali objek atau nilai
- Outlier mungkin diharapkan untuk mendeteksi terjadinya pencurian atau kemungkinan terjadinya penyusupan

MISSING VALUES

- Alasan missing values
  - Informasi tidak lengkap
  - Atribut tidak dapat diaplikasikan terhadap semua kasus
- Cara mengatasi masalah missing values
  - Eliminasi data objek
  - Eliminasi missing values
  - Abaikan missing values selama analisis
  - Gantikan dengan semua nilai yang mungkin yang dibobotkan menggunakan probabilitas

NILAI INKONSISTEN DAN DUPLIKASI DATA

- Sangat penting untuk mendeteksi dan membetulkan masalah inkonsistensi data
  - Contoh:
    - Kode pos dan kota tidak konsisten
- Data set bisa berisi data objek ganda, atau hampir sama satu sama lain
  - Isu utama ketika menggabungkan data dari sumber yang berbeda
  - Contoh:
    - 1 orang memiliki beberapa alamat email
- Data cleaning
  - Merupakan proses untuk menyelesaikan masalah duplikasi data