

### What is Data?

Collection of data objects and their attributes.

Attribute is a property of an object

Variable, field, characteristic, **feature**

Collection of attributes describes an object

Record, point, case, sample, entity, instance, **observation**

### Type of attributes

Discrete

Finite (countably)

Integer

Zip, Counts

Continuous

Real numbers

Floating points

Temp., height, weight

### Hierarchy of attributes types

Qualitative

Nominal

Category (=, !=)

ID, zip, eye, color

Ordinal

Ranked (>, <)

Grades, {low, med., high}

Quantitative

Interval

Distance (+, -)

Dates, temp (C/F)

Ratio

Zero means absence

Length, time, temp(K) (\*, /)

### Type of data sets

Record

Collection of dataobjects and their attributes

Table

Relational

Collection of data objects and their relation

Graph

Ordered

Ordered collection of data objects

Sequence

### Data quality

High quality

Are fit for their intended use

Correctly represent the phenomena they correspond to

Problems

Noise

Outliers

Missing values

### Noise

Definition

Unwanted perturbation to a signal

Unwanted data

Reasons

Limits in measurement accuracy

Interference from other signals

Measurement of attributes not related to the data modeling task

Handling

Exclude noisy attributes

Remove noise by filtering

Include a model of noise

### Outliers

Definition

Data objects which are significantly different from most others

Reasons

Measurement errors

Natural property of data

Handling

Identify & exclude outliers

Model the outliers

### Missing values

Definition

No value is stored for an attribute in a data object

Reasons

Information is not collected

People decline to give their age

Attribute is not applicable

Annual income is not applicable to children

Handling

Eliminate data objects

Estimate missing values

e.g. average

Ignore the missing value in analysis

