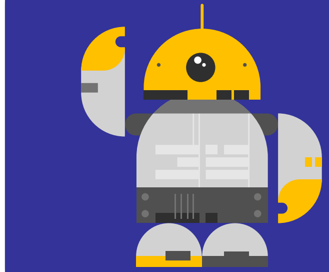

Database Systems

Lecture 3 Cont. - Getting to Know Your Data

Dr. T. Akilan
takilan@lakeheadu.ca

This Section

- Data:
    - Different types of attributes
    - Basic statistical analysis of attribute values
    - Graphical representations of data and attributes
    - Similarity and dissimilarity of data objects

What is Data?

- A data object represents an entity and contains information about it
  - “Data object” also called “sample”, “example”, “instance”, “data point”, “tuple”
  - **Entity** represented is **application-specific**
    - ✓ Customers, patients, employees, merchandise, websites, customer reviews, text documents, …
  - **Information** contained is all relevant **facts known about** the **entities**
    - ✓ Stored in **attributes** of the object

What is an Attribute?

- A data **field**, **representing information** (a characteristic or feature) of data objects
  - Also called "dimension", "feature", "variable"
  - Has a value
  - Allows us to perform some **statistical analysis**, to **model** and **understand** the data
    - ✓**Typical analysis**: mean & deviation, median, mode (most frequent value)

Data object	Attributes
Person	Hair colour, gender, income bracket, body temperature…
Car	Maker, year, used car, km on odometer, CAA ranking, …
Text document	Word count, topic, spell-checked, difficulty level, date written…
Customer review	Item reviewed, peer popularity, would recommend, date joined, time since joining, …

Attribute Types

- Nominal
- Binary
- Ordinal
- Numeric:
  - Interval-scaled numeric
  - Ratio-scaled numeric

Attribute Types - Nominal

- Nominal (categorical)
  - Name of thing
  - Descriptive (qualitative) attribute with no inherent quantitative value
    - ✓ We cannot compute mean or median
    - ✓ Mode still makes sense (we will study this later)

Data object	Attributes	Possible values
Person	Hair colour	Black, brown, blond, red, white, …
Car	Maker	Toyota, VW, Acura, GM, …
Text document	Topic	Politics, science, news, fiction, …
Customer review	Item reviewed	Movie, song, toy, car, …

Attribute Types - Binary

- Binary (Boolean)
  - Descriptive (qualitative) attribute
    - ✓ Statistical analysis: mode
  - **Symmetric**: both states are equivalent
    - ✓ No preference between them
    - ✓ Either one could be 1 or 0
  - **Asymmetric**: one state is more important than the other
    - ✓ Most important/rarest is 1, least important/most common is 0

Data object	Attribute	Possible values
Person	Gender	Male, female
Car	Used car	True, false
Text document	Spell-checked	Yes, no
Customer review	Would recommend	0,1

Attribute Types - Ordinal

- Ordinal
  - Relative order/ranking of objects
  - Descriptive (qualitative) attribute
    - ✓ Ordered category names
    - ✓ Numeric value broken up in intervals
    - ✓ Integer value (aka 2nd, 3 stars, …) is category number, not a measurable quantity of the object
  - **Statistical analysis**: median, mode

Data object	Attribute	Possible values
Person	Income bracket	[0-20,000],[20,001-40,000],…
Car	CAA ranking	1st, 2nd, 3rd, 4th, …
Text document	Difficulty level	Easy, Average, Hard
Customer review	Peer popularity	*, **, ***, ****, *****

Attribute Types - Numeric Attributes

- **Interval-scaled numeric**
  - Quantitative value in a scale **without an absolute zero**
    - ✓ Dates: year 0 is not the beginning of time (in fact it doesn't exist)
    - ✓ Celsius: 0 degrees is not absence of temperature
  - **Ratios** between values **are meaningless**
    - ✓ Year 2000 is not two times year 1000
  - But they are quantities that we can measure, and we can know exactly the difference between measures
    - ✓ We can measure what year it is
    - ✓ We can measure that 2010 is 8 years after 2002

Data object	Attribute	Scale
Person	Body temperature	Celsius
Car	Year	Time
Text document	Date written	Time
Customer review	Date joined	Time

Attribute Types - Numeric Attributes Cont.

- **Ratio-scaled numeric**
  - Quantitative value in a scale **with an absolute zero**
  - Ratios between values are meaningful
    - ✓ 20 Celsius is not 2 × 10 Celsius
    - ✓ 20 Kelvin is 2 × 10 Kelvin
  - Statistical analysis
    - ✓ Mean, median, mode

Data object	Attribute	Scale
Person	Body temperature	Kelvin
Car	km on odometer	Real
Text document	Word count	Integers
Customer review	Time since joining	Time

Attribute Types Cont.

- Discrete attributes
    - ○ Set of possible values is finite
        - ✓ Nominal, binary, ordinal attributes
    - ○ or countably infinite
        - ✓ Numeric attributes on integer/natural numbers scale

- Continuous attributes
    - ○ Set of possible values is uncountably infinite
        - ✓ Numeric attributes on float/real scale

Next → Statistical Analysis