Queen Mary
University of London

ECS766 Data Mining

Week 2: Data

Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

October 2021

School of EECS, Queen Mary University of London

Last week: Introduction

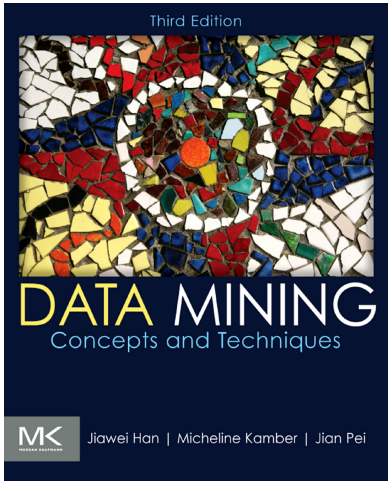
- Preliminaries
- Why data mining?
- What is data mining?
- Models in data science
- Data mining tasks
- Challenges in data mining

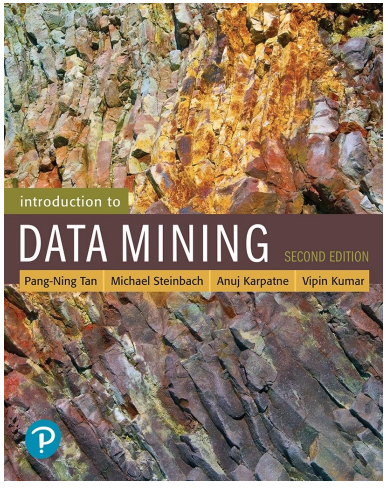
This week's contents

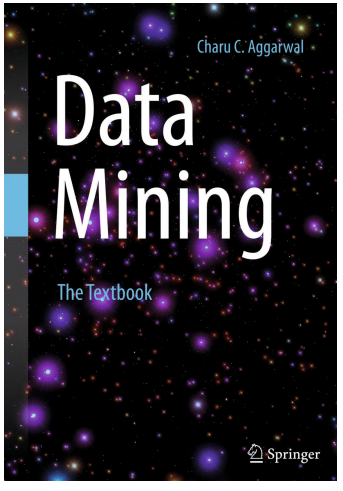
1. Attributes and Objects
2. Characteristics of Data
3. Types of Data
4. Data Quality
5. Basic Statistical Descriptions of Data
6. Similarity and Distance

Reading

- **Chapter 2 of J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", 3rd edition, Elsevier/Morgan Kaufmann, 2012**
- Chapter 2 of P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to Data Mining", 2nd edition, Pearson, 2019
- Chapter 3 of C. C. Aggarwal, "Data Mining: The Textbook", Springer, 2015





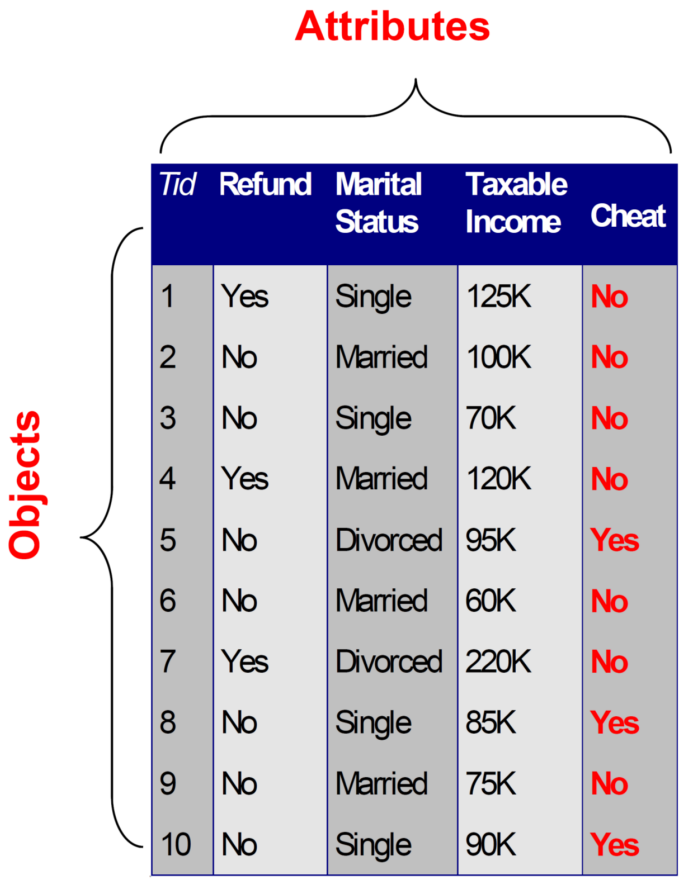


Attributes and Objects

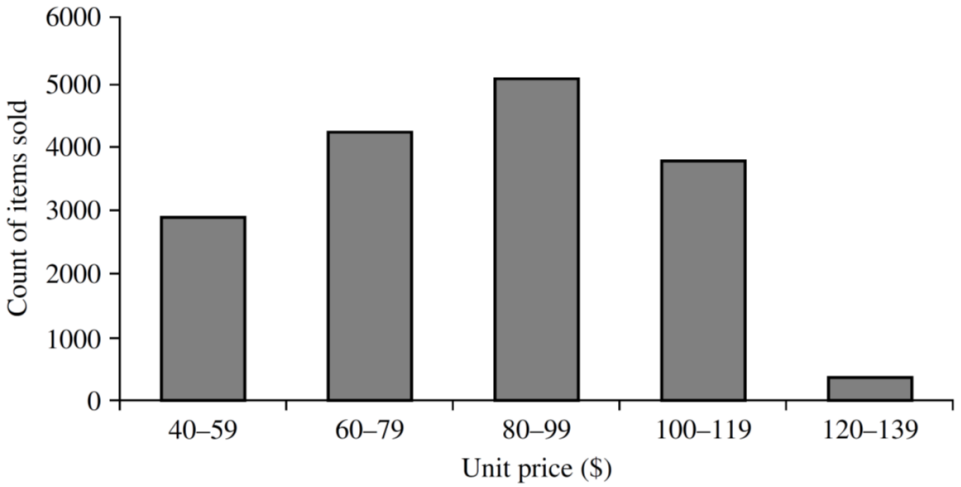
Getting to Know Your Data

A data set can be viewed as a collection of data objects. Other names for a data object are *record*, *data point*, *vector*, *pattern*, *event*, *case*, *sample*, *example*, *instance*, *observation*, and *entity*.

Data objects are described by a number of attributes that capture the basics characteristics of an object. Other names for an attribute are *variable*, *characteristic*, *field*, *feature*, and *dimension*.



Getting to Know Your Data



- Data may have parts
- Attributes (objects) may have relationships with other attributes (objects)
- Data may have structure
- Data can be incomplete

What Is an Object?

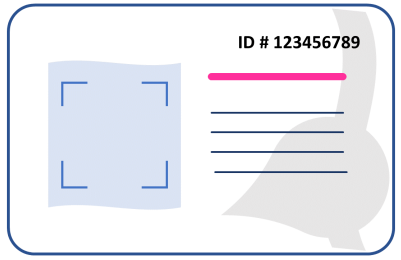
A data object represents an entity. If the data objects are stored in a database, they are *data tuples*: rows of a database correspond to data objects, and columns correspond to attributes.

- e.g. in a medical database, the objects can be patients
- e.g. in a university database, the objects can be students, professors, and courses
- e.g. in a sales database, the objects may be customers, store items, and sales

What Is an Attribute?

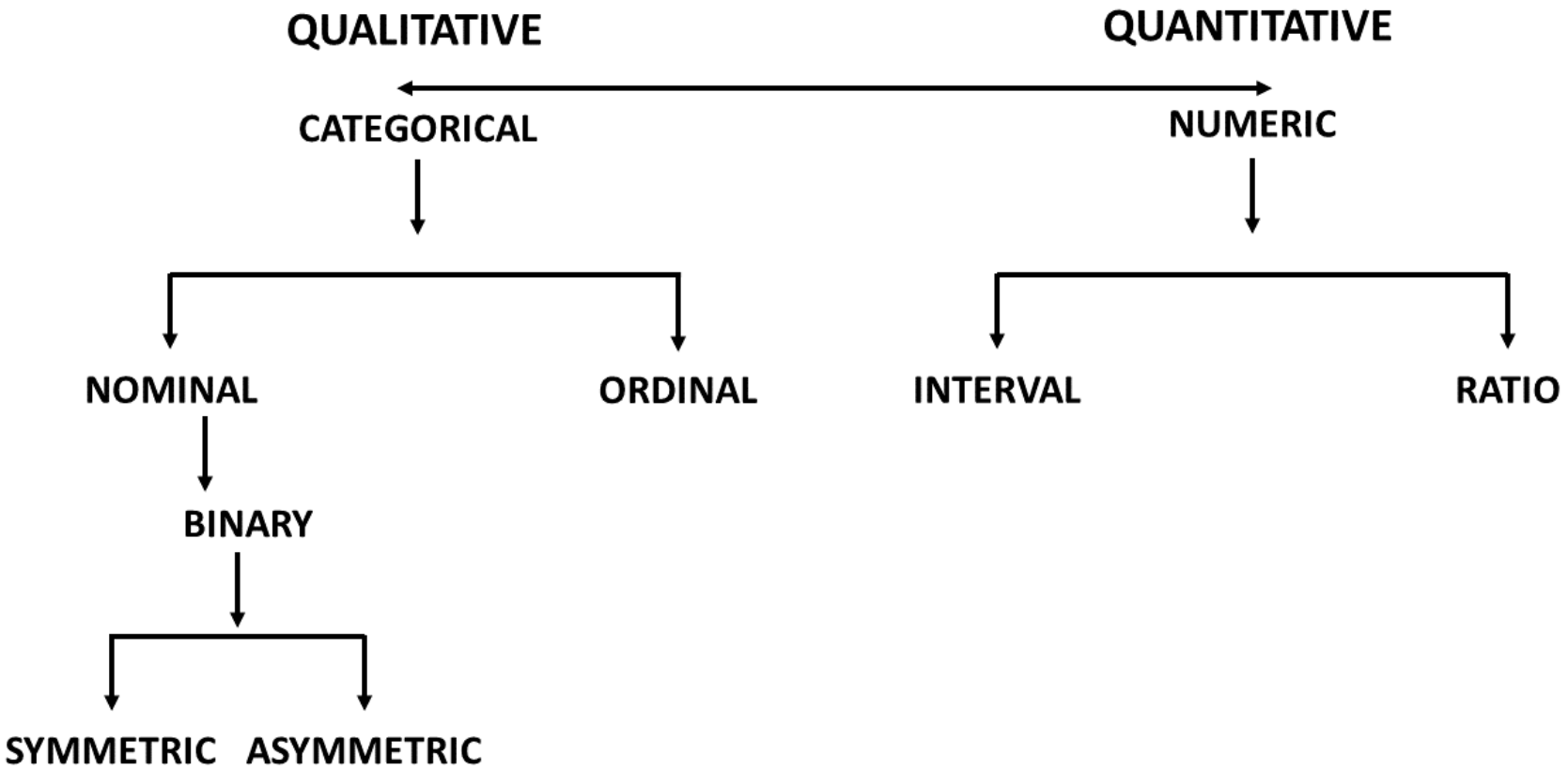
An attribute is a property, characteristic, or feature of an object that may vary, either from one object to another or from one time to another.

- e.g. attributes describing a customer object can include *customer ID*, *name*, and *address*



Observed values for a given attribute are known as *observations*. A set of attributes used to describe a given object is called an *attribute vector*.

What Is an Attribute?



Nominal Attributes

Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things, which do not have a meaningful order. Each value represents some kind of category, code, or state.

Although nominal attributes are not quantitative, it is possible to represent such symbols or “names” with numbers.

However, mathematical operations on values of nominal attributes are not meaningful. For example, it makes no sense to subtract one customer ID number from another.

Binary Attributes

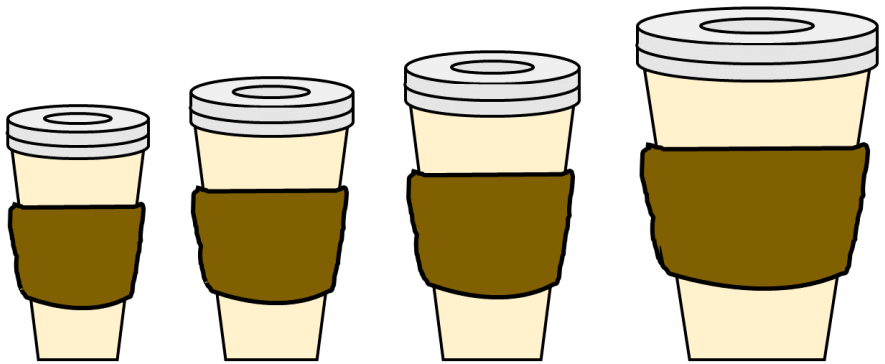
A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as **Boolean** if the two states correspond to *true* and *false*.

The attribute *role* with the states *student* and *staff* is binary and **symmetric** because its states are equally important and carry the same weight.

The attribute *outcome* of a medical test with the states *positive* and *negative* is binary and **asymmetric** because the outcomes of the states are not equally important.

Ordinal Attributes

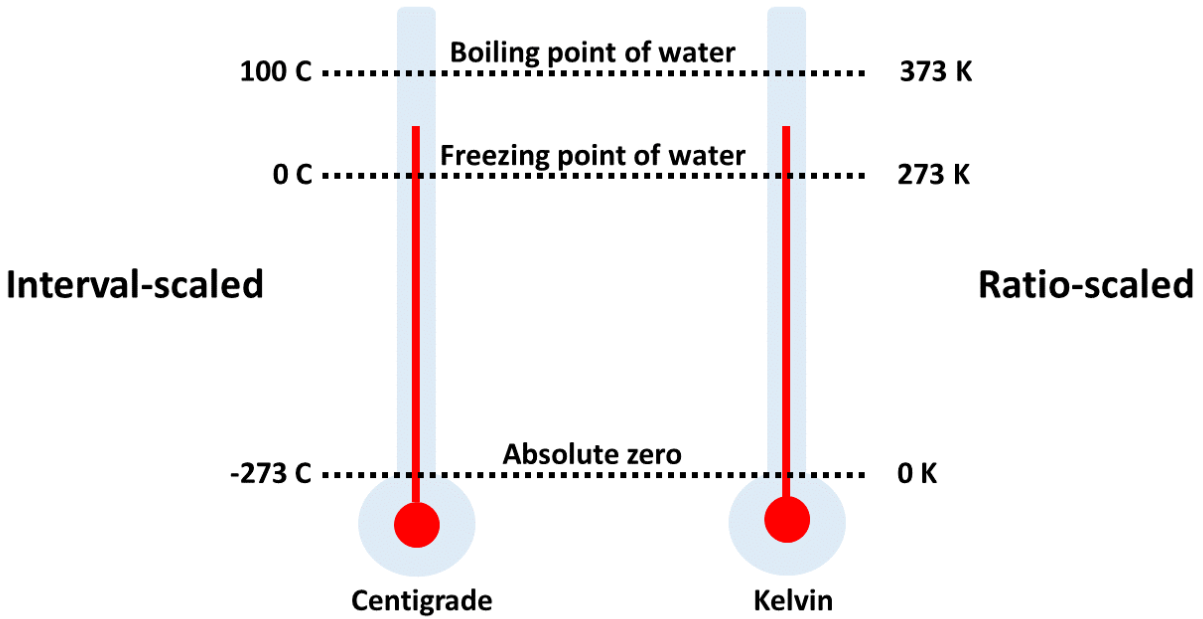
An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.



Numeric Attributes

A numeric attribute is quantitative and is represented in integer or real values. Numeric attributes can be:

- Interval-scaled attributes are measured on a scale of equal-size units
- Ratio-scaled attributes are numeric attributes with an inherent zero-point



Discrete versus Continuous Attributes

A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers. Such attributes can be categorical or numeric. Binary attributes are discrete.

A continuous attribute is one whose values are real numbers. Continuous attributes are typically represented as floating-point variables. Practically, real values can only be measured and represented with limited precision.

Asymmetric Attributes

An asymmetric attribute is an attribute that regards only presence (a non-zero attribute value) as important. We discussed already asymmetric binary attributes. Asymmetric attributes can be discrete and continuous as often used in association analysis.

- e.g. words present in documents
- e.g. items present in customer transactions

It is possible to have discrete or continuous asymmetric features. For instance, if the number of credits associated with each course is recorded, then the resulting data set will consist of asymmetric discrete or continuous attributes.

Characteristics of Data

Characteristics of Data

As we will discuss later, there are many types of data sets. However, few characteristics apply to many data sets and have a significant impact on the data mining techniques that are used:

- Dimensionality: high dimensional data brings a number of challenges
- Sparsity: only presence counts
- Resolution: patterns depend on the scale
- Size: type of analysis may depend on size of data

Dimensionality

The dimensionality of a data set is the number of attributes that the objects in the data set possess.

The difficulties associated with analyzing high-dimensional data are sometimes referred to as the *curse of dimensionality*. Because of this, an important motivation in preprocessing the data is *dimensionality reduction*.

Sparsity

The sparsity of a data set refers to any data set with asymmetric features which has a very large group of zero values and very little non-zero values.

In practical terms, sparsity is an advantage because usually only the non-zero values need to be stored and manipulated. This results in significant savings with respect to computation time and storage. Furthermore, some data mining algorithms work well only for sparse data.

Object/Attribute	A1	A2	A3	A4
O1	0	1	0	0
O2	0	0	2	0
O3	0	0	0	1
O4	3	0	0	0
O5	0	0	0	2

Resolution

Data sets can have different level of resolutions and the data attributes can be different at different resolutions.

For instance, the surface of the Earth seems very uneven at a resolution of a few meters, but is relatively smooth at a resolution of tens of kilometers. The patterns in the data depend on the level of resolution. If the resolution is too fine, a pattern may not be visible or may be buried in noise; if the resolution is too coarse, the pattern may disappear. For example, weather predictions in the scale of hours vs months.

Size

Data sets can have different sizes.

Due to the increasing sizes of the data in modern-day applications, *scalability* is an important concern in many data mining applications. There are two important scenarios for scalability:

- The data is stored on one or more machines, but it is too large to process efficiently.
- The data is generated continuously over time in high volume, and it is not practical to store it entirely. This scenario is that of *data streams*, in which the data need to be processed with the use of an online approach.

Types of Data

Types of Data

- **Record**: Data set that consists of a collection of records (data objects), each of which consists of a fixed set of data fields (attributes)
- **Graph**: A graph can capture relationships among data objects and/or data objects can be represented as graphs
- **Ordered**: The attributes of a data set have relationships that involve order in time or space

Record: Flat Files

For the most basic form of record data, there is no explicit relationship among records or data fields, and every record (object) has the same set of attributes. Record data is usually stored in flat files:

- Data in a plain text format

TID	Refund	Marital Status	Income
1	Yes	Single	125K
2	No	Married	100K
3	No	Single	70K
4	Yes	Married	95K
5	No	Divorced	120K

Record: Transaction Data

Transaction Data is a special type of record data, where each transaction (record) involves a set of items.

- e.g. a set of products purchased by a customer during one shopping trip constitutes a transaction, while the individual products that were purchased are the items

TID	Items
1	Bread, eggs, milk
2	Juice, bread
3	Juice, eggs, butter, milk
4	Juice, bread, butter, milk
5	Eggs, butter, milk

Record: Data Matrix

If the data objects in a collection of data all have the same fixed set of numeric attributes, then the data objects can be thought of as points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object:

- A data set represented by an $m \times n$ matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

A sparse data matrix is a special case of a data matrix in which the attributes are of the same type and are asymmetric, i.e. only non-zero values are important.

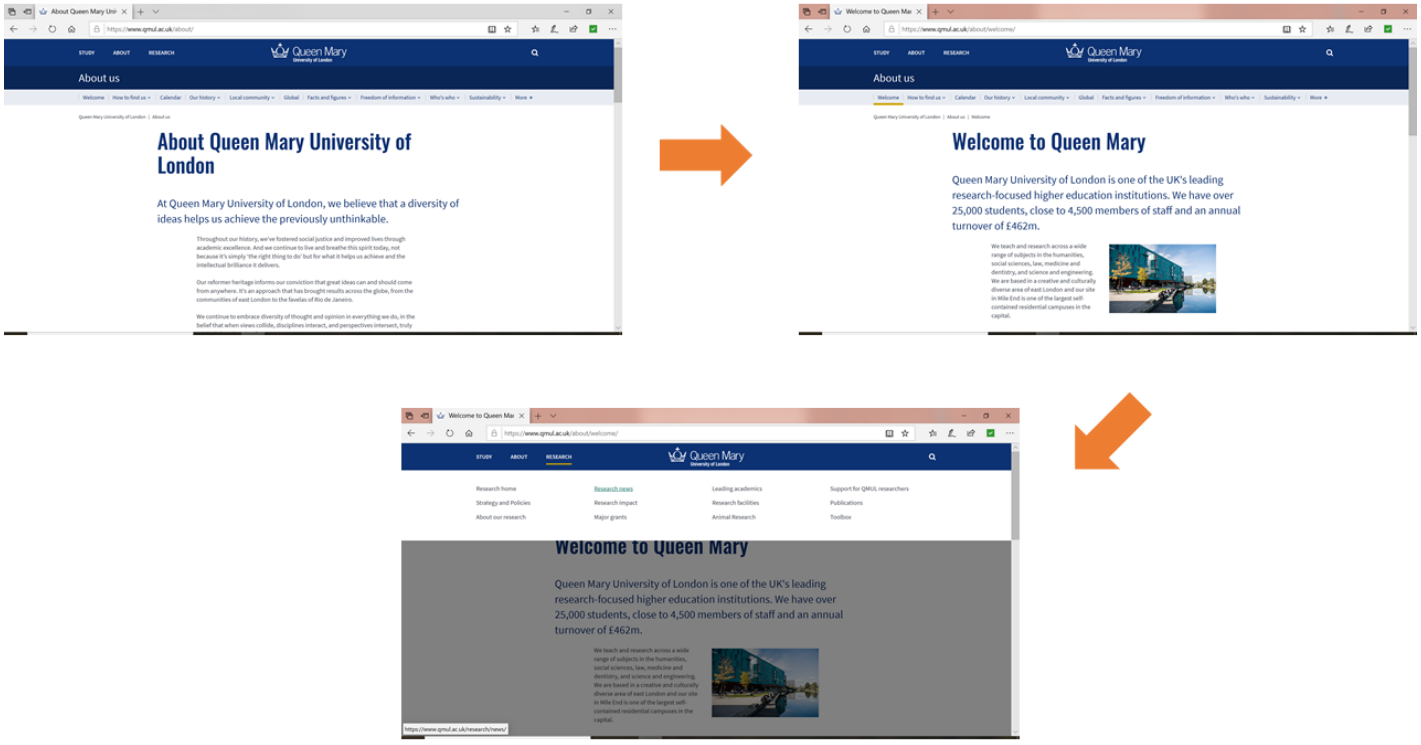
Record: Document Data

If the order of the terms (words) in a document is ignored, then a document can be represented as a *term vector*, where each term is a component (attribute) of the vector and the value of each component is the number of times the corresponding term occurs in the document. This representation of a collection of documents is often called a document-term matrix.

	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

Graph: World Wide Web

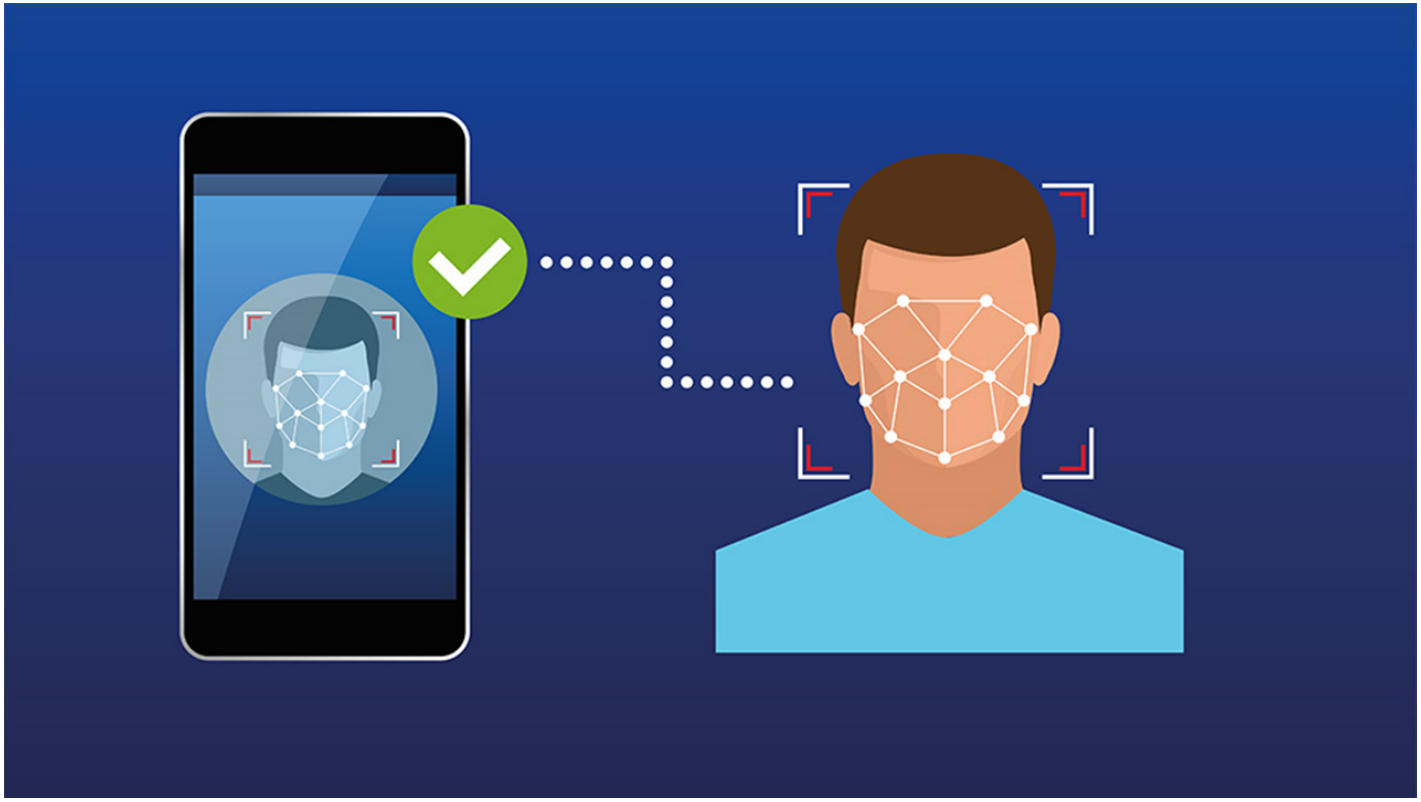
Data with Relationships among Objects.



- e.g. QMUL webpage

Graph: Multimedia Data

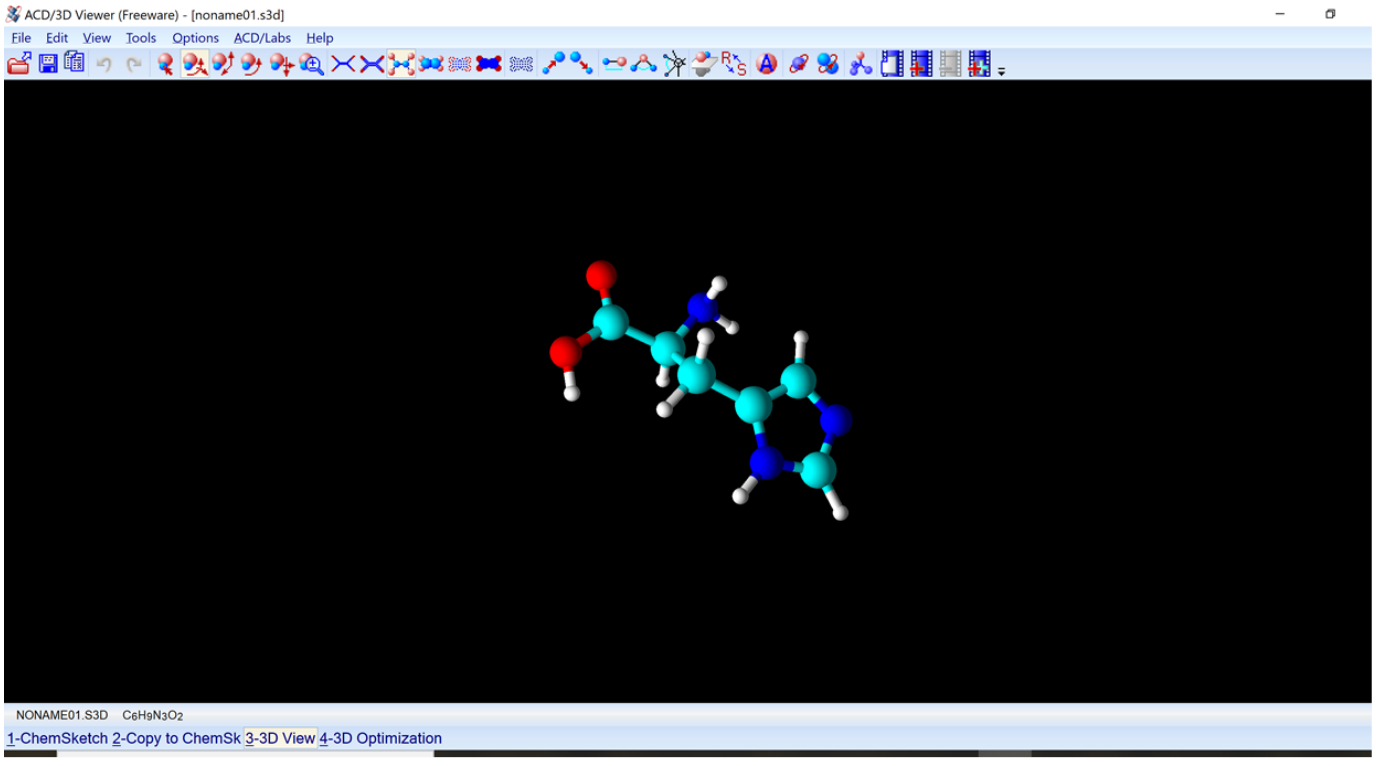
Data with Relationships among Objects.



- e.g. facial recognition

Graph: Molecular Structures

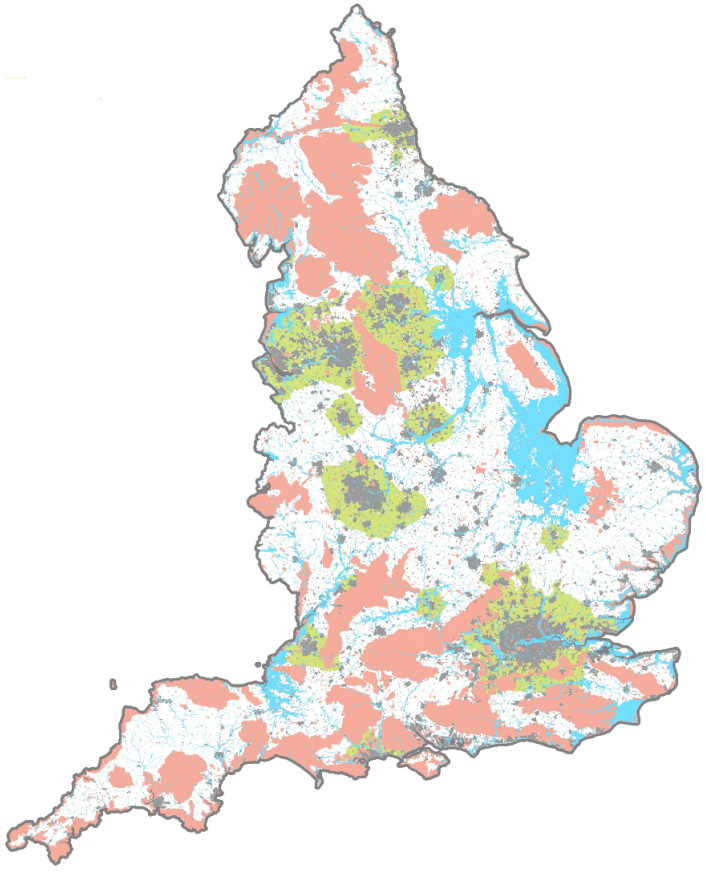
Data with Objects That Are Graphs.



- e.g. molecular structure - L-histidine

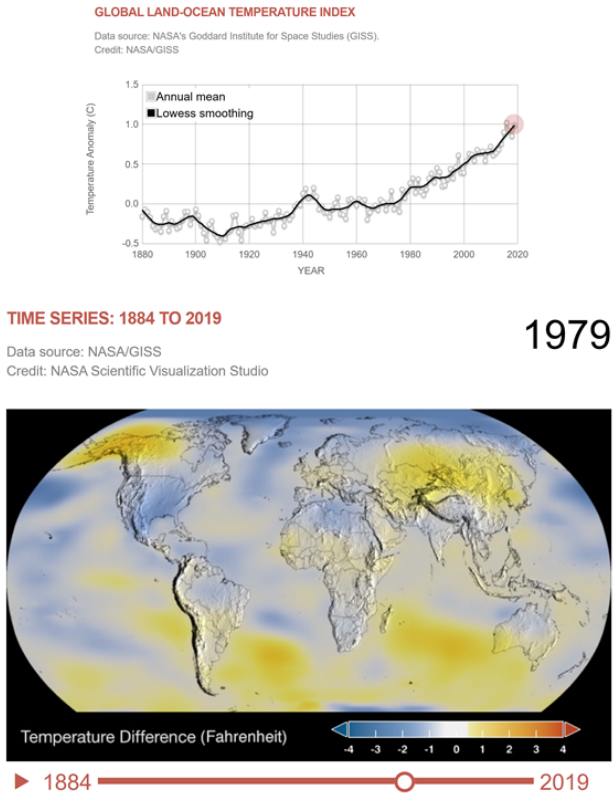
Ordered: Spatial Data

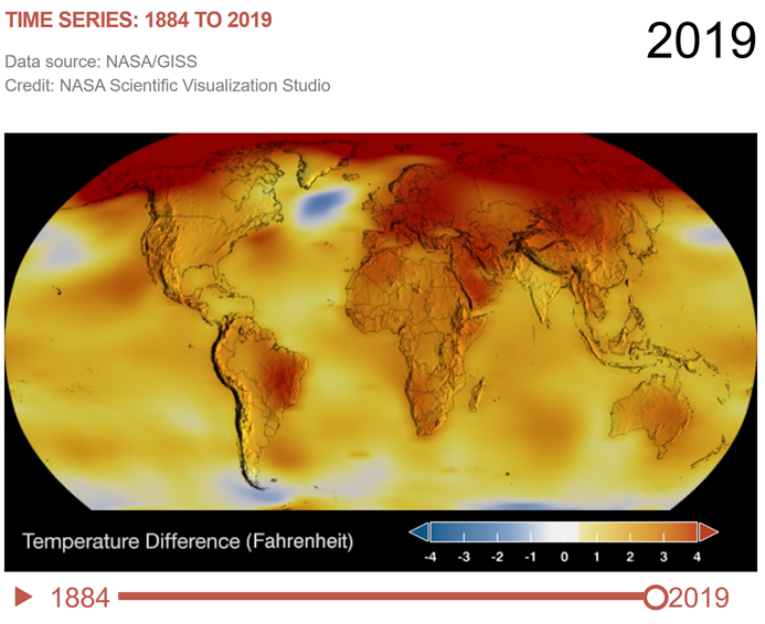
In the case of spatial data, some objects have spatial attributes, such as position or areas, as well as other types of attributes.



Ordered: Temporal and Spatiotemporal Data

In temporal (or sequential) data attributes have relationships that involve order in time. Spatiotemporal data have relationships that involve order in time and space.





- e.g. global land-ocean temperature changes in time

Ordered: Sequence Data

Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

- e.g. genetic information of plants and animals

Data Quality

Data Quality Definitions

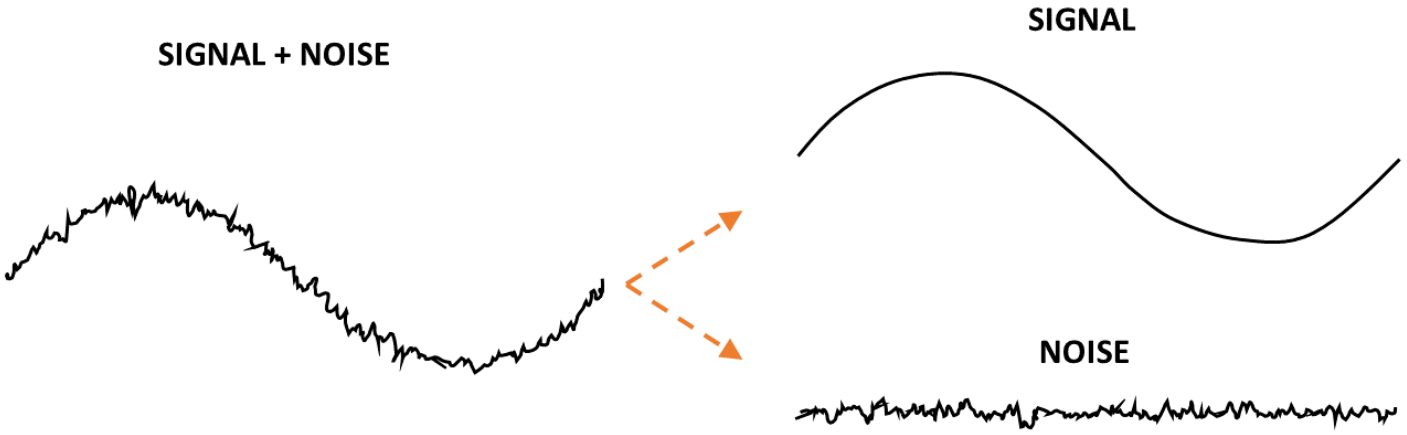
Data quality is a comparison of the actual state of a particular set of data to a desired state, with the desired state being typically referred to as ‘fit for use’ or ‘meeting requirements’. Requirements are defined in terms of characteristics or dimensions of the data:

- Accessibility and Availability
- Accuracy and Correctness
- Comparability
- Completeness and Comprehensiveness
- Consistency, Coherence, and Clarity
- Credibility, Reliability, and Reputation
- Relevance, Pertinence, and Usefulness
- Timeliness and Latency
- Uniqueness
- Validity and Reasonableness

Data Quality Problems: Noise and Outliers

Noise or poor data quality negatively affects many data processing efforts.

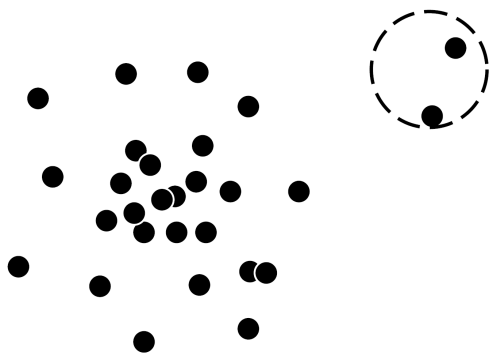
- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values



Data Quality Problems: Noise and Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.

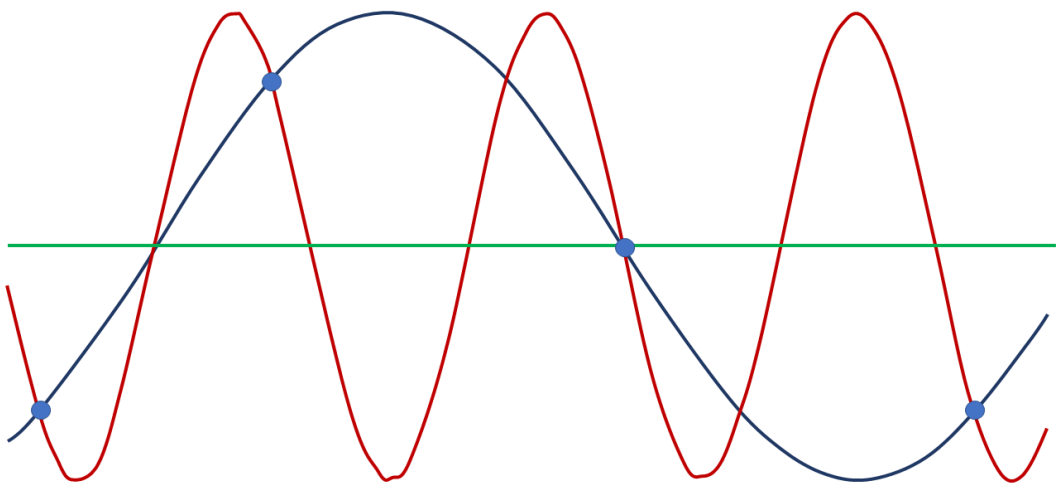
- Outliers are noise that interferes with data analysis
- Outliers are the goal of our analysis



Data Quality Problems: Missing Values

A data set has missing values when:

- Information is not collected
- Attributes may not be applicable to all cases



Data Quality Problems: Duplicate Data

A data set may include data objects that are duplicates (e.g. a person with multiple email addresses receiving same emails), or almost duplicates of one another (e.g. two distinct people with identical names).

- Major issues when merging data from heterogeneous sources
- Part of data cleaning is the process of dealing with duplicate data issues

Data Quality Problems: Wrong, Inaccurate and Fake Data

Data sets (information) have not been entered correctly or maintained.

Large volumes of fake news, i.e. those news articles with intentionally false information, are produced online for a variety of purposes, such as financial and political gain.

Fake News Detection on Social Media: A Data Mining Perspective

Kai Shu[†], Amy Sliva[‡], Suhang Wang[†], Jiliang Tang [♮], and Huan Liu[†]
[†]Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
[‡]Charles River Analytics, Cambridge, MA, USA
[♮]Computer Science & Engineering, Michigan State University, East Lansing, MI, USA
[†]{kai.shu,suhang.wang,huan.liu}@asu.edu,
[‡]asliva@cra.com, [♮] tangjili@msu.edu

https://arxiv.org/abs/1708.01967

Basic Statistical Descriptions of Data

Central Tendency

For data preprocessing to be successful, it is essential to have an overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

Measures of central tendency include the *mean*, *median*, *mode*, and *midrange*. The most common and effective numeric measure of the 'center' of a set of data is the arithmetic mean. Let $x_1, x_2, \ldots, x_N$ be a set of N values or observations, such as for some numeric attribute X. The *mean* (or average) of this set of values is the sum of values of a data set divided by number of values

$$mean(x) = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Central Tendency

Sometimes, each value x_i in a data set may be associated with a weight w_i for $i = 1, 2, \ldots, N$. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, the weighted arithmetic mean or the weighted average is

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

Central Tendency

A major problem with the mean is its sensitivity to extreme (e.g. outlier) values. Even a small number of extreme values can corrupt the mean.

- e.g. the mean salary at a company may be substantially pushed up by that of a few highly paid managers

The trimmed mean is the mean obtained after chopping off (e.g. 2%) values at the high and low extremes. However, trimming too large a portion (e.g. 20%) at both ends can result in the loss of valuable information.

Central Tendency

For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values.

$$median(x) = \begin{cases} x_{(r+1)} & \text{if } N \text{ is odd; i.e. N=2r+1} \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } N \text{ is even; i.e. N=2r} \end{cases}$$

Median can be applied to numeric and ordinal data. Suppose that a given data set of N values for an attribute X is sorted in increasing order. If N is odd, then the median is the middle value of the ordered set. If N is even, then the median is not unique; it is the two middlemost values and any value in between. If X is a numeric attribute, the median is taken as the average of the two middlemost values.

Central Tendency

The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.

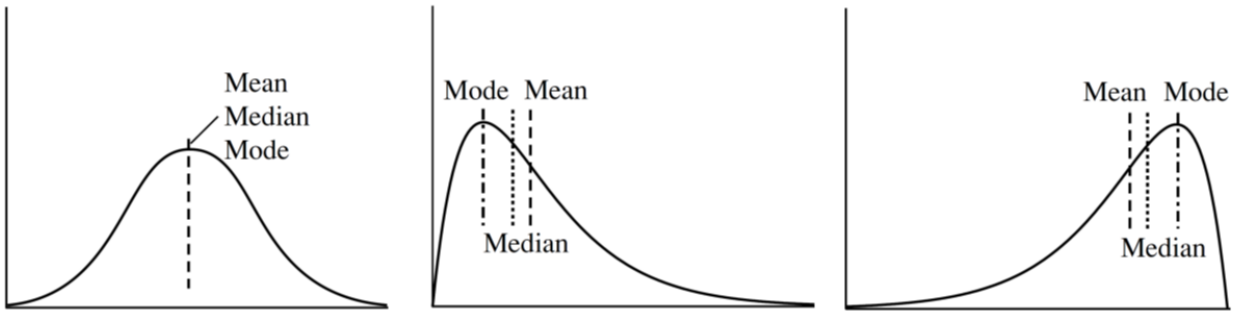
Data sets with one mode are called *unimodal* and data sets with two or more modes are *multimodal*. At the other extreme, if each data value occurs only once, then there is no mode.

The midrange is the average of the largest and smallest values in the set:

$$midrange(x) = \frac{max(x) + min(x)}{2} = \frac{x_{(N)} + x_{(1)}}{2}$$

Central Tendency

In a unimodal frequency curve with perfect *symmetric data* distribution, the mean, median, and mode are all at the same center value. Data in most real applications is asymmetric. It may be either *positively skewed*, where the mode occurs at a value that is smaller than the median, or *negatively skewed*, where the mode occurs at a value greater than the median.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Dispersion of Data

Let $x_1, x_2, \ldots, x_N$ be a set of observations for some numeric attribute, X. The range of the set is the difference between the largest (max(x)) and smallest (min(x)) values.

$$range(x) = max(x) - min(x) = x_{(N)} - x_{(1)}$$

Suppose that the data for attribute X is sorted in increasing numeric order. Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

- 2-quantile is the data point dividing the lower and upper halves of the data distribution, i.e. *median*
- 4-quantiles are the three data points that split the data distribution into four equal parts, i.e. *quartiles*
- 100-quantiles divide the data distribution into 100 equal-sized consecutive sets, i.e. *percentiles*

Dispersion of Data

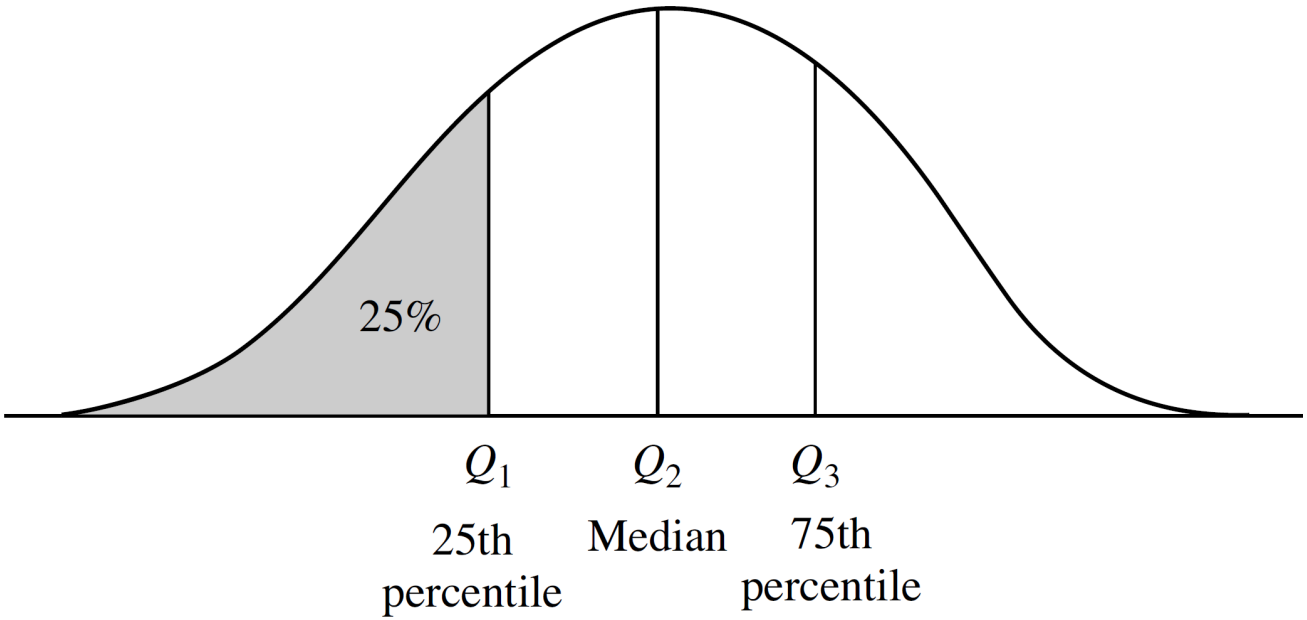


Figure: plotted percentiles for the data distribution of an attribute.

Dispersion of Data

The variance of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

where $\bar{x}$ is the mean value of the observations. The standard deviation, σ, of the observations is the square root of the variance, σ^2. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

Dispersion of Data

The mean can be distorted by outliers, and since the variance is computed using the mean, it is also sensitive to outliers. More robust estimates of the spread of a set of values are:

- the absolute average deviation (AAD)

$$AAD(x) = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

- the median absolute deviation (MAD)

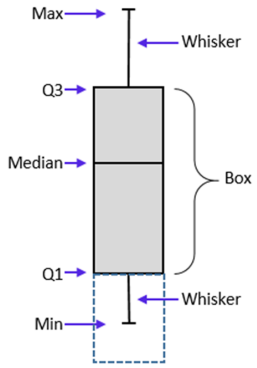
$$MAD(x) = median\left(\{|x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_N - \bar{x}|\}\right)$$

- the interquartile range (IQR)

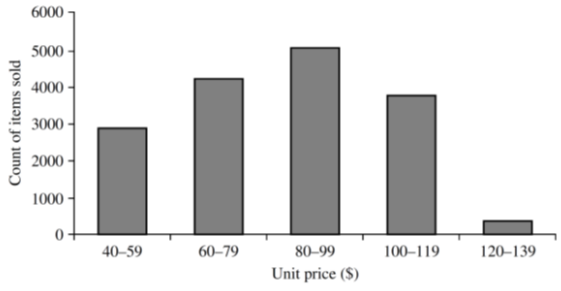
$$interquartile\ range(x) = x_{75\%} - x_{25\%}$$

Dispersion of Data

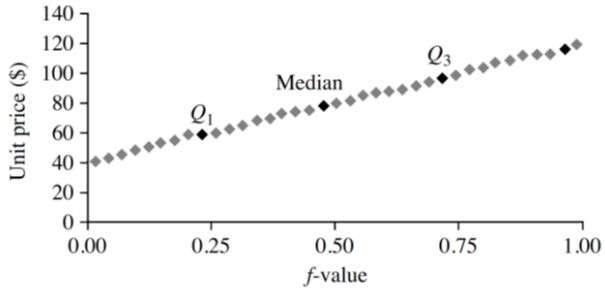
- Boxplot: graphic display of five-number summary



- Histogram: x-axis are values, y-axis are frequencies

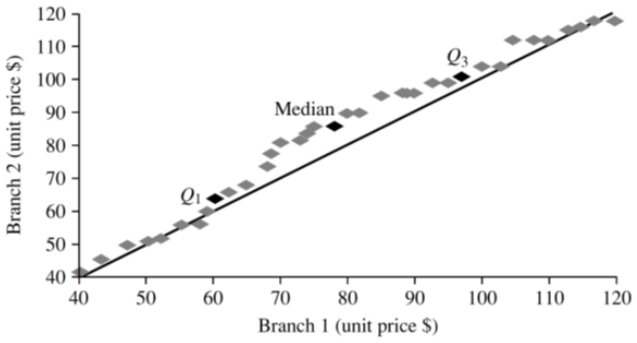


- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$

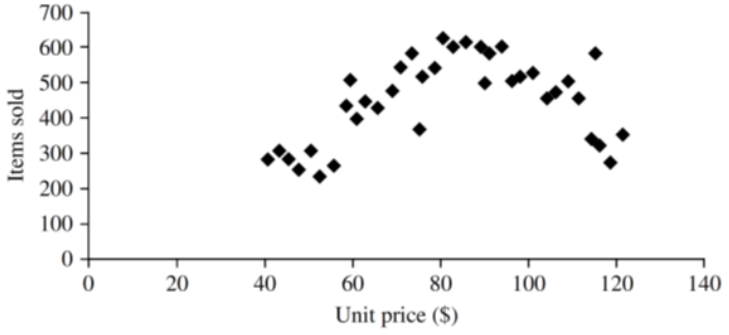


Dispersion of Data

- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another



- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane



Similarity and Distance

Similarity and Distance

Similarity and dissimilarity measures are referred to as measures of proximity. In data mining applications, such as *clustering*, *outlier analysis*, and *nearest-neighbor classification*, we need ways to assess how alike or unalike objects are in comparison to one another.

Data mining algorithms use the distance function as a key subroutine, and the design of the function directly impacts the quality of the results. Distance functions are highly sensitive to the type of the data, the dimensionality of the data, and the global and local nature of the data distribution.

Similarity and Dissimilarity Measures

Similarity between two objects is a numerical measure of the degree to which the two objects are alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

Distance or dissimilarity between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities fall in the interval [0, 1], but it is also common for them to range in the interval 0 to ∞.

Transformations

Transformations are often applied to convert a similarity to a dissimilarity, or vice versa, or to transform a proximity measure to fall within a particular range, such as [0, 1].

The transformation of similarities to the interval [0, 1] is given by the expression $s' = (s - min_s)/(max_s - min_s)$, where max_s and min_s are the maximum and minimum similarity values, respectively.

Likewise, dissimilarity measures with a finite range can be mapped to the interval [0, 1] by using the formula
$d' = (d - min_d)/(max_d - min_d)$.

Similarity/Dissimilarity for Simple Attributes

In the following Table, x and y are two objects that have a single simple attribute. Also, $d(x, y)$ and $s(x, y)$ are the dissimilarity and similarity between x and y, respectively.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)^*$	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d, s = \frac{1}{1+d}, s = \mathrm{e}^{-d},$ $s = 1 - \frac{d - min(d)}{max(d) - min(d)}$

*values mapped to integers 0 to $n - 1$, where n is the number of values.

Euclidean Distance

The most popular distance measure is Euclidean distance (i.e. straight line). Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ be two objects described by p numeric attributes. The Euclidean distance between objects $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

The Euclidean distance satisfies the following mathematical properties:

- **Non-negativity**: $d(\mathbf{x}, \mathbf{y}) \geq 0$
- **Identity of indiscernibles**: $d(\mathbf{x}, \mathbf{x}) = 0$
- **Symmetry**: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- **Triangle inequality**: $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{k}) + d(\mathbf{k}, \mathbf{y})$

A measure that satisfies these conditions is known as a *metric*.

Minkowski Distance

The Minkowski distance is a generalization of the Euclidean distance. It is defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt[h]{|x_1 - y_1|^h + |x_2 - y_2|^h + \cdots + |x_p - y_p|^h}$$

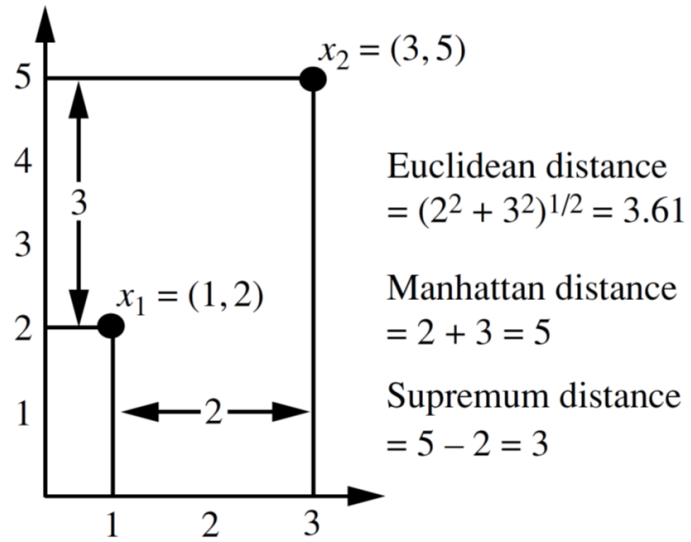
where h is a real number such that $h \geq 1$. It represents the Euclidean distance, when $h = 2$ (also known as L_2 norm), and the Manhattan distance, when $h = 1$ (i.e. L_1 norm). The Manhattan (or city block) distance is named so because it is the distance in blocks between any two points in a city:

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p|$$

Supremum Distance

The supremum distance, also referred to as L_{max} norm or L_∞ norm and as the Chebyshev distance, is a generalization of the Minkowski distance for $h \to \infty$. The attribute f gives the maximum difference in values between two objects. This difference is the supremum distance, defined formally as:

$$d(\mathbf{x}, \mathbf{y}) = \max_f(|x_f - y_f|) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_f - y_f|^h \right)^{\frac{1}{h}}$$



Similarity Measures for Binary Data

Similarity measures between objects that contain only binary attributes are called similarity coefficients, and typically have values between 0 and 1. A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.

Let x and y be two objects that consist of n binary attributes. The comparison of two such objects, i.e. two binary vectors, leads to the following four quantities

- f_{00} = the number of attributes where x is 0 and y is 0
- f_{01} = the number of attributes where x is 0 and y is 1
- f_{10} = the number of attributes where x is 1 and y is 0
- f_{11} = the number of attributes where x is 1 and y is 1

Simple Matching Coefficient

The simple matching coefficient (*SMC*) is defined as

$$SMC = \frac{number\ of\ matching\ attribute\ values}{number\ of\ attributes} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

This measure counts both presences and absences equally.

e.g. *SMC* can be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

Jaccard Coefficient

The Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes. The Jaccard coefficient, which is often symbolized by J, is given by the following equation

$$J = \frac{number\ of\ matching\ presences}{number\ of\ attributes\ not\ involved\ in\ 00\ matches} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Example: Calculate SMC and J for the following two binary vectors:
$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and
$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$.

Cosine Similarity

The Cosine similarity can be used to compare sparse vectors, such as "bag of word" representations in documents:

$$sim(x, y) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||}$$

where $||\mathbf{x}||$ is the Euclidean norm of vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$ and $\mathbf{x} \cdot \mathbf{y}$ is the inner product of $\mathbf{x}$ and $\mathbf{y}$.

The measure computes the cosine of the angle between vectors $\mathbf{x}$ and $\mathbf{y}$. A cosine value of 0 means that the two vectors have no match. The closer the cosine value to 1, the smaller the angle and the greater the match between vectors.

Summary

Attributes and Objects Data sets are made up of data objects that are described by attributes.

Characteristics of Data Data sets have some well defined characteristics.

Types of Data Data types are record, graph, and ordered.

Data Quality Data sets are never perfect.

Basic Statistical Descriptions of Data Basic statistical descriptions provide the analytical foundation for data preprocessing.

Similarity and Distance Different measures of proximity can be computed for specific attribute types.

Questions?
also please use the forum on QM+