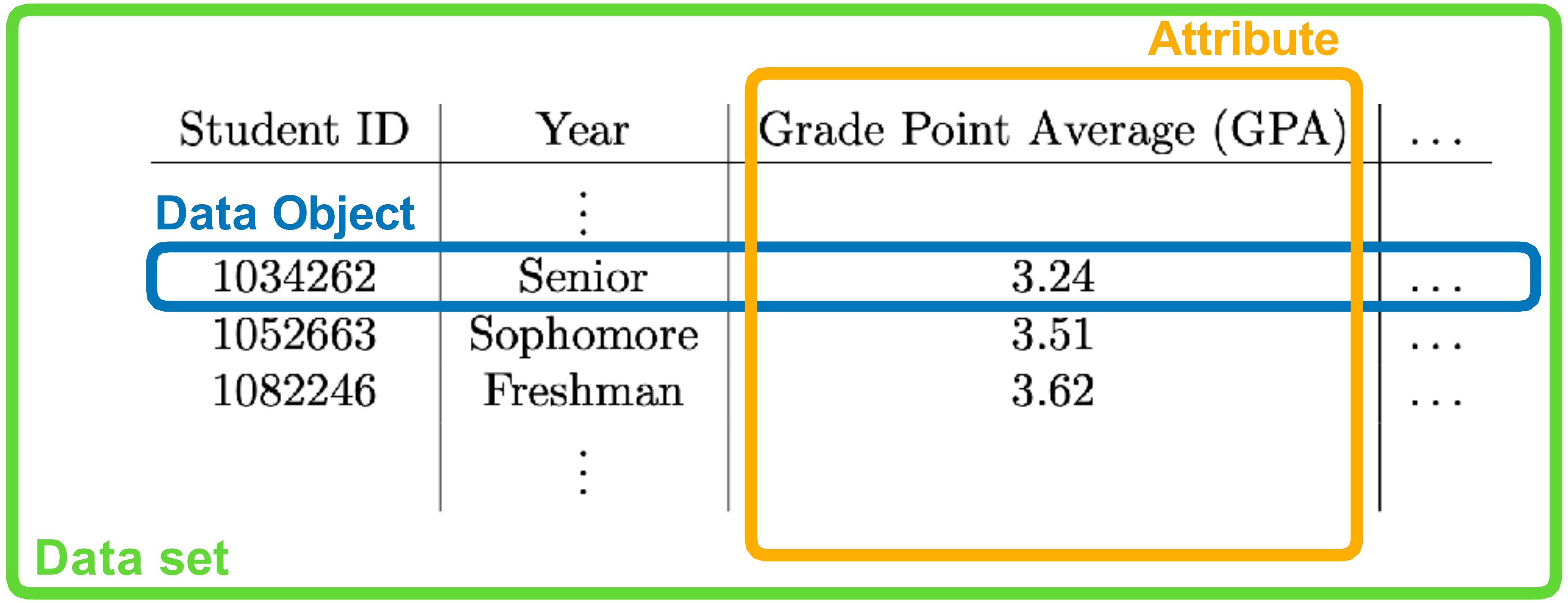
Data

CI603 - Data Mining

Lecture 2

Types of Data

- A **data set** can be viewed as a collection of **data objects** (also known as *record*, *vector*, *pattern*, *event*, *case*, *sample*, *instance*, *observation*, or *entity*)



Student ID	Year	Grade Point Average (GPA)	. . .
	⋮		
1034262	Senior	3.24	. . .
1052663	Sophomore	3.51	. . .
1082246	Freshman	3.62	. . .
	⋮		

Attribute
Data Object
Data set

- **Data objects** are described by a number of **attributes** (also known as *variable*, *characteristic*, *field*, *feature*, or *dimension*) that capture the characteristics of an object.

Attributes

- An **attribute** is a property or characteristic of an object that can vary, either **from one object to another** (i.e. *eye colour*) or **from one time to another** (i.e. temperature).

- **Attributes** can be described using **different measurement scales**, which in turn have **different properties**.

    1. **Distinctness** $= \ and \ \neq$
    2. **Order** $>, \geq \ , <, and \ \leq$
    3. **Addition** $+ \ and \ -$
    4. **Multiplication** x *and* /

- Given these properties, four types of attributes can be defined: **nominal**, **ordinal**, **interval**, and **ratio**.

Types of Attributes

Attribute Type		Decription	Examples	Operations
Categorical (qualitative)	**Nominal**	Values to provide only enough information to **distinguish one object from another**.	postal codes, employee ID numbers, eye colour, gender.	mode, entropy, contingency correlation, χ^2 test
	Ordinal	Values to provide enough information to **order objects**.	clothes sizes, grades, {*good*, *better*, *best*}, street numbers	median, percentiles, rank correlation
Numeric (quantitative)	**Interval**	**Differences** between values are meaningful	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio	Both **differences** and **ratios** are meaningful.	monetary quantities, counts, age, mass, length, electrical current, temperature in Kelvin	geometric mean, harmonic mean, percent variation

Attributes

- **Qualitative attributes**, such as *employee ID*, **lack most of the properties of numbers**. Even if they are represented by numbers, *i.e., integers*, they **should be treated more like symbols**.

- **Quantitative attributes** are **represented by numbers** and have **most of the properties of numbers**. They can be **integer-valued** or **continuous**.

Number of Values

- An independent way of distinguishing between attributes is by the **number of values** they can take.
  - **Discrete**. A discrete attribute has a **finite or countably infinite set of values**. Such attributes can be **categorical** (*i.e. postcodes or ID numbers*), or **numeric** (*i.e. counts*).
    - **Binary attributes** are a special case of discrete attributes.
  - **Continuous**. A continuous attribute is one whose values are **real numbers** (*i.e. temperature, height, or weight*).

Asymmetric Attributes

- **Only presence**—a **non-zero attribute value**—is regarded as important.
- **Binary attributes** where **only non-zero values are important** are called **asymmetric binary attributes**.
- This type of attribute is particularly important for **association analysis** (*Next session - Week 3*).
- It is also possible to have **discrete or continuous asymmetric features**.

General Characteristics of Data Sets

Dimensionality

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

- Analysing data with a **small number of dimensions** tends to be **different** from analysing **moderate or high-dimensional data**.

- As dimensionality increases, the **data becomes increasingly sparse**, affecting mainly **clustering** and **classification** algorithms (**the curse of dimensionality**).

- An important motivation in preprocessing the data is dimensionality reduction.

Distribution

- **Distribution**: the **frequency of occurrence** of various values or sets of values for the attributes of data objects.
- The **distribution of a data set** can be considered as a description of the **concentration of objects** in various regions of the data space.
- Many data sets have distributions that are **not well captured by standard statistical distributions**.
- As a result, many data mining algorithms do **not assume a particular statistical distribution** for the data they analyse. However, some **general aspects of distributions** often have a **strong impact**.

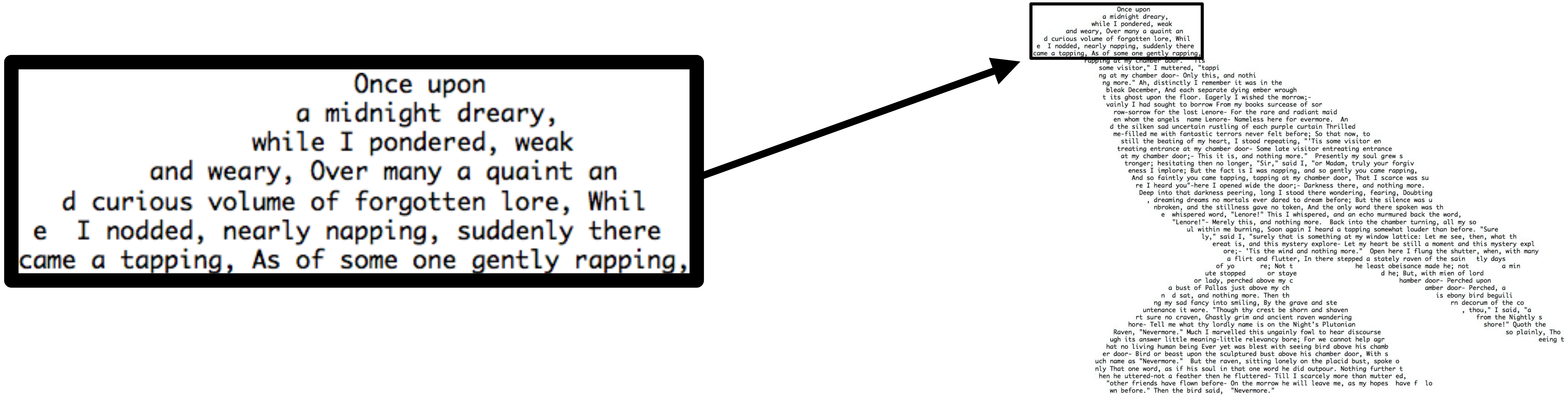
Distribution

- A **special case of skewed data** is **sparsity**. For **sparse binary**, **count** or **continuous data**, most attributes of an object have values of 0.
- In practical terms, **sparsity** is an **advantage** because usually **only the non-zero values need to be stored and manipulated** (computation time and storage savings).
- Indeed, some data mining algorithms, such as the *association rule mining algorithms*, **work well only for sparse data**.

Resolution

- **Resolution**. It is frequently possible to obtain data at **different levels of resolution**, and often the **properties of the data are different at different resolutions**.

- The **patterns in the data** also 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 can disappear**.



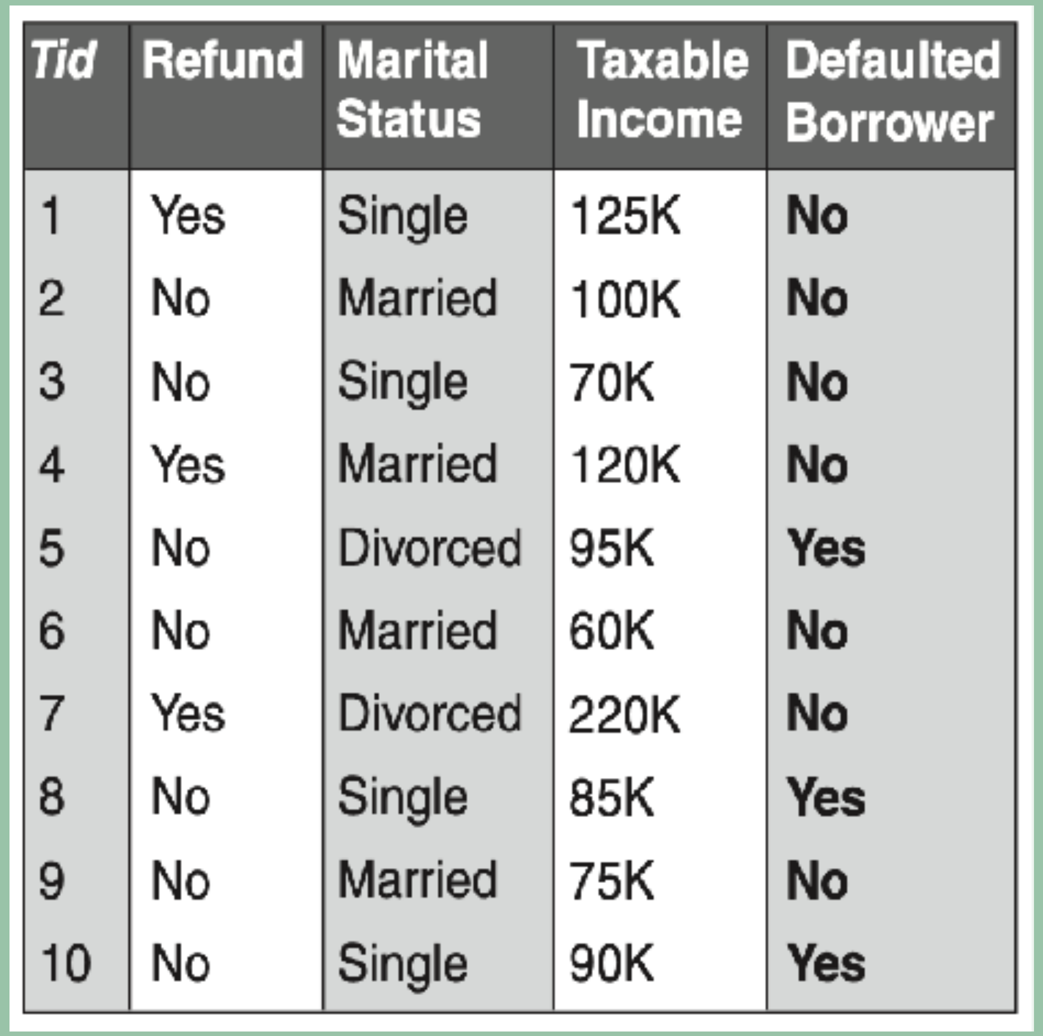
```
                 Once upon
              a midnight dreary,
           while I pondered, weak
      and weary, Over many a quaint an
  d curious volume of forgotten lore, Whil
e  I nodded, nearly napping, suddenly there
came a tapping, As of some one gently rapping,
```

```
                 Once upon
              a midnight dreary,
           while I pondered, weak
      and weary, Over many a quaint an
  d curious volume of forgotten lore, Whil
e  I nodded, nearly napping, suddenly there
came a tapping, As of some one gently rapping,
            rapping at my chamber door. "'Tis
              some visitor," I muttered, "tappi
              ng at my chamber door- Only this, and nothi
               ng more." Ah, distinctly I remember it was in the
                bleak December, And each separate dying ember wrough
               t its ghost upon the floor. Eagerly I wished the morrow;-
                vainly I had sought to borrow From my books surcease of sor
                  row-sorrow for the lost Lenore- For the rare and radiant maid
                  en whom the angels  name Lenore- Nameless here for evermore.  An
                 d the silken sad uncertain rustling of each purple curtain Thrilled
                  me-filled me with fantastic terrors never felt before; So that now, to
                    still the beating of my heart, I stood repeating, "'Tis some visitor en
                   treating entrance at my chamber door- Some late visitor entreating entrance
                    at my chamber door;- This it is, and nothing more."  Presently my soul grew s
                     tronger; hesitating then no longer, "Sir," said I, "or Madam, truly your forgiv
                      eness I implore; But the fact is I was napping, and so gently you came rapping,
                       And so faintly you came tapping, tapping at my chamber door, That I scarce was su
                        re I heard you"-here I opened wide the door;- Darkness there, and nothing more.
                         Deep into that darkness peering, long I stood there wondering, fearing, Doubting
                           , dreaming dreams no mortals ever dared to dream before; But the silence was u
                            nbroken, and the stillness gave no token, And the only word there spoken was th
                              e  whispered word, "Lenore!" This I whispered, and an echo murmured back the word,
                                 "Lenore!"- Merely this, and nothing more.  Back into the chamber turning, all my so
                                     ul within me burning, Soon again I heard a tapping somewhat louder than before. "Sure
                                        ly," said I, "surely that is something at my window lattice: Let me see, then, what th
                                           ereat is, and this mystery explore- Let my heart be still a moment and this mystery expl
                                              ore;- 'Tis the wind and nothing more."  Open here I flung the shutter, when, with many
                                               a flirt and flutter, In there stepped a stately raven of the sain   tly days
                                            of yo      re; Not t                 he least obeisance made he; not         a min
                                         ute stopped     or staye                      d he; But, with mien of lord
                                      or lady, perched above my c                           hamber door- Perched upon
                               a bust of Pallas just above my ch                                   amber door- Perched, a
                             n  d sat, and nothing more. Then th                                      is ebony bird beguili
                           ng my sad fancy into smiling, By the grave and ste                              rn decorum of the co
                        untenance it wore. "Though thy crest be shorn and shaven                              , thou," I said, "a
                      rt sure no craven, Ghastly grim and ancient raven wandering                                from the Nightly s
                    hore- Tell me what thy lordly name is on the Night's Plutonian                                  shore!" Quoth the
                Raven, "Nevermore." Much I marvelled this ungainly fowl to hear discourse                                so plainly, Tho
               ugh its answer little meaning-little relevancy bore; For we cannot help agr                                          eeing t
              hat no living human being Ever yet was blest with seeing bird above his chamb
             er door- Bird or beast upon the sculptured bust above his chamber door, With s
            uch name as "Nevermore."  But the raven, sitting lonely on the placid bust, spoke o
            nly That one word, as if his soul in that one word he did outpour. Nothing further t
             hen he uttered-not a feather then he fluttered- Till I scarcely more than mutter ed,
               "other friends have flown before- On the morrow he will leave me, as my hopes  have f  lo
                wn before." Then the bird said,  "Nevermore."
```

Types of Data Sets

Record Data

- Much data mining work assumes that the data set is a **collection of records** (data objects), each of which consists of a **fixed set of data fields** (attributes).
- In which 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** either in **flat files** or in **relational databases**.



Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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**

Transaction (or Market Basket) Data

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

- This type of data is called **market basket data** because the items in each record are the products in a person's "market basket."

TID	Items
T1	{Bread, Milk, Eggs}
T2	{Milk, Beer}
T3	{Milk, Nappies}
T4	{Bread, Milk, Beer}
T5	{Bread, Nappies}
T6	{Milk, Nappies}
T7	{Bread, Nappies}
T8	{Bread, Milk, Nappies,
T9	{Bread, Milk, Nappies}

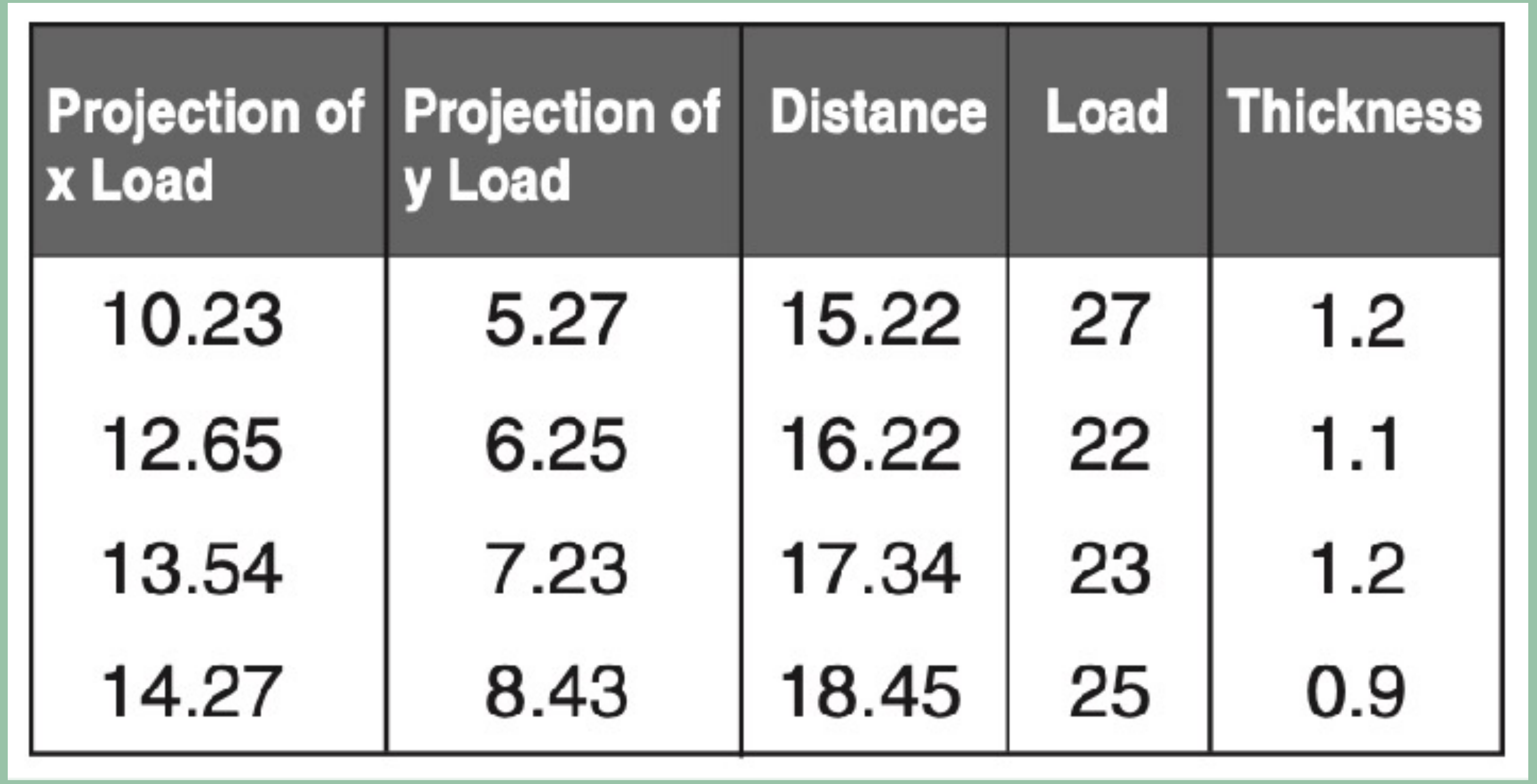
TID	Bread	Milk	Nappies	Beer	Eggs
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

The Data Matrix

- If all the data objects have the same fixed set of numeric attributes, then the data objects can be thought of as **vectors in a multidimensional space**, where **each dimension represents a distinct attribute**.

- A set of such data objects can be **interpreted** as an **m by n matrix**, where there are **m rows**, one for each object, and **n columns**, one for each attribute.

- **Standard matrix operation** can be applied to transform and manipulate the data (standard data format for most statistical data)



Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

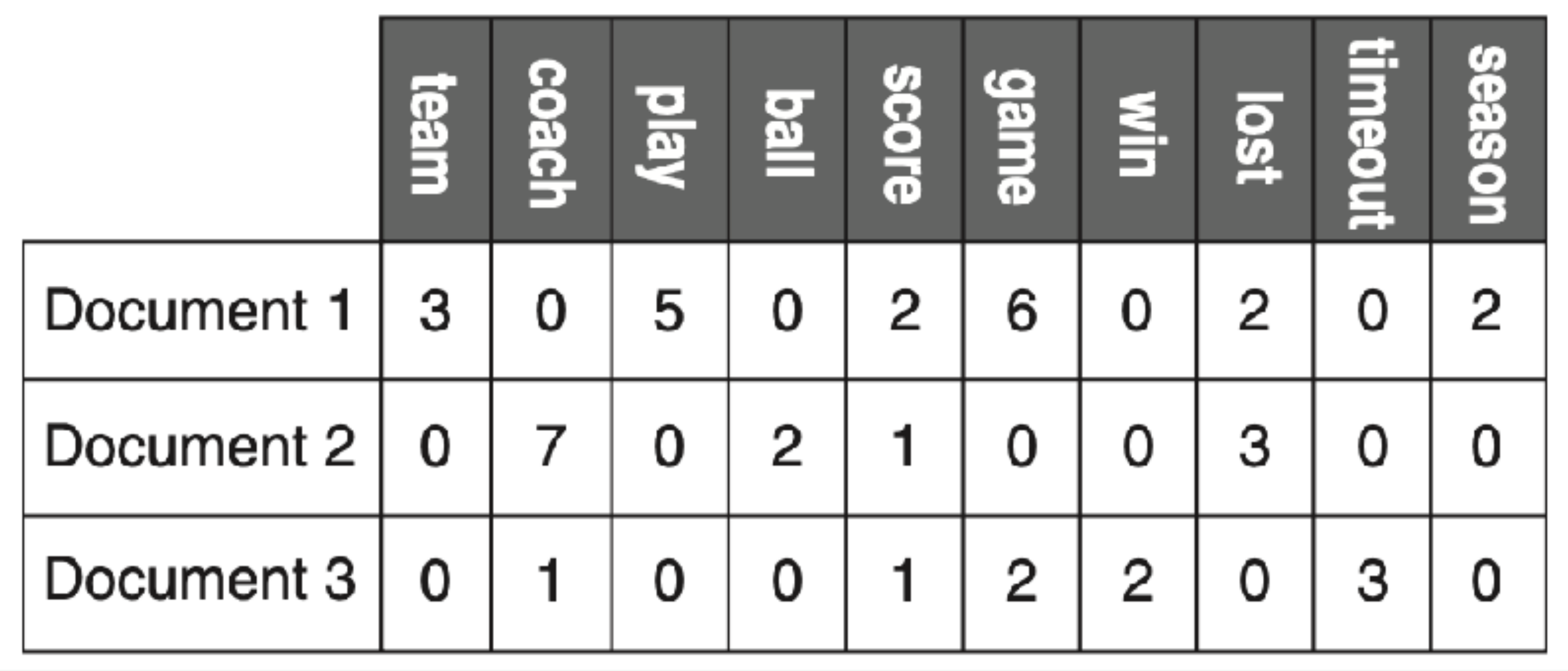
The Sparse Data Matrix

- A **sparse data matrix** is a special case of a data matrix where the **attributes** are of the **same type** and are **asymmetric.**

- **Transaction data** is an example of a **sparse data matrix** that has only 0–1 entries.

- When the **order of the terms** (**words**) in a document **is ignored** (*bag of words*), then a **document** can be represented as a **term vector** (**document-term matrix**).

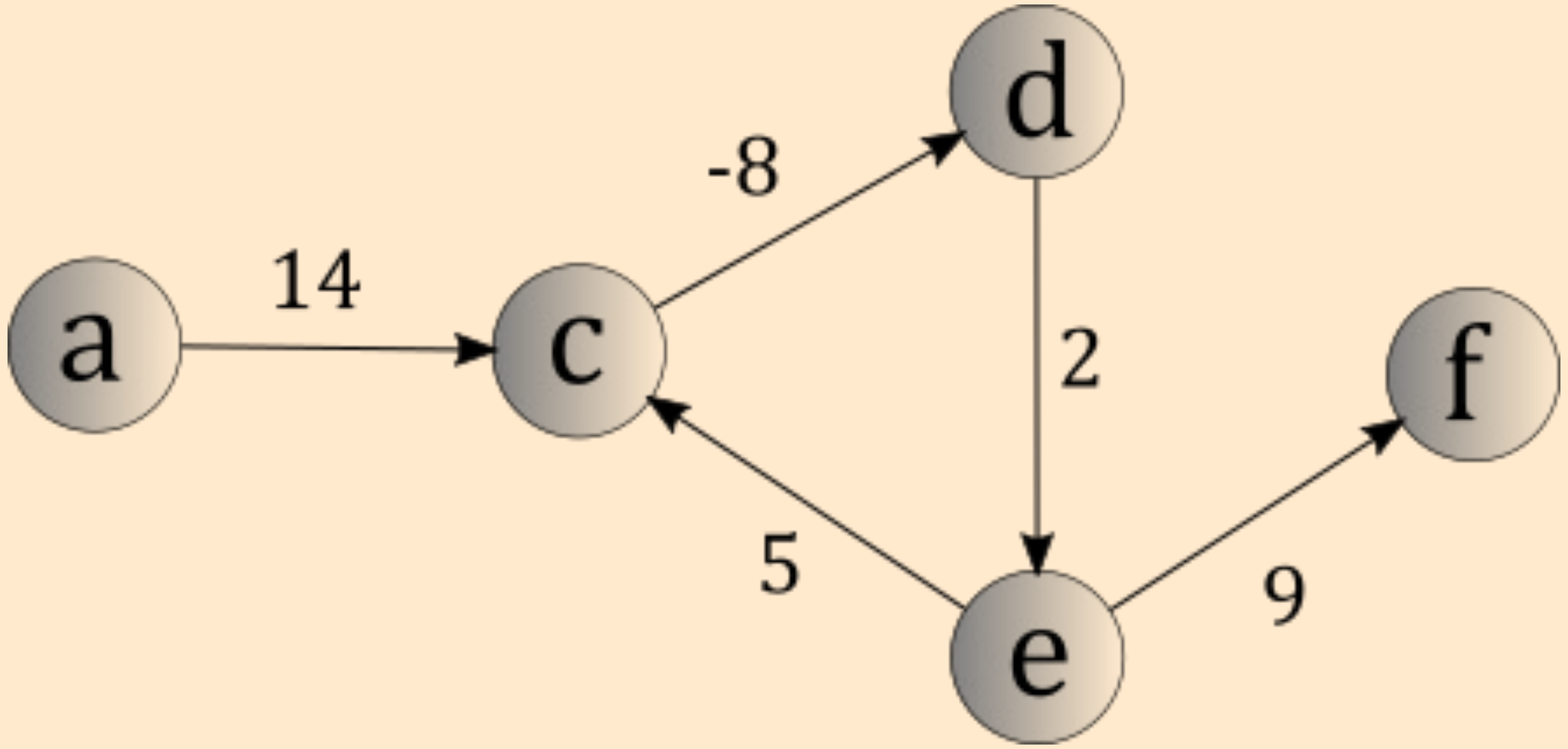
- In practice, **only the non-zero** entries of sparse data matrices are **stored**.



	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-Based Data

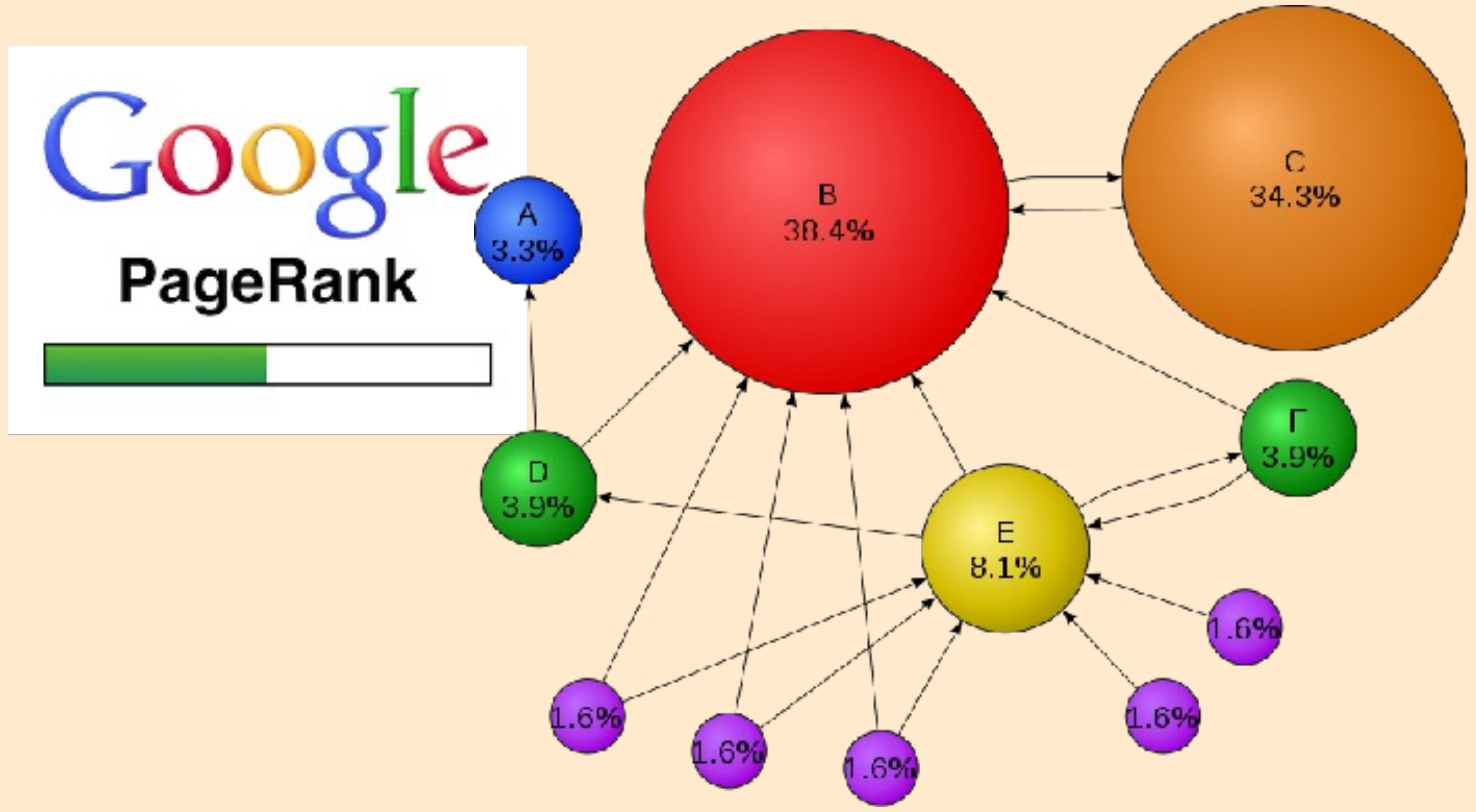
- A **graph** can sometimes be a convenient and powerful representation for data. We consider two specific cases:
  - (1) the **graph** captures **relationships among data objects** and
  - (2) the **data objects themselves** are represented as **graphs**.



Graph-Based Data

a) Data with Relationships among Objects.

- In particular, the **data objects** are mapped to **nodes** of the graph, while the **relationships among objects** are captured by the **links between objects** and **link properties**, such as **direction and weight**.



Google's PageRank

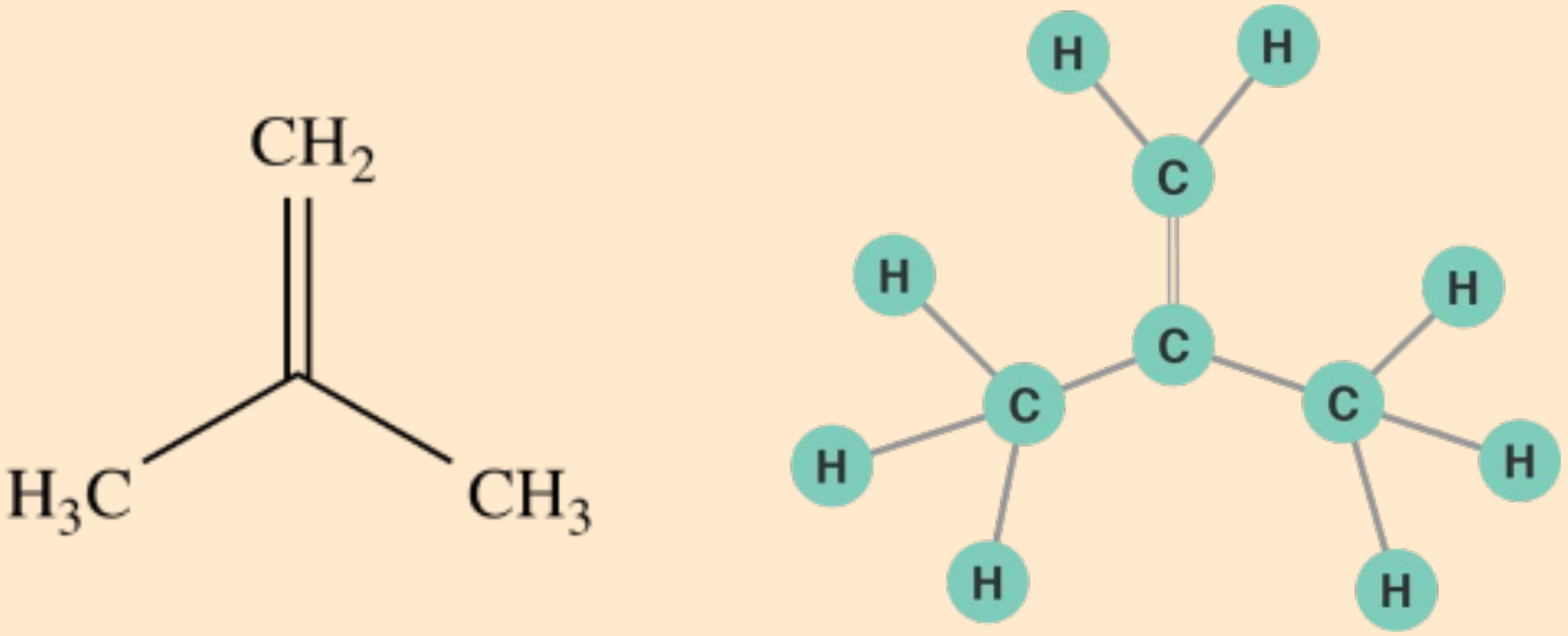
Social Network

Graph-Based Data

b) Data with Objects that are Graphs.

- If the **objects** contain **subobjects** that **have relationships**, then such **objects** are frequently **represented as graphs**.

- For example, the **structure of chemical compounds** can be represented by a **graph**, where the **nodes** are **atoms** and the **links between nodes** are **chemical bonds**.



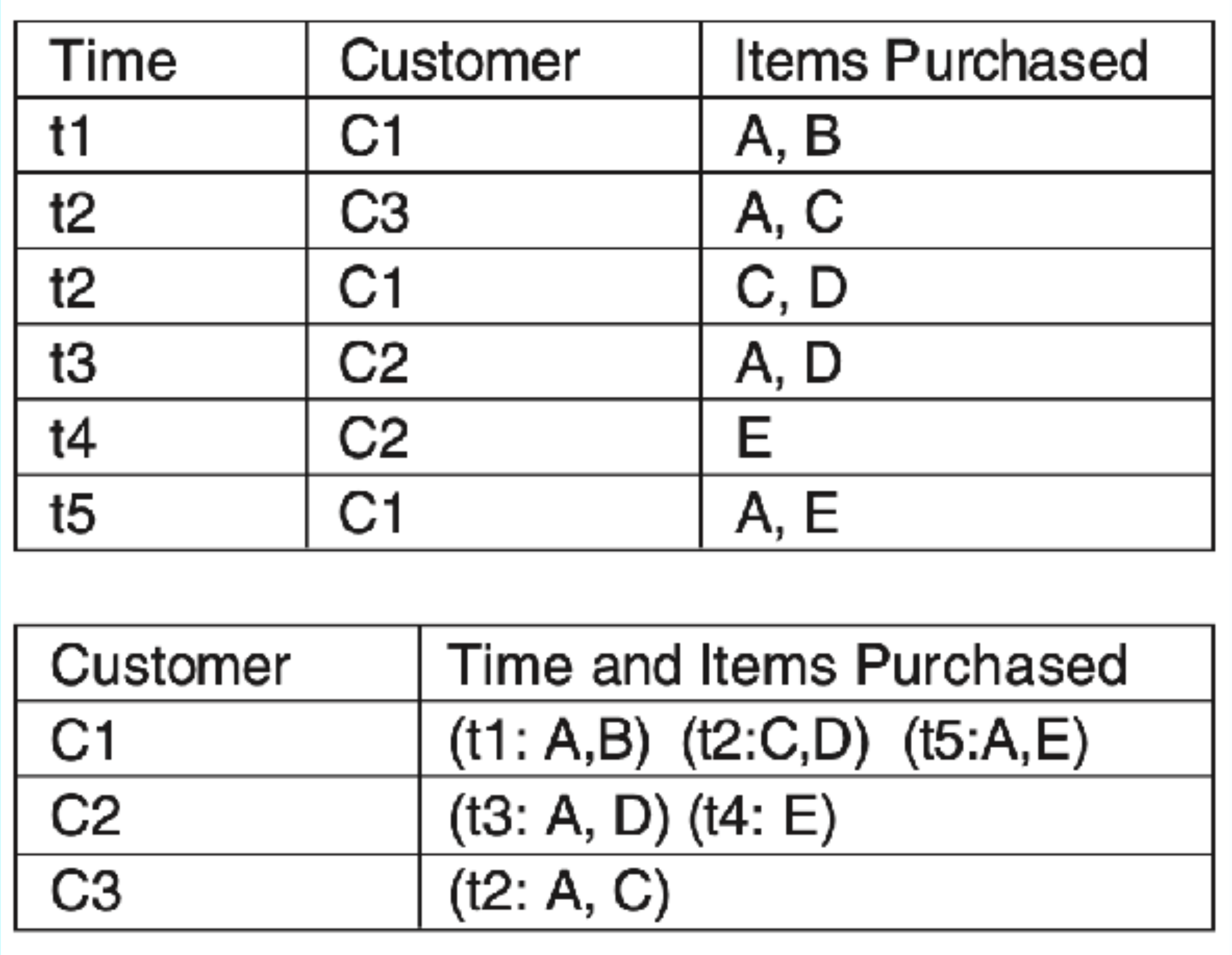
The molecular graph of an isobutylene molecule

Ordered Data

- For some types of data, the **attributes have relationships** that involve **order in time** or **space**.

Sequential Transaction Data

- Sequential transaction data can be thought of as an extension of transaction data, where **each transaction** has a **time associated with it**.
- Retail transaction data sets can store the time at which the transaction took place.
- A **time** can also be associated with **each attribute**.



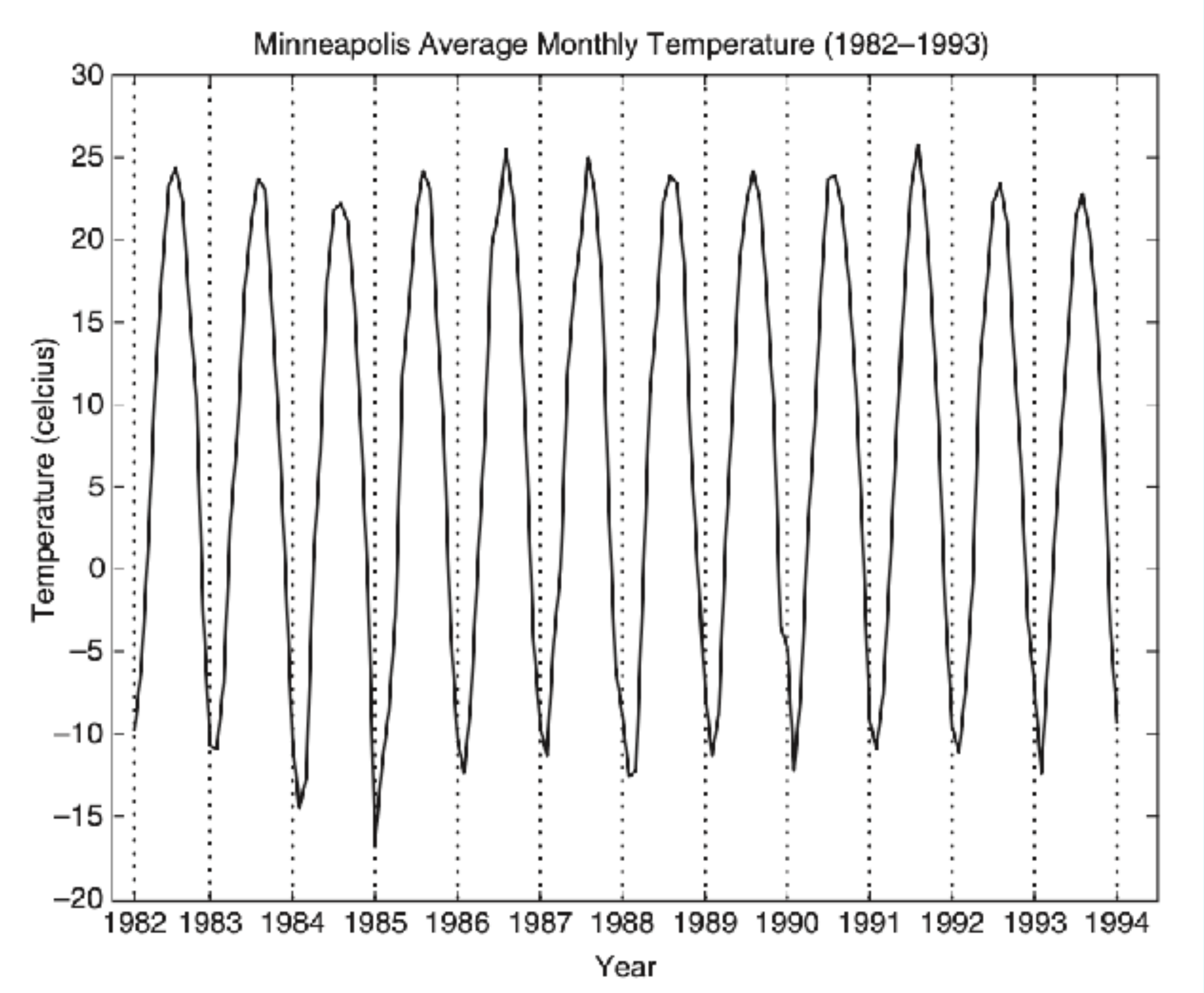
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Time Series

- **Time series data** is a **special type of ordered data** where each **record** is a **time series**, i.e., a series of measurements taken over time.

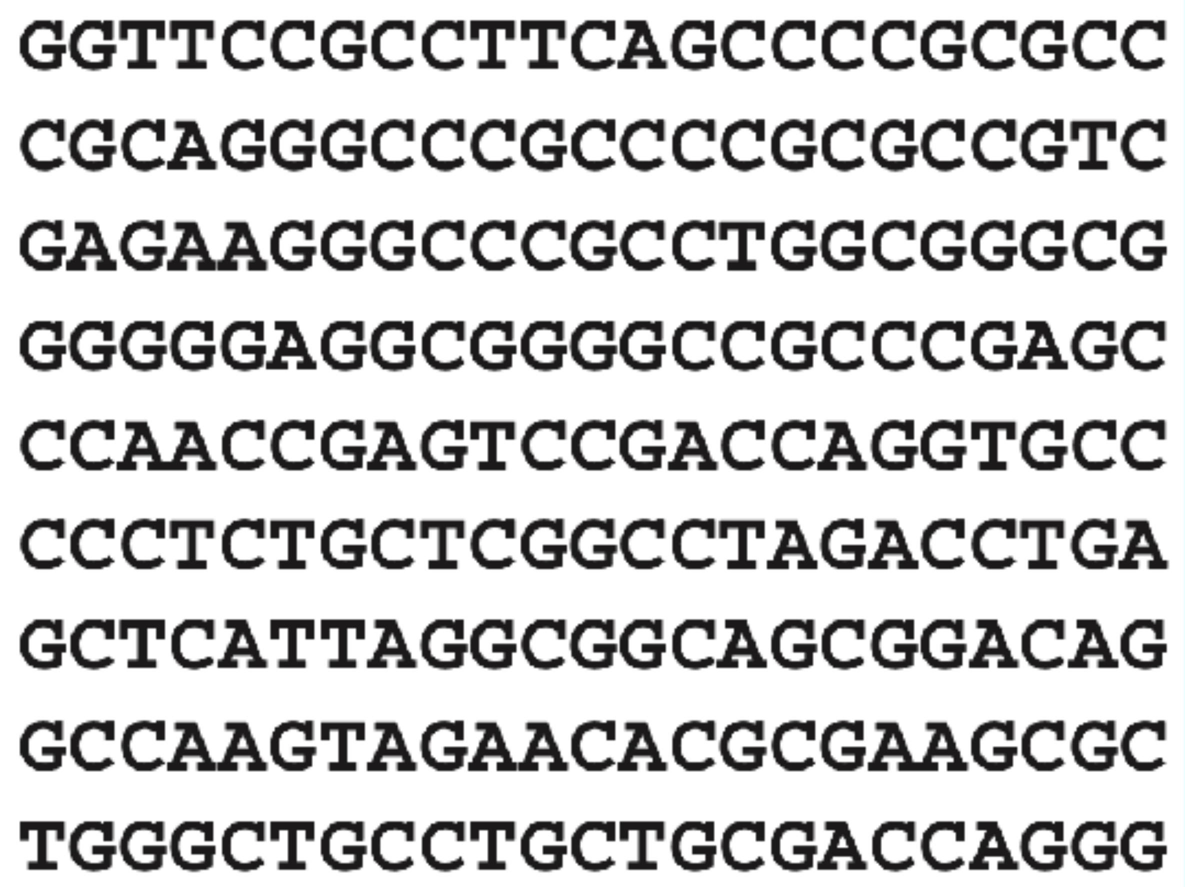
- When working with temporal data, such as time series, it is important to consider **temporal autocorrelation**.



Sequence Data

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

- It is quite similar to sequential data, except that there are **no time stamps**; instead, there are **positions in an ordered sequence**.



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

Genetic information

Spatial and Spatio-Temporal Data

- Some objects have **spatial attributes**, such as **positions** or **areas**, in addition to other types of attributes.

- i.e. **weather data** (precipitation, temperature, pressure) that is **collected for a variety of geographical locations**.

- Often such measurements are **collected over time**, and thus, the data consists of **time series** at **various locations** (*spatio-temporal data*).

- A complete analysis of *spatio-temporal data* requires **consideration of both the spatial and temporal aspects** of the data.

- An important aspect of spatial data is **spatial autocorrelation**.

Handling Non-Record Data

- **Most data mining algorithms** are **designed for record data** or its variations, such as transaction data and data matrices.
- Record-oriented techniques **can be applied** to **non-record data** by **extracting features from data objects** and using these features to **create a record corresponding to each object**.
- In some cases, it is easy to represent the data in a record format, but this type of representation does **not capture all the information in the data**.

Data Quality

Data Quality

- **Data mining algorithms** are sometimes applied to **data that was collected for another purpose**.

- Because **preventing data quality problems** is typically **not an option** data mining focuses on:

    - (1) the **detection and correction** of data quality problems (**data cleaning**), and

    - (2) the **use of algorithms that can tolerate poor data quality**.

Measurement and Data Collection Issues

- There may be problems due to **human error**, **limitations of measuring devices**, or **flaws in the data collection process**.
- As a consequence, **values** or even **entire data objects** can be **missing**.
- In other cases, there can be **spurious** or **duplicate objects**.
- Even if all the data is present and “looks fine,” there may be **inconsistencies**.

Measurement and Data Collection Errors

a) **Measurement error** refers to **any problem resulting from the measurement process**.

  - For **continuous attributes**, the numerical difference of the measured and true value is called the **error**.

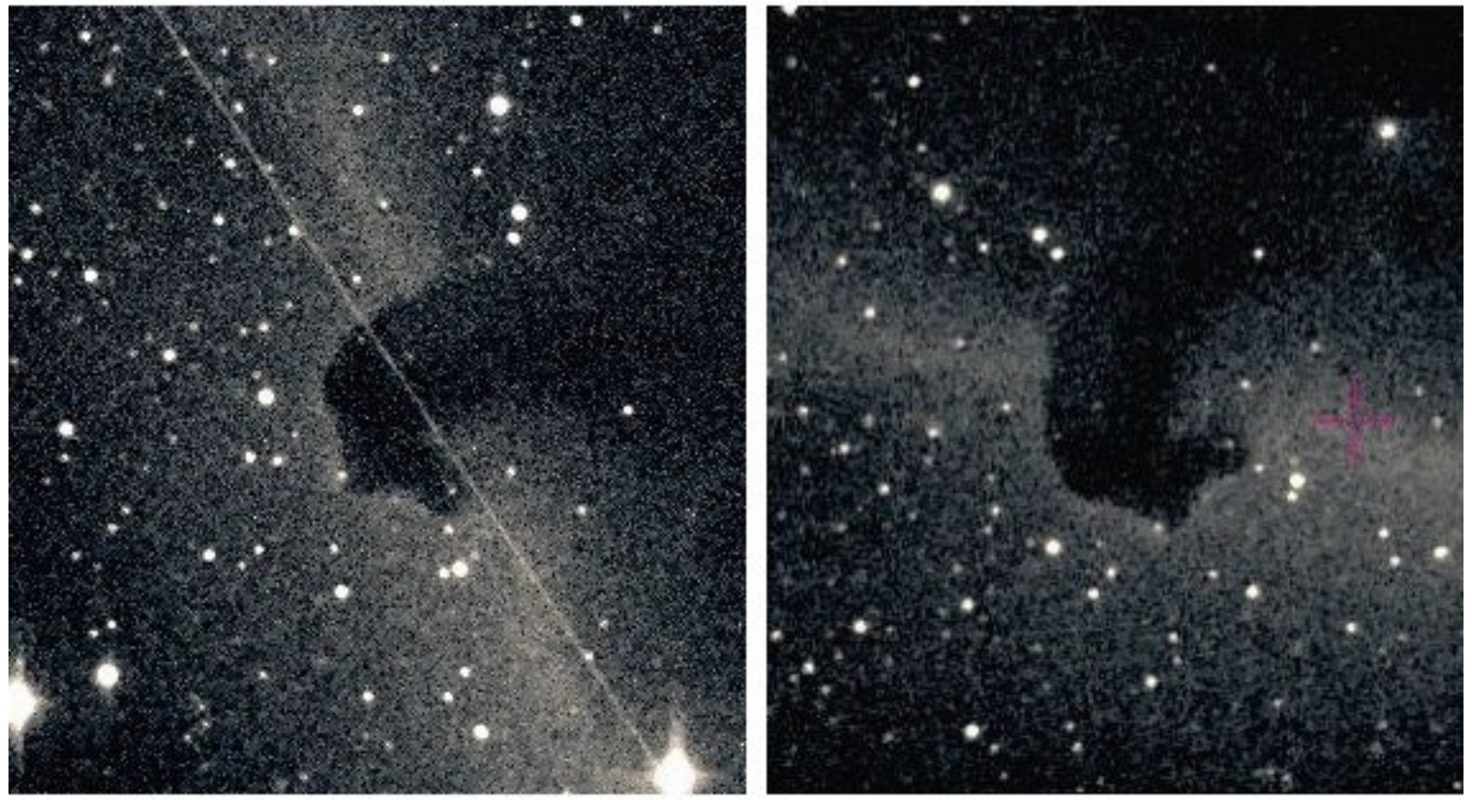
b) **Data collection error** refers to errors such as **omitting** data objects or attribute values, or **inappropriately including a data object**.

- Both **measurement errors** and **data collection errors** can be either **systematic** or **random**.

- **Certain types of data errors** are **common**, and **well-developed techniques** often exist for **detecting and/or correcting** these errors.

Noise and Artefacts

- **Noise** is the **random component** of a **measurement error**.
- The term **noise** is often used **in connection with data** that has a **spatial or temporal** component.
- **Techniques from signal or image processing** can frequently be used to **reduce noise** and thus, help to **discover patterns** (signals) that might be "lost in the noise."
- **Data errors** can be the result of a **more deterministic phenomenon**, such as a **streak** in the same place on a set of photographs (**artefacts**).



Streak artefact caused by a satellite

Precision, Bias, and Accuracy

- In statistics and experimental science, the **quality of the measurement process** and the **resulting data** are measured by **precision** and **bias**.
  - **Precision**: The closeness of repeated measurements (of the same quantity) to one another.
  - **Bias**: A systematic variation of measurements from the quantity being measured.

- **Accuracy**, is common to refer to the **degree of measurement error in data**.
  - **Accuracy**. The closeness of measurements to the true value of the quantity being measured.

- **Accuracy** depends on **precision** and **bias**, but **there is no specific formula for accuracy** in terms of these two quantities.

- One important aspect of **accuracy** is the use of **significant digits**.

- The goal is to **use only as many digits to represent the result of a measurement or calculation** as are **justified by the precision of the data**.

Precision, Bias and Accuracy

- A standard laboratory weight with a mass of 1g
- We weigh the mass five times, and obtain the following five values: {1.015; 0.990; 1.013; 1.001; 0.986}.

$$mean = \frac{\Sigma x_i}{n} = 1.001$$

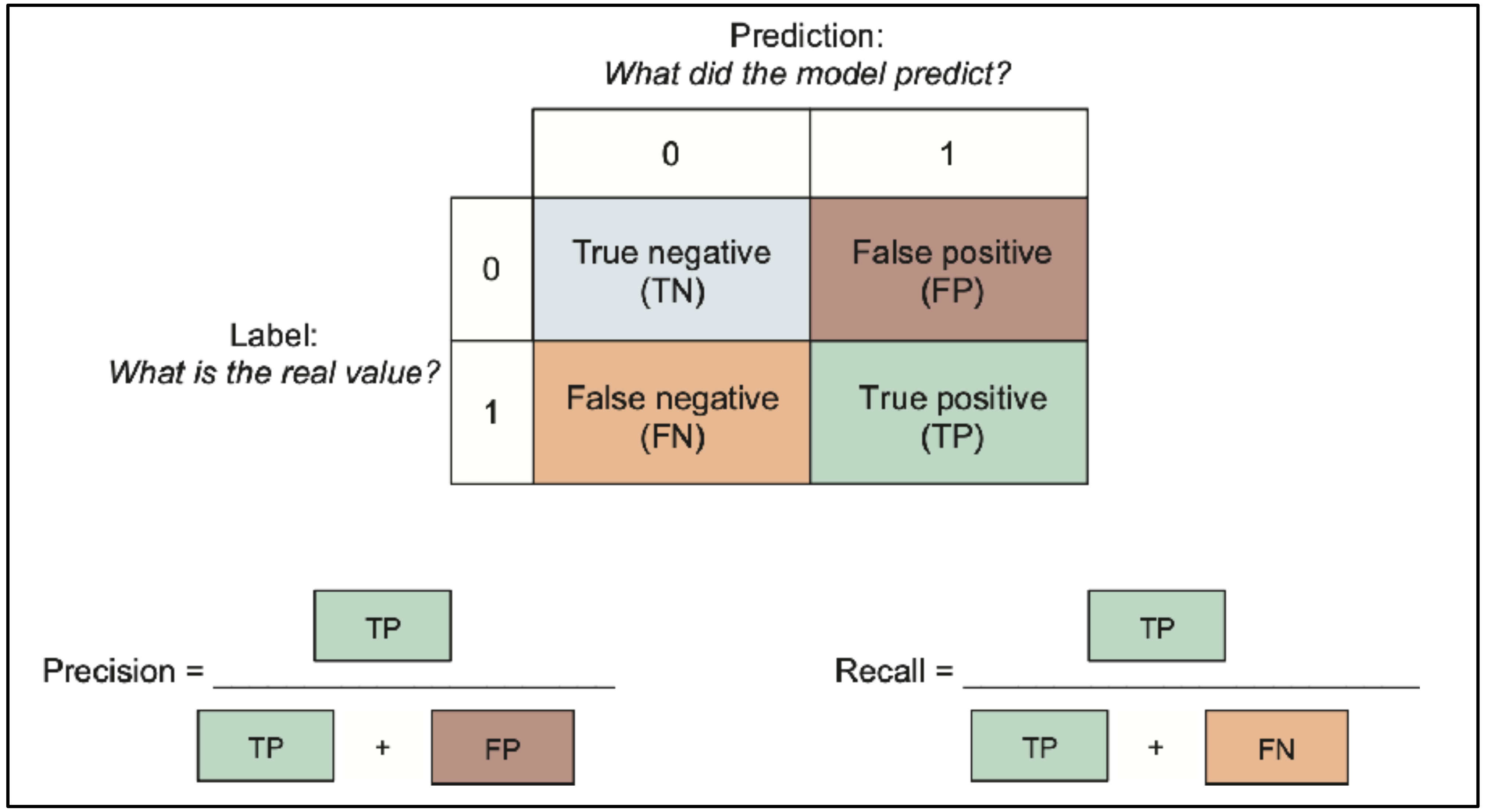
$$bias = 1.000 - 1.001 = 0.001$$

$$precision = \sigma = \sqrt{\frac{1}{N}\Sigma(x_i - \mu)^2} = 0.013$$

Precision (and Recall) in Classification Context

Visualising the performance of classification

Confusion Matrix



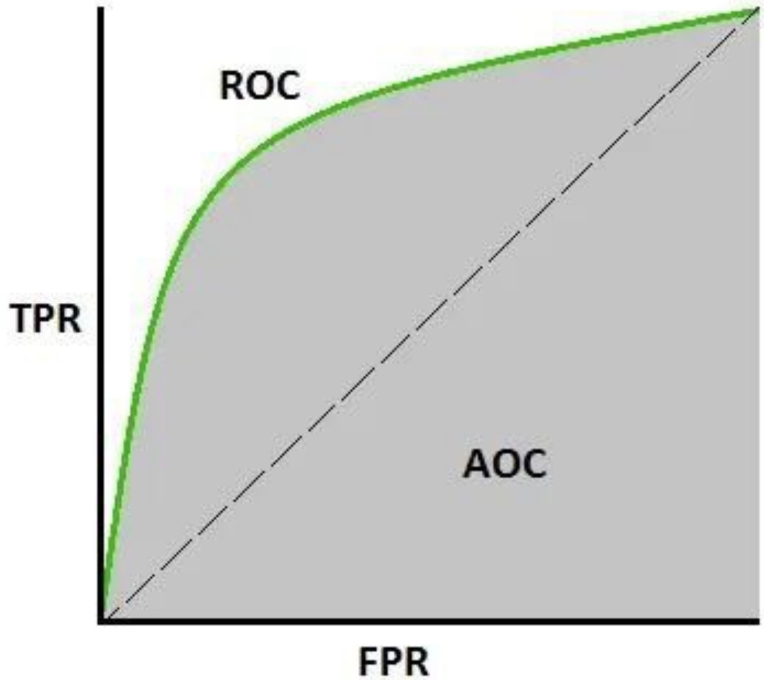
Prediction:
What did the model predict?

Label: *What is the real value?*		0	1
	0	True negative (TN)	False positive (FP)
	1	False negative (FN)	True positive (TP)

$$\text{Precision} = \frac{TP}{TP + FP}$$

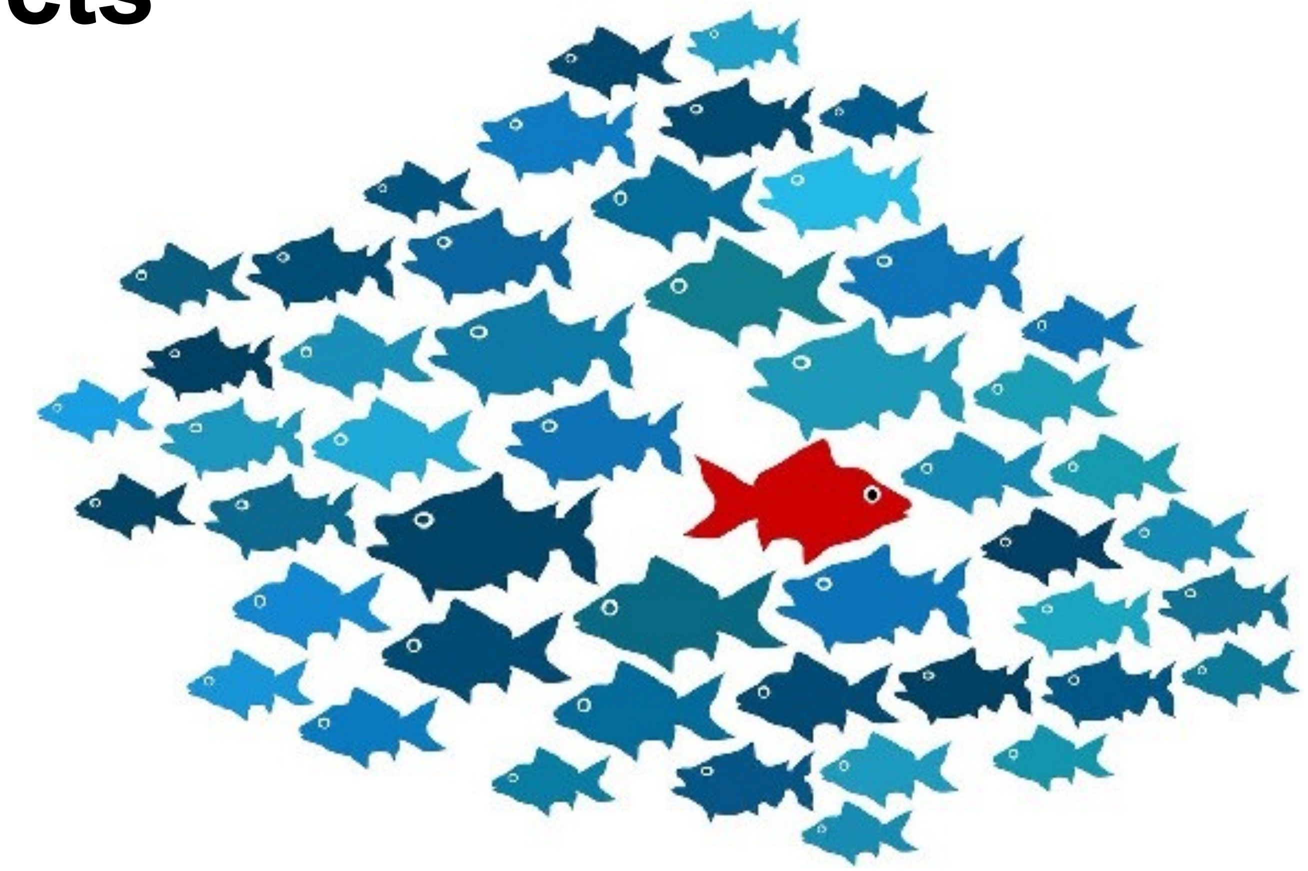
$$\text{Recall} = \frac{TP}{TP + FN}$$

AUC - ROC Curve



Outliers

- Outliers are either
  - **data objects** that have **characteristics that are different** from most others, or
  - **values of an attribute** that are **unusual** with respect to the typical values.
- Alternatively, they can be referred to as **anomalous objects** or **values**.
- Unlike noise, **outliers** can be **legitimate data objects** or **values** that we are interested in detecting.



Missing Values

- In some cases, the information was **not collected** or some attributes are **not applicable to all objects**.

- There are several strategies for dealing with missing data:

  a) **Eliminate Data Objects or Attributes**. This should be done with caution, however, because the eliminated attributes may be the ones that are critical to the analysis.

  b) **Estimate Missing Values**.

  - If the attribute is **continuous**: the **average attribute value of the nearest neighbours**.
  - if the attribute is **categorical**: the **most commonly occurring** attribute value (**mode**).

  c) **Ignore the Missing Value during Analysis**. Many data mining approaches can be **modified to ignore missing values**.

Inconsistent Values

- Data can contain **inconsistent values**. Regardless of their cause, it is important to **detect** and, if possible, **correct** such problems.

- Some types of inconsistencies are easy to detect (*i.e. a negative person's height*).

- The correction of an inconsistency may be complex and require **additional** information.

- The analyst **should consider the potential impact of such discrepancies** on the data mining analysis.

Issues Related to Applications

- Data quality issues can also be considered from an application viewpoint. Some general issues to consider:

  a) **Timeliness**. Some data **starts to age** as soon as it has been collected. In particular, if the data provides a snapshot that may represent reality for only a limited time.

  b) **Relevance**. The available data must contain the information necessary for the application. The **objects** in a data set must also be relevant (*i.e., age & gender in car accident rate prediction; survey bias*).

  c) **Knowledge about the Data**. Ideally, data sets should be accompanied by **documentation** that describes **different aspects** of the data.

Data Preprocessing

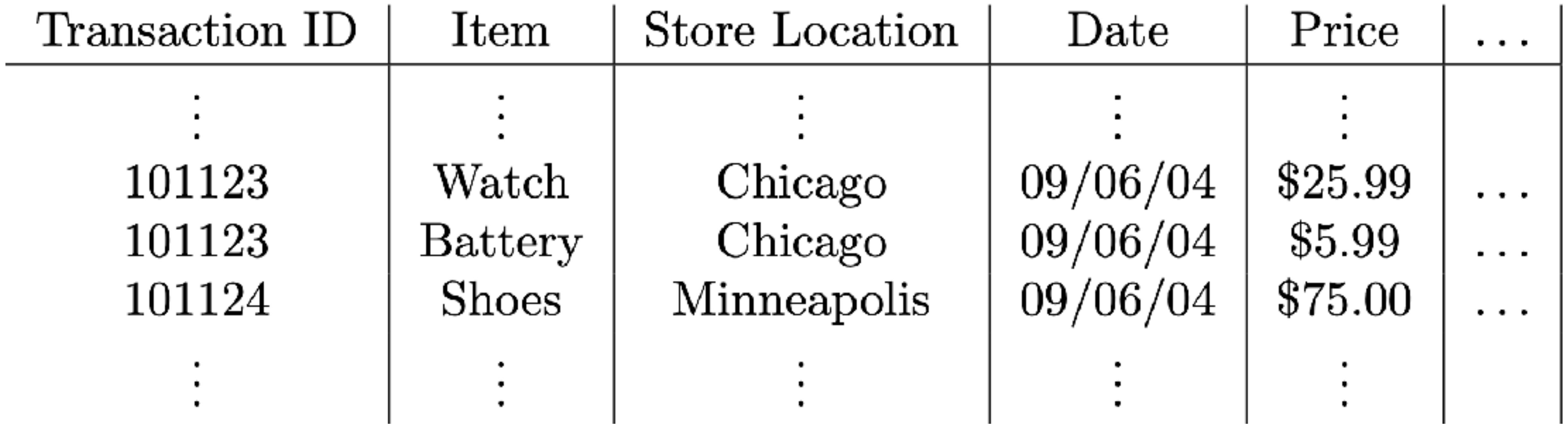
Data Preprocessing

- The goal of **data preprocessing** is to make the data more suitable for data mining.
- There are some different strategies and techniques:
    - *Aggregation*
    - *Sampling*
    - *Dimensionality reduction*
    - *Feature subset selection*
    - *Feature creation*
    - *Discretization and binarization*
    - *Variable transformation*
- There are two main categories. Selecting **data objects** and **attributes**:
    - a) for the **analysis** or
    - b) for **creating/changing the attributes**.

Aggregation

- Aggregation is the **combination of two or more objects** into a **single object**.
  - **Quantitative attributes**, are typically aggregated by taking a **sum** or an **average**.
  - **Qualitative attributes**, can either be **omitted or summarised** in terms of a **higher level category**.

- Therefore, aggregation is the process of **eliminating attributes**, or **reducing the number of values** for a particular attribute.



Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

Data set containing information about customer purchases

Aggregation

- There are **several motivations for aggregation**:
  - The use of **less expensive data mining algorithms** (**less memory and processing time**).
  - Change of **scope or scale** by providing a **high-level view of the data** instead of a **low-level view**.
  - The **behaviour of groups of objects or attributes** is often **more stable than that of individual objects or attributes** (lesser variability).

- A **disadvantage** of aggregation is the **potential loss of interesting details**.

Sampling

- **Sampling** is a commonly used approach for selecting a subset of the data objects to be analysed.
- The **motivations** for sampling in **statistics** and **data mining** are often **different**.
- A sample must be **representative**. That is, have **approximately the same property (of interest)** as the original set of data.
- Choosing the appropriate **sample size** and **sampling technique** are key to guarantee a high probability of getting a representative sample.

Sampling Approaches

- The simplest type of sampling is simple **random sampling**: **equal probability** of selecting any particular object.
  - **sampling without replacement**—as each object is selected, it is removed from the set of all objects,
  - **sampling with replacement**—objects are not removed from the population as they are selected.

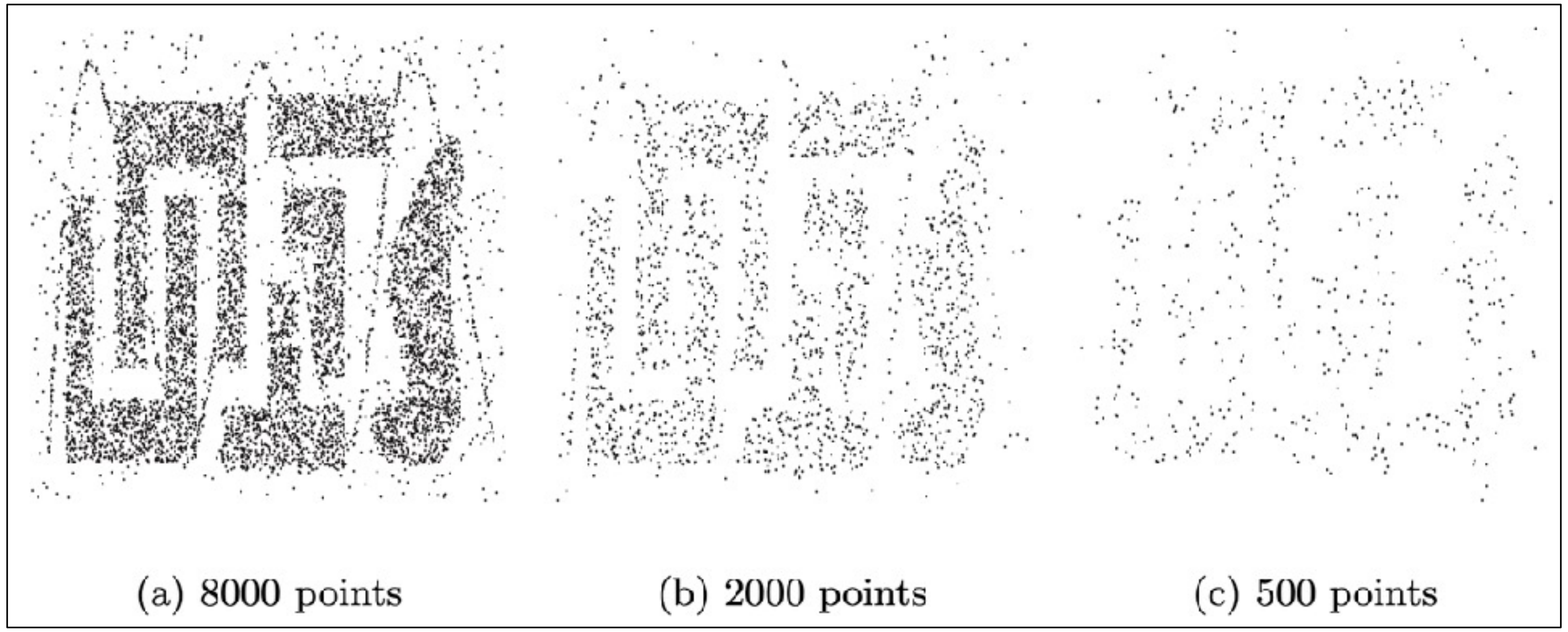
- When the population consists of **different types of objects** that are **not equally represented**, simple **random sampling** can **fail**.

- **Stratified sampling**: **equal numbers of objects** are drawn from **each group** even though the groups are of different sizes.

Sample Size

- Once a sampling technique has been selected, it is still necessary to determine the proper sample size.

- There is a **trade-off** between **larger sample** and **smaller sample sizes**.



Example of the loss of structure with sampling

Proper Sample Size

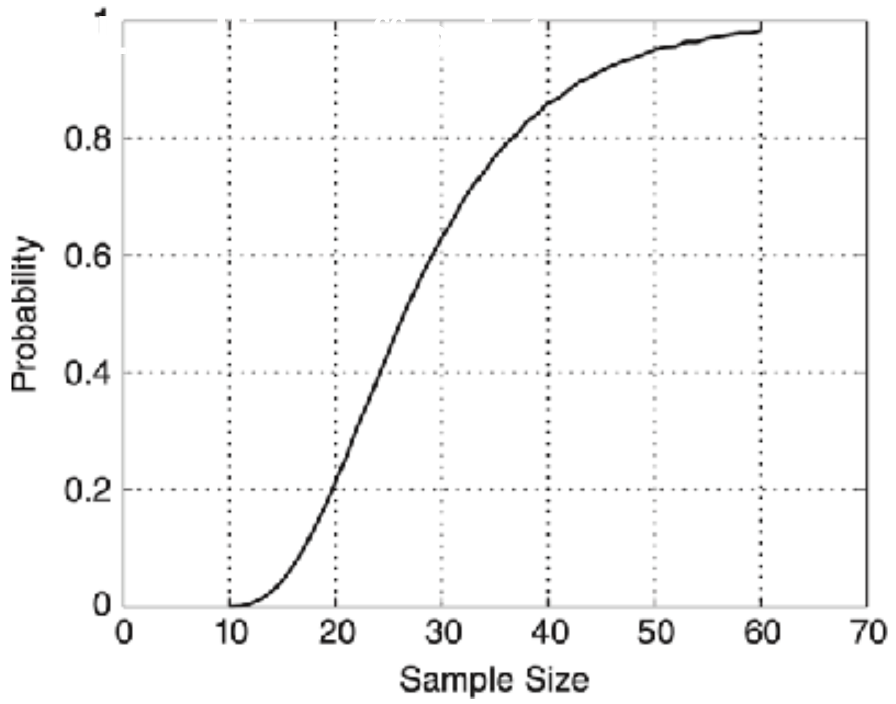
- One approach is to take a **small sample** of data points, **compute the pairwise similarities between points**, and then **form groups of points that are highly similar**.

- The **desired set of representative points** is then obtained by taking **one point from each of these groups**.

- The goal is to **determine a sample size** in which **at least one point will be obtained from each cluster**.

- Alternatives exist to eliminate the need to initially determine the correct sample size (**progressive sampling**).

Ten groups of points



Probability a sample contains points from each of 10 groups

Dimensionality Reduction

- Data sets can have a **large number of features**.
- There are a variety of **benefits to dimensionality reduction**:
  - Many data mining algorithms work better if the dimensionality is **lower**. This is partly because **dimensionality reduction** can **eliminate irrelevant features** and **reduce noise** and partly because of the curse of dimensionality.
  - **More understandable models** because the model usually involves fewer attributes.
  - The **amount of time** and **memory** required by the **data mining algorithm** is reduced.
- There are two main approaches to reduce the dimensionality:
  - by **creating new attributes** that are a **combination of the old attributes** (**dimensionality reduction**)
  - by **selecting attributes** that are a **subset of the old** (**feature subset selection**)
- Most common approaches for dimensionality reduction, particularly for **continuous data**, use techniques from linear algebra such as **principal components analysis** (**PCA**).

Feature Subset Selection

- Another way to reduce the dimensionality is to use only a subset of the features, removing those that are redundant and irrelevant.
  - **Redundant features** duplicate much or all of the information contained in one or more other attributes.
  - **Irrelevant features** contain almost no useful information for the data mining task at hand.
- Some irrelevant and redundant attributes can be eliminated immediately by using **common sense** or **domain knowledge**.
- Selecting the **best subset of features** frequently requires a **systematic approach**.
- The **ideal approach** is to try all possible subsets of features as input to the data mining algorithm of interest, and then take the subset that produces the best results. Unfortunately is impractical in most situations. (number of possible subsets is 2^n).
- There are three standard approaches to feature selection: **embedded**, **filter**, and **wrapper**.

Feature Subset Selection

- **Embedded approaches**. Feature selection occurs naturally as part of the data mining algorithm. The **algorithm itself** decides which attributes to use and which to ignore (*i.e. algorithms for building decision tree classifiers*).

- **Filter approaches**. Features are selected using some approach that is independent of the data mining task (*i.e. sets of attributes whose pairwise correlation is as low as possible*).

- **Wrapper approaches**. These methods use the target data mining algorithm as a black box to find the best subset of attributes.

- **Feature Weighting** is an **alternative** to keeping or eliminating features. **More important features** are assigned a **higher weight**, while less important features are given a lower weight.

Feature Creation

- A **new set of attributes** that captures the **important information** much more **effectively** can be created from the **original attributes**.

- Two related methodologies:

  - **Feature extraction** is a **complex** and **highly domain-specific** approach that enables the creation of a new set of features from the original raw data. It requires domain expertise and theses techniques have limited **applicability to other fields** (i.e. *BMI*)

  - **Mapping the data to a new space**. A different point of view of the data can reveal important and interesting features. Transformations such as *Fourier transform* and the *wavelet transform* have proven to be very useful for time series and other types of data.

Discretization and Binarization

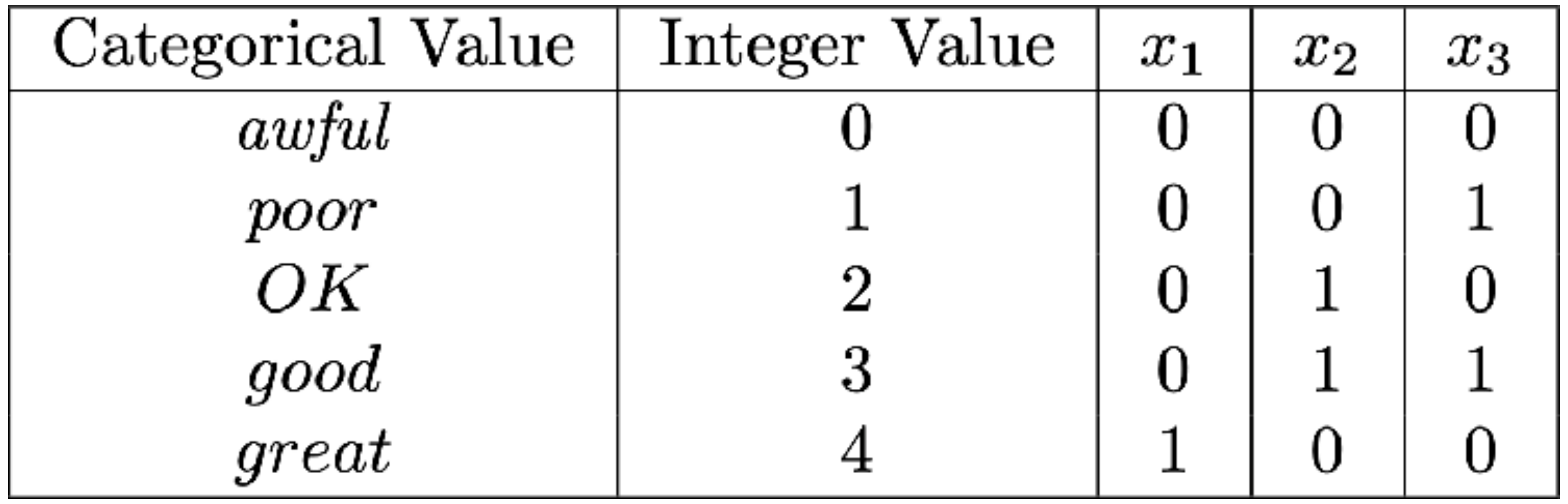
- Some data mining algorithms require some kind of transformation.
  - Certain **classification algorithms** require that the data be in the form of **categorical** attributes. **Discretization**: continuous → categorical attribute.
  - **Association analysis** requires that the data be in the form of **binary** attributes. **Binarization**: continuous/discrete → one or more binary attributes.

- If a **categorical attribute** has a **large number of values** (categories), or some values **occur infrequently**, **reducing** the number of categories combining some of the values is beneficial.

Binarization

- A simple technique to binarize a categorical attribute is the following:
  - If there are **m categorical values**, then uniquely assign each original value to an integer in the interval [0, m − 1].
  - If the attribute is **ordinal**, then **order** must be maintained by the assignment.

$$n = [log_2(m)]$$



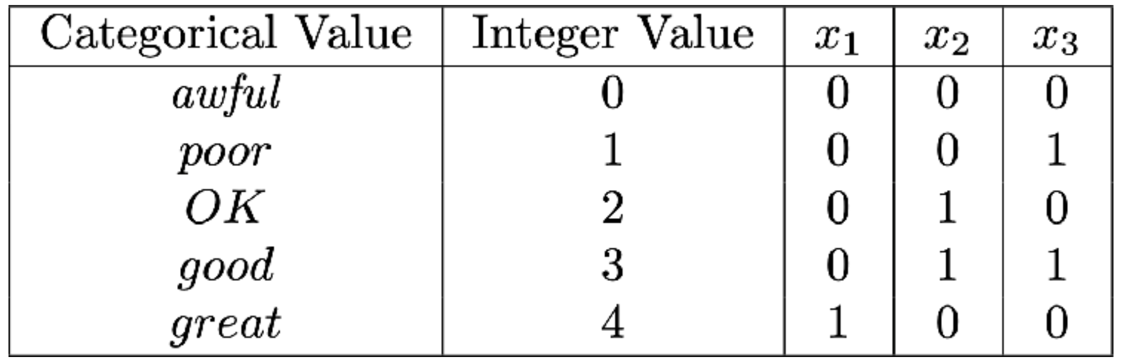
Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

  - Such a transformation can create **unintended relationships** among the transformed attributes.

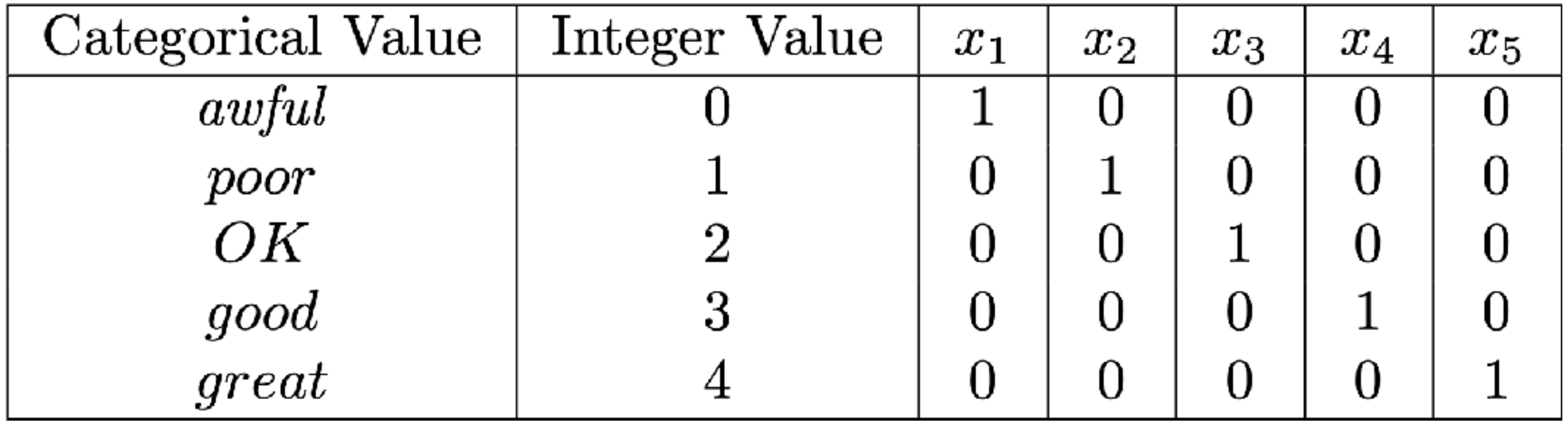
Binarization

- **Association analysis** requires **asymmetric binary attributes**, where only the presence of the attribute (non zero) is important.

- It is therefore necessary to introduce **one asymmetric binary** attribute for **each categorical** value.



Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0



Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

- For asymmetric binary attributes, the information representation is somewhat **inefficient**.

Discretization of Continuous Attributes

- **Discretization** is typically applied to **attributes** that are used in **classification** or **association analysis**.

- Transformation of a **continuous attribute** to a **categorical attribute** involves **two subtasks**:

  a) Deciding the **number of categories** (n) to have, and

  b) Determining **how to map the values of the continuous attribute** to these categories.

- In the first step, after the values of the continuous attribute are sorted, they are then **divided into n intervals** by specifying **n−1 split points**.

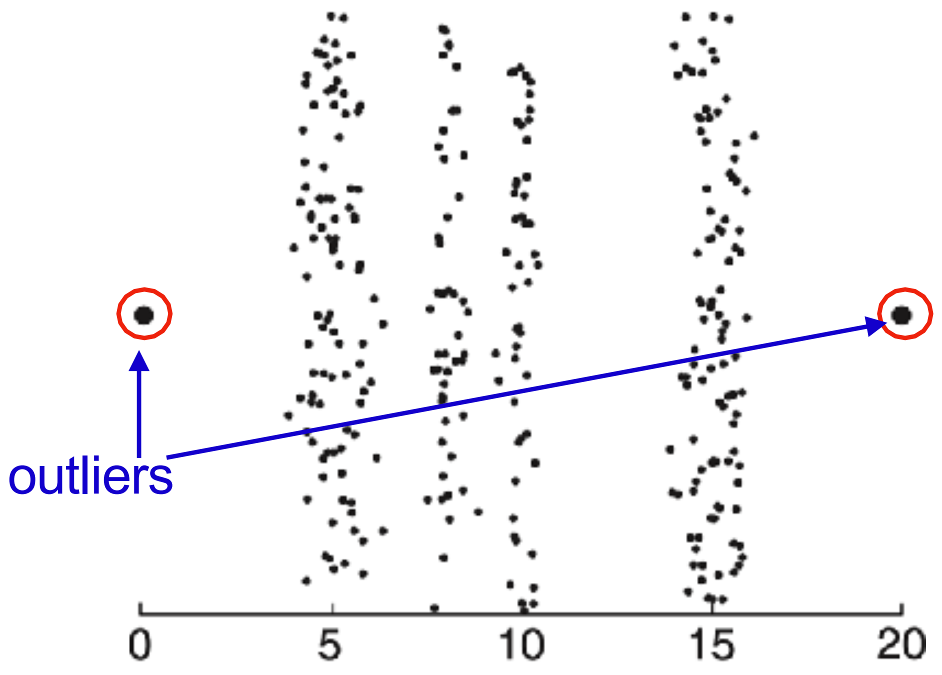
- In the second, all the **values in one interval** are **mapped** to the **same categorical** value.

- The result can be represented as a set of intervals:

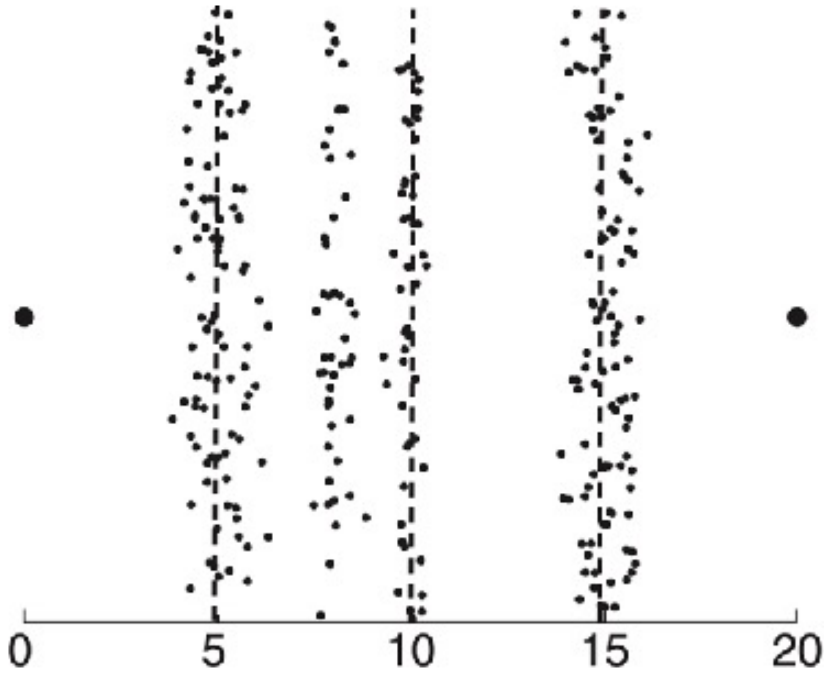
$\{(x_0, x_1], (x_1, x_2], \ldots, (x_{n-1}, x_n)\}$, where x_0 and x_n can be -∞ or $+\infty$.

Unsupervised Discretization

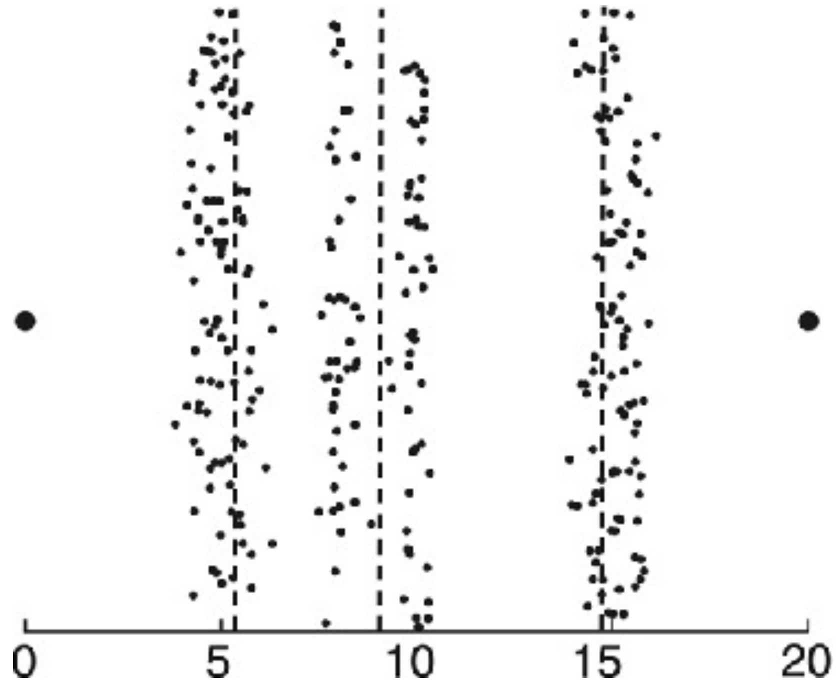
- **Discretization** methods for **unsupervised classification**:
    - **Equal width** approach divides the **range of the attribute** into a **user-specified number of intervals** each having the **same width**. This approach can be **badly affected by outliers**.
    - **Equal frequency** approach, which tries to put the same number of objects into each interval, is often preferred.
    - A **clustering method**, (i.e. K-means).
    - **Visually inspecting** the data can sometimes be an effective approach.



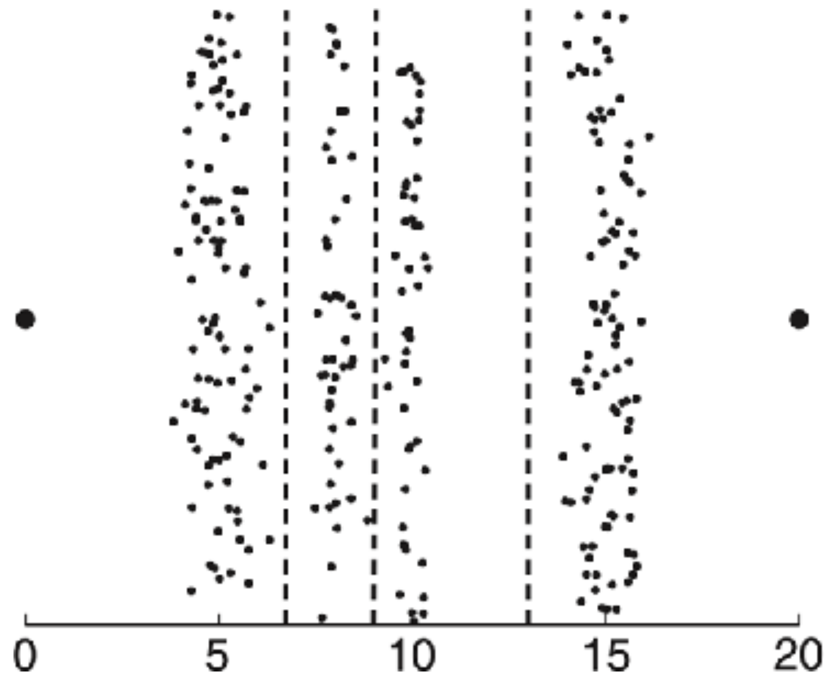
Original data



Equal width discretization



Equal frequency discretization



K-means discretization

- Overall, the best discretization will **depend on the application** and often involves **domain-specific** discretization.

Categorical Attributes with Too Many Values

- **Categorical attributes** can sometimes have **too many values**.
  - If the **categorical** attribute is an **ordinal** attribute, then techniques similar to those for continuous attributes can be used to reduce the number of categories.
  - If the **categorical** attribute is **nominal**, then, domain knowledge may be an approach.
- If **domain knowledge** does **not serve as a useful guide** or **results in poor classification** performance, a **more empirical approach** might be necessary.

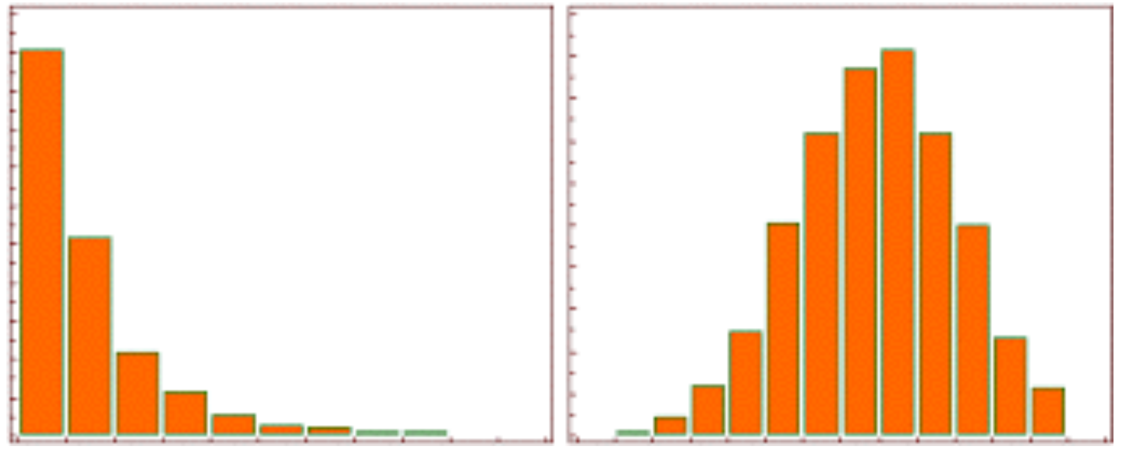
Variable Transformation

- A variable transformation refers to a **transformation** that is applied to **all the values of a variable**. Two important types of variable transformations:

1. **Simple functional transformations.** A simple **mathematical function** is applied to **each value** individually.

   If x is a variable, such transformations include: $x^k, logx, e^x, \sqrt{x}, 1/x, sinx, or\,|x|$.

- In statistics, $\sqrt{x}, logx,$ and $1/x$ are often used to transform into **Gaussian (normal) distributions**.



X and log x

Variable Transformation

- In data mining, these functions can be used to compress variables with a huge range of values. *i.e. 1 billion bytes* $= 10^9, log_{10}10^9 = 9$

- Variable transformations **change the nature of the data**. *e.g. the transformation* ***1/x****:*
  - **reduces the magnitude** of values that are **1 or larger**,
  - but **increases the magnitude** of values **between 0 and 1**, reversing the original order.

2. **Normalisation or Standardisation**. The goal is to make an entire set of values have a particular property.

   If $\bar{x}$ is the mean and σ_x is the standard deviation, then the transformation $x' = \dfrac{(x - \bar{x})}{\sigma_x}$ creates a new variable x' that has a **mean of 0** and a **standard deviation of 1**.

- If there are **outliers** then *mean* can be replaced by **median.**

Measures of Similarity and Dissimilarity

Basic Concepts

- **Similarity** and **dissimilarity** are used by a number of data mining techniques, such as *clustering*, *nearest neighbour classification*, and *anomaly detection*.

- **Proximity** is used to refer to either **similarity** or **dissimilarity**.

- The **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).

- The **dissimilarity** between two objects is a **numerical measure** of the **degree** to which the **two objects are different**. **Dissimilarities** are **lower** for more similar pairs of objects.

- **Dissimilarities** sometimes fall in the **interval [0, 1]**, but it is also common for them to range from 0 to .

Transformations

- Transformations are often applied:

  1) to **transform a proximity measure** to fall **within a particular range**, such as [0,1], or

  2) to **convert a similarity** to a **dissimilarity**, or **vice versa**, or

1) Transform a proximity measure to the interval [0,1]

a) **Transforming similarities** with a **finite range:**

- $s' = (s - min\_s)/(max\_s - min\_s)$**,** where $\max\_s$ and min_s are the maximum and minimum similarity values, respectively.

  - i.e. Transform the similarities between objects that **range** from **1** to **10** to [0,1]:

  - For s = 1, $s' = (1 - 1)/9 = 0$; for s = 2, $s' = (2 - 1)/9 = 0.11$; …
    for s = 9, $s' = (9 - 1)/9 = 0.89$; for s = 10, $s' = (10 - 1)/9 = 1$

- Likewise, for **dissimilarity** measures: $d' = (d - min\_d)$ / $(max\_d - min\_d)$.

Transformations

b) **Transforming measures** when proximity measures take in the interval $[0,\infty)$:

- $d' = d/(1+d)$
  - i.e. Transform the dissimilarities 0, 0.5, 2, 10, 100, and 1000 to the interval [0,1]:
  - For d = 0, $d' = 0/(1+0) = 0$; for d = 0.5, $d' = 0.5/(1+0.5) = 0.33$; …
    for d = 100, $d' = 100/(1+100) = 0.99$; for d = 1000, $d' = 1000/(1+1000) = 0.999$

NOTE: Larger values on the original dissimilarity scale are compressed into the range of values near 1.

2) Transforming similarities to dissimilarities (and vice versa)

- If the similarity (or dissimilarity) falls in the interval [0,1]: **d = 1 - s** and **s = 1 - d**.
- If in a different range:

$$s = -d \qquad s = \frac{1}{d+1} \qquad s = e^{-d} \qquad s = 1 - \frac{d - min\_d}{max\_d - min\_d}$$

- i.e. *d = 0, 1, 10, 100*

s = 0, -1, -10, -100 *s = 1, 0.5, 0.09, 0.01* *s = 1.00, 0.37, 0.00, 0.00* *s = 1.00, 0.99, 0.90, 0.00*

Similarity and Dissimilarity between Simple Attributes

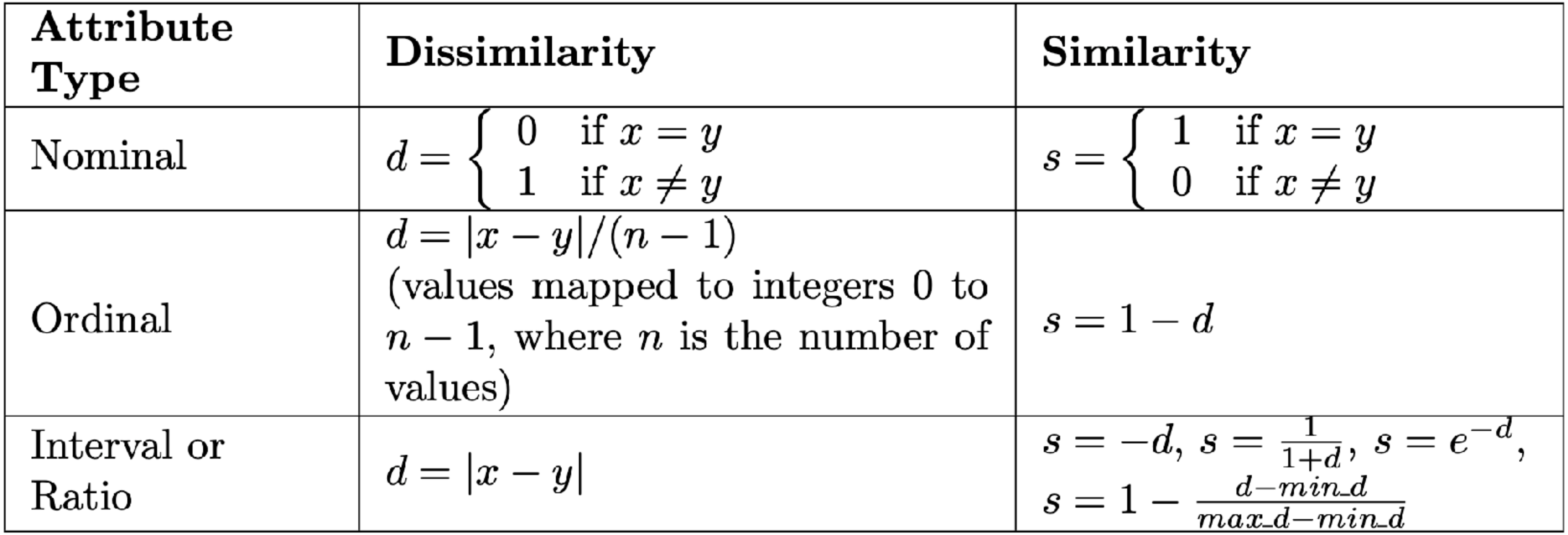
- The proximity of objects with a number of attributes is typically defined by combining the proximities of individual attributes.

- Consider objects described by **one nominal** attribute.
  - **Similarity** is traditionally defined as **1**, and as **0** otherwise (*dissimilarity* is the opposite).

- For objects with a **single ordinal** attribute, **information about order** should be **taken into account**. $d = |x - y| \ / \ n - 1$ (**n** number of elements mapped to integers 0 to n-1)
  - i.e. An attribute that measures the quality of a product: e.g., {*poor*, *fair*, *OK*, *good*, *wonderful*}.
  - Values can be mapped to integers (0 to n-1) {*poor=0*, *fair=1*, *OK=2*, *good=3*, *wonderful=4*}
  - If P1 is rated *wonderful*, P2 *good* and P3 *OK*, then d(P2, P3) = |3 - 2| / 5 - 1 = 0.25 [0,1]

- For **interval or ratio** attributes, the natural measure of dissimilarity between two objects is the absolute difference of their values. $d = |x - y|$

Similarity and Dissimilarity between Simple Attributes



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	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n - 1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between Data Objects

Euclidean Distance

- **Euclidean distance** is a proximity measure for objects with multiple attributes.
- Useful for **non-sparse** (dense) data, such as *time series* or *multi-dimensional* points.
- The Euclidean distance, **d**, between **two points**, x and y, in **one-, two-, three-, or higher-dimensional space**, is given by the following formula:

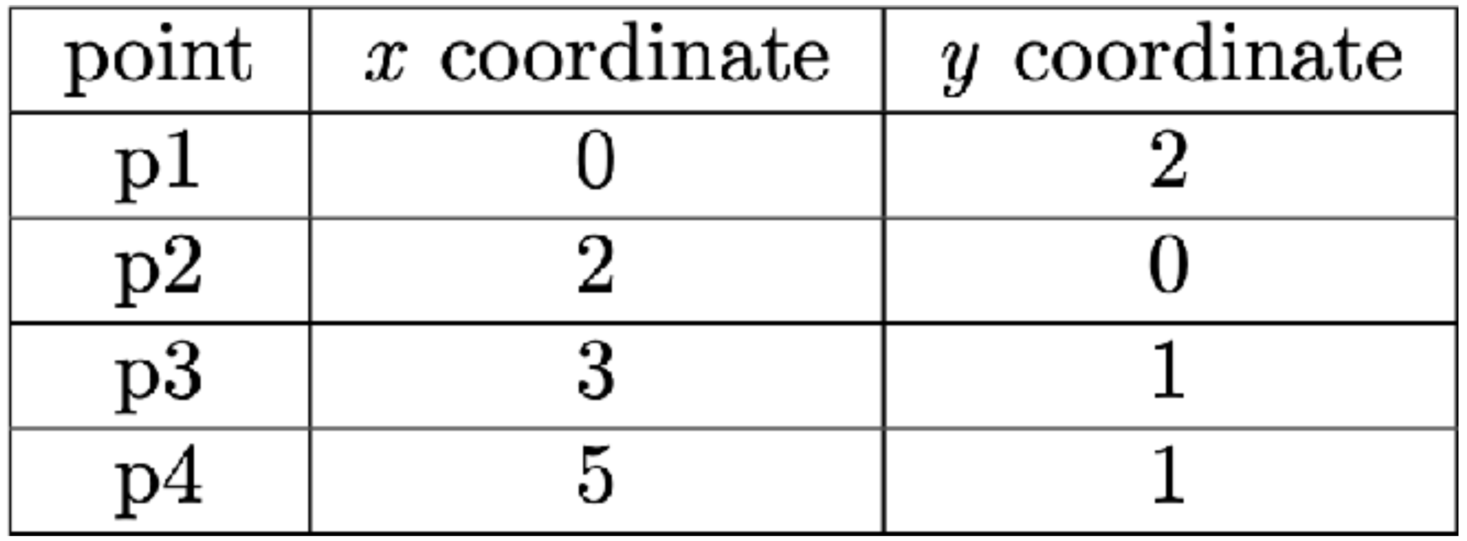
$$d(x, y) = \sqrt{\Sigma_{k=1}^{n}(x_k - y_k)^2},$$

where is the number of dimensions and x_k and y_k are, respectively, the k^{th} attributes (components) of and .

- The **greater the distance** between the objects, the **more different** they are.

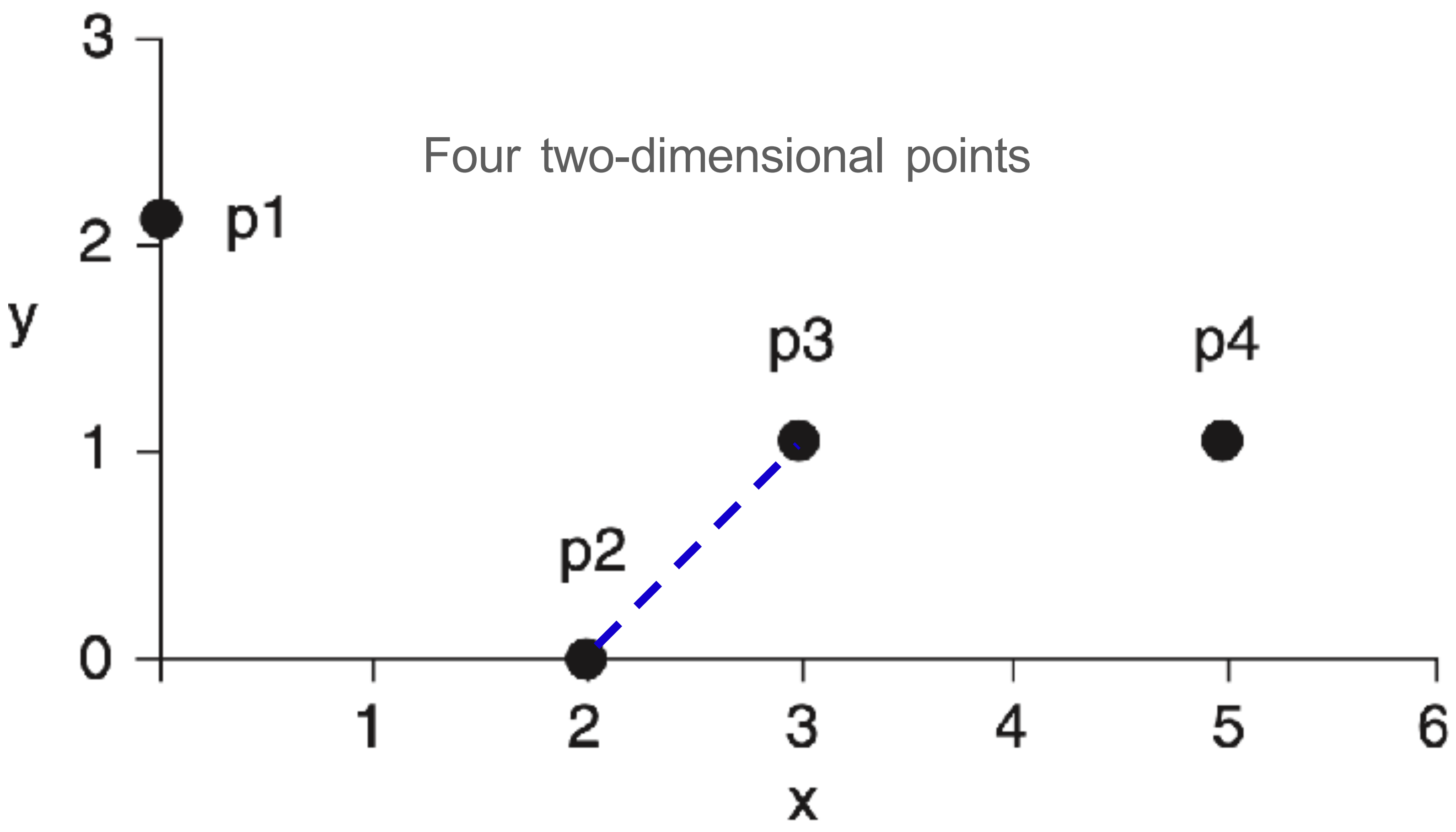
Euclidean Distance

- i.e. **Euclidean distance** between p2 and p3:



point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

and coordinates of four points



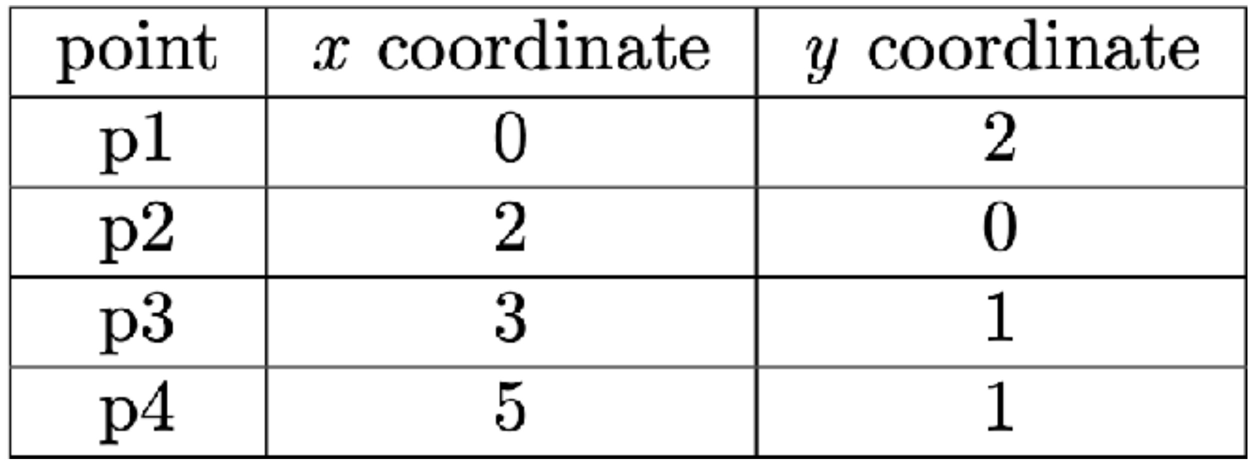
Four two-dimensional points

$$d(p2,p3) = \sqrt{\Sigma_{k=1}^{2}(p2_{xk} - p3_{xk})^2} = \sqrt{(p2_{x1} - p3_{x1})^2 + (p2_{x2} - p3_{x2})^2} = \sqrt{2} = 1.4$$

Euclidean Distance

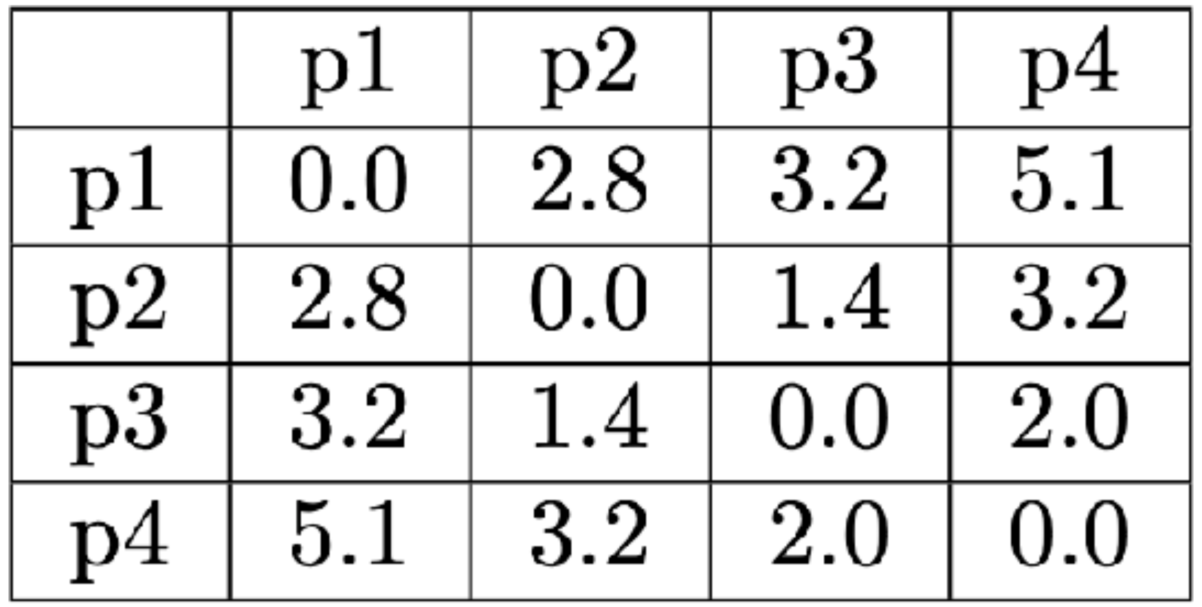
- i.e. **Euclidean distance** distant matrix:

  - $d(p1,p2) = \sqrt{(0-2)^2 + (2-0)^2} = 2.82$
  - $d(p1,p3) = \sqrt{(0-3)^2 + (2-1)^2} = 3.162$
  - $d(p1,p4) = \sqrt{(0-5)^2 + (2-1)^2} = 5.09$



point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

and coordinates of four points



	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Euclidean distance matrix

Minkowski Distance

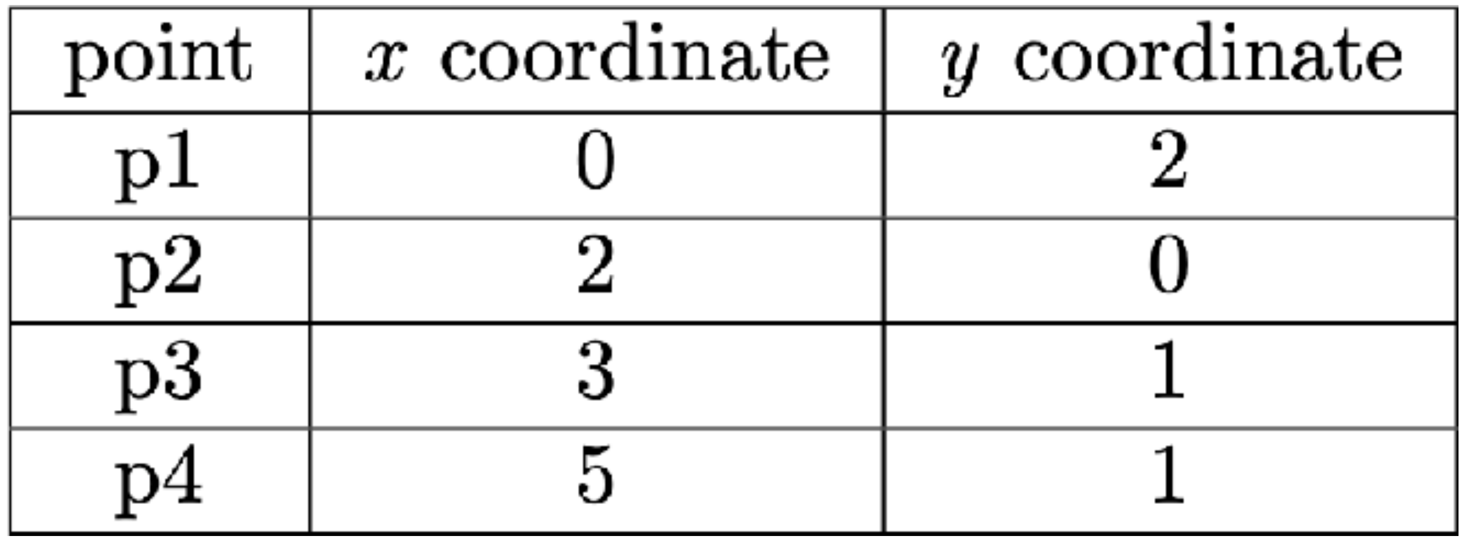
- The **Euclidean distance** measure is generalised by the **Minkowski distance** metric a proximity measure for objects with multiple attributes.

  $d(x, y) = (\Sigma_{k=1}^{n} |x_k - y_k|^r)^{1/r}$, where **r** is a parameter.

- The three most common examples of Minkowski distances:
  - **r = 1**. City block (Manhattan, taxicab, L_1norm) distance.
  - **r = 2**. Euclidean distance (L_2 norm).
  - **r =**∞ . Supremum (L_{max} or L_∞norm) distance.

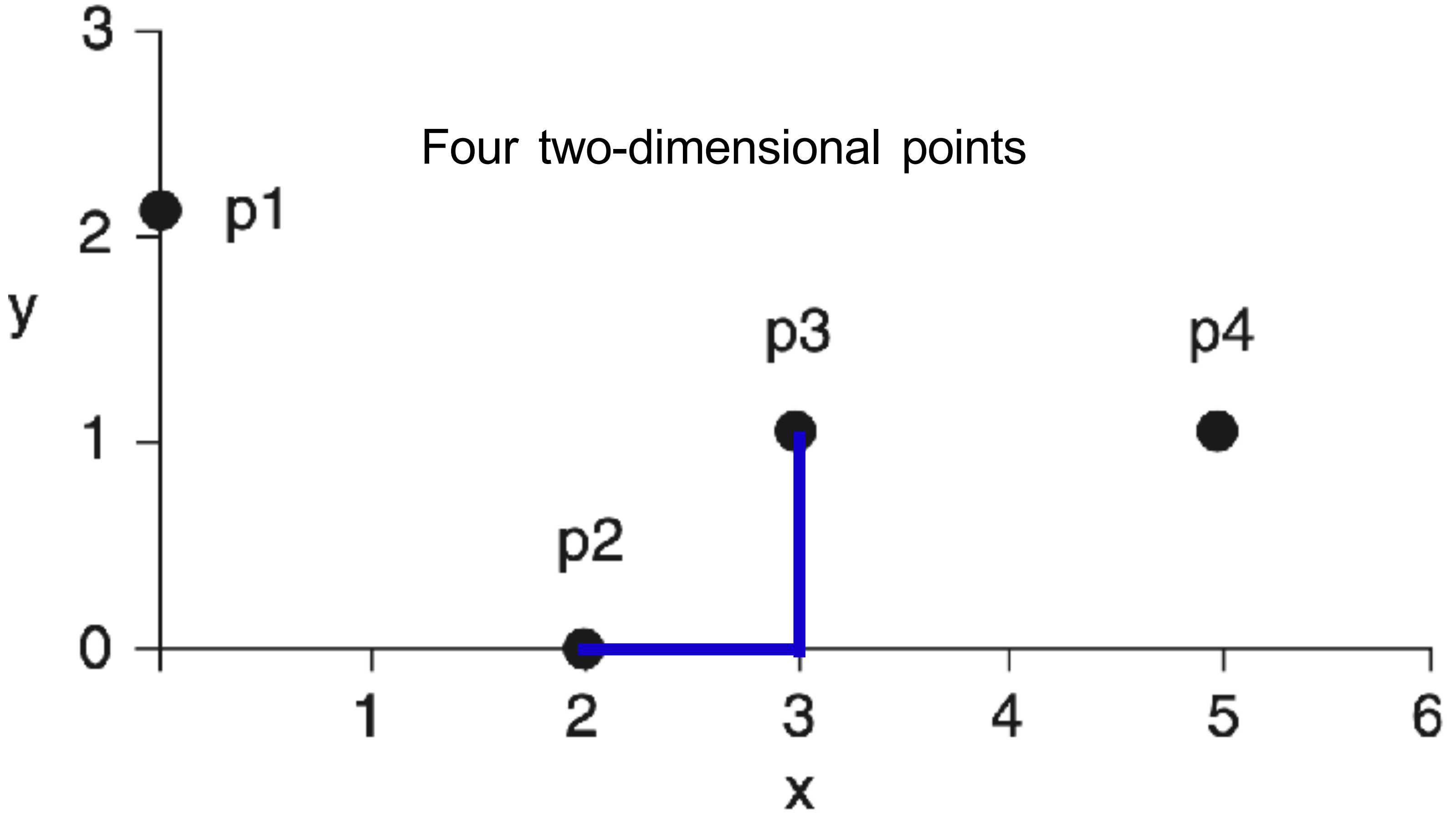
Manhattan Distance

- i.e. **Manhattan distance** between p2 and p3:



point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

and coordinates of four points

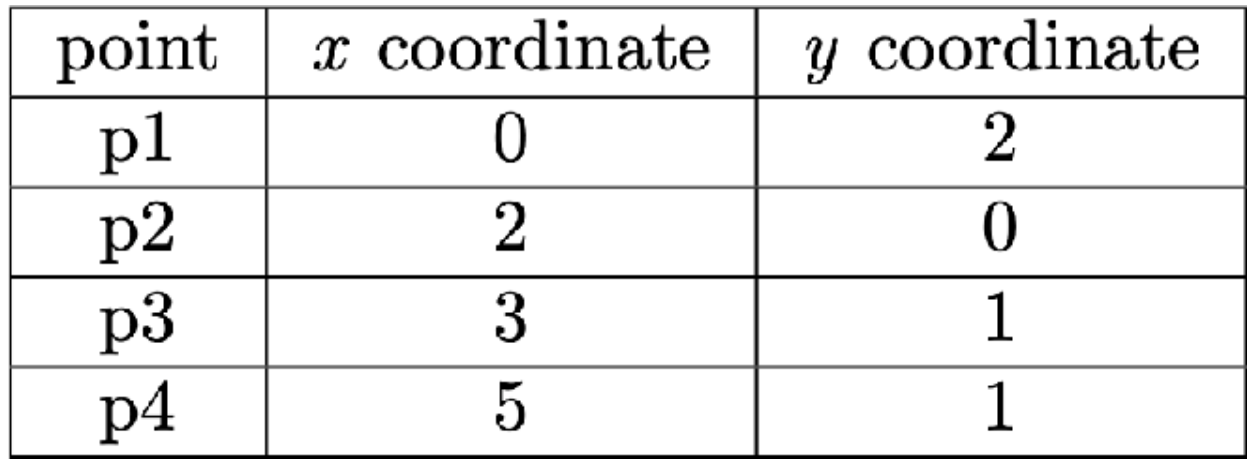


Four two-dimensional points

$$d(p2,p3) = \Sigma_{k=1}^{2} |p2_{xk} - p3_{xk}| = |p2_{x1} - p3_{x1}| + |p2_{x2} - p3_{x2}| = 2$$

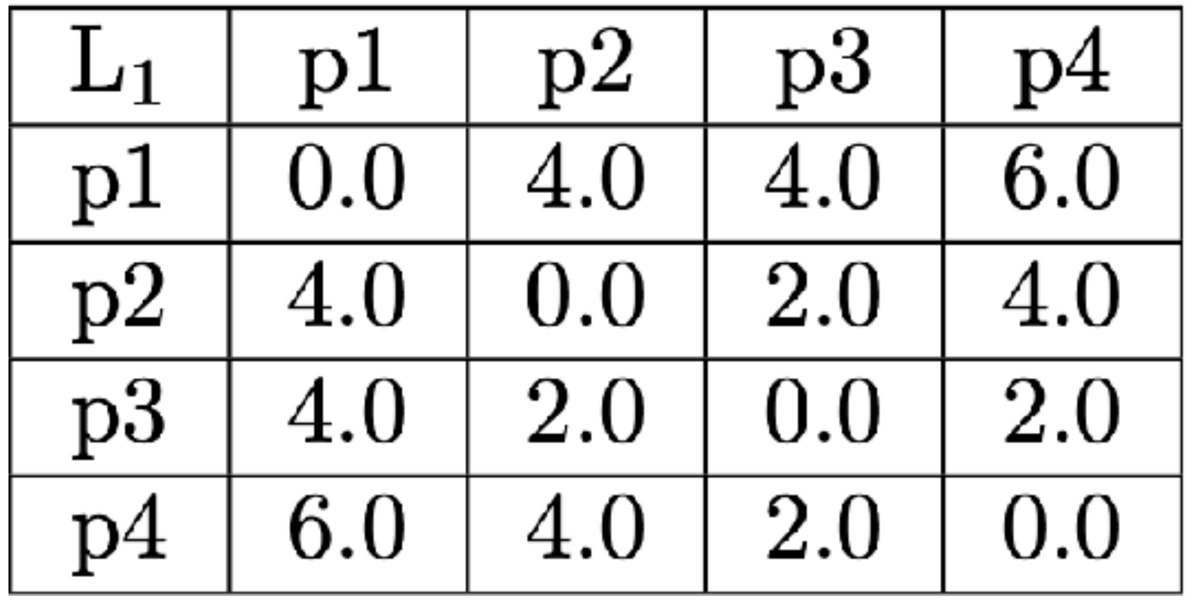
Manhattan Distance

- i.e. **Manhattan distance** distant matrix:
  - $d(p1,p2) = |0 - 2| + |2 - 0| = 4$
  - $d(p1,p3) = |0 - 3| + |2 - 1| = 4$
  - $d(p1,p4) = |0 - 5| + |2 - 1| = 6$



point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

and coordinates of four points



L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

Manhattan distance matrix

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**.
  - **1** indicates that the two objects are completely similar,
  - **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 (frequencies):
  - 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

- **Simple Matching Coefficient** One commonly used similarity coefficient is the simple matching coefficient (SMC), which is defined as:

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

- This measure counts both **presences** and **absences** equally.

Jaccard Coefficient

- The **Jaccard Similarity Coefficient** is frequently used to **handle objects** consisting of **asymmetric binary attributes**.

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

- This measure only focuses on the **non-zero** values.

The SMC vs Jaccard Similarity Coefficients

- i.e:
  - **x** = (**1**,0,0,0,0,0,0,0,0,0)
  - **y** = (0,0,0,0,0,0,**1**,0,0,**1**)

  - f_{01} = 2
  - f_{10} = 1
  - f_{11} = 0
  - f_{00} = 7

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$$

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0$$

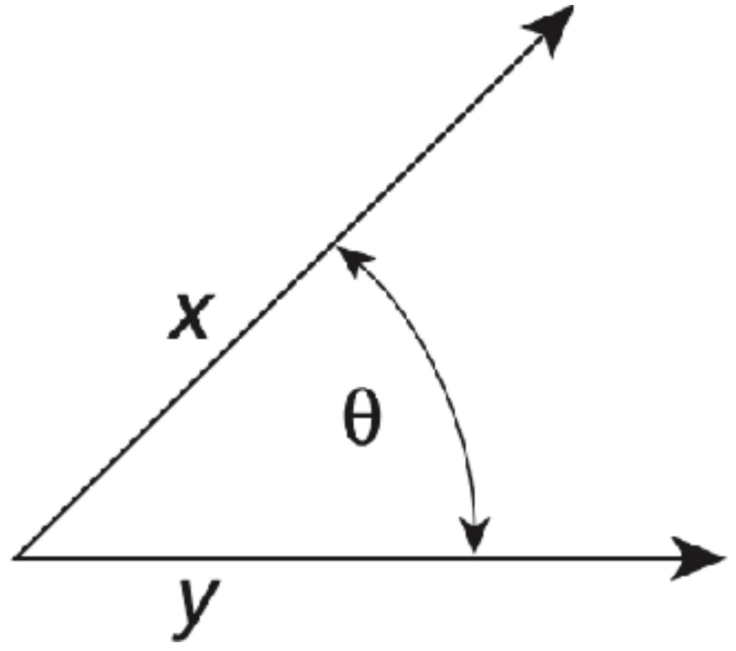
Cosine Similarity

- Documents are often **represented as vectors**, where **each component** (attribute) represents the **frequency** with which a **particular term** (word) **occurs in the document**.

- A similarity measure for documents needs to **ignores 0–0 matches** like the *Jaccard measure*, but also must be able to **handle non-binary vectors**.

- The **cosine similarity** is one of the most common measures of document similarity.
  - If x and y are two document vectors, then:

$$cos(x, y) = \frac{\langle x,y \rangle}{||x||\ ||y||} = \frac{x' y}{||x||\ ||y||} = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2}\sqrt{\sum_{k=1}^{n} y_k^2}}$$



Cosine Similarity

- i.e:
  - **x** = (**3**,**2**,0,**5**,0,0,0,**2**,0,0)
  - **y** = (**1**,0,0,0,0,0,0,**1**,0,**2**)

$$\langle \text{x,y} \rangle = 3 \times 1 + 2 \times 0 + 0 \times 0 + 5 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 2 = 5$$

$$||x|| = \sqrt{3 \times 3 + 2 \times 2 + 0 \times 0 + 5 \times 5 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.48$$

$$||y|| = \sqrt{1 \times 1 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 2 \times 2} = 2.45$$

$$cos(x, y) = \frac{5}{\text{6.48 x 2.45}} = 0.31$$

- **Cosine similarity** is a measure of the **cosine** of the **angle** between **x** and **y**.
  - If the **cosine similarity** is **1**, the **angle** between **x** and **y** is **0**°, and **x and y** are **the same** except for length.
  - If the **cosine similarity** is **0**, the **angle** between **x** and **y** is **90**°, and they **do not share any terms** (words).

Correlation

- Correlation is frequently used to measure the **linear relationship** between **two sets of values** that are observed together.

- Correlation can measure the **relationship** between **two variables** (height and weight) or between **two objects** (a pair of temperature time series).

- Correlation is used much more frequently to measure the **similarity between attributes** since the **values in two data objects** come from **different attributes**, which can have **very different** attribute types and scales.

- There are many types of correlation, we will focus on a measure appropriate for **numerical values**.

Pearson’s Correlation

- **Pearson’s correlation** measures the **correlation** between two sets of numerical values, i.e., two vectors, x and y:

$$corr(x, y) = \frac{covariance(x, y)}{standard\_deviation(x) \text{ x } standard\_deviation(y)} = \frac{S_{xy}}{S_x \, S_y}$$

$$covariance(x, y) = S_{xy} = \frac{1}{n-1} \Sigma_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$\bar{x} = \frac{1}{n} \Sigma_{k=1}^{n} x_k \text{ } is \text{ } the \text{ } mean \text{ } of \text{ } x$$

$$stardard\_deviation(x) = S_x = \sqrt{\frac{1}{n-1} \Sigma_{k=1}^{n} (x_k - \bar{x})^2}$$

- Correlation is always in the **range −1 to 1**.

Pearson's Correlation

- **Perfect correlation**. A correlation of **1** (**−1**) means a **perfect positive** (**negative**) linear relationship. i.e.
  - **x** = (-3, 6, 0, 3, -6)
  - **y** = (1,−2,0,−1, 2)

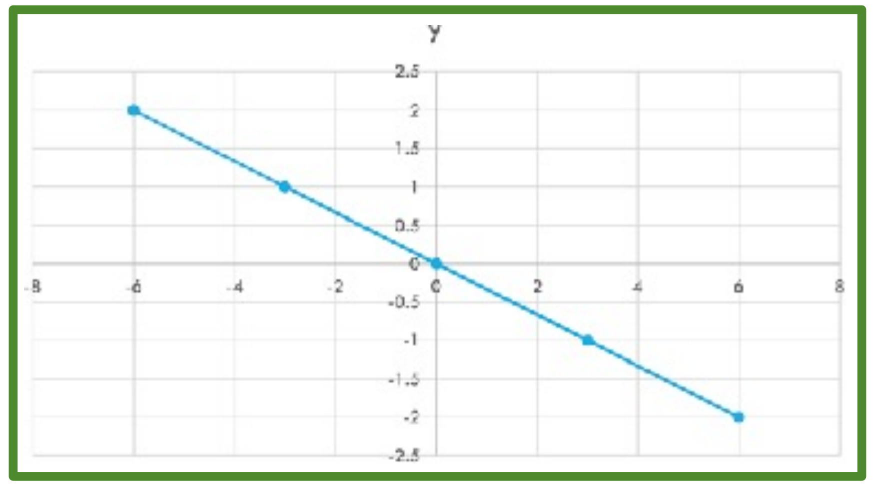
$$\bar{x} = \frac{1}{5} \text{ x } (-3 + 6 + 0 + 3 + (-6)) = 0 \qquad \bar{y} = \frac{1}{5} \text{ x } (1 + (-2) + 0 + (-1) + 2) = 0$$

$$S_{xy} = \frac{1}{5-1} \text{ x } ((-3-0) \times (1-0)) + ((6 - 0 \times -2 - 0)) + ((0-0)x(0-0)) + ((3-0) \times (-1-0)) + ((-6-0) \times (2-0)) = -7.5$$

$$S_x = \sqrt{\frac{1}{5-1}((-3-0)^2 + (6-0)^2 + (0-0)^2 + (0-0)^2 + (3-0)^2 + (-6-0)^2)} = 4.74$$

$$S_y = \sqrt{\frac{1}{5-1}((1-0)^2 + (-2-0)^2 + (0-0)^2 + (0-0)^2 + (-1-0)^2 + (2-0)^2)} = 1.58$$

$$corr(x, y) = \frac{S_{xy}}{S_x\, S_y} = \frac{-7.5}{4.74 \text{ x} 1.58} = -1$$

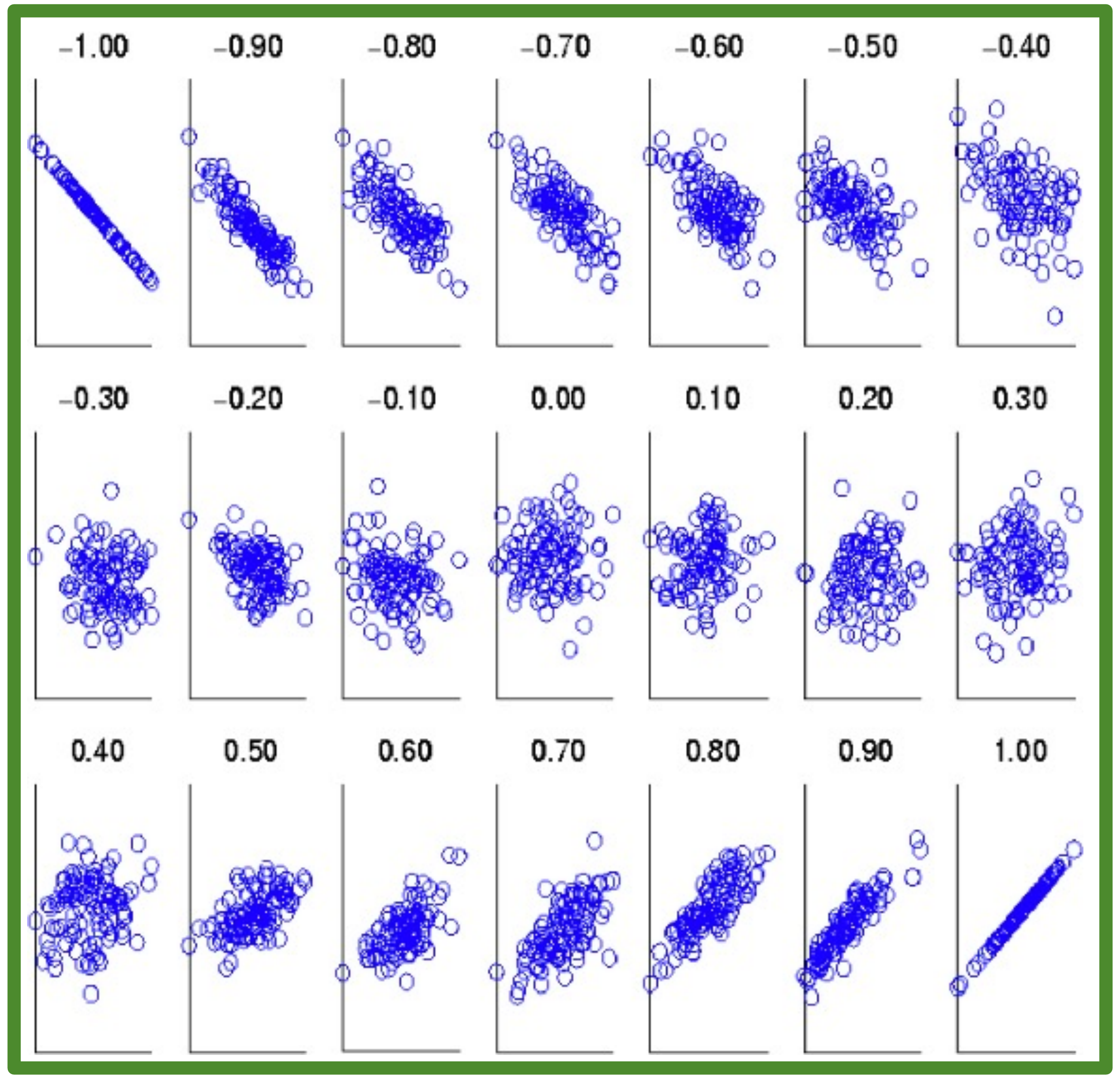


- **Nonlinear Relationships**. If the correlation is **0**, then there is **no linear relationship** between the two sets of values. i.e.

Visualising Correlation

- It is also easy to judge the correlation between two vectors x and y by **plotting pairs** of corresponding values of x and y in a scatter plot.

- The correlation of **x** and **y** ranges from **−1** to **1**.



x and y consist of a set of 30 pairs of values that are randomly generated (with a normal distribution)

Module Schedule

1. Introduction to Data Mining
2. Data
3. Association Analysis
4. Cluster Analysis
5. Classification
6. Naïve Bayes Classifier
7. Anomaly Detection
8. Information Retrieval
9. Exploratory Data Analysis
10. Data Mining Methodology
11. Different Approaches
12. Catch-up, Questions