CS685/785 Foundation of Data Science

Lecture 2: Data Fundamentals

Xi Li

Fall 2024

Table of Content

- 2.1 Attributes and Objects
- 2.2 Types of Data
- 2.3 Data Quality
- 2.4 Similarity and Distance
- 2.5 Data Preprocessing
- 2.6 Data Visualization

2.1 Attributes and Objects

Q: What is Data?

Name	Eye Color	Race	Age	Job	Height
Alice	Blue	White	28	Engineer	5'6"
Bob	Green	Hispanic	34	Architect	5'10"
Clara	Brown	Asian	45	Professor	5'5"
Dave	Hazel	Black	22	Student	6'0"
Eve	Grey	Spanish	30	Artist	5'7"

2.1 Attributes and Objects

Q: What is Data?

- Dataset: A collection of data **objects** and their **attributes**.

Attributes

Objects

Name	Eye Color	Race	Age	Job	Height
Alice	Blue	White	28	Engineer	5'6"
Bob	Green	Hispanic	34	Architect	5'10"
Clara	Brown	Asian	45	Professor	5'5"
Dave	Hazel	Black	22	Student	6'0"
Eve	Grey	Spanish	30	Artist	5'7"

2.1 Attributes and Objects

Q: What is Data?

- Dataset: A collection of data **objects** and their **attributes**.
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as *variable, field, characteristic, dimension, or feature*

Attributes

Objects

Name	Eye Color	Race	Age	Job	Height
Alice	Blue	White	28	Engineer	5'6"
Bob	Green	Hispanic	34	Architect	5'10"
Clara	Brown	Asian	45	Professor	5'5"
Dave	Hazel	Black	22	Student	6'0"
Eve	Grey	Spanish	30	Artist	5'7"

2.1 Attributes and Objects

Q: What is Data?

- Dataset: A collection of data **objects** and their **attributes**.
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Name	Eye Color	Race	Age	Job	Height
Alice	Blue	White	28	Engineer	5'6"
Bob	Green	Hispanic	34	Architect	5'10"
Clara	Brown	Asian	45	Professor	5'5"
Dave	Hazel	Black	22	Student	6'0"
Eve	Grey	Spanish	30	Artist	5'7"

Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- A **measurement scale** is a rule (function) that associates the attribute value with an attribute of an object.

Attributes v.s. Attribute Values

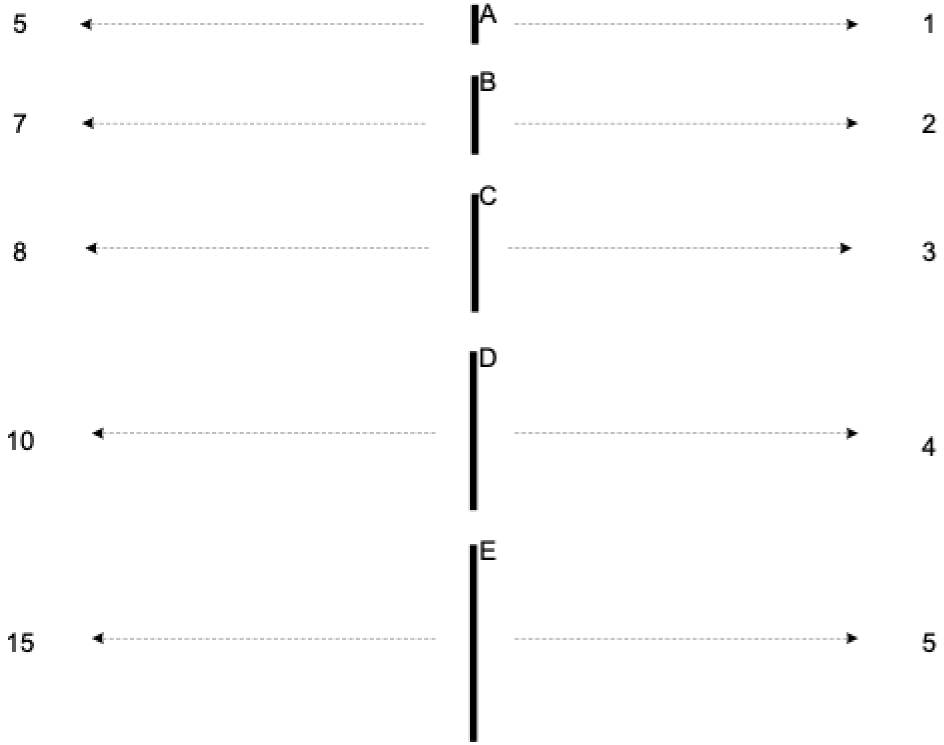
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers

Attributes vs. Attribute Values

- Distinction between attributes and attribute values
    - Same attribute can be mapped to different attribute values
        - Example: height can be measured in feet or meters
    - Different attributes can be mapped to the same set of values
        - Example: Attribute values for ID and age are integers
    - Properties of attribute can be different than the properties of the values used to represent the attribute

Example: Length of Line Segments

- All the line segments are multiples of the first.
- An attribute can be measured in a way that does not capture all the properties of the attribute.



This scale preserves only the ordering property of length.

This scale preserves both the ordering and additivity properties of length.

Types of Attributes

- **Nominal**: Data is categorized **without a specific order**.
  - Examples: ID numbers, eye color, zip codes
- **Ordinal**: Data is categorized **with a specific order** but **without consistent intervals**
  - Examples: rankings (e.g., rate from 1-10), grades, height {tall, medium, short}
- **Interval**: Data is **ordered**, and the **intervals** between each value are **equal**, but there is **no true zero point**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **Ratio**: Data is **ordered**, **intervals** are **equal**, and there is a **true zero point**, allowing for meaningful ratios between data points
  - Examples: temperature in Kelvin, height, weight

Properties of Attribute

- The type of an attribute depends on the following properties/operations:
  - Distinctness: $=$ and $\neq$
  - Order: $<, \leq, >,$ and $\geq$
  - Addition: $+$ and $-$
  - Multiplication: $*$ and $/$

Properties of Attribute

- The type of an attribute depends on the following properties/operations:
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & Addition
  - Ratio attribute: all 4 properties/operations

Difference Between Interval and Ratio

Is it physically meaningful to say that a temperature of 2° is twice that of 1° on

- the Celsius scale (Interval)?
- the Fahrenheit scale (Interval)?
- the Kelvin scale (Ratio)?

Difference Between Interval and Ratio

Is it physically meaningful to say that a temperature of 2 ° is twice that of 1° on

- ○ the Celsius scale (Interval)?
- ○ the Fahrenheit scale (Interval)?
- ○ the Kelvin scale (Ratio)?

Temperature can be either an interval or a ratio attribute, depending on its measurement scale.

- When measured on the Kelvin scale, a temperature of 2° is, in a physically meaningful way, twice that of a temperature of 1°.
- Physically, when measured on the Fahrenheit (Celsius) scale, a temperature of 2° is not much different than a temperature of 1◦.

Discrete and Continuous Attributes

- **Discrete** Attribute
  - Has only a finite or countably infinite set of values
  - Categorical (e.g., zip codes, ID numbers) or Numeric (e.g., counts)
  - Often represented as integer variables.
  - Special case: binary attributes represented as Boolean or 0/1 (True/False, Yes/No, male/female)
- **Continuous** Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Typically represented as floating-point variables.
  - Practically, real values can only be measured and represented with limited precision.

Combination of Attributes

- Theoretically, any of nominal, ordinal, interval, and ratio attributes could be combined/matched with any of the binary, discrete, and continuous attributes.
- Typically
  - nominal and ordinal attributes are binary or discrete
  - Interval and ratio attributes are continuous

Asymmetric Attributes

- Asymmetric attributes: only **presence** (a non-zero attribute value) is regarded as important
  - E.g., Items bought by customers, courses took by students
  - Can be discrete or continuous
- If we met a friend in the grocery store, would we ever say that "I see our purchases are very similar since we didn't buy most of the same things."

Key Messages for Attribute Types

- The types of operations you choose should be “meaningful” for the type of data you have
  - Distinctness, order, addition, and multiplication are only four (among many possible) properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not present
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - In the end, what is meaningful can be specific to domain

2.2 Types of Data

- Record Data
- Graph-based Data
- Ordered Data

Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence (non-zero attribute values) counts
- Resolution
  - Patterns depend on the scale

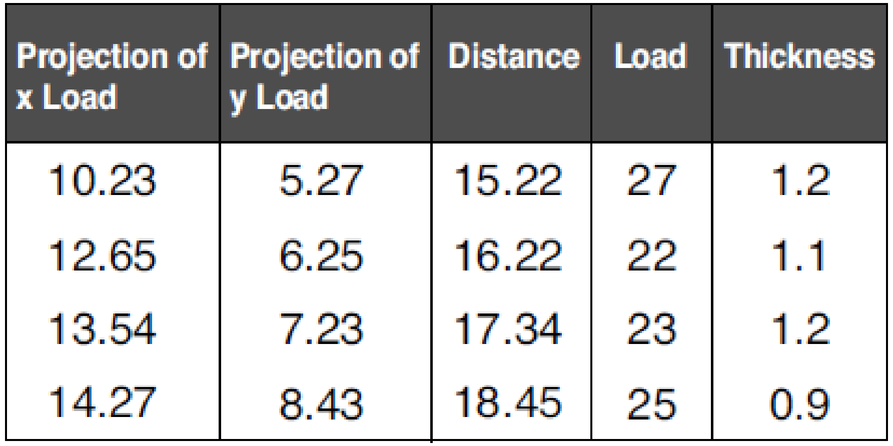
Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes
- Record data is usually stored either in flat files or in relational databases.

Name	Eye Color	Race	Age	Job	Height
Alice	Blue	White	28	Engineer	5'6"
Bob	Green	Hispanic	34	Architect	5'10"
Clara	Brown	Asian	45	Professor	5'5"
Dave	Hazel	Black	22	Student	6'0"
Eve	Grey	Spanish	30	Artist	5'7"

Record Data: Data Matrix

- If data objects have the same **fixed set of numeric attributes**, then the data objects can be thought of as **vectors** in a multi-dimensional space, where each dimension represents a distinct attribute
- Such a data set can be represented by an **m by n matrix**, where there are m rows, one for each object, and n columns, one for each attribute

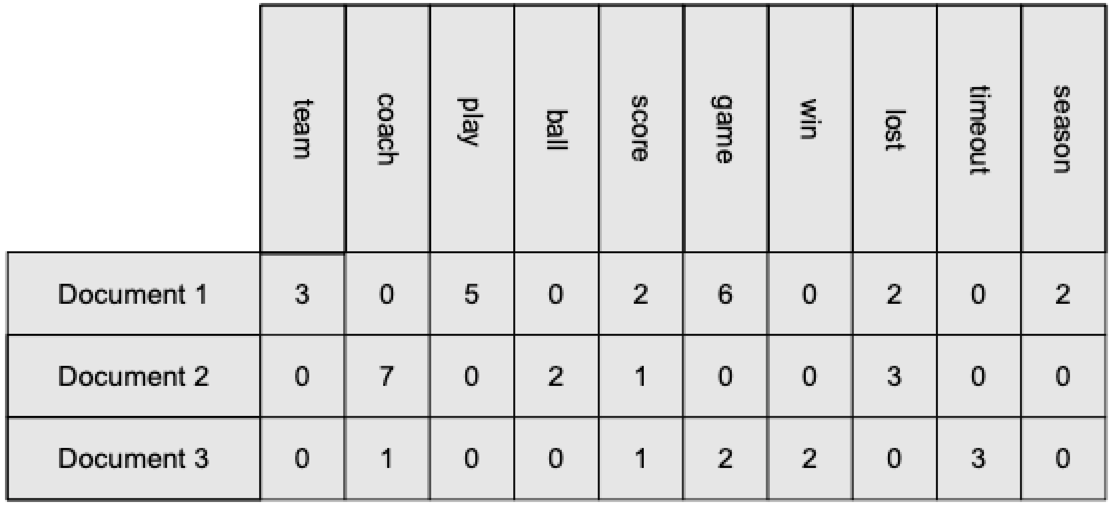


Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

$$\begin{bmatrix} 10.23 & 5.27 & 15.22 & 27 & 1.2 \\ 12.65 & 6.25 & 16.22 & 22 & 1.1 \\ 13.54 & 7.23 & 17.34 & 23 & 1.2 \\ 14.27 & 8.43 & 18.45 & 25 & 0.9 \end{bmatrix}$$

Record Data: Document Data

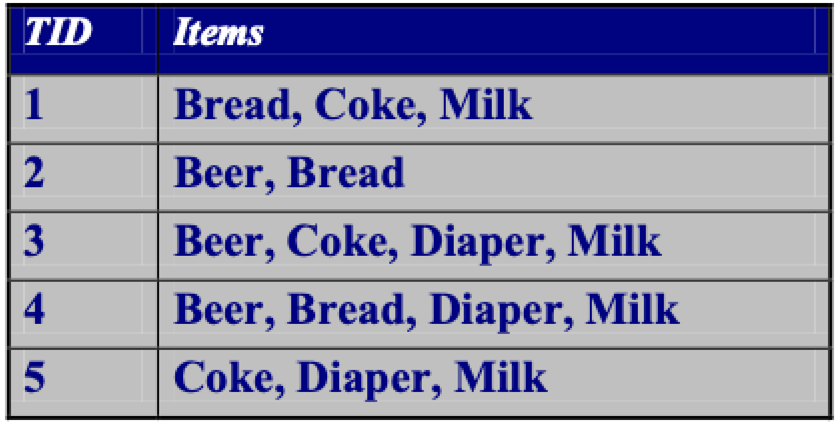
- Each term is an attribute.
- The value of each attribute is the frequency of the corresponding term in the document.
- Each document becomes a “term” vector.



	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

Record Data: Transaction Data

- Consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
- Transaction data is a special type of record data.
- It is a collection of sets of items, can also be viewed as a set of records with asymmetric attributes.



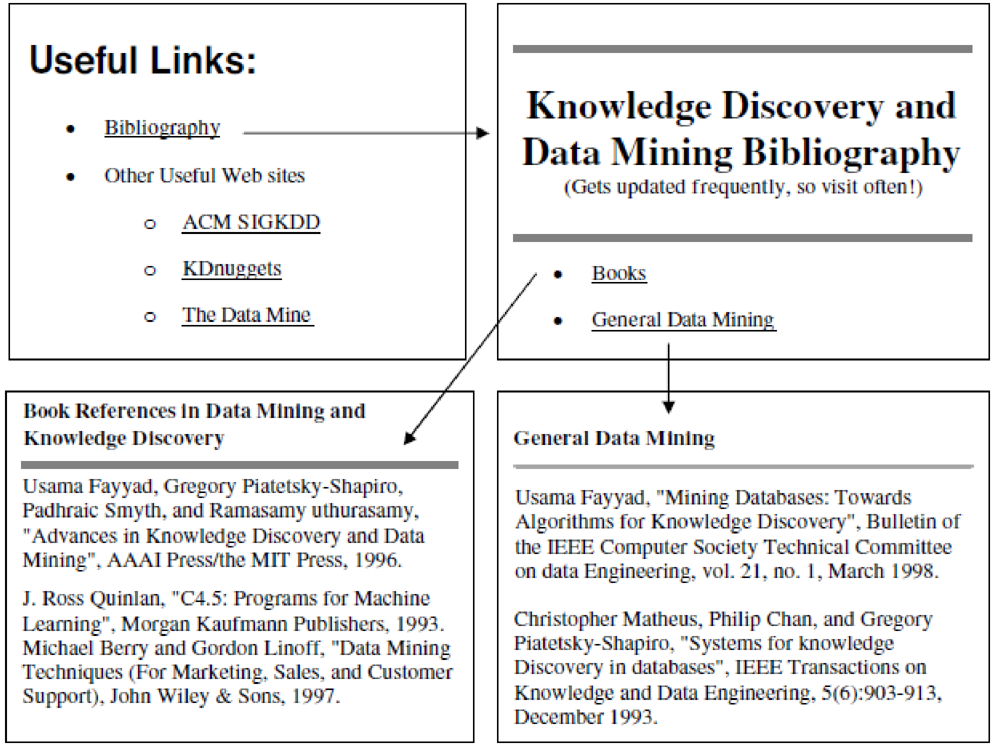
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

- The graph captures relationships among data objects
- The data objects themselves are represented as graphs

Graph Data

- The graph captures relationships among data objects:
  - Data objects -> nodes
  - Relationships among objects -> edges and weights
  - Example: linked web pagges



Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

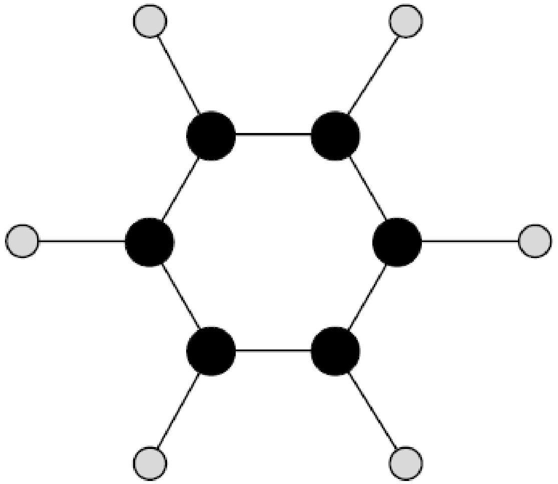
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Graph Data

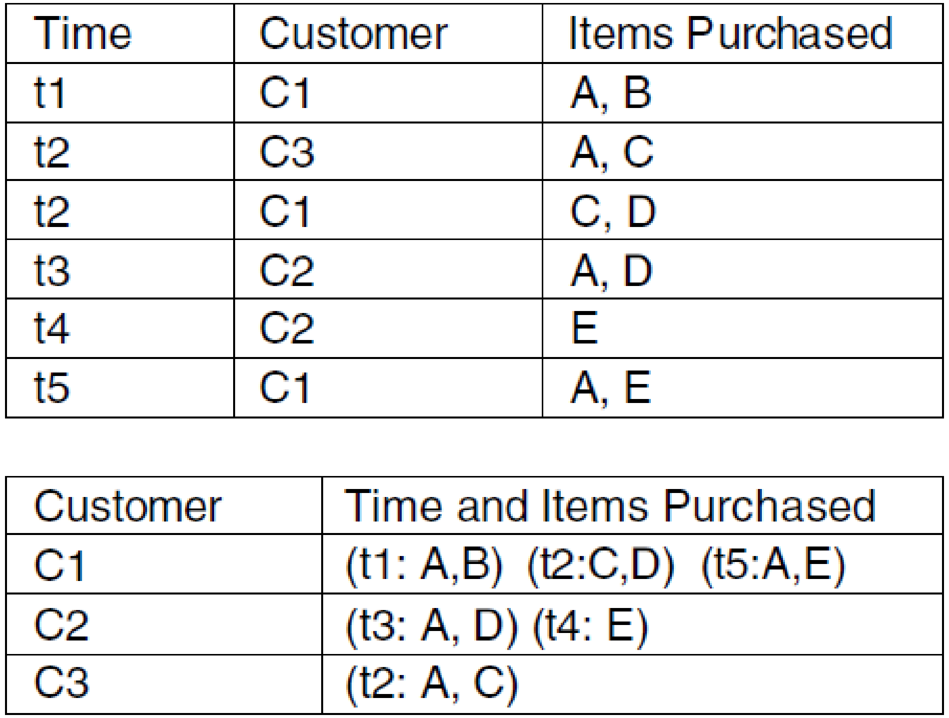
- The data objects themselves are represented as graphs
  - Objects that have structure are usually represented as graphs
  - Example: Chemical compounds



Benzene Molecule: $C6H6$

Ordered Data

- Sequential Data (temporal data): each record has a time associated with it.



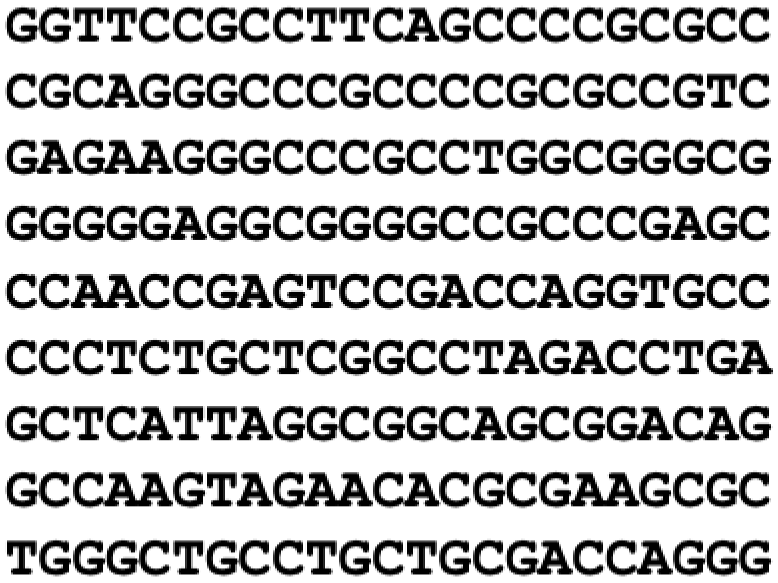
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

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)

Sequential transaction data

Ordered Data

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

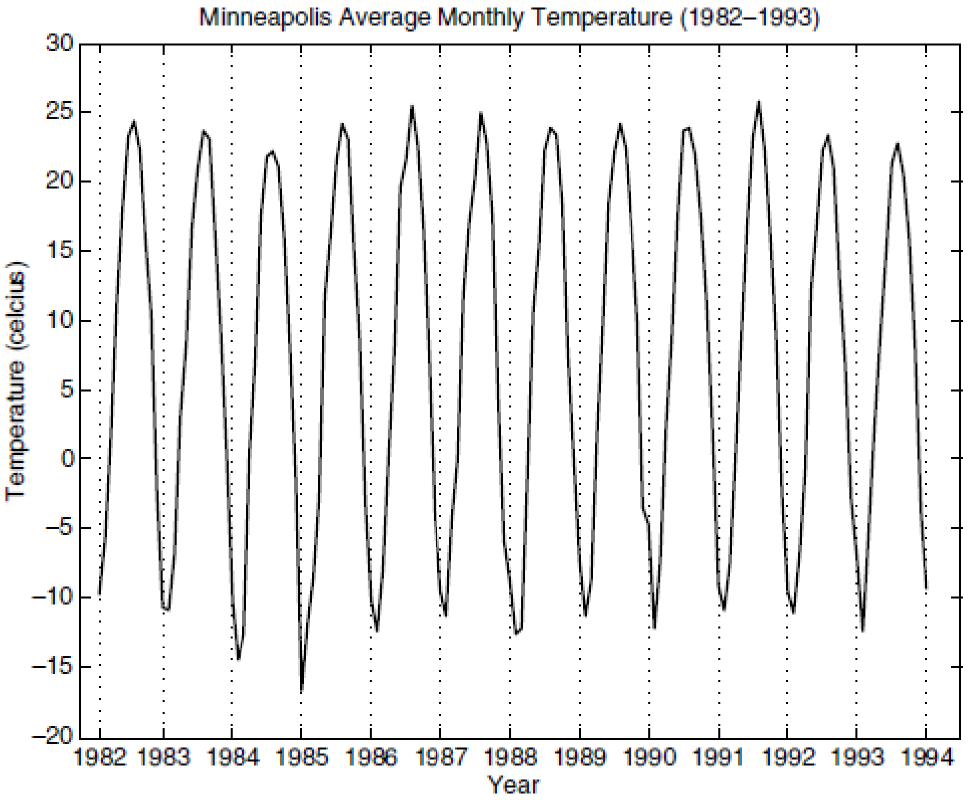


```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Genomic sequence data

Ordered Data

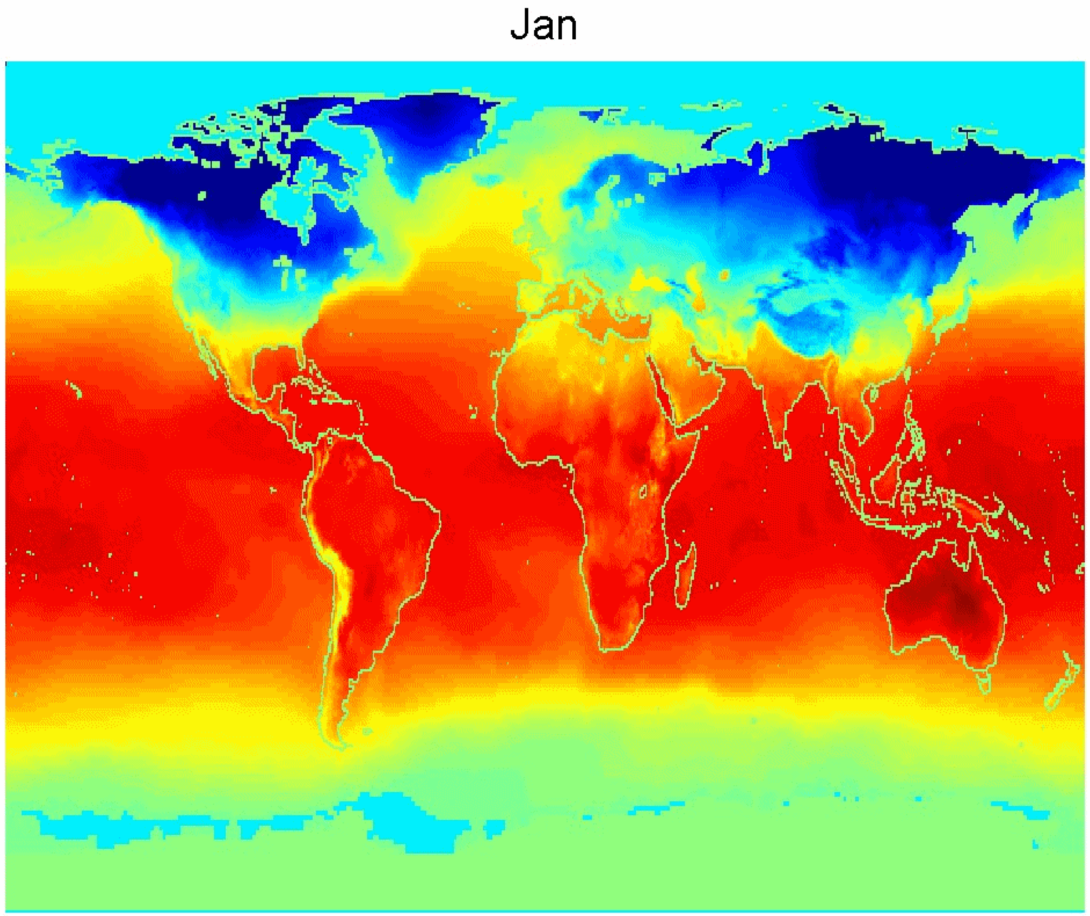
- Time Series Data: each record is a time series (a series of measurements over time).



average monthly temperature

Ordered Data

- Spatial Data: objects have spatial attributes (e.g., positions or areas)



Average Monthly Temperature of land and ocean

2.3 Data Quality

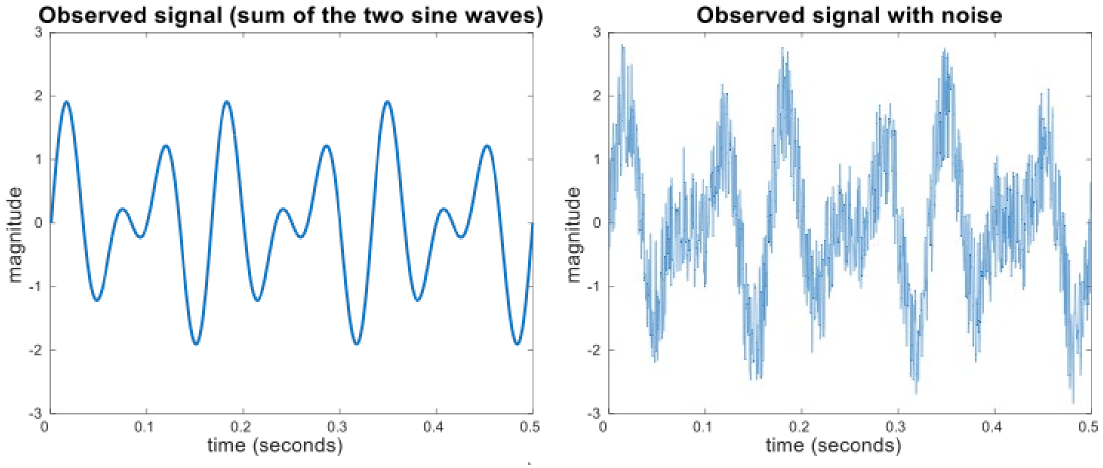
- Data is always **imperfect**.
- Poor data quality negatively affects many data processing efforts
- Example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality

- Q: What kinds of data quality problems?
- Q: What can we do about these problems?

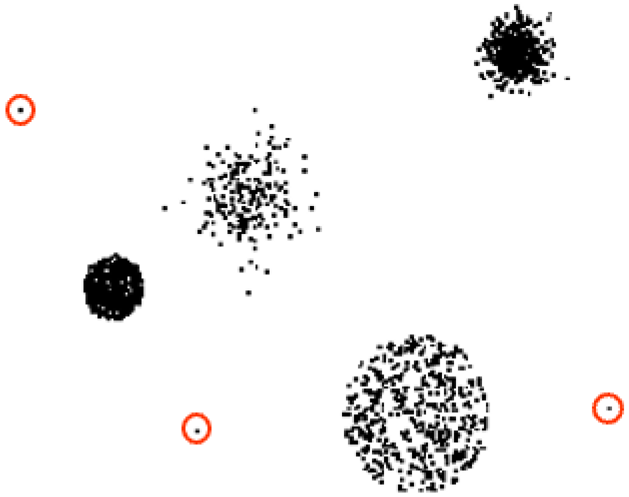
Noise

- For objects, noise is addition of extraneous objects
- For attributes, noise refers to distortion of original values
  - e.g., distortion of a person’s voice when talking on a poor phone and “snow” on television screen



Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.
- Different from noise, outliers can be **legitimate** data objects or values.
- Outliers may be of interest
  - Credit card fraud
  - Intrusion detection



Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects with missing values
  - Estimate missing values
    - Use the average attribute value or most common occurring attribute value
    - Interpolation
  - Ignore the missing value during analysis

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data deduplication
  - Challenge 1: dealing with inconsistent values from different objects
  - Challenge 2: avoid combining data objects that are similar, but not duplicates