GUC
German University in Cairo

Spring 2019

# CSEN1083: Data Mining

## *Data (1)*

Seif Eldawlatly

# Data

- A dataset can often be viewed as a collection of data objects
- Other names: Record, point, vector, pattern, event, case, sample, observation, or entity
- Data objects are described by a number of attributes that capture the basic characteristics of an object
- Other names: Variable, characteristic, field, feature, or dimension

# Data

- Examples:

**Attributes**

**Objects**

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 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 |

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

# Attributes

- An attribute is a property or characteristic of an object that may vary; either from one object to another or from one time to another

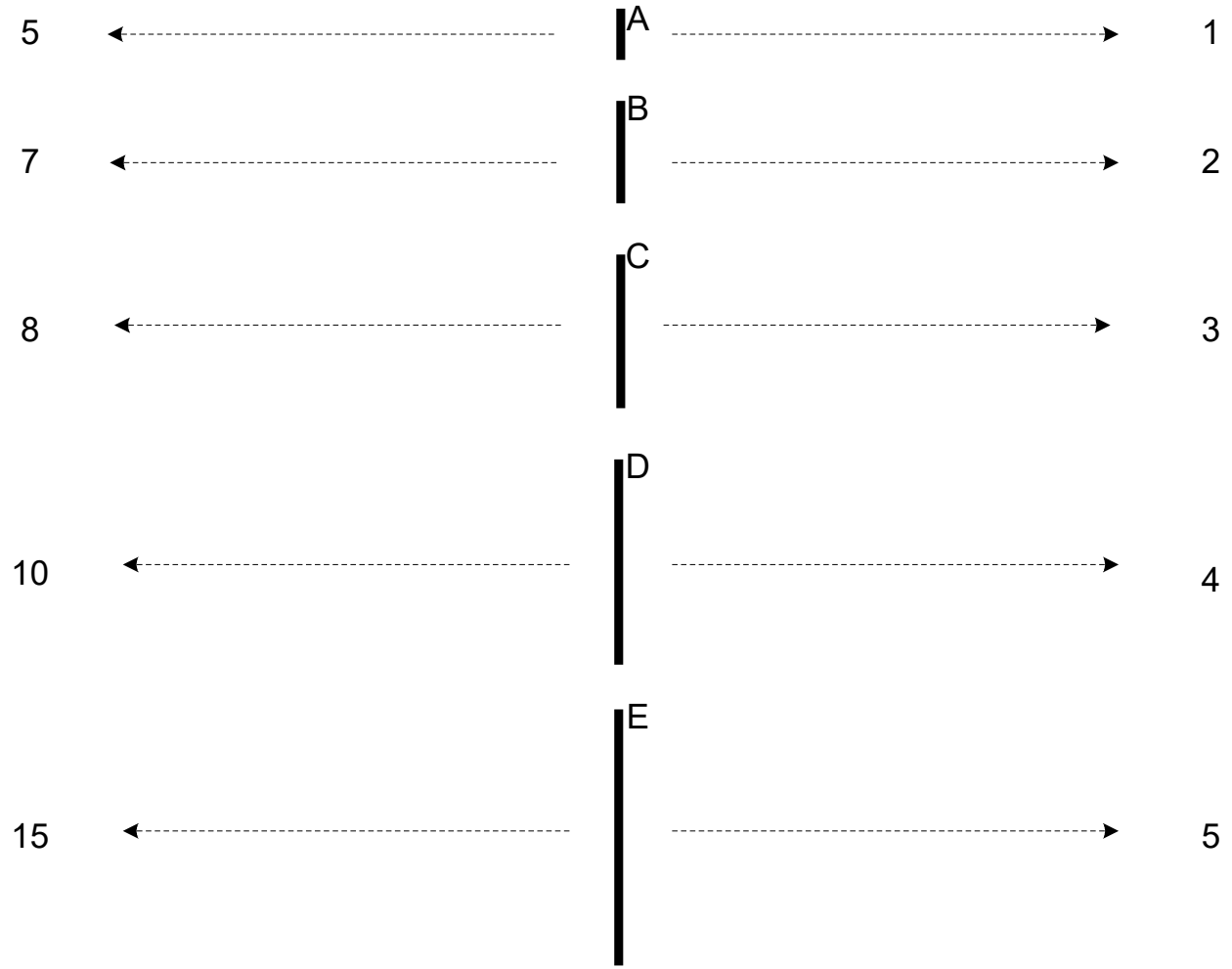
- Example: Eye color varies from person to person, while the temperature of an object varies over time

- A measurement scale is a rule (function) that associates a numerical or symbolic value with an attribute of an object

- Examples:
  - Two attributes that might be associated with an employee are ID and age (in years). Both can be represented as integers
  - Average age is reasonable, but average ID is not

# Attributes

- Example: Length of Line Segments




**This scale preserves only the ordering property of length**

**This scale preserves the ordering and additivity properties of length**

# Types of Attributes

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /

- Example: It makes sense to compare and order lengths and compute differences and ratios of length

# Types of Attributes: Method 1

- Attributes can be categorized to:
  - Nominal
    - Examples: ID numbers, eye color, postal codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

# Types of Attributes: Method 1

| | Attribute Type | Description | Examples | Operations |
|---|---|---|---|---|
| Categorical Qualitative | Nominal | Nominal attribute values only distinguish. (=, ≠) | postal codes, employee ID numbers, eye color, gender: {*male, female*} | mode, entropy, contingency correlation, $\chi 2$ test |
| | Ordinal | Ordinal attribute values also order objects. (<, >) | hardness of minerals, {*good, better, best*}, grades, street numbers | median, percentiles, rank correlation, run tests, sign tests |
| Numeric Quantitative | Interval | For interval attributes, differences between values are meaningful. (+, - ) | calendar dates, temperature in Celsius or Fahrenheit | mean, standard deviation, Pearson's correlation, *t* and *F* tests |
| | Ratio | For ratio variables, both differences and ratios are meaningful. (*, /) | temperature in Kelvin, monetary quantities, counts, age, mass, length, current | geometric mean, harmonic mean, percent variation |

# Types of Attributes: Method 1

- Transformations that define attribute levels

| | Attribute Type | Transformation | Comments |
|---|---|---|---|
| Categorical Qualitative | Nominal | Any permutation of values | If all employee ID numbers were reassigned, would it make any difference? |
| | Ordinal | An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function | An attribute encompassing the notion of good, better or best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}. |
| Numeric Quantitative | Interval | $new\_value = a * old\_value + b$ where $a$ and $b$ are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| | Ratio | $new\_value = a * old\_value$ | Length can be measured in meters or feet. |

# Types of Attributes: Method 1

- Each attribute type in the tables possesses all of the properties and operations of the attribute types above it

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

- When measured on the Kelvin scale, a temperature of $2^o$ is, in a physically meaningful way, twice that of a temperature of $1^o$

- This is not true when temperature is measured on either the Celsius or Fahrenheit scales

# Types of Attributes: Method 2

- 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, such as gender or eye color, or numeric, such as counts

  - Discrete attributes are often represented using integer variables

  - Binary attributes are a special case of discrete attributes and assume only two values, e.g., true/false, yes/no, male/female, or 0/1

# Types of Attributes: Method 2

- An independent way of distinguishing between attributes is by the number of values they can take:

- Continuous: A continuous attribute is one whose values are real numbers

  – Examples include attributes such as temperature, height, or weight

  – Continuous attributes are typically represented as floating-point variables

# Asymmetric Attributes

- For asymmetric attributes, only presence – a non-zero attribute value – is regarded as important

- Example:
  - If we met a friend in the grocery store would we ever say the following?

    *"I see our purchases are very similar since we didn't buy most of the same things."*

  - Consider a dataset where each object is a student and each attribute records whether or not a student took a particular course at a university. For a specific student, an attribute has a value of 1 if the student took the course associated with that attribute and a value of 0 otherwise

# Types of Datasets: Characteristics

- Three characteristics that apply to many datasets

- Dimensionality: The dimensionality of a dataset is the number of attributes that the objects in the dataset possess

**Two-dimensional Data**

| *Tid* | Refund | Marital Status | Cheat |
|---|---|---|---|
| 1 | Yes | Single | No |
| 2 | No | Married | No |
| 3 | No | Single | No |
| 4 | Yes | Married | No |
| 5 | No | Divorced | Yes |
| 6 | No | Married | No |
| 7 | Yes | Divorced | No |
| 8 | No | Single | Yes |
| 9 | No | Married | No |
| 10 | No | Single | Yes |

**Three-dimensional Data**

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 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 |

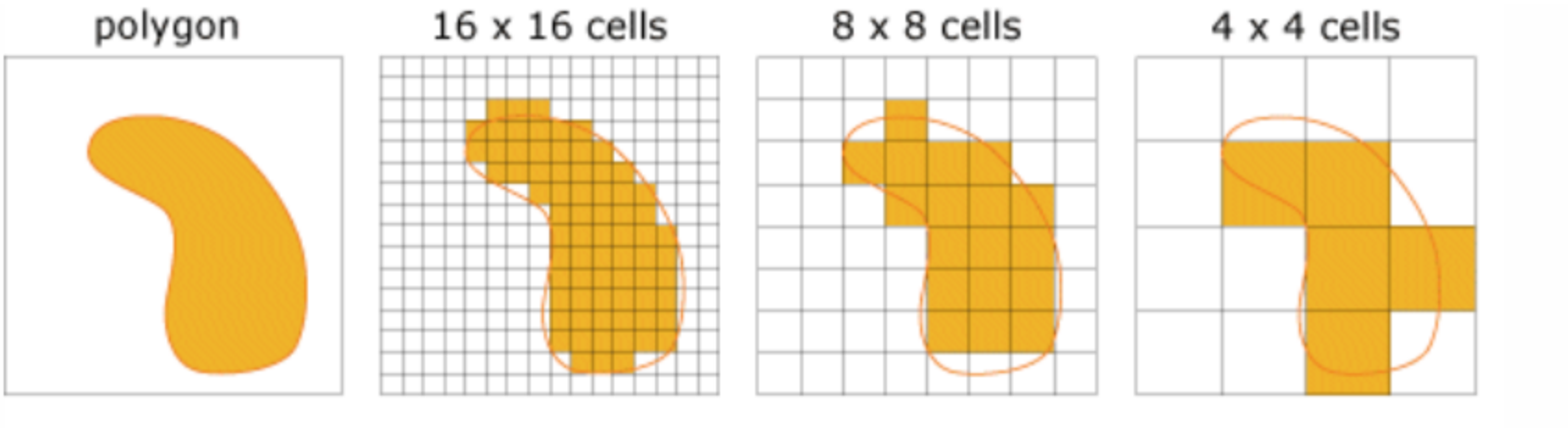
# Types of Datasets: Characteristics

- Sparsity: For some datasets, such as those with asymmetric features, most attributes of an object have values of 0

| | 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 |

# Types of Datasets: Characteristics

- 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



# Types of Datasets: Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

| *Tid* | Employed | Level of Education | # years at present address | Credit Worthy |
|---|---|---|---|---|
| 1 | Yes | Graduate | 5 | Yes |
| 2 | Yes | High School | 2 | No |
| 3 | No | Undergrad | 1 | No |
| 4 | Yes | High School | 10 | Yes |
| … | … | … | … | … |

- Record data is usually stored either in flat files or in relational databases

- Record data has different types

# Types of Datasets: Record Data

- Transaction or Market Basket Data: Each record (transaction) involves a set of items

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

- 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

| 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 |

# Types of Datasets: Record Data

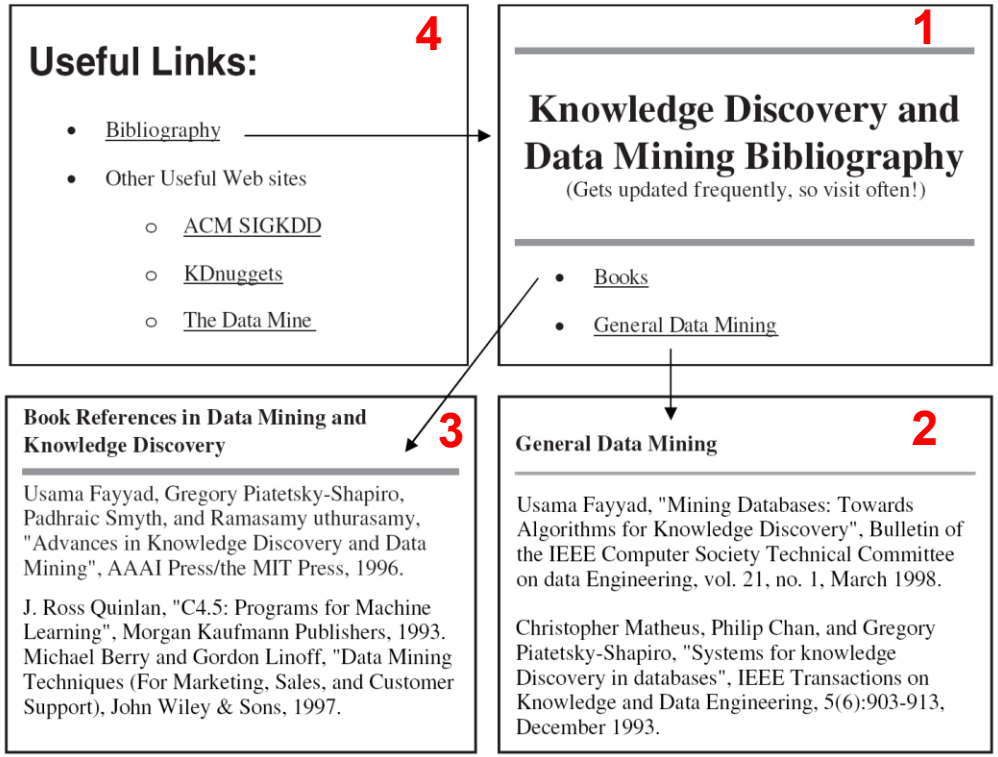
- The Sparse Data Matrix: A special case of a data matrix in which the attributes are of the same type and are asymmetric; i.e., only non-zero values are important

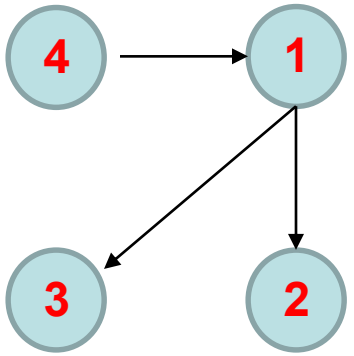
| | team | coach | play | ball | score | game | win | lost | timeout | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

# Types of Datasets: Graph-based Data

- There are two types of graph-based data

- Data with Relationships among Objects: Data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects

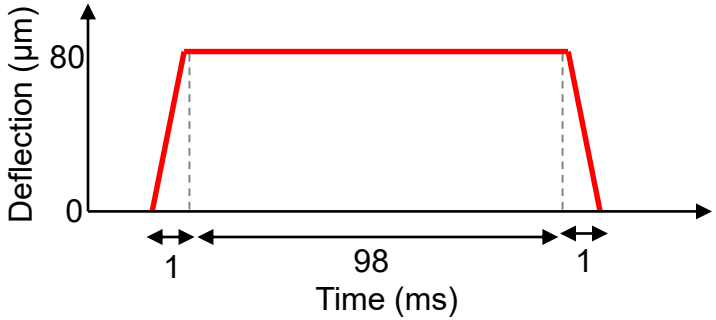


# Types of Datasets: Graph-based Data

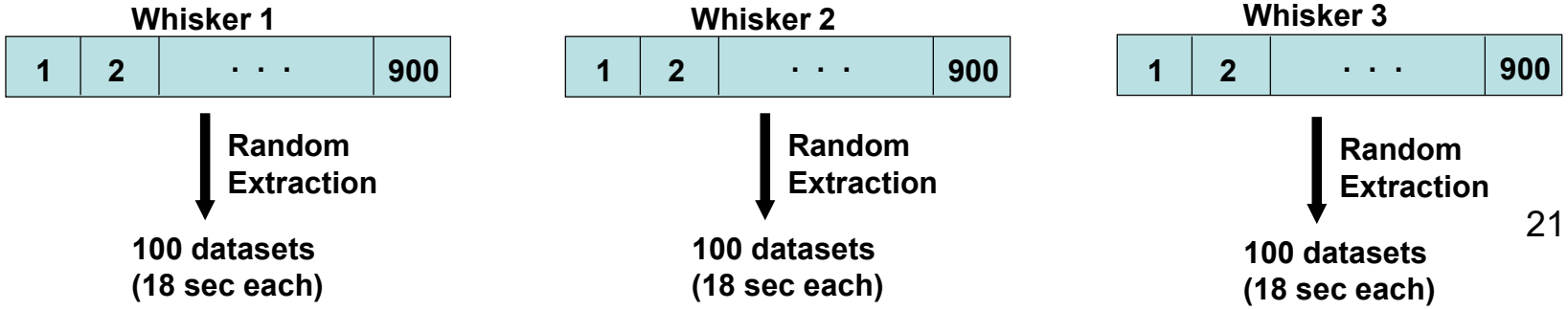
- Example: Graphical representation of brain activation

- Data Recording
  - Anesthetized rats (5 subjects)
  - Multi-electrode array with 32 channels
  - Recorded from layer V (Depth: 1100 – 1500 $\mu$m)
  - Populations size: 20 ± 7 neurons



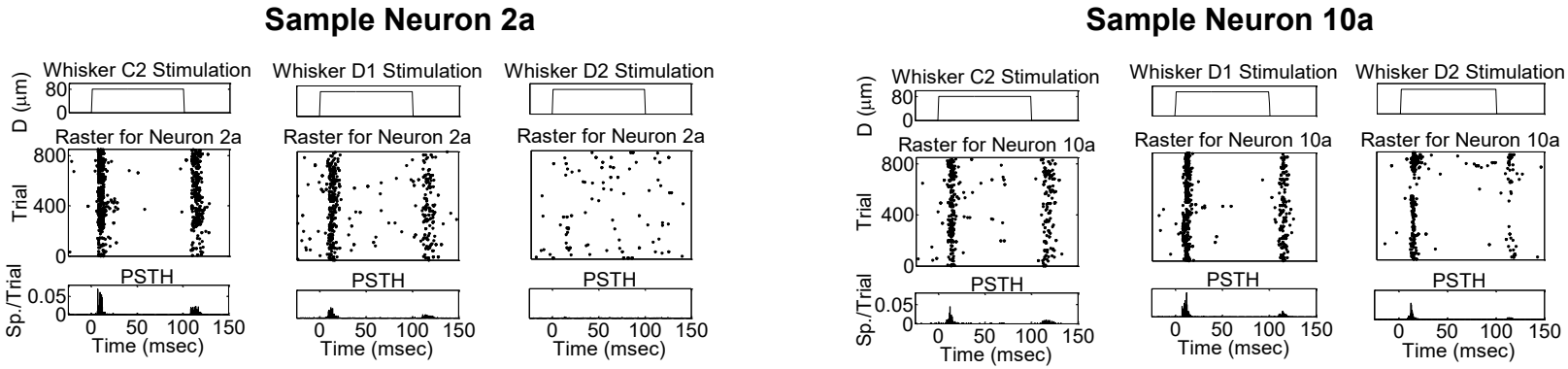

- Whiskers Mechanical Stimulation
  - 3 whiskers/rat
  - 900 horizontal deflections of 80 µm for 100 ms
  - 1 Hz frequency

- Whisker-specific Datasets Extraction

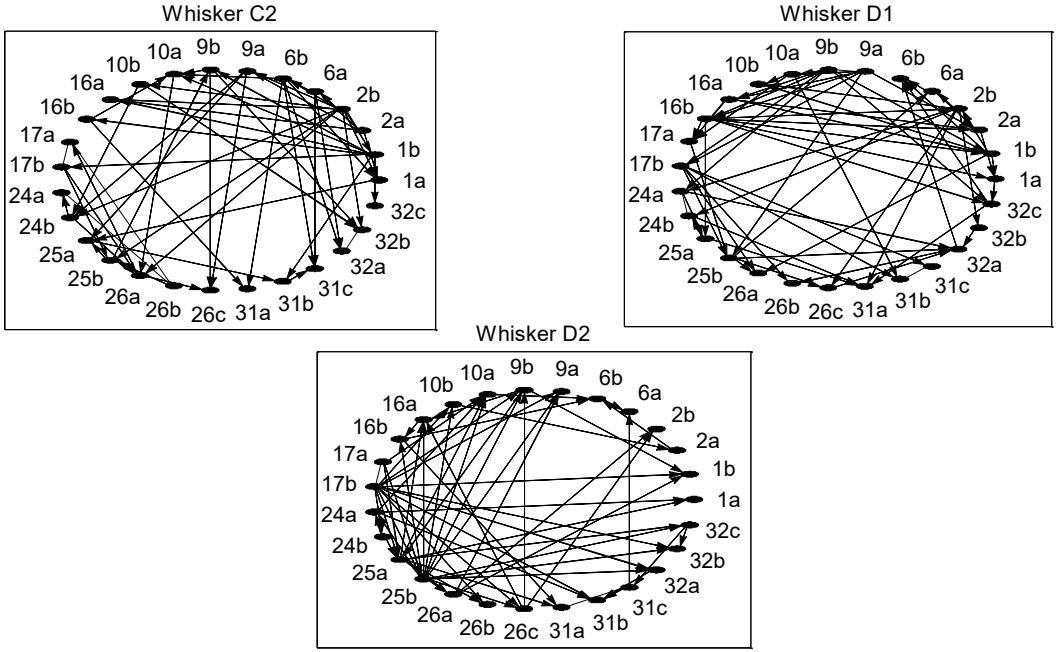


# Types of Datasets: Graph-based Data

- Variability Across Whiskers

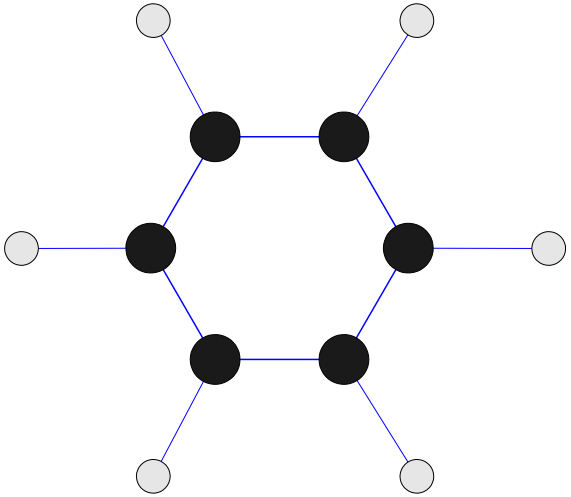



- Sample inferred graphs of whisker-specific datasets



# Types of Datasets: Graph-based Data

- Data with Objects That Are Graphs: If objects have structure, that is, the objects contain sub-objects that have relationships, then such objects are frequently represented as graphs



Benzene Molecule: C6H6

# Types of Datasets: Ordered Data

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

- Sequential Data: Can be thought of as an extension of record data, where each record has a time associated with it

- Example: Consider a retail transaction dataset that also stores the time at which the transaction took place

| 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) |

# Types of Datasets: Ordered Data

- Sequence Data: Consists of a dataset that is a sequence of individual entities, such as a sequence of words or letters

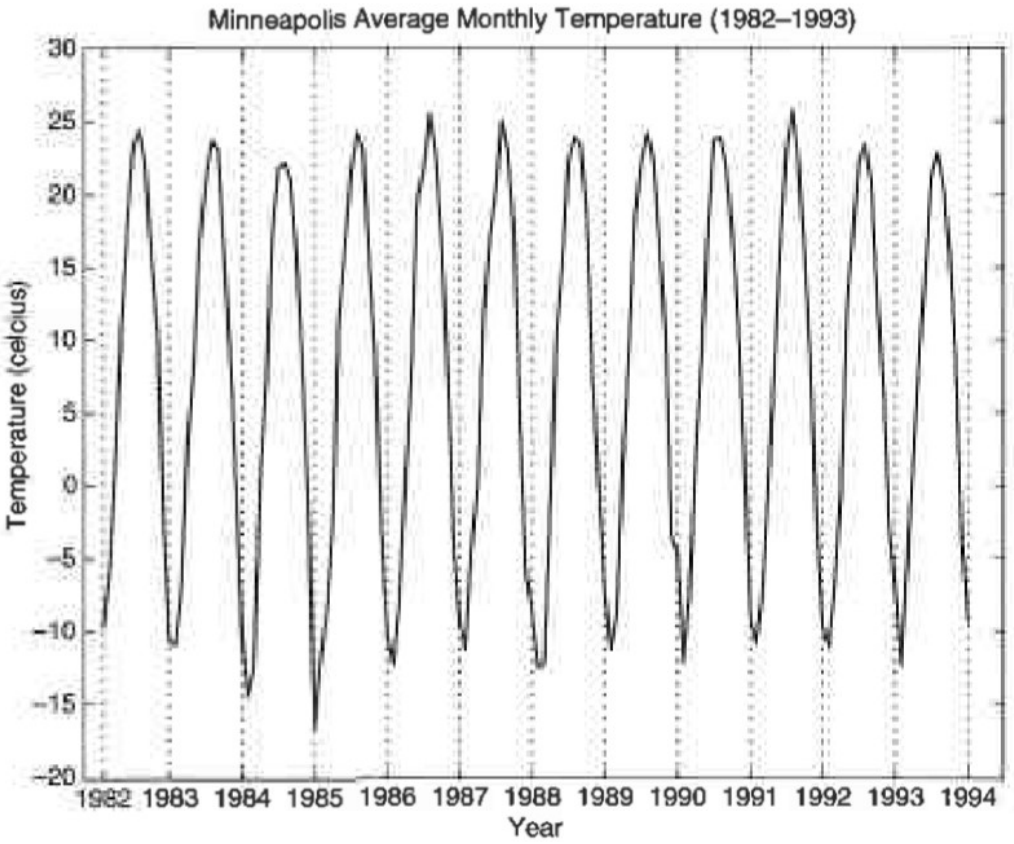
- Same as sequential but with no time information

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

**Gene Sequence Data**

# Types of Datasets: Ordered Data

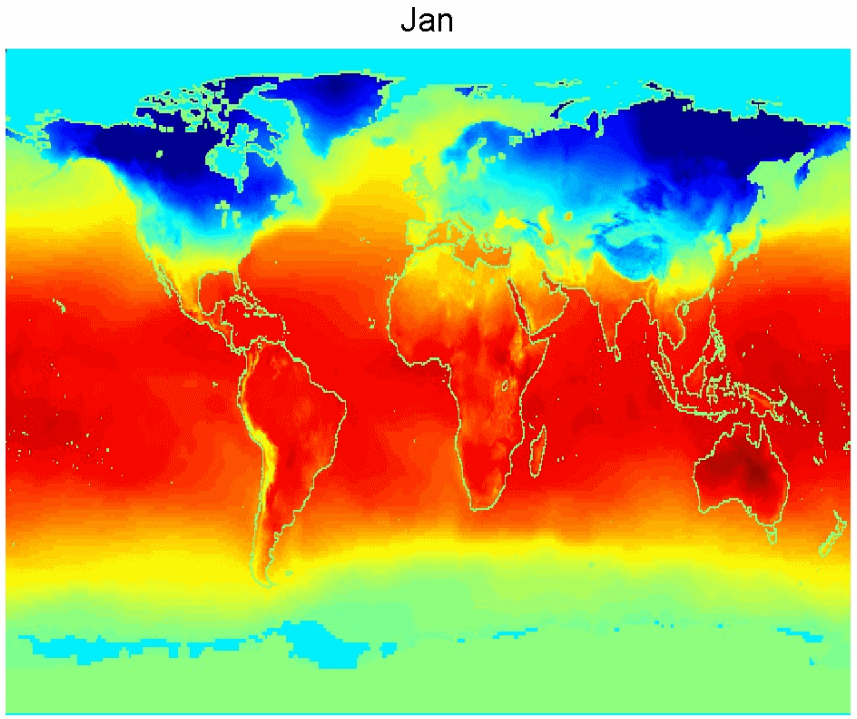
- Time Series Data: Each record represents a series of measurements taken over time




**Monthly Temperature with Time**

# Types of Datasets: Ordered Data

- Spatial Data: Some objects have spatial attributes, such as positions or areas, as well as other types of attributes




**Average Temperature with Location**