

# Data Mining

COIS 4400H / AMOD 5440H

## Data: Data Types (Chapter 2)

# What is Data?

## Attributes

The data matrix is a collection of data objects and their attributes

An attribute (feature, field, characteristic) is a property or characteristic of an object

- examples: eye color of a person, temperature, etc.

## Objects

A collection of attributes describe an object (record, point, case, sample, entity, instance)

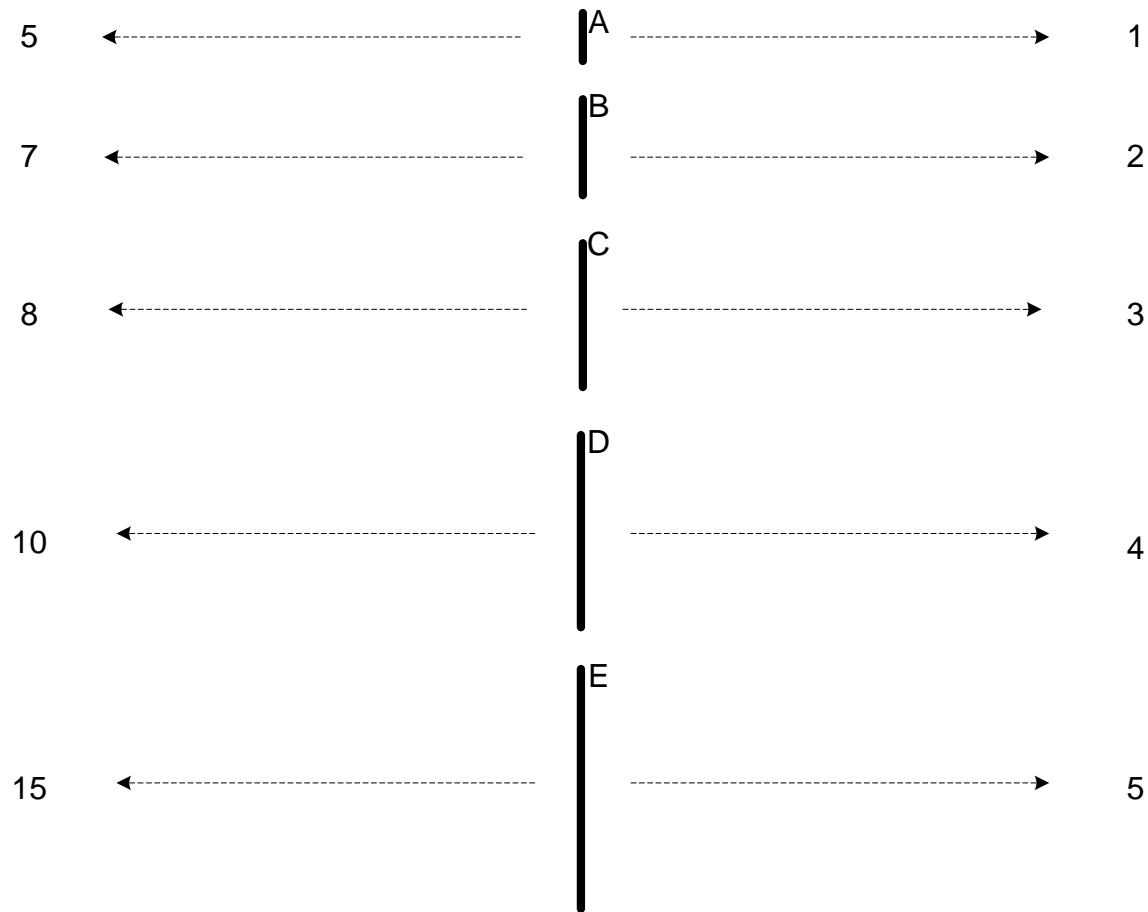
- Examples: people, customers, stars, genes, medical records, cars, etc.



<i>Tid</i>	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

# Measurement of Length

**The way you measure an attribute may not match the attribute's properties.**



# Types of Attributes

There are different types of attributes

- **Nominal**

Examples: ID numbers, eye color, zip codes

- **Ordinal**

Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

- **Interval**

Examples: calendar dates, temperatures in Celsius or Fahrenheit.

- **Ratio**

Examples: temperature in Kelvin, length, time, counts

# Properties of Attribute Values

The type of an attribute depends on which of the following properties it possesses:

- Distinctness:  $= \neq$
- Order:  $< >$
- Addition:  $+ -$
- Multiplication:  $* /$

Nominal attribute: distinctness

Ordinal attribute: distinctness & order

Interval attribute: distinctness, order & addition

Ratio attribute: all 4 properties

# Types of data sets

## Record

- Data Matrix
- Document Data
- Transaction Data

## Graph

- World Wide Web
- Molecular Structures

## Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

# Record Data

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

<i>Tid</i>	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

# Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such 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

<b>Projection of x Load</b>	<b>Projection of y load</b>	<b>Distance</b>	<b>Load</b>	<b>Thickness</b>
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1



# Document Data

Each document becomes a `term' vector,

- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

	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

# Transaction Data

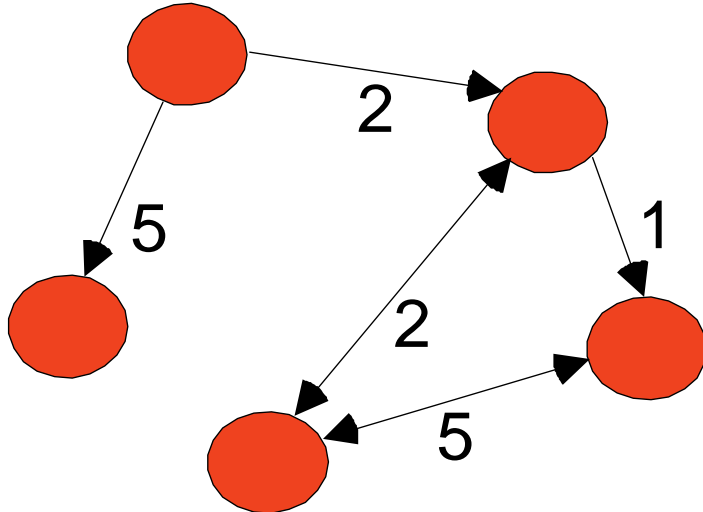
A special type of record data, where

- each record (transaction) involves a set of items.
- For example, 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.

<i><b>TID</b></i>	<i><b>Items</b></i>
<b>1</b>	<b>Bread, Coke, Milk</b>
<b>2</b>	<b>Beer, Bread</b>
<b>3</b>	<b>Beer, Coke, Diaper, Milk</b>
<b>4</b>	<b>Beer, Bread, Diaper, Milk</b>
<b>5</b>	<b>Coke, Diaper, Milk</b>

# Graph Data

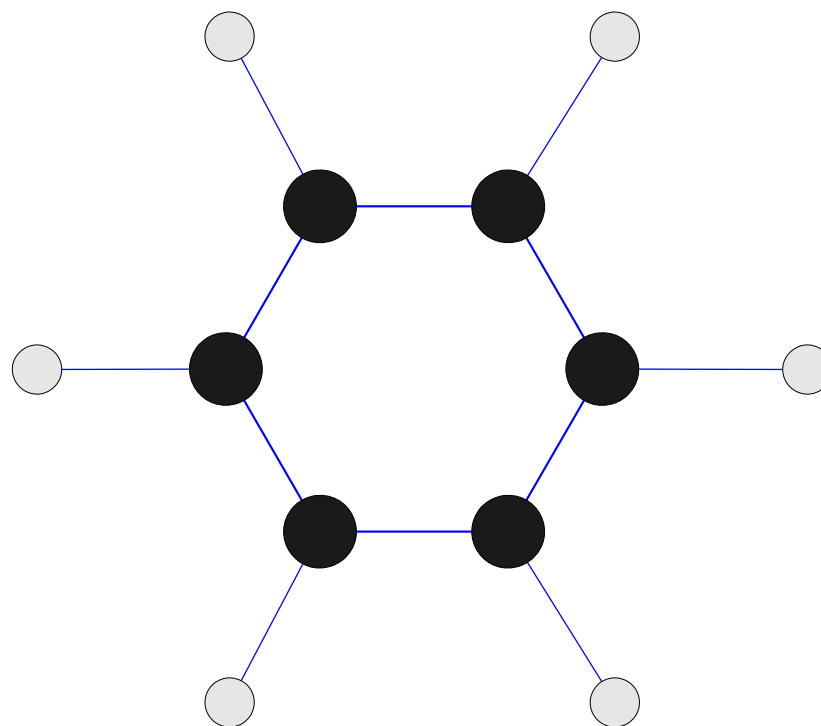
Examples: Generic graph and HTML Links



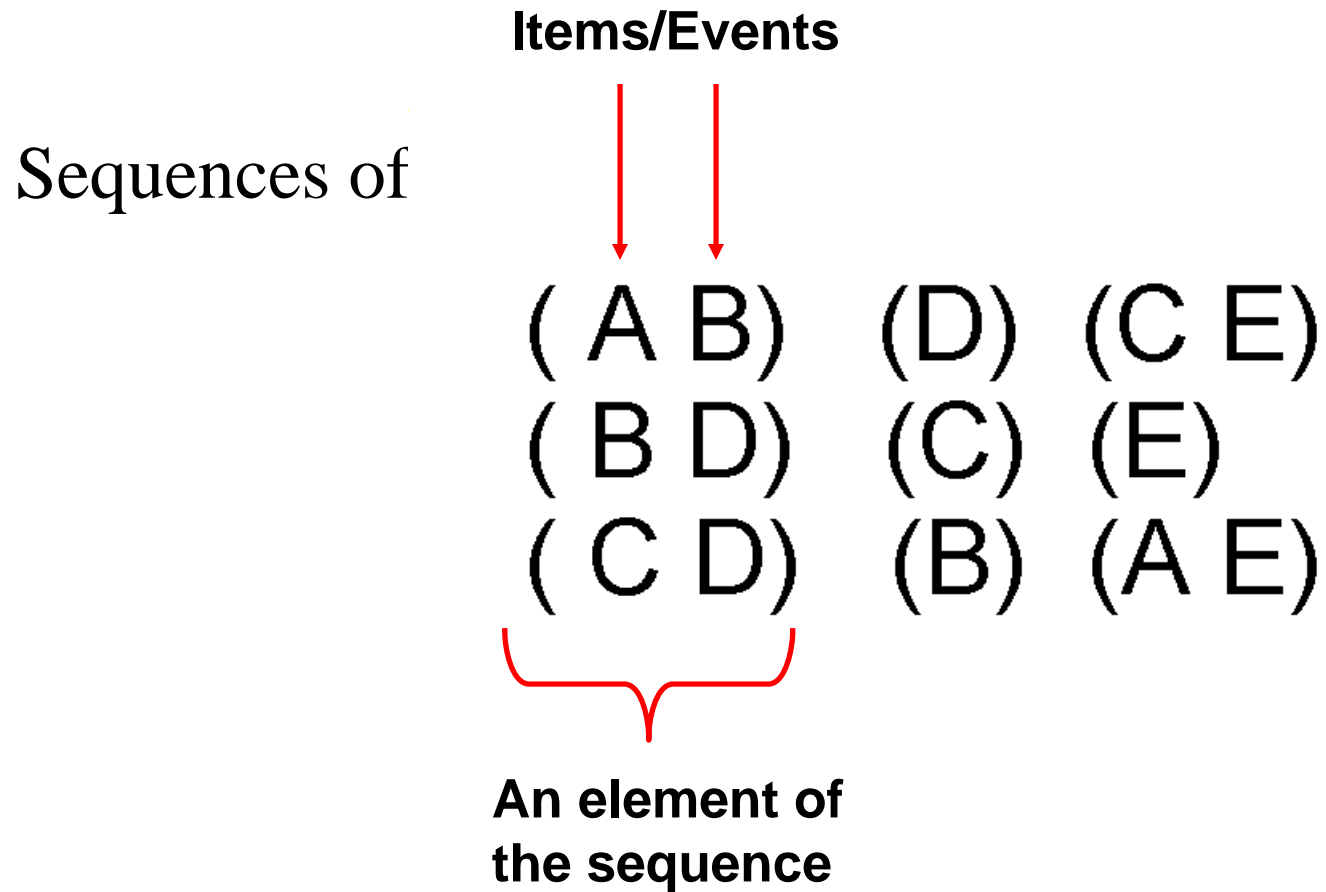
```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

# Chemical Data

Benzene Molecule:  $\text{C}_6\text{H}_6$



# Ordered Data



# Ordered Data

Genomic sequence data

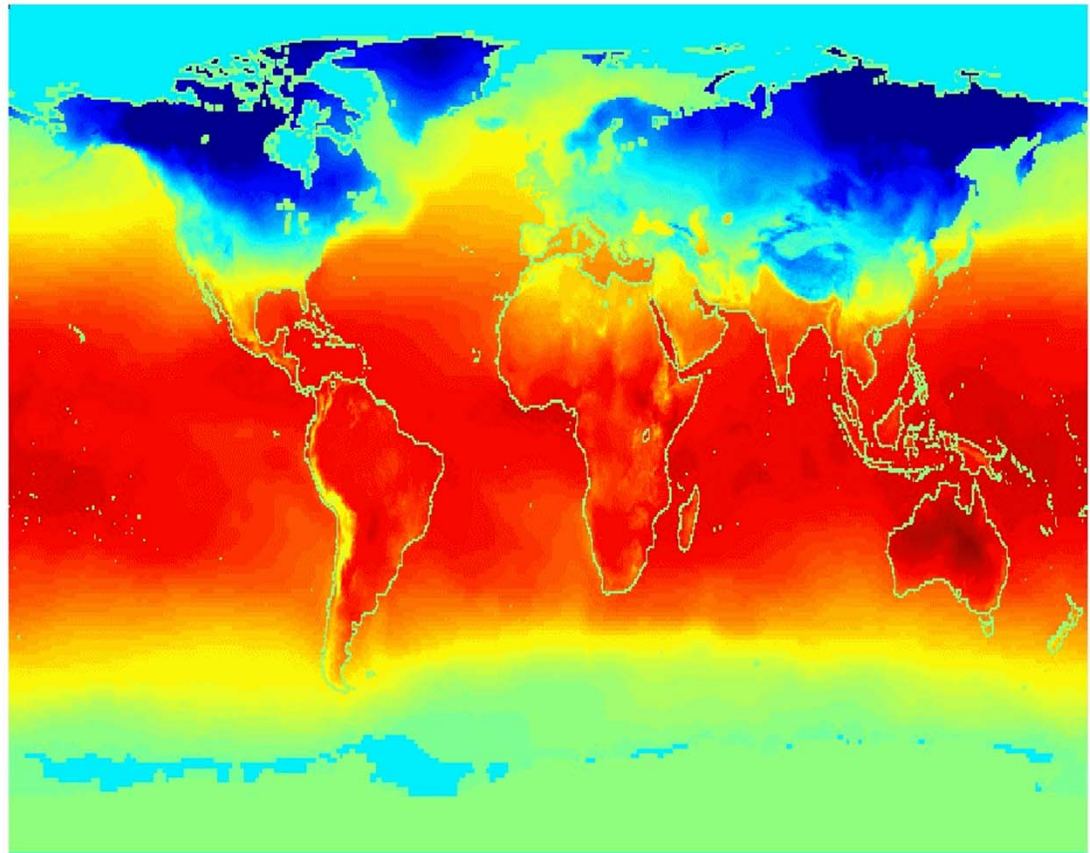
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

# Ordered Data

Jan

Spatio-Temporal Data

**Average Monthly  
Temperature of  
land and ocean**



Classify the following attributes by number of values (binary, discrete or continuous) and applicable operations (nominal, ordinal, interval, ratio)

- time of day as measured in AM and PM
- brightness as measured by you
- brightness as measured by light meter
- height above sea level:
- number of students at Trent