DATA MINING: DATA

LECTURE 7 - OUTLINE

- WHAT IS A DATA SET?
- ATTRIBUTE VALUES
- TYPES OF ATTRIBUTES
- PROPERTIES OF ATTRIBUTE VALUES
- DISCRETE VS CONTINUOUS ATTRIBUTES
- TYPES OF DATASETS
- IMPORTANT CHARACTERISTICS OF STRUCTURE DATA

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

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

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

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

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

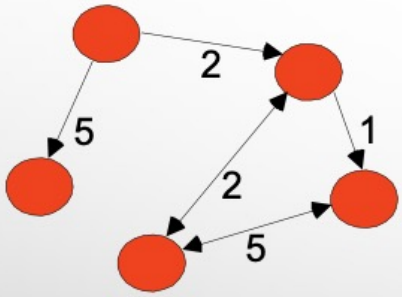
TRANSACTION DATA

- 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.

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

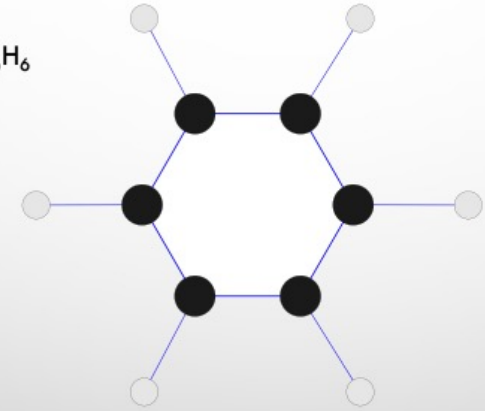
- EXAMPLES: GENERIC GRAPH AND HTML LINKS




```
<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
```

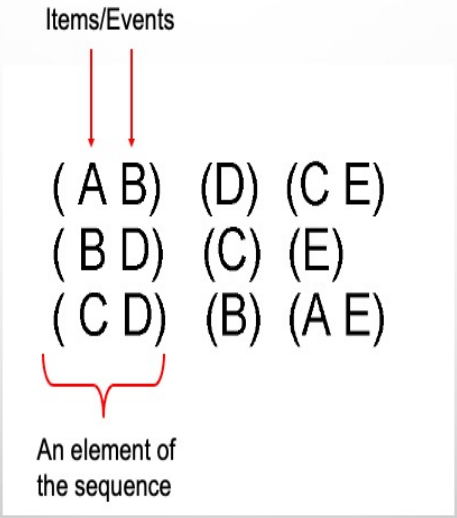
CHEMICAL DATA

- BENZENE MOLECULE: C_6H_6



ORDERED DATA

- SEQUENCES OF TRANSACTIONS



ORDERED DATA

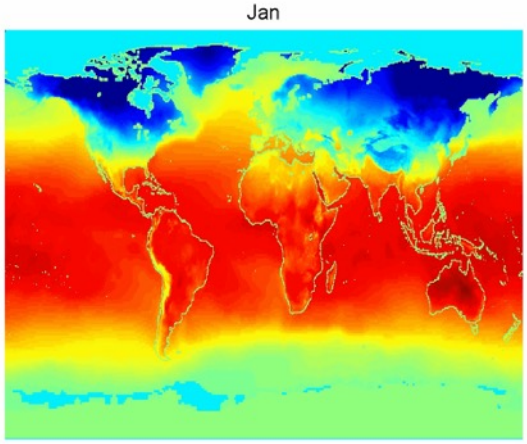
- GENOMIC SEQUENCE DATA

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

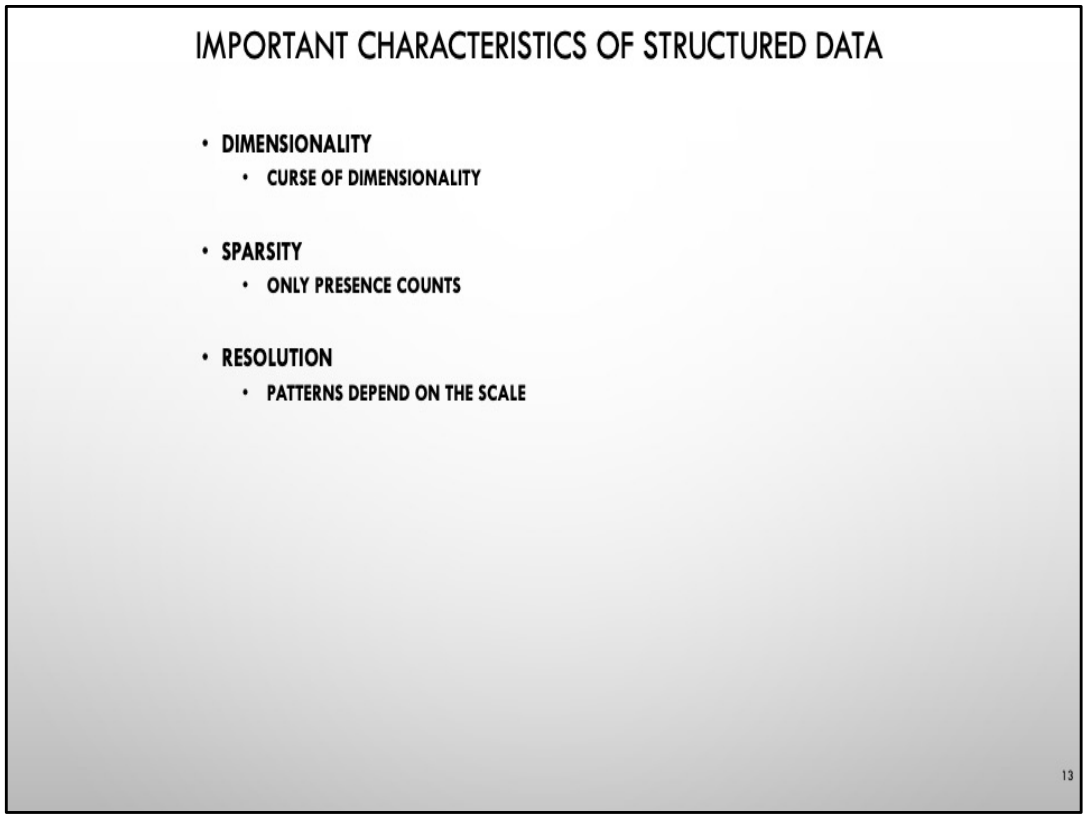
ORDERED DATA

- SPATIO-TEMPORAL DATA

Average Monthly Temperature of land and ocean



12



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

2. For some data sets, such as those with asymmetric features, most attributes of an object have values of 0; in many cases, fewer than 1% of the entries are non-zero. In practical terms, sparsity is an advantage because usually only the non-zero values need to be stored and manipulated. This results in significant savings with respect to computation time and storage. Furthermore, some data mining algorithms work well only for sparse data.

3. It is frequently possible to obtain data at different levels of resolution, and often the properties of the data are different at different resolutions. For instance, the surface of the Earth seems very uneven at a resolution of a few meters but is relatively smooth at a resolution of tens of kilometres. 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 may disappear. For example, variations in atmospheric pressure on a scale of hours reflect the movement of storms and other weather systems. On a scale of months, such phenomena are not detectable.

REFERENCE

- PANG-NING TAN, VIPIN KUMAR, MICHAEL STEINBACH: **INTRODUCTION TO DATA MINING,** CHAPTER 2, PEARSON, 2012.