



University of
Reading

CSMDM21 - Data Analytics and Mining

An Overview of Input Data

Module convenor

Dr. Carmen Lam

carmen.lam@reading.ac.uk

Department of Computer Science

Lecture notes and videos powered by

Prof. Giuseppe Di Fatta

What is “Data”?

Input data: a set of instances

- instances, aka: records, objects, points, cases, samples, entities, etc.
- Individual, independent examples of the concept to be learned.
- Single relation DB, flat file.
- 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, or feature

<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

Attribute Values

- Attribute **values** are numbers or symbols assigned to an attribute
- Distinction between **attributes** and **attribute values**
 - Same attribute can be mapped to different attribute values
 - Example: height can be measured in feet, meters, or categorical values (tall, medium, short)
 - Different attributes can be mapped to the same set of values
 - Example: Attribute values for ID and age are integers
 - But properties of attribute values can be different
 - ID has no limit but age has a maximum and minimum value

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}, ranking in a marathon

➤ **Interval**

- Examples: calendar dates, temperatures in Celsius or Fahrenheit, level of happiness (e.g. rated from 1 to 10)

➤ **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

Attribute Type Description

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=$, \neq)	zip codes, employee ID numbers, eye color, sex: $\{male, female\}$	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ($<$, $>$)	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ($+$, $-$)	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, electrical current	geometric mean, harmonic mean, percent variation

Attribute Value Transformation

Attribute Type	Transformation	Comments
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 best can be represented equally well by the values $\{1, 2, 3\}$ or by $\{0.5, 1, 10\}$.
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.

Discrete and Continuous Attributes

- Discrete Attribute
 - Has only a finite or countably infinite set of values
 - Examples: zip codes, counts, or the set of words in a collection of documents
 - Often represented as integer variables.
 - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
 - Has real numbers as attribute values
 - Examples: temperature, height, or weight.
 - Practically, real values can only be measured and represented using a finite number of digits.
 - Continuous attributes are typically represented as floating-point variables.

Types of Data Sets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Multi-Relational
 - Star or snowflake schema
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data

Important Characteristics of Structured Data

- **Dimensionality**
 - Number of attributes each object is described with
 - Challenge: high dimensionality (curse of dimensionality)
- **Sparsity**
 - Sparse data: values of most attributes are zero
 - Challenge: sparse data call for special handling
- **Resolution**
 - Data properties often could be measured with different resolutions
 - Challenge: decide on the most appropriate resolution (e.g. “Can’t See the Forest for the Trees”)

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

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	player	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

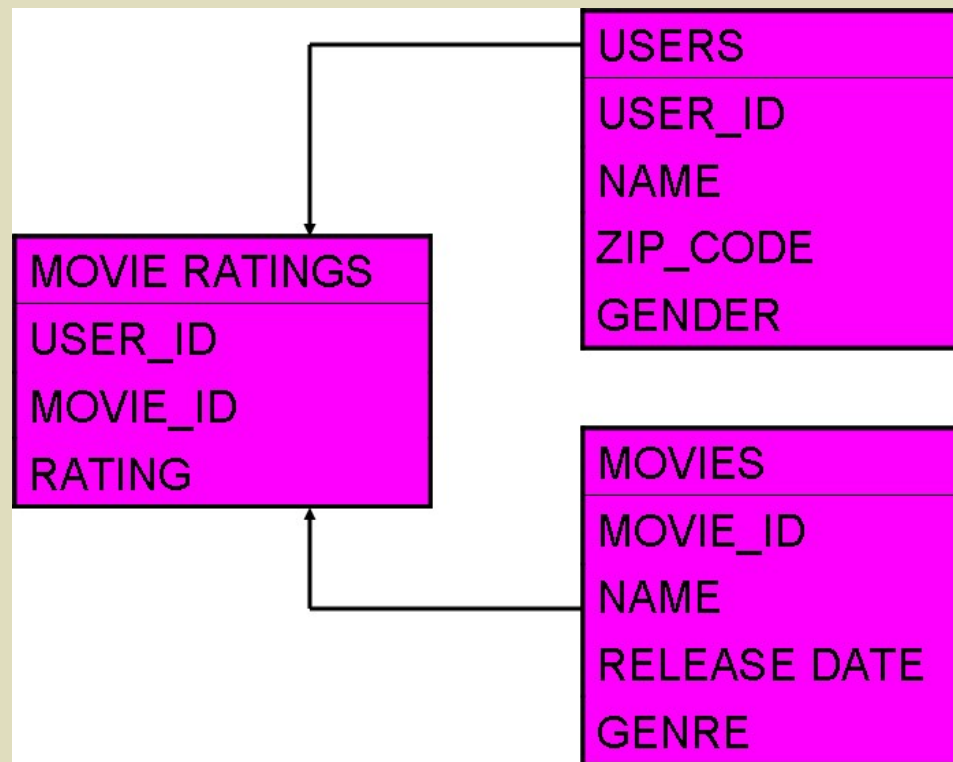
Transaction Data

- A special type of record data, where
 - each record (transaction) involves a set of items.
 - E.g., 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>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

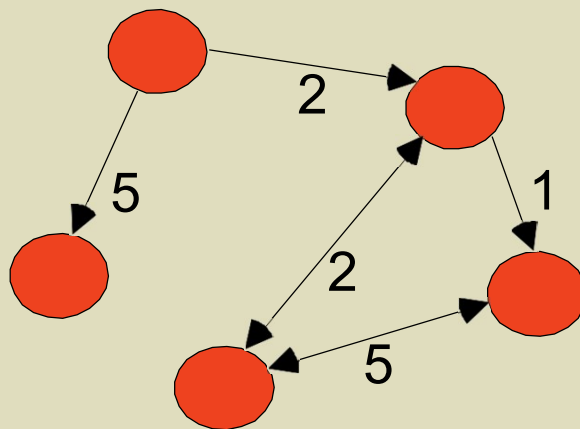
Multi-Relational Data

- Attributes are objects themselves



Graph Data

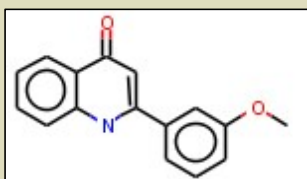
- Examples: Generic graph and HTML Links



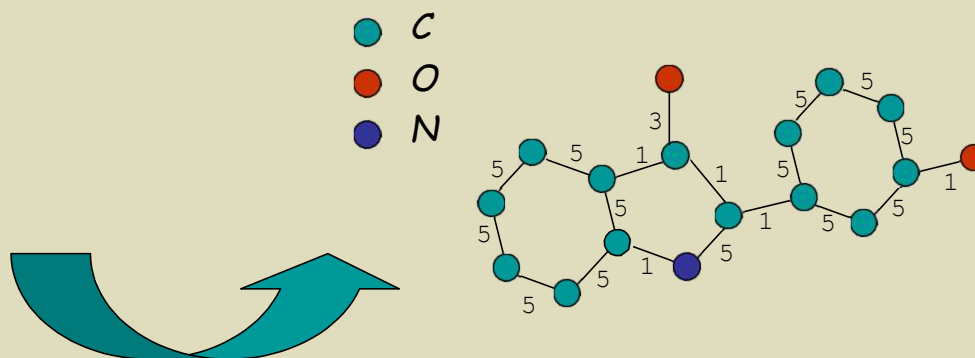
```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and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


Chemical Data

Chemical compound



Graph



Graph representations:

➤ adjacency matrix

➤ edges list

- Source atom type
- Source atom
- Bond type
- Dest. Atom type
- Dest. atom

	0	1	2	3	4	5	6	7
0		1			1			1
1	1		3				3	
2		3		5				
3			5		5			
4	1			5		5		1
5					5		5	
6		3				5		1
7	1				1		1	

(C, 0, 1, C, 1)

(C, 0, 5, N, 4)

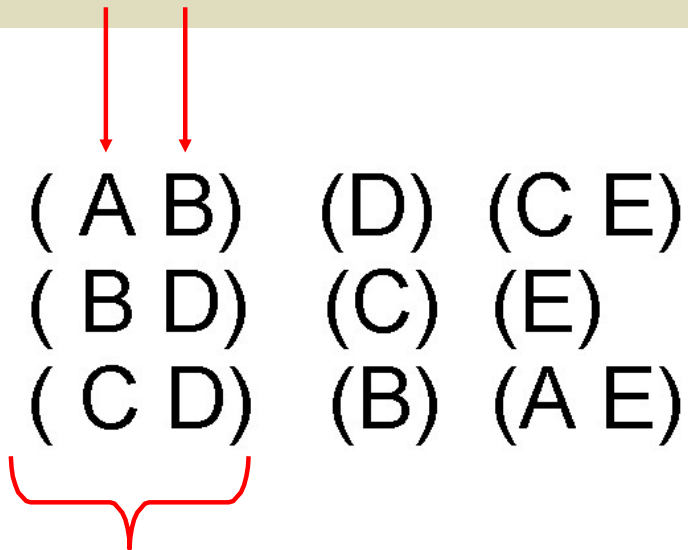
(C, 0, 1, O, 7)

...

Ordered Data

❑ Sequences of transactions

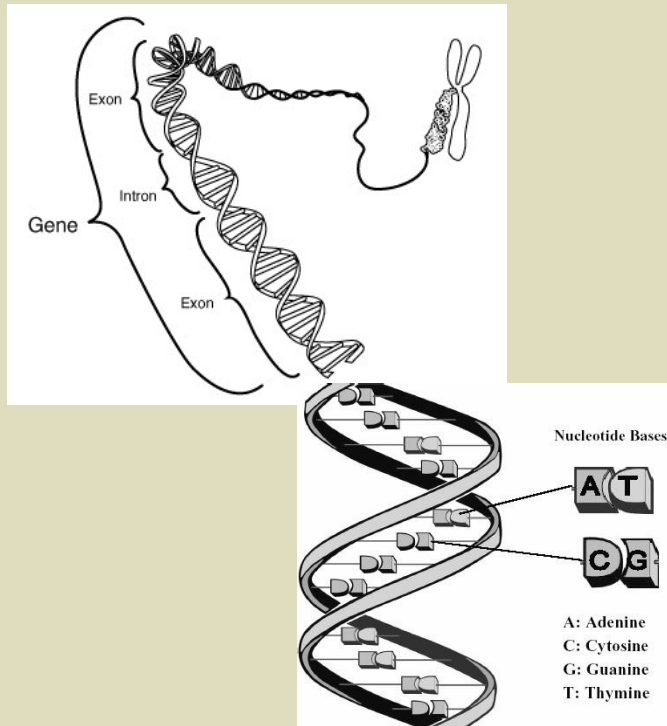
Items/Events



**An element of
the sequence**

Ordered Data

- Genomic sequence data

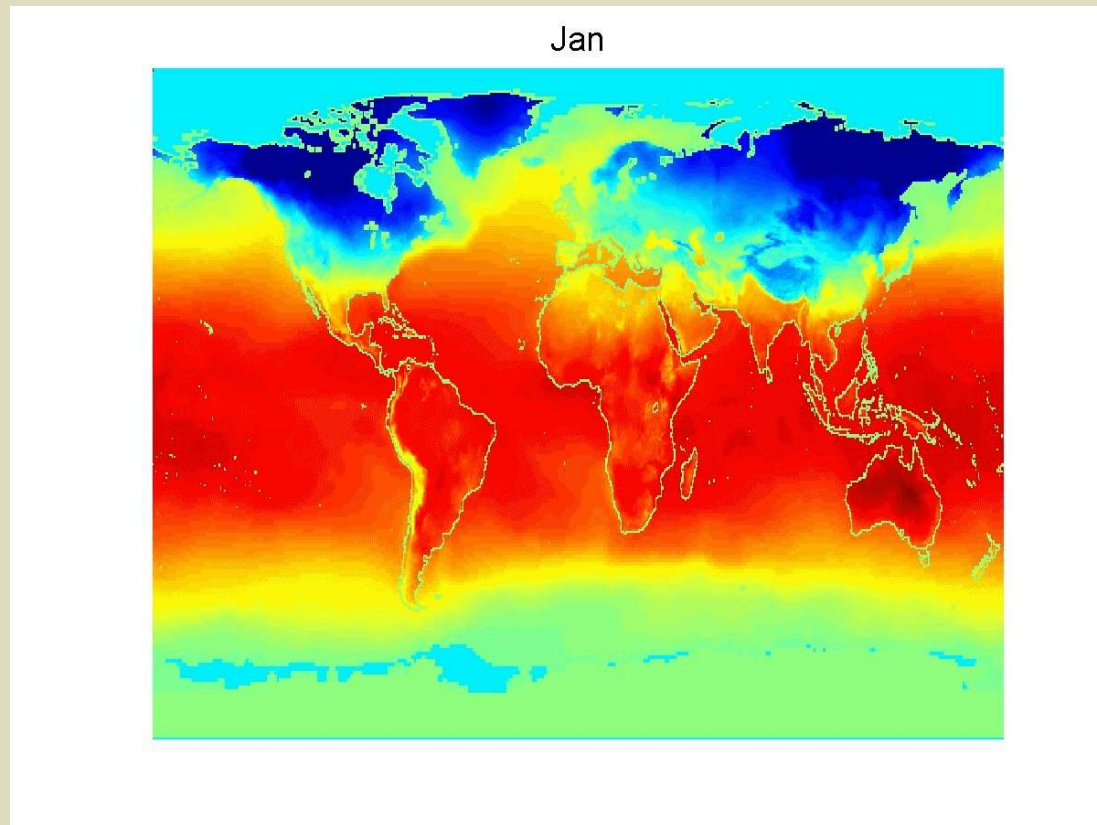


GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- Spatial-Temporal Data

Average monthly
temperature of land
and ocean

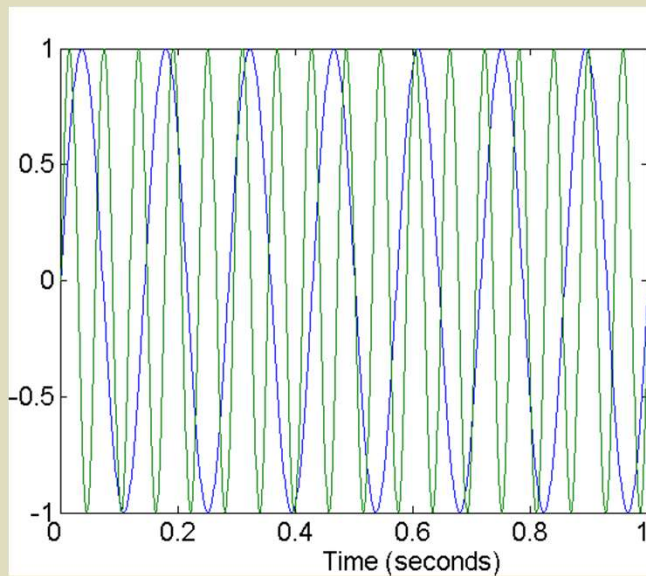


Data Quality

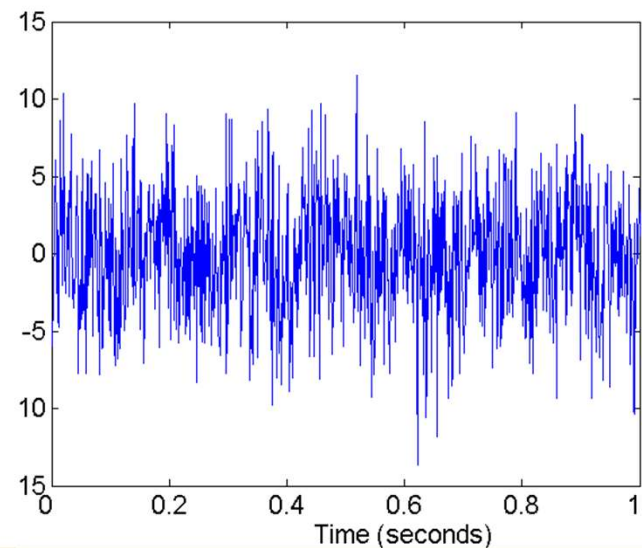
- ✓ What kinds of data-quality problems?
 - ✓ How can we detect problems with the data?
 - ✓ What can we do about these problems?
-
- Examples of data quality problems:
 - noise and outliers
 - missing values
 - duplicate data

Noise

- Noise refers to modification of original values
 - Examples: distortion of a person's voice when talking on a poor phone and “snow” on television screen



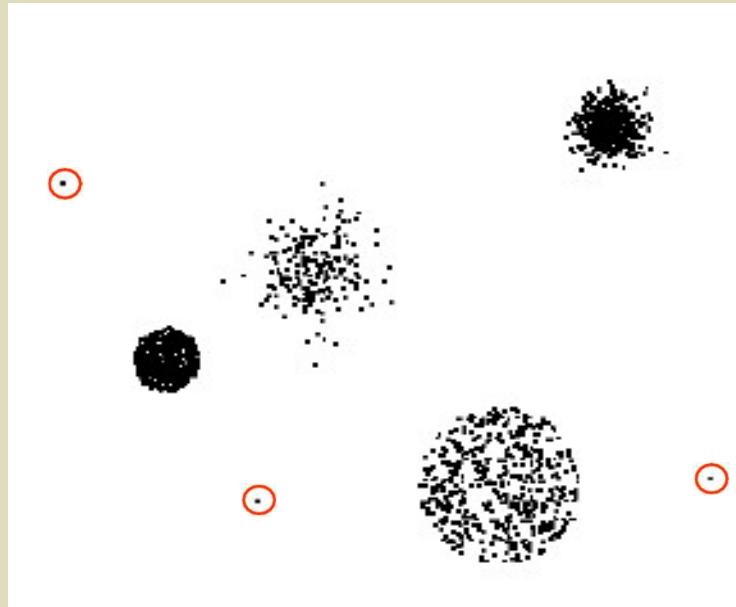
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



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
 - Estimate Missing Values
 - 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 cleaning
 - Process of dealing with duplicate data issues

Next video lecture:

➤ Data Preprocessing