Master Degree in Artificial Intelligence for Science and Technology

Types of Data

Fabio Stella

Department of Informatics, Systems and Communications

University of Milano-Bicocca

fabio.stella@unimib.it

OUTLOOK

- **DATA OBJECT** and **ATTRIBUTE**
- **TYPES OF ATTRIBUTES**
- **IMPORTANT CHARACTERISTICS OF DATA**
- **TYPES OF DATA SETS**
- **DATA QUALITY**

What is data?

- 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

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

- **ATTRIBUTE VALUES** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attribute and attribute values

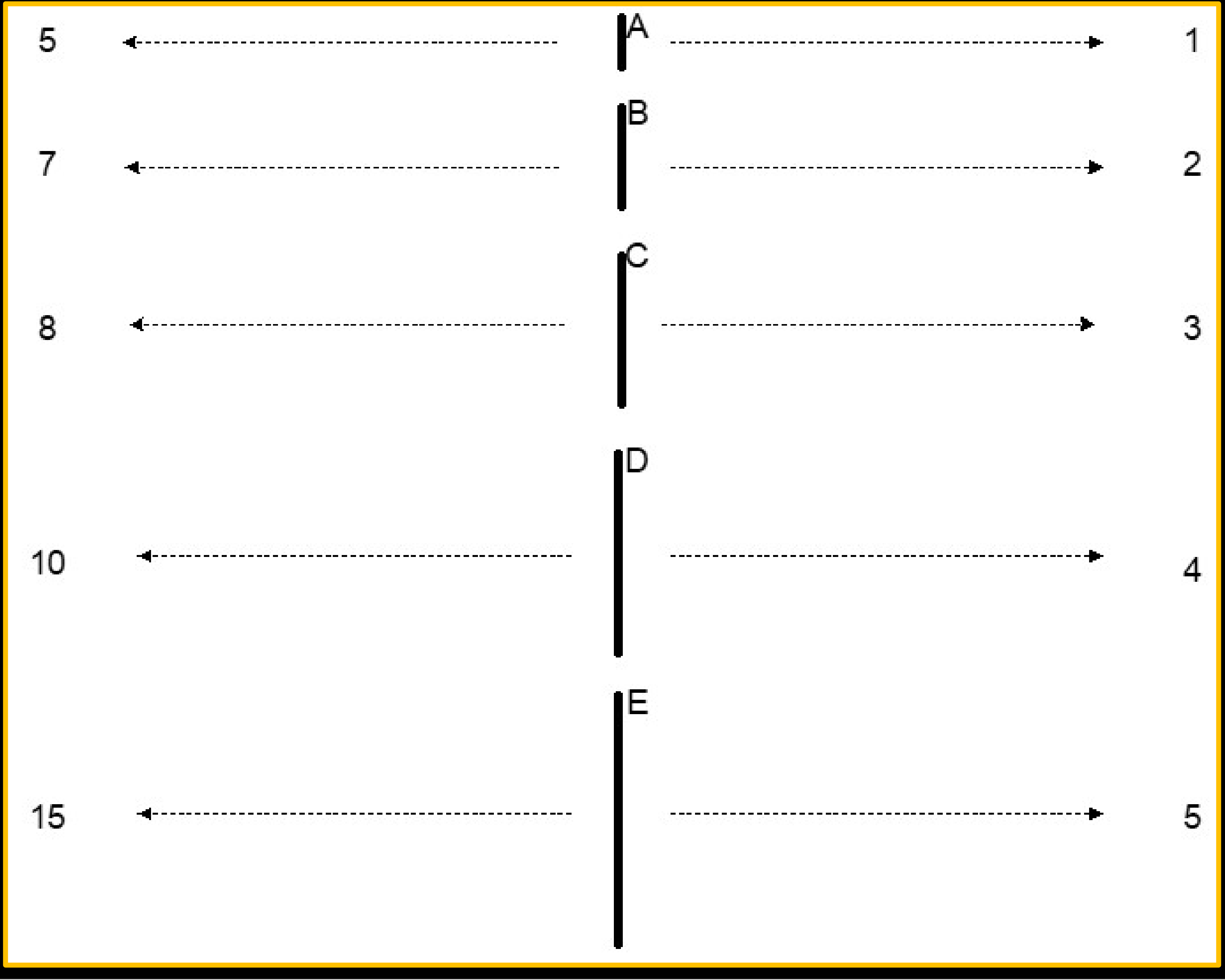
  - same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters

  - different attributes can be mapped to the same set of values
    - Example: attribute values for ID and age are integers

  - but properties of an attribute can be different than the properties of the values used to represent the attribute

The way you measure an attribute may not match the attributes properties.

This scale preserves only the ordering property of length.



This scale preserves the ordering and additivity properties of length.

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 to 10), grades, height {tall, medium, short}
- **INTERVAL**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **RATIO**
  - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Types of Data

The type of an attribute depends on which of the following **PROPERTIES/OPERATIONS** it possesses:

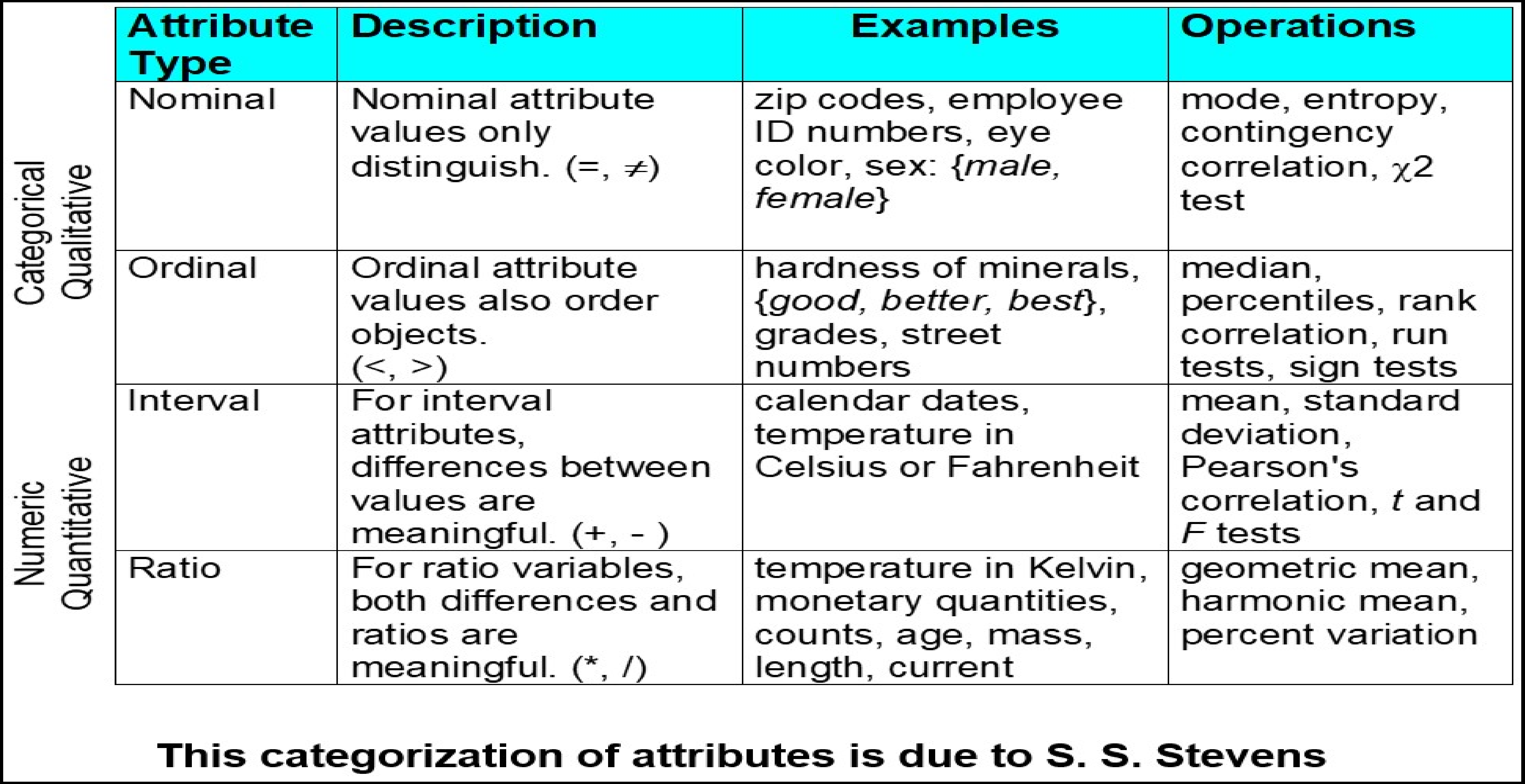
▪ **DISTINCTNESS**	=	≠
▪ **ORDER**	<	>
▪ **DIFFERENCES ARE MEANINGFUL**	+	–
▪ **RATIOS ARE MEANINGFUL**	*	/

— nominal attribute:	distinctness
— ordinal attribute:	distinctness & order
— interval attribute:	distinctness, order & meaningful differences
— ratio attribute:	all 4 properties/operations

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

- Consider measuring the height above average
  - if Bill’s height is three inches above average and Bob’s height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - is this situation analogous to that of temperature?

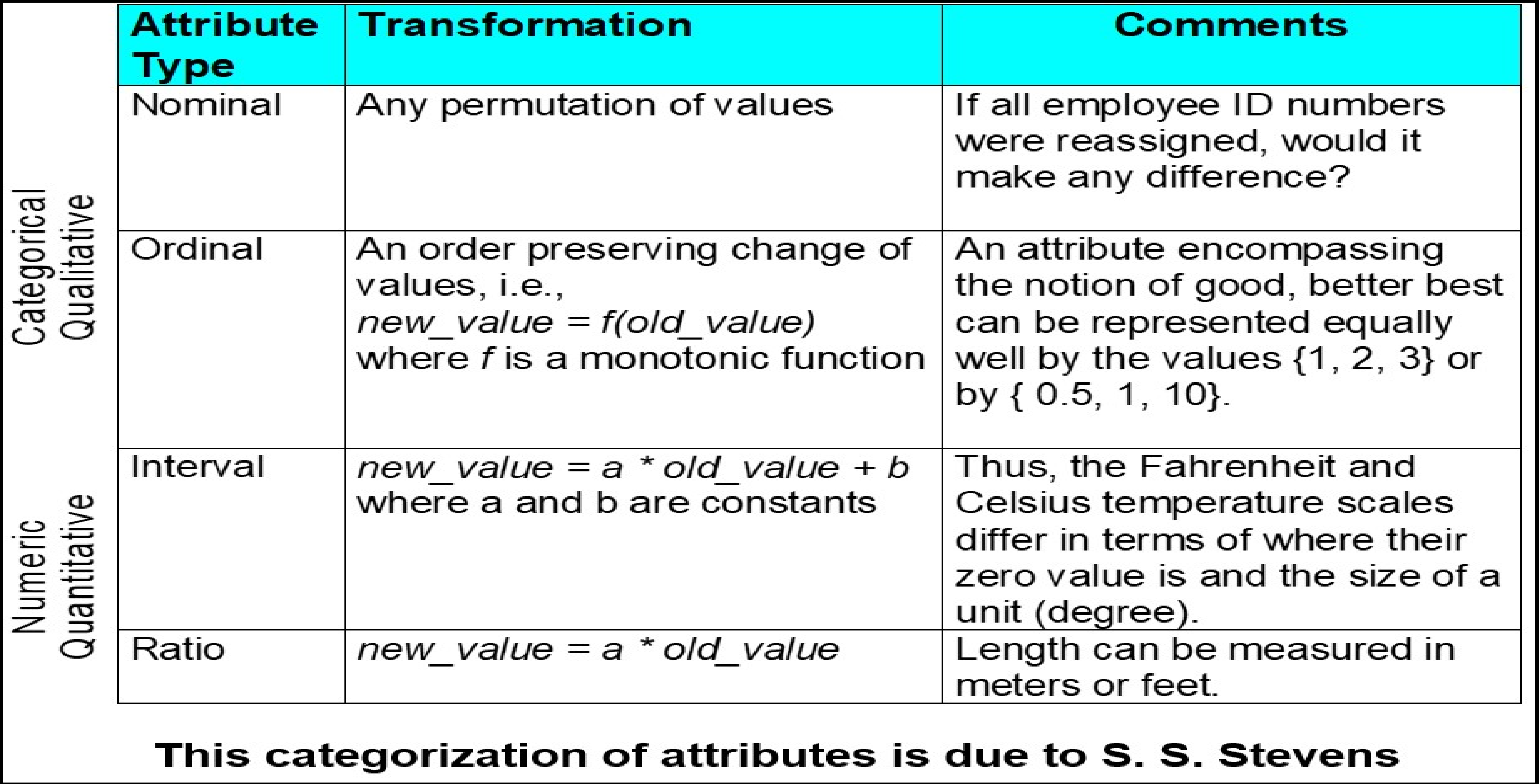
Types of Data



	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*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

This categorization of attributes is due to S. S. Stevens

Types of Data



	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 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.

This categorization of attributes is due to S. S. Stevens

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

ASYMMETRIC ATTRIBUTE

- only presence (a non-zero attribute value) is regarded as important
  - words present in documents
  - items present in customer transactions

- 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.”

Important Characteristics of Data

- **DIMENSIONALITY** (number of attributes)
  - — high dimensional data brings a number of challenges (complexity, …)
- **SPARSITY**
  - — only presence counts (values different from 0 need not to be recorded)
- **RESOLUTION**
  - — patterns depend on the scale (averaging, summarizing, zoom factor, …)
- **SIZE**
  - — type of analysis may depend on size of data (complexity, algorithm, metric, …)

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

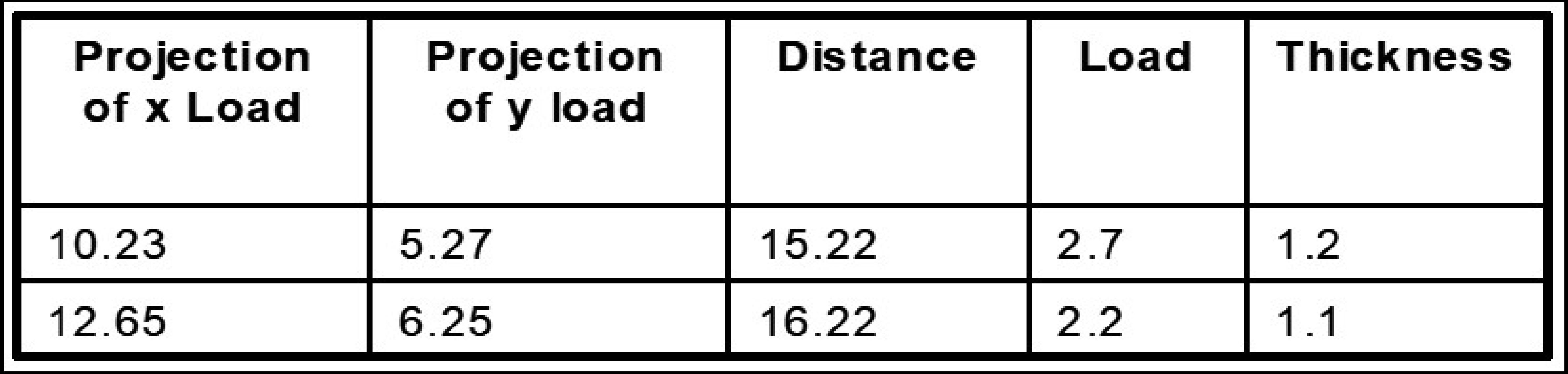
RECORD DATA

- 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 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	2.7	1.2
12.65	6.25	16.22	2.2	1.1

DOCUMENT DATA

- Each document becomes a 'TERMS VECTOR'
  - each term is a component (attribute) of the terms 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

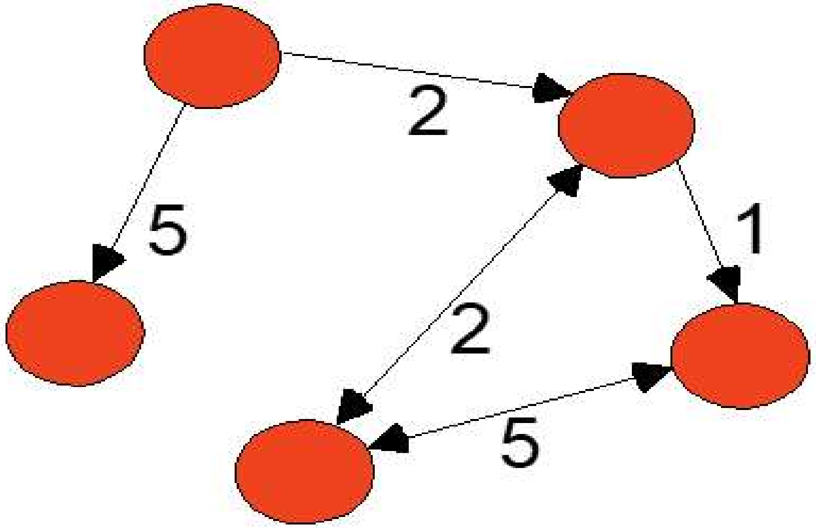
TRANSACTION DATA

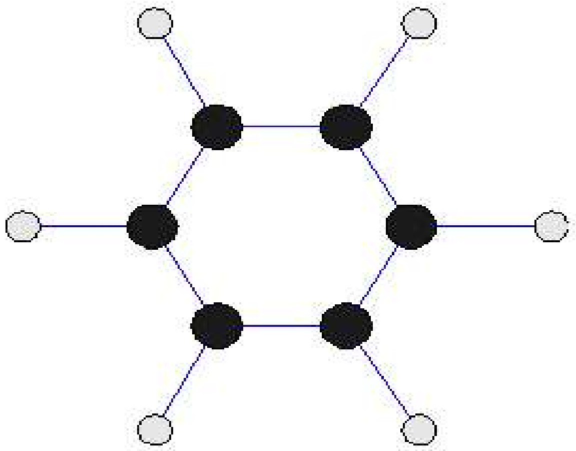
- A special type of data, where
  - each transaction involves a set of items
  - for example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items
  - can represent transaction data as record data

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

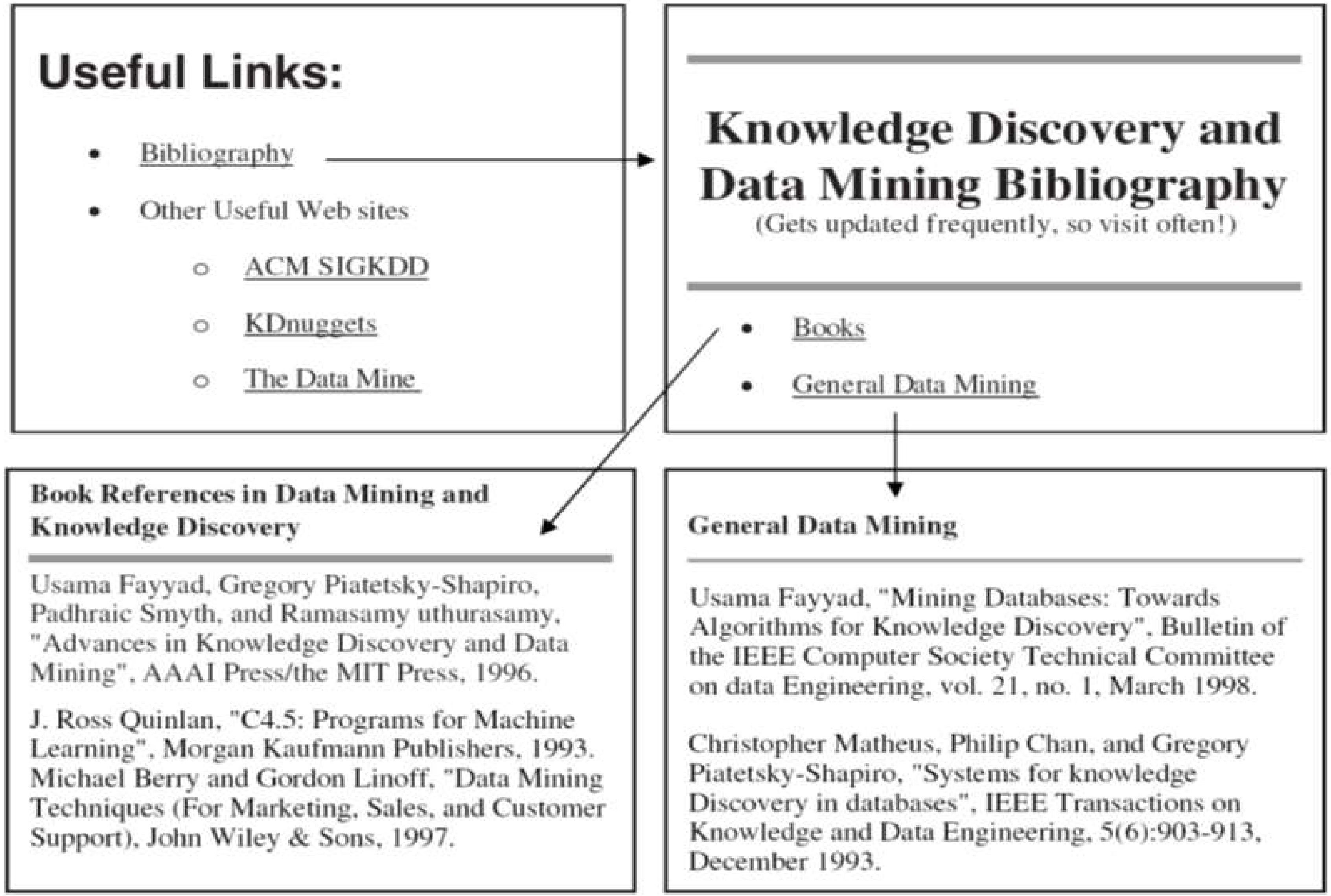
GRAPH DATA

Examples: generic graph, a molecule, and webpages





Benzene Molecule: C6H6



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.

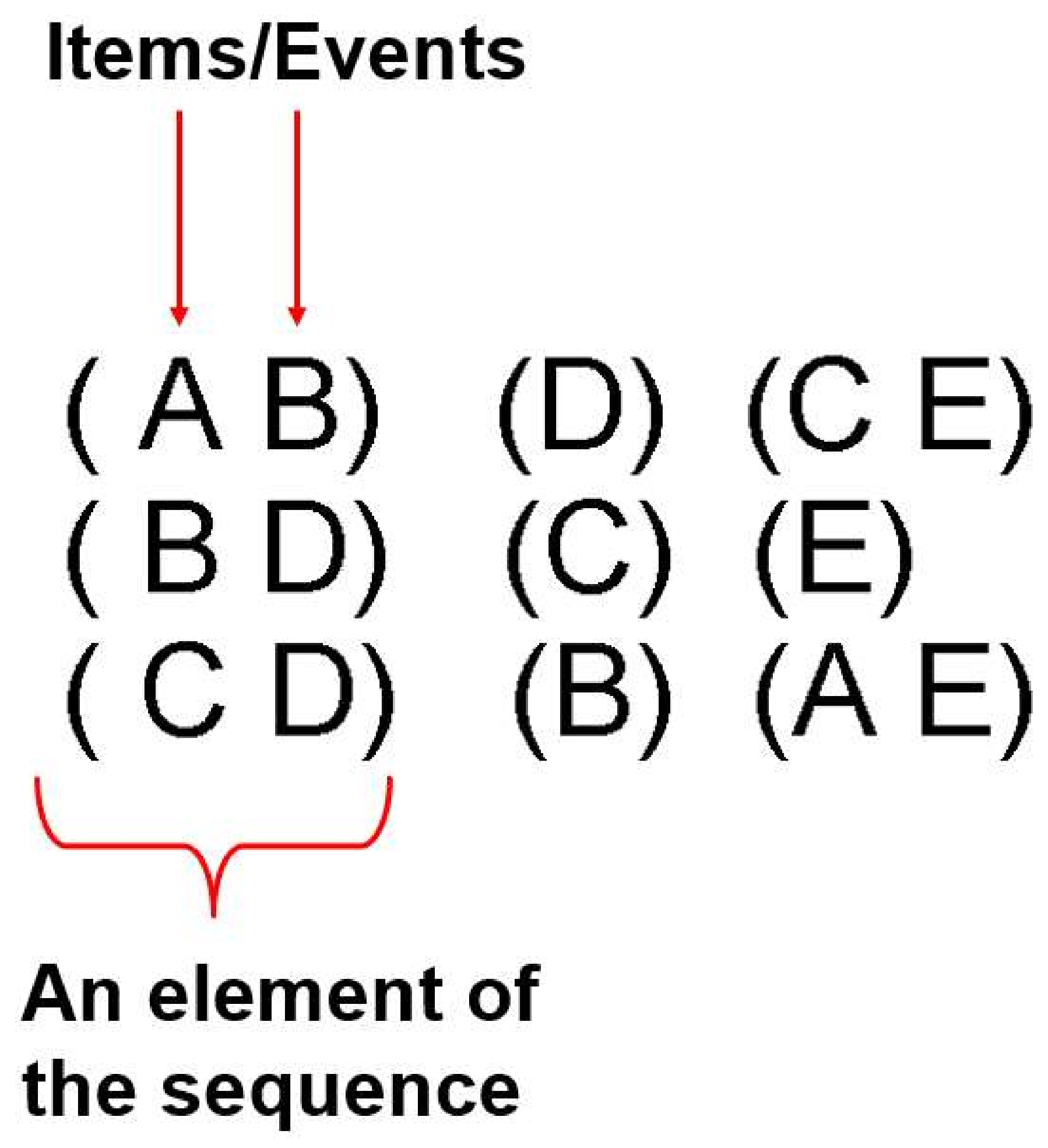
General Data Mining

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.

ORDERED DATA

- Sequences of transactions



Items/Events

(A B) (D) (C E)

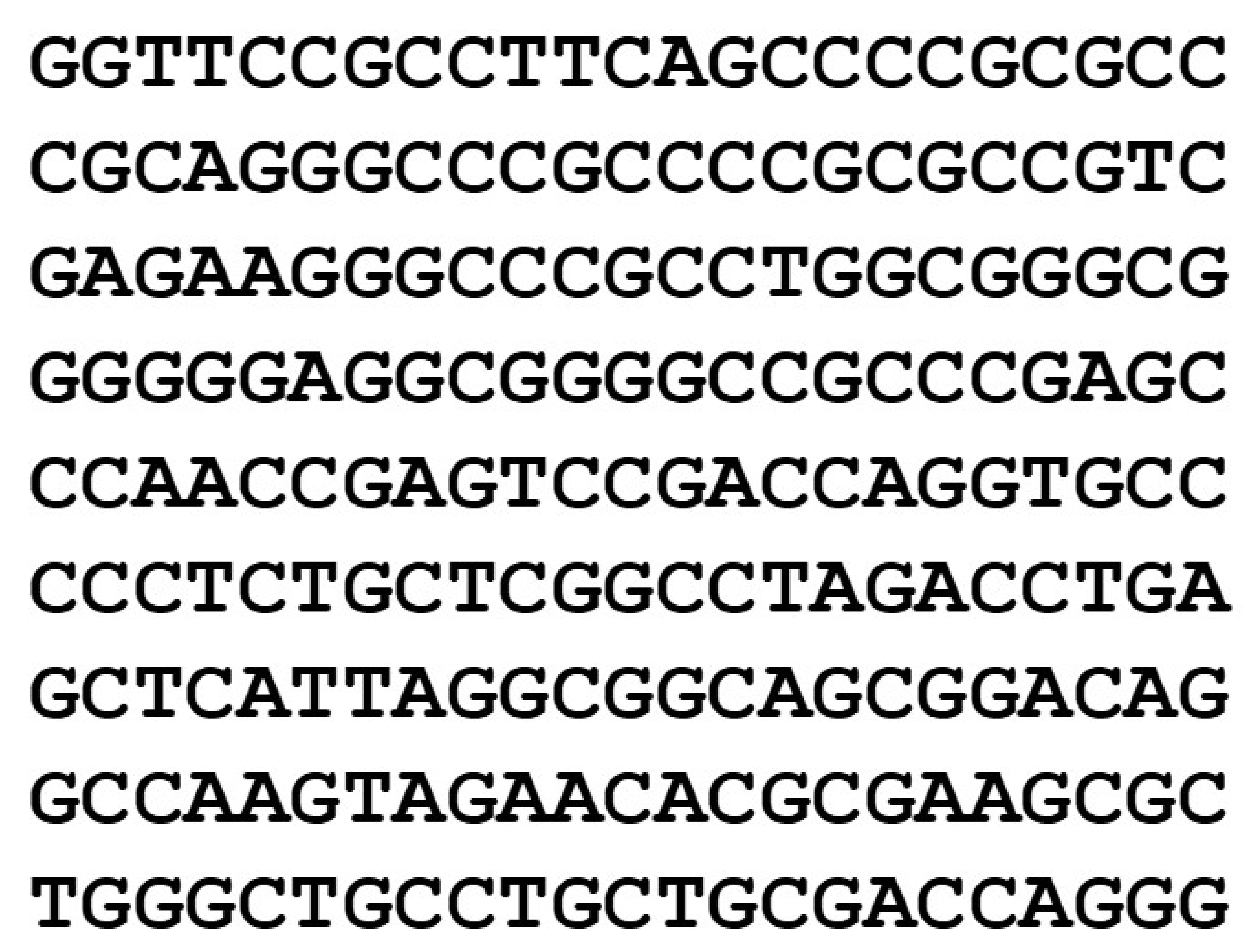
(B D) (C) (E)

(C D) (B) (A E)

An element of the sequence

ORDERED DATA

- Genomic sequence data

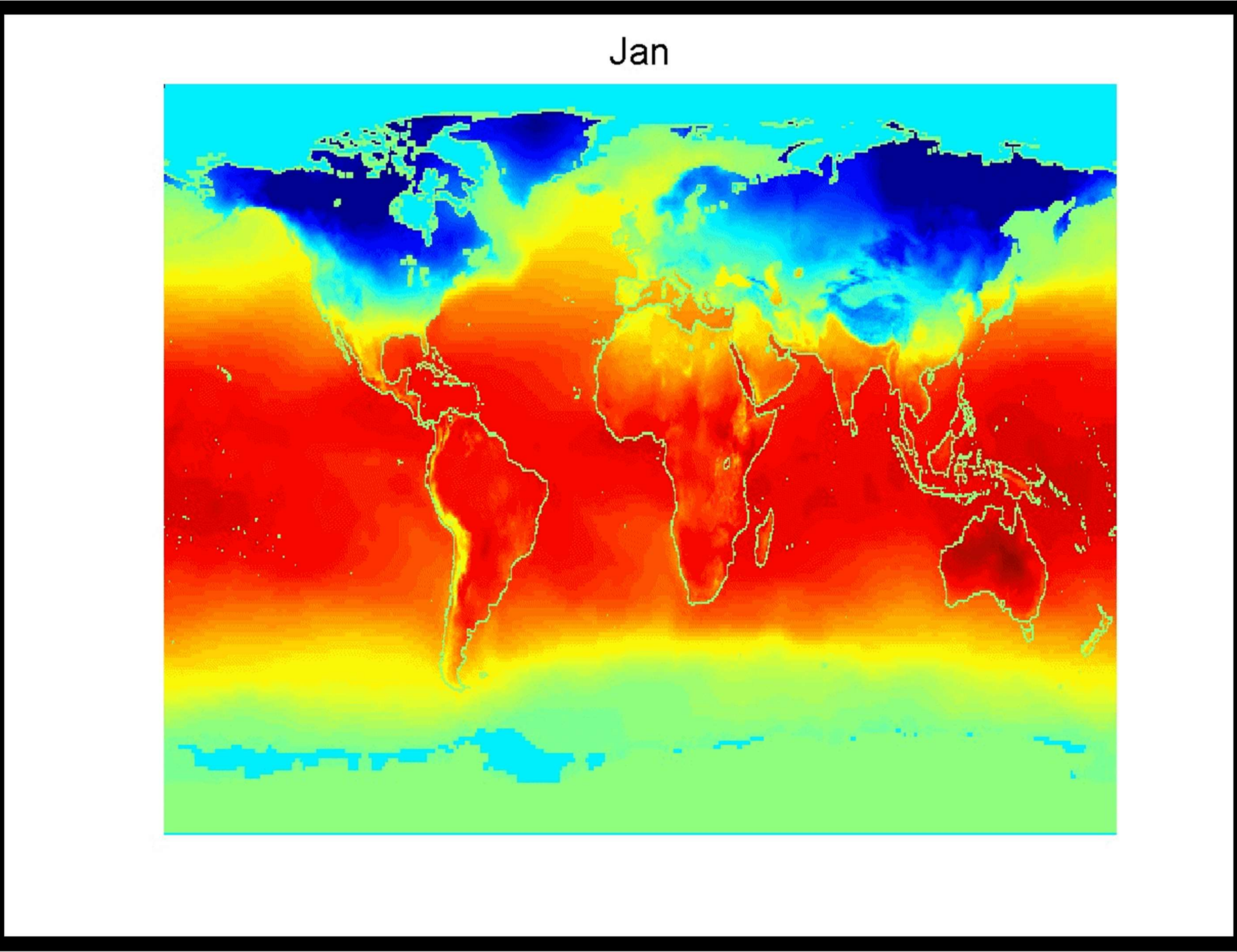


```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

ORDERED DATA

- Spatio-Temporal data

average monthly temperature of land and ocean



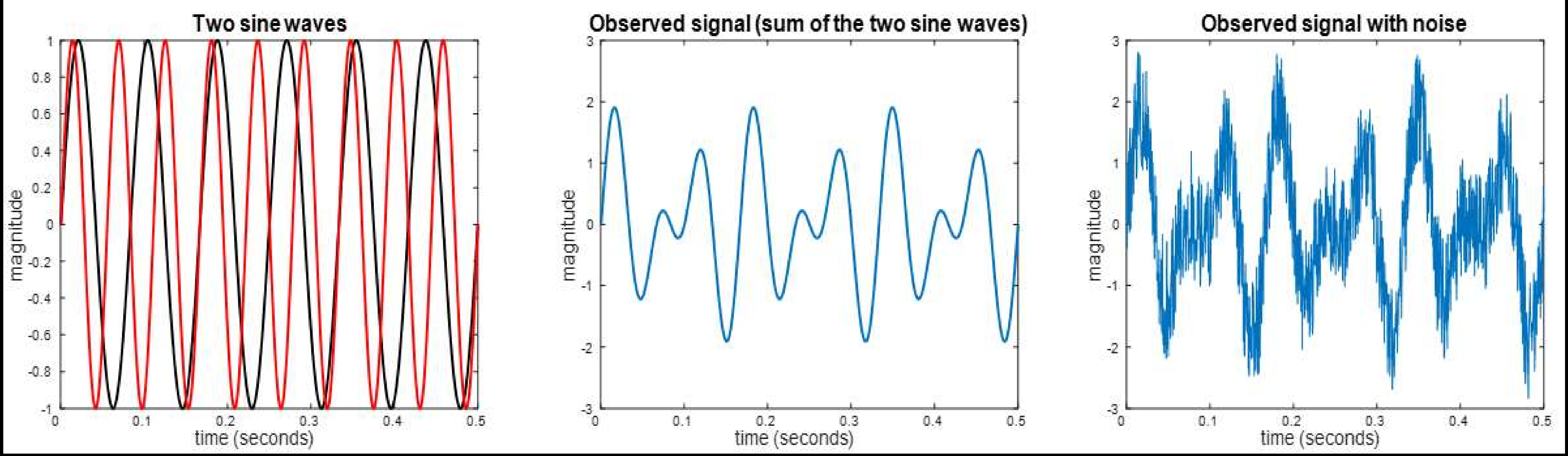
DATA QUALITY

- 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
  - wrong data
  - fake data
  - missing values
  - duplicate data

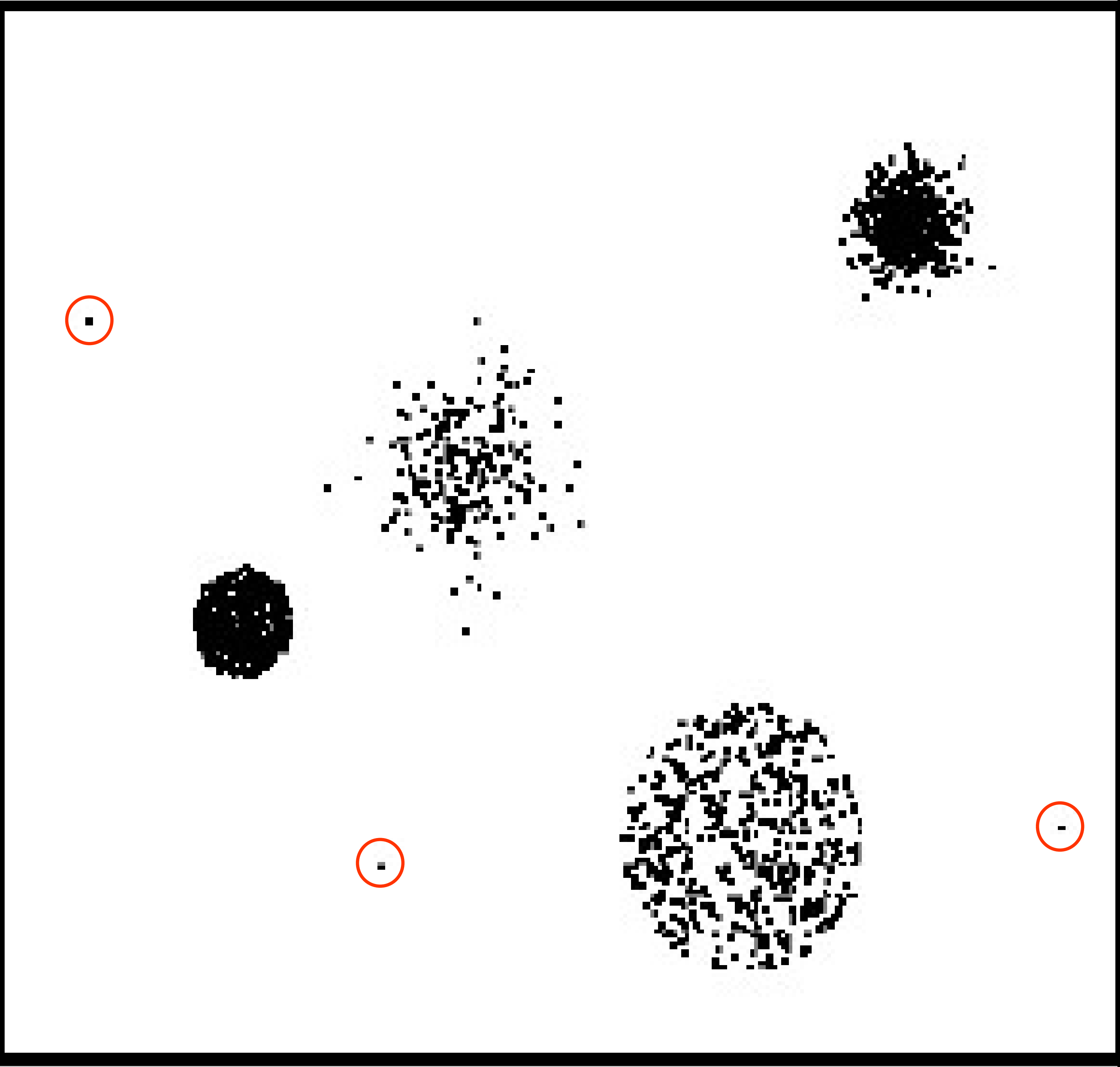
NOISE

- for objects, noise is an extraneous object
- for attributes, 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
  - the figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - the magnitude and shape of the original signal is distorted



OUTLIERS

- are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **case 1:** outliers are noise that interferes with data analysis
  - **case 2:** outliers are the goal of our analysis
    - credit card fraud
    - intrusion detection
- causes?



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 or variables
  - — estimate missing values
    - Example: time series of temperature
    - Example: census results
  - — 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 (entity linking)

- When should duplicate data not be removed?

RECAP

- **DATA OBJECT** and **ATTRIBUTE**
- **TYPES OF ATTRIBUTES**
- **IMPORTANT CHARACTERISTICS OF DATA**
- **TYPES OF DATA SETS**
- **DATA QUALITY**