Data Mining and Machine Learning

LECTURE 2: TYPE OF DATA

Dr. Edgar Acuna
Department of Mathematical Science

Universidad de Puerto Rico- Mayaguez
academic.uprm.edu/eacuna

Structured Data

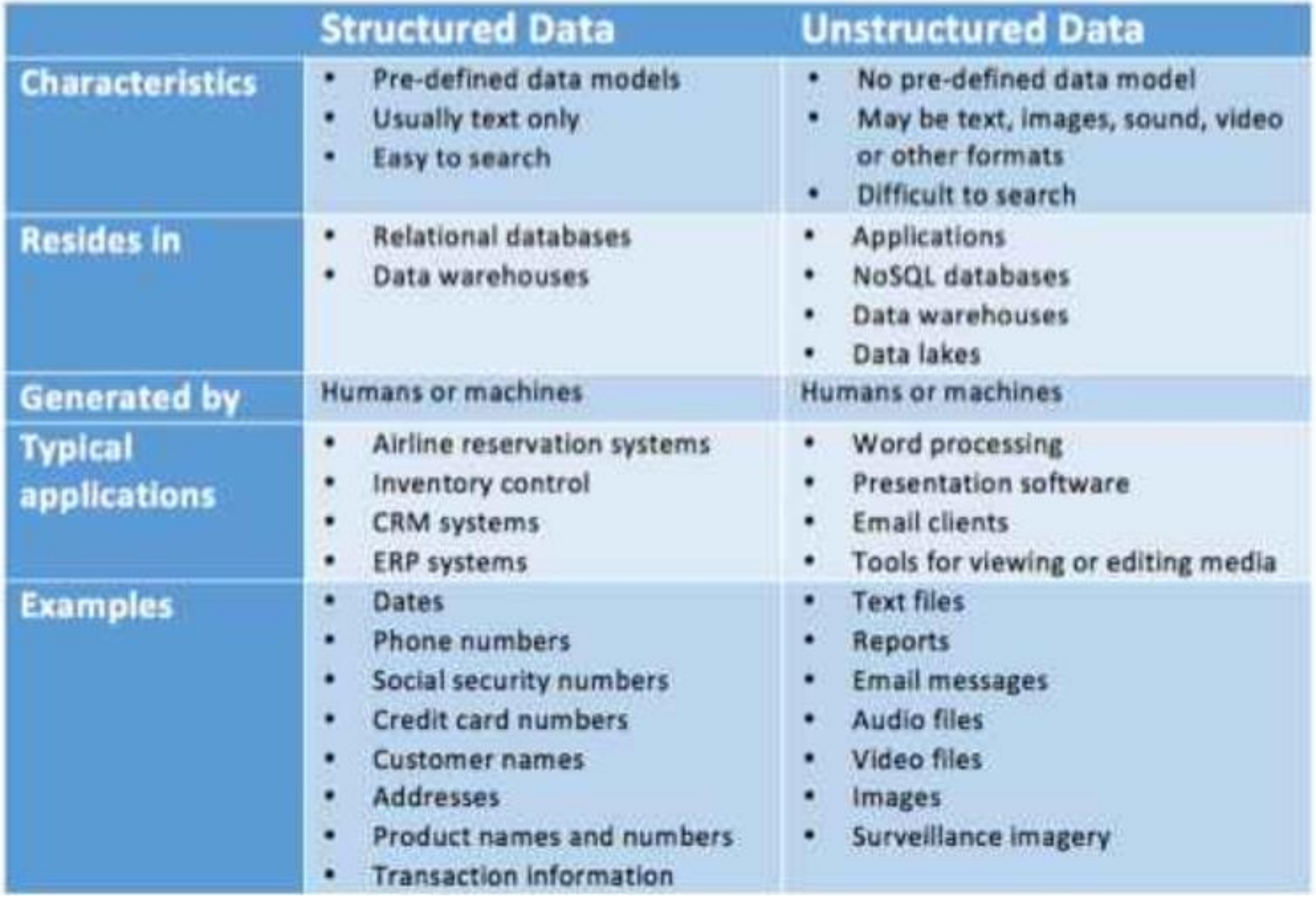
Data with a high degree of organization, such that inclusion in a relational database is seamless and easily searchable by simple search engine algorithms or other search operations. This kind of data accounts for about 20 percent of the data that is out there.

Sources of structured data can be computer-or-machine generated: sensor data, weblog data, point-of-sale data, financial data, ATM data, or human generated imput data, click-stream data, etc.

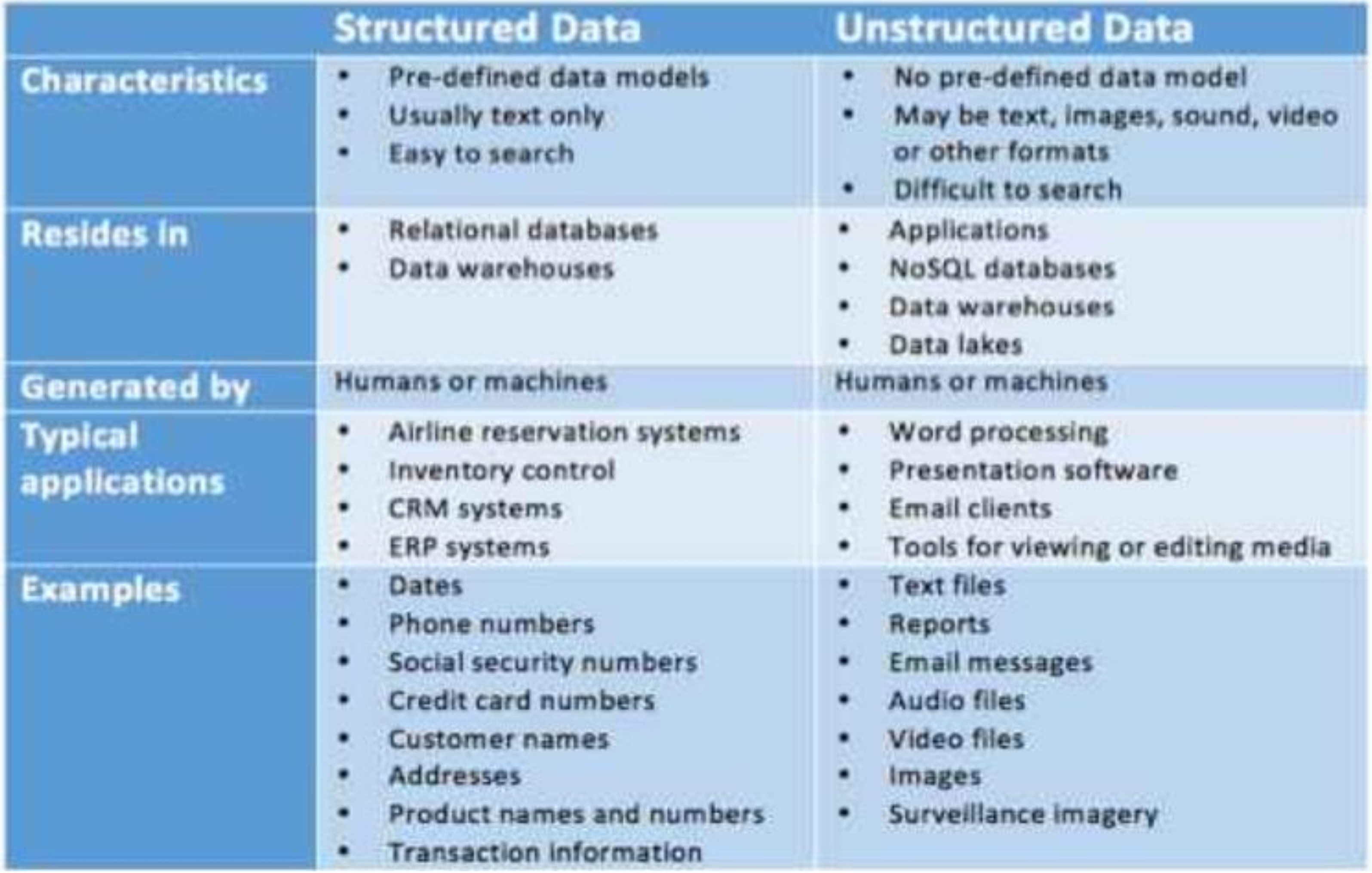
Data Mining mostly uses structured data.

Unstructured Data

- Unstructured data has internal structure but is not structured via pre-defined data models or schema.
- It may be textual or non-textual, and human- or machine-generated.
- Human generated include: Text Files, e-mail (the message field) social media(Twitter, Facebook, LinkedIn, Instagram, etc), mobile media, communications, media (audio and videos files, etc).
- Machine generate include: Sattelite imaginery, scientific data, digital surveillance, sensor data, etc.
- This may also be stored within a non-relational database like: MongoDB, HBase, Cassandra



	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery



	Structured Data	Unstructured Data
Characteristics	- Pre-defined data models - Usually text only - Easy to search	- No pre-defined data model - May be text, images, sound, video or other formats - Difficult to search
Resides in	- Relational databases - Data warehouses	- Applications - NoSQL databases - Data warehouses - Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	- Airline reservation systems - Inventory control - CRM systems - ERP systems	- Word processing - Presentation software - Email clients - Tools for viewing or editing media
Examples	- Dates - Phone numbers - Social security numbers - Credit card numbers - Customer names - Addresses - Product names and numbers - Transaction information	- Text files - Reports - Email messages - Audio files - Video files - Images - Surveillance imagery

From www.datamation.com

Semi-structured data

This data maintains internal tags and markings that identify separate data elements, which enables information grouping and hierarchies. Both documents and databases can be semi-structured. This type of data only represents about 5-10% of the whole data world, but has critical business usage cases.

Email is a very common example of a semi-structured data type. Email's native metadata enables classification and keyword searching without any additional tools.

- The use of semi-structured development centers on easing data transport issues. Sharing sensor data, as are Web-based data sharing and transport: electronic data interchange (EDI), many social media platforms, document markup languages, and NoSQL databases.

Semi-structured data[2]

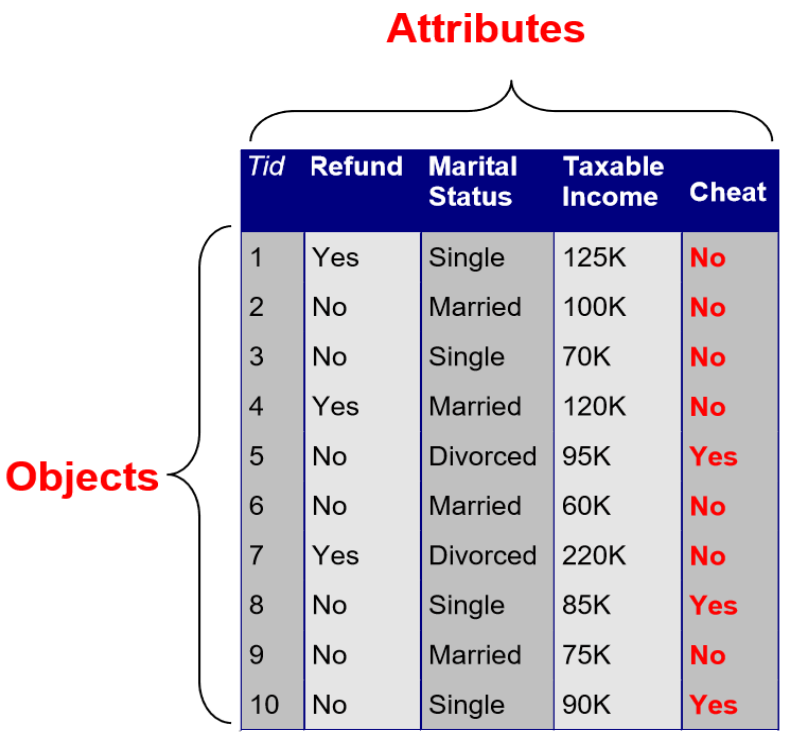
Examples of semi-structured data include: Documents in Markup Language XML, in JSON (JavaScript Object Notation) and NoSQL databases(MongoDB, Hive, Cassandra, Couch DB, etc). These databases area ideal for transmitting data between web applications and servers.

In Data Science any kind of data is considered for analysis.

The process to convert semi-structured/unstructured data in data ready to be analyzed is called "Data Wrangling".

More on structured 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, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, 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**

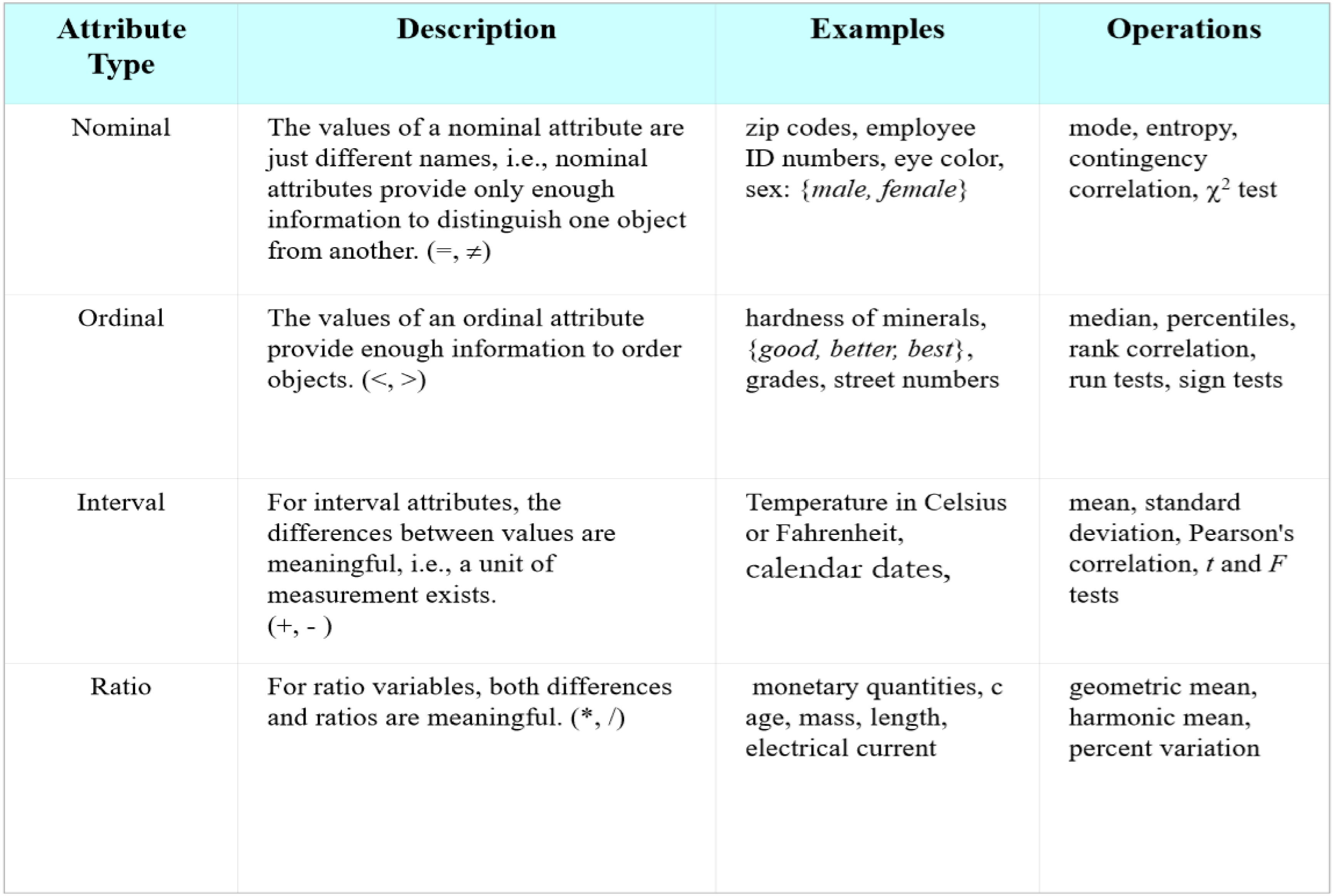
Attributes

- Attribute values are numbers or symbols assigned to an attribute
- According to their scale of measurement, there are four different types of attributes: nominal, ordinal, Interval and Ratio.

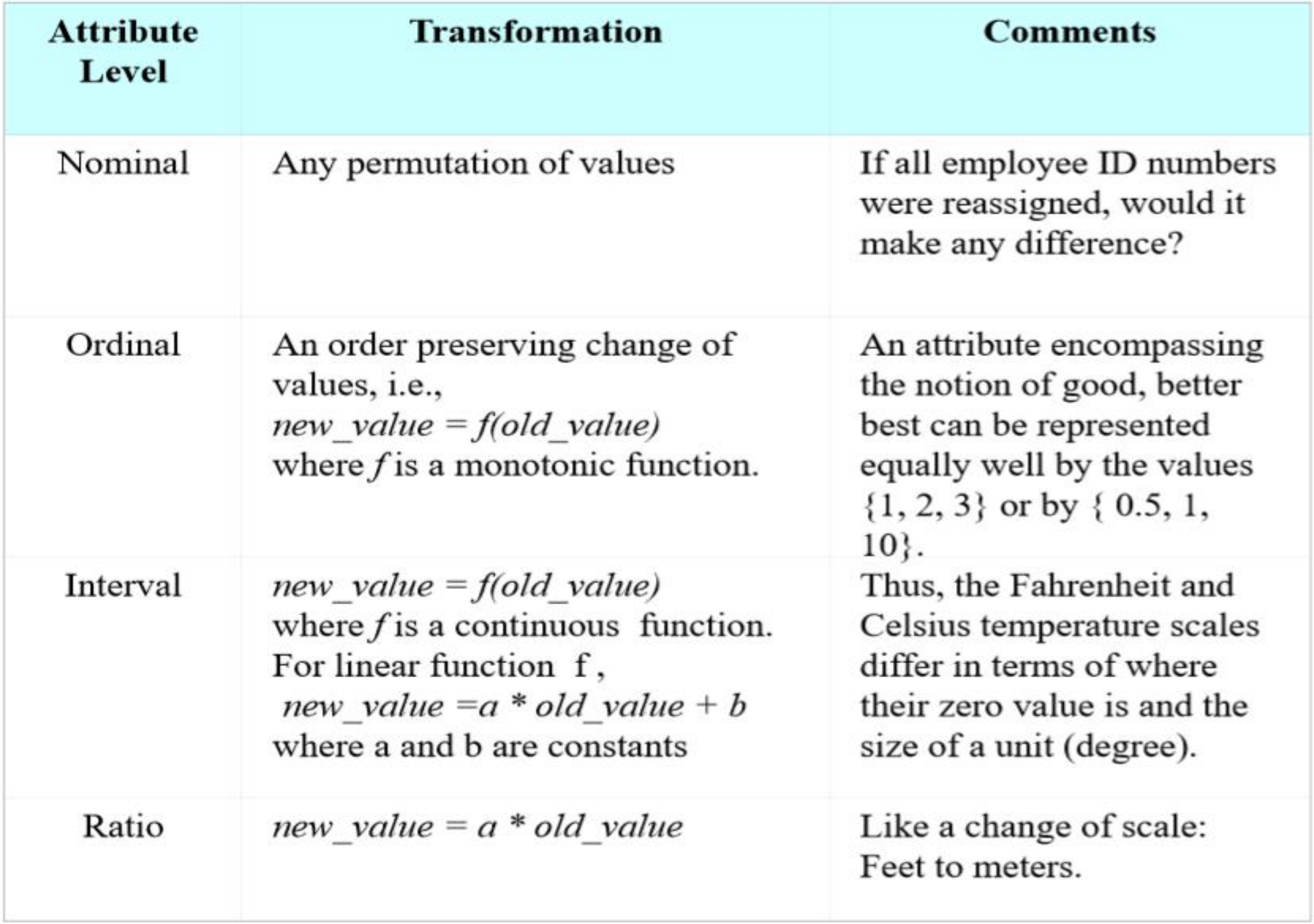
Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties



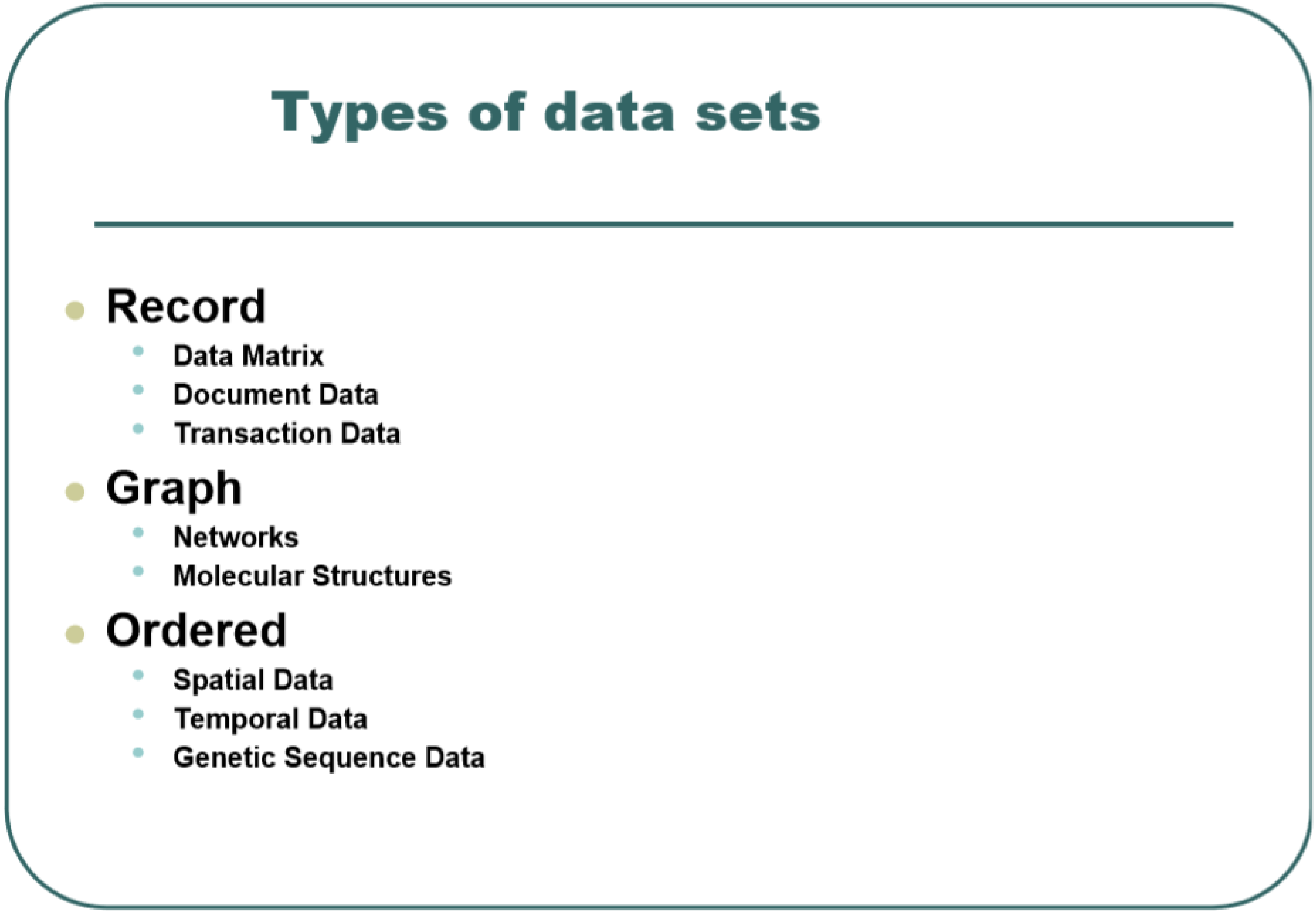
Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency 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. (+, -)	Temperature in Celsius or Fahrenheit, calendar dates,	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	monetary quantities, c age, mass, length, electrical current	geometric mean, harmonic mean, percent variation



Attribute Level	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 = f(old_value)* where *f* is a continuous function. For linear function f , *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*	Like a change of scale: Feet to meters.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: number of car sales per day , number of children in a family, number of certain word in a collection of documents.
  - Often represented as integer variables.
  - Binary attributes are a special case of discrete attributes. Example: fail, Pass.
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - In practice, real values can only be measured and represented using a finite number of digits.

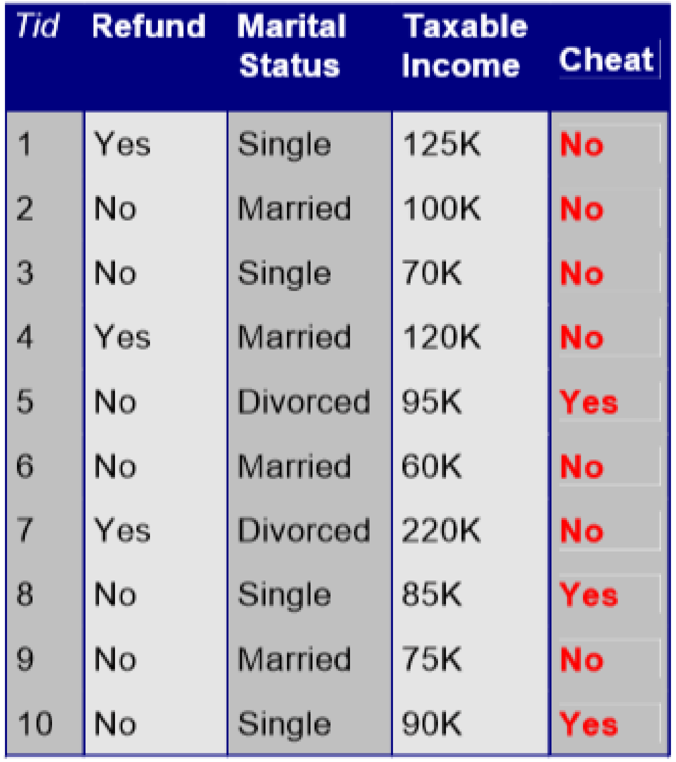


Types of data sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data
- **Graph**
  - Networks
  - Molecular Structures
- **Ordered**
  - Spatial Data
  - Temporal Data
  - Genetic Sequence Data

Record Data

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

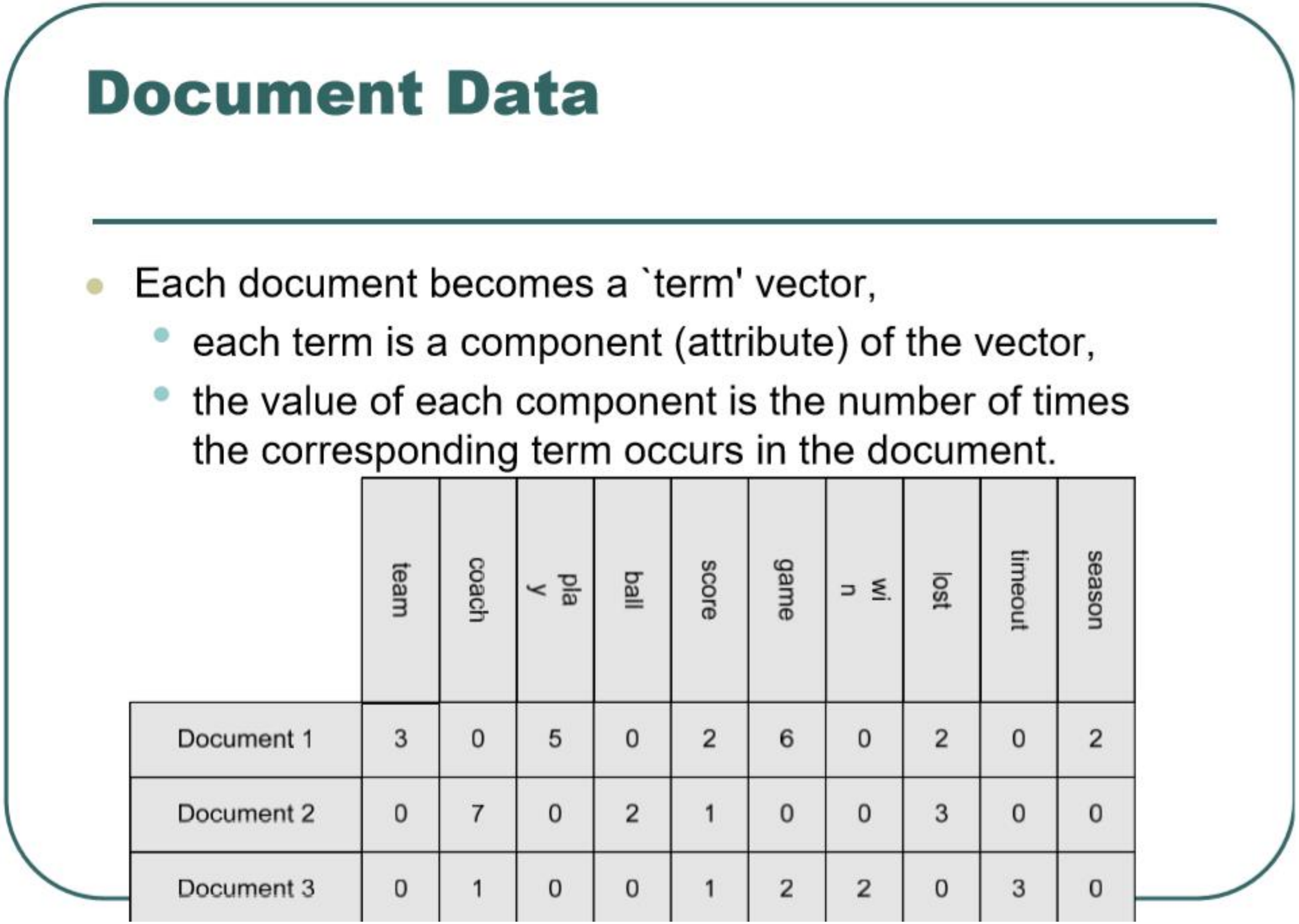


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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



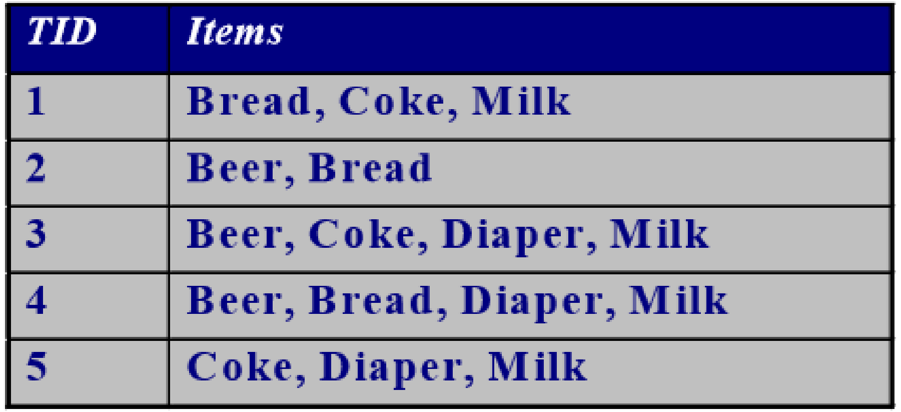
Document Data

- 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

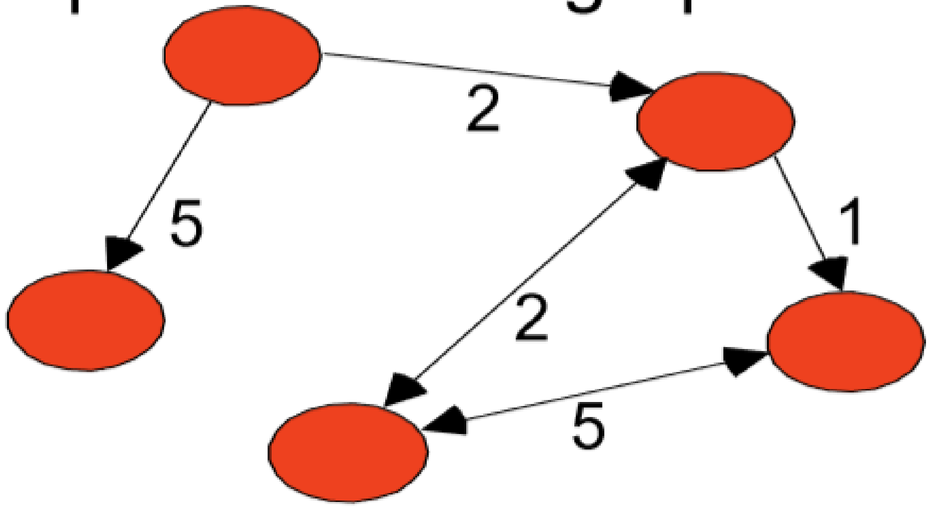
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.



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



Ordered Data

- Genomic sequence data

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

Basis

A=Adenina

C=Citosina

G=Guanina

T=Tianina

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean