Data Analytics

Lecture CS1AC16

Dr Varun Ojha
University of Reading

23/02/2022

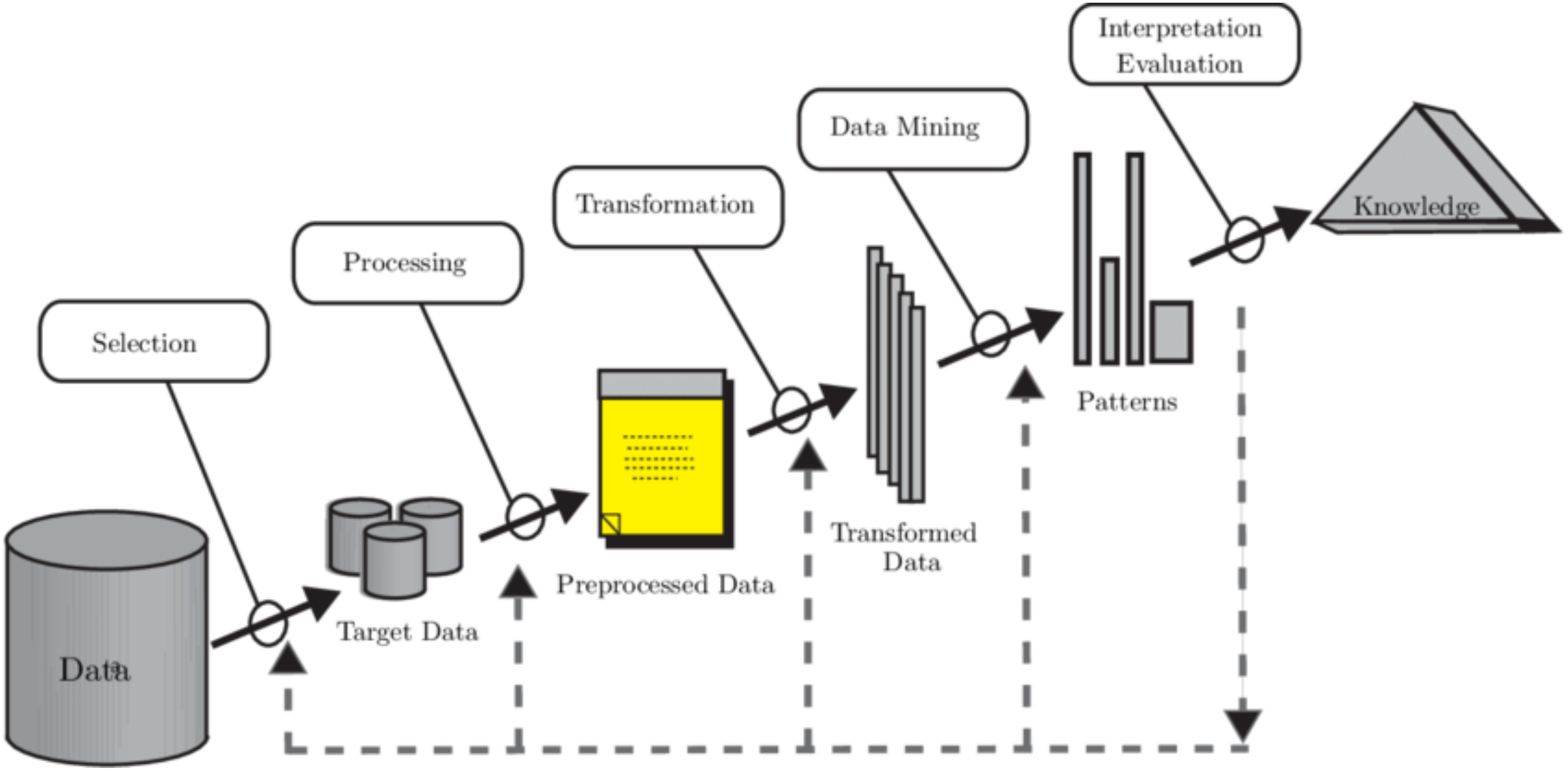
About the Module

- **Lectures:**
  - Week 7, Week 8, Week 9

- **Practical session:**
  - Week 10 and Week 11

- **Assessments**
  1. One Blackboard Class-Test in the last week of term
  2. Exam question in CS1AC16 paper (two in Data Analytics section)

Module Outline



Why data analytics?

- **The world produces tremendous volume of data**

VOLUME of DATA

- **We need to interpret this large volume of data to extract knowledge**
  - Data mining help us do that

DATA MINING

What is Data Analytics?

- **Data analysis** is a process of inspecting, cleansing, transforming, and modelling data into useful information.
- **Data mining** is a process used by companies to turn raw data into useful information.
- **Data science** is an inter-disciplinary field that uses scientific methods, processes, algorithms to extract knowledge from data.

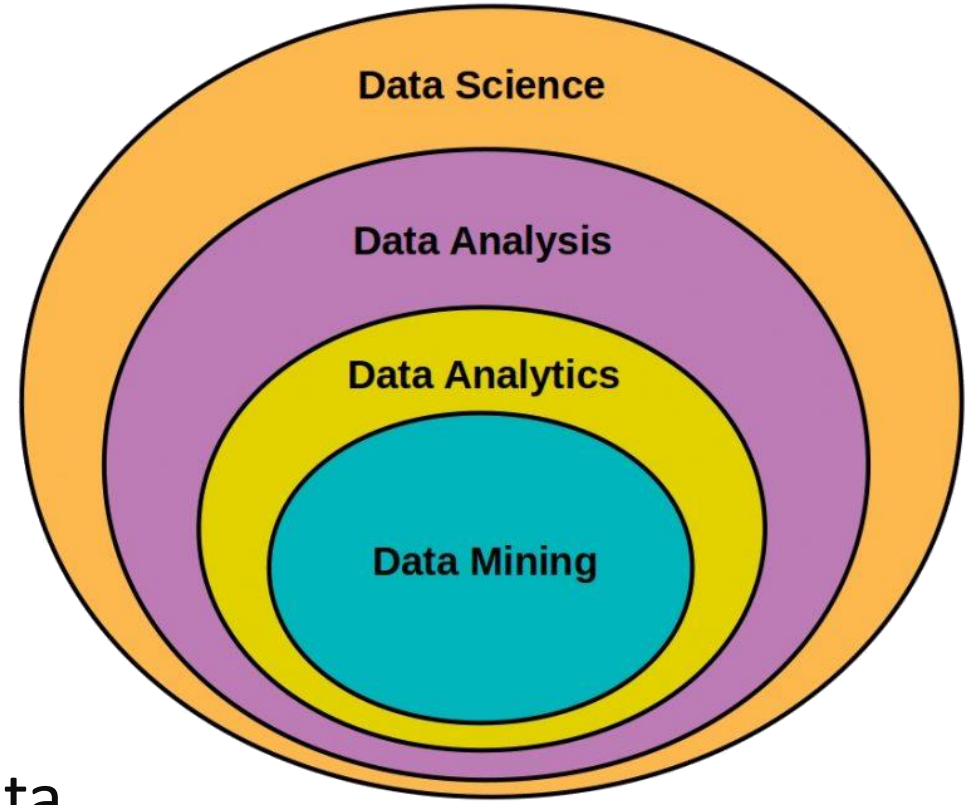
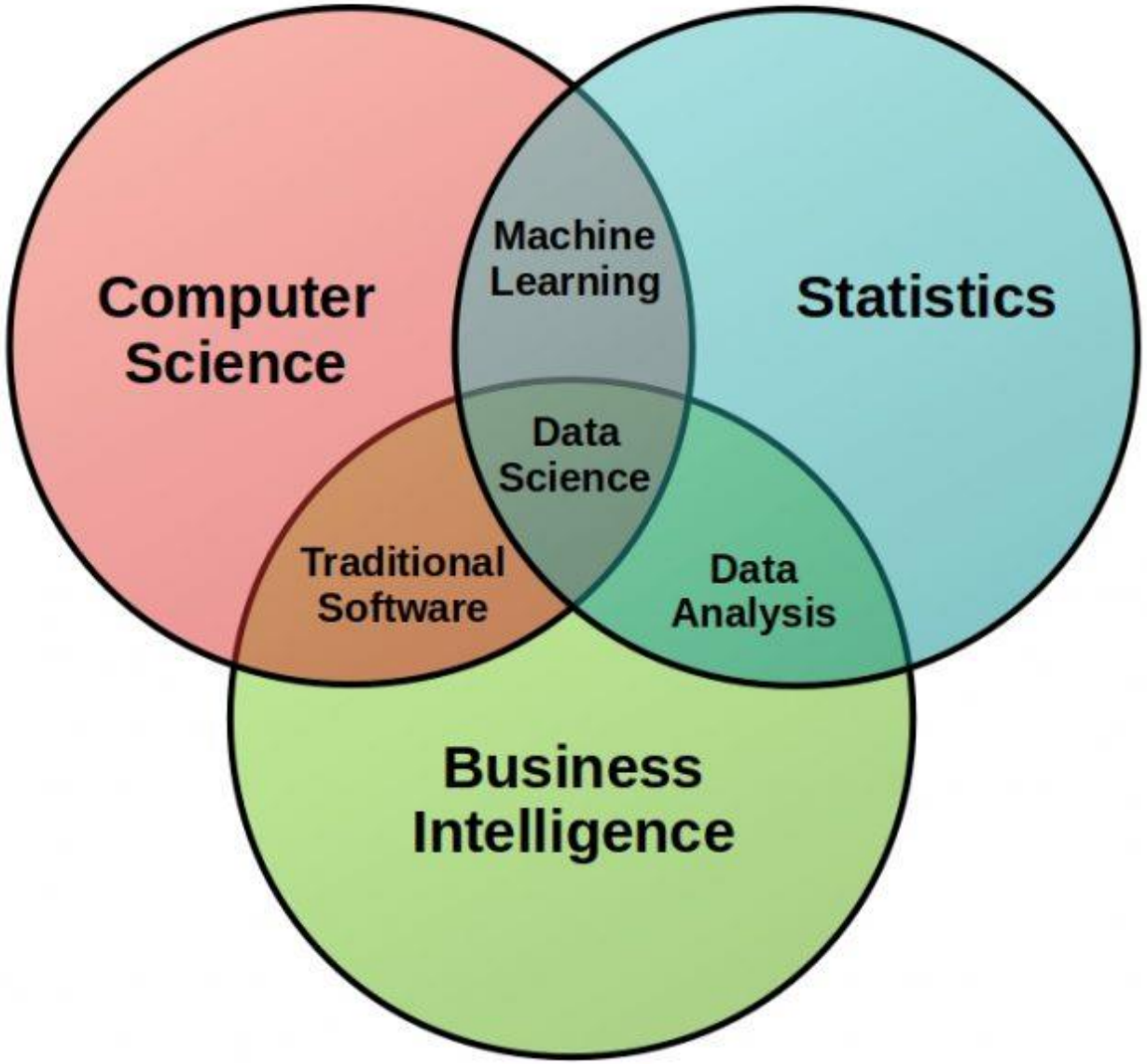


Image source: https://starship-knowledge.com/data-science-vs-data-analysis



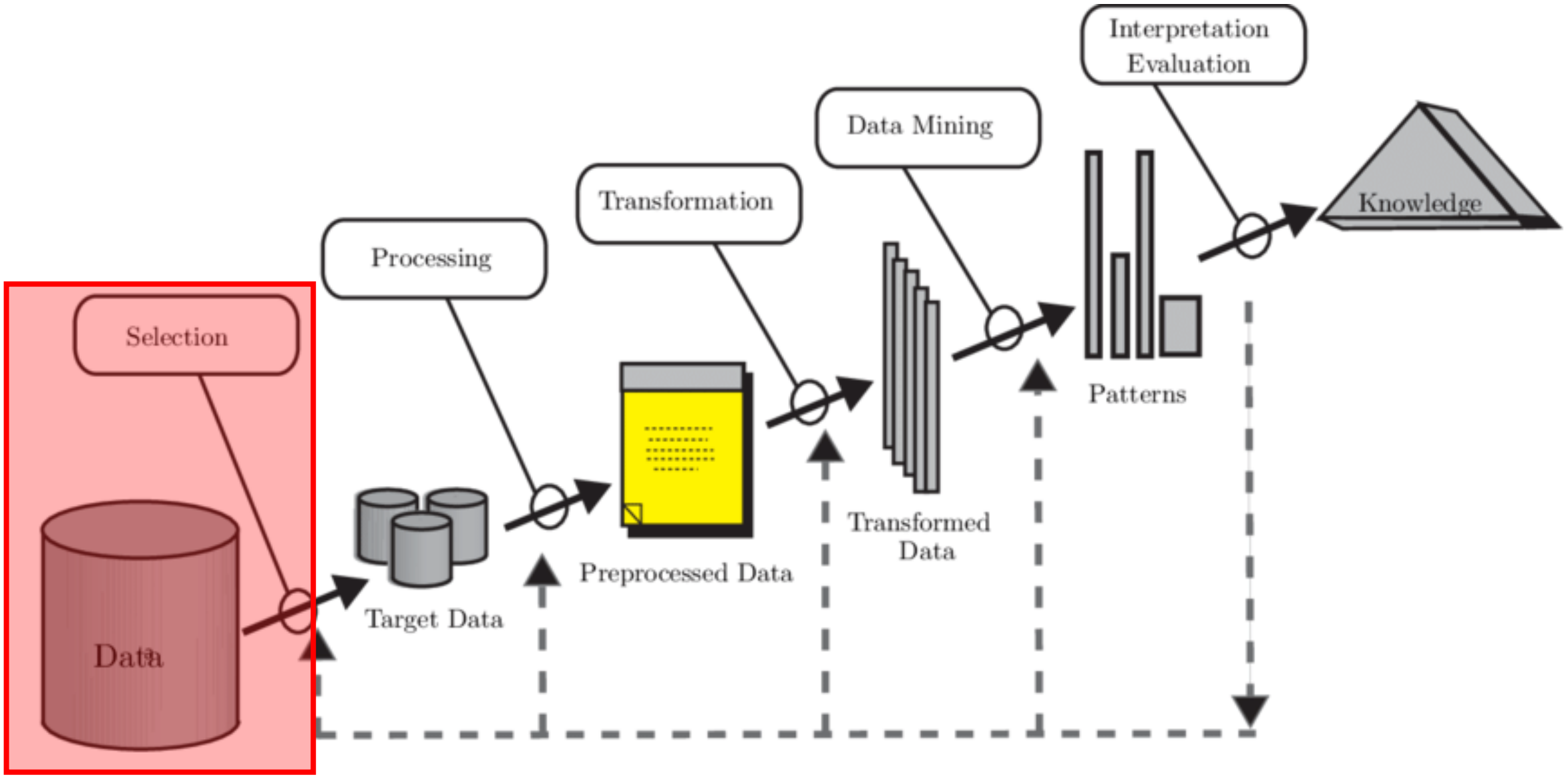
Hugh Conway (2010)

Applications of Data Analytics

- **Scientific Computing**
  - Healthcare – Medical Imaging (A single MRI scan is 21 MB)
  - Biological Experiments - The human genome project (2.2 GB)
  - Cosmology – Imaging a Black hole (5 petabytes)
  - Physics Experiments – The Large Hadron Collider (330 petabytes of data)

- **Commercial Applications**
  - Banking – Credit card use, fraud detection, product recommendation
  - Retail – Product recommendation, logistics
  - Social Networks – Ad targeting, product recommendation

Module Outline



What is Data?

- Noun, a plural of datum (as appear in dictionary)

- (used with a plural verb) individual facts, statistics, or items of information:
  - These data represent the results of our analyses.
  - Data are entered by terminal for immediate processing by the computer

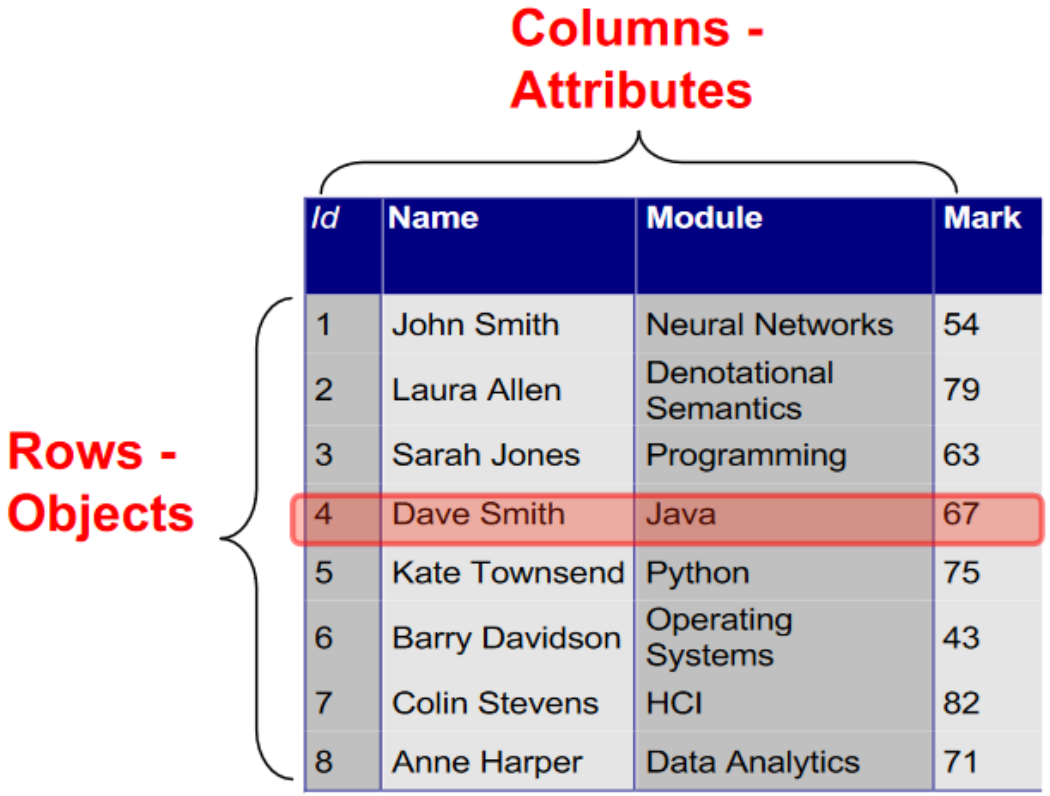
- (used with a singular verb) a body of facts; information:
  - Additional data is available from the president of the firm.

How to Store Data?

- Tabular form. D
  - Tabular data is made up of a table with rows and columns
- Each row consists of **entities**, **objects** or **instances**
  - Each student in RISIS database
- Each column describes **attributes** or **features**
  - Age, course, grade
- Each feature will have a value
  - 19, computer science, 70

Instances (or objects)

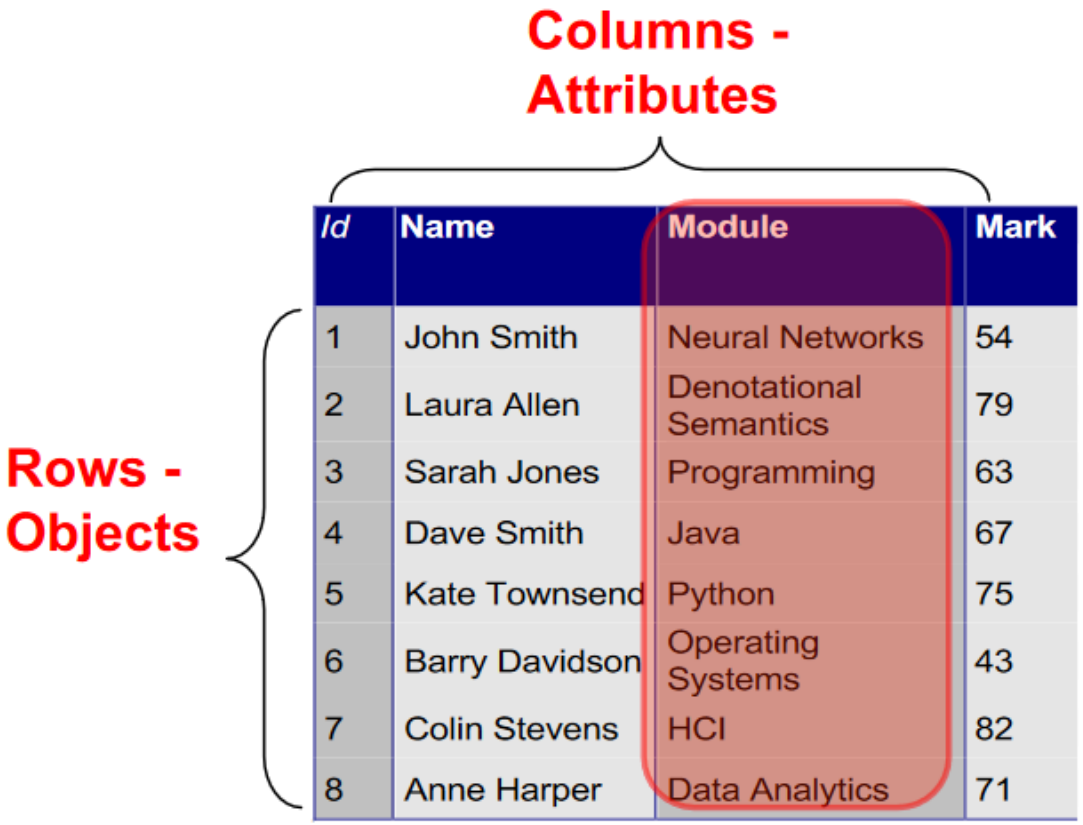
- In tabular data, one **row** will correspond to a single **object** or **instance**, or **sample.**



Id	Name	Module	Mark
1	John Smith	Neural Networks	54
2	Laura Allen	Denotational Semantics	79
3	Sarah Jones	Programming	63
4	Dave Smith	Java	67
5	Kate Townsend	Python	75
6	Barry Davidson	Operating Systems	43
7	Colin Stevens	HCI	82
8	Anne Harper	Data Analytics	71

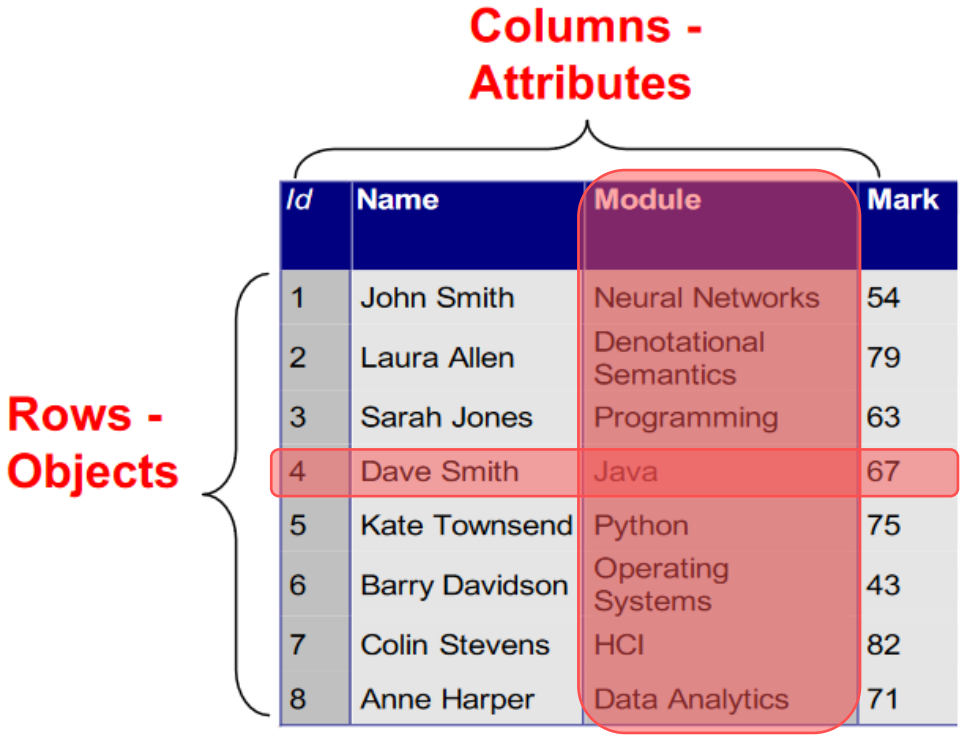
Features (or Attributes)

- A column corresponds to an attribute, also called a **feature** or a **variable**, or an **attribute**.



Features (or Attributes)

- **Attributes** are therefore properties of **objects** that we would like to record.
- For this RISIS example the features are id, name, course and grade of each student



Example of a Real World Data

- The Vancouver street trees dataset

Information | Table | Map | Analyze | Export | API

	TREE_ID	CIVIC_NUMBER	STD_STREET	GENUS_NAME	SPECIES_NAME	CULTIVAR_NAME	COMMON_NAME	ASSIGNED	ROOT_BA
1	40060	1906	W 43RD AV	BETULA	PENDULA		EUROPEAN WHITE BIRCH	N	N
2	40062	1906	W 43RD AV	BETULA	PENDULA		EUROPEAN WHITE BIRCH	N	N
3	40073	1928	W 43RD AV	PRUNUS	CERASIFERA	ATROPURPUREUM	PISSARD PLUM	N	N
4	40082	1956	W 43RD AV	PRUNUS	CERASIFERA	ATROPURPUREUM	PISSARD PLUM	N	N
5	40088	1968	W 43RD AV	PRUNUS	CERASIFERA	ATROPURPUREUM	PISSARD PLUM	N	N
6	40099	2005	W 43RD AV	AESCULUS	HIPPOCASTANUM		COMMON HORSECHESTN...	N	N
7	40106	2028	W 43RD AV	AESCULUS	HIPPOCASTANUM		COMMON HORSECHESTN...	N	N
8	40108	2038	W 43RD AV	AESCULUS	HIPPOCASTANUM		COMMON HORSECHESTN...	N	N
9	40112	2057	W 43RD AV	AESCULUS	HIPPOCASTANUM		COMMON HORSECHESTN...	N	N
10	40113	2060	W 43RD AV	AESCULUS	HIPPOCASTANUM		COMMON HORSECHESTN...	N	N

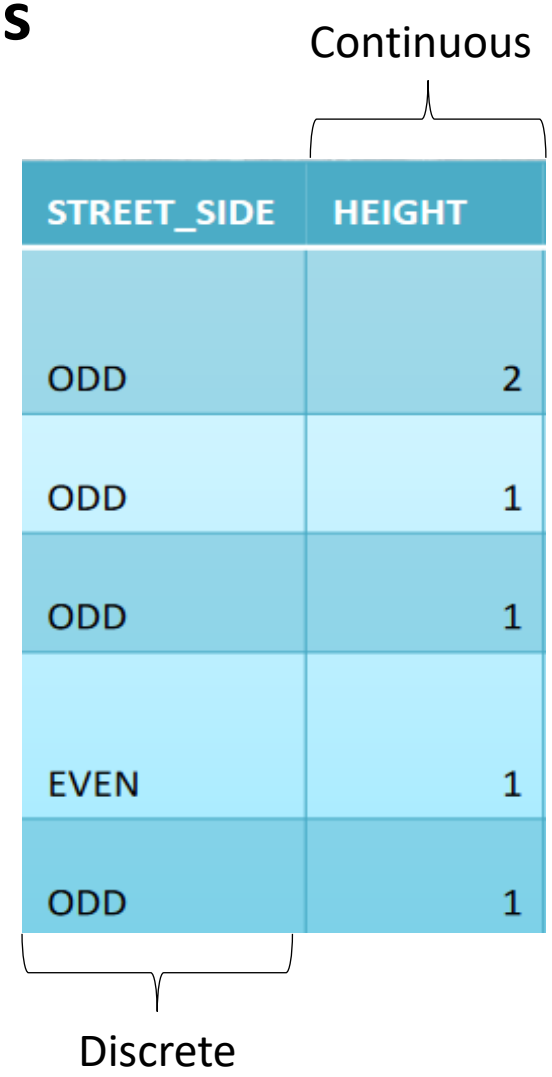
https://opendata.vancouver.ca/explore/dataset/street-trees

Types of Values

- Values of features can be either **discrete** or **continuous**

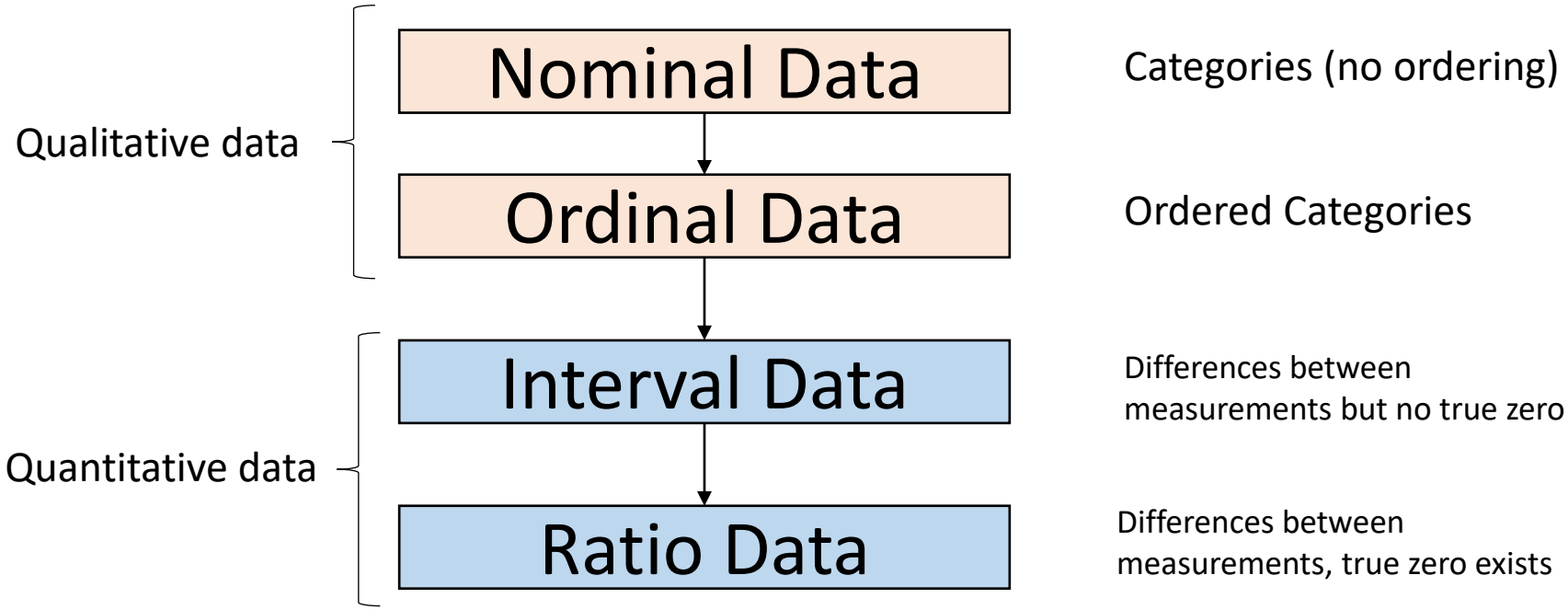
- **Discrete** data can only take certain values
  - Counts, set of words, postcodes
  - Example: the number of students in a class

- **Continuous** data can take any value (within a range)
  - Can be termed as an infinite set of floating point values
  - Examples: Heights, weights, temperatures



Types of Values

- In addition to being discrete or continuous, the characteristic of a variable can be described as:



Nominal Data

- A nominal scale describes a variable with categories that do not have a natural order or ranking.

- Can either be equal or not equal:
  - Polly Vacher building = Polly Vacher building
  - Mauritius != Faroe Islands

- Can be transformed/renamed so long as uniqueness is preserved.

COMMON_NAME
FLOWERING ASH
GOLDEN WHITEBEAM
SHIROFUGEN CHERRY
AKEBONO FLOWERING CHERRY
SHIROFUGEN CHERRY

Ordinal Data

- An ordinal scale is one where the order matters but not the difference between values.
- Can tell if one attribute is **smaller** or **larger** than another
  - E.g. small < medium
- To transform between mappings, order must be preserved

cut
Ideal
Premium
Good
Premium
Good
Very Good

Interval Data

- increases in set steps
- the difference between two values is meaningful.
  - the difference between 90°C and 100°C
- interval data can be added or subtracted

depth
61.5
59.8
56.9
62.4
63.3
62.8

Ratio Data

- has all the properties of an interval variable,

  +

- has a clear definition of 0.0.

- Examples: scores of randomly selected students

  30, 50, 70, and 90.

  - Order in this data?
  - Meaningful difference?
  - Can calculate ratio?

Source of Data

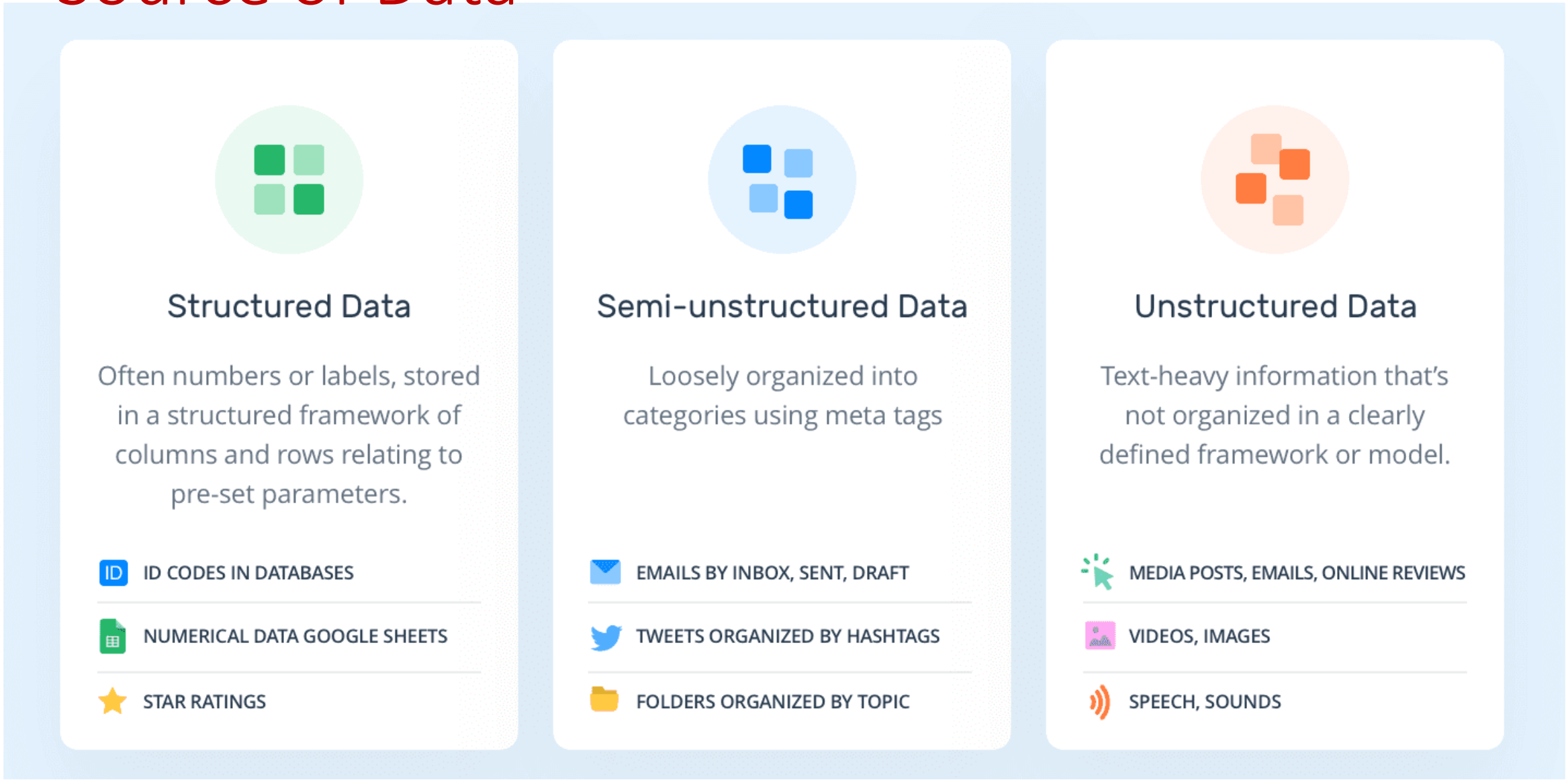


Image source Monkey Learn

Structured Data

- **Structured data** is **data** that adheres to a pre-defined **data** model

- **Structured data** conforms to a tabular format with relationship between the different rows and columns.

- **Examples include:**
  - Excel files, Relational Databases, Graph data

- **Properties of structured data are**
  - Data has Dimensionality, Sparsity and Resolution

Structured Data Examples: 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	Self Assesment
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

Structured Data Examples: Data Matrix

- If every record has same attribute type the data can be represented as a data matrix
- m by n matrix
  - m rows; one per object
  - n columns; one per attribute
- Matrix operations can then be performed
  - multiplication, inverse, eigenvalues, etc

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

Structured Data Examples: Transaction Data

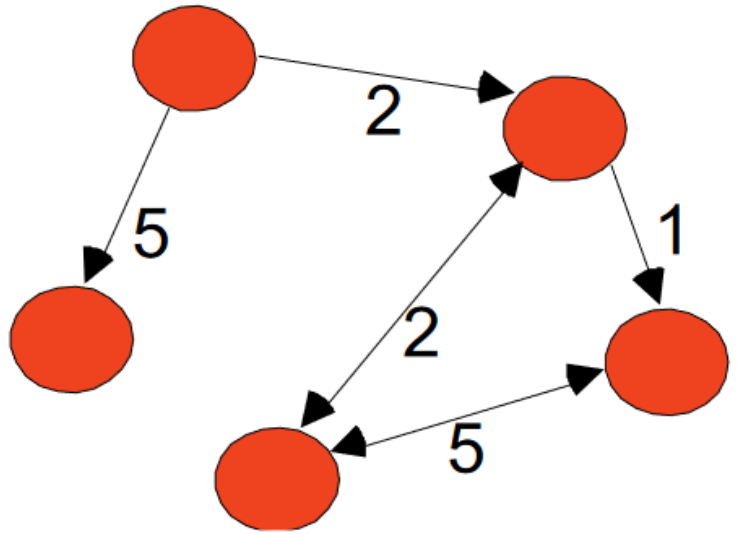
- A special type of record data where
  - Each record is observed as a transaction involving a set of items.
- Example; transactions at a supermarket

TID	*Items*
1	Energy drink, Bleach, Milk
2	Milk, Dog food, Matches
3	Coke, Tissues, Washing up liquid
4	Pasta, Milk, Cheese, Bananas
5	Chips

Structured Data Examples: Graph Data

- A graph database uses graph structures for semantic queries
- The graph relates the data items in the store to a collection of nodes and edges
  - The edges then represent the relationships between the nodes
  - Explicitly lays out any dependencies between nodes of data
- Examples: Generic graph and HTML Links

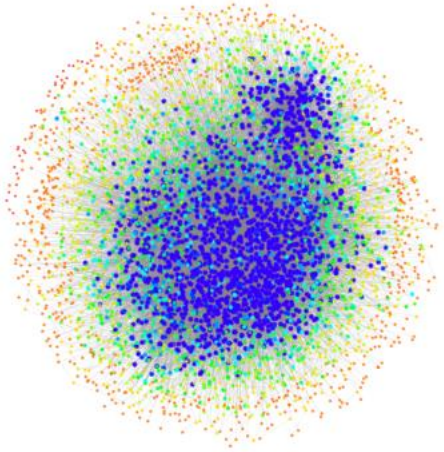
```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System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



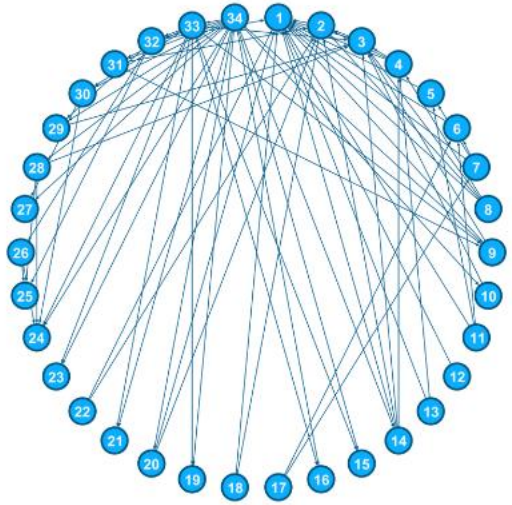
Uses of Graph Data

- In a graph database:
  - Nodes can have attributes
  - Edges can also have attribute

- These characteristics allow statistics to be calculated in the context to the rest of the network, this is known as network analysis



Biological networks



Social networks

Unstructured Data

- **unstructured data** does not adhere to a pre-defined **data** model
- more **difficult to understand using traditional data mining** algorithms.
- **Examples of "unstructured data"**
  - books, journals, documents, metadata, health records, audio, video, images,
- **Properties of unstructured data**
  - Data has Dimensionality, Sparsity and Resolution

Unstructured Data Examples: Document Data

- Text data is often unstructured e.g. emails, academic papers.
- Natural Language processing (NLP) methods are used to convert text into ordered data structures
  - Often the text is converted to numeric data
  - Each document becomes a ‘term’ vector,
  - Each term is a component (attribute) of the vector,

	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

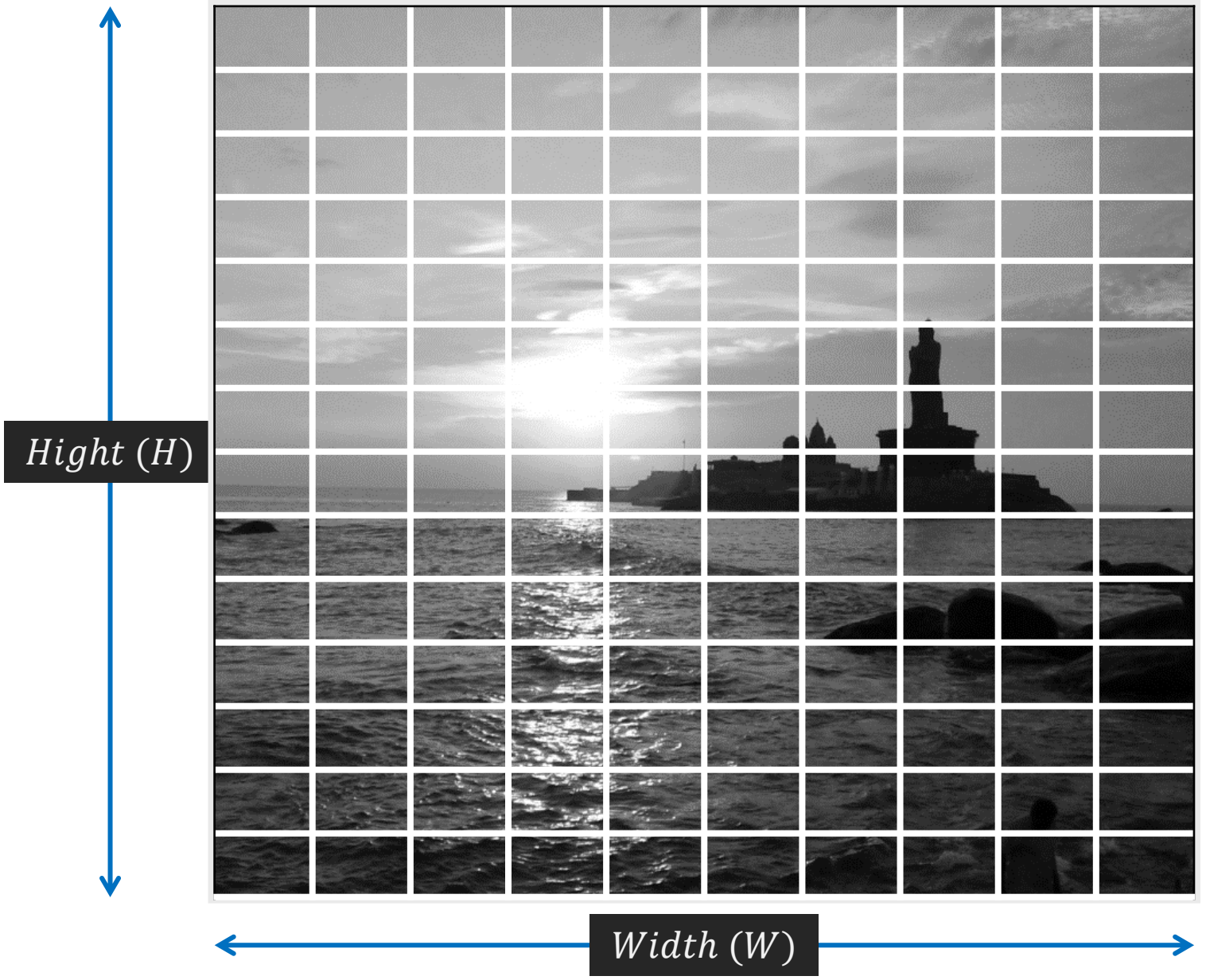
Image Data

Image: Gary scale

$$I = \begin{bmatrix} \boldsymbol{p_{11}} & \cdots & \boldsymbol{p_{1,W}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p_{H,1}} & \cdots & \boldsymbol{p_{H}, p_{W}} \end{bmatrix}$$

For $Hight = 256, Width = 256$

$$I = \begin{bmatrix} p_{11} & \cdots & p_{1,256} \\ \vdots & \ddots & \vdots \\ p_{256,1} & \cdots & p_{256}, p_{256} \end{bmatrix}$$



Data Properties: Data Dimensionality

- **Data Dimensionality -> how many attributes** a dataset has
  - For Example: Every point in a cube has an x, y, z value
- **High dimensionality data is usually bad.**
- The more dimensions -> more difficulty.
  - Known as the '*Curse of Dimensionality*'
  - Where each added attribute results in an exponential *decrease* in predictive power.

	Reddish	Bluish
	1	0
	1	0
	1	0
	1	0
	0	1
	0	1
	0	1
	0	1

	Red	Maroon	Pink	Flamingo	Blue	Turquoise	Seaweed	Ocean
	1	0	0	0	0	0	0	0
	0	**1**	0	0	0	0	0	0
	0	0	**1**	0	0	0	0	0
	0	0	0	**1**	0	0	0	0
	0	0	0	0	**1**	0	0	0
	0	0	0	0	0	**1**	0	0
	0	0	0	0	0	0	**1**	0
	0	0	0	0	0	0	0	**1**

Data Properties: Data Sparsity

- **Data sparsity -> not observing enough data in a dataset.**
- **Data sparsity is usually bad**
  - it means that we are missing information that might be important.
- Essentially, how much of the data is non-zero?
- If the data is sparse, how do we deal with missing data?

Data Properties: Data Resolution

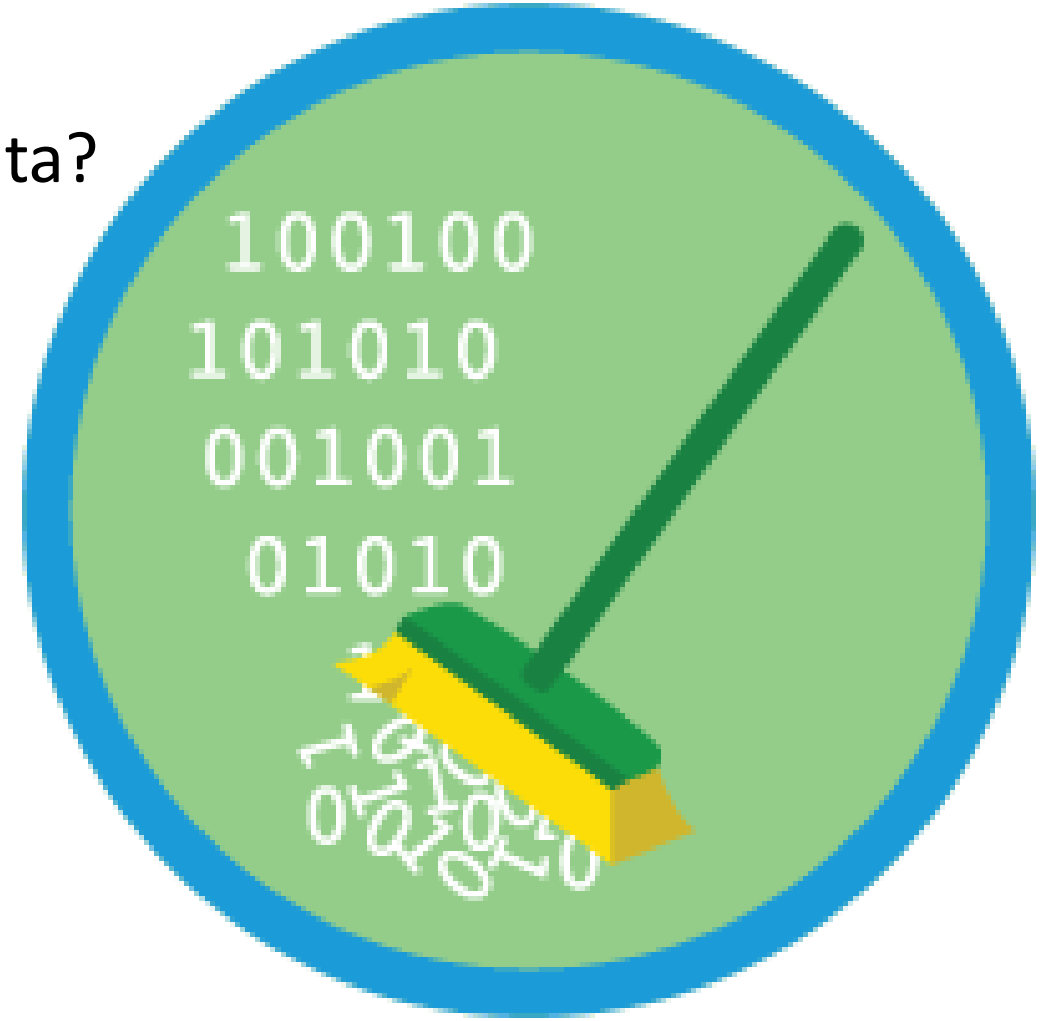
- **Data resolution** can have 2 meanings:
  - It is the ratio between the maximum signal value to the minimum
  - It is the degree to which a change can be theoretically detected

- Therefore, for any application a resolution has to be chosen
  - Too large and you may lose important features
  - Too small and there will be too much data to process

Data Quality

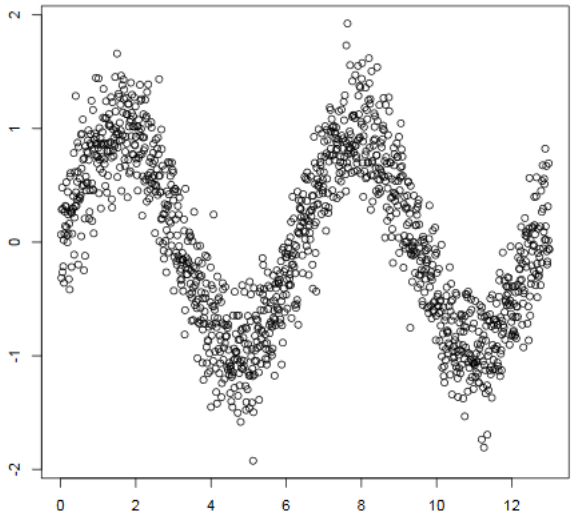
- What kinds of data-quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - **Noise**
  - **Outliers**
  - **Missing values**
  - **Duplicate values**



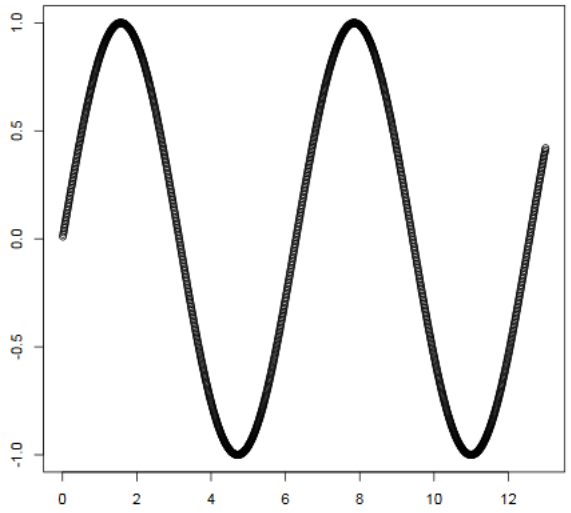
Data Quality: Noise

- **Noise** is a general term for unwanted signals when capturing data.



Sine Wave + Noise

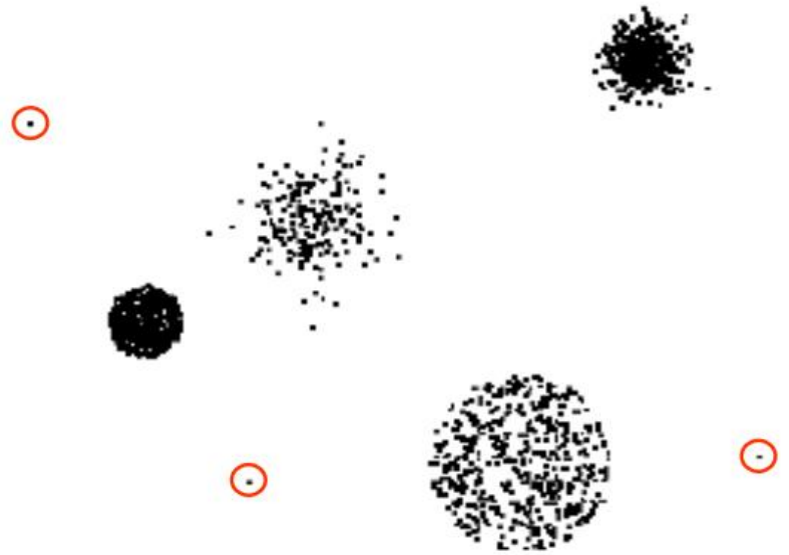
- **Noise reduction**; the recovery of the original signal from the noise.



Sine Wave

Data Quality: Outliers

- **Data objects with unusual or extreme** characteristics outside of the normal distribution
  - Why do they occur?
  - Are they significant or should they be removed?



Data Quality: Missing Values

- **Reasons for missing values:**
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be relevant in every case (e.g. a student who has no grade)
  - Data may have been discarded as an outlier

- **Handling missing values:**
  - Remove all objects that have missing data
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis.

Summary

- **Various data value types**: Discrete and continuous, nominal, ordinal, interval, ratio
- **Structured** (records, relational, graph) and **Unstructured** data (text and images)
- Properties of data: Dimensionality, sparsity and resolution
- Data Quality Issues: Noise, Outliers, Missing values, Duplicate values
- Exercises – KNIME