Everything Starts with Data

Week 3 64/1

Where are we now?

Week	Topics
1	Data Scientist Foundation
2	Basic data analytics: KDD
3	Basic data analytics: Data to Data Product
4	What is Data (Str - eg. nomi, unstr - img, text)
5	Dataset (Basic manipulation)
6	Data quality (e.g., outlier, inconsistency, duplication, etc.)
7	Processes - history, e.g., turn kdd to crisp-dm -> modern
8	Processes - in action
Midterm	

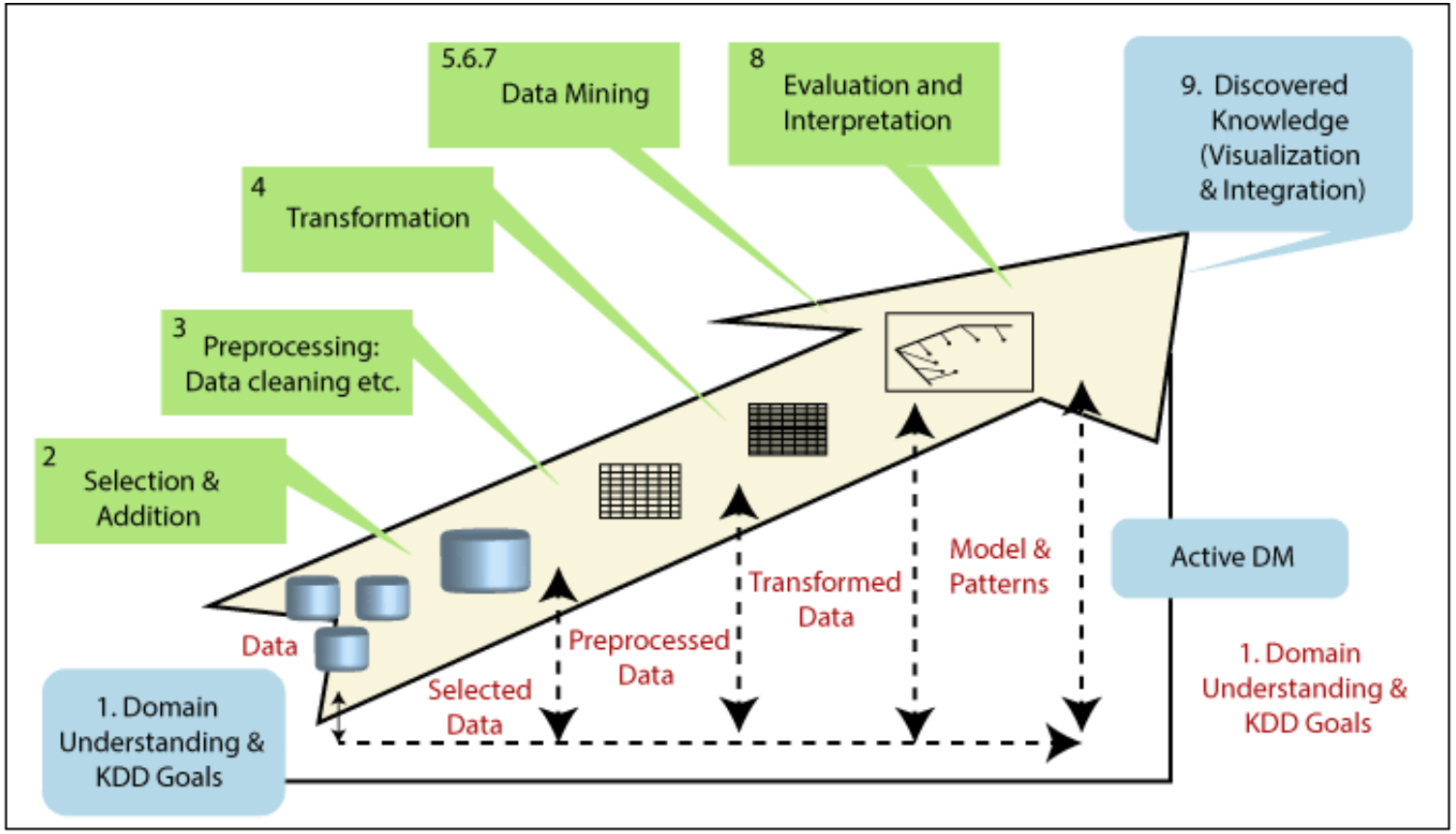
Review what have we learned from last week.

- Basic data analytics: KDD
  - DS process in business
  - Identification of problem, pain points, activities, outputs, outcomes, etc.
  - KDD process overview
  - 3 cases studies (Which one is the hardest? And why?)

Agenda

- What is Data?
- Closer look at KDD process
- From data to data product with WEKA software

Knowledge Discovery in Databases Process



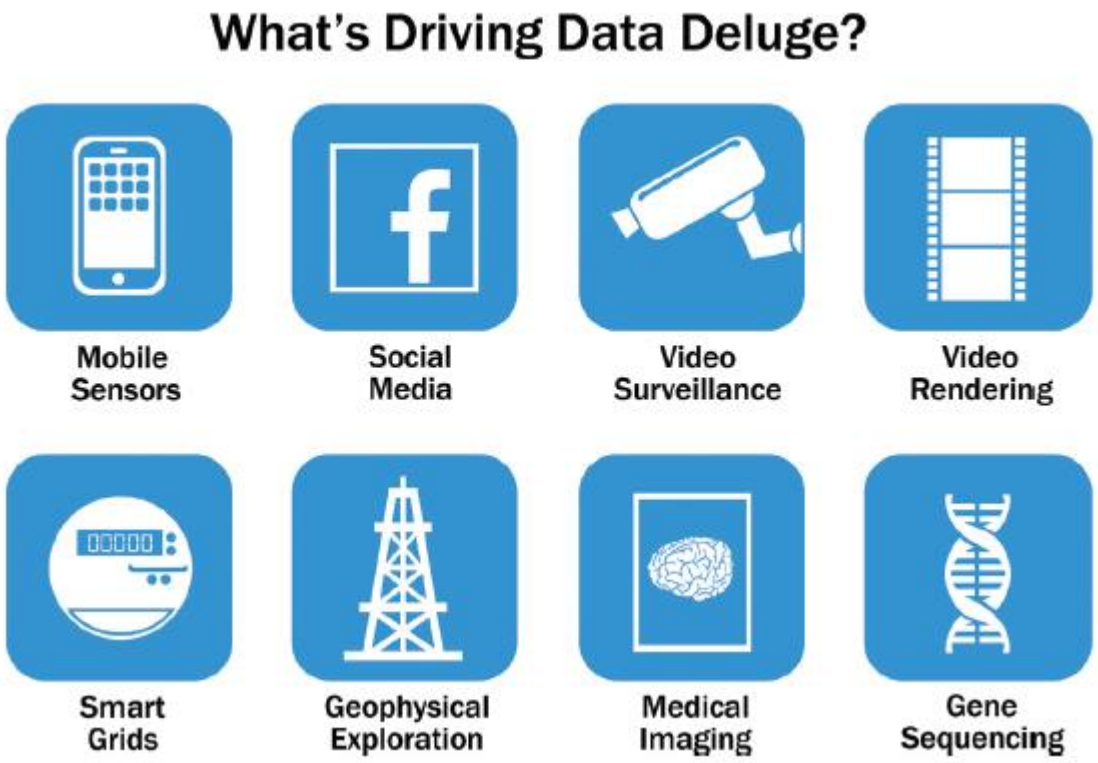
Definition

- Open question : Where are the data come from?

Where the data come from?

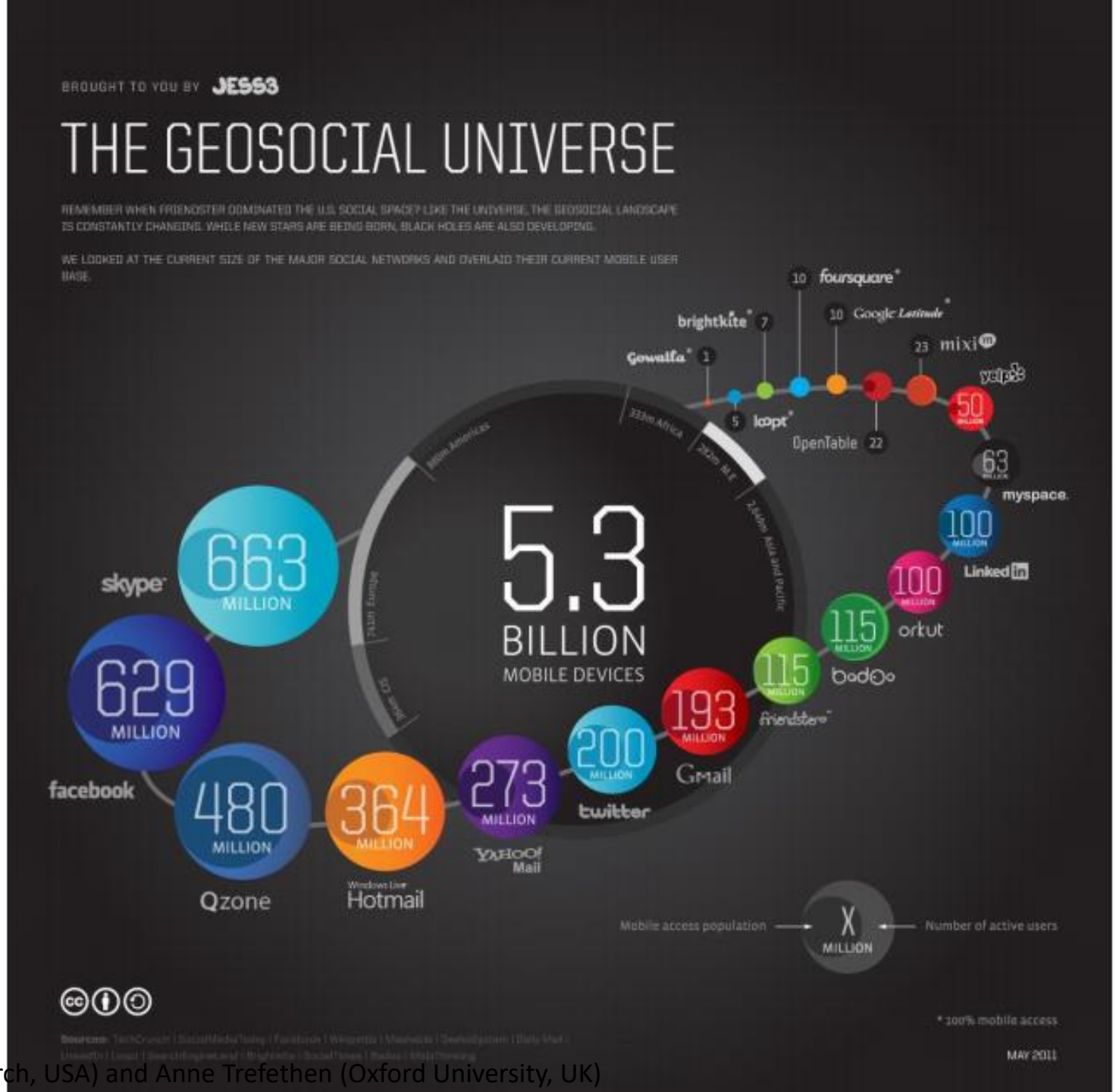
- Massive of digital information
- Credit cards
- Social networking
- Capturing traffic flow
- Measuring pollution
- Interviews, Surveys, Questionnaires
- Etc.

Where the data come from?



Where the data come from?

- 30 billion pieces of content shared on FB every month
- 30 million networked sensors deployed in the transportation, industrial, retail, and utilities sectors, increasing by more than 30%/year
- 215 million domain name registrations, 8.6% over the previous year



Geoffrey Fox (Indiana University, USA), Tony Hey (Microsoft Research, USA) and Anne Trefethen (Oxford University, UK)

Primary data vs Secondary data

- Primary data – first hand data or raw data
  - Expensive
  - Various methods (e.g., surveys, interview, focus groups, case studies, etc.).
- Secondary data – been collected by someone else (published data)
  - Easily available
  - Irrelevance, redundant, and less accuracy
  - Books, reports, censuses, government publications, etc.

Example

- John’s experiment used data from a book
- Marry conducted her experiments through questionnaire by surveying his organization

Comparison

Metrics	Primary	Secondary
Accuracy	High	Low
Control	High	Low
Relevancy	High	Low
Ownership	?	?
Accessibility	?	?
Bias	?	?
Up-to-dated	?	?

Data definition

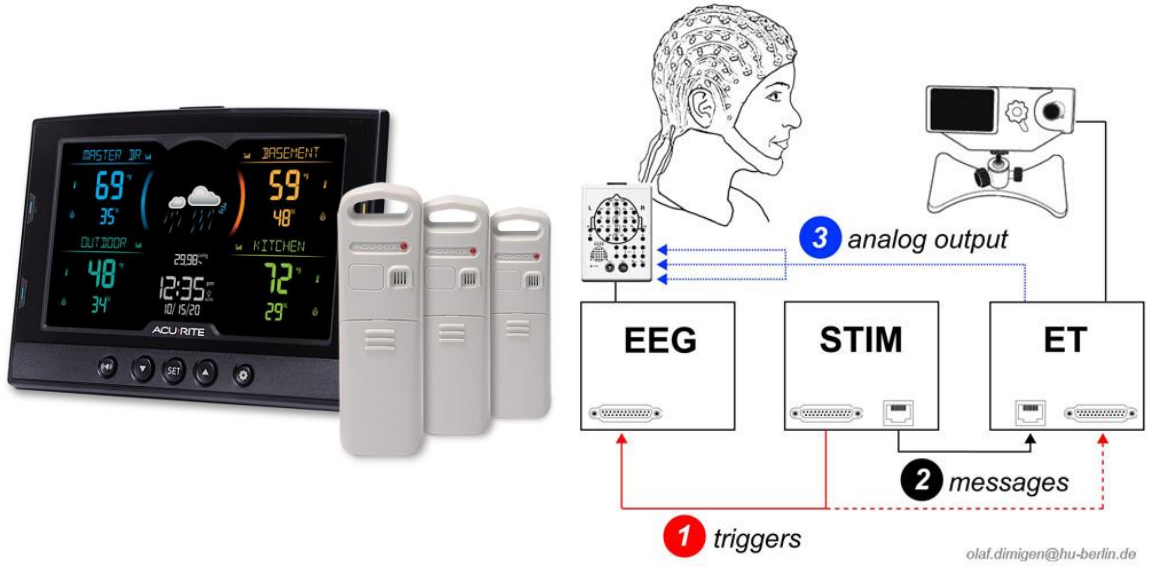
- **Data**: “stored representations of meaningful objects and events”

- Data : i**nformation, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer:** (Cambridge dictionary)

  - Structured: numbers, text, dates
  - Semi-structured: HTML, XML, JSON
  - Unstructured: images, video, documents

Data acquisition

- Hardware, Software, questionnaire, interview
  - To allow us to measure something in the real world.
  - Weather station (Temp. and humidity)
  - EYE-EEG (operate between 500-1000mhz)



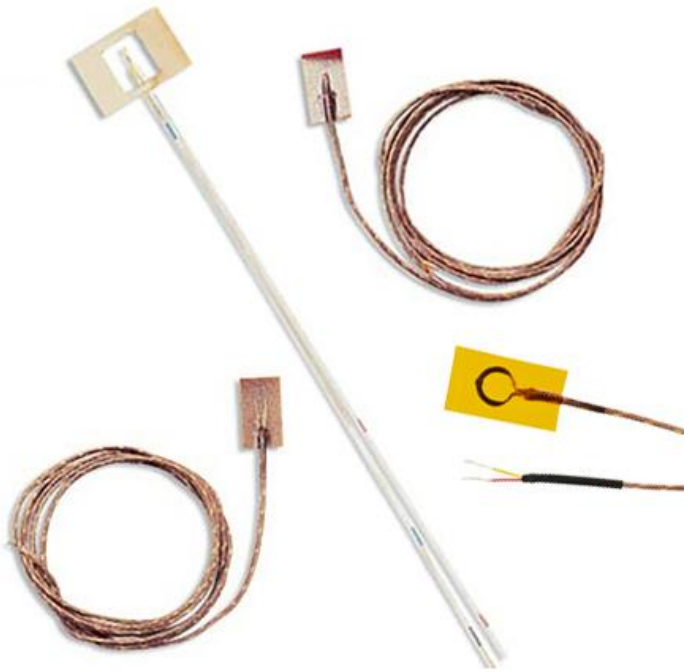
Open question: There should be way to collected data, could you explore more method?

Surveys/ Questionnaires

- Usually deployed by utility companies, building management companies, energy analysis companies or government.
- **Field Interviewer** from the Energy Information Admiration (EIA)
  - Use questionaire to collect data from selected housing uniting.
  - Data includes
    - Building characteristic
    - Energy consumption and expense
    - Household demographics
- Data from interview + data from energy suppliers
  - Estimate energy costs
  - Usage for heating and cooling

Sensor measurement

- Both survey and measurement takes very long-time recording real consumption data
- Lead to risk of inaccuracy in practice.
- Thermocouples
  - measuring the temperature for each house
  - Both inside and outside



The Household Energy End-Use Project (HEEP)

- Long term study (11 months)
- How energy is used in New Zealand household
- 400 houses throughout the country (from large and small cities)
- Two type of measurements
  - Energy end –user (11 sets of measure equipment) (55 houses)
    - Hot ware, lighting, cooking, etc.
  - Whole building energy level (345 houses)

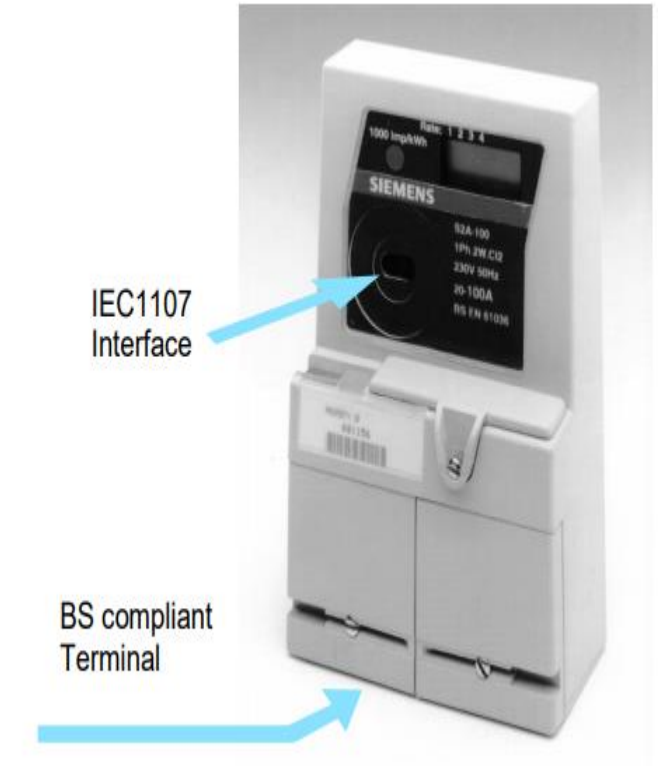


Fig.1 General view of meter

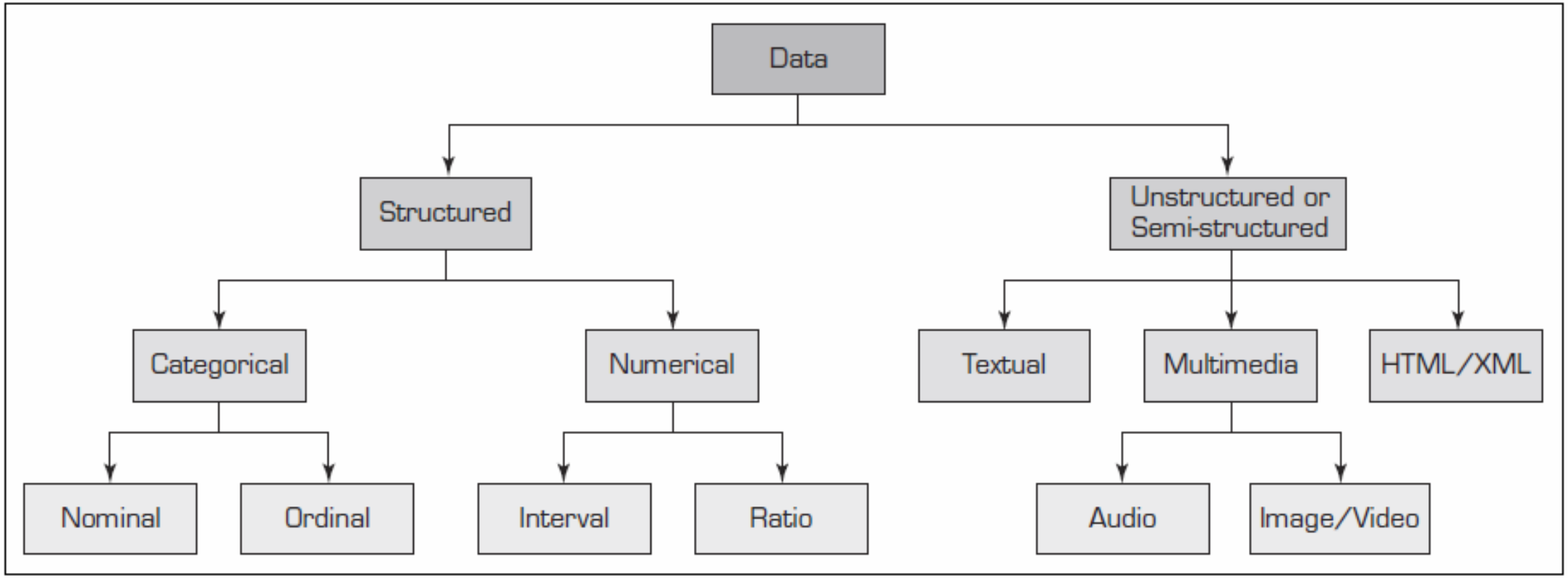
Social media

- Twetter
- Facebook
- Pantip
- Ecommerce website like JD.com, Wongnai, etc
- WebCrawler /Scraper

Simulation

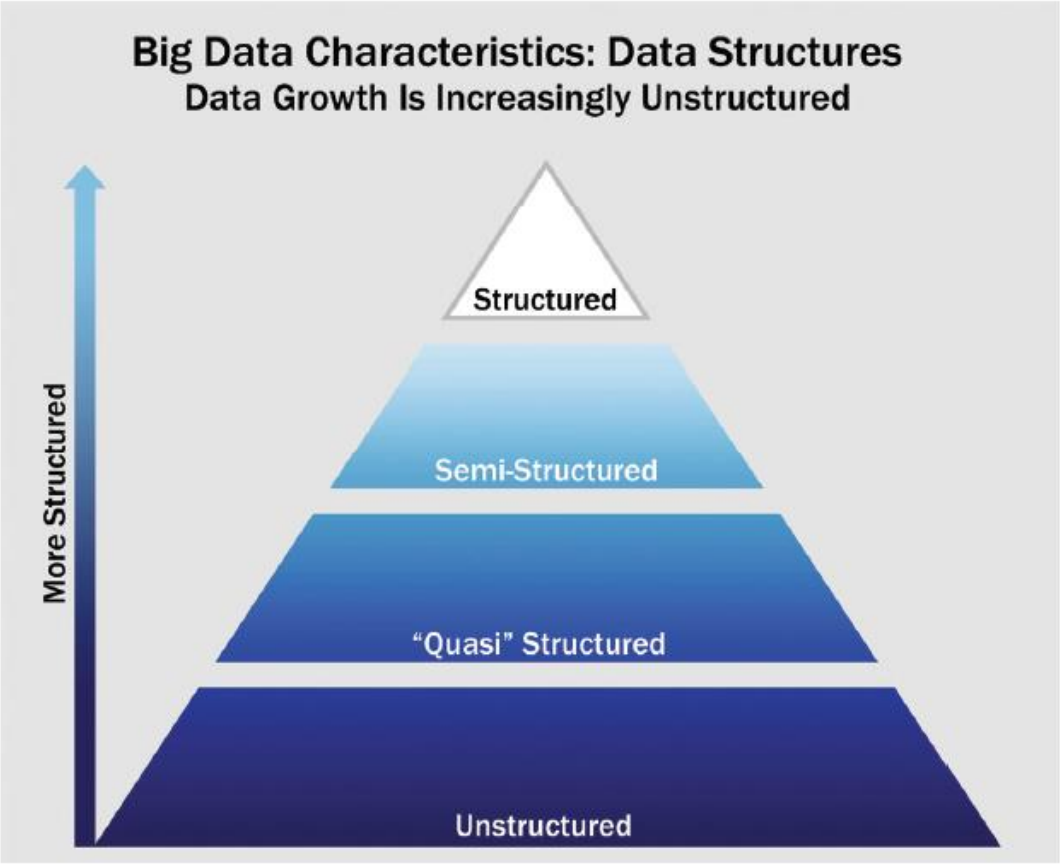
- Less expensive and less time consumption
- Recommend for complicated systems, structure or facilities.
- Building Energy Software Tools Directory (google)

Data Mining – Taxonomy of Data in DM



(Turban, Sharda, & Delen, 2014)

Big data growth increase in unstructure



Structure data

- Data containing a defined data type, format, and structure
- E.g. Transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets).

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Data Type - Qualitative

- Extract from field notes, interview transcripts
- Data can be expressed in discrete (i.e. categorical, enumerated) as follows:
  - **Nominal**- variable with no inherent order or ranking sequence (e.g. gender, nationality)
  - **Ordinal** (socio-economic status)

Open question Qualitative vs Quantitative?

Qualitative vs Quantitative data

Qualitative	Quantitative
Origin = SC	Origin = NS
Sample size = Small	Sample size = Large
Cost = Low-High	Cost = Low-High
Style = personal voice, literary	Style = formal, scientific
Type = Description	Type = numerical
Source = Interviews	Source = Instruments
2+3 more..	

Data type - Nominal data

What is your gender?

- (•) M – Male
- ( ) F – Female

What is your hair color?

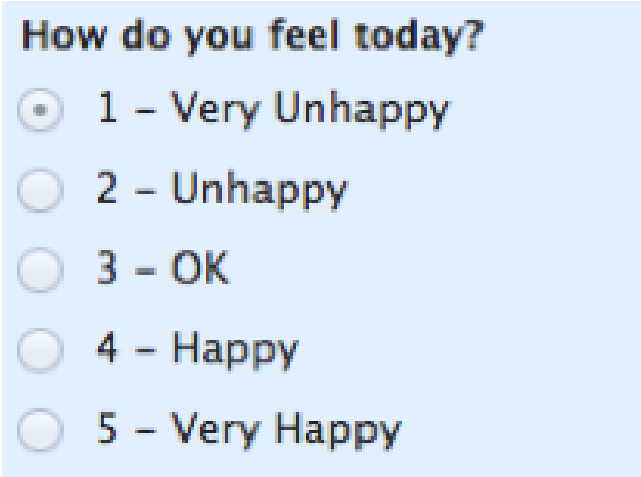
- (•) 1 – Brown
- ( ) 2 – Black
- ( ) 3 – Blonde
- ( ) 4 – Gray
- ( ) 5 – Other

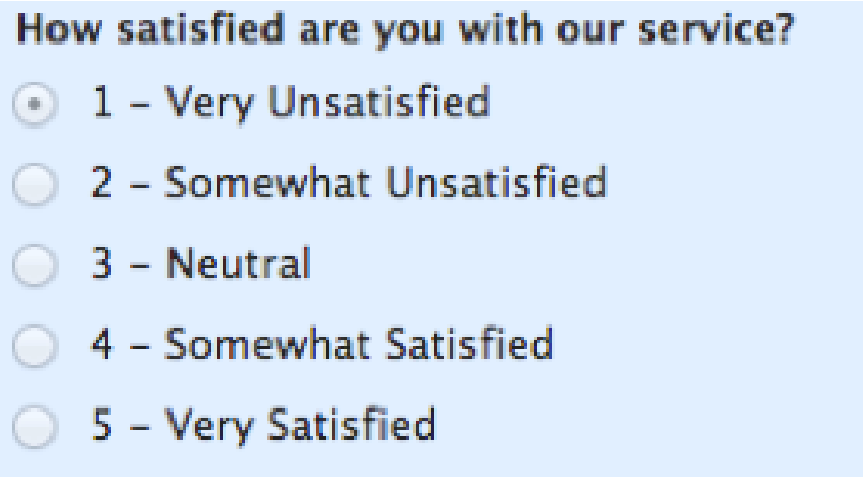
Where do you live?

- (•) A – North of the equator
- ( ) B – South of the equator
- ( ) C – Neither: In the international space station

Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/ (Accessed August, 2018)

Data type - Ordinal data





Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/ (Accessed August, 2018)

Dataset with attribute and class

DATASET WITH ATTRIBUTE AND CLASS

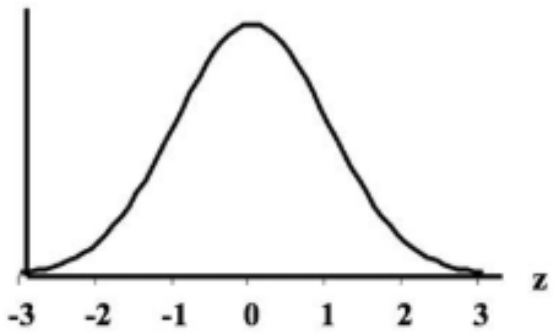
Attribute

Class/Label

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
...					

Data - Quantitative

- **Discrete** – based on counting (ordinal), and measurement
- **Continuous** (i.e. numerical)
  - Interval – temperature
  - Ratio – height, weight, length



Scaling for numerical input

- To range (-1, +1)
- To speed up the convergence

[0] 30 male 38000.00 urban democrat

[1] 36 female 42000.00 suburban republican

[2] 52 male 40000.00 rural independent

[3] 42 female 44000.00 suburban other

Scaling for numerical input (cont.)

- [0] -1.23 -1.0 -1.34 (0.0 1.0) (0.0 0.0 0.0 1.0)
- [1] -0.49 1.0 0.45 (1.0 0.0) (0.0 0.0 1.0 0.0)
- [2] 1.48 -1.0 -0.45 (-1.0 -1.0) (0.0 1.0 0.0 0.0)
- [3] 0.25 1.0 1.34 (1.0 0.0) (1.0 0.0 0.0 0.0)

Flat files

ff_tab_delimited - Notepad

File Edit Format View Help

EMP_ID	FIRST_NAME	LAST_NAME	PHONE	HIRE_DATE	JOB_ID	SALARY	COMMISSION
100	Nanette	Cambrault	6505052876	9-Dec-06	SA_REP	3800	0.35
101	Oliver	Tuvault	6505053876	23-Nov-07	SA_REP	3600	0.35
102	Janette	King	6505054876	30-Jan-04	ST_MAN	2900	0.35
103	Patrick	Sully	6505011876	4-Mar-04	ST_MAN	2500	0.3
104	Allan	McEwen	6505012876	1-Aug-04	PU_CLERK	4000	0.3
105	Lindsey	Smith	6505013876	10-Mar-05	PU_CLERK	3900	0.25
106	Louise	Doran	6505014876	15-Dec-05	HR_REP	3200	0.25
107	Sarath	Sewall	6505079811	3-Nov-06	HR_REP	2800	0.15
108	Clara	Vishney	6505079822	11-Nov-05	AC_MGR	3100	0.1
109	Danielle	Greene	6505079833	19-Mar-07	AC_MGR	3000	0.1

- Is a way of describing a simple text file, containing no structure whatsoever — data is simply dumped in a file.
- Consisting of a single Table

Advantages:

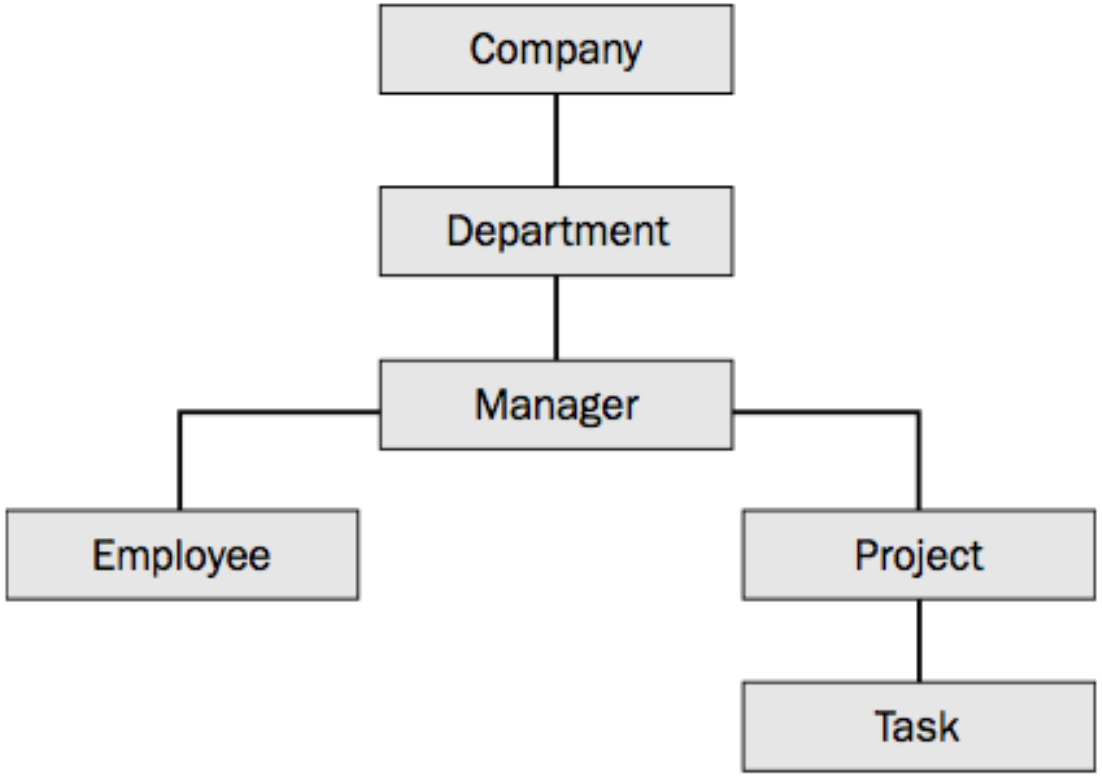
Simple to create, easy to use, inexpensive

Disadvantages:

Increased data redundancy and inconsistency

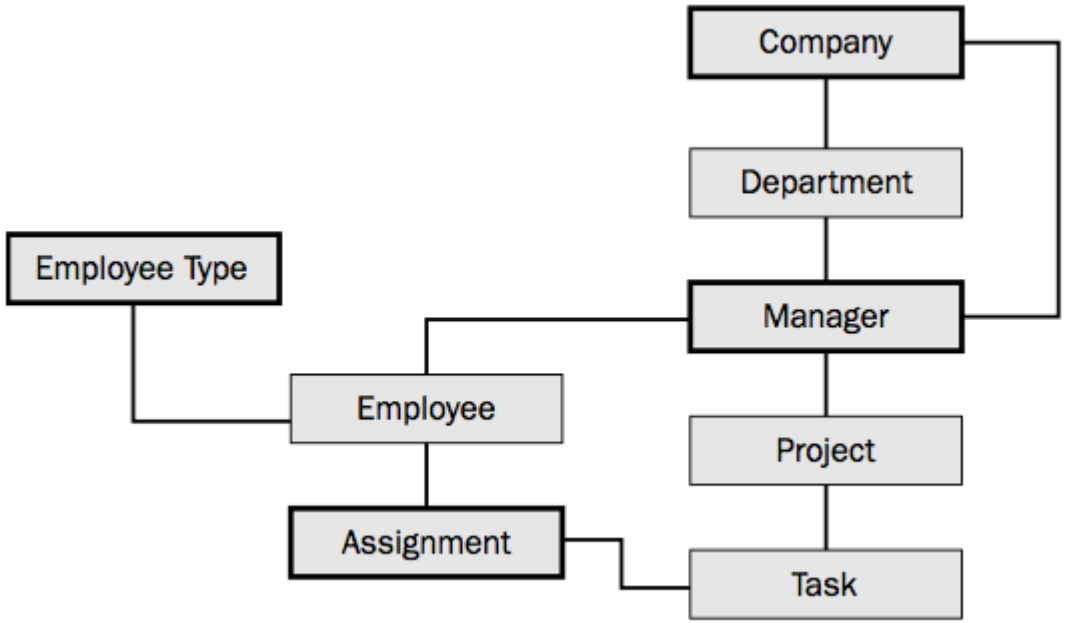
The Hierarchical Model

- The earliest databases
- Records arranged in a hierarchy much like an organization chart
- is an inverted tree-like structure. The tables of this model take on a child-parent relationship. Each *child table* has a single *parent table*, and each parent table can have multiple child tables. Child tables are completely dependent on parent tables; therefore, a child table can exist only if its parent table does.



The Network Model

- The network database model evolved at around the same time as the hierarchical database model
- The network model provided greater flexibility, but—as is often the case with computer systems—with a loss of simplicity.
- The network model allows child tables to have more than one parent, thus creating a networked-like table structure. Multiple parent tables for each child allows for *many-to-many* relationships



The Relational Model

Project

PROJECT_ID	DEPARTMENT_ID	PROJECT	COMPLETION	BUDGET
1	1	Software sales data mart	4-Apr-05	35,000
2	1	Software development costing application	24-Apr-05	50,000
3	2	Easy Street construction project	15-Dec-08	25,000,000
4	1	Company data warehouse	31-Dec-06	250,000

Task

TASK_ID	PROJECT_ID	TASK
1	1	Acquire data from outside vendors
2	1	Build transformation code
3	1	Test all ETL process
4	2	Assess vendor costing applications
5	3	Hire an architect
6	3	Hire an engineer
7	3	Buy lots of bricks
8	3	Buy lots of concrete
9	3	Find someone to do this because we don't know how

Figure 1-7: The relational database model — a picture of the data.

The Relational Model

Customer Table

Customer ID	Company Name	Contact First Name	Contact Last Name	Job Title	City	State
6	Company F	Francisco	Pérez-Olaeta	Purchasing Manager	Milwaukee	WI
26	Company Z	Run	Liu	Accounting Assistant	Miami	FL

Order Table

Order ID	Customer ID	Employee ID	Order Date	Shipped Date	Shipping Fee
51	26	9	4/5/2006	4/5/2006	$60.00
56	6	2	4/3/2006	4/3/2006	$ 0.00
79	6	2	6/23/2006	6/23/2006	$ 0.00

Employee Table

Employee ID	First Name	Last Name	Title
2	Andrew	Cencini	Vice President, Sales
5	Steven	Thrope	Sales Manager
9	Anne	Hellung-Larsen	Sales Representative

Figure 1-8 Relational table contents for Northwind

The database approach

- **Data Model** – Graphical system used to capture the nature and relationship among data

- **Entities** – A person, a place, a object, in the user environment
- **Relationships** – relationship between entities (1:1, 1:M, M:M)

Database System

Advantage

- Minimized data inconsistency
- Reduced data redundancy
- Sharing data
- Accurate and reliable data
- Use the same standard
- Security
- Greater independence of the data and programs

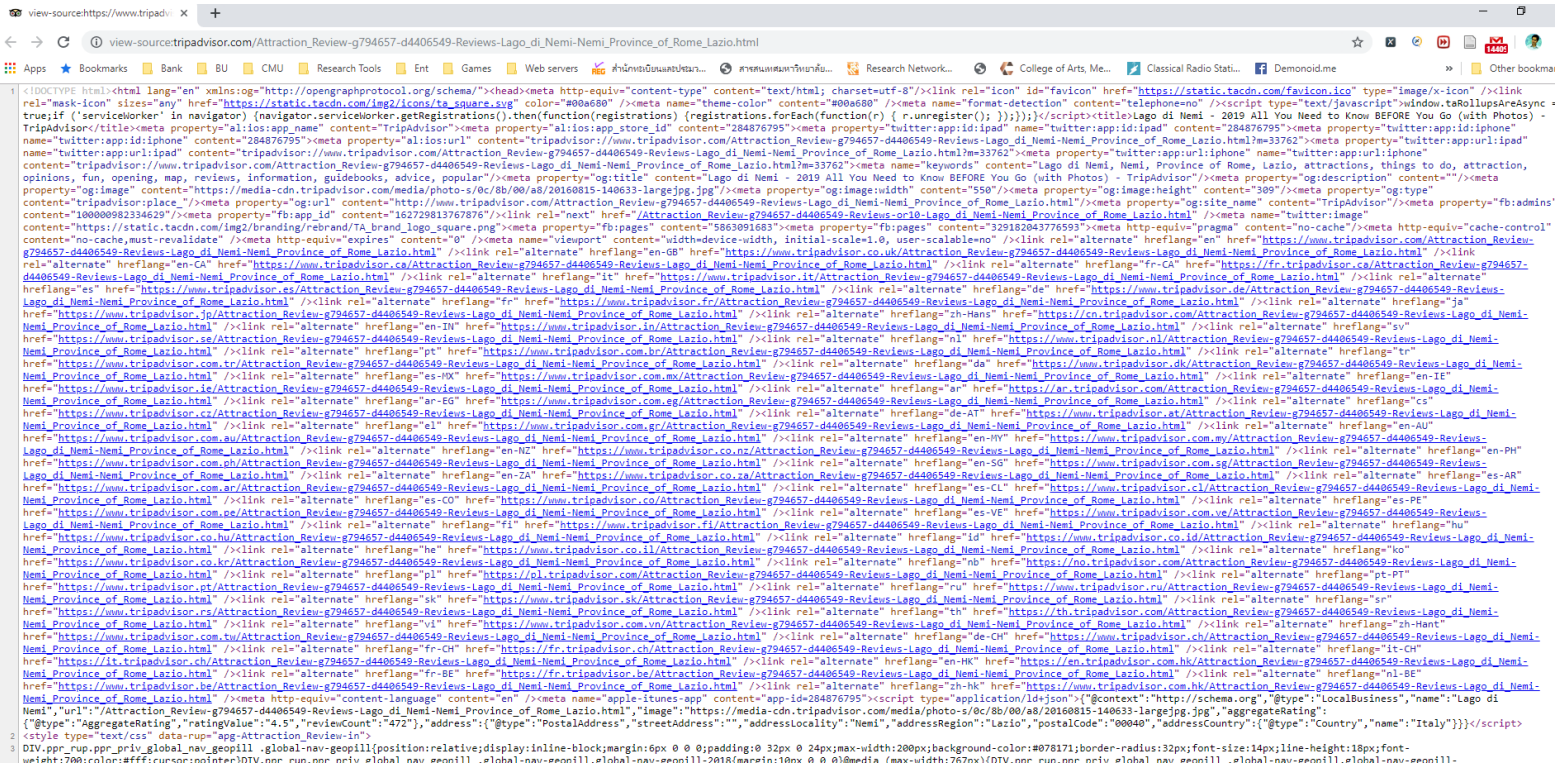
Disadvantage

- Have higher costs of hardware, software, and others
- The application and programs are complex for the users.
- High risk and high impact of system failure

Semi-structured data - HTML

- **Textual data files with a discernible pattern that enables parsing**
- **H**yper**T**ext **M**arkup **L**anguage HTML
- For display data

```html
<html>
    <head>
        Basic HTML page
    </head>
    <body>
        The body
    </body>
</html>
```



Semi-structured data -XML

- e**X**tensible **M**arkup Lanague (XML)
- For storing-carrying data
- Human and machine readable
- Parsing xml take large amount of memory

```xml
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Semi-structured data - JSON

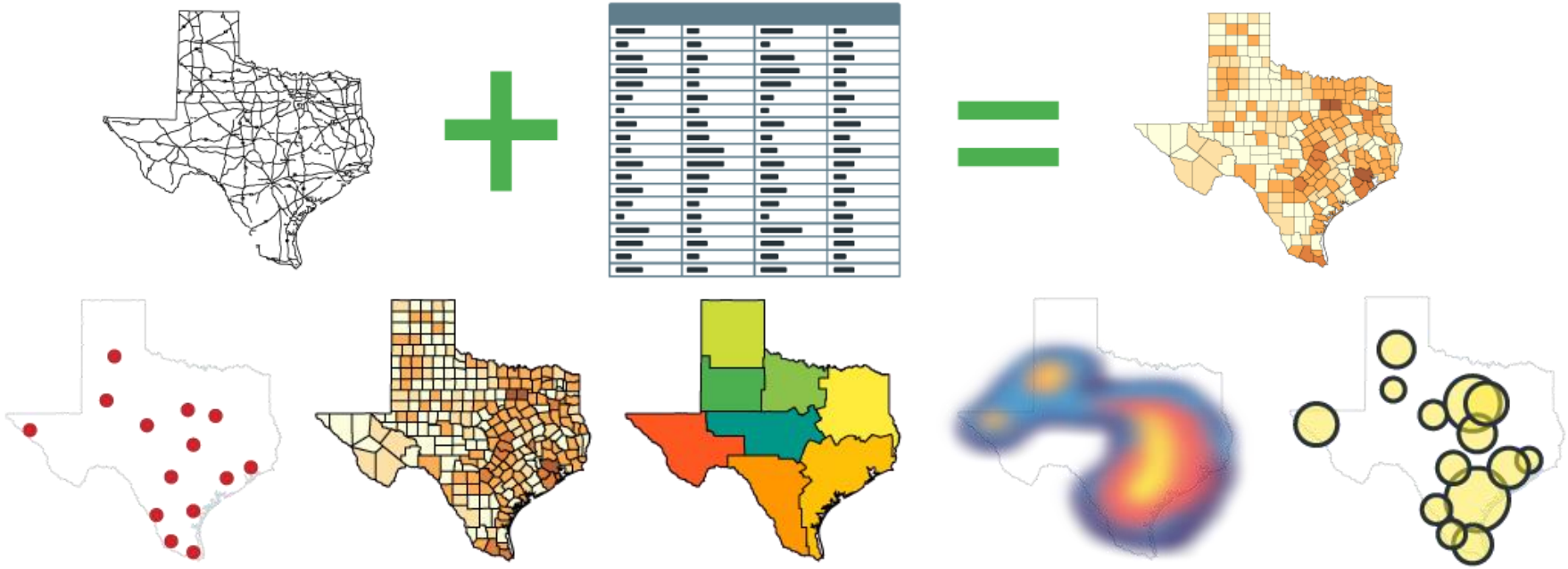
- **J**ava**S**cript **O**bject **N**otation (JSON)
- Successor of XML
- Lightweight

```
{                                                   Object Starts
    "Title": "The Cuckoo's Calling"
    "Author": "Robert Galbraith",
    "Genre": "classic crime novel",
    "Detail": {                                     Object Starts
        "Publisher": "Little Brown"                 Value string
        "Publication_Year": 2013,                   Value number
        "ISBN-13": 9781408704004,
        "Language": "English",
        "Pages": 494
    }                                               Object ends
    "Price": [                                      Array starts
        {                                           Object Starts
            "type": "Hardcover",
            "price": 16.65,
        }                                           Object ends
        {                                           Object Starts
            "type": "Kindle Edition",
            "price": 7.03,
        }                                           Object ends
    ]                                               Array ends
}                                                   Object ends
```

Image Source: w3resource

Semi-structured data - GIS

- **G**eographic **I**nformation **S**ystem (GIS)



Source: https://mangomap.com/what-is-gis

Deciding on data format

Type of Data	Common format
Tabular data, small data	Delimited flat file
Tabular data, large amount with lots of searching/querying	Relational database
Plain text, small amount	Flat file
Plain text, large amount	Non-relational database
Transmitting data between components	JSON
Transmitting document	XML

Quasi-structure data

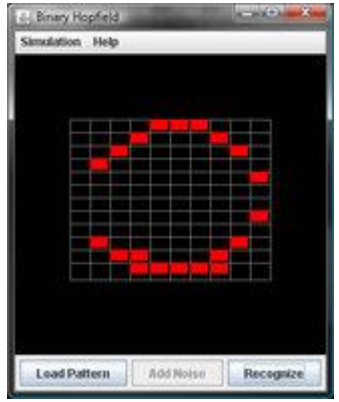
- Textual data with erratic format
- Can processed with effort and software tools

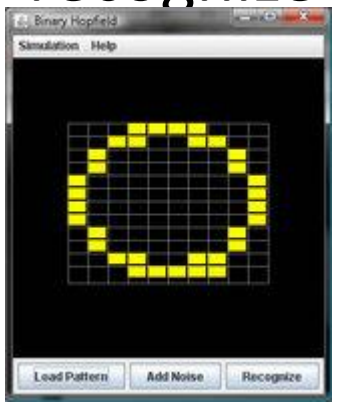
- **E.g. Clickstream data**
  - Clickstream data are a detailed log of how participants navigate through the Web site during a task. The log typically includes the pages visited, time spent on each page, how they arrived on the page, and where they went next.
  - https://www.sciencedirect.com/topics/computer-science/clickstream-data

Unstructured data- Image data

- Data that has no inherent structure, which may include text documents, PDFs, images, and video.
- http://cmtourism.org/ds/CSC5542/p5

- Binary Hopfield Network to recognize image





Unstructured data - Textual data

- Fact- 80% of data from internet in textual data
- Speech, text databases, etc.
- **Linguistic approach** – syntax, morphology, semantic, analysis stylistics, etc.
- **NLP approach** – Taggers, parses, spell checkering, word lists.

WebCrawler / scraper

- https://www.tripadvisor.com/
- Beautiful soup
- Scrapy

WebCrawler / scraper

- Requirements
  - Name of the products
  - URL of the product
  - Item code
  - Nutrition detail per 100g
    - Energy in kilocalories
    - Energy in kilojoules
    - Fat
    - Saturates
    - Fibre
    - Salt

WebCrawler / scraper

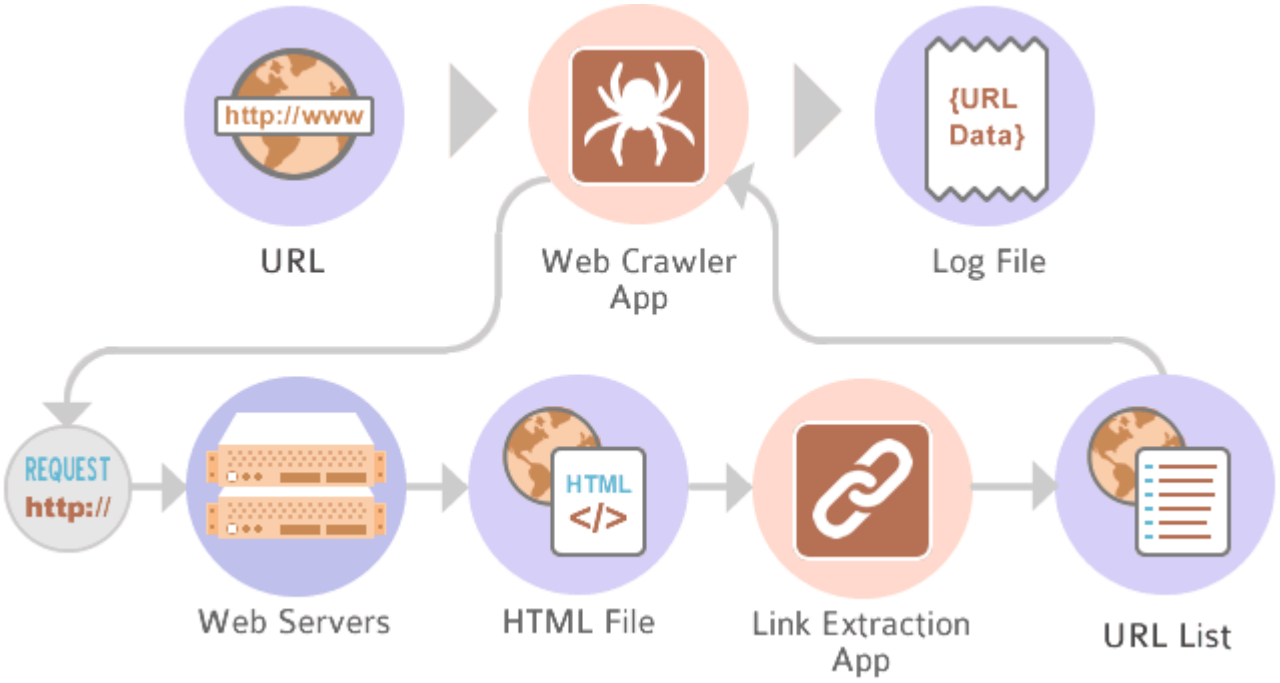


Image source: https://www.geekboots.com/story/what-is-web-spider-and-how-does-it-works

Step 1 KDD (Or most of the DM/ML Process)

- **Initial Selection (data understanding)**
  - Use your domain knowledge
  - Remove feature(s) that is not relevant for your model
  - http://archive.ics.uci.edu/ml/datasets/Adult

Your first real-world data set (adult)

- https://www.cs.waikato.ac.nz/ml/weka/

References

- Course text book (Data science and big data analytics)