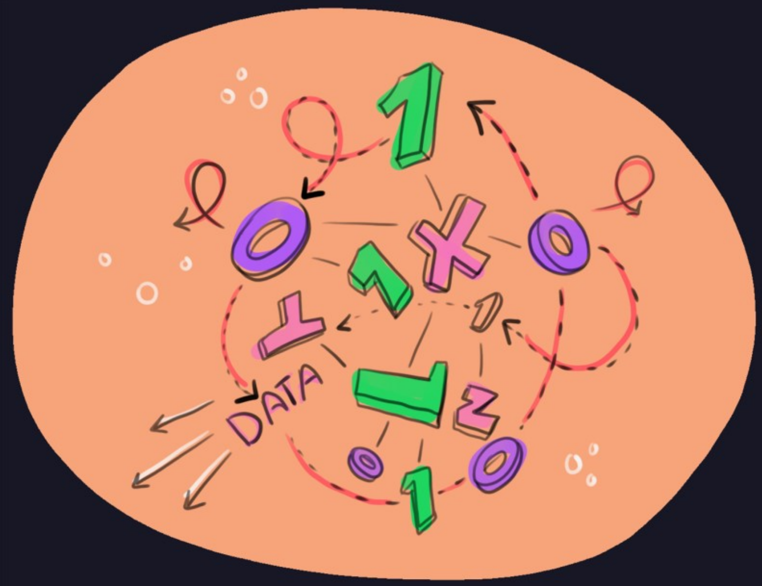
Data for Media - Spring 2024

Lesson 01:
Understanding Data

Data Types & Common Formats



Professor:
Amr Eleraqi

aeleraqi@aucegypt.edu

What is Data?

- Basic values or facts.
- Data is a collection of distinct pieces of information, which are recorded and structured in a manner that makes it easy for analysis.
- Data is also generated by mobile devices, sensors, the internet, and satellites (such as GPS data) - and many other technologies.
- Data is generated when something is voted on (such as elections results data), when something is registered (such as a birth records data), when something is purchased (such as sales records for an online store), etc.

Data vs. Information

Data can be defined as plain facts or statistics, presented in the form of numeric values or text, devoid of any specific framework or interpretation that gives it meaning. Data can be even seem useless until it is analyzed, organized, and interpreted.

Input

Information refers to understanding or comprehension derived from processing, interpreting, organizing, or analyzing data within a relevant context. It involves presenting data in a manner that offers value, relevance, significance, or utility by adding meaning to the initial data points.

Output

Data

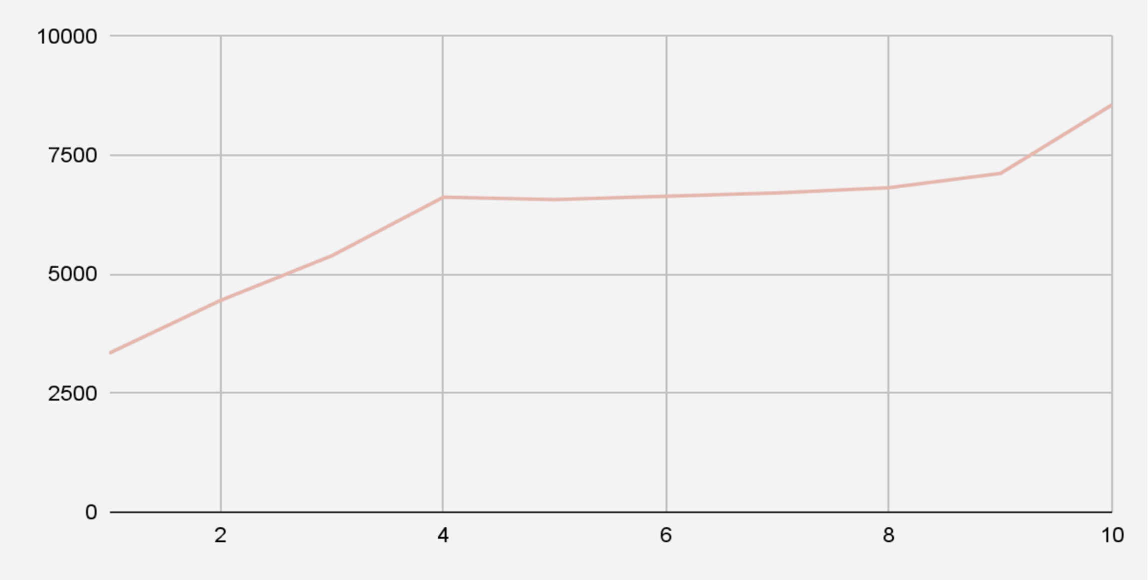
Sales Value Over the Last Ten Days

Day	1	2	3	4	5	6	7	8	9	10
Sales	3340	4450	5390	6620	6570	6640	6710	6820	7120	8560

Information

Remarkable Uptrend in Sales Value During the Last 10 Days

Observing a Monumental 156% Expansion Between first and last Day



From Data to Information

Data

10,12,13,20,09,20,09

Summarizing

Averaging

Sorting

Adding Value

Adding Context

Graphing

Information

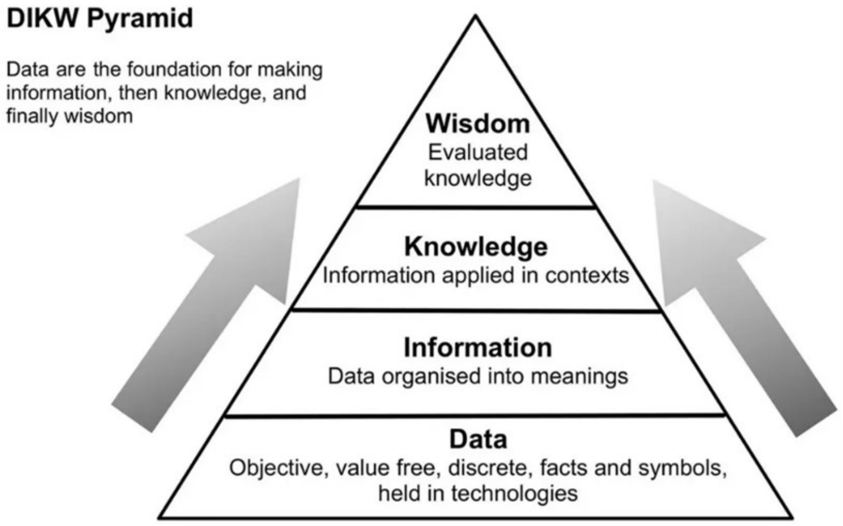
The Key Differences Between Data vs Information:

- Data is a collection of facts, while information puts those facts into context.
- Data, on its own, is meaningless. When it’s analyzed and interpreted, it becomes meaningful information.
- Data does not depend on information; however, information depends on data.
- Data isn’t sufficient for decision-making, but you can make decisions based on information.

Data is a mess and needs to be processed to make it meaningful. The DIKW hierarchy describes how data evolves into information, knowledge, and wisdom respectively.

From Data to Wisdom, The process of turning data into wisdom involves several steps: collecting, processing, analyzing, and interpreting data in order to gain insights and make informed decisions.

- Wisdom is the ability to use knowledge.
- Knowledge is the understanding and interpretation of information.
- Information is processed data.
- Data refers to raw, unprocessed facts.



DIKW Pyramid

Data are the foundation for making information, then knowledge, and finally wisdom

Wisdom
Evaluated knowledge
Knowledge
Information applied in contexts
Information
Data organised into meanings
Data
Objective, value free, discrete, facts and symbols, held in technologies

Q1. Which is Data, Information, Knowledge, Wisdom decisions?

1

Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
20	22	20	24	19	5	0

2

It seems like the temperatures are in a descending trend throughout the week, with the highest temperatures on Sunday and Tuesday, and notably cooler temperatures towards the end of the week.

3

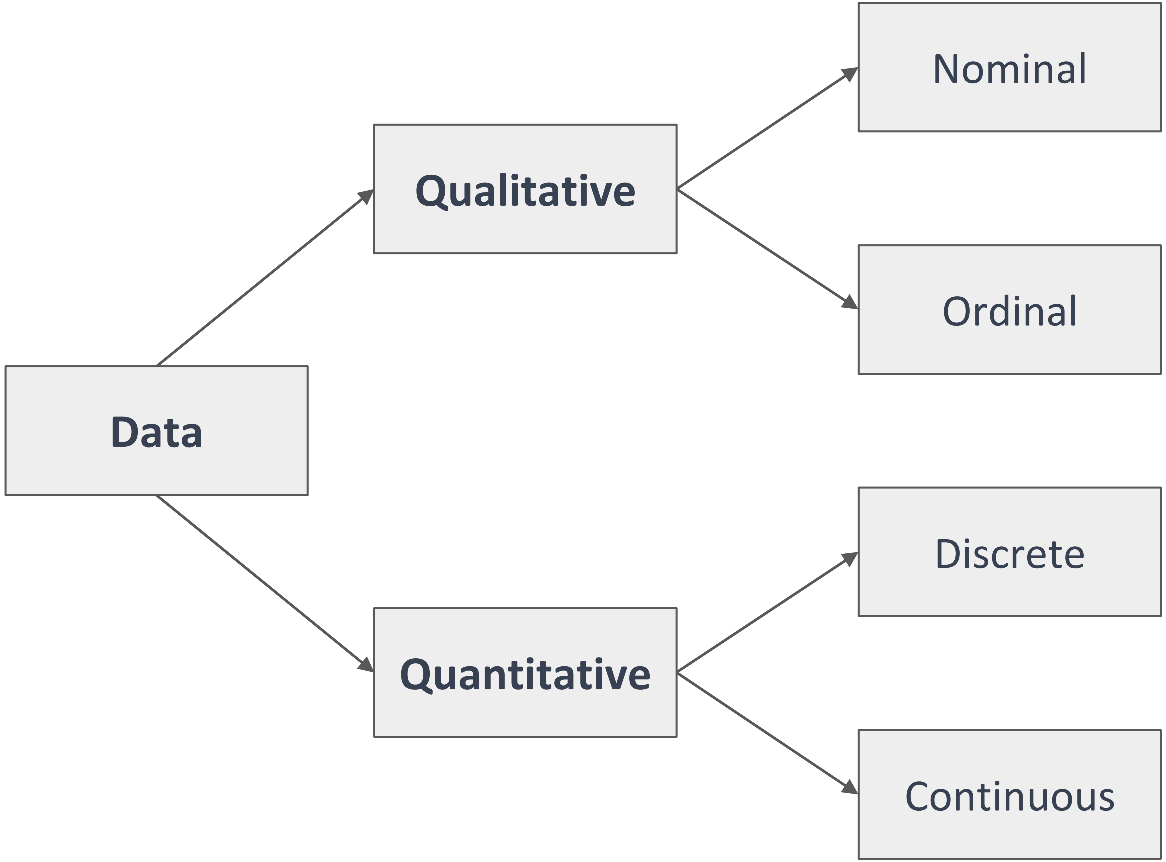
The weather will not be entirely stable throughout the week. Although mild temperatures are predicted for most of the week, there are fluctuations indicated each day.

4

Saturday, Sunday, and Monday: Mild temperatures, so you may want to wear light and breathable clothing.

- Tuesday: Slightly warmer, so similar to the first three days, light and comfortable clothing should be suitable.
- Thursday: A significant drop in temperature, you may need warmer layers.

Types of Data:



Qualitative data, also known as categorical data, is non-numerical and describes qualities or characteristics. It includes data that can be observed but not measured, such as colors, textures, sounds, smells, tastes, appearances, attitudes, opinions, and beliefs.

Types of Qualitative Data:

- Nominal data: This type of qualitative data consists of labels that cannot be ranked or ordered. Examples include gender, race, religion and occupation.
- Ordinal data: This type of qualitative data has a natural order or ranking. Examples include Likert scales (strongly agree, agree, neutral, disagree, strongly disagree), educational level (elementary school, high school, college, graduate school), and star ratings (one star, two stars, three stars, etc.).

Quantitative data, is numerical data that can be measured and analyzed statistically. It includes data that can be counted, compared, and manipulated mathematically, such as height, weight, temperature, time, distance, speed, and quantities.

Types of Quantitative Data:

- Discrete data: This type of quantitative data consists of whole numbers that do not have decimal points or fractions. Examples include the number of students in a class, the number of cars in a parking lot, and the number of books on a shelf.

- Continuous data: This type of quantitative data consists of values that fall between whole numbers and can have decimal points or fractions. Examples include weight, height, temperature, and time intervals.

Not All Numerical Data is Quantitative

Examples:

- Phone Numbers
- Credit Card Numbers
- Security Codes
- National ID
- Student ID
- Player Kit number

The sum or total of these numbers doesn't hold any meaningful significance in most cases. For instance, summing up phone numbers, or security codes would not provide any useful information. These numbers serve as identifiers or codes rather than quantities that can be added or averaged.

Ex. Patient Dataset in a Hospital:

Qualitative Nominal: Blood Type

Blood type is a nominal categorical variable. It represents categories (A, B, AB, O) with no inherent order or ranking.

Qualitative Ordinal: Patient Satisfaction Rating

Patient satisfaction rating is an ordinal categorical variable. It represents ordered categories (e.g., Poor, Fair, Good, Excellent) where there is a meaningful order or rankin.

let's consider a hypothetical patient dataset in a hospital setting:

Quantitative Discrete: Number of Medications Prescribed

The count of medications prescribed is a discrete quantitative variable. It represents whole, distinct numerical values.

Quantitative Continuous: Body Temperature

Body temperature is a continuous quantitative variable. It can take on any value within a range (e.g., 98.6°F to 99.5°F) and can be measured with a high level of precision.

Q2. Identify the following variable as either qualitative or quantitative:

- The number of people on a jury.
- The color of your house.
- A person's height in feet.
- The speed of a car in miles per hour.
- Outcome of tossing a coin.
- The amount saved monthly towards retirement.
- Type of vehicle owned.
- Preferred genre of music.

Q3. Classify the data as either nominal or ordinal:

- Colors of flowers in a garden: Red, Blue, Yellow, Orange, White .
- Types of fruits in a basket: Apples, Bananas, Oranges, Grapes.
- Genres of movies available in a streaming service: Action, Comedy, Drama.
- Education levels attained by employees in a company: High School Diploma, Bachelor's Degree, Master's Degree, PhD.
- Customer satisfaction survey scores rated on a Likert Scale: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree.
- Level of difficulty chosen for hiking trails in a state park: Easy, Moderate, Challenging, Strenuous.

Q4. Classify the data as either discrete or continuous:

- The average speed of cars passing a busy intersection between 4:30 P.M. and 6:30 P.M. on a Friday is 32.3 mi/h.
- The temperature in Manhattan at 1 p.m. on New Year's Day was 34.1 °F.
- Number of books borrowed from a library in a month: 1200 books.
- Weight of apples harvested from a single tree in a season: 25 kg.
- Time taken by participants to complete a marathon: 3 hours, 4 minutes, and 21 seconds.
- Total number of votes cast for candidate X in an election: 12,568 ballots.
- Quantity of mobile phones sold worldwide in Q1 2022: 302 million devices.

Q5. In a chemistry lab, a combination of nitrogen dioxide (N_2O) and hydrogen peroxide (H_2O_2) was blended, and the outcomes were documented.

Classify the subsequent variables as either qualitative (nominal or ordinal) or quantitative (discrete or continuous):

- The number of observed gas bubbles produced during the experiment ________
- Initial weights of chemicals used (Sodium Nitrate & Calcium Nitrate) ________
- Color of chemicals inside various containers used in the experiment ________

- The smell resulting during the reaction process (noticeable - not noticeable)

  ____
- Measured temperature fluctuations inside the lab during the experiment

  ____
- Resultant liquid classification (acidic or alkaline) _________
- Detected sound level during the execution of the experiment, in scale (very high – high – low – very low) _________
- Time elapsed for completion of the whole procedure was 30 minutes and 15 seconds _________
- The number of attempts that were made before, until the experiment was successful_________

Q6. As a motorsports journalist, you find yourself covering a significant race segment. Throughout the competition, you keep track of several variables to include in your post-race analysis. Kindly Classify these variables based on the following options:

qualitative (nominal or ordinal) or quantitative (discrete or continuous):

- The type of fuel used by each car (e.g., gasoline, diesel, electric)_________
- The position in which each car finishes the race (e.g., first place, second place, third place)_________
- The top speed reached by each car during the race_________
- The level of experience each driver has had to compete in races (such as beginner, intermediate, advanced)_________
- The amount of time it takes each car to complete one lap around the track_________
- The customer satisfaction surveys and owner reviews for each car_________

- The number of penalty flags received by each team during the race_________
- The altitude of the racetrack_________
- The pitch and roll angles of each car during cornering_________
- The fan reactions and social media discussions about each race outcome_________

Disaggregated data, refers to data that has been broken down into its smallest meaningful components. Unlike aggregated data, disaggregated data retains all the original detail and granularity, allowing analysts to examine differences and patterns within specific subgroups or categories.

Ex. Patient Information and Medical History Dataset:

Patient name	Blood Type	Gender	Floor Number	Diseases
Xxxxxx xxxxxx	A	Male	1	Influenza
Xxxxxx xxxxxx	B	Male	3	Type 1 diabetes
Xxxxxx xxxxxx	O	Female	3	Breast cancer
Xxxxxx xxxxxx	A	Female	1	Breast cancer
Xxxxxx xxxxxx	B	Male	2	Influenza

Aggregated data is a form of data that combines multiple data points into a single value, often representing a summary statistic or measure of central tendency. When data is aggregated, details about individual observations are lost, and only the general pattern or trend remains. Common ways to aggregate data include calculating means, medians, sums, proportions, and percentages.

Ex. Patient Information and Medical History Dataset:

Option 1:
Grouping by
Blood Type

Blood Type	Number of Patients
A	2
B	2
O	1

Option 2:
Grouping by
Floor Number

Floor Number	Number of Patients
1	2
2	1
3	2

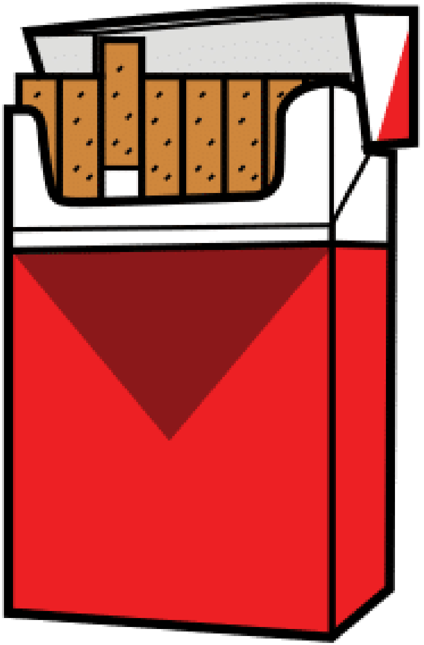
Option 3:
Grouping by
Gender

Gender	Number of Patients
M	3
F	2

When can the Aggregated data be misleading?

Aggregated data can be misleading when it hides variations within the data set. In the case of the statement ***"Smoking rate has increased by 10% in the Middle East."*** it can be misleading for several reasons:

The increase in smoking rates might be driven by a significant increase in one or a few countries, while others may have experienced a decline or remained stable. Aggregated data may not reflect the varied situations within the diverse population of the Middle East.



Machine-gathered data is generated automatically by machines, sensors, devices, or software applications without direct involvement from humans. Some advantages of using machine-gathered data include its ability to capture real-time events, reduce human bias, and scale cost-effectively. However, machine-generated data can sometimes lack context or interpretation, requiring manual intervention or processing to extract meaningful insights.

Ex. Traffic Radar Camera:

- Operates independently without direct human involvement.
- Provides consistent and precise speed measurements.
- Monitors traffic continuously, 24/7, capturing violations consistently.
- Generates automated records and photographic evidence for documentation.

Human-gathered data, is collected directly by individuals who observe, record, or measure phenomena manually. Compared to machine-generated data, human-generated data tends to be less structured, less precise, and more prone to error due to subjective interpretations or biases. However, human-generated data can offer richer insights, greater context, and more nuanced perspectives than machine-generated data alone.

Ex. Traffic officer:

- Requires direct human involvement, introducing the potential for human factors like fatigue.
- Accuracy may be influenced by various factors.
- Presence is selective, as officers are present based on factors such as time, location, and reported incidents.
- Allows for discretion in enforcement, considering factors like road conditions and driver behavior.
- Relies on the officer's testimony and manual documentation for evidence.

Structured data is highly organized and stored in predefined fields or tables, making it easy to search, sort, and analyze. Structured data usually follows a rigid schema, meaning that each piece of data must fit into a predetermined category or field. Examples of structured data include relational databases, spreadsheets, and tabular data formats like comma-separated values (CSV) or Excel files.

Ex.

Patient name	Blood Type	Gender	Floor Number	Diseases
Xxxxxx xxxxxx	A	Male	1	Influenza
Xxxxxx xxxxxx	B	Male	3	Type 1 diabetes
Xxxxxx xxxxxx	O	Female	3	Breast cancer

Semi-structured data falls somewhere in between structured and unstructured data. Semi-structured data contains elements of both formats, with metadata tags or markup languages providing some degree of organization and hierarchy. Semi-structured data is easier to work with than unstructured data since it has some inherent structure and consistency.

Ex.

Patient 1:

- Blood Type: A
- Gender: Male
- Floor Number: 1
- Diseases: Influenza

Patient 2:

- Blood Type: B
- Gender: Male
- Floor Number: 3
- Diseases: Type 1 diabetes

Patient 3:

- Blood Type: O
- Gender: Female
- Floor Number: 3
- Diseases: Breast cancer

Unstructured data, does not follow a strict format or organizational scheme. Unstructured data includes freeform text documents, images, audio and video recordings, emails, chat messages, and other forms of multimedia content. Since unstructured data lacks a fixed structure, it is challenging to search, filter, and analyze programmatically.

Ex.

The first patient has a blood type of A, is male, is on floor number 1, and is diagnosed with influenza.

- Does not follow a model, can't be contained in rows and columns.
- Difficult to search and organize.
- Usually text, sound, pictures or videos.

Machine-readable data is designed to be processed and interpreted by computers, with minimal or no human interaction required. Machine-readable data typically adheres to standards, schemas, or protocols that define its format, structure, and semantics. Examples of machine-readable data include CSV, XLS files.

- Excel files (XLS): data is saved as a table readable by Microsoft Excel
- Comma separated values (CSV): Plain text file with each data entry separated by a comma.

These formats are typically the best suited for analysis, and you can easily work with them in a spreadsheet program - like Excel. When searching for data, if you can find Excel or CSV formats, this is a good sign that you won't have to spend a lot of time cleaning and formatting.

Non-machine-readable data, includes analog formats such as print publications, handwritten notes, physical maps, film negatives, and audio cassettes. Digital file formats that are not optimized for machine consumption, such as proprietary document formats, scanned images, or PDF files with embedded texts and graphics, can also be considered non-machine-readable.

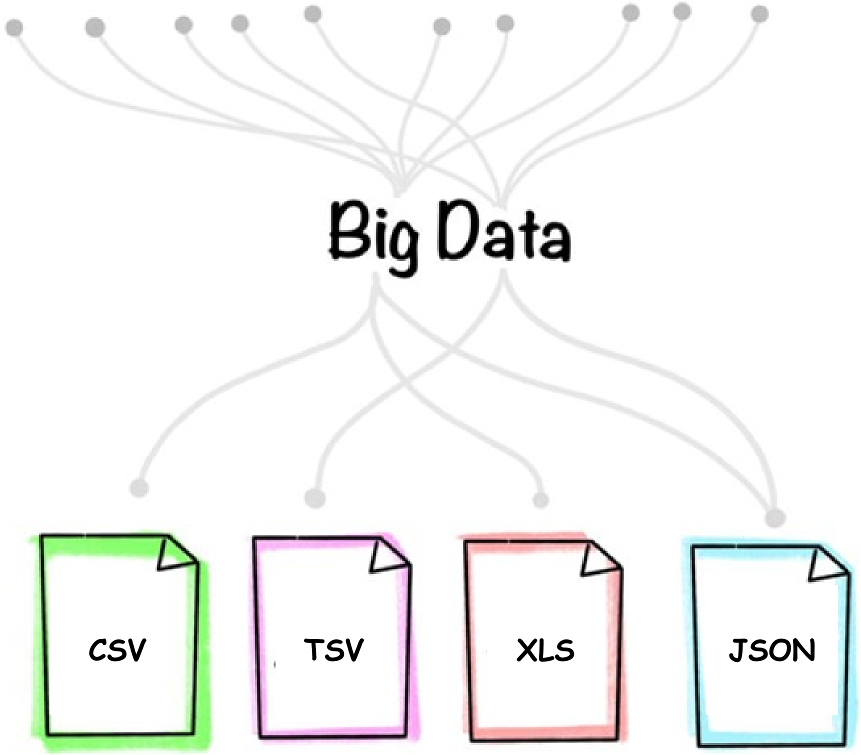
Examples of non-machine readable data formats:

- Word documents (.docx)
- PowerPoint presentations (.pptx)
- Image files (.bmp, .gif, .tiff, .raw) Video files (.avi, .wmv, .mov, .flv) Audio files.
- Scanned images of handwritten notes or drawings
- Photocopied (Scanned) documents or photographs
- Physical objects with no digital representation, such as a book or a piece of artwork.

Q7. Suppose you work as a technology reporter for a tech news outlet, and your editor requests you to visit a local artificial intelligence conference to gather details about AI innovations presented there. Consider the statements below and indicate whether they are true or false:

1	Data gathered via direct observation and documented in spreadsheets like Microsoft Excel can be regarded as both machine-gathered and machine-readable.	
2	Scanning physical brochures from exhibitors featuring product specs using Optical Character Recognition software enables such content into machine-readable form.	
3	Transcripts of interviews conducted at the conference that were previously audio recorded but later transcribed into text format are now structured.	
4	The number displayed on each company's booth is merely an identification code and does not represent any quantifiable value.	
5	The Excel file obtained from the official conference website contains participant names along with their corresponding identification numbers. is a structured and disaggregated file.	
6	After requesting information from the conference organizers, you received a confirmation stating that approximately half of the attendees hailed from North America. This information is disaggregated.	
7	Given a comma-separated values (CSV) file received from an exhibitor containing detailed information about their developed AI models and associated accuracy metrics, the separator (delimiter) in such file type is a semicolon.	
8	Portable Document Format (PDF) file containing only tables about exhibitors emailed from the media center represents a structured data type.	

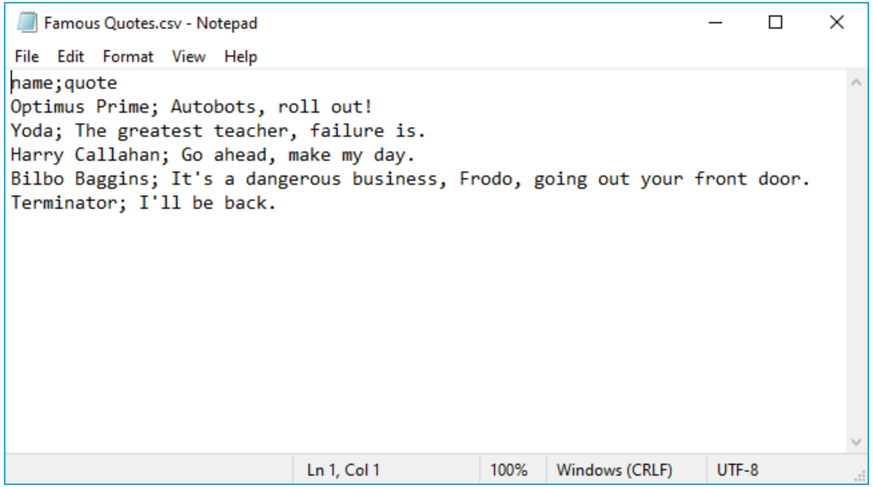
Common Data files formats



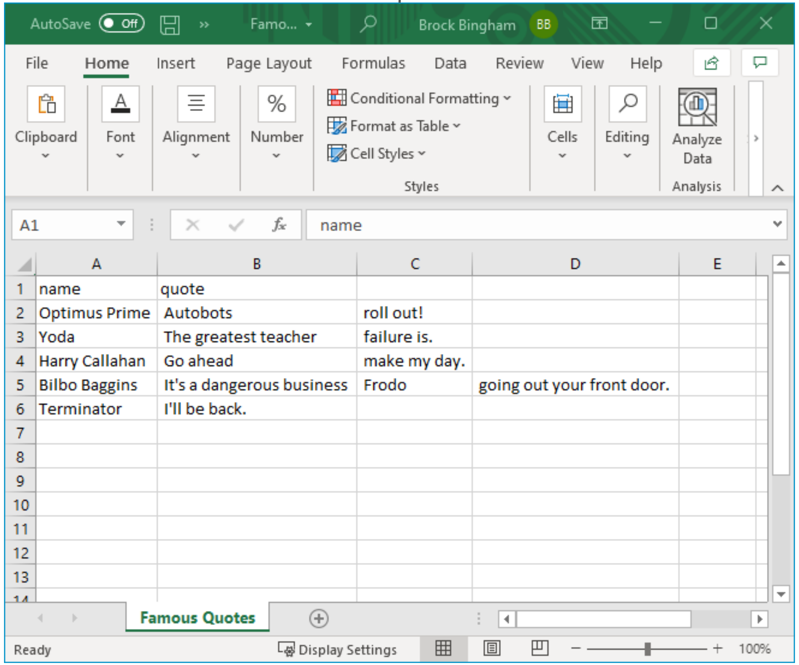
CSV, CSV stands for ‘Comma-separated values’. It is a standard format for spreadsheet data and is widely used in the public and private sectors to produce datasets. Data is represented in a plain text file, with each data row on a new line and commas separating the values on each row. As a very simple open format it is easy to consume and is widely used for publishing data. You can import CSV files into Excel.

This type of data structure (CSV) allows programs - such as spreadsheets, a variety of cloud applications, and databases - to read your data in an organized way. For example, when you view data in Google Sheets, you see data structured in tabular form - that is, in the columns and rows that make up a table.

From CSV to Excel



```
name;quote
Optimus Prime; Autobots, roll out!
Yoda; The greatest teacher, failure is.
Harry Callahan; Go ahead, make my day.
Bilbo Baggins; It's a dangerous business, Frodo, going out your front door.
Terminator; I'll be back.
```

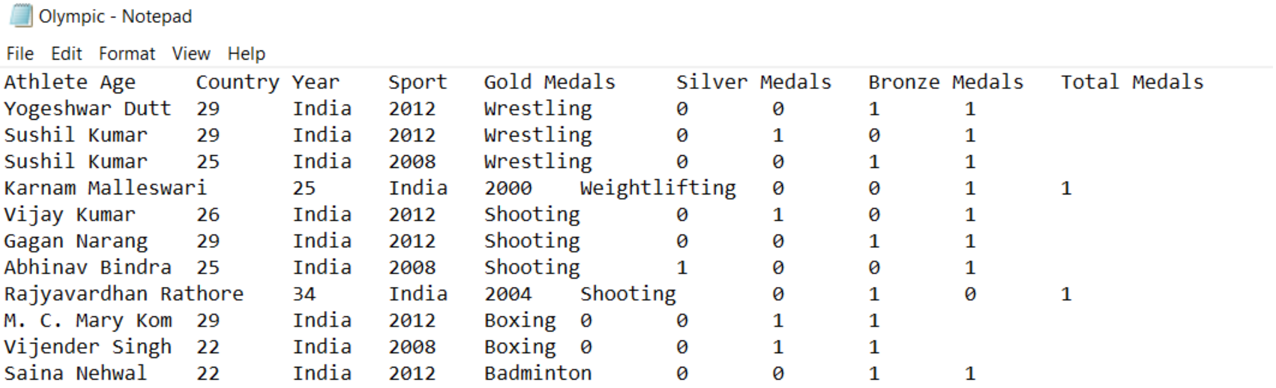


	A	B	C	D	E
1	name	quote			
2	Optimus Prime	Autobots	roll out!		
3	Yoda	The greatest teacher	failure is.		
4	Harry Callahan	Go ahead	make my day.		
5	Bilbo Baggins	It's a dangerous business	Frodo	going out your front door.	
6	Terminator	I'll be back.			
7					
8					
9					
10					
11					
12					
13					

1. With Excel open, click on the Data menu option.
2. In the Get & Transform Data ribbon section, click From Text/CSV.
3. Navigate to the CSV file, select it, and click Import.
4. click Load. you'll notice that all of the data has been imported correctly.

TSV, Tab-separated values (TSV) are a very common form of text file format for sharing tabular data. TSV files can be imported into and exported from spreadsheet software. TSV files are essentially text files and can be viewed by text editors.

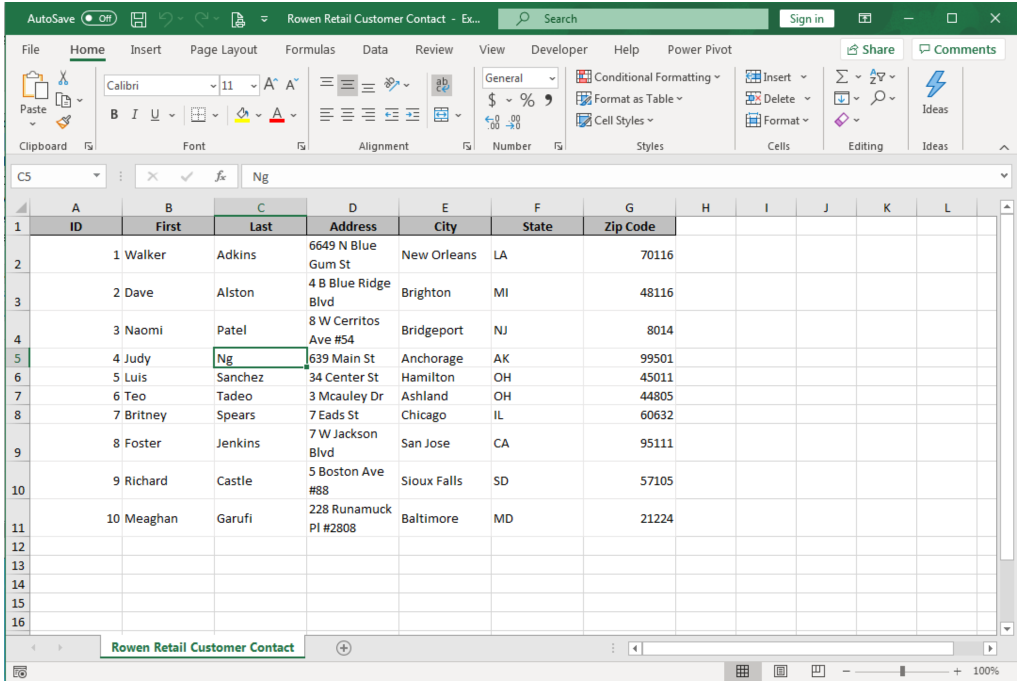
TSV field can be imported to Excel or Word based on the structure you want!



```
Athlete Age     Country Year    Sport   Gold Medals     Silver Medals   Bronze Medals   Total Medals
Yogeshwar Dutt  29      India   2012    Wrestling       0       0       1       1
Sushil Kumar    29      India   2012    Wrestling       0       1       0       1
Sushil Kumar    25      India   2008    Wrestling       0       0       1       1
Karnam Malleswari       25      India   2000    Weightlifting   0       0       1       1
Vijay Kumar     26      India   2012    Shooting        0       1       0       1
Gagan Narang    29      India   2012    Shooting        0       0       1       1
Abhinav Bindra  25      India   2008    Shooting        1       0       0       1
Rajyavardhan Rathore    34      India   2004    Shooting        0       1       0       1
M. C. Mary Kom  29      India   2012    Boxing  0       0       1       1
Vijender Singh  22      India   2008    Boxing  0       0       1       1
Saina Nehwal    22      India   2012    Badminton       0       0       1       1
```

XLS(X)

Microsoft Excel’s spreadsheet file format. Older versions use .xls files



ID	First	Last	Address	City	State	Zip Code
1	Walker	Adkins	6649 N Blue Gum St	New Orleans	LA	70116
2	Dave	Alston	4 B Blue Ridge Blvd	Brighton	MI	48116
3	Naomi	Patel	8 W Cerritos Ave #54	Bridgeport	NJ	8014
4	Judy	Ng	639 Main St	Anchorage	AK	99501
5	Luis	Sanchez	34 Center St	Hamilton	OH	45011
6	Teo	Tadeo	3 Mcauley Dr	Ashland	OH	44805
7	Britney	Spears	7 Eads St	Chicago	IL	60632
8	Foster	Jenkins	7 W Jackson Blvd	San Jose	CA	95111
9	Richard	Castle	5 Boston Ave #88	Sioux Falls	SD	57105
10	Meaghan	Garufi	228 Runamuck Pl #2808	Baltimore	MD	21224

Spreadsheet Structure:

This table is **structured** like many common spreadsheet programs, with rows and columns used to organize and display tabular data.

	A	B	C	D
1				
2				
3		B3		
4				

Header row: The first row of the table functions as a header, providing labels for each column.

Each cell can contain text, numbers, formulas, or other types of data.

- **Columns** in a spreadsheet represent variables or features.
- **Rows** contain observations or records of those variables.
- Each variable corresponds to a specific attribute or characteristic, and each observation captures the value of that attribute for a particular unit or entity.

1. Create a clear and concise header row: Your header row should describe each column's contents accurately and succinctly. Avoid abbreviations and jargon that may confuse users.
2. Remove duplicate rows: Before adding new data to the table, check for duplicate rows and remove them to prevent redundancy and maintain accuracy.
3. Ensure no missing values: Check for blank cells and fill them with appropriate values, or indicate "unknown" or "missing" if necessary. Missing values can lead to incorrect conclusions and skewed results.
4. Enter one value per cell: Make sure that every cell contains only one value, even if it seems repetitive or obvious. Multiple values in a single cell can cause problems during analysis and manipulation.