



# DSC3108: Big Data Mining and Analytics

## Lecture 01 (BSCS\_3:1)

Topic: *Fundamentals of Big Data*

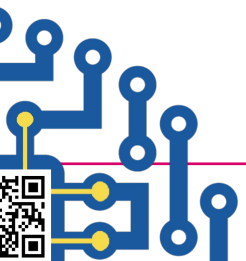
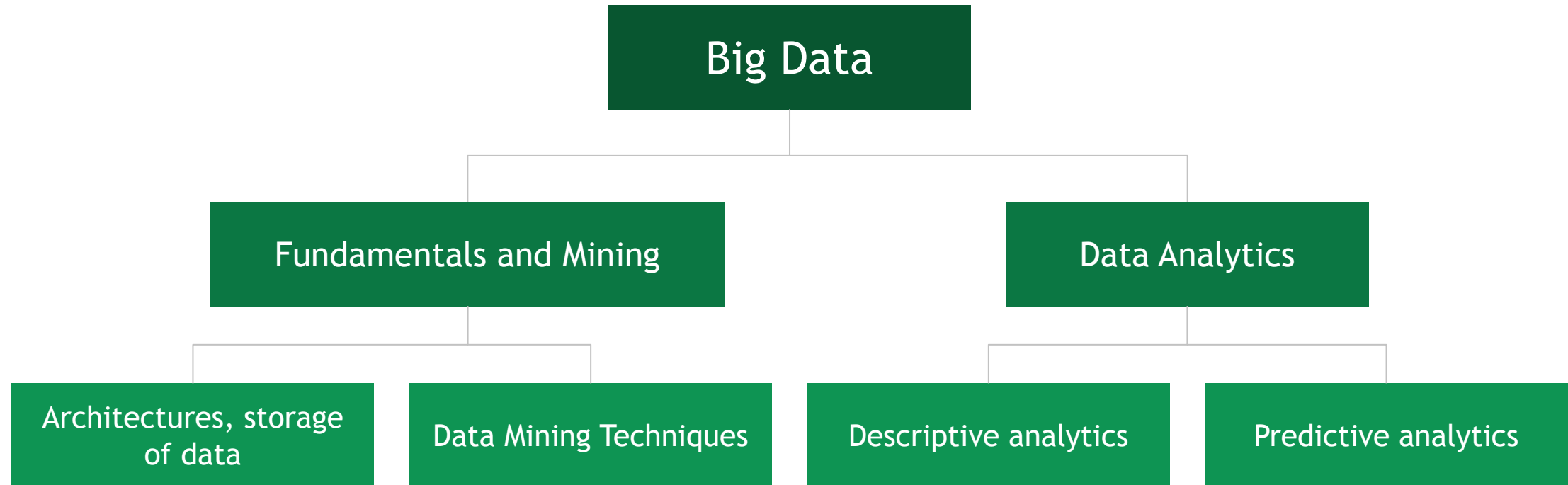
Dr. Daphne Nyachaki Bitalo  
Department of Computing & Technology  
Faculty of Engineering, Design & Technology

Fri 6<sup>th</sup> Sept 2024





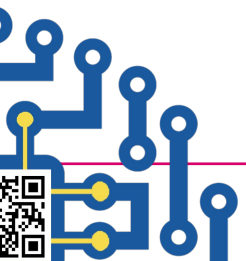
# COURSE OVERVIEW





# COURSE OVERVIEW

1. Semester-based module
2. 2 hours of lectures per week
3. 2 hours in-person practicals
4. Python libraries
5. Assessment via coursework and exams
6. Coursework contributes 60% to final grade
7. Exam contributes 40% to final grade
8. Recommended literature: See course outline on Moodle





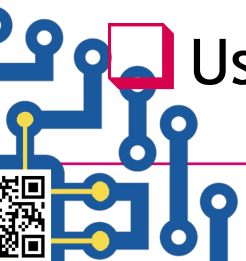
# Lecture Objectives and Learning outcomes

The Objectives of this lecture are :

- ☐ Understand the characteristics, challenges, and opportunities of big data
- ☐ Learn about the big data architectures
- ☐ Learn about big data storage and management.
- ☐ Understand the principles of data pre-processing.

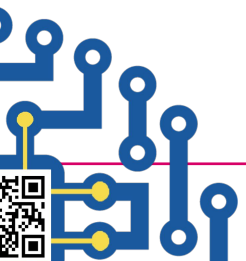
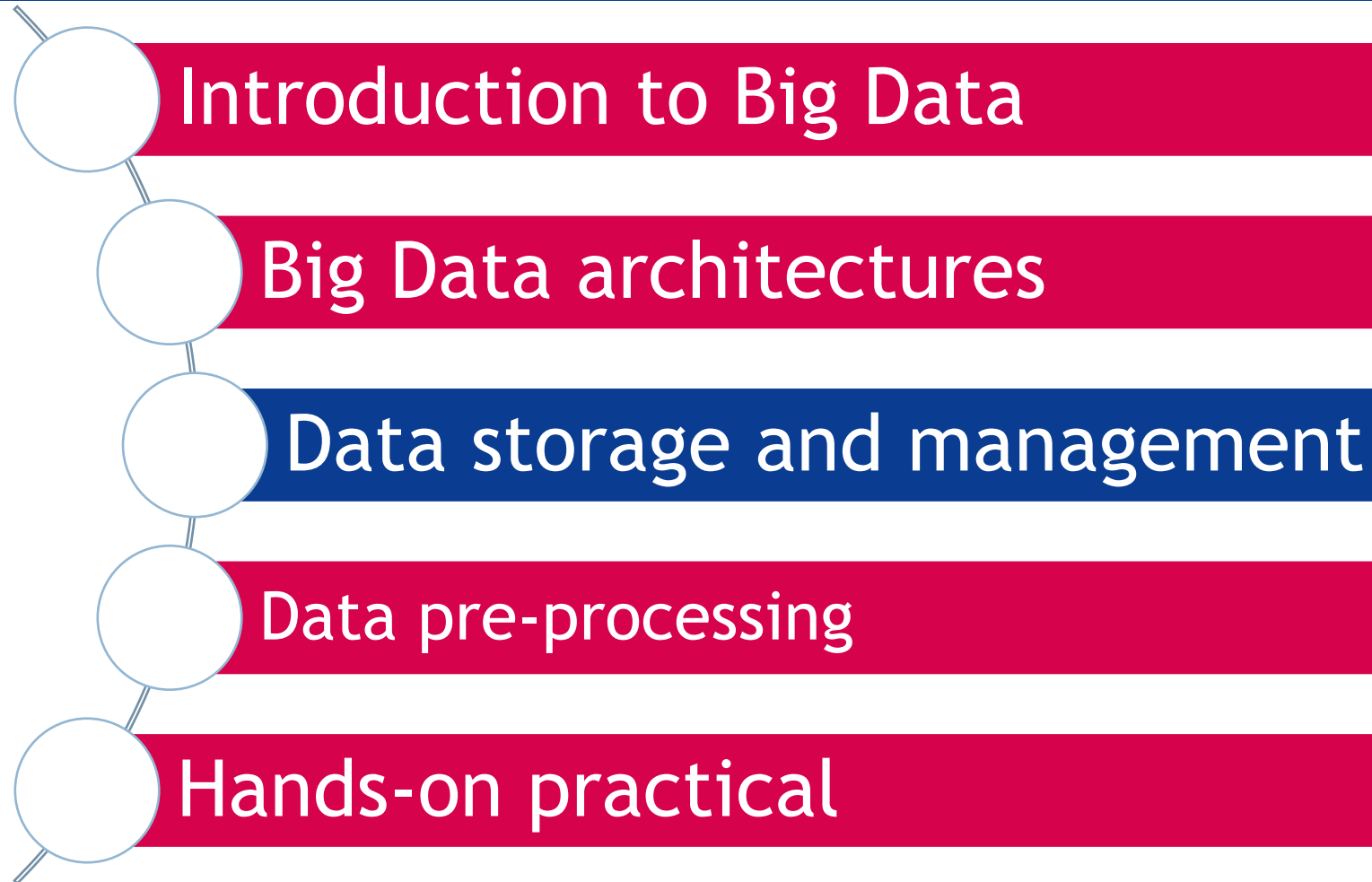
By the end of this lecture, students should be able to:

- ☐ Have an understanding of the applications of big data.
- ☐ Use python libraries to manage and pre-process big data.





# Lecture Overview





# What is Data?

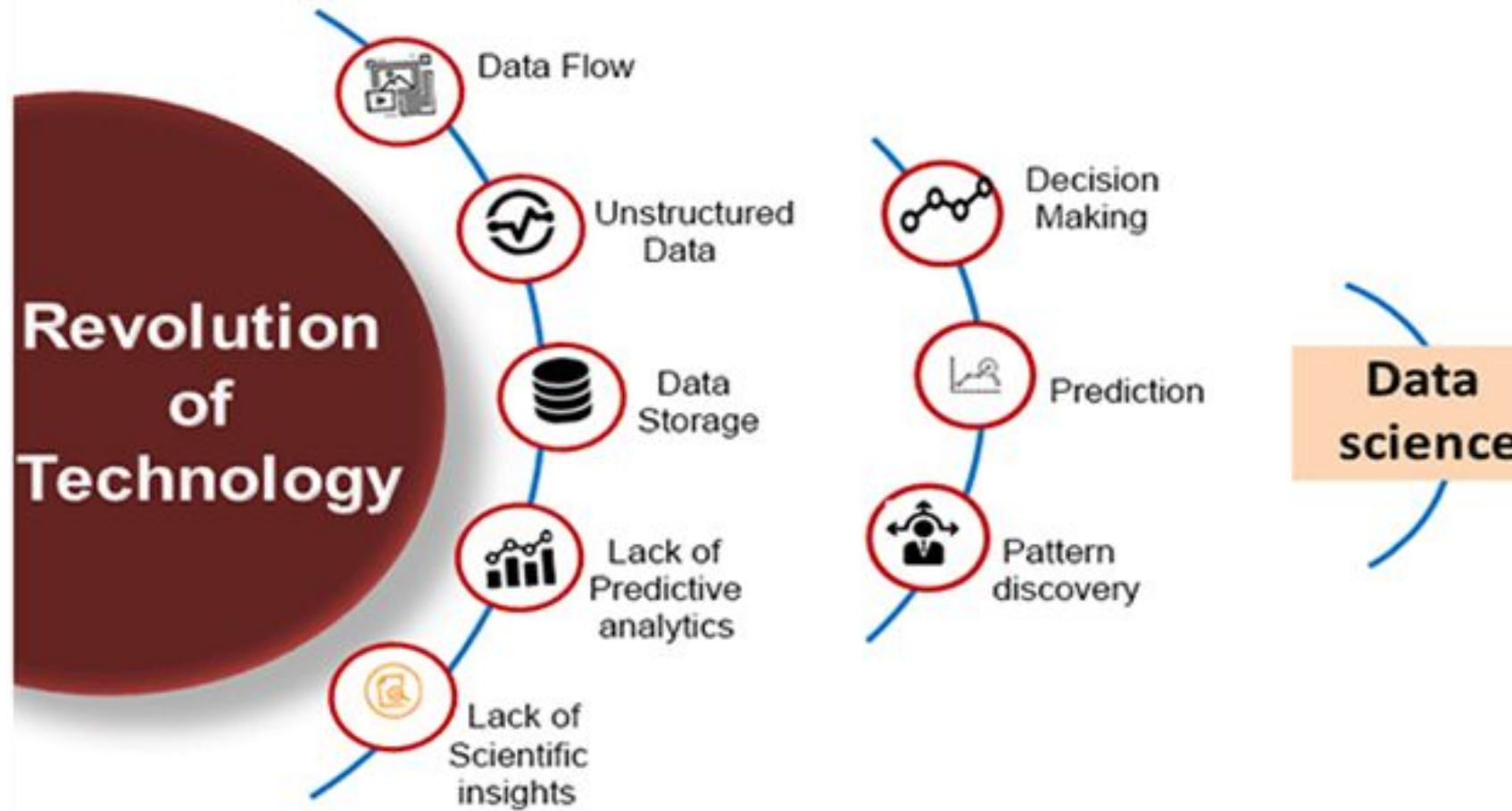
Data can be defined as any piece of facts that can be collected, analyzed, and interpreted.

It can come in many different forms, such as text, numbers, images, audio, and video.

For example, the number of hours spent studying for an exam, the temperature outside, or the number of likes on a social media .



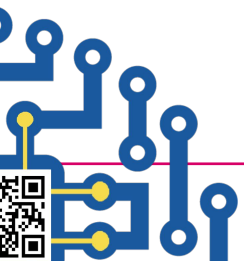
# Need for Data Science







# Need for Data Science







# Introduction to Big Data

- Big Data can bring “big values” to our life in almost every aspect.
- Technologically, Big Data is bringing about changes in our lives because it allows diverse and heterogeneous data to be fully integrated and analyzed to help us make decisions.
- Today, with the Big Data technology, thousands of data from seemingly unrelated areas can help support important decisions. This is the power of Big Data.

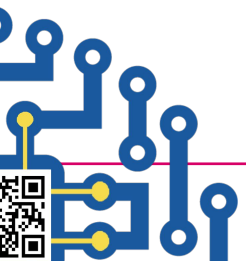




# Introduction to Big Data

## Areas of Applications

- Health and Wellbeing
- Policy making and public opinions
- Smart cities and more efficient society
- New online educational models: MOOC and Student-Teacher modeling
- Robotics and human-robot interaction
- Much of this power hinges on Research on Analytics





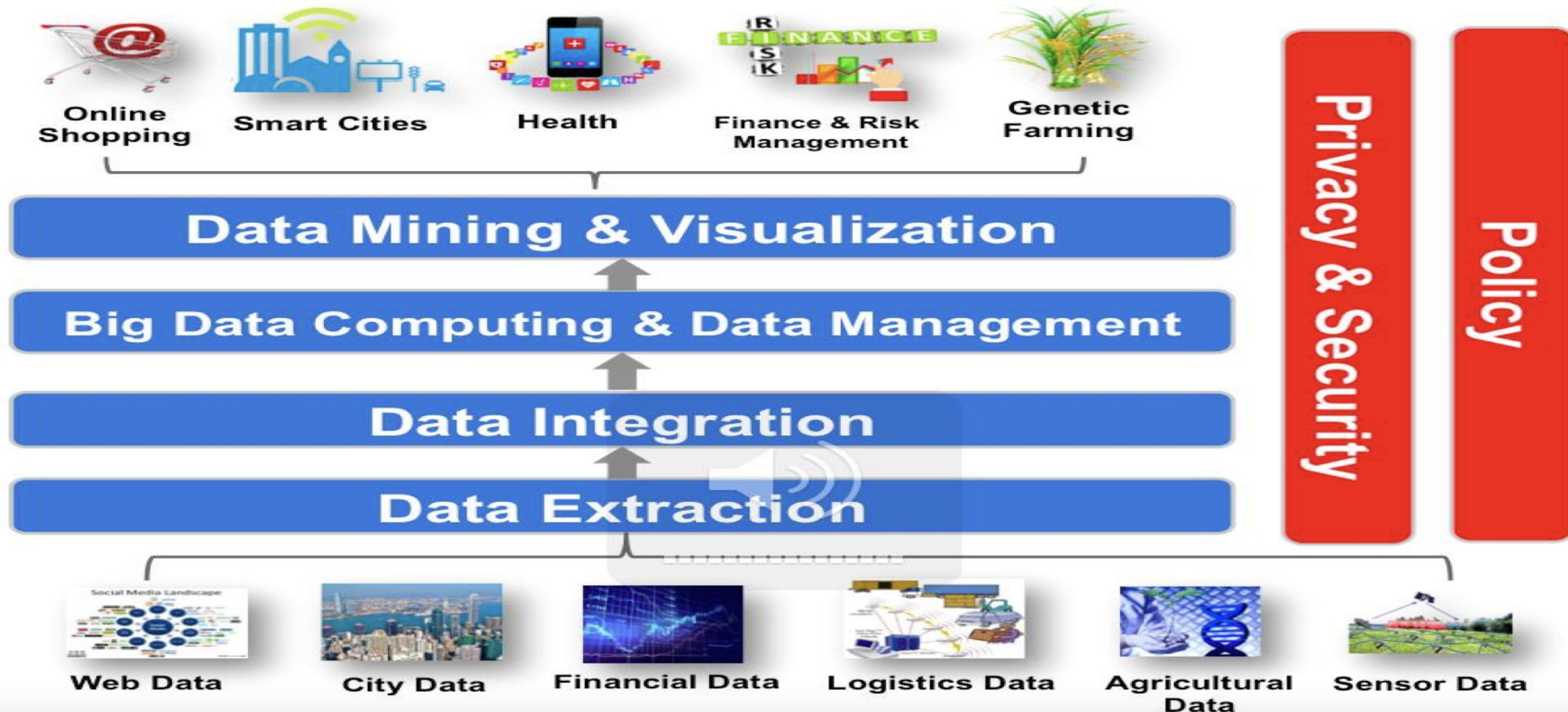
# Introduction to Big Data

- **Big Data** refers to extremely large datasets that are difficult to process using traditional data processing applications and tools. These datasets are characterized by their **volume**, **velocity**, **variety** and **veracity**.
- **Volume:** The sheer amount of data generated is massive.
- **Velocity:** Data is generated at a high speed, often in real-time.
- **Variety:** Data comes in various formats, including structured and unstructured data.
- **Veracity:** Quality and robustness of data





# Objectives of Big Data

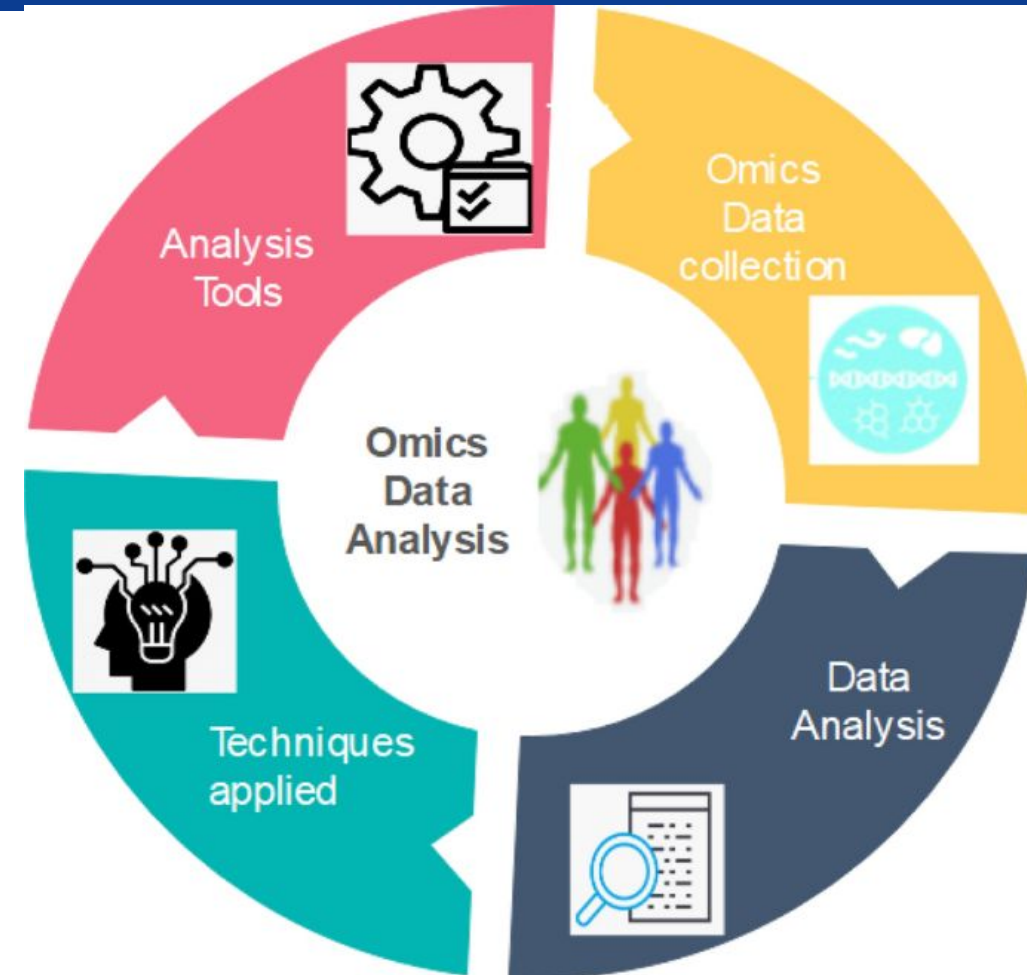


# Objectives of Big Data

**Big Data Analytics:** data mining, visualisation and machine learning

**Big Data Computing:** data center support for Analytics

Big data collection and transformation, integration and distributed data management and computing





# Objectives of Big Data

## Big Data Theory, Privacy & Security issues on Analytics

Big data sampling and statistical theory,  
Big data security and privacy

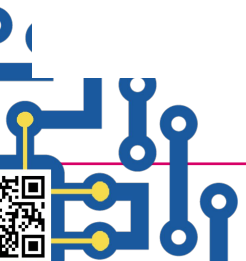
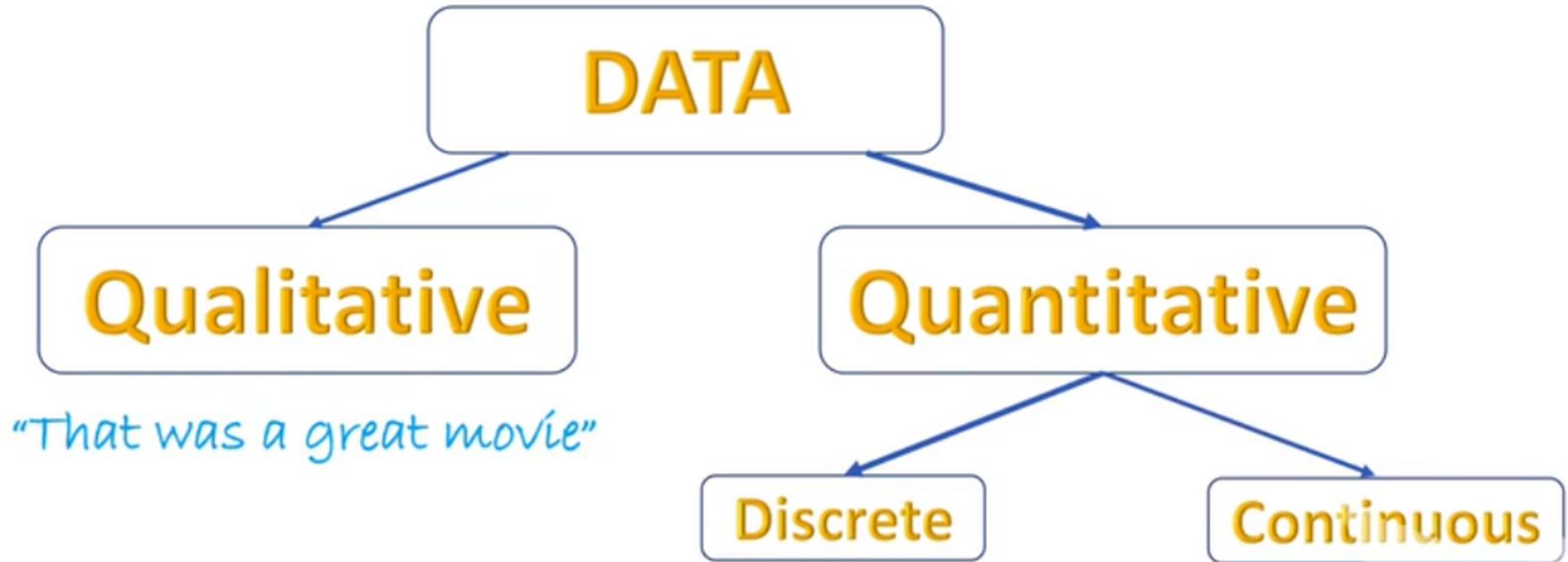
## Big Data Science: 4th Paradigm - Analytics for Science and Engineering

Big Data and Multi-disciplines (Bio, Chemistry, Engineering, Social, Business)





# Types of Data

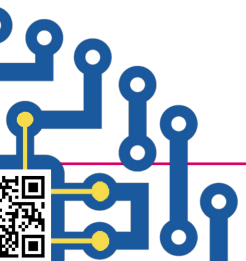






# Two major categories of Data

**1. QUALITATIVE DATA** (Non-Numerical): This is data that is descriptive in nature ie texts and is non numerical in nature. It is collected through surveys, questionnaires, interviews or observations, your social media posts.



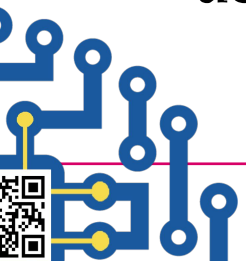


# Two major categories of Data

## 2. QUANTITATIVE DATA (Numerical)

Consists of;

- i) **Discrete:** Takes on values that are typically whole numbers. For example; Number of people who visit a website each week, Total amount of transactions processed in a day.
- ii) **Continuous:** This is a type of numerical data that can take any value within a certain range, often with infinite possibilities for example;
  - i) Average distance travelled by a moving car by which is constantly updated as the car continues to move.
  - ii) Temperature measures as these keep varying according to the time.
  - ii) Height measures of an individual





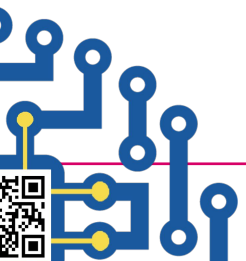
# Other types of Data

All pieces of data that are collected have a feature or characteristic known as a **data type** that helps to identify the data whether it is a character, integer which is helpful in analytics.

The different Data types.

i)String: String data is treated as text composed of letters and numbers for example room number "room 22".

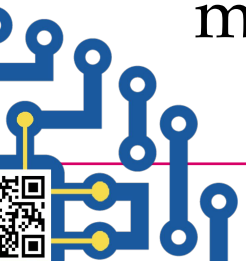
ii)Integer: Consists of whole numbers or numbers that do not include decimals or fractions. For example 0, 1, 2, 3





# Other types of Data

- iii) Floating point: Consists of numbers with decimals ie 1.2, 0.0003, -4.56
- iv) Date and time: Stores an instant time that is expressed as a calendar date and time in the formats of 2001-03-15, YYYY-MM-DD, h:mm:s, 2001-03-15 11:24:09
- v) Boolean: Data is treated as true or false and can be represented as "yes" or "1" (for true) and "No" or "0" ( for false). For example; 12 is greater than 10 = true. For example; Assume we have a database where we store employee details and a particular user Paul is not a manager this will eventually mean that; " Paul is not a manager=yes".





# Structured Data

In order to process, store, and analyze all of these different types of data, it is important to think about whether they are structured data or unstructured data.

**STRUCTURED DATA:** Makes up about 10%-20% of generated data and has clearly defined data types and patterns that make them easily stored and organized into columns and rows. This organization makes structured data easy to search and analyze.

**Sources of structured data** include sales records, airline reservation systems, and inventory control. Structured data is usually stored in relational databases such as Structured Query Language (SQL) databases or in spreadsheets such as Microsoft Excel.







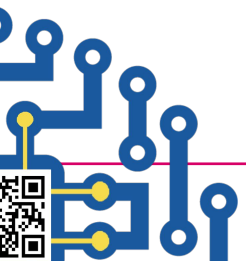
# Structured Data

	A	B	C	D	E	F	G
1	Purchase ID ▼	Last name ▼	First name ▼	Birthday ▼	Country ▼	Date of purchase ▼	Amount of purchase ▼
2	1	Davidson	Michael	04/03/1986	United States	10/12/2016	37
3	2	Vito	Jim	09/01/1994	United Kingdom	02/02/2016	85
4	3	Johnson	Tom	23/08/1972	France	02/11/2016	83
5	4	Lewis	Peter	18/10/1979	Germany	22/11/2016	27
6	5	Koenig	Edward	13/05/1983	Argentina	26/03/2015	43
7	6	Preston	Jack	16/06/1991	United States	06/11/2016	77
8	7	Smith	David	11/03/1965	Canada	15/11/2016	23
9	8	Brown	Luis	03/09/1997	Australia	03/07/2015	74
10	9	Miller	Thomas	07/01/1980	Germany	07/11/2016	13
11	10	Williams	Bill	26/07/1960	United States	20/11/2015	80
12	11	Gemini	Alexia	12/09/1995	Canada	11/03/2017	35
13	12	Bond	James	25/02/1975	United Kingdom	12/08/2017	40
14	13	Burgle	Patricia	01/12/1990	United States	18/01/2015	55
15	14	Reding	Michelle	07/04/1985	Canada	23/02/2017	28
16	15	Harvey	Billy	14/07/1971	United Kingdom	12/01/2016	41
17							



# Structured data tools

## Tools for working with structured data



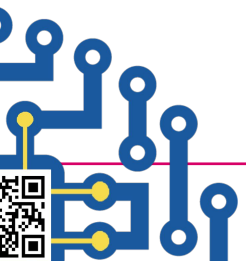




# Unstructured Data



**UNSTRUCTURED DATA:** Makes up most data that is generated, about 80%, and cannot be organized into row and columns. This makes unstructured data difficult to search, manage, and analyze.

**Sources of unstructured data** include images, PDFs, sensor data, and social media posts. Unstructured data is usually stored in a non-relational database also known as NoSQL Database.



# Unstructured Data

## Unstructured data types


 <b>Text files and documents</b>	 <b>Server, website and application logs</b>	 <b>Sensor data</b>	 <b>Images</b>
 <b>Video files</b>	 <b>Audio files</b>	 <b>Emails</b>	 <b>Social media data</b>

# Unstructured Data

Introducing one of Australia's greatest treks and one of our newest trips! 🌄

The Cradle Mountain Overland track offers some of Tasmania's most stunning scenery – dramatic valleys, temperate rainforests, beautiful lakes and more. And this 6-day camping trip is the perfect way to experience it, alongside like-minded adventurers and experienced guides.


Why travel there with us? ... [See More](#)



[View Similar Products](#)

116 11 Comments 12 Shares

**Jessica Chapman**  
\*sigh\* I miss travel. With COVID travel restrictions here it's going to be a while.  
Like · Reply · 1w  
↳ 1 Reply

**Abderrahman Chafiq**  
snow in the Atlas mountains from Morocco 🇲🇦🇲🇦🇲🇦🇲🇦  
  
Like · Reply · 1w 2

**Dian Clayton**  
Loved it, but snow on Mt Osser in November! 😊  
Like · Reply · 1w

**Hany Sayed**  
Beautiful 🌟 1  
Like · Reply · 1w

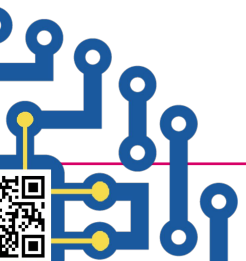
**Kelly McCarthy**  
Cradle Mountain is one of my favorite places. So many wild wombats to observe!!  
Like · Reply · 1w 1

Most Relevant is selected, so some comments may have been filtered out.



# Human-generated unstructured data

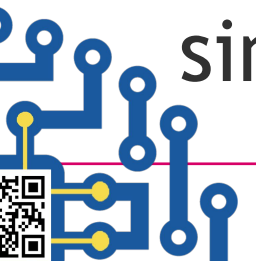
- ❑ A collection of social media posts about a popular TV show, with varying formats and no clear structure.
- ❑ •"OMG just finished the finale of the show and I am SHOOK! 🤯 #mindblown #bestshowever"
- ❑ •"Finally got around to watching the show and it did not disappoint! Already looking forward to the next season #bingewatching"
- ❑ •"I don't get the hype around this show, I thought it was boring and predictable 🙄 #unpopularopinion"
- ❑ •"Just finished the show and now I don't know what to do with my life 😭 #postshowdepression"





# Human-generated unstructured data

- ❑ Websites: YouTube, Instagram etc
- ❑ Mobile data: text messages, locations.
- ❑ Communications: IMs, dictaphone recordings.
- ❑ Media: MP3, digital photos, audio recordings and video files.
- ❑ Business applications: MS Office documents, PDFs and similar.



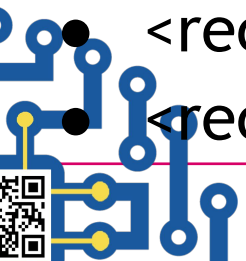


# Semi-structured Data

- Contains both structured and unstructured data.
- We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in a database management system (DBMS).
- Example of semi-structured data is a data represented in an XML file.

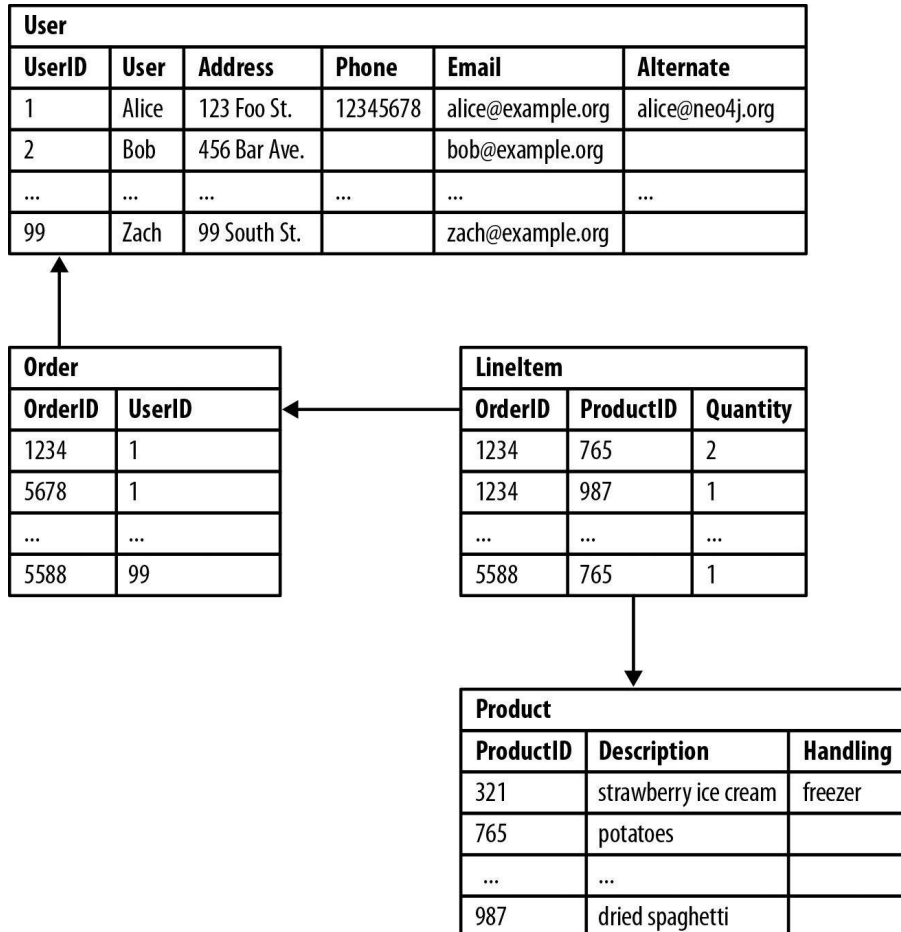
## Personal data stored in an XML file-

- `<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>`
- `<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>`
- `<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>`
- `<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>`





# Example of a Relational Database



From the top-down, we can see that UserID 1 refers to the customer Alice, who had two Order IDs of '1234' and '5678'.

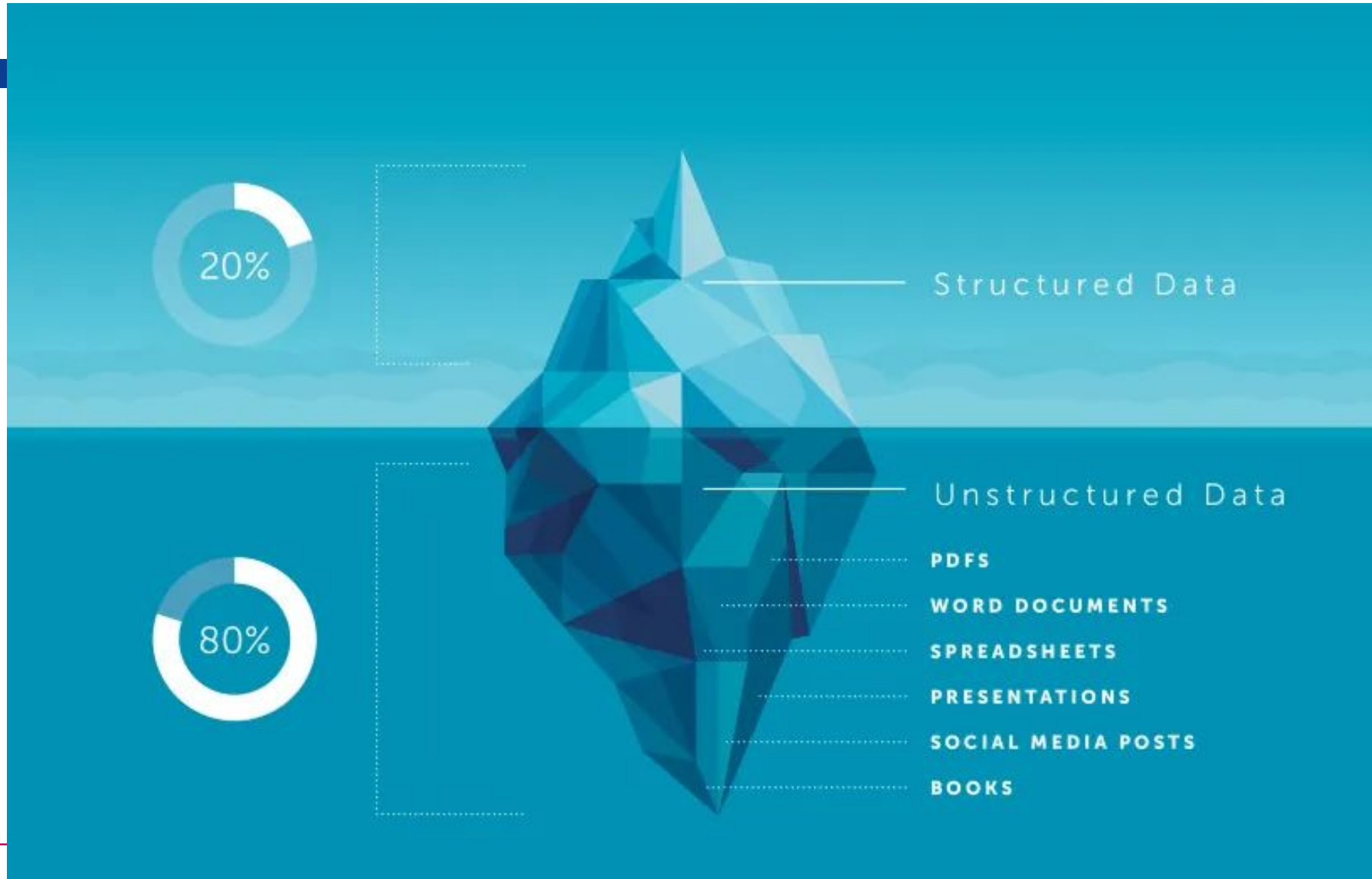
Next, Alice had two ProductIDs of '765' and '987'.

Finally, we can see Alice purchased two packages of potatoes and one package of dried spaghetti.





# Structured Data vs. Unstructured Data





# Challenges of Big Data

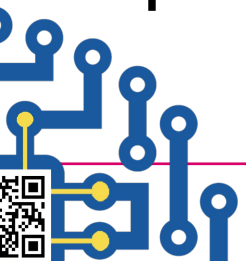
- **Storage:** Storing massive amounts of data requires efficient and scalable storage solutions.
- **Processing:** Processing large datasets can be computationally intensive and time-consuming.
- **Analysis:** Extracting meaningful insights from big data requires advanced analytics techniques/skills.
- **Integration:** Integrating data from multiple sources can be challenging due to inconsistencies and varying formats.
- **Security:** Protecting sensitive data from unauthorized access is a major concern





# Tools for Big Data

- **Hadoop:** A distributed computing framework for processing large datasets.
- **Spark:** A fast and general-purpose cluster computing system.
- **NoSQL Databases:** Databases designed to handle large-scale, unstructured data.
- **Cloud Computing:** Utilizing cloud-based infrastructure to store, process, and analyze big data.





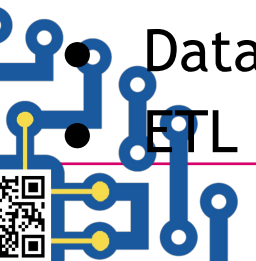
# Big Data architectures

## Hadoop

- **Distributed File System (HDFS):** Hadoop's core component, providing a scalable and fault-tolerant distributed file system.
- **MapReduce:** A programming model for processing large datasets in parallel across a cluster of computers.
- **YARN:** A resource management system that allocates resources to applications running on the Hadoop cluster.

## Use Cases:

- Batch processing of large datasets
- Data warehousing and analytics
- ETL (Extract, Transform, Load) processes





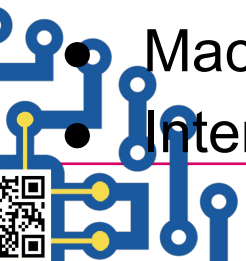
# Big Data architectures

## Spark

- **In-Memory Processing:** Spark processes data in memory, which can significantly improve performance compared to disk-based processing.
- **Unified Architecture:** Spark provides a unified architecture for batch processing, streaming, and machine learning workloads.
- **Rich Ecosystem:** Spark has a rich ecosystem of libraries and frameworks for various data processing tasks.

## Use Cases:

- Real-time data processing
- Machine learning and artificial intelligence
- Interactive data analysis





# Big Data architectures

## NoSQL Databases

- **Key-Value Stores:** Store data as key-value pairs, high performance for simple data access.
- **Document Stores:** Store data as documents, suitable for semi-structured and unstructured data.
- **Wide-Column Stores:** Store data as columns,
- **Graph Databases:** Store data as nodes and relationships, making them ideal for graph-based analysis.

## Use Cases:

- Real-time applications
- Content management systems
- Social networking platforms
- Recommendation systems

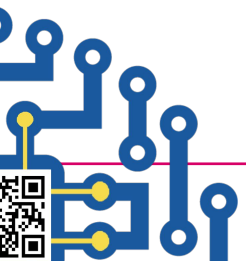




# Big Data Pre-processing

Data preprocessing is a crucial step in the data science pipeline, involving various techniques to prepare raw data for analysis. It ensures data quality, consistency, and suitability for modeling.

**It involves Cleaning, Integration, Transformation, and Reduction**







# Big Data pre-processing

## Cleaning

- **Handling missing values:** Imputation techniques (mean, median, mode, regression, etc.) or deletion.
- **Dealing with outliers:** Identification (statistical methods, visualization) and removal or correction.
- **Noise reduction:** Smoothing techniques (e.g., moving average) to remove noise or inconsistencies.
- **Data correction:** Identifying and correcting errors or inconsistencies in the data.



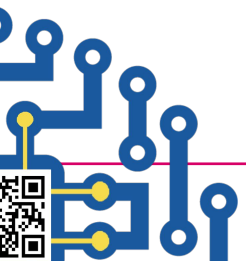


# Big Data pre-processing

## Integration

- **Merging datasets:** Combining multiple datasets based on common keys or identifiers.
- **Data standardization:** Ensuring consistency in data formats, units, and encoding.
- **Entity resolution:** Identifying and merging duplicate records representing the same entity.

Key for descriptive and predictive analysis





# Big Data pre-processing

## Transformation

- **Normalization:** Scaling data to a specific range (e.g., 0-1) to improve model performance (e.g. converting categorical data).
- **Feature engineering:** Creating new features from existing ones to capture relevant information.
- **Aggregation:** Combining multiple data points into a single value (e.g., calculating averages or sums).
- **Discretization:** Converting continuous data into discrete categories.



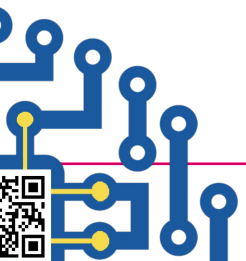


# Big Data pre-processing

## Reduction

- **Dimensionality reduction:** Reducing the number of features while preserving essential information.
- **Feature selection:** Choosing the most relevant features for analysis.
- **Sampling:** Selecting a subset of data for analysis to reduce processing time.

Key for predictive analysis

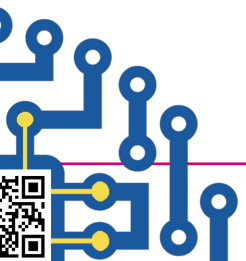




# Data pre-processing using python

## PRACTICUM USING PANDAS AND NUMPY

- Generate a dataset in excel with 8 variables/columns and call it Student\_performance.xlsx
- (Access\_no, gender, age, course, DSC3108, DSC3114, MTH3108, SYE3104)
- Populate the rows with data that you will use in the practical





UGANDA CHRISTIAN  
UNIVERSITY

A Centre of Excellence in the Heart of Africa



## Uganda Christian University

P.O. Box 4 Mukono, Uganda

Tel: 256-312-350800

 <https://ucu.ac.ug/> Email: [info@ucu.ac.ug](mailto:info@ucu.ac.ug)

 @ugandachristianuniversity  @UCUniversity

 @UgandaChristianUniversity



## Department of Computing & Technology FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY

Tel: +256 (0) 312 350 863 | WhatsApp: +256 (0) 708 114 300

 @ucuc Computeng  @ucu\_ComputEng

 <https://cse.ucu.ac.ug/> Email: [dct-info@ucu.ac.ug](mailto:dct-info@ucu.ac.ug)

A Complete Education for A Complete Person

P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: [info@ucu.ac.ug](mailto:info@ucu.ac.ug) Web: <https://ucu.ac.ug>

Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda