Advanced Data Science

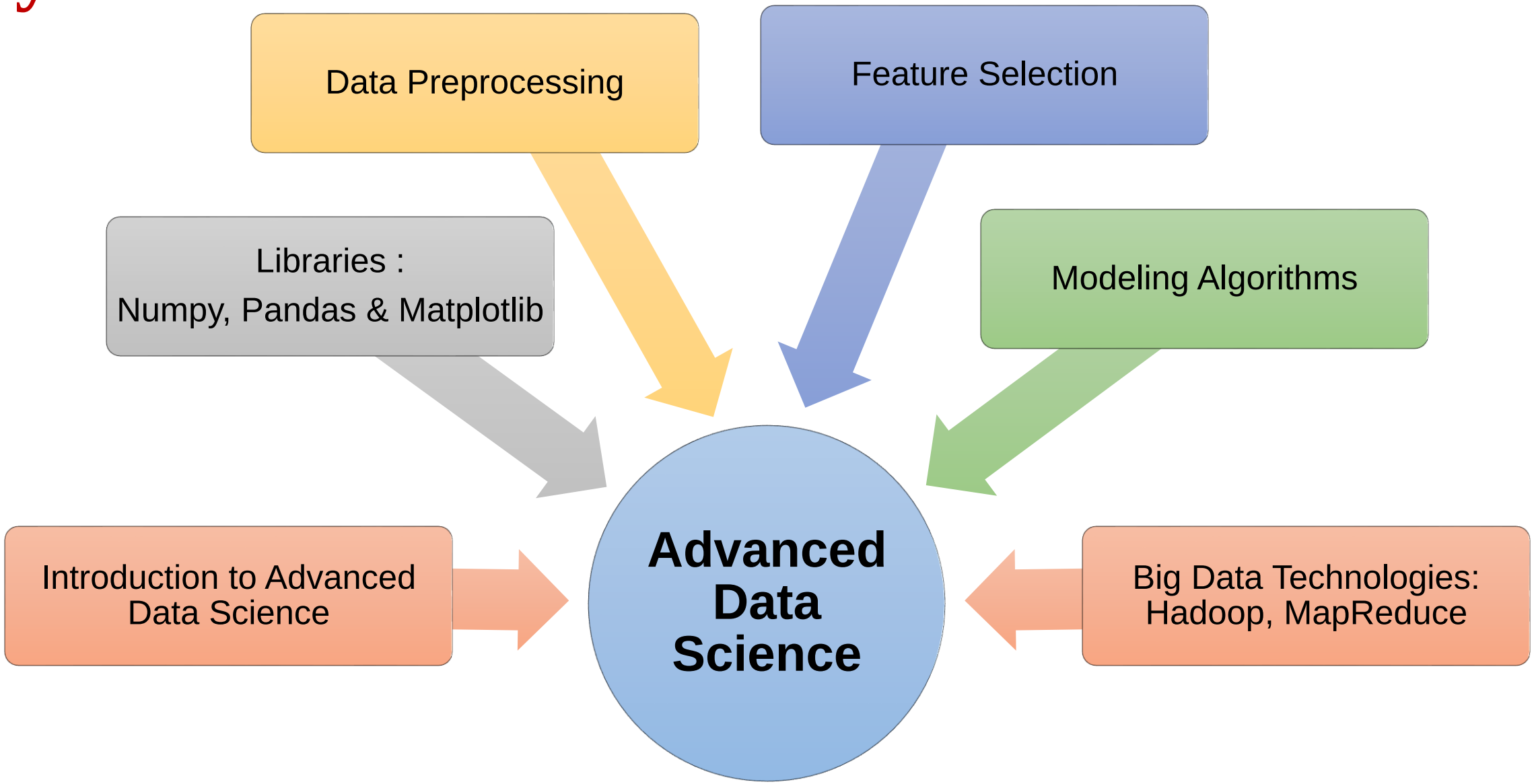
Techniques, Tools, and Applications

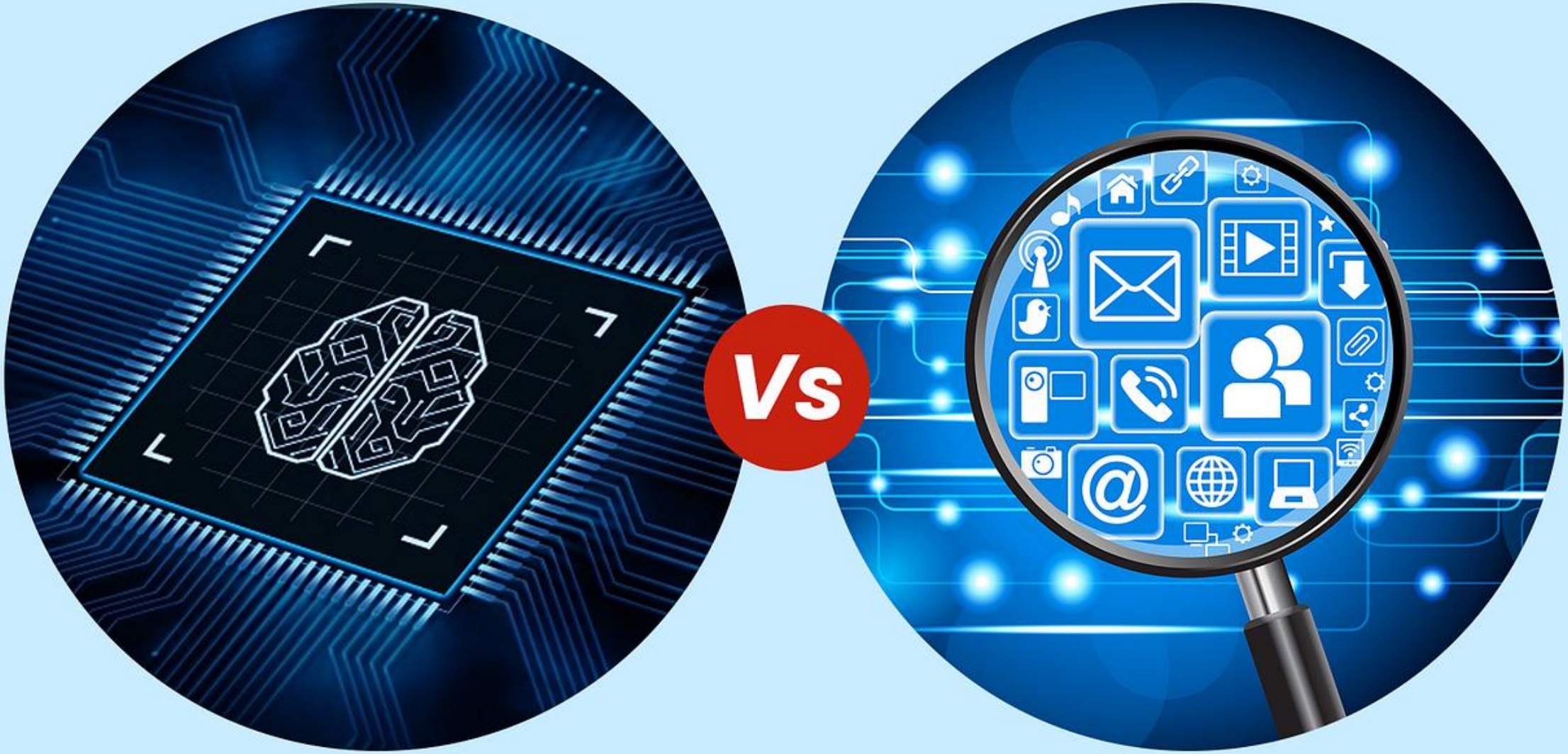
Session 1

Kiran Waghmare
Program Manager
C-DAC Mumbai

Schedule : BIA - 703 sessions		
Session	**Date**	**Time**
1	20-Jul-24	9:30AM to 12:30PM
2	27-Jul-24	9:30AM to 12:30PM
3	3-Aug-24	9:30AM to 12:30PM
4	10-Aug-24	9:30AM to 12:30PM
5	17-Aug-24	9:30AM to 12:30PM
6	24-Aug-24	9:30AM to 12:30PM
7	14-Sep-24	9:30AM to 12:30PM
8	21-Sep-24	9:30AM to 12:30PM
9	28-Sep-24	9:30AM to 10:30AM

Syllabus


Data Preprocessing
Feature Selection
Libraries :
Numpy, Pandas & Matplotlib
Modeling Algorithms
Introduction to Advanced Data Science
Advanced Data Science
Big Data Technologies: Hadoop, MapReduce



Data Information

Data → Information

Context + Processing

Data

- Raw facts, figures and statistics
- No contextual meaning
- Data can be in characters, numbers, images, words

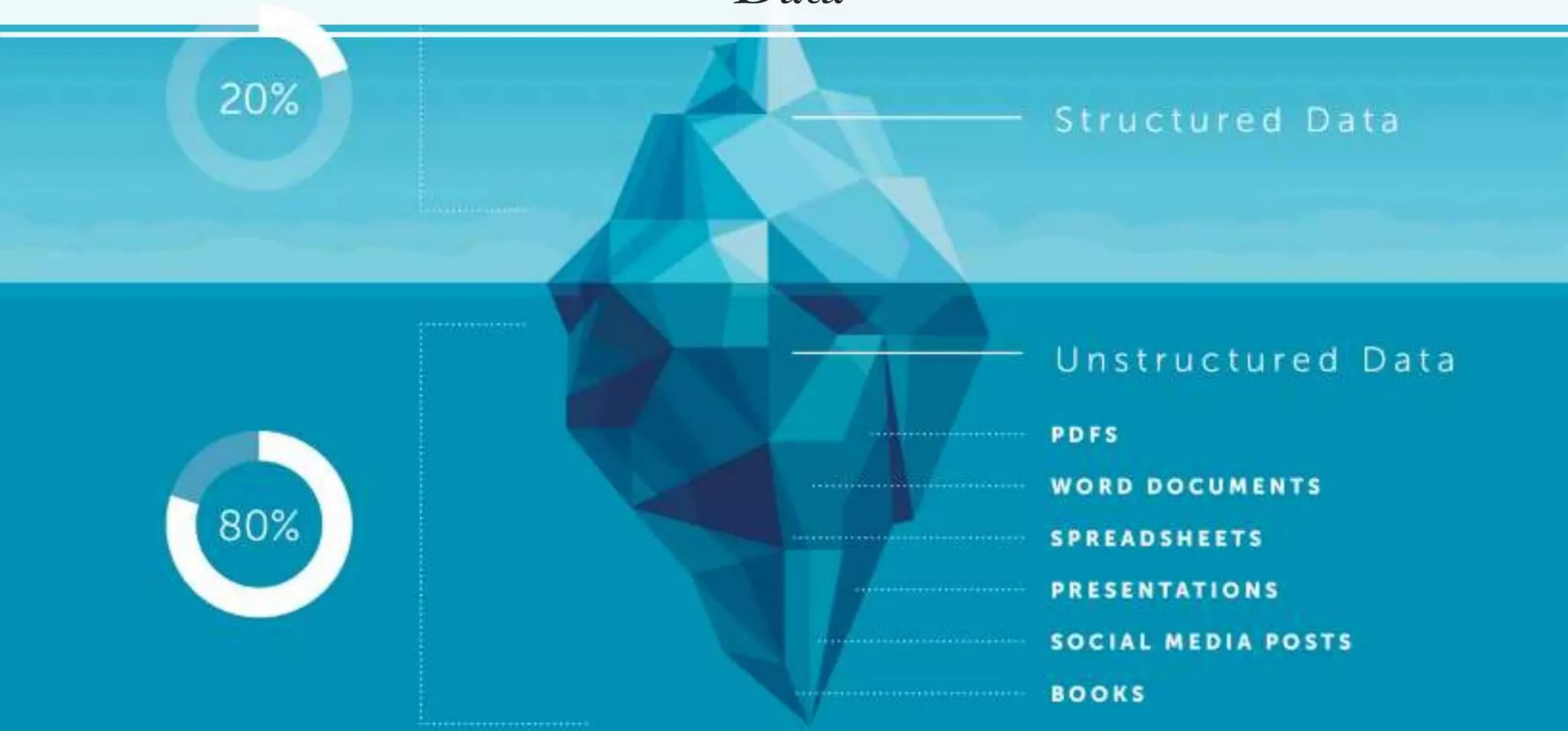
Information

- Processed / Organized Data
- Exact meaning and organized context
- Organized and presented in context – Value added to data

Measure of Data in Files – File Size

Name	Equal To	Size(In Bytes)
Bit	1 Bit	1/8
Nibble	4 Bits	1/2 (rare)
Byte	8 Bits	1
Kilobyte	1024 Bytes	1024
Megabyte	1, 024 Kilobytes	1, 048, 576
Gigabyte	1, 024 Megabytes	1, 073, 741, 824
Terrabyte	1, 024 Gigabytes	1, 099, 511, 627, 776
Petabyte	1, 024 Terabytes	1, 125, 899, 906, 842, 624
Exabyte	1, 024 Petabytes	1, 152, 921, 504, 606, 846, 976
Zettabyte	1, 024 Exabytes	1, 180, 591, 620, 717, 411, 303, 424
Yottabyte	1, 024 Zettabytes	1, 208, 925, 819, 614, 629, 174, 706, 176

Data


20%
Structured Data
Unstructured Data
80%
PDFS
WORD DOCUMENTS
SPREADSHEETS
PRESENTATIONS
SOCIAL MEDIA POSTS
BOOKS

Types of Data and it's Representation

Structured Data

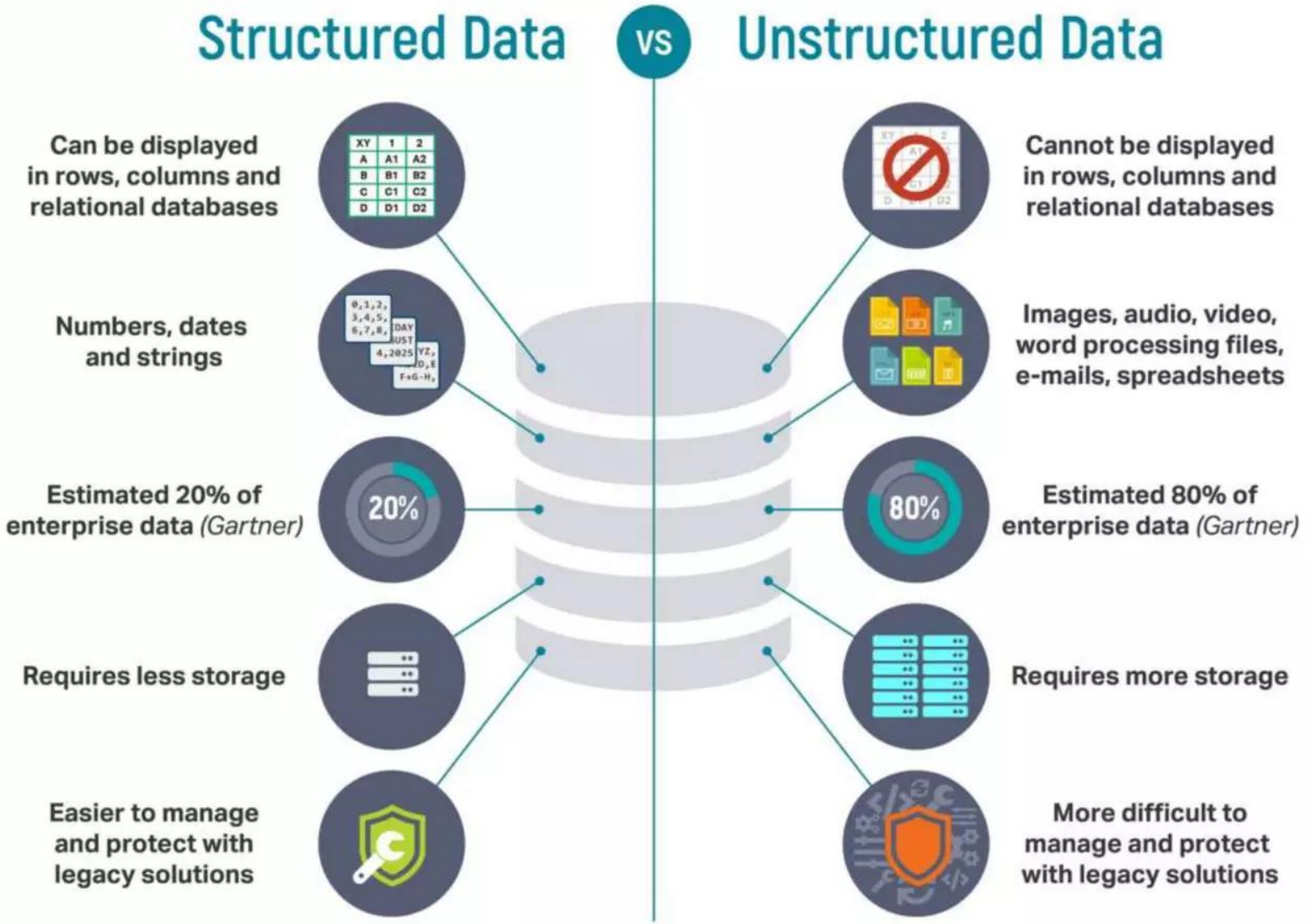
- Predefined data models
- Stored in Rows and Columns
- Examples: Dates, Phone Number, Names

Semi-Structured Data

- Loosely organized into categories using meta tags
- Stored in abstract and figures – HTML, XML, JSON
- Examples: Server Logs, Tweets organized by Hashtags

Unstructured Data

- No predefined data models
- Stored in various forms – image, audio, video, text
- Examples: Documents, Image Files, Emails & Messages


Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Challenges in Advanced Data Science

- **Content:**
  - Data
  - Quality and Cleaning
  - Model Interpretability
  - Scalability
  - Ethical Considerations

Need of Advance data Science

- **Historical Data:**
  - Earlier, data was less and structured, easily processed with BI tools.
- **Current Data Explosion:**
  - Approximately 2.5 quintillion bytes of data generated daily.
- **Future Prediction:**
  - By 2020, 1.7 MB of data created every second per person.
- **Organizational Need:**
  - Companies require data to grow and improve businesses.

Definition

- **Definition of Advanced Data Science**
  - Advanced Data Science refers to the use of
    - complex algorithms,
    - large-scale data processing, and
    - cutting-edge technologies
  - to analyze and interpret large volumes of data.
  - It goes beyond traditional data analysis
  - by incorporating techniques from
    - machine learning,
    - artificial intelligence, and
    - big data technologies.

Understand the Advanced Data Science

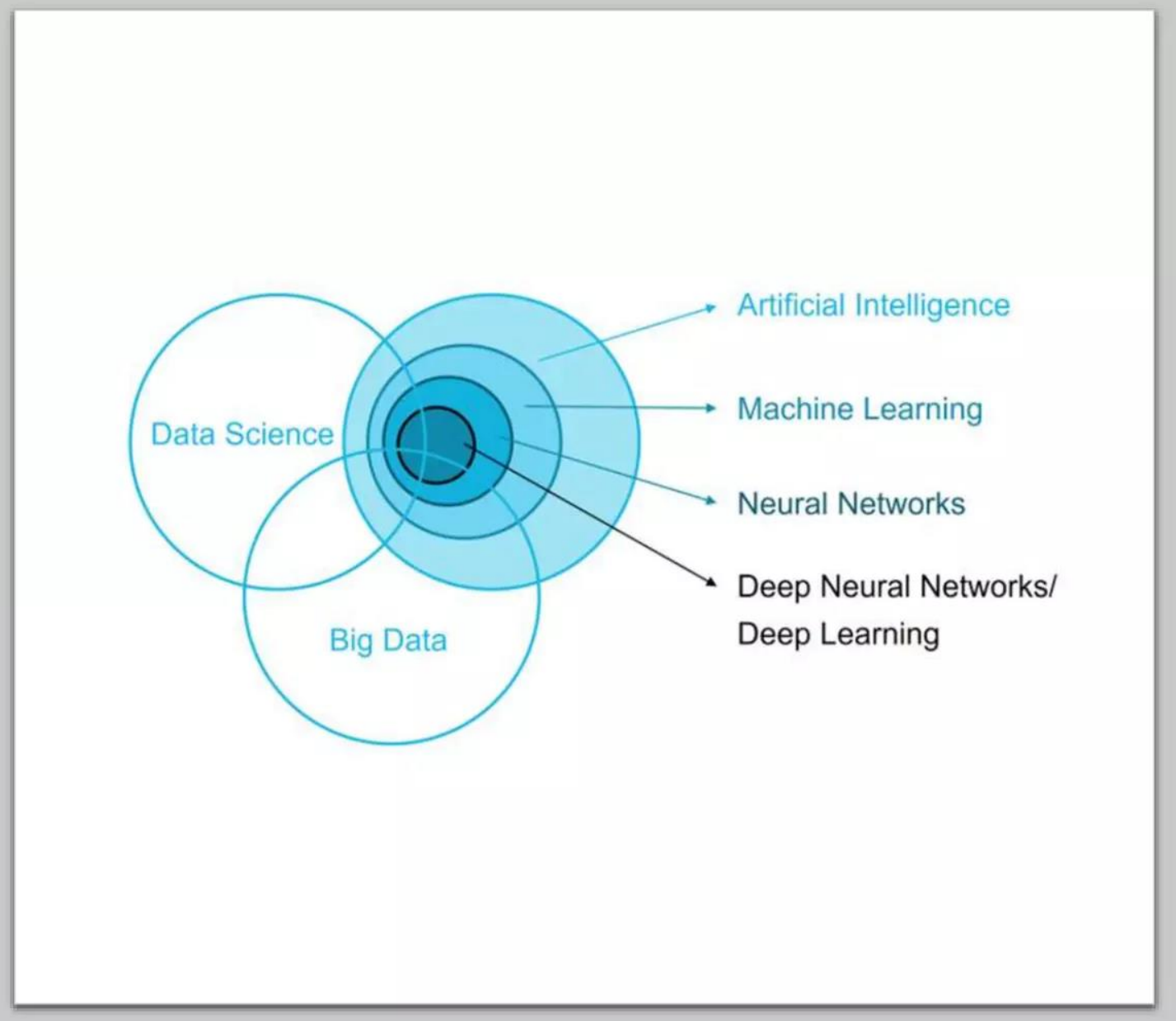
- **Definition:**
  - Advanced data science involves leveraging sophisticated analytical techniques to extract deeper insights from data.
- **Scope:**
  - Encompasses various fields such as machine learning, deep learning, and natural language processing.
- **Objective:**
  - Focuses on predictive modeling, pattern recognition, and decision-making support.
- **Interdisciplinary Nature:**
  - Combines expertise in statistics, computer science, and domain-specific knowledge.
- **Application Areas:**
  - Includes healthcare, finance, marketing, and more, driving innovation and efficiency across industries.

Importance:

- **Informed Decision-Making:**
  - Provides a factual basis for making strategic and operational decisions.
- **Problem-Solving:**
  - Helps identify root causes of issues and evaluate potential solutions.
- **Trend Identification:**
  - Uncovers trends and patterns that can inform business strategies and operations.
- **Risk Management:**
  - Assists in predicting and mitigating risks by understanding past behaviors and outcomes.

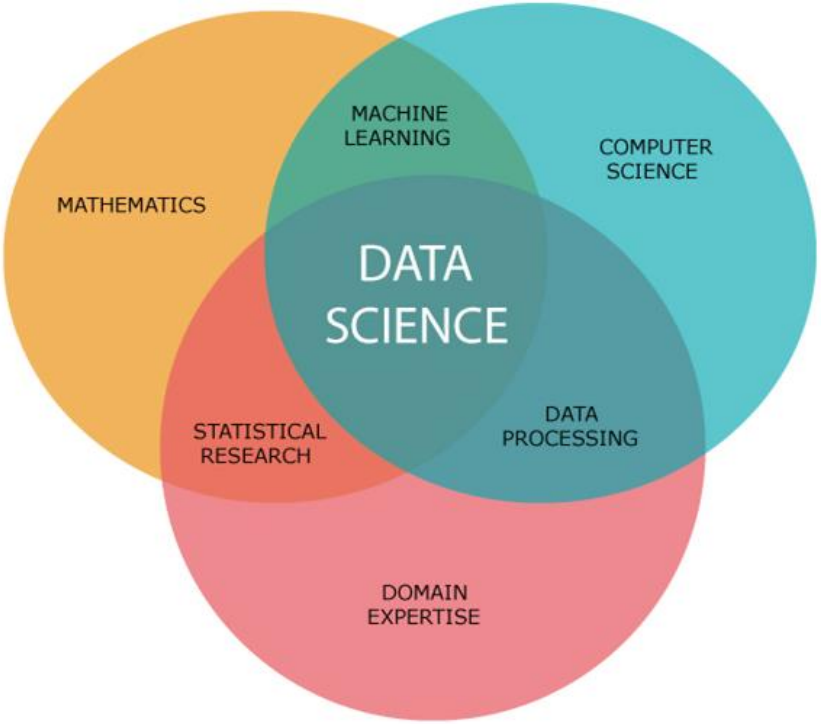
Data Science

- Data science enables businesses to **Process** huge amounts of structured and unstructured **Big Data** to detect patterns
- Alexa or Siri for a recommendation demands data science
- Operating a self-driving car
- Search Engine
- Chatbot for customer service

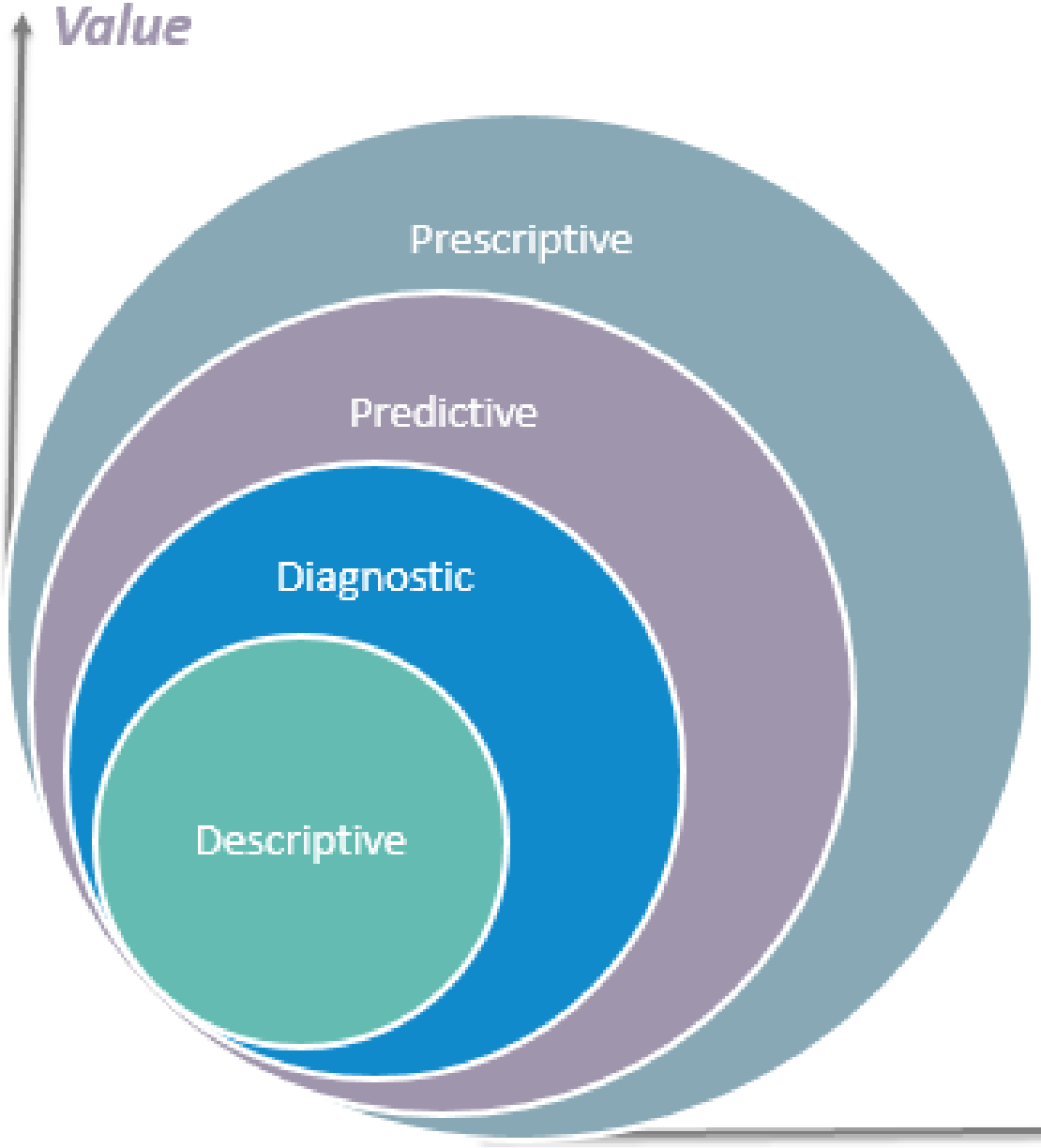


Key Components

- **Machine Learning:**
  - Utilizes algorithms that allow computers to learn from and make predictions based on data.
  - Examples: Decision Trees, Support Vector Machines (SVM), and Neural Networks.
- **Deep Learning:**
  - A subset of machine learning focused on neural networks with many layers (deep neural networks).
  - Applications: Image and speech recognition, natural language processing.
- **Natural Language Processing (NLP):**
  - Involves the interaction between computers and human language.
  - Techniques: Sentiment analysis, language modeling, and text generation.
- **Big Data Technologies:**
  - Tools and frameworks designed to handle and process large datasets efficientl
  - Examples: Hadoop, Apache Spark, and NoSQL databases.
- **Data Visualization:**
  - The graphical representation of data to help understand trends, patterns, and insights.
  - Tools: Tableau, Power BI, and D3.js.



4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Techniques

- **Descriptive Statistics:**
  - Summarizing data using measures such as mean, median, mode, and standard deviation.
- **Inferential Statistics:**
  - Making inferences about a population based on a sample (e.g., confidence intervals, hypothesis testing).
- **Predictive Analytics:**
  - Using historical data to predict future outcomes (e.g., regression analysis, time series forecasting).
- **Prescriptive Analytics:**
  - Recommending actions based on predictive models to optimize outcomes (e.g., optimization algorithms).

Tools

- **Software:**
  - Excel, R, Python (with libraries like pandas, NumPy, SciPy), SQL.
- **Visualization:**
  - Tools like Tableau, Power BI, Matplotlib, and Seaborn for creating charts and dashboards.

Tools and Technologies

- **Programming Languages:**
  - Python, R
- **Data Manipulation:**
  - Pandas, NumPy
- **Data Visualization:**
  - Matplotlib, Seaborn, Plotly
- **Machine Learning:**
  - Scikit-learn, TensorFlow, Keras
- **Big Data Technologies:**
  - Hadoop, Spark
- **Databases:**
  - SQL, NoSQL
-

Model Deployment and Monitoring

- **Content:**
  - Deployment Strategies (Batch, Real-time)
  - Tools (Docker, Kubernetes, MLflow)
  - Monitoring and Maintenance
  - A/B Testing and Feedback Loops

Data Science Pre-Requisites

Machine Learning

Modeling

Statistics

Programming

Databases

Project Phase

Discover / Acquisition

Prepare

Plan

Model

Operationalize

Communicate Results

Data Analytics

Step 1: Determine the criteria for grouping the data

Step 2: Collecting the data

Step 3: Organizing the data

Step 4: Cleaning the data

Step 5: Analyze and Derive Insights

5 COMPONENTS OF DATA SCIENCE

Supervised Un-supervised

Structured Un-structured

Machine Learning

Statistics & Probability

Data

Reinforcement

Big Data

Programming Languages

Python

R

Java

NoSQL

Volume

Variety

Velacity

Value

Variability

Data Science Process

- Problem Definition
- Data Collection
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Model Building
- Model Evaluation
- Model Deployment

Advanced Data Science Process

- **Data Collection:**
  - Gathering data from various sources, such as databases, web scraping, and IoT devices.
- **Data Cleaning:**
  - Preprocessing data to handle missing values, outliers, and inconsistencies.
- **Data Transformation:**
  - Converting data into a format suitable for analysis, such as normalization and aggregation.
- **Exploratory Data Analysis (EDA):**
  - Using statistical methods and visualization tools to understand the data's structure and relationships.
- **Model Building:**
  - Applying machine learning algorithms to develop predictive models.
- **Model Evaluation:**
  - Assessing the performance of models using metrics such as accuracy, precision, recall, and F1-score.
- **Model Deployment:**
  - Implementing models into production environments for real-time predictions and decision-making.
- **Monitoring and Maintenance:**
  - Continuously monitoring model performance and making necessary updates.

Applications of Advanced Data Science

- **Healthcare:**
  - Predictive modeling for patient outcomes, personalized medicine, and medical image analysis.
- **Finance:**
  - Fraud detection, algorithmic trading, and risk management.
- **Marketing:**
  - Customer segmentation, sentiment analysis, and targeted advertising.
- **Manufacturing:**
  - Predictive maintenance, quality control, and supply chain optimization.
- **Retail:**
  - Inventory management, recommendation systems, and sales forecasting.

Skills and Tools for Advanced Data Science

- **Programming Languages:**
  - Python, R, and SQL.
- **Machine Learning Libraries:**
  - Scikit-learn, TensorFlow, and PyTorch.
- **Big Data Tools:**
  - Apache Hadoop, Spark, and Kafka.
- **Data Visualization Tools:**
  - Tableau, Power BI, and Matplotlib.
- **Cloud Platforms:**
  - AWS, Google Cloud, and Microsoft Azure.

Future Trends in Advanced Data Science

- **AutoML:**
  - Automation of machine learning tasks, making it more accessible to non-experts.
- **Explainable AI (XAI):**
  - Techniques to make AI models more transparent and interpretable.
- **Federated Learning:**
  - Training models across decentralized data sources without sharing raw data.
- **Edge Computing:**
  - Processing data closer to where it is generated for real-time analytics.

Conclusion

- Advanced data science is a powerful discipline that combines various techniques and tools to analyze complex data and derive actionable insights.
- It plays a crucial role in driving innovation and efficiency across different industries.
- By understanding the key components and processes involved, one can harness the full potential of data to make informed decisions and solve complex problems.