Introduction to Data Science

Er. Prem Chandra Roy
Er. Dhiraj Pyakurel

What is Data Science?

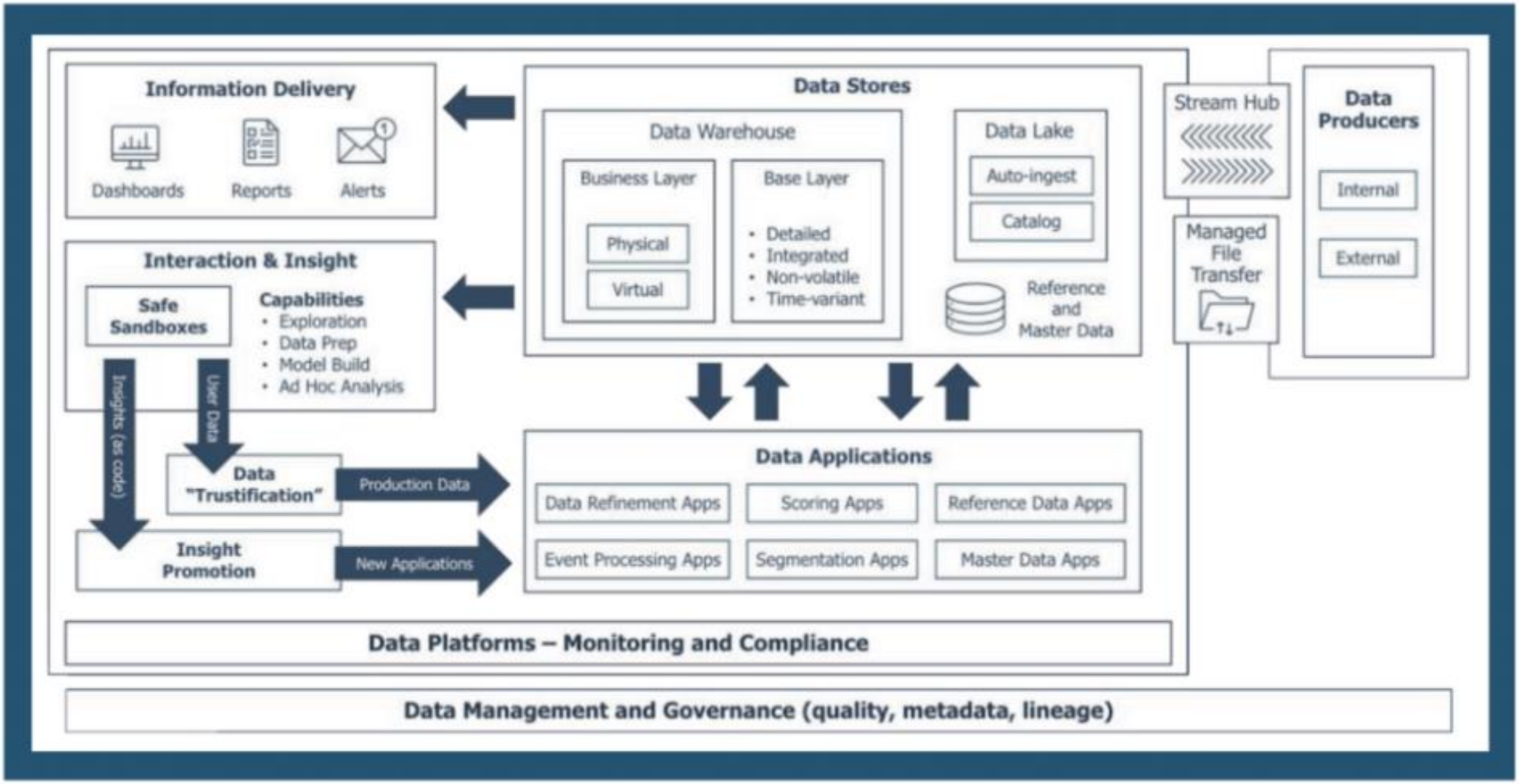
- Data science is the study of data to extract meaningful insights for decision-making.
  - Combines statistics, computer science, and domain knowledge.
    - Involves data collection, analysis, and interpretation.
  - The goal of data science is to turn raw data into meaningful information that can be used to make decisions, solve problems, and forecast trends.

Jargons of Data Science

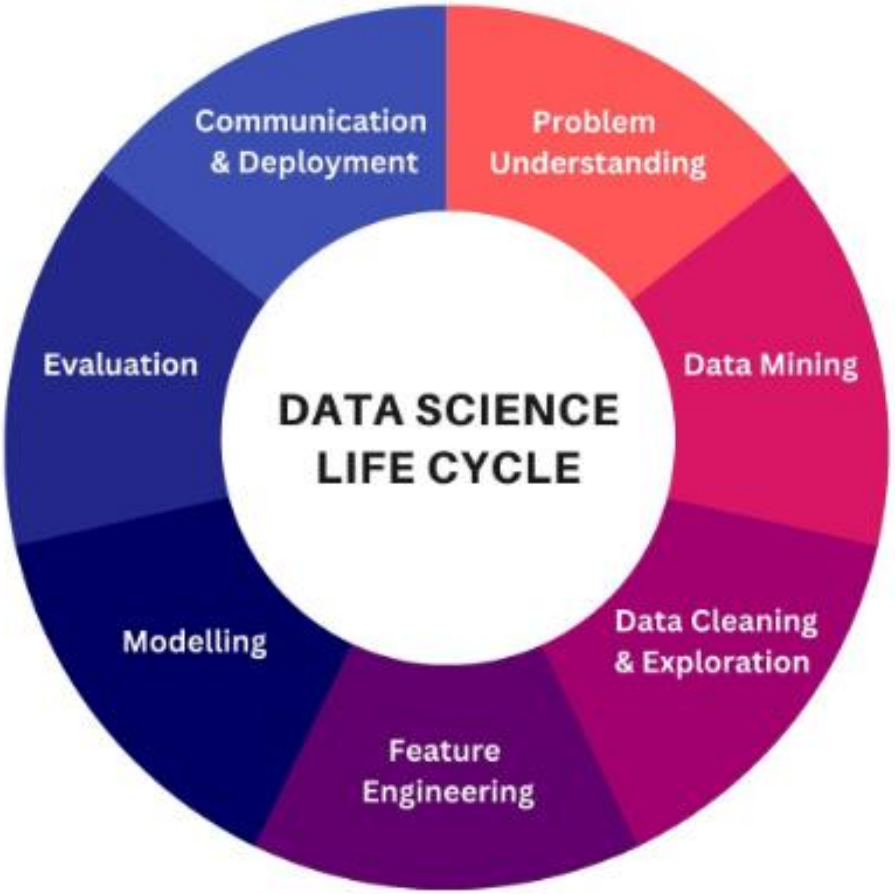
- **Big Data**
- **Machine Learning (ML)**
- **Artificial Intelligence (AI)**
- **Data Mining**
- **Predictive Analytics**
- **Data Wrangling (Data Munging)**
- **Feature Engineering**
- **Deep Learning**
- **Supervised Learning**
- **Unsupervised Learning**
- **Neural Network**
- **Natural Language Processing (NLP)**
- **Clustering**
- **Classification**
- **Regression**
- **Dimensionality Reduction**
- **ETL (Extract, Transform, Load)**
- **A/B Testing**
- **Hyperparameter Tuning**
- **Confusion Matrix and performance analysis**
- **Overfitting**
- **Underfitting**
- **Cross-Validation**
- **Bias-Variance Tradeoff**
- **Data Lake**
- **Data Warehouse**
- **SQL (Structured Query Language)**
- **Data Pipeline**
- **Data Governance**
- **NoSQL**
- **Anomaly Detection**
- **Time Series Analysis**
- **Reinforcement Learning**
- **Bagging and Boosting**
- **Gradient Descent**
- **Random Forest**
- **Principal Component Analysis (PCA)**
- **Data Imputation**
- **Data Augmentation**
- **Data Normalization**
- **Data Standardization**
- **Text Mining**

Etc…….

Modern Data Eco-System



Data Science Lifecycle



Data Science Process

- Stages:

  1. Data Collection: Gathering raw data.
  2. Data Cleaning: Preparing data by handling missing values and errors.
  3. Data Exploration: Initial insights through descriptive statistics and visualizations.
  4. Modeling: Applying algorithms to make predictions or classifications.
  5. Interpretation: Evaluating results to make actionable insights.

Data Science: Trends and markets

- Artificial Intelligence (AI) Integration
- Automated Machine Learning (AutoML)
- Explainable AI (XAI)
- Big Data Technologies
- Real-time Data Processing
- Data Privacy and Ethics
- Augmented Analytics
- Natural Language Processing (NLP) Advancements
- DataOps and MLOps
- Edge Computing
- Collaborative and Citizen Data Science
- Data Storytelling
- Cloud-based Data Platforms
- Integration of IoT and Data Science

Applications of Data Science

- Healthcare
- Finance
- Retail
- Transportation and Logistics
- Manufacturing
- Telecommunications
- Marketing and Advertising
- Sports Analytics
- Energy Sector
- Government and Public Sector
- Education
- Insurance
- Real Estate
- Social Media
- Human Resources etc…

Key Skills in Data Science

- Technical Skills:
  - - Programming (Python, R).
    - Machine learning and statistical analysis.
    - Data wrangling and visualization.
- Soft Skills:
  - - Problem-solving.
    - Communication and storytelling with data.

Tools and technologies in data science

Programming Languages
- Python
- R
- SQL
- Java
- Julia

Data Visualization Tools
- Tableau
- Power BI
- Matplotlib
- Seaborn
- ggplot2

Machine Learning Frameworks
- Scikit-learn
- TensorFlow
- Keras
- PyTorch
- XGBoost

Data Manipulation and Analysis
- Pandas
- NumPy
- Dplyr
- Tidyverse

Big Data Technologies
- Apache Hadoop
- Apache Spark
- Apache Kafka
- NoSQL Databases (e.g., MongoDB, Cassandra)

Data Integration and ETL Tools
- Apache NiFi
- Talend
- Informatica
- Apache Airflow

Tools and technologies in data science

Cloud Platforms
- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- IBM Cloud
- Microsoft Azure

Data Storage Solutions
- Data Lakes (e.g., AWS S3)
- Data Warehouses (e.g., Google BigQuery, Snowflake)
- Relational Databases (e.g., MySQL, PostgreSQL)

Collaboration Tools
- Jupyter Notebooks
- R Markdown
- Google Colab
- GitHub

Deployment and Monitoring Tools
- Docker
- Kubernetes
- MLflow
- TensorBoard

APIs and Data Services
- REST APIs
- GraphQL
- Data as a Service (DaaS) providers

Tools and technologies in data science

Business Intelligence Tools
- Looker
- Sisense
- QlikView

Natural Language Processing (NLP) Libraries
- NLTK
- spaCy
- Transformers (Hugging Face)

Robust Data Governance Tools
- Collibra
- Alation
- Talend Data Fabric

Data Security Technologies
- Encryption Tools
- Data Masking Solutions

Data Scientist

- A data scientist is a professional who utilizes their expertise in statistics, mathematics, programming, and domain knowledge to analyze and interpret complex data sets.
- Their primary goal is to extract meaningful insights from data to inform business decisions, improve processes, and solve problems.

Data Scientist : Characteristics

Skill Set

- Statistical analysis
- Programming (Python, R, SQL)
- Machine learning
- Data visualization

Domain Knowledge

- Understanding of the specific industry they work in

Problem-Solving Skills

- Strong analytical and critical thinking abilities

Communication Skills

- Ability to present findings to non-technical stakeholders

Collaboration

- Works with cross-functional teams (data engineers, business analysts, etc.)

Curiosity and Continuous Learning

- Natural curiosity to explore data and seek innovative solutions

Ethical Considerations

- Awareness of data privacy and responsible use of algorithms

Data Scientist : Roles

Data Collection
- Gather data from various sources, including databases, APIs, and external datasets.

Data Cleaning and Preparation
- Clean and preprocess data to ensure its quality and usability for analysis.

Exploratory Data Analysis (EDA)
- Analyze and visualize data to identify patterns and trends.

Feature Engineering
- Create and select relevant features to improve model performance.

Model Development
- Build and train machine learning models to solve specific business problems.

Model Evaluation
- Assess model performance using appropriate metrics and validation techniques.

Model Deployment
- Implement models in production environments for real-world application.

Monitoring and Maintenance
- Continuously monitor model performance and update as needed.

Data Visualization
- Create visual representations of data and insights for stakeholder communication.

Collaboration
- Work with cross-functional teams to understand requirements and develop solutions.

Stakeholder Communication
- Present findings and recommendations to non-technical stakeholders.

Research and Development
- Stay updated on data science trends and explore new methodologies.

Business Acumen
- Understand business objectives to align data science initiatives with goals.

Ethics and Data Governance
- Ensure ethical use of data and adherence to privacy regulations.

Continuous Learning
- Engage in lifelong learning to enhance skills and knowledge.

Challenges in Data Science

- Common Challenges:
  - - Data privacy and ethical issues.
    - Handling unstructured or missing data.
    - Building interpretable and reliable models.

Future of Data Science

- Trends:
  - - Increased automation and AI integration.
    - Enhanced focus on data privacy and ethics.
    - Advancements in machine learning and deep learning.

Conclusion

- Recap: Data science transforms raw data into actionable insights across industries.
  - Closing Thought: The potential of data science is vast, with continuous advancements shaping the future.