SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT I – INTRODUCTION TO DATA SCIENCE

FACETS OF DATA

FACETS OF DATA

In data science and big data, you encounter **many different types of data**, and **each type often requires different tools, storage methods, processing techniques, and analysis strategies**..

1. Empathy

- Data scientists face difficulty handling diverse data formats. Organizations lack clarity on which tools suit each data type.
- Decision-makers need combined insights from multiple data sources.
- Traditional databases fail to process unstructured and streaming data. There is confusion in selecting analytics methods for different data facets.

Problem Statement: Facets of Data in Data Science

- Modern organisations collect data in multiple forms such as structured tables, unstructured text, images, audio, machine-generated logs, network graphs, and real-time streams.
- Each facet of data differs in format, velocity, volume, and analytical requirements.
- However, many organisations struggle to integrate and analyse these diverse data types using a unified approach.
- This leads to fragmented insights, inefficient data utilisation, and poor decision-making.
- There is a need for a systematic framework to understand, manage, and analyse different facets of data effectively within the data science lifecycle.

FACETS of Data in Data Science

- Structured Data
- Unstructured Data
- Natural Language Data
- Machine-Generated Data
- Graph-Based / Network Data
- Audio, Video, and Image Data
- Streaming Data

1. Structured Data

- Data that *fits a predefined model* and has a predictable format (e.g., rows and columns in tables).

- Easy to store, query, and analyze using SQL and relational databases.

- Examples: spreadsheets, transaction records.

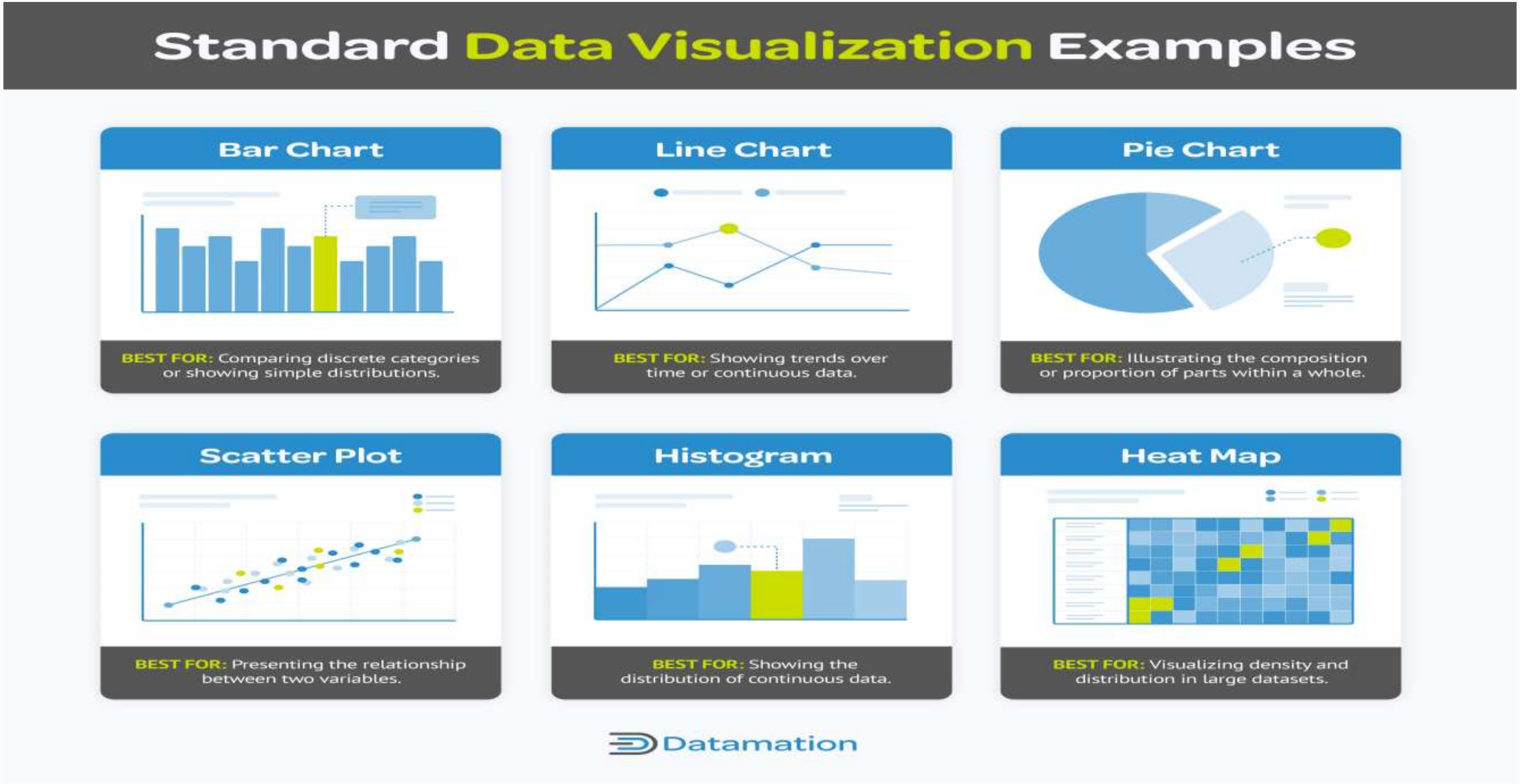
2. Unstructured Data

- Data that *does not conform to a fixed schema* and is not easily stored in traditional databases.
- Often text-heavy or context-specific, such as emails, documents, logs.
- Requires specialized tools and techniques for meaningful analysis.

2. Define

Defined Problem:

How can data science effectively handle and analyse different facets of data—such as structured, unstructured, machine-generated, graph-based, multimedia, and streaming data—to produce meaningful and actionable insights?



Standard Data Visualization Examples

Bar Chart

BEST FOR: Comparing discrete categories or showing simple distributions.

Line Chart

BEST FOR: Showing trends over time or continuous data.

Pie Chart

BEST FOR: Illustrating the composition or proportion of parts within a whole.

Scatter Plot

BEST FOR: Presenting the relationship between two variables.

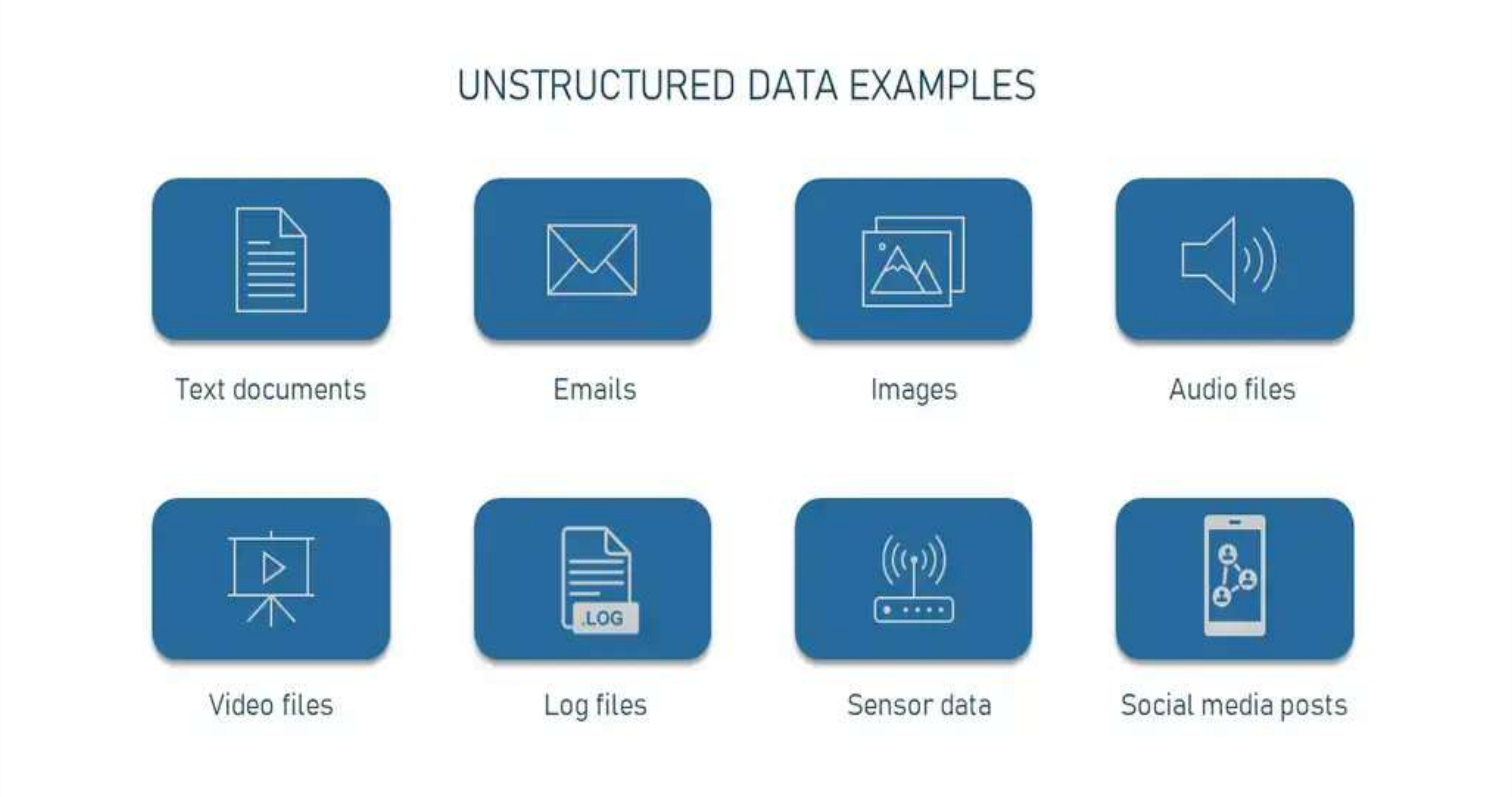
Histogram

BEST FOR: Showing the distribution of continuous data.

Heat Map

BEST FOR: Visualizing density and distribution in large datasets.

Datamation



UNSTRUCTURED DATA EXAMPLES

Text documents

Emails

Images

Audio files

Video files

Log files

Sensor data

Social media posts

3. Natural Language Data

- A *special type of unstructured data* that consists of human language (text or speech).
- Analysis often uses Natural Language Processing (NLP) techniques
- for tasks like sentiment analysis, entity recognition, summarization, etc.

4. Machine-Generated Data

- Automatically produced by systems, sensors, or machines.

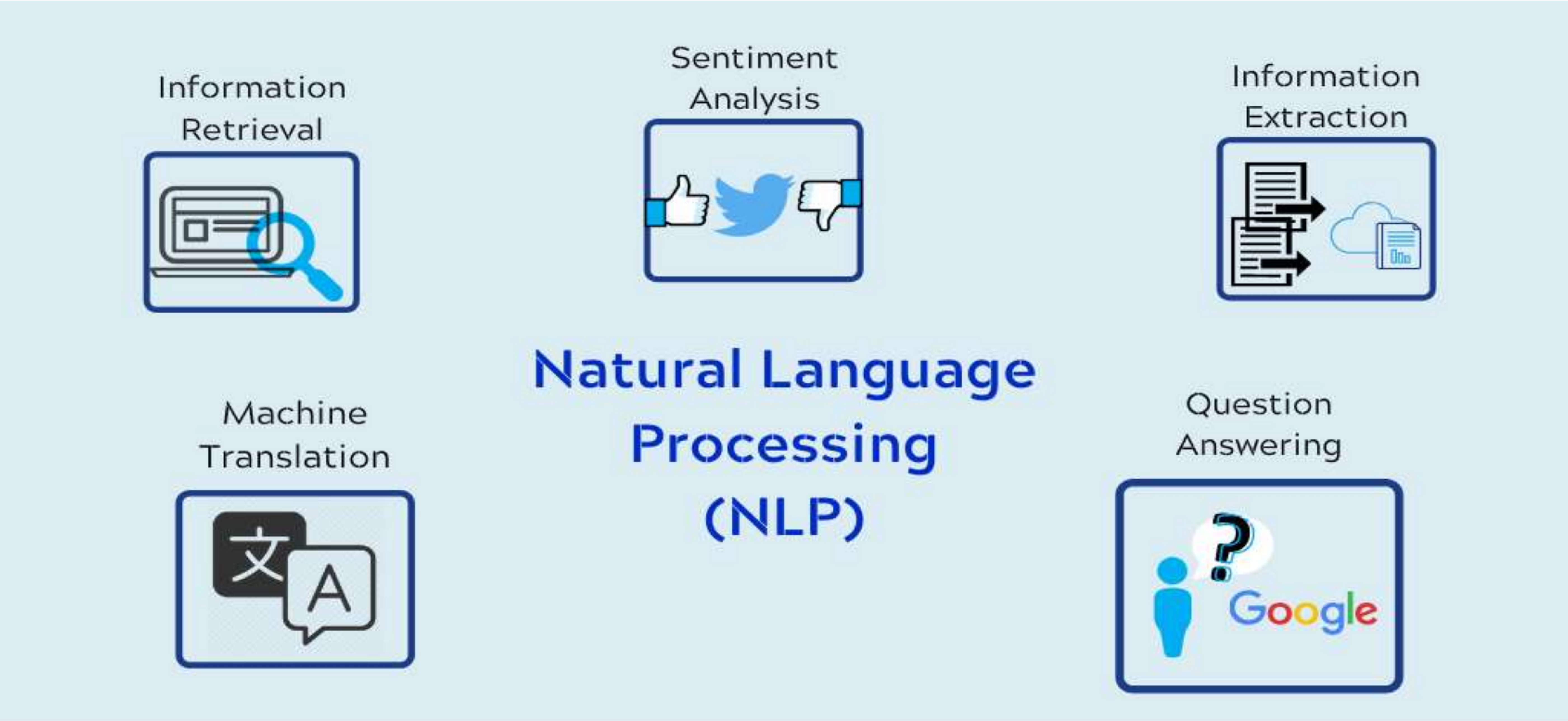
- Includes logs, telemetry from devices, clickstreams, and server logs.

- Often high-volume and streaming in nature.

3. Ideate

- Classify data based on structure, source, and behaviour. Apply suitable storage systems (SQL, NoSQL, graph databases).

- Use NLP for text data and computer vision for image/video data. Apply graph analytics for relationship-based data.Use stream processing frameworks for real-time data analysis.



Natural Language Processing (NLP)

Information Retrieval

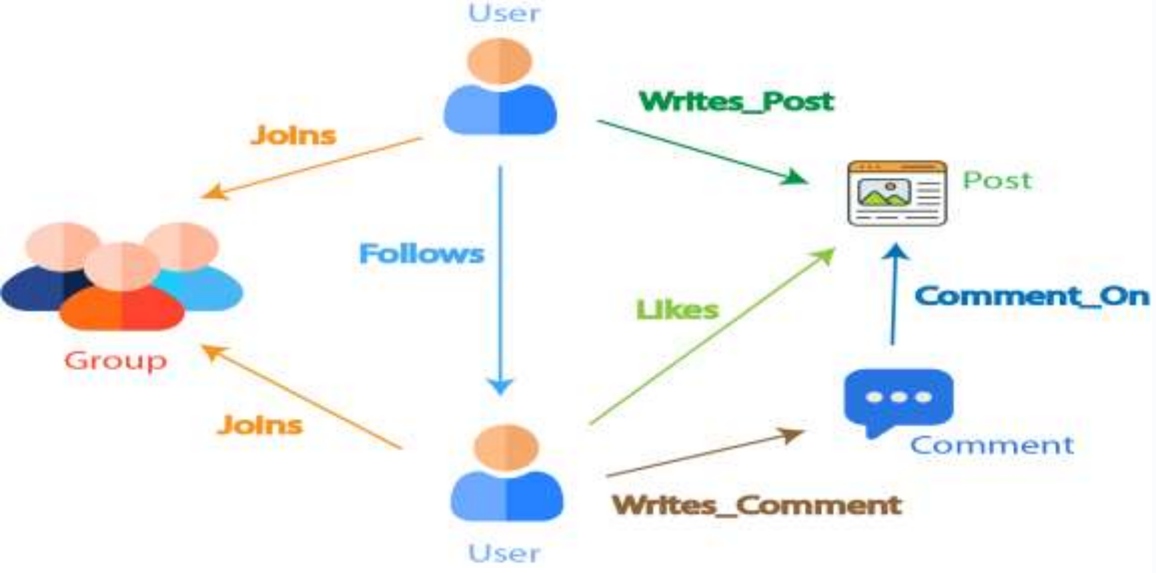
Sentiment Analysis

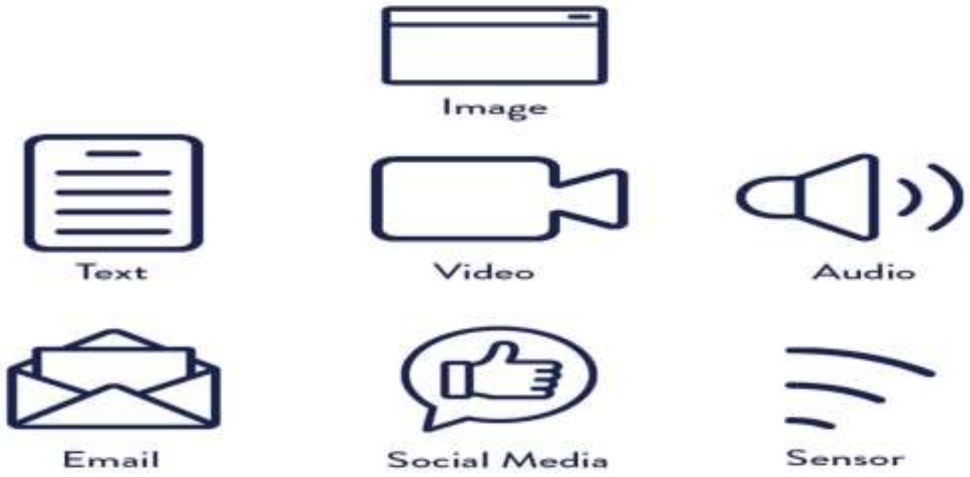
Information Extraction

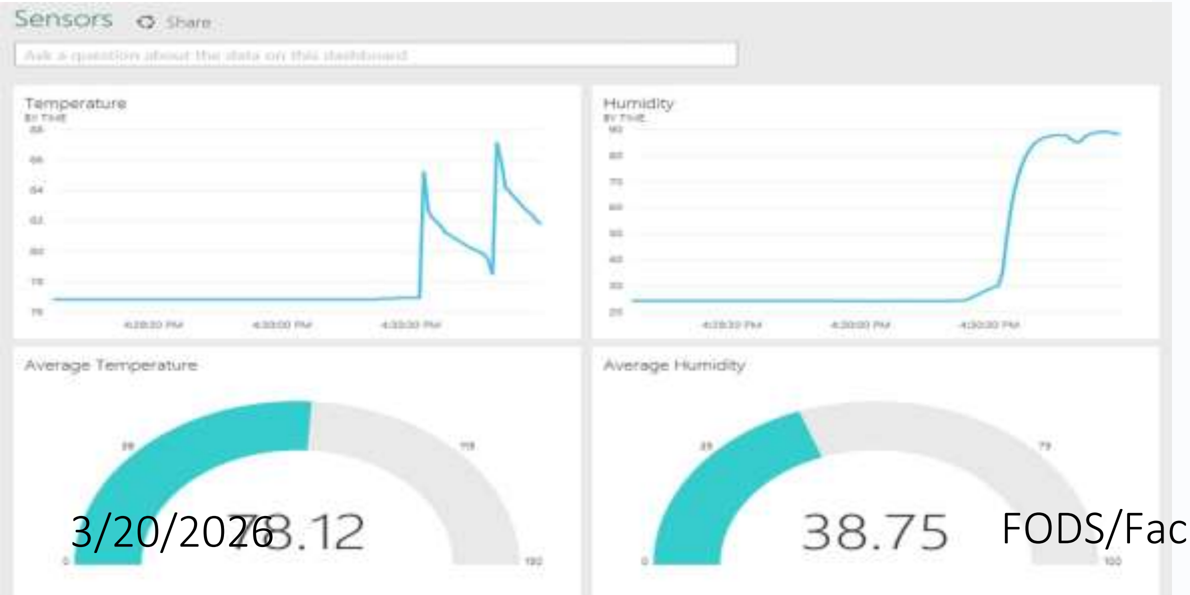
Machine Translation

Question Answering

Real-time Insights







5. Graph-Based / Network Data

- Data that represents *relationships* between entities (nodes and edges).
- Useful in social network analysis, recommendation systems, and network optimisation.

6. Audio, Video, and Image Data

- *Multimedia data*, which is unstructured and often requires computer vision, pattern recognition, or signal processing techniques for interpretation.

7. Streaming Data

- Data that *flows continuously in real time*, rather than being stored first and processed later.
- Examples include IoT sensor streams, live user interactions, and financial tick data.

4. Prototype

- Design a multi-facet data pipeline. Implement structured data analysis using relational databases.

- Build NLP models for text and language data. Develop dashboards for machine-generated and streaming data. Integrate graph visualisations for network-based insights.

Why These Facets Matter in Data Science

The **diversity of data formats and sources** directly impacts how data scientists approach analysis:

✓ Determines *data storage and management systems* (e.g., SQL vs NoSQL)

✓ Affects *data preprocessing and transformation strategies*

✓ Guides the selection of *analysis and machine learning techniques*

✓ Influences *tools and libraries* used (e.g., NLP vs image processing frameworks)

5. Testing

- Validate accuracy and performance for each data type.
- Test system scalability with increasing data volume. Evaluate real-time response for streaming data. Compare insights generated from individual vs integrated data facets.
- Collect user feedback and refine analytical models.

Expected Outcomes / Benefits

- Clear understanding of different facets of data. Improved data integration and analytics efficiency. Better tool and technique selection.

- Enhanced decision-making from diverse data sources. Scalable and flexible data science solutions