Unstructured System In the Acquisition & Structuring Of Data

Aditi Singh, Mayin Kajaria

INTRODUCTION

Understanding Data

• Structured Data: Organized, predefined format

(e.g., databases, spreadsheets)

• Unstructured Data:No predefined format (80% of enterprise data)

Why It Matters:

• 90% of digital data created in the last two years is unstructured

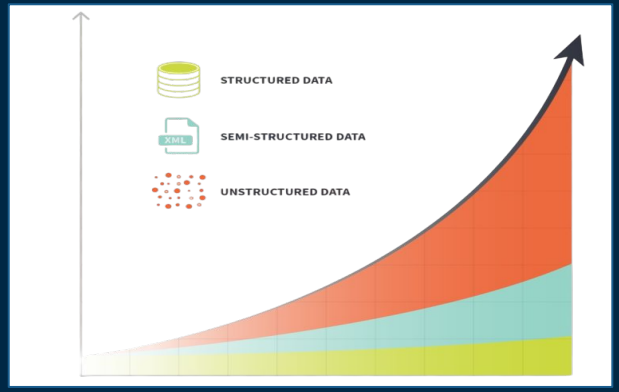
• Critical for modern decision-making and analytics

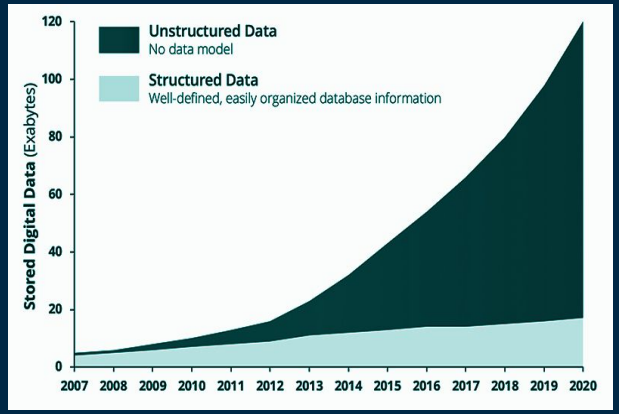
• Drives innovation in AI and machine learning

The Data Evolution:

• From simple spreadsheets to complex, multi-format data

• Rising importance of unstructured data analysis





WHAT IS UNSTRUCTURED DATA?

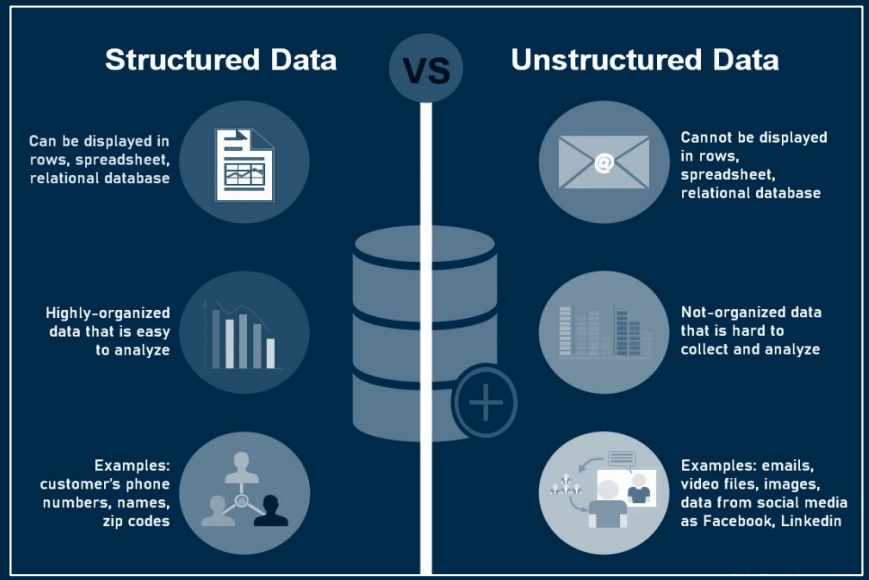
Definition:

"Data that doesn't conform to a specific, pre-defined data model"

Common Examples:

- Text: Emails, documents, social media posts
- Media: Images, videos, audio files
- Sensor Data: IoT devices, logs
- Scientific Data: Satellite imagery, weather data

Comparison with Structured Data:



KEY CHARACTERISTICS OF UNSTRUCTURED DATA

No Predefined Mode

-Lacks consistent organization
-Variable formats and structures

Heterogeneous Nature

-Multiple formats coexist
-Different sources and types

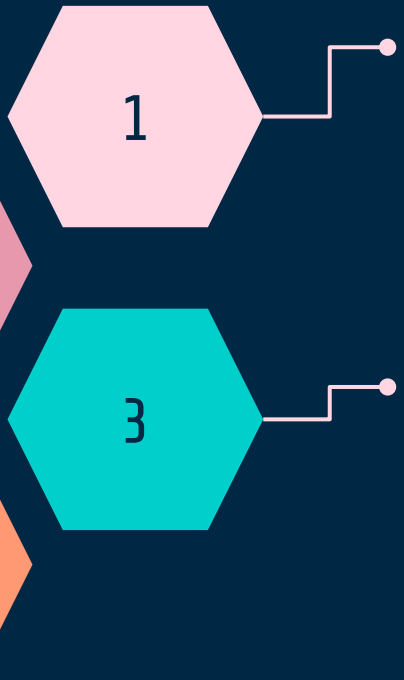
Scalability Challenges

-Exponential growth
-Storage intensive
-Processing complexity

Analysis Complexity

-Requires specialized tools
-Context-dependent interpretation

ROLE OF UNSTRUCTURED DATA IN DATA SCIENCE



Audio Processing

- Speech Recognition
- Music Classification
- Voice Authentication

Natural Language Processing (NLP)

- Sentiment Analysis
- Text Classification
- Language Translation

Impact on AI/ML

- Training Data for Deep Learning
- Realtime Decision Making
- Automated Content Understanding

Computer Vision

- Face Recognition
- Object Detection
- Medical Image Analysis

ACQUISITION OF UNSTRUCTURED DATA

What is acquisition of Unstructured Data?

The process of collecting unstructured data from various sources for further processing and analysis.

01

SOURCES

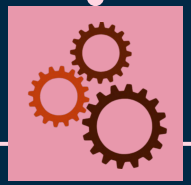
Emails, reports, PDFs, Images, videos, audio recordings, etc

02

COLLECTION

Manual Collection
Automated
Realtime
Streaming

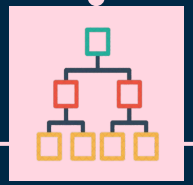
03



PRE-PROCESSING

Standardization
Handling unnecessary/ incomplete data

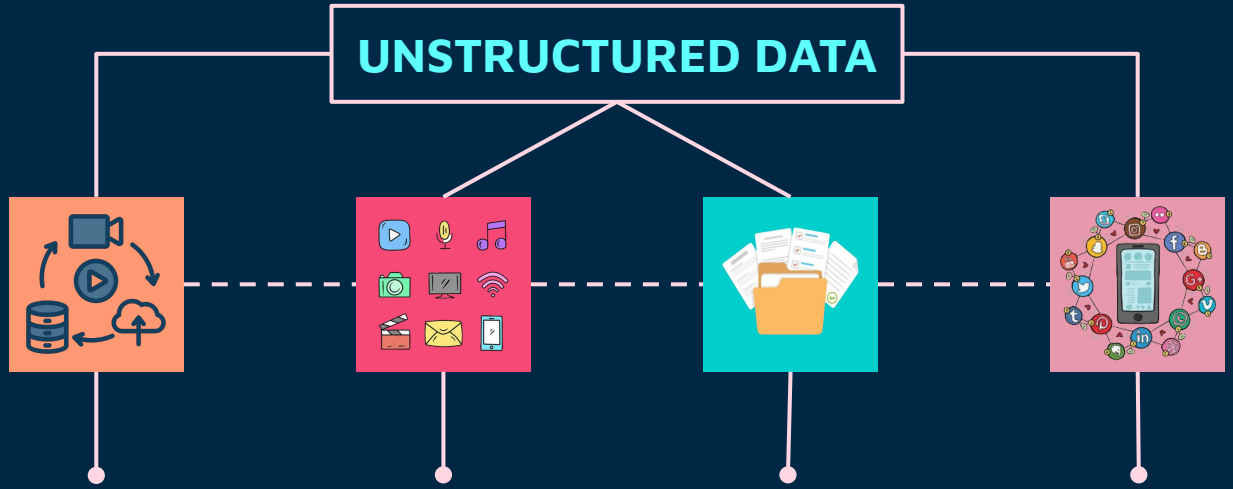
04



STRUCTURING

Storage and Organization of the data

SOURCES OF UNSTRUCTURED DATA



METHODS OF COLLECTION

Manual Collection 📝

- Involves human effort to gather data, such as downloading files, filling out surveys, or scanning documents.
- Example: A company manually collects customer feedback from survey responses.

Automated Extraction 🤖

- Uses technology to extract data from sources like websites, databases, and APIs.
- Example: A news aggregator uses web scraping to collect articles from multiple websites.

Real-Time Streaming 🌐

- Captures continuous data from sensors, social media, or transaction logs in real-time.
- Example: Smartwatches continuously collect and send health data like heart rate.

PRE-PROCESSING OF TEXTUAL DATA

Text Preprocessing Steps:

1. Tokenization
2. Stopword Removal
3. Stemming/Lemmatization

Example:
Sentence: “I am running and jumping in the park with my friends”

Tokens: Tokens: ['I', 'am', 'running', 'and', 'jumping', 'in', 'the', 'park', 'with', 'my', 'friends']

Tokens after stopword removal: ['I', 'running', 'jumping', 'park', 'friends']

Tokens after stemming: ['I', 'run', 'jump', 'park', 'friend']

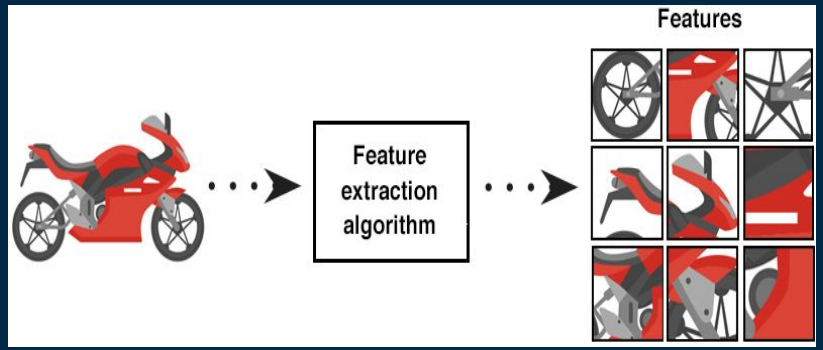
```python
from nltk import word_tokenize, stem
text = "Processing multiple words"
tokens = word_tokenize(text.lower())
stemmer = stem.PorterStemmer()
stems = [stemmer.stem(token) for token in tokens]
```

```
Tokens: ['processing', 'multiple', 'words']
Stems: ['process', 'multipl', 'word']
```

PRE-PROCESSING OF IMAGE/VIDEO DATA

Image Processing Pipeline

- Loading & Resizing
- Normalization
- Feature Extraction



Example of feature extraction

Classification Example (Using OpenCV):

```python
import cv2
import numpy as np

img = cv2.imread('image.jpg')
resized = cv2.resize(img, (224, 224))
normalized = resized/255.0  # Normalization
```

Output:

```
Original Image Shape: (1024, 768, 3)
Resized Image Shape: (224, 224, 3)
Normalized Image Shape: (224, 224, 3)
Normalized Pixel Range: Min = 0.0 Max = 1.0
```

STRUCTURING UNSTRUCTURED DATA

Challenges Faced in Structuring Data

- **Data Variety**
  Unstructured data comes in many forms (text, image, video, etc.), making it hard to apply a consistent structure.
- **Data Size**
  The volume of unstructured data is massive, which makes processing and structuring challenging.
- **Noise & Redundancy**
  Unstructured data often contains irrelevant information or duplication, requiring careful cleaning and filtering.
- **Tools & Techniques** Specialized tools like Natural Language Processing (NLP) and image recognition help in structuring, but they come with their own challenges.

TEXT DATA STRUCTURING

Natural Language Processing (NLP): NLP techniques like tokenization, part-of-speech tagging, and named entity recognition help break down and categorize text data.

Sentiment Analysis: Analyzing the sentiment of text (positive, negative, neutral) to derive meaning and structure from reviews, social media, etc.

Topic Modeling: Using algorithms like LDA (Latent Dirichlet Allocation) to discover topics within a set of unstructured text data.
Text Classification: Categorizing text data into predefined categories (e.g., classifying emails into spam or non-spam).

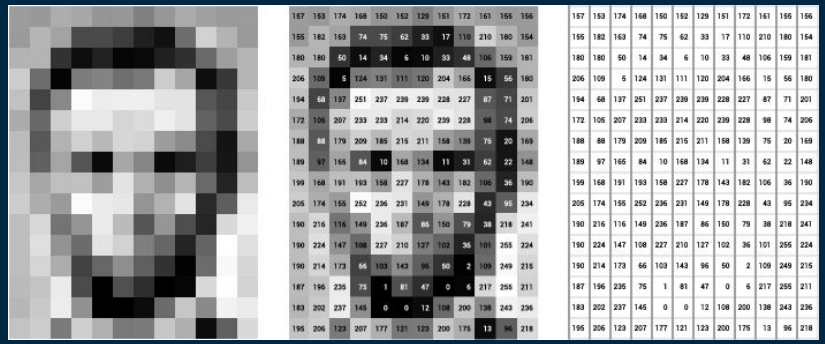
IMAGE DATA STRUCTURING

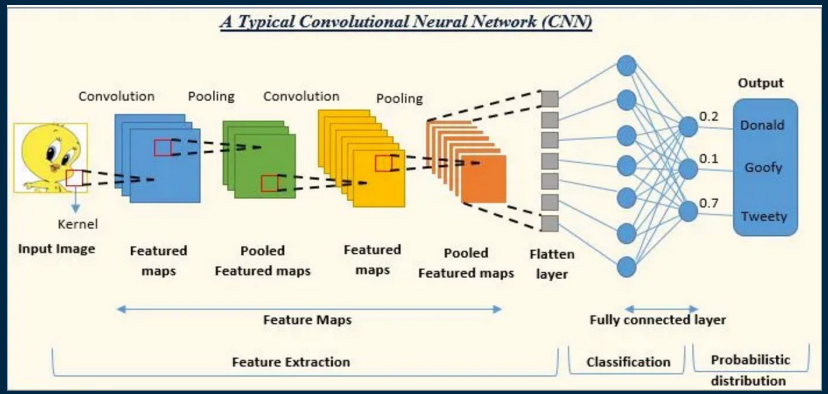
Representation: Image data is stored as a 2D-Array of values.

Image Recognition: Identifies and labels objects in images using models like CNNs.

Image Segmentation: Divides an image into regions for analysis.

Feature Extraction: Identifies key visual features (edges, colors) to classify images.



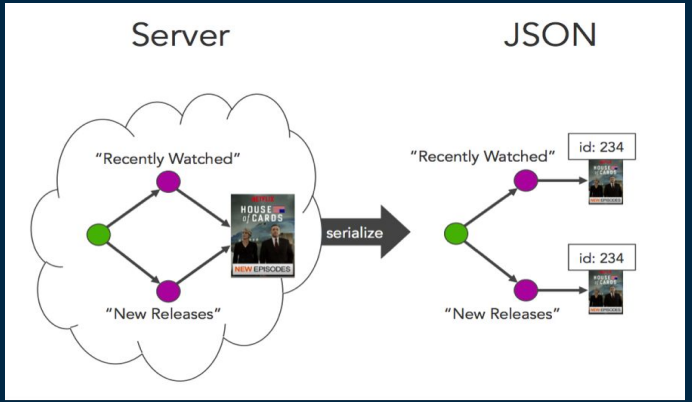


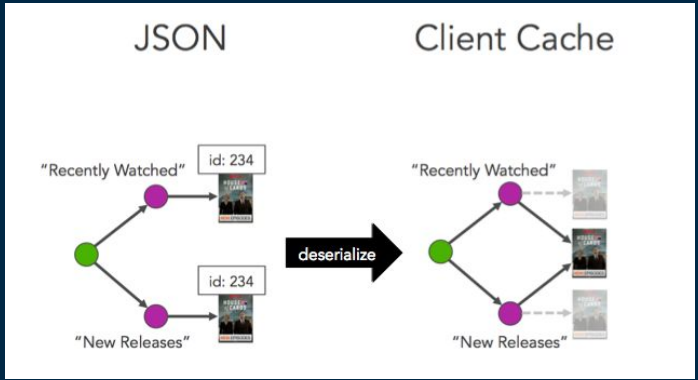
SEMI-STRUCTURED DATA AND ITS ROLE

What is Semi-Structured Data?: Data that has some structure but isn't strictly organized (e.g., JSON, XML files).

How Semi-Structured Data Helps: Easier to integrate with structured data, bridging the gap between unstructured and structured data systems.

Use Cases: APIs, log files, and user-generated content



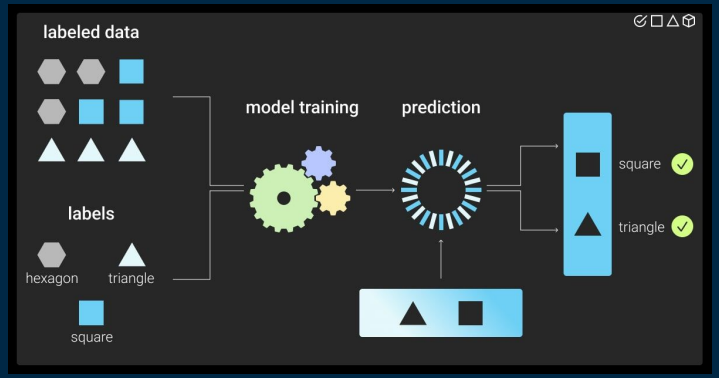


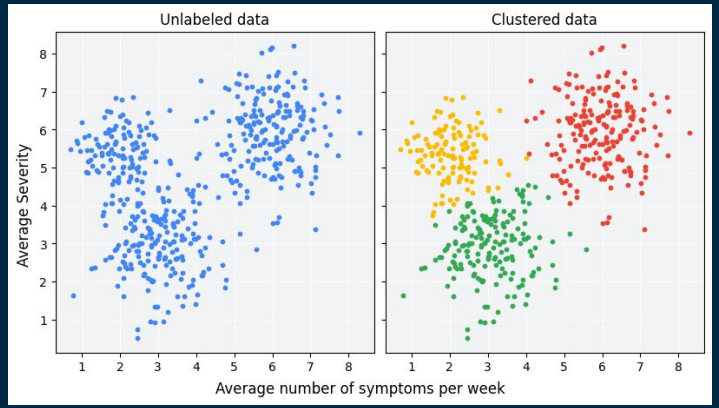
MACHINE LEARNING FOR STRUCTURING UNSTRUCTURED DATA

Supervised Learning: Training models on labeled datasets to categorize and organize data (e.g., classifying news articles by topic)

Unsupervised Learning: Using clustering and dimensionality reduction techniques to find hidden patterns in unstructured data.

Deep Learning: Advanced neural networks (e.g., CNNs and RNNs) used for tasks like image classification and text generation.



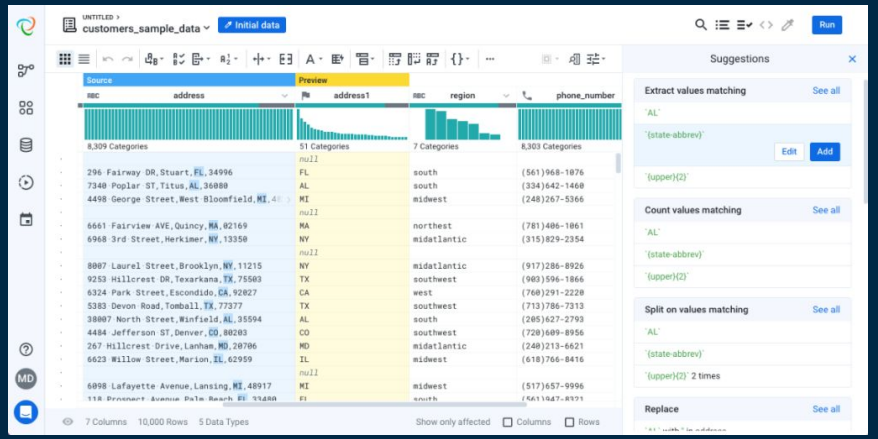


AUTOMATION IN STRUCTURING DATA

Data Wrangling Tools: Automated tools like Trifacta, Talend, and Alteryx clean, transform, and structure unstructured data.

AI & NLP-based Tools: Automation in structuring involves AI models to classify, tag, and organize large volumes of unstructured data.

Challenges: Requires high-quality training data and fine-tuning of algorithms to work effectively.



MCQs

Sentence: "She is walking and talking with her colleagues in the office."

1. After stopword removal, what will be the remaining tokens for the sentence:

A) ['She', 'is', 'walking', 'and', 'talking', 'with', 'her', 'colleagues', 'in', 'the', 'office']
B) ['She', 'walking', 'talking', 'colleagues', 'office']
C) ['She', 'is', 'walking', 'talking', 'colleagues', 'office']
D) ['She', 'walking', 'and', 'talking', 'colleagues', 'office']

Answer : B)

2. After stemming, what will be the transformed tokens?
Original Tokens (after stopword removal): ['She', 'walking', 'talking', 'colleagues', 'office']

A) ['She', 'walk', 'talk', 'colleagu', 'offic']
B) ['She', 'walking', 'talking', 'colleague', 'office']
C) ['She', 'walk', 'talk', 'colleague', 'office']
D) ['She', 'walk', 'talk', 'colleagues', 'office']

Answer : A)

CASE STUDY

Context: A retail company receives thousands of customer reviews daily in text form, which are unstructured.

Challenge: The company needs to identify key insights, detect sentiments, and categorize the feedback (positive, negative, suggestions).

Solution: Pay college freshers under 2Lakhs/year to manually go through the data and categorize it

Actual Solution: Using NLP for sentiment analysis, keyword extraction, and topic modeling. A machine learning model automatically tags feedback as positive, negative, or neutral and categorizes it into relevant topics (e.g., delivery, quality, customer service).

Outcome: Improved customer experience by prioritizing issues, identifying common complaints, and acting on suggestions.

MCQs

3. What percentage of today's enterprise data is estimated to be unstructured?

A) 30%
B) 50%
C) 80%
D) 90%

Answer : C)

4. Which of the following preprocessing techniques is commonly used to adjust the pixel values of an image to a standard range before feeding it into a machine learning model?

A) Edge Detection
B) Image Normalization
C) Object Detection
D) Feature Engineering

Answer : B)

MCQs

5. What is a key characteristic of semi-structured data?

A) It is strictly organized and follows a rigid format.
B) It has some structure but is not strictly organized.
C) It is completely unorganized and lacks any format.
D) It only exists in plain text format.

Answer : B)

6. Which of the following techniques is commonly used for identifying and labeling objects in images?

A) Image segmentation
B) Feature extraction
C) Convolutional Neural Networks (CNNs)
D) 2D-Array representation

Answer : C)

MCQs

7. What was the main challenge faced by the retail company in the case study?

A) Collecting customer reviews
B) Categorizing unstructured customer feedback
C) Analyzing sales data
D) Improving the website user interface

Answer : B)

8. What was the solution implemented to process customer feedback efficiently?

A) Using NLP for sentiment analysis and topic modeling.
B) Paying college freshers to manually categorize the reviews.
C) Relying on automated software for keyword extraction.
D) Collecting only positive reviews and ignoring others.

Answer : A)

THANK YOU