Ethnotech Academic Solutions Pvt Ltd ★

IT SPECIALIST

ARTIFICIAL INTELLIGENCE

Pearson

ARTIFICIAL INTELLIGENCE

Session Content

1. Data Collection, Processing and Engineering
2. Choose the way to collect data
3. Assess data quality
4. Ensure that data are representative
5. Identify resource requirements
6. Convert data into suitable formats
7. Select features for the AI model
8. Engage in feature engineering
9. Identify training and test data sets
10. Document data decisions

Data Collection, Processing and Engineering in AI

Data Collection, Processing, and Engineering are fundamental aspects of data-driven projects and involve various steps to transform raw data into meaningful insights.

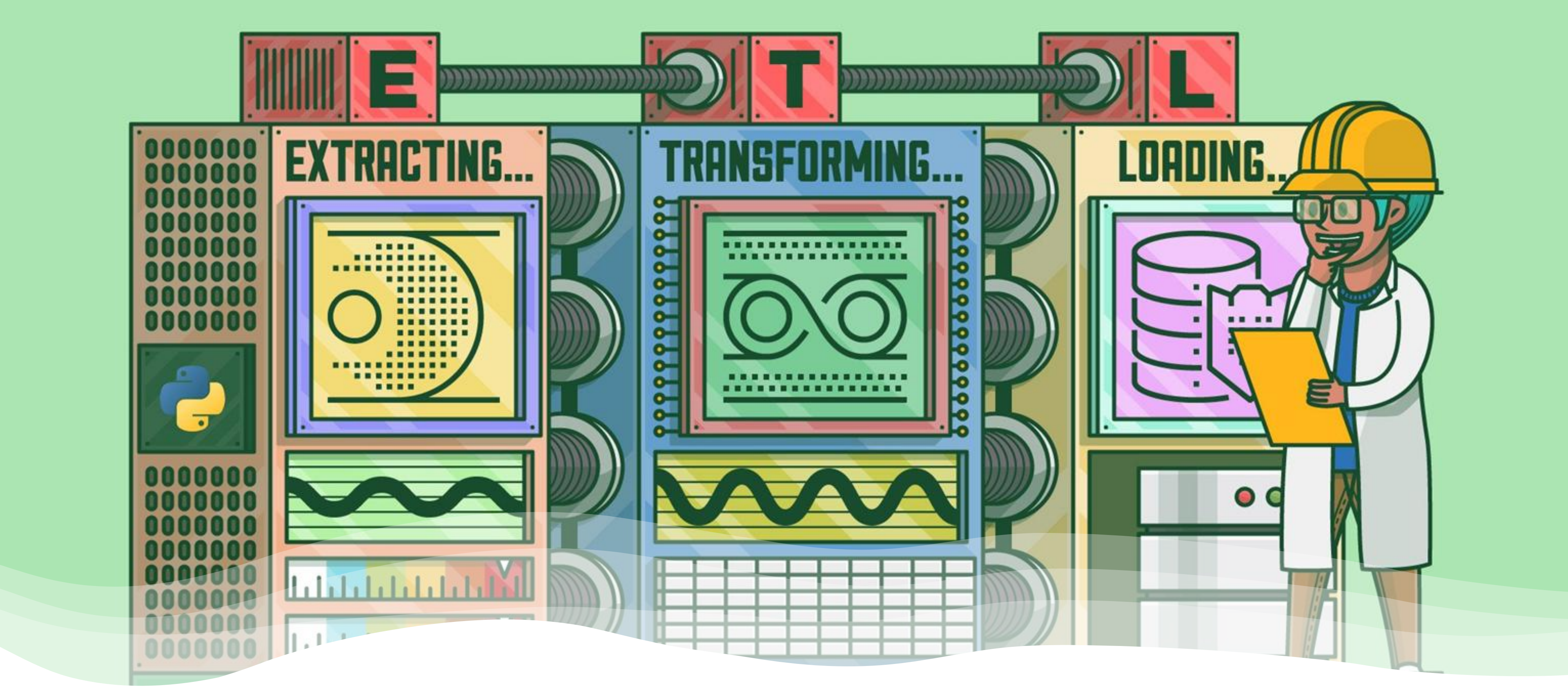
•Data Collection:

- Surveys
- Observations
- Sensors and IoT devices
- Web scraping
- Publicly available data
- Private data sources

Data Collection, Processing and Engineering in AI

Data Processing:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Aggregation
- Data Sampling
- Data Validation



Data Collection, Processing and Engineering in AI

Data Engineering:

- Data Storage
- Data Transformation and ETL
- Data Modeling
- Data Governance
- Workflow Orchestration
- Scalability and Performance

Choose the way to collect data

The choice of data collection method depends on various factors such as the nature of the data, the research objectives, available resources, and ethical considerations.

1. Surveys and Questionnaires

1. Observations

1. Interviews

1. Experiments

1. Existing Data

1. Sensor-based Data Collection

1. Web Scraping

Choose the way to collect data

Surveys and Questionnaires

A company wants to gather customer feedback on their new product.
Suitable Method: Surveys or questionnaires can be distributed to customers via email or through online survey platforms to collect their opinions, satisfaction levels, and suggestions.

Observations

A researcher wants to study the behavior of shoppers in a retail store.
Suitable Method: Observations can be conducted by physically observing and recording shopper behavior, interactions, and purchasing patterns within the store.

Choose the way to collect data

Interviews

A researcher wants to understand the experiences of individuals who have recently immigrated to a new country.
Suitable Method: Interviews can be conducted with immigrants to gather qualitative information about their challenges, opportunities, and cultural adaptation experiences. These interviews can be conducted face-to-face or through video conferencing.

Experiments

An organization wants to evaluate the usability of their website.
Suitable Method: Usability testing can be conducted by observing users interacting with the website and gathering their feedback and suggestions through a combination of direct observation, think-aloud protocols, and post-task interviews.

Choose the way to collect data

Existing Data

A researcher wants to investigate the impact of a new teaching method on student performance in a specific subject.
Suitable Method: Experiments can be conducted by dividing students into control and experimental groups, applying the new teaching method to the experimental group, and comparing their performance with the control group using pre and post-tests.

Sensor-based Data Collection

A government agency wants to analyze population demographics and socioeconomic indicators for policy planning.
Suitable Method: Utilizing existing data sources such as census data, government records, or public datasets can provide the required demographic and socioeconomic information without the need for primary data collection.

1. Web Scraping

Choose the way to collect data

Web Scraping

A researcher wants to monitor environmental conditions in a nature reserve.
Suitable Method: Sensor-based data collection can be implemented by deploying environmental sensors in different areas of the nature reserve to collect real-time data on temperature, humidity, air quality, and other relevant parameters.

Note:

The choice of data collection method should align with the research objectives, the target population, the type of data required, and the feasibility of the method within the available resources and ethical considerations.

Assess data quality

Assessing data quality is crucial to ensure the reliability and validity of the data being used for analysis and decision-making.

1. Accuracy
2. Completeness
3. Consistency
4. Validity
5. Reliability
6. Timeliness
7. Data Integrity
8. Data Documentation
9. Data Consistency with Business Rules
10. Data Relevance

Assess data quality || Example: Sales Data in an E-commerce Company

Accuracy: Cross-checking sales data with transaction records and financial statements to ensure they match.

Completeness: Identifying missing values or fields in sales records and investigating the reasons behind them.

Consistency: Comparing sales data from different sources, such as the website, point-of-sale system, and customer database, to ensure consistency.

Validity: Verifying that sales data falls within expected ranges, such as price limits or product categories.

Reliability: Assessing the consistency of sales data collection methods and training staff on proper data entry procedures.

Timeliness: Reviewing the frequency of data updates to ensure sales data is up-to-date.

Assess data quality || Example: Healthcare Patient Records

Accuracy: Verifying that patient records accurately reflect medical diagnoses, treatments, and procedures performed.

Completeness: Checking for missing information in patient records, such as allergies, medical history, or contact details.

Consistency: Examining consistency across different healthcare systems or departments to ensure patient information is accurately synchronized.

Validity: Assessing the adherence of patient records to medical coding standards, such as ICD-10 or SNOMED CT.

Reliability: Evaluating the reliability of data entry processes and implementing checks for errors or inconsistencies during data capture.

Data Integrity: Implementing access controls and encryption mechanisms to protect patient records from unauthorized access or tampering.

Assess data quality || Example: Customer Survey Data

Accuracy: Reviewing survey responses for any inconsistencies or contradictory answers.

Completeness: Checking for missing responses or unanswered questions in the survey data.

Consistency: Identifying patterns or discrepancies in survey responses across different respondents or survey waves.

Validity: Assessing the relevance and appropriateness of survey questions to ensure they capture the intended information.

Reliability: Examining the reliability of the survey administration process, including survey distribution, instructions, and data collection protocols.

Data Documentation: Ensuring that the survey data is properly documented, including information on the survey design, sampling methodology, and any data cleaning or preprocessing steps.

Ensure that data are representative

Ensuring that data is representative means that the collected data accurately reflects the characteristics and diversity of the population or phenomenon under study.

1. Define the Target Population
2. Sampling Method
3. Sample Size
4. Avoid Biases
5. Consider Demographic Factors
6. Geographical Considerations
7. Data Validation
8. Inclusion and Accessibility
9. Monitoring and Evaluation
10. Transparency in Reporting

Ensure that data || Example: Political Opinion Survey

Define the Target Population: Determine the specific demographic characteristics, such as age, gender, and geographic location, that represent the population of interest

Sampling Method: Use random sampling techniques to select participants from the target population, ensuring each individual has an equal chance of being selected.

Demographic Representation: Ensure the sample includes participants from various demographic groups, such as different age ranges, genders, ethnicities, and regions.

Inclusion and Accessibility: Make efforts to include underrepresented groups, such as minority communities or marginalized populations, to capture a diverse range of opinions.

Data Validation: Cross-validate the survey results with existing public opinion polls or census data to ensure the collected data aligns with known population statistics.

Ensure that data || Example: Market Research for a New Product

Define the Target Market: Identify the specific market segment or consumer group for the new product.

Sampling Method: Utilize stratified sampling, dividing the target market into relevant subgroups.

Sample Size: Determine an appropriate sample size that provides sufficient representation of the target market and allows for meaningful analysis.

Geographic Representation: Ensure the sample includes participants from different geographic locations within the target market to capture regional variations.

Product Usage Patterns: Consider participants with diverse usage patterns or behaviors related to similar products or competitors to understand the full market landscape.

Data Validation: Compare the collected data with market research reports or customer databases to validate the representativeness of the sample.

Identify resource requirements

Identifying resource requirements is essential for planning and executing data collection, processing, and engineering activities.

•**Human Resources**
- Data Collection
- Data Processing
- Data Engineering

•**Time**
- Data Collection
- Data Processing
- Data Engineering

•**Technology and Tools**
- Data Collection
- Data Processing
- Data Engineering

Identify resource requirements

- **Infrastructure**
  - Data Storage
  - Computing Resources
- **Financial Resources**
  - Budget Allocation
- **Data Privacy and Security**
- **Training and Skill Development**
- **Collaboration and Communication**

Convert data into suitable formats in AI

In AI, data formats often depend on the specific AI tasks and frameworks being used.

Image Data

Convert from: Various image formats

Conversion Method: Libraries such as OpenCV or PIL (Python Imaging Library) can be used to read and process image files.

Text Data

Convert from: Plain text, word documents, PDFs, or HTML files.

Conversion Method: Text data can be read and processed using programming languages like Python. Libraries like NLTK (Natural Language Toolkit) or spaCy can be used to handle text processing tasks.

Convert data into suitable formats in AI

Tabular Data

Convert from: Excel spreadsheets, CSV files, or databases.

Conversion Method: Libraries like pandas in Python provide functions to read and manipulate tabular data. They allow you to read data from various file formats, perform data cleaning and preprocessing, and save data in different formats if needed.

Audio Data

Convert from: Audio file formats (e.g., WAV, MP3, FLAC).

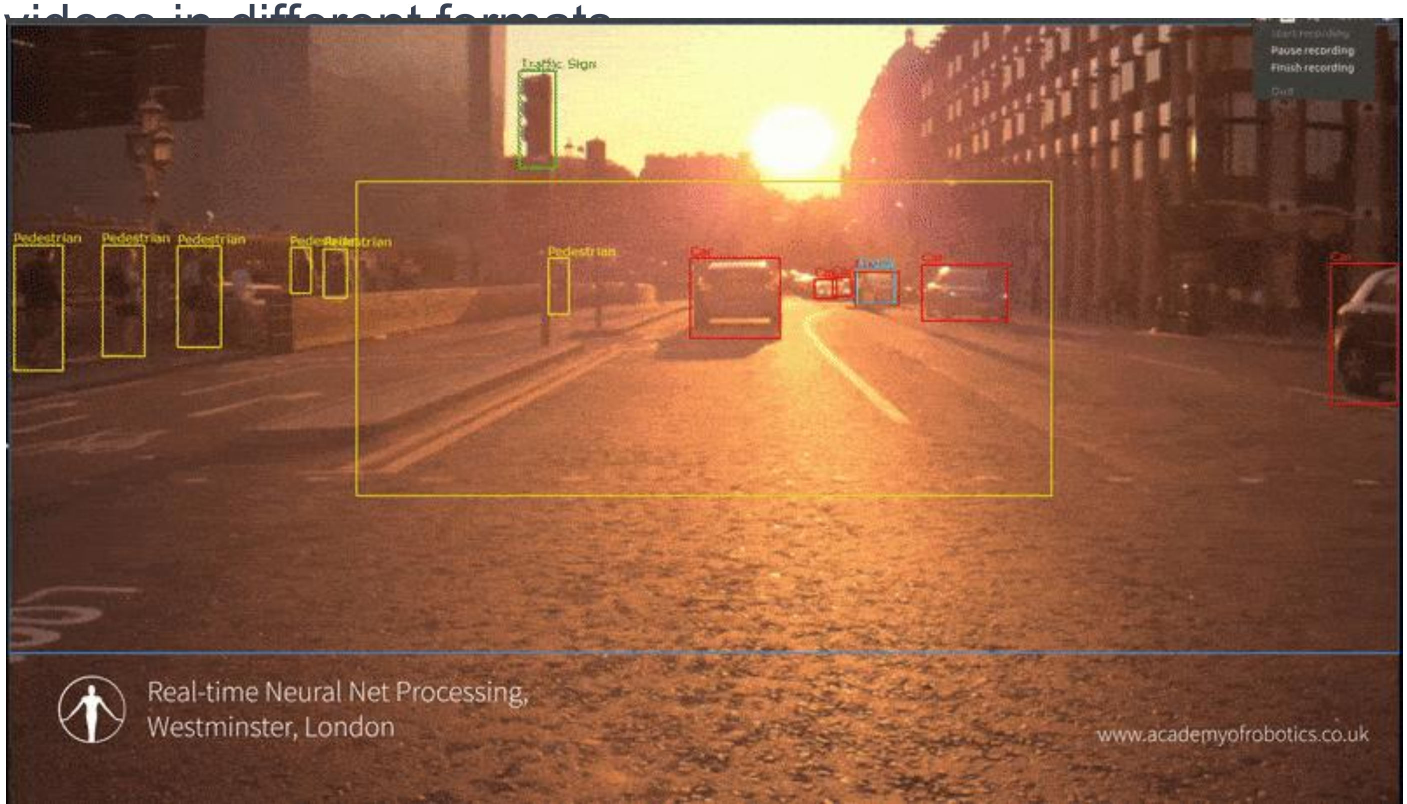
Conversion Method: Libraries like librosa in Python can be used to read and process audio files. They provide functions to load audio data, perform audio signal processing operations, and save audio files in different formats.

Convert data into suitable formats in AI

Video Data

Convert from: Video file formats (e.g., MP4, AVI, MOV).

Conversion Method: Libraries like OpenCV in Python can handle video data. They provide functions to read video files, extract frames, perform video processing tasks, and save videos in different formats.



Question 1:

Which two actions are performed during the data ingestion and data preparation stage of an Machine Learning process? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Calculate the accuracy of the model.
B. Score test data by using the model.
C. Combine multiple datasets.
D. Use the model for real-time predictions.
E. Remove records that have missing values.

Question 2:

You need to predict the animal population of an area.
Which Machine Learning type should you use?

A. regression
B. clustering
C. classification

Question 3:

Which two languages can you use to write custom code for Machine Learning designer? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. Python

B. R

C. C#

D. Scala

Question 4:

Your company wants to build a recycling machine for bottles. The recycling machine must automatically identify bottles of the correct shape and reject all other items.
Which type of AI workload should the company use?

A. anomaly detection
B. conversational AI
C. computer vision
D. natural language processing