Data Collection and Preparation

IPAI 2024

Master SDIA, University of Constantine

Pr. Layeb

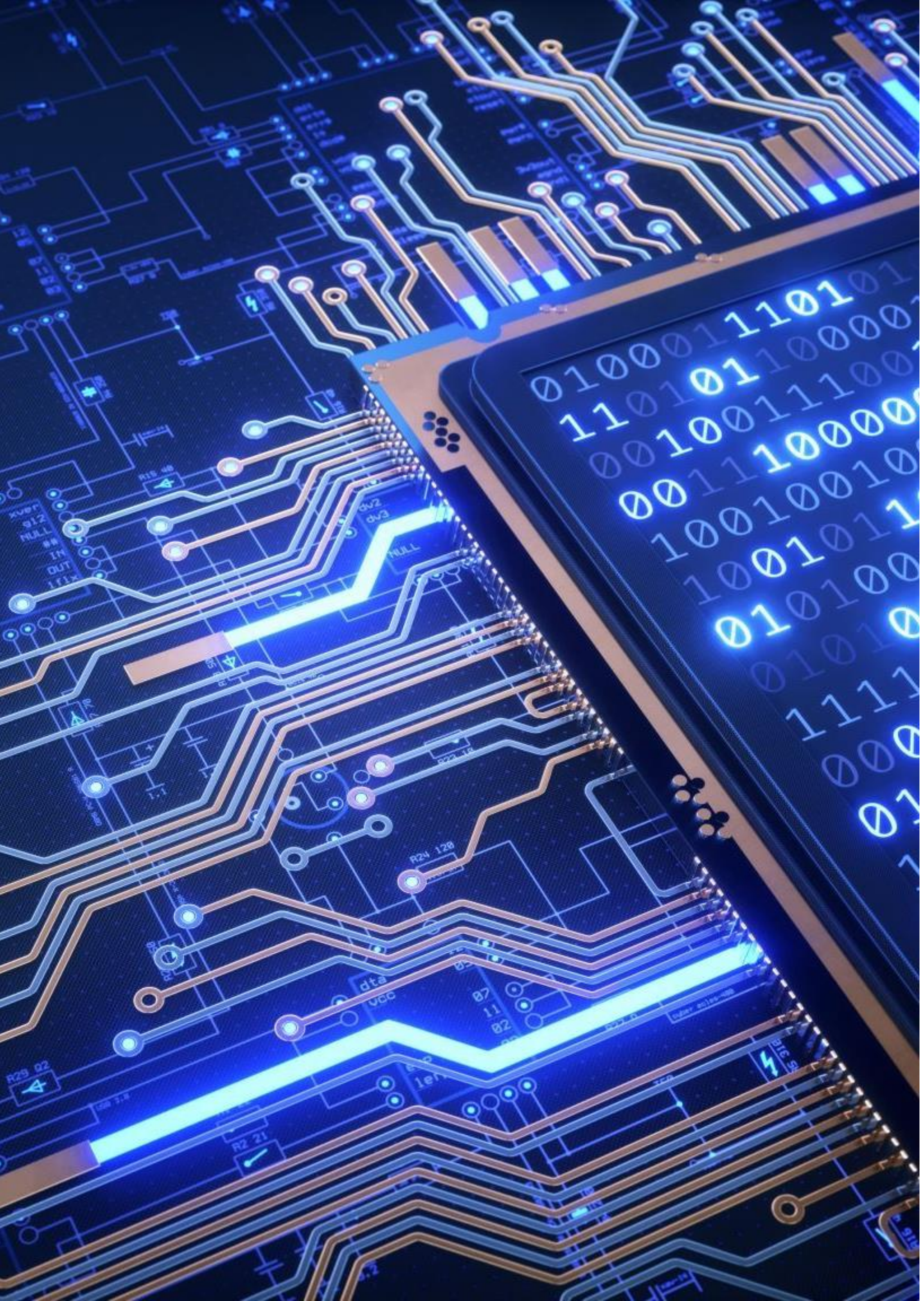
Introduction

- Data collection and preparation are crucial steps in the design thinking process for an AI project.
- These steps ensure that the data used to train and test the AI model is representative, relevant, and of high quality.

Types of data

Structured Data

- **Definition:**
  - Data organized in a fixed format with a clear schema.
  - Presented in tables with rows and columns.
- **Real-world Examples:**
  - Relational databases (e.g., SQL databases).
  - Spreadsheets (e.g., Excel).
  - CSV (Comma-Separated Values) files.
- **Implications for AI Model Development:**
  - Suitable for traditional machine learning models.
  - Efficient for querying and analysis using standard database operations.
  - Well-suited for scenarios with clearly defined and organized data.



Types of data

Unstructured Data

- **Definition:**
  - Data lacking a predefined structure.
  - Typically includes text, images, audio, and video.
- **Real-world Examples:**
  - Text documents (e.g., articles, social media ...).
  - Images and videos.
  - Audio recordings and speech data.
- **Implications for AI Model Development:**
  - Requires advanced techniques like natural language processing (NLP) and computer vision.
  - Challenges in organizing, indexing, and querying compared to structured data.
  - Valuable for applications such as sentiment analysis, image recognition, and speech-to-text.

Types of data **Semi-Structured Data**

- **Definition:**
  - Data that is not purely structured but contains some level of organization.
  - Typically includes tags, labels, or hierarchies.
- **Real-world Examples:**
  - JSON (JavaScript Object Notation) files.
  - XML (eXtensible Markup Language) documents.
  - NoSQL databases.
- **Implications for AI Model Development:**
  - Combines some benefits of both structured and unstructured data.
  - Requires flexible data processing techniques.
  - Commonly used in web applications, data interchange, and document storage.

Data Collection

a. Identify Data Sources:

Depending on your project, this could involve user surveys, interviews, existing company data, or publicly available datasets

b. Data Quality Assessment:

Evaluate the quality of available data. Assess factors such as completeness, accuracy, consistency, and reliability. Identify and address any issues with missing or erroneous data.

c. Legal and Ethical Considerations:

Ensure compliance with legal and ethical standards related to data collection. Consider privacy regulations, data ownership, and consent requirements.

Data Collection

d. Sampling Strategies:

Decide on appropriate sampling strategies based o n the dataset size and distribution. Random sampling, stratified sampling , or other techniques may be employed.

e. Data Acquisition:

Collect data from identified sources using suitable methods. This could involve web scraping, API requests, database queries, or manual data entry.

Data Collection Methods for AI Projects

1. User-Generated Data:

- **Surveys and Questionnaires:** Great for gathering a large amount of quantitative data from a broad audience. You can use online survey tools like Google Forms
- **Interviews:** Ideal for in-depth qualitative data collection. Interviews allow you to probe deeper into user experiences, motivations, and pain points. Conducting user interviews can be done in person, over video calls, or even asynchronously through text-based formats.

Data Collection Methods for AI Projects

- **2. Internal Data Collection:**
- **CRM Systems:** If your project focuses on customer interactions, data from your Customer Relationship Management system can be a goldmine. This might include customer demographics, purchase history, and support interactions.
- **Website Analytics:** Provides valuable insights into user behavior on your website or app. Tools like Google Analytics track user actions, clicks,
  and browsing patterns, helping you understand how users interact with your digital products.
- **Server Logs:** Record user activity and system events on your servers. While not always user-friendly data, server logs can reveal usage patterns, potential errors, and system performance metrics.
- **3. External Data Collection:**
- **Web Scraping (with permission):** Involves extracting data from websites. This
  can be useful for gathering publicly available information relevant to your project. However, it's crucial to obtain permission from website owners before scraping data and ensure compliance with terms of service.

Data Preparation:

- **a. Data Cleaning:**
- Cleanse the dataset to handle missing values, outliers, and inconsistencies. Impute missing data, correct errors, and ensure uniform formatting.
- **b. Data Transformation:**
- Transform data into a suitable format for analysis. This may involve normalization, standardization, or encoding categorical variables.
- **c. Feature Engineering:**
- Create new features or modify existing ones to enhance the model's performance. This step involves domain knowledge and understanding of the problem.
- **d. Data Splitting:**
- Divide the dataset into training, validation, and test sets. The training set is used to train the model, the validation set helps fine-tune it, and the test set assesses its performance on unseen data.

Data Preparation:

- **e. Handling Imbalanced Data:**
- Address imbalances in class distribution if applicable. Techniques like oversampling, undersampling, or using synthetic data can be employed.
- **f. Dealing with Categorical Data:**
- Convert categorical variables into a format suitable for machine learning models, such as one-hot encoding or label encoding.
- **g. Data Scaling:**
- Scale numerical features to ensure that they have a similar scale. Common techniques include Min-Max scaling or standardization.

Data Preparation:

- **h. Data Augmentation (for Image Data):**
- If working with image data, consider data augmentation techniques to increase the diversity of the training dataset. This includes random rotations, flips, or zooms.
- **i. Addressing Data Security:**
- Implement measures to ensure the security of sensitive data. This includes encryption, access controls, and adherence to data protection regulations.
- **j. Data Documentation:**
- Maintain comprehensive documentation describing the dataset, including variable descriptions, data sources, and any preprocessing steps performed. This is crucial for transparency and reproducibility.

Data collection and preparation :

- **Key Considerations:**
- **Iterative Process:**
  - Data collection and preparation are often iterative processes. As the project progresses, revisit these steps based on insights gained during model development.
- **Domain Expertise:**
  - Collaborate with domain experts to better understand the data and make informed decisions during the preparation phase.
- **Version Control:**
  - Implement version control for datasets to track changes and facilitate reproducibility.
- **Continuous Monitoring:**
  - Establish mechanisms for continuous monitoring of data quality and update procedures as needed.
- **Bias and Fairness:**
  - Be aware of biases in the data and take steps to address them, particularly when working with machine learning models that can perpetuate or exacerbate biases.