Introduction to Machine Learning

Zeham Management Technologies BootCamp
by SDAIA

July 28th, 2024

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Before we start in Machine learning

SDAIA

الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Let’s Practice

Notebook path :

1-Introduction to Machine Learning/

LAB/How to find Correlations in Data?ipynb

Dataset :

1-Introduction to Machine Learning/

LAB/Dataset/Expanded_data_with_more_features.csv

Introduction to ML

Let’s start together…

Agenda

Problem Definition

Data Collection

Data Preparation

Exploratory Data Analysis (EDA)

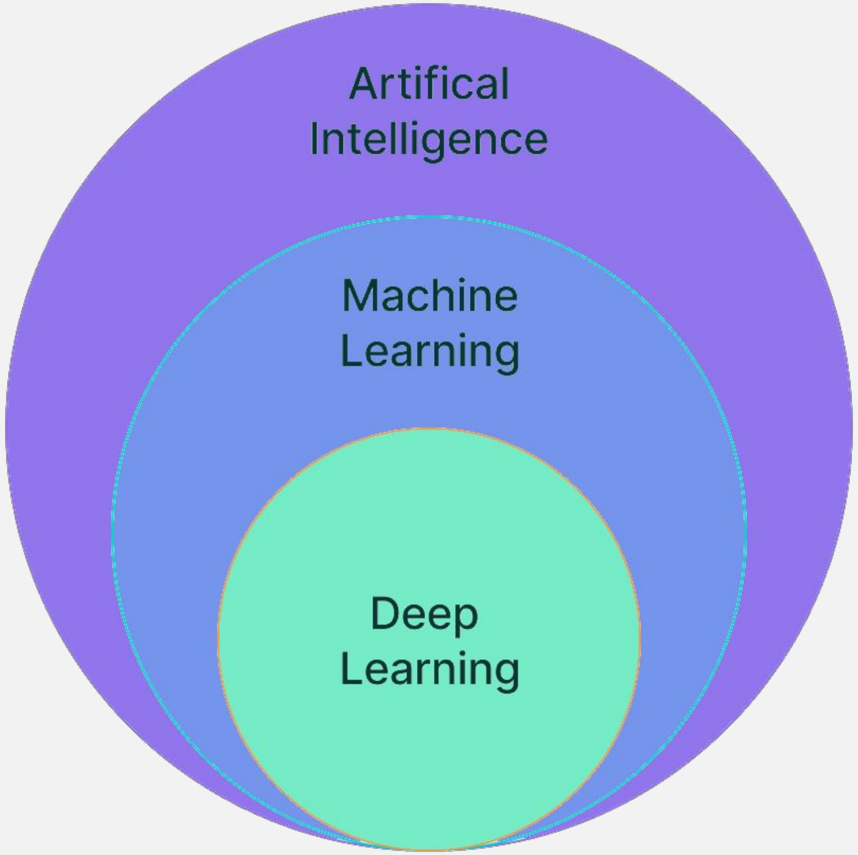
Problem Definition

Problem Definition

Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) technique, which uses statistical methods to enable . machines to improve with experience

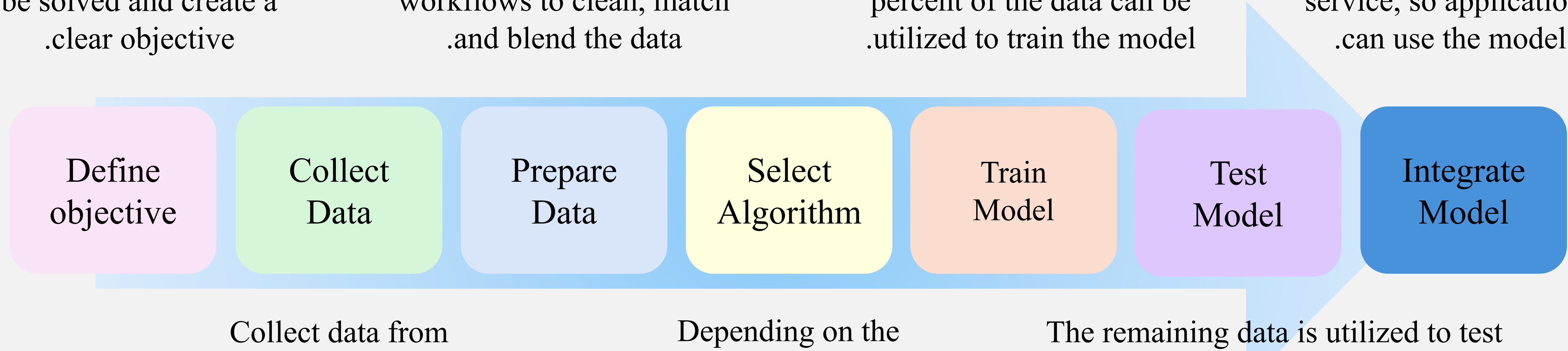
Machine Learning is a kind of Artificial Intelligence (AI) that provides computers the ability to learn without being . explicitly programmed



SOURCE

Problem Definition

How does Machine Learning work?



Define objective

Identify the problem to be solved and create a .clear objective

Collect Data

Collect data from hospitals, health insurance companies, social service agencies, .police and fire dept

Prepare Data

Preparing data is a crucial step and involves building workflows to clean, match .and blend the data

Select Algorithm

Depending on the problem to be solved and the type of data, an appropriate algorithm .will be chosen

Train Model

Data is fed as input and with the required parameters. A percent of the data can be .utilized to train the model

Test Model

The remaining data is utilized to test the model for accuracy . Depending on the results, improvements can be performed in the 'Train model' and/or .'Select Algorithm' phases, iteratively

Integrate Model

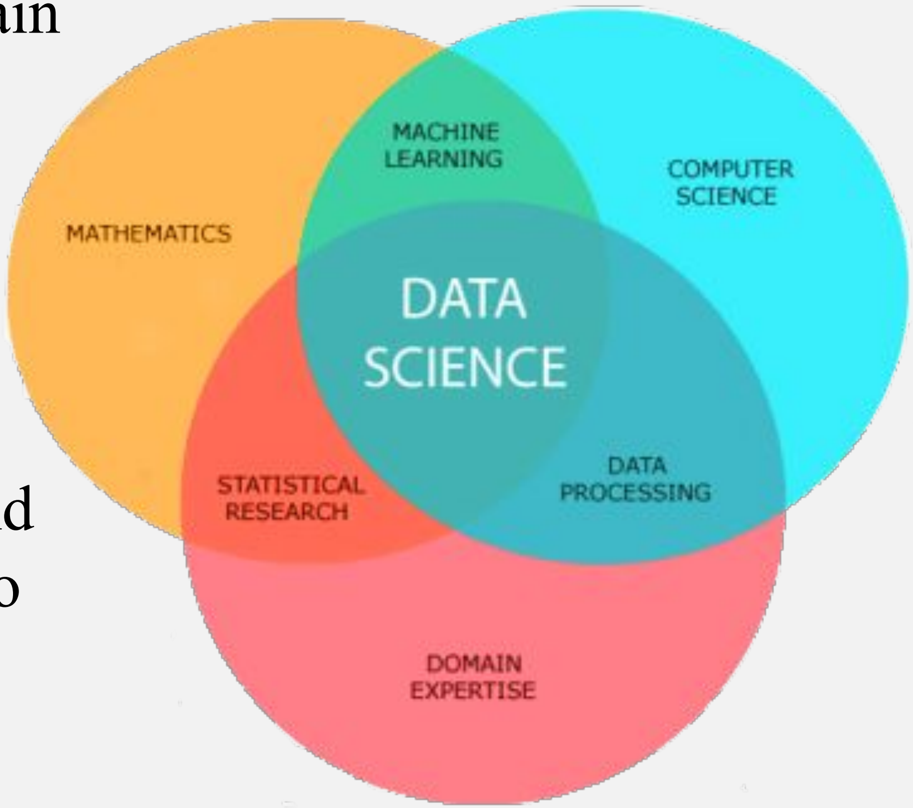
Publish the prepared experiment as a web service, so applications .can use the model

Problem Definition

What is data science ?

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful .insights from data

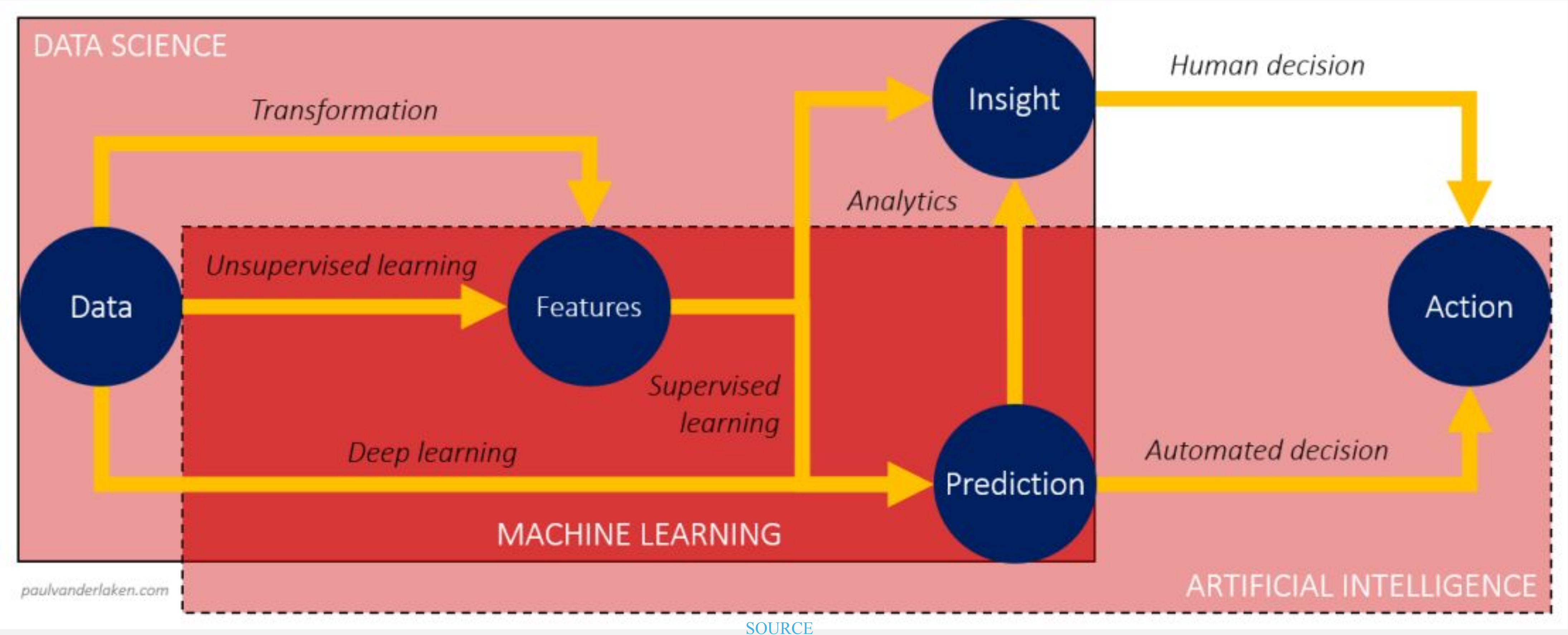
Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into .tangible business value



Problem Definition

What is data science ?

Data Science VS AI



SOURCE

Problem Definition

Why use Machine Learning ?

: Machine Learning is great for

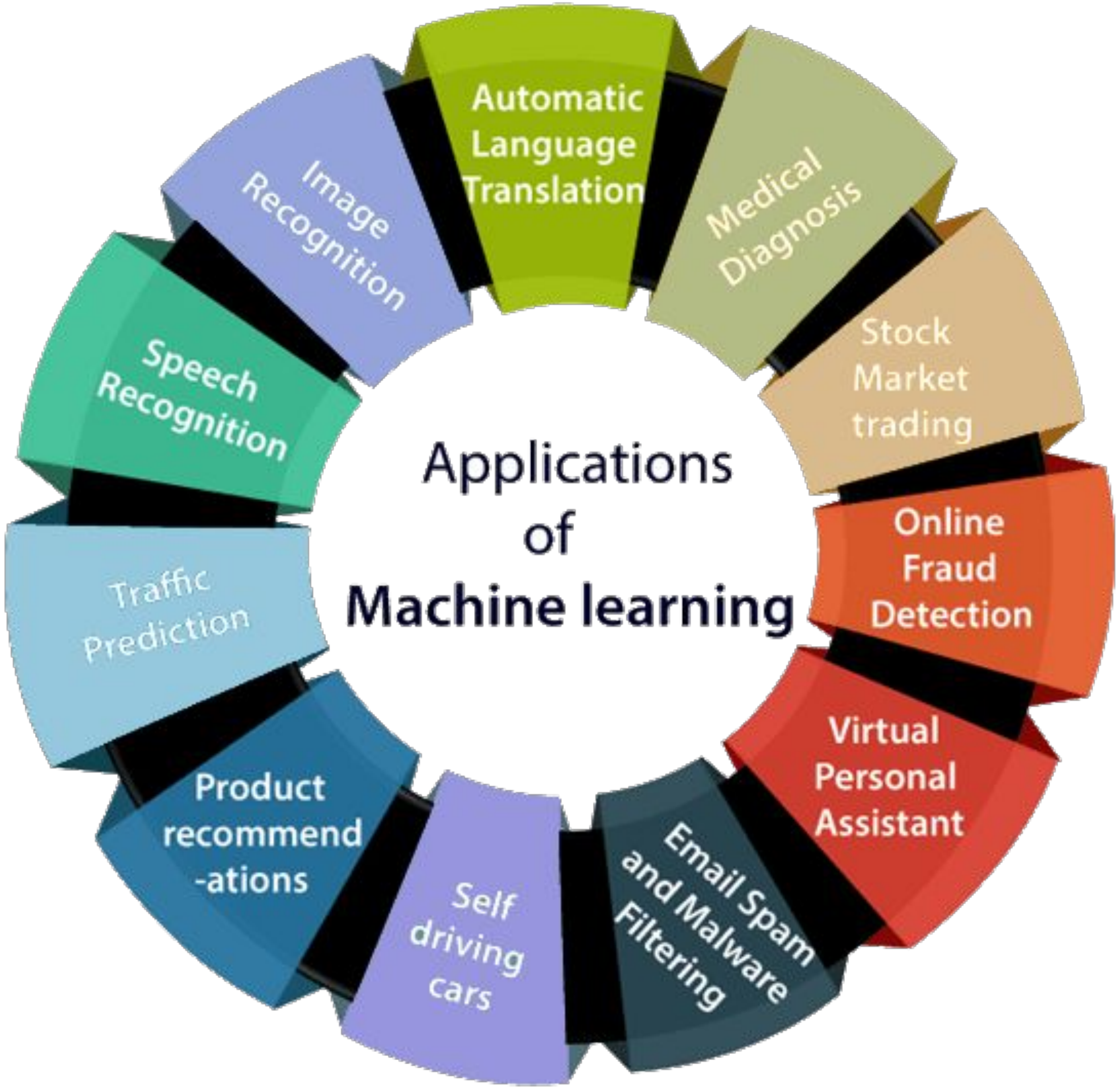
Problems for which existing solutions require a lot of hand-tuning or long lists of rules –

Complex problems (no good solution at all using a traditional approach) –

Fluctuating environments (adapt to new data) –

.Getting insights about complex problems and large amounts of data –

Machine Learning Applications



Data Collection

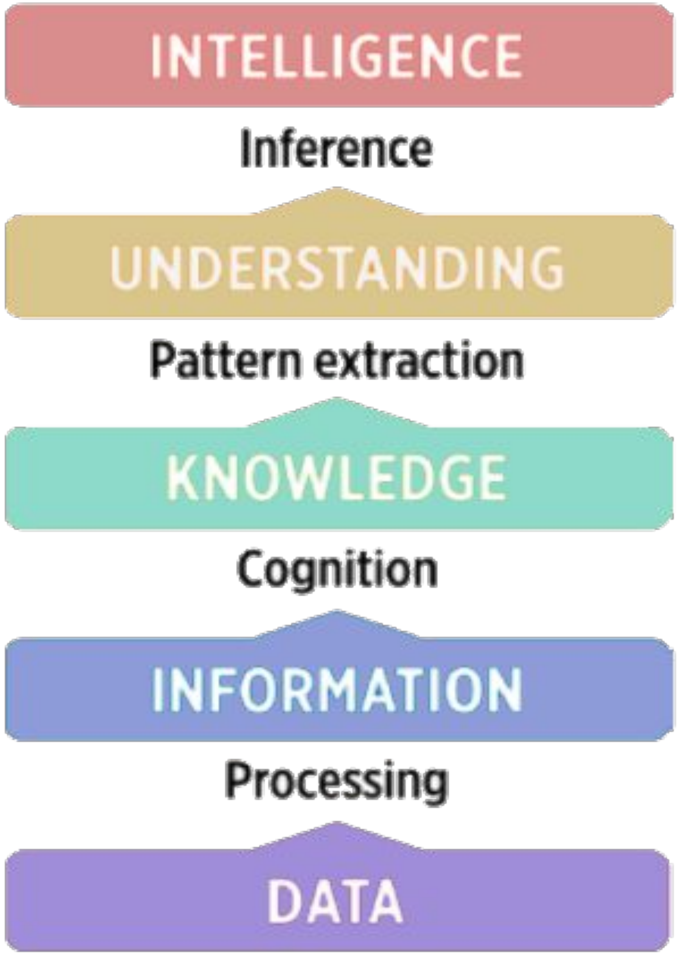
Data Collection

What is Data?

- Data is a collection of raw, unorganized facts and details like text, observations, figures, symbols and descriptions of things etc. In other words, data does not carry any specific purpose and has no significance by itself.

What is Information?

- Information is processed, organized and structured data. It provides context for data and enables decision making. For example, a single customer's sale at a restaurant is data – this becomes information when the business is able to identify the most popular or least popular dish



Data Collection

In machine learning, converting data to information is essential for extracting useful patterns .and knowledge for decision-making, predictions, and understanding phenomena
:Here are several reasons why this conversion is important

- **Understanding Patterns and Trends** : Raw data often contains hidden patterns and trends that are not immediately apparent. By converting data into information, machine learning algorithms can identify these patterns and provide valuable insights.

- **Making Informed Decisions** : Information derived from data enables better decision-making. For instance, businesses can use information from customer data to improve their products, optimize marketing strategies, and enhance customer satisfaction.

- **Improving Predictions** : In predictive modelling, the quality and relevance of the information extracted from data directly impact the accuracy of predictions. Well-processed information helps in building more robust and reliable models.

- **Enhancing Efficiency** : Information is more manageable and interpretable than raw data. This makes it easier to process, analyse, and visualize, leading to more efficient use of computational resources and time.

Data Collection

- **Facilitating Communication :** Information is easier to communicate to stakeholders than raw data. It provides a clear and concise understanding of what the data represents and its implications.

- **Supporting Learning Algorithms :** Machine learning algorithms require structured and meaningful information to learn effectively. Converting raw data into features that capture the essence of the data helps in training better models.

- **Identifying Outliers and Anomalies :** Converting data to information helps in identifying outliers and anomalies, which can be critical for tasks such as fraud detection, quality control, and system monitoring.

- **Data Reduction :** Often, raw data contains redundant or irrelevant information. Data processing and transformation help in reducing the data to a more compact form that still retains the essential information, making analysis more efficient.

Data Collection

data collection is the process of gathering, measuring, and analysing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and .probabilities

The goal of data collection in data science is to amass data that is **relevant**, .**accurate**, and of **high quality**

Data Collection

The data are usually divided into two types:

- ***Structured***
- ***Unstructured.***

The simplest example of structured data would be a .xls or .csv file where every column stands for an attribute of the data

Unstructured data could be represented by a set of text files, photos, or video files. For example, if the task is to build a system that could detect pneumonia from an image of the lungs, you need specialized equipment to create a catalog of digital images

Where can you “borrow” a dataset? Here are a couple of data sources you could try:

OPEN DATA PLATFORM

منصـــة البيـــانات المفتـــوحة

SOURCE

Data Collection

BY SDAIA

SOURCE

Data Collection

Kaggle Website

SOURCE

Data Collection

Dataset search by Google

SOURCE

Data Collection

Hugging Face Website

SOURCE

Data Collection

OpenML

SOURCE

Data Collection

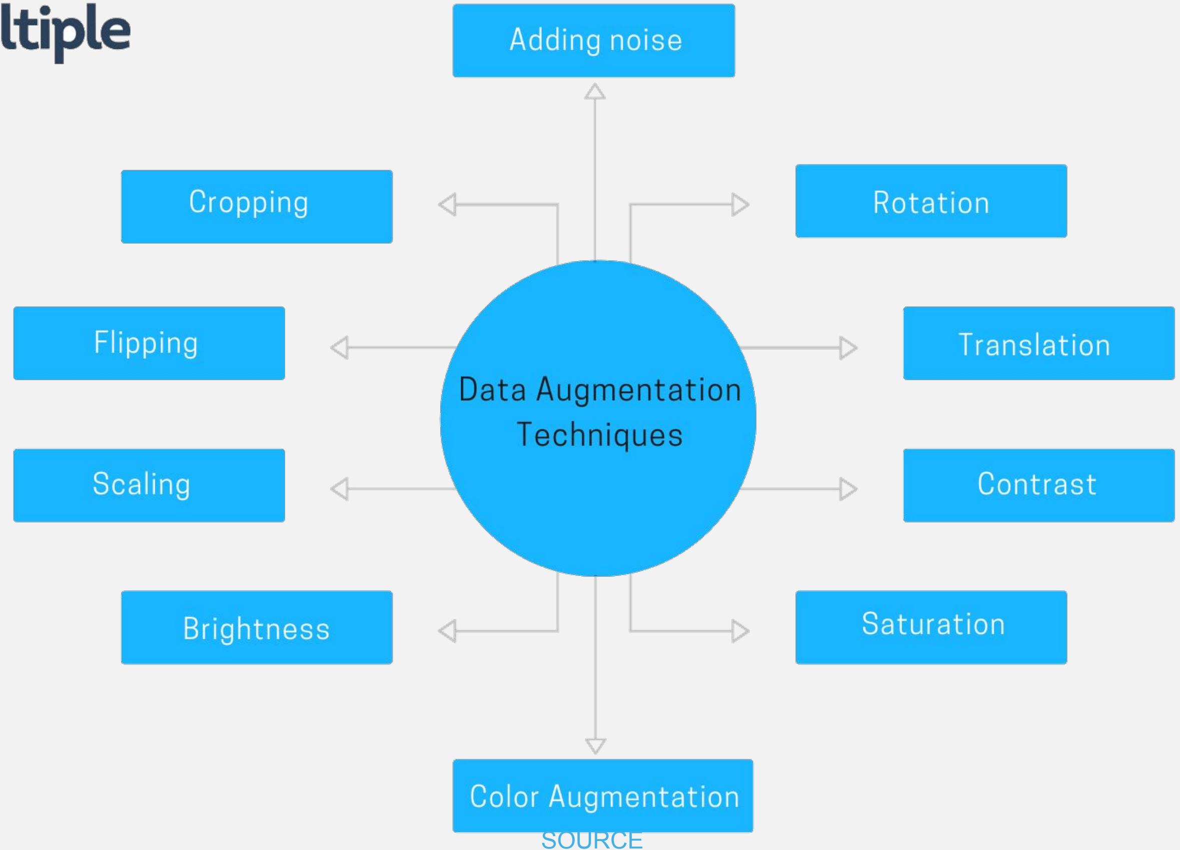
Amazon Datasets

Data Collection

…Still lacking sample data? You might need

Data augmentation *is the increase of an existing training dataset's size and diversity without the .requirement of manually collecting any new data* SOURCE

AI Multiple



Data Augmentation Techniques

Adding noise

Rotation

Translation

Contrast

Saturation

Color Augmentation

Brightness

Scaling

Flipping

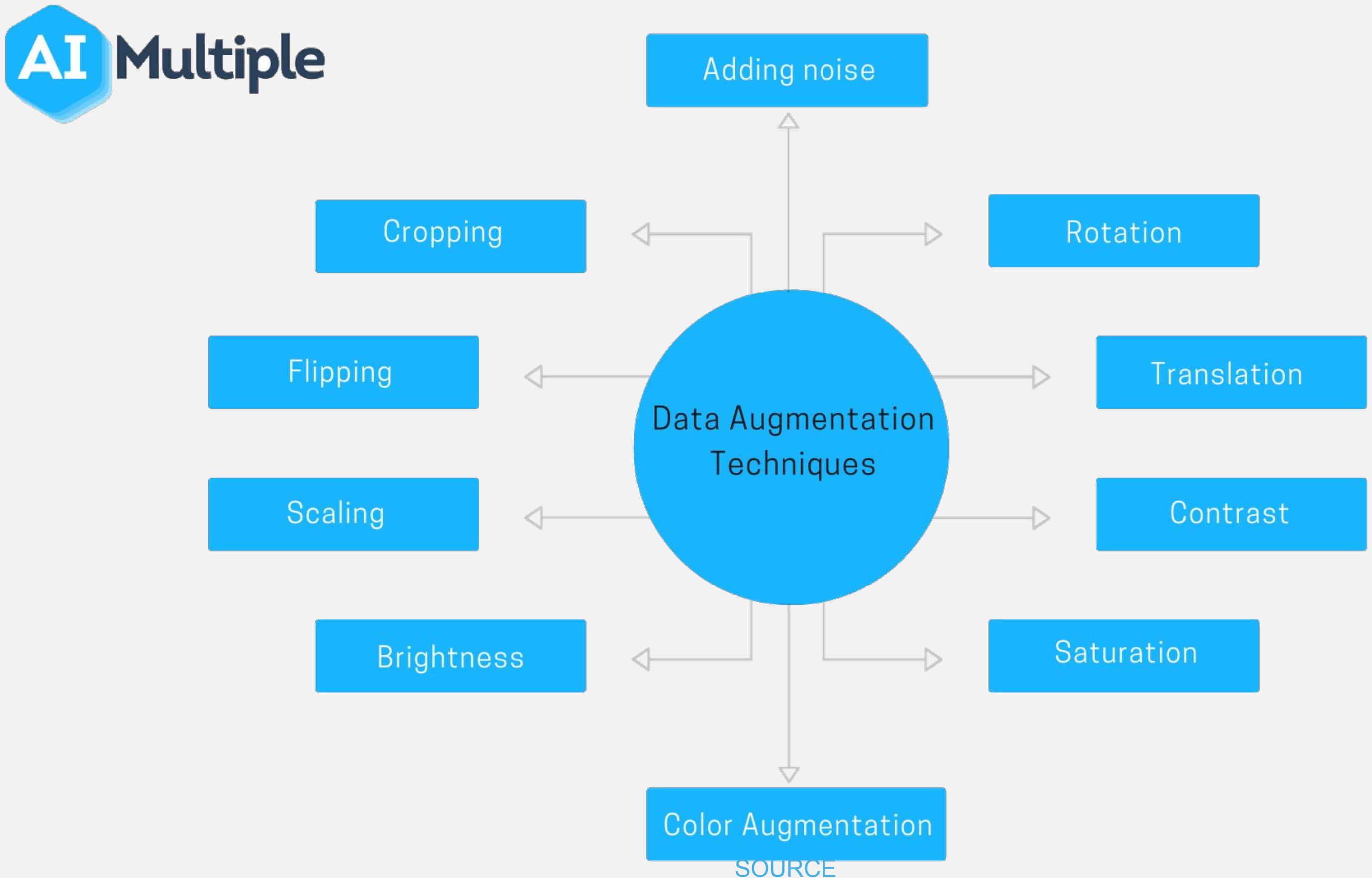
Cropping

SOURCE

Data Collection

…Still lacking sample data? You might need

Data augmentation *is the increase of an existing training dataset's size and diversity without the .requirement of manually collecting any new data* SOURCE

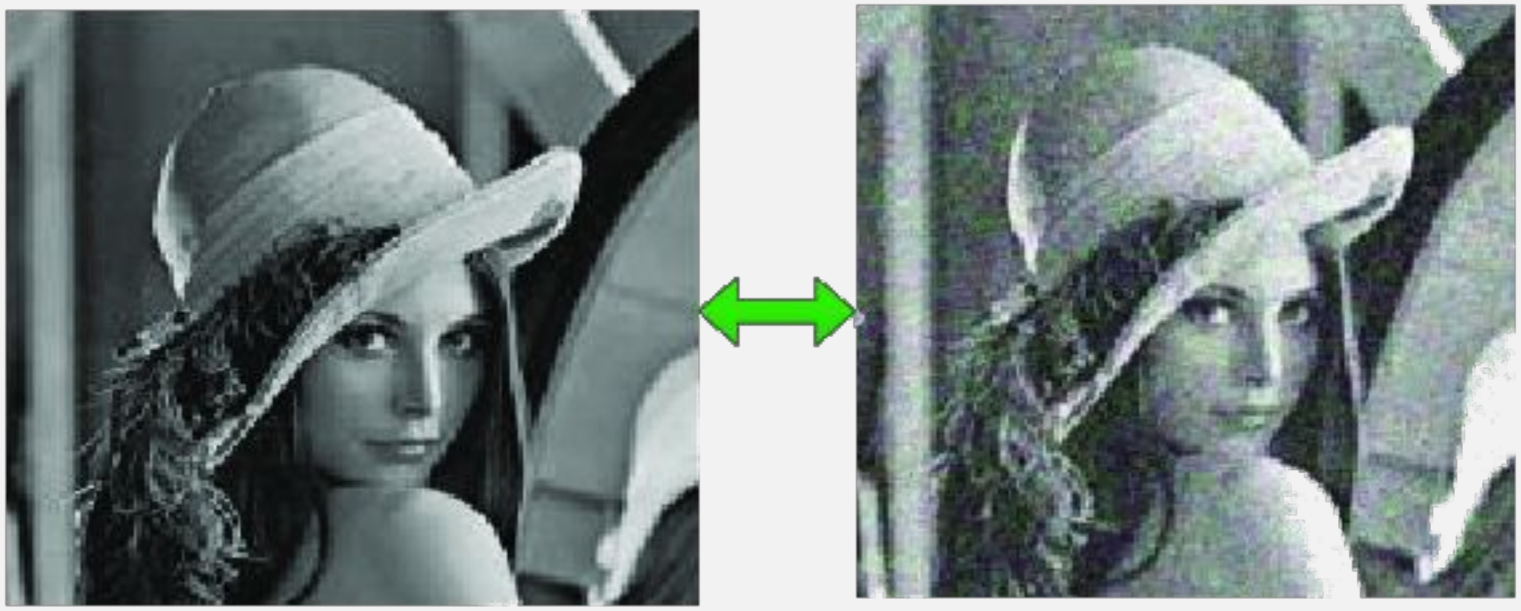


SOURCE

Data Collection

Data Augmentation Techniques

Adding Noise to images in machine learning is a technique used to improve model robustness, prevent overfitting, simulate real-world conditions, and enhance generalization by introducing random .variations into the training data



Original graph

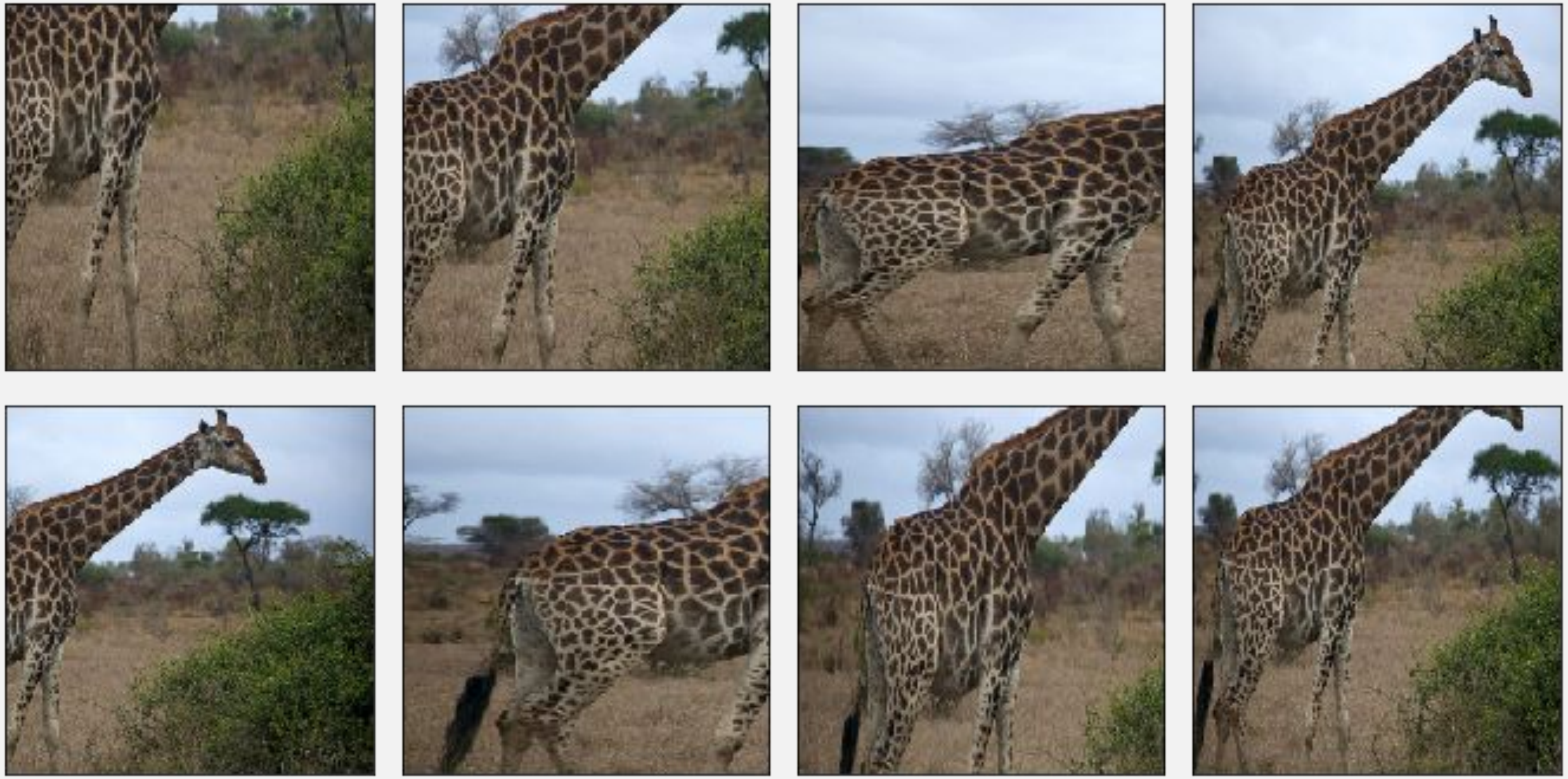
Image processed with noise

SOURCE

Data Collection

Data Augmentation Techniques

new **Cropping** images in data augmentation involves creating smaller sections of the original images to be used as training samples. This technique can significantly enhance the diversity of the dataset and help the model .generalize better

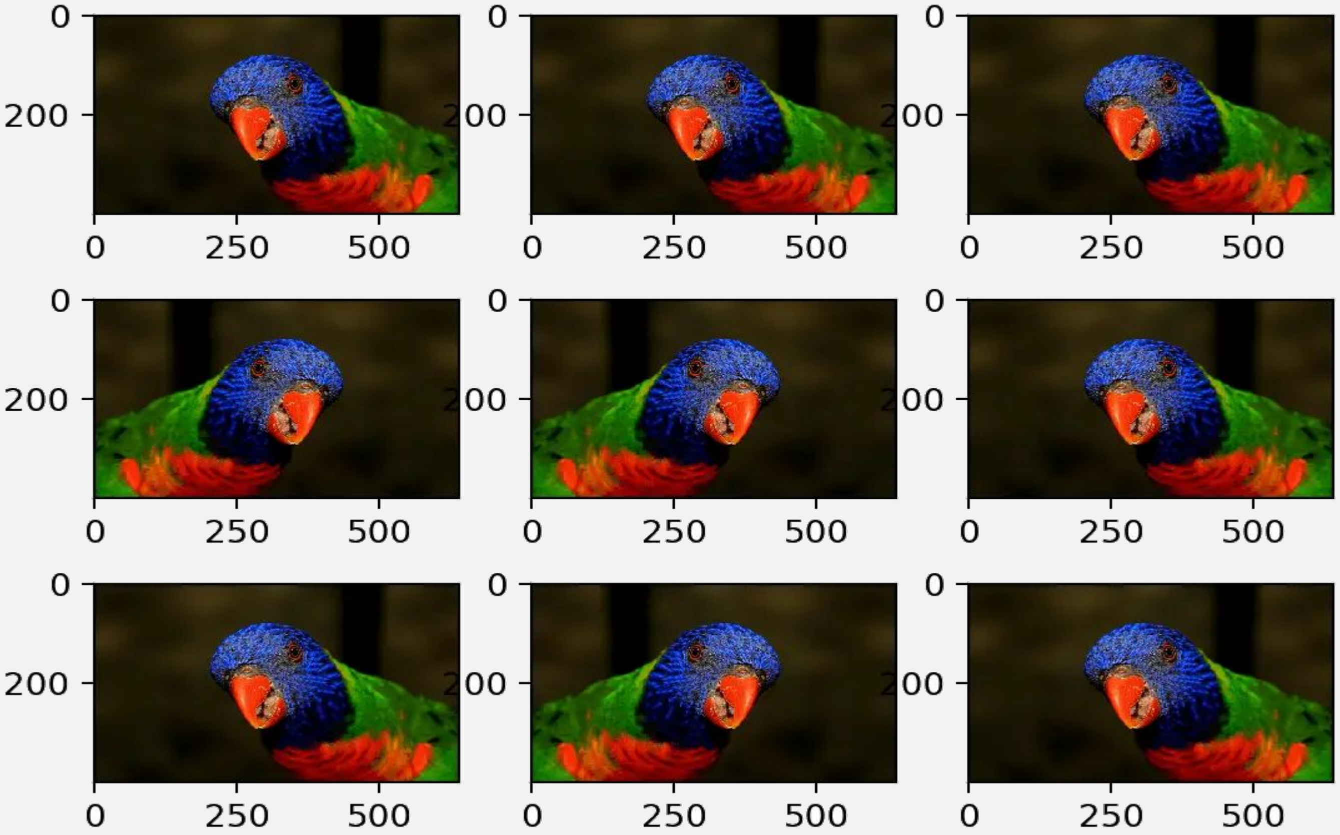


SOURCE

Data Collection

Data Augmentation Techniques

Flipping images is a common data augmentation technique used to artificially increase the size and diversity of the training dataset. This technique involves creating new images by flipping the original images horizontally, vertically, or both

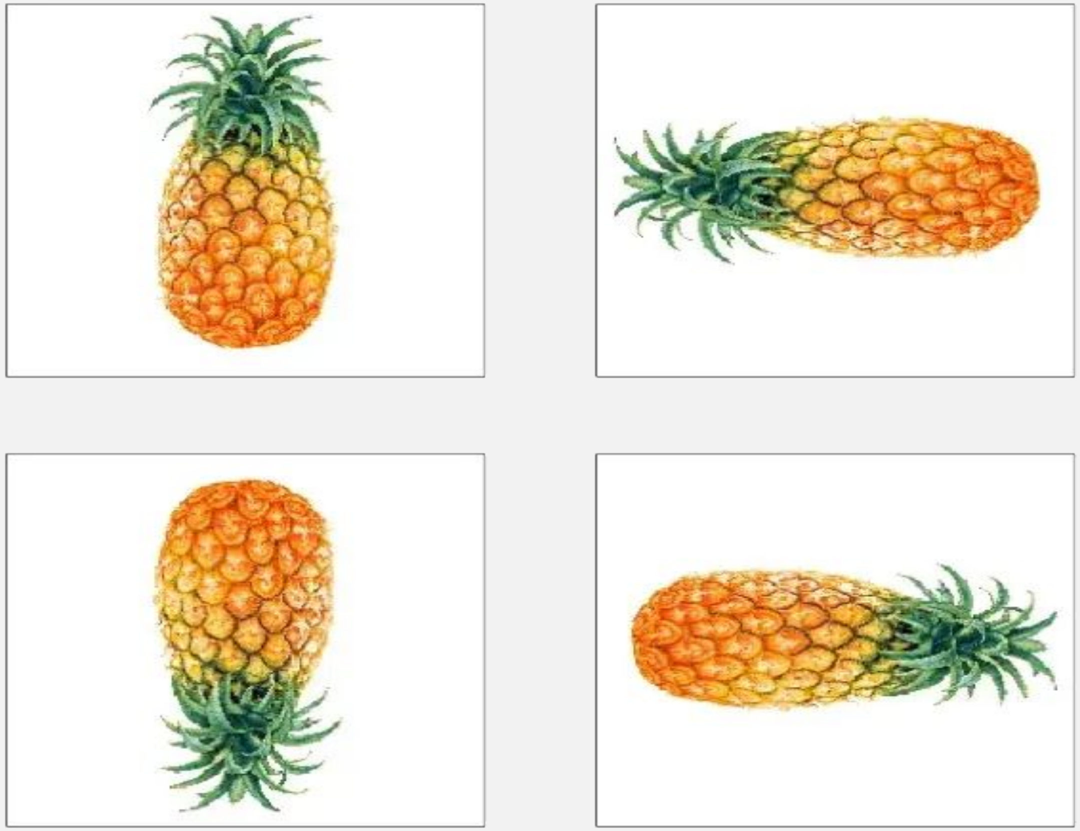


SOURCE

Data Collection

Data Augmentation Techniques

Rotation is another powerful data augmentation technique used in machine learning to increase the diversity of the training dataset and improve model robustness. It involves rotating the original images by various angles, .creating new training samples that help the model become invariant to the orientation of objects

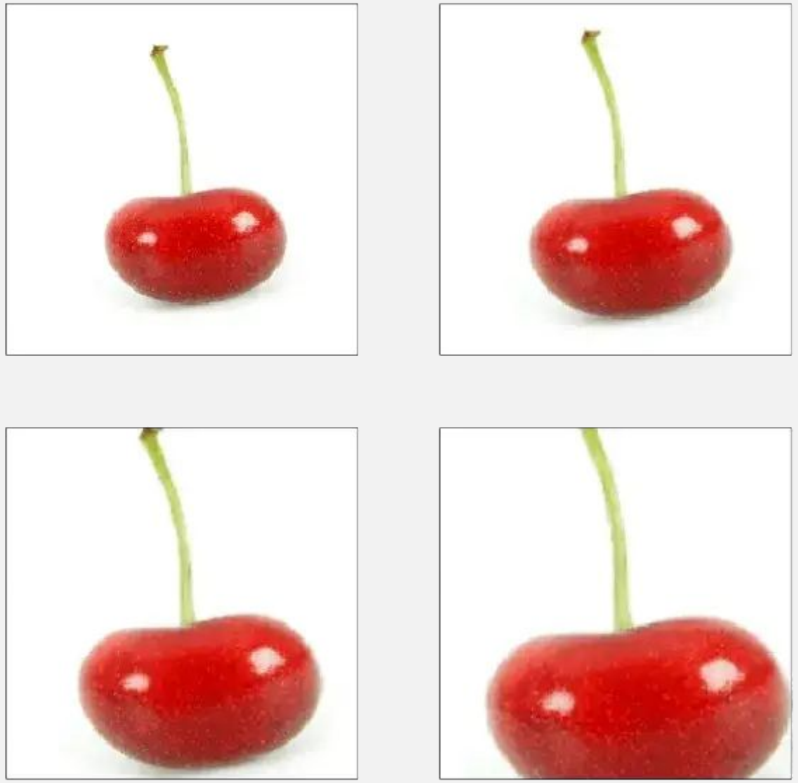


SOURCE

Data Collection

Data Augmentation Techniques

Scaling in data augmentation refers to the process of resizing or transforming data, typically images or other types of data, as part of the augmentation process. It involves changing the size, orientation, or appearance of data samples to increase the diversity of the dataset without fundamentally changing its underlying .characteristics

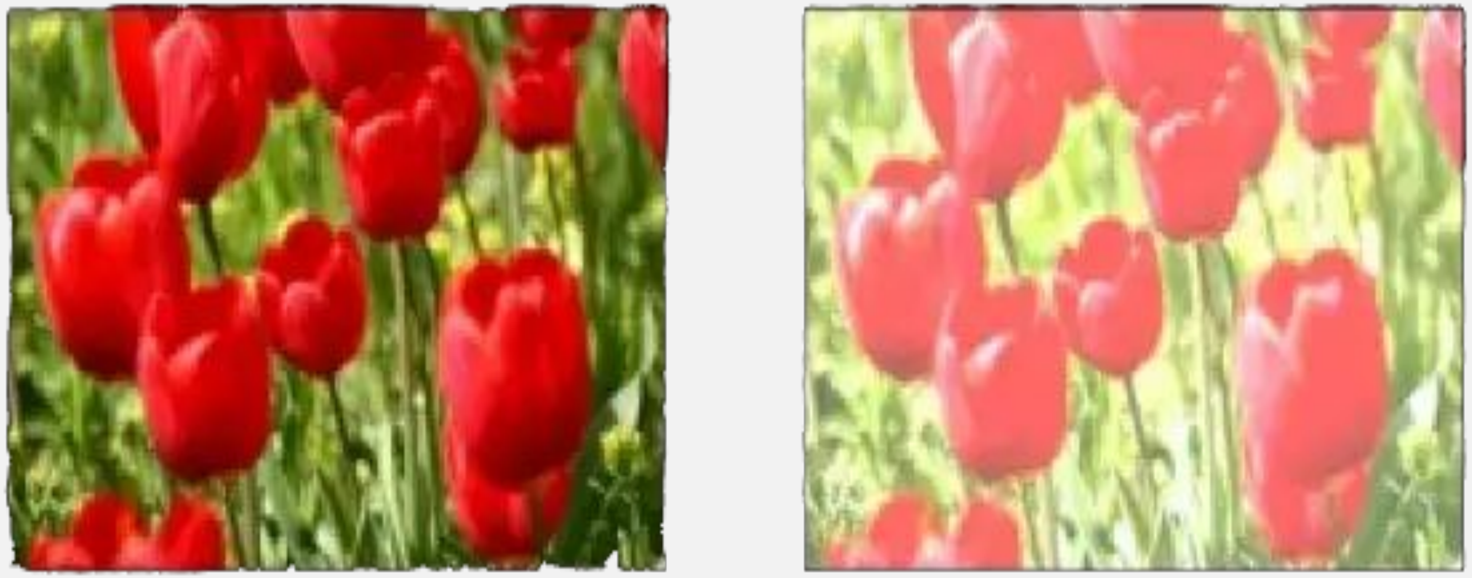


SOURCE

Data Collection

Data Augmentation Techniques

Brightness adjustment in data augmentation refers to modifying the intensity of light in an image to create variations of the original data. This technique is particularly useful in scenarios where lighting conditions may .vary in real-world applications



SOURCE

Data Collection

Planning Data Collection

.In this stage, we start with the selection of **data collection methods**

:Common methods include

.Surveys . 1

.Experiments . 2

.Web scraping . 3

APIs (Application Programming Interfaces) . 4

.Choose the appropriate methods to collect data from the identified sources SOURCE

Data Collection

Planning Data Collection

The Direct Data Collection Approach refers to a method of gathering information .straightforwardly from the source for the first time

Surveys and experiments are two fundamental methods for data collection in various fields, .including *social sciences*, *marketing*, *health research*, and *many areas of data science*

Surveys

Experiments

Data Collection

Planning Data Collection

Surveys gather self-reported data through formats like online questionnaires and interviews. They're useful for collecting opinions and behaviours.

Their advantages include:

- **Scalability:** Efficiently reach many respondents.
- **Versatility:** Collect diverse data, from demographics to opinions.
- **Comparability:** Standardized questions facilitate cross-group analysis.

Limitations include biases from question phrasing, respondent interpretation, and accuracy of responses

Data Collection

Planning Data Collection

Experiments manipulate variables to study effects and infer causality, often under controlled settings like labs, though they can also be in the field or online.

Key features include:

Control: Ability to manage conditions and isolate variables.

Randomization: Random assignment to groups to limit bias and support causal conclusions.

Repeatability: Can be replicated to confirm findings.

Valuable for exploring cause-and-effect

Limitations include high costs, time demands, and practical or ethical limitations.

Data Collection

Planning Data Collection

Indirect(Secondary) Approach: Involves using data that has already been collected by someone else for a different purpose and leveraging existing resources to gather information that can be applied to the current research.

Web scraping and API Calling are two fundamental secondary methods for data collection:

API

Web Scraping

Data Collection

Planning Data Collection

Web scraping is the process of extracting data from websites, automating the collection of information available online. It serves as a powerful tool in data collection, enabling analysts and scientists to gather vast amounts of data quickly, which is essential for analysis, research, and decision-making processes.

Key tools and technologies for web scraping include:

- **Beautiful Soup:** A Python library for parsing HTML and XML documents. It's widely used for simple projects and tasks that require quick data extraction from websites.
- **Selenium:** Originally a tool for testing web applications, Selenium can automate web browser interaction, making it suitable for scraping dynamic content that requires interaction with the webpage.
- **Scrapy:** An open-source and collaborative framework for extracting the data you need from websites. It's designed for web scraping. Scrapy is highly efficient, scalable, and versatile, making it suitable for large-scale web scraping projects.

Data Collection

Planning Data Collection

Typical Steps to handle a website in Beautiful Soup

- Fetching the web page content using requests.
- Parsing the content with Beautiful Soup to create a parse tree.
- Using Beautiful Soup's searching and navigation methods to find relevant data.
- Extracting and processing the data you need from the elements found.
- Iteratively refining your approach based on the specific requirements of your web scraping project and the structure of the web pages you're working with.

Data Collection

Planning Data Collection

What is Selenium?

- An open-source automation tool primarily used for automating testing web applications.
- Allows for browser automation, enabling tasks to be performed as if a real user is navigating the site so it can also render websites Dynamically.

Why Use Selenium for Web Scraping?

- **Dynamic Content:** Selenium can interact with webpages that load content dynamically, making it ideal for scraping modern sites.
- **Real Browser Interaction:** Performs operations in a real browser environment, allowing for actions like clicking buttons, filling forms, and scrolling.

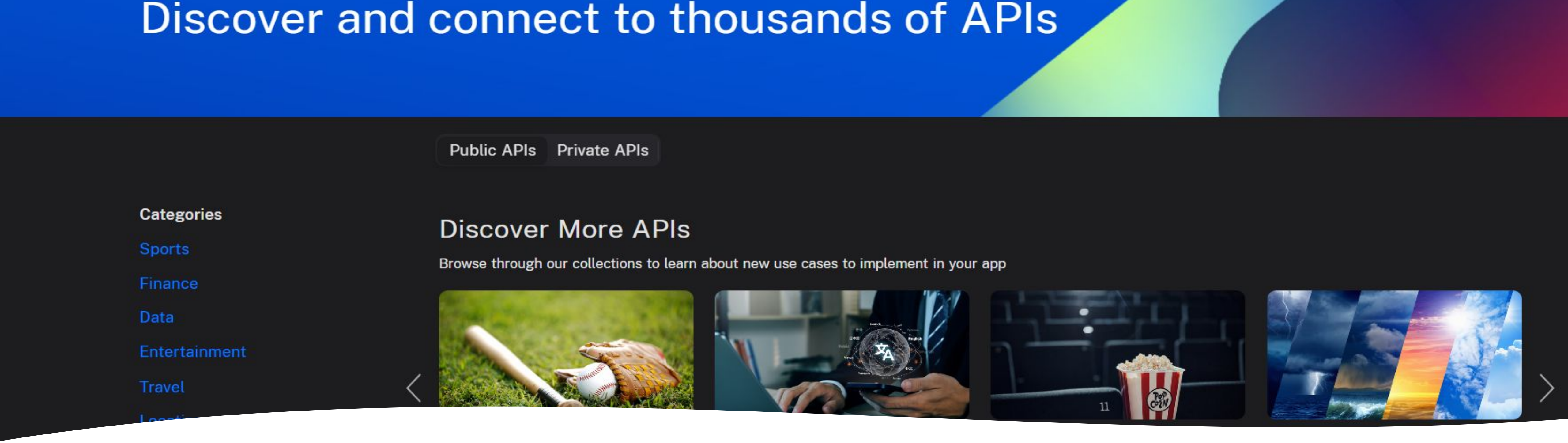
Data Collection

Planning Data Collection

APIs (Application Programming Interfaces) are software are tools that allow different software applications to communicate with each other. They acts as intermediaries allowing different software applications to communicate, simplifying the process of data collection by providing structured ways to request and receive data.

Advantages of Using APIs:

- **Efficiency:** Streamlines data access and functionality.
- **Real-Time Data:** Offers access to live data, crucial for up-to-date application needs.
- **Scalability:** Eases handling of growing data or demand with minimal infrastructure adjustments.
- **Cost-Effectiveness:** More affordable than developing custom data collection systems.



Discover and connect to thousands of APIs

Discover More APIs

Browse through our collections to learn about new use cases to implement in your app

Data Collection

Planning Data Collection

- **RapidAPI** [link] is a comprehensive platform that aggregates thousands of APIs across various domains
- It presents a unified platform for developers to discover, connect, and manage APIs through a single, standardized interface.
- It offers access to diverse data sources across various categories, including finance, sports, entertainment, weather, and more.

RapidAPI Considerations

In-Direct Approach

API Limits: Be aware of rate limits and quotas to avoid service interruptions.

Costs: Understand the pricing model of the API and usage charges.

Security: Keep your API key confidential to prevent unauthorized usage.

Performance: Test response times and reliability.

Documentation: Read the API documentation thoroughly.

Updates: Stay informed about any changes or updates to the API.

Support: Check the support options and community forums for help for Q&A.

Data Collection

Ensuring Data Quality

Ensuring data quality means apply quality assurance techniques to ensure the data is reliable and

.suitable for analysis, after that, you might need to go with data cleaning and preprocessing

Prepare the data for analysis by cleaning and preprocessing it. This involves handling missing values,

.removing duplicates, and transforming data into a suitable format

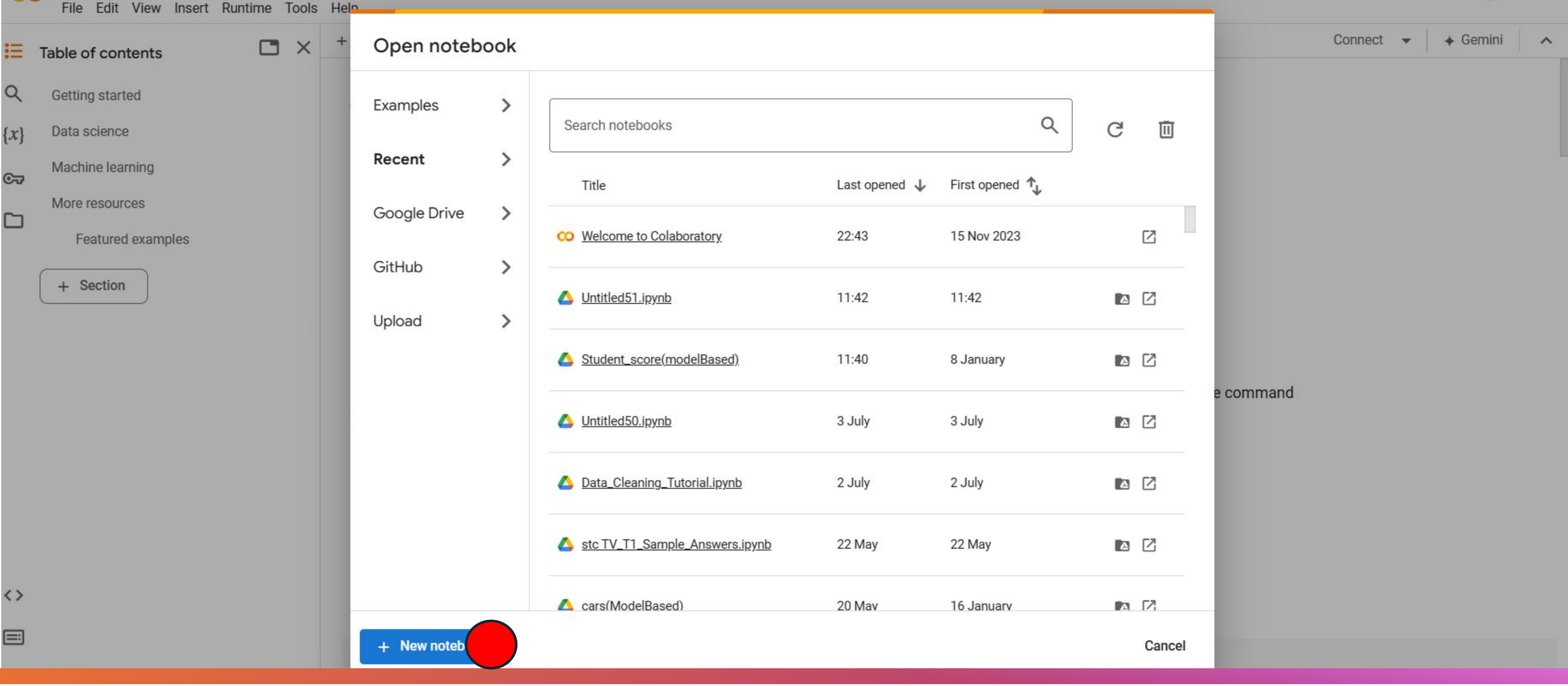
SOURCE

Data Preparation

Data Preparation

?What is Data Preparation

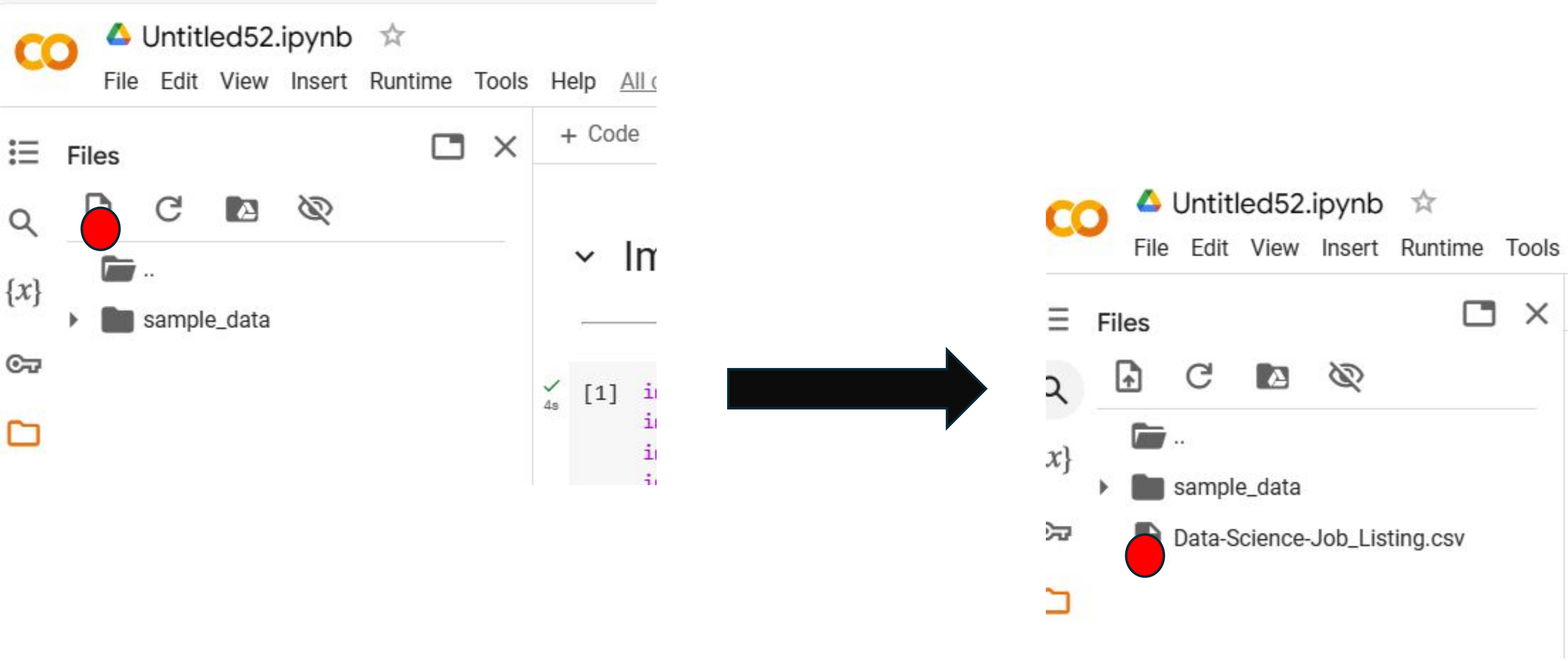
Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labelling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data. Data preparation can take up to 80% of the time spent on an ML project. Using specialized data preparation tools is .important to optimize this processSOURCE



Title	Last opened	First opened
Welcome to Colaboratory	22:43	15 Nov 2023
Untitled51.ipynb	11:42	11:42
Student_score(modelBased)	11:40	8 January
Untitled50.ipynb	3 July	3 July
Data_Cleaning_Tutorial.ipynb	2 July	2 July
stc TV_T1_Sample_Answers.ipynb	22 May	22 May
cars(ModelBased)	20 May	16 January

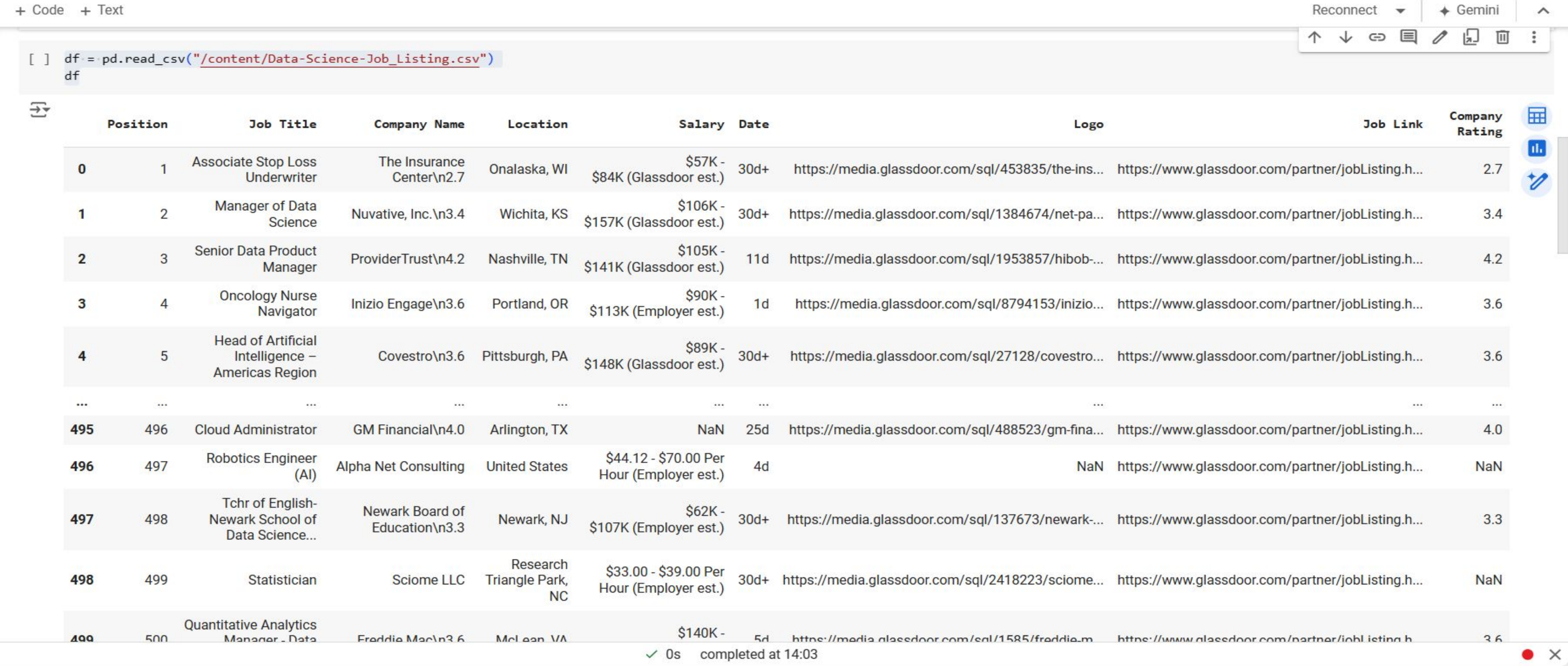
Data Preparation

Open Google Colab to Start View Data: Welcome to Colaboratory - Colab (google.com)


Untitled52.ipynb
File Edit View Insert Runtime Tools Help
Files
+ Code
..
sample_data
[1]
Untitled52.ipynb
File Edit View Insert Runtime Tools
Files
..
sample_data
Data-Science-Job_Listing.csv

Data Preparation

Click here to upload your dataset to googlecolab



```
[ ] df = pd.read_csv("/content/Data-Science-Job_Listing.csv")
    df
```

	Position	Job Title	Company Name	Location	Salary	Date	Logo	Job Link	Company Rating
0	1	Associate Stop Loss Underwriter	The Insurance Center\n2.7	Onalaska, WI	$57K - $84K (Glassdoor est.)	30d+	https://media.glassdoor.com/sql/453835/the-ins...	https://www.glassdoor.com/partner/jobListing.h...	2.7
1	2	Manager of Data Science	Nuvative, Inc.\n3.4	Wichita, KS	$106K - $157K (Glassdoor est.)	30d+	https://media.glassdoor.com/sql/1384674/net-pa...	https://www.glassdoor.com/partner/jobListing.h...	3.4
2	3	Senior Data Product Manager	ProviderTrust\n4.2	Nashville, TN	$105K - $141K (Glassdoor est.)	11d	https://media.glassdoor.com/sql/1953857/hibob-...	https://www.glassdoor.com/partner/jobListing.h...	4.2
3	4	Oncology Nurse Navigator	Inizio Engage\n3.6	Portland, OR	$90K - $113K (Employer est.)	1d	https://media.glassdoor.com/sql/8794153/inizio...	https://www.glassdoor.com/partner/jobListing.h...	3.6
4	5	Head of Artificial Intelligence – Americas Region	Covestro\n3.6	Pittsburgh, PA	$89K - $148K (Glassdoor est.)	30d+	https://media.glassdoor.com/sql/27128/covestro...	https://www.glassdoor.com/partner/jobListing.h...	3.6
...	...	...	...	...	...	...	...	...	...
495	496	Cloud Administrator	GM Financial\n4.0	Arlington, TX	NaN	25d	https://media.glassdoor.com/sql/488523/gm-fina...	https://www.glassdoor.com/partner/jobListing.h...	4.0
496	497	Robotics Engineer (AI)	Alpha Net Consulting	United States	$44.12 - $70.00 Per Hour (Employer est.)	4d	NaN	https://www.glassdoor.com/partner/jobListing.h...	NaN
497	498	Tchr of English-Newark School of Data Science...	Newark Board of Education\n3.3	Newark, NJ	$62K - $107K (Employer est.)	30d+	https://media.glassdoor.com/sql/137673/newark-...	https://www.glassdoor.com/partner/jobListing.h...	3.3
498	499	Statistician	Sciome LLC	Research Triangle Park, NC	$33.00 - $39.00 Per Hour (Employer est.)	30d+	https://media.glassdoor.com/sql/2418223/sciome...	https://www.glassdoor.com/partner/jobListing.h...	NaN
499	500	Quantitative Analytics Manager - Data	Freddie Mac\n3.6	McLean, VA	$140K -	5d	https://media.glassdoor.com/sql/1585/freddie-m...	https://www.glassdoor.com/partner/jobListing.h...	3.6

✓ 0s completed at 14:03

Data Preparation

Read dataset

Data Preparation

?How do you prepare your data

Data preparation follows a series of steps that starts with collecting the right data, followed by cleaning, labelling, .and then validation and visualization

Collect Data : Gathering all necessary data for ML, which can be challenging due to the diverse sources (e.g., **. 1** laptops, data warehouses, cloud, applications, devices) and increasing data volumes. Different data formats and .types (e.g., video vs. tabular) add complexity

Clean Data : Correcting errors and filling in missing data to ensure quality. This involves transforming the **. 2** data into a consistent, readable format by adjusting field formats (e.g., dates, currency), modifying naming .conventions, and standardizing values and units of measure

Data Preparation

Label Data : Identifying raw data (images, text, videos, etc.) and adding informative labels to provide **. 3** context for ML models to learn from. Labels indicate features like objects in photos, words in audio recordings, or irregularities in X-rays. Essential for tasks in computer vision, natural language processing, and speech .recognition

Validate and Visualize : After cleaning and labelling, ML teams explore the data to ensure its accuracy **. 4** and readiness. Visualizations (e.g., histograms, scatter plots, box plots) help confirm correctness and aid in exploratory data analysis to discover patterns, spot anomalies, test hypotheses, or check assumptions without .formal modelling

Exploratory Data Analysis(EDA)

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a technique for examining datasets to comprehend their primary attributes. It entails summarizing data features, identifying patterns, and uncovering relationships using visual and statistical methods. EDA aids in deriving insights and developing .hypotheses for subsequent analysis

THANK YOU

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority