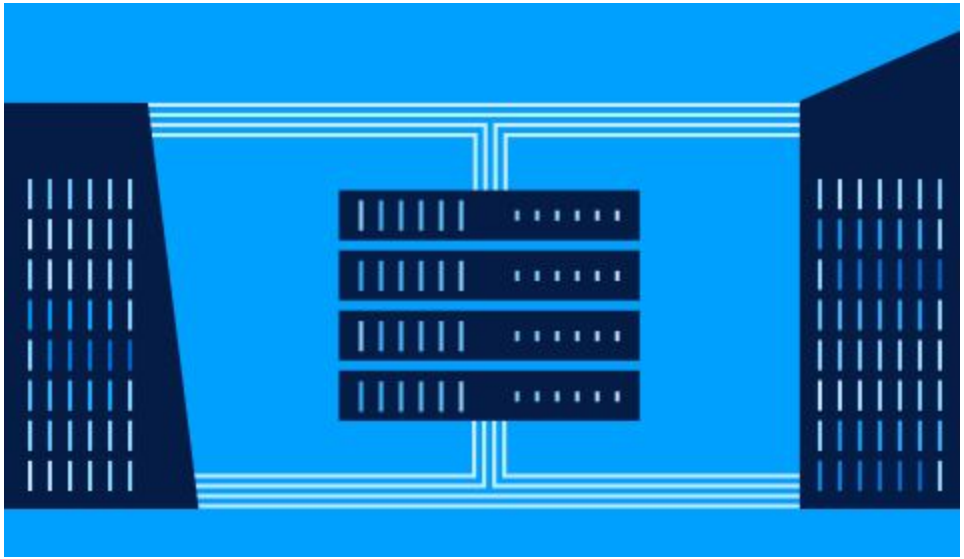
Data Wrangling

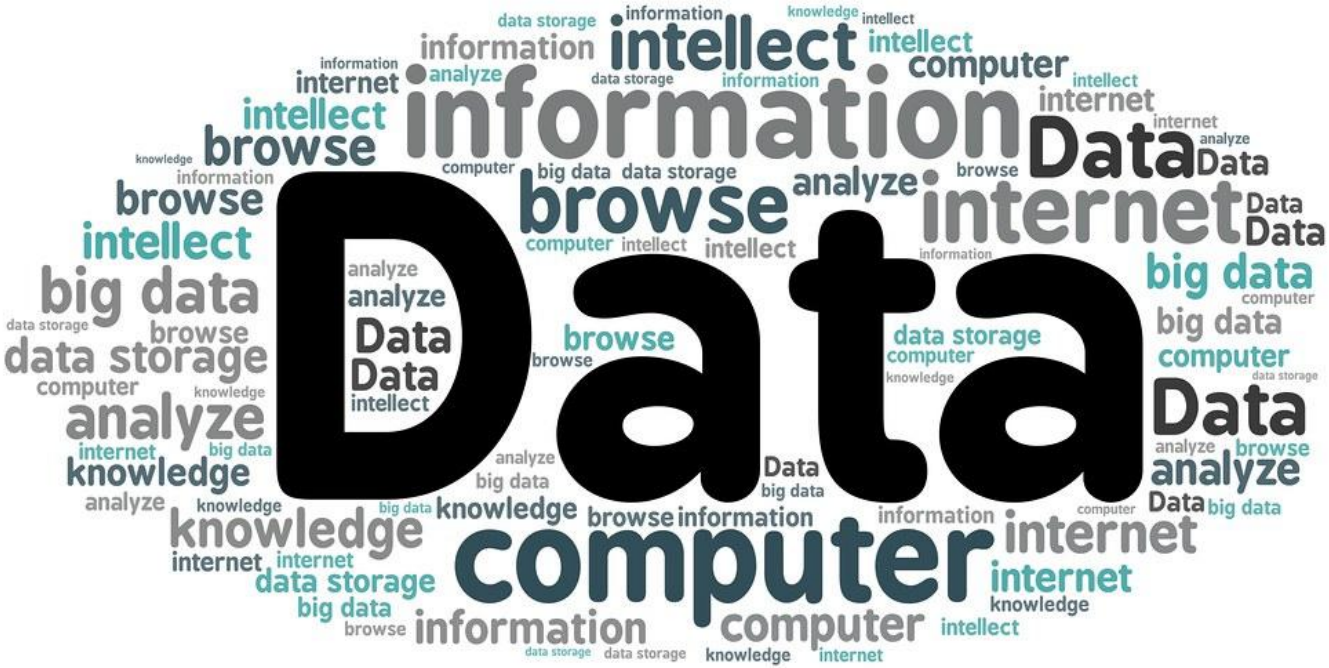
Professor Ernesto Lee

Introduction to Data Wrangling



Why Data Wrangling?

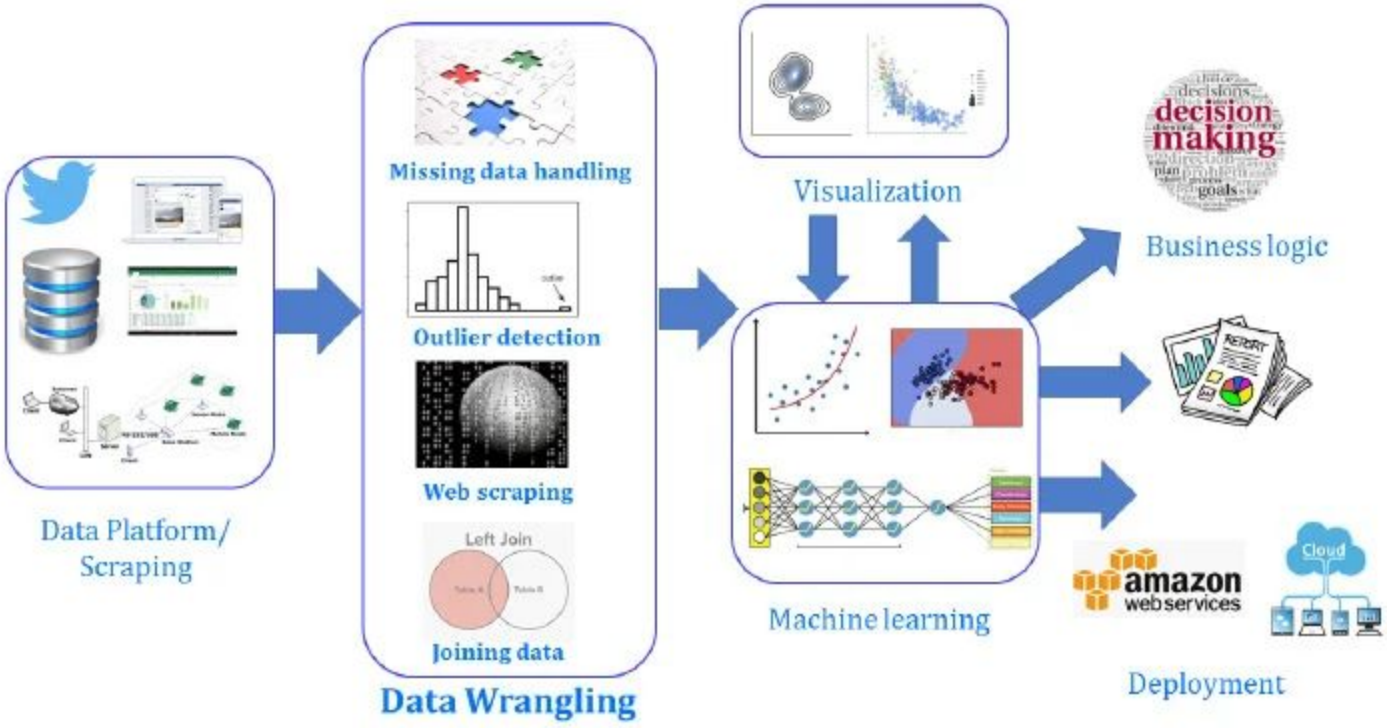
- There is a big gap between data and clean data…



Importance of Data Wrangling

Generally, the task of data wrangling involves the following steps:

1. Scraping raw data from multiple sources (including web and database tables)
2. Imputing (replacing missing data using various techniques), formatting, and transforming - basically making it ready to be used in the modeling process (such as advanced machine learning)
3. Handling read/write errors
4. Detecting outliers
5. Performing quick visualizations (plotting) and basic statistical analysis to judge the quality of formatted data


Missing data handling
Outlier detection
Web scraping
Left Join
Joining data
Data Wrangling
Data Platform/
Scraping
Visualization
Machine learning
decision
making
Business logic
REPORT
amazon
web services
Cloud
Deployment

Tools for Data Wrangling

- General-purpose data analysis platforms, such as Microsoft Excel (with add-ins)
- Statistical discovery package, such as JMP (from SAS)
- Modeling platforms, such as RapidMiner
- Analytics platforms from niche players that focus on data wrangling, such as Trifacta, Paxata, and Alteryx
- Tableau based tools
- Python

Let's Look at an Example

https://github.com/fenago/Intro2ML

- Redfin
- MurderData
- Financial Data

Business Context

Apply your knowledge

In India, did the enrollment in primary/secondary/college education increase with the improvement of per capita GDP in the past 15 years?

To provide an accurate and analyzed result, machine learning and data visualization techniques will be used by an expert data scientist.

The actual modeling and analysis will be done by a senior data scientist, who will use machine learning and data visualization for analysis.

As a data wrangling expert, your job will be to acquire and provide a clean dataset that contains educational enrollment and GDP data side by side.

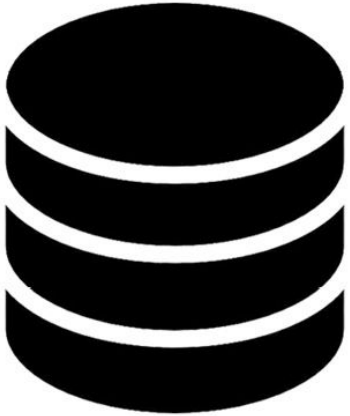
Education Data

Economic data

THE WORLD BANK

Data Wrangling

Clean data stored for further analysis



	Region / Country / Area	Unnamed: 1	Year	Series	Value	Footnotes	Source
0	1	Total, all countries or areas	2005	Students enrolled in primary education (thousa...	678,990	NaN	United Nations Educational, Scientific and Cul…
1	1	Total, all countries or areas	2005	Gross enrollement ratio – Primary (male)	104.8	NaN	United Nations Educational, Scientific and Cul…
2	1	Total, all countries or areas	2005	Gross enrollment ratio – Primary (female)	99.8	NaN	United Nations Educational, Scientific and Cul…
3	1	Total, all countries or areas	2005	Students enrolled in secondary education (thou…	509,100	NaN	United Nations Educational, Scientific and Cul…
4	1	Total, all countries or areas	2005	Gross enrollment ratio – Secondary (male)	65.7	NaN	United Nations Educational, Scientific and Cul…

Exercise

Download the dataset from the UN data from GitHub from the following link:
https://github.com/fenago/Intro2ML/blob/main/India_World_Bank_Info.csv

The UN data contains missing values. Clean the data to prepare a simple final dataset with the required data and save it to your local drive as a SQL database file.

Load into Excel for now.

Since the first row does not contain useful information, remove it.

Drop the column region/country/area and source.

Assign the following names as columns of the DataFrame: Region/County/Area, Year, Data, Value, and Footnotes.

Check how many unique values are present in the Footnotes column.

Check the type of the value column.

Create a function to convert the value column into floating-point numbers.

Print the unique values in the data column.