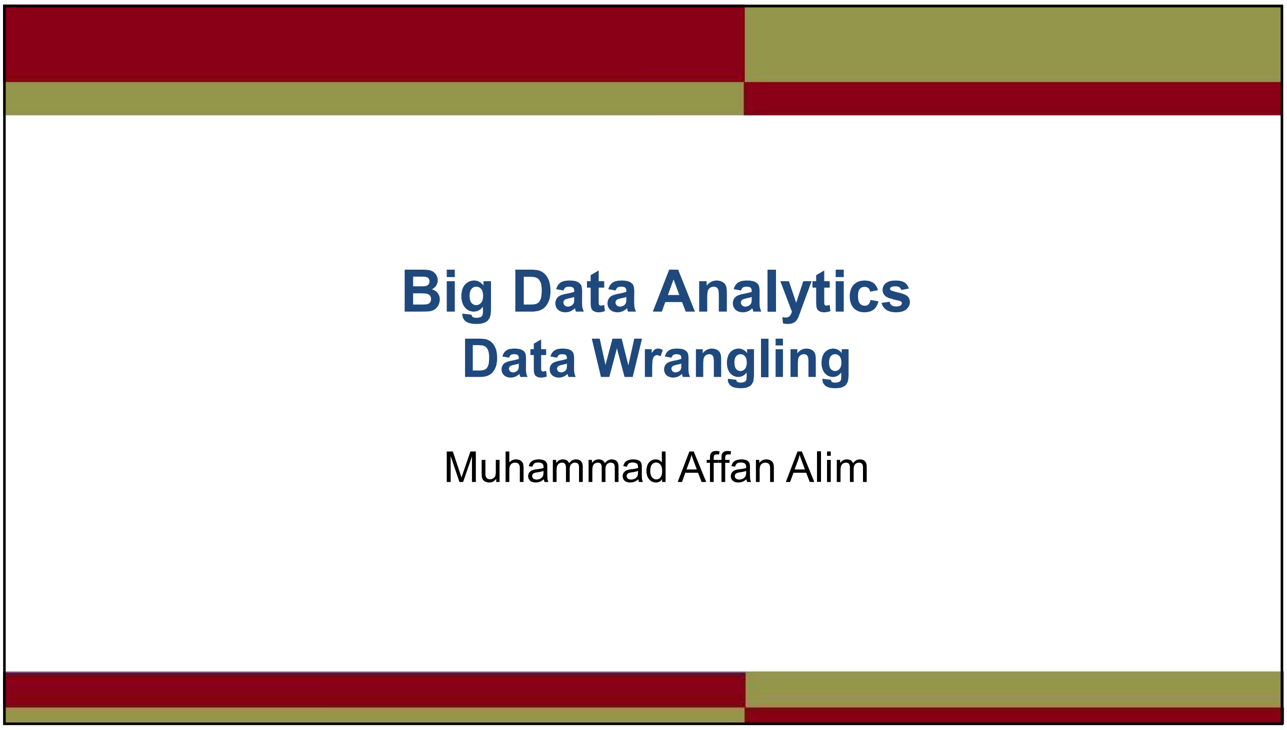
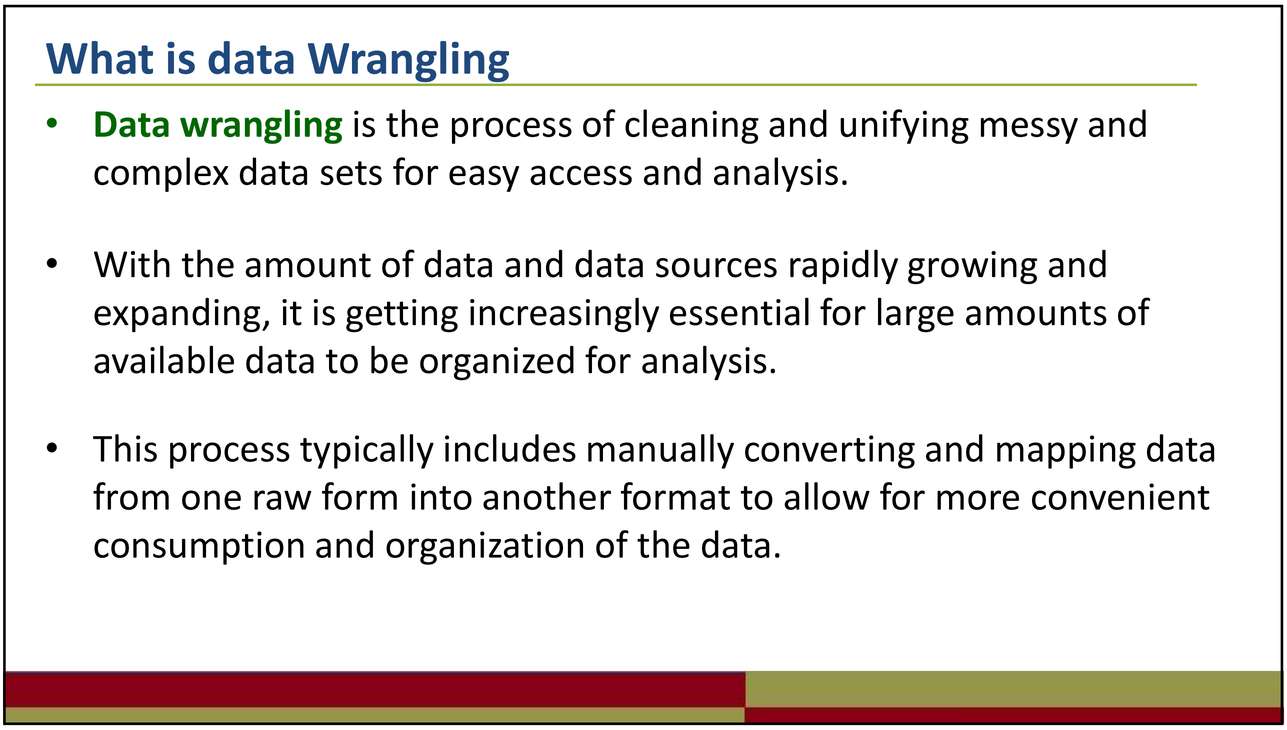


Big Data Analytics
Data Wrangling

Muhammad Affan Alim



What is data Wrangling

- **Data wrangling** is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.

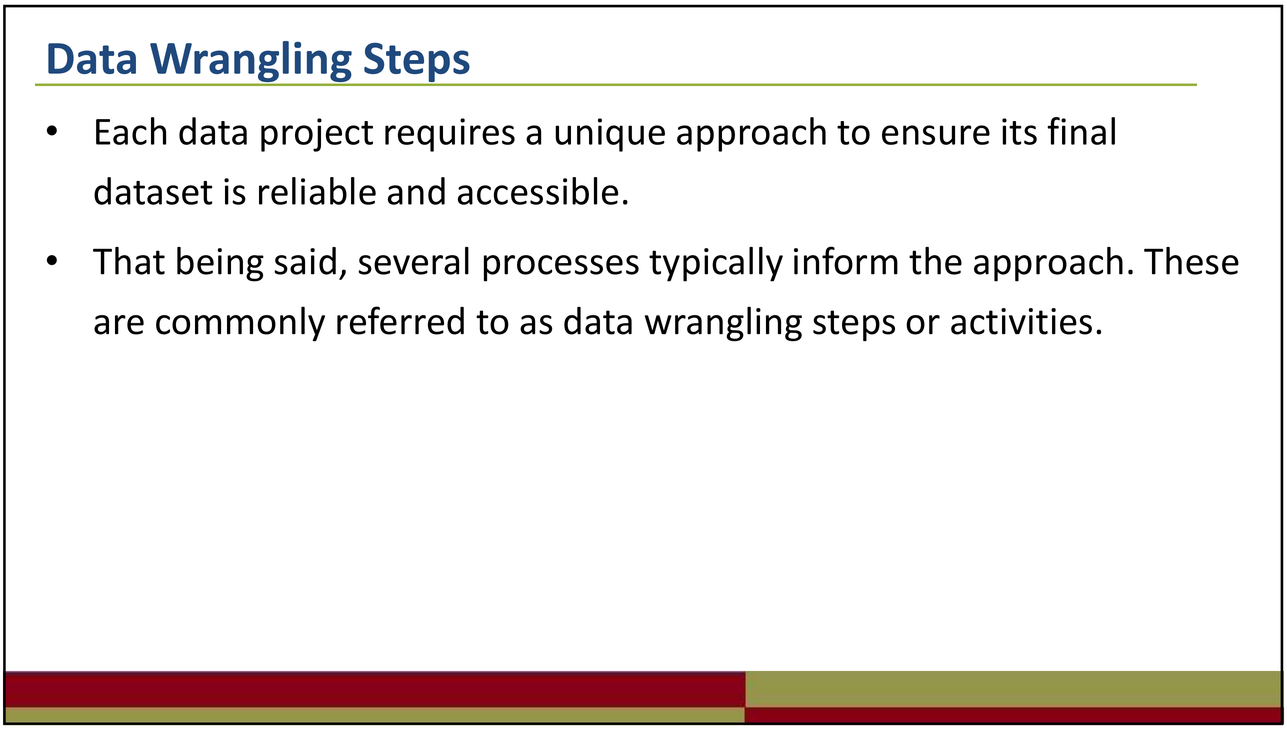
- This process typically includes manually converting and mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

What is data Wrangling-second

- **Data wrangling**—also called data cleaning, data remediation, or data munging—refers to a variety of processes designed to transform raw data into more readily used formats.
- The exact methods differ from project to project depending on the data you’re leveraging and the goal you’re trying to achieve.

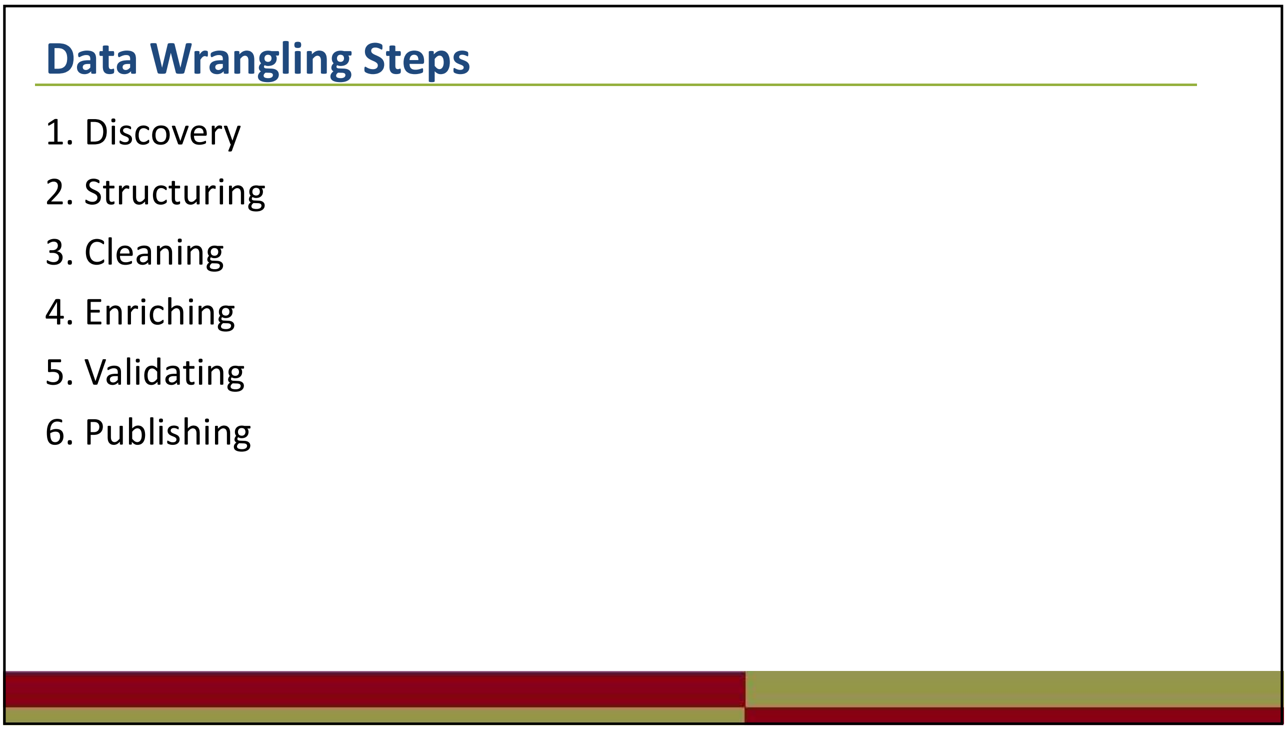
What is data Wrangling-second

- **Some examples of data wrangling include:**
  - Merging multiple data sources into a single dataset for analysis
  - Identifying gaps in data (for example, empty cells in a spreadsheet) and either filling or deleting them
  - Deleting data that’s either unnecessary or irrelevant to the project you’re working on
  - Identifying extreme outliers in data and either explaining the discrepancies or removing them so that analysis can take place



Data Wrangling Steps

- Each data project requires a unique approach to ensure its final dataset is reliable and accessible.
- That being said, several processes typically inform the approach. These are commonly referred to as data wrangling steps or activities.



Data Wrangling Steps

1. Discovery
2. Structuring
3. Cleaning
4. Enriching
5. Validating
6. Publishing

Data Wrangling Steps

1. Discovery

- Discovery refers to the process of familiarizing yourself with data so you can conceptualize how you might use it. You can liken it to looking in your refrigerator before cooking a meal to see what ingredients you have at your disposal.

- During discovery, you may identify trends or patterns in the data, along with obvious issues, such as missing or incomplete values that need to be addressed. This is an important step, as it will inform every activity that comes afterward.

Data Wrangling Steps

2. Structuring

- Raw data is typically unusable in its raw state because it’s either incomplete or misformatted for its intended application.
- Data structuring is the process of taking raw data and transforming it to be more readily leveraged. The form your data takes will depend on the analytical model you use to interpret it.

Data Wrangling Steps

2. Structuring-example

- Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization.
- These inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

Data Wrangling Steps

2. Structuring-example

- For example, the “purchase date” column name variations across sources may include:
  - purchaseDate
  - transaction_date
  - txDate
  - prchsdt
- And the values themselves are likely not rationalized:
  - 6-20-2018
  - 06/20/2018
  - 20-JUN-2018 08:03
  - 20/06/18

Data Wrangling Steps

3. Cleaning

- Data cleaning is the process of removing inherent errors in data that might distort your analysis or render it less valuable.
- Cleaning can come in different forms, including deleting empty cells or rows, removing outliers, and standardizing inputs.
- The goal of data cleaning is to ensure there are no errors (or as few as possible) that could influence your final analysis.

Data Wrangling Steps

4. Enriching

- Once you understand your existing data and have transformed it into a more usable state, you must determine whether you have all of the data necessary for the project at hand.
- If not, you may choose to enrich or augment your data by incorporating values from other datasets.
- For this reason, it’s important to understand what other data is available for use.
- If you decide that enrichment is necessary, you need to repeat the steps above for any new data.

Data Wrangling Steps

5. Validating

- Data validation refers to the process of verifying that your data is both consistent and of a high enough quality.
- During validation, you may discover issues you need to resolve or conclude that your data is ready to be analyzed. Validation is typically achieved through various automated processes and requires programming.

Data Wrangling Steps

5. Validating-tools

Key Data Validation Testing Tools | Data Validation Software

- Various Data Validation Testing tools are available in the market for data validation. Some of them given below -
  - Datameer
  - Talend
  - Informatica
  - QuerySurge
  - ICEDQ
  - Datagaps ETL Validator
  - DbFit
  - Data-Centric Testing

Data Wrangling Steps

5. Validating-How to Adopt Data Validation Testing?

- There are various approaches and techniques to accomplish Data Validation testing.

1. Data Accuracy testing to ensure that the provided data is correct.
2. Data Completeness testing to check whether the data is complete or not.
3. To verify that the provided data go successfully through transformations or not is by Data Transformation Testing.
4. Data Quality testing to handle bad data.
5. Database comparison testing to compare the source DB and target DB.
6. End to End testing.
7. Data warehouse testing.

Data Wrangling Steps

6. Publishing

- Once your data has been validated, you can publish it. This involves making it available to others within your organization for analysis. The format you use to share the information—such as a written report or electronic file—will depend on your data and the organization’s goals.

The Goals of Data Wrangling

- Reveal a "deeper intelligence" by gathering data from multiple sources
- Provide accurate, actionable data in the hands of business analysts in a timely matter
- Reduce the time spent collecting and organizing unruly data before it can be utilized
- Enable data scientists and analysts to focus on the analysis of data, rather than the wrangling
- Drive better decision-making skills by senior leaders in an organization

Data Wrangling Tools

Basic Data Munging Tools

- **Excel Power Query / Spreadsheets** — the most basic structuring tool for manual wrangling.
- **OpenRefine** — more sophisticated solutions, requires programming skills
- **Google DataPrep** - for exploration, cleaning, and preparation.
- **Tabula** — swiss army knife solutions — suitable for all types of data
- **DataWrangler** — for data cleaning and transformation.
- **CSVKit** — for data converting

Data Wrangling Tools

Data Wrangling in Python

1. **Numpy (aka Numerical Python)** — the most basic package. Lots of features for operations on n-arrays and matrices in Python. The library provides vectorization of mathematical operations on the NumPy array type, which improves performance and accordingly speeds up the execution.
2. **Pandas** — designed for fast and easy data analysis operations. Useful for data structures with labeled axes. Explicit data alignment prevents common errors that result from misaligned data coming in from different sources.

Data Wrangling Tools

Data Wrangling in Python

3. **Matplotlib** — Python visualization module. Good for line graphs, pie charts, histograms, and other professional grade figures.
4. **Plotly** — for interactive, publication-quality graphs. Excellent for line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axis, polar graphs, and bubble charts.
5. **Theano** — library for numerical computation similar to Numpy. This library is designed to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently.

What is Feature Engineering

- Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

What is Feature Engineering-Inroduction

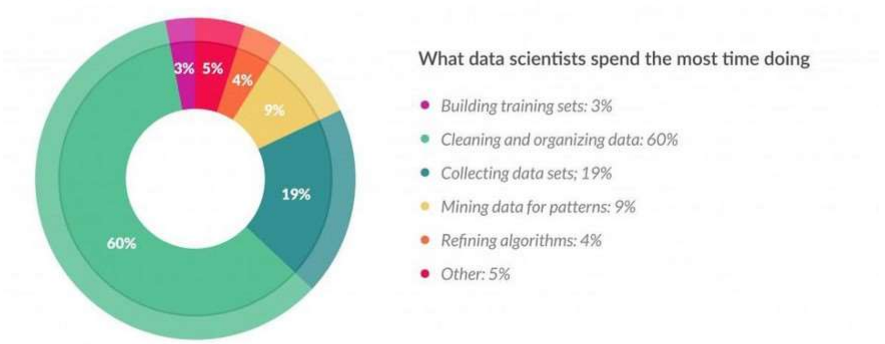
- Basically, all machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristic to work properly. Here, the need for feature engineering arises

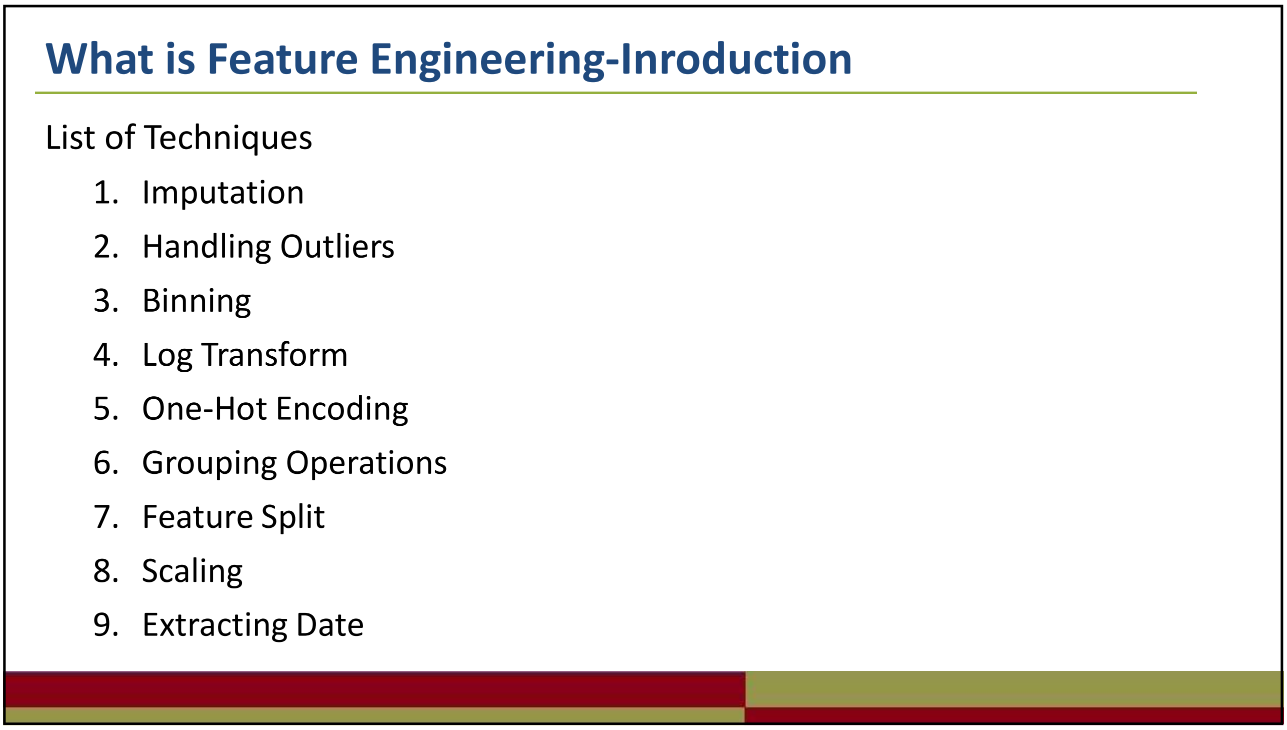
What is Feature Engineering-Inroduction

- I think feature engineering efforts mainly have two goals:
  1. Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
  2. Improving the performance of machine learning models.

What is Feature Engineering-Inroduction

According to a survey in Forbes, data scientists spend 80% of their time on data preparation:





What is Feature Engineering-Inroduction

List of Techniques

1. Imputation
2. Handling Outliers
3. Binning
4. Log Transform
5. One-Hot Encoding
6. Grouping Operations
7. Feature Split
8. Scaling
9. Extracting Date