Aasif Codes

DATA CLEANING

What is Data Cleaning?

Data Cleaning is the process of finding and *removing errors*, *inconsistencies, duplications*, and *missing entries* from data to increase data consistency and quality—also known as **data scrubbing** or **Cleansing.**

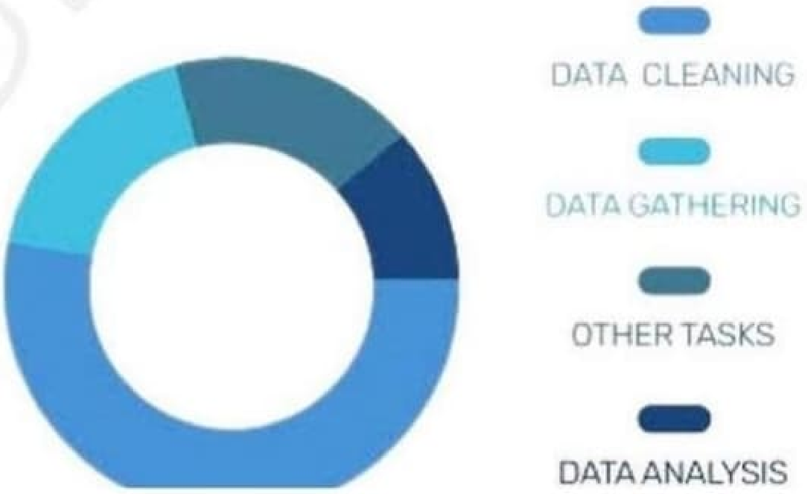
Steps in the data cleaning process are:

1. Removing duplicates
2. Remove irrelevant data
3. Standardize capitalization
4. Convert data type
5. Handling outliers
6. Fix errors
7. Language Translation
8. Handle missing values

Why is Data Cleaning so Important

Real-world data is noisy and contains a lot of errors. They are not in their best format.So, it becomes important that these data points need to be fixed.

It is estimated that **data scientists** spend between **80 to 90 percen**t of their time in data cleaning



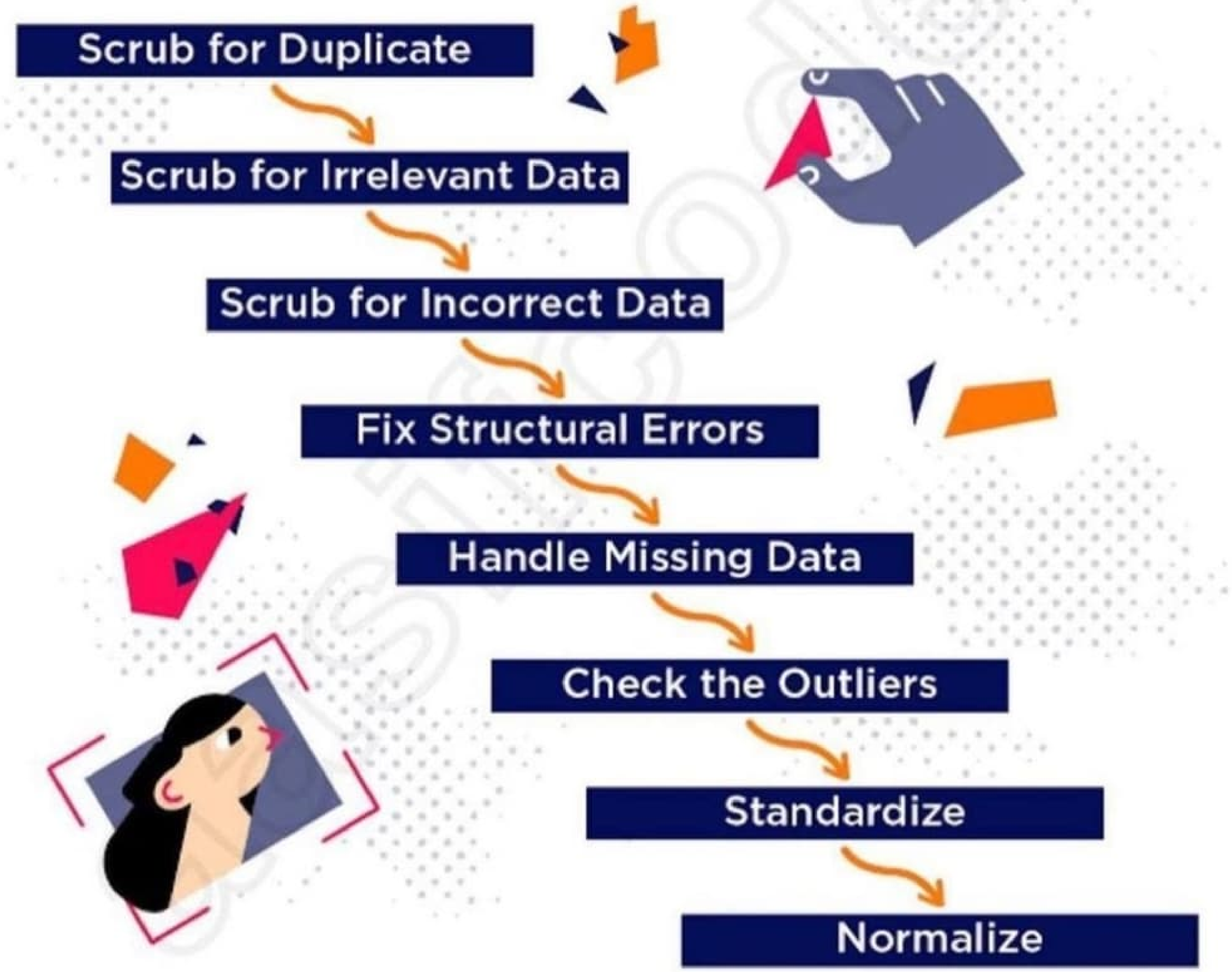
Data Cleaning Tools:

- **Microsoft Excel** (Popular data cleaning tool)
- **Programming languages** (Python, Ruby, SQL)
- **Data Visualizations** (To spot errors in your dataset)
- **Proprietary software** (OpenRefine, Trifacta, etc)

Benefits of Data Cleaning:

- Avoiding mistakes
- Improving productivity
- Avoiding unnecessary costs and errors
- Staying organized
- Improved mapping

• Data Cleaning Process •



Scrub for Duplicate

Scrub for Irrelevant Data

Scrub for Incorrect Data

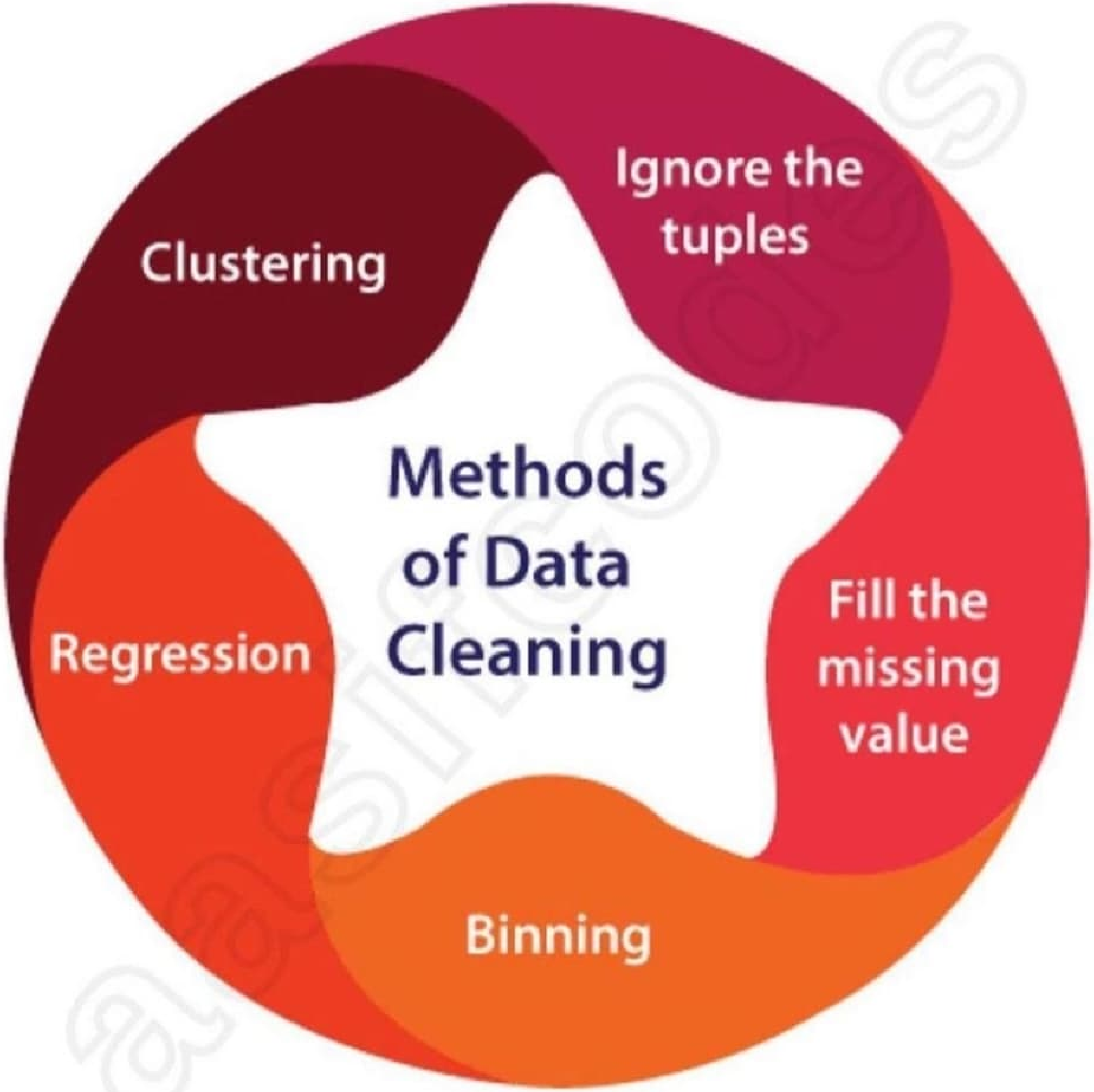
Fix Structural Errors

Handle Missing Data

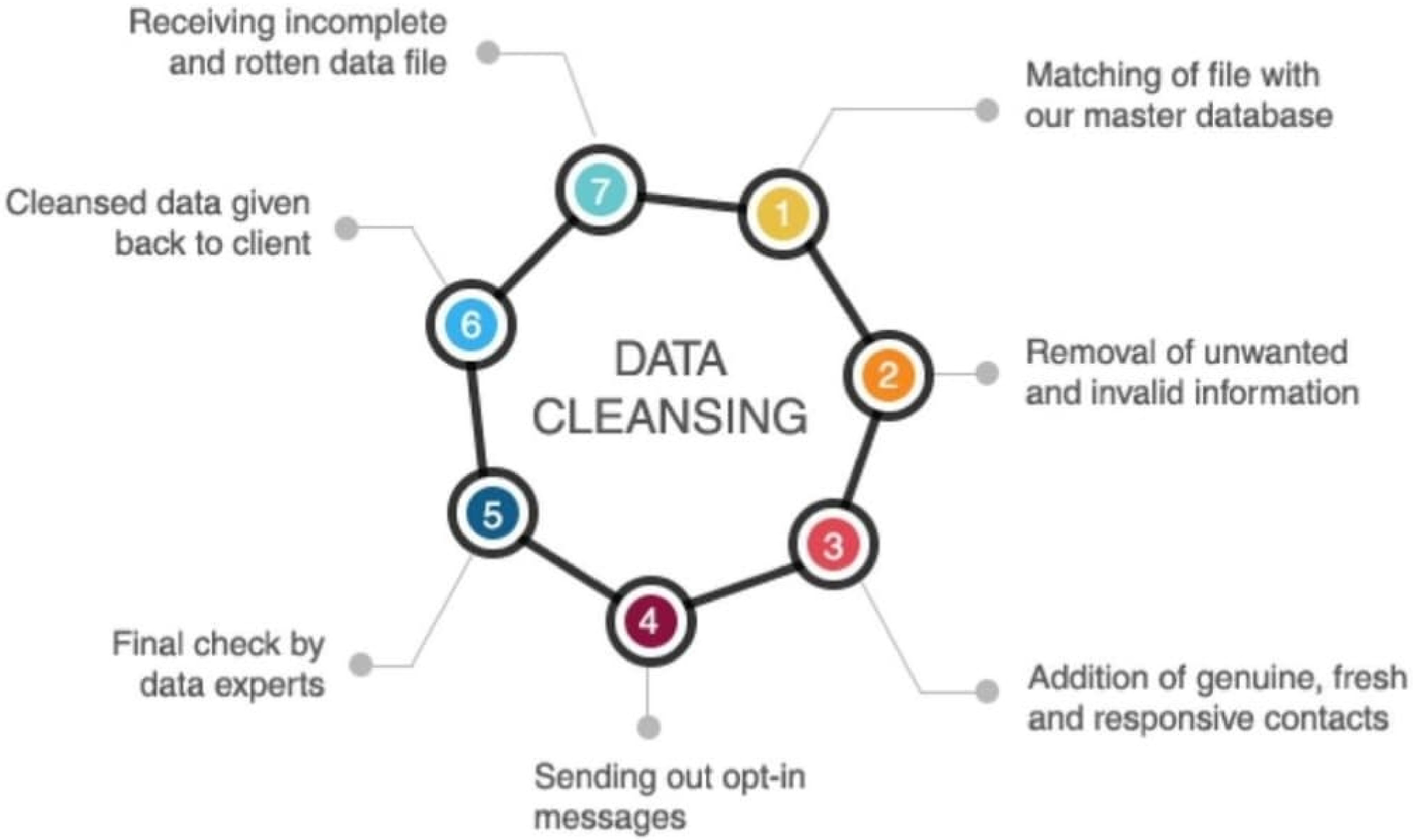
Check the Outliers

Standardize

Normalize



Methods of Data Cleaning
Ignore the tuples
Fill the missing value
Binning
Regression
Clustering


Receiving incomplete and rotten data file
7
Matching of file with our master database
1
Cleansed data given back to client
6
DATA CLEANSING
Removal of unwanted and invalid information
2
5
3
Final check by data experts
4
Addition of genuine, fresh and responsive contacts
Sending out opt-in messages