16 January 2023

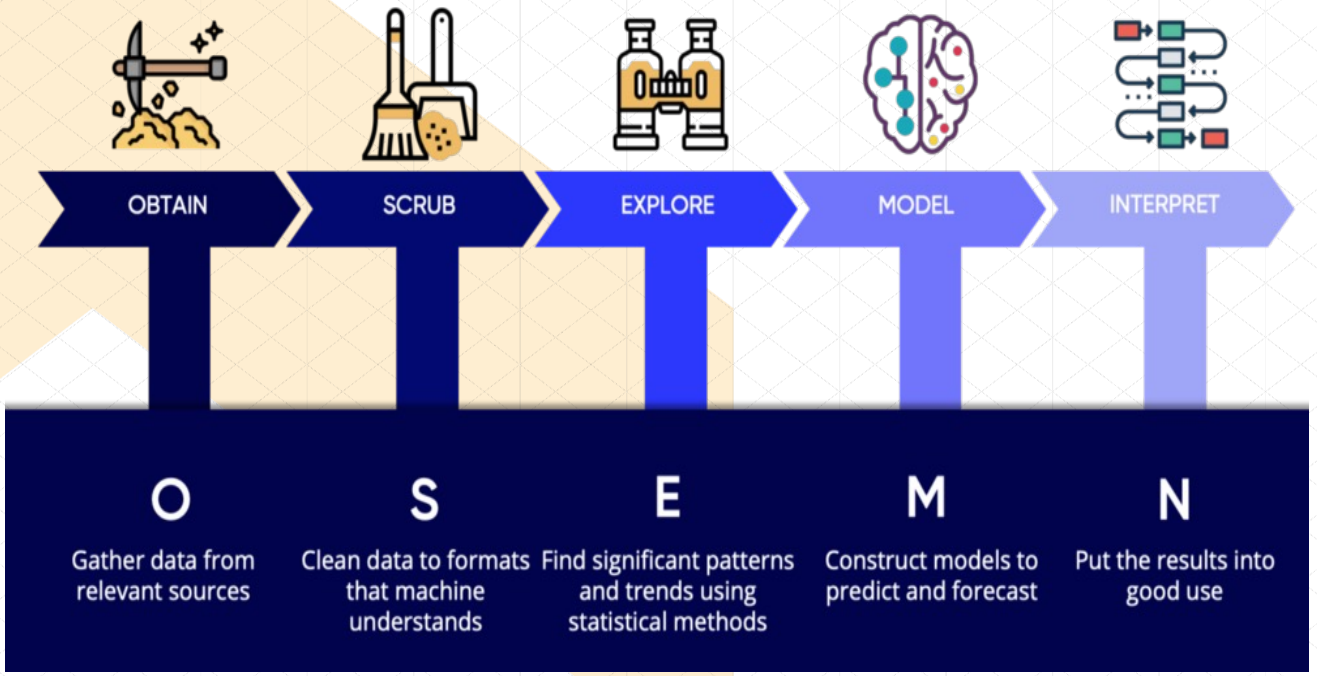
UNIVERSITY OF GLOUCESTERSHIRE

CT4031
Maths for Data Science

Week 3 – Data Cleaning

Data science process

- a systematic process used by Data Scientists to analyze, visualize and model large amounts of data.
- helps in discovering hidden patterns of structured and unstructured raw data.
- helps in turning a problem into a solution by treating the business problem as a project.

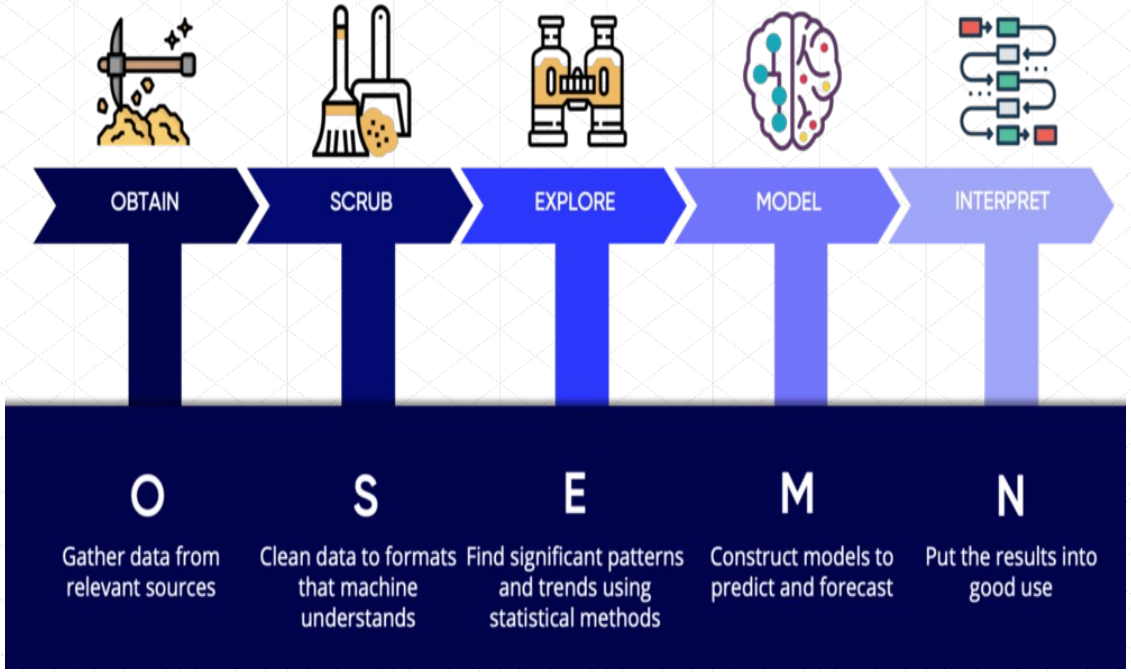


https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Source: https://bit.ly/3rlxxqV

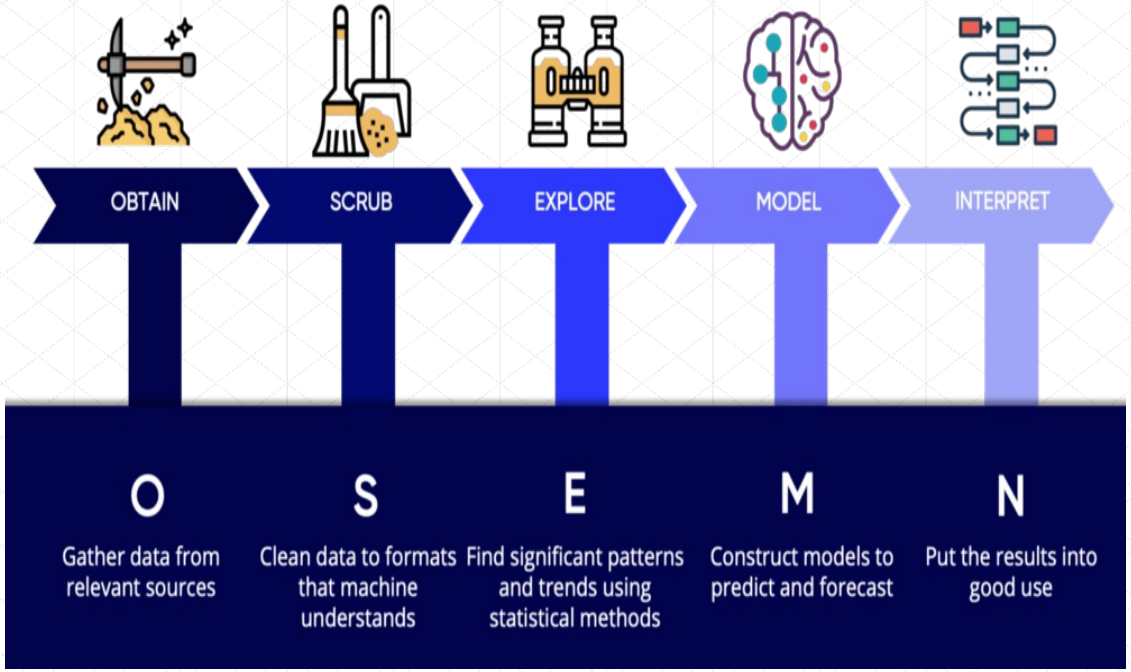
Data science process: Obtain Raw Data

- First Step
- Collect data from variety of sources
  - query databases
  - Excel sheets
  - scrape from the websites using web scraping tools such as Beautiful Soup.
  - Web APIs.
  - Kaggle
  - Flat text files
    - CSV (Comma Separated Value)
    - TSV (Tab Separated Values)
- **Skills required**
  - Database management (MySQL, PostgreSQL or MongoDB)
  - Big data tools (Apache Hadoop, Spark or Flink.)



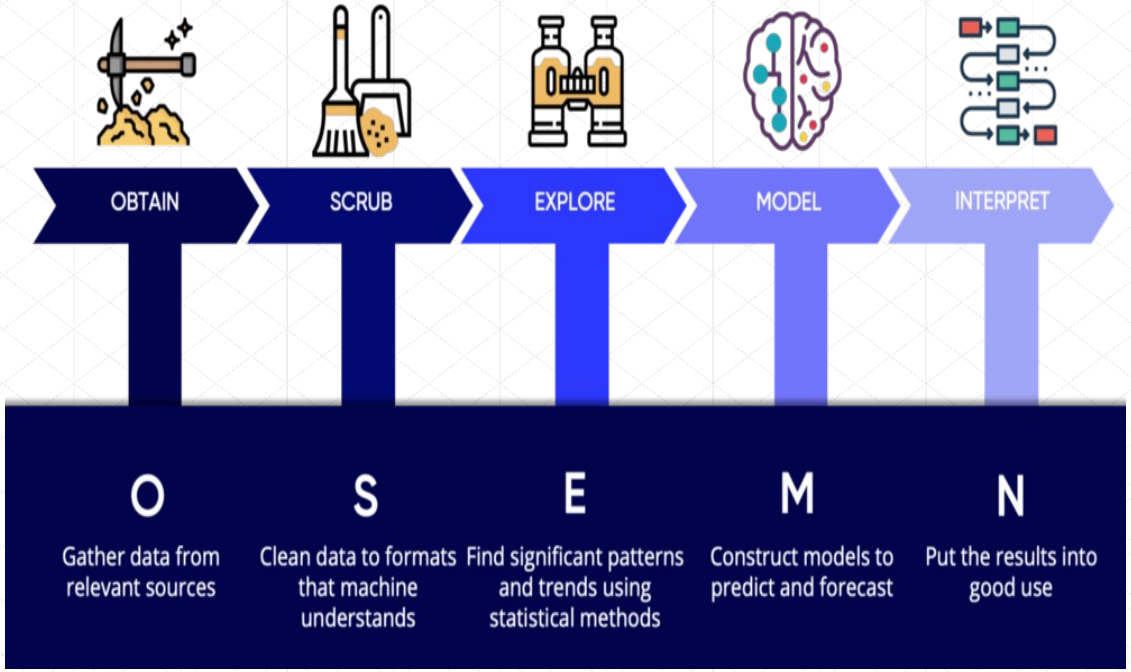
Data science process: Scrub Data

- Garbage in Garbage out
- “clean” and filter the data
- Convert the data from one format to another and consolidate everything into one standardized format across all data.
- extracting and replacing value
- Missing data
- split, merge and extract columns
- Skills Required
  - Python or R
  - Open-sourced tool: OpenRefine
  - Enterprise software: SAS Enterprise Miner
  - Data mining tools: Hadoop, Map Reduce or Spark



Data science process: Explore data

- Foundation stone
- a critical first step in analyzing the data from an experiment.
- Like an investigation carried out by a detective.
- Reveals the true nature of data.
- In EDA, the role of the researcher is to explore the data in as many ways as possible until a plausible “story” emerges.
- Confirm if the data is making sense in the context of a problem domain.
- If not collect more data and change the strategy.



Data science process: Explore data

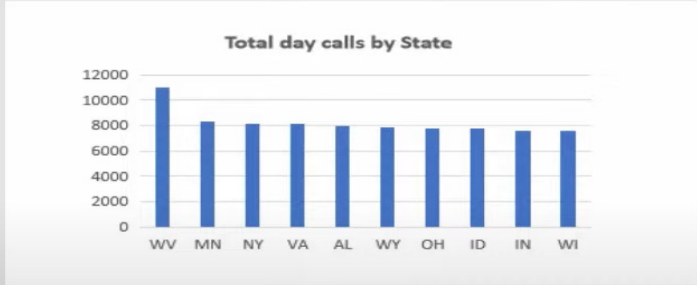
- Uncover and resolve data quality issues.
- It is highly unlikely that you will receive a clean data. So EDA is helpful in fixing the data cleaning issues such as missing data, duplicate values, incorrect values, Anomalies, incorrect data types.
- Get information about the data summary.
  - Mean, Median, Mode, Variance, Skewness, Range, Minimum, Maximum, sum, count, standard deviation
- Drop unwanted columns and derive new variables
  - Age=current year-DoB

Simple Tabulation

Area code	Customer Count	Customer Count Percent
408	838	25.14%
415	1655	49.65%
510	840	25.20%
Grand Total	**3333**	**100.00%**

Area code 415 has the highest number of customers.

Bar charts for categorical data.



West Virginia has a strong customer base.

Data science process: Explore data

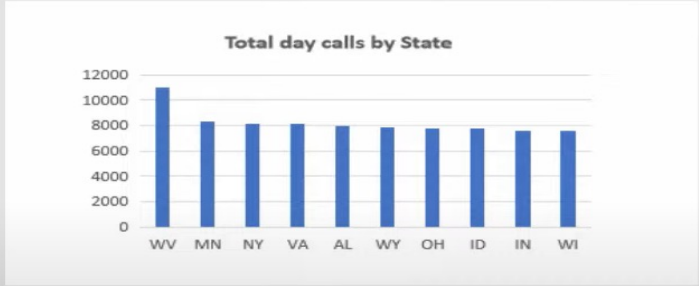
- Skills Required
  - Python
  - Numpy
  - Matplotlib
  - Pandas
  - Scipy
  - Statistics
  - Data visualization.

Simple Tabulation

Area code	Customer Count	Customer Count Percent
408	838	25.14%
415	1655	49.65%
510	840	25.20%
Grand Total	**3333**	**100.00%**

Area code 415 has the highest number of customers.

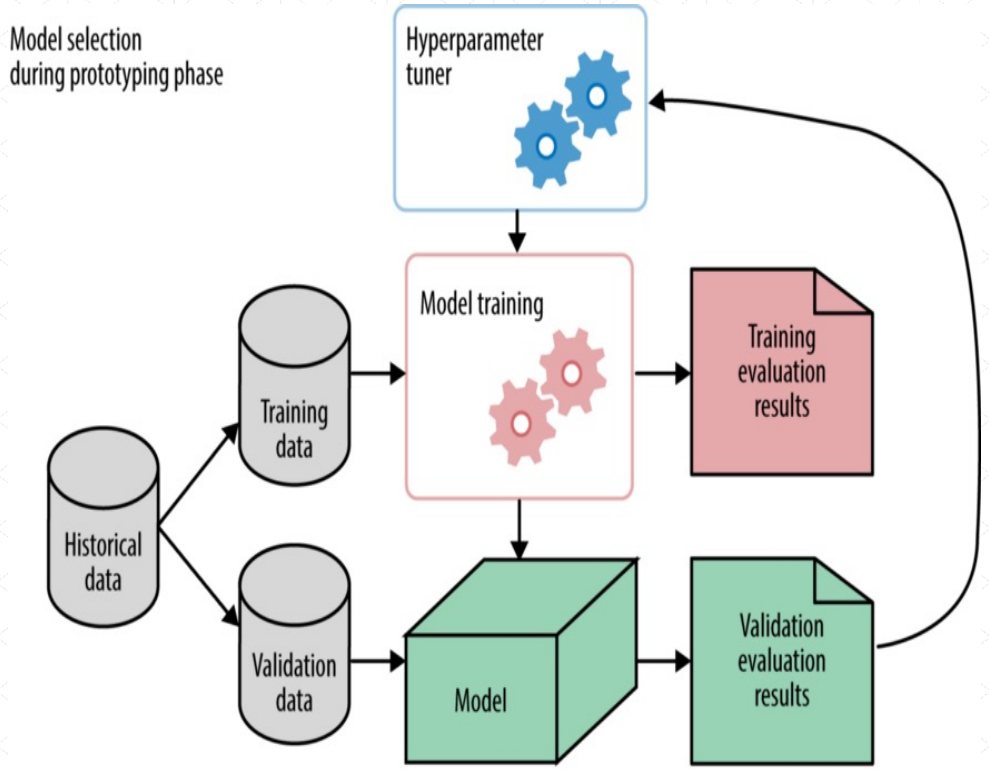
Bar charts for categorical data.



West Virginia has a strong customer base.

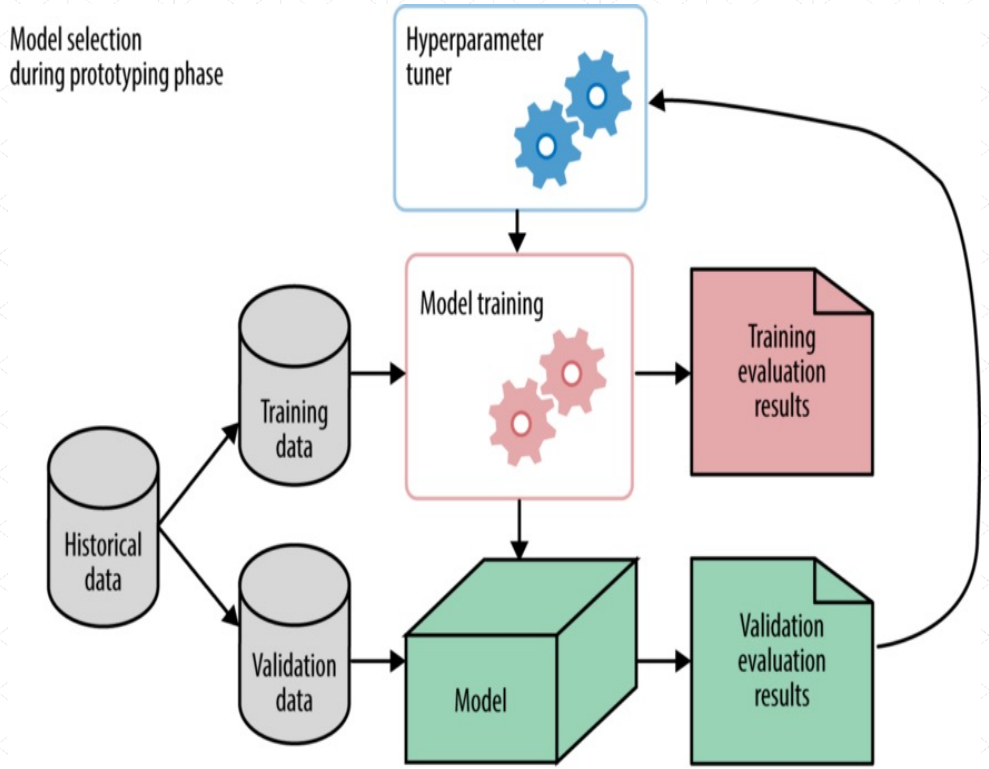
Data science process: Model Data

- "where the magic happens".
- Prediction model
- Machine learning Algorithms
  - Supervised
  - Unsupervised
  - Weather forecast system using Naïve Bayesian Network.
- Sci-kit Learn (Python)

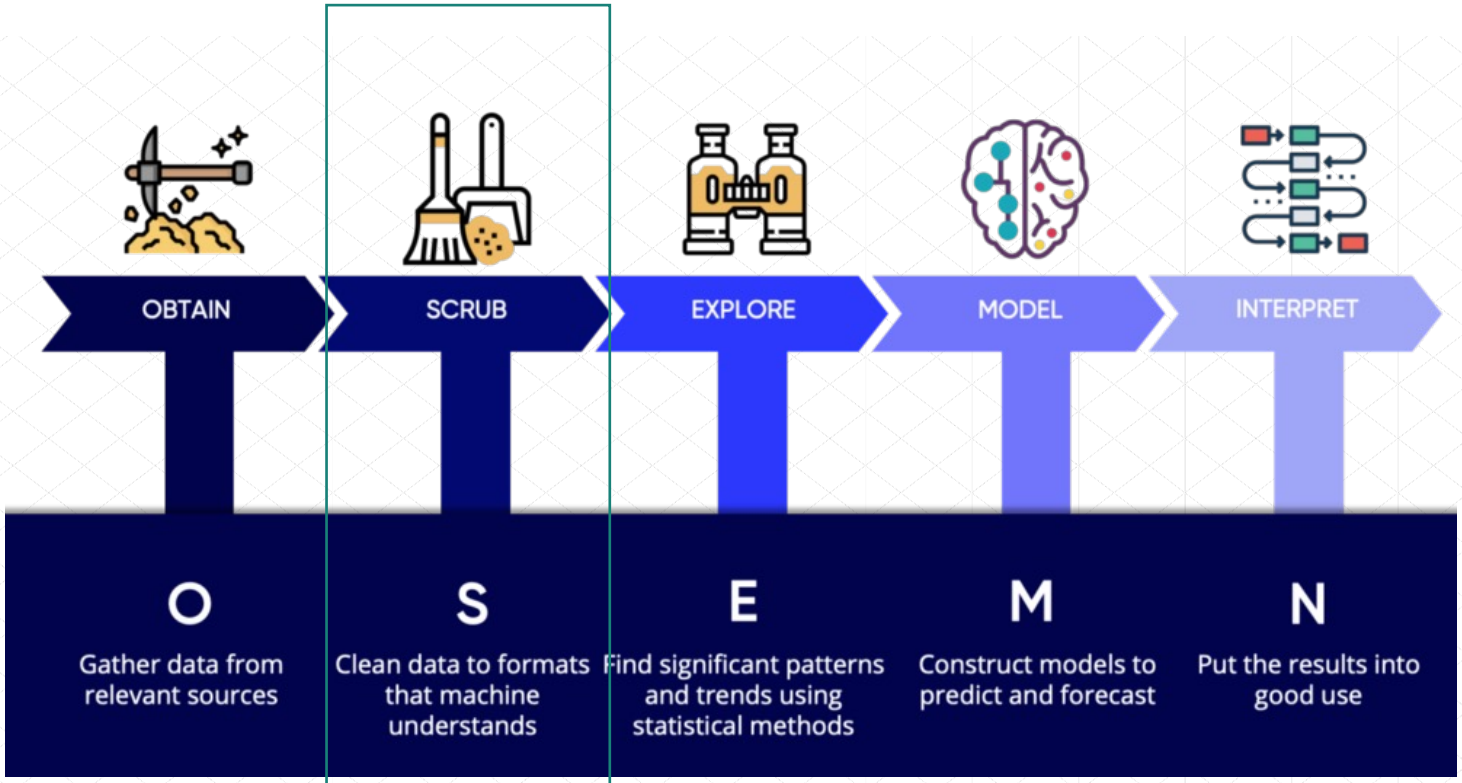


Data science process: Interpreted Data

- Final and most crucial step
- Interpreting data refers to the presentation of your data to a non-technical layman
- visualise your findings accordingly
- Skills Required: To be able to tell a clear and actionable story.
- need strong business domain knowledge to present your findings in a way that can answer the business questions you set out to answer, and translate them into actionable steps.
  - If your presentation does not trigger actions in your audience, it means that your communication was not efficient.
  - Remember that you will be presenting to an audience with no technical background, so the way you communicate the message is key.



Data science process



Source: https://bit.ly/3rlxxqV

- Data Quality
- Data Cleaning
- Data Transformation

Data quality Features

- Data quality is the measure of how well suited a data set is to serve its specific purpose.
- Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness, and timeliness.

Data quality Features: Accuracy

- **Accuracy:** The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.
- **Completeness**: Completeness is a measure of the data's ability to effectively deliver all the required values that are available.
- **Consistency:** Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in difference locations should not conflict with one another.
- **Validity**: Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.
- **Uniqueness**: Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.
- **Timeliness:** Timely data is data that is available when it is required. Data may be updated in real time to ensure that it is readily available and accessible.

Data quality

Generally, you have a problem if the data doesn't mean what you think it does, or should.

Many sources and manifestations.

Data quality problems are expensive and pervasive

- DQ problems cost hundreds of billion £££ each year.
- Resolving data quality problems is often the biggest effort in a data science study.

example

T.Das|97336o8327|24.95|Y|-|0.0|1000

Ted J.|973-360-8779|2000|N|M|NY|1000

Can we interpret the data?

- What do the fields mean?
- What is the key? The measures?

HOW/WHY data loses quality

??

HOW/WHY data loses quality

Loss of data quality can occur at many stages:

- At the time of collection
- During digitisation
- During documentation
- During storage and archiving
- During analysis and manipulation
- At time of presentation
- And through the use to which they are put

HOW/WHY data loses quality

Loss of data quality can occur at many stages:

- At the time of collection
- During digitisation
- During documentation
- During storage and archiving
- During analysis and manipulation
- At time of presentation
- And through the use to which they are put

What else?

Common problems in a dataset

Problem	Example
Illegal Values	DoB = 30_2_21
Uniqueness violations	Name: John, Id: 3 Name: Peter, Id: 3
Noise	Network monitoring
Different patterns	Age: 18, eighteen, old enough
Missing values	Unavailable values
Abbreviations	John Smith, J. Smith
Environment problem	Operation Vs Analytical
Different NULL values	NULL, 0 and " "

Common problems in a dataset

Problem	Example
Illegal Values	DoB = 30_2_21
Uniqueness violations	Name: John, Id: 3 Name: Peter, Id: 3
Noise	Network monitoring
Different patterns	Age: 18, eighteen, old enough
Missing values	Unavailable values
Abbreviations	John Smith, J. Smith
Environment problem	Operation Vs Analytical
Different NULL values	NULL, 0 and " "

What else?

Data cleaning

How do we clean a dataset?

Data cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

Data cleaning steps

- Analysis
- Defining transformation and/or mapping rules
- Apply transformation
- Verification

Analysis

Data profiling

- Analyse the different data types
- Examine the dataset to find out how the attributes vary

Detect errors and inconsistencies

Manual and automated inspections

- When is manual a better choice?
- When is automated a better choice?

Defining transformation and/or mapping rules

Plan what to do with:

Problem	Example
Illegal Values	DoB = 30_2_21
Uniqueness violations	Name: John, Id: 3 Name: Peter, Id: 3
Noise	Network monitoring
Different patterns	Age: 18, eighteen, old enough
Missing values	Unavailable values
Abbreviations	John Smith, J. Smith
Environment problem	Operation Vs Analytical
Different NULL values	NULL, 0 and " "

Defining transformation and/or mapping rules

Any ideas?

Problem	Example
Illegal Values	DoB = 30_2_21
Uniqueness violations	Name: John, Id: 3 Name: Peter, Id: 3
Noise	Network monitoring
Different patterns	Age: 18, eighteen, old enough
Missing values	Unavailable values
Abbreviations	John Smith, J. Smith
Environment problem	Operation Vs Analytical
Different NULL values	NULL, 0 and " "

Apply transformation

Implement the plans for cleaning the dataset.

Verification

- In this phase we test and evaluate the transformation plans we made
- Without this, we may end up making the data dirtier rather than cleaner
- Manual and automated

Common practices

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

Locate and identify the different data types and change when needed.

Parsing

ID	AGE	FEES	SITE	ACTIVE	TEMPERATURE
1	18 years	52	London	YES	68
2	23 years	51.1	Landon	YES	20
3	33 years	67.7	Cheltenham	NO	32
4	22 years	90	Cardiff	ACTIVE	0
5	44 years	16	Bristol	ACTIVE	212
6	56 years	88	Bristol	NO	100
7	22 years	90	Cardiff	ACTIVE	0
8	44 years	16	Bristol	ACTIVE	212
9	22 years	90	Cardiff	ACTIVE	0
10	44 years	16	Bristol	ACTIVE	212
11	22 years	90	Cardiff	ACTIVE	0
12	44 years	16	Bristol	ACTIVE	212

Correcting

Adjusting incorrect, missing or invalid data.

Correcting

ID	AGE	FEES	SITE	ACTIVE	TEMPERATURE
1	18 years	52	London	YES	68
2	23 years	51.1	Landon	YES	20
3	33 years	67.7	Cheltenham	NO	32
4	22 years	90	Cardiff	ACTIVE	0
5	44 years	16	Bristol	ACTIVE	212
6	56 years	88	Bristol	NO	100
7	22 years	90	Cardiff	ACTIVE	0
8	44 years	16	Bristol	ACTIVE	212
9	22 years	90	Cardiff	ACTIVE	0
10	44 years	16	Bristol	ACTIVE	212
11	22 years	90	Cardiff	ACTIVE	0
12	44 years	16	Bristol	ACTIVE	212

Standardizing

Transform data into a preferred format/structure to work.

Standardizing

Examples:

- Change Boolean values to 0 and 1
- Change temperature to ºC
- Change KM to Miles
- Change inches to cm

Standardizing

ID	AGE	FEES	SITE	ACTIVE	TEMPERATURE
1	18 years	52	London	YES	68
2	23 years	51.1	Landon	YES	20
3	33 years	67.7	Cheltenham	NO	32
4	22 years	90	Cardiff	ACTIVE	0
5	44 years	16	Bristol	ACTIVE	212
6	56 years	88	Bristol	NO	100
7	22 years	90	Cardiff	ACTIVE	0
8	44 years	16	Bristol	ACTIVE	212
9	22 years	90	Cardiff	ACTIVE	0
10	44 years	16	Bristol	ACTIVE	212
11	22 years	90	Cardiff	ACTIVE	0
12	44 years	16	Bristol	ACTIVE	212

Matching

- Search for duplicated values to eliminate them.
- Usually based on key attributes, however, it could be used with a combination of rules.

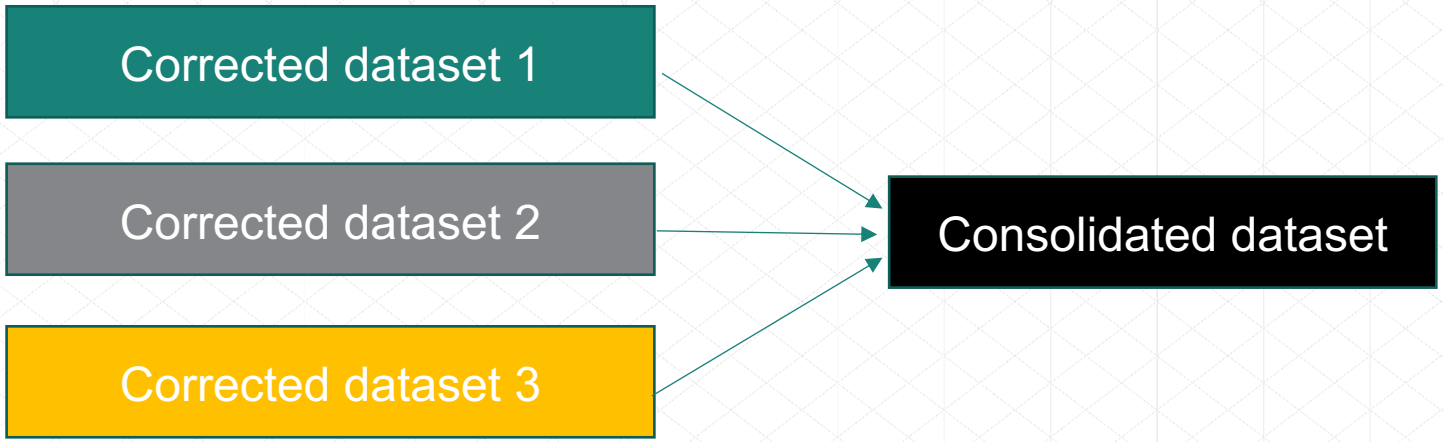
Matching

ID	AGE	FEES	SITE	ACTIVE	TEMPERATURE
1	18 years	52	London	YES	68
2	23 years	51.1	Landon	YES	20
3	33 years	67.7	Cheltenham	NO	32
4	22 years	90	Cardiff	ACTIVE	0
5	44 years	16	Bristol	ACTIVE	212
6	56 years	88	Bristol	NO	100
7	22 years	90	Cardiff	ACTIVE	0
8	44 years	16	Bristol	ACTIVE	212
9	22 years	90	Cardiff	ACTIVE	0
10	44 years	16	Bristol	ACTIVE	212
11	22 years	90	Cardiff	ACTIVE	0
12	44 years	16	Bristol	ACTIVE	212

Consolidating (AKA merging)

Combining different adjusted datasets to create a valid and consolidated dataset.

Consolidating (AKA merging)



Post sessional work

- What is Exploratory Data Analytics?
- Discuss three different ways of visualising a dataset.

References

Dasu, T., Johnson, T., 2003. Exploratory data mining and data cleaning. John Wiley & Sons.

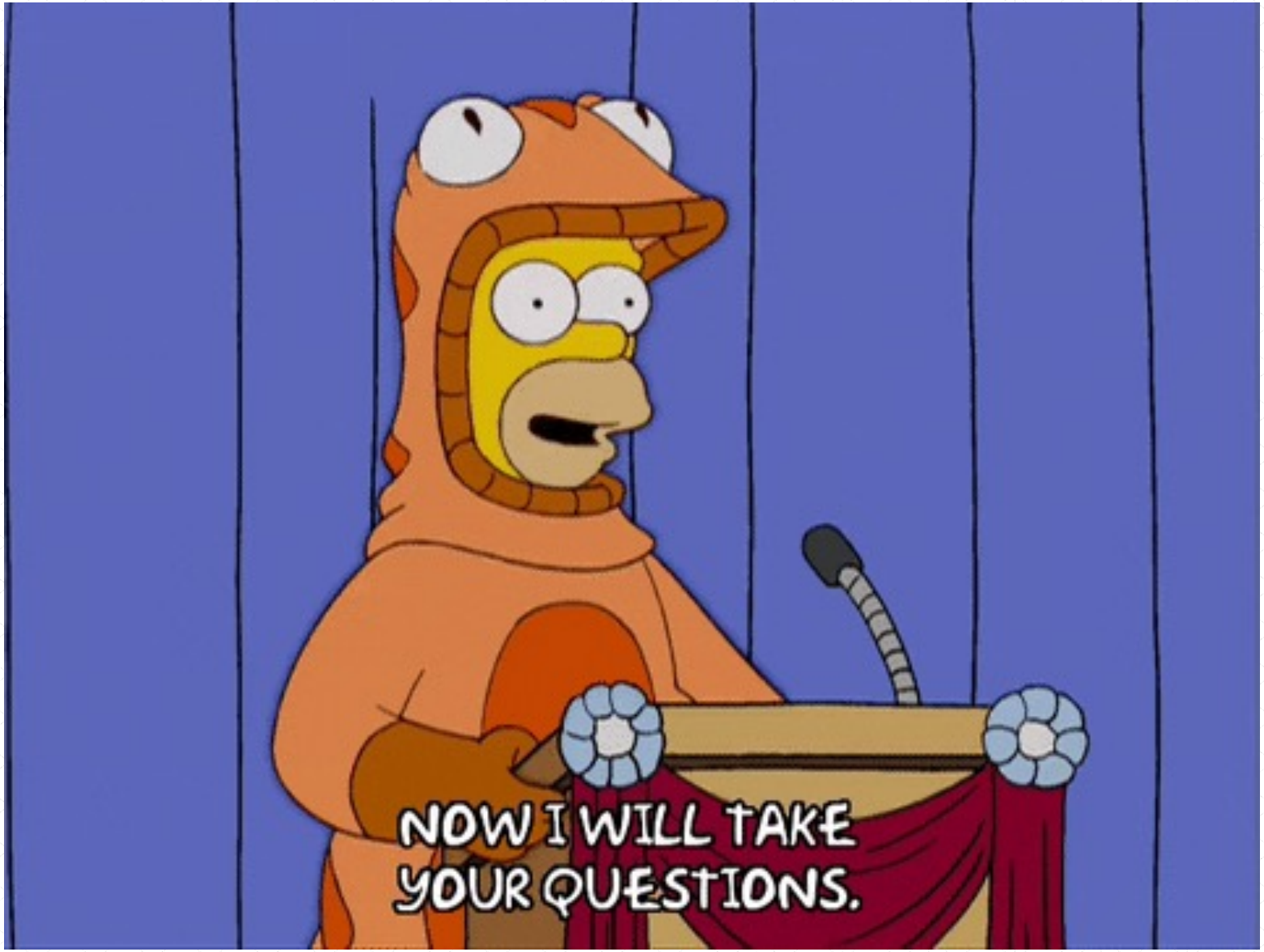
Johnson, T., Dasu, T., 2003. T3: Data Quality and Data Cleaning: An Overview.

What is Data Cleaning? [WWW Document], n.d. . Sisense. URL https://www.sisense.com/glossary/data-cleaning/ (accessed 2.22.21).

https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Next Session!

- EDA


NOW I WILL TAKE
YOUR QUESTIONS.