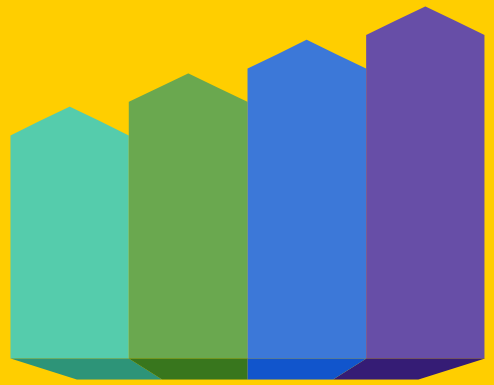


Introduction to Data Wrangling

WRANGLING?

What is Data Wrangling?
Who will do ?
When will we do?
Where is doing it?
Why must we do?
How is the process?

About data wrangling

What is Data Wrangling?

Processing data from raw data until it is neat and ready to be analysed/modelled.

About data wrangling

Who will do?

Data Analyst/Data Scientist/Data Engineer

About data wrangling

When will we do?

From the beginning of getting the data to before analyzing the data and modeling it.

About data wrangling

Where is doing it?

We use programing language, As Python or R.

About data wrangling

Why must we do?

- Not all raw data is ready for analysis.
- Not all models can be implemented on raw data.
- Make it easy for someone to analyse data
- Organize data that is difficult to understand

About data wrangling

How is the process?

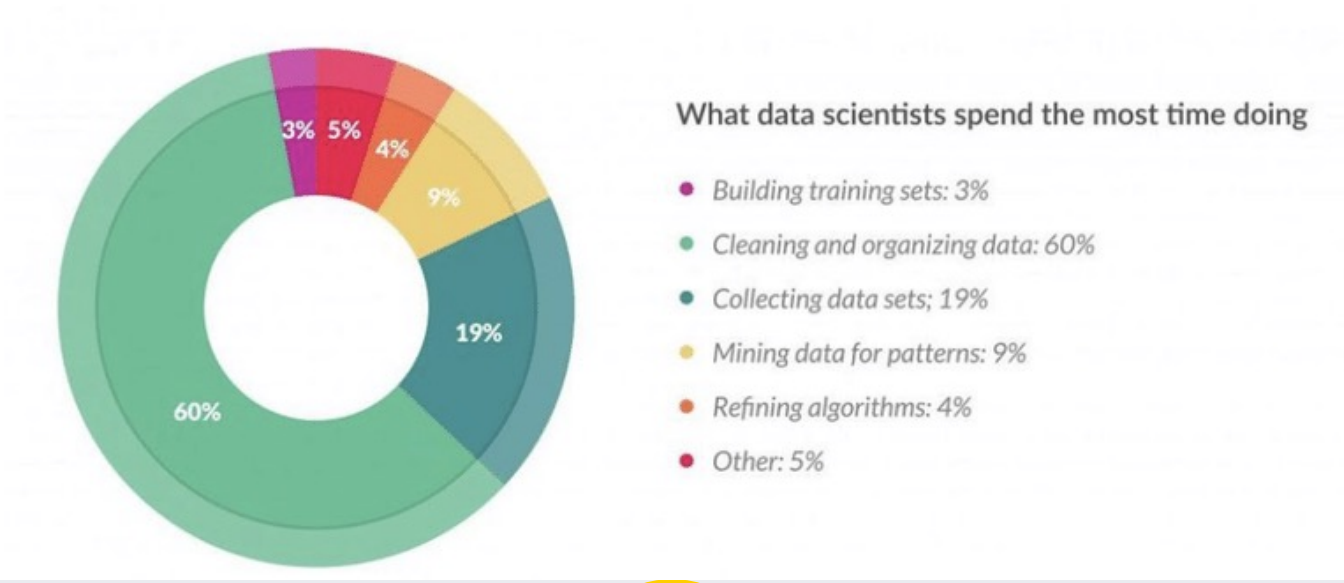
- Spot variables and observations
- Derive new variables and observations
- Reshape into best format
- Join multiple dataset
- Group-wise summarize

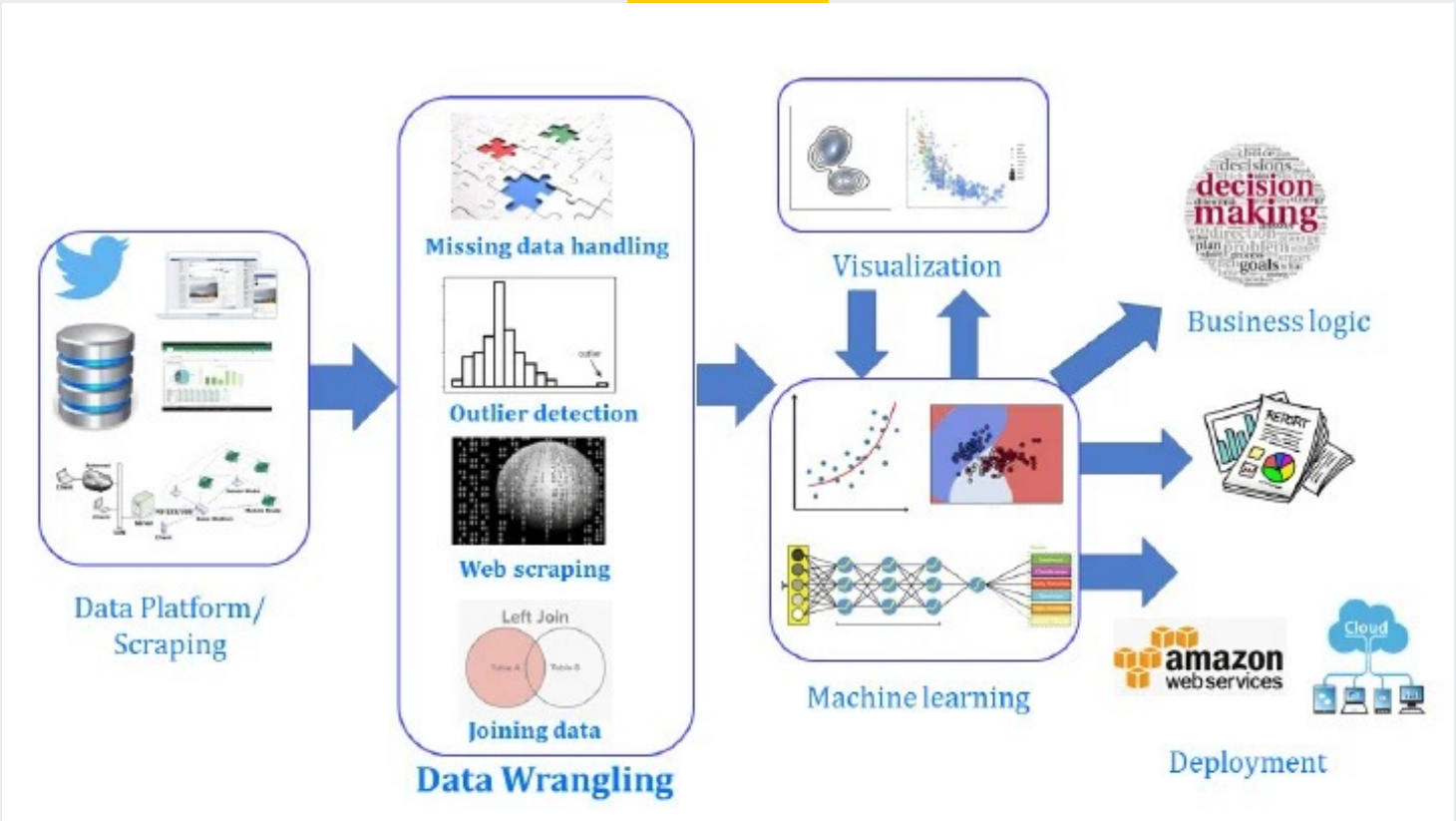
The New York Times

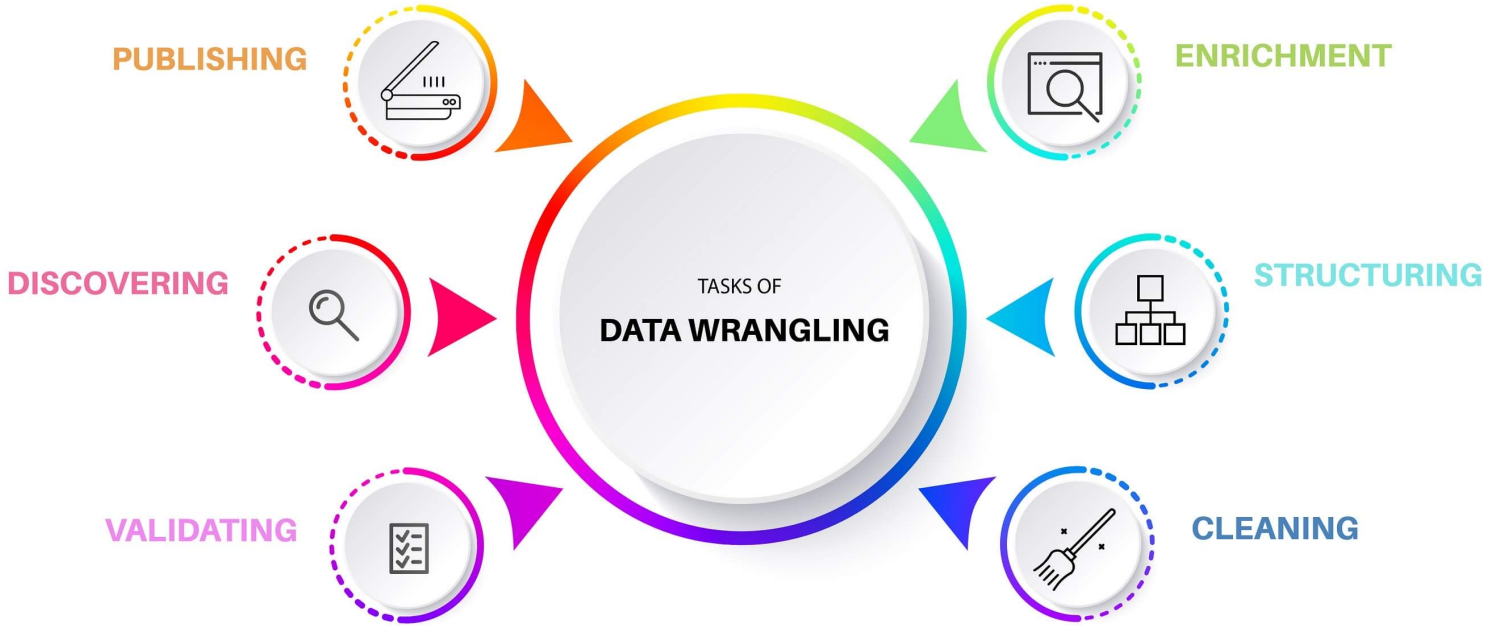
For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html


3%
5%
4%
9%
19%
60%
What data scientists spend the most time doing
Building training sets: 3%
Cleaning and organizing data: 60%
Collecting data sets; 19%
Mining data for patterns: 9%
Refining algorithms: 4%
Other: 5%


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


PUBLISHING
ENRICHMENT
DISCOVERING
TASKS OF
DATA WRANGLING
STRUCTURING
VALIDATING
CLEANING


Data Wrangling
DEFINE GOALS
1. DISCOVERING
2. STRUCTURING
3. CLEANING
4. ENRICHING
5. VALIDATING
6. PUBLISHING
ANALYZE
VISUALIZE

1 Discovering

Scraping and understand the pattern or trend

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

Discovering or discovery is the first step to data wrangling – it's about getting an overview of your data

2 Structuring

Structuring File, Variable and Sample

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

Next you'll need to organize or structure your data. Some data, that's entirely entered properly into spreadsheets. But sometimes data is presented in unstructured

3 Cleaning

Handling Missing Value, Data Transformation, Remove Duplication and Discretization.

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

The process of removing irrelevant data, errors, and inconsistencies that could skew your results.

4 Enriching

Merge and Combine Dataset

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

Data enriching is necessary to add variants of the data. Rich data will produce a better model because it has a smaller sample error and is closer to the population.

5 Validating

Normalization and Standarization

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

Data validation is the process of authenticating your data and confirming that it is standardized, consistent, and high quality. Verify that it is clean and regularly structured.

6 Publishing

Shere and ready to modelled

Department of Data Science
Institut Teknologi Telkom Surabaya

Big concept

This could mean sharing across your business or organization for different analytical needs, or uploading it to machine learning programs to train new models or run through pre-trained models.

Python Library for Data Wrangling

- Numpy (already installed in anaconda)
- Pandas (already installed in anaconda)
- BeautifulSoup (you need install first)

  pip install bs4

  or

  conda install bs4

Inspact your computer?