Share of market activity
Changes in the activity of the active and passive market is uncertain. Established positive
Projected sales of main products in 2013
Passive market share
350
300
250
200
150
100
50
0
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
25%

Data Analytics

DAP is **HOT** right now!

Companies are looking for people who can maintain their data and analyze it to:

- Effectively make decisions
- Develop new revenue
- Decrease Operational costs

Who are Data Analysts?

- Take data from their organizations
- Use that data to answer questions
- Communicate the results

Some Titles

- Business Analyst; Operations Analyst
- Business Intelligence Analyst
- Marketing Analyst; Database Analyst

Python Learning Materials

- Python For Everyone -- PY4E
- This program was created by Dr. Charles Severance (a.k.a. Dr. Chuck).
- He’s a Clinical Professor at the *University of Michigan School of Information Technology*.
- **https://www.py4e.com/**

The end project: Your Capstone

Showcases what you have covered over this 12-week period

15 – 20 minutes in length

You need to come up with a topic

Create a question with that topic

Locate raw data sets on that topic

Clean and analyze data to get an answer

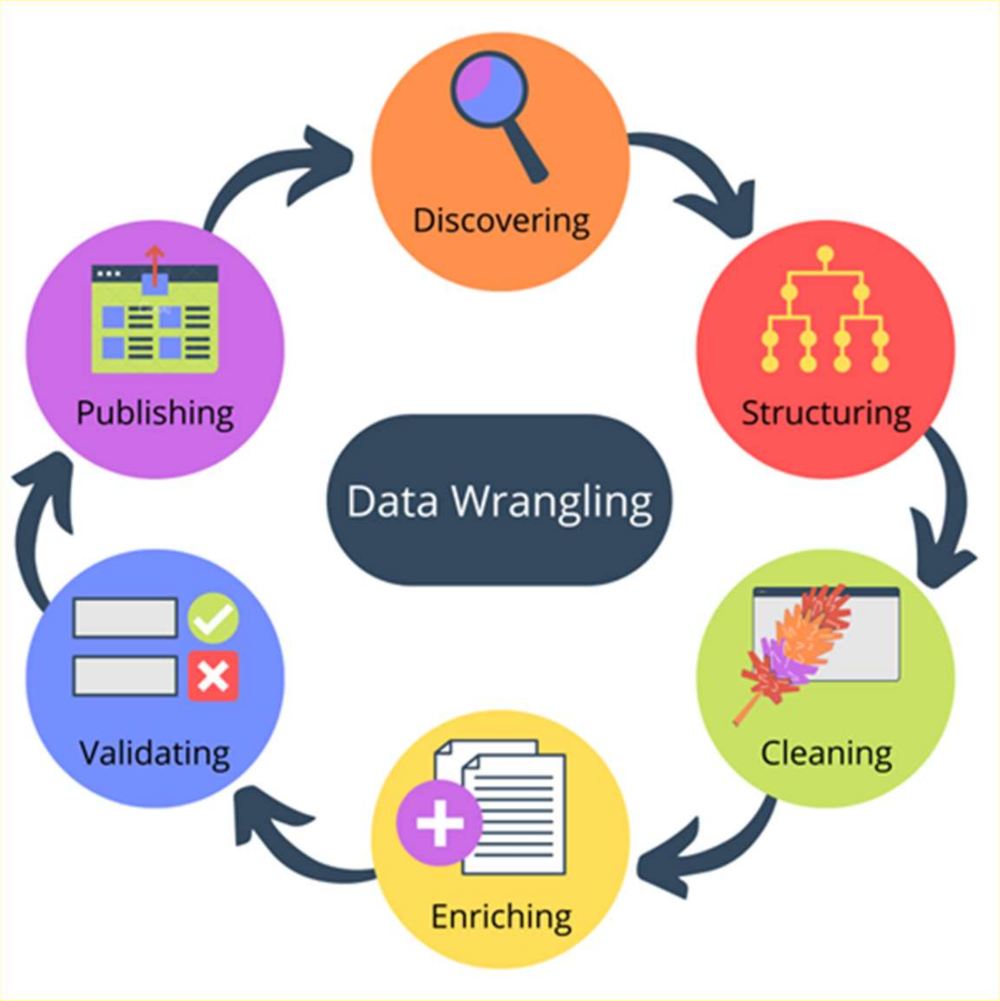
Create and present your story!

Data Wrangling

The process of cleaning and unifying messy and complex data sets for easy access and analysis.

Organizing and processing data

Data professionals spend as much as **80% of their time** in the data wrangling process.



Data Wrangling

Step #1 – Discovery

Step #2 – Structuring

Step #3 – Cleaning

Step #4 – Enriching

Step #5 – Validating

Step #6 – Publishing

Examples of Wrangling

1. Joining together multiple data sets into one
2. Finding gaps in data and filling/deleting them
3. Getting rid of data that is unnecessary
4. Identifying extreme outliers and either explaining them or getting rid of them

Step #1 - Discovery

- Find data that addresses your question
- Become familiar with your data so that you know how you will end up using it.
- Identify trends, patterns and some data cells / sections that might cause issues in analysis.

Step #2 - Structuring

- Take your raw data and transform it to what you can work with

- Unstructured data is often text-heavy and contains things such as Dates, Numbers, ID codes, etc.

- Example: - When using info scrapped from a website, you might parse HTML code, pull out what you need, and discard the rest.

Step #3 - Cleaning

- Removes outliers that can potentially skew your results when analyzing the data

- Changes any null values and standardizes the data format to improve quality and consistency

- Identifies duplicate values, standardizes systems of measurements, fixes structural errors and typos, and validates the data to make it easier to handle

Step #4 - Enriching

- Deciding if you need to add to the data by combining raw data with additional data from other sources.

- Example: - Combining two or more databases of customer information to fill in gaps in the data

- Enriching the data is an optional step that you only need to take if your current data doesn’t meet your requirements.

Step #5 - Validating

- Making sure that the data that you have is of the quality necessary to complete your project.

- The rules of data validation require repetitive programming processes that help to verify the – Quality, Consistency, Accuracy, Security, and Authenticity of data

Step #6 - Publishing

- Creating your analysis and presenting it to the public.
- You can deposit the data into a new architecture or database.
- We will display our data story using Tableau.

Goals of Efficient Data Wrangling

- Show "deeper intelligence" by gathering data from several different sources
- Provide accurate, actionable data to clients, on time
- Reduces time spent collecting and organizing raw data
- Allow data scientists and analysts to focus on the analysis
- Provide intelligence for better decision-making by leaders

What's the Difference?

DATA WRANGLING

Changes the data's format by making the raw data into something more useable.

** Prepares data's structure for modeling **

DATA CLEANING

Removing data that will not help in analysis because it contains errors or misinformation.

** Enhances data's accuracy and integrity **

Crisp-DM

CRoss

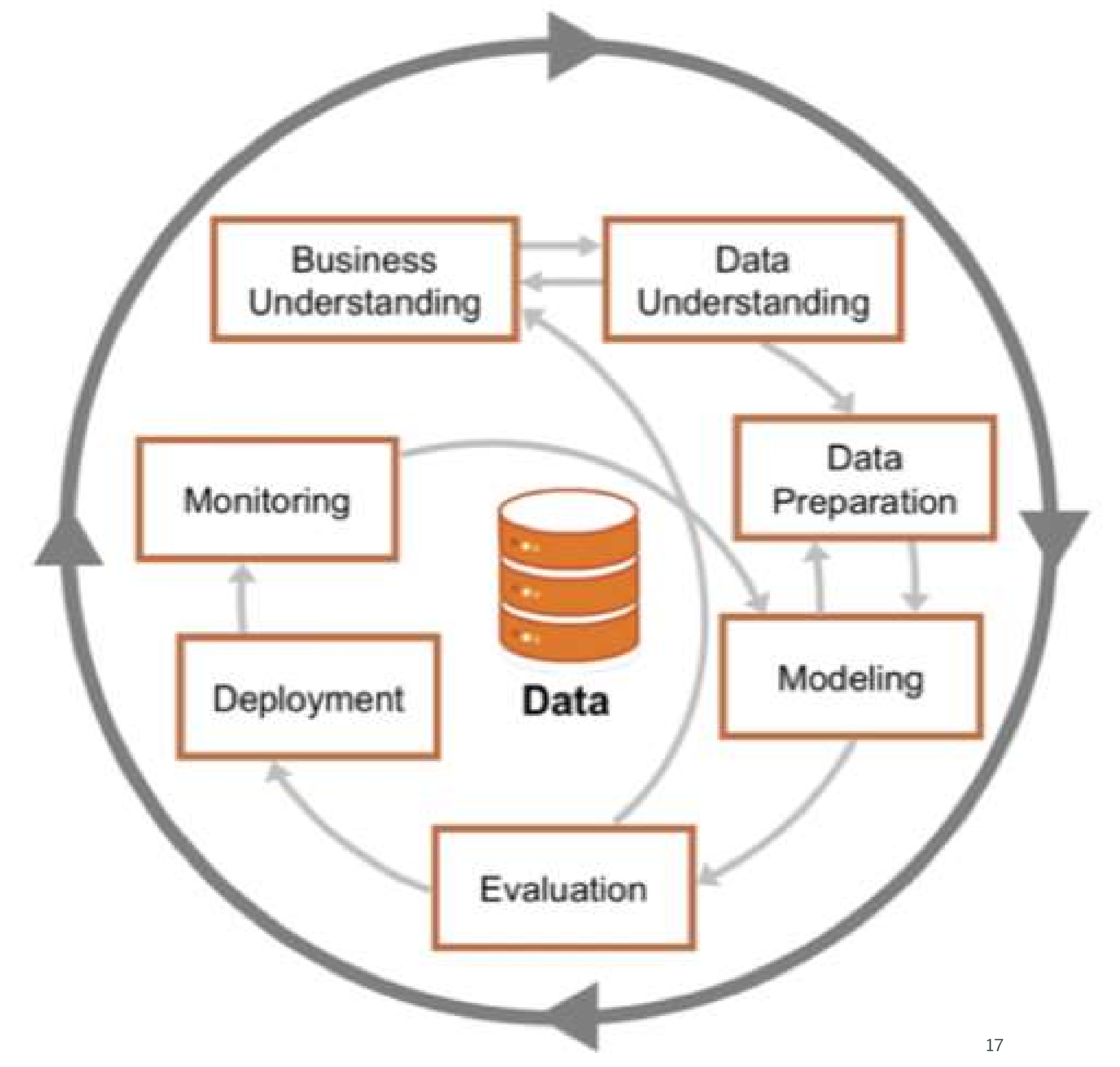
Industry

Standard

Process for

Data Mining

A process model with six phases that naturally describes the data science life cycle.



What is Crisp-DM?

https://www.datascience-pm.com/crisp-dm-2/

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a model that serves as the base for a data science process.

Published in 1999 to standardize data mining processes across industries, it is the most common methodology for data mining, analytics, and data science projects.

Data science teams that combine a loose implementation of CRISP-DM with overarching team-based **Agile** project management approaches will likely see the best results.

Data Science Life Cycle

Business Understanding
What does the business need?

Data Understanding
What data do we have / need? Is it clean?

Data preparation
How do we organize the data for modeling?

Modeling
What modeling techniques should we apply?

Evaluation
Which model best meets the business objectives?

Deployment
How do stakeholders access the results?

The end project: Your Capstone

Showcases what you have covered over this 12-week period

15 – 20 minutes in length

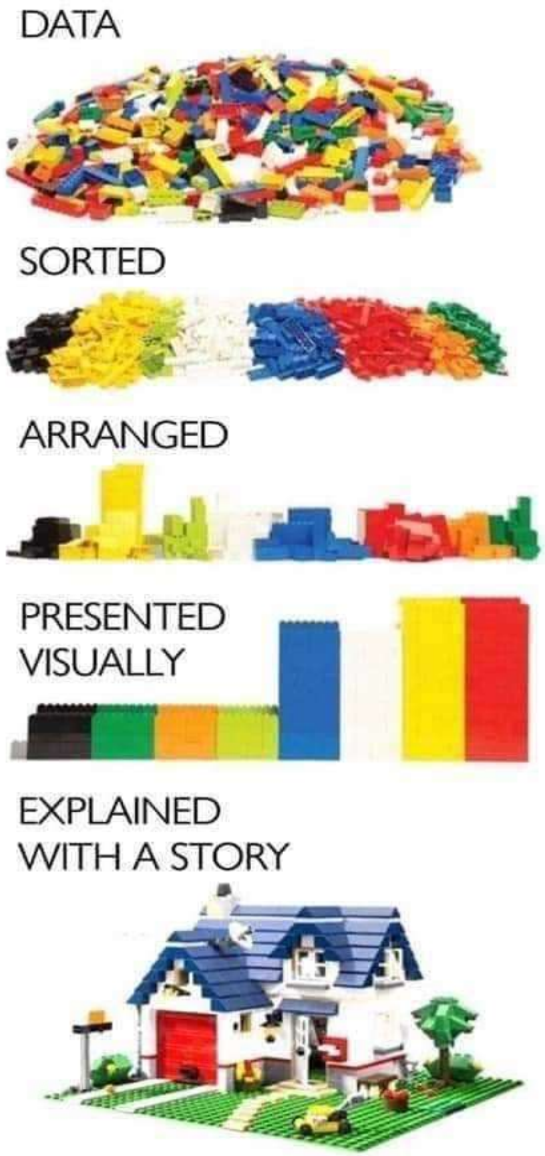
You need to come up with a topic

Create a question with that topic

Locate raw data sets on that topic

Clean and analyze data to get an answer

Create and present your story!



“Data are just summaries of thousands of stories – tell a few of those stories to help make the data meaningful.”

Quote by Chip and Dan Heat, authors of five best selling books, among them -- *“Making Numbers Count: The Art and Science of Communicating Numbers”, and “The Power of Moments: Why Certain Experiences Have Extraordinary Impact”,.*