Your Last Introduction to Data Science

By / Omar Abd El-Sameaa
Reservoir Engineer at Bapetco

Contents

- Why to Learn Data Science as a Petroleum Engineer.
- Data Science Lifecycle.
- Learning Roadmap.
- Discussion.

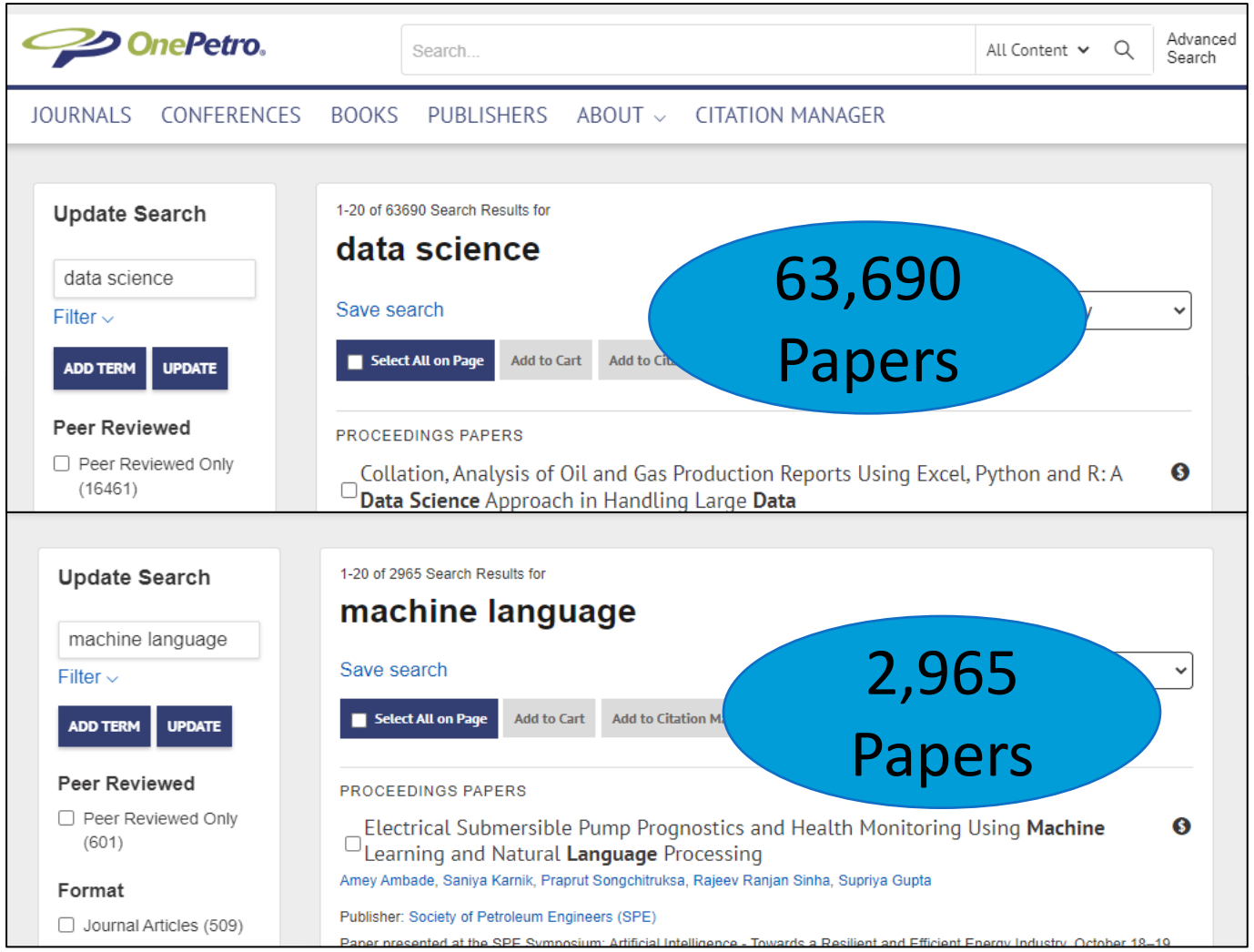
Why to Learn DS?

- No career shift.
- High level vs. low level programming languages.
- No coding platforms.

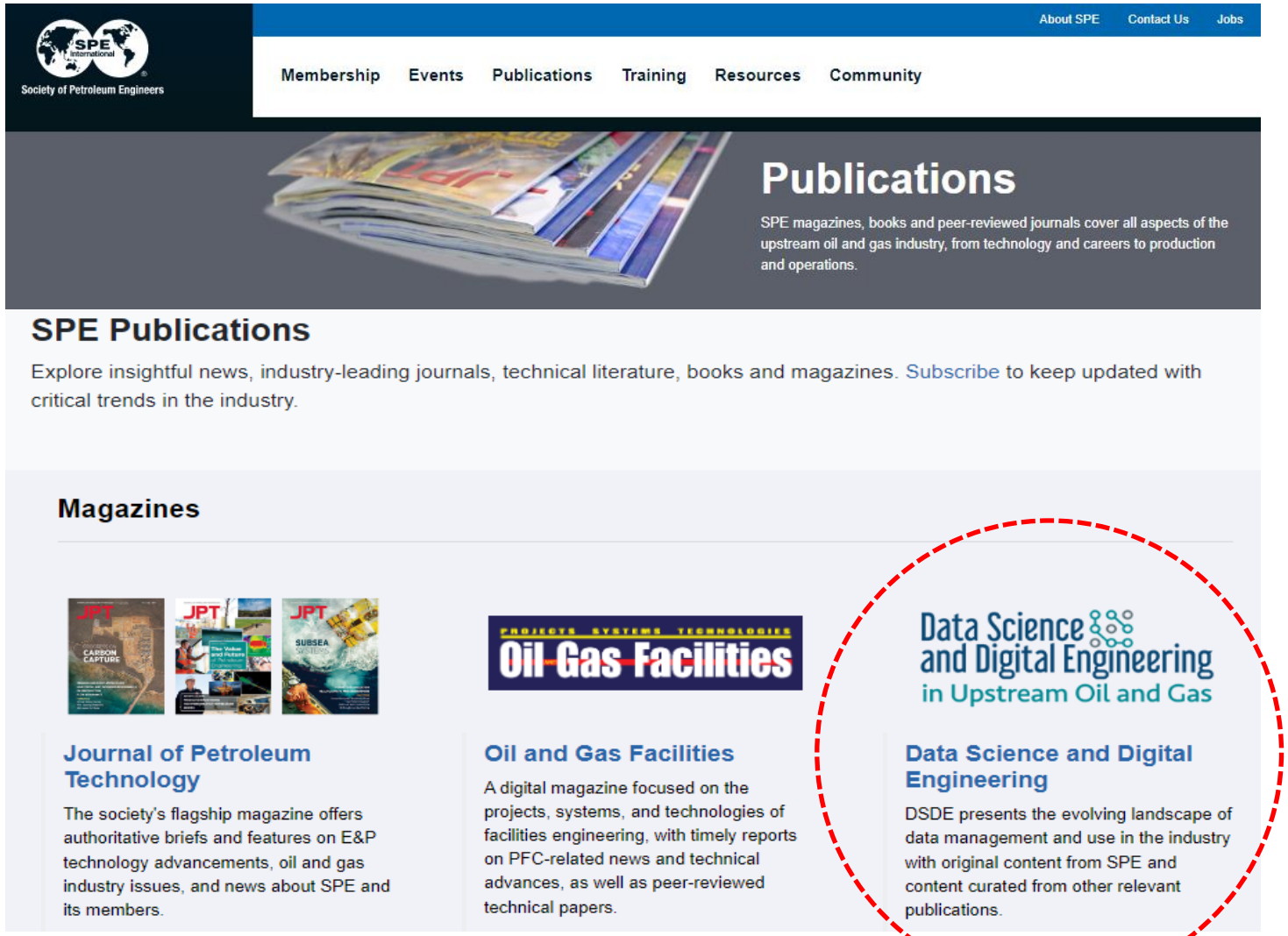
data
iku

DS365.ai

Why to Learn DS?



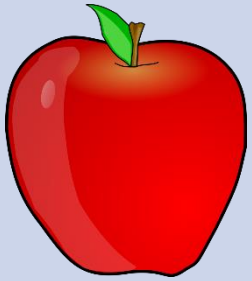
Why to Learn DS?



Why Machine Learning?

Machine Learning vs Explicit programming

Data Driven vs Physics Driven

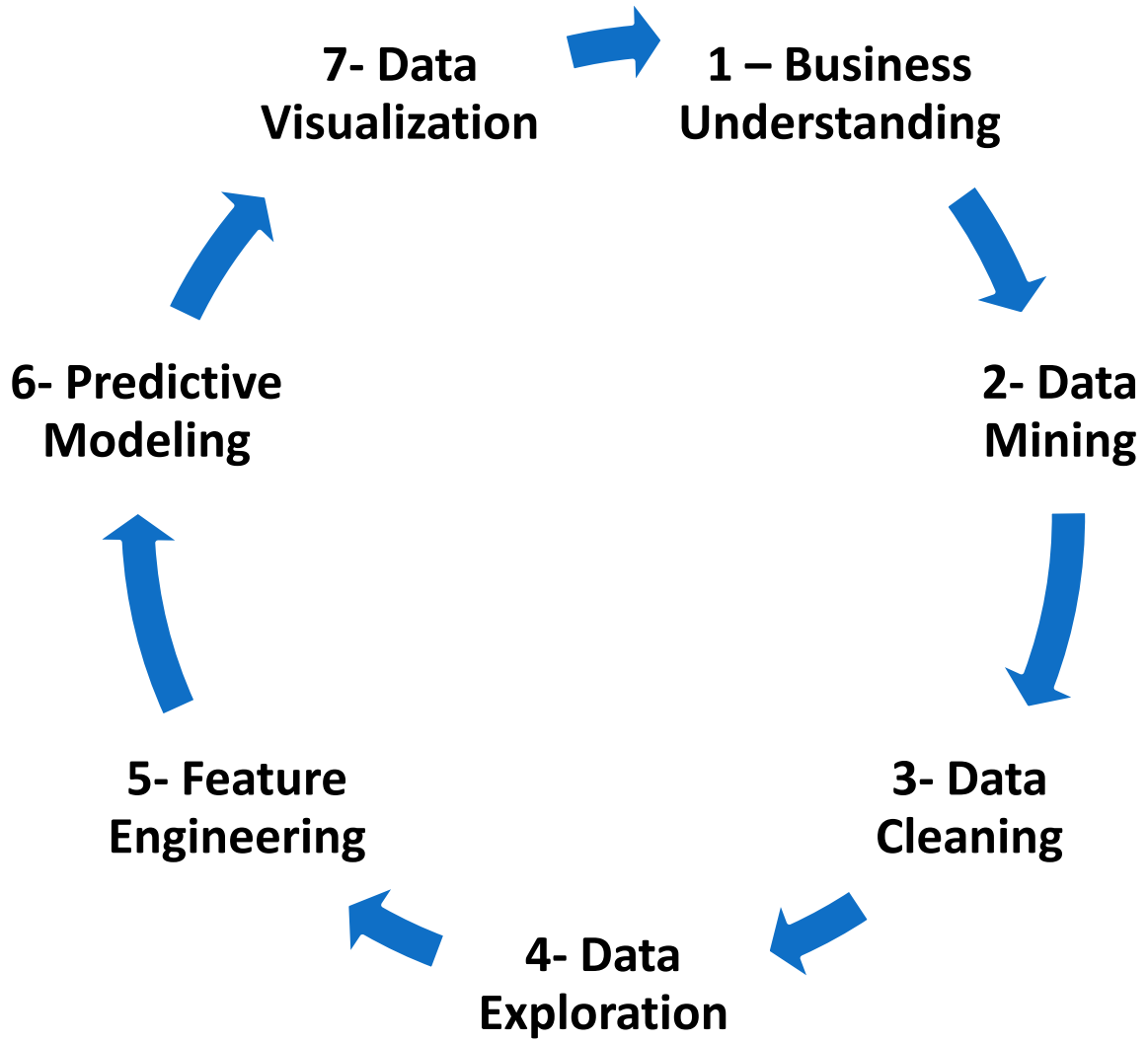
Machine Learning	Explicit Programing
• Identifying an apple? 	• Calculating a car speed given the distance and time?
• Calculating the bubble point pressure given the Rs and API?	

Why Machine Learning?

Golden Rule

Whenever it’s an art more than a science Machine Learning comes into play!

DS Lifecycle



1- Business Understanding

Data Analysis always starts with asking **Questions**

- What am I trying to find out?
- Is there a problem I am trying to solve?

Right question help you focus on relevant parts of your data and direct your analysis towards meaningful insights

2- Data Mining

Gather and scrape the data necessary for the project.

Tools:

- Database (MS SQL Server, Oracle,…) → SQL
- Web Data (Facebook, Twitter, Amazon, …) → web Crawling techniques (Python)
- Excel Sheets
- Unstructured Data (images, documents, …)

3- Data Cleaning

Fix the inconsistencies within the data and handle the missing values

Tools:

- Excel
- Python (NumPy, Pandas, …)

Data Wrangling

Another terminology for step 2 and 3

Means, making sure you have all the data you need in a great quality that you can work with.

Three Steps:

1. **Gather the data**, that help you answer your questions.
2. **Assess the data**, identify any problem in your data quality (missing and wrong values) or structure (data types).
3. **Clean the data**, modifying, replacing, or moving data. Ensure your dataset is in a high quality.

4- Data Exploration

EDA, Exploratory Data Analysis

Form hypotheses about your defined problem by visually analyzing the data.

Exploring involves:

- Finding Patterns
- Visualizing Relationships
- Building intuition about what you're working with

4- Data Exploration

After exploring you can do things like:

- Remove Outliers
- Feature Engineering

Do not be disappointed in the process!

As you get familiar with your dataset in EDA, you will often revisit previous steps

- New quality issues.
- New unexpected patterns and decide to refine your questions.
- Need more data.

Tools: Excel – Python (Matplotlib, Seaborn) – Tableau – MS Power BI

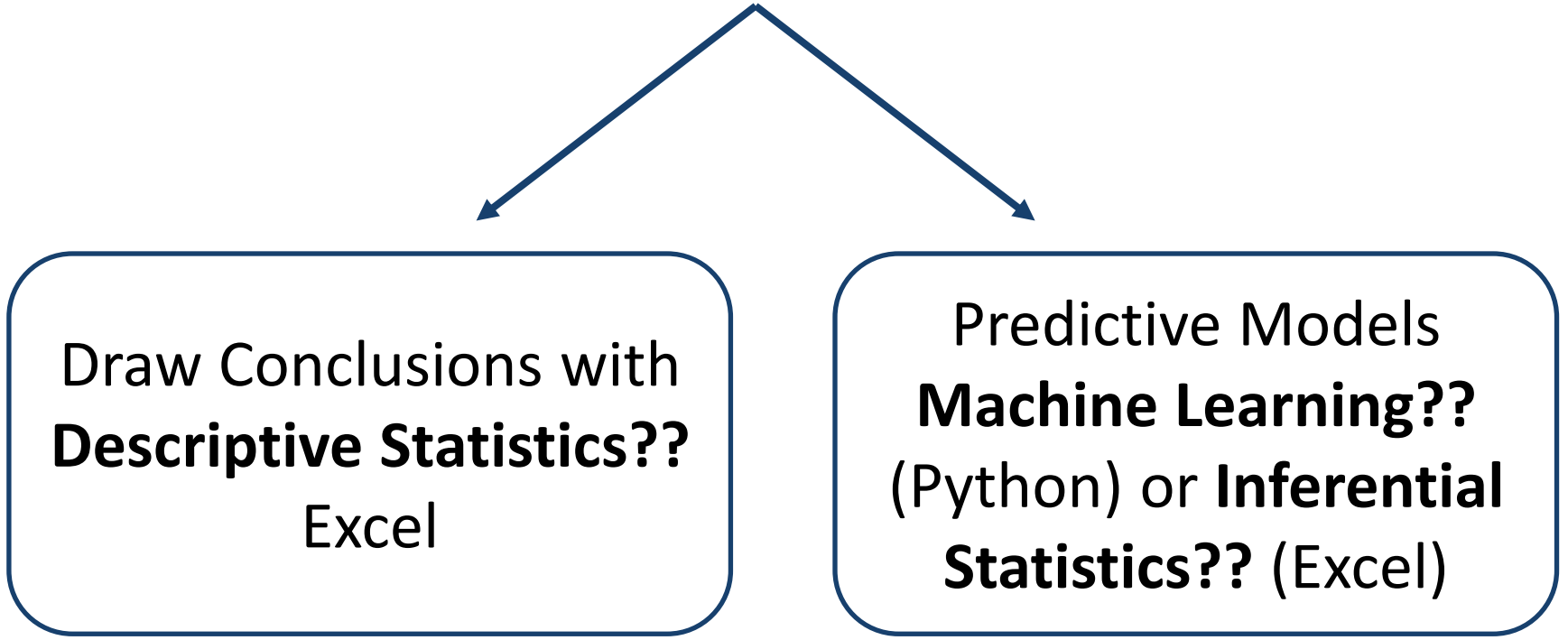
5- Feature Engineering

Select important features and construct new meaningful ones using the raw data that you have.

Algorithms:

- Principal Component Analysis (PCA)
- Correlation coefficients

6- Conclusions OR Predictive Modeling



7- Data Visualization

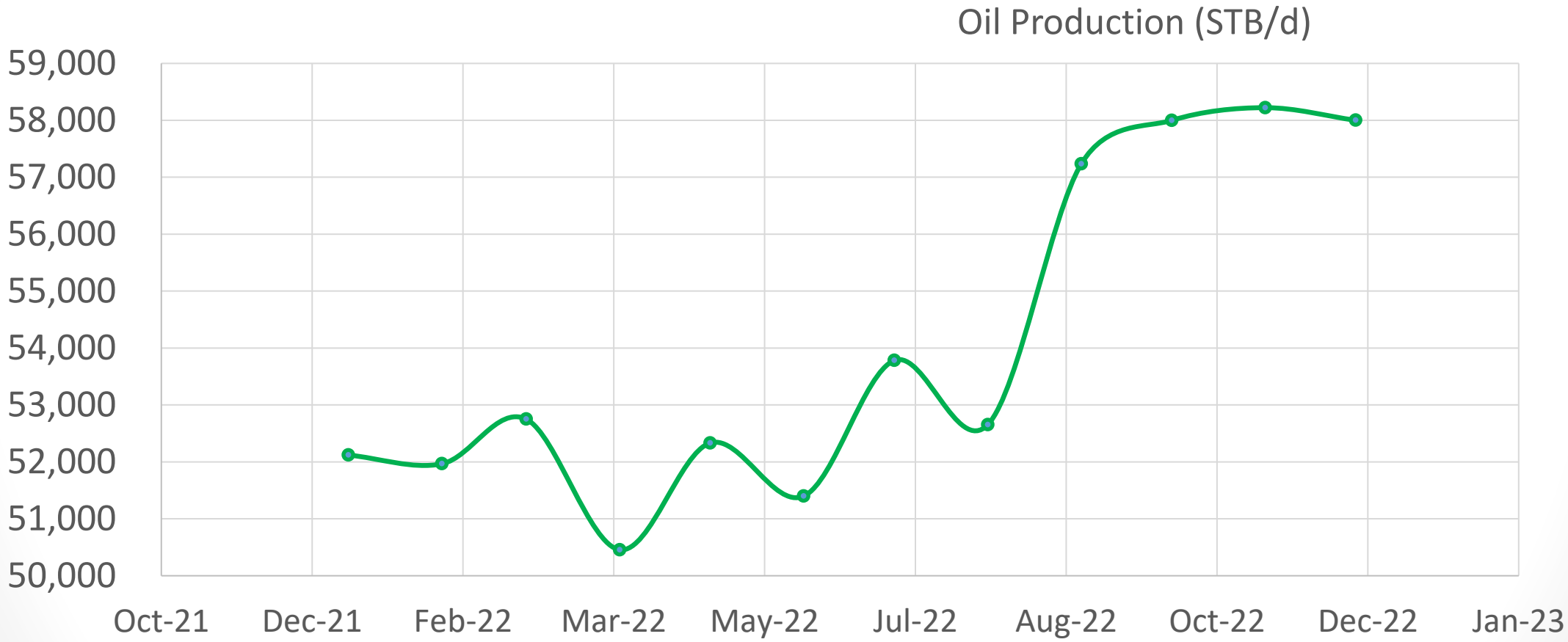
Communicating Results through:

- Charts
- Reports
- Dashboards and so on

Data Visualization is a **science**

7- Data Visualization

A Tremendous Increase in Oil Production in Our Company!



7- Data Visualization

$$\text{Lie Factor} = \frac{\text{Size of the effect shown in the graphic}}{\text{Size of the effect shown in the data}}$$

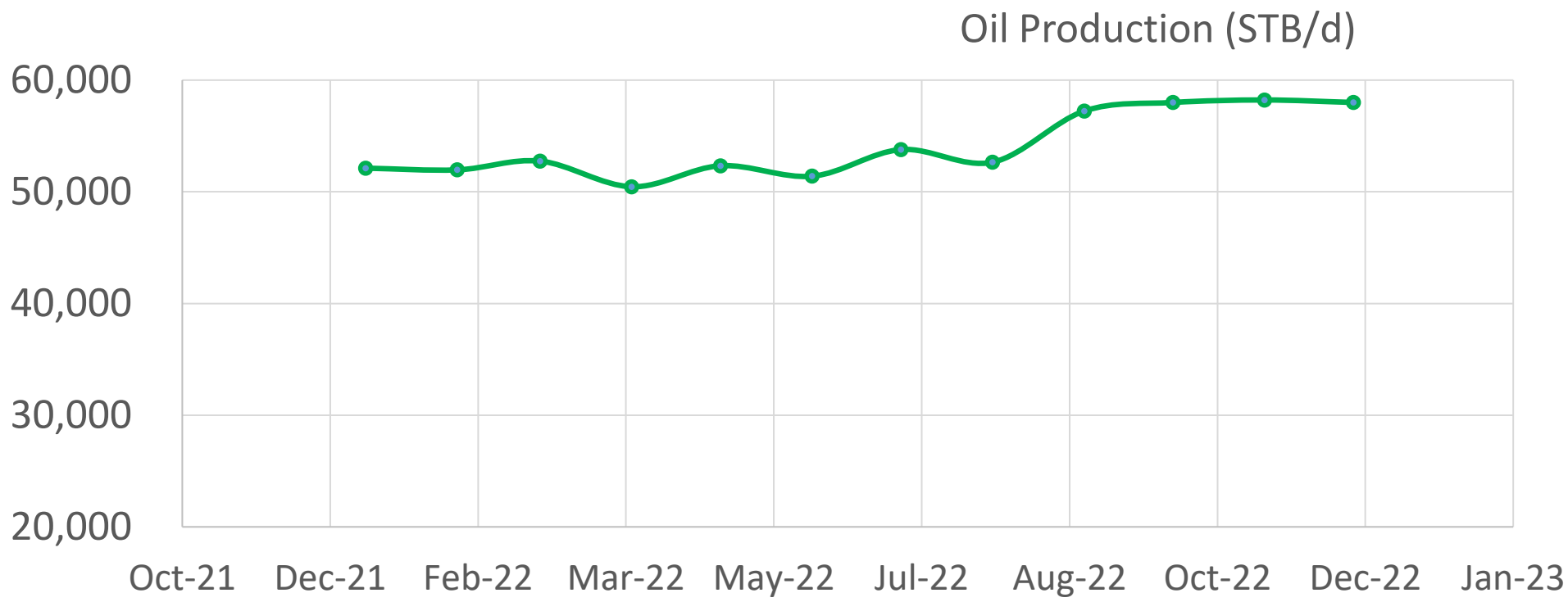
$$\text{Size of the Effect} = \frac{\text{Second value – first value}}{\text{First value}}$$

$$\text{Lie Factor} = \frac{(7\text{-}1.5)/1.5}{(58{,}000\text{-}52{,}000)/52{,}000} = \frac{367\%}{12\%}$$

$$= 31.8$$

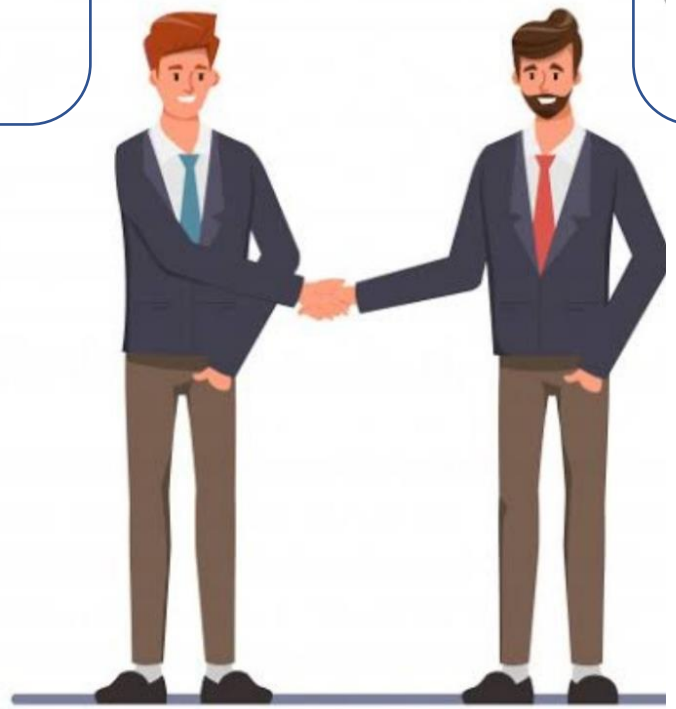
7- Data Visualization

Our Company Showed a Good Improvement in The Last Quarter!



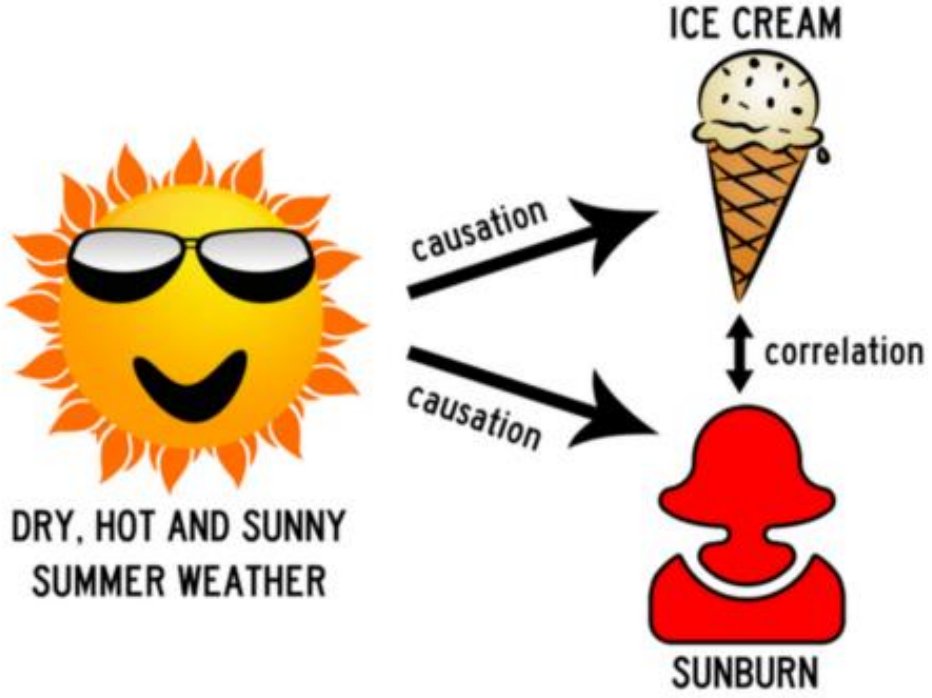
DS Team

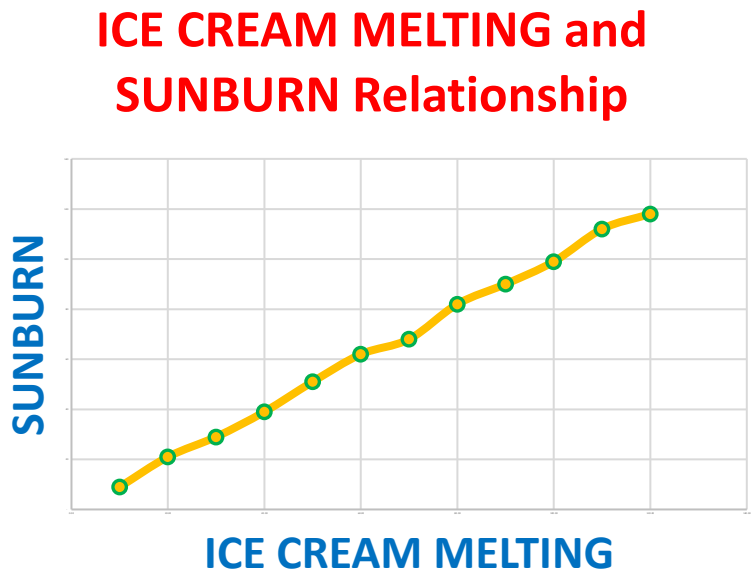
Data Scientist
Implement the algorithm



Business Expert
Have the technical knowledge put the algorithm

Correlation Doesn't Imply Causation





Platforms


coursera


egyptfwd
initiative


udemy


UDACITY


lynda.com


DataCamp

Thank you