

# Your Last Introduction to Data Science

By / Omar Abd El-Sameaa  
Reservoir Engineer at Bapetco



# Contents

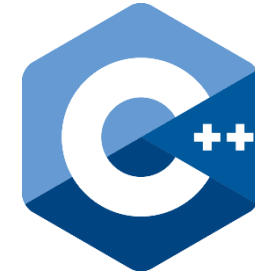
---

- 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.



**DS365.ai**

# Why to Learn DS?

The image displays two screenshots of the OnePetro search results page. The top screenshot shows search results for 'data science' with 63,690 papers. The bottom screenshot shows search results for 'machine language' with 2,965 papers. Both screenshots have a blue oval highlighting the total number of papers.

**Top Screenshot: data science**

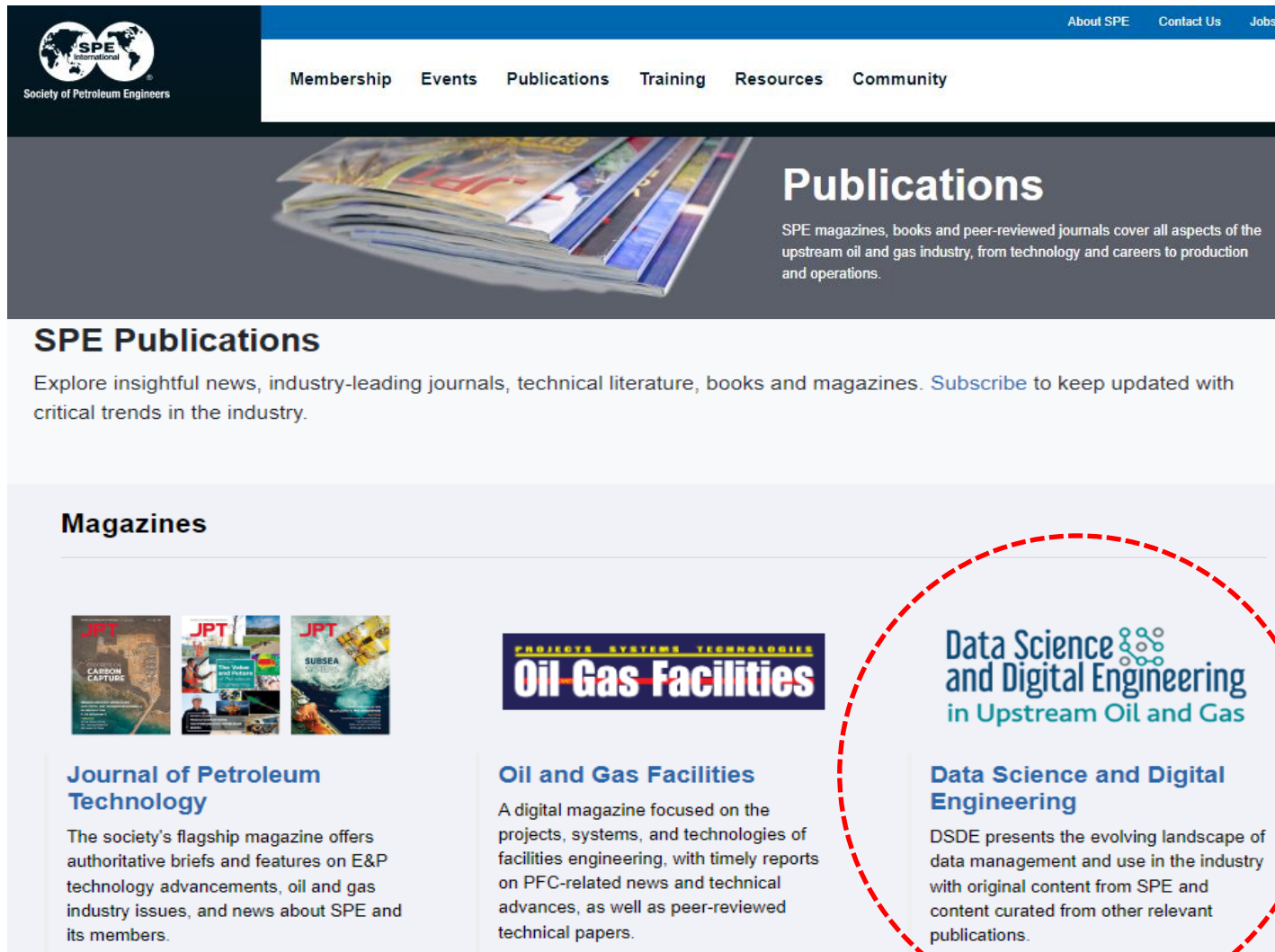
- Search term: data science
- Results: 1-20 of 63690 Search Results for data science
- Buttons: ADD TERM, UPDATE, Filter, Select All on Page, Add to Cart, Add to Citation Manager
- Peer Reviewed:  Peer Reviewed Only (16461)
- Sample result: Collation, Analysis of Oil and Gas Production Reports Using Excel, Python and R: A **Data Science** Approach in Handling Large Data

**Bottom Screenshot: machine language**

- Search term: machine language
- Results: 1-20 of 2965 Search Results for machine language
- Buttons: ADD TERM, UPDATE, Filter, Select All on Page, Add to Cart, Add to Citation Manager
- Peer Reviewed:  Peer Reviewed Only (601)
- Format:  Journal Articles (509)
- Sample result: Electrical Submersible Pump Prognostics and Health Monitoring Using **Machine** Learning and Natural **Language** Processing  
Amey Ambade, Saniya Karnik, Praprut Songchitruksa, Rajeev Ranjan Sinha, Supriya Gupta  
Publisher: Society of Petroleum Engineers (SPE)  
Paper presented at the SPE Symposium: Artificial Intelligence - Towards a Resilient and Efficient Energy Industry, October 18-19



# Why to Learn DS?



The screenshot shows the SPE Publications website. At the top left is the SPE International logo with the text 'Society of Petroleum Engineers'. To the right are navigation links: 'About SPE', 'Contact Us', and 'Jobs'. Below this is a main menu with 'Membership', 'Events', 'Publications', 'Training', 'Resources', and 'Community'. The main content area features a stack of magazines and a 'Publications' section with a description: 'SPE magazines, books and peer-reviewed journals cover all aspects of the upstream oil and gas industry, from technology and careers to production and operations.' Below this is a 'SPE Publications' section with a sub-header and a paragraph: 'Explore insightful news, industry-leading journals, technical literature, books and magazines. [Subscribe](#) to keep updated with critical trends in the industry.' The 'Magazines' section lists three publications: 'Journal of Petroleum Technology' (with three cover images), 'Oil and Gas Facilities' (with a logo), and 'Data Science and Digital Engineering in Upstream Oil and Gas' (with a logo and a red dashed circle around it). Each publication has a brief description.

**SPE Publications**

Explore insightful news, industry-leading journals, technical literature, books and magazines. [Subscribe](#) to keep updated with critical trends in the industry.

**Magazines**

**Journal of Petroleum Technology**

The society's flagship magazine offers authoritative briefs and features on E&P technology advancements, oil and gas industry issues, and news about SPE and its members.

**Oil and Gas Facilities**

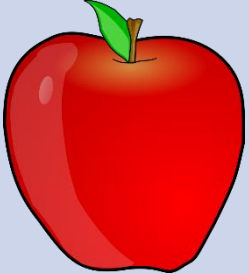
A digital magazine focused on the projects, systems, and technologies of facilities engineering, with timely reports on PFC-related news and technical advances, as well as peer-reviewed technical papers.

**Data Science and Digital Engineering in Upstream Oil and Gas**

DSDE presents the evolving landscape of data management and use in the industry with original content from SPE and content curated from other relevant publications.

# Why Machine Learning?

## Machine Learning vs Explicit programming Data Driven vs Physics Driven

Machine Learning	Explicit Programming
<ul style="list-style-type: none"><li>Identifying an apple?</li></ul> 	<ul style="list-style-type: none"><li>Calculating a car speed given the distance and time?</li></ul>
<ul style="list-style-type: none"><li>Calculating the bubble point pressure given the Rs and API?</li></ul>	

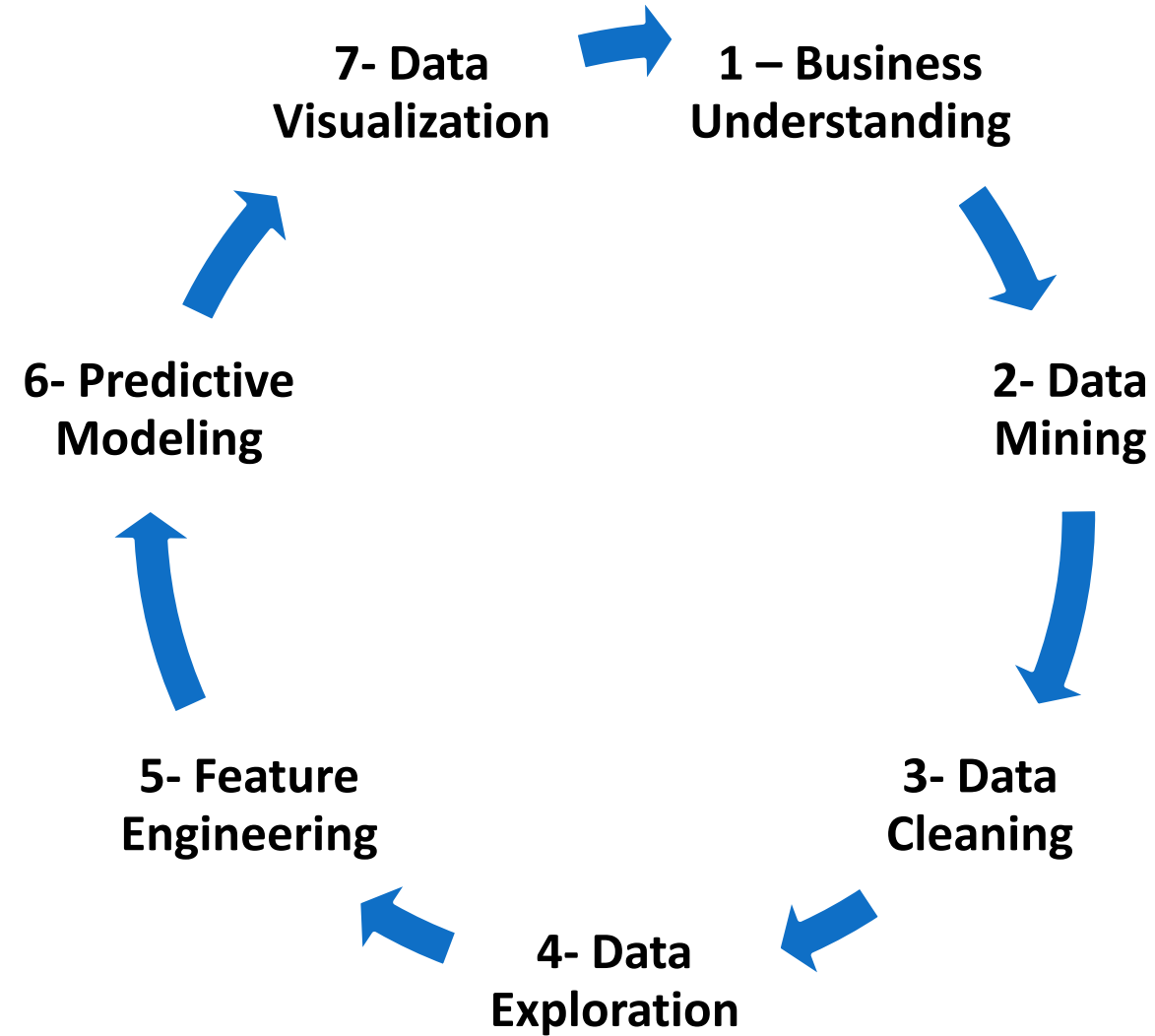
# Why Machine Learning?

---

## 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

---

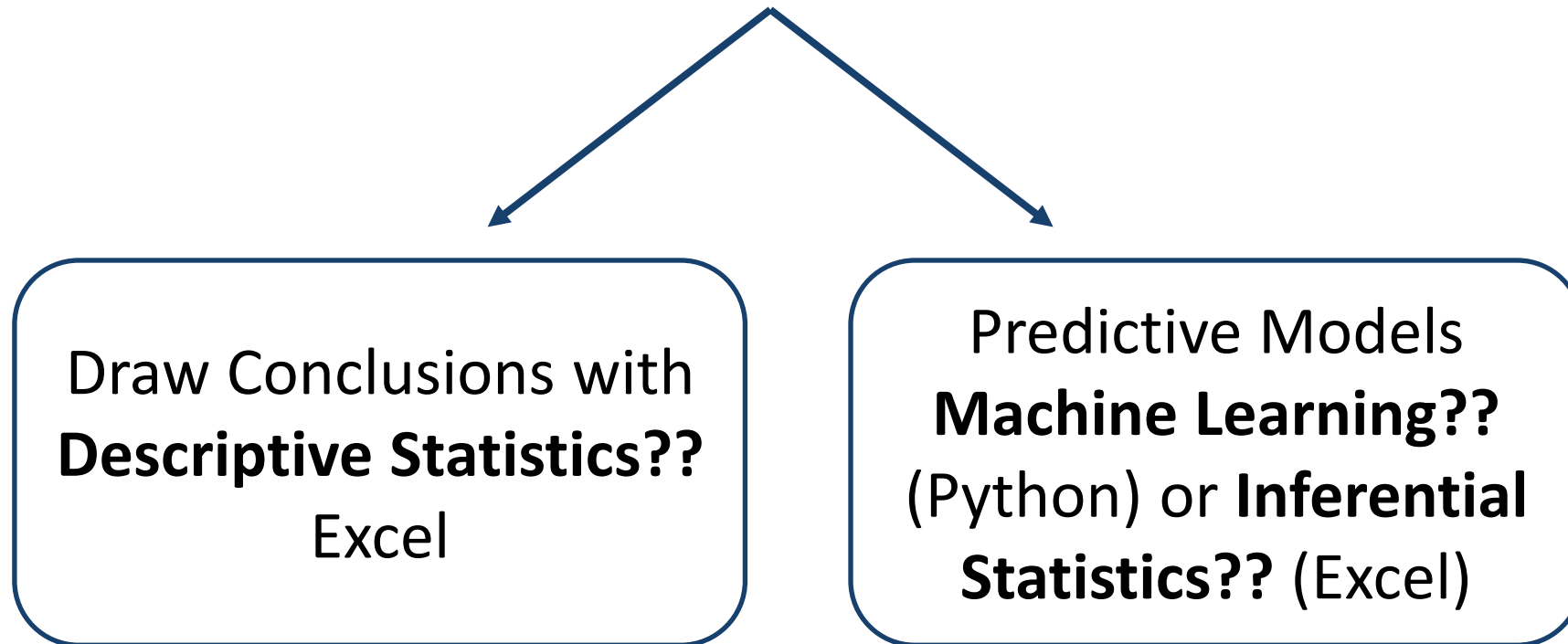
Select important features and construct new meaningful ones using the raw data that you have.

## Algorithms:

- Principal Component Analysis (PCA)
- Correlation coefficients

# 6- Conclusions OR Predictive Modeling

---



# 7- Data Visualization

---

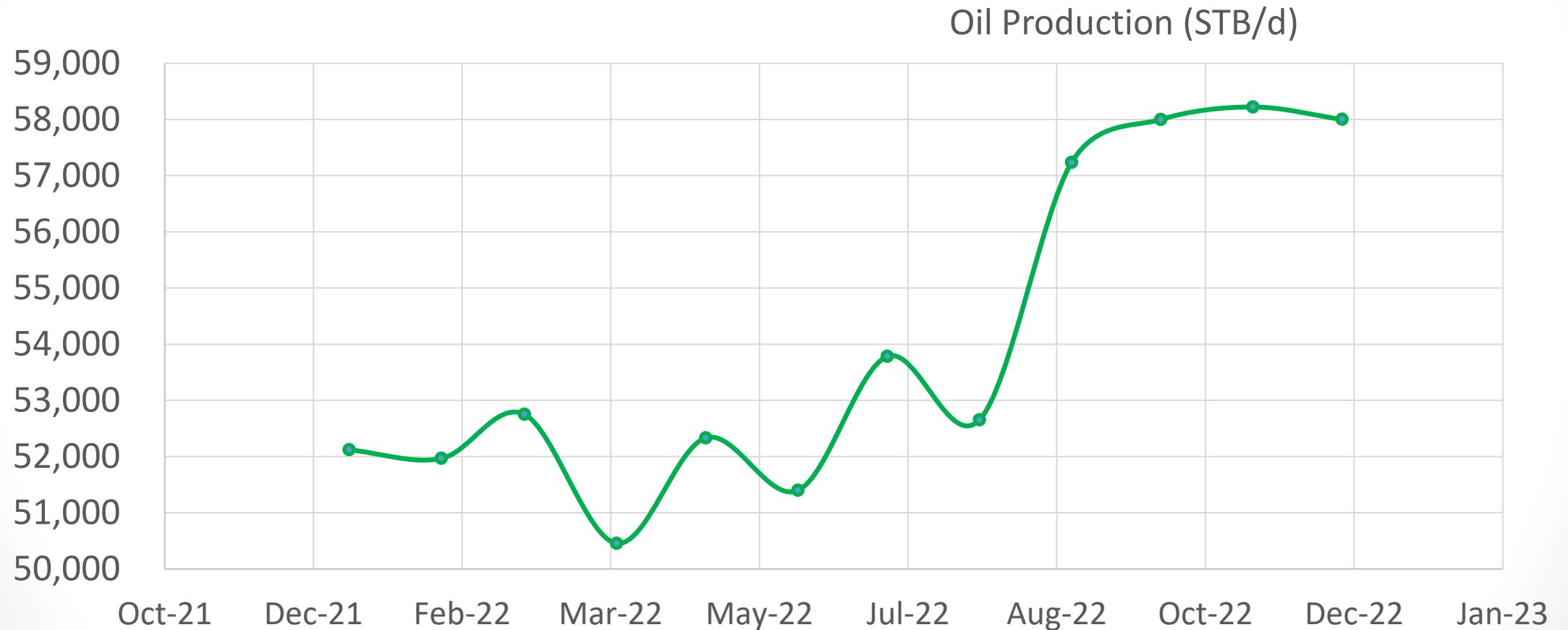
## Communicating Results through:

- 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

---

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

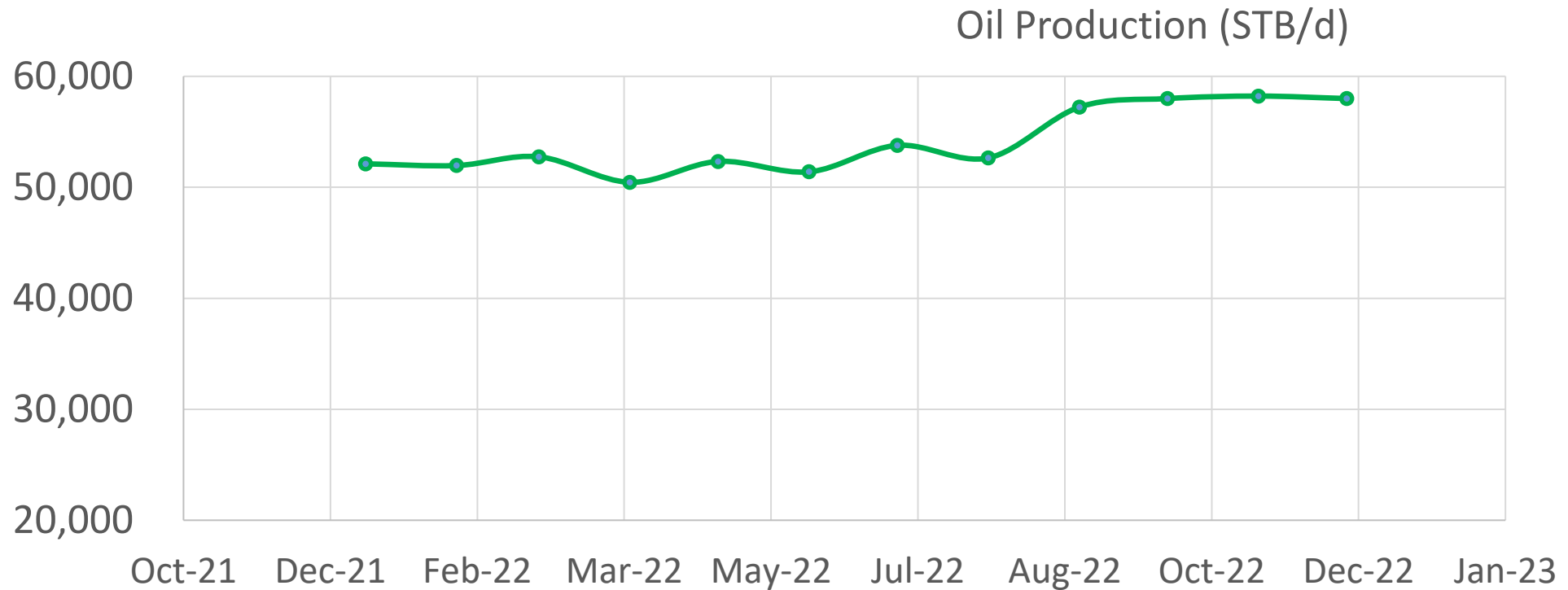
$$\text{Size of the Effect} = \frac{\text{Second value} - \text{first value}}{\text{First value}}$$

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

$$= 31.8$$

# 7- Data Visualization

**Our Company Showed a Good Improvement in The Last Quarter!**

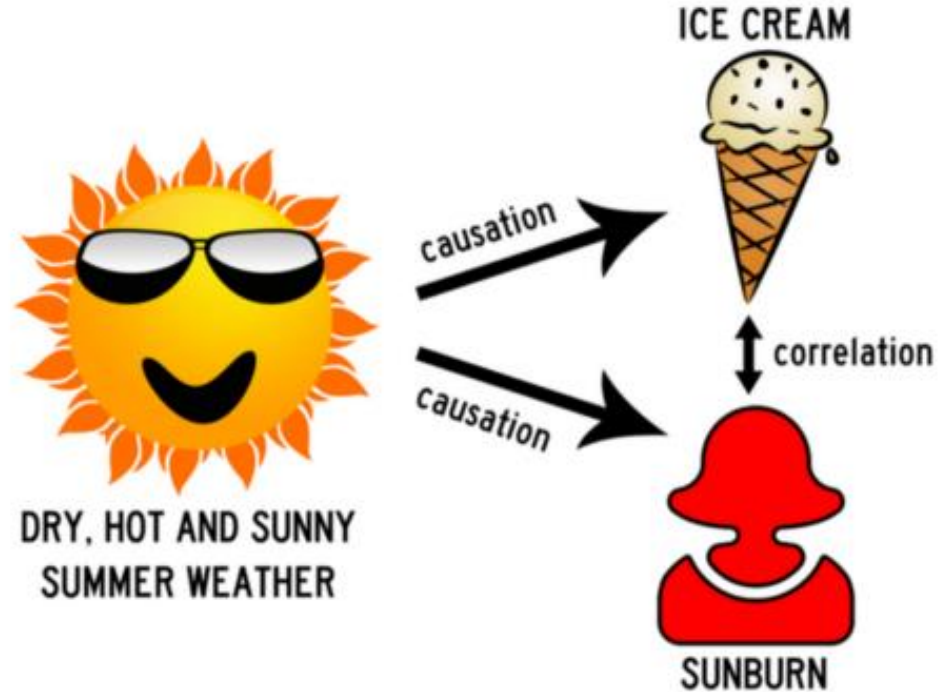


**Data Scientist**  
Implement the  
algorithm

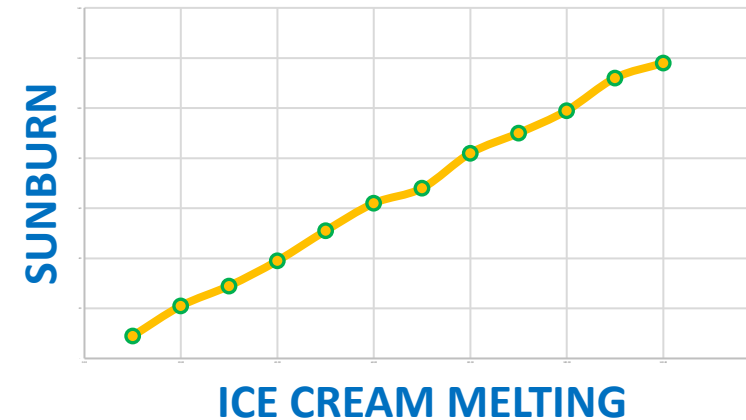
**Business Expert**  
Have the technical  
knowledge  
put the algorithm



## Correlation Doesn't Imply Causation



**ICE CREAM MELTING and SUNBURN Relationship**



# Platforms

---

**coursera**

egypt**fwd**  
initiative

**udemy**

  
UDACITY



  
**DataCamp**

# Thank you

