Data Science in Practice

Ikenna Ivenso

30 May 2020

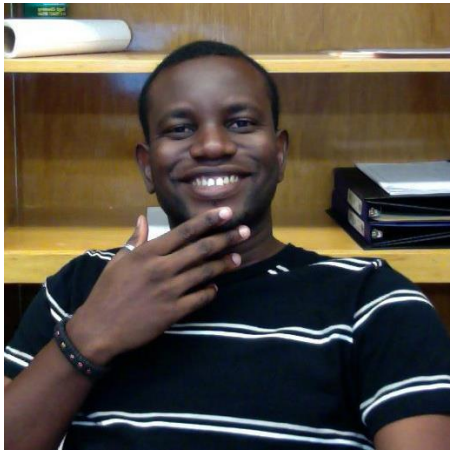
www.tqstem.org

Objectives

1. Introductions
2. What Is Data Science ?
3. What is Data ?
4. Example: Lifecycle of a Simple DS Project
5. What is Machine Learning ?
6. Two Frequently Asked Questions about a DS Career

Introductions

Ikenna Ivenso



EDUCATION:

- *Ph.D. in Mechanical Engineering (Computational Biophysics)*
- *MSc in Mechanical Engineering (Robotics and Automation)*
- *BSc in Mechanical Engineering*

PROFESSIONAL EXPERIENCE:

- *Senior Principal Data Scientist at Dell-EMC*
- *Senior Software Engineer at Intel Corporation*
- *Postdoctoral Fellow in Data Science at Insight Data Science*
- *Instructor in Computational Mechanics at Texas Tech University*

Data Science

What is it ?

Data Science uses various **tools** to extract **insights** from **data**

- Tools: math and stats, algorithms, experimentation, visualization, etc.

Some use cases

- Airlines save hundreds of millions by using DS to reduce aircraft downtime
- Amazon generates up to 35% of its revenue from its recommendation engine
- Banks have saved billions by using DS to detect and prevent fraud and cyber-crime
- DS is used in healthcare to improve patient management and medical image analysis
- Advertising firms use DS for more effective targeting of their campaigns

Data

What is it ?

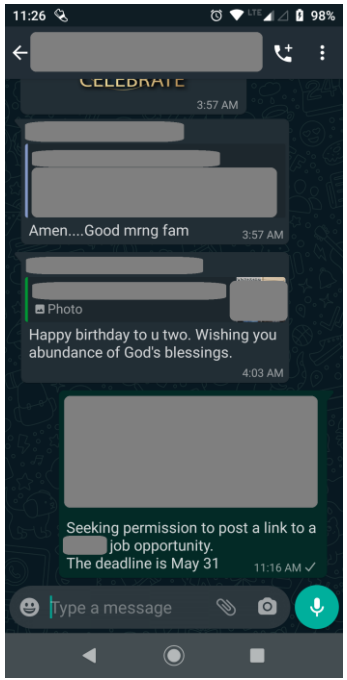
Data is anything we can observe and/or measure

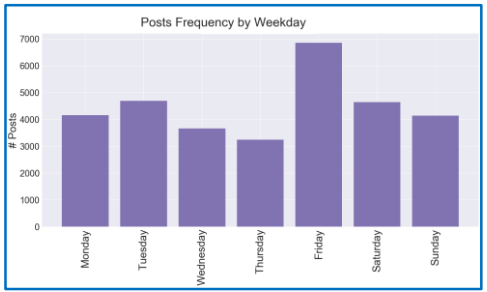
- Data empowers us to make better decisions
- Decision making improves if we can store, retrieve and analyze data

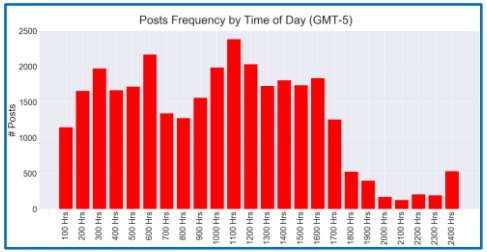
Some common data sources

- Social media: Facebook, Twitter, Instagram, LinkedIn, etc.
- Machine generated: operating conditions, sensors, computer networks, etc.
- Transactions: online payments, receipts, inventory, etc.
- Government data: census, elections, GDP, etc.
- ***Non-traditional sources: web scraping, news, books, call and chat history, etc.

Example: Non-Traditional Data Source

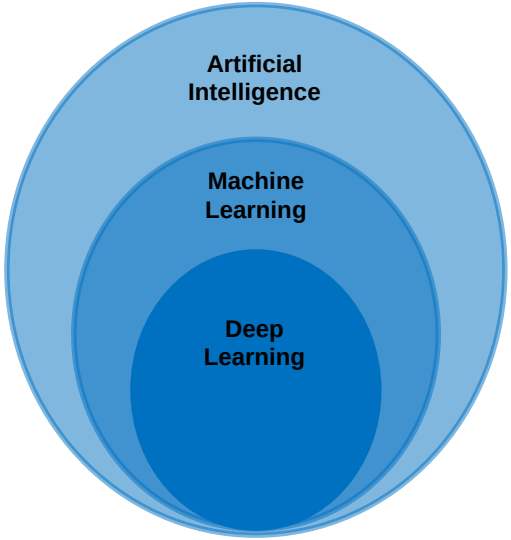




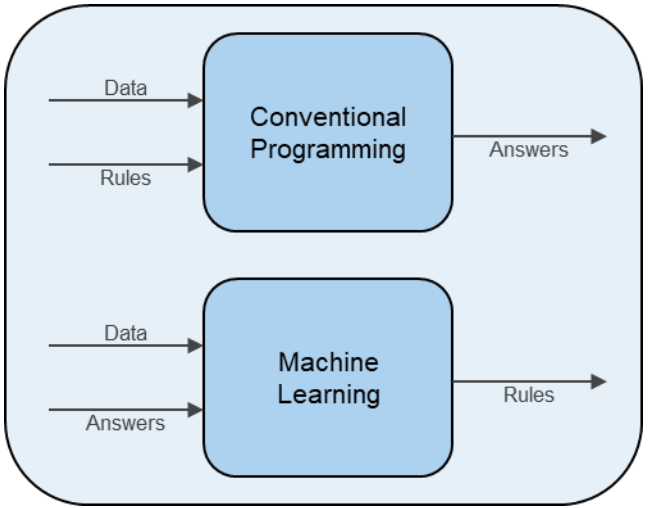


Machine Learning

- **Artificial Intelligence** (AI) is the ability of systems to *receive* information, *analyze* it and *carry out actions* in response (e.g. car collision avoidance systems)
- **Machine Learning** is the use of Artificial Intelligence to enable systems *learn from experience* without being *explicitly programmed* (e.g. spam detectors and recommendation engines)
- **Deep Learning** is a Machine Learning technique that mimics the brain's ability to *represent information in different abstract levels* (e.g. image classifiers and voice recognition)



Machine Learning



Input 1	Input 2	Output
0	0	0
0	1	1
1	0	1
1	1	1

Quant.	Verbal	Writing	Admitted
750	570	3.0	No
780	600	5.0	Yes
800	590	3.5	No
720	630	4.5	No
780	620	5.0	Yes
780	580	6.5	No

Machine Learning is useful when

- "Rules" are not obvious OR too difficult to code

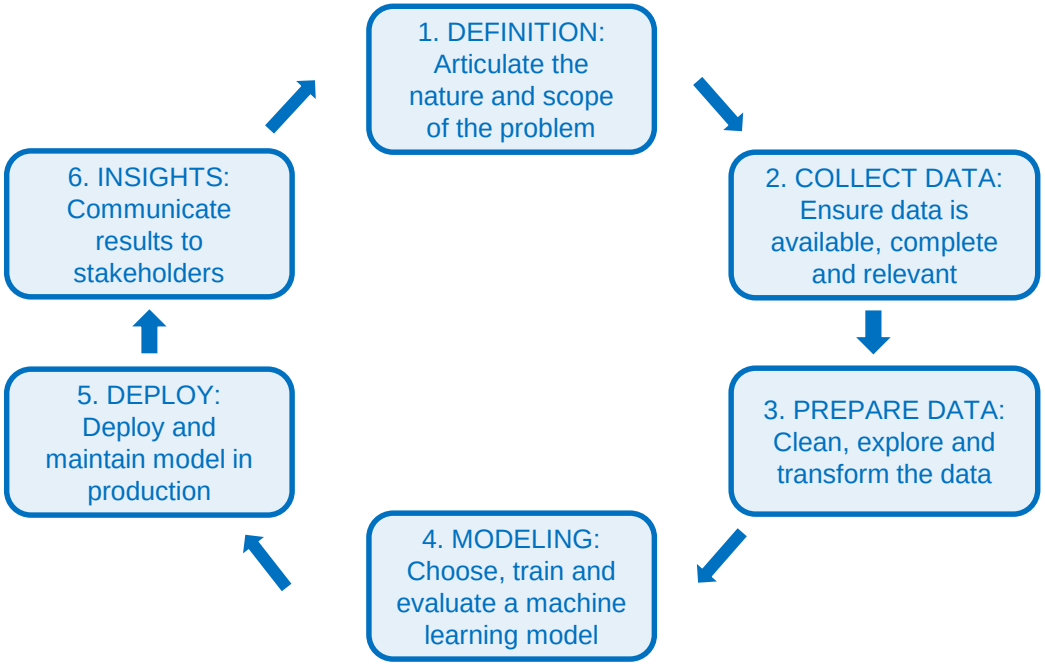
A problem is a good candidate for Machine Learning if

- A reasonable amount of data exists
- A pattern exists in the data
- The pattern is difficult to figure out

www.tqstem.org

Pause for Questions

Data Science Project Lifecycle



Data Science Project Lifecycle

1. DEFINITION: Articulate the nature and scope of the problem

Task 1:
Formulate and articulate the problem(s) to be solved

Skills:
Domain knowledge and/or access to stakeholder's business subject matter experts

To Do:

- What question/problem does the stakeholder need you to answer?
- Formulate the question so that it can be answered with data.
- What outcome(s) does the stakeholder hope for?
- What are the stakeholer's business processes/operations?
- How do they measure the success?
- etc.

¶

Business Case Overview:

- The business spends a lot of money on targeted advertising, promotions and other activities aimed at customer retention.
- They would like to know which customers are likely to cancel their subscriptions
- They would also like to know what factors drive customer churn
- This will enable them to save money on customer retention activities targeting these specific customers
- They could also, possibly, be able to reach at-risk customers before they churn
- The business is particularly concerned about false negatives (i.e. classifying customers as NOT churning who will churn)
- The business is not as concerned about false positives (i.e. classifying customers as churning who will NOT churn)

Data Science Project Lifecycle

Task 2:
Identify the data needed to answer these questions

Skills:
Access to the data and the ability to retrieve it (e.g. SQL, pyodbc, Excel, flat file manipulation, etc.)

To Do:

- Is there data and enough of it?
- Is the data accessible?
- Is the data relevant?
- Is the data complete?
- Is the data representative of the problem or is it imbalanced?
- etc.

2. COLLECT DATA:
Ensure data is available, complete and relevant

Data Science Project Lifecycle

Task 3:
Explore the data

Skills:
Data manipulations skills

To Do:

- Is the data structured or unstructured?
- What fields exist in the (structured) data?
- Are there outliers? (e.g. patient_age=250)
- Are there strong correlations between any two fields? (e.g. call_duration and call_charge)
- What data types exist in each field?
- Can new features/fields be engineered? (e.g. evaluate age given birth_date)
- etc.

3. PREPARE DATA:
Clean, explore and transform the data

Data Science Project Lifecycle

Task 4:
Select and build model

Skills:
Machine learning knowledge and skills

To Do:

- What kind of technique best suits the problem?
- Does the data need to be further transformated? (e.g. encoding)
- Can the results of the model be easily explained if necessary?
- Choose the most suitable technique...not just the coolest or newest one
- ALWAYS test and evaluate the performance of your model
- Does the model generalize well to new/unseen data? ***
- etc.

4. MODELING:
Choose, train and evaluate a machine learning model

Data Science Project Lifecycle

6. INSIGHTS: Communicate results to stakeholders

5. DEPLOY: Deploy and maintain model in production

Task 5:
Communicate the results (or deploy the model into production) ¶

Skills:
Communication/Storytelling skills

To Do:

- Target audience is most likely the executives?
- Avoid temptation to use DS jargon; communicate to them in their language
- Be sure to address the problem you set out to solve
- etc.

Pause for Questions

Frequently Asked Questions

How do I become a Data Scientist ?

- Develop deep curiosity and strong passion for DS
- Take advantage of available (free?) resources
- Start small, practice consistently
- Seek out new challenges (best way to improve)
- Keep up with the DS community
- Keep learning
- Internalize your learning by teaching others

Frequently Asked Questions

How do I get a job in Data Science ?

- Build your (online?) portfolio
- Work on interesting /challenging problems
- Build up your DS toolbox
- Master effective communication and storytelling

THANK YOU!
Any Questions

www.tqstem.org