Serving Data Science to a Rookie

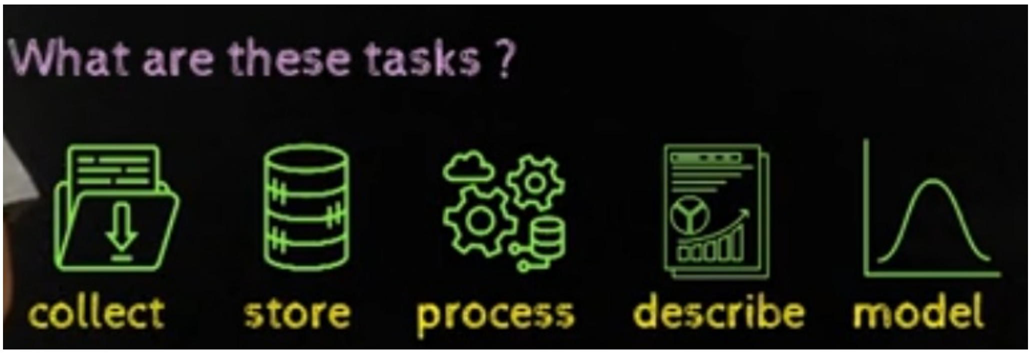
So, last week my team head asked me to interview some of the possible interns for the team for the role of data science and machine learning, and he forwarded me their resumes. He asked me to select at most 2 candidates from 8 eligible candidates. That was pretty usual, right?

Now here comes the twist, I called everyone and asked just one fundamental question "**What is Data Science?**", one replied it is the science of extracting data and then modeling it, another one responded it is equivalent to machine learning, next one told me it is a part of AI to predict/classify.

These definitions might be correct collectively, but what exactly is it? No one was able to provide a decent pretext and its legit use. I was shocked as their knowledge was utterly contrary to their fancy resumes.

Now coming to the point, What is Data Science?

Data Science is the science of *collecting, storing, processing, describing, and modeling data*.



5 ingredients help to make a delicious Data Science dish:

1. Collecting
2. Storing
3. Processing
4. Describing

5. Modeling

The critical thing to note here is Machine Learning, and Data Science is not the same—you can learn the difference between them from my previous article i.e., What exactly Machine Learning is?

Collecting: This is the first step of Data Science, and it is a collection of relevant or irrelevant data from different sources. The data can be structured (that would require the use of SQL) or unstructured (that would need skills of crawling/scraping).

Skills required:

1. Programming knowledge
2. Database knowledge
3. Statistical knowledge

Storing: This includes storing the collected data so that it is readily available for further computation and predictions. The data can be stored in data lakes or data warehouses (e.g., Hadoop).

There are 3 characteristics to big data: High Volume, High variety, High velocity (the 3V's).

Skills required:

1. Programming knowledge
2. Database knowledge (SQL, NoSQL)
3. Data warehouse/ Data lake knowledge

Processing: This is when we start preparing the data for the leading cause i.e., prediction/classification. It includes data wrangling, filling missing data, and data normalization. In novice style, it is removing/replacing every unnecessary/NaN value and only storing the pertinent data.

Skills required:

1. Programming knowledge
2. Map Reduce (Hadoop)
3. Database knowledge (SQL, NoSQL)
4. Basic Statistical knowledge

Describing: This consists of visualizing the processed data for better understanding and summarization. This stage is decisive as it helps you model the algorithms accordingly.

Skills required:

1. Statistical knowledge
2. Spreadsheet knowledge (MS Excel)
3. Visualization tools (Python, Tableau, Power-BI)

Modeling: This is drawing inferences from the processed data. It includes identifying the relation between data, testing hypotheses, and providing the statistical guarantee.

There are further 2 types of modeling;

1. Statistical modeling—Includes a simple, intuitive model and is suited for low dimensional data.
2. Algorithmic modeling—This is also called machine learning, which contains a sophisticated, flexible model and can work with high dimensional data.

Skills required:

1. Programming knowledge
2. Statistical knowledge
3. Domain knowledge

But, there is a limitation to machine learning: you can't solve every problem using it. Suppose you have a large amount of high dimensional data, and you want to learn some complex relationships between value and labels. There, we will use Deep Learning.

Skills required for Deep Learning:

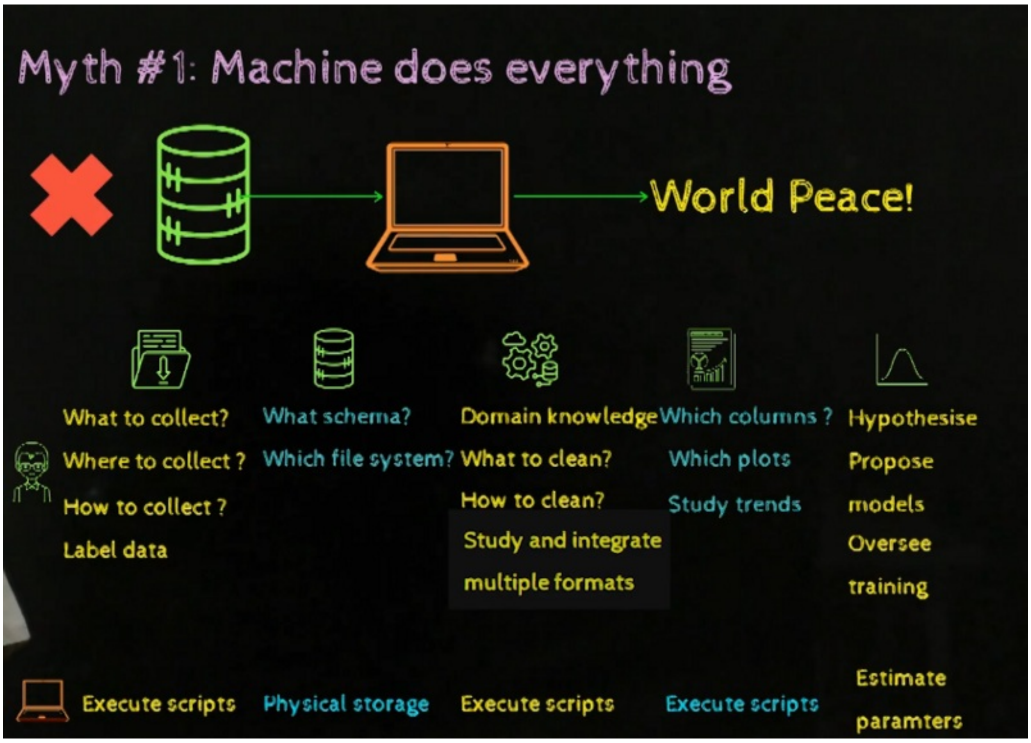
1. Inferential statistics
2. Probability theory
3. Calculus
4. Optimizing algorithms
5. Machine Learning
6. Programming skills

So what are some collective skills that Data Science demands?

1. Domain knowledge (Intermediate to Expert)
2. Programming skills (Intermediate to Expert)
3. Mathematical/Statistical knowledge (Intermediate to Expert)
4. Hacking skills (Novice to Intermediate)

Myths About Data Science

1. A machine can do everything.

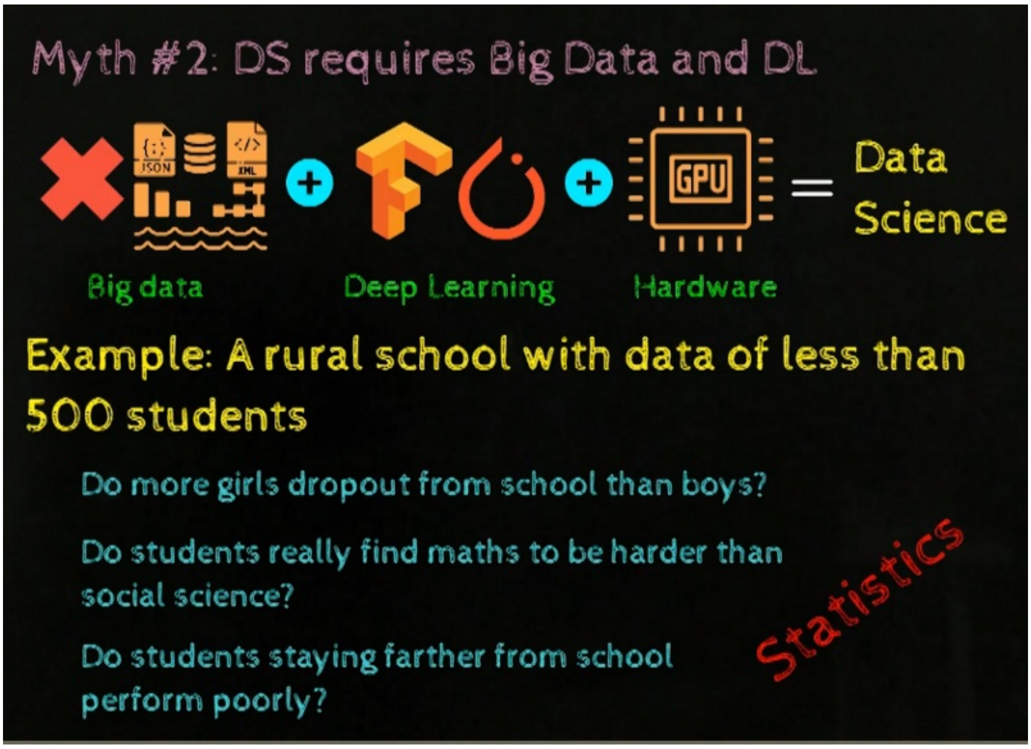


Myth #1: Machine does everything

World Peace!

What to collect? Where to collect ? How to collect ? Label data	What schema? Which file system?	Domain knowledge What to clean? How to clean? Study and integrate multiple formats	Which columns ? Which plots Study trends	Hypothesise Propose models Oversee training
Execute scripts	Physical storage	Execute scripts	Execute scripts	Estimate paramters

2. Data Science requires Deep Learning and Big Data.



Myth #2: DS requires Big Data and DL

Big data + Deep Learning + Hardware = Data Science

Example: A rural school with data of less than 500 students

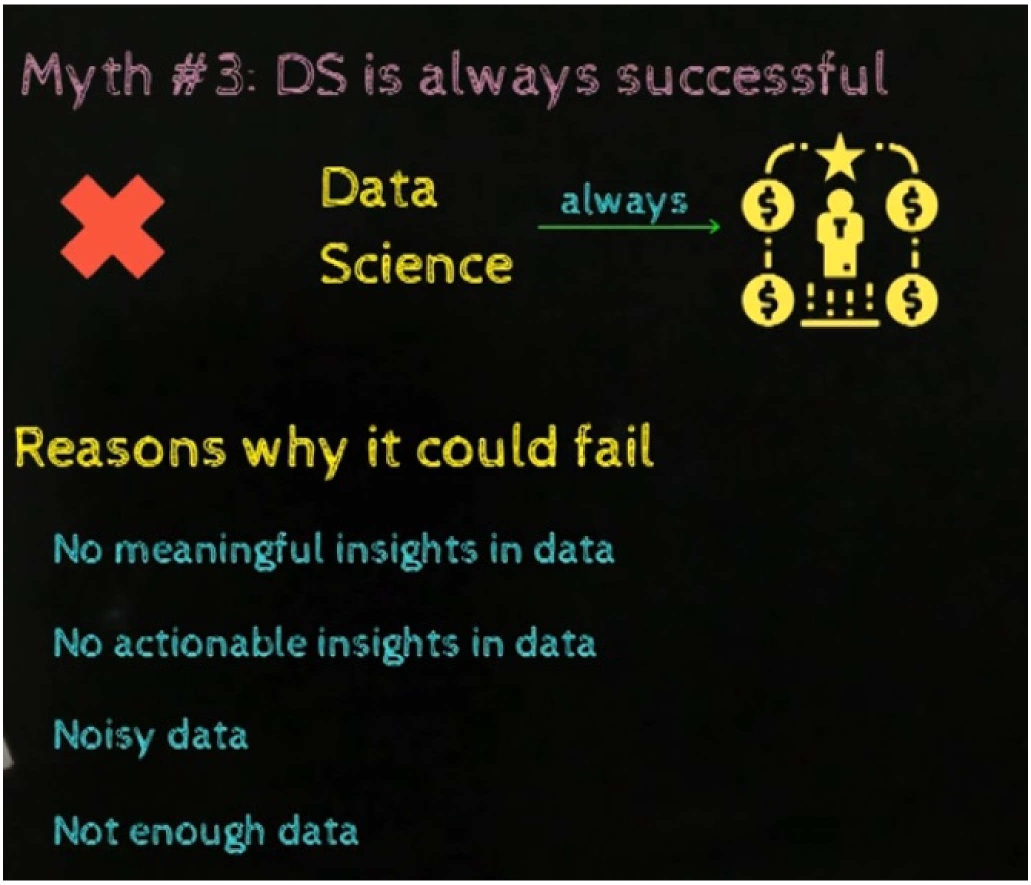
Do more girls dropout from school than boys?

Do students really find maths to be harder than social science?

Do students staying farther from school perform poorly?

Statistics

3. Data Science is always successful.



Myth #3: DS is always successful

Data Science $\xrightarrow{\text{always}}$

Reasons why it could fail

No meaningful insights in data

No actionable insights in data

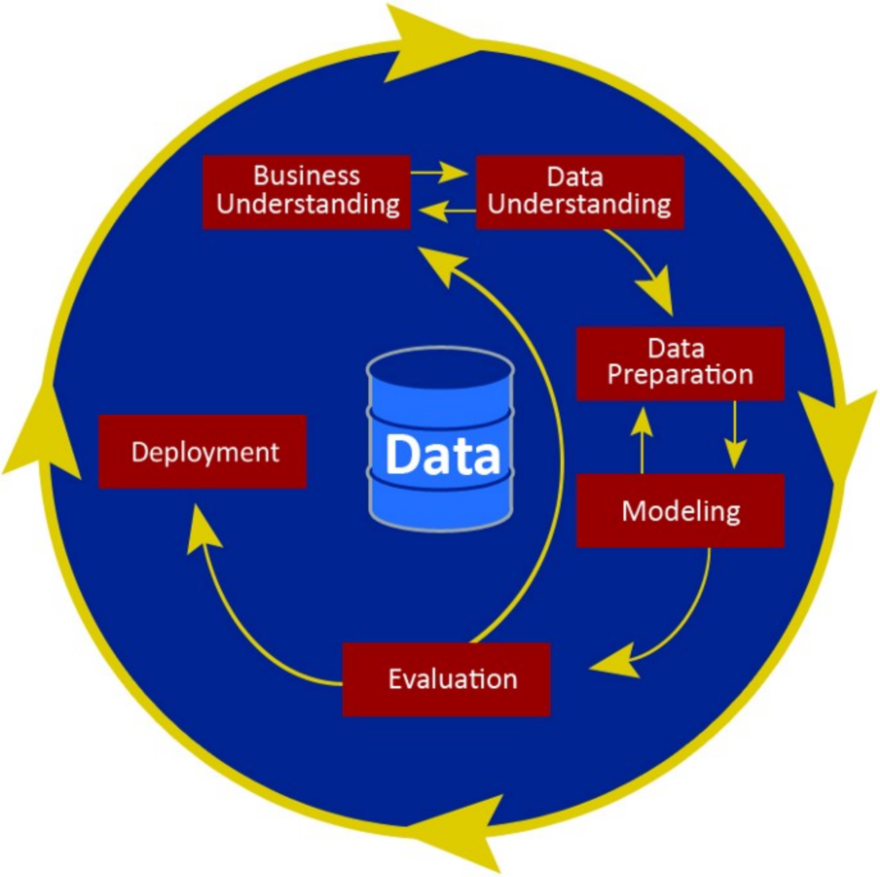
Noisy data

Not enough data

Now we know a lot about Data Science, but what about its implementation and tools it requires?

Crisp DM (Data Science pipeline)

It is an open standard process model that describes conventional approaches used to solve business analytical problems.



1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Tools for Data Scientist

1. No code environment (Novice users) — H2o.ai, Amazon Lex
2. Spreadsheets/BI tools (Novice users) — MS Excel, Tableau, Power BI
3. Programming language (Intermediate users) — Python, R, MATLAB
4. High-Performance Stack (Highly skilled users) — Hadoop, Apache Spark

Conclusion

Hopefully, this article helped you understand *What Data Science is?*, and what all skills you might require to become superb Data Scientist. And it's pretty unusual to find that just with few clicks and a little statistical and programming knowledge, we can manage many data. But again, all we care about is a model good at predicting/classifying :)

A unique token of appreciation to:

1. Dr. Mitesh Khapra and Dr. Pratyush Kumar from OneFourthLabs ~ https://www.linkedin.com/company/one-fourth-labs/
2. Gaurav Chatterjee (machinelearningman) ~ https://www.linkedin.com/in/gaurav-chatterjee-857813137/
3. Megan Dibble ~ https://www.linkedin.com/in/megandibble1/

Feel free to connect:

LinkedIN ~ https://www.linkedin.com/in/dakshtrehan/

Instagram ~ https://www.instagram.com/_daksh_trehan_/

Github ~ https://github.com/dakshtrehan

Follow for further Machine Learning/ Deep Learning blogs.

Cheers.

By Daksh Trehan on April 15, 2020.

Canonical link

Exported from Medium on July 15, 2020.