Introduction to Data Science

Understanding Data

When we are ready to learn Data Science, It is obvious that we should know what actually a data is. We must also know the various sources of data.

Definitions:

First, let's look at what a few trusted sources consider data to be.

First up, we'll look at the **Cambridge English Dictionary**, which states that data is:

Information, especially facts or numbers, collected to be examined and considered and used to help decision-making.

Second, we'll look at the definition provided by **Wikipedia**, which is:

A set of values of qualitative or quantitative variables.

In both the definition, we got some value that maybe facts(qualitative) or maybe numbers(quantitative).

So for more simplicity, Data is raw fact which can be represented in a meaningful way to form some information. Data mainly comes in the following forms: text, graphics, numerical, voice, and video.

Structural variation of Data in action

In real time scenario, the data we found can be broadly categorize in the the basis of its form:

1. **Structured data –** Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.

2. **Semi-Structured data –** Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to

analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. *Example*: XML data.

3. **Unstructured data –** Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. *Example*: Word, PDF, Text, Media logs.

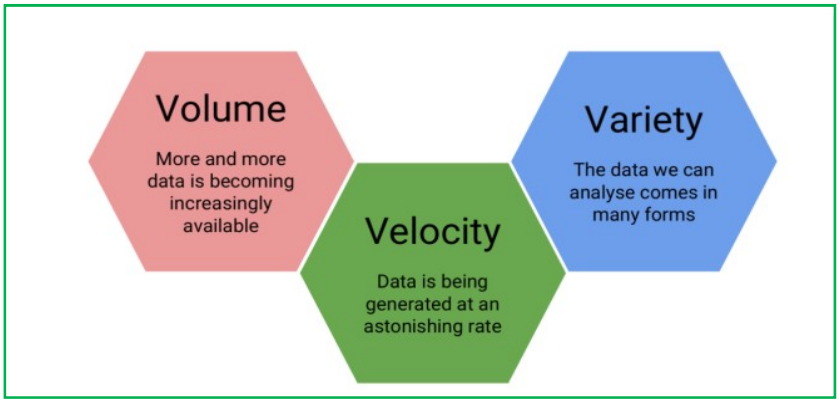
The big data

We'll talk a little bit more about big data in a later lecture, but it deserves an introduction here - since it has been so integral to the rise of data science. There are a few qualities that characterize big data.

The first is **volume**. As the name implies, big data involves large datasets - and these large datasets are becoming more and more routine. For example, say you had a question about online video - well, YouTube has approximately 300 hours of video uploaded every minute! You would definitely have a lot of data available to you to analyse, but you can see how this might be a difficult problem to wrangle all of that data!

And this brings us to the second quality of big data: **velocity**. Data is being generated and collected faster than ever before. In our YouTube example, new data is coming at you every minute! In a completely different example, say you have a question about shipping times or routes. Well, most transport trucks have real time GPS data available - you could in real time analyse the trucks movements… if you have the tools and skills to do so!

The third quality of big data is **variety**. In the examples I've mentioned so far, you have different types of data available to you. In the YouTube example, you could be analysing video or audio, which is a very unstructured data set, or you could have a database of video lengths, views or comments, which is a much more structured dataset to analyse.



A summary of three qualities that characterize big data

Data Science

There are no fixed definition for data Science but we can answer what data science is. Here we will discuss some of the common understanding of Data science definition. Data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

We can form some definition using the important tasks related to Data Science:

Data Science is an interdisciplinary field that allows you to extract knowledge from structured, semistructured, or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from structured, semi structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

Data scientist can be broadly defined as someone:

“who combines the skills of software programmer, statistician and storyteller slash artist to extract the nuggets of gold hidden under mountains of data”

Importance of Data Science

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

Here, are significant advantages of using Data Analytics Technology:

- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinctive business advantage
- Data Science can help you to detect fraud using advanced machine learning algorithms
- It helps you to prevent any significant monetary losses
- Allows to build intelligence ability in machines
- You can perform sentiment analysis to gauge customer brand loyalty

- It enables you to take better and faster decisions
- Helps you to recommend the right product to the right customer to enhance your business

Data Science Components

Data science comprises with various components

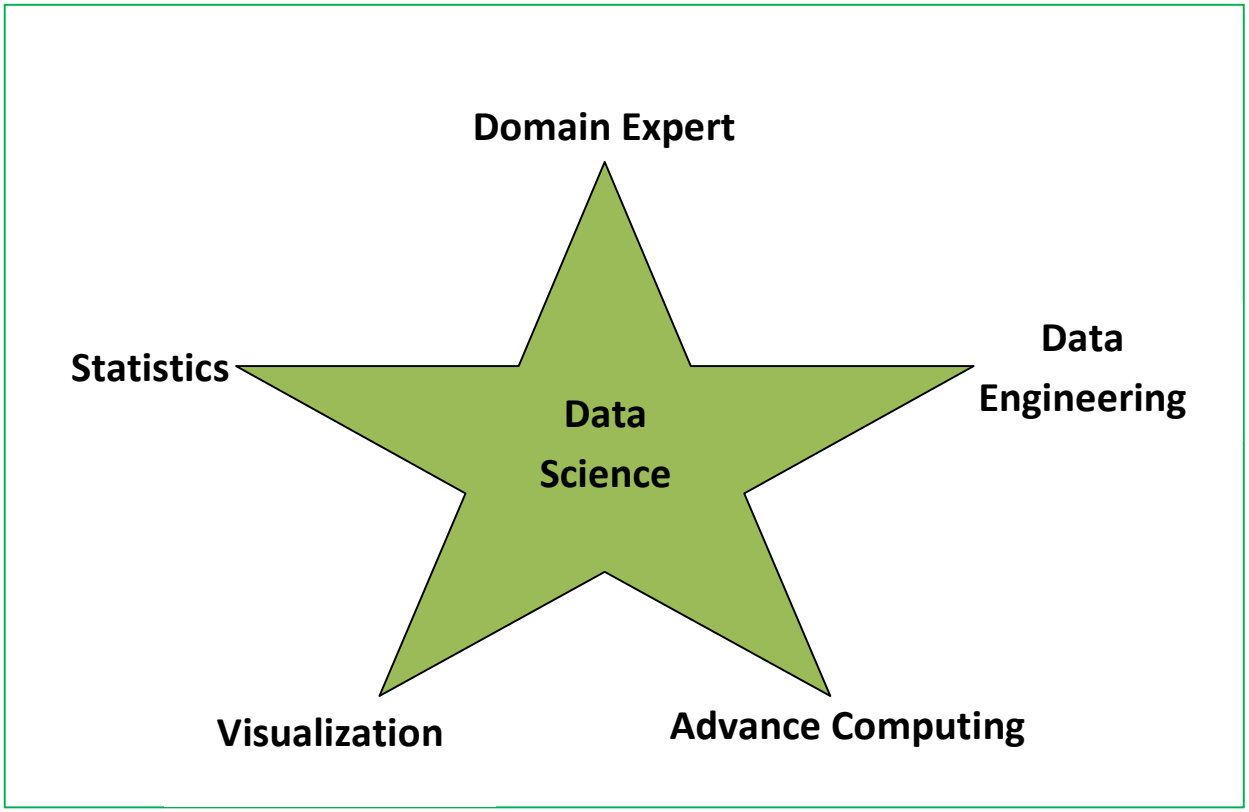


Figure: data Science Components

Statistics: Statistics is the most critical unit of Data Science basics. It is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

Visualization: Visualization technique helps you to access huge amounts of data in easy to understand and digestible visuals.

Machine Learning: Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

Deep Learning: Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.

Data Science Process

The Data Science activity starts with the inquisitiveness of the Data Scientist. A Data scientist must form a question form a problem domain, that can lead a impactful significance. Forming proper question is really a challenging and passionate task. One must have good domain knowledge or passion about the domain to form good question.

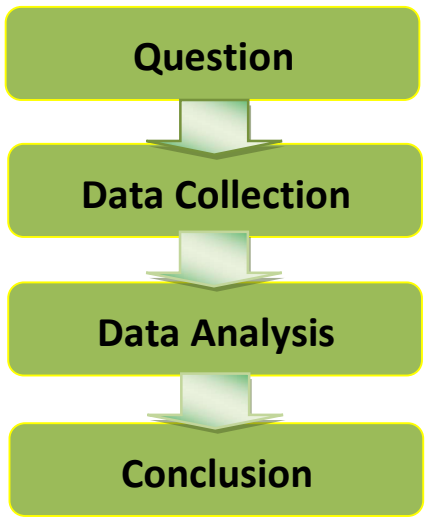


Figure: Data Science Process

The next task is to find related data to solve the question in a convenient and significant way. The data may be structured, Unstructured or semi structured. May be sufficient or incomplete.

Finally, the Model. Using statistical, Mathematics, and computer science, a good model to be prepared to analyze the data to reach to a conclusion.

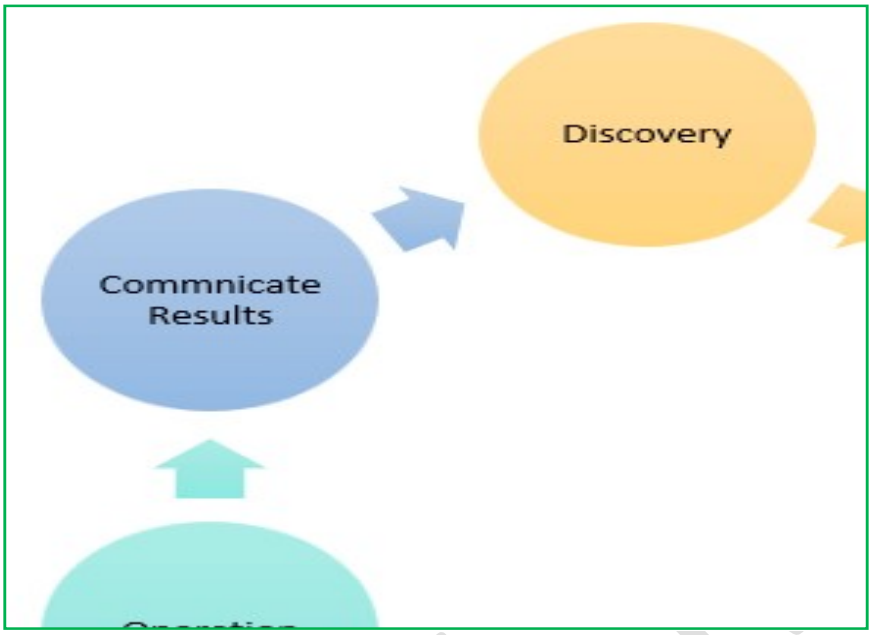


Figure: Data Analysis tasks

1. Discovery:

Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.The data can be:

- Logs from webservers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

2. Preparation:

Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned. You need to process, explore, and condition data before modeling. The cleaner your data, the better are your predictions. In this phase, we need to perform the following tasks:

- Data cleaning
- Data Reduction
- Data integration
- Data transformation

3. Model Planning:

In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas

and visualization tools. SQL analysis services, R, Python, and SAS/access are some of the tools used for this purpose.

4. Model Building:

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

5. Operationalize:

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

6. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

Data Analysis Life Cycle

A typical data science life cycle consists of the following stages:

1. Data acquisition: The primary step in the life cycle of any data science project is to acquire the right data from multiple sources. Data acquisition involves acquiring data from different internal and external sources that can help answer business questions. Data can be extracted from various sources, such as logs from web servers, social media data, online repositories, or databases.
2. Data preparation: Often referred to as data cleaning or data wrangling, it is a critical step in the life cycle. The data collected from different sources is frequently messy and is typically missing various values. Therefore, it is crucial to clean this data to derive value from it.
3. Data exploration: After cleaning the data, you can perform hypothesis testing and visualize the data to understand the data better. Data exploration is sometimes called data mining. It is used to identify patterns in your data set and find important potential features with statistical analysis.
4. Predictive modeling: To train your machine to make predictions, you need to build predictive models. For this, you have to choose the right algorithm on which the machine is to be trained. Historical data is then split into training and validation sets. The model is trained using the training set. The trained model is validated using the validation dataset, and the model is then evaluated for accuracy and efficiency.
5. Model interpretation and deployment: After a rigorous evaluation of the model, you can deploy into a production-like environment for final user acceptance. You'll want to present your model to a non-technical person and convey the actionable insights derived from the data.

Data Science Jobs Roles

Most prominent Data Scientist job titles are:

- Data Scientist
- Data Engineer
- Data Analyst
- Statistician
- Data Architect
- Data Admin
- Business Analyst
- Data/Analytics Manager

Now in this Data Science Tutorial, let's learn what each role entails in detail:

Data Scientist:

Role: A Data Scientist is a professional who manages enormous amounts of data to come up with compelling business visions by using various tools, techniques, methodologies, algorithms, etc.

Languages: R, SAS, Python, SQL, Hive, Matlab, Pig, Spark

Data Engineer:

Role: The role of data engineer is of working with large amounts of data. He develops, constructs, tests, and maintains architectures like large scale processing system and databases.

Languages: SQL, Hive, R, SAS, Matlab, Python, Java, Ruby, C + +, and Perl

Data Analyst:

Role: A data analyst is responsible for mining vast amounts of data. He or she will look for relationships, patterns, trends in data. Later he or she will deliver compelling reporting and visualization for analyzing the data to take the most viable business decisions.

Languages: R, Python, HTML, JS, C, C+ + , SQL

Statistician:

Role: The statistician collects, analyses, understand qualitative and quantitative data by using statistical theories and methods.

Languages: SQL, R, Matlab, Tableau, Python, Perl, Spark, and Hive

Data Administrator:

Role: Data admin should ensure that the database is accessible to all relevant users. He also makes sure that it is performing correctly and is being kept safe from hacking.

Languages: Ruby on Rails, SQL, Java, C#, and Python

Business Analyst:

Role: This professional need to improves business processes. He/she as an intermediary between the business executive team and IT department.

Languages: SQL, Tableau, Power BI and, Python

Tools for Data Science

Applications of Data Science

Now in this Data Science Tutorial, we will learn about Applications of Data Science:

Internet Search: Google search use Data science technology to search a specific result within a fraction of a second

Recommendation Systems: To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

Image & Speech Recognition: Speech recognizes system like Siri, Google assistant, Alexa runs on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

Gaming world: EA Sports, Sony, Nintendo, are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.

Online Price Comparison: PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

Challenges of Data science Technology

- High variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Data Science results not effectively used by business decision makers
- Explaining data science to others is difficult
- Privacy issues
- Lack of significant domain expert
- If an organization is very small, they can't have a Data Science team

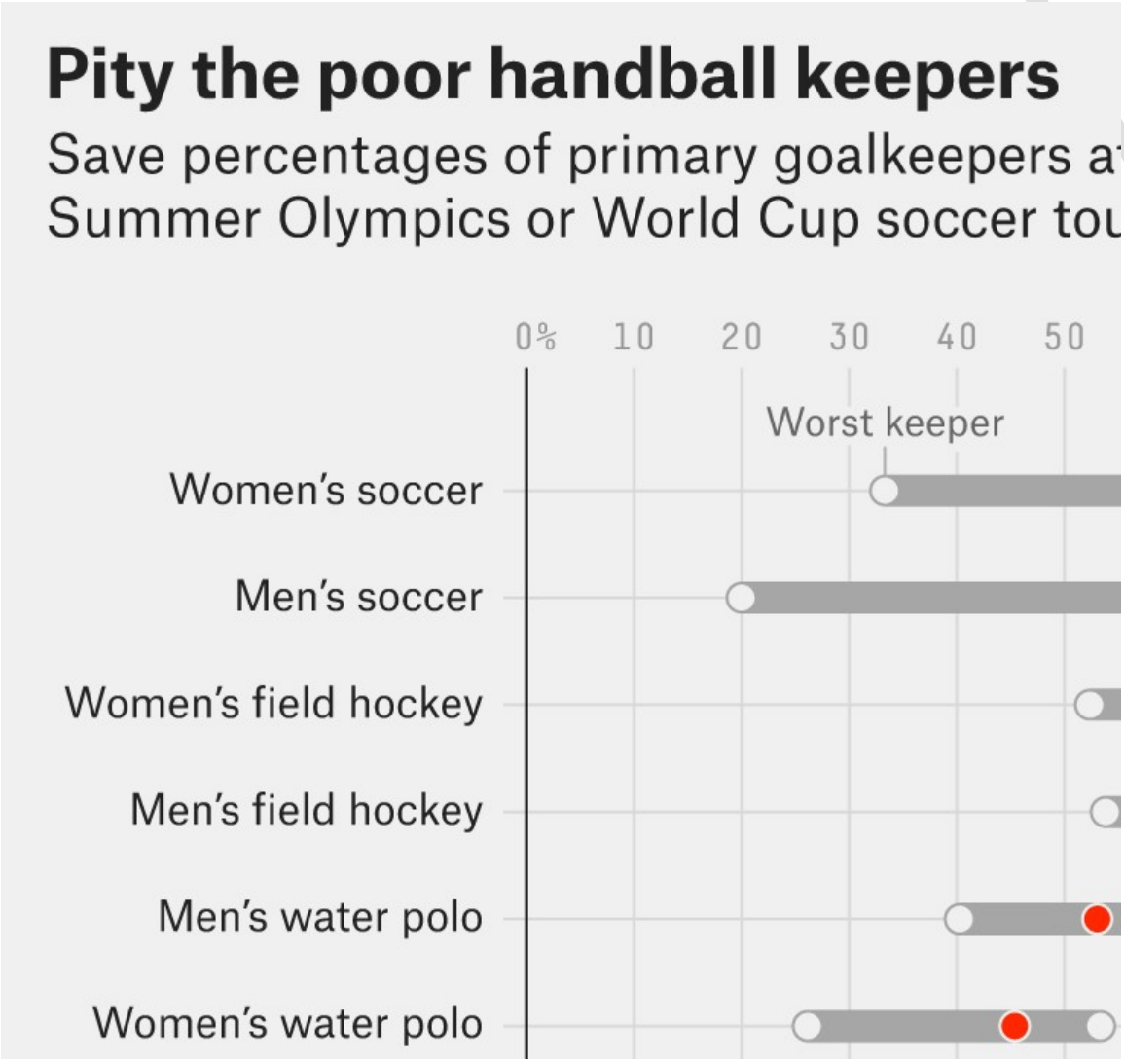
Data science in action

Example 1:

Here is case study of Data analysis, finding the sport where it was harder for Goal Keepers (https://fivethirtyeight.com/features/which-olympic-sport-is-hardest-on-its-goalies).

Here the problem statement (question) is pretty clear “Which Olympic Sport Is Hardest On Its Goalies?” Now they collected summer Olympics data from 2012, Soccer world cup data from 2014 and 2015 for men and women. Data source: SPORTS-REFERENCE.COM, BARRIESVIEW.COM.

Women’s soccer goalkeepers have some of the highest save percentages of any Olympic sport:



Soccer and field hockey keepers have the “easiest” jobs, generally saving around 70 percent of shots. Soccer goalies see far fewer attempts, though, leading to much more variance throughout a tournament (that low point for men’s soccer is English keeper Joe Hart, who saw only five shots at the 2014 World Cup but let in four). Water polo goalkeepers let in shots at a

much greater clip, especially on the women's side, but handball goalies have it worst of all, letting in around seven in 10 shots. But the saves, when they come, can be spectacular:

Example2:

As like most of the Indian we are a more or less interested about cricket. Here let us know, can we get answer of the following question applyiing simple Data Science techniques:

1. What is the team average when batting first and second?
2. How frequently does the Indian team win a match?
3. What is the probability of India winning a match against a particular team?
4. What is the target to be set by the Indian team to win the match?
5. How many times has the Indian team defended a low scoring target?
6. Which was the most successful year for Team India?

Ofcourse we do have solution for all of them.

Collect the commentary of the last 4 years of the T20 matches played by India. Download a sample dataset from

"https://drive.google.com/file/d/1gkQ_IoqJBGIpqdWqBTdEc9umh-Tal8Rh/view".

It's time to analyze the commentary and find some appealing insights. Let's do it!

(The example is taken from https://www.analyticsvidhya.com/blog/2020/02/sports-analytics-generating-actionable-insights-using-cricket-commentary/. For solution and code the link canbe visited)

Summary

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.
- Statistics, Visualization, Deep Learning, Machine Learning, are important Data Science concepts.
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results.
- Important Data Scientist job roles are: 1) Data Scientist 2) Data Engineer 3) Data Analyst 4) Statistician 5) Data Architect 6) Data Admin 7) Business Analyst 8) Data/Analytics Manager.
- R, SQL, Python, SaS, are essential Data science tools.
- The predictions of Business Intelligence is looking backward while for Data Science it is looking forward.
- Important applications of Data science are 1) Internet Search 2) Recommendation Systems 3) Image & Speech Recognition 4) Gaming world 5) Online Price Comparison.
- High variety of information & data is the biggest challenge of Data Science technology.

Reference

1. https://www.javatpoint.com/data-science
2. https://www.tutorialspoint.com/python_data_science/index.htm
3. https://www.guru99.com/data-science-tutorial.html
4. https://www.analyticsvidhya.com/
5. Python for Data Analysis, by Wes McKinney, O'Reilly 2012, First Edition.