Session – Summary

Introduction to Data Analytics – Jargon Busting

With the fast-paced growth of data science in the past decade, a lot of confusing jargon with overlapping meanings has been introduced in the industry. In this session, we looked into this jargon in great detail.

Data Mining

Data Mining: "non-trivial extraction of implicit, previously unknown and potentially useful information from data"

The term 'data mining' is a misnomer in the sense that the aim of data mining is not to mine data but to mine insights from the data.

It's important to remember the key adjectives:

- Implicit: Information that is not provided intentionally but gathered from available data streams

  e.g. From your travel history on Yatra.com, one can identify your hometown
- Previously unknown: The data is collected rather than explicitly given
- Potentially useful: It may be useful for drawing actionable insights

While dealing with implicit information, you must always be careful as there is no certainty regarding these insights, so you always need to be a bit careful about what you conclude.

Machine Learning

In machine learning, instead of explicitly telling the machine what to do, we let the machine learn on its own, by providing what we call the training data set.

Machine learning is not just the computer's ability to do coded arithmetic calculations or any other processing instructions that you give. Thus, a computer program using a mathematical formula to estimate the unemployment rate for next five years is not a machine learning system.

Popular examples of machine learning:

- Google's self-driving cars
- Google programme, AlphaGo. It was able to defeat the Korean champion in the game of Go, after it learnt from data from millions of past games
- Blocking of suspicious credit cards
- Recommendation engines on an e-commerce site
- Automatic classification of your email into regular email versus spam. To train a machine learning algorithm you can run it with data about past classification of regular and spam mails

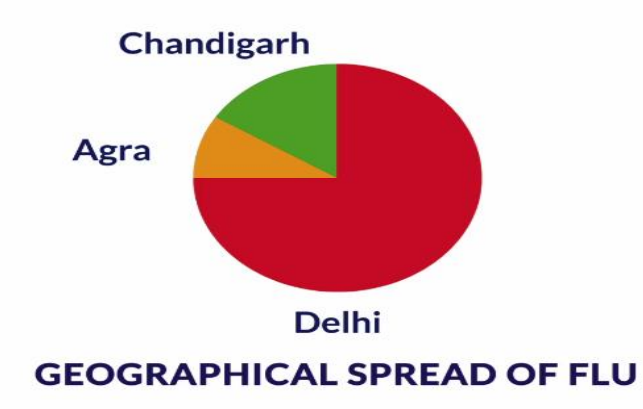
Descriptive, Exploratory and Confirmatory Data Analytics

Descriptive analysis consists of describing or summarising the data for you. It tells you about various summary statistics such as mean, median, quartiles, variance and many more. Descriptive analytics is important to gain an idea of the data you are dealing with.

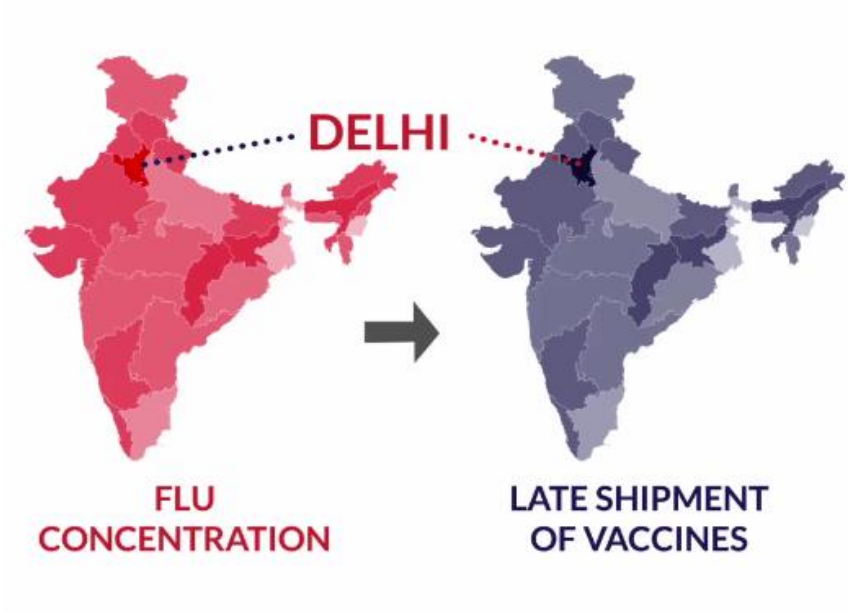
Exploratory data analysis, or EDA, is when new features in the data are discovered. It also helps us formulate basic hypotheses about the data.

Let's revisit the example we discussed in the lecture.

How can descriptive analytics and EDA help in analysing the data gathered from around the country to improve the response time of National Centre for Disease Control during flu season?
Suppose you want to know about the geographical spread of flu for the last winter. Descriptive analytics can help you in generating basic reports about the spread of flu such as the areas most affected by it.



Knowing that information, you might want to better prepare the state for the next winter. These types of queries examine past events, are most widely used, and fall under the descriptive analytics category.



In the graphic above, it can be visually detected that there is a direct correlation between the intensity of flu outbreak with the late shipment of vaccines. So essentially, exploring the data and figuring out some patterns based on which a certain hypothesis can be formulated, i.e. delayed vaccine shipments cause flu outbreak. This is called exploratory data analysis.

Confirmatory Data Analytics:

Confirmatory analysis is when existing hypotheses are proven or confirmed to be true or false.

In the example above, Confirmatory analytics can be performed to test and confirm the hypothesis through statistical means that a factor is causing a change in the outcome. The statistical analysis shows that there is an inverse relationship between the availability of vaccines and flu incidence. Such analyses are called confirmatory analyses because they help you to confirm the hypothesis that you made initially.

Predictive Data Analytics

Predictive Data Analysis:

Predictive data analysis is used to predict the future in cases where the past patterns are understood.

Let’s understand it with an example discussed in the lecture:

If a credit card company is trying to sell different credit cards to different prospects or different customers, first they would like to understand the behaviour of the customer segment. For example, they have built four-five different clusters or segments of the target population. Let's say that a particular cluster is a high transactive one, or a segment has customers with a high revolving nature. A third might consist of very frequent travellers and a fourth might be high shoppers. Let's say a particular segment or a particular cluster has all the frequent flyers, then the co-branded credit card of the company with an airline can be given to that particular segment. So, selling the right particular product to the right customer at the right time is very important. And that's how predictive analytics can help banks.

Big Data

In short, big data + analytics = big data analytics. Big data has three elements: velocity, volume and variety. Velocity means that data is generated extremely fast and often continuously processed. Volume simply means large amounts of data, like a million rows in an Excel sheet. And variety means different types of data, like a really large number of columns in an Excel sheet containing text, numbers, pictures, symbols in Greek and Hebrew, etc. For example, Google processes 20 petabytes of data every day.

We call it big data not only because the amount of data is large, but that it is acquired very fast, often continuously and the type of data varies a lot. And also, maybe, because of the huge impact it has on our world today.