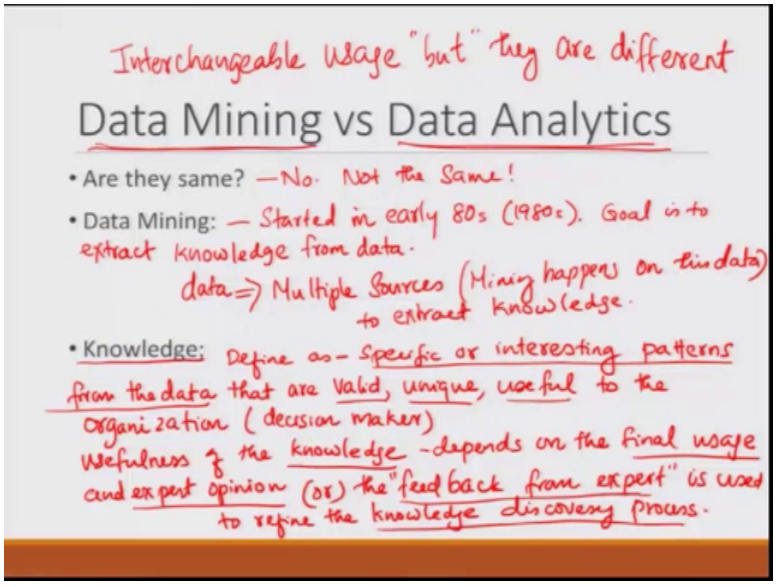
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 01
Introduction

Good afternoon. Today, I welcome you guys to the first lecture or the first module of the analytics course, the applied analytics which is a practitioners approach in descriptive prescriptive and predictive analytics. So, we will see how this course is contains a lot of materials and we will try to cover them as quickly as possible and we will try to cover this from the applied side, mostly focusing towards what industry is currently looking for. So, today is kind of the introduction topic and we will try to cover the major aspects of it.

(Refer Slide Time: 00:51)



So, the first place where we actually want to start is the question about data mining and data analytics. These are two terms that a lot of people use interchangeably they look at it. So, we can say these as the interchangeable usage, but they are different. In this regard should be very clear that both data mining and data analytics are different things, but people do use them interchangeably. So, the question is, are they the same? The answer is, no. They are not the same, first part.

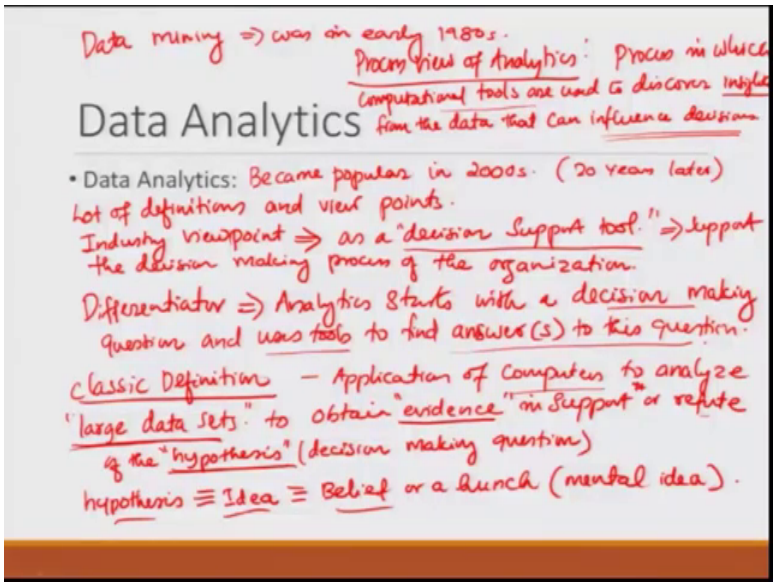
So, what is data mining? This was something like started in early eighties, I mean 1980's and what this is the, its goal is to extract you have to extract knowledge from data. So, your aim is to extract knowledge from data. So, data of you can say multiple sources or from different locations you get data and then this data from different sources is looked for, is studied or mined; here is mining happens on this data from various sources to extract knowledge.

So, what do we mean by knowledge here? The term knowledge in this case we can define this. We define it in the context of our analytics course we kind of define as, what do we define as specific or interesting talking about specific or interesting patterns from where, patterns from the data that are valid these patterns are valid, they are unique and they are useful to whom, the organization or they can talk about it as the decision maker. So, here we are talking about specific or interesting patterns, that these patterns where do you get them these patterns are obtained from the data and what are the characteristics of these patterns? They are valid; they are unique and useful to the organization or the decision maker.

The important part is the usefulness of the knowledge, whether it is useful to the organization is not? It depends on the usage, on the final usage and expert opinion all right. So, the expert who is dealing with this knowledge that is obtained from the data depends on the final usage of this knowledge and the expert decides whether this is useful or not. In other words, else or the feedback from expert is used to refine the knowledge discovery process.

So, here what we are talking about it is whatever the feedback the expert gives about the usefulness of the knowledge or whether it is the valid unique and useful to the organization that feedback, this feedback from the expert is taken and used to refine the knowledge discovery process or the data mining process. So, this data mining to a large extent is about looking for interesting patterns from the data, that is the data mining aspect.

(Refer Slide Time: 06:01)



So, what is data analytics? This became popular in 2000's, about 20 years later, 20 years after the data mining. So, data mining came first. So, remember data mining was in early 1980's. So, you are talking about 2000's, there are many definitions out of this lot of definitions and viewpoints about it what we are going to present here is the industrial the industry viewpoint. What is it? The industry views analytics as a decision support tool. This is the tool that we use for supporting the decision making process or here you are saying support the decision making process of the organization. So, the main thing is in a decision making support that they think the difference, the main differentiator is analytics starts with a question, with a decision making question and uses tools to find answer or answers to the this question.

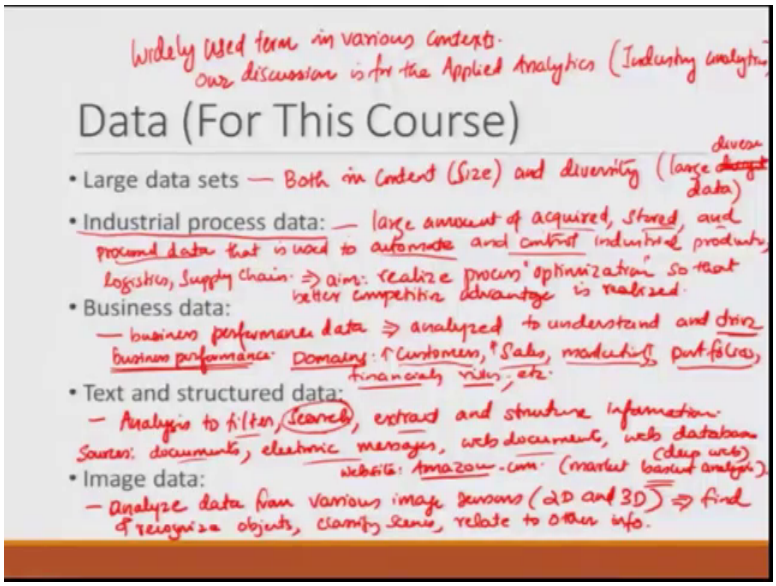
So, you have a decision making question and you want to use various tools to find answers to this question. So, the major differentiator is in mining you do not really know what the data is about or what the there is no question you are just looking for interesting patterns whereas, in data analytics you have a specific question as a decision maker and you are using the data to find answer to the question. The classic definition of this of analytics, it kind of mentions that it is actually an application of computers to analyze it is an application of the computers to analyze large data sets this is important you are looking at specifically at large data sets to obtain instead of the patterns you are looking for evidence here in support or refute of the hypothesis.

So, the new term here is evidence in support or in evidence not in support evidence. So, it is a process of analyzing large data set where you are looking for evidence to either support or refute, support or oppose the hypothesis. Hypothesis, this is actually the decision making question. What is hypothesis? It is a decision making question or what some people also call this for this course we can also take hypothesis we can equate it to equivalent to an idea or you can equate it to a belief or a hunch even like a mental idea you get a hunch. So, there are many ways you can think about it.

So, the decision maker have an hunch or an idea or a hypothesis or a belief to which he is going to use a large set of data tools and uses computer to various computer tools. So, analyze this data tools to find evidence in support or refute of this hypothesis. That is the major difference between analytics and data mining and other aspect of this is also some people also say that analytics you know they look at it the process view of analytics. What is the process view of analytics? So, it is a, we will say it is a process in which what are you doing, computational tools are used to discover insights. So, it is a process in which you are using computational tools to discover insights from the data that can influence decisions. So, the insights here are more towards influencing the decision or the question.

So, the process in which computational tools are used to discover the insights from the data to the, so that the decision maker the decision can be influenced or the initial question of the decision maker can be answered. So, the most important aspect is we start with a hypothesis in the data analytics whereas, in data mining there is literally no hypotheses are such.

(Refer Slide Time: 12:27)



So, from this we move now to the major concept is the data and data are such it is a widely used term in various contexts. Here, we are using it for our discussion; we need to remember this very clearly is for the applied analytics side, but what we call as the industry analytics. So, for the data of this course we are predominantly focusing on large data sets I mean you know both it is large both in content can talk about the size and diversity. People talk about big data and we really do not like the term actually you can re rather talk about us large divergent data or instead of divergent, large and diverse data. So, diverse amount of information is available and there is large amount of it.

So, some of the examples of data that we will be dealing in this class will be industrial process data. So, what is an industrial process data it is typically a large amount of data acquired, stored and processed data that is used to automate and control industrial production manufacturing or logistics supply chain etcetera. So, you have a large amount of acquired stored on record processed data that is used to automate and control the industrial production logistics and supply chain because typically you are dealing with the processes of the industry processes the main aim an aim of this is to realize process optimization. So, that better competitive advantage is realized.

So, you are talking about reaching optimizations, the main aim here is optimization. So, when you have large scale of industrial data which is stored and processed to data which is acquired stored and processed to data, it is used for automate and control the industrial

production logistics supply chain etcetera with the main aim of realizing process optimization.

Then there is something called as a business data which is mostly called as business performance data and why is it used and this data is analyzed to understand and drive business performance. So, the aim here is analyze so that you can drive the business performance that is the most important aspect of it. So, the major thinks the domains here where are these data applicable to? The domains is like customers, sales, marketing, portfolio, financials, risks, etcetera. So, these are the major domains in which business data is used, where you are looking at driving performance of.

In the case of customers, you will be talking about increasing the number of customers; sales, obviously, increasing the number of sales or the volume of sales or the amount of sales; marketing is how to do if actively do marketing or the product; portfolio is what are the product portfolios you need to carry and financials will be like what is the cheap best way of obtaining maximum money out of what we have the best way to investment, how to reduce the risk in the industry etcetera. So, these are the aspects of the business data. So, the main aim is dry business performance.

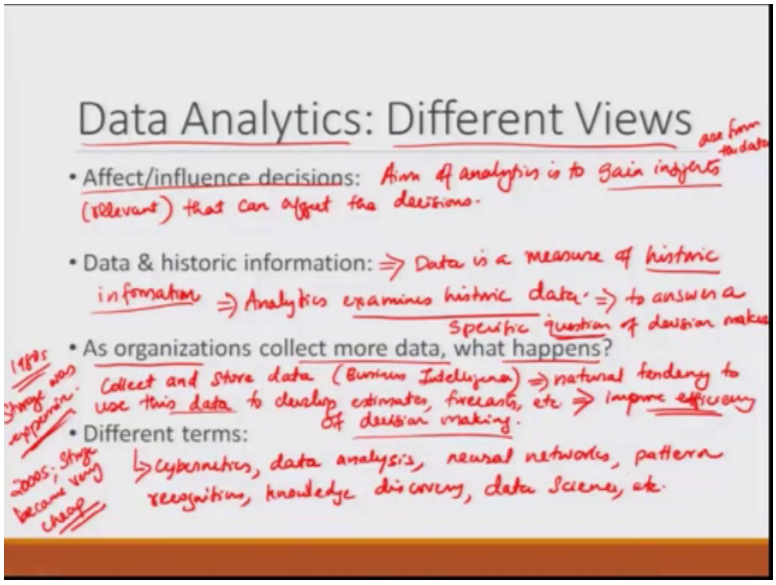
Then, the third part is structured and text and structured data that we talk about. It is the main aim is to do analysis to filter, search, extract and structure information. So, here most of the time you are talking about textures and other things and sources like we have data sources like text documents, we have electronic messages, we have web documents and we have web databases sometimes also people call us deep web, this new term. So, here we are talking about data from these kinds of sources like documents, electronic messages, web documents, web databases etcetera is then filtered searched and extracted.

So, on a classic example of this is if you go to the website of Amazon and you search for something Amazon dot com and you search for a product and then you get similar searches it is called as market basket analysis, which is a new analytics tool and the aim here is using the text data that you are using for searching your particular product it is able to show you what similar other processes are available. Similarly, the image data is you know analyze data from various image sensors you have both 2D and 3D images alright and why are you analyzing? You are analyzing to find is basically find and

recognize objects, then you are also trying to classify the scenes relate to other information.

So, somebody says here is a traffic jam that is going on in a particular part of the road and then you have a unmanned aerial vehicle that is flying on the top of it and you get the image data and if you see that there is a huge crowding of vehicles in a particular place and yes, that corroborates the plane then there is a traffic jam and then you can also use the data to find out what are the best or most efficient ways to clean the jam. So, that type of analysis is also real time analysis is also possible with this.

(Refer Slide Time: 20:28)



Now, we talked about there are data analytics, there are many different views. We kind of talked about analytics and why is the hypothesis an important aspect of analytics, but main aspect of it is we said that it can affect or influence decisions. So, the aim of the analytics is to gain insights or you can say these insights they are relevant insights that can affect the decisions that is the plan. So, the aim of analytics is to gain insights from the data. So, these insights are from the data, then these insights can affect the decisions and typically what is data, there are many ways people define data and we also have seen different ways of data, but in this course we took at data as a measure of historic information.

So, you are kind of saying that is something we are measuring the historic information of this which implies analytics examines historic data. So, an examination of the historic

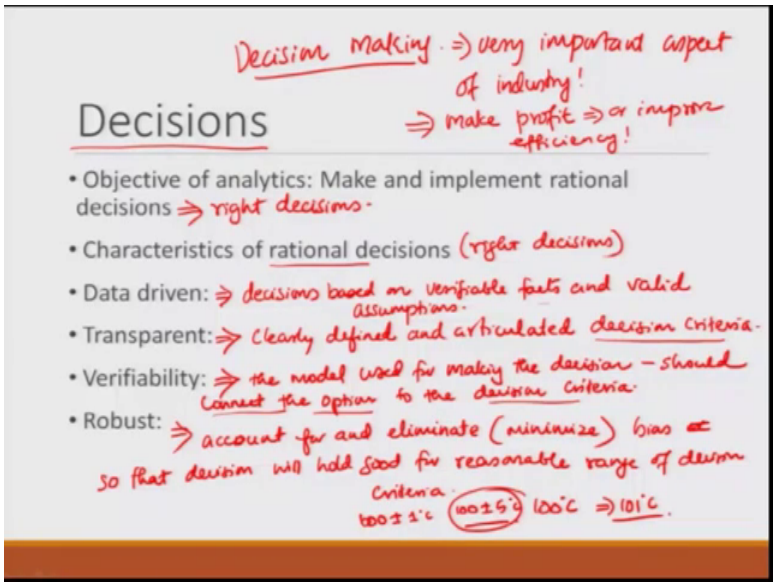
data can also be told as analytics. Why are you examining this? To answer a specific question of decision maker or you can instead of this question you can use the words like idea, hypothesis etcetera. So, you analyze this or you examine historic data to answer a specific question of the decision maker.

So, the funniest part is when the organization started to collect more data then what happened; so, when you start collect more data and store it. So, collect and store data certain people call this business intelligence, it is a terminology. Let us not worry about too much about it, let us say that the organizations are collecting and storing data when you have that then only have we say natural tendency to use this data to develop estimates, forecasts etcetera for why improve efficiency of decision making.

So, the process started by organization starting to store data. So, historically if you look at it in 1980's storage was expensive, when we reach the 2000's storage became cheaper, very cheap. So, when the organization started to store a lot of data the natural tendency of uses data to develop estimate forecast the ultimate aim is to improve the decision efficiency of the decision making process. So, there are many terms in the meantime that came out of this and these terms like you know cybernetics, data analysis, neural nets or neural networks then you have pattern recognition, then you have knowledge discovery, then you have something called as data science etcetera.

There are many terms that are used loosely in this regard, we use all these terms into different aspects of this and the main aim is again as I said earlier this started with the process of collecting more data once we started collecting large amount of data then we had a natural tendency to do use the data to improve your efficiency of the decision making process in which, this historic data amounts to providing new tools for the data analytics.

(Refer Slide Time: 25:24)



Can we now talk about something called decisions or something called as decision making which is an important this is a very important step, important aspect of industry. You are making a lot of decisions and the main aim of industry is to make let us assume that the main aim is here in this case is make profit or improve efficiency. If you are a nonprofit organizations then you would definitely want to improve your efficiency otherwise make profit, that means, you have to make more money so that you can do better business.

So, the one of the aim is you should be able to make the right decisions or the some people call it as right decisions we or call it as the rational decisions make an implement rational decisions and what are rational decisions or what are these right decisions we kind of using this term right decision here on rational decisions interchangeably, but they are different in the semantics for the meaning of the word, but what we are going to talk about here is the decisions that are for the industry when you are making the decisions they should be data driven. Data driven means decisions based on verifiable facts and valid assumptions.

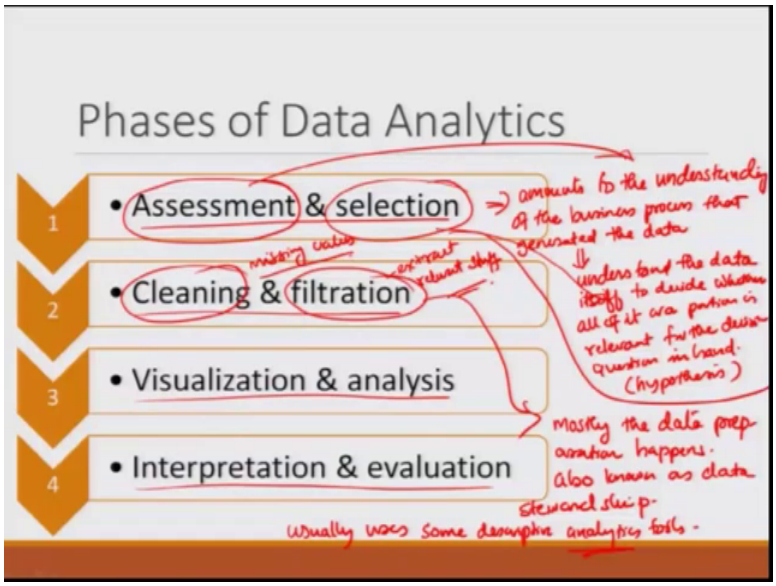
So, for example, is if you are somebody ask you what is going to be the production of wheat in India for the next year, you will probably look at the last 10 years production of wheat and from there we would generate forecasts for the next year and we assume maybe or the assumptions provider will be like we will have a normal monsoon or 80

percent monsoon, 90 percent monsoon, but if you make a production estimator and say that the monsoon will double next year, probably that assumption will not be a valid assumption, so, you have a valid assumption. Valid assumption is kind of valid as far as we are concerned and then we also have verifiable facts. So, that is one prospect of the rational decision process.

Second one is what we call as it should be transparent. Transparent means there should be clearly defined and articulated decision criteria. The criteria that you are using to make the decision what should be clearly defined and articulated, you cannot make a whimsical design it should be based on the, it should be data driven and from there once you have the data when you have the verifiable facts and valid assumption then your decision making criteria should be clearly articulated and once that it is done you also should have what we called as a verifiability; that means, the model that you use in making the decision model, used for making the decision. What decision? The rational decision, it should connect the options to the decision criteria. So, the criteria that you used the transparent criteria, transparency part that you would used it should connect with the option whatever the choice you are going to make as the decision that is important here.

Now, the robustness the last part of the robustness it is, that means, the decision the rational decision should also account for and eliminate; if you can not completely eliminate, minimize the bias. Why do we eliminate or reduce the bias? So, that decision will hold good for reasonable range of decision criteria. If you say that this decision will work at 100 degree Celsius and you can say this will not work 101 degree Celsius, then it might not be a really good decision. So, you probably say that fine we designed a particular part that should work at 100 degrees plus or minus 1 degree and somebody says I can do it a plus or minus 5 degree then, obviously, the customer would prefer this because you have a wider range of temperature within which the product will work. So, that is the idea well this also good it is just a very loose example, but still we can understand that the range you should also be the decision if it works on a larger wider range and the decision holds good is typically called as a robust decision as well.

(Refer Slide Time: 31:04)



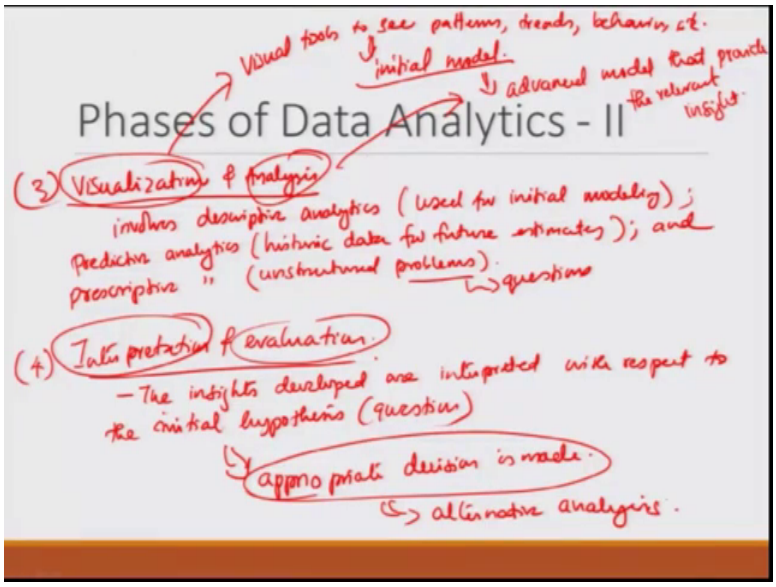
Now, there are certain phases of data and analytics data analytics is done in this you can think about the entire processes into these 4 phases. The assessment and the selection being the first phase, cleaning and filtration the second phase, visualization and analysis the third phase and interpretation and evaluation the fourth phase. So, if we talk about the assessment phase, the first phase, what is it about this it is basically amounts to the understanding of the business process that generated the data. So, the first part is it amounts to understanding the business process that generated the data then what does it amounts to? It also can understand the data itself to decide whether all of it or a portion is relevant for the decision question in hand.

So, here we are assuming that for this aspect this particular aspect we are talking about first part is understand the business process that is assessment, when you are assessing that part you are understanding the business process from where the data was generated and from there once you understand the business process then you are going to decide what part of the data whether the whole data or a subset of it should be used before deciding whether the decision should be made on and what whether it fits the decision criteria the question in hand or this is your hypothesis.

Then you have the second part which is called as a cleaning and the filtration, where this is a mostly the data preparation step happens here. So, what do we do? This also known as data stewardship and what do we do is here you kind of clean the data so, here is like

you know if there is missing values or there are errors in the data all those kinds of things are done there. And, then you filter the data in such a way that you extract relevant stuff you filter it. So, that whatever is needed because the data might be very large, but you would only require a small aspect of it. So, you take that and you do this. Sometime usually uses some descriptive analytics tools. So, in the second phase we have the usage of certain descriptive analytic tools we will see large later in the course.

(Refer Slide Time: 34:44)



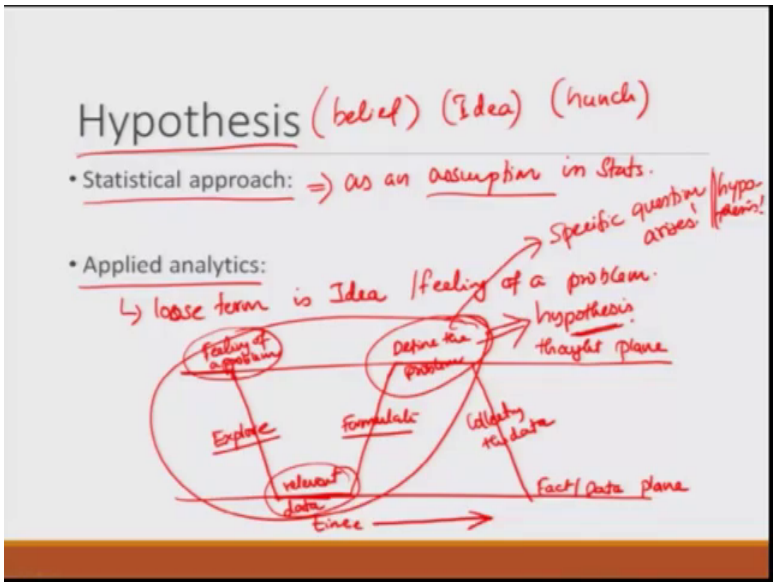
Then, also we talked about the visualization. If you talk about visualization the third step being visualization and analysis, this is our third step. So, what are we doing here the visualization and analysis is it involves descriptive analytics. So, this is used for initial modeling or building the models; then we have is the predictive analytics which is used for you know use the historic data for future estimates and prescriptive analytics where what we are talking about is unstructured problems, how do we model them and analyze that unstructured problems or questions, decision questions, that kind of things.

So, somebody says how can I double the profit in the next year that is kind of a very unstructured question which might require a lot of analyzing lot of different aspects of it. So, the visualization is first looking at the data we should be finding out using of visual tools to see patterns, trends, behavior etcetera. So, that what happens is with these you get the initial model from which using that model then you use the analysis from where

you derive the advanced model that provide the relevant what are we looking at relevant insight. So, that part gets out, it comes here.

Then, we want have the last step which is called as interpretation and evaluation. So, what are we doing here, we are interpreting and evaluating. So, here the insights developed are interpreted with respect to the initial question, initial hypothesis or question, we are doing that. Why are we doing this? So, that appropriate decision is resulted, appropriate decision is made. So, you are interpreting the insights that are developed so that the appropriateness of the decision is made, so that is where the evaluation. So, this is also called as alternative analysis as well. We will see all of this in the later classes in a much bigger way, but we are trying to talk about the major steps of this.

(Refer Slide Time: 38:19)



And, now we talk about something the last part which is called as the hypothesis which is also an important aspect of this and we mentioned hypothesis, some people call this in simple ways as a belief, some people call this an idea, some people call this as a hunch. It does not matter what you are going to call, but in statistics it is considered as an assumption in stats. So, statistics consider the hypothesis as an assumption.

So, it is based on something it is based on some information that is available to you. In the applied analytics, in analysis or in analytics we can de call this the loose term is idea or some people have also call it as a feeling of a problem. So, this usually comes from

the prior experience and other things. So, if you think about it the way we think about this or if you think about this as your thought plane this credit goes to this idea of this actually goes to process Shoji Shiba from Japan. So, you think about two parallel lines; one is the thought plane and another is the fact or the data plane, think about there is a fact than there is a thought. At some point of time somewhere as the time progresses you think about here is a time is going like this, some point of time you get a feeling some place some particular instance.

So, this is where a feeling of a problem or an idea of a problem from there once you have that then you move yourself to the data plane with the process called explore, to see whether the problem the feeling about the problem is right and then you look for what equal as relevant data, you travel in time and you try to find relevant data. Once you have that once you are convinced by this stuff then you go up again back to the thought plane by something a process called to formulate. You are formulating this and what you are doing is, the formulation in the thought process along with the data you end up defining the problem. So, once you define the problem, then you again move back to the data plane and you start collect collecting the data.

So, somewhere here in this much aspects are where you can think about the hypothesis concepts comes into picture. You have some prior experience and you had a gut feeling from there you explored with the relevant data and you formulated. So, this is where the specific question arises, and then from there this aspect this gives you the hypothesis.

So, we hope this makes sense to you and to an extent. So, we conclude this lecture right here and the basic concepts of the analytics is being explained to you and from tomorrow onwards we will start to get into different aspects of it, different types of data and other details of it and how we can use statistics to do this. So, in the meantime, continue your learning and how fun while you are learning.

Thank you.