Course Name:Business Intelligence and Analytics

Professor Name:Prof. Saji.K.Mathew

Department Name:Department of Management Studies

Institute Name : Indian Institute of Technology Madras

Week:01

Lecture:03

VOCABULARY OF BUSINESS ANALYTICS | BI&A

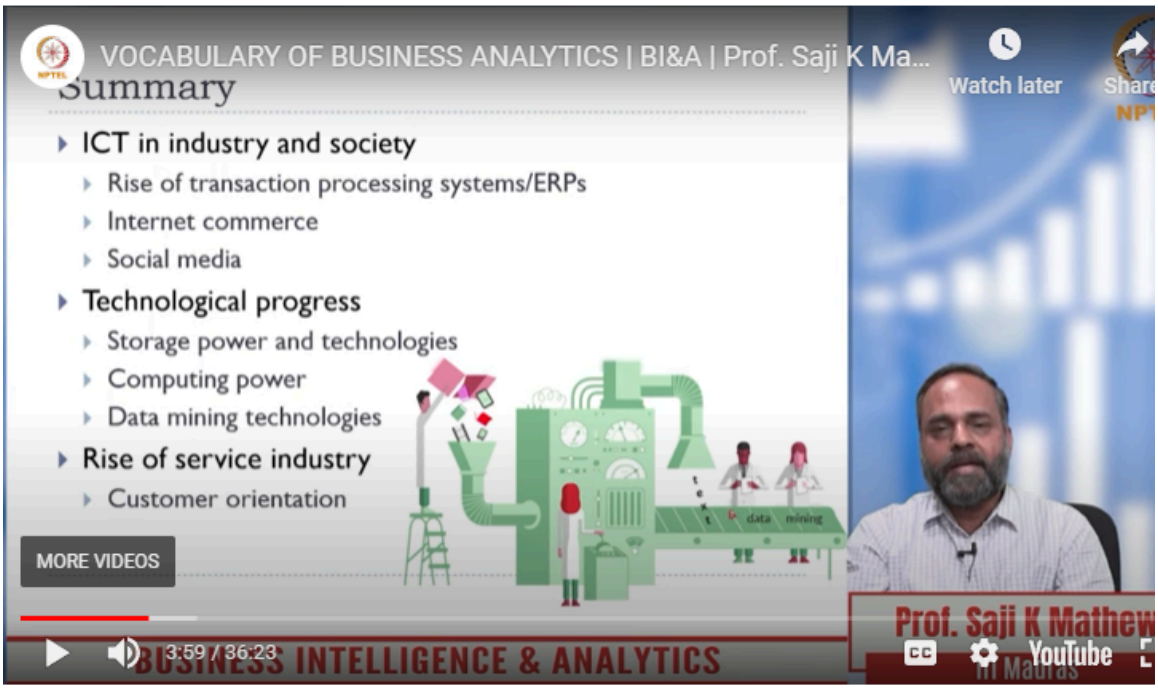
Okay, so I would summarize it here as to what drives if somebody ask you what are the drivers of business intelligence and analytics? Why are these terms very popular today? I was trying to give you a picture, a historic overview of the growth of this concepts, not just concepts, growth of these practices in industry as valuable value adding practices, increasingly value adding practices, not only to support business decisions, but also to conduct business. So, algorithms have, not only algorithms are supporting business decisions, but they are also today taking decisions. You know, so that capability is something that is very new, thanks to data. So, data driven decisions and data driven analytics is driving both technology and non-technology industry. So, historically this has come from transaction processing systems automation through ERPs with database as a byproduct and then internet commerce and social media, we have seen this.

The other three drivers that has enabled analytics and data science today is the drop in cost of storage power and technologies or the advances in the technology itself in storage power and computing power and data mining technologies or analytics technologies or analytics algorithms which have become very resourceful and powerful today to particularly analyze large volumes of data, methods to data management techniques and analytic techniques that could handle and analyze large volumes of data; that those are the advances in data mining technologies in the current era that is another driver for business intelligence and analytics.

And the third driver is the rise of service industry. Now all this may exist, you know, high technology exists does not mean that it is used. Just because you have high computing power or data mining technologies or storage does not mean that the industry will use it, industry will use a technology if it is valuable to them. So, the rise of service

industry in particular, particularly competition is driving the use of analytics today.

Data driven decisions make you smarter. You are able to make smarter decisions than your competitors. And if you have access to valuable data that itself become a valuable, rare and inimitable resource in terms of strategic theories. So, when you have resources that are unique and inimitable and rare, then you have a competitive advantage over your rivals because rivals do not have access to your data. Rivals do not have access to the algorithms that you may develop or enhance or train using your data and therefore, your decisions may be better than your rivals decisions, in principle.



So, the ability of analytics to give you competitive advantage has led to adoption of this systems in large scale in the industry. And you look at the retail industry, for example, look at any industry, use of analytics is widespread in different areas. Analytics has become a source of competitive advantage for organizations. So, that is another important aspect when you look at analytics and data science. Let me present you some terms that are related to, some terms which are related, I do not say they are related to this or that, but I am just throwing some words or some keywords on to the screen.

So, you will hear about these terms in this course like data warehousing, data marts, databases, ETL, which stands for extraction, transformation and loading. And you also hear about big data and data engineering. That is one category in the vocabulary of analytics. And then there is another category, which is mentioned subsequently, which involves data mining, knowledge discovery in databases or in short KDD, OLAP, which stands for online analytical processing, then the term analytics itself. And advanced

analytics, if you go to some websites of supply of, vendors like SAP, you will hear, you will come across this term, then DSS, which is more conventional decision support systems, then you also hear this terms business intelligence, data science, and so on and so forth.

So, it is quite natural that someone who reads about these terms in literature would get confused as to what are they and where do you place them. Too many words, which may be related, but what is an architecture? Where each of these terms fit in? And what do they do? What are their functions? So, that becomes important. So, very soon we are going to work on a case analysis, a short case analysis, which would give you a sense of what is an analytics, what is analytics, what is data and how analytic solutions evolve. So, it gives you a sense of the process involved in analytics. And it also would help you understand these different terms.

Where do they fit in? For example, I said there is one category of the vocabulary up till big data, data engineering, which is related to data, which is related to data and data storage and data pre-processing. I said data, data capture, data storage and data pre-processing, data warehousing, data marts, databases etc are management techniques for storing data in a structured way. And those are actually data storage and data management technologies. ETL is a data pre-processing software. It is another technology.

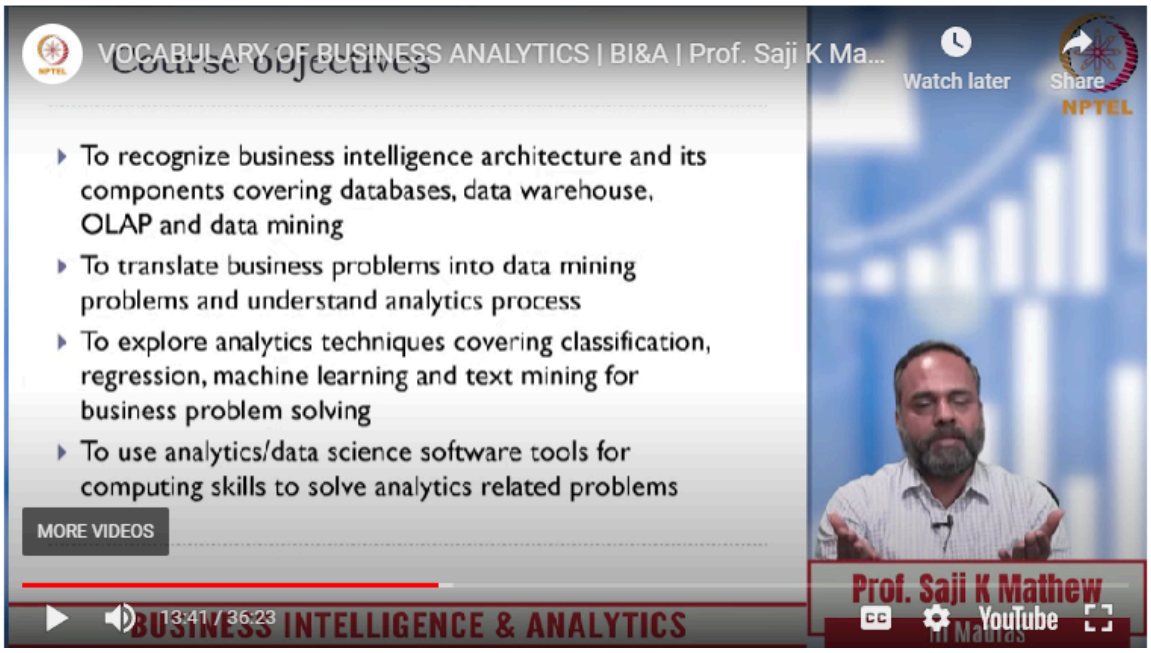
Big data, data engineering have recently been coined in relation to large volumes of data, particularly pumped from social media type of sources. But all this up till data engineering is about data, is about data. But and then you also would hear about technologies that process data, that analyze data, that model data. So, these are software, which are technologies that are used for processing data right from data mining, KDD, OLAP analytics, advanced analytics, DSS, BI, data science are all about analyzing large volumes of data to bring insights, to provide certain insights, to create certain outputs that is useful for decision makers or end users. So, broadly in analytics, I am using the term analytics to explain analytics, which is not good.

But I would say there are two types of technologies in the infrastructure that is required for analytics, that is data storage and data preparation technologies, and data analysis and presentation technologies. So, there are two categories, that is what this slide illustrates. And as I said, the analytics has a dark side. And the dark side is that in analytics, data is used and oftentimes, where do this data come from? This data come from individuals or users or people like you and me. And it is your data that is often used to profile you and your profile is something that gets shared with potential marketers or advertisers or merchants.

And therefore, you get scared, my buying patterns is known to someone. So, that is the privacy or invasion into privacy. So, there is a bright side where you get recommendations that are useful to you. There is a dark side where your data get captured systematically and get analyzed and you get profiled. But it is that you can see this is a trade off; to get something, you have to lose something.

So, there is dark side, which is concerned for privacy, which is another aspect of important attention, particularly academic and research, thanks to the growth of digital technologies. So, in summary, this course is about business intelligence and analytics, closely working with business problems, business problems that could be addressed and solved using technology, data storage, data capture, data storage, data preparation, and data analysis techniques that are available in analytics. So, that is the broad philosophy of this course. The underlying philosophy, the most important principle that is used in this course is that data driven decisions are better than non data driven decisions. They are better and there is evidence to show that data driven decisions lead to better decision quality, which leads to better business performance.

And I will show this in a subsequent session as to what is the basis for business value of business intelligence or business value of analytics. That is a separate stream of research, I will show you some insights from that research. So, in essence, this course has four objectives and let me explain those objectives to you in the session. The first objective of this course is to recognize business intelligence, architecture and its components, covering databases, data warehouse, OLAP and data mining. So, as I said in the previous slide, there are various technologies, not just one that is required to generate useful insights for business decisions.



And it involves a gamut of technologies from data storage to data analysis and presentation or visualization. And how are these technologies organized or orchestrated? And that is the architecture. So, what is the architecture or BI and analytics architecture to sort of appreciate and visualize the constituents of hardware and software? That is the effort and that is one objective of this course. Or in other words, the participants will familiarize themselves with a BI and analytics technology architecture. Maybe that is one aspect where this course stands different from a pure analytics course, which will start with some sort of algorithms and how to use the algorithms, which is very important.

And this course also dwells into algorithms. But it takes a holistic picture that you need technology, which is one of the pillars for analytics. I would say there are three pillars for analytics. One is technology, which covers databases, queries, algorithms and visualization tools. The other important pillar is the basic sciences, I would say, it is data analysis is to be done scientifically.

Sometimes it requires theories to analyze data. So, where do this scientific basis for data analytics come from? It comes from statistics, it comes from computer science, it could also come from psychology.If you are actually analyzing behavioral data, one should be aware of theories in psychology to sort of correlate or, you know, hypothesize, causation, etc. You need theoretical knowledge. So, these are sciences that inform analytics. So, one pillar is technology, second is science and third is the domain.

So, I call it the domain. So, if data, if analytics is here, if I see it as a platform, is as this is analytics, you are standing here, you are all either users or consumers of analytics, but

it stands on three pillars. It, its foundation is on three pillars, I call it technology. And I, the next is science and the third is domain. And in science, I said psychology, I said statistics, I said CS, etc. Technology could be about databases, software, and it could also be algorithms, etc.

All that is about technology or visualization tools and so on. So, what is domain here, that is something which I did not explain. Domain stands for the business domain. Every problem comes from some context, some business context or some context. It could be healthcare, it could be manufacturing, it could be retail, it could be aerospace.

There are diverse domains in humanity and particularly in business, there are different business domains, where different areas or domains where business is situated. That is a business domain. So, suppose you are analyzing retail data, that is a particular domain. And therefore, how retail business works is a domain knowledge that is required. We will be discussing a case related to retail banking, retail banking.

And therefore, in order to work on that problem, you need necessary knowledge about how retail banking works. If you do not know how banking works, you cannot address that problem. So, therefore, knowledge of the domain is another important pillar on which analytics stand. And if you are unaware of a domain and if you have good skills in data analysis, and software, etc, one must make an attempt to go and immerse oneself into a domain to understand the processes, the nuances, the decision making process, etc. in that particular domain. Only then you can actually explain a problem, convert problems into analytics problems and propose solutions to solve the problem. So, therefore, that becomes important.

So, you can see the first objective is basically to familiarize you with technology, without technology. And if you are technology averse or technologically challenged, etc. it is very hard for you to be in analytics profession.

You must feel familiar with databases, you must feel familiar with spreadsheets at a very fundamental level. And you should also be familiar with statistics. And that is also a pre-requisite to do analytics. If you do not know basic sciences, for example, you do not know descriptive statistics, or you do not know what is correlation, it becomes very hard.

So, that is considered pre-requisite. But to a large extent, the technology that is required, our understanding about technology for this course is covered in the first objective. The second objective is very important; to translate business problems into data mining problems and understand analytics process. That is the heart of the matter. As to someone or a business manager or somebody from the industry explains a problem with you. That problem is explained in the business language or in the domain language.

And that is a pure business problem. And there is no analytics there, as is visible. So, therefore, it is up to you or it is for you, as an analyst or as a researcher, I believe an analyst is a researcher. So, as a researcher, you try understand the problem in depth and think and we call it thought process. This is called thought process. You think about the problem or think through the problem or there is a thought process in analytics, that is to convert that business problem into an analytics problem or a data mining problem, where you can apply algorithms or statistics or techniques to model that data, to analyze that data, so that certain results from that analysis would suggest what is a potential solution to solve the problem.

So, as an analyst, one needs to know how to convert business problems into analytics problem. So, that is the second aspect. So, how is it done in this course? We use case studies. We use case studies, we meaning, I would show you cases and we would together along with you, I start thinking and help you convert the problems from business domain to analytics domain. The third objective is to explore analytics techniques covering classification, regression, machine learning and text mining for business problem solving.

So, we specifically pick certain algorithms and techniques. So, for example, I use this terms, techniques and algorithms separately, because algorithms come from the computer science community and techniques, I generally refer to statistics. For example, regression is a technique that is developed by the statistical community. So, in a course like analytics, we use techniques from both sciences. And we may combine them sometimes and use them to model data for certain problems.

So, we will be covering classification techniques, like decision tree is a technique that we are going to discuss. We will be also using clustering as an algorithm to, of course, to group data or to cluster data and then use them for insights. We also would be using machine learning technique called neural networks, artificial neural networks in this course. And towards the end, we would be using text mining as another technique to handle non-numeric data, textual data is not numeric data, non-numeric data. So, that is a very interesting aspect, today widely used, especially in big data analytics, how you handle text data for drawing insights.

So, all these are actually techniques that you are going to learn during this course. And the finally, this course also would introduce you to certain programming skills. Programming, I would case software programming. Well, is this a part of learning from this course, I would say yes, but is programming taught in this course, I would say no.

So, it is more like a prerequisite. It is a prerequisite that the learners of this course should either before or during the course, get familiar with tools of programming languages like Python and R. These are programming languages for data science and analytics that I will be using in this course. And you may ask, can I do this course without programming? Of course, you can listen to lectures, but in order for you to credit this course, you will have to learn Python and R. Will Python and R be taught in this course? The answer is no. And then how do you effectively do this course? I will give you learning resources for Python and R, right from how you install the software, how you start using or how you start programming with them from an elementary level to reasonable level of knowledge to use programming to analyze data.

The purpose of programming in this course is data analysis. The purpose of programming in this course is not to build software. The purpose of programming is to use the libraries and functions offered by Python and R community, which are open source community to analyze data. Data analysis is the objective of programming, not building software.

So, therefore, feel comfortable. You will only be calling certain useful functions to analyze data. Your purpose is not to build a software in most of the analysis that you will be using. And therefore, it is, I would say it is much easier. But you need to have some programming fundamentals. For example, how to use a for loop, how to use conditional loops, like if or case functions, etc., which are very fundamentally there in programming classes, which I am sure my young audience would have learned in school classes today. You learn RDBMS in school, you learn Python programming in 11th and 12th today in CBSE. And therefore, you have, as compared to the old generation, you have got the opportunity to learn them in school. So, this is an avenue for you to develop those skills further into analytics and data science. I am using again these terms analytics and data science synonymously, but I will explain what they mean, how they are similar or how they are different in a subsequent session.

Sir, one question, are you going to use any database like Oracle or MySQL? Yes, let me come to what am I going to use in this course. So, in terms of technology, I would be using MySQL, which is a database, essentially to demonstrate to you and to some extent show you how a structured database looks like, what is that structure, table structure of a database and what is something known as SQL, which is structured query language, how do you talk to a database, how do you sort of select and get data records from a database, how do you actually list it, etc. So, you need to develop an appreciation for SQL, which is a very traditional, conventional way of storing data in a structured format. And we will be demonstrating that in the class, reason being, these are widely used, database or particularly relational databases continue to be a major source of data for analytics.

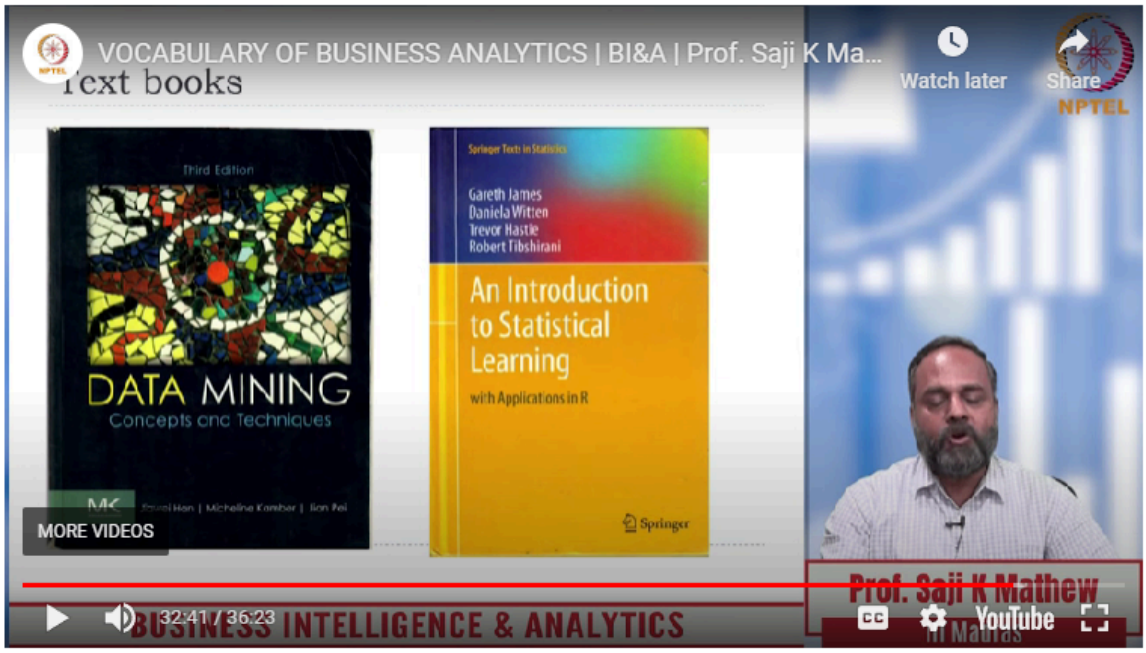
Thanks to ERPs, if business is functioning on ERPs, database, relational database is very a part of ERP systems. And therefore, all organizations which use ERP have relational databases running alongside. And therefore, database is a source of data. So, we will be using MySQL as a database example, SQL queries as a sort of programming language to talk to the database, we will be using Python. So, Python in Jupyter notebook, the Anaconda distribution, that is what I would be using, I would be using R, R programming, I want to believe that many of you are familiar with it.

So, these are software that you can install. And I will give you guidelines for installing and getting started with the software in the next class. And some of you have done this already. And in terms of pedagogy, how the course will be delivered, you could expect lecture sessions, where concepts will be taught in lecture format. And when it comes to say use of databases, use of queries, and also use of algorithms, we will be working on, you know, cases and examples. So, cases and examples, and cases and examples would have a description of the case, case problem.

And as we saw, we will convert that into analytics problem. And there will be some data that will be associated with a case, okay, data will be associated with the case. And we will be analyzing, preparing the data and analyzing the data using analytical techniques, or algorithms. And would this data be available to you? The answer is no, because these are proprietary data. And therefore, this data cannot be accessed in the course, but it will be demonstrated to you. In a class exercise, I will be demonstrating how a set of data can be analyzed using a set of a Python script or an R script.

So, you would be viewing, you will be viewing or seeing how data analysis can be done using programming techniques on a set of data to solve a given case problem. That is how the course will be. So, some of you may be wondering, what are the materials I am going to get in this course. In a course like this, you get access to the video lectures, the slides which are used and often, you know, the transcripts of the lectures, all that will come to you, but the reading resources should be procured separately by the participants. In the next session, I would show you what are the resources that you should try to procure, what are the resources you can access online, etc.

And obviously, you cannot buy data, all the data that I will be using, but you will be, you will be shown how to use data. That is how the course will be and in the subsequent sessions. And being the first class, let me also show you what are the books that will be useful for you in this course. I am just giving you the cover pages of the two books I recommend as textbooks for this course. One is Data Mining Techniques, Concepts and Techniques by Han and Kamper and Pei.



This is a very popular textbook published by LCVS Science. Now, fourth edition is available. I recommend that you procure the fourth edition, I guess the Indian edition is available too or for all the editions, you may find PDFs available online that is up to you to search and find if it is available. But that is one way to procure the resource. And so is the second book, An Introduction to Statistical Learning or ISL in short form, it is called, authored by scholars from Stanford. And this is also a very highly recommended book, a textbook for analytics and data mining.

And I strongly recommend this book, you can order this book, buy a hard copy or if you are lucky, you may find this course, soft copies of this course in some form available online. So, I recommend these two books as textbooks. We will not be providing you textbooks, you have to find ways to get these books. And other readings or other books as complimentary reading, I recommend are these books, Data Mining Techniques by Berry and Linoff. This is a widely used book in the practitioner community, because they provide a lot of examples from business and management, where the application of analytics.

They do not provide details of algorithmic techniques. But they actually provide a lot of insights about business and business problems and analytics. So, the focus is on business problems in this book. And data mining in practice. So, the Data Analysis using SQL and Excel, this book is very useful for descriptive analytics. I told you in descriptive, you only describe data in terms of certain indicators or visualization of data.

SQL is a tool for descriptive analysis. It presents you data, but it does not model data. So, that is a book if you want to learn SQL and Excel in analytics. The third book, which

is having traction in the data science community is Data Science for Business by Provost and Fawcett. This book is, they are also researchers and their research papers are also available. And this book is, again, very practical, and to a reasonable extent, they explain algorithms.

I would be using this book also, or certain parts of this book in my course. These are very useful books. And for those who are very serious in analytics as a career, I want to, I would strongly suggest you be very strong in your foundations. So, statistics is an important foundation for analytics and you would, you can learn statistics from very foundations using Newbold's book on Statistics for Business and Economics, Applied Statistics. The other book I strongly recommend to understand statistics from behavioral aspect. Human behavioral aspect is Kerlinger and Lee's book on Foundation of, Foundations of Behavioral Research.

And these books you have to buy separately or these are reference books. And here is the TA. So, each course is a TA and Subisha R is a PhD student in DoMS and she will be the TA for this course. And she will be constantly interacting you with through the, through the Swayam portal. And so, so this is how the course is going to be delivered to you. I will be giving you lectures and Subisha will be assisting me and also giving you tutorials and some sessions that is very much related to the course.

And thank you very much for listening and participating and look forward to seeing you in the next class. Thank you very much.