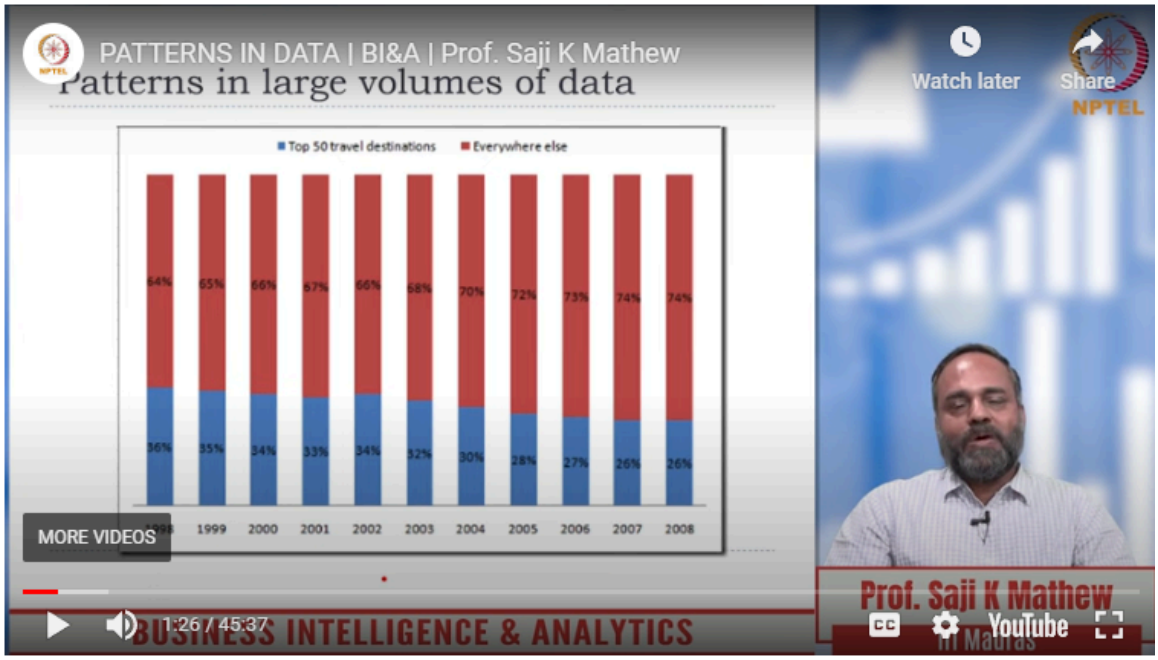
**Course Name:Business Intelligence and Analytics**
**Professor Name:Prof. Saji.K.Mathew**
**Department Name:Department of Management Studies**
**Institute Name: Indian Institute of Technology Madras**
**Week:01**
**Lecture:02**

## PATTERNS IN DATA |BI&A

Now, we come to the next slide, which shows you the travel volumes; air travel volumes analyzed by an analyst quite some time back. This analyst collected data about air travel volumes between 1998 and 2008. So, since both the years are included, this is 11 years of air travel volume data. And so the x axis is about the years and y axis is about the volume of travel. But the volume of travel is not given in absolute numbers, it is given in relative terms, in relative percentages. So, the blue bars represent travel volumes to top 50 travel destinations and the red bars show travel volumes to everywhere else.




So, this graph shows that the travel volumes or travel preferences of people or travelers have been changing over the years, over the 11 years that is depicted in the graph. So, somebody made an attempt to collect this data and visualize this data in the form of bar graphs. So, are there some quick observations from this bar graph in front of you? You can look at it for a while and make your observations. Sir, when we see this graph, it is

clearly showing that the top 50 travel destinations, the number of people who arrive, it is continuously decreasing and the other places are becoming popular.

Those places which are not in the top 50 travel destination. Maybe because of infrastructure development or more airports coming to those places and such things. Okay. Maybe Sir, there is maybe a shift towards the virtual environment like that, I do not know. Okay.

Virtual infrastructure development, like development of virtual tourist place like that, because in 1998 at the Fianna, there was lack of the internet connectivity. Okay. So, it is going like declining phase like that. Okay. So, I am getting two suggestions.

One is, well, there is a increase in travel volumes, relative increase in travel volumes to everywhere else other than the top 50 over the years that is depicted here. And so that could be number one, because of development of infrastructure world over. The other reason that is suggested is because of internet, people or travelers are able to do their online booking and search for, you know, tourist destinations etc. online.

Okay. So, essentially you are suggesting that travelers have more access to information today and that is why the air travel volume pattern is changing. Okay. Alright, good. Sir, through social media and other platforms, people are coming to know about different kind of new places and the ease of travel that is available now.

Okay. So, there are a lot of social media platforms which inform. So, there are a lot of online platforms meant for travel which are accessible today as compared to, you know, decades back. Okay. So, yeah, more access to data and information. So, people have started traveling to destinations other than the so called top 50 like the Mumbai, Delhi, Munich, New York, or you know, and top 50 cities of the world, but people are traveling to know the so called non-popular locations.

Okay. And the reasons are what you are actually suggesting. Okay. Well, there is a problem here. One, what is displayed here is information. Raw data should have been in databases, databases of travel agents, databases of airline companies who have data about how many people are traveling from where to where.

This data must have been available in the databases and somebody has sort of got this data and analyze the data and visualize the data. Okay. So, what is shown in the graph strictly speaking, is information of relative air travel volumes from 1998 to 2008. It is a valid observation from the data that travel volumes to top 50 destinations have been declining.

Okay. Or to be very specific, it has declined from 36% to 26% or 10% decrease over the last 11 years, both years included. Okay. So, that is a very valid observation, that is a pessimistic observation. Of course, you can say that differently by saying that travel volumes to destination other than the top 50 have been growing.

Okay. And it has grown by 10%. Okay. So, everywhere else travel has grown by 10%, top 50 travel has declined by 10%. Both are valid statements. And of course, we cannot say that if absolute volumes of travel has increased or not; from this data we cannot.

We can only say that relatively travel volumes have changed. That is also something that we can keep in mind when we interpret this graph. But what is not given in the graph is something that you are trying to suggest. You are suggesting the travel volumes have changed because, so you are actually bringing a 'because'here. Why or this is actually 'why' you are trying to explain.

You are, you have gotten into explaining why there is a decline or why there is an increase in the travel volumes. And is that data available in the graph that is shown? Does any data here suggest that travel volumes have increased because of internet or infrastructure or anything of that kind? No, the answer is no. There is no data or information about infrastructure improvement, be it internet, be it physical infrastructure. There is nothing, there is no such data that this analyst has collected or analyzed. So, what makes you actually suggest this from this data? Strictly speaking, one should not, one should keep mum.

Let the data speak. That is the principle in data analysis and data interpretation. You should only talk about based on or you should interpret based on the data that is available. And data about infrastructure is not available, therefore, you should keep mum. But we try to explain. And that is our nature, that is our human nature.

We want to have an explanation. Whenever we look at a phenomenon, we have an observation. And what we observe, we can describe. You can describe what we observe. And this graph, this graph is an excellent example of a good descriptive information.

And we also call it descriptive analytics. In terms of analytics, we can say, it only describes data. It does not say why the data is like this, or why a phenomenon of this kind is happening. This only say, what, what is happening? A descriptive analytics or a descriptive analysis of data is answering the question, what is going on? What is happening? Tell me what is it? So, you describe the data, you do not add any masala to the data by saying, oh, somebody is doing something like this and why that is why it is

happening.

No, absolutely no. In a descriptive analytics, do justice to the data, do not say anything else. But there is explanatory analytics, explanatory or you try to explain, explaining or can be called explanatory analytics. In explanatory analytics, you try to explain why some phenomena is happening. So, here in this case, when you look at this data, you have a natural curiosity.

The 'why' is a curiosity. In research, as a instructor of research methodology as well, I encourage students to be curious about the phenomena they observe. I did not mean to suppress your curiosity when you look at this data and try to give an explanation. You need to give explanation. So, explanation seeking research is known as explanatory research. And that is an integral part of analytics also.

If there is no curiosity to explain things, so then why do we exist? And why do we get into analytics, etc. So it is that curiosity and therefore, it is a valid curiosity when you look at this graph to ask, why this pattern, what explain this decline of air travel volumes to top 50 travel destinations or what explains why is this happening? So, let me remove that what question, let me ask, why is travel volumes to everywhere else increasing over the last 11 years, over the 11 years from 1998 to 2008? That is a very accurate and that is a very relevant and curious question. Imagine if this graph is presented to practicing managers who are decision makers, say in the airline industry, say in infrastructure industry or say in travel industry, they will be very curious when they look at this data and they would immediately try to answer like what you try to suggest, like it is because of, it is because of. A why question is answered with the word 'because'. So therefore, we have a situation here.

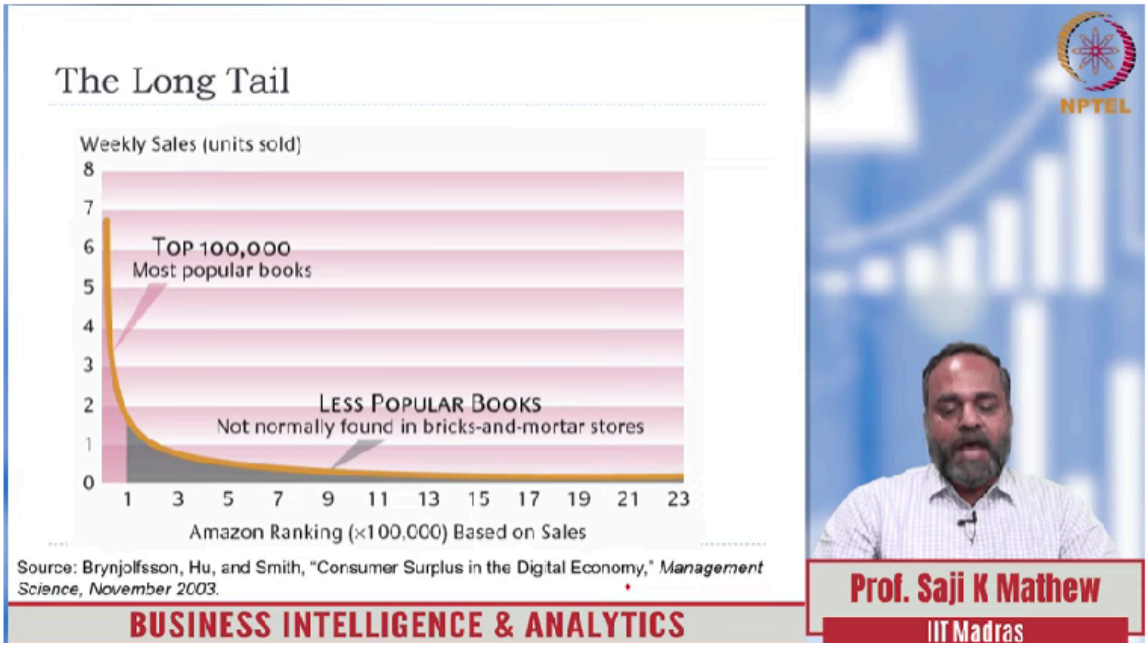
There is only descriptive data. There is only, as far as data is concerned, there is only descriptive data available. But as far as questions are concerned, we have further questions that is generated from this descriptive analytics, which is the 'why' question, why is it happening? But what we immediately did is, we try to give an explanation as if that is conclusive, as if what is suggested, like it is because of internet infrastructure, it is because of airports and so on, may be true. I do not deny that, but there is no data here. So therefore, we cannot conclude, we cannot conclusively state, it is because of this, we can, what we call in research literature as we can conjecture or we can hypothesize. You can actually say that it appears, or it could be, or a reason could be.

So, you argue, a hypothesis is something that you argue, you do not conclude, you have to go and collect more data and see if there is a correlation between, increase in infrastructure and this travel volume. So, you have to establish correlation first and then

causation would require further more analysis. So, correlation and causation are not the same, we will see as we go in our analytics lessons in the future classes. So in sum, a descriptive analysis is only describing data, it answers 'what' question, a what question like, what is the pattern of air travel volumes between top 50 destinations and other destinations for the period 1998 to 2008? That question is answered in this graph. But the question why is travel volumes increasing to destinations other than the top 50 is not answered in this graph, it is not answered.

But yeah, we can speculate like that since virtual reality is becoming popular these days. Yeah, that speculation is called a conjecture, called a hypothesis. And so therefore, we say that a descriptive analysis is very useful for generating research questions or for generating hypothesis. It is very useful for asking useful further questions. And those questions would be very useful for further data collection and analysis, which may lead to further more useful conclusive results useful for decision making, etc.

So in analytics, we say a descriptive analysis is the starting point, which generates insights, which generates insights, or it generates certain patterns, it brings certain patterns, certain order, certain structure in the data. When you look at this, here the pattern is, that there is an increase in travel volume for the period 1998 to 2008, that is a 10 percent increase. And therefore, that is a pattern that you observe from the data. That is a pattern that you observe. And you try, you further try to explain it leads to further analytics.



So a descriptive analysis actually helpful, it is helpful in generating useful hypothesis for further studies. Let us move on. And as this graph shows, this particular graph or data is taken from a book written by Chris Anderson. The title of the book is The Long Tail. And this is known as the long tail phenomena.

This is long tail phenomena. It is an old book and those who are interested can read this book, I read it long back. So the long tail is a phenomena, which is observed not only in this graph, but in online business in general. Let me actually bring you to that concept at the end of this graphical analysis of data. The long tail is a specific phenomena, wherein researchers and analysts observed from online sales data, that online sales follow a different pattern from the 80-20 principle or the Pareto principle of sales, which generally says, states that 80 percent of sales come from 20 percent of products. 80 percent of sales comes from 20 percent of products, that is the Pareto principle, but online sales, as observed in this graph, does not follow the 80-20 principle, but it follows a long tail principle.

What is the long tail? So in internet, say internet business, or in internet e-commerce, like what an Amazon does, the number of products, you can see the product volumes, the variety of products, the variety online are too large. And you must be observing this as a regular phenomenon. Because if you want your music, or if you want your favorite book, if you go to a regular book store, or a music store, today nobody goes to music store, you get it all online. But if you go to a bookstore, for example, you may or may not find the book that you want. Suppose your interest is in books related to Indian history, maybe in the Gupta Empire or, you know, or in the ancient times, ancient Indian history is your interest.

How many books you will find in a bookstore, in a general bookstore, chances are that you do not find any book related to ancient history at all. And that is a niche interest that you have. But you know, that is not others interest. So a bookstore would stock books that is based on the local demand, local popular demand. So how many books are stocked typically in a bookstore? They say depending on the square feet area 50,000 to 1 lakh or 100,000 is what a bookstore can typically store, that is the space that is available.

But how many books are in print? So this research shows when they did the research, of course, it is dated. So more than 3 million books. So you can now imagine of the 3 million books, if 1 lakh books are available in a local store, that is a very, very small fraction of the number of books that have been actually printed. And but that is the limitation of the so called brick and mortar business. But the non-ecommerce is known as brick and mortar or physical stores.

But when you go to a virtual store and an online store, when you search for the favorite book that you want, chances are that Amazon or a Flipkart will display the book that you search for. And they will make that book available to you from any part of the world. And if the book is not printed and readily available on stock, they can even print based on on demand, what is known as on demand printing. On demand printing is today possible, so they can print it for you and then ship it for you. So that has actually increased the tail or the number of products that are available for sale.

And the observation from this particular online sales data is that the large range of products that are very niche, the niche products, the large, the long range of niche products actually constitute more than 20 percent. So they actually, they say it is 60-40, the ratio is 60-40. So suggesting that the top sellers or the blockbusters may be contributing 60 percent of the sale that is fine, but the niche products do not contribute just 20 percent, but they contribute more than 20 percent up to 40 percent. So the large niche base of products contribute much more than 20 percent. That is known as the long tail, the long tail of variety of products that are available online.

And you can see there is a change on the demand side and the supply side that has happened thanks to internet, that you can search for products that you want, the niche products that you want. So the ability to search and find online, that is a demand side change thanks to internet. And on the supply side you can know, you know that the on demand printing and ability to ship materials or products from different locations of where the inventory is available. All that changes on the supply side, all that has led to a new phenomenon known as the long tail. One more thing sir, when we are suppose we are going to Amazon and searching for a popular book, the Amazon may recommend us some non popular books also.

So we may come to know about those and we may end up buying based on our liking. These facilities are not available in the recent model. Yeah, recommender systems and search. So these are all that has actually contributed to the long tail phenomenon thanks to internet. And therefore, you can see that this, the sales as a phenomena is, has undergone change from the conventional 80-20 to a different ratio, a different phenomenon.

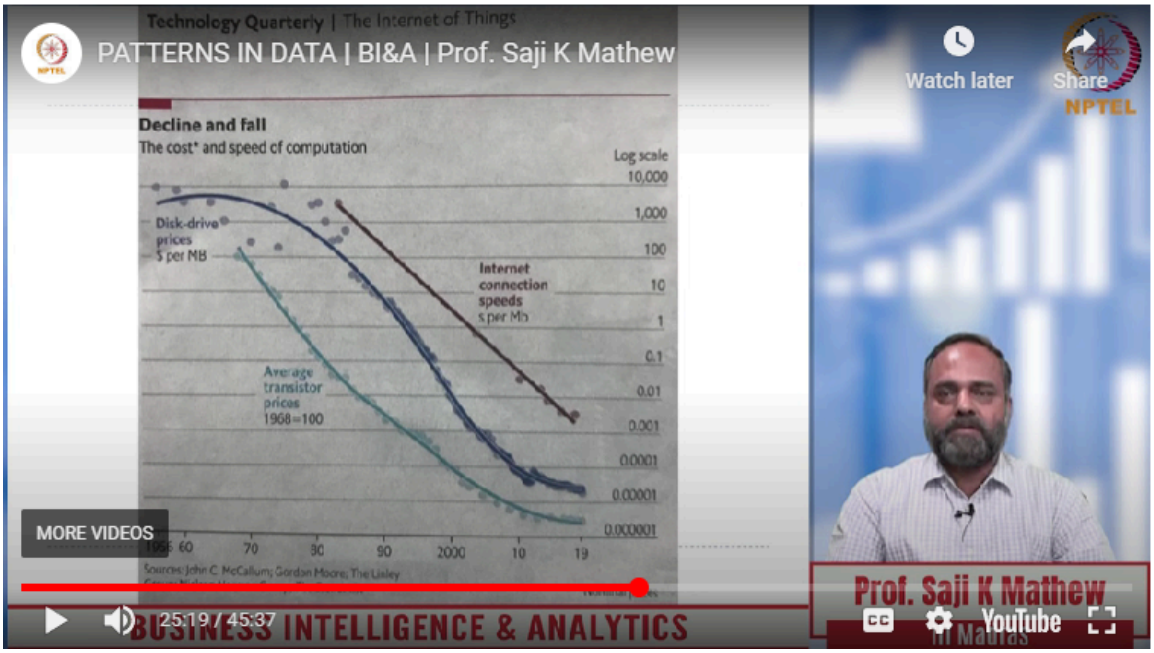
And the important thing to notice here is that this change may have happened, but we came to know or we come to know about this change thanks to analytics or thanks to data. Thanks to data that is captured in databases. And when you do the analysis of this data, you get new insight. This is called insight.

The long tail is an insight that comes from data. We saw long tail in another format.

This is nothing but long tail. That is everywhere else is the long tail. The non popular locations, people are traveling. So therefore, the travel volumes and of course the revenues, from the non top 50 is contributing largely to revenues or volumes.

And that is also a long tail phenomena you can observe in travel. So this phenomena or these insights were derived from databases or from data. Or thanks to data, today we are able to generate new insights, which otherwise are hidden inside the large volumes of raw data. Thanks to analytics, thanks to databases, this became visible. So I am just sharing some data that is captured in the research paper.

I just showed you this particular insight is available in the research paper of Brynjolfsson,Hu and Smith published in Management Science in 2003. And the title of the paper is From Niches to Riches, the Anatomy of the Long Tail. This slide is again from a research done by external agency. I have just copied it from the economist technology review.




And I found it very useful. It is not, it is more recent. And this consists of three graphs. So you can see x axis provides timelines in years and y axis is in log scale, in logarithmic scale. And it is plotting the prices. So price and speed, so for, so you know, it is a bit confusing because for transistor prices and disk drive prices, which are the two bottom graphs, it is in currency units.

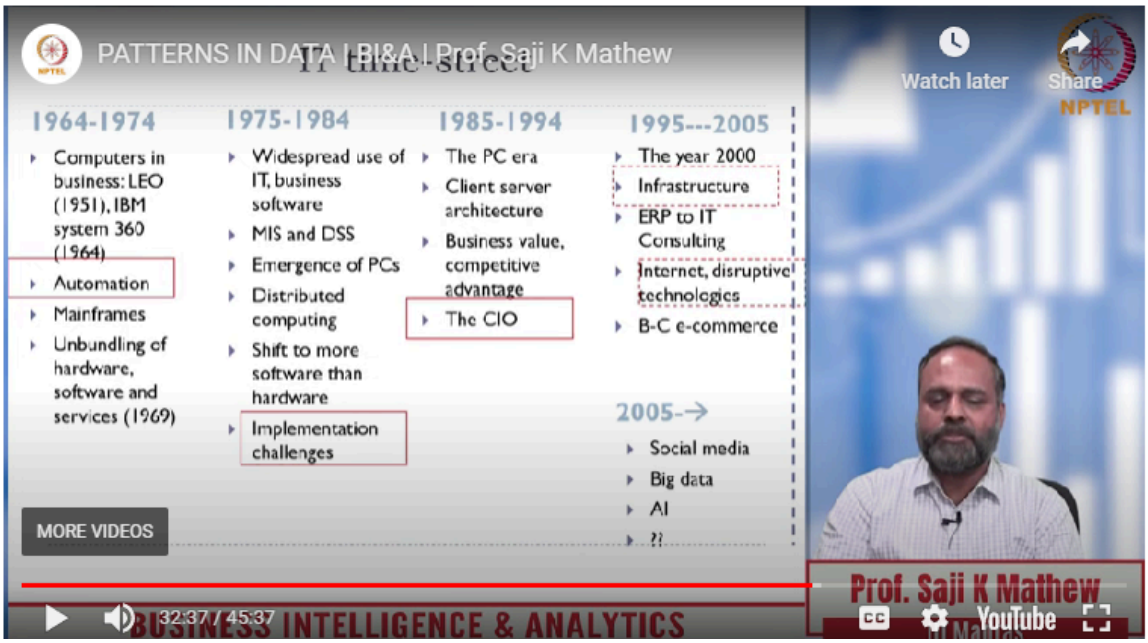
So it is in dollar per MB for disk drive prices. And it is in dollars as far as transistor

prices are concerned, average transistor prices. And you can also see for internet speed, it is dollar per MB, megabytes, megabits, it must be megabits. So that is about speed. So essentially giving us a sense of cost of what? Cost of transistor, cost of disk drive and cost of internet speed. And what are these three graphs about? These are very interesting three dimensions, I would say of infrastructure.

I would say of infrastructure. Why is, I would call this as infrastructure? Because infrastructure is a basis for delivering certain service. Infrastructure is the underlying base, base through which a service is delivered. So look at railroads. Railroads is an infrastructure, but transport is a service. In a similar way, when you, if you look at, if you look at services on internet, you can treat analytics as an example, or data science and analytics, or analytics based solutions as a service.

But in order to run those service, you need infrastructure. What are the infrastructure that you need? You need processing power. You need capacity to process data. And what is the infrastructure that actually supports processing of data? The basic infrastructure is the hardware and software. So the average transistor price is related to the processing power, processing power that is required to process data.

Disk drive, disk drive is another infrastructure for storing data. If transistor is for data processing, disk drive is for data storage. And third dimension is the internet speed. So you, we talked about processing power, we talked about storage power, and third is data transmission. The infrastructure for data transmission is represented by internet connection speed. So three major shifts, three major changes that has happened in the recent times, particularly post internet, in processing power, storage and transmission.




You can see the drastic downward slopes of these three curves. They have been falling, the prices, the cost of storage, the cost of processing and the cost of transmission have been falling over the last few decades. And you can see from in a log scale, if you look at 1960s, it was 1000 and it has become in the range of 0.001 to 0.00001 in the 2020s or 2019. So this is major change.

So what is making analytics possible today or how there is an analytics talk today? What is driving the big data talk today or the big data analytics today? Or what is making storage, transmission and analysis of large volumes of data, data collected and stored in databases, data also collected and stored in external sources like the social media. How was the analysis of that data possible today? What is driving it? This graph actually explains it. This graph very well explained that the cost of doing analytics using large volumes of data have substantially fallen. And that explains why the analytics industry is growing. If it is extremely costly to store data, audio, video and textual data, then it is not possible to do textual analytics, it is not possible to do video analytics, etc.

But that is possible today, thanks to the drop in cost. Excellent analysis by analysts and which provides us a reason for this fall or this rise of a phenomenon known as business intelligence and analytics. Let me also give you a picture of how information systems and its recent history suggest that analytics and data science have really matured and what explains it through, the through these timelines. This is my own depiction of major events in the history of Information Systems in the past decades. If you look at Information Systems , they gained importance in the industry, particularly in manufacturing and production from 1960s onwards.

Of course, computers became important in industry with industrial revolution. When production processes became automated, it was also important that the need for computation be met by computers. So computers were introduced in industry to support production processes because large scale data analysis, particularly computation was required for process automation to be carried out. So therefore,you know a company called IBM was founded in the early 1900s to support industrial revolution and automation. And the earlier of computation was about mainframes and legacy systems and so on, where a lot of data from business transactions were produced in the form of files, flat files. And flat files carried data about business transaction, but there was no concept of database or a relational database and that comes much later in the 80s.

So when you look at the history of computing in the industry between 1975 and 1984, you see along with automation systems, along with a, the so called, the ERP systems, there was, so ERPs came a bit late, there was material resource planning like the MRP and manufacturing resource planning. These were business software used by production, for production between 1975 to 84. So all this resulted in a lot of data.

And that data were analyzed to produce reports in the form of MIS reports. So it is known as management information system reports. And DSS was another byproduct of data, decision support systems, which did advanced analysis of data from files and also from databases. By mid 80s databases became very popular. Relational databases were available in industry along with ERP systems, distributed computing began. Personal computers became popular.

Microsoft and Apple and Oracle were all founded during this period. And you can see software becoming very important industry to support other business or economic activities. So mid 80s could be seen as the era of software industry and database industry, I would say. And when you move from 80s to 90s, you can see business intelligence, business intelligence as a particular concept started developing in addition to DSS. And what is the importance of it? Importance of it is that databases grew as a technology to capture data and databases grew in terms of volumes of data they captured from business.

So data in digital format became available thanks to MRPs and ERPs. What are MRPs and ERPs? These are large scale software that were used by industry for process automation, for business process automation. And we know that most businesses today would run on an ERP because it brings efficiency to business. It automates business processes. So one contribution of information systems to business is in terms of efficiency, in terms of process automation, automation systems. So you go to a bank,

every business process in a bank is automated using ERP, be it opening of an account, be it withdrawal of a money, be it deposit of a money, be it closing of an account, be it starting a FTP or some other form of investment.

All this is actually done through software in a very efficient way. And of course, today you have thanks to internet you can do it all based at home. So it also bring convenience to users. So efficiency, convenience, these are all key contributions of information systems to business. But then there is another important value creating activity, which is the intelligence part of it.

Business requires to make decisions and decision support is based on information. So the information required for business decisions comes from insights from the data. So therefore, as you can see from 70s to 80s and 90s, the particular focus of information systems on data and analyzing large volumes of data, creating information and presenting that in a way, in a format that is useful for decision makers. Right from descriptive analytics to explanatory analytics, through predictive analytics, in terms of predicting what would be the next point of data, based on historic data. So all that actually became part of the business activity.

So business intelligence and analytics is a new phenomena, thanks to the growth of databases. That is what I want to stress here. Thanks to the growth of databases which captures data and store in databases. And you can also see I have defined a new era starting from 1995. 1995 Netscape Navigator or the starting of internet, opening of internet for public use.

So 95 onwards, you can see that as a new phenomena in information systems. In the beginning of electronic commerce, lfor in the B2C format, in the business to consumer format, all began around 1995 and it never stopped, you know that. So internet commerce not only captures business transaction data, it also captures user behavior online. So for example, in e-commerce, a company like Amazon knows what you have, what you have purchased from Amazon. What are the items you purchased and what is the date of purchase and what was the value of purchase.

All that is available just like any other company which is selling products to customers. The database captures all that you have purchased and the transaction data, you call it as transaction data. But internet commerce in addition to transaction data is able to capture lot more data about consumers and their behavior, consumers and their online behavior. For example, Jeff Bezos, the founder of Amazon says, we do not throw away any data, meaning that they have, they own, as soon as you sign into your account and start using Amazon, they of course use cookies or they deposit a small piece of soft, it is not a

software, a method to capture data about your browsing history, as to which are the pages you visit, how much time you spent on each page and what are the other nature of your activities in terms of pages and time is captured well using Amazon cookies and that data is stored in your machine itself, when you sign into your machine. So therefore, these cookies are very useful sources of data to capture your behavior, as to what products interest you, what is your browsing behavior, what products you actually place in a shopping cart and what products you drop from shopping cart, a lot of details of your behavior on the internet is captured to profile you and as we discussed some time back, to recommend products to you. Amazon's recommendation systems have contributed hugely to Amazon's business value and thanks to the relevance of recommendations to users, users find it useful.

For example, to suggest that customers who bought product A also bought product B, sometimes you find it very useful because they are complementary products. That is an information that Amazon generates from its historic data, through analytics. So therefore, what you can see here is that with internet, the extent of analytics expanded, the scope of analytics expanded from database data to online behavioral data. So you see a change or an expansion of the scope of analytics and I have placed 2005 as a subsequent era, where social media, big data, AI and what more, robotics and so on are actually the new topics of interest in the world of internet. Web 2.0, interactive platforms, where you see, you look at growth of companies like Facebook in, I guess around 2004 and subsequently Twitter, which is today X and then we also talked about platforms like WhatsApp, then and several other social media platforms, which attracts and engages people, not for business.

You just come to those platforms to spend your time, to engage with platforms, oftentimes to engage with others. It is a society, it is a virtual society they have formed, where people exchange their interest. And that is a data, all that actually the platforms do is they keep registering your data, what you are talking about, what are you posting? And that talks about you, that is a data that is useful to profile you, to market products to you. So, you may not be really serious about your privacy oftentimes, because you get benefits by participating in the social media.

The privacy is a dark side of customer analytics, where your data is captured and stored and transmitted, shared, etc. and used for profiling etc. oftentimes with your consent, but you do not realize that you have given your consent. That is the dark side of the analytics, particularly with social media, but that is of enormous value in business. So, the discussion about analytics has again subsequently changed from e-commerce to social media. So, this IT time sheet provides you a picture of the growth of use of data for insights in business, from the early era of 1960s to the current era of social media from 2005 onwards, where the scope of data analysis has moved from MIS, DSS to business

intelligence, analytics, big data and data science today.

So, that is a growth story and that is a non-stoppable growth story. And today with artificial intelligence and machine learning, we are creating algorithms which are machines, but manlike. So, manlike behavior or manlike characteristics are induced into machines or algorithms where algorithms learn or they get trained over a period of time with data just like children actually learn and become more intelligent, algorithms become more intelligent with more data or with large volumes of data over a period of time and they become intelligent, that is artificial intelligence. Artificial intelligence derived from data, the intelligence comes from training data, we call it training data. So, these are aspects we will touch upon in this course, but this course is not about ML or AI or even big data analysis, we scope this course in and around business intelligence and analytics.