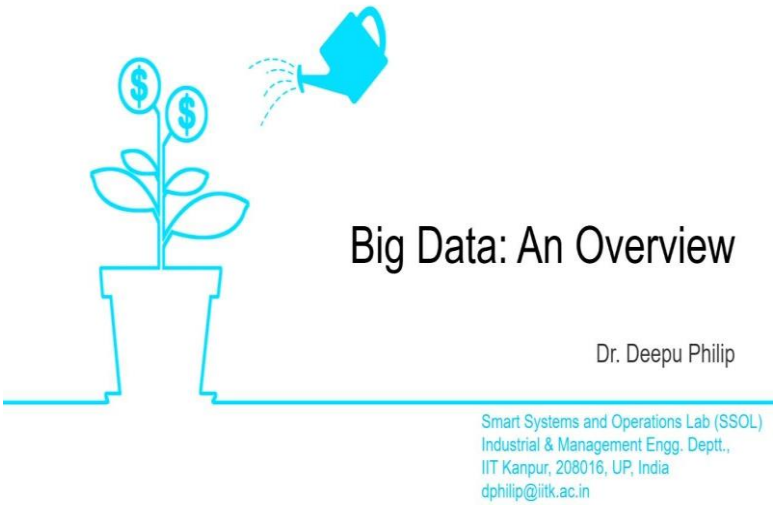
**Computer Aided Decision System**
**Industrial Practices using Big Analytics**
**Professor Deepu Philip**
**Department of Industrial & Management Engineering**
**Indian Institute of Technology, Kanpur**
**Professor Amandeep Singh**
**Lecture 17**
**Big Data: An Overview**

Good evening, everyone. Welcome to yet another lecture of the Web-based Decision Support Systems course for which is meant for both academicians, practitioners alike. And, we have been discussing various aspects of the Decision Support Systems. And we have been delving deeply into aspects related to Database, specifically Database Management Systems. And today we are going to get into a new advanced topic of Decision Support Systems, especially with the current trend on people, everybody jumping, thumbing on the Big Data.

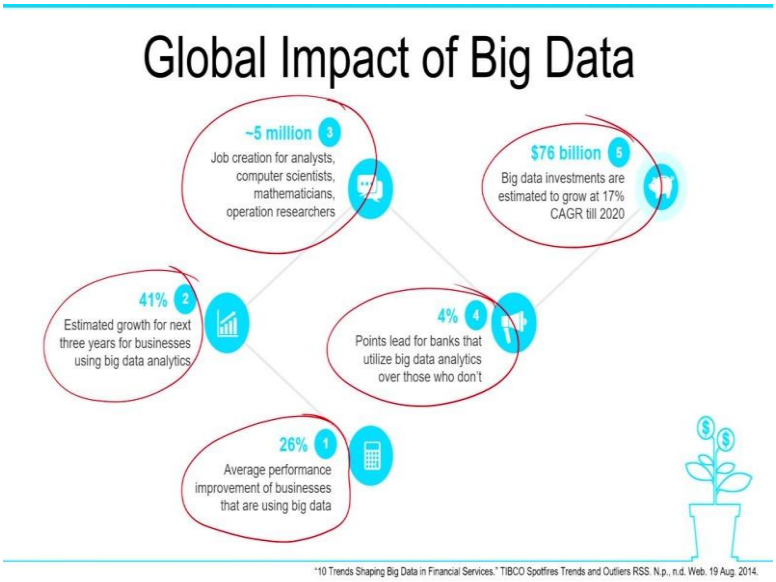
And today, what we will look into is give you a brief introduction about Big Data and its pros and cons at the same time. And, it is not that everything should be the most because I do not think the life continues, whether Big Data is there or not. So, it is something that you need to understand. And, then you are to make as a decision maker, you need to decide whether I need to use Big Data, big noise, whatever you want to call it, and then go from there.

(Refer Slide Time: 1:18)




So, today what this is just An Overview, so, the slides if you look into it, So, it is a Big Data Overview and I am Doctor Deepu Philip with the popular from IME Department, IIT Kanpur. So, without wasting much time, let us get into what is the some of the examples.

(Refer Slide Time: 1:33)

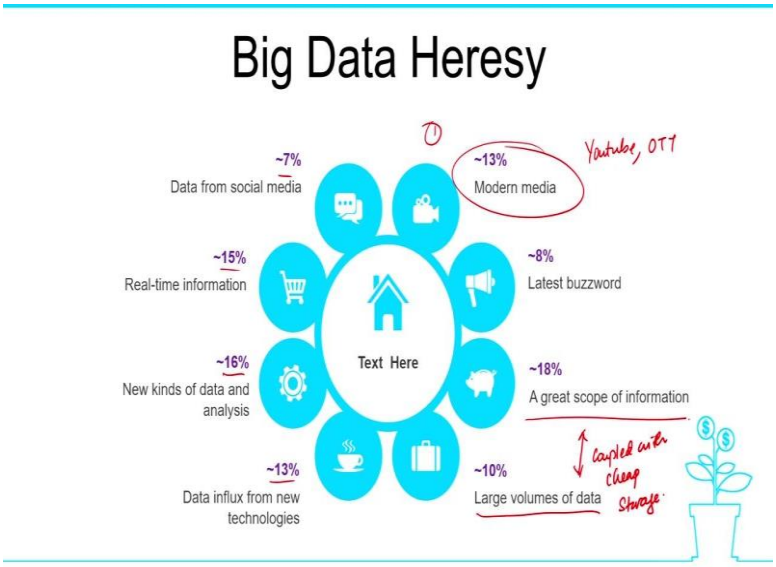



So, these are some of the grantees given right here. The major aspect here is, let us go from the first point.

1) There are various studies that are said that there is about 26 percent average performance improvement of businesses, it is not true for every business.
2) But on average, whoever has implemented using Big Data, they have seen something close to 26 percent improvement in the average performance improvement.
3) And, then the businesses, that are focusing on Big Data Analytics, it is about also 41 percent of the estimated growth in the next three years. That is what the current estimate after COVID that has come out in this picture. And an estimated of 5 million jobs. The third point is 5 million jobs will be to be created in the form of Analysts, Computer Scientist, Mathematicians, Operation Researchers, etc. In the field of Big Data, especially. And, the banks that use Big Data Analytics to identify fraud detection, and money laundering, all these kinds of things, especially in the financial sector, assessing risk, both operational risks, then also market risk, etc. They are supposed to get a point benefit compared to the banks that do not do it. So, a lot of the time, because of such kind of benefits. And also, associated insurance cost reductions, banks tend to do this a lot.
4) And, from 2020, the estimated investment was something like 76 billion in Big Data, and it is supposed to grow 17 percent cumulative annual growth rate from 2020 onwards, this number was before COVID. And now after COVID, these numbers are going to come out soon. But before COVID, this was supposed to be one of the most

bubbling growing field at one point of time. So, this is the big, big, big aspect of Big Data. So, search and find a nice pictures of Big Data.

(Refer Slide Time: 3:28)




So, now let us take a look into this what people think about Big Data. So, let us start from right side. We start from right and move to left side.

1) 13 percent of the people literally think that. This is modern media. Modern media that includes people says that, YouTube, then you have these OTT streams, etc. So, they are saying that the modern media is what we they called us Big Data, So, much of videos on the internet, So, much of audio, these kinds of things, So, that creates what they call us.
2) 8 percent of the people believe that is the latest buzzword. It is just another buzzword created by some budget management people or something like that.
3) 18 percent of the people believe that is a great scope of information, there is an opportunity for a great scope of information may not be true, but we will talk about this.
4) 10 percent believes large volumes of Data. So, to a large extent, these two things. are coupled to you should understand that coupled with cheap storage.

So, this is an example I tell all my students when I was growing up, we used what we call us BBC Micro computers with 256 kilobyte floppy disk in it. And let us say if I had to store pictures of someone, let us say my favorite actress in that, I could store maybe two to three pictures. So, if I take my floppy drive particular computer and take a look, we can say, there is

three pictures of this particular actress or two pictures of this favorite actress, so, his favorite actress is without any doubt this particular person.

Now, the floppy disks from 256 KB went to 512 kilobytes. So, instead of the 2 to 3 pictures, I could store 6 to 8 pictures. So, still, it was not very difficult because you could still sell the rate pictures of 8 pictures 6 of them belong to one particular actress and to belong to another one. So, he is still his favorite actress is this one based on the number of images he has stored. Now, go away floppy disks, come to ZIP disk. So, the kilobytes became megabytes.

So, 100 megabytes, 250 megabytes, So, pretty soon I am not storing 5 or 6 pictures and storing 50, 80, 100 pictures. So, now you require an Excel sheet. There are 100 pictures or 80 pictures of it. Of the 80 pictures, 40 belong to this particular person, 15 belongs to this, 8 belongs to this, 11 belongs to this etc. And say among all the 100 pictures, 60 percent of the pictures belong to this one particular actress. So, hence, there is a very high chance that, this person prefers or likes this particular actress and followed by this and that person.

Now, before when you had a smaller data set, that you could say with more definiteness and clarity, now some murkiness come into picture. Now, go away floppy disk, come back pen drives. Now we are talking about megabytes goes away to gigabytes and terabytes. So, instead of pictures of 100, or 200, now we are looking at 1000s of pictures, 2000, 5000, 10,000, million pictures, whatever it is, now you are stuck. You are like, who is his favorite actor, there is 20 percent this lady, 18 percent this lady.

So, when storage became cheaper, what I ended up doing was, I have only 256 kilobytes of floppy storage space. So, let me decide what to store. I have 2 gigabytes of storage. I will store whatever the heck I will want to store. So, in that process, what has happened is, I also started restoring all unwanted things, maybe I might have preferred Sri Devi or Madhuri or something like that. But now in this process, I have added all 250 different actresses from Poland, some from US, some from Japan, all we are located.

So, now we are like, who is the favorite actors? How do I figure this out? So, that adding of noise, in my opinion, has complicated the decision-making process. Now, the clarity or definiteness is not there. So, when you store, you also need to be very clear. That is what I was telling, emphasizing the previous thing also. You need to make sure that what you are going to store and what is important for you, not just store everything for the sake of storage.

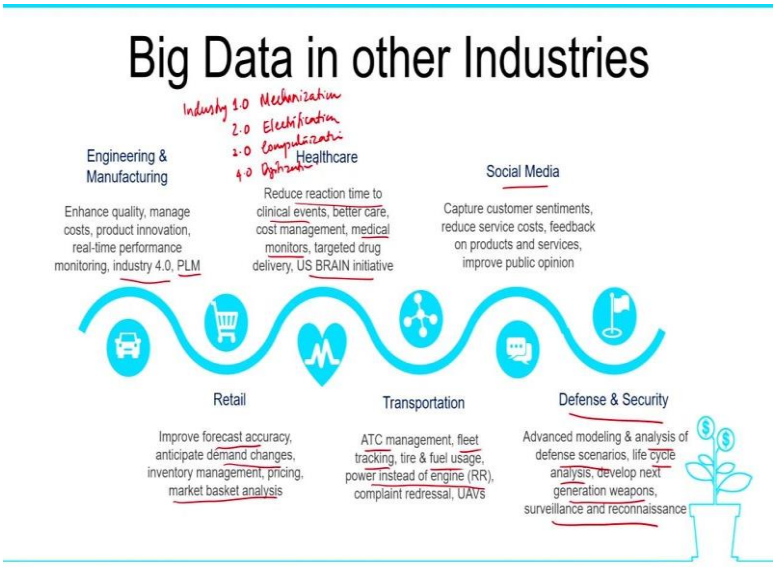
So, that is I wanted to always say this to people purely because you want to see the other side of it. It is not all rosy pictures as much. I am the perpetual pessimist in that regard.

5) Then comes the 13 percent of the people, who believe that this Data influx from new technology, that is very true. When we design, for example, one of the things that we work on is unmanned aerial vehicles, and you put in a GPS sensor and it is a 10 hertz. So, in a second 10 data points come in. And this Data keeps on coming in and you keep on storing it. So, more sensors, advanced sensors, more frequency, cheaper storage, Data starts coming at a higher frequency and there you go. You have no larger set of Data than previous what it is.
6) 16 percent of the people believe in just a new kind of Data and Analysts. So, like for example, earlier days, you wanted to measure the temperature, you ended up using what you called a thermocouple or thermometer or something like that.

Now, you have hand-held laser parameters, infrared cameras, etc. that precisely measure to the level of 0.1 degrees Celsius, what the current temperature of some location, when non-contact measurement of the temperature and you put that thermal camera and reads it and from there gives you the temperature reading and keeps on coming in a continuous Data format to your system. So, now, we know instead of the temperature reading, you may even get a thermogram varying from different color codes.

7) So, now you are having new kinds of Data. The other one is what we call real time information. So, earlier, in the 90s, into the early 2000s. That is what we use Google Maps now for going from place A to place B. We used to print the directions and then follow the directions on a printout, there was nothing like a smartphone or hand-held device available. And those things in Google Maps came into picture, things became much easier. And now everybody turns from their mobile and goes from place A to place B. You do not stop and ask for directions anymore. So, that type of information, time to figure out where you are embedding a GPS sensor in your mobile phone starts giving you real time information, that can tell you okay you are here, if you go there do you get traffic jams, so, go through this way. So, that type of thing also starts to come into picture.
8) 7 percent of the people start thinking about Data from social media and Facebook, this book, that book, I do not know what other nonsense is available. So, that also creates a lot of Data as part of it.

(Refer Slide Time: 10:23)




So, certain industries, what happens?

1) Engineering and Manufacturing. There are a lot of applications going on in Big Data. One of the major things that is being done there is PLM or industry 4.0, etc. So, industry 4.0 is like the newer version, because industry

i) 1.0 is Mechanization

ii) 2.0 is Electrification

iii) 3.0 is Computerization

iv) 4.0 is Digitization.

So, the idea here is that earlier, if you want to design a car, you have to first draw it on a piece of paper. From there, you give it to a modeler, who takes a clay, makes a clay model out of it, fine tune it, then take it to a wind tunnel, do all of those things, make sure that your aerodynamic profile is correct, bring it back, make it into the using wood, you make the shape out of it, do all those kinds of things, then make prototype test them, those kinds of things.

Now, you have Product lifecycle management software, and Unigraphics systems, Advanced computer simulation software. You basically put the design into it and put the CFD (Competition Fluid Dynamics, Colorful Fluid Dynamics), put it in and run it and that is it, your design, the time taken, they reduce you take four to five years to do it. Now, you can do it in few months. So, by replacing the expensive materials and money with cheap information, you have been able to reduce time to for a product design.

So, that is one big thing, that is going on in engineering. Another thing is similarly quality, especially product innovations etc. So, a lot of work is going on in that area. So, that kind of thing makes sense. And 'Retail', there is a lot of effort that goes on what you call as forecast, anticipating the demand changes, market basket analysis is a very common thing that when you go to Amazon and order something, they say people who bought this also ended up buying this. So, do you want to buy it?

2) Then the 'HealthCare' we all know recently what happened with COVID. Then how the country's oxygen demand was managed, targeted drug delivery, you have the people who are working on ventilators, etc. US Brain initiative, these are all like places where you use Advanced Data and models and computer applications to actually deliver better health care to people.
3) 'Transportation' has been there for a long time. ATC stands for Air Traffic Control. So, it is where so many planes are flying in the sky. So, managing them makes sure that they do not collide with each other, fly safely and all those kinds of things you track where your flight fleet is fuel usage etc. See when a new business model, Rolls Royce, you think that Rolls Royce make fancy cars, rarely not. The Rolls Royce cars are now made by BMW, Rolls Royce no longer make cars, they make aircraft engines.

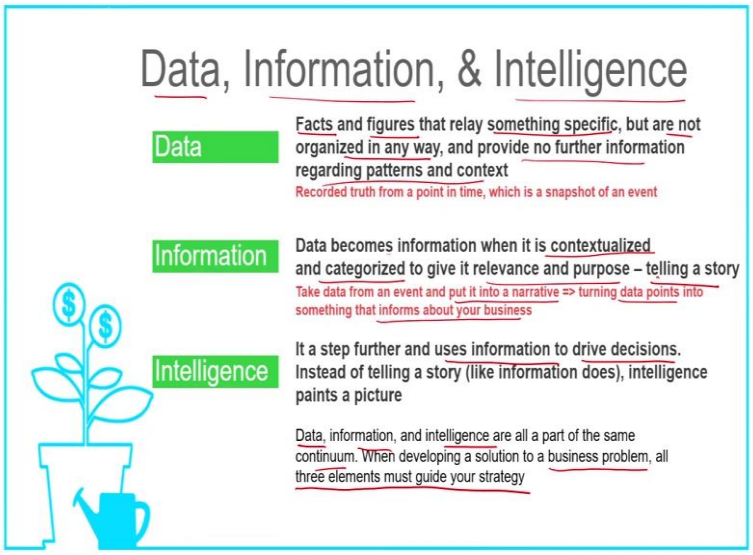
Now, Rolls Royce says we are not interested in selling you an aircraft engine, instead we will sell you power. So, each of the engines that is flying Rolls Royce monitors it from an under place and it measures how much thrust it has generated. So, if an aircraft flies from Lucknow to Delhi, when the aircraft takes off from Lucknow and lands in Delhi, by the time it lands.

Rolls Royce will send an invoice to the particular airline, 'Hey man, you flew by this aircraft it used to two Rolls Royce Trent 1000 CC engine and regenerated 600-kilogram force of thrust, as per agreed rate 1-kilogram force of thrust is 2-dollar, so, pay 1200 dollars' something like that. So, delivering a product as a service for that you require collecting data, measuring it, quantifying it and finally compiling it to make it into a product as a service offering. So, for these kinds of things, you require Big Data. Then there is social media. I kind of stay away from social media so you can read about it. That is, I have no clue about that.

4) The 'Defense and Security', there is a lot of these things especially, war games, different modeling, different scenarios, analyzing the lifecycle of the defense products, next generation weapons like the 5th generation fighter, stealth aircrafts using unmanned aerial vehicles for surveillance and reconnaissance, 24-hour global surveillance etc.

Satellite surveillance, automatic camouflage vehicle or resource detection, identifying as a human being targeting you. You name it. There are so many advancements that are going on the Defense and Security areas specifically. So, there are areas where I have taken a few of them, where it was very common, easily understandable.

(Refer Slide Time: 15:28)




So, some fundamentals and I will tell you, that Dr. Amandeep Singh will teach you better on this, I am just giving you a bird's eye view point. So, three important aspects I would like to discuss with you: Data, Information and Intelligence. And I want this concept to be very clear in your mind, because it is important. So, the first one is:

1) Data- So, verifiable flags, collectible flags, recorded flags, so many of those things. So, let us be a little bit more. So, these are 'facts and figures that relay something specific, that contain some specific, but are not organized in any way, and provide no further information regarding patterns and the context'. It is certain facts and figures that relay something specific. It gives you or points to something specific, but it is not organized, rarely organized, no further information regarding existing patterns, or a context.

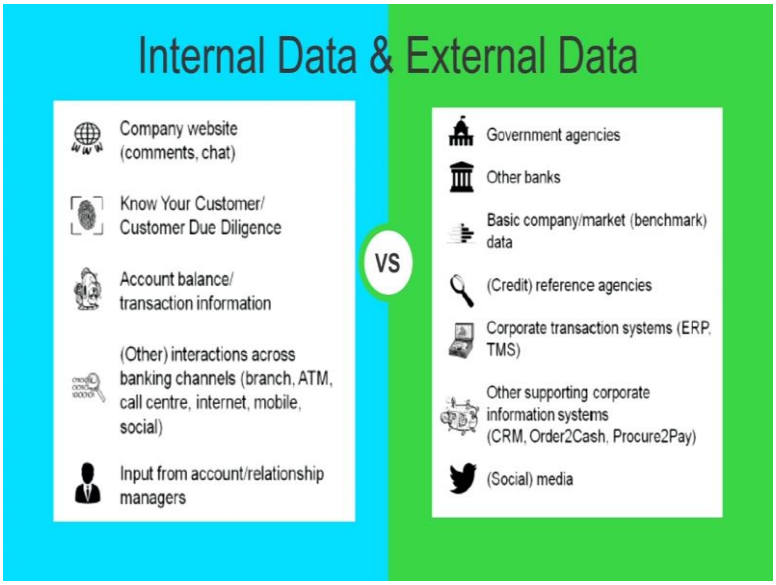
So, it is more like recorded truth from a point of time, you are taken a snapshot at a point in time, and that is it. So, there is a picture. So, there is a person surfing and you take a picture of a person surfing in the sea. So, you do not know why the person is surfing, whether he is enjoying it, or whether he is escaping from a shark, or running away to save life from pirates. You do not know. You only know, is that, there is an individual surfing and there are waves, and that is it.

2) Information- Data becomes Information, when it becomes contextualized and categorized to give relevance and purpose. When you put it in context, and categorize, then it gives you the relevance. So, instead of taking the snapshot, you are talking about telling a story. So, you imagine a bigger picture where somebody sees a person surfing. And behind that you see the shark fin coming and like, oh, this person is surfing purely to escape from the shark. And you see the Pirates of the Caribbean where Johnny Depp is chasing him, they are like, oh, he is escaping from pirates, whatever, So, the idea is that the Information is taking the Data of an event and put it into a narrative, So, Information in a way is, it is putting the narrative contextualizing. So, turning Data points into something that informs about your business. you have Data, and then from there it tells you something that gives you Information about doing something with the business. So, that is your information. So, you contextualize the Data, and categorize it to give relevance. So, that is the Information.
3) Intelligence- It is one step further the Information, and what it does is, it uses Information to drive decisions. That is why Decision Support Systems are important because it matures you from Management Information System (MIS), to what we call as Intelligence. It gives you the Intelligence, the alternative Analysts Alternatives whereabouts.

The idea is that instead of telling you a story, I am trying to paint a picture. So, it is like somebody is telling you a story, and you are listening to it. And you are also reading Amar Chitra Katha. So, you can basically visualize it, understand it, and probably be clear about it.

So, the idea is that they are not independent of each other. So, the Data, Information and Intelligence are all part of the same continuum. So, the idea is this, if you are a business decision maker, or if you are looking into the business decision, or making a business decision for a business problem, all three elements must guide your strategy or a decision. You should focus on how all the three will take you to make the ideal decision. That is the important part of it.

(Refer Slide Time: 19:21)

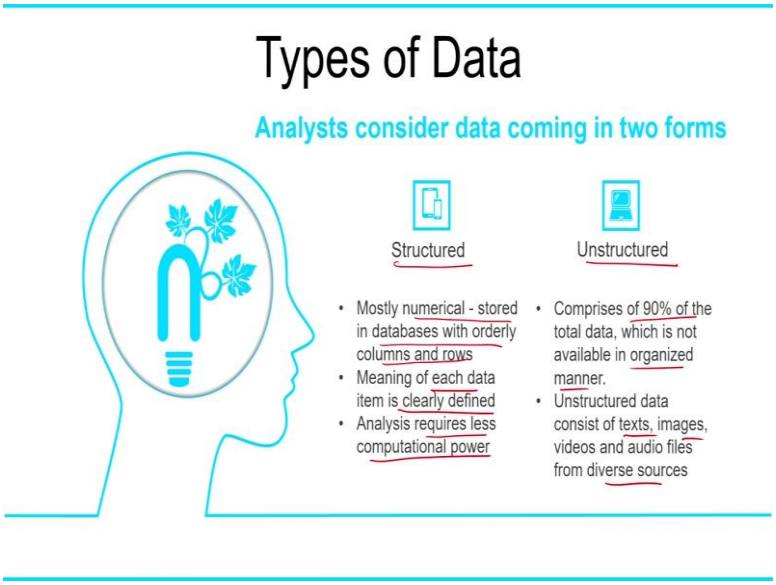



So, there are two major types of things in an organization, when you are looking into Data, there is Internal Data and External Data.

- And Internal Data is within the company, your account balance, your manufacturing, all those kinds of stuff. If it is for a bank, that is the branch, ATM, call center, etc. Customer relationship managers, blah, blah, blah.
- The external Data is government agencies, your lending agencies, credit, all those kinds of stuff. Social media, ERP, whatever you want to call it. So, there are so many other sources coming out of it. So, there is Internal and External Data.

Internal Data is something that you generate and you have some control over, whereas External Data is something that comes to you and most of the time we are forced to take it as it.

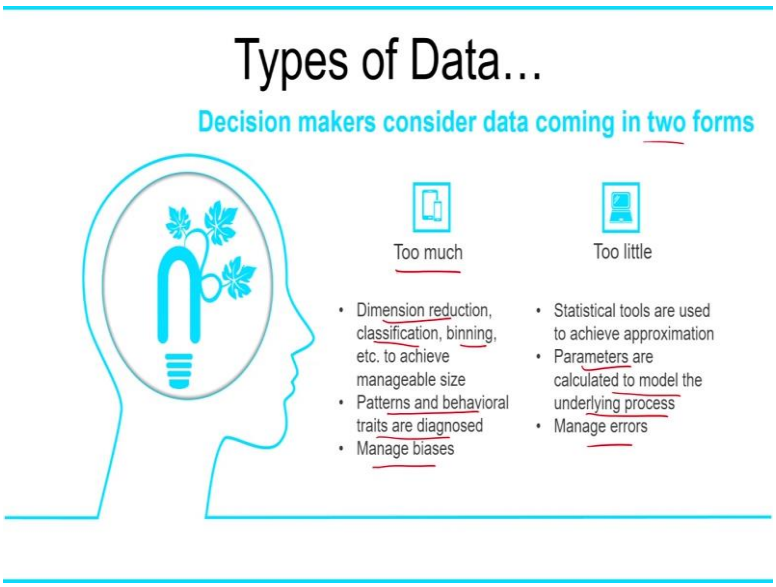
(Refer Slide Time: 20:06)




So, there are two types of Data according to Analysts, the people who analyze Data.

1) Structured Data- Mostly numerical. Stored in Databases with orderly rows and columns. So, that is what we have been discussing so far in the class, mostly numerical. And the meaning of each Data item is clearly defined. And unless this requires less computational power. So, if you store it properly, well-defined and everything, then ambiguity is reduced, you can clearly use it to make decisions. So, your computational power also reduces, you do not require a supercomputer to do this.
2) Unstructured Data- Mostly 90 percent of the total Data, which is not available in the organized manner. This is typically what is External to you, most of the External Data comes to you in an unorganized fashion. They are mostly text, images, videos, audio files from diverse sources and analyzing that is anybody's ballgame. you are like, I have no clue what to do. Or maybe I am looking for a needle in a hairstyle, kind of thing. This is not the decision-making problem; this is the availability of the Data. Look, we are trying to make a decision on Unstructured problems, not Unstructured Data. Data is different in this regard. So, most of the time, we would like to make decisions on Unstructured problems using Structured data, or whatever Data we have collected, organized and stored properly.

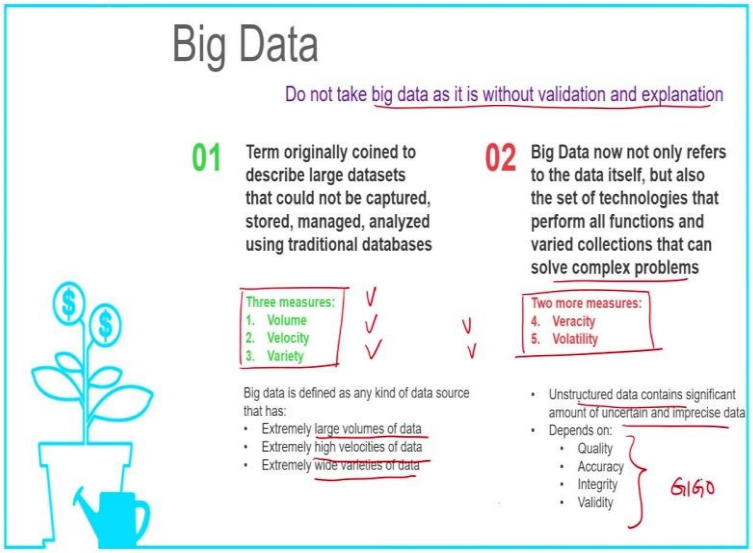
(Refer Slide Time: 21:31)




Decision makers, on the other hand, analyze that Data comes in two forms.

1) Too Much- This means too much Data, a lot of things are there, then you need to focus on reducing the dimension. Specifically, like classification, binning etc. to achieve size. So, if you want to eat an elephant, how do you eat an elephant, one bite at a time. To take a bite that you can chew, then you chew, then you swallow, take another bite, chew, swallow that kind of thing. And the aim of this is your reduction and another kind of thing is patterns and behavioral traits to be diagnosed. You want to diagnose what are the patterns and behavioral traits, and also to manage biases, inherent biases in the Data that you want. So, because somebody might have collected Data so much, and then that personal bias also might have caught into the picture. Sometimes Data comes in too little. There is not enough Data.
2) Too Little- You use statistical tools to achieve approximation. So, you use probability distribution, or approximate probability functions, etc. To do with that, and you try to calculate parameters, for an underlying model, or parameters to create a model that mimics the underlying process, and then use that model to study the Data that also is there and then when you do this, when you do not have too much of Data, and you create models or extrapolate or interpret how do you tackle the errors that is the other aspect of this problem.

(Refer Slide Time: 22:59)




So, Big Data is one of the fundamental things I want everybody to take, do not take Big Data as it is without validation and explanation, do not just take somebody who gave you a large amount of Data, does not mean that you should just take it as it. So, the first one is:

1) Originally this term was coined to describe large datasets. that could not be captured, stored and managed by traditional Databases. So, initially it was this Big Data. So, like for example, when the US Defense started using UAVs or drones for surveillance, the video feed from the drones is so huge. So, it is like something you find very difficult for you to store in a particular Database and analyze it. So, that became what we called as and can be taken as an example of Big Data.

So, initially, there were three volumes, three measures three V's. V, V, V is written as Volume, Velocity and Variety. So, large volumes of Data, extremely large volumes of Data, large quantities of Data, high velocities of Data. So, Data is coming in at an extremely fast rate. It is probably higher than what your personal computer can manage. And different varieties of Data. Some of them will be videos, some of them will be audio, some of them will be GPS feed, some of them will be object feed, Thermal Graham, Cassidy detection, you name it, like all sorts of different varieties of Data comes into picture. So, how all of them can be mixed and matched and make any relation between them. That is what the original Big Data problem turned out to be.

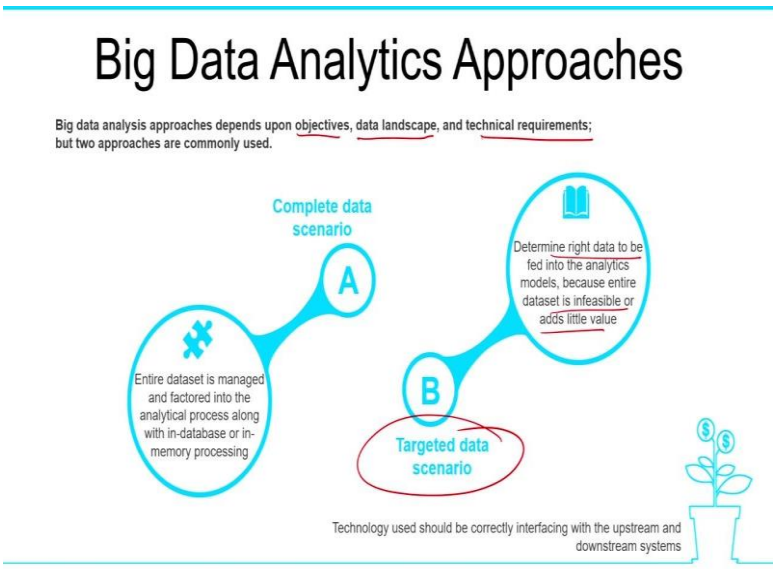
2) Currently, Big Data now not only refers to the Data itself. but also refers to the technologies to solve this complex problem. So, that introduced two more V on to the system. V and V Second V now it is 5 Vs, Veracity and Volatility. Veracity means you

need to be sure the source of the Data is where this Data is coming from. Who is generating this Data? How much of this Data I can believe that can and have an inflammation?

So, that is the major aspect of Big Data, then the other one is Volatility, can you store this Data? How much of this Data is after some point of time going to go away? How much of this context of the Data that you are able to understand? And after some point of time, you are not going to get the context later. So, Volatility is also another important aspect.

But it all depends upon these three things, quality, accuracy, integrity and validity. This is not tracing these four things. And remember, when you are talking about Unstructured data, it contains a significant amount of uncertain and imprecise Data, So, when you use, that is where the principle of guy go comes into, use garbage in garbage out that is very true.

(Refer Slide Time: 25:48)

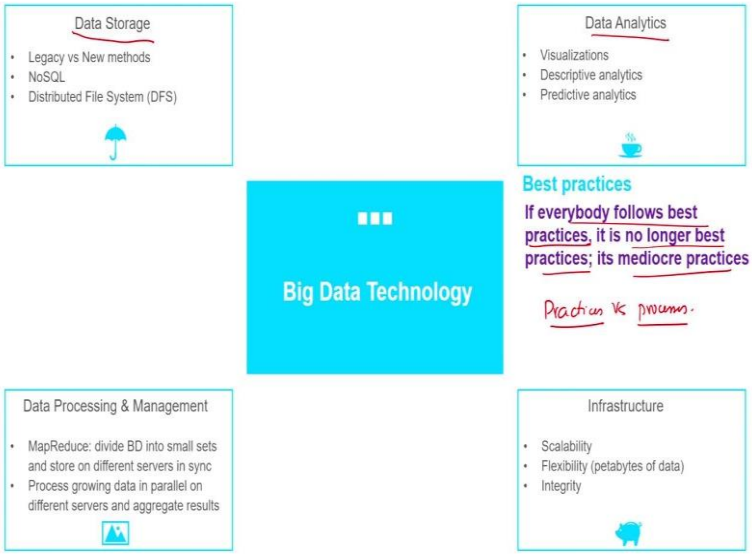



So, the two major approaches to do Big Data Analytics, either use the complete Data scenario, So, the entire Data set is managed and factored into the analytical process. in which you also do a lot of in memory processing. So, you try to use the entire complete set of Data. This Analysts depends upon the objectives, why are you doing this? What is the type of Data landscape, how spread it is? And what is the technical requirements? And what are you intending to achieve by this?

But these are the two common approaches one is the entire Data set, complete Data scenario, the other one is:

1) Targeted Data Scenario- The targeted data scenario is trying to find out what is relevant with you. And then you are trying to take it forward from there.
2) Entire Data Scenario- The right Data, because using the entire Data set is infeasible and adds a little value, where you are identified, there is too much noise, I do not want to spend too much of fortnight dealing with this.

(Refer Slide Time: 26:43)




So, there is Big Data technology, just to give you an update of this.

1) In Data storage, there are so many technologies coming in. So, three main things are:

- Legacy old school techniques, raid and other things with the new methods.
- NoSQL
- Distributed file system

2) In that Data Analytics, you have:

- Visualization
- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics

3) Data processing & management, the Big Data is divided into small sets, like an example:

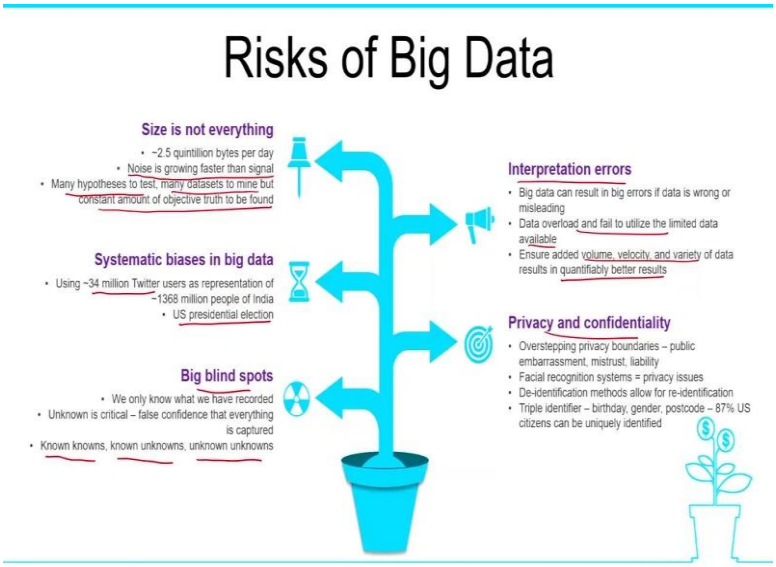
- MapReduce an excerpt or something like that. And also, Data in parallel and etc., like that is another aspect.

4) Then the infrastructure, the server farm or storage talk about cloud or whatever it is, but at the end of the day, you have to store Data at some place. So, how do you handle that storage?

- Scalable
- Flexible
- Integrity

And one of the common terms is what people say is industry best practice, everybody says I am following industry best practice. So, one of the questions that you should remember is, if everybody follows the best practices, it is no longer the best practice. It is a mediocre practice. So, I am a person who distinguishes between practices versus processes. So, processes are scripted, there is a very definite script and you take that script, follow it independent of who is following the script, you will get a proper outcome, as practices, it is heavily dependent on the individual who's doing it. So, you may not get the same output for two different individuals as part of it.

(Refer Slide Time: 28:18)




So, what are the risks of Big Data? These are major risk and I think that is pretty much the end of our discussion. But the major risk is:

1) Size is not everything, one fundamental aspect that I want you to take out of this is:

- Noise is growing faster than the signal.
- Many hypotheses to test many Data sets to mind, but constant amount of objective truth to be found.

So, if you want to say that I want to improve the productivity of the company, and that is pretty much you want to make the decision of that you may can talk about the average education qualification of the people who are working into that and try to model the improve the impact of this one on the of the education qualification on your productivity. Or you can look at the ratio of output versus input and look at how you can increase inputs and increase outputs and reduce inputs. So, that your productivity can be improved.

2) Systematic biases in Big Data-

- So, one example is the US presidential election, there is a phenomenon that was observed at one point of time in the La mayor election. So, it is like, when Donald Trump was running as the president of the US, a lot of the people who are asked, whom do you vote for? Will you vote for Trump or something? Or whether it is Hillary who was standing against it.

So, if you say that yes, I will only vote for Trump. People will look down upon you or somebody who is and the feminist anti against woman kind of a thing. So, many other people choose to keep quiet.

- So, people say that there are 34 million Twitter users, and we have, what 130, 140 crore people who ridiculous amount of people with that. So, using the decision based on the Twitter users is not really a factual representation of the population of the country. So, that is it. So, that biases come into picture.

3) Big Blind Spots-

- We only know what is recorded. So, the thing is, we have a usual thing, it is not the things that you do not know. But it is the things that you do not know that are going to hurt you. So, the unknown unknown is very critical. So, it is the known knowns, if that, it is not good, we know that it is unknown, that is also good, you can handle that. If you do not know that it is, I do not know it, then that is the difficult part. So, that is the blind spot that we are talking about.

4) Interpretation error aspect of it, because you have too much amount of Data then can bias you into one direction. So, and with the overload of Data, So, it is very common among pilots, you can see that when an aircraft is going through a distress or something and so, many alarms with the stall alarm going the master caution, beeping autopilot beeping, airspeed indicator, landing gear, flaps, blah, blah. So, the pilot would like, What the hell am I supposed to do.

And that fall process, he or she forgets that the fundamental law is to fly the aircraft. And that sensory overload comes into picture, So, that is very common.

- So, one of the things is when you are interpreting Data, you have to make sure that your added volume, velocity and variety of Data do help you in quantifying better results. There is a positive impact, there is not like I have, instead of looking at 10 kilobytes of Data now and looking at 100 megabytes of Data. So, my decision should be better. That may not be true, make sure that whatever you are looking for is necessary for you.

Then comes the Privacy and Confidentiality. This is a discussion for later down the road. We can talk about it later. But that is not an important discussion. But then a lot of court rulings, legal battles, fights, etc. talks about the need for privacy and confidentiality. And many times, now pitch, see, you may not want a picture of yours to be taken. But there is no guarantee that you may be part of somebody's selfie or somebody's image that is being taken. And with the digital cameras, images go all over the internet very quickly.

So, with this, we conclude the Big Data introduction or primer, and Dr. Amandeep Singh will for sure take you to the important aspects various case studies related to manufacturing industry 4.0, additive manufacturing, etc. So, that you have an idea what is going on, and how that can be used to make better decisions. And we cannot cover all those aspects. So, whatever is relevant, we will do that. And we will go from there.

So, this pretty much concludes most of our aspects related to the Database or DBMS Database Management Systems aspect of the Decision Support Systems. I will now later move towards both the User Interface and as well as the Application Program. And then we will come back to a model-base and knowledge base and then probably conclude the course as part of it. Thank you very much. Thank you for your patient hearing. We will see you soon. Thank you.