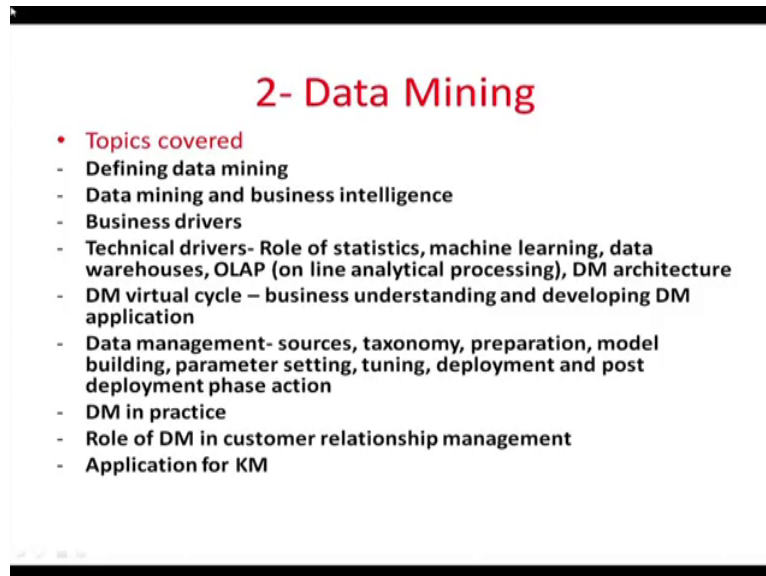


Knowledge Management
Prof. K B L Srivastava
Department of Humanities and Social Science
Indian Institute of Technology - Kharagpur

Lecture – 28
Data Mining.

Okay so we are moving to a new topic that is data mining.

(Refer Slide Time: 00:27)

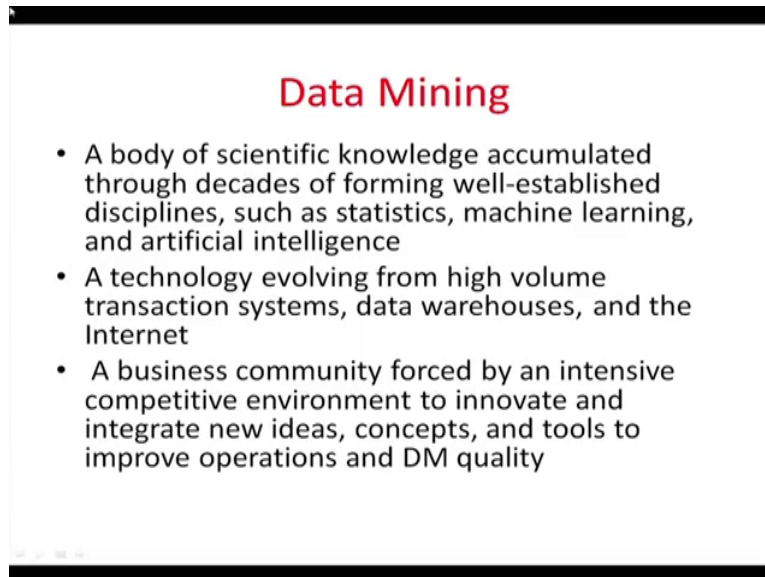


And data mining basically is used to infer certain decisions based on data analysis and other things and here we will discuss some of the issues relative data mining that how we can apply it in context of business. What kind of decisions can be taken based on the data analysis and other things and then we will also discuss about data mining and drivers what are the why we are moving to data mining.

What are the business drivers the technical drivers okay? What is the cycle and how we can better understand business using data mining and then we will also talk about on the second part about the data management okay. How we get the sources what are the sources how we prepare. What are the models that is used and how we identify certain criteria for decision making and then how we deploy.

And see that what needs to be done in the later stages then we will take some examples to show you that how data mining has been used in business context okay. So we are moving to data mining now first part that is data mining.

(Refer Slide Time: 01:33)



Data Mining

- A body of scientific knowledge accumulated through decades of forming well-established disciplines, such as statistics, machine learning, and artificial intelligence
- A technology evolving from high volume transaction systems, data warehouses, and the Internet
- A business community forced by an intensive competitive environment to innovate and integrate new ideas, concepts, and tools to improve operations and DM quality

Now what is data mining if you look at data mining basically it gives introduced English subjects and the input come from a number of areas maybe computer science maybe mechanical engineering maybe statistics material sciences right. So the inputs are coming from different disciplines like statistics which basically talks about what you called how to analyze the data using certain statistical tools.

And techniques then we are machine learning that is how you are going to simulate machines to infer design certain techniques for data analysis and then we also use artificial intelligence where we develop certain programs to make decisions based on computers, so when you are going to have used data in your organization it could be related to transactions various transactions that you make or it could be the data that is already in your archives or repositories okay, are in the data is that is coming to internet.

So that it could be different kind of things that has evolved the and that is how have been created a huge databases okay and since you have this database, but it is lying idle in your repositories because you are not able to make use of it or taking certain decisions okay. And that is why it is very, very important for you to use the data apply the data takes certain influences from them and then meet certain business decisions which could be advantageous to you.

Because you know that today you are going to work in a very competitive environment unless you innovate and continuously improve and develop yourself as an organization it is

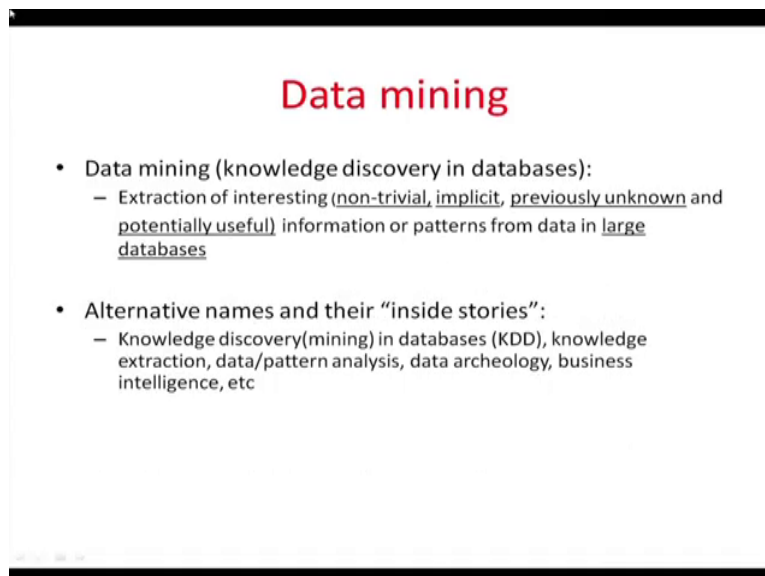
not possible for you to have competitive advantage because in this competitive world only those who are going to continuously improve and innovate themselves are going to survive and grow.

So what needs to be done is that how we can use this data and different kind of data that is available the organization, it could be related to the employees, it could be related to the sales, it could be related to the profit or any kind of databases which are there so what you need to do is you would need to mine the data. Mining is extracting patterns identifying say certain things out of it then making certain decisions which can add value to the business.

So it is very, very important to go for these kinds of mining activities mining in the sense that you are going to extract something okay. And that mining is to be done from the data right, for example when you go to a coal mine what you are going to extract is coal. So similarly when you are going to extract data you are not going to extract data but you are going to identify certain trends, patterns, values which help you to decision-making okay.

And then you have to see what kind of tools and techniques are to be used for decision-making using data mining right.

(Refer Slide Time: 04:45)



Data mining

- Data mining (knowledge discovery in databases):
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names and their “inside stories”:
 - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc

So moving further we will see that how data mining is used and what are the things that is done in case of data mining so data mining if you look at the definition is nothing else but discovering knowledge which is available with the databases. See databases are basically

information okay. Now if you are going to use and apply in a particular context it becomes knowledge as we have already talked about it.

So when you are going to have some kind of discovery from the data it means that you are going to identify something that is new something that is innovative, something that is known to you right so that is what we call data mining so basically you are trying to extract something which is not very trivial but it is there okay.

Which is not known earlier but now you will be able to know it and it could be useful for you right or you can identify certain information of items from the data which is lying with the large database. Let me give an example say for example this further say for example IIT Kharagpur has a huge database of Alimonies okay.

Alimonies means that the number of individuals who are passing out every year in different disciplines an area now they are passing out the database is there with the illumination now many of them are moving to different roles positions higher positions right. Now this database is not going to be useful unless you are going to source these aluminous and ask them to contribute to the growth and development of the organization right.

Now many of them are willing to give back to the Alma mater but what you need to do is that you need to identify who are the potential contributors okay. So you have to search the databases find out who are working different positions who are in a position to afford to contribute to the alma mater okay. So you need to extract this data based on certain rules okay.

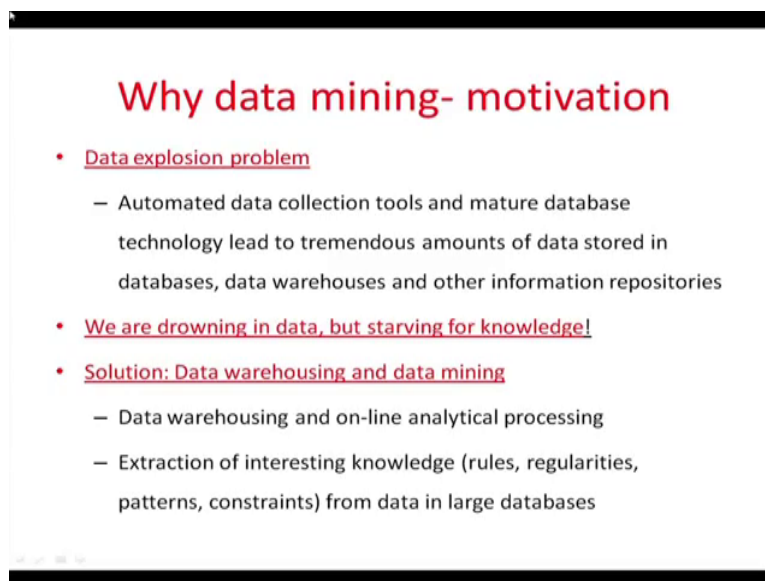
And then you can approach them and ask them to do it and that is going to be a potential usefulness you can say what the organization. So even if you have a large databases it is not going to be used why you then it is useless so what you need to do is that you see that how we are going to make use of data by extracting patterns, trends, values and these kind of things it means identifying something that is already there but you do not know about it okay.

So you are going to extract data you are going to do some kind of analysis to identify the patterns and trends, Say for example I taken and the ratio see we have to see that how that placement of IIT students are taking place every year okay. In terms of the quality of

placement, in terms of compensation, apart from the number of students that are being placed.

Now you have this database so can you identify certain patterns okay say for example what is the average salary trend, what is the quality of placement, right and you can use them to further strengthen your placement process right. So that is a decision that you are going to use based on the data of related to the placement right and that is where it is very important to use this data for potential application in the context of business and then based on it you are going to take certain decisions right. So that is what exactly data mining is.

(Refer Slide Time: 08:25)



Why data mining- motivation

- Data explosion problem
 - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
 - Data warehousing and on-line analytical processing
 - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Why we are going for determining, what are the motivations, why it is important for data mining? If you remember in these two examples that I have taken data related to the aluminum I mean the databases available how it could be a potential use or the placement data right. Why I am giving this examples to see that these data are lying idles you are having huge data of alimony or huge data placements year to year that is piling up in your repositories.

But why you want to make your graph because you want to extract something and you want to use your alimony for the benefit or you want to strengthen your compensation process or system you come into streamline it you want to see that IIT student's passing out will get better jobs with better salaries and a better prospects in life and later on they contribute to the alma mater.

Now what are the issues that are coming at data mining right? Now even if you have a huge database okay, and this database is increasing day by day okay. So what you need to do is that the data is exploding at a rate that cannot be matched. So every year the database is increasing right, so storage is another issue. And then you are going to store tremendous data in your databases right or repositories.

It could be other repositories and say archives that you have it could not be you are say websites where you have lot of data. You could have data in your warehouse is basically data warehouses or repositories right. So these databases warehouses and repositories which have stored huge amount of data okay, these might explode tomorrow because you are not going to make use of it.

And then there would be a time where you will not have enough place to store the data. So once that data is used in identify certain trends and patterns okay. That data might not be necessary to have it. So data explosion has become a big problem with most of the organization's they do not know what to do with this data. How to make use of this data how they can take certain decisions and that is where data mining comes to help you okay.

Another issue is that yes, data is there, information is there, but it has not been converted into knowledge okay, So the statement says that we are drowning in data but starving for knowledge though we have a huge database, but you are not going to make use of it significant use of it or apply it in business context right. So what is the benefit of having a data if you are not going to convert them into knowledge?

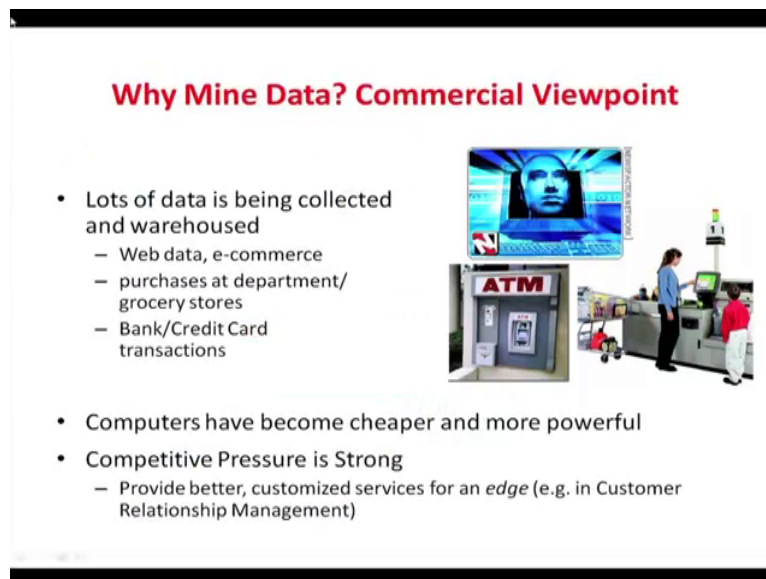
Which is relevant, which is actionable, which could be put into context to make certain business decisions. Now what is the solution if you have huge data, you are not able to convert them into knowledge which could be used in business context okay, and then the solution is that go for data warehousing and data mining. Now data warehousing basically is something that we are going to make use of data mining.

But not at the deeper level compared to what you call data mining right. And what actually happens in data warehouse data warehousing that you frame certain rules regulations and all kind of things so you can get some interesting always out of it, but if you are going for data

mining you can extract better things you can use statistical models and other form of say rules and regulations and from the databases right.

So it is very important for us today to go for some kind of data mining activities so that at least whatever data is there in the repositories or in the warehouses can be meaningfully useful.

(Refer Slide Time: 12:33)



Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/ grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

The slide includes two images: one of a person's face on a screen with the text 'DATA MINING' and another of an ATM with a person standing next to it.

Now if you look at a certain perspective like it has a business application commercial view point right. See lot of data is being collected okay. It could be related to web data it could be related to e-commerce or purchase at department or grocery stores or by bank and credit card transactions. So can you use this data to take certain decisions, say for example you have ATM for money exchange right.

Now if you look at how much money is being dispersed by ATM every month on an average for a year can you make use of this data yes you can identify the trends and patterns. How many people are using, how much amount is being dispersed and accordingly you can take certain decisions that took you how much money you need to feed into the ATM, so that people find it more convenient and comfortable to use right.

So there could be certain decisions like you are going for online transactions okay. So you can track your sales, you can track your growth that how much how many what are those items which are beings sold more, what are those items which are being sold less right in a

particular store. And then accordingly you can take certain decisions related to inventory management.

What are the items that you need to store more, what are the items that you need to store less, so these decisions can be taken based on the data analysis that data mining right. So these are the decisions which are commercial decisions which are the interest of the business and that is why you need to go for distilling right. And that is where IT and computers have come there in a big way to help you to identify these trends and patterns right.

And it is very, very important because you need to make profits as a commercial establishment your objective is not to only sell goods and for services but also to make profit. And when you want to make profit and have to take certain business related to this then you have to use these kind of databases to make certain decisions okay.

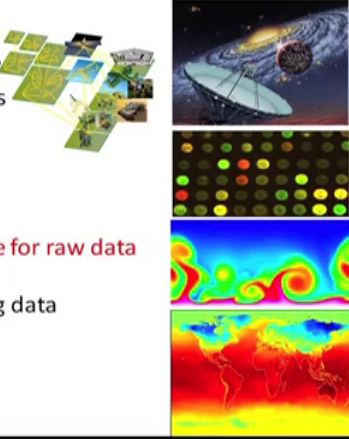
It could be related to online transactions, it could be related to inventory management, it could be related to tracking your profits and growth or even managing your customers also right. You can track how many customers you have, how many customers are leaving and then according you can go for some kind of relation management to retain customers to make them loyal and committed to your organization okay.

Because all these are the things which are going to provide you some kind of competitive advantage right, so it has a purpose and most of the commercial organizations are going for this kind of activities today to make certain businesses decisions.

(Refer Slide Time: 15:34)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
 - remote sensors on a satellite
 - telescopes scanning the skies
 - microarrays generating gene expression data
 - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
 - in classifying and segmenting data
 - in Hypothesis Formation



So that was the commercial view point now we will discuss about the scientific prospect that why we are going to have this kind of thing. You know that the data that is collected and stored in the repositories is huge we are still not sure that how many GB of data is being collected or being archived per hour so you can imagine. Every year how much data is being stored right like the data that is coming through satellites?

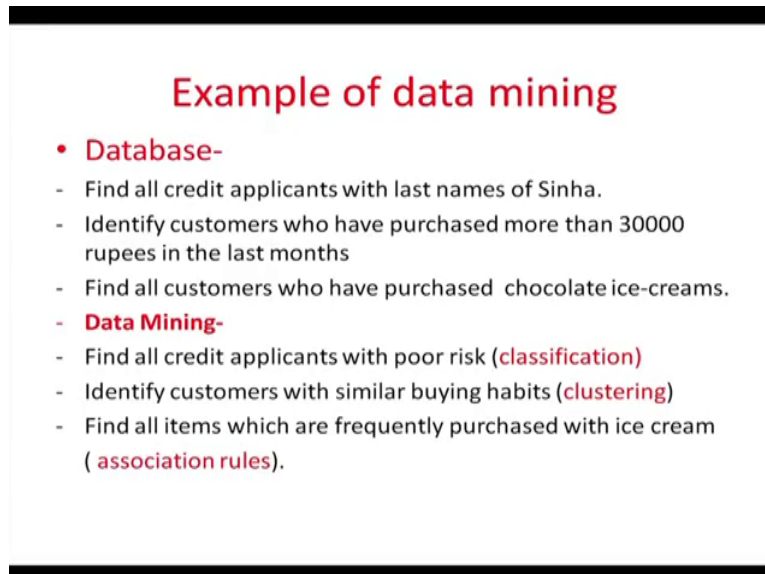
Or that when you are going to scan the skies or when you are going to see that how the genes are being expressed all a lot of examples that you have seen here right. So you are going to save accumulate terabytes of data and data and data but unless you make use of them to identify certain trends and patterns it becomes useful.

And that is why if you look at from the scientific perspective also it is being used by scientists, by engineers, by microbiologists, by genetic engineers and if you look at these examples you can see that how they have been able to identify trends in patterns okay. By simulating genetic as generating terabytes of data based on the set. So that they can identify trends and patterns and the taking they take takes certain decisions basis of this.

Now why we are going for data mining why not other techniques because it is not possible to structure raw data classifying and organizing them into meaningful information using other techniques and that is why it is important that we should go for data mining. Which basically helps you to classify in segment data based on certain parameters and you can also check certain hypothesis based on the data that you have achieve.

And then you can go for some kind of deductive reasoning okay if this still happening then this will happen right.

(Refer Slide Time: 17:28)



Example of data mining

- **Database-**
 - Find all credit applicants with last names of Sinha.
 - Identify customers who have purchased more than 30000 rupees in the last months
 - Find all customers who have purchased chocolate ice-creams.
- **Data Mining-**
 - Find all credit applicants with poor risk (**classification**)
 - Identify customers with similar buying habits (**clustering**)
 - Find all items which are frequently purchased with ice cream (**association rules**).

Now I am giving some examples of a data mining here for example say if you look at this particular example. It says that find all credit applicants with the last name Sinha. Another example, customers who have purchased more than thirty thousand rupees in the last month okay. And then all customers who have purchased chocolate ice cream. These are related to say vendors okay.

Now if you are going to use databases for data mining what you can do, you can go for classification. What does it mean okay, it means that you can identify all those credit applicants which I have poor risk, because based on this you can see whether they have been defaulters or not defaulters. So you can classify them as default or not defaulter's right. And then you can also buy identify those customers which are spending more than thirty thousand rupees.

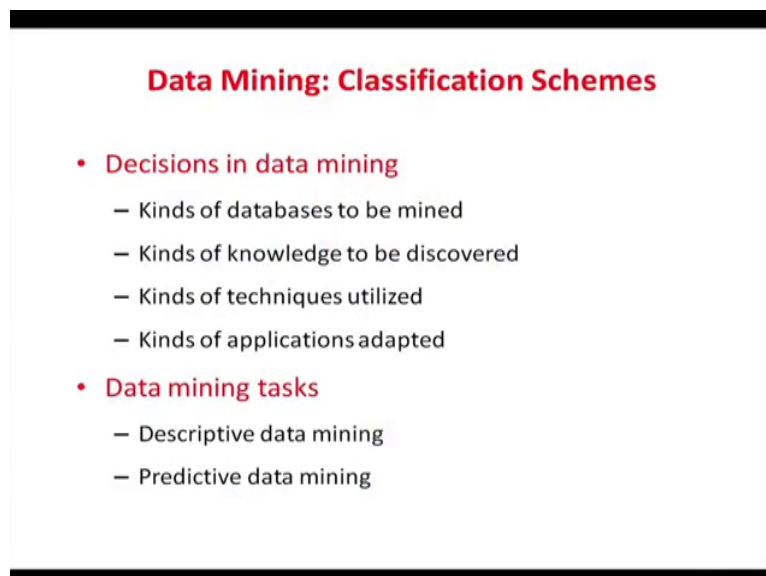
So you can cluster them together and you can see what kind of things there buy and take certain decisions. So that that is another technique that which is used in data mining is clustering. So classy classifieds classification we have already talked about if you remember the last part when we go for realizing the data we classify the data based on certain decision rules.

And then we go for clustering of the data, so you can identify clusters of groups or people with similar habits attitudes and behavior okay. And we can also identify certain rule association rules like okay this is relate to this. I mean all items which are frequently purchased with ice cream if a person is purchasing ice cream, chocolate ice cream and what are the other items they are purchasing okay.

So you know that if the person these people who are those people who are purchasing ice creams they are going to purchase these items. So you go for betray inventory management, okay we also need to keep these items because people are going to buy ice cream along with these things right. So these are some of the techniques which we will discuss further like classification, clustering and association.

And some of them we have already talked about it like association rules and clustering which is used in case of data mining.

(Refer Slide Time: 19:43)



Data Mining: Classification Schemes

- **Decisions in data mining**
 - Kinds of databases to be mined
 - Kinds of knowledge to be discovered
 - Kinds of techniques utilized
 - Kinds of applications adapted
- **Data mining tasks**
 - Descriptive data mining
 - Predictive data mining

Now how you go about decisions based on data mining right. Now if you look at the classification scheme we have to identify what are the databases. What are the different type of databases that need to be mined right. Then you also need to know what kind of knowledge you are going to discover right. So you make certain hypos okay you are going to use this data to identify certain things right.

Then what are the techniques that are used for data mining whether you are going to certain statistical technique or non statistical techniques okay. And what are the applications of this

information that is being converted into knowledge in that context of it right. So when you are going to take decisions based on data mining first of all you need to identify the database that is to remind.

Second, what is the use of the database for what purpose you are going to mind the database okay. Third, what are the techniques that you are going to use for mining the databases and finally what would be its applications. Where you are going to apply this information sorry knowledge that is being coming out of this database right.

So it could be in two forms right it could be descriptive or it could be predictive it means you are going to simply describe the databases right and second is that you are going to make certain predictions based on the database let me give you some examples of both. Say you want to identify what is the trend of admissions in the B tech program right.

So once all the admissions in all IIT's have been completed you can identify how many students have taken admissions across different IIT's okay, right. Then you can also describe this data based on gender right. How many girls' students have taken admissions how many boys students have taken and what is the ratio of boys to girls.

You can also describe this data based on and the factor that is rank wise okay based on the IIT ranks JEE ranks where they have taken admissions you can give a descriptive information okay that yes and this IIT has got students with better ranks. This IIT has got students with not so better ranks right. So rank can also be used as a data on inside for description of the data to identify which IITs are attracting better students.

Which IIT which IIT's are not able to attract better students but this information is only descriptive in nature because what you have done you have used certain parameters to describe the data right. Say for example based on the socio-economic status you can identify. You can also identify state wise how many schools have cleared JEE right. So but all this data is going to be based on certain parameters whether it is rank, whether it is status.

That is socio economic status, or whether it is going to be something else. But that is going to be more descriptive in nature right. So that is what we call descriptive data mining coming to predictive data mining. Predictive data mining is that when you are going to use this data

based on the analysis you are going to predict certain things okay. So for example you can predict based on the trends of the admissions based on the rank okay.

That it is likely that the students with higher ranks are going to take admission in this particular IIT right. So this kind of predictions can be made based on the past data and then it becomes a more predictive analytics, so predictive data mining what happens you are going to identify the trends and patterns for the future. So if you have databases say up the top hundred our top 200 students where they have taken admissions.

So you use certain statistical techniques and can identify that top rank holders are likely to take admission in which IIT. That is, you are going to predict that yes, and they would be taking admission in these IIT's. What next, once you are able to identify this then you can take certain decisions. That what needs to be done by other IIT's, who are not able to attract students okay? And that is going to the business decisions right.

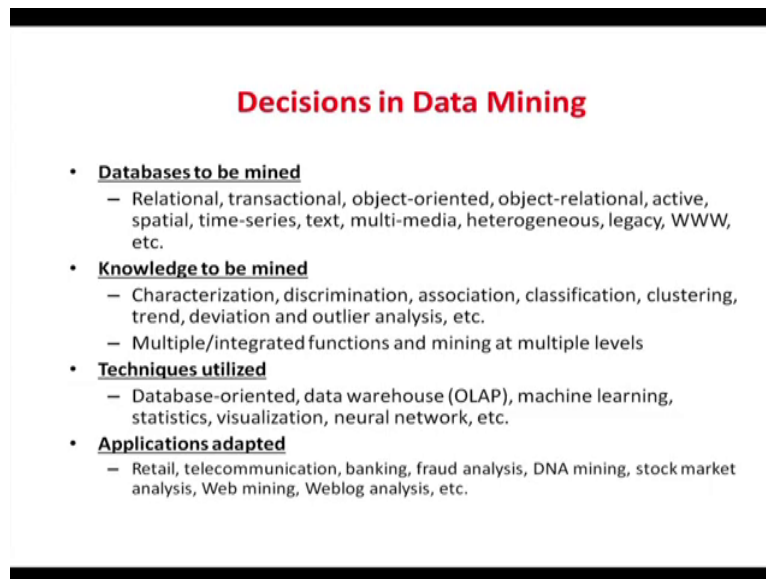
So when you are going to take decisions based on data mining. So you have to identify what kind of databases you are going to use? What for you are going to use this decision based on that data bases of JE admissions right. You are going to what you are going to discover that is what is the sex ratio? What is there, what are the status of the students who qualify JE. Right, from which states they are coming?

What is the right different kind of ranks? And based on that you are going to analyze certain techniques either descriptive statistics are influencing statistics to describe the data or you are going to predict certain things right. That is I am talking about the technique and finally the application and then you are going to make use of this data, once you have identified the trends and the patterns right. How you are going to make use of this data?

It means say for example to find that sex ratio is very skewed it means there are more male students coming to the system then female students. Now I can need to encourage and motivate Girl students to come to IIT's right to study science and engineering so the mystery and IITs can take a decision that how they need to encourage female students to join IIT's. So that is the application part right.

And that is where you are going to take certain decisions based on data mining right.

(Refer Slide Time: 25:51)



Decisions in Data Mining

- **Databases to be mined**
 - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
 - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
 - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
 - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
 - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

So further description databases that is be in mind you have to identify it could be relational databases, it could be transactional data bases or it could be object-oriented relational activities it could be days and the databases based on time, text databases, multimedia databases okay. Or worldwide web databases so you have different kind of databases the only things that you need to identify which database is to be used by you.

Then what kind of knowledge is to mined. Whether you are simply looking into the characteristics of the data or you want to discriminate between different categories and groups, you want to make some kind of association between different factors which are there right. Or you want to classify the time to different categories I want to cluster them.

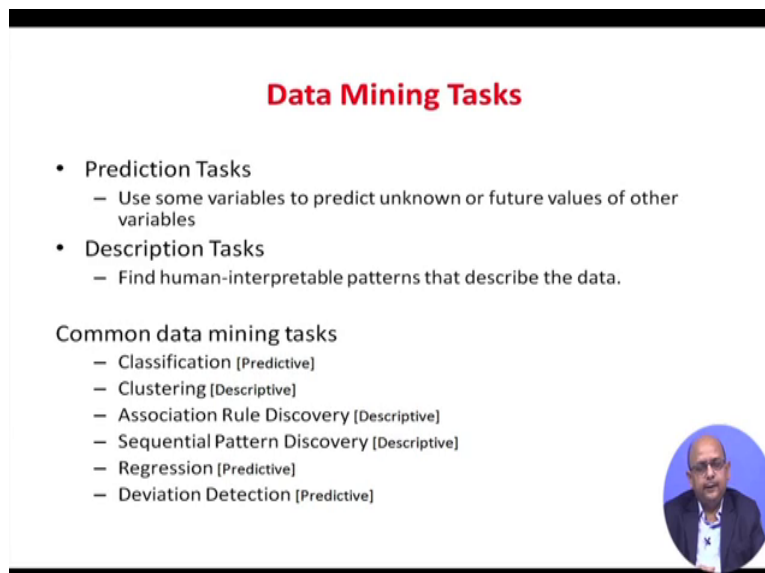
Say for example state-wise data if you have of JE students who have qualified then you can cluster them okay. For each state how many students have qualified right. So you can find identify clusters of students who are coming from a particular state right. So that is how you can identify the cluster you can go for trends you can identify okay from which states most students are qualifying.

You can also deviate it and go for out layer analysis because you will find that okay from some states more students are qualifying compared to the average okay. And then you can also go for multiple integrated functions and mining at multiple levels. You can not only at one particular level but you can move further to different level to identify and do the same kind of activities okay. Then what kind of technology you are going to use okay.

Whether it is database oriented or you are using online electrical processing especially what is known as data warehousing. Or you are going to use machine learning methods more simulated environment right. Or you are going for data visualization or using certain statistical techniques for description in prediction okay. Or you can also use neural networks we have already talked about it.

Then what are the application business applications. Especially in case the banking, fraud analysis telecommunications okay, stock markets, web mining all kind of things that can be taken up from the commercial perspective, so when we are talking about decisions and resubmitting it is very, very important and that is how you are going for data mining okay.

(Refer Slide Time: 28:09)




Data Mining Tasks

- Prediction Tasks
 - Use some variables to predict unknown or future values of other variables
- Description Tasks
 - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]



So what are the different tasks that is done that is prediction task you are going to use certain variables to predict unknown or future behavior of those variables right. Other you are going to simply describe the characteristics of the data, based on the analysis of the data you can identify certain patterns that is going to describe the decision.

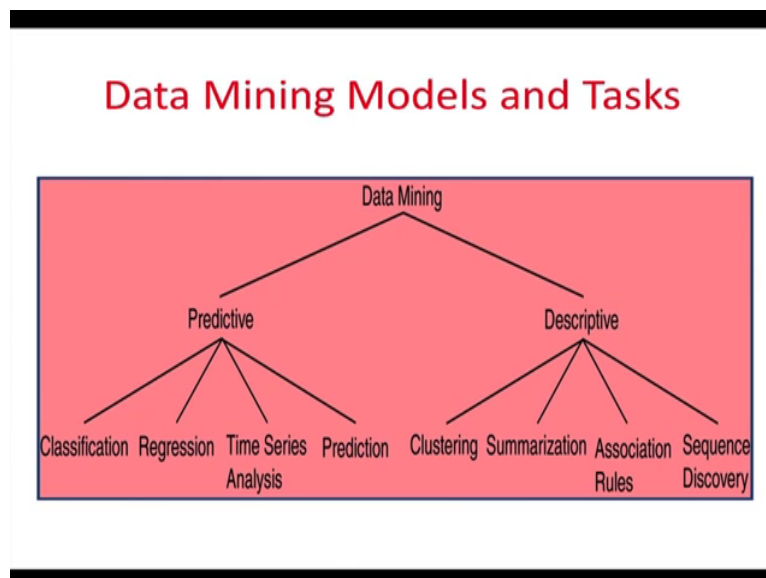
For example based on the JE admissions you can identify okay how many students are qualifying right. From which to state their qualifying? What is the ratio of the boys to the girls so this is basically the description of the data now how we go about data mining different activities that is you go for classification that is you want to use some kind of predictive analytics then clustering.

Where you want to describe the data based on the patterns, then you are going to again do certain Association rules discovery for example you can say that okay those students who are having good ranks will take admission to on this particular IIT. So you are going to try to associate but that is again descriptive, then you can also identify certain patterns okay. What will happen that is again descriptive.

Then you can use certain predictive analyst analytics like using certain statistical forms of analysis like regression analysis and these kinds of things where you try to predict future pattern okay. Like we are going for multiple regression analysis based on certain factors how you are going to predict then you can also use other kind of techniques basically the idea is to predict okay.

Then you can also identify deviations like what are the outliers and those kinds of things based on the data. So these are the different kind of data mining data's that is used okay.

(Refer Slide Time: 29:54)



Look at it this, this talks about how we are going to use data mining models and tasks. Now if you look at this it talks about to kind of determining techniques that is descriptive and productive in case of predictive we have classification, Regression, time series analysis and predict regression and prediction.

In case of descriptive techniques we have Clustering, summarizing the data to identify the characteristics, identifying certain association rules or discovering the secrets right. So these

are the data mining models that are used in terms of techniques and activities that take place in case of data mining.

(Refer Slide Time: 30:33)

Data mining and business intelligence

- Data mining helps in producing new knowledge and discovering new patterns to describe the data using intelligent automated systems.
- BI is a global processes, techniques, and tools that support business decision making based on IT.
- Approaches can range from a simple spreadsheet to an advanced decision support system
- **DM and the three bodies of knowledge-**

Source: Awad and Ghaziri: Knowledge Management, 2007)

The slide features a Venn diagram with three overlapping circles. The left circle is labeled 'Scientific Knowledge', the right circle is labeled 'Information Technology', and the central circle, which overlaps with both, is labeled 'Business Community'. To the right of the diagram is a small circular inset image of a man with glasses, wearing a dark jacket over a light-colored shirt.

Now we will talk about data mining and business intelligence right. So as you remember we are talking about that there has to be some kind of business applications of data mining so that you can identify certain new knowledge discover certain new patterns okay which could be potentially useful in case of business.

Now you are trying to develop certain software such an automatic system which is going to be fed with the data and then you are getting certain output okay. And that is what we know as business intelligence. Business intelligence is nothing else but it includes tools and techniques which helps you to take certain business decision which is based on information tech systems right.

So it these are basically automated committed systems using software which analyzes the data and then give certain outputs which could be helpful for taking certain business decisions right. It could be like a spreadsheet, excel sheet okay. The beginner's level and then you are moving to a more advantage and support systems.

Where you feed the data and then it also not only gives out some but tells okay these are the decisions that could be taken okay and then you have to decide which decision you have to go far right. So if you look at data mining it includes three different type of knowledge that is it includes scientific knowledge, it is going to also help this business community.

And it is it high the basis of IT because eighty IT is going to provide you the technology okay. For automation, for analysis of the data right, it is supports the software so that you can use to software to analyze return from get the pattern right and that is why we find that a data when is very important to make certain business decisions. Thank you.