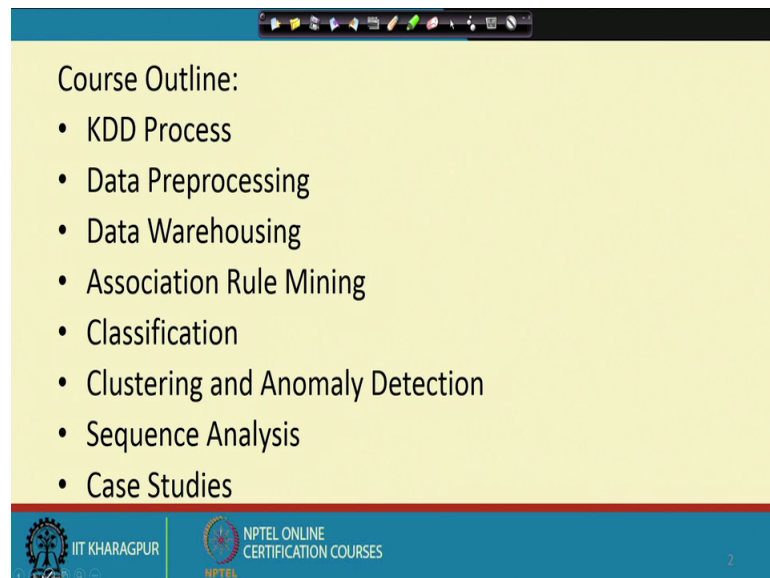


**Data Mining**  
**Prof. Pabitra Mitra**  
**Department of Computer Science & Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 01**  
**Introduction and Knowledge Discovery process**

Welcome to lecture one, of the course on Data Mining. In this lecture I will provide an overview of the topic of data mining, including the motivation, various aspects and the steps. I am Pabitra Mitra from the Computer Science and Engineering Department at IIT, Kharagpur.

(Refer Slide Time: 00:48)



The slide displays a course outline for Data Mining. It features a yellow background with a blue header and footer. The header contains a navigation bar with icons for back, forward, and search. The main content is a list of topics under the heading 'Course Outline:'. The footer includes the logos of IIT Kharagpur and NPTEL Online Certification Courses.

Course Outline:

- KDD Process
- Data Preprocessing
- Data Warehousing
- Association Rule Mining
- Classification
- Clustering and Anomaly Detection
- Sequence Analysis
- Case Studies

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as you know data mining is; this is the outline of the course, first I will talk about the of the entire course of the KDD process knowledge discovery process, and then I will talk about its different parts like data preprocessing, data warehousing and various talks like association rule mining classification, clustering and anomaly detection, sequence analysis and finally I will present some industrial case studies.

(Refer Slide Time: 01:26)



The slide is titled "Why Data Mining?" and features a bulleted list of points. The first point is "The Explosive Growth of Data: from terabytes to petabytes", which is further broken down into "Data collection and data availability" (including automated tools, databases, and the web) and "Major sources of abundant data" (including business, science, and society). A second point states, "We are drowning in data, but starving for knowledge!". The final point is a quote: "Necessity is the mother of invention"—Data mining—Automated analysis of massive data". The slide footer includes the IIT Kharagpur logo and the NPTEL Online Certification Courses logo.

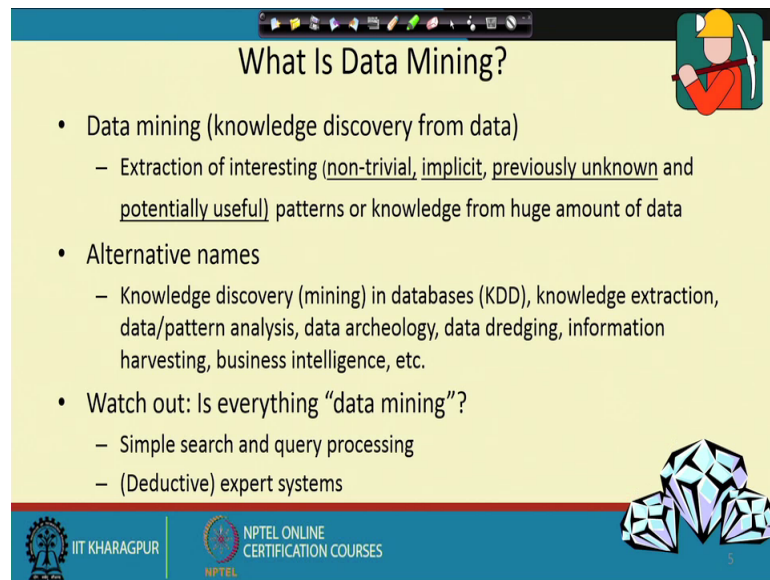
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, ...
    - Science: Remote sensing, bioinformatics, scientific simulation, ...
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data

So, let me go to the motivation of the topic; there have been an explicit growth in the volume of data, because of ease of storage because of ease of data collection, because of more computerization of various companies and daily life.

So, business like e-commerce, Amazon, Flipkart whatever you use the web, Google or any search engine bank transactions stock markets; various scientific areas like remote sensing images, biological data, scientific simulation various social media sites like news digital photography, the flip kart and others YouTube video. So, there is a huge volume of data, but it is the data is not giving us enough knowledge.

So, this gave rise to the method of automated analysis of such massive data together meaningful knowledge, and this gave birth to the subject of data mining. There is an alternate name to data mining which is knowledge discovery in databases. So, both of these names are used in industry as well as research KDD knowledge discovery in database as well as data mining. I will come to my previous slide ok.

(Refer Slide Time: 03:23)



**What Is Data Mining?**

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let me define first the term data mining. Here is the definition it is the non trivial process of extracting interesting previously unknown, which brings about the discovery part and potentially useful pattern or knowledge from huge amount of data. So, each of these terms in this definition is important.

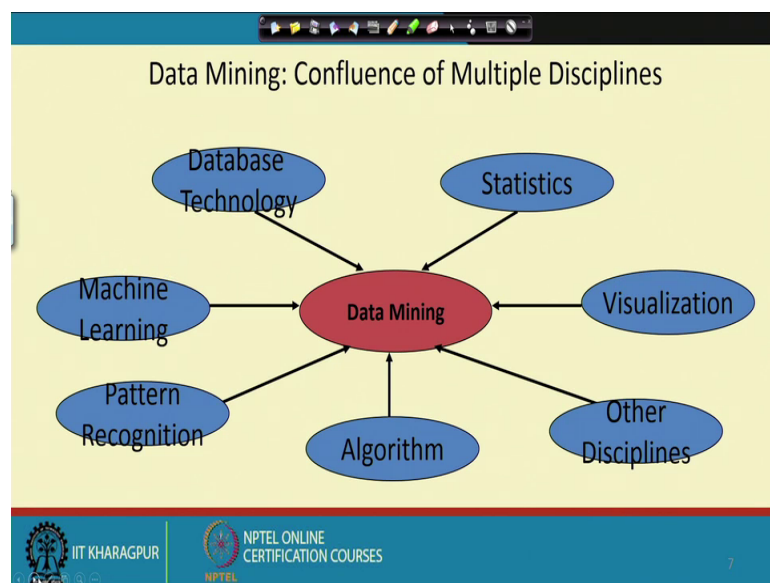
First it is a non-trivial process; which means that it is not obvious the knowledge is not obvious it has to be extracted. It is implicit in the sense that the knowledge is inbuilt in the data you extract it only from the data. There is a novelty part which means that the knowledge has to be a new knowledge and unknown knowledge previously. And finally, it has to be potentially useful this has to be useful knowledge depending on the application. So, the knowledge often takes the form of patterns in data, some regularity or some kind of structure in the data and from huge amount of data that is also an important aspect.

This definition separates this process from two other data analysis techniques. So, the plain search like we do in Google or any search engine; similarly query processing in a r DBMS system relational database management system, which you do in a transaction value. So, for example, query how much balance you have in your account these are not data mining. So, the plain querying in Google as well as say you are booking a ticket in Indian railway irtc dot com and you want to find out how many reservations are freely

available in this train on this day. This is query this is not data mining. Data mining would be from historical data, not from the existing data and would be on historical data.

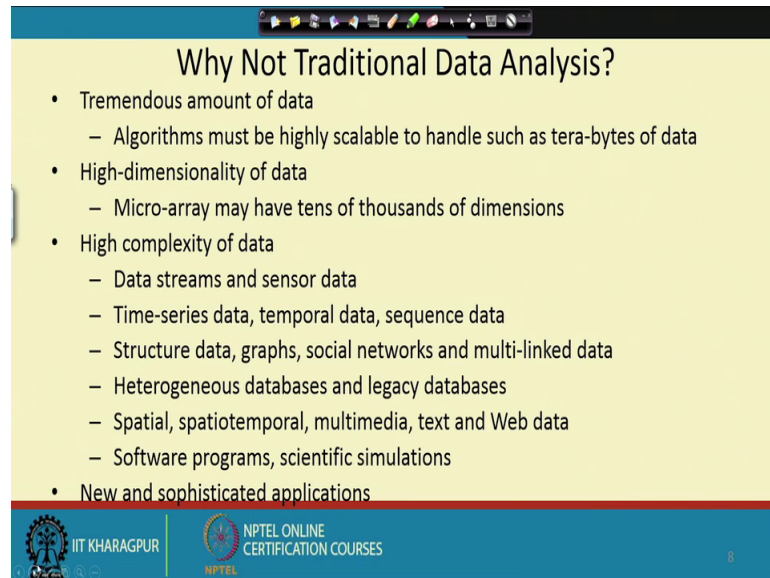
So, this similarly we also know there is another extreme besides this query processing is the expert system, where you do a logical inference. So, you know what deductive or expert systems do is that given a set of axioms or statements it tries to derive new statements. So, using a process of logical reasoning so, that is also not data mining because that is a concrete well defined process that is also not data mining.

(Refer Slide Time: 06:53)



So, these two are not data mining; data mining as we have defined is the non trivial process of extracting useful and novel patterns or knowledge in the data and how you do that what we do in that process that of course, in the course of the these lectures I will explain them. Before going as I had mentioned in the previous lecture also this entire study is very much interdisciplinary, it borrows from database technology, it borrows from statistics, it borrows from machine learning pattern, recognition algorithms cognitive theory visualization lot of aspect.

(Refer Slide Time: 07:51)



The slide is titled "Why Not Traditional Data Analysis?" and lists several challenges of traditional data analysis. The background is light yellow with a blue header and footer. The footer contains the logos of IIT Kharagpur and NPTEL Online Certification Courses, along with the number 8.

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

So, the reason that course in this course I would like to highlight is that, how does techniques from all of this varying technologies come together to solve this important problem of knowledge discovery. I would also like to highlight some of the drawbacks or limitations rather of traditional data analysis. Statistical data analysis where people have been doing in any reporting or industrial thing which is inadequate for certain purposes.

For example, traditional algorithms would fail when there is a large volume of data, when there is a high dimensionality there are very many aspects of the data, when the data is very complex where you have heterogeneity in the data, various types of complexity in the data, these and many more sophisticated applications that data mining a standard data analysis is inadequate and data mining comes into the picture.

(Refer Slide Time: 08:49)

The slide is titled "Data Mining: On What Kinds of Data?". It lists two main categories of data sets and applications:

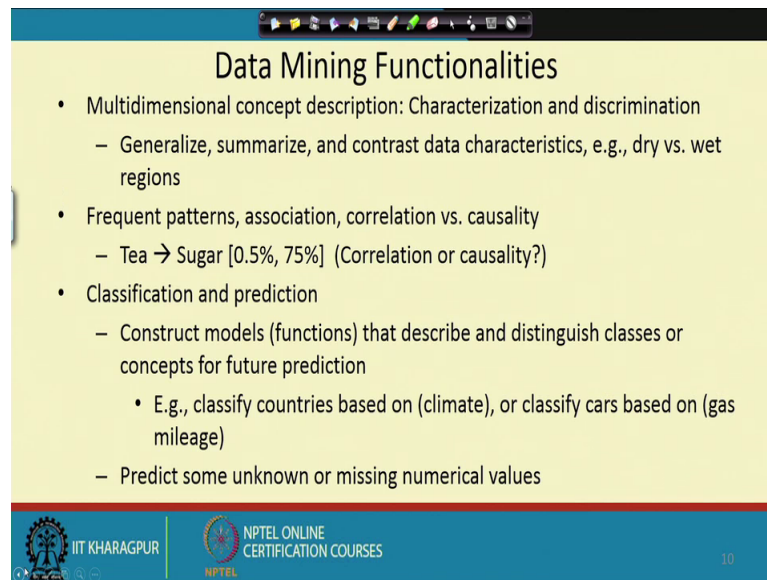
- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

The slide also features the IIT Khargapur logo and the NPTEL Online Certification Courses logo at the bottom left, and a small video inset of the presenter at the bottom right.

I would also like to outline because after the end of the course when you actually implement a data mining system, it will be possibly on one of this following kind of data. So, it will be the data will come from either a relational database, data warehouse or a transactional database or it will come from more advanced applications say for example, sensor streams, you have an industrial sensors measuring say temperature and pressure you can have sequences like say stock markets. You can have graphs and social networks like Facebook or any other linked structure, you can have the object oriented database you can have spatial data geographical information system special temporal data.

You can have multimedia data, you can have text data, you can have web data. So, each of these applications have their own research issues, they have their own adaptation of the basic techniques that I will teach in the course, each of them have their own challenges and own requirements and I in the beginning of the course I want to just make you aware of what kind of applications you might face, when you actually do an data mining; exact application of course, I explained it.

(Refer Slide Time: 10:34)



**Data Mining Functionalities**

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation vs. causality
  - Tea → Sugar [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 10

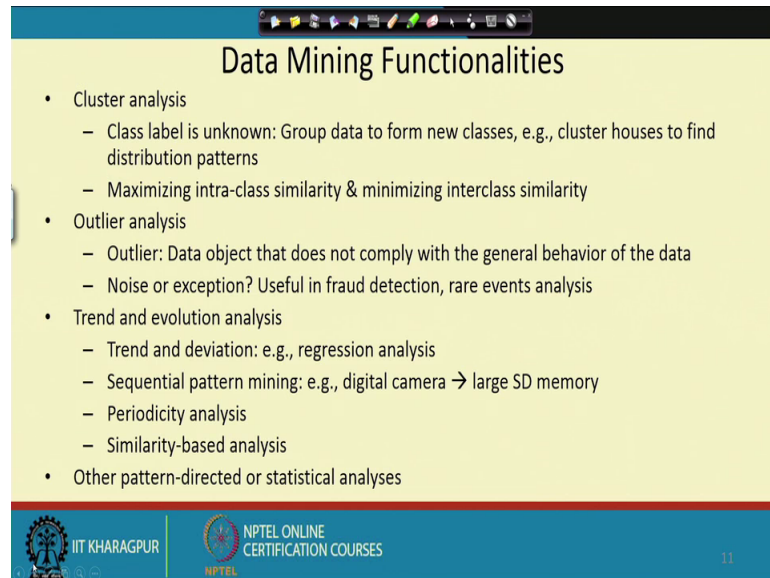
And also let me quickly summarize, what are the patterns or what are the types of knowledge that we would like to discover in the data mining process from all these types of data that I had hinted that can be applied to. The simplest pattern is kind of a concept description you just summarize the data, you tell some characteristics.

So, So, what is a market segment for example, if I have the log of all the transactions in say a big bazaar a retail chain, I want just want to describe; what is the segments of customers that come just a description. A slightly more complex pattern would be something called an association, which says that between different events or different attributes what can be the correlation or causality for example, in a retail store maybe whenever a person buys tea, see most of the time buy sugar also. So, this is an association, such patterns are indeed very useful and this is the topic of something called association rule analysis which of course, will we will cover in the actual lectures. This was the first pattern which was which was studied when the data mining subject came into picture. IBM research labs first proposed a system for doing this and many of the early developments of this kind of data mining tools, happened in the IBM research labs besides many universities.

Then you have classification, where you classify an event or tag an event depending on groups. For example, somebody has applied for a credit card whether I press that person in a fraud group or actual group proper ordinary group, this is a classification problem

we have an email do I classify it as spam or a non-spam I have a news article, which is a sports news or a political news or a science news that is a classification problem.

(Refer Slide Time: 13:26)



The slide is titled "Data Mining Functionalities" and lists several key areas of data mining:

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

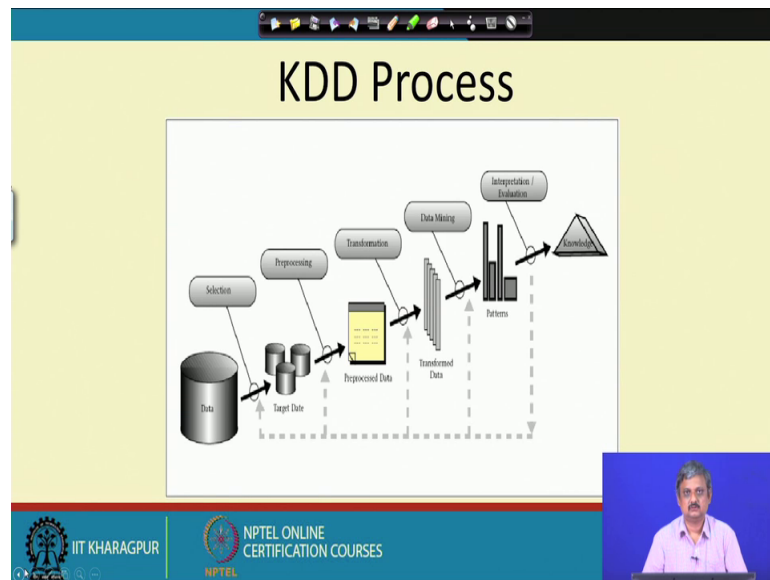
The slide footer includes the IIT KHARAGPUR logo, the NPTEL ONLINE CERTIFICATION COURSES logo, and the number 11.

Next we have a slightly variation of this classification problem called the exploratory analysis or cluster analysis where these groups are not previously defined, you have to discover these groups by doing what is called a cluster analysis. A related term with this cluster analysis related tasks is what is called a outlier or they anomalies or they are or they are unexpected variations that is an outlier discovery of that. And then you have a trend or a sequence or a evaluation analysis.

How does things change with time. So, analysis of that finally, you have more complex patterns we have more complex patterns, more wait more complex patterns involving graph structures and other kind of structures.



(Refer Slide Time: 14:42)



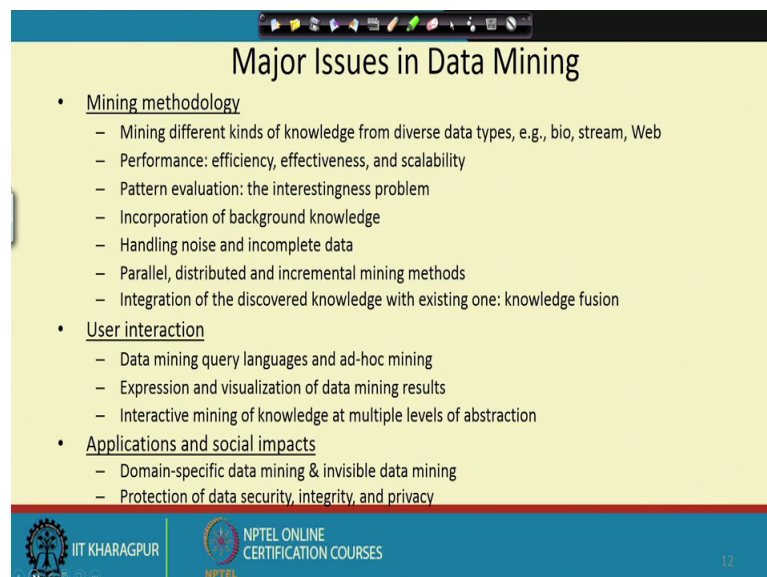
So, these are the talks let me tell you about the process of knowledge discovery in databases. So, in the beginning of the process you have a database which stores the data, not just one you may have multiple databases each contributing. And then what you do is from the database you first do a process of selection, which gives you only the data you are interested in and putting together such data you get a selected database often put in a data warehouse. On this data warehouse you apply various preprocessing techniques. So, data warehouse is kind of a normal relational data modified and stored and integrated in a form, which may be useful for further mining data mining from it.

You can then pre process the data and then you have the pre processed data; you might also want to do a data transformation various methodologies for that we will also discuss, you can do a data transformation which will change the data into a suitable form. So, you must have done this in many of your data analysis tasks, when directly whatever variable we measure we do not take we transform it into a form which is more insightful which gives us more information, and then process that data. Finally, after selection preprocessing and transformation our data is ready for applying a data mining core data mining process. What do you do in the core data mining process? You discover pattern what is a pattern? Pattern is nothing, but a succinct representation; that means a few statements for example, you are measuring voltage.

And current across a resistance for different value of voltages you will get different values of current, maybe for 1000 different values of voltages you will get one thousand different values of current, but you can summarize all this 1000 pair of voltage and current into one single information, that voltage equal to current into resistance the ohms law. So, the ohms law is a pattern in this data, it is a mathematical description a model of this data. So, this data mining process would give you a mathematical model, but we do not stop there, what we do mathematical models are mathematical models often they are not interpretable to human beings, whereas the final user of the knowledge is a human being of an a manager of an industry. So, we interpret we visualize we convert this mathematical model into a human interpretable form; by a process of visualization, by a process of actionable you convert it to an actionable form.

And a mathematical model with the visualization and an associated actionable item is a knowledge, that is what is finally, used. And I have already listed some of these mathematical models that would better be used it can be a concept description, it can be association rule, it can be a classification all these are different mathematical models that we try to fit to the data.

(Refer Slide Time: 18:45)



The slide is titled "Major Issues in Data Mining" and is presented in a yellow box with a blue header and footer. The content is organized into three main bullet points, each with sub-bullets:

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

The footer of the slide contains the logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, along with the number 12.

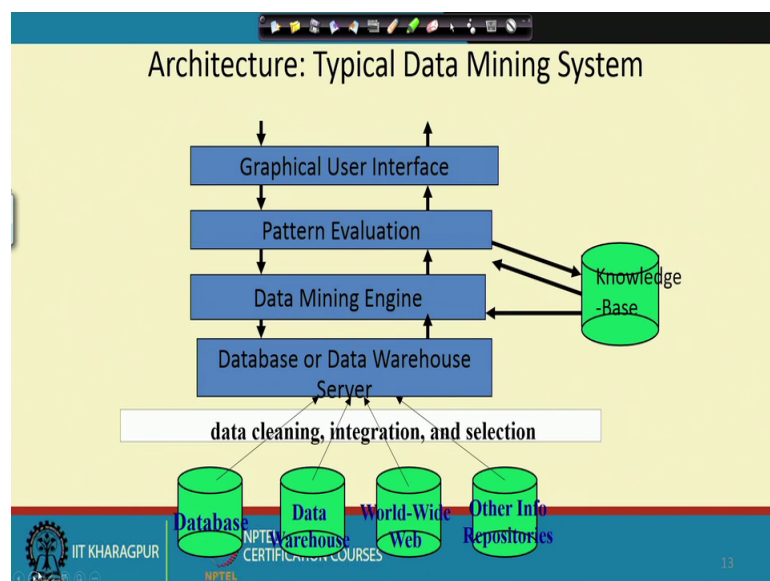
In this entire process another thing I would like to mention that this process of preprocessing selection model, formation, visualization, this is not a one way this is a iterative process one often you do a knowledge discovery, you come back you find that

this data is not enough you select a separate data and again repeat the entire process. So, it is a iterative process. This entire set of steps whatever methodology you propose there are some issues you have to take care. First thing is that efficiency performance scalability does it work on large data, heterogeneity does it work on heterogeneous data.

Human in the loop can you handle a presence of a human in the loop? Real life noise and incompleteness can you handle real life effects like noise and incompleteness finally, can you integrate this with your existing business process, these are all the challenges that one needs to take care. There are various applications, there are user interactions there are social impacts that one need to take care.

So, there are various other aspects also like privacy whenever you do data mining, I give all my credit card data and somebody minds it and finds out something harmful against me that is not allowed. So, you have to take care of this ethical and privacy issues, while you do your data mining the societal issues.

(Refer Slide Time: 20:52)



So, in summary this would be a structure of a typical data mining system, you would have a data base in the bottom most layer.

Then we have integration selection cleaning, then you have a data warehouse, then you have a data mining engine, which would find out the patterns of the model and finally,

you evaluate the patterns and visualize them through an user interface and generate more knowledge and store it an knowledge base for further action.

So, this is the overall scheme, in from the next lecture onward I will delve deeper into each of the steps of this knowledge discovery process. So, if you see the slides you I will I will first go through different layers. First I will go through the data integration preprocessing and cleaning process, then I will go to warehousing, then I will spend a significant amount of time in the data mining engine where I describe the tasks that I had mentioned classification and clustering algorithms for doing them, finally I will touch upon visualization. And evaluation and graphical interface, and in the end of the course I will touch upon how to use this knowledge by means of a case study I will illustrate it to into actual business or scientific use.

So, thank you for today's lecture for your attention, I hope I provided an overview; if you have any doubts or questions about this process about the different steps of this process. Of course, this is just the concepts the actual implementation will become clear, when I discuss the case studies and also of course, before that I will discuss the methodology. So goodbye for today, I will join you in the next lecture starting with data preprocessing and cleaning.

Thank you.