Spatial Informatics
Prof. Soumya K. Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 11
Spatial Database: An Overview

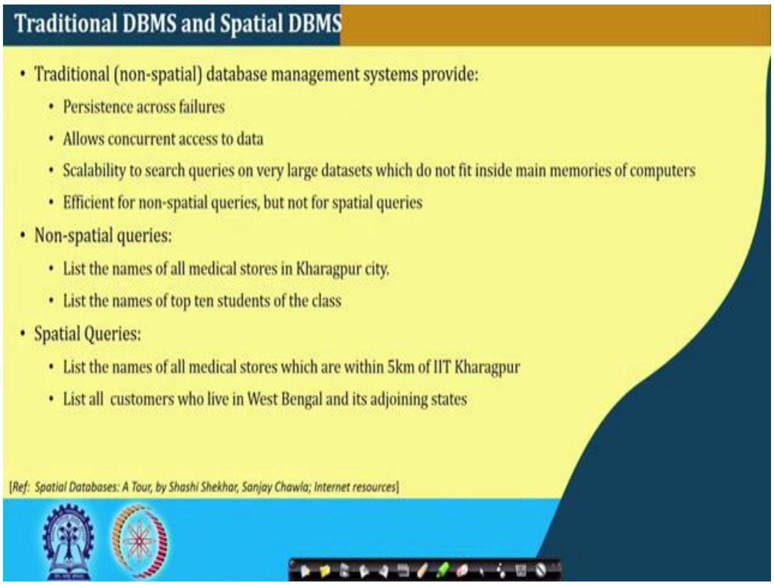
Hi. So, we will be today discussing on aspects of this a special aspect of this database or more of a spatial database right. So, I believe that most of you are accustomed with database management system or DBMS or more specifically the today's relational database management systems.

And we will try to look at a overview of this how this spatial database how it differs from our traditional database systems or non-spatial data handling type of things and why we require a little bit special type of treatment for this. So, that is our bottom line and slowly from this database we will go for the query engines and other type of things. So, that we have the data spatial database of presence into them.

(Refer Slide Time: 01:12)

So, mostly we will be looking at some of these concepts and we most of the things we will be many of the things will be expanding in the in our subsequent lectures.

(Refer Slide Time: 01:21)



So, if you look at the our traditional database management systems so, it is very robust these days right. So, studying from banking to say library erp systems and other etcetera it is works pretty robust. So, it is persistence across failure, allows concurrent access of users right. So, it is scalable to search queries for very large data sets which does not fit in the main memories. So, that is still a challenge, but it can handle, efficient for non-spatial queries not for spatial queries why we will see right.

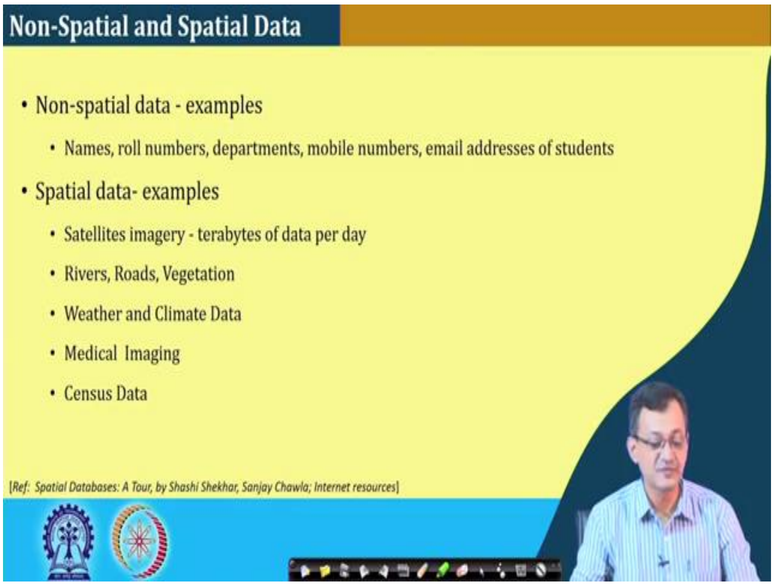
So, this traditionally our database systems or DBMS or relational database systems are mature to a great extent or where it has some sort of a full proof mechanisms right. So, there are non-spatial several type of non-spatial queries like list all students of a department list say all in the medical stores in Kharagpur or list all students of a particular hall or all residential students etcetera, now which are typically non - spatial or what we say traditional queries we are used to it again right. Find out the libraries books in the libraries with this particular author on a particular subject or something like that or a particular publisher etcetera these are there.

But we have to we may look into the some of the spatial queries names of like main all medical stores which are 5 kilometer within the IIT Kharagpur or like we had some of the queries like if this particular river or a drainage is flooded. So, which are the areas likely to be if which are the villages likely to be inundated or partially inundated and type of things right or find out that which are the say something schools which are within say x kilometer 2 kilometer from say state highways or etcetera right.

So, these are the things which has a spatial context like if I want to say that within 5 kilometer of a IIT Kharagpur which are the bookstores or which are the medical halls or medical centers then I have to find out this buffer of 5 kilometer at. So, it has a context on the things I have a 5 Kharagpur and a buffer and then need to overlay find out which are the medical stored in that locations right. So, there is a coordinate systems coming into play right and this you as you understand it is pretty cumbersome to handle with non-spatial or traditional data sets which we are used to right.

So, like list all customer who live in West Bengal and it is adjoining states so, what is the definition is our adjoining state right. So, I can say adjoining states are the states which touches West Bengal, I can say adjoining states are the states whose boundary is within so much kilometer of that. So, it is something which is has some geometry some coordinate systems into play right. So, these are spatial things and that handling these with our non - spatial database systems is a challenge.

(Refer Slide Time: 04:49)



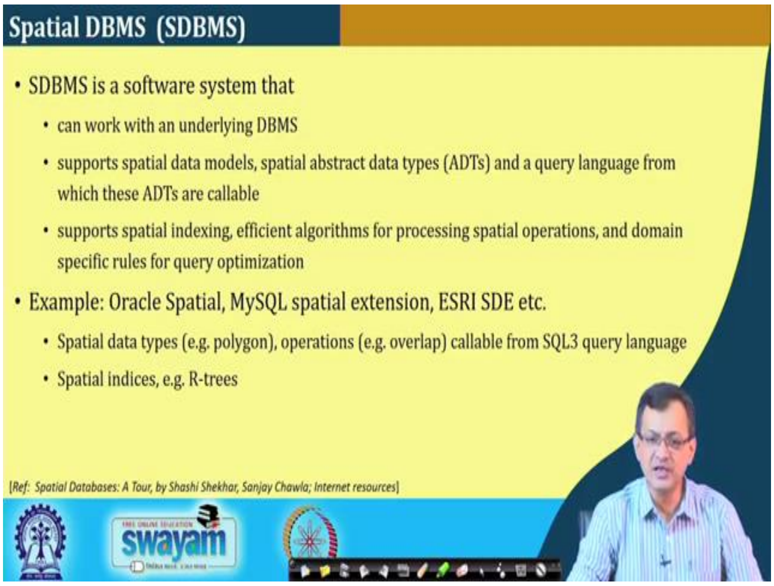
So, example of non - spatial like names of students name, roll number or department, etcetera etcetera or for a library book name, type of a say authors or a authors name, publisher name, etcetera which are accession number these are all spatial non-spatial data sets right. Or even any anything for that matter like if you look at the banks there are all attributes etcetera.

Whereas, if we look at the spatial data to be stored like satellite imagery right say images taken by these satellites right or even airborne anything or a drone etcetera, how do I store that, how do I search that right, how do I query on that. That is a challenge even road, river, vegetation cover, these are all things like road is a linear thing, this polyline river may be a polyline or a polygon, vegetation cover may be a polygon weather and climate data right.

It is we come with a spatial context right. Even if we leave the if we go little broader from the our geospatial concept to a basic spatial like considering a human body as a space then medical imaging type of things are there. Even census data that is very typical right census data like if say population, map or a different demographic data these are not seen from the space right, but they are related to something right, I say the population of India is so much, population of Kharagpur is so much, population of say West Bengal is so much or population of say Bihar is so much right population of a particular mouza etcetera.

So; that means, this data is a context or I say literacy level of this is this or this much population is affected by these and other type of things right. They are so much is above this edge and type of things of this particular region. Whenever you come to a regional location etcetera then we have a context into the plane. So, in that sense the census though as such may be a non-spatial or traditional data set, but connecting with a some coordinate systems we can a spatial context.

(Refer Slide Time: 07:09)



So, spatial depth database management system or DBMS or SDBMS if you in short if you want to tell so, can work with underlying a DBMS right. So, DBMS is a standard right, it has grown over years, proven over years and then it is a underlying DBMS supports spatial data models right. We have seen the spatial data model this field supports a data model, spatial abstract data types, query language from which these data types can be callable all those things right. So, I require a database spatial database which understands spatial data models and spatial ADTs or abstract data types and etcetera, like it should understand what is point polygon etcetera, it should understand topological relationships.

Now, if I am query that whether this touches this overlap this these type of things right supports spatial indexing right this is important we will see in subsequent things spatial indexing, if you see an algorithm for processing spatial operations right and domain specific rules for query optimizations right. So, that we will see when we look at the things, but it should have a spatial indexing right like there is a inherent locational property of the things right.

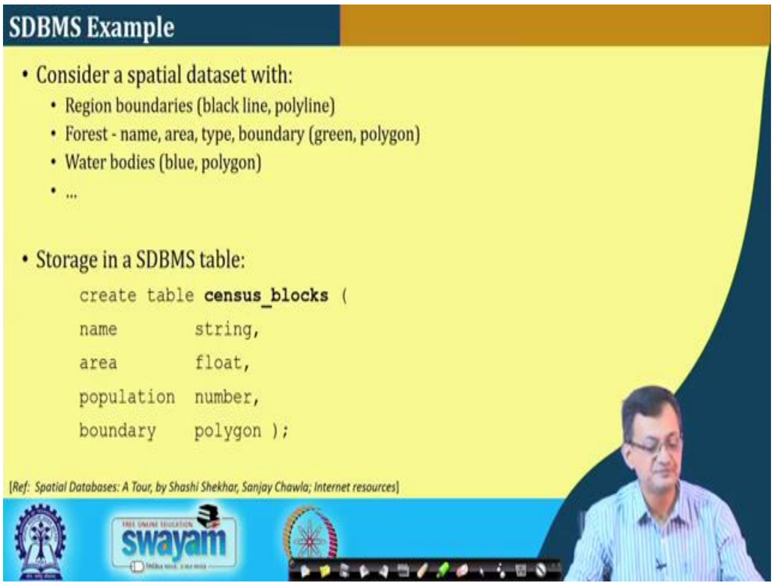
Like we say that nearby objects are likely to be more close well whose then distance object, like I say temperature of this particular outside this particular building is likely to be same or nearby then it is a subsequent building. Then this temperature of IIT

Kharagpur a particular outside on some things that and will be same as Kolkata or Darjeeling or Shimla and type of things right.

So, the more distance I more nearer things, now when I search type of things whether it is necessary that this type of which are in the same cluster the data should be in the same place, but there is no as such this type of things in case of a; in case of a non-spatial data like on this spatial context right, different tuples are not that way related they may have a better relationship, but that way not related that nearness and type of things right.

So, there are several examples like we have different database players practice starting from open source and etcetera like MySQL to postgre to Oracle spatial and there are specific tools like ESRI SDE etcetera that specific proprietary things which are able to capture the spatial context. Like spatial data types like polygon, poly line point, spatial operations like overlap, join, touches, intersects and etcetera and those are callable by SQL or SQL 3 and there are spatial indexing strategies we will see that R- trees and type of things right.

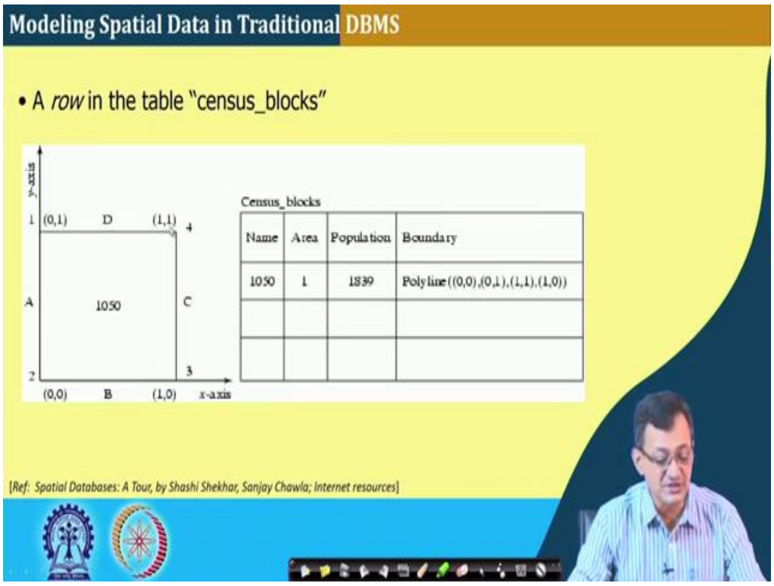
(Refer Slide Time: 10:12)



So, I can have spatial data sets like region boundaries black anyway that figure is not there so, forget about that within the bracket black line and all those things. So, region boundaries, forest name and boundaries and water bodies etcetera so those things are there. Storage may be like looking at a census block we if you see that Shashi Shekhar's book.

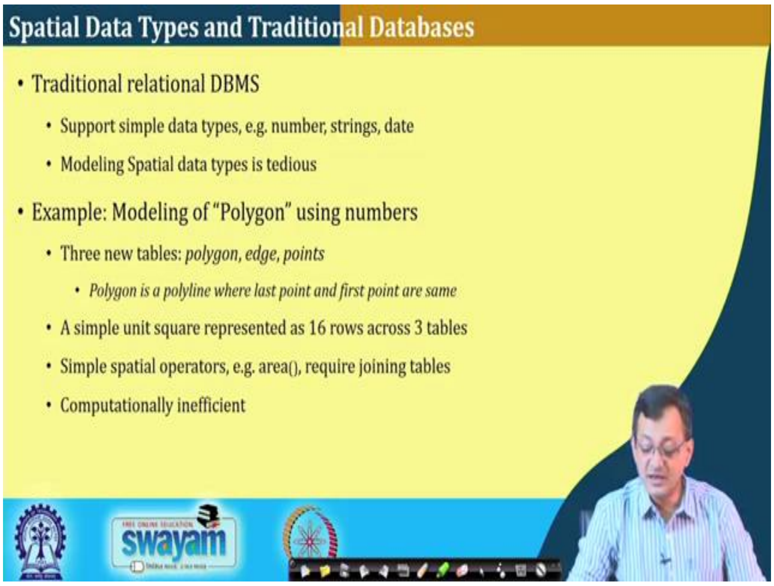
So, has a something like you see it has a name, string, float, number and polygon see I consider a river. So, river may be having a name, a particular some other properties and then has a feature like polyline right I say a IIT Kharagpur region right. So, it has some of the non-spatial attributes and some spatial context right, that means, whether it is a like in this case boundary is a polygon all right a road may be a polyline and type of things.

(Refer Slide Time: 11:15)



Now, how do I represent this, like if we consider a square block in this case is interest block, then one is that I give a number area population and I define the poly as a polyline like this. Now my polyline if the source and destination as same we say it is a polygon so, if the source and destination touches each other. So, this is one way I can represent. So, we can add we can argue that whole thing I could have done in a standard database also right, like I some way I represent that these are the points they are in that particular polygon and type of things, it could have been nice that if the polyline is defined or polygon is defined in a direct sense.

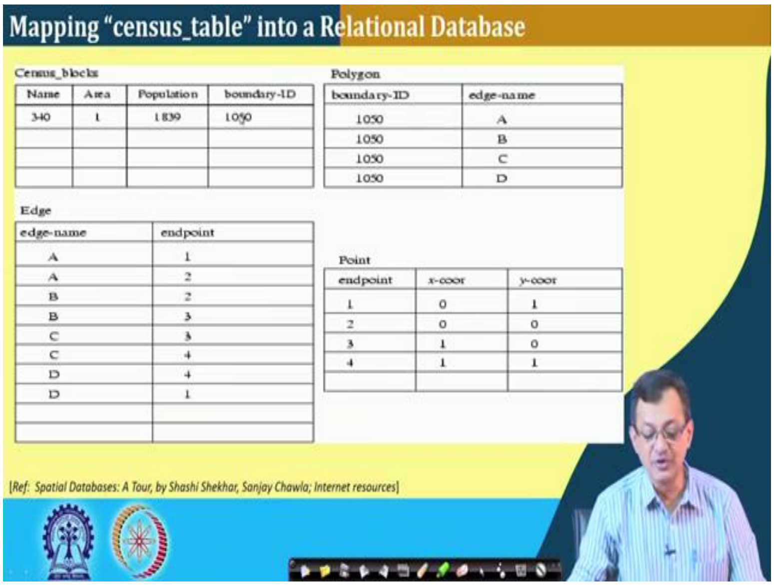
(Refer Slide Time: 11:53)



Now, traditional relational database has supports simple data types number, strings, dates and there may be a little bit of complex data types. Modeling spatial data types is a very very tedious on those things right. It becomes more a tedious if like that like making a polygon and I have a number of poly line, from the polyline every has to be defined by different points which connects to the lines and it has become a tedious things.

So, modeling of polygon using numbers so, polygon or edge or polyline and points so, polygon is a polyline where last points and first point or starting point and ending points are same. A simple unit square represented by 16 rows across 3 tables what we will see, operations like area requires joining of these tables and it is computationally intensive.

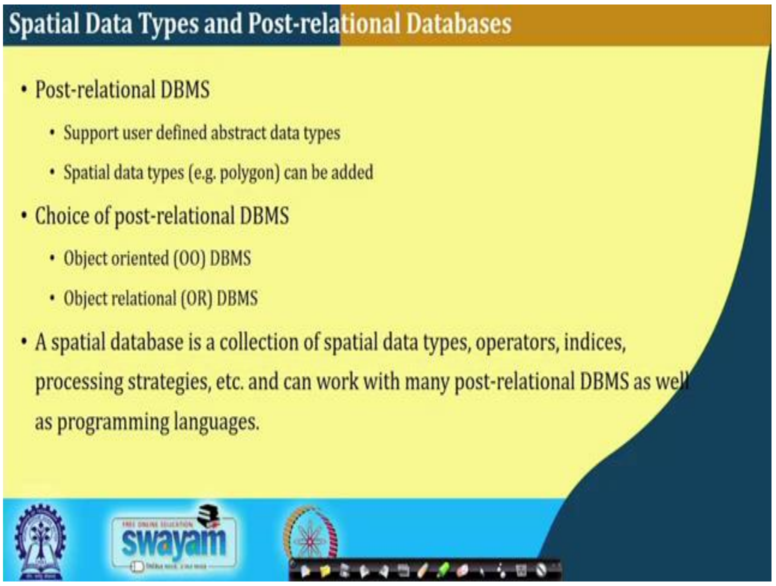
(Refer Slide Time: 12:47)



Like as we have seen so, some polygon so it has a boundary, that boundary has to be written by the edges the edges has some start point end point every these points has a coordinate. Now, see a single unit square I require so many things, now if I have some operation like finding the area of the things now I require joining these tables and doing calculation and type of things right.

So, it is possible by traditional data set to model this, but whenever I want this type of query which are like area of a particular region or length of a things I require so many tables to be joined etcetera. So, I require a more a smarter solution, underlining DBMS is there over that I require smarter solutions which will give me quicker results.

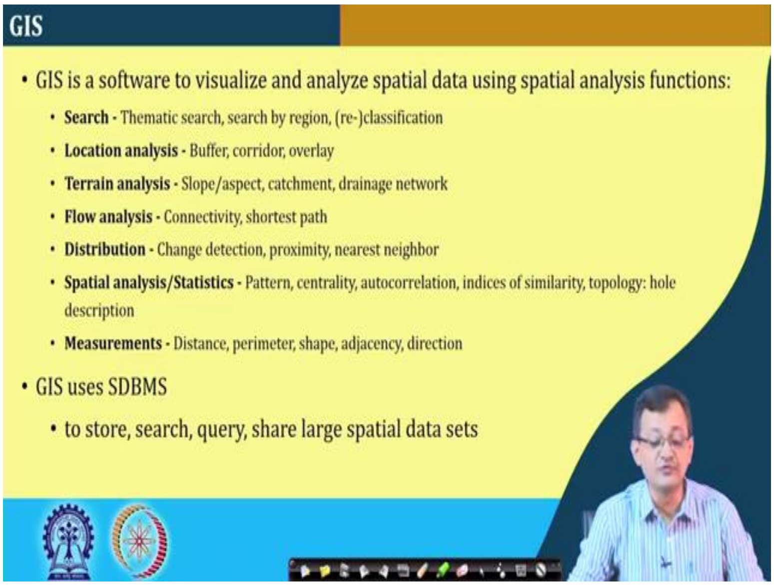
(Refer Slide Time: 13:40)



So, we have some of the things who are post relational databases like support user defined abstract data types right, we have that user defined abstract data types spatial data a polygon can be added to that. So, I can say the user defined data type is a polygon and the definition of the polygons is come into play. So, choice of there are several choices of post relational database we can have. So, a spatial database is a collection of the spatial data types, spatial operation, spatial indices, processing strategies of this and anything on the spatial and that can work with many post relational DBMS as well as programming languages right.

I can have different programming languages like java etcetera etcetera which can work on the things right. So, first of all it can handhold with the spatial data types, operators, indices and the processing strategies, it may have different type of processing strategies for the things we will see that when we look at the query processing etcetera right. We can have different processing strategies like how do I calculate area, how do I calculate intersections. So, there should be a strategy some algorithm should be there how do I find shortest path right on a network of roads or network of transportation network, how do I find a shortest path so, there should be strategies so that is important ok.

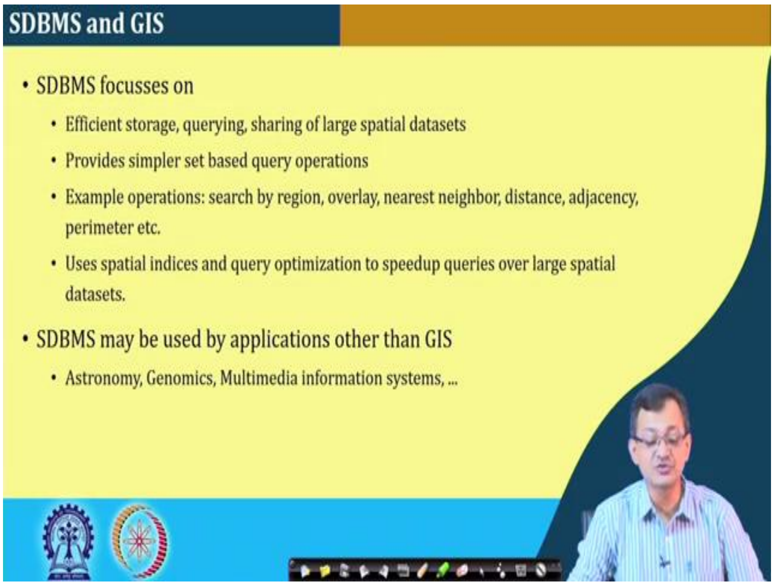
(Refer Slide Time: 15:00)



So, GIS is a software to visualize and analyze spatial data using spatial analysis functions right. So, another part we will come to that in a little detail. So, there is a geographical information system there are different variation of the acronym people sometimes think it is geographic information sciences and several other things, but primarily it is a software tool sometimes, why we are mentioning this sometime we will try to mix up with these GIS with the spatial database SDBMS or spatial database management system, but they are not exactly same.

GIS uses spatial database management system, but GIS is a much higher level right. So, it is a software to visualize analyze spatial data using spatial analysis functions like search, locational analysis, terrain analysis, flow analysis, or distribution, spatial analysis and statistics measurements or several type of measurement. So, these are the properties of GIS. So, it uses the underlying SDMS, but they are not the synonyms like I cannot say GIS is this and type of things right. So, it is uses, spatial database management since store, search, query, share large spatial datasets and so on and so forth.

(Refer Slide Time: 16:17)

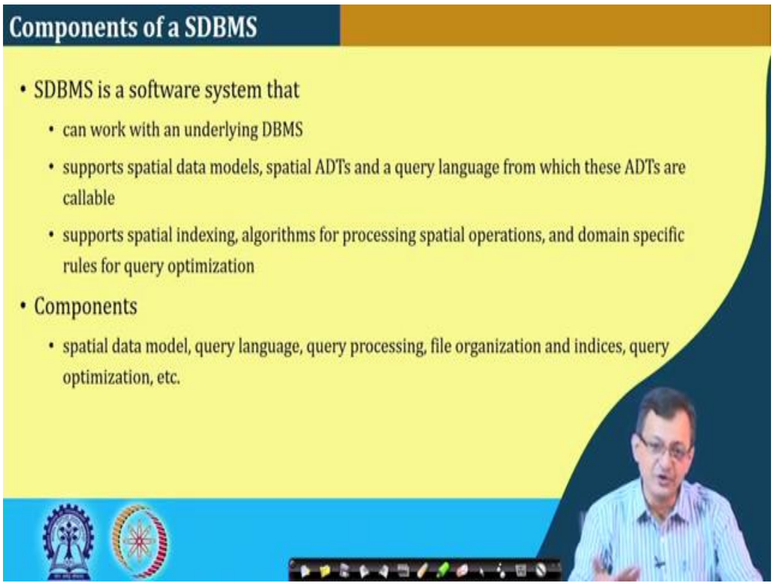


So, SDBMS focusses on efficient storage, querying, sharing of large spatial datasets, typically this spatial data set is not only have this type of geometry property they are usually pretty large right. So, handling this data is a big challenge provides simply simpler set based operations like search by region, overlay, nearest neighbor, distance, adjacency, perimeter etcetera right.

So, if you look at the spatial context if you want to find out how much near I am to even I have to find out that from this to this I may not have the whole data, but I or that last mile connectivity and find out the nearest things which are there right or within 1 kilometer of the road, within 5 kilometer of the river, within 10 say 2 kilometer radius of a factory and there are different that pollution will be so much human I say I can say that this is a chemical factory and within 10 kilometer there should not be any population or human habitation. If it is there then we need to trigger medical checkup and other things for them on a regular basis that may be a regulation of the government right.

So, how to do this 2 kilometer overlay of the region finding out from the population match who are the effected who are likely to be affected by these particular factory or any type of things right. So, there are different operations which to do.

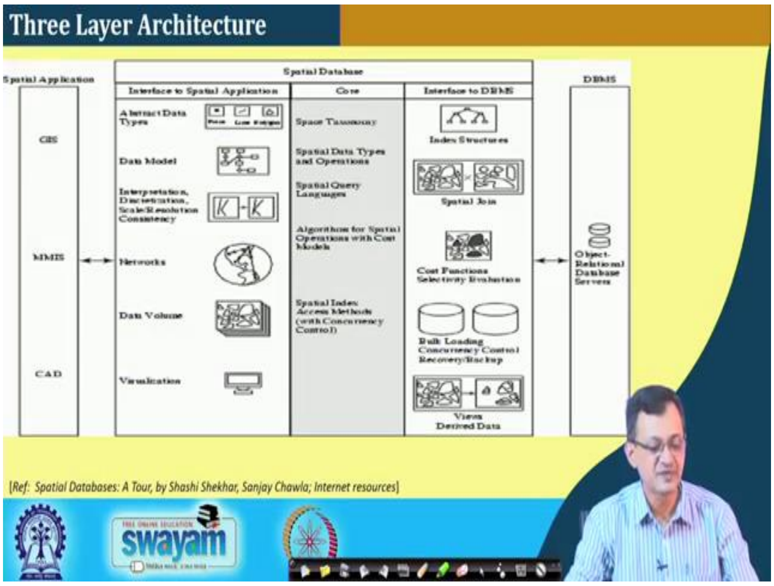
(Refer Slide Time: 17:46)



So, it is a software system if we recap again it can work with underlining DBMS supports spatial data models, spatial ADTs, query language, supports indexing algorithm etcetera already we did just to recap so, there are several components right. So, spatial data models, query language, query processing, file organization indices, query optimization etcetera these are the different components.

If you try to relate with that if for in for our normal databases also different components are there, but there is a spatial context into it here. So, need to be dealt in a little more span I should say smarter way on a or more in a spatial way quote I will quote spatial way into the thing.

(Refer Slide Time: 18:29)

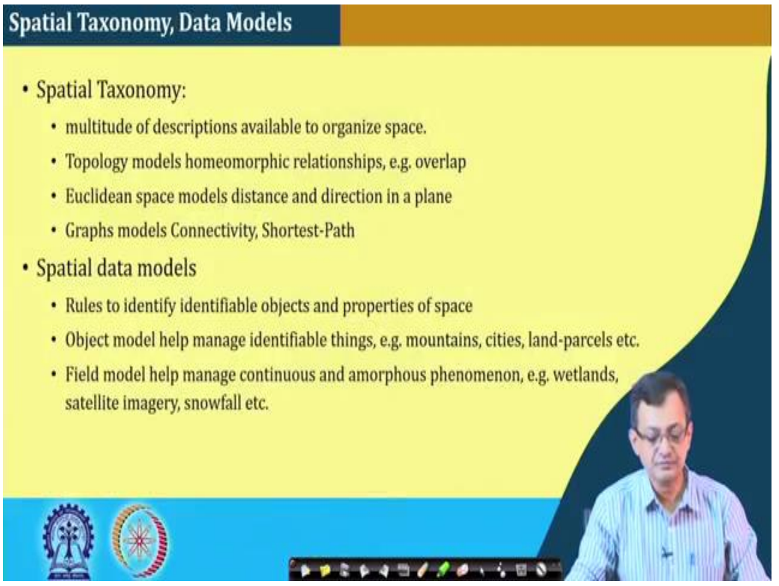


So, this figure is if we look at it so, they again 3 layer architecture. So, it is broadly no need of immediately break your head over this, one side is this different type of elegant GIS, MMIS, CAD which takes this input from the things and other side is the basic database management systems right, so, underlining database because this is a working thing.

Here we have the spatial things in the it is core is spatial taxonomy, spatial data types, spatial query language these are the core one side, we have the abstract data types, data models, interpretation, networks, data volumes, visualization, other side we have the interface with the data database how do I. So, there is a index structure has to be there, there are issues of spatial join whether the same as that our normal join right, how do I spatially join 2 polygon joining and 2 a student table or a department table student table and a hall table joining are not the same thing right.

So, it is a 2 spatial data joining how they are so, spatial joining. Cost function of selectivity evaluation when we do some selection what should be the cost function views of derived data sets. So, bulk uploading, contrasting, control etcetera all this. So, there are one is that interfacing with the database, one is interfacing with the spatial data analysis tools and another is the core right. So, this is the whole stuff which comes into play.

(Refer Slide Time: 20:06)



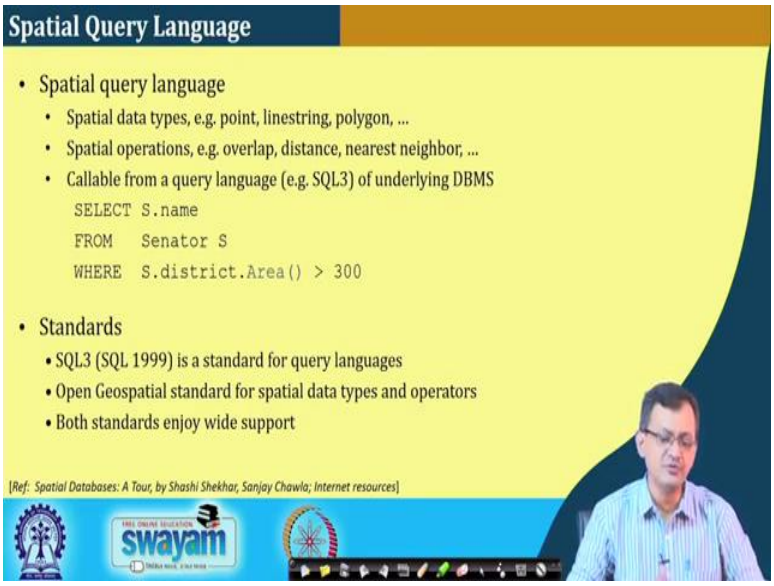
So, what we require is if we look at that there are spatial taxonomy and models right. So, what do you mean by spatial taxonomy, their multitude of description available to organize in space that how it organizes in the space. So, topological model homeomorphic relationships like overlap and we have seen the topological models and how they behave. Euclidean space model like distance some of the Euclidean space model like distance direction in a plane right. Graph model for connectivity, shortest path and type of things right see.

Like see if you look at this overlap and other type of things or adjacency connect touching each other etcetera that is one context. Other there are different graph models right when we try to look at say shortest path all right. So, that is not the crow find path right. So, I like if you I have a building here a building at the other side in between there is a wall right. So, the distance to between this Euclidean distance to building maybe 200 meters right, but if I want to find shortest path I cannot do scale this wall, I have to go to this through this path etcetera which may be 700 meters right.

So, though this is other way near, but when we consider this road network to finding the travel time etcetera the it is much far maybe the a something which is 300 meters away on the road is nearer to these things right. So, that in other sense the spatial networks need to be dealt in a different fashion right. So, this so, that is why we will see here also that we need to we will look at these spatial network operations in it in a different vessel.

Then we have that spatial data model rules to identify identifiable objects and properties of space right. Object model may manage identifiable things like mountains, cities, land parcel etcetera. So, field model if you recollect we discussed some of this object model and field model in our previous some previous lecture. Field model helps to manage continuous and amorphous phenomena, like wetlands, satellite imaginary so, where the continuous changes are there and type of things right. So, we need to what we mean to say these are the things need to be accommodated by the spatial database management system right so, that we can work with the things.

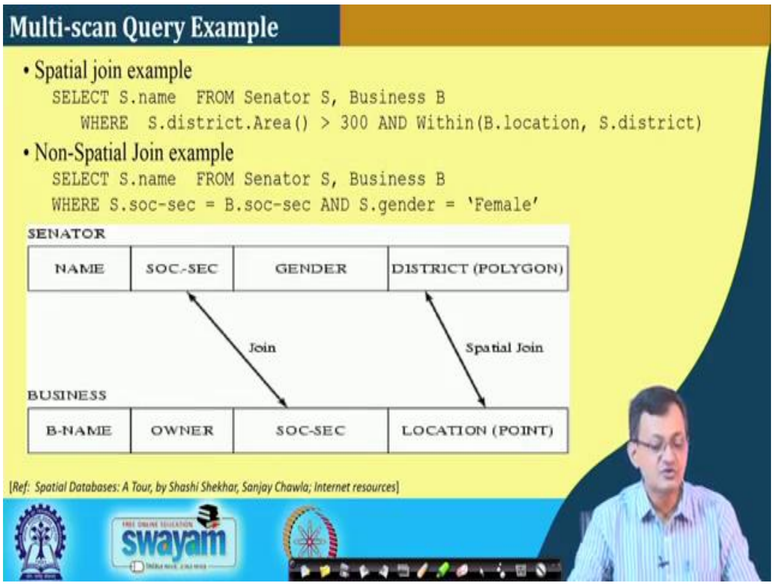
(Refer Slide Time: 22:56)



One aspects of the things we will be dealing this in much bigger details is the spatial query language all right like point, linestring, polygon or point polyline polygon not different data types, spatial operation overlap, distance, nearest neighborhood are the different operations. So, callable from a query language SQL 3 underlining DBMS, SELECT S dot name, FROM Senator say particular senator or administrator A is WHERE district name area is greater than 300 etcetera that may be a query. Now, you see here the last line is doing some operations like a particular distance more than some 300 units only that I require the district administrator name in or something some MLA or something maybe in this case senator etcetera.

Standards so, there are several standards which also evolved SQL 3 is a standard for query language, open geospatial standards or for spatial data types and operation, otherwise there will be problem of having a standard set of the things. So, there they both the standards enjoy wide support, both from industry, both from different federal or government agencies which follow these standards.

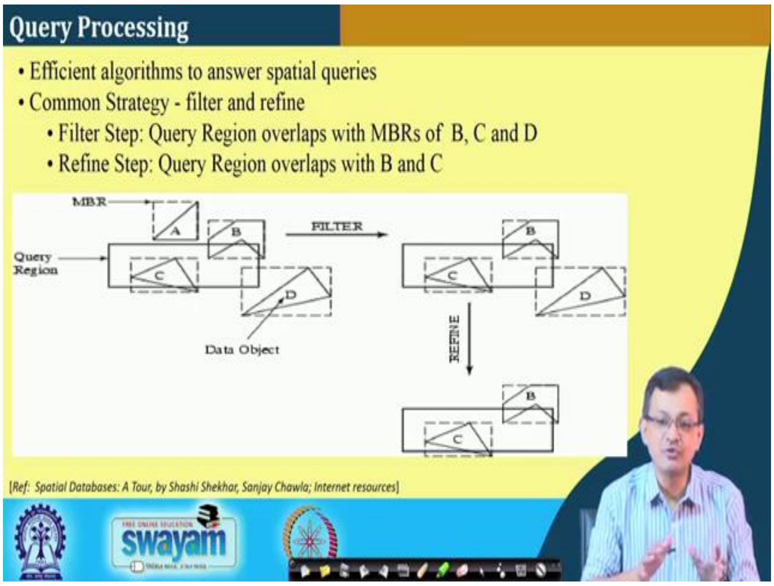
(Refer Slide Time: 24:10)



So, like that query is SELECT some S name FROM Senator, etcetera, WHERE the Area is 300 Within B location and S district right. So, see these functions area within these are typically not there in our standard traditional database systems. So, these need to be accommodated, like non-spatial join these senator where something social security code and gender female these are non-spatial things.

Now, if we want to join I have to join like this so, district polygon with point location the spatial join has to be there this is a normal join. So, wherever the spatial context is there I need to go for a spatial type of join. So, how does how it looks like right.

(Refer Slide Time: 24:53)



So, we will come to that just to just what you are trying to do, we are trying to open up this scenarios where these things are make sense or make important. So, which try to impress upon that there is a need to look at the spatial database management system in a different way, not as it is something more than our traditional data base management systems.

So, whatever the traditional things are there so, that is there, but we need to look into the things. So, like if you say we need to feel the efficient algorithm to spatial queries right. So, there are two things; one is to filter the things right, another is refinement right like say for example, I want to have a overlap of the things right. So, we will see that what is the MBR? MBR is the Maximum Bounding Rectangle suppose A is like this. So, what is the maximum bounding rectangle? This is the rectangle right. So, B is like this, this bounding rectangle is this one, see C is like this and the bounding rectangle is like this right.

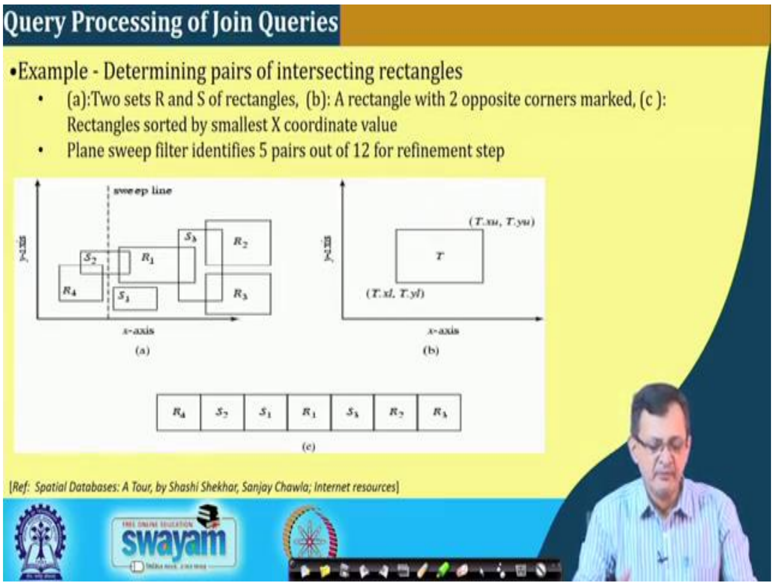
So, if my query region is something like this, I want to find out that where is the query region etcetera, now these are polygon ok. So, this querying all these that I what I have to do that, if I want to find out whether which are the regions which falls within these things A, B, C, D whether they are fall. So, first of all what I have to do, I need to basically in a very crude method every coordinate I have to find out whether this is my query region or region of interest whether they falls into the things right and if it is an irregular shape it is a very very difficult or tedious job right.

I can do a filter step, I say A maximum bounding rectangle or MBR or some or a MOBR sometimes we call so, that is rectangle. So, I for every region A I find out the rectangle, B I found out the rectangle, C I find out the rectangle, D I find out. If my query region is not overlapping with these rectangles right query with their MBRs then it will not overlap with the particular object also, other way it is not true all right.

So, if it is overlapping with the MBR that does not mean that it will overlap with the actual object also, but nevertheless in doing so, I have I could bring down the search space right. In this case I can easily eliminate A, A is nowhere in the things because A is MBR is not overlapping and MBR is a rectangle. So, comparing 2 rectangles or multiple rectangle is much easier than doing a irregular polygon shape right. So, this is the filtering step I eliminate B right still this things are there, but what we can see that finally, in the refined state I can eliminate C.

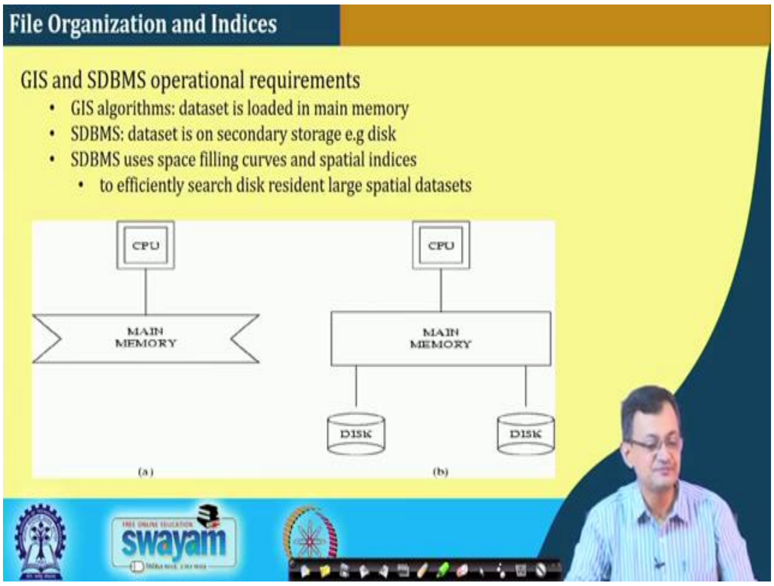
So, how we can do go about this so, is a another challenge, but nevertheless my first one is a there filtering bring down the search space and second one is go for a refined main stage or more detailed thing. So, that I actually find out the things right so that is the way of looking at it.

(Refer Slide Time: 28:34)



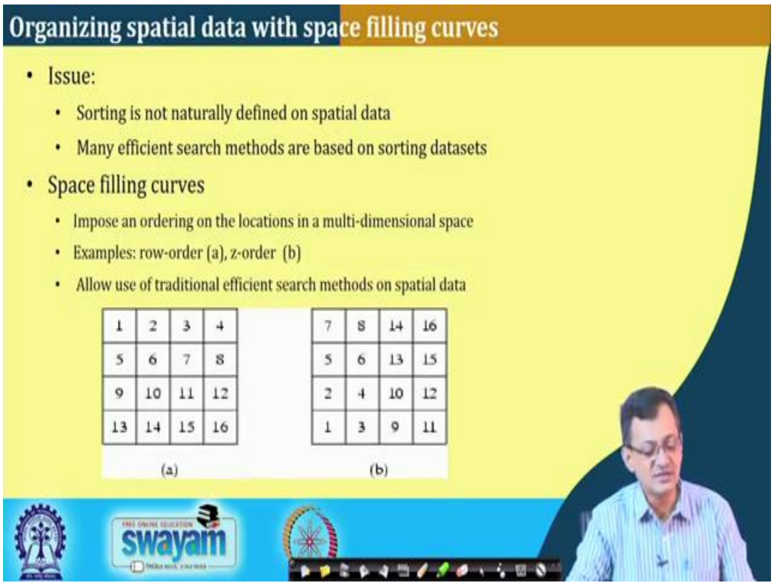
And there is also the query processing of spatial of join queries, when there is a join like this then there should be a way of things to left, one typically I have a sweep line right. So, the sweep line says that where if it is passing through that it is cutting the things then I say there is a possibility that there will be a likely there may be a chance of this 2 rectangle overlapping there is a question of join right. So, I can have horizontal vertical sweep and type of things and then can find out which are the potential candidates where who are likely to be the participant in the join operations. So, there is this is another challenge.

(Refer Slide Time: 29:21)



File organization is a another major challenge right. So, GIS and SDBMS operations requires GIS algorithms, dataset is loaded in the main memory. So, GIS algorithm typically no one that to be main memory, SDBMS data set is on secondary storage in the disk. SDBMS uses space filling curves or spatial and spatial indices to efficiently search the disk resident large datasets. So, it is a space filling curve it is not like row wise etcetera I may have a space filling curve. That means, nearby objects should have a conscience consecutive sequence number; that means, this should be localized in the disk space right.

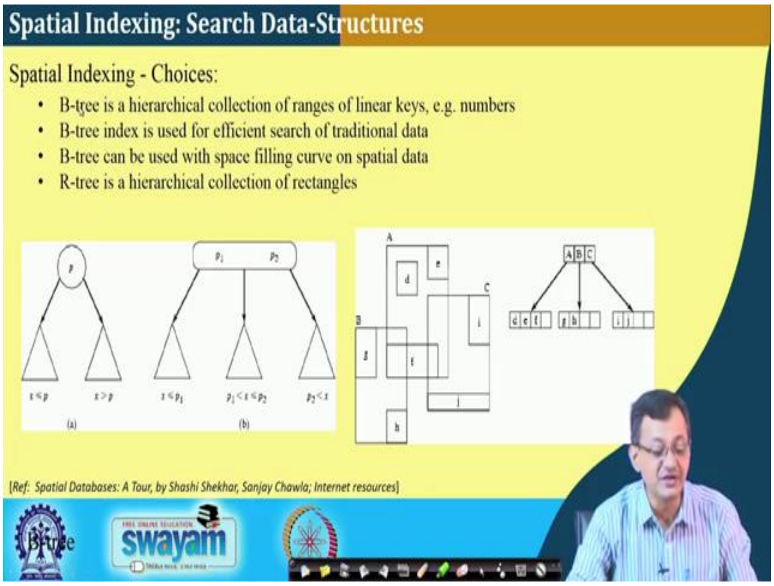
(Refer Slide Time: 30:06)



So, there are several strategies we will look into the things one is this is a row order, but this is a z order right, 1 2 3 1 2 3 4 5 6 7 8. So, what you 5 6 7 8, what you see that these are localized; that means, these instead of this in a continuous line. So, in the discard will be there 1 2 3 4 5 6 7 8 like this right, but see what is likely to be there these are likely to be nearer, but if it is in the row filling in this case after fourth it will come right. But if I can have this type of space filling then I can have this data much quicker than on a elongated space right. So, this is one of the things we will discuss.

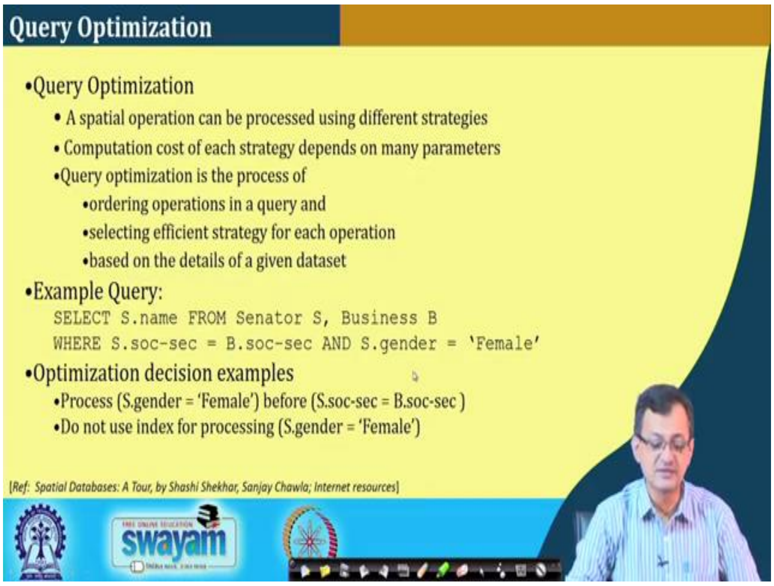
So, sorting is not naturally defined on spatial data, I cannot sort on the spatial data naturally right which is there in the traditional dataset. I can sort on something student name, student marks or etcetera etcetera some sorting algorithm, but how do I sort on the polygons right. So, there should be some way to look at right, whether there is a nearness in the things or sorting the polylines, poly points etcetera what should be the basis of the things, many efficient search methods are based on sorting data sets right whatever the search etcetera that is basically based on a traditionally right. So, this is has to be there some of the ordering stage.

(Refer Slide Time: 31:30)



There is another that indexing scheme like, B-tree an hierarchy collection is linear key is a very popular B plus trees are very popular, whether that is also applicable here. So, there is a concept of R-tree in a hierarchical collection of rectangles right. So, like this I can represent as a R-tree, R plus tree and type of things. So, indexing is also a challenge.

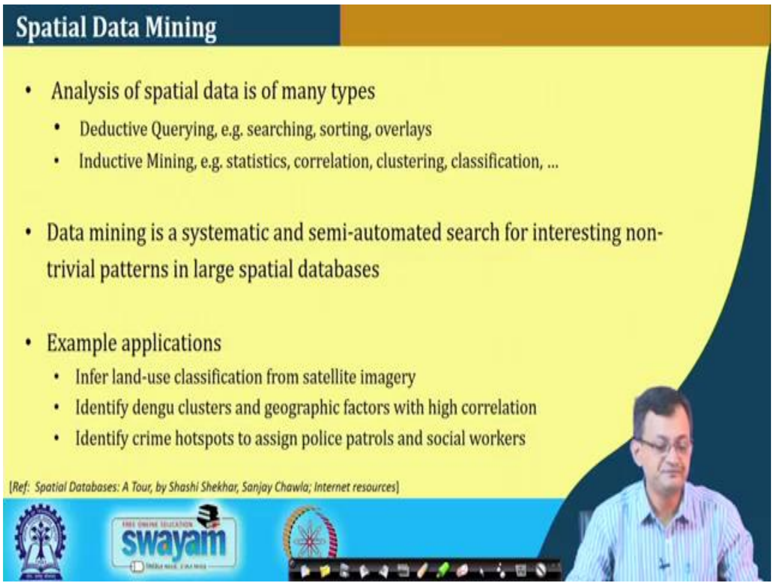
(Refer Slide Time: 31:53)



Then finally, come as the query optimizations right, like if I say that say which operation I should do first right. So, if I have a number of select project and other spatial operations whether that I can do what should be my sequence or what should my ex query execution tree that is important, but that brings down the data load that. This is important for our traditional data or traditional systems or database systems also where which need to be there, like whether I should check this one first or this one first to be so process gender 'female' before this may give me a much faster things, because it may be more the data set will reduce and type of things.

And here the things are there if I have; if I have spatial operations then whether I will be the non-spatial operations into much lower into the things right so, we will see that. So, this query optimization finding efficient query execution trees are important.

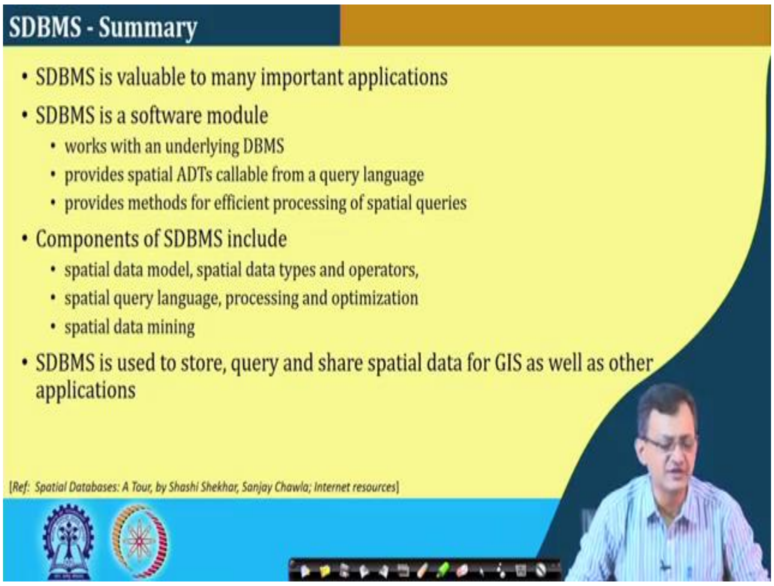
(Refer Slide Time: 32:59)



And another aspect is what I why I want to do means what next right query is one of the things, another is that I want to do some sort of a data mining right spatial data mining into the things, analysis of spatial data of many types, deductive queries, inductive mining, statistical, correlation, clustering, classifications right. So, data mining is a systematic and semi-automated search for interesting non-trivial pattern in the large data sets right.

For spatial dynamic spatial data sets right, how do I mine those things what are the spatial mining systems like, which are the correlated operation co occurred operations, co occurred situation or events I need to mine and find out. So, these are non trivial, if it is trivial then I do not need all those things like, identifying dengu cluster and geographic factors with a high correlation in a particular region right or crime hotspot to assign police patrol and social workers or something like that right. So, what are the crimes hotspot in a city or a region. So, that because my police resources or the security resources which are much constant not that is can be resources.

(Refer Slide Time: 34:10)



So, what we look at in look at a overview of the things is a valuable to many important applications with the spatial context works with underlining database management, systems provide spatial ADTs which are callable by spatial queries, provides method for efficient processing of spatial queries. And there are several components like spatial data models, spatial data types operators, spatial query language, processing, optimization, data mining and there are several I can go we can go on adding these things.

So, SDBMS is used to store, query, share spatial data for applications like GIS right, as well as several other applications. So, what we tried to today to have a go over overview of the whole how what is this spatial database management system. So, that is this builds our foundation for several other things when you go for spatial queries, optimizations, spatial networks etcetera right. So, that we will cover in our subsequent lectures during this course of during this lecture series.

Thank you.