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

Lecture - 28
Spatial Analysis - III

Hi, so, we will continue our discussion on Spatial Informatics, we are discussing on Spatial Analysis. In this particular analysis, what we mean to say that we will be discussing on primarily on spatial data mining and other related areas, right; that we usually see in coming one or two lectures.

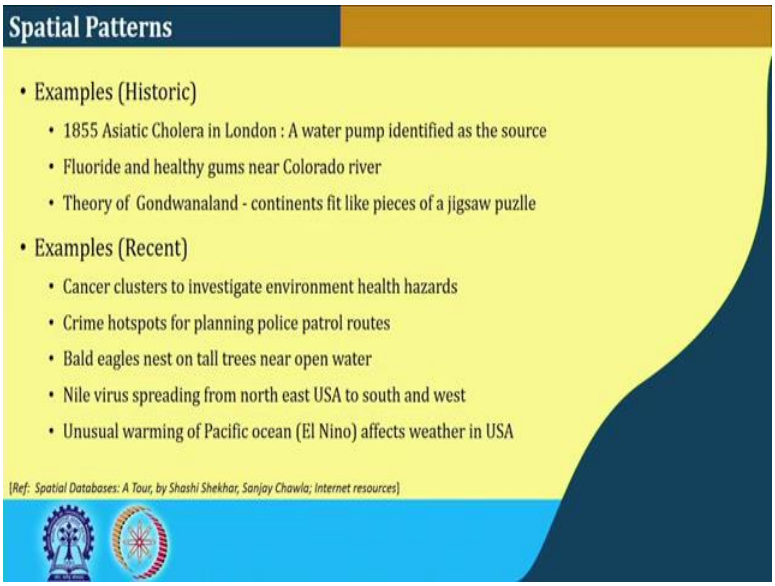
So, as we have seen in the last two lectures, there is more of a overview of what is data warehousing, data mining, how data mining is different from our standard query processing or standard OLTP type of things. So, today we will be more, look more focus on the spatial data mining, right? So, that it has all the properties or feature set of a generic data mining.

Along with that, it, it, also have some other properties or should say that that spatial context that we will take up in this today's talk and our subsequent talks, right? So, why it is little bit different from our traditional non-spatial things, right?

(Refer Slide Time: 01:31)

So, mostly we start with spatial data mining.

(Refer Slide Time: 01:33)



So, what we do when we try to mine data, we primarily are looking for patterns, right? Unless a specific queries, in case of standard OLTP or SQL type of things, here also where we are looking for a some pattern, which may help us in prediction, forecasting, or finding some anomaly or something like that. Like, overall doing some analysis or something some knowledge discovery from the data set, right?

And, but we should keep in mind, like I told you in previous lectures also, that nevertheless our traditional data set this query systems lies, because that is the source of the data, right? So, that is the source of the data what we extract from the things, right? And then we build up some, say model, maybe some functional model or what we see some of the things like association rule and type of things which helps us in finding out that whether there is a inherent pattern into the things, right?

Like, as we discussed in non-spatial, like whether there is a pattern of purchase of items in a departmental store on a particular day of a week or on weekdays, weekends, or on a particular season or it varies over the region. So, these are, these want to look at the patterns. If there is a inherent pattern then I want to do my overall mechanism, in case of departmental store it may be the business goal is the maximize profit or maximize revenue income.

Then I want to place in such a way that those things in such a way that will, like as there is a common thing what we say that, that bread-butter, right? If I buy bread I am likely to

butter buy butter, right? So, the bread butter racks or the storage in a departmental store should be nearby, right? So, it go on picking up the things. So, these are closely associated, right? It is unlikely that the person who is buying bread will buying buy shoes at the same time, it is no as such association per se, right?

So, keeping those two racks together may not help in maximizing the, filling up the what we say that market box or the cart of the our customer. So, that is why in a spatial context where there is a, whether there is a relationship between thing. So, some of the things which we have seen that some of the example, historic examples, which are very popular in literature; one is 9 1855 Asiatic Cholera in London.

So, it a water pump identify as a source, right? So, there was a outbreak of cholera in London and it was very very difficult for them where, where, it is coming from, right? The search space is pretty large it may come from waterborne, it may come from some other type of things in environmental or any, any, such type of epidemic there may be multiple sources. So, there are difference search space we want to boil down to the things, right? Even if it is waterborne, which part of the water is causing problem is a problem.

So, what they did consider one of the, one of the very earliest application of spatial analysis is that they plotted one, so, some geographer plotted these where are the cholera hotspots are there and they could find that they are concentrating more, originating more, on a place and then could be able to find out there is a water pump which is creating the other. That water is contaminated they would examine and say, because of maybe connecting with some contaminated water or sewage line etcetera and things are there.

Now, that is the, now doing this analysis can boil to the things, right? So, again that in near Colorado River it is people are having healthy gum and as there are much fluoride in the water. So, fluoride and healthy gum relationship maybe there in doing analysis. Or what we say that continents, continental drift theory or continents fits into another pieces of a jigsaw puzzle maybe a large scale spatio temporal datasets.

So, some of the recent thing like finding a particular academic cluster like cancer cluster to investigate environmental health hazards, that may be one of the things. Crime hotspots for planning police patrol is another major aspects, which are used extensively these days. Because number of police resources or number of security resources, a

particular say, city or particular region have is, may be, may not be sufficient to cover the whole region in a, what we say in a regular or in a giving importance regular giving similar importance to the things, right?

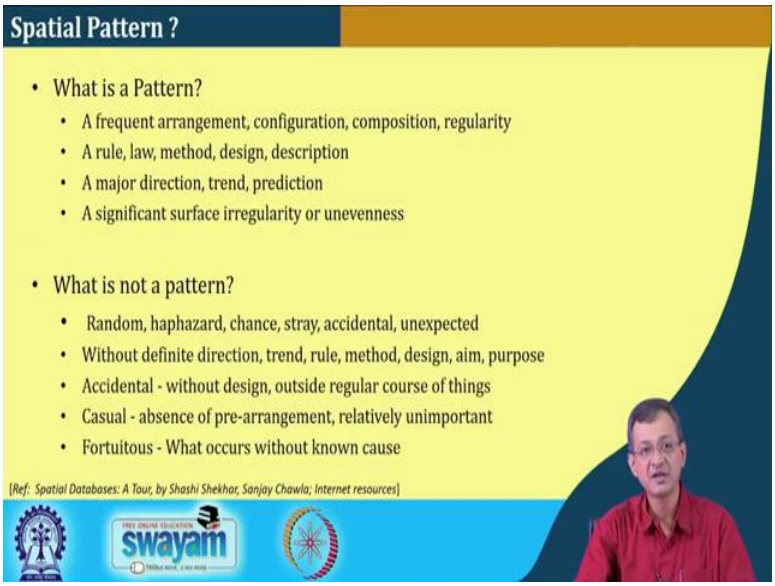
So, there will be, there will always a disparity between the demand to cover the things and the supply of the things, right? But on the other hand, crimes are not happening everywhere, right? So, even if you are having, say, 10 police team the crime maybe 7, but you need to cover a region of 100, right? So, if we, if there is a, if I can have a predictive model that this or I can say these are the crime hotspots and these are the likely to be the next crime which is going to happen in coming, coming days, coming week, coming months, then the patrolling can be rearranged like that, right?

Even a after a crime has occurred what is the possible pass this culprit may take to exit from that place, maybe thing which may help the things to guard the region, right? So, and there are other things like environmental like bald eagles nest on tall trees near open water, that may be a things which is a interesting fact. Nile virus spreading north east USA to southwest, unusual warming of a Pacific Ocean that it El Nino affects weather in USA or weather in our India some part of the things.

So, these are some of the things which are something, events which are correlated, some of them are co-located, or co-evolved, right? If this happens, this happens right on the things; like some of the things may not be very pretty interesting. Like, if I say the temperature of next to, next town to IIT, Kharagpur is very high, so, it is likely that here also temperature will be high, right?

But, if it makes sense that, if the, if there is a temperature fluctuations deep in the south and it has a some pattern that it will be get after 2 days it is gets affected in south of Bengal the same type of feature; if there is a this sort of pattern is there, that is, that may be interesting to look at, right? So, some maybe trivial which are not so called interesting patterns, there maybe pattern but not so called interesting pattern but we are looking for interesting pattern which our data mining or spatial data mining may help.

(Refer Slide Time: 09:10)



So, what is a pattern? A frequent arrangement configuration, composition and with regularity it happens; we can set up a rule, law, method, design, description of the things; a major direction, trend and prediction; a significant surface irregularity or unevenness, right? So these are the things pattern, which is not a pattern, which is purely random, haphazard, chance, stray, accidental, unexpected or without any definite direction. Or it can be accidental a, casual like absence of pre-arranged, relatively unimportant and for what the causes etcetera I cannot investigate or find out.

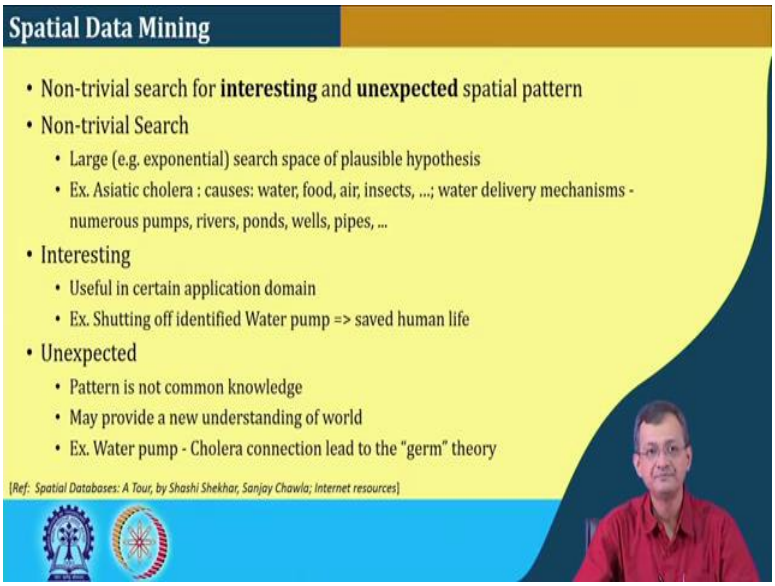
(Refer Slide Time: 09:52)

So, when we talk about now the data mining, spatial data mining per se, so, there are some metaphors like mining nuggets of informations embedded in large database; if it is a very large database as we have seen that spatial database is a pretty large. So, there are nuggets of informations which need to be from interesting useful and unexpected spatial pattern nuggets of large database and mining looking for those nuggets into the things, right?

So, these are the things and when you talk about spatial data mining, we search for the spatial patterns; that whether there is a spatial patterns are there or not and non-trivial search as automated as possible reduce human effort. So, that human intervention must be minimum and the search should be non-trivial, right? So, what we are looking for, so, it should be interesting useful and unexpected spatial mean unexpected patterns, right for.

Or what in other sense it should not be trivial thing right, which is expected that thing sun rises in the east may not be a, it may be a regular pattern but may not be a interesting unexpected pattern, right? If there is unexpected pattern that may help in my overall business process. So, that what exactly these miners or spatial data mining activity we want to look at that.

(Refer Slide Time: 11:14)



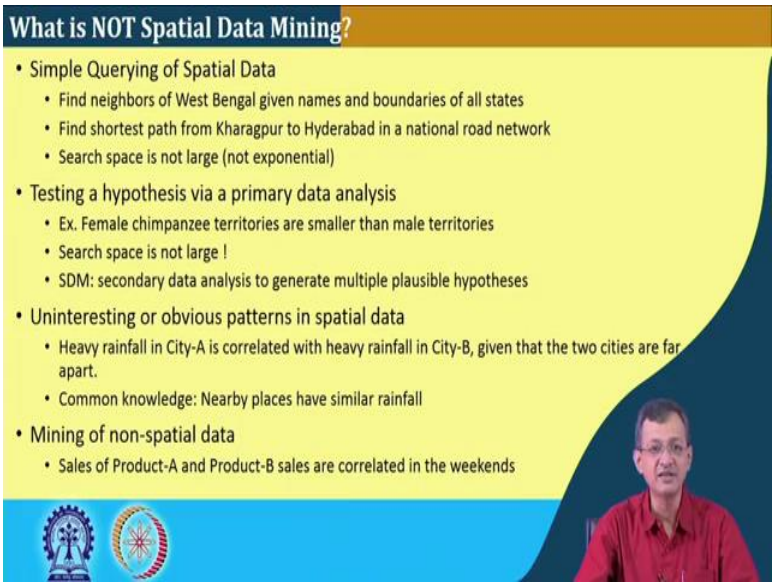
So, just to consolidate; one is that non-trivial for search interesting and unexpected spatial pattern; non-trivial search large search space was plausible hypothesis will say

pretty a pretty large has search space like for Asiatic Cholera; that the things we there if the cause may be water, food, air, insects, right? So, there are several parameters for each of them. So, the if you look at the overall search space it is quite high, right?

So, water delivery mechanisms, numerous pumps; even if we concentrate on water there may be different type of water delivery mechanisms like say pumps, rivers, ponds, wells pipes etcetera. So, it is difficult to handle the search space. So, we require some mechanism, some hypotheses, some heuristics to look into the things right, to how to handle this pattern.

Interesting useful in certain application domains shutting off, like in this case shutting off these identified water pumps saved human life right, so this is a direct effect. So, unexpected pattern is not common knowledge, right? So, this sort of patterns should not be a common knowledge then it is not a, we do not have to dig so much data and to find out that what is that pattern is; may provide a new understanding of the world, right? Like cholera connection leads to the germ theory and etcetera; different type of things.

(Refer Slide Time: 12:47)



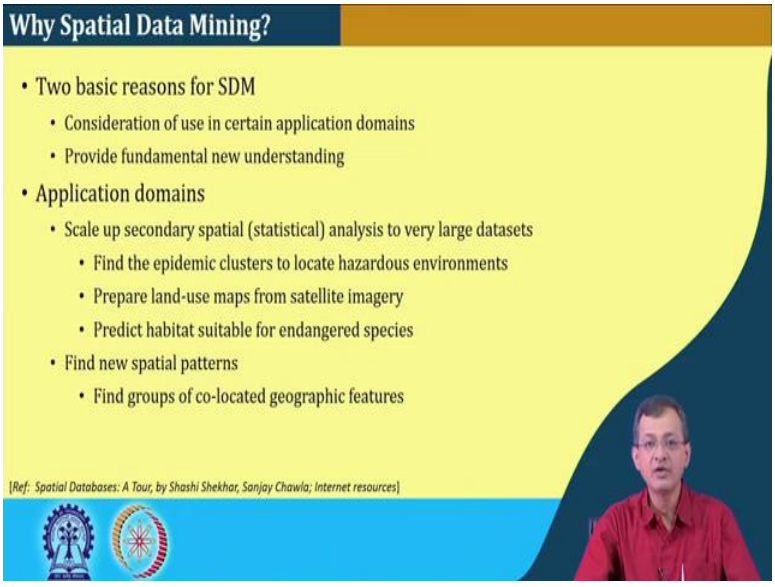
So, what is not spatial data mining? We have already seen, simple query on the spatial database; find neighbors of West Bengal given the names of the boundary of all states, that may not be spatial data mining, right? Or testing a hypothesis via a primary data analysis; like if the space is not large; and the secondary data analysis to generate

multiple plausible hypothesis; then it may not be a considered to be a problem or case for spatial data mining.

Uninteresting or obvious pattern in spatial database, right? Heavy rainfall in city-A is correlated with the rainfall in city-B, which two cities are not so far; there is a small typo; those, those cities are not so far apart or those cities are quite nearby or next to next cities and then it is obvious, right? So, common knowledge that nearby region will have similar rainfall, similar temperature and type of things or similar weather condition, right?

So, mining non-spatial data can be there but that we do not say a spatial data mining, but definitely mining there is non-spatial data can be there that can be useful for our context. Also, that may help us in giving some supportive secondary information. Like, in like mining, like products A and B are correlated etcetera, that may not be spatial but in some cases it may give some of the things which help us in better mining process.

(Refer Slide Time: 14:30)

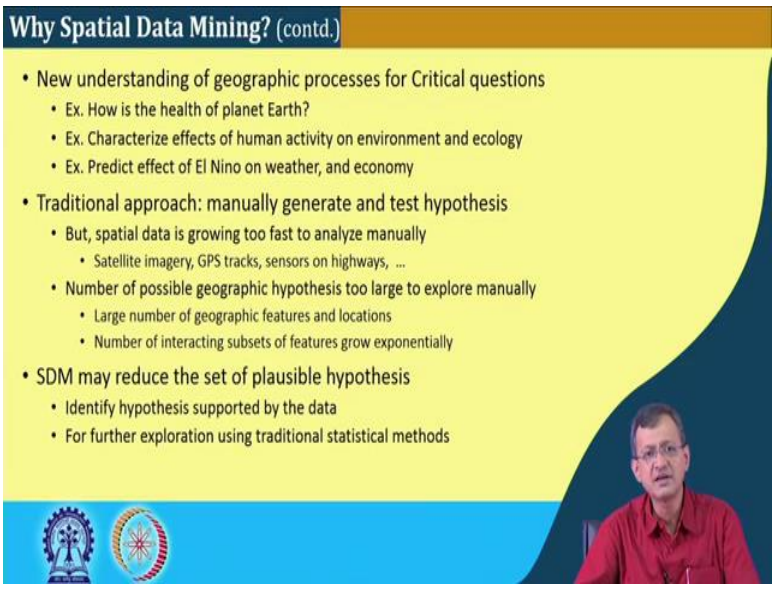


So, why we need spatial data mining? Two basic or fundamental reason; consideration of use of certain application domains, right? So, that that demands, like if I look at traffic or whether type of things they look for that whether we can find out some interesting pattern, even rainfall pattern etcetera. Provide fundamental new understanding, right? So, looking at finding a fundamental new understanding of the things like whether there is, this is the way it goes on type of things right, of a particular phenomenon.

So, there can be very various application domain; scale up secondary spatial statistical analysis to large database datasets, right? So, our standard statistical analysis we can scale up to large datasets; find the epidemic cluster to locate hazardous environment, right? So, we find out that where the clusters forming for the epidemic and go on; prepare land-use maps for satellite imagery, preparing land use land cover maps and predict habitat suitability for endangered species.

So, these all these may have my this ours spatial data mining or what we are discussing by helping supporting this application domain. A finding new spatial patterns is or maybe also interesting right; find groups of co-located geographic features can be interesting to look at. So, there are several application domains for spatial data mining; several there is several needs from the application domain and it helps in finding interesting pattern from large scale spatial datasets.

(Refer Slide Time: 16:27)



So, few more new understanding of geographic process for critical questions; that how is the health of the earth planet, or predict effect of El Nino on weather economy of a say on, on India of India. Traditional approach: manually generate and test hypothesis, so it was there. Spatial data is growing, it is increasingly growing in a very fast rate, right? So, it is very difficult if not impossible to have some traditional analysis approach for analyse this data.

So, spatial imagery GPS tracks, sensors highways and type of thing; number of possible geographic hypothesis too large to explore manually, right? Large number of geographic features and locations, a number of interacting subsets of the feature now grow exponentially and those things are there, right? So, SDM may reduce a set of plausible hypotheses.

So, it is it may reduce what you are discussing that the overall search space that will help in reducing the search space; identify hypothesis supported by data for further exploration using traditional statistical. So, initially reducing the suspects and then I even if I want to do a specialized exploration or more granular exploration, I can we can always fall back to our statistical, traditional statistical methods, to look into the things. So, these are things which are possible.

(Refer Slide Time: 18:08)

So, next things which come up that who are the major actors in this SDM or spatial data mining. So, one definitely the domain expert right; if I am looking for meteorological data it require a met scientist right; if we are looking for say something on traffic analysis, I require a traffic management person, traffic engineers and type of things we can held we can interpret the data.

So, identify SDM goals also what we are looking for. It is a huge data and for a particular day, particular things; which are the, which type of things we are looking for is

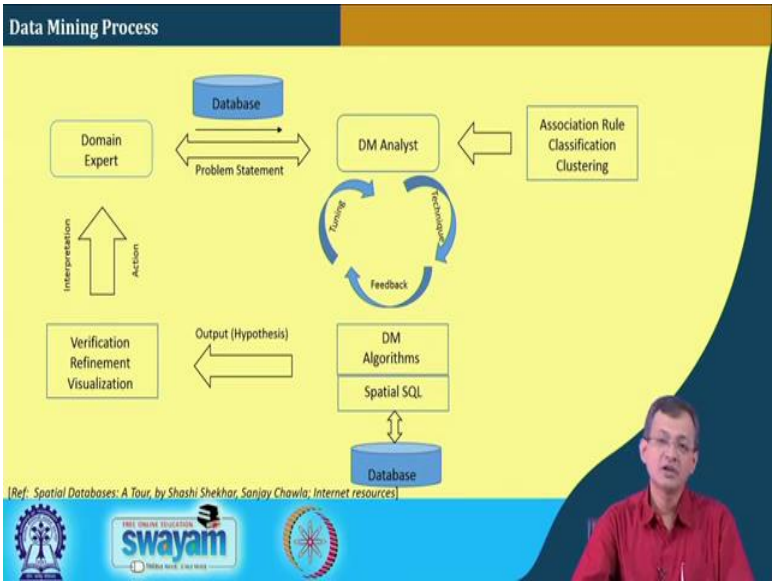
needed for these things. So, it is a identifying SDM goals statistical datasets experts is required.

Describe the domain knowledge, well known patterns that is, correlates etcetera that can describe, validate new patterns if there are a new pattern is there he or see the expert needs to validate the pattern, so, this is pretty important. On the other hand we require a data mining analyst, right? So, based on the expert, domain expert, this analyst will frame these algorithms right, or different procedures, data mining procedures; to work on these data and find interesting patterns.

So, that is needed, right, so, I require a data mining analyst on the other hand. So, helps identify pattern families, SDM techniques to be used, explain the SDM output to the domain expert, so, this is required and it is a feedback loop, I, between the things, right? domain expert explain something and implement see that whether it is correct, if there is something is required and type of things.

By that what it happens, that you able to set up the whole very precisely and the also the algorithm and process to require to reach that goal from the input data set also set on the or precisely set. Now, this can be used for that mining thing. So, it is a joint effort feature selection of the patterns for feature exploration. So, these are needs to be looked into.

(Refer Slide Time: 20:25)



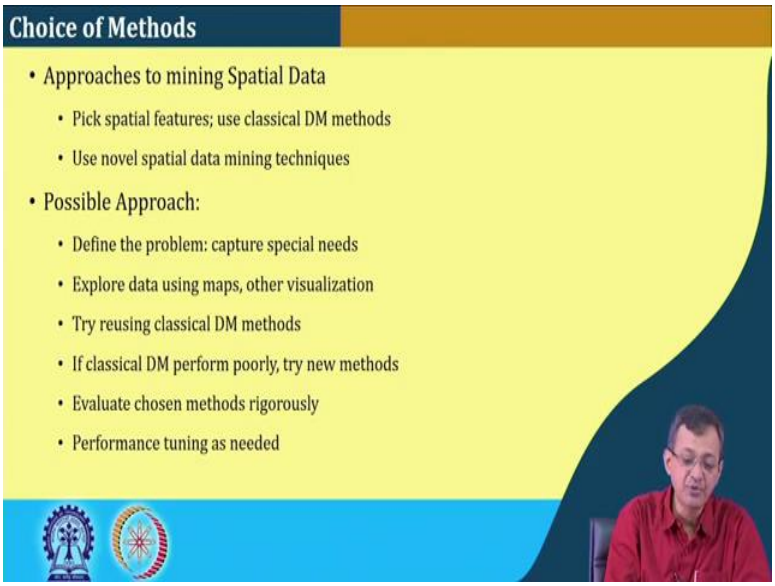
So, keeping all this in mind if we see, so, there is a domain expert, right, there is a data mining analyst, right? So, domain expert sets the goal that this is the thing is required and the based on that data discusses, discuss with the data mining analyst and said the what is the problem to be solved, based on this goal what is the exact problem to be solved and that is definitely based on what is available in the database or the spatial database things into the things.

Once that is there, then the data mining analyst go on using different data mining algorithms, which in turns involved spatial query to extract data from the things right; and this goes on type optics then this is output as the output will be the hypothesis where verification refinement visualization will be there, right?

So, this is a, in a circular fashion so, it is a technique used a feedback tuning the things and go on this, so, long it is satisfying that goal etcetera. Once is there then the output is that based on this hypothesis is basically pushed to the for verification, refinement and visualization and that interpretation and an action to be taken by taken the domain expert, right?

And, here again different type of basic techniques like what data mining aspects if we want to look at; whether is association rule mining or classification problem or clustering problem, that inputs get from the to this DM analyst, right? So, this is the, so, to say the overall big story or overall picture of that how these data mining activity will go on and type of things.

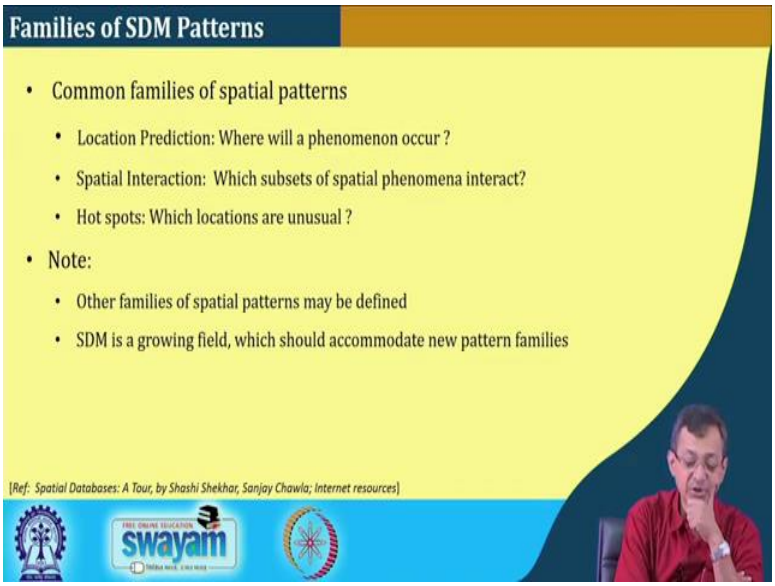
(Refer Slide Time: 22:23)



So, choice of methods, approach to mining spatial data; pick spatial feature, use classical data mining methods, right? So, approaches to mining spatial data. So, picks spatial. So, in to look at what are the feature set, what are the classical mining methods and use novel spatial data mining techniques. So, it is how I what are the different techniques, what are the methods need to be adopted for this meaning.

Possible approach, define the problem, what, what are the different possible approach; define the problem capture the in either a spatial need, explored data using map and other visualization. Try using classical data mining methods, if it is not there then we need to develop or incorporate new methods, evaluate chosen method rigorously and performance use tuning as needed, right? So, these are the, possible approaches.

(Refer Slide Time: 23:31)



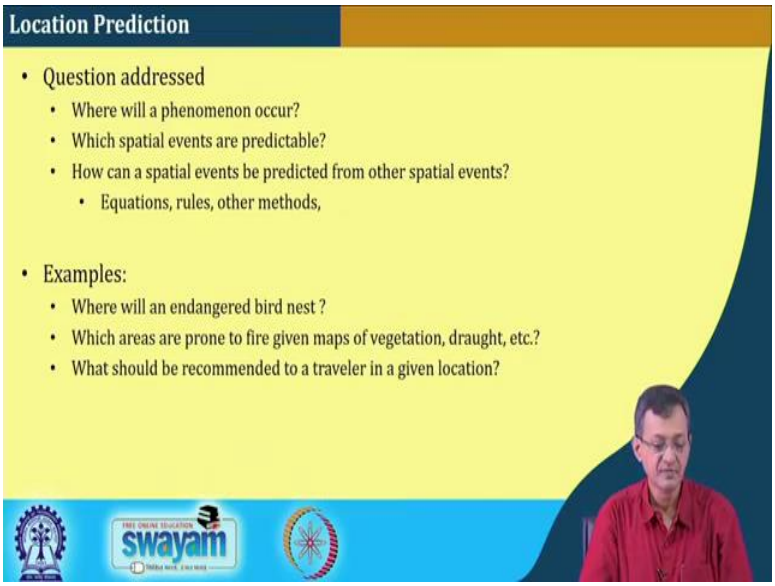
Families of SDM patterns, common prediction where is a, where is a phenomenon occur, right. These are the location prediction, there are spatial interaction and hotspots. So, if we look at the families of spatial patterns, so, what are the different groups; one is the location prediction, what, where next the things will happen, right?

So, it is more of a location prediction or finding that what next or what next location things will be there, there can be spatial interaction, which subsets of the spatial phenomena interact, right? So, there can be different spatial phenomena, out of that which spatial phenomena interact and other interesting work is on the hotspot, which locations are usual for some particular activities or particular species etcetera.

So, these are the hot hotspots, so like, like we are talking about crime hotspots, we are talking about say in a city of traffic the congestion hotspot and so, on and so, forth, so, these are things. So, it may be noted that other families have spatial pattern may be defined. So, you can always define based on your domain.SDM is growing field right, so it is there are lot of new things, new techniques, are coming up.

So, we should accommodate new pattern families, right; so that means, it say what we are seeing that there are it is a new field, new techniques technologies are coming up. The volume and the type of data being collected changes changing; like now we are having lot of sensory data for different IOTs or drone and type of things have different resolutions. So, the volume of data in a different scale and type of things.

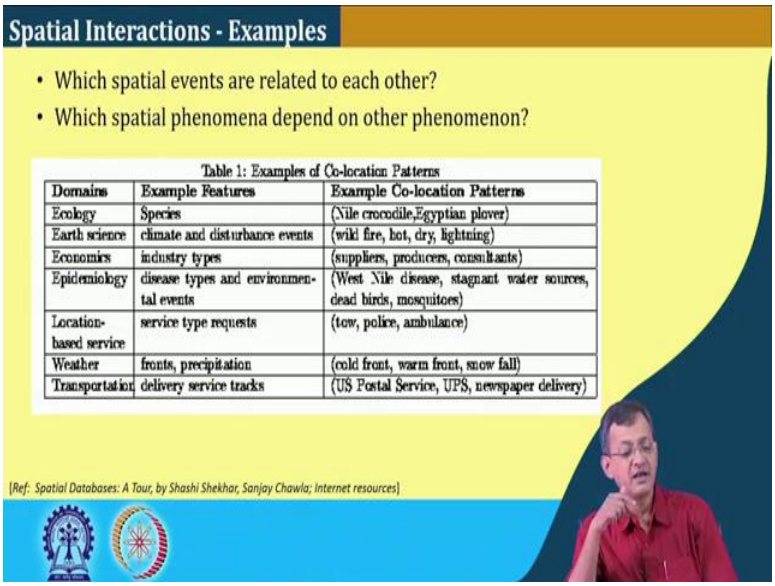
(Refer Slide Time: 25:26)



So, if we look at the location prediction as we told that there are three typical category of things. If we look at the location prediction questions like, where will the phenomena occur; which spatial events are predictable; or how can a spatial event be predicted from others spatial events, like equation, rules and other methods etcetera.

So, it is the location prediction or where a endangers like example, like where an endangered bird nest or which areas are prone to fire, given maps of vegetation drought etcetera, which should be recommended to a traveller in a given location and type of things, right?

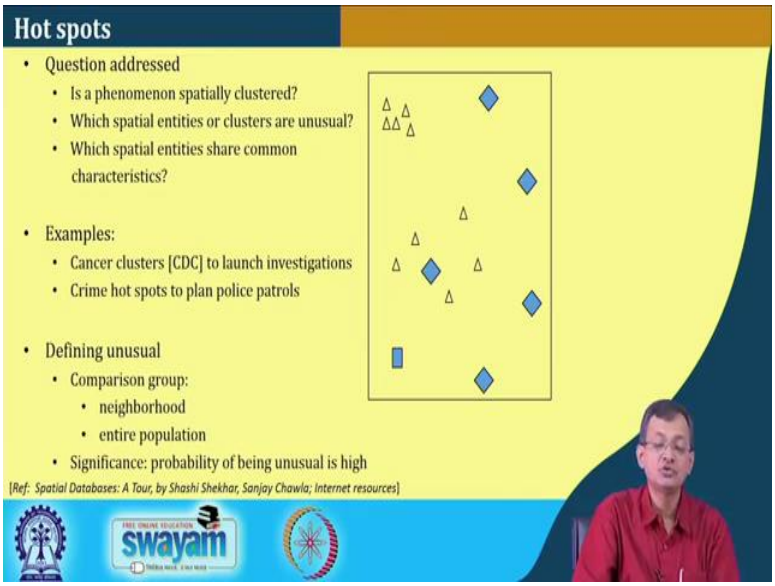
(Refer Slide Time: 26:14)



So, these are different type of location related and there can be spatial interactions, which spatial events are related to each other, which spatial phenomena depends on other phenomena. Like, here we have seen that, like ecology species that Nile crocodile, Egyptian plover and type of things. Similarly, art science if the feature set climate disturb disturbance events like wildfire and type of things.

So, we what we see, so, there are different, different type of feature sets and we want to look at that how the who, which spatial phenomena depends on other phenomena and type of things, right? Like, wildfires spread may be related to the dry forest, may be related to lightning, and high temperature why windy weather and type of things. So, these are the things which are there. So, there are some few examples out here.

(Refer Slide Time: 27:21)



So, and finally, the hot spot, the question address is the phenomena spatially clustered, right? So, if it is happening whether they are clusters right, like as we see that some sort of a things like as that cholera epidemic or the swine flu and type of things whether there is a cluster, right?

So, which spatial entities or cluster are usual a unusual. So, this we it forms a cluster and then it is unlikely there that should be a cluster at that, at that particular spatio-temporal location; means at that location on a particular time, which spatial entities share common characteristics. So, if there is a cluster whether these type of things are common characteristics between the difference spatial entities.

So, this helps us or this enables that some hotspots in a region right like maybe say in IIT, Kharagpur region where different hotspots are there. So, that is the phenomena has something else. So, like example cancer cluster to launch investigation, crime hotspot to plan police guards and type of things, right? So,, so, there can be disease type of cluster which is going in a particular things when investigate that whether these waterborne or some sort of a other type of issues are there.

Or, like in case of a crime hotspot it will help in police patrolling. Also, try to look at that how this is happening of the or these kinds manifested in over a time and so. So, defining unusual comparing group neighborhood or entire population. How do I say unusual?

Say, I say the temperature is unusually high so, that means, with respect to the particular region, right?

So, it is a neighborhood or on the entire population based on that we need to do need to be enter. Significance probability of being unusual is high, so, it is if I want to find out that which are the features which are pretty high in become unusual those need to be identified. So, what we, what we have seen or discuss today's class today's lecture is basically looking at the basics of spatial data mining.

Why that is required, right? Why required, which I can say there is a spatial pattern; which is not a pattern. Where, which are not considered to be spatial data mining and different aspects of the things. Like, where, what are the different choices and methods and to find out or to for mining or for spatial data mining what are the different type of approaches and so and, so forth, right?

So, this with this basic or fundamentals of the spatial data mining let us conclude our discussion today. What we will do in the subsequent lectures, we will deal some more aspects of some of the related to the spatial data mining or more or less all are related to the thing or spatial analysis in our subsequent lectures.

Thank you.