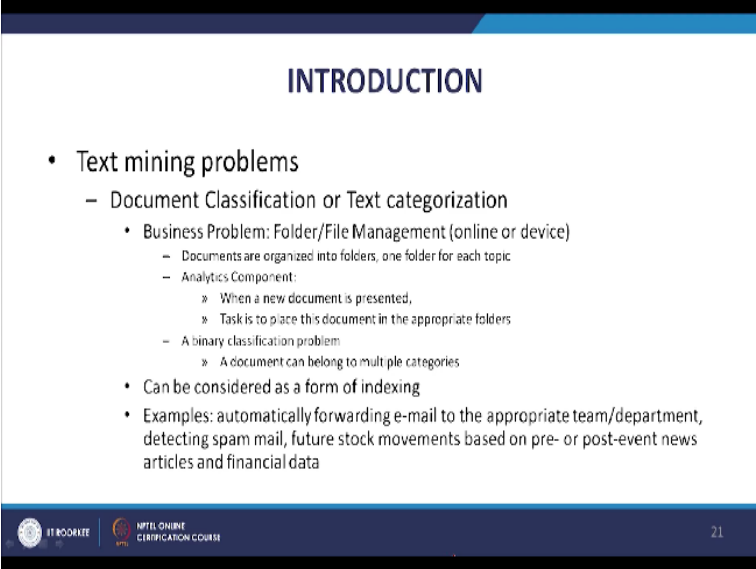


**Business Analytics And Text Mining Modeling Using Python**  
**Prof. Gaurav Dixit**  
**Department of Management Studies**  
**Indian Institute of Technology-Roorkee**

**Lecture-03**  
**Introduction-Part III**

Welcome to the course business analytics and text mining, modeling using python. So, let us do a recap of what we discussed in the previous lecture. So we started our discussion on text mining problems, different types of text mining problems. So one of the first one that we discussed is about our document classification text categorization.

**(Refer Slide Time: 00:44)**



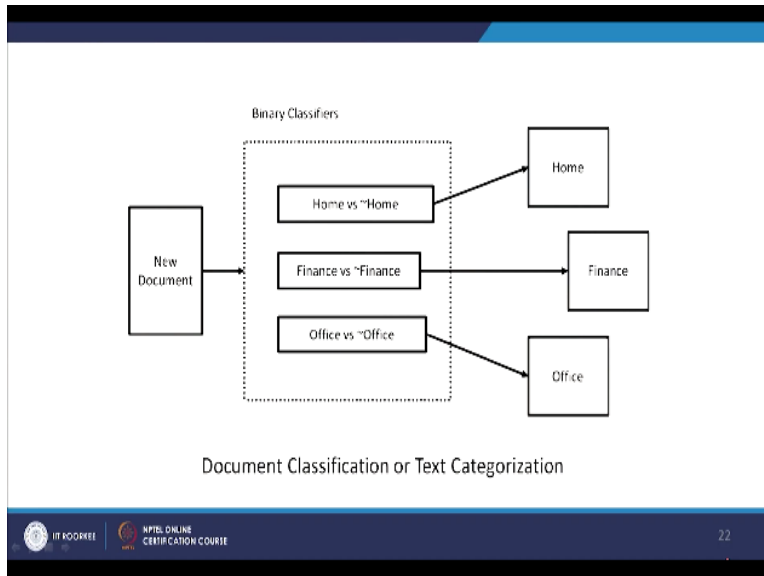
**INTRODUCTION**

- Text mining problems
  - Document Classification or Text categorization
    - Business Problem: Folder/File Management (online or device)
      - Documents are organized into folders, one folder for each topic
      - Analytics Component:
        - » When a new document is presented,
        - » Task is to place this document in the appropriate folders
      - A binary classification problem
        - » A document can belong to multiple categories
    - Can be considered as a form of indexing
    - Examples: automatically forwarding e-mail to the appropriate team/department, detecting spam mail, future stock movements based on pre- or post-event news articles and financial data

IIIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE | 21

So we discussed a different aspect like what is the relevant business problem, what is the analytics component, so, that aspect we discussed, we took the help of this example, given in this in this figure.

**(Refer Slide Time: 00:59)**



That given new document, and it is to be classified among you know, one of these 3 categories home, finance and office. So, how we can have a group of binary classifier, classifiers and it could be done.

**(Refer Slide Time: 01:19)**

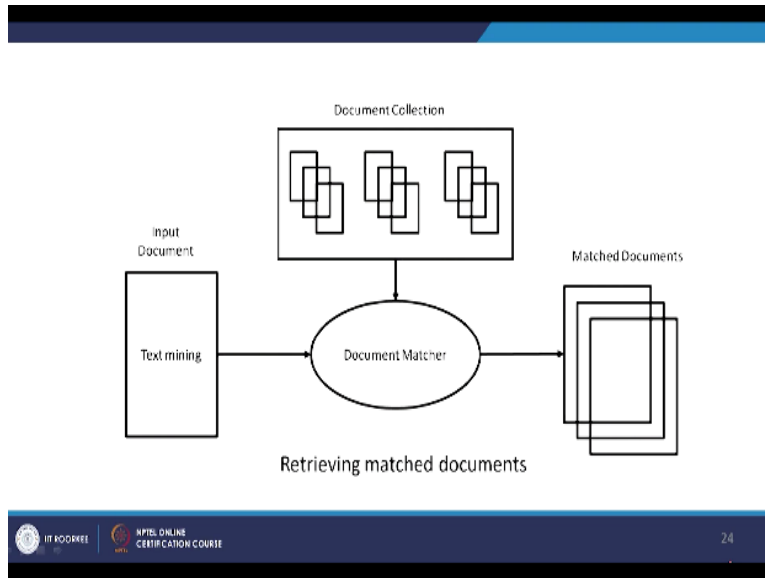
## INTRODUCTION

- Text mining problems
  - Information Retrieval
    - Business Problem: Document matcher (online or device)
      - Given a large collection of documents, finding relevant documents
      - Analytics Component
        - » Task is to retrieve the relevant documents based on the best matches of input document with the collection of documents
        - » New document is compared to all the other rows (documents), and the most similar rows and their associated documents are the answers
    - Similar to a search engine function
      - A few words are presented, and these words are matched to others
      - Best matches are presented as the responses
    - Based on measuring similarity as in nearest-neighbor methods

At the bottom of the slide, there are logos for IIT BOORNEE and NPTEL ONLINE CERTIFICATION COURSE, and the number 23.

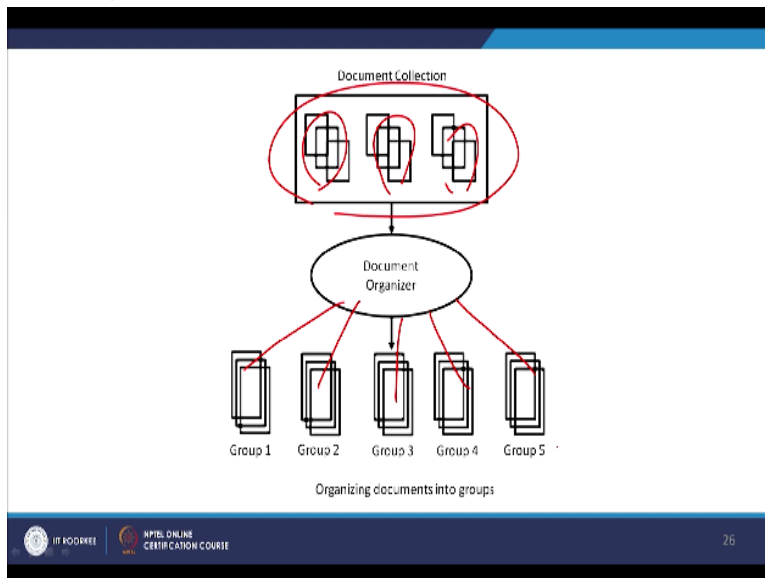
Then we talked about the next problem type that is information retrieval, where we talked about the business context meaning the document matcher and then we talked about the analytics component also. We took the help of this particular figure to explain this particular type of problems.

**(Refer Slide Time: 01:33)**



So, again input document text mining and this is then we have a document matcher here and then it is going to match this input document on text mining with the document collection and the matched documents are going to be presented. So, this is the next problem type that we discussed, then we talked about clustering and organizing documents.

**(Refer Slide Time: 02:03)**





So, the business context was unknown document structure, and we also discuss the relevant analytics component.

**(Refer Slide Time: 02:09)**

**INTRODUCTION**

- Text mining problems
  - Clustering and Organizing Documents
    - Business Problem: Unknown document structure (online or device)
      - Given a collection of documents with no known structure, find a set of folders such that each folder holds similar documents
      - Analytics Component
        - » Task is to cluster the similar documents in the collection and assign labels to each cluster
    - Examples: learn about the categories and types of help-desk complaints
      - Might lead to identification of complaints which have no existing solution





25

And we took the help of this figure where we are talking about this a document selection of this kind, you know you can see different groups here and you know, document organized, it will help us create different you know, groups: group 1 to group 2, group 3 to group 4, group 5. So, this is a different type of problem that we discussed, then we aim to the next text mining problem type that is information extraction.

**(Refer Slide Time: 02:35)**

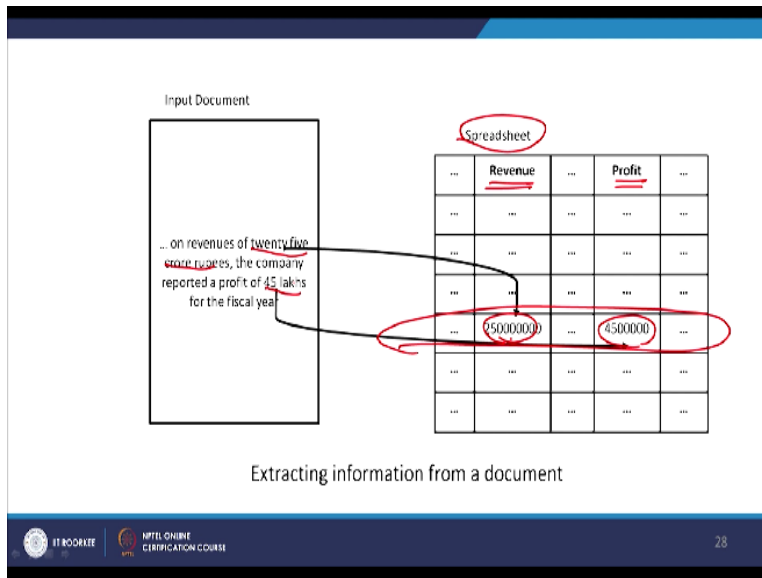
**INTRODUCTION**

- Text mining problems
  - Information Extraction
    - Business Problem: Populating database from unstructured data
      - Given a collection of documents, automatically filling the relevant values associated with certain defined variables in a database
      - Analytics Component
        - » Task is to extract data from an unstructured format based on words which can be higher-level concepts or real-valued variables
        - » The variable that is being measured will not have a fixed position in the text and may not be described in the same way in different documents
    - Examples: extracting the sales volumes and industry codes from company documents



27

We are business problem and being the populating database from an unstructured data, we talked about the analytics component as well.

**(Refer Slide Time: 02:44)**



So this was the figure that we use to explain this example, here as we can see in the input document, we have you know, certain information like twenty-five crore rupees and forty five lakhs being presented in different format and how this information can be extracted, you know, and put into a document like this, this a spreadsheet and in the relevant you know, columns in this fashion.

**(Refer Slide Time: 03:15)**

## INTRODUCTION

- Prediction and Evaluation
  - Text mining modeling process is similar to data mining modeling process
    - Process is about building models based on prior cases (from training partition)
    - Then the built model is used to predict the unseen cases (from test partition)
  - Evaluation of the model success is
    - Based on its performance on the test partition which is not part of the model building process
  - This mechanism works well for most of the text mining scenarios
    - However, there might be few special scenarios

So, we talked about these different types of, you know, text mining problems that we can deal with, then the next part is about the prediction and evaluation aspect of text mining. So, we discuss that in this text mining modeling also the process is quite similar to a data mining

modeling something we have discussed in our previous courses. So, that process remains the same.

So, again here also we discussed that we use prior cases and that form that are going to be part of our you know, training partition. So, we use prior cases to build our model. And then once this model is built, then we use this built model to predict the unseen cases, which are going to be part of our test partition. So, that is the text mining modeling process, very similar to what we do in data modeling process.

Now, we talked about few other aspects of a prediction evaluation as well that you know, it is the tests partition and our performance of the model on tests partition that is important. So, evaluation is always going to be on that partition the unseen observations. So, that part also be explained, we talked about that you know most of the text mining scenarios, this is the process that is going to work quite well other there can be some you know, special scenarios.

**(Refer Slide Time: 04:36)**

The slide is titled "INTRODUCTION" and contains the following content:

- Prediction and Evaluation
  - Example: Topic assignment
    - Assigning topics to news stories, such as financial or sports stories
    - However, news stories might change over time
      - News stories for test partition should be selected taking into account this sensitivity towards dates of publication
        - » Since model training process typically won't account for changes over time
  - Measurement of error
    - Typically, classical measures of accuracy work well if all errors are to be evaluated equally
    - However, as in topic assignment problem, not all errors will be evaluated equally
      - Measures of accuracy such as "recall" and "precision" are especially important in such scenarios

At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL ONLINE CERTIFICATION COURSE, and the number 30.

So, you know that will be discussing as we go along in this course, one example to discuss this, these prediction and evaluation you know aspect we talked about topic assignment and you know, assigning so, an example of this topic assignment could be assigning topics to news stories such as financial stories. So, as we understand there are so many news websites. Nowadays and you can see different, different tabs on those new websites with different news stories are categorized into different tabs.

You know, as we have learned from our previous discussion, that one new story can go into you know, more than one category also. So, how these topics you know, how these news stories are going to be assigned to different topics like you know, one user story being financial a story or a sport story, that could be one example. So, in this example, as you can see, the typical, you know, text mining modeling process that we discussed, might not work so well.

Because news stories might change over time. So, the same point is mentioned here that news stories might change over time. And when we talk about you know, text test partition which is going to be used to you know, to test the built model and it might so happen that by you know, the test partition might be having news, stories, which are you know, maybe from a later time period and therefore, the built model might not work, so, well on that period.

So, it is very important for us that you know, whatever model we have built, well you know, using the training partition observations, you know, the test partitions would also have similar kind of stories. So, the same point is mentioned here that news, stories for test partition should be selected taking into account the sensitivity towards date or dates of publication, because, you know, two years down the line if we try our model built model on news, stories, two years down the line.

Maybe the way the you know, articles are published the stories are, you know, categorize that might change in a slightly so, that might bring down the performance of you know, the built model, which might not be justified. So, therefore, these changes over time, you know, so, in the training process, we typically do not account for these changes over time. So, the text mining modeling process that we discussed, might not be able to cover such cases.

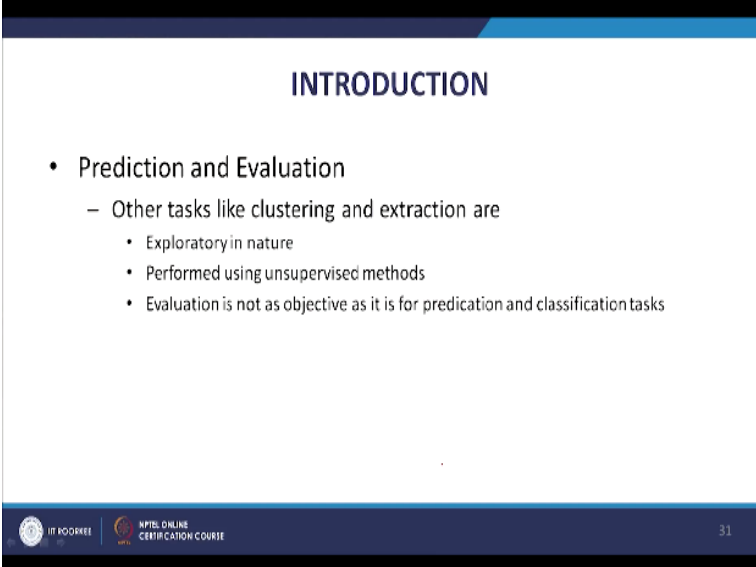
So we have to you know, take a special caution of such scenarios, then we talked about measurement of error. So, we discuss that the classical measures of accuracy, they work well if all errors are to be evaluated equally. For example, you know, if we have two classes and you know, if class one and class zero or let us say class one and class two, if you know their performance, there is no special class.

And just like we discussed in our previous courses also, if there is no special class, so, it would not matter us you know, we would look for overall, you know, overall classification accuracy, rather than looking you know, focusing on one special class and having more observations, you know, classified to that particular special class. So, if that is not the scenario, then the classical measures of accuracy, they were quite well.

So, that is what we are mentioning here, now, if we look at the here topic assignment problem, so, in this case, not all errors will be evaluated equally. So, as we saw this so, many different accuracy measures, we will talk about these measures, in more detail in incoming lectures. So, recall and precision so, these are the majors, which could be especially important in such scenarios, where if we are looking to identify you know, a particular you know, in observations or you know, cases belonging to a particular class.

So, these majors can really help us in that context, similar kind of discussion, we have done in our previous data mining courses as well. Now if we look at the other tasks, which are more complimentary in nature in text mining context, the clustering and extraction.

**(Refer Slide Time: 09:08)**



The slide is titled "INTRODUCTION" and contains a bulleted list of points. The first point is "Prediction and Evaluation", which has a sub-point: "Other tasks like clustering and extraction are". This sub-point has three further sub-points: "Exploratory in nature", "Performed using unsupervised methods", and "Evaluation is not as objective as it is for predication and classification tasks". The slide footer includes the IIT ROORKEE logo, the text "NPTEL ONLINE CERTIFICATION COURSE", and the number "31".

- Prediction and Evaluation
  - Other tasks like clustering and extraction are
    - Exploratory in nature
    - Performed using unsupervised methods
    - Evaluation is not as objective as it is for predication and classification tasks

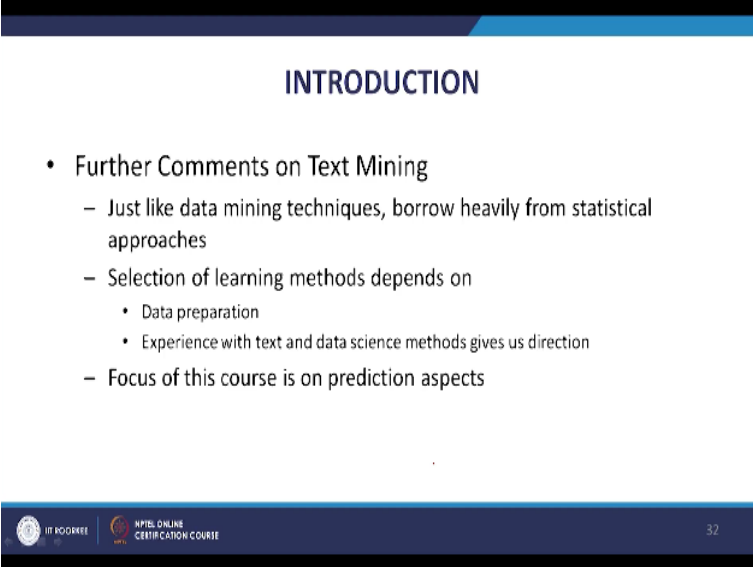
So, these are more explorative in nature. So, there you know, the metrics the evaluation part is, you know, not as objective as we might think because, in clustering it is as we have understood from our previous discussion in the in the previous lecture, that we are trying to create new



groups. So, we have unstructured you know, collection, and we are trying to create new groups. So, that is quite subjective, you know, it depends upon how that those newly formed groups are to be we used later on.

So, our clustering process would also be slightly focused on you know, focus on that direction. So, it is going to be slightly more subjective. So, the evaluation aspect is you know, not as discussed in these tasks like clustering and extraction, mainly them being exploratory in nature and mainly, you know, being very similar to you know, what we discussed as unsupervised learning methods in our data mining courses.

**(Refer Slide Time: 10:15)**



The slide is titled "INTRODUCTION" and contains the following content:

- Further Comments on Text Mining
  - Just like data mining techniques, borrow heavily from statistical approaches
  - Selection of learning methods depends on
    - Data preparation
    - Experience with text and data science methods gives us direction
  - Focus of this course is on prediction aspects

At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL ONLINE CERTIFICATION COURSE, along with the number 32.

Few more comments on text mining. So, just like data mining techniques, here also, we borrow very heavily from a statistical you know, approaches. So, the whole process itself and you know, many techniques also the way you know, they have been developed the way they are applied. So, lot more, you know, in a lot more has been borrowed from the statistical approaches, now, we look at the selection of learning methods.

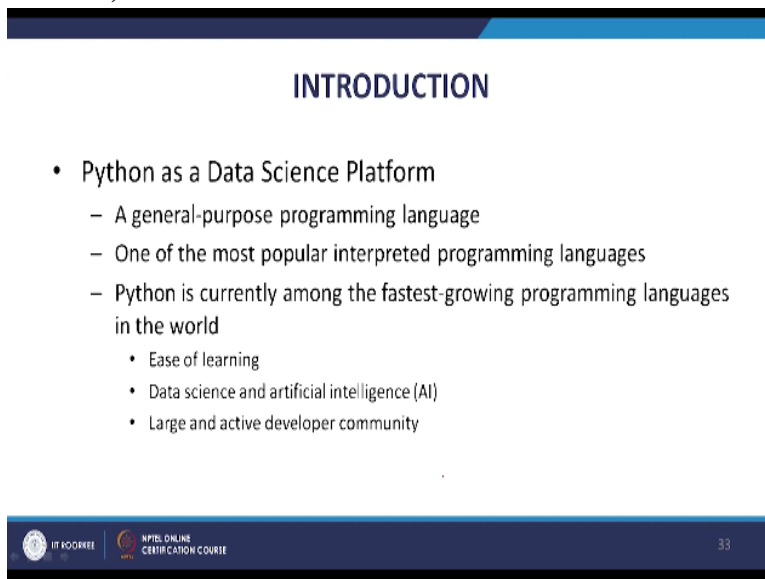
So, this is going to depend on you know, 2 aspects mainly the data preparation aspect, so, the way data is going to be prepared in text mining. So, as you might have understood from our discussion in last 2 lectures, that data preparation is going to take a lot more time in text mining modeling scenario in comparison, so, what in comparison to what we do in you know, data mining modeling.

There we typically deal with a structured data, our in text mining modeling, we start with the unstructured data and then we try to you know, you know, process it transform it into a slightly a structure form. So, data preparation effort is very time consuming and time intensive, and what kind of learning methods we are going to apply, you know, certain aspects of our data preparation will also have a role in that, experience with text and data science methods.

So, that will also give us direction, we will get to as we get more experienced in text mining modeling process, we get to know what kind of technique, what kind of method is going to work well, with what kind of you know data, so, that experience that learning will also help us in terms of selecting a appropriate learning method. Now, as far as this course is concerned, our focus is going to be on prediction aspects.

So, the problems, different types of text mining problems, all of that we have discussed. So, they are also you would have, you know, noticed that initial few problem types, they are more into your prediction aspects. So, focus of the scores is also going to be in that direction. Now, let us talk about since, we are going to use python for text mining modeling for this course. So, let us talk about a few you know, few things about python as a data science platform.

**(Refer Slide Time: 12:38)**



The slide features a dark blue header with the word "INTRODUCTION" in white, bold, uppercase letters. Below the header is a bulleted list. The first bullet point is "Python as a Data Science Platform", which is followed by three sub-bullets: "A general-purpose programming language", "One of the most popular interpreted programming languages", and "Python is currently among the fastest-growing programming languages in the world". The last sub-bullet has three further sub-bullets: "Ease of learning", "Data science and artificial intelligence (AI)", and "Large and active developer community". At the bottom of the slide, there is a dark blue footer containing the NPTEL logo, the text "NPTEL ONLINE CERTIFICATION COURSE", and the number "33".

## INTRODUCTION

- Python as a Data Science Platform
  - A general-purpose programming language
  - One of the most popular interpreted programming languages
  - Python is currently among the fastest-growing programming languages in the world
    - Ease of learning
    - Data science and artificial intelligence (AI)
    - Large and active developer community

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 33

So, as we know that python is a general purpose programming language, so, it is you know, being used for software development, application development, and there are certain aspects of

python as a programming language, which are making it very popular in the software development domain, another aspect of python is that it is one of the most popular interpreted programming languages.

So, there are typically 2 types of programming language 2 categories of programming language, one is you know, compiled programming language and the other one is interpreted programming language. So, in compiled programming language, what happens is, once we write our lines of codes, once we complete our coding aspect coding part, then that code is actually compiled.

So, that compiles into machine level instructions, so, that when the that, you know, when it comes to execution, less time is taken in executing that code. So, any code once it is developed, it is compiled and then it is executed. So, time taken in compiled code time taken to execute the compiled code is typically less so, those kind of you know, compile you know, programming language, they have their own set of advantages.

Because of the process that is followed over there. So, they are you know more into where we are developing systems. So, more into systems programming, so, these kind of you know, compiled programming language are more popular where we are trying to develop you know larger systems, while there is another category interpreted programming languages, there the code whatever code that we write is not compiled, rather it is you know, in that process of converting the code into machine level instruction is done at the runtime itself.

So, as you would understand, because, you know, converting the code into machine driven instruction, that being done, for runtime takes the runtime time is a bit longer in comparison to compile programming language. So, this is one of the you know, in a sense, you know disadvantage with interpreted programming language that time taken runtime is more however, because of many things, you know, this translation to machine level instruction being done in the runtime, many things can be done dynamically.

So, whenever we are going to change a certain part of our code, which is written in interpreted programming language, you know, not much you know, have to be done while we are running it, because the process is still remains same. However, when we compare this aspect with the compiled programming language, they are we have to recompile are whatever code it be changed certain part of code that has to be you know.

And we had already compiled the previous version, now, we will have to recompile and only then it is going to be executed. However, in this case, you know, that flexibility is there. So, you would see that of python and other you know, such interpreted language, they are more useful in the application development, you know, domain. So, while the other programming compiled programming language, they are more into the systems development.

These and interpreted language are more into application development, so, different types of programming language, they have developed their own utility in that sense. Now, we will be talking about it and another important aspect about python is that, it is currently one of the fastest growing programming language in the world. The main reason being is that not just this is a general purpose programming language.

So, it can be you know easily integrated into a production setup, so, you can write your reports and easily you can do your production setup and production setup and deploy the software, the developed software that programme. So, that advantage is the python however, it is it has, you know, another few features with associated with it, for example, ease of learning.

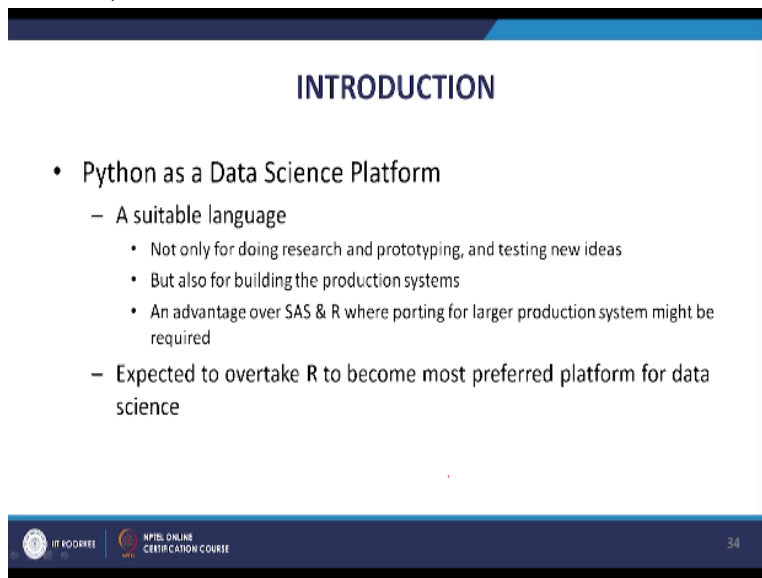
So, python is something which is a learning curve, that learning period is smaller. So, because of that many people can easily pick up python, and it is increasingly being used by data science and artificial intelligence community. So, in this domain, the usage of python is increasing, which has led to large and active developer community. So, software community which they are, you know, on a regular basis.

They are building new libraries, new functionalities, new features, which could be used by data science and artificial intelligence community, and overall, you know, general application

development also. So, because of this growing community, because of this growing development in this in the python platform, as such, this is becoming one of the fastest growing programming language in the world, as of now.

Few other aspects about python as a data science platform. So, this is a suitable language, not only for doing research and prototyping, but testing new ideas also. So, because ease of learning, you can easily write your code and, you know, execute. And so, the research and prototyping and the way different libraries have been developed for python, it becomes slightly easier for scientists to, you know, to do their research while do their prototyping and, you know, in an efficient manner, test out their ideas.

**(Refer Slide Time: 18:37)**



The slide is titled "INTRODUCTION" and contains the following content:

- Python as a Data Science Platform
  - A suitable language
    - Not only for doing research and prototyping, and testing new ideas
    - But also for building the production systems
    - An advantage over SAS & R where porting for larger production system might be required
  - Expected to overtake R to become most preferred platform for data science

At the bottom of the slide, there are logos for "IT KODKREE" and "NPTEL ONLINE CERTIFICATION COURSE" on the left, and the number "34" on the right.

So, because of this is ease of doing, you know, this says, you know, this particular platform python, is you know suitable language for data science platform as well. Also useful as I discussed, also useful for building the production systems, because it is as good as any general purpose programming language. So, in that sense, easily it can be integrated into a production systems and this kind of these are few advantages, which are giving python as an edge, over SAAS and R where once we do something.

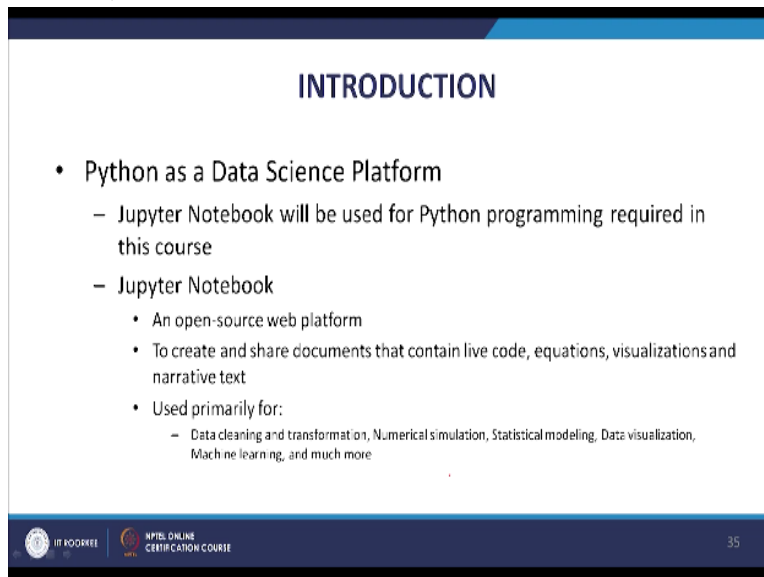
Once we developed something in SAAS, and R that has to be ported into another language into another platform, so that it could be used for larger production system, however we developed something in the python platform that can be directly integrated with the production systems.

That is one advantage is that python is having over SAAS and R and that is leading to its increased popularity.

So that is why another point about python is that sooner it might overtake R to bigger most preferred platform for data science. So many of the libraries many of the contributions that have gone through in the in the large community of R, they are being ported to python platform as well. So what is available there is also being poured into python platform and python has its own, you know, its own advantages, which can make it you know, the most preferred platform for data science.

Now, as far as this course is concerned, we will be using Jupyter notebook. We will be talking about what this notebook is about.

**(Refer Slide Time: 20:23)**



The slide is titled "INTRODUCTION" and contains the following content:

- Python as a Data Science Platform
  - Jupyter Notebook will be used for Python programming required in this course
  - Jupyter Notebook
    - An open-source web platform
    - To create and share documents that contain live code, equations, visualizations and narrative text
    - Used primarily for:
      - Data cleaning and transformation, Numerical simulation, Statistical modeling, Data visualization, Machine learning, and much more

At the bottom of the slide, there are logos for IIT ROORKEE and NPTEL ONLINE CERTIFICATION COURSE, along with the number 35.

So, we will be using Jupyter notebook for our python programming, whatever programming is required in this course. So, this is the platform that we will be using, and a few points about Jupyter notebook. This is an open source, a web platform. And the main idea behind this is to be able to create and share documents and that might contain no live code, equations, visualizations and, you know, narrative text also.

So, the main idea is, you know that sharing because, this is not a web platform, that makes it, it is suitable for, you know, platform for sharing with other people, not just within the team across

the world, any place because all, you know, every, the whole document is being developed in the web platform. And it is this platform in a way combines the coding part, the equations part, the narrative part of English visualization part into one container, which is the unique feature of this particular platform.

And that is why this platform is also gaining increasing popularity. Now, this Jupyter notebook though it is started as a part of python projects, and was mainly developed for, you know, python programming, but other programming languages including R and Java and other those, you know, codes written in those programming language can also be and also be, you know, done in this particular, you know, web platform.

So, primarily being used for data cleaning, transformation, numerical simulation, statistical modeling data visualization machine learning, and much more. So, these are the primary users where Jupyter notebook is being used, though, the developers, they are continuously increasing its functionality and more number of language are being into being brought into the, you know, Jupyter notebook fold.

And, you know, not limited to nowadays not limited to just data science specific, you know, tasks that we have mentioned here. Now, if we talk about this course, you know, we will focus on python programming language, and its data oriented library ecosystem, because we are going to use python for our analytics course, and this course, we are focusing on the text mining part of analytics.

So, therefore, how the data oriented library ecosystem of python works. So, our focus is on that and the relevant programming skills in python. So, that is also so those things are our main agenda here in this course, and we are not looking to you know, provide you know, serious expertise in python programming language as such, rather, focuses on the data science aspect.

**(Refer Slide Time: 23:14)**

## INTRODUCTION

- Python
  - This course focuses on using Python programming language and its data-oriented library ecosystem for analytics
  - Suitable for application development (Higher productivity language)
    - Due to it being an interpreted programming language
    - Run substantially slower in comparison to compiled language like Java or C++



Few more points about python is that as I have discussed that suitable for application development, and this is this makes it higher productivity language. So, you know, if because, it is the application domain software application development is the area where a lot more productivity is require, a lot more applications are developed, because systems you know systems software that we have, they are typically, you know, a one software is you know.

There are going to be few software and they are going to be competition among those softwares for in the market, however when it comes to the application, those systems software can be used by different by large pool of applications developers to cater to different needs of the customers and develop you know, different applications. So, in the application development domain, higher productivity is required. So, the python is one of such you know, programming language, which provides this.

And the main reason for this language being productive is it being an interpreted programming language. So, this advantage that run substantially slower in compile into compiled language like Java or c++.

**(Refer Slide Time: 24:31)**



## INTRODUCTION

- Python
  - Not suitable for highly concurrent, multithreaded applications, particularly applications with many CPU-bound threads
    - Due to global interpreter lock (GIL) mechanism
      - Prevents the interpreter from executing more than one Python instruction at a time
- Python data ecosystem
  - Important library packages
    - NumPy, pandas, and matplotlib

Now, another disadvantage of python is that it is, not suitable for highly concurrent multi threaded applications, particularly application with many CPU bound threads. So, if we have a you know, parallel, multi threaded kind of code, so, python is not suitable for those kind of scenarios. And main reason being that it has something called global interpreter lock mechanism.

So, this mechanism in a sense prevents the python interpreter from a executing more than one python instruction at a time. So, it is one instruction you know, being executed at a time and in some scenarios, this you know, serial execution is preferred over parallel execution and those scenarios, python is definitely you know, having the higher advantage, but if we are looking for, you know, paralyzing our core or parallel execution of the code, we have parallel code where we are looking to expire the, you know, many CPU bound threads.

So, in that scenario, it might not be that suitable, however, some developments are being made to make it you know, work can even in those scenarios as well, but primarily it is not suitable. Now, we will be talking about the python data ecosystems. So, we have important some important library packages, that you will have to familiarize ourselves during this course, and you will be using them a lot.

So, some of these packages are NumPy, which is NumPy, pandas and matplotlib. So, we will be talking about these packages in a bit more detail. So, let us start with NumPy. So, this is short for numerical pattern. So, this is mainly for a numerical computing in python. So most of the numerical computing calculation that are required that is part in any programming language, that is, you know, provided by this particular library.

**(Refer Slide Time: 26:45)**

**INTRODUCTION**

- NumPy
  - Short for Numerical Python
  - For numerical computing in Python
  - Contains
    - Arrays for storing data (used as primary data structure), functions for manipulating data
- pandas
  - Name derived from **panel data**, an econometrics term
  - For working with tabular or structured data

IIT Koorkee NPTEL ONLINE CERTIFICATION COURSE 38

And it is apart from other things, but our focus what you know, based on our focus based on this course, we have mentioned a few things here. So, apart from other thing, it contains arrays for storing data, so, that is going to be the primary data structure that we will be using in the context of this course, and functions for manipulating this data, then next library is pandas. So, this is the name of panda is derived from panel data.

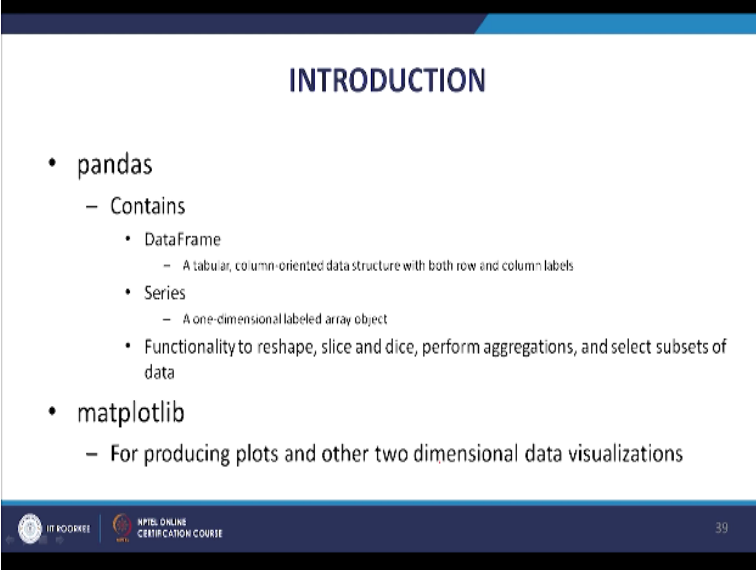
So you can see pan and da. So, it is coming from there, this is a so this is standardize and econometric term and this typically refer refers to multi dimensional highly structure multi dimensional data set. So, pandas as a library provides, you know, that kind of functionality, where we can easily work with you know, highly structure multi dimensional data set. So, the same thing is mentioned here for working with tabular or structured data.

So, this particular library is especially useful for that. So, pandas is something that you know, brings the python platform at equal footing in comparison to R platform, R platform was mainly developed for you know, statistical modeling and therefore, the way data is treated in

statistical modeling, that kind of functionality was built into R and so, the same thing, similar thing is being provided in python platform also.

And that is why python is quickly gaining over R so, it has certain advantages of its own as we have discussed, and, you know, whatever is there in R is also being built into python platform.

**(Refer Slide Time: 28:20)**



The slide is titled "INTRODUCTION" and contains a bulleted list of Python libraries. The first bullet is "pandas", which has a sub-bullet "Contains" with two items: "DataFrame" (described as a tabular, column-oriented data structure with both row and column labels) and "Series" (described as a one-dimensional labeled array object). The second main bullet is "matplotlib", with a sub-bullet "For producing plots and other two dimensional data visualizations". At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL Online Certification Course, along with the number 39.

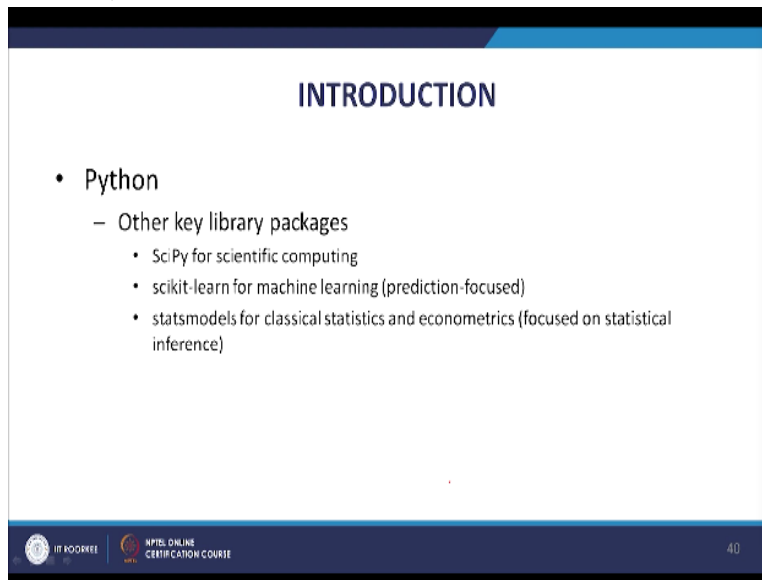
- pandas
  - Contains
    - DataFrame
      - A tabular, column-oriented data structure with both row and column labels
    - Series
      - A one-dimensional labeled array object
    - Functionality to reshape, slice and dice, perform aggregations, and select subsets of data
- matplotlib
  - For producing plots and other two dimensional data visualizations

So, apart from other things, pandas contain data frame. So, which is very similar to what we have you know, very similar kind of functionality in terms of functionality very similar to what we have in R, so, data frame is a tabular column oriented data structure with both row and column labels, series, this is another one dimensional label array object that is part of this. So, functionality that is there in pandas is to reshape, slice and dice and you know, perform aggregation and select subsets of data.

So, these are the kind of functionality that we would be interested in for our this course, then there is another library matplotlib. So, this is mainly for producing plots and other 2 dimensional data visualization. So, that is also a very important part of analytics. So, even in text mining modeling we would be you know, producing many of such you know, plots for our analysis purposes.

So, this particular library is going to be useful in that, other key library packages that we would be using from time to time SciPy this is for scientific computing.

**(Refer Slide Time: 29:25)**



**INTRODUCTION**

- Python
  - Other key library packages
    - SciPy for scientific computing
    - scikit-learn for machine learning (prediction-focused)
    - statsmodels for classical statistics and econometrics (focused on statistical inference)

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSE 40

And then we have scikit learn, this is for machine learning. So, we have the algorithms in this package, which are mainly developed, you know, in the prediction focused context. And then we have another package stats model this is for classical statistics and econometrics you know, kind of, you know, algorithms have been developed. So, the main focus is on the statistical influence there.

So, as you can see, the whole platform provides enough libraries, where, you know, machine learning, you know, focuses there were statistical modeling focuses there, scientific computing, numerical computing, dealing with data, the way it is typically done in the analytics area. So, all those features functionalities are built into python platform through these libraries, other consideration that, you know, one might consider is that, you know, IDE integrated development environment.

**(Refer Slide Time: 30:27)**

The slide is titled "INTRODUCTION" in a bold, dark blue font. Below the title, there is a bulleted list. The first bullet point is "Python: Other considerations", which has two sub-bullets: "Integrated Development Environments (IDEs) and Text Editors" and "In this course, we shall be using Python 3.7 or later versions". The second sub-bullet has a further sub-bullet: "Spyder (free), an IDE currently shipped with Anaconda", which is followed by "Similar to RStudio that was used in previous courses". The slide has a dark blue header and footer. The footer contains the IIT Kharagpur logo, the text "NPTEL ONLINE CERTIFICATION COURSE", and the number "41".

- Python: Other considerations
  - Integrated Development Environments (IDEs) and Text Editors
    - Spyder (free), an IDE currently shipped with Anaconda
      - Similar to RStudio that was used in previous courses
  - In this course, we shall be using Python 3.7 or later versions

And text editor, so, Spyder is the free IDE that is currently separate with an anaconda distribution, that is the distribution that will be using, as we will see in upcoming lectures. So, this IDE is spyder is very similar to Rstudio that we have used in our previous you know, data mining courses. So, however this time we are using going to use Jupyter notebook, which is a web you know, platform.

So, this is going to be a different experience in the analytics space. So, we are one experiences with it, we have we have different panels, escape for one banner for escape and console the help and plot and environment section and all those things are there however web platform as we see that everything is integrated into one thing. So, that is why we have decided to go with Jupyter notebook and also its popularity.

So, in this course, as far as python versions are considered we will be using python 3.7 or later versions. So much of our code that will be using it would be compatible with this or forward compatible with the later versions. In terms of course roadmap so, this is the roadmap that will follow.

**(Refer Slide Time: 31:43)**

## INTRODUCTION

- Course Roadmap
  - Module I: General Overview of Text Mining
  - Module II: Python for Analytics
  - Module III: Data Preparation
  - Module IV: Predictive Models for Text
  - Module V: Retrieval and Clustering of Documents
  - Module VI: Information Extraction
  - Module VII: Conclusion



Our first module is going to be on general overview of text mining. So that is something that we have covered. Now, next module is on python for analytics. So that is something that will start from the next lecture. Then the after that third module is on data preparation, where we focus on the data processing aspects, then the we have module 4 where we will focus on predictive models for text.

And then we have module fifth, which will focus on retrieval and clustering of documents, which is the you know, supplementary tasks of what we have discussed in text mining problems. And then we have module 6 model VI, where we will focus on information and extraction. And in the last module, we like to conclude. So this is the overall roadmap in this course that we are going to follow. So we would like to stop at this point. And we will take up the next module in the next lecture, thank you.

**Keywords:** Spyder, SciPy, Rstudio, Jupyter notebook, NumPy, pandas, matplotlib, information extraction.