

Python for Data Science
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 01
Why Python for Data Science?

Welcome to this course on Python for Data Science. This is a 4-week course; we are going to teach you some very basic programming aspects in python. And since this is a course that is geared towards data science, towards the end of the course, based on what has been taught in the course, we will also show you two different case studies; one is what we call as a function approximation case study, another one a classification case study.

And then tell you how to solve those case studies using the programming platform that you have learned. So, in this first introductory lecture, I am just going to talk about why we are looking at python for data science.

(Refer Slide Time: 01:10)



What is Data Science?

- **Data Science** is the science of analyzing **raw data** using statistics and machine learning techniques with the purpose of drawing insights from the data
- **Data Science** is used in many industries to allow them to make better business decisions, and in the sciences to test models or theories
- This requires a process of inspecting, cleaning, transforming, modeling, analyzing, and interpreting raw data

The slide includes three images: a blue background with binary code, a magnifying glass over a bar chart, and a man speaking in front of a screen displaying 'Conclusions & Recommendations'. A navigation bar at the bottom shows 'Python for Data Science'.

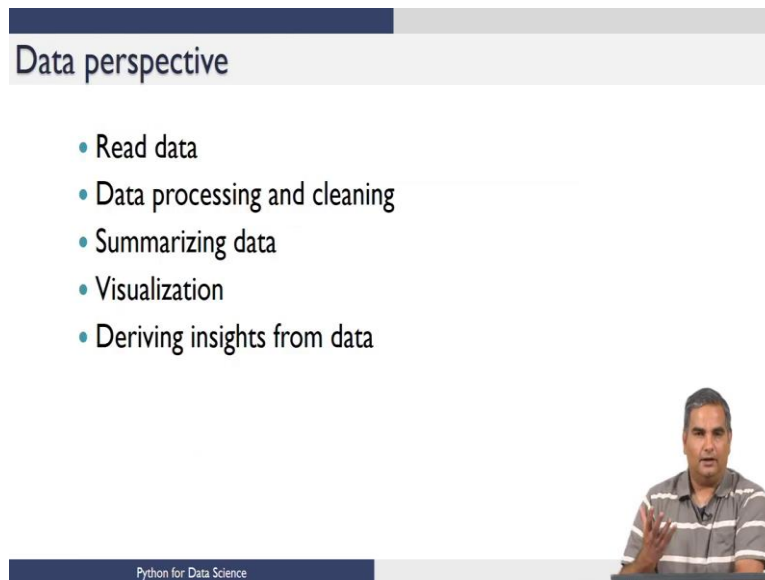
So, to look at that first, we are going to look at what data science is. This is something that you would have seen in other videos of courses in the NPTEL and other places. Data science is basically the science of analyzing raw data and deriving insights from this data. And you could use multiple techniques to derive insights; you could use simple statistical techniques to derive insights, you could use more complicated and more sophisticated machine learning techniques to derive insights and so on.

Nonetheless, the key focus of data science is actually deriving these insights using whatever techniques that you want to use. Now there is a lot of excitement about data science, and this excitement comes because it's been shown that you can get very valuable insights, from large data and you can get insights about how different variables change together, how one variable affects another variable and so on with large data which is not very easy to simply see by very simple computation.

So, you need to invest some time and energy into understanding how you could look at this data and derive these insights from data. And from a utilitarian viewpoint, if you look at data science in industries, if you do proper data science, it allows these industries to make better decisions. These decisions could be in multiple fields; for example, companies could make better purchasing decisions, better hiring decisions, better decisions in terms of how to operate their processes, and so on.

So, when we talk about decisions, the decisions could be across multiple verticals in an industry. And data science is not only useful from an industrial perspective, but it is also useful in actual science as themselves. So, where you look at lots of data to model your system or test your hypotheses or theories about systems and so on. So, when we talk about data science, we start by assuming that we have a large amount of data for the problem of interest. And we are going to basically look at this data we are going to inspect the data; we are going to clean and curate the data then we will do some transformation of the data modeling and so on before we can derive insights that are valuable to the organization or to test a theory and so on.

(Refer Slide Time: 03:47)



The slide features a dark blue header with the text 'Data perspective' in white. Below the header is a list of five bullet points, each preceded by a blue dot. At the bottom of the slide, there is a video inset showing a man in a grey and white striped shirt speaking. The text 'Python for Data Science' is visible in the bottom left corner of the slide area.

- Read data
- Data processing and cleaning
- Summarizing data
- Visualization
- Deriving insights from data

Now, coming to a more practical view of what we do once we have data. I have these four bullet points, which roughly tell you, supposing you were solving a data science problem, what are the steps you will do? So, you will start with just having data someone gives you data; and you are trying to derive insights from this data. So, the very first step is really to bring this data into your system. So, you have to read the data. So, the data comes into this programming platform so that you can use this data. Now data could be in multiple formats so you could have data in a simple excel sheet or some other format.

So, we will teach you how to pull data into your programming platform from multiple data formats. So, that is the first step, really. If you think about how you are going to solve a problem, these steps would be first to simply read the data. And then, once you read the data many times, you have to do some processing with this data; you could have data that is not correct. For example, we all know that if you have your mobile numbers, there are 10 numbers in a mobile number, and if there is a column of mobile numbers and then say there is one row where there are just five numbers, then you know there is something wrong. So, this is a very simple check I am talking about in real data processing; this gets much more complicated.

So, once you bring the data in when you try to process this data, you are going to get errors such as this. So, how do you remove such errors? How do you clean the data? It is one

activity that usually precedes doing you more useful stuff with the data. This is not the only issue that we look at there could be data that is missing.

So, for example, there is a variable for which you get a value in multiple situations, but in some situations, the value is missing. So, what do you do with this data do you throw the record away? Or you do something to fill that data and so on. So, these are all data processing cleaning steps. So, in this course, we will tell you the tools that are available in python so that you can do this data processing cleaning and so on.

Now what you have done at this point is you have been able to get the data into the system, you have been able to process and clean the data and get to a certain data file or data structure that is complete so that you think you can work with this data set at which point what you will do is you will try to summarize this data. And usually, summarization of this data, a very simple technique would be very very simple statistical measures that you will compute; you could, for example, compute a median, mode, mean of a particular column.

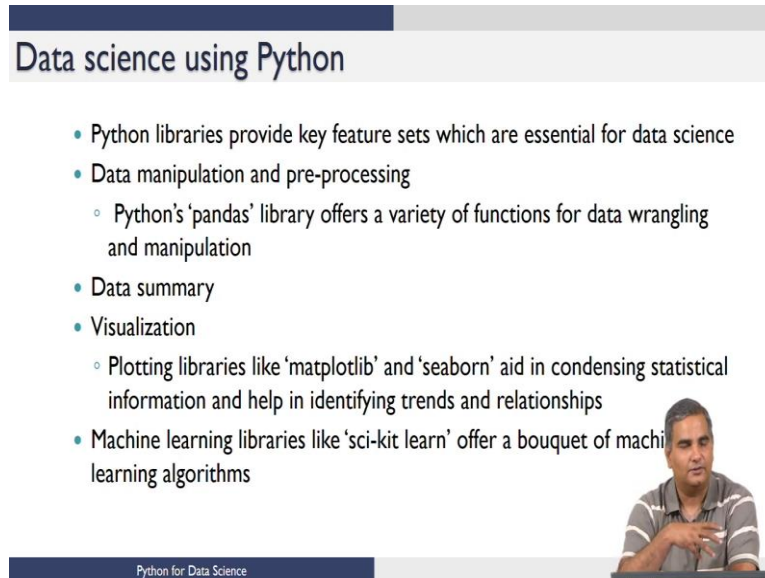
So, those are simple ideas or summarizing the data you could compute variance and so on. So, we are going to teach you how to use these notions of statistical quantities that you can use to summarize the data. Once you summarize the data, then another activity that is usually taken up is what is called visualization. So, visualization means you look at this data and more pictorially to get insights about the data before you bring in heavy-duty algorithms to bear on this data. And this is a creative aspect of data science; the same data could be visualized by multiple people in multiple ways. And some visualizations are not only eye-catching but are also much more informative than other types of visualization.

So, this notion of plotting this data so that some of the attributes or aspects of the data are made apparent is this notion of visualization. And there are tools in python that will teach you in terms of how you visualize this data. So, at this point, you have taken the data, you have cleaned the data, got a set of data points or data structure that you can work with, you have done some basic summary of this data that gives you some insights. You also looked at it more visually, and you have got some more insights, but when you have a large amount of big data, the last step is really deriving those insights which are not readily apparent either through visualization or through a simple summary of data.

So, how do we then go and look at more sophisticated analytics or analysis of data so that these insights come out? And that is where machine learning comes, and as a part of this

course when you see the progress of this course, you will notice that you will go through all of this so that you are ready to look at data science problems in a structured format and then use python as a tool to solve some of these problems.

(Refer Slide Time: 08:57)



Data science using Python

- Python libraries provide key feature sets which are essential for data science
- Data manipulation and pre-processing
 - Python's 'pandas' library offers a variety of functions for data wrangling and manipulation
- Data summary
- Visualization
 - Plotting libraries like 'matplotlib' and 'seaborn' aid in condensing statistical information and help in identifying trends and relationships
- Machine learning libraries like 'sci-kit learn' offer a bouquet of machine learning algorithms

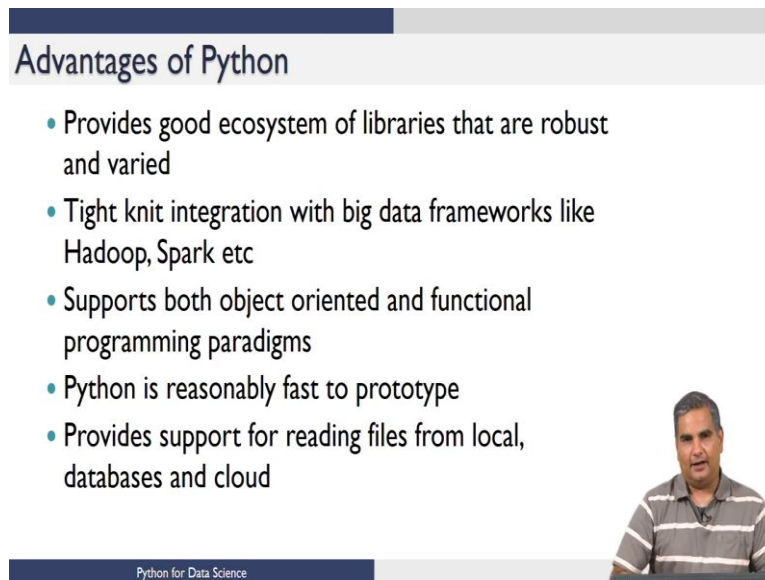
Python for Data Science

Now, why python for doing all of this? The number one reason is that there are these python libraries, which are already geared towards doing many of the things that we talked about so that it becomes easy for one to program, and very quickly, you can get some interesting outcomes out of whatever we are trying to do.

So, there are, as we talked about in the previous slide, you need to do data manipulation and pre-processing. There are lots of functions libraries in python where you can do data wrangling manipulation and so on. From a data summary viewpoint, there are many of these statistical calculations that you want to do are already pre-programmed, and you have to simply invoke them with your data to be able to show data summary. The next step we talked about visualization, there are libraries in python, which can be used to do the visualization.

And finally, for the more sophisticated analysis that we talked about all kinds of machine learning algorithms are already pre-coded available as libraries in python. So, again once you understand some bit about these functions and once you get comfortable working in python, then applying certain machine learning algorithms for these problems becomes trivial. So, you simply call these libraries and then run these algorithms.

(Refer Slide Time: 10:29)



The slide is titled "Advantages of Python" and features a list of five bullet points. To the right of the text is a small video inset showing a man in a striped shirt. At the bottom left of the slide, the text "Python for Data Science" is visible.

- Provides good ecosystem of libraries that are robust and varied
- Tight knit integration with big data frameworks like Hadoop, Spark etc
- Supports both object oriented and functional programming paradigms
- Python is reasonably fast to prototype
- Provides support for reading files from local, databases and cloud

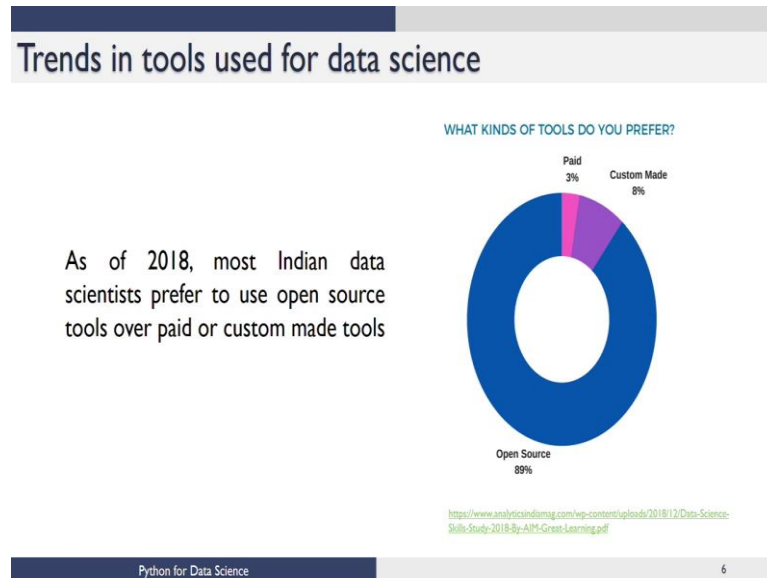
In the previous slide, we talked about the flow process for how I get the data in clean it and all the way up to insights, and then parallelly, we said why python makes it easy for us to do all of this. If you go back if you go forward a little more and then ask in terms of the other advantages of python, which are little more than just very simple data science activities. Python provides you several libraries, and it's continuously improved so, anytime there is a new algorithm that is coming into the set of libraries. So, in that sense, it's very varied, and there is also a good user community.

So, if there are some issues with new libraries and so on and those are fixed so that you get a robust library to work with and we talk about data, and data can be of different scale. So, the examples that you will see in this course are data of reasonably small size, but in real-life problems, you are going to look at data that is much larger, which we call big data. So, python has an ability to integrate with big data frameworks like Hadoop spark and so on.

And python also allows you to do more sophisticated programming object-oriented programming and functional programming. Python, with all of these sophisticated tools and abilities, is still reasonably a simple language to learn its reasonably fast to prototype. And it also gives you the ability to work with data which is in your local machine or in a cloud and so on. So, these are all things that one looks for when one looks at a programming platform that is capable of solving problems in real life right.

So, these are real problems that you can solve; these are not only toy examples, but real applications that you can build data science applications that you can build with python.

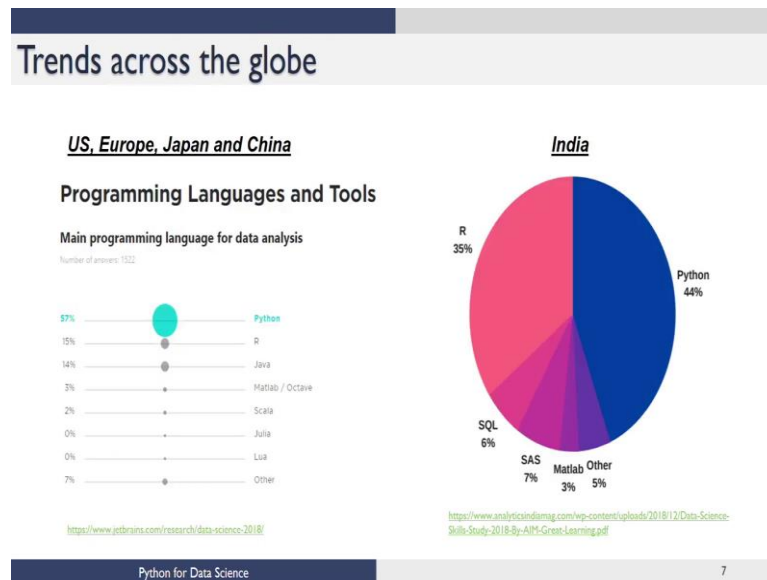
(Refer Slide Time: 12:49)



And just as another pointer in terms of why we believe that python is something that a lot of our students and professionals in India should learn. As you know, there are tools which are paid tools for machine learning with all of these libraries, and so on.

And there are also open-source tools and in India, based on a survey, most people, of course, prefer open-source tools for a variety of reasons cause being one because its free to use. But also if it is just free to use, but it does not have a robust user community, then it's not really very useful; that is where python really scores in terms of a robust user community, which can help with people working in python. So, it is both open-source, and there is a robust user community, both of which are advantageous for python.

(Refer Slide Time: 13:48).



And if you think of other competing languages for machine learning, if you look at this chart in India, about 44 percent of the people who were surveyed said they use python, or they prefer python. And of course, a close second is R. In fact, R was much more preferred a few years back, but over the last few years in India, a python is starting to become the programming platform of choice. So, in that sense, its a good language to learn because of the opportunities for jobs and so on or a lot more when you are comfortable with python as a language.

So, with this, I will stop this brief introduction on why python for data science. I hope I have given you an idea of the fact that while we are going to teach you Python as a programming language, please keep in mind that each module that we teach in this is actually geared towards data science. So, as we teach python, we will make the connections to how you will use some of the things that you see in data science; and all of this, we will culminate with these two case studies that will bring all of these ideas together. In terms of both are giving you an idea and an understanding of how the data science problem will be solved and also how it will be solved in python, which is a program of choice currently in India.

So, I hope this short four-week course helps you quickly get on to this programming platform. And then, learn data science, and then, you can enhance your skills with a much

more detailed understanding of both the programming language and data science techniques.

Thank you.