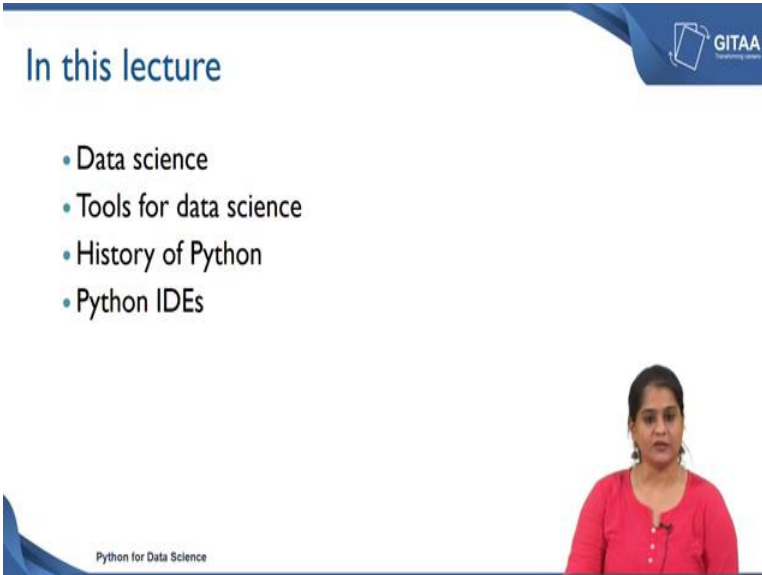


Python for Data Science
Prof. Ragnathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 02
Introduction to Python

Welcome to the lecture of Introduction to Python.

(Refer Slide Time: 00:17)



The slide features a blue header with the text "In this lecture" and a logo for "GITAA" (Global Institute of Technology and Advanced Research) in the top right corner. Below the header is a bulleted list of topics: "Data science", "Tools for data science", "History of Python", and "Python IDEs". In the bottom right corner, there is a video inset showing a woman with dark hair, wearing a red top, speaking. The text "Python for Data Science" is visible in the bottom left corner of the slide.

- Data science
- Tools for data science
- History of Python
- Python IDEs

In this lecture we are going to see what data science is in brief, we are also going to look at what are the commonly used tools for data science, we will also look at the history of python and followed by that will look at what an IDE means.

(Refer Slide Time: 00:33)



The slide is titled "Introduction" and features a blue header with the GITAA logo. The main content is a bulleted list of data sources. A woman in a red top is visible in the bottom right corner of the slide frame. The text "Python for Data Science" is at the bottom left.

- We live in a world that's drowning in data
- Data is generated from various sources
 - Websites track every user's every click
 - Your smartphone is building up a record of your location
 - Sensors from electronic devices record real time information
 - E-commerce websites collect purchasing habits

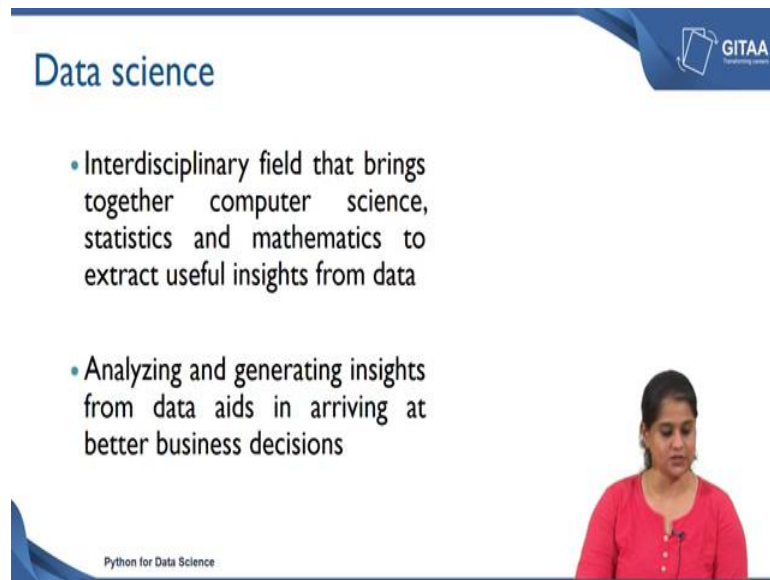
Python for Data Science

So, we live in a world that is drowning with data, wherever you go wherever we are data is getting generated from various sources.

Now, when you browse through few websites, the websites track every users click and this forms a part of web analytics. The other instance of where data gets generated is when you use a smartphone; you are basically building up a record of your location. So, all these information go and sit somewhere and get collected in the form of data.

We also have sensors from electronic devices that record real time information and you also have e-commerce website that collect purchasing habits. So, whenever you log into any of these e-commerce sites, you will see some recommendations based on your previous purchase history or previous view history.

(Refer Slide Time: 01:19)

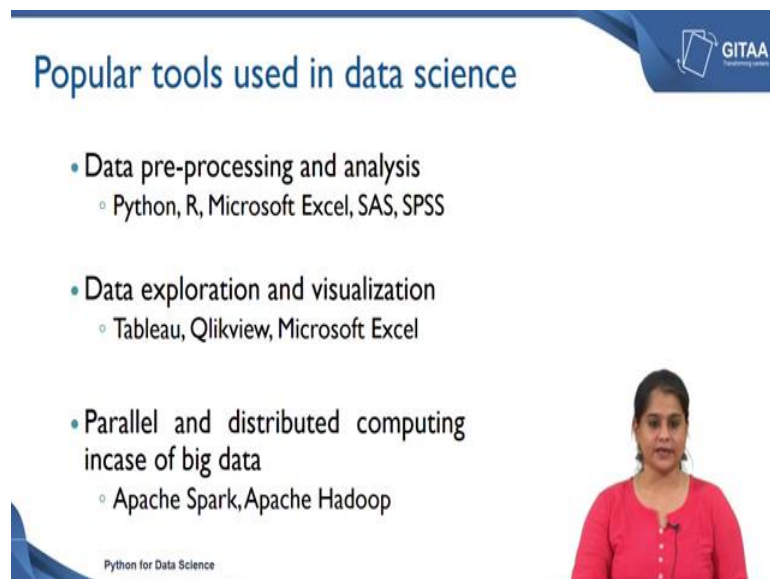


The slide is titled "Data science" and features the GITAA logo in the top right corner. It contains two bullet points: "• Interdisciplinary field that brings together computer science, statistics and mathematics to extract useful insights from data" and "• Analyzing and generating insights from data aids in arriving at better business decisions". A presenter in a red top is visible in the bottom right corner. The text "Python for Data Science" is at the bottom left.

- Interdisciplinary field that brings together computer science, statistics and mathematics to extract useful insights from data
- Analyzing and generating insights from data aids in arriving at better business decisions

So, now let us see what data science is all about. So, it is an interdisciplinary field that brings together computer science, statistics and mathematics to get useful inferences and insights from a data. Now these insights are very crucial from a business perspective because, it will help you in making better business decisions.

(Refer Slide Time: 01:40)



The slide is titled "Popular tools used in data science" and features the GITAA logo in the top right corner. It contains three bullet points: "• Data pre-processing and analysis" (with sub-points "◦ Python, R, Microsoft Excel, SAS, SPSS"), "• Data exploration and visualization" (with sub-points "◦ Tableau, Qlikview, Microsoft Excel"), and "• Parallel and distributed computing incase of big data" (with sub-points "◦ Apache Spark, Apache Hadoop"). A presenter in a red top is visible in the bottom right corner. The text "Python for Data Science" is at the bottom left.

- Data pre-processing and analysis
 - Python, R, Microsoft Excel, SAS, SPSS
- Data exploration and visualization
 - Tableau, Qlikview, Microsoft Excel
- Parallel and distributed computing incase of big data
 - Apache Spark, Apache Hadoop

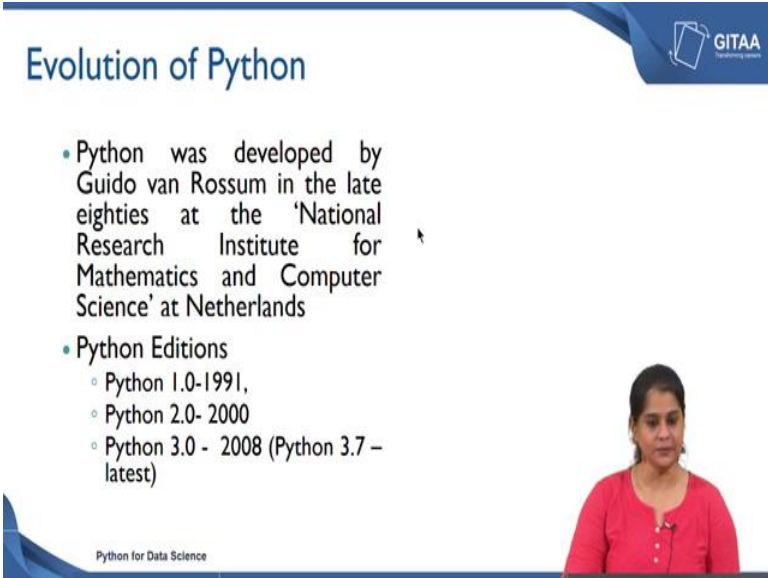
Now, currently there are many tools that are being used in data science; now these tools can be bucketed into 3 categories. The first category is where you are going to be looking at data preprocessing and analysis, now tools and software's that fall under this category

are python, R, MS Excel, SAS and SPSS. Now all these tools are required for you to preprocess and analyze the data. So, apart from data preprocessing and analysis, there is a fair share of effort that is given for data exploration and visualization and these are done even before you analyze your data.

Now, the commonly used data exploration and visualization tools are Tableau, Qlikview and of course, you always have your MS Excel. So, the next bucket that we are going to look into is when you have huge chunks of data, now when your collecting data on a real time basis you are going to be collecting data over every second every minute. Now if you want to store all these data and preprocesses it the regular desktop or computing systems that you have might not be useful.

So, that is when you use parallel or distributed computing, where you distribute the work across different systems popular tools that are being used for big data Apache Spark and Apache Hadoop. So, in this course we are going to be mainly focusing on tools that are required for data preprocessing and analysis and in specific we are going to look into python.

(Refer Slide Time: 03:08)



The slide features a blue header with the title "Evolution of Python" and the GITAA logo. The main content is a bulleted list. In the bottom right corner, there is a video inset of a woman in a red top. The footer contains the text "Python for Data Science".

Evolution of Python

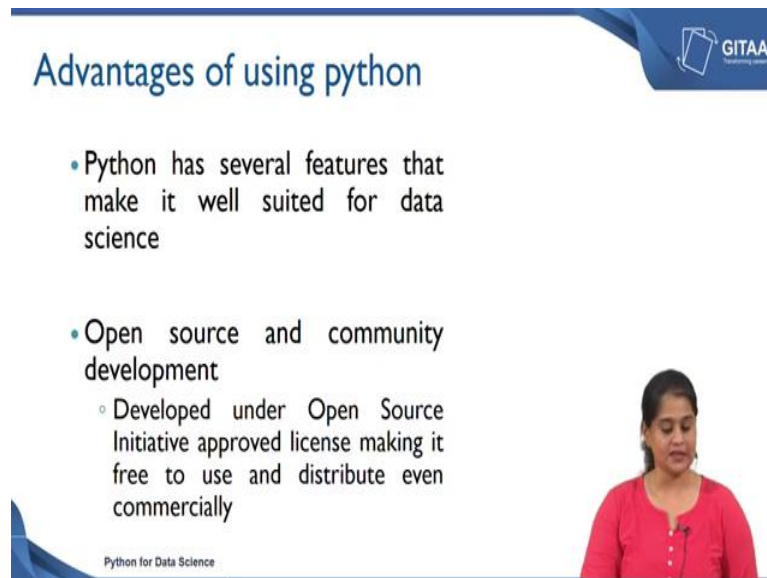
- Python was developed by Guido van Rossum in the late eighties at the 'National Research Institute for Mathematics and Computer Science' at Netherlands
- Python Editions
 - Python 1.0-1991,
 - Python 2.0- 2000
 - Python 3.0 - 2008 (Python 3.7 – latest)

Python for Data Science

So, let us look at the evolution of python. So, python was developed by Guido van Rossum in the late eighties at the national research institute for mathematics and computer science and this institute is located at Netherlands.

So, there are different versions of python, the first version that it was released was in 1991; the second version was released in 2000 and the third version was released in 2008 with version 3.7 being the latest. So, let us look at the advantages of using python.

(Refer Slide Time: 03:41)



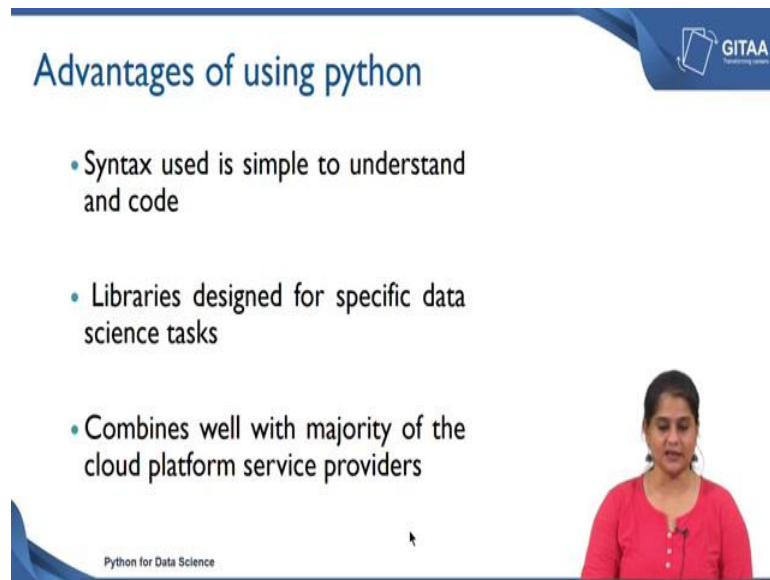
The slide is titled "Advantages of using python" and features a blue header with the GITAA logo. The content is organized into a bulleted list. A presenter in a red top is visible in the bottom right corner of the slide frame.

- Python has several features that make it well suited for data science
- Open source and community development
 - Developed under Open Source Initiative approved license making it free to use and distribute even commercially

Python for Data Science

So, python has features that make it well suited for data science. So, let us look at what these features are. So, the first and foremost feature of python is that it is an open source tool and python community provides immense support and development to its users. So, python was developed under the open source initiative approved license thereby making it free to use and distribute even if its for commercial purposes.

(Refer Slide Time: 04:05)



The slide features a blue header with the title 'Advantages of using python' and the GITAA logo. The main content is a bulleted list of three advantages. A presenter in a red top is visible in the bottom right corner of the slide frame. The text 'Python for Data Science' is located at the bottom left of the slide.

Advantages of using python

- Syntax used is simple to understand and code
- Libraries designed for specific data science tasks
- Combines well with majority of the cloud platform service providers

Python for Data Science

The next feature is that the syntax that python use fairly simple to understand and code and this breaks all kinds of programming barriers if you are going to switch to a newer programming language. So, the next important advantage of using python is that, the libraries which are contained in python get installed at the time of installation and these libraries are designed keeping in mind specific data science task and activities.

Python also integrates well with most of the cloud platform service providers; and this is a huge advantage if you are looking to use big data. So, if you are going to download python from the website and install it, you will see that most of the scripting is done in shell. So, there are applications that provide better graphical user interfaced for the end users and these are taken care by the integrated development environment.

(Refer Slide Time: 04:57)



Integrated development environment (IDE)

- Software application consisting of a cohesive unit of tools required for development
- Designed to simplify software development
- Utilities provided by IDEs include tools for managing, compiling, deploying and debugging software

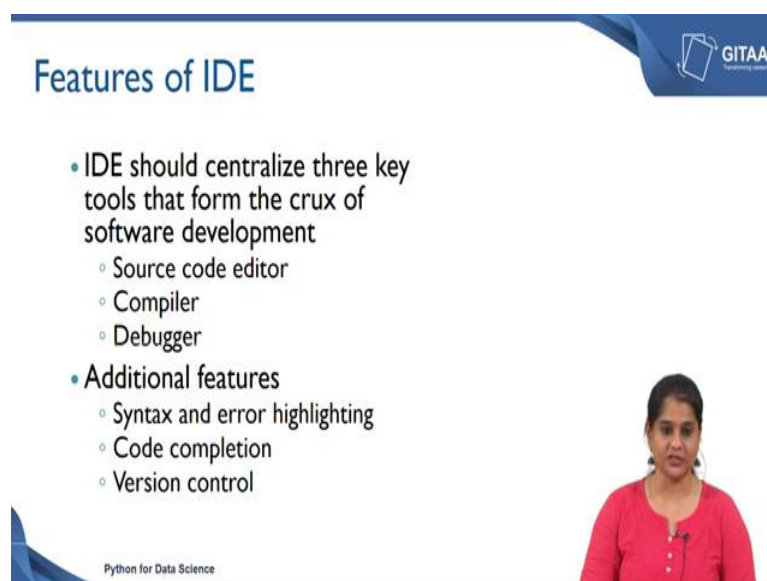
Python for Data Science

GITAA

The slide features a blue header with the title 'Integrated development environment (IDE)' and the GITAA logo. The main content is a bulleted list. A presenter in a red top is visible in the bottom right corner. The footer includes 'Python for Data Science' and the GITAA logo.

So, now, let us see what an integrated development environment is, an IDE as how its abbreviated is a software application and it consists of tools which are required for development. All these tools are consolidated and brought together under one roof inside the application. IDEs are also designed to simplify the software development this is very useful because as an end user, if you are not a developer you might want all the tools available at a single click. Using an IDE will be very beneficial in that case also the features provided by IDEs include tools for managing, compiling, deploying and debugging a software. So, these also form the core features of any IDEs.

(Refer Slide Time: 05:44)



Features of IDE

- IDE should centralize three key tools that form the crux of software development
 - Source code editor
 - Compiler
 - Debugger
- Additional features
 - Syntax and error highlighting
 - Code completion
 - Version control

Python for Data Science

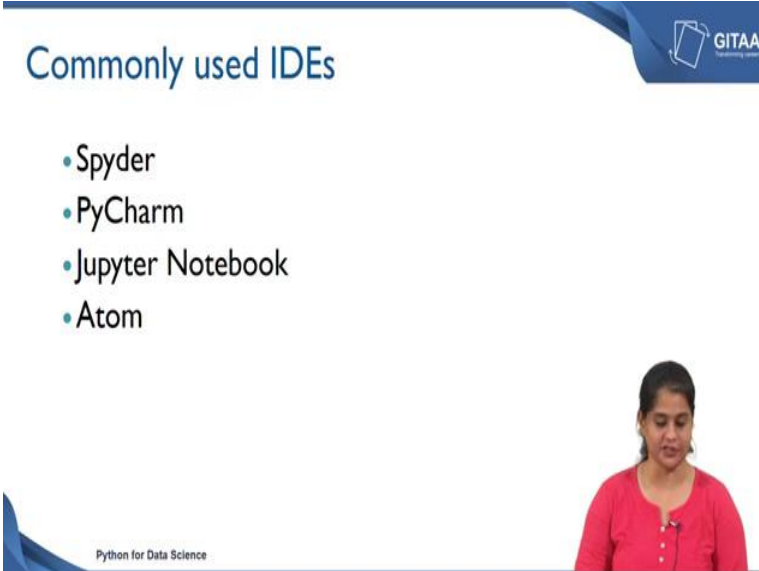
GITAA

The slide features a blue header with the title 'Features of IDE' and the GITAA logo. The main content is a bulleted list with sub-bullets. A presenter in a red top is visible in the bottom right corner. The footer includes 'Python for Data Science' and the GITAA logo.

So, now let us look at what are the features of an IDE in depth. So, any IDE should consist of three important features; the first is the source code or text editor, the second is a compiler and the third is a debugger. Now all these three features form the crux of any software development.

The IDEs can also have additional features like syntax and error highlighting, code completion and version control.

(Refer Slide Time: 06:09)

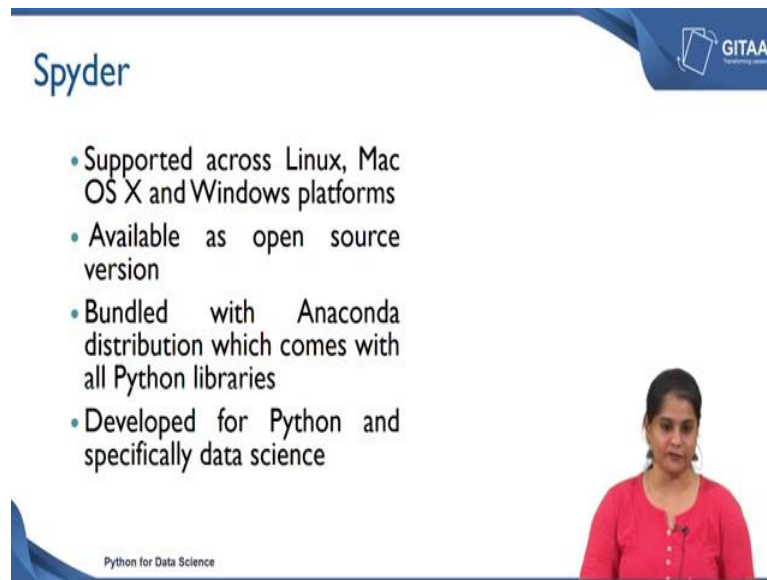


The slide is titled "Commonly used IDEs" and features a blue header with the GITAA logo. Below the title, a bulleted list identifies four IDEs: Spyder, PyCharm, Jupyter Notebook, and Atom. In the bottom right corner, a woman in a red shirt is visible, likely the presenter. The bottom left corner of the slide contains the text "Python for Data Science".

- Spyder
- PyCharm
- Jupyter Notebook
- Atom

So, let us see what are the commonly used IDEs for python, the most frequently used as spyder, PyCharm, Jupyter Notebook and Atom. And these are basically from the endpoint of the user, depending on what he or she is comfortable with.

(Refer Slide Time: 06:24)



Spyder

- Supported across Linux, Mac OS X and Windows platforms
- Available as open source version
- Bundled with Anaconda distribution which comes with all Python libraries
- Developed for Python and specifically data science

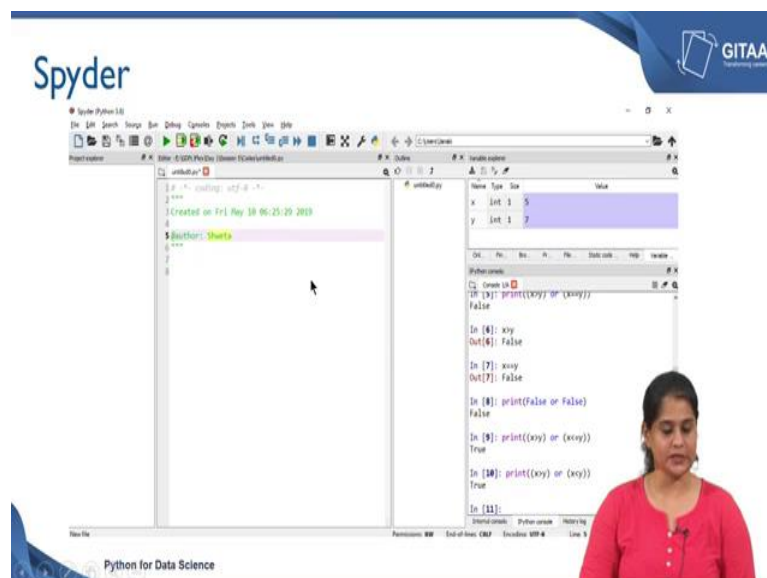
Python for Data Science

The slide features a blue header with the 'Spyder' logo and the 'GITAA' logo. A presenter in a red top is visible in the bottom right corner.

So, now let us look at spyder, the spyder is an IDE and it supported across Linux, Mac and Windows platforms. It is also an open source software and it is bundled up with Anaconda distribution which comes up with all inbuilt python libraries.

So, if you want to work with spyder you do not have to install any of the libraries. So, all the necessary libraries are taken care by Anaconda. So, another important feature of spyder is that it was specifically developed for data science and it was developed in python and for python.

(Refer Slide Time: 06:57)



Spyder

```
#!/usr/bin/env python
# coding: utf-8
"""
Created on Fri May 18 06:25:29 2018
@author: shweta
"""
x = 5
y = 7
```

Name	Type	Size	Value
x	int	5	5
y	int	7	7

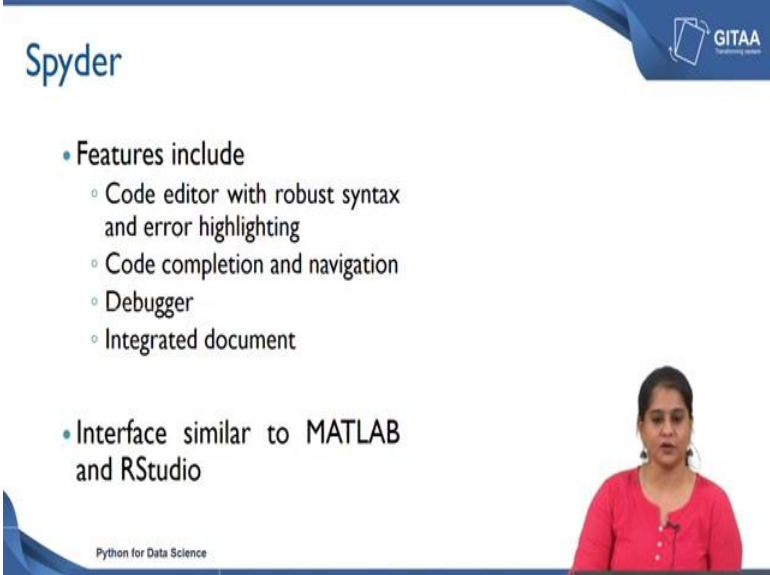
```
In [6]: x*y
Out[6]: False
In [7]: x==y
Out[7]: False
In [8]: print(False or False)
False
In [9]: print((x*y) or (x==y))
True
In [10]: print((x*y) or (x==y))
True
In [11]:
```

Python for Data Science

The screenshot shows the Spyder IDE interface with a code editor, a variable explorer, and a console. A presenter in a red top is visible in the bottom right corner.

So, this is how the interface of spyder looks, you have the scripting window and you have other console output here, you have a variable explorer here. All these features we are going to be looking at in the next few lectures to come.

(Refer Slide Time: 07:11)



The slide features a blue header with the 'Spyder' logo on the left and the 'GITAA' logo on the right. The main content is a bulleted list of features. At the bottom right, there is a video inset of a woman in a red top. The footer contains the text 'Python for Data Science'.

- Features include
 - Code editor with robust syntax and error highlighting
 - Code completion and navigation
 - Debugger
 - Integrated document
- Interface similar to MATLAB and RStudio

Python for Data Science

The other features of spyder includes a code editor, with robust syntax error highlighting features; it also helps in code completion and navigation it consist of a debugger, it also consist of an integrated documents that can be viewed within the python interface on the web. Another advantage of using spyder is that it has a interface which is very similar to MATLAB and RStudio's. So, if you are a person who is already work with these two programming languages and are looking to switch to python, then the transition is also going to be seamless.

(Refer Slide Time: 07:44)



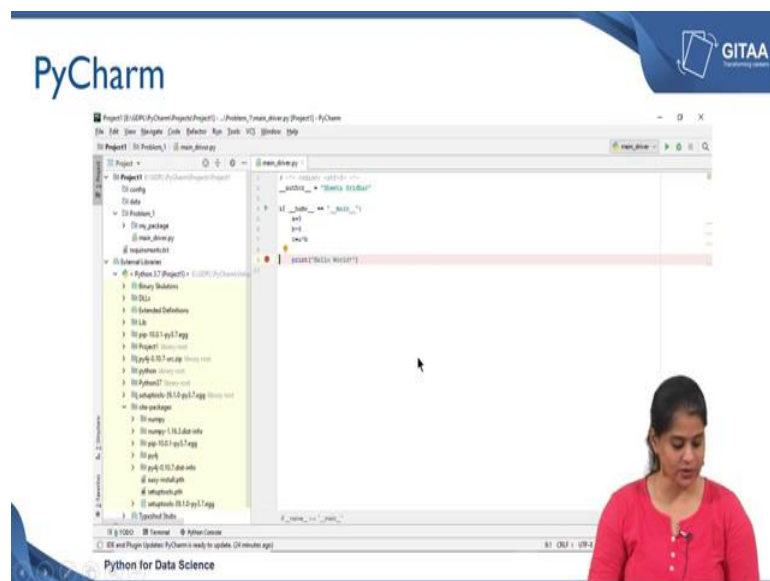
The slide features the PyCharm logo at the top left and the GITAA logo at the top right. A presenter in a red top is visible in the bottom right corner. The slide content is as follows:

- Supported across Linux, Mac OS X and Windows platforms
- Available as community (free open source) and professional (paid) version
- Supports only Python
- Bundled with Anaconda distribution which comes with all Python libraries
 - Can also be installed separately

Python for Data Science

So, now let us look at the second IDE which is pyCharm. So, pyCharm is also supported across all OS systems which is Linux, Mac and windows. It has two versions to it one is the community version which is an open source software; the other is the professional version which is a paid software. So, pyCharm supports only python and it is bundled up and packaged with Anaconda distribution which comes with all the inbuilt python libraries. But; however, if you want to install pyCharm separately then that can also be done.

(Refer Slide Time: 08:14)



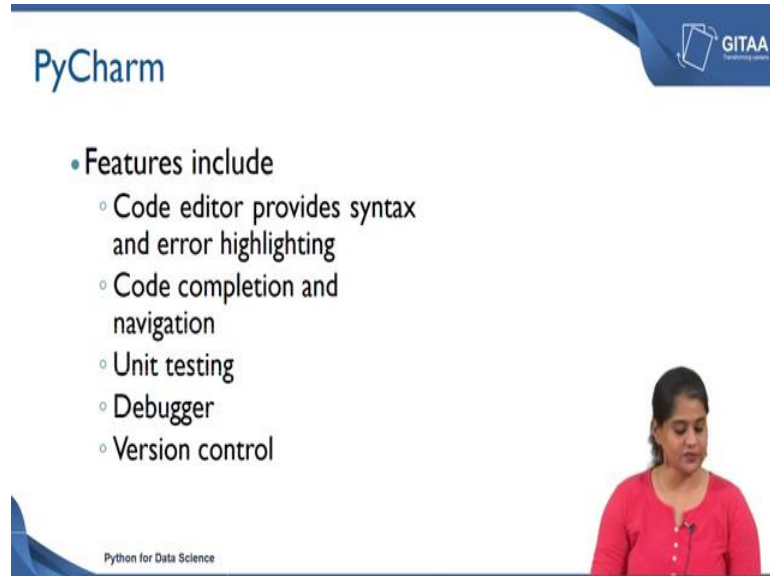
The slide shows a screenshot of the PyCharm IDE interface. The presenter in the red top is visible in the bottom right corner. The IDE window displays a project structure on the left and a code editor on the right. The code editor shows a Python script with the following content:

```
if __name__ == '__main__':  
    print("Hello World!")
```

Python for Data Science

So, this is how the interface of pyCharm looks, you have a very very well define structure for naming your directories and you have the scripting window here.

(Refer Slide Time: 08:25)



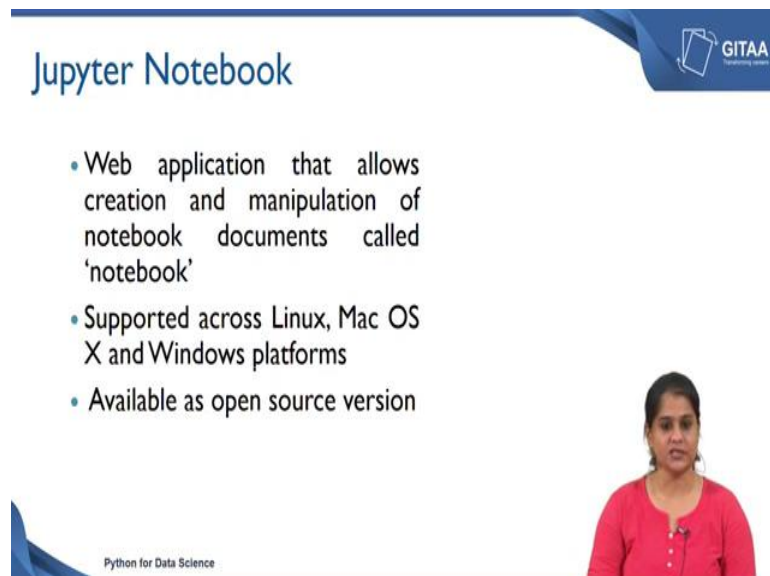
The slide features a blue header with the 'PyCharm' logo on the left and the 'GITAA' logo on the right. The main content is a bulleted list of features. A woman in a red top is visible in the bottom right corner of the slide frame.

- Features include
 - Code editor provides syntax and error highlighting
 - Code completion and navigation
 - Unit testing
 - Debugger
 - Version control

Python for Data Science

So, let us look at some of the features that pyCharm consists of. The first is that it consists of a code editor which provides syntax and error highlighting; then it consists of a code completion and navigation feature it also consists of a unit testing tool which will help the compiler go through each and every line of the code. It also consists of a debugger and controls the versions.

(Refer Slide Time: 08:48)



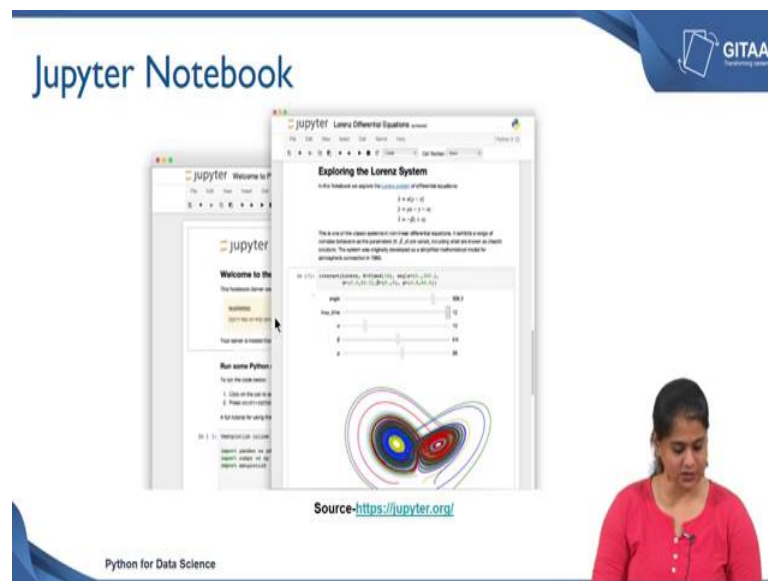
The slide features a blue header with the 'Jupyter Notebook' title on the left and the 'GITAA' logo on the right. The main content is a bulleted list of features. A woman in a red top is visible in the bottom right corner of the slide frame.

- Web application that allows creation and manipulation of notebook documents called 'notebook'
- Supported across Linux, Mac OS X and Windows platforms
- Available as open source version

Python for Data Science

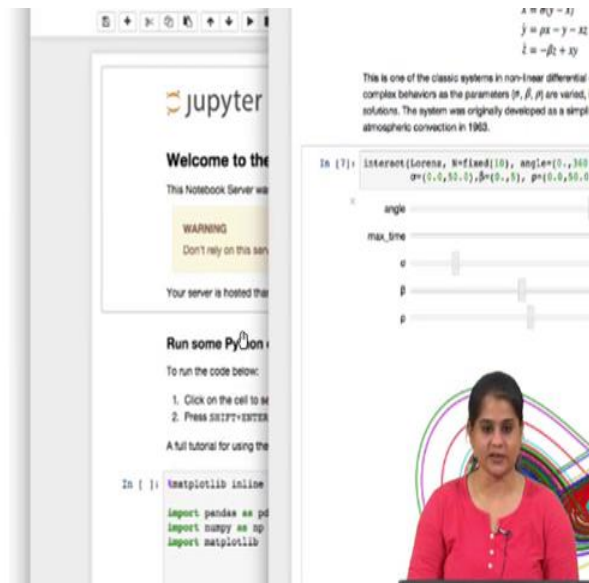
So, now let us look at the next IDE which is Jupyter notebook. So, now, Jupyter notebook is very different from the earlier two IDEs in the sense that it is a web application which allows creation and manipulation of the codes; now these codes are called notebook documents and hence that is how Jupyter gets its name Jupyter notebook. Now Jupyter is supported across all operating systems and it is available as an open source version.

(Refer Slide Time: 09:18)



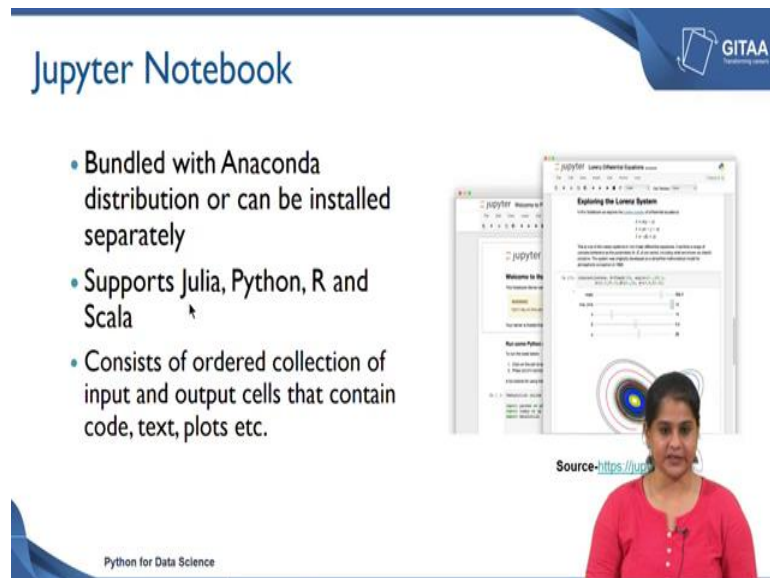
Now, this is the interface of Jupyter, you can see that you have few cells here as an input you also have some output let me just zoom in and show you how the interface looks.

(Refer Slide Time: 09:33)



So, here you can see some of the codes that is written, if you just scroll up and see this is some narrative about whatever you have written.

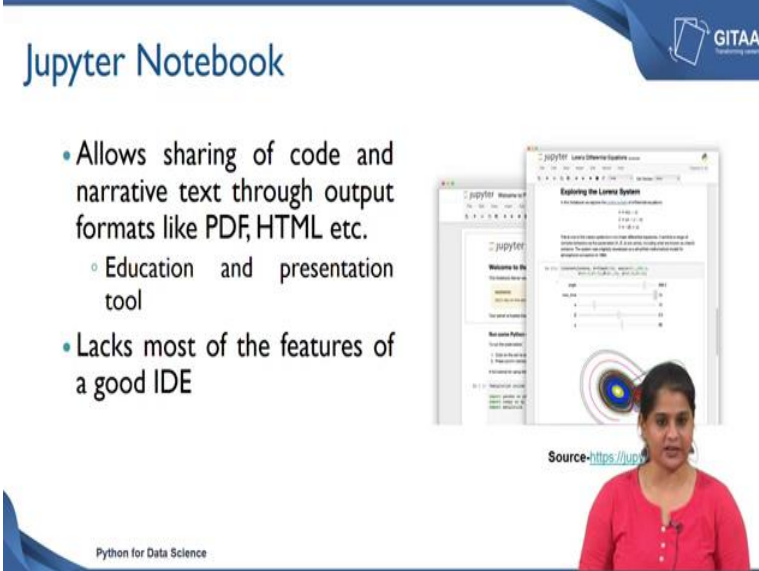
(Refer Slide Time: 09:41)



So, Jupyter is bundled with Anaconda distribution, but it can also be install separately. It primarily supports Julia, python, R and Scala. So, if you look at the name Jupyter it basically takes the first two letters from Julia the next two from python and then R.

So, that is how Jupyter gets its name as Jupyter it also consists of an ordered collection of input and output cells like how we earlier saw; and these can contain narrative text, code, plots and any kind of media.

(Refer Slide Time: 10:13)



Jupyter Notebook

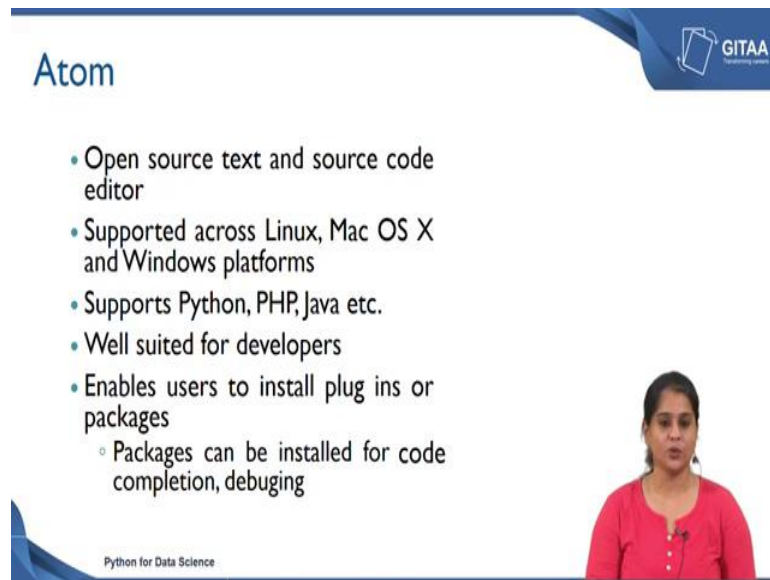
- Allows sharing of code and narrative text through output formats like PDF, HTML etc.
 - Education and presentation tool
- Lacks most of the features of a good IDE

Source: <https://jupyter.org/>

Python for Data Science

One of the key features of Jupyter notebook is that, it allows sharing of code and narrative text through output formats like HTML markdown or PDF. If you are working in an education environment or if you would like to have a better presentation tool, then you can use these kind of output formats to present. So, though Jupyter consist of features that give a very good aesthetic appeal to it, it is deficit of the important features of a good IDE. So, by good IDE, I mean it should consist of a source code editor and compiler and a debugger; and all three of these are not provided by Jupyter.

(Refer Slide Time: 10:50)



The slide features a blue header with the word 'Atom' in white. In the top right corner, there is a logo for 'GITAA' with the tagline 'GROWING TOGETHER'. The main content is a bulleted list of features for Atom. At the bottom right, a woman in a red top is visible, likely the presenter. The bottom left corner of the slide has the text 'Python for Data Science'.

Atom

- Open source text and source code editor
- Supported across Linux, Mac OS X and Windows platforms
- Supports Python, PHP, Java etc.
- Well suited for developers
- Enables users to install plug ins or packages
 - Packages can be installed for code completion, debugging

Python for Data Science

So, the next IDE that we are going to look into is atom. So, atom is an open source text and source code editor and is supported again across all over systems; it again supports programming languages like python, PHP, Java etc. And it is very very well suited for developers, it also helps the users to install plug-ins or packages. So, one common drawback with all these text editors and source code editor is that these do not come installed with basic libraries of any programming languages; you have to install these kind of packages as and when you have a need for them.

So, that is one major drawback for using any kind of text editor or the source code editor. But; however, atom does provide packages or libraries that are suited for data science and code completion or code navigation or debugging. So, you can install it, so if you are a developer and if you want to code an text editor environment then you can go ahead with atom. But you will have to install all these packages as and when you require.

(Refer Slide Time: 11:52)

Atom

```
Project
├── .git
├── .io
├── .buffer-binding.js
├── .editor-binding.js
├── .event-portal-binding.js
├── .file-portal-binding.js
├── .normalize-url.js
├── real-time-package.js
├── .node_modules
├── .script
├── .styles
├── .test
├── .gitignore
├── .travis.yml
├── index.js
├── package-lock.json
├── package.json
└── README.md
```

```
1  const {CompositeDisposable} = require('a
2  const {allowUnsafeFunction} = require
3
4  let Client
5  allowUnsafeFunction(() => { Client })
6
7  const BufferBinding = require('./buffer-
8  const EditorBinding = require('./edito
9
10
11  module.exports =
12  class RealTimePackage {
13    constructor (options) {
14      cons
15
```

Source: <https://atom.io/>

Python for Data Science

So, this is the interface of atom, this is how it looks, it is a proper text editor interface.

(Refer Slide Time: 12:00)

How to choose the best IDE?

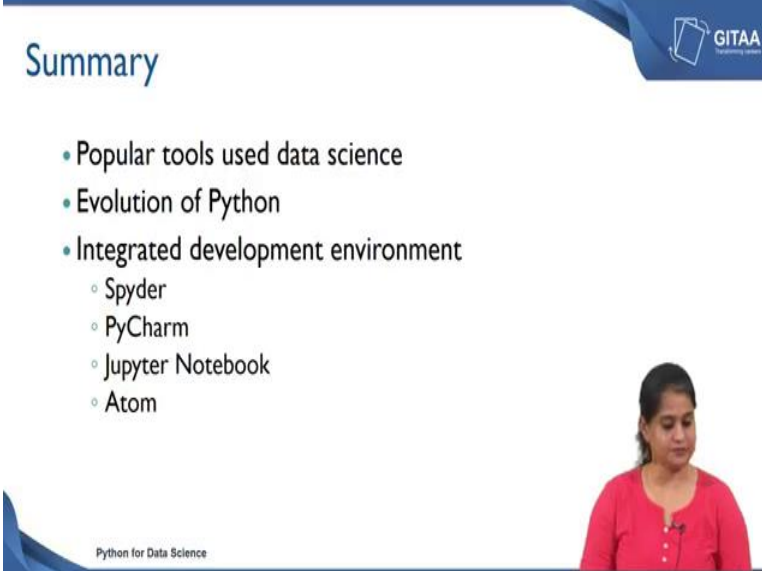
- Requirements
- Working with different IDEs helps us understand our own requirement
- In this course, Spyder will be used

Python for Data Science

So, how will you choose the best IDEs then important question. So, it basically depends on your requirements, but it is a good habit to work first with different IDEs to understand what your own requirements are. So, if you are new to python then it is better that you work across all these IDEs and there are several other IDEs out there you can work with all these IDEs see what suits you and then take a call on which IDE to use.

But in this course we are going to be looking at spyder; and that is primarily because it is a very good software that has been developed only for data science and python; and it as an interface that is very very appealing and easy to use for beginners.

(Refer Slide Time: 12:43)



Summary

- Popular tools used data science
- Evolution of Python
- Integrated development environment
 - Spyder
 - PyCharm
 - Jupyter Notebook
 - Atom

Python for Data Science

So, to summarize in this lecture we saw what are the popular tools used in data science environment. We also saw how python evolved and what are the commonly used integrated development environment. We also looked at what each of these IDE have to offer us and some of the common pros and cons of each of these.

Thank you.