An Introduction of Proteogenomics

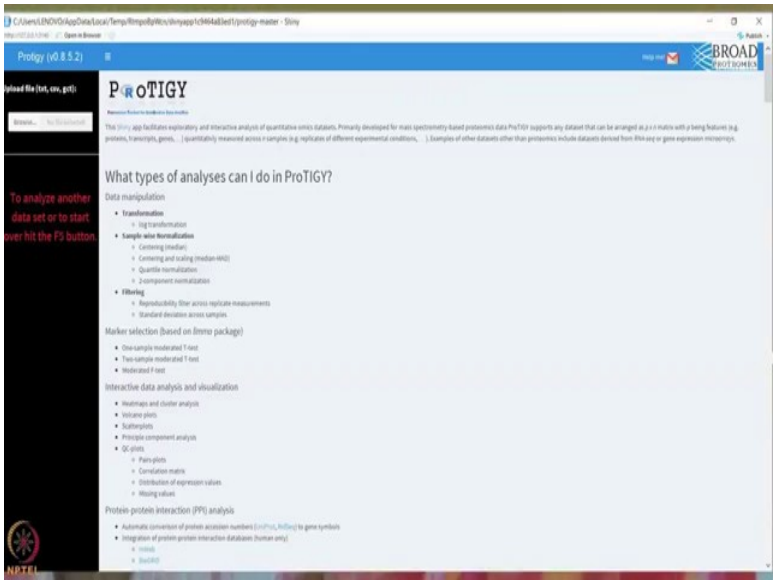
Dr. Sanjeeva Srivastava
Dr. D. R. Mani
Prof. Deeptarup Biswas
Department of Bioscience and Bioengineering
Principal Computational Scientist, Proteomics Platform
Indian Institute of Technology, Bombay
Broad Institute of MIT and Harvard, USA

Lecture - 23
Hands-on: Protigy I

Welcome to MOOC course on Introduction to Proteogenomics. To enable proteomics researchers, to interactively explore the acquired data matrices of quantified proteins or post-translational modifications and to facilitate an integrative set of analysis tools; the broad institute team has devolved a software known Protigy which is proteomics tools set for integrative data analysis. Protigy is primarily developed for the proteomics platform and now it is open access for the broader audience. Protigy streamlines the entire proteomic data analysis pipeline, provides an institutive interface for the lab researchers to analyze and explore proteomics data sets and ensure the reproducible data analysis by the keeping track of workflows and various parameters.

Today we have Dr. D R Mani who is going to conduct the Hands-on session on Protigy. In the first session he will primarily focus on installation of Protigy and explaining its various parameters. Next he will show how to load data sets and choose annotation. He will also demonstrate running the analysis and then exploring results. So, let us welcome Dr. D R Mani for today's hands on session on Protigy

(Refer Slide Time: 02:04)

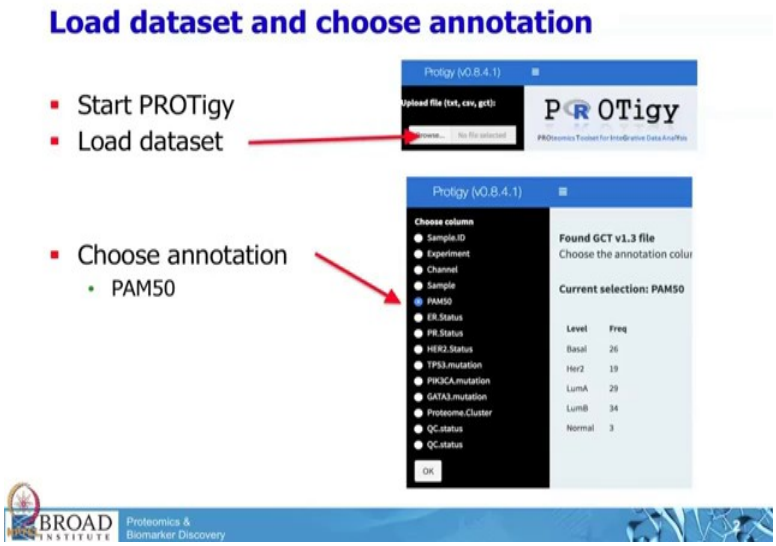


If you have installed Protigy and it went through completely you will have a page that looks like this in your web browser, so that is the tool. So, the history of this is I have talked about t tests and moderated t tests and F tests, like two component normalization and lot of these things that you would applied to pretty much any projects that you work on. So, previously in our group I had all these tools written up as R code and then people would ask me when they have a project to run it, and some of the more adventurous people would take my code and run it in R themselves and kind of work on their project.

And, then after Karsten join we figured that it would be a good idea to make a tool that it is easy to use, encapsulate all the algorithms that our group things that we should use on a regular basis and then make it available in such a way that anybody can use it. And, they would need to come to us only if their experiment or project did not fall into the standard set of tests and analysis that you would do. There are several lot of those, but large fraction of projects in our group kind of just use Protigy and then they have their results which they sent to the collaborators. And, they look a date and say which is biologically relevant or look at other follow up experiments and so forth.

So, this is the tool that we use at the broad, it is also freely available for anybody to use anywhere, its available on GitHubs. So, you can download it and use it which is what all of you are doing now. And, I will just quickly go through how to use the tool.

(Refer Slide Time: 03:51)



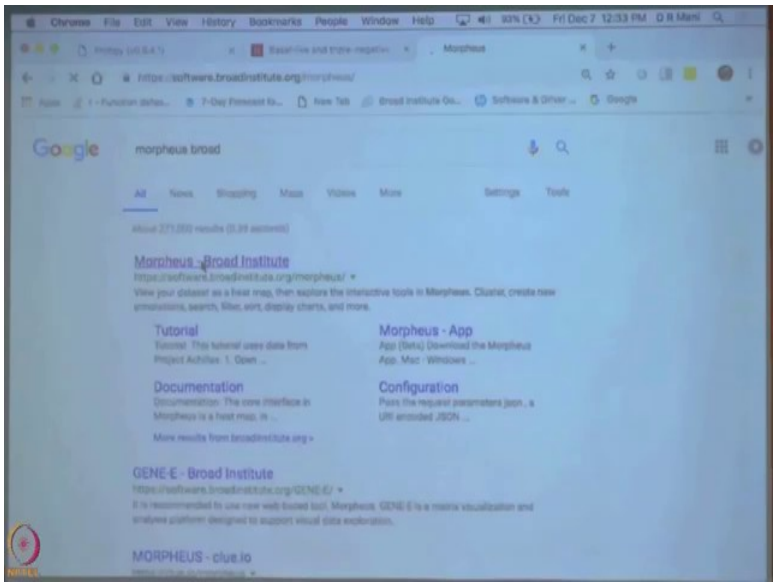
The first thing we want to do is load a dataset. So, in the hands on part on piazza there was a dataset that I had put up; if you can download it or get it or if you have already got it then we can try to load the dataset into Protigy. So, the first thing you need to do is load the dataset and choose any annotations that you want to use for your analysis and then you go and do the analysis. So, loading the dataset is the first part and very essential part, you cannot do an analysis without it. So, the most people have what we need?

Student: Can we use example dataset?

Yes. So, the dataset is the you can see the extension is .gct. So, gct is stands for Gene Clustered Text, it is a format that we came up at the broad. The cool thing about this is not only does this have the data table, it also has sample annotations and gene or protein annotations included in the table. So, you have the data and additional rows and columns that provide annotations. It is basically a text delimited file which you can open in excel, if you ignore the top two lines.

So, if you ask excel to open it as a text file you can see the table. The top two lines are description of how many annotation columns there are and how many data columns there are. So, that is only for software that reads it, but for you, you can ignore those two and just look at it. If you want to take your data and create a GCT file, there is a broad software called Morpheus.

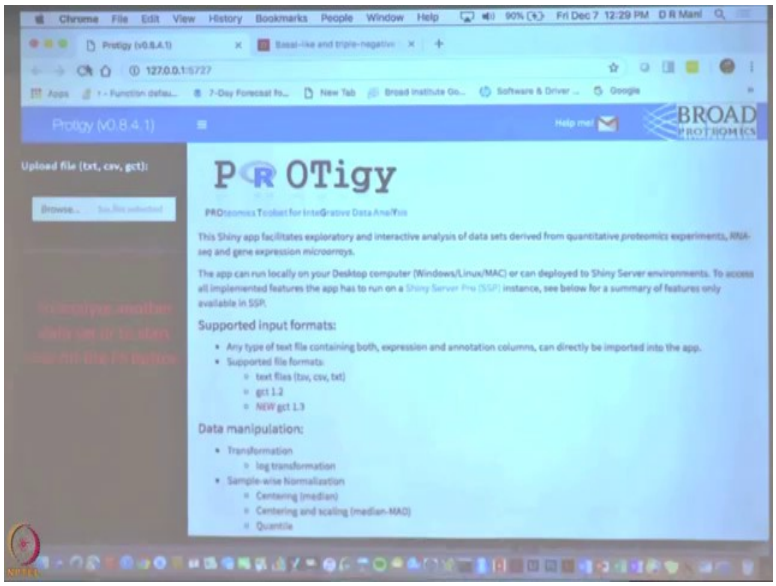
(Refer Slide Time: 05:37)



So, if you go to Google and search for Morpheus; now it is taking a while, but the thing is Morpheus can read a lot of different format. So, it can read comma separated text files, all kinds of formats and then write out GCT 1.3, and once you read it into Morpheus if you want you can add annotations. You can add annotations from a separate file or in a given file you can say these rows or, these columns are my annotations and then you can kind of get everything set up in Morpheus and you can export it as GCT 1.3 file and then you can read it in to Protigy.

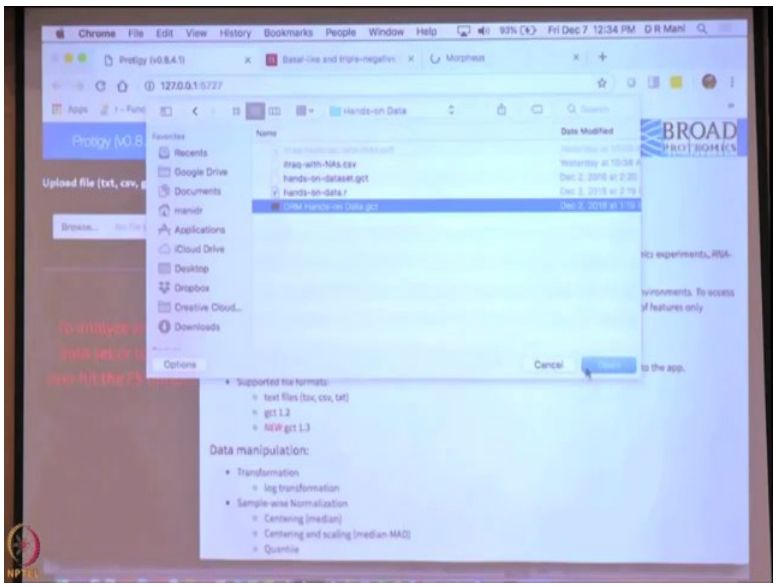
Protigy also can read other formats, but if you have GCT 1.3, it is easy to say which annotations you want to use, because the annotations are included in it; otherwise including annotations is a little more complex and I will not go in it today.

(Refer Slide Time: 06:33)



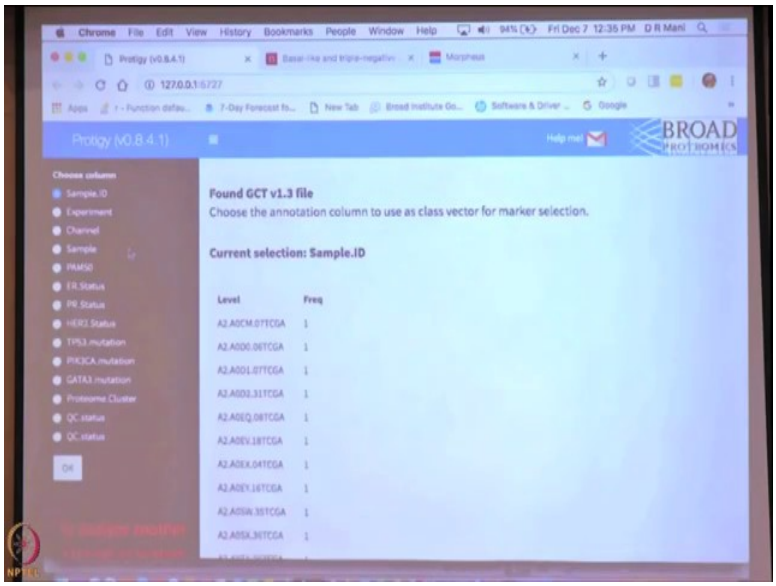
So, in the slides I have tried to pictorially display what we need to do, but I will actually do it on the screen. So, you click on browse on the left side, you will get a file browser.

(Refer Slide Time: 06:46)



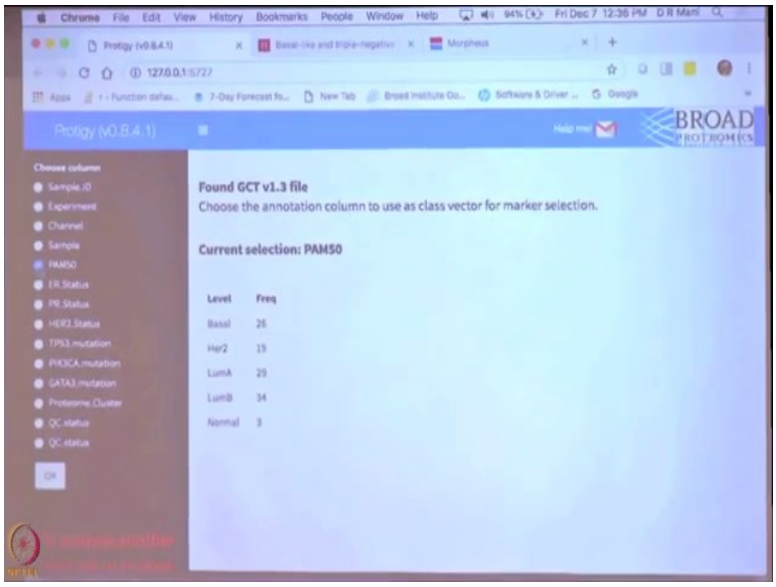
In the file browser you pick the data set that you want to use. So, I am going to pick a DRM hands on data and you just say open.

(Refer Slide Time: 07:00)



So, you will say upload complete and then just wait a couple of seconds you will get a new screen. So, it says what type of data file it form, say it is found GCT 1.3. It says what the samples are and how many. So, these are like the sample levels and its say how many times it found the sample. So, ideally you want sample levels to be unique and so you want the frequency to be 1.

Now, here on the left it showing all the columns, annotation columns that it found in the GCT 1.3 file. So, you can see there is the PAM50 status for the cancers, for samples ER PR HER2 status T53 mutation status and so forth, and there is also some other experimental detail. So, we have experiment number which is the TMT plex in which it was run. There is the channel which is the sorry this is all iTRAQ data, so it is a iTRAQ experiment and the iTRAQ channel that it was run. So, all those are included in the dataset. So, for this or kind of hands on let us pick PAM50 as the annotation that we want to use.(Refer Slide Time: 08:06)



So, when you pick PAM50 it shows you the various levels in PAM50 and how many samples there are. So, it says there are 26 basal samples, 19 HER2 samples and so on. You can also see we have three normal samples in there which were actually normal breast samples, that where included in the study and it shows that rows also.

So, I will just start over again. So, I think some people wanted me to repeat. So, this is the Protigy opening page, you click on browse. So, it looks like in, you have to wait for the full file to download. If you do not have the full file and you try to load it, I think Protigy is going to complain that the file was not complete or there were not enough rows or you will get an error and Protigy will close. So, you have to wait for the file to fully download and once it downloaded, then you can load it into Protigy and you will get to the screen that says you want to pick annotation column for which one.

Student: For mutation.

Yeah.

Student: 253.

0 means not mutated, 1 means mutated. So, you could can designate on what type of mutation or which site was mutated, but that only results in subsets that are way too small to analyze. So, we just used mutated or not mutated. And, I think there is also NA which is missing. So, for some of the mutation status if you look at it there would be three groups.

Student: Yes.

0; that means, it is not mutated, 1 means its mutated, NA means we do not know. So, remember we have three normal samples. For normal samples we did not measure whether the thing was mutated or not and so for those we basically mark it as NA. So, when you do an analysis you want to exclude those samples and work only on things that are mutated or not mutated, I will show you how to do it, but let us just see if everyone is in reasonable shape yeah.

So, there is a way to look at the data in Protigy. So, once we get there I will show you otherwise, I can show you the data separately.

Student: Ok, you will.

You can open the GCT file in excel also.

Student: No, that I will see.

Yeah.

Student: But, how you reach to the that GCT file.

So, that we use Spectrum Mill. So, we use Spectrum Mill to create the log ratios.

Student: Ok. So, that takes the.

And, then we went through normalization.

Student: Ok.

If I will, you can do normalization with this if you want, but you need at least output from Spectrum Mill or.

Student: So.

We use Spectrum Mill, you can use anything.

Student: So, Spectrum Mill take the input of raw.

Mass spec data.

Student: iTRAQ data.

Yeah.

Student: Ok.

Yeah.

Student: So, if you have any raw data iTRAQ data.

All the data for that paper should be on the nature website.

Student: Ok, it is already available.

It is all there yeah.

Student: Ok

Sorry. So, let us actually.

Student: Excuse me, Sir.

Yeah.

Student: So, I need to say ok here.

Yeah, I will get to that, we will.

Student: Another file is coming. So, just to want to show Morpheus; this one.

Yes, that one.

Student: Then how to see this one.

I will show you.

Student: All the csv file vcd file and these file can be through, we can convert excel into csv and you can.

Yeah you can use, but the annotation is little thickly like I mentioned so.

Student: What would be the QC status? What would that mean?

Student: QC pass failed.

Yeah. So, we use QC pass failed to decide which one to include and which ones to exclude. So, in the breast cancer paper there where set of samples that we excluded.

Student: Yeah.

So those who were marked as QC failed so.

Student: What was the criteria that was put?

That you have to read the paper, it is relatively complex, so I do not want to go into it now.

We can talk later if you want.

Student: Sir, both are same, there are two QC status.

Yeah I think that was because of some processing, I think it was included multiple times, it is the same thing.

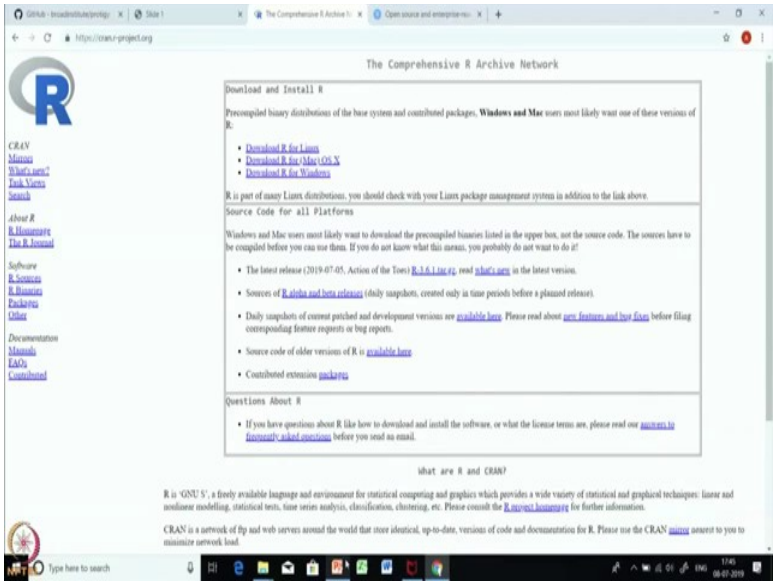
Student: Ok.

Yeah I think it is possible one might have 0 1, the other one has QC failed QC passed something like that.

(Refer Slide Time: 12:15)

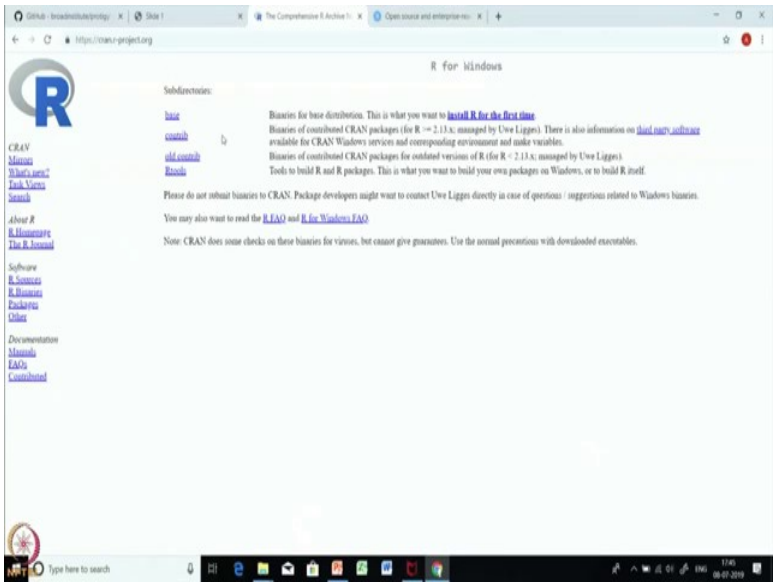
Protigy is completely based on R programming. So, we need to learn a basic workflow how to run the scripts to get the wave interface of Protigy. Today, I will like to show you a very basic work flow how to install R and R Studio and to run the scripts that required for prodigy. This hands on is only for the people who do not have the R installed in their system.

(Refer Slide Time: 12:54)



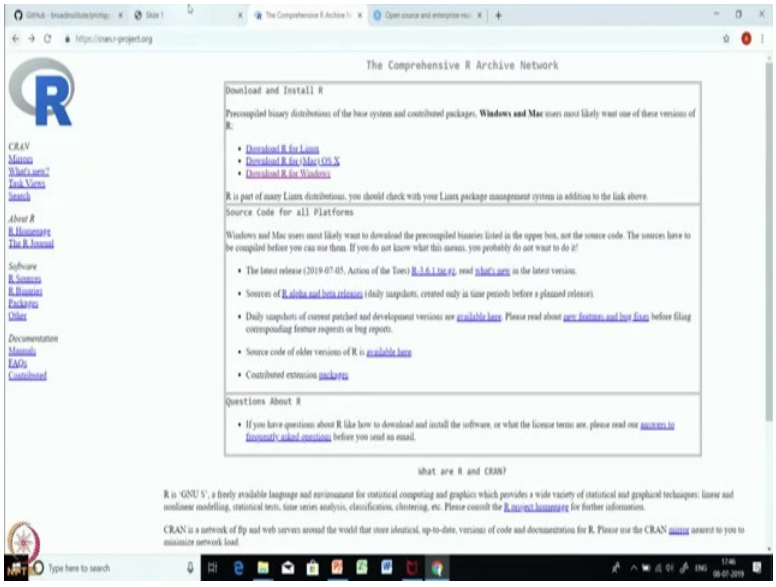
First we need to install R; the version is 3.5.1 and R Studio. So, on the basis of your system compatibility if you are using Linux or Mac or Windows you need to download R on the basis of that.

(Refer Slide Time: 13:09)

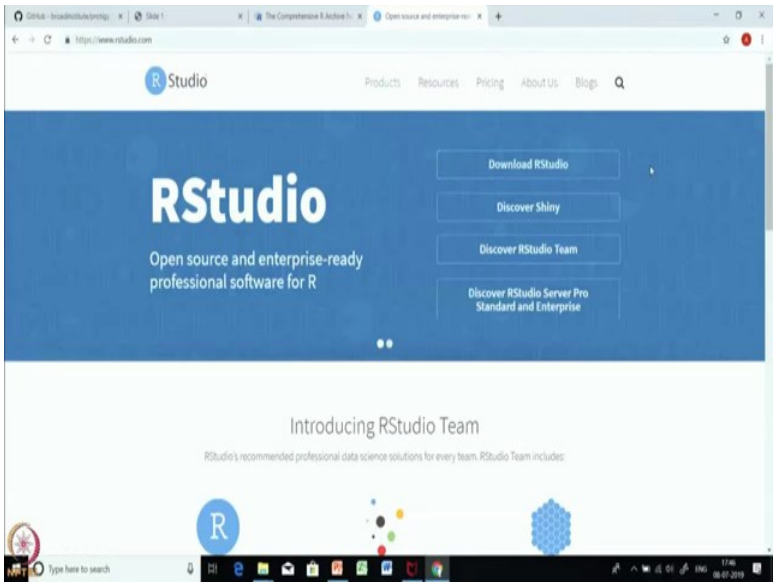


So, I will show all the downloading and installation based on Windows.

(Refer Slide Time: 13:17)

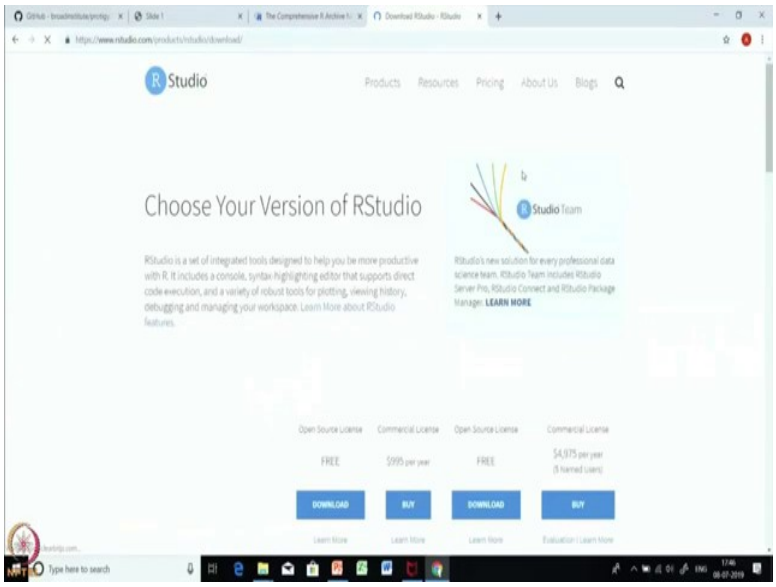


(Refer Slide Time: 13:19)

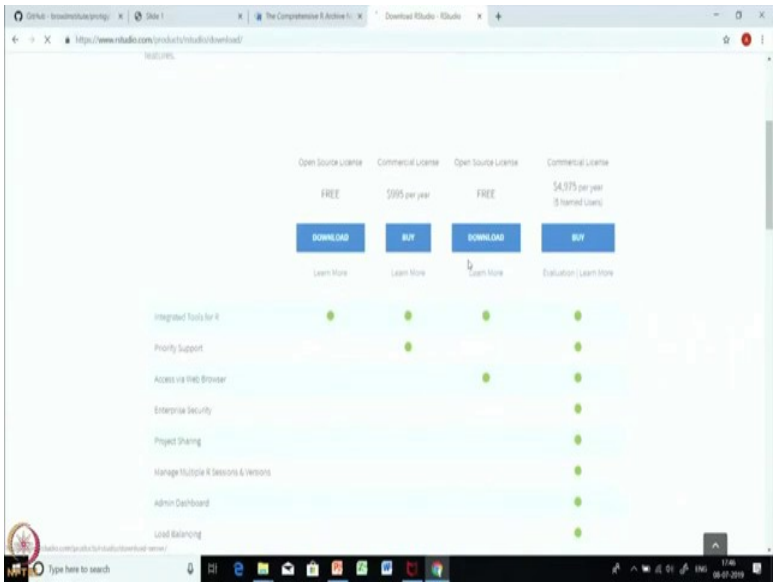


So, you need to download R from here and R Studio from this website.

(Refer Slide Time: 13:23)

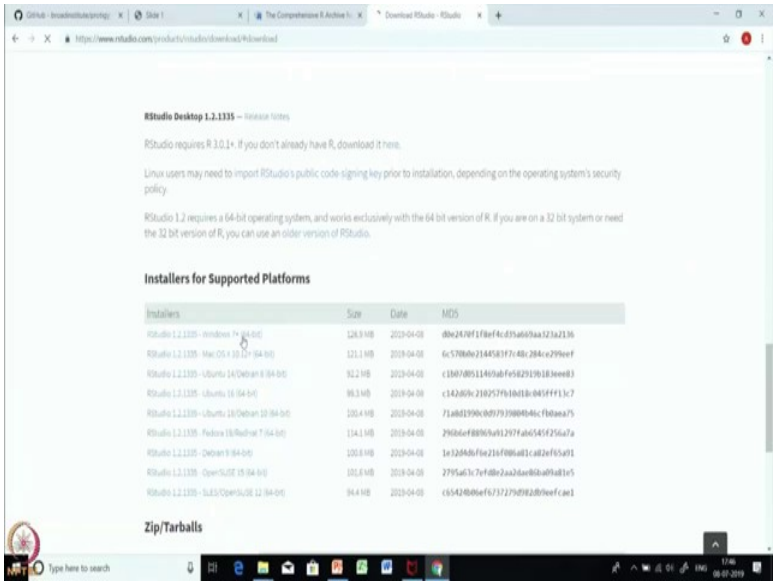


(Refer Slide Time: 13:25)

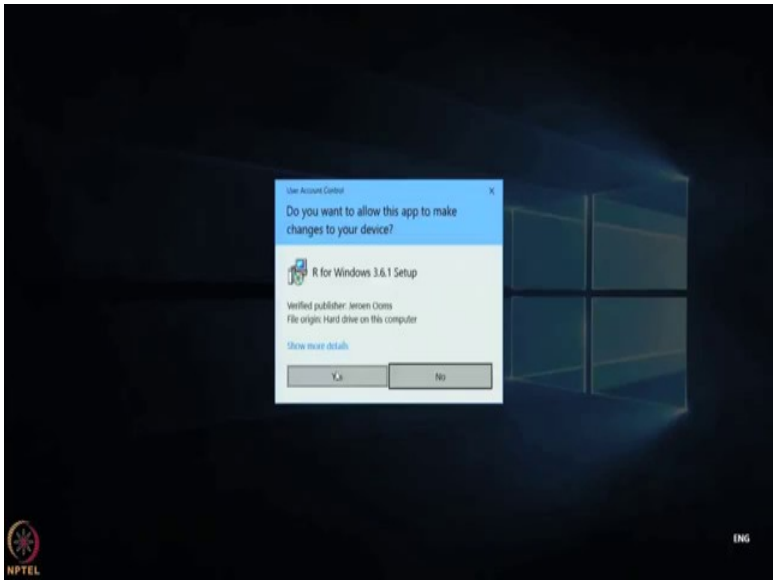


In the while downloading the R Studio, you need to keep in mind that you need to go for the free version that is available here. After downloading both the R and the R Studio you need to choose the installers on the basis of whatever operating system you are using.

(Refer Slide Time: 13:30)

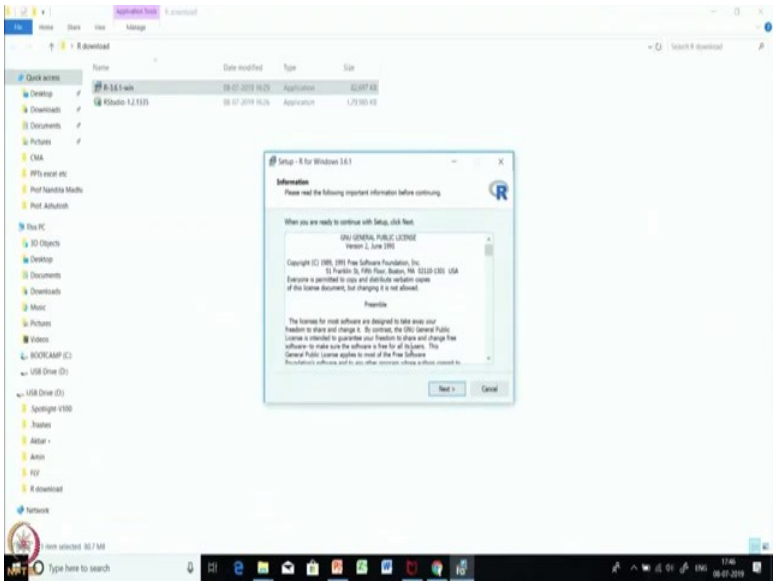


(Refer Slide Time: 13:43)

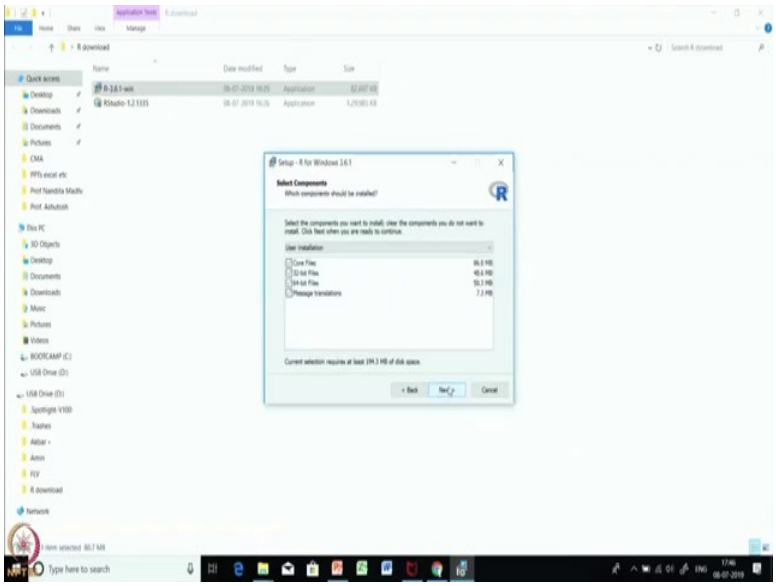


So, now so after download the both R and R Studio we need to first install the R.

(Refer Slide Time: 13:47)

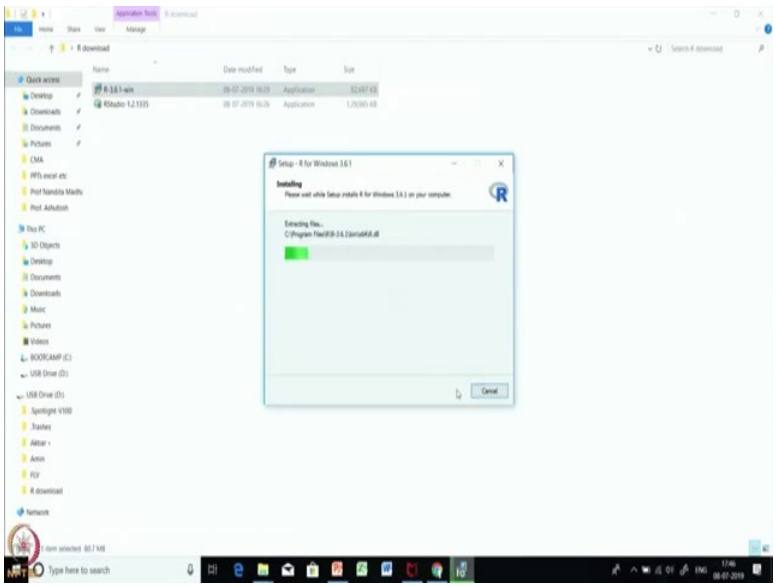


(Refer Slide Time: 13:50)



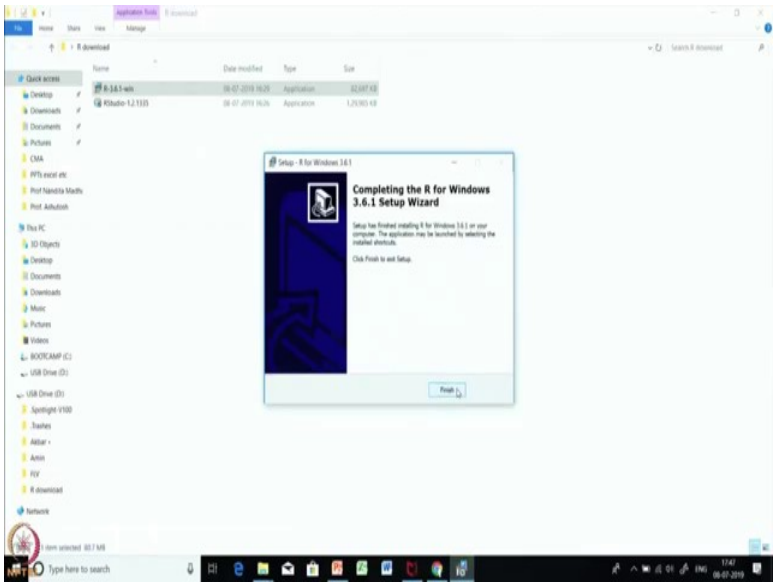
We need to keep in mind that we will only follow the default installation rather any kind of customize installation.

(Refer Slide Time: 13:54)



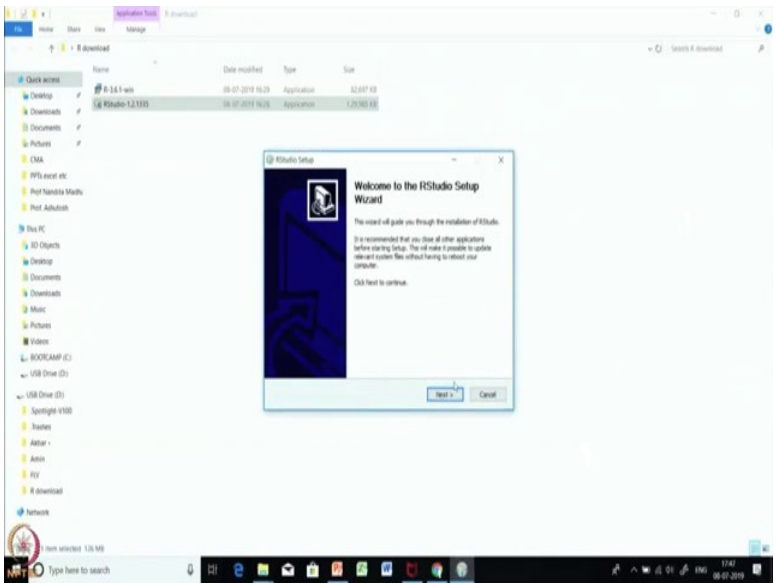
The installation might take some time.

(Refer Slide Time: 13:57)



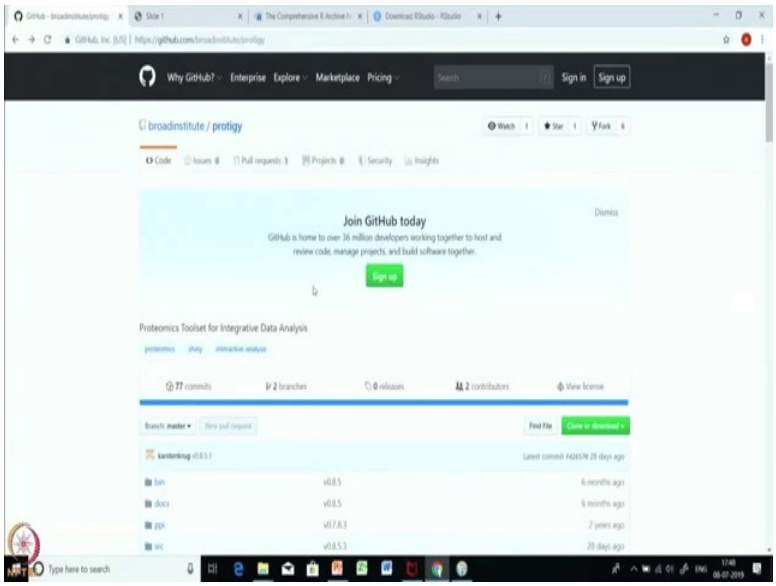
So, after the completion of the first installation, we need to do the installation for the R Studio.

(Refer Slide Time: 14:06)

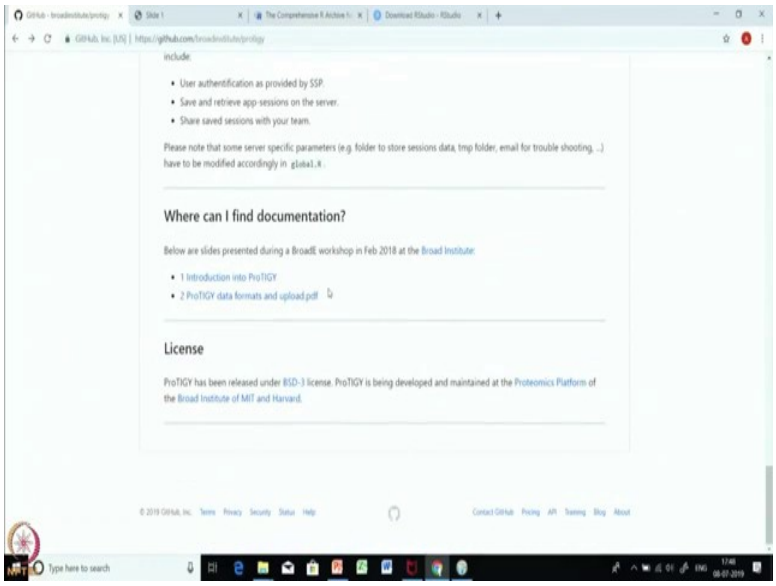


So, here also I will recommend that do not go for any kind of customize installation, rather just click all the tabs as the default installation and install the R Studio also. So, while the software's are getting installed, the Protigy broad institute in GitHub.

(Refer Slide Time: 14:23)

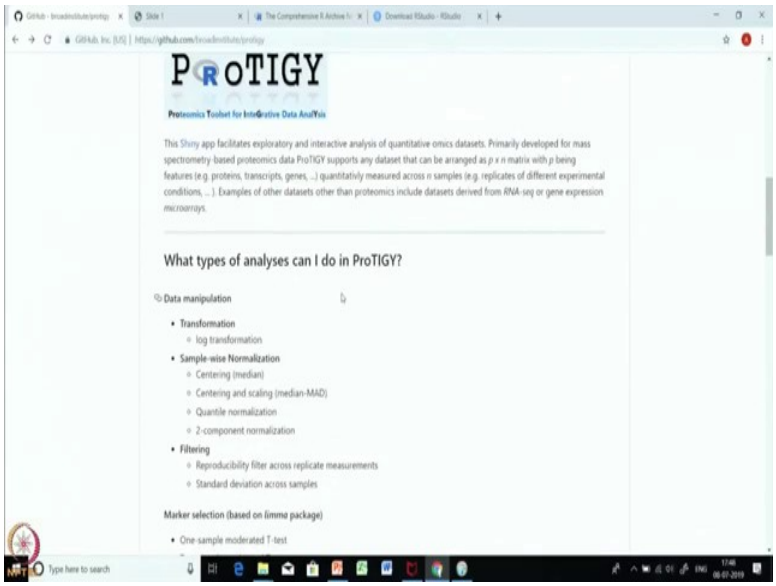


(Refer Slide Time: 14:33)



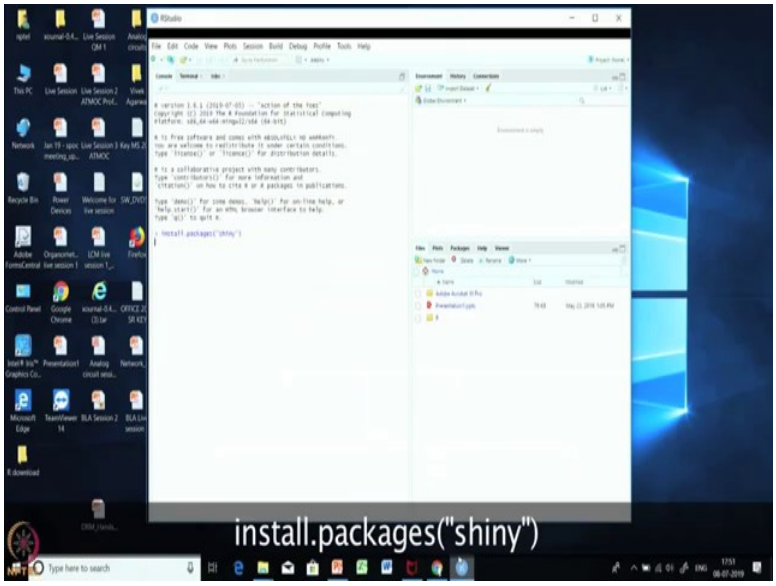
So, this is the software Protigy. So, from here you will get the couple of information about the Protigy and even the slides that you will be needing to understand and to install or to upload the data, and what kind of data format you need to do the analysis in Protigy.

(Refer Slide Time: 14:44)



So, I will recommend you to go through the complete web page to understand and to get important information about the tool. The kind of dataset require for to run Protigy is in p*n matrix. So, where p is the features that is, that can be may be the number of proteins or genes and n is the number of samples. So, this kind of small-small information you will get after reading this software web page. The installation of R Studio also got completed.

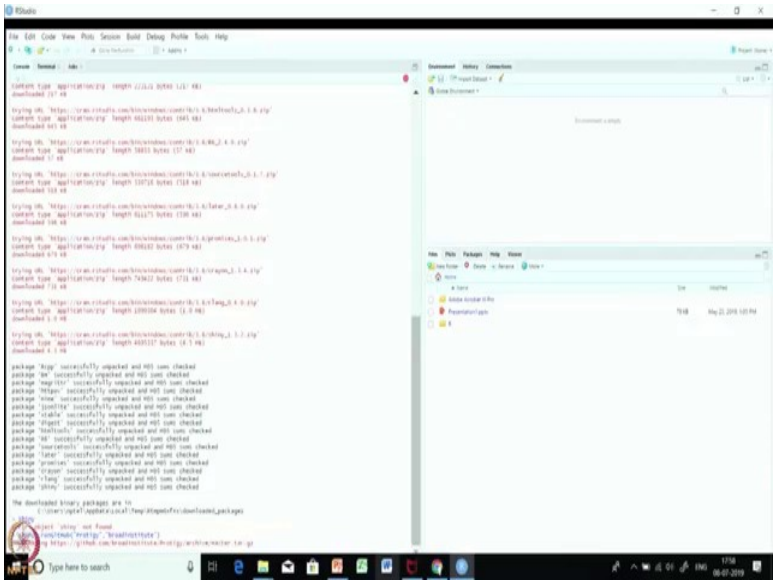
(Refer Slide Time: 15:13)



As Protigy is a R based software, but after giving some basic command it will give you a web interface. So, most of the R based command which is giving web interface to you are always scripted on the shiny. So, first we need to install the shiny packages. So, for installing any packages in R, we need to keep in mind that the basic code for installing is installed.packages and in inverted comma you need to give the name of the software.

So, let us try to install the shiny first. So, as you can see after just writing install that is coming in the drop down menu. So, from here just you can choose the install packages and here in inverted comma you can write shiny and click enter.

(Refer Slide Time: 16:06)

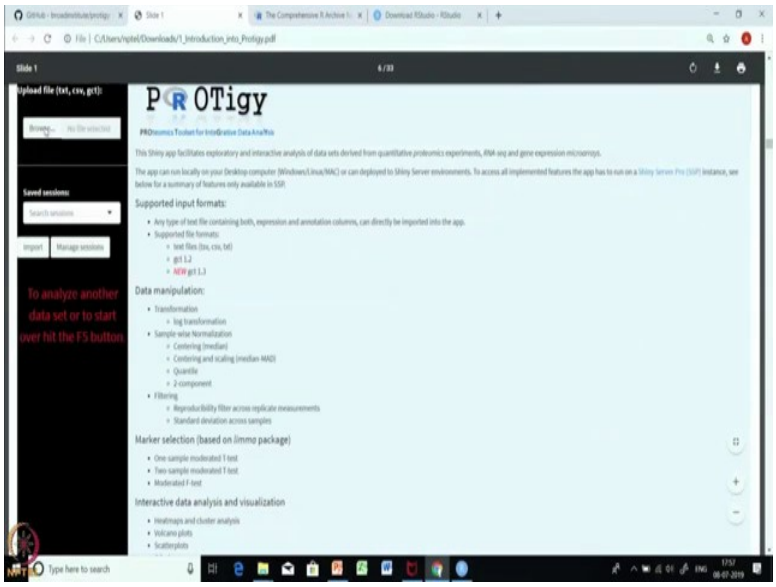


So, as you can see after clicking enter, it has start downloading the packages for shiny. So, it will take some time and after the package get installed, we will try to install the Protigy software. The installation might take some time. So, finally, the shiny package got installed. So now, we will try to install Protigy software. So, to install the Protigy software we need to write as we have already install shiny, now all the command will be based on shiny only.

So, let us see how to install Protigy software. So, the basic command for installing the Protigy software is shiny two colons. So, the basic command for running the Protigy software is need to have the shiny and followed by two colon, then run GitHub Protigy and broad institute. So, after that we will click enter and it will show that Protigy is getting downloaded. So, it will take some time, sometimes it might take 10 minutes, for

some user it might take more than an hour. So, after the completion of the Protigy software you will find there will be web interface of Protigy will be opening and from there you can upload your data. The data that is available in the Google drive link that has already been shared to you.

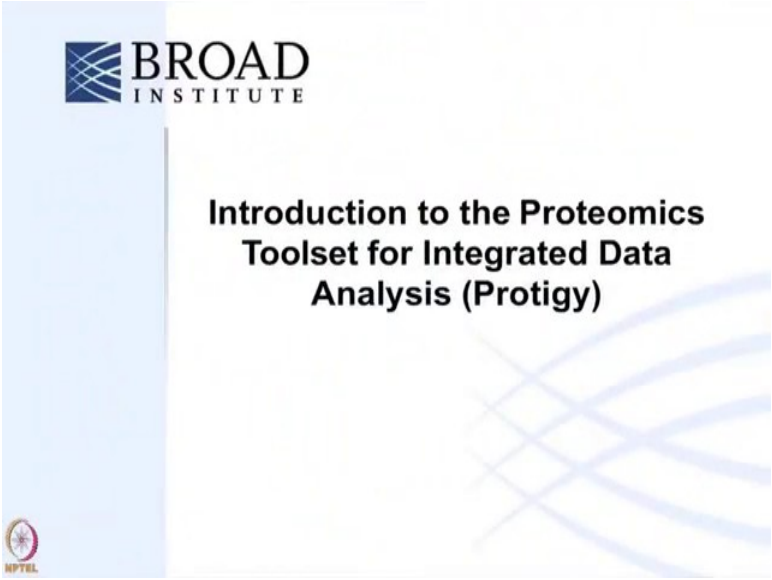
(Refer Slide Time: 18:20)



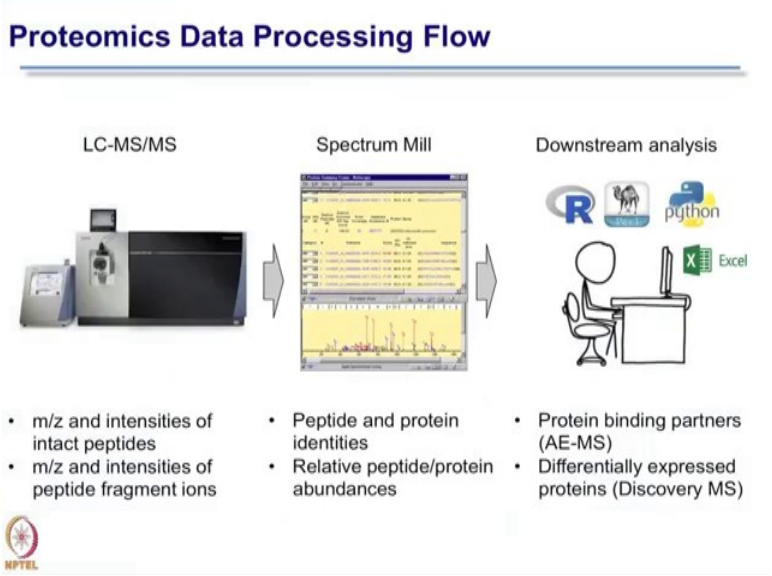
So, after completion of the Protigy installation, the web interface will look like this and in the left hand side there is a browse option, where you need to click browse and you need to upload your dataset. For some user there might some error comes, but to troubleshoot the error you need to read the error what is the problem that is coming, if it is something that is linked with installing a software, just writing install.packages inverted comma and the name of the software which the error asking for and click enter. You will see that it will help you to download the Protigy.

Thank you.

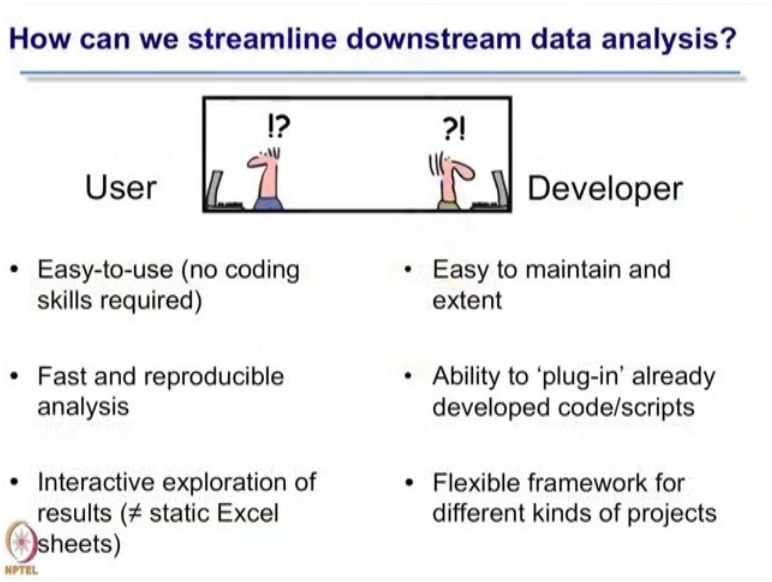
(Refer Slide Time: 18:58)



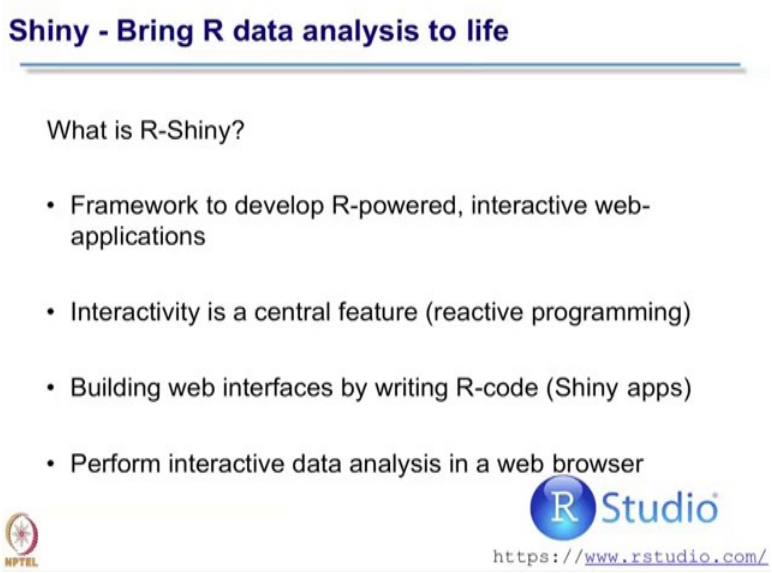
(Refer Slide Time: 19:03)



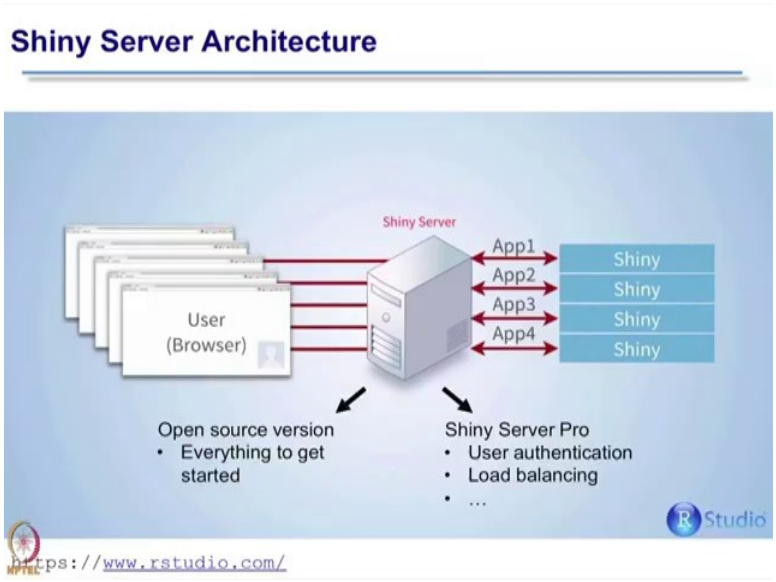
(Refer Slide Time: 19:08)



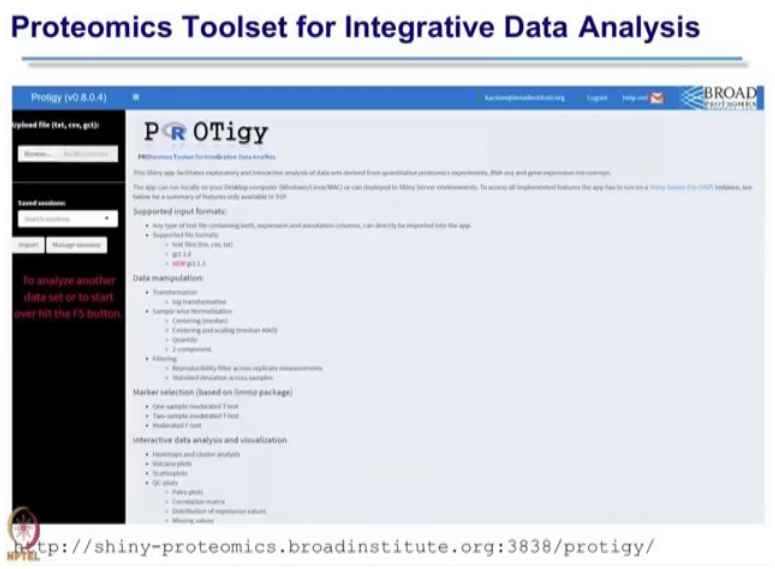
(Refer Slide Time: 19:13)



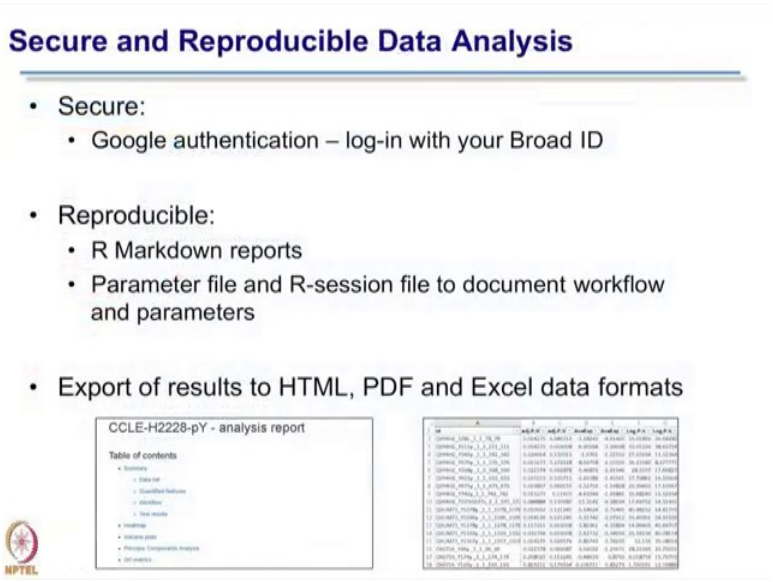
(Refer Slide Time: 19:18)



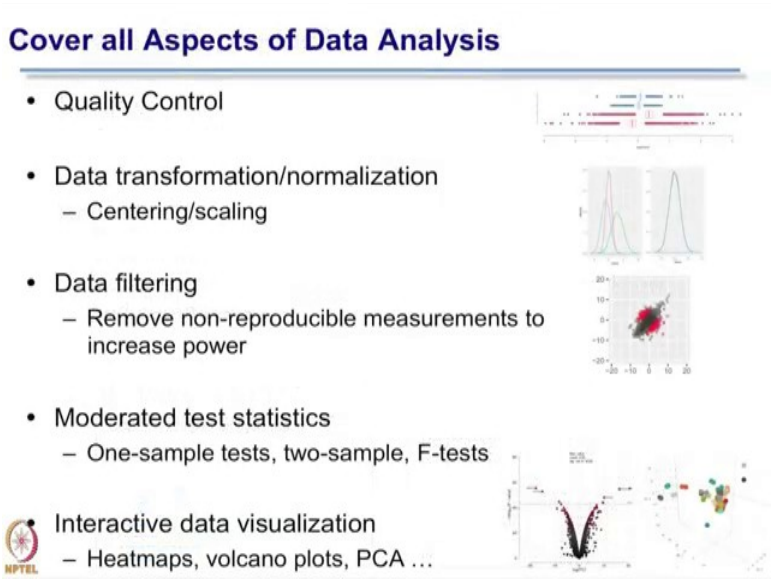
(Refer Slide Time: 19:23)



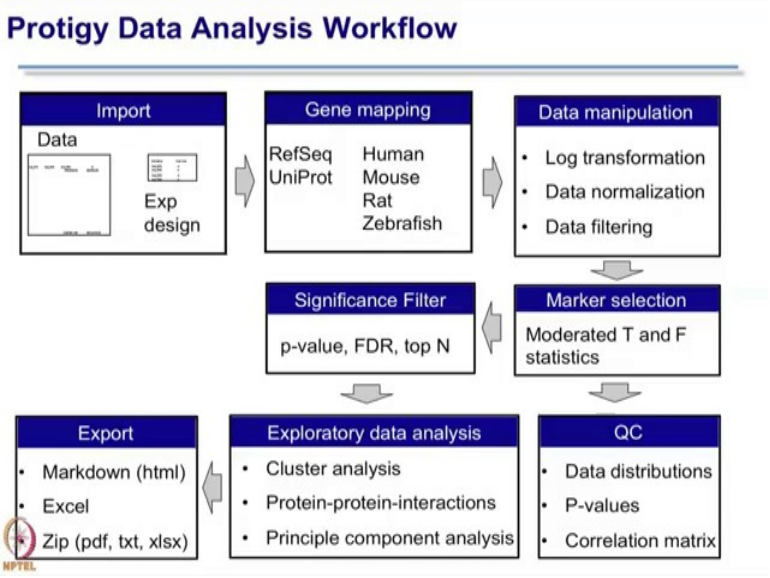
(Refer Slide Time: 19:28)



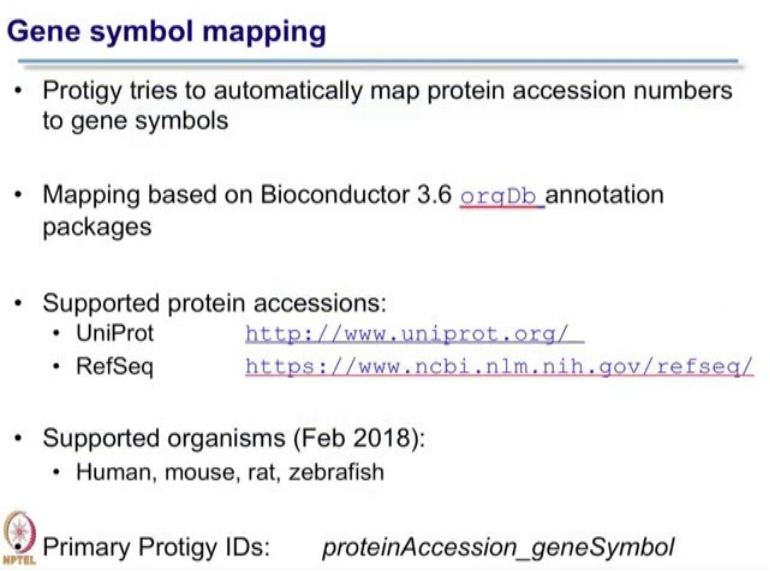
(Refer Slide Time: 19:32)



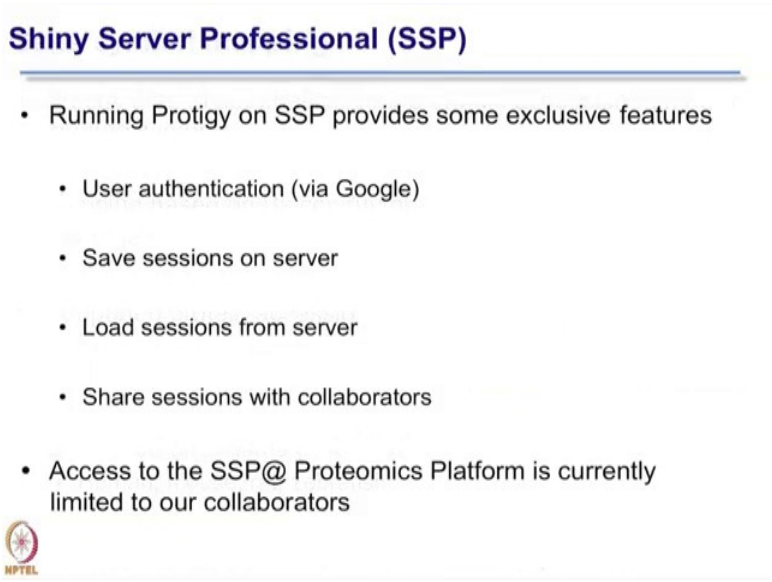
(Refer Slide Time: 19:38)



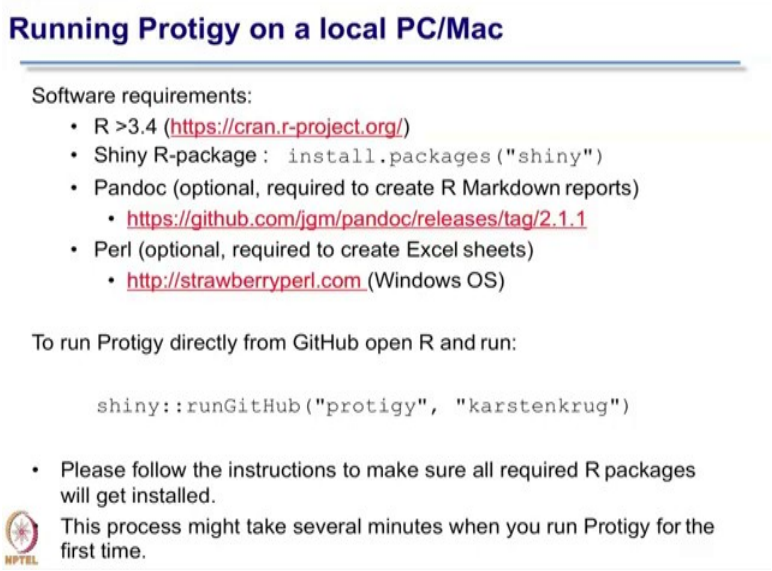
(Refer Slide Time: 19:43)



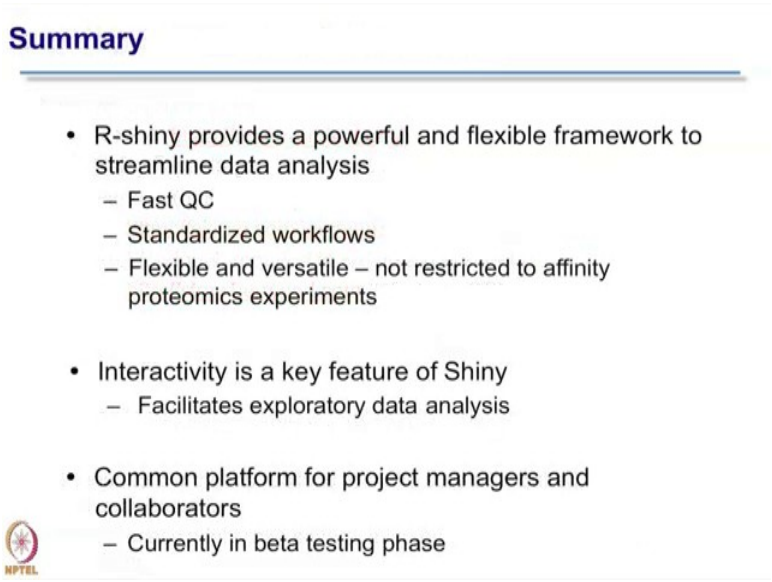
(Refer Slide Time: 19:47)



(Refer Slide Time: 19:53)

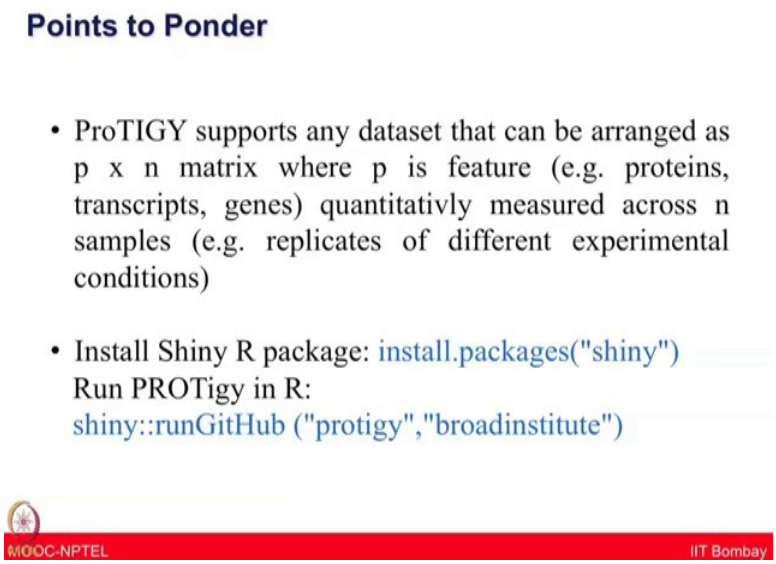


(Refer Slide Time: 19:58)



(Refer Slide Time: 20:02)

(Refer Slide Time: 20:07)



I hope today's session was useful and you are introduced more details about protigy. You must have got now a fair bit of idea how to go to the broad institute portal and explore the software Protigy. The session of annotation is the crucial step before running the analysis. In the next hands on session, we will learn more about different options; like log 2 transformation, normalization, data filtering, and test selection in Protigy.

Thank you.