Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Pratik Jagtap
Dr. Ratna Thangudu
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
University of Minnesota

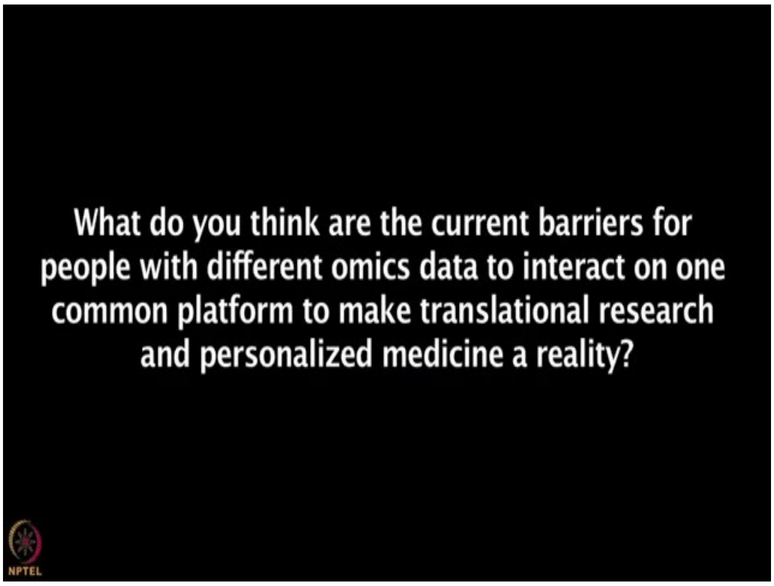
Supplementary Lecture – S20

A perspective on Proteogenomics – V

Hello, my name is Pratik Jagtap, I am a Research Assistant Professor at the University of Minnesota part of the Galaxy P team, wherein we developed proteomics tools to be deployed within galaxy. The team at university of Minnesota led by a Professor Tim Griffin basically works on adding tools and workflows for proteogenomics research as well as meta protegenomics research.

And we have been doing this for last 6 years, we conduct tutorials as well as we have certain using this workflows for multiple research projects and I am very excited to be part of this cancer proteogenomics workshop here in Mumbai. Basically because of the fact that, that is the research that we work on and gives the ability to reach out to the audience here in India as well as the rest of the world, right.

(Refer Slide Time: 01:21)



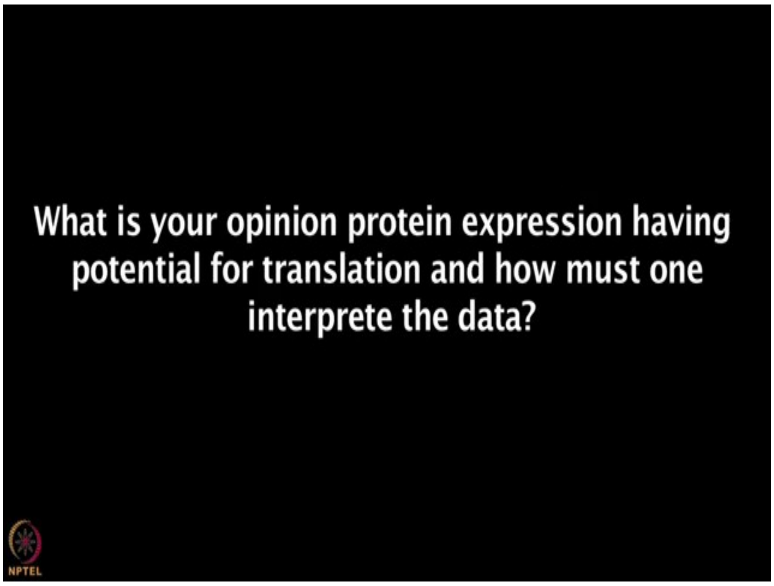
What do you think are the current barriers for people with different omics data to interact on one common platform to make translational research and personalized medicine a reality?

So, I will try to answer this question in two parts; the first part you mentioned was about a common platform and that is exactly what we do with galaxy platform and I will be giving a talk very soon about that. But, galaxy platform basically helps both genomics transcriptomics as well as proteogenomics researchers to use a common tool or platform to using it.

But, in general I mean one of the main barriers that I see is you know most the time to researchers or developers a kind of specialized in one field and not the rest of them. And hopefully a platform like galaxy or anything else helps to bring that common playground or common place varying all these researchers or a developers can develop tools and help integrate the data. I think there is also need for researchers to understand the fact that developers and users need to work together because things that are developed by a developer might or might not be useful to use it.

So, it could be a great algorithm does fantastic things, but if it is not something that aligns with the question that is asked by the researcher, then it just becomes an academic tool right vice versa. The user also has to understand the possibilities and challenges that a developer faces and try to achieve tools that work and give you a multi omic or a systems biology perspective to the data.

(Refer Slide Time: 02:59)



What is your opinion protein expression having potential for translation and how must one interprete the data?

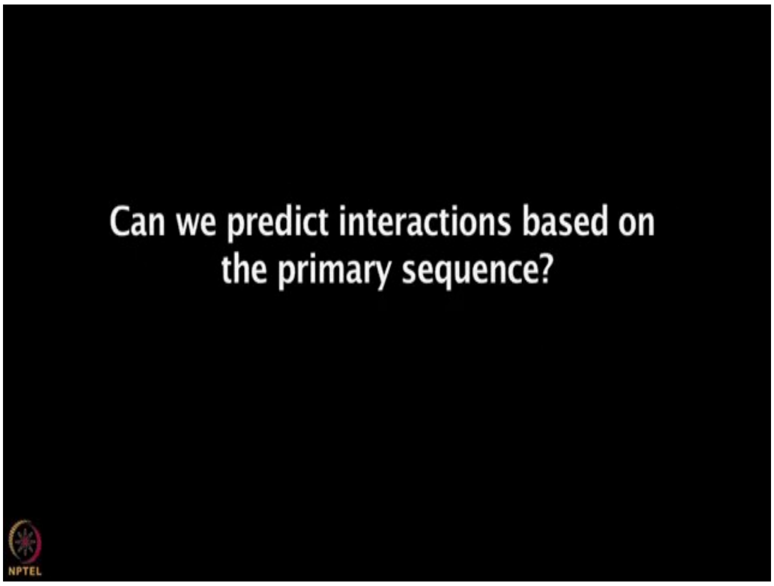
One of the observations that the researchers have started making now with multi-omics or trans-omics research as it is called when you comparing let us say a transcriptomic data to proteomic data. At least in early days one found at the correspondence to each other in terms of quantitative expression was not exactly a 100 percent.

In fact, the concurrence was much much lower, that was little bit of a concern earlier. But now it is understood that the way RNA expression works or protein expression works is not exactly instantaneous in the sense. You could have a RNA expression and the protein expression could lag behind right or you could have the stability of your RNA molecule determining you know how much of protein is going to be expressed.

So, I think it is really important that researchers start undertaking temporal or time dependent expression studies for both transcriptomics and proteomics to kind of make a much more studied conclusion on the expression of both protein as well as RNA. Because if you find that the RNA is low in protein is more it does not mean that you know it is just giving you a particular snapshot and not the cycle of that particular expression. So, I guess the answer to that is time dependent studies.

And the technology is getting there I mean RNA-seq is already there and I think with your developments in mass spectrometry, the scan speeds are getting really fast. We hope that we can get deeper as well as lot more data from mass spectrometry data. So, that one can match that with transcriptomic data.

(Refer Slide Time: 05:01)

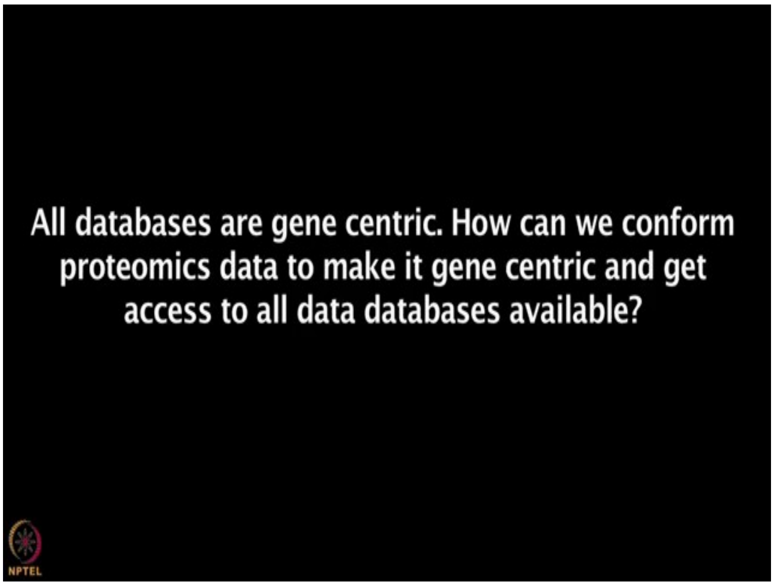


Can we predict interactions based on the primary sequence?

I have a little bit of concern with that given that you know and again I have personally I have not really studied that as much I have worked out domains and I have looked at. Using domains how one can predict the function of a protein.

But, in terms of interaction I either one do not have enough information or secondly, if you have predictive models and if these predictive models are backed up by experimental data then yes one can see that, but until we actually have a good correlation between those two. I think experimental data is going to be a lot more dependent or lot more determinant of what actually interacts with each other than you know computational modeling, but that is my opinion.

(Refer Slide Time: 06:03)



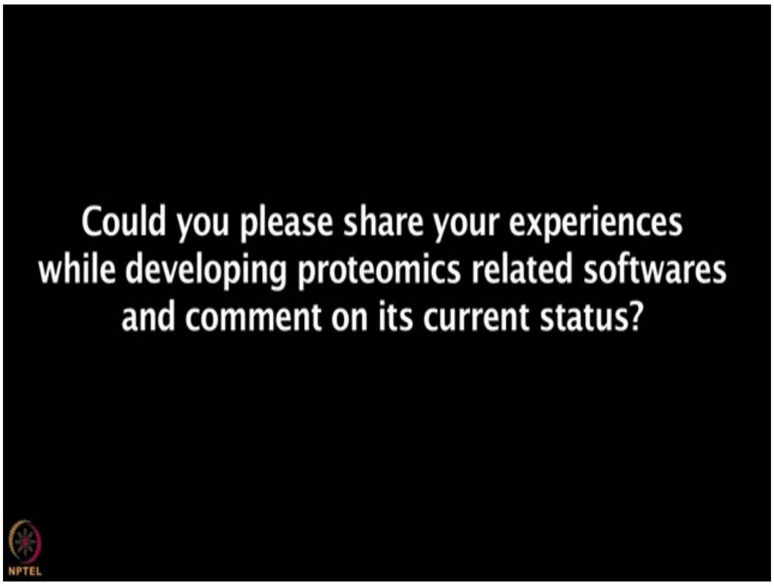
All databases are gene centric. How can we conform proteomics data to make it gene centric and get access to all data databases available?

I think it is important to start with RNA-seq data and I will be covering a little bit of that in our in the talk that I am giving today. So, if you start with a RNA seq data; obviously you have your genomic coordinates or you can go and get back your genomic coordinates. And if you use that as a template and transfer that information to your proteomics data or at least have a database scheme are two kind of go back and find out your genomic coordinates or a gene centric approach to that.

I think that is possible we have shown that and; obviously, there are tools and workflows that need to be first developed and then optimize make made robust enough. So, that one can do this on a more consistent basis not only for known organisms for, but also for organisms that are you know getting sequenced.

But, it definitely is possible going through the RNA-seq data, I think for the proteogenomics field to develop, one would actually have to make this almost a requirement, because if you do not correlate your protein to your DNA or to your RNA, you almost losing that information and you want to maintain that. Because, you kind of know it is coming from you know from DNA to RNA to protein just the fact that if you do not have the tools available or the coordinates available is not a good enough excuse to lose that information. So, I think it is going to be necessary as the field of proteogenomics becomes more established fields as you know as it is emerging.

(Refer Slide Time: 07:48)



Could you please share your experiences while developing proteomics related softwares and comment on its current status?

So, I have basically worked on two workflows or two areas of research one is proteogenomics. So, when we started working on galaxy using galaxy for proteomics what we did not want to do was just develop another platform for proteomics research, you know taking your mass spec data your protein FASTA file you get peptides and proteins.

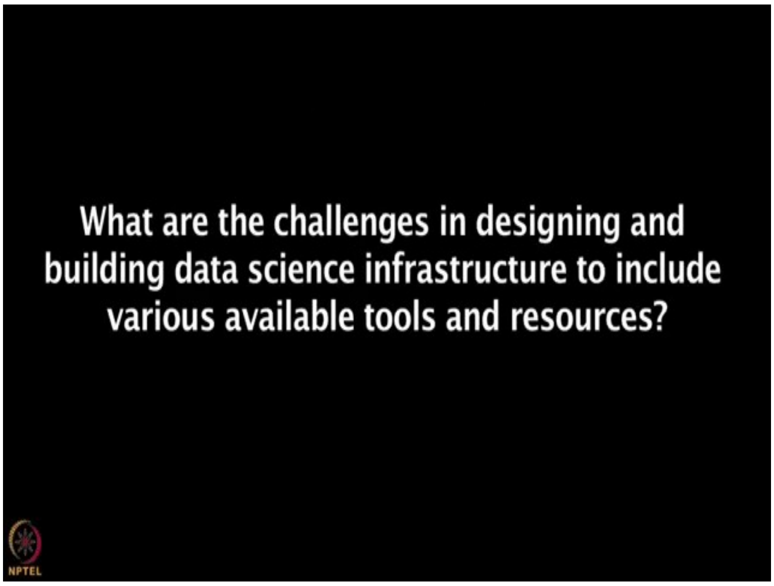
There are many tools which do that, what we wanted to do was take some challenging areas and this was 6 years ago. So, proteogenomics was very new, metaproteomics was very new and I will not say they were very new, but they were emerging and we saw a promise as well as a challenge there.

So, what we did there was we work with the post processing of the peptides identified and tried to make it easier for a user to use it. And that is where I mean the challenge there is working with the user or a project and a developer. And as I was mentioning during the break you know the developer is extremely you know enthusiastic about his work, the user is very focused on his questions asked. And sometimes these do not need and that leads to a program or a workflow which you know which is great in it is own field, but it is not usable, right.

So, I think the developers and users if they work together on a project with a specific questions in mind. And then creativity starts coming in once you have the basic blocks in place, I think that is how you know tools are going to develop. In terms of it is current status I think it is in pretty good shape. I would say mzML format mzXML format are kind of getting accepted the only part I think I see a need for development is the mz quant or the quantitation portion of the mzML format.

And I know there are developments taking place there, but I think making this more robust and optimized is going to be the need. Because they are I think are going to be many quantitative studies and especially quantitative studies that correlated to RNA-seq data or any other quantitative analysis data. So, getting the quantitative portion of the proteomics or mass spectrometric data is going to be important.

(Refer Slide Time: 10:28)



What are the challenges in designing and building data science infrastructure to include various available tools and resources?

I will again answer this in two parts, one is you know you want to develop something to show it works right and we have been doing that we take small datasets generate workflows develop it on a cloud. So, you know build a docker instance share it with the world we give presentations saying this works, but that is almost like playing a sport in a small little you know backyard or something.

The real data that comes out is not going to be just a few raw files or a few FASTQ files which going to be many and maybe multiple replicates and, but hopefully many time points as well. So, you need to have something at the back end the infrastructure needs to be such that it can support that. But, it is also important that the tools and the workflows can run on that they use the ability to use you know the vast resources that are available either in the cloud like Amazon and Google or you know any supercomputing infrastructure that you might have.

So, I think it would need to go through steps you need to make it work first and then you know just like child develops you want to see that it graduates from that you know it school to college and maybe into the real world. So, there is definitely need for that, in academic studies sometimes it is not possible because teams get funded for 5 years and then focus changes. But, I hope the field in general kind of understands that and makes it possible because it is really not much fun to go back developer new workflow and start again and again. So, hopefully there is you know this momentum keeps going on.

(Refer Slide Time: 12:20)

According to you, what are the capabilities that graduate students should acquire to enable them to analyse multi omics data?

At multiple levels definitely the ability to ask good questions and that does not come easily right and some people are naturally talented they will ask good questions after you know after in really earlier in the education. But, one thing that you developed as you do your advanced studies is you start asking really good sharp questions. I mean as a young student you always have 10 questions, but you know you are not able to decide which of these 10 are good you think all 10 of these are good, right.

So, asking good questions and then secondly, designing experiments; it is always good to have great ideas, but to put it into practical stepwise manner is point. The third part is you know sample preparation I mean, I know I mean I work in the bioinformatics area, and I know that many researchers kind of or take it as sample preparation is going to be good or you either blame the sample preparation or you know you kind of kind of taking it for granted.

But, I think there is a lot of quality control that needs to be done lot of things that need to be you know considered that’s true for data acquisition as well. You need to have qc parameters to ensure that if you have replicate you know generated on day 1 and generated on day 10, you can compare them or if you cannot compare them what is the reason you cannot you know.

So, the measures to have that and there are tools in place to do that as well, but I think most importantly it is important that the researcher gets to interpret the data, right. And interpretation could be on various levels it could be by using programming you or it could be just the biological interpretation where in saying good I have seen this data I know what it means.

But, I need tools to do that and that is where the person can work with the developer which is what I do, I work with developers, because I kind of get a sense of where data is going or what could be important. But you know you can start in one area and then develop into any other area or you could just become an expert in one area. And then you know once you start having that ability to look at overview of the project asking good questions and publishing good science.

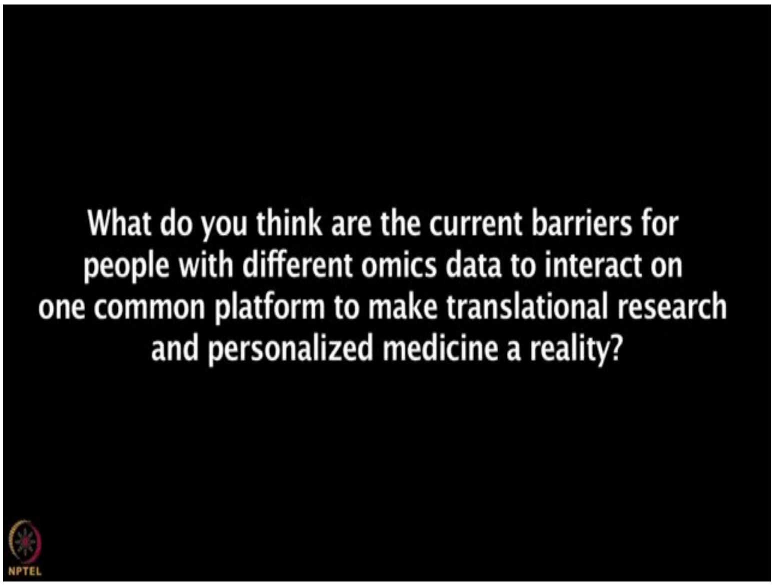
Then I think you achieved the ability to do know do things. While I was answering one of the things I kind of learnt during my career was also ability to communicate and there is not just through manuscript and through you know through anything that is public or I think you need to clearly mention to the person you are collaborating with because team science is going to be very important.

Now, you cannot be an expert in mass spec and an expert in and you might be, but you might not have time to do that. So, you need to communicate your expectations to your you know collaborator and get the best out of them and also offer the best that they want from you. So, I think communication is very important I think; obviously, the younger generation already has kind of strength in that because of the amount of social media that is available.

But, I think it is important to have effective communication while avoiding the noise, you know how do I get this across to somebody which is a signal that could be useful rather than giving you 10 pages of data and say go find your answer. So, I think these are a few things that that develop and I am sure there will be new skill sets that will come up as the you know as the field develops.

Hi, my name is Ratna Rajesh Thangudu. I am from a company called ESAC, we work closely with the National Institutes of Health USA. So, we are a bio informatics and health IT company, we do provide a lot of services to both the National Institutes of Health and also the office of National Coordinator in the USA.

(Refer Slide Time: 16:33)



What do you think are the current barriers for people with different omics data to interact on one common platform to make translational research and personalized medicine a reality?

I think we can start with the sheer volume of the data and then there is lack of resources and also the infrastructure to manage that volume. So, everyone has all their kind of tools and the processes and pipelines, but not to handle the big volume of data. And the other thing I would stress is the lack of data standards or the lack of adherence to it. So, what happens is you have a large amount of data, but there is no data standard. So, you can actually come you do not you miss the ability to actually combine that analysis with other datasets that are creates the string or that are coming from the other programs. So, I think these are the main three things that the strikes my mind to begin with.

(Refer Slide Time: 17:26)

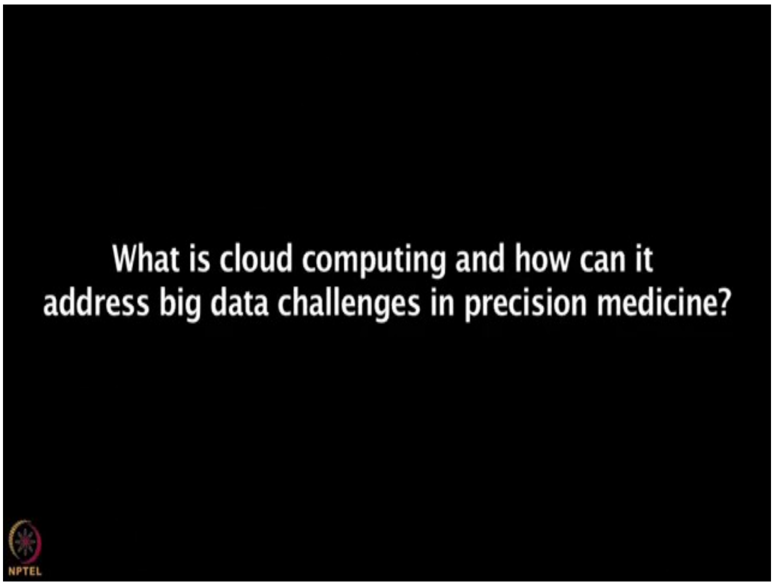
What are data standards and what is their need in data organisation for multiOMICS integration?

Data standards are basically a set of rules or agreed upon rules. So, that you represent your data in certain way so, annotate that and also represent it. So, that actually helps the data harmonization part.

So, what happens when you start generating data which is specific to a particular program or a particular country or a particular disease type or particular population and when you want to actually integrate the direct data into a bigger platform and bring the data from other platforms, for example, you call the same disease or a same gene by different names and you know that it is the same. But the computer does not understand until you tell that.

So, data standards actually help you get to that point and the other point is in the in the data harmonization. So, you analyze all of the data through a particular pipeline. So, that allows you to see all the data to the same eyes. So, even though you can analyze each of those data set independently with a different pipeline, harmonization actually brings them together and helps understanding on a larger scale.

(Refer Slide Time: 18:34)



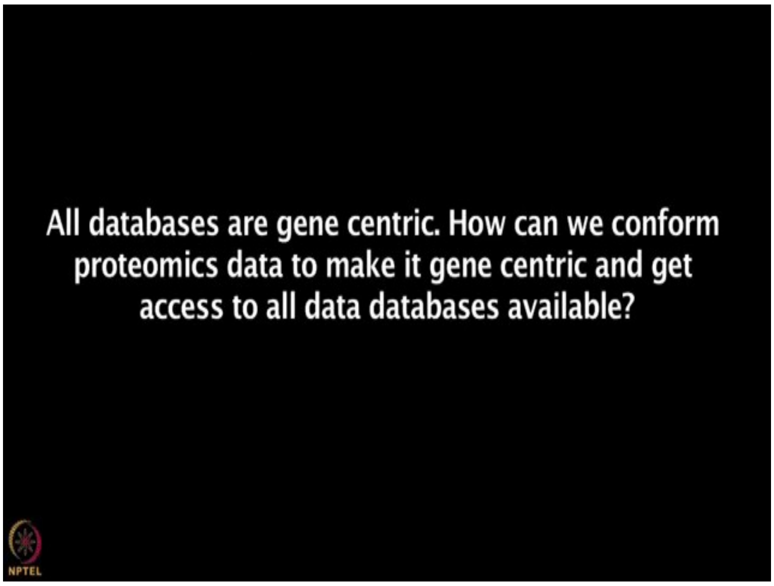
What is cloud computing and how can it address big data challenges in precision medicine?

Earlier I mentioned about the lack of infrastructure or the lack of resources. So, cloud computing actually removes that barrier. So, it is on demand pretty elastic, So, there are a lot of companies out there which are pretty established something like Amazon, Google cloud and now Microsoft Azure is there. So, for that you do not need to have a data center on your premises. So, you do not need to have a set of IT crew helping you out adding more disk space and networking and all these things. So, what you need is basically a good internet connection.

So, you based on the volume of the data it will scale up pretty quickly. So, you can increase the size of the disk space that we are using and also kind of resources to the compute power that you need. So, it is pretty helpful in that sense, so in the precision medicine so the question is specific to that.

So, what happens is like it is not the data about a particular program, but it is data about the individual patient or individual candidate or a subject it depends on how you call that. So, within the personalized genomic space so, every day the data is growing exponentially. So, the Moore's law it is does not apply anymore. So, cloud computing comes into picture to handle that kind of data.

(Refer Slide Time: 19:59)



All databases are gene centric. How can we conform proteomics data to make it gene centric and get access to all data databases available?

I do not see it as such a big problem because the connections between the genes and the proteins in the level at the level of action numbers between the most common data databases out there are for example, refseq and uniprot they are very valid annotated. And there are very well maintained and most of the proteomic pipelines are actually rolling up the final protein parsimony deserves to the gene level. So, it is pretty easy to actually map back to the individual isoforms and also see if you start from isoform you can easily come back to the genes.

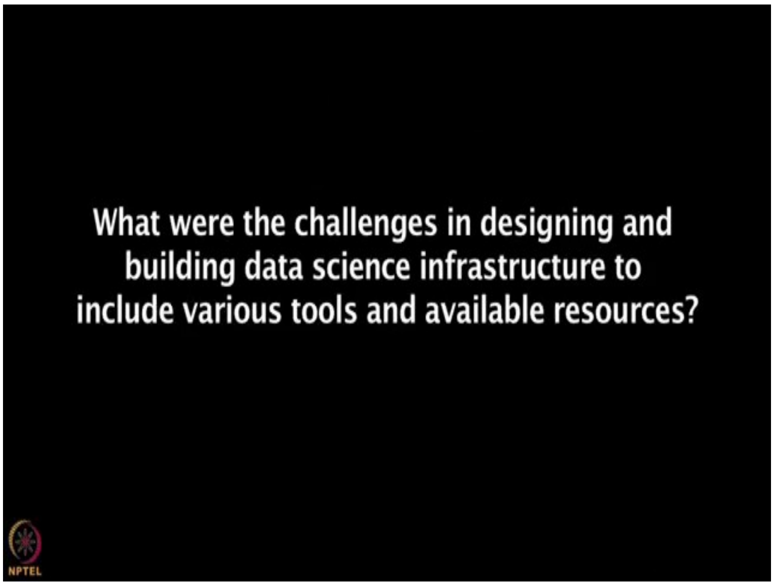
(Refer Slide Time: 20:36)

Can you share with us your experience about developing large scale data management systems and its current status?

So, now the big buzzword everywhere at least at the level of the governments that are involved the US and I would say the European Union is the big data and the data commons. So, data commons actually brings together everything at one place. So, the resources in terms of the storage the tools, the compute power everything at one place that is called data commons.

So, what the user needs is simple login. So, like the way you login to your email account. So, the user can actually just log in and it does not have to bring anything to the table and the only thing that is there is he can take back some results from the analysis directly from the cloud computing platform.

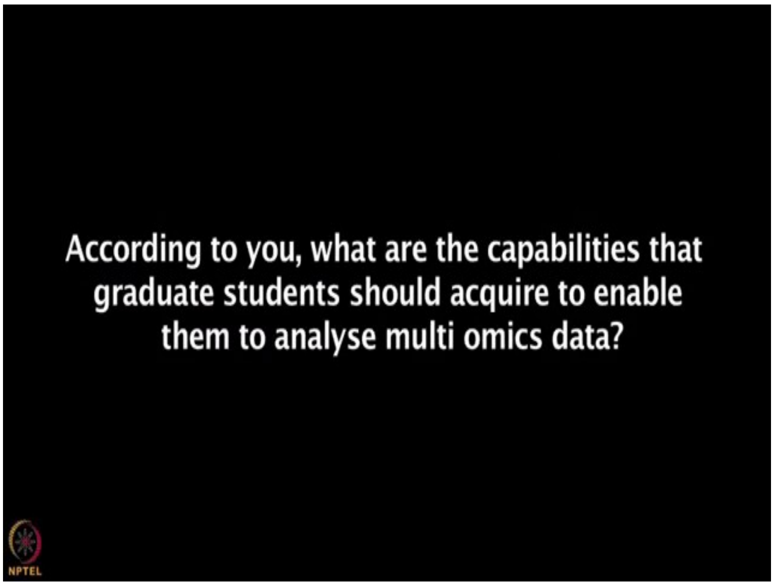
(Refer Slide Time: 21:25)



What were the challenges in designing and building data science infrastructure to include various tools and available resources?

Like I said earlier data standards, standards-standards that is very very important in achieving the goal that we have in front of us. There is precision medicine the large volumes of data and if you and this is the silo nature of the data actually never helps. So, you have this genomics proteomics, imaging the immunology all these data are sitting there side by side. But they cannot talk to each other because there are no standards there. I am not let me take that back there are standards there, but they are not adhered to. So, what happens is your calling the same thing with different names like I said earlier. So, you cannot actually do the integrated analysis.

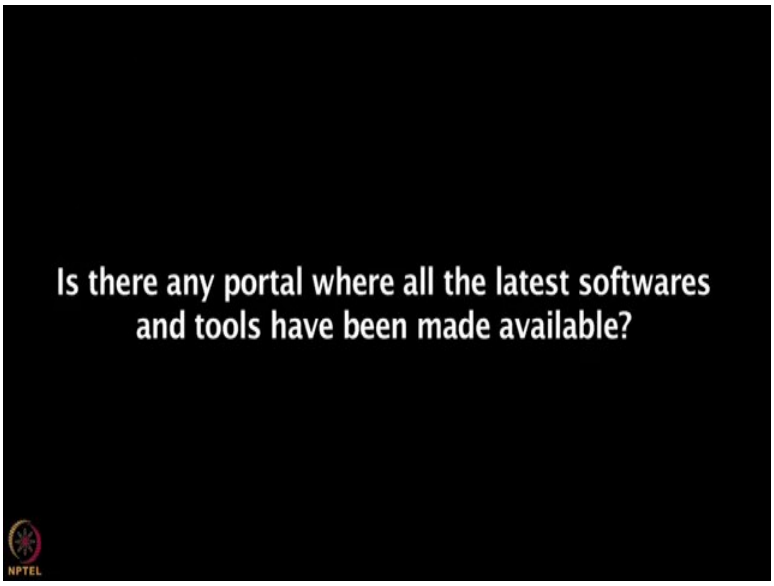
(Refer Slide Time: 22:10)



According to you, what are the capabilities that graduate students should acquire to enable them to analyse multi omics data?

Keep your eyes wide open because there is a lot of open source data already available. So, you do not have to actually generate the data, you can start looking into the existing resources bring the data onto your laptop and start analyzing that. So, to analyze the simplest tools I would recommend currently are R and python. So, they pretty easy to learn and all the public data that is available there you could process that data through. So, those kind of parsing tools and the statistical packages and we are ready to go.

(Refer Slide Time: 22:53)

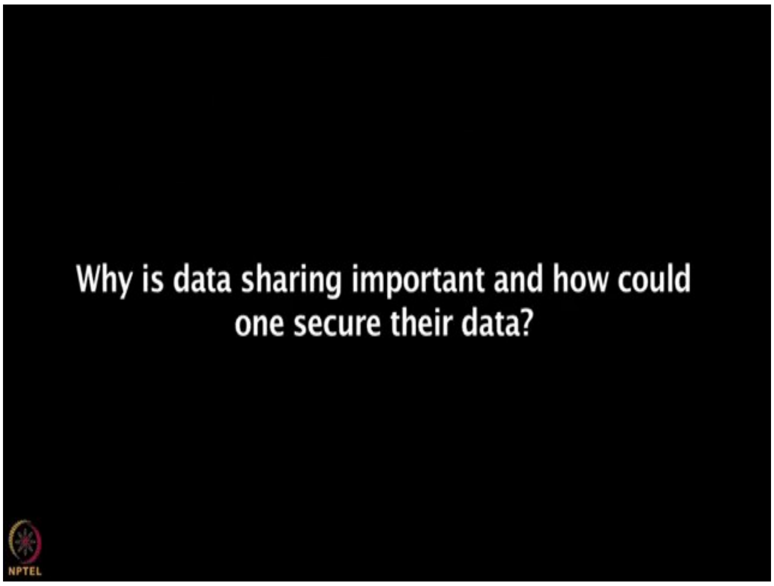


Is there any portal where all the latest softwares and tools have been made available?

There are some portals that I know of which collate all the data from the literature and also the nucleic acid research is journal which publishes the available databases and the tools on a yearly basis. But at the same time now, it is a requirement with most journals that you make your data available somewhere on a public repository and also the tools that are, that you have used and also the versions of the tools.

So, it actually lot of things that are already there that provide the metadata of the analysis that is performed on the particular data set. So, the data is available the tools that are used are available and the versions are available. So, it actually helps to reanalyze the data on different settings or on different data set or just simply reanalyze in to see cross check the validity of the reports on the publication.

(Refer Slide Time: 23:56)



Why is data sharing important and how could one secure their data?

That is critical to the all of the data common supports that we talked about, right. So, if you do not share the data, so there is no data commons. So, if you want to share the data that is always welcome. And now the governments actually require you to I mean as long as the research funding is coming from the public money they are required to submit their data to one of these resources. For example, in US there is got something called genomic data commons, where they make available all of the genomic data. And now they are building a proteomic data commons that we are building in our group.

And then the other point is so it actually helps innovation. So, sometimes you start with a small data set for example, the TCGA pan cancer analysis is one such an example. So, there are about 30 plus different cancer types that they have analyzed and now after so many years the program closed. I think at least 3 or 4 years ago, I but now there is research coming going into that and just because the data is shared everything is available in the public domain and people are using the data. So, lot of proteomic data is it is not protected it is in the in other words is open access. So, that is a good thing.

But the, if you are looking at specifically at the proteogenomic kind of integration; so, some of the genomic data is actually a controlled access. So, there are data access committees so you need to submit an application they will review your application and it there is no cost factor involved here here other than your research interest.

So, once you submit that so they will review your application and see the research statement and they come back to you with their decision. And there is the expectation that you adhere to the guidelines proposed by those kind of repositories. So, what happens this is all patient related data, so patient privacy is a primal to the all of the data sharing.