Interviewer: On, alright, so we're recording now. So, before we actually dive into the drawing part, can you just give me kind of a high-level, what is kind of the major topic of your research and some of the main problems you're trying to solve?

Researcher: So I'm trained as a basic bench scientist and I do immunology related work. So I collect cell samples and usually I analyze it, I used to analyze it in SAS a little bit while during my PhD years. My current lab employs R and I'm completely new to it, but we have a data analyst who helps me and she's going for a PhD. So in the next six months I'm going to be in charge of you know, doing the analysis myself too and I want to learn the skills so it's easier when you're actually doing the analysis too. So I have tried doing R and a little bit of basic coding and things there is a R tutorial that I have tried, it's called Swirl, which has been useful, and I've attended a couple of seminars and all that but it's still pretty basic. And I want to improve so that I can actually do the analysis, basic stats.

Researcher: I'm also using some unsupervised clustering machine learning these things. And now I'm actually dependent on GUI, so any package I use I need a GUI. If I want to move to basically, I can code and I can do it. It improves my chances of getting a better data out of it.

Interviewer: Great, and we'll talk about all of that in more detail. So the immunology work that you're doing, what is kind of the main goal of that, what kinds of problems are you trying to solve?

Researcher: So, I'm studying a disease called [disease], it's a lung disease. It causes breathing difficulty in people. It usually affects people who are cigarette smokers and also if you're exposed to smoke, outside or indoors, can also cause [disease]. And I'm studying the lung immune cells in relation to [disease]. So, there are a lot of lung immune cells which get exposed to the smoke that we inhale and they act differently in [disease] patients compared to a healthy, non-[disease] or non-smoker. So I'm trying to differentiate what are the types of immune cells in our lung which are reacting to it, how are they changing and can we figure out which cells are affected most so we can actually start and you know develop therapeutics directed to that.

Researcher: There's very less known about it because these cells are difficult to deal with so there's this relative new technique called CyTOF or mass cytometry. The advantage is that there are more markers so more in-depth data. The side effect is that it is high dimensional data so you need data analysis tools to do it.

Interviewer: Mm-hmm (affirmative), you can't just open an Excel file and yeah?

Researcher: Yes. I mean if it would have been easy but it's like too many markers and clinical parameters does it all so we have like three-four questionnaires which go with the data, a lot of different types of data like: functional data, the descriptional data, all that. So it all needs to be compiled together and analyzed together and you can't do it manually. You need a machine.

Interviewer: Yeah, no that makes sense. Great, okay so let's get started with the drawing. So basically the idea here is we've got this kind of rough template. Covers data acquisitions; all the data that you're gathering; processing if you do any kind of cleaning, more you know, QA kind of checks; analysis, your statistical, any kind of analysis you do; and then your outputs, looking at you know are producing a paper, a poster, is there code or data that you're sharing. But this is just kind of a suggestion, so it might be that you don't do anything here, it's all in the analysis stage. You know it's really up to you. Just to give you some ideas of what these can look like this is an example for more kind of data/science pipeline but you know we've got two data sets that they're merging together. They're doing some data cleaning here. They're processing it in kind of a inter-loop with Stata, "Do I have enough data, do I need more data?" Then they're writing their paper using LaTeX which is a programming language.

Researcher: Yeah, I use LaTex too.

Interviewer: Oh great, and then they submit the paper, and then fame and fortune is rained upon them. So that's kind of the-but yours might look very, very different. It's actually quite interesting to see how different everyone else's is looking so far. But yeah so, and we'll just talk about it as you draw and here we're kinda paying attention to what are the tools that you are using for each of these parts. You know, are they open source, are they closed? That kind of stuff. If there's code, where does that come from?

Researcher: Yeah

Interviewer: Yeah, so feel free to dive in, you can start wherever you want.

Researcher: Uh-huh (affirmative). So this is what I am currently doing, and not what I want to do. So, data acquisition, it's basically mass cytometer output file which is... basically called a fcs .fcs file. I would also have a Questionnaires. What else do I have? So I would also have other experimental data.

Interviewer: Mmm hmm....So the questionnaires, those are from-

Researcher: Clinical data-

Interviewer: Okay in person.

Researcher: Yeah in person...Kind of spell questionnaires...Experimental date, what else do I have?

Interviewer: What kind of experimental data?

Researcher: So mass cytometry is one experiment. I also have other functional experiments for, I actually treat the cells with something, see whether they take up bacteria. Things like that... I'm sorry. We also have like the patient information as more than in the clinical data, we have like lung function test. So those kind of clinical data. So, this is me, my main input, and I try and merge them all in a-well actually my data analyst in lab does this, she merges them into a format which is basically a CSV file with all of them in it...

Researcher: And then we actually-main things that we are looking for is looking in the clinical characteristics and categorizing subjects on patients and then we look at how they change with the functional data or the experimental data. So basically group correlated clinical data and experimental data and tried to figure out, okay, is there a particular group of subjects with a particular clinical characteristic? So, basically a phenotype.

Interviewer: I'm so sorry, so this merging all the data together, that's your data analyst in your lab does that? Okay.

Researcher: And in the data processing, basically all these files need to be quantified, so we have standards for it. So, this particular experiment is the main computationally intense one, because we have several patients and we can't run all of our data together. So we barcode them and run sets of them, so we have to normalize across and there are packages that we use for it.

Interviewer: And what kind of packages?

Researcher: Two of them are MATLAB based packages, but GUI.

Interviewer: So, they all have some kind of interface that you kind of click and-

Researcher: Yes, Mm-hmm (affirmative)...

Researcher: For basically data normalization and also what we call de-barcoding; because we basically cluster like ten or twenty samples together and then run them all together. It goes in as a set and then you have to de-convolute the data.

Interviewer: Got it. Okay when it comes out.

Researcher: ...

Researcher: So that's the data processing and from there we take the normalized data and then try to look at data analysis platform. So this is a discovery project so we don't know what kind of cells we are expecting. So, I did it using the normal data analysis process which is a classic flow cytometry package called FlowJo. Are you-

Interviewer: No I am not familiar with that one.

Researcher: It's basically, there is a UCSF license that we buy every year, and it’s what immunologists use usually. So that's the basic one.

Interviewer: Mm-hmm (affirmative), so this is a commercial tool that people have bought?

Researcher: Yeah commercial...

Researcher: So it gives us an overview of what our data looks like and then we use another package which is also commercially available called Cytobank. Cytobank is very specific for usually mass cytometry when you have multiple things. What is lets you do is, it lets you do what FlowJo does; it also lets you do clustering. So again we-like you give them- "Okay, I want a cluster based on these characteristics." And it does some unsupervised clustering basically like a t-SNE. It’s like, I want to say, it basically tries to connect all the cells which have similar characteristics. There are multiple methods to do it. Sometimes its density based clustering, so they try and separate based on the density of the particles; if there are more particles it gets separated. Sometimes its force director, sometimes its hierarchical. So different types of clustering. There are specific packages on the Cytobank to do that.

Interviewer: Got it. OK.

Researcher: So that's what I have been doing. Cytobank is great but it has its limitations so the kind of package I was looking for it wasn't there. So we decided to use R package for that. So I don't know whether all this makes much sense?

Interviewer: Yeah... so you were mostly using Cytobank but it was missing a package that you needed.

Researcher: Also, Cytobank I can't actually upload all my clinical data. It lets me analyze the cell data, the .fcs cell file data, but if I'm using R I can actually input other battery tests.

Interviewer: It just doesn't have that capability. Okay.

Researcher: And it’s also like cloud-based. I don't know how the privacy parts are I never tried doing more.

Interviewer: So the clinical piece, it's not really going to work.

Researcher: So right now I'm actually using Cytofkit package, which is R-based and it's free. All of these, FlowJo and CytoBank I have to pay for. The R-based packages it’s all on GitHub and it's just downloading from there. Luckily Cytofkit has a GUI, but I need more output from than that. GUI doesn't give that so I'm trying to do the coding for-

Interviewer: Got it.

Researcher: But no success.

Researcher: Yeah, so that's for the experimental part of it, but there is also a stats parts of it where actually we put all the-so I'll get all of the unsupervised clustering, I get phenotypes of cells, and then I correlate with the clinical data. So that's the basic stats that we use. So usually I write the tests but it's not very useful, so we do regressions with multiple clinical parameters of interest.

Interviewer: Mm-hmm (affirmative),and what tool do you use for that?

Researcher: So my data analyst does that, I want to do that myself too. She uses R, sometimes STATA. So those are the two.

Interviewer: Yeah, write that down. No I think that it is important to distinguish kind of what you do, versus what someone else does that you'd like to do too.

Interviewer: I mean we can mark that the data analyst does that piece or, and this one too, just so we remember. It's nice to have a data analyst.

Researcher: Yeah, I'm so sad that she is going to be going away. So during my PHD my advisor was really particular that I need to do all my analysis. I was forced to learn SAS because of that, because I had no idea how computing works and I took classes and you know I got an idea and now I am dependent on my data analyst because, hey she can learn it right away...bother so. It's kind of sad, but-

Interviewer: You are doing all of this other stuff so, it's easy to see.

Researcher: So this is where we are in terms of research output I have a couple of abstracts and presentations and I am writing a manuscript.

Interviewer: And what tool do you use for that?

Researcher: This is the first time I am using Latex, I am trying to covert my boss for it. But he still wants a word document so its been DOC. But yeah my husband does computation stuff and he introduced me to Latex. It's really cool.

Interviewer: Nice. Very cool. But still kinda having to do WORD at the same time?

Researcher: It doesn't actually go together.

Interviewer: Yeah, no that's pretty common. And of any of this code that you are producing for the analysis or for merging the data, is any of that every shared with anyone or is it kinda kept within the lab?

Researcher: I think it's nothing novel about the analysis that we have done. So we have no reason to share it or anything. The analysis methods will be published with paper but I don't see any code by itself.

Interviewer: And that's not something that your journals ever asked for?

Researcher: Not so far.

Interviewer: Okay. Cool.

Researcher: But I think we would be happy to share if somebody asked for it. So I am trying to use, I can't remember, there's a page on R where you can actually write the description and then record.

Interviewer: Mm-hmm (affirmative), Are you talking about CRAN?

Researcher: Not CRAN Its like within R there is a window where you can actually do this. Somebody showed it to me, I have yet to try out.

Interviewer: Yeah, there's lots of great R packages out there. Alright so looking back over this are there any other steps or tools that haven't talked about?

Researcher: I think this is mostly it. And all of this data, it’s like, these two are clinical.

Interviewer: The questionnaires and the lung function. Uh huh.

Researcher: And the other two are experimental.

Interviewer: And where do you get that clinical data, is it from like the EHR?

Researcher: No, we have in-person interviews for questionnaires and lung function data have a pulmonology, ours is a pulmonology lab so we have a pulmonary function test we do.

Interviewer: Okay, so you're actually, are you doing those in-person?

Researcher: No somebody else.

Interviewer: Alright with some of this code that you are using for the cleaning do you do any kind of version control with any of that?

Researcher: Um. No not really. Well its only with a particular R version. I think its 3.4.

Interviewer: In terms of like tracking, you know as you make changes, tracking different versions of it?

Researcher: No I haven't actually. Is that something that I should be doing?

Interviewer: Well you'll learn about it in the workshop.

Researcher: I know some of them don't work but I have a couple of Cytofkit when I started and I couldn't use it in my lab computer because it was a 3.4 on there, it talks only with a 3.3 so I had to, so those things, but now I have to figured out how to use different versions.

Interviewer: Well that's good. Yeah I know that kind of stuff can be really frustrating.

Researcher: But I haven't actually paid attention whether they updated it or-

Interviewer: Mmm hmm, any other tools or processes?

Researcher: No I don't think so.

Interviewer: Okay, so now I just want to get a sense of how you feel about most of this, so is there anything that feels frustrating or something that you maybe want to learn or do more of or

Researcher: Mmm hmm, so I am at this stage. I would definitely want to know how to do, especially, I can do coding for a small sample set, but when there are multiple parameters, how to prepare the data, I'm still confused. Somebody has to load the data for me and then I can do the code. So I am like a total big nerd so I want to do the basis statistical analysis. This is something I need to learn so then I know what kind of analysis goes in and you know, that it's actually adjusted for smoking or what ever I want to and I think my stats is kinda weak now it’s been several years since I had a course so.

Interviewer: Yeah and here its not so much knowing which tool its actually knowing which test to run and stuff like that, or it is-

Researcher: Well I kinda know the tests but I don't know this platform to run the tests. I can probably do it in SAS.

Interviewer: Mmm mm-hmm (affirmative) and is SAS the one you are thinking of learning or doing more with?

Researcher: No I think R is more user friendly than SAS. And there are more packages and helpful like Github. And you know if there are issues, they are more responsive. SAS was never that great. I want to convert to an R person.

Interviewer: Great, any other pieces of this?

Researcher: Maybe, this part too.

Interviewer: So in the data processing merging on the data-

Researcher: And you know of the way I enter the data is sometimes, my data analyst goes like, Uh I have to change manually now because it doesn't actually recognize it in the current format so it’s like there is a gap between acquisition and normalization or you know processing.

Interviewer: Mm-hmm (affirmative) So that's something about the way you are structuring it isn't quite working. Trying to learn more about that. Any other pieces, that-

Researcher: I would like to learn more about the Latex, it's new so.

Interviewer: Yeah, and what are you hoping to do with that?

Researcher: Actually write my entire paper there so I don't have to go back and forth. And I am still struggling for the table insertion and things like that.

Interviewer: Oh yeah, like putting in your figures and stuff. No I haven't used Latex myself but people love it so.

Researcher: It actually great, it looks like an actual paper from the first draft. It's just awesome, it makes sense.

Interviewer: Yeah, yeah. It definitely feels like professional when you are doing it, which is fun.

Interviewer: Alright so then yeah the last piece of this is you know, we kinda talked about some of these things. But is there anything else here or in general that you are hoping to learn in the workshop?

Researcher: I think basic R programming is what I am looking for, so I am comfortable with doing R myself and you know. It used to be like six months back if you asked me, I would have said how to insert, like how to basically download packages. Because I was struggling with that but now I have all the packages, I just understand the system more. Now I have to actually learn how to program.

Interviewer: Okay so general, how to program.

Researcher: Yes, yes.

Interviewer: Anything else?

Researcher: I think that's about it.

Interviewer: That's not at all. Awesome. Well you definitely get a lot of that in-

Researcher: Okay.

Interviewer: Okay so the last piece of this is just a little checklist. So this is basically just some behaviors that we kinda try and teach in the workshop. So trying to get a sense of where you are at now. And we talked about a lot of these things already but so in your current workload do you use any programming languages like R, Python or the Command Line?

Researcher: R

Interviewer: Do you have any kind of step by step work flow that you have transformed into a script? So instead of like change this one thing and then upload this other, change this other thing, it's all in one script.

Researcher: So usually I just follow the Github portion and they have it step by step but I just follow it. I never have to make it myself.

Interviewer: Okay. So we'll just say no. Do you use any version control for your code.

Researcher: No.

Interviewer: Do you use any open source software?

Researcher: Yes. R.

Interviewer: Do you ever share your code publicly, like outside of your lab?

Researcher: I've never had to.

Interviewer: And do you share any of your, kind of overall computational workflow or protocols?

Researcher: I think we have collaborators who have asked for it. But that's not me it’s the data analyst who does that.

Interviewer: Okay. So you yourself haven't like ever had to publish any of that? Alright. Well that is pretty much it. So I'm excited to come back in June and we can talk more about this and see if anything major has changed, you know if your Latex , your PI is all on board with your Latex or not and all the other pieces of it. But thanks so much for chatting with me. I think you will have fun in the workshop its two very intense days. So you will feel like drained but, you know its super worth it so.

Researcher: I want to learn it so. My father-in-law is like a statistician and he works with R and he teaches R and he is like its a simplest ever. And I am like, no.

Interviewer: Yeah its like, if you have never done any programming, its very like-

Researcher: I wish I had actually started when I was a little younger. In my early college days.

Interviewer: Its really important and thats you know, you're kind of in the same boat as everyone else in the other workshop. They come to use USCF, especially and they are like, Oh, I'm supposed to like work with all of this huge data and I can't do it, so it feels that way. Definitely a lot ready to learn. Awesome. So this is for you.

Interviewer: Thanks so much.

Researcher: Thank you.

Interviewer: Turn this-