P1

Author 0:01
Okay, so yeah, via recording. So I would just like to start with, can you talk about your usual data analysis process? Like, what sort of tools do you use are like any specific environments you work in?

Participant 0:14
Yeah. So typically, I'll run in lots of different environments. But I do a lot more in Python. When I'm doing more data, managing things, data, linking all that kind of stuff, my statistical analysis will either be in R, or with SPSS, and most of my data visualization is in Tableau. So I kind of go through all of those, usually, it's a lot of Python or R to get a data set, then I usually will drop that to Tableau to start looking at the relationships visually. And then when I'm actually doing modeling, then I'll switch it back to our SPSS to do the analysis.

Author 1:01
So since you mentioned you use this multiple tools, do you switch? So when switching between this multiple tools? What are the challenges you face? When you carry over your analysis from? Like, let's say you do something in Tableau, and now you want to move on to like an R node to correct Patenaude? What are the challenges you face? Like? Are there things you have to redo in this new environment? Again?

Participant 1:25
Yeah, you know, I think the big ones are for, it's definitely dropping it down to an analyzable data set. So and replicating essentially a lot of the filters and the sorting and all that kind of stuff. So at least in you know, Tableau, you can kind of get this visual sense of I want these categorical variables, and you've got these facets, you know, that you're looking at gender, males, females, whatever. And then having to essentially write the code, obviously, to model in our that, okay, these are the ones any filtering, you know, because Tableau has got the filtering So, right, you know, when you bring that back in dark, then you have to use your filtering your sub filtering, whatever to get it to an analyzable. So that's, you know, all replication, you know, right. Good thing, again, is that Tableau, at least I can translate it, because this, here's the visualization. I know what filters there are placed on the data, and then replicating that in our SPSS or something. Right. Yep. That's kind of the drag.

Author 2:42
Yeah, that's yeah, I like that. So usual challenge when working with this multiple tools. So okay, I have like a few series of questions, but we'll break them down the first let's focus on the script part like where you were with where you actually write the code.

Participant 2:57
In R, or Python, in any any

Author 3:00
any language like basically, where you write code, versus like using something like Tableau, like will speak broadly to different paradigms? Or do you think there are certain tasks that you would find easier to do in Tableau versus in your our Python environment?

Participant 3:22
Yeah, I mean, again, it's all about, like regrouping data, that's always been a huge challenge. Well, not a challenge, you know, but it's kind of a drag to, you know, if you have five different races, and you only want to cut it in a to race, you know, whites, non whites type of thing. Yeah, you do regrouping in python, but then you always forget your variable name. And then you're always looking through your data frame for what is it type of thing, where Tableau, it's easy to do the groupings, and it just kind of makes a new variable right underneath it. So it's a little bit easier to manage. When you need to manage, like in a spreadsheet than Tableau is much better. But yeah, that that those typical coding types of things. It's a bit of a drag in our Python, because you know, it's syntax, and it's just a long script. And I think I could be better in my coding for sure. But I'm always lazy, so

Author 4:29
that and it's not it's like a common thing, right. Like, programming shouldn't be a problem. Like the task is the main thing. Yeah. Okay. So do you think it would be helpful for you if there was a way to integrate this, for example, to integrate the see if somehow we could integrate Tableau and its feature within your, let's say, an AR notebook, where you can just quickly do your interactive things and Have those results directly usable in the next code snippet?

Participant 5:05
Oh, gosh, that would be Yeah, that would be great to be able to do that. Right. You know, it'd be, it'd be kind of like, if I was grouping, grouping variables, you know, type of thing, then it's essentially would be a great code generator, right? Yep. That it would, you know, I want these three grouped together in this one group, these two in this group? And to be able to actually have it code that in our Python, yeah, that would be lovely to do that. So I can kind of see that as a recode. Section. Right. Right. So, you know, kind of following your workflow? It, I can see it chronologically, right. So sometimes I would, you know, and that's pretty typical, that we're looking at a variable like race, and we look at it and we go, oh, yeah, there's not enough Asians, so I'm going to combine groups, write a little code to do that, then I would look at it and say, Okay, that looks good. And then I can go down the road for a little while, right? With that one variable. And then then I'll want to recode some other like, demographic, you know, ses income or something like that. But I don't want to lose all of my breadcrumbs from after doing the race recode. And I'm looking at stuff and I'm, you know, doing some panel graphs, and then, oh, let's bring income into the equation. I have to regroup at that point. Yeah, if you did it chronologically, then you would have that, but it would also make sense that any kind of recodes whether it's raised or income or whatever, go into kind of the Recode syntax section type of thing. Right, you know, so, you know, having both, I think would be great. But also, but also being able to functionally go, here's the Recode section, here's the analysis section. Here's some panels, you

know, visualizations section, and then a chronological one as well. I mean, that would be super bowl.

Author 7:27
Yep. Okay, so now, let's come back to like the small discussion on the tableau part. So, our Okay, so this question actually applies for both your working environments like Tableau and computational like code. Do you keep it? Like, do you keep a track or a record of your analysis sessions in some way?

Participant 7:53
Further, yeah, for the most part, yes. So again, it's kind of like this exploration piece. Are the variables working? Like, they should be kind of working? Yeah, at some point, the analysis is like, Yep, this is going to be the published one, the manuscript the astral.

Author 8:14
And how do you do that? How do you keep like, basically, like, in Tableau? How do you track like, these are the 10 apps? ID?

Participant 8:23
I don't do it in Tableau, because I don't know, you know, is not, has does not have a great tracking ability. Right. So yeah, I would not do that in Tableau. Tableau really is much more for the exploration and or final product of graphene, you know, I've, I know what I want to produce. I'll either produce that in Tableau or SPSS or gg plot or something like that. There. Yeah, there. The history for in Tableau is no good. And I would not do history.

Author 8:57
But would you want a way to track your exploration in Tableau, let's say, Tableau, let's say there was a tableau plugin that could show you a graph of, Hey, these are the steps you did. And this is where you branched off to do something else. Would you like a feature like this?

Participant 9:16
I think it would be great. Yeah, absolutely. If there was some, you know, I think having that function would be great. Whether or not I would use it, it would be the same with like Git. GitHub, right? It's like, when do you fork when do you not, you know, it's like, oh, it's like now I'm not going to do it. And I'll go a long ways, and it's like, oh, shoot, I should have done it. forked it, you know, two days ago type of thing. I'm too lazy to go back and do it. So it's a lot of that same thing, having that, you know, functionality would be great. Whether or not I would actually be cognizant and use it that way. Hmm. I'm not sure because usually what I do instead is do essentially the same sort of thing by just copying it. Right. So, yeah, I think this is pretty good. I want to go a little bit different copy paste it, and then, you know, work for a while and whether or not it becomes a fork or not. I don't know. So that's, that's kind of how I typically would work with with something like Tableau and history.

Author 10:27

Right? Yep. What I want it. Oh, boy.

Participant 10:31
It's a good question. I mean, I? I think it would be good. Yeah, I think that it could be a really great function, especially to kind of cobble them together, right? When you copy that, that Tableau spreadsheet, then you start the history for that spreadsheet, right? So then you could kind of at least get back to where you were, when you forked it from the right spreadsheet, or something like that, I think would work pretty dang, well. Yeah. Right.

Author 11:08
And okay, so my next question is follow up with the support, would you? So let's say there was a way that you were able to track your analysis session in Tableau? Like, what information would you like for it to track? Like, just what your interactions were? Like? What sort of things we would you'd like to keep a track of this history? Something that would you you know, you would like to revisit, often, you know, maybe read jog your memory?

Participant 11:39
Yeah, the, the recodes. Again, the recoding of variables would definitely be the different types of graphs cuz I'll switch, you know, from a scatter to a line to bar charts to right, you know, paneling with the different variables. Yep, definitely. Things like that.

Author 12:02
Okay. Um, so, yeah. So now, do you have parts of your analysis sessions across like any of the platform, which, you know, like, there are certain things that you have to repeat, like, they are basically, like, recurring parts of your analysis that you have to do every time your data set updates?

Participant 12:27
Yeah, which is table one, the demographics table, that is like, you know, the, the one that always changes, because, you know, you're always retreating, another subject and percentages, and accounts always need to be modified. You know, the, again, a good cutter will probably do that, and are in Latech. And, you know, be able to do that I, again, get lazy and do that. Plus the Latech tables, I just, I'm not the biggest fan of them. Yeah. So, yeah. So I don't know if it's, you know, obviously, it's not a data visualization. But gosh, Table Table, one demographics table is absolutely something that absolutely needs to be, you know, syntax driven. And if we could do it on dynamic data, right. That's the big one.

Author 13:29
And so, okay. Is this like, Is this one of those take table? Or is one of those data set that keeps updating over time? Or, you know, get updated frequently?

Participant 13:41
Absolutely. Because what tends to happen is, is this even when the study is done, and we're doing like secondary data analysis, what, what always happens is that even if if we recruited

100 people, right, you would think that that demographics table is always going to be 100. And everything's going to be static. But what happens is that the PI's will always do conditional stuff, they'll they'll always say, you know, I only want to look at breast cancer. So right there, your 100 goes to 40. And then you'll do the analysis, and then they'll go, oh, you know what, we need to exclude these two because they're actually geriatric patients. So you're 40 goes to 38. So even though you weren't, you're not recruiting anymore, what happens is the analysis always changes because of conditions being placed on the sample. So and again, those the visualizations will change. But again, the the reporting of the dimmer graphics table will always change and even with one one small person, it's, it's a drag to have to redo all of those.

Author 14:53
Oh, and that's exactly what's going to be next question like, does this take even these changes like, how much time How much of your analysis you have to redo like, I would assume something like R or Python. Most of the time, it could be as simple as rerunning the script, again, like there might be parts of your analysis where if you change the input data, you can rerun it. But even in those cases, like, change of data might basically compel you, like, hey, this piece of code or this, like, this condition I used for filtering doesn't make sense. They updated it. Now I have to change it. How often does that happen? A

Participant 15:31
lot all the time, right? Yep. All the time. The filters, the again, the groupings will run the model and go Geez, why?

Author 15:45
Sorry, Participant, you cut out there, I can't hear you,

Participant 15:48
again, that the variability in the three and Indians together doesn't make sense. You know, let's, let's regroup that again. And once you do that, one, your your demographics table totally changes, right? Because now you had race as three variables. And now you're going to go to two variables. And all of those numbers change. So we're going to do it in late tech, once, when we recreate a new variable, then we have to go back into the link tech, take out the race, three categories put in the race to categories. Yep. So it's yes. So you can kind of see the the issue is that we'll do the analysis, we'll think through things we'll get to this thing, then we have to essentially go back and recreate all the the demographics, tables, and any of the work we did beforehand. Because you can you can run, you know, essentially one analysis on one outcome, let's call it I don't know, blood pressure. And, and then you do something for like, diabetes, a one see a different outcome. And you you slightly change your your people again, then you have to go back and rerun the blood pressure analysis again. So that that is all of the process of redoing everything. And even with coding. Again, it's it's this categorization of variables, having to propagate it go back recursively, right to go back up, and then propagate it through your first analyses. Yep. Unless you're really good coder and can keep great, but that's not that case.

<demo>

Author 38:59
Right. So in general, like so like, first, like a very general question, like, what do you think of such a technique to capture and potentially, like, share your analysis workflows?

Participant 39:11
I think it's great, I think the the, so you're going to get a lot of, you're going to get people that are not going to like this, because essentially, you are using k means clustering or some clustering algorithm, right? And people are going to essentially say, you're essentially making up your groups on your own. So from a real true statistician viewpoint, you should never be creating your own groups because that's, you know, you're just doing it willy nilly type of thing and with good research design, your group should always be established type of thing. And understanding that, you know, the first thing would be outliers and you see these two or three and use Some sort of, again, some k means clustering or some clustering algorithm to say, these are our outliers type of thing. It's, I love how quick and easy it is. Right? I think it could be a little bit too easy from the aspect of you know, it's almost, when it's too easy to, to, to, to redo the data, and you see, oh, yeah, we're going to filter out these outliers. The first question to me is always, why is that an outlier? You know, as opposed to allowing the algorithm take it out? I want to dive in more and make sure that outlier should be an outlier type, right. So that that would be one of the first things that I would, it's almost, it almost makes it too easy to do. You know, especially with outlier detection, people would just use it way fast, which, you know, helps them, but it also could be bad, because they really don't have a good reason. Right, that they're outliers. So that would be kind of my first comment on on being able to do that. I really like it for the data wrangling. It's all I really think this could be good in data cleaning peace, you know. And one thing, especially with this label, a label be how, how I could see it is, again, it's almost utilizing the brushes to create your groups, right? So you brush these, these would be label a, these will be label B, right? The what would be, what would be great is instead of you have label a true false labeled V, true false, is that you've got a new category, right? Right, is a or b? Yes. type of thing. So it's one column one variable that is distinguishing A and B as opposed to two variables, right? Binary, so that

Author 42:25
we actually just, I'll just load something to show you an example here. Yeah. So in there, we have some different datasets here. And we support things like this, for example, we call those as categories. Yeah. And for example, here, like we already have some categories repre assigned, but what we love is, for example, we can create a test category. Great. And then you can add options to it. So let's call options A, B, C, and D, just for now. Great. And then basically, now all the points are circles, which are unassigned. And then you can like, select some points, brushing them. Yeah. And that categories them at certain things. Yeah. And then when you load this in Python, this appear as a single column with either ABC or D or unassign. Great. Yeah, like that is so that can be supported by the technique. But like, again, so directed, like, this is how we chose to implement this in this prototype. But overall technique just supports, supports that inherently right, the technique itself doesn't rely on this particular implementation, you could

potentially do this in Tableau, and then extract that information from Tableau and save it as our workflow. Like our contribution is the concept of saving, or capturing the correct information to save as a workflow, rather than like this specific prototype tool, which, like, you know, which has, like, this is how you do it, this tool, you might do it differently, but the end result will be the same.

Participant 44:09
Yeah, so show me what the workflow is for that recategorize ation that we heard. workflow is that

Author 44:17
so for example, here is how the workflow looks like right, we added a scatterplot selected point, we categorize selection. Yeah. And we could like create a workflow, like minder

Participant 44:31
categorize like this.

Author 44:35
And then once we can add this and then like, it saves it automatically. And then to load it in Python, we could do something like for I don't have the data set loaded here, but I could just load the workflows. Yeah. Let me see if I get this correct. Are Gapminder work? I don't actually Remember the name of the project here? Yeah. So it has our Gapminder categorize workflow here, which we just created. And then when we load it, it basically shows up like this. And then if we want to reapply it, let me just load the actual dataset.

Yeah. This Yep. And then let's load it in our target.

And now if we, we didn't update the original Sorry, just commented for, and then when we reapply it

so yeah, when we reapplied it, essentially create this test category, and then assigns it a CB or unassigned. Here, like here is the new column we added per test category

Participant 46:17
that that n and we we are doing that on the data set that you brushed, correct.

Author 46:27
We are doing that on the data set. We brushed like a Gapminder. Here.

Participant 46:31
But what if you took Gapminder? Yeah. 1980. And ran that?

Author 46:38
Yes. So we can do that. So here, I'm actually loading 2010. And let's try to load 1980

Participant 46:46
here, right? Yes.

Author 46:49
So we just rerun this, and then okay, let's actually, you know, load both of this in original at Target. So we can just compare. 2010 is the original and the target is so. So we have for our 1980 dataset, we have 129 rows, and for our 2010 we have 178 rows. So now if we rerun this here, I'll just uncomment this because now we have both datasets, okay, update the label and rerun this, we can see that this runs correctly on die, it tries to basically reassign this to it tries to label whatever the countries are in 1980 data. Right? So, and again, like here, the workflow is basically selections and categorization, right? If we replace one of the selection with something like algorithmic, like, let's say, a cluster of countries, it will try to calculate the same cluster and then label it.

Participant 48:00
Yeah, so on that 1980 You're just going to get more unknown signs cuz we, we didn't brush it,

Author 48:07
right. But yeah, like so the, basically the technique, the underlying technique does support doing something like this. Mm hmm. Yep. But yeah, so my, so the questions I have are more on lines of like, if we can imagine this technique to be integrated in let's say, Tableau, where now, instead of using this, like clergy tool to generate workflows, you could use all the features Tableau has and you get your analysis saved like this. Yes. So basically, like I have, I want to like see, like, how some of these features if applied to your work and your environment? How what what are your thoughts on that, for example, if you could capture your tableau analysis as like this branching series of analysis graph, how useful that could be in your current setup

Participant 49:06
yet that's a great question hmm man, I do like it. Again, I think it's, it's a little bit I know that that true statisticians true biostatisticians and not gonna like it because again, it is to have the ability to create groups so quickly.

Author 49:54
Right? So like for this question, like I'm not focusing on like creation of groups. So I just basically want to focus on anything you interact with is stored as this like graph. And you can at any point, go back and do something else. And it creates a new branch. Yeah, that that is sweet. I really like Yeah. Yeah. So basically, I want your thoughts on like, how having something like this in Tableau, let's say, would like, what are your thoughts on having something like this in Tableau? And how would it affect your workflow? Would you use it?

Participant 50:27
No, I like branching for sure because I think that's kind of what I typically do with, copy the spreadsheet and then go from there. And Tableau has the ability to go backwards, right. So essentially, I can go forwards and I can go backwards. But I do like the, the branching piece

from here, because it's visually a lot easier than to click the back button, or forward button, because it also is telling me a little label of what changed with that. With that Tableau piece, otherwise, I click the back button. And it's like, I know, something changed. And I can't remember what I did before. So having a little label there of how it changed from, you know, the mouse click, that I think is really nice. So selected three points, like 19, you know, add a label. However, those things that that branching with a label of what changed is, is really good compared to Tableau, which is really the forward backward button, right?

Author 51:38
So I want to come go back to before we did the demo, we had a discussion on similar things, right, like, so you mentioned that, for example, something like good you get, you basically sometimes have problems, or you sometimes don't decide, hey, I don't want to for now I just do my analysis linearly. And then you think that, oh, I wished I had created a four to, you know, have a parallel analysis running. So would an interface like this, make it easier or entice you more to use this branch chain analysis, where you basically essentially just have to click and do something else? And the branches are added automatically? You don't have to think about I should create a Branch at this point. Would that basically entice you more to use features like this?

Participant 52:35
I think once you get used to it, for sure. Because again, I kind of like that, I could go back up to select 53 and branch off of that, right, and still leave select 19 points for later, right. So I could essentially, click that plus and right, go back there work there for a little while, and then go back to select 19, you know, venture down 53 for a while ago? No, that's not quite right. Right, you want to go back to and start on 19. And work in that and branch from that one. So yeah, being able to visually see these demarcation points, and be able to then, you know, dive down that rabbit hole if I want to, for a while. And, you know, I kind of like the big X, or, you know, which is telling me, No, we don't need to go any further down this road.

Author 53:34
Something here? Yeah, we actually had that something like this, right? Where you do the

Participant 53:39
exactly, it's like this, this is about as good as I could do from this one. It's not quite, quite right. Let's see root Yeah. Because this is yeah, these are all different data sets, right?

Author 53:53
Yeah, these are two different datasets. So, okay. So, next. So, you already spoke about this, but basically, I think initially you're trying to make this point about how statisticians would like this automatic generation. So coming back to it. So, one of the features like why these algorithms are predicted automatically is what we call semantic selection, right? Use your selections to detect a pattern automatically. Yeah, yeah. Like now, like, this is where like, I would like to hear your thoughts on like, when this would be useful or if not like, What will your concerns be using something automatic like this?

Participant 54:37
Yeah, just like what I was saying before and if I do it algorithmically, with some sort of with any kind of algorithm, calculate that. Then I, if I switch to data sets, I may be losing something Some, some something, you know, if you switch data sets from from this thing and you go to four or something like that, or, you know, as you said, the red ones up there are being used as one single Kmeans cluster. Right, right. And it'd be like, I don't, to me, there's something interesting about there, there really should be two clusters, not just one cluster. And I would want to understand why that data set that algorithm that Kmeans cluster changed, what why don't order? Why didn't it change? You know, what are the characteristics of those, those dots? So yeah, I understand semantically, I'm trying to do some data cleaning, using the K means cluster. But sometimes the K means cluster is going to group things where I don't want it to group I want to know why. What distinguishes that that cluster from another cluster?

Author 56:06
So would it be helpful in this context? If, for example, and again, like these are recorded predictions are just as a way to refine your selections, right? So would you be likely to trust them more? If you could see, what exactly were the parameters that were used while doing this clustering? Or that's not enough?

Participant 56:31
Yeah, you know, what might be helpful is essentially doing, you know, here, here, you you're picking the K means cluster, and you go to a different data set five, or something like that. Now, it's hard to because the K means pick those read, right? Yeah. Because you're just using the same algorithm gone from one data set

Author 56:59
to Well, yeah, so that's what we do. So basically, if you pick an algorithm for this step, we try to apply the suit. That's what happens when you switch here, it applies the same algorithm, and it like detects the clusters incorrectly. Like or maybe like, it treats this whole thing as one big cluster. Whereas you may not want it to treat as one whole big cluster. So that's where you the review step comes in, where you say that, hey, this automatic application is not correct. And maybe like, for example, this particular prototype doesn't support analyzing of, you know, by the clusters will done this way. But like an advanced tool may have that feature as well. So,

Participant 57:43
so if I, if I go to here, let's see where, where are we on the branch,

Author 57:50
the OCR on the Apply clusters?

Participant 57:53
So can I can I? Can I brush from here? Yes, I could, I could brush that top. Top. So

Author 58:00
if you just want to brush the top one, yeah, you could go just back and just start by brushing the top one. Just add a new branch, right there. Okay. And then the predictions will now basically be related to this cluster. Like it would try and select clusters near this. near you. So basically, what the way this the way our technique works is it takes your initial selection in consideration, and tries to basically tell you these are the patterns that are the most near to your selection, right? So it doesn't add like it doesn't automatically determine that, hey, this, these are the five clusters, it takes your initial cluster initial selection as a seed to find what is the best possible cluster? Well, not the best, but what are the possible clustering and tries to like rank them somehow? Yeah.

Participant 58:56
Yeah, it is definitely nice. It's always a step four, for me and for us using these visualization tools. So go back to yeah, get rid of that. That.

Author 59:16
Perfect. Yeah. Just this all

Participant 59:19
go back to like cluster sample for. Okay. Yeah. Yeah. So the, the clustering we, we would brush this for an example. Right? And, and there's something with this. So we've developed this clustering algorithm right here, right? So then you would use that to use on a different data set like simple V five,

Author 59:54
right? So why don't I select one of this? Yeah, Slack Yeah, let's just look this one.

Participant 1:00:01
Yeah, it's like that one. So now we've got a clustering algorithm loaded up. Yeah. Yeah. So then we could use that on a different sample. Yeah. And that is true. It is using that algorithm.

Author 1:00:18
Yes, it is using the same algorithm. So like, so it is using like a DB scan with epsilon of point one and min samples of phi. Yeah. Like we do. So this is like, since it's a prototype, we don't have a nice way to show the parameters here. We just show it as a huge dumb, but yeah. But yeah, so essentially, when you switch it tries to apply the same. So it does actually. So like it does basically try to apply the same thing at that side get something wrong. And this is the part of the technique where we are trying to think through, should we apply the same clustering algorithm, or should we judge. So essentially, what happens the first time is you select something, we run like 20, different clustering algorithms with different heuristics, and then try to find the one that closely matches your selection. Right now, if you select one of those, the way we currently approach this is we basically, when you switch a data set, we apply the one you selected directly to this new data set. The other approach we are thinking of is, should we just run our 20 algorithms or the new data set again, and then pick one of the then pick one of those which

closely matches your initial selection? So these are like two different ways, right? Like you, once you pick an algorithm, you stick to it, or you try to match the points? Or try to match your selection in general. So between both this tooling, and like, at some point, you know, we might we figured that we might just give users an option. Like, if this was an actual data analysis tool, we might think of, you know, give user like, the way we have this last year, give user an option, hey, this is what happens if you apply the algorithm directly. But this different algorithm actually fits better. Which one do you want to pick now?

Participant 1:02:20
Yeah. Yeah, and I don't you know, cuz I don't I don't know if you would need the the clustering and peace across all of the different data sets. Because to me, to me, it's almost these datasets to me would be essentially, like different studies or different. Totally different timeframes, right? For three months, six months, nine months type of thing. I don't know if I would say I want to really use the clustering algorithm from dataset one on dataset two, because they're so different. Right? Ours, our thinking is very linear, have one data set. Let's do all of this kind of stuff. Let's subset let's do the filtering, faceting regroup data, go down that for a while, go back up, you know, as you know, using the history again, and go down that one for a while and then up, go back up and then down another one while but actually go from class, you know, taking from simple one, V one to V two, three. I'm not sure if that is as important data set the data set, right? Yep. But within within the data set, right. Having that history is really good.

Author 1:04:07
Right? Yeah, that makes sense. Okay, so Okay, so now like just moving on from semantics directions. The next question I have is, What are your thoughts on the way we generate, like we we create workflows. So for example, like a common approach is basically something like a Python script, that you start from scratch and then you add blocks of code. You have as you know, this part filters this part facets as part visualizes. Yep, that is one way to create a workflow. Whereas the way we propose creating workflows in our approach is you do open ended exploration, you have multiple branches, and then when you're happy with one of the branch you curate that to a workflow to be used in future. So what do you think of this approach of creating workflow where, you know, you use your past analysis as a basis for creating a workflow.

Participant 1:05:08
Yeah, I think I think that's a good approach, I think. So just like this one, and let's go back up to the first add brush. So dance right there. And let's branch off of that for for a second. And add something off of that.

Author 1:05:31
And you can just create a label for

Participant 1:05:34
cheer. Great. So what this tell is telling me is that the green is essentially everything that I've done, right? Yes. And if I go to the gray one to the left, right, now that becomes green, right?

Yeah. But it's telling me there is another branch, we branched off of the first one. Yes. Yeah. And I do like, this green work flow path, is if I curate curate, yeah, curating the green one, right? Yeah.

Author 1:06:10
So you're creating, you're basically curating the one which is shown vertically, like everything else is a branch. So when you go to create, you can like pick which one of the branches you want. And then like, basically, start curating that one. So then when you are, when you go to create a workflow, it essentially strips out all the other branches and gives you just your core branch so that, you know you can decide to, like remove some things from this if you want. Yeah. So yeah, like so the question mostly is like, like this approach of using your previous steps to curator workflow, like, what do you think of like this approach, like visually create a workflow from something you did earlier? You know, like, remove unnecessary pieces?

Participant 1:06:58
Yeah, you know, I like this, because it's much more natural, right? We, we do this, what and also, what I do like, though, is that if I curate this and drop all the Python code for this, right, then I actually could say, You know what, I'm just going to do all my labeling first. So I'll just pull my section, assign label a, and assign label B, and put them in a Python section. Right? So it's more about, yeah, this is how I got here. And maybe I can clean up my code. If you code generate this, then I can clean my code a little bit better. Right? You know, so yep, I think I think it's good because it gives the user the ability to work naturally, in an on the fly, be able to generate the code for it, and then be able to rearrange it, because that's how I work anyway, is that all just all my Python code, at some point, I need to stop. And, you know, get things a little bit more organized anyway. But here, it's like, I've, I'm generating the code while I'm doing it, then pop it to Python, bring the labels together and put the brushing together? Well, well, it depends. Because it has, sometimes it will have to be in the order. Right? Because if you had, you know, brush one, brush two, and then label labeled it from there, then I might have to do do it in order and not labels together.

Author 1:08:40
Yep. Um, but yeah, like so for example, if something like this, right, in Python, for example, you could treat this as one single block. And then instead of like, say, this was the brushing and the labeling, you can call all of this single label block in Python, and just use that directly. That's true. Yeah. So that brings me to like, my next question regarding this feedback is, so the way we currently have this implemented is, when you bring in a workflow and apply it, we give you two potential outputs, right? Like, we first give you this huge data frame, which has like all this extra stuff added to it. And then a clean version of the dataset, which basically has removed all the internal IDs, blah, blah, blah. Yeah. So the goal is like this raw data frame allows you to tweak the generated data frame before you clean it up. yourself. This is something like you could directly visualize, without, like all the extra information. So the question is, would you like is this approach enough for you to let's say, do something custom with this data frame? Or do you actually would like the Python code to create this data frame generated for you as a new block? Like the Python by Python code, I mean, like, for example, like all the Python code to select

these points and assign them a label a the Python code to add a new column, you like to have access to that code generated for you. So now, not just the final result, you also get ready made code which you cannot do.

Participant 1:10:22
That's a great question. I'm trying to think of a reason I would want a Python code because if it's creating the variable for me, I'm happy with this and then can go forward, you know, quick, I'm just trying to think when I would actually want that code.

I don't know, because that code is going to be very specific to each data element. Right? So I don't

Author 1:11:02
hmm, maybe Okay, so let me rephrase this question in some in a different way. Would this code being generated for you, where you can actually look at the exact code that generates this data frame, give you more trust on this automatic application process?

Participant 1:11:21
Oh, absolutely. Yeah, having it available, whether it's right

Author 1:11:24
there, you can just look and verify whether, okay, yeah. When I selected these three points in the UI, I could see those three points before I labeled them. I don't know if those exact points are like, for example, so the generated code might show you, hey, you brush from this range. So the generated code says has like a conditional from less than three to greater than five or whatever. Yeah, yeah. Like, would that help with the trusting of this generation?

Participant
Absolutely.

Okay. Yep. So yeah. And then like, yeah, so that basically brings me to my final question, which is like, very broad and open ended, like, basically. And we have probably covered some of this before, but any thing you like or dislike about this approach, particularly?

Participant 1:12:15
Oh, I think it's great. I would love love to see how in the world you would implement this in Tableau. It would be crazy to kind of see how you would do that. Because it's an it's Tableau is, is closed,

Author 1:12:33
they have like, Tableau does have a public API, really very VCAT, like very have access to things like you know, so basically, what our approach does is our approach is very agnostic. So our main contribution in this paper is we created a set of, like environment agnostic representation for all the interactions. So when you represent a selection, we created a

representation that allows us to express selections. Yeah, they got rectangular selection of point selection, or an algorithmic selection, independent of where it came from. Right? Like, it doesn't matter how Tableau implements a brush, or brushes, a brush, it will always have left, right, top and bottom. Yeah. So we created a representation to capture those. And I think like the tableau public API should allow us enough access to it to actually try and implement this.

Participant 1:13:28
Yeah, I would love to see that. I, you know, there's so much here, I don't know what else I can. What else I could essentially, say, Yeah, this would be great. I think you've done outstanding work, you know, I think this is really great. The I kind of like, you know, this Python coding and see the hierarchy of things, you know, which is nice. And maybe that's where, under, you know, again, this is Python, though, would be great to be able to click on the Select 53 points, and then see the all coding, you know, to that would do the 53 points, right type of thing. Yep.

Author 1:14:23
Yeah, that's actually that's actually really good. At like that is. So this is the part like where, you know, like, once we bring this in Python, like our core contribution is the fact that we can do this. Now our next challenge is where, what is the best way to show all this information to a Python user? Like when they load a workflow, right? How do we show them what the workflow does? What are the details? And yeah, this is like a really useful feedback that yeah, maybe some way to interactively see for exactly

Participant 1:14:55
that, you know, you click that and it pops up the code As you were doing earlier, that's where the, you know, you can change your JSON string to be the Python string, right? Or K means cluster, type it right? Yep. At that point, they could see oh, yeah, that's one two or you know, point two or whatever. Right, right. Yep. Oh, that would be nice.

Author 1:15:20
Yeah, that's, that's Yeah, that's actually really useful. That's probably something we should pick. We'll explore doing

Participant 1:15:29
though I think it's great. I think it's very cool.

Author 1:15:31
Yeah. Thank you. So okay, I just stopped my screenshare Yeah. So unless you have any more specific feedback, I can stop the recording.

Participant 1:15:42
Yeah, please.

P2

Author 0:00
Okay. So what I would like to do today is have a brief chat first about your data analysis process, and what sort of tools you use, etc, then I would do a short presentation on the techniques we have been developing, what do we mean by workflow? How do we capture it and how we can use it, and then give you a short demo, we have been made for like, just to show this technique can be implemented in some way. And that we can have a chat about like, what do you think of the technique, how it can be applied to your work? And what challenges might be there when we adopted to your work? So to start with, like, just like a very broad question of like, can you just talk about your data analysis process in general from like, when you get the data? How, what sort of tools to use for analyzing what you do with it?

Participant 0:58
Yeah, sure, I'm trying to think it will probably be best to kind of do an example. Right, like, okay, so recently acquired a data set that has only its pilot data. So it's not a very big data set. But it came from a collaboration we have with <*medical institute*>. And we got pilot funding to analyze hair samples from suicide, yes, for different analytes. So we're analyzing hair for trace elements, the presence of up to 12-13, trace elements, and then also cortisone. Cord, no, I'm sorry, cortisol, sorry, cortisol, and nicotine. So now I have that data back. So for each individual, they're represented, you know, multiple times for each of these analytes. And so I got that back. And, you know, first step is to just look at the distribution, you know, to load the data into SAS, you know, so I'll convert it to, I think it came in an Excel file, I'll convert it to a CSV file, I will import it into SAS, and then I'll do some basic, descriptive look at the data and plot it, right. And so this will really take, you know, really at first because this data, which, like a lot of my data, is interval data, so follows an interval, you know, is, is that type of biological data. So I can do it, you know, plot it using, you know, histogram, so I'll probably say start there, look at the different histograms for each of the analytes. And just get a sense that they're normally distributed if they're not normally distributed. So that's usually what I'm thinking of at first, maybe then I will then do some descriptive statistics overall, consider the you know, ask for the mean, the minimum maximum, the median, the interquartile range, this is all right, really easy in SAS, because it is via Proc Univariate type of record, right? And remind me, are you a SAS user?

Author 3:37
No, I actually don't use SAS a lot. But like, I'm comfortable with, like Python and R. So but yeah, like, so I have, like, a general idea of like, what the SAS UI looks like, but not super into, like, what sort of features it has.
But you can relate, if I say, this procedure, then you know, it'd be some are, like, function or something.

Participant 4:03
Yep, exactly, exactly. That just runs it off. And then maybe I'll look, you know, organize the data by you know, I looked at samples from males and females. So then I'll look separately. Well, what's the descriptive statistics like for males versus females? Is anything starting to pop out?

This is really agnostic, because it's pilot data, right? I have no hypotheses about it. So that's really probably just thinking what I do there to begin with the dataset where I didn't have, you know, specific expectations. But probably Overall, I'd like to make sure if it were a data set, we're right, we're actually testing hypotheses, I need to make sure the data is coded the way we want it to be coded and meets the assumptions with we think it should meet so in another data set where you need another analysis, we're using natality files, which are birth files nationwide from a bunch of different states in the country. That's much bigger data set, right. You know, if it's 12 states 16 states and it's their entire birth cohorts for, you know, a big part of their population. So they're pretty big. And you we have to be really sensitive to the coding and make sure we pulled out what we want. We coded it correctly, it fits if it's categorical that we see the categories we expect, though, you know, it's a lot of first data. Right? Yeah. And then we'll get once we know the data, so when we want it, then we'll get into the model fit.

Author 5:24
Okay. So, okay, this is very interesting. Could you just talk about like, what sort of tool like so the SAS, you basically have to write certain scripts in SAS? Or is it more drag and drop to create charts?

Participant 5:40
You know, SAS has a JMP it has, it's called JMP, which is more a GUI interface, drag and drop and but no, SAS itself is a programmable.

Author 5:51
Oh, so that's interesting. So do you use this GUI interface? Or not at all? Yeah, well, okay. So I have like two questions, because it's interesting that you use SAS as your means of, like, analyzing the data. So like two questions like, Is there any reason you prefer scripting within SAS, when you have a GUI? And the second question is, why SAS compared to something like Python? Or R, which is? Well, I don't know, like, in general data science domain, they are popular, but maybe in your domain, SAS is the most popular thing.

Participant 6:29
The first question is reproducibility. So you know, I have a master's degree in statistics. So my statistical training, we always use SAS in our and so that came, you know, it's just my comfort zone for statistics. And then now definitely, with the push to I never gotten comfortable, you know, with using any GUI interface GUI interface program, so, so that for me, like SPSS never used that really, I have colleagues who do, right, and friends, don't let friends use Excel for statistics that kind of, you know, that kind of thing. So, um, so yeah, I've always, you know, always use SAS or R, then my training in my master's degree, definitely heavily, like three quarters of the program was SAS, because coding in SAS, right, with a quarter of it used to are, so I just have a strong foundation in SAS. And then because I'm an epidemiologist, I think a lot of epidemiology is also like, my work with the CDC is very SAS, right, I guess they use SAS at the CDC we do, right? So it's kind of just been, but then, you know, my data, my bigger data, kind of data science colleagues, yes, they're more than our anything I do that spatially related or

geographically related is done in our do all my plotting, and are any figures that I send to publish I do an RMA export the data out of SAS, but then I plot in R. So

Author 8:01
that's really interesting. You mentioned like with CDC, like when you work with their, like, your colleagues there, they use SAS, you use SAS. So I would sharing your results or sharing your analysis kind of is easy in that case, because you're on the same platform. Now, do you have colleagues who you frequently collaborate with where they don't use SAS, and then when you have to share your analysis, they somehow have to replicate or you have to replicate that in, let's say, R? For them to be able to, like, reproduce the same results? Like does this happen?

Participant 8:35
is more of a challenge where it's come up is? Uh, no, I have, you know, a few grants that I've needed some help with support with. And when I find, you know, staff and team members to support my projects, they may come and say, No, I'm more of an R user than SAS. So then, therefore, either it's some combination of, well, I'll send you the SAS code, and I think, you know, you're, you know, sophisticated savvy enough, you can take the SAS code and convert it to R. Or you may just do it from the stage one and R that's, you know, I believe them some leeway. So that's more where it's been, or just a collaboration with other colleagues or faculty, who are more are programmers, we can talk the same language when it comes to the analysis, Oh, these are the kinds of models I formulated the output, um, but less being able to share the maybe,

Author 9:27
right? So you're basically like, when collaborating with like non SAS user share to kind of the conversation shift to using more statistical terms on like, SAS specific things, because then you have to, you have to be slightly more abstract.

Participant 9:44
Okay, yeah, exactly. Yeah, I will say this, you know, we're fitting we're doing this kind of analysis, we're doing a causal inference analysis where we're, you know, do estimating the propensity scores and then waiting with them and, you know, doing it stepwise, but just not you know, say, well, here, I'll shoot you my code. So you can Yeah, that kind of thing. Yep. Yeah, and I would say, you know, for me what I feel like, um, it hasn't been limiting yet. I don't use PI. I don't Python use Python at all, which is probably we may be limitation someday it hasn't been yet. But, um, that, you know, from a data management perspective and manipulation perspective, I think SAS is a little bit better than R. And I still do a lot, we still do a lot of that with R.

Author 10:27
Okay. So, okay, so this next question is slightly more in the abstract sense, where I want you to think of like, the kind of tasks, especially the initial data wrangling tasks where you know, you have to select some points and maybe code them or like, tag them in some way. Do you think doing them in SAS versus? Let's say you had a GUI interface to do it, where you could select the points manually from a chart? Do you think that would be easier? Or do you think like, in

SAS currently, the way your workflow is set up? That's way faster than if you had to do it in a GUI?

Participant 11:06
Probably fast, pretty fast, and SAS pretty straightforward, because I've got code already. Right? I've got a little bit of that pipeline already developed, you know, so yeah, yeah. So I can say no, now with SAS, what procedure do you use to find those points that may not convert conform to what I expect the data.

Author 11:25
right. So um, so do you have any like, Do you have a method for basically keeping a record or a track of your analysis, for example, like, you might have a script where you're doing something, but you decide, let me try this one other thing which deviates from this, like, how do you go about keeping track of things like this?

Participant 11:48
Yeah, so I try to be pretty good about my notation. And really notate the code and say what I'm doing when, you know, like, Oh, now I'll try this. And so what I'll often do is if I go and deviate, and I say, now let me try this, and I try something didn't work the way I want what I'll often do, and it makes the code a little messy, but also keeps everything I've done is, you know, SAS makes it really easy to just take a piece of code. And, um, you know, what's the word for me? inactivate it, you know, the codec? Well, I'm sorry, say that, again,

Author 12:20
like folding, like where it like, collapses, like, it doesn't run as part of regular,

Participant 12:25
right? I'm less than more, he makes it inactive. You know, I can, you know, bracket it and say, Yeah, I did it. But it no longer if I run the whole piece of code, it won't run that piece,

Author 12:39
right? Something like basically turning it ditto comments so that the SAS ignores it when it runs. Okay, so, okay, so this is, okay, so doing this, like, when you go back over your analysis, like, what are the challenges of this approach?

Participant 12:59
Sometimes the challenges of this approach is I don't have a clean sometimes, I'm not always great about doing that. So I may have a piece of code that stays active. And right, I don't comment it out. But I don't really want the output like I finalize the analysis, finally got the data, maybe I found something wrong, go back to clean the data, okay, I want to rerun it again, export, you know, results, the output. And I, if I ran the whole entire script, then I'm going to end up with output that I don't want, right? And so instead, you know, a SAS, like, you know, our, instead I'll often do, which is probably inefficient, if I haven't gone back to clean it up, and make

it exactly final, I'll just run parts at a time, right? You know, because I'll say, oh, yeah, this part I want. Now, this part, I don't want this part I want and so right, that's the

Author 13:49
business like so. So doing, like running parts of the code in this way? Does it kind of make it hard to know, if you're run all the sections correctly? whether, you know, like, did you like maybe you wrote a code to do some transformation? Now, when you run parts of it? Is there a challenge that hey, did I do did I run this part, which transformed the code or the digestion rate output without? Like, is that a challenge? Sometimes?

Participant 14:19
Yes, definitely. Definitely. You could skip the step. Maybe, actually, you know, I do a lot of like, you know, updating the datasets at each step, but like, often at each step by so if I skip a DATA step step, will give me an error, right? Yes, certainly, I could end up running some code where I did skip a step. And it's doing something I didn't want to or doing something or you know, the timing is off or something. So yeah,

Author 14:47
yep. Yeah, that makes sense. Okay, so the next question is about like parts of your analysis, which repeat like other parts where you know, like, every time you Get a type of like, some data for, let's say, a certain type of data set. These are the things you have to do every time before you start your actual analysis. And they might not be related to it's mostly like, you can think of things like data cleaning, or just, you know, sanity checks to make sure are these things for din Rangers? Or, you know?

Participant 15:22
Yes, yes, definitely. Those are typically by they may not be exactly the same every time. But certainly, yeah, they're I think I mean, I see. Right, I think this is a lesson learned early on by statisticians just should be that really, right. You have to know what your data looks like, before you apply, do anything else with it? And you know, I'm sure there was a stage where I wasn't very good at that, and learn that lesson. Or whatever. But I've seen other junior people do that. Right. Right. Just apply an analysis and, and like, within analysis able to do look, I mean, let's and you know, they didn't they didn't do those checks. So yeah, definitely. I'm doing spending have those.

Author 16:01
Right?

Participant 16:03
Stages. Yeah.

Author 16:04
So okay. So do you see some, do you see any challenges in automating this part of the analysis like this recurring parts, you know, especially the initial ones? Like, have you automated? Or do

you see any challenges in these parts? You know, something like, I'll run the script, if it doesn't throw an error, I can assume everything is correct. You think any challenges and doing that? Or maybe you already have done it somewhere?

Participant 16:34
Well, I think my challenge is I don't have you know, I have scripts, and have a lot of them. And I don't have any more polished anything. That's all that polished, you know. So that's probably the biggest challenges and probably an efficient that way. Because, you know, I think it's so great. You know, I work I have a team member who's with the study design and biostatistics center that she mostly works on my data sets. And you know, they're really working on trying to have these good data kind of hygiene, macros that that do all of this, so that it's efficient.

Author 17:11
I like that term, like data hygiene is a good term. And okay, so, one last questions before we like move on to a presentation is, do you switch between different tools frequently, like me, like, not, when you collaborate, essentially, but like, when you work on your own? Like, do you have to switch between different tools? And that it's like pen and paper? Right?

Participant 17:38
Right, right, I'd say the biggest tools I switched from is probably very raw data prep in Excel that gets converted to see the CSV, right, like, so I'll even do if I don't want it, sometimes it is easier to just change a variable name in Excel, then rather than SAS, depending on where we're at, you know, so there's that, and then, you know, then switching the SAS, then like I said, I do all my figures in R. So then taking the output from SAS and importing into R, and where I have a whole different use, if I want to use GG, two for my figures, they do a lot of forest plot type figures. I have that coop script already. And then you know,

Author 18:22
so Okay, so this, like, when when when you put it like this, it almost seems like the Excel when you use Excel to do this, Excel is kind of like the GUI part, right? Because you get to interact with the data directly versus writing the code, which changes the data. So in that Excel is like a GUI. So in terms of that, like what are like Do you have some challenges when you shift between these tools? Or like, are there certain things for example, you know, you find something interesting in your Excel analysis, and you want to go to SAS, but now you have to do the same thing in SAS to start.

Participant 18:59
Well, this comes up and it's less finding something interesting. It's just a formatting thing. Most of my issues are the for formatting, okay, in Excel, because I really don't look at the data in Excel. It's more Oh, do I think this column is a numeric, but actually, it's a character column? Because there is unexpected entry somewhere, but it's 20. You know, it's 100,000 rows. And so I didn't I don't know, you know, so that's the kind of thing and then I go to import it into SAS and oh, wait, why is it a character about you know, field and so then I have to go a good maybe go back to Excel because it didn't import the way I wanted it to. Because and then it gets especially

dates are especially tricky, right? Going from Excel to SAS. Yep. Yep. That's the kind that's where I run into. Yeah.

<demo>

Participant 37:29
Um, no, it was, I think I followed it. Yeah. Um, and I could see sort of them just thinking about like SAS, and how, you know, sort of do some of those similar steps that you showed in Python in, in SAS. And I know, you, you know, are using lino sort of the scatterplot as the sort of start approach. Do you see here, turn an application to sort of a different, you know, kind of pipeline as well. I'm just curious.

Author 38:10
Yes. Yes. So the basically the under like, techniques, so this prototype is not, this is prototype is definitely not like, good, complete data analysis tool, right. So, but the underlying technique, which is capturing this interactions, it essentially allows us to like, basically have pretty much anything you want, like any kind of plot you want, maybe you want to start off with like a series of histograms instead of like a scatter plot. But yes, because essentially, what the technique does is when when I say add a scatter plot, it essentially represents that I want to focus on these two features, like a scatterplot represents two features of the dataset. So instead of scatter plot of your histogram, that means you are now focusing on one feature and looking at this distribution. Yeah, so yeah, like, Sure. So you can think of this as, for example, this, let's say this was somehow integrated in the SAS GUI interface where whenever you do something in the GUI, it remembers what you did, and then allows you to use that in the code automatically. Mm hmm. Yeah. Yeah. So yeah, so yeah, so it is possible. So like, so you're saying something at that time.

Participant 39:27
Now, I'm also looking at I see, you know, here you've got you we really focused on cluster piece, but I see you've got other, you know, down at the bottom, you've got polynomial regression and linear regression not to get you off topic, but yep, these just different. Yeah,

Author 39:41
so these are different patterns like so every time you do a selection. The system tries to compute a lot of different patterns which can best represent your selections, but it tries clusters. It tries because computer doesn't know right? It tries to guess me, trying liberation, maybe you're trying to follow Don't be regression, it tries a lot of those. And then here like, our selection doesn't take definitely doesn't fit a polynomial regression. So it ranked that as a lower choice. And it ranked the cluster as a higher choice, right? Right. Under regression, like, especially polynomial regressions are really hard to, you know, detect, like linear regression sustain, you know, easy, but with polynomial regression, it becomes tricky. Yeah.

Participant 40:29

So what algorithm do you have for identifying the clusters themselves? What are you using behind the?

Author 40:34
Yes, so we use, like multiple algorithms. So we have some isolation, flourish, clustering, and k means clustering as two of our primary algorithm. But what we do is, for each of those, we run a multiple parameters of them. So, we try clusters from like two to 20, and then try to match which one of those best fits the selection you made. And then try to rank it. So, you So, is there, there is a, we definitely try a limited number. So, if you load a data set with 50 clusters, probably the results won't be as good. But that's a limitation of this implementation. Right? You could always have a better implementation. Right? Yeah. Yeah, need. Okay.

So I would like to, like go through pieces of this tool are basically more like, I want us to think about sub, we will talk like, I'll point out a particular part of the tool. And then I want us to think about basically, how useful that part would be a flick when it was if it was basically part of your workflow, if you got to use this feature somehow, how helpful would it be? Or would it not apply to your current work? or would there be any challenges?

So like, the first part I want to talk about is this history part rights, and we capture this history. So anything you do in the GUI, basically, you get to keep track of it. If you go back and do something else, it creates a branch, because that's similar to like how our thought process works, right? Like, we want to try two different things, and then maybe compare them the so what are your thoughts on like, way, like this approach of keeping track of your work?

Participant 42:24
Yeah, I like it. I think it's one of these things that I'm looking at it and I understand the logic behind it. And it would be familiarizing myself with how to navigate through, right, like understanding which parts of the branches apply to a previous step. And the changes that were done at each step right. And not just that takes some time to just, you know, get comfortable with the tool, right. So,

Author 42:49
right, yeah. Yeah. So So for example, here, like, basically. So like this, like this particular analysis session like this, there are two aspects to this, right. First is basically when you don't switch between data sets, but just have two different analysis, or, you know, like, here you are brushing. And like in this particular branch, you're filtering out things. But maybe in this branch, you're like, labeling things. And maybe you have four or five different branches, where you add different labels and how you want to compare which categorization or which basically, you can use think of this as a way to do different types of coatings for the data. But you do the coatings visually rather than the code.

Participant 43:33
Right, right. Sure. So should I see the different brush types that you can choose from?

Author 43:39
Yeah, so this rectangle, this is like a free form brush, where you can just find urine, select, and then you have different sizes, if you just want to quickly select some things.

Participant 43:52
See, cool. And what if you're doing the brushing, you have control over color choices?

Author 44:00
When brushing, no bar, for example. Okay, so let me first ask this question. Why do you want color choices?

Participant 44:10
I'm just thinking and I don't know if I were but if maybe and in addition to a label if they're, you know, some people are very stepwise, very particular, and colors could reflect thought process and chose steps made? I don't know, necessarily that way. But I think

Author 44:31
you Yeah, well, I mean, it is like adding a way to basically like adding a color choice is something trivial, right? So that you can definitely do that. And like if you add labels here, the labels themselves are color coded. So if we add multiple likes to okay, they all get like a different color. So you can have this multiple colors and represent multiple labels. But one other read too. basically capture thought processes at any point in this graph, in this history graph you get added. I don't know if this comes as part of this. Yeah, it you can basically add an annotation here, which allows you to document your thought process right here. Okay, nice. Yeah, like, interesting or something like this, like, you can read this. I just Yeah. It takes time. But yeah, it allows you to annotate your thoughts while industry. So this tree, or this, like history doesn't just represents what you did, but potentially also why you did it. And you can so do we have this as part of the tool itself, where you if you share this URL with someone, and they let's open it or their computer, they get exactly the same things you did? Today, like anything, if you added like an annotation here, and you gave this link to someone, they can see what you did stepwise as well as the annotations of why you did it.

Participant 46:04
Yeah, nice. Yeah. That's, that's a good idea. Yeah.

Author 46:08
So that does that. So like, yeah. So in general, if you see something like this, as, you know, beneficial for your analysis process and sub point.

Participant 46:19
Yeah, I could see that how it certainly could be, I think it would take a little, just like anything else using when you're using a new tool, right, a little bit of thought, transition terms of process, right? So that's where right like, and I go, I mean, it's just like, recognizing where I am in

inefficient, and no way I am at times, and not doing anything to rectify. So So yeah, right. It's one of those things where, yeah,

Author 46:51
Okay, so the next question is more about the predictions the system makes to help you with your selections initially, rather than this providence? So how do you what are your thoughts on system predicting some patterns to help you? And do you see any challenges with this approach? Or why you would be hesitant to use this approach to let's say, improve your selections? For example? Like, if you selected certain points here, and computer made this prediction? What are your thoughts on it? Like? Would you like to have those as part of the system? Or would you be hesitant in using them for some reason?

Participant 47:37
Well, I think, you know, that's why it kind of asking about what was they were what algorithms it was using, right? Because I wanted to understand where what was happening in the background. So Right. So I think part of it would be knowing some kind of, you know, any, any prediction is kind of like the validation piece of prediction. Right? So I think that's the, that's where it would come in trying to know make sure I understood and that whatever prediction was happening was a valid prediction. Right?

Author 48:07
So for this, so for right now, we have a very basic way to tell you what the algorithm was, like, what algorithm we used, what are the parameters for it, like, we are still trying to think of a good way to show this information where it's like, useful, right now. It's just like, we just show a JSON dump, like, this is everything. But yeah, like it definitely we can show this information about a desus algorithm. Like for this prediction, this was the algorithm. These were the parameters. This is the, let's say, the clustering parameter we used. So yeah, so will having this information make it more likely for you to use this predictions?

Participant 48:53
I think so. cuz otherwise, like, I hesitate, you know, um, yeah, yeah. Just cuz I being a statistician, right, like, yeah, yeah. Exactly. I don't tend to use the you know,

Author 49:10
and that makes perfect sense. Right, like, so, like, here, like, so. Like, even for this, like, the goal here is to, you know, for this is a very, how do I say, this is like a very generic prototype to show this techniques. But what I would assume is, if we wanted to implement this techniques for you, like, like, this would be the interface, right? Like, we would create a custom interface based on your tasks, the plots you need, and also the patterns you have in your data commonly, like we weren't, we would use the prediction system to, you know, kind of the data you have, versus like this right now these algorithms take they train on a generic data set. They don't and that's why our examples here are simple, but we will definitely like so The actual implementation for particular use case would have all the customization cygnets for the use case.

Participant 50:06
Hmm, yeah, <unclear>. And then you're kind of good because I'm thinking of even where we are doing, you know, we're doing a bunch of classification work right now, in our suicide work. And so how much of this within like, just maybe simplify later? What else we're doing give us a sense a priori of what how the data is maybe clustering on certain dimensions, right, before we're actually applying our, you know, actual machine learning algorithms.

Author 50:34
Like that's, like, this is like one of the lead sources this, like, our main contribution is not the tool, right? It's more of like, the way you can keep a track of what you did. Yeah. And that definitely, like, would really benefit from something like customizing it for a particular use case. Because creating a tool like this, which is generalizable for everything is, well, it's near impossible, right, like, yeah, there is no way one could do that. But yeah, so yeah, that's why I chose examples, specifically, which are very generic. Because that's, like, if I do something very specific, that would involve customizing, like tuning the parameters for that domain a lot. And yeah, so. So my next question is around, so Okay, so my next question is around this capturing of workflows, which are described, like using this interface. So a usual usually, we can think of workflows, like you can think of, for example, your SAS file, one SAS file can be your one pipeline, right? It can be a one workflow from start to end. And the way we usually start when we have like script based workflows, is we might have existing scripts we copy over, but the usual approach is we start with a blank slate, and we add one piece at a time, this piece will normalize our data, this piece will facet the data, this piece will, let's say, like, record our data. Right? So that is an approach to create workflow from scratch, the approach we use in this technique is you do a free form exploration using the GUI. And then using the history of your exploration, you curate it into a workflow. Essentially, you take this history you have, yeah, take the history, you. And then you add and remove parts of it to create a workflow. Hmm. This approach basically says, feel free to do whatever you want, like, take multiple branches, explore different things. And when you like, something you have, then create a workflow out of it. So So what do you think about this approach of creating a workflow versus like starting from scratch and building each block at a time?

Participant 53:03
Yeah, so yeah, it's just a different sort of, instead of being sent the text based, right, you've got the visual based. And so that's, that's neat. It's a different way of building your workflow. Yeah, just looking at this workflow makes me wonder, like, you know, applying, like you said, your, when you showed in your presentation, like, how, what would happen if I applied this workflow to a slightly different version of this dataset? Right, like, so yeah, seeing how it would play out that way. Right. Yeah, would be would be informative.

Author 53:35
Okay, and does this like? And does this approach kind of match your mental model of what you think about when you think of workflows? Or is this something is just not how you think of workflows? Usually?

Participant 53:50
I do. I think it's a specific type of workflow, right? Like, it's not Wednesday. It's a generic workflow, right. It's a specific workflow. So So yeah, I think so. Just with a Yeah, knowing that it's a type of workflow. Yeah.

Author 54:04
Right. And then the like, this is like a pretty open ended last question I have, is there any thing like basically anything you particularly like or dislike about capturing workflows in this way?

Participant 54:22
What I like is, I'm thinking about, you know, which data sets this may just naturally sort of work better with and others, I could definitely see I work with a big range of datasets, and some maybe when I don't see the applicability as readily as with other datasets like the one I once I said, where we want to do clustering, like then I could see you know, how we've made you want to apply this type of workflow from an early stage.

Author 54:52
And can you give me an example of fun uses some type of data sets you don't see this working with like, can you describe the quarter data set that could be add maybe what tasks you would want to do on it?

Participant 55:04
Yeah, I think I'm thinking of, you know, maybe like the natality files, I was telling you about that data where we have a lot of categorical data. And so I just, you know, maybe, I'm not sure, and it's very, very large, we just have to see how, how applying this type of workflow, from what I've see visually so far, and it just may, I may be missing, you know, something, you know, where the applicability would be to, to that type of data, where I need to mean, make sure, like, like I said, early on, when I'm interested, if, if there were clusters, that would be, you know, thinking that they're, you know, if you're thinking of a two by two table, you know, your clusters and 80% of the event are fall into one cell, right, like, um, are making sure that your, you know, unknowns are coded as nines, and not nine nines, and like making that kind of thing, make, you know, right away now that I talk it through, I could see how that could show up,

Author 56:06
I think, one of the, like, basically, one way to address that would be essentially, to think about our initial encoding, right? Maybe scatterplot is not the best way to represent that kind of data. Like, we might want to use something else. But yeah, but ya know, that that makes sense, like this is our demo here is restricted to scatter plot and like a parallel coordinate plot. It's, like, it's fair to say that this prototype definitely won't cut it.

Participant 56:37
Well, but I can also see now that I thought about it, like you would just plot maybe, you know, if you're doing, you know, you're thinking about it as a two by two table, or a three by three table, you know, you could still plot it this way. Right? Right, you know, and if you found that you

thought you had, you're supposed to have a three by three, and you have a five by three, then you know, there's data in there, that's not adhering to your schema. So

Author 56:59
know that that is something interesting. Like that is something we do want to think about. And then just, this is something actually forgot to get a get your feedback on is what do you think about the use of workflows to move between a GUI and a code based environment?

Participant 57:19
Yeah, no, I think that is, I mean, like you said, I do, like maybe it's not as explicitly gooey, when I do Data Prep and excels. But I already have some familiarity with that, right? And then SAS isn't that different, where you've got a piece of code where you're just in using running that piece of code, import your data from Excel, and instead it would be your Yeah, your GUI interface,

Author 57:43
right. Do you think like, there is applicability, applicability of this, for example, like one example would be, you have collaborators who are not familiar with SAS, but let's say they have a tool, a GUI tool where they can, you know, tag the data or code the data, and then save it as a workflow. And when you get the workflow, you just run it in SAS? To get Yeah.

Participant 58:08
Because I do have I work with clinical colleagues who are not the ones that are not comfortable. I know, they're not coding in SAS, but they're looking at the original data. So so in that way, yes, yeah.

Author 58:23
Okay, um, yeah, so I will just stop the screen share because yeah, like and I'll stop the recording.

P3

Author 0:02
Okay, so I would like to start with, can you just talk about your data analysis process? How do you usually get your data? What you do with it? What sort of tools to use?

Participant 0:16
Yeah, so right now I'm, uh, since we last talk, I kind of tuned your role. I'm a population health researcher. And so I used a few different types of data, but it's it's mostly retrospective, health related data. So it can be things like, one sort of dataset I use a lot is a large conglomeration of insurance claims. So anytime somebody has any kind of healthcare encounter, and they submit that to their insurance company, that gets compiled into the insurance company records, the date the person, some demographic information, what, like, what were they seen for, like, what procedures did they get? What diagnoses did they get? How much money did they spend? And that all kind of gets compiled. And so we then go and look for, you know, okay, we're interested in people who are maybe have a diagnosis of diabetes, and we want to see how many of them are getting this specific drug or something like that.

So the data is all most of the data used is something similar to that where it's records from somebody sort of healthcare encounters, right. In terms of tools I use, it all usually starts, I don't know exactly what you want to know. But it usually starts in like a SQL database, right? Okay, so it starts usually in a little bit of a messy format and an SQL database, and then either I or a member of our data team will, will run some queries to kind of clean it up and find the data that we're actually interested in analyzing. And from there, I typically take that sort of cleaned dataset, and put it into our to do any kind of analysis and causations. To begin with, I do some quality checks. So I make sure I look for weird things like, you know, somebody's getting a like, like we did a study looking at outcomes of pregnant women. And a lot of the pregnant women we found were four years old, which is some kind of problem, right? So looking for things like that. And making sure people aren't like duplicated. And just kind of doing some general quality control stuff. Then looking at sort of the primary variables that I'm interested in. So it could be age, it could be a certain disease characteristics. And just plotting those in some way, maybe a scatterplot, maybe histogram, just to get a sense of what those data look like and how they're distributed. And then deciding sort of a modeling approach from there, what types of models make the most sense for what the data look like?

Author 3:43
Okay. So you basically, so you're, so you, basically, you work with SQL queries, plus a combination of our scripts, right?

Participant 3:55
It's usually sequel aren't SAS?

Author 3:57

Okay. Do you have so like, just to just to be clear, like, Do you use any sort of BI tool or like, visuals like an interactive visual analysis tool like Tableau in your work? Okay, is there a specific reason for that, like use prefer like script and SQ SQL environments over something more visual.

Participant 4:23
So, I prefer again, are because it's very, very flexible, and it's very, I can control everything that's going on. It's not like a clicking point and click kind of built in County functions. And I can still do any kind of visualizations I want, right? I can right? Look at the data however I want. It just takes a little bit more coding.

Author 4:46
Yep. Sure. Do you think there are certain tasks like for example, selecting so for example, if we want to select a few points in our, we can like if the selection is Basically, union or intersection of different conditions, it's trivial to do. But when we want to select some complex pattern, do you think it would be easier to work with a visualization tool where you can visually select points using your mouse or some other interface? And then bring that to our to work on it?

Participant 5:24
Um, I would be very hesitant to select points like that. If it I mean, we probably need to look at a specific example to think through exactly. But that seems like we're going to be introducing a lot of bias by sort of select points like that.

Author 5:54
Can you? So for example, let's take an example. Right? We have, let's say we have a data set, which has a lot of crushers, we visualize the data set in a scatter plot, and we see certain clusters emerging, right? So wouldn't it be so would you say, like silica, would selecting these clusters from the scatterplot directly? Is more easier than versus let's say we wanted to select the same, like, we want to detect these clusters, using our like, we can visualize them in our but how, like, is there a way in our we can select those clusters directly?

Participant 6:35
You mean, select them from a visualization

Author 6:37
as well? Not from the visualization. But in our for example, we can visualize a plot and see that there are clusters? How would you basically select those clusters using our code?

Participant 6:50
So yeah, so I, you can visualize it to kind of look to see oh, yeah, my data do have these clusters, right? So I'm going to choose a an analytic approach to try to identify these clusters in a rigorous and repeatable way. Right? If I, if I look at these clusters, and I was like, Okay, here's a cluster here, the cluster, here's a cluster, I sent that data to somebody else, they might not pick, or they might not put the same thoughts in the same clusters, always because they're the

person and maybe even if I come back two weeks later, I might do it differently. Why? So if I'm, if I do see that there's some clustering, and I want to account for that, I'll use something like k means clustering, right? Some kind of approach that's rigorous and repeatable and is it doesn't depend on my bias,

Author 7:37
right? Yep. Okay. Yeah, that makes sense. Then, what I would like to do now is before, like, I have a few more questions, which, basically, we will talk more about? Well, okay, let's just go through this question. Before I do a demo. So you mentioned that you have your data sets, mostly your data set mostly come from like a SQL database, right? Are the do those data sets change over time?

Participant 8:10
Or will? Some do? Some don't?

Author 8:12
Okay? And for the datasets that change? Do, you basically have to rerun your data cleaning scripts or your analysis scripts every time they update? Or is there some automated way you do it?

Participant 8:29
So from the raw data. So so some of the a lot of these datasets,

the raw data will get updated maybe every other year. And when that happens, we update the database, we have brand new data, and we go through the same QC steps as we did before, which is the same scripts. After that, that data is locked, and it doesn't change. Okay. And so when I have a SQL script that then pulls data for a project I, I run that SQL script get get those data and then I usually don't go back and touch it unless, you know, halfway through the analysis, we realize, oh, we should have also, you know, pulled these data to like we've forgotten variable. Now go back and change the script. Every

Author 9:23
right. And then so for the for basically, for situations when this happens, do you have part? So for example, do you have to update, for example, when the data changes, right, like when the data updates the raw data or like data at any point updates? Do you have to update your data cleaning scripts to account for the variations or is the variation in the data pretty limited that your scriptures work every time?

Participant 9:52
For the most part, they should work? It sort of depends on what got updated so Like if we realize, oh, you know, we didn't account for whether this person had high blood pressure or something like that. So we go back, we pull those new data. And we have this new file. What I will usually do is, look to make sure that the old data and the new data match each other, except for this new like variable we've added. And then the data cleaning is all the same. And maybe,

you know, the modeling will change a little bit because we have a new variable, but data cleaning doesn't change. Okay.

<demo>

Author 36:09
So I would like to just get your general feedback on what do you think of an approach like this to basically bridge the gap, like, for example, using technique technique for like capturing workflows like this, and like, you can try to think of this as, like, for example, here, we implemented this technique in like a basic tool, like, this tool doesn't really have a lot of features. But like, for example, we could like the technique, in theory supports multiple different interactions. So it can be implemented as plugin of some professional data analysis tool or BI tool like Tableau. What do you think of this approach of using a Power BI or a visual analysis tool to start with your analysis or to do a part of your analysis and then taking it with you to a computational environment where you get more flexibility?

Participant 37:08
It's really interesting. I, I wonder. One, one concern I would have is you brought up the fact that one of the reasons that this would be really helpful, is because if you're going to do something in a script environment, you need to know how to code you need to know Python Rs. But it sort of seems like this is meant to sort of visualize and and start to analyze your data in a visual way. But you still have to move it over to a scripting environment to kind of complete that analysis. Is that right?

Author 38:02
Well, so. So like, in the demo, I do it to show that this can be done. But for example, if this, instead of this prototype, we had something like Tableau, which supports a lot more analysis features, you could just keep using Tableau, and it will basically create a workflow for you. And you can share that analysis sessions with someone else. For example, it's one thing, the BI tools don't like BI tools like Tableau, the lag is they don't store your analysis session as a graph, right?

Participant 38:36
Yeah, I

Author 38:37
see. I see. So like, moving between environments is one feature. Yes. But the core feature is being able to capture the analysis in a meaningful way in the first place.

Participant 38:50
I see. No, I mean, I think that's if someone's going to be running analysis in something like Tableau. I think I think you're right, it's very important to have something like this, that's that's defined steps that are repeatable. That makes a lot of sense. And

Author 39:17
so does, so does this. So basically, it does, do does the like, does this way of capturing workflows, like the way we described you? So basically, the there are like, if you look at like the way we can create a workflow for something, there are the usual approaches, we start with like a blank slate, and we create a flowchart of a ledge to this first step, then a second step, the third step, versus our approach where you do an open ended exploration first, and then you read that exploration into a workflow. Do you think that matches your mental model of creating a workflow or Using the first one where you start from a blank slate and do like a flowchart and create a workflow out of it. No,

Participant 40:09
my work was definitely more exploratory because I, I always look at how the data look like and how they're distributed. And that sort of informs how I approached the analysis. And that's, that's, you know, there, there are some things that I could write out from a blank slate like, you know, okay, I'm going to check to make sure people aren't duplicated, I'm going to check to make sure there aren't obvious errors, like somebody with an age of 200 ad or something. But before sort of moving on to a model, it's definitely just like you describing it's much more exploratory. Okay, let's see, what does this look like, you know, with this, or what if we bought transform this one or this?

Author 40:57
Okay. And then, so, we can just go over some of the features of the basically, in this prototype, like we some parts of the technique, which, like, I just want your thoughts on how this can be adopted as part of your analysis process, or, like, this doesn't really fit in your analysis process? And why do you think that is? So for example, this capturing of the history or like what we call provenance of your analysis steps? How, like, is that something that can be adapted to the way you do your analysis? Would that be something useful to have? Or is there a way you already do it in some way?

Participant 41:45
I'd say I already do it. Because I sort of have this record of the script that I've written and kind of seen, okay, I've looked at it this way, this way. And usually, you know, if, if something is important to kind of, even if it's not going to end up being part of the analysis, but it's important to keep it around, I'll keep it in, in just not evaluated section of script.

And so in that way, it's sort of similar. I think in terms of the applicability of this tool.

I, personally, I would be a little bit hesitant to analyze my data, just through like a visual medium, right. And I think probably, for most people who do big data biostatistics that are sort of familiar with coding, I think they probably feel burned. I think a lot of journal reviewers with to think this is probably really powerful. For people who have other sort of, right now I work with a lot of clinicians, a lot of doctors, and they are interested in research, but they don't have much research background, right. And this would be something that I think would be really, really beneficial for them, because they are going to be, they're going to want to do a lot more kind of looking at the data and sort of touching the data last year, and it's going to be really important that they they have that log where they can come back, give me data and just sort of reflect on.

Author 43:42
So yeah, I guess so like this. Okay, let's talk about this in the context of like, so you have worked with conditions who, like, would probably benefit from using like a visual tool, right? So from that context, so what do you think about this feature of automatically suggesting some patterns that, like, so what do you think of this feature of where basically, based on your selection, the suggests some patterns that the user potentially might be trying to select, and then they can use one of these to, like, update their selection?

Participant 44:20
Yeah, I think that's super smart. And I think that'd be really helpful for a lot of people cuz I know I work with people who they, you know, they know like a little bit of stats or something like that. And they'll try and create a dataset and like, it all just goes crazy. And then they try and come back and do it. It goes like crazy again, and like, they come back and like a whole new dataset, but like, it's just they did something weird in their code. So something like this, I think would be really helpful for them. And and especially where, like you're describing where, you know, they're saying, Here's what I think is a cluster and in the program is saying, Okay, looks like this is what you're trying to define. Is that correct? That would be that's really helpful.

Author 45:03
And then, like, another part of the SEC similar to, for example, do you think if the clinicians, they could create, like, they could take their analysis? They they do this analysis, right? And I'm assuming they share some of the results with you at some point. So basically, how would what would Dec Do you see the ability to create workflows like for to create like this workflows out of their analysis session, and then being able to share this workflows? Would you like aid in your communication? Compared to like what you currently do?

Participant 45:42
Yeah, I think that would be helpful. Because especially if these the workflows came with, like metrics, like if, you know, they chose a cluster, and then, you know, I've just thinking about this. And yeah, except clustering, since that's what we're looking at here. But if they chose a cluster, and the program, you know, recorded that they did that brush recorded the within cluster sum of squares, and then you had the estimating cluster and kind of gave the same thing, right, if they were able to show me their workflow, and I was able to go through and see, right, or like, analytic standpoint, how it progressed, I think that would be really helpful. Okay. No, I don't think they're gonna care about that. You know, I don't think all right. Yeah. But thinking about it sort of, in a robust analytic sense. I think being those numbers and report that to reviewers is going to be important.

Author 46:40
Right? Yeah. I also think of the question the same way, like, one of the aspects, which I actually did include in the presentation is the, this is also a means of communication, right? Where you

create something repeatable, and then use that to communicate the results where your recipient can independently go through the steps and verify if it works, as you said, it does. Yeah, yeah. Absolutely. Yes. And then like, basically, like the last question of anything you particularly like, like or dislike about this approach, or you think something that doesn't make sense in context of the way you do your analysis?

Participant 47:27
Um, I don't think much that that doesn't make sense, but I definitely liked the way it branches. I think that's a super cool aspect of it. And then being able to kind of settle on one sort of branch analysis I'd be able to explore like, that is really powerful.

Author 47:45
Okay. Okay. Um, then yeah, so, like, if unless you like have any other thoughts, I will stop the recording.

Participant 47:57
Now, I think that sounds great. Okay,

Author 47:59
we just stopped recording

P4

Author 0:01
Okay, so I would like to start by just Could you talk about your usual data analysis process? Start with your analysis.

Participant 0:12
So the software part, I mostly use Matlab, I hardly use Python. And I have been doing mostly image processing. Without any machine learning or anything, it's like basic image processing that I've been doing. And once I get data from, like, segmentation and tracking it mostly, I don't know, if you remember this I worked on, I do it. So I worked on analyzing the growth and motility of cells as a whole, or like inside cells, the components inside the cells. So I use image processing to actually do a segmentation and tracking of the cells, or do an image registration for computing the velocities inside cells. So it kind of depends based on whichever project I was looking at. And then I used after I have those measurements, I use them to compute trends of like, how much characteristic motility that particular cell has, or what is the deformation I see inside the cells based on local velocities. Kind of like take a bunch of huge amount of data and like make sense out of it. That's, like, the main pipeline of how I get things done in my project.

Author 1:36
Right. And you mentioned MATLAB, so the other any other tools like are there any, like specialized domain specific tools that you have you use or like some general tools, or you just use Matlab,

Participant 1:50
I basically created all my code from scratch. I used a little bit of Image Processing Toolbox. But I didn't use it like manually as a toolbox, I mostly use the functions that come with the toolbox.

Author 2:06
Right? Can you talk about like, what is this image procedure? Is it like MATLAB or Python library? Or is it like a tool you can use?

Participant 2:15
So there are some things like, you can do a Sobel filter with MATLAB, right? Like it's just a simple function they have in like return inbuilt function they are returned, which can do a Sobel filter on your image. So using that, I can figure out the borders of my cells, and then use a filling method to fill that portion which is indicated by the Sobel filter. And then I can use VW label, it's like a function, which can actually label the different sections that you can figure out inside an image like small segmented portions, so making a label in that, or even watershed so that you can like, figure out the center of your segmented portions and divide. If there is like two cells sticking to each other, you can like separate them using a watershed. So simple inbuilt functions like that, mostly.

Author 3:13

Okay. And so, so you mostly so like this dislike leads me to believe you mostly work with something like a scripting environment where you write or run MATLAB scripts to analyze your image data, right? Yes, that's correct. So, so this might be so because I don't know a lot about your domain. So is this like a usual approach? Or is this? Or does your domain have some like tools, which you can use, but you still prefer using scripts? Because that's how that's what you're comfortable with?

Participant 3:50
Yeah, actually, there. I don't know that there are a lot of tools are available for microscopic images as such, but we use a computation method called Q pi, which is not very widely used. It's not a very well established microscopy, Legacy computation, microscopic technique. And most of the labs who use q pi actually do their scripts from scratch. So there is not a lot of standard tools available, which is like convenient to use. So we mostly, that's the reason why we use our own scripts.

Author 4:22
Right? And so they do so do the tool should do the tools which are present, but you don't use like do they have some mechanism where you can expand them by adding your methods to it or like they don't allow, like the it's not possible to like basically add your techniques to the existing tools so that you can use the tools instead.

Participant 4:50
I think, slick, we have kind of tried out some tools from stuff like image check. And what we realized We need to do a lot of modification on our data to actually fit to that role. Right? Or we should be able to actually modify that tool for us to like, get a copy working. So yeah, we just I guess we were just lazy to actually go in and do that.

Author 5:16
No, that makes perfect sense, right? Like, if you have to spend time doing something, might as well, you already have the script, so just use those. So that that makes sense. Also,

Participant 5:27
I think we feel comfortable actually, knowing every element of what we're actually doing inside the tool, right. That also made the a lot of sense for us.

Author 5:36
Right. So basically, like, working with a script allows you more fine grained control. Yeah, right. Okay. Um, so, during your analysis session, when you run your analysis, is there a way you record your analysis session in some way? Like, maybe keeping lab notes or maybe using some programmatic method of these are the five steps I did and this is the result or maybe something like that.

Participant 6:09

We try to actually make the whole data analysis pipeline into like, one continuous flow, right. But it was collected

Author 6:19
by one continuous flow, you be like one file, like what MATLAB script to do everything related to this data?

Author 8:24
right. Okay, so the topic back to the year approach. So you have one side to track like, basically, you equate like one MATLAB file, like just to simplify, you equate one meta file with like one analysis pipeline, right? So in this approach, that say you feel the need that you want to keep this current script, but make a small change and try something new. Do you have to like create a new file? Or do you just edit the same file? And then how do you go back? How do you keep a track of these, like, branching approaches?

Participant 9:03
Yeah, so we actually, we have this weird method of saving a copy. Like, if there is one specific project for which we have developed one script, then we keep a copy of that, along with all the data that comes in that script. At least I do that. And then if I have to try something else on the same script, which will work with a different data set, I make a copy and like, into that data set, like that folder of the data, and start modifying on that.

Author 9:31
Right. And do you have to revisit this previous copies frequently? And are there any challenges you faced with this approach? Or, like your, you hope like something better comes along like a better way to do this? Or are you happy with this approach?

Participant 9:49
Yeah, actually, we have had trouble with that, especially because our code has a lot of parameters usually. And we have to adjust those parameters. So we keep going back into the old wish to see if we have changed the parameters too much? What is like the optimum number we can like a combination of number we can actually give to get something done. And I think a one of that was one of the reasons why there has been a lot of talk on machine learning. Because machine learning, we believe me, might don't have a lot of expertise and expertise on machine learning, but think that might solve the problem.

Author 10:26
Okay, so like the problem, the challenge is basically going back to your previous things and finding, which was the, like, what parameters you used earlier, were the best ones.

Participant 10:38
Yeah. Okay, because sometimes it does work for a specific data set, right? And then you went back and like, check for us data set, which looks very similar and it doesn't work, then actually,

you feel like something has changed from like, the time you actually took the same copy of that script and change happens.

Author 10:56
Right, right. Actually, that brings me to my next question was basically, do you have this happen frequently, where your datasets are updated? Like the datasets are similar? But there is a slight update to it? Or do you have? So basically, okay, so I'll rephrase my questions, like, do you do your data sets update over time? Or do you get like completely new data sets, which are similar to old one, but are not related?

Participant 11:25
Yeah, we do get a lot of new data set, because we mostly work with aberrant cells, which are like cells, which sticks to the dish. And a lot of cells we work with looks very similar. But there is like some difference in how much features they have, or whatever the size of the cells are, or how much they move. But in like one, like when you just sequence, it looks exactly the same. But we always encounter differences in how they can be processed.

Author 11:57
Right? And so I think you mentioned this, but just to reiterate, do you have so when you get like a new dataset or an updated data set, you try to rerun your analysis scripts? Or? Yeah, how frequently do you have to update the scripts to account for the changes in this dataset?

Participant 12:20
So I have, like, in my five years of PhD, I think I have done like eight different projects with eight different kinds of cells. So I would say like, once or twice a year,

Author 12:34
okay, so, like, so? Yeah. And what about within the data set? So for example, within one project, let's see, you take cell reading, so you take like, image images for one particular cell? And then like, you have a similar like, basically, you have an observations captured for a different time for the same cell? Do you are scripts were in that scenario? Do you still have to make some minor edits?

Participant 13:02
So we try to keep the datasets similar, the thing thing is some, it also depends on small things like how much dense we are plating the cells, and it never comes out to be that perfect every time. So yeah, we always have some issues with keeping up and like, and also the cells actually can get older, when you store them. And then it makes it lesser, like growth and health wise, it will be weird. And that also makes a difference. It keeps waiting. Okay, it doesn't stay the same.

Author 13:37
Okay. Yeah. Okay, so that makes sense. Um, so my next question would be like, do you have certain parts of your analysis that, like, you have to keep doing again and again, between multiple projects, like common parts of the analysis that you do this every time?

Participant 13:57
Yeah, actually. One thing that I kept doing in common, like in a lot of collaborative projects was, this is just one of the things was tracked, like measuring the growth of the cells, which always comes down to segmenting the cells and tracking them. So my own projects didn't need that, because I was working on intracellular things like going into a cell and like doing computations. So it was a completely different kind of, like algorithms. Even the other projects that I worked by myself, it was like that, but most of the collaborators actually want to know the growth of the cells, they that's where they that's why they come to us for grad collaborations because we have that computation microscopy technique, right.

Author 14:42
Okay. And so, last question before I switch to the very start with the because before I start to show you the technique we are working on is Do you switch between different tools like between, let's say MATLAB By turn or some other tool frequently during your analysis sessions, or is it usually like once you start with MATLAB, like, you do everything in there.

Participant 15:10
So I personally didn't like.

Author 15:14
Just like, when I mean tools, like you can also include things like pen and paper observation Cinetic. Basically, anytime you switch between, from your current environment to like, let's say what pen and paper to do something.

Participant 15:28
Yeah, so I have personally only, like, Software wise, I only use Matlab. I mostly stick with that. I tried to keep everything in my on my end uniform. And, yeah, I definitely use pen and paper. A lot of times, like I keep doing my computations on that sometimes, just to like, kind of see manually if I'm getting the same answers. My lab needs to they sometimes switch between Python and MATLAB. So sometimes they do the data, initial collection and processing through Python. And then they take that data and like do post processing in MATLAB. So they do keep switching things.

Author 16:10
Right. So okay, now I have two questions on this front one. First is like, why don't you do this? Like, are there challenges to doing this or is using MATLAB just easy. And the second is, do you know from your lab mates experience that if they have challenges actually switching between these things.

Participant 16:35
So I never used Python, because I always got stuff working on MATLAB, like I had, I built a different microscope. And I could get the speed I needed to buy just using MATLAB, so I just like stick, stick with it without switching to Python. But my lab mates, they built a microscope from

scratch, like using Arduino and small optical paths. And they figured out that if they using Arduinos, they could get way better speeds when they were doing the programming in Python. And it also evolved because one of the undergrads started using it. And it was just like, taken over by my lab mates from like the rudimentary stage. And they also saw that Python was actually nice to work with when it came to controlling the microscope and doing initial processing. So they stuck with Python. But the MATLAB scripting was already like kind of there, which I was already using. And they just stuck with that on the post processing part.

<demo>

Author 47:34
Yeah. Okay. Perfect. Then in general, so like, my first is like, is there a general feedback? Or what do you think about this technique of basically capturing workflow from one environment, using your analysis using your interactions and then applying it to some other environment like Python?

Participant 47:56
I think it does great. Because one of the things like I'm not like a heavy computation person, I'm more like, kind of in the middle of everything. So as you mentioned before, it's actually kind of more difficult for a person like me who's not exactly from a computing environment to kind of figure out what exactly is happening, even if it's like, building code from scratch. So it's always like, one of the things I really like about MATLAB has been always like, you can pull out things and like, actually visually see what is happening in the data? Like, it's a lot of time consumption, because you're actually, it's time consuming, because you're actually pulling out things and like, like, we are, like, analyzing it and figuring out how to do the image. Like how to actually see it, visualize it and all that. Right. Yeah, but definitely, like, actually be able to see what is going on actually definitely helps. So I think this is great in that aspect, especially for a person like me,

Author 49:01
right? So next, I want to know, basically, what are your thoughts on and be and if it is applicable to, if it might, in some way be adapted for your workflow, like your particular pipeline? So what do you think about this approach of keeping a record of all your analysis? For example, when you deviate from your analysis, create a new branch rather than, you know, basically, deleting everything and starting from scratch again, what do you think about this approach of keeping a track of everything you do? And can this be modified? Like for here, it's a UI tool. But do you think of a way that it can be modified to fit your workflow where you work mostly with scripts?

Participant 49:46
Yeah, I think I can see using this tool on my work, like especially my main projects, because it's like a ton of numbers and I have to figure out a trend inside them. So Yeah, like figuring out, like, I have a workflow and then I kind of deviate to see what else I can do. So having a tree like structure like this can definitely help and being able to go back into something I have been seeing before the other, I think that is definitely helpful.

Author 50:20
Right? And what about the approach we have here of using computer predictions to refine your initial selections? So basically, like, when you select something and computer says, hey, the things you're selected are part of a cluster, do you want to redefine them as a cluster selection? What do you think of this approach?

Participant 50:41
I think the separate will be creative, we are able to even add our own models to because we can like as a preliminary analysis, we can try out some of the things which are already provided. And then we have like, new things we want to try if we are able to add things into it, like my own models, I think that'd be a nice thing.

Author 51:03
Okay, you can let us know what you think can be added as, and you can like, basically, when you think about this, like feel free to reimagine, like, it doesn't have to be this tool, right? Essentially, any tool you use has this feature. Now what kind of models would you like to add? You can think of this as basically, like, when you try to segment like, let's say you have a tool, which allows you to visually segment things. And then when you load and particular image, the computer tries to help you by automatically segmenting some parts, and then you can say they use it segmented this incorrectly or you segmented this correctly.

Participant 51:42
You know, so, yeah, like, the different kinds of segmentation which already is available. As I mentioned before, our microscope technique has not been really established. So most of the time, we actually have to work with tools, which has like gradient images and all that, right. So most of the models come in with like, that kind of segmentation tools, that kind of models. And, like, sometimes we try to actually do a simple, more simplified model or write something, which actually fits our images, like specifically. So yeah,

Author 52:25
okay. Yeah, that makes

Participant 52:26
sense. Yeah, we do use the models, which are already, like, given for different microscopic methods. So in that way, I think you, like if I'm, like, really visualizing this whole tool, I would see like a library of models, which, right, it's the other microscope images, and then like, we can take our own and they kind of slightly modify one of the models you have already given and like, try it on our data or something.

Author 52:54
Yeah. Yeah. That that, that makes perfect sense. Yeah. So my next question is. So when you think of creating a workflow, so the you the way you create workflows, or your analysis pipeline is to usually start usually start by modifying one of your existing MATLAB scripts, right. But essentially, it is you're basically you start with a blank page, and then you add steps as MATLAB

code to it. Yeah, right. So that is one approach to create workflows. Whereas the other approach, which we see here in this tool is to basically you do your free from analysis, you have all these multiple branches. And then when you're happy with result of one of the branch, you turn it into a workflow. So essentially, you have this different branches, and then you explore different things. And then when you're happy with one of the things, you turn it into a workflow. So this is another approach of creating a workflow. which one fits like do you think this this approach is applicable to your work style? Or would your work style will be benefited in some way, by using this approach of using exploration first, and then selecting one of the paths you did? or prefer your existing way of like starting from scratch?

Participant 54:19
I definitely think it will be applicable because most of the time, we actually don't inherently change the method itself. We erase the methods or omit some steps in the method and like he, like basically, try a different workflow. So I definitely can see this to be helpful.

Author 54:38
Okay. And does Does this match your mental model of what you think of as a workflow?

Participant 54:45
Yeah, I think so. I can actually see my segmentation code and that workflow, right there. Yeah. Okay.

Author 54:52
Perfect. Yeah. So okay, so yeah, so this was good. And I just have like a last question. about like, anything you particularly like or dislike about this approach, like, of course, and I mean that in context of your work, right? If it is not applicable to your work, you can say so.

Participant 55:13
Yeah. I think I like how simple the tool is, and the advantages of the tool for sure. Like, it records the workflow, it actually helps you visualize what is going on with the data and all that, right. Yeah, when it comes to maybe this is not something like I'm really understanding or something. Or maybe it's just like, kind of brain being wired that way. Like, we always have this thing of being very, like, we have the trust issue, I feel like, we always want to see what is happening inside the glue that actually does the data analysis. So

Author 55:59
So basically, you like you would want transparency?

Participant 56:06
Yeah. So I guess your core does that on GitHub, right? Like,

Author 56:11
for example. Okay, so just let me give you an example. And just to confirm with this, so for example, when computer automatically selected a cluster, they predicted Hey, select this cluster

instead of europaische. So we have like a details about what the algorithm was used to detect this cluster here. So like, we use DB scan with these five parameters. Okay, these are the points. So do you mean information like this?

Participant 56:39
Yeah, I guess. Like, if you for instance, if you see a small building function in MATLAB, you can just like open that function and see okay, exactly what the return on the quarter.

Author 56:52
Okay, yeah.

Participant 56:53
Because sometimes we see this really weird things like, we do similar kind of computation and our data looks horrible the results. And these use a tool and it looks really great. So we want to see what is the difference that means

Author 57:08
okay, yeah, I understand. Yeah. So that is so that that that because basically, if the code for this, like tools like this, let's say someone created a tool similar to this, which has like tracking and workflow for like segmentation, if the source code is available to you to see why their segmentation works better, you would be

Participant 57:30
yes, exactly. Okay.

Author 57:32
Yeah, that that's, that's good. I'll just stop sharing the screen. Okay. But yeah, essentially, yeah. So, unless, like, unless you have any other feedback, I'll just stop the recording. Okay. Okay.