All right.

In the interest of time, let’s go ahead and get started.

Hey everybody,

thank you so much for coming today and this week seminar.

It’s my pleasure to introduce Stephen Larsson and Adria Haimann from Metacell.

This is a few words of context here.

We’ve talked about, we’ve had people, we started this semester with somebody from the hospital.

We’ve had people from academia, we’ve had people from pharmaceutical companies.

And so very excited to present something different.

So Metacell is a company that works in sort of the research space.

Near and dear to my heart.

They’ve been, from their beginning, I think, very active in the computational neuroscience community.

We both contributed to a project called NetPyNE for building models of computational neurons.

But more broadly, they work in the greater health informatics space.

And they’re going to tell us a little bit about how we can enhance biostatistics and health informatics research.
25 00:01:12.750 --> 00:01:15.960 through collaborative cloud-based data science tools.
26 00:01:15.960 --> 00:01:17.403 So let’s welcome them.
27 00:01:19.500 --> 00:01:21.870 <v ->Thank you very much. Good afternoon everyone.</v>
28 00:01:21.870 --> 00:01:23.700 I can see some of the back of your heads,
29 00:01:23.700 --> 00:01:25.650 so I can imagine that I’m also, you know,
30 00:01:25.650 --> 00:01:27.250 virtually looking at your faces.
31 00:01:28.200 --> 00:01:29.580 Thanks so much for having us.
32 00:01:29.580 --> 00:01:32.940 I’m Adria Haimann and I work alongside Stephen at MetaCell.
33 00:01:32.940 --> 00:01:35.370 And as already mentioned, today we’re gonna share with you
34 00:01:35.370 --> 00:01:39.060 some insights into how academics are using cloud-based
35 00:01:39.060 --> 00:01:41.610 collaboration tools to enhance their research.
36 00:01:41.610 --> 00:01:43.230 But before I kind of begin with this,
37 00:01:43.230 --> 00:01:45.240 I wanna provide you with some context.
38 00:01:45.240 --> 00:01:48.390 So, 10 years ago I was in your position,
39 00:01:48.390 --> 00:01:50.310 I was studying health economics
40 00:01:50.310 --> 00:01:51.690 at the London School of Economics,
41 00:01:51.690 --> 00:01:53.730 and I had joined a research team
42 00:01:53.730 --> 00:01:55.560 at the European Observatory for Health.
43 00:01:55.560 --> 00:01:57.210 And I was relatively new to this field
44 00:01:57.210 --> 00:01:59.730 and kind of found myself in a Catch 22
45 00:01:59.730 --> 00:02:01.590 that maybe you can relate to.
46 00:02:01.590 --> 00:02:04.920 So I wanted to know how can someone or a student or postdoc
47 00:02:04.920 --> 00:02:07.710 or researcher discover the best way to collabo-rate
48 00:02:07.710 --> 00:02:09.630 on their research and use new tools
49 00:02:09.630 --> 00:02:11.790 if you have fairly minimal experience,
50 00:02:11.790 --> 00:02:14.340 neither academia or in industry.
So that’s essentially what we want to show you today and what we’d love to share with you, if you could go to the next slide, which is kind of a collection of key topics of how researchers are doing just that, while also getting the most out of their data. So during this seminar, we’re gonna cover different methods that you can share data analysis and introduce you to a specific cloud-based collaboration platform that we’ve created called Cloud Workspaces. And then we’ll run you through some examples of how researchers are using this platform, as well as how we’ve formed an industry partnership. And then lastly, we wanna show you kind of other ways that this tool can be used in academic settings. And then of course, we’ll open it up to you guys and encourage you to ask us questions on any of these topics. So I’ll hand over to Stephen now. Thanks Adria for that great introduction.

I currently see you as tiny, tiny pixels on my screen because of the way this is viewed. So as much as I’d love to be there in person...
and looking into the whites of your eyes, I’m not gonna get that chance. But, I think we have a really good robust discussion. for you guys that I hope you’ll find very interesting. And thank you very much again to Robert for the invitation. So similar backstory on myself, I went through undergraduate training at MIT in computer science, did a master’s in AI before it was cool again, and then shipped off to UCSD for a PhD in neuroscience with a computational specialization. So very much familiar with the academic experience and I’m really excited to share with you some of the things that I’ve learned since leaving academia. And one of those things has been to start this company, MetaCell, which I basically started as I was wrapping up my PhD and I kind of realized that I wanted to serve science in a different way than was gonna be possible just within the confines of academia because I realized that I was a builder and to build software that could software tools that could be useful to, you know, tools that I would wanted to have had as myself, a graduate student.
I would need to kind of put a professional team of folks together that, you know, really came out of industry and that are kind of high hard to higher end academia.

So the story of this slide is, since then, all the different great groups that we've had a chance to work with, and you'll see a really kind of motley crew of logos that are present here from, you know, really, really big pharma companies like Yale, you guys are on here, other universities that we've had the chance to work with, and then biotech companies, med device companies that we work with some, some of the US lots internationally.

And realizing that, you know, the core thing that unifies all the work that we've been doing over time is the way that sort of math and computation can help us understand the life sciences. So hence I come to you today in a biostatistics seminar to talk about, you know, some of the other pieces of the puzzle that go into advancing the life sciences in that way.

So, let's start with a really simple, simple example, right?
So let’s say you’re doing some kind of analysis on some kind of bio data, okay?
Perhaps in the statistics context, you’re using SaaS.
In a computational neuroscience context, you may be using Python and the Python suite of tools.
Some in the statistics field are using R open source,
you know, statistics packages.
Whatever it is, you’ve got some data, you know,
maybe you’re analyzing it on behalf of yourself,
maybe you’re analyzing on behalf of your lab,
the group that you’re working with.
Maybe you’re analyzing it in terms of a company.
Whatever it is,
you wanna share that data analysis with somebody else.
You’re probably gonna have to gather some history of those commands together.
Maybe it’s packaged up as a script, maybe not.
You’re gonna send that file to somebody else very often.
And then you’re also gonna wanna somehow collect the outputs of that, right?
The figures, the diagrams, the summary statistics,
the result of T-tests, you know,
things like this, right?
And send that output somewhere, right?
So, you know, that is a problem time immemorial.

And you know, as long as I’ve been, you know, working in this space still, you know, it’s very common to just do this and it’s maybe send this over email, right?

It’s still a practice that I’m sure you know, happens.

And so, and that’s probably just fine, you know,
in many small circumstances.

But as that scales up, there’s problems of reproducibility,
there’s problems of, you know,
keeping track of who sent what.

Email is not a great file management system.

So we’ve been thinking a lot over the course of our company,
which is, we’ve been around now,
this is our 13th year about how, you know,
the cloud and the internet basically can come into that in any better way than sending email along.

And so we’ve thought a lot about, you know,
what starts to happen when there’s a computer that lives
in the cloud that multiple people can jump into and join.

And what is, you know, how does that work in general?

It’s something that we’re not only just us doing, right?

This is an idea that’s been there for a while.
Anybody familiar with, say Python Notebooks, right, are aware of this idea. There’s tools like Google Colab, and then we’ve even been talking to major universities, like we’ve been having a conversation with Harvard Medical School, where they’ve been working collaboration with Amazon to kind of work together with them to set up computers that are in the cloud. Similarly, of course, there’s gonna be what happens with, at like, at your local university with your local computing infrastructure. Typically that’s based around supercomputers that are there for doing like really powerful computations or calculations. Things that are very data intensive. A workspace in the cloud is sort of in between. So it’s kind of like, you know, just a laptop that isn’t your physical laptop, but it’s like a laptop that’s somewhere else in the cloud that you can log into and do some analysis with. And it basically lives as long as you wanna do that analysis and then it goes away if you don’t need that analysis anymore.
or it can stay there as long as your lab is around, right?

And then go away if you don’t need it anymore.

So the idea is then in this story, instead of just gathering the history of commands, sending the file and sending the output of the file, what if, right you could do all that in the context of a computer that multiple people can join and look at, right?

Work in that same environment.

When you log out, it’s exactly where you left it, right? Like if you know your computer gets misplaced or you drop it, you know, off a bridge into a river, doesn’t matter ‘cause all this stuff is preserved, right?

So, how does that idea start to change the basic practice of interacting with data and doing analysis like this if you were to change that one variable okay?

So that’s sort of the starting premise for our chat today. So, you know, what that might look like is, you know, a session one-on-one or two-on-one with multiple people where you get, you know, perhaps one of you in the future.
In the case that we’ve been doing in our company, one of our staff members, who has experience in doing a different kind of data analysis. In our case, we work on a variety of problems, but one of the major ones we worked on is like the imaging of calcium signals in neural tissue okay? But you know, you might be on a call like this one and just the same way that you might meet with your lab members on a Zoom call, you might meet with someone with experience in data analysis or biostatistics that is not in your lab or not in your even organization. It might be somewhere remote, maybe at another university or in a company like ours. But what they might get as the experience of that is jointly logging into this workspace that lives in the cloud. And if SaaS is the thing you wanna use, you might find a whole SaaS instance there in a desktop that you can log into. But the point being that multiple people now can type on it as opposed to like physically handing your laptop around in the lab or even just screen sharing it in some kind of a lab meeting, right?
It’s actually allowing for people to jump into the same application and literally like trade off like typing commands into it. Kind of like what you get with a Google Document or a Google Spreadsheet, right?

That real-time collaboration, but now for any kind of application. So that’s one experience you might have. Not just SaaS, right?

So a Jupyter Notebook, as I mentioned before, is another thing that you can use. And those of you who might be using, again, the more open source technologies, if you might be using R Statistics or using Python if you might be using R Statistics or using Python

or whatnot, you’d be familiar with, you know, a Jupyter Notebook. So it’s based around, you know, this idea of putting a computer in the cloud, multiple folks logging into it, and then being able to sort of transport your expertise around the world. Because in addition to the knowledge of doing analysis being shipped around,
data can also come into this workspace as an intermediate space that’s private to a given lab,
where it sort of stays under the control of the lab,
you know, whoever puts it there can take it back,
that kind of thing.
Okay so we’ve been exploring this model
and we’ve also been talking to other organizations
and universities about this model and how to use it,
how to implement it, right?
As I mentioned, we’ve been talking to folks like
at Harvard Medical School that partner with Amazon
to bring these sorts of instances into their labs and what can be done with it.
So I’m gonna wanna talk a little bit about like some of those details,
I’m saying it here in the context of our product,
but I’m not trying to sell you anything.
I’m really trying to talk about it
more in the context of what can be done.
So thinking about it, like,
so I mentioned SaaS as an example.
I mentioned Jupyter Notebooks as an example,
but there might be other kinds of software
that are more particular to a use case,
like MATLAB’s another one that could be installed.
But there might be even more specific software
that might need to be set up or run.
Sometimes, for example, survey software where you might collect data from a very particular kind of survey system and you need something to work with it. So imagine that, like for the use case that you might have, right, you could have a workspace that is set up so that all that software comes pre-built once you set it up. Much like, you know, having laptops that have come pre-configured with a certain set of tools, but instead of handing out physical laptops, it’s on the cloud. The virtual collaboration, I think I’ve gone through a lot, the multiple workspace, I think I mentioned also. Data security I kinda mentioned, you know, anybody who’s doing data analysis with anybody who has, you know, talking to somebody that they weren’t the ones to collect it, I’m sure has run into challenges where folks are reticent to, you know, share data. So that’s why in this context, it’s really important to note that like, you know, we can lock that environment down and make sure that only the people that can log into it.
317 00:14:44.310 --> 00:14:47.400 have access to it, that’s a really important point.
318 00:14:47.400 --> 00:14:49.140 So it’s not really like the data
319 00:14:49.140 --> 00:14:50.520 are going out of somebody’s control.
320 00:14:50.520 --> 00:14:51.540 Again, they’re kept in a place
321 00:14:51.540 --> 00:14:53.490 where anybody who wants to can remove
322 00:14:53.490 --> 00:14:55.563 that data again and delete it.
323 00:14:56.580 --> 00:15:00.664 And then if there were to be very computationally aggressive
324 00:15:00.664 --> 00:15:04.353 things to do, it’s very easy to scale it up.
325 00:15:05.360 --> 00:15:09.510 And that’s something that folks also like.
326 00:15:09.510 --> 00:15:13.710 So how, you know, how are ways that this kind of workspace
327 00:15:13.710 --> 00:15:16.680 can support biostatistics research
328 00:15:16.680 --> 00:15:18.270 and data analysis in general.
329 00:15:18.270 --> 00:15:20.280 So I mentioned data science as a service
330 00:15:20.280 --> 00:15:21.990 a little bit in this example.
331 00:15:21.990 --> 00:15:25.547 So this would be the case where any organization
332 00:15:25.547 --> 00:15:28.880 who say doesn’t have biostatistics
333 00:15:28.880 --> 00:15:32.082 or data science expertise local to them
334 00:15:32.082 --> 00:15:36.090 might be interested in sort of renting time
335 00:15:36.090 --> 00:15:40.020 or having some part-time person come in to help with that.
336 00:15:40.020 --> 00:15:42.401 And that’s a model that we’ve seen work well
337 00:15:42.401 --> 00:15:44.250 both for labs and for companies.
338 00:15:44.250 --> 00:15:48.510 One way in which labs really like it is new PIs
339 00:15:48.510 --> 00:15:51.150 with a startup package that just, you know,
340 00:15:51.150 --> 00:15:53.970 first few weeks into their appointment
341 00:15:53.970 --> 00:15:56.760 with an R, right, no staff yet.
342 00:15:56.760 --> 00:16:01.323 Nobody, but they’re coming in with data from their previous,
343 00:16:03.182 --> 00:16:05.744 you know, from their postdoc basically.
And what do they do, right?

They need to write grants, they need to like hire stuff,

so we've actually found labs are very happy in that circumstance just to get going, you know,

So data sciences service can be very useful for that.

So data sciences service can be very useful for that.

So data sciences service can be very useful for that.

So, you know, I'm not sure how much it's affecting folks in the room, but the NIH over time has gotten increasingly serious about making data sharing happen for real for real,

And so this year in particular,

So, you know, I'm not sure how much it's affecting folks

So, you know, I'm not sure how much it's affecting folks

So, you know, I'm not sure how much it's affecting folks

And so this year in particular,
strategy figured out upon submission. And even, you know, saying you need to set aside some budget for that. 'cause it turns out data sharing doesn’t happen for free, doesn’t happen for free, you know, for PIs for their time, right? So that’s also something where, okay, I don’t have the expertise to figure out which of the billion databases I might share my data in. Could somebody come in and help do that? Well how do you do that? You know, when I did work in the neuroinformatics space as a graduate student and I was trying to help figure out for neuroscientists how to get data that they had, you know, collected in a very laborious process of experimental collection, was trying to help them share their data 'cause they wanted to comply with these policies even back then, you know, very frequently I would get the challenge of like, "Yeah, it’s in a hard drive under my desk, right? Physical hard drive sitting under my desk, right?" Like, okay, so you can go pick it up and like take it away
and do something with it.

But you know, they don’t have the expertise, you know,

locally to even know, okay, now we’re gonna plug it in

and we gotta look through it

and like, oh, the PhD student is left three years ago.

And like, how do I do that?

So the idea of, okay, if all we can do is like take that

hard drive from under the desk

and like plug it in the cloud, share it on Dropbox,

okay, something like this or you know,

have a conduit to get it to the cloud,

share that folder in a workspace online

and then have somebody else that does this all the time

like go through all that and do their best to start,

you know, documenting what they find,

maybe raising questions that they might find, you know,

to present to the PI,

"Hey, I know your PhD student left three years ago,

but you know, can you tell me a little bit

about this experimental methodology?"

There’s now at least a hope that you can start,

you know, standardizing that data,

sharing it in a better way,

making the NIH not come kick down your door
416 00:19:09.097 --> 00:19:11.040 with the data sharing police force
417 00:19:11.040 --> 00:19:13.680 that I’m sure they’re setting up now.
418 00:19:13.680 --> 00:19:14.580 Okay probably not.
419 00:19:15.519 --> 00:19:20.519 Okay a third way is through workshops.
420 00:19:20.700 --> 00:19:22.920 And I’ll have some specific examples
421 00:19:22.920 --> 00:19:24.780 a little bit later about this one.
422 00:19:24.780 --> 00:19:26.520 But if you think about, you know,
423 00:19:26.520 --> 00:19:29.670 the experience of either physically traveling
424 00:19:29.670 --> 00:19:31.440 or doing what we’re doing here
425 00:19:31.440 --> 00:19:35.760 and then being exposed to software, right?
426 00:19:35.760 --> 00:19:37.230 It’s one thing to have slides show
427 00:19:37.230 --> 00:19:39.060 you pretty pictures of what software looks like.
428 00:19:39.060 --> 00:19:42.787 And it’s another thing to say basically like,
429 00:19:42.787 --> 00:19:47.120 "Hey, log into, like go right now on your laptops
430 00:19:47.120 --> 00:19:49.740 and go hit this address”
431 00:19:49.740 --> 00:19:52.860 and like, here’s your login and like while I’m explaining it
432 00:19:52.860 --> 00:19:56.569 to you, check it out, play with it, right?
433 00:19:56.569 --> 00:20:00.450 So we’ve actually found that also to be a really valuable
434 00:20:00.450 --> 00:20:05.450 way to do an extra level of education and demonstration,
435 00:20:05.460 --> 00:20:08.790 especially for tools built in academia,
436 00:20:08.790 --> 00:20:10.920 which generally have a pretty small audience, right?
437 00:20:10.920 --> 00:20:14.010 Not a lot of people use them maybe necessarily,
438 00:20:14.010 --> 00:20:15.780 or it’s like a very niche community.
439 00:20:15.780 --> 00:20:17.700 So the total number of humans is not great.
440 00:20:17.700 --> 00:20:21.000 So to have the ability right now in a live session
to be like, let me show you this software you log in right

now, play with it can move the needle a lot on getting folks

to use stuff that that there will really be tools that they will actually help them a lot.

And then lastly, you know,
collaborations between labs, right?
Hey, we just set up a consortia, it’s a five lab consortia and we’re all studying this thing, right?

Hey, we just set up a consortia, it’s a five lab consortia and we’re all studying this thing, right?

Okay, great, we got this really smart set of mathematicians who are gonna do all these great statistics, awesome.

How do you get the data from point A to point B?

Well email, right?

So what if you can improve that, right?

Or you know, the context of, you know, we also find companies wanna collaborate with each other’s

and then universities and companies wanna collaborate

with each other also, right?

So in ways that I haven’t already listed,

but just collaborations of whatever variety.

So when it comes down to those things, right,
it’s one step better than just sharing on Dropbox
and being like, here are the data, go check it out cause you’re keeping the analysis all together, right?

It adds a layer of reproducibility to those kinds of collaborations, which are hard to match in addition to all the other things, all the great best practices for reproducibility.

Okay so that’s four ways to use cloud workspaces support biostatistics research.

So let’s, you know, I think I’ve kind of walked through this example already verbally, but I did have a slide specifically for it. So like this happens in research all the time.

There’s a lab that needs a particular analysis completed and they don’t have the expertise in lab. What can be done? So typically the alternatives are, you know, bring in some student or a postdoc or collaborate with a lab that has some mathematical expertise to perform analysis.

But that can be quite time consuming, you know, that might not deliver the results you’re looking for. Secondly, right for folks who might, you know, be in a position, like I mentioned with early lab set up, right?
Engaging some part-time data scientists from industry could help work on particular problems as needed. And that’s interesting both perhaps from the perspective of me as a company, but also maybe interesting for yourselves thinking about a path through industry where you might be able to do biostatistics for multiple organizations at once, not just one at a time. And then it’s also interesting, as I mentioned from the perspective folks that have the problem that need to get the analysis done. Okay so some case studies, does this happen? I sort of mentioned abstractly, it does, but these are five cases that we’ve worked on in our company and they are, many of them have a, well they all have the theme of being calcium imaging data, okay? So here, you know, swap out biostatistics for looking at data that comes from a microscope. But at the end of the day, that data from a microscope is basically a video stream, generally black and white images that then have to be post-processed. And from that video stream there’s a spatial component of looking at a field of neurons under a microscope.
and a time component.

Like how did those, you know, neurons activity change over time. But there’s a lot of like statistical challenges that have to go into that.

You need to separate the neurons out from each other, okay? They kind of overlapped on each other. So looking at a video stream, you’re not always sure, right? If I’m looking at one neuron or two neurons. So you have to do some spatial analysis to separate those out.

And then you wanna do some sort of peak finding over time. What you kind of wanna extract out is a time series of however many neurons you’ve detected in your field of view and then start to do some additional analysis. And that additional analysis will be based on the specifics of the experimental setup and like, you know, what part of brain were you looking at?

What was your protocol that you applied and what kind of expectations do you have about the time series that you extracted? So these organizations that we work with, I guess, you know, four out of five are universities. So DGIST is Institute of Science and Technology.
in South Korea, McGill University in Canada, University of Penn, UPenn and University of Alabama. And then Maze, which is a small pharma company in San Francisco and they’re all doing calcium imaging work. And I think we served all of these organizations within the same span of about six months. Each one of them had brought different data to the table. They’re all generally in this form of video data with the calcium imaging to extract. All five of them were served by the same data scientist on our side, gentleman whose picture you saw earlier but they had very different scientific protocols, right? So it wasn’t necessary that one person full-time over six months worked on each of these projects, right? Instead we have one individual, who’s able to jump from project to project and check back in with multiple PIs/business leaders, managers to check in on the results of that, right? And that person never left their home, right? So our company is also fully remote, which is nice. And so I think that’s a really powerful demonstration.
of what’s possible for this kind of analysis, whereby, you know, essentially organizations in multiple different countries and different continent in one case, right, can all be served by the same person doing roughly having roughly the same skillset of data analysis but working on data that addresses very different scientific questions all at the same time. Okay, so that’s a thing. And, in each one of these, I should say been done in this collaboration model that I mentioned where there’s one workspace per organization, right? So each organization has their own workspace, they log into it, they can see the results of the data science work that happens. They have all in one way or the other, put data into the workspace, right? And, they’ve all sort of been able to pull figures back out again and direct the flow of analysis in the direction that they wanted through Zoom calls, like the one that I mentioned generally on like a weekly basis or every couple weeks check in. So yeah, a little bit more about the team behind that in terms of thinking about like what it takes
to make that happen. While there is a little bit of like finding those labs and figuring out that they have that problem, which are not taken care of by the individuals on this screen.

But I mentioned, I mentioned Phil, the PhD; another PhD, who’s worked with us as data scientist is Marcus. And then kind of orchestrating behind the scenes, the standing up of these workspaces is a software architect, Zoran.

So again, interesting to think about the different geographies where folks come from being able to serve people in different geographies, but all of them when it comes to a project, like the center organizing node is a workspace. That is the thing that helps coordinate a lot of this together. There are a few other technologies that help. Those of you familiar with like a Kanban board or just really any kind of task driven software, you can bring that to bear as well. So one of the ways you can organize work a little bit better.
than just sending emails back and forth. Is to encapsulate each task, break each task down into a card on a Kanban board. We like the tool called Trello, but there’s lots of them out there that can be used for such things. And then, you know, one card per task. Is a nice way to organize things. And then using a practice from software engineering, you can actually sort of estimate in roughly how many hours, you know, the data scientists might think it would take to do a given task and then use that as a way to figure out like how long it’s gonna take to do a certain kind of analysis. This is a practice that we actually use across my company for all sorts of tasks, not just data science, really organizing kind of everything that we do on the basis of making cards like this and moving things across. And I’m still surprised how many organizations don’t use this. I have lots of friends in academia that do this just for their labs. You guys might do this in your labs, I don’t know. But for organizing oneself,
639 00:29:43.560 --> 00:29:45.690 even if you do meet in person,
640 00:29:45.690 --> 00:29:47.875 having this sort of set up in the cloud
641 00:29:47.875 --> 00:29:50.943 can be very helpful for organizing work.
642 00:29:51.840 --> 00:29:53.610 Not sure how new or not new this is
643 00:29:53.610 --> 00:29:57.300 to those of you in the room, but something
we use.
644 00:29:57.300 --> 00:29:58.440 And then of course there’s Slack,
645 00:29:58.440 --> 00:30:01.743 which I think has pretty good adoption
amongst academia.
646 00:30:03.360 --> 00:30:06.219 We do find almost every lab that we talk to
647 00:30:06.219 --> 00:30:08.883 pretty much is on Slack or some version of it.
648 00:30:09.780 --> 00:30:12.210 Companies are using Microsoft Teams,
649 00:30:12.210 --> 00:30:13.470 which I personally like less,
650 00:30:13.470 --> 00:30:16.620 but you know, but we use that too.
651 00:30:16.620 --> 00:30:18.123 But basically, you know,
652 00:30:20.430 --> 00:30:23.490 one thing that we do that maybe others don’t
653 00:30:23.490 --> 00:30:25.800 do is to connect a Kanban board like
654 00:30:25.800 --> 00:30:28.410 the one that you saw to spit out notifications
655 00:30:28.410 --> 00:30:31.020 in a Slack channel at the same time,
656 00:30:31.020 --> 00:30:33.850 which can be really nice if you are a Slack
based person
657 00:30:34.740 --> 00:30:37.260 to just like be able to see how tasks are chang-
ing
658 00:30:37.260 --> 00:30:39.600 and evolving in the feed,
659 00:30:39.600 --> 00:30:41.880 which then doesn’t require an extra conver-
sation, right?
660 00:30:41.880 --> 00:30:45.210 Like ”Hey, so we agreed on Monday that you
were gonna,
661 00:30:45.210 --> 00:30:50.210 you know, do that t-test on this survey data,
662 00:30:50.430 --> 00:30:52.410 how’s that going right?”
663 00:30:52.410 --> 00:30:54.960 Well if they’ve moved that card,
which was like T-test on survey data from the
to-do column,
to the doing column,
a little notification’s gonna pop up in Slack.
And then when they write a comment like,
"Yep, you know,
I ran the test and wasn’t statistically signifi-
cant,"
then that’s gonna pop up also.
That comment will then be relayed into Slack.
So then when you go back to check in,
you don’t have to ask that question.
It’s like, "Yep, I saw that it happened
and by the way I saw that it happened on
Tuesday,
you know, now it’s Wednesday, you know.
I forgot to check back in with you about it."
So like that idea of asynchronous work can
happen
in this cloud-based context also, which again,
like we use also in all other parts
of our company can be really helpful
for moving projects along in lots of ways.
So yeah I’ve told you a lot
about a particular example then of doing
work.
I wanna call Adria back in here
to extend a little bit more in a partnership
example
that we’ve had some experience with.
back to you Adria.
Thanks, so one thing that Stephen
mentioned was, you know,
Another challenge we might face is, okay, where do we go find people who have data that they might need help with? And we were thinking about where does data come from, right? And so one area that data’s generated from is through devices and manufacturers make devices that are sitting in labs. So we thought of the idea of let’s have discussions with these manufacturers and see if we could form some sort of partnership. Now when you’re forming a partnership in industry, you need to think about why that would benefit both sides in order to kind of engage your perspective partner as to why they should talk to you right? So one thing that we identified was that a key aim of manufacturers is to provide additional support to their customers or make sure, hey, I have a customer or a lab that has data and then what if there’s an aspect of their data they don’t know how to do something or they don’t know what to do, maybe they’ll stop using my device down the line because the data’s just not useful to them at this point.
'cause they’re lacking a skillset. So we thought of an idea whereby we could approach device manufacturers and kind of explain what Stephen explained about our data science as a service offering, and say, "Hey look, we could form a partnership with you, whereby as an offering, in addition to extending a warranty on your device, you could offer custom analysis support or data science support to any interested customers, whereby they could use cloud workspaces to put their data that they’re collecting and then they could work with someone like Phil to solve a challenge that they might have.” And so actually successfully did form such a partnership quite recently. And if you go to the next slide, you’ll see, we are now working with a company called Neurophotometrics. They produce a device that does the imaging that Stephen previously described. And what our partnership involves is we essentially offer cloud workspaces as a solution to their customers, whereby when they collect their data, they can then work on our cloud workspaces alongside Phil or ourselves and we can work with them.
to solve any challenges they might need.

Now who are these customers of Neurophotometrics?

They are a bunch of different labs kind of all over the world as well.

Mostly academics, some in industry as well.

And so it’s that way for us as an organization to kind of find potential labs we didn’t even know had the challenge.

And then it’s also solving the problem for NeuroPhotometrics of how do you keep your customers happy if you don’t really offer a service they’re already kind of asking of you as a follow-on for providing this device.

So, so far the partnership is fairly new. It seems to be working quite well so far and we’re meeting new people and already getting kind of more projects like Stephen described for Phil to work on. So we’ll see how it goes.

But this is just one way to show you that it’s not just about kind of solving a problem for a customer, it’s about where do you find your customers and that could be through an industry partnership.

Awesome, thanks for that.

So I mentioned one other model earlier, which is workshops.
And we have done a few of them actually as well in the computational neuroscience space.

So now the space near and dear to our work with Robert.

So one of those projects was a collaboration actually Brown University on something called the Human Neocortical Neurosolver.

We have kind of a neuroscience bias in the company.

We like doing those sorts of things.

We did a workshop also.

We helped facilitate a workshop that allowed a software tool that came out of this particular collaboration to be shown.

And, let me show you a little bit more.

So in this case, I’m actually gonna switch away from the Human Neocortical Neurosolver and also show you an example with NetPyNE, which is the thing that Robert mentioned earlier.

that we work with as well.

It’s similar to HNN.

In both cases there’s a computational model of a neuron, okay?

Just think of like, you know, a spatial model of a neuron that has a cell body and has an axon and dendrite, that kind of thing.

And you wanna simulate something about it.

And so you have a specialized piece of software
that knows how to look at the model of a neuron, the way that it’s shaped and how to get signals out of it basically, right?

So in collaboration with NetPyNE also a software platform called Open Source Brain at UCL that we’ve been partnering with for a while. You might have something that looks like this. So what you can do in a workshop context with something like a workspace that’s really exciting, as I mentioned to you before is have people put hands on with the software itself.

And this is one of those pictures from one of those workshop that we did, where you can kind of see what everybody’s looking at. So everybody brought laptops in, right?

And they’re able to launch in this case they’re literally, you can see several of ’em, like this one up in front and this one over here, they literally have exactly the same screen up they literally have exactly the same screen up not because they’re logged into a Zoom, but ’cause they’re actually logged into essentially a workspace environment where they can also like, you know, change parameters around.
So you can get this hands-on tutorial effect in a workshop, in this context. That is kind of hard to do any other way if you don’t have that. If it’s deployed as web-based software, that makes it a little bit easier. But if it’s not, you know, if it’s something that’s traditionally supposed to be on a desktop, then this is kind of the only way to do something like that. And this was at an academic conference, I think CNS that gets held. So yeah, from all that today then kind of wrapping up the part where I just was just talk at you and I hope those questions that you guys have, what do we sort of talk about today? Like how can some cloud-based data science tools help enhance the ability to do biostatistics health informatics research? I’ve been, you know, leaning on some examples that are heavily neuroscience based, but we kind of think that that’s not the thing that’s particular to this, right? It’s still, you know, as I started at the beginning, you know, doing some analysis, you know, sharing the results of the commands that we’re using in the analysis.
and then sharing the output of that analysis, right?
Like that’s where we began.
I think that’s common to every technique.
We’re bringing some kind of science and math to bear on some data, right?
So what we’re finding is that, you know,
by using cloud-based platforms really can help us facilitate collaborative research,
allowing colleagues to share data and work together.
You can help labs efficiently gain access to additional data science support if that’s desirable.
That they, you know, otherwise might struggle to get or is just kind of unaffordable.
 Doesn’t make sense ‘cause there’s too much of a person.
And then finally in the last example, right,
you can facilitate, you know,
distance workshops that allow much more immediate hands-on experience with certain software.
So with all that, I will thank you all for listening
 to us for a full 40 minutes
and happy to take any questions that you have on this
or any other thing I can help directly.
Thank you very much.
Thank you so much."
Does anybody have any questions for our presenters?

I'll start if there's no questions.

So data science is a service growth industry. People want jobs.

What's your take on the industry on that?

We are about 18 months into our exploration of the market.

We have seen growth so far.

We think there's more to go.

I showed you those five labs, in total maybe served certainly more than a dozen,

I wanna say maybe like 15 and like labs plus companies or so

in those 18 months.

We had to figure out lots of other stuff along the way.

We think there's a need, you know, like I mentioned

and folks that have the skillset to, you know,

provide that data science service

that are continually in demand.

So I'm gonna say yes, it's growing.

We're always wondering in industry how fast, you know,

that's always the question,

but it's definitely not shrinking.

Alright, that's an exciting option.

Yeah just really quick,

what happens with authorship?

If you work with the lab very closely on a project,
they come out with a really good publication.

How do you deal with that in this industry?

Yeah, great question. Thank you. So as a company,

we don’t require to have our data scientists listed

as co-authors on papers.

I think from an ethical perspective

in the case where the contribution that the
data scientist has made are very significant

sometimes PIs have asked the question to us,

you know, sometimes PIs have asked the question to us,

what sort of acknowledgement

would you like of the data scientist?

And if the PI feels that, say, you know,

someone who has a PhD who works with us

has done enough work that it merits authorship,

they’re free to add that person.

We don’t require that.

Otherwise, you know, an acknowledgements nice always right?

But also not required.

I think, you know, sometimes the nature

of the contribution really matters.

So, you know, as a company it’s a little bit

like how much do you acknowledge

the vendor of your microscope, right?

You might say, okay, I did this on a Nikon microscope

or you know, but you might write that more

as a method section.
And then if like a technician came out and like helped you calibrate it, you’re probably not gonna give that person an authorship either. But you might acknowledge them if they did extensive help that like led to some novel process. So on the whole, it’s a case by case conversation that scales based on the level of the contribution, but it’s not the first thing that we think of. It’s not like, "Hey, because we did anything for you, please put us on a paper." Definitely don’t do it that way. It’s more the opposite, which is like, you know, we’re gonna do a thing for you. Probably, you don’t need to cite us. But if it gets up to a certain point and we kind of mutually agree that that’s appropriate, then we’re happy to discuss that.

Thank you for sharing Stephen. So I have a quick question too. So if you’re running on data sets, one cell may take really long time to run, then how do you solve the concurrency issue? Let’s say there’s multiple people collaborating online that when the cell is running, what if some other, another party just clicked stop.
or doing something random?

How do you solve the issue that people are on the same page when something takes really long time to run?

Yeah, great question.

So a few ways,

one nice thing about a cloud workspace is that we can expand the number of processors and the amount of memory kind of behind the scenes transparently.

So basically you can log out of the workspace and in five minutes log back into the workspace and we’ve like doubled the processing speed and like doubled the memory.

So we tend to keep our default instance at like a reasonable like laptop, like probably not a high end.

And then when we discover cases like what you’re talking about where like, yeah, no, that cell requires a lot

and we kind of know a little bit in advance, like we’re gonna wanna run that a lot, right?

We might do this, which was we might like just beef it up, right?

And that’s cool that we can do that.

And then the question becomes like, does that need to run, you know, 24/7, does it need to run every day, every week, every month right?
We think a little bit about that because then there’s some additional costs on our side.

If you’re gonna do it for like an afternoon, it’s like really not, it’s not worth making any additional, you know, requests of somebody.

But there’s another part of your question I wanna get at too, which is like maybe overriding each other, right?

So that can happen.

And that’s a little bit like software specific. So like in a Jupyter Notebook, you could, if you don’t coordinate a little bit with your lab member,

like overwrite something in one cell at one time, right?

The other person didn’t notice.

So for that, we have some best practices, you know.

By far the most common, you know, example that we see is,

is like two or fewer people collaborating, but if it were three or four, we’d probably recommend that they do a best practice.

of like, you know, while you’re doing work that’s separate and you’re not like talking to each other, do work on separate copies of the thing, right?

And then come together in a meeting and like put it back together, right?

Usually is the better practice if you’re say,
working on a Jupyter Notebook, and you know, communicate, you know, using some other method like a meeting like this. So yeah so those are the two aspects. On the one side, if it’s computation intensive, we can make it bigger. If it’s actually about people writing each other, we recommend some best practices for communicating outside of the workspace. All right, I have one more question. So like in the old days, people would buy a nice computer for their lab or maybe a couple of nice computers and like then everybody would log in at that and it was a one-time cost, right? And so how have you found, I don’t know, I mean, so it’s a very different model for both academia industry, wherever that’s trying to transition from this one time cost. Where now, you know, you might still be using this computer 10 years later for good and ill. versus sort of this continuous cloud-based thing. I don’t know, do you have any words of wisdom on this transition? Because it seems like, you know, you pay
for a cloud computer and if it’s on constantly, it eats up a lot of money.

Yeah, yeah.

So really good question.

So I think and-

Lose control of your data also, which to some extent,

like somebody else has your data.

In theory, yes.

But you know, I think some of this is just like a journey.

and a transition that, you know, scientists are making.

Those of us, like yourself, we’re more software engineer minded,

have been comfortable with the idea of say, you know,

like all of our company’s data, for example,

is kind of in Google’s clouds,

Google’s workspace technically.

None of it is sitting under my desk, right?

But we’ve gotten a level of comfort about data ownership

based on essentially trust and agreements

and our understanding of how certain sections

do kind of disk are like cordoned off, you know, for ourselves

and lying on some of those best practices.

But to get to the heart of your question,

I think the best metaphor is like

buying a house versus renting an apartment, right?
So, you know, going down to Apple and picking up a laptop or Dell or whatever you wanna use, right, is that’s the buy model. And we’re super comfortable with that. The cloud model is more the like renting the apartment. And certainly people make the choice, you know, not to rent sometimes because it’s like, doesn’t work out economically, right? It’s like, “Hey, I’m throwing money away.” Sometimes people throw, right? But what is the advantage of renting, right? The advantage of renting is, you know, if a thing breaks in your rented apartment, you call us and you’re like, “Hey, this thing didn’t work, please fix it, right?” And then there’s this scaling thing, right? Which is like, if you go back to Apple and you’re like, “Actually can you add like double the CPU and double the memory?” They’ll be like, yes, you can pay us for that, but it’s gonna take a while, right? And it’s not gonna happen flexibly and scalably.
So I think it fits into a different space, right?
Obviously these two come together,
I’m talking to you on a physical laptop that I own, right?
But I’m also using cloud instances to do things.
So I think it’s like, it fits into this niche where like,
actually the most useful computer for this purpose,
this collaborative purpose
is a rented one, right rather than an owned one.
And you know, maybe that means when I’m not using it,
I’m not paying for it at all, basically, right?
Like, if I’m like paused on this collaboration,
then I’m like actually not paying for it at all,
but then I can bring ’em back and six months and start paying for it again.
So this is what I hope that folks take away is like,
it opens up a lot of new possibilities.
And the ones that we’ve gotten are certainly not the only ones.
There’s just like lots more that you can imagine or envision.
But, but yeah, it’s a mindset change and it’s one that I think, you know,
requires some adapting, yeah.
All right. Thank you so much.
I have a question for you guys
if there’s not another question for me.

There’s a question on the screen.

Sorry, I have a question.

I think piggy-backing off of that question-

Hi hello. Hi Noelle.

Actually Hi.

I used to like physical like pieces of data
and like having physical hard drives.

So like what is the security for data that’s on the cloud?

Yeah, so folks like,

we ourselves build these cloud instances
on the back of three major providers,
whose names you’ll recognize,
Amazon, Google, and Microsoft okay?
Those are the big three cloud providers
and they make a guarantee to us
and then we make a guarantee to our customers
about the data protection.
So it’s kind of like a layer cake.
And the foundation of it begins with, do you trust Amazon?
Do you trust Google? Do you trust Microsoft?
Some people say yes, some people say no,
but fundamentally they are the ones that, you know,
build data centers, right where the physical aspect
of these computers actually live.
So, you know, this virtual computer,
maybe if you go and like,
"Hey, show me the hard drive where this lives."

You're gonna go out to like, I don't know, Washington State near some power plant basically,

where it's very economical to set this up, right?

So they then guarantee like,

how do you know that that's safe, right?

Well they guarantee that they're following industry standards to secure those facilities, to lock them down,

to like continually maintain and manage the networks

that are there to patch the servers

that they're using to keep ahead of any security faults.

So there's one layer of this

where we rely on these big providers to do their jobs.

And despite the last 15, 20 years of like hacks

that you've heard about whatnot that happened in industry,

these three providers so far have managed to avoid

being hacked in any major way.

Like you've not heard of like Amazon getting hacked,

Google getting hacked, Microsoft getting hacked.

If tomorrow Amazon gets hacked, then yeah,

we're all worried okay?

And then we probably would need to shift around.
But so there’s a fundamental guarantee that like all cloud kind of relies on and it’s like good to talk about it because like we all have to kind of trust these, you know, these large providers. But they also invest, I’d say millions or hundreds of millions of dollars in computer security. Like if you’re in the field of computer security, you know these guys because they are sort of world leaders in this sort of thing. Microsoft, you know, notably was involved in doing some forensic analysis on like Russian hacking back in 2016. Like they were some of the first people to notice that a state actor like Russia was on the scene doing the various things, taking over computers. So generally the community of software engineers that do cloud work know these things and kind of rely on Google, Amazon, and Microsoft to like make these investments in computer security. And notably like, I don’t go like set up my own data center because I know that I would have to invest millions
of dollars in having an equivalently good computer security

team to like watch out for Russia, who by the way also invests hundreds of millions of dollars

to try to hack these things.

So, the world of computer security is a problem.

So there’s that level of trust, okay?

And then on top of that, you have to trust one more level,

which is the group that like sets up the workspace.

So you kinda have to trust, like if it’s from us,

you have to kind of trust us that we’re not screwing

something up on top of all of those protections

‘cause it is possible to do that at the level of like,

you know, Jupyter Notebook that our logins are well used.

So we also invest in using industry standard

like login protocols, so that only the people that we say

can log in right?

There’s a layer of software security there that, you know,

we have to be on top of patching at one level also.

So these are all the things that make that secure.

And the last thing would be like,

do you or don’t you trust us to like not to,
to not go in and do something nefarious with your data
even though we’re the only ones that can control it.
So you trust that nobody else can get into it,
but do you trust us?
And then that becomes,
yeah a question of like, you know,
going back and checking your references, you know,
talking to other PIs, making sure that something nefarious
hasn’t happened, you know, there.
And you probably wanna gain some confidence on that.
But what we’ve found is that organizations are getting more and more comfortable with that.
Dropbox is a publicly traded company,
lots of people put stuff on Dropbox.
When you put something on Dropbox,
you’re essentially trusting Dropbox.
Dropbox is also built on one of these three providers same way, right?
So it’s that kind of idea that takes some getting used to but you know,
becomes increasingly useful to do this kind of work on.
And we see large banks and large pharma companies having taken their time to also adopt cloud
large financial institutions.
But over time there’s been increasing comfort as some of these security questions have been, you know, asked and answered. So bit of a long answer, but thank you for the question ’cause it’s important.

Alright, thanks so much. In the interest of time, I think we’re gonna have to stop it here, thanks again.

Really appreciate. Thank you guys. Thank you all for your time.

Have a great day.