**Interviewer**  
So, I have listed those questions by the research sub-questions and for the sub-question around safety. I have one topic or question.  
So when we're using machine learning in the process of discovering new materials, how could that lead to a safety issue, to a safety risk?

**Respondent**  
Do you mean safety during the making of the materials or during the use of the material?

**Interviewer**

So, if for example we developed the machine learning model that is going to predict what chemicals to use in order to create a material with certain properties.

**Respondent**  
OK. I think first, when you build your model, you have to build in some safety constraint. For example, that could be on the selections of the components.

And you want for example, a lower selection of a formulations to make use of a component that for example is not in reach, or that is toxic, these are the easy to screen out.

Now, let's say you’re building a synthesis model, there's a few things that you can represent. For example, you could compute the reactivity of your systems as well. So, you bring in a model to compute what could be the reactivity of the chemicals that you put together. And that would be a safety concern.  
How do you predict that the model won't lead you to generating a poisonous material?  
It always comes down as well to human decisions.

Machine models will make you a proposition and then you would have to make a decision out of it.  
And this decision is how you're going to try out this material. And that's where you have your management of change that would prevent any safety concerns, whether on the scale up, on the production of these materials, or on the use of these materials.

Let's say, this material is actually an assembly of compounds that are too low to be used in Europe for example. Could be something like that. So, I would say the one that is taking the risk is not the machine itself, it's the user.

Another risk that we have with machine learning is essentially on the extrapolations that are a bit on the edge of our training space domain. So, let's say we are in the process of exploring new materials. That is to say that we are trying to seek solutions in a domain space of your solutions where you have little experience.

So, your model is more likely to have been trained on a small zone of your solutions, and then you want to reach something that is outside that zone. How do you do that safely?

First of all, you need to have a kind of safe extrapolations.  
So you need to make sure that your model won't blow up in terms of craziness of its predictions as soon as it’s outside his training zone.   
And to do that, there’s several techniques and basically it means trying to avoid overfitting.  
So you will try to avoid fitting the noise in your data, which will lead ultimately in the diversions during the extrapolations.

But you can constrain your model to be more on the exploration side, OK. I want to have a discovery process. So I'm not going to use the same type of models as in the control zone. Let's say, I want to build first a model to control what would be my product discovery in certain zones, that I know I can handle with my plants for example, and that I have a lot of data for.

In that case you will try to use the model with the best accuracies which will interpolate very well in this domain space.  
But this model with a high accuracy, would have a high risk when you're outside the space.

So, you want to balance that and you say, OK, for this model where I want to be a bit more exploratory, I need to change the model.

You could define a different kind of model where you will look not only at the accuracy, but much more trying to find the underlying bias, or the fundamentals in the data, that would drive a better projection of your performance in the extrapolated space.

And that's why you have always, what we call a balance in complexity and accuracy.

Take a model, a simple linear model, and you predict the trajectory between two points. It's the minimum that you get.

OK, now let's say you have one point here and then you have plenty of points there. If you'll fit it too much, you will fit a sinusoid.

But this sinusoid here, won't be able to get to the point there. but if you fit, forced to a fit curve, a fit line through these data points, you will end up with a line that will be parallel to your points that you're looking, but you will still be in the vicinity of your explorations. You would not have developed like, OK, you have quadratic terms or whatever. So that's just to give an example of it.

But I think to really mitigate the risk, the safety risk, basically either a work predictions or absurd predictions.

That's two different things, but the absurd predictions, you really have to put a bit of human knowledge behind it as well and find out; OK, this is not going to work or this is impossible, I should not do that. And that's also a bit when people are doing some catalyst explorations, they say, OK, I have maybe selected 40 solutions, but actually I'm going to test only 10 that fundamentally makes sense to me to test. It could be a good idea *[to test]*. That's a bit how you mitigate that. So, you mitigate always by preventing your systems to over predict.

Basically, you should not ask more from your data than what is in the data. What is really in the data? And actually, you have to think about, OK when I'm certain to what is in the data.

And I see a lot of people that are developing a machine learning models that actually are just focusing on the accuracy. Yeah, I get the “we have the best accuracy of the models” and select the ones that is very good in the test data. And you know then that it's validated.  
OK, come on. So, you train well your models. And then you test it and I agree, you selected the models that actually perform very well in the testing. Now guess what happened when they receive a new batch of test data, the model performs poorly.  
And the reason is simple. There, they just have overfitted their models and then selected the OVERFITTED model, that actually works fairly well by chance for this test data.  
But there’s a chance if you go a bit on the other space of test data, that has not captured the right physics. And I've seen a lot of different people were wasting years on developing models because of that. So my approach is always to take the sample path.  
Be careful about accuracy and the dominating dogma of being accurate. I think about what do you learn from the models, what is the trend that models predict? Does it have its physics? And that's why how safety can be managed.

**Interviewer**  
And that actually relates pretty closely to one of the other questions that I had around the model evaluation/model validation, where indeed having a high accuracy model is wanted. But how does it perform outside of your data set? So, when we look at the model validation, do we accept within Dow to have that model validated by test-data only or will that always need to be backed by an experiment?

**Respondent**  
What I see is that very often the solutions proposed by models are just a kind of information for the next DEU space for example.  
I've never seen people really considering the solutions; OK that's the solution I need to do. But that's on matter productions. Now if you go about matter predictions, like properties predictions, people tend to believe more of what comes out of the model.

When the model proposes synthesis proposal or proposed formulations, the business, the formulators the chemist, he wants to try out the idea.

You have to go back to the physical world at some point. You have to measure the materials, you have to produce it or whatever, so you have to go back to the physical world.  
Now you have this range of models that actually remain in digital space. For example, they give to a JS&D *[name of a software application?],* the predicted performance in the applications of its material.   
And then he will make a business decision from it.

It's probably lower risk, but maybe not. Let's say I predict the adhesions of a glass sealant and I tell you that with this adhesion, the lifetime of your glass sealant will be at 30 years.

Now you build this sealant and you go into your skyscraper and 20 years later the windows fall off the skyscraper *[Laughter].* OK, so there's always a relative thing to it in safety.

But the problematic is that, it's the control. I mean, people tend to say, OK, I have trained models.

Look, it performs very well on your test data you gave to me. I'm done. This is your model. Use it for forever.

One thing that we need to do and I'm contributing into as well the machine learning operations *[MLOps/MLO].* We need to implement more control feedback OK, about how this model performs. How does it continue to perform, based on your data? At which point do we trust all the kind of reevaluation, retraining or remanufacturing of this models?

And this is very important. There's new data that come in and there's probably new data that we need to test. And if you rely on the model that has been built five years ago, I think you may be out of business very rapidly.

So, I assist more with my colleagues to say OK, what is the methodology? What is the pipeline? What is the workflow that you're building that allow us to develop these models? And I really hate when people say; yeah, I don't know, i did everything on my computer, I make a big Excel spreadsheet and then from this big spreadsheet I develop a model. Yeah, guess what.

What do you need to build a new model? Yeah. I need a, you know, to build my Excel spreadsheet. While this is, if you move to another role, who is going to do that? Yeah.

So, I think we have to think about the models as a being. It's alive. As long as we use it, it's alive. And we have to make sure that we feed him correctly and we treat it well and we observe it. OK, so we should not be blindfolded by the predictions. We should always look at it and then provide feedback, as soon as we start to see deviations from the models to the reality to what is.

But I don't know if this process is in place. At least not on the product developments. Maybe on the manufacturing but not on the product development.

**Interviewer**  
OK. That actually touches on the next question that I have here that’s around the process for developing models, but from what I'm hearing on the feedback about the evaluation/validation, developing that model is like a constant evolution of that particular model.

**Interviewer**  
You may have already answered it a little bit, but is there a standard process for developing that model? and if not, would there be a need for such a process?

**Respondent**  
Yes and no. So first of all, there's not yet a standard.

Depending on the researcher, I would say, on the group of researchers there's a different process, different ways, different methods to develop a machine learning models.

There are people that will use R-studio and R, to develop their models. Other people that really prefer Python. Other people that will only do TensorFlow in Python and the other will prefer Pytorch. Other people would say; II have tried out the symbol integration package and it's nice.

So, there would never be a standard. Because it's too young, it's a too young technology. It's still evolving. Today, the packages we’ll consider as standard today, like Scikit, maybe in five years’ time will be fully obsolete. Or there would be a fork of Scikit.

So, I think it's very risky if we limit ourselves to a standard. But, well, at least the standard must not be done on the choice of the methods of building the machine learning models.

We are looking into the ML OPS, and it’s little more in a standardizations of model deployments, model sources and model documentations, things like that.

More than on the specific tools. So right now, Informatic Research will support free languages, mostly R, Python and MATLAB, because I requested it *[Laughter]*.

I matter because it's very much used in the engineering world, I mean manufacturing and so on. People are really relying a lot on MATLAB. And MATLAB can interact very well with Python, with R probably as well. So, I think what they have in common is they are high level languages, and they rely on packages themselves and those packages will be maintained, will be stopped and so on. I mean, if you use package in GitHub from Python for example, most of them are abandoned after three years.

So, what's going to happen? Well, the next model you will build will need another package or different one. So, you have to think about, OK, I'm going to build the model a bit differently, achieving the same goals.  
So it's just like all the roads leads to Rome.  
You know, there's not a single best solution, a single methods that one fit all approach that we should have. And we should allow the creativity and the flexibility of the researchers to use the tools that they want. But Dow should provide the infrastructure to support those tools.

That's what I believe IR *[Information Research]* would need to do. Instead of focusing on asking us to develop everything in Python. They should care about, If I develop something in MATLAB, how you're going to distribute it? And not to tell me; Ohh but it's not possible. They don't say that, they say; Ohh we only know how to do it in Python so.

*[comment to IR]* Learn to do it with other languages. I mean there's you know an API for here, for there.  
So I think there's not really yet a standard, a lot of science community at Dow are taking initiative. I'm part of the Inside Science organization as well, that are trying to develop a bit more structure around it.

But you know, it's still booming, so. That would be a, you know, it's a jungle and then you have to clean the trees later, to make a nice forest, but it's natural evolutions so.

So, the company just needs to support that, that's all.

**Interviewer**  
That makes sense.

Next couple questions are around the data collection. Starting out with, what data is available within the organization and in what form? But then also the subsequent questions; on how to get that data? Is it easily accessible, stored in a centralized database or is it more a matter of collecting bits and pieces from various locations?

**Respondent**  
OK, yeah, so first of all, there's structured data and unstructured data. OK, the structured data are, let's say, you know number of repositories that we can find here and there.

Preferably they are coming from a design, they are designed data. But the reality is that a lot of the Dow information is contained in unstructured data and in an undesigned fashion as well.

I’ll explain. A lot of the data are actually captured in documents. Like Word, PowerPoint, ELN [*Electronic Laboratory Notebook*], CRI’s... All kind of Office repositories, that are nice to create reports, but that are very hard to extract data from. OK, and hopefully when I have a project to start and see, OK, how do we extract the data from those documents. But it's not there yet.

There's a lot of data that is actually not known by Dow. When you have some ??? guy that will assemble 10 years of data on his computer. Who knows in the company that this guy has that? Maybe two or three guys, his colleagues, you know, but what's going to happen? Let's say if the guy is fired? What's going to happen to this data? They will put it on a folder and nobody will care about it, because they don't know what to do with it and where it comes from, don’t know what it means. And so on.  
So there's a lot of data like that, and I have that as well. A lot of repositories on my own, that I own. But, it's not known to Dow.

Do we have a global, standardized and centralized way for the data? No, we do not have.

There is a lot of initiatives, whether in manufacturing or R&D, that are trying to consolidate those assets.

But there’s not a preferred method. I will only take talk about the R&D side; I don't know too much about the manufacturing side.

You know, in R&D we working with IR to develop, let's say solutions. Right now, big businesses like Plastics, for example, are relying on an Oracle database to collect data. But they are limited in the type of data they can collect.

So, for example, if you have an instrument that measures a curve. What is collected in the database is not the curve. It's only one point, like one moment like the average molecular weight, for example. Let's say it's a GPC curve and we have distributions of molecular weight.  
We only, let's say, record the average molecular weight. So that's one point, instead of having the full spectrum information, you store one point.

The same thing if you have a very logical curve, like the viscosity as a function of the shear.

We will store only maybe, what’s called the zero-shear viscosity, so the final viscosity, we’re not storing the full complex graph.

So that's the problem. The data that they store is always a very small subset of the data that is generated at Dow.  
There is an ongoing effort in IR called SDMS, which is to store all the lab instruments into a file warehouse.  
Basically there would be a system that will collect from each of the instruments, let's say the results file. And copy it and normalize its name and so on, in a file storage. I call that the file warehouse.

The advantage of that is that this file warehouse will have its own databases, so you can retrieve the document very easily. You don't need to know which computer it comes from, from which share folders and so on. OK, so that that's already a big, nice step.

The problem with that it’s still not a handy tool to make cures, to generate, let's say to do some analytics. Because you still have to read the data. So, if you read an Excel spreadsheet, it's very difficult to make a big table out of it. You have to import those spreadsheets.

What we are lacking at the moment, and we're still don't have a good solution for it, is that we are lacking a data warehouse for machine learnings and analytical purpose.   
And basically everybody that is doing data science, is building his own kind of data warehouse.

A lot of people are just collecting and building big spreadsheets.  
OK. That's a data warehouse. Or they do an access database on their own. Personally, I developed an SQL database to host this data. Because when you need data, you need to have it in a certain shape, certain format and so on. Because you need to call it. So, you need to control what’s in it. Or you need to have the documentation on what it is.

The problem is that we have poor access to documentation on how the data is currently stored.

Because IR never thought about documenting that because these guys don't care about accessing and where is the data. You just want to click the button and have the data. But a data scientist doesn't want to access just 1 sample or a couple of studies, he wants to access 200 studies. So, he cannot use the same interface. He has to have lower access level, to the data in that database.

But that means that he needs to know how the data is structured, or IR has to create a kind of pipeline to get the data to him. I very often requested; could you write for me an SQL statement, a database query to retrieve this kind of data from your systems. I never got any answer.

Because we don't have methods to make this request you know. In analytical science, we say; measure the DOC of the samples, click on the button, then you have the request sent, then you receive an email. Please send the samples there and this lab and that's it.

We don't have a request where I could just push a button that says OK, I need this data here and there and there. Please provide it, in the table formats and the supporting files in a zip.

There's no way for me to do that. I have no contact at Dow that can do that. And that's it. So I have to do it myself.  
And it's time consuming and that's very… the most frustrating part of the data science at Dow, is that you spend 90% of your project looking for the data.

And that's very handy for people, [*because]* they can blame the data. But it's not an excuse for Dow. We have to have a better data management.

That are suited for the purpose of data mining, data analytics and data modeling. And I'm not sure.  
IR is doing a lot of proof of concept on various things, but those guys are not the real the users of it.

So that should be part the data citizens community that should handle it. With proper experts, that knows how to do this kind of infrastructures. OK, we have big informatics, but we don't have really an infrastructure. An infrastructure that is something where you have an architect.

OK, we have plenty of techniques, plenty of licenses of software and so on. But we don't have an architect that makes something out of it really, for the purpose of our needs.

So, things would change. There's a lot of goodwill in there, in a lot of people.

But I think we don't have the tactic yet. We don't have the tactical plan to do that at the company level.

That's my personal opinion. So, if you share the transcript, that's personal, you know.

**Interviewer**  
I will mention it though.

**Respondent**  
Well, at least I don't see this tactical plan in in action or presented to me.

But the result is that if I take tomorrow a data science project, I know that I will have to spend a lot of time in getting the data.  
And then you know, if you have a 90% of your time in cleaning the data, presenting the data in the way you wanted and so on and sometimes… And it's difficult because sometimes you will evolve it; oh, I need the data in a different shape and so on. So, you always have to a bit refactor your data set anyhow.

But in the end, you spend so much time that you don't have a lot of time left for the modeling activities.  
So I see a lot of people doing a really poor job in what you call, the quality of the models. And I don't talk about just taking the one that is the best predicting one, but as well thinking about it; is it the right models that I need to use? And we talk about the safety risk and so on before.  
Then we don't have the time because we have to deliver the solution and now you know, I have one month now and I've spent four months on building the data sets.

Then the data curation is also some part. I mean you have to basically with your naked eyes create the data to see, OK, does it make sense to include this data or not?  
And because you have so little data, you will never have big data, you really have to pay a lot of attention to the representations. You cannot say it's OK, I'll remove, I’ll remove because I'm unsure. Because in the end you get only 10% of the data left and then you cannot do anything. So, you're forced to really include those shitty data, and really look into it; go back and says, OK, this data and then you call the requester of this data to make sure that you know it makes sense to include it and so on. And this takes a lot of time

So, we don't have an infrastructure yet that fully supports the data science activities in a more routine fashion. I think its people are still exploring what we can do and how we should do things.

**Interviewer**  
And that is also answering some of the next questions that I have, because that was also around the data cleaning, the data curation process, but indeed you already gave me a couple of those challenges that you see for the data collection, but also the data cleaning process.

**Interviewer**  
I actually wanted to go over to the next set of questions, regarding the literature and also the use of publicly available data.

Does the organization have the possibility to make use of that publicly available data and would that be of any use to, for example, enhance or enrich the existing data from within the organization?

**Respondent**  
Yeah, this is done. So, there are material open databases out there. Let's say for example atomic configurations of small molecules and things like that. That we can use as long as everything is documented, people are able to pull from those databases what they need.  
I've not experienced it myself, but I've seen people using it. It's usually at the, I would say atoms level, so we use a lot of data like atomic movement, or whatever. I mean some startup properties that needs to be downloaded at higher scales. So, this exists.

**Respondent**  
There are also other type of databases that are bit more material. But for which maybe we have to pay or we have to contribute to a license. And mostly they're hosted by a lab in the university or by a university. And so, you have access to that database only if you pay something.

Yeah, it's doable as long as there's a documented method to retrieve the data from an excellent source, we can do it.

But then again, that would face the similar challenges as what you see with the data collection from within your organization itself. You need to have clear documentation.

So at least, you need to know what is in the data. But very often there is kind of, at least if you take some data set that are serious, there has been a review process on it. So normally you can trust those data because you're not the only one that would be using it. So, if it has really a defect in the data, other people would have reported it. That's what everybody believes and that's perfectly right, it makes sense to believe it.

So, I think, yeah, we can make use of it and we’re definitely making use of it for a few projects.

**Interviewer**  
OK. And then the next question. When looking at the literature and the research that have been done and have actually resulted into machine learning models, that have been validated by the researchers themselves, obviously.  
Would it be possible for the organization to make use of those models? And if so, what would be needed from the literature to actually use those models?

**Respondent**  
OK, so those are models that are developed externally and that we need to….

**Interviewer**  
Yeah, exactly.

**Respondent**  
Well, I think, probably we need documentation of the input of the models. But most of those models will be a black box anyhow. So even if we have a kind of understanding about what the model is made of, how it has been created, it's probably OK.  
I think what needs to be well documented is the context of the data that's been used to generate the model.  
Because first we validate if the same context is applicable, or if the context is applicable to the Dow case.  
If the context is applicable to the Dow case, the next question will be….

The first question would be, can I use this model? So, is it protected by patents, or by some licensing terms as well? So, we need to be careful about that. So, having a licensing strategy and an IP [*intellectual property]* strategy is very relevant and today Dow is building his own competency in this space.  
But it’s still in its infancy, so we really need to know, OK, what can we do with this model? Are we allowed to use it in the commercial environment? Because very often… The difference with an external model is that the external models are made for university or open source or open data type of research.

And even if they are free, even if they are GPL, sometimes they have a licensing term for commercial use.  
So you could have an open-source repository, but you still have to paya fee to the university to use it for commercial purpose. So, you have to check the license terms of the model.

So, let's assume you'll find a model that is by its context of creations, looks interesting for your Dow case. You have checked out that you may be able to use these models for Dow.   
You bring the model in. So first of all, you have to see the documentation of this model; what do I need to run this model? Like what kind of data? What is the format in and out? So, this needs to be documented.  
The next step will be I'm going to try this model. I'm going to test this model.  
And that's where you will see if you going to use this model permanently or not at Dow. If I'm able to leverage in this model and actually it's quite good, I will make use of it.

And then if that's something from GitHub, you can import the GitHub to your Dow GitHub, and make use of it. But, provided again that you know what it is, and that you have the right to use it. And I think that's very often overlooked at.

**Interviewer**  
If you would compare that process of checking out that external model versus developing a model in-house. Do you think it would even have any benefit to look at that external model or would you then say, well, no, let's spend that time in developing a model in-house?

**Respondent**  
The advantage of using the external model will two-fold. The first will be they have used training data that we do not have. They have used a bigger training set than we have.  
OK, so I could make use of this well-trained model into my case, which is just a small set of data.

And then I don't need to train it. I just need to test it.

You could say, OK, this model is actually trained and validated on big data as I said. Now I'm going to test it on my own and then you decide, OK, the testing looks good. So I will use it on that. OK. You're not going to retrain the models, unless you have to…. for example, you can say, Imagenet or things like that. Those are pre-trained, that are network models on the generic databases and now you have to retrain these models on your own image datasets. But it's already pre-trained.

So, I'm just talking here about the models that have been trained, you don't need to retrain it. But already pre-trained model is also an advantage, because it has already more intelligence than the new model. So, the advantage is that you could have a model that has been trained on the higher datasets and as well maybe that has been trained with an expert.

At Dow there's probably too little experts. We have people that have good know-how about it, but probably are not expert enough. So, they could make mistakes in building the models themselves.

I think for the cases where we need to train the models on ourselves, because our context is different, what we could leverage from external is the approach. What kind of methodology those guys have been following this case, that I can reproduce?

OK, that's what we're doing into the, what we call the physics and form machine learnings. We’re trying to get inspiration from other papers where people have tried to do similar things.

And then we will have some proof of concept to internally apply to Dow cases. So, we have to rebuild the algorithm ourselves. But now we have a bit more understanding of how we should do it. That's I think what the most interesting part is for us at this, at least in terms of science.

So yeah, there's not a yes and no answer, again to that question. But the main advantage will be having access to a bigger training set of data. And as well methodology that has been validated. And that's what we want to leverage in.

The opportunity to leverage a drop in models at Dow like, this model has been applied successfully for Bayer, it has been published and then we take it and we apply it for Dow, is unlikely. We probably would have to refactor the models and use the same methods. But at least you know where to start.

**Interviewer**  
And then I do have a couple questions around the development with regards to software packages, software applications that are needed and also the hardware resources. So as an organization do you have access to all of those software packages and applications that are needed? And are the existing hardware resources sufficient for training and developing the models?

**Respondent**  
I would say rather yes. There are still some IS issues in compiling, in running data, because we have all the safeties, the security. So, the informatic securities preventing a lot of us to, let's say maybe download from SourceForge for example. We had some issue recently with even Anaconda, a Python package to just upgrade the repositories, or just update the packages and so on. So, we raised the case and now it's fixed. But that will come back.

So yeah, I think we have access to most of the open-source packages that are out there. We can do that. For example, I have a developer level to enable that. Otherwise, we would have the bit 9 pop up *[security platform alert]* all the time and you cannot do that.

So, there are some bypasses that we can request to happen. But there's not in the company a kind of data scientist profile, in terms of security where we could have more freedom.

The second is we don't have the freedom to contribute back to the open community. So, we take a lot from the open-source community. Like you know, you download the source code to your computer. Then you modify, you see there an opportunity to modify the source code, but then you cannot put it upstream back.  
Because, OK, there's a full evaluation of external release to do that. So, I think longer term Dow would have to create a kind of method or process for us to allow interacting with the open community.

The other tools, I mean, we rely on R I said, we rely on Python, we rely on MATLAB, we rely on Azure ML.

We have a big contract with Microsoft. So, we have a lot of developer tools from Microsoft that we can use, like the GitHub that we have at Dow. We have a lot of access to the machine learning tools from Microsoft.  
But the reality is, very few people really know how to use all those tools. So, the limitation is not in the portfolio of the tools that we have. It’s in the documentation on how we can use them and who is the expert to help us.

For example, I want to start with Azure ML. I don't know where to start. I don't know who to call first.  
I would have to go back through emails or presentations eventually to find out where to start.

So, we have, I think a Sharepoint for the Dow Intelligence Center or whatever it's called nowadays.  
But I don't think this kind of formation is there, it’s much more of a press release type of information.

But the documentation is being created, so it's maybe what I say is already wrong and I have not checked, but I think the IT infrastructure is very fuzzy or it’s behind a fog for me. I know it's there. But I don't know how to get there. I don't know how I can get the support and then it's always understaffed.

So anyway, even if I can reach them, who’s going to do the job for me that I need? And that's a problem. And I think Dow is aware of that and we're trying to find a solution, but there is not a single community of developers of machine learning models. We are across several organizations. We use different methodologies. So, it makes things a bit more complicated.

But I would not say that we are limited by the infrastructure, as having access to tools. It's much more a case of, how can I use this infrastructure, that is the limiting factor.

But I think that's the problem.

We need IR to develop whats called in manufacturing; most effective technologies [*MET*].  
So what is the most effective technologies to develop, or first, to get the data? What is the most effective technology to develop a model?

But you could have different most effective technologies that are proven and documented. I don't want one, we need several. The most effective technology to deploy my model is depending on the model type and the use of it. So, if those are properly documented and people are trained, then we will know how to, you know, we will be more efficient.

But right now, I've received more training on putting a helmet than on getting data… I know how to put on a helmet, but I don't know how to get the data. I’m still learning that by myself. And this has to be recurring training in the portfolio of training sets for the diamond learnings of each of the data scientist.

**Interviewer**  
And how about the hardware resources that are needed?

**Respondent**  
So yeah, hardware resources; first of all, we were limited with the Dow laptop. The default laptops are fully underpowered and as well full of IS gems [*laughter*], that prevent us to do a lot of things normally.

So, we need to rely more on servers and access to virtual machines and workstations. So, I have my own workstation where I have a bit more freedom, but again with IS package. It's still not ideal.

Ideally, I would like to have a lot of freedom in the sandbox, in a virtual machine that is a real-world sandbox workstation. But this workstation still needs to be connected to Internet because you always need to download something.

But I think we're not limited with the hardware. We have enough clusters, and those people are actually begging that we use their hardware more so they can request more.

And we have Azure and the cloud. maybe later we will have a contract where we only develop with Microsoft Azure. But again, all those tools for Microsoft have just alpha versions. So that's the problem as well. It will take time.

**Interviewer**  
And regarding the deployment of a model, once it's developed and trained, basically ready for use by the end users, do you have all the capabilities that you would need in order to deploy and hand over that model to your targeted community?

**Respondent**  
There are methods, but right now, just running scripts and then Jupiter notebooks and things like that are well supported. Now, if you want to go beyond that and develop a more complicated interfaces, for the user to use, then we have a lack of application developers at Dow.

It's a job on itself. So, the model deployment does not stop in saying OK, I have sent my machine learning models to a certain machine. But it stops when the users can act with or interact with the models.

And that's this last part, to building the, let's say the web app or building the applications for the end user that is really, we are really lagging.

So, there are methods and we are trying to. So, for example I'm running a proof of concept of deploying with MATLAB web app. Other peoples are using Streamlits to deploy models for web, Python base, Dash framework. But there's still not a most effective technology and there's still not a high level of expertise.

So, we will need to hire application engineers that develop those packages and maintain them.  
Because once you have distributed to your client, this guy will request you; OK, change the shape of the buttons or the color of the backgrounds and whatever. Or I want to have this type of inputs instead or …

I don't like this color of the graph. I prefer to have, I don't know, box plots and then this type of graph. So, you will have to make those changes continuously.

So, we need a group, we need a service that is an application engineer. And those guys are too few at Dow. And they're not really focusing on building the applications themselves. They're still playing around with which framework they have to use. They still have to pick up a framework.

So that's a bit of the problem.

Right now, the best method to package it is if you're able to push it to an Azure pipeline in Azure ML.

Or to have a text-based *[model]*. That's why I like simplification models. They are text-based models. So, you can just give it as a copy and paste in an Excel eventually and it works.

So yeah, it's 70% is there, but we're lacking the 30% to say that it's complete.

We can easily deploy. It's not yet there.

**Interviewer**  
OK, not in a very structured manner, so to say.

**Respondent**  
Well the problem I think, is really expertise and resources. The problem is that the data scientists at the moment is doing the data engineering work, the data deployment, I mean the model development, model deployment and end user interfaces.  
So this is too much for a single guy. So, we need to split that in groups. So we need to have a data engineering group, and then the data scientists can focus on really building the nicest models and make a sense of it.  
But then he needs to be seconded by somebody that knows how the model would be deployed and start building the interface and the concepts interface to the model. But that's not there.

**Interviewer**  
OK, makes sense.

**Interviewer**  
So and for my final question. Although I think you have already mentioned several gaps/challenges and changes that would be needed in the IT infrastructure and architecture. But besides what has already been mentioned in our call, can you think of anything else that would be needed at this point?

**Respondent**  
Well, first of all, we need certifications, you know, training certifications. You will motivate people to get training if they get external certifications.  
I mean, just having a Dow training is not, you know, it's not motivating. And it's not a proof of quality neither.

So, I would say that we need to have a budget to help people getting refresher courses.  
Getting evening courses and then being certified. There's a lot of online possibilities, but we need to have a budget for that.

The second would be that we need to have a group, internal groups that are highly specialized.

Whether it’s on the data engineering or on the application deployments and application building.

I think this group exists and the group specializes in building the pipeline and workflows on the models, I think this group is there. But we really liking on the data engineering and model deployment, but that's where I think we are lacking resources and expertise. So, we need to hire people and we need to…

People always believe that digital building digital solution will cost less and should cost much less, or should have a negligible cost compared to, let's say, real life development.

We need to accept the idea that building digital solution has a high cost as well at the start. It has a low cost in use, but it has a high cost to develop.

And we need to take it seriously. So, we need to have a good picture on that. I don't know what's the budget nowadays, but for example, I should not be told that, you know, €2000 for a MATLAB license is too much. Are you serious? I mean, if my work is not worth 2000 Europe per license then why do I work?

You should fire me because it costs more, much more than those €2000, you know. So that's the kind of things that needs to change.

**Interviewer**  
Alright, thank you so much. We’re through all of the questions, I'm going to stop the recording now.