# MIZUHO INDIA JAPAN STUDY CENTRE

## WEBINAR TRANSCRIPT

## Webinar Title

## DON'T WE HAVE CHATGPT? PROBLEMS AND CHALLENGES IN MACHINE LEARNING AND ROBOTICS

**DATE: 11th August 2023**

**SPEAKERS: Prof. Chintan Amrit**
**Dr. Floris Erich**

### ABOUT THE SPEAKERS

**Prof. Chintan Amrit:** He is an associate professor at the Department of Business Analytics at the University of Amsterdam. He holds a master's degree from the Indian Institute of Science, India, and did his Ph.D. in coordination in software development from a Netherlands university. His research interests include business intelligence through machine learning open-source development, and mining software repositories and he has applied all these analytical tools aligned with the UN's sustainable development goals. He has an extensive publication record with over 70 research articles published in his name and he serves as a department editor at IEEE at Transactions in Engineering Management. He is also a coordinating editor for the Information System Frontiers Journal and associate editor at Peer Computer Science Journal. He is an active participant in the academic community, and he regularly shares tracks at the European Conference on Information Systems and many other such conferences worldwide.

**Dr. Floris Erich:** He is a permanent researcher at the National Institute of Advanced Industrial Science and Technology, Japan. His work centers around bridging the gap between the virtual and physical worlds and he develops tools and techniques to model real-world conditions for verifying the correct behavior of robotic systems. He has contributed to various projects sponsored by organizations like the New Energy and Industrial Technology Development Organization as well as the Japan Science and Technology Agency. He has earned his PhD in Human Informatics from the University of Tsukuba. Dr. Floris's research focus was on reactive programming using procedural parameters for end-user development and operations of Robert's behavior control.

### ABOUT THE MODERATOR:

Saideep Rathnam is the Chief Operating Officer of Mizuho India Japan Study Centre, bringing a wealth of 47 years of industry and academic experience to the Centre. An alum of IIM Bangalore, from Hindustan Aeronautics Ltd. to British Aerospace, UK he has spent over 2 decades in the aeronautics industry and over 18 years in the automotive sector in various capacities including president of manufacturing excellence at Anand Automotive Ltd. He is also a Certified Chartered Management Accountant [CMA], UK. He wears many hats and has chaired Anand University, helping companies in the fields of management of change and innovation. Recently, he drove the Visionary Leaders for Manufacturing (VLFM) program as a Senior Advisory Committee Member of CII.

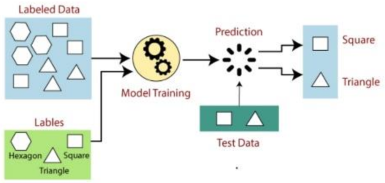
**TOPIC OF DISCUSSION:**

1) Addressing real-world challenges within a corporate business based on data-driven machine learning models having their limitations with reference to the data both in terms of quality and quantitative challenges.
2) Ideas about the nuances and nuanced distinctions between practical machine learning or other models and their deep neural network counterparts.
3) Discussion on comprehensively exploring the multifaceted landscape of data-related challenges and unraveling the inherent problems that arise while dealing with such data issues. Additionally, they discuss potential strategies to effectively overcome these challenges within the context of Machine Learning and creating models.
4) Highlight Domain Adaptation and its ability to address data-related obstacles and facilitate cross-domain learning and adaptation.
5) Dr. Floris Erich focused on the application of Generative Adversarial Networks. GANs can ingeniously create the relevant data thereby enhancing the effectiveness of these models even when faced with limited or insufficient data. Valuable insights into the challenges forced by data limitation in Machine Learning and innovative strategies for mitigating these challenges and the potential impact of GANs in the context of robotics as well as predictive modeling.

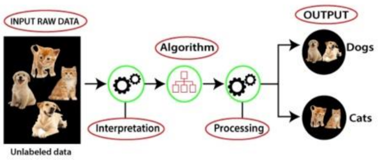
**Prof. Chintan Amrit talked about machine learning:**

**Supervised learning** basically deals with labeled data so you want to know which part of the data you're dealing with trying to predict whether the shape is a square, hexagon, or rectangular must be labeled assuming that this is a complex task could be anything like you're trying to predict if a particular patient has cancer or if a criminal is really a criminal based on data about the person.

Here he explains how they train a particular model by trying to predict the shape of the object and have labels in green such as hexagon, triangle, and square, and what we basically do is train a particular machine learning model by feeding it with label data. Eg: The label tries to learn that 6 sides are hexagons and 3 sides are a triangle and 4 sides are squares and then they give it test data and then see how well it has performed. So, if it's able to predict it correctly then it's doing well. Then it shows that we have learned something from the data and the supervised learning algorithm works.



Supervised Learning

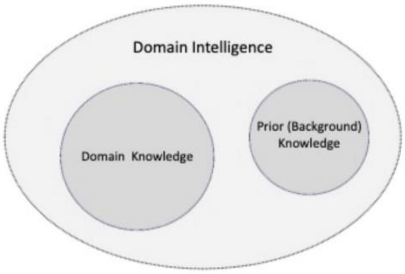


Unsupervised Learning

U**nsupervised learning** deals with more complex data which is not easy to even perhaps even label but in the case of humans it's easy to label slightly more complex with more features or more aspects of the data to explain humans. If we have all the features and it is small and simple, then we have supervised learning but also something we need the algorithm to learn by itself

by seeing the data to understand patterns by itself and then predict what kind of object it is. Then there is unsupervised learning where we give unlabeled data so we don't know which is a cat or which is a dog then it tries to understand the way the pixels of the dog are arranged and that it is really different from the cat. Then they learn from the output that this is a bunch of dogs or a bunch of cats because the pixels are arranged in a different way. The focus of this talk is more on why buying a domain is such a crucial thing in data analysis.

**Domain Intelligence**: It's a concept introduced by a Professor in Australia. It includes two important aspects of the data it has some domain knowledge and it has some prior knowledge domain knowledge.



1) **Domain knowledge**: Domain knowledge gives you some insight into what exactly is happening in that particular case.

2) **Prior knowledge**: The mathematical knowledge behind applying the machine learning model to that particular domain so it can be things like we know that this problem, this is the best algorithm that works.

**AIS Data:** AIS is data that all ships barges and boats that want to go to a harbor need to have and deal with. It’s like a transponder put on the ship with tracks its current location. It's like a GPS for the ship. Then they need a place to store the data where they can build their own server using a simple antenna. These can help to understand the current location of the ship, the size of the ship, the speed of the ship, direction of travel. These AIS data come with their own positive and negative points because it's not always accurate.

Positives: It prevents from colliding and even predicts the time of arrival at the harbor.

Negatives: Equipment quality, the server, the receiver, external conditions, and human inputs, lead to incomplete tracks. Sometimes it could be an overload of data.

In some cases, the size of the data also matters, there are two types of data “big data” and “small data.”

With big data, we can extract a lot of information learn from interesting matters, and work on new concepts like “deep planning”. But with small data, there is actually a shortage of information and a lack of accuracy. Hence when it comes to small data, domain knowledge is very critical. The problem in this industry is that there is not much “big data” available here.

Example: “Logistics Supply chain in the arrival time of trucks for distribution purpose” which is important for distribution centers when they are dealing with logistics supply chain, they want to know when the ships would come so that the trucks would be ready so that cargo is loaded and unloaded quickly and most efficiently. In this particular case, he is discussing this particular domain of arrival time of trucks but this could be potentially used it even for example in robotics and compare areas where there are data challenges and it’s a complex domain where there are people managing trucks, people managing ships and also people managing the whole process in the distribution center. He also mentions about the complexities that follow with human interactions and interventions compared to having robots interacting with each other which can be kind of preprogrammed. though you can indeed bring in some complexities there but unmatched to human complexities.

So, whenever we see a predictive model in a complex domain there is a chance, we can learn from that. They observed in their case that some of the truck drivers were not incentivized to arrive on time. These are things you understand when there is a human connection to understand what’s happening in their domain. These insights could be used in a new model where some aspects of the domain are missing. In the new model, they have no information on traffic at this particular granularity so in the first model they had traffic at an hourly level, and second model they have information only at a daily level. This difference needs to be taken into account. So, it’s possible to get traffic at an hourly level then the cross-domain analysis will tell us that that’s the way to go to improve the model. So hourly level data would make much of a difference. The same comparison can be made for the weather as an element. The same way of you don’t have weather data on a daily level will make much difference. So, this kind of understanding that we can gain from understanding or analyzing one particular case can be leveraged when we are trying to create a predictive model for the same domain which should be the arrival time of tucks but in a different setting. For creating a predictor model in any complex domain where it's not very clear what is the effect of that particular feature in creating the predictive model that is where he thinks this cross-domain analysis is most effective.

At which point the domain information is most important? During the data selection process, data pre-processing, data transformation, or data mining? Also, do we even use domain knowledge for validation? Domain knowledge works well with an exciting predictive model that has been developed.

Cross-domain learning is where we learn from one predictive model and apply it to another.

**Key insight:** Domain understating is extremely important and the type of machine learning algorithm not just for analysis but also the size of the data along with the reliability of the data. Also, comparing predictive models across domains.

**Dr. Floris Erich's talk on Robotics:**

What makes robots intelligent?

In the past century, we have seen success in industrial robotics. If we ask an Industrial robot to make a sandwich with a pre-programmed setting can make 1000 sandwiches per hour but an intelligent robot at home will face different issues. So, the robot can ask multiple questions related to making a sandwich: what is a sandwich, what kind of sandwich the user wants him to make, what goes into a sandwich and how can we compose the ingredients of the sandwich, where should I make the sandwich, many of these questions require the user to have a dialogue with the robot? The robot needs to understand words, where things are located and which ingredients we have, how to navigate through the house, and how to handle dangerous objects like knives. We cannot pre-program all the tasks because there are so many different tasks to be done not everything can be pre-programmed.

The speaker talks about the societal issue of aging. A leading issue for Japan yeah our working population is expected to decline from 87 million to 60 million so that's a massive decline in the working population and could lead to a labor shortage and a lot of difficulties in passing on the technological know-how and the center is focused on improving the situation and the way they do this is by increasing the productivity per person and also increasing the number of production workers and also effectively transferring know-how from person to person and doing our sensors that give different test beds to facilitate the adoption of technologies so we focus on factories logistics and drug discovery these are industries which are labor intensive so there can be a big benefit of introducing robotics.

Talks about, what the future holds. New technology like Chat GPT allows us to control robots through voice and dialogues and also new information model allows robots to perceive the world to understand the relationship between vision and language and also Chat GPT allows to robots to reason about complex tasks and the new development.

## QUESTION & ANSWER SESSION HIGHLIGHTS:

**Q**: What are the primary challenges in managing data to create an impactful AI model and what are the strategies to overcome this?

**Ans:** The main challenge is to understand the key insights from the domain. Many companies generally do is gather all relevant important data that they need and manage it in a suitable way. This could mean they have multiple databases, and they could store them in one large data warehouse which could be used for data analysis but this needs to be a continuous action without being a burden to the company. The strategies that would be used to overcome the same are Cross-domain analysis (learning from an existing predictive model and gaining insights and applying them to the current problem) and complex aware data analytic methods (getting insights from the relevant experts).

**Q**: Please give an example of Cross Domain learning on how it has helped in outcomes to better.

**Ans**: Cross-domain idea comes from the number of extensively made machine learning models that already exist, so the idea is to leverage that while creating a new model. Mr. Amrit gives a real-world example of a "Logistics Supply chain in the arrival time of trucks for distribution purposes" explained above.

**Q:** How does one really work out which particular model is the best one for a particular data set?

**Ans:** The large challenge in machine learning is that we have a data set to train on and these data are actually just a representation of the real data we want to apply that model on. So, he thinks we can build better experiences that are completely optimized for training data. We don't want them to be optimized for training data to be optimized for the whole data set. The whole thing in machine learning is about bias and variance. How biased is your model and how much variance it has? It is related to the concept of overtraining or undertraining your model. So, if it is over-trained then it's more biased and that's why we can't be sure because we use a part of the data for training, testing, and validation especially in deep neural networks. It is important to use a simple baseline. Many companies don't really like complex deep neural nets, it is not about only making the net but also maintaining it even if it is heavy on serve space. They prefer models that are more understandable with fewer layers that can handle in a better way. Sometimes a simple model is better than a complex model.

**Q:** How do we know the adequate training time for a robot or an application?

**Ans:** When you train a model, and you can store previous versions of the model basically you can train your model until it starts to overfit the data so it starts to perform better on the training data and on the validation or test data and then you just take the previous model which you saved and use that as your final model.

## CONCLUSION:

It has given us a brief on all the things that are happening as we discussed as you explained as a disclaimer in the beginning Chat GPT could be an overkill but it does play on large data sets and Chat GPT does use large data sets as we use more practical day to day small data sets. Data sets are limited but in these limited data sets, noisy data can be addressed and that is very refreshing to hear.