Real-Time Captioning

Note: This is not a transcript.

April 29, 2020

Artificial Intelligence Applications in Food and Cosmetic Safety

10:00 AM– 10:10 AM	Welcome Amy P. Abernethy, Principal Deputy Commissioner, Acting Chief Information Officer, US FDA, College Park, MD Overview and Speaker Introductions Ernest K. Kwegyir-Afful, US FDA CFSAN, College Park, MD
10:10 AM– 10:45 AM	From Artificial to Real: AI Stories in Government Steve Bennett, SAS Institute, Cary, NC
10:45 AM– 11:20 AM	Artificial Intelligence: Introduction, Applications, and an Overview of the Colloquium Jim Riviere, 1DATA Consortium, Institute of Computational Comparative Medicine, Raleigh, NC
11:20 AM– 11:55 AM	AI Technologies for Future Factory Cleaning Nicholas Watson, University of Nottingham, Nottingham, UK
11:55 AM– 12:30 PM	Using AI to Extend QSAR Models Chaoyang (Joe) Zhang, University of Southern Mississippi, Hattiesburg, MS
12:30 PM– 1:05 PM	Using Machine Learning for Cosmetics and Cosmetic Ingredients Tim Allen, University of Cambridge, Cambridge, UK
1:05 PM– 2:00 PM	Roundtable Discussion Moderator: Jim Riviere, 1DATA Consortium All speakers Additional Panelist: Ernest K. Kwegyir-Afful, US FDA CFSAN

Ernest K. Kwegyir-Afful: Good morning. Good afternoon depending on your location. Welcome to the 22nd session of the SOT FDA colloquia. Today's session is intended for the food and cosmetics safety and I data scientist with the Office of Food Safety and

Applied Nutrition at FDA and today we are welcome you to this session. The Commissioner options and the agency's day-to-day functions and direct special and high quality agencies with the application of drugs medical application to micro entry and oncologist and palliative medicine and we recognize clinical data expert and clinical researchers and expertise including New World evidence clinical trials health services research and patient support outcomes in clinical [inaudible] before coming to FDA we as the chief medical officer two scientific officers and the team of scientific oncologists clinical operations and data science teams and contributed to the overall strategic vision of the company, including directing their research [inaudible] new world evidence. Prior to that, Dr. Abernethy was professor of medicine at Duke University School of Medicine, where she ran the Center for [inaudible] Healthcare in Deep Clinical Research [inaudible] Institute in the Duke Council Research Program at Duke Council Research Institute. At Duke, she pioneered the development of technology platforms [inaudible] advancements in the care of people with cancer and other [inaudible] illnesses. Dr. Abernethy was formerly an appointed member of the National Academy of Medicine's National Council Policy Forum and an elected member of the American Academy of Clinical Investigation and Past President of the American Academy of Hospice and Palliative Medicine. Dr. Abernethy received her MD at Duke University, where she also did her internal medicine residency, served as chief resident, and completed her [inaudible] fellowship. She received her PhD from [inaudible] University in [inaudible] evidence-based medicine, clinical informatics, and her bachelor's degree [inaudible]. Welcome, Dr. Abernethy. Could you unmute Dr. Abernethy, please? Next slide?

Welcome
Amy P. Abernethy, Principal Deputy Commissioner, Acting Chief Information Officer, US FDA, College Park, MD

Terrific. Thank you very much for that wonderful and honorable introduction. I am very delighted to be here with you today. As mentioned, I am the Principal Deputy Commissioner here at FDA and relevant for today's conversation, I am also the Acting Chief Information Officer. I welcome the opportunity to provide opening remarks for many reasons not the least of which that artificial intelligence, AI, and machine learning have been an interest of mine and before coming to FDA my focus was on AI machine learning is one of the health care ecosystem and I asked how we could build data technology to improve clinical care delivery while also improving research in the clinical environment.

These can be applied to the food ecosystem and how do we build the data and technology solutions to leverage information and improve decision-making during the flow of products to ensure safety from farm to the consumer. In FDA's vision for safeguarding the nation's food supply digital data and its analysis including artificial intelligence plays a key role but to do this we need the infrastructure, digital data as well as the development of programs and models that can inform decision-making. We need to be able to make sense of the data.

This is one of the driving forces behind our technology modernization effort. We need to have the right infrastructure to receive data in a format that we can use at FDA at the

right time to inform regulatory decisions and ultimately the way that we use information informs how information is going to be used all the way across industry. In September we announced the technology modernization action plan also known as the T map and that provides the technological foundation for the development of FDA's ongoing strategy around data itself a strategy for the stewardship, security, quality control, analysis, and real-time use of data and this is going to accelerate the path to better and more efficient use of data including food safety data. This data of the future can be used to predict events like foodborne outbreaks and will accelerate responsiveness and focus on activities in order to make sure we best protect American consumers.

Domestically we can see how such data might be used to inform food production practices, help us predict shortages or bottlenecks, or quickly identify all the food products associated with contaminated ingredients. Digital data and AI can be put to work to better protect American consumer from adulterated or misbranded imported foods and FDA's food program is considering a number of different applications of AI and other emerging technologies to improve the agency's oversight of our country's food supply.

While these projects were in the works before the COVID-19 pandemic this experience during the pandemic is emphasizing for us the need to accelerate machine enabled processes. For instance we have been talking for some time about the importance of fostering greater traceability our food supply and the use of new and evolving digital technologies may play a pivotal role in tracing the origin of contaminated food to its source and can do so in minutes or even seconds instead of days or weeks. Moreover, this trace-back could be conducted remotely and more quickly than sending our investigators into a facility to evaluate paper records in person. These days we can see how important it is to have new and novel techniques and it is visible to us in the news every day.

AI and machine learning could enable us to better consider many hundreds of factors that could contribute to a given outbreak and incorporate these contribute factors into analytical tools to better produce predict future outbreaks. In light of COVID-19 we are now realizing that the transparency this technology provides could have been outside of an outbreak event. It could be helpful in dealing with this supply-chain issue resulting from a national emergency including efforts to prevent food waste by redirecting certain supplies when necessary.

Another exciting project is the use of AI and ML to improve FDA's import screening. As part of FDA's new air of smarter food safety, the FDA is exploring the potential for artificial intelligence to improve screening of imported foods before it's allowed into the United States for sale to our consumers. FDA uses an import screening tool called PREDICT for the Predictive Risk Based Evaluation for Dynamic Import Compliance Targeting. PREDICT helps FDA employees feed the review of import while targeting products based on risk and the tool is designed to automatically search and analyze large amounts of current and historical data. PREDICT helps FDA personnel identify patterns, flag issues, and determine potential risks of new shipments in real time. Increased number of automated decisions gives human reviewers more time to focus on higher risk entries. Therefore, PREDICT is a valuable tool to ensuring the safety of imported foods but right now it's predominantly a rules-based engine and we are continuing to evaluate ways to make it smarter and even more efficient. In a proof of

concept project PREDICT will serve as a testing comparator as FDA developed a prototype machine learning model where we are going to identify imported seafood shipments that are more likely to be violative. We will compare the PREDICT model to this new machine learning model and expect that the machine learning model will improve the sensitivity, specificity, and predictive value of the selection model for import entry review.

Americans expect FDA to make decisions based on rigorous science-based information. And at FDA were going to continue to conduct the public health mission, our public health mission, as it's been entrusted to us, for many decades but we're also going to carefully use newly designed technologies like PREDICT to help us operate more efficiently and effectively.

On that note I would like to thank you for your time, and I look forward to hearing more about your work in in AI application outside of the FDA. Thanks.

Overview and Speaker Introductions
Ernest K. Kwegyir-Afful, US FDA CFSAN, College Park, MD

Thank you, Dr. Abernethy. Next slide please. So a quick review of the FDA SOT colloquium series is a partnership between U.S. FDA Center for Food Safety and Applied Nutrition and the Society of Toxicology and are this is the safety of the colloquia and it was put together to simulate a dialogue among leading toxicology experts on future oriented toxicological science relevant to student food ingredient safety assessment and is meant to be a forum to discuss the latest toxicological science in the context of food chemical safety and not a form for soliciting regulatory advice or discussing food or food ingredient revelatory issues.

The colloquium series is open to the public to attend usually in person or via webcast and in this situation we are only via webcast and it's a global audience from all employment sectors. For example, in the 2018, 2019 year, we had participants from 34 countries. Today, we are in for a very exciting time. We have talks starting with Dr. Bennett and he will give a general overview of AI and the stories in government and then we will have Dr. Riviere who is the Chair of this colloquium to give an introduction and talk about applications and an overview of the colloquium. Next Dr. Watson will give you AI technologies for future factory cleaning and talking about how we are using AI to extend the QSAR models. Dr. Allen from the University of Cambridge will talk about using machine learning for cosmetics and cosmetic ingredients. The idea is to hold all your questions and you can send them in and we will hold them until the discussion, and we will get your questions to the speakers.

This is a note to the organizing committee for this colloquium. So, our first talk is by Dr. Bennett and he is [inaudible] at SAS, and he is passionate [inaudible]. Dr. Bennett?

From Artificial to Real: AI Stories in Government
Steve Bennett, SAS Institute, Cary, NC

Thank you very much, Ernest. Let me do a quick confirmation of sound. Can you hear me?

Kwegyir-Afful: Yes.

Bennett: Fantastic. I will let everybody take a look at this and then we will get started. Next slide.

So, the year was 1943 and Allied Command of World War II had some very hard decisions to make. Aircrews were flying missions over Germany with increasing frequency in the roof suffering these heavy aircraft losses due to German fighters and ground fire. Given the high risk of those missions the U.S. was desperate at times to find a way to increase the survivability of the aircraft they want to reduce the likelihood of catastrophic damage so military decision-makers wanted to know wanted to know where they could add armor to the aircraft to protect them they wanted to know how much to add and where to put the armor. The challenge was too little armor makes the plane vulnerable and too much makes them slow less maneuverable and decreases the range. Somatic complaints that did survive the mission were returning covered in bullet holes, but those bullet holes were not evenly distributed across the aircraft. When military officers looked at the battle damage data represented by the red dots in this diagram and by the data and the table, they so that more damage per square foot in the fuselage and wings services and they saw much less in engine. Officers saw an opportunity for efficiency, and they believed that they could get effective protection on the plane with less armor plane with less armor if they concentrated the armor on the places with the red dots. Where the returning planes had been hit the most. That [inaudible] was completely wrong. Next slide, please.

We did talk a little about this man Abraham Wald, he was a Hungarian born mathematician and professor of statistics at Columbia and part of the team that the government had created to help the Allied war effort make better decisions and let's take another look at the data to find out what he discovered. Looking at that same battle damage data he concluded that the armor should not go where the bullet holes are. It should go where the bullet holes are not on the engines. So, it was and asked why the bullet holes were not evenly distributed across the aircraft he realized that the planes that returned after the mission the one that has survived no matter how damaged they were it was the planes that did not make it back from their missions the ones that were shot down that must have suffered damage in critical areas and it was those critical areas those are the places that were undamaged in the surviving aircraft so after at his recommendation the decision was changed armor was added not where they saw damage of the returning aircraft of the red dots but where they didn't.

It was more than just to put armor on the planes they needed help with all kinds of decisions and from what the aircraft they could get close to the Germans and Marines to how to determine the training hours for flight crews and everything in between. So, the field and now we cooperation research was born, this is the definition from the late 1940s, a scientific method of providing executive departments with a quantitative basis for decisions regarding the operations under their control. This was the very first time

that quantitative information was brought together like this to help leaders make decisions better faster and cheaper and it was born in government. And today we call this analytics.

If you look at the definition of analytics the scientific process of transforming data into insights for decision-making it looks very similar and what started in government as operations research 70 years ago now helps organizations solve all kinds of heart problems and help us make decisions better across every industry.

So, again, I'm Steve Bennett, I lead the Global Government Practice at SAS and I want to talk to you about artificial intelligence and its lineage tracing back all the way to that first use of quantitative data inside of government. Today, I want to tell some stories. I want to set up our day by just probably looking across all different applications in government where artificial intelligence and machine learning can make a little difference. Let's start with a little bit of history lesson. Artificial intelligence was born not long after World War II, were in 1956 in the first and the term was used by Don McCarthy artificial intelligence at Dartmouth. The concept around artificial intelligence and the definition to the left which is not too far off of how we would describe it today really persisted for the last 60 or 70 years. And machine learning is not all that you are. 1959 is the first time that term was defined by Arthur Samuel and looking at the definition it is not unlike the definition we would use today when we talk about artificial intelligence.

The question you might have is, why are we hearing all of this buzz around artificial intelligence and machine learning now? We cannot throw a rock in a conference about without hearing artificial intelligence and machine learning, but it's been around for 60 or 70 years, why is it new? The reason it's new is this gentleman he was a South Korean grand champion in the game of go and if you don't know about this game it is the most complex human game in existence. What about 172nd power combinations and for comparison there are quote unquote only 10 to the 82nd power of atoms in the universe so very highly complex game and not to get into the computer science but the way that we have thought about teaching computers to play games previously, think about IBM and deep blue, for chess the concept was to teach the computer the basic rules of the games and then you throw it up compositional power added to look far enough ahead in the game to be able to see the feet humans.

You cannot do that with the game of as complex as go. The thing that Google did with deep mind as they created a system that turn that problem upside down. Rather than trying to teach rather than give the computer the rules and teach it to play you just give the computer all a bunch of data about what works and what does not let us figure out the rules and that's really the difference between ways of approaching problems in artificial intelligence and machine learning: machine learning turns the problem upside down and you throw a bunch of data into the system and let the learning algorithm figure out the best rules to play the game rather than just starting the other way around. You start with data instead of the human written goals. This broke everything open in 2016 with a form of machine learning that for the first time was able to deceive the seat the seat a human player which was thought never to be possible and this broke discussions around machine learning and deep learning out into the forefront.

Two things make it possible today that we did not have 60 years ago with Don McCarthy and Arthur Samuel and those are the broad availability of lots and lots of data to train and empower these machine learning tools, artificial intelligence tools, and the increasing power that we have computationally to crunch a lot of these complex algorithms weekly to those two things make it possible to do now at large-scale what John regarding Arthur Samuel were able to do 60 or 70 years ago when they wrote down the theory.

So, everybody may have their operating definition of artificial intelligence but here's what I would use for the remainder of my remarks. Artificial intelligence is a science of training systems to emulate specific human tasks through learning and automation. I can spend the rest of my time talking about why we are not talking about general human AI with the thinking robots, but we are talking about specific human tasks through learning and automation. A couple of things that machine learning are good at. Other than just beating humans. Machine learning is very good at categorizing or cataloguing things that are like. That's very, very important in a wide range of problems. A second thing it's good for the thing likely outcomes based on identified patterns. It is very good at identifying previously unknown patterns and relationships and one thing to say I got a pattern and what the likely outcome based on that and another thing to say here's a bunch of data, showing the best pattern and segment that data into something I can use and finally maximizing a reward or goal by evaluating the result.

The discussion on things like PREDICT and list scoring would fall into that category, so these are four broad categories of algorithms and machine learning are very good at if you got enough data to train the models. It turns out these four problems are all over the place in government here in the United States and around the world. What I want to do as we advanced the slide that I want to do the next few minutes is talk about opportunities and challenges for AI in government.

Just a couple of notes here for AI in government. Just a couple of notes here AI and machine learning are increasingly a priority for government around the world. The United Arab Emirates pointed out the were the first country to appoint a cabinet level secretary of artificial intelligence and in Canada the first country to release a national AI strategy and now we have one in United States out of the office of scientific policy technology so lots of energy around AI not only how to use it but how to use it ethically so 75% of government managers talk about AI helping them keep up and it's an incredible opportunity because of the type of data that government tends to have are structured text data, voice video, a lot of that is very hard to leverage without AI, so that's just a little bit of the motivation why AI is so important for government applications.

Let's talk opportunities. I said I will tell you some stories so when I tell you some things that AI can help us to, I will go very quickly each of these can be a whole talk on its own to begin each one of these projects. First let's talk about an anomaly detection, how can you find things you care about? Often, the government wants to do is find the thing that sticks out, show me the benefits claimed is likely fraudulent, show me the tax return that likely represents identity theft. At a border checkpoint, show the person based on the data I'm looking at that needs a secondary screening at an airport or the Customs and Border Protection checkpoint on the northern border. So, anomaly detection, the ability to find the thing that sticks out so you can take some investigative action or further

action in some other way that the critical area that's important for government and it's an important area where more machine learning shines. Resource optimization. Any problem that has limited cost limited materials, limited personnel or time, which that's pretty much every government problem. If you have limited resources, you can use techniques and resource optimization to make sure you're getting the most given the resource constraints you have.

And a great story out of Boston Public School where we have the picture of the school bus, they wanted to know how can we save money and increase the safety of children walking to bus stops, so they wanted to know can I optimize the bus routes so I can minimize fuel consumption but also want to have a constraint where I minimize the walking time from students from their homes to the bus stop because there's accidents risks as these children are walking across streets and the likes. Machine learning was used to optimize route planning to save money for Boston Public Schools and to help keep children safer by minimizing the walk time between where they live and where they had to catch the bus. Some research optimization can sound theoretical but has very practical outcomes when used carefully in a government problem like this for Boston Public Schools.

Very different application is inspections and maintenance. Government owns a lot of stuff whether you're at the post office and you have [inaudible] or at the airport with planes and anywhere you have a crib and you can use things like machine learning and artificial intelligence to help you and where should I target my inspection teams, how should I prioritize maintenance, and at the other end of the spectrum the complexity for the maintenance, can I use machine learning and you can we can use machine learning to help these machines tell us before they're going to break by looking at data that we understand when machine is healthy and we can start to detect the anomalies when the machine starts acting in ways that are not healthy.

Let's put some handles on that, some real examples. U.S. Air Force uses this today. So, Lockheed Martin has created both the F 35 as well as the C-130, two of the most advanced airplanes in the world, and this predictive maintenance using machine learning to help the text when these things are going to break before they break, and this goes thus two things. Keeps the equipment more operational useful for the mission a high percentage of the time, and also make sure that United States are airmen Marine and sailors come home safe at a higher probability to their families when they're done with the mission. Corrective maintenance any part of the government that has got equipment machine learning can have a huge impact on saving money and increasing the operational upkeep of the equipment.

For another travel example on the other side of the world another local example you strongly in state Western Australia there using machine learning similarly for safety. They are looking at intersections, have a whole bunch of data on where traffic accidents occur in occur in all the intersections in Western Australia and they wanted to know how can we figure out where we need to invest resources to make these things safer so they use machine learning applied to action data to understand where the best places to place resources and keep Australians safe as they travel.

Let me give you another example very different when you might not expect. This one is around case management so any government entity that has investigate cases,

machine learning can have a great impact for that as well. You can use it to prioritize work order, you should pursue cases let's say you're an officer of Inspector General most the time those entities are understaffed and overworked so what order should you pursue investigative cases by the likelihood of getting a good outcome, by the level of damage or cost, how should you prioritize those cases and how should you train.

Let me give you a real example. This story's from the country in the Middle East and I cannot tell you which one but they had a problem in their judicial system where they wanted to reduce the backlog in their judicial system like many countries and they saw that a number of cases when they finally made it to court they were closing Monday let's say a DUI or something very simple. And they wanted to know can we put up front which cases are going to close in one day or less and then we want to be able to take those and funnel those to a one-day court so we can clear out that logjam for the cases that need more time. Using a very simple machine learning approach they figured out a simple predictor and you can see the colored dots on the slide, I think, the focus on the lower left in the predictor shows that when the predictor said that we think this case is going to close in less than one day, 93.6% of the time it closed in less than one day. This is very simple machine learning algorithm.

For government we want to not just look at where we get it right. This is really important for applying machine learning and government. We don't want to look only where the machine learning algorithm does a great job and that pattern; we want to look at where we get it wrong. If you look at the data on the upper left the predictor when we predict the cases would close in less than one day, 4% of the time discuses that more than three months and why is that? Maybe there's something wrong with the algorithm. Possibly. You can always make your computer science better but sometimes these cases reveal problems in the underlying system maybe all of those cases were from one precinct or under a single judge so when you get it wrong, it's action opportunity to highlight and get the learnings out of the underlying systems. Maybe there's some things you can learn from where the predictor got it wrong beyond just how the algorithm is performing.

Let's talk about public health and healthcare. Machine learning and artificial intelligence can have a huge impact in everything from surveillance to diagnostics to outcome optimization and if you read in my bio at some point, my last job instead of government about four years ago was I was a director of national bio surveillance in the Department of Homeland Security as a senior executive there we let some of the federal response for things like Ebola and our last coronavirus outbreak from the Middle East MERS and we used machine learning and I want to get the COVID-19 but we use machine learning and Congress used ask us all the time can you look at data and look figure out whether or not we are seeing an anomalous level of health problems or impacts inside United States. And just as an aside if any of you have ever done any large data projects looking at Twitter, the more you look at Twitter, you lose your faith in humanity.

Setting that aside, so we did that and looked at Twitter data and for reasons we will not get into statistical methods using keyword search were insufficient to get a good signal for anomalies in social media data, so we did a project with a simple random machine learning approach and we got to where we would and that pilot we can pick up influenza-like symptoms in Twitter that 85% accuracy and you locate them so that was

a little bit of an initial flu surveillance tool for us and homeland security couple of years ago.

Machine learning is a practical use all over the place in public health and healthcare. I did want to talk about COVID-19. I will stop here for a minute. There's an awful lot of value in these sorts of technologies for helping in the fight against something like COVID-19. Prediction, there is groups like the global project eco-health alliance they're looking at ways to use artificial intelligence to predict hotspots around the world based on where humans are interacting with animal ecology so using machine learning we might be better prepared to start to look for and plan on hotspots geographically where these things might pop up and detection looking for anomalies in health data travel data you can use machine learning to more quickly detect these events as they appear and mitigation treatment.

I think it was last year the first AI develop vaccine for [inaudible] and a lot of the discovery of small molecules for treatment is very much trial and error and we are using seeing a huge exposure of AI applications in speeding up and making faster dose trial and error processes for the discovery of potentially active molecules. And then prevention and warning I mentioned about the hotspots work and some of the things you can do on the policy side with these technologies is you can a dry run different what-if scenarios and see what your policy impact might be when you're not in the middle of the places, so these are four large ways that we are seeing that artificial intelligence and machine learning can have a big impact on something like COVID-19 before the next one and right in the thick of things and mitigation and treatment.

So, just a little bit there on the work going on around the world and government applying these technologies on COVID-19. I'm actually going to skip this one. So, just on the optimization site this is something that's gone on not only France but for Germany in the thick of COVID-19 the ability to optimize using the thick of COVID-19 the ability to optimize using the best technology optimize critical resources like things like ICU beds in hospitals. So, France has done this across the entire hospital network in the German Ministry of Health has done this and I know it's been discussed at the state level in a number of places in the United States as also little practical example of where these techniques can be used to optimize resources.

So, just talking about citizen services can we advance must advance a couple. So, let me give you one good example around citizen services. The Ministry of Social Development in New Zealand, they have the largest expenditure of any agency in New Zealand. They spent about $22 billion for welfare every year and they saw that about a third of the people coming on benefits entered the system as children under the age of 18. They wanted to know how can we changed the trajectory of children, families and poverty New Zealand and they did something that was very creative, just as an aside, I have for my kids and when my wife was pregnant with our second child we had not told anybody yet and yet in the mail target starts sending me all of these advertisements and my wife all of these advertisements for baby formula and maternity clothes and it was kind of creepy and we were wondering how they knew that we were pregnant we had not told anybody. My wife had a Target Red Card and probably bought something on Target, so they probably set only pregnant ladies by that, so they started sending her advertisements, so they do that all the time; it's called customer intelligence.

So, the team in New Zealand said why can't we apply that same approach to citizens we know a lot about customers and retail what if we said the government knows an awful lot about citizens from citizens from our data, could we figure out what citizens need and tailor the right benefits and right services to the citizens based on what we know about them. Are in a single mom, what's her educational background, where they work and instead of a one-size-fits-all approach to welfare, let's tailor the benefits directly to those families and they saw huge impact, they saw forecasted savings of $1 billion over four years and they saw an immediate reduction in unemployment for single parents of 10%, so whatever side of the aisle you're on, if you like smaller government or better benefits, this is a win-win when you apply technology to the data the government already owns.

So, that's a lot of opportunities and I tried to skim the surface to make sure we paint a broad picture of all the applications across government where artificial intelligence and machine learning can have a huge impact, but it is not all roses as we know particularly as you are thinking about AI in government. There are some key challenges as well.

I'm going to close my last five or six minutes highlighting a couple of categories of challenges the first one is technical organizational challenges and there's three of these, data readiness, the skills gap in government and some cultural realities. Let's go to the next slide.

The first is the data readiness. All those great things I just told you about require a lot of data to make them work, so if we talk to a government agency in the say I want to talk let machine learning the problem is well-suited for machine learning and we say great, where is your data? And they say here it is on a scrap of paper in my pocket, that's not ready for machine learning, so you have to have enough data to reveal to train these models and that it has got to be in good shape so you can use these models. That's something to keep in mind these machine learning approaches are very data hungry.

The second challenge is a skills gap we see across government. We have a lot of vendors who will sell artificial intelligence to government and I worked in government and we procured AI technology windows and homeland security, but it is not good enough for the vendor to have the know-how to make these things work. The reason for that is that sometimes these models tell you something that does not make any sense and you have to have enough technical expertise on the government side to know when to say, hey, that's garbage, I believe that and I'm not going to ask on that in my agency because I think that's an error in the model. We have to have the right skills in government to be able to oversee and scrutinize these models not just sort of trust them and think the best things since sliced bread. We want to have that scrutiny and there is no replacement for good government staff having that role. Cultural realities well, I'm sure this never happens in the FDA or in my organization but sometimes you have the right data, you got the right model, you have the right problem but sometimes the agency might just say we just don't want it. We just don't make decisions based on the sorts of models. That's something that's important to know as well; you can do all the technology right, get all your data right, get your skills right, inside the government everything right, but without an appetite from senior leadership that can take the results from those things and use them, it's just wasting taxpayer money. So, think about projects in government. This is an important one and if an agency is not ready to use

results from these sorts of tools in this decision-making, it is best to wait even if all the technology lines up so just something to think about.

I promise you two categories of problems and talked about the technology and accommodation problem but there's another set of problems as well, particularly in government, and I mentioned retail use of analytics, your Amazon, wherever you buy a pair a pair of jeans, if you like that pair of jeans shows up on every website you look at. That's because retailers like Amazon and others are using machine learning to figure out what they think you want to buy, and they are showing it to you, but if Amazon gets that wrong, is to show you a pair of jeans you think is completely ugly and you would never buy in 1 million years, what is the harm? Where there is little harm, Amazon does not make any money from you say might want to improve their model but there's really no harm.

But what if machine learning model is used to impact a regulation or machine learning model is used to deny or delay a benefit to a citizen? Or extreme cases what if machine learning is used to deprive somebody the liberty in a policing context? When these models are used in government, the impact of something going wrong is a lot higher than in retail, right, so we want to talk about those challenges as well.

In my remaining four minutes I will hit these, and these are large things that we should and could be discussing in government for whole colloquium just on these, but we will be remiss if you do not at least raise them. Those are the legal, ethical, and societal challenges around artificial intelligence and there are three that will I mention: geopolitical challenges, ethics and values challenges, and the legal challenges.

Let's start with the geopolitical challenges. There is a difference between could you do something, and should you do something. That is something that is popping up the needs technologies like never before. The technology enables us to watch was going on in places like China where there is a near ubiquitous facial recognition tracking, and you could do those things but should you? If you watch the science around this, around facial recognition, about three years ago a team of scientists were able to produce a deep learning based facial recognition model that was 94% accurate and you might realize that's notable because the human race is only 93.6% accurate at recognizing faces and our face is built to recognize faces and we build technology that does it better now. So, have these challenges now that could we versus should we and if we decide not to pursue the research for values reasons, what does that do to our standing in the research community? Does that put us behind globally? All those challenges come about when you think about technology that are so far forward on the front edge that we are running into questions whether or not we should do something.

The second of these three areas is around ethics and values. We have choices. We want to make sure that the model that we build reflects our values, and these machine learning models, since they require training data, if there is a bias in your training data, guess what? There's going to be a bias in your predictive approach that comes out of it. Government in particular needs to take great care to be articulate what the values are when it goes to build one of these tools and models and make sure that those values are explicitly encoded in the modeling work that they're not going to accidentally show up. We have to make sure you talk about what your values are and quality equity racial

and sensitivity and make sure you are intentionally building those values and the model.

Finally, legal. I mentioned before often what if we're doing something that impacts regulation or has something to do with the depriving somebody the liberty in a sort of policing context? Those things have to stand up in court so the IDF chain of custody and explain ability of the models you have to be you cannot just say you denied your tax benefit because the model told us to. You have to have a way to explain to an investigator why this decision was taken, so for this reason, it is a best practice I believe that government should never take an adverse action solely on the output of a machine learning model. There always should be that trained investigator or human to take that output of the model and set it aside, set it alongside and here's the other expertise and resources we have to make human judgments, so these models are all about augmenting and enhancing human judgment, they never particularly government never should replace decision-making and authority of a human especially when it could result in an adverse action against an American citizen.

So next slide. I made it to the end, and FDA I told you I would finish at 10:45 and I did. I'm excited to take questions at 1:00 and I know whether to do that altogether as a group in the roundtable at 1:05. I'm excited to hear your thoughts and feedback and questions at 1:00 but for now, Ernest, I will turn it back over to you and thank you for your attention and for the invitation an opportunity to speak.

Artificial Intelligence: Introduction, Applications, and an Overview of the Colloquium
Jim Riviere, 1DATA Consortium, Institute of Computational Comparative Medicine, Raleigh, NC

Ernest K. Kwegyir-Afful: Thank you very much, Steve, that was a fantastic talk and fantastic overview of AI. Next we will Dr. Riviere, the Chair for this colloquium. Dr. Riviere will talk about applications, looking at traditional modeling techniques and its migration to artificial intelligence. Jim is a distinguished professor emeritus at both North Carolina State University and Kansas State University. Jim?

Jim Riviere: Thank you for the introduction and great job. I said we were agreeing on the basic underlying concepts. I have no conflict of interest with the research prescribed in this presentation and all opinions are my opinions.

So, we have an idea about AI. My focus of my presentation is to actually bridge where normal classical modeling is used and what really is different from AI and there are some major differences that I hope I will show you as I continue but what's really important is we going to this is that a lot of the underlying processes of making sure that there is no bias and that the data is balanced carryover from modeling and this is how a lot of AI approaches in some disciplines have been developing.

I'm going to show you the priorities, so have a food safety program and I've been involved with for 35 years and another consortium of migrating animal and human health databases together Kansas and my pensionable research has been modeling

studies on dermal absorption and nanoparticles and over 40 years what I have seen is a gradual ability to make much more complex and better models and decisions much faster, which can be a problem more data and AI has been migrated into this. What I want to really drive home is that AI at one end is very different from what was done but in other cases, it actually is not. I have some personal philosophy of modeling and this is tied into is that I find the best way but modeling is we create a model and that tries to explain something and make a prediction and then when it does not work, we learn something about the underlying crisis. Doing a lot of drugs and toxins, this always happens, these issues look like but really fantastic *in vitro* tests, and then we go *in vivo*, and nothing is predicted. You just drop your hands you go back and modify what your assumptions were so a major aspect of for using models if you're not a 1% coder you're just trying to create a model and you figure out how these models can be used to learn something about the system.

There's a philosophical issue that I run into a people are trying basic statistics that is experimental design. The classical statistic is really blocking subject and doing a lot of controls but what's interesting is you have no idea what the subjects are, and in some cases, when you stop looking at how to define this, it really becomes a problem. People like to reduce variability but sometimes you reduce variability by getting rid of very important factors.

Another aspect that stricken this happens with AI models is inefficiency during different levels of models. Are you going from a molecular to a cellular to an organ to an individual to population? Disease process is often nonlinear and multifactorial and disciplined by models too much in some cases didn't know how to simplify or why are we even do modeling? A lot of factors can go into what a model is. Another aspect is statistical testing. You're trying to look at the difference between four groups, you can do classic analysis of variance in computer versus versus A versus B versus C and you can look at is A greater than B greater than C greater than D or some other covariant, but the key is if you get a significant regression analysis that implies that the sectors are actually different. Or what is done pretty much in most of the models your deceiver AI and machine learning is you come up with a model prediction and make a prediction with a certain confidence interval of what that value should be and if the sectors actually fall in the respective comp confidence intervals you feel pretty good that the model is working.

So, historically what is a good traditional model traditionally mathematically, effectively elegant models, they contain very define relationships and elements and they can make predictions in the future. Mechanistic hypothesis-based studies. However, what happens is the big data aspect is different with AI is that in the seven systems it is creating a model to create and explain the model. When you're looking at unsupervised AI system. we don't pull a lot of structure into it and it actually learned potentially new relationships and there's a lot of overlap in this. However, you start looking at these factors you have to start realizing that it is essentially an area that [inaudible] by merger of different approaches in these models. So, there's a lot of proliferation of a lot of different modeling approaches and I'm going to try to touch on a number of these and go into a little bit of depth but not too much depth in the underlying subject areas.

Again, most models are model based and relatively structured but as computer power and data storage availability increase, there has been more approaches to less rigid

analytics and we can handle more data and we can create larger samples and we just cannot handle that 20 or 30 years ago some going to use my personal research in some areas that's going to jump around over three years to illustrate some aspects of what I think are important in any kind of modeling approach.

The first one is a food safety program called the Food Animal Residue Avoidance and Depletion program, FARAD, and it's been around as a consortium with four or five universities and one of the cofounders was Steve some lock the director for CBM and food safety and our goal is essentially to make sure that when you eat a hamburger that essentially it does not have any drugs or something left over that originally was in the cow. So, what are we looking at is under certain situations the drug label may not be able to be followed and accidents occur and so we have to create a database to pilot all this information together and what are the approved drugs what are the tolerances for residues that what as a realtor agency says this is safe to consume and a lot of Pharmacokinetics which is very different discipline oriented but applying time versus concentration relationships to predict concentrations and finally trying to determine whether it's very similar to the COVID situation now they have rapid test and have final test and is that test accurate.

So, the primary task is if you look at the bottom left corner is essentially how long it takes a drug to deplete after it's given to the animal. It's [inaudible] for the population but we are predicting the outlier and that is the negative percentile with a 95% confidence that an animal is going to have this tissue concentration low tolerance and tolerance is that number that says it is safe that's called withdrawal time so what kind of situations can alter that? If you look at the right hand slide and see that the withdrawal time since this is exponential kinetics is to backwards of your ex financial growth rate and bacteriology things are dropping by was called a half-life so if you double the dose, the only increase the withdrawal time by a half-life and is probably five, six, seven half-lives in withdrawal time that's easily handled. The problem is when it's underlying disease or environmental problems come into play and now that changes that depletion and that's when you get your residue violations so we have done trying to figure out when can stuff like this happen.

So, have a lot of different types of data coming in, Pharmacokinetic data, and there's a lot of different drugs in different species approved in our countries and we can access that. The fantastic aspect of machine language approaches to scrub literature in the last four or five years has made us so much easier to get access to data on a real-time basis that in the past you have to get or in the real past you have to write a letter, you have to go extract that data, put it into some kind of database you can read and now a lot of that can be automated. But I want to get into a lecture.

Pharmacokinetics we have a couple of courses on these but there's two different approaches and this goes to philosophy again. The left is what I call the plumbing diagram. Essentially, we model this work by actually modeling the blood flow between tissue. It's a great tool for species extrapolation and we feed our data into this and then we essentially predict what the concentration would be in a specific animal. The right term is called mixed effect modeling and it's a Pharmacokinetic model with a statistical approach in the human. This is called population pharmacokinetic modeling and what it does is it makes a simple site kinetic model that does not have any reality to real

tissues but it allows us to assess the variability in the population and variability of different points and allows us to come up with a much better statistical approach.

On the left side this is a diagram actually written four years ago but all the different sources of data and that really has not changed that much. We can actually type that together now and I will show you how these databases can be curated and again, this is part of this is AI partners just traditional data processing, but the tools involved are getting more sophisticated. One such tool so if you look at here is an intramuscular demonstration of an anti-body in cattle and there is references here. Guess what? All the numbers are not the same and if you use algorithms like this you can start detecting where there is one group there's another group that is adults. One group is normal and one group is disease and you look at the poor guy on number 17 that was completely wrong and that date is not good, so this is for the export mediation utilizing these types of model systems and somebody needs to look at it and again this provides a great tool to determine what's right.

So, how to fix some of these models. Again it's when you don't get model this is a model observed versus predicted and you looking like a we're missing a lot but this is statistics 101 and regression 101 and you look at the error the residual part the predicted minus the observed NUC can you make any sense out of this in risk case will yes, it's covariates is correlated to that and get a great relationship versus observed versus predicted so you need to continue to testing to make sure you have the space covered so you can make sure that there's not some other factor that you're missing.

The right side just ties together how we tie together different types of data sources and fix our model FDA and the drug approval processes and we do a lot of very specific tissue depletion studies other all done in healthy animals and all these types of studies are allow us to see what happens when the animal is not healthy and by tying these models together we can combine the two.

So, FARAD 2020 what happens logged onto a website and forms a request and correspondence take a look and see this is a legal drug application so it's not a [inaudible] being used and we have request on this rematch it and research our databases to find everything and if it is slow one we have an algorithm that allows us to convert what's called maximum residue levels to U.S. tolerance he equivalent type approaches and then we use a number of pharmacokinetic models. When we do this we create some of this models which may not have a lot of studies in it but this these anomalies is not for population study to say cattle population in United States is probably for five cattle being treated for something so the REIT risk assessment is not meant for a lifetime. We don't publish these type of approaches we are trying to mitigate risk, so we tried a model try to profiling have some variance created some synthetic populations and then subpopulation and sample in the same way that FDA would do for drug approval process using what's called an FDA tolerance algorithm and we do this on the find physiological models and on population-based models and then these models can be constantly updated and we see these guys all the time the solid cost are the training sets in the open circles are validation sets, so whenever we do this we always pull apart two sets of data to see how well a model is a doing. And then we essentially come to a decision by a committee.

This is a historical response and what's that number. We have a heuristic half-life those multiplier algorithm what you do to the machine language is where we what are the conditions and what you think might be now and if we look for foreign approvals we go ahead and modify the tolerance and see if it matches that number and interspecies to the right-hand insert this is where we can actually do what's called a comparison and the disposition of a drug. We do a rat and a horse and we consider if this is a well behaved drug and we can do extrapolations than if that was not true in the population pharmacokinetic estimate and we do the estimate and then essentially we do we decide what we should do and that's a decision and we tend to be conservative and therefore go for the longer time unless they can find a reason to go shorter and again with enough experience we can try AI to actually do this for us.

The second program I wanted to talk to you about is the one data program which is a program in Kansas that we tie together human and animal databases to essentially pull together elements that we look at one data element in my be very good on the toxicology a look at another data source might be good on the efficacy so how can we pull these things together to create synthetic entities that we can do some work with. I want to go into a lot of detail but everybody creating curated databases realizes there's a lot of different sources and the key is to identify the table of similarity between these data sets and how we can emerge them into a simple data sets that can then be used for analysis. So, jobs are different across countries, chemical compound names may help separate specific chemical entities, but it's the process of creating how to link this data to bring it into another data space on which the analysis is actually performed. You have to realize if you make a mistake on that aspect, you're going to make a mistake on everything.

This is data occupation and the way to look at this is to essentially look at three different data sources, you create the puzzle this is the food initiative to estimate what some of these missing pieces are these missing pieces are and then you allow that to be able to use to make your predictions. We have a tool that we developed in this program called one drug assist and you can use it for last three weeks to actually explore existing adverse drug events databases from the U.S. FDA from Canada databases are in the ACE inhibitor class is to be problems with inhibitors and we have been looking at very specific drugs within those classes and can we determine symptomology that actually might have a problem with COVID and as we define them and looking to medical records systems that are part of our group these are to see what we can actually help manage or avoid some problems. Interesting application but the system was ready built before the crisis hit.

I'm going to go faster on this because you will have about a talk in this area and another area we did a lot of work causes the penetration on the skin it's a modeling approach and for cosmetic safety this essentially is a compound put on the surface does it penetrate the skin. A lot of this work is done in what is called quantitative structure activity analysis in this cosmetic domain it's really quantitative structure permeation relationship and as we look into this it gives us a technique called [inaudible] relationship. I cannot go into all the chemistry but essentially it is how one compound tends to associate with another compound and be it traction bonding we can essentially look at that and come up to predict what is the permeability and it has been done for a long time.

The one we really focused on was we take a look at absorption in a single vehicle because the data is not that variable and you can build a lot of neat models but if you look at the predictive versus the observed concentration in the top left block of that four block area you see them KP effect this is the slope to the actual chemical. However, if you look, there is lines there because each of those chemicals are different vehicles in the vehicles is actually larger than the slope due to the KP. We have come up with approaches to models, what is the effect of the vehicle on the absorption and again is the residual plus and you see that the residual is on columns and those are all because of different vehicles and resort explain some of the vehicle effects we have a better estimate of absorption and a lot of data that goes into this because of a lot of vehicle work and has a lot of good tools that allow us to model is better.

We have applied that an actual compounds a lot of natural stuff is put on the skin let's say and there's a lot of stuff that might really be on the skin and candles factor actually modulate absorption and the answer is yes so we have been trying to get some work of how to predict what those natural factors and the mechanism behind that enhanced absorption. I'm going to this exercise with nanoparticles, so essentially I'm looking at this chemical interaction between the nanoparticles which can have multiple types of forms and what we need to know is biology is how does it interact with stuff in the environment and in the body and how that can be small molecules. We use AI techniques in this approach and we want to look at 20 different small molecules and essentially look at their binding and simultaneously saw 28 type relationships a whole lot of data goes into this and what you get on the right are the 16 blocks of very different types of nanomaterials not characterized by how they interact with small molecules and the molecular forces. If you want to just predict what is the partitioning behavior of that if you look at the measured versus predicted that predicted is based on the modeling system you can actually predict physically trying to partition these and again you see the training sets and the purple of the validation sets and along the right is something called a plot to share that your data set is relatively well-balanced because sometimes these relationships can be thrown out by one outlier.

Final thing I want to show you is to classify these nanoparticles into three different groups and even though they’re very, very different they tend to behave similarly. When you see how large the circles are in that diagram is because some of these binding parameters are so small that it's hard to measure analytically. Instead of just measuring concentration we predict that low concentration from the other sets on the Langmuir model and that pretty much squeezes down that prediction. So again, that data set has assumptions built into it and you have to really watch out and have expert areas to know that even if you are going to let this this go into a world classification scheme that data sometimes needs to be curated.

So, what have we learned from these projects? The one thing I learned is use models with different assumptions and it's a lot better having two or three models that agree than having one model that you have no idea if some assumptions is violated. You can use structured mechanistic models to define input parameters into more general AI type approaches. Like preprocessing the data with a pharmacokinetic model or something like this when Langmuir and these apply to more into the toxicology aspect of food safety. More data and the more the diverse the data, the better, and that melanine incident that occurred in China on milk and in the U.S. on dog food like 8 to 10 years ago never was predicted because nobody ever included a plastic byproduct in any of

the food safety databases. So, if something new comes up, you're going to be surprised, the need to update data sets and find new sources. And develop strategies, if you can use a data in a very large analysis you can actually define who can get credit for where that data came from and other people have been doing this using block chain to tag individual data elements.

I'm going to get into AI, and I'm going to skip over a lot of this because we covered this pretty well. The major thing I really want to focus in this line is really the difference I see with supervisors is unsupervised and the algorithms used currently are covered by later speakers but the key is are you telling the program how the data should be interpreted or is the program analyzing the data and telling you what the relationships are or how you combine the two and again everything you're talking in here if you look at the AI literature it's really domain specific and not getting into general. This is the stuff of super intelligence but it's not the stuff that were going to worry about. Again with AI a lot of people's mind it's really a confluence of multiple computer digital developments that allow us to access a lot of data, a lot of data fast, analyze that data fast, and come up with rapid decisions almost in real time and a lot of factors tie into this and quality computing, data AI and big learning, and along whole aspect of the Internet of things and to safety that becomes important with waiting for ID chips and sensors that you can actually get everything together and then the ultimate application of block chain that provides security in these systems. A lot of different types of artificial intelligence the machine learning areas again, you can see different aspects of these actually discussed in some of the next slide so I don't want to get into too much but just realize a lot of it has to do with how much is routine based and if you look at the level of machine you have these systems, civilized learning systems worse is unsupervised learning systems and this is touched upon on PREDICT and Dr. Abernethy mentioned this. Moving into systems that are not supervised.

Typical study design really not that much different. Organize and prepare the data. This is what I was telling you about the data curation. You have to make sure that similar things are named the same when trying to an analysis and identify what features can be used as input, identify the outcomes being modeled identifier training sets, figure out algorithms go through and refine the analysis and prove your labels and rerun and rerun and validate with a separate data set and when you feel competent confident deployed. A lot of what's going to be discussed that people think is synonymous with AI are the algorithms used to do this and these are Neural Nets in the upper layer what you're trying to protect any let this continually approve that training and it's cold so-called Bayesian approach the one you come up with the estimate you keep on improving your estimate and tweak certain areas that need tweaking and the hidden layer which is when you know what's going on in earlier these are very simple in your nets because you can get layers would hundreds of layers, and it becomes confusing as to what's happening.

The original aspects of AI learning was rule-based and logic-based systems defined by knowledge domain experts for instance graduation translation programs using grammar syntax diagnostic systems and this was the basis of AI original success in chest which was the Rubik's game and trying to decide what happens and can you predict and as we mentioned before go is a little bit too complicated forgot and particularly the systems look at the entire system and figure out how we can get to a template which is winning

the game these-based systems but we not ready for that and a lot of applications we have been discussing.

Logic-based algorithms it's defined by the program and define the relationship and this I think is the big division between AI and what has occurred in the past. Image analysis systems that were touched on really get the point of pattern recognition and you're going to have an excellent discussion on this in the next speaker. I don't want to talk much about it. It impacts self-driving cars, identifying cats versus dogs, and you have to realize how this works is that a human actually makes that association and says this is a picture of a dog, this is a picture of a cat and that's information within the algorithm and maybe some this security things when we log on and you have to go find how many stoplight are the picture you're helping train the AI system. For self-driving cars it's very large and diverse data sets in the make the surveillance system so you can focus on adversarial examples, so it's fascinating literature.

Misclassification systems which I discussed versus classifying the specific sets of data into disease of data into disease or non-disease or classifying really complicated nanostructures with complicated data set into groups of similar behaving particles and the system to define associations you try to identify clusters of similar objects and this is a major point of big data and as mentioned earlier this is basically a business bought model of Facebook, Amazon and Google and when we buy, we see that advertised everywhere. Is also chemical contamination in the food system and this has applications as we will see and it's also the basis of the systems across the chemicals based on chemical properties and outcomes and we will hear about that later.

Big data and food safety, there's a lot of platforms a look similar and these are primarily online databases containing food safety relevant data and WHO has the food safety platform and creating multiple food safety related information sources resources that can be tied together and in some cases can be analyzed together. Major AI efforts in food safety are waiting to visualization and geospatial a linkages and a lot of work to map out and track food supply and this is applied in the COVID pandemic right now trying to chase down where the bottlenecks are. Supply chains with its for the GPS with sensors and systems that allow you to get an idea where things are going were problems I have originated and where things can go to eliminate them.

Talking with Ernest and looking over what's going on within the FDA one aspect where AI is actually used a lot is in the device section because software can to medical devices and the realtor framework for some these systems are actually published by CDRH and there's pilot projects using block chain for food traceability and in the system these extensively by dermatologist across skin tumors and by radiologist and reader graphs to automatically process the presence of a mammary tumor. There's chemical hazard and intelligence programs that try to tie together to come up with can actually figure that there's an event occurring in hours rather than days and a good surveillance monitoring theological investigation and the project system for inputs. And the text mining tools to scan expected notes to see if everything is coming up and again post-market surveillance systems for drugs system, Sentinel and the adverse event reporting system, using a lot of machine learning approaches. This image analysis system has been designed to identify to look at drug-induced injuries.

So, I want to finish up here in the next minute or two with potential witness weaknesses. One aspect of AI is that it does not inherently from sampling to data set so there is an absence of sampling errors and what that means is you really need to make sure that the population that you unleash an AI system on actually represents the population you are trying to make inferences on. Another thing inferences on. Another thing that happens is we can get carried away and we offer I one complex system to the input of another conflict system where things can happen if you do that if you don't understand what's actually driving that system.

Many systems remain challenged by complex environments and you can see that with the COVID aspect right now to create certain models and you find out that there is other pre-existing conditions or something else is explaining what's happening and you have not included that. But the key about this is realizing that every model is not a complete model, you need to constantly improve it and if you get really, really large models there's a so-called brittle because they can create these instabilities in certain areas and model crashes. So, there are some specific problems that essentially if this data entry problem and if you do that you can have the perfect model but you're going to have garbage data and that garbage data results in garbage results and you can have a horrible model and you need to look at the specifications of these kind of things and what's going on in that. You have incomplete or unbalanced data, this is very important in food and cosmetics and you can obviously malicious intervention and hacking.

This is my last slide. Potential weaknesses on problem is sociological sensitive and engineers are trained by classic definition of modeling and certainly lack of details and we should be careful about that and therefore we have multiple systems actually this action makes sense that people are trained by AI but what happens in their home systems and you might get frustrated with Alexa or you might get frustrated that you cannot find your robotic vacuum but if you unleash a AI system without any feedback on the food safety system, you can have real consequences.

Now, we will move into very specific applications. There's a lot of suggested readings to follow up. And I'd like to acknowledge all the people that have done the work and the agencies in the front of this for the last 30 or so years. And thank you very much, and I will turn this back over to Ernest.

Ernest K. Kwegyir-Afful: Thank you for the linkage between traditional and AI. Next, we have Dr. Nicholas Watson for the AI technologies for future factory cleaning. Dr. Watson’s research is focused on developing [inaudible] technology [inaudible] for food manufacturing [inaudible]. Dr. Watson?

AI Technologies for Future Factory Cleaning
Nicholas Watson, University of Nottingham, Nottingham, UK

Thanks, Ernest. Next slide, please. I have no conflict of interests.

So, this is a general review of the presentation content of our group. We have tested and tackled some the challenges with them and fracturing safety traceability and when we want to do that by developing technologies and data and the challenge in the rescission or if is anything we develop has to be able to work in the factory service to

block the work, but ultimately we want to make sure that all the technology works and we found out when we started this journey several years ago that reaction using censoring and it's really, really expensive and what the main problems is that we take existing tools and trying to work them in our approach is really going back to something that we say is the long-standing questionnaire trying to answer and maybe representing high-resolution images need some separate technology that will allow you to make a decision.

So, we went with the ultrasonic and optical technology because it’s a relatively low cost, which is very important and several years ago we were faced with this problem on recross data, but it's very noisy is very variable and how do we actually turned out sensor reading into the material that we're trying to model? And this is where we got very interested in machine learning and we started working in this space. And I think one of things I want to talk today is one of the great things that machine learning can do. We will highlight some of those challenges also that we have to overcome, and we need to realize as a community it's not going to solve everything yet. And we pull up in IoT applications because it enables us to do things more effectively, then we talk a little bit about that.

And the two areas we're looking at is process optimization cleaning and also look at mixing and exclusion and then the properties in the food and we will talk about main projects that were working on one is self-optimizing and cleaning places. To adjust the current challenges with factory cleaning. There's two areas that are looking at your cleaning the internal surface of all the equipment and we make sure that there are hard unique and making sure that we remove all the material efforts clean. And the second part is actually cleaning the production environment, which is the outside in the Mecca.

There's lots and lots of challenges here if we think about the drink and when patching sector is characterized by high volume low profit margins, and generally, the people cleaning are also doing other tasks within the factory at the same time, so it's a sustainability numbers. So, if we look on the right side we can see 35% of the water in the production, and in dairy, about 30% of the energy and these are the worst ones because they have high temperature and long life and using high temperature and that's the same in the brewery when you're boiling the water and it's very difficult to clean.

The first project I'll talk about is optimizing cleaning in place. So, work on the assumption [inaudible] if you look on the right, we got some trends there. We’re keeping a factory; we’re simplifying and a lot of our work this is the process and we need to clean this and what would happen if you have to disassemble it? And we have been using cleaning places and these are autonomous systems and you have a tank which is maybe some mixers and you might be making let's say a soup in there and you've got lots of different [inaudible] and you need to clean it. So, you have what's called a springboard inside of that and that's attached to the of the chemicals and fluids in there that help it clean and it's similar to a dishwasher at home. Initially introduced that would water temperature and when you have the formulation are trying to clean and then you rinse the chemicals and then he might have gone through these stages and the target at the end of the process. Hopefully every time you open that dishwasher door you're greeted with a load of clean plates and cups and pots and pans and that's fantastic because that's what you want. That's just from that sustainability point of view.

So, cleaning is done by the mechanical action of the load of the water coming out these spray nozzles and around the tank would chemicals and is augmented and it's validated routinely with swabs or ATP tests. And looking at this project what was the cost of cleaning and we found that there's a cost associated with any if you go back to the about the food every minute that you're open and every time every time you start the cleaning equipment you not thinking about the food so that's an efficiency and this is actually a problem that's getting worse and worse.

One of the food manufacturers that working for this where they made the not true [inaudible] some of our traditional Christmas foods such as Christmas pudding, it's even back 10 years they used to make one of these so they could [inaudible] and clean the equipment. Instead of making just one batch, they now got low sugar batches, one with premium ingredients in there, so what their production schedule looks like is there making the same volume of this desert but instead of having one large batch they're having 8 smaller batches and they have to clean between each one set of cleaning once a day can be up to eight times per day and this is by just become a problem and just going back to that idea that the dishwasher every time you open it your pots are clean and that's fantastic but if you keep running that same for half an hour or an hour do you think they may be cleaner for half an hour after an hour and that's really the problem because some of them clean really quickly and the rest of the time it's pumping chemicals in, caused loss of energy, cost energy and time and this is because there is no sensors in the intelligence. And some of the dishwashers you find they can indicate concentration and the candidate's billings on the surfaces, and this is what you want to look at. We can tell when it's clean and we can stop cleaning when we washed them and normally when you design this you design for the ones that file the most because you do not want to not have a clean system but there's a real attention there.

We have sensor companies and food many factors and who work on the instances and working on the optical sensors and I will not talk about that but it works on the intra-violent sensors in the place we work on this is that you can set the sector down into tanks and it's just an interesting point on the mixing vessel. And restart taking these cameras in there with your system you can only see 50% and that's no good. So, what percentage could you see? When you use sensors in real life we have to be realistic in the capabilities of the technology and even if you see only 50%, 20%, 30% or 5% is a lot more than we could do before so we got the sensors and they're never going to build on machine learning models and use that information to control the process.

Let's talk a little bit about ultrasound because that's one of the technologies that our research group uses and return to the single-point approach where we discussed a single way and really interesting is because of the years the technology reduced cost and in size and the and what I said about IoT we don't need those expensive laptops computers and electronics. And if we look at the two images 2012 that's the picture and it's a mobile phone unit today and that includes a laptop and some electronics and there's about 25,000 pounds and that was reduced to a monitor and one of the processes. And we got is very similar to ability in the device a few thousand pounds smaller than a mobile phone now I can and connect that and that makes measurements in an environment a lot easier, more advanced and is noninvasive so some of the projects are working on, they’re actually very interested in combining differences and multi-sense data fusion.

I should've just mentioned in the previous slide at these are production environments and their abilities like variable temperatures humidity vibrations people walking around in the ultimate challenge which how to detect that measurement entered into actual information. Looking at the principle of operation we assume that we have a pipe and the orange brown that's [inaudible] if we look at the middle, what we are seeing is it will propagate through that pipe wall and will reflect from the internal pipe and it will turn to the transducer and begin.

To the far right we just have an image that reflects there is actually several [inaudible] and were interested in this sensor technology is what happens at that interface because if we are identifying those signals part of the signal that's that reflection look at that and we should be able to monitor one filings are there. We have been working on this project for several years now in a small-scale rate and two-time cleaning systems on the right and we [inaudible] material on their left and rear cleaning and last year we started doing them until production scale. So this is our small Reagan we actually have the falling material where we looked at several different [inaudible] and you can see in the reddish blob in the center underneath this is a method to replicate what you get in an industrial environment and we're just using an image here.

Can you play the demo, please? This is simply showing you the consider waters flowing for the system and it cleans away the dirt. One of the questions I hope lots of you are thinking right now why do need this [inaudible] you can just watch it through the camera for controversially that's not the case. In factories these are all closed pipes, so but what we are using is because lots of build model and we want to build a supervised machine learning model so we need something to label our data so we will talk about that in a bit.

Here we have several repeated results in a believe this is a concentrated use in the beer receptor so disconnecting what's going on, on the left we have several graphs and the images are going to show you as the clean this and on the left we have some features we calculated from the ultrasonic signal and the blue line is the energy of that reflected signal and want to go through we can see that's going to as it becomes clean we started reflecting [inaudible] we now have water there and expect to get a different reflection represented. The orange line which is the exact opposite but at the same time what we do here is we take a pipe that we know is clean and clean and has no filing and we compare all the way that we are recording to that log and then once it goes back to clean state that area should be reduced. We can see if we look at some of these if you just pick one we can see that you can see the images the stain is being removed and we can see this clean others a few points there the ultrasound is appointed measurement and only senses the filing at that position this is often one of the challenges with when they use and industrial environment.

The other thing I should've told you is actually below it looks like all the lines go to the same point that's the way we normalize the data in these figures, and we do find time limits, so this is a very different falling material. This is gravy and this clean very differently than what we find with this is this is a lot more noise than I would like to look at the one that is on the third row down and if you can start playing the video, please. You can see this clean very differently and if you look at the third down on the left what

we're noticing is that is not cleaning correctly but it's good to disappear all of a sudden since you keep watching now. We can just see the cleaning has disappeared.

So, we can attract them with this [inaudible]. Initially when we started this work, we said well, we actually need to do some machine learning because we need to look at gradients and how that gradients changes and we don't really need to do the machine learning. What we actually started as we were doing more of these configurations and different materials, we found that there was lots of areas and it did not always have the greatest impact, and this is where the machine learning is really powerful.

And you don't always need to use machine learning. There are complex techniques and just look at your data and if you see a [inaudible] in your data that sign and then using these collocated methods. We were fine and we cannot actually take this selected to supervise machine learning and we had this from the, which tells us [inaudible] and its clean and in label that data and that's all you are trying to predict is it clean.

So, just to go through some of the steps we have taken in machine learning, we need to understand that it takes those requirements so we have to collect data and range temperature that will be relevant in the second part is to collect the data. And then you label in your data and then there is splitting that up into the training part and then the part in the middle is featured engineering, so we have that but what we actually take from that that we want to use in the model under several different approaches you can use physical picture physical features or you could hire ultrasonic waves and wanting you put a data point in the actual volume that's when these problems, we tried several different models with those wave analysis and we can do principal components on K best predictors and features and we identified about 200 best places.

You can train the models and consolidate them in one of the things you're looking to is making sure our testing app that has not been used in the training validation that's really important and one of the practices and you can repeat this several times. What we like to try and do is make sure that the last [inaudible] we do the monthly to at the end is [inaudible] that in the testing and the application of this application of this the models and industry to collect the data and you try to model it and then you try to make it work for a limited amount of time so you make sure you are testing that data.

This is some of the results and you can see this our flat rate. We put these black circles in because this corresponds to the region of interest where it was [inaudible] to the other side and what you have is in the top line where the numbers are. This is from the images and each represents when this filing is still present and blue when it's clean and this is the ground data. We used some technique from classified [inaudible] we take some shallow [inaudible] and combined them together.

This is lots of individual decision tree and we got the boost in message. What we can see here, and these are two different test runs one is at high temperature and if you look at the bottom you look at the bottom you get really, really good models and most of the models really work. What this is doing is this is telling us after it's clean so after number 7 that shows is making a prediction that and this is one of the challenges we have seen you don't understand really are not always so you have to be really colorful and can go back and changed [inaudible] this is something else I want to talk about.

The K machine learning we found over the years and the quality is often the volume and it's been very difficult to make and model to look at the top and what we have done it here is all the data we have on this we are trying to use the modeling and they have the different temperatures and we feel comfortable compiling the data from the different materials because ultimately the final stage you are trying to get to is a clean pipe and we see the solutions here and all the models in the bottom table what we have done is we build models where we build models to predict only using the tomato data for training data and you can see the model performance is a lot lower than the cost of using the status of really think about that and raising standard machine learning techniques and it really needs a lot of data I need to think about how you get that because of often that's the problem on how to get that.

So, here are some results in a pipe simulation we used and again some problems here this is a different type of problem here because this is connected clean and it's actually more in the case of our problem here this is much worse than the [inaudible] when it's actually clean this is going to tell us where the safety issue might arise. These are some of the results of the SS pipes we can still see regarding model predictions and from different materials and I think your work become notices a lot of the results we can see this is a problem for us.

Once it's predicted because what we found is that [inaudible] in the different cleaning stages. Trying to predict regression models and this is showing you some of the techniques we use so we have been using your networks here and we're trying to predict the volume and the number of pixels found in a region of interest. What we can see here is that this work in preparation with the University is the filing of the region of interest which is the same and here is our training test results and which you can see again for this regression very good results so we can finally predict that it's clean but we can also predict that [inaudible] because that enables us [inaudible] in the cleaning processes of the manufacture can actually know when that can be used for production again.

Just a few concluding remarks on this for the real challenge is the training data set and we got some ideas about this and labeling that data and in our lab we can use images but how we are going to use them in the factory and this is where we are looking at solutions around the data or transfer learning so try them in the lab and in testing in the factory and looking at that in the machine learning project and then it will give us very promising results and we might in a factory environment might be better.

I'm going to summarize the work we are doing another project this is but cleaning the factory environment and this is a Robo clean project and what we're looking at here is how we can use robots to clean the factories and am aware that there is already robots to do clean factories but the errors were looking at a research perspective are how do we communicate with those robots and looking at that and the language we use and how the robot communicates back to the operators and the symbol stage and coordination so we have multiple robots, and with this we are using the infrared sensors and again, it's very good identifying changes in composition in this area and we are using a small [inaudible] low-cost.

So, the first part are the were you want to do is you put the sensors on anyone to identify what material you are cleaning and we know this is a big challenge in building

environments and overlook that here is just looking at two problems, one is open materials so that we have a small sensor and we just placed our sampling dish. We had lots of different flowers and spices lots of different spices in there. And this is showing some of the recording to get specter and reflection of that light and what we can see is those similar to see the difference between the different types of flowers and materials. And this is just a spectrum for the spices. You can also use the camera and were looking at combining the AI with a camera actually.

These are some of the results talking about clustering methods of so we just they're very similar civil but it's common methods component analysis and this reduces the resemblance because you have a lot of spectral data that represents the variation principal component analysis and what we can see in this figure this was 40 the group that went [inaudible] and you can see they're all spread out the only time we see a bit of a group between the gluten free and the gluten contained flour the red and black. We got several dots because we use different varieties and different suppliers and a lot of these as well. These are the spices and I think we can skip this slide.

And once we got that data we have some classification models and with the flurry wanted to look at you can determine how many classes you want we can just have one class of material in one for non-gluten containing materials and we decided so far gluten contained flowers, non-gluten contained nuts animal-based another. And here is the results. Again, we use different methods that were used and we got the training set and the test training data when we have huge metadata and we can get investigation up to about 95% vacancy in the table below most of these materials classified quite accurately but we a fairly small data set. Some of the projects were looking at the quality and if are looking at the quality of the cookie and how many chocolate chips it has in there and if we have external sensors and we have 90% accuracy that's good and we're trying to the text or use the right material and these allergens are not in it 91%, 99% of that success when this is what we have to be really careful in what those results mean and what the consequences again are to account for AI and trustworthy AI in these areas.

We can skip this slide that was the result of the spices. So, I just want to talk about some of the challenges in the future work. Within this particular project we have experimental parameters and looking at the production line for this is spice mixes, tea mixes, and is that part of the material what we would expect to see on the floor. It hasn't been there and some the future look we're going to look at operations for example garlic powder and non-purpose just because of the equipment used with peanut flour and what is the lowest concentration you can identify? In the teams we were a lot of the work goes into this area of infrared looking at this and we have some workshops involving people and lots of organizations and food manufacturers and their biggest problem or the biggest worry is not a material with a slight amount of alteration. But if somebody's taking the wrong flour or somebody labeling their own flour and if you have an idea about that they might make that mistake is really good and but these sensors use a giving indication right away and of all the factories will be following Goodman faction practices and the be taking samples send them of the testing but that they have to do is they need to give you information to you right away. So, real actionable information and needed someone to interpret that reading and the making of autonomous decision I think it should be put back to an individual looking at that and making a context.

This is a summary want to talk about today cleaning and contamination is an essential operation in many not just manufacturing but restaurants, in your house and you can look at the processes of cleaning. And if you mop your floor and you mop every square inch, we need to look at precision cleaning technologies and sometimes key parts of that and we're interested in small sensors easy to use and machine learning and machine learning is a powerful method because it means we have to develop these complex physical mechanistic model to understand what's going on as long as you collect the data set and the variables we are absolutely fine. And the challenge then with the machine learning is the time and cost of that data labeling understanding complexities.

Here are some references in action some of our work is a lot of the cleaning work is. Please get in touch and I can send you all the work that we have done in this area. There should be one more slide.

Lots of different people who worked on this project and in my team and University and engineering and pewter scientist and we got lots of external collaborators and they are our partners and multinationals that we've had and Internet food and drink green core and any questions I look forward to answering them at the end.

Ernest K. Kwegyir-Afful: Thank you, Nik. Thank you for giving us very good insight into the future of factory cleaning. Having worked in food allergen risk assessment before, I can see how this is very important to a lot of people.

Using AI to Extend QSAR Models
Chaoyang (Joe) Zhang, University of Southern Mississippi, Hattiesburg, MS

Ernest K. Kwegyir-Afful: Our next talk is going to be by Dr. Chaoyang Zhang.

Chaoyang (Joe) Zhang: Hi, yes. This is Chaoyang Zhang from School of Computing Sciences and Computer Engineering at USM. Go to the next slide. There is no conflicts of interest.

The objective is to introduce SAR-based chemical toxicity prediction and also present how to develop machine learning approach for QSAR modeling and how to extend the QSAR models with deep learning. And also address its challenges and identify the future efforts for predictive toxicity analysis.

Toxicology studies, they are important and critical in food safety. We need to identify those chemicals that disrupt normal cell function. The *in vitro* and *in vivo* study approach started in cell lines on the animal subject. It is time-consuming and expensive, and they may also get involved in ethical issues. In recent years, more and more effort has been put on the development of computational approach, use of data mining, machine learning, artificial intelligence approach for structure activity relationship modeling, SAR modeling.

This figure shows the framework for SAR-based predictive toxicology. Here we need to connect a data set that consists the length of a compound. So, we know the structure of each compound, so we can [inaudible] molecular descriptor, which is a [inaudible] to describe the chemical's structure. We already know the activity indicated by the labor as active or inactive. That's the data set. We can connect it to train the model and appropriate machine learning and classification and then use data to train the classifiers to train the models. Basically, establish a mapping between the structure andn the label. Then, we apply the models to the new chemical compound that we already know the structure, we have the molecular descriptor. So use this feature as an input to train the model, you can predict the activities of the chemical, active or inactive. So, that's a framework for SAR-based predictive toxicology.

The previous slide showed the framework of supervised learning. We used the trained data to train the model then you can train the model for a prediction. So, these types of machine learning are called supervised learning, and they are [inaudible] don't talk about in details here.

This is the supervised learning framework. First you must have a quantitative training data set and train the machine learning algorithm and apply the training data set to the machine learning algorithm, go through the training process, to obtain the trained model. And then, enter the new input of data to the trained model for predictions, and another important component here is the evaluation. How to evaluate the model. There are so many different evaluation methods, each one [inaudible] appropriate.

The supervised learning contains several steps. So, first, you must have good training data. And also, how to describe each data object that the feature representation. We use a molecular descriptor to describe the chemical structure and which algorithm to use. Deep learning, Bayesian, [inaudible] machine, or random forest? We need a complete design to build the model and how to evaluate the accuracy. Those are important steps.

What data can you use? Where the research works is the Tox21 data challenges. It's a federal collaboration between NIH, Environmental Protection Agency, and FDA. The goal is to create a benchmark data set, then allow individuals or independent researchers to [inaudible] prediction model and to see how and where the model performs, chemical structure data. More detail can be found from this website. So, that's the data set we use. We can compare the performance with the other researchers.

So, of the data preprocessing [inaudible] we use the data set. The data set consists of twelve bioassays. And for each assay, [inaudible]. These are total number of chemicals we selected from the data set. We removed the duplicates. And this data set consists of two classes, the majority class and the minority class. You can see a significant difference in terms of data object or data [inaudible] in these two classes. The ratio is significant; this one divided by this one. The imbalance ratio. That's IR. IR is very significant. How does IR affect the prediction? So, that's highly imbalanced data. It's the same for the [inaudible] data set. That's the question we want to answer in this research.

This is a machine learning approach we developed for this research. The data preprocessing and the feature selection, we skipped these two parts. And then, we have the data set, into three data set, and we test the data set. And here we [inaudible] highly imbalanced data. We need to use sampling techniques to improve the prediction accuracy. Here we generate a *N* stratified bootstrap sample. Why we use [inaudible]? Because if you only split one time, we get a bias. [Inaudible] get a good result or a poor, poor, poor result when we use *N* stratified bootstrap sampling. Stratified means we maintain the distribution. So, here, we consider three sampling techniques. So, the first one is the random undersampling. Undersampling means to randomly remove the data sampling instances from the majority class. And these two methods are shared on the next slide. They're the three sampling techniques combined with the random forest, the random forest, that's the base classifiers we use, so we get a [inaudible] prediction model, and then, we use a sampling learning to create a *N* classifier [inaudible] with respect to each of these classification models, and then use a majority vote of different classifiers to determine the final prediction labels for each chemical.

I think it's worth mentioning two techniques here for imbalance handling. So, one is SMOTE. That's synthetic minority over-sampling. [Inaudible] important information, right? So, it's preferable to use oversampling. The traditional oversampling, just [inaudible] repeat the selections, [inaudible]. But this may not be a good idea, so SMOTE technique [inaudible] will create an artificial instance here [inaudible] and then create an artificial one, which is just a point on this slide, random point. So, that's the idea, to increase the minority [inaudible]. That's a great idea, but if the featured space is a [inaudible] shape, then there's [inaudible] problem. [Inaudible] data in our case, it's possible to have an artificial instance like this. [Inaudible] if this is the active instance, right, but all of its labels are inactive, so all labels and this one should have a similar chemical structure. The structures should have a similar activity, but this is a contradictory result, right? So, labels are inactive, this one is active, we need to remove these kinds of artificial instances. This is called a SMOTEENN, combine the SMOTE with Edited Nearest Neighbor to create these artificial instances created by this SMOTE algorithm. That's the package we are going to use to handle that imbalance.

So, the base classifier is the random forest [inaudible] to vote to determine the prediction and label with the final prediction. Based on this random forest, combined with the three imbalance handling techniques. This one without any imbalance handling, so we have four classification models, and next we want to test how does the four classification models perform.

Another issue we need to address is evaluations. Those are standard in model evaluations have been widely used for different applications. Precision. Recall. These are two parts. We won't go into detail here. [Inaudible] mean of the appreciation and recall, so we have the specificity and the balanced accuracy. Balanced accuracy is the mean of these two terms. Now have an MCC and the Brier score, AUROC. AUROC has been widely used. And we will still consider AUPRC. P is the Precision-Recall. Precision-Recall [inaudible] we can also get a curve. Area under this curve is called an AUPRC. Here, we are more interested in minority class of active compounds. That is very important because we don't expect to predict a chemical compound that is active [inaudible]. And another thing I want to mention here is AUROC may not be good for performance [inaudible] highly imbalanced [inaudible].

Here we share part of the result [inaudible]. This is a performance comparison with respect to five evaluation methods. [Inaudible] lowest rank means they have a better performance for most of these evaluation methods. Here we have the p-values or the multiple and pair-wise comparisons of [inaudible] similar performance. Pair-wise comparisons [inaudible] three classification models. So, we can see the F1 score, this one, the MCC, and Brier score, those three methods, we've got a consistently low p-values. These three can [inaudible] the hypothesis; in other words, these three can show the differences of the different concentration models. These three are more sensitive and also provide consistent evaluations for different methods. So, that might be the evaluation methods to be considered for highly imbalanced classification problems.

So, we compared our result with the Tox21 data challenge winners. Among twelve assays we use AUROC, so five of our predications are better than Tox21 data challenge winners. Here, we used balanced accuracy as a criterion [inaudible] of our predictions are better. We calculated the ratios of all classifiers and challenge of [inaudible] accuracy [inaudible] ratios higher than one [inaudible]. We only compared these two methods because these two are the methods used int eh Tox21 data challenge, so we used the same criteria.

One question we need to answer is to address how the imbalance ration the performance. This is the log IR, this is the F1 score for four different classification models. So, we can see that there are strong correlations between the accuracy and the imbalance ration. The higher the imbalance ratio, the lower averages we got. This is true for all different methods, all concentration models. We can see the IR is greater than [inaudible]. That suggests to us that to create a [inaudible] of the data, we need to [inaudible] the data set in the future. And even if there are negative correlations, we can see these methods [inaudible] the average is consistently higher than the rest.

We summarize these applications. So, here, the imbalance handling can improve the imbalance of SAR-based toxicity analysis. There exists a strong negative correlation between prediction accuracy and the imbalance ratio. All methods become less effective, including the imbalance handling techniques, right, less effective if IR exceeds a certain threshold. This suggests that we need to increase the minority data instances. The F1 score, the MCC, and Brier score sensitive metrics. May be good for performance evaluation.

In recent years, a lot of effort has been put into starting deep learning for toxicity prediction. Dep tox, a significant effort, develop deep tox and deep learning for Tox21 challenges. Win 7 of 12 sub-challenges. [Inaudible] deep learning or shallow learning methods for QSAR modeling. And based on this work, their results showed that the deep learning [inaudible] outperform others, all data sets, it really depends on the physical data. This motivated us to have a further investigation to see what the potential of deep learning is, can deep learning outperform the others? So, that's our preliminary result. Deep learning-based SAR modeling for multi-class. The next few slides are a summary of this result.

We skipped the data preprocessing part, so we got the data we used for seven chemical compounds, four classes, the outcome, assay outcome is agonist, antagonist, inactive, and inconclusive.

This is the diagrams for deep learning models. Here, you include hidden layers. How many hidden layers? You need to decide. There are four units and output layers corresponding to four class labels, agonist, antagonist, inactive, and inconclusive. The deep learning, there are so many parameters we need to optimize.

Here, we simply use a Bayesian hyperparameter optimization tools [inaudible] deep learning library, we'll skip this.

That's the tree workflow. So, after the data preprocessing, we proposed two nested loops. The outer loop [inaudible] each data set into [inaudible] and we used the rest for tree. So, we calculated the average accuracy to reduce the possible bias foreach partition. And then for each training, we used a tenfold cross-validation using [inaudible] loop. So, the purpose of this is to optimize the parameters so [inaudible] and then use the validated models to test the performance.

There are so many different algorithms. So, to save time, we didn't compare deep learning. With all different methods, we did a simple check. So, we used default parameters. Which classification methods works [inaudible]. So, here, random forest and deep neural network, DNN, so here out inputs need to compare the deep neural network, deep learning, we used the random forest.

So, here, we are starting to optimize random forest. We are still using random forest parameter optimization and get the result based on the [inaudible] validation. Calculate the average accuracy responding to each of the evaluation methods. So, that's the result.

We also optimized parameters deep learning. So, with the result, that's the result we got form deep learning, and then this heat graph shows [inaudible] deep learning and random forest, so it's better to file evaluation methods we can show that deep learning is [inaudible] outperform random forest.

Detail compare [inaudible] the [inaudible] class classification program. So, this is the [inaudible] have been correctly predicted by the random forest model, so that's [inaudible] percent. Deep learning model [inaudible] percent. [Inaudible]. Mistakes made by the predictive model. So, second one here, [inaudible] antagonist but random forest can only correctly identify the one. So, here, deep learning can identify, works much better.

To summarize the deep learning process. Deep learning has great potential to significantly improve the accuracy of *in silicon* predictive toxicology. The hyperparameters in deep learning must be optimized. [Inaudible] get a good result [inaudible] default parameter. We must optimize the deep learning parameters. The deep learning models, with the average or macro-average f-score of 0.83 was much better than random forest with 0.56. Both deep learning and random forest had difficulty predicting antagonist outcome correctly, but deep learning did better. So, one of the reasons that the [inaudible] small for antagonists, only 17, [inaudible] problem, so in the future, we needed to have more data instances for each class.

Discussion and future efforts. So, first, we need to generate a large benchmark data set. So, here, to try to reduce the imbalance ratio, we need to test a more active chemical in order to get a good performance. Data quality is very important. Quality of the data set, a lot of instances that are inconclusive makes a model that is confused, right? So, that affects model performance. In the first part of the work, we only considered either inactive or active. [Inaudible] active, inconclusive chemicals. The model made a mistake. It's easy for the model to make a mistake. How to improve the quality? How to get a larger benchmark data with smaller imbalance ratios? So, that's' the work we need to do in the future.

In the next one, improve the QSAR modeling to use of novel descriptors derived from molecular dynamics simulation, socking, and other information. Use more features to describe the chemicals, we expect to improve the accuracy. And in QSAR modeling, we know that. So, here, we use more than 2,000 features. That's very high dimensionality. High dimensionality is creating a problem of difficulty for model development. We may consider feature engineering, for example feature selection or feature extraction, to reduce the dimensionality. That's another possible effort in this area. Here, since we may use a different type of information, molecular descriptors and other type of information, all these different type of features [inaudible] in the prediction? Maybe not. Here, this suggest to us to develop a multi-channel deep learning framework. You can develop deep learning for each type of features and then combine the prediction's results. That's maybe another effort we can try. And the last one is ensemble methods, because our research work and other research work shows that deep learning doesn't necessarily work better in all cases. There is a need to consider ensemble methods, consider the predictions from heterogeneous, from all different types of algorithms, and then, you can assemble methods to get a final prediction based on themajority vote, weighted vote.

This is a number of references. Acknowledgements. I would like to thank a number of individuals who participated in this project. [Inaudible] from ERDC, his are of expertise is chemistry, biochemistry, he provided a lot of contributions to this project. I also would like to thank Dr. Huixiao Hong from NCTR of FDA [inaudible] model development added to research. Also, thank you to two graduate students in our lab who actually did the work I presented here. Dr. Ping Gong, [inaudible], and I edited a special issue on the front of deep learning for toxicity predictions [inaudible] journal Frontier. It's an e-book published two months ago. So you can go to Frontier's website sand download that. References and papers are available in the e-book. That's all. Thank you.

Using Machine Learning for Cosmetics and Cosmetic Ingredients
Tim Allen, University of Cambridge, Cambridge, UK

Ernest K. Kwegyir-Afful: Thank you. Next we have Dr. Tim Allen, the Research Associate in the MRC Toxicology Unit at the University of Cambridge and will talk to us about using machine learning for cosmetics and cosmetic ingredients. Tim?

Tim Allen: Thank you, Ernest, for the very kind introduction. Hello, everyone, and welcome to my talk to finish off the day. I am here today to talk a little bit about machine learning, obviously, and a little bit about the cosmetics and cosmetic ingredients safety decision process and the compositional algorithms.

I work at the University of Cambridge, but I work closely with experts at Unilever Safety and Environmental Assurance Centre. They are the ones who deploy the algorithms and making the safety decisions. I need to be very careful about how they describe their work, but it is very excellent. If you go to the next slide, I have no conflict of interest to declare. Ultimately all the research done, and all the research presented here is either published or about to be published and we are happy to share the algorithms if you want to get into the contact in the future.

On the next slide I outlined some of the objectives of my talk. I need to introduce how Unilever is applying the national risk assessment in their safety evaluation pipeline. Then we can talk about where predictive computational toxicology and machine learning fit into this pipeline. Then we can share algorithms we have been developing and compare them to each other. I will talk about why machine learning efforts are good for some of these tasks and last but not least I will come back around to the start by talking about where we're going next to some of these research efforts and how it impacts the use of the NGRA.

The next slide introduces the world of Unilever in the world of consumer products. Unilever is a project, a wide variety of different products. Cleaning product, things and home care like shampoo and face cream and personal care and the cosmetics space and food and refreshment, it’s mostly tea and ice cream in the UK but also mayonnaise and things like that so it is a wide variety of products. The products are essentially assessed and evaluated in the safety context.

The safety information generated in a toxicity evaluation is used in various ways by Unilever. There is the classification and labeling space you show at the top as shown by these symbols you might find on various barrels, etc. or if you work like me in the Department of Chemistry you will find those there. These are labeling focused toward the occupational world protecting the workers in these go to hazard assessment. This is the case of this molecule for this space.

The second way safety information is used in is in this grieving procedure and this is where new products are being developed. You might have a product and we need may be a surfactant to go into this product and maybe there are a number of molecules to potentially fit this requirement, but we are thinking about which molecule to choose and which is the best choice. The best choice most of the time is the molecule that is the least toxic. If we can use computational toxicology approaches to screen those compounds and the target compounds at this stage, we can make safety decisions that will impact which compound is chosen.

Last but not least there is the risk assessment part, and this is where it comes down to making a decision based on the specific ingredient in a certain product. This involves calculations of exposures of the product to the consumer. It has a high level of scrutiny but Unilever has external regulators and the computational approaches can be used as well as the *in vitro* approaches, but we need to have more information and a high degree of accuracy required.

On the next slide I show the NGRA framework proposed by Unilever included in this publication. This shows a pipeline of how Unilever is trying to make safety evaluation

decisions in the risk assessment category using methods, moved away from animal testing in any of its decision-making and relying on *in vitro* and *in silico* methods. This diagram runs left to right. On the left is before the exposure of the molecule is calculated for the consumer and calculations regarding [inaudible] and existing information and this is where the *in silico* predictions fit in.

After you generate a hypothesis using these calculations, you move on to do a characterization *in vitro* and calculate where the departure appears to the exposure to see if there is a margin of safety for the product. Further *in vitro* studies can be performed before a risk assessment conclusion is reached based on how the margin of safety is.

If we click to the next slide, this is a scenario, we have hazard estimate. This is quite high, and the exposure estimate is quite low, and these peaks have relatively low standard deviations or uncertainties. It shows the margin of safety is quite high in this case between those two peaks.

If you click through the next slide, this scenario is a little less ideal and where the hazard and exposure estimate are little closer and also have uncertainties and there is an orange area for there is overlap which is undesirable.

Last but not least is the principles of risk assessment without animal testing. Cosmetic regulations in Europe, no testing can be done on animals. These regulations, and also a desire to have products that are not tested on animals and do science in a better and more ethical way, have led to this iterative approach primarily focused on *in vitro* and *in silico* methods in determining the safety of these products.

Let me give you a little bit of background what Unilever wants to do with not only computational tools but different tools in the toxicity space. If we start with the computational part of the talk, this is something were normally I get to places and I never have to introduce the speakers, and somebody introduces it for me, but nobody has shown this yet. This is the adverse outcome pathway and the pathway was postulated 10 years ago at the U.S. Environmental Protection Agency in a comes out of toxicology. The idea is if you understand the effect chemicals have at different levels of organization, we can make good robust mechanistic toxicology decisions. My background is in chemistry and I work in the Department of Chemistry, so I'm very interested in the left-hand side of this diagram. This is the era we are interested in. Can we go from toxic and chemical properties to the Mac group molecular interactions, something like a [inaudible] interaction for example. In the past we have done this using several computational methodologies and one is shown on the next slide.

This is a procedural for structural alerts that are relatively straightforward making decisions in the space. In the structural alerts world we take our training data and we take data from Campbell in this case and we find a number of targets [inaudible] kind of a consortium of drug companies coming up with pharmacological receptors that if you make a new drug that hits on these targets there is an off target effect that can be really bad so we thought these were good starting points. Then we use a structural algorithm to find substructures in the training data that were common to chemicals in the algorithm would output these in small strings essentially in performance data so the new compounds could be compared to those substructures. [Inaudible] that particular

target. Overall this procedure is quite popular. The way it works is very straightforward and transparent and easy for toxicologist to use this prediction tool and look to other molecules that contain the fragment.

This shows where we came to in this research. [Inaudible] that into the structure alert database and where there is a hit, the hit is for the [inaudible] receptor and that the algorithm can go back into the training data and it can find all the molecules [inaudible] that have this structural alert and output molecular activity is a PKI value.

This shows the binary project may be a slight estimate of the quantitative activity of these molecules and we are trying to apply these approaches as well as the machine learning approaches to more targets. This slide shows where we are trying to get to in terms of the number of biological targets. This is a list of 79 biological targets that we have developed [inaudible] approaches. What we have been doing with this data, rather than just relying on data from a single data source which often causes problems as we have seen in the previous talks, if the data is imbalanced it is tougher in a computational algorithm to make the predictions whether that is based in structural alerts or deep learning. To help us overcome this we combined the Campbell data we have used in the past which is heavily dominated by active compounds with [inaudible] which is mostly negative data and the result is we can create data sets that are a little more balanced and suitable for computational predictions to be applied to.

This shows a neural network, and this is the kind of approach we have been trying to apply in the machine learning space. In the neural networks we use chemicals on the left-hand side that are normally [inaudible] chemical fingerprints and binary activity values and hidden layers in between. This is us trying to link the chemistry on one side and biology on the other. We train these networks across the targets I showed on the previous slide and the average model of performance is shown on the next slide. For our training validation and test data, which of the test data outburst and we used the validation data to model to optimize [inaudible]. The model statistics show sensitivity, sensitivity and accuracy and receiver operating characteristics. Generally, the models predict well so we get accuracies of over 90% which we are happy with. The predictors do work in this binary task. What is interesting when I built these predictors is you might expect the predictors to better when we have more data. The hunger of deep learning and how much you have acquired [inaudible] best performances.

What we see in this graph with the MCC values shown on the y-axis and the number of compounds on the X axis, in the two problematic cases we found, these cases are intrinsically challenging as modeling tasks. This graph released to the neural networks but we [inaudible] structural alerts as well so yes, there is a data requirement and when we had data sets the neural networks [inaudible] but also about how much of a difference there is between the active and inactive [inaudible] these particular tasks are so challenging. The neural networks generate very good performance on their predictions. They can actually output [inaudible] values. Because you get these numerical outputs, it can give itself a confidence score which we can't do so this graph was made for the [inaudible] receptor. Experimental negatives are shown as blue and experimental [inaudible] if you had the neural network and you put your molecule in and gave a value of say 0.59 you would be very confident the network new it was going to be active and therefore you would be happy with that prediction. If it was the value of

[inaudible] that would give you a lot less confidence. It allows the network to give you a little more information you can get from the structural alert.

One of the big questions we did this was how good are the neural networks compared to the [inaudible] already constructed? We will skip to the next few slides but if you go to the next slide, I will explain what they show. We go through the different statistical performance of the neural networks and the change between what we see in the neural network and what we saw with the [inaudible]. The squares are blue where the network is performing better in the other color is [inaudible] and generally we get quite a lot of blue on this slide which is good. You can push them forward and when there are cases [inaudible] tends to be better in most of the other cases and this might be a case for the neural network has not quite been optimized. Generally, we say more blue [inaudible] generally the neural network is the best predictor. 75% of the comparisons came out in favor of the neural network. As was alluded to earlier, the idea of these models the composition isn't always what we want. Obviously it is nice to compare them to see which ones work best [inaudible] the case of certain targets it is good to have the information so we can train better in the future and make better decisions but the way we use [inaudible] so if we have new compounds were going to use the neural network and the random forest in structural alerts and put the predictions together to try to make the best safety decisions moving forward.

Another thing alluded to earlier by Steve was the algorithm confidence and toxicology; if we make that predictions it could the to bad adverse outcomes, so we don't want to do that. If we go to the next slide, we think about how we could investigate networks and a little bit more about how they think. We decided all of the neurons or this, are placeholders for numbers so once the network is trained [inaudible] are magical relationships. It means when you put the molecule in, all of those different boxes have numbers and extracted those numbers and we treat them as a vector solution molecule that goes in has its own vector and we can compare those vectors to each other to see which ones are similar. The idea being when the vectors are similar in the network is thinking about the molecules in the same way therefore maybe these molecules will be suitable for [inaudible] to support safety decision-making.

This shows the molecule which was not in the training set for this particular neural network which is meant to be [inaudible] molecule this active at, and at the molecules in the training we have got the five most similar by the neural network activation similarity. There with got the five most similar [inaudible] algorithm. The one that has the highest chemical similarity is this one on the left-hand side with the exception of the [inaudible] tends to have a lot of structural similarity. The similarity is only 0.275. This is the only molecule that appears to have the structural similarity to the target chemical.

If we click the next slide we see [inaudible] experimentally active in at least three of the five are also active whereas four chemicals are active [inaudible]. If we kick forward one more time, I found it was hypothesized the reason, is active is it has an aromatic ring as shown in the Blue Square. What is interesting among the similar chemicals is this motif has only essentially been picked up twice where's the neural network has learned this because it is featured a lot more common amongst the five most neural network similar compounds. This is really promising that the network is learning without some of the things that make compounds hERG active.

Everything I have been talking to you about has been about binary predictions or classification tasks. Is that molecule active or not active? This really isn't good enough for the risk assessment decision-making because we need to understand quantitative information about our prediction. Luckily machine learning can help with this. We just change our algorithm, so I adjusted my data set to contain only quantitative values and I got almost 5000 values. I had to change, which was appropriate, we can see the output graph from this predictive task. We have experimental activity on the y-axis and the points generally seem to cluster around the dark blue diet no and the root is the square arrow across all of these examples #the algorithm is able to do some predictions here that are pretty good.

If you consider the distribution of this data, if you click through to the next slide, I have broken this graph down on the y-axis by 10% chunks of the data. Each of these bars is a different 10% chunk of this test set. It shows about 50% of the data in the experimental activity range and making the predictions in this area are significantly easier so that is why they have a low RMSE where the RMSE is a lot higher. That makes it much more difficult for the algorithm. We are looking at how we can put a target through procedures like this to make these predictions and how we can slightly better balance this algorithm, so it does a little better prediction algorithm.

That leads me to summarize on the next slide. Neuron networks can provide binary and quantitative predictions. We have used structural alerts, random forests and networks together to predict binary activity. A combination of these models and the understanding of the workings is key to gain the highest model performance in convincing a toxicologist and regulators the use and toxicology decision-making which is key in this space because we can build models with high-performance all day long but if they are not being used in the real world they are not having the impact and not going to move toxicology forward. Less released, quantitative predictions will be the next step for the conversational toxicology in move you closer to the risk assessment space where we can compare values to exposure rather than just hazard classification.

I have included some references, not too many. First the paper about the adverse outcome pathway. Maria's paper which talks at length about the NGRA process being used. And then two of our most recent papers where we talked about structural alerts and the random forest models and how we can put them together to build a consensus for predictions. That just leads me to thank these people on my acknowledgments slide. [Inaudible] was a PhD student and work on [inaudible] the organizers for putting on this event, it has been really insightful, and I really enjoyed it. It is great to see we are moving these events online so were not missing opportunities to event [inaudible] thank you so much and I guess we're going to move on next to the questions. If you have questions, I would be happy to try to answer any of them on anything I have presented.

Roundtable Discussion
Moderator: Jim Riviere, 1DATA Consortium
All speakers
Additional Panelist: Ernest K. Kwegyir-Afful, US FDA CFSAN

Ernest K. Kwegyir-Afful: Thank you. Jim, do you want to take over for the Roundtable?

Jim Riviere: Sure. Again if people have questions the best way to do it is on the chat box to everyone. Ernest, you had questions from people from FDA, do you want ask those?

Kwegyir-Afful: I have a couple questions, this question is for Joe. It says in the summer slide you said deep learning has great potential to significantly improve the accuracy of *in silico* predictive toxicology, can you explain why deep learning would improve the accuracy?

Chaoyang (Joe) Zhang: This is based on the results from other researchers and our own results. Deep tox, the team who developed deep tox, won seven out of 12 of Tox21 data challenges. So, that's number 1. The second is to systematically assess the performance of deep and shallow learning algorithms. They show a result, not a conclusion, but a [inaudible] deep learning [inaudible] of a performance. So, our results shows if we customize the deep learning type of parameters, so there is a predictor to get a good performance. Now, for some of the cases, performance is not as satisfactory because it's limited by the imbalanced data. Data distribution and data characteristics can affect the performance. So, for example, the last few slides I showed the results for an antagonist case, deep learning only identified [inaudible] percent. That's because the [inaudible] there is only 17 instances in that class. That might be a problem. That's an imbalance problem we need to handle. So, from my point of view, in my opinion, it is not because of the performance of the deep learning. It's the data characteristics.

Kwegyir-Afful: Thank you very much, Joe. Jim, do you want to do the questions in the chat, or should I finish with the FDA questions?

Riviere: Finish the FDA questions and then we can address these.

Kwegyir-Afful: This question, Jim, is for you. You discussed your work on the metrics that could be used to characterize nanoparticles by improving the surface of the nanoparticles. Can you discuss how the [inaudible] interface [inaudible] is considered as an input into the algorithm?

Riviere: Repeat the last part of that question?

Kwegyir-Afful: Can you discuss how the [inaudible] interface [inaudible] is considered as an input into the algorithm?

Riviere: The descriptors we were using were under defined which actually pro the aqueous nanomaterial interaction. We've also done work to predict the corona from that and that medium dominates everything. The problem with the whole area of nanoparticles is you can take very well-defined conditions and map the surface and you can get an idea of what type of proteins might bind, but then there is so many types of proteins and carbohydrates *in vivo* that can bind to it. Again, it is a question of getting that data under proper conditions. In this case we really don't have enough data to work that out. You were starting to get some idea that surface interaction is really a function of not just the constituents on the service but how it interacts with the nanoparticles. If

you take a simpler aspect from a tox response and follow that one through, it completely depends on response and what kind of media you use in an *in vitro* assay and a number of work, including our group, has shown that if you put a cell culture media looking at rats, you're going to get protein coronas [inaudible] nanoparticle, but the cell culture media you're going to predict *in vivo*, it is a different situation.

Kwegyir-Afful: The next question is something you mentioned but I think something also the other speakers mentioned so anybody can take this one. You mentioned specific problems including incomplete or unbalanced data sets. Could you describe what you would consider a data set to be unbalanced? Jim or anybody else.

Tim Allen: I think saying with the did is balanced is asking how long a piece of string is. The ideal is to not have data for you have 10% of one type, for example, a classification class you have 10% active and 90% inactive, that obviously is less than ideal, but saying when is it balanced, I don't think it has to be exactly 50/50. There are certain mathematical balancing algorithms that can be used to certainly overcome, maybe more like a 60/40 data set. Just kind of away from those really unbalanced 95 to 5 is a good starting point at least.

Steve Bennett: This is Steve, a comment about maybe the broader question of the use of data sets in training. A lot of the cutting-edge computer science you read in the field is really focused on the training data approaches in how you maximize sparse or small data sets so that they are maximally effective when training machine learning problems, and then kind of a newer set of techniques coming out about how to synthetically construct applicable training data when you have little to no training data such that it meets the criteria you need. Lots of interesting work in the computer science side in how you construct data sets that are effective for training without having to have as much as we traditionally thought we needed.

Nicholas Watson: I was going to say the second comment was about creating synthetic data to augment your existing data set, but the first part is understanding the challenge areas in your data sets. For example in our cleaning project, we knew we had a classifier, which was dirty or clean, when we collected data through the cleaning process, if that cleaning process took 10 minutes we would keep recording data for another 10 minutes when the pipe work was clean, because then we know we had about half was dirty and half of it was clean. When we did the regression work, we only wanted data predicting when there was an amount of fouling on the equipment, so we didn't include any of the data when it was clean. Just understanding your model and making sure you're not unbalancing it as well.

Riviere: Do we want to take questions from the audience now? I'll just take them in the order they came in. The first is for Nik. Is [inaudible] and aspiration of the chemical drug manufacturing?

Watson: Shortest answer ever. Yes. We just worked in food and drink because that's where the majority of companies were collaborating or are intersected. The sensor itself doesn't discriminate. It is the humans that decide what applications [inaudible] on the wall, whether you eat that or brush your teeth with it, for example. We have had discussions, and there's work going to be happening in that area.

Riviere: This is back to the panel, what one development does the panel think would be a game changer in their respective fields? Steve and then down the list in order, I guess.

Bennett: I saw this in the chat window and have been thinking about it. I think quite a bit of the time how we can make this computational techniques useful to help serve citizens, help governments do a better job of serving the people they work for, and for me so many of these models are still black box and it is difficult to unroll them and the explainability and traceability and why you get an answer you get out of these models. So I think a game changer would be, and there is a lot of research ongoing in this area, a game changer for me would be a quantum leap in the explainability and traceability in these things which would make them easier for us to use in government where we have to know why we are getting the answer we are getting.

Riviere: I agree completely. What I think has become a game changer, at least in the field I'm looking at, is the ability of open data sets and sharing data on what is in real time. As more and more if this becomes available it is much easier to start pulling in data that otherwise would not have been possible. In the old days where data was really protected, I think we are seeing a more openness and hopefully that continues.

Watson: I was going to say I think it is the business model that needs to be looked at because you still have [inaudible] companies who say, we're an AI company, we're a data company, but if you look at their business models, they're actually just trying to sell you something which is a sensor, and really, if you're going to combine sensors with machine learning, you need an end-to-end solution which involves how you collect the data and train the models and retrain the models, If required. So, I think the the jump that I think needs to happen is in the actual business models.

Riviere: Joe?

Zhang: There's a question for me, right?

Riviere: I'm just trying to get everybody to address what the game changer would be.

Zhang: Okay. Could you repeat the question?

Riviere: What one development do you think would be a game changer in your field?

Zhang: Okay. It is very difficult to comment because so far deep learning has been widely used, but for QSAR modeling, learning, don't use data collaboratively, don't understand that deep learning, those are averages [inaudible] not able to get a good result. That's what we got from our research. It's very difficult to say. I think the first [inaudible] high quantity data because deep learning requires a large data set [inaudible] address that imbalance problem. So far, the Tox21 data set imbalance ratio is pretty high. [Inaudible] assess the performance of different algorithms. We might have a benchmark data set with a relatively small imbalance data, but that is the first step, and then, we can assess the performance of different algorithms. And another issue is the criterion, I already discussed that in my talk, is the evaluation criteria. Which one is best? If we only use an area under ROC, that's not, I don't want to say [inaudible], so at least we need to consider other evaluation methods, because

AUROC, that's the true positive and false positive. So, with imbalanced data, you may still get a very high AUROC value, but actually, there's a lot of false positives. And we pay a lot more attention [inaudible] reduce the false positives, or, false negatives, false negatives. But the first step, game changing, we do need to have a project, a joint effort, Tox21 data challenge create a large data set with more active chemical compounds.

Riviere: Okay. Tim, any comments?

Allen: I think Joe's point on the data is very valid. I think high-throughput screening is going to make a big difference [inaudible] capacity to give us data sets that are more balanced, some that do have consistent experiments shall we say? In terms of the application of the methods, regulatory acceptance will be huge, but I think [inaudible] and I think AI is really interesting as a field because a lot of the machine learning algorithms we talked about the day were essentially conceived, as has been pointed out earlier, about 50 or 60 or 70 years ago. And what it means is that actually, we're kind of using almost old mathematics but with modern data and modern computing power and I think if AI is going to take that job forward where it needs to go, toward more of a general intelligence, there is going to need to be some kind of development on the algorithmic side, because the current algorithms just aren't sophisticated enough.

Riviere: I agree completely. It's amazing how far some of these were developed.

Allen: Looking back at all that time.

Riviere: The next question is actually to you, and you touched upon it. What can we do to gain regulatory acceptance for such methodologies across different countries especially the EU?

Allen: Give me a magic wand. The regulatory question is huge. As much as we can build the model and we can get involved with industrial partners, we can see the models being applied in decision-making, the regulatory space is obviously entirely different. There is kind of normalization in the regulatory space where animal methods are not perfect, but they are very good, but they are accepted. We need to try to break this down. What Unilever has done is really positive. We need to accept in the computational and AI space were not going to be replace an animal method with one [inaudible] method, it's about putting them together, putting them with other tests in these case studies and presenting them to regulators and industry and industrial companies working together to build these frameworks together to essentially push them as a group so we can get weight behind them. And I think, ultimately, computational approaches have been used before, because read-across exists and has existed for a while. It is kind of about trying to put the methods we have now on the parallel with that and saying, here's the read-across and the evidence surrounding that and trying to normalize the idea that the data can be generated by AIs or by computational methods, and presenting it to regulators so that they become more familiar with that as well.

Riviere: Very good, thank you.

Allen: I think it's going to be very tricky.

Riviere: To continue questions from the audience, the next one is general to everyone, recent published work has discussed to define the so-called applicability domain and assess uncertainty estimates for individual predictions. Does anyone have any recommendations on this matter?

Allen: I'm happy to go again. I don't want to feel like I'm holding the mic. This question was submitted by [Inaudible] who used to work at [inaudible], so hi to him. Obviously, applicability domains are huge. And I think in the QSAR space, applicability domains are still a standard practice of publishing a QSAR paper. But I think as well move into the machine learning space, maybe it's become slightly less of a priority compared to building the biggest and most advanced algorithms, the largest data sets, and trying to get them all to work together. It's important we don't lose sight of all the stuff we learned we were doing QSAR in the past, because applicability domains are still very important and when chemicals come from a different space than the training data, obviously there is a much higher potential for an incorrect prediction by a machine learning algorithm. They do need to be factored into this, and whether they're done for things like chemical similarity or input similarity, etc., maybe that is kind of a way to build them in.

Riviere: I agree completely. You can get into a lot of trouble if you're taking a model from a different compound and again I have seen a lot of that on even simple approaches of dermal absorption for compounds that are using a model that is basically absorption from water and the vehicle [inaudible] and it is different interaction [inaudible].

I'm going to jump in to answer the last question because it really goes into this applicability domain. Someone didn't understand the comment that AI was a statistical inference in a traditional sense if applied to random sample from a population, why wouldn't it be? My answer to that is most the time AI is not applied to a random sample from a population. If you do get the random sample and you sure it is truly a random sample you can get some level of statistical inference that it's matching it but in general it is not applied to random sample, and discussions you just had on applicability domain or compounds that show toxicity versus compounds that don't, there has to be a lot of work on developing to a point that you really have an applicable sample. That was the only thing I made on that comment. Anyone else have input into applicability domains?

If not, the next question, I'm interested in the use of structured metadata ontologies swabs that were collected during inspection for infectious disease control of [inaudible] drug manufacturing to support AI against the whole genome sequence data from pathogens. Do any of the panel have any recommendations to support this line of inquiry? This is out of my area of expertise. I have no idea. Anyone? Well, whoever asked it, sorry, we don't have the right people here to address this and that is better than addressing it wrong.

I guess the next statement on this chat [inaudible] random pathogens were arising during production, that must be associated with that.

The next one, the model is only as good as the data behind it. Since there may be different ways to interpret the underlying data, are there any efforts in the tox world

across agencies to collectively review and consolidate the data sets so they can have been universally applied to the models we are building? Again, I think Tim or Joe could take this.

Zhang: Okay, I'll try to answer this question. [Inaudible] it's true, to model, to do the model, it's based on [inaudible] as I mentioned previously. So, we should know [inaudible] this kind of benchmark data is good enough to assess different [inaudible] future work in this community. Do we need high-quality data, the lower imbalance ratio that's [inaudible] different predictive model.

Riviere: Anyone else?
Allen: I think there are several problems in the tox world from different areas of industry. I think the industry of products and stuff, this is also true in the food industry. The lack of data or framing any kind of model makes this really challenging and I think a lot of the companies that work in this space are quite willing to share the data they have but is very, very biased toward negative data points and even the data sets are quite small.

At the other end of the spectrum we have the pharmaceutical industry where you have loads of compounds being developed by loads of different companies, but they are all proprietary and they don't necessarily want to share this compounds with each other because they don't want other people knowing what they're working on. And getting that kind of data set would be valuable for AI. And I know that Lhasa, which is a company based in Leeds, in the UK, has been doing that for a while as essentially acting as a broker that holds onto data for various companies and uses it to build models together. So, I think that is definitely a strategy that could be used, but I don't' even know if that would be enough in the food and cosmetic spaces. We just have to take the data that is available through the open-source databases and obviously through the Tox21 or the ToxCast database to do as much modeling as we can. I don't think we will be seeing a lot more kind of Tox21-style studies coming out given the current economic situation. So, it might be a while before we get those big datasets again.

Riviere: Another question just came in structural activity based predictive toxicology, can this be applied to biologics or is it already being done?

Allen: It can of the data exists. The reason we target small molecules as a first path is because there is a lot of data on small molecules, like [inaudible], ToxCast, these databases exist, you have lots of data points. If you're talking about therapies like genome-editing therapy, when you're talking about therapies like [inaudible] protein or something like that [inaudible] data, and even then, it seems very, very complex. I know AI is being used in protein-folding space and stuff like that. And that's a really large challenge, because there are so many ways the proteins are folded. Their 3D structure obviously can completely change the way in which they operate, so even going from sequences to 3D structures with proteins, it's a really, really big challenge for AI that's not quite there yet for that one, so.

Riviere: I would say that is the exact same with nano. Another question, we're often confronted with small data sets, are there any metrics or rules of thumb to know when a

date is too small to apply AI? Is there any consensus on making a judgment in regard to this?

Allen: In the past I attempt to try it and if it doesn't work it is probably because the data set's is too small. You can tell quickly if you're able to get anywhere. Sometimes you press the train button and you sit back, and you wait for it to train, and you look at it [inaudible] that was never going to work. When you've only got 200 compounds and the performance hasn't gone anywhere. I don't know about anyone else, but that's what I've found.

Watson: I was going to say we look at the problems [inaudible] we kind of train it, and then we take a bit of the data away and train it again because what you're looking for is convergence if you keep adding data you get to a point where you model performance does it really improve that much and you can kind of work back. Like Tim was saying, if you get a bad result that was probably the reason. If you're taking away [inaudible] try again, and it's getting worse, probably giving you an indication. I think also if you look at some of the models and things like deep learning needs a lot more data than some of the more of the simple learning but you can go to the resources combine and [inaudible] which models that formed their own data.

Allen: Nik's absolutely right. You might find if you go straight to the biggest deepest network you can make, and then you try to make it train, it just won't work with the data, but actually, you've dialed it back a little bit using [inaudible]. You can tend to find that maybe it's a little less sophisticated algorithm, and will be able to do a better modeling, a better job modeling, and it will be faster as well. So, there's always that advantage, too.

Bennett: A couple of comments from the theoretical side, there is a lot you can look into in the statistical learning theory around different training data sizes, there is these dimensions they give you the complexity of a model and you can then figure out the appropriate training size based on that VC dimension, how complex the model is. There are things you can do with statistical learning to get a sense of whether or not your data set, your training size is good enough and then there are things you can do to try it and watch how your performance levels off as your training data grows. A couple things there, a couple on the theoretical and practical side.

Riviere: In the QSAR work there all of these statistics, leave out 10%, leave out 25%, and just run through the data over and over to see if it is balanced or just not converging on anything. But it really depends on the domain you are trying to predict.

Zhang: I want to add a comment on this. When we talk about the data size, not only [inaudible], we also need to consider the feature space, that's the dimensionality. For example, you have a data set, you have 1,000 features, just like QSAR model, and 2,000 descriptors to describe the structure. In this case we do need a large data set [inaudible] we have so many unknowns, the parameters, the model. [Inaudible] model to optimize the parameters. So, the data size, related to dimensionality, when we consider the data set, what size of data is appropriate for machine learning algorithm, we do need to consider the feature space, how many features we have. For example, [inaudible] maybe 200 patients' data is enough for you to build a very accurate model, so we did in the past. But for QSAR, that is not going to work. For QSAR, the

dimensionality is very, very high. Thousands. So, in this case, we do need a large data set, including the minority data. They are related. [Inaudible] are related. I wanted to add a comment on this.

Riviere: I was going to ask you a question when you're using 2,000 descriptors, I have seen people submit things to see if they can predict 100 compounds.
Well yeah. So, I think in that line, there are some statistical approaches with how many parameters, how many degrees of freedom when estimating this, and what so you really need? Part of that question is for some data sets, yes. But when the endpoints get more nebulous to predict, it is tricky to mail that number down.

Kwegyir-Afful: Jim, this is Ernest, I had a question come in for Steve. Somebody wanted to know if it's possible to develop a model for foodborne disease outbreak if you have limited data? Do they have any chance of developing a model? What kind of machine learning models that would allow them to track and detect foodborne disease outbreaks?

Bennett: Thanks for that. First, I don't know, there could be work developing machine learning-based models for food borne disease surveillance. I'm not as up to date on the literature and what's going on in government as I used to be a couple years ago. But I would certainly think it is possible. One of the things, unfortunately, that we have a lot of foodborne disease outbreaks. There was quite a bit of data the community was pulling together every around foodborne disease outbreaks. You might have enough to extract some key predictive features that can be used to be predictive. As an aside, whenever you build models like this, we want to orient them for action, particularly something like this surveillance model. The action might be better placed in looking at risk, where are the supply chain model next to the place we suspect there might be enhanced risk for contamination. Ernest, some of the things that you work on. Just a long way of saying I don't know any, but I think it is something that could be done. As I understand it there are enough historical foodborne disease outbreaks to be able to build something to try it.

Watson: I think I have seen some work in this area in the UK, but I don't want to quote any names until I check but if someone sends me an email, I can definitely have a look and provide some information.

Riviere: Ernest, any more questions from FDA?

Kwegyir-Afful: Yes, there was one for Tim, I need to scroll down a little bit. Hold on a second. Tim, this is about [inaudible] machine learning algorithms to be used as a risk assessment instead of [inaudible] one of the conclusions. Can you state what you believe is the primary issue that is preventing these [inaudible] predictions from being used as a risk assessment?

Allen: At the moment, almost all the models we have constructed have been classifiers which I pointed out are useful in a safety decision-making context but they're not capable of being entered into a risk assessment. We need quantitative information, which is why we've moved on to the regression algorithms. The one I showed you is one of only a handful that I've trained up to this point. I think it shows a lot of promise and it shows there is potential for these algorithms to be used in that way. These

algorithms predict quantitative activity at specific MIEs, which is really great. But even if you can look at that from the AOP perspective, we know that activating an MIE is not necessarily the same as causing an adverse outcome so we have to think about what happens next, maybe through what we would call a quantitative adverse outcome pathway and what's going to happen after that [inaudible] and how does that relate to something more like a point of departure value, and which would be developed through the next generation risk assessment. And also, what happens before and how we go from molecules and exposures into the cells as well. A little bit more around that space but the prediction itself is not enough to be put into a box and say this is the answer, unfortunately. But we are getting closer and I think moving into the quantitative space is the right option and then after that it will come down to things like QAOPs. And there is a lot of computational capacity for predicting things like exposures and concentrations, so maybe that side is a little but more answered than the AOP side. But the question really is what happens when you expose someone to a [inaudible] active androgen receptor binder.

Riviere: Thank you. Ernest?

Kwegyir-Afful: I think that is the questions I have.

Riviere: I have a couple of questions that are pretty general, and if we don't get any more, I think we're all set to wrap up. This one is to Steve. You mention the problem is getting groups to adopt some AI platforms and decision-makers don't understand it. This happens on bench level scientists, trying to get them to work together and essentially, they publish in different journals and don't necessarily read the journals. In your experience, how do you [inaudible] bridge this?

Bennett: If I can share a story, I will share you a story of something I did completely wrong very early in my career and I learned a lesson from it. Very early in my career and homeland security I worked in the Science and Technology Directorate and we were creating a large-scale computational simulation to be able to [inaudible] different risks and I was working in bioterrorism and biological security and we had different risks and we were trying to rank them to figure out where we needed to prioritize efforts.

We then expanded that to all of homeland security's mission. If you had a marginal dollar where would it be best spent? In hurricane preparedness, in hiring another border patrol officer or building codes, all these different mission areas, where is that marginal dollar best spent? A very hard problem in government. We wanted to quantitate that by the risk. So, we spent two years building a very good simulation, we had gotten our [inaudible] right, have a better way to communicate it and we go to brief this to senior people. Somebody said, Dr. Bennett, that was a great presentation and I have learned my career when someone says that, I am in trouble. He said, in my agency, we just trust our gut. All of that two years of work, it didn't go down the drain, there was [inaudible] but we didn't get the decision impact in the agencies we wanted.

So, I learned two lessons from that. One is good old-fashioned stakeholder management. We should've had that guy and his staff in our office two years earlier as we designed the plan and built the experiments and built out the is stimulation assumptions and getting them involved because at the end, we wanted them to trust it and use it. The second lesson I learned is nothing beats an example. Pick something

small and demonstrate a so what? We work with a lot of cops and people who don't care about data science. If you want to show them some of these tools can help them one way to do it is say, you do what you would normally do for your investigative process, but share your data with us and let us come alongside you for six months and let's compare notes at the end of six months. And we will show you what you would've had along the way over the last six months if you decide if you that would have been of value to you. And invariably we can get some of the most skeptical operators to become quite open to these things if you can show the value. So, two things. Bring people along from the beginning. Don't do your work in the closet in the [inaudible] bench and come out when you're done; bring people along. And second, pick an example and start small. That was kind of long-winded, but I learned that lesson the hard way a couple of times.

Riviere: It is really difficult. You need to get early, and I think even in an academic perspective, trying to get people working in different areas but you know that they could talk together and work together, it just takes time. Thank you.

I have got one question for Nik. In your experience, how much resistance have you encountered in actually deploying these types of approaches on the factory floor?

Watson: This is really good [inaudible] and I think it's [inaudible]. We have done a number of projects in different areas and you have to get people engaged and you have to almost want them to come to you and say, we have got a problem and what can you do about it? Because if you try to send it to them, then they're kind of like, maybe it's not a priority, so I think it works best when they come with a problem. The challenge we find with industry [inaudible]. I think the food sector wants the solution to its problem and it wants that solution tomorrow and with these tech allergies it takes time to figure out where you need to take a measurement, how you're going to do it, then analyzing it, building the data. So, you know, you're not looking at one or two weeks, you're looking at year, and I don't think they appreciate sometimes the time that research takes, especially some of the companies that we've been engaging with.

The other concern which stops a lot of this is concerns around who owns the data and who might get to see it and what the data might be used for. I don't think that is unique to the food sector but if you think of an example on a film, the mum and dad go away and they tell their teenaged son or daughter no house parties and then the film continues and they all have this wonderful house party and just in the nick of time, they clean the house and no one knew about it and it's fine. But that would be different if the family had cameras in every room and monitoring them all the time. I think that is kind of some of the concerns some organizations have about collecting data related to hygiene and safety [inaudible] people being able to access that [inaudible]. There is some projects going on in the UK about data trust and getting people together and that leads in to what we said about sharing data. There lots of questions and concerns there, and I think that's one of the hurdles that we need to over. It's just about [inaudible].

Riviere: We just got another question, what is the ongoing work on combining the machine learning AI approaches to model-based approaches such as metabolic *in silico* models? Anyone? Tim?

Allen: I haven't worked myself on the metabolic prediction tasks. I know there have been in the past a number of tools built in the traditional QSAR way regarding the metabolic outcomes of molecules. I'm sure there must be an AI, someone must have tried to build an AI to predict those. I know that prediction task is extremely challenging, and maybe for that reason it is a good a potential area to apply machine learning. The methods that were produced that were more traditional struggled because of the complexity of the number of different metabolites that you could get. You'd put a molecule in, and you would get 50 different molecules out and that doesn't really help in safety evaluations. And then you've got to think about all those different molecules and how they work. A little bit more targeted, I think, where you could get maybe the top three or top five metabolic profiles out, potentially, that would be the most useful way to build an AI this space in terms of safety evaluations. I think it probably does exist; I just haven't done it.

Riviere: I'm not aware, either. I think we are all set. No more questions coming in, you have answered all the questions from FDA. I wanted to thank everyone who listened, and I want to thank all the speakers. This really tied together, you always get scared that you going to come up with conflicting talks. This is worked out very well, so I appreciate it. I got to close up some stuff, so next slide please.

There other colloquia in these series coming up, so make sure you watch with the announcement of these topics. Also, to let you know the colloquia materials, recordings, and slides, and captioning text are archived and can be found on the Society of Toxicology website.

Thank you for your participation. We appreciate your input and you will be sent a link and we'd appreciate you completing the survey. I'm not sure if everyone was aware but originally this was supposed to be handled at FDA in person and with everything happening obviously with travel and COVID, it went completely to a webinar. From this perspective, it seems to have worked fine. Again, thanks a lot, and that's the end.

Watson: Thank you, everyone. Take care.

Allen: Thank you so much. Stay safe, everyone!

Kwegyir-Afful: Alright. Bye.