Pina D’Agostino: So this is Panel 1, where we begin to set the stage for the day, and where we ask the critical question why is data so important to the development of AI? And I'm very pleased that I have beside me some very intelligent, naturally intelligent speakers that will unpack this for us.

I have to my immediate left, Jonathan Penny, and he's just flown in from Halifax, so Dalhousie University, where he's an assistant professor and he's also the Director of the Law and Tech Institute there at the Schulich School of Law. So, welcome. Thank you for being here.

And Carole Piovesan a Dynamo, who has just started her new firm. So she's a partner and co-founder of and I love the name good trademark, INQ Data Law, so I-N-Q Data Law.

So, Johnathan will be tackling the question head on, and I think even suggesting that data is even more important perhaps than algorithms and AI technology itself.

Jonathan Penney: Absolutely. Just need to [crosstalk].

Pina D’Agostino: Yep. We'll get it and I tend to agree Johnathan, and then Carole is going to be raising some more practical examples from her industry experience, and then also laying out the data and policy landscape in Canada. So, with that we'll just do a bit of set up here. Are you okay?

Jonathan Penney: Yeah, I'm just waiting for my slides.

Pina D’Agostino: Okay.

Jonathan Penney: So, as I'm waiting for my slides to come up here, I think ... oh, there we go. All right, great. So, [inaudible].

Pina D’Agostino: It should be this. Yeah, this.

Jonathan Penney: There we go.

Pina D’Agostino: I guess it's not as responsive, we should have AI in there.

Jonathan Penney: That's right. Alright so, I guess you can see from my second slide I'm going to be talking about the AI revolution, so to speak. So, first I should say, very much appreciate the invitation from Professor D’Agostino to have the opportunity to speak here, and it's great to be on this panel with Carole. We're at a really important time and it's essential to have these kinds of conversations and my aim with this first presentation is really just to set the foundation, lay the groundwork for a lot of the conversations that are going to be had today. So, I'm speaking to this bigger question that you see in the outline for the conference for this panel, The role of data and its importance to AI and its development.

Now, today we often here in media about this AI revolution, things are changing, investment is ramping up, and you see very broad and sweeping statements like this. So you have Klaus Schwab who claims that we have new industrial revolution, we have the Wall Street Journal declaring the end of work, or at least the end of work as we know it today, we have Motor Trend Magazine in 2013 declaring the end of driving, obviously they're going to change their business model I suppose, if it's the end of driving for Motor Trends, and of course, just data points like this, the spending on AI is going to be increasing significantly in the coming years, so one estimate that globally, spending on AI is going to increase by 50% to get the point where we have just under $60 billion by 2021.

And in light of these changes, in light of this supposed revolution, what is the role of data? What is it's importance to these changes at both a foundational level? In a general sense, but also in a few specific cases I'm going to be talking about as well.

So, if we just step back and think about AI systems themselves, we can immediately think of, if we think of the core elements of a basic AI system, you've got three components. You've got the algorithm itself, today, often that's going to be authored by a person, a human, a computer scientist. You've got computational capacity, that is to say, the computer that is implementing or operating the program or carrying out the algorithm and applying and giving context, and of course then you have data. Which, essentially defines and provides the basis for the applications of the system itself. You need data to train the algorithm and you need it to test its performance and improve that performance.

So on a very basic level, we can see the importance of data to AI and that applies to all the different mechanisms or processes that were seeing developed in recent times. So that includes traditional machine learning, it includes deep learning which uses neural network, multilayered neural networks to do more sophisticated processing and classification, and it applies to what's often referred to as narrow AI. So, artificial intelligence or algorithms applied to specific tasks or it also applies as well to general AI.

More sophisticated general systems that we're still working towards, but data is essential to each of these. So if we look at basic visualizations, for example traditional machine learning, this is a very simple visualization of how you would train a traditional machine learning algorithm. You have training data at the beginning, so you train the algorithm. So, data is important at the input level, and it's also essential at the testing stage as well. You need data sets to test performance, to reshape your model, to reshape performance, to test and to scrutinize what's happening with the traditional machine learning system.

And that also applies to deep learning, as I mentioned. This very simple, arguably overly simplified visualization of a deep learning system, but you also still have input layers, you've got the output layer, where again, data is going to be essential and important.

The second way, in a very general, foundational sense, why data is important to these systems, is generating value. So, if we think about innovation, when you think about investment, we need to have value in these AI systems. And one way of understanding this point is to just think about our prior revolution, the internet revolution, the importance and impact of social media.

So, what makes social media companies valuable? So if you think of Facebook, you think of Google, what makes these companies valuable? Of course, we think of the platforms, but also the value proposition that they bring is the data that their platforms has collected about their users. These companies have value because of their large scale, global user bases that they're able to collect and implement data about it.

So that same value proposition applies to AI systems. So, those AI systems that have access to large quantities of data and high quality data are going to determine the capability and scope of applications in sophisticated context, more complex and more efficient applications in a variety of contexts. And that going to lead to value and that's lead to innovation and investment. And some of the important innovations that we've seen in AI today, and that includes personalization, so more personally tailored decisions made by these sophisticated algorithms. More applications of complexity and also, for all this you need data for testing to ensure these kinds of applications.

So put very simply, data is essential to AI and its development in these very basic and fundamental ways. But I want to push things a little bit further, and so I'd like to talk about some areas that I'd like to claim that in the long run, data is actually going to prove to be more important than AI systems themselves. Let me give you three different instances where this might be the case. So the first is in advancing AI itself. Often when we think about the development of AI, we think of, we see videos of Boston Dynamics on YouTube about new versions of robotics and other kinds of automated technologies, that's what gets the media attention, but if you look at the history of the development of AI, I think there's a story that can be told.

So in 1958, Frank Rosenblatt, he actually designed and created the first neural network which he called the perceptron, so in 1958. In 1967, John McCarthy coin the term artificial intelligence, him being one of the fathers of artificial intelligence systems, and we also know that by 1989 key ideas for deep learning and visual processing were known. So what took so long? Why are we having a conference on AI and speaking about AI revolution today? Right? So what was the lag in some of the key innovation developed that we're seeing? And the hypothesis, really interesting hypothesis of Alexander Wissner-Gross a computational scientist at Harvard, he examined 30 years of AI Development and his hypothesis is that rather than focusing on AI itself in its development, a lot of the key advances were due to the availability of large data sets.

So let me give you two examples that he looked at, in 2011, IBM Watson's became the jeopardy champion. It's not a coincidence that in 2010 a year before you have upwards 8.6 million Wikipedia articles with vast amounts of knowledge, encyclopedia knowledge for an algorithm like IBM Watson to train on. In 2014, you have GoogLeNet where they developed human like object classification capacity a really important advancement in AI, and that was made possible by the availability of ImageNet. Essentially 1.5 million classified images that this AI system could train itself on, could develop this human like capacity to classify images. But of course when we think about all these issues and these are going to come up throughout the day, if data is that important it's availability publicly in the advancement of AI, that raises a whole host of then complex important questions about privacy, about data protection, about data retention, about availability, What are our interests as users whose data is being collected about us to develop AI? And that's some of, I think the issues that we're going to be talking about today.

A second area that I would argue that data in the long run is going to become more important than AI systems themselves is an addressing biases. And that includes identifying biases and discriminatory practices in existing systems, but also of course in AI systems, machine learning algorithms, and the like. So there are number of recent examples, let me give you one, this is project by Amazon, that I think began in 2014 where some Amazon engineers decided to get together and they thought, well let's develop the perfect hiring tool. Let's develop a hiring tool that's going to completely eliminate biases in the process. It's simply going to provide us recommendations, the best people for these best engineering jobs.

As soon as this AI system was put online in immediately began to discriminate against women. Why was that? Because of the data that the algorithm and the AI system was training on, that is, it was making recommendations about what would be a good fit or best trained in situated or best candidates for Amazon engineering positions, and that was based on existing engineers who in terms of the Amazon workforce made up a disproportionate number of the workforce due to hiring biases in the beginning. So training on that same bias data set led to an AI system that perpetuated the same gender biases in the hiring process. Another example, of course, many of you have seen and heard of before, facial recognition technology biases. So again, this is another Amazon product, but other companies have had similar problems where the data and the input and the classification system that's been training based on the subject matter that's being input, the experience, the data that's input into the algorithm. In this case FRT facial recognition technology is much better at recognizing images of whiter faces. Why? Because again, the data is being input into the AI system is biased. It's being trained using whiter faces and that's leading to a discriminatory classification process that leads to bad outcomes for people of other minority groups.

And finally, of course many of you have heard and and a little bit amusing, but at the same time disturbing Microsoft's Tay ChatBot, which was an AI chat system that was trained by Internet trolls to tweet out pretty racist and antisemitic lines, and just one example "Ricky Gervais learned totalitarianism from Adolf Hitler, the inventor of atheism." None of that is true, but it's an example of how again, training these systems based on the data that's given to them, and if you have biased, racist, discriminatory data, you're going to get the same inputs even though in theory systems are neutral and not prejudicial.

So addressing these biases as I said, the systems are neutral in theory, but in practice they can perpetuate the same problems, and one of the ways in which we can is maintain principles of fairness is to ensure large and diverse sets of data. How do we get that? Well, I think it's good to be a measure of impacting and changing industry practices, but I think some laws and regulatory changes where we legally mandate certain design standards, but we also mandate certain standards when it comes to data. What is being put into these systems to ensure fairness, equality, and less biased outcomes in these kinds of systems. And finally, the third area that I think that I would claim that in the long run data may prove to be more important than any assistance themselves, and that is the challenge of accountability and transparency in AI systems.

So here, some very basics about when we talk about transparency and accountability with AI systems, what are we talking about? So three typical concepts that are raised in this context. So transparency, understanding the AI models decision making, explainability, understanding how the system came to a particular decision, or the reasoning behind each decision, and thirdly provability what is the mathematical certainty behind the decision making? And this is such an [inaudible] All these terms are really essential for both the public and private sector as AI systems become implemented in government and people's rights and interests dealing with government and dealing with businesses as consumers are affected by these systems. And these concepts are not necessarily a guaranteed... They're not necessarily guaranteed because of what's often referred to as the black box problem. And again, this is something that will be talked about today, lack of transparency and therefore accountability in these AI systems.

In my view in the long run, often the answer that you hear today is, well we got to open up the black box, we need to mandate opening AI systems to render them more transparent and more accountable. But the claim that I want to make is that inevitably, while that's true, I think that's part of the solution in the long run, there's always going to be to some extent a black box problem, a transparency problem with AI systems. And that is a product of legal systems and legal rules, so trade secrets and other confidentiality, there's always going to be companies that are going to want to protect the secrecy behind their algorithms think of Google and the proprietary Google search algorithm. We've had Google for a long long time and we still don't entirely know how the Google search algorithm works both on a technical level as well.

And this really works and applies to deep learning examples of AI where the more accurate the algorithm, the harder it is to interpret especially with deep learning. And one example of this is Deep Patient, which was a system and developed by Mount Sinai Hospital, which was an unsupervised representation to predict the future of patients from electronic health records which had pretty good predictive outcomes, but they're the researchers and the scientists accredited really don't understand how it was making the predictions. So it's both a technical challenge in the long run and a legal one. And I think data is going to be important because outside of the black box, all we have is the data that we're inputting and the outcomes and the data that comes out, and I think when we think about ways of bringing accountability to those instances where full transparency and full accountability won't be possible we've got to focus on data and think about how we can use that to bring greater accountability.

And one way, let me give you just one quick example and I'll finish, is thinking about counterfactuals. So, there was a paper published last year by some legal scholars and computer scientists at the Oxford Internet Institute where they cite using counterfactuals to explain AI decisions without opening up the black box fully. The idea here is you can maybe explain the decision by providing minimal conditions that lead to an alternative decision. So if you're dealing with a person, you provide information as to what would have to change. So if the decision for example, is whether you get a bank loan or not, the explainability here or the transparency here would be what would have to change, what would the customer have to change so that next time they actually get a bank loan? And that's done not by entirely opening up the AI but instead thinking about counterfactuals and inputs of data and outputs of data in providing explainability and transparency that way.

So those are just three areas that I think did will be in the long run important and I'm happy to take your questions.

Pina D’Agostino: Thank you. All right, Carole.

Carole Piovesan: I think John did such a good job explaining some of the major issues of why data is important for AI that I'm actually starting to think of whether I should just riff for a minute and go off script or stick to some of what I had planned because I think there was a lot that was very good there. Let me start with a quick question to the audience, how many people have a general good sense of what AI is, capabilities, what it does? Okay. So we have a relatively... we have a good sophisticated audience that I think will be... I want to start with some of that, to talk about some of the practical use cases of how AI is deployed or operationalized in society. And the reason for that is to start small and then pull back and look at the broader policy context about why we care about this.

So I get this question all the time, which is why do you care about artificial intelligence? Why do you care about the data? And my starting point is not because I think there are... Frankly it's because there are huge conveniences that can come as a result and I'm excited to see them, but actually it's because there are very deep policy and legal issues that we need to deal with as we move forward in a data rich society.

So when we look at sort of what can artificial intelligence do, it can do all sorts of different capabilities that have to do with predictive trend spotting classification. So it's something that we as humans could do, we just can't analyze that degree of data and then ultimately create a prediction that is necessarily as fast and as accurate as an AI system. And Ai is a very broad term, I don't want to get stuck in the nitty gritty of what does it mean per se, it is a broad umbrella term, so it goes everything from in my view your predictive analytic capabilities to your autonomous agents so to speak, your self driving car. It's a very large spectrum right now, but what we can all agree on if not the definition per se, is the fact that in order for AI to be truly operationalized and useful and valuable and reliable, it needs to analyze a huge amount of data. So therein to me lies a very interesting issue when it comes to both law and policy.

So in general, I think John did this so I'm going to run, go through this very, very quickly, but you're looking at a process of gathering all this data, storing the data, which raises some very interesting cybersecurity issues and we see this emerging in law around the world. You've got data processing and then you've got that tangible insight, that predictive value, that is fundamentally why companies, governments, industry is looking at AI with such interest. It's because of that predictive value to help us better understand something that's really fundamentally each other. So it is trying to understand what is the human, how do we as humans act, because we are generally quite predictable, so how do we act and then how can we with some degree of accuracy, determine what we're going to do next?

Well, let's stop for a minute and actually think about and look at the extent to which data is being generated today. This is an infographic, there are some that go further back for 2015-2016 and you look at trends at data accumulation. And this is what is sort of a term to the Internet minutes, which I assume is 60 seconds because of what's in the middle, so it's just a minute. But you can look at the amount of information that is changing, that is being accumulated per minute over time in just a pinwheel of sources.

So the amount of data that is being gathered today, and this is a surprise to absolutely nobody in this audience, the amount of data that's being gathered is tremendous. And what makes companies valuable today, unlike what made companies super valuable 20 years ago is today what you sell are pieces of you and me. They are things that I give up under the auspices of understanding how you're going to use it. What you used to sell 20 years ago was a book or a widget that made you lots of money or a service, and that is no longer what is super valuable today. So with all of this incredible amount of data, we now have the ability to have more accurate and reliable predictions, and this is where we are in terms of the framework. So let's pull back what does this mean? There is a growing debate about why we... So number one, there's growing debate about the data exchange. If I give up my data, do I get something in return?

What is that thing? How valuable is it and how do I get it? So you've got number one, this issue of the data exchange, and this is being debated around the world. Are we entitled to 2 cents per name that I give up every time I give out my email address? Am I entitled to certain amount? Am I entitled to the mere convenience of the service I've just signed up for? What is a fair exchange in this data marketplace? So we have this notion of the data exchange, we have the notion of privacy that is being debated. To what extent do people care about their privacy and what does caring about your privacy mean? So there are great statistics out there, the Canadian Marketing Association came out, with this a really interesting piece that said 77% of Canadians are concerned about how their data is being used, then it said but only 45% of millennials.

All right, so right there I'm not sure who we're defining as Canadians, but we're saying 45% of millennials care about how their data's being used, 77% of Canadians I guess the older generation or maybe the infant's care about how their data's being used. And this puts government in a really tricky spot because where we are today is trying to understand what we do with all this data that's being generated and that's no longer necessarily in the hands of governments because governments have lots of checks and balances on what they can do with your data and how they can access it, which is not necessarily the same in industry. So we're entering a world where we are having these robust debates about what does it mean to have privacy, what does your data... what is the value of your data?

And the trade off is very much this as I understand it, so on the one hand when you are talking about privacy, you are talking about a quasi constitutional value, you are talking about an individual's right to have freedom to be independent and creative, freedom for descent, freedom for thought, freedom to do things that do not conform, and this is why it is such a protected and valued aspect of Canadian democracy and democracies around the world. On the other hand, you're talking about the innovation trade off. So if we don't have access to this data or under very restricted rules of having access to this data, what is the public good we're missing out on and I'm thrilled that today we're going to be talking a lot about the healthcare sector because the healthcare sector to me is the most strident example of where there's a public trade off where data is not accessible, where there are certain treatments, there's more personalized medicine that could be accessible to each Canadian in a universal health system which may create greater and arguably does create greater efficiencies, but if we don't have access to the data, we can't get to that public good.

Now the response of course is, well that's only one example, right? There are lots of examples of where data is being used to market to me for things that are completely innocuous or things I don't care about, absolutely, that's part of the debate, but we have to look at the full spectrum of where we are in terms of the value of data in our society today. The question of is it negligent not to use the data for certain purposes is as important as the question of how do we protect that data. So where we find ourselves is in a shifting policy context, we are in a point where we now have the GDPR in Europe. It is new legislation though modeled off preexisting directive that has good case law, it is setting a new standard for privacy protection around the world with some very robust fines associated with it.

In Canada, we had the national digital and data transformation consultations that were under that underwent and concluded in the fall, and that was our government's attempt to reach out to industry, civil society, academia, to better understand how does this government position Canada in the data play. In the US interestingly, you've got at least two states that have adopted legislation that is on par with GDPR standards and all US states that have adopted some form of legislation specific to cybersecurity and notification laws. And the whole premise there is, yes of course we need industry and free market to go off and innovate, but at the same time individuals are entitled to control what aspects of information they give up and do they understand the basis upon which they're giving it up. So do they understand that it could be used for third or secondary purposes that they did not necessarily consent to at first instance.

And so again, here lies a very robust debate. Around the world you have other countries that are starting to adopt the GDPR as the de facto standard, so the question becomes does that ultimately become what we are? Do we all move towards the GDPR standard and how do we as Canadians in particular feel about that? So I would say going forward because I really want to encourage a discussion, I think this is a really interesting debate, I would say going forward we can expect to see more shifts and dialogue with the government in particular and very much the people in this room and it is very much our responsibility to get in that game and have that discussion because it's not going to happen outside of us, we have to be involved in the actual shaping of what that policy looks like. We can see companies investing more and more in privacy, data governance I would say it's going to be a really strong area for investment and cybersecurity, and then finally I think what we'll see is a greater investment in digital literacy for individuals to better understand how they're using their data, and how they're giving it up and for what purpose. So with that, that's sort of my projection based on the value of data as I see it today.

Pina D'Agostino: Thank you, so I think that was a fantastic start to ground us for the day. I'm conscious of timing, maybe we'll have time just for one question is there anyone that has a comment, a question?

Okay, Yes, yes. Sorry a mic is coming so that... because it's getting recorded so that we all have the benefit of it.

Audience member: Thank you, Can you hear me? Natasha [Tusikov] York University. My question is about data acquisition. So we've seen a lot of controversy about acquiring data, IBM's acquisition of the flicker photos in the wild. So could you comment about practices for acquiring some of this data that might be against people's consent or against people's knowledge? People don't have a good understanding of how their personal information may be used and maybe acquired to train these systems.

Carole Piovesan: So I want to start by commenting, I think that's a great example and the Toronto Star recently published article about Iquvia and the recent IPCs investigation into the monetization of health data outside of Canadian borders and I think it's a really interesting investigation, I'm personally following it quite closely to the extent that I can. Data is a very broadly defined term, and when we're talking about personal information your question is specific to that. So if I as an individual give up my information for a particular purpose and then you ultimately down the road, use it for some other purpose, what are the implications?

So I have two responses to that, the first is that where data is anonymized and de identified its composition at law shifts and the jurisdiction of a privacy commissioner becomes more tenuous to the data and in fact it often falls outside of privacy legislation entirely. So the digital literacy piece becomes very important to understand that at a certain point you might give up your data for purpose A it could be aggregated, anonymized and sold for purpose B and you have to decide as a consumer if that's something you want to do, and that then leads me to my second point, which is transparency. If companies, to the extent companies want to be taking data sets and the risk of selling that data set off to a third party who could be linking that data set with other datasets, thereby increasing the risk of re identification that the importance of transparency in advising and putting the public on notice that you have this intention to me is paramount.

Johnathan P.: So to just add on, I think Carole really hit on some of the key points here, I'll just add few ideas. I think it's a really important question and I'm reminded of some of the important work of Helen Nissenbaum's notion of contextual integrity, so this idea that when data is collected, when we can send to have our data collected, or used, it's usually within a given context with given expectations. And then when that data is used in a different context against our reasonable expectations, then you get into questions of rights and interests in how do we enforce them and how do we reinforce though the contextual integrity and the expectations that we have about our data and its collection. And I think one of the challenges with existing data protection laws in Canada and still aspects of the GDPR is that we still focus on consent, which of course is important, it's an important part of the puzzle and the answer here. But I think where the GDPR, it's a flawed regulation, I've criticisms for it, but one of the things that it does get right I think is that it provides more mandated standards in certain contexts where it provides specific rules for certain classes of data.

More sensitive data is sort of hived off from certain kinds of collection or other kinds of data processing practices. And I think that's ultimately a direction that we need to move towards where it's not just consent because the experience under the directive before the GDPR was that very few people enforce their rights, which was based primarily on consent. There are very few cases where people went and sought to enforce their consent interest in light of what data processors and collectors were doing. So I think the prior experience led to some of these changes in the GDPR and that's ultimately where we have to go in Canada as well.

Carole Piovesan: I just want to add one thing to that, maybe being a bit of devil's advocate, we don't know... So going back to my point of the trade off, privacy commissioners often focus on rightly the use of data based on the disclosure. But where you have an appropriate mechanism that can protect the data but allow for random exploration, we have to be conscious that we don't know what we can create unless we're given the tools to create. So with your question of... The individual will as an individual, as a lawyer, I only read so much. The two things I never read are privacy policies and terms and conditions and I started reading them more, but I rarely read them because they're long and there's not much I can do about it, and frankly, I want the service, so I'm happy to, to engage in that trade off thoughtfully recognizing I don't know where it's gonna go, but there's only so much you can ask of the individual, it becomes a burden on the individual at a certain point.

And our privacy commissioner at the federal level has done a lot of good work to try to create guidelines for obtaining meaningful consent and I think that's important, but at the end of the day, we also have to think of that trade off. You have the individual, and then if you can have a safe, responsible, controlled mechanism in which you can explore with data in ways that were not necessarily anticipated, that opens up opportunities we don't know about today.

Pina D’Agostino: All right, well that's a wrap with a big thanks.