Okay. I think I must prepare this I'll ever be. So let's get started. Hi. Good morning, everyone. Let me go and close the door for any late comers. Is this your first lecture doing? Oh, gosh, what an honour. Okay, Aidan, welcome to the. University of Edinburgh Business School. I guess I'll be kind of your welcome committee, but you did have a welcome week last week. I guess. So. You did kind of get to know the school a little bit and how we do things here and maybe a couple of people. But you probably don't know each other and you don't know me and you don't know what is going to happen. So this is all very excited for all of us, including myself. And so welcome to this course in predictive analytics and modelling of data. Today will be kind of a relaxed lecture. So I think it's important to settle our expectations, what we want to learn, get out of this course, maybe get to know each other, talk a little bit about what we'll be covering, doing this course and do a bit of a kind of introduction to what is predictive analytics. What will we learn during this course? And yeah, all of that fun stuff. Let's start with a bit of kind of housekeeping. So this is the official course overview which you've been given in the course introduction, and it's a kind of slightly boring blurb, but it helps you. This course aims at training students in the field of predictive analytics to respond to the job market using a variety of methodologies. So the student's journey shall be a quest to distinguish the true signal from a universe of noise through the lens of predictive analytics. And to be more specific, this course covers the typical methodological steps of a prediction exercise statistical modelling, artificial intelligence methodologies, and it also covers practical issues in predictive analytics and how to address them. So this kind of balance between methodology in theory and the practical issues will be extremely important. And that's also one of the reasons why we combine lectures and computer labs the way we do. So what does that mean in practice? You'll have 20 lecture hours, so that's ten lectures, two hours each and ten tutorial hours over 11 weeks. Now there's one reading week in there. I think that's week six. That means there's no lectures and no computer labs in that week. And it's a week for you to catch up on any reading that you've been missing out. We'll also have a guest lecture in week five so you get a little bit of a break from me just standing here and talking to you all day. And we'll have an alumnus from the university talking a little bit about his experience actually applying some of those techniques that you'll learn about in this course. And I think that will give you a bit of an maybe good outlook for the future, what you might be doing with what you've learned in this course and also gives you a bit of hope for the future. Maybe so theoretical concepts will be introduced in the lecture and then we'll implement them during the computer labs. And this is a kind of combined approach, so we'll balance theory and practical implications. You'll also notice that the lectures are extremely applied, so we will cover the mathematics behind a lot of the machine learning methodologies that we'll discuss, but we'll do so from a kind of practical point of view. So not only talk about how does a method work, but what does that mean for the application of the method? What does that mean for you? How to pick a specific method for a specific data set, for example. So this is not the school of Informatics. You're not doing a degree in computer science. You're doing a very applied degree in business analytics. So the specific field of applying analytical techniques for business problems, we have two recommended readings here. One I actually brought with me because I have a paper copy, so this is kind of the core text ish for most of the lectures. It's pretty it's a pretty good book. One of the disadvantages of it, it uses all for its implementation examples. So each chapter kind of has a theoretical overview with plenty of case studies and implementation examples, but it also uses art to then show how you would actually implement those methods. Practically. Now in this course we'll actually be using Python. So that kind of is a little bit of a disadvantage of the book. That being said, I'm of a firm belief that if you know one programming language, you probably can read others as well. So you'll be still able to read the whole syntax even if you've never seen our because it is quite similar to Python. Now the other book that we'll be using is an actual book using Python as its application language. I also got a question about this book because in your library resources it still lists the version you. Losing are the reason for that is last year this course actually did use our assets code language. So that was a bit of a mistake probably on my side. When we copied over some of the information from last year, we did not change the book. That being said, the book exists in our version and in our Python version. It's the same book, The only difference the syntax and its implementation. So do try to get the Python version, but if you have the R version, it doesn't really matter that much, at least from a theoretical standpoint. You can still read the chapters and the content is the same. Talking about books, considering your new here, I will give you a brief introduction to maybe how to find books. I don't know whether you've actually covered that. Have you been given an introduction to the library system? Not really. Discovered this page a little bit. Okay, so we have two libraries here that you have access to. I mean, we have more than that, but two main libraries. We have the library here on campus, which is the main library, two buildings down. And then we also have our own business school library, the hub, which is exclusive to postgraduate students. And you can find it here on the lower ground floor. So we have our own librarians, our own library system, but we also have access to the university wide library system. This is their main website. So if we do, for example, want to check out this book. It will tell you hopefully where to find it. It does indeed. So you can see here that it says online access and full text available. That means there are probably paper copies available somewhere. They are probably more likely on Kings Buildings, which is our kind of Stem campus in the south of the city. But you have access to full online versions of these books and you have access to full online versions of both books. So there's no need to buy a paper copy if you don't want to. There's not even a need to borrow a paper copy from the library. The online versions will be completely fine. And if you look into the system with your student account, you'll be able to access those online versions. So no need to buy expensive textbooks. Yes, assessment is probably important for you as well. So assessment will be twofold. We'll do 60% of the grade will come from coursework, which will be done in groups. And then at the end of the year, sometime in December, you will have a written exam and that's an individual exam. Now, the group will consist of a report which will document the analysis of a provided data set using the techniques you learned in class. So we'll divide you up into random groups. No, you won't be able to choose them. We'll divide you up and then you will work together on a data set that I provide you with and you'll use whatever techniques you find most suitable for the problem you are given. So you can choose the techniques and then you'll write a report discussing and analysing your findings. Further details will be shared probably towards the end of this week, So we'll share the assessment brief and we'll also share your group numbers. And then you can come together in your groups and start working on this whenever you want. We'll give you a deadline, but you can start early or late. I don't really care as long as you submit by the deadline, please do that. The dates will be confirmed. Exams will happen in December. The exam timetable will hopefully be published in early December. Late November will see. It depends on how many exams they have to schedule. It's a huge operation to schedule all of the exams in a way that are not overlapping in location and time for students and gives us enough time to create whatever you've written for us. Yes, I've talked a lot without actually introducing myself, so I will do that now. So my name is Antonia Kitchen, which is probably a difficult name for you, so you may call me Antonia if you like, or you can call me Dr. Gibson, if you can pronounce it. If you can already tell from my accent, which is impossible to get rid of. And my name. I was born and raised in Germany. I've been in the UK for eight years, so I'm trying my best to emulate a BBC English, but it's not there yet. I have a PhD from this university and I wrote my dissertation on spatial temporal cluster analysis. So we'll have one lecture on cluster analysis and you'll see me very excited and happy about that topic in particular and ask a lot of questions about that topic please. So I then went to the States. I spent a year at Carnegie Mellon, did my postdoc there. I was working in the Pittsburgh Supercomputing Centre, which was really, really exciting. So the PSC is actually a joint computing centre between CMU and the University of Pittsburgh. And what we do there is provide both universities with computing resources for any really large scale computing projects that they have. So we worked a lot with kind of physics and maths and stats departments, but also medicine, psychology, Everyone who has a lot of data and wants that analysed comes to the PSC. Now, after a year in the States, I realised the states aren't for me. Even though my work environment was amazing. I did not really like the states themselves. So I decided, okay, let's come back. I really missed Edinburgh and I became a lecturer here in predictive analytics. You'll see a lot of meta semester actually, because I will be teaching both principles of data analytics and the predictive analytics module, and I'll also be doing your predictive analytics computer labs, but I will have two amazing Tas to take care of you in the principles of data analytics computer labs. I also realised those names are terribly long. Oh, okay. Let me drink something. It's really early. Good. I've already done quite a bit of talking and we've established that you don't know each other and you don't know what you're doing here. So we will first cover a bit more about myself, apparently, and then I'll hand it over to you. So it might interest you what my research areas are. My research interests are roughly in the area of computational computational social science. So I consider myself a computational social scientist. That means I'm interested in the application of various methodologies, machine learning and computational statistics in human behaviour very broadly. So a couple of areas have kind of emerged from that. My background is actually in quantitative marketing, so I still do a lot of research in kind of consumer behaviour, especially related to food consumption and also tourism and the movement of tourists throughout areas of a country. I'm interested in financial well-being and the connection of that to the mental and physical well-being of a population. So we were talking in the States actually about how. Your financial well-being is related to your mental health. So if you're not doing well financially, that will affect how well you're doing mentally, which affects how well you're doing on your job, which affects your financial well-being. So it's kind of this recursive system. But a lot of analysis in this area is actually looking at it quite separately and not really combining these different data sets in practice. So we're working on a way of combining different data sets to analyse this more holistically. I almost interested in local food systems. I'm interested in how people can access fresh produce and how affordable these produce is, especially in areas which are more underrepresented. So I'm working with colleagues in Canada who are looking at indigenous populations and how they can access fresh, fresh food in Canada. And I'm interested in rural areas of Scotland and interested in spatial inequality as a more broad concept and all of the above. So if any of that interests you, you do know you have to write a thesis or dissertation at the very end of your masters. If something of that kind of area is interesting to you, do let me know and we can talk about that. And maybe that's a nice dissertation for you in there. Now over to you. So as you don't know each other yet, you probably sit with those people that you know. I know that happens quite naturally, but you might not know everyone around you. So I would like you to turn around, maybe form group ish of maybe five people or so. I think that's a good natural group. Talk to each other, introduce yourselves if you don't know each other yet and talk about what you would like to take away from this course. Now, this is actually quite an important question for me as well, because I would like to make this course interesting and applicable for you. So after you've talked five ten minutes with the people around you and discussed what you would like to take away from this course, maybe talk a bit about your experience, why you've chosen this program, why you've chosen Edinburgh, any previous work experience or your undergrad experience? And then we'll go briefly across the room, pick a couple of people that make eye contact with me and then you can tell me what kind of topics emerged from that discussion. Don't have to be your own. That's why you talk in groups can be kind of topics emerging in your groups. And if there are some things that I pick up on that makes it possible for me to actually shape the course a little bit more towards what's interesting to you. For example, I can put my emphasis on application areas that are interesting and it kind of makes it more enjoyable for you. So do a good job discussing for ten minutes and then we'll share. Okay? Also let me start that recording again. I don't know whether you are aware you probably are that lectures are being recorded. So if there's anything that you missed during the lectures, they will be uploaded to learn afterwards and then you can actually rewatch a lecture and kind of catch up with anything that you've missed. Computer labs are not being recorded, so you should be present. And in some of the computer labs, attendance will also be taken. So we actually checking whether people are engaging and that's just engagement monitoring. So we know if something goes wrong and someone suddenly disappears off the face of the earth, we are aware of that. So don't feel like we're checking too much on you. We're just kind of trying to keep in touch and try to keep engaged. So maybe start the left so obvious. Let's start on the right. So let's start maybe in the back. What have you discussed? Have you had a chance to introduce yourselves? Anything interesting that came up? Yeah, we have discussed a lot of things. That's good. Yeah. So we share about our expectations about this course. And one of them is we we expect that we are able to, I mean, to use kind of like unsupervised or supervised learning methodologies and kind of like we discussed as well, this this course is similar with data science steps starting from domain knowledge that are the thing and the modelling and algorithm comparison. And that's what we have discussed in previously Perfect. Anything you're particularly looking forward to in that, in that list of things or just kind of the general process? I haven't decided yet. Oh gosh. Okay. Now that's good. That's good. We'll have time. We'll have 11 weeks to figure that out. So that's good. It's a long time. I know you won't think that it's long if as soon as you actually in December and the course is drawing to a close, maybe in the front. Yes. We discussed some models that we have learned during our undergraduate study and we learned some like bagging or decision tree, random forest, all exposed, for example, any so many models. But the question is that we don't know how to use these models to predict things in the real reality and we want to know how to get the parameters to predict the future, to predict the what we want to get in in the in the lab. Okay. So you kind of aware and you know the theory of these models, but you would like to know more about the practical implications and how to actually use them. That's excellent news because that's exactly what we'll be doing. And I think decision trees are in which week I should know that I decided on that week nine. So look forward towards week nine for you. Some where you part of the same group or where you're a separate group. Separate group. So another point, please. Yeah. Yeah. Any one of you. Thank you. Yeah, We have discussed some of the things and. Yeah, one of my friend here, uh, she was in a Start-Up company, and they they to. Feel it's something like sports and something. So by learning. This This is analytics and. What is the predictive model? So we can kind of like make the dashboard not just, uh, showing the current situation, but also we can like somewhat pretty or forecast things. And. Oh, so. And it's one thing and. In my case. I want also to explore that. We we we work. With a lot of data. I work in a telco company, so we have many. Uh, traffic and traffic data. So I want to know how we can use the data and get insights from it. That's what's interesting. So you have kind of live data streams coming in. I don't know. Let me write that down. That's a really interesting kind of type of data. So we will briefly talk about kind of time series data, which is which is closely related to that. But actually thinking about online learning in that context might be an interesting point as well. So let me take a note. Wonderful. Okay. My left. Aha. Surprise different. This time. Okay, so we talk a lot and we have some things in common. And I think for us, if we want to know how to predict things. And for me, I think, you know, nowadays we we are facing a lot of data, a lot of, you know, these statistics. So I want to know how to abstract like useful information from the data. Now the data is just data, and modelling is just like a kind of tool. So the most important thing is extract some useful information from. I want to know the steps. How do you do that? Yeah. Yeah, I completely agree. So it is relatively simple to actually collect a lot of data and a lot of companies are doing that. So they have just they're just collecting data because someone told them data is really important. Data is the future, but they don't think about what to actually do with that. So there's very little kind of targeted data collection. So knowing what kind of information you actually want to collect, to answer specific questions that you have, that can be extremely difficult because who knows what kind of questions you're asking yourself in five years. So to kind of think into the future and decide on what kind of questions you want to answer, maybe in the future is quite tricky. So we touch upon that a little bit, I think throughout the course. So we'll talk about what kind of data actually is suitable for different models and then also how to interpret the results that you get from the different models. And they will also be part of your coursework. So I look forward to that as well. Left back? Yes. Any one of you. I don't really. I don't know any of you. So any one of you can talk. What we were talking about was we were looking quite forward, looking forward to quite the application of it all because we're all quite new to like Python and. Um, coding part of it. So like it would be quite good for us learn how to apply and actually get conclusions from data where it's not actually quite efficient in that regard. You know what I mean? Yeah. Yeah. Perfect. So that's exactly what we're doing in the computer workshops. I think this will be very interesting for you where we go step by step through how to derive from data to actually implementing those models and then interpreting results where you pass the same group second. What do you want? What? That. That kind of group. Okay, so we have another group here in the back. I think a lot of us here share the same. How we all came from. A foreign background. Uh, probably computer science by default. Analytics in general would be probably tough for us, but we do look forward to. Uh, whether we use the data to create addictions or. And then in general, the applications in which appropriate methodology to use. And we also look forward to using real life examples and especially given that a lot of scenarios involve several outliers and how we're supposed to deal with that and the appropriate treatment. Yeah, yeah. Outliers will be, I think, coming up in week three as a kind of focus point. But we'll be talking about them throughout the whole lecture series. It's really interesting that you all mentioned different backgrounds, so let me check who has more of a computer science data science statistics, background. So let's call all Stem, maybe under the umbrella. Okay, Half ish. And who has a more kind of social science business? Similar background, the other half. So you can see what we're actually trying to do in this course is bring both sides together. So you will at times kind of think, Oh, this is more difficult for me, this is more easy to me and that's completely normal. And you will have colleagues for whom that's exactly the other way round. So do try to mix up between those groups, try to maybe talk to your colleagues, especially those with a different background. If there's a part of the lecture theorists that you find particularly challenging, someone else will find it easier to for than you and they will be able to to explain it to you. And you can learn together. And I think that's really the kind of the way to approach this lecture series is to think about it as a combination of those two areas. Okay, before I lose track, are you a separate group or do you belong to the group in the background group, the front. You belong to the group. So I covered everyone. Okay, so middle, middle, back place. Yeah. So actually 12 as well in one group. You know, the actually the problem with going last is I guess everyone have already told. Yeah. I mean. That's kind of good for me. Though. Actually, between. Us, all five of us are mostly from different backgrounds. I would say, as you said, diversity. Some of us have computational science background. The others is max, that and otherwise. And some of us, you know, also have a few of us are mostly familiar with the theoretical side of it and looking forward to the practical part. And the other is on the other way, as in some of us have started with the data data on its own and here to understand the theoretical aspect of the scene. The other thing. That came in between was, as for me, I come from a theoretical side, so implying I was always given a ready to go data set and I just had to analyse, predict whatever it is with simple programs. But the real life data is not apparently like that. Yeah. So I look forward to know how to clean data, how to make it ready to use. And finally, the other point was I said before we are we were discussing after this course, we could hopefully bring both business aspect of the thought because it was this, because that is how the collision always happens, how business people think and how an analytics person thinks. Then the knowledge doesn't match actually. But hopefully after this course we can come to that combination with both. People are happy. It's only knowledge, it's the language. So in a lot of cases you actually use completely different language for the same concepts and that's really, really tricky. So sometimes you just realise you're talking in parallel. So you're talking about the same concept using different words and you're getting angry at each other because it seems like you don't understand, but actually you're talking about the same thing. It's frustrating. So we'll try to kind of bridge that gap, maybe give you a couple of words and tools to to be able to talk to both sides and yeah. Back, Right. My right, you're left. Yeah. So, first of all, we discussed the powers. Hard for us to survive in this cold and wet weather because where are you from? I'm from Kurdistan. So that's the central region. And I guess people from my group are the same direction, from the same direction. So, I mean, like. Additionally, we've discussed comparison art programming and flight time. And we. On the language. Yes, we discussed that Python is much more simple in using, but a lot of great research papers have done. In our programming. So we just all knew in implementation of. To model that we've done. You know how to teach us how to implement. Well being freezing cold. Oh, my gosh. Yeah. Wonderful. So, yeah, this is actually I don't know what you've heard. We had a really hot summer this year. It was like 26 degrees, which is incredibly hot for Scotland. For Scotland is really, really hot. Also, you try living in Norway. That's fun. I was. I was having lectures at -27. That's cold. Okay, so you're not talking -20. We're not we're not talking about ten degrees now. Yes. I feel you. It's difficult to to adjust to a new climate. So keep your jackets on if you like. That's all right. And I have covered the back. I think. So. Were you kind of one group here in the front? Okay. Give you something. So majorly We talked about two aspects. One was, like most of us have implemented some of the other models, you know, during their undergraduate or during their work experience. But what we. Really want to learn is. Whether a particular. Model that is. To be implemented is the right. One or not. So back when I was working, I had a few folks who just. Left, right. Implemented models, but they. Did not really understand the math behind it. So that was. Kind of. Weird to me because later on when I started learning. So that is one thing that we want to learn. And the second is understanding how implementation of a particular model, uh, impacts the business. Because, you know, you just suggest a model to your kind and be like, you know, Hey, this is really fancy, this is all working. But the client has to understand the importance and whether it is relevant to their business strategy or not. So that becomes very important. So these are the two things that we are as a group looking forward to learn from this course. Excellent. Yeah, you make an excellent point here. I think like everyone always wants to learn neural networks and deep learning nowadays, so everyone always tells me, Oh, when are we covering deep learning? And then you have these companies with their their small data sets and they want to do deep learning for really simple relationships. And that makes absolutely no sense. So we'll actually cover when does it make sense to use what kind of model? And sometimes a simple model is really makes much more sense because the results, the accuracy is often better, it's more interpretable. It works on smaller data sets. So in many cases, simpler models make much more sense. There are circumstances where you need deep neural networks. If you are analysing large amounts of data with very complex non-linear relationships, they are amazing. But I feel like that's a personal pet peeve of mine. I feel like they're really being overused. So you will not hear me gushing about neural networks all throughout the lecture series. I hope that's not a disappointment to you. Okay, So great to hear from everyone. And I think it's also really valuable that you hear from your colleagues, that you're all kind of in a similar mindset and on a similar boat. So. Let's talk a little bit about what the rest of this lecture today will look like. We'll talk about what is predictive modelling and what will we cover in this course related to it. We'll have a brief look at different types of models and also the structure of the overall modelling process. And the recommended readings for this are the first couple. The first the first two chapters of both of your books, if you would like to have a look at them. So let's talk about predictive modelling now. And we've already talked about that predictive modelling really is about kind of decision making decision making in a very broad term, so that can relate to decision making within a company, but it also relates to decision making that you do in your everyday life. So every day you make some kind of decision here a couple of days, which way should I walk to go to, I don't know, campus fastest? Should I wear a jacket today or bring a raincoat or an umbrella in Scotland? The answer is always yes to that. But why is it always yes? Because based on history, based on the weather report, based on previous experience, and based on your personal feelings towards the temperature. Right now, all of those are contributing to how you make a decision in your everyday life. So how do we make decisions? We collect information and that information can come from various sources. So, for example, if we decide how do we go from A to B, we might ask Google Maps and Google Maps might tell us, Hey, this road is currently closed, so take a different route today. So based on information that Google Maps gives us, we change our decision. Similarly, if we want to buy the freshest bread at a local bakery, we might check Facebook reviews which tell us, Hey, they always make bread at 11 a.m. So it would be kind of logical to go there at shortly after 11 for you to get the freshest bread possible. Or you might think, what kind of car should I buy next? And then you might ask a parent, you might ask your mother, because they recommend a specific car point to you. And then based on her experience and you might check a couple of technical specifications that you know you want in your new car. And then based on that, you make a decision to buy a specific car model. So. All these kind of situations use information, they use data, but they all come from very different sources. Some of the data is objective. Some of that data is subjective and some of the data is kind of a mix between the two. So, for example, asking for a recommendation gives you subjective data. That's just one opinion. It might be a very important opinion for you and you might trust that kind of data and that opinion very much for that decision. But it's still just one opinion versus the technical specifications. They are objective data, so they are objectively true about that car. Now how you actually interpret that objective information is, again, up to you. In this case, you might check whether it suits your personal preferences. For example, what does that require? You have to have some kind of knowledge about those specifications. So in order to actually interpret the data, you need to know the source of the data. You need to know how reliable the data is, and you need to know how relevant it is. And you also need to know how to interpret it. So it's actually quite a difficult decision making process and these are really kind of everyday decisions that you make every day, and these are kind of simple decisions. So if we think about decisions which are even harder to make than that, at some point you reach a point where you as a human cannot answer that on your own. So here are a couple of questions. For example, what kind of ads should we run for this new product we are launching, or should we invest in this specific stock, or what will the housing prices look like in five years? And I think we all hope that there will be a little lower than at the moment because I can't afford a mortgage right now. So all of these questions are really difficult. And I, as a sole human can maybe make an educated guess ish, but I can't really reliably tell you what the stock of a specific company will look like in a few months. What I would do for that is I would collect as much data as I can for that specific question, and then I would use a predictive model to actually try and predict the future based on that amount of data, which I can't analyse myself. So that is the whole idea behind predictive modelling. We're kind of improving the decision making process that you already have in your brain and we're making it suitable for a wider and more complex array of problems. The questions I asked earlier are still the same though. Where does the information come from? How reliable is it? How relevant? How do you interpret the results? What kind of knowledge do you need to be able to interpret your results? So that is all still the same and that's all still true. So how my book, my trusty book here puts it. Predictive modelling is the process of developing a mathematical tool or model to generate an accurate prediction. Accuracy is one of these key words that you'll hear me talk about a lot, because it actually how we compare most models in terms of their performance. So we look at how close to reality models are able to predict something. That being said, sometimes higher accuracy is not the best thing that you should always strive for. In machine learning, you will often see kind of an accuracy hunting behaviour, so you'll see maybe a couple of papers and they report accuracy of 78.3% and then 78.6%. So clearly the second model is better and we should always choose the second model. What people don't really think about in that context is how expensive was that one, that model? What kind of data was necessary for that model to perform? That is the black box model. And the regulator doesn't like us to use black box models in finance. So all of these questions are also very important. It's not all about accuracy. It's a combination of factors in evaluating models. This is the idealised predictive modelling process that you'll see roughly in some form or another. In a lot of books. We start with some kind of question or problem. We collect data relevant for the question. We pre-process data, we look for outliers, we normalise, etcetera, etcetera. Select the model, train that model, evaluate the model and report the results. Now, what it really looks like is more like that, more like circular, because really when you actually start reporting your results, more questions arise. It's very rare that you actually manage to finally find the one answer that you've been looking for and you solve all of your problems. New problems will come up and the whole process starts again and it might even start again much earlier than that when you realise there is no data available to actually answer the question or the data has bad quality. So you have to go back to the data collection and kind of this whole circular process and there's a lot of questions there will ask yourself doing that time. So how do I translate a question into a problem that comes back to this language issue that we were talking about earlier? Business problems can be quite vague in their terms and trying to formulate that in a way that is a testable question is very challenging. Data collection What data is relevant? How do I find that data? Can I just use that data? What are ethical considerations and questions that have to be asked about that data? Then at some point you will ask yourself, Do I need more data? Then at some point you'll ask yourself, Do I have too much data? Because laptops are constantly crashing? So at some point you forget about the question you were actually asking yourself and think about why am I even doing this? Am I still pursuing the right path? Or what even is happening? So you have to continue on because that's what you do. You think about should I have picked a completely different model? Is that model even good for this question? And then in the end, you ask yourself, what does this all mean? How do I package that nicely? Why am I doing this? Why did I come to grad school? Something along those lines. And then you start all over again. And this is what I've been doing for the past six years. Okay, so I've talked a bit about each of these points already, starting with a question of business problem, deciding on the data that you need to do that, collecting those data, pre-processing it, selecting a model. This is actually quite tricky as well because model selection depends on a lot of different factors and in many cases those are subjective and irrational. Some people prefer some type of model over others. I mentioned earlier, I really like clustering and unsupervised learning. There is no rational explanation for that except that I spent five years of my life doing it. So I'm good at it, so I prefer to use it where possible. It's irrational. Training a model is also quite tricky, and that's actually the point a lot of people focus very much on when they think about kind of data science and the analytic process. They all think about this training step or maybe the model building step. They don't really think about the steps that come before and after evaluating results. I was already talking about the accuracy problem. Reporting results is interesting as well because you translate a question from business English into kind of data science, English, and then results have to be translated back into business English. So that's actually very interesting to think about and it will be part of your report writing in your group project. So still think about that and how to actually translate findings back into something that can be maybe visualised or communicated well to people. It's so loud outside. I thought I had to stop. But I still can keep going. Okay. Yes, that's the problem. So let's talk a bit about components of predictive modelling in kind of very formal terms. A predictive model is really trying to predict a relationship between variables. That's all we're trying to do. So we have some kind of features which we represent with X, and they can be called features, explanatory variables, independent variables. That's your data, That's what you've been collecting to solve your question and to answer your problem. And then you have some kind of target. You can call it a label, a response, variable output outcome, whatever you want to do it. That's your dependent variable because it depends on your independent variables. And then you have some kind of error term, obviously. So yes, notation. We write vectors and matrices as bold letters, matrices as capital, bold vectors as lowercase, bold letters. That's just here for kind of putting it on the wall. So you have that in your slides. Interesting is the function f that describes the relationship between your dependent variable and your independent variable. And that's what we are building and that's what we're actually trying to model and create. So the relationship between your data and the question or the answer to the question, that's what we're trying to model. And the way we usually do that is we'll do. A kind of building process with our data, with our independent variables, and then we're testing that with our with a sample of actually existing already existing outcomes. So we have a sample of YS which are given and we have a sample of X which are given, we're building a model and then we are testing how well that model performs. If we give it the same new data from the same source and no outcomes, and then we can compare the outcomes with the new data, with the outcomes from the old data. That's the very simple basic process of predictive modelling. We also talked about that terminology can differ between domains, not just between business, social science domains where maybe, for example, business professionals come from, but also within different areas. So for example, we are talking about the difference between R and Python. Earlier. A lot of that comes down to domain where people come from and what kind of area they are actually from. So R is often used by statisticians. For example, mathematicians often use Matlab and most computer scientists do use Python. So if you try to put a statistician, a machine learner and mathematician into one room and tell them to create something, that will take a while because they will be arguing non stop about the use of language and then they will be talking about the terminology. So that's actually quite tricky. During my PhD I was supervised by a statistician and a mathematician and they thankfully went along splendidly, which was good for me as a student. But we still sometimes had situations where they had to kind of talk to each other and tell tell each other. Okay, there's a different notation for matrices than I'm used to. Should be followed. This notation or this notation, What should we tell the student? What kind of notation should she be using? So there were actually a lot of discussions about that. A couple of examples are here data points. Some people call them a sample. So a singular data point is actually called a sample, but multiple data points that you have sampled is also called a sample. So I don't particularly like that word. You can also call them observations, instances or measurements. Features are called predictors, independent variables, input attributes, descriptors. Dependent variables can be called target class outcome or response variables. So all of that depends on what kind of book you're reading. What kind of paper are you reading, what your background is. I don't particularly care which kind of language you want to use, so feel free to to pick the one that is maybe the one you like best or the one that you prefer. I will most likely use machine learning ish terminology, but I might sneak in a bit of statistical statistics in there as well because that's my background. So it depends on whether I'm thinking in machine learning terms of statistics terms. But yeah, if you're confused, please do ask me and I can try to translate it into your language. The domain language that is, I only speak two and a half languages. Really? Okay. Yeah. Similarly, data types referred to differently. Categorical data can be known as nominal data, attribute data, discrete data, qualitative data, etcetera, etcetera. And sometimes you also have special data types like Likert. Scales are used in the social science a lot. They are just a subtype of ordinal variables, but because they are used so much, you sometimes see the term Likert scale data to refer to ordinal data in general, even though it refers to a very specific type of scale of seven points. For example, model training, building parameter estimation, all of that refers to the same process roughly parameter estimation is kind of strictly speaking, a sub process, but people use it to refer to the general term. Okay, so that was a lot of terminology, but we'll continue on that. We'll talk about supervised unsupervised modelling and I think we've touched upon that a bit earlier already. So supervised and unsupervised modelling are kind of the two big categories of machine learning, strictly saving this semi-supervised learning as well, and a couple of other subfields. But let's focus on the two big ones because that's the ones we will actually be covering in this course as well in supervised modelling. The process I described earlier is actually the one happening. So we have a number of observations that we've collected. So a data set that we've collected and we label part of that data set according to the target variable. And then we are fitting a model that models that relationship, that known relationship between the known data and, you know, target outcomes for each of these variables. And then you use that trained model on new data and check whether it's still performing well and still kind of putting out the correct terms and the correct the correct outcomes for each of your new variables. The purpose of that is at some point you're getting new data that's unlabelled and you still want to be able to label that. So a simple example for that is in if you have a binary outcome, spam or non spam for emails coming in, you have a bunch of emails and you know whether an email is spam or not spam. So you throw that into a model and you train that model to be able to recognise spam emails based on your sample. But then you have new emails coming, coming in and you want to use the same model to predict whether that email is spam or not. Based on the sample of emails that you had. You will already see the problem with that. Depending on how good your sample of spam or non spam emails is, they might be able to better predict new emails coming in. And you might also see that in 5 or 10 years time, spam emails might look completely different. So updating your model with new data constantly is really important and can be quite expensive depending on how fast your data landscape is changing. Unsupervised modelling is kind of a different process, so we don't have a response variable. So we're not targeting or labelling data as spam. For example, we are more interested in exploratory analysis of patterns that exist in the data. So we are trying to find, for example, in clustering groups of data points which are similar to each other and very separate and dissimilar to other data points. A popular example for that is in customer segmentation. Basically, you look at customers and their demographic data, for example, and you throw that into an unsupervised model, for example, a clustering algorithm, and the algorithm will look for patterns in that data set and groups of people who are similar to each other and very different from other people. Now, there is no target that we're trying to predict, really. We're trying to explore and find groups and patterns within data without a second step. The advantage of that is obviously that we don't have to label any data, so you don't need to know whether something is spam or non spam to train your model because we're just exploring. The disadvantage obviously, is that you're not training a model to predict a specific outcome variable. It's more about understanding your data and structure it has. So in this course, we'll cover both supervised and unsupervised models. Supervised will cover regression K and NS decision trees, support vector machines and neural networks and unsupervised learning will cover cluster analysis at domain one. We also cover PCA. Whether it's unsupervised machine learning or not depends on your definition of machine learning because it's a dimensionality reduction technique. It's not technically a learning method, it's reducing dimension. But some people count it as unsupervised machine learning. So let's let's trust their definition. Okay. What time is it coming? Would you like a break for some water? How are we doing? Are we getting tired? Right? Yeah. Okay. Me too. Let's take a five minute break and then come back and look at a couple of examples. Thank you. So. Okay, let's get started again. Settle down. Settle down. Don't worry. You're almost done. It'll be fine. Okay. Ready? Ready? Okay. So we just briefly talked about supervised learning and unsupervised learning and the difference between the two. So let's have a look at a brief example for both so it becomes maybe a little bit clearer. This is an example for a supervised problem. So we are trying to predict the likelihood of a tourist visiting some kind of visitor attraction in Scotland. And we collect the following data, which is very small, but it exists. So we have a small data set here, five records. We have the age of the visitor and we have the nationality and this is our label. So we're actually labelling each of our records, whether they visited that attraction or they haven't. And now the model is trying to find the relationship between your X, which is your age and nationality and your Y, which is whether they visited that attraction or they haven't. So who would like to take a guess? What kind of rules might that model look for? So do you see any patterns maybe in behaviour between visitors and non visitors? And the age. The age. Yeah. Good. So what would you say about the H. H. It's like. It's like. Could you repeat that? Is less likely to visit the place. So all the people less likely to visit. Yeah. So we have a look at for example, here we have these two oldest and they're both not visiting and I guess, yeah. Age and nationality. English people are less likely to visit while Scottish are. Yeah. So we have two English here and they're both also not visiting and we have Welsh and Scottish and they are visiting. This is actually quite interesting because we can see that these are both pointing into the same direction. Now let's imagine we collect new data of someone. So we have two new people, someone who's 70 and from Scotland, someone who is 20 and from England. How do you think the model will predict for these two? Will they visit or won't they? Mm. Yeah. So you think the kind of the nationality one is stronger than the H one? Yeah. Reasonable guess. Any other guesses? It's really impossible to tell, isn't it? You don't really know whether what is driving the behaviour is whether someone you up, whether who visited. So which of these variables is actually responsible for driving their behaviour? So it's actually completely reasonable guess Ted It might be nationality, it might be age. We really can't tell from that data. So this is a good example of kind of data limitations limiting what kind of relationships your model would be actually able to, to learn from. In some cases, the model will then probably assign pretty random. So this might be an example of a model where you're not really reaching a high accuracy because it's kind of a bit of a toss up unless there is actually a stronger pattern that we might observe if we collect more data. So if we can then confirm whether one of those visited or hasn't visited, that could add to our model and improve our learning for the future. So let's have a look at an unsupervised model I mentioned earlier. It's actually used quite a bit for segmenting costumers. So this is a data set that bank collected on the use of their mobile banking app and the age of the customer. So if you have a look at that, what kind of patterns do you think an unsupervised machine learning model might find? Yeah, the person is the less likely to use the mobile app, you know? They're probably referring to this kind of roughly linear relationship. So what did I say earlier about unsupervised learning? What are we looking for? Patterns and groups. So this is actually a good example for when an unsupervised machine learning model might tell you something that you're not really that interested in. Because what I think what clustering, for example, would do with this data is something like this. It would find probably three groups these points, these points and these points, because they are the closest together. It won't actually answer that kind of question. We might ask about the relationship between the two variables. It would just find patterns in the groups of similarly behaving people. Now, what you could then do with those groups is you could have a look at the blue group and you realise they are younger and are more using the app and you have a look at the red group and realise they are older and they are not using that group as much. But that is kind of a second step in the interpretation of the results. So the actual kind of modelling outcome that you get is a pure segmentation of the data. It's not actually picking up on that linear relationship directly. If you use a supervised learning model, for example, it might be able to actually pick up on the relationship between the two data points, especially if you try to kind of put a linear regression line through that. Another thing I would like to you to put your attention to is these two up here, because these are outliers. They are really kind of odd people. That's just 1 or 2 that are not behaving like the rest of the group. They're not following the trend. So a question for the company now could be what are these people doing? Why are they different from from the others? Is there maybe a collection error or are they actually a valid extra group that we might be interested in in behaving in their behaviour from a bank perspective? For example, in many cases, banks try to target specific products and communications to subgroups in their kind of target audience or their groups of customers that they have. So they might create some kind of communication, for example, for the elderly. Hey, have you tried our mobile app? It's really easy to use and they might kind of target these younger people and they push, for example, new products through the mobile app because they know they will reach the right audience. But what do we do about these these here? That's that's actually one of a business question, isn't it? Do we want to target them? I mean, we might think it's it's valuable enough or we might think it's expensive to target them very specific or specific communication devices without really getting much of a return. So these kind of have to create two questions. Why are they here? Are they real or is this a collection error and what do we do about them? And that's not a question the model can answer. That's a question that you have to answer in combination with what you know about the business. Okay. Now let's talk a little bit about another way of differentiating between different types of models, specifically regression and classification. So these are both typically under the same umbrella of supervised modelling. So we'll talk more about supervised modelling than unsupervised modelling generally. They both share the same concept. So we want to make predictions based on some kind of known data set. That's what we've already been discussing today. The difference is that in classification we're trying to predict a class label, so the belonging of a data point to a specific class and a regression, we are trying to predict a continuous quantity. So basically a numeric value, an outcome. There is sometimes a bit of overlap and you'll see that's where the lecture series. So we'll talk about, for example, linear regression, which is a regression, a regression model, and then we'll talk about logistic regression, which is a regression model which kind of in the sense that gives you a class label. Then we talk about regression trees, decision trees, predicting a number. We talk about decision trees, predicting a class label. So we always talk about the same model in two different contexts and which one to choose completely depends on what you're trying to predict. So let's have a look at four examples here, and I ask you whether something is a regression problem or a classification problem. So the first one is what will we stock price for the specific company next month? Regression? Yes, that's a number we're trying to predict. Will the customer of my company churn, So will they leave me classification? Yes, that's a specific case of classification, which is binary classification. Either they will churn or they won't. That's the only two options. Which genre does this movie belong to? So classification and how many visitors will this museum attract? Yes, that was quite simple. But these are kind of the questions you will try to answer. So in the first step, you will be given a business problem. And the second step, you try to create a measurable question based on that business problem. And then you have to decide whether it's something that you solve with a classification approach or a regression approach or some kind of segmentation clustering, unsupervised approach. So these are kind of your three options that you will have to have a look at. Let's talk a bit about types of variables. So generally we just first distinguish between quantitative data and qualitative data. Now, depending on your background, you might be more familiar with one type or the other. Quantitative data is anything that is numeric and the numbers have a numeric meaning kind of. So we're counting something. We have a ratio of something. We have a real number, something that is recorded through numeric values. Qualitative data is non numeric, and in a qualitative analysis, we also treated as non numeric. That's the big difference. I actually that's that's something I added to this definition I think a couple of nights ago when I was thinking about that, because examples for qualitative data are texts, transcripts, images, recording sounds. ET cetera. But all of that type of data can be treated as quantitative data. So you might have read about image recognition, for example, or you might have read about text mining algorithms. So the difference between kind of quantitative approaches to qualitative data is how we treat qualitative data. In the social sciences, there are two general schools of thought. You can treat qualitative data as it is for qualitative research. And that means, for example, doing in-depth interviews with someone and then analysing the codes, the themes that are rising through those interviews. Or you can collect a large number of surveys from from a group of people and then analyse the results from those surveys with various quantitative methods. Both can be used for similar questions. So which ones you choose depends on what kind of answer you're looking for really. And both are also. That's really important, really valid, valid choices of research. That being said, this course will focus on quantitative research. So please don't try to convince me that you want to do interviews in your group research because that's not relevant for the applications in this course. But generally, qualitative research and quantitative research are equally valid in the social sciences. So that's what I said here. Much qualitative data nowadays can be and is being transformed to quantitative data in image recognition. That usually takes the form of a matrix where you have zero and ones where you indicate the pixels of the image. So that's a way of transforming an image into quantitative data. And then we can analyse that with the usual techniques. So you can use neural networks for recognition of handwriting, for example. And we also do that for categorical data. So if someone takes a box what their highest educational degree is, then we can interpret that as qualitative data. So as the words and the meanings behind them. Or we can use that as a label and then we can use quantitative techniques to analyse that label. Yeah. And the big difference is really whether we analyse smaller amounts of information more in depth or larger amounts of data to generalise from that. For example, if you're interested in your customers opinions about your company or specific product, you could do in-depth interviews with them, but you can only interview so many people. So usually you will maybe interview 1020 people, but really, really in depth. You'll talk an hour to them and find out what their subjective opinions and motivations and beliefs and how that is affecting their connection to your company. That's really valuable information that you won't be able to get through surveys. On the other hand, it's difficult to actually extrapolate from that to the general audience. So using, for example, text recognition to learn general themes from a large number of transcripts, instead of having to do manual coding and thinking about them gives you maybe more information, but it also is more surface level. So let's focus on quantitative data, because that's what actually will be dealing with throughout this lecture series. So we generally distinguish between discrete data and continuous data, and within those we then distinguish between binary. So zero one categorical, also known as nominal data, ordinal data, which gives you a ranking or for example, a liquid scale and then numeric integer account data. So these are all countable numbers. If you think back to our example earlier, visitors visiting my museum, that's countable. You can count how many people actually visit. Continuous data is either interval, so the zero has no true meaning or ratio examples for that for interval. Typical example is temperature and Celsius. There is a zero, but a zero doesn't really mean anything, it's just another number. So you can go below zero in temperature. An example for a ratio with a true zero would be weight. Zero of weight means there is no weight. So non existence of that. So that's the main difference between these two. And yes, I mentioned categorical variables are typically transformed in our type of analysis. So we're looking at one hot encoding. Usually in that way we can actually use categorical variables in our modelling. Okay. We're almost done. You're doing well. I know you're getting tired. So let's have a look at a couple of data types. I write most likely data type because strictly speaking, they could be recorded in different ways, but kind of go with the most logical explanation for each of these. So I already mentioned temperature measured in degree Celsius. What was it? Yes, continuous and more specifically interval. Yes. So a number of visitors to a theme park would be discrete. Yes. And within that account, data education level. For the No. Yes, That's actually an interesting one because you could argue, depending on what kind of education levels are recorded, there might be categories on the same level which are then not ordinal, which in that case it would be categorical data. But typically in census data it is recorded as ordinal variables. The favourite floater flavour of soda of a sample of restaurant visitors. Categorical. Yeah, most likely. So the way the question would probably be asked is which of these sodas do you prefer? And people take a box. So that would be categorical price for an item over time. Yeah, continuous variation probably, and also time series data. So that's kind of the specific data type where you're collecting a value over time for the same object and whether a patient has a disease or not. Binary Exactly. So another thing I wanted to briefly mention is that in the social science you will encounter in many, many cases in your data sets, mixed data types, and that can be really challenging. So you might have some kind of categorical response variable and then have a mixture of numeric and ordinal inputs that are trying to predict that categorical response, or you might have a regression problem or you have a continuous output, but you have numeric and binary inputs. So kind of trying to combine these different data types is a whole research area of its own, to be honest. And I've actually spent a considerable amount of my research time lately thinking about different data types and how to combine them into one data source. That becomes even more difficult if you're looking at different data sources. So in a recent research project I'm working on, we are looking at combining census data from different countries which are collected in different ways. So I mentioned the education example earlier. Different countries have different ways of categorising education levels, but if you then want to to combine and be able to compare populations from different countries, you have to find ways of linking between the two. And that is mostly done manually by poor people like me, who then have to go through hundreds of pages of census documentation trying to find out how these variables are recorded, coded, sampled, and then trying to understand educational systems in Canada versus France and trying to understand in which ways you could be able to compare them. So that is really, really tricky. Thankfully, you will most likely not be forced to do that unless you choose to do a dissertation. And then that's why you will have to only look at ways of treating variables of different types, doing pre-processing. For example, we were talking about how hot encoding and then in the interpretation phase it will be really interesting and important to think about information that you can gain from categorical and ordinal variables. So, for example, an ordinary variable, the numbers don't mean anything except the ordering. So for example, high school diploma is not too smaller than a master's degree, even though it might be coded that way. It's just less educated than a master's degree. But there is not really a sense of the distance between the two. So what do ordinal variables actually tell us and which models are suitable for mixed data types? And I think someone who was excited about decision trees. So decision trees are very suitable for mixed data types. Yeah, extremely important to always, always, always check your assumptions of your model. We were talking about that earlier. When you want to have your colleagues mentioned, people just use any model without thinking about whether it's actually applicable. And a lot of people actually do that. They think, Oh, neural networks are amazing and they just throw data at them without thinking about what are the implications. So people might, for example, have a couple of binary variables and then create more variables by combining these binary variables and loaded into a model. And you can't just do that because then you actually run into the problem with Multicollinearity where the variables are running Colonia and you can't treat them as separate because they're correlated with each other. So there's a lot of kind of problems coming up where people think about this is a model I want to use. This is the data. I just put the data into the model and I interpret the results and you might get really, really good results, which don't mean anything because you didn't check your assumptions. So please check for your assumptions even though it's kind of a headache. So you have to check for like, I don't know, independence of error terms for linear regressions, etcetera, etcetera. There's a list of assumptions. You have to go through them. You have to report whether you checked all of them, and then you have to after that, use the model and report the results, because otherwise people can't really. Judge. How well you did your analysis because you might have done complete. Yeah. Yes. So in. Mixed data types. Specifically, you will encounter problems in all steps preprocessing. We already mentioned that. Which variables should you actually choose for your model? Is there a way of choosing variables which work well for the model and give you a lot of information? You might need to think about things like expert knowledge? What do these variables actually mean? That can mean asking people about that, talking to the company, reading 200 pages of documentation, all of that model section. We talked about decision trees being better with categorical data, for example, and interpretation. How are we interpreting numbers if we know mixed data types were involved in the model building? Okay. Last point. You almost done challenges very briefly talking about challenges and predictive modelling. So I took these from the book inadequate data, pre-processing, inadequate model validation, unjustified extrapolation and overfitting a model to existing data. These are kind of the core four problems that the authors saw a lot with people in their predictive modelling. I've seen all of those as well, but I would also like to add the following. For my own experience. Too little or too much data, both don't work. There is a sweet spot somewhere. I think I actually created a picture of that once which kind of looked something like a probability of this happening and then it's too little data, the right amount of data and too much data. So you either have too much data or you little data. I've never had a sweet spot of exactly the right amount. It probably won't happen to you either, so that will be tricky. You have to choose the right model for the right amount of data. I mentioned. New networks survive on a lot of data. If you have little data looking at simpler models like regression analysis makes much more sense. P Value hacking pet peeve of mine and I will talk about that multiple times in the other course actually in your principles of data analytics, because that is something I see all the time, especially in dissertation phase and sadly enough in a lot of academic papers as well. P Value hacking means that people are looking for something that isn't really there. They want to find a relationship so desperately because they spend so much time with money on something and they test for it and then they report something as marginally significant or they report something like, well, it's not significant, but we still think it is. And it's just. No, that doesn't make sense. Significance is binary. There's either significance or there isn't. And that's it. It just report the numbers as they are and then you go with that. That's that's also related to the last point, overreliance on theory and general knowledge. You can see that, for example, in economics. Oh, I hope no, none of my colleagues watches this. No, no. They are amazing. I've seen that in a paper a couple of years ago where people think that because specific things in economic theory are thought to be true, even though data sets differently, it must be that data as wrong. Not a model. Obviously, that's not only the case in economics, that's just one of my areas, because I read this paper, which was fury eating, but it also happens in Stem areas, actually in physics, for example, where you collect data and you have these theoretical models about how the universe probably works, and then you collect data and the data tells you probably not, but they think this model looks beautiful and there is a Nobel Prize attached to it or something. I have no idea. So the the data must be wrong. There must be something wrong in the data collection process. And in many cases that is true. In many cases, data quality is affecting your results and is affecting what you can actually report. But in other cases it's probably empirical evidence pointing towards the direction. So let's trust empirical evidence over what we think and believe to be absolutely true. Okay. I realised I've talked a lot. So because this is your very first lecture, I do want to have you talk a little bit more to me just for a couple more minutes. Five, ten minutes and tell me what you think about what we've discussed so far. Does this kind of fit into your expectations? And also, are there any questions about anything that I've covered? So feel free to ask as many questions as you like. That's really why I'm here. So I'm trying to put material. You could just read this book. So my job is to take this knowledge, make it understandable, presentable, applicable to you, and present it to you in a way that works for you so we can learn from this together. I'm not inventing new knowledge here. I'm kind of packaging knowledge in a way that hopefully works. So that's why it's really important for me to get your feedback and get your questions. If you think it's a stupid question, it probably isn't. Your colleagues probably wondering the same thing, but if you're a bit too shy to ask a question in class three, free to email me or ask me in a break or after the lecture and I'm equally happy to talk one on one. Any immediate questions about the course structure? Anything I've told you? Anything unrelated? Yes. So I know we have an exam. Is it just the one exam in December or will we have also, like a midterm? There's only one exam. Yes. So the assessment is just the coursework, which will most likely be submitted, I think, in November. And then the exam beginning of December. Yeah. So one coursework, one exam, no midterms. Yeah. Would we be getting monthly segments of weekly? Would be cool about that. Yeah. So there is no continued assessment. So there's no grading continued ongoing during the semester. We do have the computer labs, which I think you will have more on tomorrow and we'll have kind of exercise sheets for you to work through and you will get solutions and feedback from me on that. But I'm not grading any any of your work so you can double check what the solution that I'm providing and the feedback that I'm giving during the class. What? What is the. Rite Aid early in December. I really want to join my Christmas. And you'll find out on Friday. Yes. Yes. Then I think at least the deadline for the coursework will be will be on this Friday as soon as it comes through. Moderation. The exam timetable, I can't tell you. So exam. I think we have like a two week, two week period. So I would recommend that you wait for with scheduling any of your holidays until you have your exam dates because they can go deep into Christmas. So I've seen exams on the 23rd of December. So if you think that you are saved because it's two days before Christmas, you're not safe. Yep. How will the course. Content of this course cooperate with the course? Mhm. Yeah, that's a good question. So I think we actually touched one of your colleagues touched upon that earlier when they were talking about describing and descriptive versus predictive. So the cause and principles of data analytics is a basic statistics course, so it will be more in the descriptive analytics direction. We will cover things like hypothesis testing, analysis of variance, Anova and kind of descriptive statistics versus discourse, which is much more machine learning heavy. So we're not doing statistics as traditional as in that course, but obviously they will run in parallel, so you'll see some overlap. That cause is also running with a second cohort. So you meet your colleagues from the fintech program. Don't ask me which one. I think I changed the name recently. So one of the fintech programs. So there's a very large cohort of you. I think we have like 115 or so. So big group of students, only five weeks of lectures. It's a crash course in statistics. Yeah. Anything else? Retired. It's your first day of lectures. I do hope you're not tired already because there's many more to go. Yeah. That's your favourite dataset you've worked on? My favourite dataset I've worked on. Oh, that's really a tricky question. So currently I'm one of my research interests is I mentioned earlier in financial well-being and kind of financial health. So I was working with a data set from the Financial Consumer Agency of Canada, the FCA, recently, and we were looking at the well-being of people regarding their finances during Covid 19, which was really interesting to look at because they are collecting this type of data over years. And then Covid happened, so they changed their survey up, but still some of the variables stayed the same. So we could actually have a look at the development of financial well-being of Canadians before the pandemic, then doing the pandemic and now in the recovery phase, they're back to kind of their regular updates. And I think this kind of temporal development is really interesting to me, especially because there's a spatial component as well. So I could have a look at the spatial inequality and how well different areas of Canada recovered quicker than others. And you can connect that, for example, to socio economic disadvantage groups which are struggling more recovering. So I think that's a really interesting dataset for me because it was so topical, so new and relatively clean, which is always nice to work with. And yes, you can ask me about my research as well. I'm happy to to answer that. No, you're done. I can. I can see that you're starting to wrap up. I see. You know. All good. So let's wrap a little up, wrap up a little early today because our break was a bit shorter. And I will see you tomorrow in your lecture as well as your computer lab. And I'll also see you next week. And if you have any more questions, email me and I can set up a meeting or answer by email. Cool. Thank you. Bye bye.