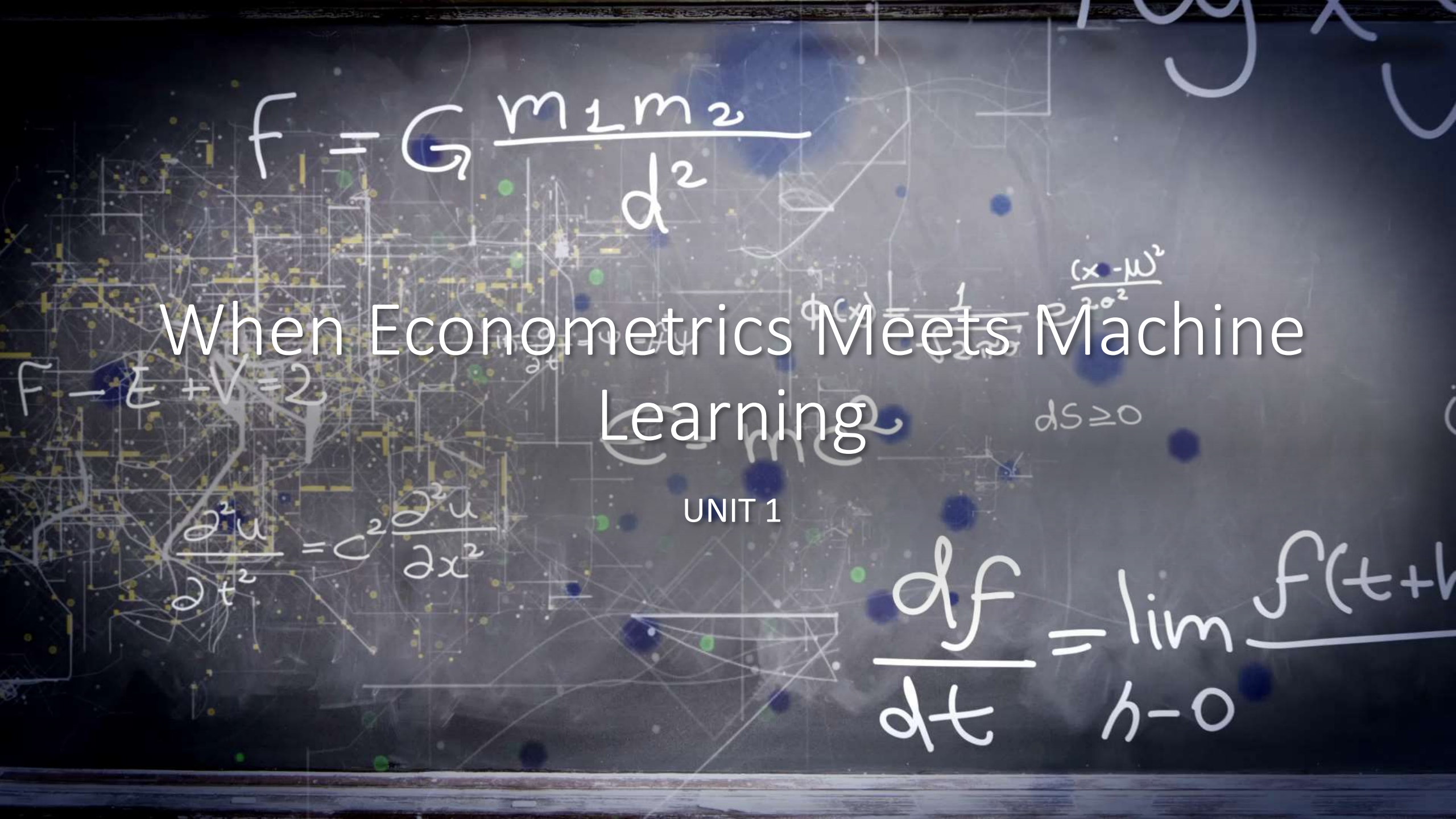

When Econometrics Meets Machine Learning

UNIT 1

INTEGRATION OF ML & ECONOMETRICS WILL BE OF GREAT VALUE

Remember that ***methodology*** itself is just a tool. We want to advance the tool in order to solve more interesting business problems.

Econometrics is used to find out the ***causal relationship*** of data

Whereas, from the ML's point of view, we want to achieve predictive accuracy.

Therefore, if we probe into the 2 approaches deeper, it can be realized that there are probably the ***same philosophical ideas*** that accurately describe the data generation process.

Obviously, once we get to know the ***way in which data*** are generated, we can get more ideas about the ***causal relationship*** in this process, and we can achieve higher ***predictive accuracy***

ML is Predictive, While Econometrics is Explanatory

ML is a way of problem-solving, which is designed in the paradigm to find a solution, a recommendation/computation system, or a classification.

It is usually in a ***predictive mode***, very computational, not necessarily in a statistical way.

In econometrics, by hypothesis testing & other traditional social science research methods, econometrics is trying to explain ***what is going on***.

ML usually goes along with ***design science*** paradigm, while econometrics always goes along with the ***economics*** of paradigm.

In ML, we use various ***computational techniques*** like classification trees, random forest, and deep learning models, and also logistic regression. ML is used as a pattern-searching tool.

Since all you want is achievement of ***great predictive*** power, there is ***rarely a linear*** relationship in ML.

However, econometrics basically explores ***linear relationships***. Sometimes, you can also find a way to incorporate nonlinear ones. It is primarily a take-away from **regression**, with different sophistication levels to pursue these linear relationships

Combination of ML & Econometrics is achievable and challenging

There seem to be 4 ways of combining ML with econometrics.

- In the 1st one we use a ***ML model*** to ***feed in*** the variables that are going to be used in an econometric model.
- The 2nd one is you have the ***ML algorithm***. Then, you can borrow some of the fundamental techniques from econometrics to enhance those of ML.
- Thirdly, you can do it the other way around, which means to ***borrow ML*** techniques to enhance econometrics models.
- The 4th one, which is also the most interesting one, is that you use both of them to explain the same phenomenon. You can take different approaches then. For example, by doing text mining & sentiment analysis, you can define different variables, and then they become part of your econometric models.

As far as I am concerned, I would like to have the approach of using ML first and then try to get the explanation of what is going on, as well as gain an understanding of the underlying correlation for the co-occurrence. Then, it is time to find their causal relationships.

ML & Econometrics are different in perception and understanding of problems

Econometricians are basically from one side, and statisticians & computer scientists are from another.

Since econometrics needs theories from economics, they develop theoretical models.

However, with the help of various techniques, statisticians/computer scientists use data to develop models.

A statistician uses data to solve problems without any analysis of structures or backgrounds of issues.

By using bootstraps & other techniques, he can develop statistical models and then find solutions for problems.

To be specific, statisticians/computer scientists do not make assumptions when they develop models. They will not suppose variables to be dependent or independent before it can be proved.

The features & relationships of data have to be figured out by computer techniques, called bootstrapping.

Thus, computer scientists and econometricians are totally different in their treatment of data.

With the help of computers, computer scientists can do data processing, such as storage, analysis, organization, and classification, which can finally lead to a model.

However, econometricians always start with an assumed structured-stochastic model, the unknown and unobservable parameters of which are estimated from a relevant sample by using observed data, and they can be used for inference and prediction.

Econometrics is theoretical foundation, and ML is technical assistance

In econometrics, estimating the ***AVERAGE*** response of the independent variable to the dependent variable is the main purpose of the prediction and regression process.

One goal of the model is to eliminate all the things that cannot really explain the dependent variables. For example, if you want to ***hide in the crowd***, you had better look like an average person.

The primary reason that all these methods have been used is that they can ***avoid high degree of complication*** in econometrics.

Nevertheless, they ignore and lose all the ***individual components*** in the data, which ML can exactly address

In my opinion, a successful combination of ML & econometrics is that you can use econometric models with the bootstrap of ML.

After all, ML cannot start from a vacuum, and it also needs a theoretical foundation.

Both econometrics and ML have advantages & disadvantages

ML is good at ***cross-validation***, which can be used for model evaluation. Cross-validation is used to assess the performance of models in practice, which is really hard for econometrics to accomplish.

Econometrics can be used to reveal the ***fundamental*** underlying process, which is helpful for grasping the economic structure.

However, ML **cannot** help contribute to the fundamentals, due to which you are not satisfied.

Statisticians like to resort to computers when dealing with changes & new problems.

It is necessary for theoretical researchers to know about the fundamental process & underlying relation, which are important for simulation, what-if analysis, and decision-making effect analysis.

Can we use both techniques in a new way and in a high level?

Econometrics & ML depend on each other

- Take the bank lending process as an ***example***: Assume there are 8 individuals with different features, such as height, income, & education level. One person is a tall guy, with high income & high school education. The 2nd one is a tall guy, with low income & undergraduate education. The 3rd one is a tall guy, with medium income & undergraduate education.
- Therefore, we just want to find out the ***logical relationship*** behind the data.
  - Is height relevant with income?
  - Is education relevant with income?
  - Is there any other connection among these features?
  - What are the decision causes in the 8 individuals?
  - What are the dependent variables?
  - And how can we get a decision tree with dependent variables?
- It can be inferred that ***income*** is the most relevant in the decision tree from the samples above.
- **So, what exactly is ML?**
  - It is a process that ***derives insights*** from a large set of data. We try to understand the distribution, which is a classification. By conducting a survey, we find a pattern, which is fully coincidental to different factors or variables.
  - Then we try to find out the ***causal effect*** of a decision. After that, we can evaluate the decision and its effects.
- **So, what essentially are we doing with ML & econometrics?**
  - In general, we tend to develop explanations inside knowledge through certain data. The example mentioned earlier are very simplistic, and you need to acquire more in-depth knowledge by yourself.

Econometrics & ML depend on each other

- Take the study with a commercial bank as an example: What we found from the decision tree generated to explain the commercial loaning process for medium-sized companies is that the loan is based on the ***financial attribute*** and the ***risk level*** of the applicant.
- The numbers 1 to 5 represent the risk level perceived for the small- and medium-sized businesses, with the score 1 meaning the most secure and 5 meaning the riskiest levels.
- In classification, we use 1 to 5 as the dependent variables, and the financial characters of the company applying for loan as the independent variables.
- Consequently, ML not only gives the causal relationship & explanatory power, but also depicts the structure, which is called ***tree*** structure.
- For example, the decision-making process of the most secure company is very short. We definitely have a very long decision-making and evaluation process for the riskiest company.

Econometrics & ML depend on each other

- As a result, this process is about generating not only knowledge that we can gain from ML but also the knowledge about knowledge, which is called ***meta-knowledge***.
- When you look at the tree, you can see characteristics about the decision tree. Therefore, you are able to re-evaluate the decision and the policy effects.
- In summary, that is why we need econometrics. We never know whether the tree generated or any other ML model generated is stable or reliable.
- Therefore, we have to do validation. Over the years, we have used many different techniques for validation. Cross-validation is also necessary. However, the reality is that the use of ML usually brings a massive result, which is a trigger for the training.
- The tree training can make the tree more understandable, which means you need some sort of techniques to provide a confidence level of the validity about the ML model generated. There are also some other useful techniques, such as statistics & econometrics.

When do Econometrics & ML need each other?

ML & Econometrics can be beneficial to each other

1. Use ML models to ***feed in*** the components of econometric models. The classification results of ML can be used as variables in econometric models, with which we can make explanation.

2. ***Enhance*** econometric models with ML techniques, or enhance ML models with econometric techniques. Another approach is to borrow techniques from one of the 2 areas to enhance those of the other one. Consequently, cross-validation, as a component of ML, is usually used when developing an econometric model. Additionally, it is possible to use econometric models when dealing with overfitting in ML.

3. Use both ML & Econometrics for comprehensive understanding of a phenomenon. Sometimes, there may be a ***tough problem***, which is difficult to begin. Then you could use ML for coherences, which emerge from exercises such as building a tree, or something similar to a decision rule. Then, you may have some ideas to determine the approach to be used for the ***causal*** influence or some kind of analysis.

Therefore, I think that after integration, they can enhance techniques mutually, and they also can be used in a parallel or hierarchical way to solve the problems.

DL is a good approach for revealing the correlation between endogenous variables & instrument variables

Discovery & explanation of ***endogeneity*** are crucial in research.

And it is also quite difficult to get the ***instrument variables***, especially when short of variables.

Moreover, instrument variables could be somewhat ***weak***, which means the correlation between the endogenous variables & instrument variables is not close.

In this situation, instrument variables are ***incapable*** of describing the variation in the endogenous variables.

Then, it is time to use Deep learning as a new method. The actual relationship between endogenous variables & instrument variables cannot be conjectured or assumed easily but can be discovered and caught through DL.

Consequently, DL is able to improve the predictive power of endogenous variables & instrument variables.

Exploitation & Exploration are essential to each other

- How exploitation can help exploration?
- Knowing how exploration actually works and how it improves the chance of exploitation is a very complicated process that involves the methodology of how we actually read & understand the exploration.
- Here, you need an ***intermediate variable***, which is unobservable. This is the ***gap*** between where you are and where you can make the exploitation.
- Traditionally, without assuming the ***latent variables***, it is very hard to capture the entire process.
- That is the common understanding of DL, and we can use the ***short-term*** or ***long-term memory*** to basically capture the immediate process.
- Then, instead of just purely using ML to train the model parameters, because we do not have the intermediate variable, we actually use the latent variable to develop a ***tree***, which comes from econometric models at the later part.
- There are actually 2 different connected processes from DL to econometrics. Then the entire model can be achieved, as well as the estimation of the entire parameters on both components.

So, in my opinion, the methodology of DL can be used to enrich the understanding of the generation of data, as well as that of the whole process.

We need to know both WHAT & WHY

- If the past data generation is ***consistent*** with the future data generation, the prediction has been proved to be extremely good, which is perfect. However, ***it is not always the case***. That is the reason why your prediction is not always so good compared to what the model explains.
- There are basically 2 fundamental reasons for that:
  - The 1st one is that the data you are using to generate the predictive model is not consistent with the future data generation.
  - The 2nd reason is that the theory does not apply.
- On the one hand, when the underlying process is ***unperceived***, ML can help with identification & recognition, and it also can be useful in the process of theory building potentially. If you want to explain ***WHY*** it is, you have to tell ***WHAT*** it is first. Something like when Amazon says people who like this would also like something else, there is no reason for why, there is no explanation for it is what it is.
- When doing explanation, people always use the theory, which contains many constraints. For instance, when doing a movie ***classification***, people assume that thriller movies are similar to horrible movies. This may come from some kind of category, which ML techniques are good at and can be of great help.
- On the other hand, econometrics can be helpful to ML with the ***quick converging*** of process. When the processes are based on economic theories, econometrics helps the process converge quickly, which is followed by good identification and bootstrap.
- Although you have a prediction model, it does not mean that you can explain everything, even if your model has a high level of ***prediction accuracy***. Prediction is not only to do with ML.

Why do ML & econometrics need each other?

- ML is very good at generating models ***systematically***. However, it is uncertain whether the cause of inference can be maintained.
- Can we use ML to generate models systematically while maintaining the causal inference?
  - Nobody knows the answer!!!.
- ML can mimic the decision process of the ***judges***. Sometimes, we think it is a little bit too much to solve the underlying societal problems.
- For example, in India, many people are potential crime suspects. The judges have to decide ***WHO*** should go to trial. Imagine, the majority of the people ending up in trial have been found to be innocent. It would have been a waste of resources if the person had been kept in prison earlier.
- Therefore, the judge is trying to use intuition, experience, and all that kind of stuff in the trial process. The ML tries to ***mimic the judge***, instead of the outcomes of the judge, such as their emotion, their eyes, and many other aspects.
- Then, a decision is made about ***WHO*** is going to be devoiced. All you have to do is to use ML to learn, to train the observables.

When you challenge yourself with these theory problems, you can finally solve some societal problems. There are many other issues that can be explored, such as health applications, government issues, and so on.

A cocktail approach can be used in the combination of ML & econometrics

- In modern medicine, there is an approach called the ***cocktail method***. This kind of approach combines 2 or more techniques for a treatment, which contributes significantly to applicable results.
- The way we combine ML & and econometrics is just like a cocktail approach from different disciplinary perspectives.
- They are not equal because they originate from different research foundations. Econometrics seems familiar to us since it has been around for a while. ML comes from a tradition that started from the so-called ***symbolic pattern recognition***, and it has already combined with statistical pattern recognition.
- Specifically, with symbolic processing capabilities, ML gives you more structured information, as well as texture information. You know a little bit about the content, and perhaps it provides explanation from a different angle.
- Consequently, the combination of the 2 can be quite useful. For example, you want to provide a decision maker, no matter what the field is, with the result of your analysis.
- With pure econometrics, most of the time in statistics, the base is a numerical base. With ML, you can provide explanation from many different angles, and you are able to provide the insightful information for the decision maker involved.

As a result, I think using different approaches to help each other out and then understanding the underlying decision-making process is important, not just relying on the predictive ability or the explanatory power.

How to get prepared for the combination of Econometrics & ML as Analyst or Researchers?

Necessary to Master domain knowledge & use tools skillfully

- Econometrics & ML are just ***methods***, and their routes are always joined together with statistics upstream.
- Within computer science, there is ***inductive*** learning. Besides, there are predictive methods, or methods that can make predictions or discover some common patterns that might be taught in data mining classes and ML classes, as well as statistical methods.
- Statistics takes one direction and ML takes one direction, while DL handles large scales of data in the process.
- We can use certain tools that we have trained and that can be applied better than the tools that we do not know. If you truly care about the benefits of combining these 2, you must know the tools very well.
- Suppose, I am trained in ML, I have to learn econometrics and many other tools that remain unknown to me. When I apply econometrics to deal with problems, I need to be very cautious. As a serious analyst/researcher, one has to ***master*** the tools, as well as the usage & application methods.
- Also, ***compare*** them with other tools and find out the similarities. When you solve a predictive analysis problem, you either use the ML method or use the econometrics method combined with cross-validation; you either evaluate model fit or evaluate predictive errors: all that depends on your goal.
- Different communities have different methods that can predict better. I think, that predictive analysis researchers from a ML background have the attitude of improving performance. If you want to have good performance, you need to know the problem very well, as well as the data, the method, and the process.
- It is necessary for you to do numerous evaluations and try different things. It does not mean that all of us need to be very knowledgeable with behavior theory, but at least you have to possess some good domain knowledge, which can help us explain well.
- I will not stop immediately after I have good performance, and I will go out and find some reasons to help explain.
- Today, I am paid to do so because I am a consultant, I definitely have to find some explanation because your stakeholders or your clients ask you that your deep learning predicts so well and why, and whether it can predict better and have hierarchical correlation relationship.

Important to choose appropriate methodologies for problems

- It is very important to choose the appropriate methodology for the problem.
- So, if you cannot do it theoretically, you can use ML techniques, because in my view, it is not true that ML does not provide explanations; ML models do provide explanations, they provide explanations through logic, you know, through logic or business domain knowledge.
- Therefore, you can actually provide good explanations. The question is whether you can interpret them with theoretical explanations.
- When you are using them, it is important whether or not you can explain why and what it is.
- They may not have some statistical relation, such as those of the sample, but they do provide you some insight that might be some potential explanations of certain things and it might be some interesting research opportunities.