**Note for research paper**

* **When and Why Model Understanding?**

Better adoption rate:

It is human tendency not to adopt something that cannot interpret or understand. We are using AI models to remove biasness from human decision making; however, the decision will be dangerous if the outcome is not justifiable, legitimate, and transparent.

* High-stake decision-making setting
  + Impact on human lives/health/finances
  + Settings relatively less well-studied, models not extensively validated
* Accuracy alone is no longer enough
  + Train/test data may not be representative of data encountered in practice
* Auxiliary criteria are also critical:
  + Nondiscrimination
  + Right to explanation
  + Safety
* Auxiliary criteria are often hard to quantify (completely)
  + E.g.: Impossible to predict/enumerate all scenario violating safety of an autonomous car
* Incompleteness in problem formalization
  + Hinder optimization and evaluation
  + Incompleteness ≠ Uncertainty can be quantified

What XML (Explainable machine learning) is trying to achieve:

* **Trust:** The prediction accuracy is a clear function of data quality, true causality, and choice of an appropriate algorithm. However, the models are subject to generating false positives in the prediction process. If the models generate a lot of false positives, then the end user will lose trust in the model. Thus it is important to convey confidence in the model to the end user.
* **Associations:** The ML or DL models learn to make predictions based on associations between various features. The associations can be correlations or mere associations. The correlations that are unexplained are spurious correlations that make the model impossible to interpret. Hence it is important to **capture the true correlations.**
* **Reliability :** Confidence in the **model, the stability of the model in predictions**, and the robustness of the model are also very important. This is required in order to have more trustworthiness in AI models and to ensure that the end user has enough confidence in the model predictions. If this is not available, then no user will trust the models.
* **Fairness:** The AI models should be fair and ethically compliant. They should not discriminate among religion, sex, class, and race in generating predictions.
* **Identity:** The AI models should be able to keep privacy considerations intact, without revealing the identity of an individual. Privacy and identity management while generating XAI is very important.

In a healthcare scenario, if a doctor relies on model predictions and the model takes a bunch of data and labels patients as healthy or sick. If the doctor just looks at the predictions, he or she has no way of knowing which predictions to rely on and which ones they should apply their own judgement to. But if they had a full understanding of what underlying model is potentially doing. Things could be different. **Model understanding helps assess if and when to trust model predictions when making decisions.** Such model understanding can also be helpful for regulatory authority in deciding if a model should be approved for broader deployment or not. **Model understanding allows us to vet models to determine if they are suitable for deployment in real world.**

* Utility
  + Debugging
  + Bias detection
  + Recourse
  + If and when to trust model prediction
  + Vet models to assess suitability for deployment
* Stakeholders benefited
  + End users
  + Decision makers (e.g., doctors, judges)
  + Regulatory agencies (e.g., FDA, European commission)
  + Researchers and engineers
* **Why Explainability is important in healthcare? What’s unique about ML in healthcare?**

Life or death decisions

So, we need robust algorithms (we need to think about whether they are safe, how do we check for safety long-term. What are the checks and balance we should put into the deployment of algorithm to make sure that it still working as it was intended?)

Need fair and accountable algorithms

(Each of the healthcare intervention has money associated to them. You can’t do them to everyone. So, when using machine learning to prioritize who do you give those interventions to. Because health is so intimately tied to the social-economic status. One could think about what happens when these algorithms are not fair. It could have really long-term implications for our society.)

Many questions are about unsupervised learning

Many of the questions we want to answer are causal in nature

**Challenges**

* Because driving labels for supervising learning prediction is very hard, one has to think through how do we automatically build algorithms to do what’s called electronic phenotyping to discover, or to figure out automatically what’s the relevant labels for a set of patients that one could then attempt to predict in the future.
* Because we only have rare data, for example, some rare diseases. Some common diseases present in very diverse ways. We need to think about how can we bring together domain knowledge and how can we bring together data from other areas or to learn sth that we could refine for the foreground question of interest
* There’s a ton of missing data in healthcare.
* Difficult of de-identifying data (need for data sharing agreements and sensitivity)
* Difficulty of deploying ML (due to the challenge of integration)
* Maybe the hospital we want to deploy your algorithms uses the EHR that is not built for your algorithm to plug into.
* **How To Achieve Model Understanding**

Method 1: Build inherently interpretable predictive models (linear regression, logistic regression, or shallow decision trees)

Method 2: Explain pre-built models in a post-hoc manner. (This method is popular in recent times) Basically what it does is passing the model through “Explainer algorithm” and the explainer will give you the important features associated with the prediction.

* **In healthcare, most of the time, we don’t have true label, how do we deal with this kind of situation with unsupervised learning model? And how do we derive the decisions we need to make via this unsupervised learning models.**
* **Why machine learning models haven’t been widely adopted in the healthcare setting?**

1. Poor generalizability

Unlike our social media data, healthcare data is rather limited. Because of the limited amount of labelled data used for training models, the models generally do not generalize very well.

1. There’s a gap between the input they expected and the current clinical workflow & hard to get enough training data

Manually data collection process. Physicians need to manually enter the data into the system, which takes a lot of time and cost a lot of money (time is money). Because data is collected manually, they are difficult to maintain🡪 difficult to generalize our model (with new data)

Explanation: If you go to a new place, and you want to apply the same algorithms, you suddenly have to redrive parts of this model from scratch. For example, the prior probability of the diseases are going to be very different from where you are in the world. Now, you might want to go into different domains outside of primary care, you have to take a lot of efforts to redrive this model. As a new discovery in medicine is made, again, you have to update this model. This has been a huge blocker to deployment.

* **The applications of machine learning in healthcare**
* Diagnosis
* Intervene at the right time
* Predict the progression of a patient’s disease over a period of years
* Tailor treatment for chronic disease
  + Predicting a patient’s future diseases progression
  + Precision medicine (This is an example of a causal question. We want to know how do we cause a change in the patient’s disease trajectory.)
* **What is the future of how we treat chronic disease?**
* **Early diagnosis (e.g., diabetes, Alzheimer’s, cancer)**
* Continuous monitoring and coaching

Diabetes example

Type 1 diabetes, as opposed to type 2 diabetes, generally develops in patients at a very early age. Usually as children when it’s diagnosed. Insulin pump attached to patients. But there’s a really challenging control problem. If you give patients too much insulin, you could kill them. If you give them too little insulin, you could really hurt them. How much insulin you give them is going to be a function of their activity and what food they ate and various other factors. It also presents very interesting opportunities for machine learning. Because right now we are doing a very good job at predicting future glucose levels, which is essential to figure out how to regulate insulin. And if we could have an algorithm that takes a patient’s phone, take a picture of the food that the patient is eating and have that automatically feed into an algorithm that predict its caloric content and how quickly that’ll be processed by the body. And based on the patient’s metabolic system, when should you start increasing insulin levels and by how much. They could have a huge impact on the quality of life.

* **Interpretable Machine Learning method**

Skater

Skater is a unified framework to enable Model Interpretation for all forms of model to build an interpretable machine learning system. It is an open-source library designed to demystify the learned structures of a black box model with globally (inferences on the basis of complete data set) and locally (inference about an individual prediction).

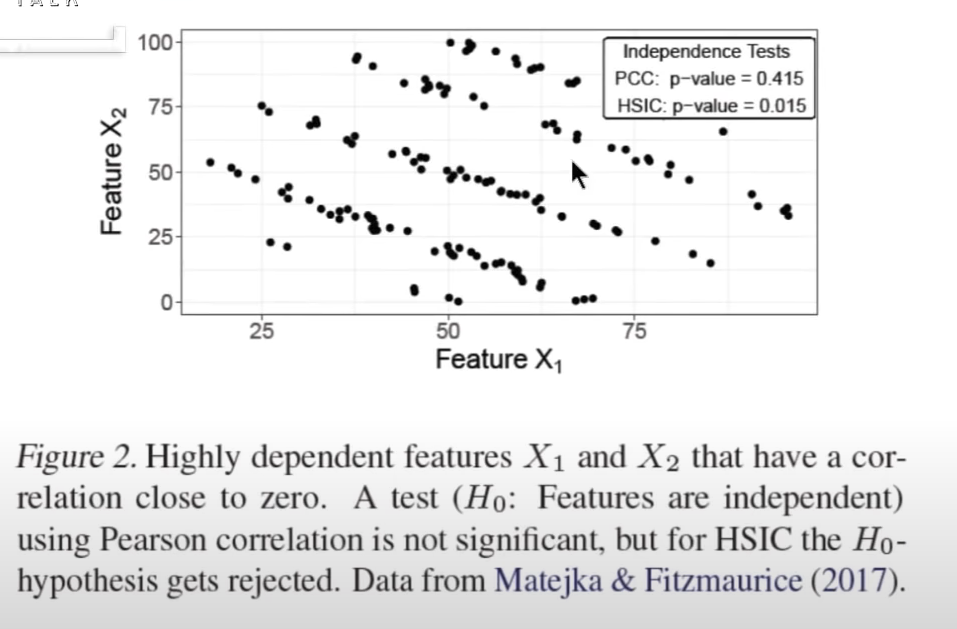
* <https://christophm.github.io/interpretable-ml-book/properties.html> (Chapter 3.5) IMPORTANT!!!)
* **Consistency:** How much does an explanation differ between models that have been trained on the same task and that produce similar predictions? For example, I train a support vector machine and a linear regression model on the same task and both produce very similar predictions. I compute explanations using a method of my choice and analyze how different the explanations are. If the explanations are very similar, the explanations are highly consistent. I find this property somewhat tricky, since the two models could use different features, but get similar predictions (also called “Rashomon Effect”). In this case a high consistency is not desirable because the explanations have to be very different. High consistency is desirable if the models really rely on similar relationships.
* **Stability:** How similar are the explanations for similar instances? While consistency compares explanations between models, stability compares explanations between similar instances for a fixed model. High stability means that slight variations in the features of an instance do not substantially change the explanation (unless these slight variations also strongly change the prediction). A lack of stability can be the result of a high variance of the explanation method. In other words, the explanation method is strongly affected by slight changes of the feature values of the instance to be explained. A lack of stability can also be caused by non-deterministic components of the explanation method, such as a data sampling step, like the local surrogate method uses. High stability is always desirable.
* **Certainty:** Does the explanation reflect the certainty of the machine learning model? Many machine learning models only give predictions without a statement about the models confidence that the prediction is correct. If the model predicts a 4% probability of cancer for one patient, is it as certain as the 4% probability that another patient, with different feature values, received? An explanation that includes the model’s certainty is very useful.
* **Degree of Importance:** How well does the explanation reflect the importance of features or parts of the explanation? For example, if a decision rule is generated as an explanation for an individual prediction, is it clear which of the conditions of the rule was the most important?
* **Novelty:** Does the explanation reflect whether a data instance to be explained comes from a “new” region far removed from the distribution of training data? In such cases, the model may be inaccurate and the explanation may be useless. The concept of novelty is related to the concept of certainty. The higher the novelty, the more likely it is that the model will have low certainty due to lack of data.
* **Representativeness:** How many instances does an explanation cover? Explanations can cover the entire model (e.g. interpretation of weights in a linear regression model) or represent only an individual prediction (e.g. Shapley Values).
* **What is a good explanation**
* Humans usually do not ask why a certain prediction was made, but why this prediction was made instead of another prediction. We tend to think in counterfactual cases, i.e. “How would the prediction have been if input X had been different?”. (Rationale: If my loan application is rejected, I do not care to hear all the factors that generally speak for or against a rejection. I am interested in the factors in my application that would need to change to get the loan. I want to know the contrast between my application and the would-be-accepted version of my application.)
* **Contrastive explanations are easier to understand than complete explanations.**
  1. A complete explanation of the physician’s question why the drug does not work might include: The patient has had the disease for 10 years, 11 genes are over-expressed, the patients body is very quick in breaking the drug down into ineffective chemicals, … A contrastive explanation might be much simpler: In contrast to the responding patient, the non-responding patient has a certain combination of genes that make the drug less effective. The best explanation is the one that highlights the greatest difference between the object of interest and the reference object.
  2. What it means for interpretable machine learning: Humans do not want a complete explanation for a prediction, but want to compare what the differences were to another instance’s prediction (can be an artificial one).
* **Explanations are selected. People do not expect explanations that cover the actual and complete list of causes of an event.** We are used to selecting one or two causes from a variety of possible causes as THE explanation. But it also means that there is more than one selective explanation why a certain prediction was made.
* **Explanations are social. They are part of a conversation or interaction between the explainer and the receiver of the explanation.** The social context determines the content and nature of the explanations. If I wanted to explain to a technical person why digital cryptocurrencies are worth so much, I would say things like: “The decentralized, distributed, blockchain-based ledger, which cannot be controlled by a central entity, resonates with people who want to secure their wealth, which explains the high demand and price.” But to my grandmother I would say: “Look, Grandma: Cryptocurrencies are a bit like computer gold. People like and pay a lot for gold, and young people like and pay a lot for computer gold.”
  1. What it means for interpretable machine learning: Pay attention to the social environment of your machine learning application and the target audience. Getting the social part of the machine learning model right depends entirely on your specific application. Find experts from the humanities (e.g. psychologists and sociologists) to help you.
  2. What it means for interpretable machine learning: If one of the input features for a prediction was abnormal in any sense (like a rare category of a categorical feature) and the feature influenced the prediction, it should be included in an explanation, even if other ‘normal’ features have the same influence on the prediction as the abnormal one. An abnormal feature in our house price prediction example might be that a rather expensive house has two balconies. Even if some attribution method finds that the two balconies contribute as much to the price difference as the above average house size, the good neighborhood or the recent renovation, the abnormal feature “two balconies” might be the best explanation for why the house is so expensive. What it means for interpretable machine learning: The explanation should predict the event as truthfully as possible, which in machine learning is sometimes called fidelity. So if we say that a second balcony increases the price of a house, then that also should apply to other houses (or at least to similar houses).
* **Good explanations are consistent with prior beliefs of the explainee.** Humans tend to ignore information that is inconsistent with their prior beliefs. What it means for interpretable machine learning: Good explanations are consistent with prior beliefs. This is difficult to integrate into machine learning and would probably drastically compromise predictive performance. Our prior belief for the effect of house size on predicted price is that the larger the house, the higher the price. Let us assume that a model also shows a negative effect of house size on the predicted price for a few houses. The model has learned this because it improves predictive performance (due to some complex interactions), but this behavior strongly contradicts our prior beliefs. You can enforce monotonicity constraints (a feature can only affect the prediction in one direction) or use something like a linear model that has this property.
* **Why switching from analyzing assumption-based, transparent model to analyzing assumption-free black box model?** 
  + Because making all these assumptions are problematic: They are usually wrong (unless you believe that most of the world follows a Gaussian distribution), difficult to check, very inflexible, and hard to automate.
  + For the black box approach, we do not make assumptions, we approximate reality as close as possible (while avoiding overfitting the training data).

[**https://www.youtube.com/watch?v=0LIACHcxpHU&t=310s&ab\_channel=MachineLearningStreetTalk**](https://www.youtube.com/watch?v=0LIACHcxpHU&t=310s&ab_channel=MachineLearningStreetTalk) **(Interpretable Learning—Christoph Molnar)**

* As we start to use these more powerful non-linear models to make decisions on real-world matters, it’s inevitable that our attention must now turn to interpretability and expandability.
* There is a whole plethora of techniques out there to explain why a model made a certain prediction, some models like low dimensional linear regression are intrinsically interpretable. You could just look at the mode coefficients and that tells you how the model is working under the hood.
* **Example-based explanations:** try to find the smallest change in the input data that would cause the output predictions to change.
* **Interpretability is often a deciding factor when a machine learning model is used in a product a decision process or research**
* **Interpretability methods can be used to discover knowledge to debug or justify a model and its prediction to control and improve the model to reason the potential biases in the model as well as increase the societal acceptance of models. But interpretability methods can be quite esoteric. They add additional layer of complexity and the potential pitfall require expert understanding.**
* Machine learning models are inherently less interpretable than classical statistical models but typically they have a better predictive performance and that’s because of their ability to handle non-linear relationships and also higher feature interactions automatically.
* Complexity of the model vs. the lack of ability to understand it
* Simplicity model approximations can often mask important information and be misleading.
* In classical statistic, there’s an entire field called model diagnostic to check the assumptions and simplifications have not been violated. This is something that does not yet exist in interpretable machine learning
* **Is it even possible to understand complex models or even human for that matter in any meaningful way?**
* He also focuses on the broader impact of interpretability and what interpretability even means.
* He points out feature dependence where you have shared information between features
* One of the most important challenges with interpretable machine learning is “feature dependence”. Molnar points out that **feature dependence makes attribution and extrapolation problematic.** This is exactly what happens in partial dependency plots. We are basically extrapolating and we are creating fictitious data points that didn’t really exists. This fictitious data points probably exist outside of the data distribution. So Molnar thinks that the model we built should reflect the causal structure in the world but of course that is not really the case most of the time. He points out that the statistical learning is just reflecting surface feature correlations not the true causal structure beneath these scenes. Causal structure would be more robust if we could actually capture them. The predictive performance and learning causal factors is a conflicting goal which I think not many people have thought about.
* We need to think about when we can make causal interpretations and a lot of work is underway in this field. But being completely frank, this is very nascent. There is not very much out there at the moment.
* **Molnar also points out the lack of statistical rigor. Most IML methods do not even give you confidence estimates, something which is completely standard in the statistical world.**
* Model and explanations are computed from data which means that they are subject to uncertainty. **We need to be making distributional and structural assumptions.** He points out the risk of p hacking, something which is prevalent in the natural sciences. This is something that could be coming to the world of iml (interpretable machine learning) very soon if we don’t start thinking about this more carefully.
* Monlar also points out there is no accepted definition of interpretable machine learning methods. It’s not entirely clear that how we can compare iml methods to machine learning models.
* It’s really easy to access machine learning models because we have benchmarks and we have true labels. But we can’t really quantify how correct an explanation is and it doesn’t really help there is taxonomy. And you need to have technical knowledge to understand these assessments.



* The setting of machine learning method is too static doesn’t reflect how these models are actually used in practice.
* **I’d like the idea about thinking the process rather than thinking about just the model.** We need to have a holistic view of the entire process. We need to think aobut how we explain predictions to folks from diverse backgrounds. How do we have interpretability at the societal level or at the institutional level. Thinking more broadly than at the moment. We need to reach out to other disciplines such as psychologists and social scientists.
* **Paper: Pitfalls to Avoid When Interpreting Machine Learning Models.**
* In this paper, he points out that there is a growing number of techniques providing model interpretations but many will lead to the wrong conclusions if used incorrectly and he pointed out many of those pitfalls. For example,
* (1) assuming the model generalizes well or assuming that the model has been fit correctly. If the model is underfit or overfit, the interpretation method will perform badly as well. The interpretation can only be as good as the model underlying it.
* (2) Unnecessary use of complex model. Which is to say the use of opaque or complex machine learning model when the interpretable model is sufficed, which is to say that the performance of interpretable model is only negligibly worse than one of these black box models. (Clinical decision support decision tree – internship slide (give an example). He recommends to start with simple, interpretable models (generalized linear models or lasso models, additive models, decision trees or decision rules) and gradually ratcheting up the complexity as required.
* One of the things I don’t like about machine learning is the laziness. I think we should always seek to understand and simplify problems wherever we can. We should always striving to create the most elegant and simple and maintainable solution. We shouldn’t overcomplicate things. (The kiss principle is very generalizable here)
* (3) Ignoring feature dependence. This is the problem many IML methods have. He gives an example of partial dependency plots where they extrapolate in areas where the model has little training data. It can cause misleading interpretations. Perturbations produce artificial data points that are used for model predictions, which in turn are aggregated to produce global interpretations.
* (4) Confusing correlation with Dependence. He gives an example. Pearson correlation coefficient (PCC) close to zero can still be dependent and cause misleading model interpretations. While the independence between two features implies that PCC is zero, the converse is generally false.



* **(5) Misleading effect due to interaction.**
* If a machine learning model performs well, why don’t we just trust the model and ignore why it made a certain decision. The problem is that a single metric such as classification accuracy is an incomplete description of the most real world tasks.
* **Paper: Interpretable Machine Learning – A Brief History, State-of-the-art and Challenges**
* **The explanation could be deceptively good. The sort of cognitive bias maybe that we have to look for contrastive explanations or counterfactual explanations like in principle it seems good.**
* What we can learn from the social sciences about what a good explanation is. They should be contrastive. They should be short. But they should also confirm to some prior knowledge that the people have. A lot of things that you wouldn’t say are good explanations in some sense. Maybe it’s not a good explanation that fits the priori knowledge. Because it’s not the correct one maybe. It’s important to think about the human side of it.
* **What we really want after the explanation.**
* **There are many dimensions that you can judge explanations.**
* If we can’t quantify how good an explanation is then where we are really? There are objective evaluations like sparsity and interactions, strength, and fidelity and human centered evaluations, which might come from domain experts or laypeople.
* **His view on interpretable machine learning:** giving interpretability or bringing interpretability to machine learning—bundle all the methods together, kind of aim to reduce this high dimensional function to something in a lower dimension. So we are kind of doing this mapping, something gets lost in a way this is fine. And part of the analysis is to find out what art gets lost. For example, when we look at the feature importance values, of course, it’s a summary of your model and a lot of information gets lost.
* Could it be dangerous? You gave an example of random forest, when you have a lot of shared information between the features. It will actually tells you that correlated features have a higher feature importance than you might otherwise expect.
* **For feature importance, you kind of measure how well your model performs and then you measure again after you shuffle one of the features.** You need to understand what happens when you shuffle the feature. You kind of break the association between the feature and the prediction because now it doesn’t carry the information about the target anymore. Because you’re shuffling it randomly in the model. So the feature importance now measures how much performance you lose because of this break of information. But you also need to understand that the shuffling also breaks the association with your other data. This is the limitation of the method. I think what is needed is that we understand in which way this methods break or in which scenarios we are allowed to use them or how we are allowed to interpret them. The situation is kind of similar to statistics where you have models and then you interpret the coefficient of the models. You still need to learn how to do the interpretation and what are the assumptions that need to be met that you’re allowed to make this interpretation.
* **He made a really good point in the book, which is that even these so-called intrinsically interpretable models are only interpretable to a certain dimensionality. As the dimensionality goes up, there is no model that is intrinsically interpretable.**
* **Partial dependence plot. At what point we are just developing a complex math model to explain complex model and we haven’t made much progress along the interpretability axis.**
* **For partial dependence plot.** The intuition is that you do some intervention on your data, so for one feature, you will replace all values with one fixed value and kind of get the average prediction that you get afterwards and do it for a lot of points and you’ll get a curve. So, you’ll get an expected change over the feature range.
* **When you have feature dependence, it makes attribution and extrapolation problematic.** A simple example in the medical field that height and weight are highly correlated.
* I want to make a distinction between we understand what’s going on inside and doing sensitivity analysis where we just try out what happens in certain scenarios. So feature importance is basically to see how does it behave if we break some features and then we rank features by this as an importance.
* Interpretability can help us construct a better machine learning model.
* We need to be more rigorous and there is no quantification of uncertainty with the current IML methods. When you have models and explanations that are computed from data and they are subjected to uncertainty and that’s just not modeled at all at the moment. We need to make some distributional and structural assumptions that we’re not making now. And you point out there’s this phenomenon of p hacking, which is a big problem in natural sciences which hasn’t quite mad its way to IML methods yet but probably will do.
* **I still think it’s better to have not only just one number or one explanation but also have the distribution to do this explanation or this number and to quantify uncertainties behind computing this number. For example, when we have linear model, we have coefficient. But we don’t just look at coefficient, usually we will also look at the confidence interval. But we don’t do it at the moment for interpretability. So, if you have like a feature importance value and you get some result. How much variance is behind it? If I were to use slightly different data or refit the model again, how similar would the number be? I think that’s something that should come to the interpretability as well.**
* If you have a table of accuracy; however, there’s no accuracy attached to it, then you should be suspicious of it. Because if you retrain your neural network with different seed, you might end up with a different accuracy at the end. If you want to day a method is better than another method. You want to quantify how large ranges of uncertainty. You need to try a lot of things such as choice of data, choice of splitting points and training and test data and weight initialization and so on.
* There are three kinds of lies. Lies, damned lies, and statistics.
* Fundamentally, whenever we go measure data and we have model. What we’re actually able to extract from that data and the model is inherently probablisitic. It’s a probability distribution. And we get into trouble any time we try to take that probability distribution and project it to numbers i.e statistics. It doesn’t matter whether it’s a mean or confidence interval. The fact is that we’re throwing away information. Oftentimes, we have to use these probabilistic things to reach a decision and as soon as we get to the point where we have to make a concrete decision, we’re forced to project it. But along the way it’s important not to lose sight of the fact that we’re throwing away information.
* People said in the reinforcement learning, you can learn causal factors. But that’s not really true. What you’re learning is a surface representation of causal factors.
* The goal of the model is that it should reflect the causal structure. Most statistics just reflect surface feature correlations, they don’t even scratch the surface of what we want.
* You also need to decide the goal of your model. Do you want a causal interpretation?
* There can also be good reasons to include “non-causal” features into your model. If your goal is prediction and some features might help you with a good prediction but it might not be causal at all.
* The problem is when we are using deep learning model. They have some structures that probably has no relationship to the real world. Causal factors do generalize much better.
* It would be nice if the machine learning methods could indicate that there may be the possibility of a causal structure.
* There is a difference between “a causal factor” and a “causal structure”. I think that challenge is that we don’t have enough fidelity in the structure.
* But even with the structure learning, you’ve got this adjacency matrix and all of those nodes that you’ve already come up with a priori. So what you want to learn is the node itself.
* If you don’t even know what the features are. If you have a convolutional neural network. There is this issue that you have this entanglement between concepts. For example, a frisbee is always on the same image as a dog. Maybe a neural network can’t even separate these two things because they are too entangled in the data to discover the structure.
* The setting of the machine learning model is too static. It doesn’t reflect how these models are used in reality. For the scientists, it really easy to use the fixed models to geek out, but in reality you have to interact with institutions and people who will be affected by the model.
* How to better explain and interpret ML models so that human beings can have that oversight because that’s the only thing that can give us comfort as a society.
* We can have a very superficial discussion. You can say well we need to be able to represent the reality better than we do and we have a whole tool set here to identify sources of bias or lack of robustness et cetra. But it’s so much more complex than that, because these models are used in a very complex process and you get this very complex dynamics emerging as a result of that. I think we’re just scratching the surface of understanding those dynamics.
* Science tends to study things in isolation. Study one type of model, study one type of adjustment for a deep neural network.
* Interpretability metrics whatever they are (could be partial dependency plot). You can actually build in some requirements of those into the objective function when you train the model.
* Inductive bias
* Interpretability vs. Cheaper to deploy
* **Going to the direction of automatic machine learning. You don’t just get the best performing one but you have this pareto set. We have multiple objective you want to hit then there’s not one model that works the best but you have a set of models that have different trade-offs between these objectives and then you have to decide what is the trade-off that you want to make.**
* I’m skeptical about Lime because it’s difficult to parameterize your local model.
* Shapley value. If you have a model that predicts someone’s income. If you have salary in the model twice, then the shapley value will be divided between the two duplicate field.
* With the IML methods, we’re kind of compressing information down into a representation. That’s the transport that can be understood by different people.
* What we really need is an operating model, or a set of guidelines on how to implement these tools. How do I identify sources of problematic correlations. We need to have a database of problematic correlations. Having a tool that allows me to identify and mitigate biases frankly is useless. What do I do with that? Many of the machine learning cloud providers whether it’s Data IQ, Azure ML, and Sagemaker. They all have this interpretability built in now and it’s just a box ticking exercise frankly. It’s completely useless, there’s no existing guidance on how these tools should be used. So, if I were a large company and I’m building an operating model around how to implement fairness techniques. Just having the technology is irrelevant. It’s about people and the process and the kind of operating model of how we implement it. And there’s basically no useful information out there to help us do that.
* The other thing is we spoke about this becoming an engineering discipline, which is to say what if we could create an interface to abstract away some of the vagaries and esoteric interpretability methods. We might come up with some primitives or some common language and then we can hide the complexity behind the interface. This is kind of what we do with ML Devops already. We automate as much as we can then we templatize and remove friction out of the process. We even create building blocks using domain specific languages.
* This might just become a box ticking exercise and this is something that we see in security and AI ethics already. We can’t really trust people to self-report that the model is behaving correctly or that the project has no concerns from an AI-ethics point of view. The whole point here is process. If we want to create an operating model and ensure best practices are followed or any kind of standardization in a large organization, we have to design a process and many eyes make shallow holes. The process will mandate that a certain number of stakeholders were involved in accessing the particular IML technique and validating it essentially and then we would need to record that assessment (who said what when) and the if the company ever became audited if they got forbidden there was some kind of IML problem where there were some problems with the IML methods, then we can rewind the clock and see how can we change the model.
* [**https://ssir.org/articles/entry/the\_case\_for\_causal\_ai**](https://ssir.org/articles/entry/the_case_for_causal_ai) **(The Case for Causal AI)**

Therefore, it is imperative that decision makers also consider another AI approach—causal AI, which can help identify the precise relationships of cause and effect. Identifying the root causes of outcomes is not causal AI’s only advantage; it also makes it possible to model interventions that can change those outcomes, by using causal AI algorithms to ask what-if questions.

Another challenge with using only predictive models is a fundamental lack of knowledge about why they make particular predictions in the first place. Deep learning was inspired by how human brain cells are organized (in “layers”) and how they communicate with each other (taking input signals from cells of one layer, transforming the signals, and outputting the transformed signals to cells of another layer). Unlike commonly used methods for predicting outcomes—such as [regression](https://hbr.org/2015/11/a-refresher-on-regression-analysis), a traditional statistical technique that maps the relationships between variables to the predicted outcome with a single best mathematical formula—deep learning can map variables to outcomes with much more complex relationships between them. By combining multiple layers between the input variables and outcomes, deep learning algorithms can learn input-output relationships far more complex than a single mathematical formula and use them to predict outcomes. However, the links and intermediaries between variables are “black boxed,” meaning that the users—and even the creators—of the algorithms cannot easily discern how the variables relate to the outcome and to each other. This means it is often impossible to know which input features deep learning models have used to make their predictions.

This opacity is unacceptable when dealing with the trajectory of people’s lives, such as in the US criminal justice system.

**Recidivism Score:** For example, low income is correlated with crime, but that does not mean it causes crime. Yet people from low-income households may automatically be assigned a high recidivism score, and as a result they are more likely to be sentenced to prison. Fixing the criminal justice system requires a focus on understanding the causes of crime, not merely its correlates. **A closer look at causal AI will show how it can open up the black box within which purely predictive models of AI operate. Causal AI can move beyond correlation to highlight the precise relationships between causes and effects.**

**The Limitation of Randomized Controlled Trials**

The importance of testing causality is not new in either the health or development sectors. A straightforward way to do it is to conduct an intervention in people randomly assigned to one population group, known as the treatment group and conduct no intervention in an otherwise identical group, known as control group. (Vitamin D example: strong association between “Vitamin D deficiency” and “increased risk of diabetes, hypertension, cardiovascular disease, and cancer”.

Limitation: Large groups of individuals are required to ensure that the results aren’t biased or affected by coincidental, outlier characteristics such as age, sex, health status, or educational level.

* 1. This tends to make such trials extremely expensive (in the millions of dollars) and time-consuming (they can take years to conduct).
  2. RCTs can test the effect of only one or at most a few bundled interventions at a time, despite the fact that health and social outcomes are complex, with many underlying drivers.
  3. They can only predict only whether an intervention will cause an effect on a typical member of the treatment group, **not a specific individual. (Hard to trust)**

This is where causal AI comes in. It offers new opportunities to test causality in individuals and population group faster and more efficiently, along with the ability to unravel the underlying complexity. It allows the researchers and program designers to **simulate** the intervention and infer causality by relying on already available data.

**Two approaches to Discovering Causality**

* **Potential Outcome Framework**
* **Causal Graph Models**

To understand the two methods and how they work—as well as their differences—consider the following hypothetical scenario: **Researchers want to discover if an antismoking advertising campaign persuade people to quit, but there was no control group because the ads were released nationally. They only had a data set showing whether individuals were exposed to the ads, whether they gave up smoking, and information on their demographics, and other health**