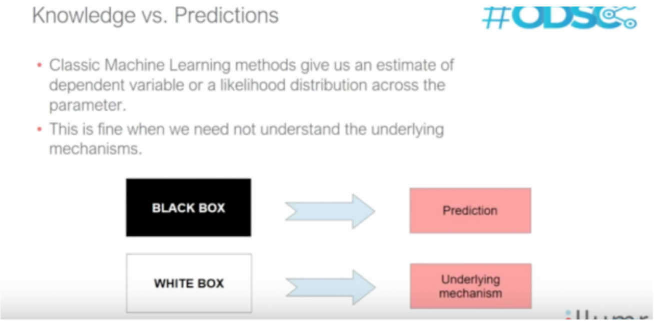
# Explainable AI: From Prediction To Understanding

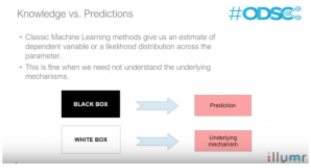
ODSC - Open Data Science Mar 20, 2019 · 5 min read ★



Knowledge vs. Predictions

- Classic Machine Learning methods give us an estimate of dependent variable or a likelihood distribution across the parameter.
- This is fine when we need not understand the underlying mechanisms.

It's not enough to make predictions. Sometimes, you need to generate a deep understanding. Just because you model something doesn't mean you really know how it works. In classical machine learning, the algorithm spits out predictions, but in some cases, this isn't good enough. Dr. George Cevora explains why the black box of AI may not always be appropriate and how to go from prediction to understanding.



Knowledge vs. Predictions

- Classic Machine Learning methods give us an estimate of dependent variable or a likelihood distribution across the parameter.
- This is fine when we need not understand the underlying mechanisms.

*[Related article: The Importance of Explainable AI]*

## Why We Need To Understand The Data

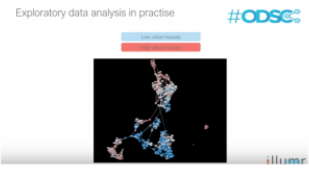
So why do you need explainable AI? Cevora outlines two primary reasons companies often need explainability

- **Human Readability:** When you're making decisions for a company, your director or CEO isn't interested in the data itself. Instead, they may be looking for reasons behind the interpretation and specifically what to do about it. Explainable AI gives the reasoning behind certain decisions, and that can both increase transparency and help offer better business understanding.
- **Justifiability:** In Europe, hiring and firing are often driven by large data sets, but employees have the right for a clear justification with any decision made that involves them. If you don't know how the machine came to a conclusion, you don't satisfy this fundamental right and could be subject to legal consequences.
- **Discrimination**: It's possible to unintentionally replicate discrimination through data sets. Working with black box makes it difficult to uncover this discrimination, and while white box isn't a guarantee against it,

transparency does help reveal areas of discrimination. Discrimination can also take the form of feedback loops.

- **Facilitating improvement**: Black box models don't always list reasons behind the predictions. That can be good for prediction's sake, but it makes it difficult to fix any problems that may come up. If your employees are leaving, black box machine learning may be able to predict who will leave, but it won't tell you why.
- **Eliminating Overfitting**: Black box models don't always pick up the right kinds of relationships. Attempting to understand what relationships actually work and which ones aren't valid can help you teach your machine to make better predictions overall.

How can we really understand this data? We can use classical scientific concepts such as inductive reasoning, but as data scientists, complex systems are too large for us to comprehend. The pieces are interacting, but we're only able to understand small pieces. Machines can pick up the bigger picture. Inductive systems don't lend themselves to these broad data sets, whereas deductive reasoning can help us derive conclusions that we may otherwise miss.



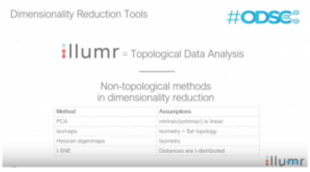
Exploratory data analysis in practise

As an example, Cevora looks at a data set outlining high value and low-value houses. In traditional machine learning, he would be attempting to predict *which* houses would be of high value. Instead, he's looking at *why* certain houses are more valuable than others. This type of white box machine learning may help reveal patterns that wouldn't be available otherwise.

## Methods For Understanding Data

There are several tools for increasing the explainability of the data.

## Non-Topological



Dimensionality Reduction Tools

illumr = Topological Data Analysis

Non-topological methods
in dimensionality reduction

| Method | Assumptions |
|---|---|
| PCA | intrinsic(extrinsic) is linear |
| Isomaps | Isometry + flat topology |
| Hessian eigenmaps | Isometry |
| t-SNE | Distances are t-distributed |

Dimensionality reduction is a crucial part of understanding the data because our minds cannot understand anything more than 3D. Four dimensionality produces data that we cannot quite comprehend, so instead, we remove the extrinsic space to understand patterns in intrinsic space. Think of all the readings of weather across all data from multiple weather stations versus the reasons behind why the data looks the way it does. Reducing the information to the simple description is sometimes called manifold learning.

When reducing these weather patterns, Cevora outlines a few different ways to reduce dimensionality in a non-topographical sense. PCA doesn't work because there are no real linear patterns with the weather. T-SNE could be a better method, but it assumes that distances are T-distributed.

Isometry could also break for something like weather because weather is highly non-linear.

## Topological Data Analysis

TDA Retains Local Features

The rabbit shown below is a 3D scan

- Illustrates how features are exposed

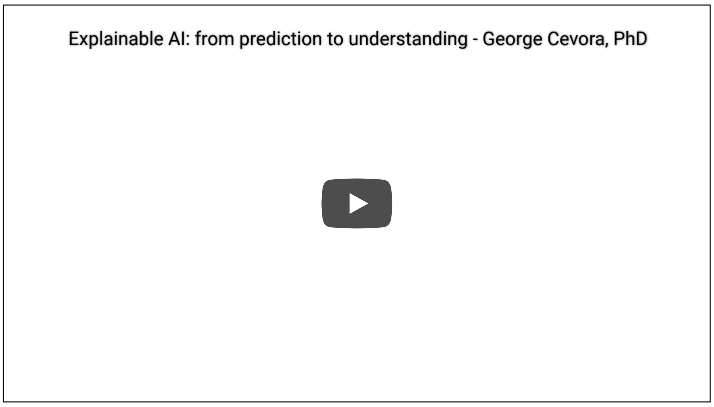
Illumr would use topological methods here instead. Non-linear data can still look linear on a very small scale giving you a better insight into data that doesn’t follow a linear pattern. Measurements are taken in arbitrary units, and there is no straightforward relationship between those small distances and what’s going on.

TDA is good at preserving local features, which could illuminate new insights into data. It structures data in a useful way and can help clarify what the fundamental drivers are with each group.

***[Related article: The 2019 Data Science Dictionary — Key Terms You Need to Know]***

## Wrapping Up

You have to understand what your machine is doing in many cases because of things like justifiability, human readability, and reducing discrimination. It’s vital to view your data sets as a clue to understanding the fundamental drivers for what’s causing the prediction or pattern. While it may not be useful all the time, it’s more and more necessary as our algorithms get smarter. We need to retain the human check on data analysis and understand, not just predict.



Explainable AI: from prediction to understanding - George Cevora, PhD

*This video was taken at ODSC London 2018 — attend* ***ODSC East 2019*** *this April 30 to May 3 for more unique content!* ***Subscribe to our YouTube channel*** *for more videos taken at past conferences.*

Machine Learning Artificial Intelligence Deep Learning Data Science Technology