**introduction of Data Science**

You know when you are doing Data Science when your predictions eventually get boring.

Hello, everybody. My name is John Thomas Foxworthy, and I am the Data Scientist on the Data Team along with Shaya, Tim, Kenley, and Ajay. I have many years of experience in the Financial Services Sector in Big Banks, Hedge Funds, and Consulting. I once had the title of Econometrician, and I have bachelor’s degree from the Department of Economics from UCLA, specializing in Econometrics and Foreign Exchange Forecasting. I am currently doing a Master of Science in Data Science at Northwestern University, on a part – time and remote basis, and I have recently qualified to specialize in Artificial Intelligence. Separately, outside of PeerStreet and School, I have my own person algorithmic trading strategy on stocks and currencies on my personal computer, where I have already coded up 8 models for prediction. (Okay, enough about me).

**Definitions**

Let’s get into it. Here is a short outline of today’s presentation. (A lot of these items are a single page, and overall, I will keep this short.) Regardless of background, I want everybody to gain from this presentation and ask questions.

We will be revisiting this topics throughout this presentation, but these are short and concise definitions for all components of the Data Science practice. Data Science sits on top of Statistics like Statistics 2.0 with more data and computing resources. ML is a bot that makes decisions and predicts with no more programming code. DL is ML, but has a more layered approach, hence depth in deep learning. Both ML and DL can have the same single output, but DL is more involved. Predicting a future number, category, text, or image is Supervised Learning, but if you cannot do that because of data issues, then you focus on describing with Unsupervised Learning. If you can predict something, but your data still has concerns, then you are stuck in between with semi – supervised learning (the half way transition point.)

The new kid on the block is RL, Reinforcement Learning, which I will get into later and AI the superset of everything above.

**Definitions and their Evolution**

So how did these definitions evolve?

To provide some historical context that you can relate to I have these moon shapes of the evolution of Data Science. We first began before modern computing and human beings trying to predict, forecast or surmise a future event, i.e., Supervised Learning. Kind of like Astrology. Next, the modern age of data coupled with prediction in Phase Two, kind of like Astronomy. Lastly, Phase Three (kind of like Astrophysics) is the current phase as of last year, 2020. I will talk about RL at the end, as it currently used for gaming, and other applications that surpasses ML. ML is a program, while RL is an agent.

**The Current State of Definitions of Data Science**

This is my world, I swim in this ocean. The children of AI is ML, DL, and the three orbs of SL, UL, and RL. Reinforcement Learning is the task of trail and error with no human being labeling the data, no human collecting the data or explicitly designing the data like UL, but then it acts and predicts like SL. There are many overlap scenarios, which explains this diagram.

Once you put a prediction model in production in Supervised Learning, then you can upskill to Deep Learning, to provide ML business and AI narrative. Or start with Unsupervised Learning, then follow the same route.

**Definitions Breakdown with (Un) Supervised Learning**

I said a lot so let me try to relate so more to you guys. Everybody has a different experience with education and it all can be traced to Data Science, because Data Science like any department of Statistics is inter disciplinary. Statisticians work in Finance, Health Care, etc.

Businesses inputs capital and labor to output a product, engineers drive water in a river dam facility to hydroelectric energy outcome, mathematicians and statisticians regress independent variable to explain regressand or dependent variable in a function. Clinical Psychologists diagnose psychological disorders by predicting a response in a psychological experiment. Global Warming explains the focus on Climate Change. Reading the description by the Federal Reserve statement can predict if they will raise or cut interest rates. In sum, Machine Learning encompasses all of the above. Any questions? Does that make sense?

**Relabeling Defines Data Science and its Purposes**

Machine Learning comes directly from Statistical Learning.

Learning is fitting equations. Remember y = mx + b in school, finding the slope, the rise over the run fits an equation. Model evolves to network for Deep Learning Neural Networks and graphs for image processing.

Density Estimation (Show GitHub)

**3. Data Science Motto**

There is no loyalty at all for models in Data Science, because the most important value is Accuracy. All models are wrong, but some are useful is from a Statistician who came up a time series model for prediction half a century ago.

**Prior Job Titles**

Data Science is not new and has culminated from a prior set of professions and their influence work up until today.

First let’s talk Pschyometrics. . . Thanks to the original Psycho movie in 1960 staring Anthony Perkins and its immense influence in cinema and television up until today. The job title Pschyometrician sounds like. . “What you kill people them you measure them. You are a monster.” So the job title changed to clinical psychologist. Yeah, no kidding.

**Where did Data Science come from?**

Yes, it is one guy who did this . . . the origin of the word of Data Science has been widely accepted by William S. Cleveland, a professor at Purdue University, in Indiana. Originally, in verbal discussions in 1999 at a Statistical Symposium with other professors, until it was voted on in 2001.

The justification at the time was more processing power because not every regression equation can be done on paper and the growth of data. Today, we would add an additional, and third reason is the rise of open source software couple with freely available data.

**Supervised Learning and Predictive Analytics**

Enough with the appetizers, this the main course. Prediction. We forecast 4 things, a future number like tomorrow’s stock price of a company, a category like loan application acceptance or not, a text like I said earlier about the Fed statement to cut or not cut interest rates and image processing like facial recognition.

Linear Regression is 200 years old, starting in the early 1800’s and the most widely used in every industry and the most common in real estate analytics. This is unfortunate because it is the least accurate, even though it is the most interpretable.

(Show Kaggle Survey 2021)

Ridge, Elastic Net, and least absolute shrinkage and selection operator are workaround solutions or commonly referred in the industry as Regularization and invented in the 1960’s, 70’s and 80’s when data is much less available. Let go to the next slide to help here.

Regression draws a line of best fit between data points and tries to minimize errors as it struggles with outliers, randomness and of course non – linear behavior like defaulting on a loan. A linear regression will underestimate a loan default and overestimate a loan recovery because it is averaging on a normal distribution.

Next, Support Vector Regression was a game changer because it can handle outliers very well. It is instance based and it classifies data by polarity. What I mean by that, it draws a line called a hyperplane to separate two clusters of data and the datapoints near the line are support data points as its measure the magnitude, i.e., the vector, of the other datapoints away from the line. It tries to interject multidimensionality when doing this . . . but it is arbitrary and difficult to interpret why it drew the line in the first place. It can struggle with large data sets.

Random Forest is excellent with random data such as what is the reason somebody who is not a customer lands on the PeerStreet website? Hard to determine why they landed on the website so don’t use determinism and then use a random processing model. Great with non – linearity, interaction with other variables and high dimension (i.e., many columns). It also can generate random number if you have missing numbers in your dataset, but it is a black box so you cannot inspect intermediate values. (ENSEMBLE)

Decision Tree is a flow chart approach to separating data into two categories at a time. Like Random Forest, it also processing with missing data. In addition, there are data convenience issues with no data normalization requirements or even scaling. Great with non – linearity.

Boosting. Boosting is the future. (ENSEMBLE) GBM, XGB and other related models is a big game changer because it introduces several intermediate models iteratively before fitting the equation, this is going from a poorly fitted weak learner to a strongly fitted strong learner.

“Is the right slope, no the error is too big. How about this slope, better, but try again.”

The vast majority of ML models is all about data, then model, and then output in a sequential, one, two, and three. Boosting iterates from two to three (model and output), . . . then back to model, output, then back to model, output. It was basically invented at the University of Washington by a PhD student in 2016. Written in C++ with a Python Wrapper.

Bayesian Models do not work well for us because it is for smaller datasets and assumes there is no correlation in your datasets. Let’s move on . . .

Deep Neural Network or DL is the most accurate, but also the least interpretable so the exact opposite of Linear Regression.

The rest are classifiers

**Reinforcement Learning Method of Artificial Learning**

In 1961, Robert Anson Heinlein (pronounced Hineline) wrote a novel called, “Stranger in a Strange Land” that influenced many people in different sectors, including music. He was an aeronautical engineer, navy officer, he studied Physics at UCLA for his bachelor’s degree.

For you heavy metal fans, Iron Maiden has a song called Stranger in a Strange Land. (He also wrote a book called the Number of the Beast, also a Iron Maiden song). The main take away is something called, “grokking”, or to grok, which describes empathy, not to confused with sympathy. Empathy leads to agency, and thus our model here. (Whether you are nice to somebody or not as a form of sympathy, is independent from you putting yourself into somebody else’s shoes like empathy.)

The last side talked about networks, and prior slides about modeling, but all models, networks are an abstraction in reality no matter how perfect they are in prediction. What we have here is no model, no abstraction, but a embedded machine, recording, learning (which is fitting equations), then replicating a behavior of an agent.

Just about everything, we have talked about so far is prediction. The perfect answer in a selection . . . so that is not human.

Reinforcement Learning is about second, third or fourth best answers to questions, that is suboptimal solutions to problems. Like a video game.

<https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b>

Text, letter

Description automatically generated

A human being collects information about customers and trains an unsupervised learning model to segment the customer information to uncover underlying relationships in customers.