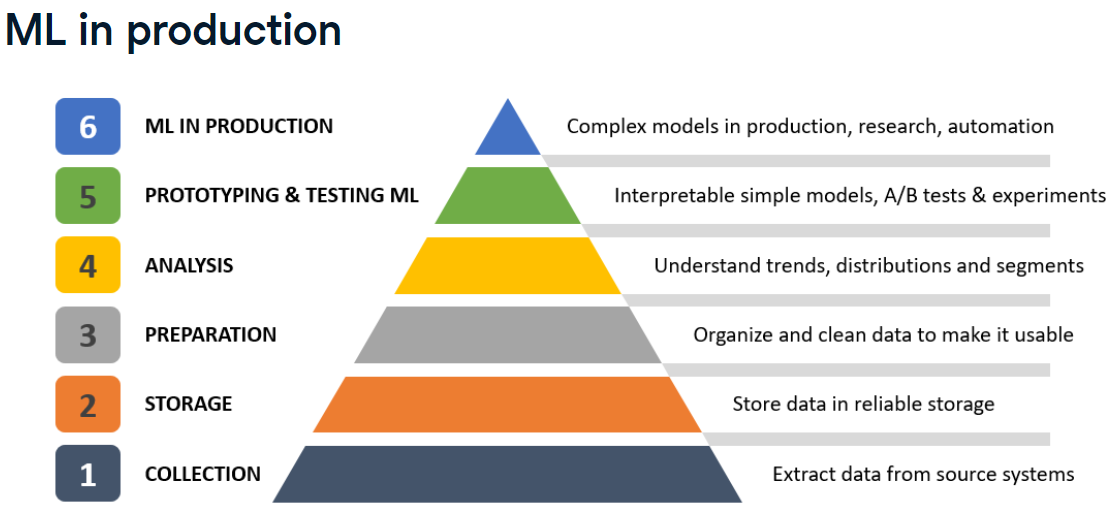
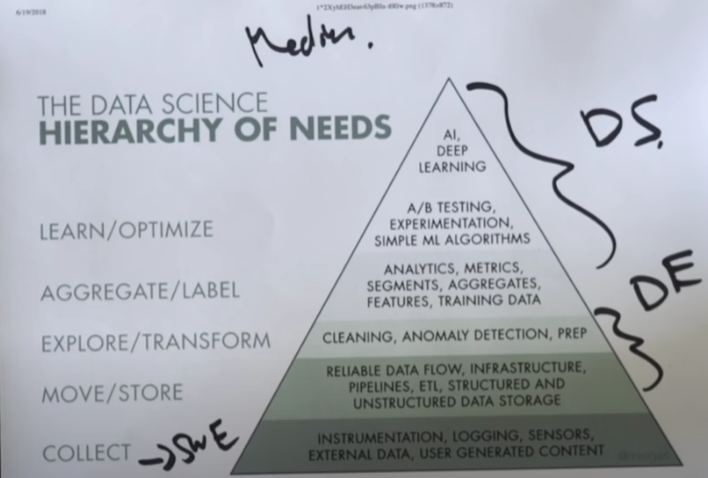
**Datascience Overview**

Here’s a general overview of the ML Pipeline,



Data scientists can be involved at all levels of production,



In small company, it’s possible the data scientists to do a little bit of everything. In larger companies, it’s more segregated. There would be a groups dedicated to the three separate tiers. The datascientists would do the top tier of the pyramid, handling the data analytics and deep learning, etc. The data engineers would do the cleaning and data pipelines. And the software engineers would create the architecture to handle running data analytics software, and machine learning stuff.

**Collection**

There’s lots of data out there. Some data sources are:

* movies watched on Netflix
* songs listened to on Spotify
* ads interacted (commented on?) in Facebook
* responses on Twitter, Instagram, TikTok?
* purchases made on Amazon
* searches made on Google
* people interacted with on dating apps
* people called, or which area codes, etc.
* credit card payments
* health outcomes in meta-studies
* images digitized for computer processing (for self-driving cars, say, or maybe paintings or medical imaging scans)
* sounds digitized for computer processing

This information will typically be stored in databases. And we’d use SQL, MongoDB, or Java to access it, and sift through it.

**Preparation**

Once we have the data, we need to prepare it. You’ll want to find missing data, merge identical data, maybe look for outliers that should be discarded, etc. This is called cleaning the data. Seems there are a few tools for this. Pandas is the major one.

**Analysis**

There are two basic types of machine learning: Supervised and Unsupervised. But from somewhere else, I added Reinforcement and Deep learning. Whatever.

**Supervised ML**. Here we have feature variables, X, and try to predict a target variable y. We can use this to predict who will cancel a certain business service. And then hopefully, determine why. If can figure this out, then we can change business model, perhaps, to prevent this. This illustrates that there are two main kind of supervised learning: *prediction* and *inference*. Prediction is only concerned with having accurate predictions, but not with ascertaining which factors in X are most important for driving those outcomes. Inference is precisely interested in figuring out what features of X are most important for driving these outcomes. Prediction is probably best with neural network models, while inference might benefit from something like decision trees, or regression. Some examples,

*Prediction*

predicting fraudulent credit card transactions.

assessing likelihood of making a purchase given exposure to certain ads or whatever.

predict which machine parts are likely to break based off of sensor reading.

predict weekly demand for items in inventory.

classify emails as spam or not.

predict shipping delays.

finding best shipping routes.

classify sounds as a particular word (speech recognition)

classify images as tumerous (medical imaging), or images as people (self driving cars)

classify emails as spam, credit card purchases as fraudulent

classify images as people, and maybe rate their features

Recommend movies to watch on Netflix, or purchases to make on Amazon.

*Inference*

What are the main drivers of fraud?

Which are the main drivers of heart attack?

What are the main drivers of customer turn over (churn)?

Furthermore, these categories can be each split by target variable. If it’s discreet, then this is roughly a *Classification* problem, and if continuous, a *Regression* problem. Apropos classification problems, one must have a clear idea of what the metric should be: accuracy? precision? recall? Are all mistakes equally bad, or are false positives worse than false negatives, etc. For example, its probably better to make false positive fraud transaction predictions than false negative ones, due to the costs of losing customers, or financial penalties vs. the cost of paying someone overtime to manually check the relevant financial data. Or something. So there might be a balance between precision and recall that we need to strike.

**Unsupervised ML**. Here we just have features, X, but no target variable y. The goal is to see if X can be segregated into meaningful groups. Might try to identify groups of customers with similar spending habits, and then perhaps business can tailor their advertising or product line to these groups. Would help also to determine the highest and lowest spending groups. For example, we

aggregate credit card transaction data and look for outlier purchases.

aggregate sensor readings from machine parts and see if we can identify outliers. Anomaly detection in general.

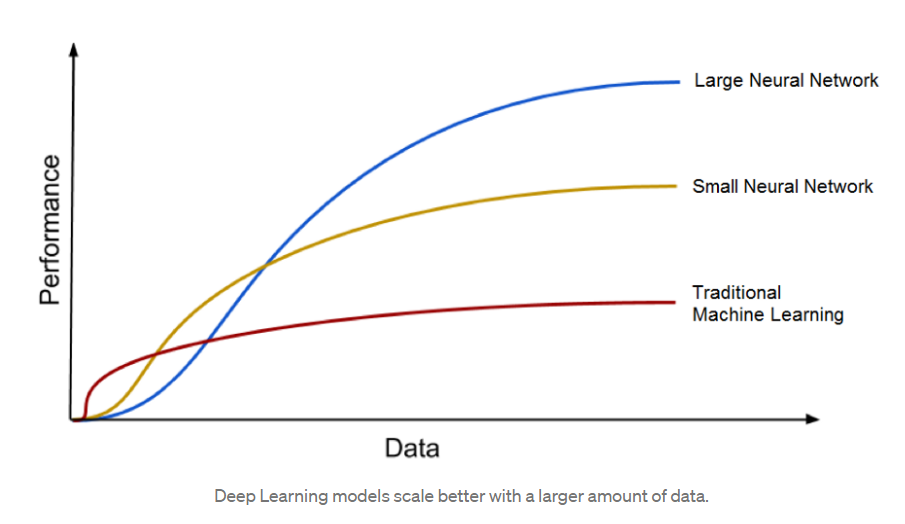
classify colors in picture into major groups in case have to display with fewer colors.

sorting purchases into classes, and then can use that scheme to recommend purchases to specific person. An example of this is movie recommendation engines

**Reinforcement Learning**

**Deep Learning**

Deep learning is a kind of Machine Learning, and is apparently much more efficacious than the usual Machine Learning. It’s used in self-driving cars, Netflix movie recommender algorithms, Alexa, etc. Apropos Classification and Clustering, basically it uses Neural Networks to construct both the manifold (feature extraction) and the metric, rather than rely on a person to do it. You would still train the AI on say half the data and test on the other half. Apparently, these networks get better and better at their jobs, the more data they have (obviously) and apparently their metrics eventually surpass the human created ones (typically? Always?)



When embarking on a ML project, should always ask if it makes financial sense in the first place. Are the possible benefits of creating the model greater than the risks its attempting to mitigate?

**Prototyping and Testing**

A/B testing is when you take whatever actionable insights you’ve found in your model and test them on a control group. For instance, if your model (inference model I guess) shows that thos with more ad exposure are more likely to buy, then you’d take a group of customers and split them into a *more-ads* and *normal-ads* group. And then you’d see if the more-ads group actually bought more stuff. This is where statistics can come into play. You’d want to know that the results of the experiment can be reliably generalized to the population.

