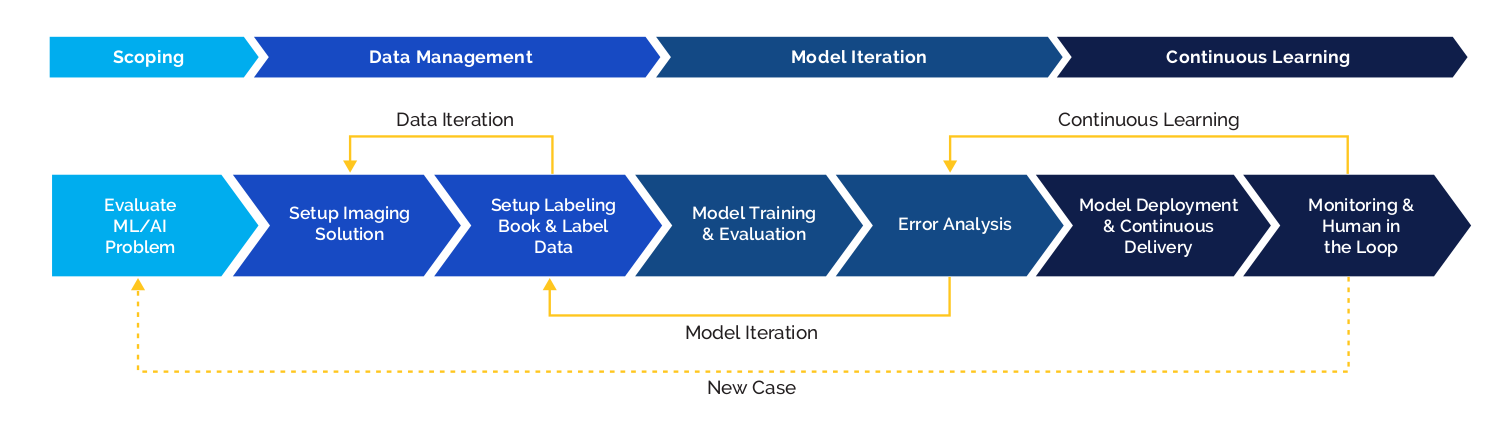
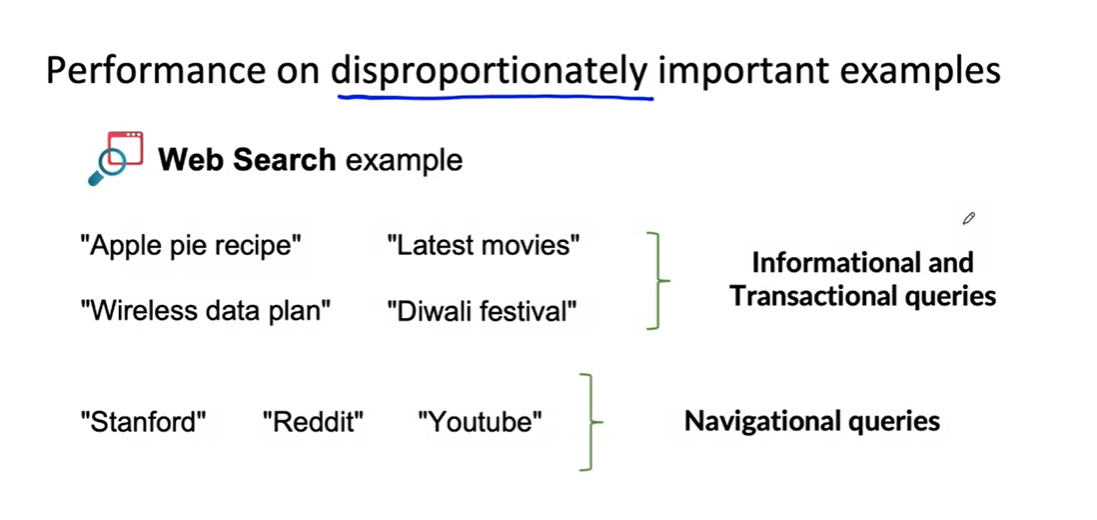
**Why low average error isn't good enough. #ML #DataScience**

****

A machine learning system may have low average test set error, but if its performance on a set of disproportionately important examples isn't good enough, then the machine learning system will still not be acceptable for production deployment. Let me use a search engine as an example.



There are a lot of web search queries like these: Apple pie recipe, latest movies, wireless data plan etc. **These types of queries are sometimes called informational or transactional queries**, where I want to learn about apple pies or maybe I want to buy a new wireless data plan and you might be willing to forgive a web search engine that doesn't give you the best apple pie recipe because there are a lot of good apple pie recipes on the Internet.

This happens a lot

For informational and transactional queries, a web search engine wants to return the most relevant results, but users are willing to forgive maybe ranking the best result, Number two or Number three. There's a different type of web search query such as Facebook, Corona , or Reddit, or YouTube. These are called navigational queries, where the user has a very clear intent, very clear desire to go to Stanford.edu, or Reddit.com, or YouTube.com. When a user has a very clear navigational intent, they will tend to be very unforgiving if a web search engine does anything other than return YouTube.com as the Number one ranked results and the search engine that doesn't give the right results will quickly lose the trust of its users. **Navigational queries** in this context are a disproportionately important set of examples and if your learning algorithm improves your average test set accuracy for web search but makes a few navigational questions incorrect, it might not be suitable for deployment. The challenge, of course, is that average test set accuracy tends to weight all examples equally, whereas, in web search, some queries are disproportionately important. Now one thing you could do is try to give these examples a higher weight. That could work for some applications, but , just changing the weights of different examples doesn't always solve the entire problem.