When to Use Machine Learning

So, we've just talked about a bunch of different ways that newsrooms have used machine learning, and hopefully that's been somewhat inspiring. But let me try to give you some guidelines as to when you should consider maybe using or not using machine learning for a project.

So the first question you should definitely ask yourself is: Whatever you're trying to build a model for, can you just do it manually? So, for example, let's say you're looking through crime. You have 300 reports of different crimes that you have to categorize. It will take you much longer to build a model that can categorize those examples and probably take more data than for just you, or you and your interns, to categorize them manually. If it's conceivable that you could just do it in one shot, always prefer doing your task manually.

Next question is: Do you have the data set you need to do what you want? So there are lots of general tasks that you can use a preexisting model for. We'll talk about this a little bit later in the tool section. But if you have just a photo and you kind of want to generally know what's in the photo – Is it a statue? Is it a chair? Or is it a dog? – You don't need a huge training data set. You can probably use an existing model to do that.

But if you want to do something very specific, like you want to identify spy planes, you're going to need hundreds or thousands of examples of flights and their labels. So, you either need to have that data, know where to find that data, or be willing to invest in creating that sort of data set.

Next question: Do you require explainability? What is explainability? A lot of people, I think, understand that when a machine learning model, especially a neural network, makes a prediction, we don't always understand how it came to that conclusion. So, for example, if a crime is classified as being a violent crime, what words do the model look at to come to that conclusion? You know, it's not always clear to even the creator.

Sometimes it doesn't matter if you understand why a model came to a decision. Because like, for example, in the L.A. Times case, when we were trying to see whether or not a crime was misclassified as being violent, the reporters and the editors could go through and say, "OK, this model made this prediction, but let me just verify that it was correct." So, you know, the model is just is helping you in a way, but the editor is always there verifying the results.

But you have to consider whether when you're reporting your story, what the model did, if that matters, because it might be that you can't really get a satisfying explanation of why the model made a certain decision.

And finally, and most importantly: Can you tolerate error? All machine learning models pretty much make errors. It's just inevitable. And the correct way to use machine learning is to account for those errors, to understand when they're really bad, to understand when they're acceptable, and to understand how to handle them.

So what I've sort of been emphasizing is you really want to use machine learning with regular reporting complementarity. And I think of it as sort of using machine learning to find needles in haystacks or to find leads in tons of tons of data, where a model says you might

want to investigate this plane or you might want to investigate whether this crime was violent or not. So you work with the model, but you know, that it will make errors. So just to account for that in your reporting process.

I hope that is a good summary of when you should and shouldn't use machine learning.