# Problem Formulation: What this pipeline phase entails and why it’s important

The problem formulation phase of the ML Pipeline is critical, and it’s where everything begins. Typically, this phase is kicked off with a question of some kind. Examples of these kinds of questions include: Could cars really drive themselves? What additional product should we offer someone as they checkout? How much storage will clients need from a data center at a given time?

The problem formulation phase starts by seeing a problem and thinking “what question, if I could answer it, would provide the most value to my business?” If I knew the next product a customer was going to buy, is that most valuable? If I knew what was going to be popular over the holidays, is that most valuable? If I better understood who my customers are, is that most valuable?

However, some problems are not so obvious. When sales drop, new competitors emerge, or there’s a big change to a company/team/org, it can be easy to say, “I see the problem!” But sometimes the problem isn’t so clear. Consider self-driving cars. How many people think to themselves, “driving cars is a huge problem”? Probably not many. In fact, there isn’t a problem in the traditional sense of the word but there is an opportunity. Creating self-driving cars is a huge opportunity. That doesn’t mean there isn’t a problem or challenge connected to that opportunity. How do you design a self-driving system? What data would you look at to inform the decisions you make? Will people purchase self-driving cars?

Part of the problem formulation phase includes seeing where there are opportunities to use machine learning.

In the following practice examples, you are presented with four different business scenarios. For each scenario, consider the following questions:

1. Is machine learning appropriate for this problem, and why or why not?
2. What is the ML problem if there is one, and what would a success metric look like?
3. What kind of ML problem is this?
4. Is the data appropriate?

**The solutions given in this document are one of the many ways you can formulate a business problem.** The first scenario has been completed for you. Remember that there are two ways to start an ML problem. The first is by addressing an obvious problem, the second is by seeing an opportunity. Lastly, be sure to consider whether this is even an ML problem at all. Take a look at scenarios 2 – 4 below and see if you can answer the questions above.

1) Amazon recently began advertising to its customers when they visit the company website. The Director in charge of the initiative wants the advertisements to be as tailored to the customer as possible. You will have access to all the data from the retail webpage, as well as all the customer data.

* 1. ML is appropriate because of the scale, variety and speed required. There are potentially thousands of ads and millions of customers that need to be served customized ads immediately as they arrive to the site.
	2. The problem is ads that are not useful to customers are a wasted opportunity and a nuisance to customers, yet not serving ads at all is a wasted opportunity. So how does Amazon serve the most relevant advertisements to its retail customers?
		1. Success would be the purchase of a product that was advertised.
	3. This is a supervised learning problem because we have a labeled data point, our success metric, which is the purchase of a product.
	4. This data is appropriate because it is both the retail webpage data as well as the customer data.
1. You’re a Senior Business Analyst at a social media company that focuses on streaming. Streamers use a combination of hashtags and predefined categories to be discoverable by your platform’s consumers. You ran an analysis on unique streamer counts by hashtags and categories over the last month and found that out of tens of thousands of streamers, almost all use only 40 hashtags and 10 categories despite innumerable hashtags and hundreds of categories. You presume the predefined categories don’t represent all the possibilities very well, and that streamers are simply picking the closest fit. You figure there are likely many categories and groupings of streamers that are not accounted for. So you collect a dataset that consists of all streamer profile descriptions (all text), all the historical chat information for each streamer, and all their videos that have been streamed.
	1. ML is appropriate because of the scale and variability.
	2. The problem is the content of streamers is not being represented by the existing categories. Success would be naturally grouping the streamers into categories based on content and seeing if those align with the hashtags and categories that are being commonly used. If they do not, then the streamers are not being well represented and you can use these groupings to create new categories.
	3. There isn’t a specific outcome variable. There’s no target or label. So this is an unsupervised problem.
	4. The data is appropriate.
2. You’re a headphone manufacturer who sells directly to big and small electronic stores. As an attempt to increase competitive pricing, Store 1 and Store 2 decided to put together the pricing details for all headphone manufacturers and their products (about 350 products) and conduct daily releases of the data. You will have all the specs from each manufacturer and their product’s pricing. Your sales have recently been dropping so your first concern is whether there are competing products that are priced lower than your flagship product.
	1. ML is probably not necessary for this. You can just search the dataset to see which headphones are priced lower than the flagship, then compare their features and build quality.
3. You’re a Senior Product Manager at a leading ridesharing company. You did some market research, collected customer feedback, and discovered that both customers and drivers are not happy with an app feature. This feature allows customers to place a pin exactly where they want to be picked up. The customers say drivers rarely stop at the pin location. Drivers say customers most often put the pin in a place they can’t stop. Your company has a relationship with the most used maps app for the driver’s navigation so you leverage this existing relationship to get direct, backend access to their data. This includes latitude and longitude, visual photos of each lat/long, traffic delay details, and regulation data if available (ie- No Parking zones, 3 minute parking zones, fire hydrants, etc.).
	1. ML is appropriate because of the scale and automation involved. It’s not feasible to drive everywhere and write down all the places that are ok for pickup. However, maybe we can predict whether a location is ok for pickup.
	2. The problem is drivers and customers are having poor experiences connecting for pickup, which is pushing customers away from the platform.
		1. Success would be properly identifying appropriate pickup locations so they can be integrated into the feature.
	3. This is a supervised learning problem even though there aren’t any labels, yet. Someone will have to go through a sample of the data to label where there are ok places to park and not park, giving the algorithms some target information.
	4. The data is appropriate once a sample of the dataset has been labeled. There may be some other data that could be included too. What about asking UPS for driver stop information? Where do they stop?