**Preview**

In this article I describe my perception of the Machine Learning (ML) industry from a Business Strategy perspective. Since the popularization of ML, companies have strived to develop end-to-end solutions, but more recently companies have focused on integrating horizontally across steps in the Data Supply Chain.

**Article**

**Structure**

I’m going to describe some Machine Learning concepts and build the Data Supply Chain. Then I’ll explain some Business Strategy concepts and propose a competitive strategy.

**Supervised Learning**

For Supervised Machine Learning algorithms to function effectively, they need to be trained on labelled data. This data is typically labelled by humans which can be a mundane task. An example of labelling data could be drawing bounding boxes on objects in images and labelling those boxes with a category of object. The bounding box coordinates could then be stored in a file like a JSON with the type of object included. Assuming the goal of this model is to identify objects, then it will be trained on (likely) thousands of images with the accompanying bounding box coordinates and object labels. If you’re interested in the logistics behind Object Detection you can read more [here](https://www.linkedin.com/feed/update/urn:li:activity:6735314387894091776/).

With the same example in mind, after the model is trained with a Machine Learning Framework, it needs to be deployed to production (ex: embedded on a site). The deployment step typically requires a lot more code than the model itself. Andrew Ng recently proposed that there is large gap from Proof of Concept to Production models that are ready to use. Ng also hypothesized that most of the aggregate time spent working on ML across all scientists is spent tuning algorithms, but productivity would improve if scientists spent more time iterating on their datasets.

The bottom line is that a disproportionate amount of time is spent developing ML algorithms, as opposed to building good datasets or deploying models so they can be used practically.

**The Data Supply Chain**

Center this and pad it:

Idea/Hypothesis 🡪 Data Collection 🡪 Model 🡪 Production

This is a high-level look at the steps required for a Machine Learning model to create value.

1. Idea/Hypothesis

For example: a team observes that there is a relationship between some variables that a computer could learn better than a human. They decide they want to make a model that makes some human tasks more precise.

1. Data Collection

Supervised models need to be trained on data that is labelled with ground truths. Collecting this data can be difficult considering it may be hard to find. If we’re dealing with object detection, then data would need to be collected and annotated.

1. Model

Given a robust data set, it needs to be split into training and testing data sets. Data can then be fed through the model/algorithm which learns relationships in the data and trains itself to predict things from data it hasn’t seen before.

1. Production

When the model is trained and tested, then it can be put into a context where it’s used based on the original hypothesis or idea. For example, deployed on a website, or implemented in a piece of hardware. People in the industry may also refer to this step as ML Ops.

**Porter’s 5 Forces**

1. Industry Rivalry
2. Bargaining Power of suppliers
3. Bargaining Power of buyers
4. Threat of entry
5. Threat of Substitutes

It may seem abstract at this point, but bear with me. Porter’s 5 forces paint the competitive landscape for an industry and can help new entrants assess the attractiveness of an industry. In my strategy class we learned about this concept by segmenting the supply chain of an industry and analyzing each of the 5 forces in that industry.

Picture it like this:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dairy Farms | Ice Cream Manufacturers | Ice Cream Distributors | Grocery Store |
| Threat of Entry |  |  |  |  |
| Threat of Substitutes |  |  |  |  |
| Bargaining Power of Suppliers | **Vertical** | **Horizontal** |  |  |
| Bargaining Power of Buyers |  |  |  |  |
| Rivalry |  |  |  |  |

Each of the sections across the top represents a step in the Ice Cream supply chain from dairy farms to grocery stores. Some companies might try to establish themselves vertically by purchasing their own farms, manufacturing their own Ice Cream, and selling it directly to end customers. This could significantly increase their margins. On the other hand, a company could try to establish themselves horizontally within the scope of one of the steps in the process. For example, they could introduce a new way to manufacture Ice Cream that makes it taste a lot better and instead of expanding to manufacture their own dairy or sell directly, they could acquire other ice cream manufacturers and expand their manufacturing capabilities. This constitutes horizontal integration, and it can be powerful because it creates bargaining power with suppliers/buyers and effectively insulates the industry from new competition (3 of Porter’s 5 forces). The Ice Cream manufacturer can now buy more dairy in bulk or pressure farms to lower prices. They can also scale their product and differentiate from competition or new entrants.

**Opportunities in Machine Learning**

So, what does Porter’s 5 forces say about the Data Supply Chain? Most companies that want to provide ML capabilities focus on an end-to-end solution where their client has minimal interaction. These companies would be considered vertically integrated across the Data Supply Chain.

But has any company tried to integrate across one step of the Supply Chain horizontally? I think that Scale AI is a good example of horizontal integration across the Data Collection step. They created a two-way platform for contractors and companies who need high quality labelled data. As of 2019 they had over 30,000 contractors labelling data. This is also great example of [Low-end Platform Disruption](https://hbr.org/2016/04/why-platform-disruption-is-so-much-bigger-than-product-disruption). Scale AI is reducing the cost of labelling high quality data (the incumbent method being an Engineer labelling the data). This approach has been successful for them; they are sourcing high quality data for some of the largest companies in the world.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Idea/Hypothesis | Data Collection | Model | Production |
| Threat of Entry |  |  |  |  |
| Threat of Substitutes |  |  |  |  |
| Bargaining Power of Suppliers | **Common ML Solutions** | **Scale AI** |  |  |
| Bargaining Power of Buyers |  |  |  |  |
| Rivalry |  |  |  |  |

I think the next big opportunity is to horizontally integrate across the Production step in my Data Supply Chain (ex: a tool that makes ready to use). It’s a daunting task for sure, but anyone who can make a robust ML Ops platform for many types of models and deployment, will be successful IMO.