Overview of Machine Learning Systems

MLOps is a set of tools and best practices for bringing ML into production.

“ML algorithms” is usually what people think of when they say machine learning, but it’s only a small part of the entire system.

A diagram of a computer

Description automatically generated

Machine learning is an approach to   
(1) learn   
(2) complex patterns from   
(3) existing data and use these patterns to make   
(4) predictions on   
(5) unseen data

*------------------------------- (8/1/2024) Continue*

1. Learn: the system has the capacity to learn

if you want to build an ML system to learn to predict the rental price for Airbnb listings, you need to provide a dataset where each input is a listing with relevant characteristics (square footage, number of rooms, neighborhood, amenities, rating of that listing, etc.) and the associated output is the rental price of that listing. Once learned, this ML system should be able to predict the price of a new listing given its characteristics

*------------------------------- (8/2/2024) (Pg 3)Continue*

1. Complex patterns: there are patterns to learn, and they are complex ML solutions are only useful when there are patterns to learn. Sane people don’t invest money into building an ML system to predict the next outcome of a fair die because there’s no pattern in how these outcomes are generated.4 However, there are patterns in how stocks are priced, and therefore companies have invested billions of dollars in building ML systems to learn those patterns

*------------------------------- (8/3/2024) (Pg 4)Continue*

there might be a pattern in how Elon Musk’s tweets affect cryptocurrency prices. How‐ ever, you wouldn’t know until you’ve rigorously trained and evaluated your ML models on his tweets. Even if all your models fail to make reasonable predictions of cryptocurrency prices, it doesn’t mean there’s no pattern.

*------------------------------- (8/4/2024) (Pg 4)Continue*

A diagram of software

Description automatically generated

*------------------------------- (8/5/2024) (Pg 5)Continue*

1. Existing data: data is available, or it’s possible to collect data Because ML learns from data, there must be data for it to learn from. It’s amusing to think about building a model to predict how much tax a person should pay a year, but it’s not possible unless you have access to tax and income data of a large population. In the zero-shot learning (sometimes known as zero-data learning) context, it’s possible for an ML system to make good predictions for a task without having been trained on data for that task. However, this ML system was previously trained on data for other tasks, often related to the task in consideration

*------------------------------- (8/6/2024) (Pg 5)Continue*

1. Predictions: it’s a predictive problem ML models make predictions, so they can only solve problems that require predictive answers. ML can be especially appealing when you can benefit from a large quantity of cheap but approximate predictions. In English, “predict” means “estimate a value in the future.” For example, what will the weather be like tomorrow? Who will win the Super Bowl this year? What movie will a user want to watch next?

*------------------------------- (8/7/2024) (Pg 6)Continue*

1. Unseen data: unseen data shares patterns with the training data The patterns your model learns from existing data are only useful if unseen data also share these patterns. A model to predict whether an app will get downloaded on Christmas 2020 won’t perform very well if it’s trained on data from 2008, when the most popular app on the App Store was Koi Pond. What’s Koi Pond? Exactly

*------------------------------- (8/8/2024) (Pg 6)Continue*

6. It’s repetitive

Humans are great at few-shot learning: you can show kids a few pictures of cats and most of them will recognize a cat the next time they see one. Despite exciting progress in few-shot learning research, most ML algorithms still require many examples to learn a pattern. When a task is repetitive, each pattern is repeated multiple times, which makes it easier for machines to learn it.

*--------------------- 8/9/2024 up to page 06*

1. The cost of wrong predictions is cheap Unless your ML model’s performance is 100% all the time, which is highly unlikely for any meaningful tasks, your model is going to make mistakes. ML is especially suitable when the cost of a wrong prediction is low. For example, one of the biggest use cases of ML today is in recommender systems because with recommender systems, a bad recommendation is usually forgiving—the user just won’t click on the recommendation.

*--------------------- 8/10/2024 up to page 06*

1. It’s at scale:

Having a problem at scale also means that there’s a lot of data for you to collect, which is useful for training ML models

1. The patterns are constantly changing:

Today an indication of a spam email is a Nigerian prince, but tomorrow it might be a distraught Vietnamese writer.

Machine Learning Use Cases

Recommendation system, assisting people in many of their daily activity.

Enterprise applications might have stricter accuracy requirements but be more forgiving with latency requirements.

improving a speech recognition system’s accuracy from 95% to 95.5% might not be noticeable to most consumers, but improving a resource allocation system’s efficiency by just 0.1% can help a corporation like Google or General Motors save millions of dollars.

*--------------------- 8/11/2024 up to page 10*

Deciding how much to charge for your product or service is probably one of the hardest business decisions; why not let ML do it for you? Price optimization is the process of estimating a price at a certain time period to maximize a defined objective function, such as the company’s margin, revenue, or growth rate.

for example, internet ads, flight tickets, accommodation bookings, ride-sharing, and events.

*--------------------- 8/12/2024 up to page 11*

The cost of acquiring a new user is approximated to be 5 to 25 times more expensive than retaining an existing one.12 Churn prediction is predicting when a specific customer is about to stop using your products or services so that you can take appropriate actions to win them back.

*--------------------- 8/13/2024 up to page 11*

Machine Learning in Research Versus in Production

As ML usage in the industry is still fairly new, most people with ML expertise have gained it through academia: taking courses, doing research, reading academic papers. If that describes your background, it might be a steep learning curve for you to understand the challenges of deploying ML systems in the wild and navigate an overwhelming set of solutions to these challenges

*--------------------- 8/14/2024 up to page 12*

A black and white text on a black background

Description automatically generated

*--------------------- 8/15/2024 up to page 13*

ML engineers   
Want a model that recommends restaurants that users will most likely order from, and they believe they can do so by using a more complex model with more data.

Sales team   
Wants a model that recommends the more expensive restaurants since these restaurants bring in more service fees.

Product team   
Notices that every increase in latency leads to a drop in orders through the service, so they want a model that can return the recommended restaurants in less than 100 milliseconds.

ML platform team   
As the traffic grows, this team has been woken up in the middle of the night because of problems with scaling their existing system, so they want to hold off on model updates to prioritize improving the ML platform.

Manager   
Wants to maximize the margin, and one way to achieve this might be to let go of the ML team

*--------------------- 8/16/2024 up to page 14*

Production having different requirements from research is one of the reasons why successful research projects might not always be used in production. For example, ensembling is a technique popular among the winners of many ML competitions, including the famed $1 million Netflix Prize, and yet it’s not widely used in produc‐ tion. Ensembling combines “multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.”15 While it can give your ML system a small performance improvement, ensembling tends to make a system too complex to be useful in production,

*--------------------- 8/17/2024 up to page 14*

The misalignment of interests between research and production has been noticed by researchers. In an EMNLP 2020 paper, Ethayarajh and Jurafsky argued that benchmarks have helped drive advances in natural language processing (NLP) by incentivizing the creation of more accurate models at the expense of other qualities valued by practitioners such as compactness, fairness, and energy efficiency

*--------------------- 8/18/2024 up to page 15*

Computational priorities

When designing an ML system, people who haven’t deployed an ML system often make the mistake of focusing too much on the model development part and not enough on the model deployment and maintenance part.

*--------------------- 8/19/2024 up to page 15*

The average latency of Google Translate is the average time it takes from when a user clicks Translate to when the translation is shown, and the throughput is how many queries it processes and serves a second. If your system always processes one query at a time, higher latency means lower throughput. If the average latency is 10 ms, which means it takes 10 ms to process a query, the throughput is 100 queries/second. If the average latency is 100 ms, the throughput is 10 queries/second. However, because most modern distributed systems batch queries to process them together, often concurrently, higher latency might also mean higher throughput.

*--------------------- 8/20/2024 up to page 16*

A diagram of a number of queries

Description automatically generated

*--------------------- 8/21/2024 up to page 17*

To reduce latency in production, you might have to reduce the number of queries you can process on the same hardware at a time. If your hardware is capable of processing many more queries at a time, using it to process fewer queries means underutilizing your hardware, increasing the cost of processing each query

In 2017, an Akamai study found that a 100 ms delay can hurt conversion rates by 7%

*--------------------- 8/22/2024 up to page 17*

Use percentile to avoid the outliers and find the average of the latency. Like 50th percentile.

*--------------------- 8/23/2024 up to page 18*

Data During the research phase, the datasets you work with are often clean and wellformatted, freeing you to focus on developing models. They are static by nature so that the community can use them to benchmark new architectures and techniques

*--------------------- 8/24/2024 up to page 18*

means that many people might have used and discussed the same datasets, and quirks of the dataset are known. You might even find open source scripts to process and feed the data directly into your models

*--------------------- 8/25/2024 up to page 18*

In research, you mostly work with historical data, e.g., data that already exists and is stored somewhere. In production, most likely you’ll also have to work with data that is being constantly generated by users, systems, and third-party data.

*--------------------- 8/26/2024 up to page 19*



*--------------------- 8/27/2024 up to page 19*

Fairness During the research phase, a model is not yet used on people, so it’s easy for research‐ ers to put off fairness as an afterthought: “Let’s try to get state of the art first and worry about fairness when we get to production.” When it gets to production, it’s too late. If you optimize your models for better accuracy or lower latency, you can show that your models beat state of the art. But, as of writing this book, there’s no equivalent state of the art for fairness metrics

*--------------------- 8/28/2024 up to page 19*

You or someone in your life might already be a victim of biased mathematical algorithms without knowing it. Your loan application might be rejected because the ML algorithm picks on your zip code, which embodies biases about one’s soci‐ oeconomic background. Your resume might be ranked lower because the ranking system employers use picks on the spelling of your name. Your mortgage might get a higher interest rate because it relies partially on credit scores, which favor the rich and punish the poor. Other examples of ML biases in the real world are in predictive policing algorithms, personality tests administered by potential employers, and college rankings.

*--------------------- 8/29/2024 up to page 19*

ML algorithms don’t predict the future, but encode the past, thus perpetuating the biases in the data and more. When ML algorithms are deployed at scale, they can discriminate against people at scale. If a human operator might only make sweeping judgments about a few individuals at a time, an ML algorithm can make sweeping judgments about millions in split seconds. This can especially hurt members of minority groups because misclassification on them could only have a minor effect on models’ overall performance metrics

*--------------------- 8/30/2024 up to page 20*

If an algorithm can already make correct predictions on 98% of the population, and improving the predictions on the other 2% would incur multiples of cost, some companies might, unfortunately, choose not to do it. During a McKinsey & Company research study in 2019, only 13% of the large companies surveyed said they are taking steps to mitigate risks to equity and fairness, such as algorithmic bias and discrimination

*--------------------- 8/31/2024 up to page 20*

Interpretability

“Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

While most of us are comfortable with using a microwave without understanding how it works, many don’t feel the same way about AI yet, especially if that AI makes important decisions about their lives.

First, interpretability is important for users, both business leaders and end users, to understand why a decision is made so that they can trust a model and detect potential biases

*--------------------- 9/1/2024 up to page 21*

Machine Learning Systems Versus Traditional Software

many challenges are unique to ML applications and require their own tools. In SWE, there’s an underlying assumption that code and data are separated. In fact, in SWE, we want to keep things as modular and separate as possible.

On the contrary, ML systems are part code, part data, and part artifacts created from the two. The trend in the last decade shows that applications developed with the most/best data win. Instead of focusing on improving ML algorithms, most companies will focus on improving their data

*--------------------- 9/2/2024 up to page 22*

In traditional SWE, you only need to focus on testing and versioning your code. With ML, we have to test and version our data too

*--------------------- 9/3/2024 up to page 22*

The size of ML models is another challenge. As of 2022, it’s common for ML models to have hundreds of millions, if not billions, of parameters, which requires gigabytes of random-access memory (RAM) to load them into memory. A few years from now, a billion parameters might seem quaint—like, “Can you believe the computer that sent men to the moon only had 32 MB of RAM?”

*--------------------- 9/4/2024 up to page 22*

As ML models get more complex, coupled with the lack of visibility into their work, it’s hard to figure out what went wrong or be alerted quickly enough when things go wrong.

*--------------------- 9/5/2024 up to page 23*

ML systems are complex, consisting of many different components. Data scientists and ML engineers working with ML systems in production will likely find that focusing only on the ML algorithms part is far from enough. It’s important to know about other aspects of the system, including the data stack, deployment, monitoring, maintenance, infrastructure, etc

*--------------------- 9/6/2024 up to page 23*

Introduction to Machine Learning Systems Design

ML systems design takes a system approach to MLOps, which means that we’ll consider an ML system holistically to ensure that all the components—the business requirements, the data stack, infrastructure, deployment, monitoring, etc.— and their stakeholders can work together to satisfy the specified objectives and requirements

*--------------------- 9/7/2024 up to page 25*

Before using ML algorithms to solve your problem, you first need to frame your problem into a task that ML can solve. The difficulty of your job can change significantly depending on how you frame your problem

*--------------------- 9/8/2024 up to page 25*

Business and ML Objectives

When working on an ML project, data scientists tend to care about the ML objectives: the metrics they can measure about the performance of their ML models such as accuracy, F1 score, inference latency, etc. They get excited about improving their model’s accuracy from 94% to 94.2% and might spend a ton of resources—data, compute, and engineering time—to achieve that.

*--------------------- 9/9/2024 up to page 26*

A pattern I see in many short-lived ML projects is that the data scientists become too focused on hacking ML metrics without paying attention to business metrics. Their managers, however, only care about business metrics and, after failing to see how an ML project can help push their business metrics, kill the projects prematurely (and possibly let go of the data science team involved).

*--------------------- 9/10/2024 up to page 26*

The ultimate goal of any project within a business is, therefore, to increase profits, either directly or indirectly: directly such as increasing sales (conversion rates) and cutting costs; indirectly such as higher customer satisfaction and increasing time spent on a website

*--------------------- 9/11/2024 up to page 26*

Example of Comparing ML metric to business metric

1. Increased Recommender system’s click through rate increase in purchase and revenue

For example, Netflix measures the performance of their recommender system using take-rate: the number of quality plays divided by the number of recommendations a user sees.4 The higher the take-rate, the better the recommender system. Netflix also put a recommender system’s take-rate in the context of their other business metrics like total streaming hours and subscription cancellation rate. They found that a higher take-rate also results in higher total streaming hours and lower subscription cancellation rates.

The same ML model can also solve their problems faster, which makes them spend less money on your services.

To gain a definite answer on the question of how ML metrics influence business metrics, experiments are often needed. Many companies do that with experiments like A/B testing and choose the model that leads to better business metrics, regardless of whether this model has better ML metrics.

*--------------------- 9/12/2024 up to page 27*

Yet, even rigorous experiments might not be sufficient to understand the relationship between an ML model’s outputs and business metrics

Actual threats will then go through another, different process aimed at stopping them. When this process fails to stop a threat, it might be impossible to figure out whether the ML component has anything to do with it.

Many companies like to say that they use ML in their systems because “being AIpowered” alone already helps them attract customers, regardless of whether the AI part actually does anything useful.

*--------------------- 9/13/2024 up to page 28*

Returns on investment in ML depend a lot on the maturity stage of adoption

*--------------------- 9/14/2024 up to page 28*

A graph with numbers and text

Description automatically generated with medium confidence

*--------------------- 9/15/2024 up to page 29*

Requirements for ML Systems We can’t say that we’ve successfully built an ML system without knowing what requirements the system has to satisfy. The specified requirements for an ML system vary from use case to use case. However, most systems should have these four charac‐ teristics: reliability, scalability, maintainability, and adaptability.

*--------------------- 9/16/2024 up to page 29*

Reliability

The system should continue to perform the correct function at the desired level of performance even in the face of adversity (hardware or software faults, and even human error).

*--------------------- 9/17/2024 up to page 29*

“Correctness” might be difficult to determine for ML systems. For example, your system might call the predict function—e.g., model.predict()—correctly, but the predictions are wrong. How do we know if a prediction is wrong if we don’t have ground truth labels to compare it with?

*--------------------- 9/18/2024 up to page 29*

ML systems can fail silently. End users don’t even know that the system has failed and might have kept on using it as if it were working. For example, if you use Google Translate to translate a sentence into a language you don’t know, it might be very hard for you to tell even if the translation is wrong

*--------------------- 9/19/2024 up to page 29*

Scalability There are multiple ways an ML system can grow. It can grow in complexity. Last year you used a logistic regression model that fit into an Amazon Web Services (AWS) free tier instance with 1 GB of RAM, but this year, you switched to a 100-millionparameter neural network that requires 16 GB of RAM to generate predictions.

*--------------------- 9/20/2024 up to page 30*

Your ML system can grow in traffic volume. When you started deploying an ML system, you only served 10,000 prediction requests daily. However, as your company’s user base grows, the number of prediction requests your ML system serves daily fluctuates between 1 million and 10 million

*--------------------- 9/21/2024 up to page 30*

An ML system might grow in ML model count. Initially, you might have only one model for one use case, such as detecting the trending hashtags on a social network site like Twitter. However, over time, you want to add more features to this use case, so you’ll add one more to filter out NSFW (not safe for work) content and another model to filter out tweets generated by bots

*--------------------- 9/22/2024 up to page 30*

This growth pattern is especially common in ML systems that target enterprise use cases. Initially, a startup might serve only one enterprise customer, which means this startup only has one model. However, as this startup gains more customers, they might have one model for each customer. A startup I worked with had 8,000 models in production for their 8,000 enterprise customers

*--------------------- 9/23/2024 up to page 30*

For example, at peak, your system might require 100 GPUs (graphics processing units). However, most of the time, it needs only 10 GPUs. Keeping 100 GPUs up all the time can be costly, so your system should be able to scale down to 10 GPUs

*--------------------- 9/24/2024 up to page 30*

However, with one hundred models, both the monitoring and retraining aspect will need to be automated. You’ll need a way to manage the code generation so that you can adequately reproduce a model when you need t

*--------------------- 9/25/2024 up to page 31*