## Introduction to Machine Learning Systems Design

### Business and ML Objectives

most companies don’t care about the fancy ML metrics.

The ultimate goal of any project within a business is:

* directly such as increasing sales (conversion rates) and cutting costs
* indirectly such as higher customer satisfaction and increasing time spent on a website.

The effect of an ML project on business objectives can be hard to reason about. So 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.

Additionally, many companies pursue ML for its appeal to customers rather than its functional utility.

### Requirements for ML Systems

reliability, scalability, maintainability, and adaptability

#### Reliability

الصحة او المصداقية

يجب التأكد ان النظام وعملية التنبؤ تعمل بشكل صحيح ودقيق بعملية التنبؤ

With traditional software systems, you often get a warning, such as a system crash or runtime error or 404. However, ML systems can fail silently.

#### Scalability

ML systems can expand in various ways, each impacting system demands and requiring scalable solutions:

1. **Complexity**: A system may evolve from a simple model, like logistic regression, to a highly complex model.
2. **Traffic Volume**: as your company’s user base grows, the number of predictions requests your ML system serves daily fluctuates between 1 million and 10 million.
3. **Model Count**

Whichever way your system grows, there should be reasonable ways of dealing with that growth.

Important example: 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.

Autoscaling, a critical feature in cloud services, adjusts resources automatically. However, it is complex to manage

However, handling growth isn’t just resource scaling, but also artifact management.

#### Maintainability

It’s important to structure your workloads and set up your infrastructure in such a way that different contributors can work using tools that they are comfortable with, instead of one group of contributors forcing their tools onto other groups.

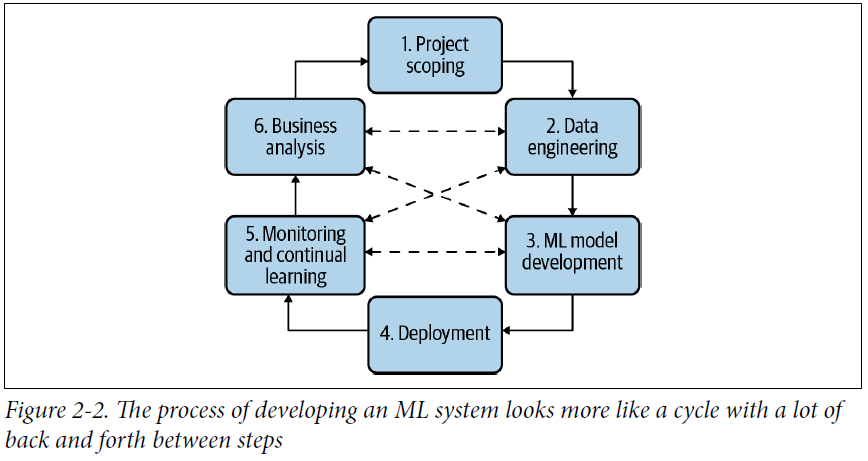
Code should be documented. Code, data, and artifacts should be versioned.

#### Adaptability

Because ML systems are part code, part data, and data can change quickly, ML systems need to be able to evolve quickly. This is tightly linked to maintainability.

### Iterative Process

the process looks more like a cycle with a lot of back and forth between different steps.



1. **Project Scoping**: Define goals, objectives, and constraints. Identify stakeholders and allocate resources.
2. **Data Engineering**: Process and curate raw data into training data, including sampling and labeling.
3. **ML Model Development**: Perform feature engineering, model selection, training, and evaluation with the curated data.
4. **Deployment**: Make the model accessible to users, acknowledging that continuous improvements may be necessary.
5. **Monitoring and Continual Learning**: Track model performance to adapt to changes in the environment and requirements.
6. **Business Analysis**: Assess model impact on business goals to refine projects and identify new opportunities.

### Framing ML Problems

we’ll focus on two aspects: the output of your model and the objective function that guides the learning process.

#### Types of ML Tasks

1. **Classification versus regression:** Classification models classify inputs into different categories and A regression model can easily be framed as a classification model by adding threshold to the output
2. **Binary versus multiclass classification:** binary classification, where there are only two possible classes and multiclass more than two labels (When the number of classes is large, hierarchical classification might be useful. In hierarchical classification, you have a classifier to first classify each example into one of the large groups. Then you have another classifier to classify this example into one of the subgroups.)
3. **Multiclass versus multilabel classification:** In multiclass classification, where an item belongs to only one category, you select the category with the highest probability, in multilabel classification, since an item can belong to multiple categories, you may choose the top two or three categories with the highest probabilities.

**Changing the way you frame your problem might make your problem significantly harder or easier.**

#### Objective Functions

An objective function is also called a loss function, choosing an objective function is usually straightforward, though not because objective functions are easy.

In Arabic:

عندما يكون لدينا اهداف كثيرة لاستخدام الML نكون امام حالتين: اما التعامل مع كل هدف على حدى لان الأهداف عند جمعها قد تكون متعاكسة في حالة جمعها كمثال بناء نموذج لكل هدف وهذا الامر يسهل علينا عملية الصيانة والتطوير في حالة فصل الأهداف، وقد يكون هذا افضل من التعامل مع الأهداف جميعها على اننا يمكن تحققيها باستخدام نموذج واحد كمثال ذكر في الكتاب ترتيب المنشورات حسب التفاعل (جودة المنشور والتفاعل)

### Mind Versus Data

Progress in the last decade shows that the success of an ML system depends largely on the data it was trained on. Instead of focusing on improving ML algorithms, most companies focus on managing and improving their data.

***Mind****:* might be disguised as inductive biases or intelligent architectural designs.

***Data:***might be grouped together with computation since more data tends to require more computation.

When asked how Google Search was doing so well, Peter Norvig, Google’s director of search quality, emphasized the importance of having a large amount of data over intelligent algorithms in their success: “We don’t have better algorithms. We just have more data.”

Christopher Manning suggest that systems with intelligent structures that learn from less data may be more efficient in certain cases.

Regardless of which camp will prove to be right eventually, no one can deny that data is essential, for now. Both the research and industry trends in the recent decades show the success of ML relies more and more on the quality and quantity of data.

### Summary

Before building an ML system, we need to understand the requirements that the system needs to meet to be considered a good system. The exact requirements vary from use case to use case. The success of systems including AlexNet, BERT, and GPT showed that the progress of ML in the last decade relies on having access to a large amount of data.24