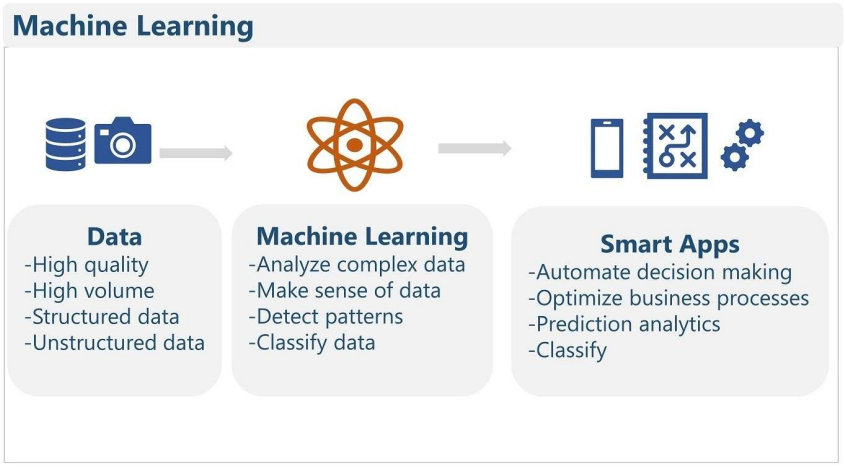
2 Machine Learning

Machine learning (ML) is the foundation of artificial intelligence and is gaining momentum in business applications. As ML gains momentum, professionals must explore ML, understand the concepts and terminology to be in a position to use it as a tool for business. This chapter will help you get familiar with foundation ML concepts, requirements and processes to provide you more than a cursory knowledge of ML.

Machine learning is making it possible for companies to transform collected & real time data into rich insight, enabling them to optimize existing business processes and create totally new services.

Lemonade insurance uses machine learning algorithms to automate and craft the best insurance policy for customers. It has replaced paper and brokers, with chatbots and algorithms. UPS analyzes data from historic truck routes to optimize delivery routes for their trucks saving them time and money. LinkedIn is practically an AI company, with a mission to match people and organizations to opportunity.

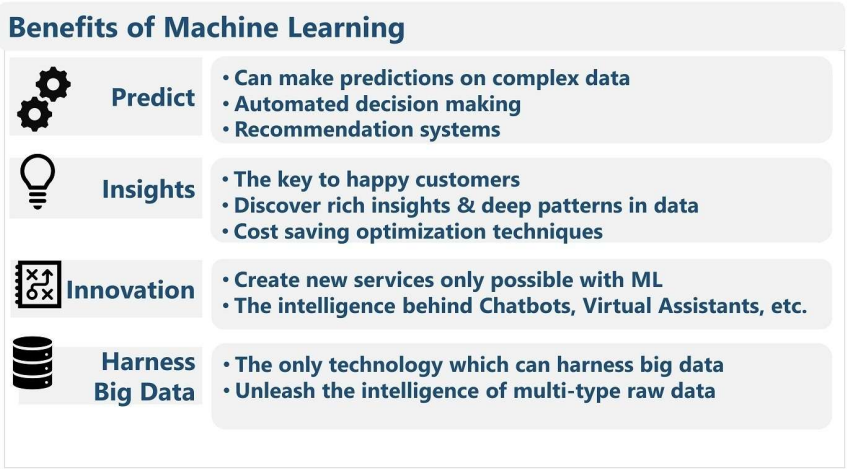


Machine learning makes it possible for algorithms to learn from data without programming and then make predictions, classify data in similar groups and extract hidden structures from data.

ML is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns within the data and make decisions with minimal human intervention.

ML has the potential to make apps, websites, business processes and machines smart. Data is the fuel to create these smart systems. The huge amount of data being generated by systems, IoT sensors, apps and social media is overwhelming the analytical capabilities of business. Businesses cannot rely on traditional tools and humans to analyze this information. The "after-the-fact" reporting approach to data is not working anymore. Businesses realize that by using ML, they can extract insights from large volumes of data and adopt a "forward thinking" predictive approach.

With the ubiquitous availability of cloud computing as a utility, data management has become commoditized. Now cloud providers have started offering ML as a service. Creating smart systems that use machine learning can be transformative for your business. The customer of tomorrow, armed with AI powered smartphones will expect and demand businesses to have a more predictive and consultative approach to service, not a "after the fact" approach.



Machine Learning Explained

In machine learning, we "teach" computers to make predictions, or inferences. *How is this done? I will explain ML in several ways, and then explain the important concepts related to ML.*

Machine learning is a subfield of Artificial Intelligence, and it enables computers to learn patterns in data through algorithms without the need for explicit programming.

In simple words:
Using **know data** as input, develop a **model** to **predict** unknown data. Learning a function f to map input X to output Y in Y=f(X).

Known data: The data you currently have
Base algorithm: Start with a base algorithm y = B0 + B1*x . Run the data through this algorithm to derive the best values of B0 & B1
Model: Is the specific algorithm created based on known data. So it could be y=4 + .5X
Predictions using unknow data: Use the model with new data to predict values of Y

Raw data is sometimes called as “dark data” because it does not offer a lot of value in its unprocessed state. By applying the right set of ML algorithms, we can extract powerful insights from this data. This is exactly what machine learning does.

- In traditional programing we create the algorithm using hand written business rules.
- In machine learning we are essentially generating the algorithm (called the model) based on known data.
- But why do we need to generate the model? Because the data is so complex that it cannot be programmed ahead of time.

Machine learning examines large volumes of data looking for patterns, then generates the model that enables you to find those patterns in new data. Let us step back and understand the different type of analytics and where ML fits in:

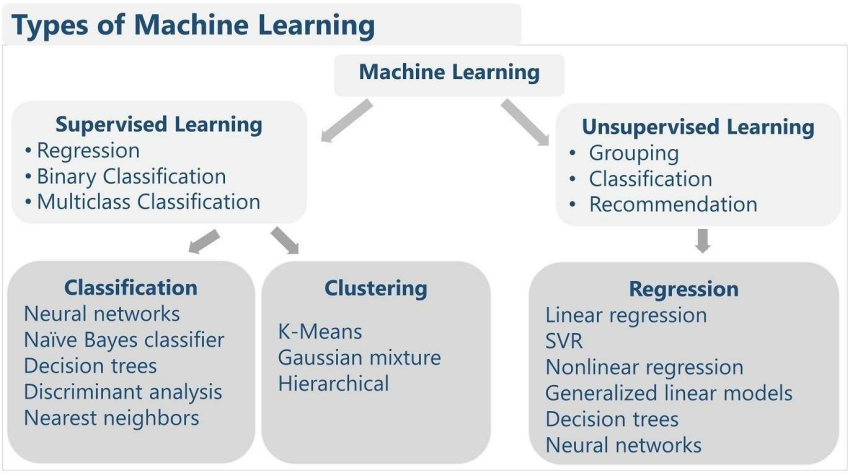
Description Analysis:
This is backward data analytics used by many systems today. This is employed heavily today and provides insights into "what has happened".

Predictive Analysis:
What **might** happen? Predictive analytics is probabilistic in nature and can forecast what might happen in the future. It is based on machine learning, analyzes historical data to create models which can be used to make predictions. For ex., you want to predict the median price of a home based on a number of variables such as number of rooms, area, number of bathrooms, area, zip code, etc.

Prescriptive Analysis:
Prescriptive analytics goes beyond predicting events by also advising one or more courses of action: and displays the likely outcome of each decision.

There are three categories of machine learning which we will discuss later in this chapter.

- Supervised machine learning
- Unsupervised machine learning
- Reinforcement machine learning



Let me explain machine learning (ML) using a very simple example.

Description	Card Present	Location	Date	Amount	Fraud
Target Store	Yes	Sugarland, TX	6/21/2018	$89	N
Dollar Store	Yes	Austin, TX	5/23/2018	$109	N
AMC Cinema	Yes	Sugarland, TX	5/23/2018	$56	N
Kikkinice2.com	No	Ukraine	5/24/2018	$799	Y
Kroger	Yes	Sugarland, TX	5/25/2018	$56	N
On2Auction.jp	No	Japan	4/28/2018	$590	Y

Examining 6 instances of credit card spending data of a person, you can eyeball the following patterns immediately:

- Maybe the person resides around Sugarland, TX. Because Target, Dollar Store and Kroger are grocery stores in Sugarland, TX. AMC Cinema is also located in Sugarland, TX. But we cannot say (or predict) for sure since **we do not have sufficient data**.
- This person may be vacationing in Sugarland, TX for 2-3 months. We would need more data, say for the past 2-3 years to predict his state of residence from credit card spending data.
- The 4th and 6th transaction look suspicious because the card was not present during the transaction, was from merchants of far off countries *(Japan & Ukraine)*, and had a high dollar amount.
- You can predict, with a fair amount of certainty (although maybe not 100%) that the Ukraine and Japan transactions may be "fraudulent" transactions.
- In the future if the bank received a new payment transaction from Kikkinice2.com, the bank can predict with high probability that it may be "fraudulent" transaction.

Now, just imagine if you had 20 columns & 6 million records. It would be very difficult for a human to examine that data and derive any patterns or meaning out of it. That's where machine learning comes in, and uses statistical techniques to find patterns in large amounts of data. This was just a simple example to explain the concept of ML and prediction using fraud detection use case.

To summarize:

(a) Machine learning uses statistical techniques and algorithms on large amounts of data to find patterns.
(b) It then generates "code" that can recognize those patterns. This code is called the "model".
(c) This generated code or "model" can be used by applications on new data, to make predictions or detect patterns.

Because ML can make predictions and can detect patterns in large volumes of data, it is opening new opportunities for businesses. Businesses can use ML to automate existing business processes and also creating new business models which could not be conceived before.

Generally speaking machine leaning algorithms fall in one of the below 3 categories. **It is important to understand this and become familiar with this vocabulary.**

Regression: means **predicting** the output value based on input variables. For ex. predicting the price of a house based on several input variables.

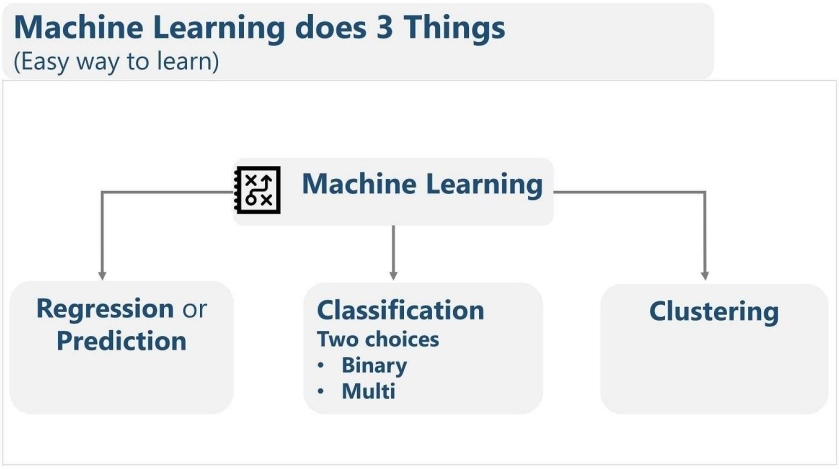
Classification: means we are trying to group data into a predefined class.

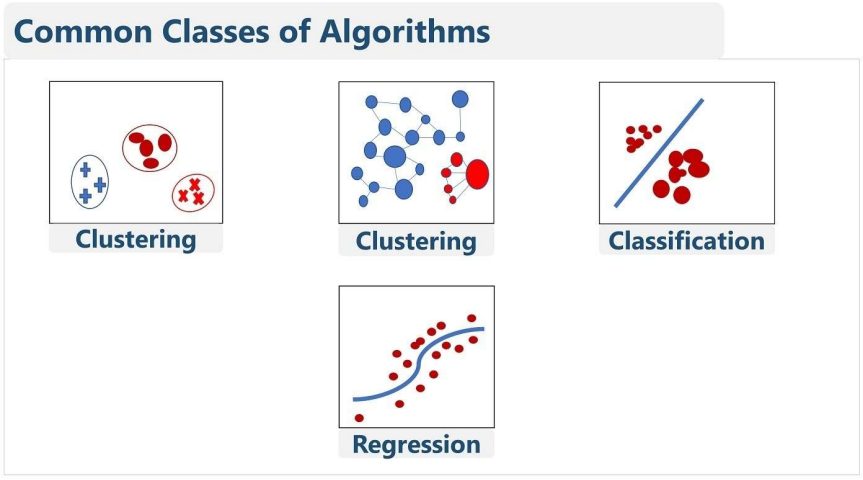
- Binary classification: predict values that can only have one of two values. For ex., classify customers who will "stay" or "leave" in the next financial year.
- Multiclass classification: when there are more than two labels, it is called multi-class classification. For example, based on the content of an email it could either be of type "primary"," promotion", "spam" or "social".

Clustering: The model groups data into categories by analyzing attributes. These logical groupings help filter out noise and derive insights from otherwise random data. Below are a few examples:

- Grouping of customers for marketing purposes
- Search results grouping
- Anomaly detection

In simple words, machine learning does three things **prediction**, **classification** and **clustering**. Machine learning algorithms are also classified in a similar manner. This is the mantra of ML, which you need to remember. That is the reason, I am trying to unpack this concept from different angles.





How do you do machine learning?

I will try to explain the machine learning process with a very specific audience in mind: business or technical professional, someone who is not a data scientist and is just getting started.

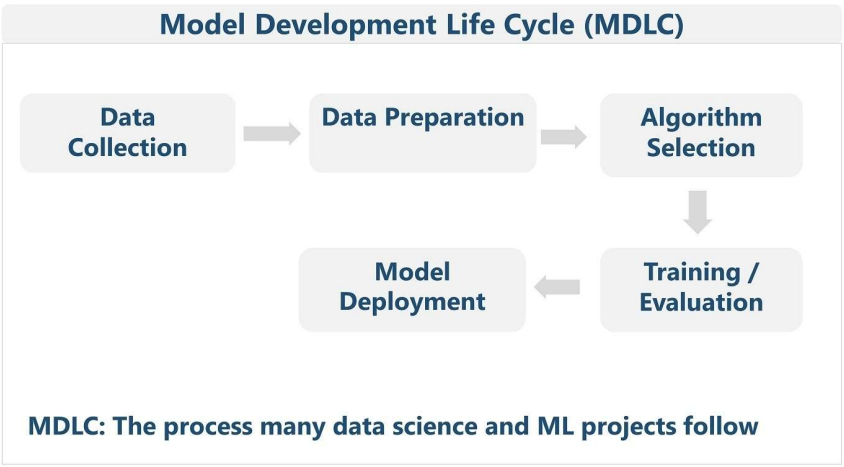
Objective + Use Case Selection:
Work with your stakeholders to identify the business problem you want to solve. Try to link your business problem to machine learning language**. Do you want to predict a value, classify data or group data?**

ML eligibility and guidelines:
Are your data sources & data attributes ML eligible? Identify data sources and determine if your data is available & suitable for machine learning. How do you decide that? Here are a few guidelines:

1. You have a lot of data and it is not possible for humans to derive intelligence out of it. This may be good candidate for ML.
2. Data credibility: Is your data credible and can it be consolidated into one source? If yes, then ML may work for your use case.
3. **Can you write your objective in terms of what has been observed *(input data)* and what answer *(target label)* you want the model to predict?**
4. Here are the things machine learning does. At a high level, you need to understand this ground truth about ML:
    - [Regression]: **Predict a target value**
    - [Classification]: **Which category**
    - [Clustering]: **Data grouping**
    - [Anomaly detection]: **Segregate the abnormal or weird**
    - [Recommendation]: **Which future path will be the best**

    At a high level, can you link your objective to one of the above
5. Define success metrics for your project early on. For example, you want to predict the correct delivery time with more than 95% accuracy.

Model development life cycle (MDLC): is process many data science projects follow. There are five major steps which we will cover next.



Step 1: Data Collection

To train a model, you need historical data. First collect all the right data. The type of data you need, would depend on the business problem you are trying to solve. The data could come from internal systems such as CRM, ERP, data warehouse, accounting, etc. as well as external systems such IoT sensors, social media, weather, etc.

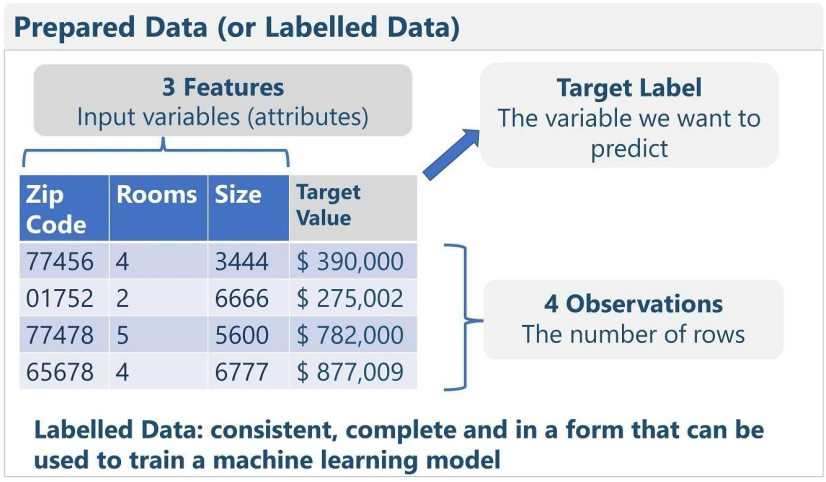
This is a common task in today's data warehousing or data analytics projects. For example, in a fraud predictive system, what data should you use to detect fraudulent transactions? **Select attributes which provide good signals to the end goal.**

You need to have business acumen to collect the right data. This is an area where business and functional experts have a big role to play.

- The more data you have, the better the results are going to be.
- You may need to consolidate data from multiple sources into one central place.
- Clean and prepare the data to make it ready for feeding to a machine learning algorithm.
- Eyeball and analyze the data, run sanity checks to validate the quality of the data and understand the data.

Step 2: Data Preparation

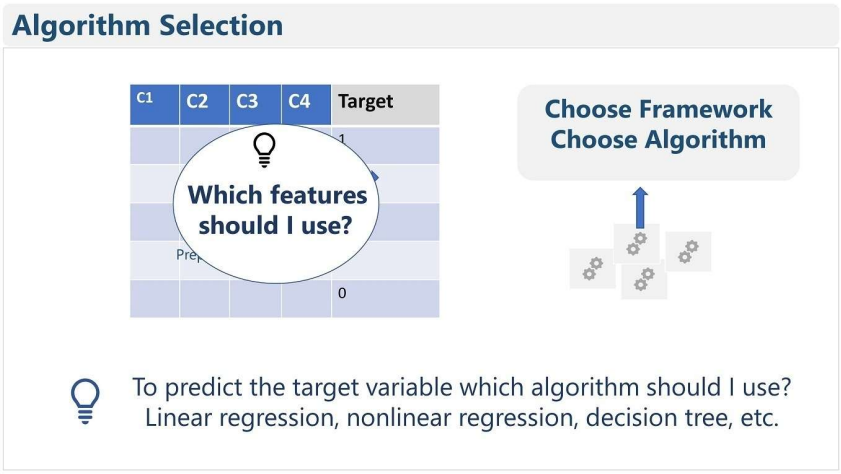
This step involves making the consolidated data "machine learning" ready. The art & science of representing data in the most clear and best way possible is called "**Feature Engineering**". It involves aggregation, cleansing and transformation of the raw attributes to "meaningful and consistent" attributes. Team members must consist of both domain experts *(who understand data)* and data engineers *(who can massage data)*.



- Use case: predicting the price of a house: The dataset shown above is called labelled data (or prepared data). Every row (**or observation**) has a **target value**, which is the price of the house. All other columns are called "**features**", and are relevant to deciding the target value and have values in a consistent format.
- Cleanse and massage your input data, remove duplicates and label it properly. You will need to preprocess your raw data, and convert it into a **labelled dataset**.
- Remove the **features** (or variables) which are redundant or provide no signal in the determination of the target value.
- This is the hardest part of the data engineer or scientist job. The data engineers will need to work hand-in-hand with business experts to create "**prepared data**".

Step 3: Algorithm Selection

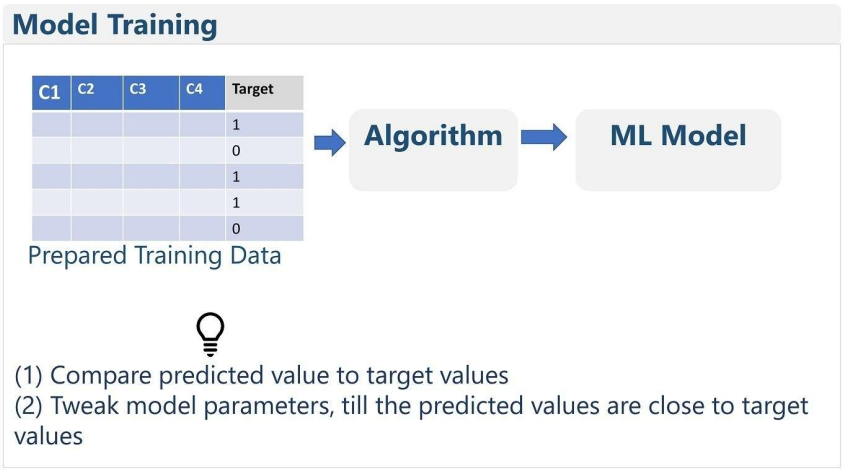
Once the data is prepared, the data scientist will need to determine which algorithm to use. For prediction use cases, you will need to use a regression or classification type of algorithm. For grouping and cluster analysis, clustering algorithms will need to be used.



- A ML algorithm uses training data to create a solution (a model) for the business problem you are trying to solve.
- The algorithm that the data scientist chooses will depend on the business problem. For common use cases, the data scientist can work with standard open source algorithms. The data scientist may have to experiment with a few algorithms, before finalizing the best once.
- In complex use cases, the data scientist may need to create a customized algorithm.
- The objective of this step is to iterate and determine the best algorithm.

Step 4: Model Training and Evaluation

The prepared data is then run through the algorithm, and the result is called the model. During training compare the predicted values to target values. Tweak model parameters, till the predicted values are close to the target values. **The process of generating the best model is called "model training". The data used is called "training data".**

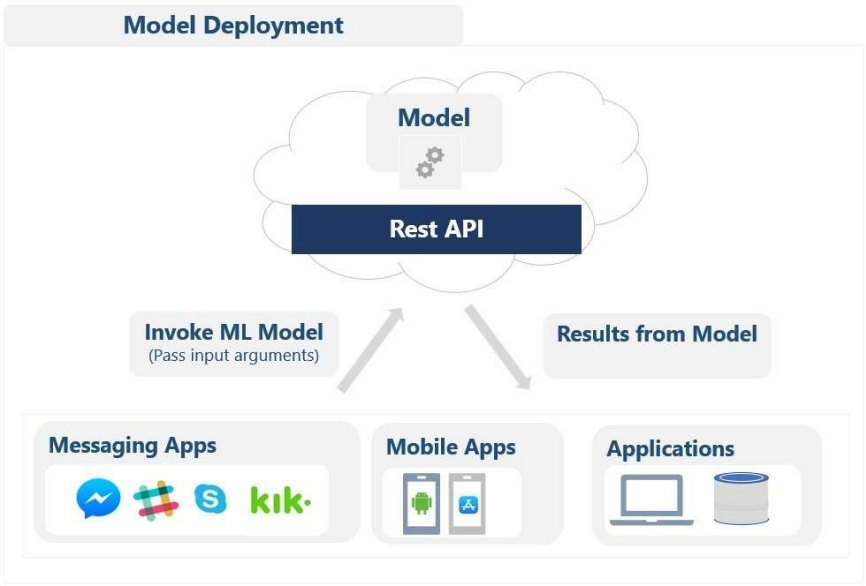


- After you've trained your model, you need to evaluate (test) it to determine the accuracy of the model. Reserve 80% of your prepared data to train the model. Use 20% of the prepared data to evaluate the model.
- Evaluate the model with the remaining 20% of the prepared data to determine whether its accuracy allows you to achieve your business goals.
- Continue tweaking the parameters till the predicted results are as close as possible to the target labels. The model needs to be fine-tuned till it produces the most accurate results.

Step 5: Model Deployment

Once the team is happy with the accuracy of the model, it needs to be deployed so that it can be used by applications. It can be invoked by various apps using a REST API in real-time. Batch predictions can also be executed or scheduled that can run all at once.

Inference: This is the part of machine learning where the ML model carries out the task *(prediction, classification, anomaly detection, etc.)* it was trained to do.



High-level concept diagram showing apps using the model.

Machine learning process is a continuous cycle. After deploying a model, you monitor the predictions, and re-evaluate the model on a consistent basis. You collect the newly collected data and retrain the model to further increase it's accuracy. **Machine learning systems continuously learn and get smarter over time, as new data becomes available.**

Designing and executing an end-to-end machine learning project requires stakeholders and team members made up of subject matter experts, data scientists and traditional IT people.

Supervised Machine Learning

This is the most common type of machine learning found in commercial applications today. Machine learning algorithms are given labelled data, characterized as input and target variables. The task here is to learn y =f(x) from a set of labelled data.

Provided enough data, machine learning algorithms can learn the data patterns, and create a ML model which can be used in the future to make predictions on new data. A human is needed to examine & validate the model till it is sufficiently accurate.

How it works:

1. The algorithm is trained on the prepared data to find the relationship between the input and output variable. This is an iterative process: a human (data scientist) examines and tweaks the parameters till the model is sufficiently accurate.
2. A test dataset is usually used to validate the model.
3. The model can then be applied to new data. It can be used for real-time predictions by your apps and can also be used for batch predictions.
4. The ML model needs to be periodically evaluated and optimized as more data becomes available.

In supervised learning, you know the target value you are trying to predict: so, both classification and regression algorithms can be used for supervised learning. (*It is beyond the scope of this book to cover each algorithm.)*

Classification algorithms	Regression algorithms
Support vector machines (SVM)	Linear regression
Neural networks	Nonlinear regression
Naïve Bayes classifier	Generalized linear models
Logistic regression	Decision trees
Decision trees	Neural networks
Discriminant analysis	
Nearest neighbors (kNN)	

Applications of supervised learnings:

Supervised learning is used in financial applications for credit scoring, spam detection, pattern recognition, algorithmic trading, bond classification, in health applications for tumor detection and in the energy industry for price/forecasting.

Anomaly detection is another common application of supervised learning. A group of data are labelled as anomalies and the task is to spot them. So "Anomaly" is just a label in the classification. To summarize:

The task of supervised learning is regression and classification. This alone can have a transformative impact on your business.

Uses of supervised ML:

- regression (predicting a numeric value)
- binary classification (target can have one of two values)
- multiclass classification (target can have than two values)

It is getting easier and easier to do supervised machine learning. For example, Amazon ML makes it easy for any developer to create prediction models based on supplied data. Azure Machine Learning Studio provides drag-and-drop tools which any developer can use to build predictive analytics solutions.

Examples of **regression** classification problems:

- Based on the number of rooms, bathrooms, size, neighborhood, city, etc. what price will this house sell for?
- Predict annual income based on education level

Examples of **binary** classification use-cases:

- Will the customer cancel account?
- Will the borrower ever pay back the loan?
- Is this image a cat or a dog?

Examples of **multiclass** classification problems:

- What type is this roof: Open Gable, Dormer, Box Gable or Hip?

Unsupervised Learning

In this case you start with unlabeled data (there is NO target value). The most common use of unsupervised learning is in clustering and grouping of data. Unsupervised learning is used to explore data attributes and find hidden patterns in data.

To put it simply, say you want to group customers into 3 groups based on income, assets & spending behavior. A clustering algorithm can be used to detect patterns in the data & create the 3 groups. The team can examine the results, iterate this process till they decide the model is good enough.

Problem: Too much unlabeled data. Target value not know.
Solution: Dimension reduction and clustering.

The task of unsupervised learning is:

- Clustering: find patterns in your data, and grouping the input data into different groups or clusters
- Dimensionality reduction: is similar to compression. Reduce the number of dimensions, with minimum loss of useful information.

How it works

1. Run the algorithm against data
2. The algorithm analyzes the structures & patterns within the data
3. The algorithm classifies the data into different groups based on similar attributes
4. The team can tweak algorithm parameters, iterate this process till they are happy with the groups or clusters.

Common algorithms include:

- k-Means clustering
- Gaussian mixture models
- Hierarchical clustering
- Recommender systems

Applications

Examples of where unsupervised learning methods might be useful:

- Segment customers into smaller groups by similar attributes (same zip code, buying patterns, etc.) for marketing purposes.
- Anomaly detection: discover unusual data patterns which can be useful in discovering faulty pieces of hardware
- Product association: in retail identify sets of products that are usually bought together. Retailers can use this intelligence for side-by-side placement of products in the store or for marketing.

Reinforcement Learning

An algorithm learns to perform a task based on evaluations that are given about how good or bad it's action/result is. It is used when you do not have a lot of training data and you cannot define the final output. The only way to learn is to run the algorithm and provide feedback in real time. It stores & remembers the training examples ("this action was good, that action was bad") through trial-and-error as it performs its task, with the goal of maximizing long-term reward. Self-driving cars, computer games are few examples.

Use cases: Optimize pricing for online auction, calibrate stock and pick robots, etc.

When to use Machine Learning

Machine learning is not a solution for every type of data analysis problem. If your data is not very complex, and you can determine the target value by using business rules and predetermined steps that can be programmed, then you do not need machine learning. Below situations are good candidates for use of machine learning:

Data:
Is the raw material for machine learning. If you do not have or cannot create high volume and high-quality sets of labelled data, your machine learning algorithms may not be very accurate.

High frequency rule-based human processes:
Machine learning is not good for something you do, say 20 times a year. It is good for manual **high volume tasks or decisions you make repeatedly.** For example, you get approx. 500 loan applications every single day on your website. These loan applications have to be manually approved in 24 hours. And you have a team for 25 loan officers manually approving such cases. This is a good ML use case.

There is a relationship between inputs and well-defined output:
ML can be used when you have good labelled data.
1. ML can learn to predict the Y value given input X *(prediction)*
2. ML can classify data *(classification)*

A good example is approving a loan application based on several input variables.

You have a complex task and cannot code the rules:
ML can be considered when you have a task that cannot be solved using a programmed rule-based solution. For example,
- creating a recommendation system for your websites
- determining if an email is spam or not
- fraud detection
- anomaly detection

Complex data:
ML can be used when you have complex data and it keeps changing. For ex you want to predict sales based on a combination of user data, weather data and event data. You want to derive insights by joining internal, external and historical data.

Clearly defined goals:
ML works well when you can clearly define your goals and metrics

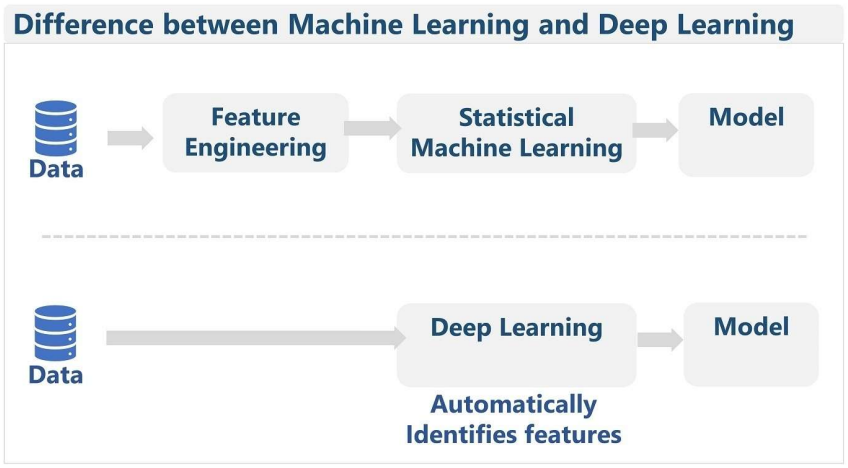
Tolerance for error:
It's almost impossible to train ML algorithms to be 100% accurate. ML works well, if you have a certain tolerance for error.

Deep Learning

The machine learning techniques we covered so far, provided the algorithm prepared data with all the relevant features. The algorithm then uses statistical methods to generate the model. But most real world problems usually consist of unstructured data, such as images, videos, text, audio, etc. That's where deep learning comes in.

Deep learning is a subset of machine learning that mainly deals with unstructured data. What we learnt so far is often called "shallow machine learning" which is task based, such as predicting the price of a house based on various variables. Deep learning is designed to copy the way our brains learn by creating artificial "neural networks" that can understand complex concepts, features and relationships from data. Deep learning uses artificial neural networks, which have multiple layers of many connected artificial neurons

In deep learning, all you do is feed the data to the algorithm and it figures out the features. So for image recognition, you just provide the algorithm several labelled images of dogs, cats, zebra, horses, etc. You do not need to provide the features of the animals such as a dog has 2 eyes, zebra has stripes, tail, shape of tail, etc. The algorithm figures that out and generates the model.



Deep learning is used in complex tasks such as:

Computer Vision:
Using large datasets of labeled images, algorithms can be trained to recognize objects on their own. Deep learning neural networks can identify objects & subjects with human like accuracy. This capability is currently being used in facial recognition AI applications, robots, security systems, self-driving cars, etc. and more.

Speech Recognition:
Deep learning has made it possible for computers to understand human voice. Even though voice is a complex dataset, due to varying speech styles, patterns and accents: deep learning algorithms can be trained to understand what is being said. This capability is used today in Siri, Alexa, Google Assistant and other virtual assistants.

Natural Language Processing:
Deep learning makes it possible for computers to understand natural human conversations and large volumes of text. In conversations, deep learning algorithms can detect language, emotions, intent and context of the conversation. This capability is being used in creating customer service bots that can converse with users in a natural way.

Deep learning is used in many products such as face recognition, self-driving cars, smart voice assistants, precision drones, etc. We will cover these capabilities of AI later in this book.