MACHINE LEARNING

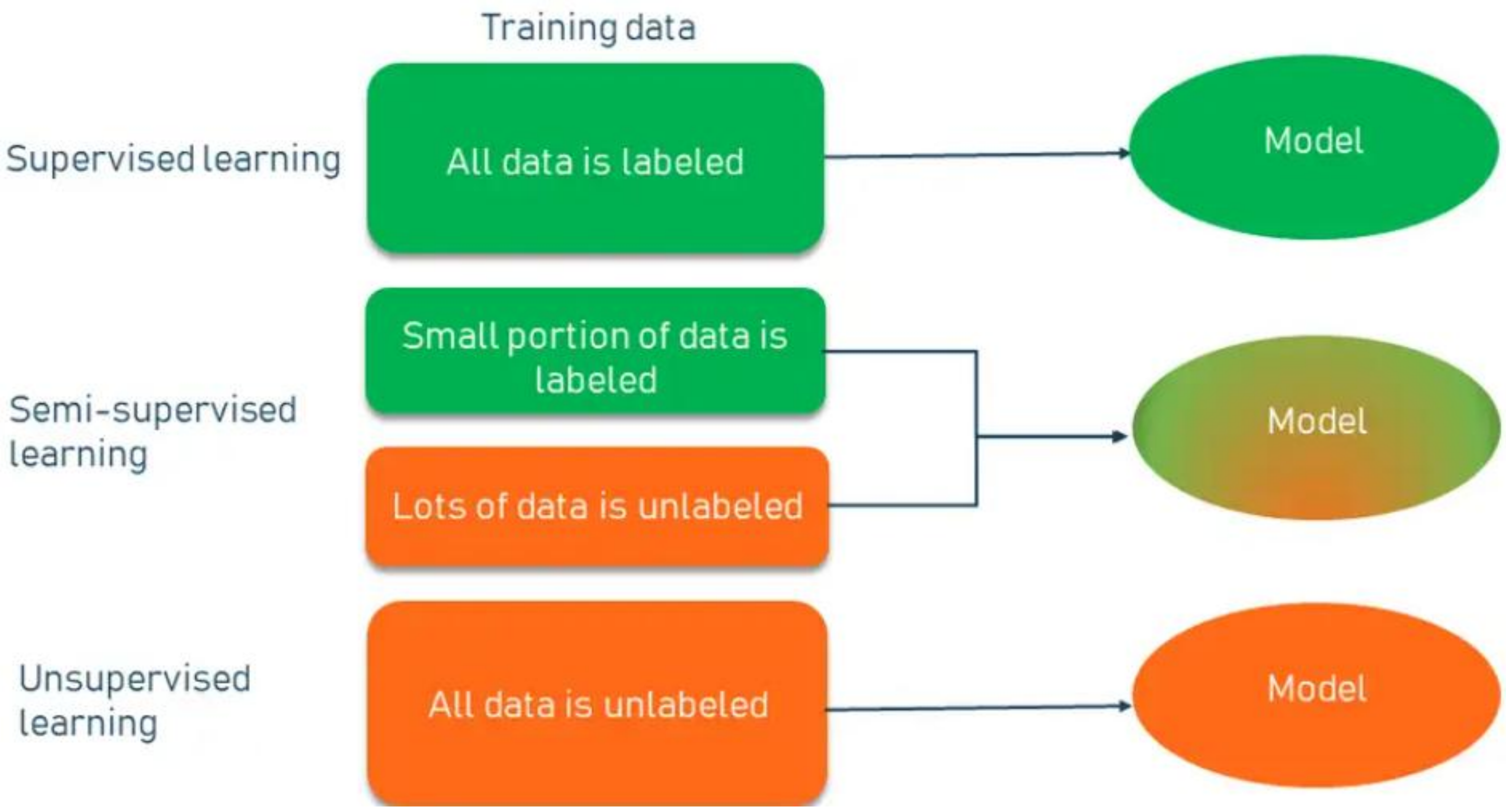
Semi Supervised Learning

Semi Supervised Learning

- On the one hand, supervised learning is the bread-and-butter of machine learning (ML) techniques, but is powered by labeled data which is tedious and expensive to annotate. Alternatively, unsupervised learning uses unlabeled data, which without human-made annotations is often plentiful.
- When used alone, either of these strategies is often impractical for training a model to deployment-ready benchmarks. Labeling an entire dataset is time-consuming and expensive, and unlabeled data may not provide the desired accuracy.
- What if we have access to both types of data? Or what if we only want to label a percentage of our dataset? How can we combine both our labeled and unlabeled datasets to improve model performance?
- We can use **SEMI-SUPERVISED** intuition to answer these questions, as it leverages both labeled and unlabeled data to bolster model performance.

- Semi-supervised learning is a broad category of machine learning techniques that utilizes both labeled and unlabeled data; in this way, as the name suggests, it is a hybrid technique between supervised and unsupervised learning.
- In a nutshell, **semi-supervised learning (SSL)** is a machine learning technique that uses a small portion of labeled data and lots of unlabeled data to train a predictive model.

SUPERVISED LEARNING vs SEMI-SUPERVISED LEARNING vs UNSUPERVISED LEARNING



Training data

Supervised learning

All data is labeled

Model

Semi-supervised learning

Small portion of data is labeled

Lots of data is unlabeled

Model

Unsupervised learning

All data is unlabeled

Model

- **Semi-supervised learning** bridges supervised learning and unsupervised learning techniques to solve their key challenges.
- With it, you train an initial model on a few labeled samples and then iteratively apply it to the greater number of unlabeled data.
- Unlike unsupervised learning, SSL works for a variety of problems from classification and regression to clustering and association.
- Unlike supervised learning, the method uses small amounts of labeled data and also large amounts of unlabeled data, which reduces expenses on manual annotation and cuts data preparation time.

How semi-supervised learning works?

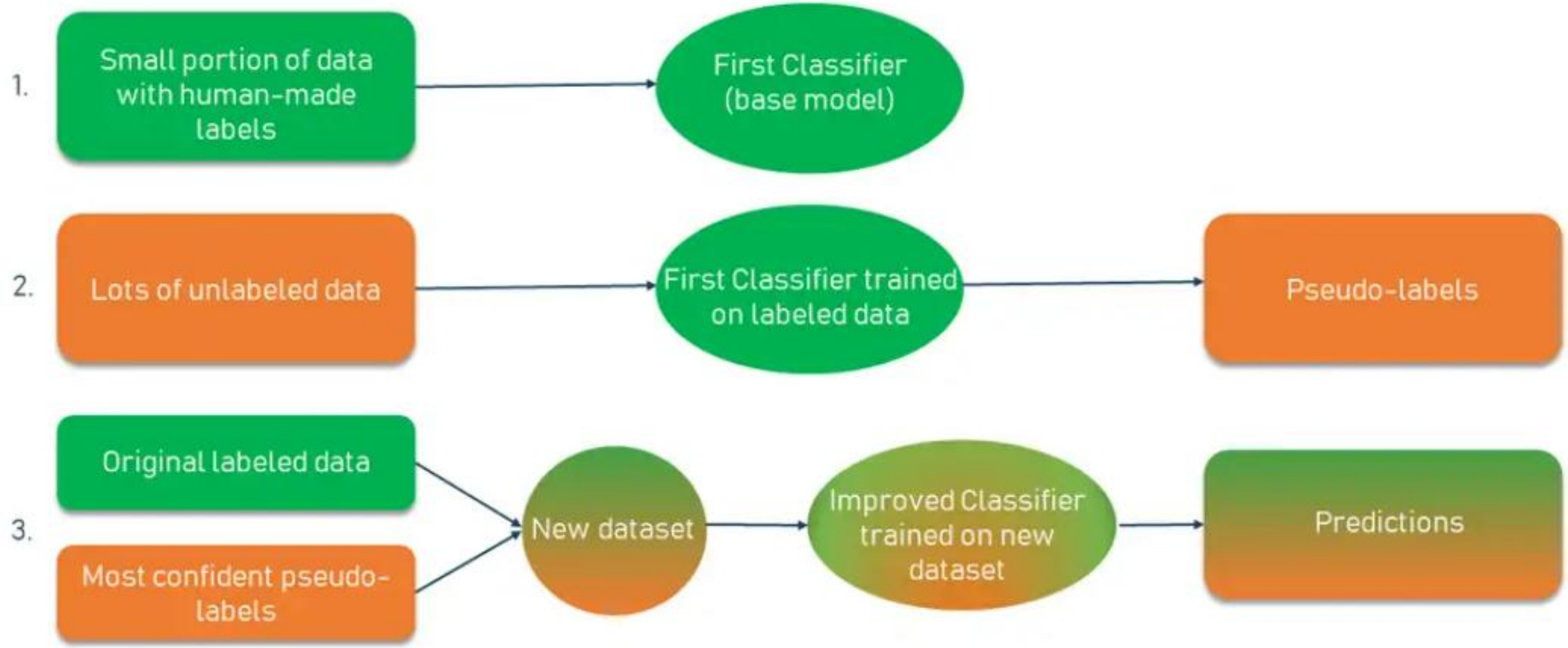
- Imagine, you have collected a large set of unlabeled data that you want to train a model on. Manual labeling of all this information will probably cost you a fortune, besides taking months to complete the annotations. That’s when the semi-supervised machine learning method comes to the rescue.

- The working principle is quite simple. Instead of adding tags to the entire dataset, you go through and hand-label just a small part of the data and use it to train a model, which then is applied to the ocean of unlabeled data.

Self training.

- One of the simplest examples of semi-supervised learning, in general, is **Self-training** is the procedure in which you can take any supervised method for classification or regression and modify it to work in a semi-supervised manner, taking advantage of labeled and unlabeled data. The standard workflow is as follows.

SEMI-SUPERVISED SELF-TRAINING METHOD



1. Small portion of data with human-made labels → First Classifier (base model)

2. Lots of unlabeled data → First Classifier trained on labeled data → Pseudo-labels

3. Original labeled data, Most confident pseudo-labels → New dataset → Improved Classifier trained on new dataset → Predictions

Scenario

- You pick a small amount of labeled data, e.g., images showing cats and dogs with their respective tags, and you use this dataset to train a base model with the help of ordinary supervised methods.
- Then you apply the process known as *pseudo-labeling* — when you take the partially trained model and use it to make predictions for the rest of the database which is yet unlabeled. The labels generated thereafter are called *pseudo* as they are produced based on the originally labeled data that has limitations (say, there may be an uneven representation of classes in the set resulting in bias — more dogs than cats).

Scenario

- From this point, you take the most confident predictions made with your model (for example, you want the confidence of over 80 percent that a certain image shows a cat, not a dog). If any of the pseudo-labels exceed this confidence level, you add them into the labeled dataset and create a new, combined input to train an improved model.
- The process can go through several iterations (10 is often a standard amount) with more and more pseudo-labels being added every time. Provided the data is suitable for the process, the performance of the model will keep increasing at each iteration.

Semi-supervised learning examples

- Speech recognition
- Web content classification
- Text document classification

When to use and not use semi-supervised learning

- With a minimal amount of labeled data and plenty of unlabeled data, semi-supervised learning shows promising results in classification tasks while leaving the doors open for other ML tasks. Basically, the approach can make use of pretty much any supervised algorithm with some modifications needed.
- On top of that, SSL fits well for clustering and anomaly detection purposes too if the data fits the profile. While a relatively new field, semi-supervised learning has already proved to be effective in many areas.
- Semi supervised learning is applicable to all tasks. If the portion of labeled data isn’t representative of the entire distribution, the approach may fall short. Say, you need to classify images of colored objects that have different looks from different angles.
- Unless you have a large amount of labeled data, the results will have poor accuracy. But if we're talking about lots of labeled data, then semi-supervised learning isn’t the way to go. Like it or not, many real-life applications still need lots of labeled data, so supervised learning won’t go anywhere in the near future.