**Step One: Define the Problem**

**How do You Start a Machine Learning Task?**

* ***Define a very specific task.***
  + Think back to the snow cone sales example. Now imagine that you own a frozen treats store and you sell snow cones along with many other products. You wonder, "‘How do I increase sales?" It's a valid question, but it's the **opposite** of a very specific task. The following examples demonstrate how a machine learning practitioner might attempt to answer that question.
    - “Does adding a $1.00 charge for sprinkles on a hot fudge sundae increase the sales of hot fudge sundaes?”
    - “Does adding a $0.50 charge for organic flavors in your snow cone increase the sales of snow cones?”
* ***Identify the machine learning task we might use to solve this problem.***
  + This helps you better understand the data you need for a project.

**What is a Machine Learning Task?**

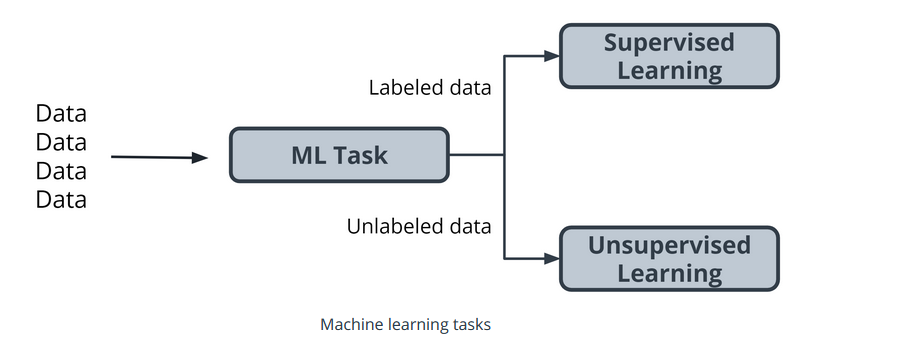
All model training algorithms, and the models themselves, take data as their input. Their outputs can be very different and are classified into a few different groups based on the *task* they are designed to solve. Often, we use the kind of data required to train a model as part of defining a machine learning task.

In this lesson, we will focus on two common machine learning tasks:

* **Supervised** learning
* **Unsupervised** learning

**Supervised and Unsupervised Learning**

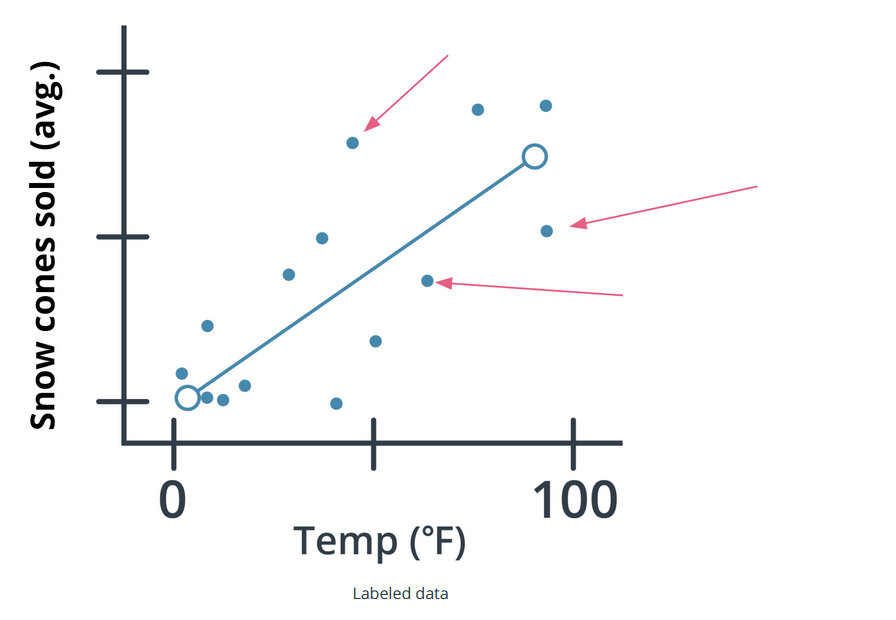
The presence or absence of labeling in your data is often used to identify a machine learning task.



**Supervised tasks**

A task is *supervised* if you are using labeled data. We use the term *labeled* to refer to data that already contains the solutions, called *labels*.

For example: Predicting the number of snow cones sold based on the temperatures is an example of supervised learning.

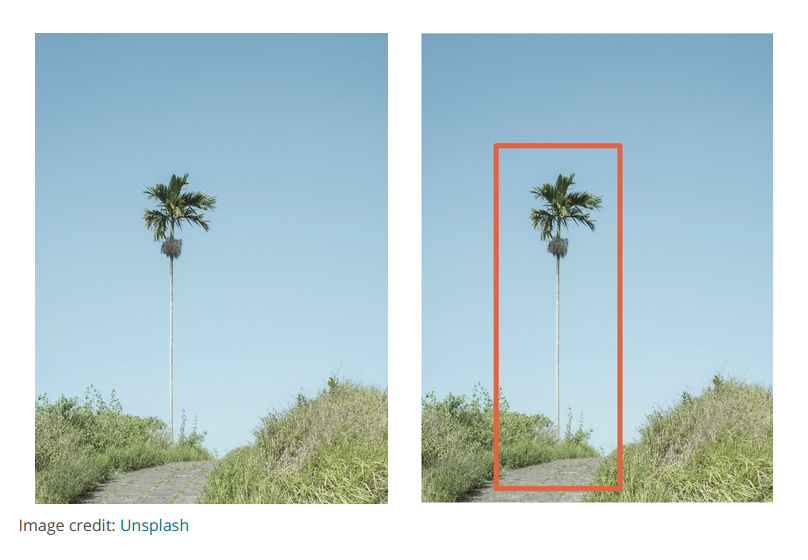


In the preceding graph, the data contains both a temperature and the number of snow cones sold. Both components are used to generate the linear regression shown on the graph. Our goal was to predict the number of snow cones sold, and we feed that value into the model. We are providing the model with labeled data and therefore, we are performing a *supervised machine learning task*.

**Unsupervised tasks**

A task is considered to be *unsupervised* if you are using *unlabeled data*. This means you don't need to provide the model with any kind of label or solution while the model is being trained.

Let's take a look at unlabeled data.



* Take a look at the preceding picture. Did you notice the tree in the picture? What you just did, when you noticed the object in the picture and identified it as a tree, is called *labeling the picture*. Unlike you, a computer just sees that image as a matrix of pixels of varying intensity.
* Since this image does not have the labeling in its original data, it is considered *unlabeled.*

**How do we classify tasks when we don't have a label?**

Unsupervised learning involves using data that doesn't have a label. One common task is called **clustering**. Clustering helps to determine if there are any naturally occurring groupings in the data.

Let's look at an example of how clustering in unlabeled data works.

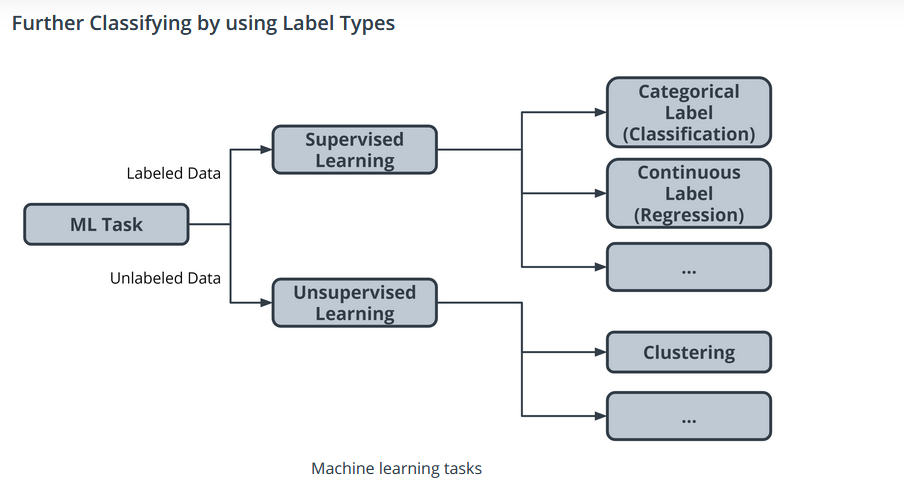
**Identifying book micro-genres with unsupervised learning**

Imagine that you work for a company that recommends books to readers.

*The assumption*: You are fairly confident that micro-genres exist, and that there is one called *Teen Vampire Romance*. Because you don’t know which micro-genres exist, you can't use **supervised learning** techniques.

This is where the **unsupervised learning** clustering technique might be able to detect some groupings in the data. The words and phrases used in the book description might provide some guidance on a book's micro-genre.

**Further Classifying by using Label Types**

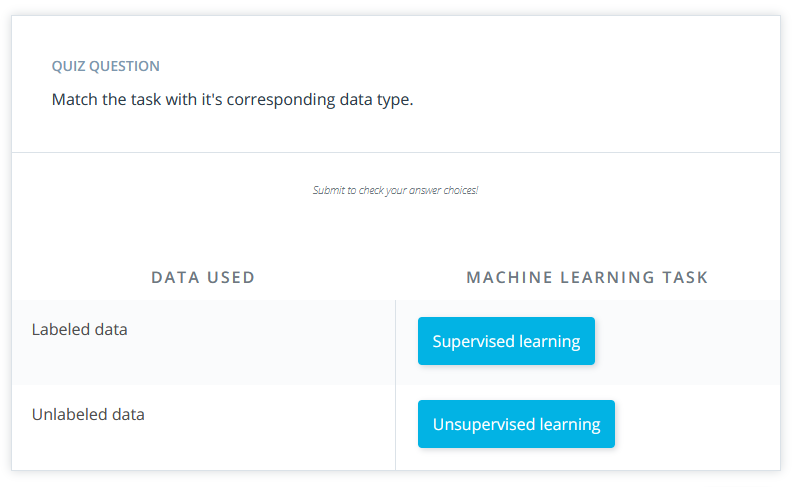


Initially, we divided tasks based on the presence or absence of labeled data while training our model. Often, tasks are further defined by the type of label which is present.

In **supervised** learning, there are two main identifiers you will see in machine learning:

* A **categorical** label *has a* discrete *set of possible values. In a machine learning problem in which you want to identify the type of flower based on a picture, you would train your model using images that have been labeled with the categories of flower you would want to identify. Furthermore, when you work with categorical labels, you often carry out* classification tasks\*, which are part of the supervised learning family.
* A **continuous** (regression) label *does not have a discrete set of possible values, which often means you are working with numerical data. In the snow cone sales example, we are trying to predict the* number\* of snow cones sold. Here, our label is a number that could, in theory, be any value.

In unsupervised learning, **clustering** is just one example. There are many other options, such as deep learning.



Terminology

* Clustering. Unsupervised learning task that helps to determine if there are any naturally occurring groupings in the data.
* A *categorical label* has a discrete set of possible values, such as "is a cat" and "is not a cat."
* A continuous (regression) label does not have a discrete set of possible values, which means possibly an unlimited number of possibilities.
* Discrete: A term taken from statistics referring to an outcome taking on only a finite number of values (such as days of the week).
* A label refers to data that already contains the solution.
* Using unlabeled data means you don't need to provide the model with any kind of label or solution while the model is being trained.

Additional Reading

* The [AWS Machine Learning blog](https://aws.amazon.com/blogs/machine-learning/) is a great resource for learning more about projects in machine learning.
* You can use Amazon SageMaker [to calculate new stats in Major League Baseball](https://aws.amazon.com/blogs/machine-learning/calculating-new-stats-in-major-league-baseball-with-amazon-sagemaker/).
* You can also find an article on [Flagging suspicious healthcare claims with Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/flagging-suspicious-healthcare-claims-with-amazon-sagemaker/) on the AWS Machine Learning blog.
* What [kinds of questions and problems](https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-problems-in-amazon-machine-learning.html) are good for machine learning?