Exploration: What is Machine Learning? (w1)

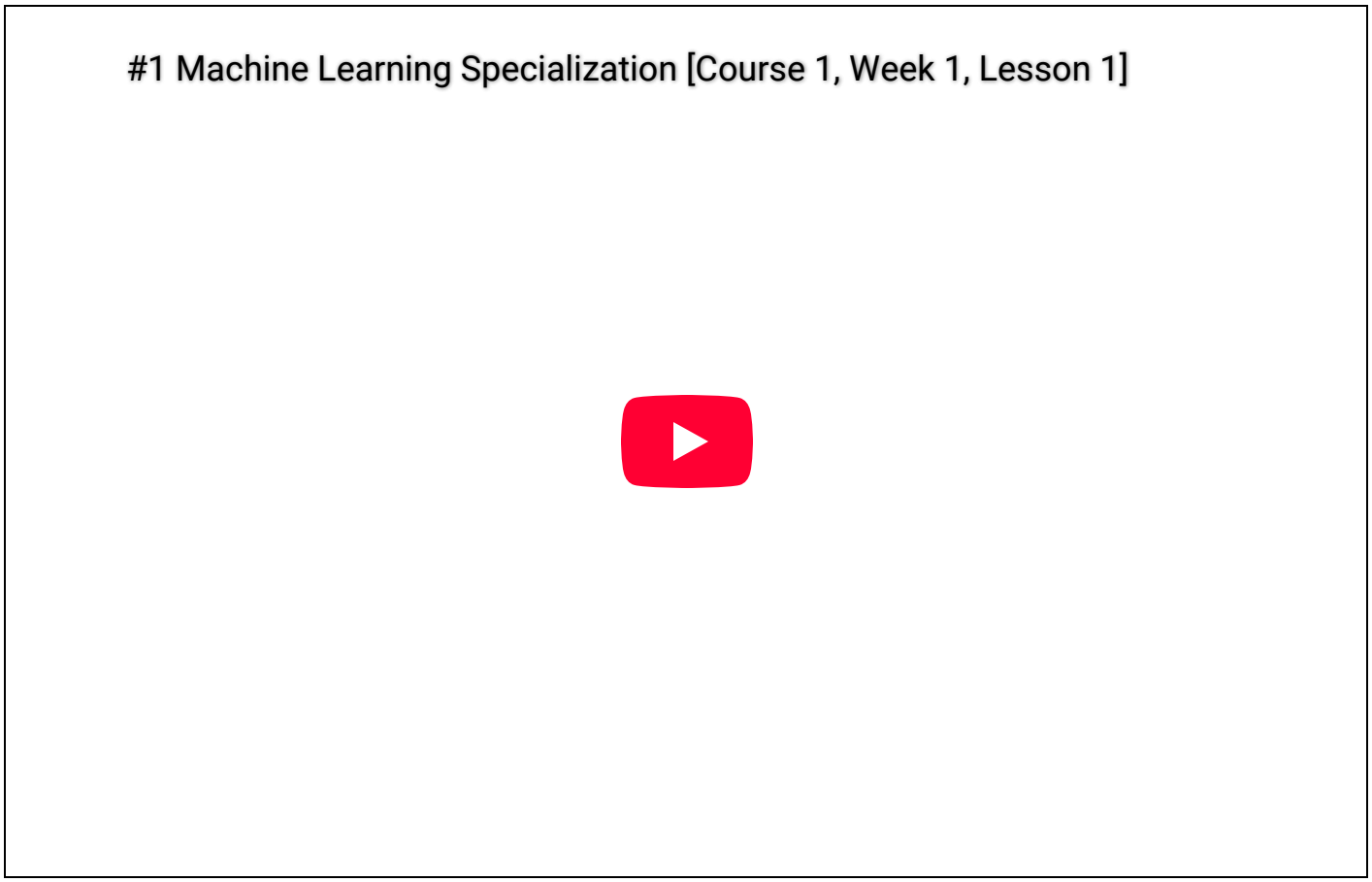
Introduction

Take a moment to *really* answer this question: how can we use machine learning to solve real-world problems? As end users of software systems, we can probably describe many experiences that, somehow, rely on machine learning, perhaps without us even noticing. For example, any time we conduct a Google search, we are experiencing an application of machine learning. Or, perhaps we have communicated with "smart" chat agents, that have been trained by machine learning. In this course, we hope to enable all students to apply machine learning to solve problems in their workplace and in the world. We'll begin our learning journey by exploring what machine learning is.

In this Exploration, we'll identify some applications of machine learning in action, define machine learning and its components, and describe the steps of the machine learning process.

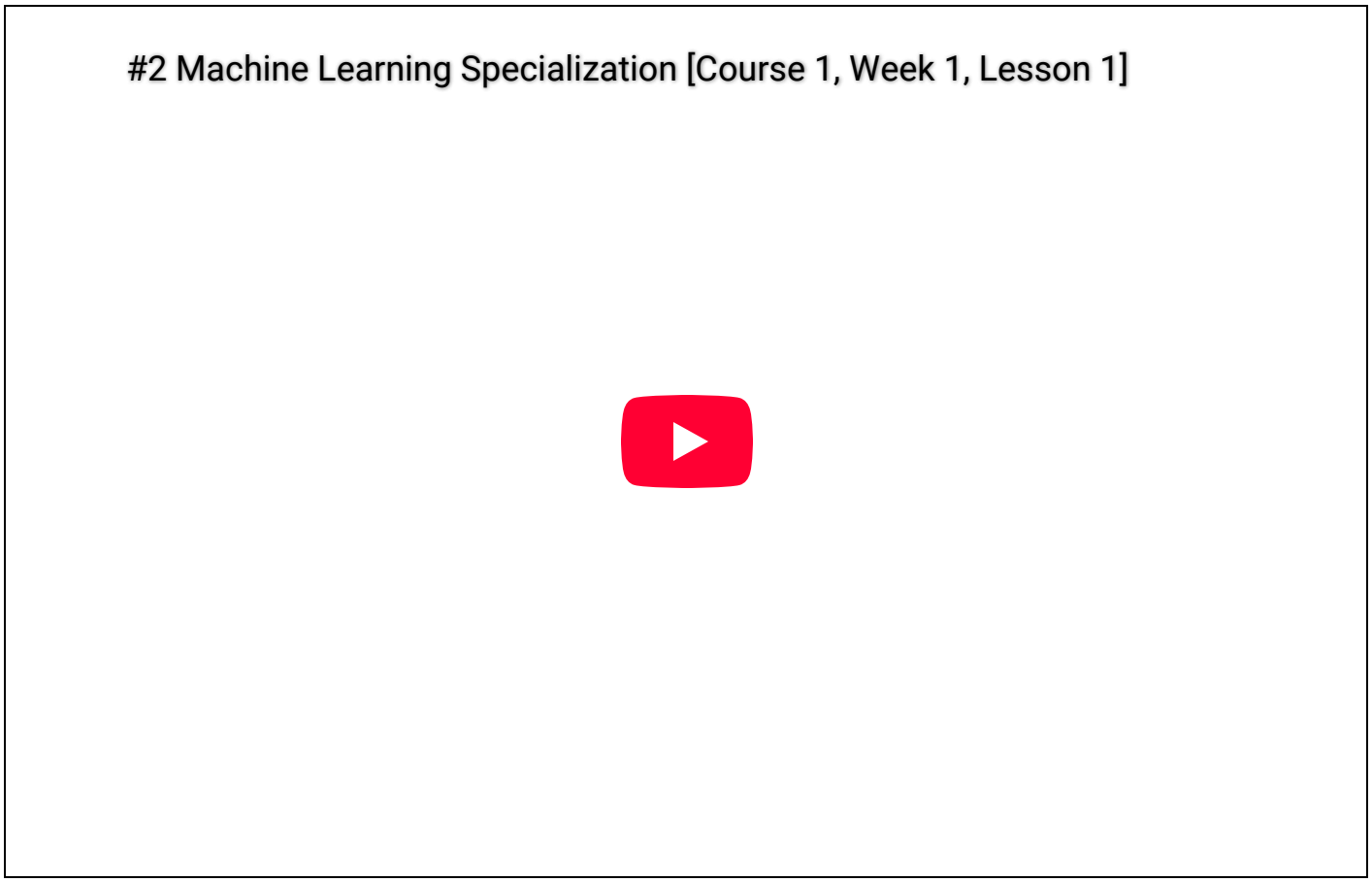
Applications of Machine Learning

Even if you don't know anything about machine learning, you have probably **heard about machine learning in the news (https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html)** , or **used an app that, behind the scenes, uses machine learning (https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0)** . Before we define machine learning, let's take a look at some important, constructive applications of machine learning in action, presented by Dr. Andrew Ng, one of the most well-known experts in machine learning.



Far beyond a novelty, and beyond consumer products, we can see how machine learning enables us to solve problems in business, industry, climate change, health care, and manufacturing. If you pay attention to Dr. Ng, you'll notice he drops a brief definition of machine learning: "the science of getting computers to learn without being explicitly programmed." We'll investigate this definition in a moment.

For now, let's continue to explore some applications of machine learning, and see if we can describe where machine learning came from.



We can say that the field of machine learning emerged from the problems within the field of artificial intelligence, and that machine learning has enabled us to implement higher quality AI systems. For decades, computer scientists have been attempting to create intelligent software, but found that *explicitly* programming intelligence did not always yield the desired results. For example, we have found that we can design and implement explicit path-finding algorithms, that feel "smart" enough to give you directions from point A to point B. But designing explicit algorithms for many tasks, such as speech recognition or tumor detection, has been very hard to do.

By *explicitly* programming, consider this simple program that recommends whether or not to wear a coat.

```
def wear_a_coat():
    if is_cold() or is_raining() or is_mild() and you_tend_to_feel_cold() ...
```

Here, we must decide *ahead of time* all of the possible factors that go into recommending whether or not one should wear a coat; and we have *explicitly programmed* this algorithm. It would be fantastic if, somehow, a computer (a "machine") could learn, on its own, all of the deciding factors that go into making a recommendation on whether to wear a coat - or how to interpret human speech, make recommendations, detect health issues, or make predictions - without being explicitly programmed about the details.

If we pay attention to the world around us, we'll notice machine learning continuously at work: interacting with intelligent voice agents, driving automobiles, face detection for security systems,

and yes, even watching Netflix.

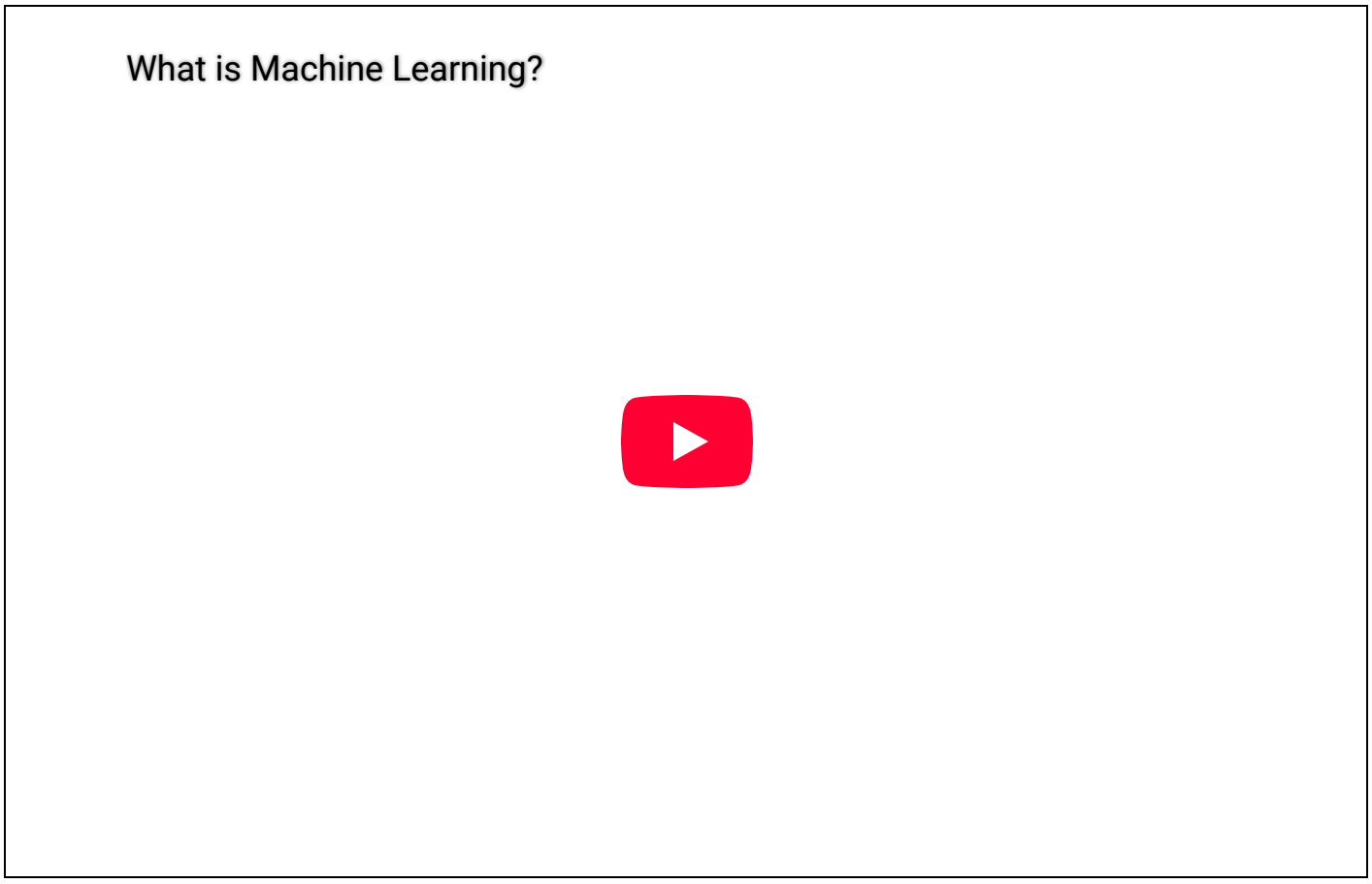
Key Points

- Machine learning emerged from needs within the larger field of artificial intelligence
- Machine learning is used in consumer products, such as search and recommendation systems
- Machine learning is used in business, industry, manufacturing, health care, finance, security, and the sciences, and has broad, real-world impacts

What is Machine Learning?

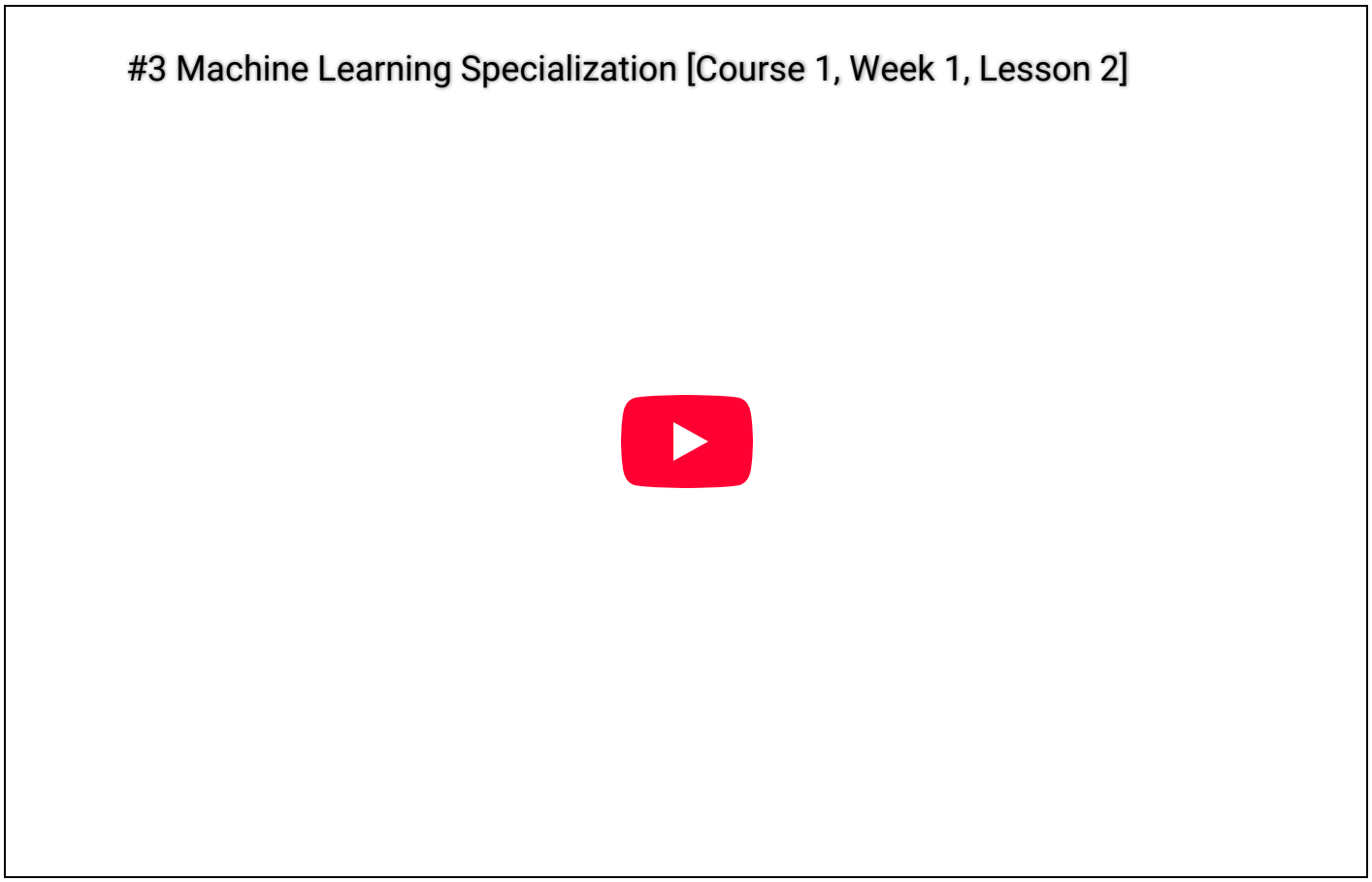
As a student of statistics, perhaps you have heard "Machine learning is really just glorified statistics." On the one hand, it is comforting to know that machine learning "just" relies on mathematics. But even if we know statistics and mathematics, we do not necessarily know machine learning, so there *must* be more to machine learning than just glorified math.

Let's take a look at a generalized, but incomplete, definition of machine learning.



"Using data to make predictions" is a good start to understanding what machine learning is, but this falls short of the details we need to become machine learning practitioners. After all, statisticians and data analysts have been using data to craft explanations and make predictions for many years - without machine learning. But notice the high-level process of machine learning: using *data* to *train* a *model* to make *predictions*.

Let's return to expert Dr. Andrew Ng to see a more formal definition of machine learning.



In this course, we agree with Dr. Ng and define machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed," as defined by Arthur Samuel in 1959. Notice how Arthur Samuel did not create an *explicit* program containing all the tactics of how to win a game of checkers. Instead, he created a system that enabled a machine (a computer system) to play checkers and let that machine play thousands of games of checkers. The machine "learned" to play checkers well based on its experience of playing those thousands of games - not by being explicitly programmed to play checkers well.

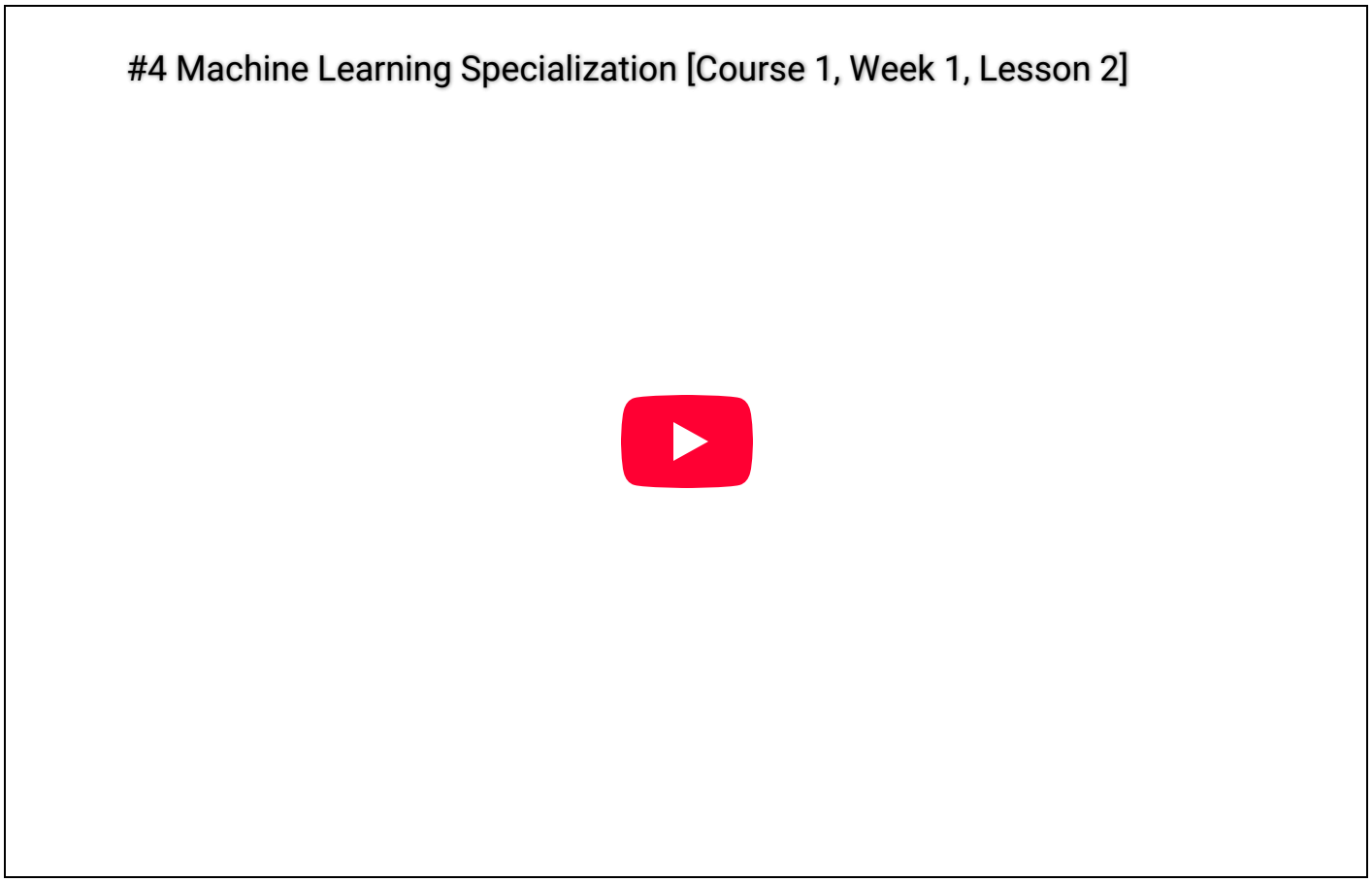
Turning back to our "coat wearing recommendation system," we would avoid crafting an explicit logic statement in code to make a recommendation. Instead, we would collect data from thousands of people who wear a coat or not, and pay attention to temperature, precipitation, humidity, preference, time of day, and so on, for each example of coat wearing / no coat wearing. We would use that data to create or "train" a statistical model, informed by that past data, that we can integrate into a *new* computer program that, given the *current* temperature, precipitation, humidity, etc, propose a coat recommendation for us.

We would like to emphasize that machine learning is a *process* that gives computers the ability to learn. This process involves writing code that analyzes data to create a "model," often backed by statistics, and then applying that model to make predictions about new data. We will dig deeper into the machine learning process later in another Exploration.

Before we continue, we find it helpful to create a mental framework about machine learning, by dividing it into two main subtopics: *supervised learning* and *unsupervised learning*.

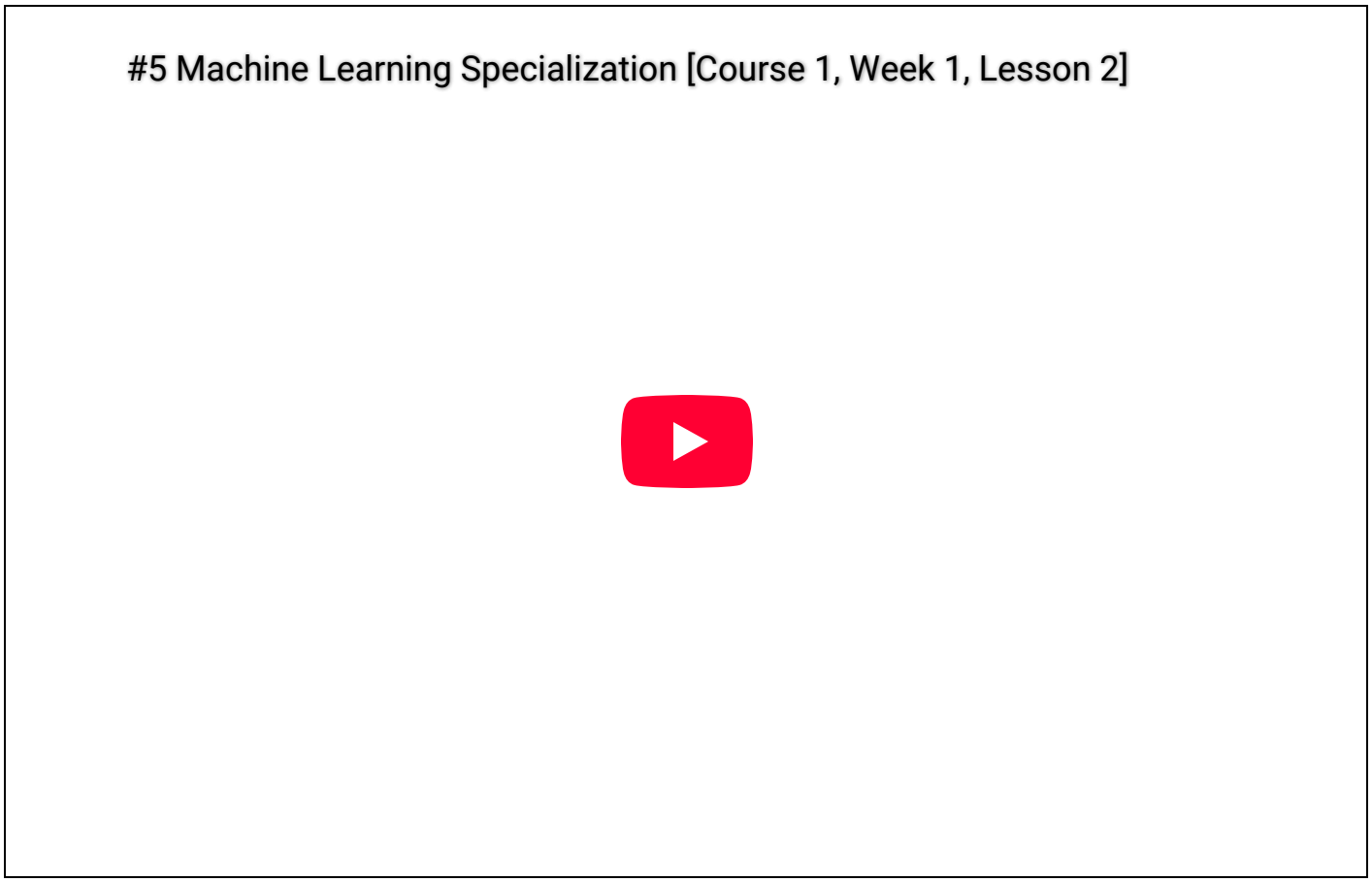
Supervised Learning

Supervised learning is one "family" of algorithmic techniques in machine learning. As a student of statistics, you probably already know of one supervised learning technique: regression.



The key concept of supervised learning is that we have a "label" or "output" we are trying to predict, and the training data we have to work with has some labeled data that we can learn from. When we have an existing dataset of examples that are labeled, we can train a machine to create a statistical model that is able to predict a label for new, unlabeled datum. When the "labels" are continuous numeric data, with an infinite number of possible outputs, such as housing prices, we tend to apply a *regression* model. In this course, we will learn to apply **linear regression** and **logistic regression**.

Classification is another supervised learning technique. Let's investigate what *classification* is, and how it is related to *regression*.



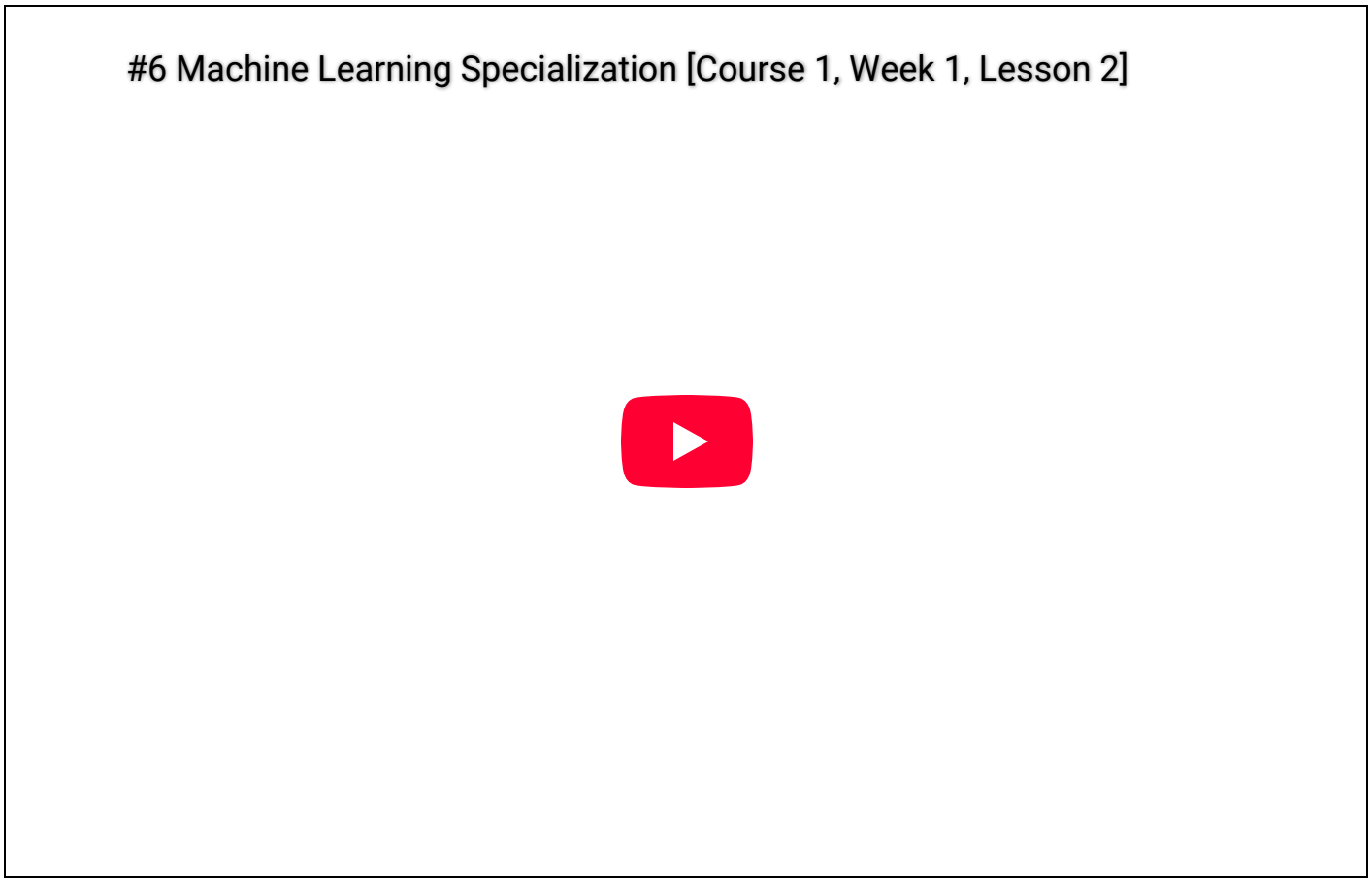
Classification serves the goal of predicting a label to apply to new, unlabeled datum. Similar to regression, we use labeled data - a corpus of "right answers" - to build a model that is able to predict a label for new, unlabeled datum. In classification, however, we often have a finite, discrete, and smaller number of possible labels to produce as output. For example, *benign*, *malignant 1* and *malignant 2*.

Some common classification models include decision trees, rule systems, nearest-neighbors, bayesian classifiers, artificial neural networks, support vector machines, and ensemble methods (bagging, boosting, random forests). In this course, we will learn to apply **k-nearest neighbors**, **perceptrons**, and **support vector machines**.

Unsupervised Learning

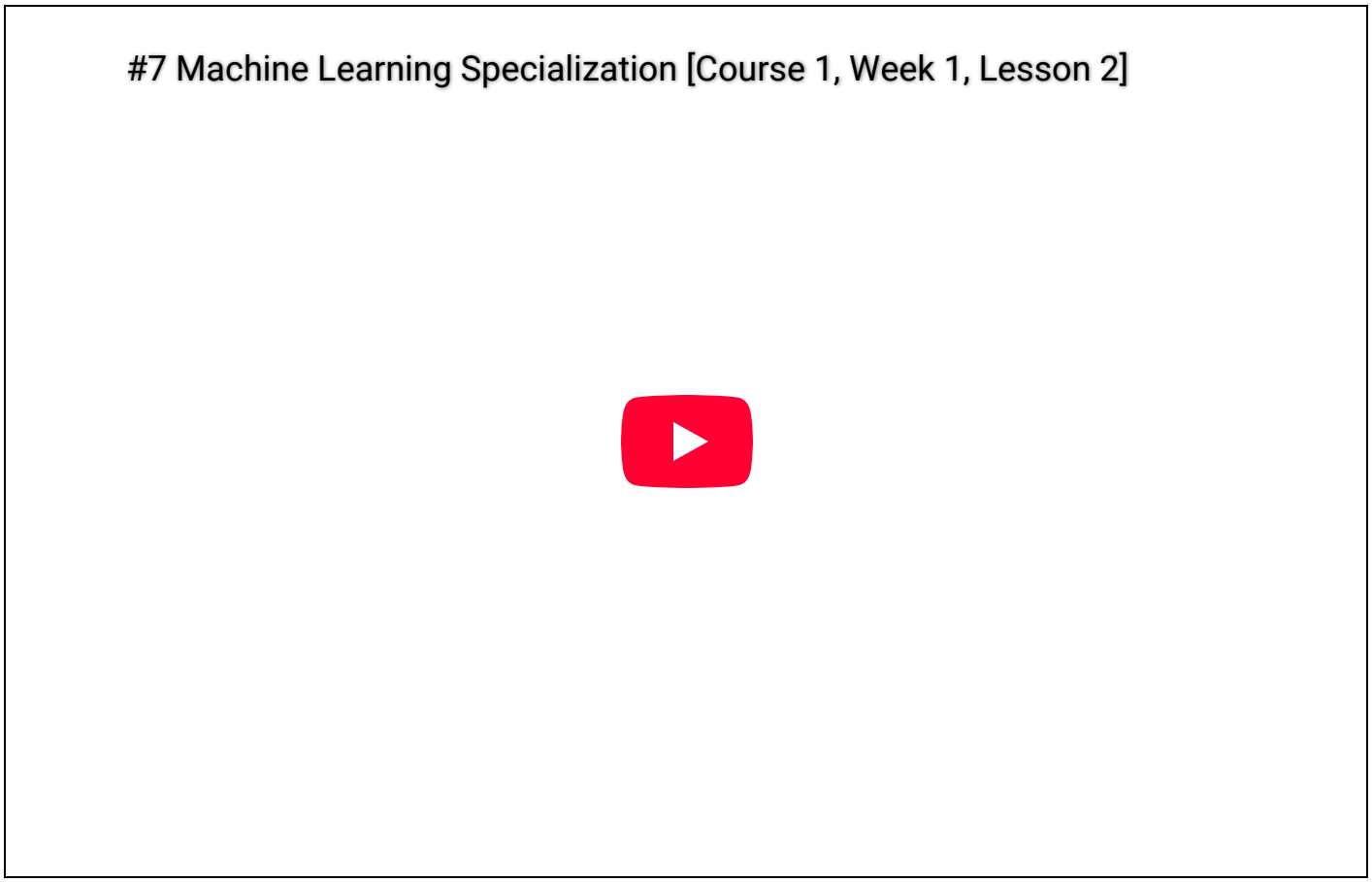
Supervised learning with regression and classification make predictions of values or labels, based on data consisting of already-labeled examples. In contrast, unsupervised learning does not involve readily-labeled data. As such, unsupervised learning answers different kinds of questions, or makes different kinds of predictions.

Clustering is an unsupervised learning technique that tells us if there are interesting, often hidden, patterns or similarities in the data.



As we have seen here, unsupervised learning informs us about interesting things in *unlabeled* data. Some use cases include media recommendation systems, customer segmentation analysis, genetic research, and more. Some common clustering models include k-Means clustering, agglomerative hierarchical clustering, density-based clustering, and graph-based clustering. In this course, we will **not learn about clustering**.

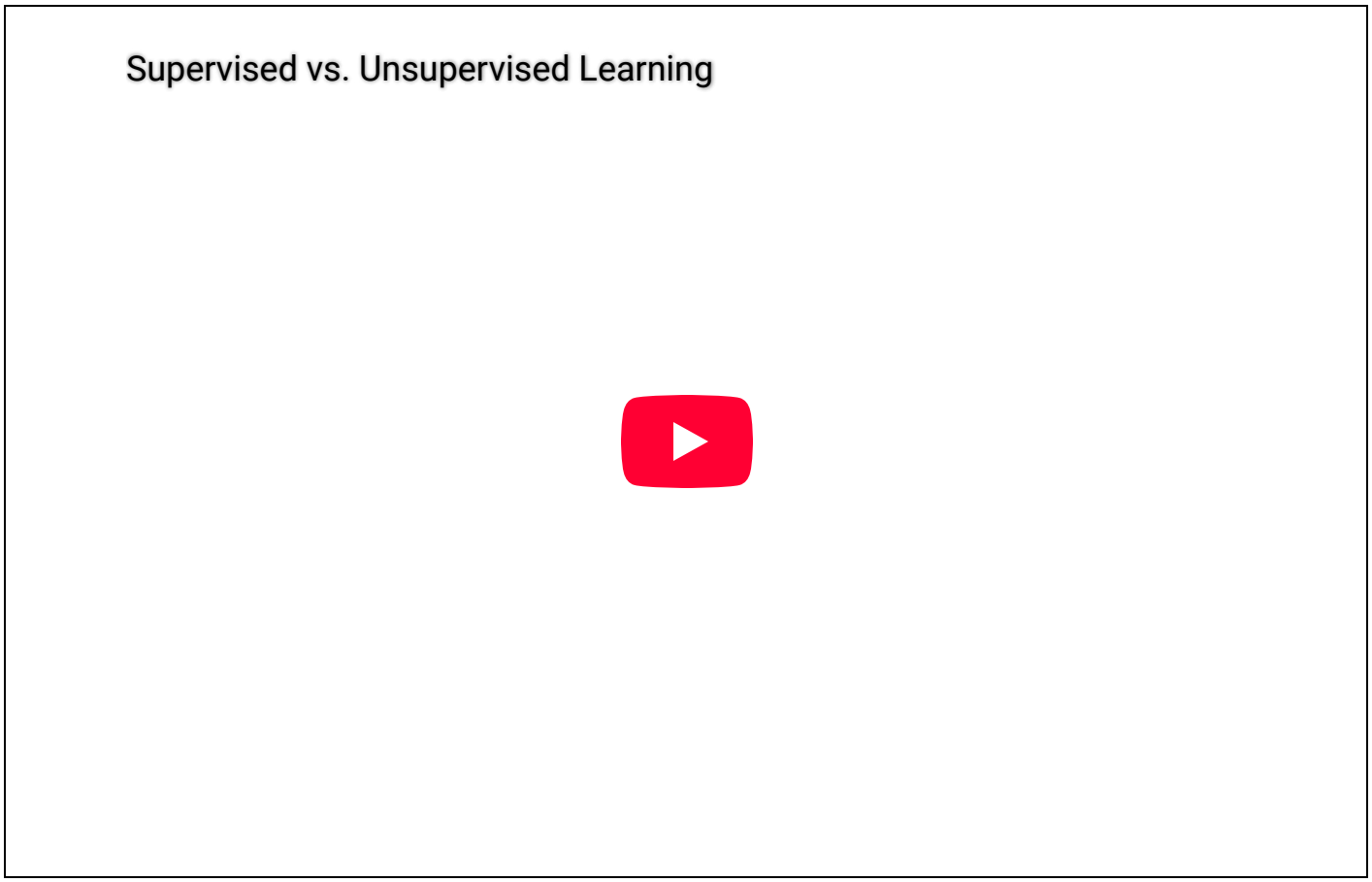
Anomaly detection, *association analysis*, and *dimensionality reduction* are also three additional unsupervised learning techniques.



Anomaly detection enables a system to identify an event or datum as being very different from past data, and is a technique that banks and security companies use to detect fraud. *Association analysis* is a means of creating a model that reveals interesting relationships between data, especially in the domain of market basket transactions. For example, using association analysis, a large retailer may discover that people who buy bread, milk and diapers also tend to buy beer. Lastly, *dimensionality reduction* helps practitioners automatically separate the features in a dataset that are most important from those that are not. We often use dimensionality reduction to simplify or "compress" our datasets as an intermediate step prior to training for our primary machine learning goal.

Review: Supervised vs Unsupervised Learning

Machine learning is a process through which we use existing data to "teach" models that can make predictions or tell us about interesting things in our data. Machine learning has proven to be a valid means of building artificial intelligence systems, and consists of many well-known techniques and models.



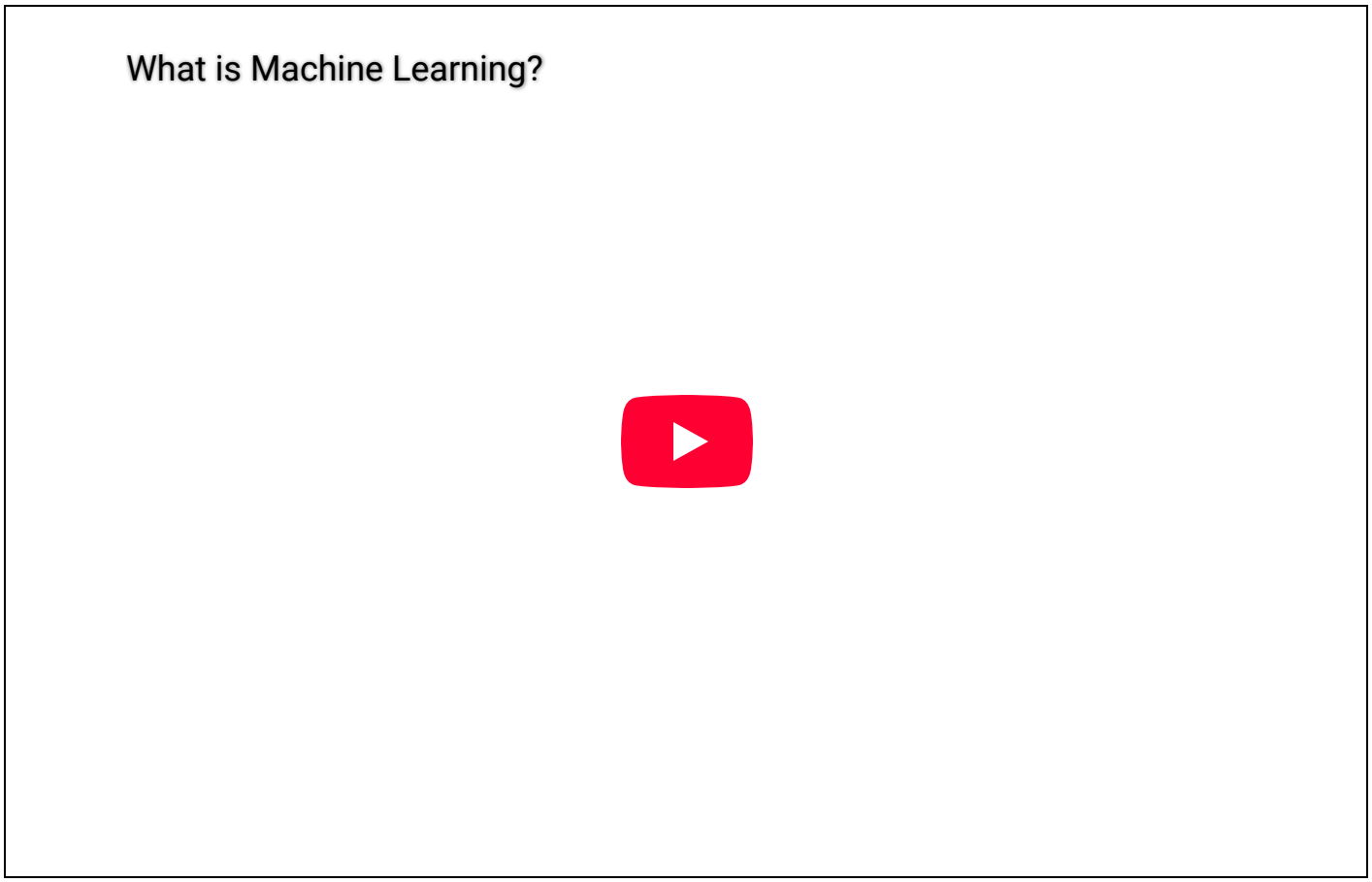
It is helpful to divide machine learning into two sub-areas: supervised learning and unsupervised learning. Supervised learning consists of regression, and classification. Regression problems make use of well-known statistical models. Classification problems make use of decision trees, k-nearest-neighbors, artificial neural networks, and other models. Unsupervised learning consists of clustering, association analysis, anomaly detection, and dimensionality reduction; each of which make use of a number of important models.

Key Points

- Machine learning is more than just using data to make predictions, and is more than just statistics and mathematics
- We can use data to "train" statistical and algorithmic models that we can further use to make predictions
- Machine learning is the process of enabling computers (machines) "the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)
- Machine learning can be subdivided into supervised learning and unsupervised learning
- Supervised learning involves labeled data and predicting the label of new, unlabeled datum
- This course focuses on supervised learning - regression, k-nearest neighbors, perceptrons, and support vector machines
- Unsupervised learning involves unlabeled data and identifies interesting patterns and structures in data

Review: What is Machine Learning?

Let's validate our mental framework of machine learning by observing the following presentation that summarizes what machine learning is.



At this point in our Exploration, much of the content in this summary should feel familiar. If some concepts still feel a bit vague, you are in the right place! In this course, we will focus on the most widely-used aspects of machine learning: supervised learning, regression and classification. Together we will learn from first principles about how to apply machine learning to solve problems in your own work, and within our modern world.

Key Points

- Artificial intelligence (AI) is the field of enabling computers (machines) to mimic the decision-making capabilities of the human mind
- Machine learning is a subset of AI using algorithms that derive knowledge from data, to predict outcomes
- Machine learning can be subdivided into supervised learning and unsupervised learning (and, semi-supervised learning, deep learning, and reinforcement learning)

Check Your Understanding

Before you continue, please respond to the following questions and prompts to verify your understanding. Click any ▶ disclosure arrow to view a possible answer. ***[Accessible version for screen-reading software] (https://canvas.oregonstate.edu/courses/2025514/pages/exploration-questions-and-answers)***

1. ▶ Consider our coat recommendation system example. What kind of machine learning technique does this example seem to resemble the most?
2. ▶ Why is regression a form of supervised learning?
3. ▶ What is the main difference between unsupervised and supervised learning?
4. ▶ True or False: This course focuses on unsupervised learning
5. ▶ What machine learning models will we learn about in this course?
6. Define the following terms.
    1. ▶ Machine learning
    2. ▶ Training data
    3. ▶ Model
    4. ▶ Label

Additional Resources

Consider these curated resources an *essential*, if not mandatory, starting point for a deeper investigation into the topics within this Exploration.

- Reading: **What is Machine Learning? (https://www.wolfram.com/language/introduction-machine-learning/what-is-machine-learning)** Introduction to Machine Learning by Etienne Bernard, free from Wolfram
- Reading: **Machine Learning Paradigms (https://www.wolfram.com/language/introduction-machine-learning/machine-learning-paradigms)** Introduction to Machine Learning by Etienne Bernard, free from Wolfram
- Reading: **SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning (https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/)** by Isha Salian, Nvidia
- Reading: **What is Machine Learning? The Ultimate Beginner's Guide (https://www.v7labs.com/blog/machine-learning-guide)** by Pragati Baheti, Microsoft
- Video Lecture: **Stanford CS229 Machine Learning Course, Lecture 1 by Dr. Andrew Ng (https://www.youtube.com/watch?v=jGwO_UgTS7I&t=2173s)** (36:15 - end)
- Video Lecture: **MIT 6.0002 Introduction to Computational Thinking & Data Science, Lecture 11 Introduction to Machine Learning by John Guttag (https://www.youtube.com/watch?v=h0e2HAPTGF4&ab_channel=MITOpenCourseWare)**
- Tutorial: **Intro to Machine Learning (https://www.kaggle.com/learn/intro-to-machine-learning)** by Kaggle