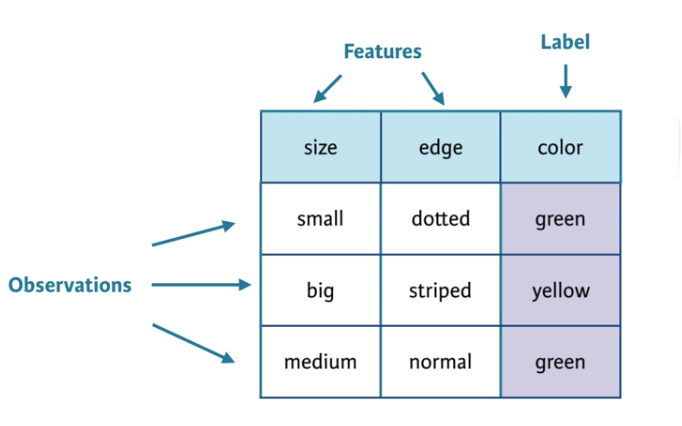
Introduction to Machine Learning

**Course Goals**

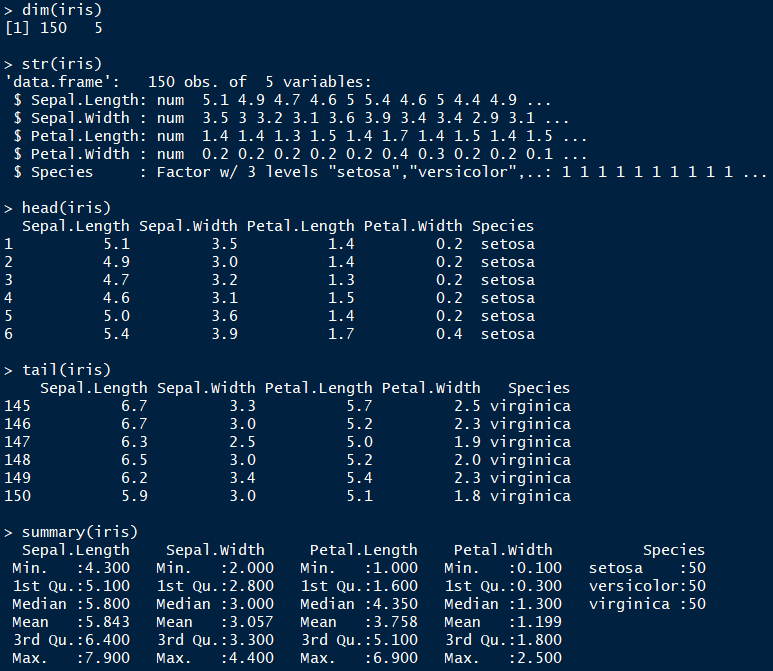
* Identify a machine learning problem
* Use basic machine learning techniques
* Think about your data and results

**What is machine learning?**

* Construct and use algorithms that learn from data
* A machine improves its performance when it receives more information
* Information or experience is gained from observing previous solutions

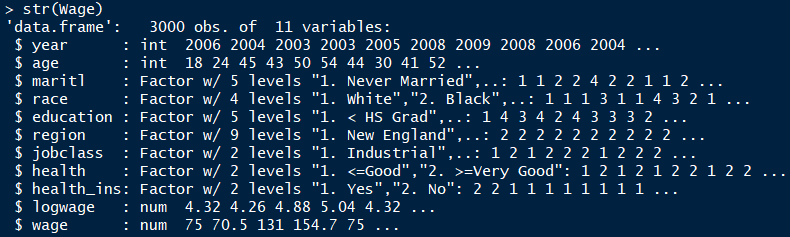
The dataframe to the left is an example of a classification problem. When a new observation comes in with a set of given features (input), the machine learning problem, in this case, would be to predict the output (label) by pushing the input through an estimated function.

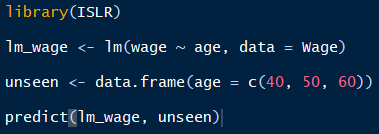
**Understand your data…**



**A Basic Prediction Model**

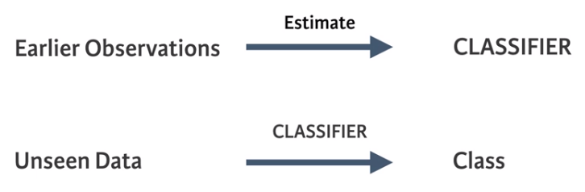
The Wage dataset is in the ISLR package. Linear regression can be viewed as a machine learning algorithm…



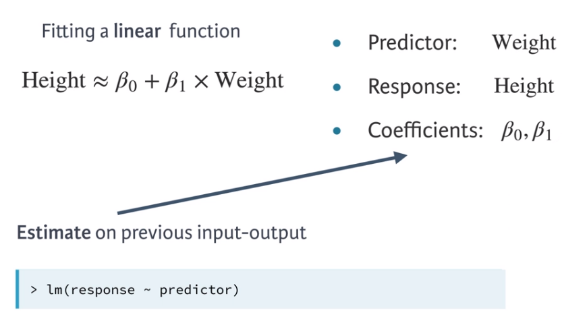


**Classification, Regression, and Clustering**

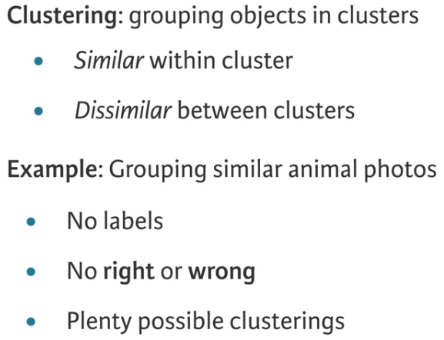
**Classification** The goal is to predict category (class) of new observation. Outputs are qualitative and can be put into predefined classes.

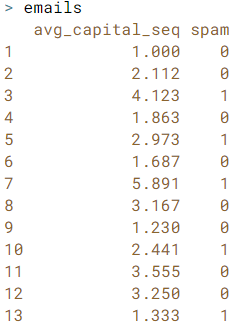


**Regression**



Outputs are quantitative and the regression function  
is estimated using previous input-output observations.

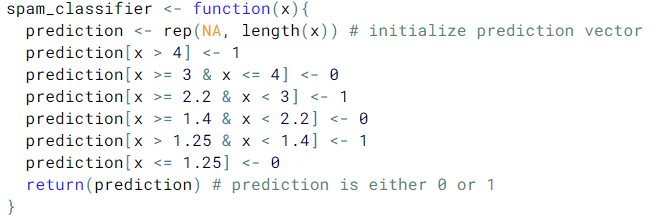
**Clustering** Clustering groups observations. It does this without any prior knowledge of what these groups could or should look like. Hence, prior knowledge and unseen observations are less meaningful than for classification and regression.

**Classification Example**

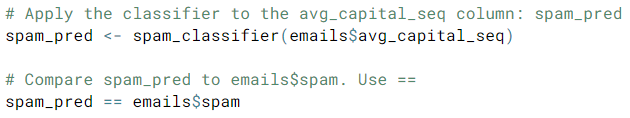
The emails dataset represents the observed data…

**avg\_capital\_seq** is the feature

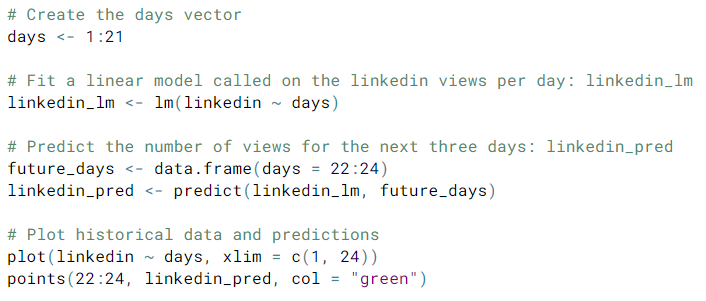
**spam** is the human labeled class indication if the email  
was actually spam (1) or not spam (0)



spam\_classifier is the estimated function that takes the feature as the input and outputs the predicted class

emails$avg\_capital\_seq is put through spam\_classifier and the output class predictions per observation are stored in the spam\_pred vector

**Regression Example**

**days** is the independent variable

**linkedin** is the dependent variable

linkedin\_lm has the linear regression coefficients stored in it and is used to generate predictions via the predict() function on the future\_days input vector

plot(linkedin ~ days, …) same as plot(x = days, y = linkedin, …)