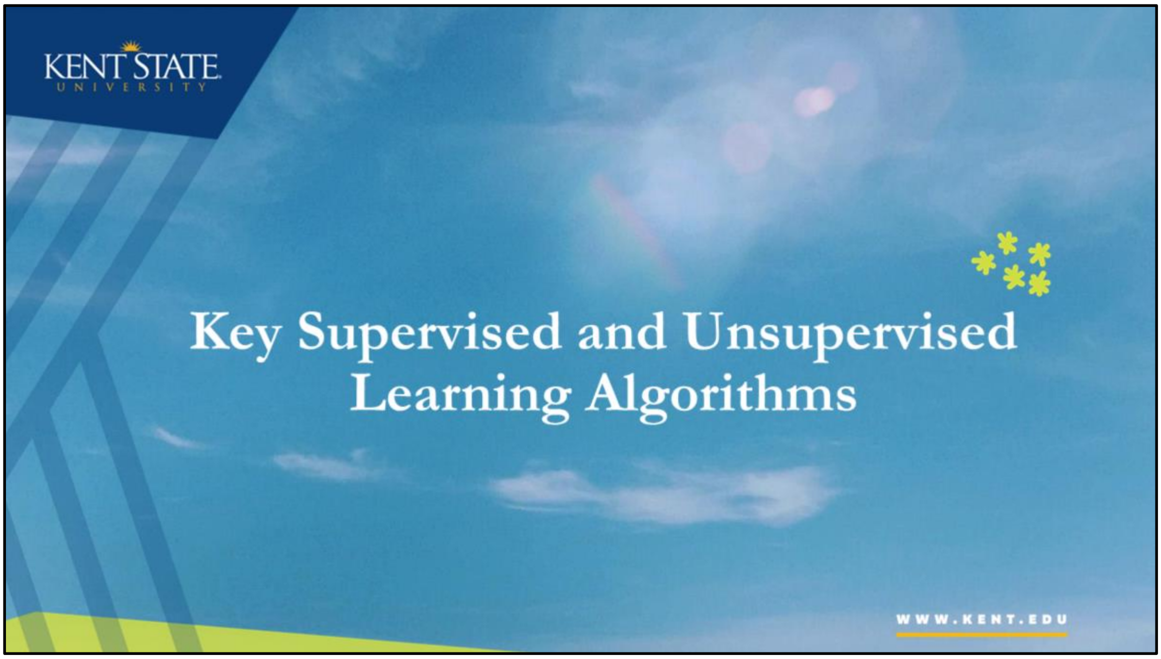
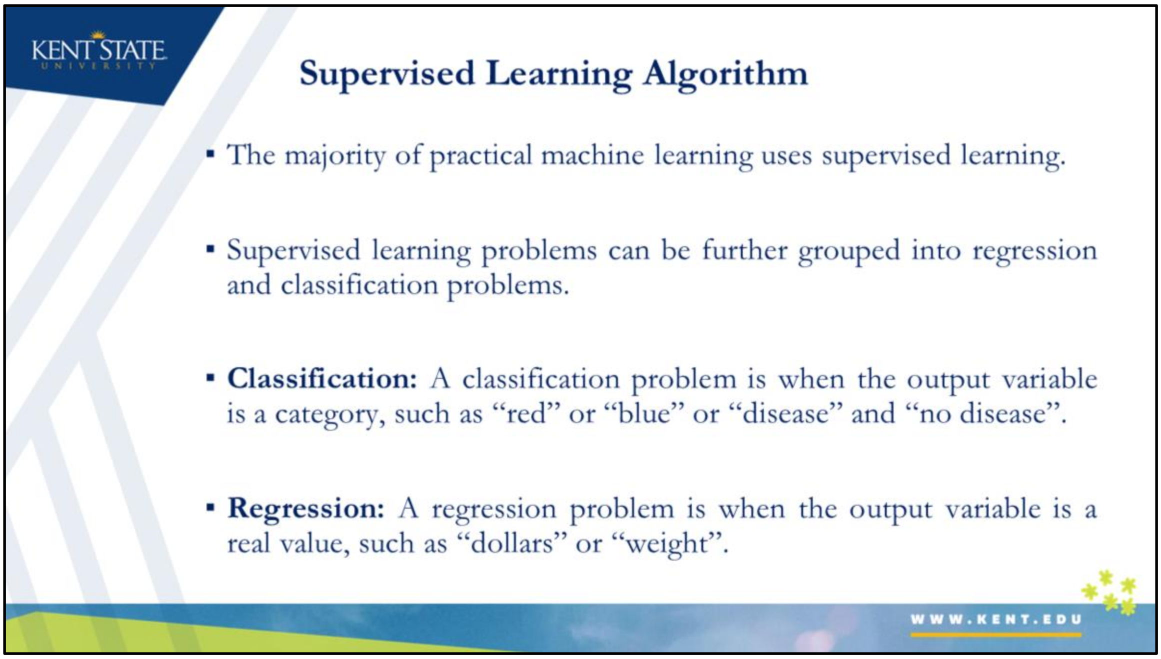


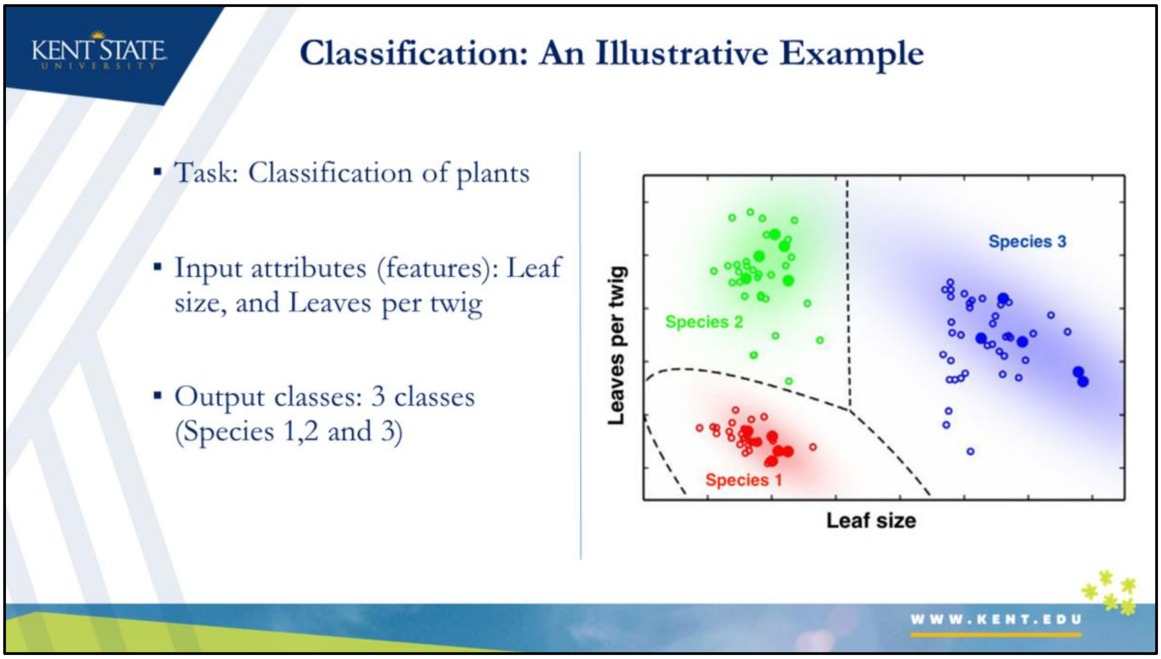
In this lecture, we briefly discuss some important supervised and unsupervised learning algorithms.



As we mentioned earlier, supervised algorithms comprise the bulk of applications currently. There are two main classes of supervised learning problems: Classification, and Regression.

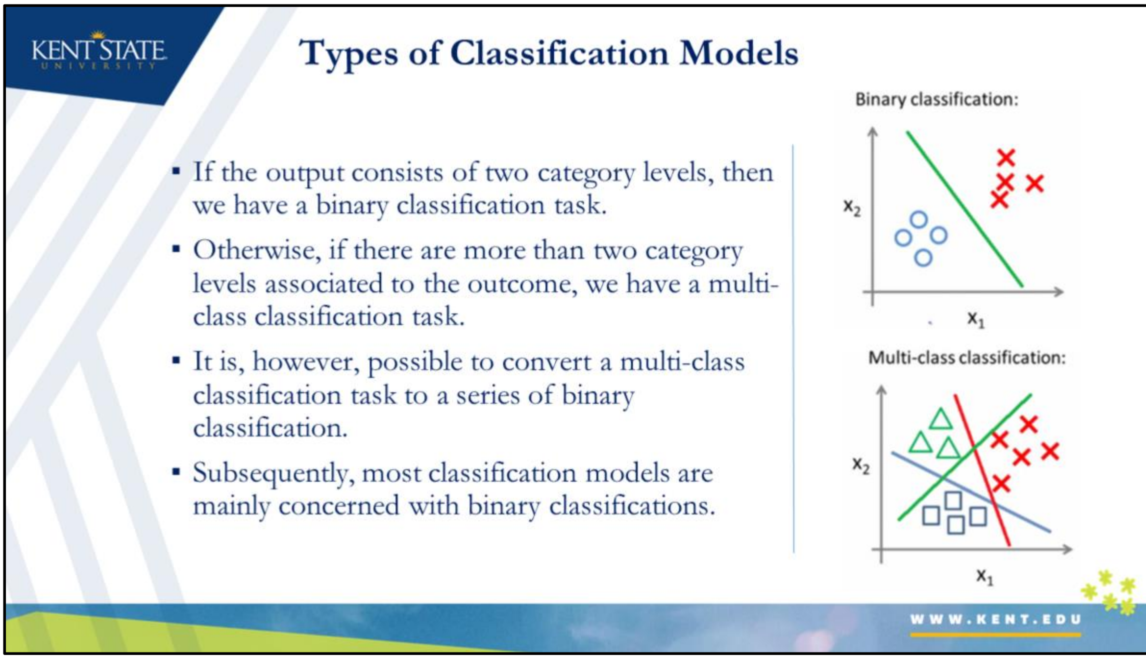
In classification, the objective is to group the output into a category. The category can be represented by a non-numeric values, for example, malignant, or benign, or can be represented by a numeric value, say 0 or 1. In either case, the output variable is a category. While many classification problems are based on binary classification, that is, the output only has two values, multi-classification problems are also quite common. In later classes, we will see how a multi-class problem can be represented and solved.

In Regression, the output variable is usually quantitative and continuous. For example, annual salary. The objective is to formulate and predict the annual salary based on a number of factors, such as, education, race, sex, etc. The output variable here, annual salary, is usually called as the response or dependent variable, and the other variables, education, etc., are called as predictors or independent variables. In notation, we represent the response by Y, and the predictors by X.

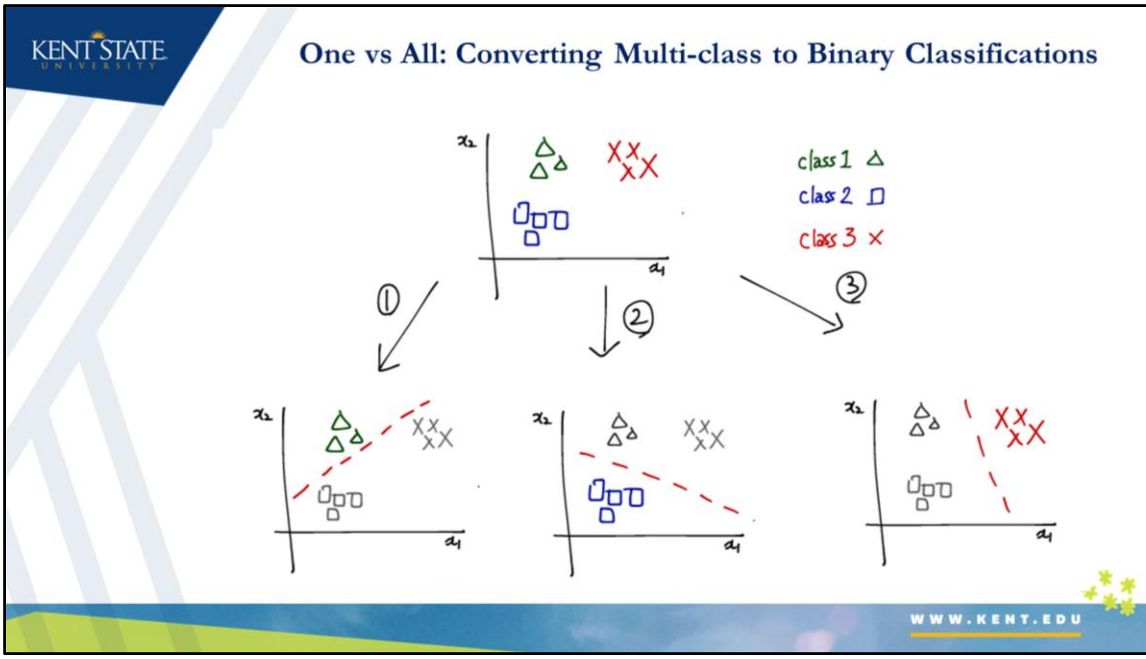


Let’s consider an example of a classification problem. In this example, we are trying to classify a leaf into one of three classes. This is a multi-class (>2) classification problem. We use leaf size and leaves per twig as the attributes to determine the class. These attributes are the predictor or independent variables.

An important question to answer here is “how do you measure the performance of the classifier?” Would it simply be the leaves correctly classified? As you will see in later classes, the choice of objective (performance) is vital to both the accuracy and efficiency of the classifier. For most classification problems, we can actually output multiple measures of performance, but this is something to which you should pay attention.



While there are several multi-class ML algorithms that will directly predict the probability of belonging to a class when we have more than two classes, we could also treat a multi-class classification problem as a series of binary classification tasks. For example, consider we have a classification problem with three classes, A, B, and C. As a first step, we create a binary classification with classes A, and Not A (B, C). Next, a binary classification task with B, and Not B (A, and C). Finally, C and Not C (A, B). The next slide provides a more visual representation of this explanation.



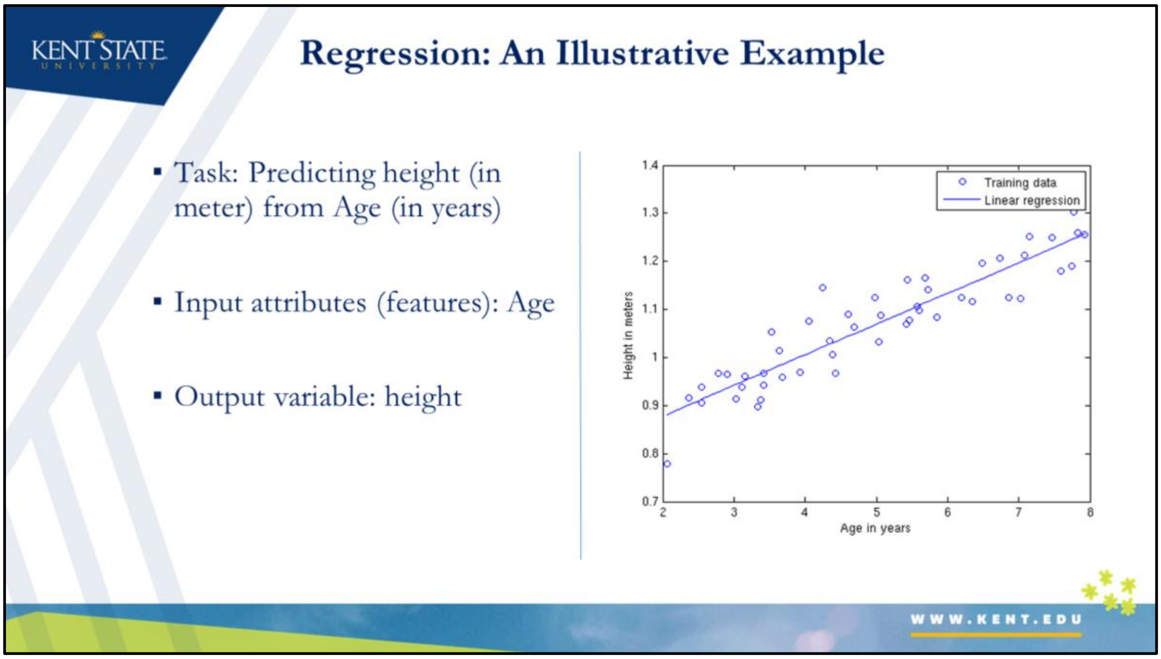
There are several points on which you should ponder.

1. For an initial problem with N classes, how many such binary classification problems will you solve?
2. How will you define performance?
3. How will you combine the results of these binary classification problems?
4. What would be the time complexity to solve such a problem? For those of you with background in CS, what is the O()?

The above approach, One-Vs-All (OVA) is just one way to treat a multi-class problem. One could, on the other hand, also use a One-Vs-One (OVO) approach to multi-class problem. Here, the classifier will create one classifier per pair of classes. For N classes, this translates to how many classifiers?

Ans: N * (N-1) / 2.

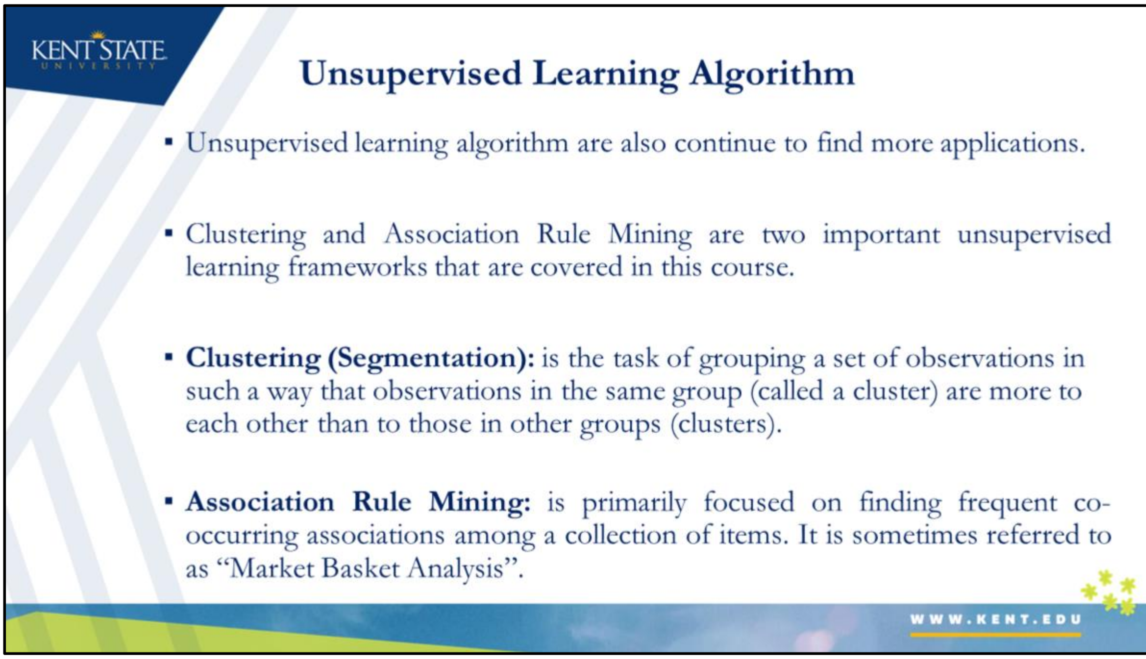
For more discussion on this topic, here is one post https://medium.com/value-stream-design/machine-learning-reductions-mother-algorithms-part-ii-multiclass-to-binary-classification-1dad599147b



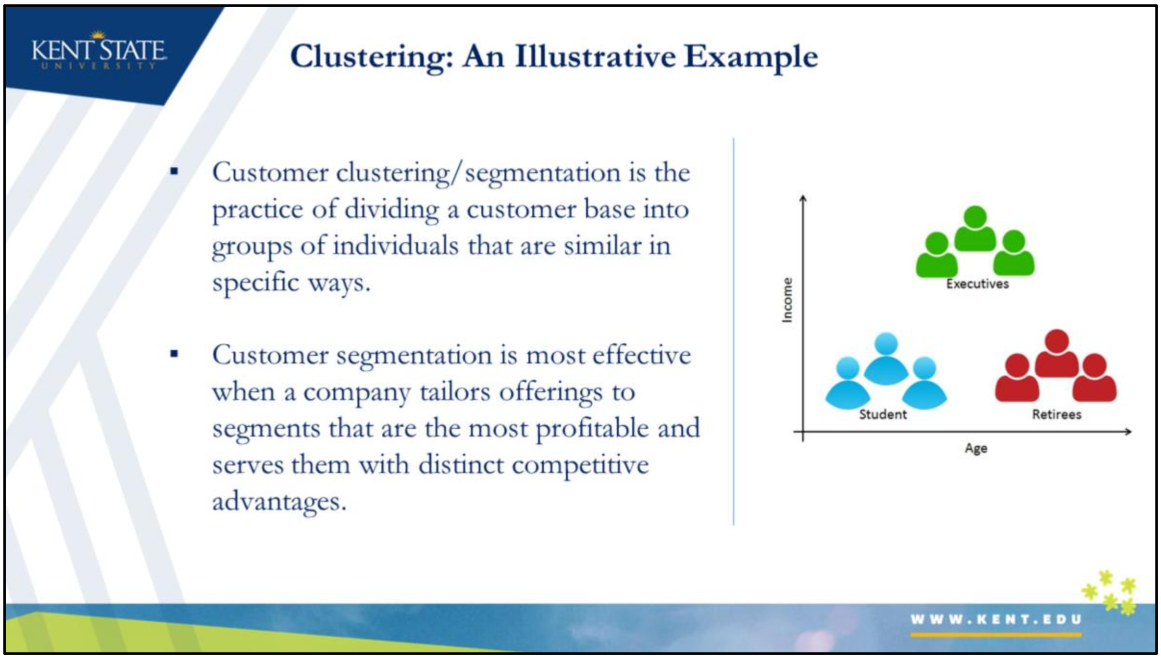
In regression, the response variable is usually continuous (quantitative and real, rather than categorical). The objective is still the same. That is, to predict the value of the response, for our example, Height, from data given to independent, or predictor variables. In our example, Age. Note in the graph that we see that as Age increases, Height also tends to increase. Also, the line there represents what is called as the least-squares line, or regression line. Our best guess at the relationship between Age and Height. There are a few points that you should consider further:

1. Clearly, the regression line does not pass through all the points. What is the line then predicting?
2. How do you measure the performance of our line? Is there another line that might give better performance?
3. How is the line constructed?

These are all questions that we will answer in later lectures.

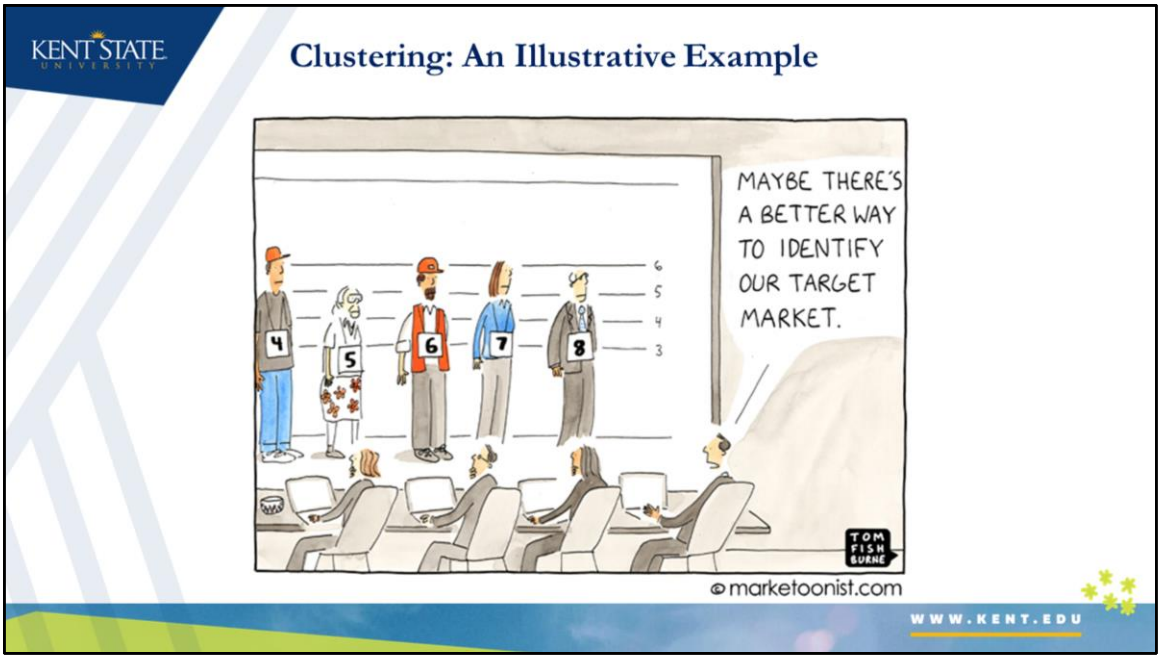


Let us now briefly consider unsupervised algorithms. Clustering and dimensionality reduction, e.g., PCA, are the most common unsupervised algorithms. Here, we will cover clustering and Association Rule mining.

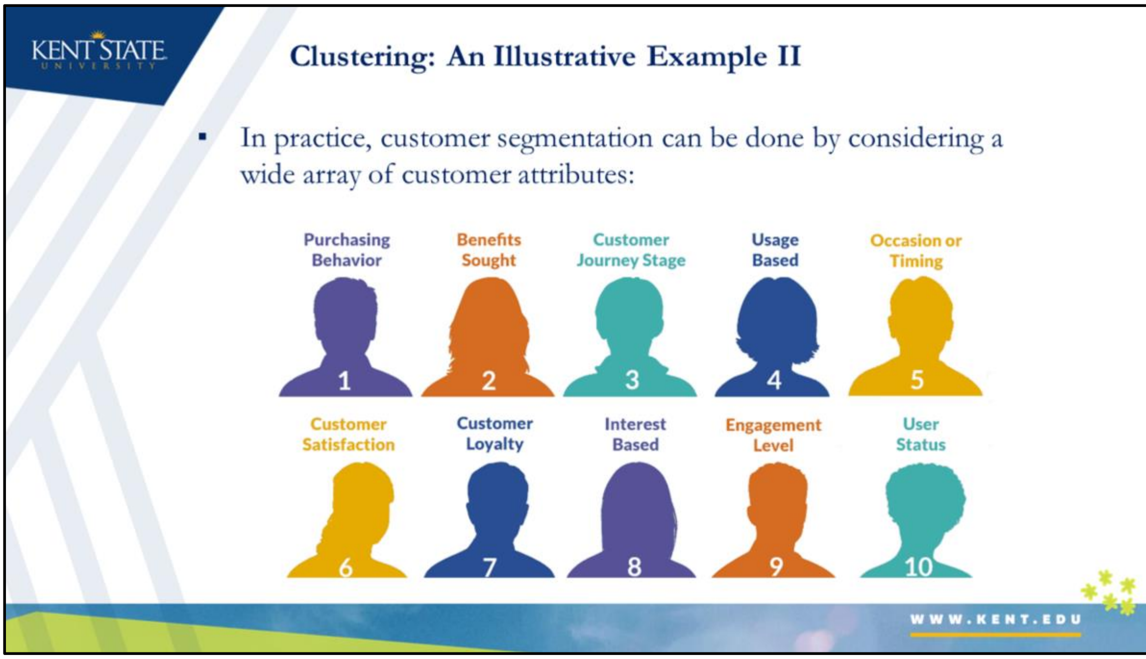


The idea of clustering is very common, something that we do frequently. In a formal setting, clustering or segmentation is a way of dividing data into groups where there is homogeneity within a group. As in the example illustrated above, people are grouped based on their status. Within a group, there is homogeneity according to the attribute of grouping.

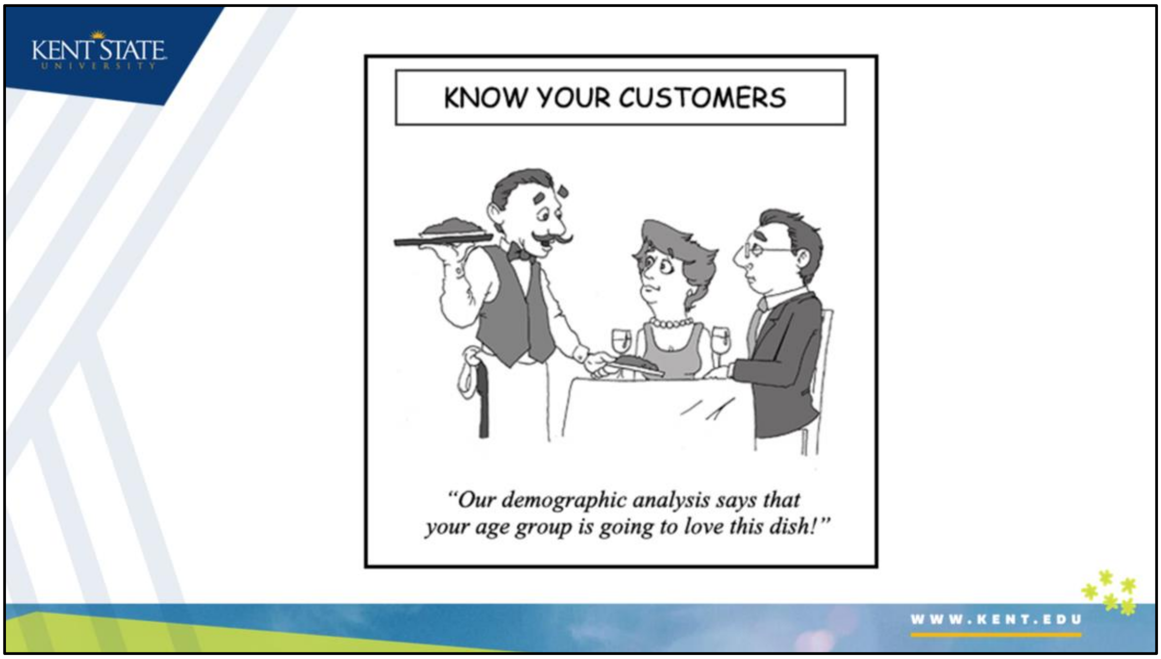
For a company, this provides an effective way of grouping customers on different attributes, thus allowing them to target these groups differently.



Have you ever noticed how online ads are targeted? That is one form of clustering.



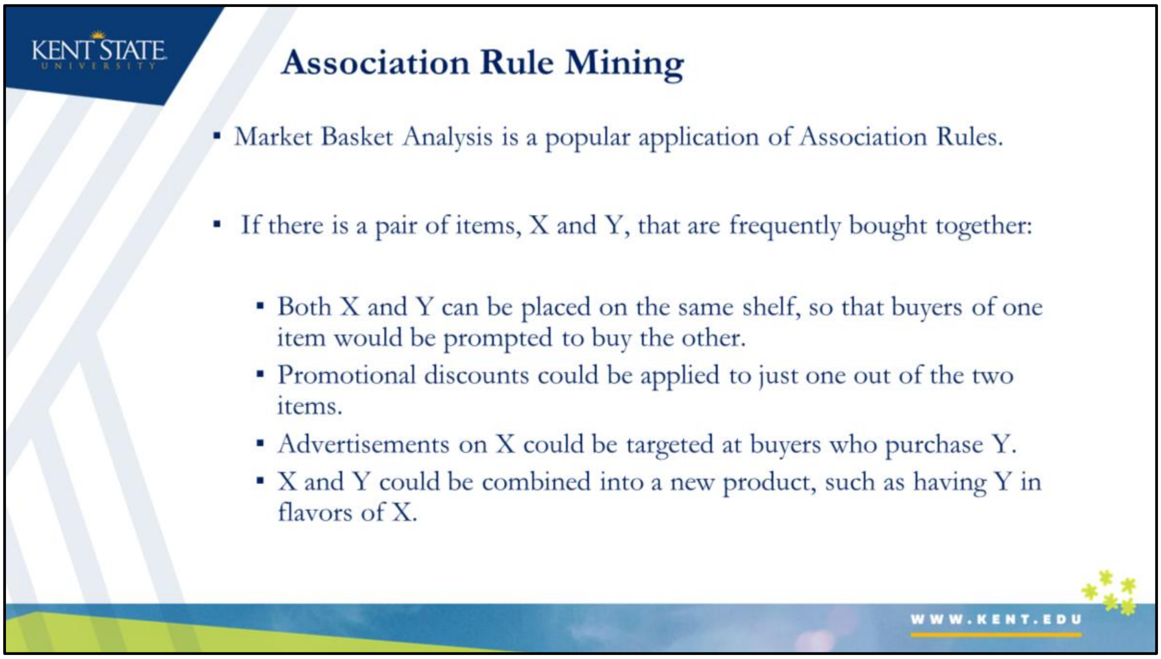
Clustering, and by definition, customer segmentation is usually done using multiple attributes. This allows companies to tailor policies for segments that are meaningful to them.



Of course, this can lead to disastrous results.

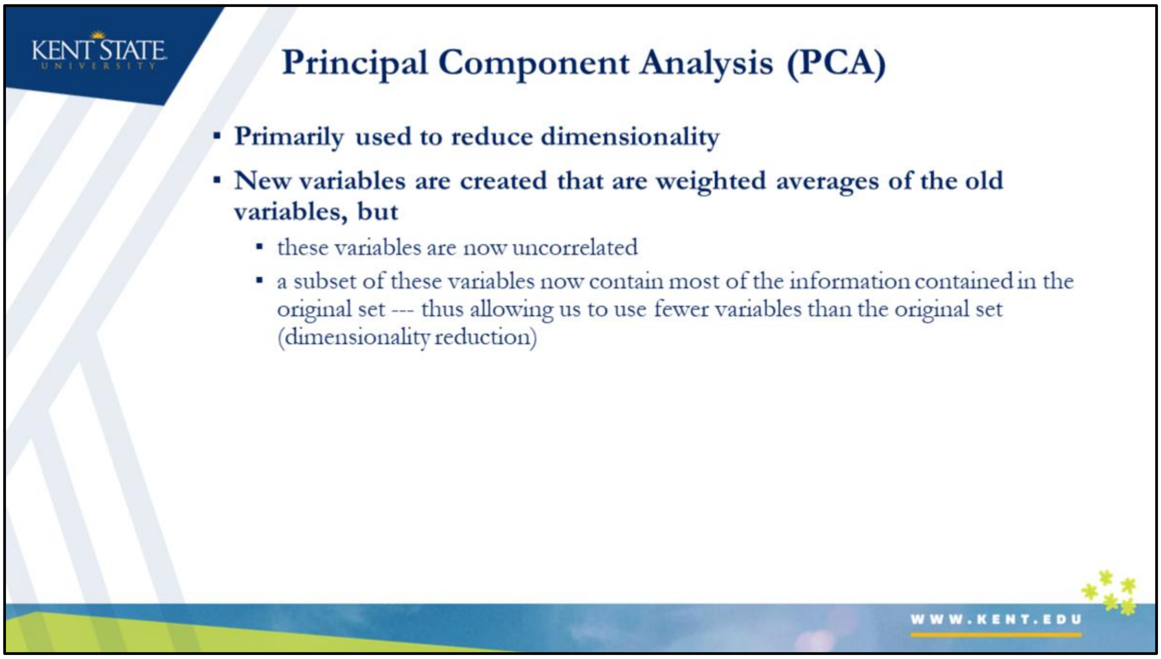
Association rule mining is now very common. Most online shopping and many recommender systems use this to provide suggestions. One of the earliest examples of this was the case where a supermarket noticed an increase in beer sales when diaper sales went up. It turned out, that men, when they came to shop for baby diapers were also buying beer. So the supermarket chain moved beer displays next to diapers, which resulted in further sales. Notice the commonality here. It was on the subject, rather than on the product.

Association rules are much more sophisticated and targeted now, as anybody who has shopped online can attest.



Market basket analysis is an application of association rules. It is an old idea, but enhanced with the availability of not just large amounts of data, but specific enough to target individual customers.

Think about it: Can you find examples to illustrate each of the above scenarios for MBA?



PCA can be thought of a way of dimensionality reduction. This is an important application of unsupervised learning.

Here, a new set of variables is created from the original set of variables, which may be quite large. While the total number of new variables created will always be equal to the old set, these new variables have some properties that make it more useful:

1. These variables are uncorrelated. For practical purposes that means the effect of one variable can be determined independently of the other variable
2. A small subset of the new variables contain most of the information of the old set. This allows dimensionality reduction, as we can then use the small subset and retain most of the information, rather than use the large original set of variables.

PCA can be usually be used independently of other techniques. For example, PCA can be used before regression is applied, and then a regression model can be created based on the variables from PCA.