Machine learning

Presented by : Dr. Hanaa Bayomi

Lecture 5 : KNN K-nearest neighbor

What is KNN

1. A very simple classification and regression algorithm
   a. in case of classification, new data point get classified in a particular class
   b. in case of regression, new data point get labeled based on the AVR.
   Value of KNN
2. It is a lazy learner because it doesn't learn much from the training data (most of learning happens from a live data)
3. It is a supervised learning algorithm
4. Default method is Euclidean distance
5. Non- parametric method used for classification

Eager vs lazy learner

Eager learner	Lazy learner
The distinction between easy learners and lazy learners is based on when the algorithm abstracts from the data.	
1. When it receive data set it starts classifying (learning) 2. Then it does not wait for test data to learn 3. So it takes long time learning and less time classifying data	1. Just store Data set **without** learning from it 2. Start classifying data when it receive **Test data** 3. So it takes less time learning and more time classifying data
Ex. Linear Regression, Decision tree	Ex. K. Nearest Neighbors

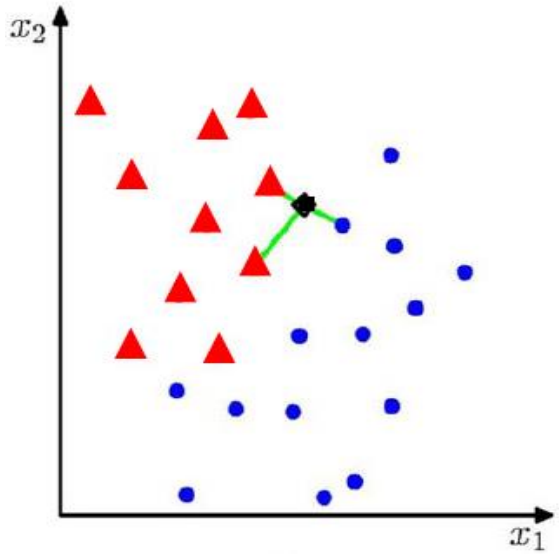
K-NN algorithm

- 1 NN
  - Predict the same value/class as the nearest instance in the training set
- k NN
  - find the k closest training points (small $\|x_i - x_0\|$ according to some metric, for ex. euclidean, manhattan, etc.)
  - predicted class: majority vote

Distance functions

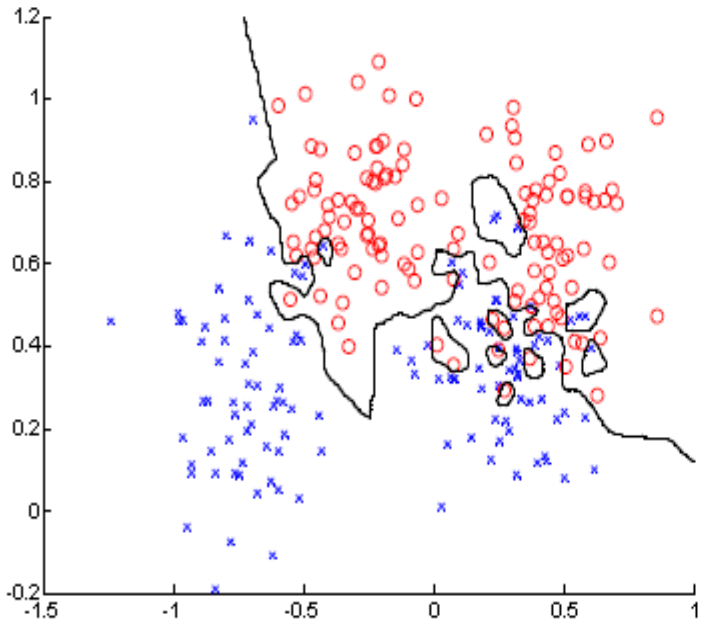
Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$



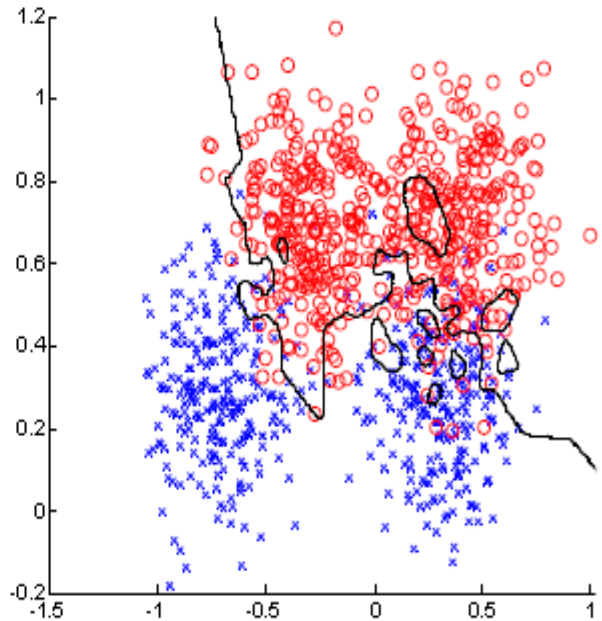
K = 1

Training data



error = 0.0

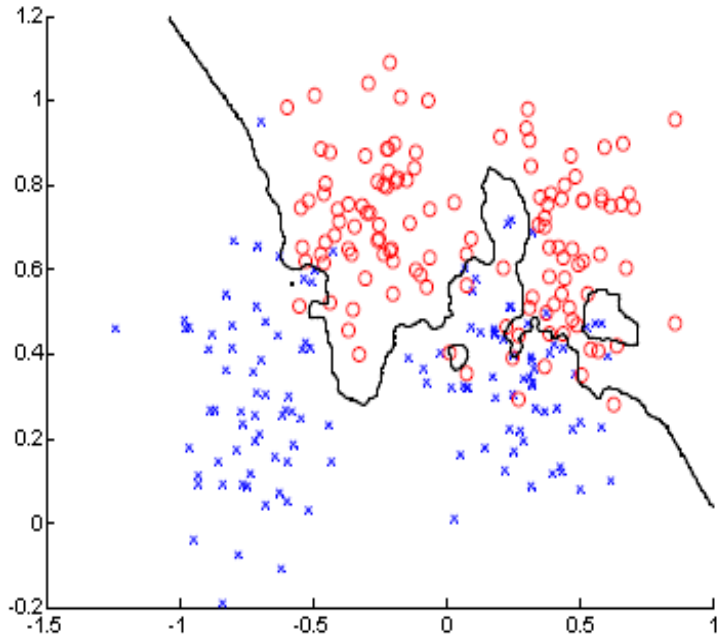
Testing data



error = 0.15

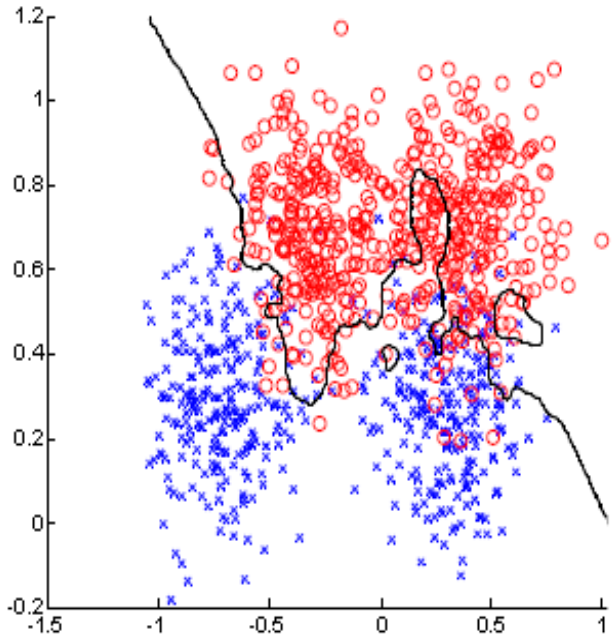
K = 3

Training data



error = 0.0760

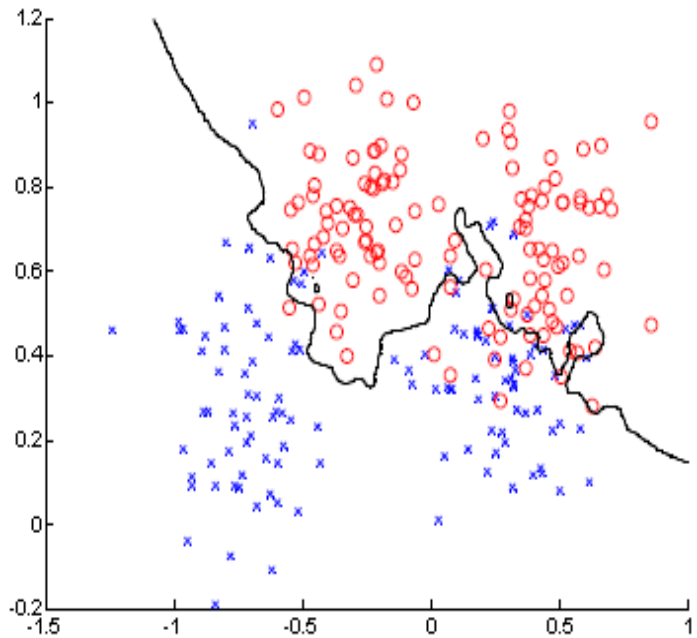
Testing data



error = 0.1340

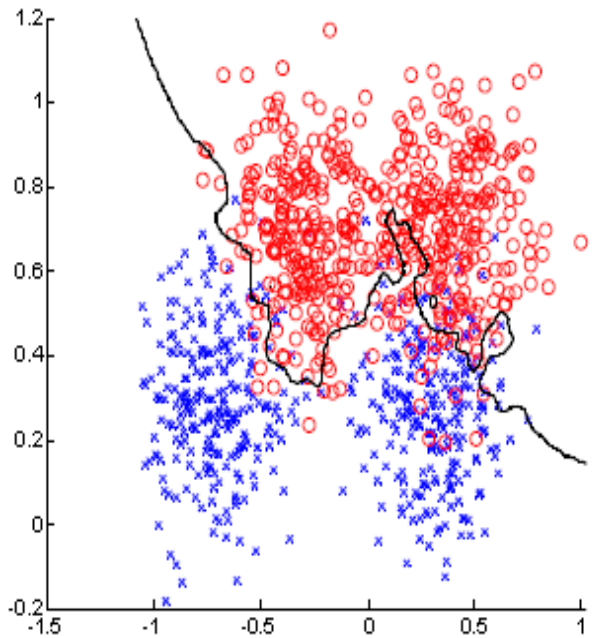
K = 7

Training data



error = 0.1320

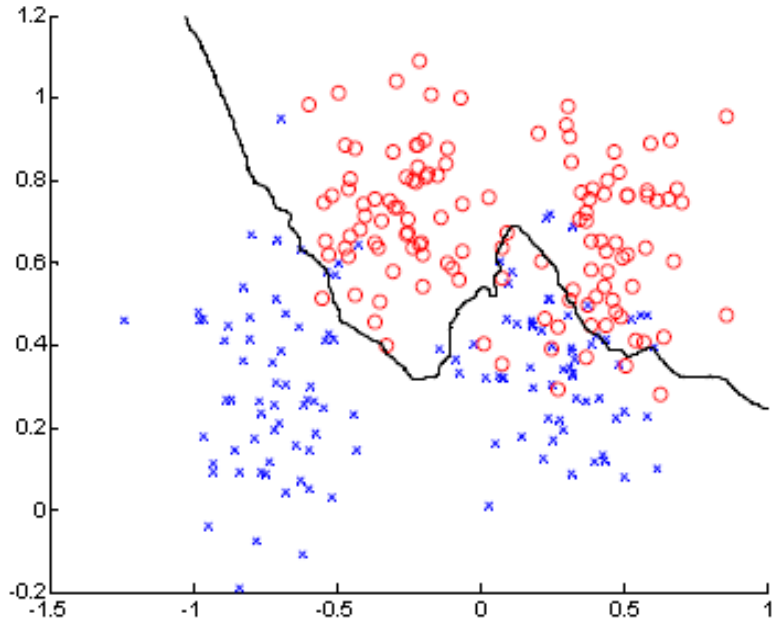
Testing data



error = 0.1110

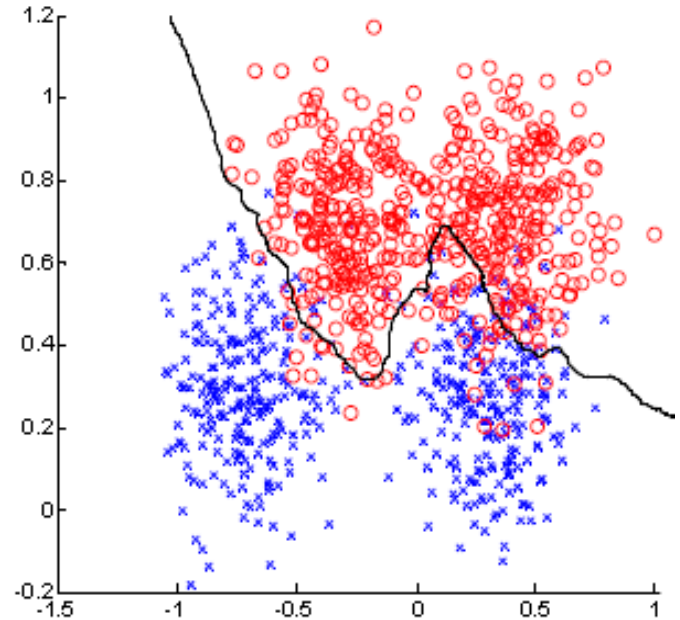
K = 21

Training data



error = 0.1120

Testing data



error = 0.0920

Example

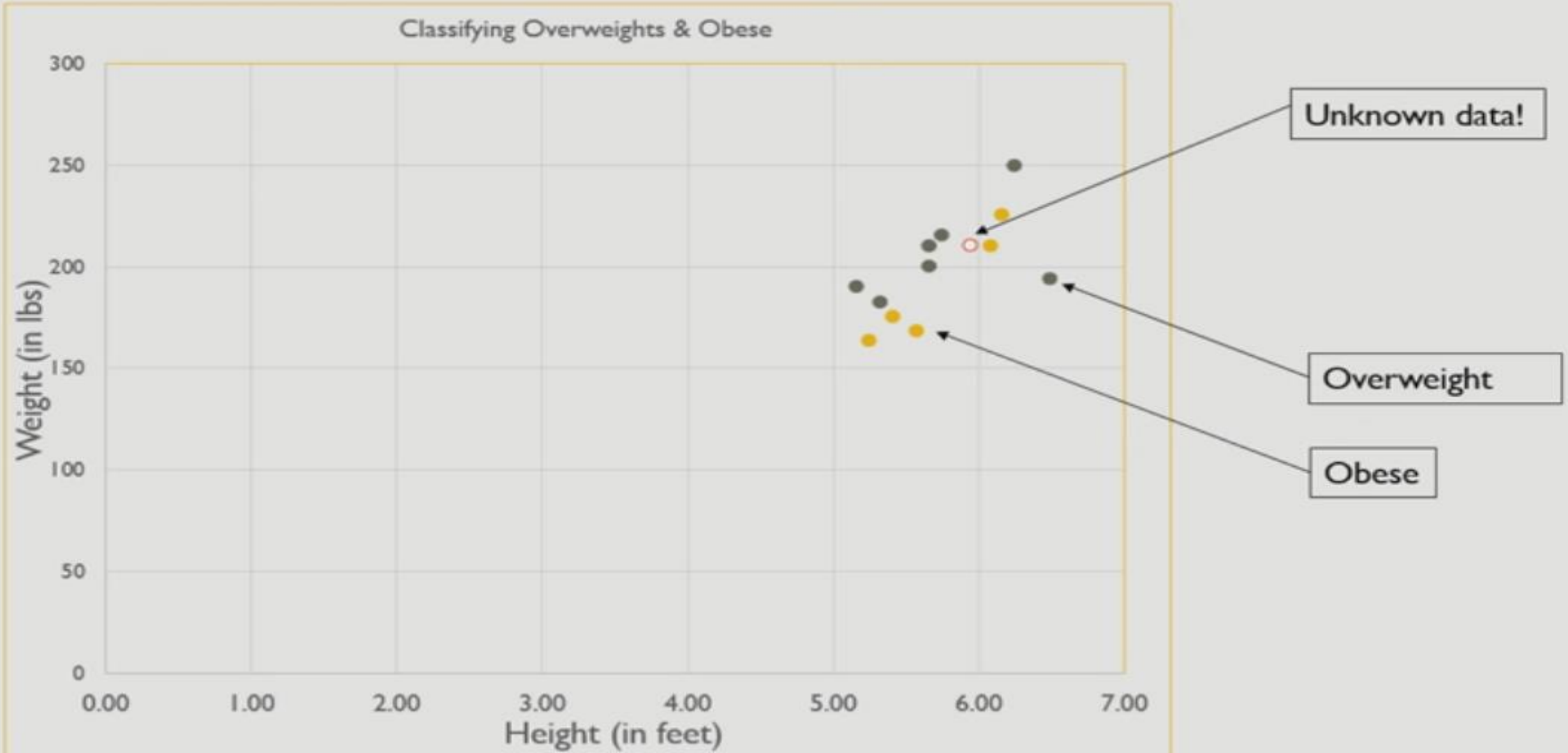
Height (feet)	Weight (pound)	Obesity
5.33	182	Obese
5.17	190	Obese
6.50	193	Overweight
5.67	210	Obese
6.17	225	Overweight
5.58	168	Overweight
5.75	215	Obese
6.25	249	Obese
6.08	210	Overweight
5.25	163	Overweight
5.42	175	Overweight
5.67	200	Obese

6 obese
&
6 overweight

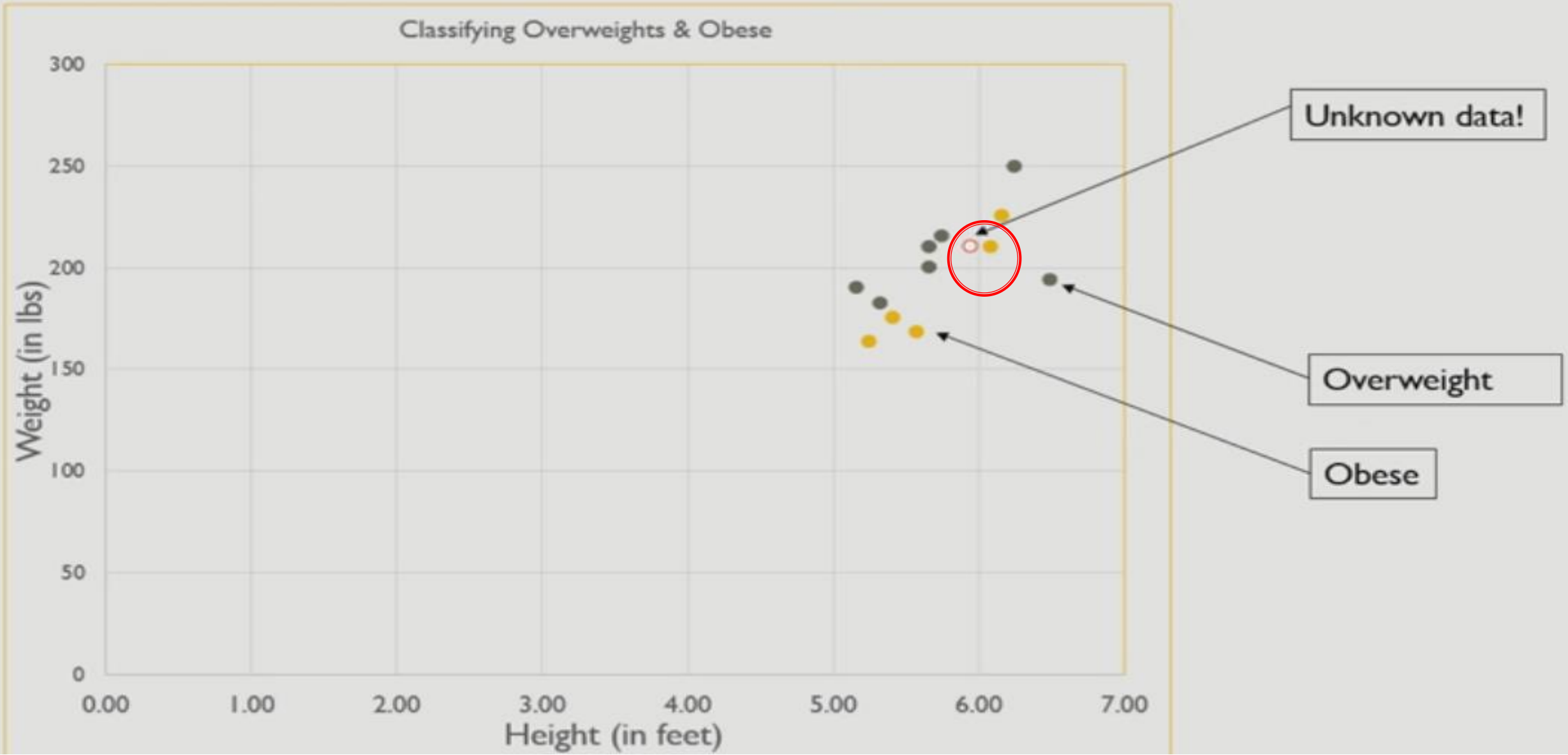
New Data

5.95	210	?

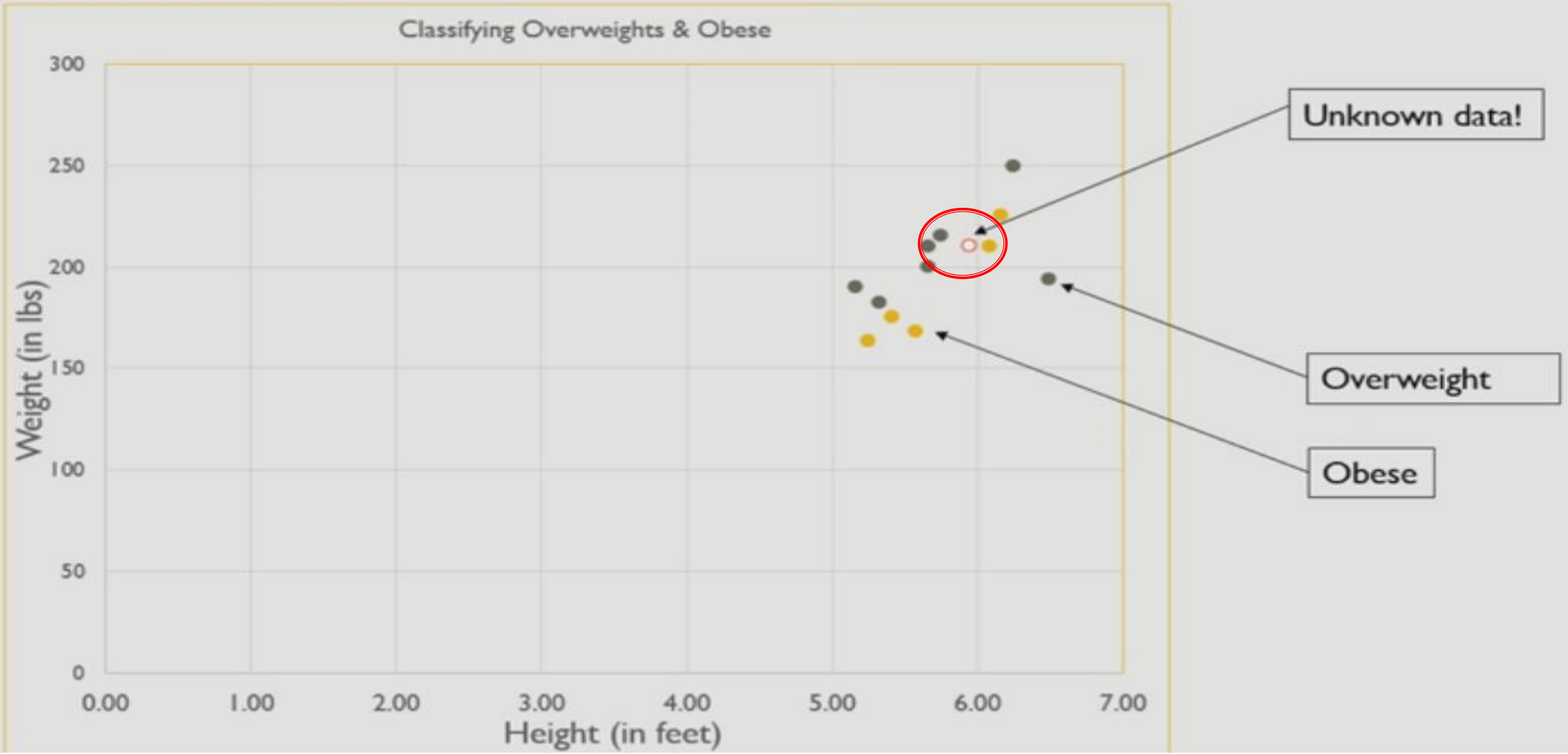
VISUALIZING THE DATA



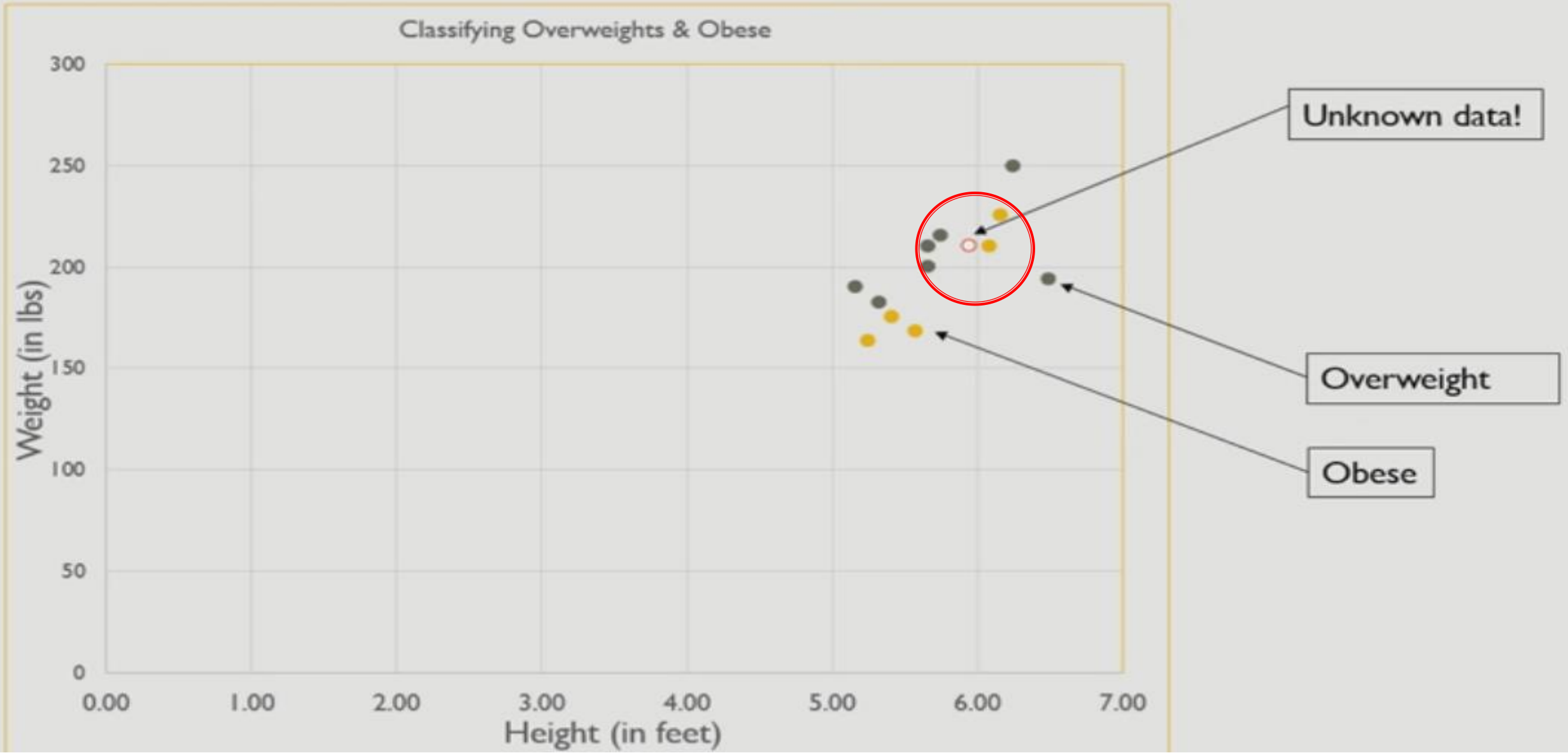
VISUALIZING THE DATA



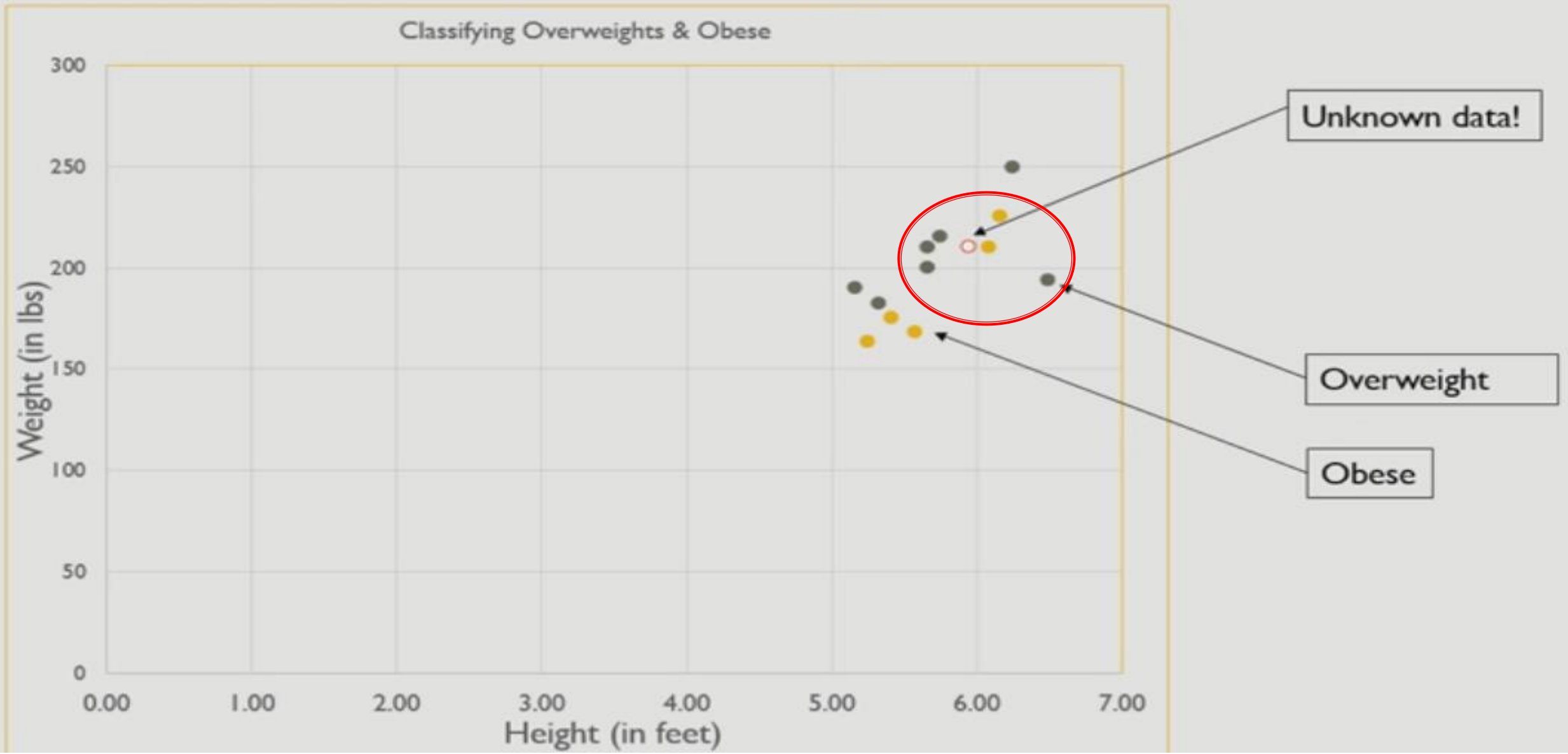
VISUALIZING THE DATA



VISUALIZING THE DATA



VISUALIZING THE DATA

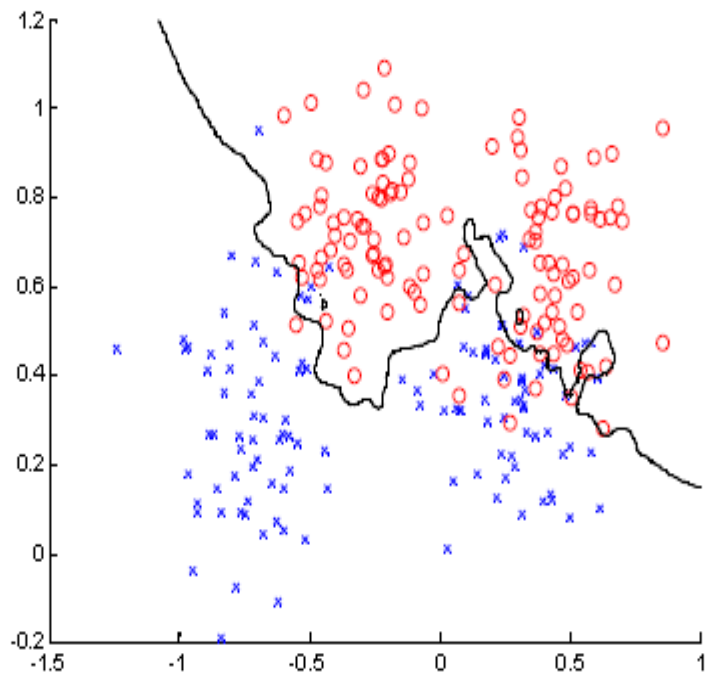


How to choose K

1. Choose an **odd K value** for the two clases
2. K must not be a ***multiple of the number of the classes***
3. If K too small , then the nearest neighbor classifier may be susceptible to **over fitting because of noise in training data**
4. If K too big , then the nearest neighbor classifier may mis-classify the test instance because its list of nearest neighbor may include data points that are located far away from the neighbor
5. Usually a value between 5-10 is taken as reasonable value of K.
6. Choose (learn) K by cross-validation
   - Split training data into training and validation
   - Hold out validation data and measure error on this

Advantages:

- K-NN is a simple but effective classification procedure
- Applies to multi-class classification
- Decision surfaces are non-linear
- Quality of predictions automatically improves with more training data
- Only a single parameter, K; easily tuned by cross-validation



Disadvantage

- - affected by local structure
- - sensitive to noise, irrelevant features
- - computationally expensive O(nd)
- - large memory requirements

ASSIGNMENT

WHAT IS THE DIFFERENCE BETWEEN?

bias and variance

Bias refers to the error that is introduced by approximating a real-life problem with a simplified model.

Variance refers to the error that is introduced by a model that is overly complex and fits the training data too closely