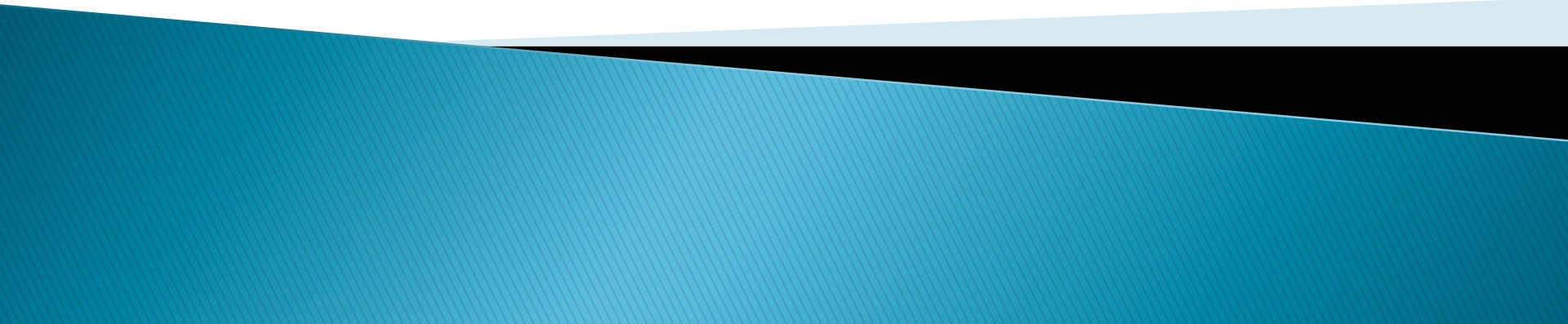
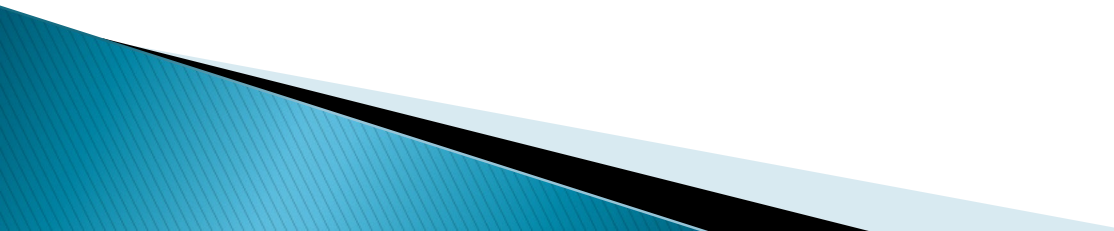


# Introduction to Clustering



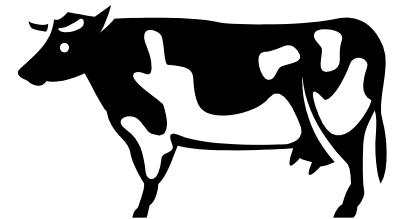
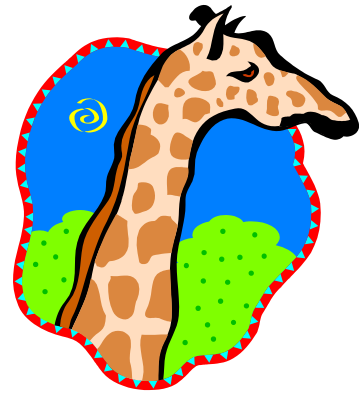
# What is clustering?

- ▶ Refers to methods that are used to look for groups of similar observations
  - ▶ Can also isolate “outliers”
  - ▶ Note that the user gets to determine what “similar” means and there are many options.
  - ▶ Much more to this than we see in just section 12.1.
- 

# Clustering Intuition

- ▶ How would you cluster the following 18 animals into 2 groups? Use your own personal criteria.


▶ Giraffe	Horse	Elephant
Sheep	Cow	Brown Bat
Human	Squirrel	Tiger
Chimp	Duck	Dog
Jaguar	Dolphin	Lion
Rat	Cat	Mouse



# Clustering Intuition II

- ▶ You might have clustered into farm vs. not farm or small vs. large.
- ▶ You could have clustered based on # hours sleep required in 24 hour day, if you had that information.
- ▶ Clearly, we have choices to make about what similarity/dissimilarity we use.
- ▶ Also have to pick # of clusters! Two was just for example.

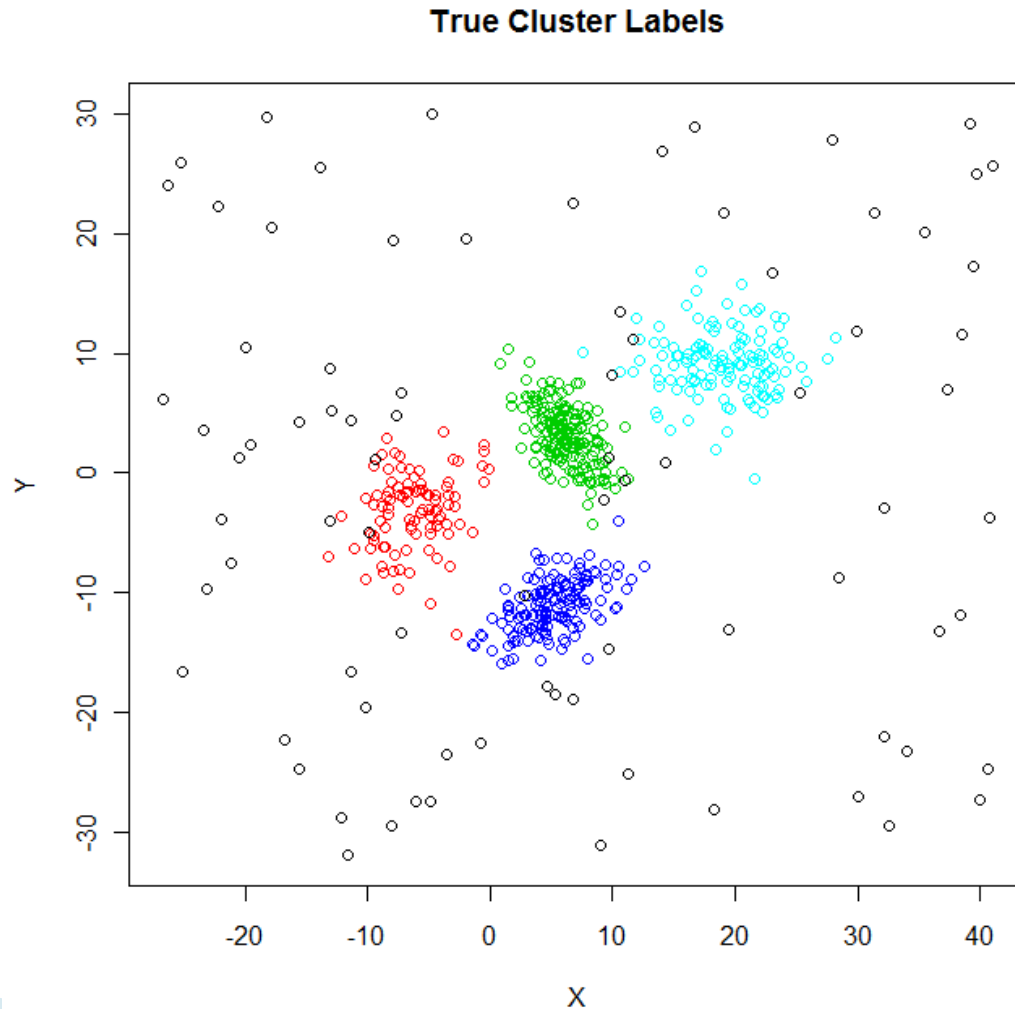
# Clustering Methods

- ▶ Need a distance (or similarity) measure to cluster.
  - ▶ Need to pick # of clusters, or various other input parameters
  - ▶ Variety of methods exist
    - Hierarchical – hclust/agnes/diana
    - Partitioning – kmeans
    - Graph-based methods
    - Density-based methods
    - Model-based methods
  - ▶ Need to consider scaling/standardization of variables input to distance measure.
- 

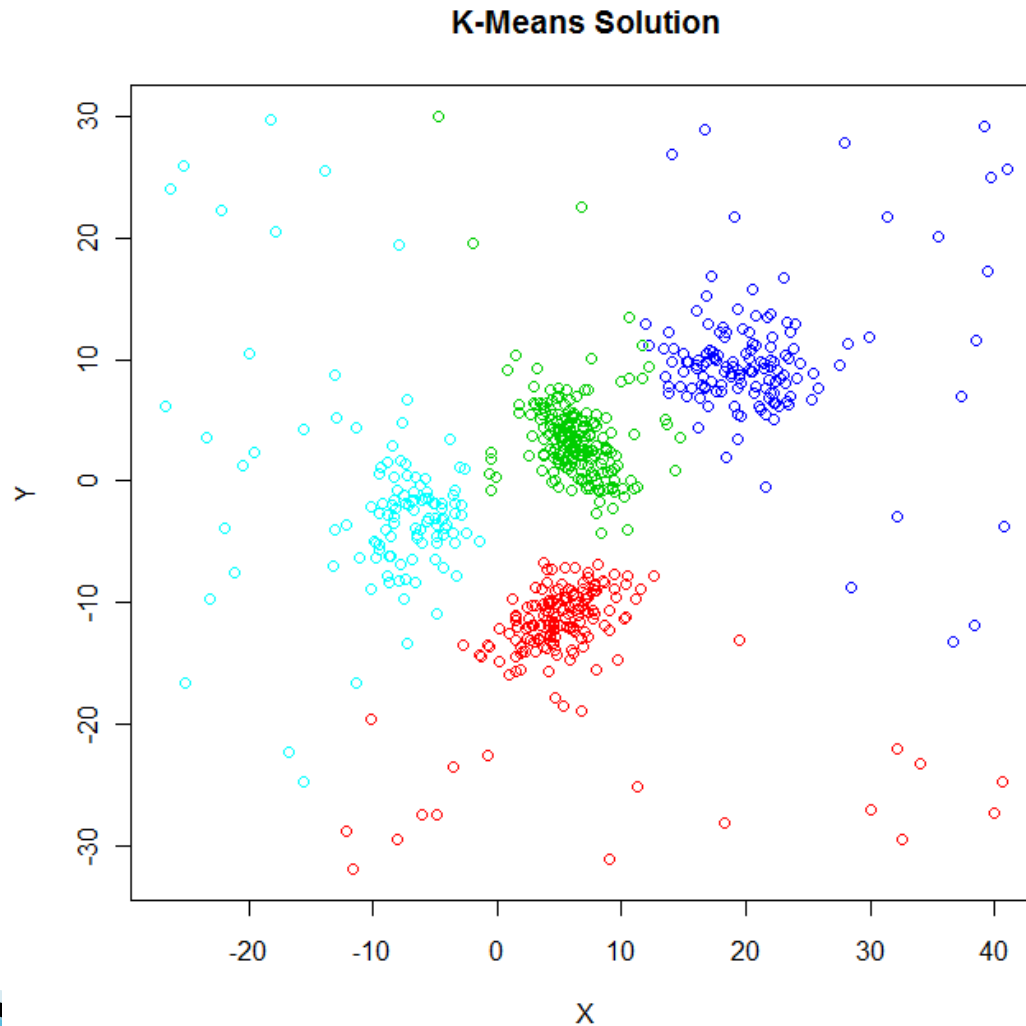
# Simple Example

- ▶ Let's consider an example where we have points in 2D (so we can visualize), and we just use Euclidean distance between them as their dissimilarity.
- ▶ We'll generate starting data that has clusters in it (neat R package), as well as some outliers, and I will show you the clustering results, according to some different methods.
- ▶ We will be “nice” here and tell the algorithms to look for the correct # of clusters, but really that's a problem all on its own.

# Simple Example – True Labels

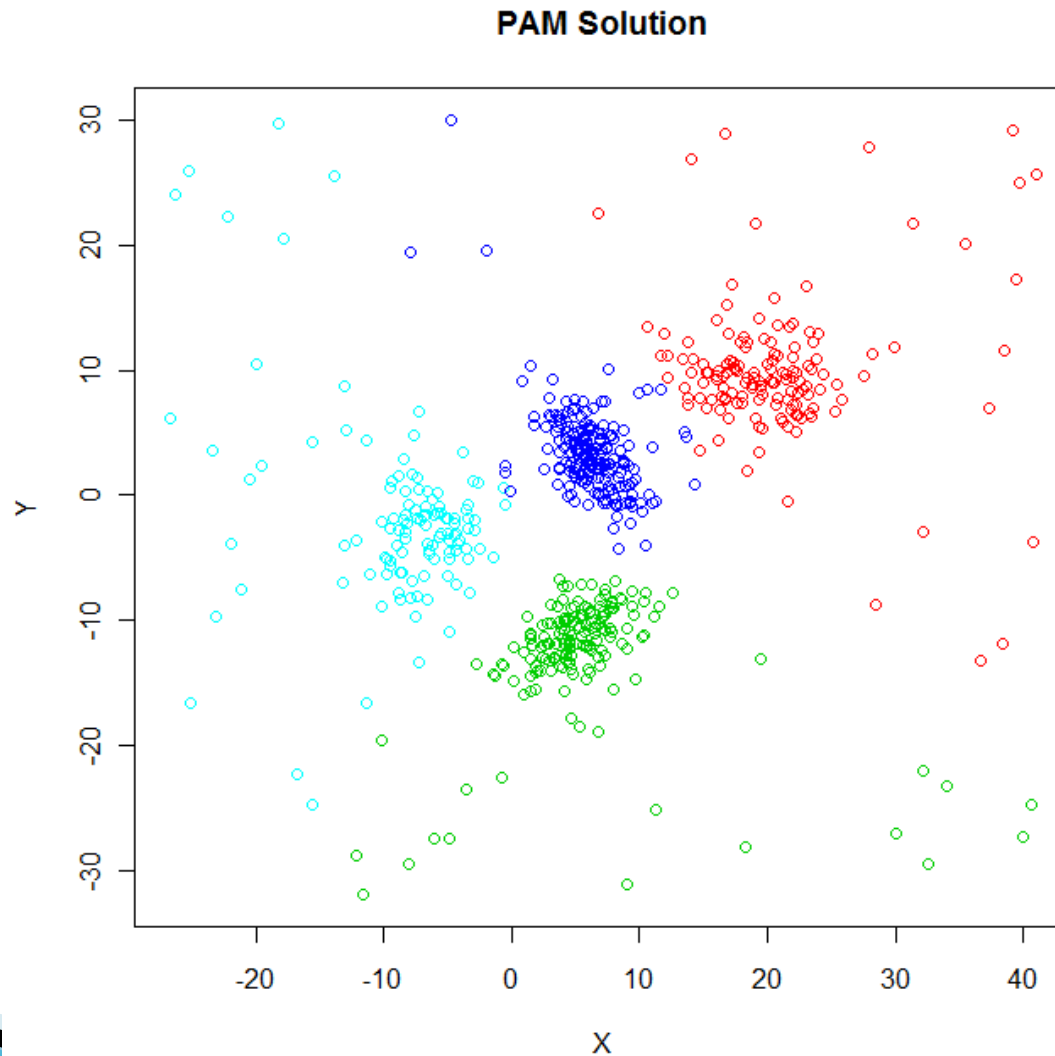


# Simple Example – K-means

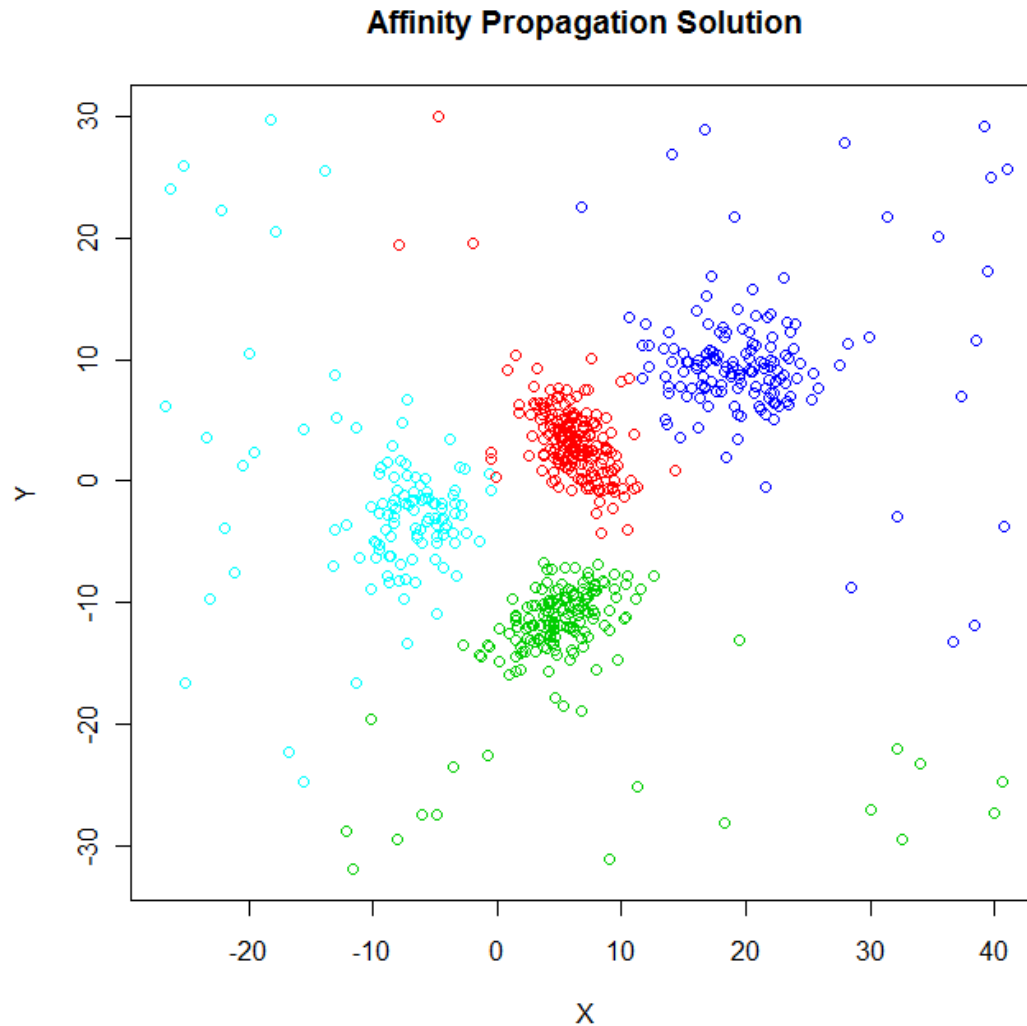




# Simple Example – PAM Solution



# Simple Example – Affinity Propagation



# Simple Example – Challenges

- ▶ This was only 2 variables, with both variables being important to see the clusters.
- ▶ Dealing with many variables, with some “noisy” variables is more challenging.
- ▶ Choosing # of clusters is challenging.
- ▶ The outliers are included in the clustering solutions.
- ▶ These clusters were fairly distinct, but that may not happen for your data.

# Clustering Methods

## Partitioning k-Means

- multi-cluster membership → Fuzzy k-Means
- speed, scalability → FFT
- categorical → COOLCAT
- outliers → k-Medoids
- scalability via data sampling → CLARA
- random sampling → CLARANS
- categorical → k-Modes
- speed, scalability → squeezer
- mixed datatypes → k-Prototypes

## Hierarchical

- linear algebra basis → Spectral
- both rows and columns → Biclustering
- speed, scalability → Birch
- arbitrary shapes → Cure
- speed, scalability → Sting
- categorical, interconnectivity of clusters → ROCK
- closeness of clusters → Chameleon
- speed, scalability → Limbo

## Density-based

## Density-based DBSCAN

- no input parameters → OPTICS
- speed, arbitrary shape(convex) → DENCLUE
- speed, insensitivity to order input → CLIQUE, WaveCluster
- clusters' relevant attribute sets → *Projected*: CACTUS, STIRR, CLICK, CLOPE
- categorical, minimum parameters, outliers, insensitivity to order input → HIERDENC
- speed → MULIC multi-layered

## Model-based

- arbitrary shapes → Self-org.maps
- classification tree:node is categor. concept → COBWEB
- mixed datatypes → BILCOM empir. Bayes
- prior distributions, no #cluster parameter → AutoClass
- minimum parameters, no local minima → SVM clustering

## Graph-based

- highly connected subgraphs → MCODE
- robust to edge removal → SPC
- insensitivity to parameters → RNSC
- robust to graph alterations → MCL

- ▶ From Andreopoulos et al (2009), some of these were specifically developed for biological applications
- ▶ Many other methods and variants exist!

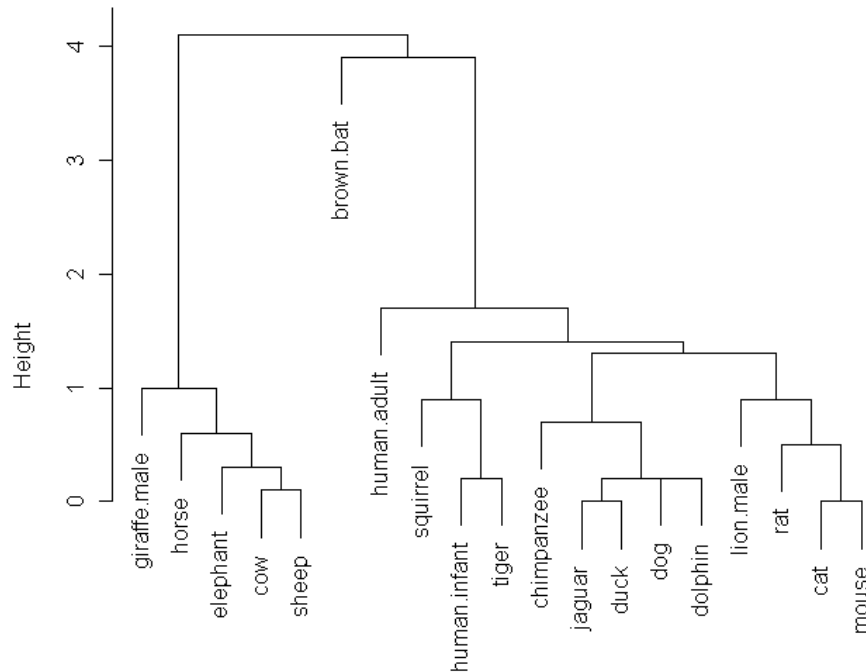
# Key Questions

- ▶ Which method?
  - Depends on application
  - Each method has pros/cons
- ▶ How many clusters to look for?
  - Many methods need an input #, subjective
  - Some methods let the data “decide”
- ▶ How do you evaluate the final clustering solution?
  - Silhouette plots
  - Goodness of fit measures
  - Other metrics
- ▶ How do you select a representative object from a cluster? When would you want to do that?

# Hierarchical Clustering

- ▶ Agglomerative – start with all observations individually and slowly merge them to form one giant cluster
  - Use linkage to determine how to update distances after each merge
  - A distance cutoff determines the number of clusters found
  - Results typically displayed in a dendrogram
- ▶ Divisive – start with all observations in one big cluster and gradually split it till all observations are separate

# Agglomerative Animal Dendrogram

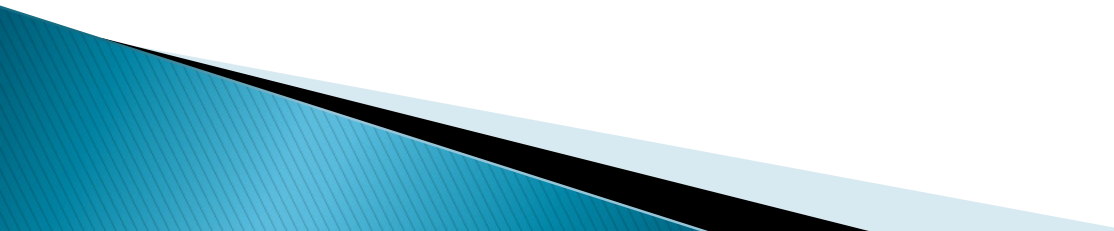


Note if you wanted 3 clusters, brown bat would be all by itself.

- ▶ This used single linkage - meaning the distance was updated at each step as the minimum distance between any two objects in the clusters.
- ▶ What variable was used?

Average amount of sleep required in 24 hour day

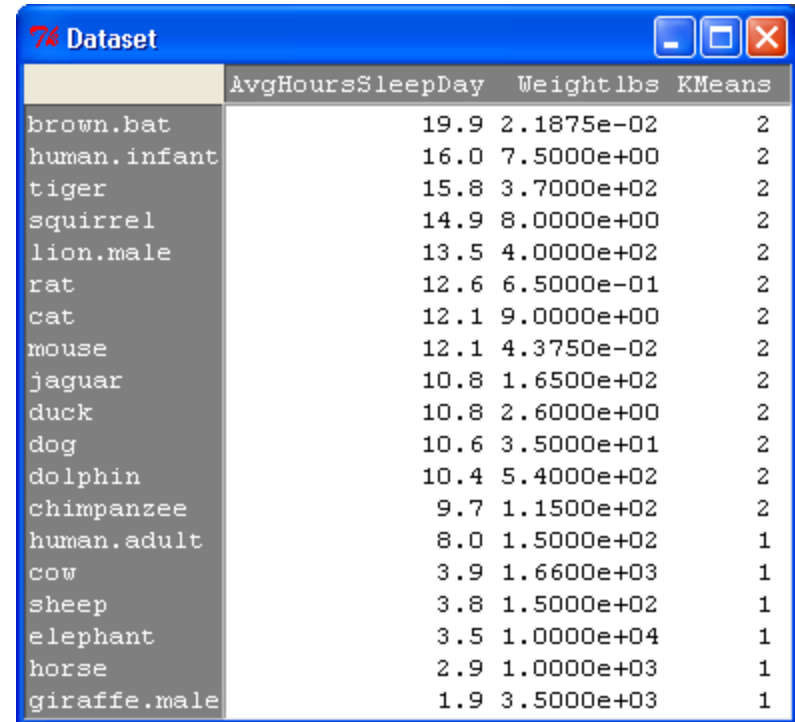
# K-means

- ▶ Partitions the observations into a pre-specified number of clusters,  $k$ , based on the provided distance measure.
  - ▶ Iterative procedure where an observation may be assigned to one cluster and moved later.
  - ▶ Can be sensitive to starting centers that are chosen (more robust versions have been developed as a result).
- 



# K-means Solution on Animal Data

- ▶ The k-means  $k=2$  solution here is almost the same as the agglomerative hierarchical solution.
- ▶ The difference is that human adults were added to the cow/sheep, etc. group in this solution.



	AvgHoursSleepDay	Weightlbs	KMeans
brown.bat	19.9	2.1875e-02	2
human.infant	16.0	7.5000e+00	2
tiger	15.8	3.7000e+02	2
squirrel	14.9	8.0000e+00	2
lion.male	13.5	4.0000e+02	2
rat	12.6	6.5000e-01	2
cat	12.1	9.0000e+00	2
mouse	12.1	4.3750e-02	2
jaguar	10.8	1.6500e+02	2
duck	10.8	2.6000e+00	2
dog	10.6	3.5000e+01	2
dolphin	10.4	5.4000e+02	2
chimpanzee	9.7	1.1500e+02	2
human.adult	8.0	1.5000e+02	1
cow	3.9	1.6600e+03	1
sheep	3.8	1.5000e+02	1
elephant	3.5	1.0000e+04	1
horse	2.9	1.0000e+03	1
giraffe.male	1.9	3.5000e+03	1

The solution was run just using sleep. But think about scaling here if you included weight too! If you didn't standardize, which variable would be driving your solution?