LECTURE 22:
K-MEANS AND HIERARCHICAL CLUSTERING

December 4, 2017
SDS 293: Machine Learning

Announcements 1/2

Consider submitting final write-ups to the Undergraduate Statistics Project Competition!

ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

CAUSE

Deadline: December 22, 2017
www.causeweb.org/usproc

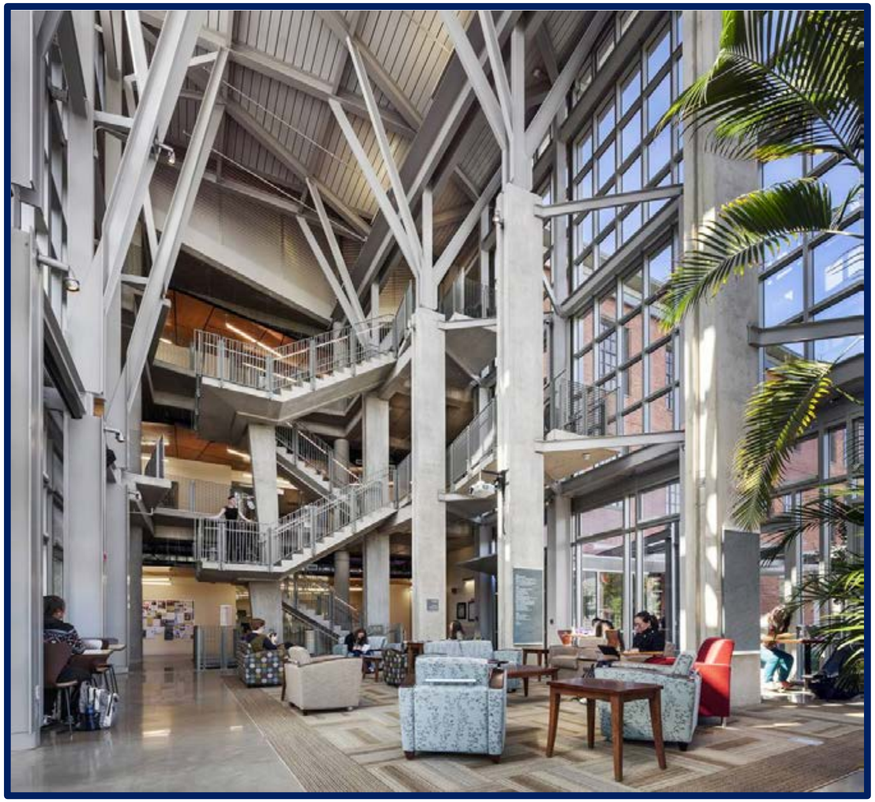
Announcements 2/2

One last change to Final Presentations:

December 13th

9:00 - 10:20am

Ford Hall Atrium



Final Project Deliverables

✓Nov. 8th - FP1: Data Appendix

✓Nov. 27th – FP2: Initial Model

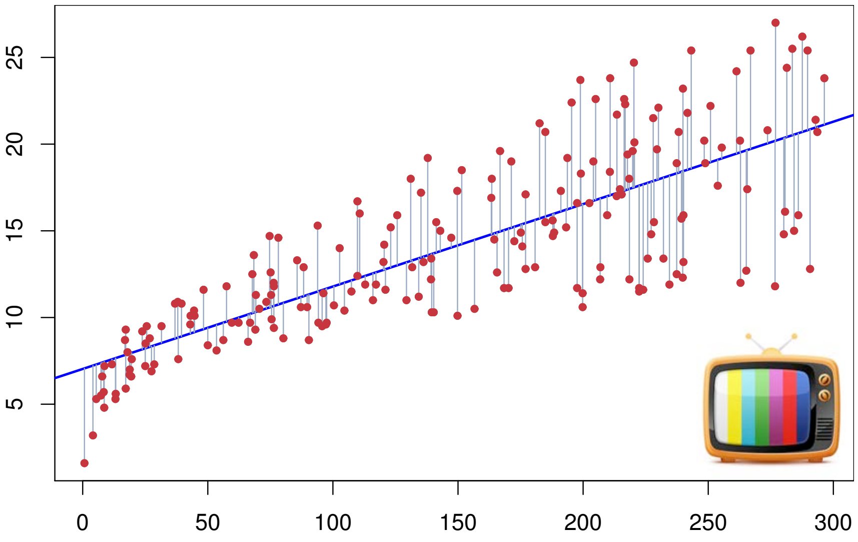
- **Dec. 6th (← new date!) – FP3: Revised Model**
- Dec 13th – Final Project Reception

  (posters due 5pm Friday Dec. 8th if you want Jordan to print it for you)
- Dec. 22nd - FP5: Final Write-Up

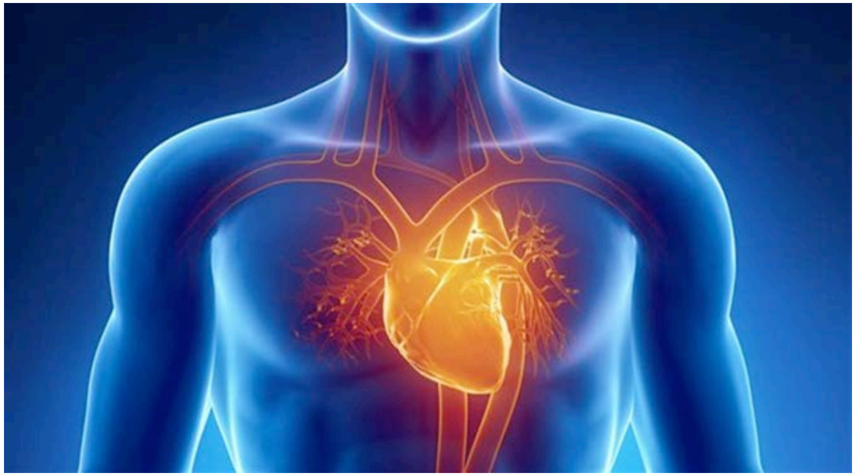
Outline

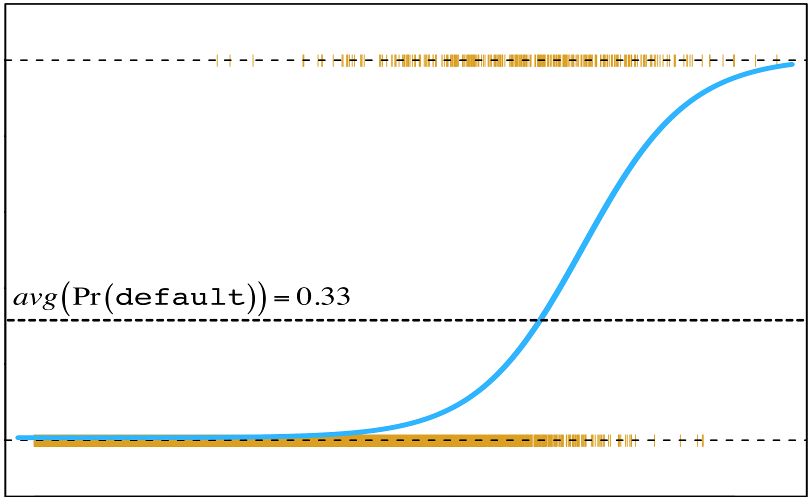
- Supervised vs. unsupervised learning
- Clustering methods
  - K-means
  - Hierarchical
- Lab

Recap


25
20
15
10
5
0
50
100
150
200
250
300


MARVEL
DRAGON BALL Z

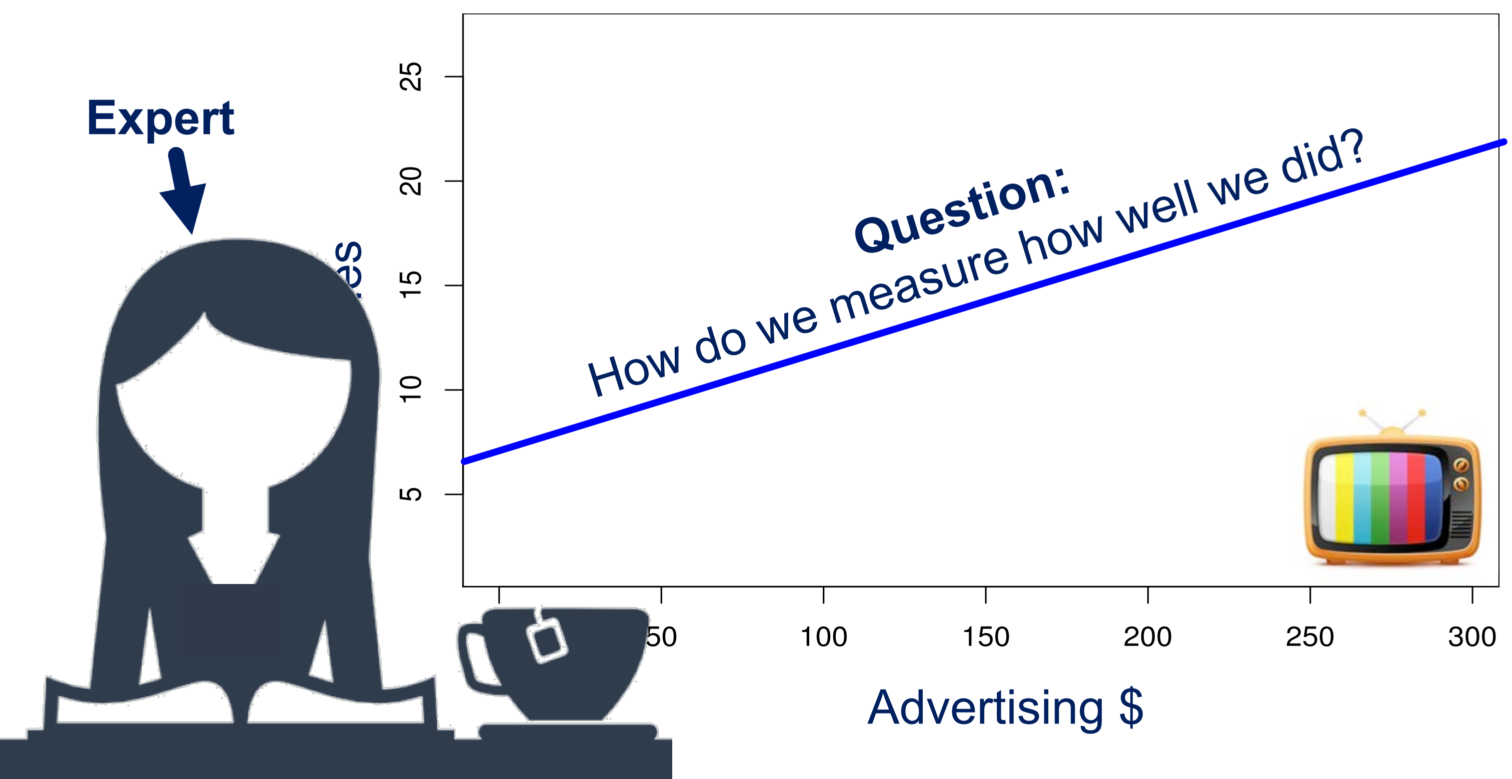



$avg(\Pr(\texttt{default})) = 0.33$

Supervised methods

- **Big idea**: estimate the value of the response using some function of the predictors

$$\hat{y} = f(X)$$



Supervised methods

- When we know the **true value** of the response, we can check our work by seeing how well our model predicts it
  - cross-validation
  - independent test set
  - adjusted R^2, Cp, AIC, BIC, etc.

When we **don't have a response**,

things get a little messier…

Unsupervised methods

- **Goal**: look for structure / patterns in the data **without** having a clear goal (i.e. predict y from X)

- **Examples:**
  - Shoppers with similar browsing and purchase histories
  - Subgroups among tissue samples from 100 breast cancer patients
  - Individuals with similar click patterns when using a search engine

Discussion

- **Question:** what makes this kind of analysis challenging?
- **Answer:** tends to be more subjective, since we don't have a clear measure of "success"

Unsupervised learning is often performed as part of **exploratory data analysis**

Clustering

- **Big idea:** partition observations into distinct groups s.t.
  - observations **within** each group are similar to each other
  - observations in **different** groups are different from each other

- **What we need:** a clear idea of what it means for two or more observations to be *similar* or *different**

*this is usually a domain-specific consideration that must be made based on additional knowledge of the data

K-means clustering

- **Goal**: partition* the observations into a **pre-specified** number of groups

"apples"

"oranges"

*each observation is assigned to **exactly one** group

K-means clustering

- **Big idea**: good clustering = small **within-cluster variation**

- Mathematically, we want to solve the problem:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

- We often use **Euclidean distance**:

$$W(C_k) = \underbrace{\frac{1}{|C_k|} \sum_{i,i' \in C_k}}_{\text{Average over all pairs of obs. in cluster}} \underbrace{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2}_{\text{Euclidean distance}}$$

Discussion

- **Question:** what's the problem?
- **Answer:** there are $O(K^n)$ ways to partition n observations into K groups: a huge number unless K and n are tiny!

Luckily, a very simple algorithm
can be shown to provide **a local optimum***

*a "pretty good" solution

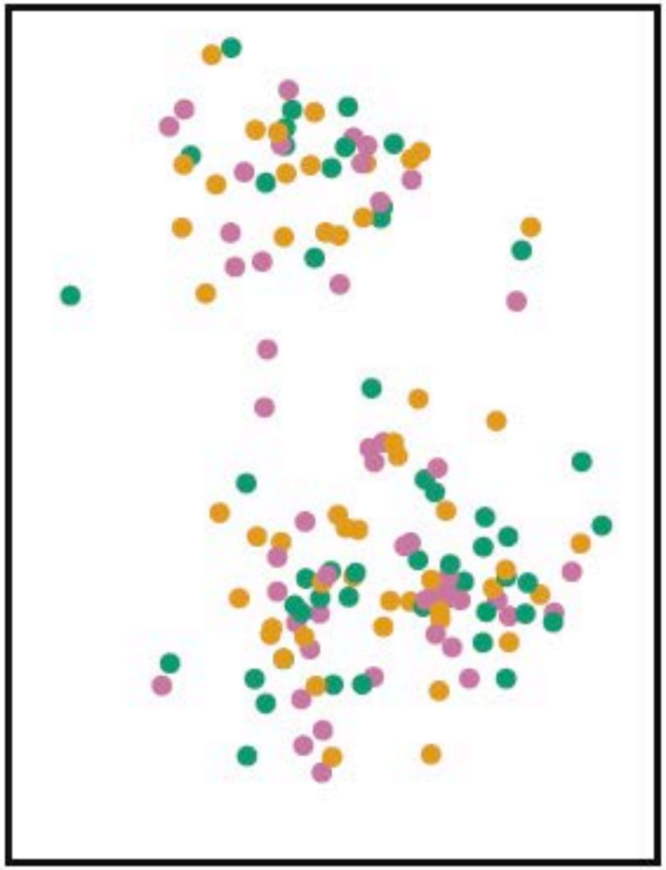
K-means algorithm

1. Randomly assign each observation to a cluster.
2. Iterate until the cluster assignments stop changing:
   a) Compute the vector of the p feature means for the observations in the k^{th} cluster (this is called the **centroid**)
   b) Assign each observation to the cluster whose centroid is closest (where "closest" is defined using Euclidean distance)

Example: k=3

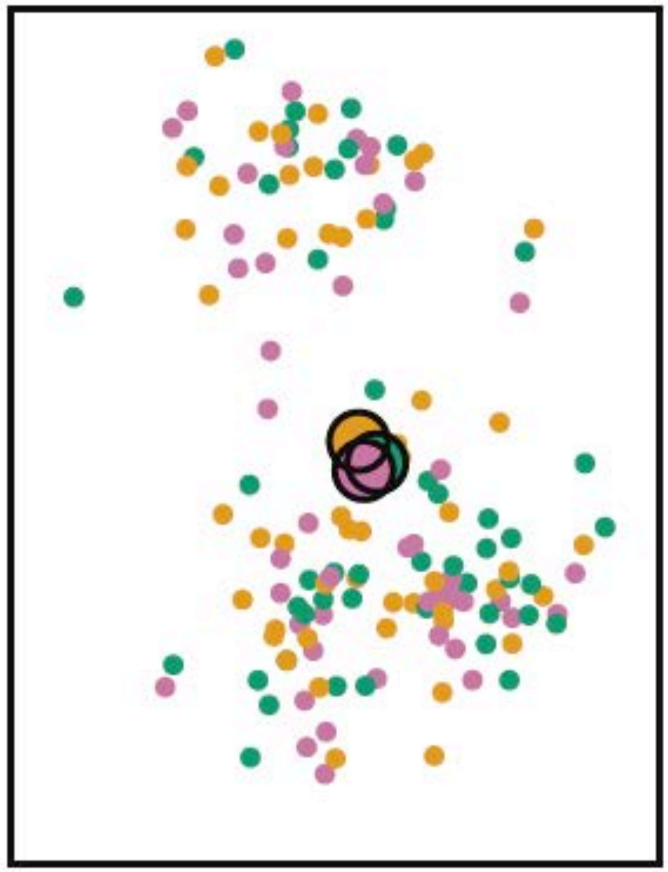
Example: k=3

Randomly assign clusters

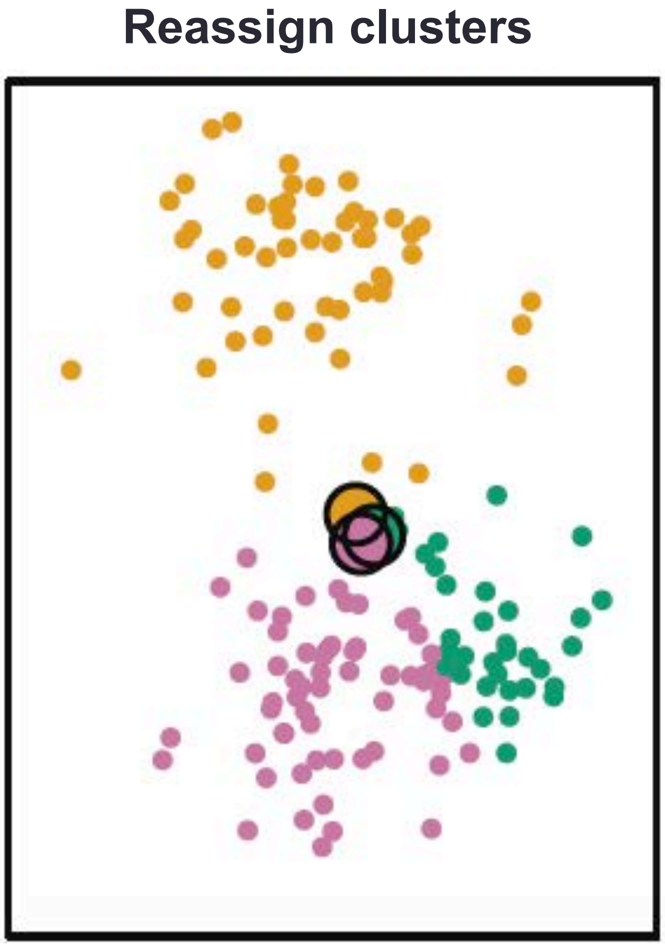


Example: k=3

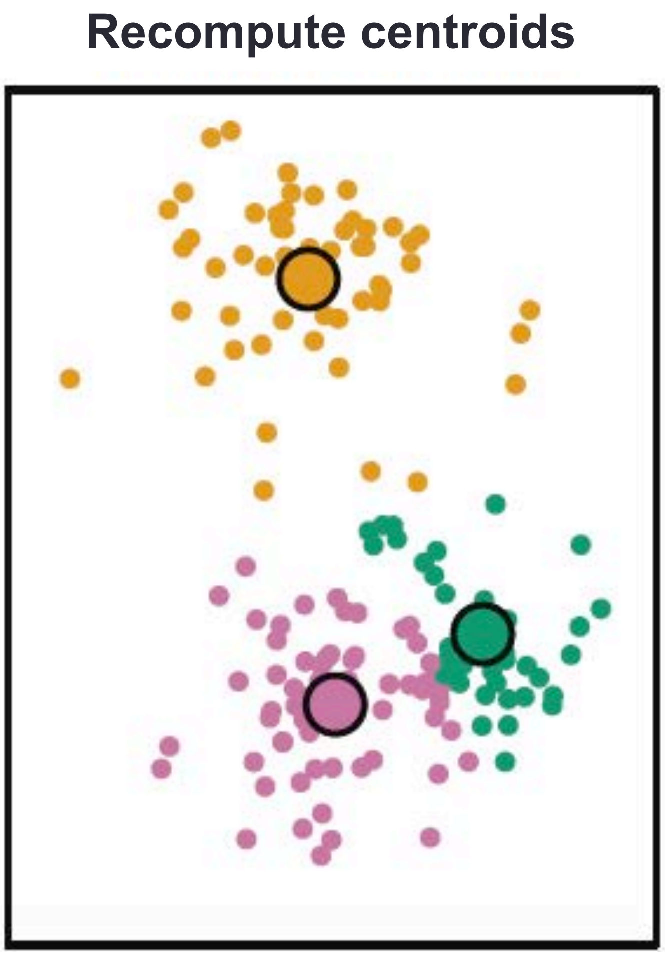
Compute centroids



Example: k=3

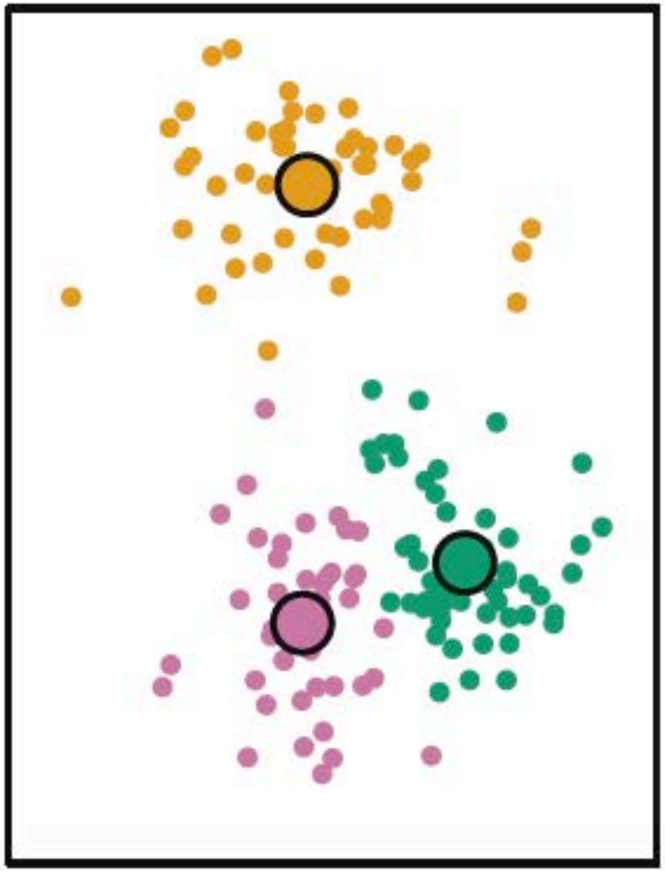


Example: k=3



Example: k=3

Repeat until clusters stabilize



Discussion

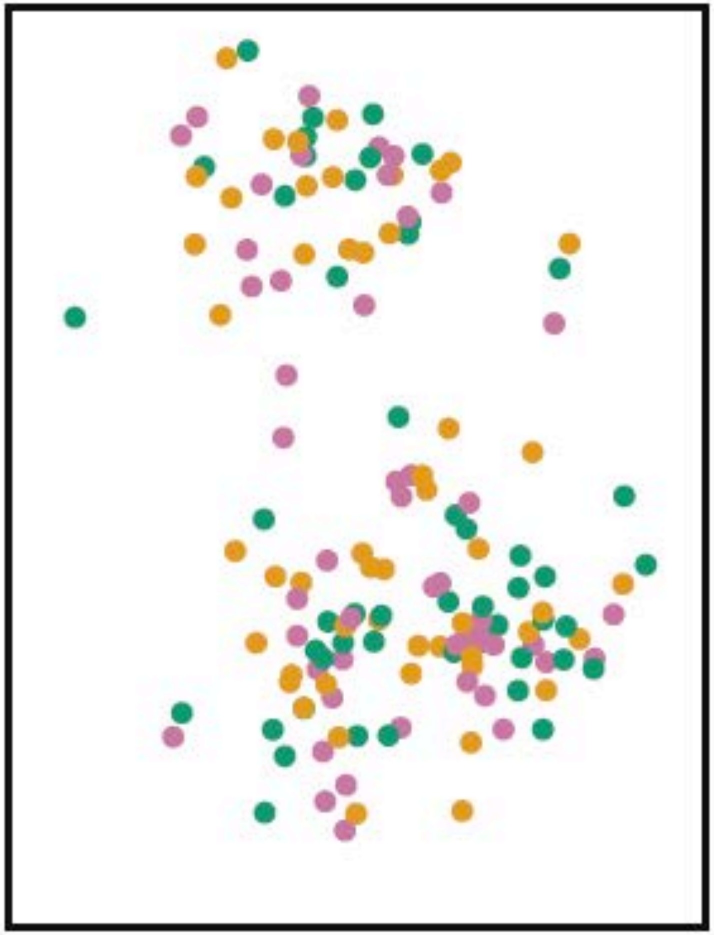
- **Question:** this process is guaranteed to decrease the cluster variation at each step - why?
- **Hint:** the following identify is helpful:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

the **cluster means** are the constants that minimize the sum-of-squared deviations, so reassigning can only help!

K-means clustering

- The K-means algorithm finds a **local** rather than a global optimum
- The results obtained will depend on the initial (random) assignment
- **Important**: run the algorithm multiple times from different initial configurations to avoid getting "stuck"

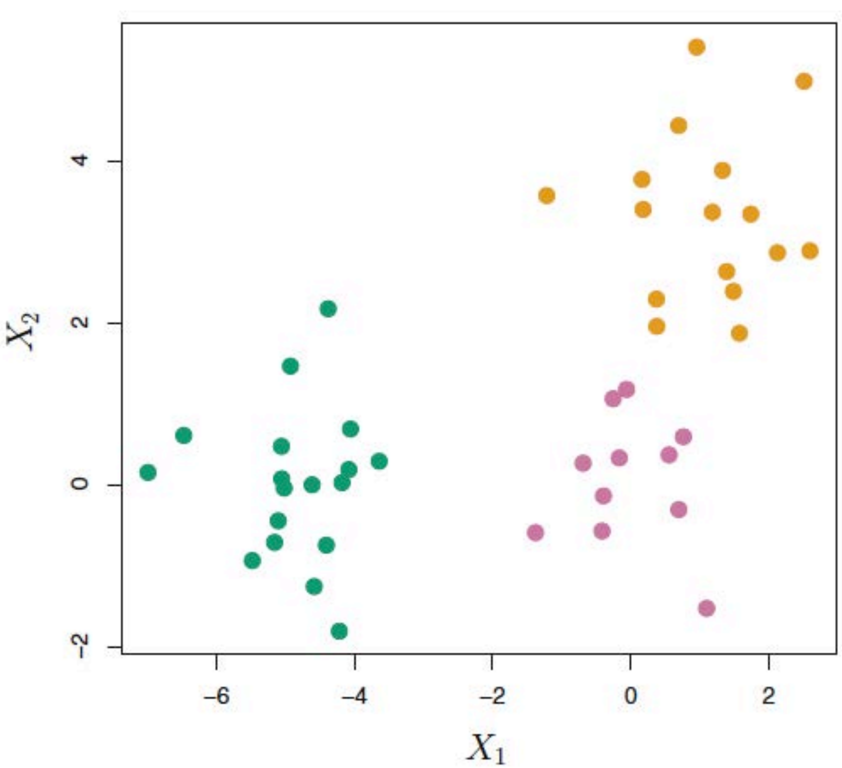


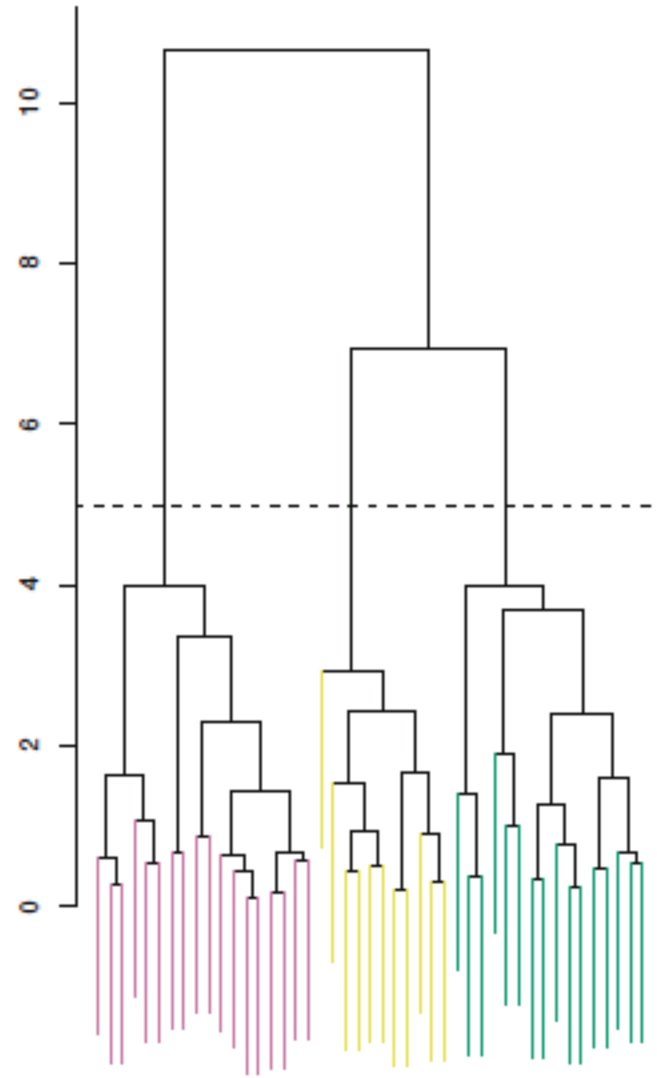
Discussion

- **Question:** so what's the problem with k-means?
- **Answer:** …how do we pick the right number of clusters?
  - Could do a parameter sweep
  - Maybe we have domain knowledge

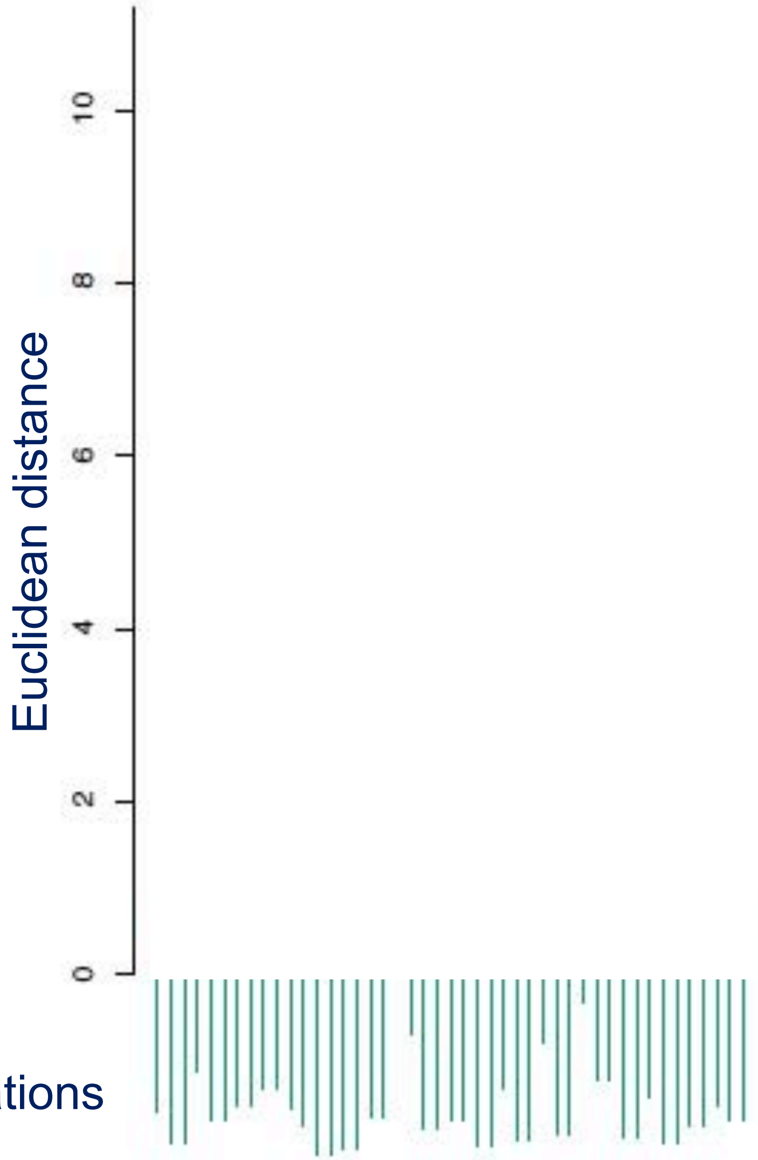
Hierarchical clustering

- **Big idea**: adapt tree-based methods to perform clustering **without** having to pre-specify # of clusters





Dendrograms



Less similar

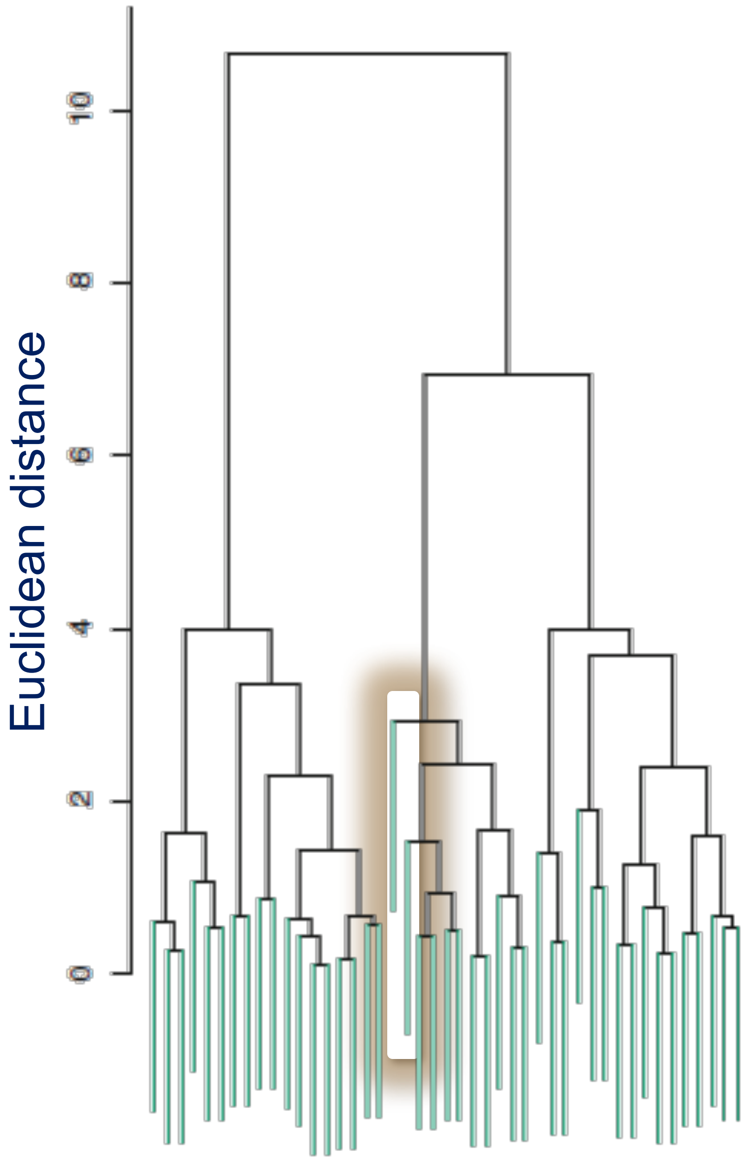
As we move upward, we group observations that are sufficiently similar

More similar

Leaves = individual observations

Dendrograms

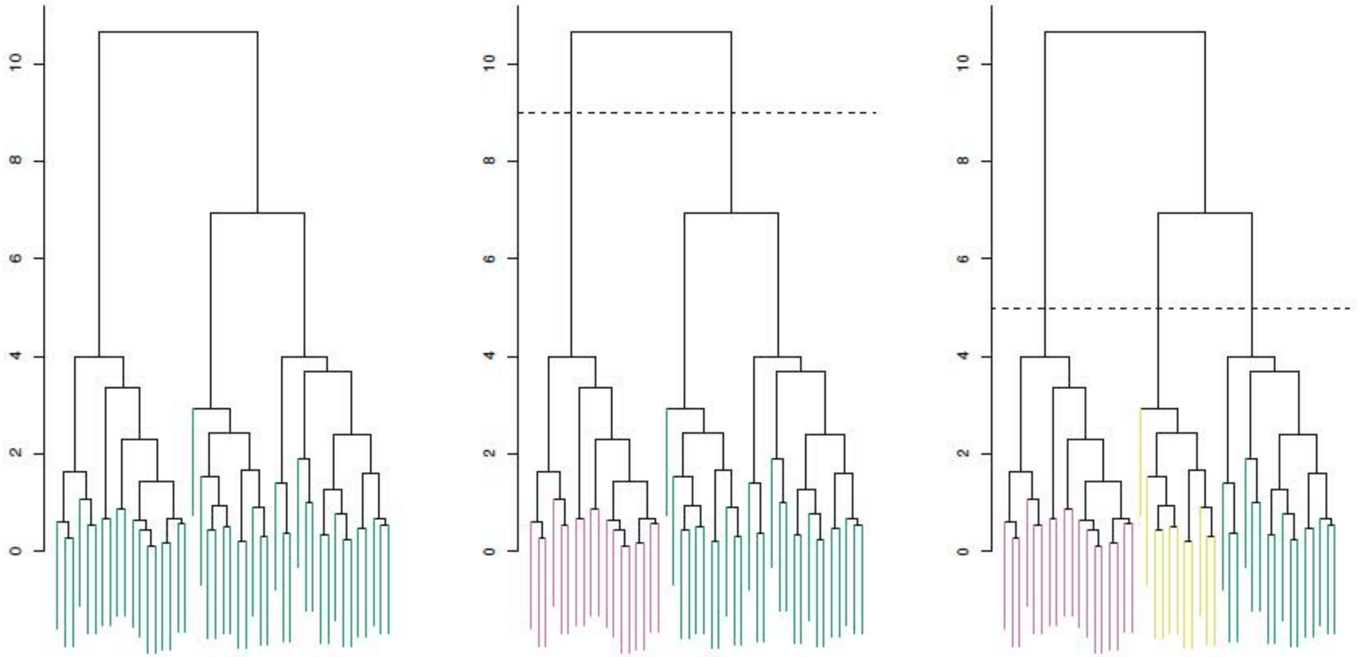
Similarity of observations can be inferred based on the location they first fuse on the **vertical axis**



Important: proximity along **horizontal axis** doesn't tell us anything about similarity!

Hierarchical clustering

- To go from a dendrogram to actual clusters, just cut!



- The **height** of the cut serves the same role as the K in K-means clustering: it controls the number of clusters

Building the dendrogram

- Begin with n observations and a measure of all the $(n \text{ choose } 2)$ pairwise distances. Treat each observation as its own cluster.
- For $i = n, n-1, \ldots, 2$:
  - Examine all pairwise inter-cluster distances and identify the pair of clusters that are most similar.
  - Fuse these two clusters. The distances between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
  - Compute the new pairwise inter-cluster distances.

Discussion

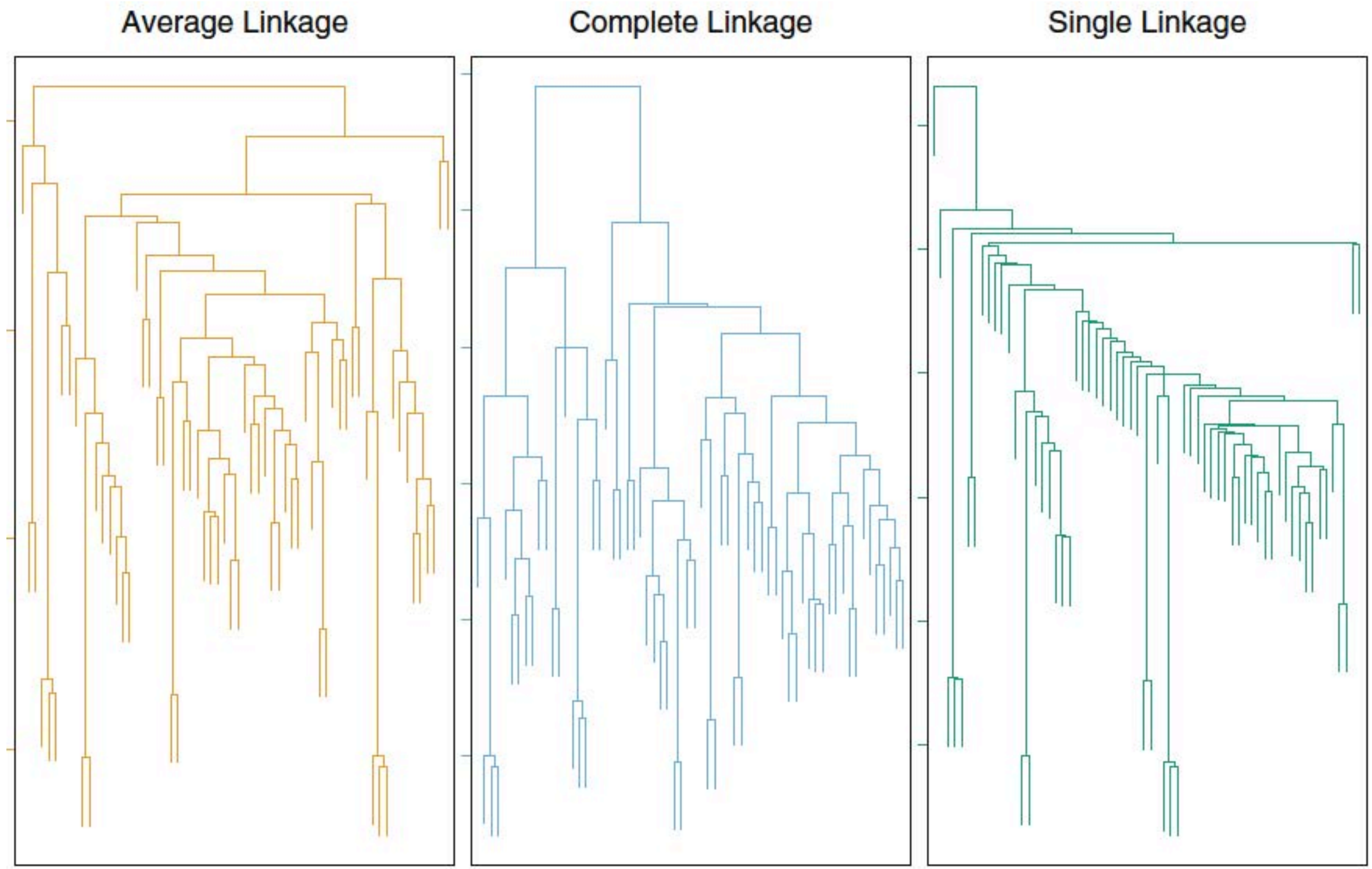
- **Question:** what's missing?
- **Answer:** how do we measure distance between clusters?

Linkage types

- **Complete**: maximal intercluster distance (all pairs)
- **Single**: minimal intercluster distance (all pairs)
- **Average**: mean intercluster distance (all pairs)
- **Centroid**: distance between cluster means (inexpensive, but can result in problematic inversions)

Average, complete = generally more **balanced**

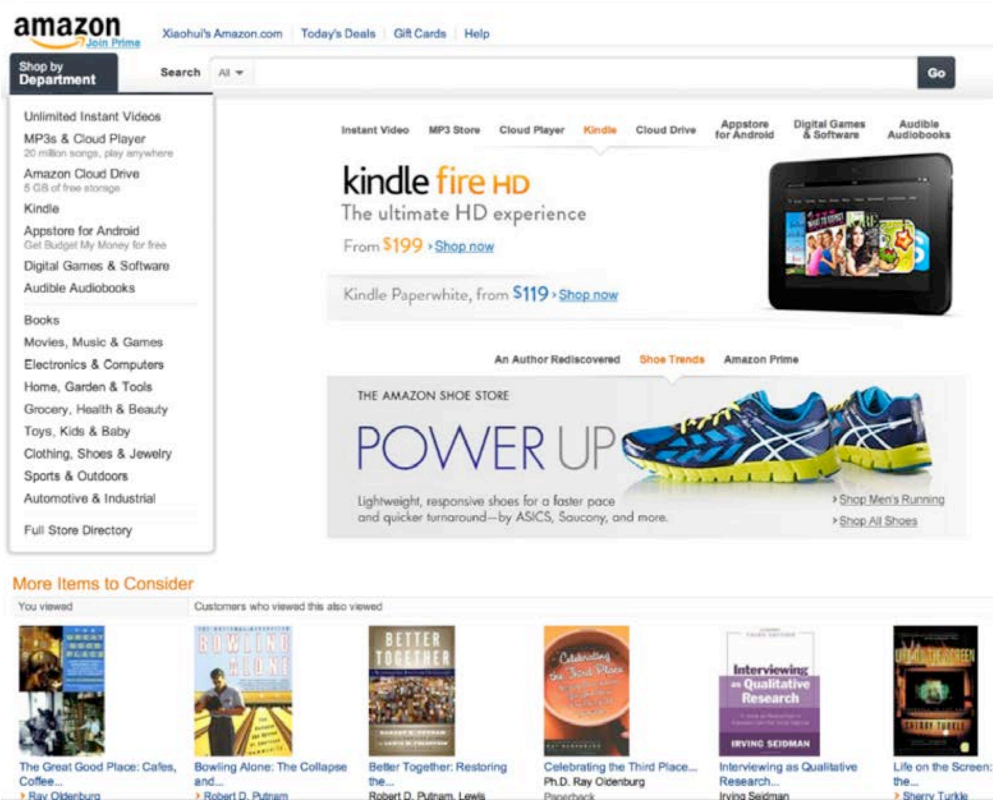
Linkage types



Practical considerations: distance measure

- The choice of **distance measure** is very important, as it has a strong effect on the resulting clusters
- Pay attention to the **type of data** being clustered and the **question** you're answering
- Example:

Online shopping



Shopper	Item 1	Item 2	Item 3	Item 4	…
Alice	1	1	1	0	…
Bob	0	1	1	1	…
Cindy	1	0	0	0	…

Practical considerations: scaling

- Question: should we scale the data to have **standard deviation 1** before measuring similarity?
- If so, then each variable will be given **equal importance** when clustering is performed
- Example:

Small decisions, big consequences

- Each of these decisions can have a **large impact** on the results obtained
- In practice, we usually try **several different choices**, and look for the one that seems the most useful
- Any solution that exposes **some interesting aspect** of the data should be considered!

Lab: clustering

- To do today's lab in R: **`broom`** (just for data wrangling)
- To do today's lab in python: **`scipy`**
- Instructions and code:

  [course website]/labs/lab16-r.html

  [course website]/labs/lab16-r.html
- Full version can be found beginning on p. 404 of ISLR
- If you finish early, take some time to work on your project!

Coming up

- A8 out tonight
- Wednesday 12/6:
  - Jordan in NC
  - Guest lecture: Neural Networks (G. Grinstein)
  - FP3 due
- Monday 12/11:
  - Final lecture: open research questions in ML
  - A8 due
- Wednesday 12/13: **FINAL PROJECT RECEPTION**