Recommendations Worth a Million

An Introduction to Clustering

15.071x – The Analytics Edge

Netflix

- Online DVD rental and streaming video service

- More than 40 million subscribers worldwide

- $3.6 billion in revenue

- Key aspect is being able to offer customers accurate movie recommendations based on a customer's own preferences and viewing history

The Netflix Prize

- From 2006 – 2009 Netflix ran a contest asking the public to submit algorithms to predict user ratings for movies
- Training data set of ~100,000,000 ratings and test data set of ~3,000,000 ratings were provided
- Offered a grand prize of $1,000,000 USD to the team who could beat Netflix's own algorithm, Cinematch, by more than 10%, measured in RMSE

Contest Rules

- If the grand prize was not yet reached, progress prizes of $50,000 USD per year would be awarded for the best result so far, as long as it had >1% improvement over the previous year.
- Teams must submit code and a description of the algorithm to be awarded any prizes
- If any team met the 10% improvement goal, last call would be issued and 30 days would remain for all teams to submit their best algorithm.

Initial Results

- The contest went live on October 2, 2006
- By October 8, a team submitted an algorithm that beat Cinematch
- By October 15, there were three teams with algorithms beating Cinematch
- One of these solutions beat Cinematch by >1%, qualifying for a progress prize

Progress During the Contest

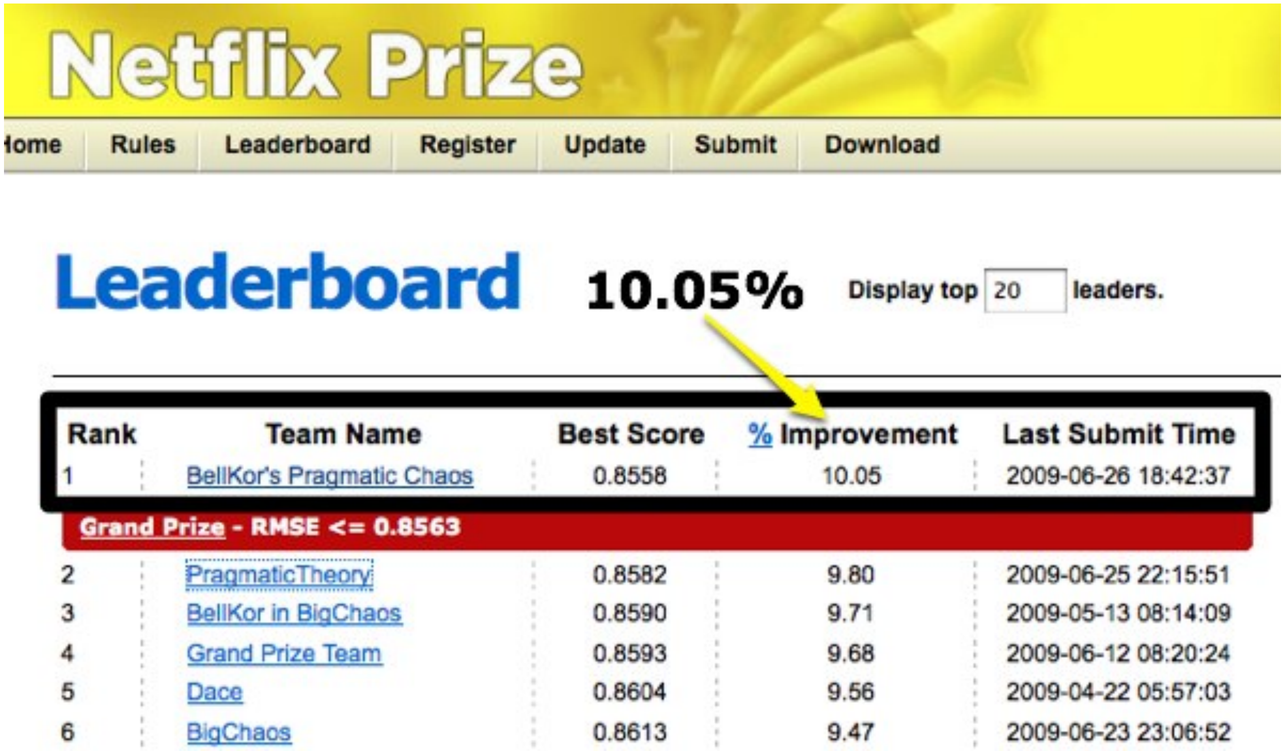
- By June 2007, over 20,000 teams had registered from over 150 countries
- The 2007 progress prize went to team BellKor, with an 8.43% improvement on Cinematch
- In the following year, several teams from across the world joined forces

Competition Intensifies

- The 2008 progress prize went to team BellKor which contained researchers from the original BellKor team as well as the team BigChaos
- This was the last progress prize because another 1% improvement would reach the grand prize goal of 10%

Last Call Announced

- On June 26, 2009, the team BellKor's Pragmatic Chaos submitted a 10.05% improvement over Cinematch



Leaderboard 10.05% Display top 20 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

Predicting the Best User Ratings

- Netflix was willing to pay over $1M for the best user rating algorithm, which shows how critical the recommendation system was to their business

- What data could be used to predict user ratings?

- Every movie in Netflix's database has the ranking from all users who have ranked that movie

- We also know facts about the movie itself: actors, director, genre classifications, year released, etc.

Using Other Users' Rankings

	Men in Black	Apollo 13	Top Gun	Terminator
Amy	5	4	5	4
Bob	3		2	5
Carl		5	4	4
Dan	4	2		

- Consider suggesting to Carl that he watch "Men in Black", since Amy rated it highly and Carl and Amy seem to have similar preferences
- This technique is called **Collaborative Filtering**

Using Movie Information

- We saw that Amy liked "Men In Black"
  - It was directed by Barry Sonnenfeld
  - Classified in the genres of action, adventure, sci-fi and comedy
  - It stars actor Will Smith

- Consider recommending to Amy:
  - Barry Sonnenfeld's movie "Get Shorty"
  - "Jurassic Park", which is in the genres of action, adventure, and sci-fi
  - Will Smith's movie "Hitch"

This technique is called **Content Filtering**

Strengths and Weaknesses

- Collaborative Filtering Systems
  - Can accurately suggest complex items without understanding the nature of the items
  - Requires a lot of data about the user to make accurate recommendations
  - Millions of items – need lots of computing power
- Content Filtering
  - Requires very little data to get started
  - Can be limited in scope

Hybrid Recommendation Systems

- Netflix uses both collaborative and content filtering
- For example, consider a collaborative filtering approach where we determine that Amy and Carl have similar preferences.
- We could then do content filtering, where we would find that "Terminator", which both Amy and Carl liked, is classified in almost the same set of genres as "Starship Troopers"
- Recommend "Starship Troopers" to both Amy and Carl, even though neither of them have seen it before

MovieLens Data

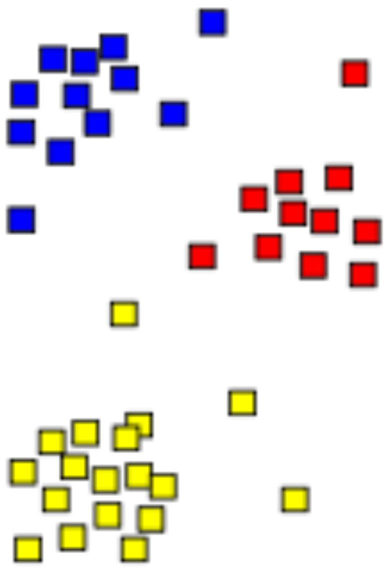
- www.movielens.org is a movie recommendation website run by the GroupLens Research Lab at the University of Minnesota
- They collect user preferences about movies and do collaborative filtering to make recommendations
- We will use their movie database to do content filtering using a technique called clustering

MovieLens Item Dataset

- Movies in the dataset are categorized as belonging to different genres
    - (Unknown)
    - Action
    - Adventure
    - Animation
    - Children’s
    - Comedy
    - Crime
    - Documentary
    - Drama
    - Fantasy
    - Film Noir
    - Horror
    - Musical
    - Mystery
    - Romance
    - Sci-Fi
    - Thriller
    - War
    - Western
- Each movie may belong to many genres
- Can we systematically find groups of movies with similar sets of genres?

Why Clustering?

- "Unsupervised" learning
  - Goal is to segment the data into similar groups instead of prediction
- Can also cluster data into "similar" groups and then build a predictive model for each group
  - Be careful not to overfit your model! This works best with large datasets



Types of Clustering Methods

- There are many different algorithms for clustering
  - Differ in what makes a cluster and how to find them

- We will cover
  - Hierarchical
  - K-means in the next lecture

Distance Between Points

- Need to define distance between two data points
  - Most popular is "Euclidean distance"
  - Distance between points i and j is

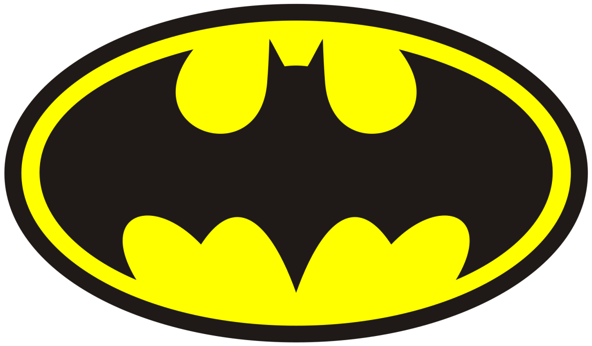
$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}$$

where k is the number of independent variables

Distance Example

- The movie “Toy Story” is categorized as Animation, Comedy, and Children’s
  - Toy Story: (0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0)

- The movie “Batman Forever” is categorized as Action, Adventure, Comedy, and Crime
  - Batman Forever: (0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0)



Distance Between Points

- Toy Story: (0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0)
- Batman Forever: (0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

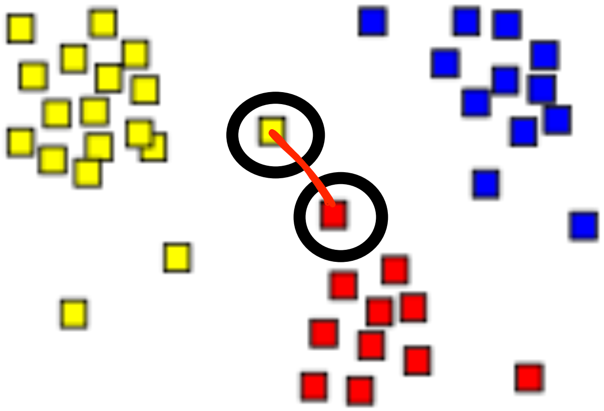
$$d = \sqrt{(0-0)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + \ldots}$$

$$= \sqrt{5}$$

- Other popular distance metrics:
  - Manhattan Distance
    - Sum of absolute values instead of squares
  - Maximum Coordinate Distance
    - Only consider measurement for which data points deviate the most

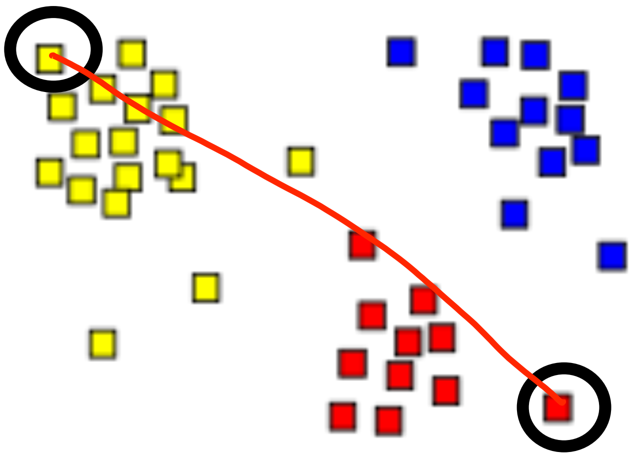
Distance Between Clusters

- Minimum Distance
  - Distance between clusters is the distance between points that are the closest



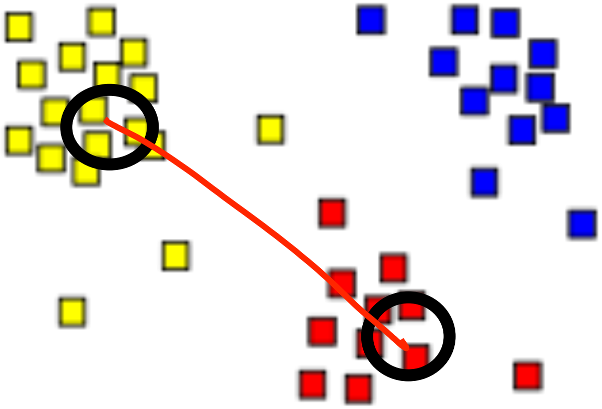
Distance Between Clusters

- Maximum Distance
  - Distance between clusters is the distance between points that are the farthest



Distance Between Clusters

- Centroid Distance
  - Distance between centroids of clusters
    - Centroid is point that has the average of all data points in each component



Normalize Data

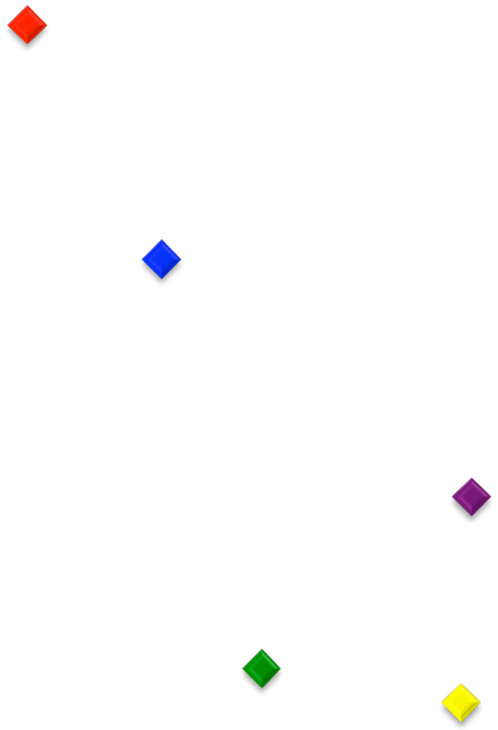
- Distance is highly influenced by scale of variables, so customary to normalize first

- In our movie dataset, all genre variables are on the same scale and so normalization is not necessary

- However, if we included a variable such as "Box Office Revenue," we would need to normalize.

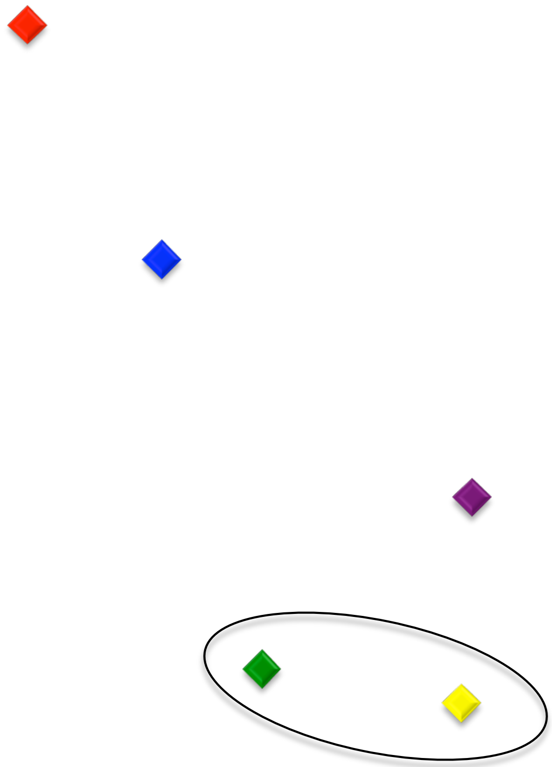
Hierarchical

- Start with each data point in its own cluster



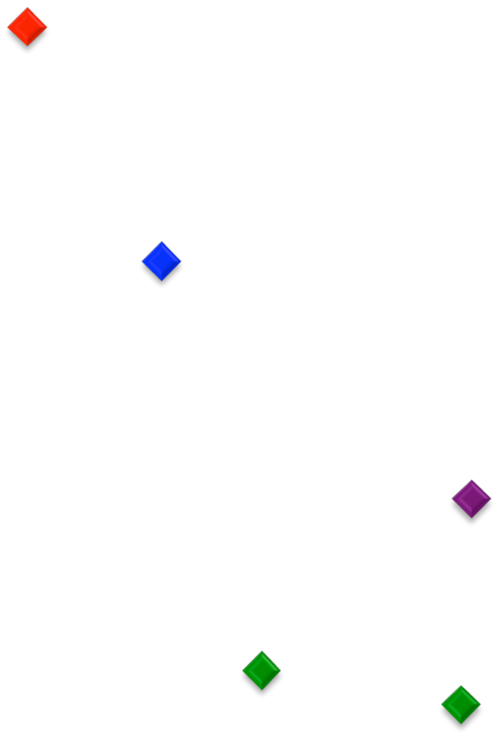
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



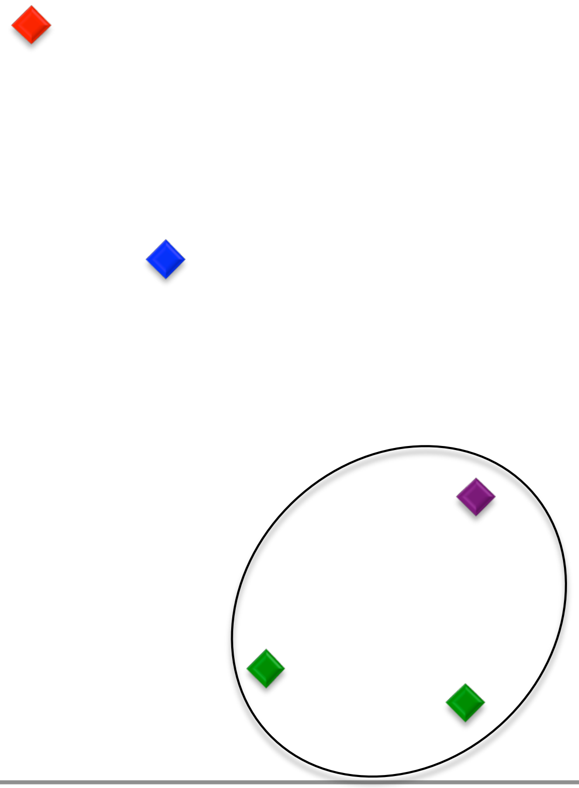
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



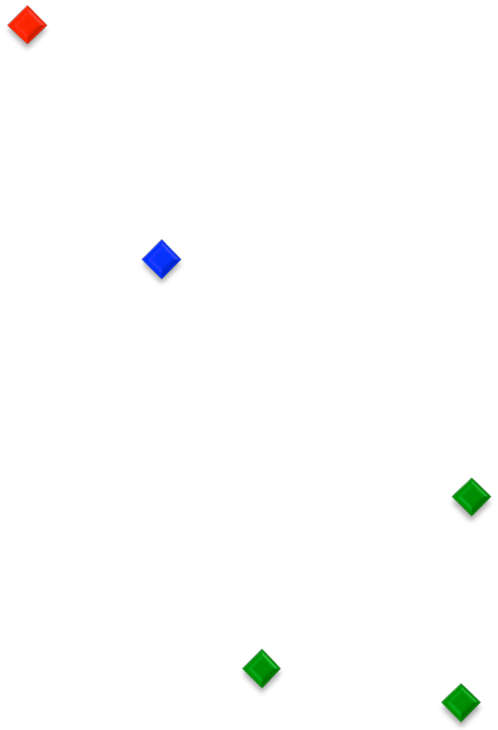
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



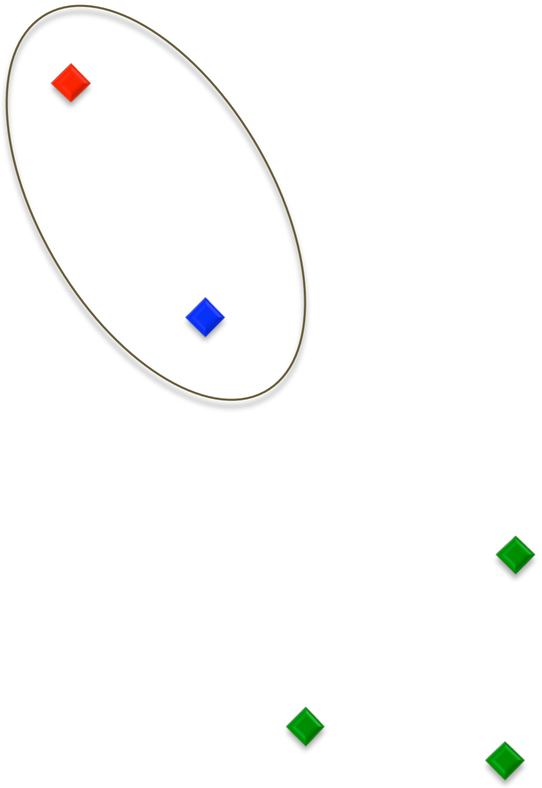
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



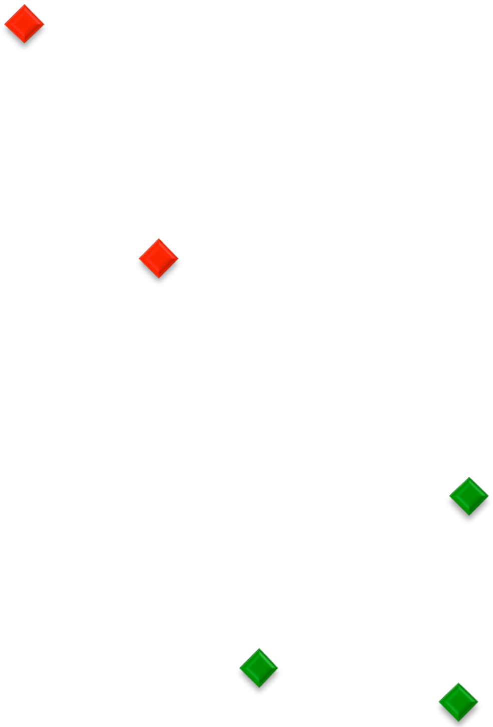
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



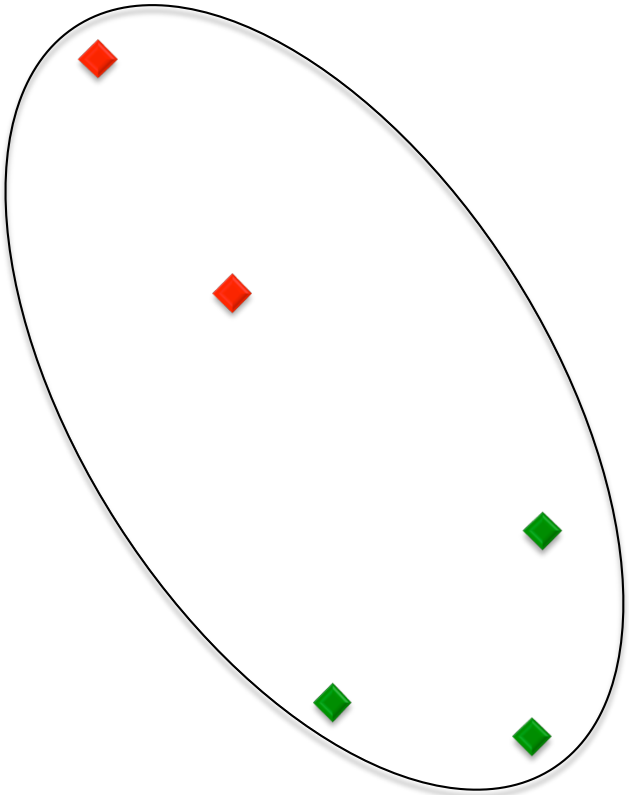
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



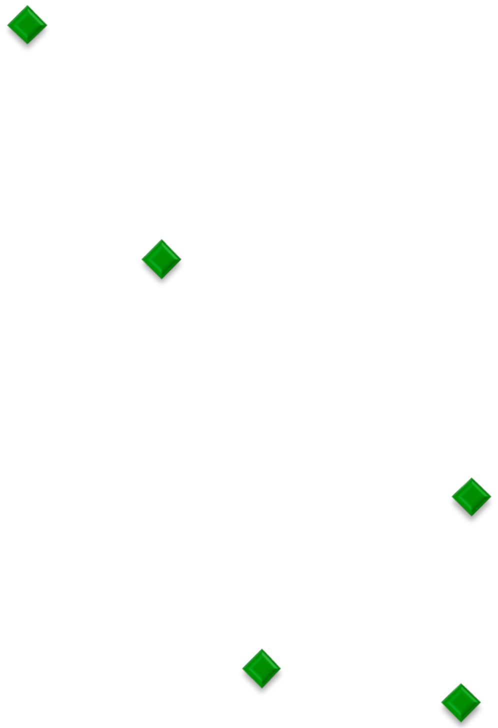
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



Hierarchical

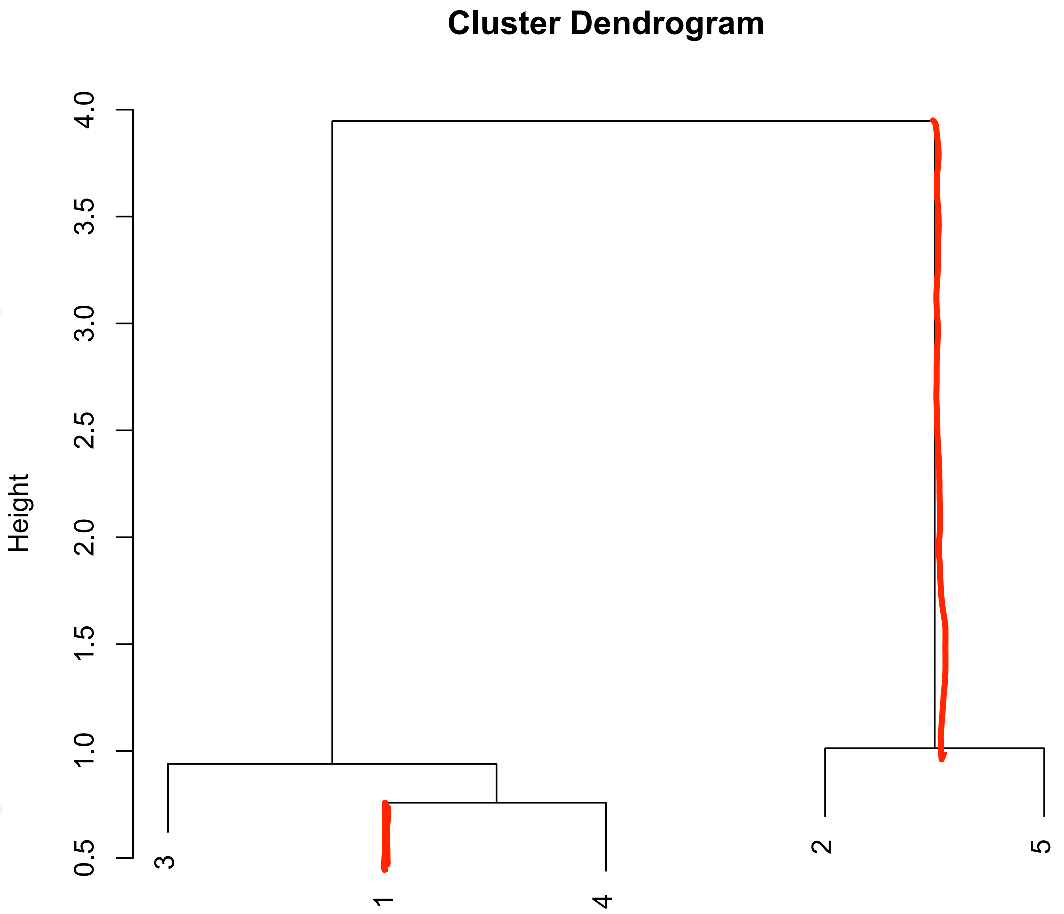
- Combine two nearest clusters (Euclidean, Centroid)



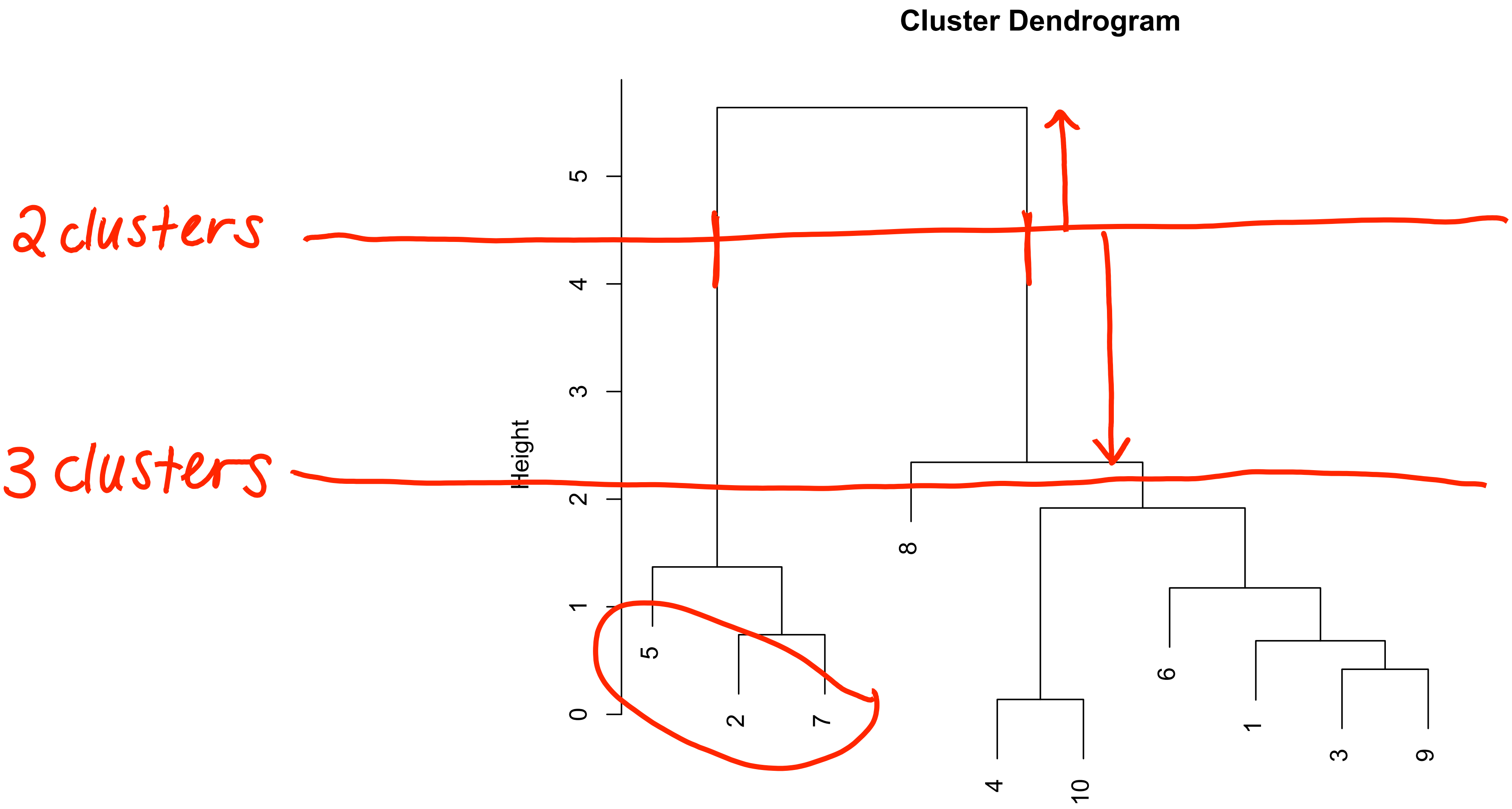
Display Cluster Process

Height of vertical lines represents distance between points or clusters

Data points listed along bottom



Select Clusters



Meaningful Clusters?

- Look at statistics (mean, min, max, . . .) for each cluster and each variable

- See if the clusters have a feature in common that was not used in the clustering (like an outcome)

Beyond Movies: Mass Personalization

- "If I have 3 million customers on the web, I should have 3 million stores on the web"

  – Jeff Bezos, CEO of Amazon.com

- Recommendation systems build models about users' preferences to personalize the user experience

- Help users find items they might not have searched for:
  - A new favorite band
  - An old friend who uses the same social media network
  - A book or song they are likely to enjoy

Cornerstone of these Top Businesses

Recommendation Method Used

- Collaborative Filtering
  - Amazon.com
  - Last.fm
  - Spotify
  - Facebook
  - LinkedIn
  - Google News
  - MySpace
  - **Netflix**
- Content Filtering
  - Pandora
  - IMDB
  - Rotten Tomatoes
  - Jinni
  - Rovi Corporation
  - See This Next
  - MovieLens
  - **Netflix**

The Netflix Prize: The Final 30 Days

- 29 days after last call was announced, on July 25, 2009, the team The Ensemble submitted a 10.09% improvement

- When Netflix stopped accepting submissions the next day, BellKor's Pragmatic Chaos had submitted a 10.09% improvement solution and The Ensemble had submitted a 10.10% improvement solution

- Netflix would now test the algorithms on a private test set and announce the winners

Winners are Declared!

- On September 18, 2009, a winning team was announced

- BellKor's Pragmatic Chaos won the competition and the $1,000,000 grand prize

The Edge of Recommendation Systems

- In today's digital age, businesses often have hundreds of thousands of items to offer their customers
- Excellent recommendation systems can make or break these businesses
- Clustering algorithms, which are tailored to find similar customers or similar items, form the backbone of many of these recommendation systems