GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 11: K-MEANS CLUSTERING

I. CLUSTER ANALYSIS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

supervised	*making predictions*
unsupervised	*discovering patterns*

supervised	*labeled examples*
unsupervised	*no labeled examples*

Q: What is a cluster?

Q: What is a cluster?

A: A group of **similar** *data points.*

Q: What is a cluster?

A: A group of **similar** *data points.*

The concept of similarity is central to the definition of a cluster, and therefore to cluster analysis.

Q: What is the purpose of cluster analysis?

Q: What is the purpose of cluster analysis?

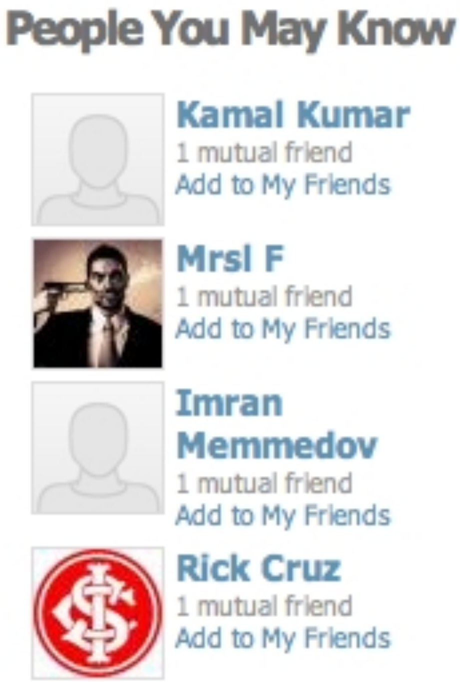
A: To enhance our understanding of a dataset by dividing the data into groups.

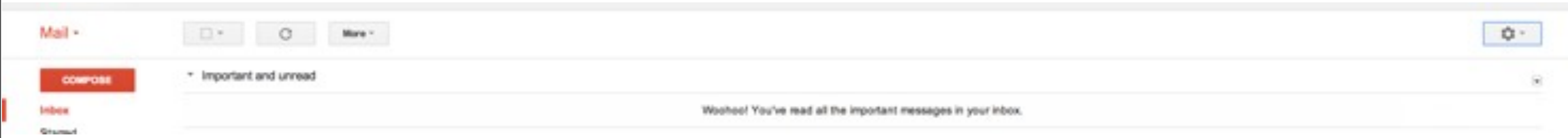
Q: What is the purpose of cluster analysis?

A: To enhance our understanding of a dataset by dividing the data into groups.

Clustering provides a layer of abstraction from individual data points.

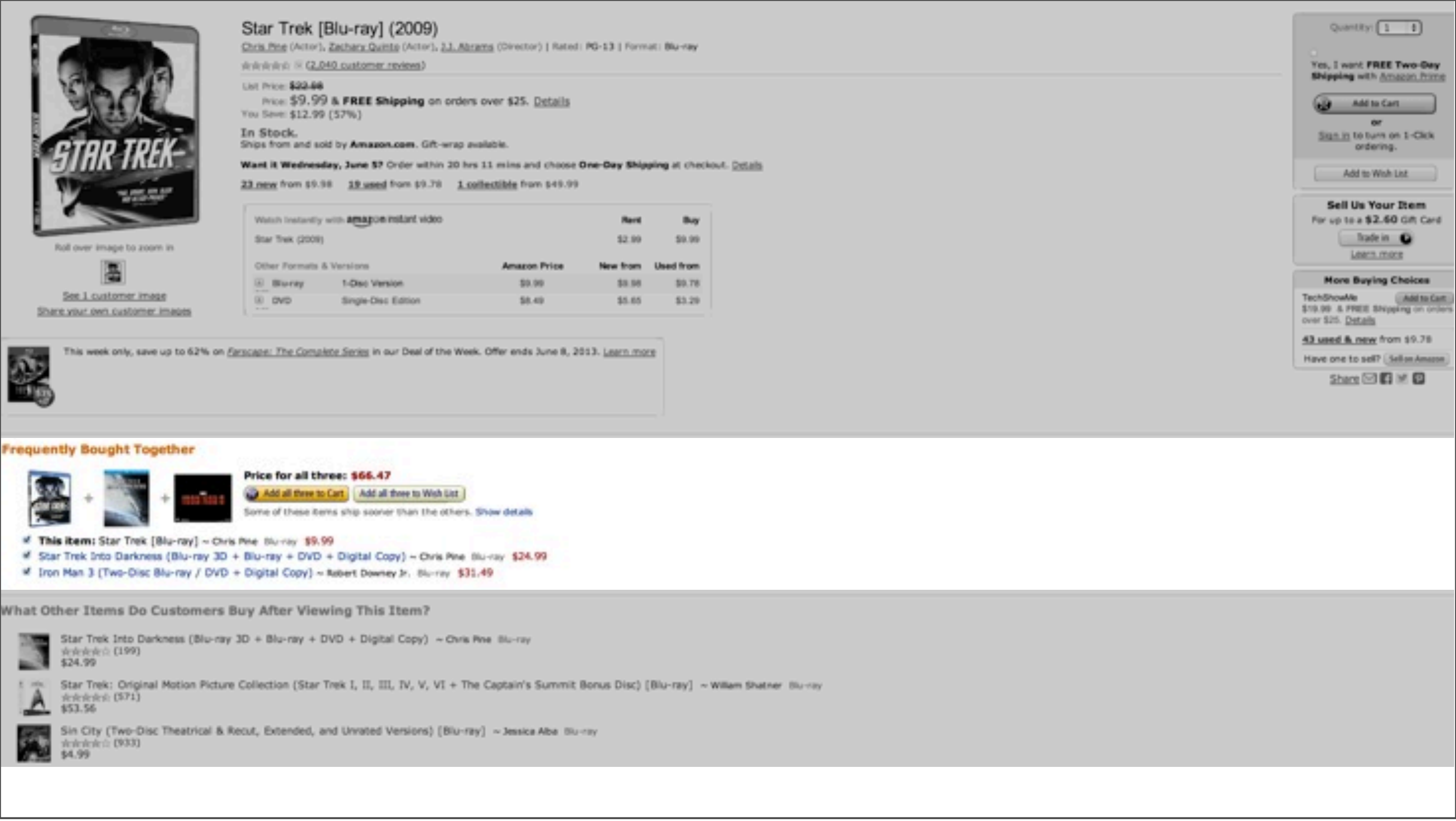
The goal is to extract and enhance the natural structure of the data (not to impose arbitrary structure!)

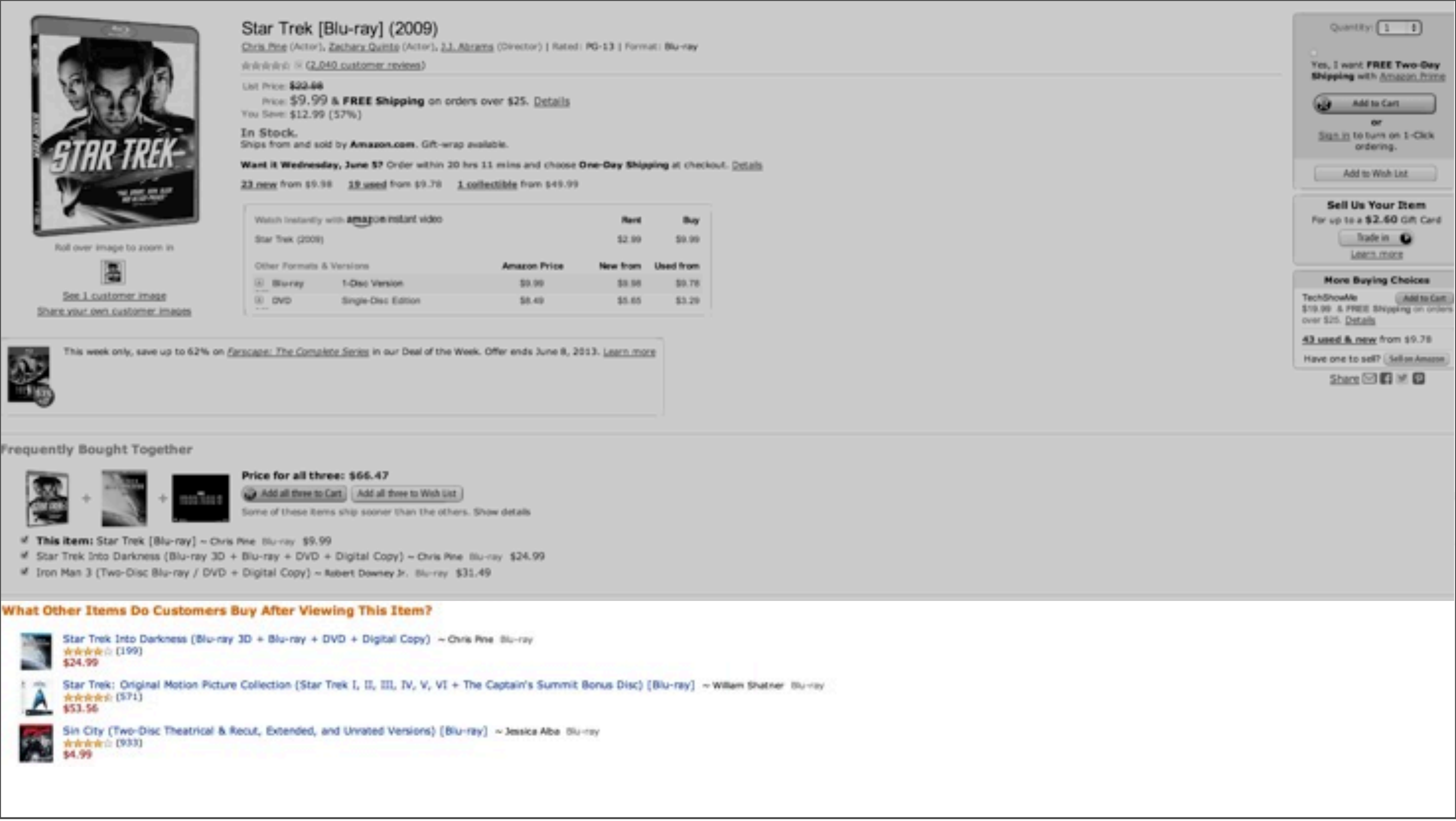

People You May Know
Kamal Kumar
1 mutual friend
Add to My Friends
Mrsl F
1 mutual friend
Add to My Friends
Imran Memmedov
1 mutual friend
Add to My Friends
Rick Cruz
1 mutual friend
Add to My Friends



Priority Inbox: Unsupervised Learning

Group mails into groups and decide which group represents important mails


Star Trek [Blu-ray] (2009)
Chris Pine (Actor), Zachary Quinto (Actor), J.J. Abrams (Director) | Rated: PG-13 | Format: Blu-ray
(2,040 customer reviews)
List Price: $22.98
Price: $9.99 & FREE Shipping on orders over $25. Details
You Save: $12.99 (57%)
In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.
Want it Wednesday, June 5? Order within 20 hrs 11 mins and choose One-Day Shipping at checkout. Details
23 new from $9.98 19 used from $9.78 1 collectible from $49.99
STAR TREK
Roll over image to zoom in
See 1 customer image
Share your own customer images
Watch Instantly with amazon instant video
Rent
Buy
Star Trek (2009)
$2.99
$9.99
Other Formats & Versions
Amazon Price
New from
Used from
Blu-ray 1-Disc Version $9.99 $9.98 $9.78
DVD Single-Disc Edition $8.49 $5.65 $3.29
This week only, save up to 62% on Farscape: The Complete Series in our Deal of the Week. Offer ends June 8, 2013. Learn more
Quantity: 1
Yes, I want FREE Two-Day Shipping with Amazon Prime
Add to Cart
or
Sign in to turn on 1-Click ordering.
Add to Wish List
Sell Us Your Item
For up to a $2.60 Gift Card
Trade in
Learn more
More Buying Choices
TechShowMe
Add to Cart
$19.99 & FREE Shipping on orders over $25. Details
43 used & new from $9.78
Have one to sell?
Sell on Amazon
Share
Frequently Bought Together
Price for all three: $66.47
Add all three to Cart
Add all three to Wish List
Some of these items ship sooner than the others. Show details
This item: Star Trek [Blu-ray] ~ Chris Pine Blu-ray $9.99
Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray $24.99
Iron Man 3 (Two-Disc Blu-ray / DVD + Digital Copy) ~ Robert Downey Jr. Blu-ray $31.49
What Other Items Do Customers Buy After Viewing This Item?
Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray
(199)
$24.99
Star Trek: Original Motion Picture Collection (Star Trek I, II, III, IV, V, VI + The Captain's Summit Bonus Disc) [Blu-ray] ~ William Shatner Blu-ray
(571)
$53.56
Sin City (Two-Disc Theatrical & Recut, Extended, and Unrated Versions) [Blu-ray] ~ Jessica Alba Blu-ray
(933)
$4.99


Star Trek [Blu-ray] (2009)
Chris Pine (Actor), Zachary Quinto (Actor), J.J. Abrams (Director) | Rated: PG-13 | Format: Blu-ray
(2,040 customer reviews)
List Price: $22.98
Price: $9.99 & FREE Shipping on orders over $25. Details
You Save: $12.99 (57%)
In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.
Want it Wednesday, June 5? Order within 20 hrs 11 mins and choose One-Day Shipping at checkout. Details
23 new from $9.98 19 used from $9.78 1 collectible from $49.99
Roll over image to zoom in
See 1 customer image
Share your own customer images
Sell Us Your Item
More Buying Choices
This week only, save up to 62% on Farscape: The Complete Series in our Deal of the Week. Offer ends June 8, 2013. Learn more
Frequently Bought Together
Price for all three: $66.47
Add all three to Cart
Add all three to Wish List
Some of these items ship sooner than the others. Show details
This item: Star Trek [Blu-ray] ~ Chris Pine Blu-ray $9.99
Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray $24.99
Iron Man 3 (Two-Disc Blu-ray / DVD + Digital Copy) ~ Robert Downey Jr. Blu-ray $31.49
What Other Items Do Customers Buy After Viewing This Item?
Star Trek Into Darkness (Blu-ray 3D + Blu-ray + DVD + Digital Copy) ~ Chris Pine Blu-ray
(199)
$24.99
Star Trek: Original Motion Picture Collection (Star Trek I, II, III, IV, V, VI + The Captain's Summit Bonus Disc) [Blu-ray] ~ William Shatner Blu-ray
(571)
$53.56
Sin City (Two-Disc Theatrical & Recut, Extended, and Unrated Versions) [Blu-ray] ~ Jessica Alba Blu-ray
(933)
$4.99

Q: How do you solve a clustering problem?

Q: How do you solve a clustering problem?

A: Think of a cluster as a "potential class"; then the solution to a clustering problem is to programatically determine these classes.

Q: How do you solve a clustering problem?

A: Think of a cluster as a "potential class"; then the solution to a clustering problem is to programatically determine these classes.

The real purpose of clustering is data exploration, so a solution is anything that contributes to your understanding.

II. K-MEANS CLUSTERING

Q: What is k-means clustering?

Q: What is k-means clustering?

A: A **greedy** *learner that* **partitions** *a data set into k clusters.*

Q: What is k-means clustering?

A: A **greedy** *learner that* **partitions** *a data set into k clusters.*

greedy – captures local structure (depends on initial conditions)

Q: What is k-means clustering?

A: A **greedy** *learner that* **partitions** *a data set into k clusters.*

greedy – captures local structure

partition – each point belongs to exactly one cluster

Q: How are these partitions determined?

Q: How are these partitions determined?

A: Each point is assigned to the cluster with the nearest **centroid***.*

Q: How are these partitions determined?

A: Each point is assigned to the cluster with the nearest **centroid***.*

centroid – the mean of the data points in a cluster
→ requires continuous (vector-like) features
→ highlights iterative nature of algorithm

One important point to keep in mind is that partitions are not scale-invariant!

One important point to keep in mind is that partitions are not scale-invariant!

This means that the same data can yield very different clustering results depending on the scale and the units used.

One important point to keep in mind is that partitions are not scale-invariant!

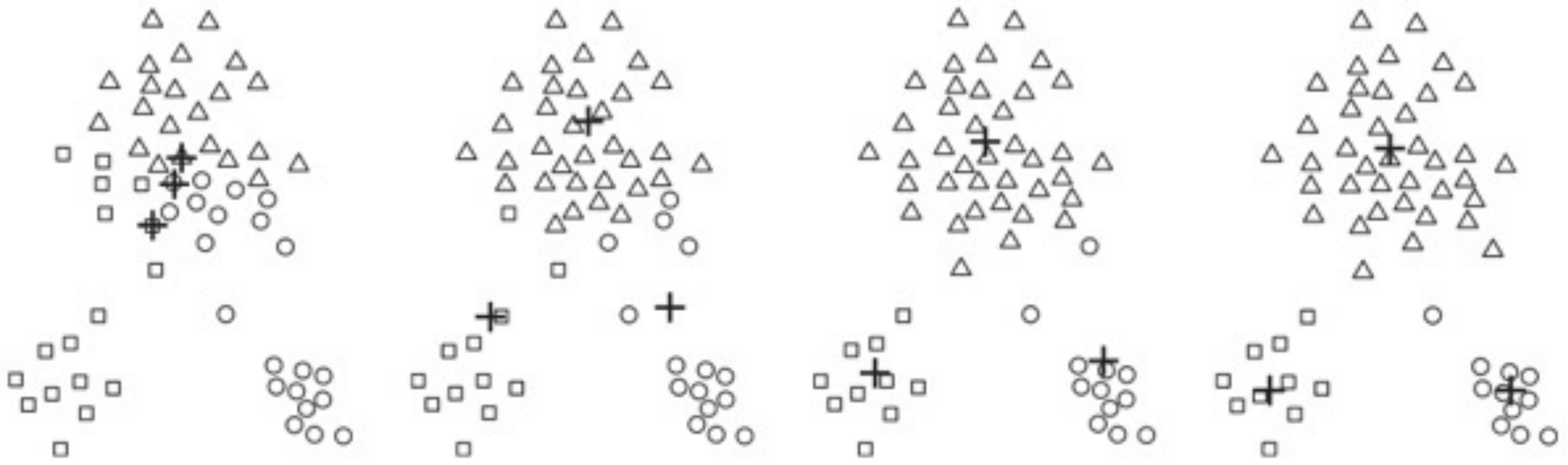
This means that the same data can yield very different clustering results depending on the scale and the units used.

Therefore it's important to think about your data representation before applying a clustering algorithm.

1) *choose k initial centroids (note that k is an input)*

2) *for each point:*
   - *find distance to each centroid*
   - *assign point to nearest centroid*

3) *recalculate centroid positions*
4) *repeat steps 2-3 until stopping criteria met*



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Figure 8.3. Using the K-means algorithm to find three clusters in sample data.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

K-means is algorithmically pretty efficient (time & space complexity is linear in number of records).

Q: How do you choose the initial centroid positions?

Q: How do you choose the initial centroid positions?

A: There are several options:

Q: How do you choose the initial centroid positions?

A: There are several options:

- randomly (but may yield divergent behavior)

Q: How do you choose the initial centroid positions?

A: There are several options:

- *randomly (but may yield divergent behavior)*
- *perform alternative clustering task, use resulting centroids as initial k-means centroids*

Q: How do you determine which centroid is the nearest?

Q: How do you determine which centroid is the nearest?

The “nearness” criterion is determined by the similarity/distance measure we discussed earlier.

Q: How do you determine which centroid is the nearest?

The "nearness" criterion is determined by the similarity/distance measure we discussed earlier.

This measure makes quantitative inference possible.

There are a number of different similarity measures to choose from, and in general the right choice depends on the problem.

There are a number of different similarity measures to choose from, and in general the right choice depends on the problem.

For data that takes values in R^n, the typical choice is the **Euclidean distance***:*

$$d(x, y) = \sqrt{\sum(x_i - y_i)^2}$$

There are a number of different similarity measures to choose from, and in general the right choice depends on the problem.

For data that takes values in R^n, the typical choice is the **Euclidean distance***:*

$$d(x, y) = \sqrt{\sum(x_i - y_i)^2}$$

We can express different semantics about our data through the choice of metric.

The matrix whose entries D_{ij} *contain the values* $d(x, y)$ *for all* x *and* y *is called the* **distance matrix**.

The matrix whose entries D_{ij} contain the values $\mathrm{d}(x, y)$ for all x and y is called the **distance matrix**.

The distance matrix contains all of the information we know about the dataset.

The matrix whose entries D_{ij} *contain the values* $d(x, y)$ *for all* x *and* y *is called the* **distance matrix***.*

The distance matrix contains all of the information we know about the dataset.

For this reason, it's really the choice of metric that determines the definition of a cluster.

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

A: By optimizing an **objective function** *that tells us how "good" the clustering is.*

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

A: By optimizing an **objective function** *that tells us how "good" the clustering is.*

The iterative part of the algorithm (recomputing centroids and reassigning points to clusters) explicitly tries to minimize this objective function.

Ex: Using the Euclidean distance measure, one typical objective function is the **sum of squared errors** *from each point x to its centroid c_i:*

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2$$

Ex: Using the Euclidean distance measure, one typical objective function is the **sum of squared errors** *from each point x to its centroid c_i:*

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2$$

Given two clusterings, we will prefer the one with the lower SSE since this means the centroids have converged to better locations (a better local optimum).

We iterate until some stopping criteria are met; in general, suitable convergence is achieved in a small number of steps.

III. CLUSTER VALIDATION

supervised	*test out your predictions*
unsupervised	*...*

In general, k-means will converge to a solution and return a partition of k clusters, even if no natural clusters exist in the data.

In general, k-means will converge to a solution and return a partition of k clusters, even if no natural clusters exist in the data.

We will look at two validation metrics useful for partitional clustering, **cohesion** *and* **separation***.*

Cohesion *measures clustering effectiveness within a cluster.*

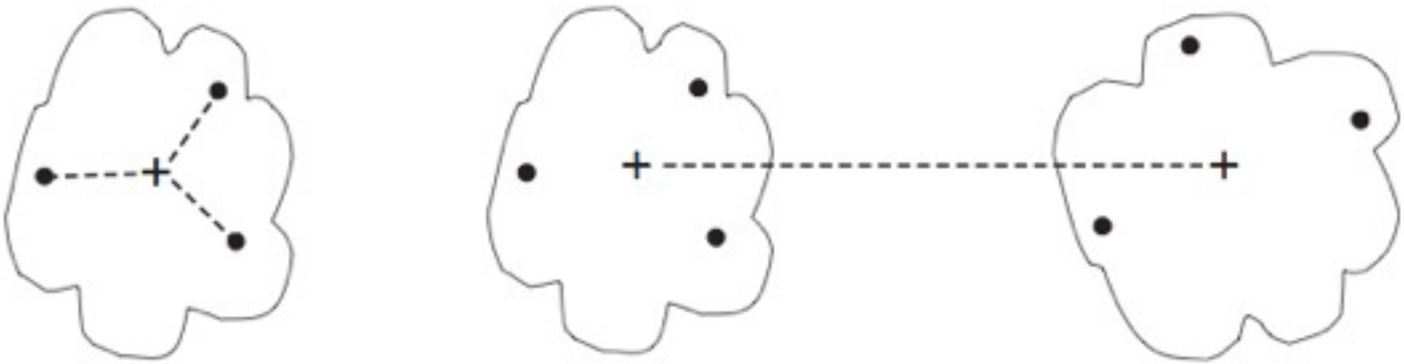
$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

Cohesion *measures clustering effectiveness within a cluster.*

$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

Separation *measures clustering effectiveness between clusters.*

$$\hat{S}(C_i, C_j) = d(c_i, c_j)$$



(a) Cohesion. (b) Separation.

Figure 8.28. Prototype-based view of cluster cohesion and separation.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

We can turn these values into overall measures of clustering validity by taking a weighted sum over clusters:

$$\hat{V}_{total} = \sum_{1}^{K} w_i \hat{V}(C_i)$$

Here V can be cohesion, separation, or some function of both.

We can turn these values into overall measures of clustering validity by taking a weighted sum over clusters:

$$\hat{V}_{total} = \sum_{1}^{K} w_i \hat{V}(C_i)$$

Here V can be cohesion, separation, or some function of both.

The weights can all be set to 1 (best for k-means), or proportional to the cluster masses (the number of points they contain).

Cluster validation measures can be used to identify clusters that should be split or merged, or to identify individual points with disproportionate effect on the overall clustering.

One useful measure than combines the ideas of cohesion and separation is the **silhouette coefficient**. *For point x_i, this is given by:*

$$SC_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

such that:

a_i = *average in-cluster distance to* x_i

b_{ij} = *average between-cluster distance to* x_i

$b_i = min_j(b_{ij})$

The silhouette coefficient can take values between −1 and 1.

In general, we want separation to be high and cohesion to be low. This corresponds to a value of SC close to +1.

A negative silhouette coefficient means the cluster radius is larger than the space between clusters, and thus clusters overlap.

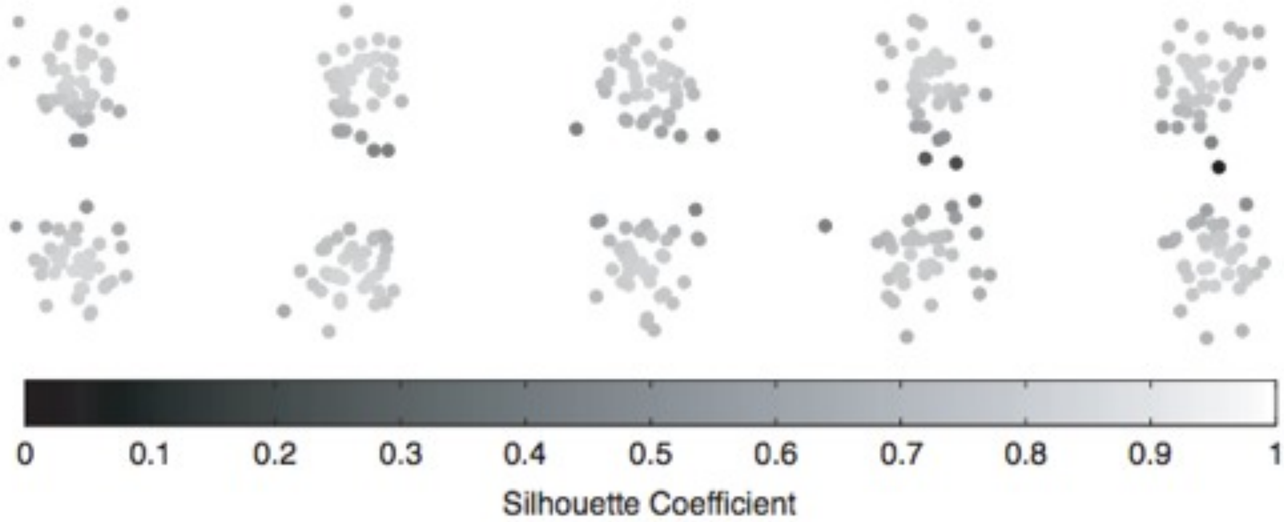


Figure 8.29. Silhouette coefficients for points in ten clusters.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

The silhouette coefficient for the cluster C_i is given by the average silhouette coefficient across all points in C_i:

$$SC(C_i) = \frac{1}{m_i} \sum_{x \in C_i} SC_i$$

The silhouette coefficient for the cluster C_i is given by the average silhouette coefficient across all points in C_i:

$$SC(C_i) = \frac{1}{m_i} \sum_{x \in C_i} SC_i$$

The overall silhouette coefficient is given by the average silhouette coefficient across all points:

$$SC_{total} = \frac{1}{k} \sum_{1}^{k} SC(C_i)$$

The silhouette coefficient for the cluster C_i is given by the average silhouette coefficient across all points in C_i:

$$SC(C_i) = \frac{1}{m_i} \sum_{x \in C_i} SC_i$$

The overall silhouette coefficient is given by the average silhouette coefficient across all points:

$$SC_{total} = \frac{1}{k} \sum_{1}^{k} SC(C_i)$$

NOTE

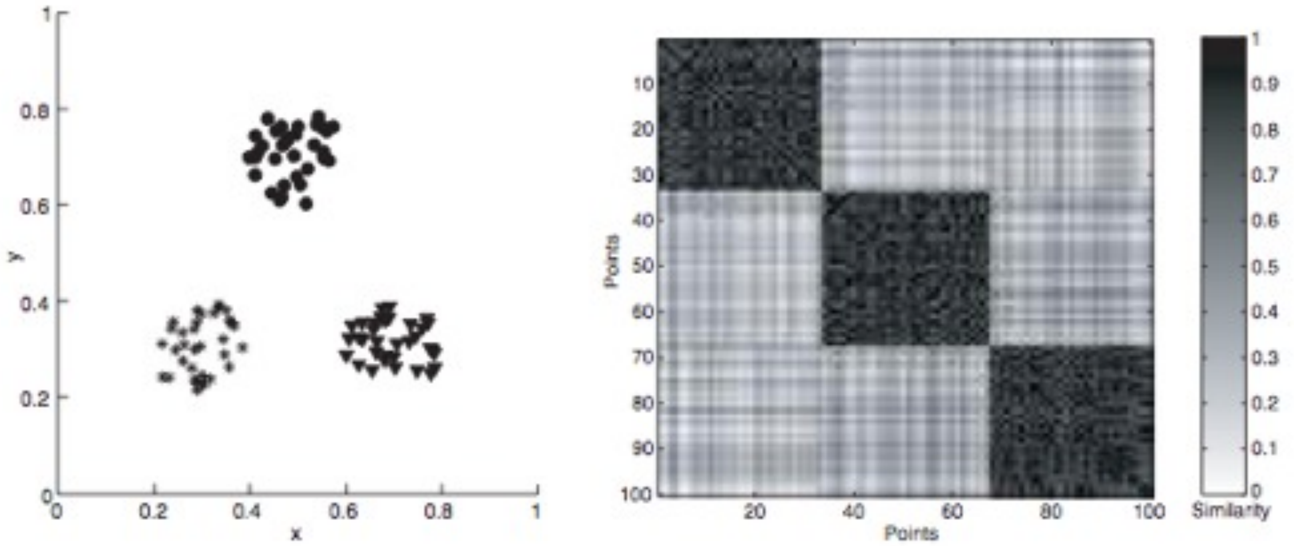
This gives a summary measure of the overall clustering quality.

An alternative validation scheme is given by comparing the similarity matrix with an idealized (0/1) similarity matrix that represents the same clustering configuration.

An alternative validation scheme is given by comparing the similarity matrix with an idealized (0/1) similarity matrix that represents the same clustering configuration.

This can be done either graphically or using correlations.

CLUSTER VALIDATION



(a) Well-separated clusters.

(b) Similarity matrix sorted by K-means cluster labels.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

One useful application of cluster validation is to determine the best number of clusters for your dataset.

One useful application of cluster validation is to determine the best number of clusters for your dataset.

Q: How would you do this?

One useful application of cluster validation is to determine the best number of clusters for your dataset.

Q: How would you do this?

A: By computing the overall SSE or SC for different values of k.

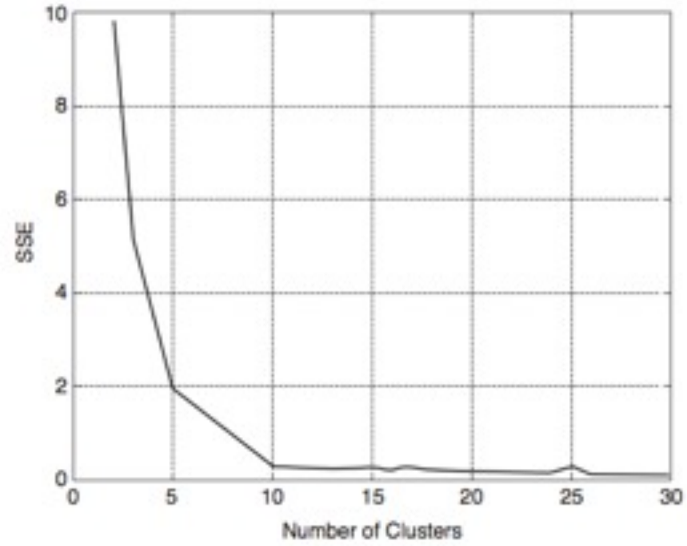


Figure 8.32. SSE versus number of clusters for the data of Figure 8.29.

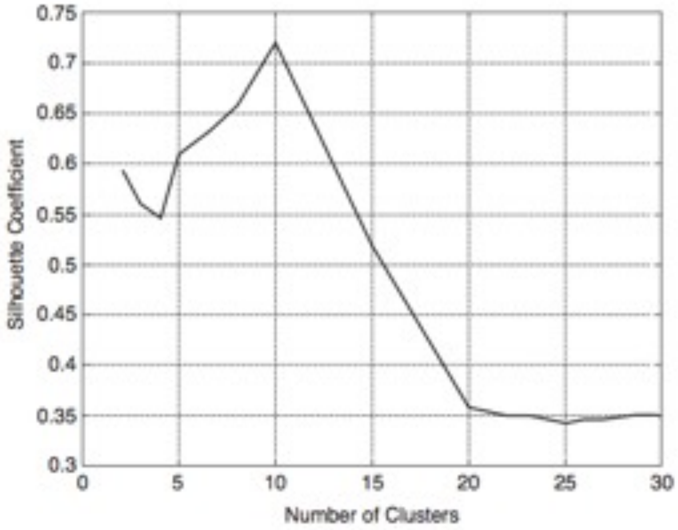


Figure 8.33. Average silhouette coefficient versus number of clusters for the data of Figure 8.29.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Ultimately, cluster validation and clustering in general are suggestive techniques that rely on human interpretation to be meaningful.

EX: K-MEANS CLUSTERING