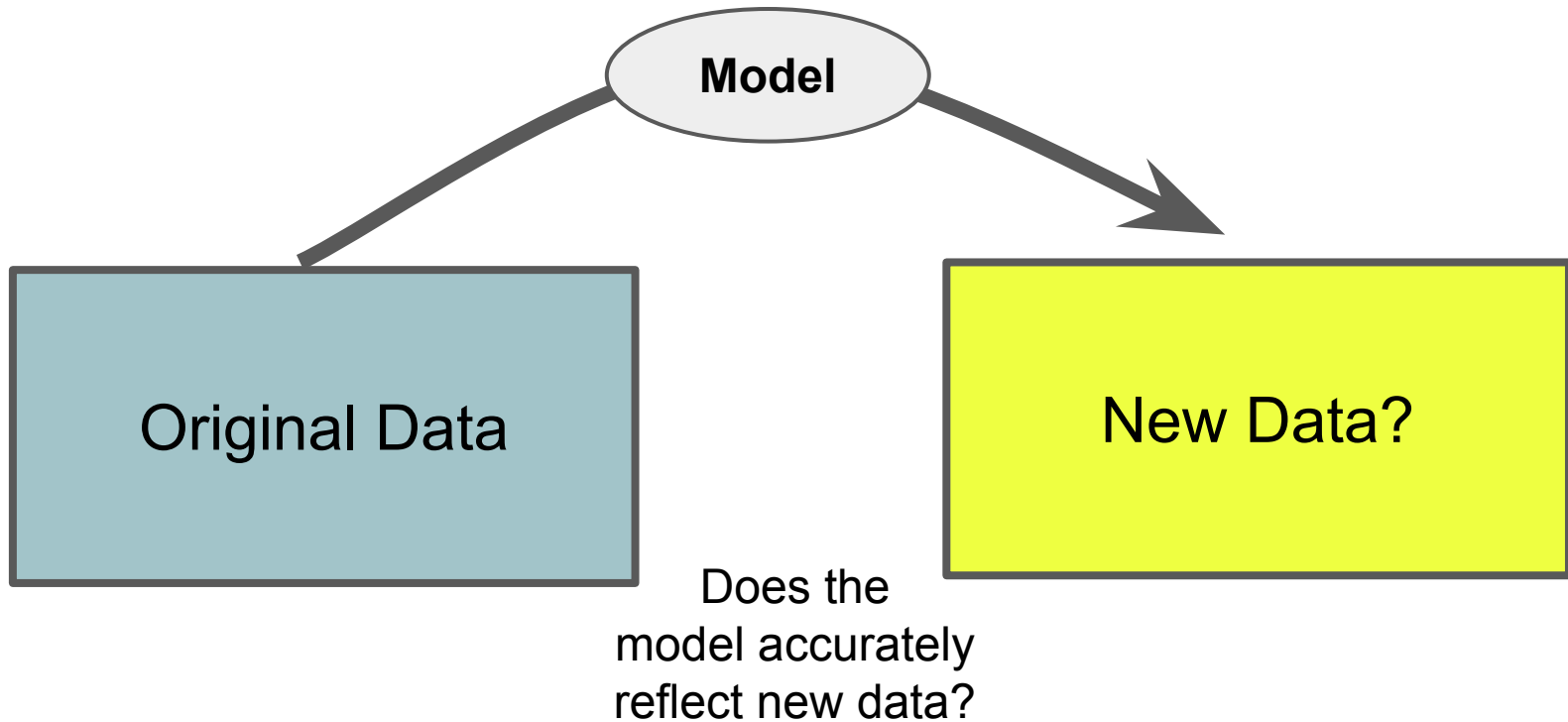


Clustering and Dimensionality Reduction

Stony Brook University
CSE545, Fall 2017

Goal: Generalize to new data



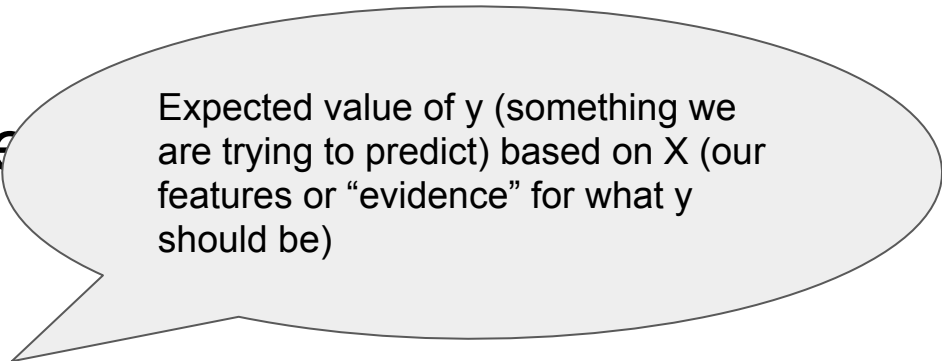
Supervised vs. Unsupervised

Supervised

- Predicting an outcome: $E(y|X)$
- Loss function used to characterize quality of prediction

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Supervised vs. Unsupervised



Expected value of y (something we are trying to predict) based on X (our features or “evidence” for what y should be)

Supervised

- Predicting an outcome:
- Loss function used to characterize quality of prediction

$$E(y|X)$$

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Supervised vs. Unsupervised

Supervised

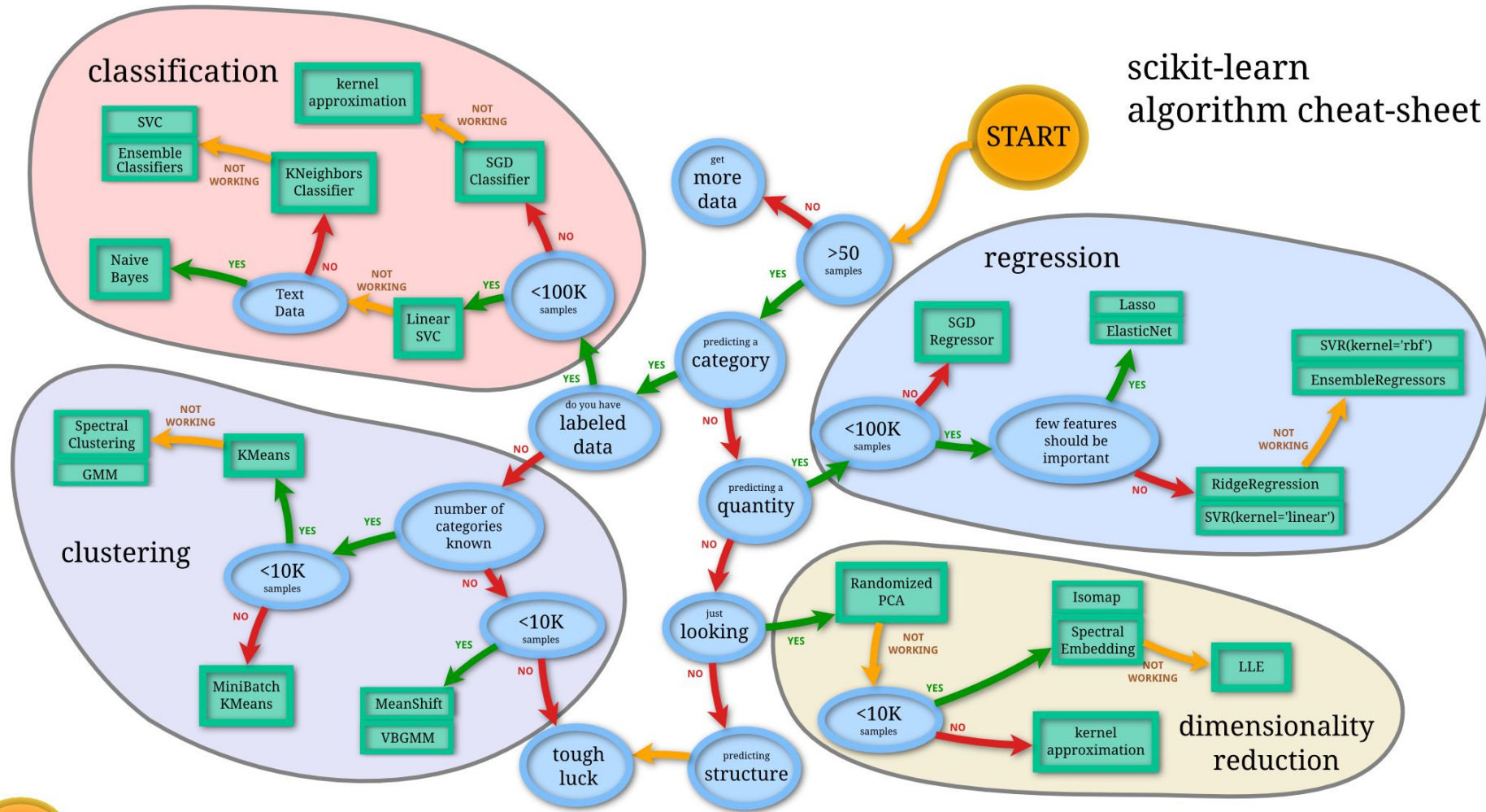
- Predicting an outcome $E(y|X)$
- Loss function used to characterize quality of prediction

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Unsupervised

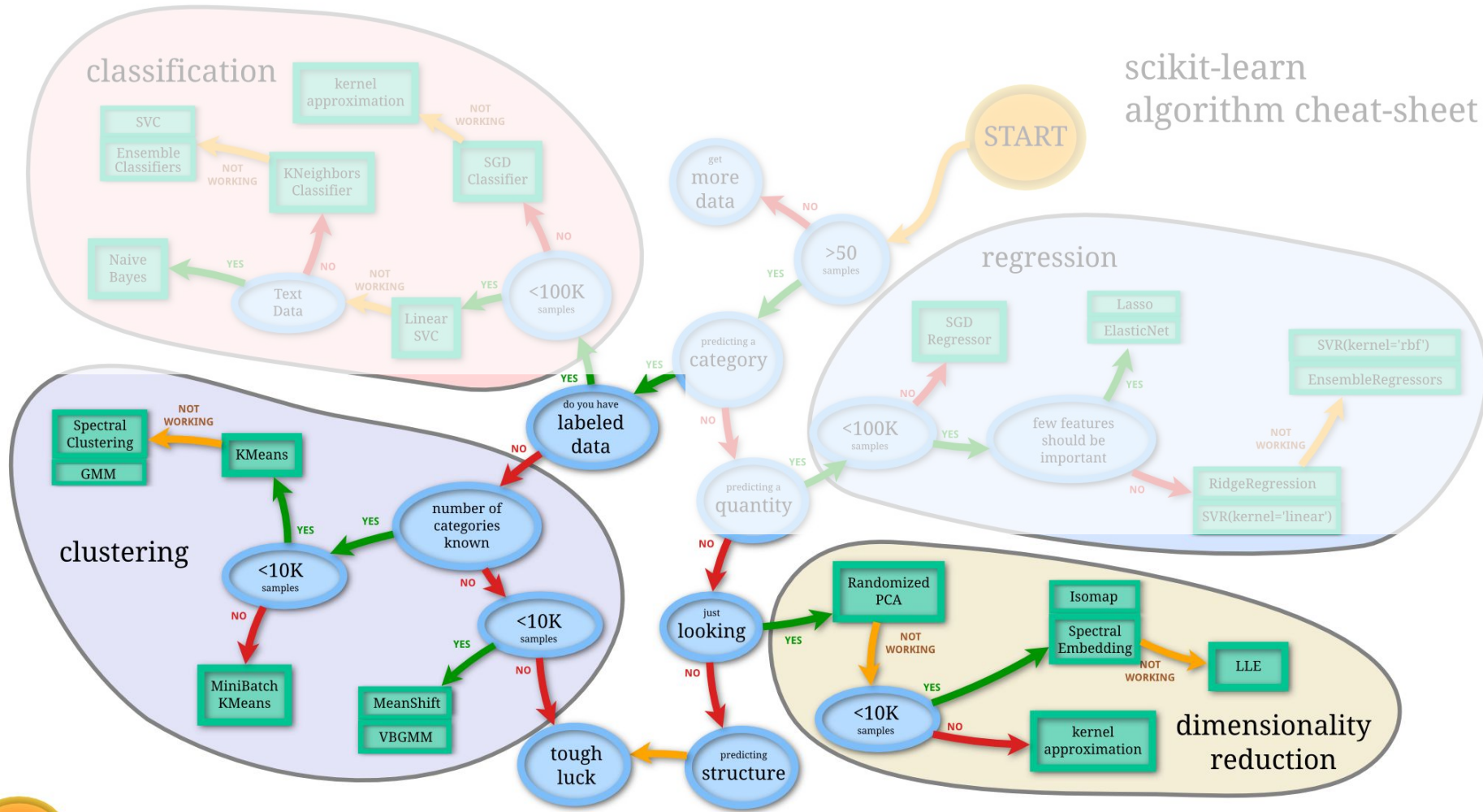
- No outcome to predict
- Goal: Infer properties of X without a supervised loss function.
- Often larger data.
- Don't need to worry about conditioning on another variable.

scikit-learn algorithm cheat-sheet



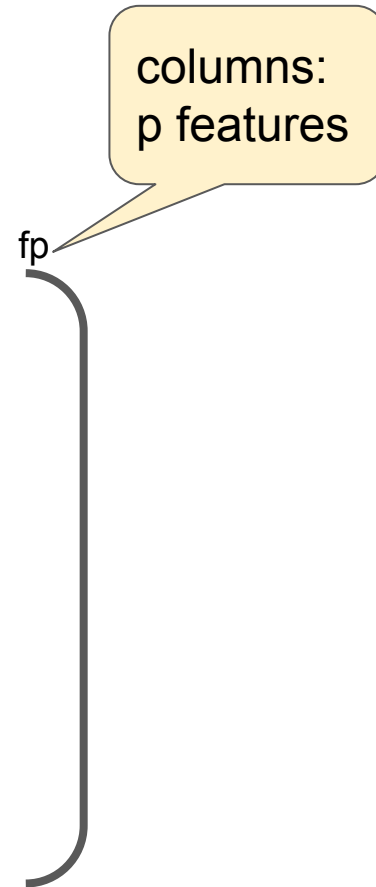
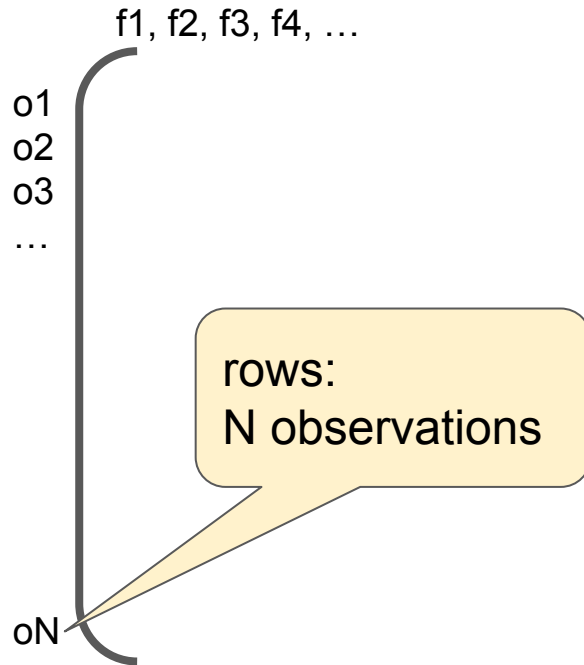
Back

scikit-learn algorithm cheat-sheet



Back

Concept, In Matrix Form:




Concept, In Matrix Form:

	f1, f2, f3, f4, ...																										fp			
o1	1	2	0	2	3	7	5	0	8	4	2	0	6	4	5	2	3	6	7	3	4	2	2	9	2	0	3	9	8	7
o2	1	2	5	3	1	8	9	7	3	1	0	5	4	2	7	8	2	6	9	1	0	4	9	5	3	7	5	9	4	9
o3	1	6	7	9	8	3	7	4	7	2	4	1	5	0	4	8	3	6	7	5	6	2	0	4	1	5	5	4	9	0
...	1	9	5	2	2	0	8	4	9	0	2	6	8	8	1	1	9	4	9	4	5	4	2	4	5	4	8	5	1	7
	1	1	6	9	4	9	6	0	0	4	4	1	8	8	5	4	8	2	5	8	3	2	3	5	4	9	6	3	8	1
	1	4	8	5	0	1	1	9	7	4	3	8	5	4	3	3	5	6	5	5	5	3	3	6	4	2	9	4	3	2
	1	5	5	8	0	9	3	4	1	0	4	9	8	9	6	3	7	6	1	6	4	7	4	0	0	0	7	9	7	1
	1	8	7	2	0	3	6	8	2	4	2	7	6	2	6	3	7	6	1	6	4	7	4	0	0	0	7	9	7	1
	1	7	9	1	4	9	7	1	1	1	7	5	3	0	6	8	2	9	8	2	5	3	1	9	4	0	2	5	5	4
	1	6	5	6	0	7	4	5	1	7	0	3	3	4	0	2	4	3	7	1	7	4	2	6	8	7	7	1	6	8
	2	2	2	3	4	4	1	9	4	3	0	4	0	4	7	5	6	3	2	6	9	0	1	9	4	9	6	3	1	2
	7	0	8	8	7	7	6	3	8	9	0	5	4	6	6	9	1	4	3	7	5	8	6	4	0	4	1	1	4	7
	9	5	2	9	2	8	5	5	2	9	9	8	4	2	5	5	7	9	7	2	2	6	7	2	0	8	6	5	0	9
	7	0	0	5	8	9	8	6	7	8	1	5	3	8	9	4	6	9	5	0	9	6	0	5	4	2	2	3	0	7
	9	1	7	0	6	9	8	5	9	7	6	6	9	7	2	0	6	6	9	6	3	7	7	1	8	8	1	6	9	9
	1	0	1	1	4	8	9	5	7	8	1	1	5	1	4	0	8	5	4	7	8	1	5	0	9	5	8	6	5	1
oN	1	5	5	2	2	7	2	3	1	9	2	0	5	6	5	2	3	8	1	5	3	1	5	4	0	9	5	9	5	5

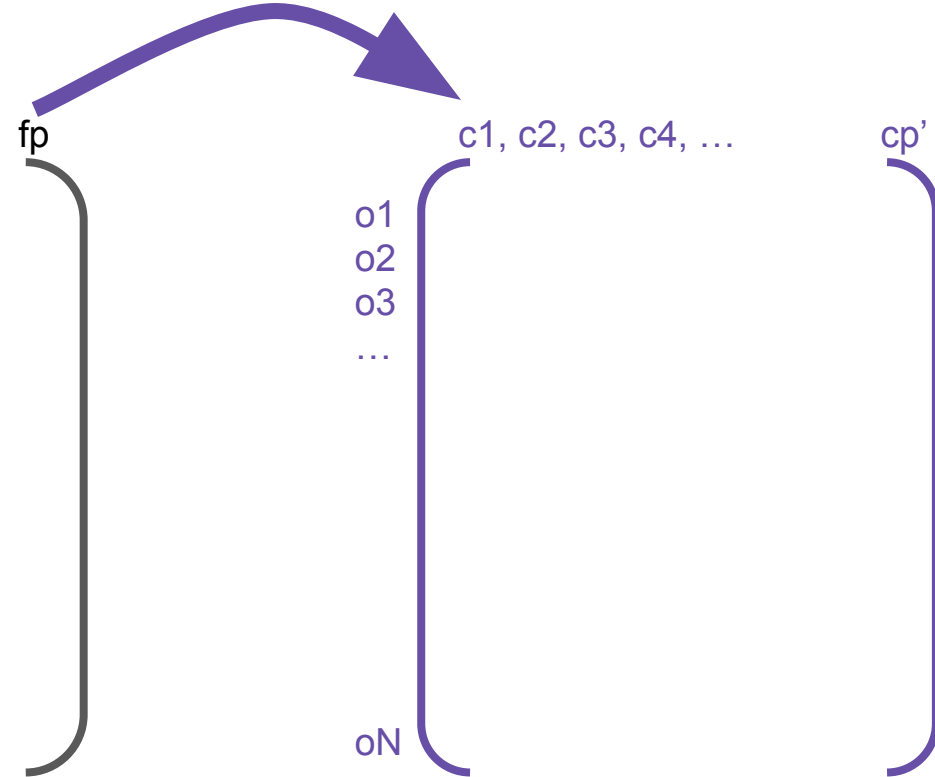
Concept, In Matrix Form:

f1, f2, f3, f4, ...

o1
o2
o3
...
oN

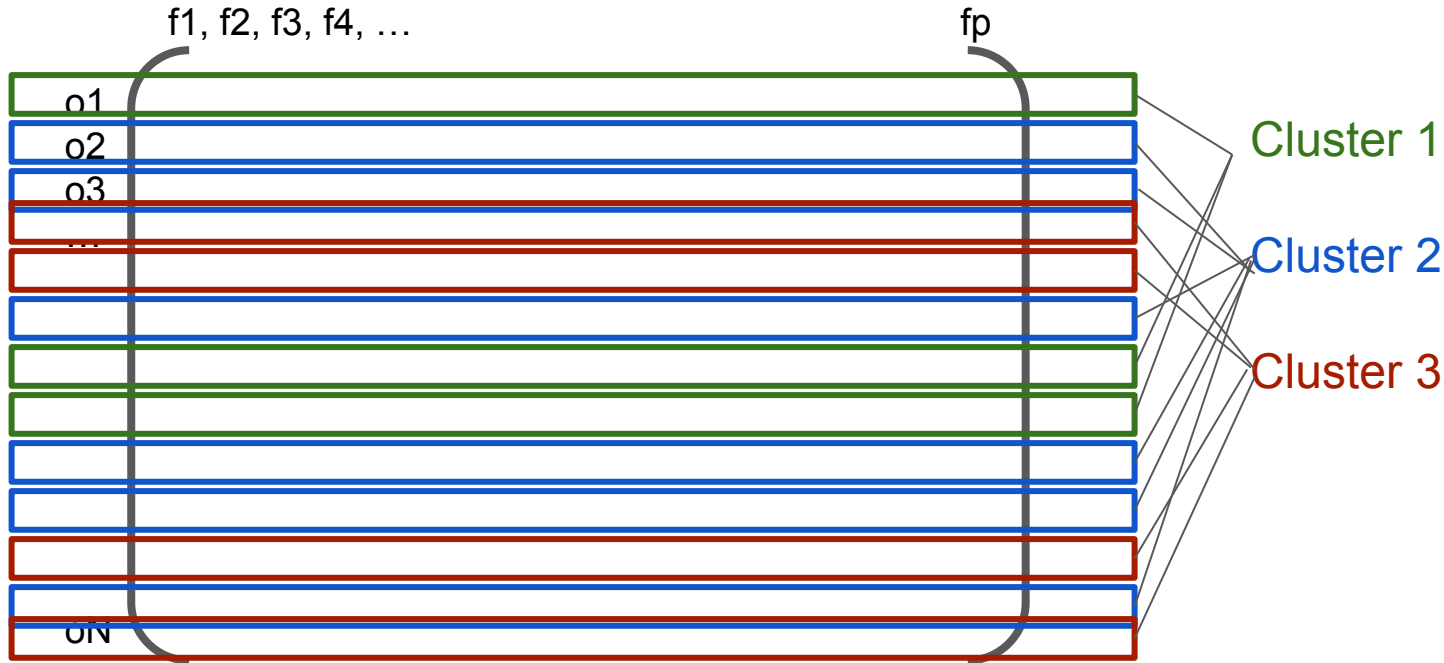


Dimensionality reduction
Try to best represent but with on p' columns.

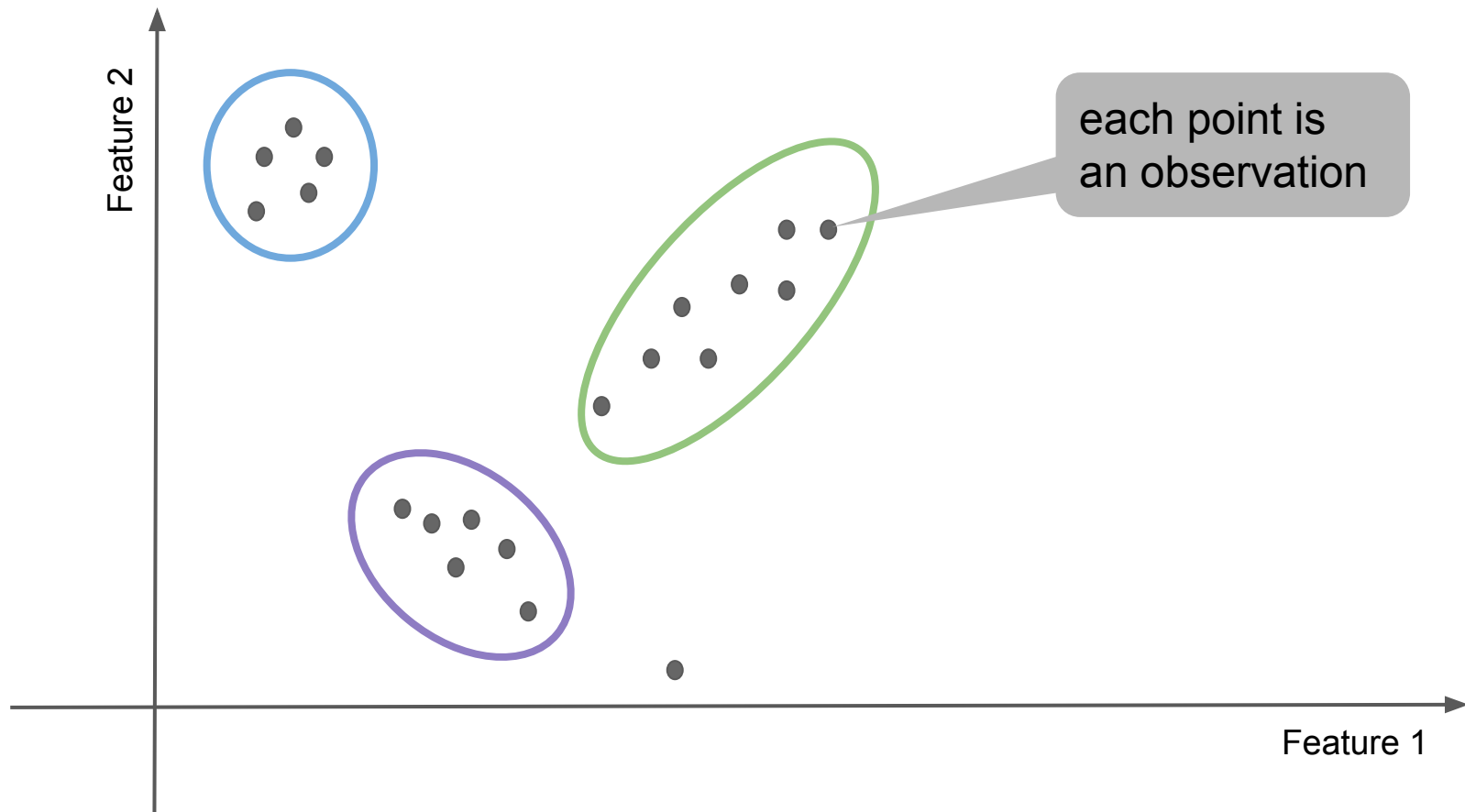


Concept, In Matrix Form:

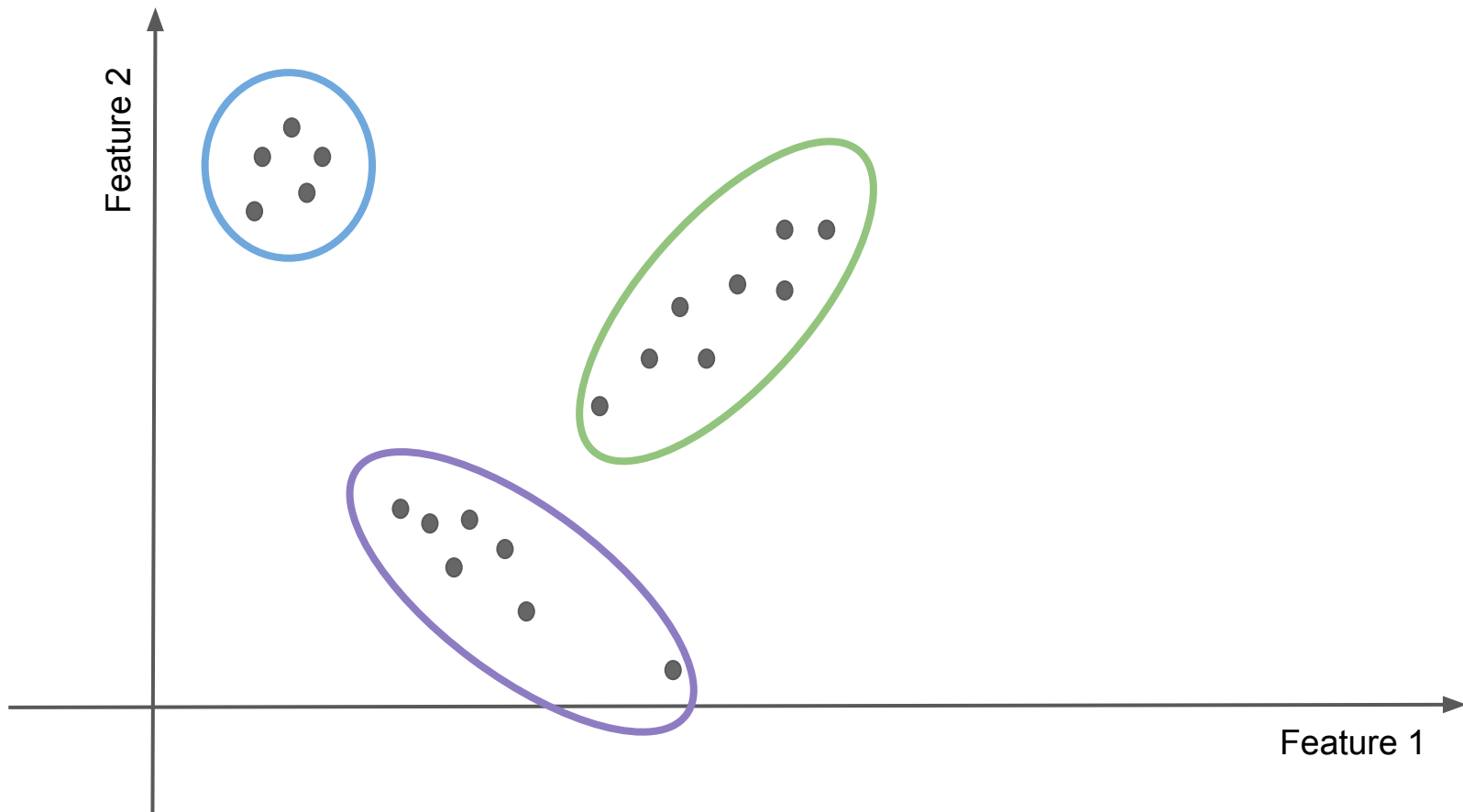
Clustering: Group observations based on the features (i.e. like reducing the number of observations into K groups).



Concept: in 2-d (clustering)



Concept: in 2-d (clustering)



Clustering

Typical formalization:

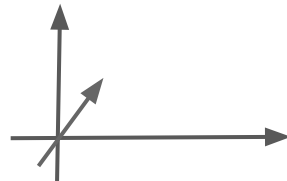
Given:

- set of points
- distance metric (Euclidean, cosine, etc...)
- number of clusters (not always provided)

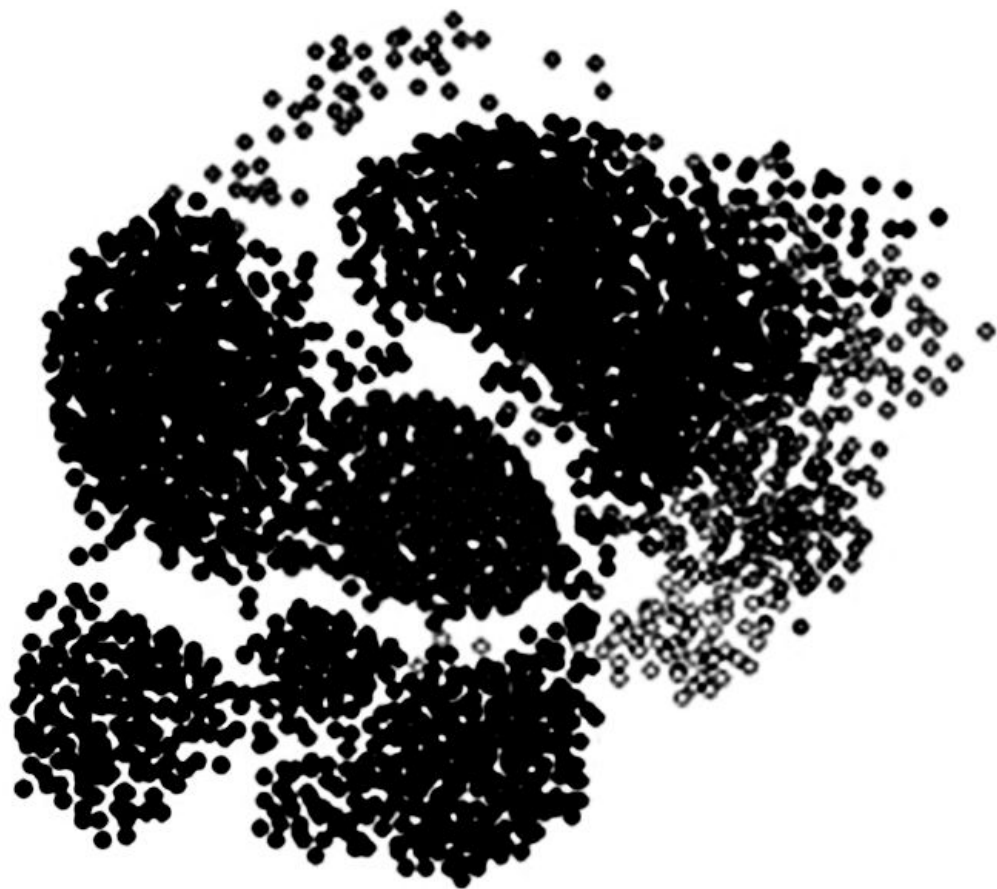
Do: Group observations together that are similar. Ideally,

- Members of same cluster are the “same”.
- Members of different clusters are “different”.

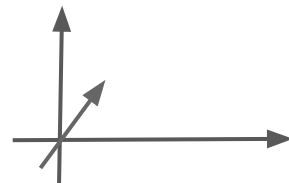
Keep in mind: usually many more than 2 dimensions.



Clustering



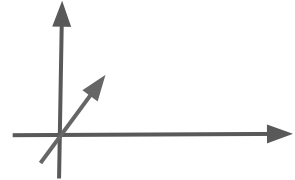
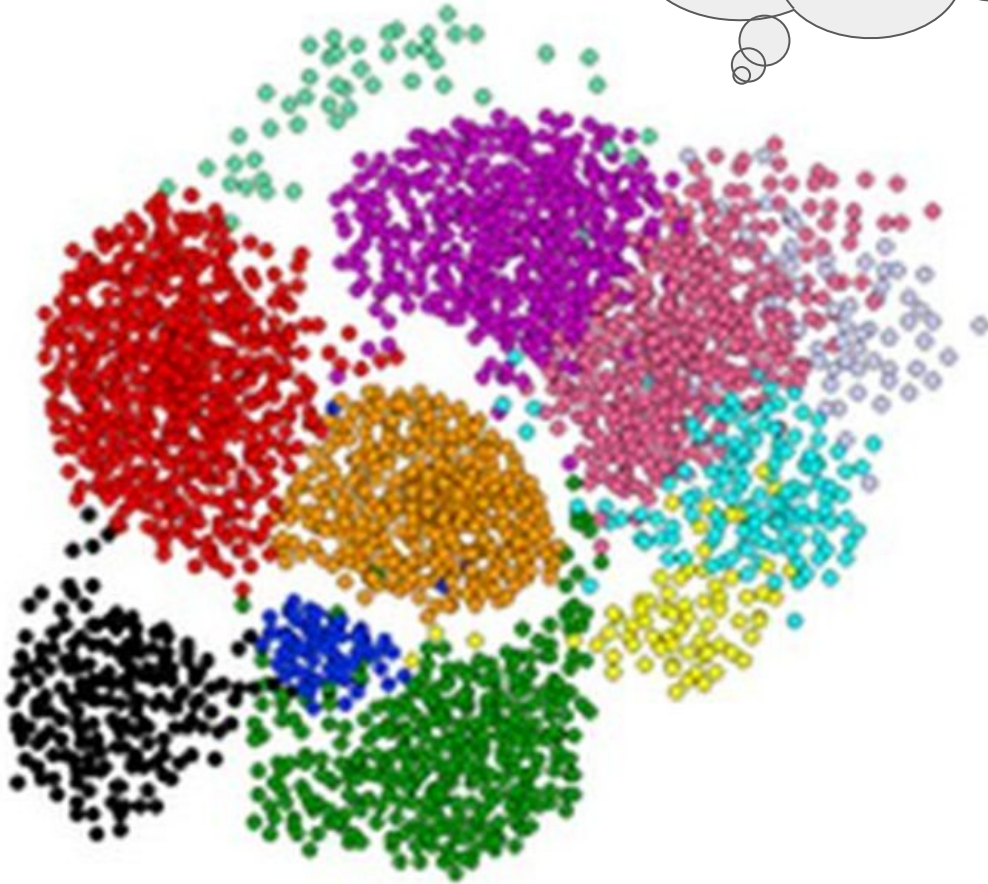
Often many dimensions and
no clean separation.



Clustering

Supposes
observations have a
“true” cluster.

Often many dimensions and
no clean separation.



K-Means Clustering

Clustering: Group similar observations, often over unlabeled data.

K-means: A “prototype” method
(i.e. not based on an algebraic model).

Euclidean Distance:
$$d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^m (x_{ij} - x_{i'j})^2} = ||x_i - x_{i'}||$$

K-Means Clustering

Clustering: Group similar observations, often over unlabeled data.

K-means: A “prototype” method
(i.e. not based on an algebraic model).

Euclidean Distance:
$$d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^m (x_{ij} - x_{i'j})^2} = ||x_i - x_{i'}||$$

centers = a random selection of k cluster centers
until centers converge:

1. For all x_i , find the closest center (according to d)
2. Recalculate centers based on mean of euclidean distance

K-Means Clustering

Clustering: Group similar observations, often over unlabeled data.

K-means: A “prototype” method
(i.e. not based on an algebraic model).

Euclidean Distance:
$$d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^m (x_{ij} - x_{i'j})^2} = ||x_i - x_{i'}||$$

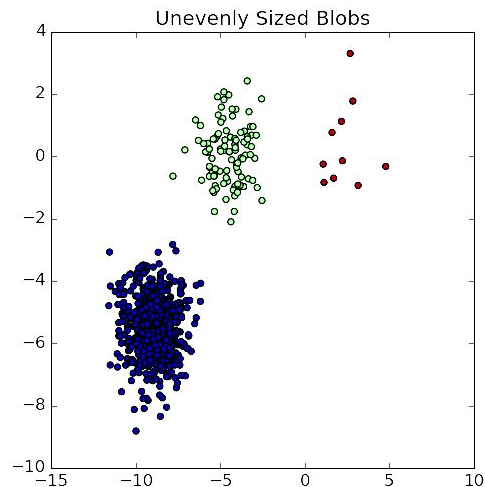
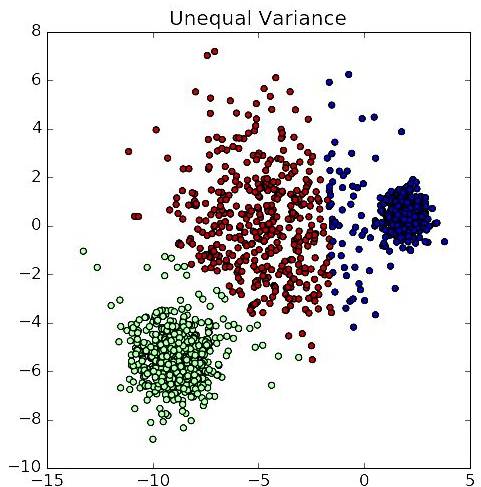
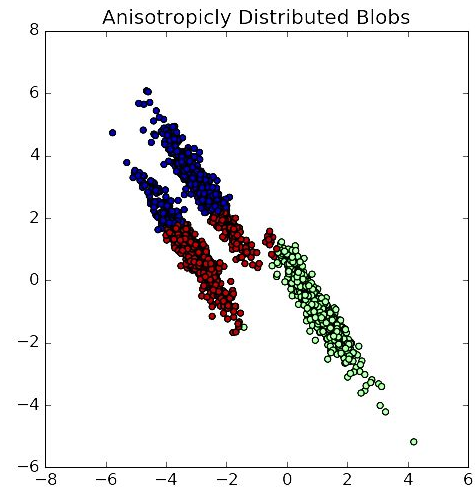
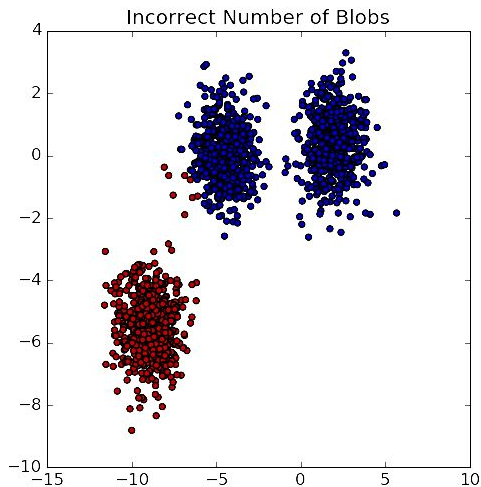
centers = a random selection of k cluster centers
until centers converge:

1. For all x_i , find the closest center (according to d)
2. Recalculate centers based on mean of euclidean distance

Example: <http://shabal.in/visuals/kmeans/6.html>

K-Means Clustering

Understanding K-Means

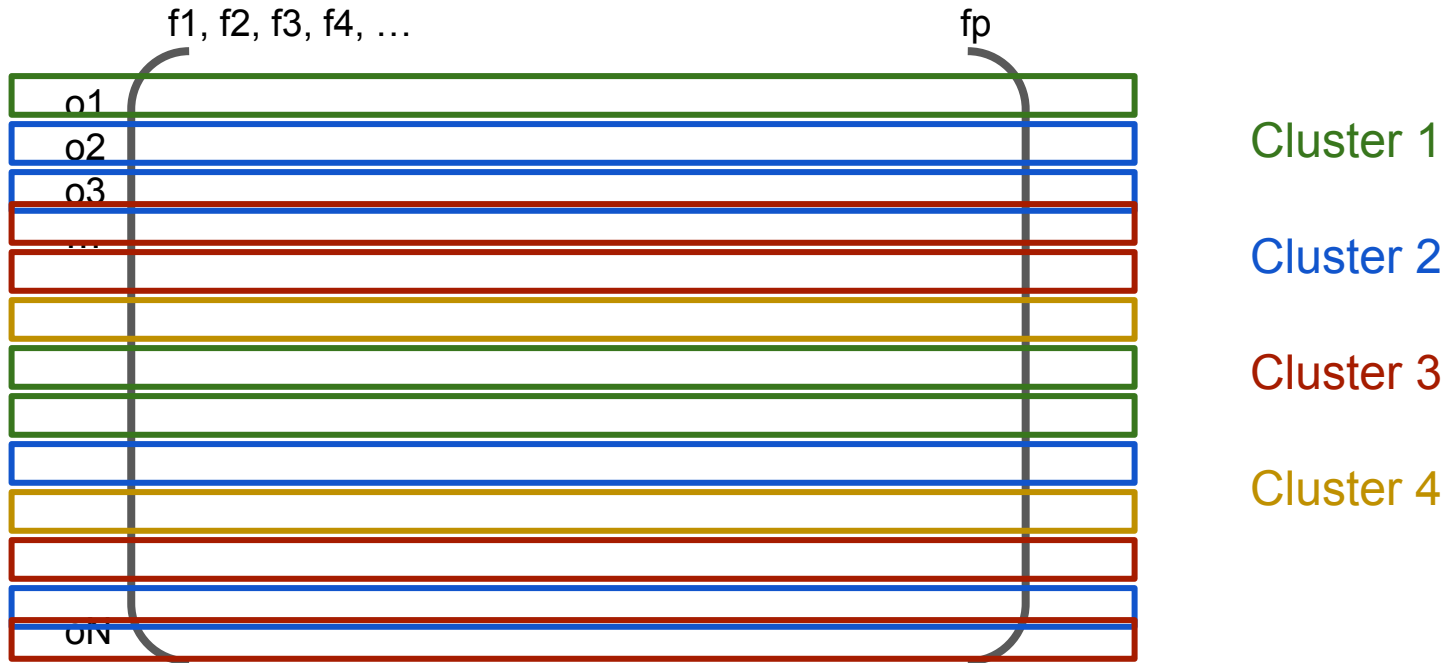


The Curse of Dimensionality

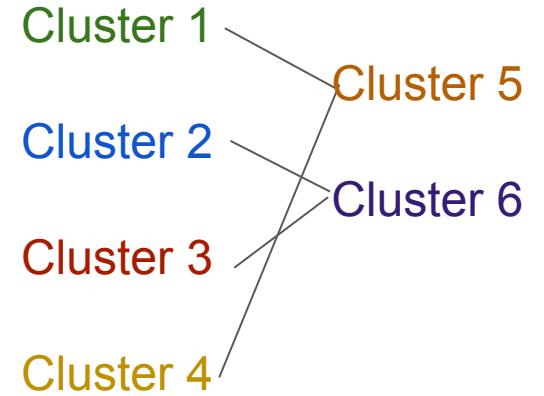
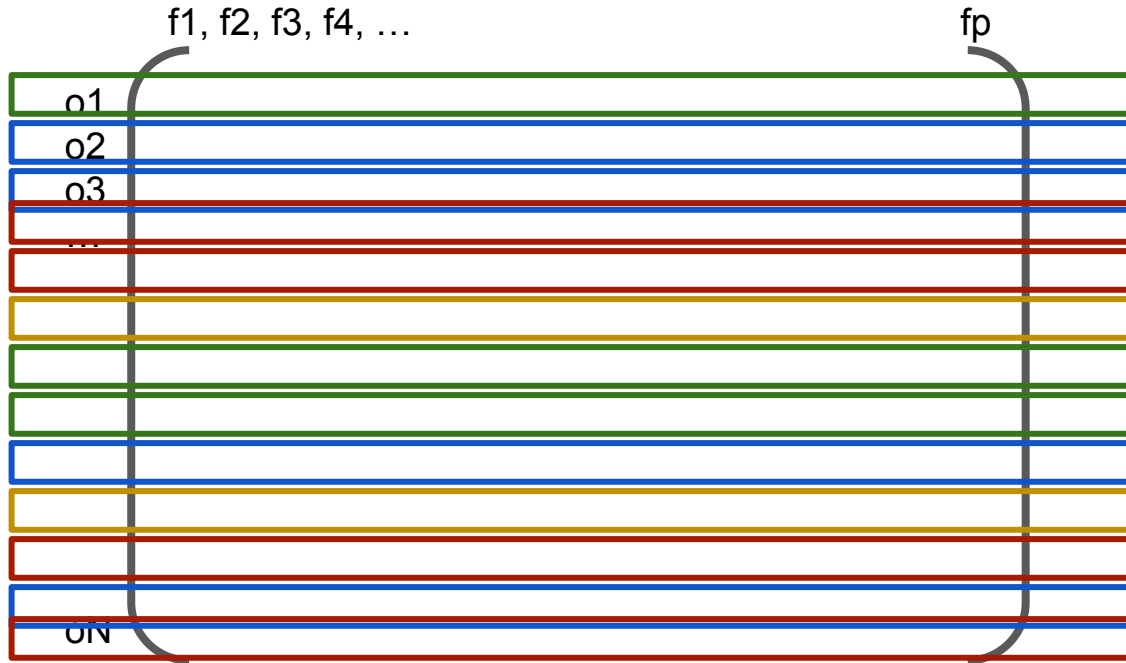
Problems with high-dimensional spaces:

1. All points (i.e. observations) are nearly equally far apart.
2. The angle between vectors are almost always 90 degrees (i.e. they are orthogonal).

Hierarchical Clustering



Hierarchical Clustering



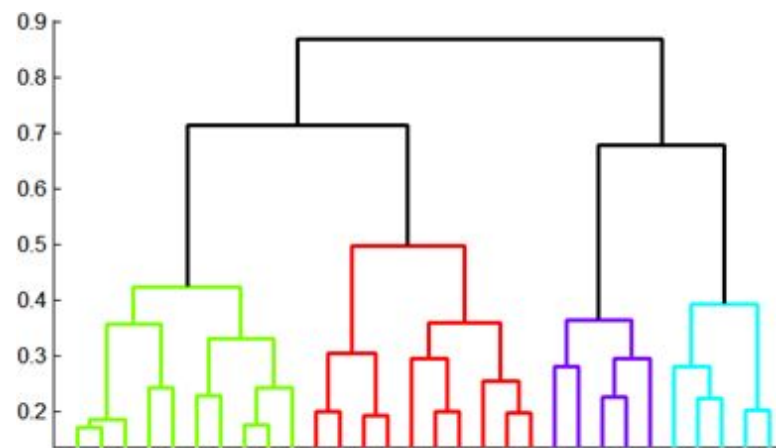
Hierarchical Clustering

- **Agglomerative** (bottom up):

- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one

- **Divisive** (top down):

- Start with one cluster and recursively split it



Hierarchical Clustering

- **Agglomerative** (bottom up):

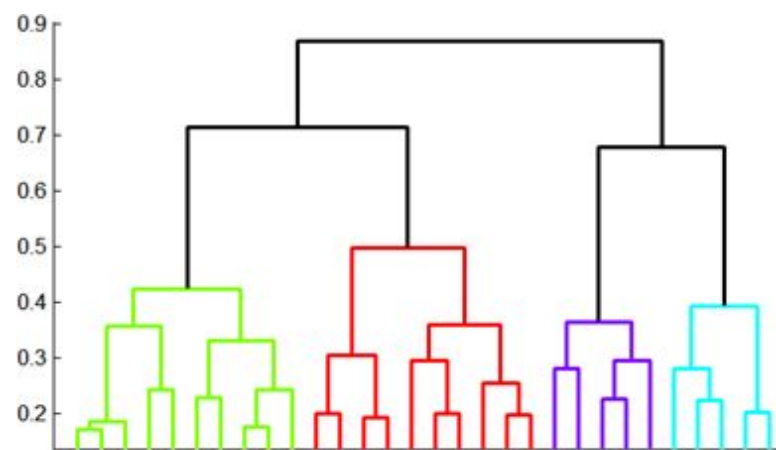
- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one

- **Divisive** (top down):

- Start with one cluster and recursively split it

- **Regular K-Means is
“Point assignment clustering”:**

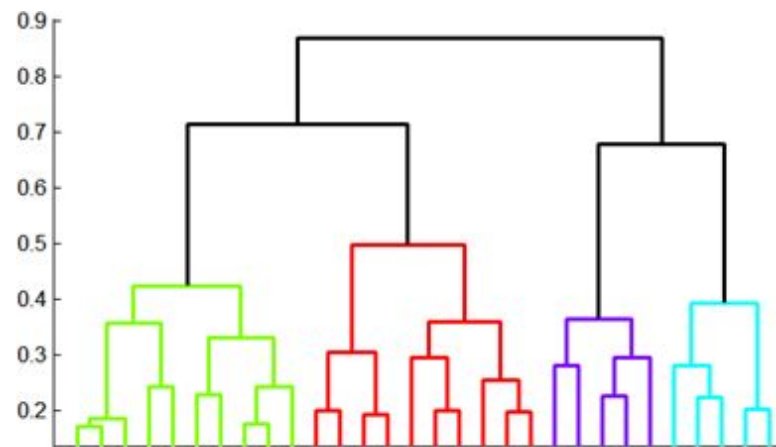
- Maintain a set of clusters
- Points belong to “nearest” cluster



Hierarchical Clustering

- **Agglomerative** (bottom up):

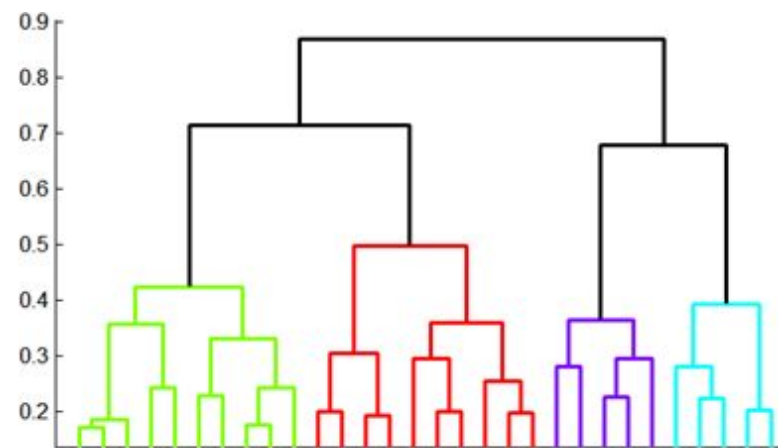
- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one



Hierarchical Clustering

- **Agglomerative** (bottom up):

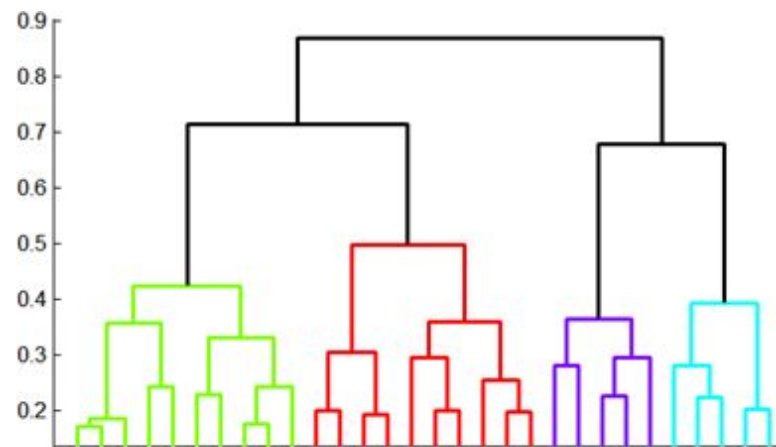
- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one
- Stop when reaching a threshold in
 - Distance between points in cluster, or
 - Maximum distance of points from “center”
 - Maximum number of points



Hierarchical Clustering

- **Agglomerative** (bottom up):

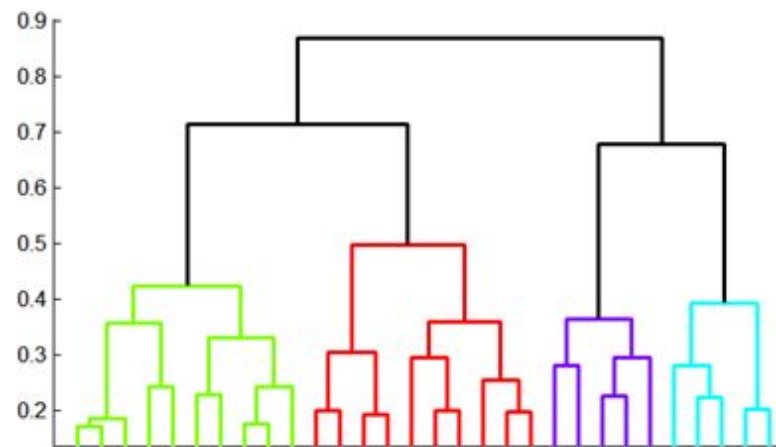
- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one
- Stop when reaching a threshold in
 - Distance between points in cluster, or
 - Maximum distance from “center”
 - Maximum number of points



Hierarchical Clustering

- **Agglomerative** (bottom up):

- Initially, each point is a cluster
- Repeatedly combine the two “nearest” clusters into one



But what if we have no “centroid”?
(such as when using cosine distance)

Clustering: Applications



Clustering: Applications



Excellent treatment. The nursing staff, both day and night, is absolutely first rate. My wife is an experienced RN and applauds the staff for its professionalism. The room techs are well-qualified and quite pleasant. The docs visit regularly and communicate clearly. The C-T, MRI and EEG techs are highly skilled. The occupational, physical and speech therapists are personable and extremely competent. The food is passable but nobody stays at a hospital for either food or rest. The only glitch was the ER. At 4PM, the initial admit and blood tests were handled expeditiously and skillfully. After that it was downhill. At 10PM it was determined I'd be admitted. At 3AM I was finally moved to a room and only then because I demanded directions to the room with the intent of walking there myself. In sum, Washington Hospital itself is worthy of 5 stars. This is the place if you need a hospital in DC.

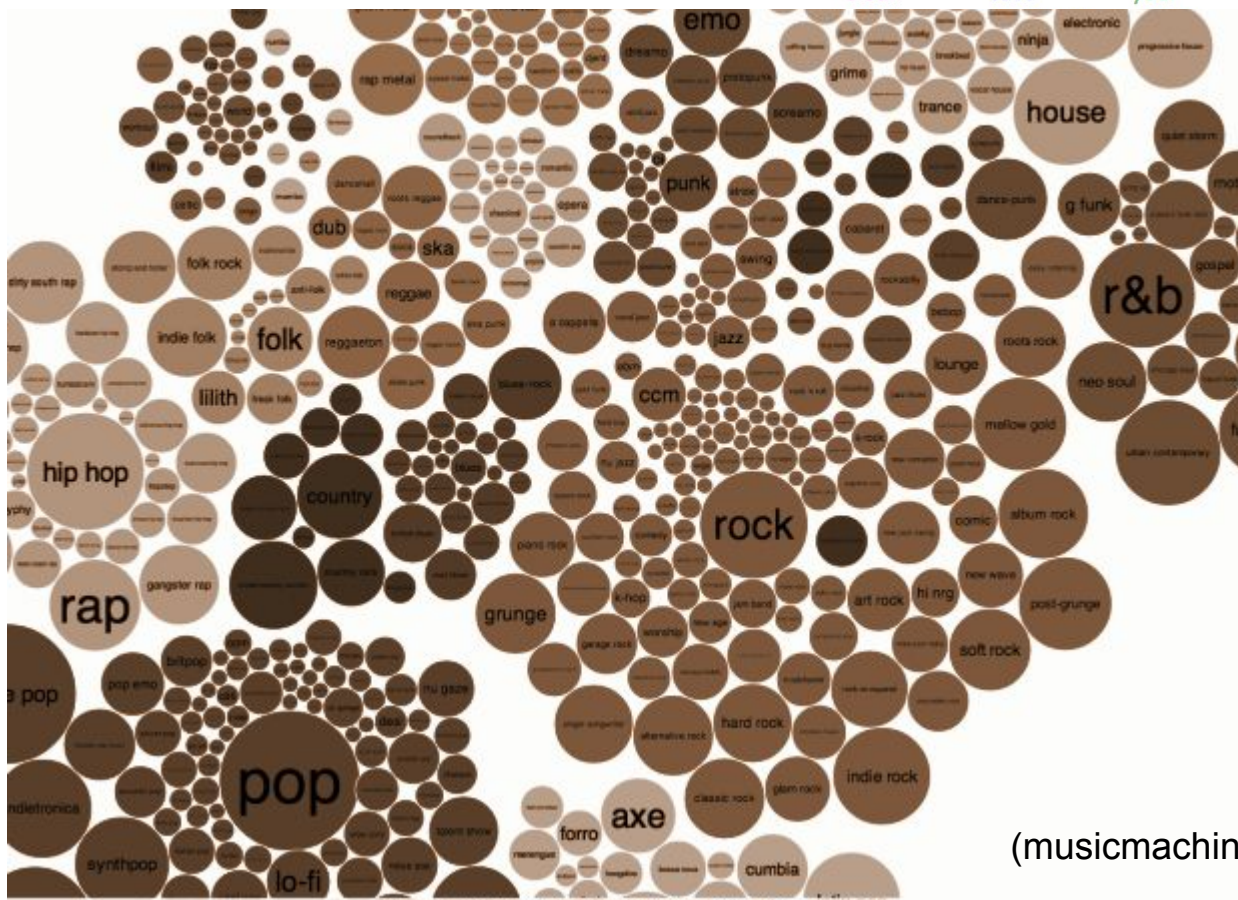
Clustering: Application



24	27	30	36	42	47	48
hospital	patient	dr	care	nurses	hospital	er
staff	medical	physical	staff	day	hospitals	staff
great	hospital	back	nurses	nurse	doctors	friendly
nice	information	surgery	hospital	night	city	emergency
nurses	treatment	mri	doctors	room	medical	visit
friendly	physician	therapy	great	hours	area	room
clean	case	year	caring	hospital	queens	experience
good	received	weeks	excellent	stay	live	professional
place	review	knee	experience	days	methodist	wait
helpful	told	therapist	received	husband	good	quick
doctors	condition	brain	wonderful	admitted	center	nice
super	staff	replacement	made	call	top	quickly
experience	lack	found	professional	time	place	efficient
people	health	great	kind	didnt	hill	pleasant
caring	due	left	helpful	morning	mt	night

Excellent treatment. The nursing staff, both day and night, is absolutely first rate. My wife is an experienced RN and applauds the staff for its professionalism. The room techs are well-qualified and quite pleasant. The docs visit regularly and communicate clearly. The C-T, MRI and EEG techs are highly skilled. The occupational, physical and speech therapists are personable and extremely competent. The food is passable but nobody stays at a hospital for either food or rest. The only glitch was the ER. At 4PM, the initial admit and blood tests were handled expeditiously and skillfully. After that it was downhill. At 10PM it was determined I'd be admitted. At 3AM I was finally moved to a room and only then because I demanded directions to the room with the intent of walking there myself. In sum, Washington Hospital itself is worthy of 5 stars. This is the place if you need a hospital in DC.

Clustering: Applications

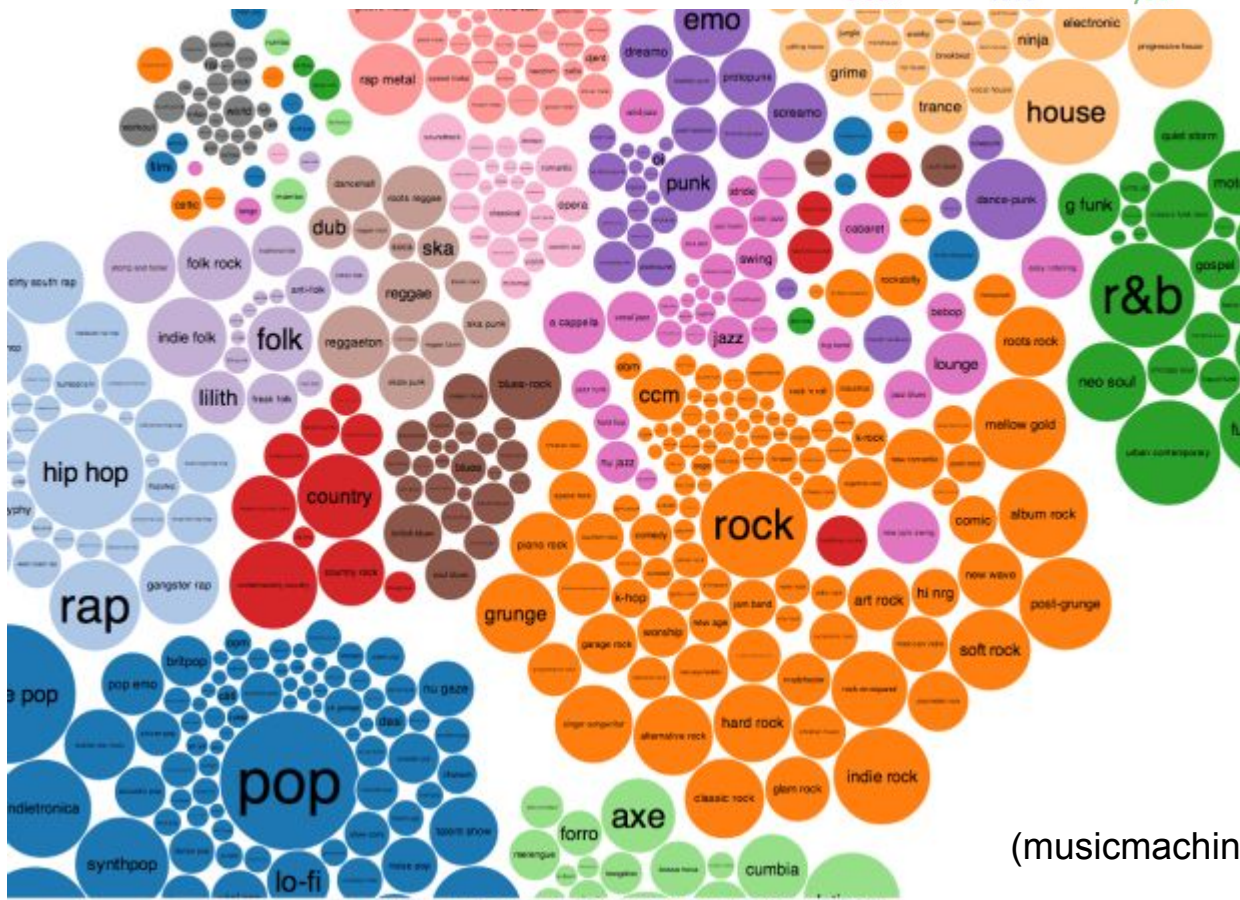



24	27	30	36	42	47	48
hospital	patient	dr	care	nurses	hospital	er
staff	medical	physical	staff	day	hospitals	staff
great	hospital	back	nurses	nurse	doctors	friendly
nice	information	surgery	hospital	night	city	emergency
nurses	treatment	mri	doctors	room	medical	visit
friendly	physician	therapy	great	hours	area	room
clean	case	year	caring	hospital	queens	experience
			excellent	stay	live	professional
			experience	days	methodist	wait
			received	husband	good	quick
			wonderful	admitted	center	nice
			made	call	top	quickly
			professional	time	place	efficient
			kind	didn't	hill	pleasant
			helpful	morning	mt	night

g staff, both day and night, is absolutely first
RN and applauds the staff for its
s are well-qualified and quite pleasant. The
hicate clearly. The C-T, MRI and EEG techs are
physical and speech therapists are
etent. The food is passable but nobody stays
est. The only glitch was the ER. At 4PM, the
re handled expeditiously and skillfully. After
was determined I'd be admitted. At 3AM I
d only then because I demanded directions to
king there myself. In sum, Washington
rs. This is the place if you need a hospital in

(musicmachinery.com)

Clustering: Applications

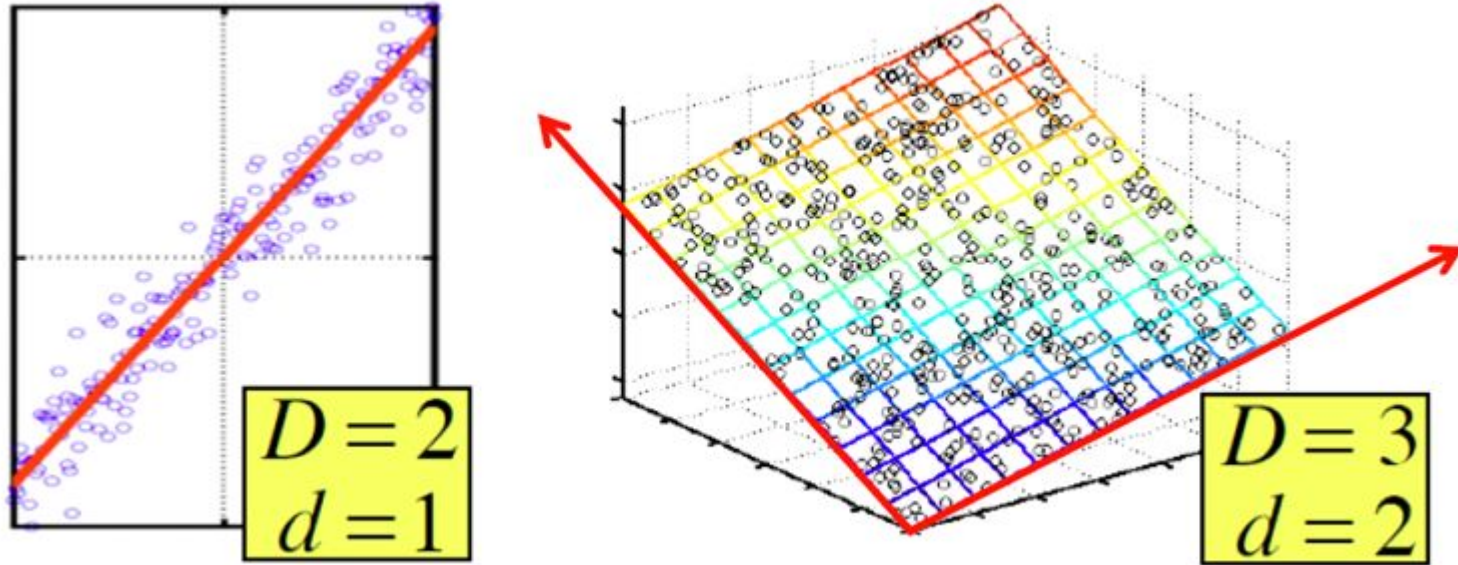


24	27	30	36	42	47	48
hospital	patient	dr	care	nurses	hospital	er
staff	medical	physical	staff	day	hospitals	staff
great	hospital	back	nurses	nurse	doctors	friendly
nice	information	surgery	hospital	night	city	emergency
nurses	treatment	mri	doctors	room	medical	visit
friendly	physician	therapy	great	hours	area	room
clean	case	year	caring	hospital	queens	experience
			excellent	stay	live	professional
			experience	days	methodist	wait
			received	husband	good	quick
			wonderful	admitted	center	nice
			made	call	top	quickly
			professional	time	place	efficient
			kind	didn't	hill	pleasant
			helpful	morning	mt	night

g staff, both day and night, is absolutely first
RN and applauds the staff for its
s are well-qualified and quite pleasant. The
nicate clearly. The C-T, MRI and EEG techs are
physical and speech therapists are
etent. The food is passable but nobody stays
est. The only glitch was the ER. At 4PM, the
re handled expeditiously and skillfully. After
was determined I'd be admitted. At 3AM I
d only then because I demanded directions to
king there myself. In sum, Washington
rs. This is the place if you need a hospital in

(musicmachinery.com)

Concept: Dimensionality Reduction in 3-D, 2-D, and 1-D




Data (or, at least, what we want from the data) may be accurately represented with less dimensions.

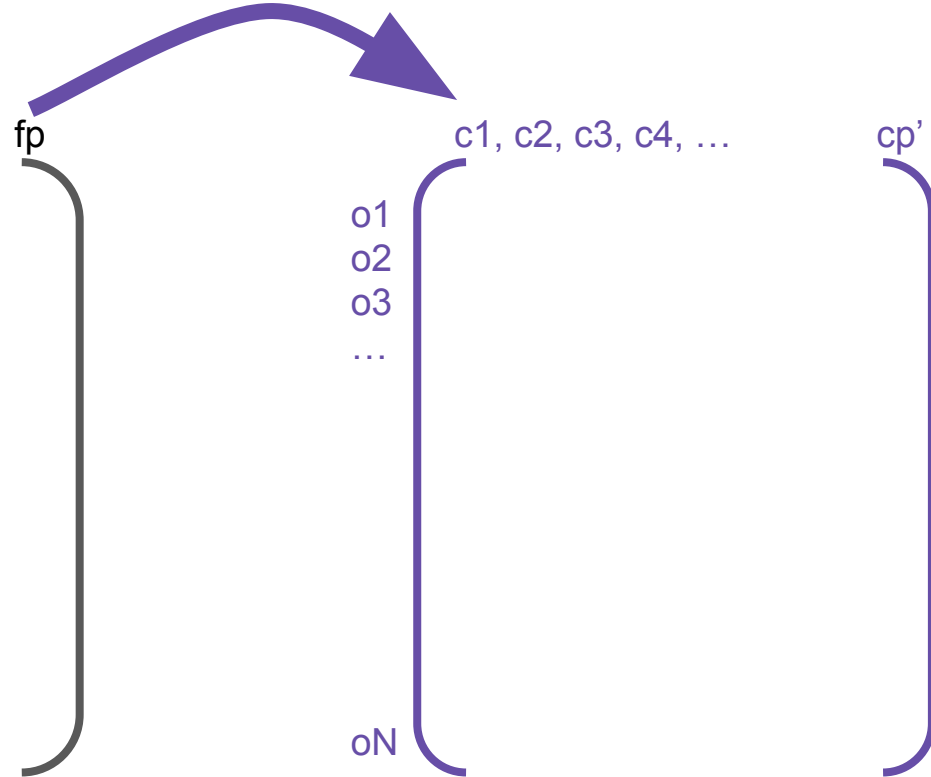
Concept, In Matrix Form:

f1, f2, f3, f4, ...

o1
o2
o3
...
oN



Dimensionality reduction
Try to best represent but with on p' columns.



Dimensionality Reduction

Rank: Number of linearly independent columns of A.
(i.e. columns that can't be derived from the other columns through addition).

Q: What is the rank of this matrix?

$$\begin{pmatrix} 1 & -2 & 3 \\ 2 & -3 & 5 \\ 1 & 1 & 0 \end{pmatrix}$$

Dimensionality Reduction

Rank: Number of linearly independent columns of A.
(i.e. columns that can't be derived from the other columns).

Q: What is the rank of this matrix?

$$\begin{pmatrix} 1 & -2 & 3 \\ 2 & -3 & 5 \\ 1 & 1 & 0 \end{pmatrix}$$

A: 2. The 1st is just the sum of the second two columns

... we can represent as linear combination of 2 vectors:

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} -2 \\ -3 \\ 1 \end{pmatrix}$$

Dimensionality Reduction - PCA

Linear approximates of data in r dimensions.

Found via *Singular Value Decomposition*:

$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

X: original matrix,

U: “left singular vectors”,

D: “singular values” (diagonal),

V: “right singular vectors”

Dimensionality Reduction - PCA

Linear approximates of data in r dimensions.

Found via *Singular Value Decomposition*:

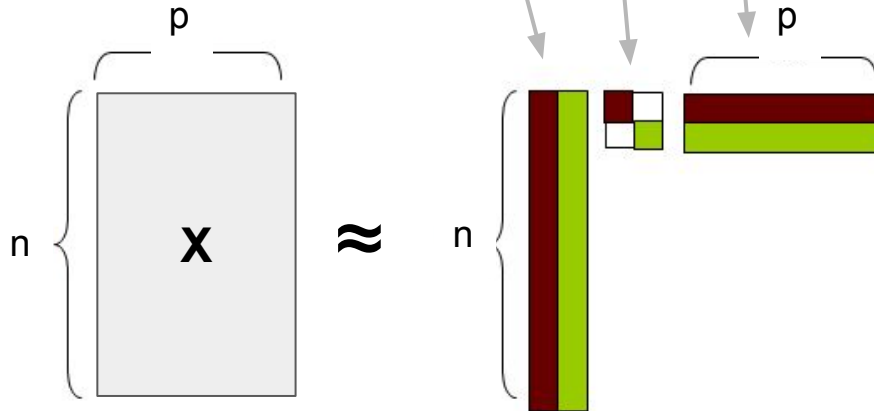
$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

X : original matrix,

D : “singular values” (diagonal),

U : “left singular vectors”,

V : “right singular vectors”



Dimensionality Reduction - PCA - Example

$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

Users to movies matrix

SciFi

Romance

	Matrix	Alien	Serenity	Casablanca	Amelie
1	1	1	0	0	0
3	3	3	0	0	0
4	4	4	0	0	0
5	5	5	0	0	0
0	2	0	4	4	4
0	0	0	5	5	5
0	1	0	2	2	2

=

0.13	0.02	-0.01
0.41	0.07	-0.03
0.55	0.09	-0.04
0.68	0.11	-0.05
0.15	-0.59	0.65
0.07	-0.73	-0.67
0.07	-0.29	0.32

x

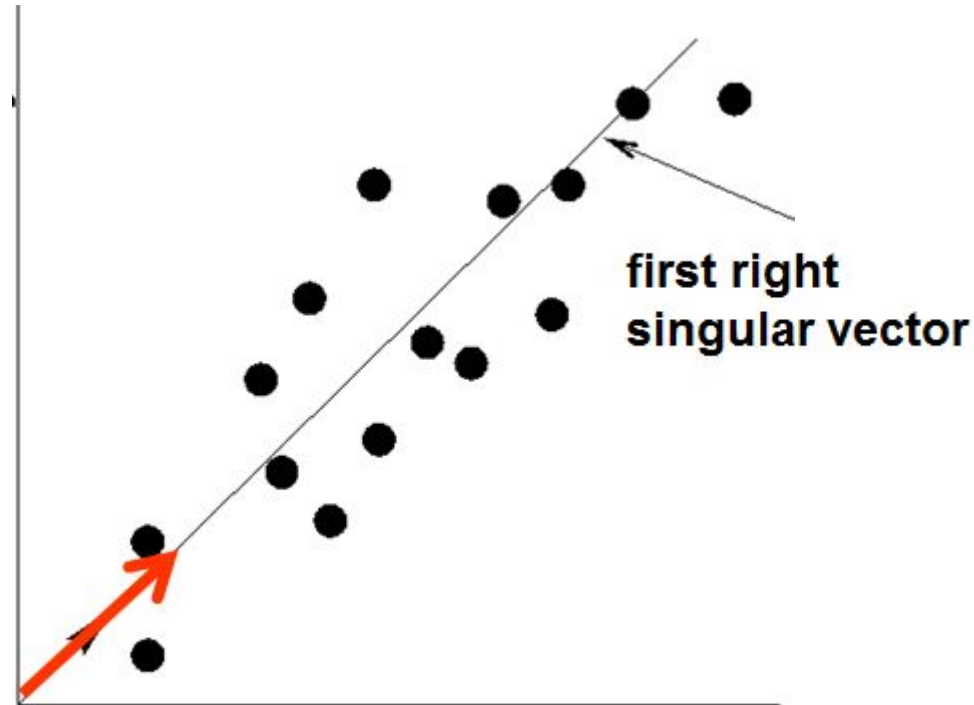
12.4	0	0
0	9.5	0
0	0	1.3

x

0.56	0.59	0.56	0.09	0.09
0.12	-0.02	0.12	-0.69	-0.69
0.40	-0.80	0.40	0.09	0.09

Dimensionality Reduction - PCA - Example

$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$



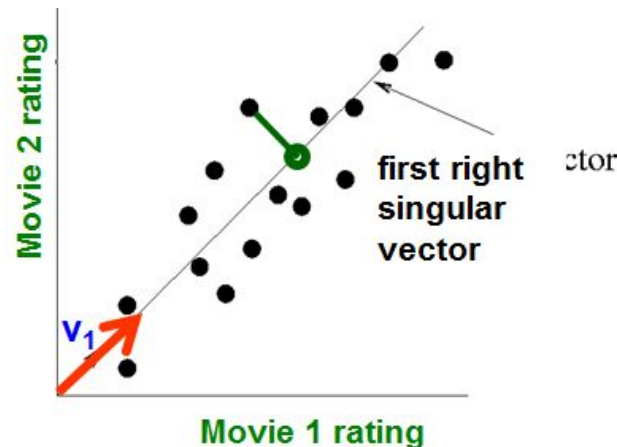
Dimensionality Reduction - PCA - Example

- **Goal:** Minimize the sum of reconstruction errors:

$$\sum_{i=1}^N \sum_{j=1}^D \|x_{ij} - z_{ij}\|^2$$

- where x_{ij} are the “old” and z_{ij} are the “new” coordinates

- **SVD gives ‘best’ axis to project on:**
 - ‘best’ = minimizing the reconstruction errors
- In other words, **minimum reconstruction error**



Dimensionality Reduction - PCA - Example

- **Goal:** Minimize the sum of reconstruction errors:

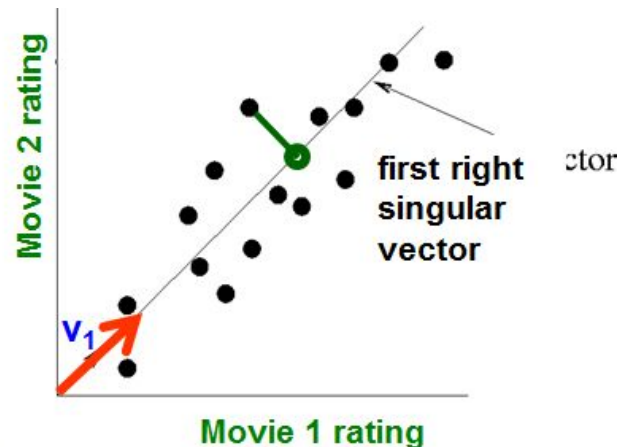
$$\sum_{i=1}^N \sum_{j=1}^D \|x_{ij} - z_{ij}\|^2$$

- where x_{ij} are the “old” and z_{ij} are the “new” coordinates

- **SVD gives ‘best’ axis to project on:** $V =$

- ‘best’ = minimizing the reconstruction errors

- In other words, **minimum reconstruction error**



0.56	0.59	0.56	0.09	0.09
0.12	-0.02	0.12	-0.69	-0.69
0.40	-0.80	0.40	0.09	0.09

Datasets, <http://www.mmms.org>

13

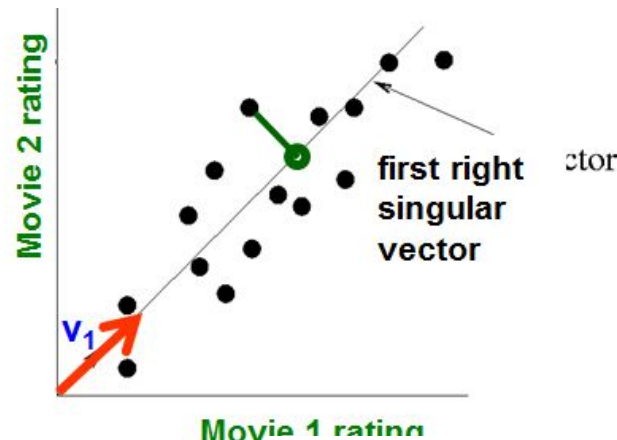
Dimensionality Reduction - PCA - Example

- **Goal:** Minimize the sum of reconstruction errors:

$$\sum_{i=1}^N \sum_{j=1}^D \|x_{ij} - z_{ij}\|^2$$

- where x_{ij} are the “old” and z_{ij} are the “new” coordinates

- **SVD gives ‘best’ axis to project on:** $(UD)^T =$
 - ‘best’ = minimizing the reconstruction errors
- In other words, **minimum reconstruction error**



1.61	0.19	-0.01
5.08	0.66	-0.03
6.82	0.85	-0.05
8.43	1.04	-0.06
1.86	-5.60	0.84
0.86	-6.93	-0.87
0.86	-2.75	0.41

Dimensionality Reduction - PCA

Linear approximates of data in r dimensions.

Found via *Singular Value Decomposition*:

$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

X: original matrix,

U: “left singular vectors”,

D: “singular values” (diagonal),

V: “right singular vectors”

Projection (dimensionality reduced space) in 3 dimensions:

$$(U_{[n \times 3]} D_{[3 \times 3]} V_{[p \times 3]}^T)$$

To reduce features in new dataset:

$$X_{\text{new}} V = X_{\text{new_small}}$$

Dimensionality Reduction - PCA

Linear approximates of data in r dimensions.

Found via *Singular Value Decomposition*:

$$X_{[n \times p]} = U_{[n \times r]} D_{[r \times r]} V_{[p \times r]}^T$$

U, D, and V are unique

D: always positive

Dimensionality Reduction v. Clustering

Clustering: Group n **observations** into k *clusters*

Soft Clustering: Assign **observations** to k *clusters* with some weight or probability.

Dimensionality Reduction: Assign m **features** to p *components* with some weight or probability.

Dimensionality Reduction v. Clustering

Clustering: Group n **observations** into k *clusters*

Soft Clustering: Assign **observations** to k *clusters* with some weight or probability.

Dimensionality Reduction: Assign m **features** to p *components* with some weight or probability.

Can often use one to do the other with one extra step. Examples

- From Dimensionality Reduction to Clusters:
 - Use U instead of a V from SVD = mapping observations to soft clusters
 - Project based on V , Apply a threshold on U = mapping observations to clusters
 - Threshold v (or use sparse PCA) = soft clustering of **features**
- From Clusters to Dimensionality Reduction:
 - Use soft cluster ids as features