Advanced Data Mining
(basic concepts as a starting points)

Lecture 1

Yao-Chung Fan

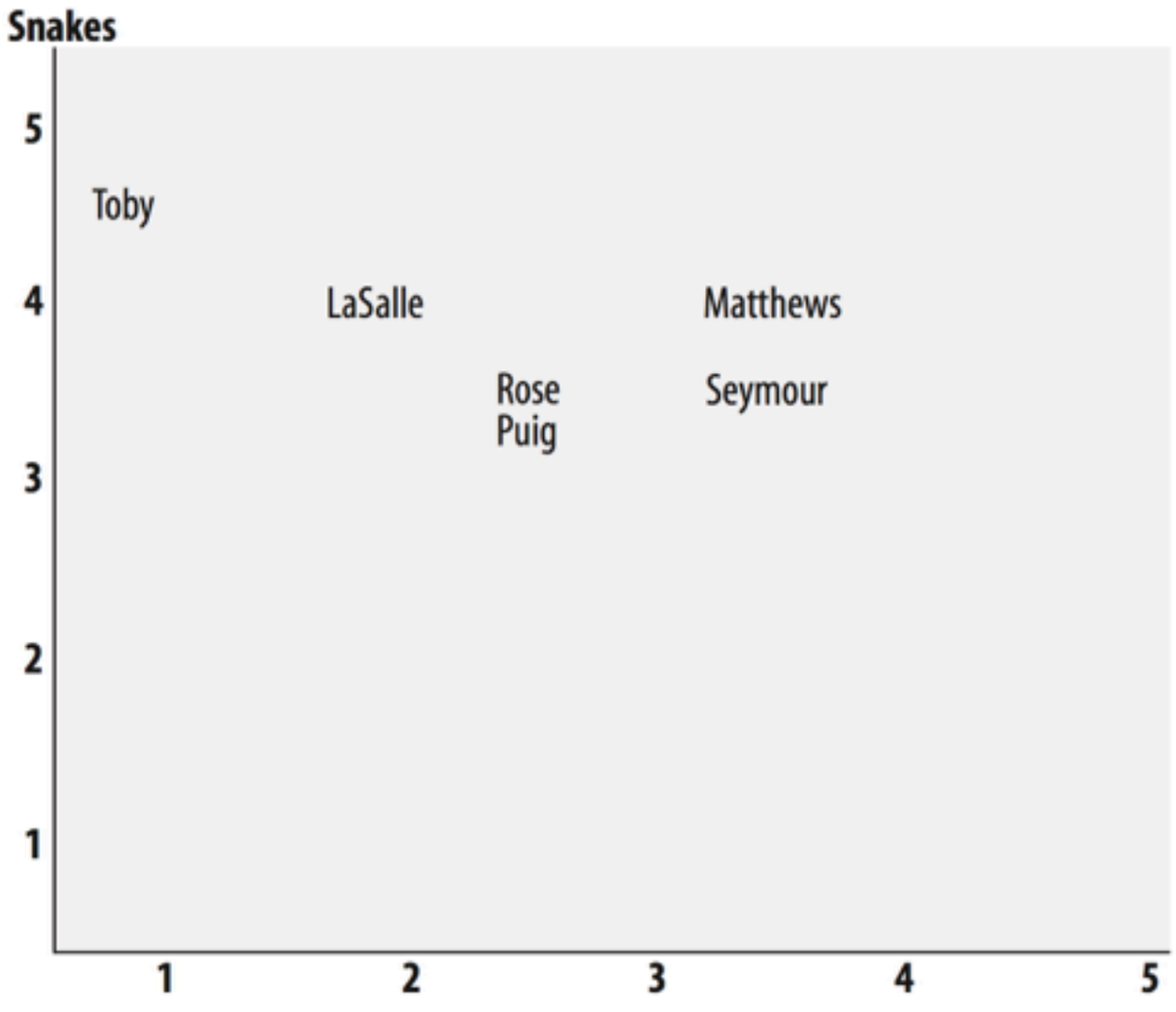
Data = Money

A very fundamental step for all data mining techniques:

- Finding similar items

Concept: User Space

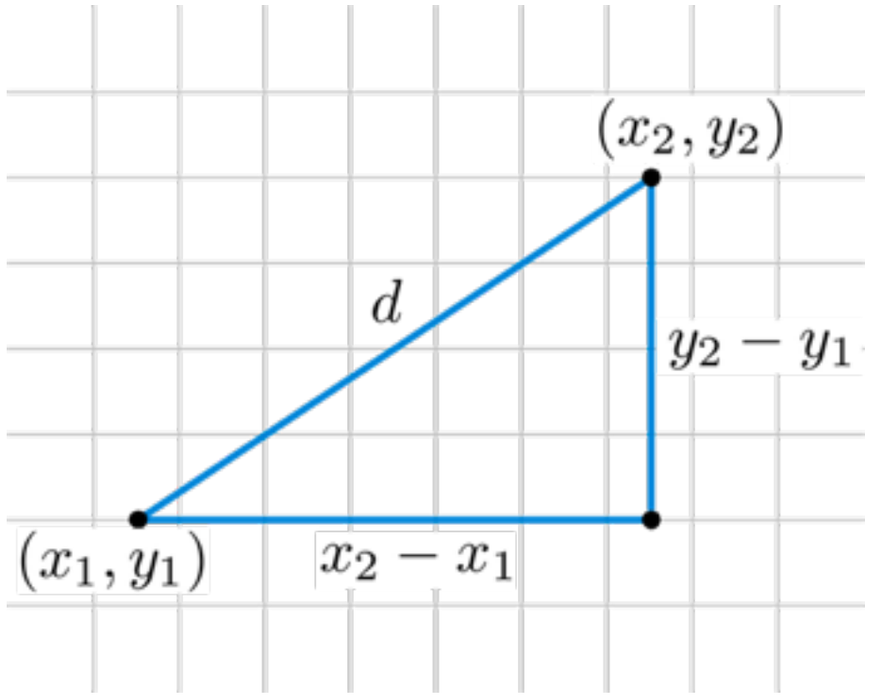
	Snakes on a Plane	Superman Returns
Lisa Rose	3.5	3.5
Gene Seymour	3.5	5
Michael Phillips	3.0	3.5
Claudia Puig	3.5	4.0
Mick LaSalle	4.0	3.0
Jack Matthews	4.0	5.0
Toby	4.5	4.0



Concept: Similarity/Distance

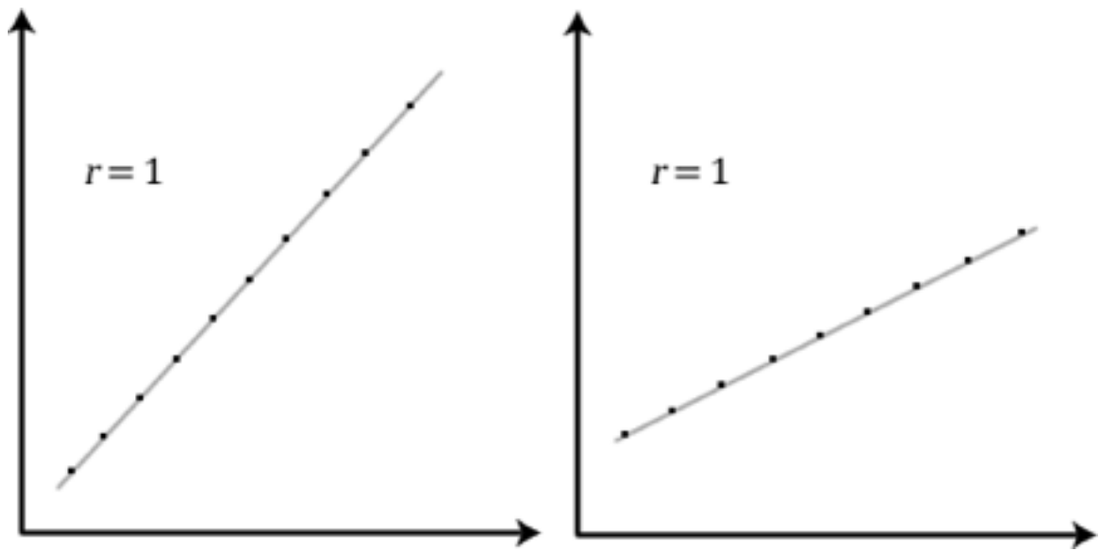
Euclidean Distance

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

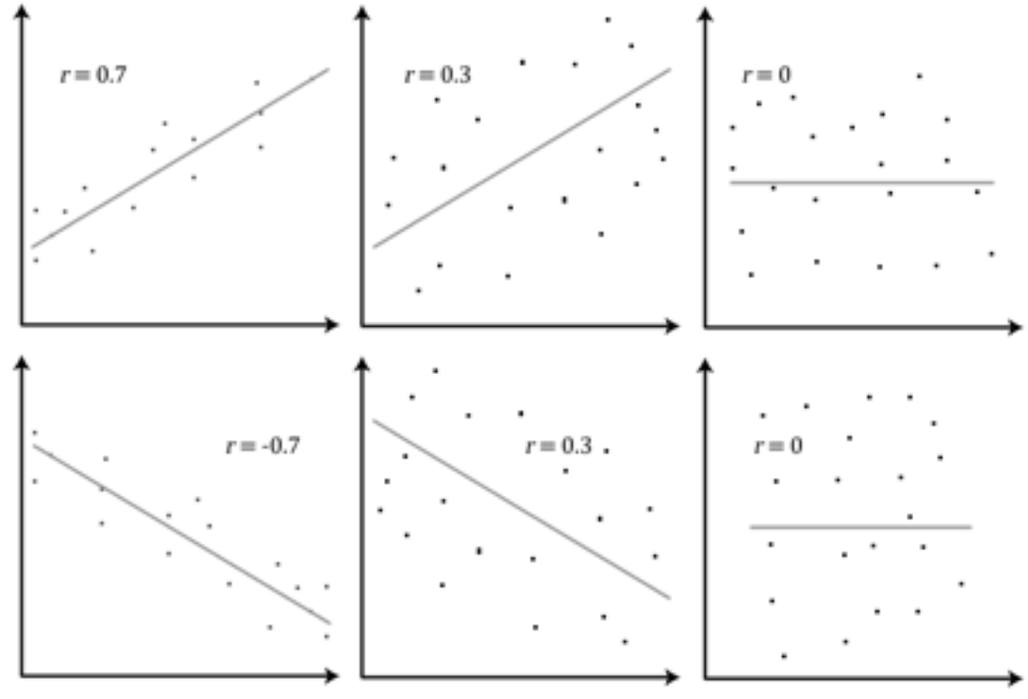


Pearson Distance

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$




```
def pearson(x,y):
  n=len(x)
  vals=range(n)

# Simple sums
sumx=sum([float(x[i]) for i in vals])
sumy=sum([float(y[i]) for i in vals])

# Sum up the squares
sumxSq=sum([x[i]**2.0 for i in vals])
sumySq=sum([y[i]**2.0 for i in vals])

# Sum up the products
pSum=sum([x[i]*y[i] for i in vals])

# Calculate Pearson score
num=pSum-(sumx*sumy/n)
den=((sumxSq-pow(sumx,2)/n)*(sumySq-pow(sumy,2)/n))**.5
if den==0: return 0

r=num/den

return r
```



Concept: A Recommendation

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Lisa Rose	2.5	3.5	3.0	3.5	2.5	3.0
Gene Seymour	3.0	3.5	1.5	5	3.5	3.0
Michael Phillips	2.5	3.0		3.5		4.0
Claudia Puig		3.5	3.0	4.0	2.5	4.5
Mick LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Jack Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

Find similar users and use their ratings to predict rating for a target ?

Critic	Similarity	Night	S.xNight	Lady	S.xLady	Luck	S.xLuck
Rose	0.99	3.0	2.97	2.5	2.48	3.0	2.97
Seymour	0.38	3.0	1.14	3.0	1.14	1.5	0.57
Puig	0.89	4.5	4.02			3.0	2.68
LaSalle	0.92	3.0	2.77	3.0	2.77	2.0	1.85
Matthews	0.66	3.0	1.99	3.0	1.99		
Total			12.89		8.38		8.07
Sim. Sum			3.84		2.95		3.18
Total/Sim. Sum			3.35		2.83		2.53

User Similarity ?

Item Similarity ?

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Lisa Rose	2.5	3.5	3.0	3.5	2.5	3.0
Gene Seymour	3.0	3.5	1.5	5	3.5	3.0
Michael Phillips	2.5	3.0		3.5		4.0
Claudia Puig		3.5	3.0	4.0	2.5	4.5
Mick LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Jack Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

Find the items that a use rated and use the ratings and the item similarity to predict the ratings that not yet rated by the user

Movie	Rating	Night	R.xNight	Lady	R.xLady	Luck	R.xLuck
Snakes	4.5	0.182	0.818	0.222	0.999	0.105	0.474
Superman	4.0	0.103	0.412	0.091	0.363	0.065	0.258
Dupree	1.0	0.148	0.148	0.4	0.4	0.182	0.182
Total		0.433	1.378	0.713	1.764	0.352	0.914
Normalized			3.183		2.598		2.473

Recap

Collaborative Filtering recommendation

* **User-Based Recommendation ?**
* **Item-Based Recommendation ?**
* **Similarity**
* **Pearson Distance**
* **Euclidean Distance**

Assignment 1:
為我推薦個電影吧？
我預先勾好幾部我喜歡的電影，以及其評價。
看看同學有沒辦法精準預測出我的喜好。

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please take a short survey about the MovieLens datasets

MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: http://grouplens.org/datasets/movielens/1m/

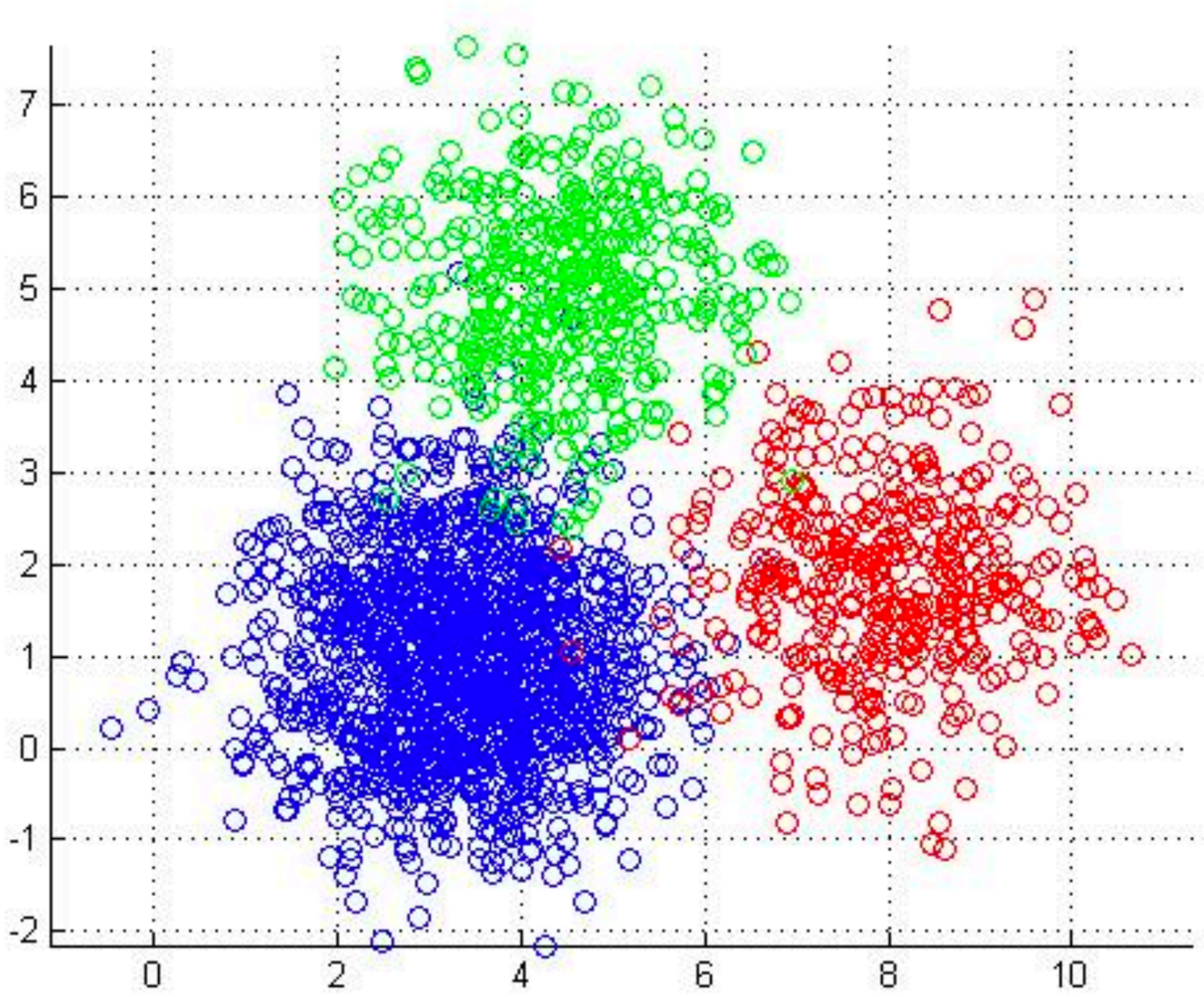
https://grouplens.org/datasets/movielens/

My Rating for the Following Movies

1::Toy Story (1995)::Animation|Children's|Comedy, 5
2::Jumanji (1995)::Adventure|Children's|Fantasy, 4
9::Sudden Death (1995)::Action, 2
10::GoldenEye (1995)::Action|Adventure|Thriller, 2
13::Balto (1995)::Animation|Children's, 1
14::Nixon (1995)::Drama, 1
17::Sense and Sensibility (1995)::Drama|Romance, 1
22::Copycat (1995)::Crime|Drama|Thriller
23::Assassins (1995)::Thriller, 3
47::Seven (Se7en) (1995)::Crime|Thriller, 2
356::Forrest Gump (1994)::Comedy|Romance|War, 5
3147::Green Mile, The (1999)::Drama|Thriller, 5
593::Silence of the Lambs, The (1991)::Drama|Thriller, 2
2028::Saving Private Ryan (1998)::Action|Drama|War, 5
838::Emma (1996)::Comedy|Drama|Romance, 1
1721::Titanic (1997)::Drama|Romance, 5
2628::Star Wars: Episode I - The Phantom Menace (1999)::Action|Adventure|Fantasy|Sci-Fi, 4
1608::Air Force One (1997)::Action|Thriller, 4
165::Die Hard: With a Vengeance (1995)::Action|Thriller, 4
589::Terminator 2: Judgment Day (1991)::Action|Sci-Fi|Thriller, 2

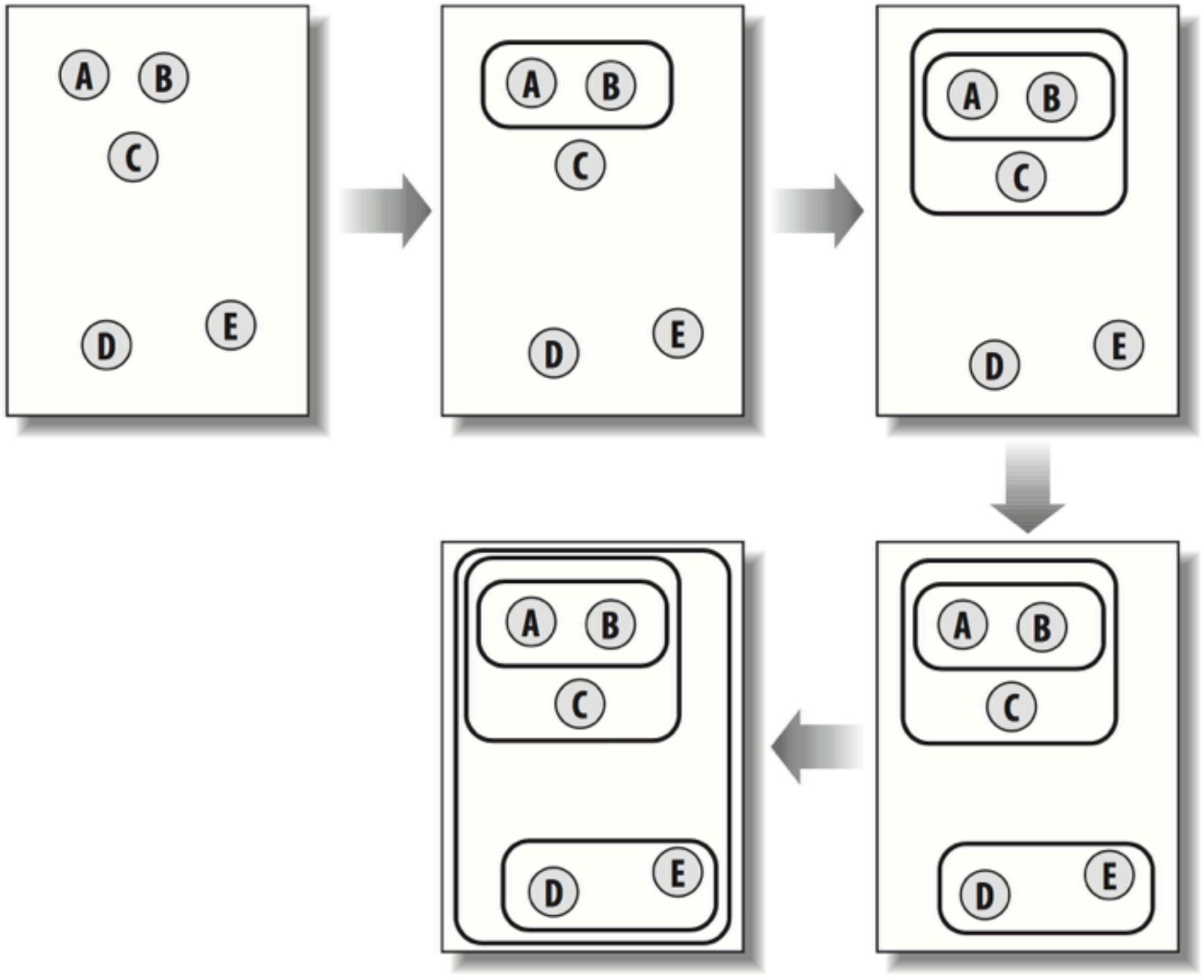
318::Shawshank Redemption, The (1994)::Drama, ?
527::Schindler's List (1993)::Drama|War, ?
2959::Fight Club (1999)::Drama, ?
393::Street Fighter (1994)::Action, ?
3285::Beach, The (2000)::Adventure|Drama, ?
2571::Matrix, The (1999)::Action|Sci-Fi|Thriller, ?
1270::Back to the Future (1985)::Comedy|Sci-Fi, ?
3578::Gladiator (2000)::Action|Drama, ?
1200::Aliens (1986)::Action|Sci-Fi|Thriller|War, ?
2858::American Beauty (1999)::Comedy|Drama, ?

Discovering Groups

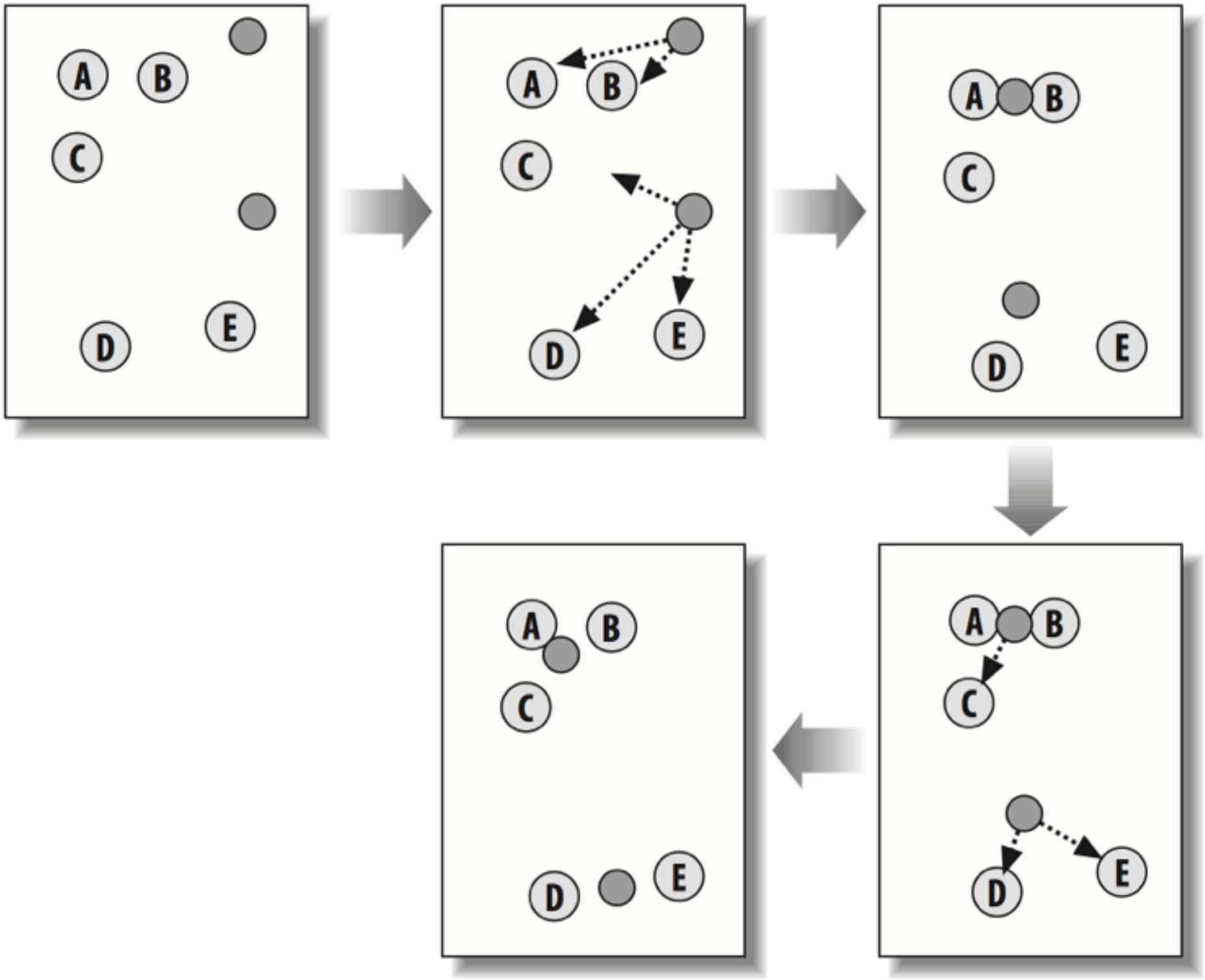


Hierarchical Clustering Algorithm
K-Means Clustering Algorithm

Hierarchical Clustering Algorithm

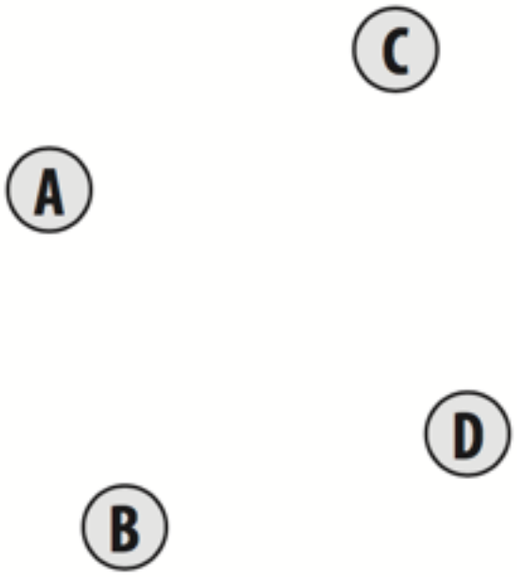


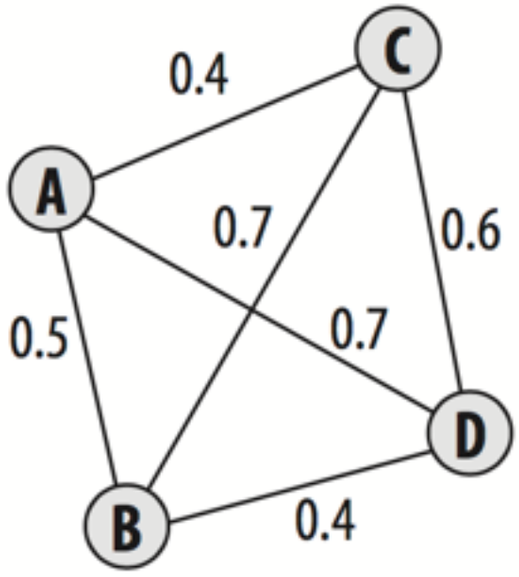
K-Means Clustering Algorithm

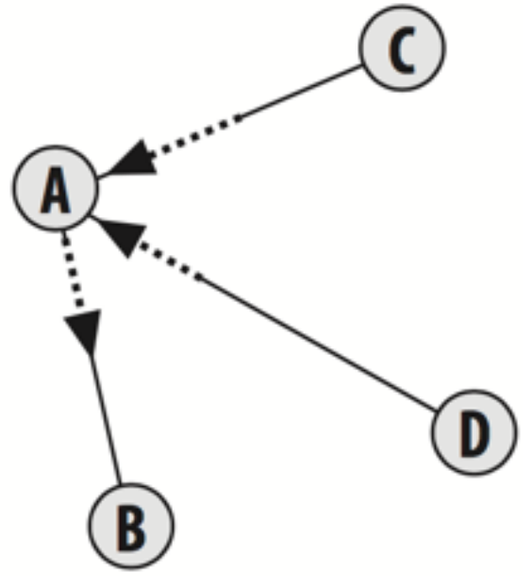


Multidimensional Scaling (How Visualize Multidimensional Space)

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Lisa Rose	2.5	3.5	3.0	3.5	2.5	3.0
Gene Seymour	3.0	3.5	1.5	5	3.5	3.0
Michael Phillips	2.5	3.0		3.5		4.0
Claudia Puig		3.5	3.0	4.0	2.5	4.5
Mick LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Jack Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	







	A	B	C	D
A	0.0	0.2	0.8	0.7
B	0.2	0.0	0.9	0.8
C	0.8	0.9	0.0	0.1
D	0.7	0.8	0.1	0.0

Assignment 2:
利用Hierarchical Clustering Algorithm 畫一個dendrogram

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please take a short survey about the MovieLens datasets

MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: http://grouplens.org/datasets/movielens/1m/

https://grouplens.org/datasets/movielens/