

Advanced Data Mining (basic concept as a starting points)

Lecture 1
Yao-Chung Fan

Data = Money

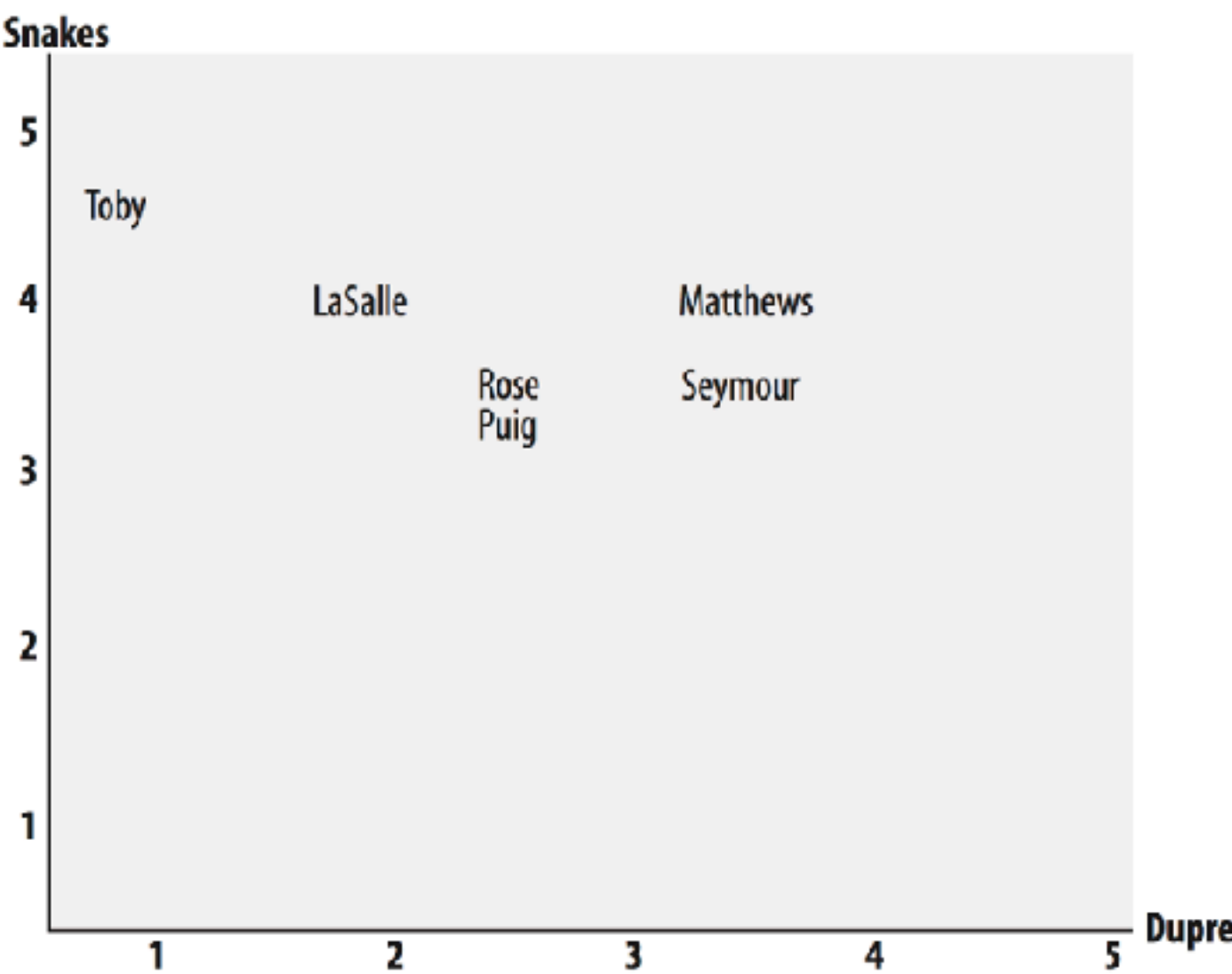
A very fundamental step for all data mining techniques:

- Finding similar items



Concept: User Space

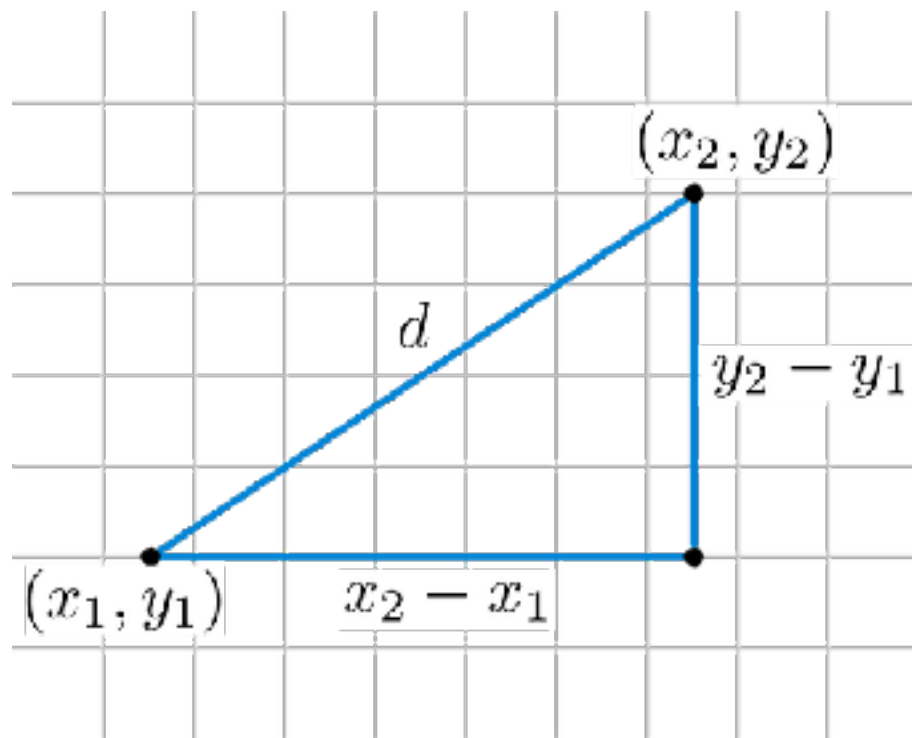
	Snakes on a Plane	Superman Returns
Lisa Rose	3.5	3.5
Gene Seymour	3.5	5
Michael Phillips	3.0	3.5
Claudia Puig	3.5	4.0
Mick LaSalle	4.0	3.0
Jack Matthews	4.0	5.0
Toby	4.5	4.0



Concept: Similarity/Distance

Euclidean Distance

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

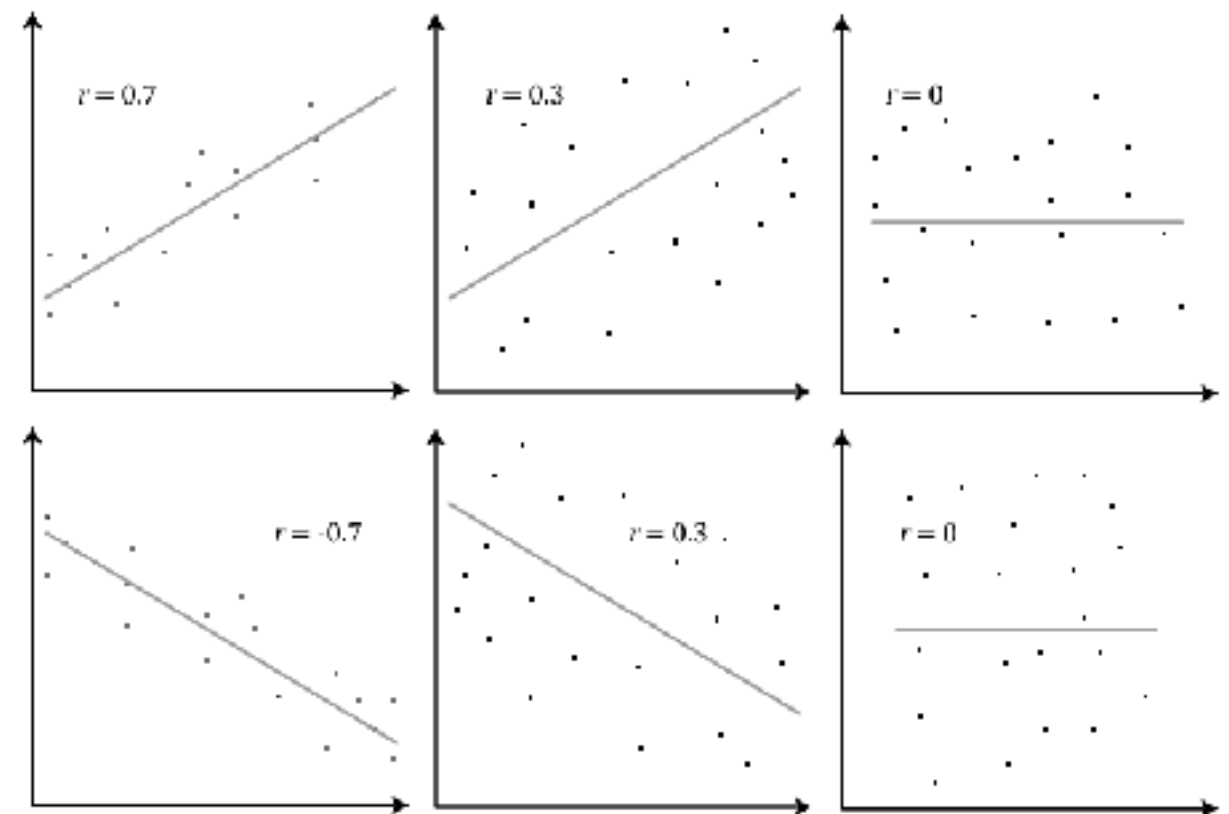
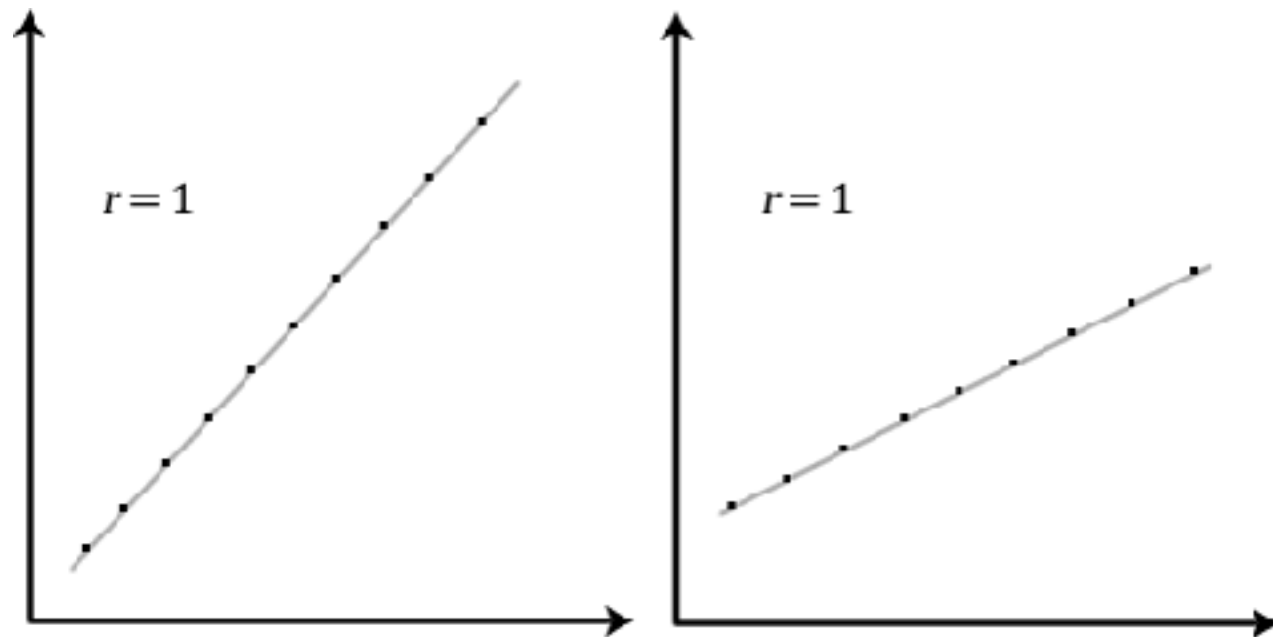
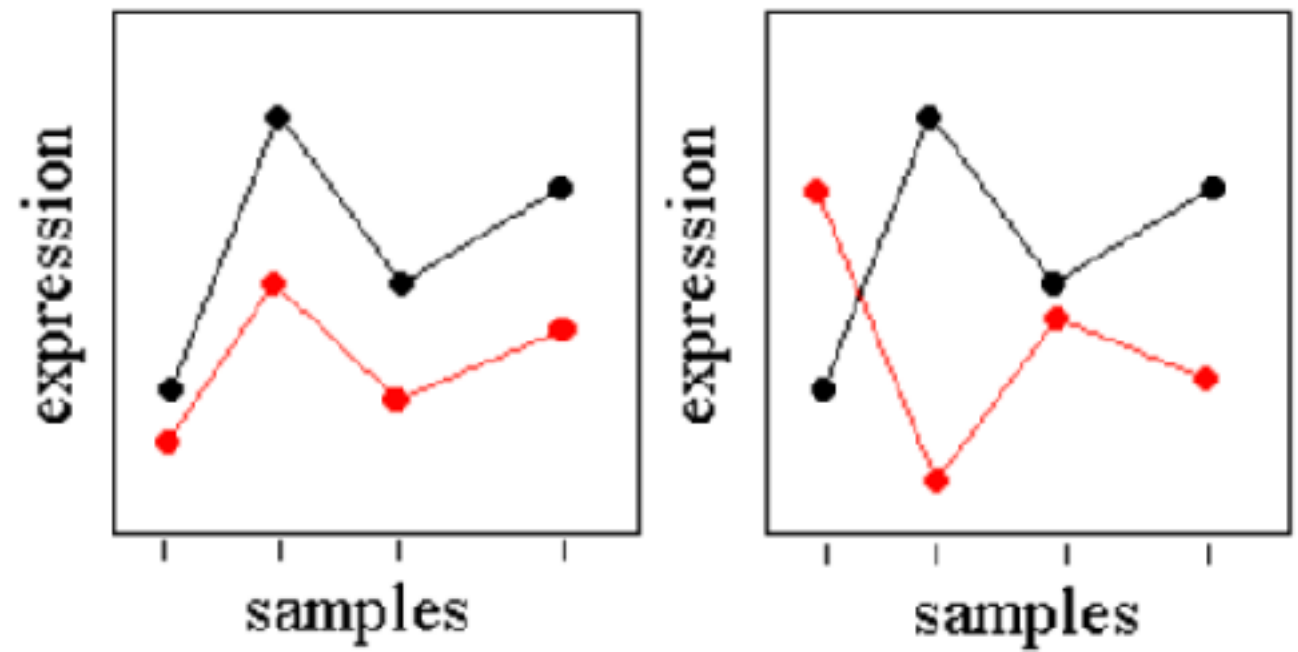


Distance/Similarity

[0-1] ???

Pearson Distance

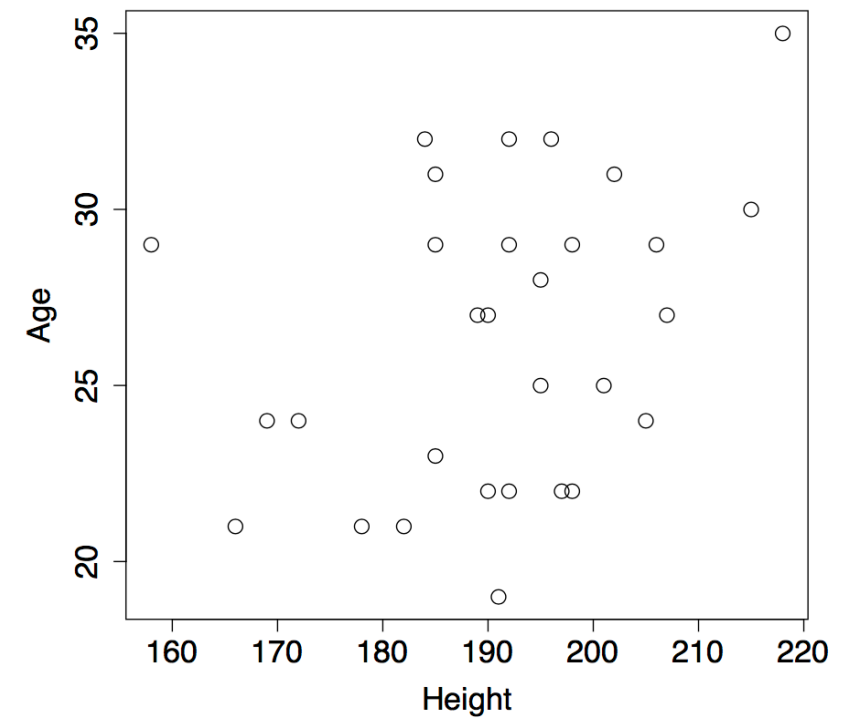
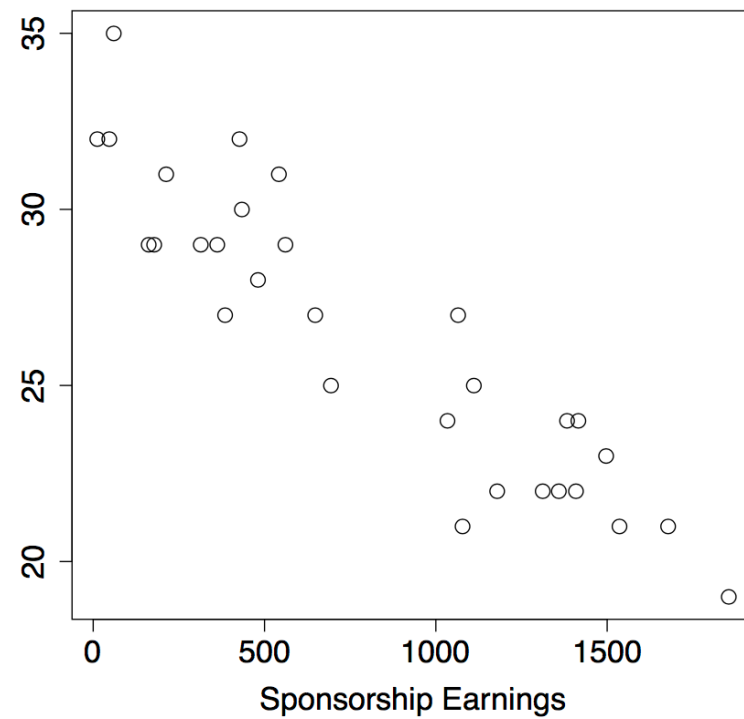
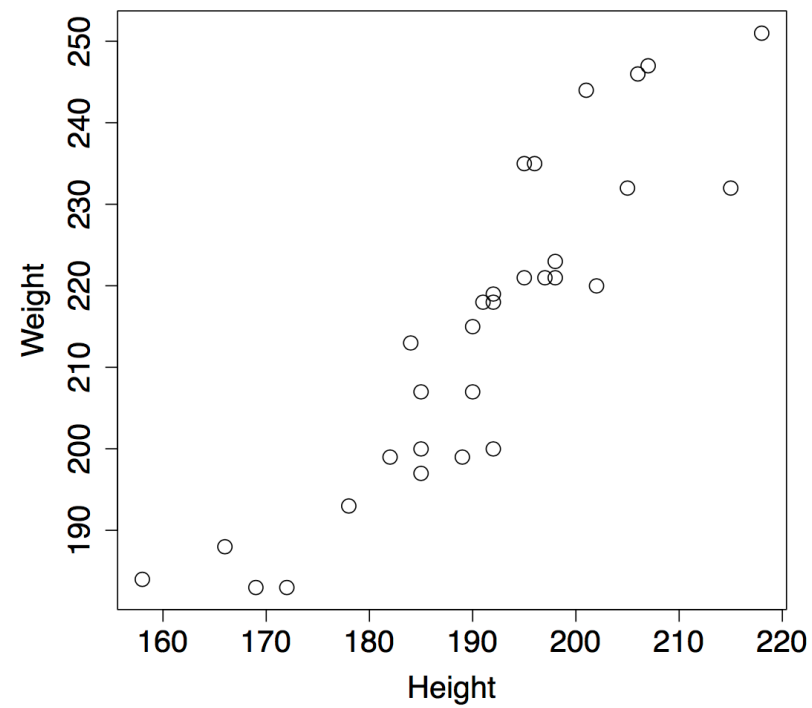
$$\rho_{xy} = \frac{\sum((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}}$$



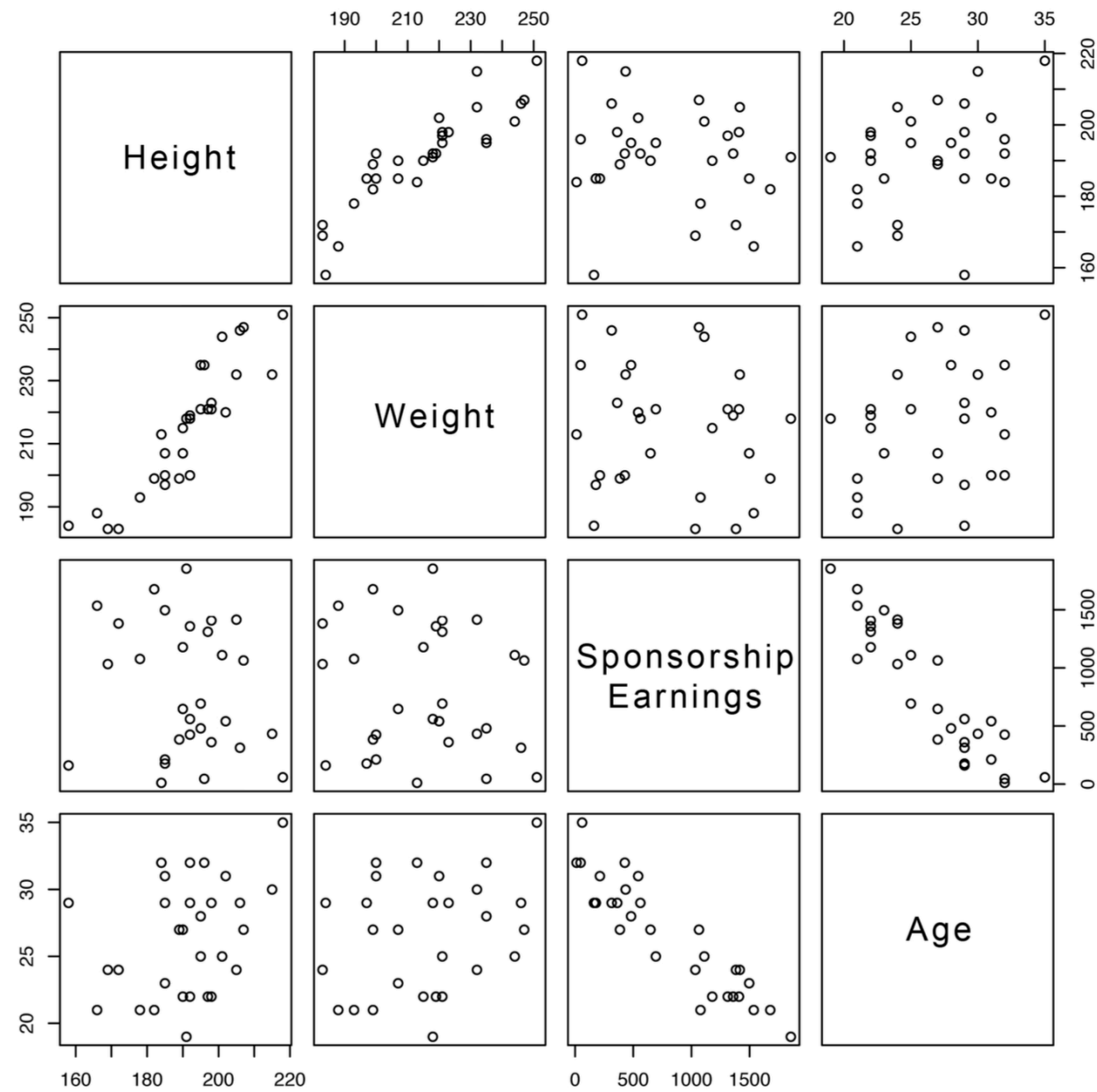
NBA Player Data Set (An Example for Pearson Correlation)

ID	POSITION	HEIGHT	WEIGHT	CAREER STAGE	AGE	SPONSORSHIP EARNINGS	SHOE SPONSOR
1	forward	192	218	veteran	29	561	yes
2	center	218	251	mid-career	35	60	no
3	forward	197	221	rookie	22	1,312	no
4	forward	192	219	rookie	22	1,359	no
5	forward	198	223	veteran	29	362	yes
6	guard	166	188	rookie	21	1,536	yes
7	forward	195	221	veteran	25	694	no
8	guard	182	199	rookie	21	1,678	yes
9	guard	189	199	mid-career	27	385	yes
10	forward	205	232	rookie	24	1,416	no
11	center	206	246	mid-career	29	314	no
12	guard	185	207	rookie	23	1,497	yes
13	guard	172	183	rookie	24	1,383	yes
14	guard	169	183	rookie	24	1,034	yes
15	guard	185	197	mid-career	29	178	yes
16	forward	215	232	mid-career	30	434	no
17	guard	158	184	veteran	29	162	yes
18	guard	190	207	mid-career	27	648	yes
19	center	195	235	mid-career	28	481	no
20	guard	192	200	mid-career	32	427	yes
21	forward	202	220	mid-career	31	542	no
22	forward	184	213	mid-career	32	12	no
23	forward	190	215	rookie	22	1,179	no
24	guard	178	193	rookie	21	1,078	no
25	guard	185	200	mid-career	31	213	yes
26	forward	191	218	rookie	19	1,855	no
27	center	196	235	veteran	32	47	no
28	forward	198	221	rookie	22	1,409	no
29	center	207	247	veteran	27	1,065	no
30	center	201	244	mid-career	25	1,111	yes

Feature Correlation



Scatter Plot Matrix



Covariance and Correlation

The way to measure the correlation between features

For two features, a and b , in a dataset of n instances, the **sample covariance** between a and b is

$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{i=1}^n ((a_i - \bar{a}) \times (b_i - \bar{b})) \quad (1)$$

where a_i and b_i are values of features a and b for the i^{th} instance in a dataset, and \bar{a} and \bar{b} are the sample means of features a and b .

	HEIGHT		WEIGHT		$(h - \bar{h}) \times$	AGE		$(h - \bar{h}) \times$
ID	(h)	$h - \bar{h}$	(w)	$w - \bar{w}$	$(w - \bar{w})$	(a)	$a - \bar{a}$	$(a - \bar{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
...								
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

Covariance and Correlation

The way to measure the correlation between features

ID	HEIGHT (h)	$h - \bar{h}$	WEIGHT (w)	$w - \bar{w}$	$(h - \bar{h}) \times (w - \bar{w})$	AGE (a)	$a - \bar{a}$	$(h - \bar{h}) \times (a - \bar{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
...								
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

$$\text{cov}(\text{HEIGHT}, \text{WEIGHT}) = \frac{7,009.9}{29} = 241.72$$

$$\text{cov}(\text{HEIGHT}, \text{AGE}) = \frac{570.8}{29} = 19.7$$

Covariance and Correlation

The way to measure the correlation between features

Correlation is a normalized form of covariance that ranges between -1 and $+1$.

The correlation between two features, a and b , can be calculated as

$$\text{corr}(a, b) = \frac{\text{cov}(a, b)}{\text{sd}(a) \times \text{sd}(b)} \quad (2)$$

where $\text{cov}(a, b)$ is the covariance between features a and b and $\text{sd}(a)$ and $\text{sd}(b)$ are the standard deviations of a and b respectively.

Correlation values fall into the range $[-1, 1]$, where values close to -1 indicate a very strong negative correlation (or covariance), values close to 1 indicate a very strong positive correlation, and values around 0 indicate no correlation.

Covariance and Correlation

The way to measure the correlation between features

ID	HEIGHT (h)	$h - \bar{h}$	WEIGHT (w)	$w - \bar{w}$	$(h - \bar{h}) \times (w - \bar{w})$	AGE (a)	$a - \bar{a}$	$(h - \bar{h}) \times (a - \bar{a})$
1	192	0.9	218	3.0	2.7	29	2.6	2.3
2	218	26.9	251	36.0	967.5	35	8.6	231.3
3	197	5.9	221	6.0	35.2	22	-4.4	-26.0
4	192	0.9	219	4.0	3.6	22	-4.4	-4.0
5	198	6.9	223	8.0	55.0	29	2.6	17.9
...								
26	191	-0.1	218	3.0	-0.3	19	-7.4	0.7
27	196	4.9	235	20.0	97.8	32	5.6	27.4
28	198	6.9	221	6.0	41.2	22	-4.4	-30.4
29	207	15.9	247	32.0	508.3	27	0.6	9.5
30	201	9.9	244	29.0	286.8	25	-1.4	-13.9
Mean	191.1		215.0			26.4		
Std Dev	13.6		19.8			4.2		
Sum					7,009.9			570.8

$$\text{corr}(\text{Height}, \text{Weight}) = \frac{241.72}{13.6 \times 19.8} = 0.898$$

$$\text{corr}(\text{Height}, \text{Age}) = \frac{19.7}{13.6 \times 4.2} = 0.345$$

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	

Idea: 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 user 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

合理嗎？

	Dupree	1.0	0.148	0.148
---	--------	-----	-------	-------


Recap

Collaborative Filtering recommendation

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

Other Movies You Might Enjoy

[Amélie](#)




Add

★★★★★

Not Interested

[Y Tu Mama Tambien](#)




Add

★★★★★

Not Interested

[Guys and Girls](#)




Add

★★★★★

Not Interested

[Mostly Martha](#)




Add

★★★★★

Not Interested

[Only Human](#)




Add

★★★★★

Not Interested


[Russian Dolls](#)



Add

★★★★★

Not Interested



Eiken has been added to your Queue at position 2.

This movie is available now.

Move To Top Of My Queue

[Continue Browsing](#) [Visit your Queue](#)

NETFLIX

Netflix Prize

COMPLETED

Home

Rules

Leaderboard

Update

NETFLIX

Browse

Recommendations

Friends

Queue

Buy DVDs

Home

Genres

New Releases

Previews

Netflix Top 100

Crit

Movies For You

Randy, the following movies were chosen based on your interest in:
[Bowling for Columbine](#)
[Carnivale, Season 1](#)
[Fahrenheit 9/11](#)

The Big One

★★★★☆

over subversive

by from

on /

angel

OT

RIGHT

Lena Black, B

and Goro

Add

★★★★☆

Not Inter

By Decade

By Studio

By Genre

By Year

By Rating

By Country

By Language

By Director

By Cast

By Plot

By Theme

By Style

By Mood

All Discs
Guaranteed

You really
liked it..

Now over \$5.99

Shop

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

as low

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

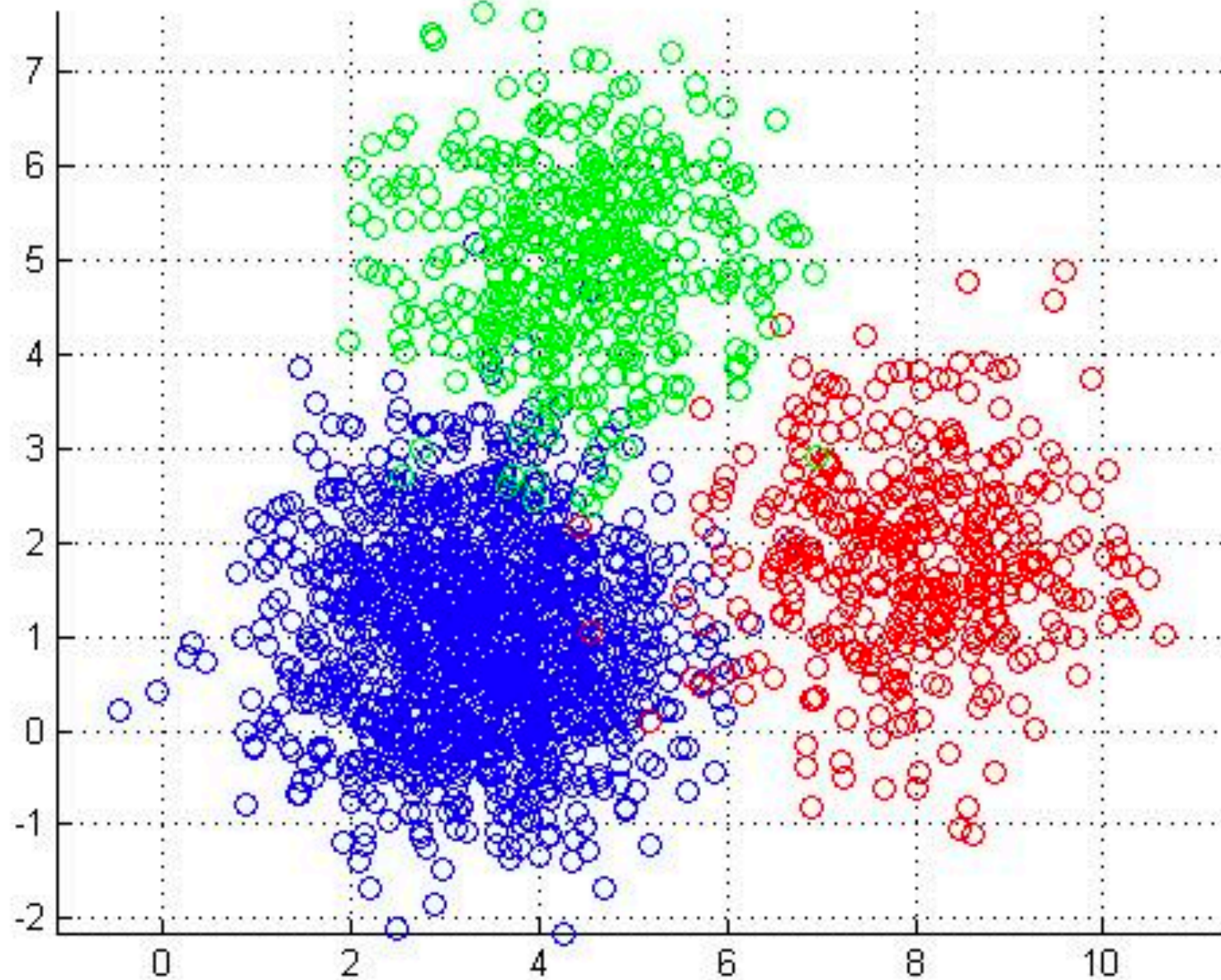
1::Toy Story (1995)::Animation|Children's|Comedy, 5
2::Jumanji (1995)::Adventure|Children's|Fantasy, 4
9::Sudden Death (1995)::Action, 2
10::GoldenEye (1995)::Action|Adventure|Thriller, 2
13::Balto (1995)::Animation|Children's, 1
14::Nixon (1995)::Drama, 1
17::Sense and Sensibility (1995)::Drama|Romance, 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

How to Measure the Similarity between blogs/documents?

A: Word Vectors

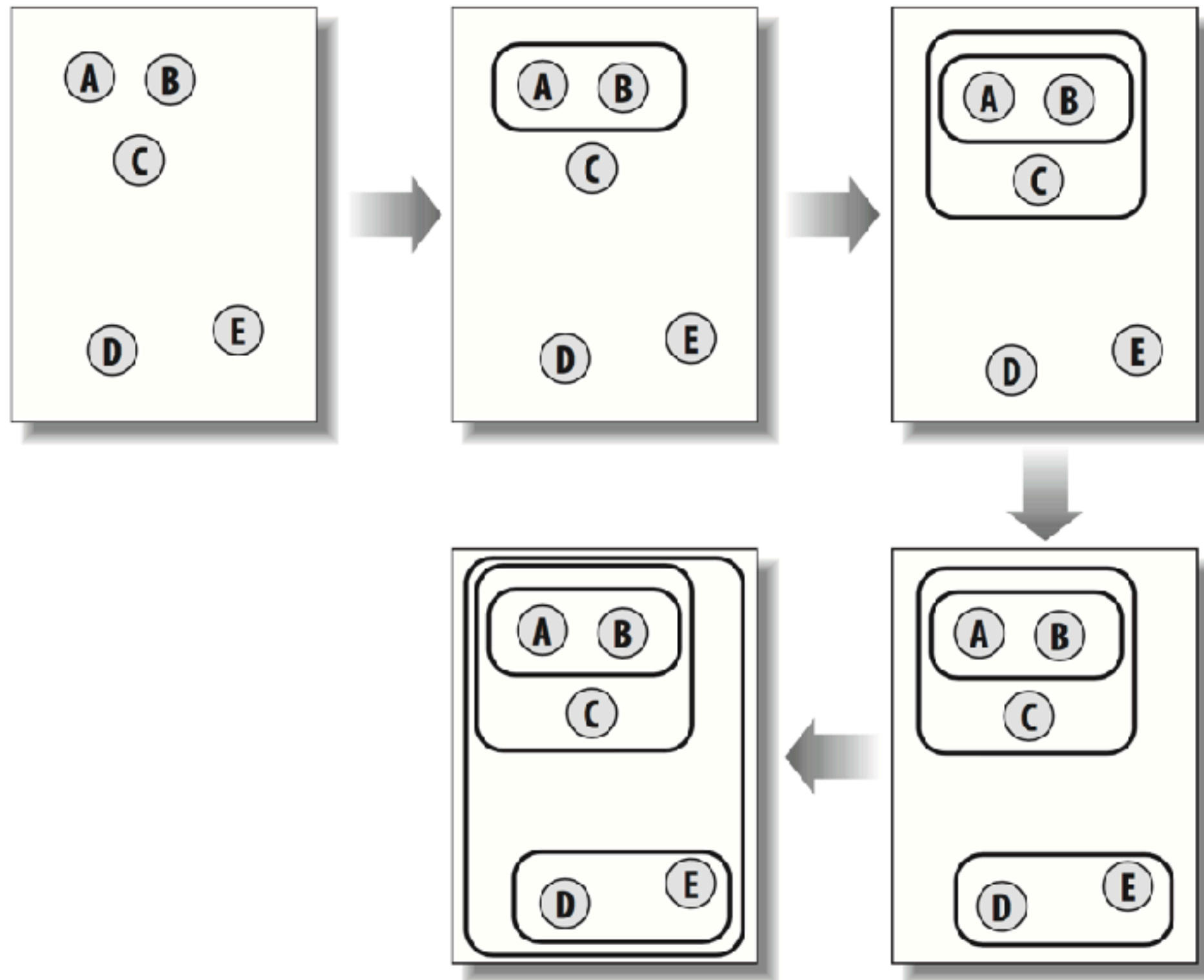
Blog	china	kids	music	yahoo	want	wrong	service	tech	saying	lots	had	address	working	following	
The Superficial – Because You're Ugly					0	1	0	0	3	3	0	0	3	0	6
Wonkette		0	2	1	0	6	2	1	0	4	5	25	0	0	0
Publishing 2.0		0	0	7	4	0	1	3	6	0	0	1	0	0	0
Eschaton		0	0	0	0	5	3	2	0	1	0	1	0	1	0
Blog Maverick		2	14	17	2	45	11	8	0	4	7	24	2	4	3
Mashable!		0	0	0	0	1	0	1	0	0	0	0	0	0	0
we make money not art			0	1	1	0	6	0	0	1	0	1	21	3	20
GigaOM 6		0	0	2	1	0	3	1	0	0	1	0	0	1	1
Joho the Blog		0	0	1	4	0	0	1	0	0	0	4	1	0	0

words

documents



Hierarchical Clustering Algorithm



How to Measure the Similarity between words?

A: Document Vectors

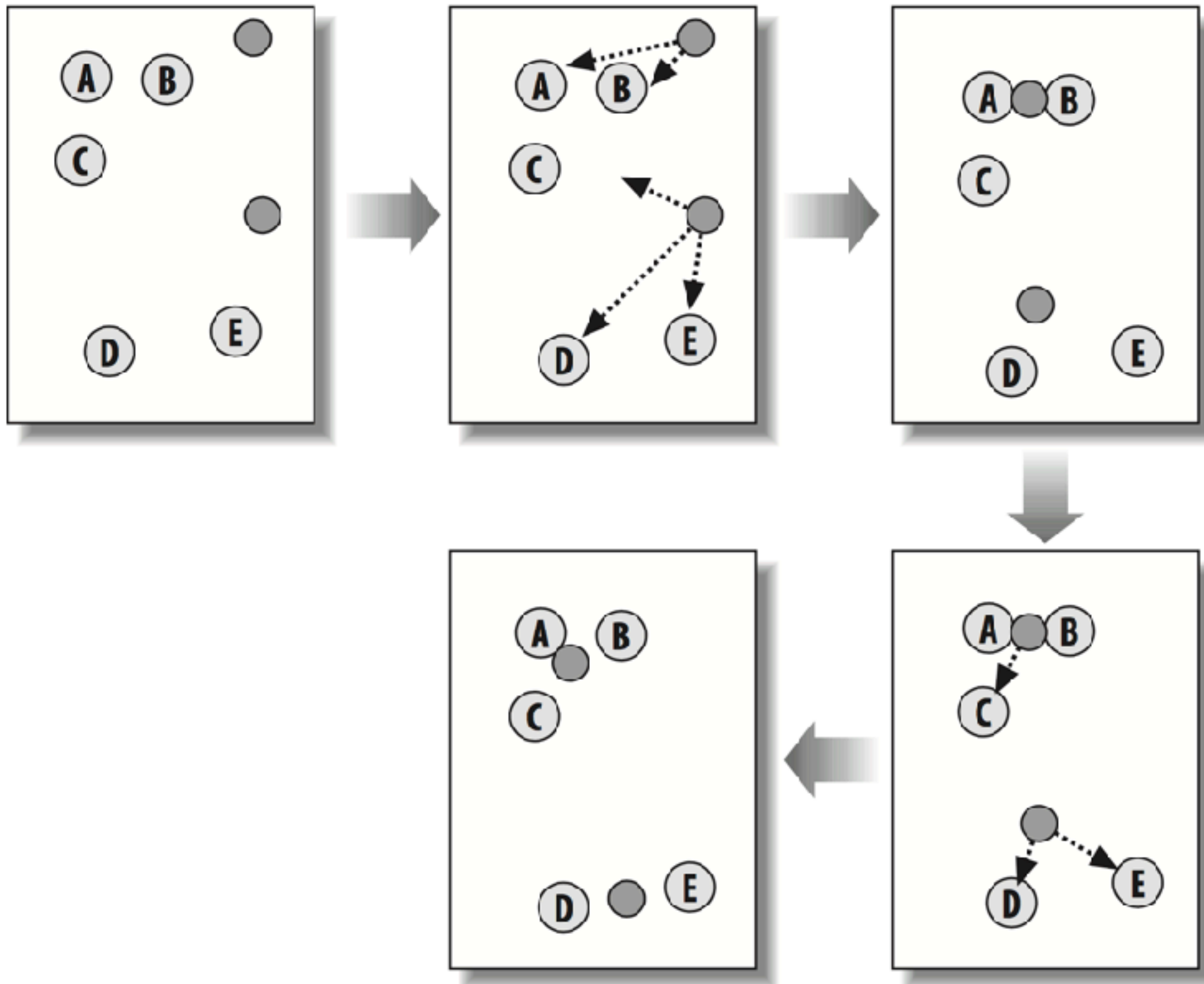
Blog	china	kids	music	yahoo	want	wrong	service	tech	saying	lots	had	address	working	following
The Superficial	–	Because You're Ugly	0	1	0	0	3	3	0	0	3	0	6	
Wonkette	0	2	1	0	6	2	1	0	4	5	25	0	0	0
Publishing 2.0	0	0	7	4	0	1	3	6	0	0	1	0	0	0
Eschaton	0	0	0	0	5	3	2	0	1	0	1	0	1	0
Blog Maverick	2	14	17	2	45	11	8	0	4	7	24	2	4	3
Mashable!	0	0	0	0	1	0	1	0	0	0	0	0	0	0
we make money not art	0	1	1	0	6	0	0	1	0	1	21	3	20	
GigaOM 6	0	0	2	1	0	3	1	0	0	1	0	0	1	1
Joho the Blog	0	0	1	4	0	0	1	0	0	0	4	1	0	0

words

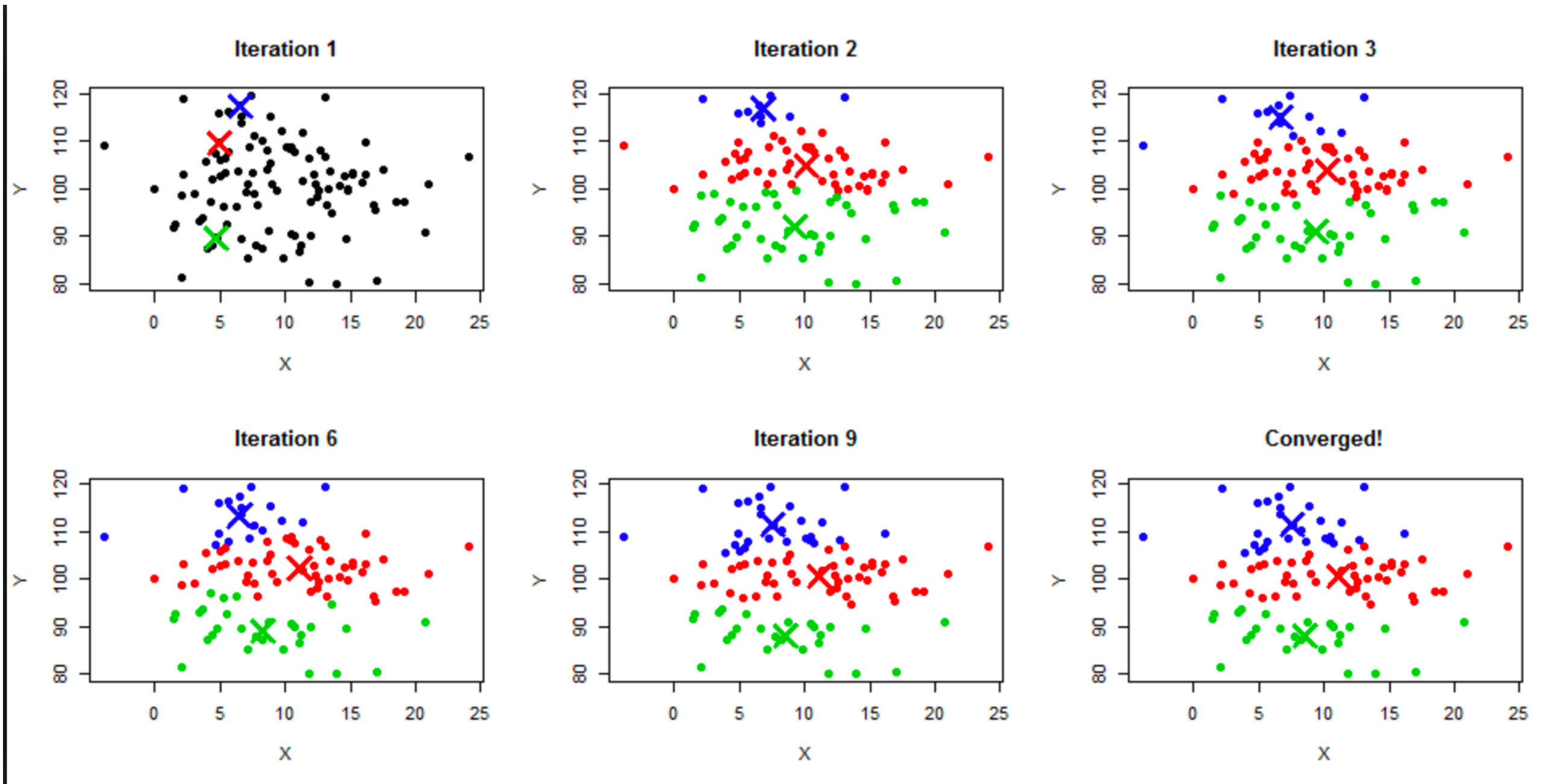
documents



K-Means Clustering Algorithm



K-Mean Algorithm Running Example



Multidimensional Scaling

(How to Visualize Multidimensional Space)

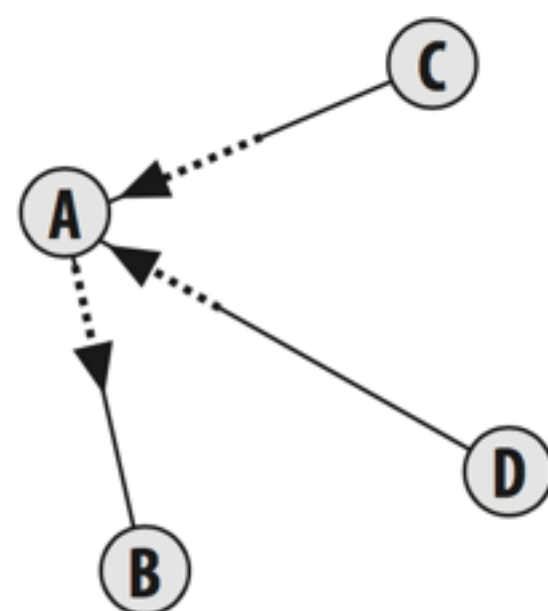
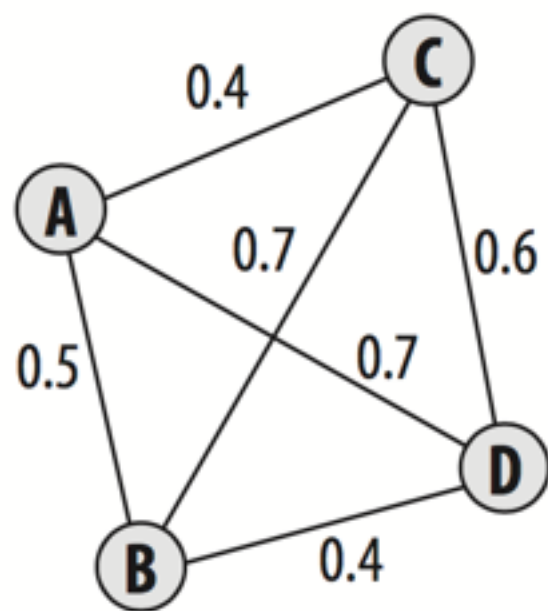
	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Lisa Rose	2.5	3.5	3.0	3.5	2.5	3.0
Gene Seymour	3.0	3.5	1.5	5	3.5	3.0
Michael Phillips	2.5	3.0		3.5		4.0
Claudia Puig		3.5	3.0	4.0	2.5	4.5
Mick LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Jack Matthews	3.0	4.0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

A

C

B

D



A	0.5	0.0	0.3	0.1
B	0.4	0.15	0.2	0.1
C	0.2	0.4	0.7	0.8
D	1.0	0.3	0.6	0.0

	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


```

for k in range(n):
    for j in range(n):
        if j == k:
            continue
        # The error is percent difference between the distances
        errorterm = (fakedist[j][k] - realdist[j][k]) / realdist[j][k]

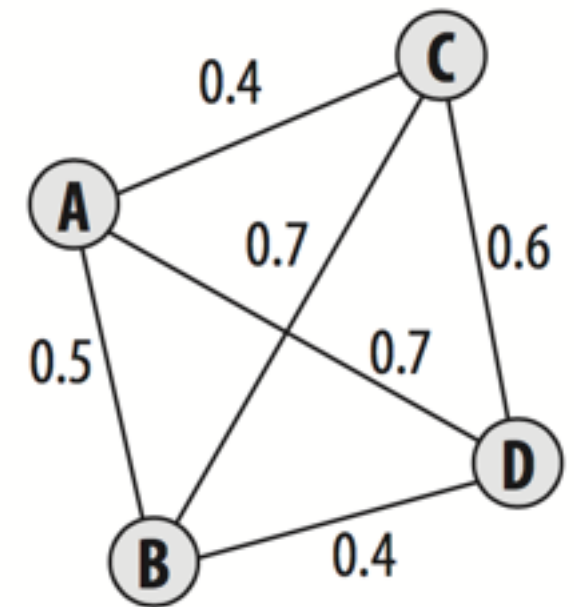
        # Each point needs to be moved away from or towards the other
        # point in proportion to how much error it has
        grad[k][0] += (loc[k][0] - loc[j][0]) / fakedist[j][k] \
            * errorterm
        grad[k][1] += (loc[k][1] - loc[j][1]) / fakedist[j][k] \
            * errorterm

        # Keep track of the total error
        totalerror += abs(errorterm)
    print totalerror

    # If the answer got worse by moving the points, we are done
    if lasterror and lasterror < totalerror:
        break
    lasterror = totalerror

    # Move each of the points by the learning rate times the gradient
    for k in range(n):
        loc[k][0] -= rate * grad[k][0]
        loc[k][1] -= rate * grad[k][1]

```



	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 for Movie Similarity

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/>

List Comprehension

```
S = {x2 : x in {0 ... 9}}  
V = (1, 2, 4, 8, ..., 212)  
M = {x | x in S and x even}
```

Math Set Notations

```
>>> S = [x**2 for x in range(10)]  
>>> V = [2**i for i in range(13)]  
>>> M = [x for x in S if x % 2 == 0]  
>>>  
>>> print S; print V; print M  
[0, 1, 4, 9, 16, 25, 36, 49, 64, 81]  
[1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096]  
[0, 4, 16, 36, 64]
```

Python List Comprehension