Clustering : K-means

Clustering

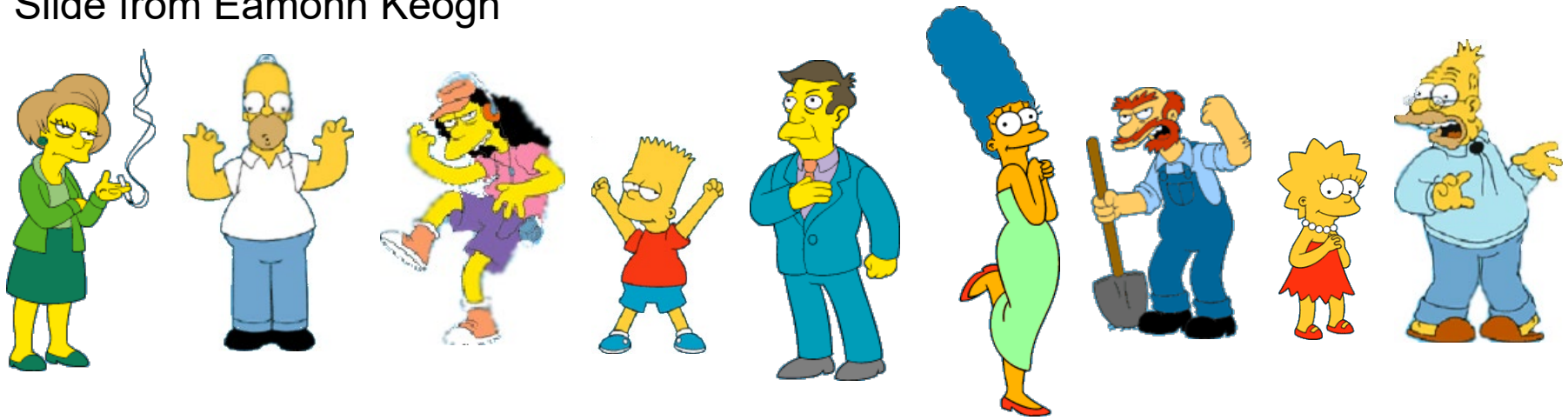
- Cluster : เป็นการจัดกลุ่มของข้อมูลประเภทต่างๆ โดยสามารถนำมาจัดกลุ่มกันที่มีคุณสมบัติคล้ายกันโดยวัดตามลักษณะของข้อมูลที่มีความคล้ายคลึงกัน (Similarity)

- Cluster Analysis เป็นกระบวนการจัดวัตถุต่างๆ ให้อยู่กลุ่มที่เหมาะสม ซึ่งมีคุณสมบัติที่วัตถุที่อยู่ในกลุ่มเดียวกันจะคล้ายกัน แต่มีความแตกต่างจากวัตถุในกลุ่มอื่น

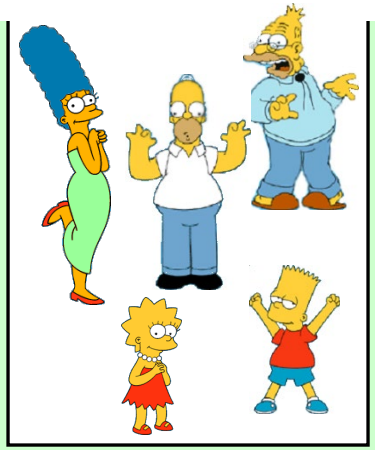
- Clustering การจัดกลุ่มจะแตกต่างจากการแบ่งประเภทข้อมูล (Classification) โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดคลาสประเภทข้อมูลไว้ก่อนหรือไม่ทราบจำนวนกลุ่มล่วงหน้า เป็นการเรียนรู้แบบไม่มีผู้สอน (unsupervised Learning)

What is a natural grouping of these objects?

Slide from Eamonn Keogh

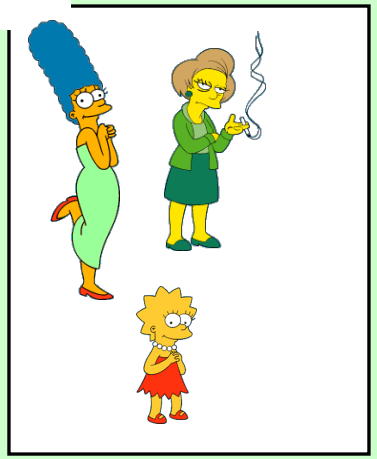


Clustering is subjective



Simpson's Family

School Employees



Females

Males

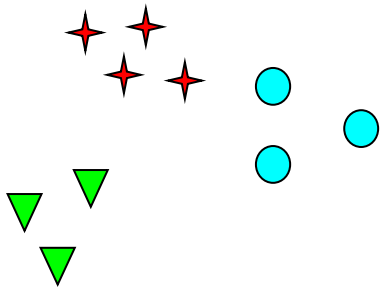
What is Similarity?

Slide based on one by Eamonn Keogh

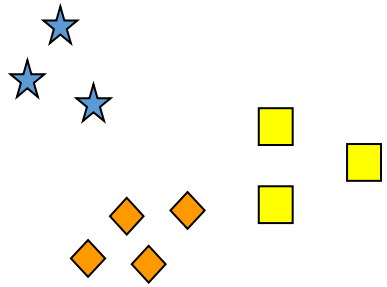
Similarity is hard to define, but…
"*We know it when we see it*"

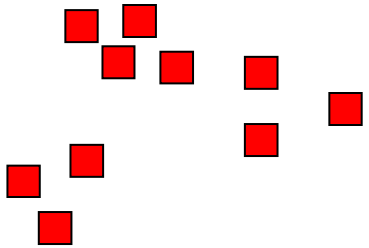
ตัวอย่าง clustering

How many clusters?

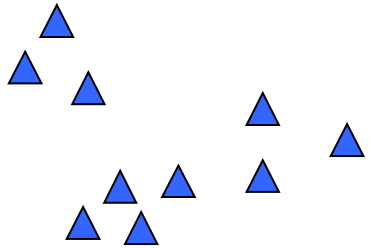


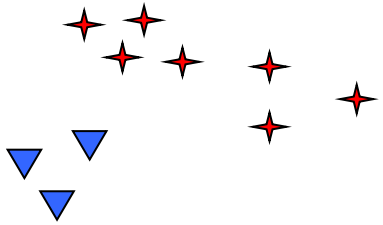
Six Clusters



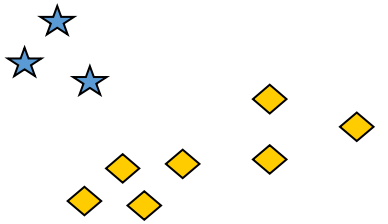


Two Clusters



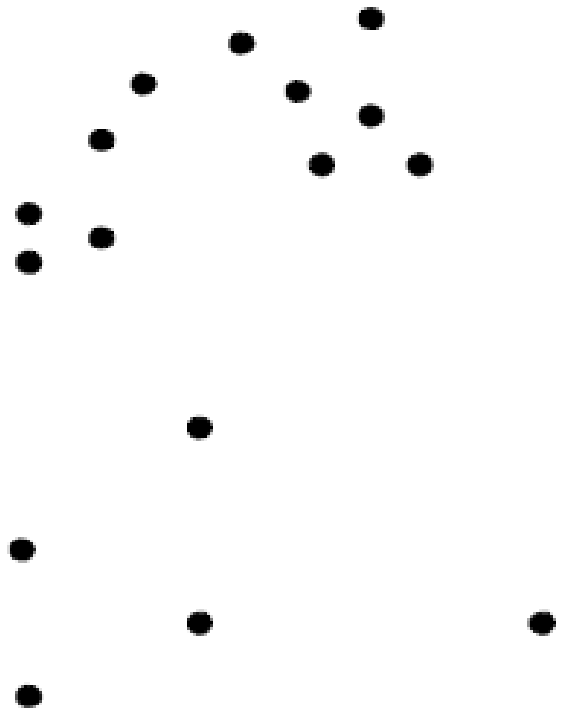


Four Clusters

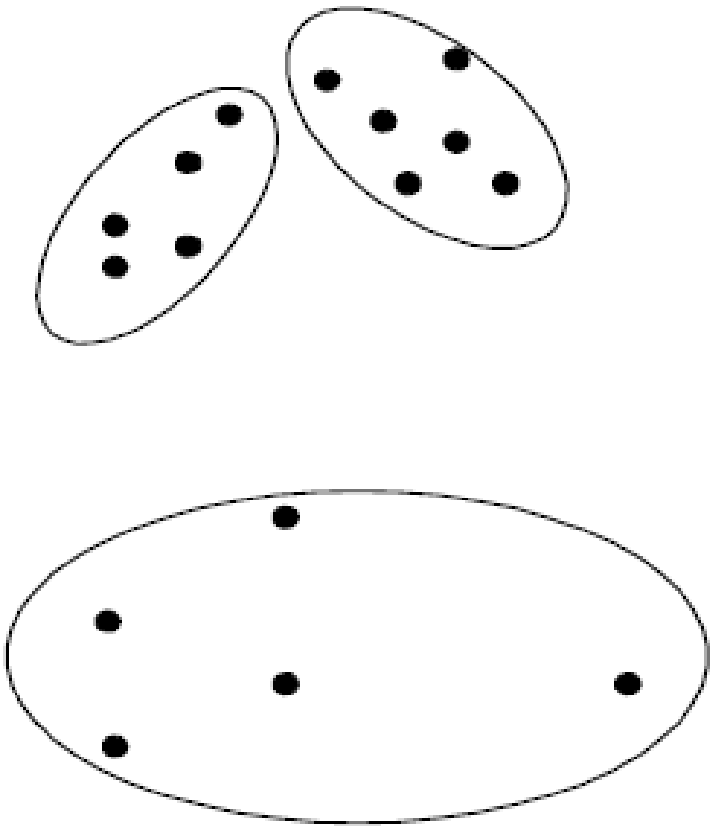


ประเภท Clustering

- Partitional Clustering คือการแบ่งกลุ่มอย่างชัดเจนโดยไม่มีกลุ่มใดซ้อนทับกันอยู่

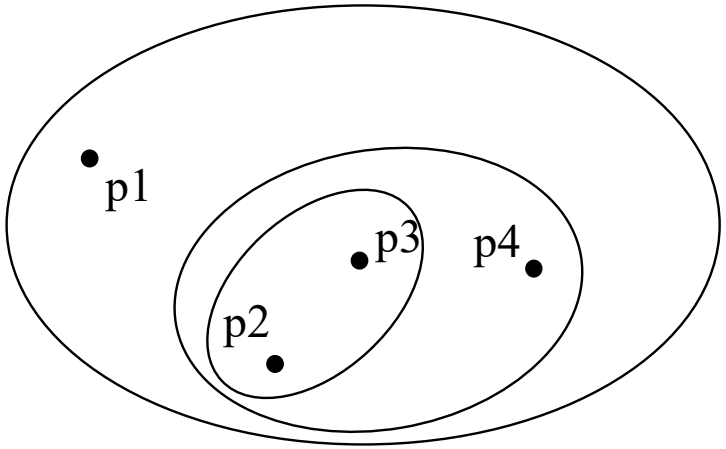


Original Points

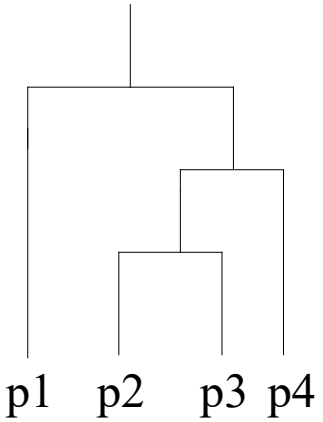


A Partitional Clustering

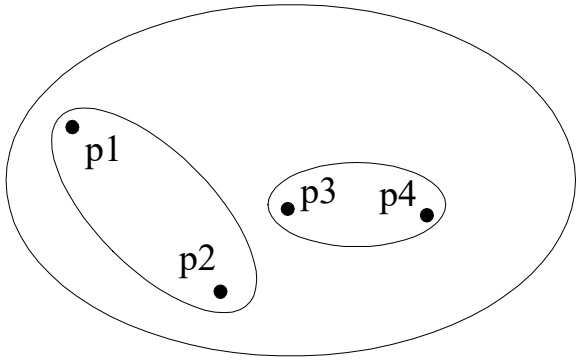
ประเภท Clustering : Hierarchical Clustering



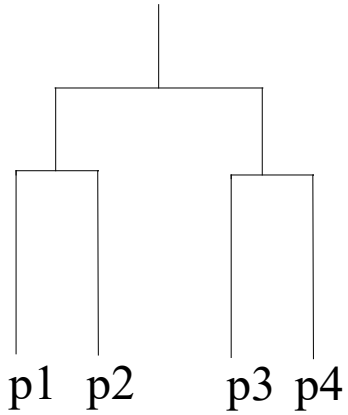
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



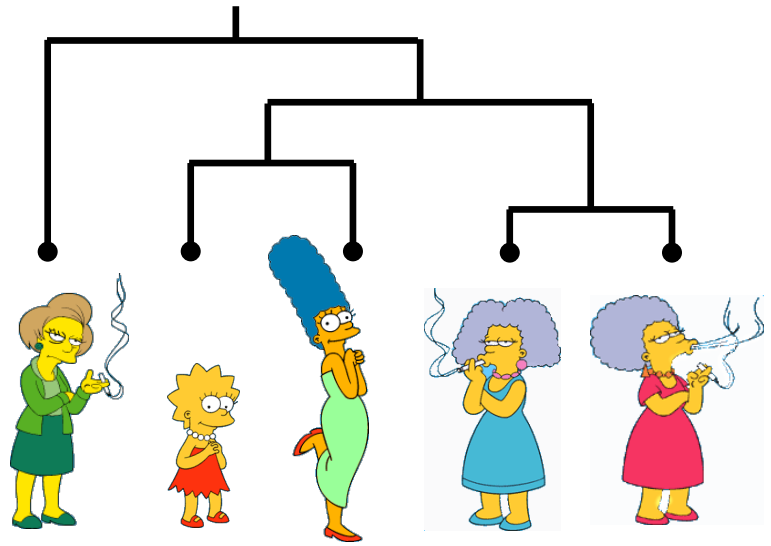
Non-traditional Dendrogram

Slide based on one by Eamonn Keogh

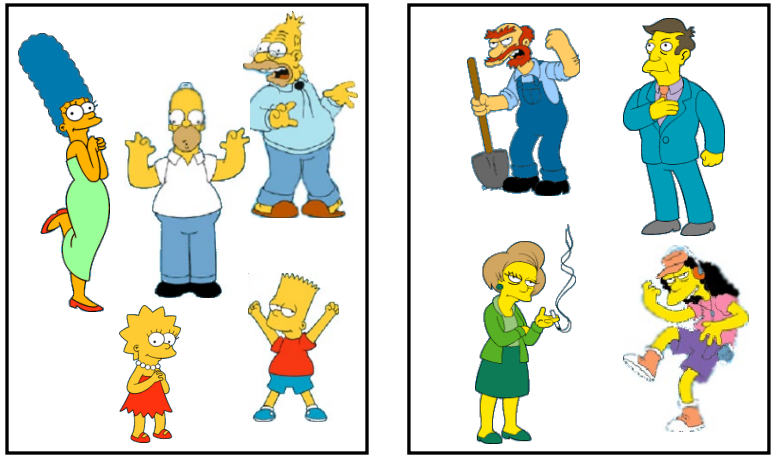
Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical

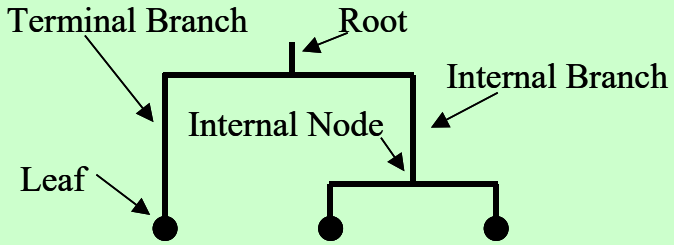


Partitional

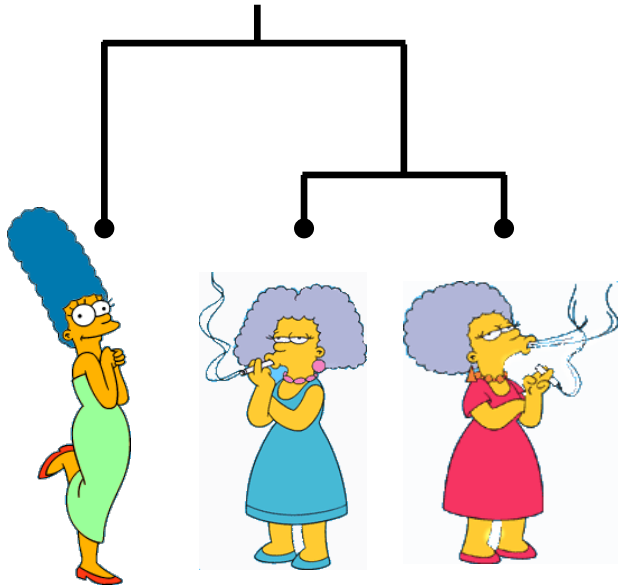


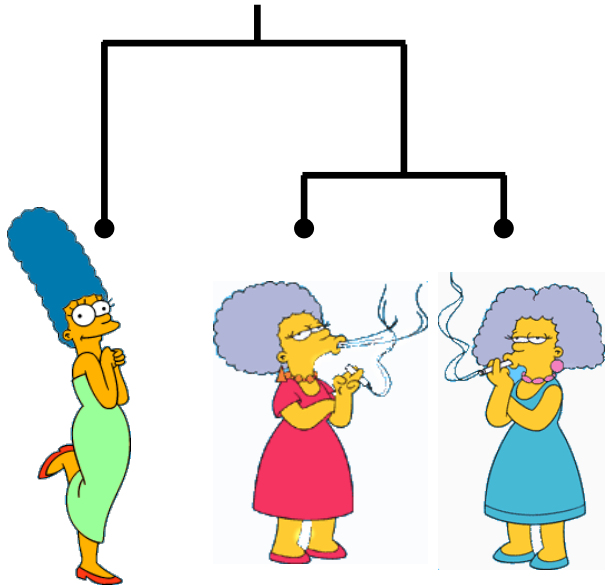
Slide based on one by Eamonn Keogh

Dendogram: A Useful Tool for Summarizing Similarity Measurements



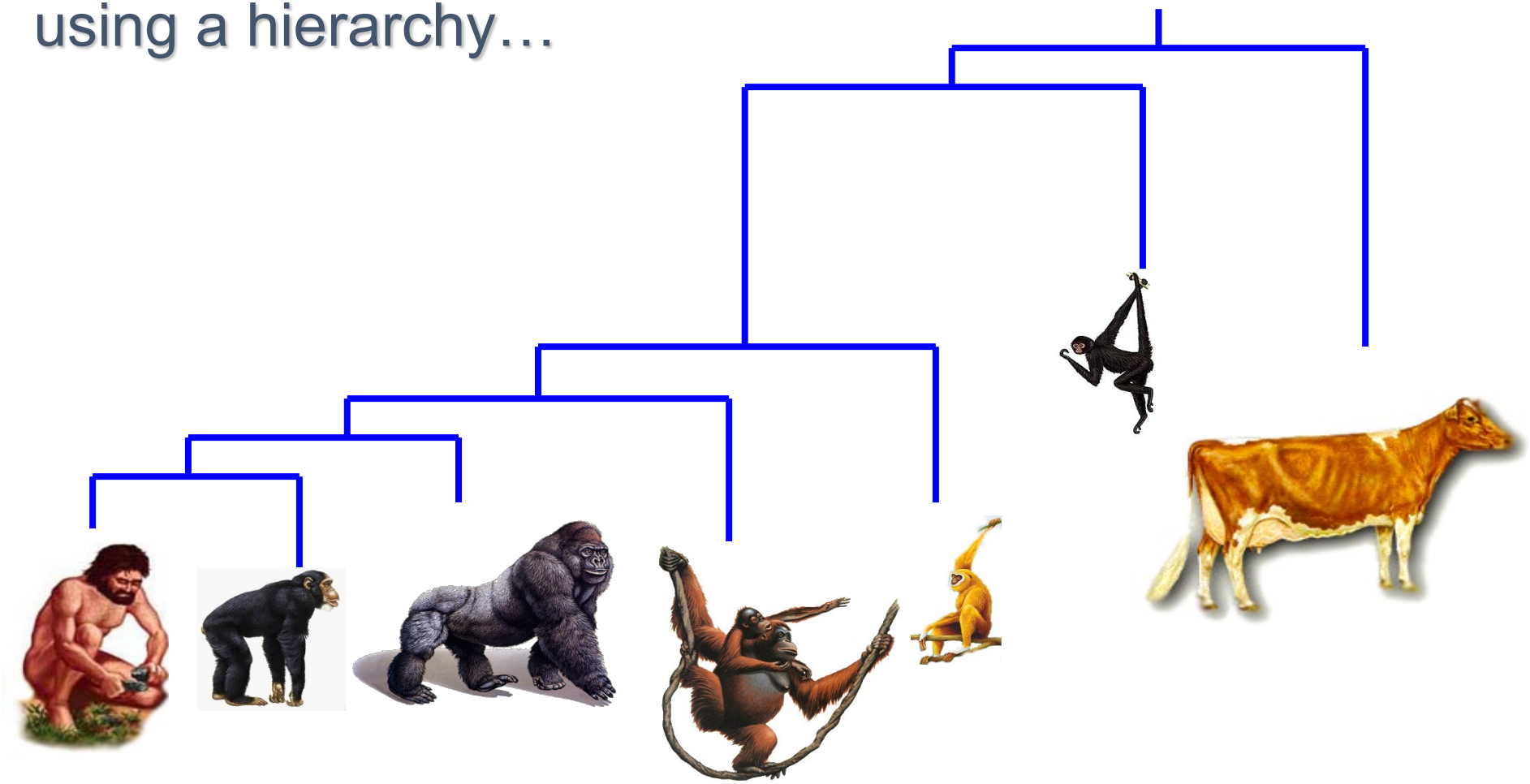
The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.





Slide based on one by Eamonn Keogh

There is only one dataset that can be perfectly clustered using a hierarchy…



(Bovine:0.69395, (Spider Monkey 0.390, (Gibbon:0.36079,(Orang:0.33636,(Gorilla:0.17147,(Chimp:0.19268, Human:0.11927):0.08386):0.06124):0.15057):0.54939);

ประเภท Clustering : K-Mean

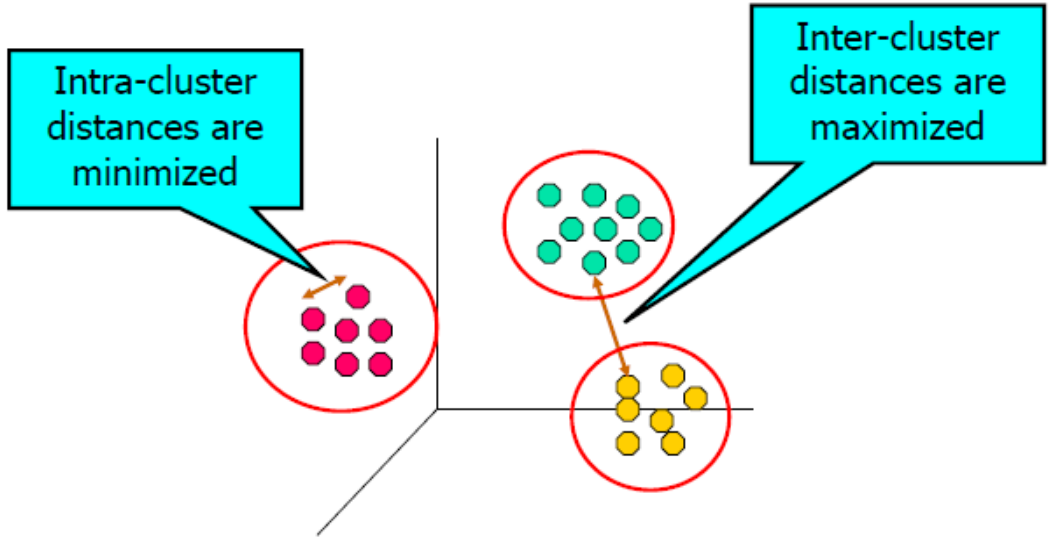
- **K-means clustering**
  - K-means หรือเรียกอีกอย่างหนึ่งว่า การวิเคราะห์กลุ่มแบบไม่เป็นลำดับชั้น (Nonhierarchical Cluster Analysis) หรือ การแบ่งส่วน (Partitioning)
  - เป็นอัลกอริทึมเทคนิคการเรียนรู้โดยไม่มีผู้สอนที่ง่ายที่สุด เพราะเป็นการแก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป โดยอัลกอริทึม K-Meansจะตัดแบ่ง (Partition) วัตถุออกเป็น K กลุ่ม
  - แทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน

ประเภทของตัวแปรที่ใช้ในเทคนิค K-Means Clustering

- ต้องเป็นตัวแปรเชิงปริมาณ (Quantitative)

  สเกลอันตรภาค (Interval Scale)

  สเกลอัตราส่วน (Ratio Scale)



- จัดกลุ่มโดยพยายามให้ระยะห่างของสิ่งที่อยู่ในกลุ่มเดียวกันอยู่ใกล้กันให้มากที่สุด (Minimize Intra-Cluster Distances) และระยะห่างที่อยู่ต่างกลุ่มมีความห่างแตกต่างกันมากที่สุด (Maximize Inter-Cluster Distances)

K-means Clustering Algorithm

1) กำหนดหรือสุ่มค่าเริ่มต้น จำนวน k ค่า(กลุ่ม) และกำหนดจุดศูนย์กลางเริ่มต้น k จุด เรียกว่า cluster centers หรือ (centroid)

2) นำวัตถุทั้งหมดจัดเข้ากลุ่ม โดยทำการหาค่าระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง หากข้อมูลไหนใกล้ค่าจุดศูนย์กลางตัวไหนสุด อยู่กลุ่มนั้น

3) หาค่าเฉลี่ย (Mean) แต่ละกลุ่ม ให้เป็นค่าจุดศูนย์กลางใหม่

4) ทำซ้ำข้อ 2-3 จนกระทั่งจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลง

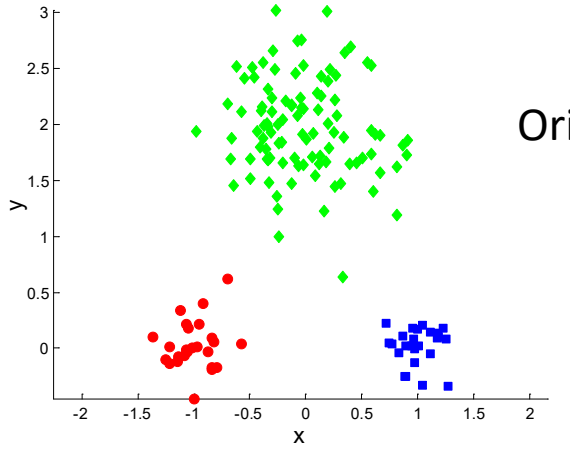
สมมติฐานหลักของ K-Mean คือ

1. มีอยู่ k กลุ่ม
2. Sum Square Error (SSE) ผลรวมของข้อผิดพลาดกำลังสอง เพื่อลดค่าความผิดพลาดของการจัดกลุ่มให้น้อยที่สุด
3. กลุ่มทั้งหมดมีค่า SSE เดียวกัน
4. ตัวแปรทั้งหมดมีความสำคัญเหมือนกันสำหรับทุกกลุ่ม

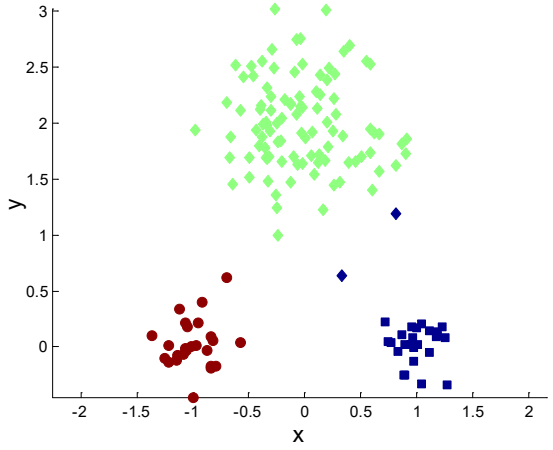
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

http://shabal.in/visuals/kmeans/4.html

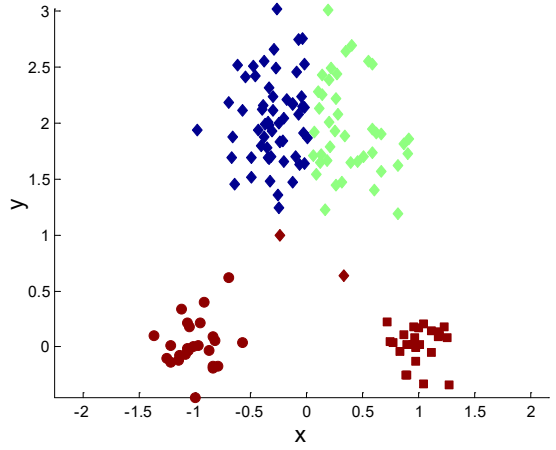
Two different K-means Clusterings



Original Points

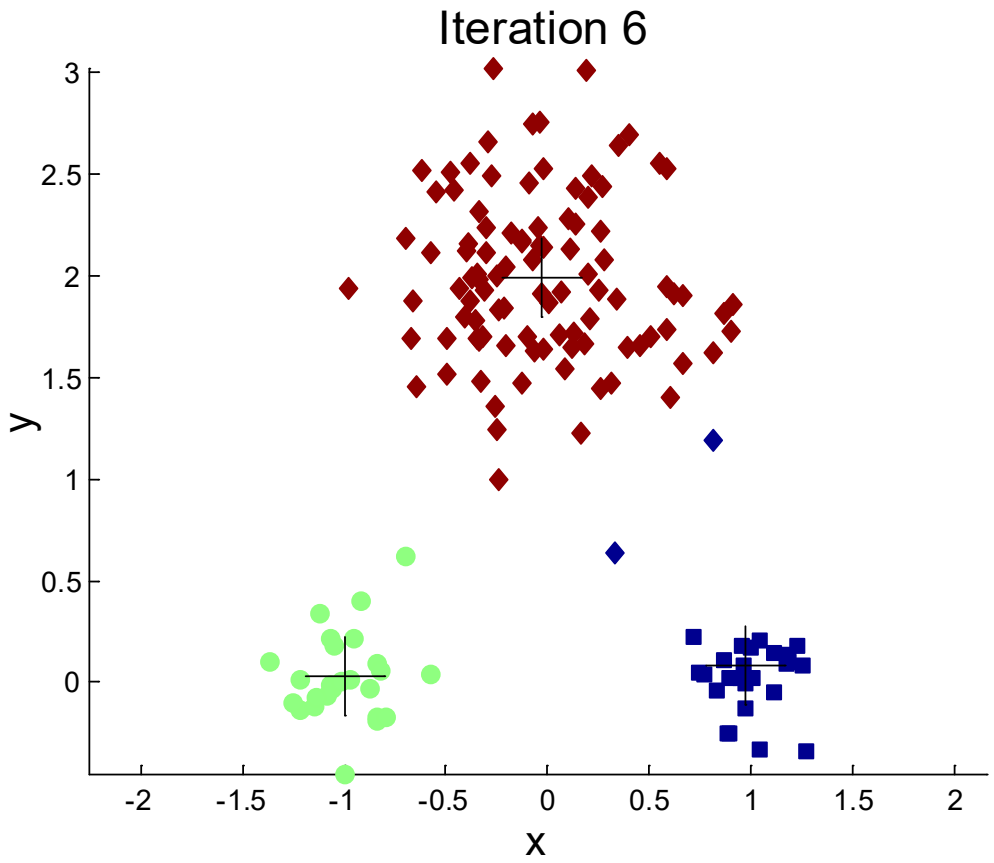


Optimal Clustering

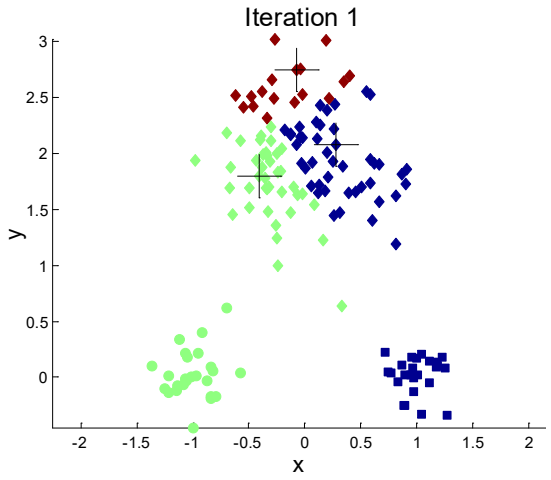


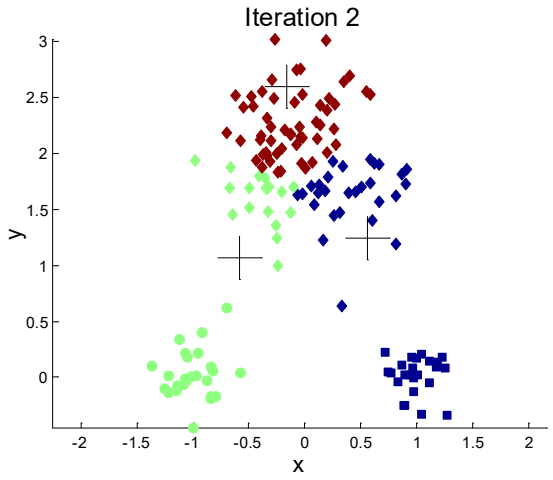
Sub-optimal Clustering

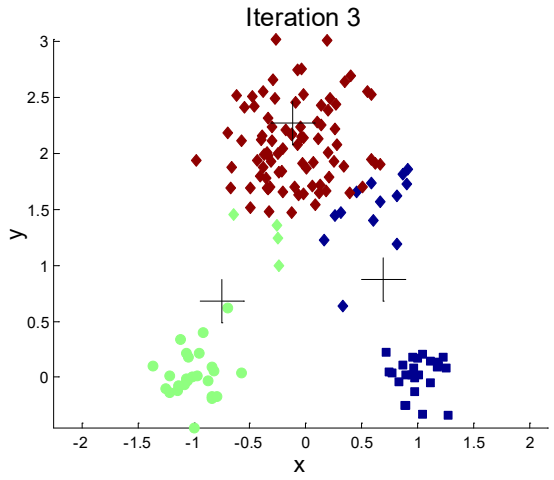
Importance of Choosing Initial Centroids

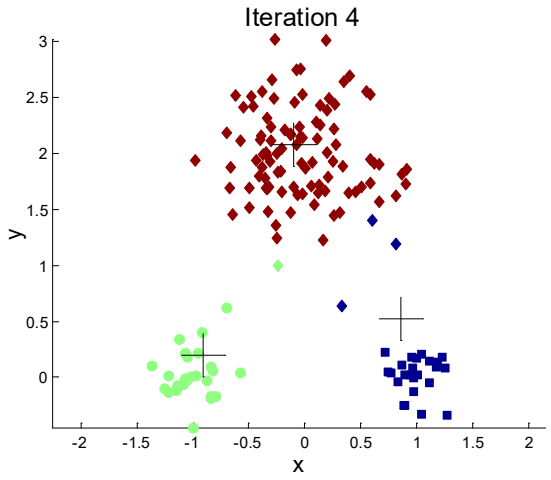


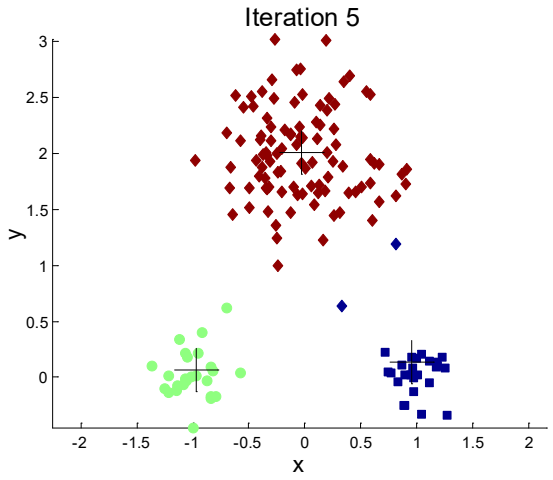
Importance of Choosing Initial Centroids

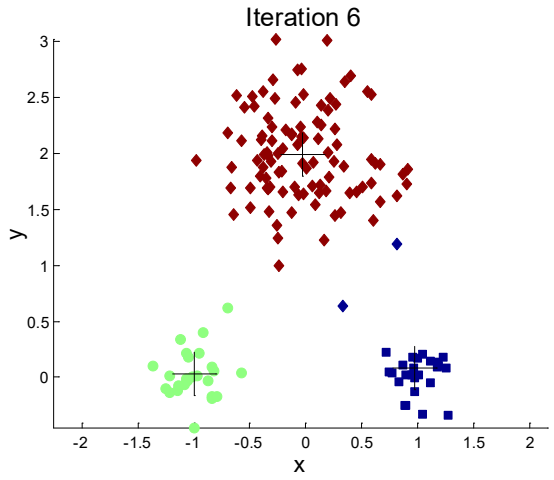




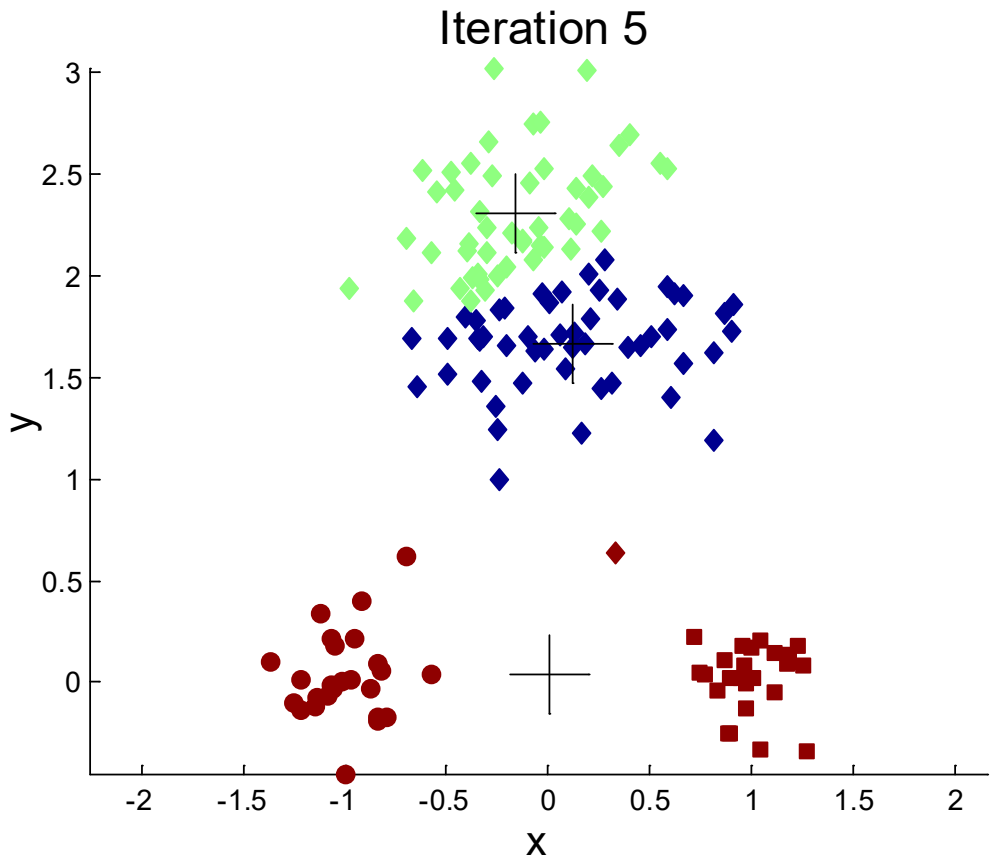




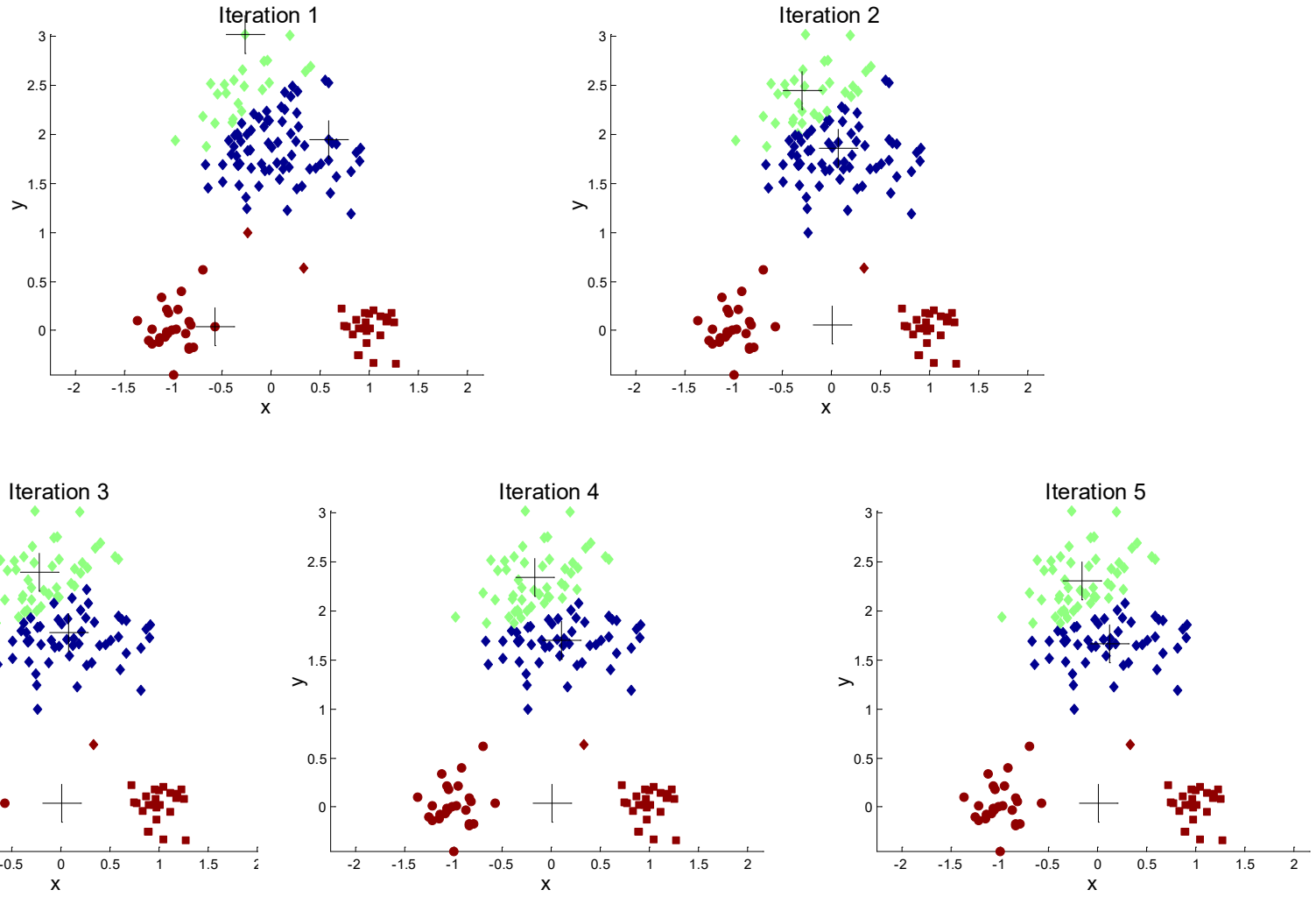




Importance of Choosing Initial Centroids …



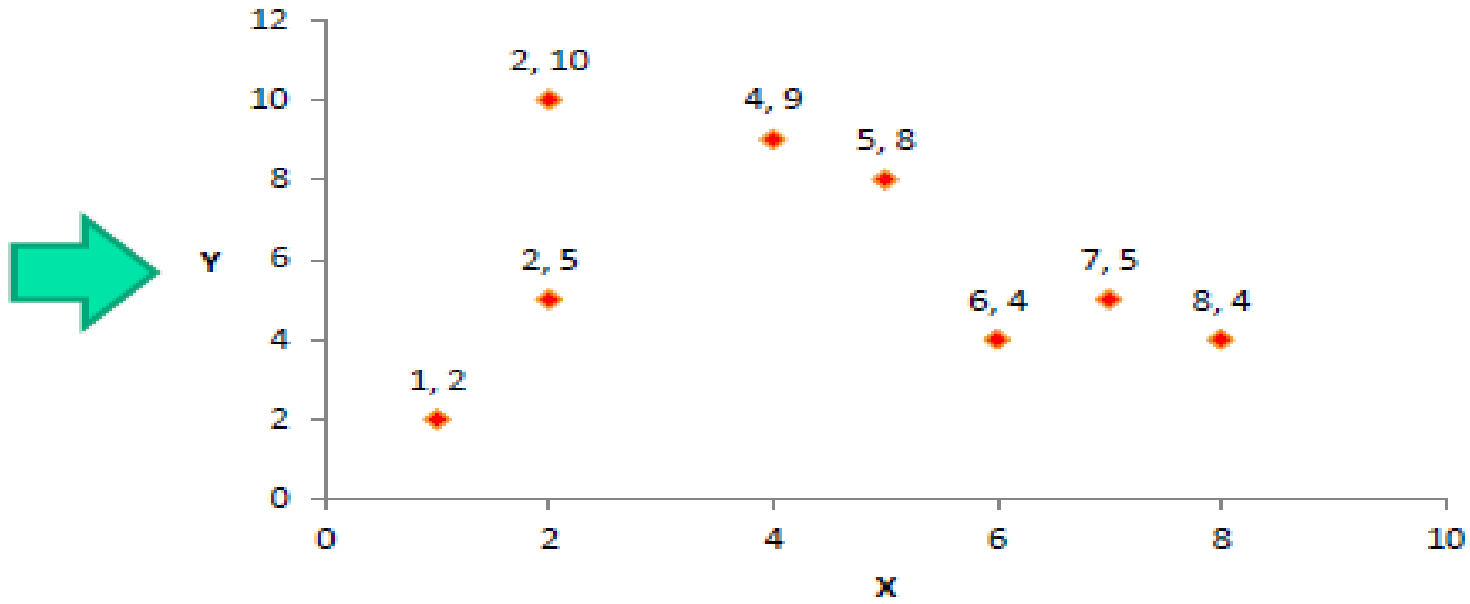
Importance of Choosing Initial Centroids …



ตัวอย่าง K-Mean Clustering

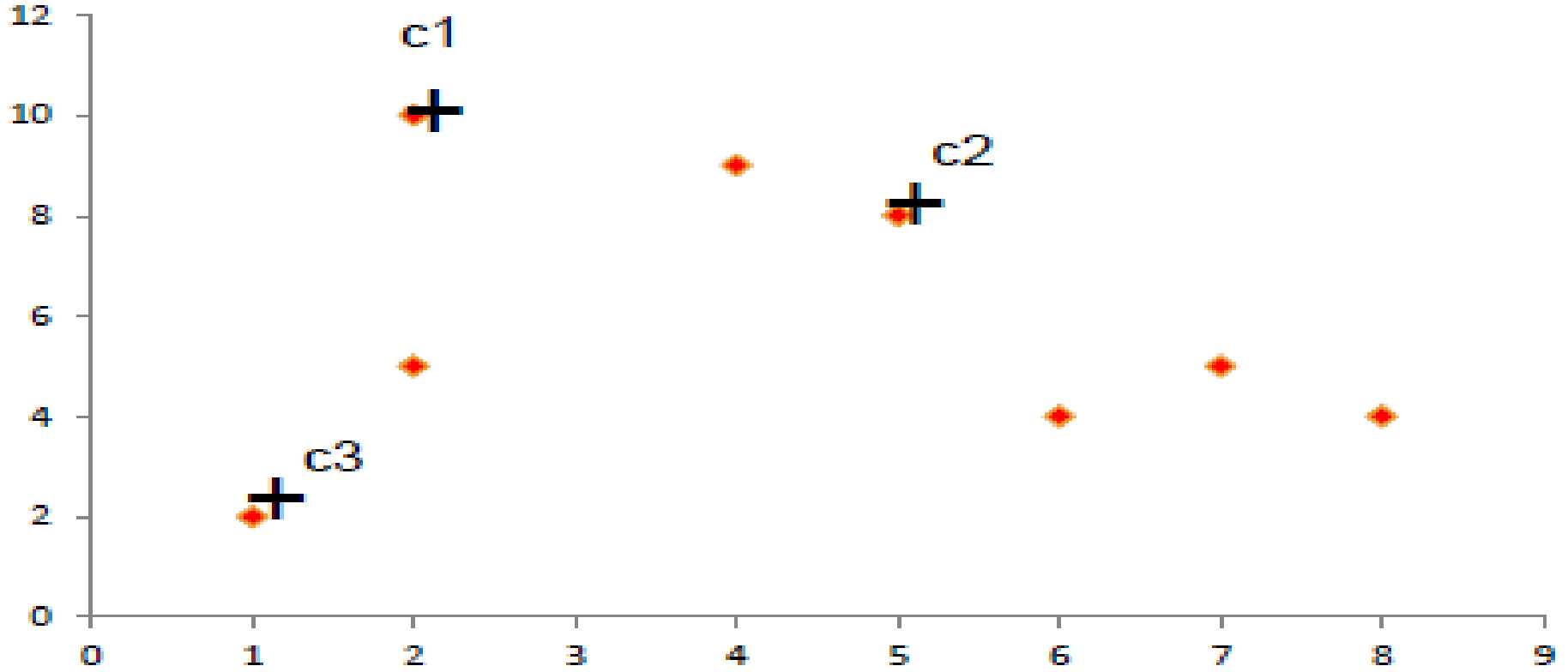
ID	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

k =3



Example: K-Mean Clustering

- สุ่มค่าเริ่มต้น จานวน k ค่า เรียกว่า cluster centers (centroid)
- สมมติ k =3 แสดงว่า c1, c2 และ c3 เป็น centroid ที่เราสุ่มขึ้นมา c1(2, 10), c2(5, 8) and c3(1, 2).



Example: K-Mean Clustering

- ขั้นตอนที่ 1 หาความห่างกันระหว่างข้อมูล 2 ข้อมูล คือ หาความห่างจากข้อมูล A =(x1, y1) และ centroid =(x2, y2) โดยใช้สูตร Euclidean

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} \qquad \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

		c1(2, 10)	c2 (5, 8)	c3(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

Example: K-Mean Clustering

- ขั้นตอนที่ 2 หาระยะห่างระหว่างข้อมูล กับจุดศูนย์กลาง (ตัวอย่างบางชุดข้อมูล)

point	mean1
x1, y1	*x2, y2*
(2, 10)	(2, 10)
distance(point, mean1)	$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ $= \sqrt{(2-2)^2 + (10-10)^2}$ $= 0$
point	mean2
x1, y1	*x2, y2*
(2, 10)	(5, 8)
distance(point, mean2)	$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ $= \sqrt{(2-5)^2 + (10-8)^2}$ $= 3.61$
point	mean3
x1, y1	*x2, y2*
(2, 10)	(1, 2)
distance(point, mean3)	$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ $= \sqrt{(2-1)^2 + (10-2)^2}$ $= 8.06$

รอบที่ 1 ได้การจัดกลุ่มข้อมูลดังต่อไปนี้

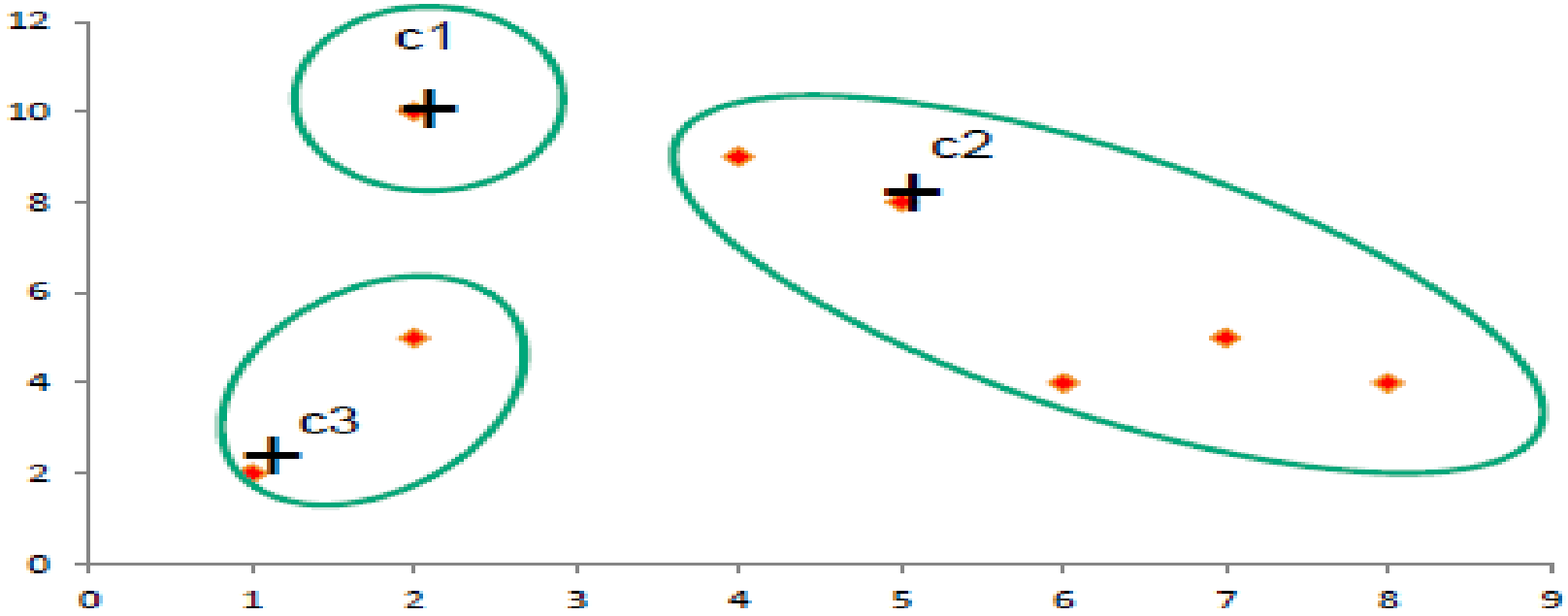
	Point	c1(2, 10)	c2(5, 8)	c3(1, 2)	Cluster
A1	(2, 10)	0.00	3.61	8.06	1
A2	(2, 5)	5.00	4.24	3.16	3
A3	(8, 4)	8.49	5.00	7.29	2
A4	(5, 8)	3.61	0.00	7.21	2
A5	(7, 5)	7.07	3.60	6.71	2
A6	(6, 4)	7.21	4.12	5.39	2
A7	(1, 2)	8.06	7.21	0.00	3
A8	(4, 9)	2.24	1.41	7.62	2

นำมาสร้างกลุ่มใหม่

Cluster 1
A1(2, 10)

Cluster 2
A3(8, 4)
A4(5, 8)
A5(7, 5)
A6(6, 4)
A8(4, 9)

Cluster 3
A2(2, 5)
A7(1, 2)



Example: K-Mean Clustering

- ขั้นตอนที่ 3 หาค่าเฉลี่ยแต่ละกลุ่ม ให้เป็นค่าจุดศูนย์กลางใหม่

Cluster 1	Cluster 2	Cluster 3
A1(2, 10)	A3(8, 4)	A2(2, 5)
	A4(5, 8)	A7(1, 2)
	A5(7, 5)	
	A6(6, 4)	
	A8(4, 9)	

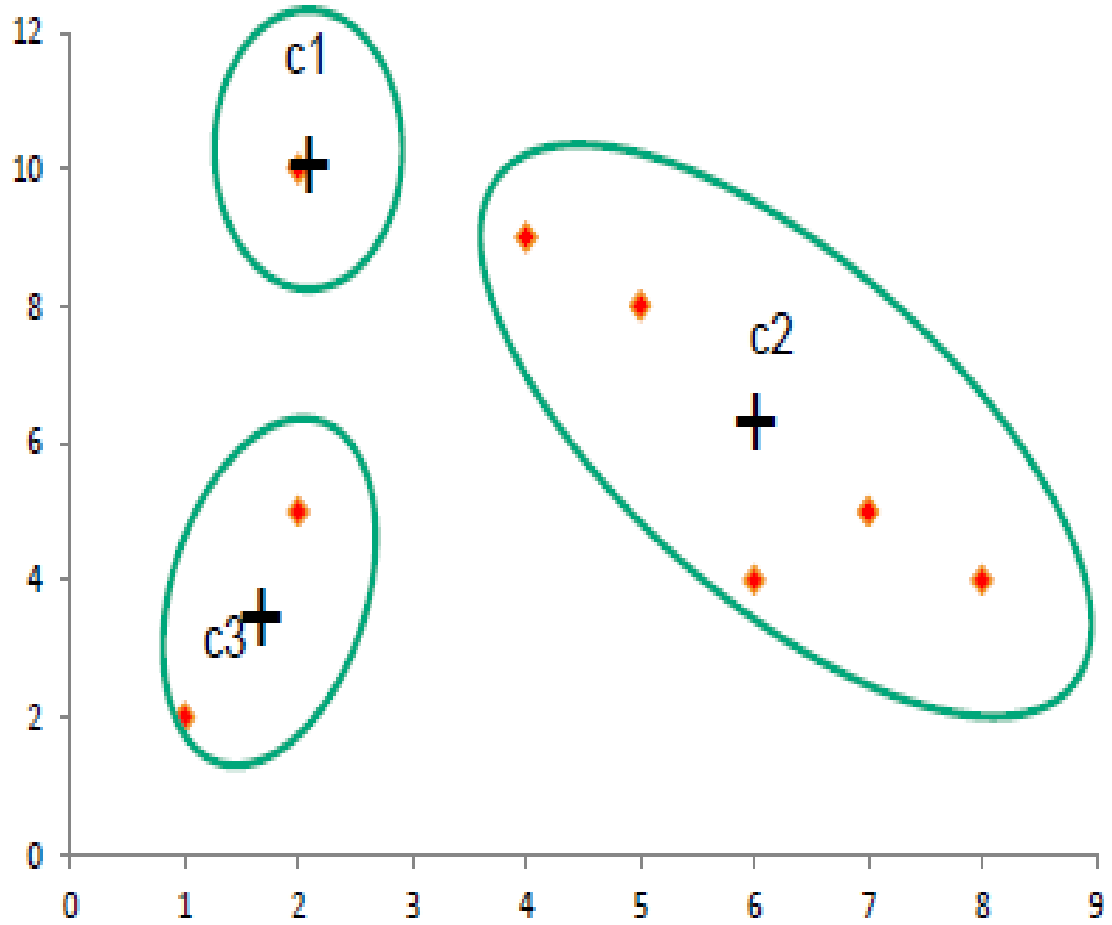
- สำหรับ Cluster 1 มีจุดเดียวคือ A1(2, 10) แสดงว่า C1(2,10) ยังคงเดิม
- สำหรับ Cluster 2 มี 5 จุดอยู่กลุ่มเดียวกัน เพราะฉะนั้นหา C2 ใหม่

  ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = C2(6, 6)

- สำหรับ Cluster 3 มี 2 จุดอยู่กลุ่มเดียวกัน ((2+1)/2, (5+2)/2) = C3(1.5, 3.5)

Example: K-Mean Clustering

- รอบที่ 2



	Point	c1(2, 10)	c2(6, 6)	c3(1.5, 3.5)	Cluster
A1	(2, 10)	0.00	5.66	6.52	1
A2	(2, 5)	5.00	4.12	1.58	3
A3	(8, 4)	8.49	2.83	6.52	2
A4	(5, 8)	3.60	2.24	5.70	2
A5	(7, 5)	7.07	1.41	5.70	2
A6	(6, 4)	7.21	2.00	4.53	2
A7	(1, 2)	8.06	6.40	1.58	3
A8	(4, 9)	2.24	3.61	6.04	1

Example: K-Mean Clustering

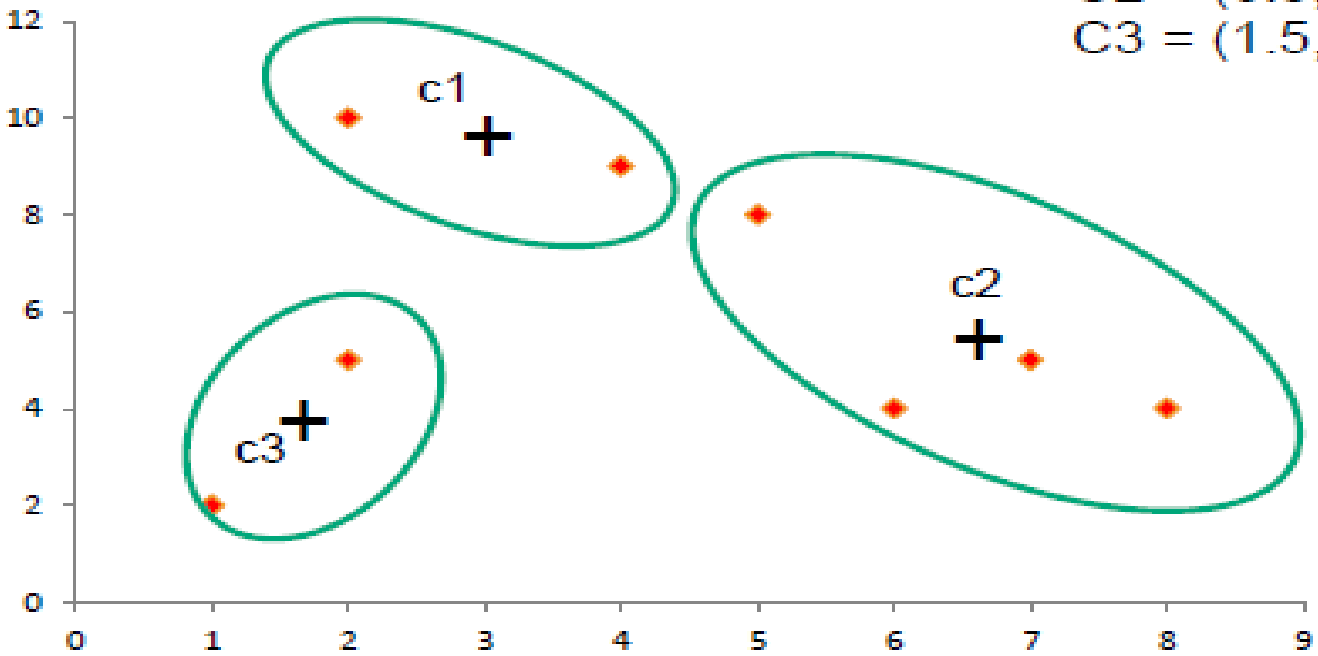
Cluster 1
A1(2, 10)
A8(4, 9)

Cluster 2
A3(8, 4)
A4(5, 8)
A5(7, 5)
A6(6, 4)

Cluster 3
A2(2, 5)
A7(1, 2)

คำนวณจุดศูนย์กลางใหม่

C1 = (2+4/2, 10+9/2) = (3, 9.5)
C2 = (6.5, 5.25)
C3 = (1.5, 3.5)



Example: K-Mean Clustering

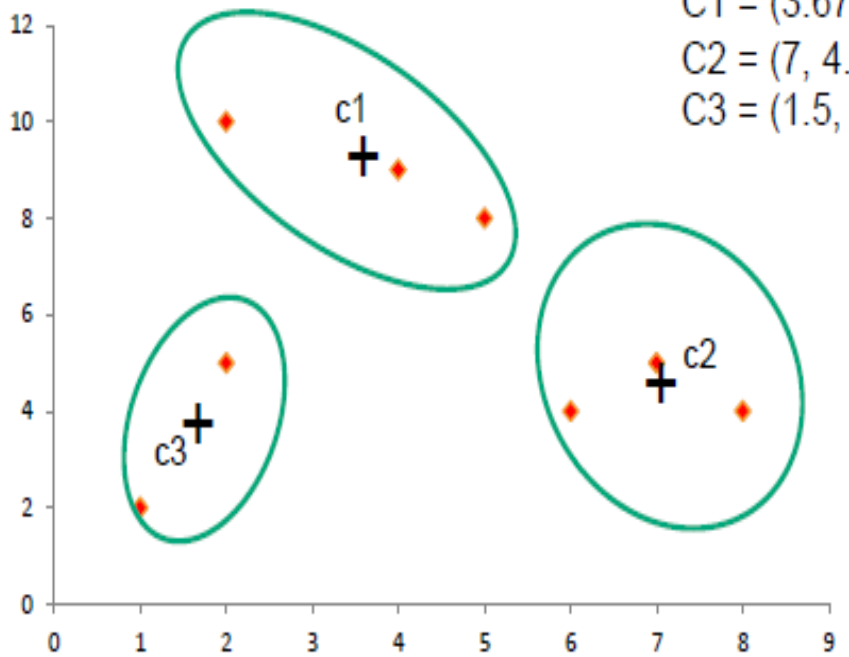
- รอบที่ 3

	Point	c1(3, 9.5)	c2(6.5, 5.25)	c3(1.5, 3.5)	Cluster
A1	(2, 10)	1.11	6.54	6.52	1
A2	(2, 5)	4.61	4.50	1.58	3
A3	(8, 4)	7.43	1.96	6.52	2
A4	(5, 8)	2.50	3.13	5.70	1
A5	(7, 5)	6.02	0.56	5.70	2
A6	(6, 4)	6.26	1.35	4.53	2
A7	(1, 2)	7.76	6.39	1.58	3
A8	(4, 9)	1.12	4.50	6.04	1

Cluster 1	Cluster 2	Cluster 3
A1(2, 10)	A3(8, 4)	A2(2, 5)
A8(4, 9)	A5(7, 5)	A7(1, 2)
A4(5, 8)	A6(6, 4)	

คำนวณจุดศูนย์กลางใหม่

C1 = (3.67, 9)
C2 = (7, 4.33)
C3 = (1.5, 3.5)



Example: K-Mean Clustering

- รอบที่ 4

	Point	c1(3.67, 9)	c2(7, 4.33)	c3(1.5, 3.5)	Cluster
A1	(2, 10)	1.94	7.56	6.52	1
A2	(2, 5)	4.33	5.04	1.58	3
A3	(8, 4)	6.62	1.05	6.52	2
A4	(5, 8)	1.67	4.18	5.70	1
A5	(7, 5)	5.21	0.67	5.70	2
A6	(6, 4)	5.52	1.05	4.53	2
A7	(1, 2)	7.49	6.44	1.58	3
A8	(4, 9)	0.33	5.55	6.04	1

Cluster 1	Cluster 2	Cluster 3
A1(2, 10)	A3(8, 4)	A2(2, 5)
A8(4, 9)	A5(7, 5)	A7(1, 2)
A4(5, 8)	A6(6, 4)	

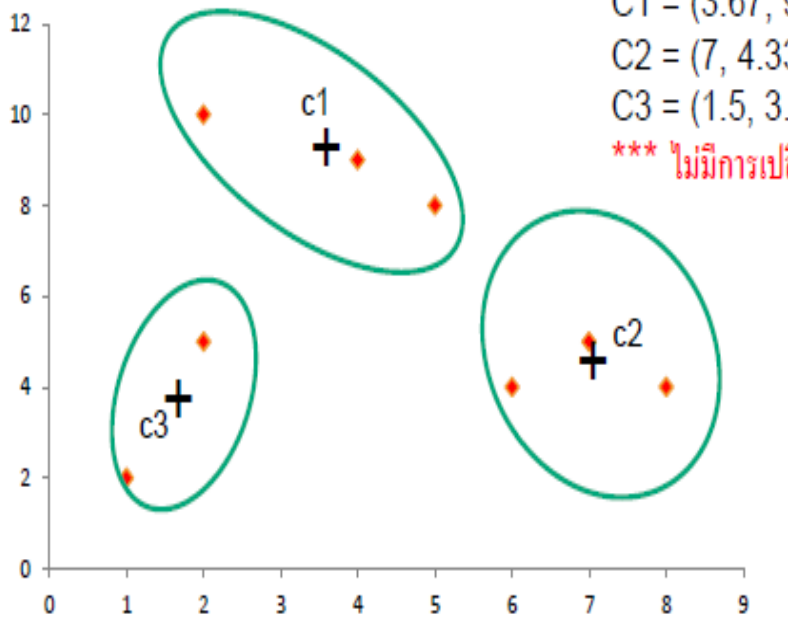
คำนวณจุดศูนย์กลางใหม่

C1 = (3.67, 9)

C2 = (7, 4.33)

C3 = (1.5, 3.5)

*** ไม่มีการเปลี่ยนแปลง **** STOP



Um, what about k?

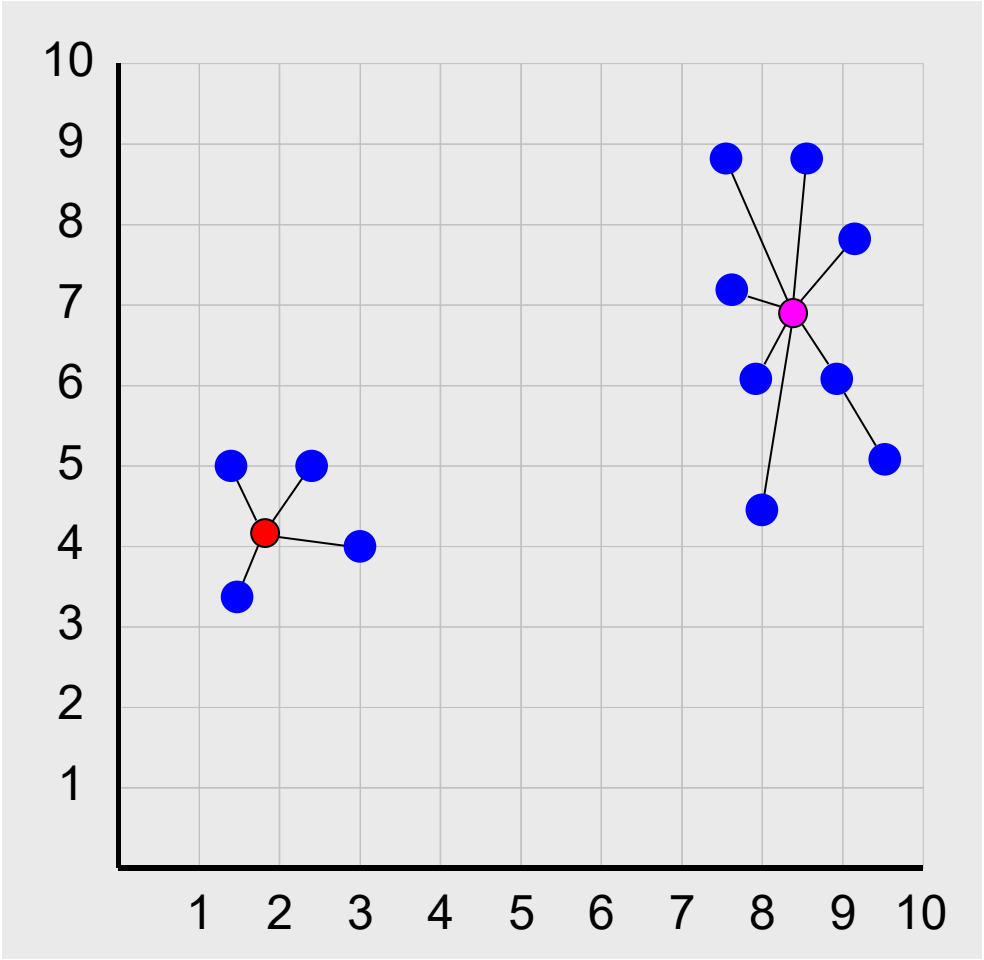
- Idea 1: Use our new trick of cross validation to select k
  - What should we optimize? SSE? Trace?
  - Problem?
- Idea 2: Let our domain expert look at the clustering and decide if they like it
  - How should we show this to them?
  - Problem?
- Idea 3: The “knee” solution

Slide based on one by Eamonn Keogh

Squared Error

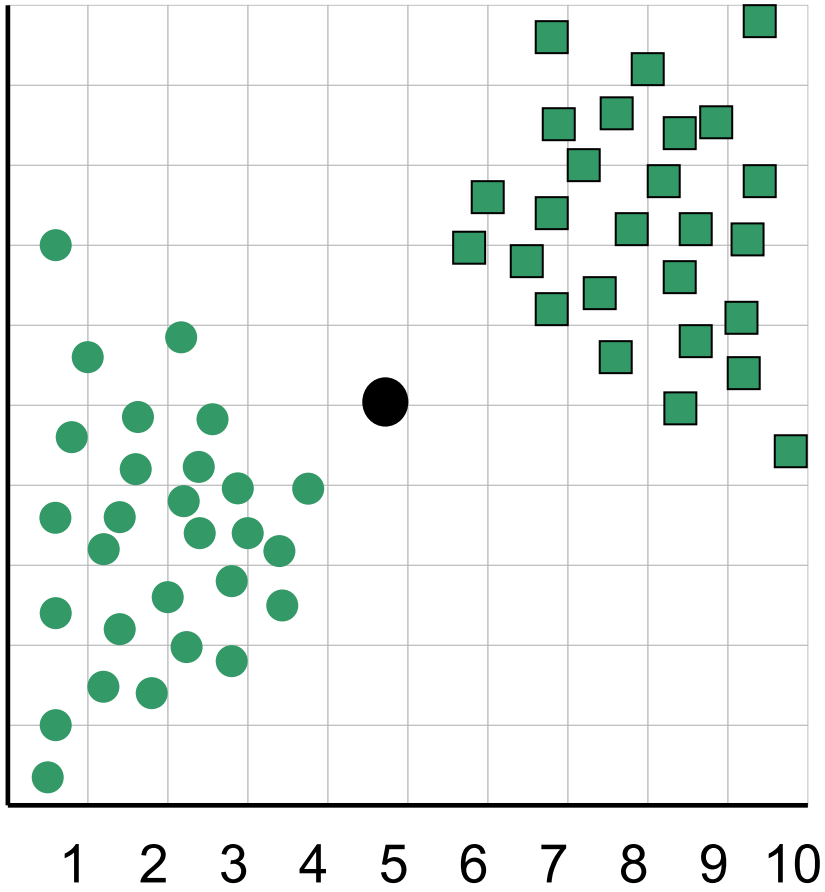
$$se_{K_i} = \sum_{j=1}^{m} \|t_{ij} - C_k\|^2$$

$$se_K = \sum_{j=1}^{k} se_{K_j}$$



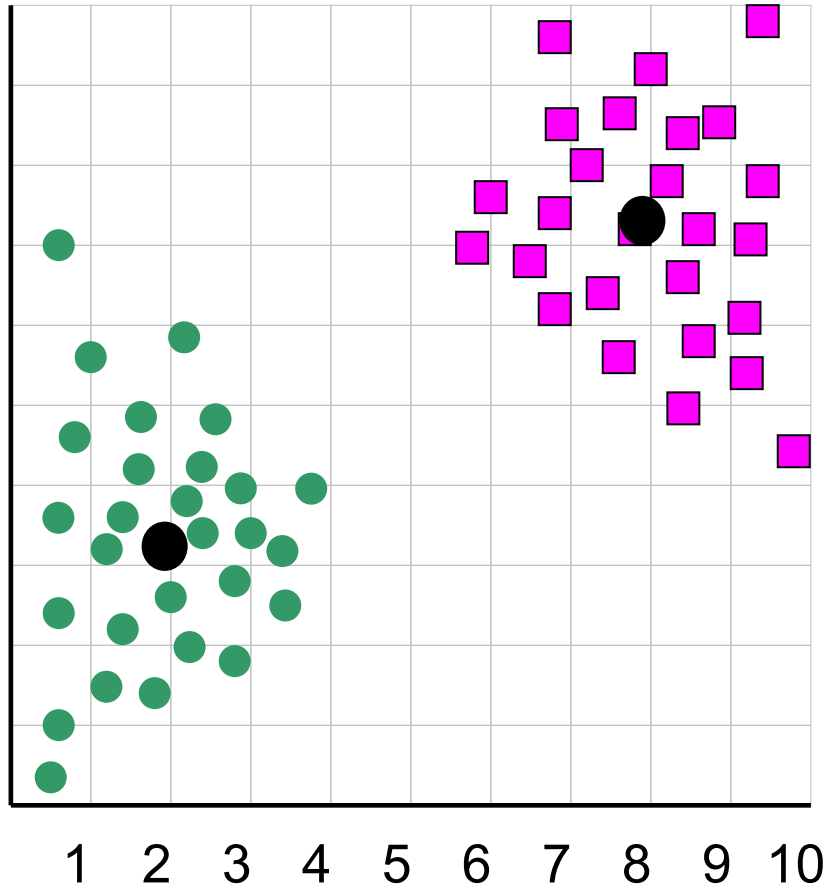
Objective Function

When k = 1, the objective function is 873.0



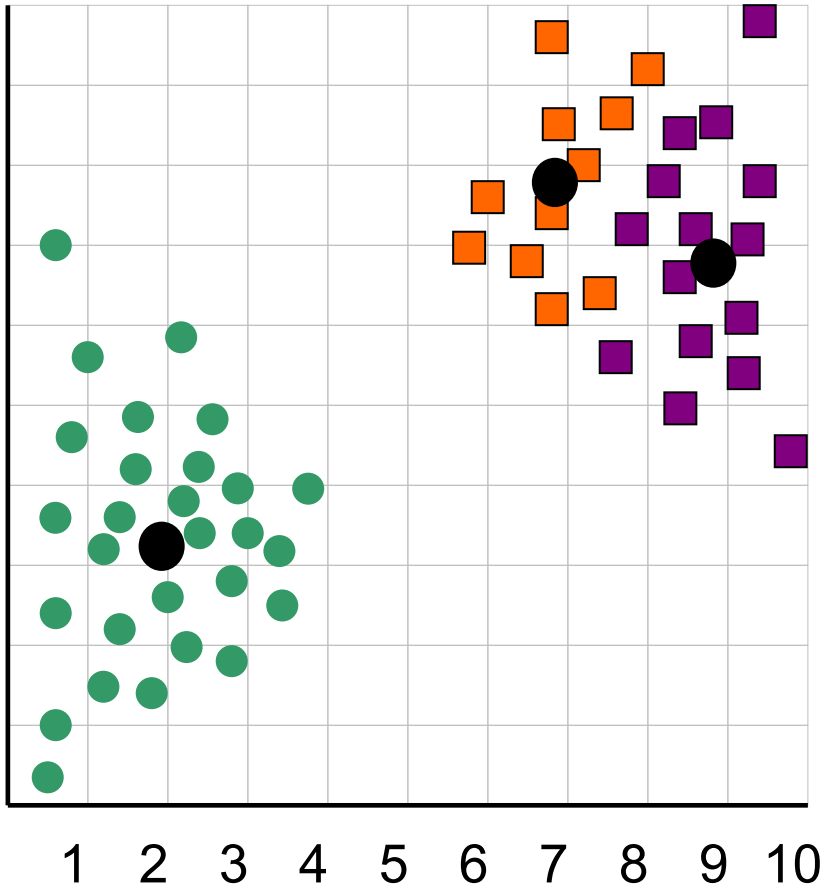
Slide based on one by Eamonn Keogh

When k = 2, the objective function is 173.1



Slide based on one by Eamonn Keogh

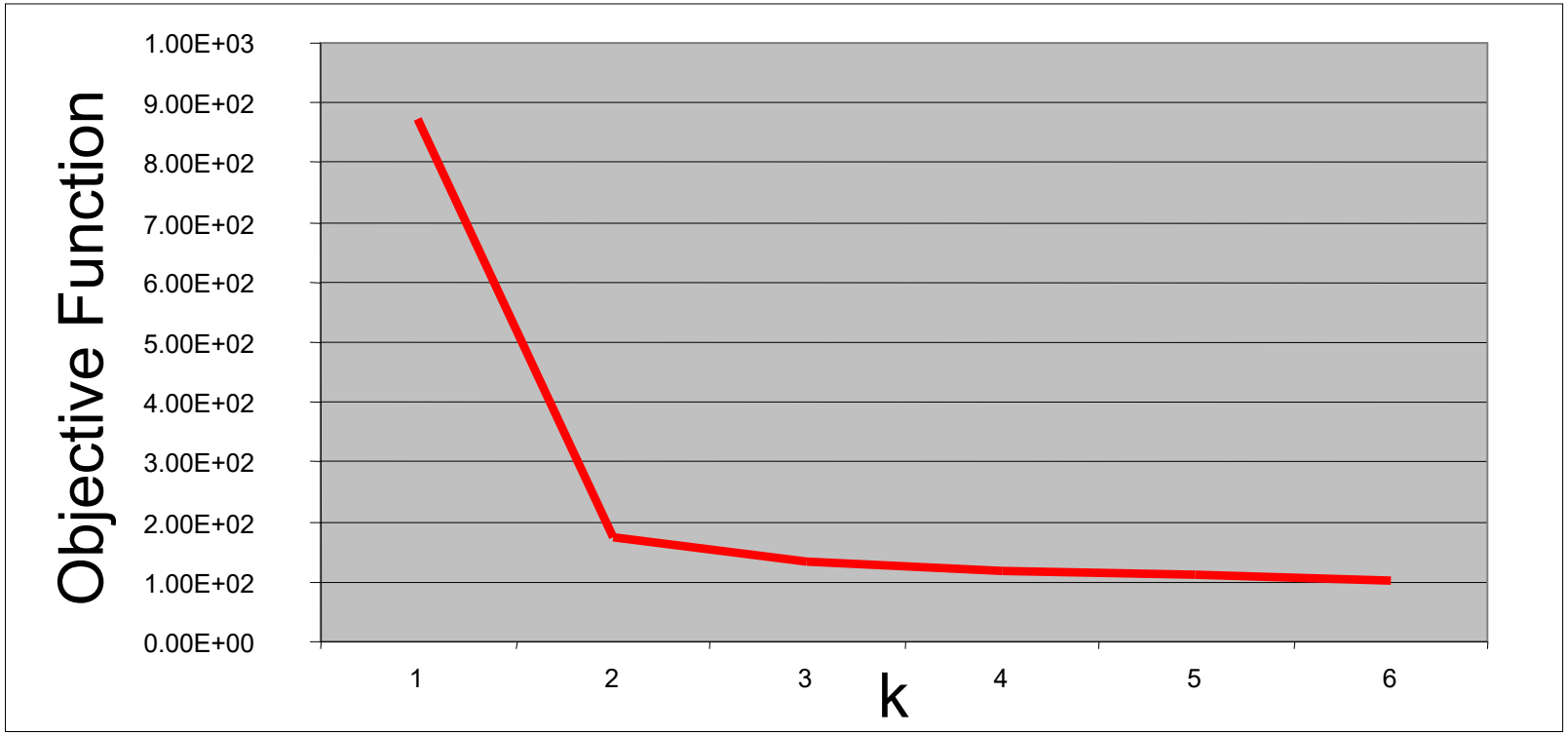
When k = 3, the objective function is 133.6



Slide based on one by Eamonn Keogh

We can plot the objective function values for k equals 1 to 6…

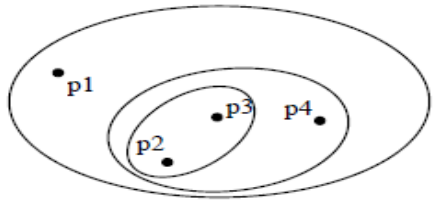
The abrupt change at k = 2, is highly suggestive of two clusters in the data. This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.



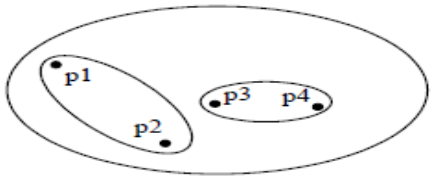
Slide based on one by Eamonn Keogh

ข้อดีของ K-Means Clustering

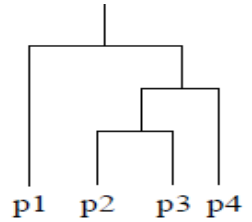
- เมื่อจำนวนข้อมูลมีจำนวนมาก และมีจำนวนกลุ่มน้อย การหาค่าเฉลี่ยแบบ K-means อาจจะคำนวณได้เร็วกว่าการจัดกลุ่มแบบอื่น ๆ เช่น การจัดกลุ่มแบบลำดับชั้น (Hierarchical)



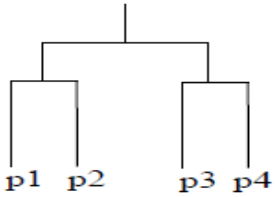
Hierarchical Clustering#1



Hierarchical Clustering#2



Traditional Dendrogram 1



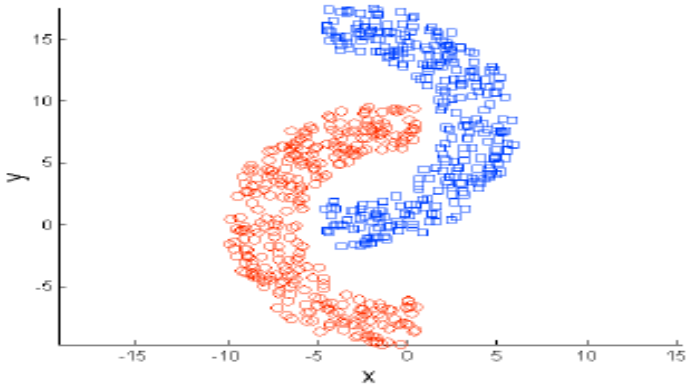
Traditional Dendrogram 2

- ขั้นตอนการหาค่าเฉลี่ยแบบ K-means อาจจะได้สมาชิกภายในกลุ่มหนาแน่นกว่าการจัดกลุ่มแบบ Hierarchical โดยเฉพาะถ้ากลุ่มเป็นวงกลม

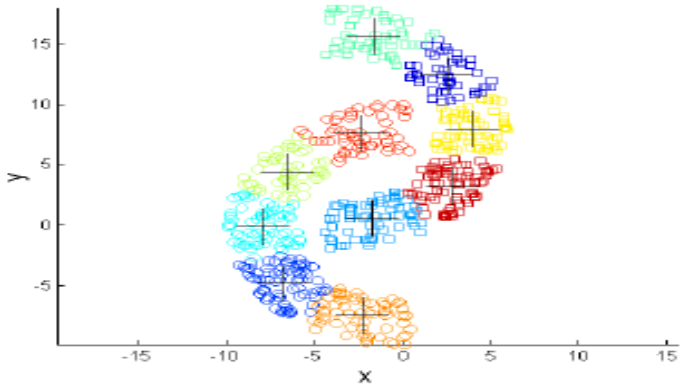
ข้อเสียของ K-Means Clustering

- การหาค่า K ที่เหมาะสมคาดเดาได้ยาก
- ทำงานได้ไม่ดีถ้ากลุ่มข้อมูลไม่เป็นรูปวงกลม
- มีข้อจำกัดในเรื่องของ
  - ขนาด ถ้าจำนวนข้อมูลน้อย จะใช้ได้ผลไม่ดี
  - ความหนาแน่น ถ้าความหนาแน่นน้อยจะจัดกลุ่มได้ไม่ดี
  - รูปร่างของข้อมูลที่ไม่มีการกระจายเป็นรูปร่างวงกลม

K-means Limitations :รูปร่าง ขนาด ความหนาแน่น

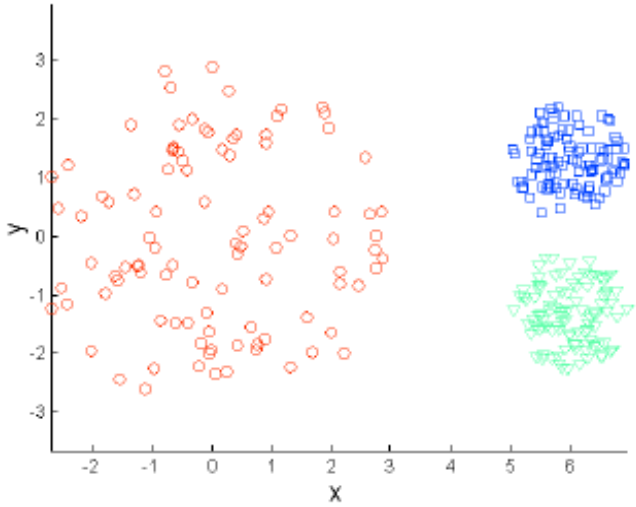


Original Points

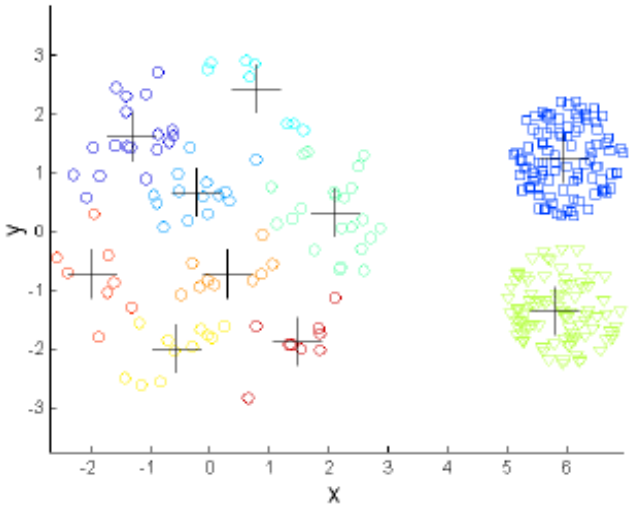


K-means Clusters

รูปร่าง



Original Points



K-means Clusters

ขนาด และ ความหนาแน่น