Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

Remind

Uh, Mom?
Remember that paper
you wrote for me about nuclear
reactors? What does it mean
when the big, red “meltdown”
light is flashing?

AI Awareness Poster, Brigham Young University
https://live-academicintegrity.pantheonsite.io/wp-content/uploads/2017/12/462c19_1dc7602ba4ec4f4586d803e7c43904cc_mv2-1.gif

Remind about the assignments

- They are "**individual assignments**": you can check slides/books/Internet for the answer, but not referring to your friends, classmates or anyone else.

- **Any** (public and private) discussion in any form about the assignment **before due** are **not** allowed.

- If you have any question regarding the assignment, please send an email to the teaching group (only involving you and the teaching group).

More information: https://ppl.app.uq.edu.au/content/3.60.04-student-integrity-and-misconduct

Last week: Association Rule Mining

- For a dataset of transactions
  - Each transaction is a set of items

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- Mining strong rules from the dataset

MIN_SUP count = 2

MIN_CONF= 60%

	support	confidence
Beer→ Bread, Diapers	40%	66.7%
Bread, Diapers→ Beer	40%	66.7%
Beer, Diapers→ Bread	40%	66.7%
Beer, Bread→ Diapers	40%	100%
…	…	…

More about the “items”

Name	Eat/Drink	Price	Weights	State	Fresh	…
Bread	Y	6$	200g	Solid	Some	…
Milk	Y	5$	1200g	Liquid	Very	…
Diapers	N	10$	3000g	Solid	A bit	…
Beer	Y	3$	300g	Liquid	Some	…
Cola	Y	2$	300g	Liquid	A bit	…
Eggs	Y	4$	600g	Solid	Very	…
…	…	…	…	…	…	…

With these “descriptions” of items, can we find (mining out) more relationships among the items?

Lecture 3: Clustering 1

Clustering

Given a set of ***objects***, each object with descriptions of its properties, ***partition*** them into groups based on the descriptions, such that objects within one group are similar, and objects from different groups are dissimilar.

Colour	Taste
Green	Sour

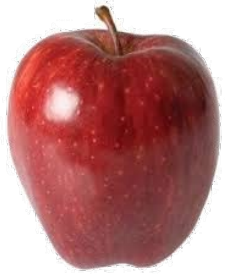
Colour	Taste
Red	Sweet

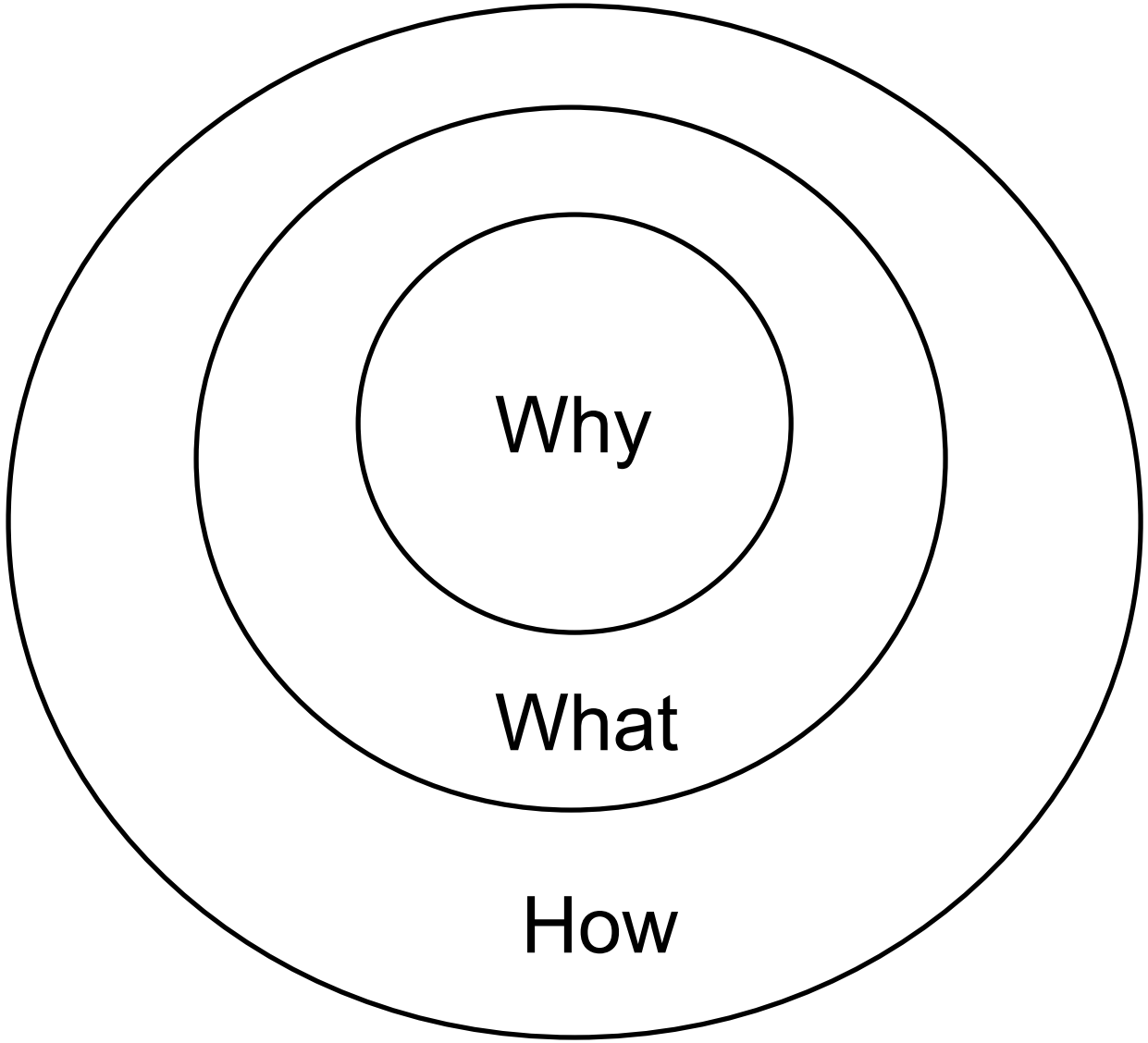
Colour	Taste
Red	Sweet

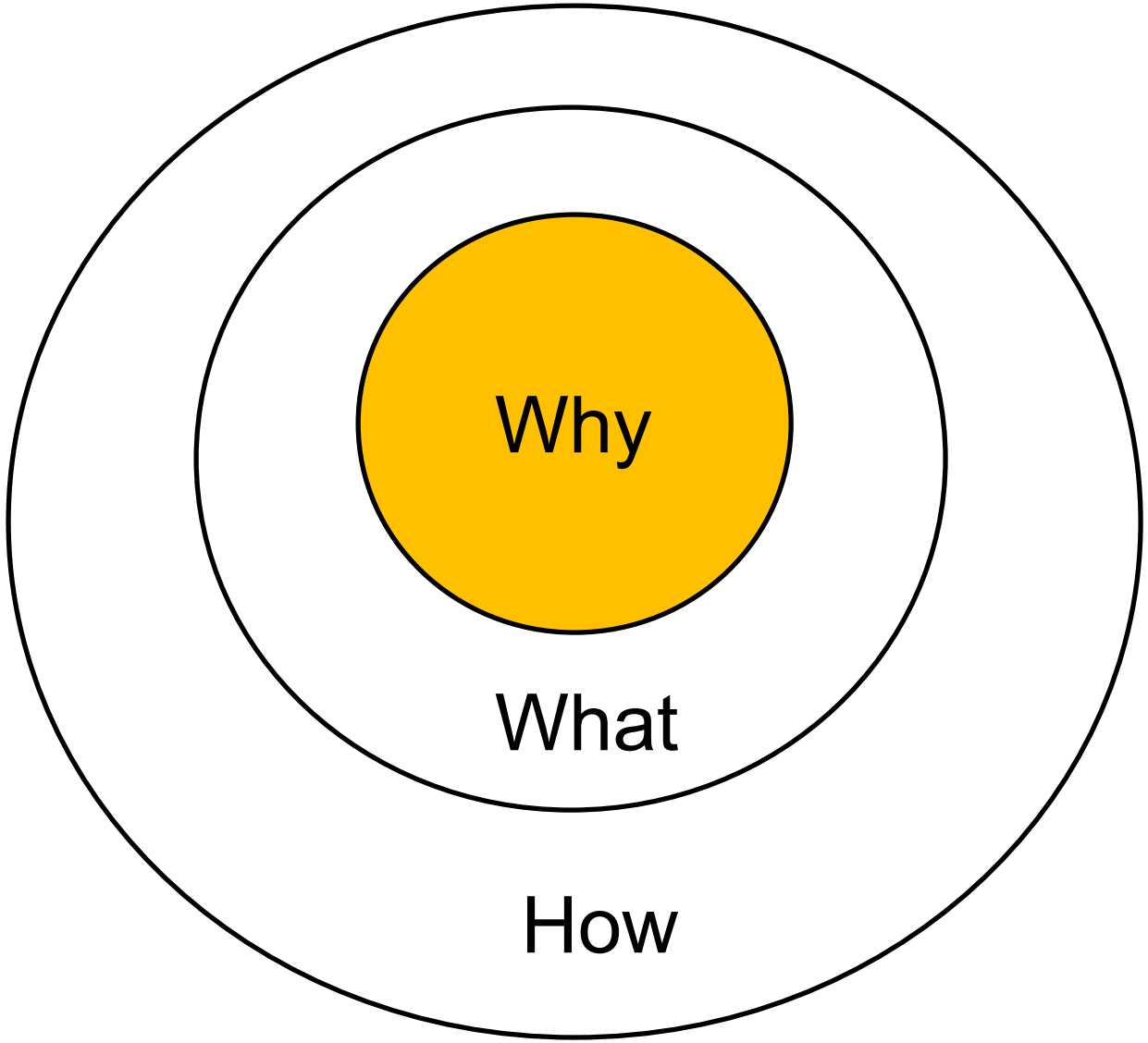
Colour	Taste
Red	Sweet

Colour	Taste
Green	Sour

Clustering

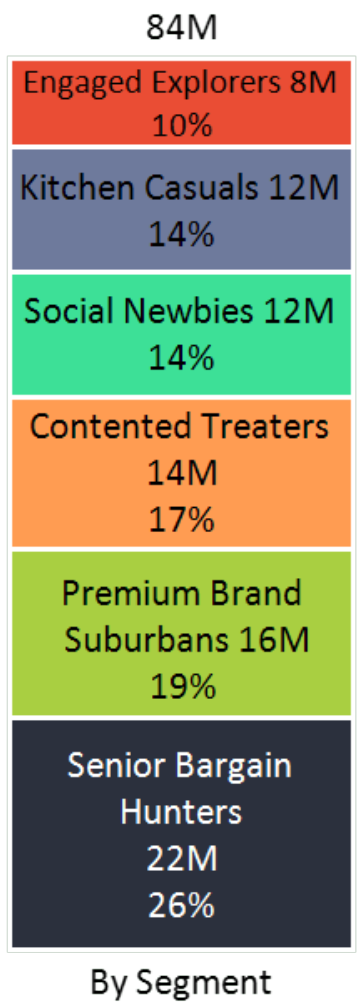


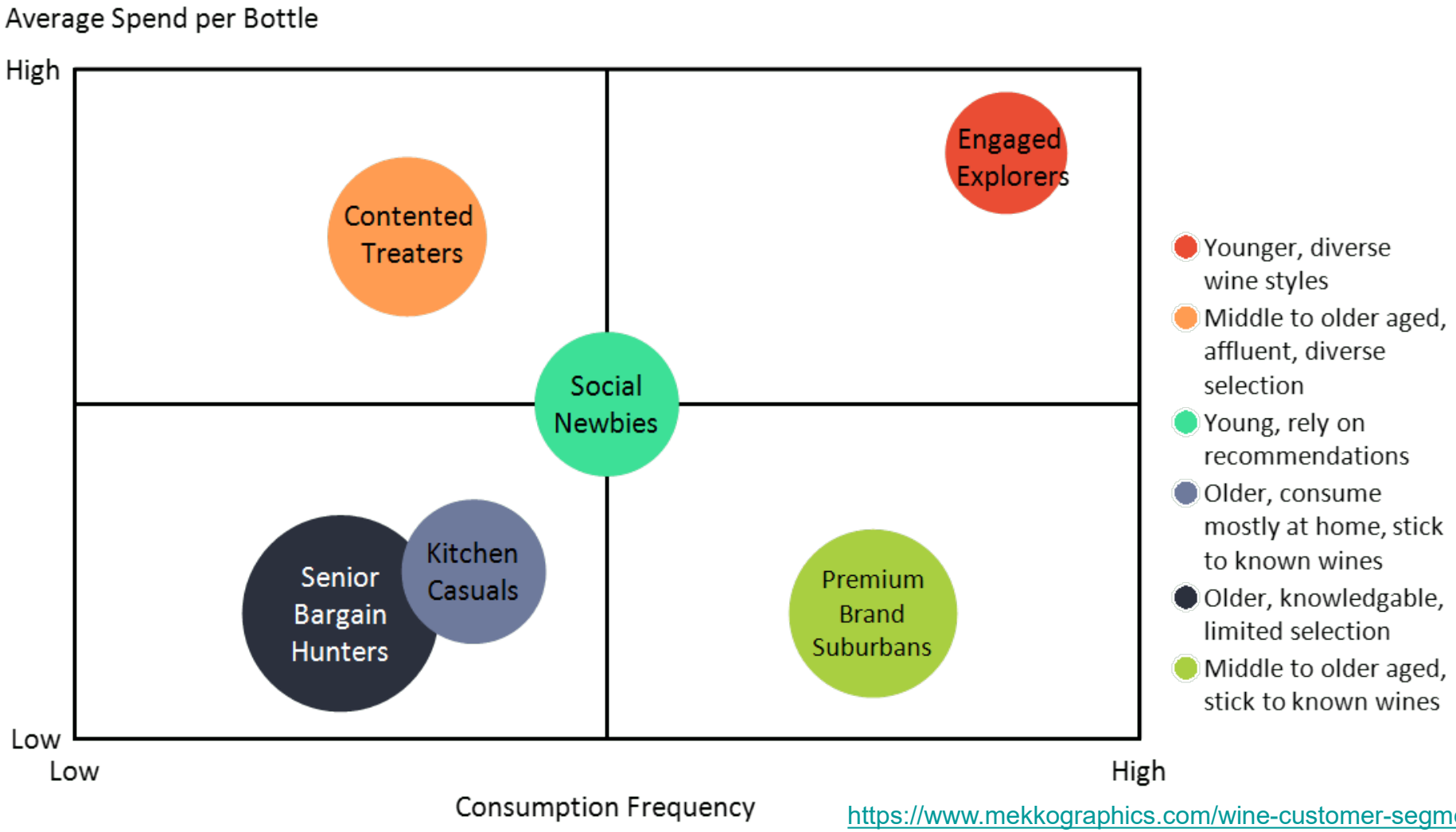

Why
What
How


Why
What
How

Customer segmentation

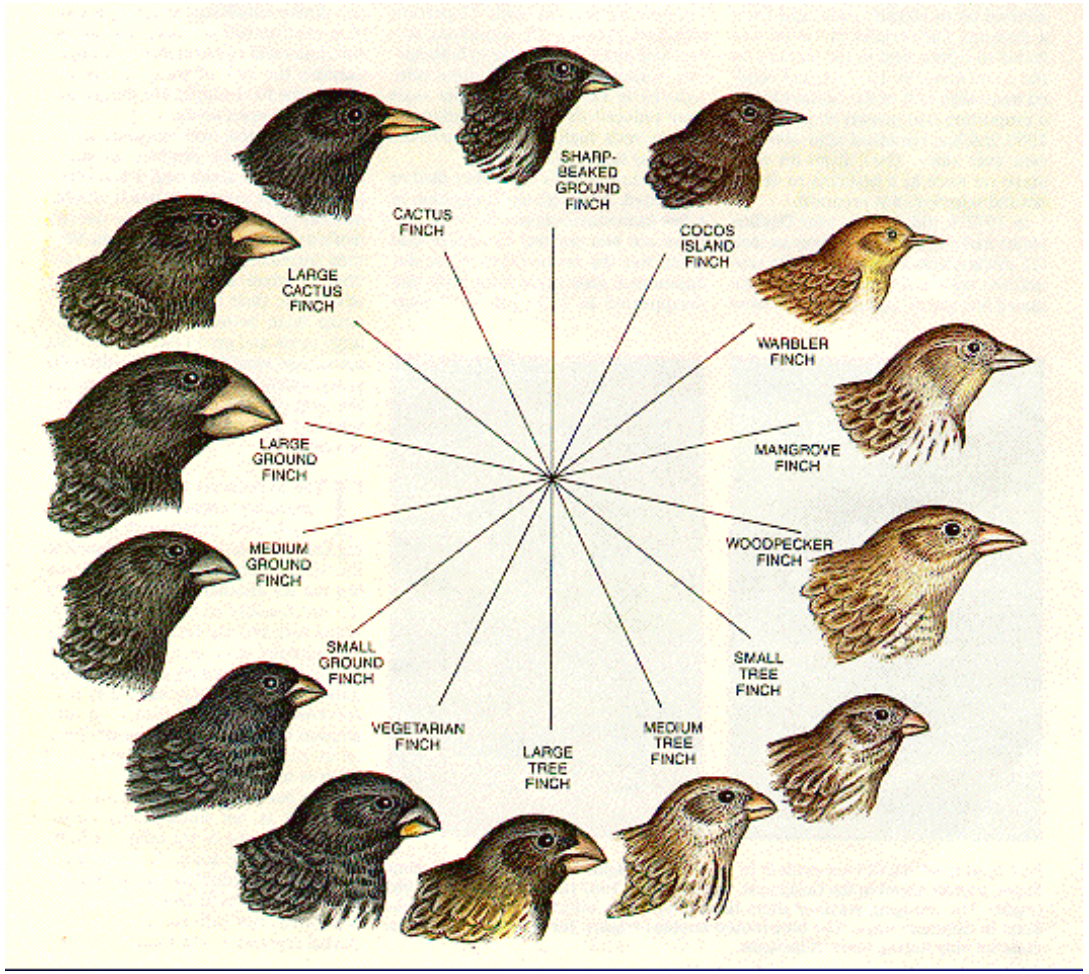
- Describe by: *age*, *income*, *location*, *behaviour pattern*…





https://www.mekkographics.com/wine-customer-segmentation/

Biology



Reprints from "Booker et al., Detecting positive selection in protein-coding genes. BMC Biology, 2017"

Community detection

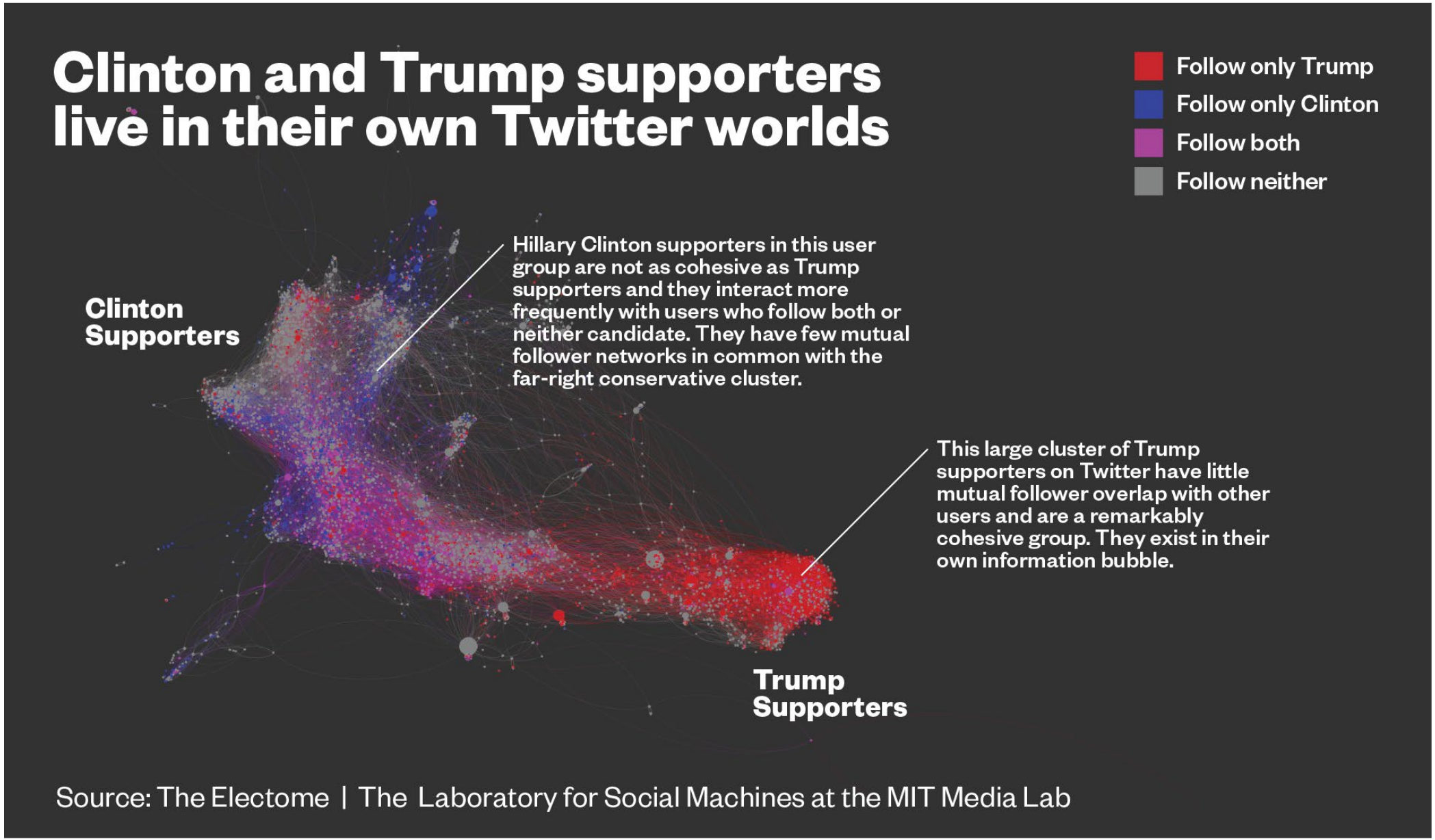


Image compression

original image
(396*396)

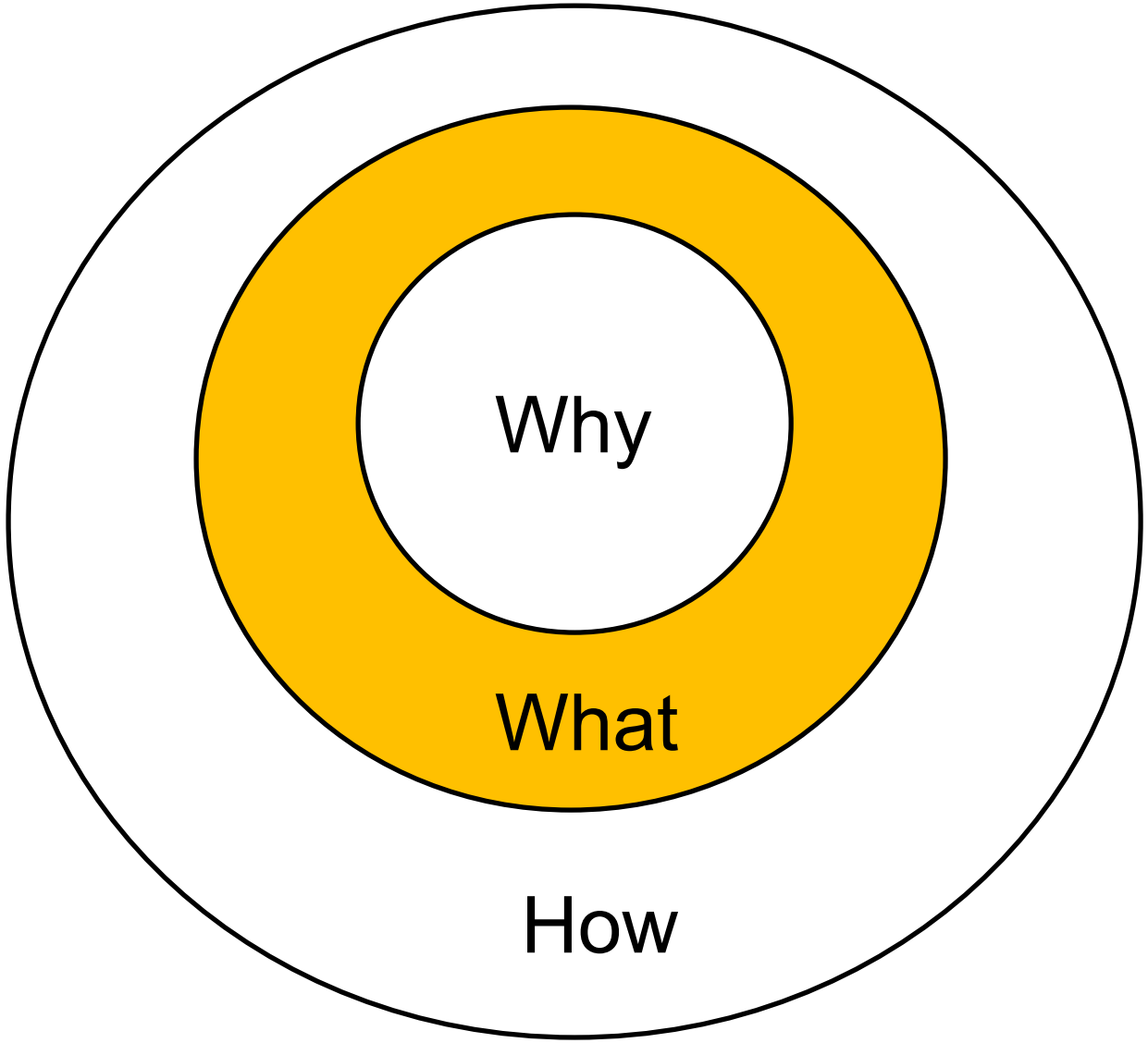
corrupted image
(30 colors)

Reprints from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Why clustering

- Understanding
  - Understand customers
  - Understand birds
  - Understand twitter users

- Summarization
  - “Summarize” the image to a compressed form


Why
What
How

Clustering

Given a set of *objects*, each object with descriptions of its properties, *partition* them into groups based on the descriptions, such that objects within one group are similar, and objects from different groups are dissimilar.

- What are the descriptions?
- What are the groups?
- What are the similar/dissimilar between **two** objects?
- What are the similar/dissimilar within **groups** of objects?

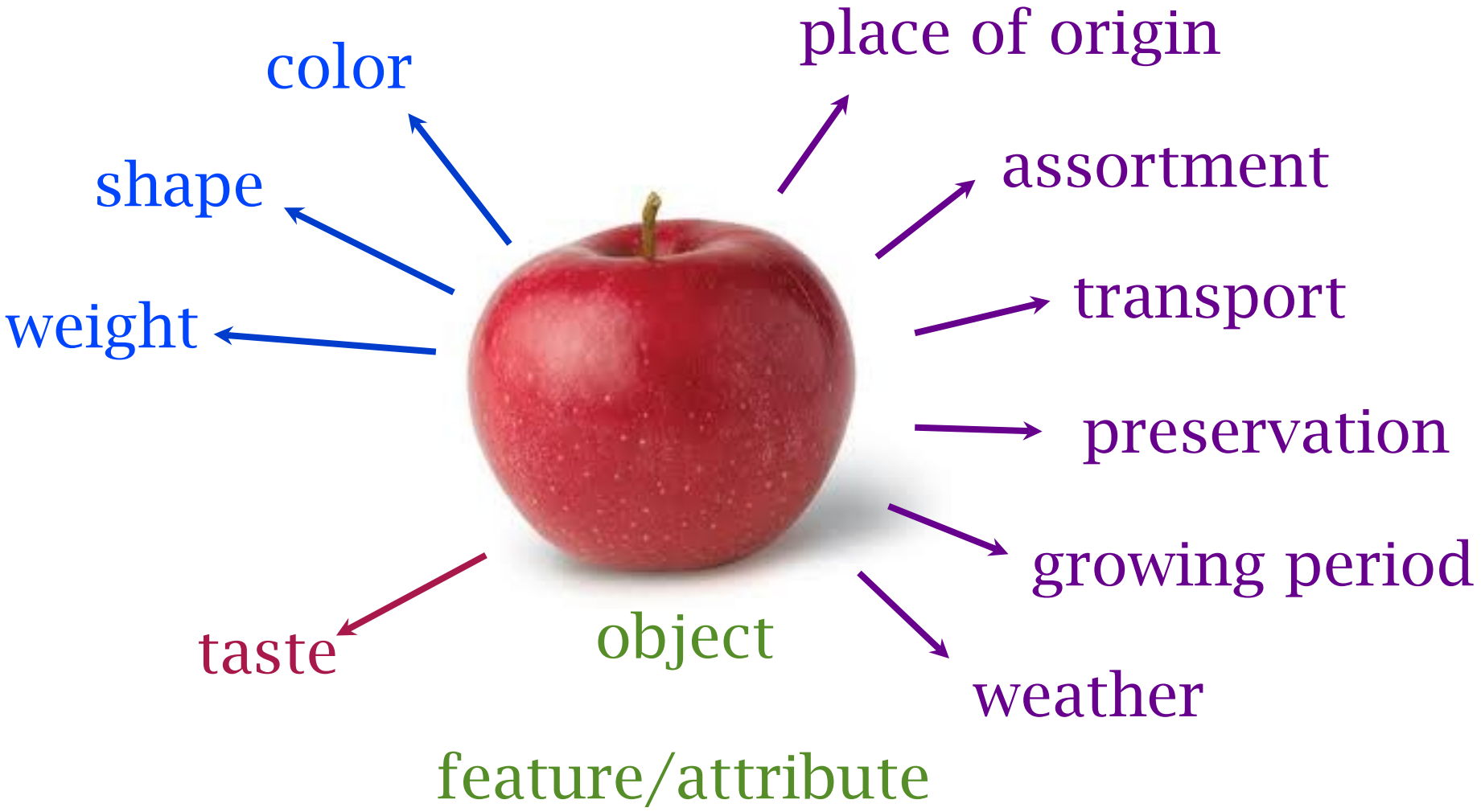
- What are the descriptions?
- What are the groups?
- What are the similar/dissimilar between **two** objects?
- What are the similar/dissimilar within **groups** of objects?

What are the descriptions

- Attribute/feature/input variable
  - Each associated with a given property of an object

Colour	Taste
Green	Sour

Feature engineering is important!



name	color	shape	weight	PoO	assortment	transport	preservation	growing	weather	taste
A1	red	round	200	Fuji	H	express	frozen	150	sunny	sweet

What are the descriptions

- Attribute/feature
  - Each associated with a given property of an object

Item	Eat/Drink	Price	Weights	State	Fresh	…
Bread	Y	6$	200g	Solid	Some	…
Milk	Y	5$	1200g	Liquid	Very	…
Diapers	N	10$	3000g	Solid	A bit	…
Beer	Y	3$	300g	Liquid	Some	…
Cola	Y	2$	300g	Liquid	A bit	…
Eggs	Y	4$	600g	Solid	Very	…
…	…	…	…	…	…	…

Are these attributes in the same form?

Forms of attribute

- Numerical:
  - The values of the attribute is to indicate the **quantity** of some predefined unit.
- Nominal
  - The values of the attribute are **symbols**, which is used to distinguish each other.
- Ordinal
  - The values of the attribute is to indicate certain **ordering relationship** resided in the attribute.

- Which attributes are numerical/nominal/ordinal?

Item	Eat/Drink	Price	Weights	State	Fresh	…
Bread	Y	6$	200g	Solid	Some	…
Milk	Y	5$	1200g	Liquid	Very	…
Diapers	N	10$	3000g	Solid	A bit	…
Beer	Y	3$	300g	Liquid	Some	…
Cola	Y	2$	300g	Liquid	A bit	…
Eggs	Y	4$	600g	Solid	Very	…
…	…	…	…	…	…	…

Nominal *Numerical* *Numerical* *Nominal* *Ordinal*

Operations on attributes

- Numerical:
  - Some algebraic operation: 6$+5$ = 11$
- Nominal
  - Transfer into numerical form: one-hot (solid: 1/0; liquid: 1/0)
  - No algebraic operation -- Comparing
- Ordinal
  - Transfer into numerical form: very: 5; some: 3; a bit: 1
  - No algebraic operation -- Sorting

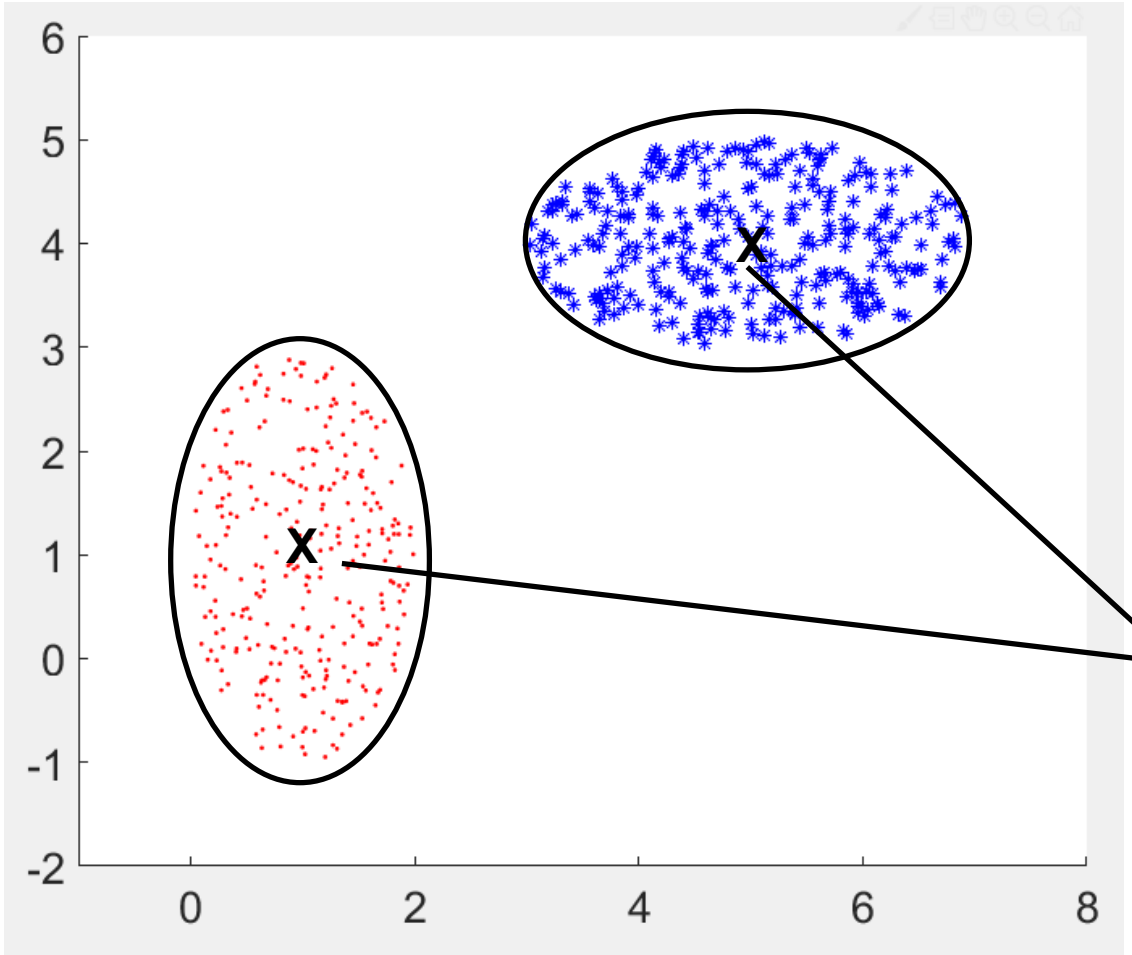
- Transfer into numerical

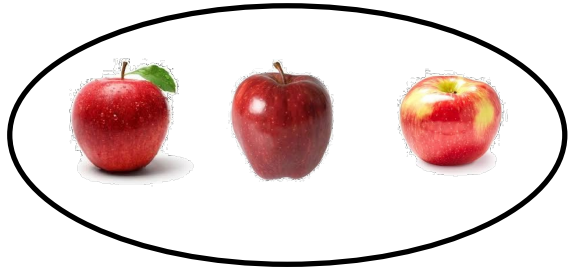
Item	Eat/Drink	No Eat/Drink	Price	Weights	Solid	Liquid	Fresh	…
Bread	1	0	6	200	1	0	3	…
Milk	1	0	5	1200	0	1	5	…
Diapers	0	1	10	3000	1	0	1	…
Beer	1	0	3	300	0	1	3	…
Cola	1	0	2	300	0	1	1	…
Eggs	1	0	4	600	1	0	5	…
…	…	…	…	…	…	…	…	…

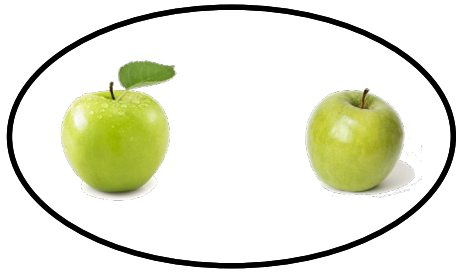
- What are the descriptions?
- What are the groups?
- What are the similar/dissimilar between **two** objects?
- What are the similar/dissimilar within **groups** of objects?

What are the groups

- A group is called a “cluster”



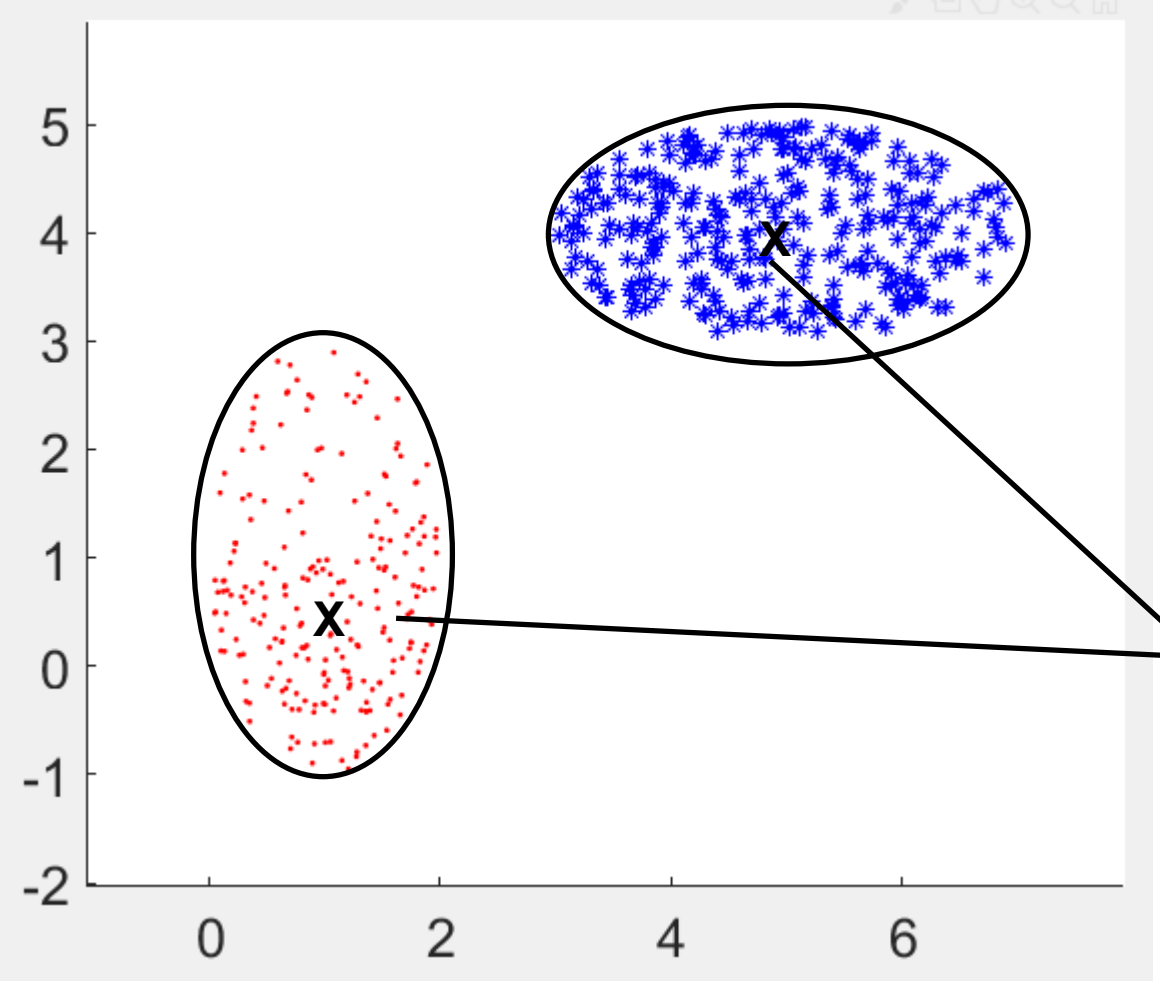




A centroid of a cluster: the mean of all points in the cluster.

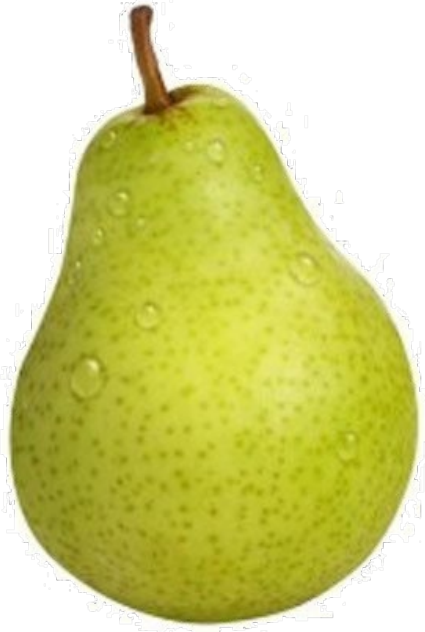
What are the groups

- A group is called a “cluster”

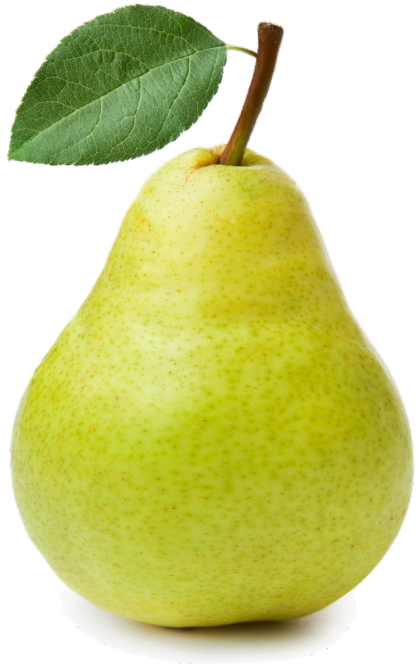


A centroid of a cluster: the mean of all points in the cluster.

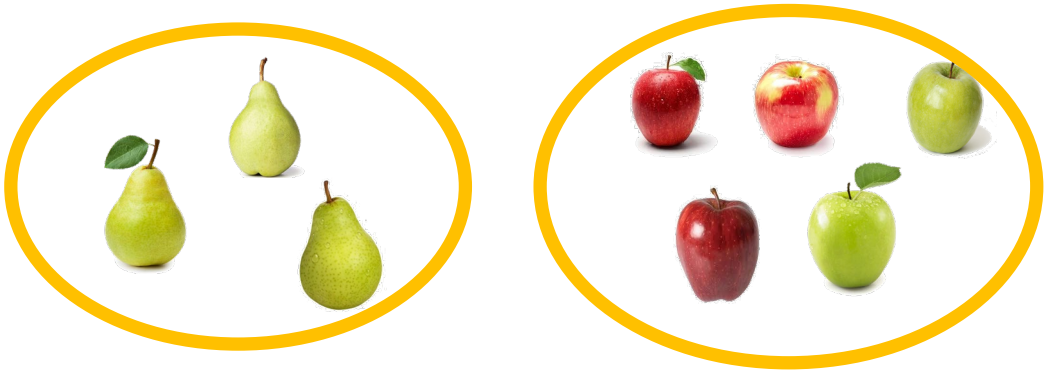
- What are the descriptions?
- What are the groups?
- What are the similar/dissimilar between **two** objects?
- What are the similar/dissimilar within **groups** of objects?



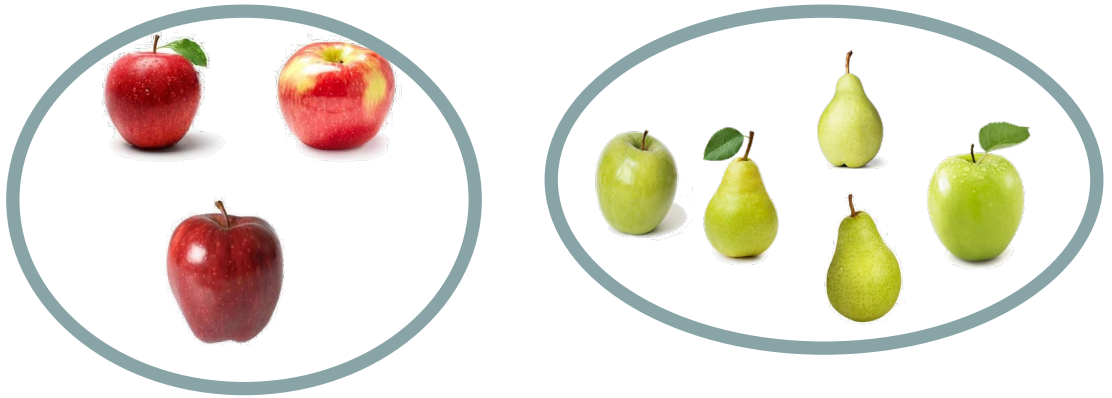


Which is correct clustering?



Similarity measured by *genre*

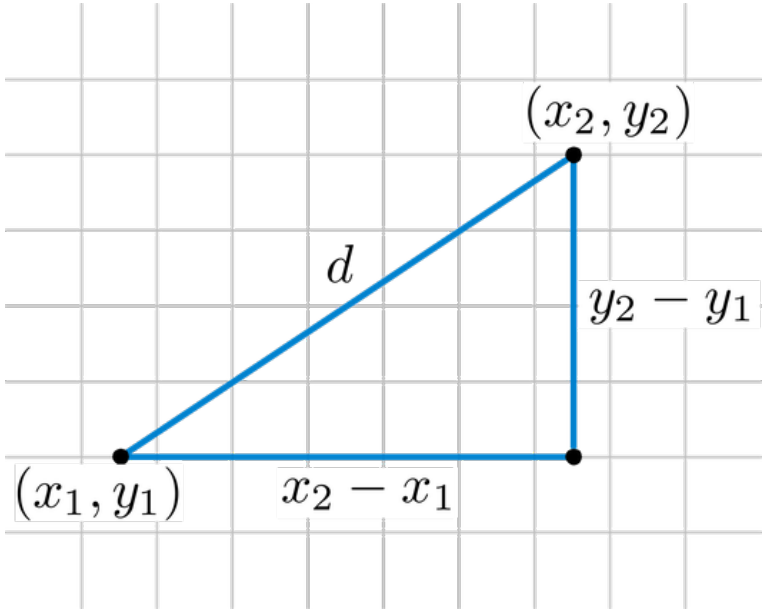


Similarity measured by *color*

How we measure the similarity is important and impacts the clustering results!

Distance

- "Similarity" is measured through distance
- The more similar two objects are, the less distance between them
- What kind of "distance" do you know?



$$d = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$

Pythagorean theorem

Are this the only way to measure distance?

Minkowski distance

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}|^2 \right)^{1/2}$$

Euclidean Distance

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}|^p \right)^{1/p}$$

Minkowski Distance

Minkowski distance: different p

- p = 1: Manhattan distance (L_1 distance)

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}| \right)$$

- p = 2: Euclidean distance (L_2 distance)

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}|^2 \right)^{1/2}$$

- $p = \infty$: Chebyshev distance (L_∞ distance)

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max_{u=\{1,2,\ldots,d\}} |x_{iu} - x_{ju}|$$

Manhattan distance: the name

Metric

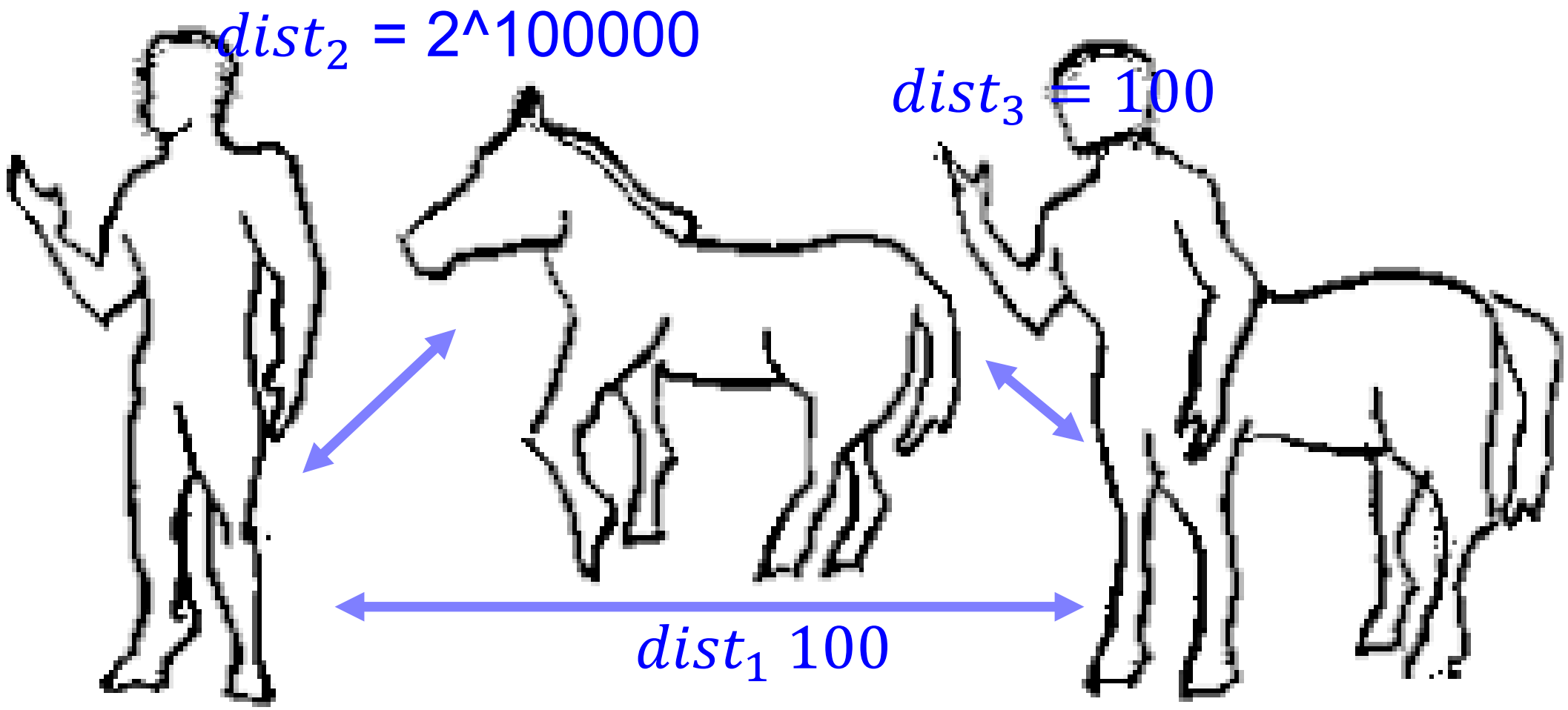
- Non-negativity: $d(i,j) \geq 0, d(i,i) = 0$

- Symmetry: $d(i,j) = d(j,i)$

- Triangle inequality (Transitivity):

$$d(i,j) \leq d(i,k) + d(k,j)$$

Non-metric distance



~~$Transitivity$~~

Activity: calculate the distance

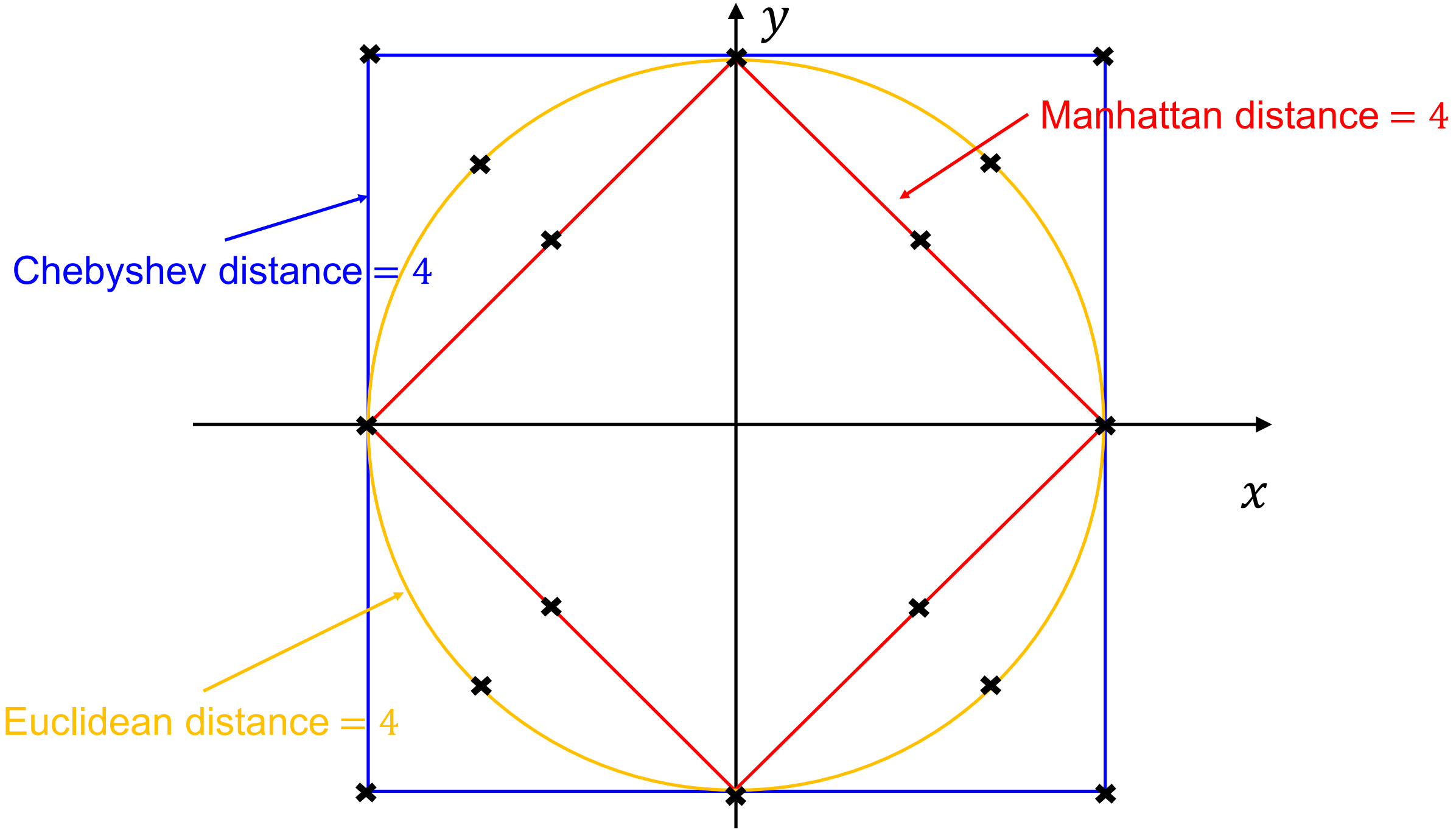
- Calculate the distance to (0, 0)

A	B	C	D	E	F	G
	manhattan distance	Euclidean distance	Chebyshev distance	manhattan distance	Euclidean distance	Chebyshev distance
(0, 4)	Group 1	Group 1	Group 1	Group 17	Group 17	Group 17
(2, 2)	Group 2	Group 2	Group 2	Group 18	Group 18	Group 18
(2√2, 2√2)	Group 3	Group 3	Group 3	Group 19	Group 19	Group 19
(4, 4)	Group 4	Group 4	Group 4	Group 20	Group 20	Group 20
(4, 0)	Group 5	Group 5	Group 5	Group 21	Group 21	Group 21
(2, -2)	Group 6	Group 6	Group 6	Group 22	Group 22	Group 22
(2√2, -2√2)	Group 7	Group 7	Group 7	Group 23	Group 23	Group 23
(4, -4)	Group 8	Group 8	Group 8	Group 24	Group 24	Group 24
(0, -4)	Group 9	Group 9	Group 9	Group 25	Group 25	Group 25
(-2, -2)	Group 10	Group 10	Group 10	Group 26	Group 26	Group 26
(2√2, 2√2)	Group 11	Group 11	Group 11	Group 27	Group 27	Group 27
(-4, -4)	Group 12	Group 12	Group 12	Group 28	Group 28	Group 28
(-4, 0)	Group 13	Group 13	Group 13	Group 29	Group 29	Group 29
(-2, 2)	Group 14	Group 14	Group 14	Group 30	Group 30	Group 30
(-2√2, 2√2)	Group 15	Group 15	Group 15	Group 31	Group 31	Group 31
(-4, 4)	Group 16	Group 16	Group 16	Group 32	Group 32	Group 32

https://docs.google.com/spreadsheets/d/1a5FuwA4_hcQq85L8XMZIdw5fvOi6vlZ7yqwYc38ZJTE/edit?usp=sharing

Answer

	Manhattan distance	Euclidean distance	Chebyshev distance
(0, 4)	4	4	4
(2, 2)	4	$2\sqrt{2}$	2
($2\sqrt{2}$, $2\sqrt{2}$)	$4\sqrt{2}$	4	$2\sqrt{2}$
(4, 4)	8	$4\sqrt{2}$	4
(4, 0)	4	4	4
(2, -2)	4	$2\sqrt{2}$	2
($2\sqrt{2}$, $-2\sqrt{2}$)	$4\sqrt{2}$	4	$2\sqrt{2}$
(4, -4)	8	$4\sqrt{2}$	4
(0, -4)	4	4	4
(-2, -2)	4	$2\sqrt{2}$	2
($2\sqrt{2}$, $2\sqrt{2}$)	$4\sqrt{2}$	4	$2\sqrt{2}$
(-4, -4)	8	$4\sqrt{2}$	4
(-4, 0)	4	4	4
(-2, 2)	4	$2\sqrt{2}$	2
($-2\sqrt{2}$, $2\sqrt{2}$)	$4\sqrt{2}$	4	$2\sqrt{2}$
(-4, 4)	8	$4\sqrt{2}$	4


y
x
Manhattan distance = 4
Chebyshev distance = 4
Euclidean distance = 4

What's the effect of these distances in clustering

- For two instances, if most of the attributes are close, but a few are quite different (have a large gap).

- Which distance metric is more likely to measure the two as "similar"? *Manhattan distance*

- Which distance metric is more likely to measure the two as "dissimilar"? *Chebyshev distance*

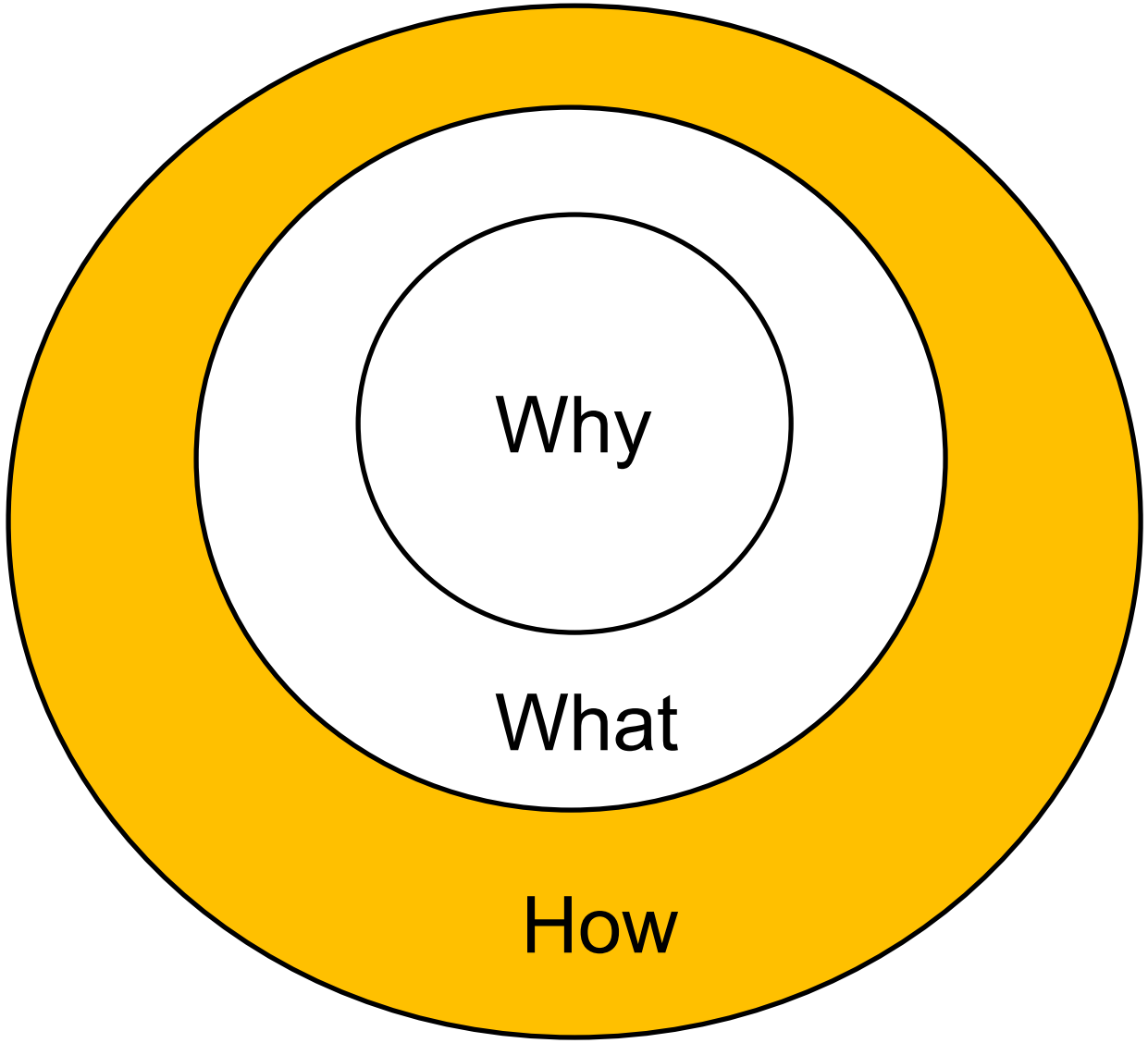
Robustness: Manhattan distance> Chebyshev distance

- What are the descriptions?
- What are the groups?
- What are the similar/dissimilar between **two** objects?
- What are the similar/dissimilar within **groups** of objects?

Clustering validity index

- Sum of Squared Error (SSE):
  - If the data are clustered into k clusters, each represented as C_i
  - Each cluster's centroid is c_i

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


Why
What
How

Objective of clustering

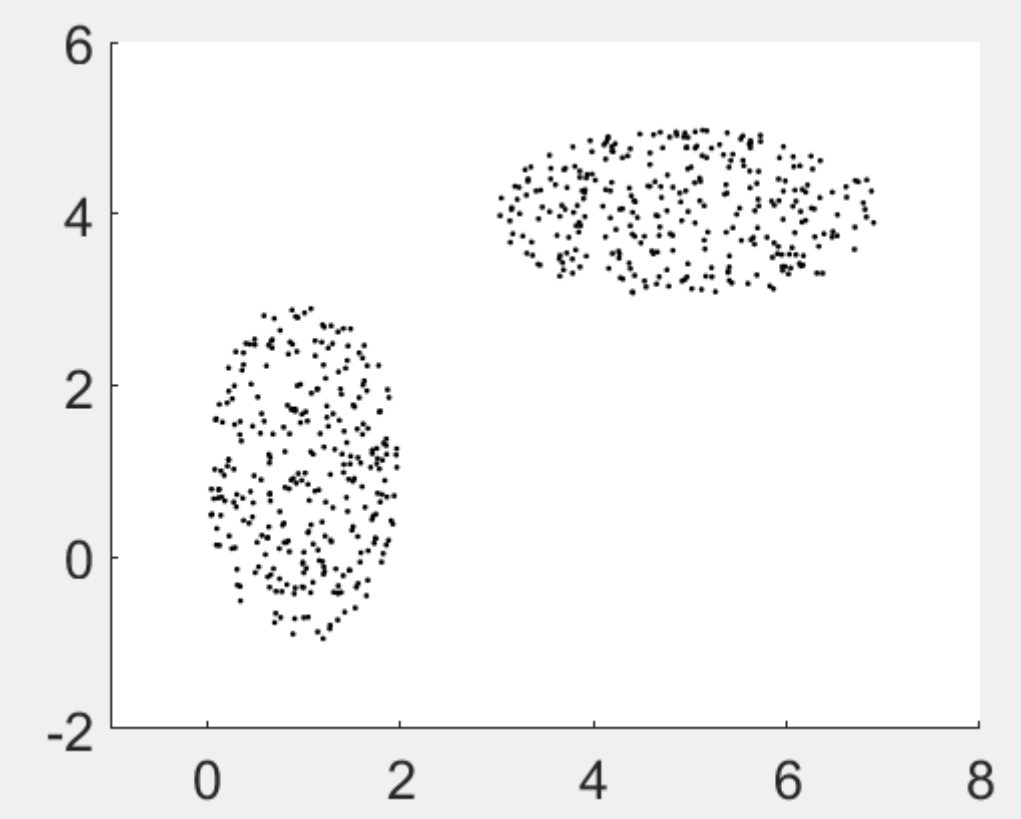
- Objective

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

- Fix k, minimize SSE in a greedy manner
  - Random pick up centroids
  - Optimize cluster assignments
  - Optimize centroids

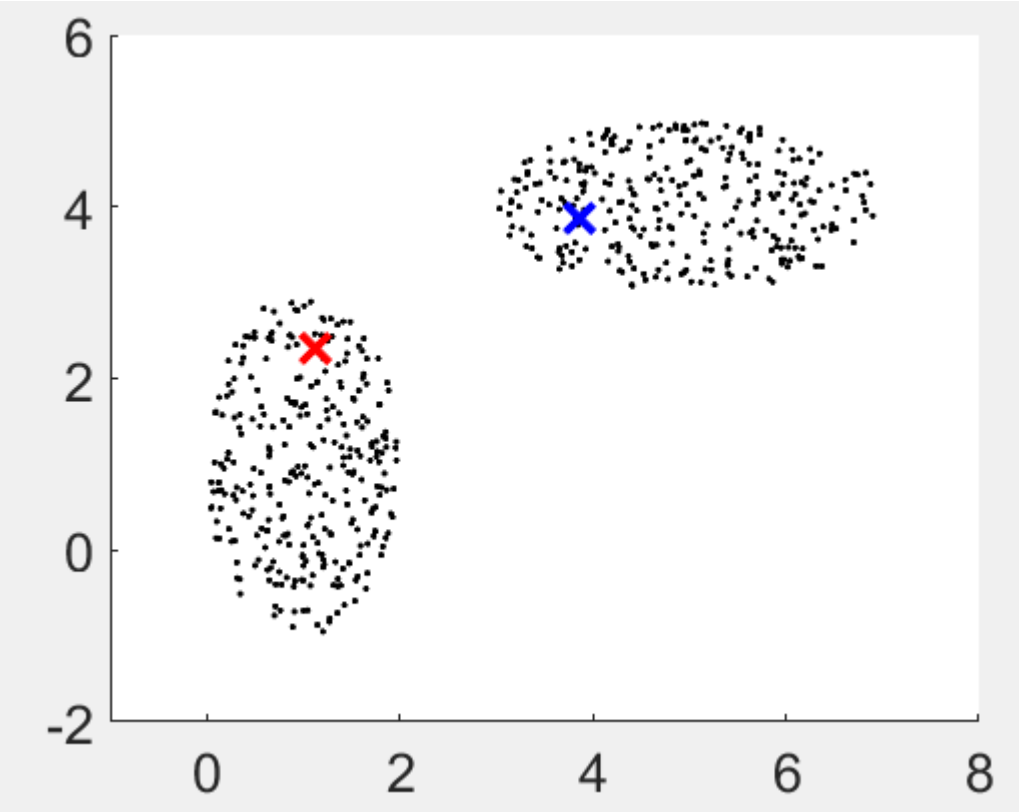
Case 1: optimize cluster assignments

- Fix k to 2, we *randomly* select two points as initial centroids



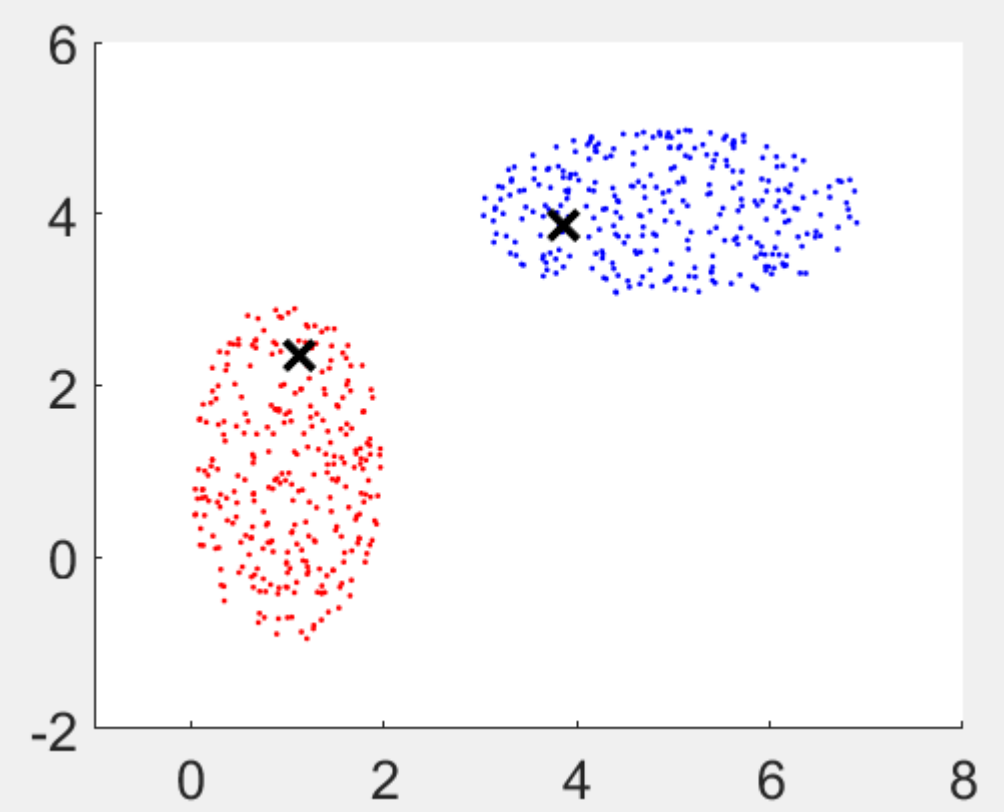
Case 1: optimize cluster assignments

- Fix k to 2, we *randomly* select two points as initial centroids



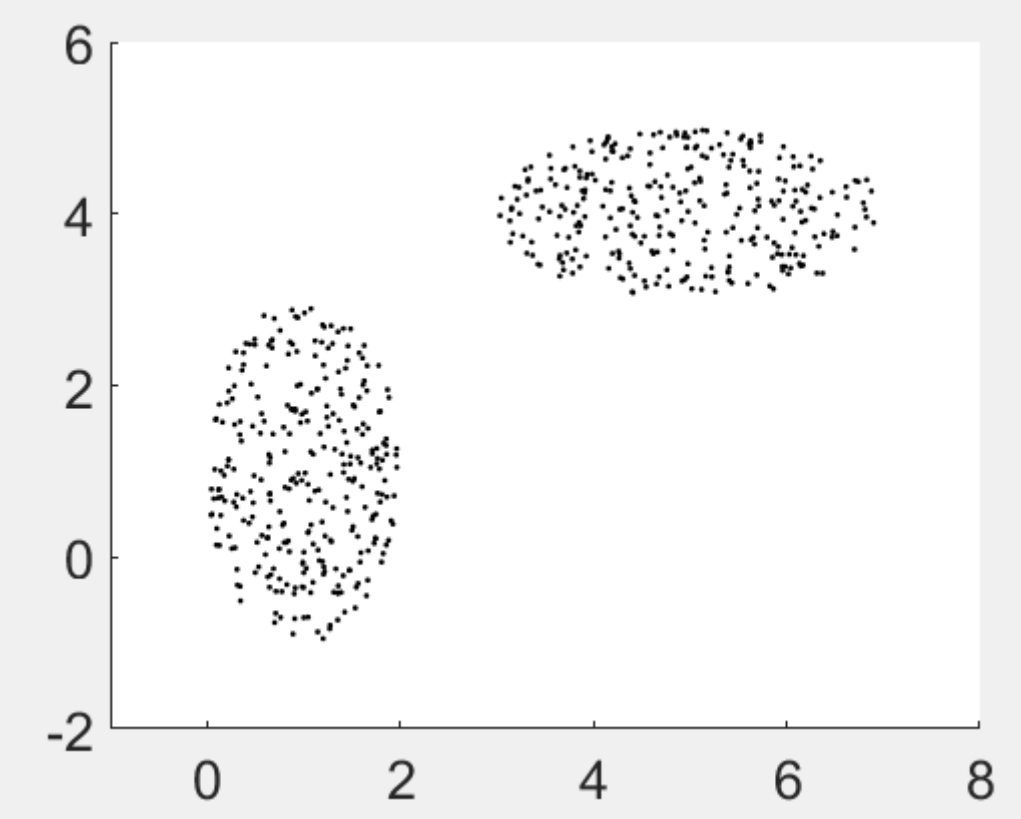
Case 1: optimize cluster assignments

- Calculate all points' distances to the two centroids
- Assign point to a closer centroid as a cluster



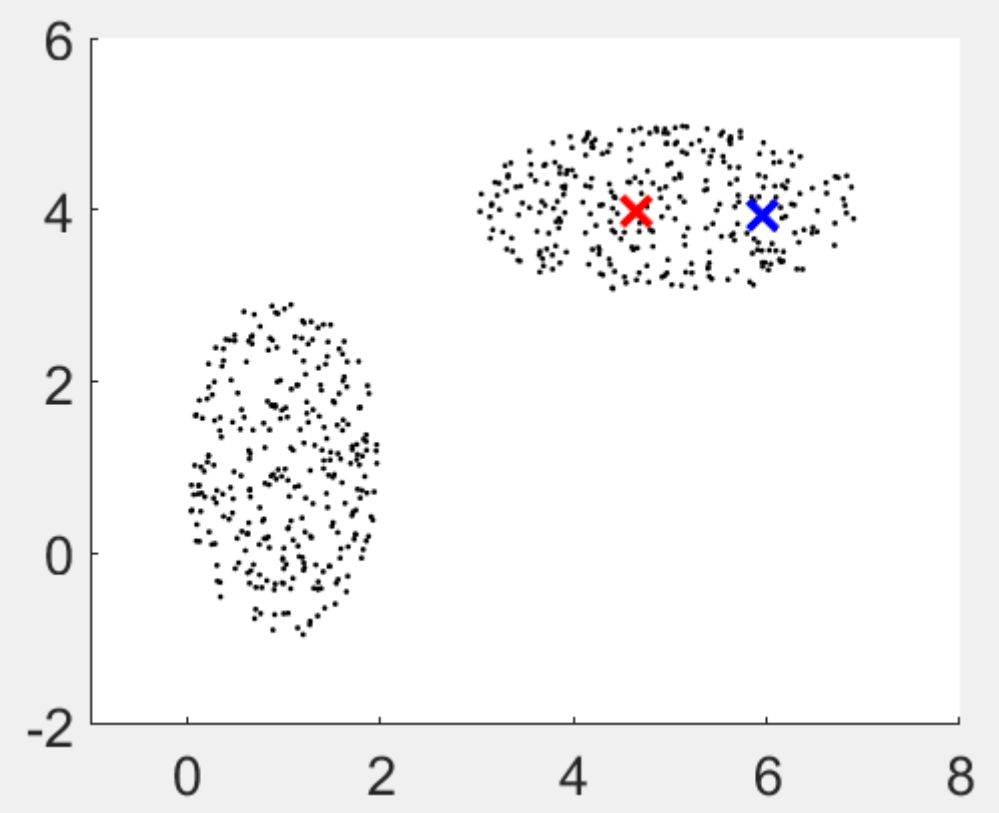
How about two initial centroids are in the same "cluster"?

Case 2: two initial centroids close



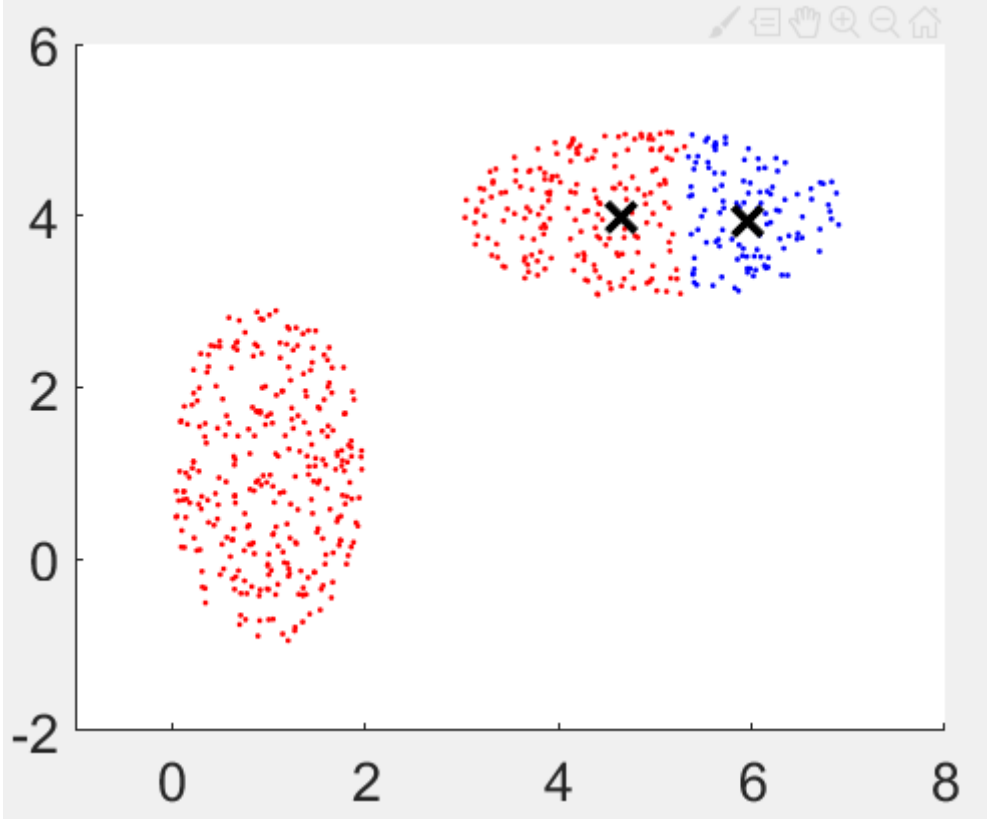
Case 2: two initial centroids close

- Random select two centroids



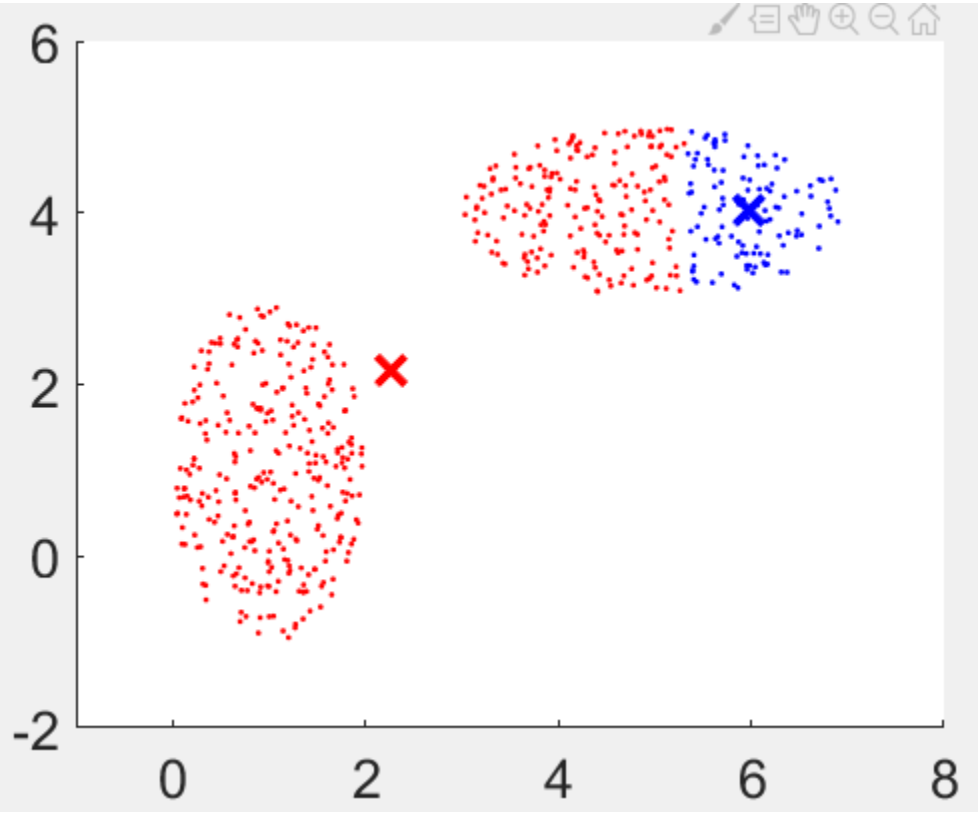
Case 2: two initial centroids close

- Calculate the distances and assign points



Case 2: two initial centroids close

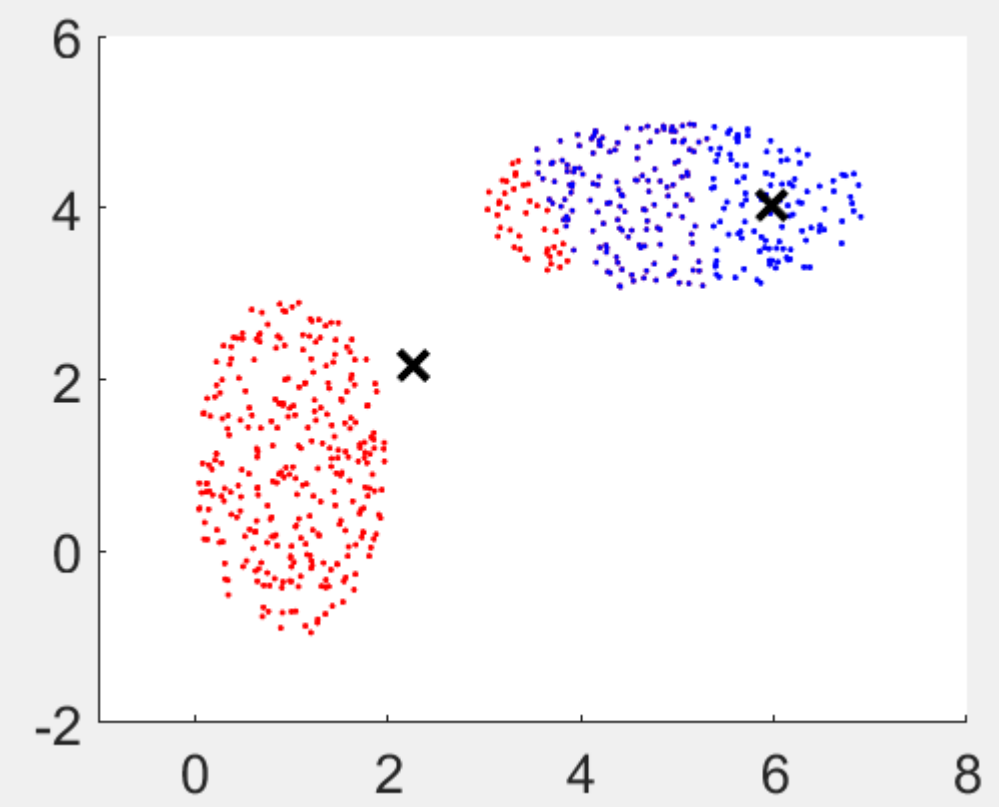
- Update the centroid by the “mean” of each cluster



Case 2: two initial centroids close

Round 2

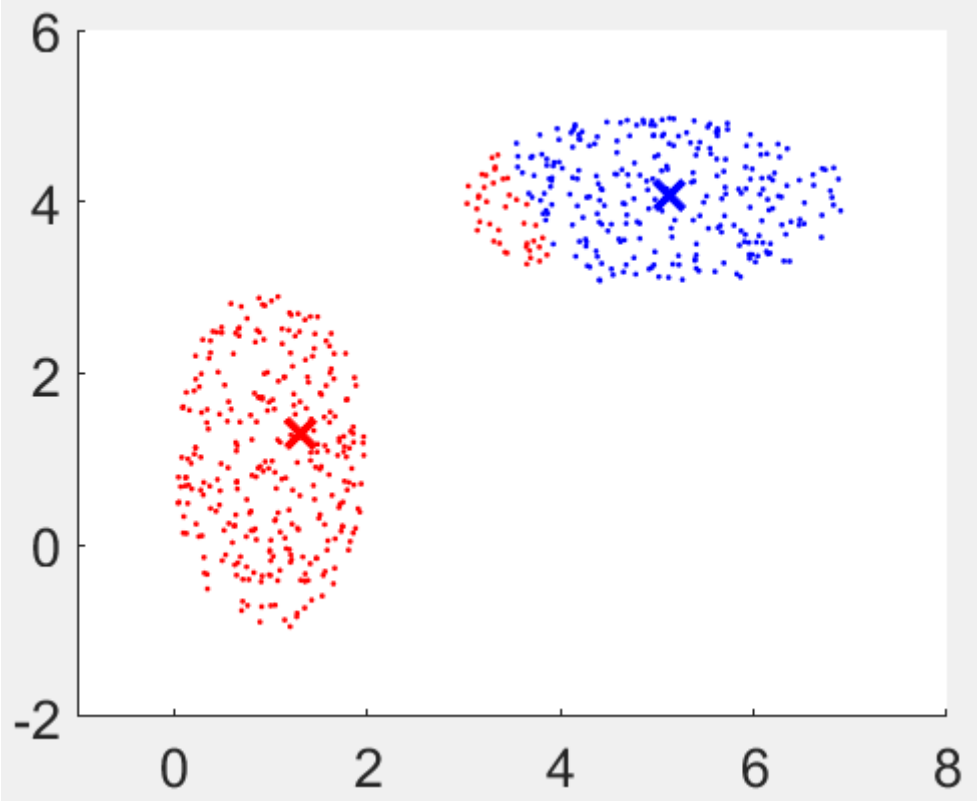
- Calculate the distance again and re-assign the cluster



Case 2: two initial centroids close

Round 2

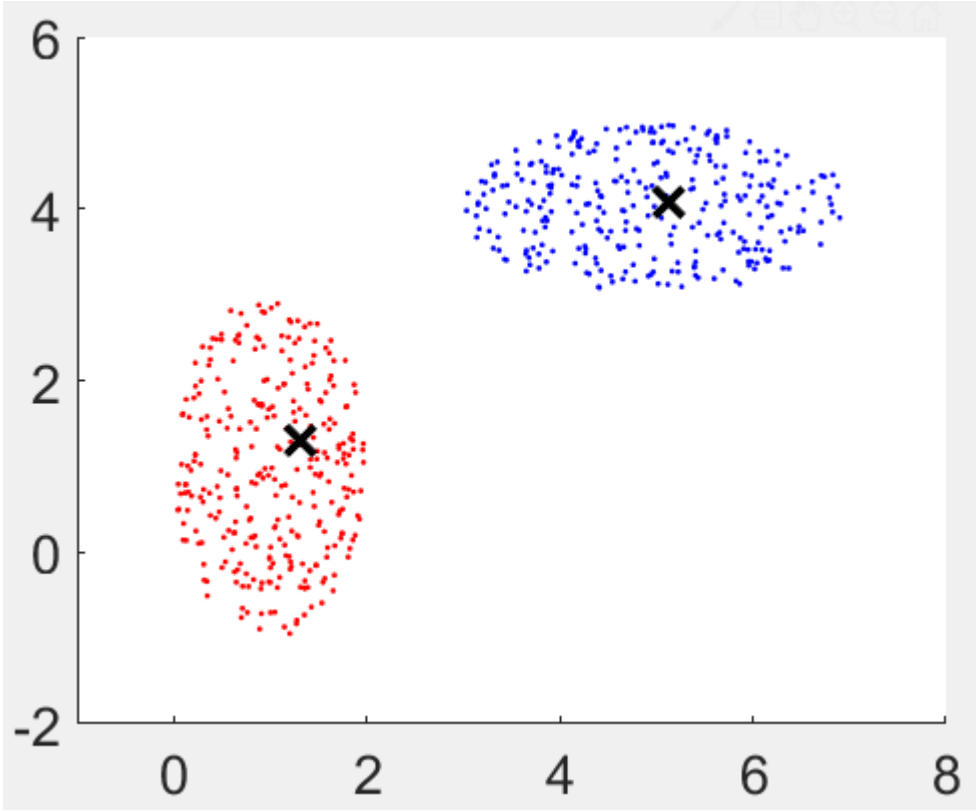
- Update the centroid



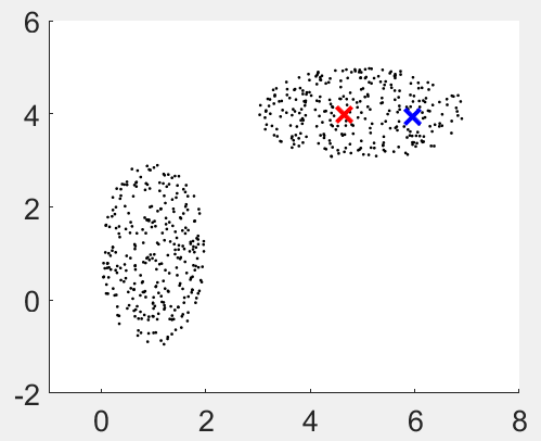
Case 2: two initial centroids close

Round 3

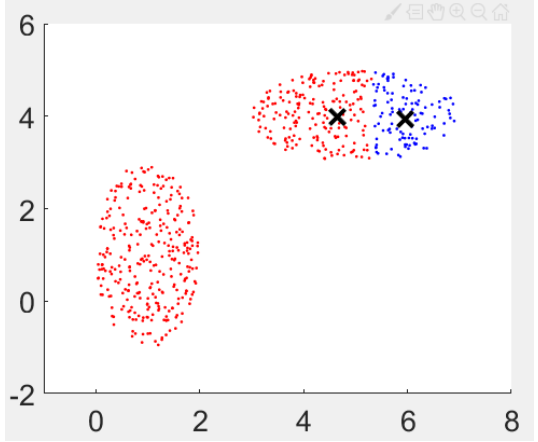
- Calculate the distance and re-assign the cluster



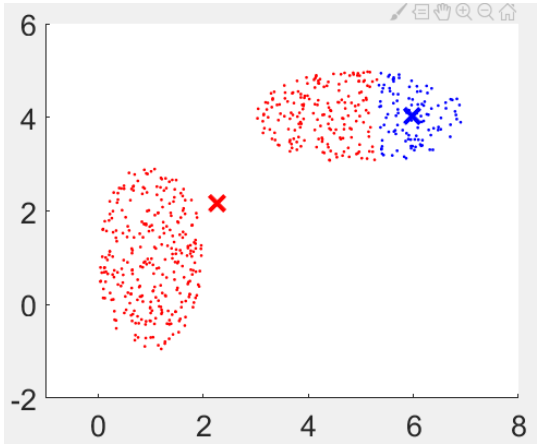
Summarize the three rounds



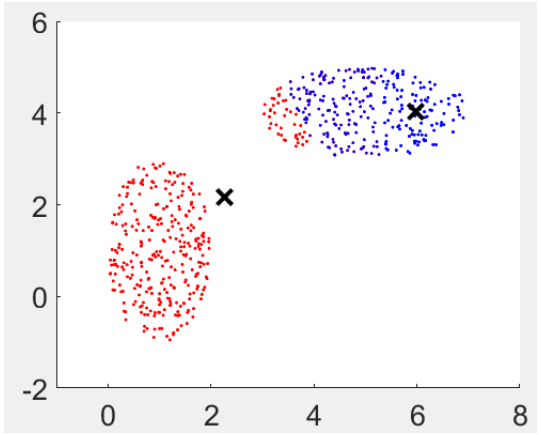
Select initial centroids



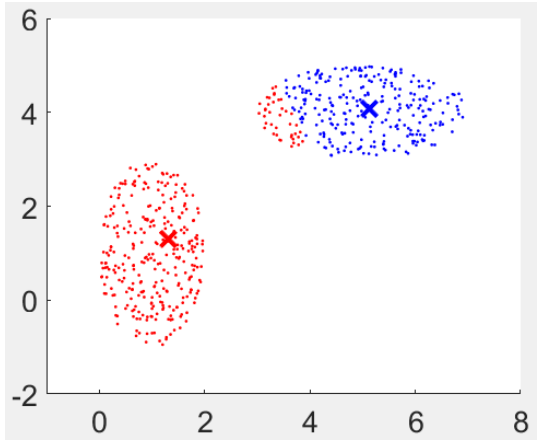
Assign points to closest centroid



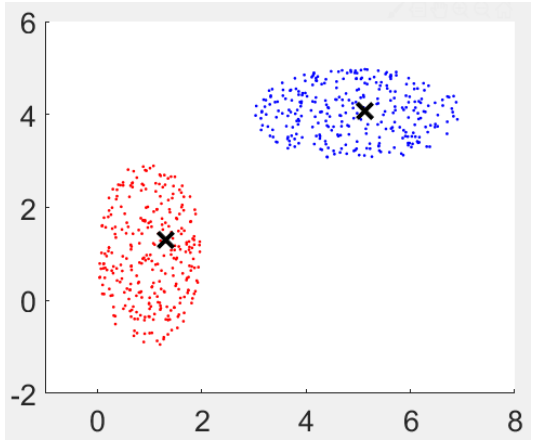
Update centroids



Assign points to closest centroid



Update centroids



Assign points to closest centroid

The *k*-means algorithm

1. --Select *k* points as the initial centroids
2. Repeat
3. --Form *k* clusters by assigning points to the closest centroid
4. --Update the centroid of each cluster
5. Until --The cluster assignment don't change

One representative of
prototype-based clustering

Activity

- The k-means procedure for one-dimensional data

We have the following one dimensional data

18, 22, 25, 42, 27, 43, 33, 35, 56, 28

If k = 3 and the initial centroids are 22, 35, 43. Using the Euclidean distance, what will be the result round-by-round of k-means?

Answer

Round 1

Centroid	Cluster assignment	Updated centroid
22	18, 22, 25, 27, 28	24
35	33, 35	34
43	42, 43, 56	47

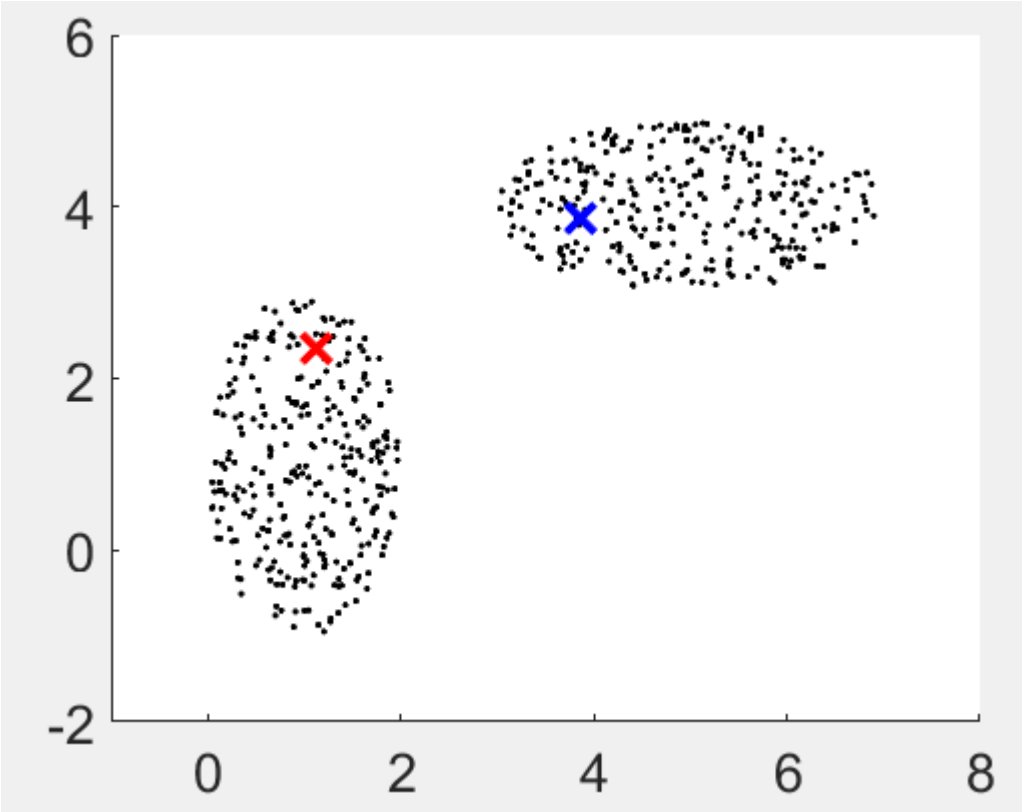
Round 2

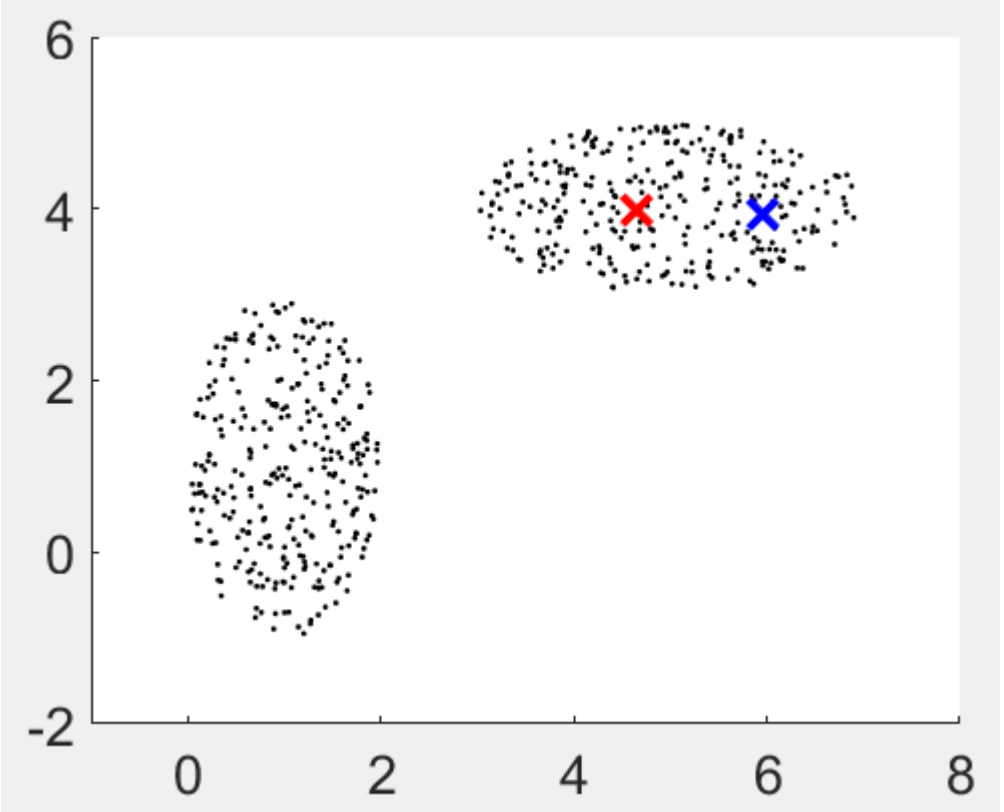
Centroid	Cluster assignment	Updated centroid
24	18, 22, 25, 27, 28	24
34	33, 35	34
47	42, 43, 56	47

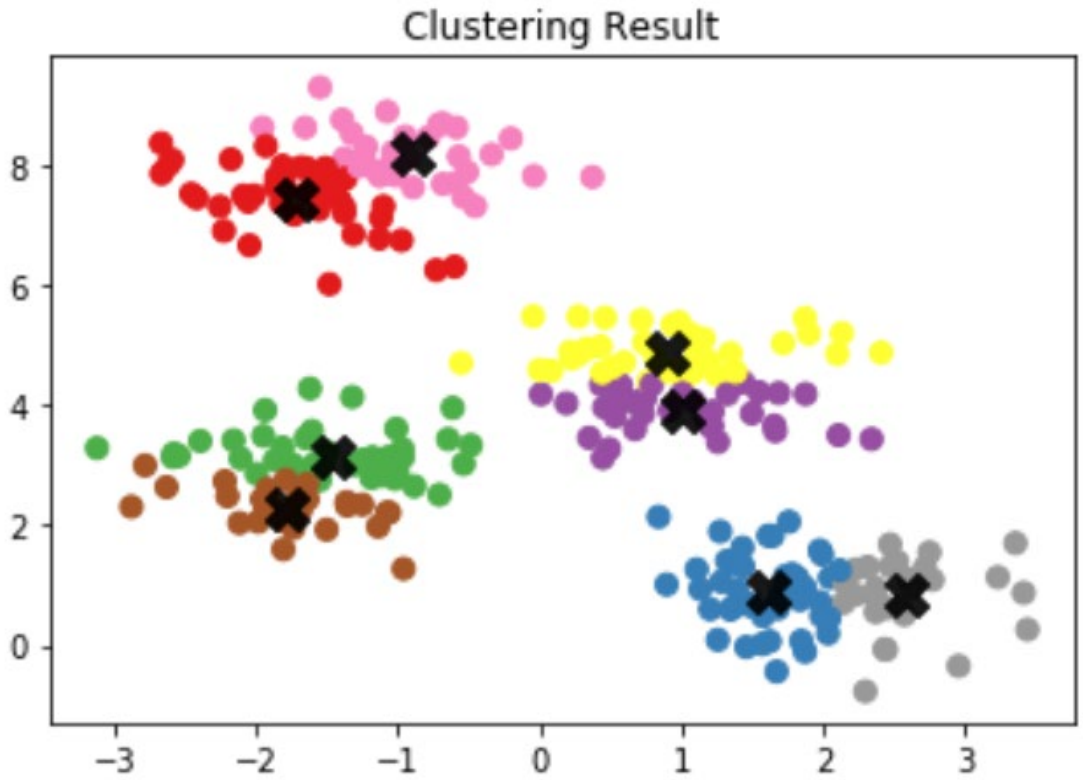
More consideration on *k*-means

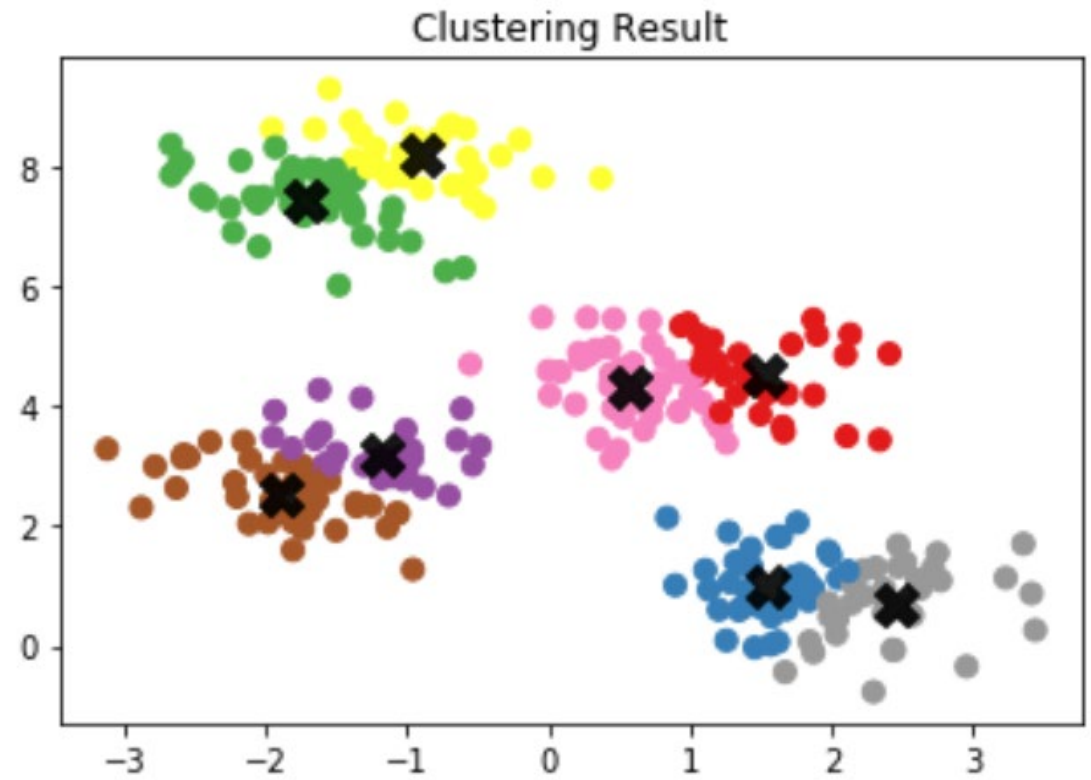
- The impact of initialization
- How to set *k*
- Limitation
- Can it converge

The impact of initialization





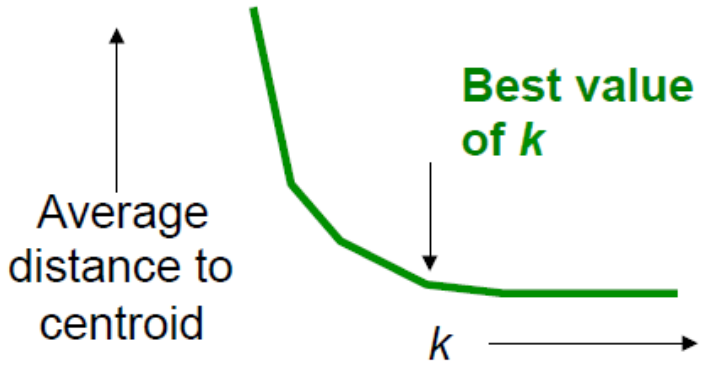

Clustering Result
8
6
4
2
0
−3
−2
−1
0
1
2
3


Clustering Result
8
6
4
2
0
−3
−2
−1
0
1
2
3

How to set k

- Try different k and see the average distance to centroid

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i} dist(x, c_i)^2$$

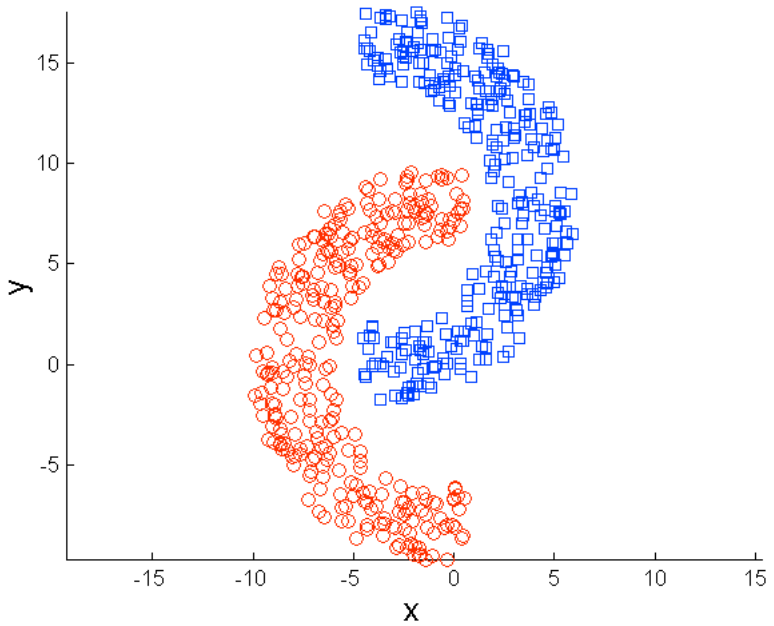


- Try k from 2 to $\sqrt{n}$ (n is the number of all data points)

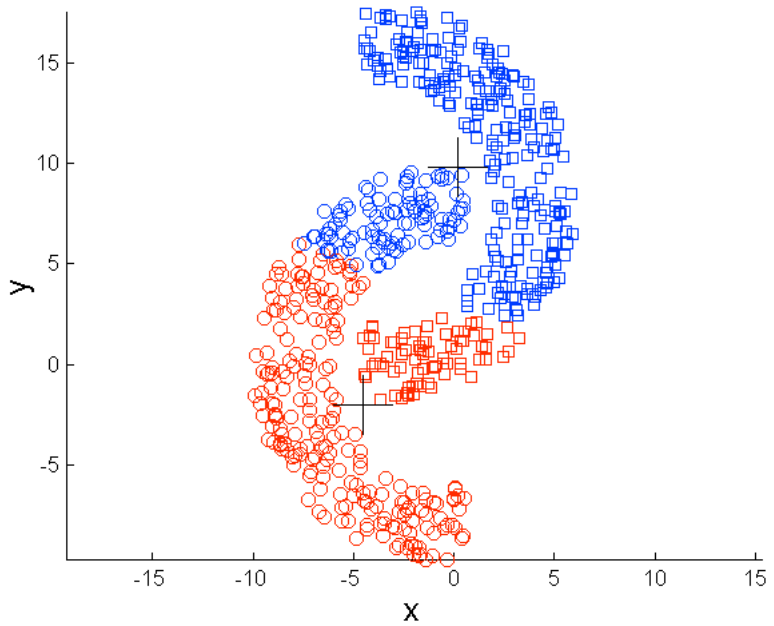
Limitation of *k*-means

- NFL theorem:
  - **No free lunch** theorem
  - If you work well on some data, there must exist some data you cannot work well (you need to pay for what you achieve).

- What kind of data *k*-means cannot do well?



Original Points



k-means (2 Clusters)

Can *k*-means converge?

- Yes! It can be proved. (please try!)

- How about if it cannot converge within reasonable amount of time?
  - Early stopping

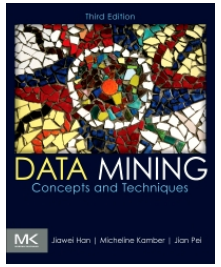
Summary

Given a set of *objects*, each object with descriptions of its properties, *partition* them into groups based on the descriptions, such that objects within one group are similar, and objects from different groups are dissimilar.

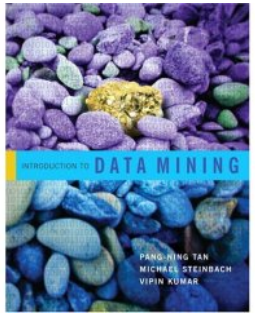
- Why?
  - Understanding, summarization
- What?
  - Numerical, nominal, ordinal; cluster, centroid; distance; validity index
- How?
  - *k*-means

Recommended reading (not required)

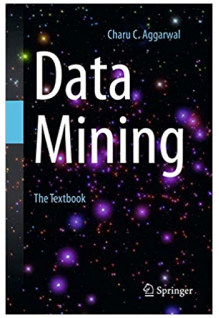
- [Han et al., 2012]
  - Sec. 10.1, 10.2, 10.6



- [Tan et al., 2005]
  - Sec. 8.1, 8.2, 8.5



- [Aggarwal, 2015]
  - Sec. 3.1-3.2
  - Sec. 6.1, 6.3, 6.9



Next week

- More clustering algorithms