Clustering: how to and Applications

John Monroe
10/26/18

Clustering Objective

Complex Data

Simple conclusion

Biology

Rich bacterial variability

Taxonomic classification

Machine Learning

Haphazardly collected data

Meaningful patterns

General Science

Many degrees of freedom

Essential underlying parameters

Basic Ingredients

1. Data

Spatial Information

	A	B	C	D	E
1	Specs of the ROM	CyanogenMod	PAC-man	Slimbean	Carbon
4	Android version	4.2	4.2	4.2	4.2
5	Link to the video review	Youtube Review	Youtube Review	Youtube Review	Youtube Revi
6	**Navigation**				
7	**Navigation bar/rings**				
8	Number of buttons	See Notes	See Notes	See Notes	See Notes
9	Navigation rings	See Notes	See Notes	See Notes	See Notes
10	Color options	✗	✓	✓	✓
11	auto-hide navbar	✗	✗	✗	✓
12	**PIE**				
13	Paranoid Android PIE	✗	✓	✗	✗
14	other PIE	See Notes	✗	✓	✗
15	**PIE Controls**				
16	Trigger Area	✓	✓	✓	✗
17	Gravity	✓	✓	✓	✗
18	Middle Alignment	✗	✓	✓	✗
19	PIE Style	✗	✓	✓	✗
20	PIE size	✓	✓	✓	✗
21	PIE color	✗	See Notes	✓	✗
22	Bottom Angle	✗	✓	✓	✗
23	Item gap	✗	✓	✓	✗
24	**Targets**				
25	Show menu button always	✗	✓	✓	✗
26	Show search button	✗	✗	✓	✗
27	Last app button	✗	✗	✓	✗

Shopping habits

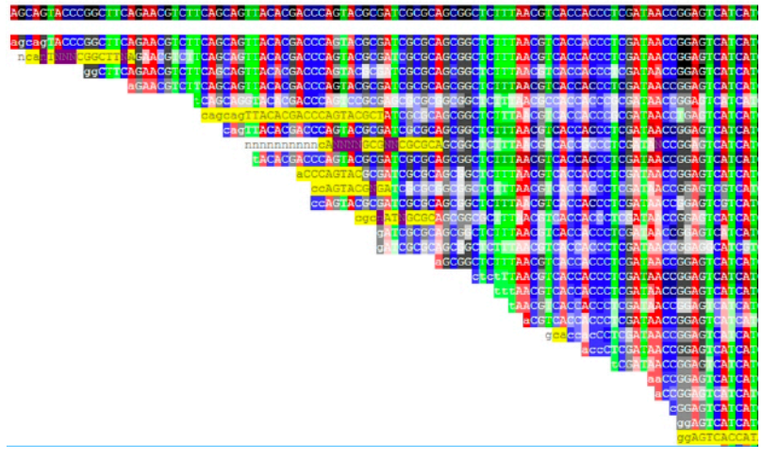
Predicted mutation

evidence	seq id	position	mutation	annotation	gene	
RA	NC_007779	3,449,011	A→C	I774S (ATC→AGC)	*rpoC* ←	RNA polymerase, beta prime subunit

Read alignment evidence...

	seq id	position	change	freq	score	reads	annotation	genes	
*	NC_007779	3,449,011	0	A→C	100.0%	278.6	93	2321	*rpoC*	RNA polymerase, beta prime subunit

Reads supporting (aligned to +/- strand): **new** base (47/46): **ref** base (0/0): **total** (47/46)



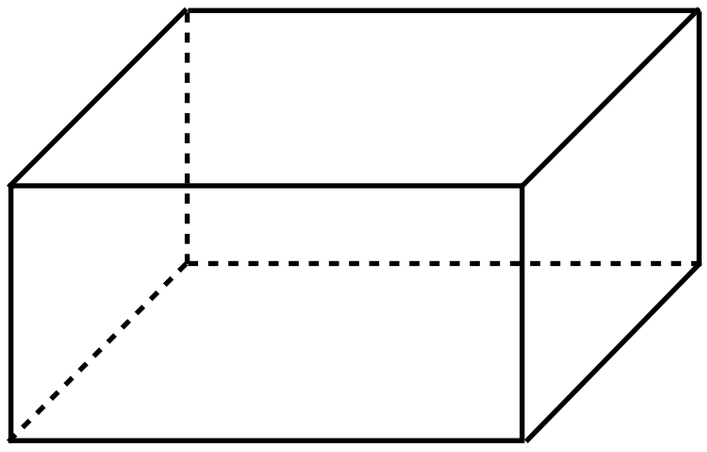
Sequences

- Similarity sets are not always immediately obvious
- High dimensional data is hard to visualize

The Verge 2018/2/9/16994638

Basic Ingredients

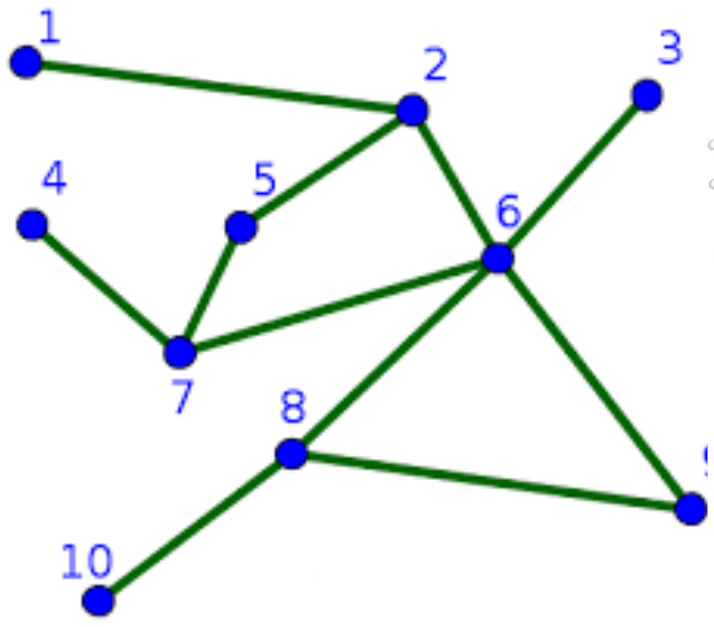
2. Distance Measure



Spatial Information

+

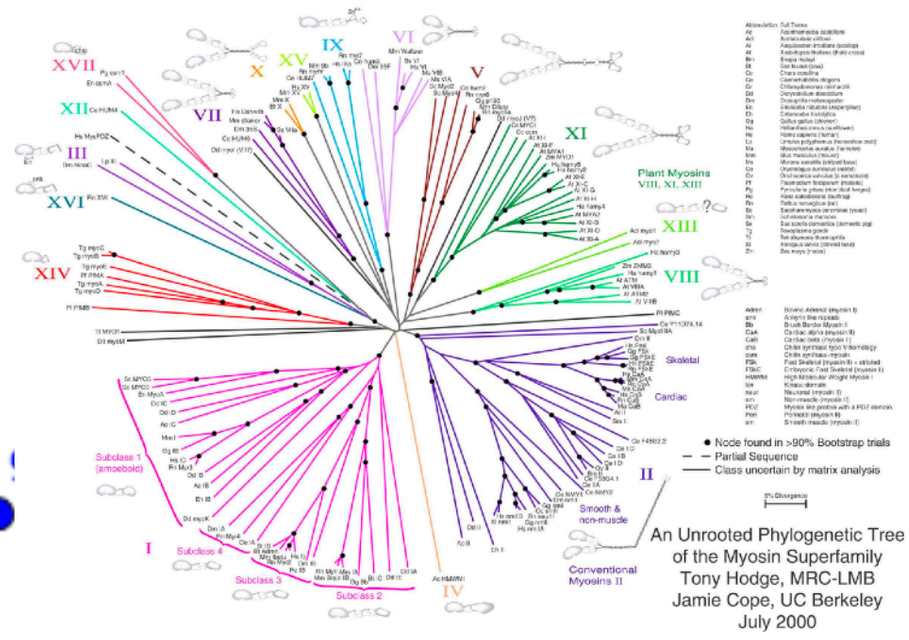
Euclidean Distance



Shopping habits

+

Similarity graph



Sequences

+

Genome counts

Distance measures are generally applicable to many datasets.

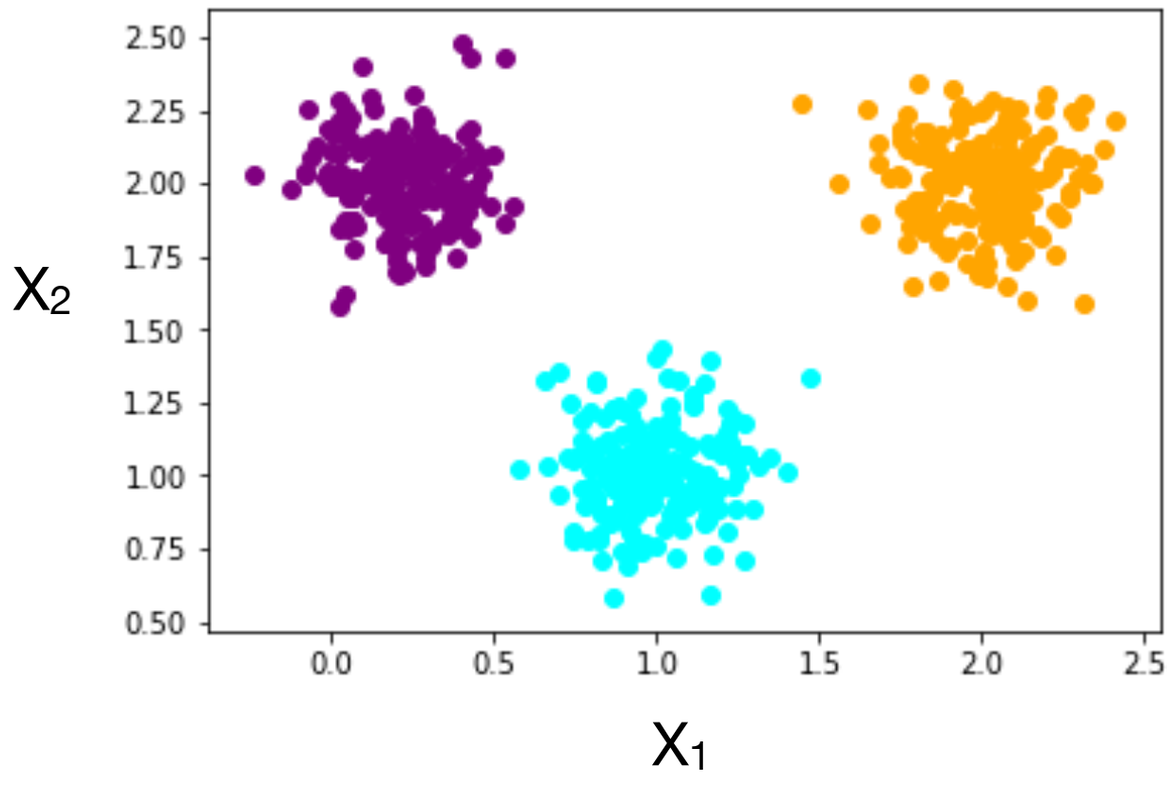
ncbi.nlm.nih.gov/Structure/cdd

The k-means Algorithm

The algorithm:

```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total_distance += distance
```

The data:

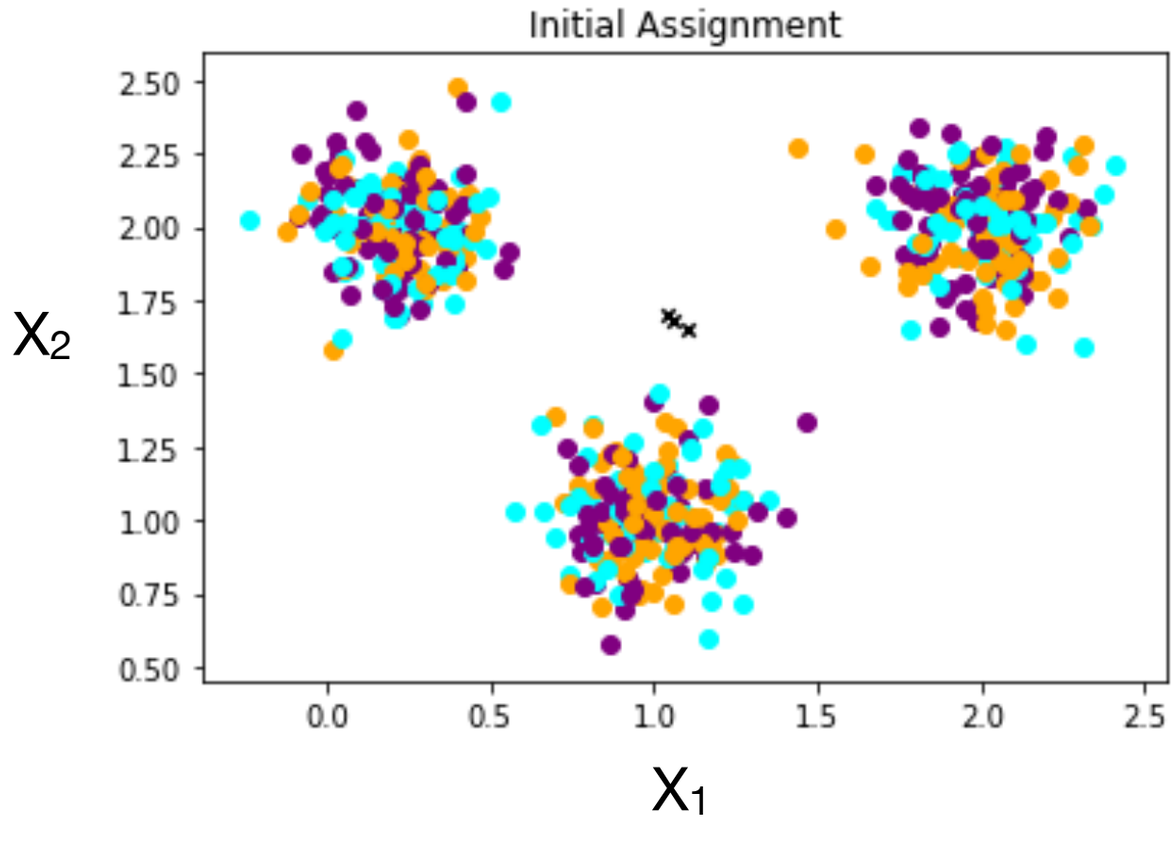


The k-means Algorithm

The algorithm:

```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total_distance += distance
```

The data:

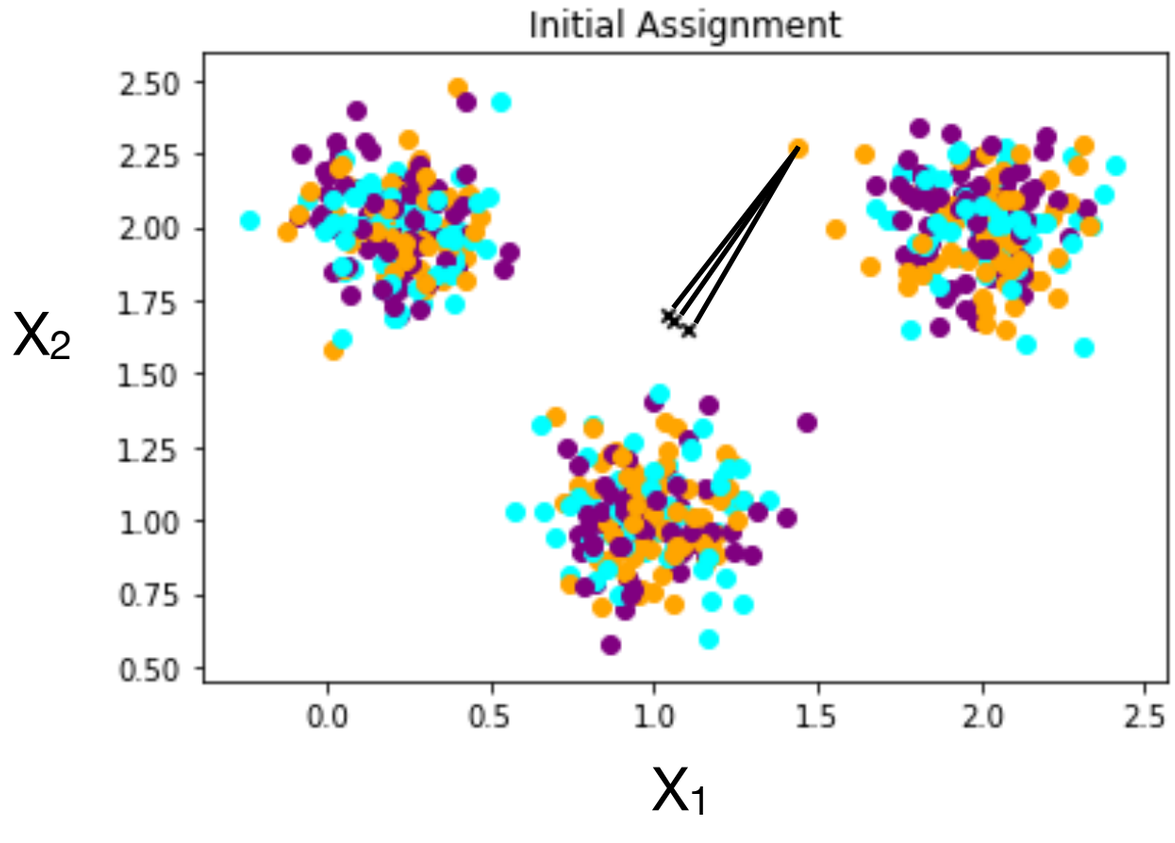


The k-means Algorithm

The algorithm:

```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total_distance += distance
```

The data:

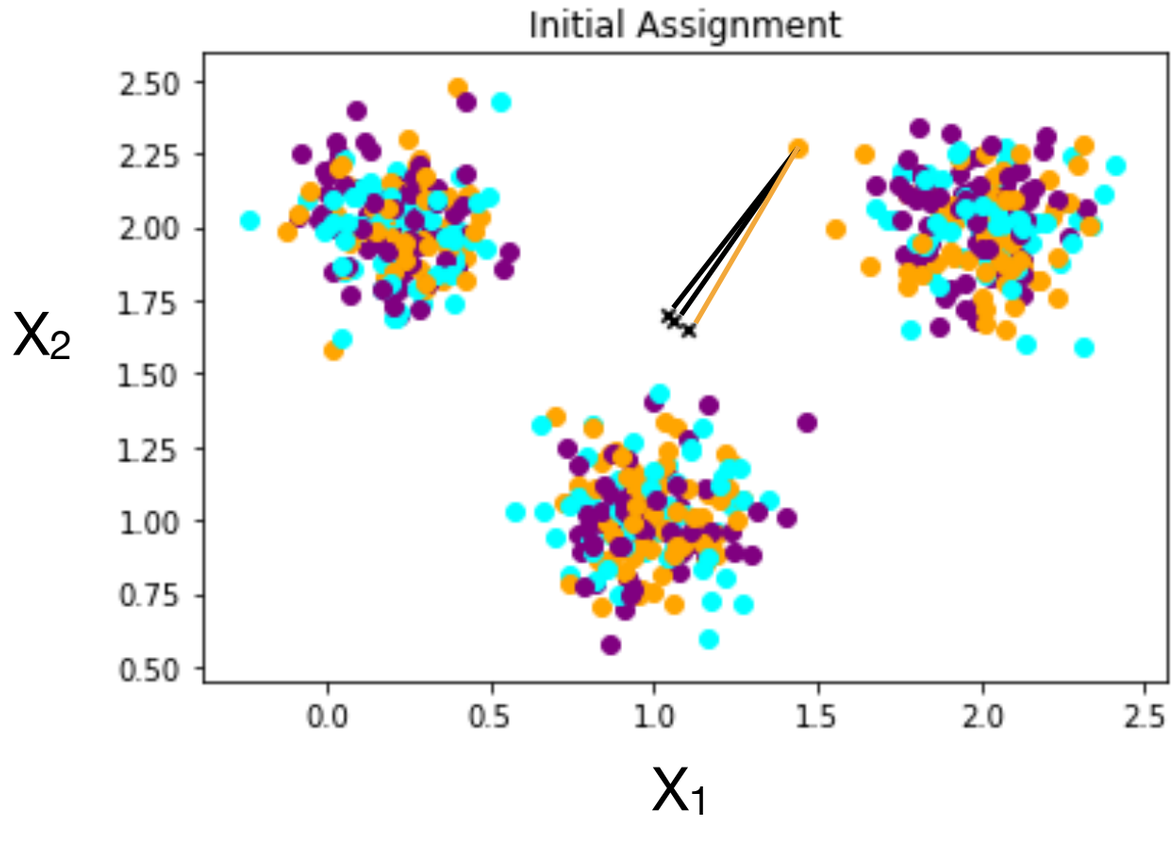


The k-means Algorithm

The algorithm:

```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total_distance += distance
```

The data:

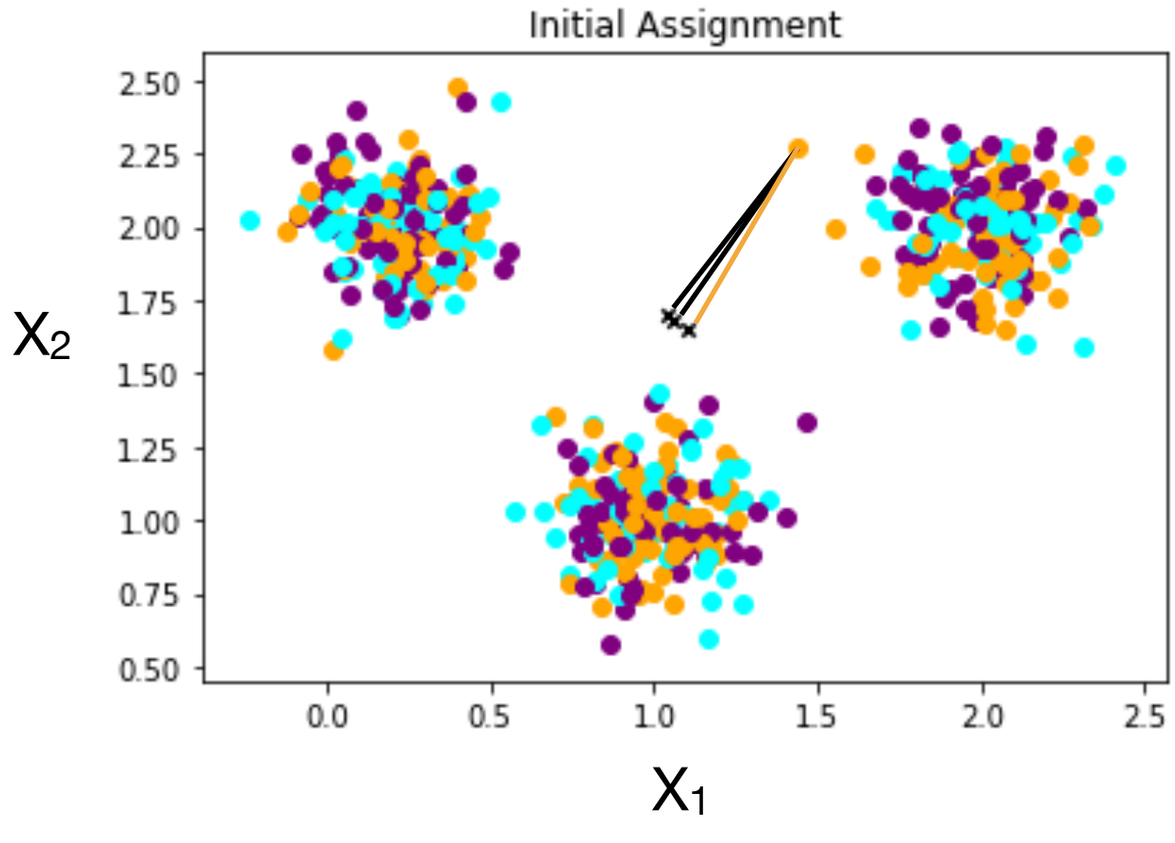


The k-means Algorithm

The algorithm:

```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total distance += distance
```

The data:

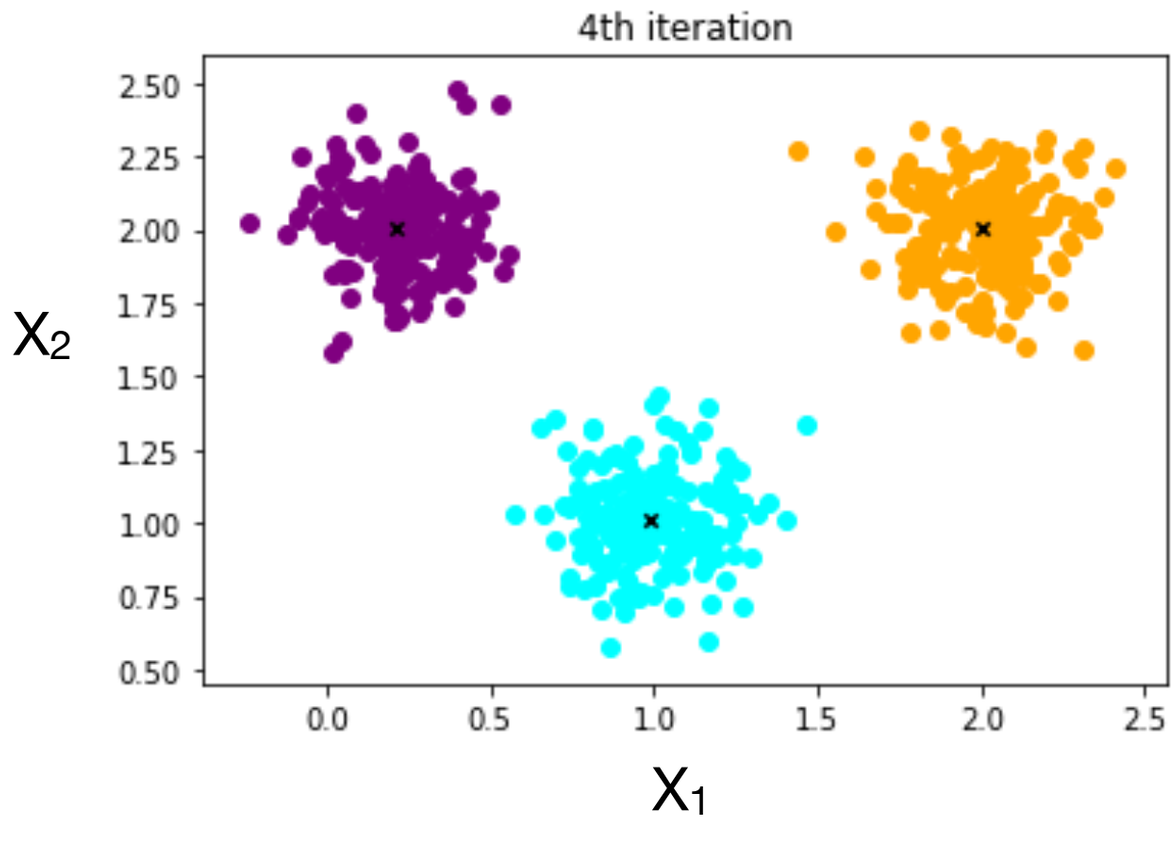


The k-means Algorithm

The algorithm:

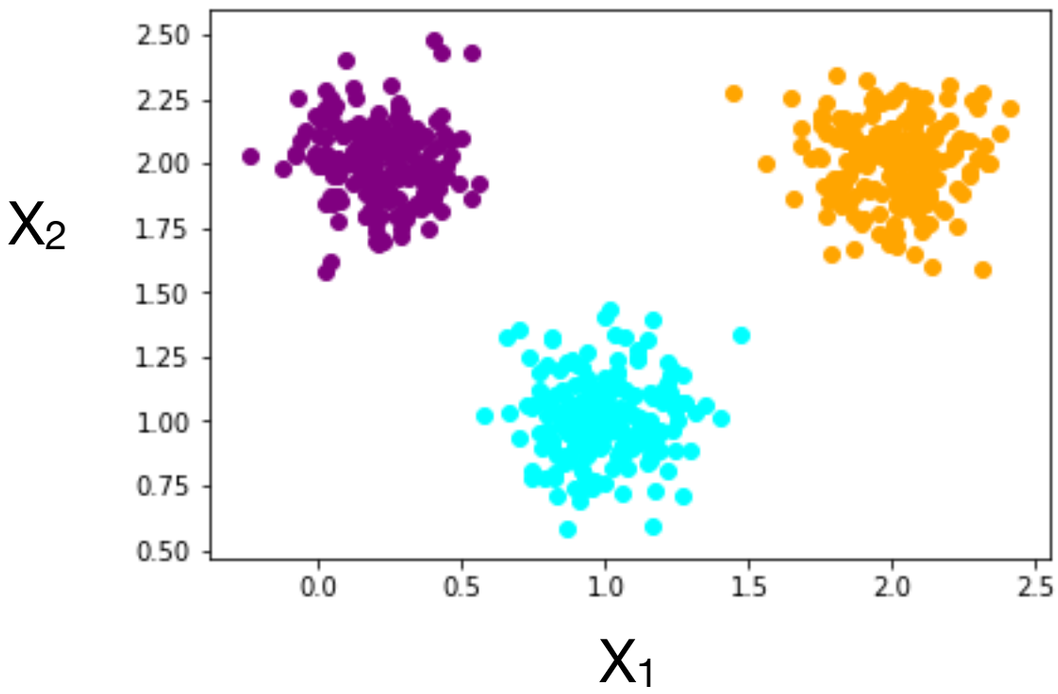
```
> Initialize random cluster indices
> Repeat until convergence:
>     For each data point:
>         Distance( point, center_j )
>         new_index ← min_j(distance)
>         total_distance += distance
```

The data:

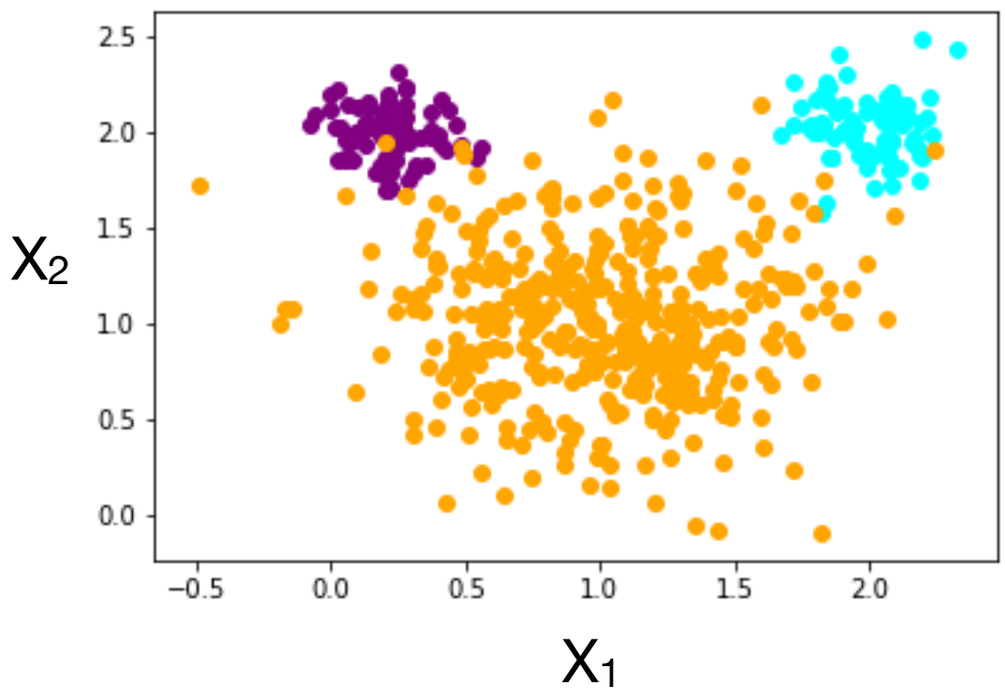


Problems with k-means

Previous data:

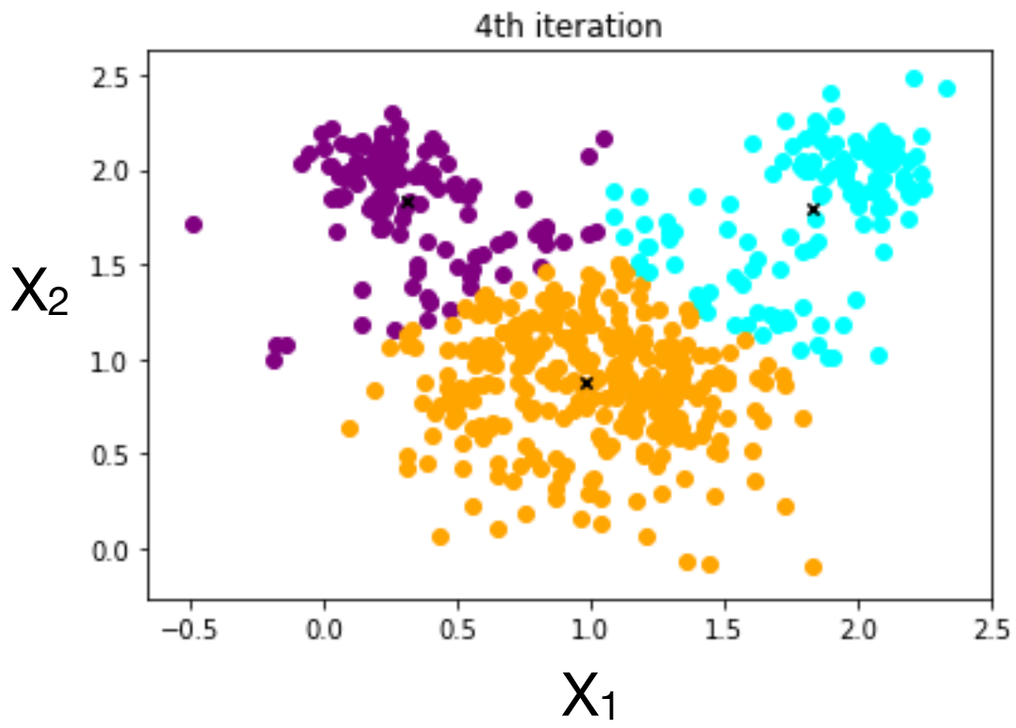


New data:



Poor fits to dissimilar clusters

Solution:
Soft k-means



Soft k-means Algorithm

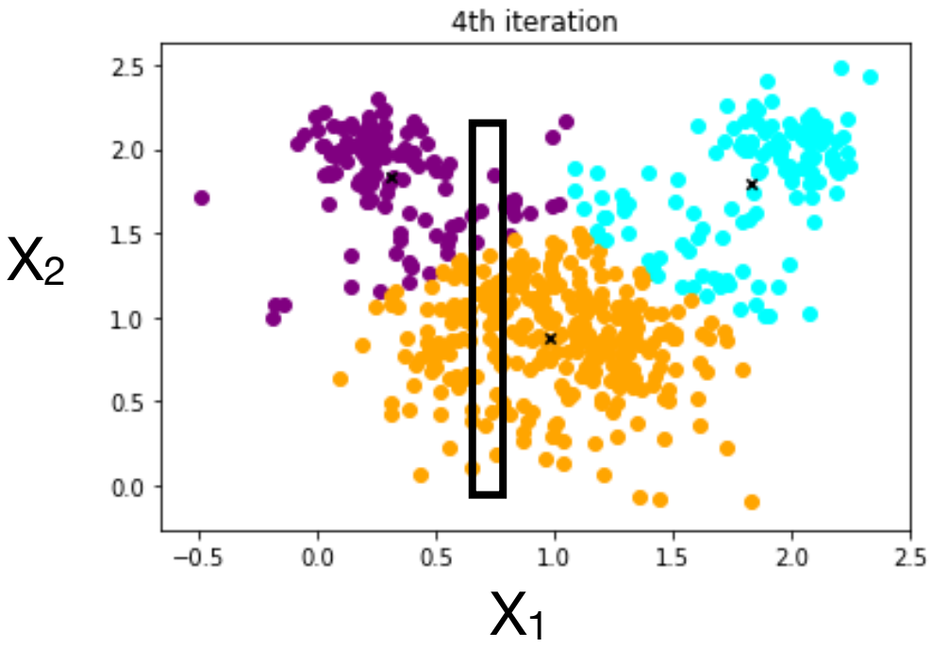
Hard k-means:

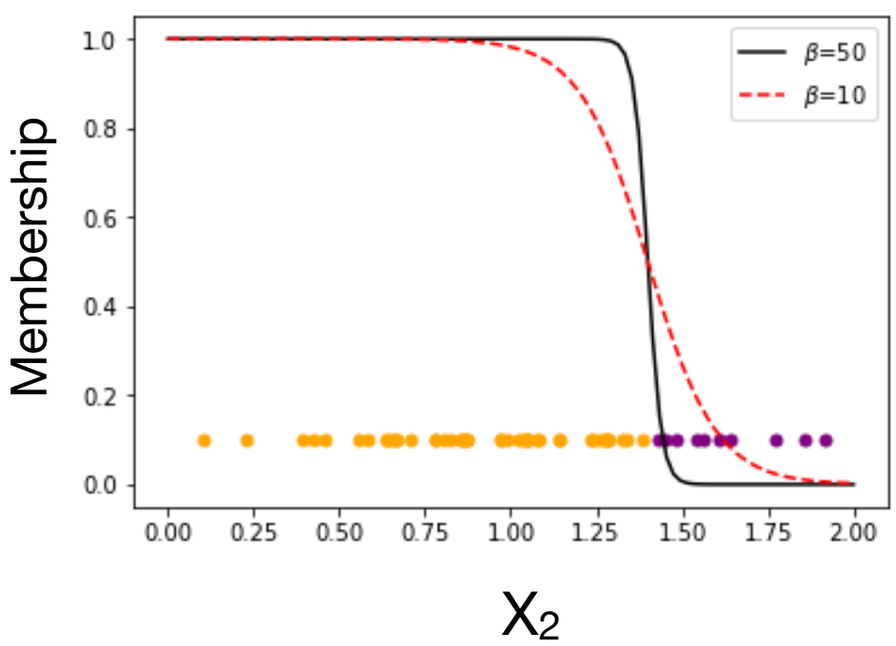
```
> new_index ← min_j(distance)
```

Soft k-means:

```
> membership ← r_j(distance)
```

$$r_i = \frac{\exp\left[-\beta \ \mathrm{dist}(x_i, c_j)\right]}{Z}$$

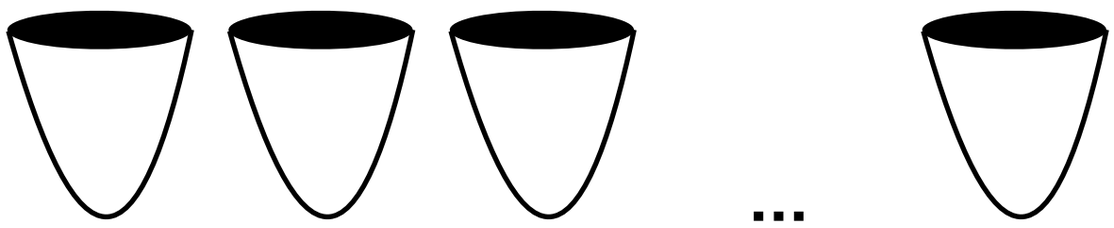




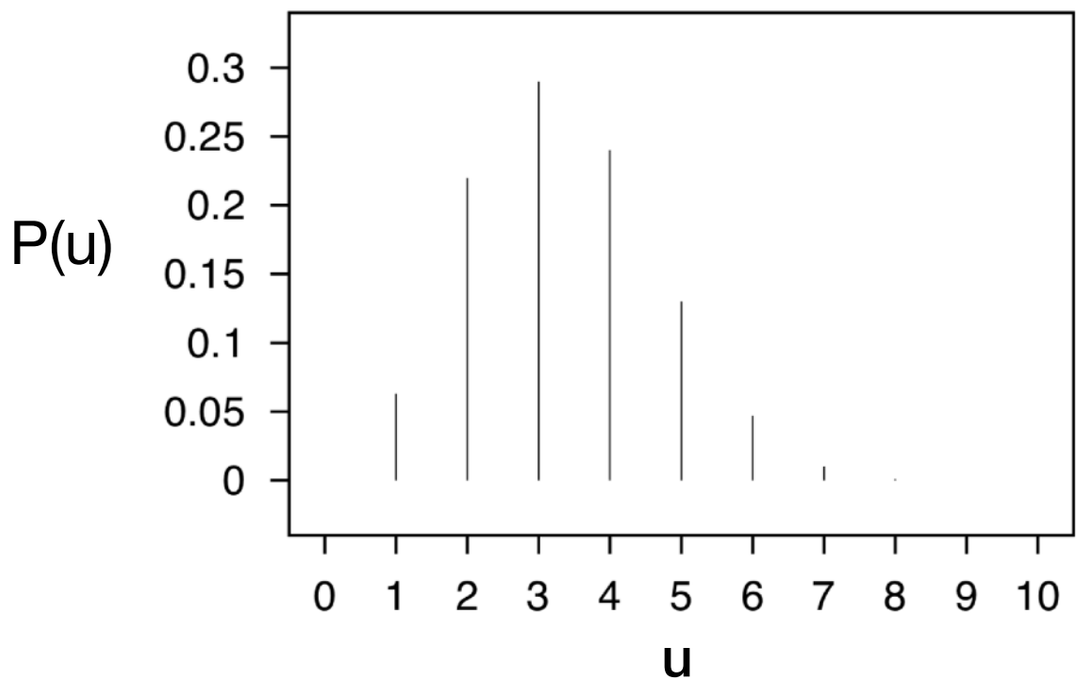
Soft k-means Algorithm

Utilizes full posterior distribution rather than approximate solution

Recall the urn problem:

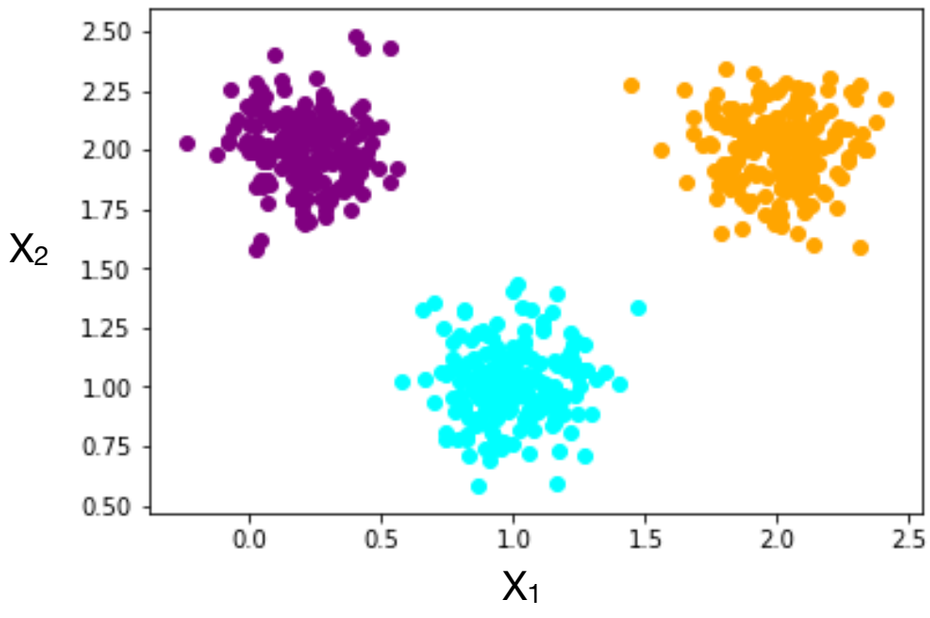


N=10 draws ● ● ● ○ ○ ... ○ → Cluster = urn 3

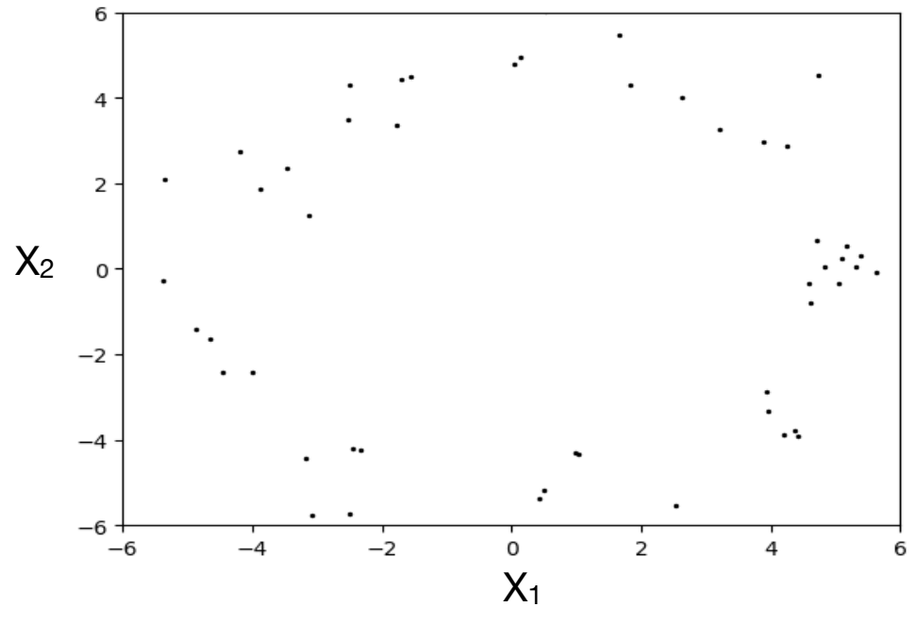


→ Cluster distribution

Another Problem: Choosing k



k is “obvious”



What is k?

Few clusters

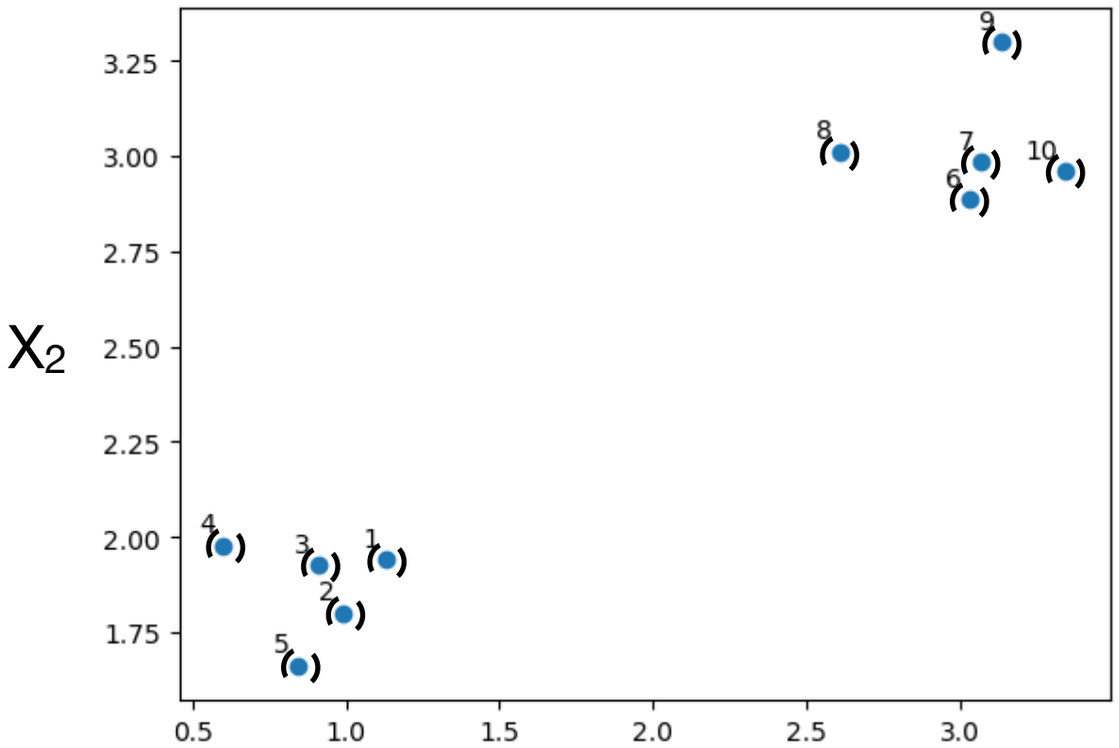
Many clusters

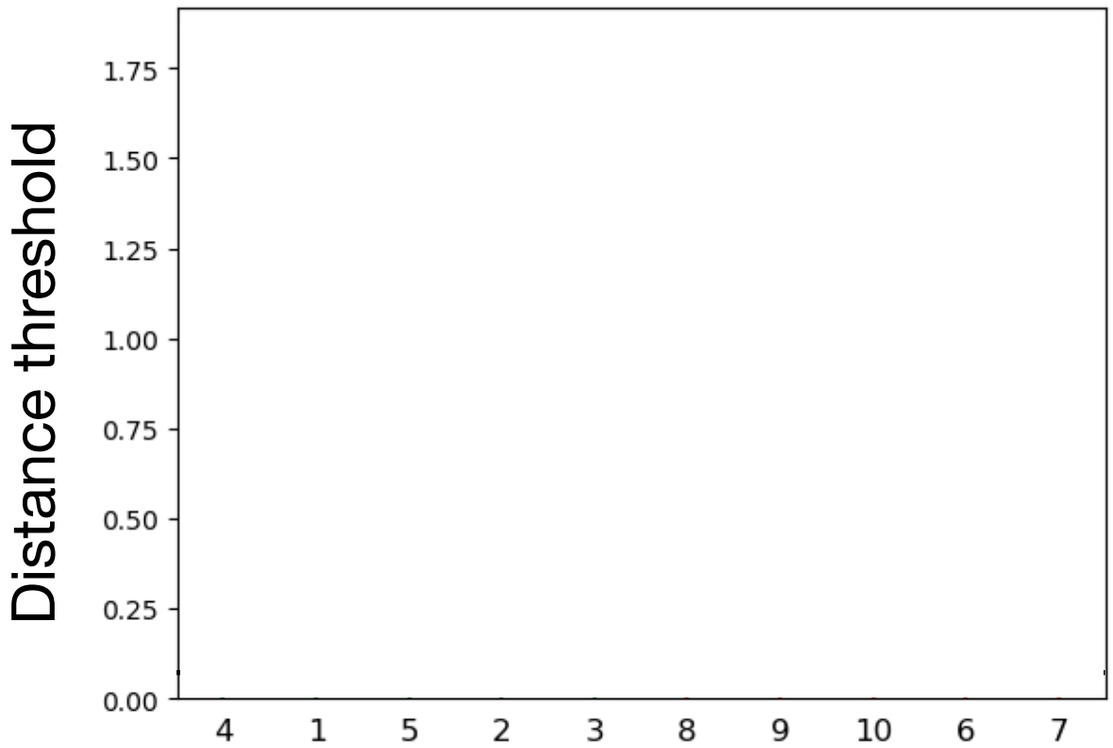
Compression

Accuracy

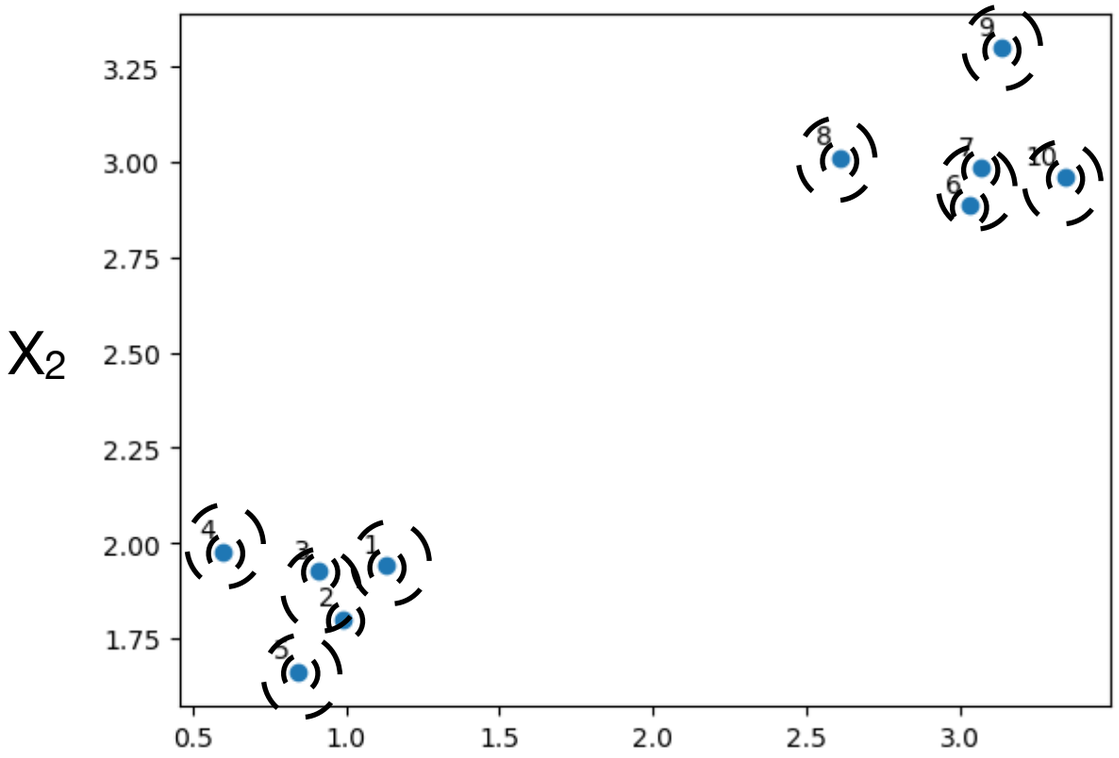
Recall hypothesis comparison with many parameters

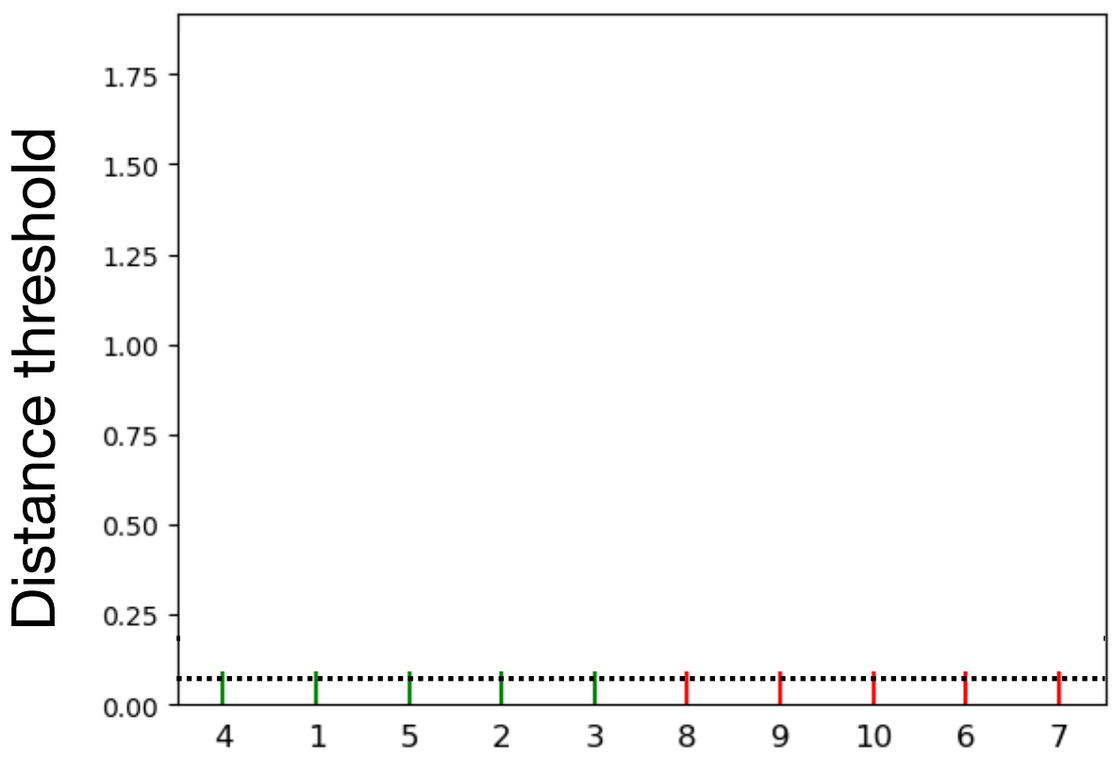
Solution: Dendrogram



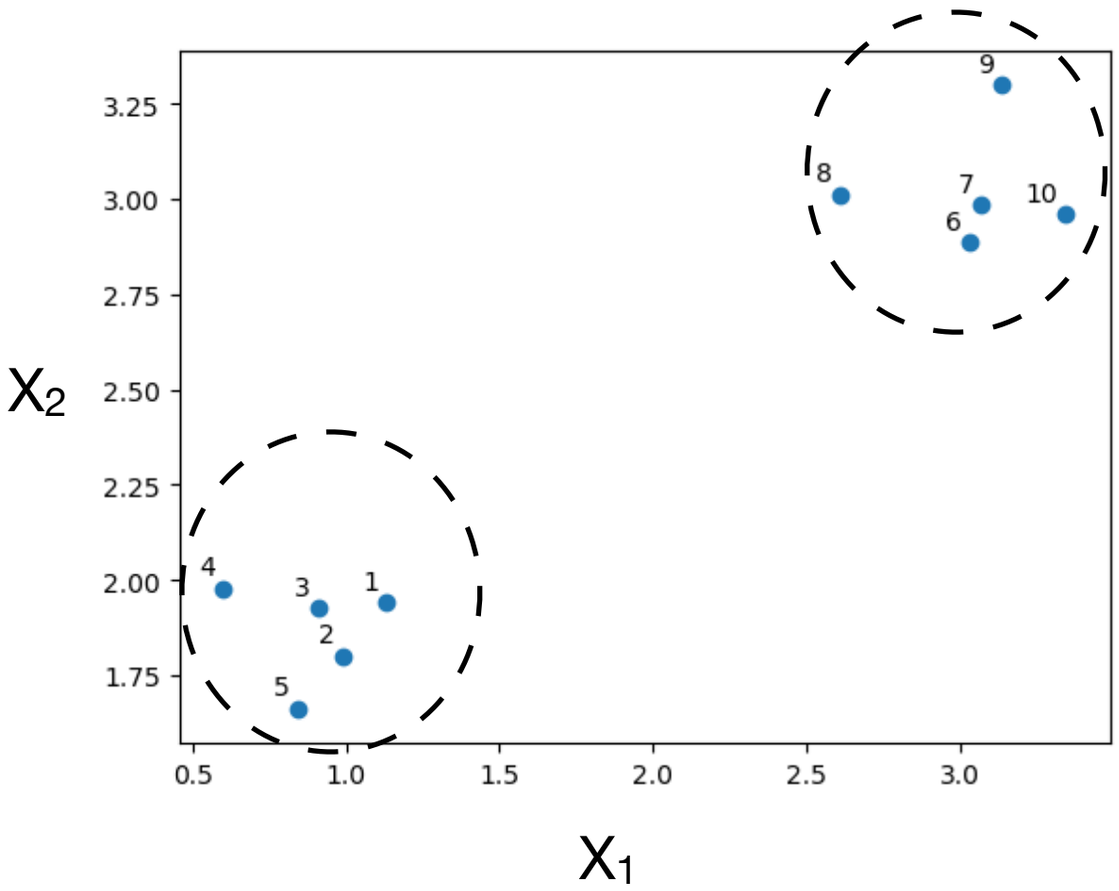


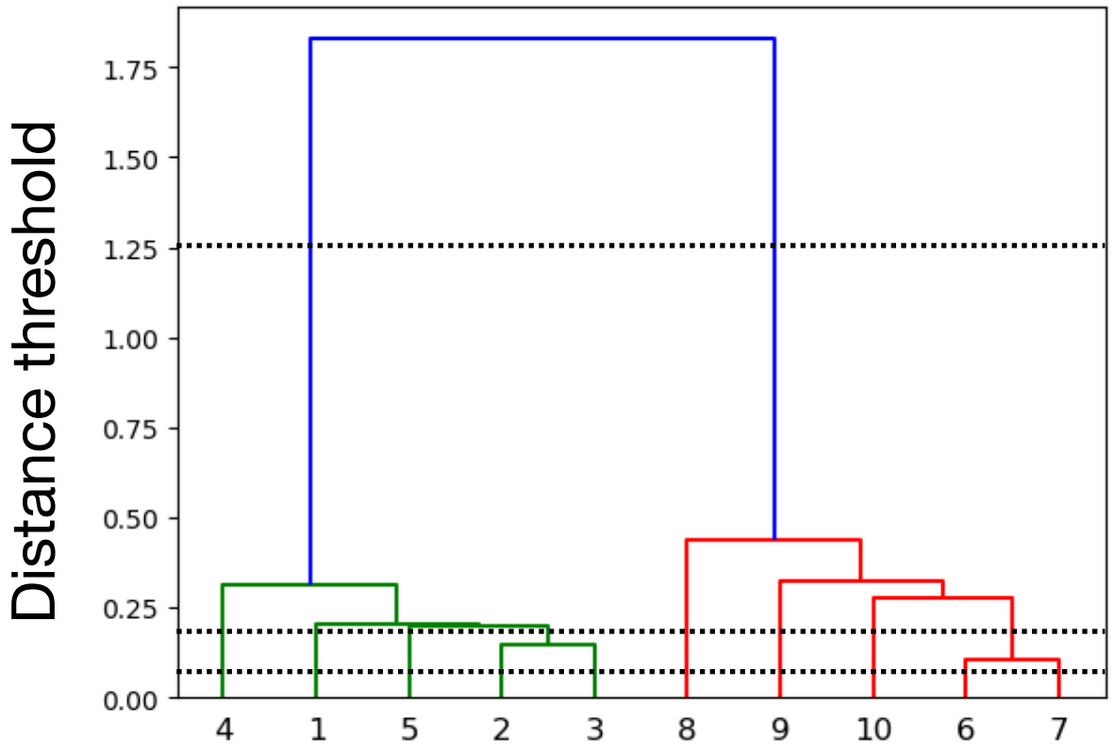
Solution: Dendrogram





Solution: Dendrogram

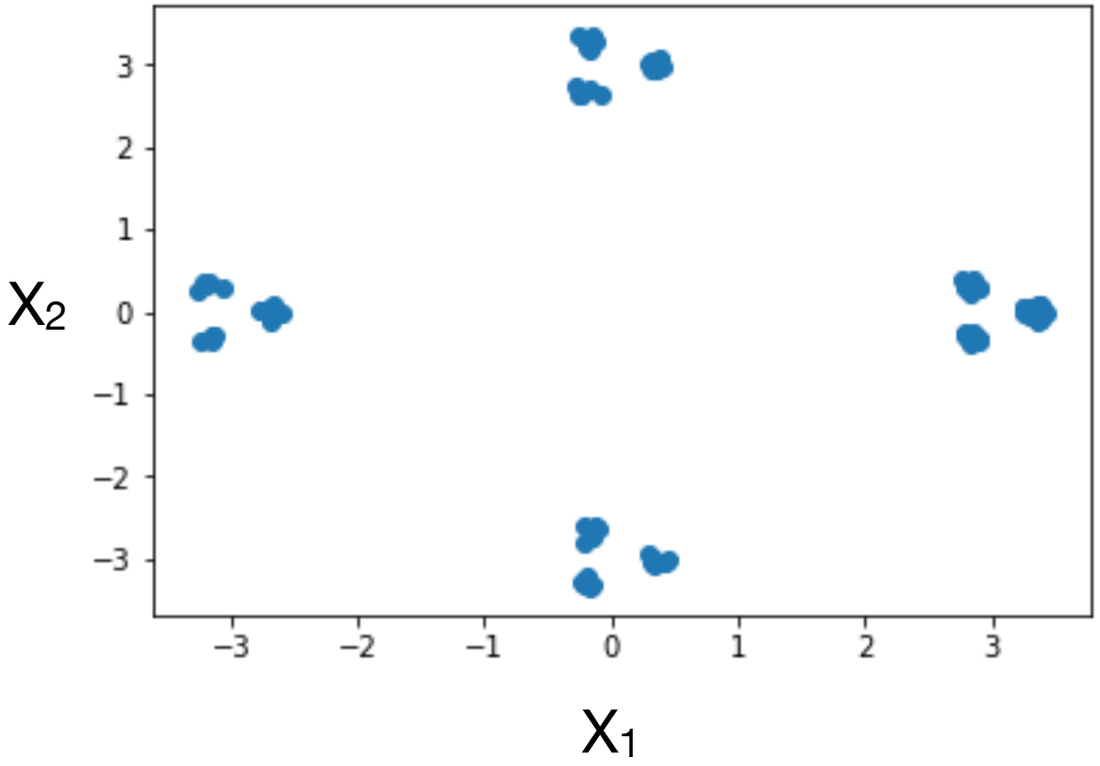


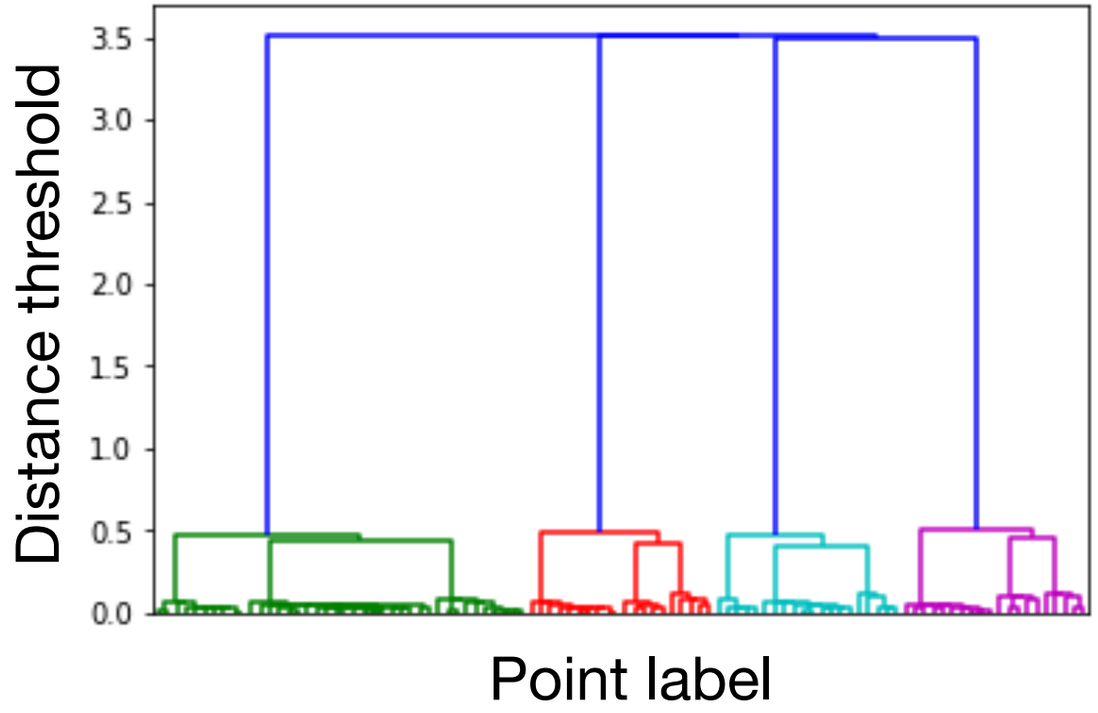


Definite criteria for k:
Choose maximum change in distance threshold

Hypothesis comparison:
Allow the data to speak for itself

No Silver Bullet



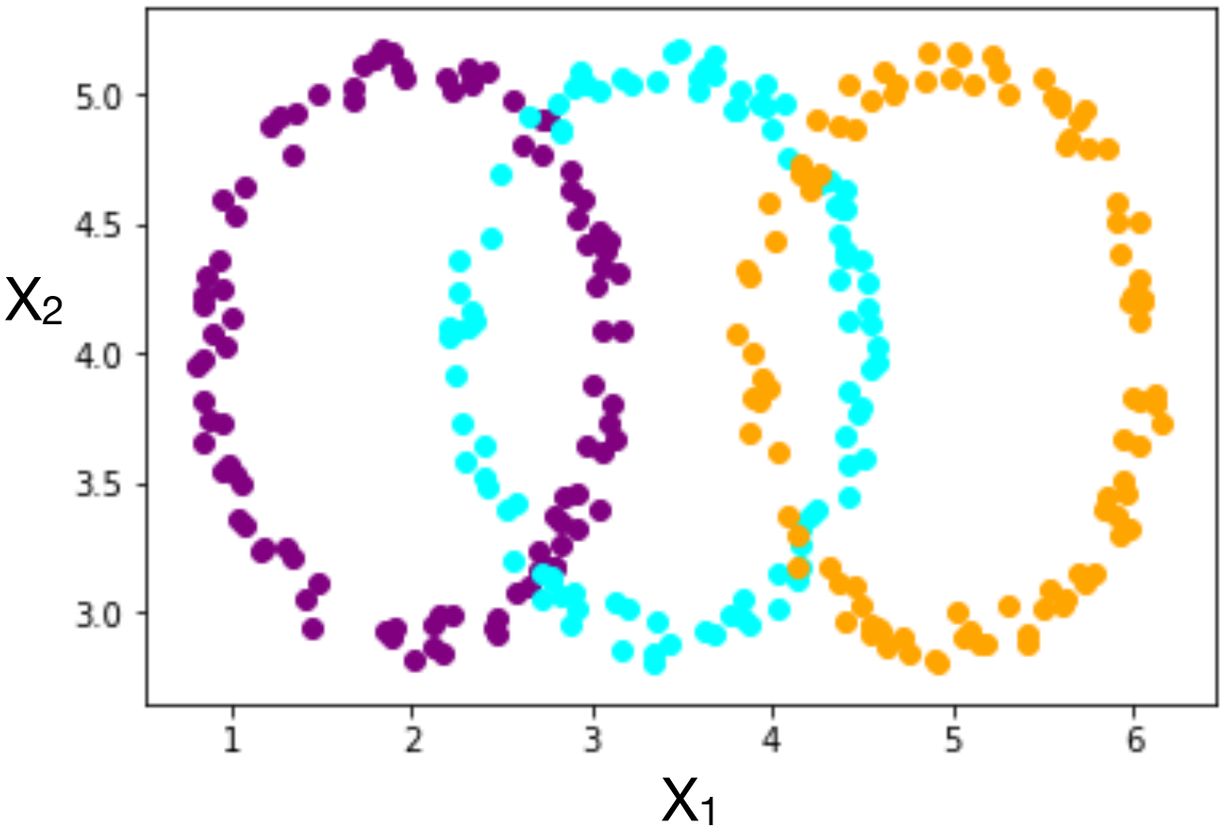


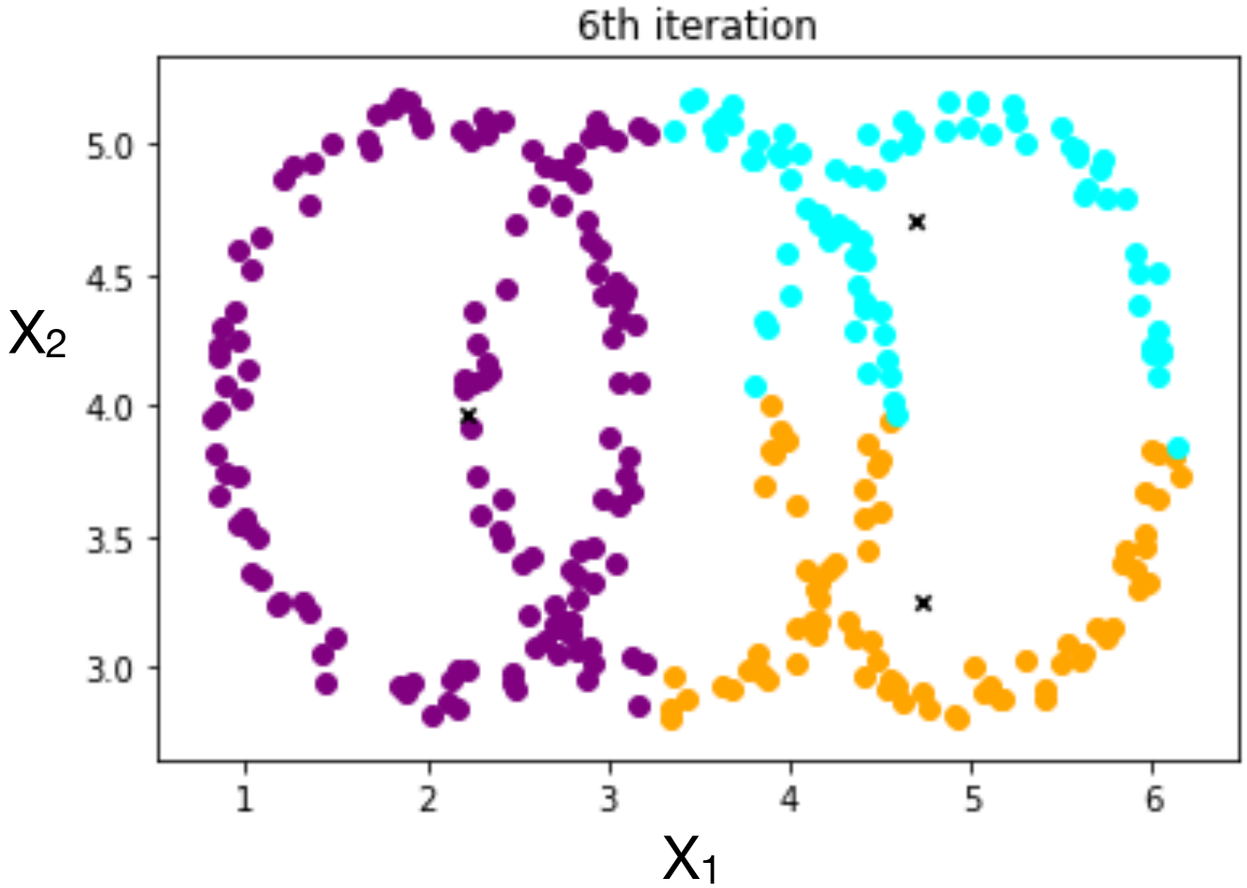
Which threshold is “correct?”
k=4?
k=12?

The question is not “which k should I choose?”
but rather “how is my data structured?”

No Silver Bullet Continued

Some data is very difficult to cluster





Clustering is

Takeaways

Clustering is a valuable tool for dimensional reduction.

Modified k-means is a good tool for clustering.

Clustering strategy is ultimately sensitive to data’s structure.