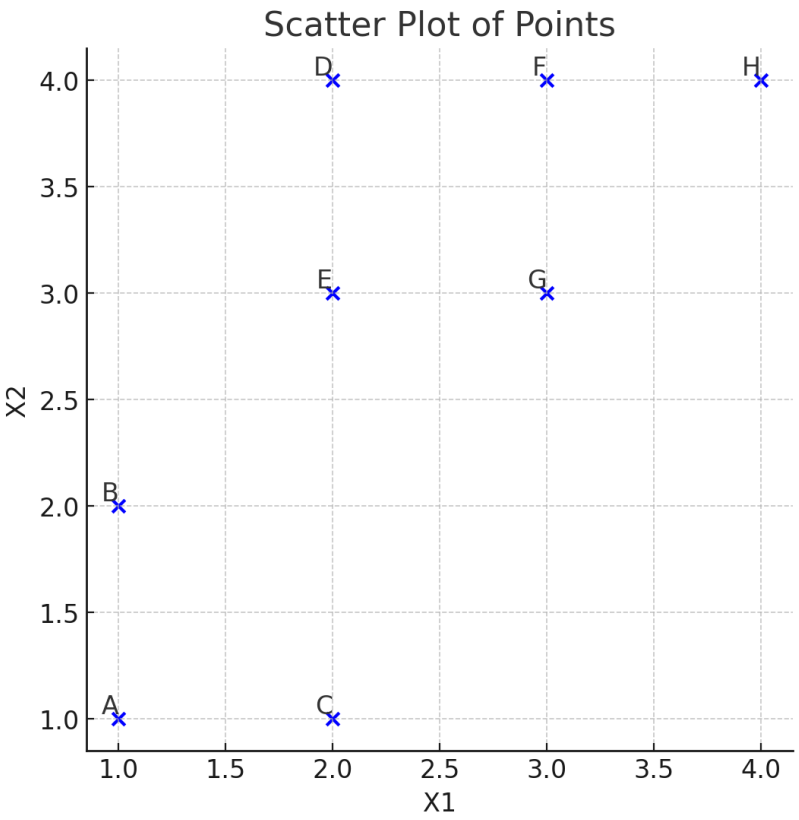
K-Means Exercise

Point Data			
Point	X1	X2	Color
A	1	1	
B	1	2	
C	2	1	
D	2	4	
E	2	3	
F	3	4	
G	3	3	
H	4	4	

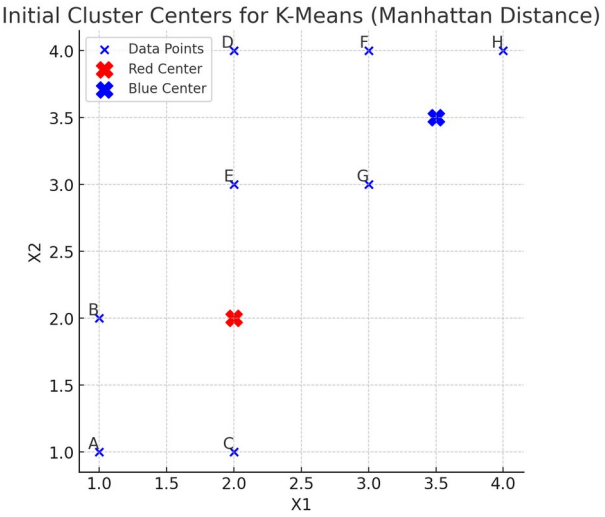
Red Center		
Blue Center		



K-Means requires an arbitrary (randomly) selected set of centers to start

Iteration	1	
Red Center	2	2
Blue Center	3,5	3,5

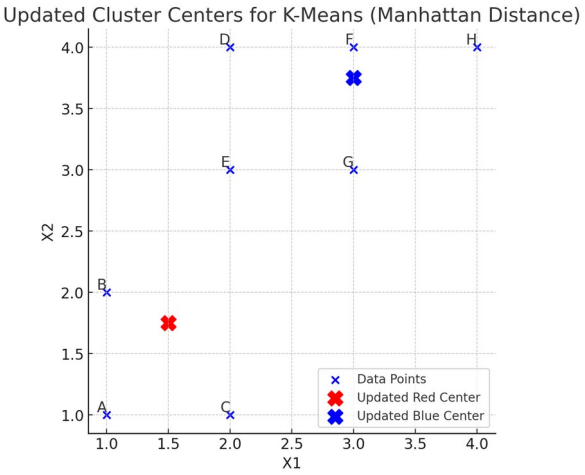
Point Data					
Point	X1	X2	D_{RED}	D_{BLUE}	Color
A	1	1	2	5	Red
B	1	2	1	4	Red
C	2	1	1	4	Red
D	2	4	2	2	Red
E	2	3	1	2	Red
F	3	4	3	1	Blue
G	3	3	2	1	Blue
H	4	4	4	1	Blue



In the next iterations, we calculate the centers based on the classification of the previous iteration.

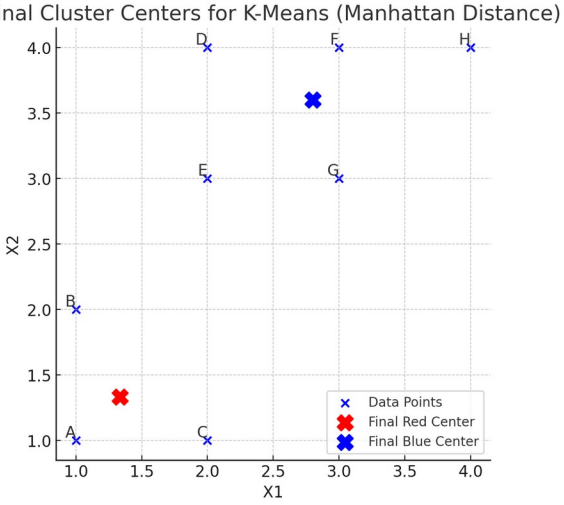
Iteration	2	
Red Center	1,5	1,8
Blue Center	3	3,8

Point Data					
Point	X1	X2	D_{RED}	D_{BLUE}	Color
A	1	1	1,25	4,75	Red
B	1	2	2,75	3,75	Red
C	2	1	1,25	3,75	Red
D	2	4	2,75	1,25	**Blue**
E	2	3	1,75	1,75	Red
F	3	4	3,75	0,25	Blue
G	3	3	2,75	0,75	Blue
H	4	4	4,75	1,25	Blue



Iteration	3	
Red Center	1,3	1,3
Blue Center	2,8	3,6

Point Data					
Point	X1	X2	D_{RED}	D_{BLUE}	Color
A	1	1	0,66	4,4	Red
B	1	2	1	3,4	Red
C	2	1	1	3,4	Red
D	2	4	3,33	1,2	**Blue**
E	2	3	2,33	1,4	**Blue**
F	3	4	4,33	0,6	Blue
G	3	3	3,33	0,8	Blue
H	4	4	5,33	1,6	Blue



Iteration	4	
Red Center	1,3	1,3
Blue Center	2,8	3,6

Centers have stabilized, our existing classification is stable.

Notes:

1- Because decimal numbers are not stored exactly in computers, but in a particular floating point format, sometimes rounding errors make the canter calculations move around a point but not exactly stabilize. So we might define a small tolerance and check if the change in the center calculations within the tolerance.

2- An alternate method would be to check if there is any actual change in the classifications themselves. If the classifications are the same, then we have achieved stability.

Question:

If we keep the centers fixed from now on, and then start adding points (non-updating k-means prediction, can we manage “concept drift”?

Sudden Drift	A rapid and abrupt change in the data distribution, where the previous patterns become immediately obsolete. Example: Consumer behavior changes drastically after a major event
Gradual Drift	The data distribution slowly shifts over time, and both old and new concepts may coexist for a while. Example: A new product feature slowly gains popularity, while the older one still sees some use.
Incremental Drift	Small, continuous changes accumulate over time, leading to a significant shift in the data distribution. Example: A user's preferences slowly evolve based on long-term behavior trends.
Recurring Drift	Previously seen data distributions return periodically. Example: Retail shopping habits during holidays like Black Friday or New Year re-appear each year.

Methods to check for "drift"

Kolmogorov–Smirnov (K–S) Test

Jensen–Shannon Divergence (JSD)

Population Stability Index (PSI)