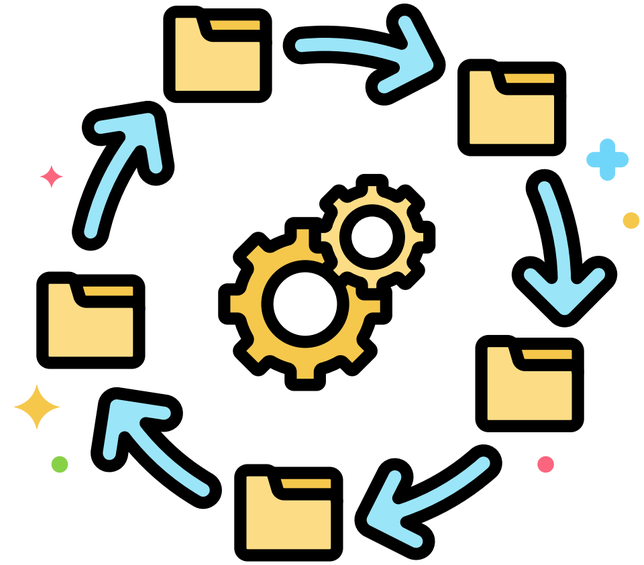
Data Pre-processing

codefinity

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes



Data Pre-processing

codefinity

Missing Values

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Missing Values Filled

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	63777.8	Yes
France	35	58000	Yes
Spain	38.7778	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Data Pre-processing

codefinity

Categorical Features

Age Group	Education Level	City
Young	High School	New York
Young	Bachelor	Los Angeles
Middle-Aged	High School	Chicago
Senior	Master	New York
Middle-Aged	PhD	Los Angeles
Young	High School	Chicago
Senior	Bachelor	Miami
Middle-Aged	Master	New York
Young	PhD	Miami
Senior	PhD	Chicago

Ordinal Encoding

Age Group	Education Level	City
0.0	0.0	New York
0.0	1.0	Los Angeles
1.0	0.0	Chicago
2.0	2.0	New York
1.0	3.0	Los Angeles
0.0	0.0	Chicago
2.0	1.0	Miami
1.0	2.0	New York
0.0	3.0	Miami
2.0	3.0	Chicago

Data Pre-processing

codefinity

Categorical Features

Age Group	Education Level	City
Young	High School	New York
Young	Bachelor	Los Angeles
Middle-Aged	High School	Chicago
Senior	Master	New York
Middle-Aged	PhD	Los Angeles
Young	High School	Chicago
Senior	Bachelor	Miami
Middle-Aged	Master	New York
Young	PhD	Miami
Senior	PhD	Chicago

One Hot Encoding

Age Group	Education Level	Chicago	Los Angeles	Miami	New York
0.0	0.0	0.0	0.0	0.0	1.0
0.0	1.0	0.0	1.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	0.0
2.0	2.0	0.0	0.0	0.0	1.0
1.0	3.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0	0.0
2.0	1.0	0.0	0.0	1.0	0.0
1.0	2.0	0.0	0.0	0.0	1.0
0.0	3.0	0.0	0.0	1.0	0.0
2.0	3.0	1.0	0.0	0.0	0.0

Data Pre-processing

codefinity

Data Normalization

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Data Pre-processing

codefinity

Data Normalization

MinMax = (x - min) / (max - min)

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Country	Age	Salary	Purchased
France	0.73913	0.685714	No
Spain	0	0	Yes
Germany	0.130435	0.171429	No
Spain	0.478261	0.371429	No
Germany	0.565217	nan	Yes
France	0.347826	0.285714	Yes
Spain	nan	0.114286	No
France	0.913043	0.885714	Yes
Germany	1	1	No
France	0.434783	0.542857	Yes

Data Pre-processing

codefinity

Data Normalization

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

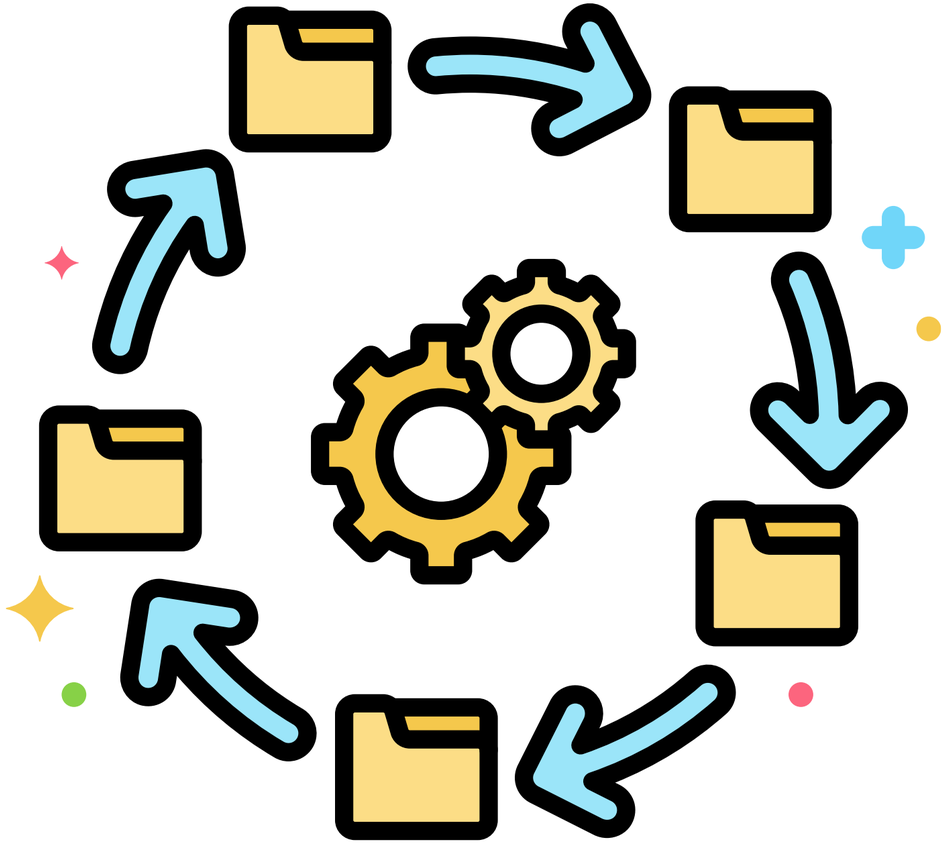
Standard Scaler

Country	Age	Salary	Purchased
France	0.719931	0.711013	No
Spain	-1.62368	-1.36438	Yes
Germany	-1.2101	-0.845529	No
Spain	-0.107224	-0.240207	No
Germany	0.168495	nan	Yes
France	-0.520801	-0.499631	Yes
Spain	nan	-1.01848	No
France	1.27137	1.31633	Yes
Germany	1.54709	1.66223	No
France	-0.245083	0.27864	Yes

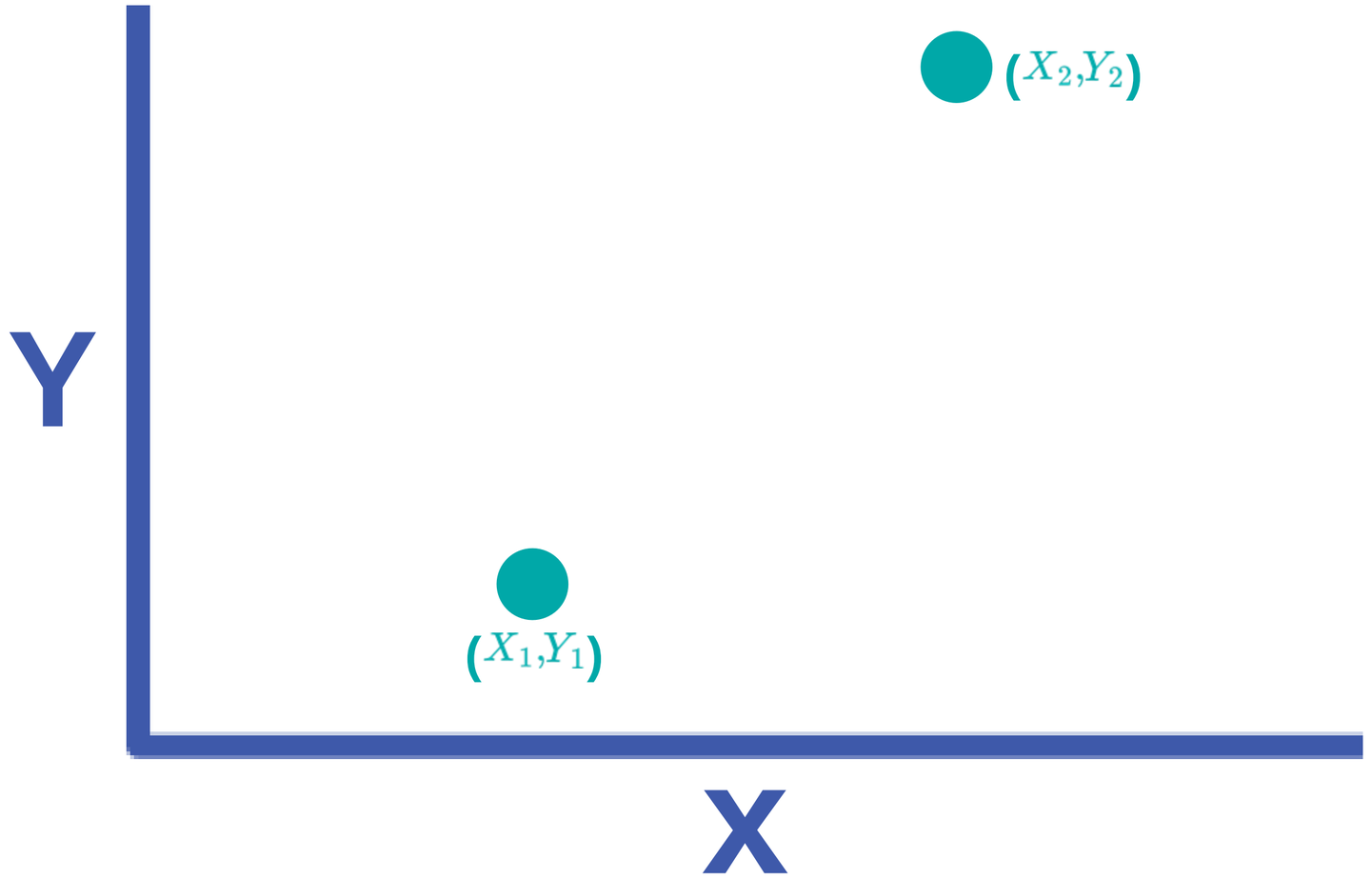
(x - mean) / standard deviation

Data Pre-processing

codefinity

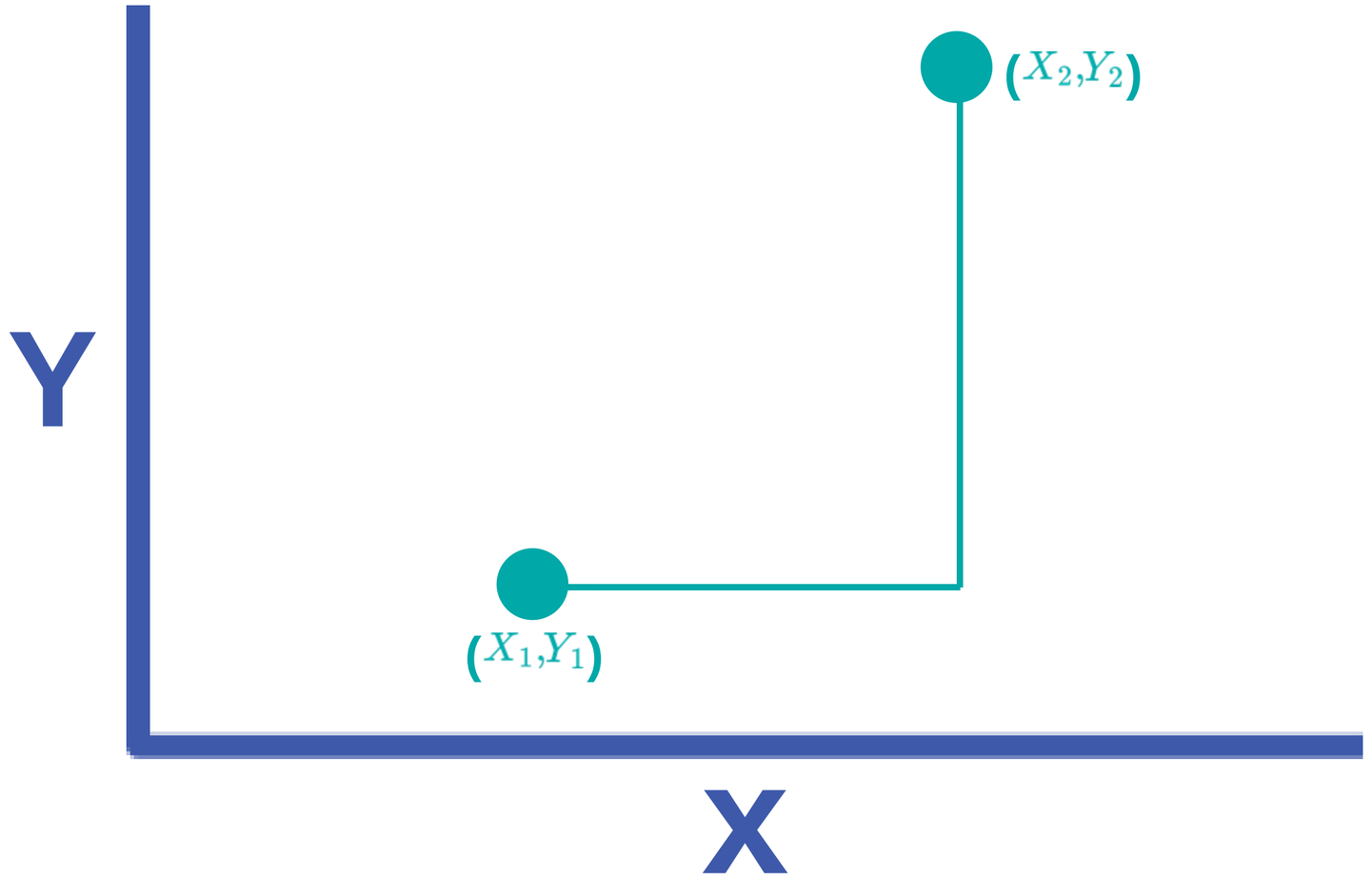


Distance Measures

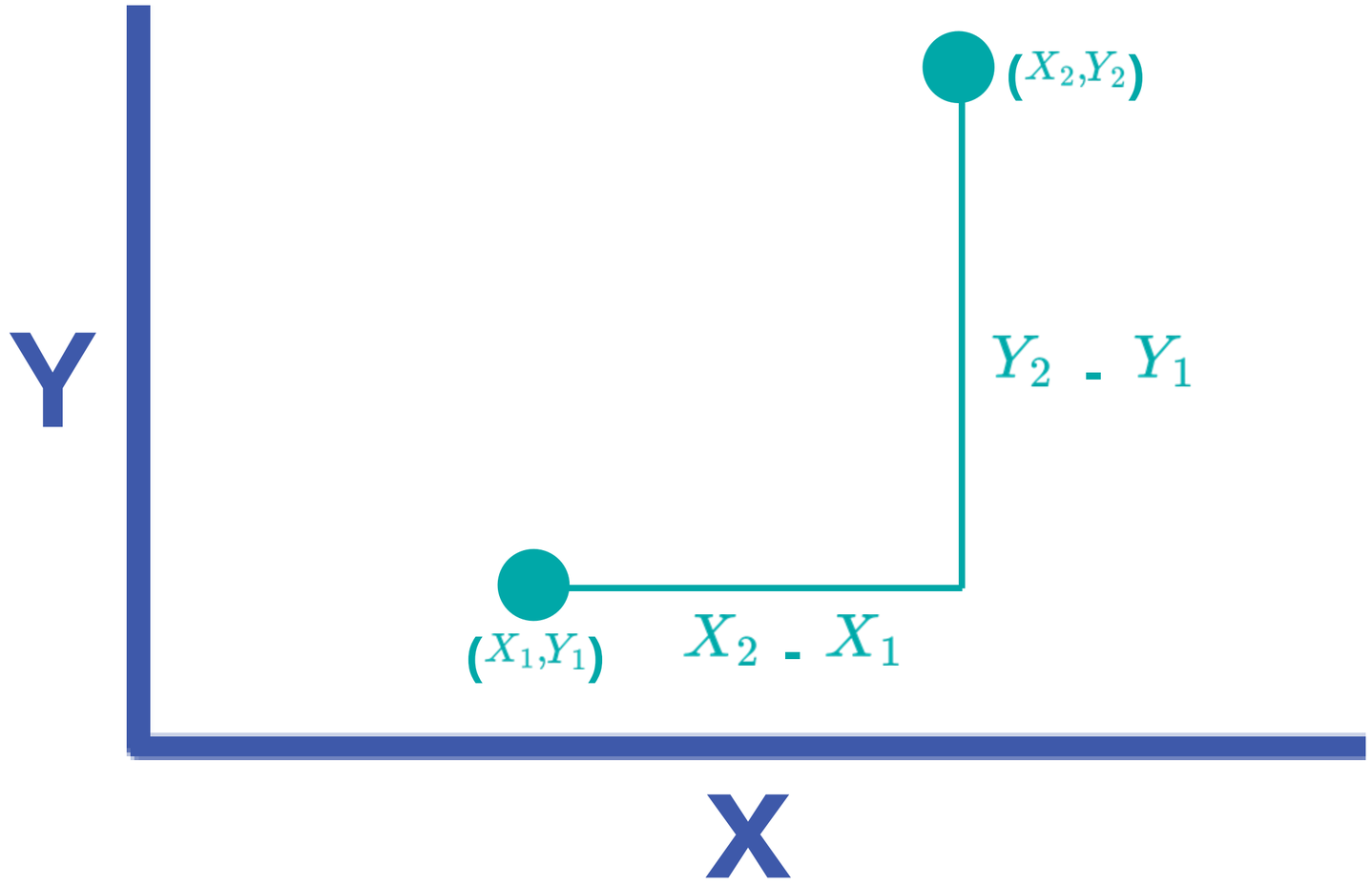


codefinity

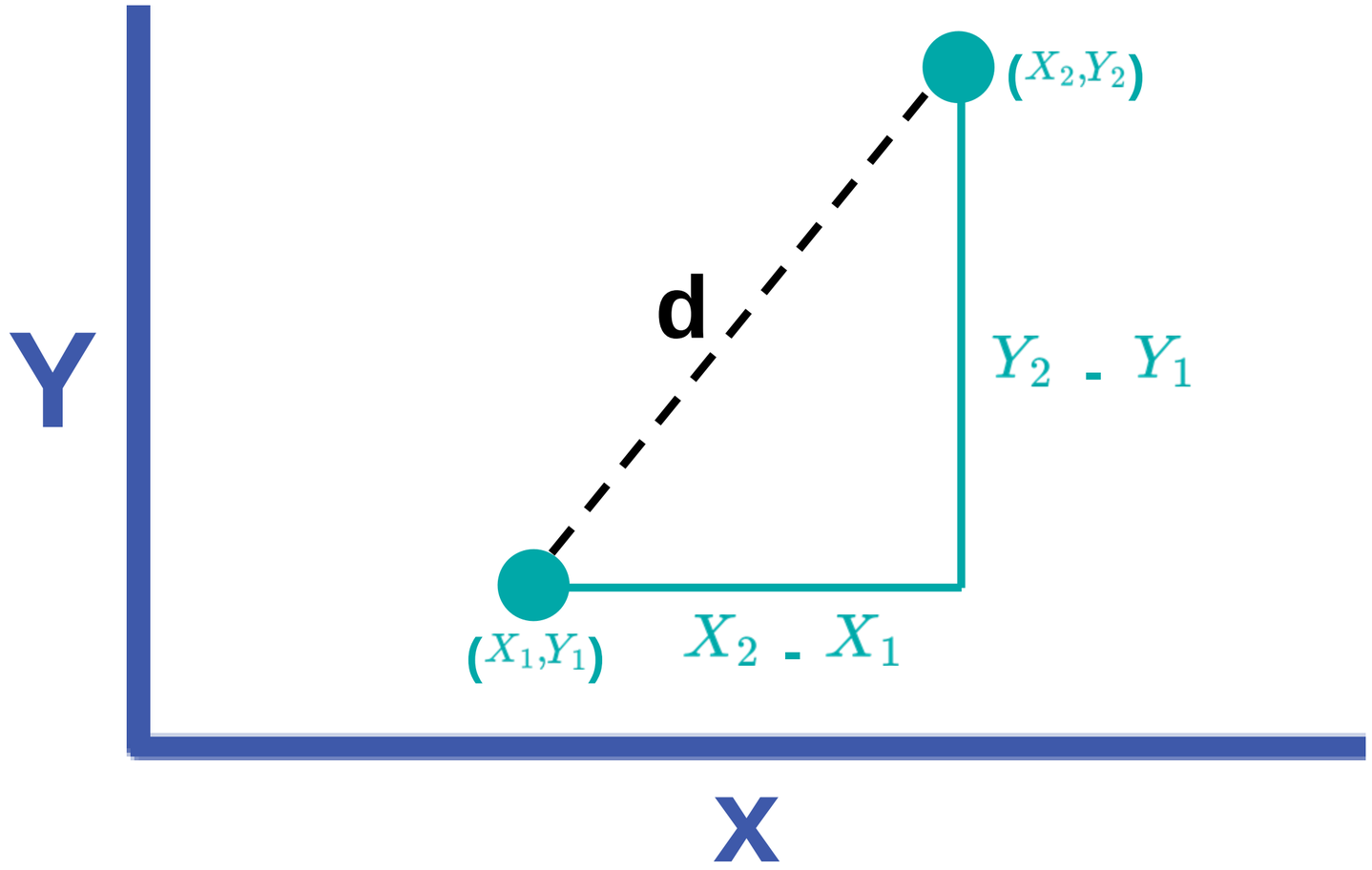
Distance Measures



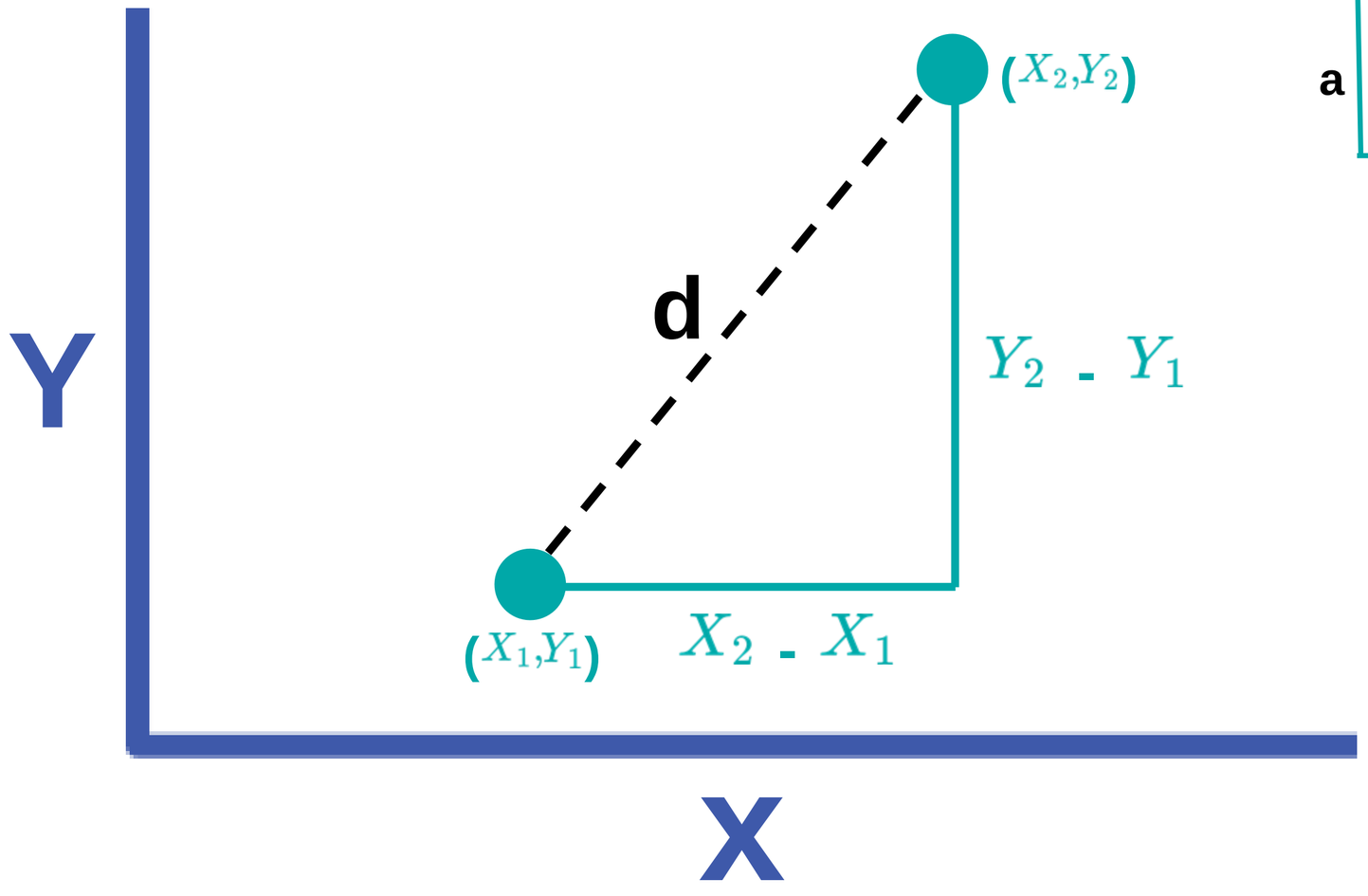
Distance Measures

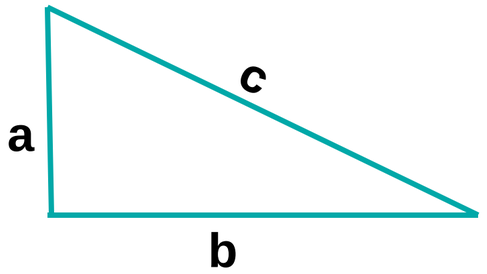


Distance Measures

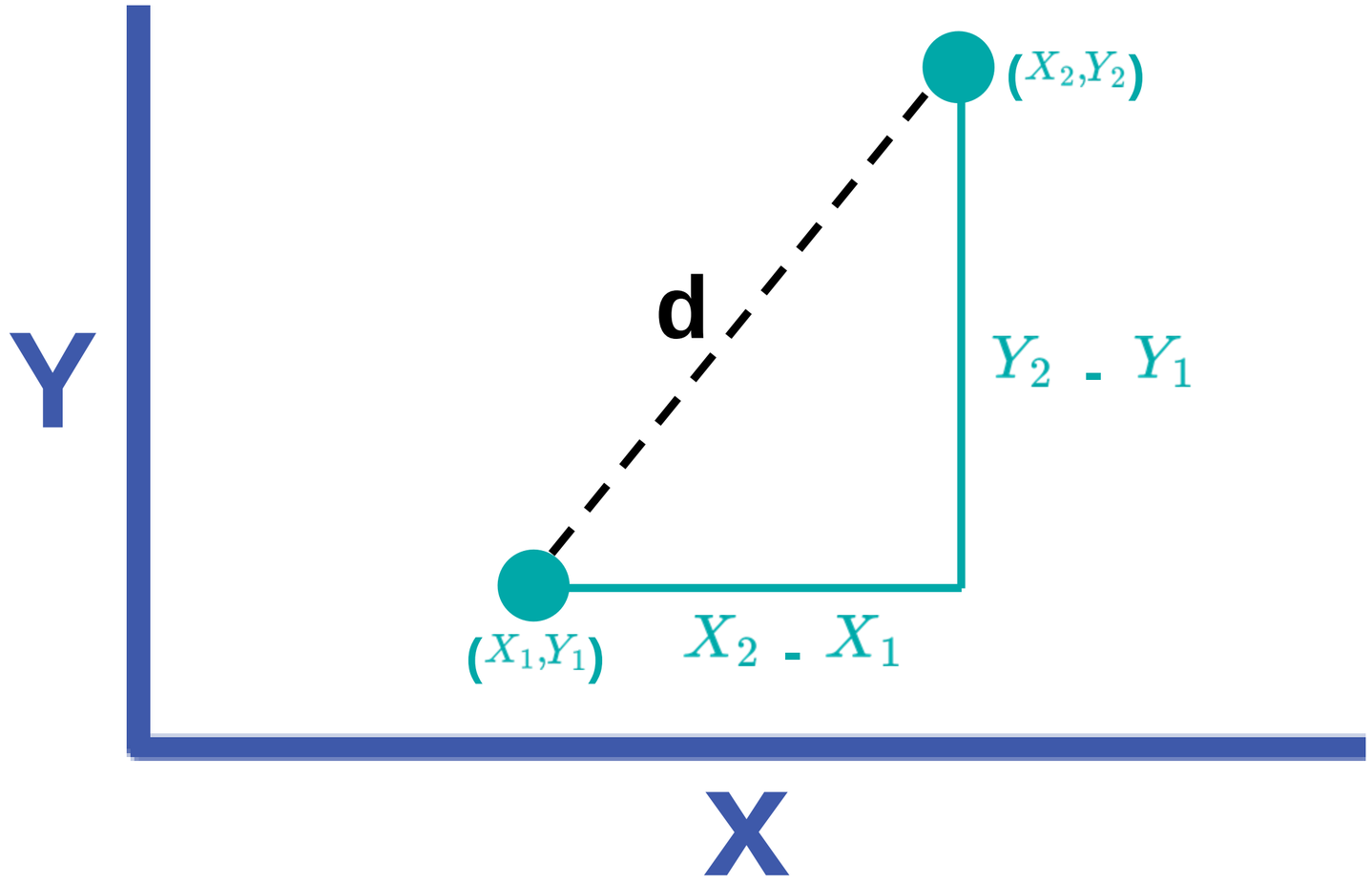


Distance Measures





Distance Measures

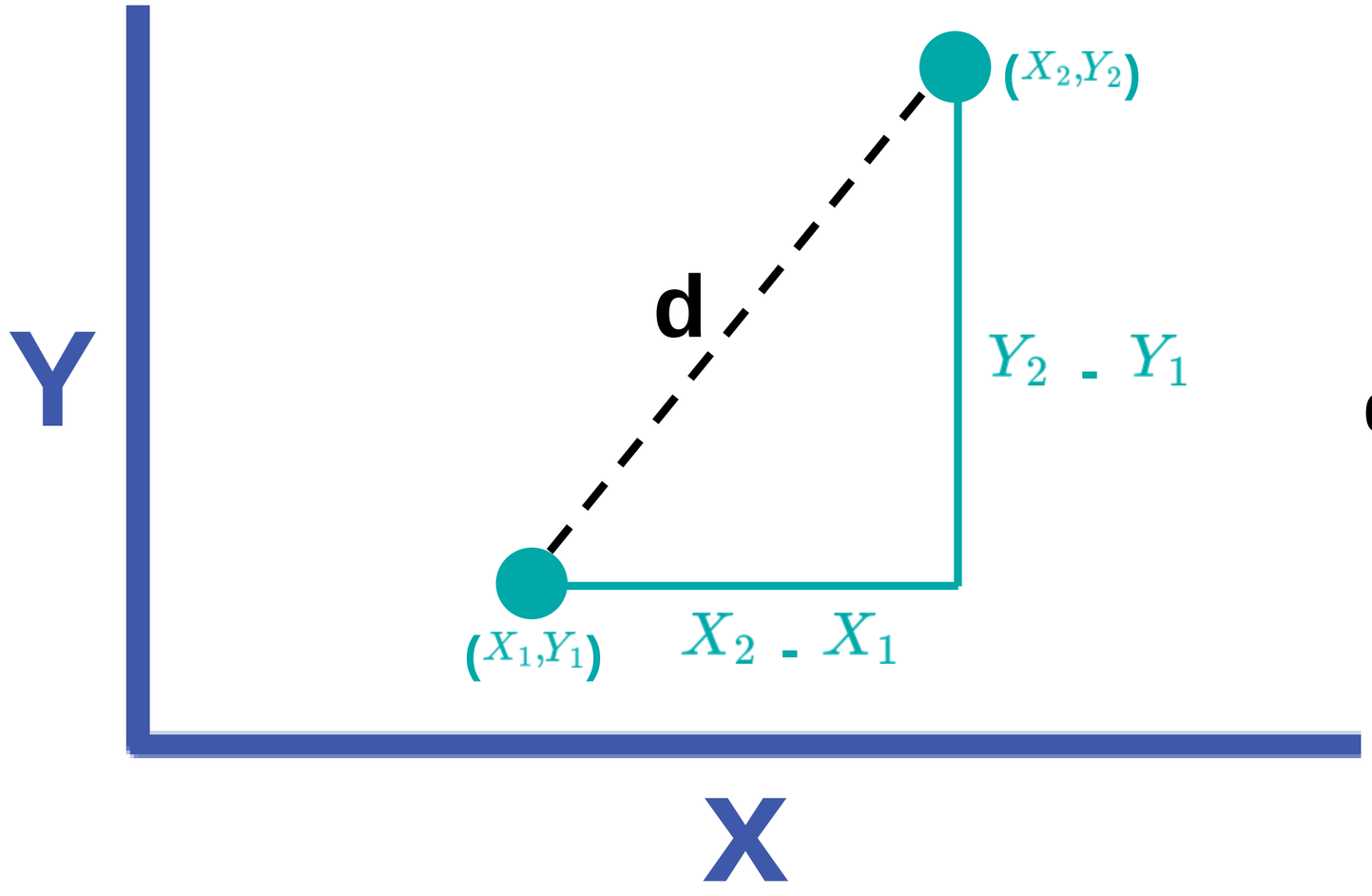


Euclidean Distance

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

codefinity

Distance Measures



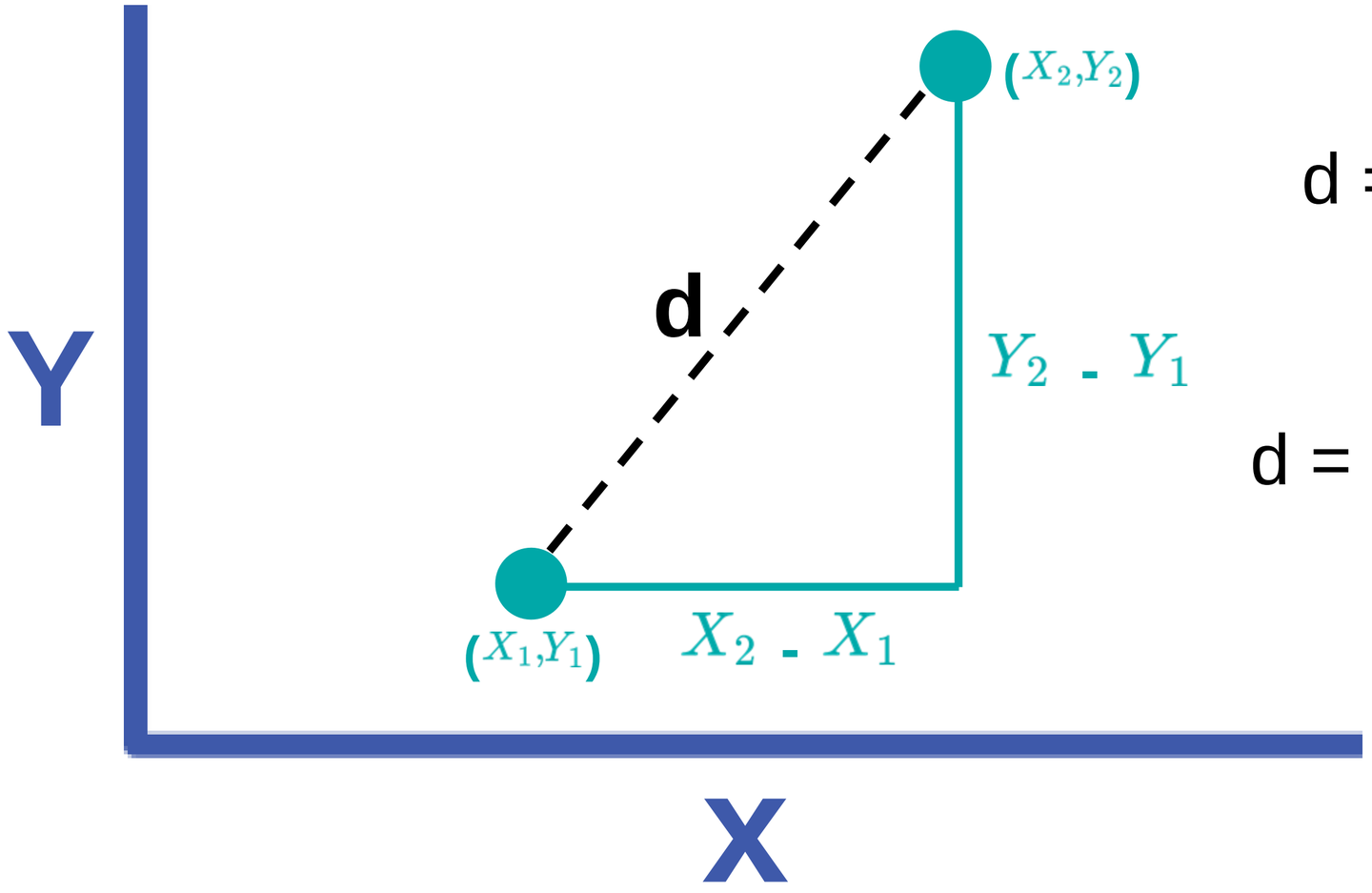
Euclidean Distance

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Manhattan distance

$$d = |X_2 - X_1| + |Y_2 - Y_1|$$

Distance Measures



Euclidean Distance

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Manhattan distance

$$d = |X_2 - X_1| + |Y_2 - Y_1|$$

Distance Measures

Euclidean Distance

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Manhattan distance

$$d = |X_2 - X_1| + |Y_2 - Y_1|$$

Distance Measures

Euclidean Distance

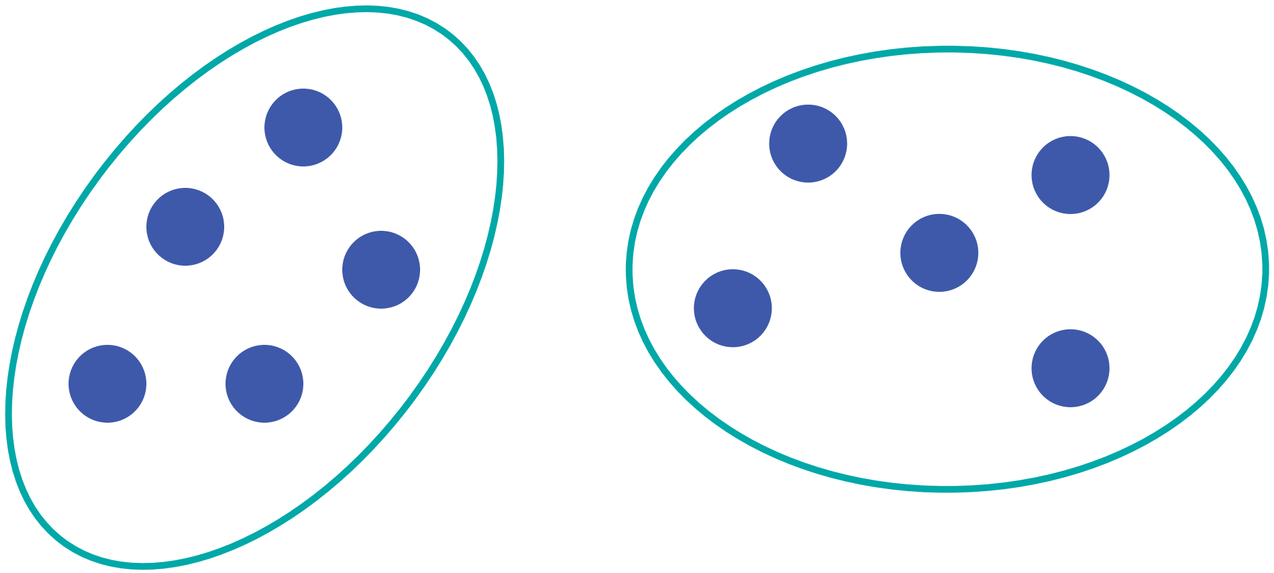
$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

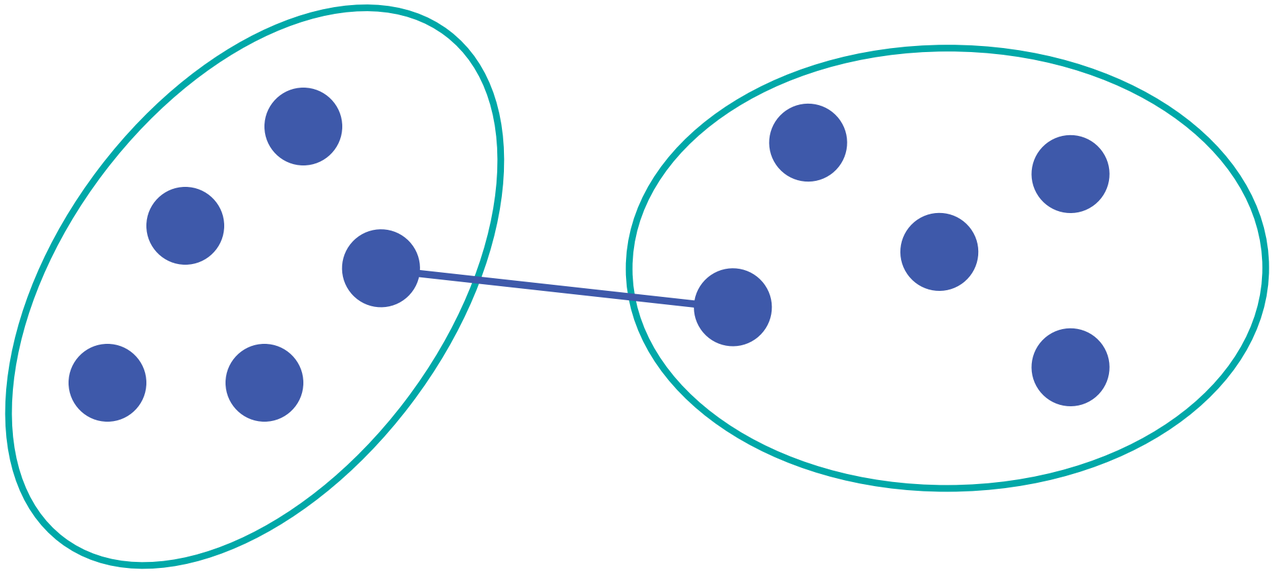
Manhattan distance

$$d = |X_2 - X_1| + |Y_2 - Y_1|$$

codefinity

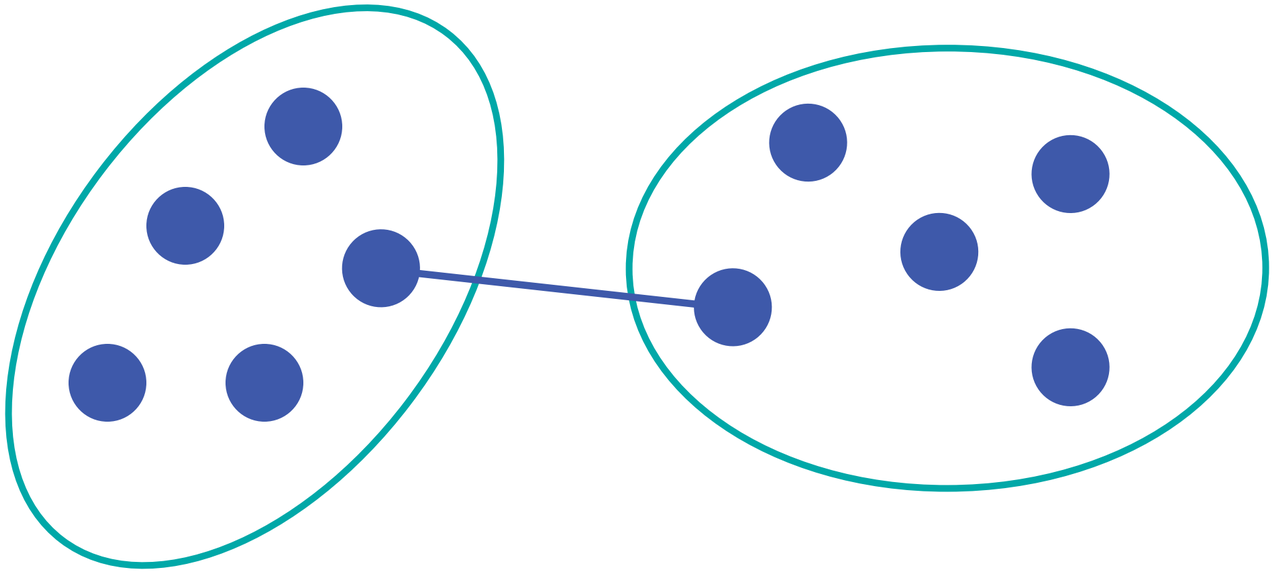
Linkage





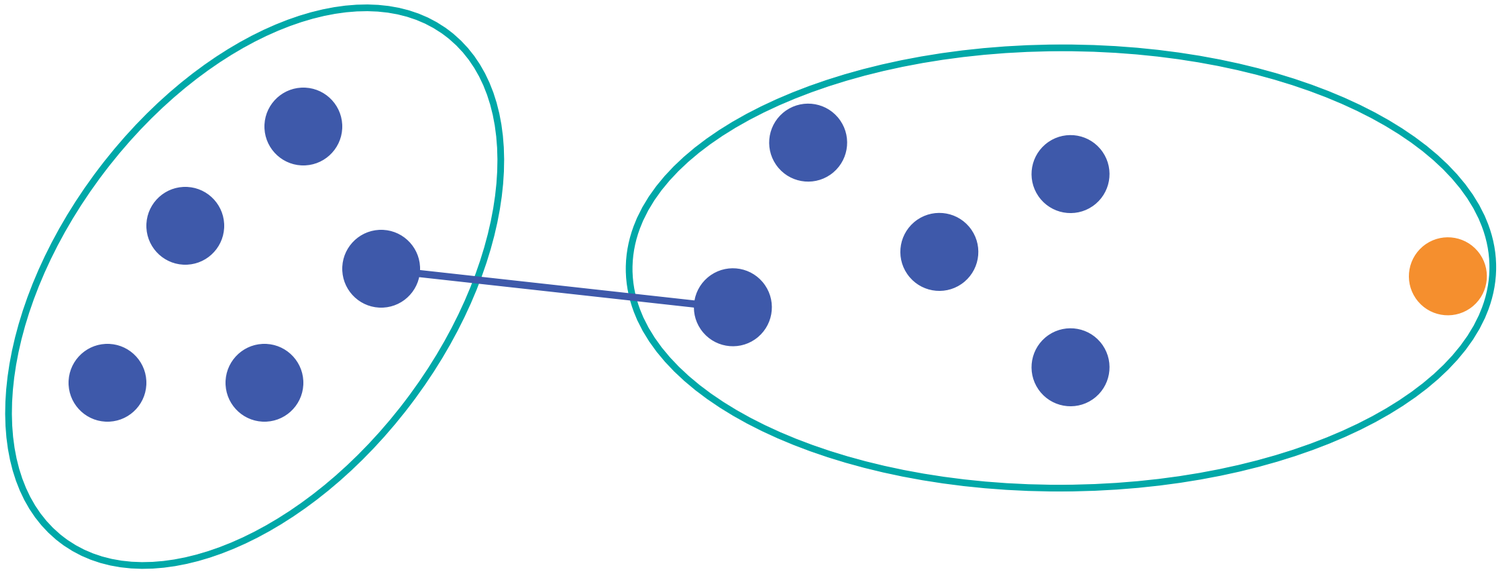
codefinity

Single Linkage



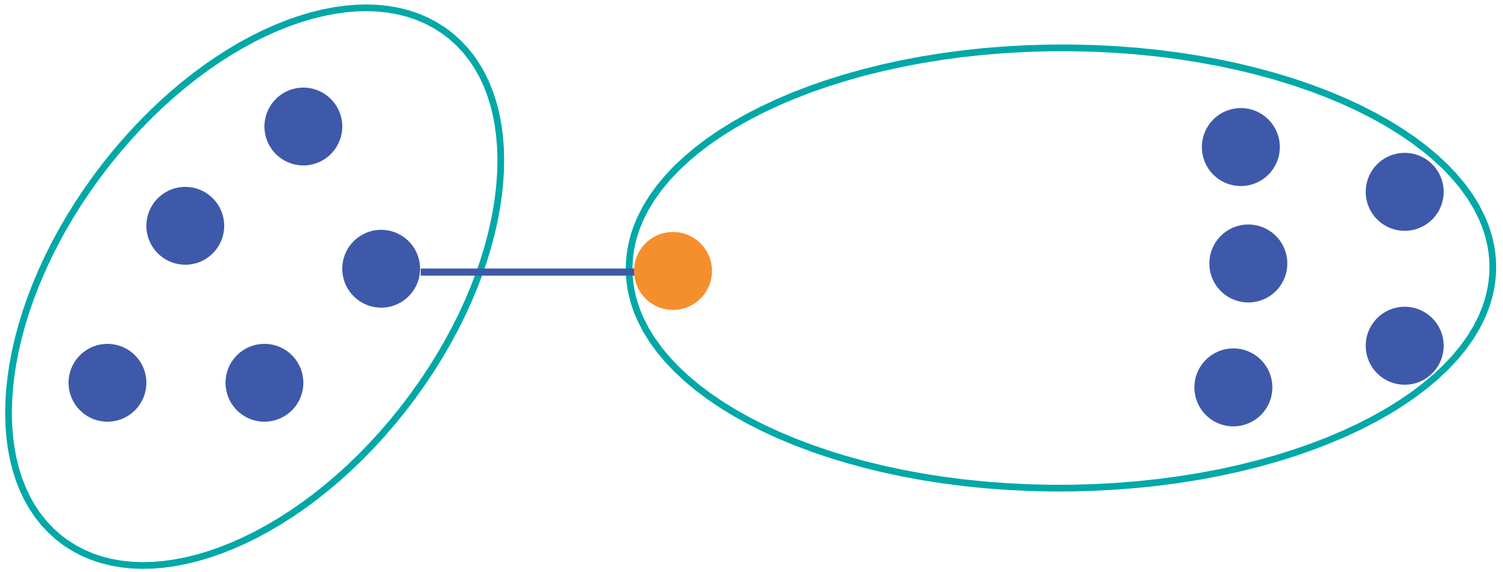
codefinity

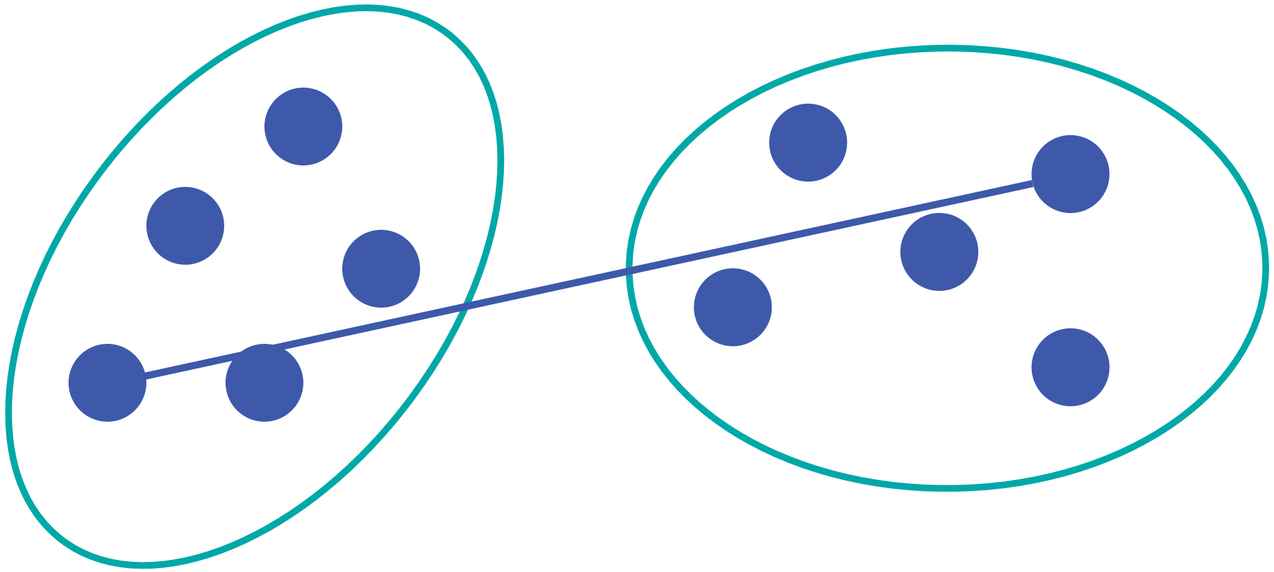
Single Linkage



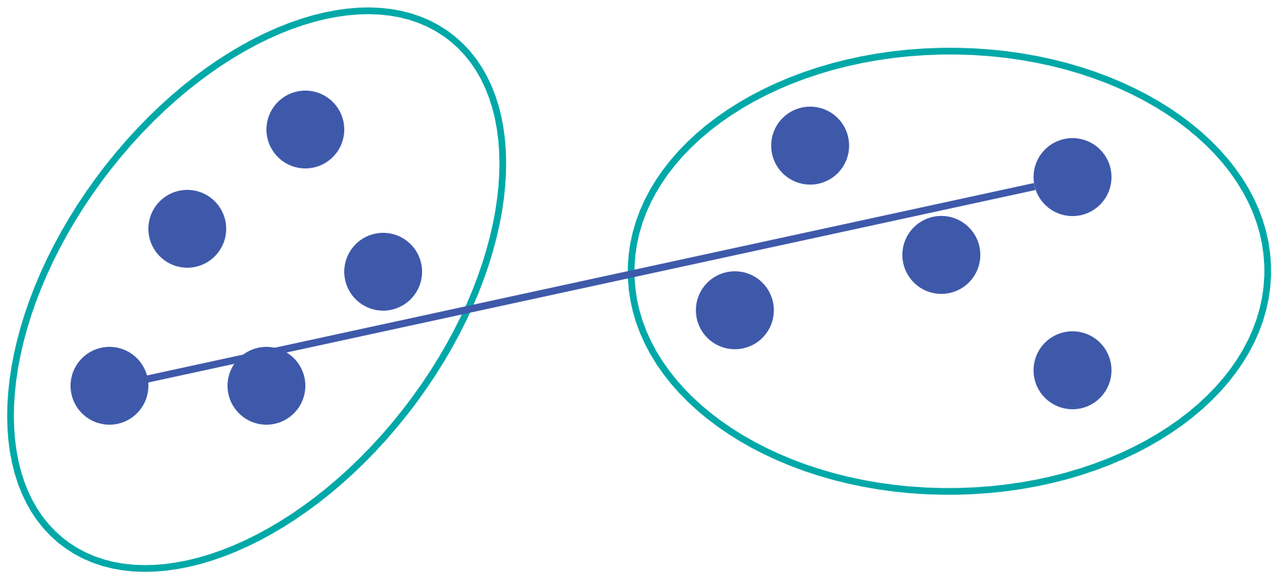
codefinity

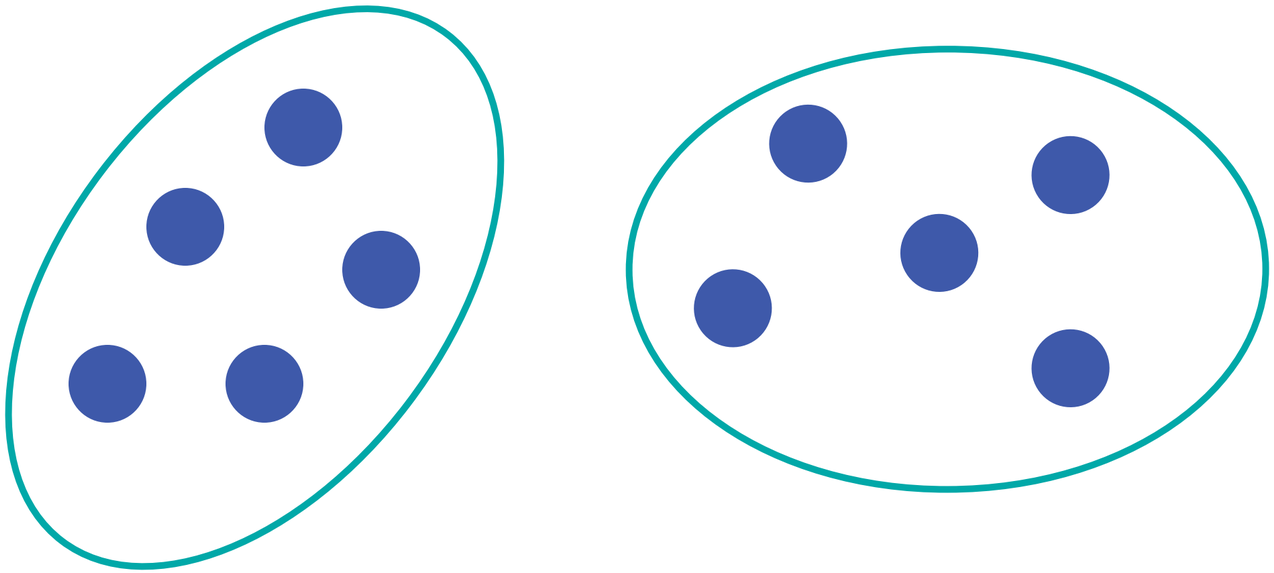
Single Linkage

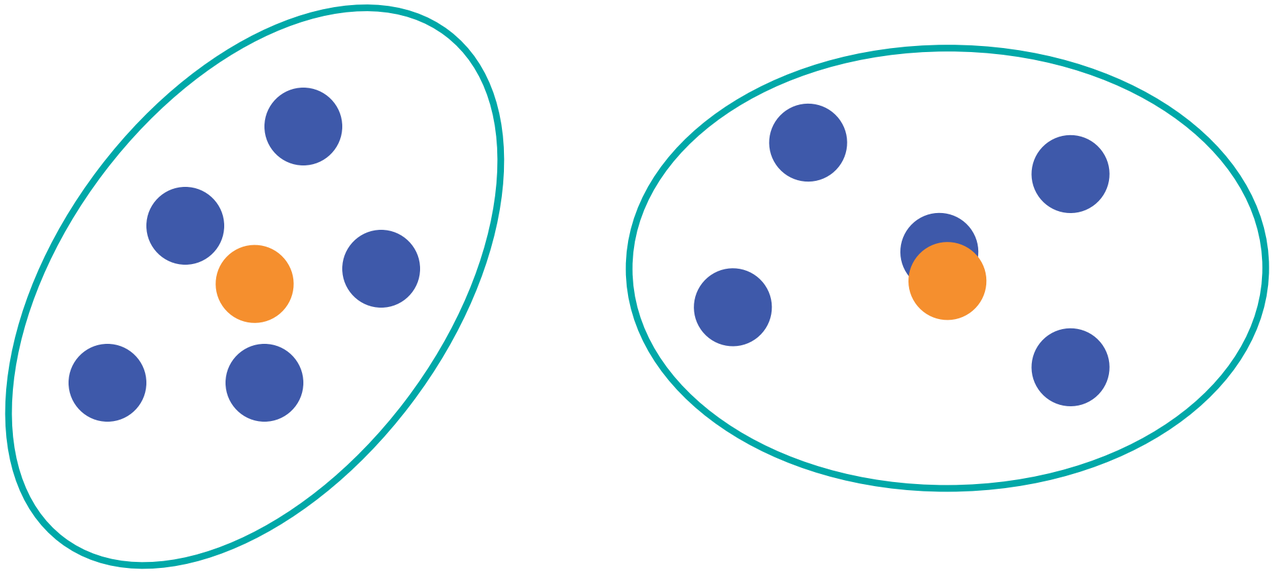




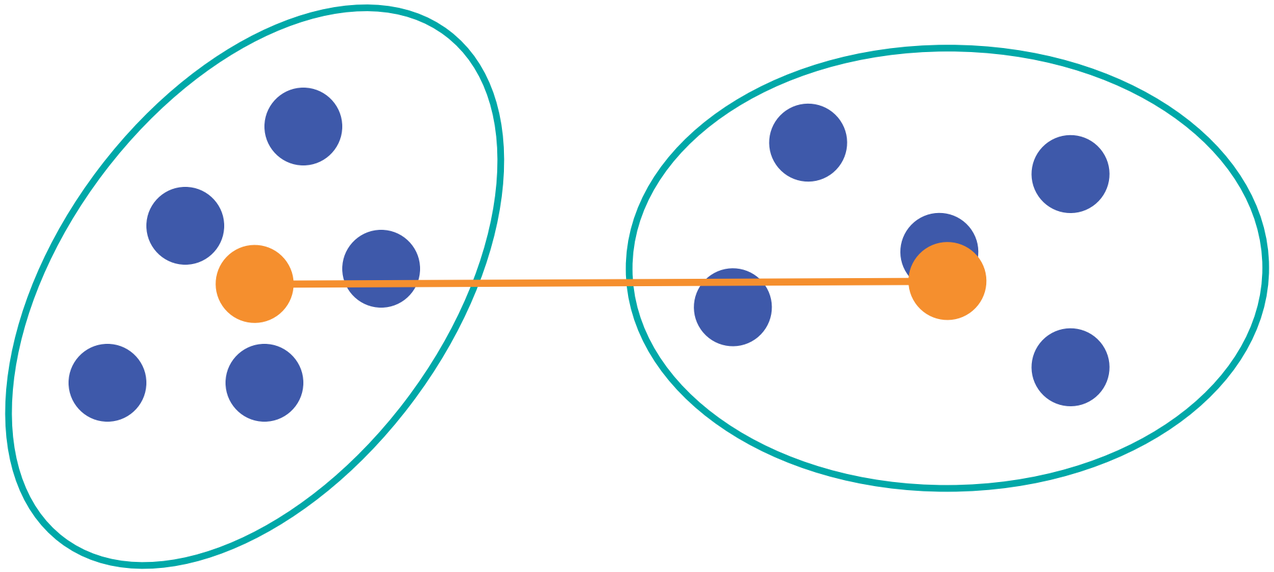
Complete Linkage

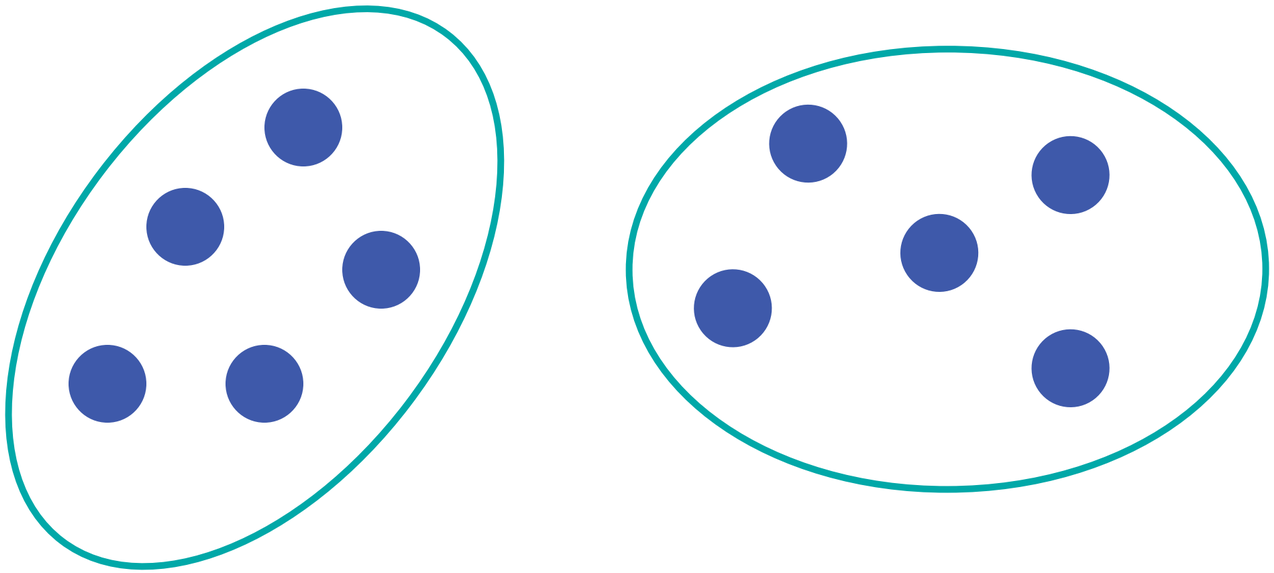




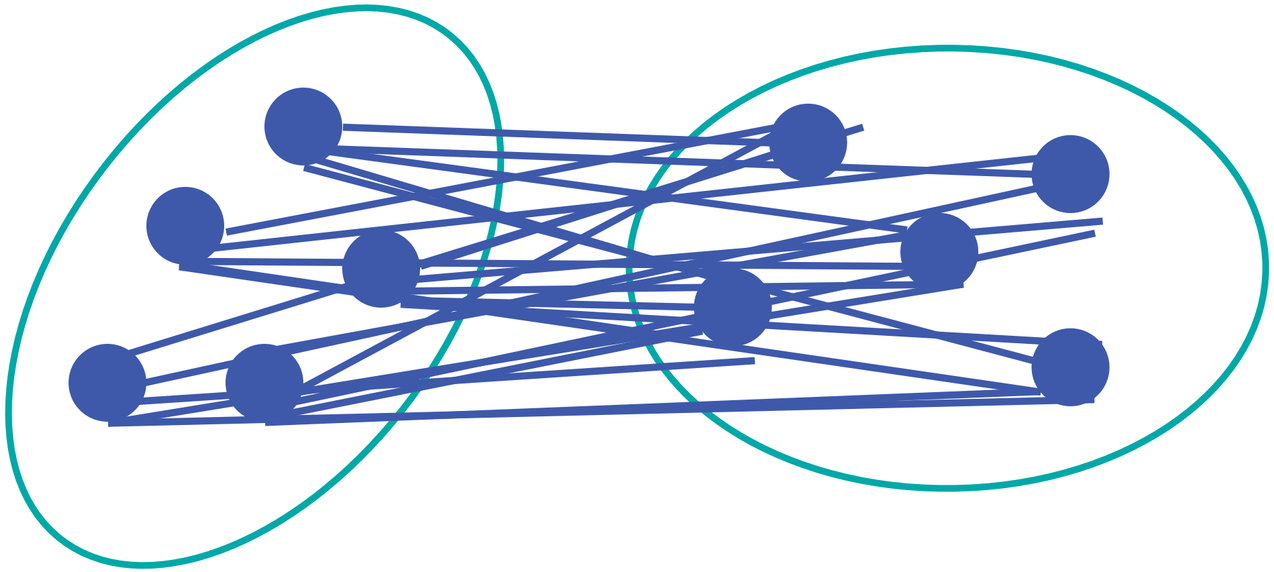


Centroid Linkage





codefinity



codefinity

Average Linkage