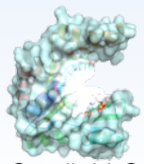
DataWknds.

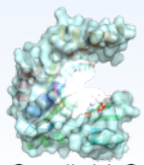
CLUSTERING



Catalit LLC

CLUSTERING

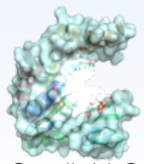
	CONTINUOUS	CATEGORICAL
SUPERVISED	?	?
UNSUPERVISED	?	?



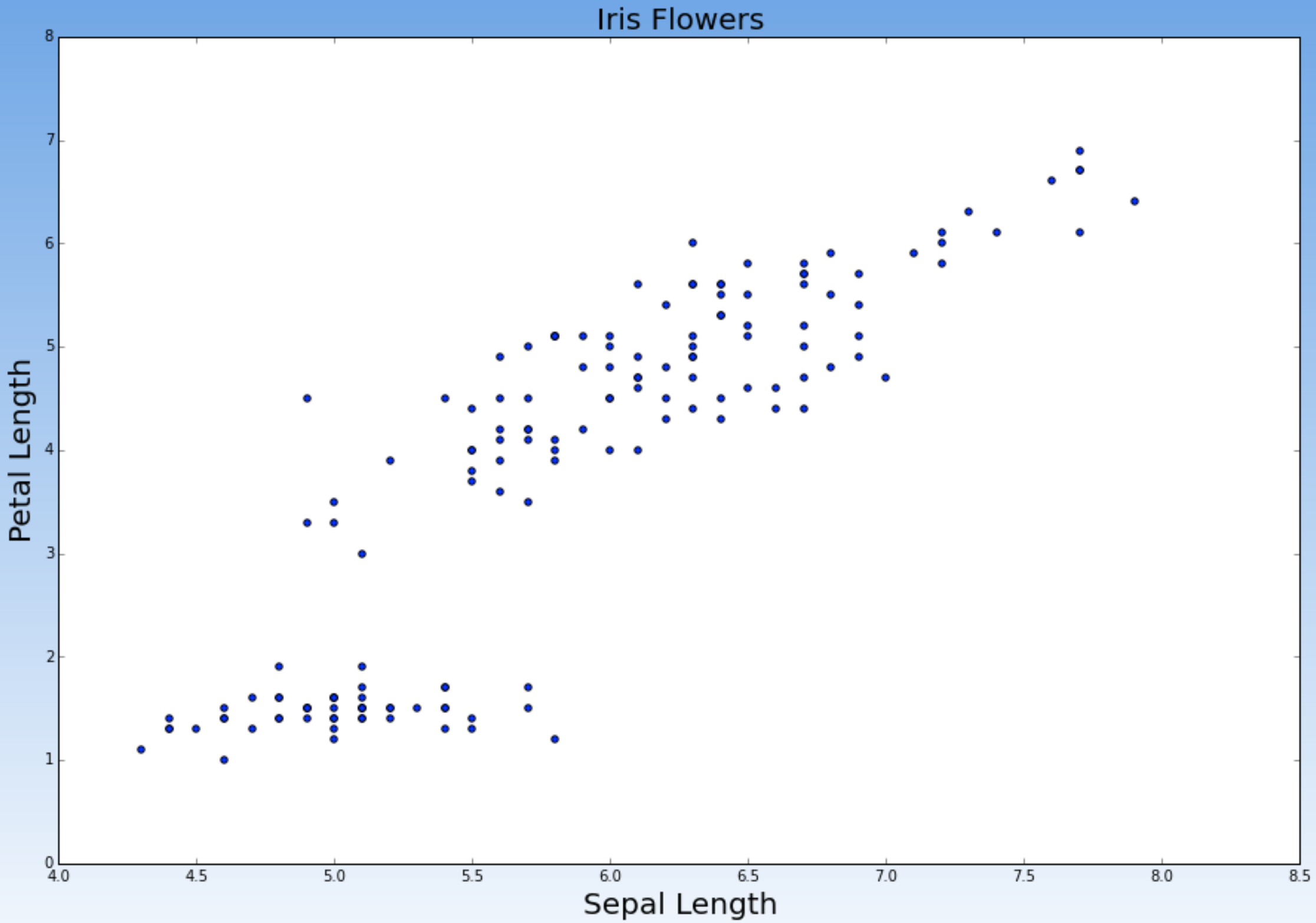
Catalit LLC

CLUSTERING

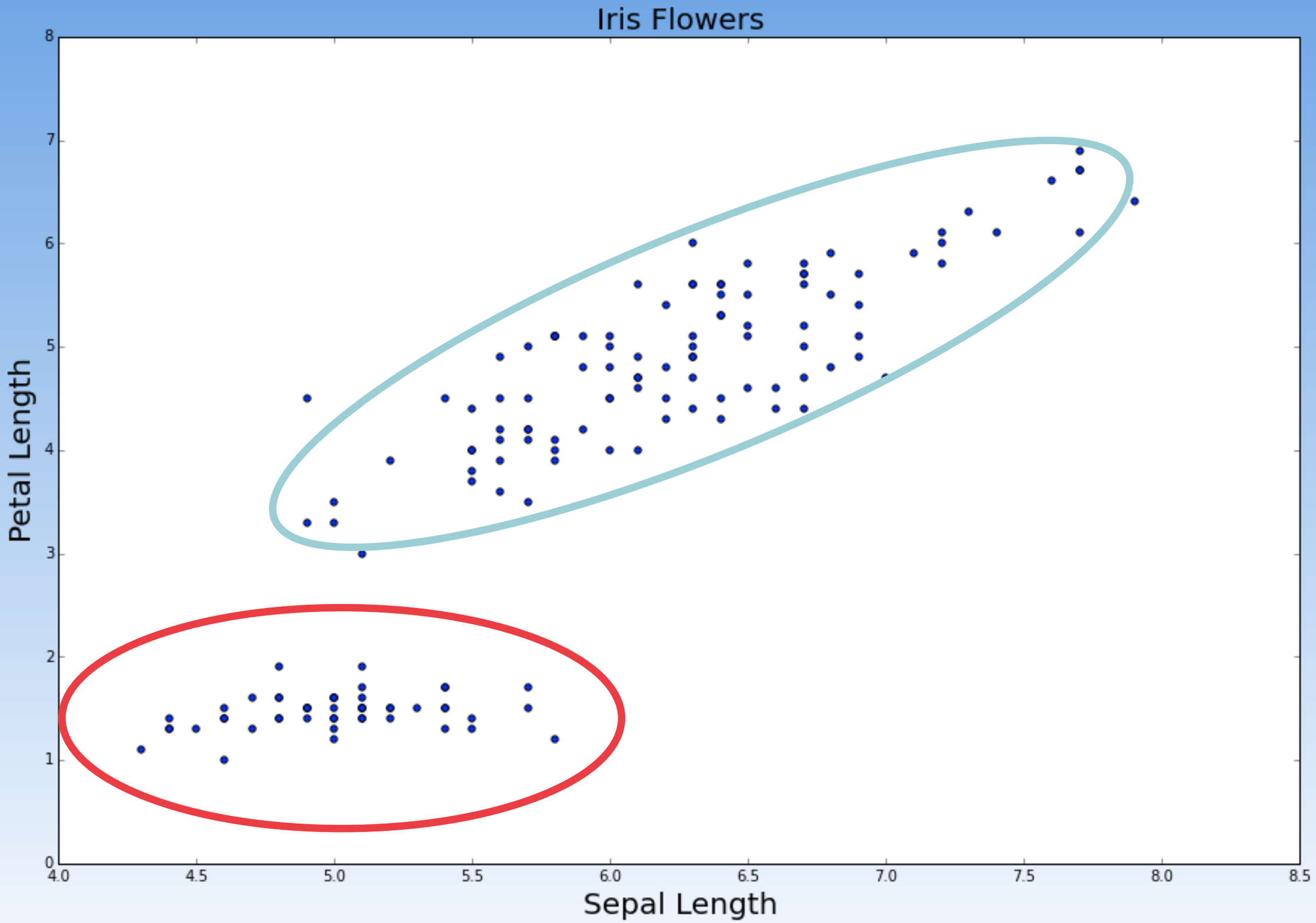
	CONTINUOUS	CATEGORICAL
SUPERVISED	REGRESSION	CLASSIFICATION
UNSUPERVISED	DIMENSION REDUCTION	CLUSTERING



Catalit LLC


Iris Flowers
Petal Length
Sepal Length
0
1
2
3
4
5
6
7
8
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

Catalit LLC


Iris Flowers
Petal Length
Sepal Length
0
1
2
3
4
5
6
7
8
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

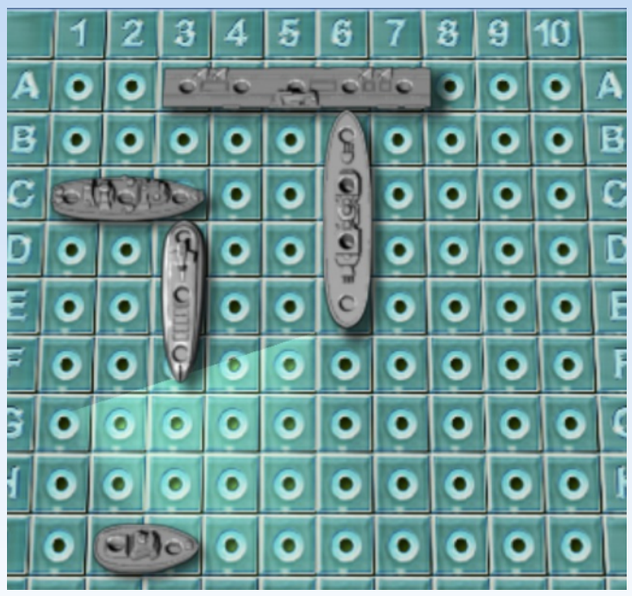
Catalit LLC

GOAL OF CLUSTERING

- GOAL: find data points that are similar to one another

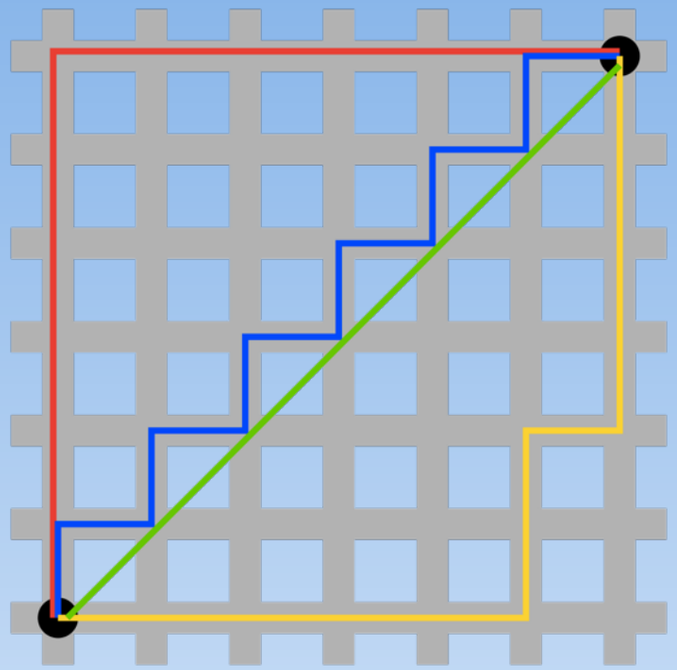
DISTANCE & SIMILARITY

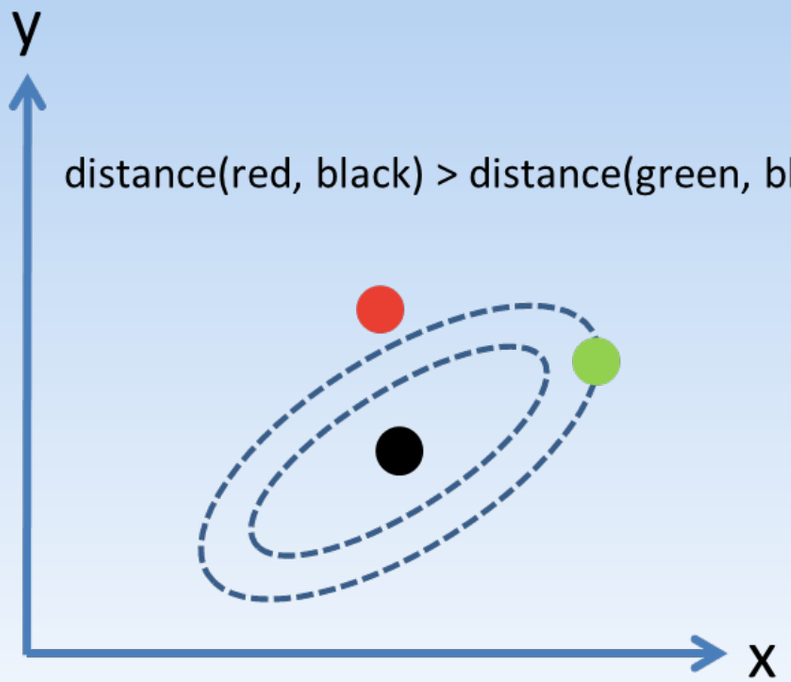
	Age	Gender	Annual Salary	Months in residence	Months in job	Current Debt
Client 1	23	M	$30,000	36	12	$5,000
Client 2	30	F	$45,000	12	12	$1,000
Client 3	19	M	$15,000	3	1	$10,000

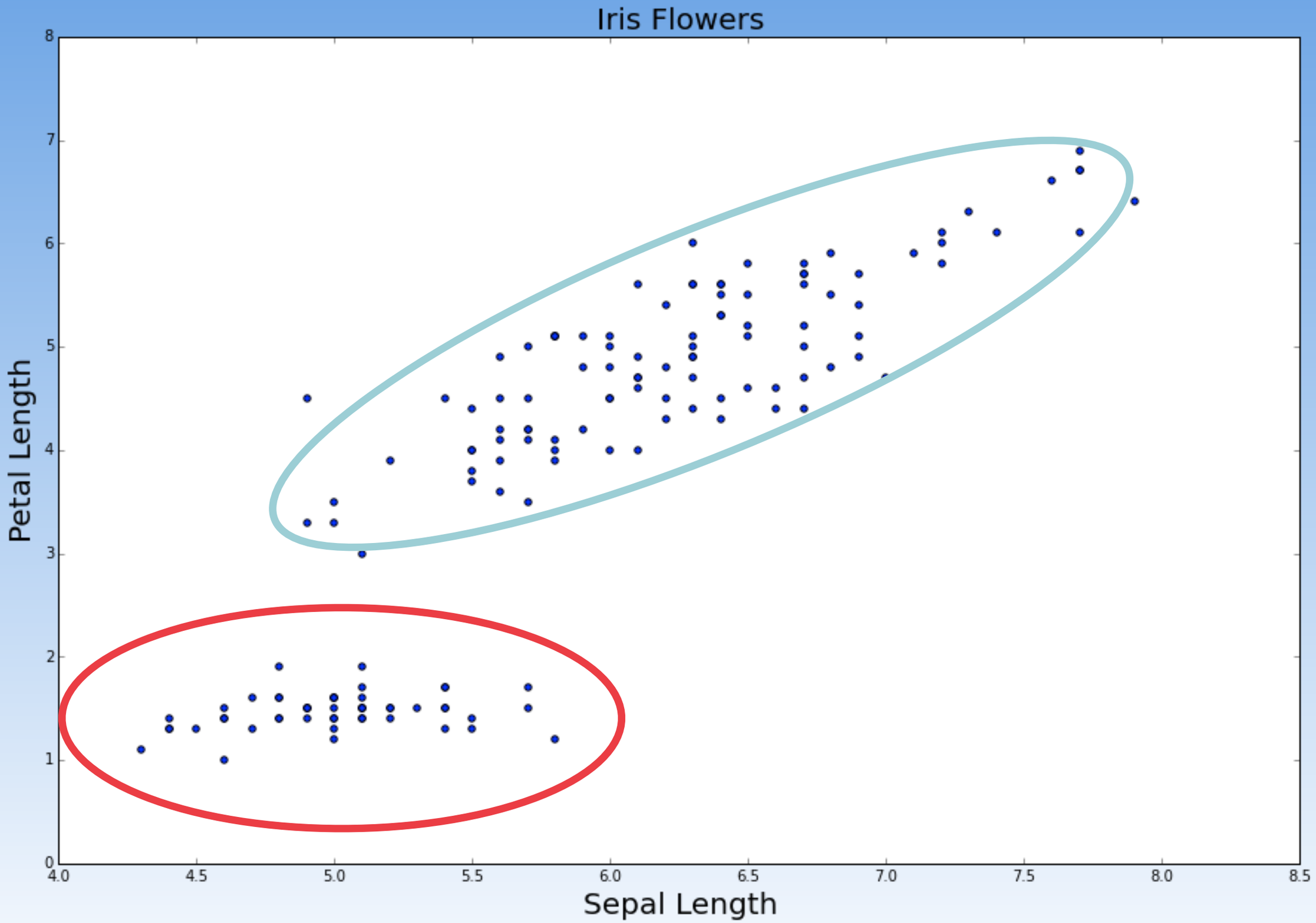


Catalit LLC

DISTANCE & SIMILARITY





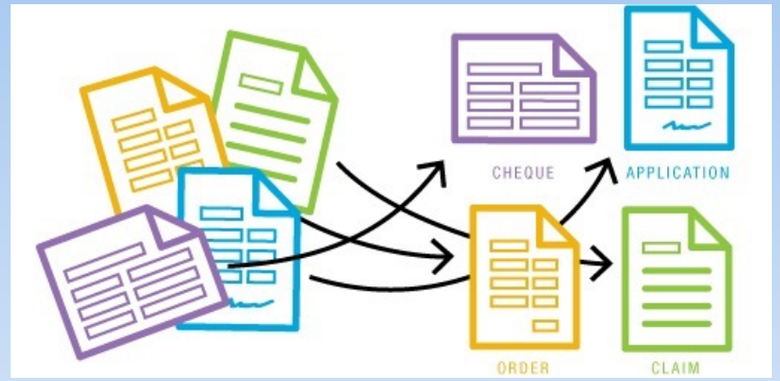

Iris Flowers
Petal Length
Sepal Length
0
1
2
3
4
5
6
7
8
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

Catalit LLC

APPLICATIONS

- Malware detection

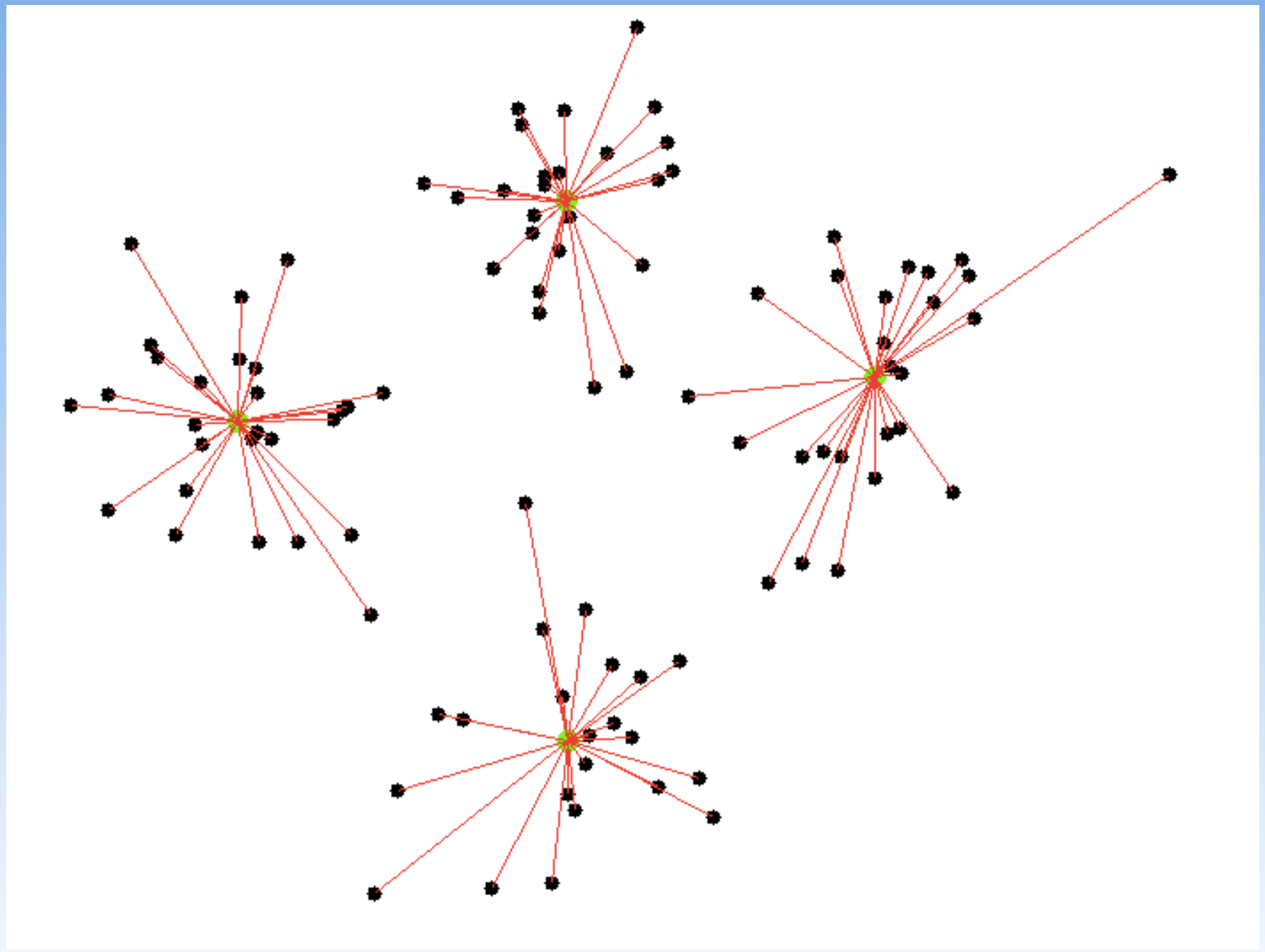
- Intrusion detection



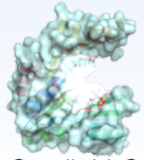
- Document clustering
- Product placing
- Customer segmentation

- Gene discovery

K-MEANS

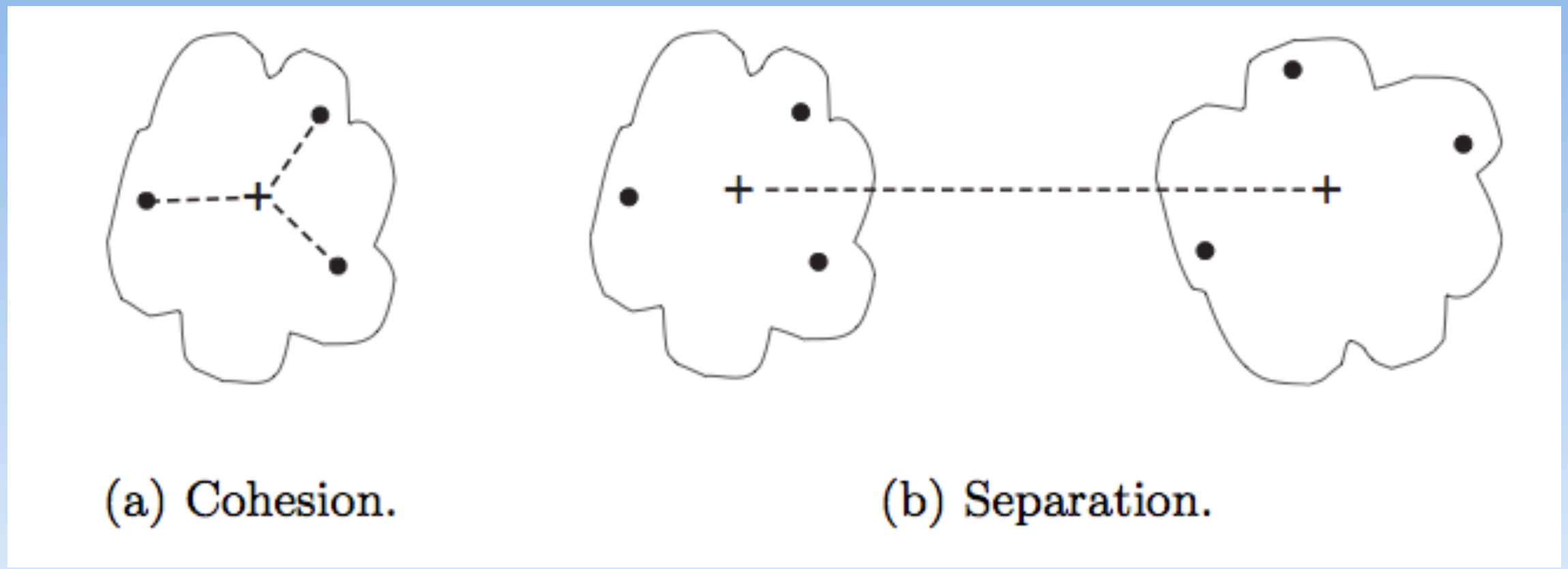


http://shabal.in/visuals/kmeans/1.html



Catalit LLC

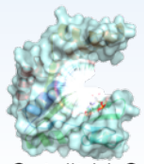
CLUSTER VALIDATION



(a) Cohesion.

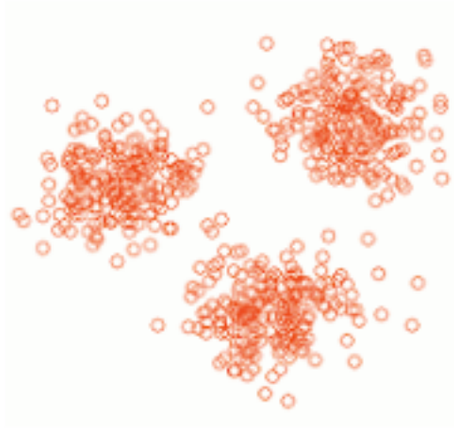
(b) Separation.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

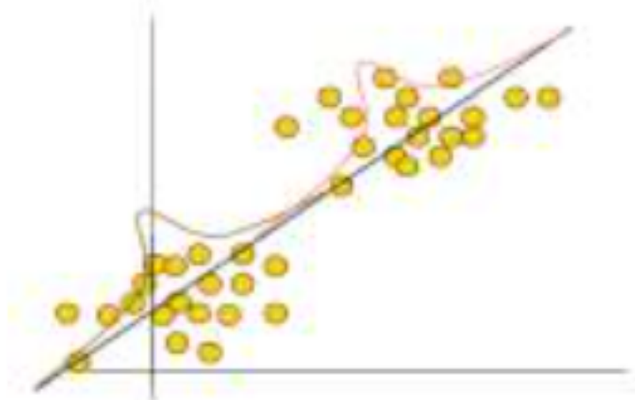


METHODS FOR CLUSTERING

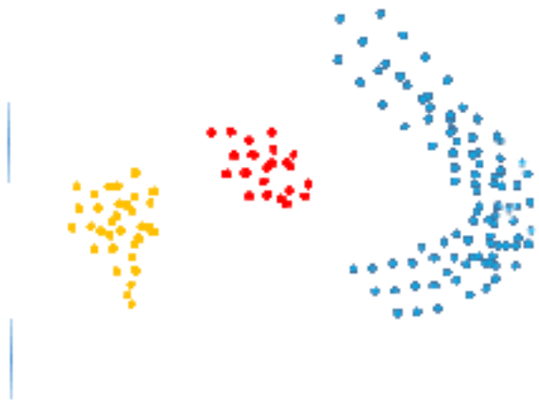
- K-means



> Expectation Maximization

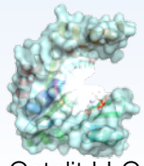


- Density-based



> Hierarchical

LAB: K-MEANS



Catalit LLC