Creative Data Mining (FS 2017)

Introduction to Machine Learning

Discrete data	Classification	Clustering
Continuous data	Regression	Clustering and dimensionality reduction
	Supervised learning Labeled data	Unsupervised learning Unlabeled data

Instructor: Dr. Varun OJHA

Course Instructor:

Dr. Dani Zünd

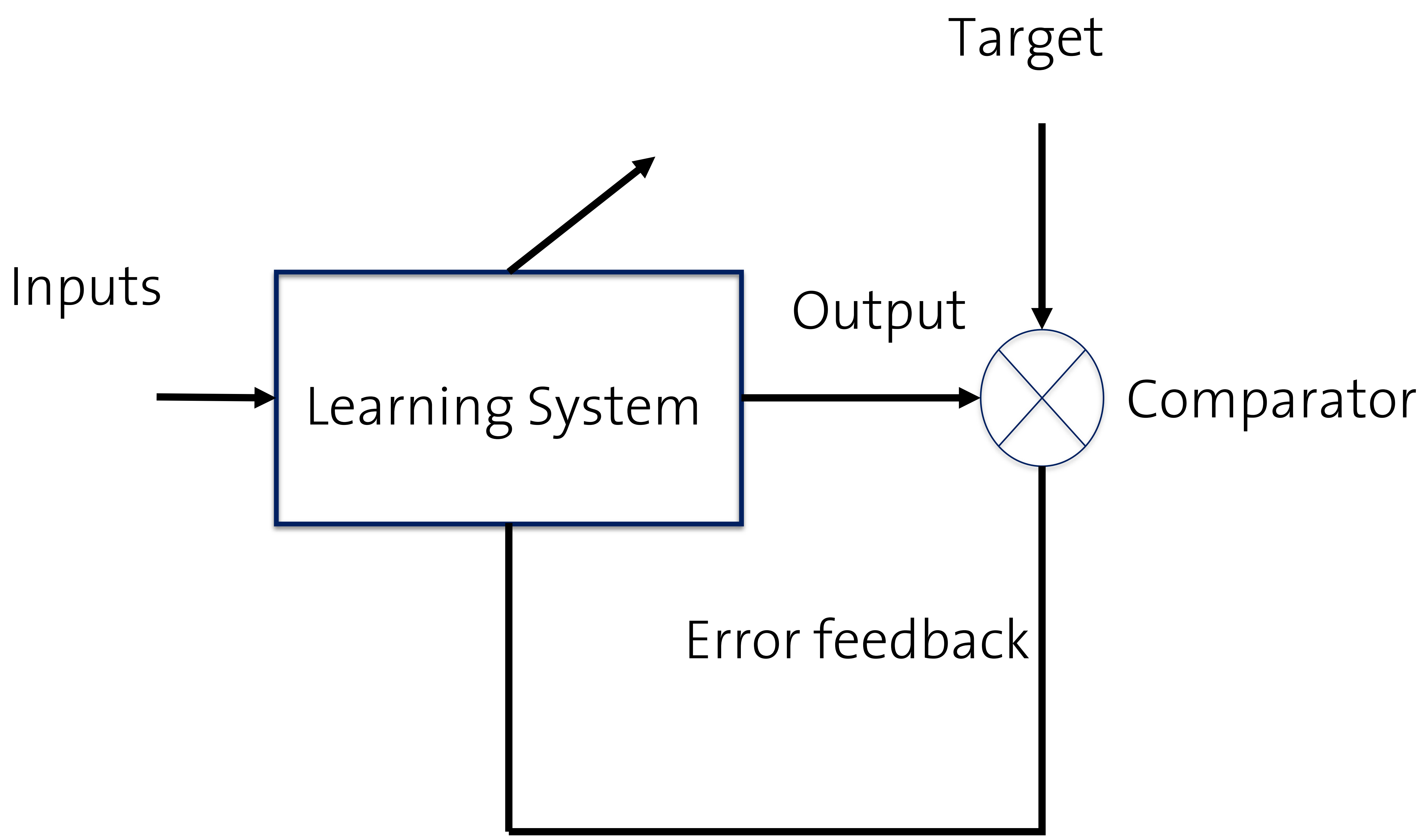
Danielle Griego

Artem Chirkin

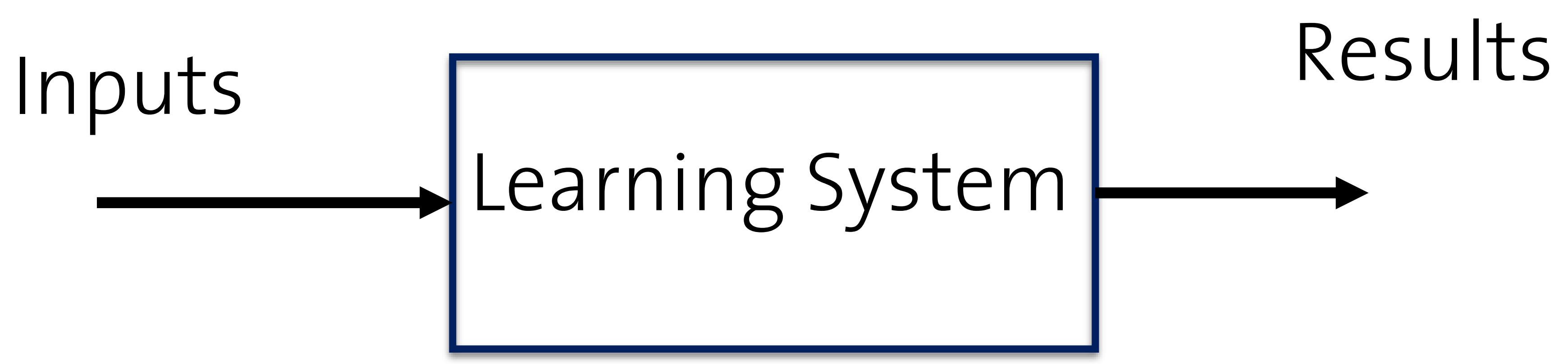
ETH zürich DARCH iA Chair of Information Architecture

Learning Systems

Supervised



Unsupervised



Data (unlabeled)

#	Inputs			
	HR	BVP	EDA	TempBF
1	76.85	17.27	0.22	34.63
2	76.97	19.54	0.22	34.64
3	77.10	18.51	0.22	34.64
4	85.28	46.09	0.22	34.61
5	85.42	35.83	0.22	34.61
6	88.02	2.59	0.22	34.63
7	77.25	6.34	0.22	34.65
8	77.49	6.98	0.22	34.63
9	85.81	12.18	0.22	34.61

We do not know what is target/output.

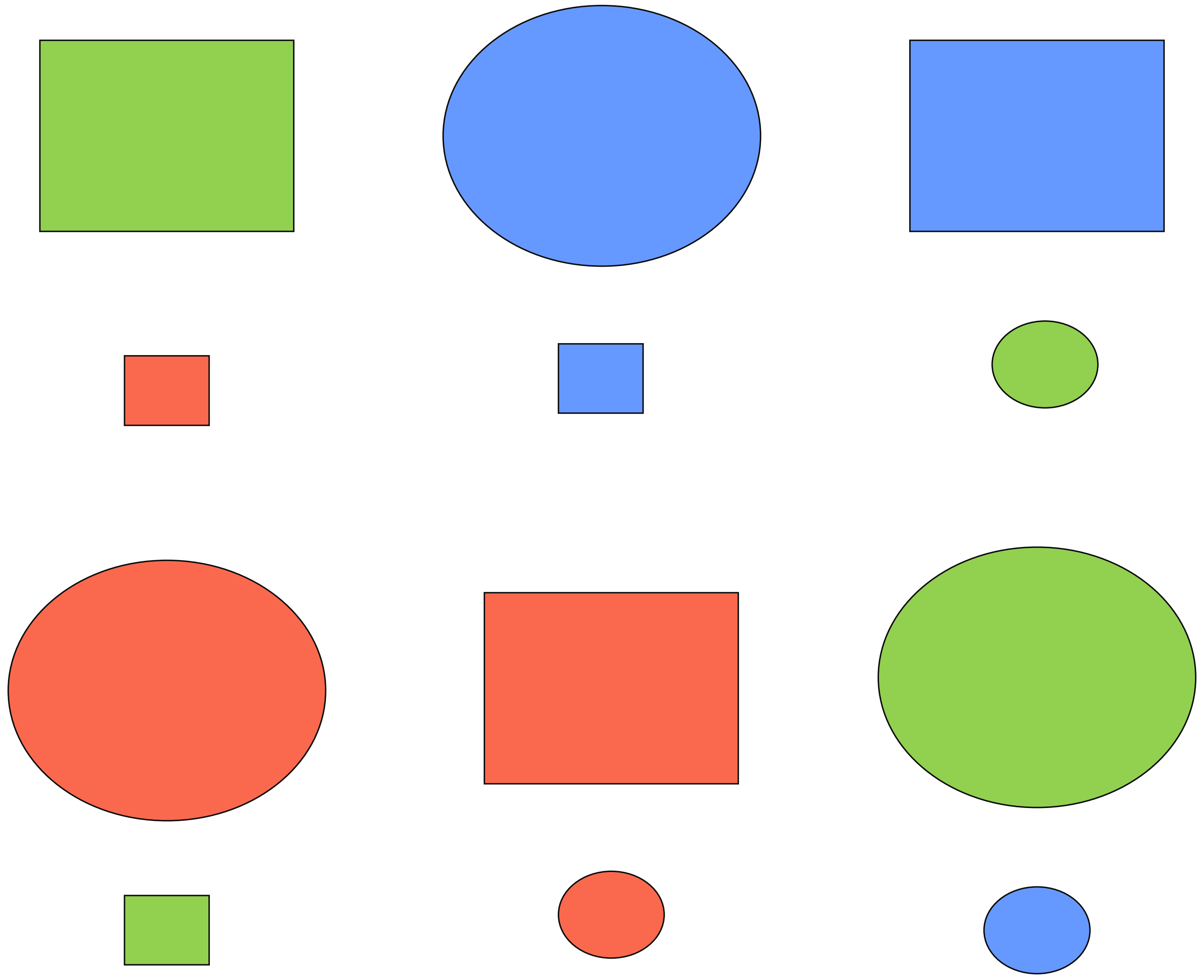
We just want to understand some pattern or some sort of grouping of data points.

Data (unlabeled)

Example: Natural Concept

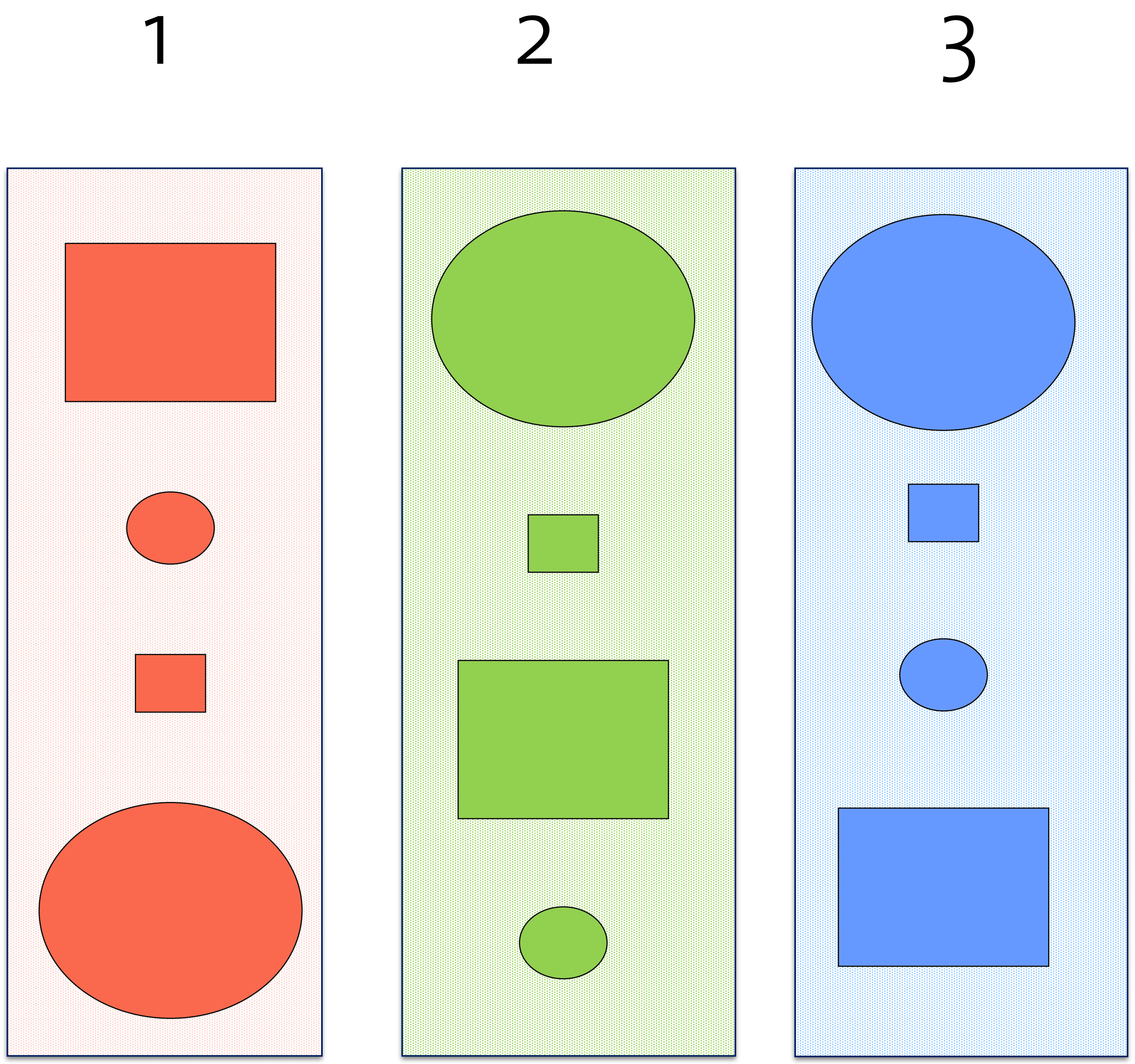
Can we form groups of these objects?

What features these objects has?



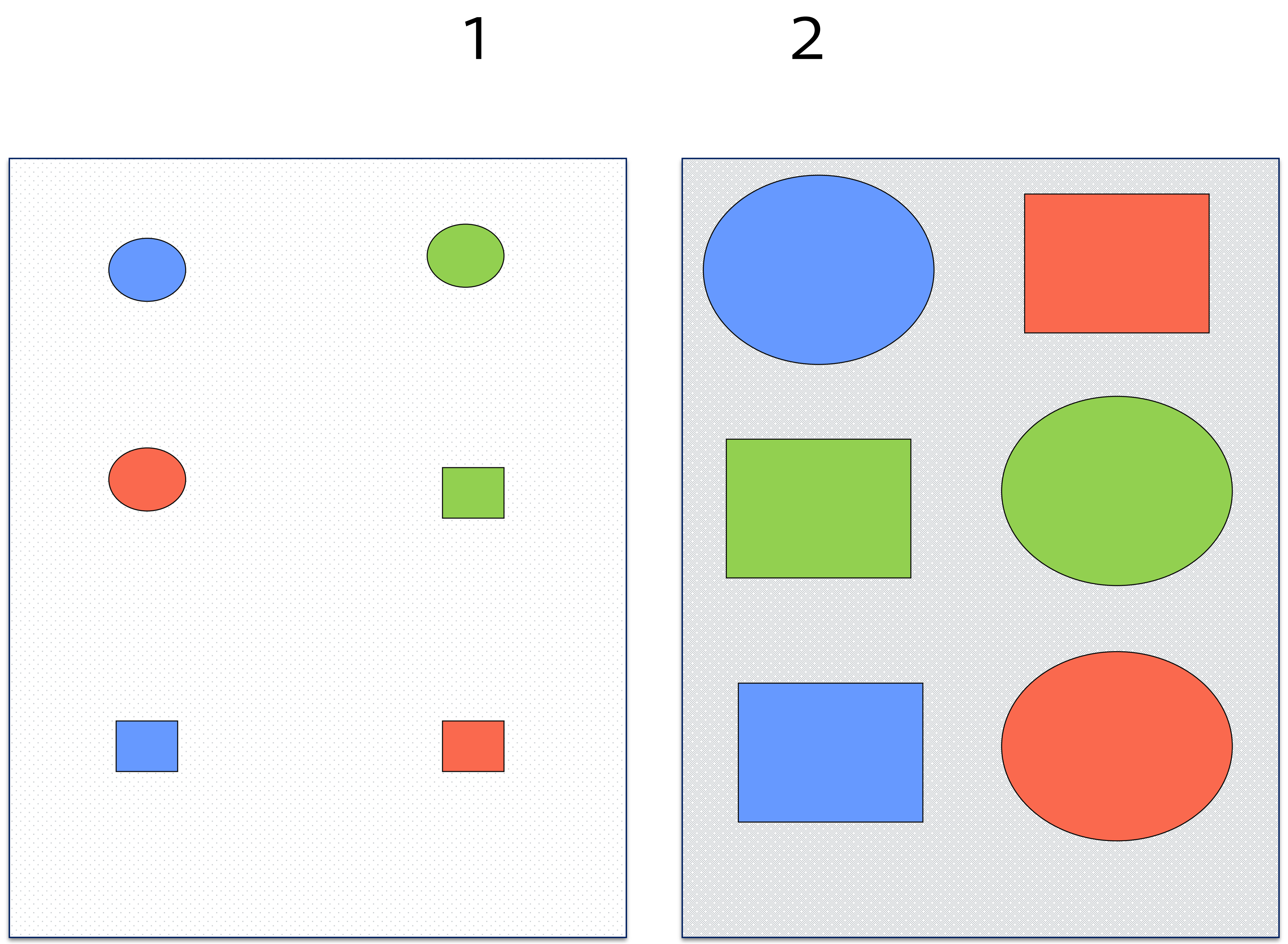
Clustering

Criteria: Color



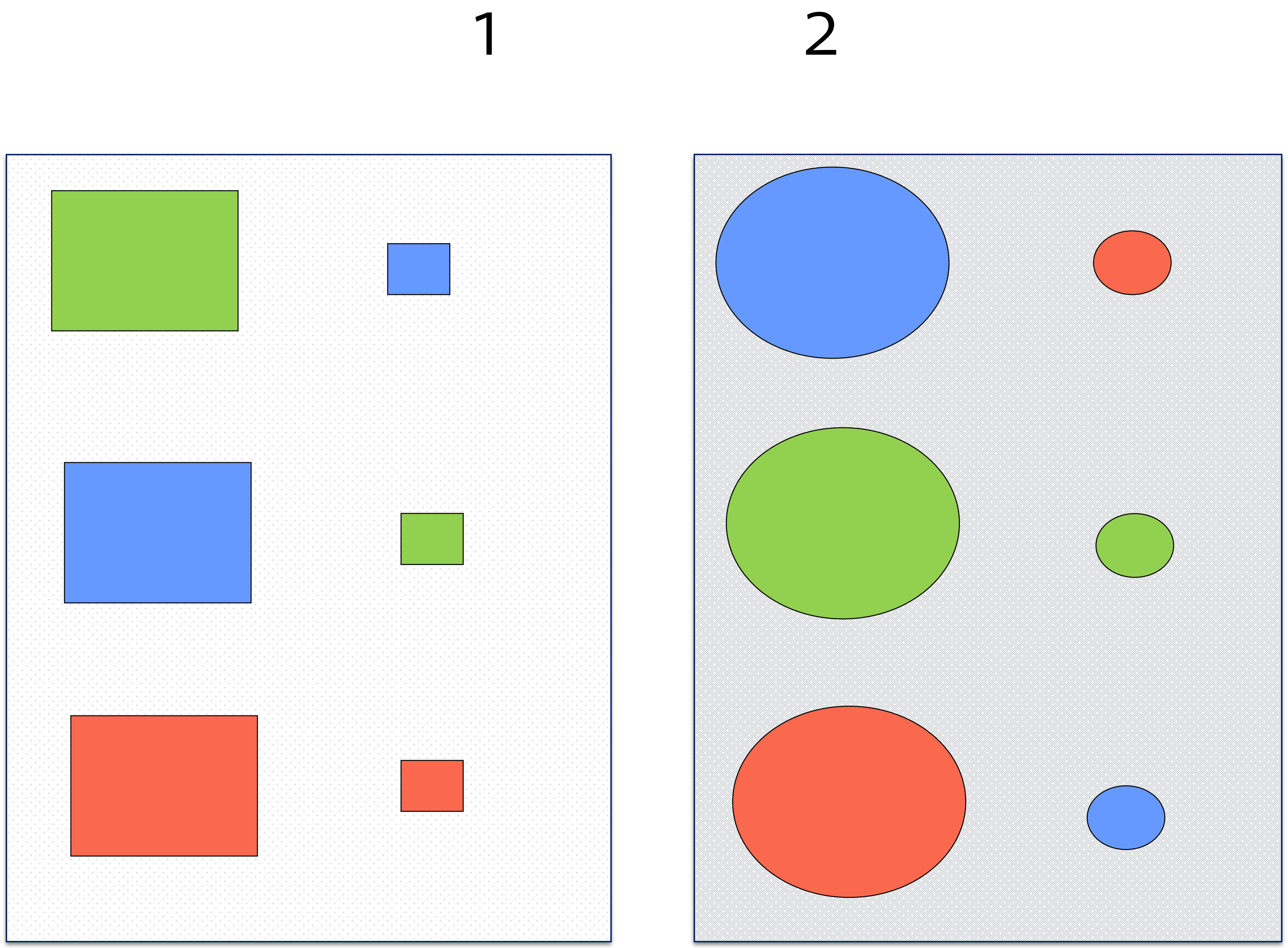
Clustering

Criteria: Size



Clustering

Criteria: Shape



ESUM Data (unlabeled)

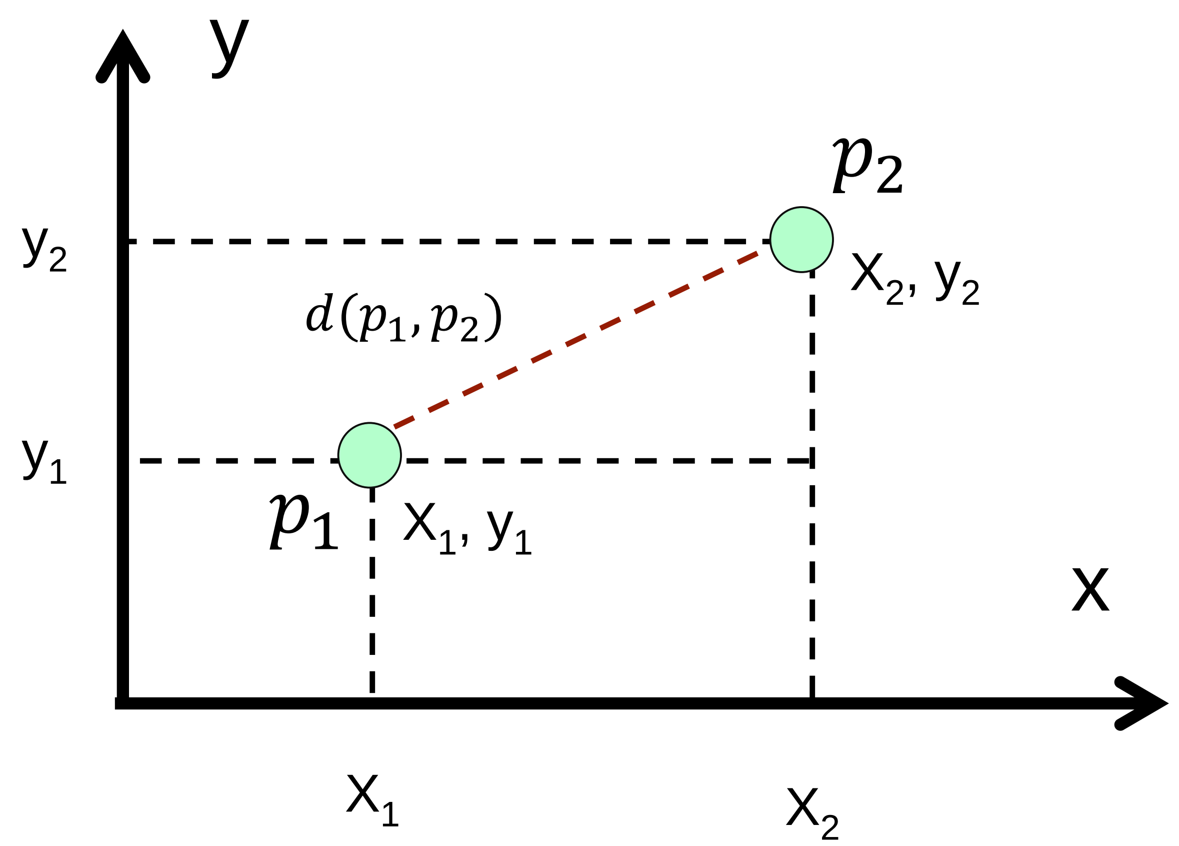
Can we form groups of these samples?

What are the criteria for grouping?

#	Inputs			
	HR	BVP	EDA	TempBF
1	76.85	17.27	0.22	34.63
2	76.97	19.54	0.22	34.64
3	77.10	18.51	0.22	34.64
4	85.28	46.09	0.22	34.61
5	85.42	35.83	0.22	34.61
6	88.02	2.59	0.22	34.63
7	77.25	6.34	0.22	34.65
8	77.49	6.98	0.22	34.63
9	85.81	12.18	0.22	34.61

Clustering

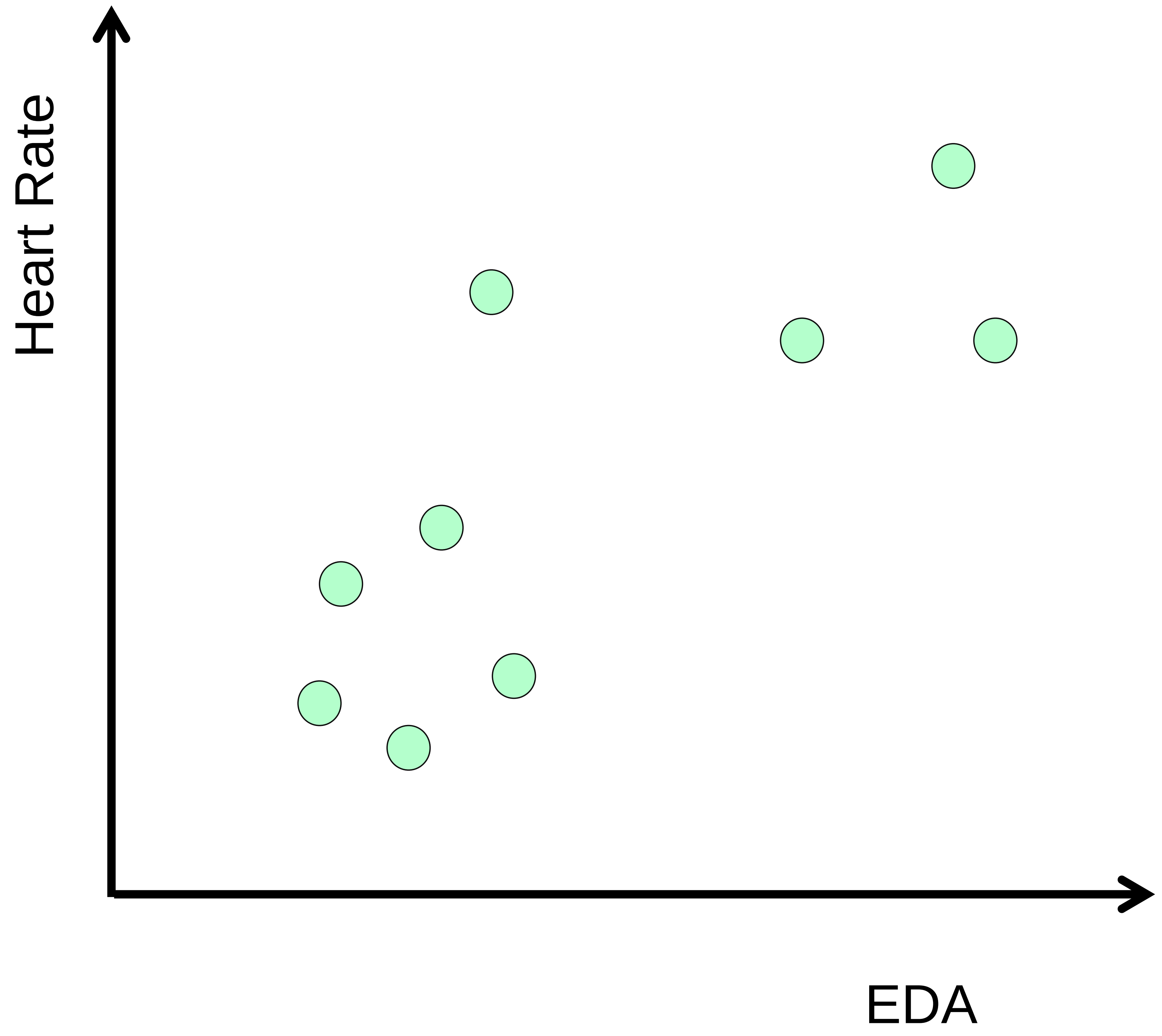
Distance based: K-Means



Euclidean distance:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

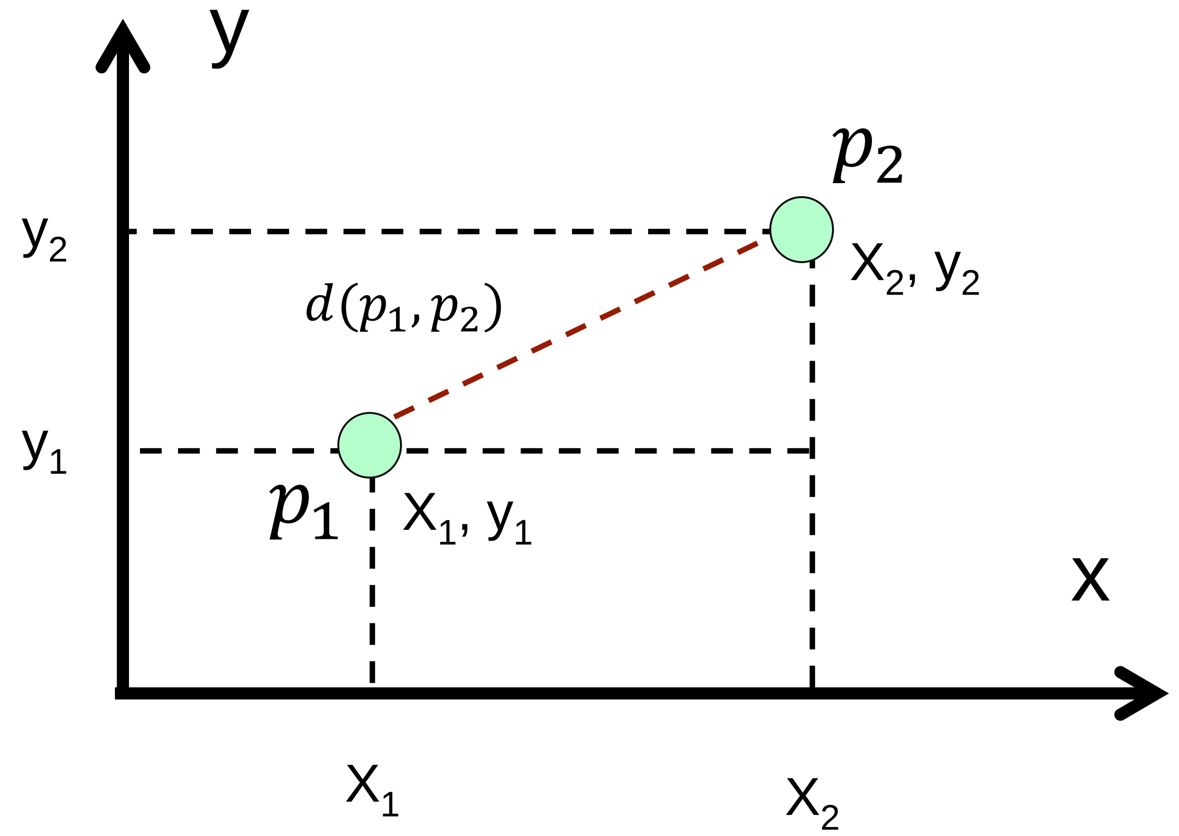
ESUM data



Clustering

K = 2

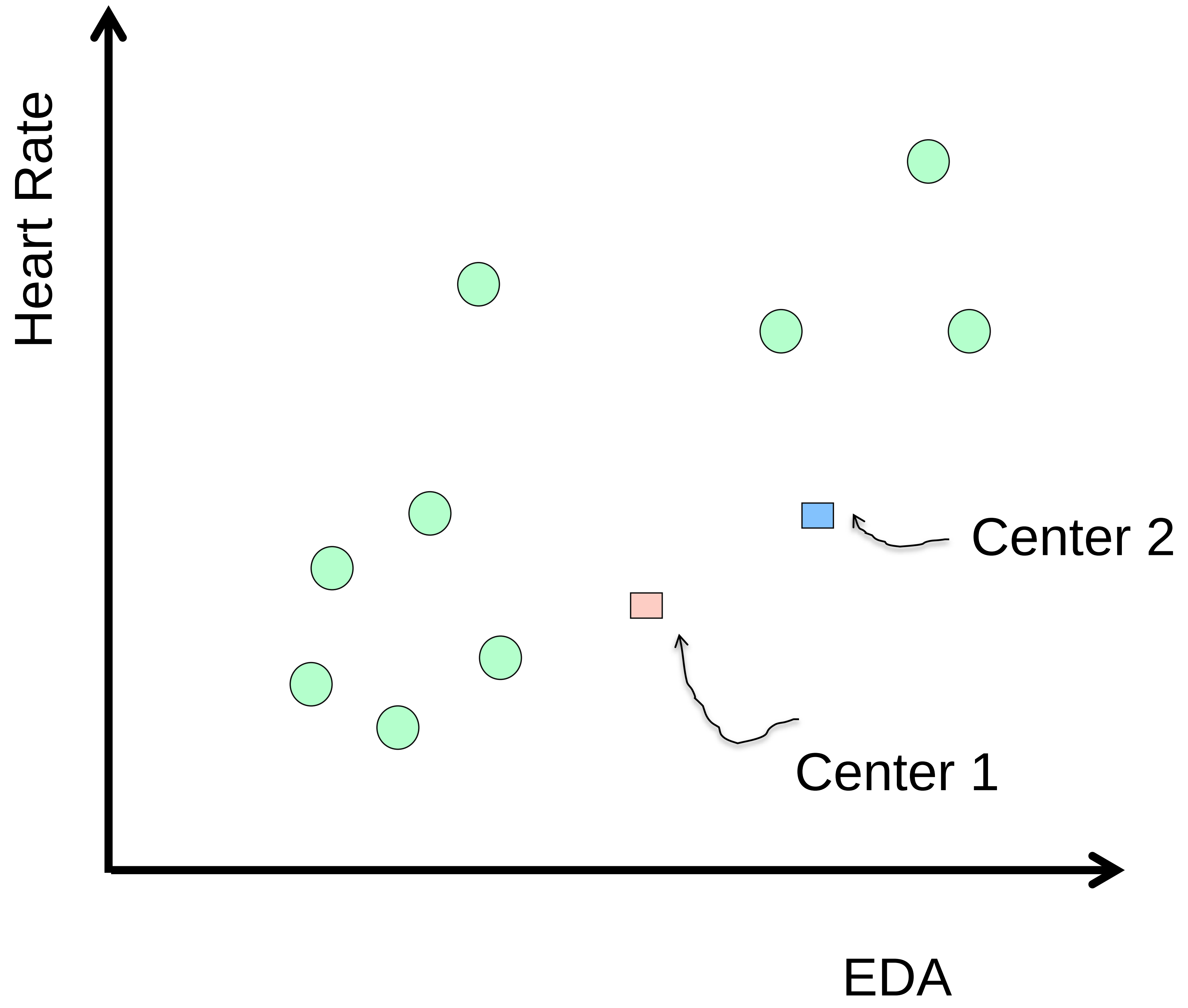
Distance based: K-Means



Euclidean distance:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

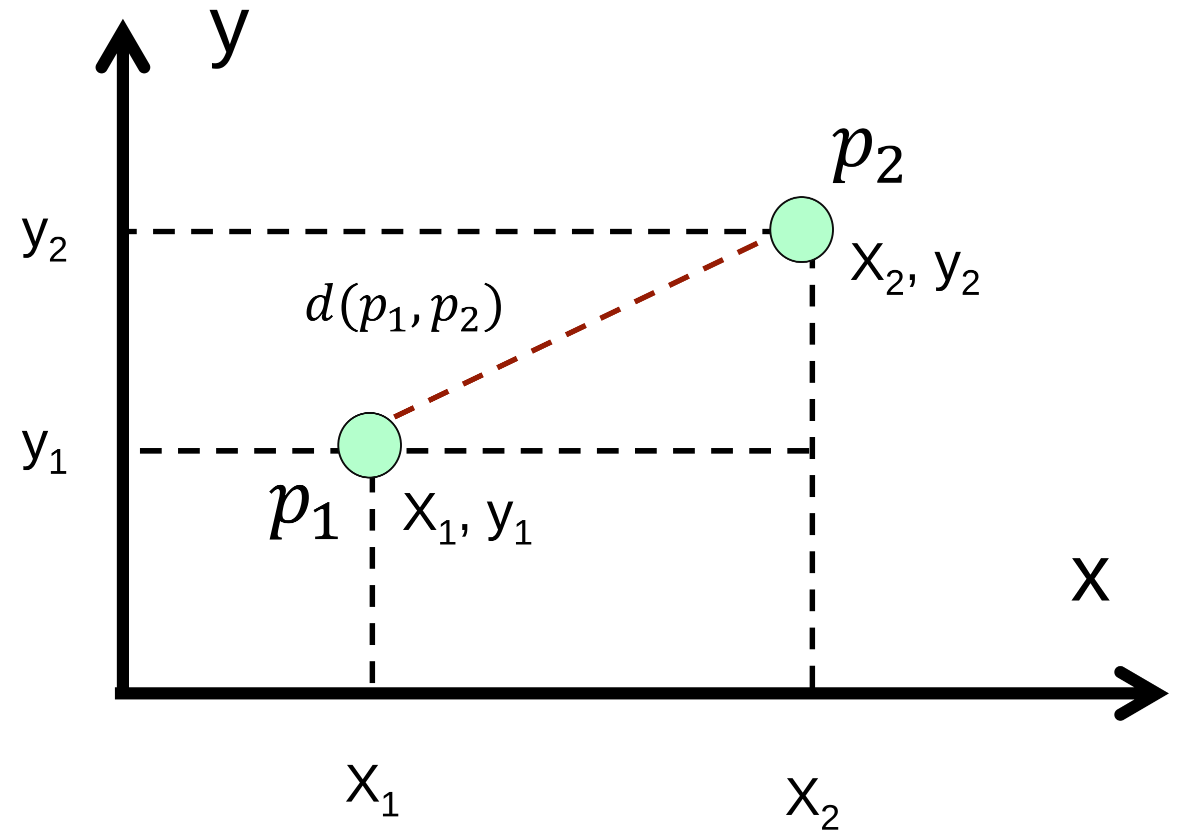
Defining centers



Clustering

Distance based: K-Means

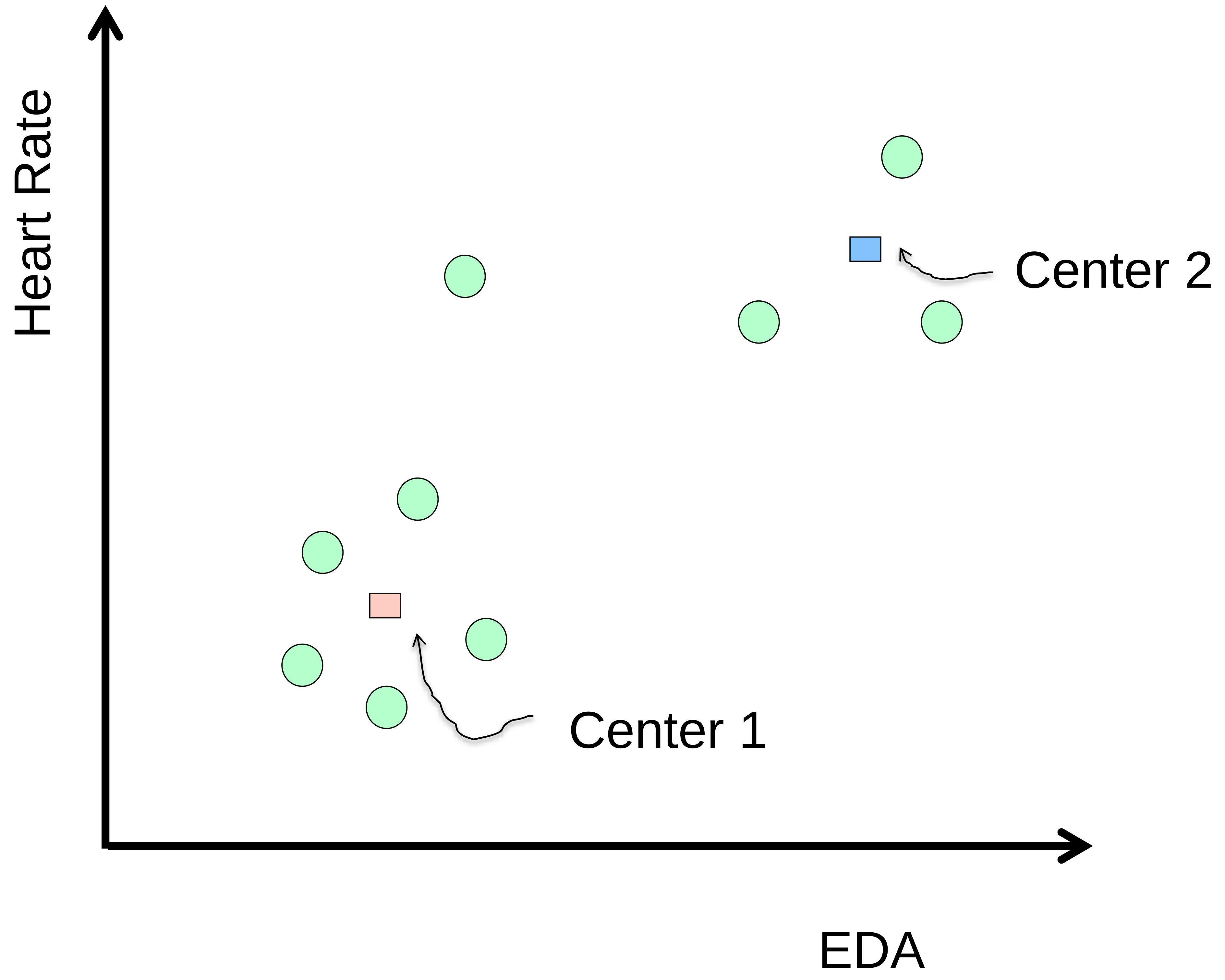
K = 2



Euclidean distance:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

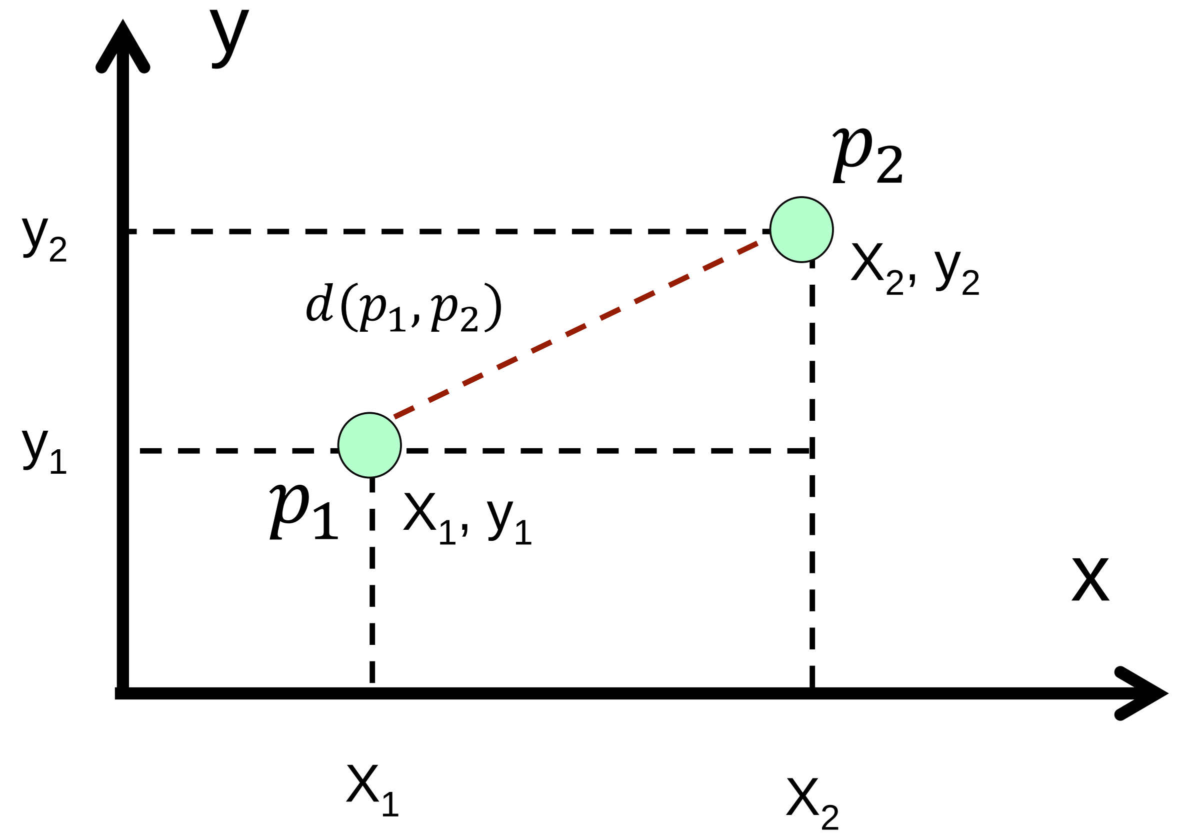
Results



Clustering

K = 2

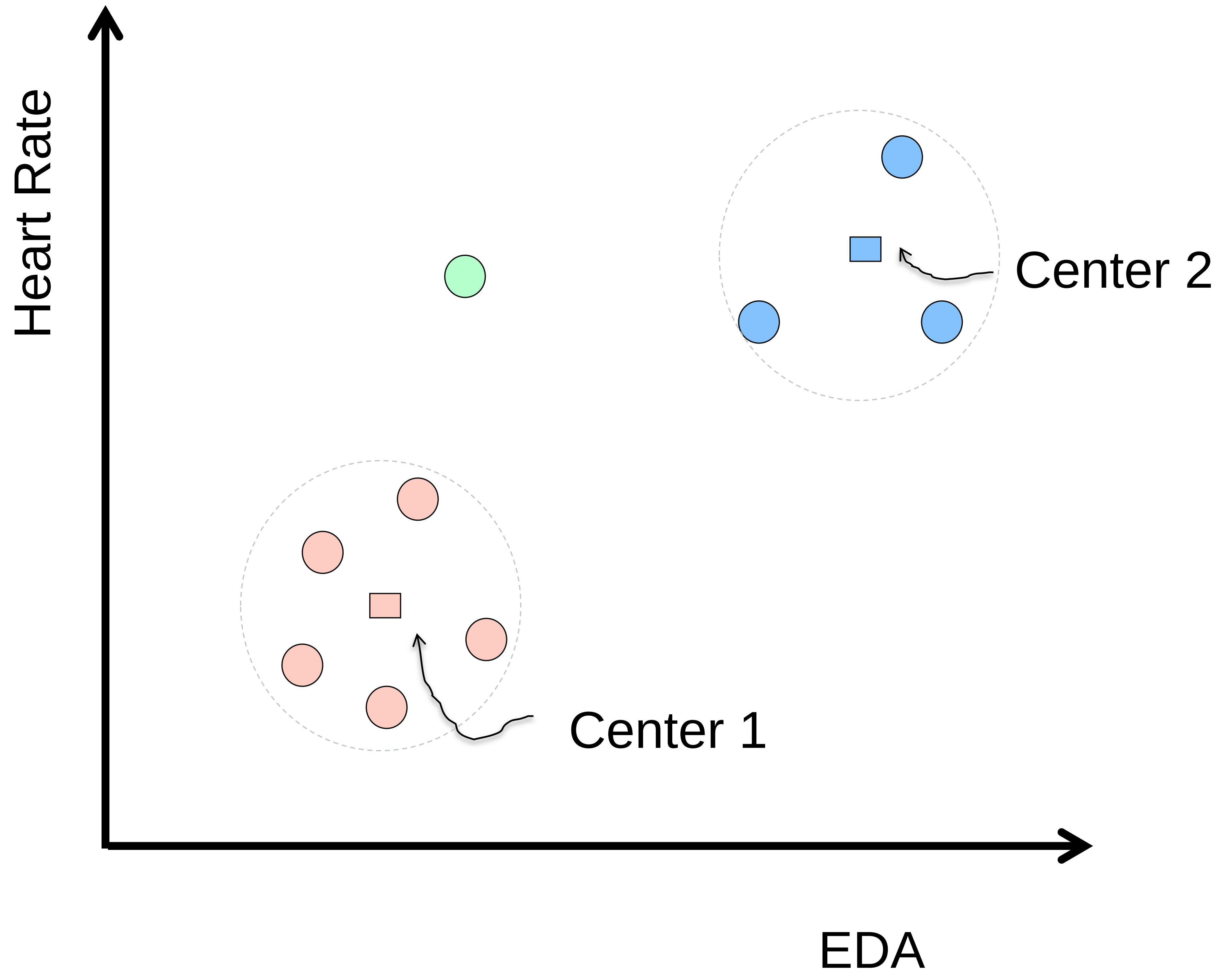
Distance based: K-Means



Euclidean distance:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Results

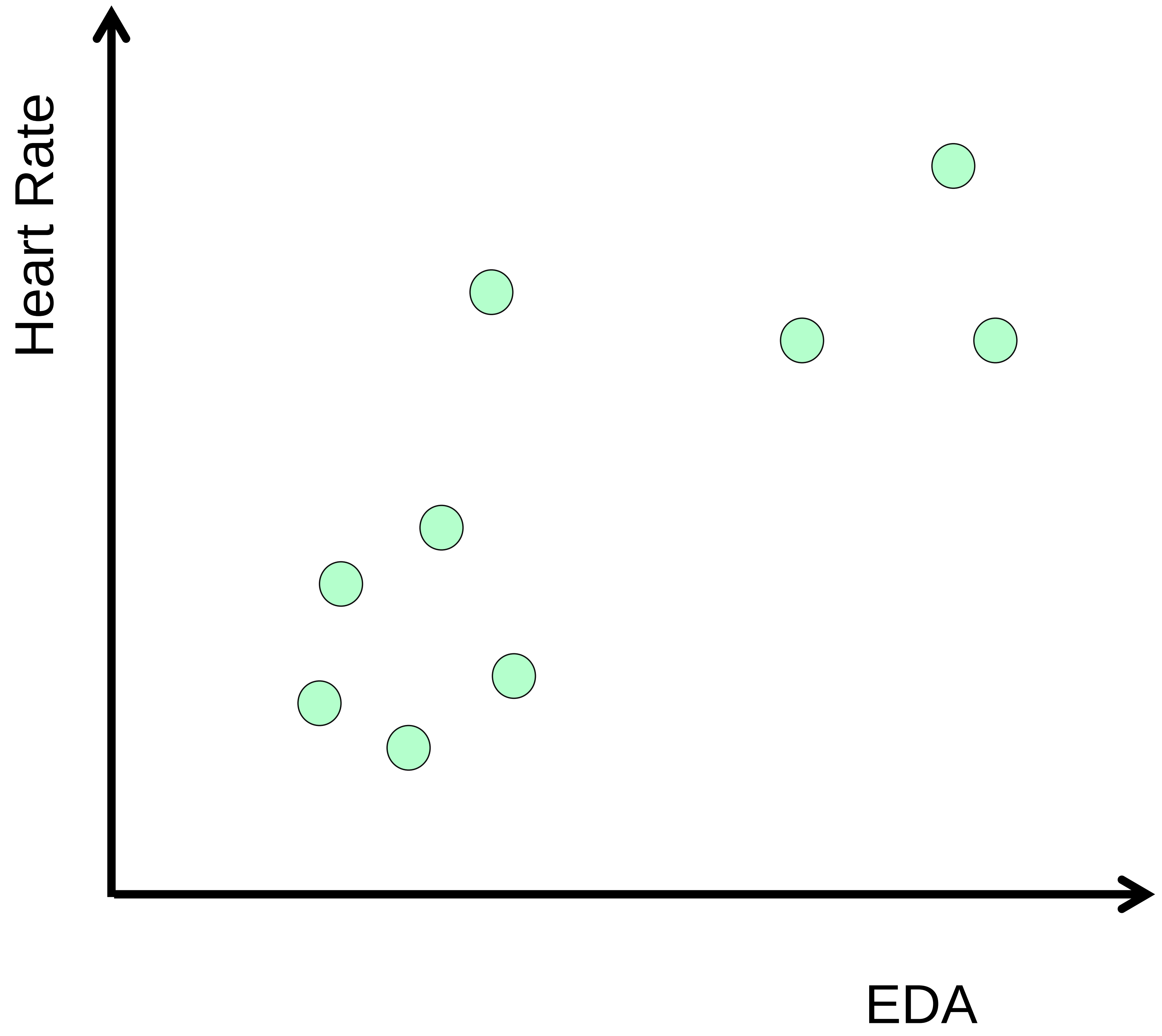


Clustering

Density based : DBSCAN

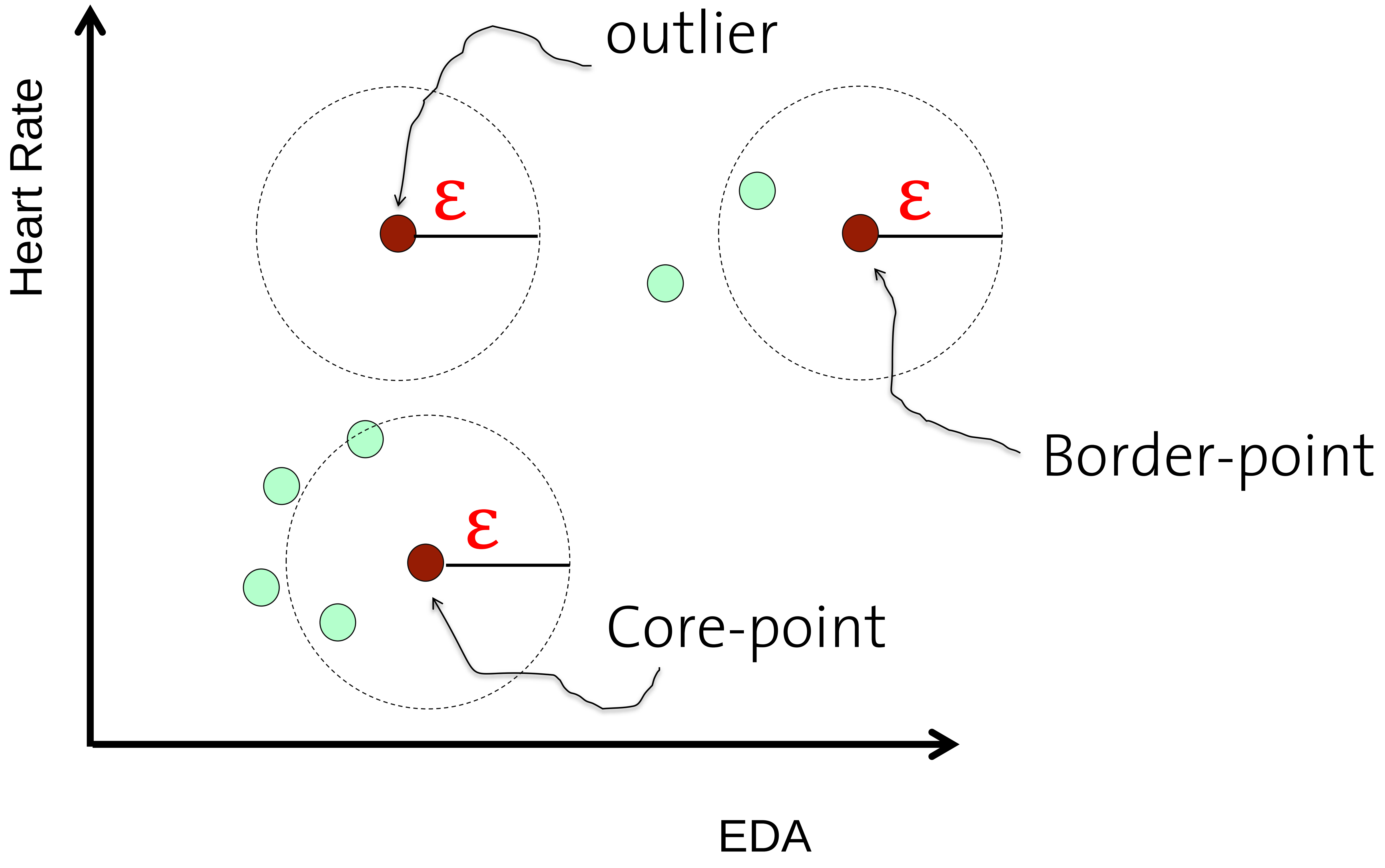
- Select two parameters:
  - ε-Neighborhood
  - m-minimum number of points
- Select a point *p* randomly
- Find all points reachable from *p* based on ε and m
- If *p* is a *core point*, mark points as a cluster.
- If p is a *border point*, no points are reachable from *p*. Select a new point *p*.
- Continue until all of the points are visited.

ESUM data

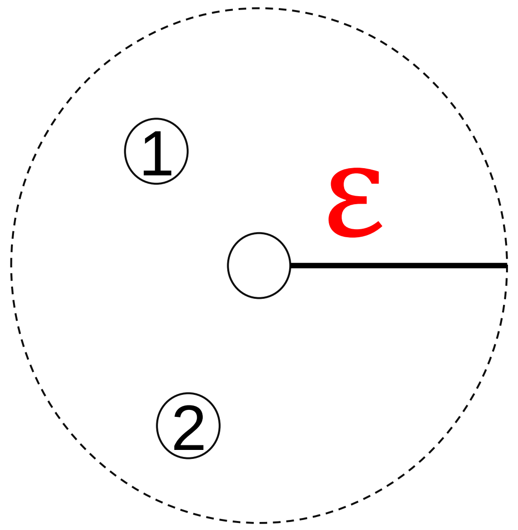


Clustering

Density based : DBSCAN



Define a center

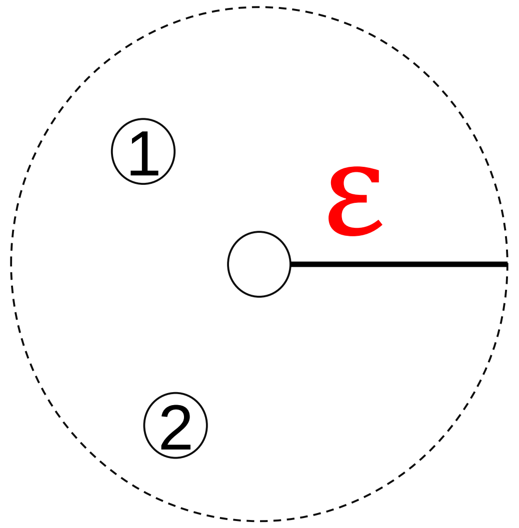


ε = 1 cm

m = 2

Clustering (DBSCAN)

For a center



ε = 1 cm

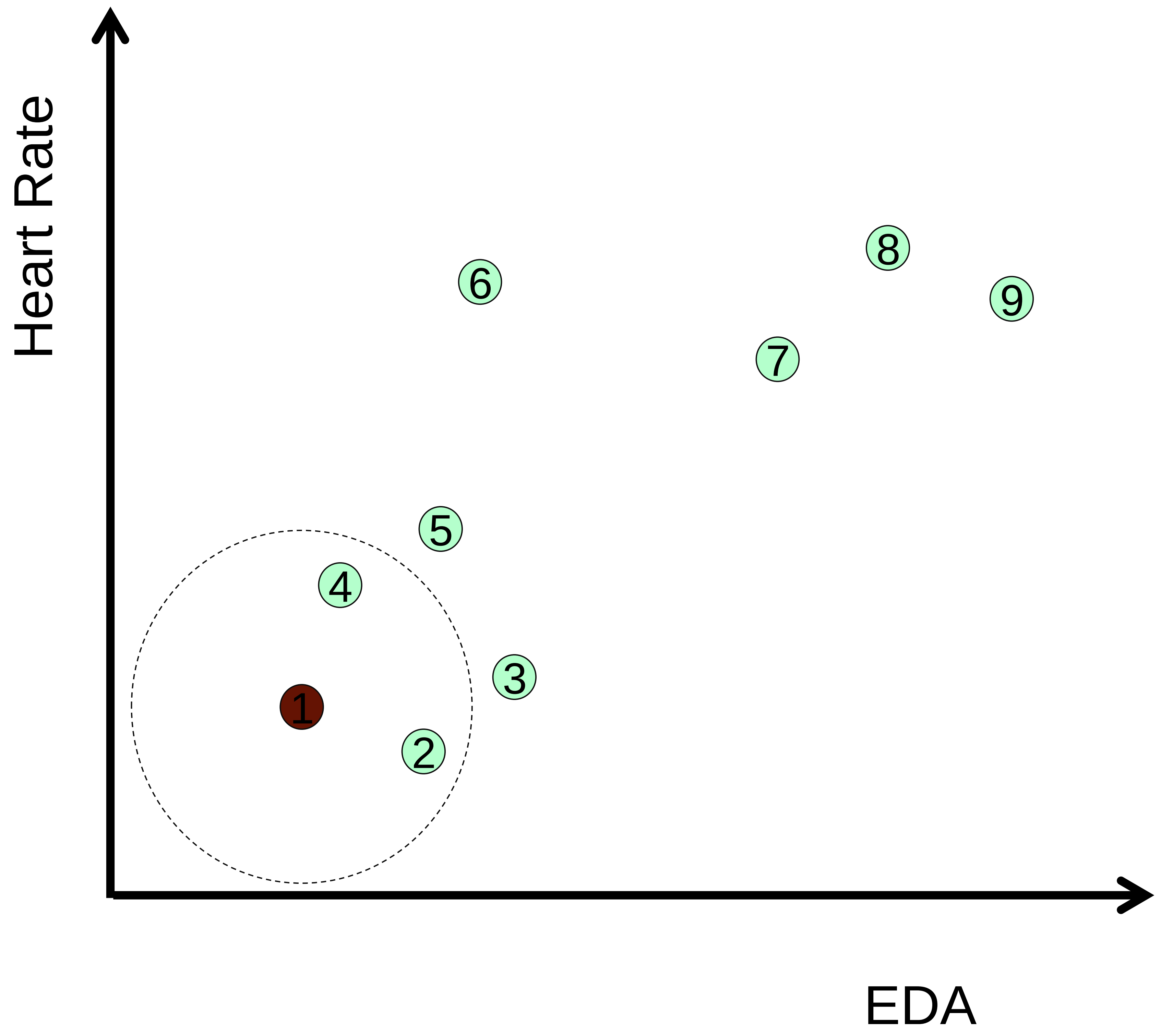
m = 2

Scanning point

1

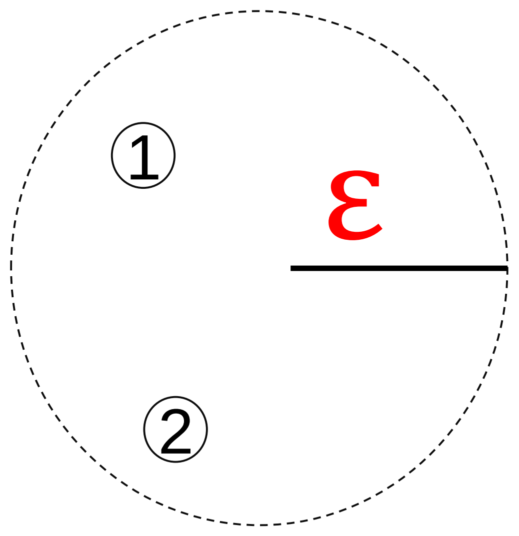
Points in cluster:

1 -> 2,4



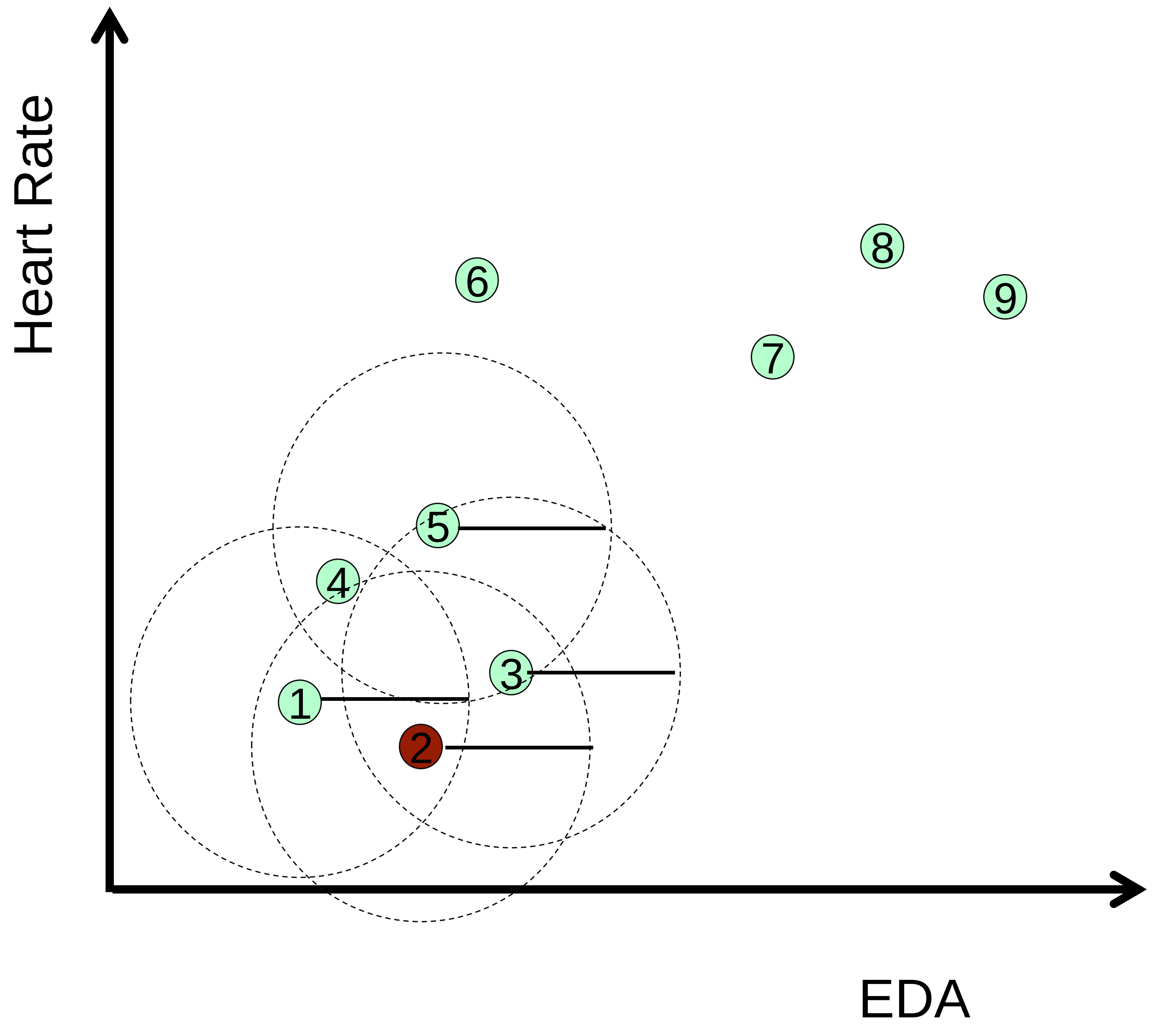
Clustering (DBSCAN)

For a center



ε = 1 cm

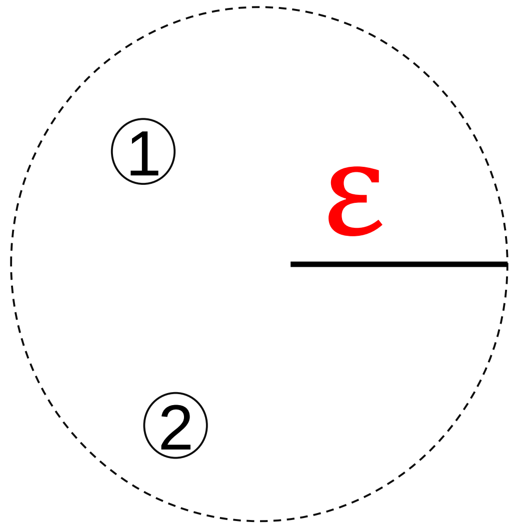
m = 2



Scanning point	Points in cluster:
# 2,	2 -> 1, 3,4
#3,	3 -> 2, 5
#5	5 -> 4,3

Clustering (DBSCAN)

For a center



ε = 1 cm

m = 2

Scanning point

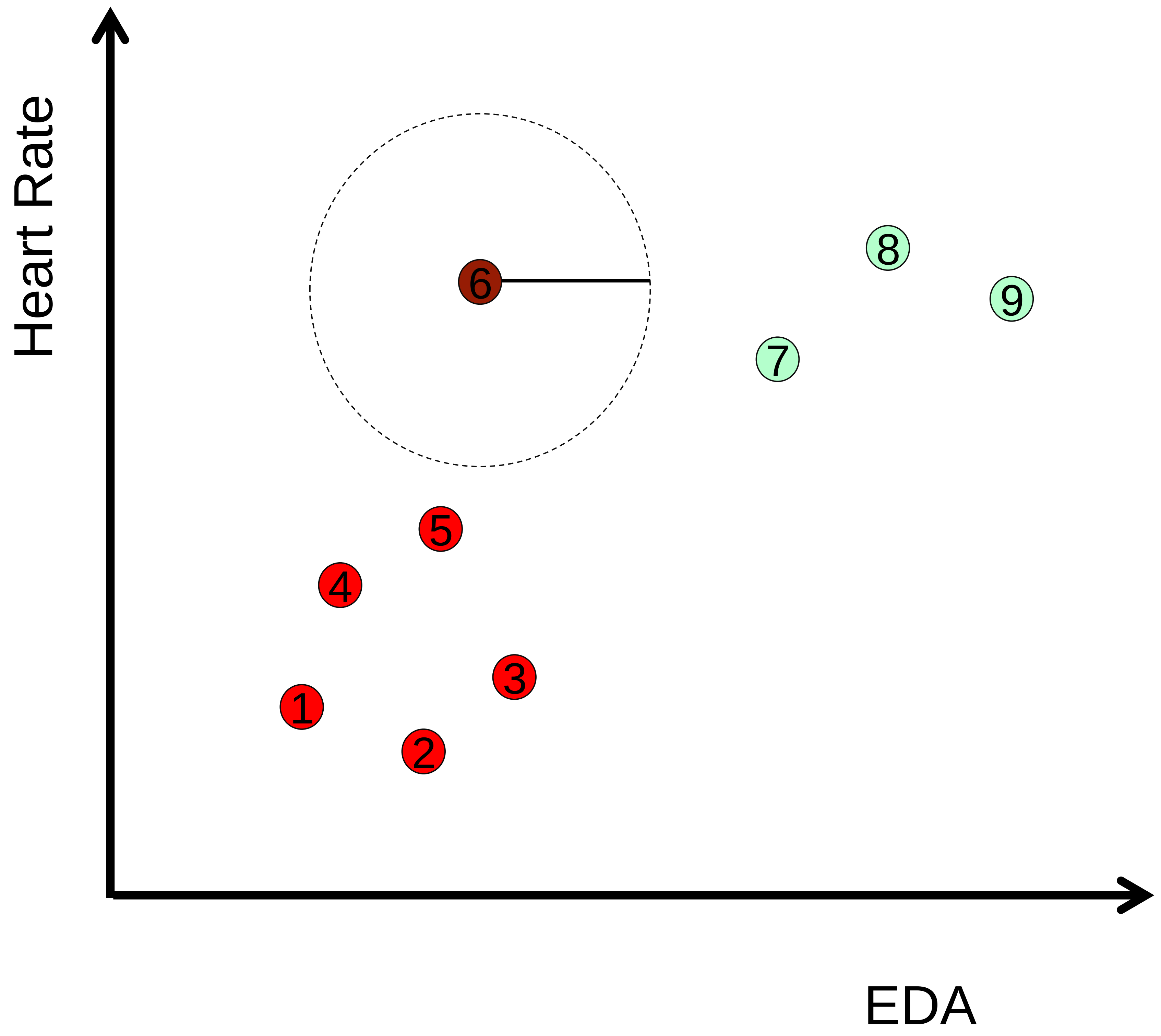
6,

Points in cluster:

6 ->

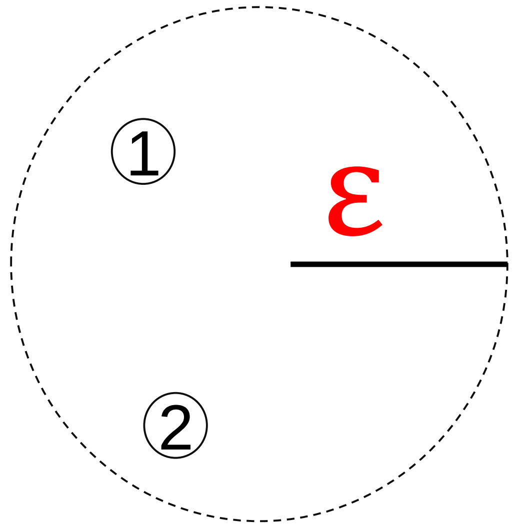
Cluster found:

red



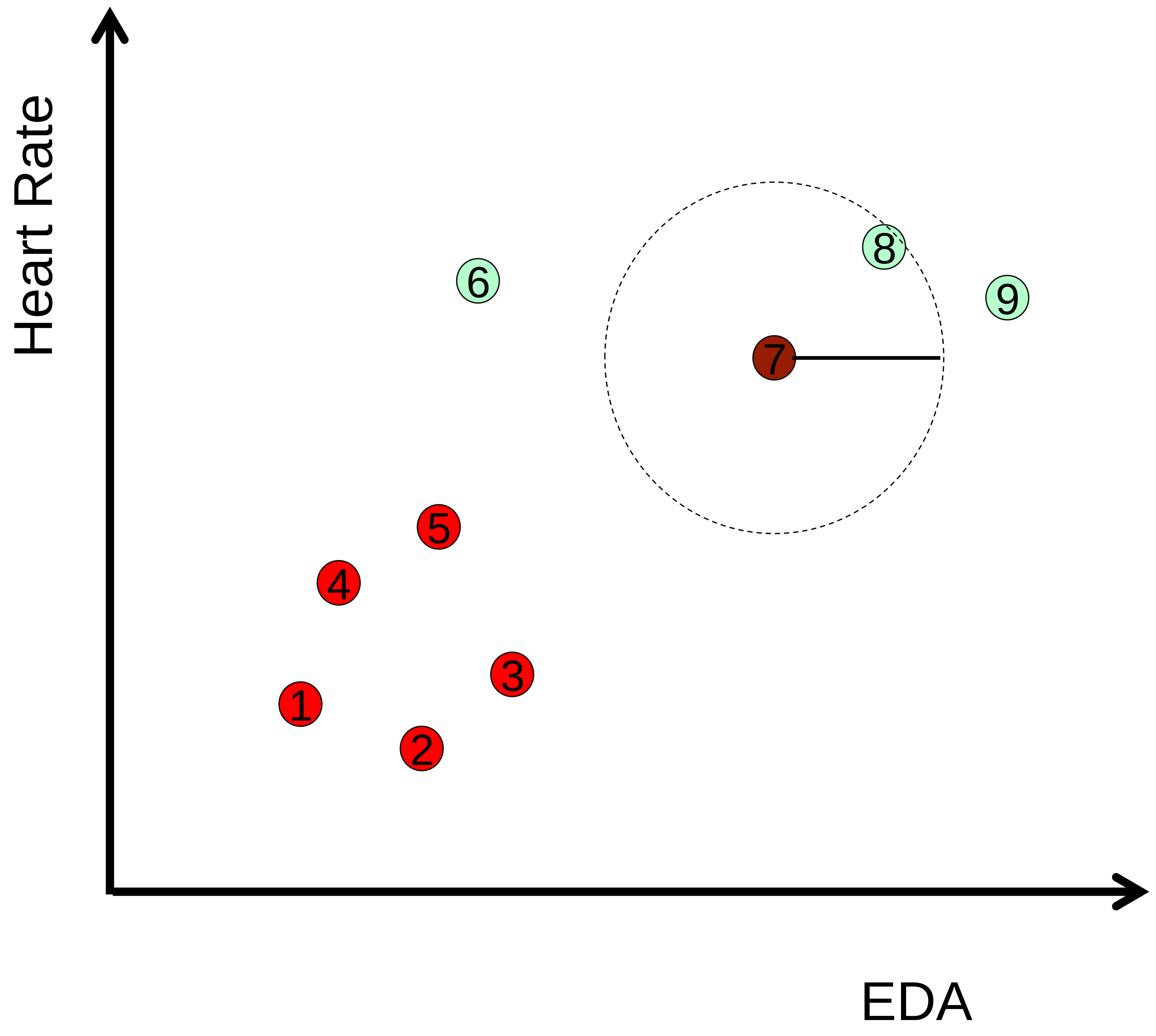
Clustering

For a center



ε = 1 cm

m = 2



Scanning point

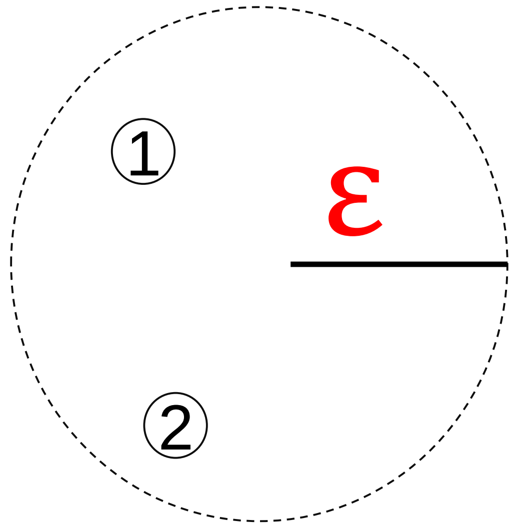
7,

Points in cluster:

7 -> 8

Clustering (DBSCAN)

For a center



ε = 1 cm

m = 2

Scanning point

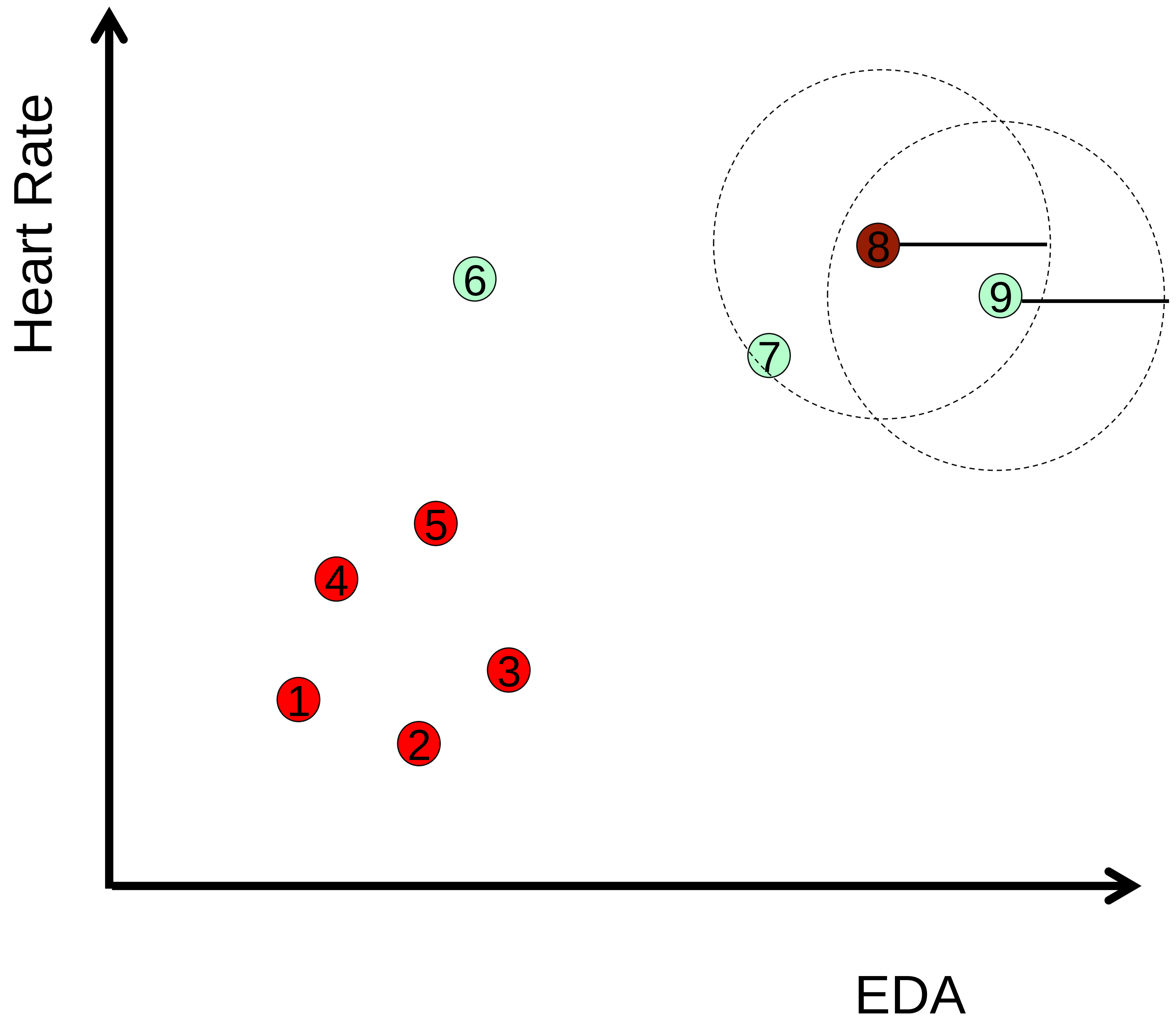
8,

#9

Points in cluster:

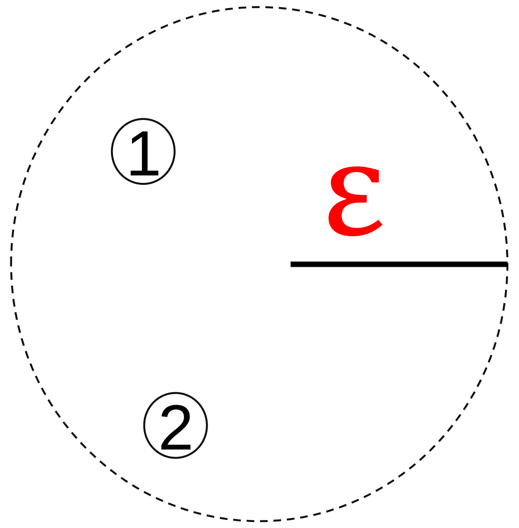
8 -> 7, 9

9 -> 8



Clustering (DBSCAN)

For a center



ε = 1 cm

m = 2

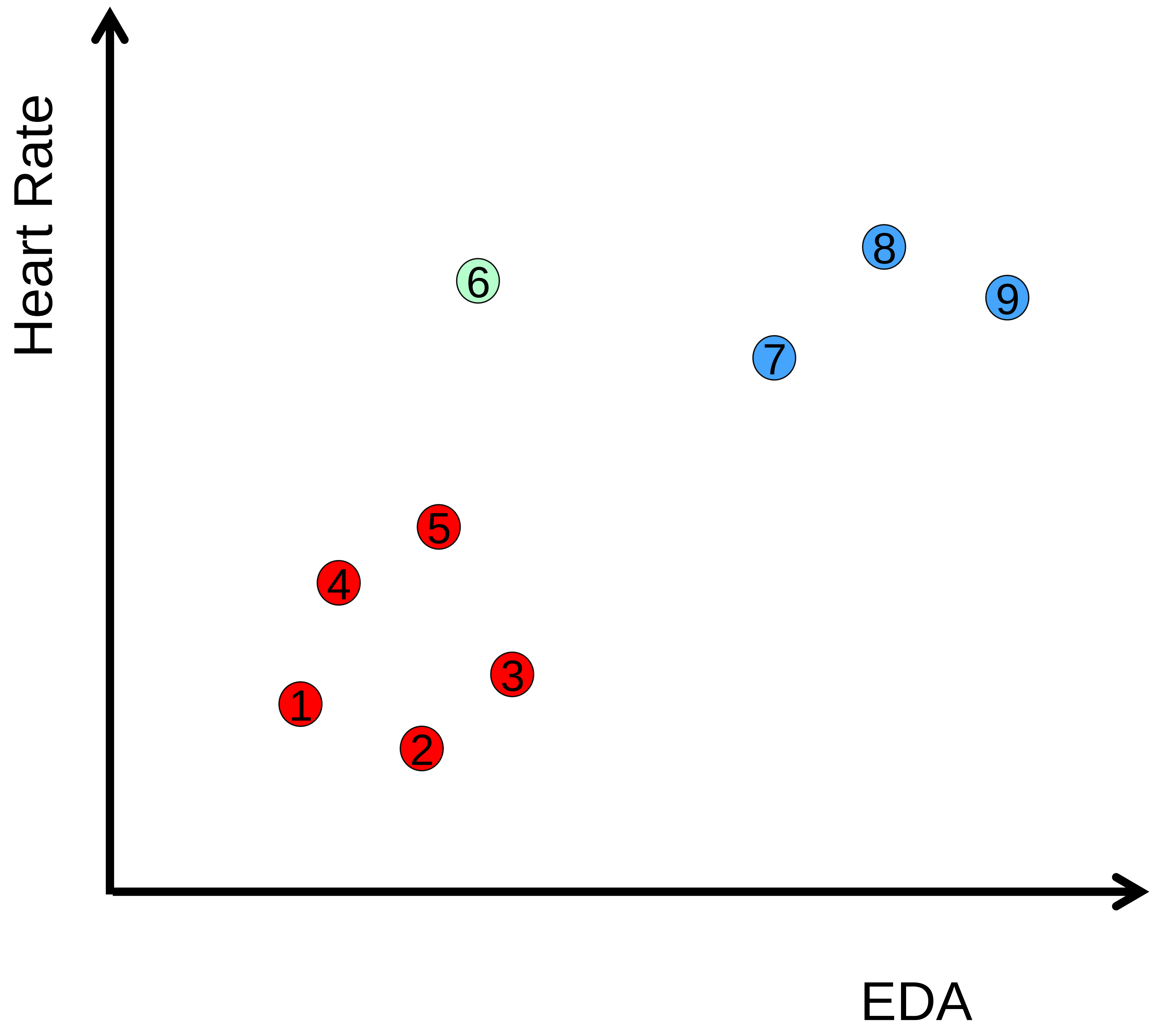
Scanning point

Finish

Total cluster 2

red

blue

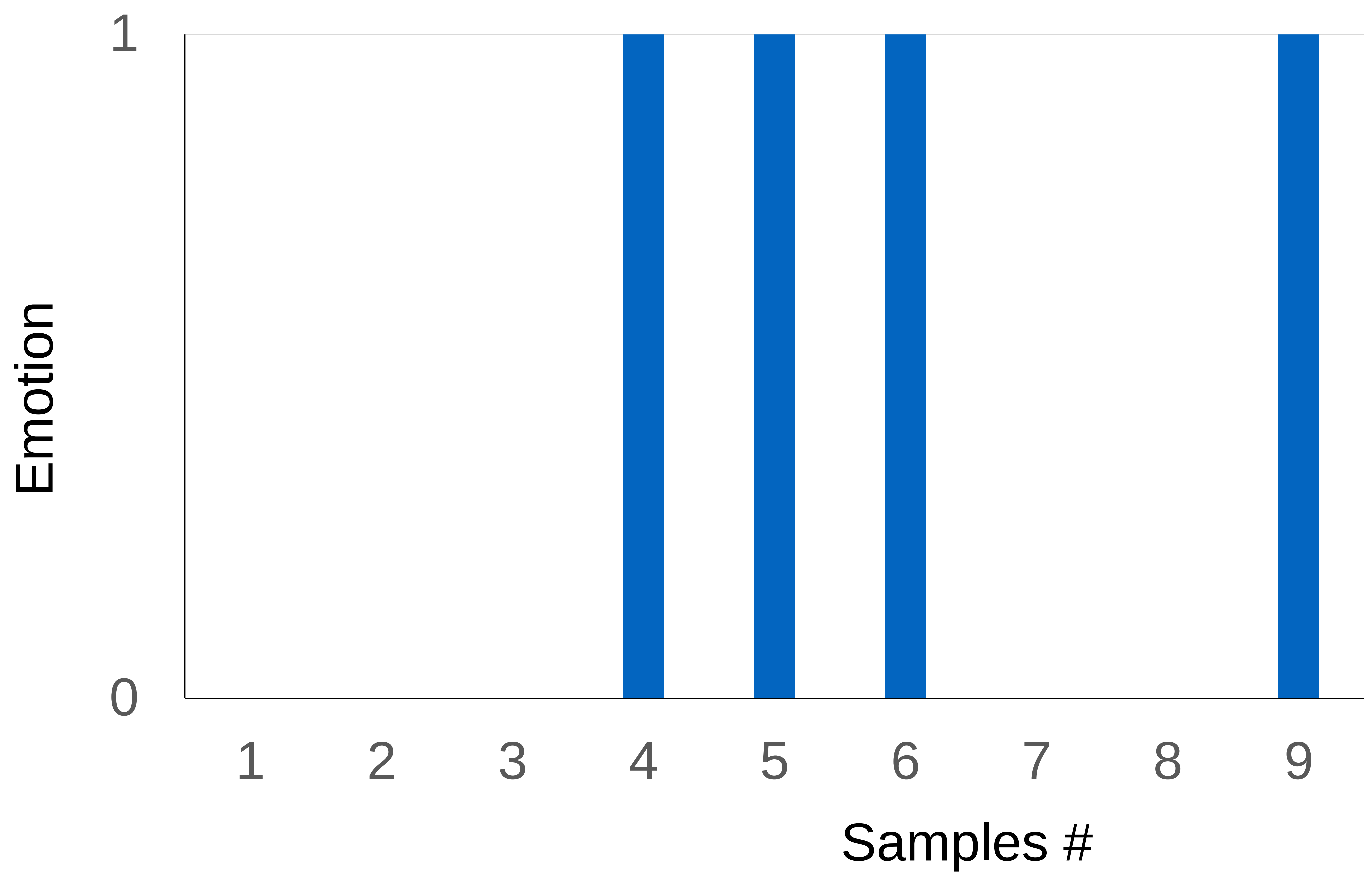


Data (labeled)

Data samples are labeled

Discrete labeled data

#	Inputs				Target
	HR	BVP	EDA	TempBF	Emotion
1	76.85	17.27	0.22	34.63	Negative
2	76.97	19.54	0.22	34.64	Negative
3	77.10	18.51	0.22	34.64	Negative
4	85.28	46.09	0.22	34.61	Positive
5	85.42	35.83	0.22	34.61	Positive
6	88.02	2.59	0.22	34.63	Positive
7	77.25	6.34	0.22	34.65	Negative
8	77.49	6.98	0.22	34.63	Negative
9	85.81	12.18	0.22	34.61	Positive

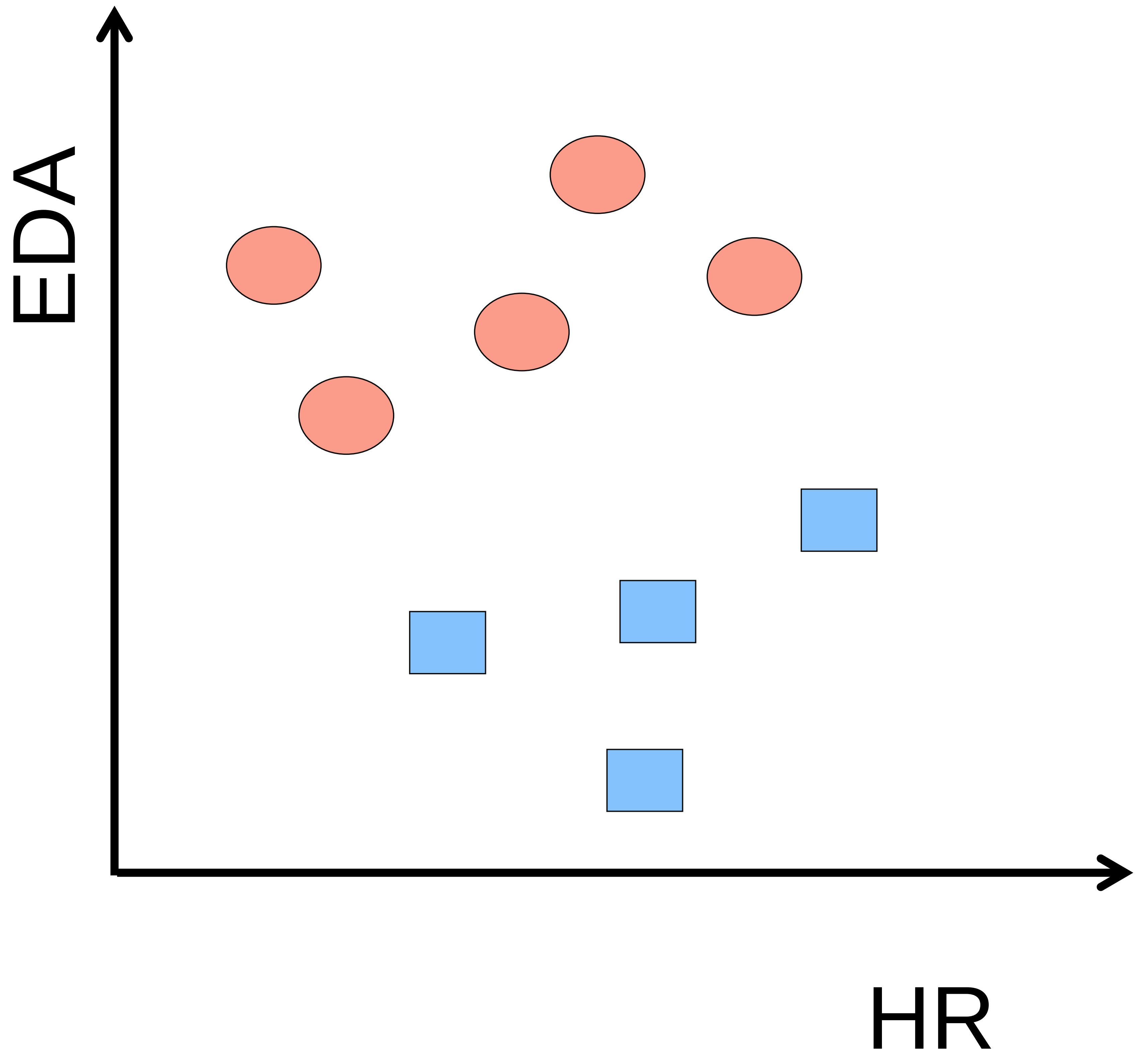


1 = positive

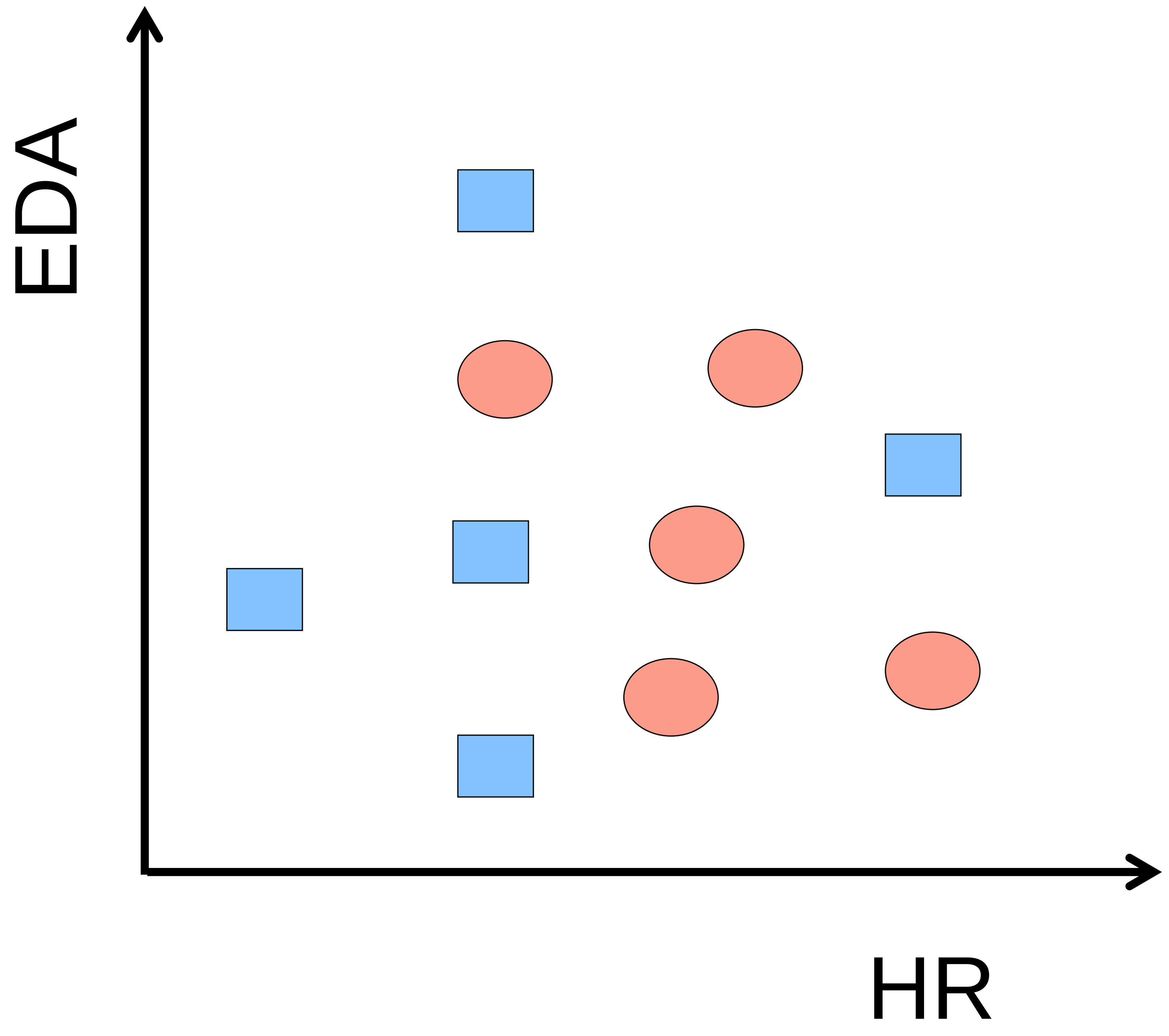
0 = negative

Classification

Linear classification

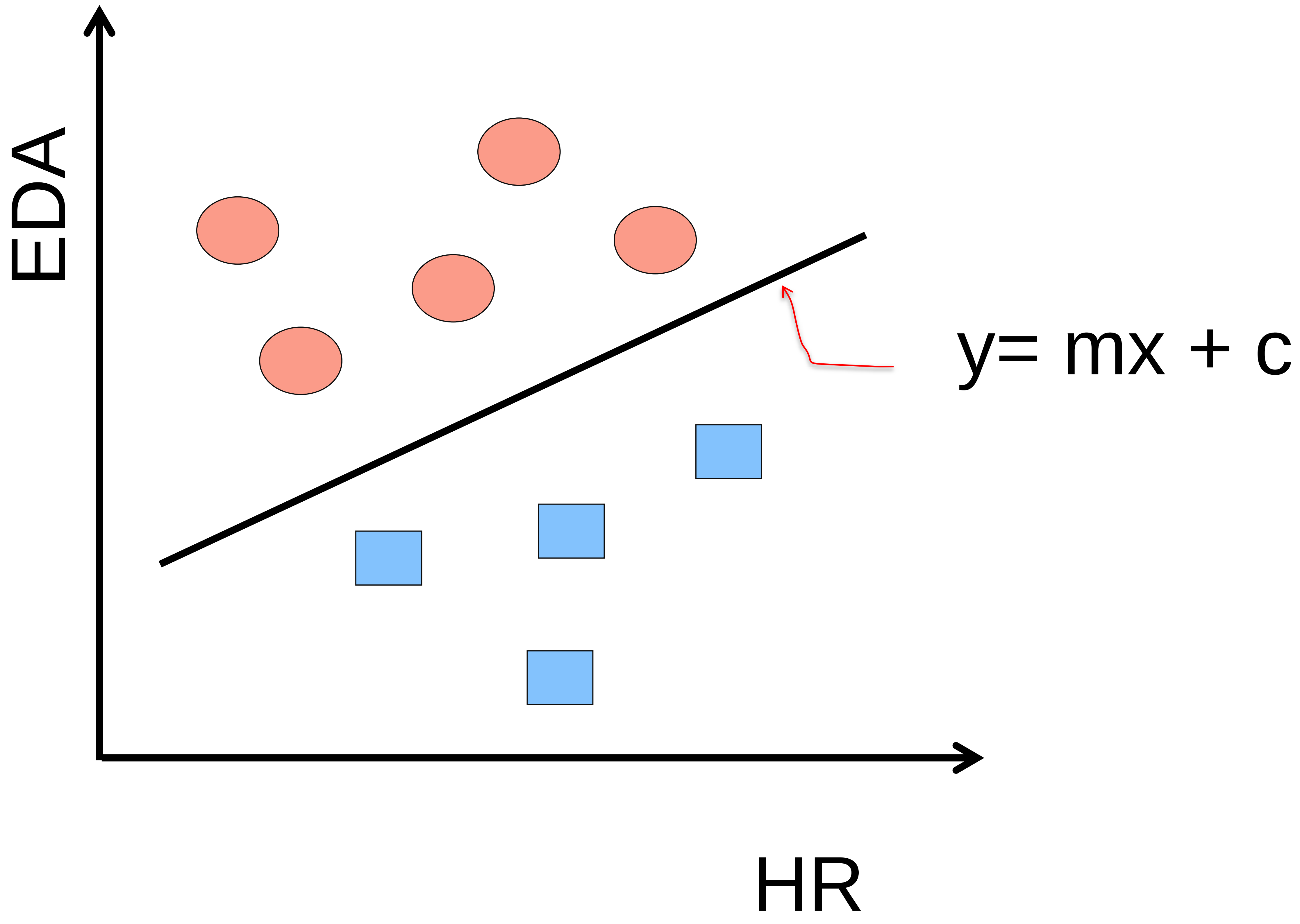


Non-linear classification

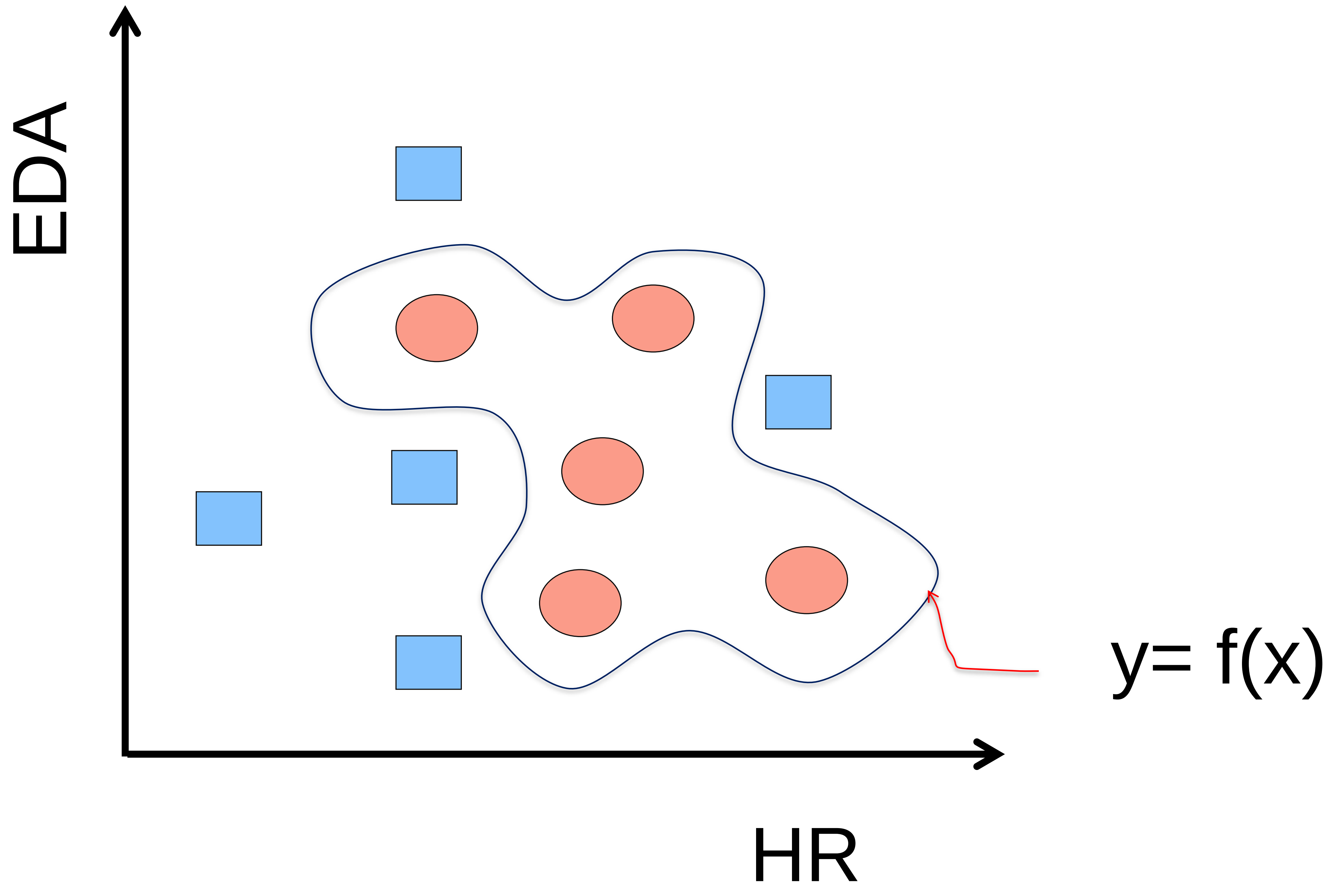


Classification

Linear classification



Non-linear classification



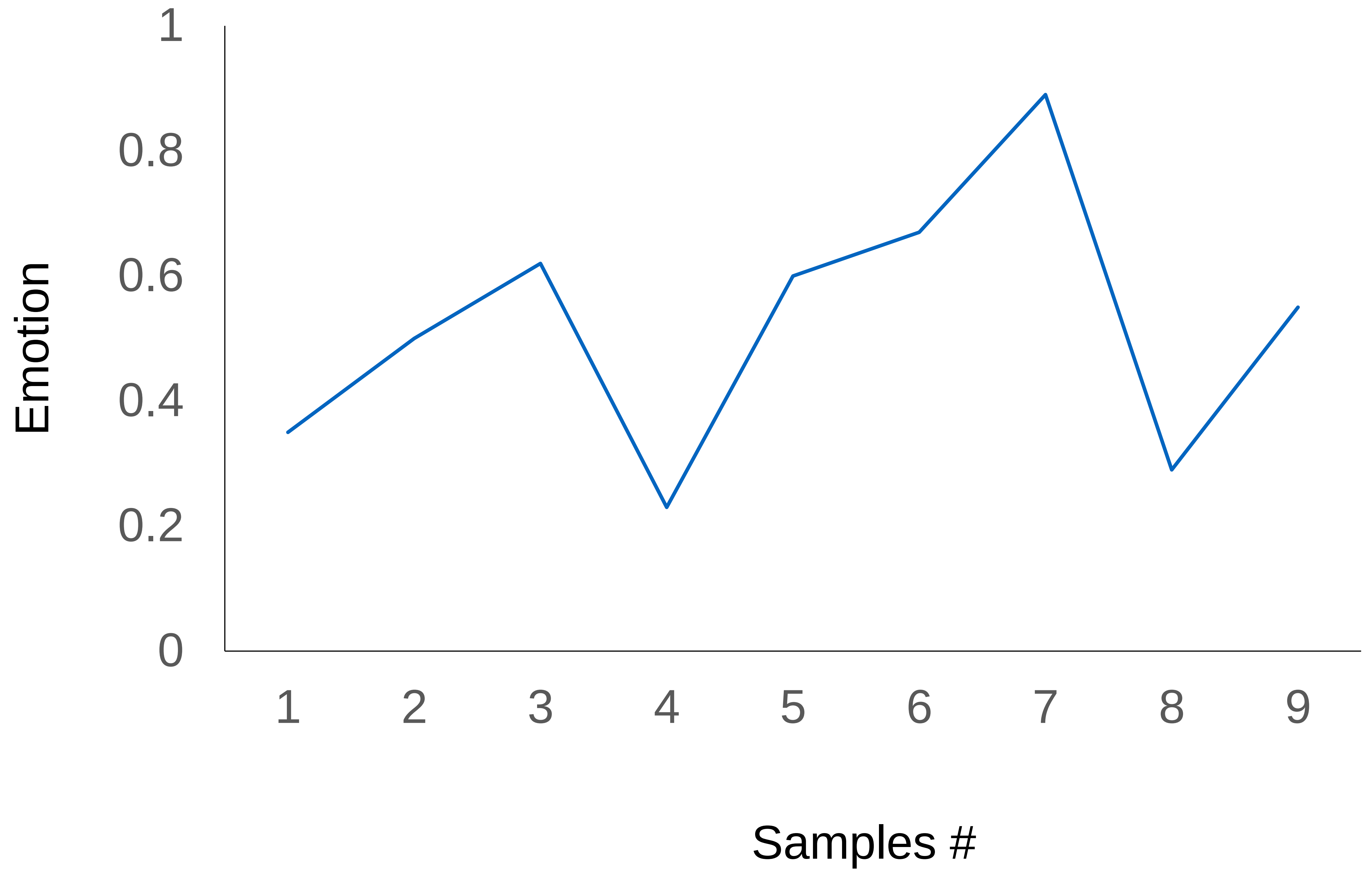
What is f(x)?

Data (labeled)

Target has continuous value

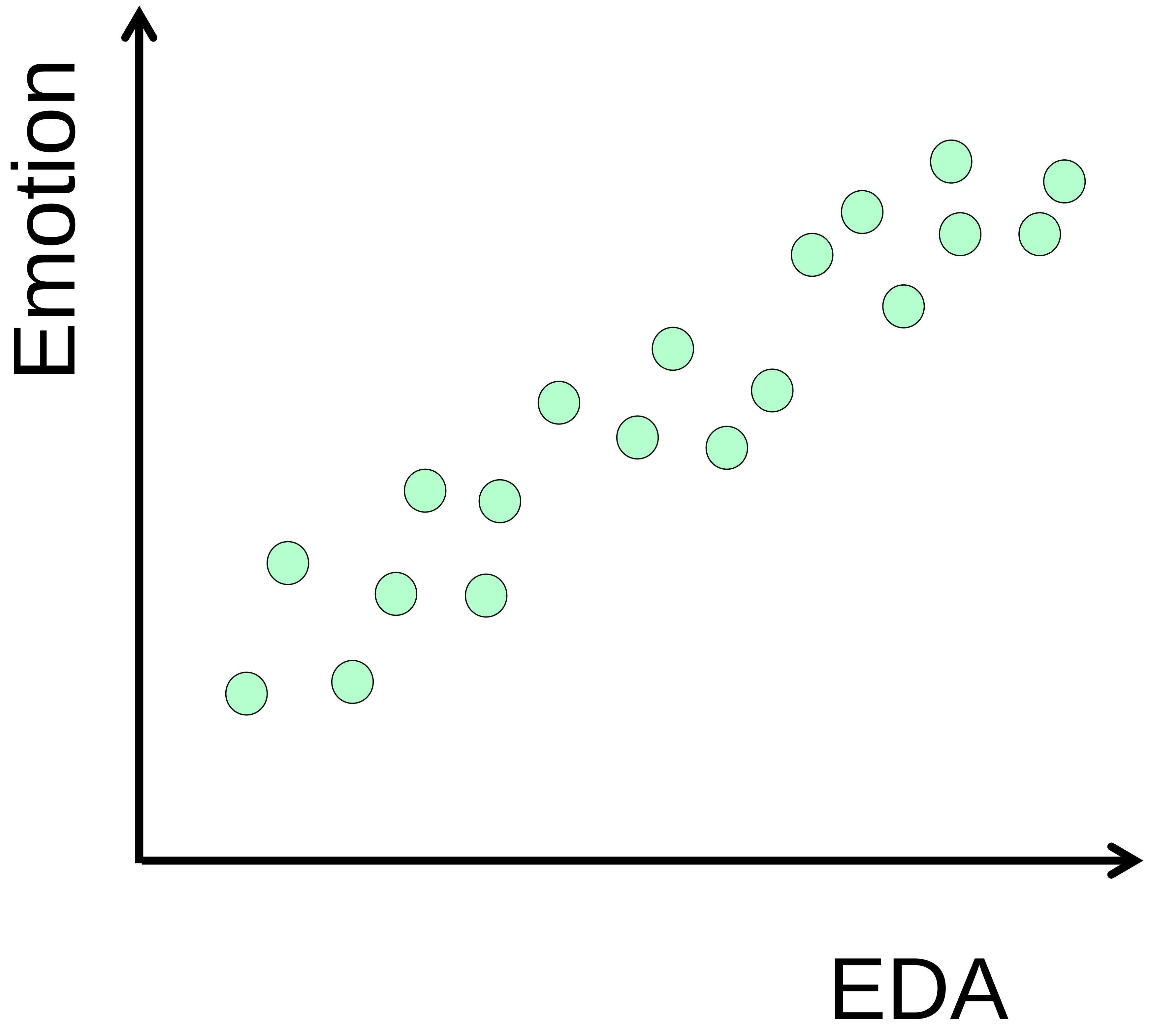
Continuous labeled data

#	Inputs				Target
	HR	BVP	EDA	TempBF	Emotion
1	76.85	17.27	0.22	34.63	0.15
2	76.97	19.54	0.22	34.64	0.5
3	77.10	18.51	0.22	34.64	0.76
4	85.28	46.09	0.22	34.61	0.23
5	85.42	35.83	0.22	34.61	0.6
6	88.02	2.59	0.22	34.63	0.67
7	77.25	6.34	0.22	34.65	0.89
8	77.49	6.98	0.22	34.63	0.2
9	85.81	12.18	0.22	34.61	0.55

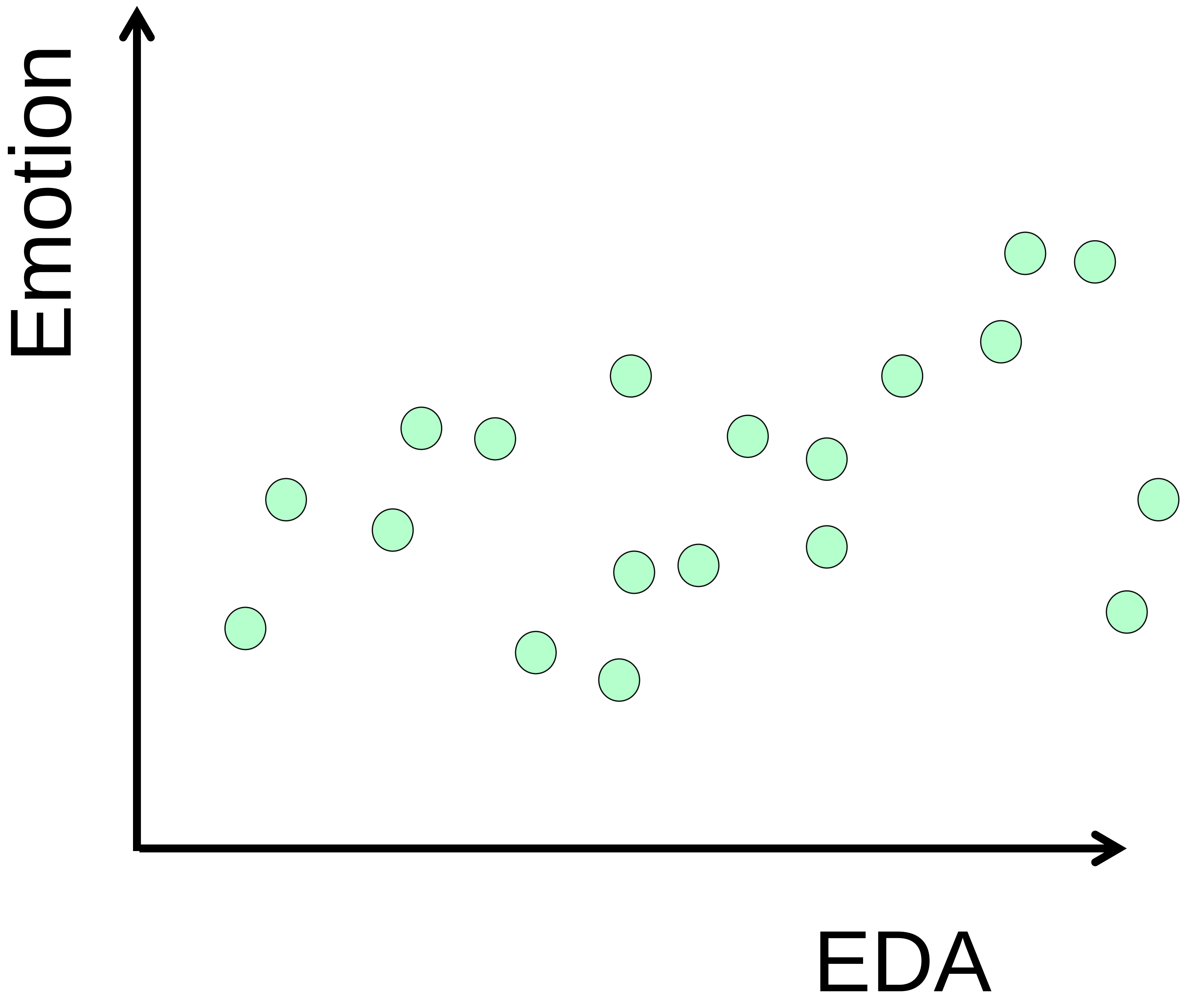


Regression

Linear regression

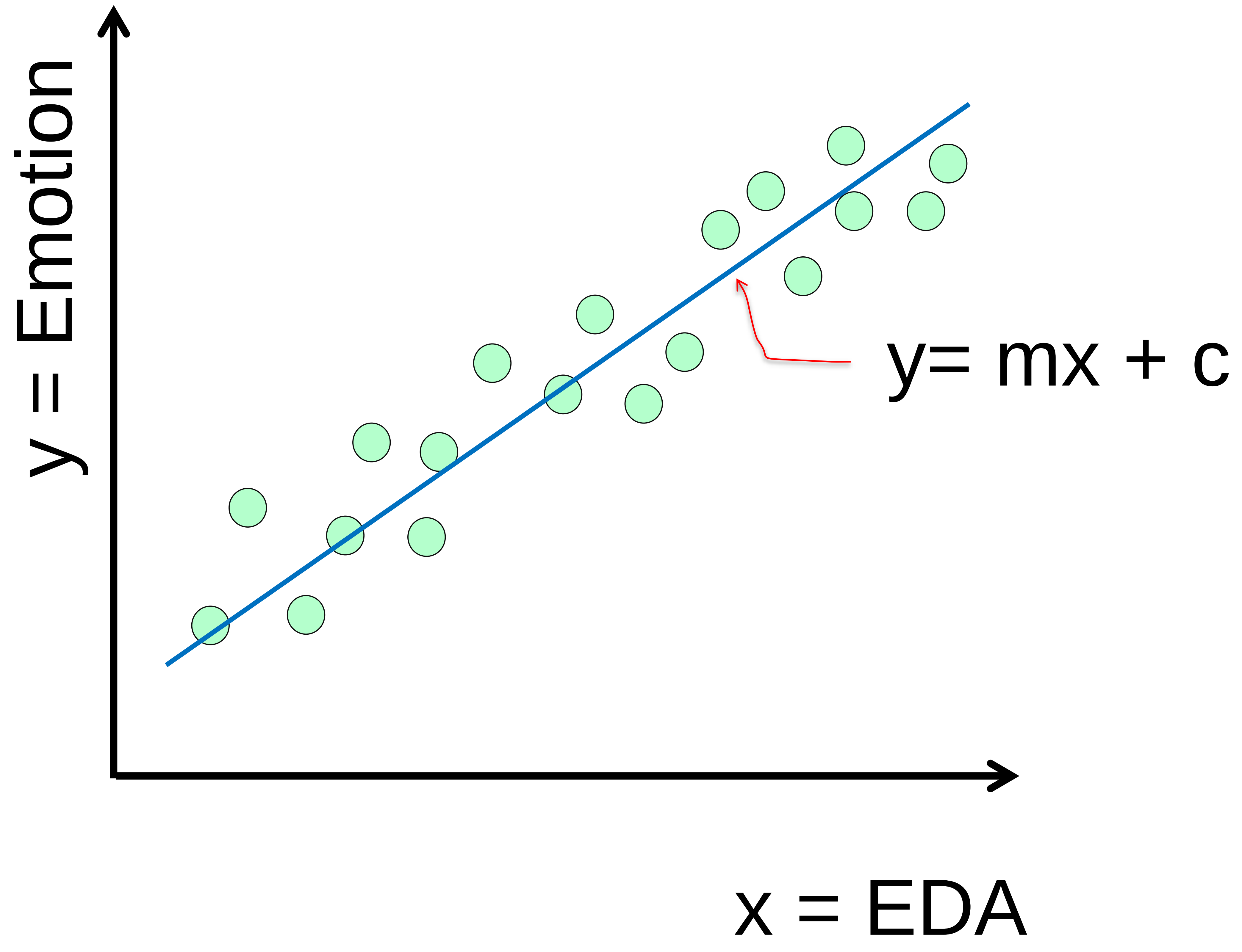


Non-linear regression

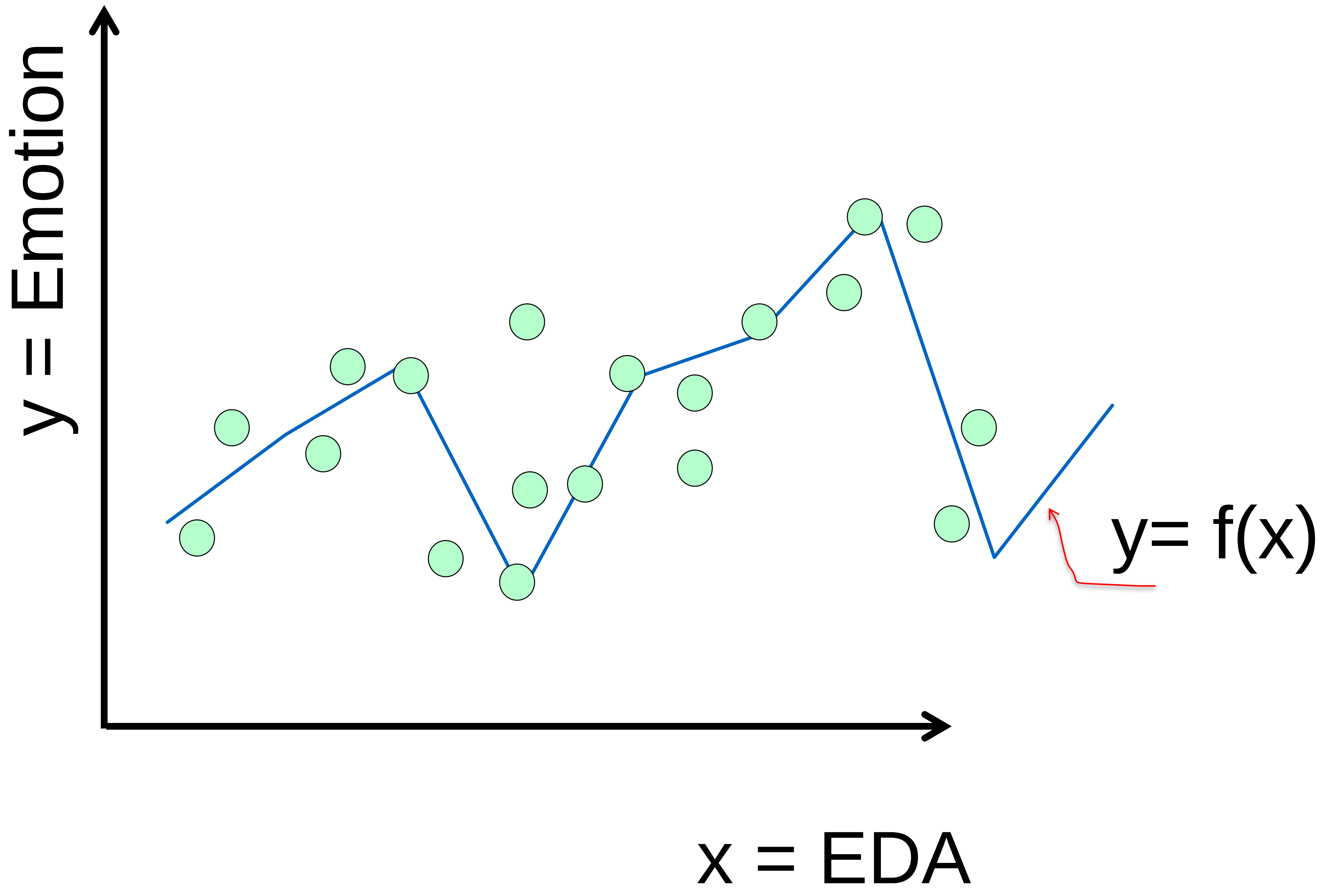


Regression

Linear regression



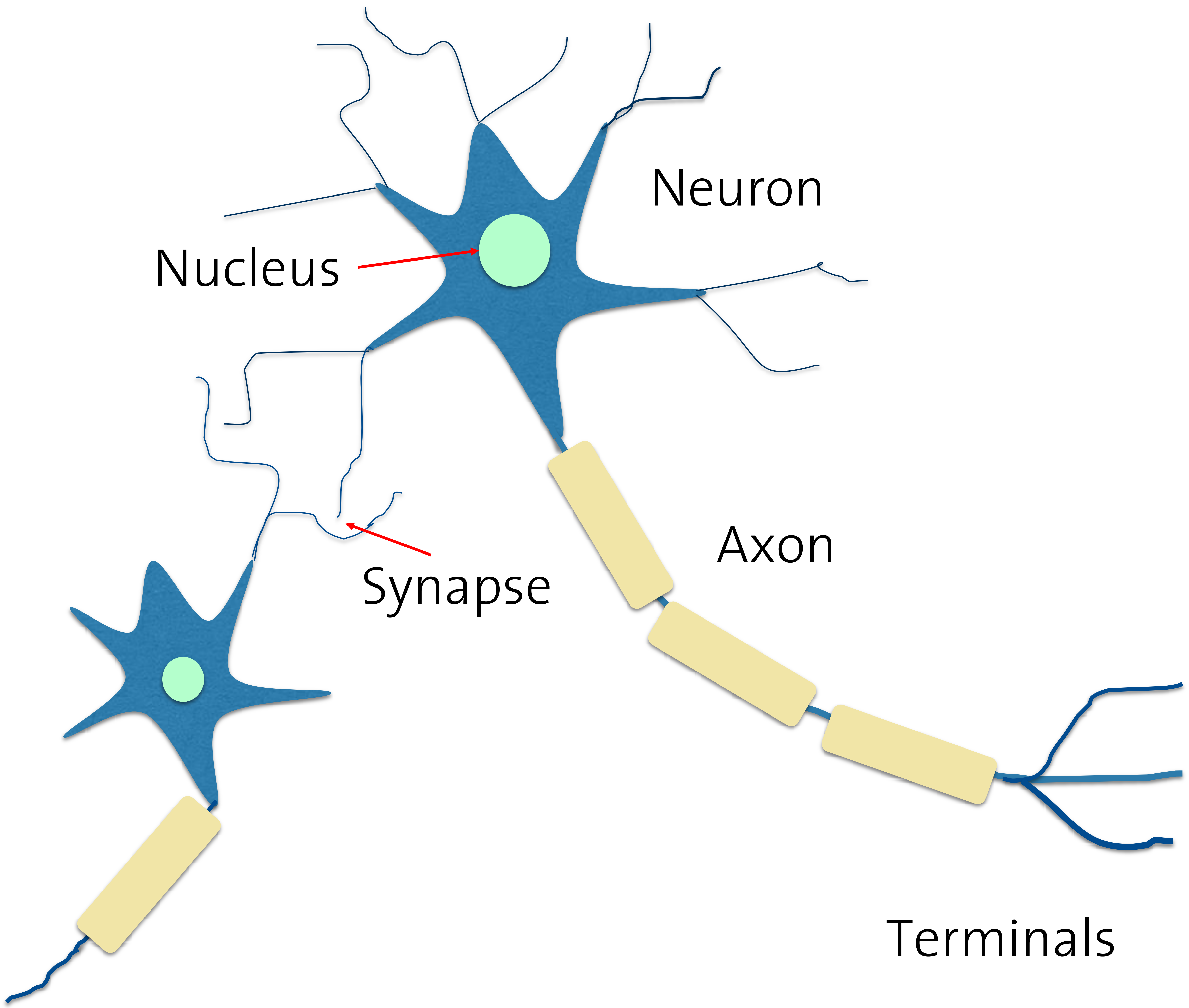
Non-linear regression



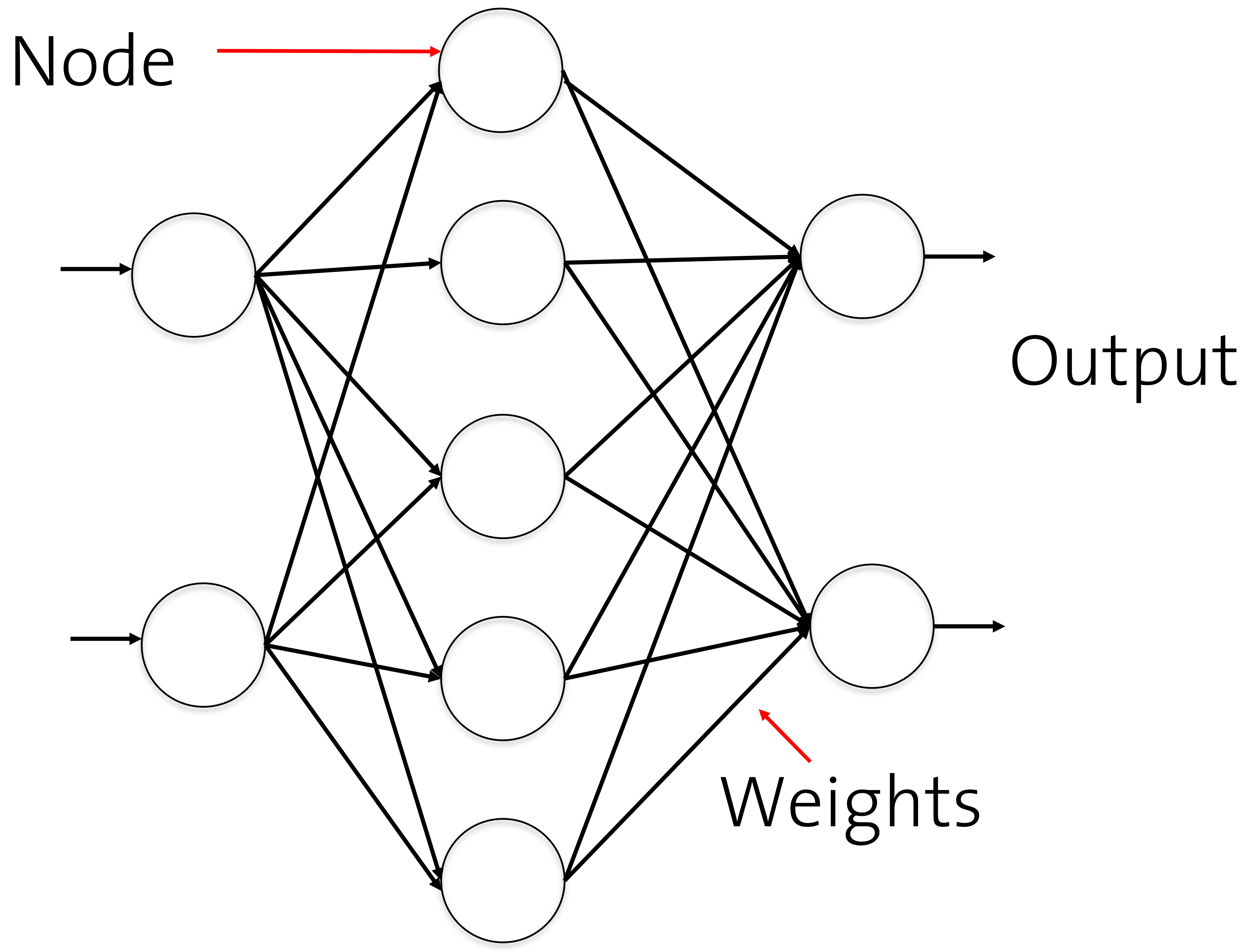
What is f(x)?

Neural Network

Biological neural network



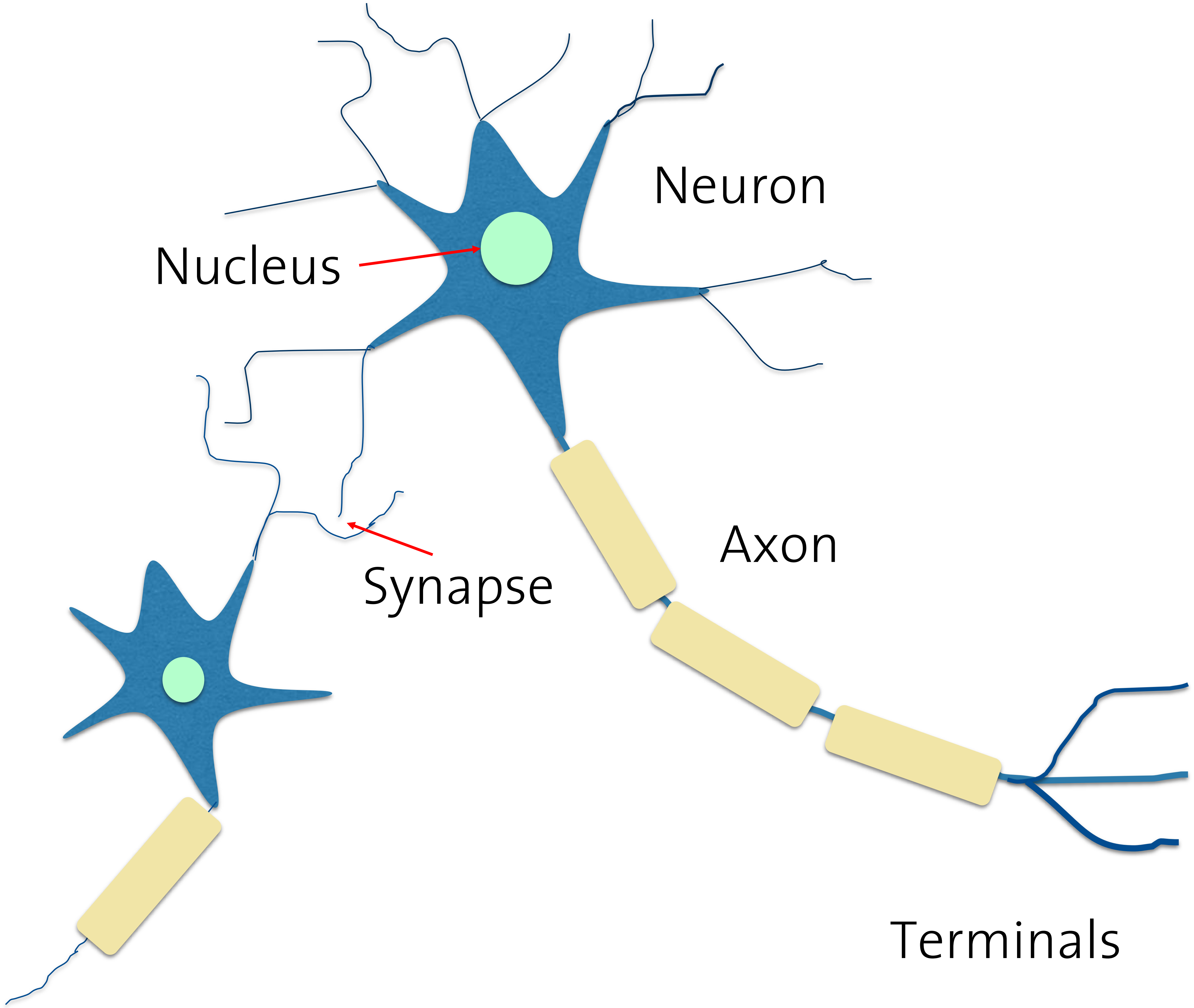
Artificial Neural Network



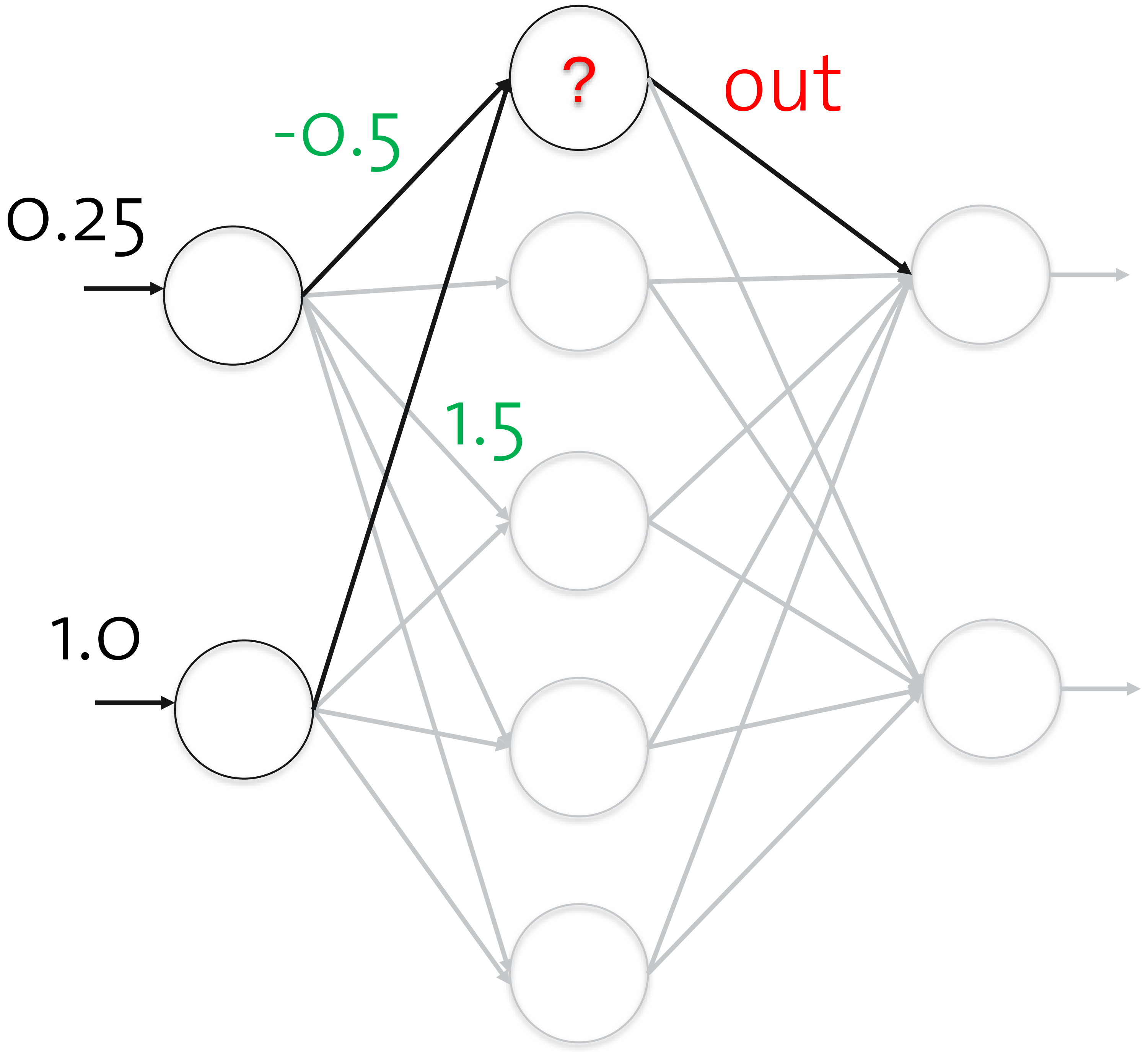
Neuron -> Node
Nucleus -> Squashing function
Synapse -> Weights
Terminals -> Output

Neural Network

Biological neural network



Artificial Neural Network

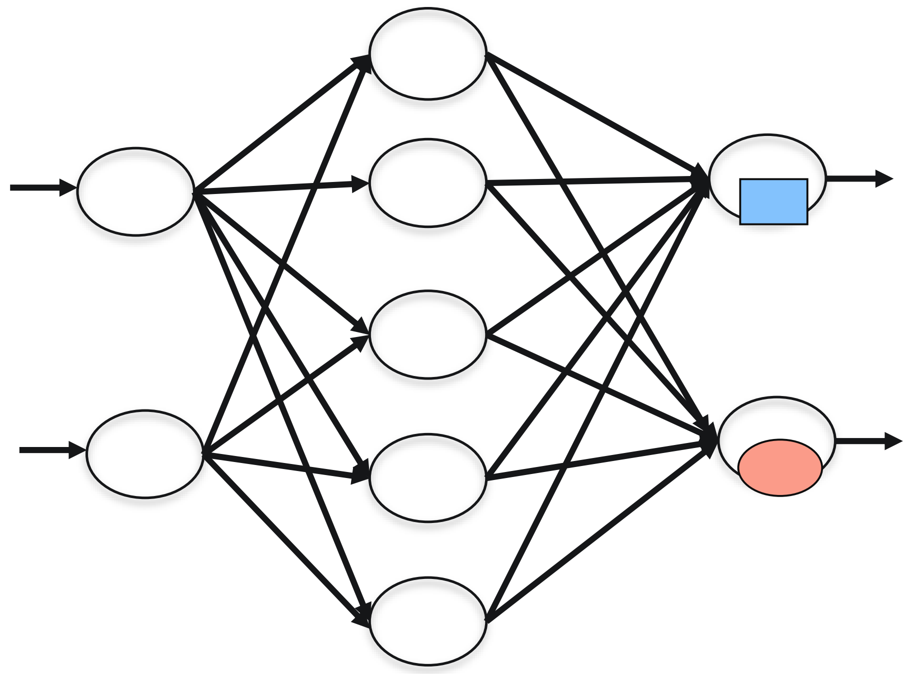


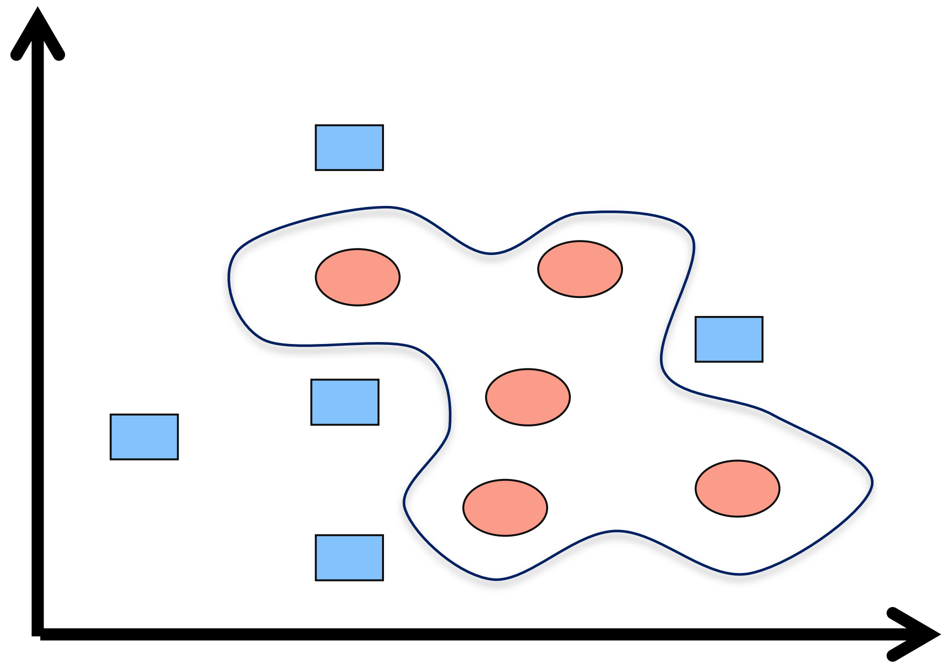
$$\text{sum} = (0.25 \times -0.5) + (1.0 \times 1.5) = 1.625$$

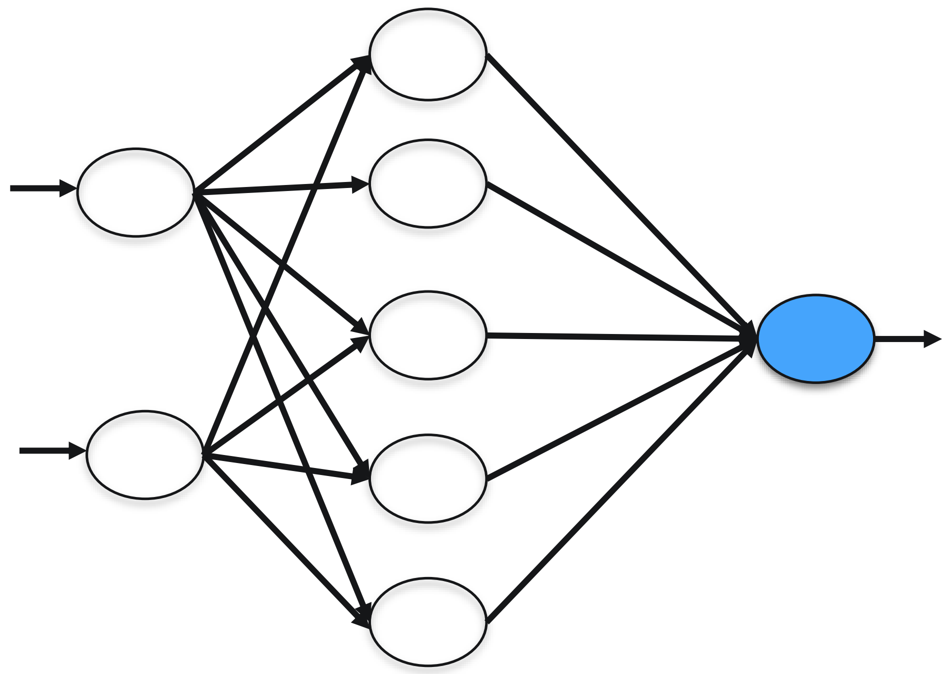
$$\text{out} = \frac{1}{1+e^{1.525}} = 0.165$$ Squashed value

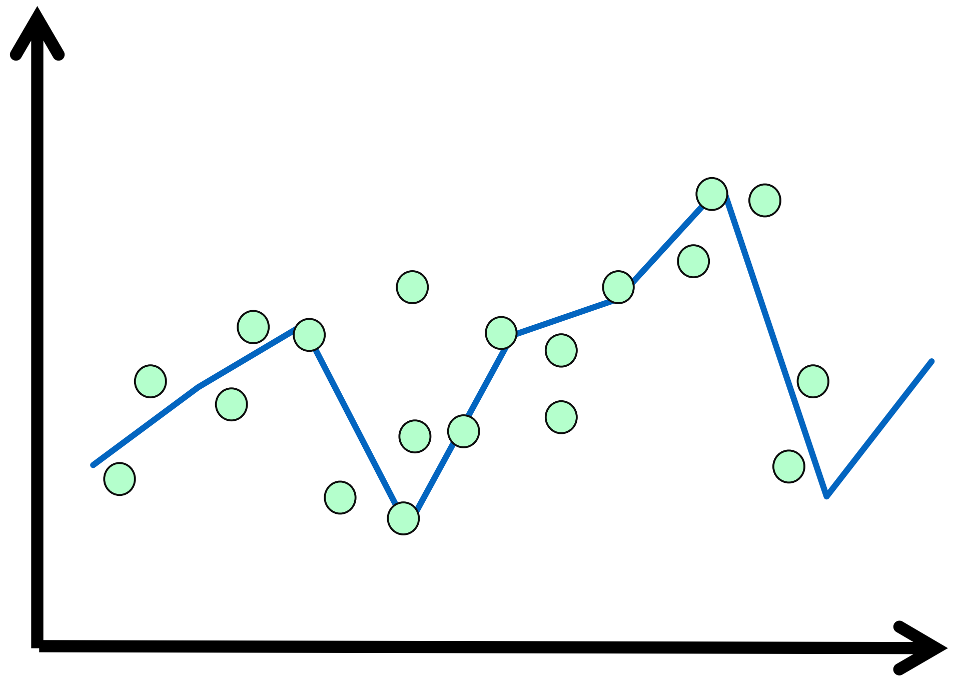
What is f(x)?

Neural Network -> f(x)





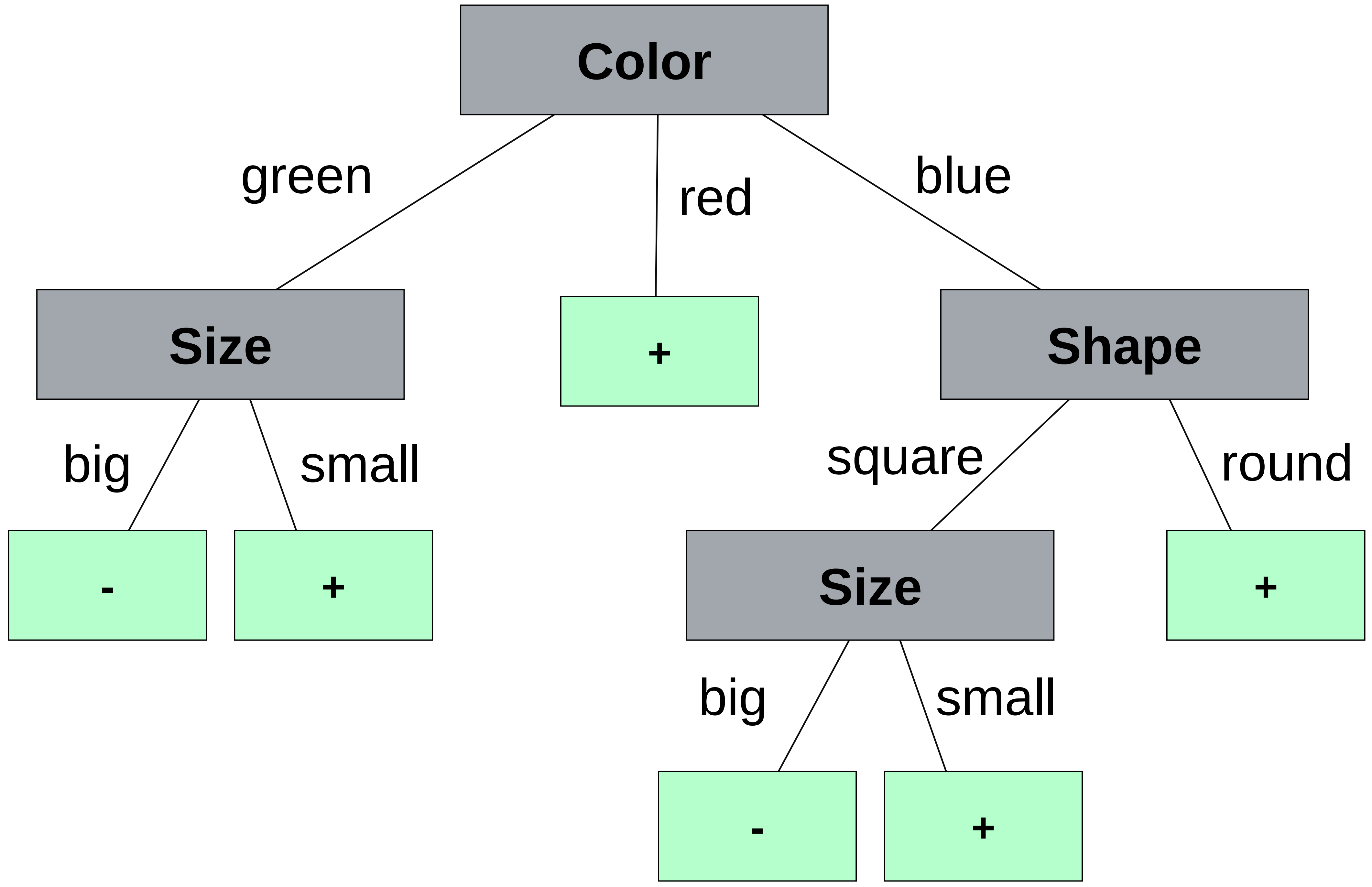


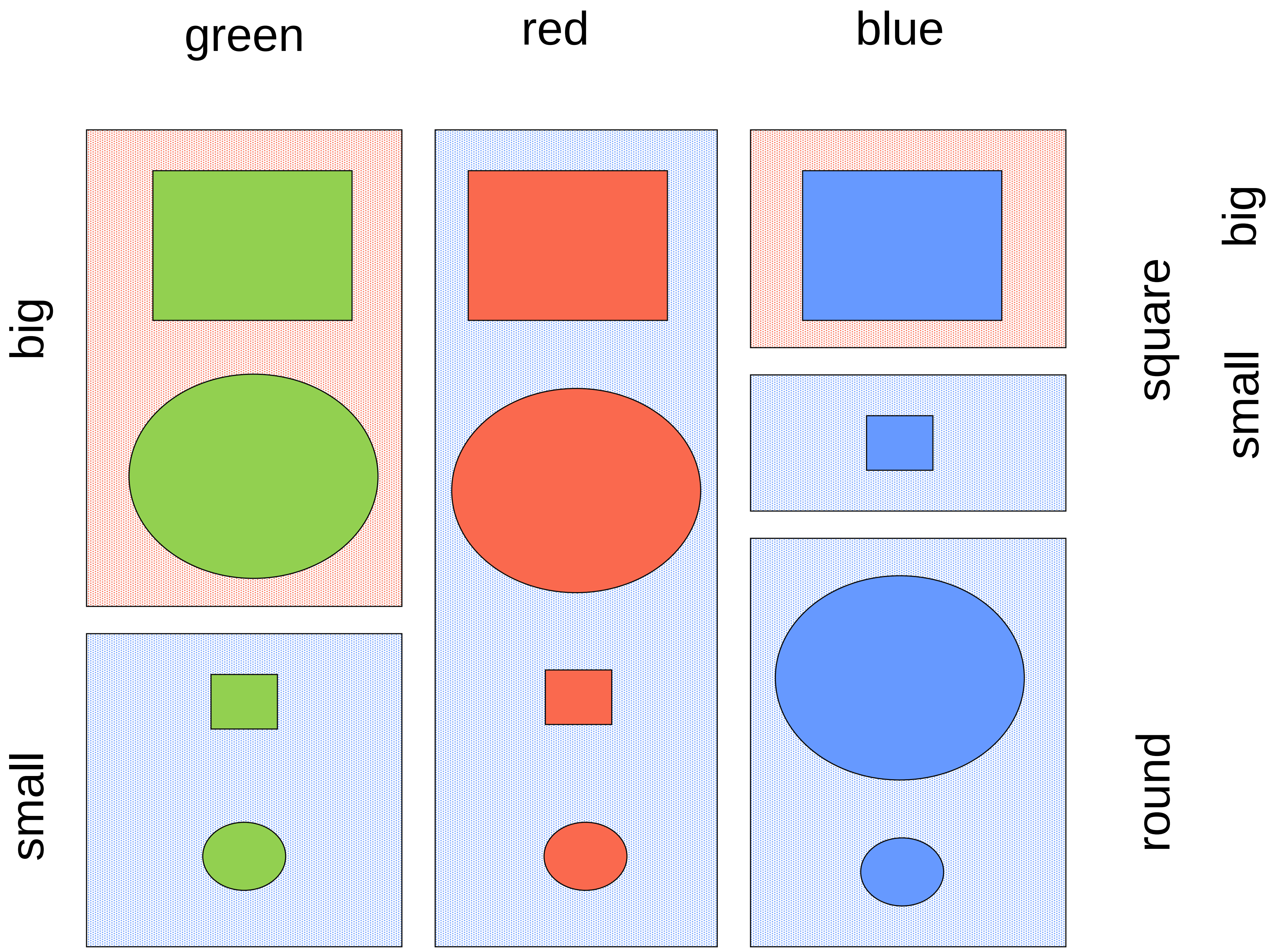


Neural Network will give a function f(x) for which f(x) will separate two classes with a maximum accuracy

Neural Network will give a function f(x) for which f(x) will fit within the data points with less error.

Decision Tree





Slide credit: Anonymous author