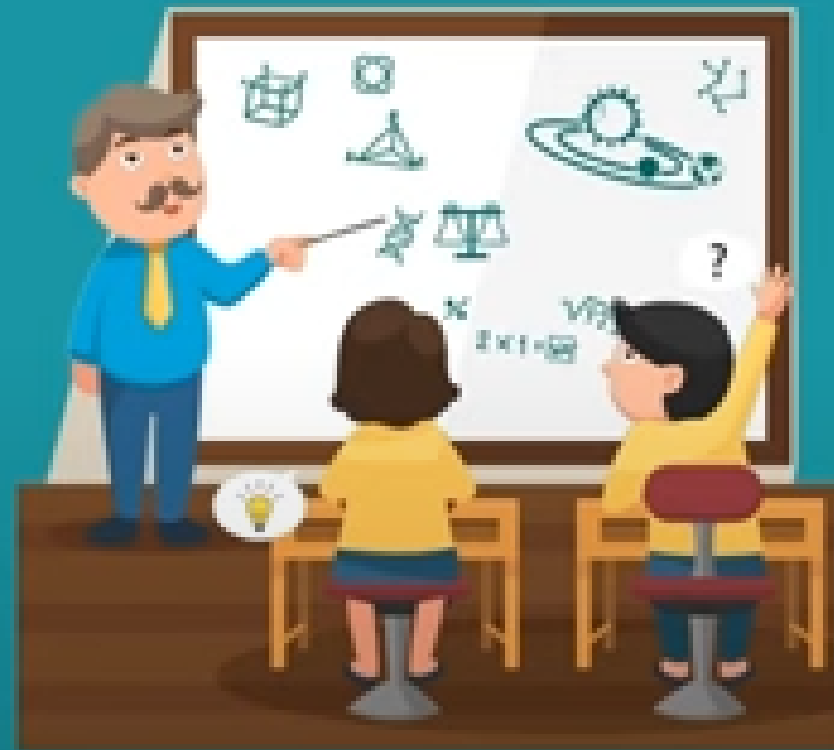


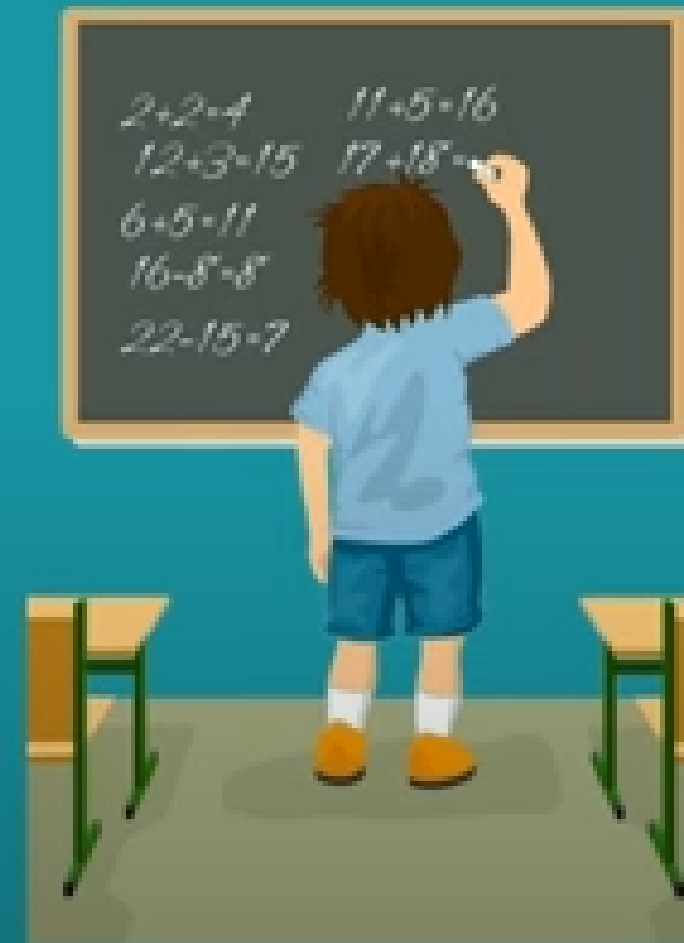
Supervised và Unsupervised Learning

The background features abstract, wavy, concentric line patterns in orange and blue, creating a modern and artistic aesthetic. The orange lines are primarily located in the top right and bottom left corners, while the blue lines form a central, organic shape behind the text.

Supervised learning is a method in which we teach the machine using labelled data

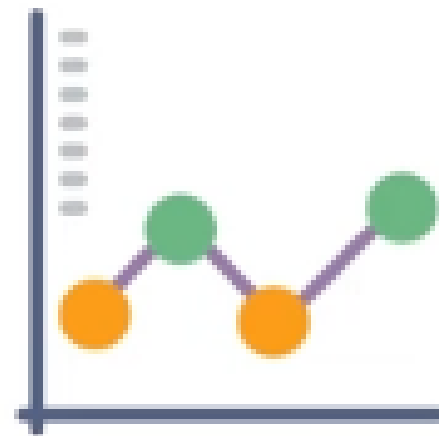


In unsupervised learning the machine is trained on unlabelled data without any guidance



Supervised Learning

Regression



Classification

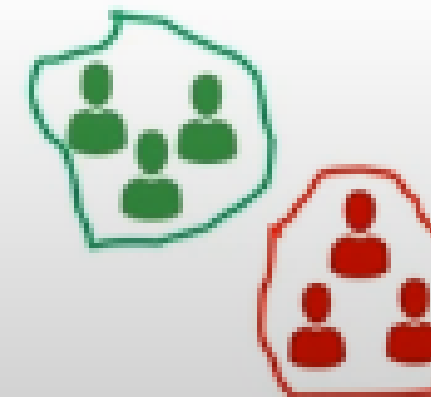


Unsupervised Learning

Association

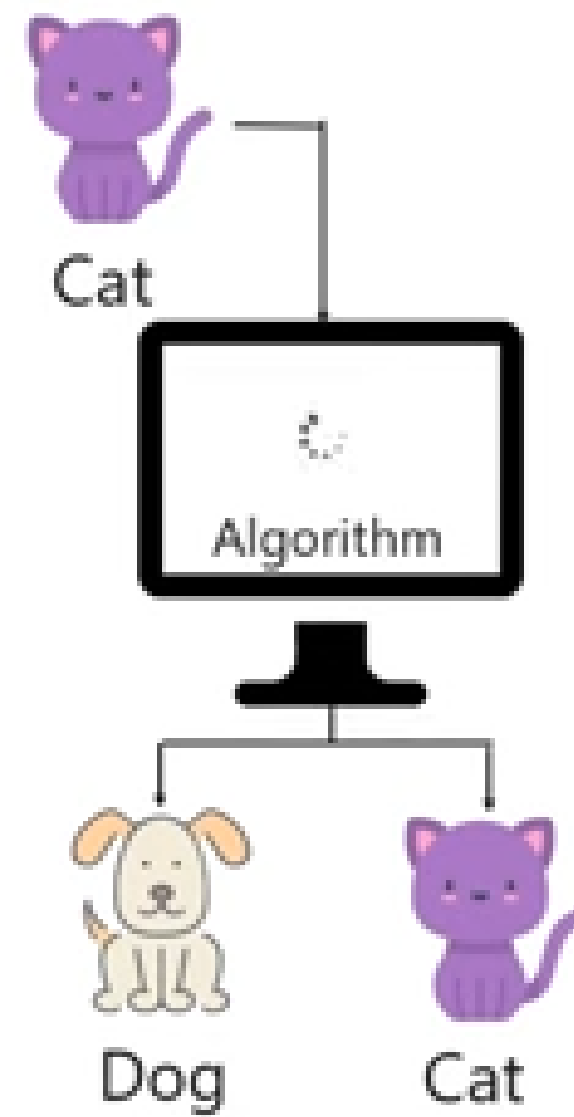


Clustering



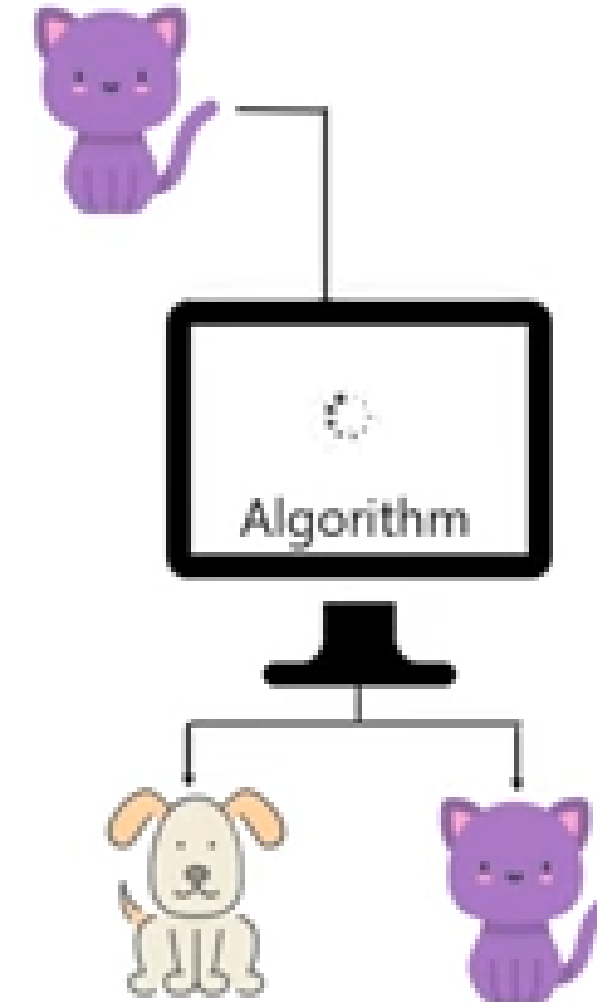
Supervised Learning

Labelled Data



Unsupervised Learning

Unlabelled Data



Supervised Learning

Forecast outcomes



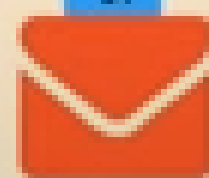
Unsupervised Learning

Discover underlying patterns



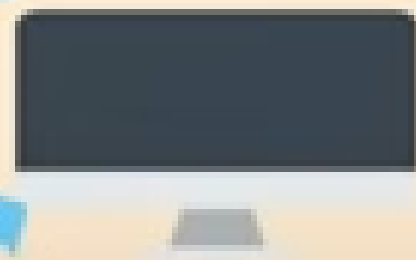
Classification

Not spam



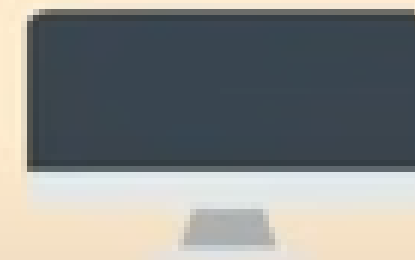
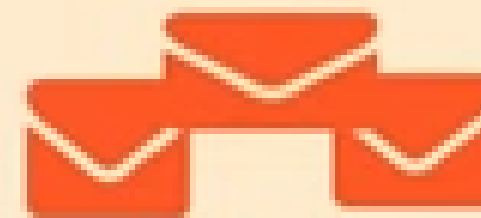
Spam

Spam Filtering



Learns

New mail



Scans the content

Categorical Separation



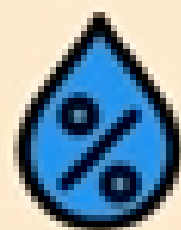
Not spam



Spam

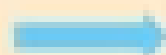
Regression

Humidity



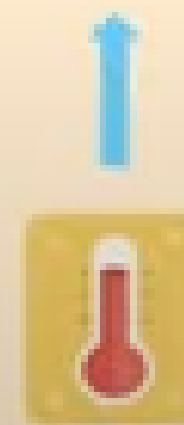
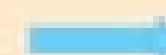
Temperature

Past Data



Learns

Prediction



New data

Clustering

The method of dividing the objects into clusters which are similar between them and are dissimilar to the objects belonging to another cluster

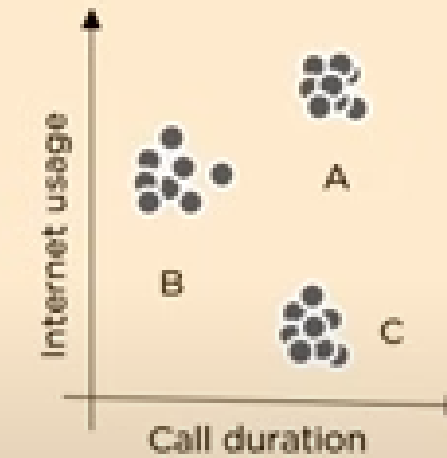
Clustering

Suppose a telecom company wants to reduce its customer churn rate by providing personalized call and data plans

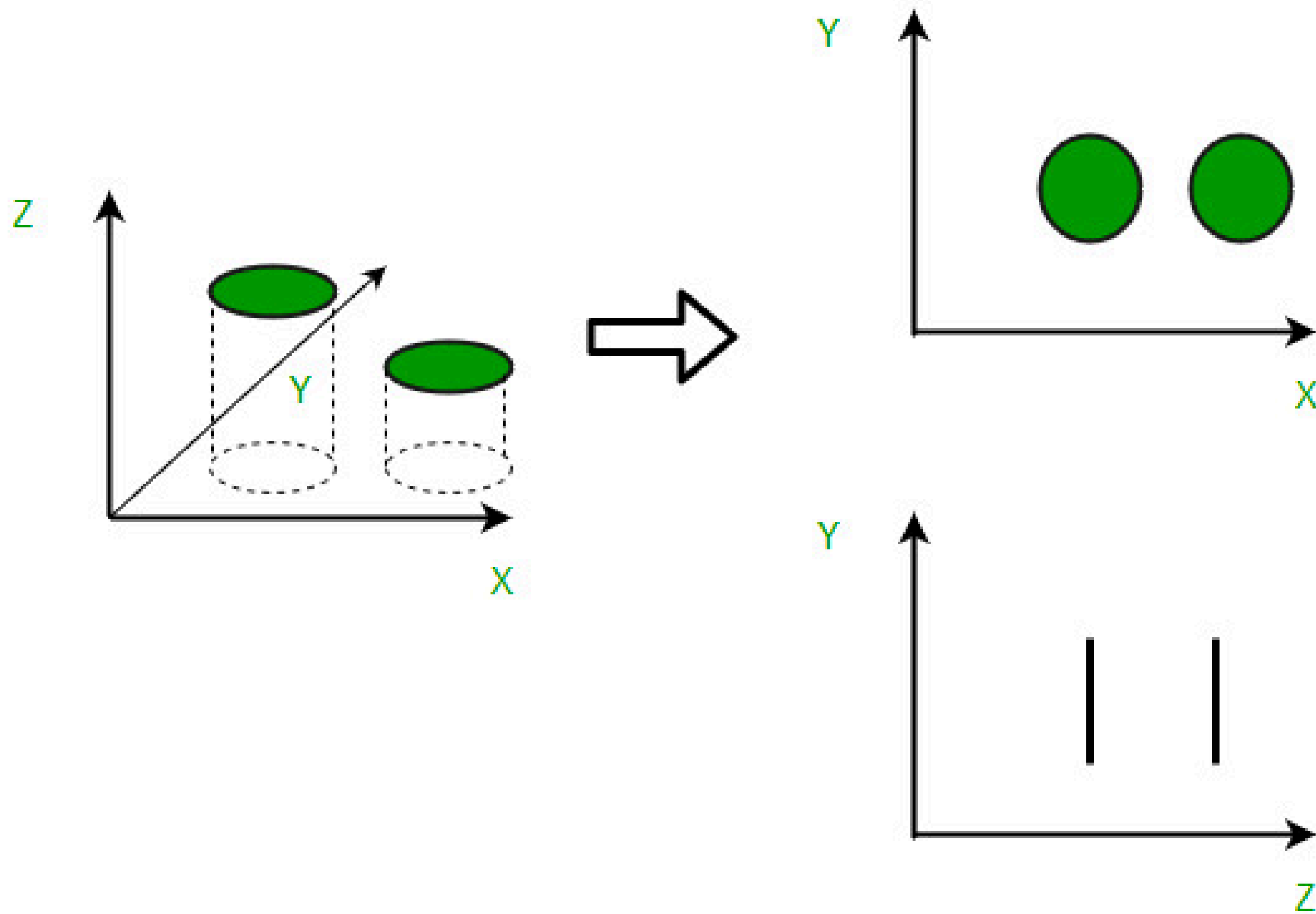

Total Call duration



Internet Usage



Dimensionality Reduction



Logistic Regression Model

Want $0 \leq h_{\theta}(x) \leq 1$

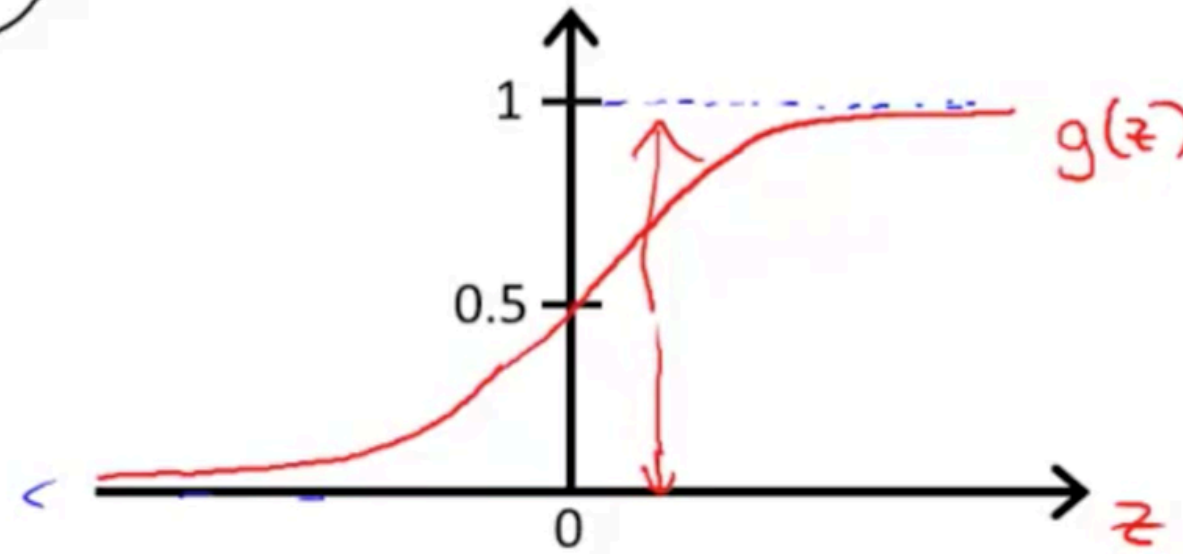
$$h_{\theta}(x) = g(\theta^T x)$$

$$\rightarrow g(z) = \frac{1}{1 + e^{-z}}$$

$\theta^T x$

→ Sigmoid function
→ Logistic function

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$



Parameters $\underline{\theta}$.

LINEAR REGRESSION

The thing we want
to explain

DEPENDENT
VARIABLE

y

i.e 77% of the variance in y is
explained by x. Below c.30% means
they're hardly connected. Above 95%
and they're practically the same.

$$R^2 = 0.77$$

If you only had data on x, this line
provides your best estimate of y. If the
fit is strong and no major outliers, x could
be used as a surrogate or forecast of y.

LINE OF BEST FIT

DATA
POINT

95% CONFIDENCE BAND

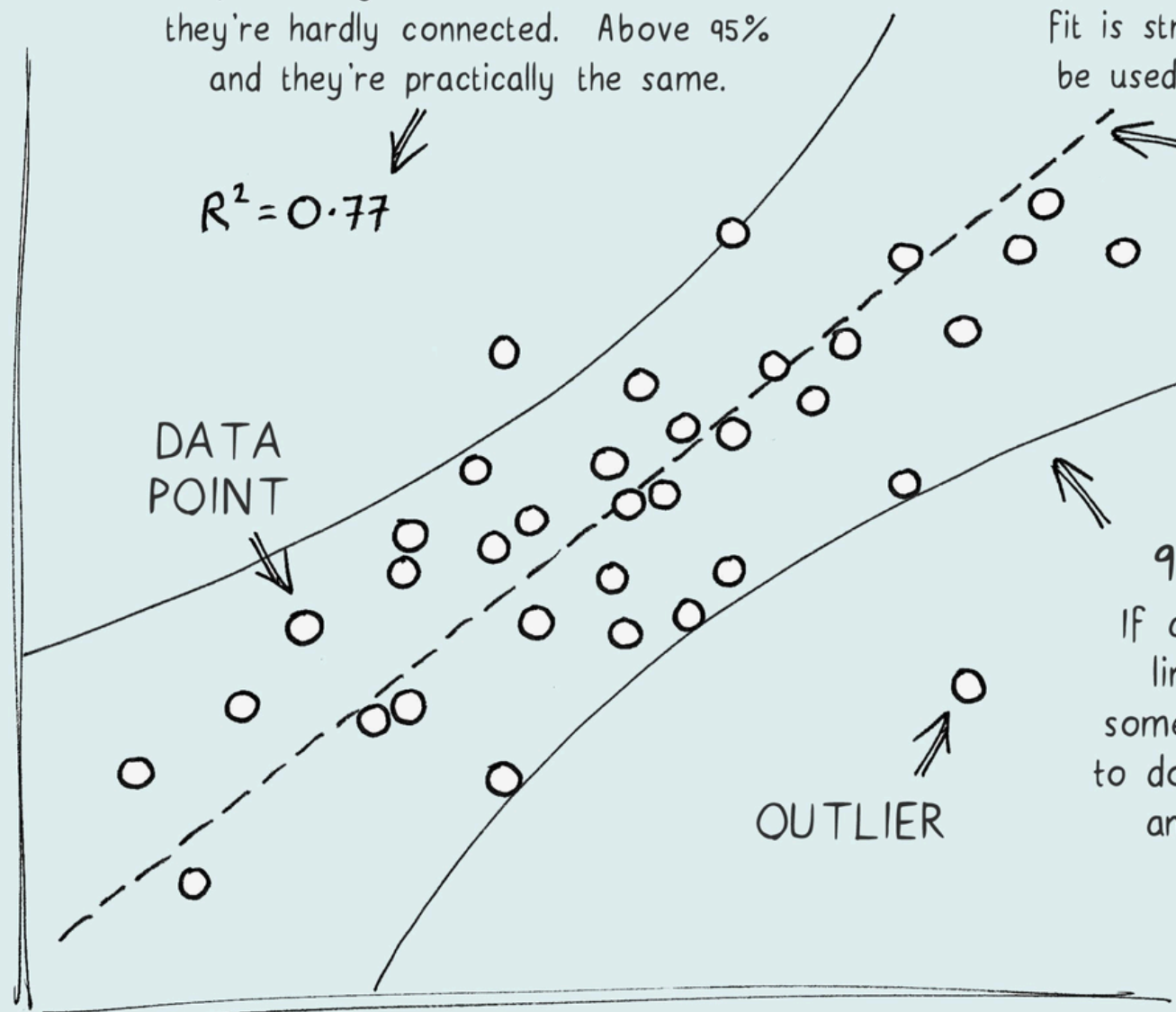
If a data point falls outside these
lines, you're 95% sure there is
something special about it causing it
to do better or worse than others -
an 'outlier' worth understanding

OUTLIER

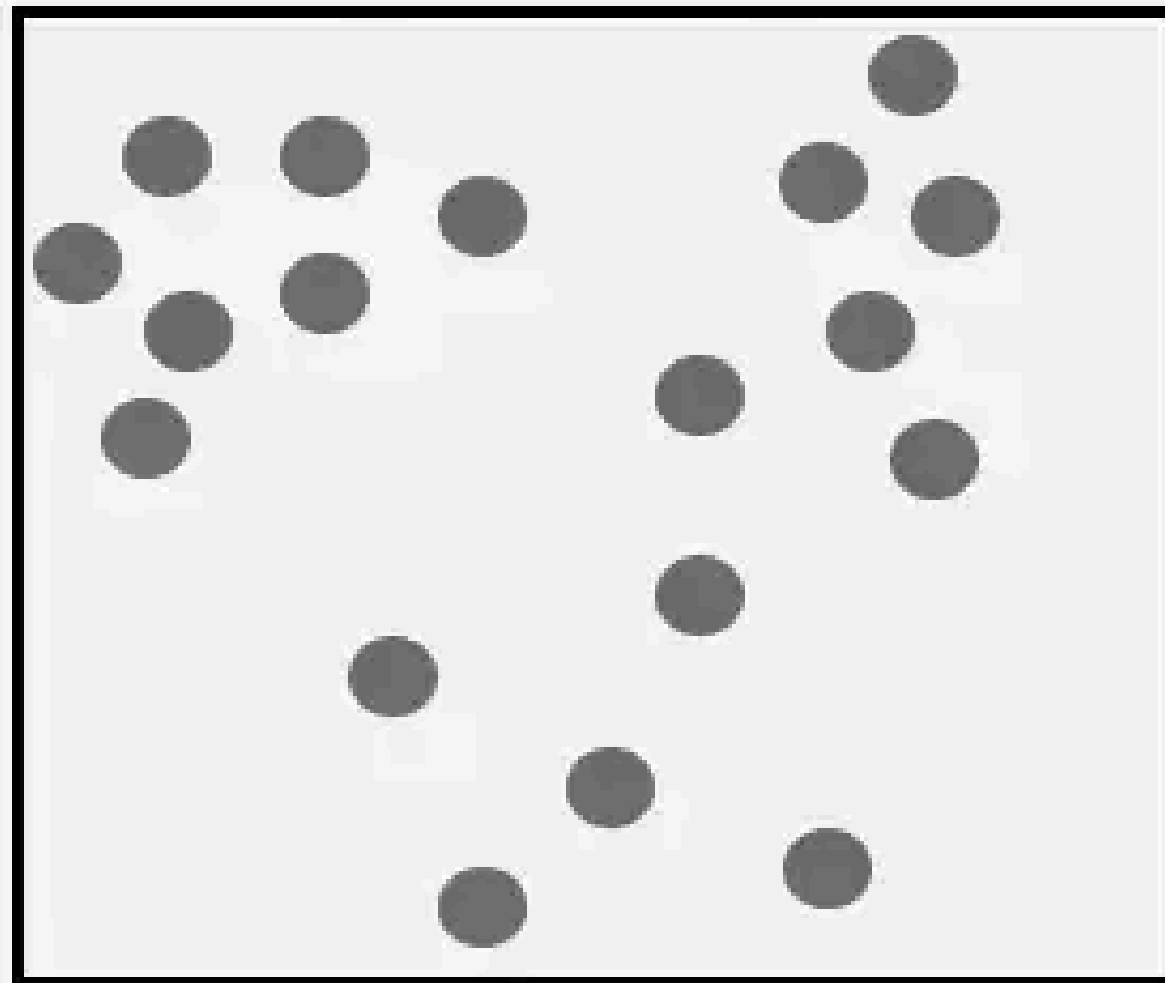
INDEPENDENT
VARIABLE

x

The factor we think
might influence the
dependent variable



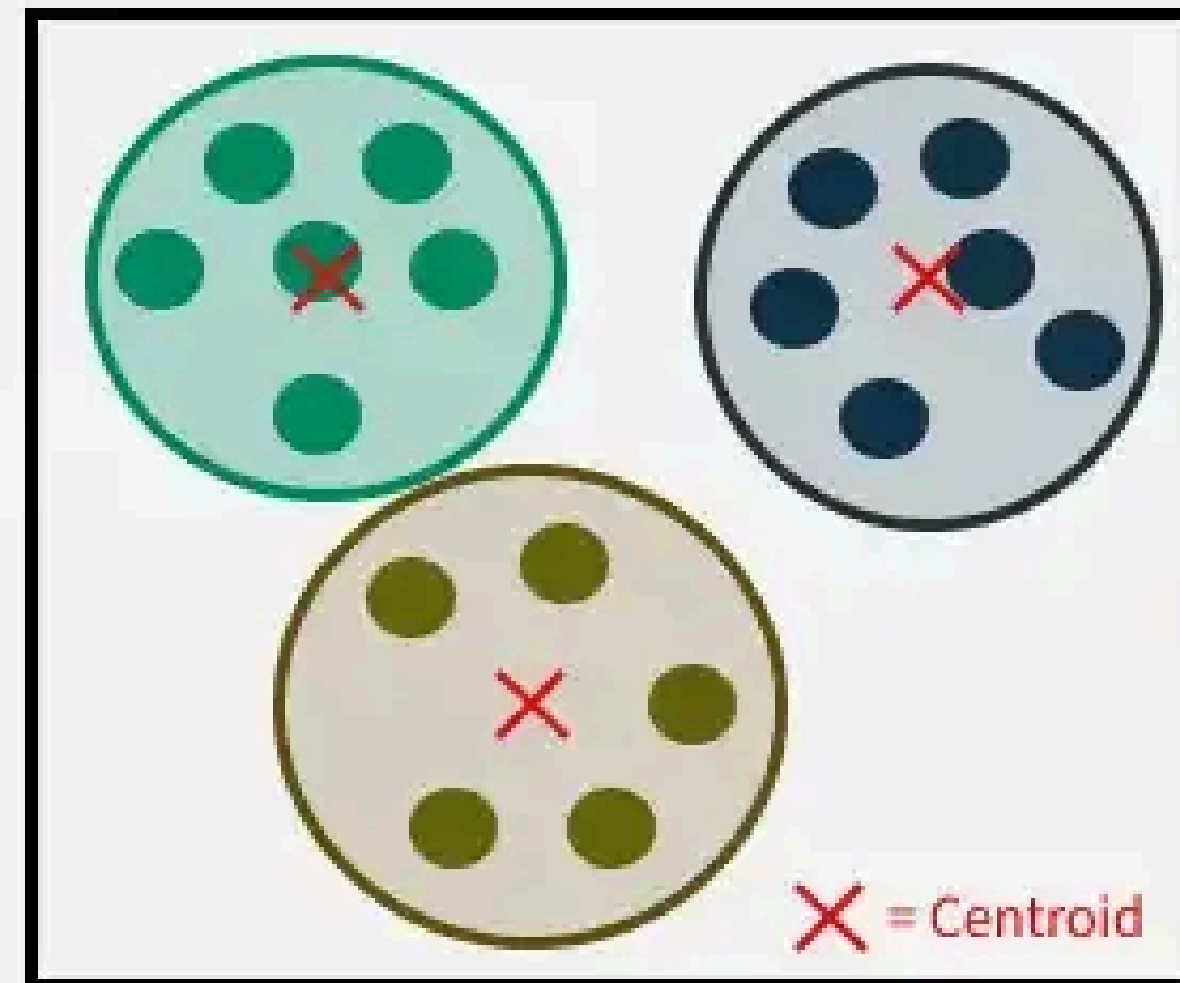
Unlabelled Data

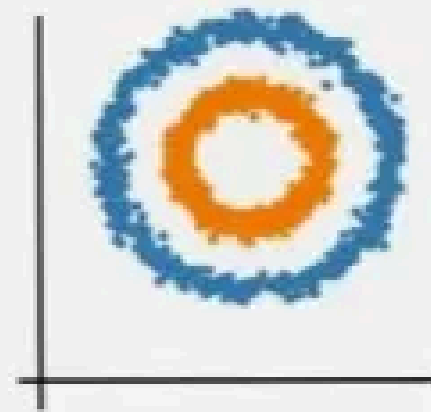


K Means

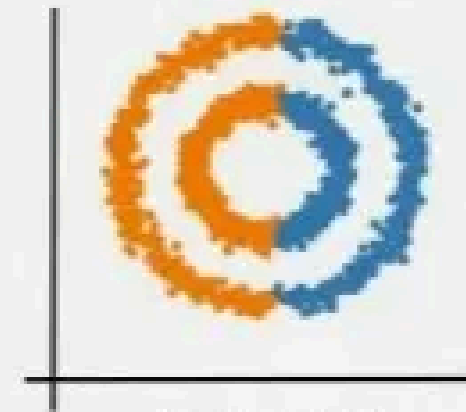


Labelled Clusters

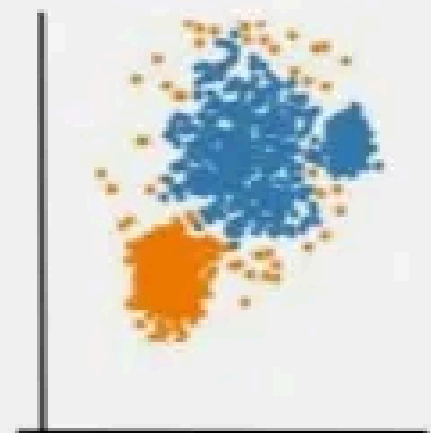




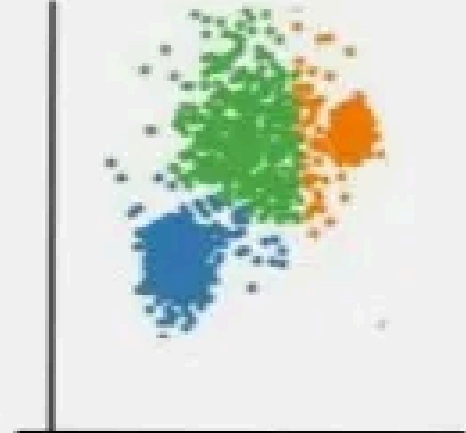
DBSCAN



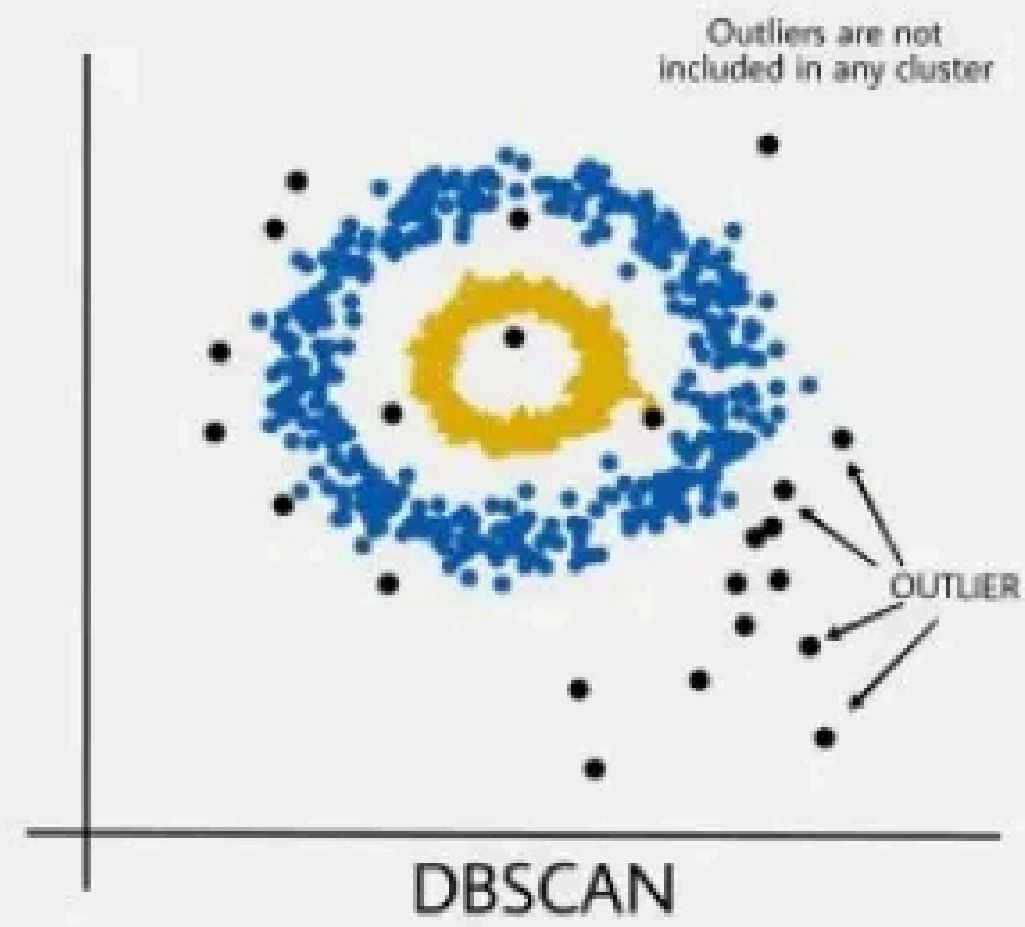
K-MEANS



DBSCAN

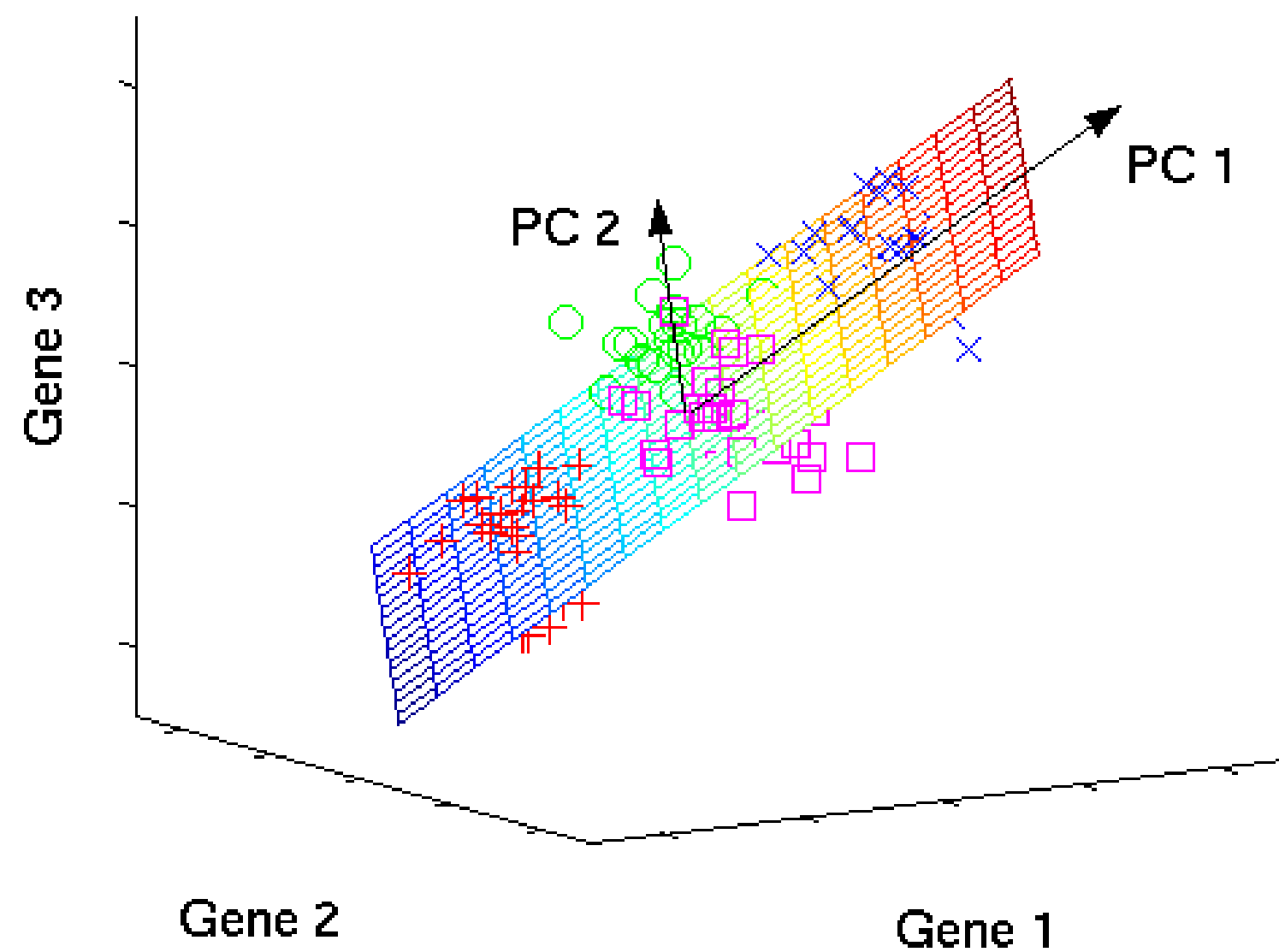


K-MEANS



DBSCAN

original data space



PCA



component space

