# Machine Learning

Plamen Kokanov, Miroslav Nenov
September 04, 2018

PUBLIC

THE BEST RUN SAP

# Agenda

Overview
- Definition
- Why now?
- Types of Data

Statistics

Supervised Algorithms
- Linear Regression
- Classification
- Neural Networks

Unsupervised Algorithms
- Clustering
- Ohters

# What is machine learning?

# Definition

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.”

*Arthur Samuel, 1959*

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

*Tom Mitchell, 1997*

# Why now?

- Increased computational power
  - According to Shane Legg, a cofounder of Google DeepMind: “Training run that takes one day on a single TPU device would have taken a quarter of a million years on an 80486 from 1990.”

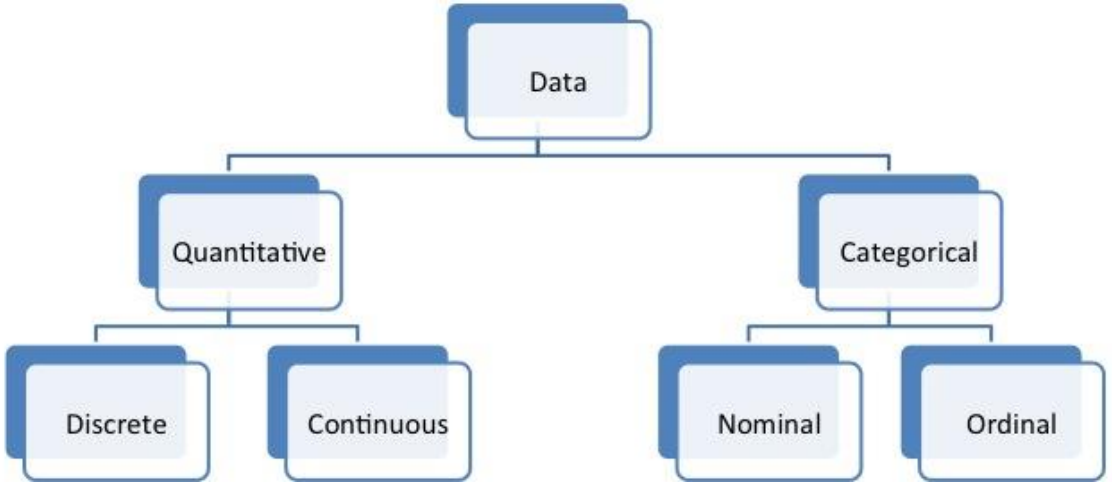
- A lot of data
  - There are 2.5 quintillion bytes of data created each day at our current pace, which will grow
  - Over the last two years alone 90 percent of the data in the world was generated.
  - Every minute we:
    - Users watch 4,146,600 YouTube videos
    - 456,000 tweets are sent on Twitter
    - Instagram users post 46,740 photos

- Improved algorithms
  - The algorithms and approaches that now dominate the discipline — such as deep supervised learning and reinforcement learning — share a vital basic property: Their results improve as the amount of training data they’re given increases.

# Types of Data

- **Qualitative (Categorical) Data** is data which represents characteristics like a person's gender, language etc.
  - **Nominal Data:** values represent discrete units and are used to label variables, that have no quantitative value;
  - **Ordinal Data:** values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters;

- **Quantitative (Numerical) Data**
  - **Discrete Data:** We speak of discrete data if its values are distinct and separate. This type of data **can't be measured but it can be counted**.
  - **Continuous Data:** it represents measurements and therefore their values **can't be counted but they can be measured**.

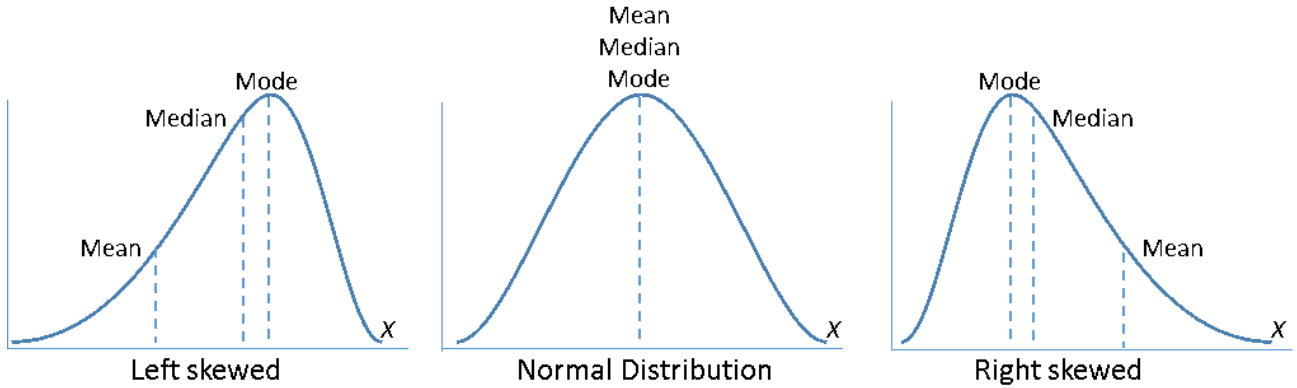


# Statistics

- Mean is given by the total of the values of the samples divided by the number of samples.

$$\bar{x} = \frac{1}{n}\Sigma x_i$$

- Median is the middle value of a sorted array.
- Mode represents the most common value in a data set.



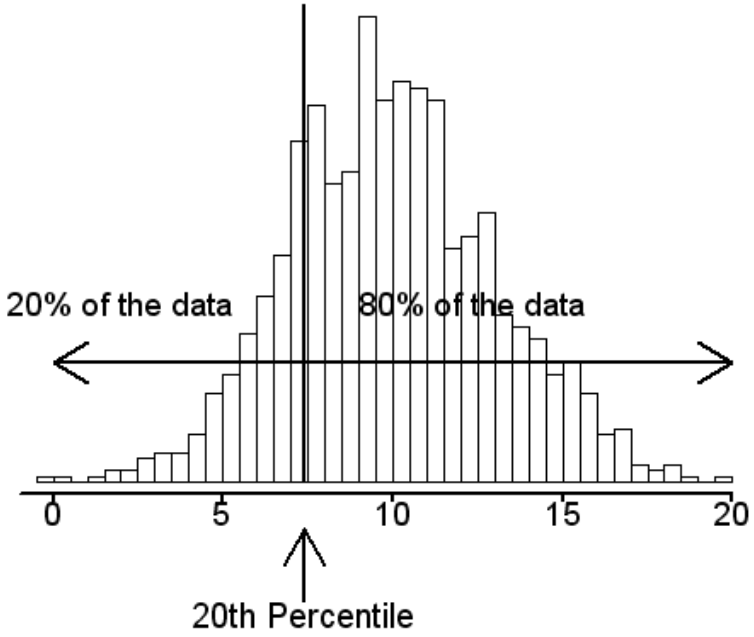

- Variance and Standard Deviation are essentially a measure of the spread of the data in the data set.
- Variance is the average of the squared differences from the mean.

$$\sigma^2 = \frac{1}{N}\Sigma(X - \mu)^2$$

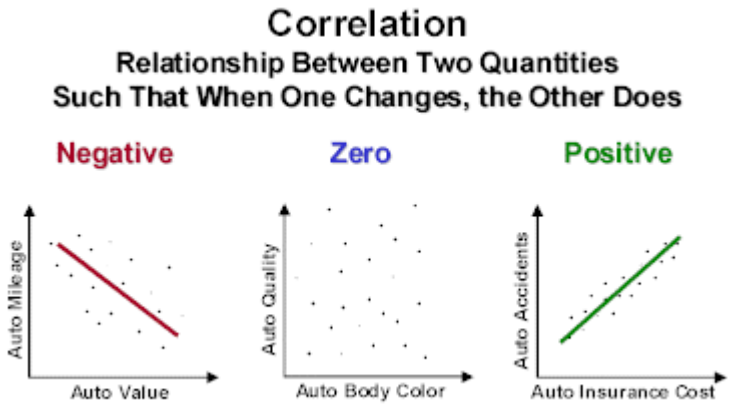
- Standard Deviation is the square root of the variance. It is an excellent way to identify outliers. Data points that lie more than one standard deviation from the mean can be considered unusual.

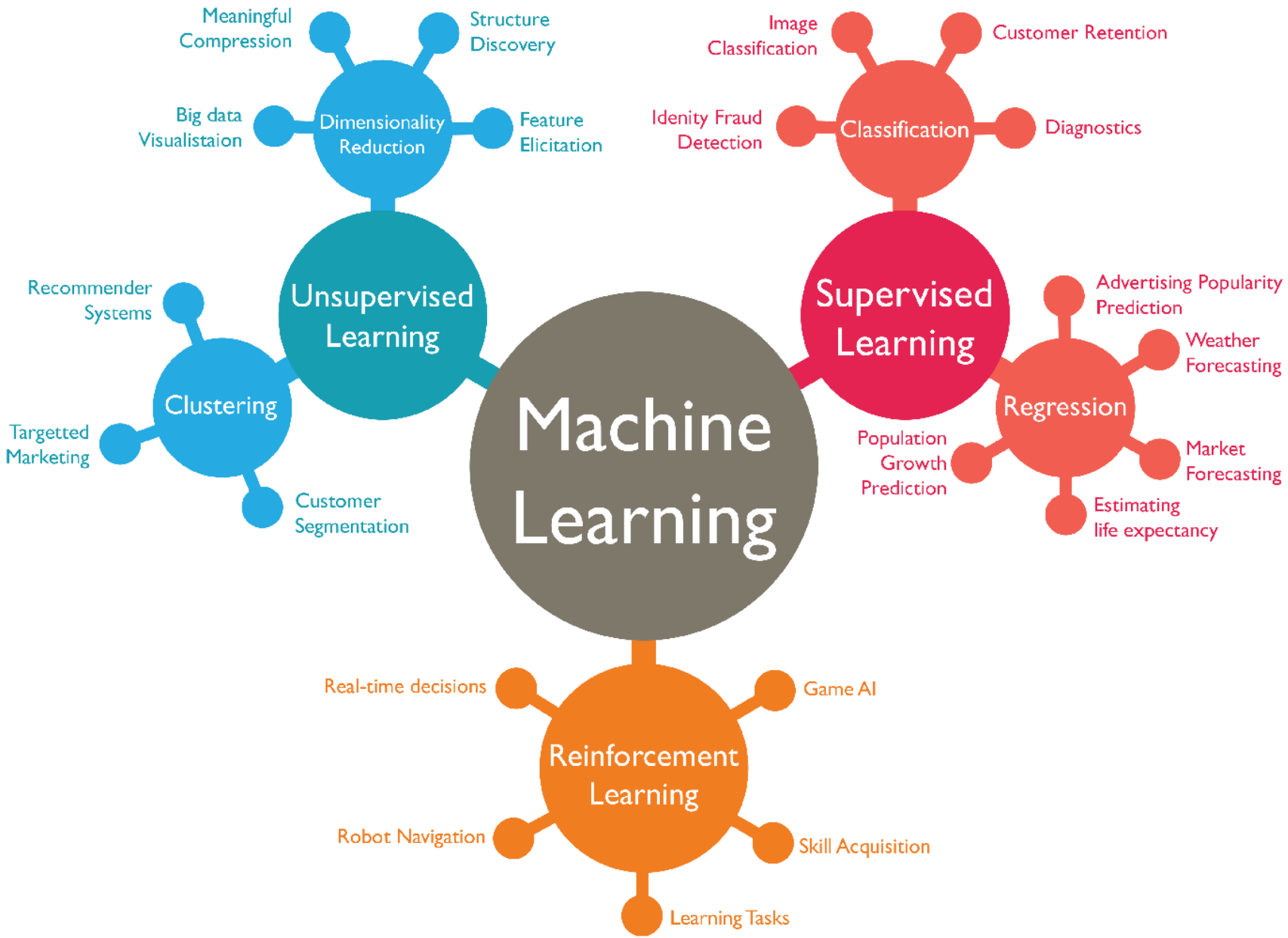
# Statistics

- A **percentile** is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.




- **Correlation** also measures how two variables move with respect to each other. A perfect positive correlation means that the correlation coefficient is 1. A perfect negative correlation means that the correlation coefficient is -1. A correlation coefficient of 0 means that the two variables are independent of each other



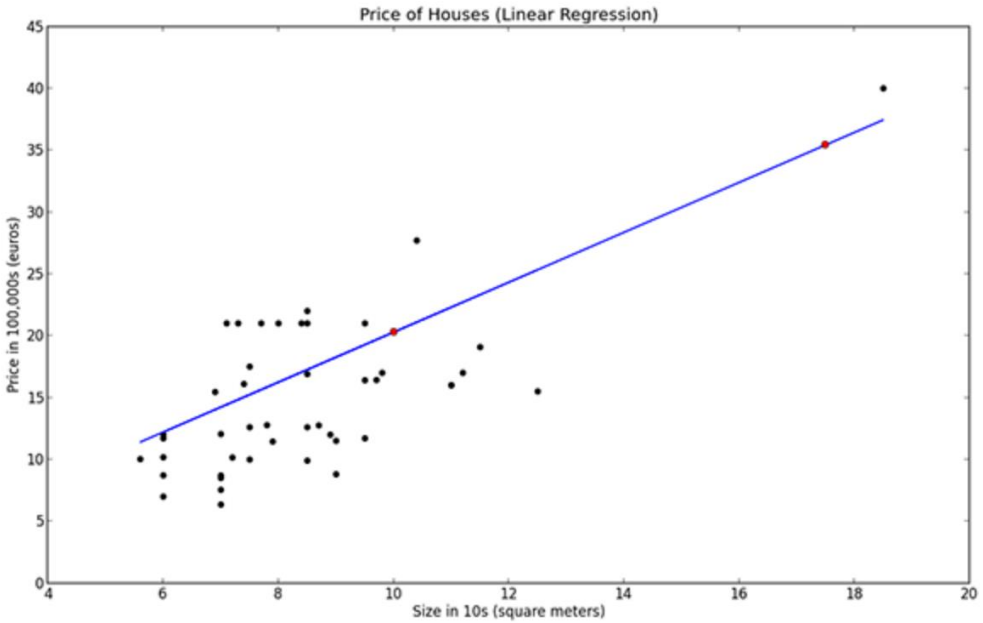


# Supervised Learning

# Linear Regression

- Finds the linear dependency between an input vector X (**feature**) and an output vector Y (**label**)
- Multiple **feature** vectors can be combined into a **matrix**
- **Labels** are most likely numerical

$y$



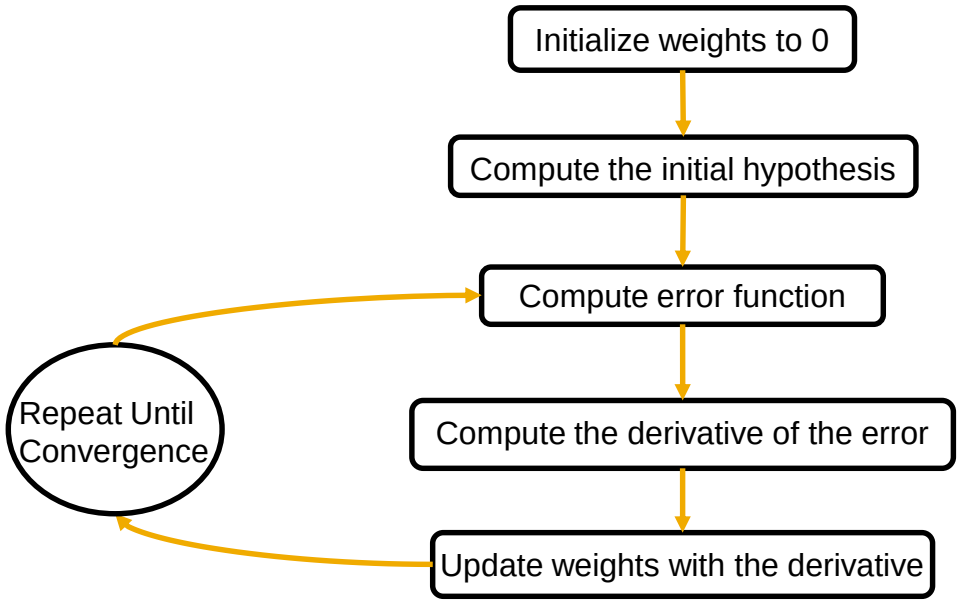
# Linear Regression

- Linear regression behind the scenes
  – Optimization objective

$$h(x^i) = w_0 + w_1 * x_1^i$$

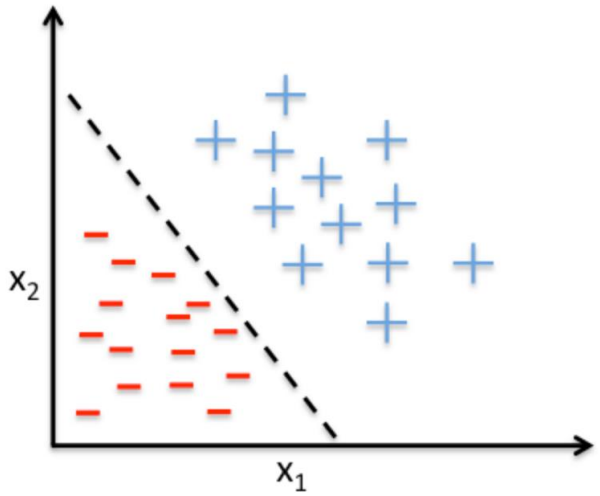
$$J(w_0, w_1) = \frac{1}{2m} * \sum_{i=0}^{i=m} \left(h(x^i) - y\right)^2$$

  – Gradient Descent Algorithm



# Classification (Logistic Regression)

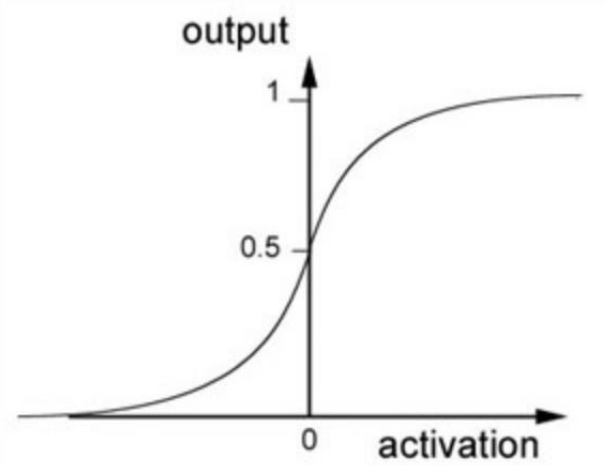
- Assigns classes (**labels**) to the data instances depending on their input **features**
- Classes **(labels)** are most commonly categorical (encoded with **0** and **1**)
- Can be either **binary** or **multiclass** classification



# Classification (Logistic Regression)

- Classification behind the scenes
  - Sigmoid Function

$$sig(x_1 x_2) = \frac{1}{1+e^{-(w_0+w_1*x_1+w_2*x_2)}}$$




  - Optimization Objective

$$h(x_1 x_2) = \text{sig}(x_1 x_2)$$

$$J(w_0 w_1 w_2) = \frac{1}{m} * \sum_{i=0}^{m} [y^i * \log(h(x_1^i x_2^i)) + (1 - y^i) * \log(1 - h(x_1^i x_2^i))]$$

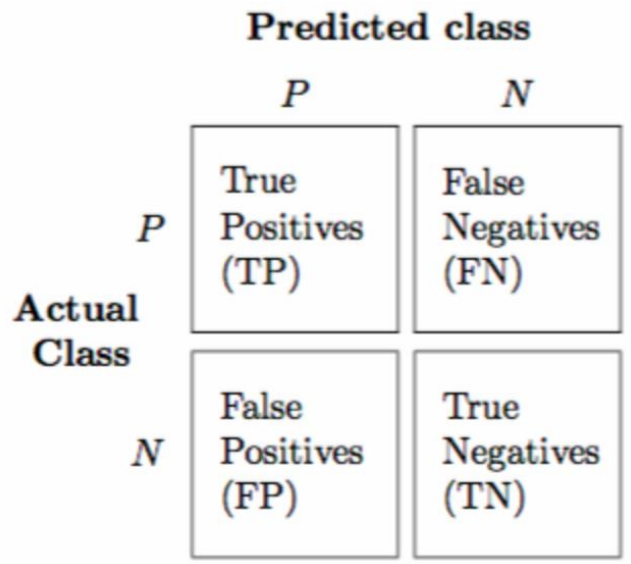
  - Decision Boundary
    - Separates the positive and negative classes

  - Gradient Descent
    - Same as with linear regression

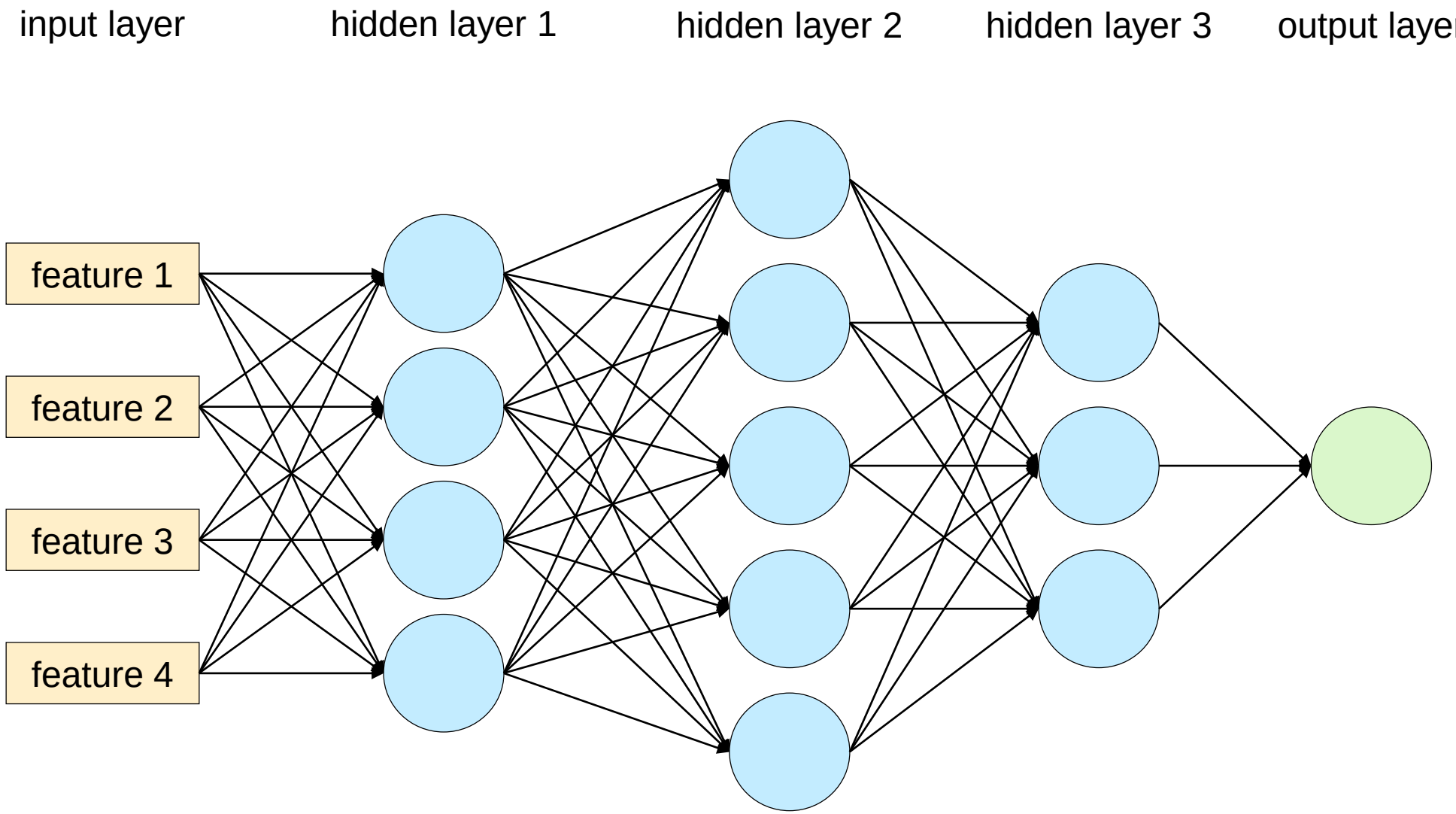
# Demo

# Evaluating A Classification Model

- Loss
- Confusion Matrix
- True Positive and False Positive
- True Negative and False Negative
- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1 Score = $2 * \frac{Precision * Recall}{Precision+Recall}$



# Densely Connected Neural Networks



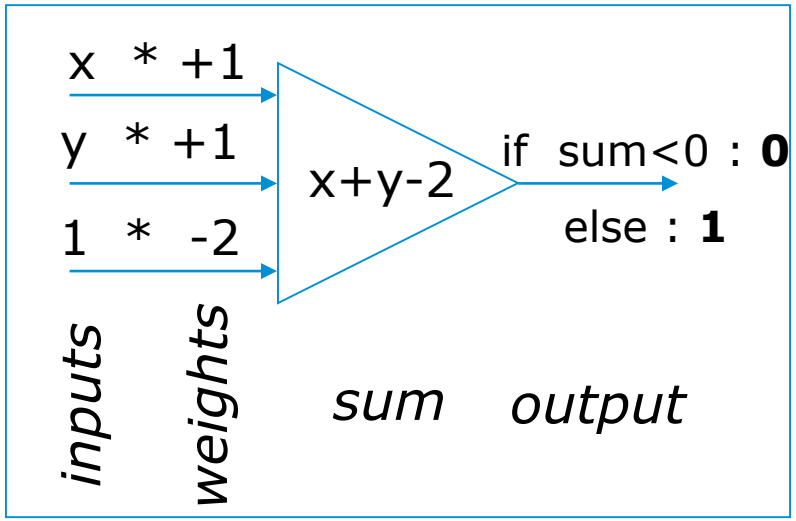
# The Neuron

The general artificial neuron model has five components, shown in the following list. (The subscript i indicates the i-th input or weight.)

1. **A set of inputs, $x_i$.**
2. **A set of weights, $w_i$.**
3. **A bias, u.**
4. **An activation function, f.**
5. **Neuron output**, y

# The Neuron

Simple artificial “neurons” could be made to perform basic logical operations such as AND, OR and NOT.



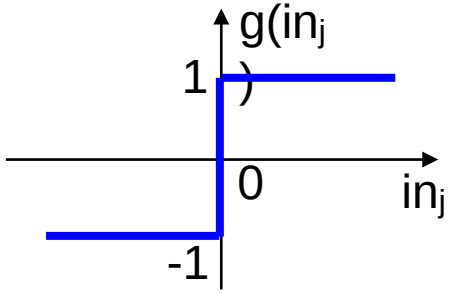

**Truth Table for Logical AND**

| x | y | x & y |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

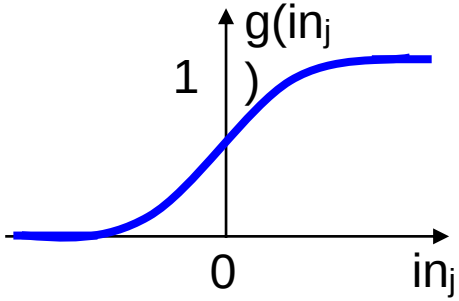
*inputs* *output*

# Activation functions

- Sign function (sometimes step function or threshold)



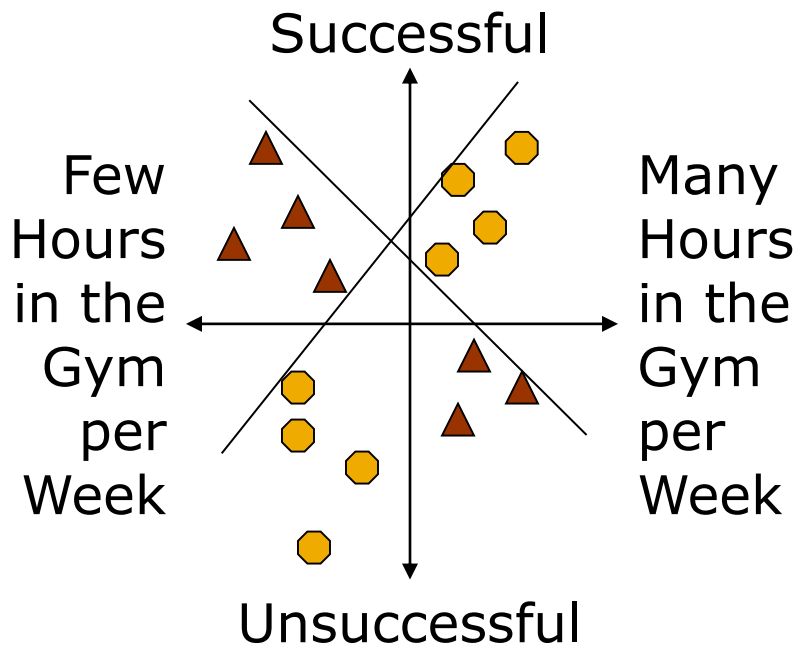

- Sigmoid function $1/(1+e^{-x})$



# The Fall of the Perceptron

- Before long researchers had begun to discover the Perceptron's limitations.
- Unless input categories were "linearly separable", a perceptron could not learn to discriminate between them.
- Unfortunately, it appeared that many important categories were not linearly separable.
- E.g., those inputs to an XOR gate that give an output of 1 (namely 10 & 01) are not linearly separable from those that do not (00 & 11).

# The Fall of the Perceptron




- Athletes
- Academics

…despite the simplicity of their relationship:

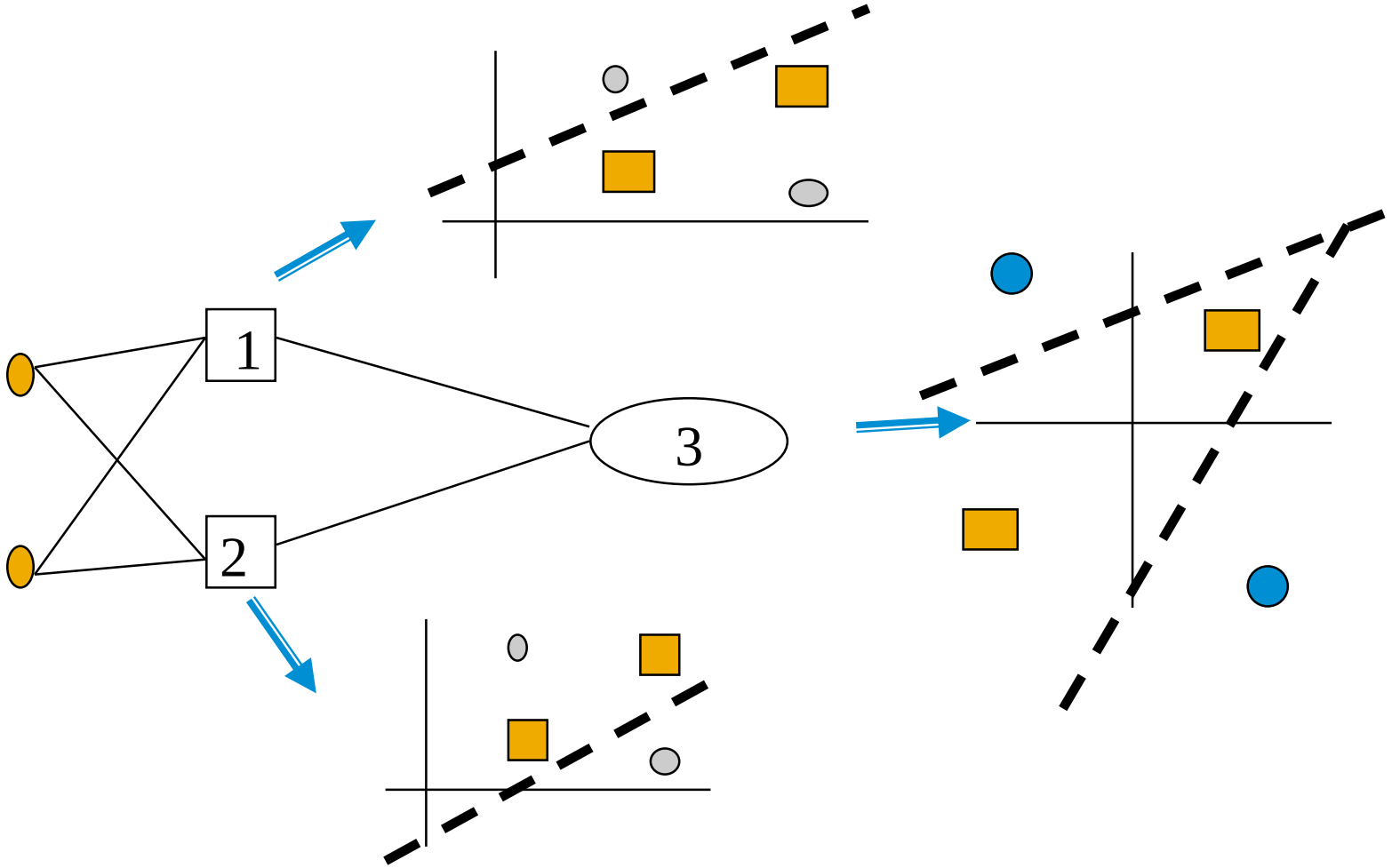
Academics = Successful XOR Gym

In this example, a perceptron would not be able to discriminate between the athletes and the academics…

This failure caused the majority of researchers to walk away.

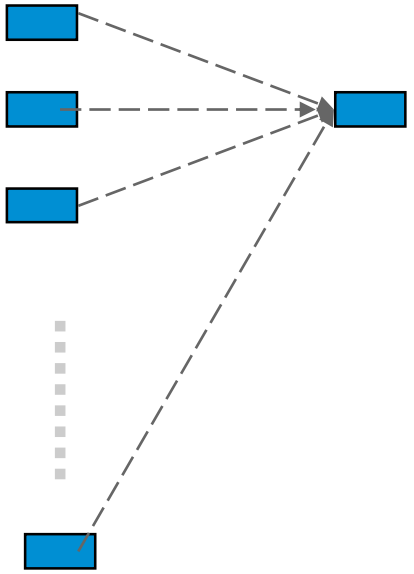
# The Multi-layer Perceptron

Minsky & Papert (1969) offered solution to XOR problem by combining perceptron unit responses using a second layer of units

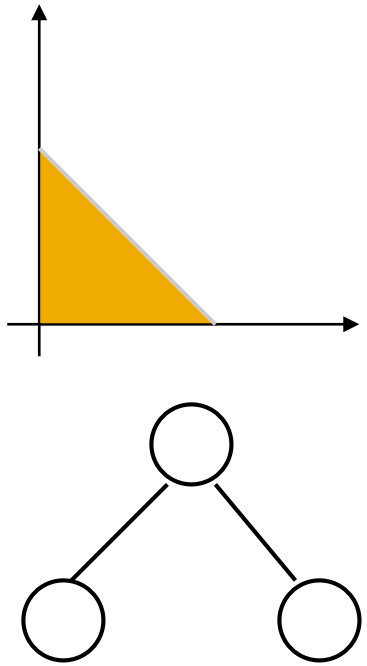


# Properties of architecture

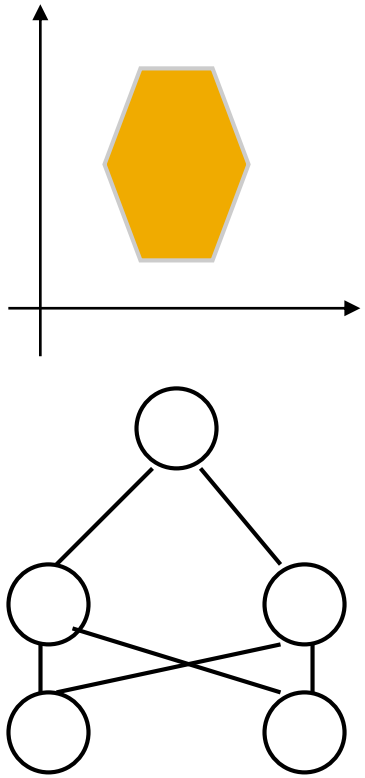
- No connections within a layer
- No direct connections between input and output layers
- Fully connected between layers
- Often more than 3 layers
- Number of output units need not equal number of input units
- Number of hidden units per layer can be more or less than input or output units



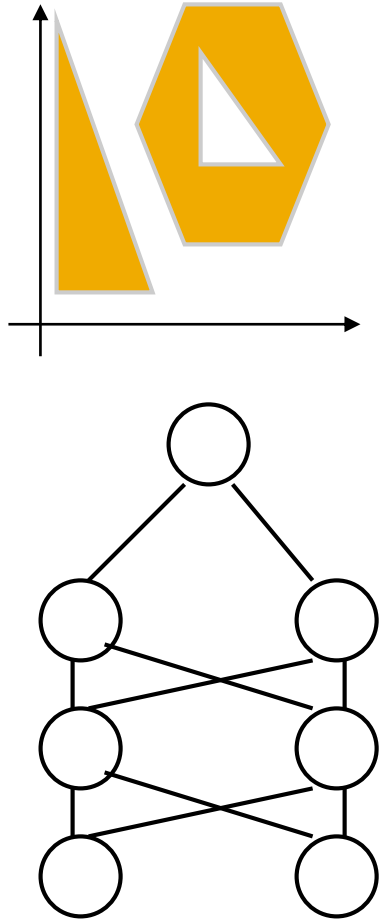
# What do each of the layers do?



1st layer draws linear boundaries



2nd layer combines the boundaries



3rd layer can generate arbitrarily complex boundaries

# Backpropagation used for training

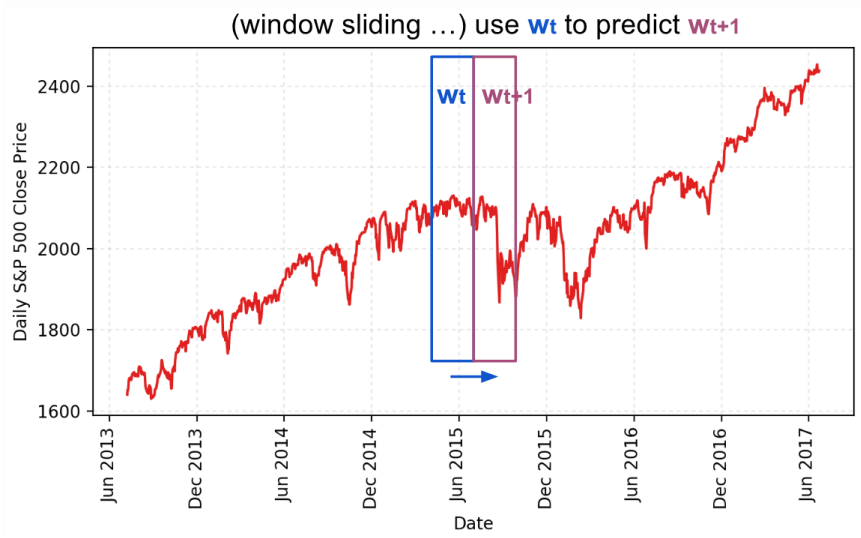
**The Backpropagation has two phases**:

**Forward pass phase:** computes ‘functional signal’, feedforward propagation of input pattern signals through network

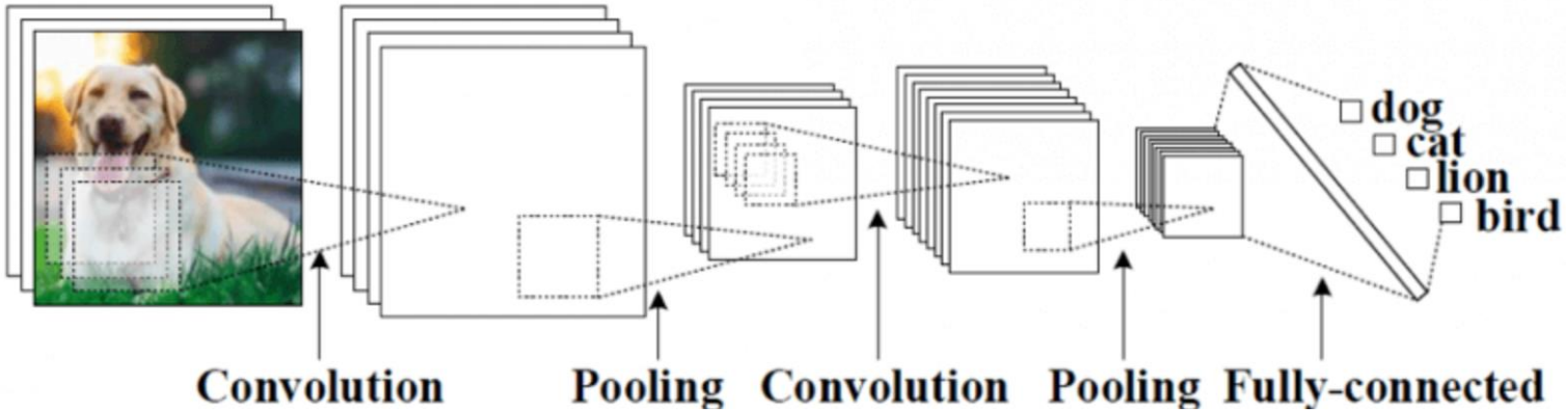
**Backward pass phase:** computes ‘error signal’, *propagates* the error *backwards* through network starting at output units (where the error is the difference between actual and desired output values)

# State of the Art Neural Networks

- Recurrent Neural Networks
  - Long Short-Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)



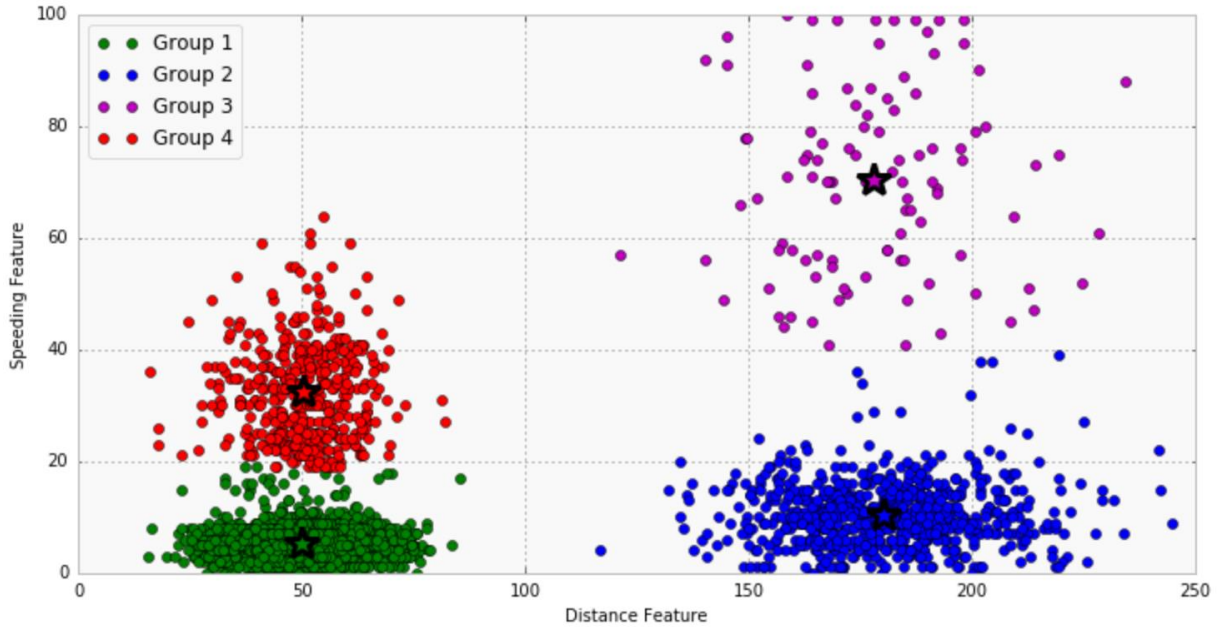

- Convolutional Neural Networks



# Unsupervised Learning

# K-means Clustering

- Separates the data instances into **k** clusters
- Distance between each point and the centers of the clusters is calculated via **Euclidian Distances**
- During training the centers move so that these **distances** are minimized



# Demo

# Other Unsupervised Algorithms

- Anomaly Detection
- Principal Component Analysis
- Independent component analysis

# Thank you.

Contact information:

**Plamen Kokanov**
plamen.kokanov@sap.com

**Miroslav Nenov**
miroslav.nenov@sap.com