La Trobe Business School

BUS5PA - Predictive Analytics

Topic 5 – Predictive Modelling with K-Nearest Neighbours and Neural Networks

Learning Objectives

- Understand the need for learning different types of data modelling techniques
- Learn the basic ideas of K-Nearest Neighbours
- Understand how K-Nearest Neighbours can be used for predictive modelling
- Learn the basic ideas behind artificial neural networks
- Understand how neural networks can be used to model different situations represented by data
- Introduce Deep Neural Networks

LA TROBE UNIVERSITY

Predictive Models

Predicts target variable using statistics

Target variable can be the future event or any unknown target

PREDICTIVE MODELS/ALGORITHMS

Regression Models

Target is a continuous variable

1. Linear Regression (parametric)
2. Decision Tree (non-parametric)
3. K nearest neighbours
4. Artificial neural network
5. ...

Classification Models

Target is a categorical/discrete variable

1. Logistic Regression (parametric)
2. Decision Tree (non-parametric)
3. K nearest neighbours
4. Artificial neural network
5. ...

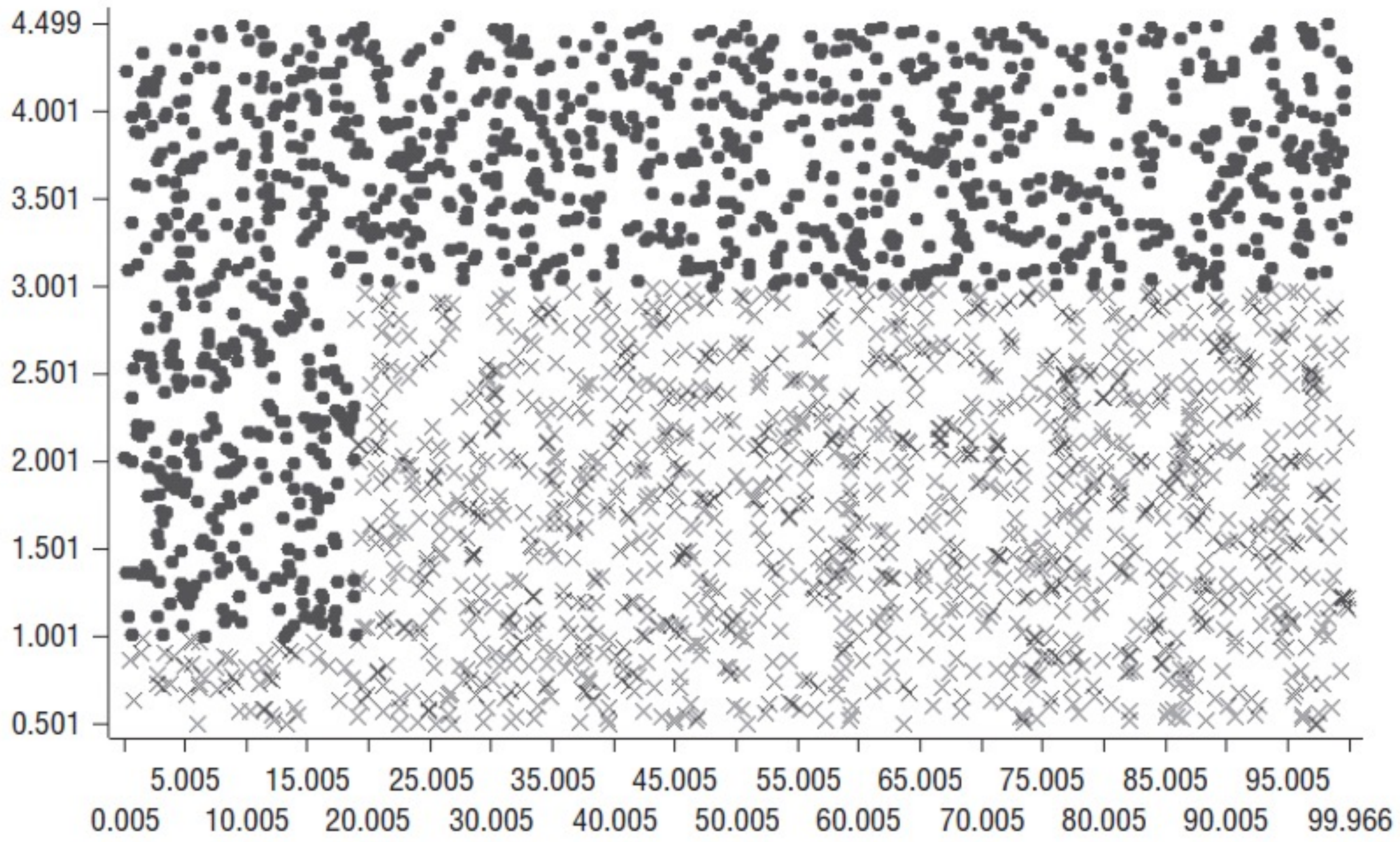
Predictive Modelling Techniques

- So far we have looked at supervised segmentation (decision trees) and fitting a model to data using the numeric functions most commonly used in data analytics linear and logistics regression models.

- Different data can have diverse distributions and underlying associations and relationships which cannot be represented by the above techniques.

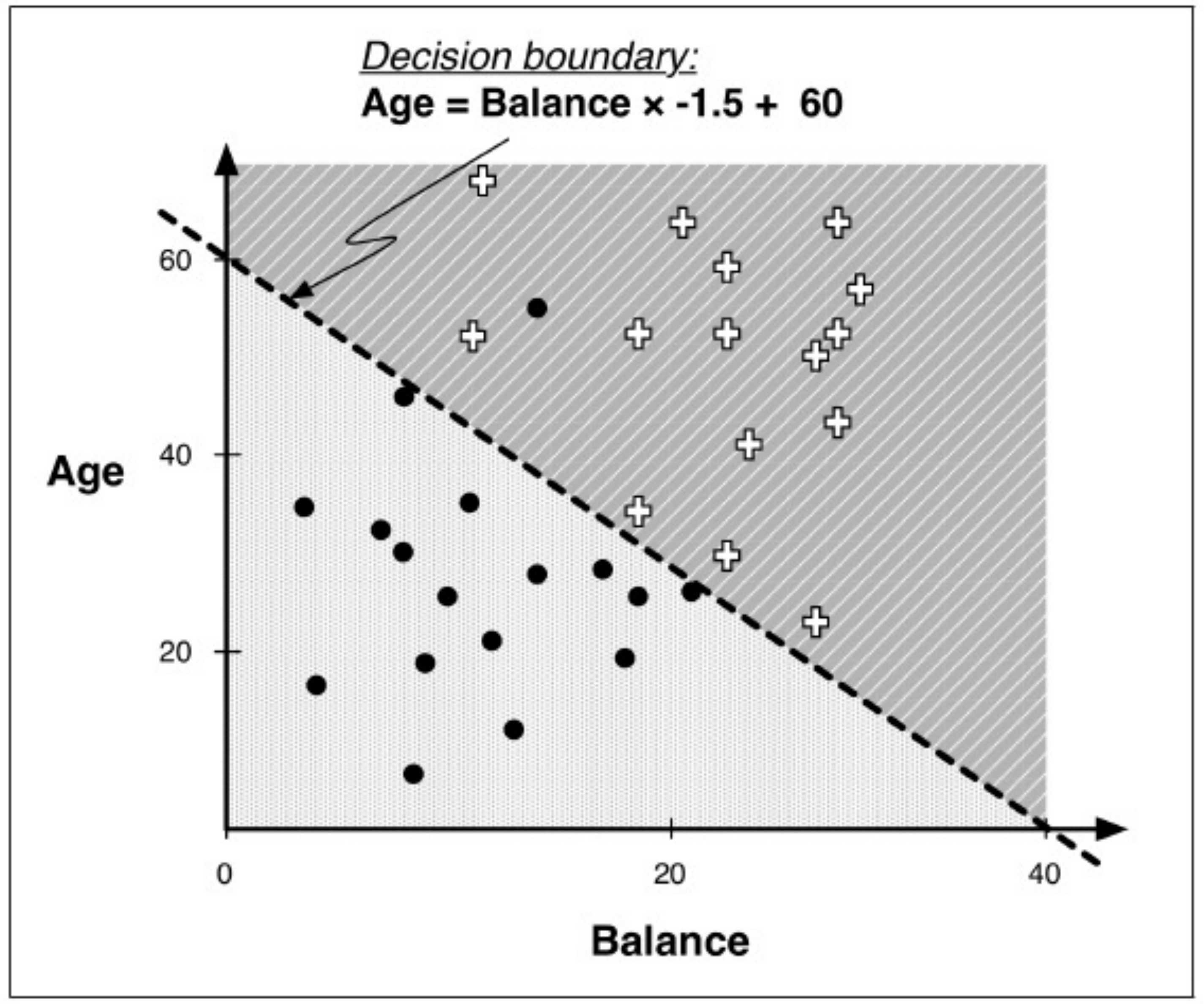
- It is useful to have a variety or a range of tools and techniques at your disposal.

Modelling Different Situations



A classification tree and the partitions it imposes in instance space.

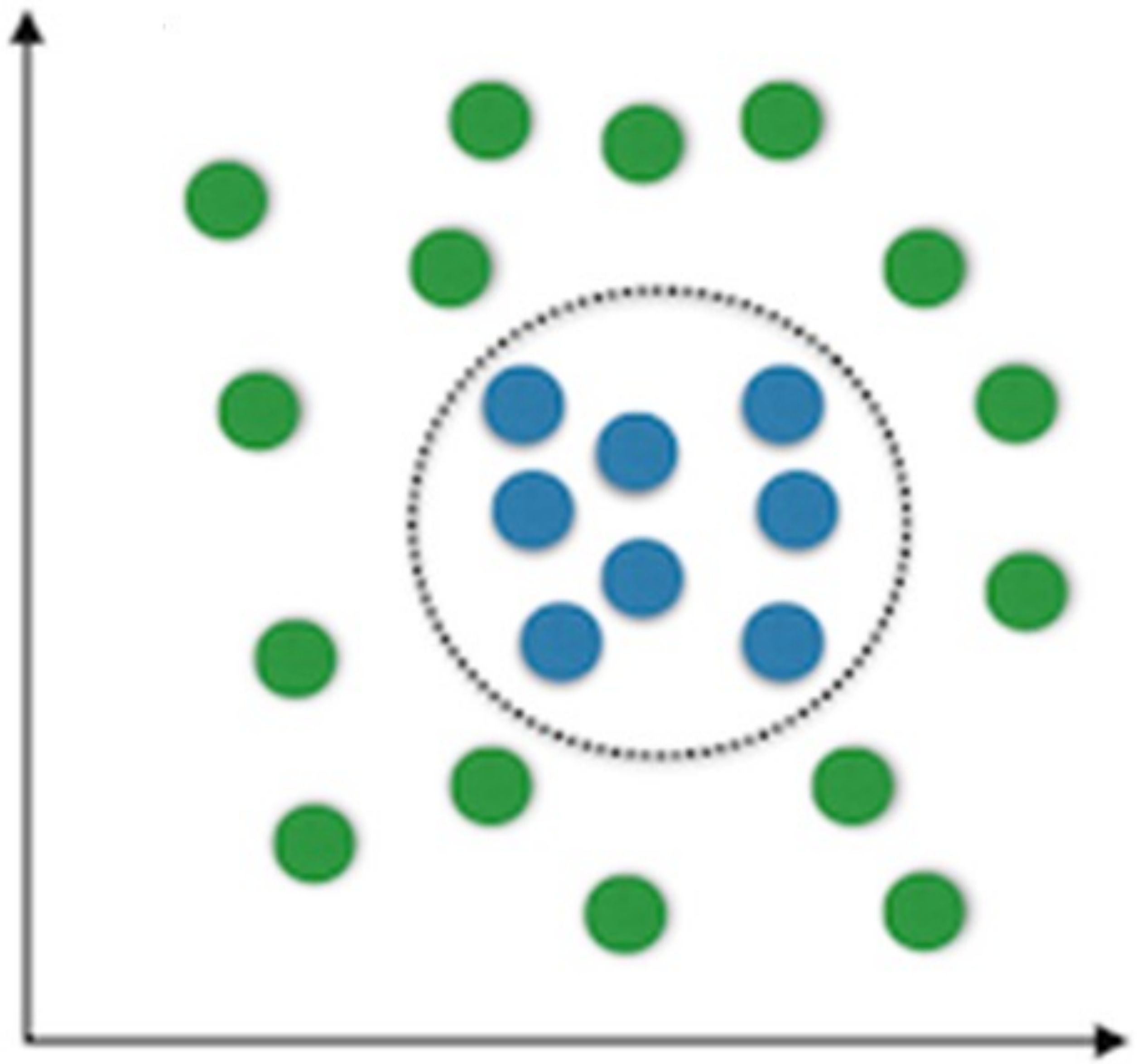
We can use decision tree to model this situation



The dataset with a single linear split.

We can use linear model to model this situation

How about this ?



1. K Nearest Neighbour

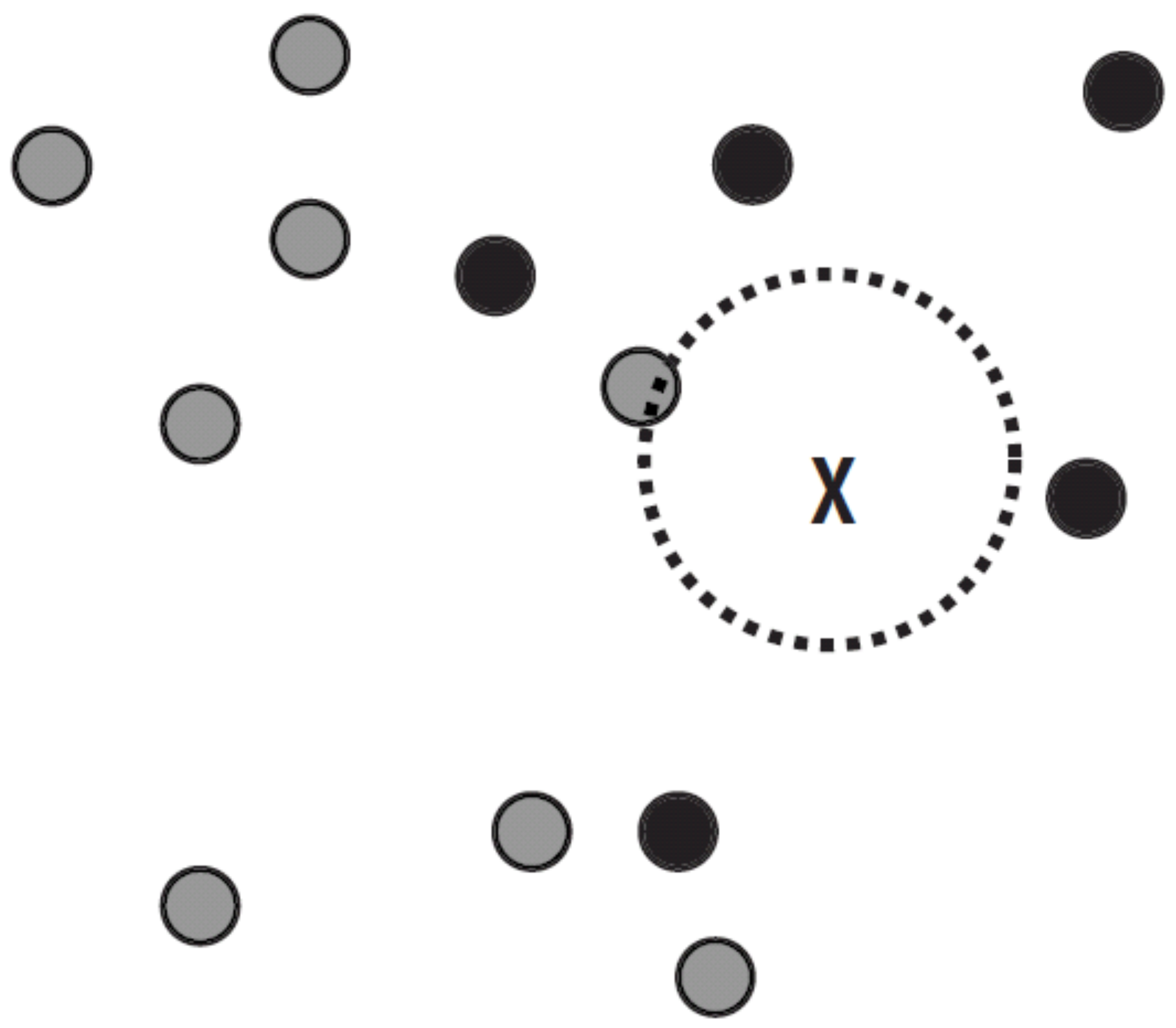
- The K nearest neighbour algorithm is a non-parametric algorithm

- k-Nearest neighbour is an example of instance-based learning, in which the training data set is stored, so that a classification for a new unclassified record may be found simply by comparing it to the most similar records in the training set.

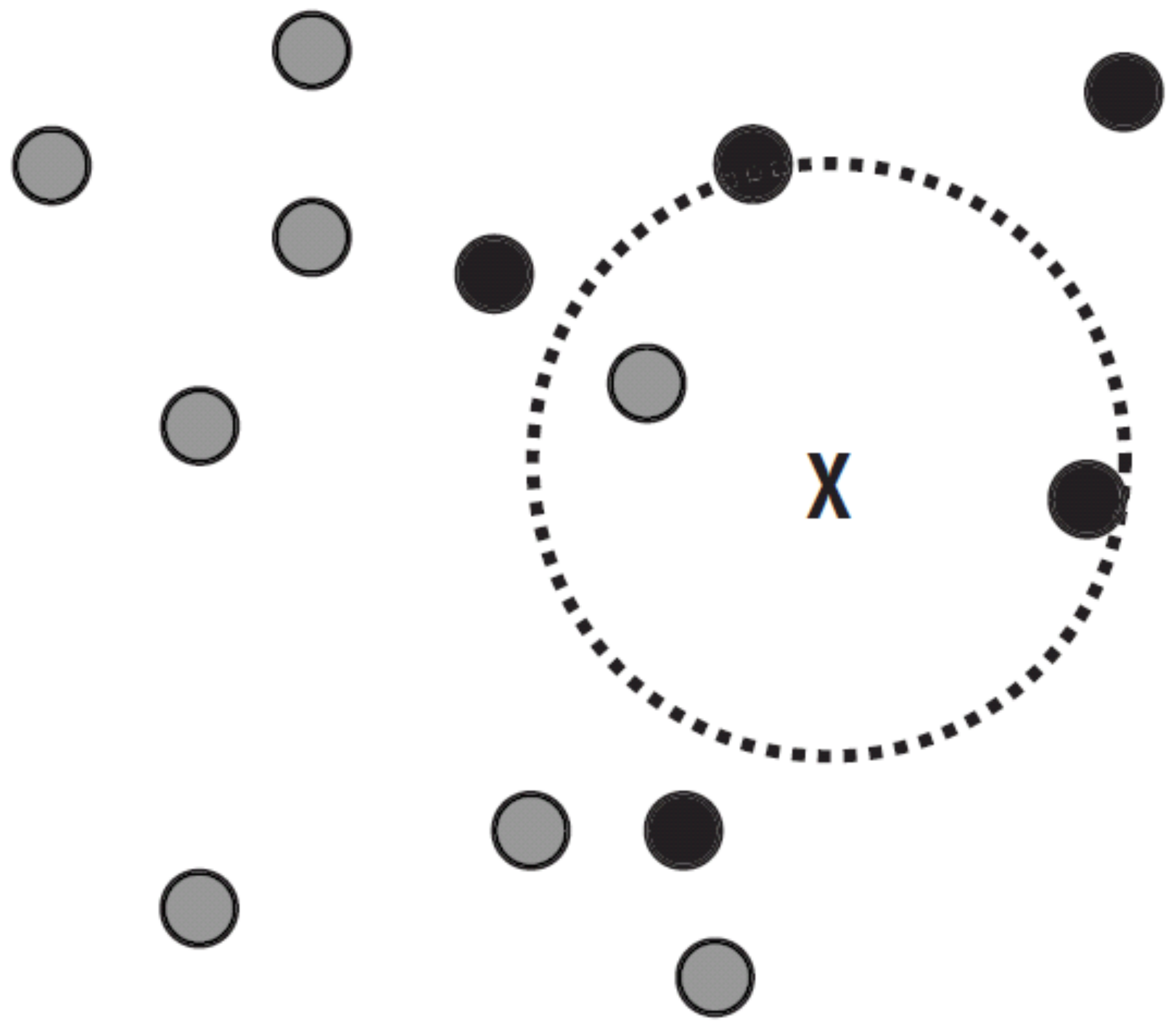
- It is a so-called “lazy learner,” meaning that there is little done in the training stage. The training data is the model. In essence, the nearest neighbour model is a lookup table.

K Nearest Neighbour

- For a new record, the k-nearest neighbour algorithm assigns the classification of the most similar record or records.

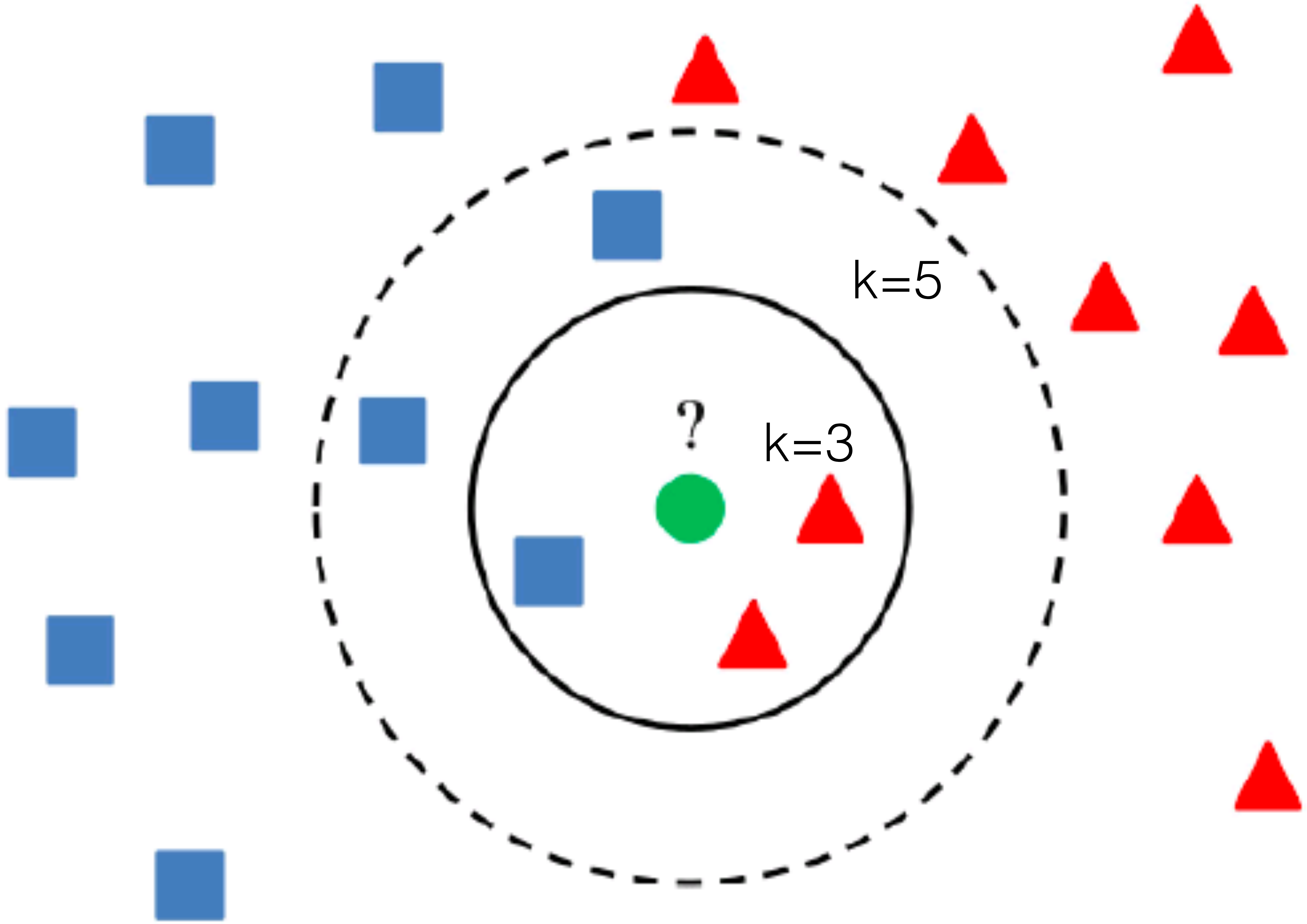


k=1

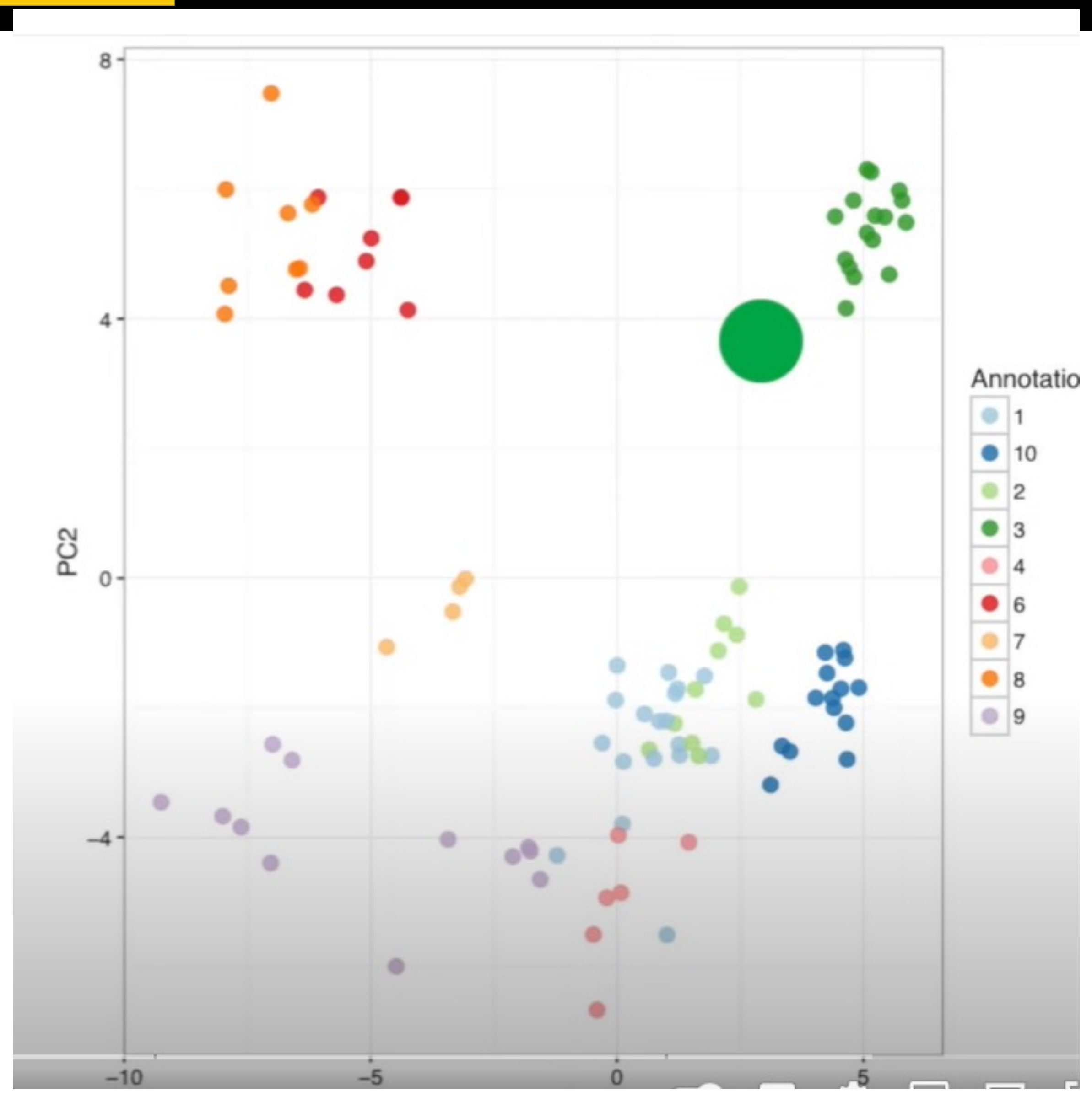


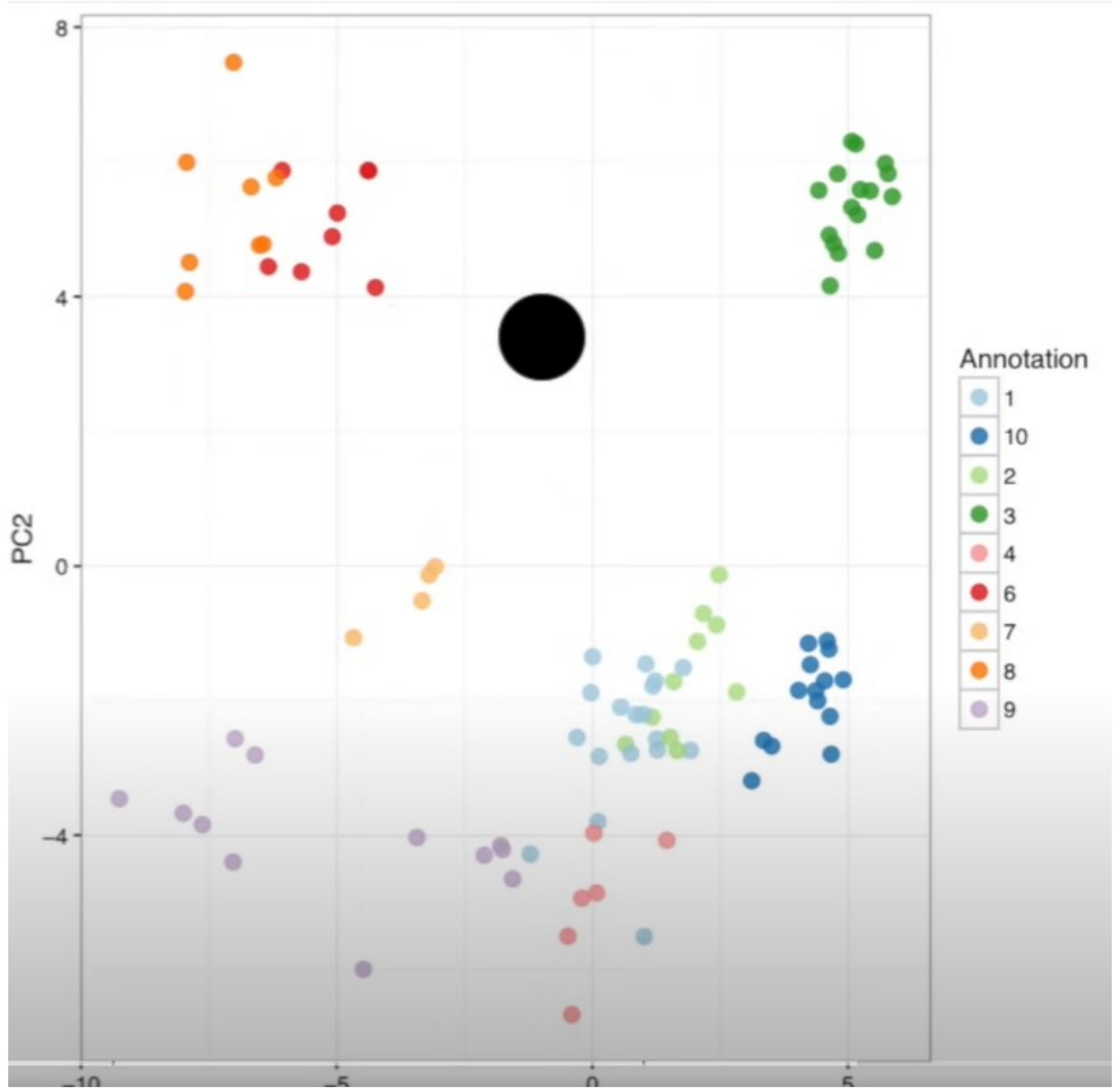
k=3

K Nearest Neighbour



K Nearest Neighbour





If K=11 and the new cell is between two (or more) categories, we simply pick the category that "gets the most votes".

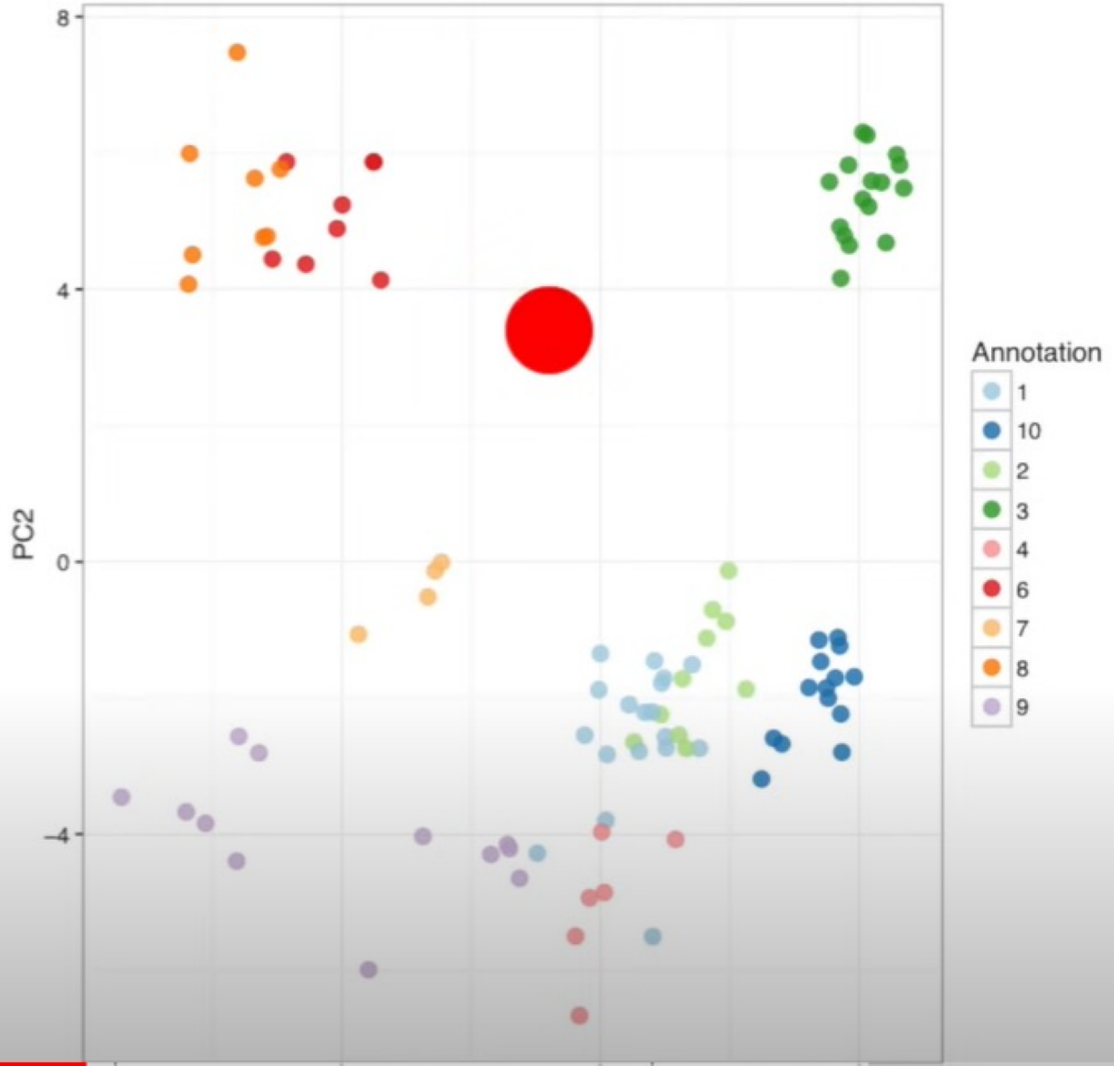
In this case....

7 nearest neighbors are **RED**.

3 nearest neighbors are **ORANGE**.

1 nearest neighbor is **GREEN**.

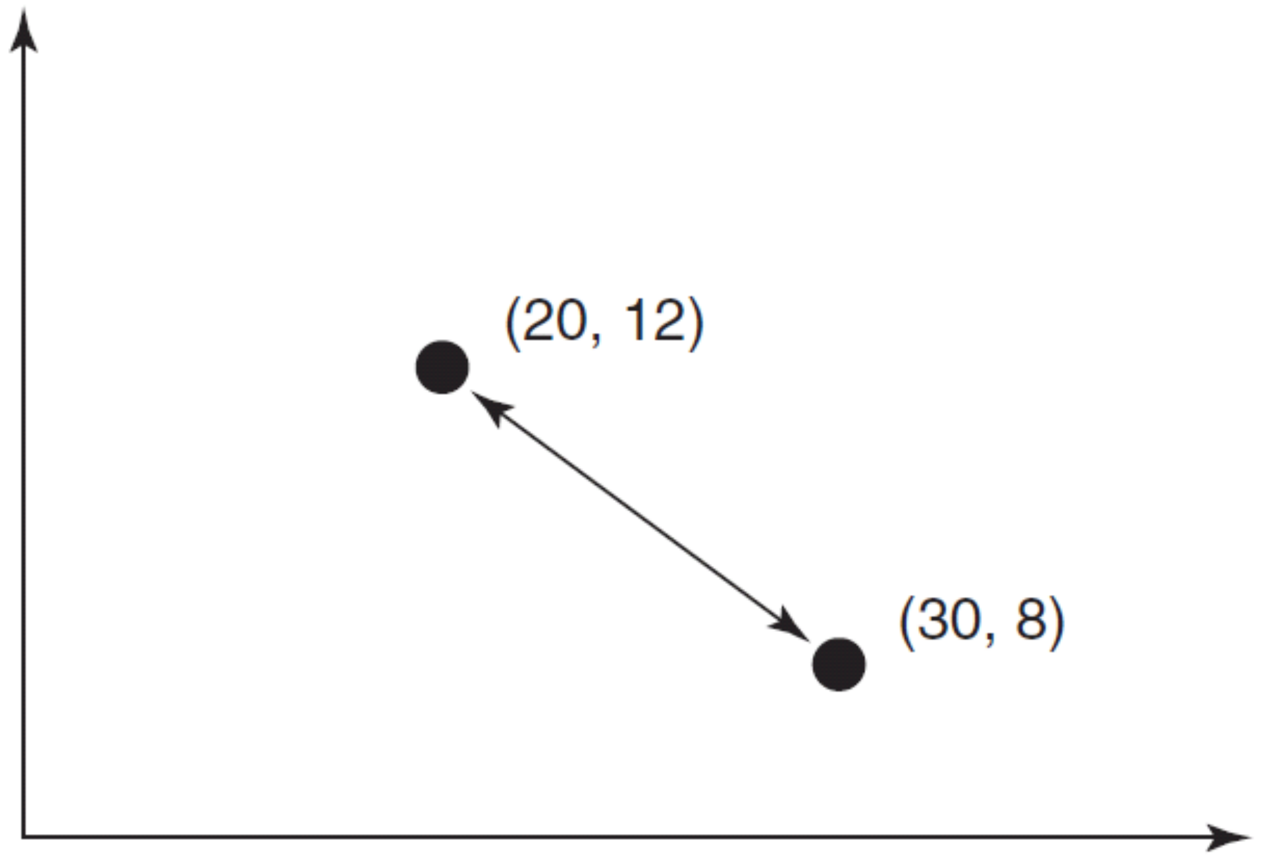
Since **RED** got the most votes, the final assignment is **RED**



How do we define similar?

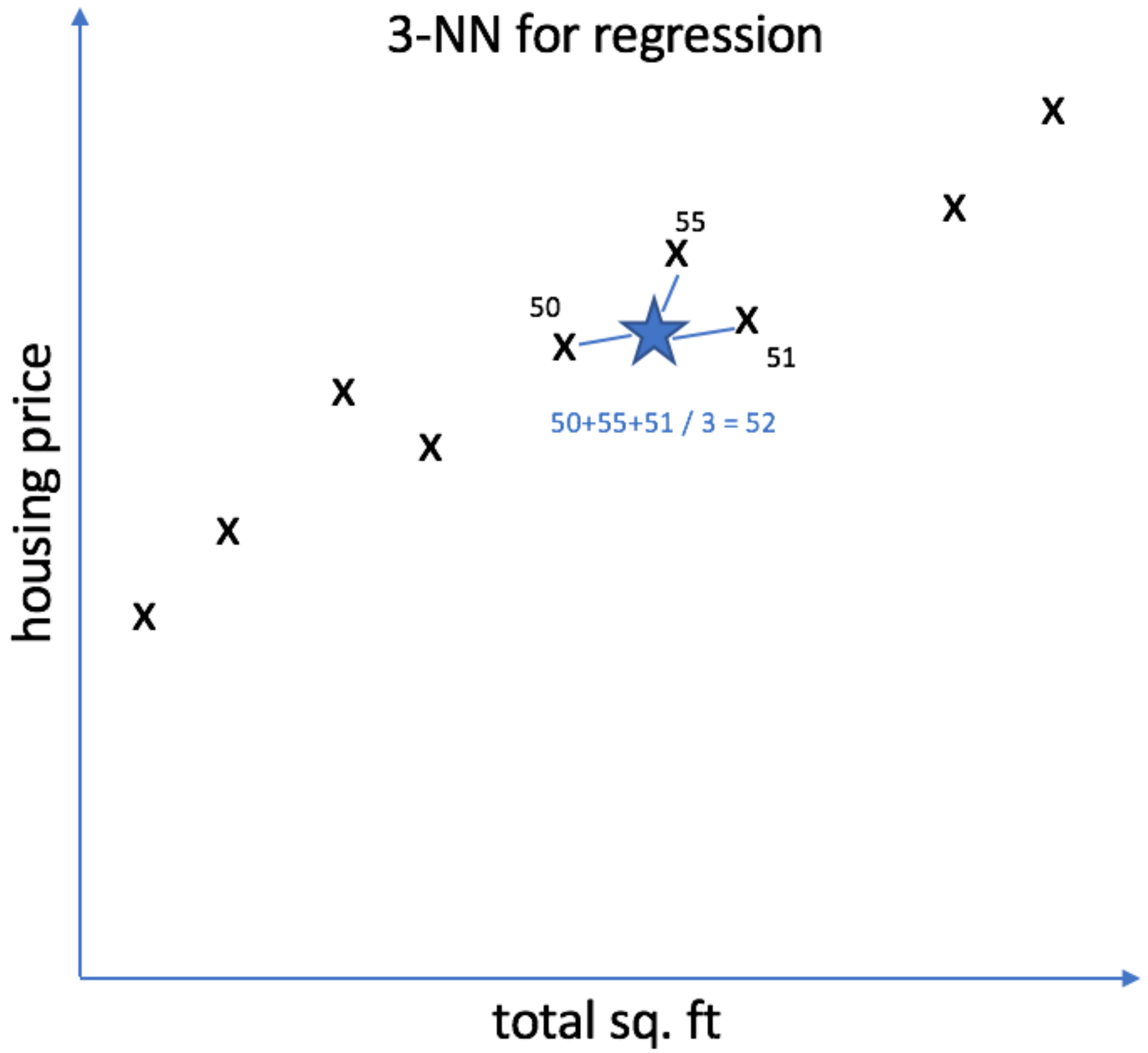
- Data analysts define distance metrics to measure similarity.
- The most common distance function is Euclidean distance, which represents the usual manner in which humans think of distance in the real world

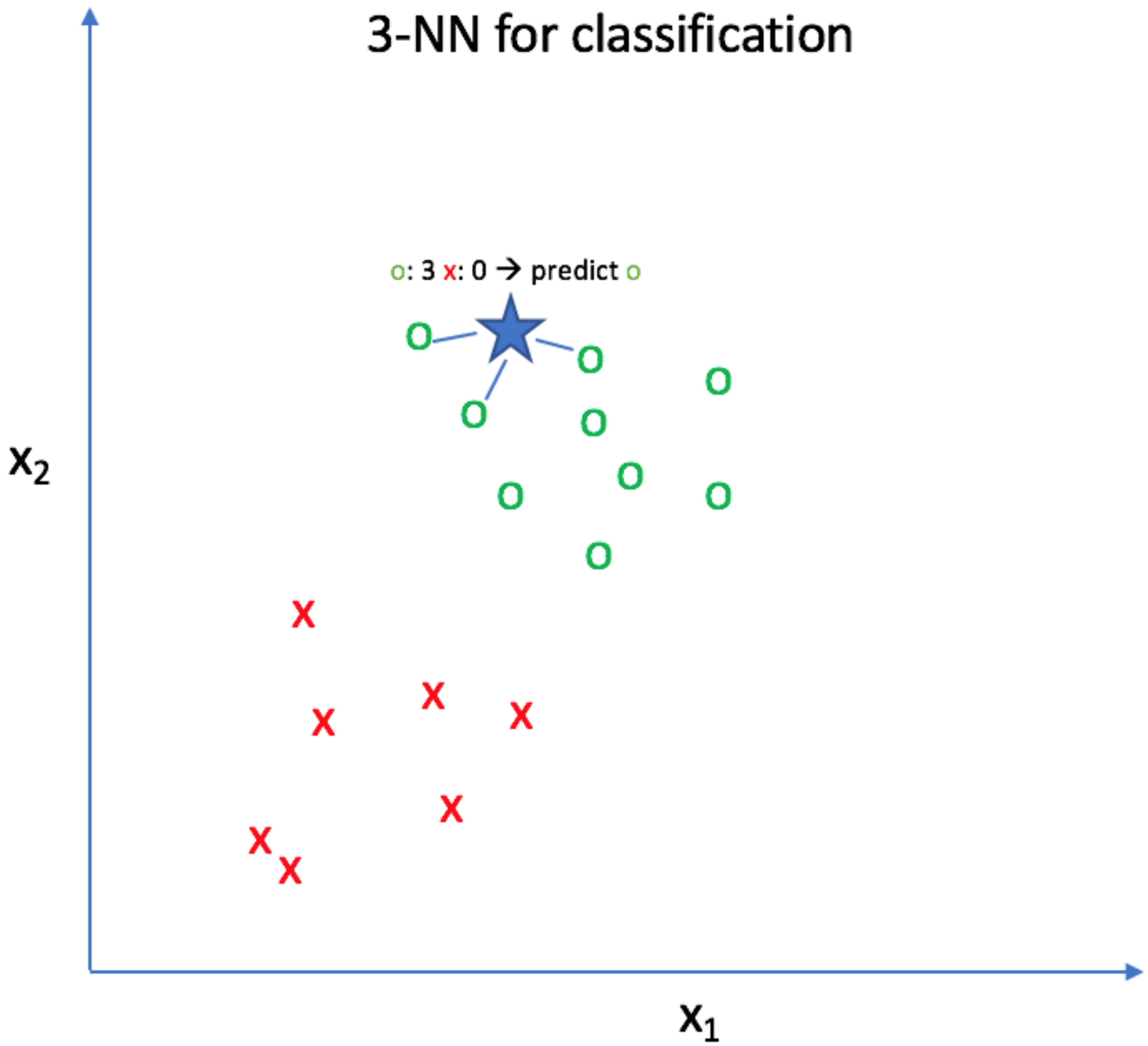
$$d_{\text{Euclidean}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$



- Other distance measures can include
  - the Manhattan distance,
  - the Hamming distance, and
  - the Mahalanobis distance.

K Nearest Neighbour





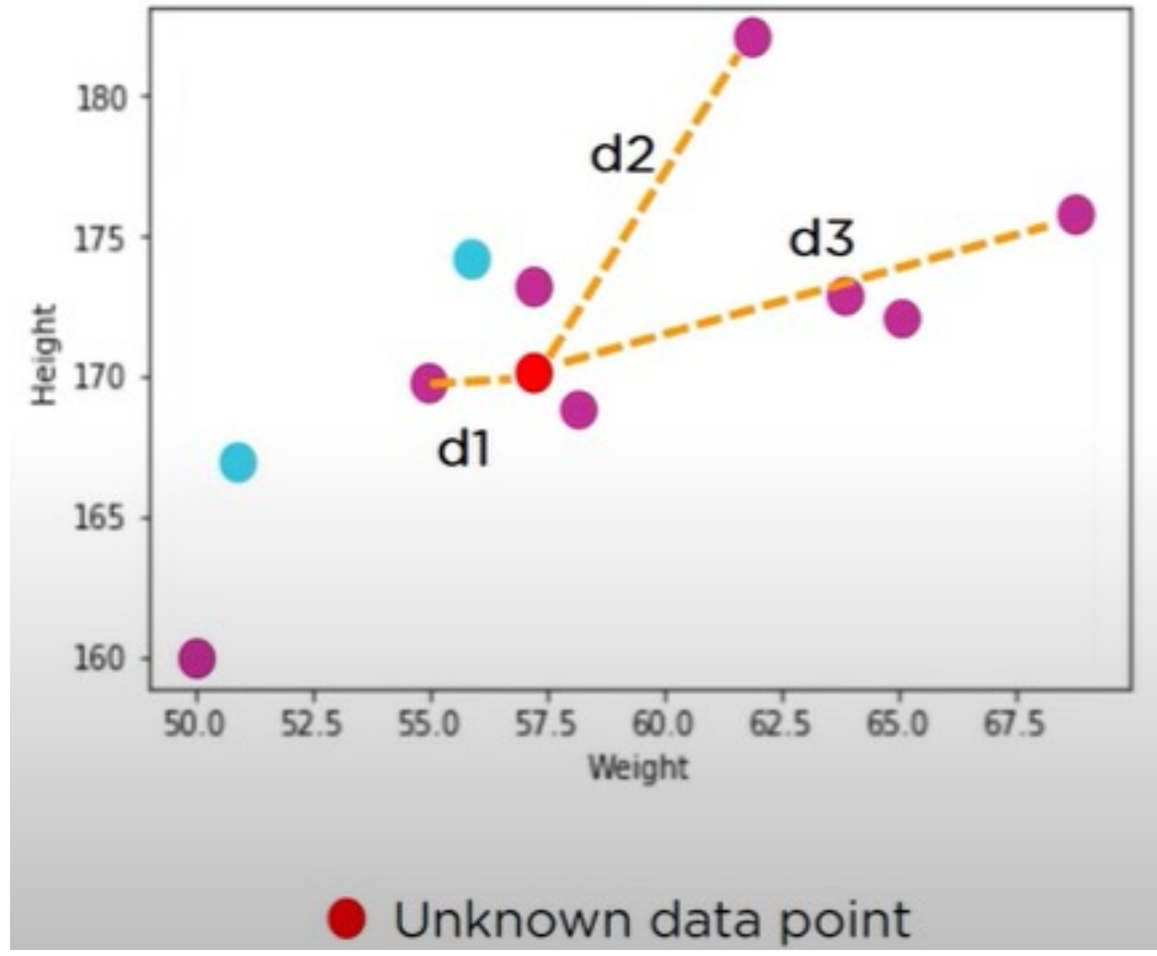
How does the KNN algorithm work?

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

New data set appear Weight 57 Kg and height 170 cm.

Class could be ??

Let's calculate it to understand clearly:



$dist(d1) = \sqrt{(170-167)^2 + (57-51)^2} \approx 6.7$

$dist(d2) = \sqrt{(170-182)^2 + (57-62)^2} \approx 13$

$dist(d3) = \sqrt{(170-176)^2 + (57-69)^2} \approx 13.4$

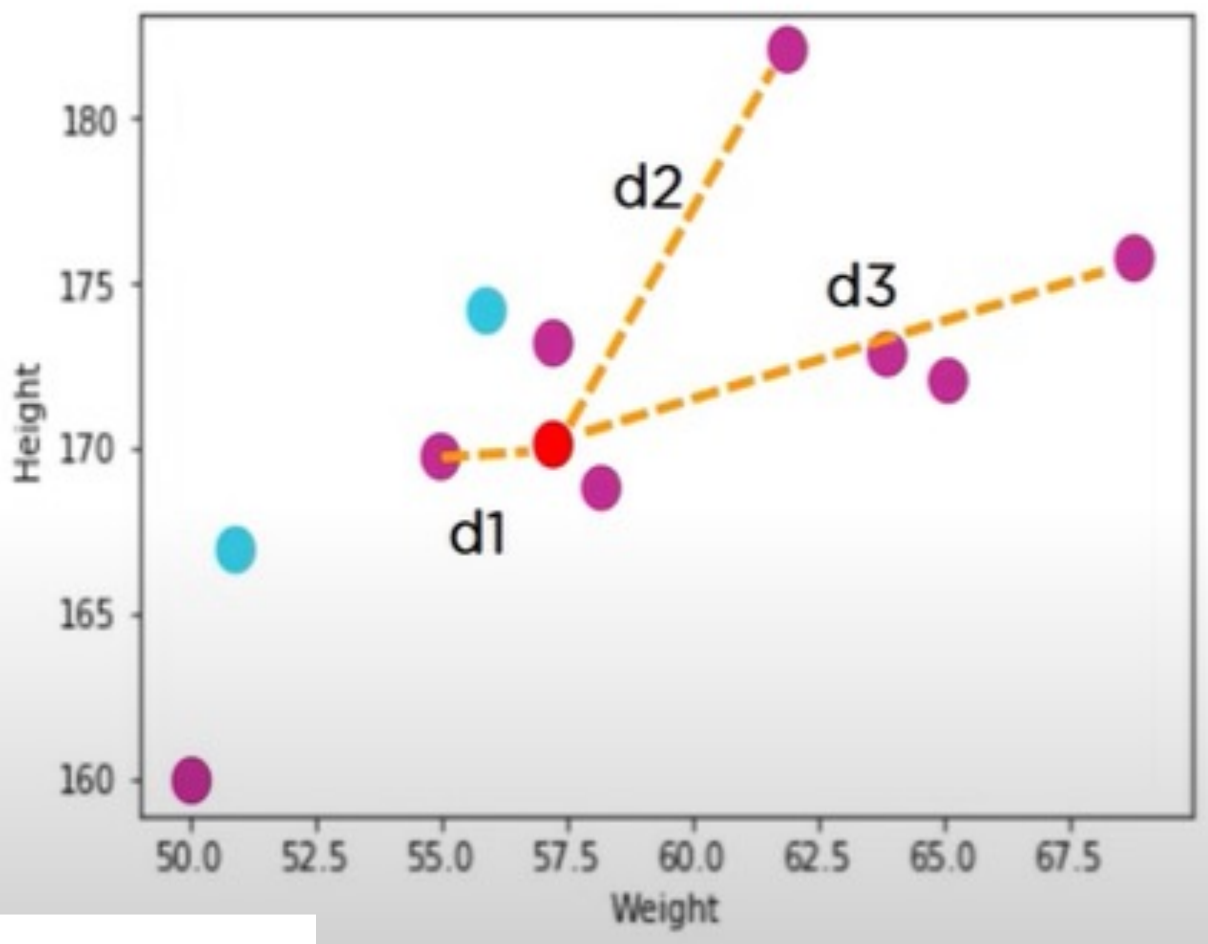
Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

How does the KNN algorithm work?

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

Let's calculate it to understand clearly:



dist(d1)= √(170-167)² + (57-51)² ~= 6.7

dist(d2)= √(170-182)² + (57-62)² ~= 13

dist(d3)= √(170-176)² + (57-69)² ~= 13.4

Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

● Unknown data point

Now, lets calculate the nearest neighbor at k=3

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

k = 3

Check the accuracy score.
If it is above 75% a fairly good model

Challenges of K Nearest Neighbour

- The data set would need to be balanced, with a sufficiently large percentage of the less common classifications. It is especially important that rare classifications be represented sufficiently, so that the algorithm does not only predict common classifications.

- Another challenges with k-NN and other distance-based algorithms is the number of inputs used in building a model.

2. Artificial Neural Networks as a Modelling Technique

- Neural networks implement complex nonlinear numeric functions, based on the fundamental concepts of fitting a model to data.

- A neural network can be considered as a “stack” of models.

- On the bottom of the stack are the original features and from these features are learned a variety of relatively simple models – e.g. Logistics regressions.

- Each subsequent layer in the stack applies a simple model (let’s say, another logistic regression) to the outputs of the previous layer.

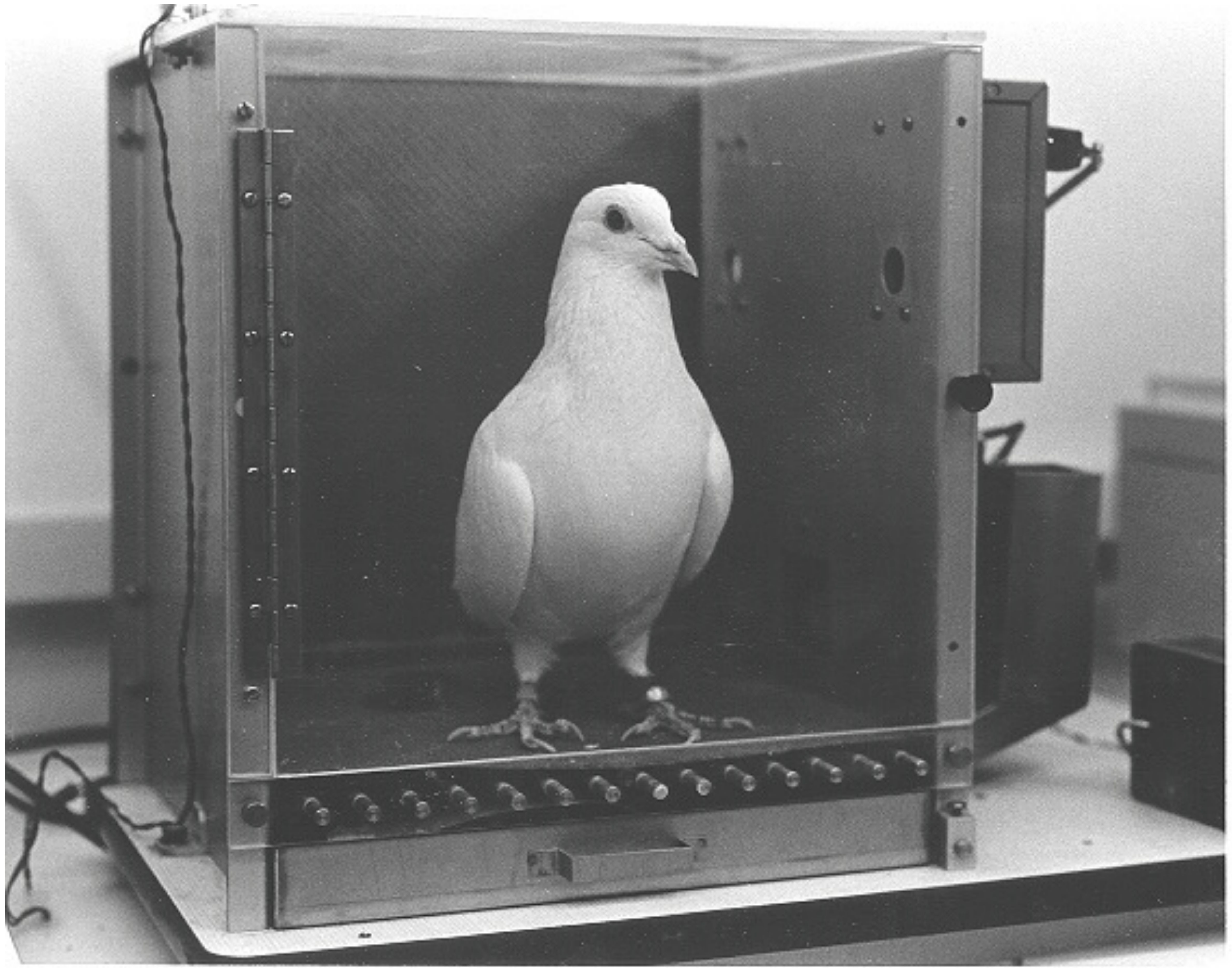
- Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal one another.

- Artificial neural networks (ANNs) are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

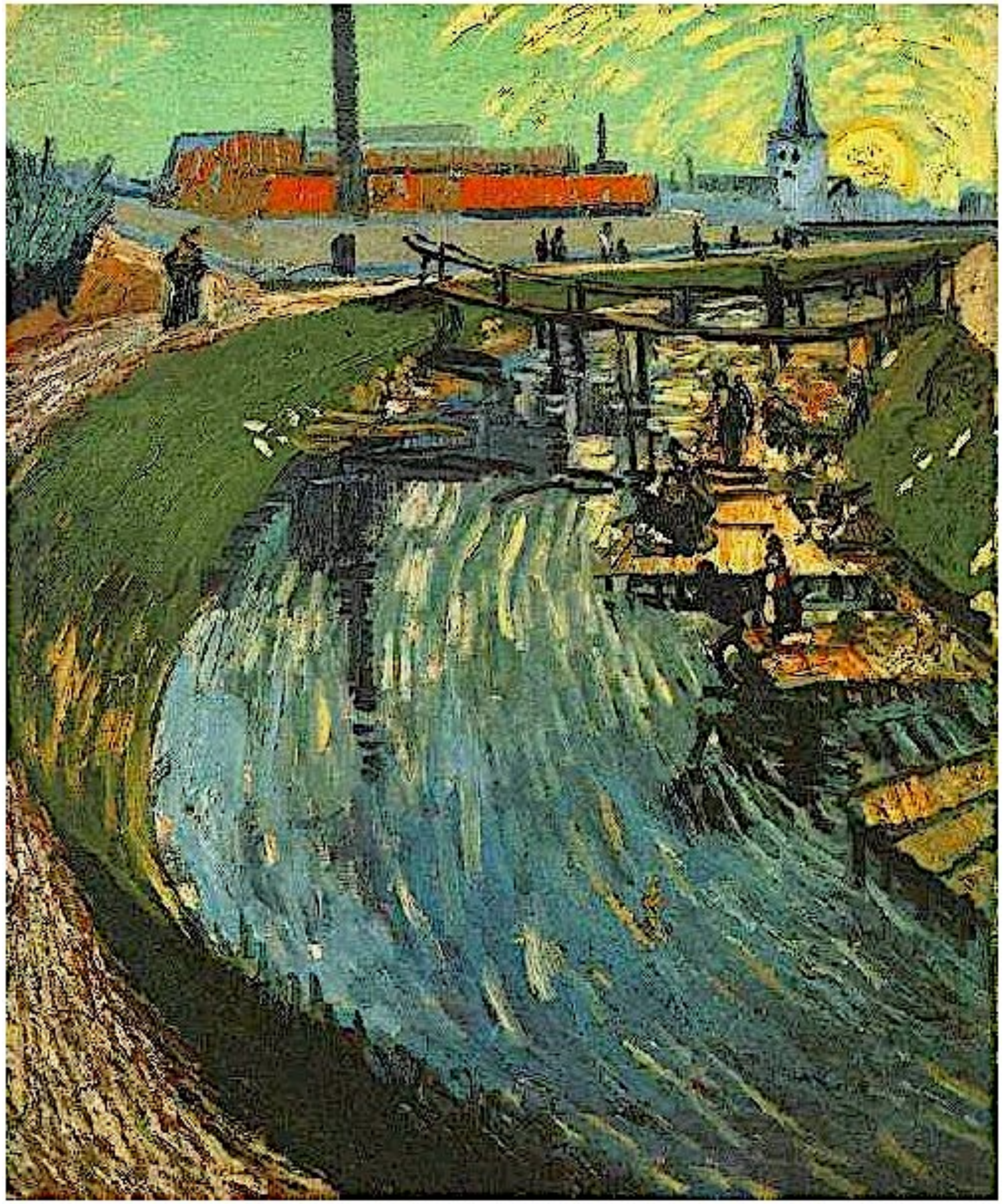
$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias$$

Biological Neural Nets at Work

- Pigeons as art experts (Watanabe *et al.* 1995)
    - Experiment:
        - Pigeon in Skinner box
        - Present paintings of two different artists (e.g. Chagall / Van Gogh)
        - Reward for pecking when presented a particular artist (e.g. Van Gogh)



Pigeons as Art Experts



Chagall / Van Gogh

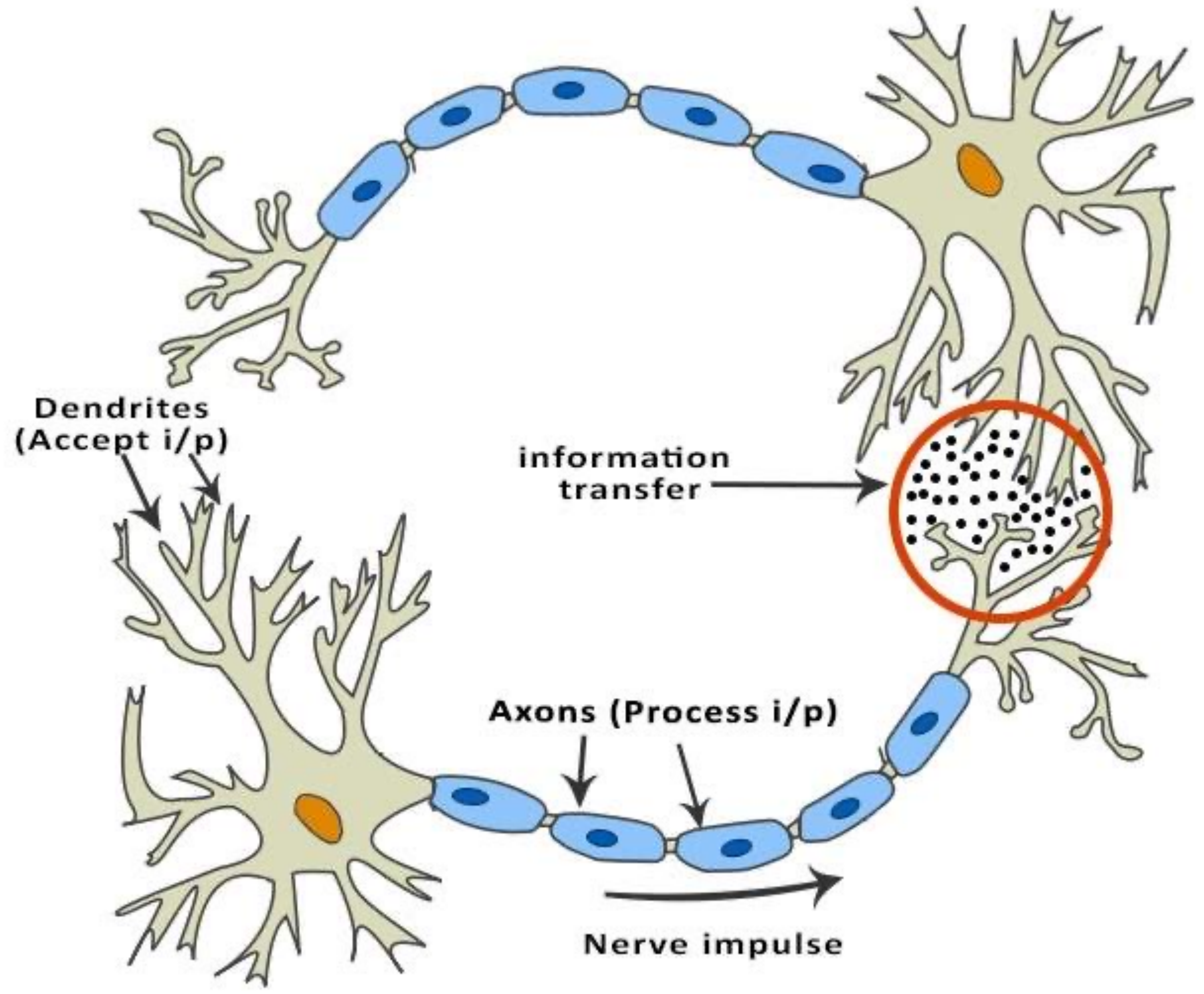
Pigeons as Art Experts

- Pigeons were able to discriminate between Van Gogh and Chagall with 95% accuracy (when presented with pictures they had been trained on)

- Discrimination still 85% successful for previously unseen paintings of the artists

- Pigeons do not simply memorise the pictures
- They can extract and recognise patterns (the ‘style’)
- They generalise from the already seen to make predictions
- This is what neural networks (biological and artificial) are good at (unlike conventional computers)

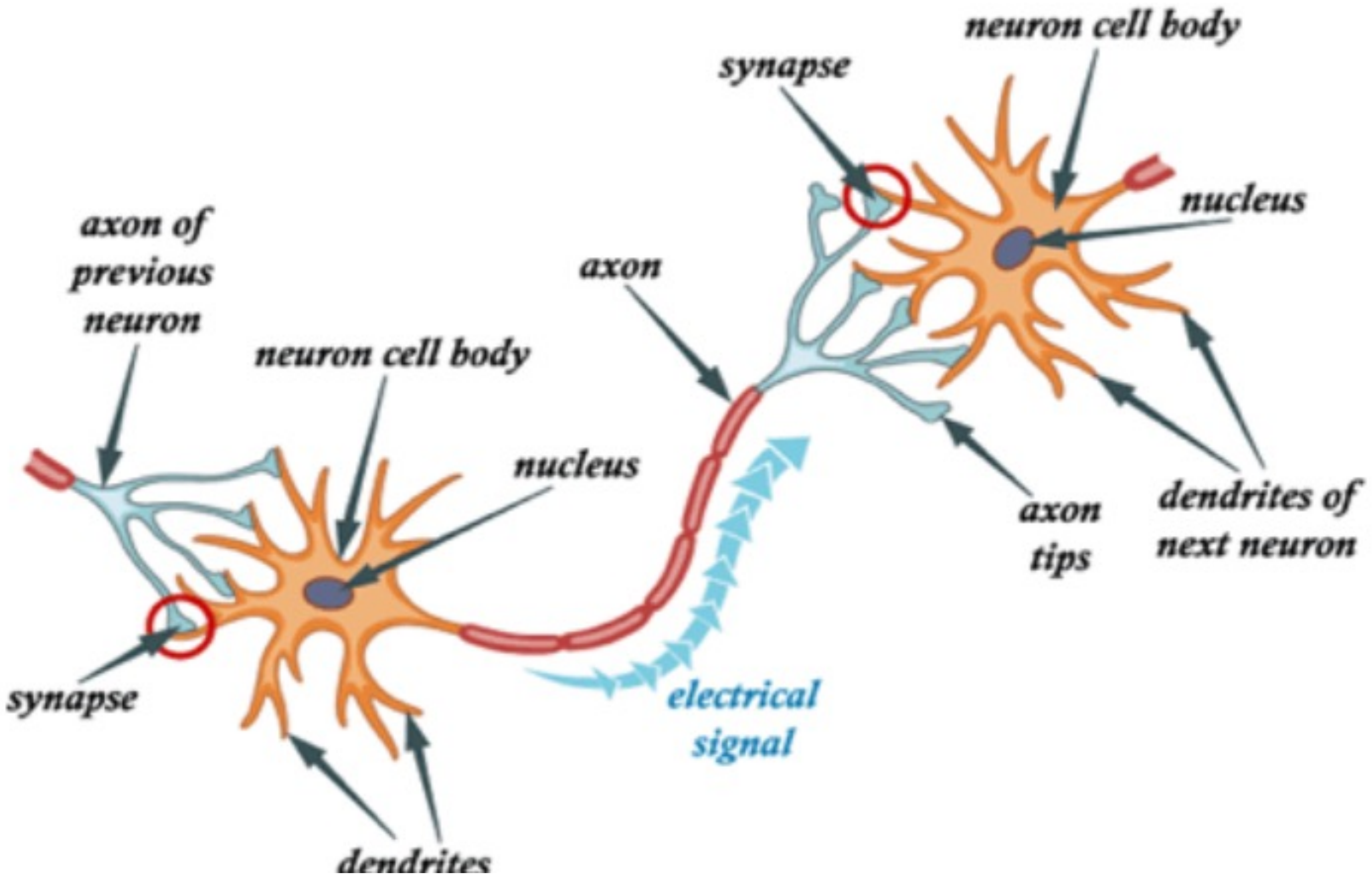
Why Artificial Neural Networks?



- Are we more intelligent than computers?
- What is intelligence anyway?
- Computers are high-speed serial machines
  - suited to tasks such as arithmetic operations; database creation, manipulation and maintenance; word processing etc.
  - hopeless at simple tasks like reasoning, generalizing ("thinking"), etc. (things that any 2 year old child can do easily)

Biological Inspiration

Artificial Neural Network (ANN) uses the processing of the brain as a basis to develop algorithms that can be used to model complex patterns and prediction problems.



Step 1: External signal received by dendrites

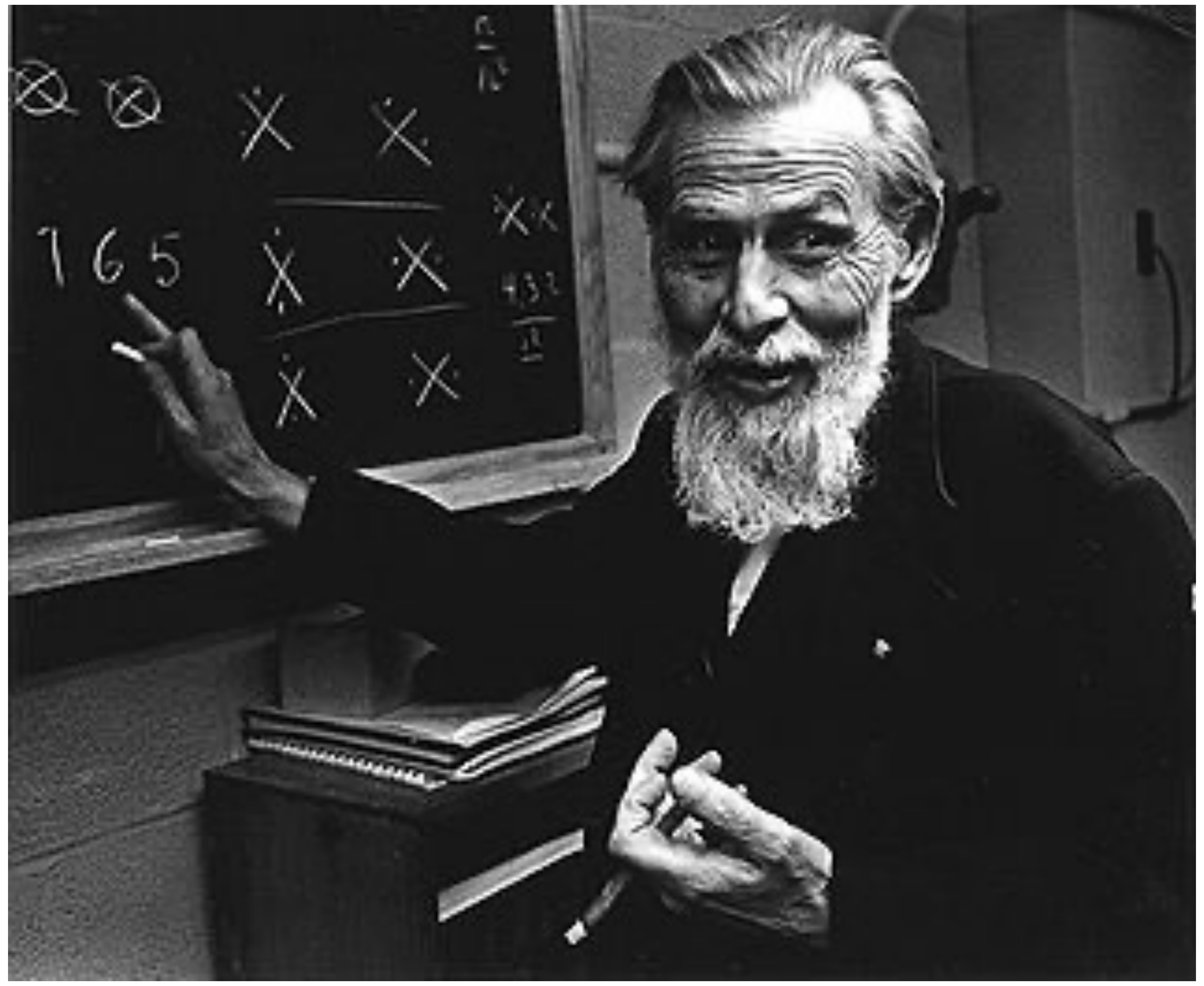
Step 2: External signal processed in the neuron cell body

Step 3: Processed signal converted to an output signal and transmitted through the Axon

Step 4: Output signal received by the dendrites of the next neuron through the synapse

A Bit of History

- In 1943, Warren McCulloch (neurophysiologist) and Walter Pitts (logician) –Yale - developed a simple model to explain how biological neurons worked
- Took place in the 1930s and 40s – before the digital computer
- Original work was carried out to understand, and later simulate, the biological brain



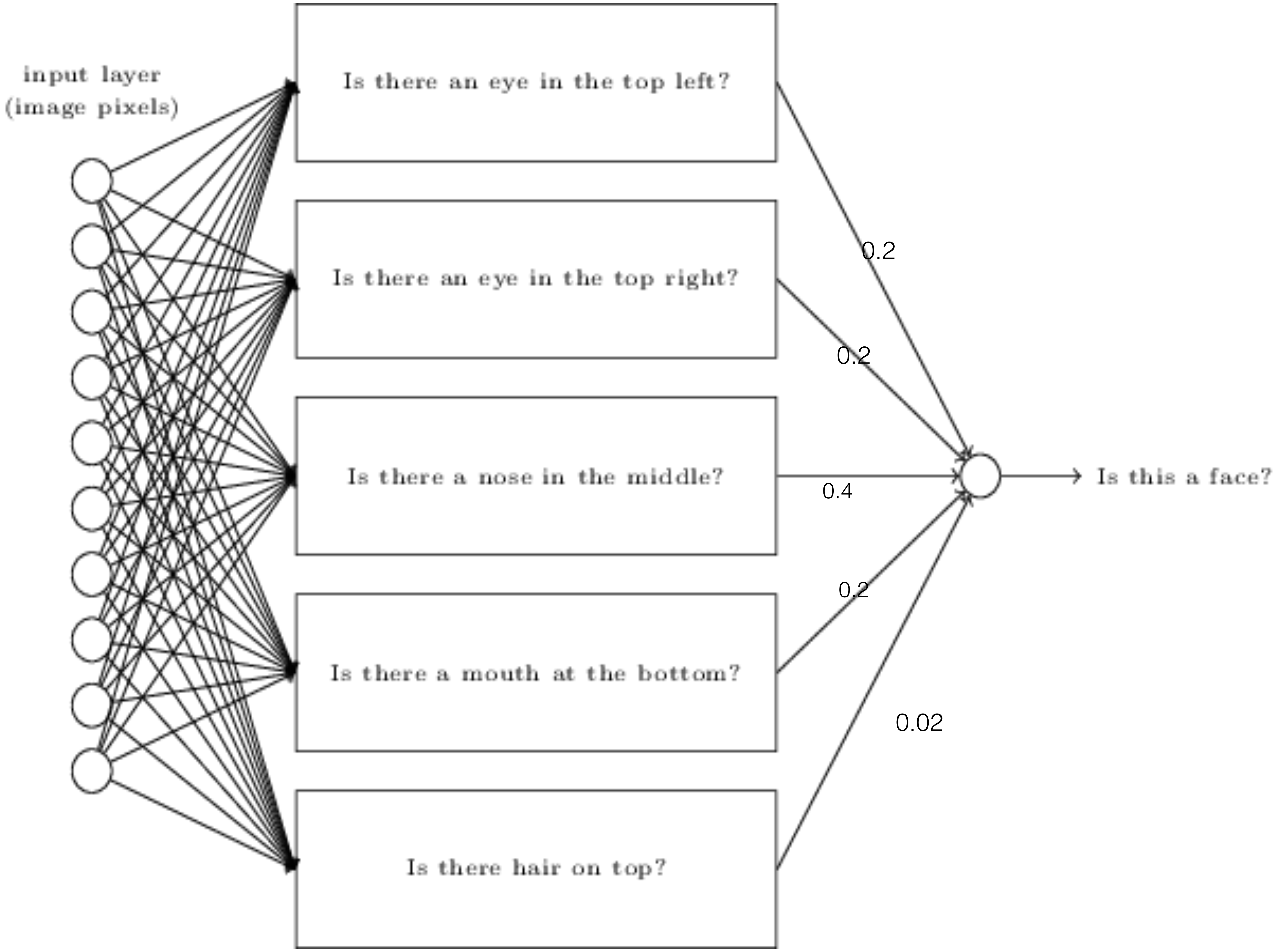
Warren McCulloch

Walter Pitts

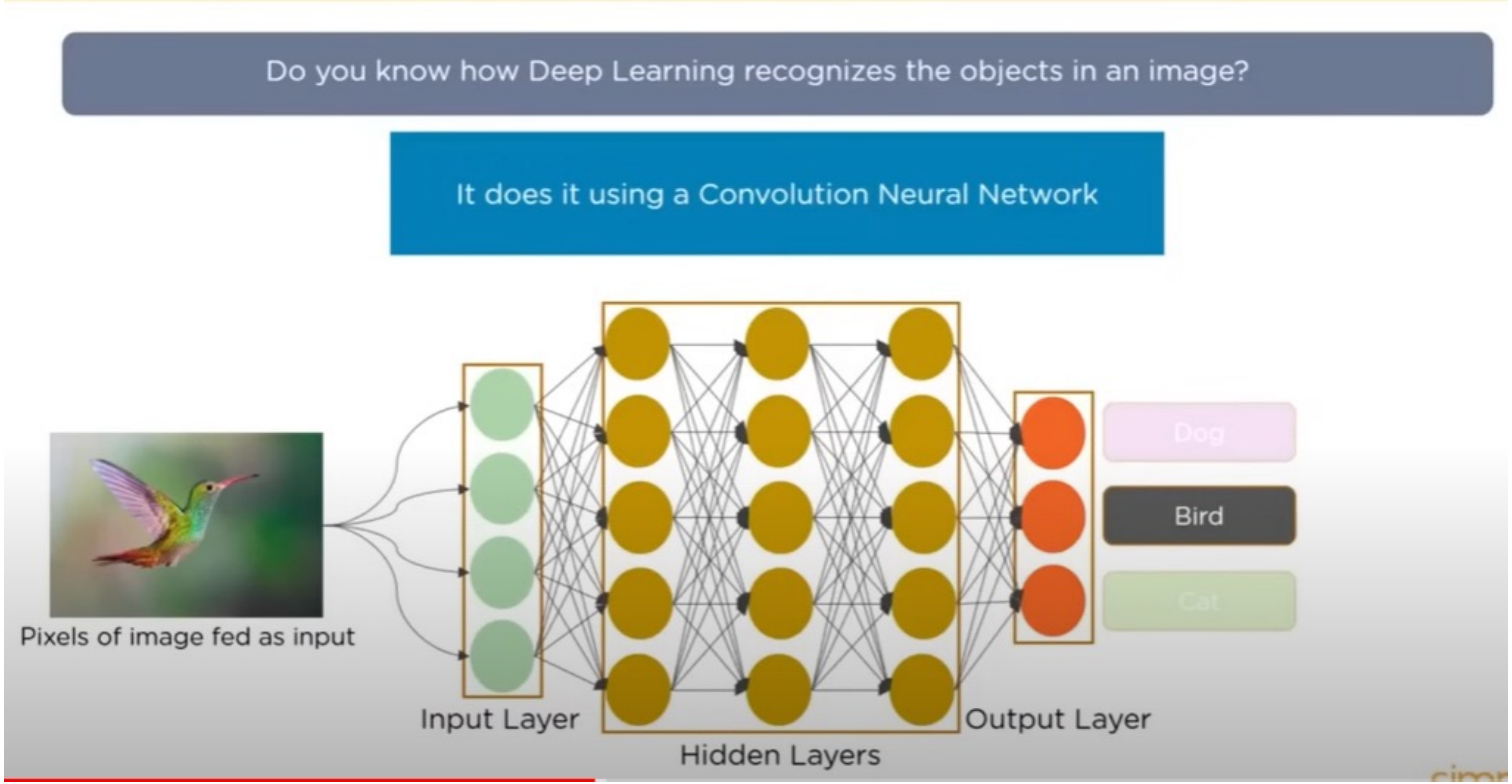
Artificial Neural Networks as a Modelling Technique

- So in a two-layer stack, we would learn a set of logistic regressions from the original features, and then learn a logistic regression using as features the outputs of the first set of logistic regressions.

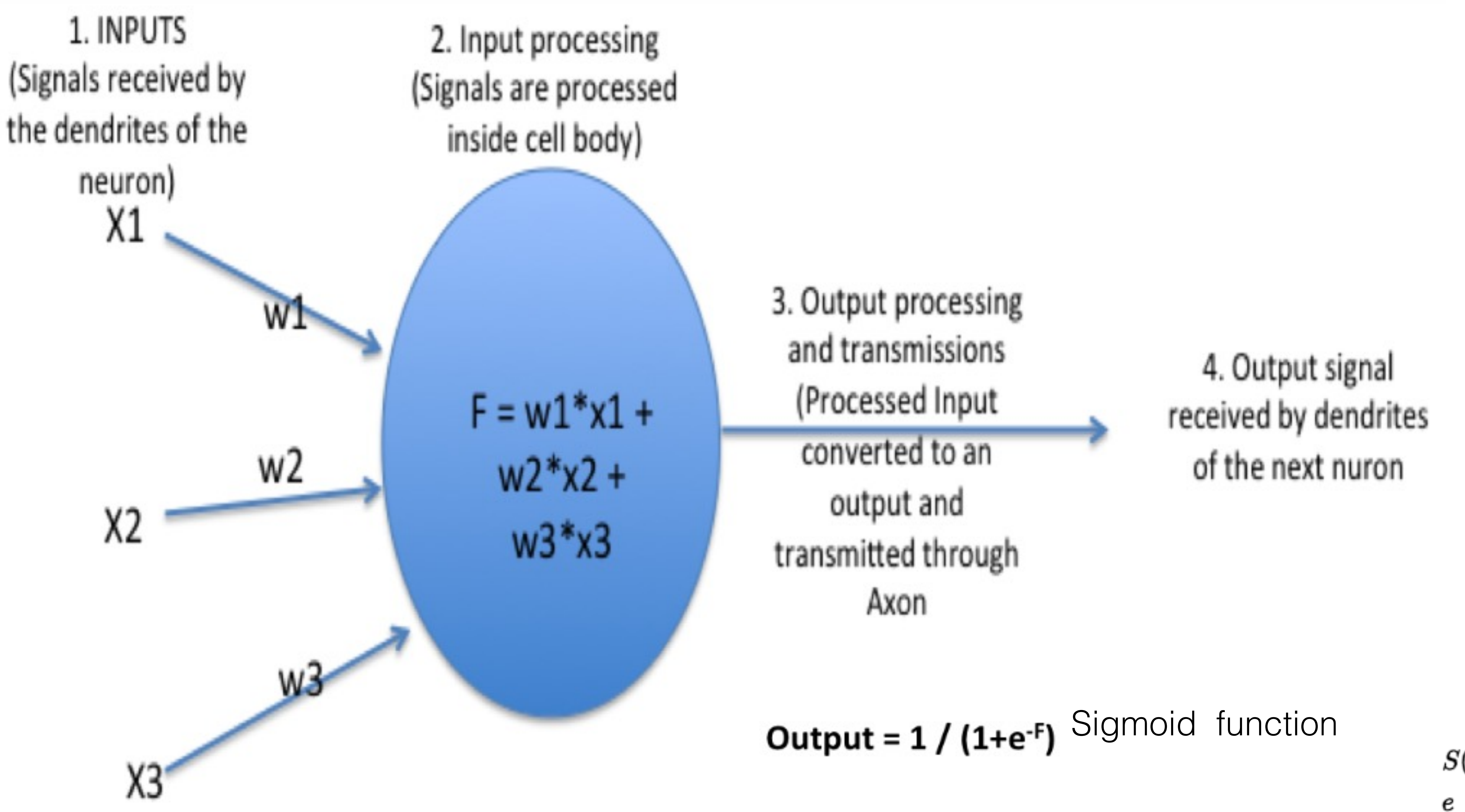
- We could think of this very roughly as first creating a set of “experts” in different facets of the problem (the first-layer models), and then learning how to weight the opinions of these different experts (the second-layer model).



How ANN works



Neural Network Architecture



Output = 1 / (1+e^{-F}) Sigmoid function

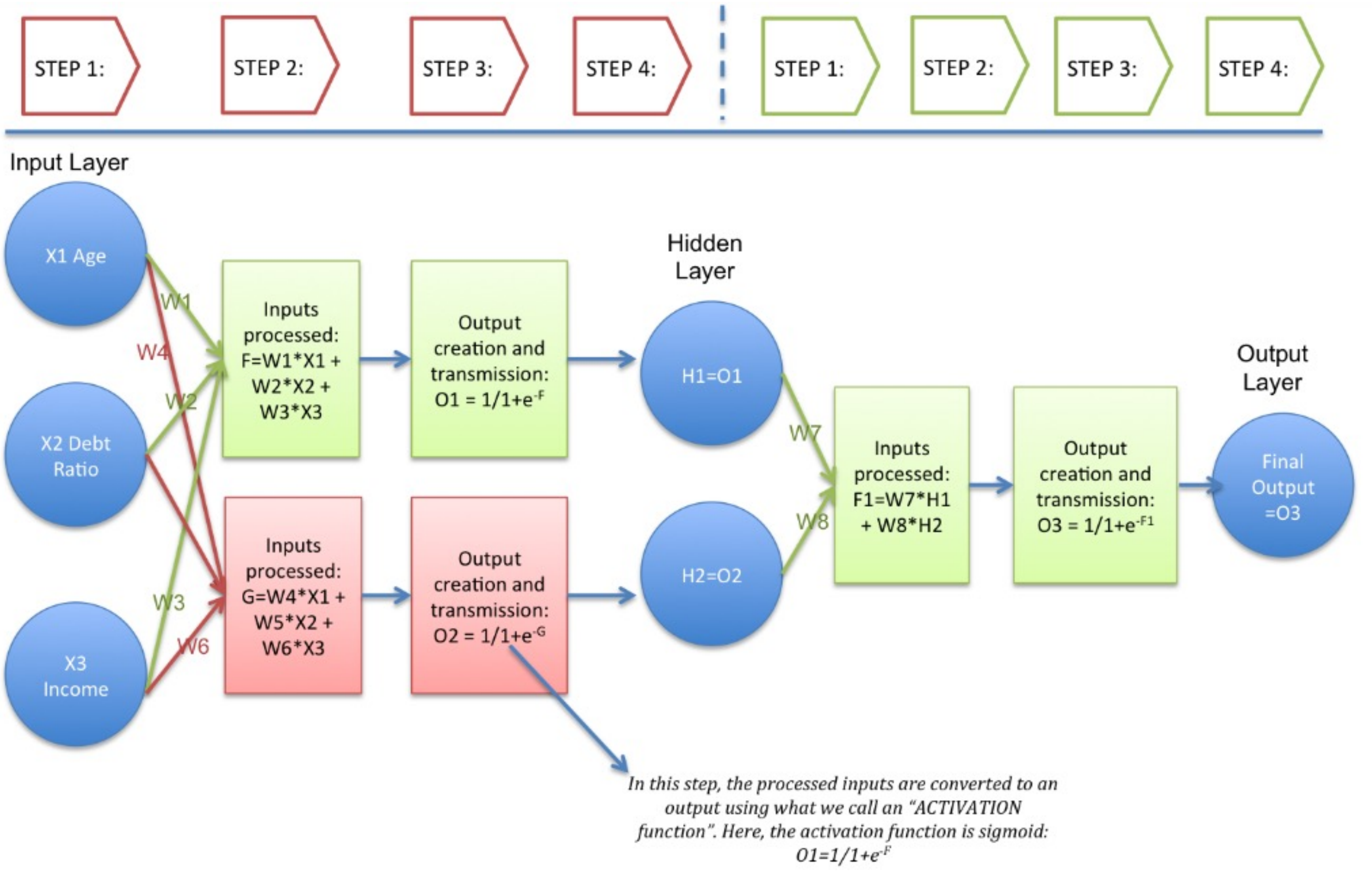
$$S(x) = \frac{1}{1+e^{-x}}$$

$S(x)$ = sigmoid function

e = Euler's number

w1, w2, w3 gives the strength of the input signals

Example 1: Loan Application Assessment

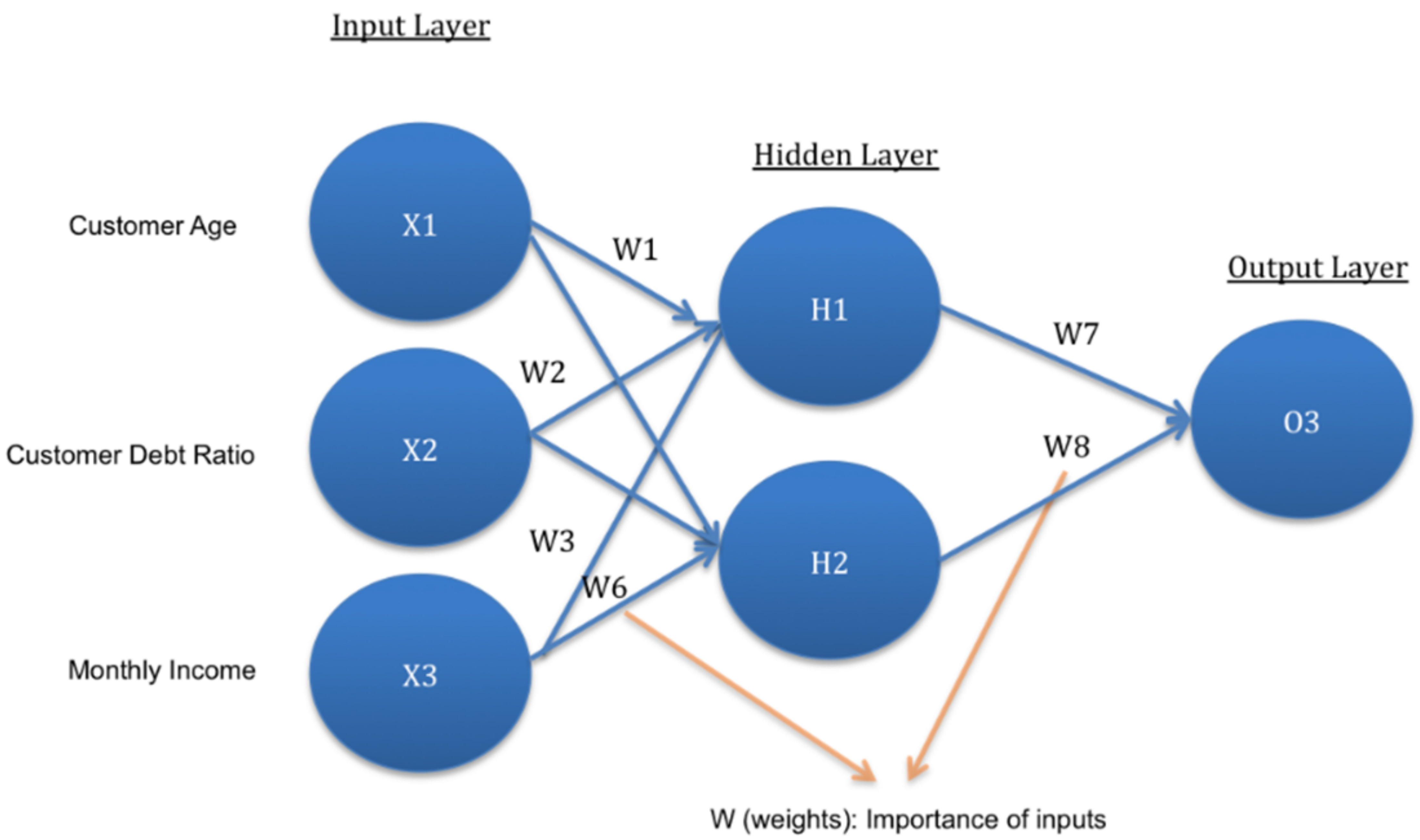


This network architecture is called "feed-forward network", as you can see that input signals are flowing in only one direction (from inputs to outputs).

Example 1: Loan Application Assessment

A bank wants to assess whether to approve a loan application to a customer, so, it wants to predict whether a customer is likely to default on the loan.

Customer ID	Customer Age	Debt Ratio (% of Income)	Monthly Income ($)	Loan Defaulter Yes:1 No:0 (Column W)	Default Prediction (Column X)
1	45	0.80	9120	1	0.76
2	40	0.12	2000	1	0.66
3	38	0.08	3042	0	0.34
4	25	0.03	3300	0	0.55
5	49	0.02	63588	0	0.15
6	74	0.37	3500	0	0.72



So, we have to predict Column X. A prediction closer to 1 indicates that the customer has more chances to default.

La Trobe Business School

Example 1: Loan Application Assessment

Customer ID	Customer Age	Debt Ratio (% of Income)	Monthly Income ($)	Loan Defaulter Yes:1 No:0 (Column W)	Default Prediction (Column X)	Prediction Error
1	45	0.80	9120	1	0.76	0.24
2	40	0.12	2000	1	0.66	0.34
3	38	0.08	3042	0	0.34	-0.34
4	25	0.03	3300	0	0.55	-0.55
5	49	0.02	63588	0	0.15	-0.15
6	74	0.37	3500	0	0.72	-0.72

- A good model with high accuracy gives predictions that are very close to the actual values.
- In the table Column X values should be very close to Column W values.

- The key to get a good model with accurate predictions is to find “optimal values of W — weights” that minimizes the prediction error.
- This is achieved by “Back propagation algorithm” and this makes ANN a learning algorithm because by learning from the errors, the model is improved.

LA TROBE UNIVERSITY

Advantages of ANNs

1. ANNs have the ability to learn and model non-linear and complex relationships, which is really important because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex.

2. ANNs can generalize — After learning from the initial inputs and their relationships, it can infer unseen relationships on unseen data as well, thus making the model generalize and predict on unseen data.

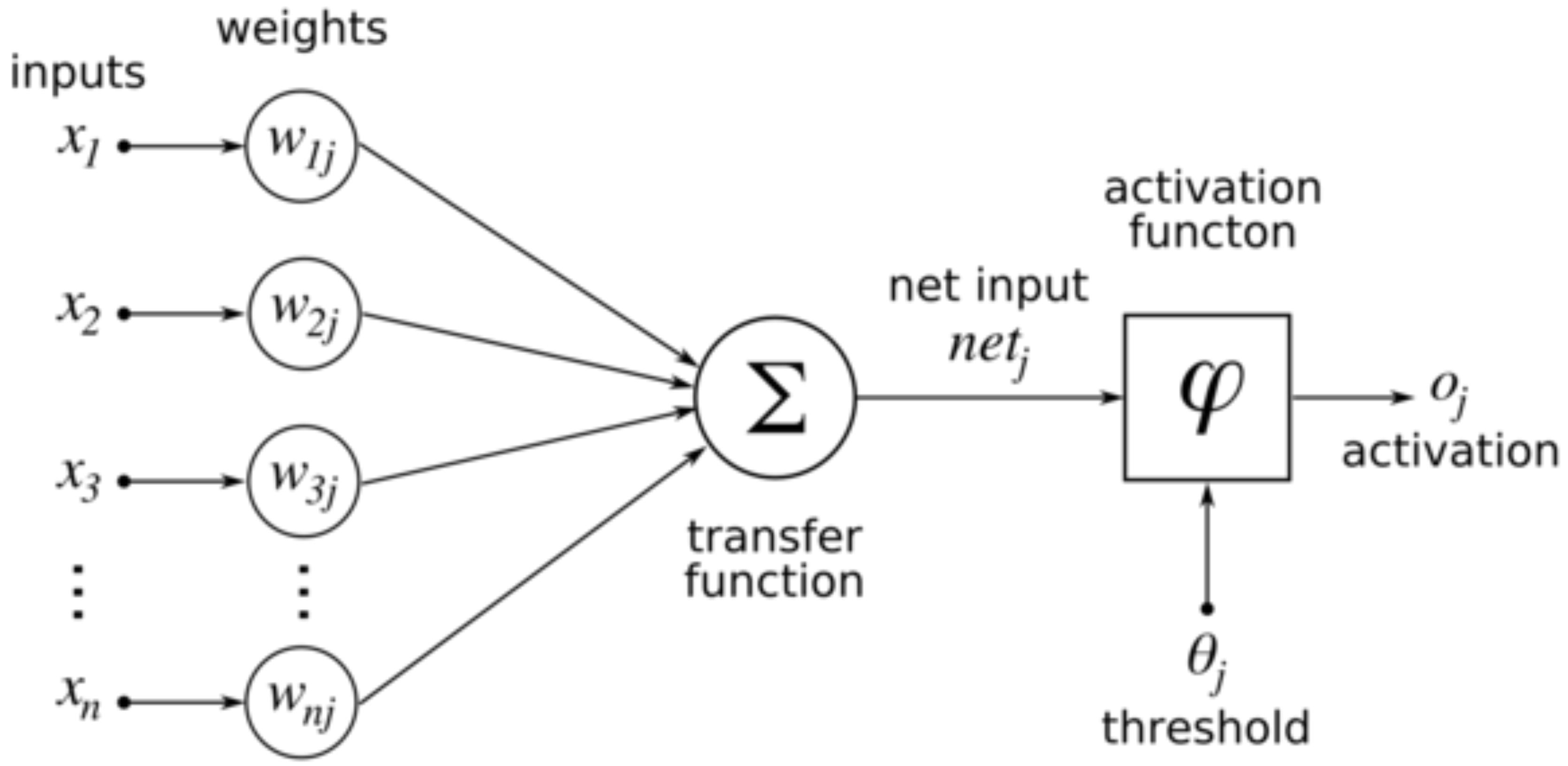
3. ANN does not impose any restrictions on the input variables (like how they should be distributed).

4. Additionally, many studies have shown that ANNs can better model heteroskedasticity i.e. data with high volatility and non-constant variance, given its ability to learn hidden relationships in the data without imposing any fixed relationships in the data (financial time series forecasting - e.g. stock prices where data volatility is very high).

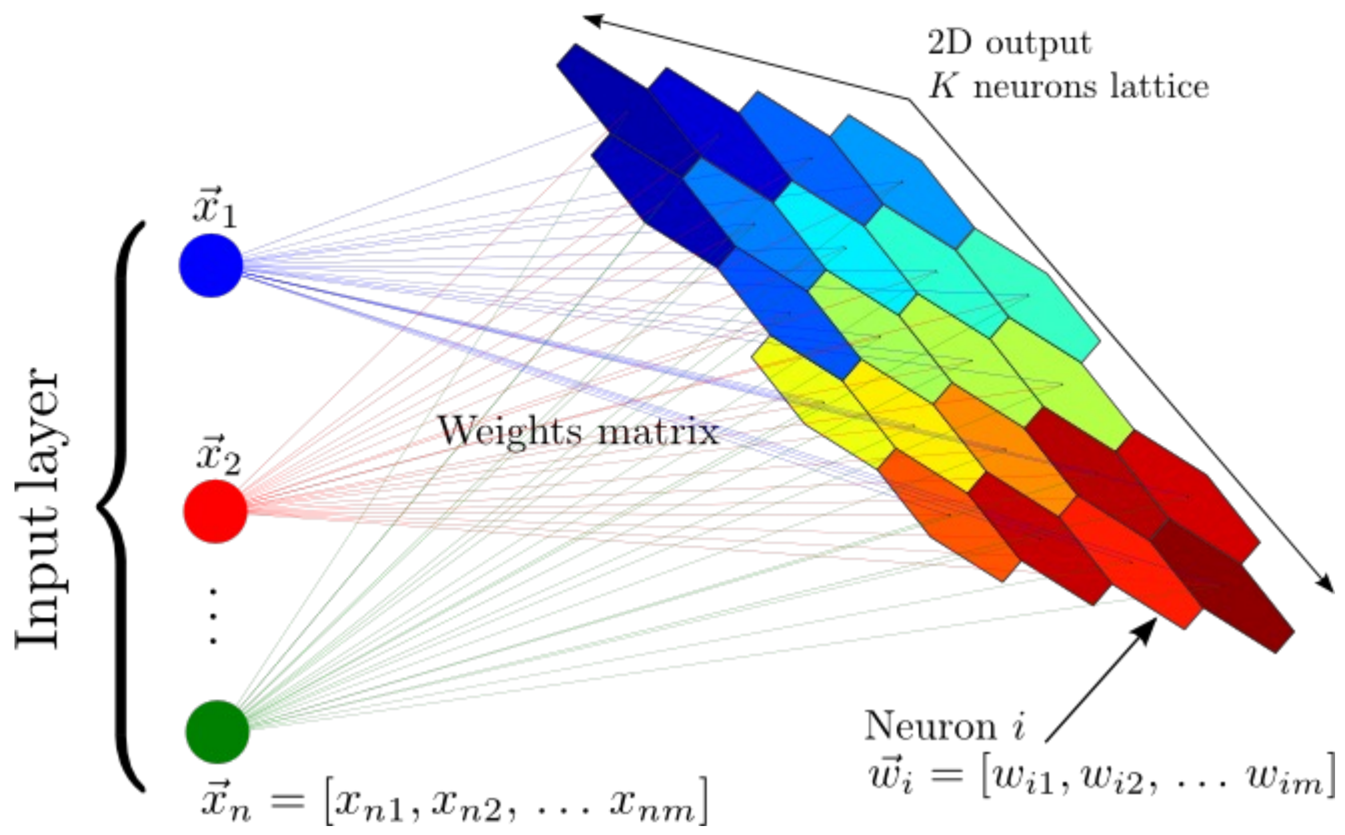
Types of Neural Nets

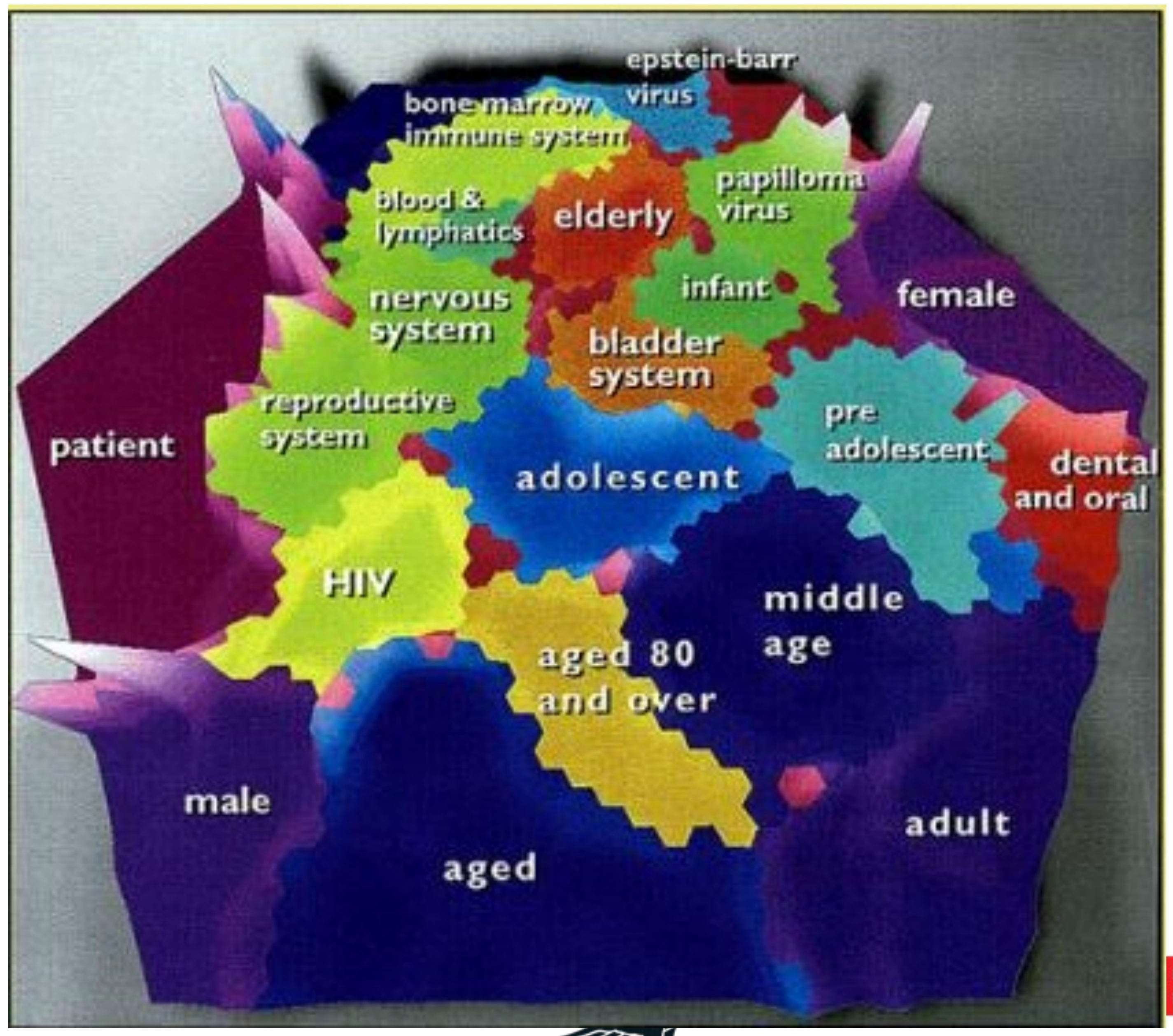
Two main types

- Supervised NNs
    - Feed forward with back propagation

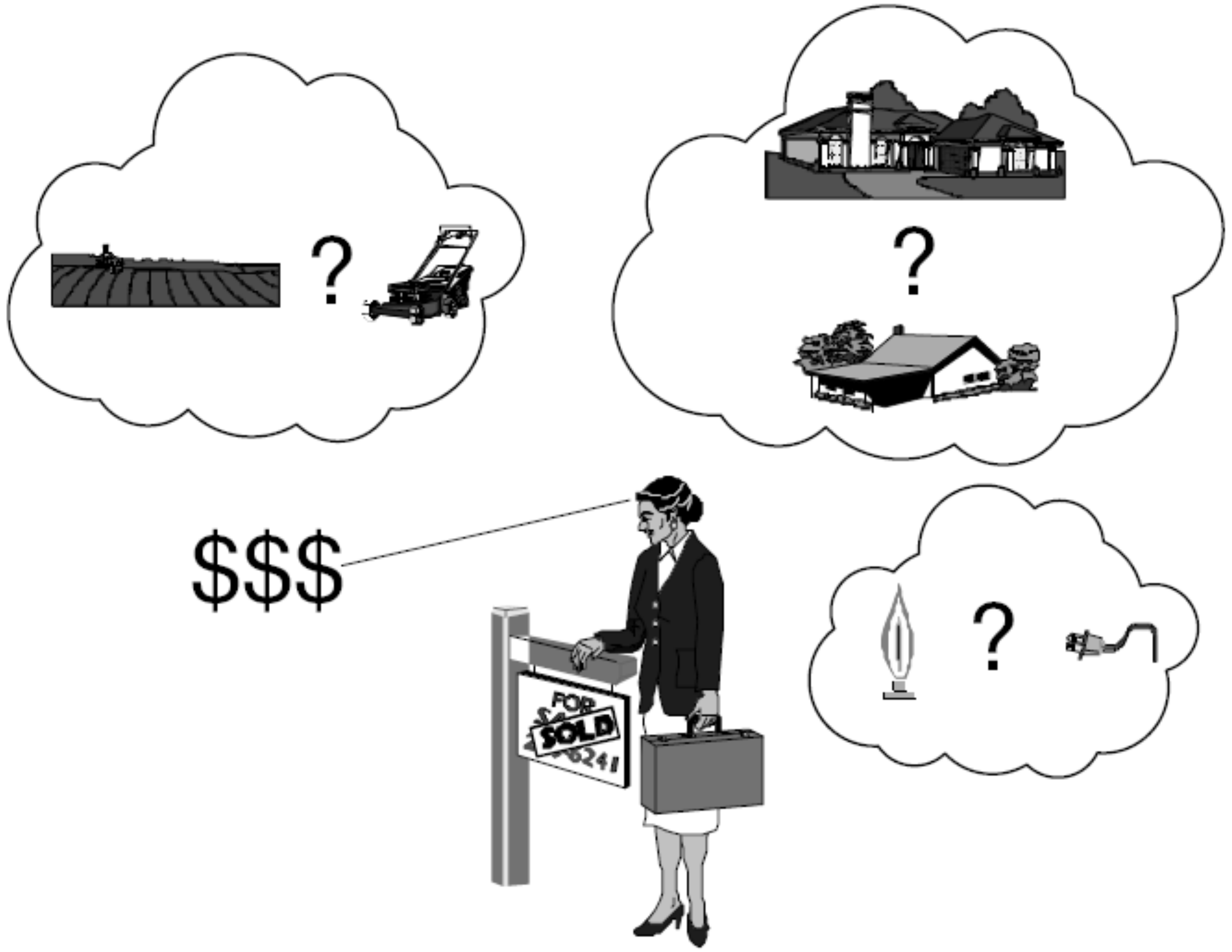


- ~~Un-supervised NNs~~
    - ~~Self Organizing Maps~~





Example 2: Real Estate Appraisal

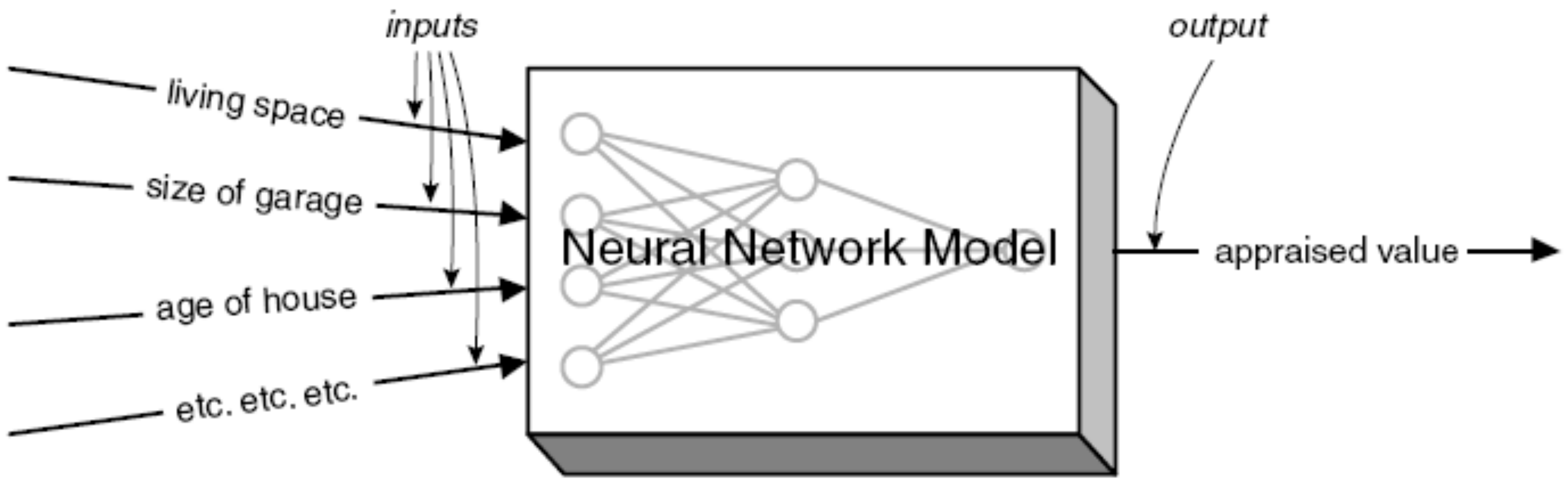


Location, bedrooms, larger garage, style, lot size etc will be considered when estimating the value

- Automated appraisals could help real estate agents better match prospective buyers to homes, improving the productivity of inexperienced agents
- Automated web based feedback to prospective buyers
- The neural network mimics an appraiser who estimates the market value of a house based on features of the property

Example 2 : Real Estate Appraisal

- A set formula is not applied – her experience and knowledge about sales prices of similar homes is used
- She is aware of recent prices in the region as well as trends over time – which are used to fine tune her calculations to fit the latest data
- This is an example of a human expert in a well defined domain – a good problem for NNs



Example 2: Real Estate Appraisal

- Some features for real estate appraisal are shown below
- Further information such as demographics of neighbourhood, and other qualities etc. will be useful

FEATURE	DESCRIPTION	RANGE OF VALUES
Num_Apartments	Number of dwelling units	Integer: 1–3
Year_Built	Year built	Integer: 1850–1986
Plumbing_Fixtures	Number of plumbing fixtures	Integer: 5–17
Heating_Type	Heating system type	coded as A or B
Basement_Garage	Basement garage (number of cars)	Integer: 0–2
Attached_Garage	Attached frame garage area (in square feet)	Integer: 0–228
Living_Area	Total living area (square feet)	Integer: 714–4185
Deck_Area	Deck / open porch area (square feet)	Integer: 0–738
Porch_Area	Enclosed porch area (square feet)	Integer: 0–452
Recroom_Area	Recreation room area (square feet)	Integer: 0–672
Basement_Area	Finished basement area (square feet)	Integer: 0–810

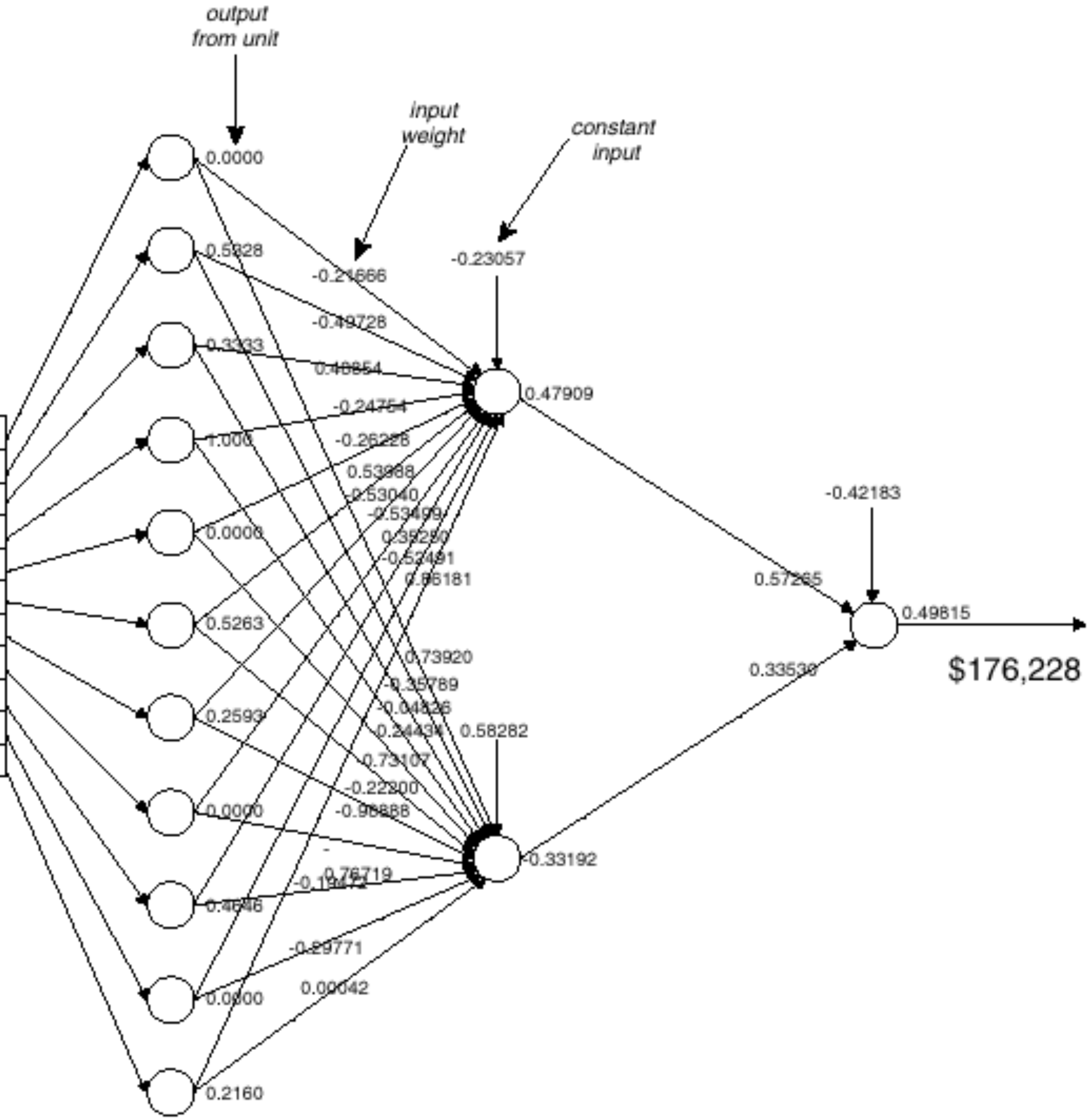
Example 2: Real Estate Appraisal

- A sample record from the training set can be:

FEATURE	RANGE OF VALUES	ORIGINAL VALUE	SCALED VALUE
Sales_Price	$103,000–$250,000	$171,000	–0.0748
Months_Ago	0–23	4	–0.6522
Num_Apartments	1-3	1	–1.0000
Year_Built	1850–1986	1923	+0.0730
Plumbing_Fixtures	5–17	9	–0.3077
Heating_Type	coded as A or B	B	+1.0000
Basement_Garage	0–2	0	–1.0000
Attached_Garage	0–228	120	+0.0524
Living_Area	714–4185	1,614	–0.4813
Deck_Area	0–738	0	–1.0000
Porch_Area	0–452	210	–0.0706
Recroom_Area	0–672	0	–1.0000
Basement_Area	0–810	175	–0.5672

Example 2: Real Estate Appraisal

Num_Apartments	1	0.0000
Year_Built	1923	0.5328
Plumbing_Fixtures	9	0.3333
Heating_Type	B	1.0000
Basement_Garage	0	0.0000
Attached_Garage	120	0.5263
Living_Area	1614	0.2593
Deck_Area	0	0.0000
Porch_Area	210	0.4646
Recroom_Area	0	0.0000
Basement_Area	175	0.2160

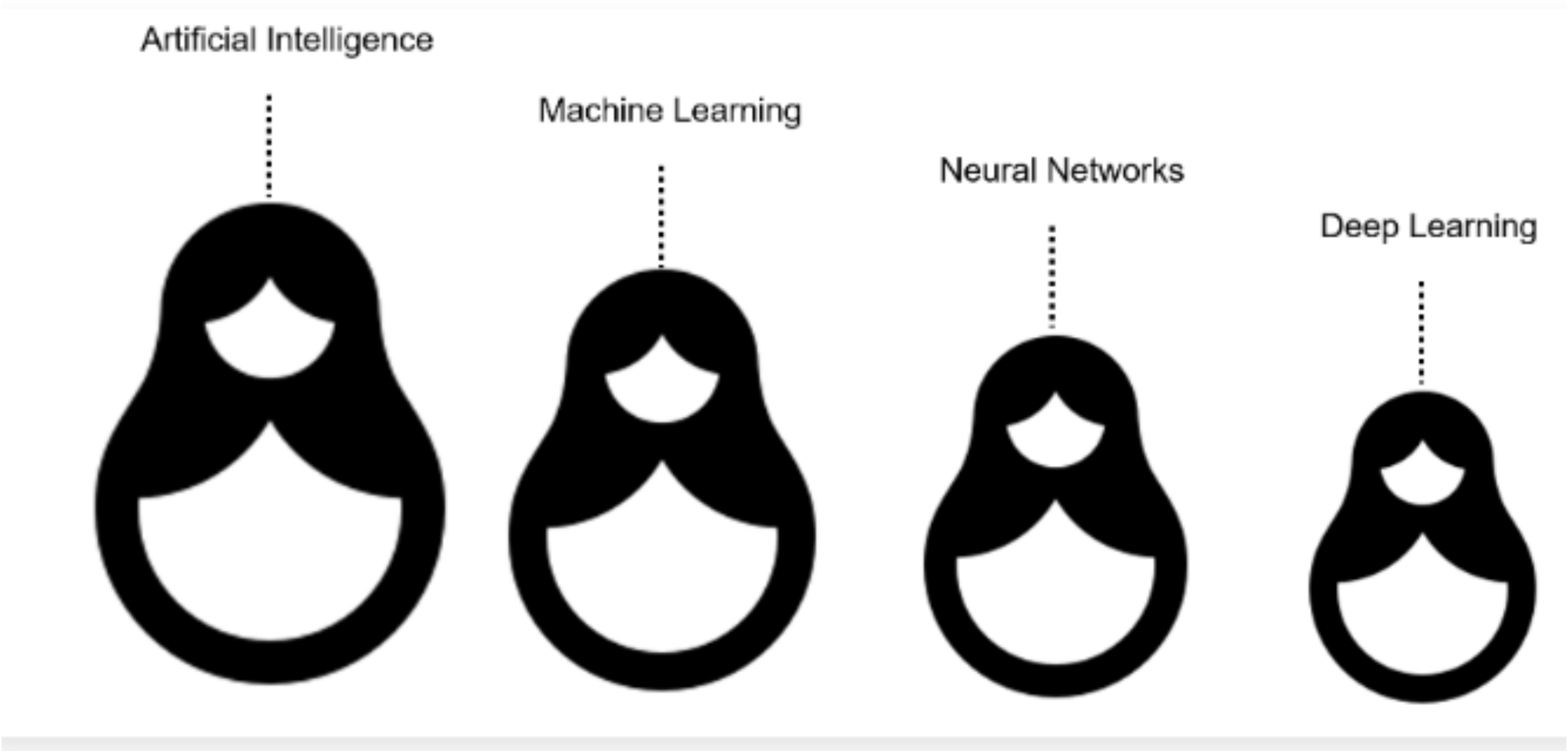


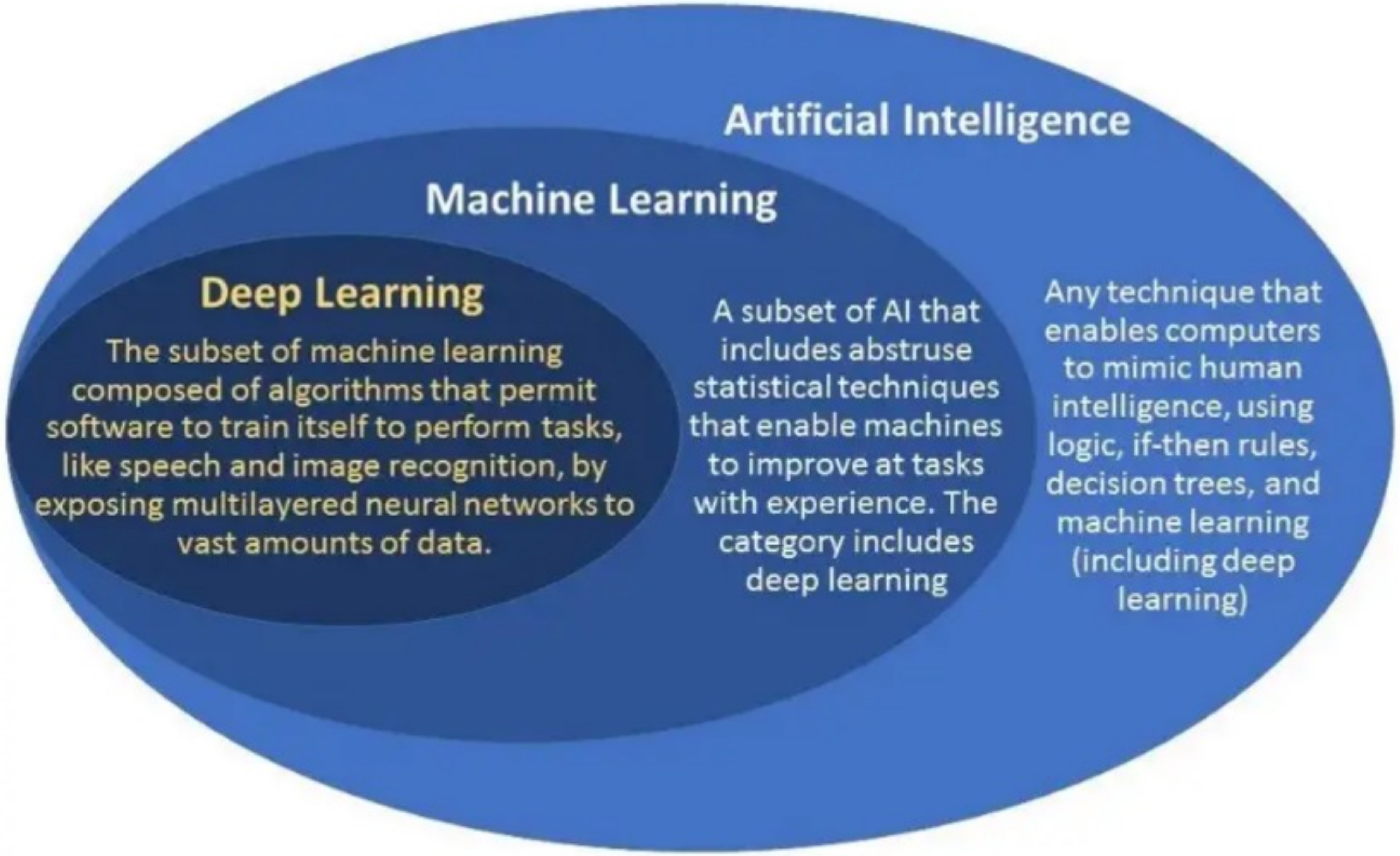
Steps in Using a Neural Network as a Predictive Model

1. Identify the input and output features
2. Transform the inputs and outputs so they are in a small range (0 to 1)
3. Set up a network with an appropriate topology Train the network on a representative set of training examples
4. Use the validation set to choose the set of weights that minimizes the error
5. Evaluate the network using the test set to see how well it performs
6. Apply the model generated by the network to predict outcomes for unknown inputs

How do artificial intelligence, machine learning, neural networks, and deep learning relate?

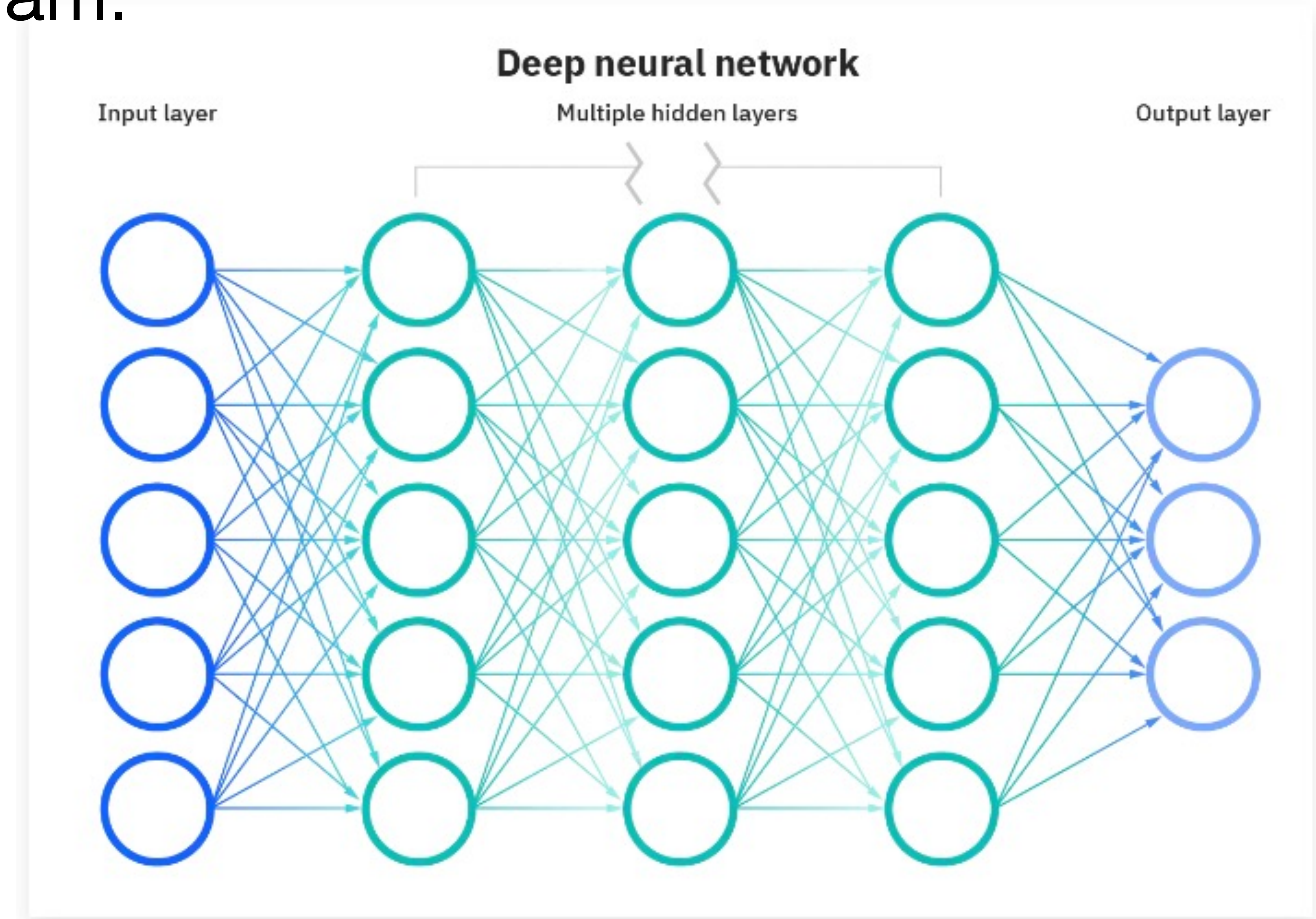
- NN are a type of machine learning technique – other types being: regression, classification, clustering …
- What is Artificial Intelligence – is it same as machine learning?
- Deep learning? – how does deep learning relate to all above?





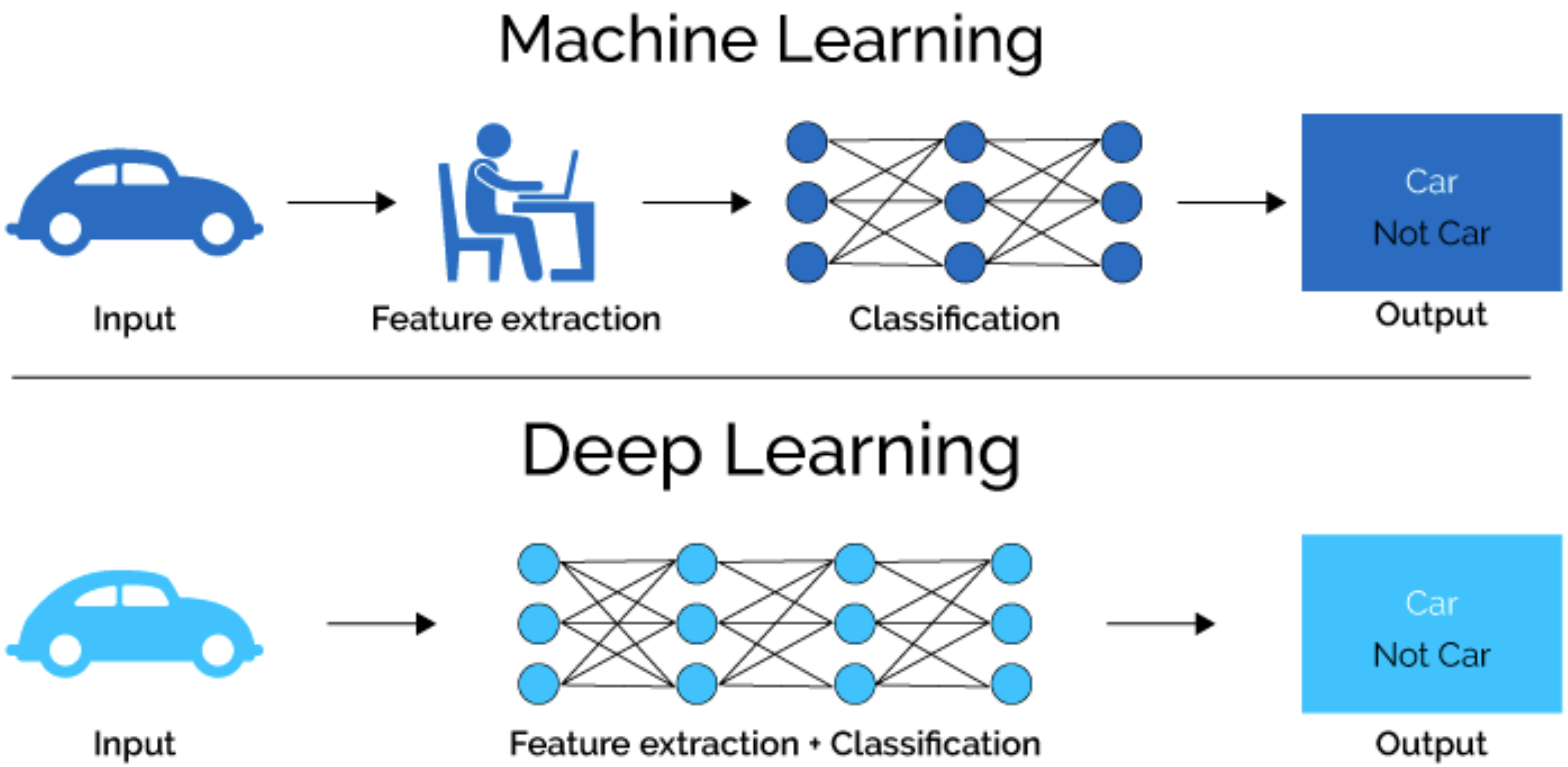
How is deep learning different from neural networks?

- While it was implied within the explanation of neural networks, it's worth noting more explicitly. The "deep" in deep learning is referring to the depth of layers in a neural network. A neural network that consists of more than three layers—which would be inclusive of the inputs and the output—can be considered a deep learning algorithm. This is generally represented using the following diagram:



- Most deep neural networks are feed-forward, meaning they flow in one direction only from input to output. However, you can also train your model through backpropagation; that is, move in opposite direction from output to input. Backpropagation allows us to calculate and attribute the error associated with each neuron, allowing us to adjust and fit the algorithm appropriately.

Automatic feature learning



References

- Data Science for Business, Foster Provost and Tom Fawcett, 1st ed.
- Gordon Linoff and Michael Berry, Data Mining Techniques, 3rd edition, Wiley, 2011

References

- Data Science for Business, Foster Provost and Tom Fawcett, 1st ed.
- Gordon Linoff and Michael Berry, Data Mining Techniques, 3rd edition, Wiley, 2011

- Video links:
- https://www.youtube.com/watch?v=4HKqjENq9OU - KNN Algorithm In Machine Learning | KNN Algorithm Using Python | K Nearest Neighbor | by Simplilearn
- https://www.youtube.com/watch?v=ob1yS9g-Zcs
- https://www.youtube.com/watch?v=9dFhZFUkzuQ - Neural Network Full Course | Neural Network Tutorial For Beginners | Neural Networks | Simplilearn

- Machine Learning vs Deep Learning vs Artificial Intelligence | ML vs DL vs AI | Simplilearn

Artificial intelligence (AI) Vs M/C Learning

- Artificial intelligence (AI) is the broadest term used to classify machines that mimic human intelligence. It is used to predict, automate, and optimize tasks that humans have historically done, such as speech and facial recognition, decision making, and translation.
- AI is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable

- Machine learning allows experts to “train” a machine by making it analyze massive datasets. The more data the machine analyzes, the more accurate results it can produce by making decisions and predictions for unseen events or scenarios.
- Machine learning models need structured data to make accurate predictions and decisions. If the data is not labeled and organized, machine learning models fail to comprehend it accurately, and it becomes a domain of deep learning.
- Deep Learning : Machine learning models need human intervention to improve accuracy. On the contrary, deep learning models improve themselves after each result without human supervision. But it often requires more detailed and lengthy volumes of data.
- The deep learning methodology designs a sophisticated learning model based on neural networks inspired by the human mind. These models have multiple layers of algorithms called neurons. They continue to improve without human intervention, like the cognitive mind that keeps improving and evolving with practice, revisits, and time.