Machine Learning - Part 2

Apr. 10, 2025

Recap question:

April 10, 2025

You are a new employee at Apple and are given the task of designing a regression algorithm to predict how much people in the Bay Area spend on groceries every month. What are some plausible independent variables you could use in the regression?

Recap question:

Some examples of plausible independent variables are

- Size of household
- Neighborhood they live in
- Income
- Age
- Number of hours spent working every week
- …

Machine Learning - Part 2

April 10, 2025

By the end of this lecture, you will be able to:

1. Give examples of supervised and unsupervised learning
2. Define a neural network
3. Train a neural network

Two categories of ML algorithms

1. **Supervised learning**: algorithms are trained based on example inputs that are labeled with their desired outputs by humans.

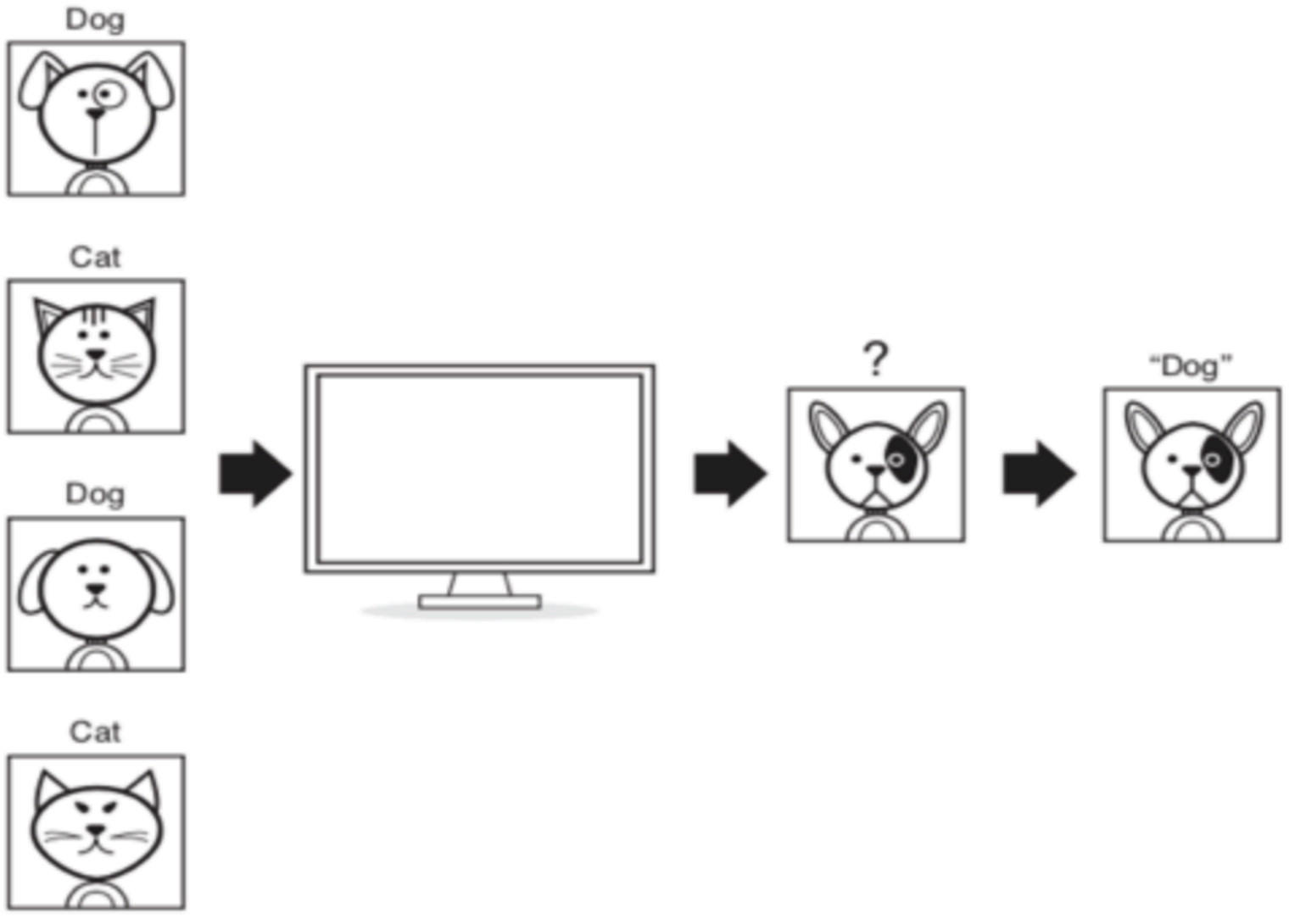
Coins vs cats

Two categories of ML algorithms

1. **Supervised learning**: algorithms are trained based on example inputs that are labeled with their desired outputs by humans.

2. **Unsupervised learning**: the input data is not labeled and algorithms are expected to find structure within the input data by itself.

Example 1: Image classification



Example 2: Real estate pricing

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

3	2000	Hipsterton	???

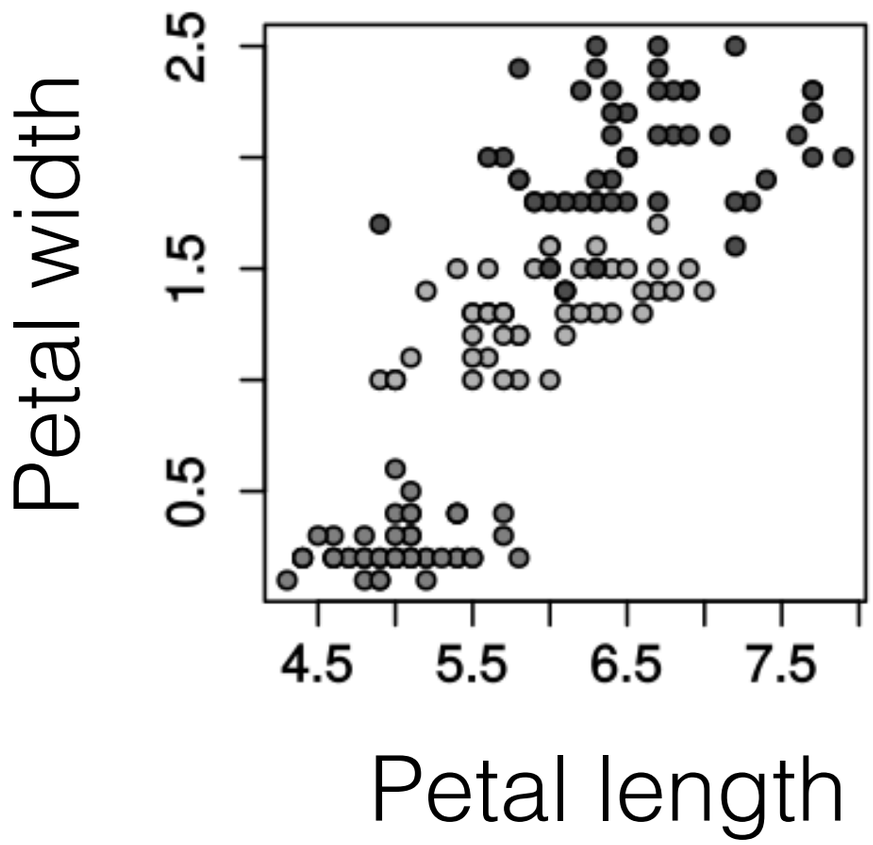
Unsupervised learning

Example 1: Clustering unlabeled data into groups. We have a database of pictures of flowers. Can we find out how many different species there are in the database?



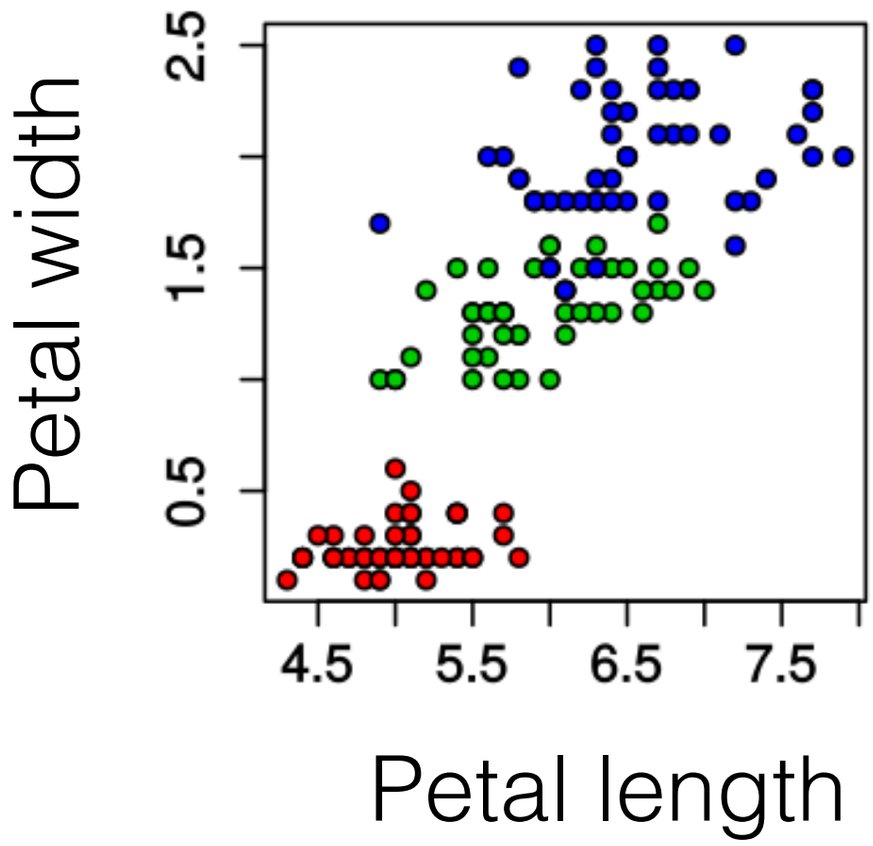
Unsupervised learning

Example 1: Clustering unlabeled data into groups. We have a database of pictures of flowers. Can we find out how many different species there are in the database?



Unsupervised learning

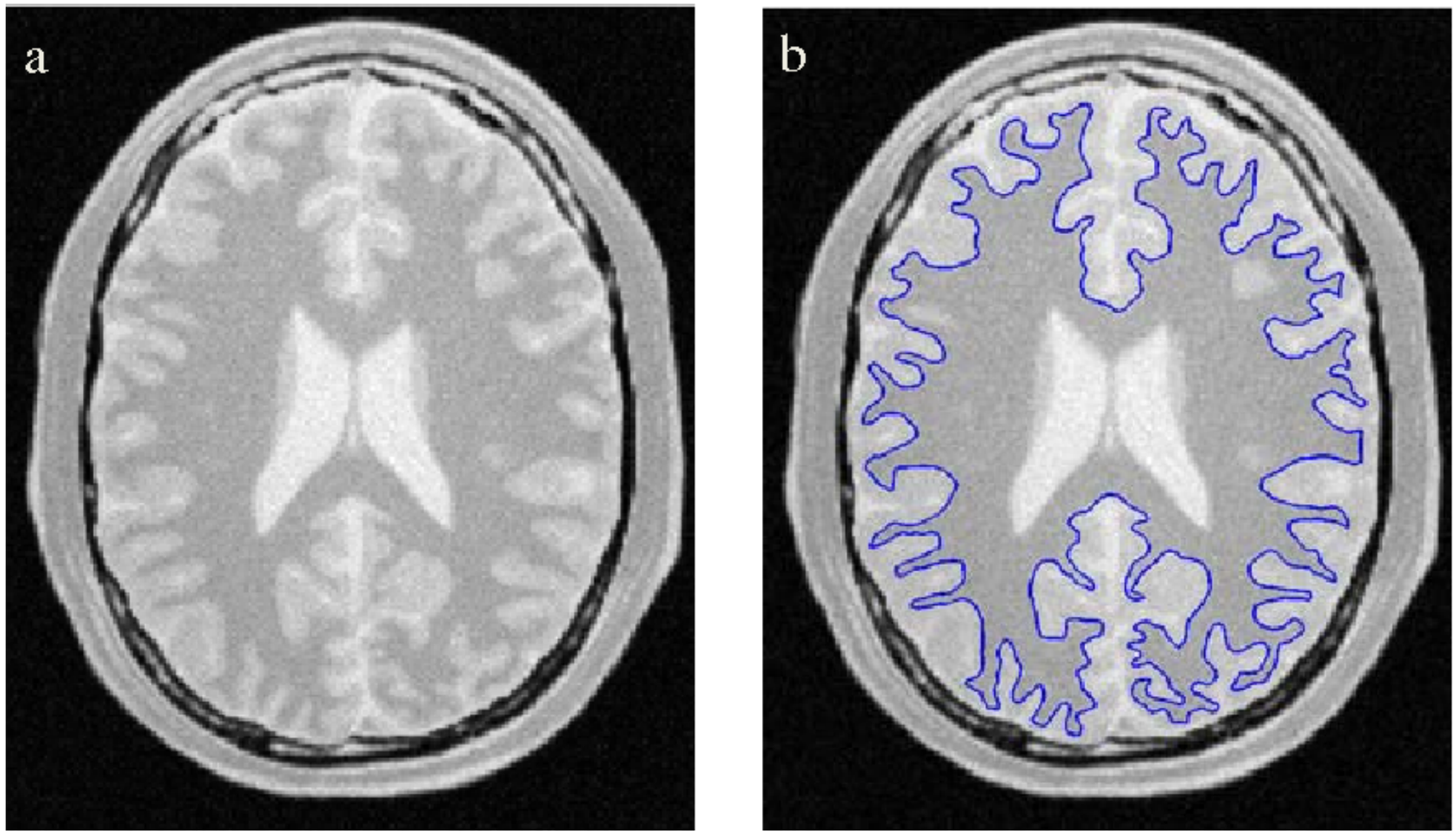
Example 1: Clustering unlabeled data into groups. We have a database of pictures of flowers. Can we find out how many different species there are in the database?



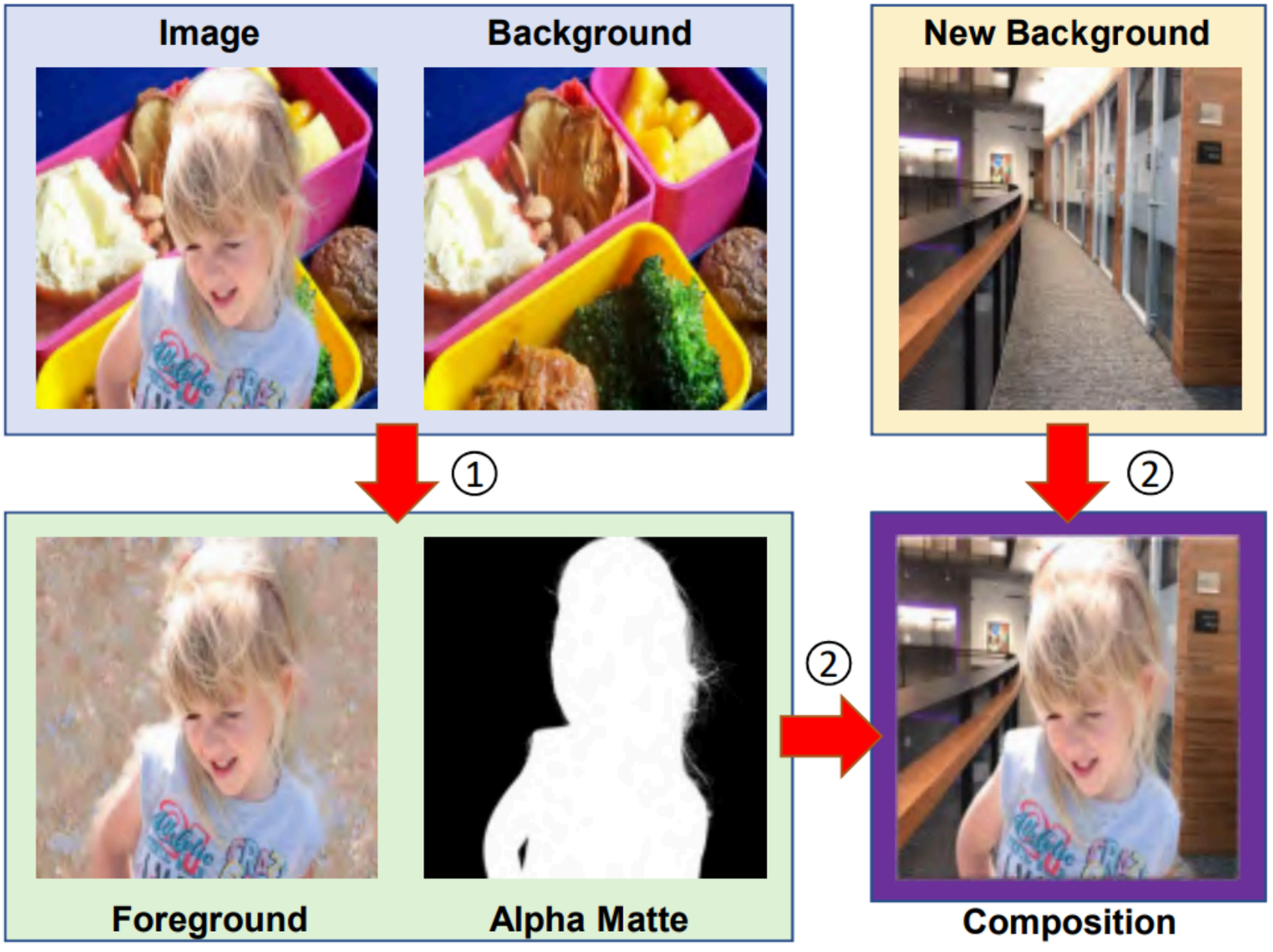
3 species!

Example 2: analyzing medical images

Distinguish regions of interest in medical images, without giving labels to the algorithm.



Example 3: extracting regions of interest of pictures



Reinforcement Learning

Example: AlphaZero: a Google project to make computers better than humans at board games, by making the computer play against itself repeatedly.

Reinforcement Learning

Example: AlphaZero: a Google project to make computers better than humans at board games, by making the computer play against itself repeatedly.

Results against previous best computer programs:

Example: AlphaZero: a Google project to make board games, against itself

BBC Sign in Home News Sport Reel Worklife Travel

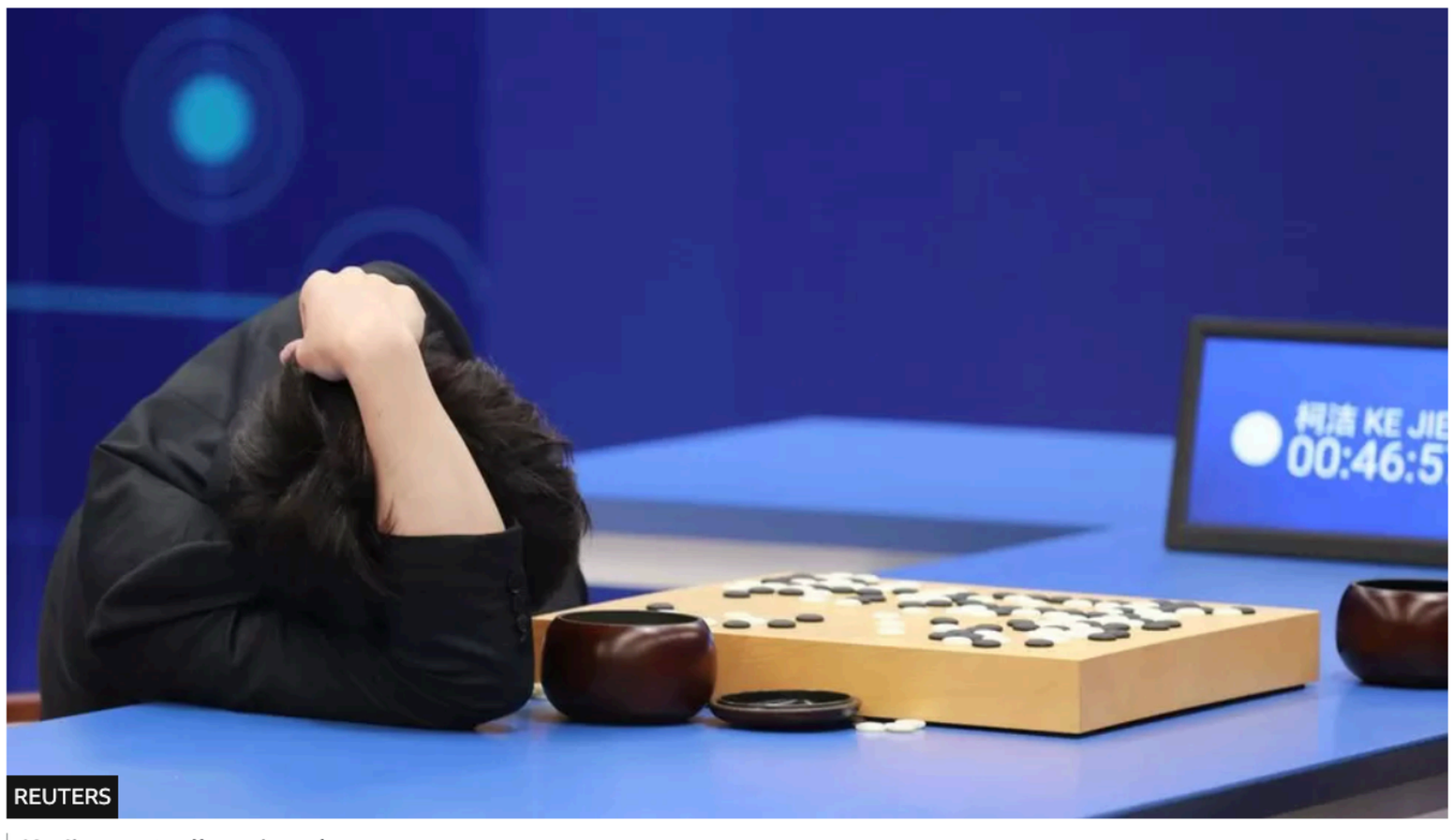
NEWS

Home | War in Ukraine | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech

Tech

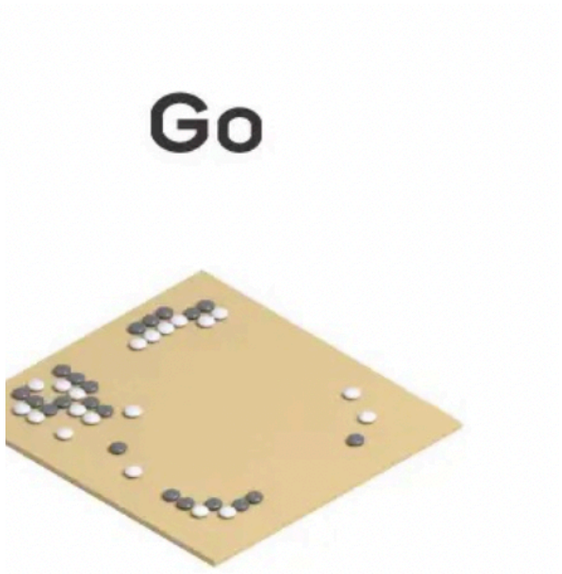
Google AI defeats human Go champion

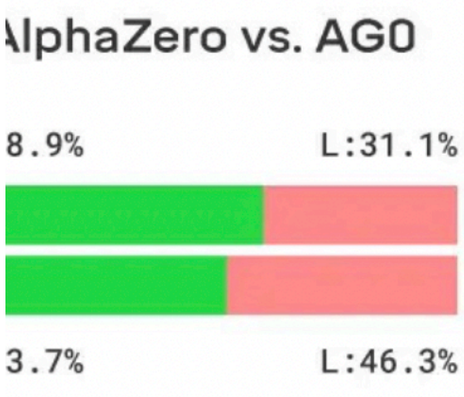
25 May 2017



REUTERS

Ke Jie eventually resigned





AZ wins AZ draws AZ loses AZ white AZ black

Neural networks

Back to the example of home value estimates

We use regression to estimate the sale price from the independent variables.

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

x_1 x_2 x_3 y

Back to the example of home value estimates

We use regression to estimate the sale price from the independent variables.

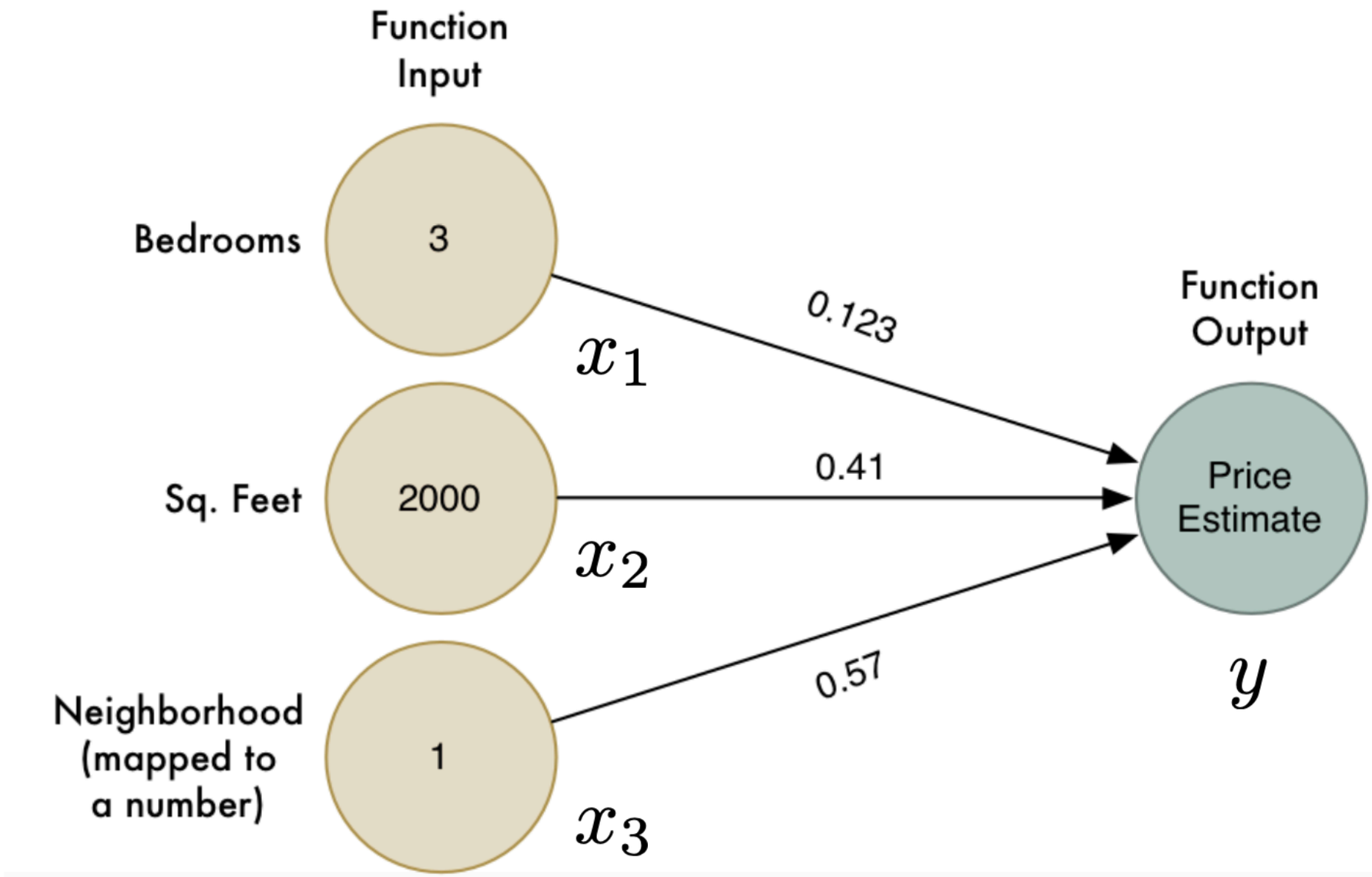
Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

x_1 x_2 x_3 y

Linear regression uses a formula of this form:

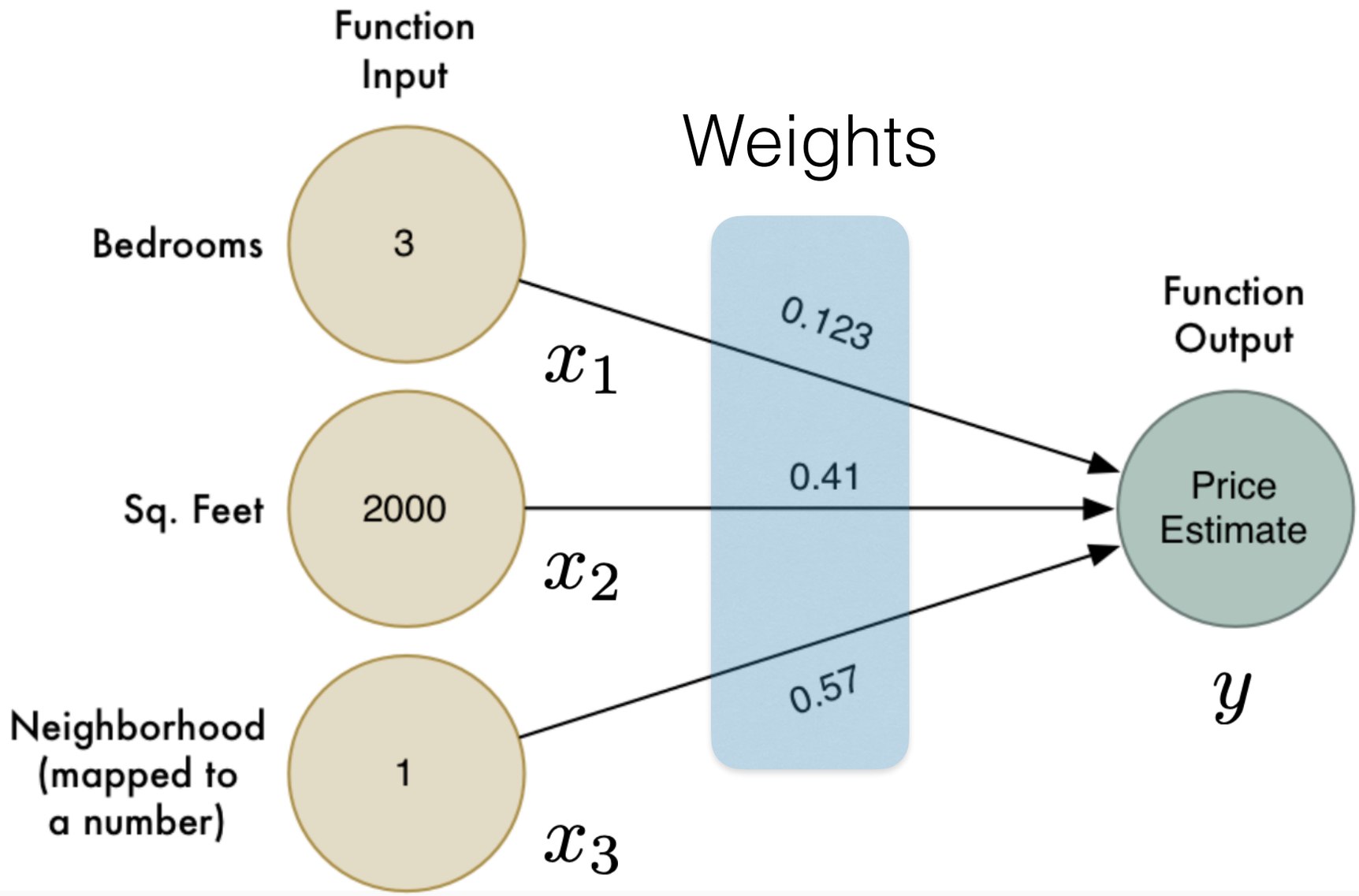
$$y = 0.123x_1 + 0.41x_2 + 0.57x_3$$

Price estimate from one real estate agent



$$y = 0.123x_1 + 0.41x_2 + 0.57x_3$$

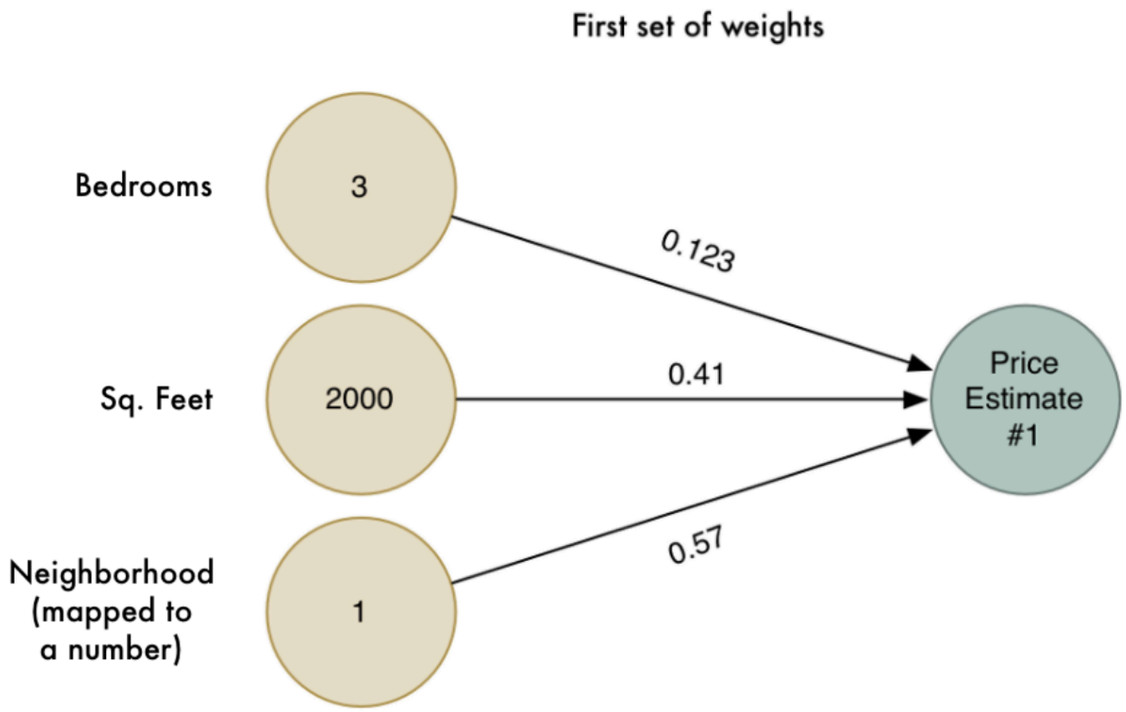
Price estimate from one real estate agent



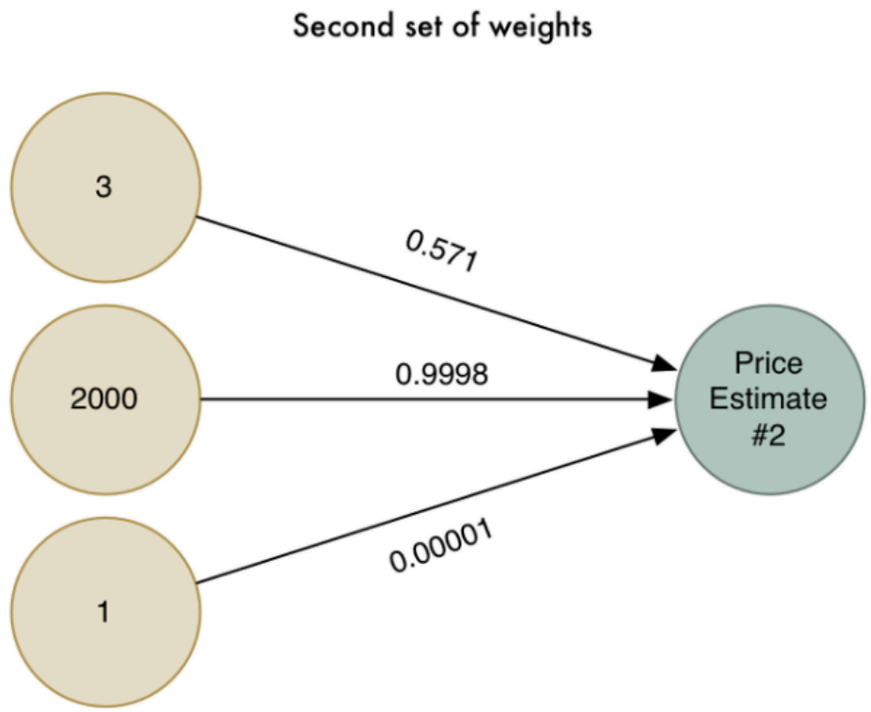
$$y = 0.123x_1 + 0.41x_2 + 0.57x_3$$

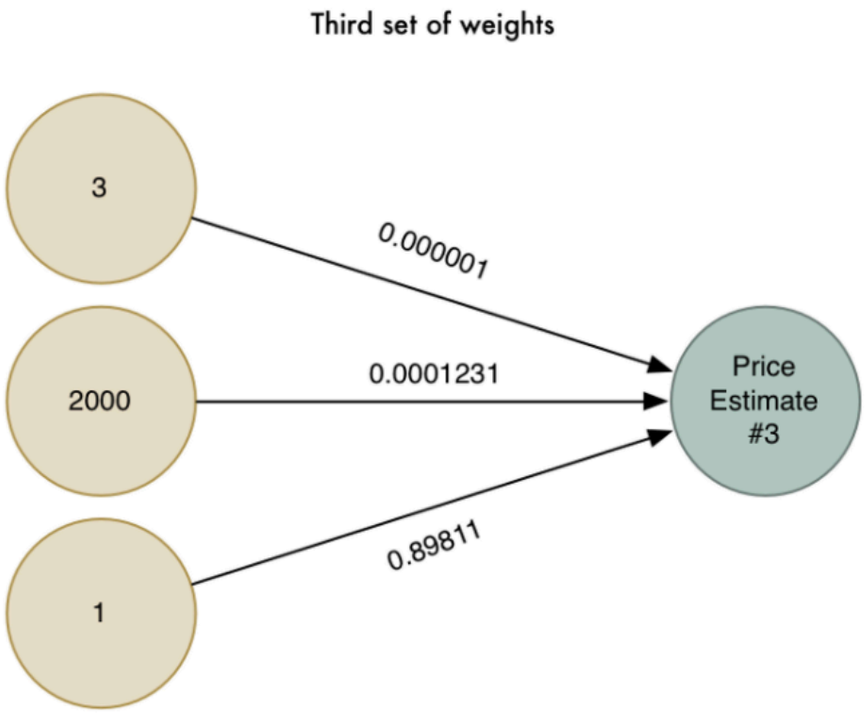
A single real-estate agent can give an inaccurate answer. It is probably a good idea to ask multiple agents to get a second opinion, third opinion, … .

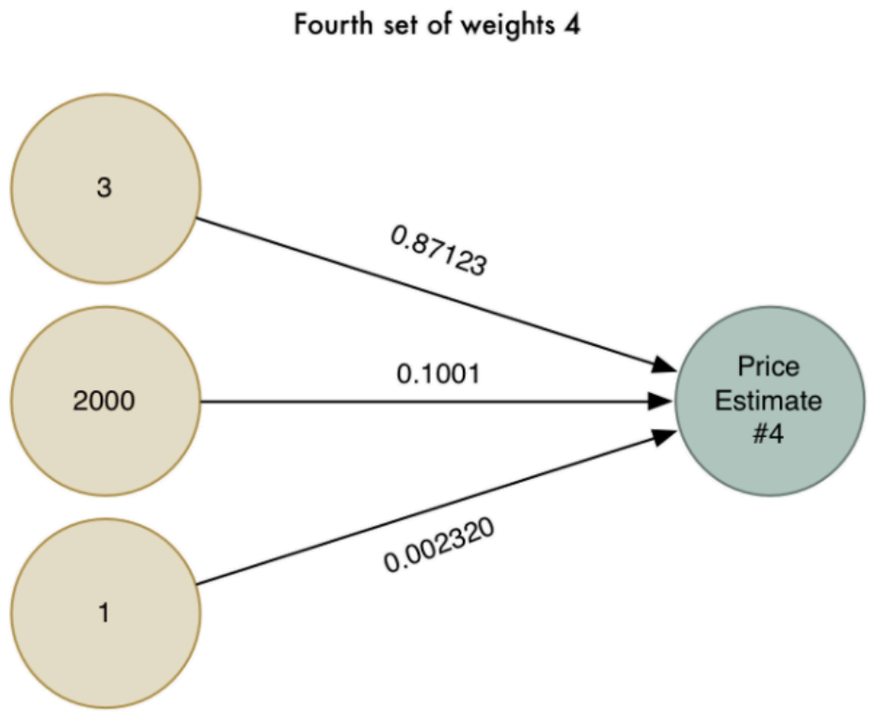
Price estimate from agent 1



Price estimate from agent 2





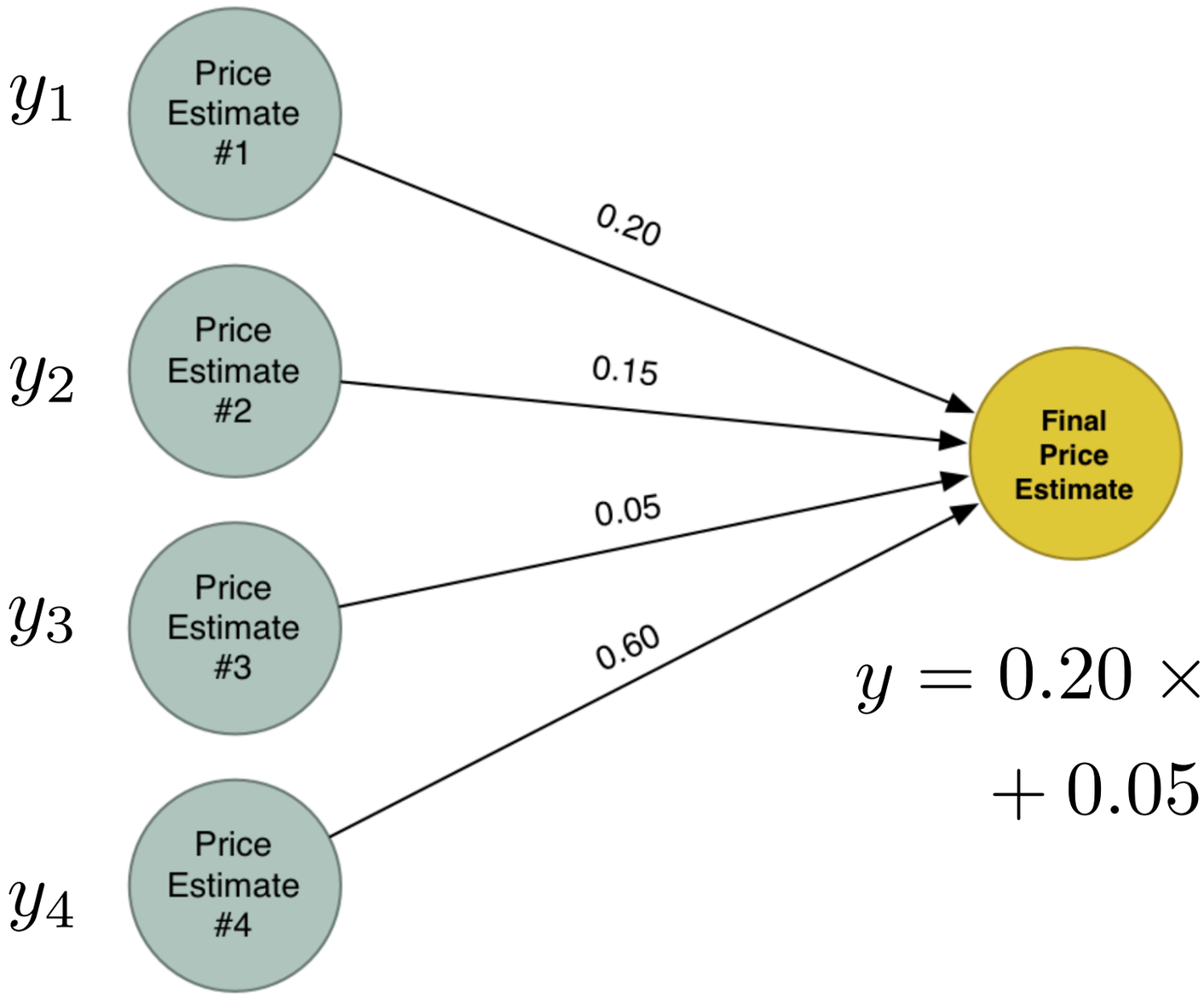


Price estimate from agent 3

Price estimate from agent 4

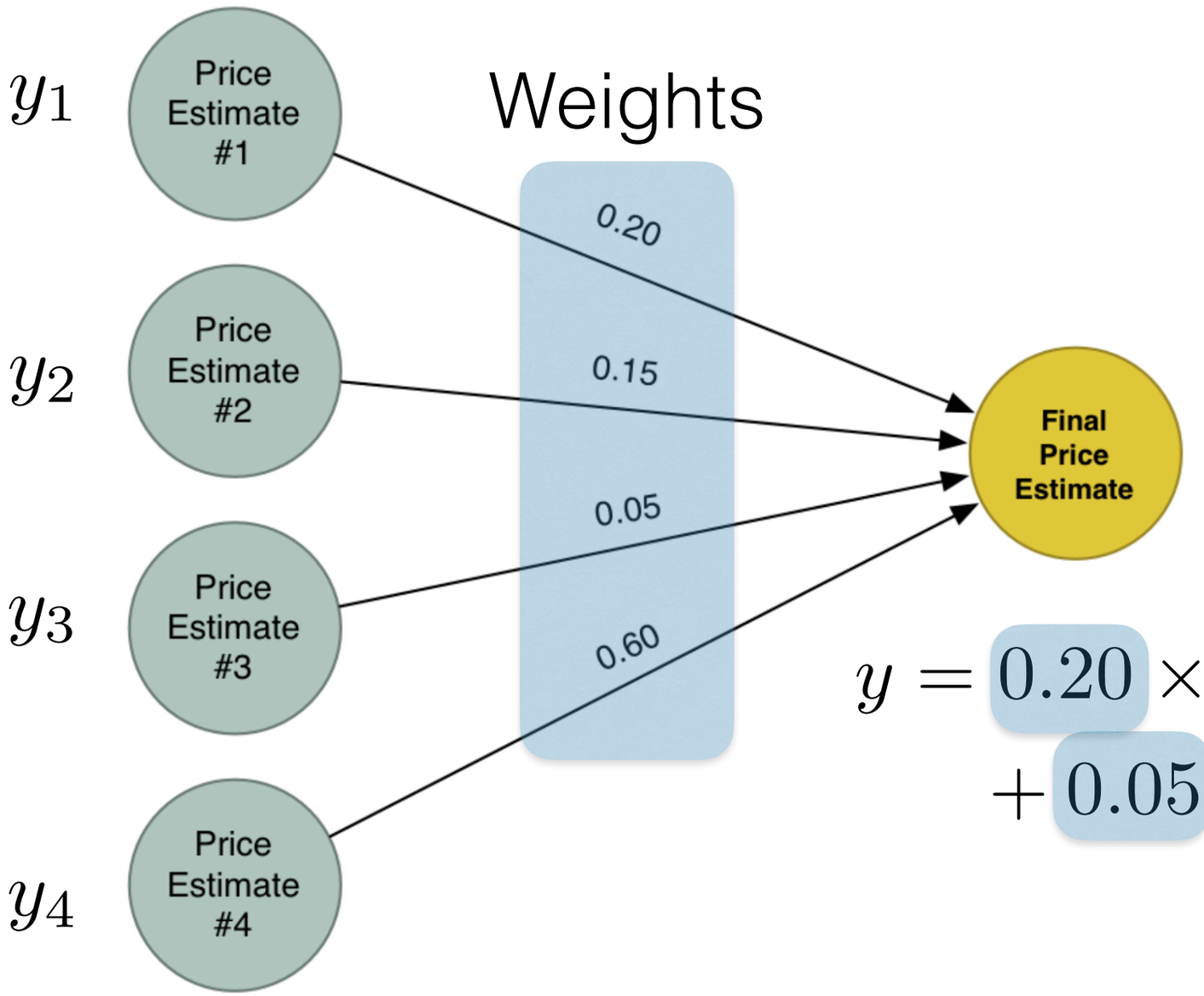
But how do we combine these estimates? One agent could be more reliable than another, so we can assign different weights to their opinions.

Combine their estimates to get an accurate price

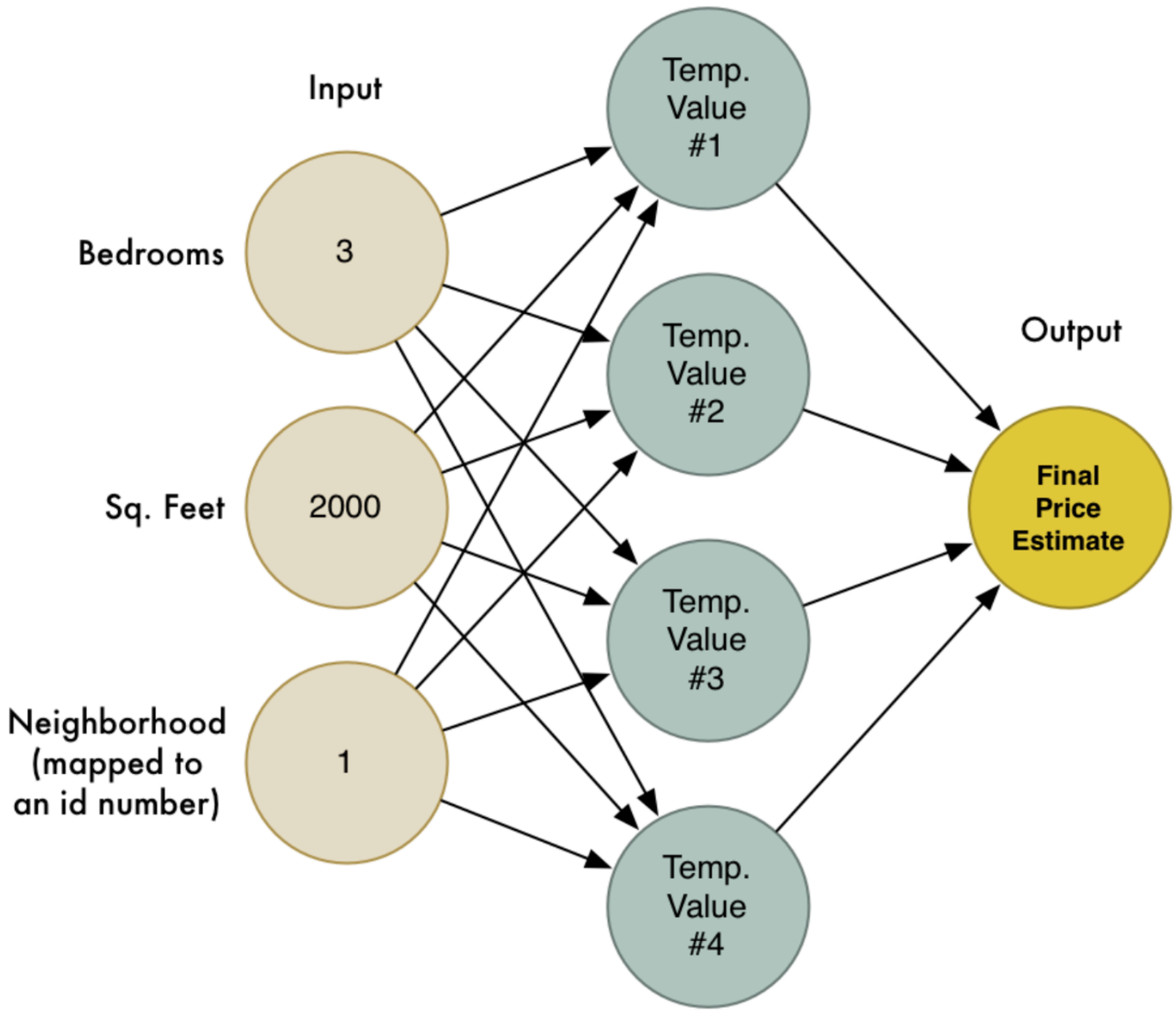


$$y = 0.20 \times y_1 + 0.15 \times y_2 + 0.05 \times y_3 + 0.6 \times y_4$$

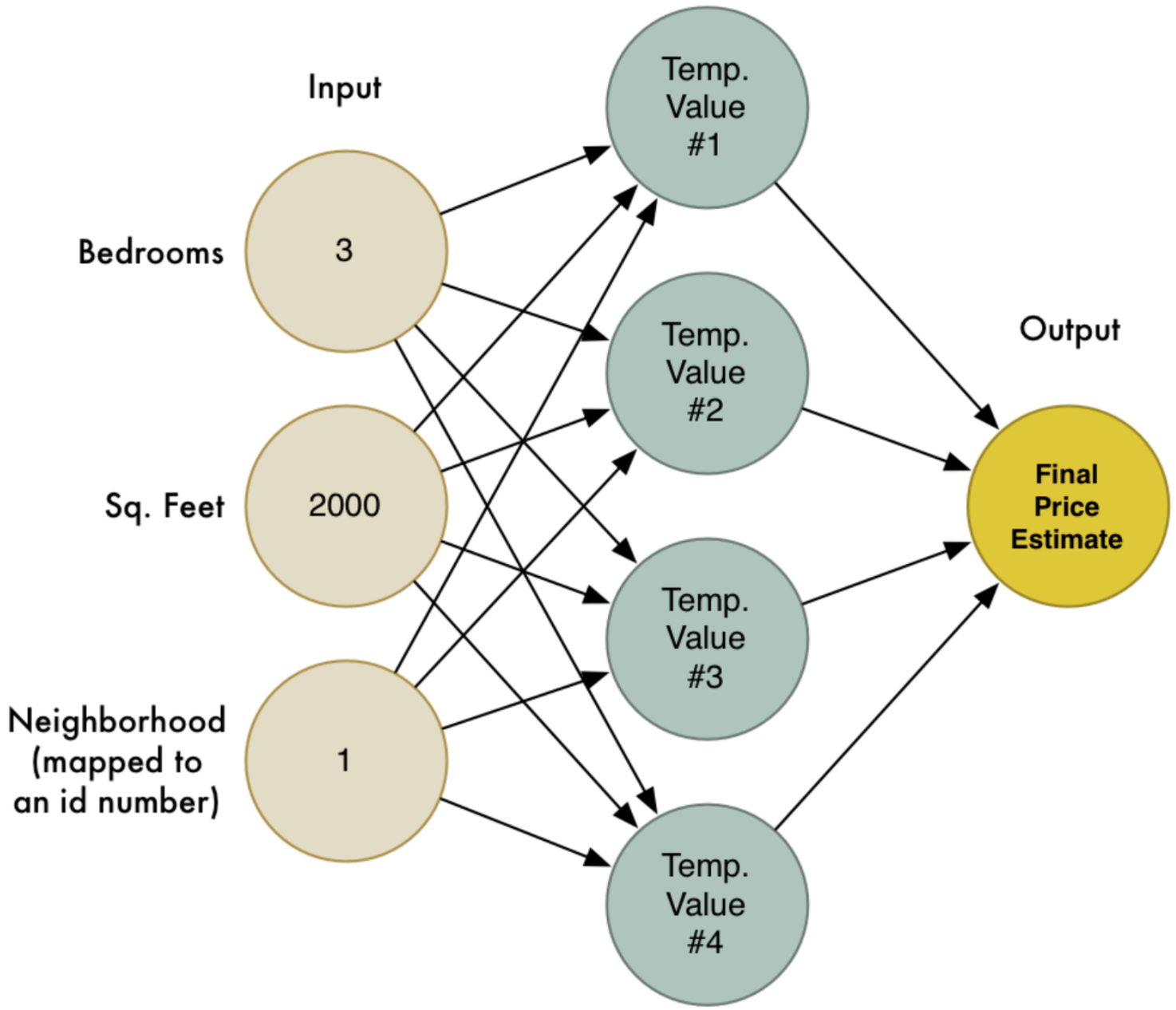
Combine their estimates to get an accurate price



$$y = 0.20 \times y_1 + 0.15 \times y_2 + 0.05 \times y_3 + 0.6 \times y_4$$



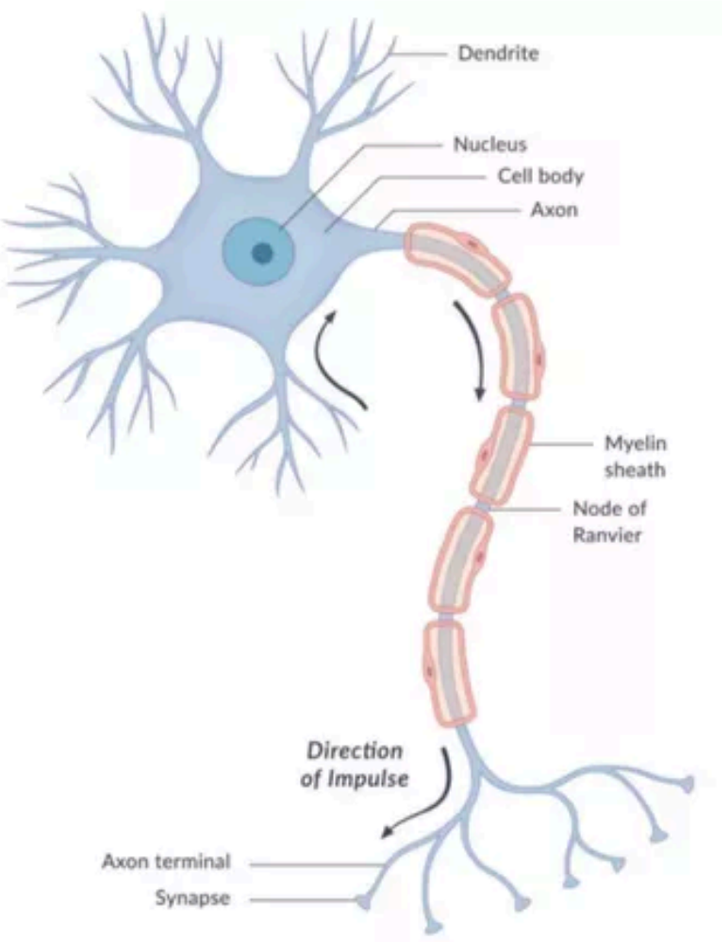
By chaining lots of real estate agents together, we can model functions that are too complicated to be modeled by a single real estate agent.



This is (almost) an example of a neural network!

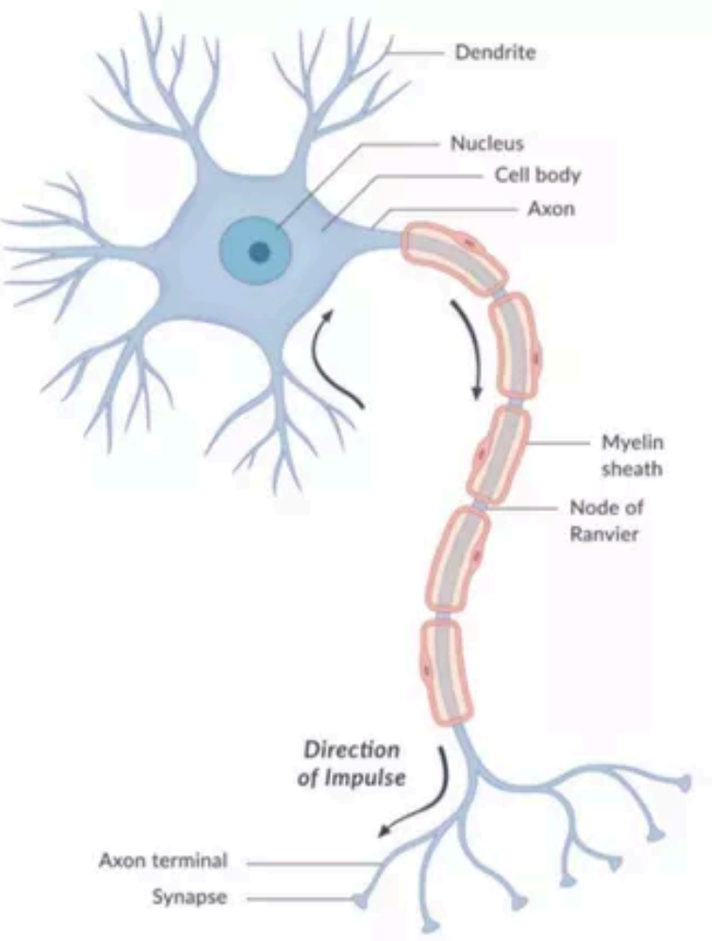
What is a Neural network?

Artificial neural networks (ANN) began with Warren McCulloch and Walter Pitts[1] (1943). It is a technique for building algorithms that learn from data. It is very loosely inspired by how we think the human brain works.



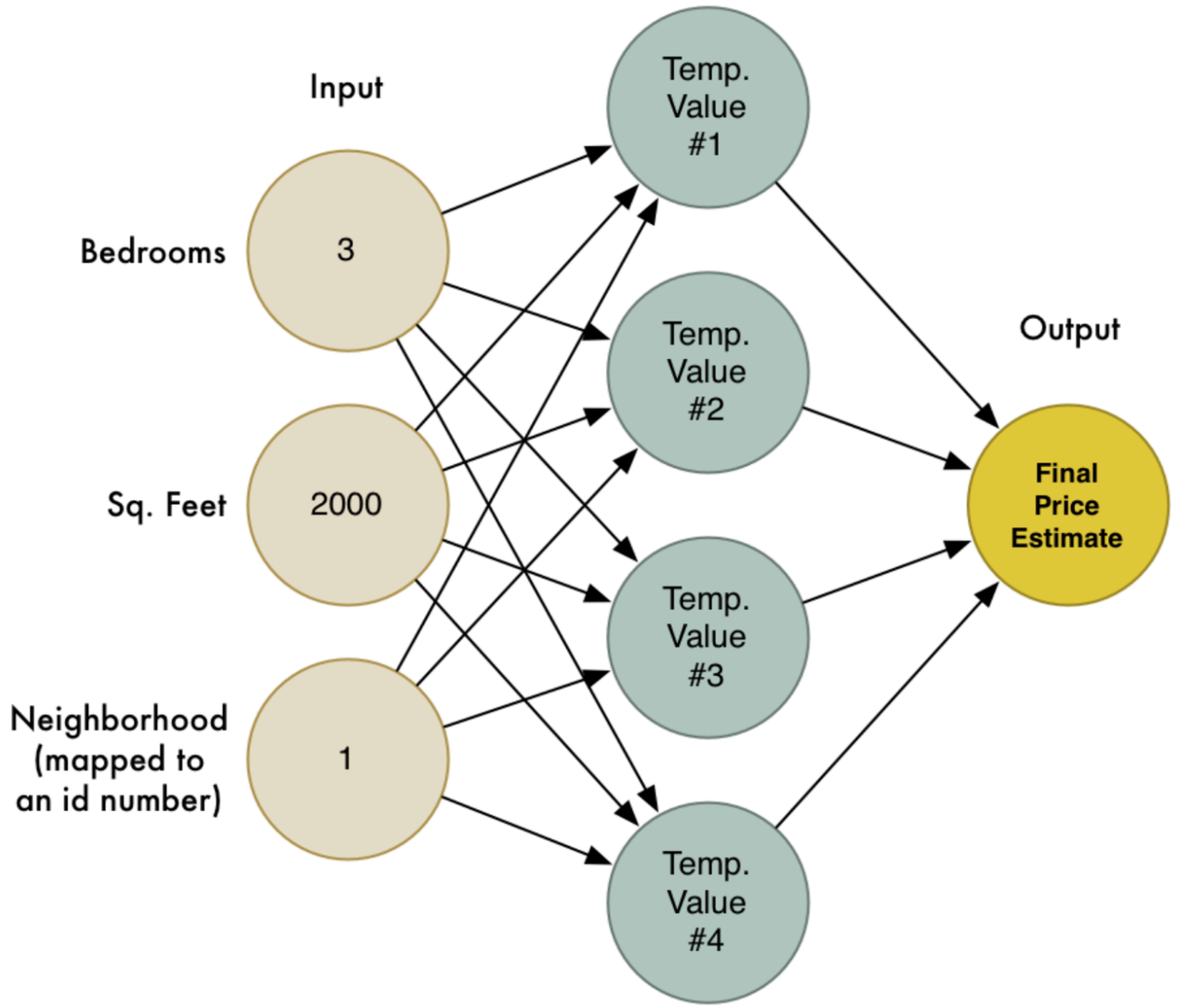
What is a Neural network?

Artificial neural networks (ANN) began with Warren McCulloch and Walter Pitts[1] (1943). It is a technique for building algorithms that learn from data. It is very loosely inspired by how we think the human brain works.



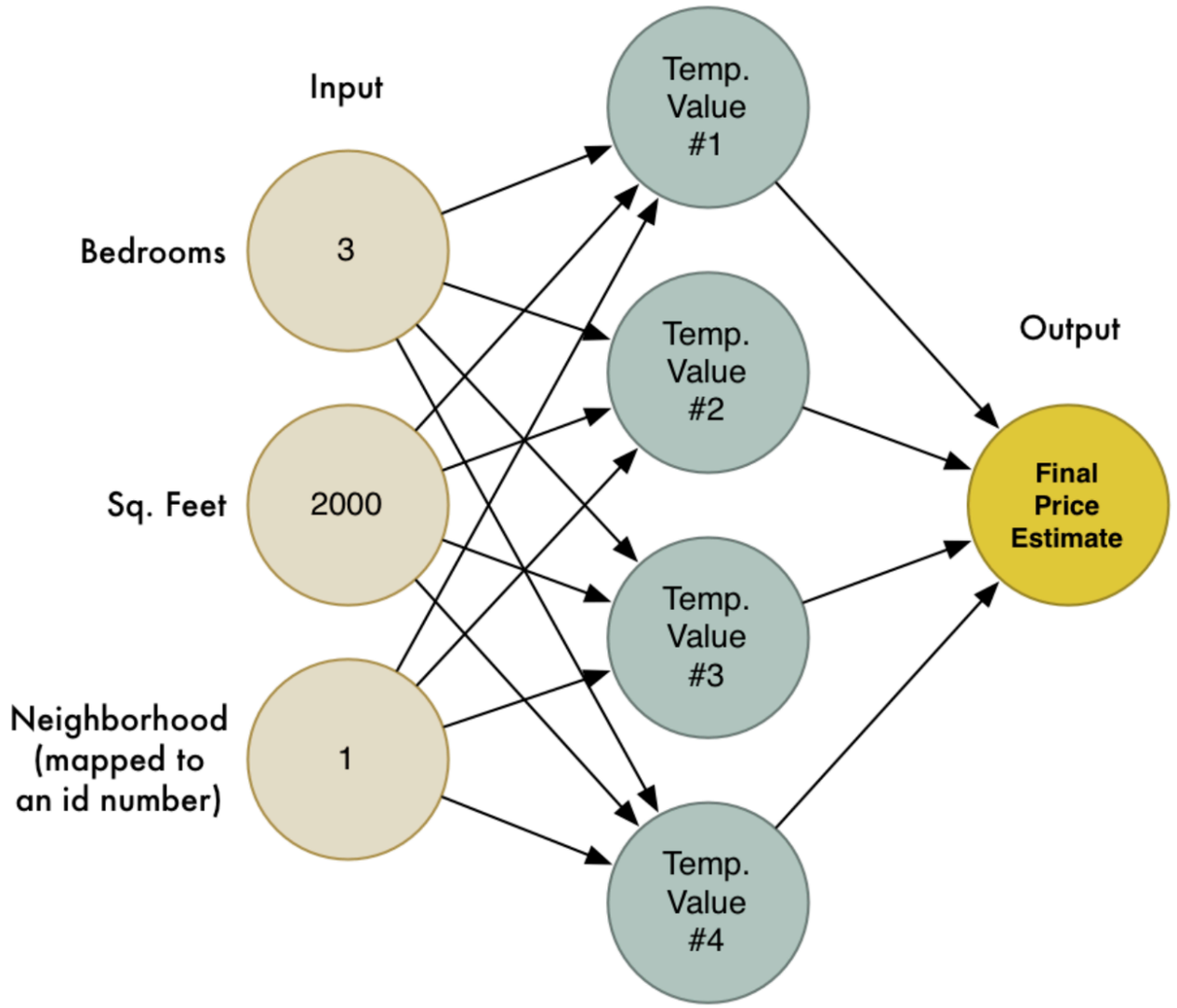
A neural network takes inputs (numbers), and outputs numbers.

An example of neural network



Each circle is called a **neuron**

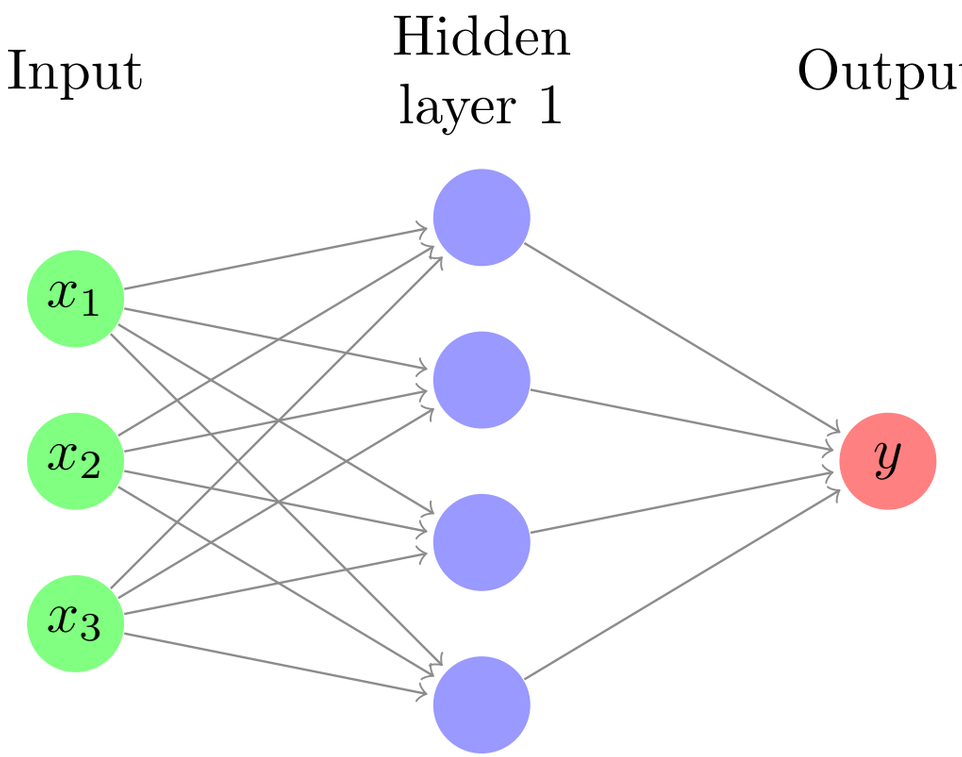
An example of neural network



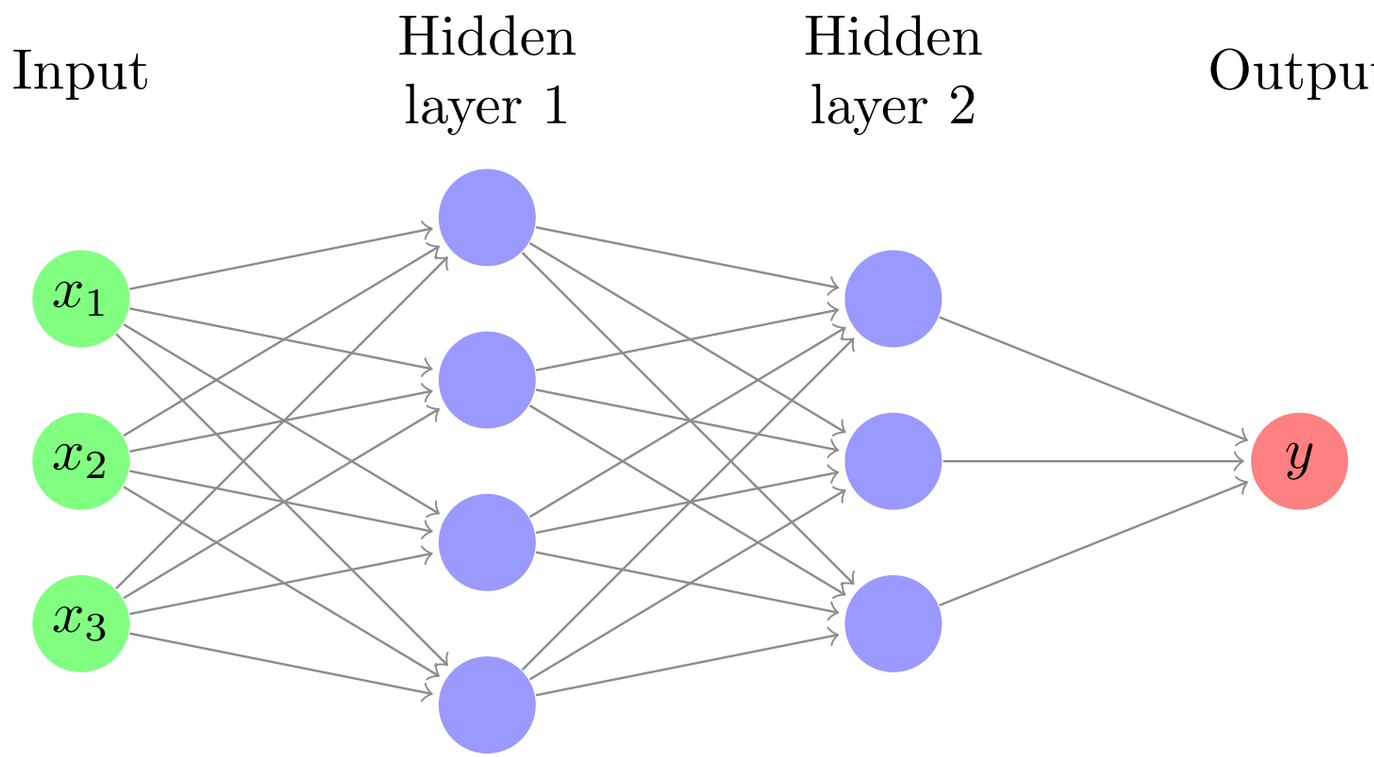
The neurons between the input and output are arranged in so called **hidden layers**

An example of neural network

One hidden layer

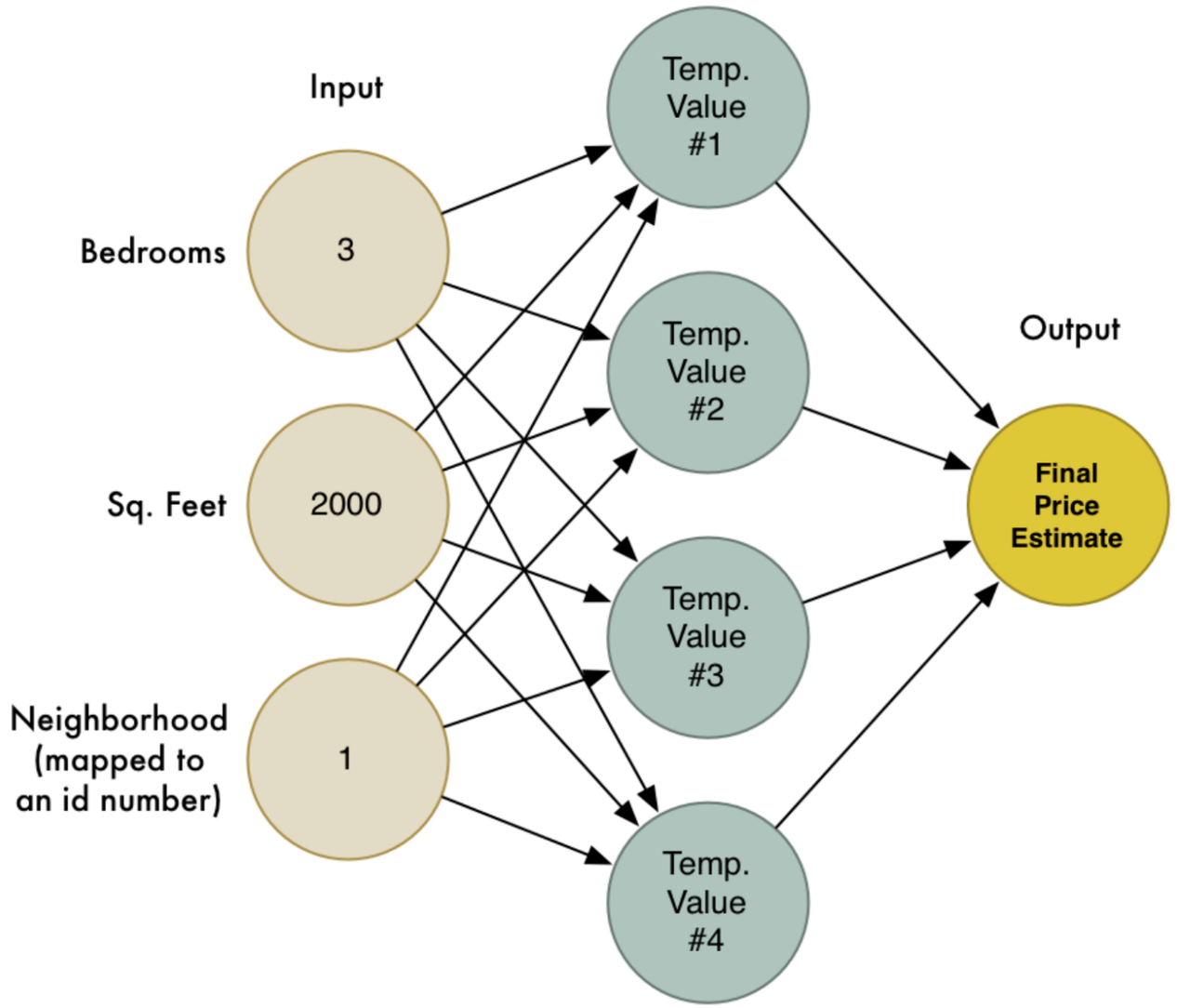


Two hidden layers

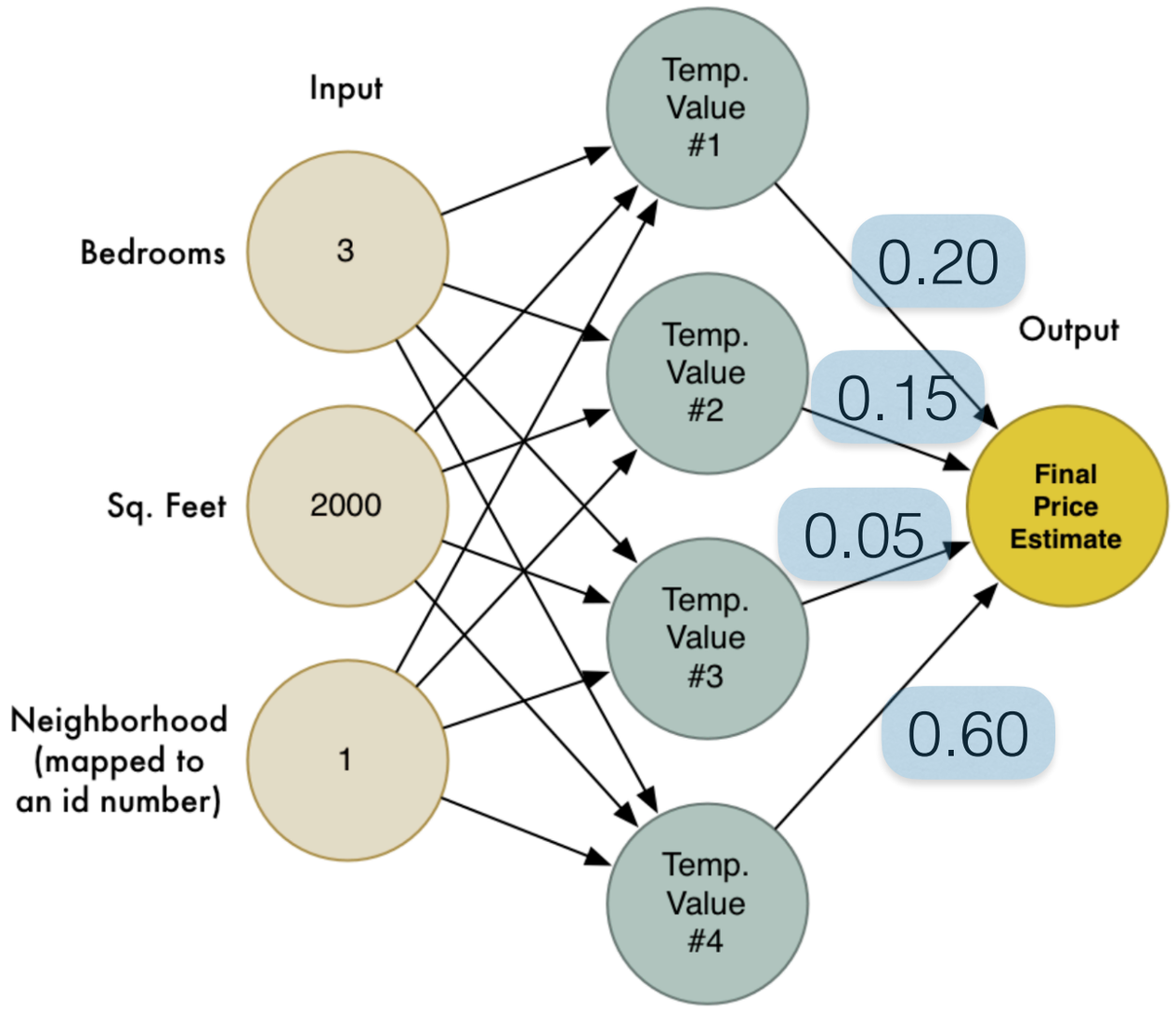


The neurons between the input and output are arranged in so called **hidden layers**

An example of neural network



Each arrow has its own **weight** (a number)



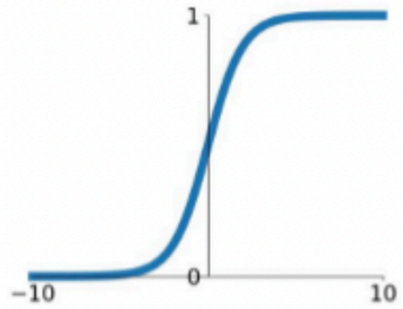
For each neuron, we multiply the value coming in from each arrow by its corresponding weight. We add these together and then apply an **activation function** to it. This becomes the **value** of that neuron.

Use nonlinear activation functions

Activation Functions

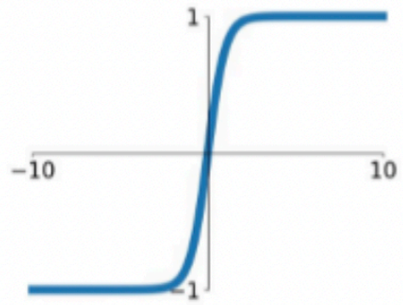
Sigmoid

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



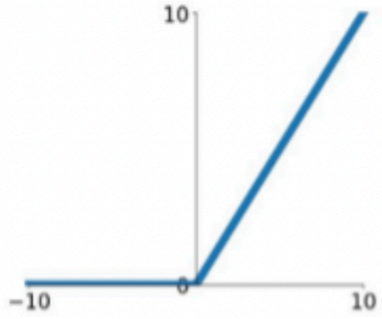
tanh

$$\tanh(x)$$



ReLU

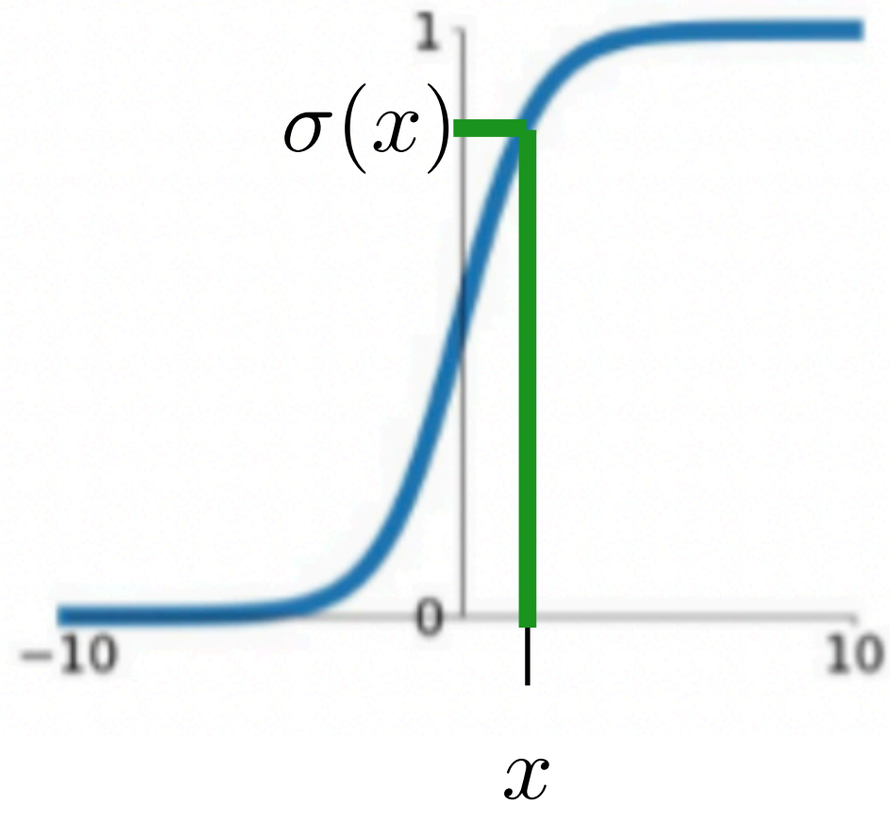
$$\max(0, x)$$



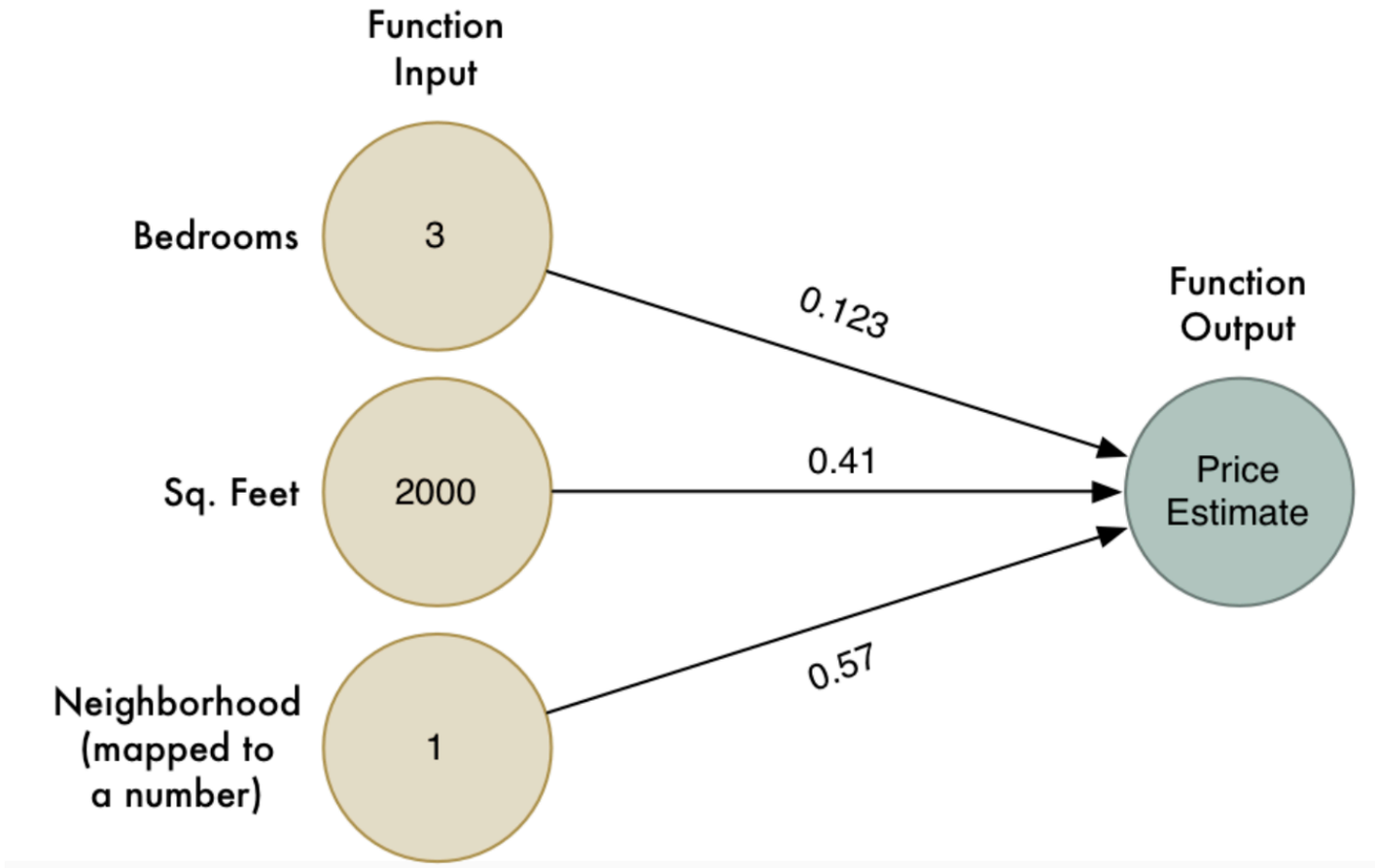
Use nonlinear activation functions

Sigmoid

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



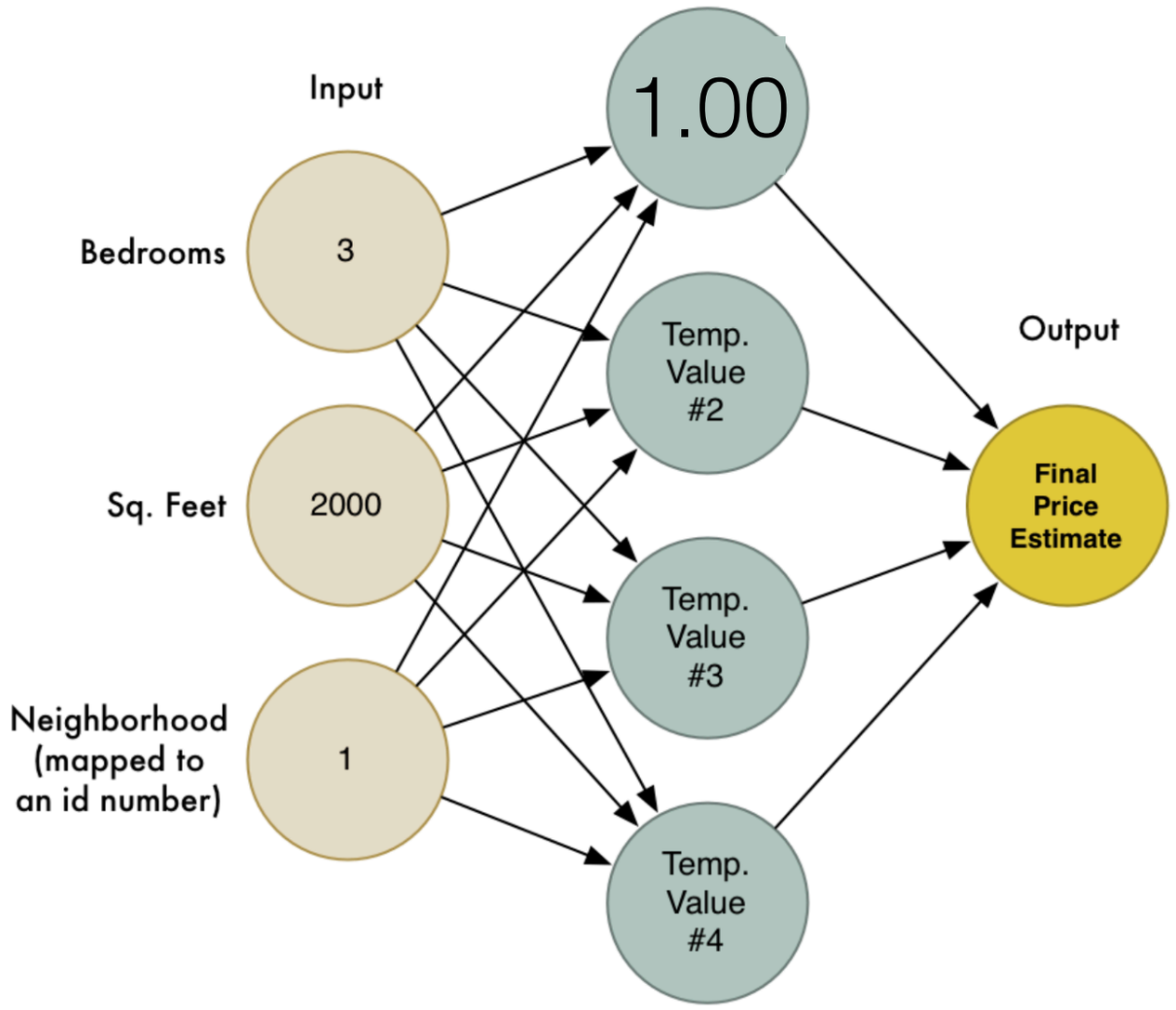
Example: computing the value of the first neuron for the price estimate



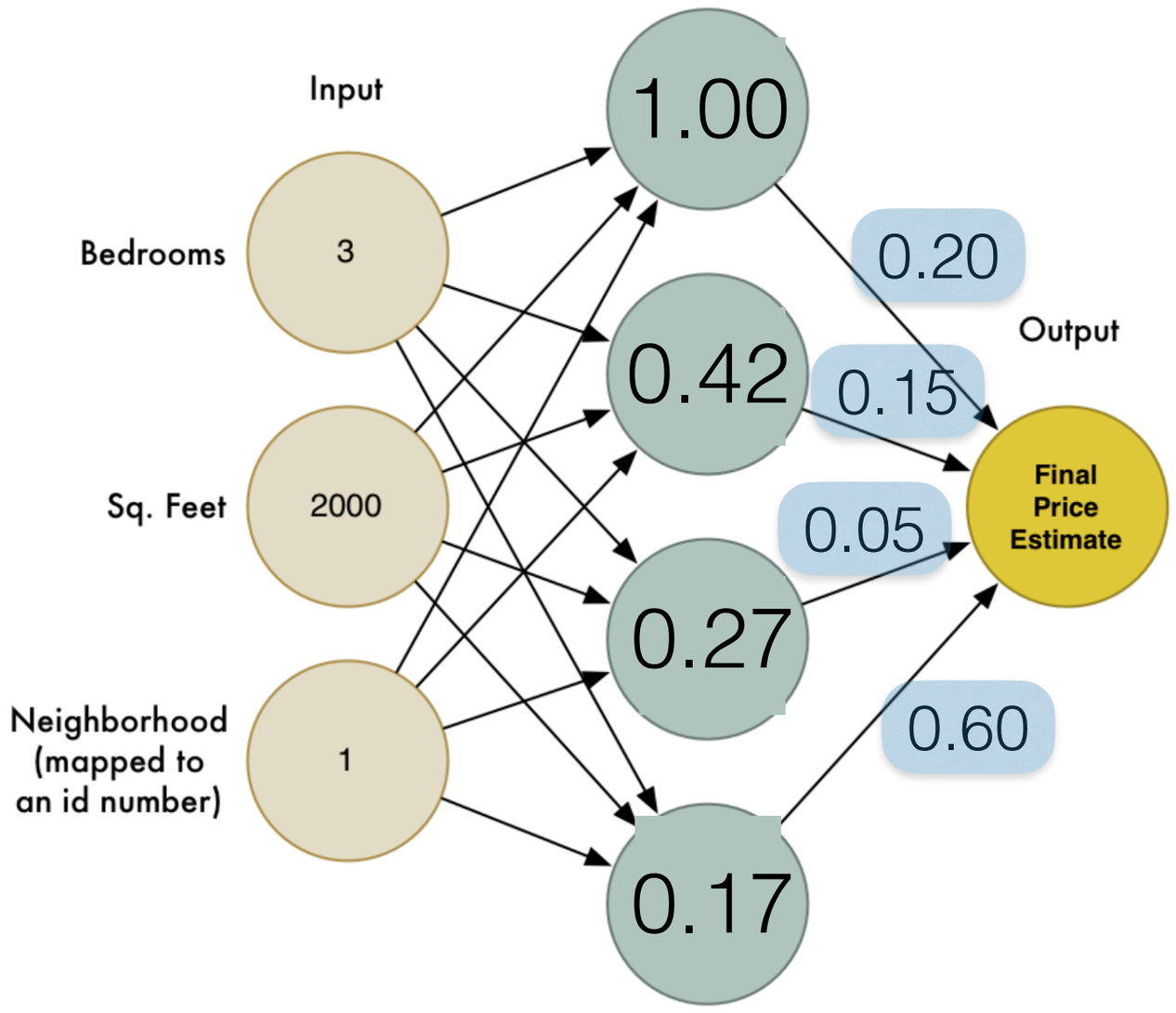
$$\text{Price estimate } = \tanh(0.123 \times 3 + 0.41 \times 2000 + 0.57 \times 1)$$

$$= \tanh(820.939) = 1.00$$

Example:


Input
Bedrooms
3
Sq. Feet
2000
Neighborhood
(mapped to
an id number)
1
1.00
Temp.
Value
#2
Temp.
Value
#3
Temp.
Value
#4
Output
Final
Price
Estimate

Example:



Final price estimate $= \tanh(0.20 \times 1 + 0.15 \times 0.42 + 0.05 \times 0.27 + 0.60 \times 0.17)$

$= \tanh(0.3785) = 0.36$

How do we use neural networks?

An engineer choses the structure of the neural network (number of hidden layers, the number of neurons in each hidden layer and which activation function to use).

How do we use neural networks?

An engineer choses the structure of the neural network (number of hidden layers, the number of neurons in each hidden layer and which activation function to use).

We use **training data** to come up with weights for all arrows. We then test how good our network is by applying it to the **testing data**

Example 1:

Training data

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000

Testing data

Bedrooms	Sq. feet	Neighborhood	Sale price	Predicted price
2	850	Normaltown	$150,000	$133,335
1	550	Normaltown	$78,000	$91,000
4	2000	Skid Row	$150,000	$145,000

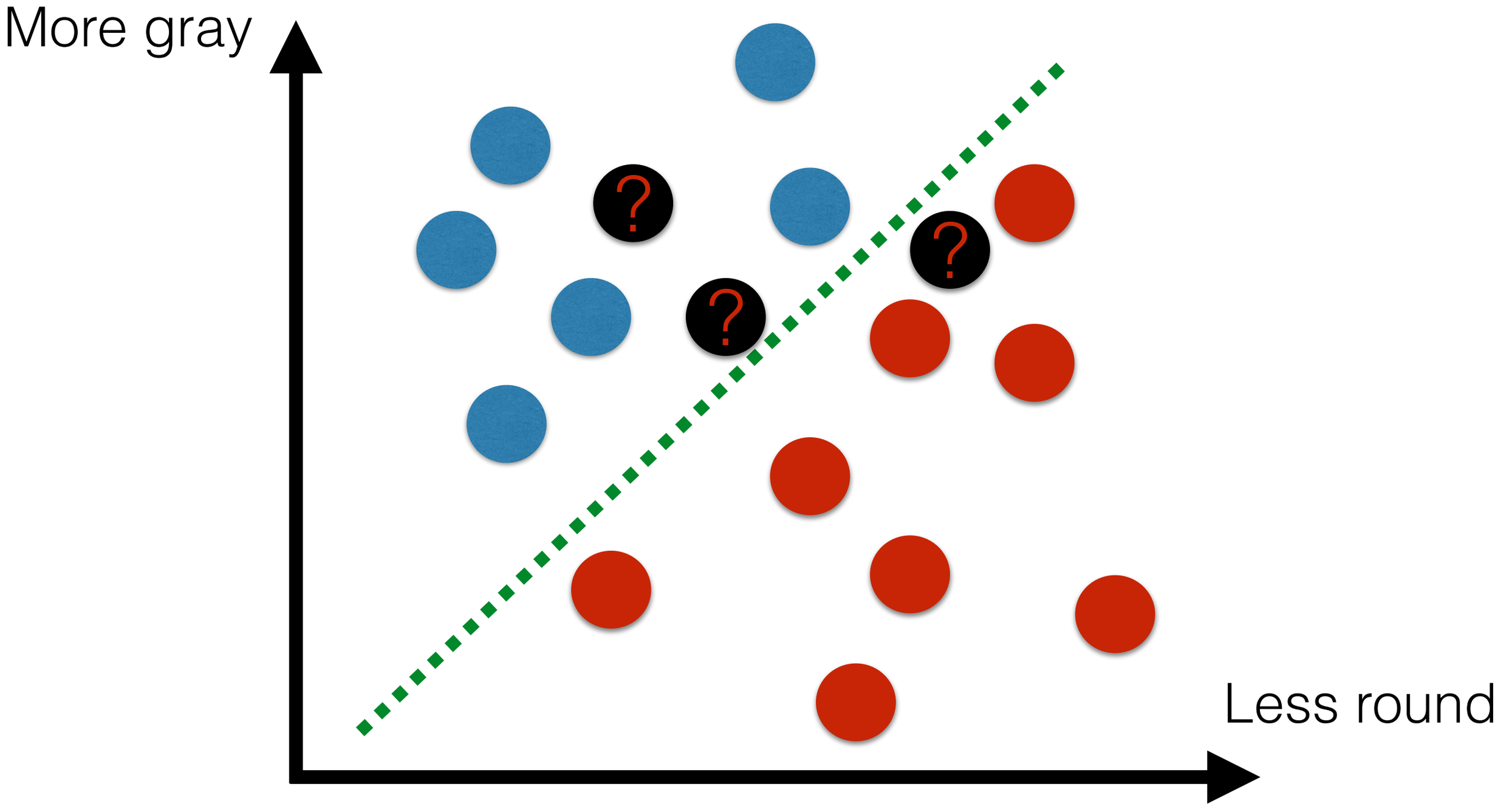
Testing loss: how far away the predicted prices are from the sale prices

Example 2:

and : training data

: testing data

Testing loss: measures how many we got wrong

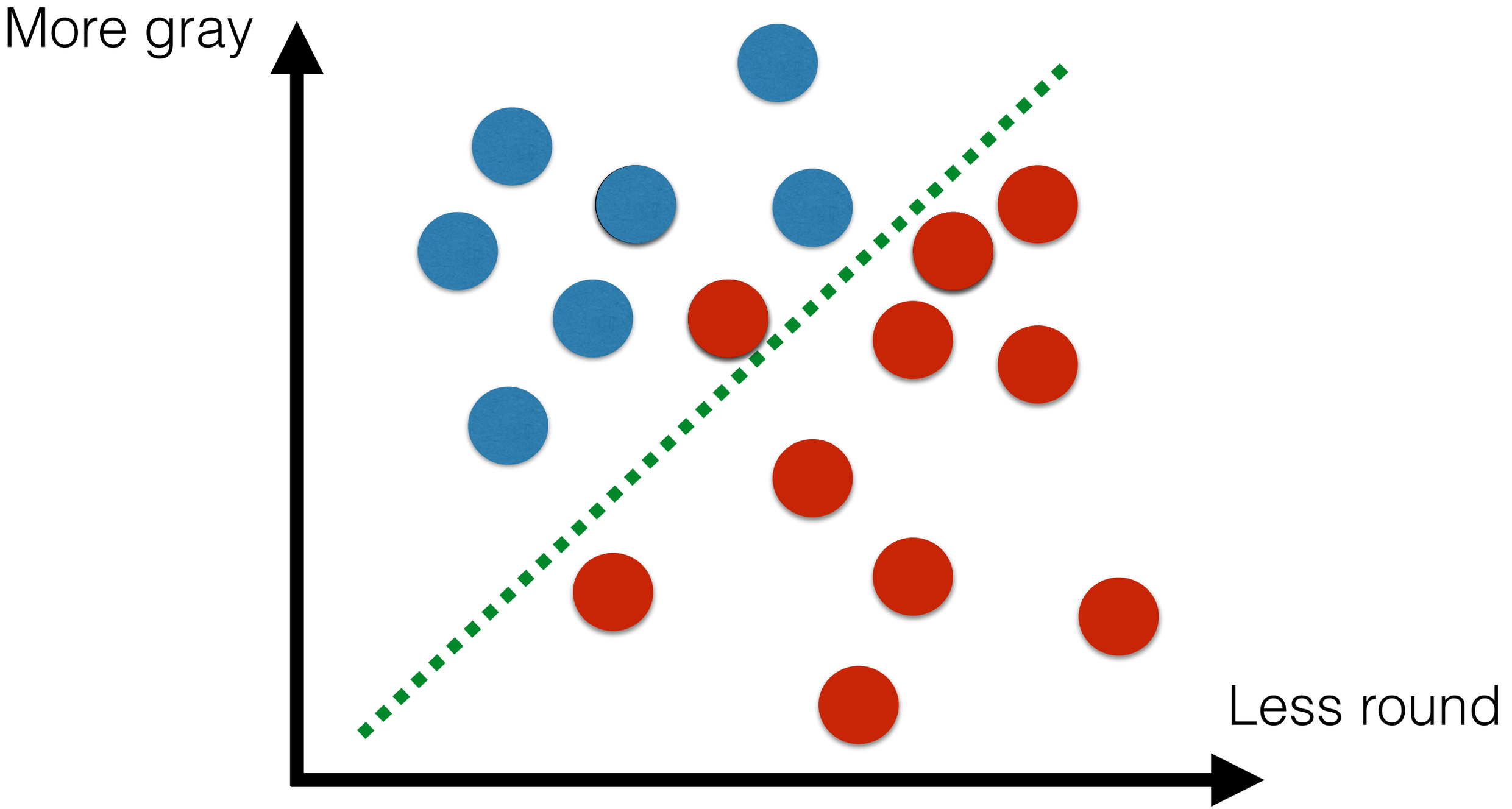


Example 2:

and : training data

: testing data

Testing loss: measures how many we got wrong



https://playground.tensorflow.org/

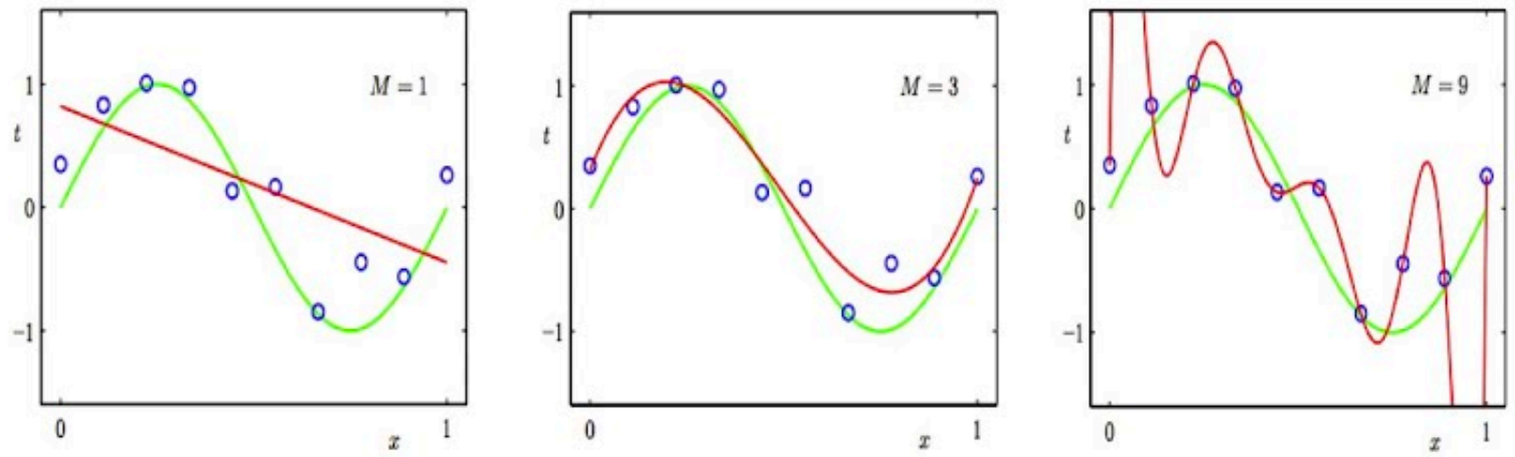
https://playground.tensorflow.org/

Let’s have a competition!

Who can get the lowest test loss on classifying spiral data points (the 4th dataset) ?

Under- and Over-fitting examples

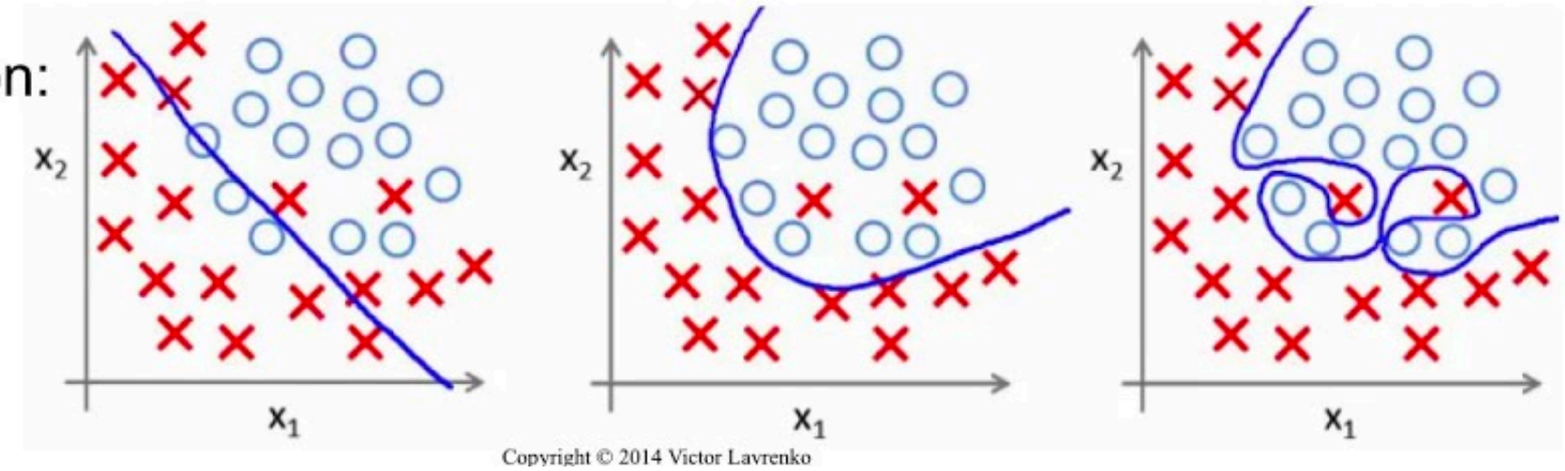
Regression:

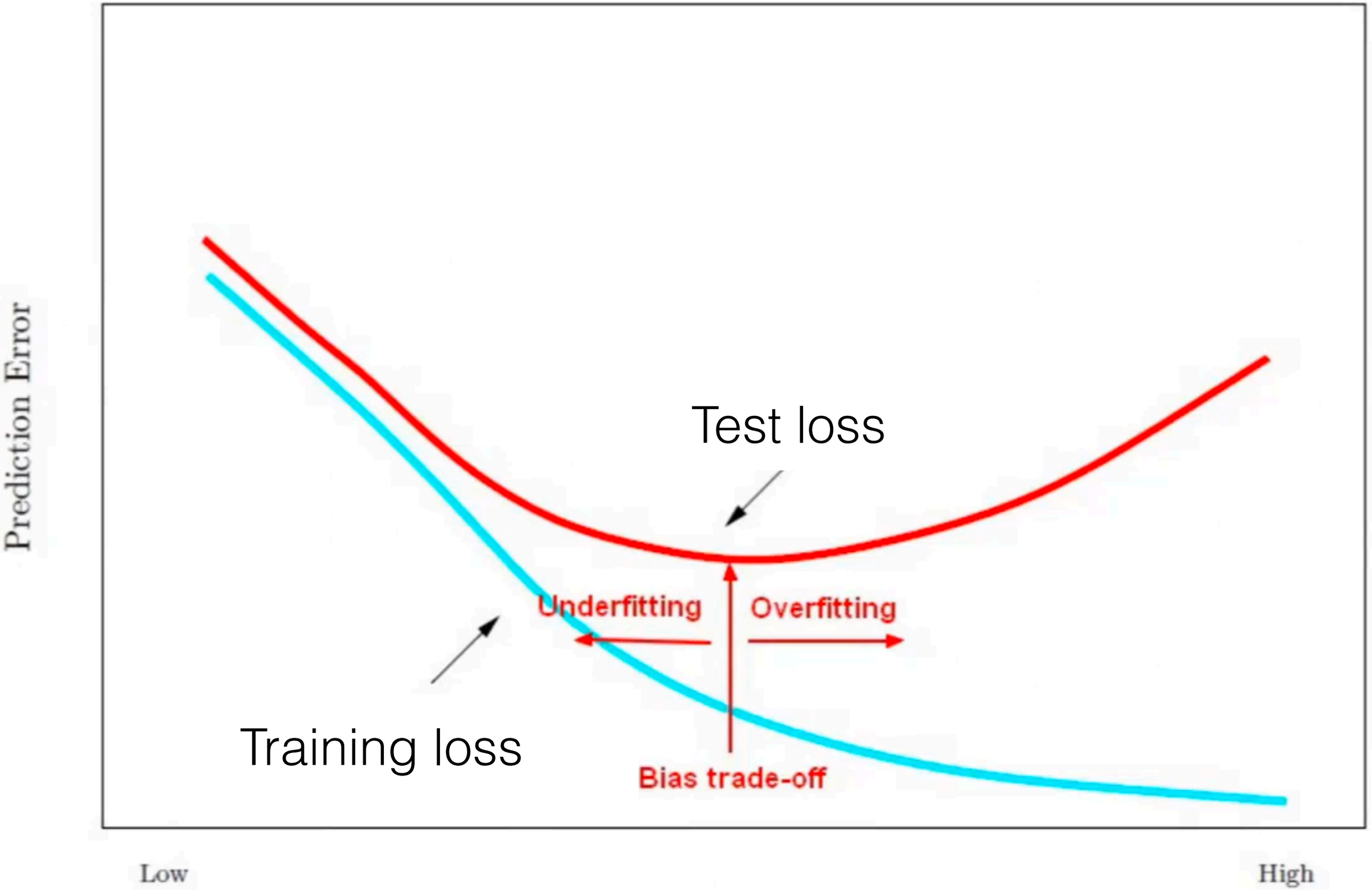


predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:




Prediction Error
Test loss
Underfitting
Overfitting
Training loss
Bias trade-off
Low
High
Model Complexity

Deep Neural Networks (DNN)

Neural Networks, Vol. 2, pp. 359–366, 1989
Printed in the USA.

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

Multilayer Feedforward Networks are Universal Approximators

KURT HORNIK

Technische Universität Wien

MAXWELL STINCHCOMBE AND HALBERT WHITE

University of California, San Diego

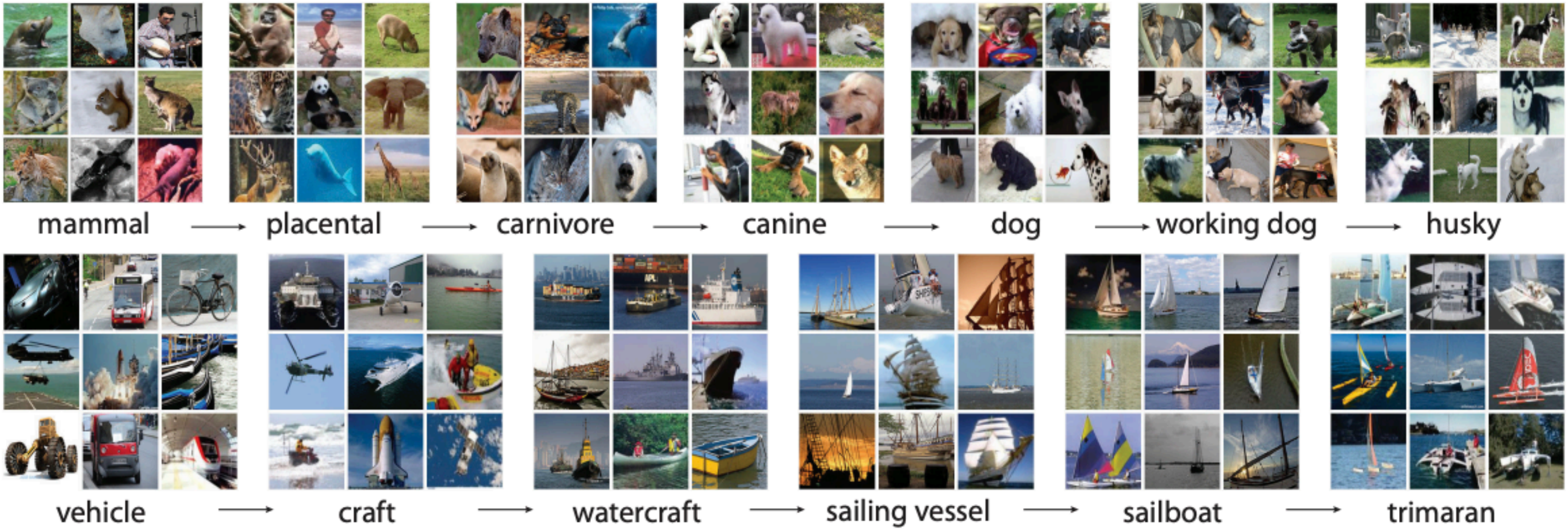
(*Received* 16 *September* 1988; *revised and accepted* 9 *March* 1989)

Abstract—*This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.*

Keywords—Feedforward networks, Universal approximation, Mapping networks, Network representation capability, Stone-Weierstrass Theorem, Squashing functions, Sigma-Pi networks, Back-propagation networks.

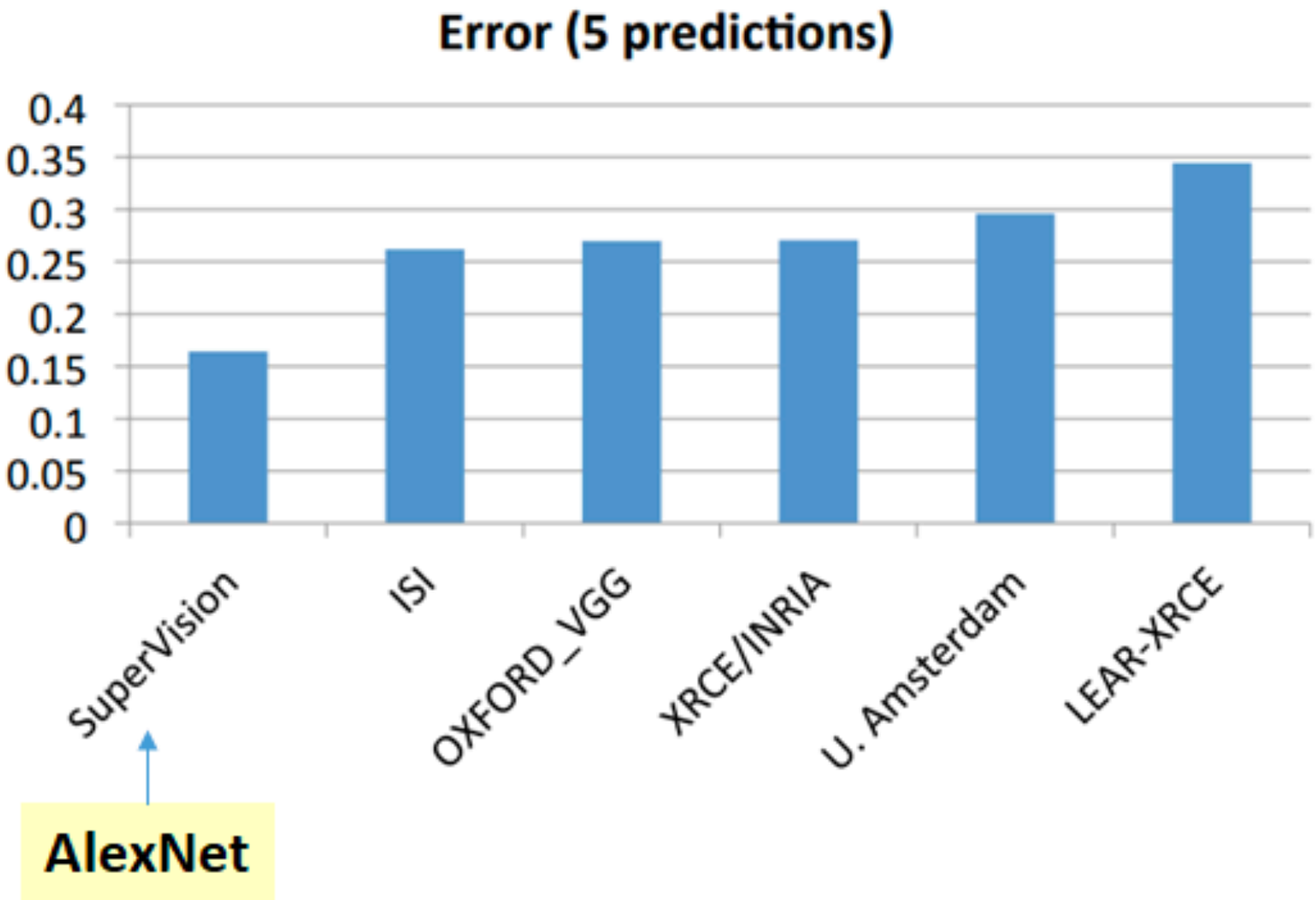
Many researchers did not believe that neural networks are useful, until the ImageNet Challenge in 2012.

More than 14M images divided into 20 000 categories.
Task: perform classification!

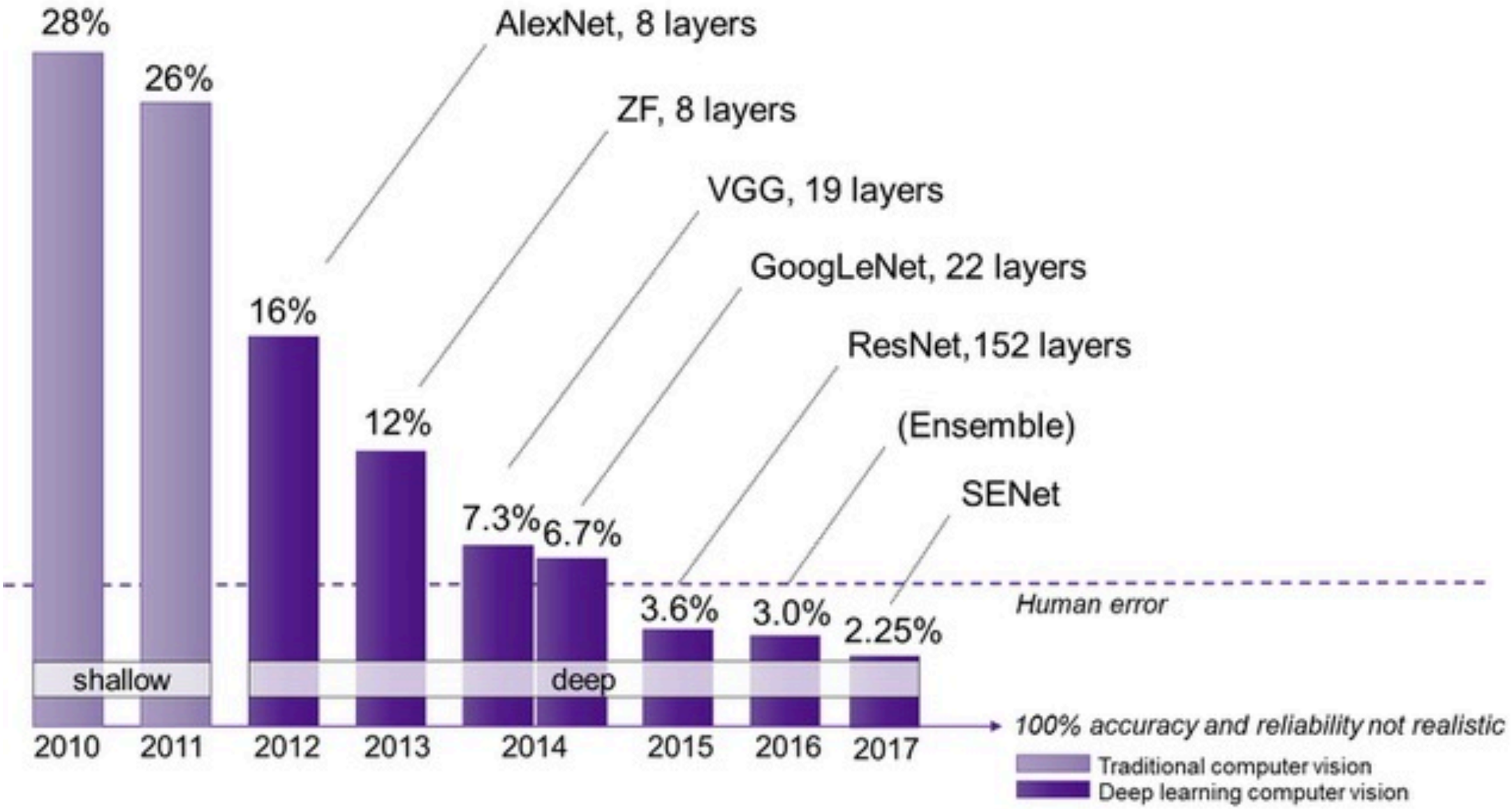


ImageNet Challenge (2012)

Ranking of the best results from each team



ImageNet Challenge

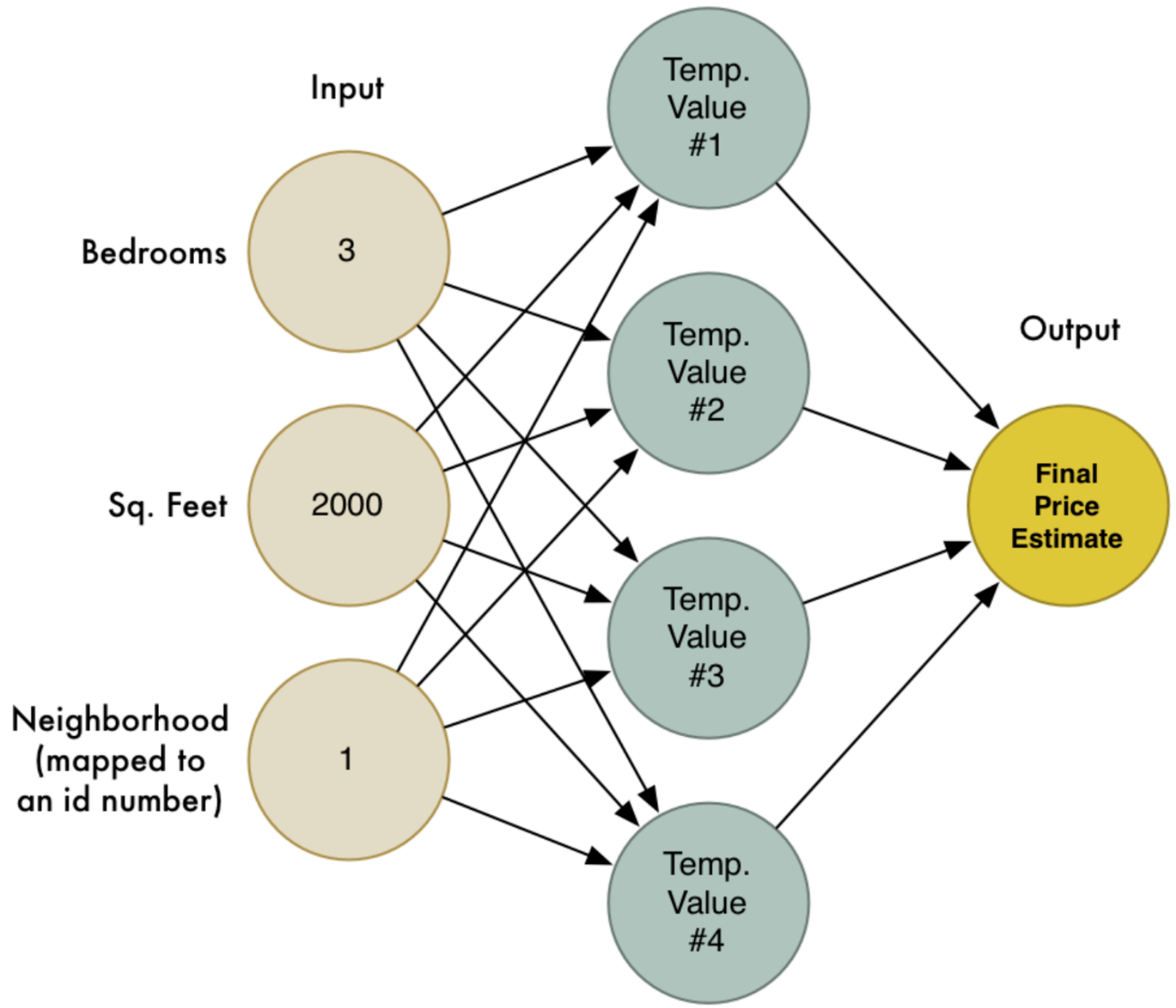


What are the drivers of the improvements?

1.Better hardware
2. Better training algorithms

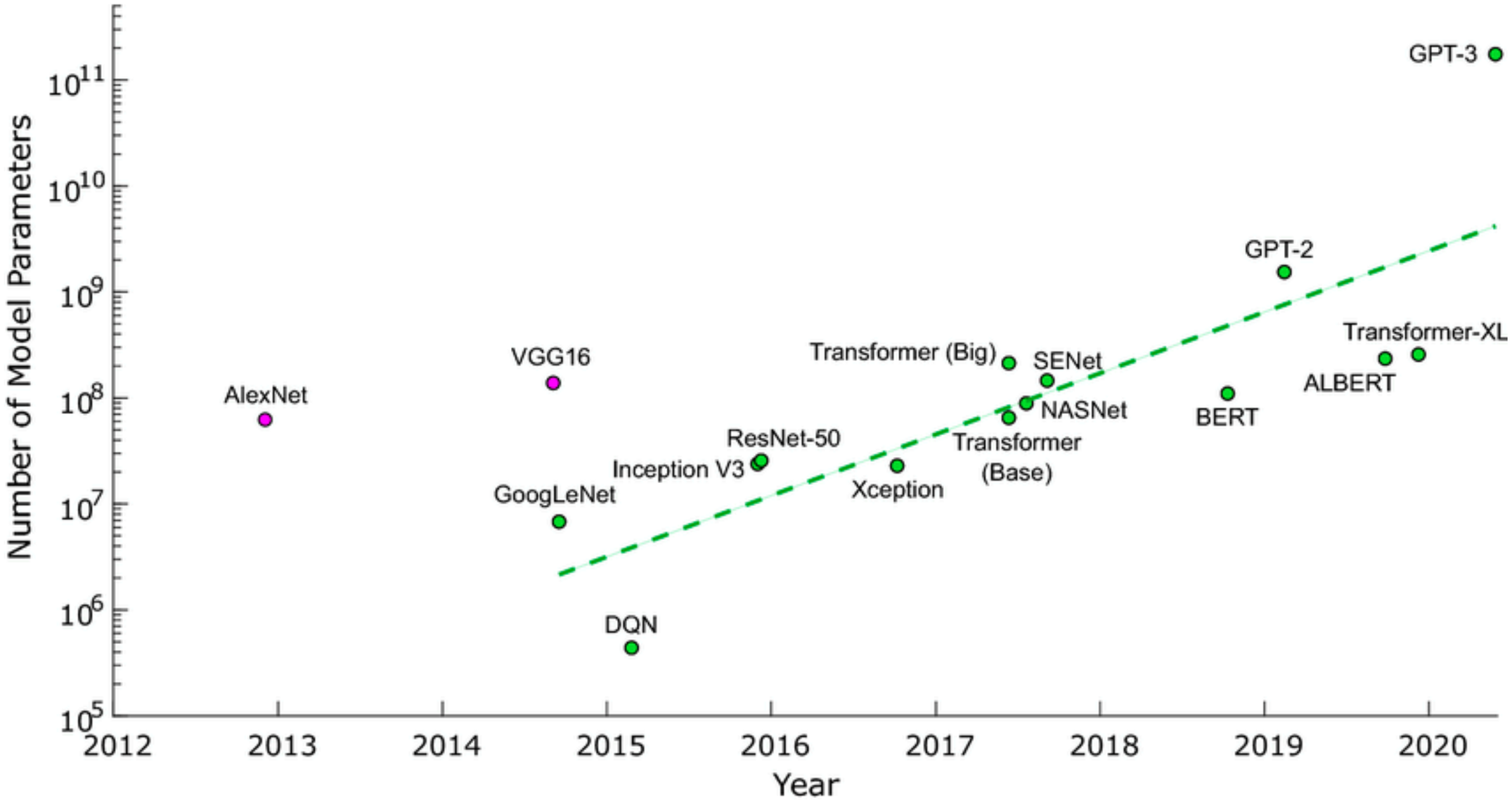
Together, they allow us to train larger and larger networks, using more parameters, which gives better performance.

The number of weights increases

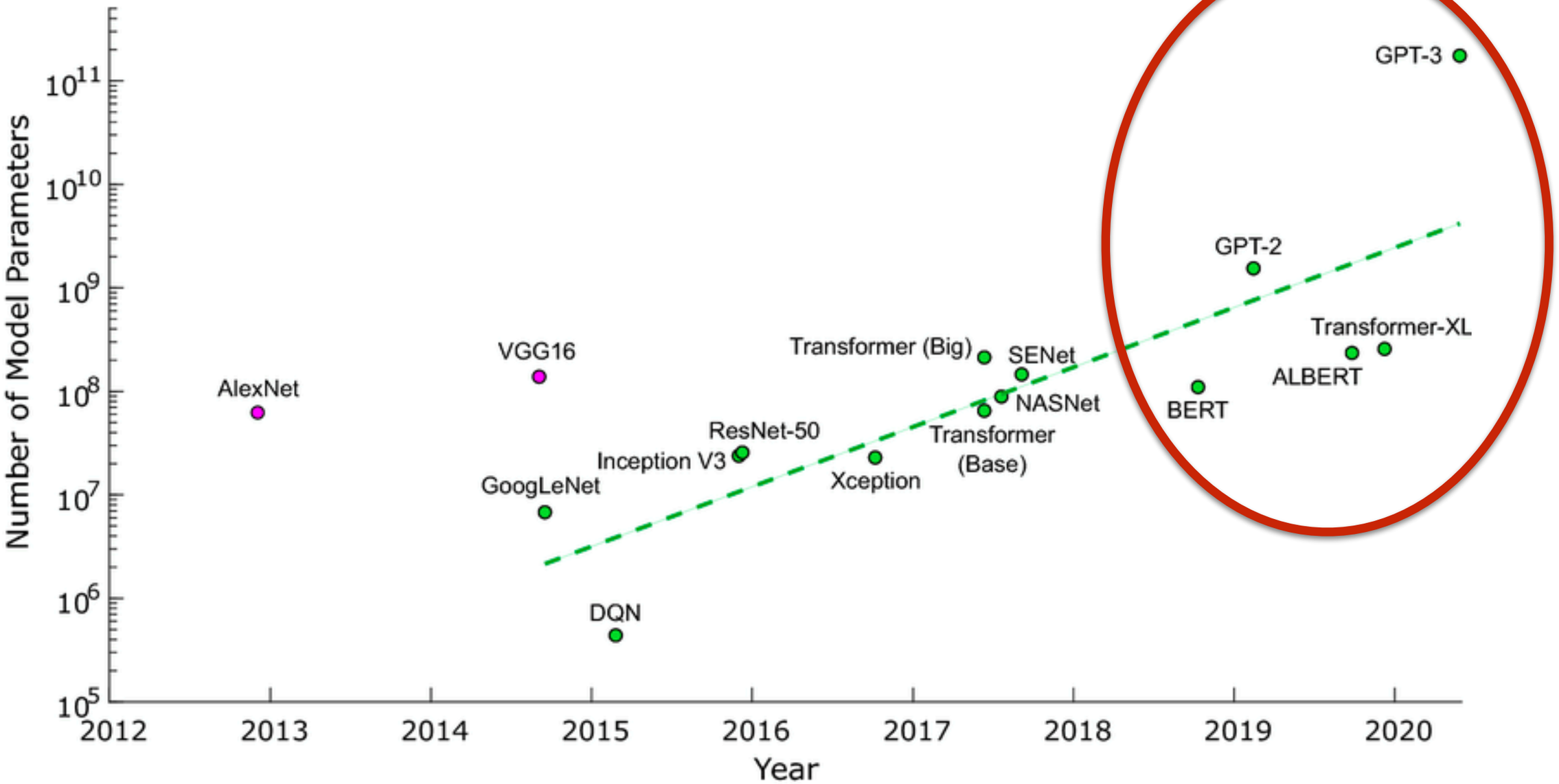


16 weights (since there are 16 arrows)!

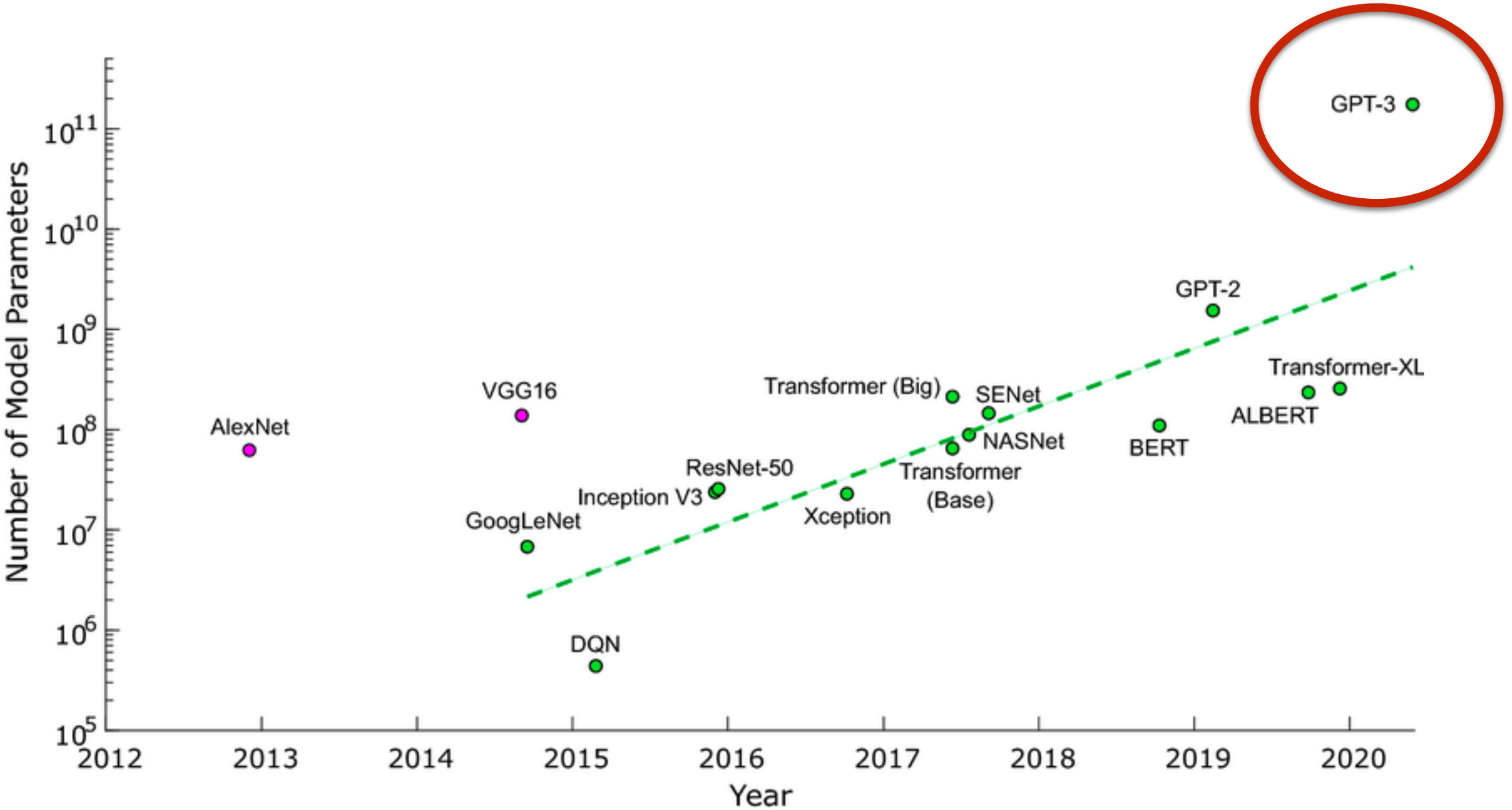
The number of weights increases



The number of weights increases



The number of weights increases



MIT News
ON CAMPUS AND AROUND THE WORLD

Shrinking deep learning's carbon footprint

Through innovation in software and hardware, researchers move to reduce the financial and environmental costs of modern artificial intelligence.

Kim Martineau | MIT Quest for Intelligence
August 7, 2020

The training of GPT-3 cost $4.6 million and 355 years in computing time. Training smaller models than GPT-3 releases around 626,000 pounds of CO2.