Gene2vec : distributed representation of genes based on co-expression

Vasudev R
Empid - 10073
Software developer trainee in machine learning
Machine learning department

Agenda for discussion

Machine Learning Algorithms working

- Machine learning classification
- A simple example
- Modelling process in SL
- Real world example
- Overall Architecture of SL

Deep Learning Algorithms working

- What is ANN
- What’s Artificial Neuron
- How relationship is captured
- What’s weight and bias ?

Natural Language Processing

- Types of input in NLP tasks ?
- Whats RNN ?
- Whats word Embeddings in NLP ?
- What’s with word2vec

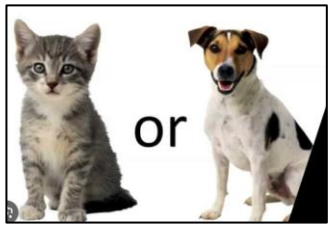
Gene2vec Paper explainaton

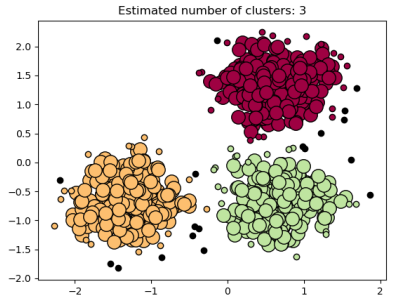
Machine Learning algorithms working

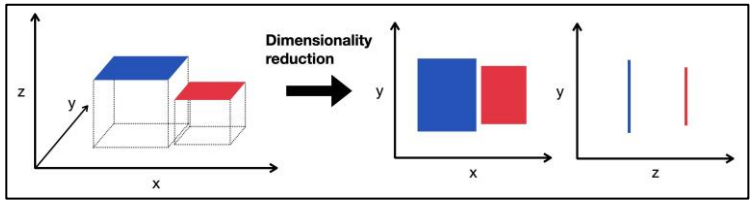
MACHINE LEARNING ALGOS CLASSIFICATION

Machine learning algorithms
Classifications

- Supervised learning
  - Regression problems
  - Classification problems
- Unsupervised learning
  - Clusturing
  - Dimensionality reduction
- Reinforcement learning







- Can you find the values of y for 4 and 5

Input (X)	Output (Y)
1	2
2	5
3	10
4	?
5	?

- Modelled the equation is y = f(x)= x^2 +1
- And substituted x with 4 and 5

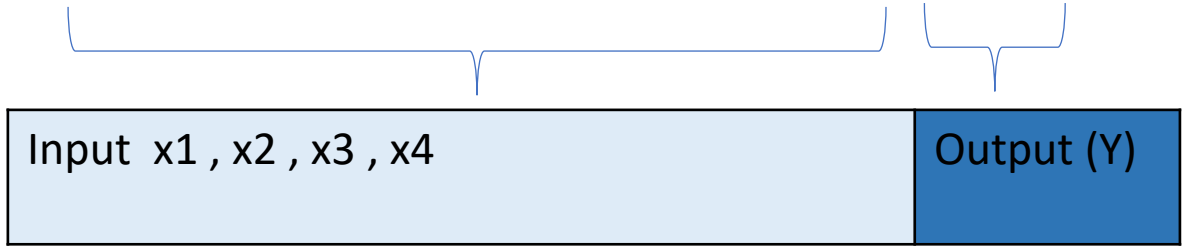
Input (X)	Output (Y)
1	2
2	5
3	10
4	17
5	26

Training data

Test data

- Example Boston House price prediction dataset

	A	B	C	D	E
1	area	bedrooms	balcony	age	price
2	1200	2	0	2	500000
3	2300	3	2	5	620000
4	2500	4	2	1	122500
5	3650	5	3	3	6000000
6	1800	3	1	5	2122000
7	3000	3	1	4	120000
8	1222	1	0	2	450000
9	4600	5	3	1	6500000
10	2050	2	2	2	1530000
11	1450	2	2	3	1563330



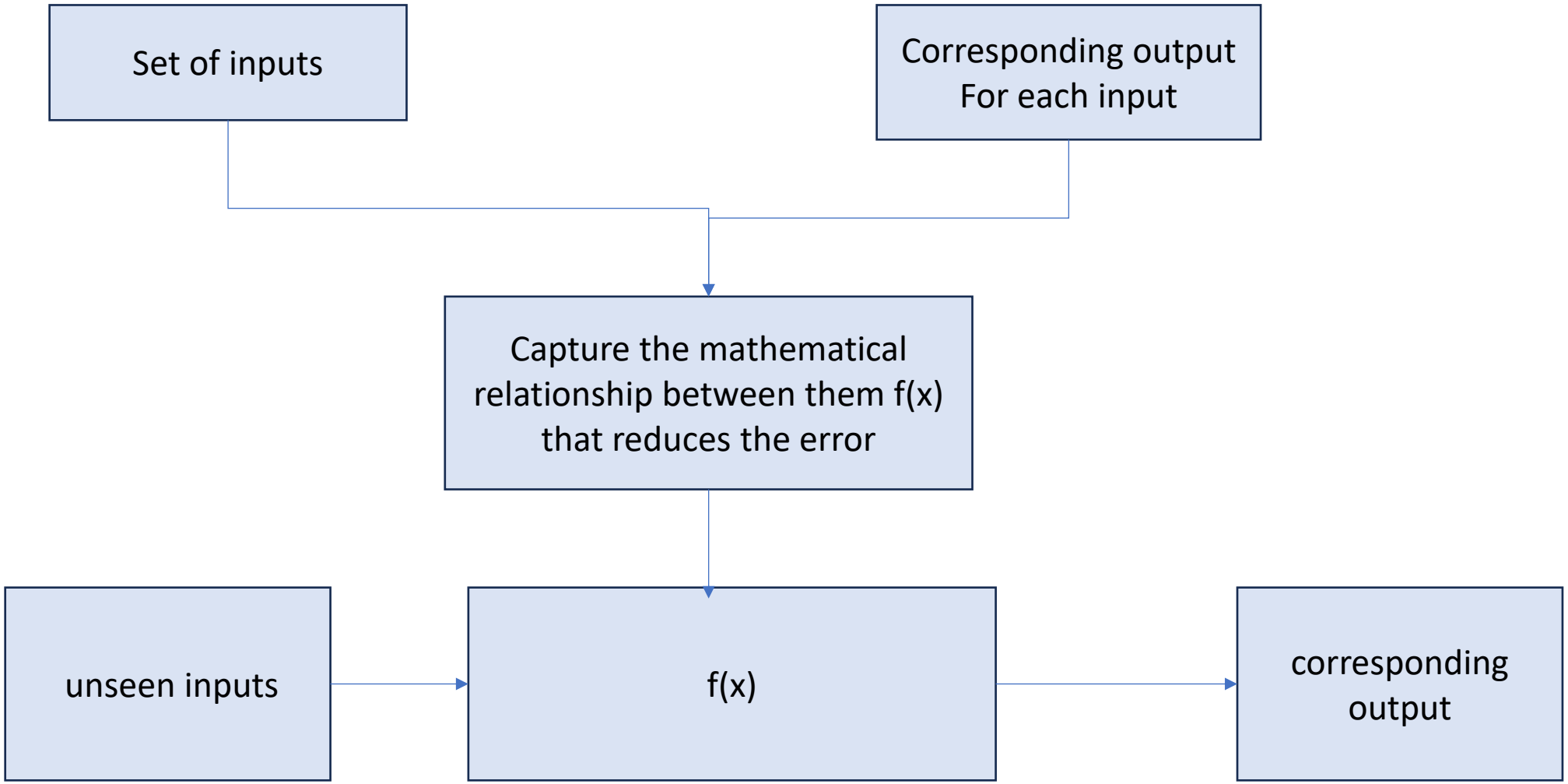
- In real life the output and input wont have a perfect relationship

Input (X)	Output (Y)
1	~~2~~ 3
2	~~5~~ 4
3	~~10~~ 11
4	
5	

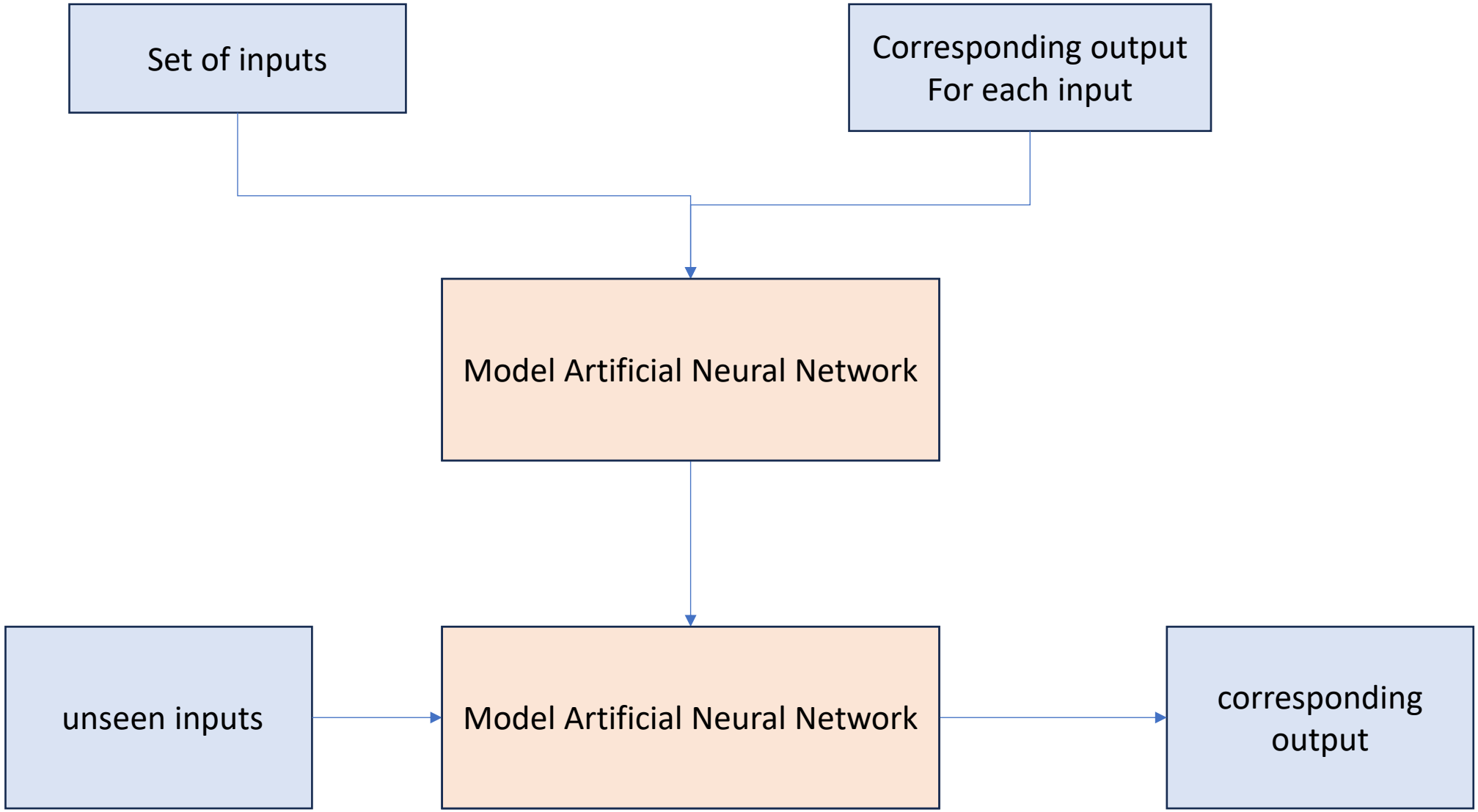
It approximates to a function that minimizes the error as small as possible

Here y = f(x)= x^2 +1

Overall Architecture of Supervised Learning Algorithms



Deep Learning algorithms working

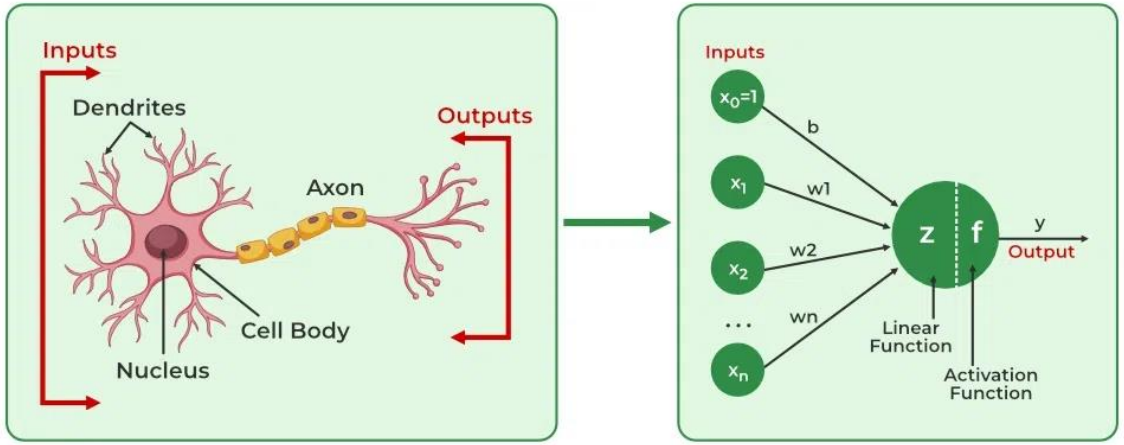

Set of inputs
Corresponding output
For each input
Model Artificial Neural Network
unseen inputs
Model Artificial Neural Network
corresponding
output

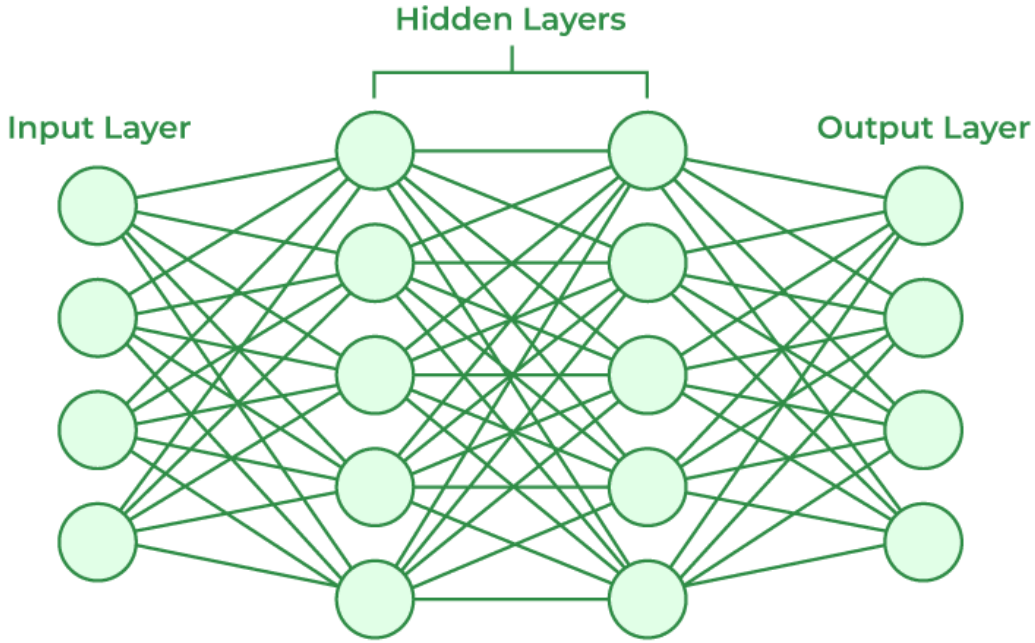
What’s inside an ANN ?

- Inside ANN there are many number of neurons
- Neurons are inspired by biological neural network

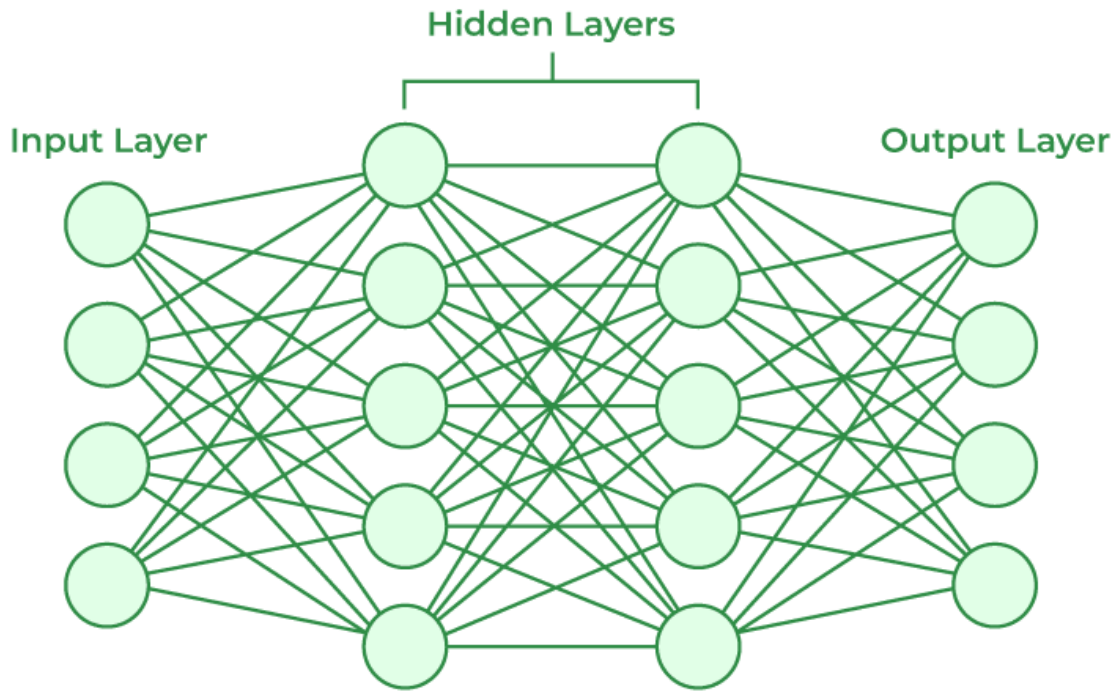
Neurons

- Input layer and an activation layer
- have weights and bias

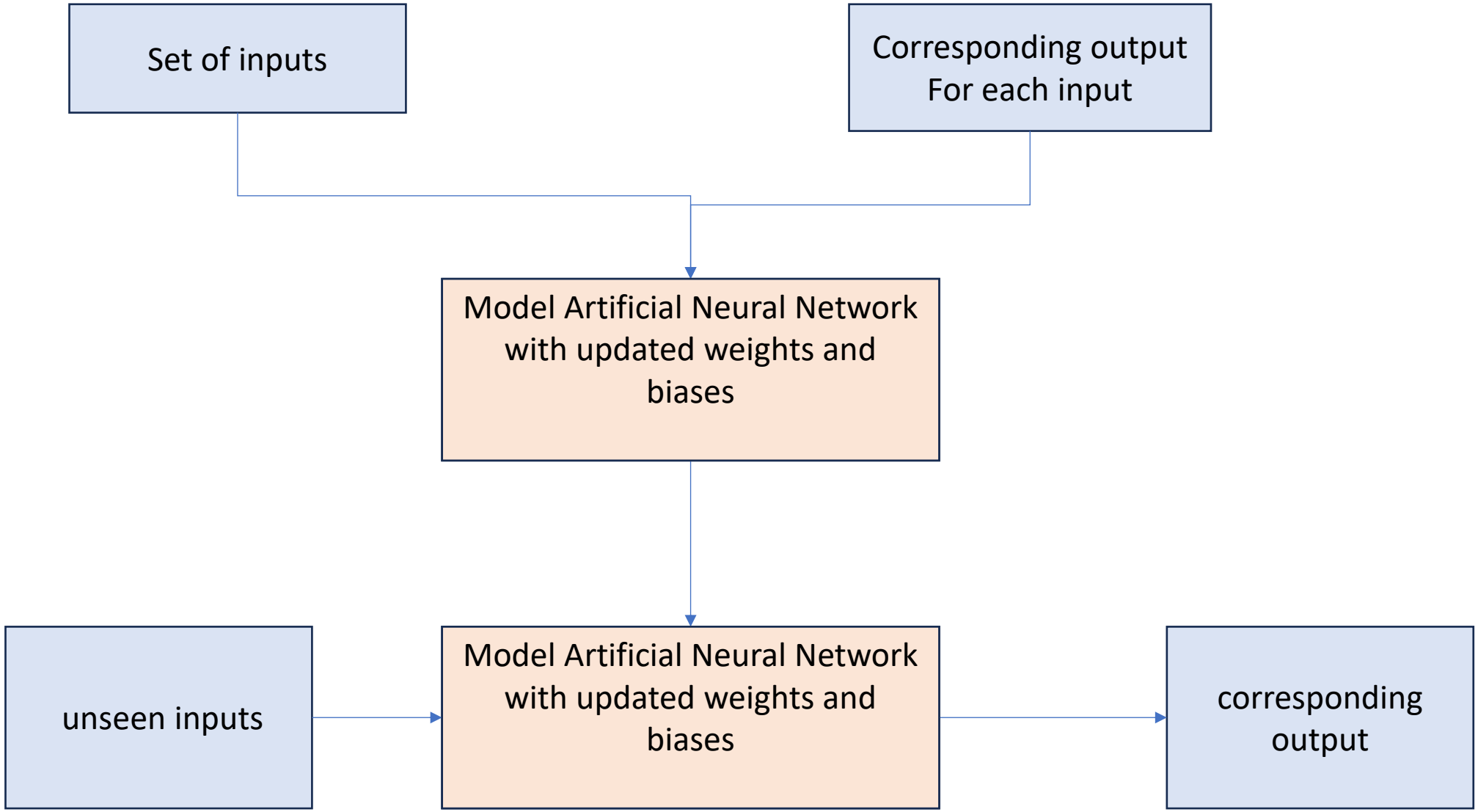




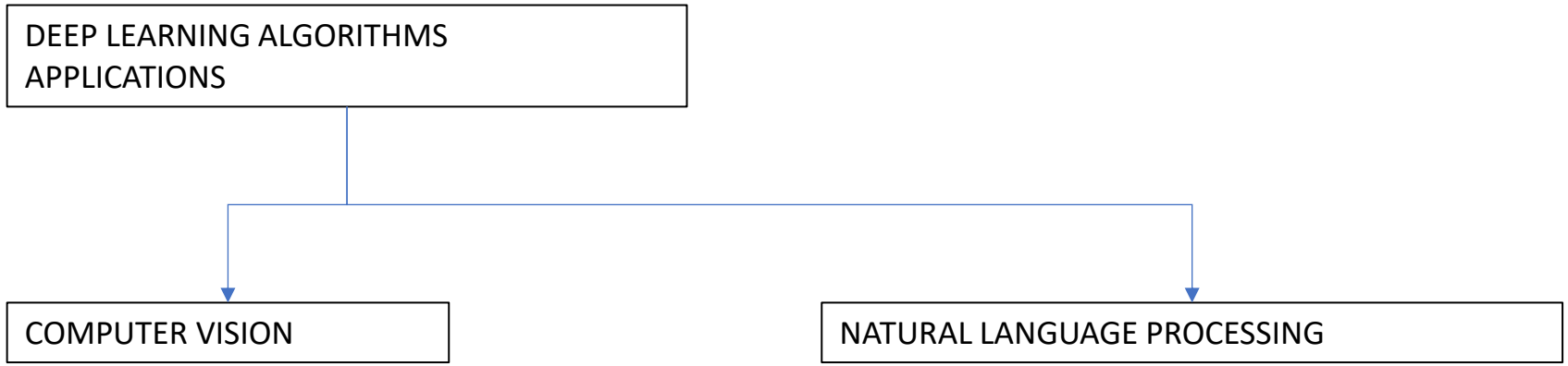
How the mathematical relationship is captured ?



- Instead of calculating a mathematical function
- Based on inputs and their corresponding output labels
- Weights and bias values are updated from a small initial value through a process called **Backpropagation**


Set of inputs
Corresponding output
For each input
Model Artificial Neural Network with updated weights and biases
unseen inputs
Model Artificial Neural Network with updated weights and biases
corresponding output

Deep learning algorithms applications



Aims to replicate how humans see and understand the world around them.

- Facial recognition.
- Self-driving cars.
- Robotic automation.
- Medical anomaly detection.
- Sports performance analysis.
- Manufacturing fault detection.
- Agricultural monitoring.
- Plant species classification.

Technology that allows computers to understand, process, and manipulate human language.

- Interpret the semantic meaning of language
- Translate between human languages
- Recognize patterns in human languages
- Read text
- Hear speech
- Measure sentiment
- Determine which parts of a text are important

NATURAL LANGUAGE PROCESSING

Difference in Input type in NLP vs ML tasks

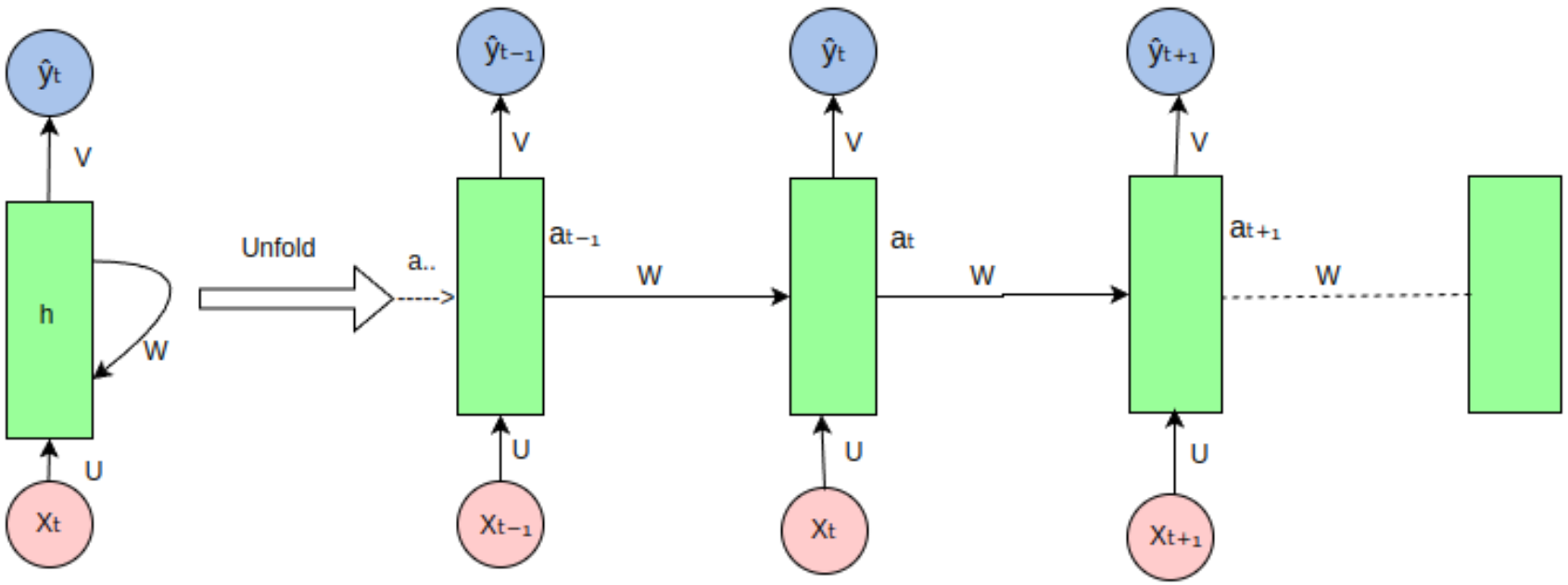
	A	B	C	D	E
1	area	bedrooms	balcony	age	price
2	1200	2	0	2	500000
3	2300	3	2	5	620000
4	2500	4	2	1	122500
5	3650	5	3	3	6000000
6	1800	3	1	5	2122000
7	3000	3	1	4	120000
8	1222	1	0	2	450000
9	4600	5	3	1	6500000
10	2050	2	2	2	1530000
11	1450	2	2	3	1563330

	news_headline	news_category
0	saudi arabia open air space land maritime bord...	world
1	indian railway use facial recognition end desp...	tech
2	u pursuing seditious conspiracy case unprecede...	world
3	coronavirus outbreak hertha berlin striker sal...	health
4	google former ceo urge u govt invest artificia...	tech
5	amid record u case new white house outbreak tr...	world
6	india inoculates cr report covid upsurge five ...	india
7	u rescinds trump administration claim u n sanc...	world
8	govt working relief plan sc order vodafone ide...	tech
9	india first indigenous pneumonia vaccine devel...	health

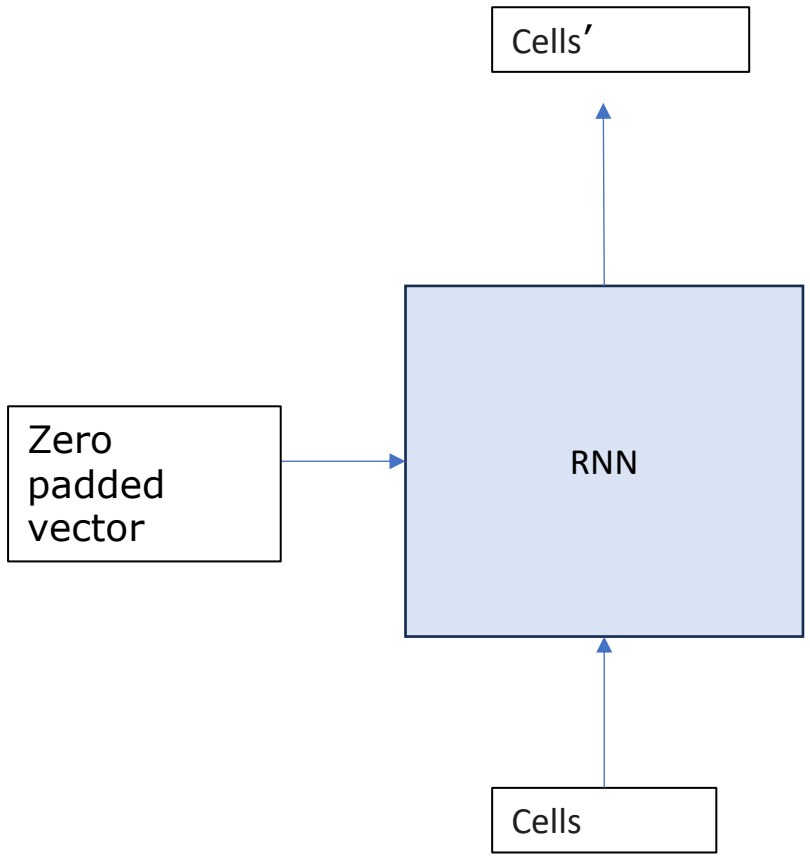
- Input sizes (of sentences) are varying
- Sequential Information is also present

To Resolve this We use RNN (Recurrent Neural Network)

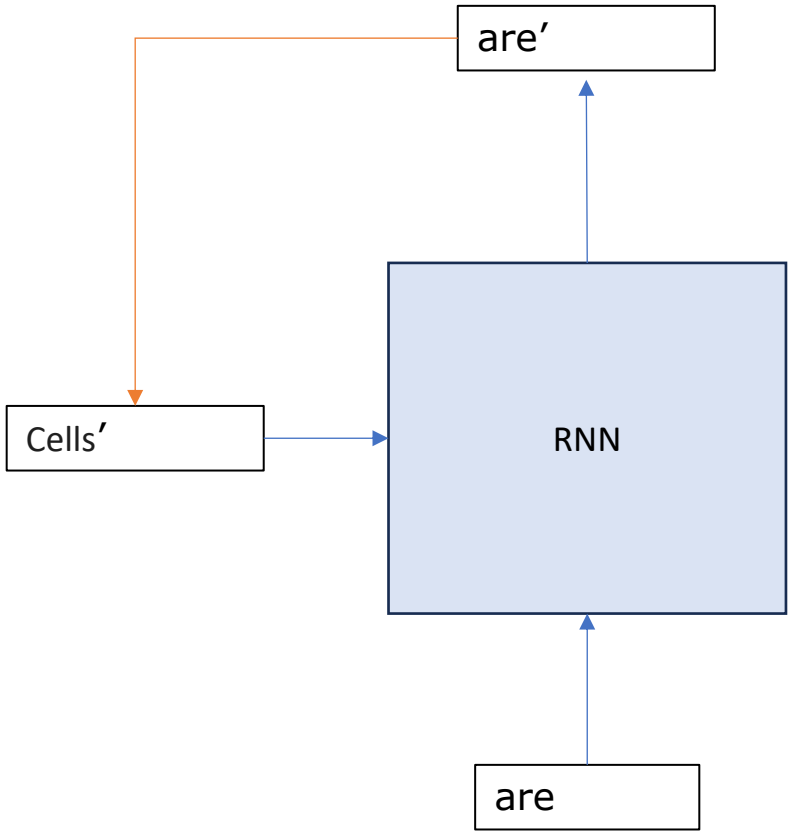
Recurrent means repeating nature



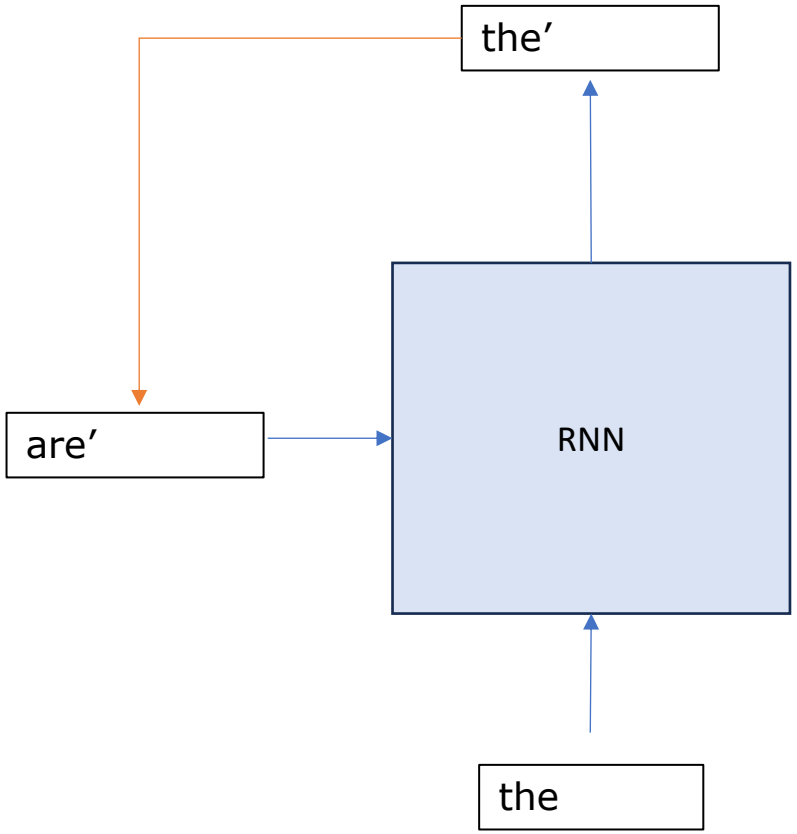
E.g. input : **Cells** are the basic building blocks.



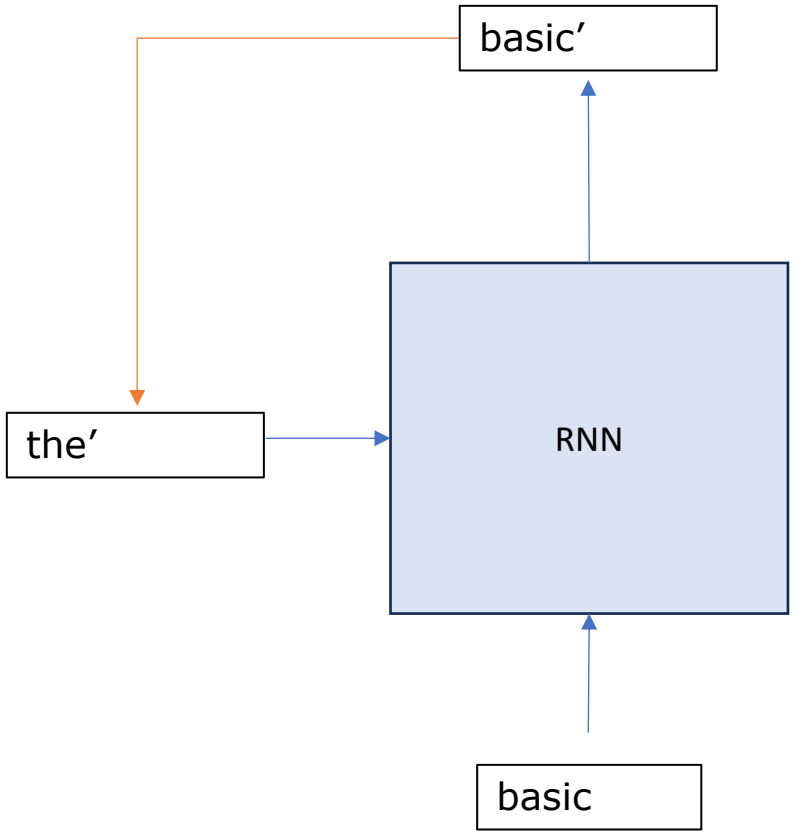
E.g. input : Cells **are** the basic building blocks.



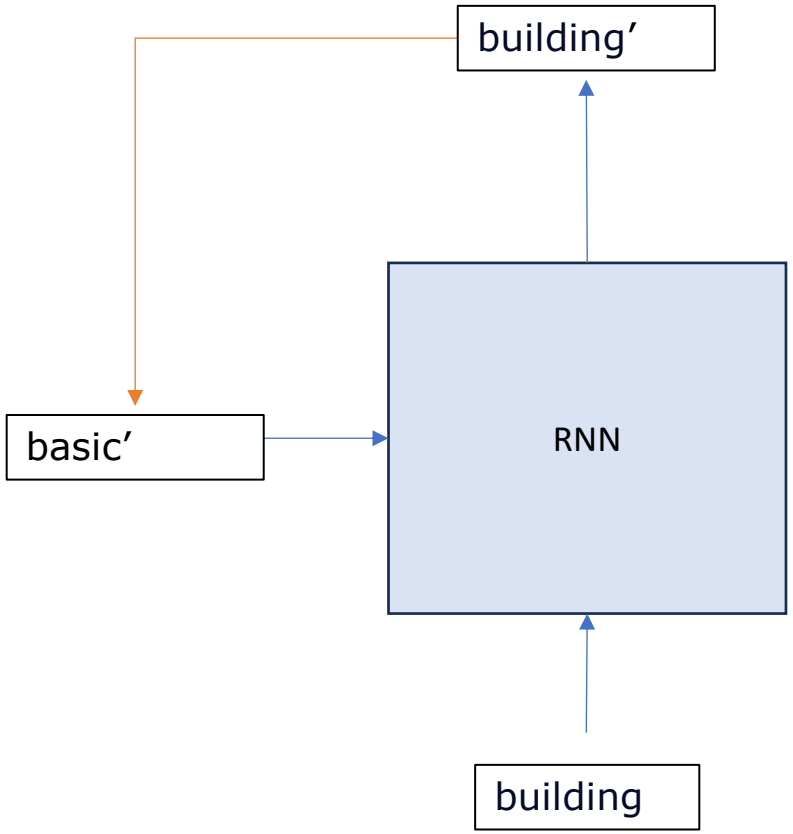
E.g. input : Cells are **the** basic building blocks.



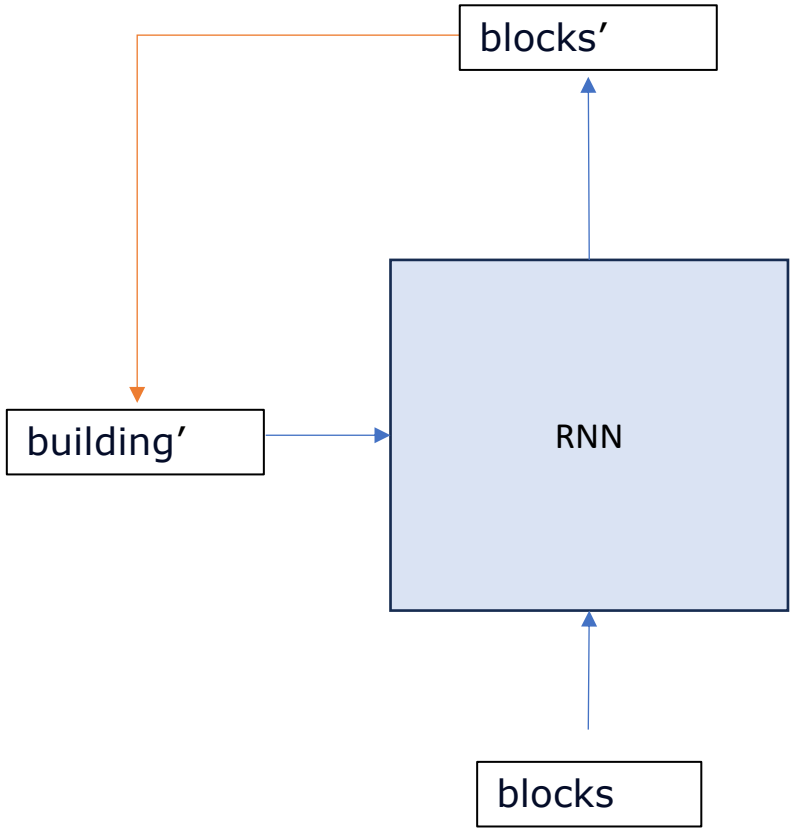
E.g. input : Cells are the **basic** building blocks.



E.g. input : Cells are the basic **building** blocks.

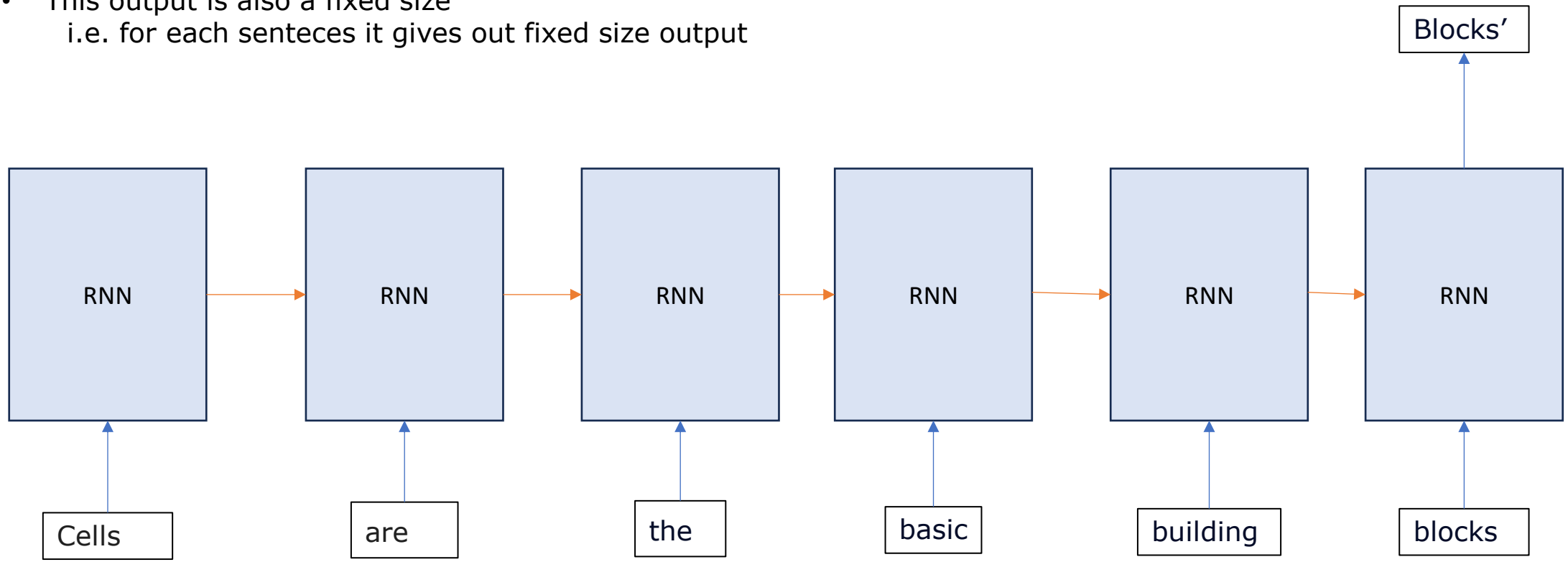


E.g. input : Cells are the basic building **blocks**.



E.g. input : Cells are the basic building blocks.

- This final output carrys the sequential information of each sentence.
- This output is also a fixed size
  i.e. for each senteces it gives out fixed size output



Overall architecture of RNN + ANN

Cells are the basic building blocks of all living things. The human body is composed of trillions of cells. They provide structure for the body, take in nutrients from food, convert those nutrients into energy, and carry out specialized functions. Cells also contain the body's hereditary material and can make copies of themselves.

Cells are the basic building blocks of all living things.

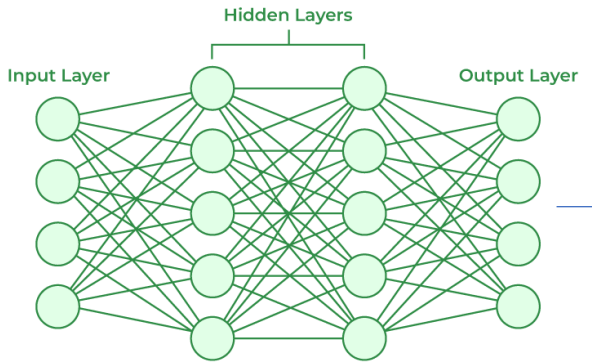
The human body is composed of trillions of cells.

They provide structure for the body, take in nutrients from food, convert those nutrients into energy, and carry out specialized functions.

Cells also contain the body's hereditary material and can make copies of themselves.

RNN

Forward propogation



backward propogation

Test output For each sentence

Data Preprocessing
The Encoding process
Word Embeddings
Word2vec
Gene2vec

Why Data Preprocessing is needed ?

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	F
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	N
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	N
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	N
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	N
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	N
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	...	0	N
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	...	0	N

Some Data Preprocessing techniques

- Scaling
- Normalization
- Filling N/A values
- Outlier treatment
- Data Encoding
- Attribute selection
- Dimensionality Reduction

Data Encoding

- Common techniques include
  - One hot encoding
  - Ordinal encoding
  - Label Encoding
  - Target Encoding
  - Hash Encoding
  - Categorical Encoding