National Research Tomsk State University

Natural Language Processing. Neural Networks and Text Processing. Part 2

(A Very Short Introduction)

Sergey V. Axyonov,

PhD, Associate Professor,

Department of Fundamental Computer Science, Institute of Applied Math & Computer Science,

Tomsk State University

Tomsk-2024

Summary from the Previous Presentation

1. Tokenization is the transformation of a raw text into meaningful lexical tokens.

2. We can use stemming or lemmatization to get fewer unique tokens.

3. A computer asks numbers to conduct deep analysis of data. So all tokens (letters, words, or sub-words) need to be converted to numbers.

4. We can use One-hot encoding procedure to code each token.

5. Bag-of-Words is a very simple procedure for representing any sentence.

Part of Speech Tagging. PyMorphy2 library

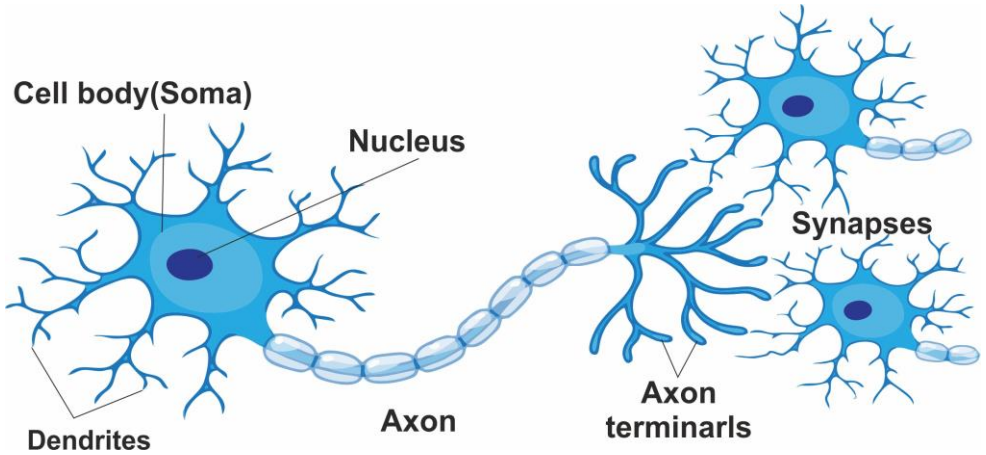
Example:

"Сибирские кошки являются местными лесными кошками России и они, как известно, долгое время обитали в густых лесах Сибири."

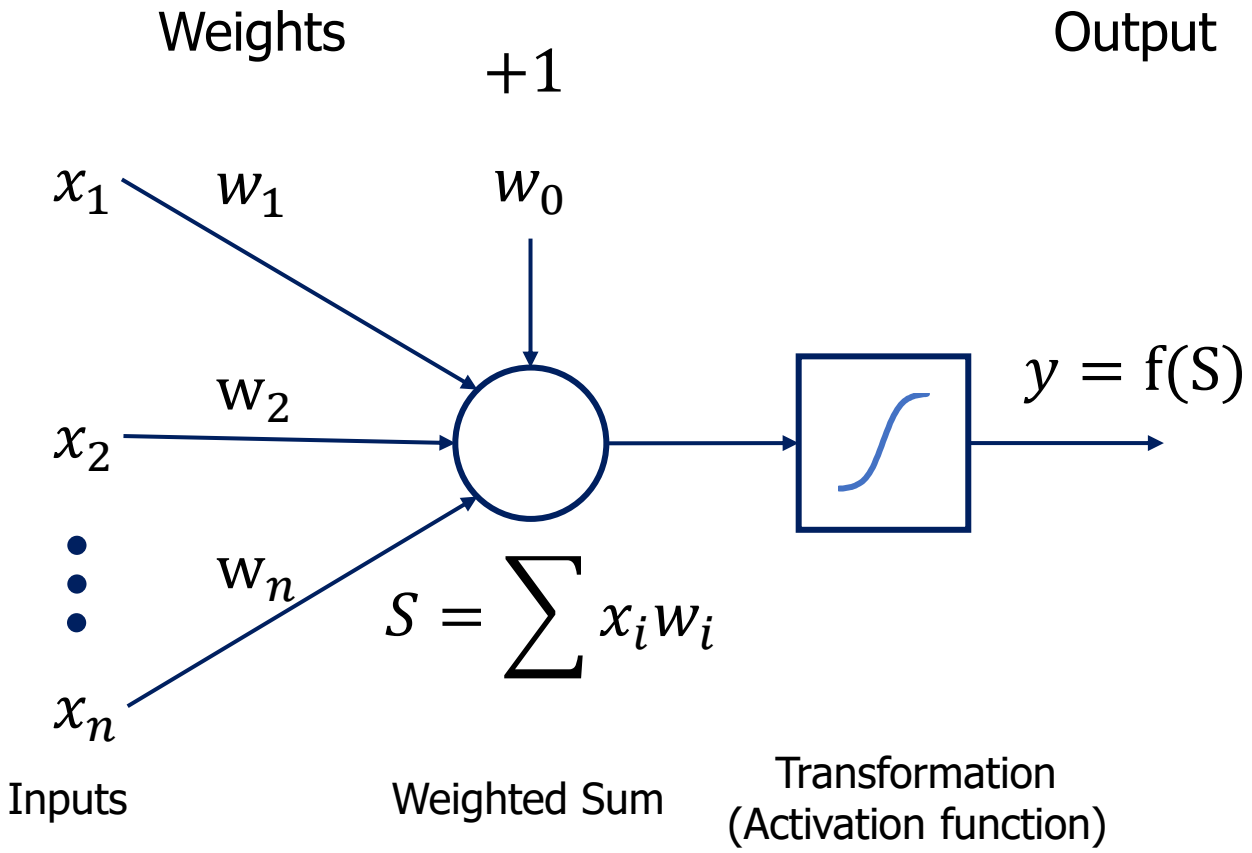
Tagging:

word='сибирские', tags=('ADJF plur, nomn'), normal_form='сибирский’
word='кошки', tags=('NOUN, anim, femn plur, nomn'), normal_form='кошка’
word='являются', tags=('VERB, impf, intr plur, 3per, pres, indc'), normal_form='являться'
word='местными', tags='ADJF, Qual plur, ablt'), normal_form='местный'
word='России', tags=('NOUN, inan, femn, Sgtm, Geox sing, gent'), normal_form='Россия'
word='и', tags=('CONJ'), normal_form='и'
word='известно', tags=('PRED, pres'), normal_form='известно'
word='в', tags=('PREP'), normal_form='в'

Biological & Artificial Neurons



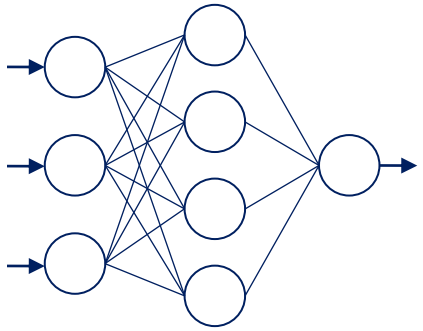
Biological Neuron



Artificial Neuron

Neural Networks & Text Processing Tasks 1

Image-to-Text



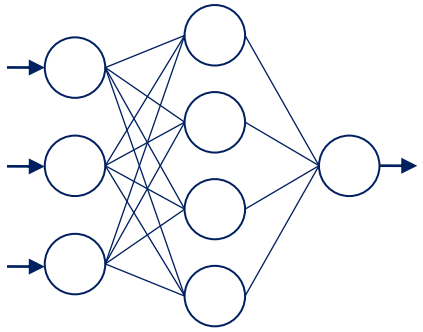
«A woodpecker sits on a fallen birch tree.»

Input

Output

Text Classification

"This was definitely the best bag I’ve bought."



«A positive review»

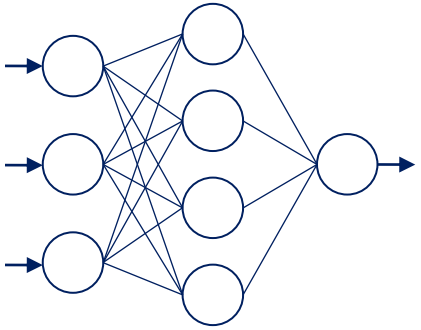
Input

Output

Neural Networks & Text Processing Tasks 2

Text-to-Image

Cat barista, making coffee.

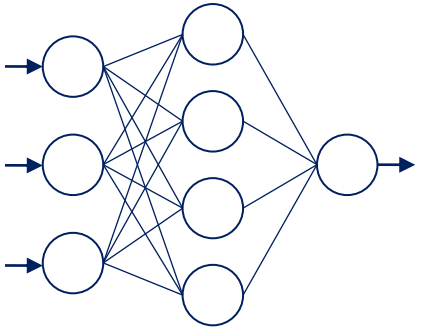


Input

Output

Text-to-3D

A blue poison-dart frog sitting on a water lily.



Input

Output

Source: https://www.midjourney.com/

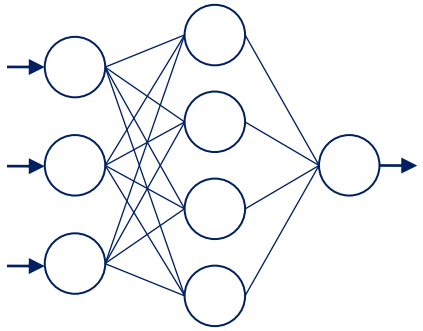
Source: https://research.nvidia.com/labs/dir/magic3d/

Neural Networks & Text Processing Tasks 3

Text Translation

Cats have excellent night vision and can see at one sixth the light level required for human vision.

Input



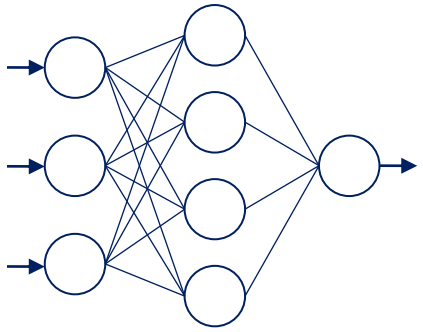
«Кошки обладают превосходным ночным зрением и могут видеть на уровне одной шестой уровня освещенности, необходимого для человеческого зрения.»

Output

Text Generation

Describe this fantastic game.

Input



«One of the best moments in the game comes after a lengthy puzzle section that requires you to transport between different time periods to alter the present.»

Output

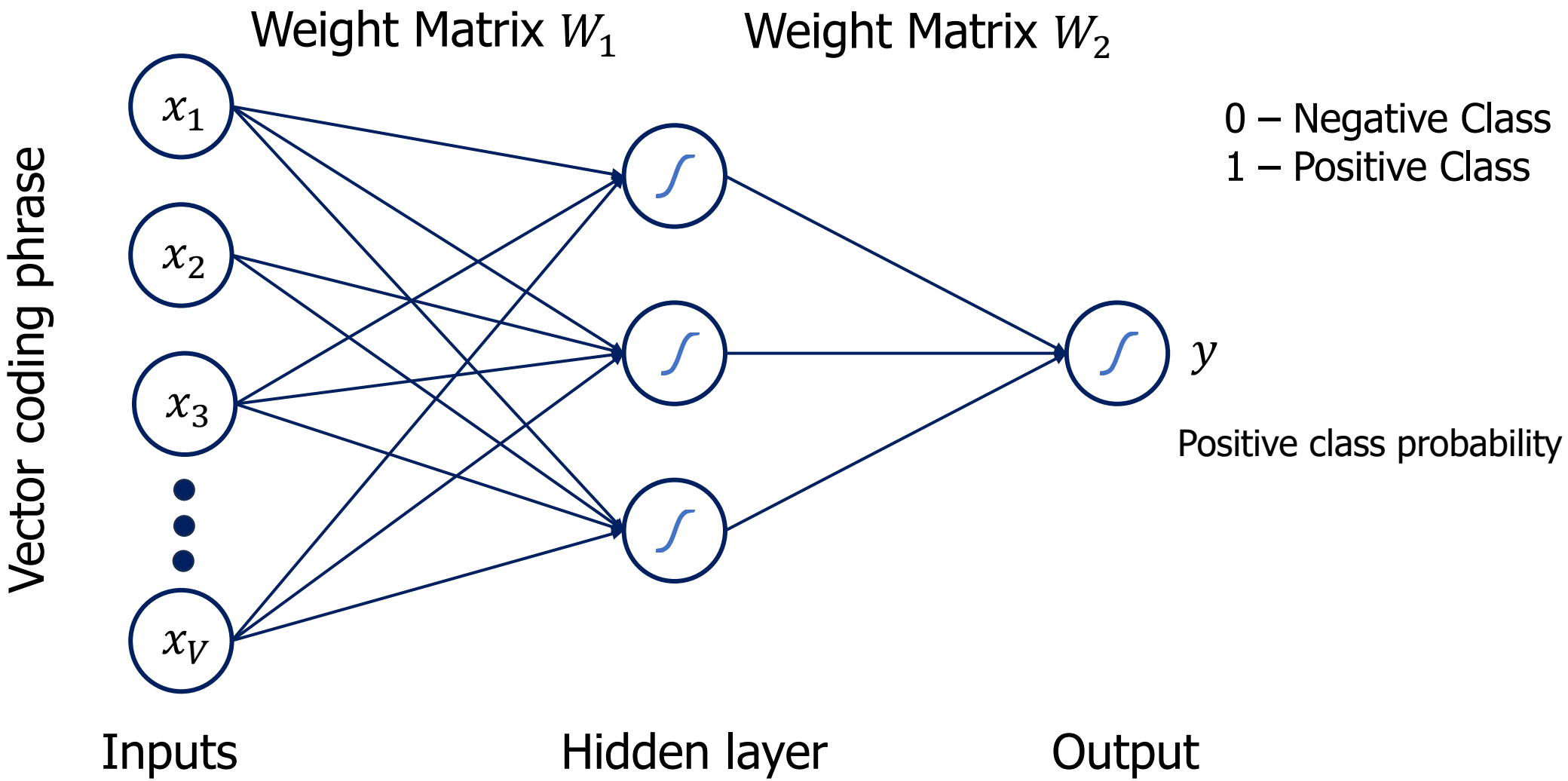
Source: https://openai.com/

Source: https://translate.google.com/

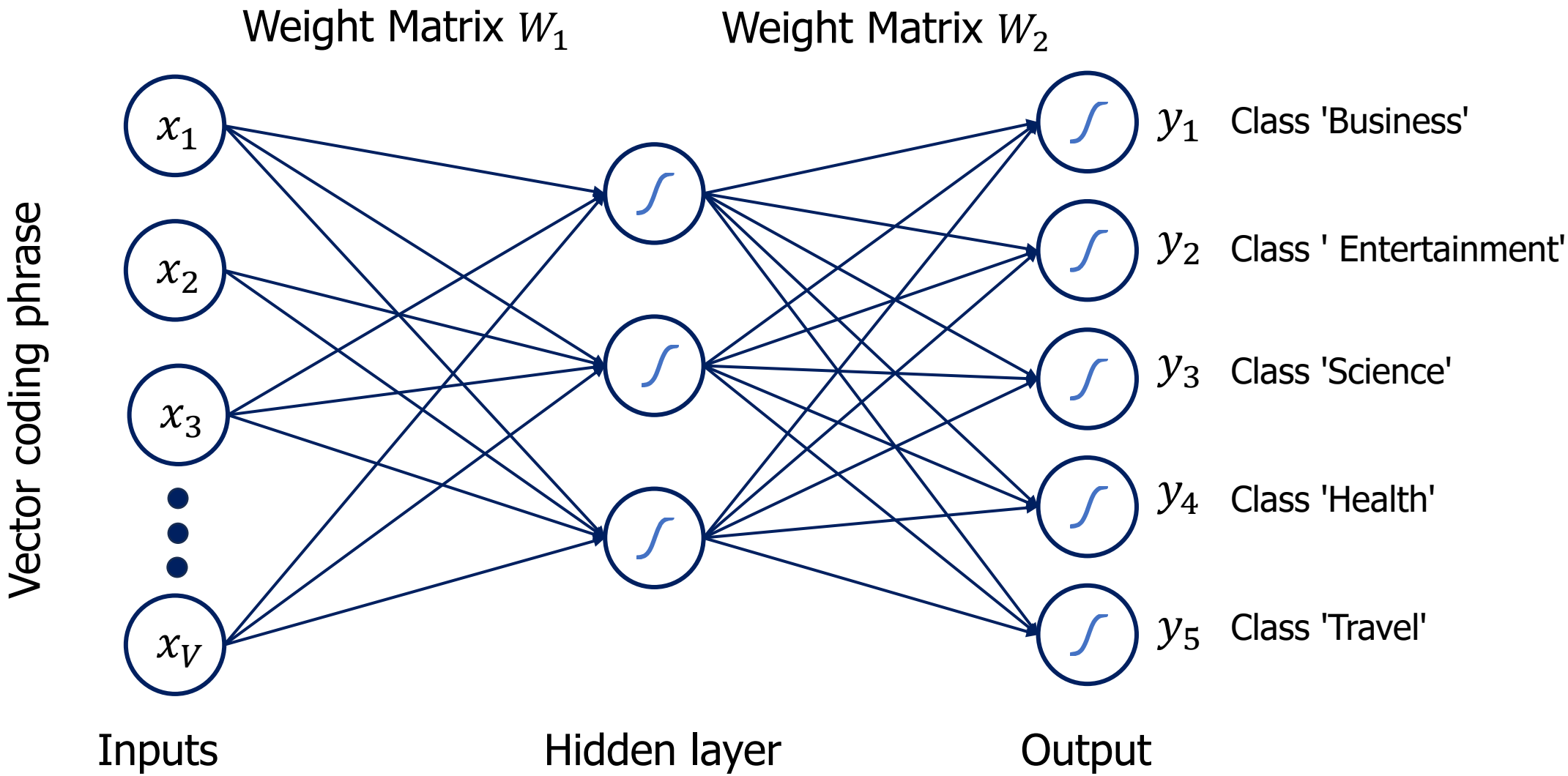
Datasets Used for Training. Coding

Input	Labels
«Extremely rude staff. They don’t resolve any customer issues effectively and I had a bad experience with this company! I don’t highly recommend them to anyone looking for quality products and excellent service.»	Negative Review
Bag-of-Words Representation: (0, 1, 0, 0, 1, 0, … , 0)	0
«Five stars all the way! The service, the product, and the overall experience were outstanding.»	Positive Review
Bag-of-Words Representation: (0, 0, 1, 0, 0, 1, … , 0)	1

Feed-Forward Networks: Binary Classification

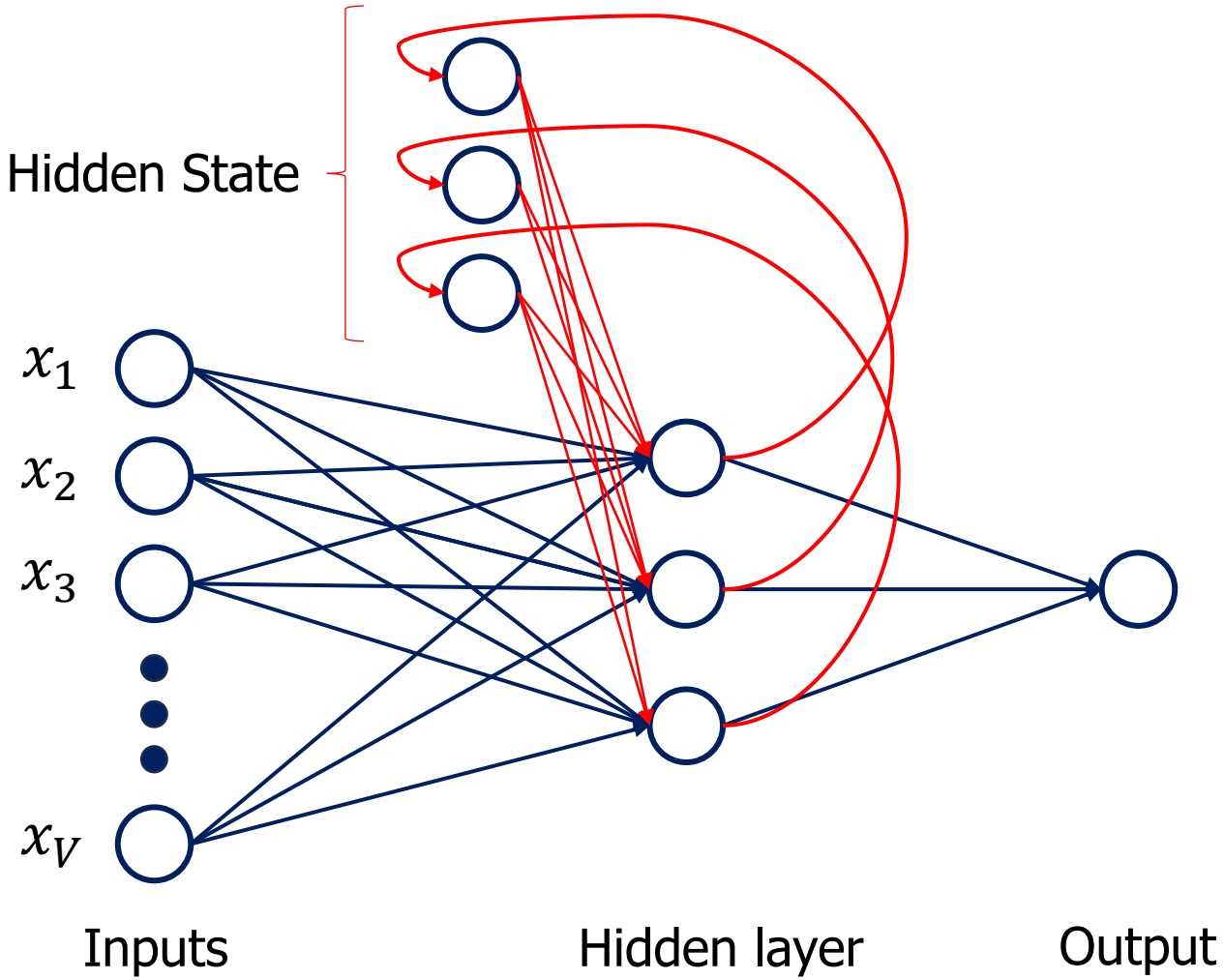


Feed-Forward Networks: Multi-Class Classification



Recurrent Networks

- ✓ Time-Series/ Sequence Processing
- ✓ Feedback
- ✓ Hidden State – information about previous activations



An Example of Dataset 2

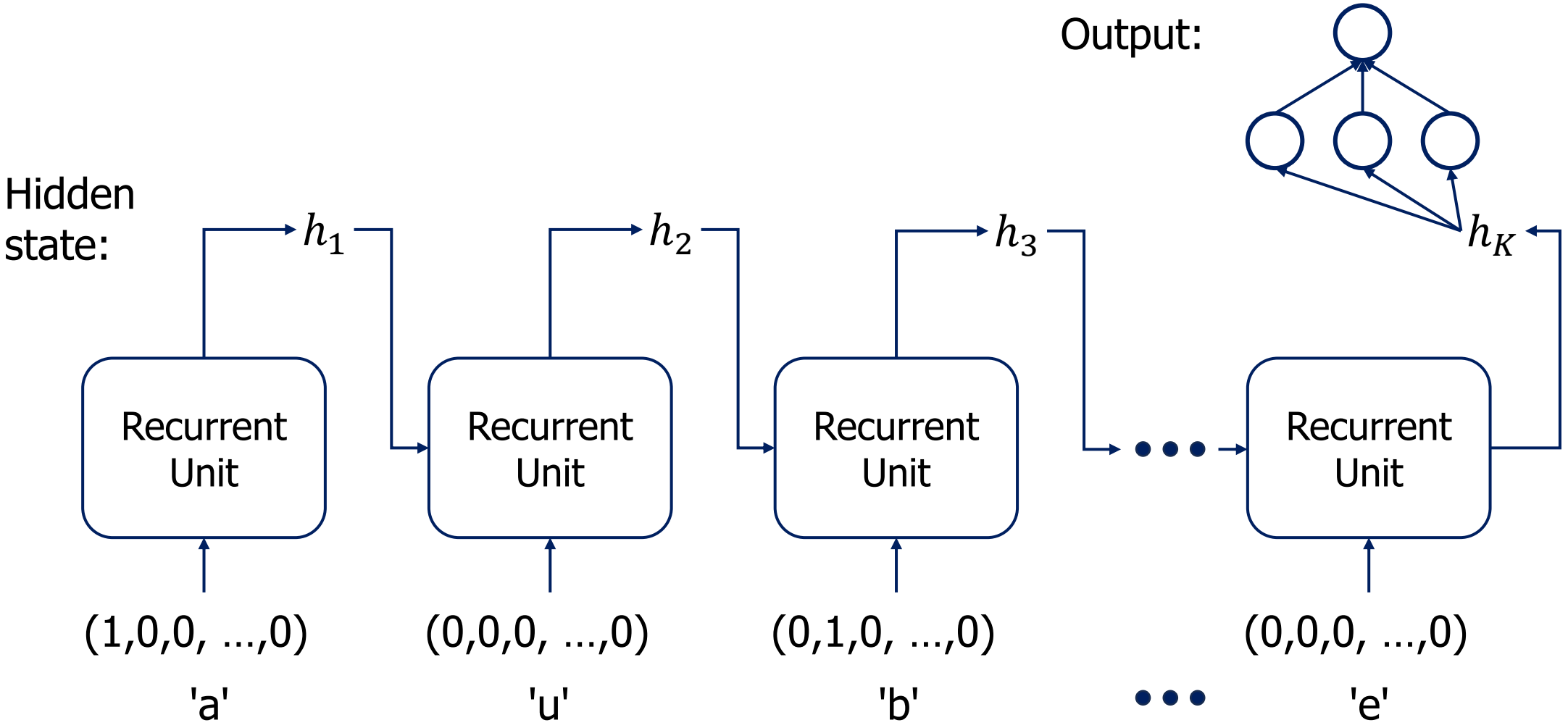
Names: 'Aubree' → Female, 'Jane' → Female, 'Ann' → Female, 'Jack’ → Male

'Aubree' (Tokenization) → 'a', 'u', 'b', 'r', 'e', 'e'

'a' → 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'u' → 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
'b' → 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'r' → 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
'e' → 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'e' → 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Class Labels: Female → 1, Male → 0

Recurrent Network Activation



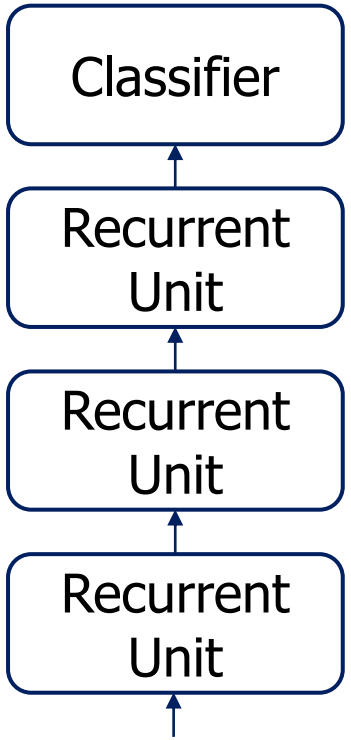
More Difficult Dependencies in Data

joined_hidden_state =
hidden_state1 & hidden_state2

hidden_state2←'a'←'u'←'b'←'r'←'e'←'e'

'a'→'u'→'b'→'r'→'e'→'e'→hidden_state1

Bidirectional Representations



Stack of Recurrent Units